PYTHON DATA SCIENCE

THE ULTIMATE GUIDE ON WHAT YOU NEED TO KNOW TO WORK WITH DATA USING PYTHON

Python Data Science

How to Work with Data with Python Programming Language The Ultimate Guide on What You Need to Know to Work with Data Using Python

BY

Oliver Soranson

Table of Contents

Introduction

Congratulations on purchasing *Data Science Python* and thank you for doing so. The following chapters will discuss all the steps that we need to use to start our Data Science project and finally get some insights and good information out of all that data we have been collecting. Even better, we are going to take a look at how we can complete this project with the help of the Python programming language!

To start this guidebook, we are going to take a look at some of the basics that come with the Python language, and how it can work so well with the process of Data Science. We will also add in some information about what Data Science is all about, and how Python and Data Science can come together to provide us with amazing results in the process. We can then spend some more time on the Python language and what comes with it before moving on to more about getting started with Data Science.

Next on the list is a look at some of the best libraries that we can use to help handle our Data Science project. We will start out with a look at the NumPy library, and the Pandas library, since these are the two most commonly used programming libraries to help with the different parts of a Data Science project. We can later expand out to some of the other options, like Jupyter and TensorFlow as needed.

With this information under our belt and some of the Python libraries set up and ready to go, it is time to take a look at some of the different parts of the puzzle we can explore in the Data Science project. We will look at the basics of collecting and preparing the data, working with data cleaning and preparation, what is data wrangling, how to take all of our information and plot it to make a visual, how to work with data aggregation and group operations, and a look at what the time series is and how it relates back to our work in Data Science.

To end this guidebook, we are going to take a look at a few other topics as well, ones that will ensure that we get a full understanding of Data Science and the steps that we need to take to make this process work. We will take an in-depth look at Machine learning and how it fits in with Data Science, and even explore some of the practical examples of Python Data Science at work so we can finally see the results that we want.

There is so much that we are able to do with the help of Data Science, and when we put it to work, with the help of the Python Programming Language, we can really dive deep into our data and learn some interesting insights that never were available to us in the past. When you are ready to learn some more about Python Data Science, make sure to check out this book to get started.

There are plenty of books on this subject on the market, thanks again for choosing this one! Every effort was made to ensure it is full of as much useful information as possible; please enjoy it!

Chapter 1: Why Python Works So Great for Data Science

The first topics that we need to take a look at here are Python, and how it can work with Data Science. There is so much that we can learn about the Python language, and because of all the extensions and libraries that come with this coding language, we can put it to work with helping out with the process of Data Science. There are a lot of parts that come together for each of these topics, so let's dive right in and see how they both works, separately, and together.

The Basics of Python

The first topic that we need to spend a little bit of time on is the Python Coding Language. As we go through some of the different processes that come with Data Science and other similar topics, it is likely that you will need to create a model. These models are useful because they can be trained to take on data and provide the results that you need in the process. And while there are some choices that you can make when it comes to working with these models, Python is one of the best options out there.

Python is simple and easy to use coding language. It was designed to be fun and to open the world of programming and coding up to novices and beginners. Just because it is easier than some of the other coding languages to learn, though, doesn't mean that you aren't getting a strong and powerful option from the start. Thanks to some of the power that does come from the Python language, we are able to work on complicated tasks and codes, including Machine learning and Data Science.

With this in mind, we need to take a look at some of the benefits of Python, along with some of the basics that come with this coding language over some of the others. The first benefit is that this coding language is easy to learn. The whole design of this coding language was to make sure that even beginners would be able to jump on board with coding and see some great results in the process. Whether you just want to add in a new coding language to your arsenal, or you have never done any coding in the past, the Python language is the right choice for you.

Even with all of the ease of use, Python has a lot of power behind it. Many people are worried that getting started with this language because they think it will be too easy, and it won't be able to handle some of the difficult tasks that are needed with Machine learning and Data Science. But when we add in the right functionalities and the right libraries, along with the Python language, we are able to make this all work well for our needs, and Python will have all the power for programming anything.

There are a number of libraries that come with the Python language that make it stronger and helps you to accomplish all of the tasks that you want. The traditional library that comes with the download of Python is going to have a lot of neat classes and functions to it as well, and many of the codes and programs that you want to write out will work just fine with this.

However, there may be times when the library just doesn't have the functionality that you are looking for. The traditional Python library is not good at handling some things like scientific work and mathematics. This doesn't mean that you are out of luck though. It simply means that we need to find another Python library that we can add to the Python language and get these types of projects done. There are a lot of options here, so you just need to find the one that will help you to finish up your own project.

Another benefit that comes with the Python language is that it is considered an Object, Oriented Program, or OOP. This means that the language is going to rely mainly on classes and objects to keep things organized and to ensure that all of the parts are able to stick together and stay in the locations where you want them. This makes the language and all of the codings that you do with it easier to work with overall and will ensure that you see some great results in the coding that you do even as a beginner.

The Python language has a large community that you can rely on when things get tough to code. Sometimes, a new programmer is going to work hard on their code, and it is just not turning out the way that they want. Or maybe there is something brand new that they want to learn, and they are not sure how to make this happen. The fact that Python has a large community is a big benefit for these very reasons.

Because of all the benefits that come with Python, and because so many programmers throughout the world are working with this coding language, it has developed a large community over the years. This is great for those who

get stuck on a project, who have questions, or those who just want to learn something new along the way.

This coding language also does a great job of helping out on more complex coding projects including Data Science and Machine learning. Even with all of the other choices out there, Python is still the number one choice to handle Data Science and Machine learning tasks. We are able to bring out Python to provide us with all of the models that are needed for Data Science, and since many of these models are created with Machine Learning in mind, it makes sense that we would be able to use this coding language to help with these kinds of projects as well. No matter what data you need to get through, or what you are hoping to find while creating your model, the Python coding language is going to help us to get it all done.

There is just so much to love when it comes to working with the Python language, and that is why it is often the thing that is brought up any time that we need to create a model with Machine learning. Regardless of the questions that you need to be answered, or the business problem that you want to solve, Python is one of the easiest ways to get it all done.

What Is Data Science All About?

Now that we know a bit about the Python language, it is time to explore how it can be used with the process of Data Science. And the best way to explore this a little bit is to see what Data Science is going to be all about. This can give us a better understanding of how these two topics are going to work with one another, and why a business would want to work with Data Science in the first place to help them succeed.

When we look at Data Science, we will see that as time goes on, it continues to evolve as one of the most promising and in-demand career paths for those who are skilled professionals. This says a lot about how companies all over are working with this process when they are working with all the data they have collected. And today, those data professionals who are successful understand that they have to be able to go past some of the traditional skills of analyzing large amounts of data, programming skills, and data mining.

In order to uncover some of the useful intelligence for their organizations, data scientists need to master the full spectrum of the Data Science life cycle, and it needs to possess a level of understanding and flexibility to maximize

returns for each of the phases of this process.

As the world has entered into a time of big data, the need for a lot of storage has grown as well. This is actually one of the machine challenges and concerns for many industries until about 2010. The main focus of this was to build up a framework and some solutions to storing all of this data that weren't hard to use or too expensive. Luckily, there are now a few different frameworks out there that can handle this kind of process, which makes it easier to focus more on how to process all of that data, rather than on where to store it.

Data Science is basically the process that we need to follow in order to gather, clean, wrangle, analyze, and visualize our data. The hope of doing this one is to learn some of the best predictions and insights that are found in all of that data. This can help a company to keep ahead of the competition, figure out which products to release, how to find a new niche, how to increase customer satisfaction and so much more.

Now, it is time to take a look at some of the reasons that we need Data Science. Traditionally, the data that we worked with would be smaller in size, and it would be structured. This allowed us to use some simple tools from business intelligence to help us analyze all of this information. Unlike the data that is found with some of those traditional systems, which is mostly unstructured, most of the data that companies are able to get ahold today is going to be semi-structured or unstructured. And there are often large amounts of data that we need to focus on as well.

This is going to complicate things quite a bit. When the data is so large, and we are not able to get it all in a structured form, this is going to make it harder to read through the data and get any insights out of it. But with the help of Data Science and a good data analysis in the process, you will find that we can still get through this data, but we need to use different methods.

The data that we work with to gather predictions and insights will come from a lot of different sources, and the sources that you work with will often depend on what kind of question you would like to answer or what business problem you would like to solve. Simple tools of BI are not going to be able to process the huge volume and variety of data. And this is why we need to make sure that we pick out some analytical tools that are more advanced and complex. This ensures that we are able to process, analyze, and draw insights

out of the data we have.

Of course, this is not going to be the only reason why so many companies are starting to take a look at Data Science. Nor is it the only reason that this Data Science has become so popular. Some of the reasons that companies are going to rely on this Data Science, and all of the tools and techniques that come with it will include:

Being able to learn the precise requirements that your customers have with your business. This is possible if you just use the existing data you hold onto, such as the customer's age, income, browsing and purchasing history and learn from this. This is all data that you have been collecting for some time now. But with Data Science, we are able to take all of that data and use it to train our own Machine Learning models more effectively. Over time, and with the right kind of data, it is easier to use the model to effectively recommend new products to customers, with the precision that you need to make the sale.

That is just one example of what we are able to do when working with Data Science. We can look at the idea of the smart car that is developed by Google right now. These kinds of cars have to collect a ton of data, from many sources, in order to learn how to behave on the road. A self-driving car is going to collect live data from things like lasers, radars, and cameras to create a good map of its surroundings.

Based on the data that the self-driving car is able to read from, the car will then learn when to make different decisions. It will learn when to go faster or slower when to overtake other cars, were to make a turn, and more. And all of this is done with the help of your big data, and some great Machine Learning algorithms.

Data Science is also a big part of a process that is known as predictive analysis. A good example of how this one works is with weather forecasting. Data that comes from a variety of sources, such as ships, radars, aircraft, and satellites, can be collected and then analyzed before building up some weather models. These models, once they have gone through some of the proper training, will not only be able to forecast the weather for us, but they will be great at predicting the occurrence of a natural calamity and the likelihood that this calamity is going to happen. If it is used in the proper manner, and it is trained to show some accuracy, it could help us to take the

right measures ahead of time to save many lives.

We have spent a lot of time talking about Data Science so far in this section, but we need to go a bit further into some of the basics of this process, and not just how we can benefit from this process. Many companies are increasingly using the term of Data Science, but it is hard to know exactly what this term means? What are the skills that are required to become a data scientist? And is there a difference between Data Science and business intelligence? And how are we meant to work on the decisions and predictions within this process?

These are just a few of the questions that can be raised when it comes to working with Data Science. But first, Data Science is going to be a blend of principles of Machine learning, algorithms, and other tools that all have the goal of discovering some of the hidden patterns that can show up in the data. The difference that comes between the process of Data Science and what statisticians have been doing for years is found between explaining and predicting.

Data Analysis is going to spend their time explaining what is going on by processing the data and learning the history that comes with it as well. But a data scientist is not only going to do what is called an exploratory analysis to learn the insights in the data, but it will also use a variety of algorithms from Machine learning to identify the occurrence of a particular event in the future. This means that a data scientist is going to be able to look at the data from a lot of angles to see what is there, and sometimes these angles are brand new.

So, to keep it simple here, Data Science is primarily going to something that we use to make good decisions and predictions. And this is all done with the help of several different tools and techniques including Machine learning, prescriptive analytics, and predictive causal analytics. Let’s take a look at each of these and see how they will help us through the process.

The first type that we are going to take a look at is known as Predictive Causal Analytics. If you are working with a model and you want it to predict the possibility that an event is going to happen in the future, then we need to apply the Predictive Causal Analytics to it. A good example of this is when a bank provides money on credit. They would use this model to help them figure out the probability of the customer making their credit payments on time. If they are highly likely to pay the amount each month, and on time,

then they are more likely to get the loan.

With Predictive Causal Analytics, we can build up a model that will perform the process of predictive analytics. But these analytics are going to be done on the payment history of the customer, allowing the bank or the financial institutions to predict of the payments on this debt will be paid out on time or not. It may not be 100 percent accurate because people are hard to predict all of the time. By analyzing some of the past behavior of that particular person, though, we are able to get a better idea of how they are going to behave in the future.

Then we have prescriptive analytics. If you are working on a project and you want to create a model that has the intelligence of taking its own decisions and the ability to modify it with some dynamic parameters, then it is time to pull out the prescriptive analytics to get it done. This is a new field that comes with Data Science, and it is all about providing advice based on the data that you have.

This model is not only going to predict for us, but it is going to help suggest a range of prescribed actions and shows the associated outcomes that we can get with it. The best example that we can look at for this one is the self-driving car from Google. The data that the vehicle is able to gather is going to help it to learn how to drive on its own. You are capable of running some algorithms on this data to ensure there is some intelligence brought into it. This enables the car to take some decisions, like when to turn, which path to take, and when to speed up or slow down.

Another option here is Machine learning. Machine Learning can come in at several points of the process, but first, we are going to look at how Machine Learning can be used to make some good predictions. If you are working with a lot of transactions, such as what we see with a financial company, and we want to build up a new model that can determine a future trend, then the algorithms that come with Machine learning are some of the best to make all of this happen.

These Machine Learning Algorithms are going to fall under the category of supervised Machine learning, basically, because we already have the data that we want to base our models on, and that we can use to train all of our machines. For example, a model that detects fraud is going to be trained using some of the historical records that are available for purchases that

turned out fraudulent in the past.

Another way that we can use Machine learning is with pattern discovery. If you are working on a model or a project, but you don't have any parameters based on which you can make some predictions, then it is time to do some work. We need to still go through our set of data and find some of the hidden patterns that are inside. These patterns are going to be used to help us make some meaningful predictions overall.

This is a good example of what we are allowed to do with an unsupervised Machine learning Model as there aren't going to be any predefined labels for the grouping when we get started. There are a lot of algorithms and techniques that we can use for this one, but one of the most common to help with pattern discovery will be Clustering.

Let's say that your job is to work with a telephone company and you want to be able to expand out your network. Your goal is to put towers in a new region and you must make sure that the towers are placed in the right areas and locations in order to reach the maximum number of customers at the same time. The Clustering Machine learning Algorithm is going to be the perfect technique that we can use to find these tower locations while ensuring that all of the customers receive an optimal amount of signal strength.

We are going to go through some of these steps a little bit later in this guidebook, but it is important to know the steps that come with completing our own Data Science process. Some of the main steps that are found in data analysis, and in the process of Data Science, will include:

1. **Discovery:** This Is the time where we look for the data that we want to use, often from a variety of sources.
2. **Data preparation:** Since the data comes to us from many different sources, we use this stage to prepare it and organize everything so it works with the model later on.
3. **Model planning:** This is where we are going to determine the best techniques and methods that we can use on the data, helping us to draw up the best relationships between the variables.
4. **Model building:** With this phase, we are going to develop the sets that we need for training and the ones that we need for testing.

This ensures that our model is going to be accurate and work the way that we want.

5. **Operationalize:** With this phase, we will be able to deliver some of the final reports, briefings, code, and technical documents, and more.
6. **Communicate the results:** At the end of it all, we need to be able to come out and communicate what we were able to find within the data. Often this is done with some visualizations, such as a graph or a chart, to get the work done.

Data Science can easily pair together with the Python Coding Language in order to get things done. Python is adept at handling the large amounts of data that are needed for a Data Science project. This language will help us to clean and organize the data, plan out our model, and even build up the model so that we can get the best results and really learn what insights are out there for us within all of this data.

There are many companies that are going to rely on the basics of Data Science and all that comes with it. There are a lot of processes that come with Data Science, and it isn't just about gathering up the data. This is a good place to start, but we also need to spend some time exploring how to clean and organize the data, how to wrangle the data, working with Machine learning and various algorithms within it to create models to look at the data and provide you with insights and predictions as needed. We also have a step that will take some time working with data and making it appear in a graph or chart, ensuring that it is easier to understand complex types of data in just a few minutes.

There are a few other coding languages that can do the same kind of thing. But none of them are going to combine the ease of use, and the power, that we can find with the Python Coding Language. When we can add all of these together, it becomes so much easier to really get through that data in a quick and efficient manner and to ensure that we will see the results that we want.

Chapter 2: The Basics of Python That Everyone Needs to Know

Before we dive into some of the other things that we need to know to get started with Data Science and how Python is able to handle some of the complexities that come with Data Science, it is now time for us to take a look at some of the basics that come with the Python coding language. There are a lot of simple parts that are in this library, and being able to learn about them, and understanding when they will show up in some of the codings that you do, can make you more efficient as time goes on as well. With that said, let's dive in and learn some of the basics of the Python code to make it easier to get started.

The Statements

Statements are pretty simple in Python. These are just the strings of code that you write out and that you want the compiler to list out on the string. When you tell the compiler the instructions that you want it to work on, you will find that those are the statements of your code. As long as you write them out properly, the compiler will read them and show up the message that you want on your screen. The statements can be as short or as long as you would like, depending on the code that you are working on.

The Comments

As you are writing out the code, you may find that there are times you want to include a little note or a little explanation of what you are writing inside the code. These are little notes that you and other programmers are able to read through in the code and can help explain out what you are doing with that part of the code. Any comment that you write out in Python will need to use the # symbol ahead of it. This tells the compiler that you are writing out a comment and that it should move on to the next block of code.

You can add in as many of these comments as you need to explain the code that you are writing and to help it make sense. You could have one very another line if you would like, but you should try not to add in too many or you may make a mess of the code that you have. But as long as the # symbol

is in front of the statement, you can write out as many of these comments as you would like and your compiler will just skip out of them.

What are Classes?

There are a lot of different things that we can talk about when it comes to the Python Language. But one of the most important topics that we are going to take a look at is the classes in Python. These classes, to keep things as simple as possible, are going to be containers that can hold onto the objects and other parts of the code. You have to take some time to name all of the classes in the proper manner and put them in the right spots of the code to help get them to work the right way, and of course, we are going to take each of these classes with the right objects to get the code to work.

It is possible to store pretty much anything that you would like inside a new class that is designed. But the major rule with this one is that you have to make sure that the objects you choose for one class are pretty similar in one way the items don't have to be identical to one another, but when someone looks inside one of these created classes and understand why you put all of those objects together in that class.

For example, you don't have to just put cars into the same class, but you could have different vehicles that show up in the same class. You could have some items that are seen as food in one class. You can even have items that are all going to be the same color. As the programmer, you are able to get some freedom when creating the class and storing objects into these classes, but when another programmer looks at the code, they should understand how all of these items or objects go with each other.

Classes are going to be very important when it is going to come time to write out the code. These are going to hold onto the various objects that you want to write into the code and will ensure that each of the parts will be stored properly. They will also make it easier to call out these objects, and other parts of the code when it is time to use these during the execution of the code.

The Operators

Another thing that we can work on with some of our codings are the

operators. These are often ignored or forgotten because we assume that they are not that important or that they play too small of a role in the work that we are trying to do. But while the operators are simple to use, this does not mean that we should ignore them and not get the full benefits from them as well.

There are several different types of operators that we can work with on a regular basis. We can work with the arithmetic operators to help us add together, or subtract, or do another mathematical equation on our code. There are the logical operators, the comparison operations, and more, that can help us make sure that we will get the results that we want.

Assigning a Value to a Variable

One thing that we can work with while writing some codes is assigning a value to a variable. A variable is basically going to be a piece of storage on the computer. We can leave it blank, but at some point, to make sure that our code is going to work the way that we would like. This means that we need to be able to take some kind of value and assign it back to the variable, or that space in the memory, so we can use it later on.

The process of adding value to the variable is pretty easy. You just need to use the equal sign between the two. So, pick out the value that you want to assign over to a specific variable, and then make sure that the equal sign shows up between the two. You can technically add in as many values to the same variable if you would like, as long as we make sure that the equal sign is there.

The Control Flow

The idea of the control flow in the work you do with coding is not meant to be too difficult to work with. It is just going to tell us how to read through the codes we write, to ensure that the compiler is going to be able to handle the different parts. To keep with the ease that we talked about earlier with Python, we have to remember that this language is going to write out the codes from left to right. We will look at some of the examples of how this code will look later on, but you will find that the codes are written out in the same manner that we see the words on a page in a book or on a website.

The Python Functions

Another topic that we need to take a look at when it comes to working on our codes in Python is known as the functions. These functions are basically a set of expressions that can be like statements as well. These are going to be some of the very first-class objects that we are going to work within the code, which is a great thing because we will not have to add in a lot of restrictions in order to use these as we go. You will be able to work with the functions in a similar manner that we can with some other values, like strings and numbers, and they are going to have attributes that we are capable to bring out with the prefix of dir.

Now, these functions are very diversified and there are many attributes that you can use when you try to create and bring up those functions. Some of the choices that you have with these functions include:

- **__doc__:** This is going to return the docstring of the function that you are requesting.
- **Func_default:** This one is going to return a tuple of the values of your default argument.
- **Func_globals:** This one will return a reference that points to the dictionary holding the global variables for that function.
- **Func_dict:** This one is responsible for returning the namespace that will support the attributes for all your arbitrary functions.
- **Func_closure:** This will return to you a tuple of all the cells that hold the bindings for the free variables inside of the function.

There are different things that you can do with your functions, such as passing it as an argument over to one of your other functions if you need it. Any function that is able to take on a new one as the argument will be considered the higher-order function in the code. These are good to learn because they are important to your code as you move.

These are just a few of the different topics and options that you can work with when it comes to writing out codes in the Python language. Learning how all of these can work and how we can use them to write our own codes. With this in mind, let’s move on to some of the other things that we need to work with when it comes to creating our own models of Data Science.

Chapter 3: The NumPy Library and How It Can Help with Data Science

The library that we are going to start our journey with Data Science is the NumPy Library. This is going to be an open-source package of Python. It can be used for scientific and numerical computing. In many cases, though, it is a library that is used for more efficient computation on arrays when it is needed. This library is going to be based, as well as written, on two main programming languages, including Python and C, and the programmer can decide which of these two languages they would like to use at any time.

While NumPy can be used with the C Coding Language, it is considered a Python package, and the name of this library stands for Numerical Python. There are a variety of things that you are able to do with this kind of library, but it is going to be used in many cases to help us process multidimensional arrays that are homogeneous. It is going to be one of the core libraries to help out with some of the scientific computations.

Hence, we can see how this library is going to be powerful when it comes to multidimensional array objects and integrating tools that can be useful any time that we want to work with these arrays. It is important in almost all of the scientific programming that we can do with Python, whether this includes a project in bioinformatics, statistics, and Machine learning.

The NumPy Library is also going to provide programmers with some good functionality that is written out well and will run in an efficient manner. It is mostly focused on helping us to perform some operations that are mathematical on a contiguous array, much like what we see in lower-level languages that handle arrays as well, such as C. To make this easier, it is going to be used to handle the manipulation of numerical data.

Thanks to this language, there are a lot of times when we can add in this language to our coding, and get more work done. This is a very good computational library to work with and specifically works with arrays and projects that are more scientific. With this in mind though, Python can also be used as an alternative to MATLAB with the help of the NumPy Language.

One thing that we will notice when it comes to using the NumPy language is that it is really going to be one of the most used. This is due to the fact that

Data Science techniques need all of that work done on matrices and large-size arrays. And some of the heavy numerical computation needs to be done as well to extract out the predictions and the insights we are looking for. There may be a few methods that we can use to make this happen, but all of the mathematical functions that are needed for it fall under the NumPy library, so why not just use that to make things easier?

Some programmers worry that the NumPy Library is too basic and won't be able to meet all of their needs with Data Science. Yes, this library is one of the most basic of the Python statistical and scientific libraries, but it is still considered one of the most important when it comes to working with Python and scientific computing.

We can also add to this that there are a lot of other libraries that work with Data Science, Machine learning, and other tasks, that will depend on the arrays from NumPy as their basic inputs and outputs. This means that if we ignore all that the NumPy Library is able to accomplish, then we are missing out on a lot of other neat things that we can do with this library as well.

To add to all of this, the NumPy Library is going to provide us with some neat functions that will allow a developer to perform basic and sometimes more advanced functions of statistics and mathematics on multi-dimensional arrays and matrices with fewer lines of code to get it all done. You will find that the n-dimensional array, which is going to be seen by the name of Ndarray, is the main functionality that comes with NumPy. These are going to be homogeneous arrays and all of the elements that come inside that array will need to be of the same type for this to work.

Now, if you have worked with Python in the past, you may see the array from NumPy, and think that it is simply the same thing as a Python list. And there are some similarities that come with the two of these. One thing to note with it though is that the arrays of NumPy are going to be much faster than what we can find with a Python list. However, the Python list is going to be more flexible than the arrays because you can only store the same type of data in each column. You have to decide which one is the best for your projects or your models and work from there.

There are a lot of different features that come with the NumPy Library as well, and many of these are going to assist Python in providing some Data Science help as well. Some of the top features that come with this particular

Python Library will include:

1. It can help us reshape our arrays. Because it is so good at doing this, it often allows Python to become a more efficient, and easier to use, alternative to MATLAB.
2. It provides us with a variety of functions for arrays.
3. It includes some homogeneous arrays that are multidimensional. The Ndarray is going to be a part of this as well, which is just going to be a n-dimensional array.
4. It is able to combine the Python and C languages together, which can add in some more flexibility and power to the coding that you do.

The next question that comes with this library is why we should consider using this language for some of our scientific computing, or our coding needs. There are a few options here but remember that the array from NumPy is sometimes considered the same thing as a list from Python. The good news is there are three main reasons why we would choose to work with the array rather than the list, and that includes that there is less use of the memory, the array can perform faster, and they are more convenient to work with.

The first reason with this one though is that the array is not going to take up as much space in the memory as we see with the Python list. Then, it is going to also be pretty fast when it is time to execute the codes and models that you want to use. And finally, the arrays are often a lot easier and more convenient to work with compared to some other options, including Python lists, so adding these into your toolbelt can be important.

There are many advantages that will show up when we choose to work with the NumPy library with Python, rather than some of the other options. Keep in mind though that there are a ton of different coding libraries out there that also work with Python, so it is all going to depend on what works the best for you, and what kind of process of coding you would like to work on the most. With this in mind, some of the advantages that we can enjoy when it comes to working with the NumPy library includes:

The NumPy array is going to take up less space than we see with some of the other options, including the Python list. The arrays that come with this library

are going to be smaller than the Python list. To start, the Python list has the ability to take up at least 20 MB of space on the memory, based on how big the list is to start with. And then the array will take up much less at just 4 MB. The arrays are often going to be a lot easier to access when it is time to read them or write them out.

And the second benefit that comes with this library is that the performance of speed is great. The NumPy array is able to perform computations at a much faster rate than what we will see with the Python list. Because this library is open-sourced, it is not going to cost anything to use it, and it is going to rely on Python to get things done. And since we can combine the C code, especially existing codes from this language from before, it is easier than ever to get some important projects done.

Now, as we can see from this chapter, NumPy is going to be a really strong library when we look at all of the high-quality functions that this library holds onto. Anyone is able to perform large computations or calculations, and thanks to this library, it can all be done with just a few lines of code. This is one of the things that makes the NumPy library perfect for those who want to work on numerical computations. If you are going to do some work with Python and Data Science, then it is worth your time to download the NumPy Library and get familiar with some of the features and functions that come with it.

Chapter 4: Manipulating Data with the Pandas Library

The next library that we are going to take a look at is known as the Pandas Library. This library is one of the best when it comes to Machine learning and Data Science, and will stand for Python Data Analysis Library. According to many sources on this library, Pandas is going to be the name because it is derived from the term of panel data. This is basically an econometrics term that handles data sets that are multidimensional in structure.

Pandas are going to be seen as a bit game changer when it is time to analyze the data that you have used the Python Language, and it is often going to be the number one Python Library to use when it is time to handle data munging and wrangling. It can also handle all of the other aspects of Data Science that you want as well, making it an all in one library for your needs. Pandas are also going to be open-sourced, free for any programmer to use, and is one of the best Data Science libraries to focus on.

There are a lot of cool things that come with the Pandas Library, so taking some time to look it over and figure out what it all entails will help you with a lot of Data Science projects. One thing that is cool with Pandas is that it is able to take almost any kind of data and will then create an object in Python with rows and columns. These are going to be called the data frame and it is going to look pretty much like what we see with Excel. If you have worked with the R Programming Language before, then you will see some similarities here as well.

However, compared to working with the dictionaries or lists that come with Python, or through loops or list comprehensions, the Pandas Library is going to be so much easier overall. The different functions that come with Pandas can make it a much easier library to work with, especially when it comes to some of the complexities of working with Data Science.

Installing the Pandas Library

The next thing that we need to take a look at here is how to actually install the Pandas Library and get it all set up. To install this library, we need to

have a Python version that is at least 2.7 or higher. The Pandas Library is not designed to work with any of the older versions of Python, so if you have one of the older versions, it may be time to upgrade. At this time, you need to make sure that some other deep learning libraries are in place. Pandas are going to be dependent on a few other libraries, based on what you would like to accomplish. It really needs to have at least NumPy associated with it, and if you want to do something like plotting with your information, then you need to work with Matplotlib.

Because you need a few extras that go with this library, we may want to consider installing a package to make sure that all the extras are there when you need them. The Anaconda distribution is a good option to work with and it can work on all of the major operating systems including Linux, OS X, and Windows systems.

Pandas is able to work with the Python IDE, including options like Spyder or the Jupyter Notebook. But to get these to work, the Pandas library has to be installed and ready to go. The Anaconda extension will come with both of the IDE or Integrated Development Environment, so that can make things easier to handle.

Importing one of these libraries means that you need to first load it into your memory, and once the installation is all done, you will be able to open up the needed files and work with them at any time. to make sure that you can import Pandas in the right manner, all that you need to do is run the following code below:

- *Import pandas as pd*
- *Import numpy as np*

If you would like to make some of the codings that you do a bit easier, you would add in the second part (as pd) because it allows you to access Pandas with just the pd.command, rather than having to go through the process of writing out pandas.command each time that you wish to use it. As we can see with the code above as well, you need to import NumPy at this time. NumPy is a useful library to work with any scientific computing in Python, and often the Pandas library will need to pull out functions and other parts from this to get things done. At this point, Pandas is up and running and ready for us to use.

The Benefits of Using Pandas

With that work done, it is time to take a look at some of the many different benefits that come with using the Pandas library. There are a lot of benefits to this one, and it is one of the most popular options that come with this kind of Data Science work. With this in mind, let's take a look at some of the benefits that we are able to see with the Pandas library.

The first benefit that comes with the Pandas library is the data representation. Pandas are going to provide programmers with a streamlined form of data representation. This is going to be important as you analyze and work to understand the data that you hold onto a bit better. When you can simplify some of the data representation that you have, it is going to facilitate better results for some of your projects in Data Science.

The second benefit of this library is that it provides us with a way to get more work done, without having to do as much writing. This is actually one of the biggest advantages that we are able to see with this library. With the traditional form of Python, we may have taken many lines of code to get the work done, without any support libraries, but with Pandas, we can get that same work done in just one or two lines of code. This means that by using Pandas, we can shorten up the procedure of handling the data that we have. When we can save all of that time, we are allowed to focus more on the algorithms that we need for the data analysis.

An extensive set of useful features is next on the list of Pandas benefits. Pandas are going to be seen as really powerful in the coding world. They are able to provide us with a big set of commands and features that are important, and which can be used to easily look through the data and analyze it. We can use Pandas in order to perform various tasks including filtering out the data based on conditions that we set, or segregating and segmenting the data according to the preferences that we would like to meet.

The next benefit that comes with working with Pandas is that this library is able to handle a large amount of data in an efficient manner. When the Pandas Library was originally created, its goal was to handle large sets of data in an efficient manner. Pandas can really help us to save a lot of time and hassle because it can import large amounts of data quickly and efficiently.

The Pandas Library is also able to make data customizable and flexible.

There is a huge set of features in Pandas that can be applied to the data that you have. This can be great for beginners because it helps us to customize, edit, and pivot that data according to what we want to see happen. This is going to ensure that we can get the most out of our data each time.

And finally, the last benefit that we will see with the Pandas library is that it is made for Python. Python is one of the biggest and most sought-after programming languages in the whole world, and it has an extensive amount of features that we can enjoy. And with just the amount of productivity that is offered, it is no wonder that many people want to learn how to code in this language.

Because of this, and all of the great features that come with Python, the fact that we are able to code with the help of Python in Pandas is going to be a great thing. It allows the programmer to tap into the power of many libraries and features that work with Python, which adds in some of the strength and power that we need with our coding.

Now, there are a few disadvantages that come with this library compared to some of the others, but often there are ways to work around these. Some of the disadvantages that can come with the Pandas library that programmers need to be aware of include:

1. **The learning curve is steeper:** Pandas was thought to have mild learning slow in the beginning. But the more that you explore the library, the steeper the learning curve is going to become. Sometimes the functionality of this library is going to get confusing, and for beginners, this is going to bring on some challenges.
2. **The syntax can be hard:** While Pandas is going to work with the Python language, sometimes it is going to add in some challenges when it comes to the syntax that has to be used. Switching back and forth between Python and Pandas codes can cause some problems.
3. **It doesn’t work well with 3D matrices:** If this is something that you want to work with, it can be a drawback of this library. If you are planning on just creating a 2D matrix, then this will not be a problem at all.

4. **Bad documentation:** Without a good amount of documentation to go along with the project, it can be difficult to learn a new library. The documentation that comes in Pandas isn't going to do much to help us get the harder functions of the library done. This is going to slow down our learning procedure and can make coding difficult.

Viewing and Inspecting the Data

One thing that Pandas is able to do to help with our Data Science project is to work with viewing and inspecting the data. In reality, Pandas is able to help with all of the various processes that you may want to do with Data Science, but right now, we are just going to focus on this part. You can use a variety of the functions that come with Pandas in order to take a look at what is in the data, figure out if there are any missing or duplicate values, and then make the changes as needed to work on your data analysis.

With this in mind, it is also possible for us to get some statistics on the entire series, or an entire data frame. Some of the codes that you would need to use to make this happen includes:

1. **Df.mean():** This one is going to help us return the mean of all our presented columns.
2. **Df.corr():** This one is going to return the correlation between the columns that are in your frame of data.
3. **Df.std():** This one is going to help us see what the standard deviation ends up being between each of the columns.
4. **Df.median():** This one is going to help us see what the median of each column is like.
5. **Df.min():** This one is going to help us see the lowest value that is present in each of our columns.
6. **Df.max():** This one is going to help us see the highest value in each of our columns.
7. **Df.count():** This one is going to help us by returning the number of all of the non-null values that show up in each column of the data frame we are using.

There is just so much that we are able to do with the help of the Pandas Library, and learning how to make all of this come together and work for your needs can make a difference in how well the Data Science project works for you. Depending on the kind of data that you want to sort through, and what your end goal is with the data, you may choose to go with another library to help out with the various parts. But if you want to just go through the whole process of Data Science on your own, without having to switch back and forth between the libraries and the processes that you are using, then the Pandas library is the right one for your needs.

Chapter 5: Collecting and Manipulating Data

When we are working on the process of Data Science, there are a few steps that need to happen in order to actually see the insights and predictions that will offer you some sound business decisions. The first step out of this process is going to be collecting and manipulating the data. You can't go through and complete analysis if you don't first have some data to look through. And that is where this chapter is going to come in and help.

Collecting the data that we need to help train and test our models for use can be so important. Having it set up to find the right information, the information that will help us solve a business problem, or answer some questions, can ensure that the model we choose is going to complete the job. With that in mind, let's take a look at some of the best topics that we need to consider when it is time to collect the high-quality and pertinent information needed in Data Science.

Where Should I Collect Data From?

The first question that we need to take a look at when it is time to work on this step is where a company should collect its data from. This is often going to depend on what your company is hoping to achieve when they work with the data. Are you looking to gather information on what would make your customers happy? Then you may want to go straight to the customer and ask them some survey questions. If you want to learn more about what the competition is doing, then you may want to look for another source to help you with this.

There are a lot of different ways that a company is able to collect data, and many sources that they are able to work with as well. Some of the more popular options that are available for collecting data include industry papers, surveys, research, social media, transactional information, recommendation sites, and anywhere else that would hold onto the data and information that you think will answer your biggest business challenges.

Unstructured vs. Structured Data

While you are searching for the data that you want to use during this process, you are going to run across the idea of unstructured and structured data.

Understanding how these two types of data are similar, and how they are different will make it easier to know which kind you want to collect as you work with this process.

Structured data is usually the kind that is thought of as a more traditional form of data. It is going to consist mainly of many types of text files, and the information is going to be well-organized in the process. Structured data is going to be stored inside a warehouse for data, and it can be pulled out for analysis at any time that we need it. Before we add in the era of big data and all of the new sources of data that we can use, structured data was the only kind available and would help organizations make many of their important business decisions.

Structured data is nice because it is easy for us to digest and really organized. This means that analytics would be possible through the use of legacy solutions in data mining. To be more specific, structured data is going to be made up for things like customer data, including addresses, the names, and contact information of your customers. In addition, businesses can also collect up some of the transactional data they need as their structured data source, which can consist of things like financial information. But this financial information needs to be stored up appropriately to meet compliance standards.

Most companies would prefer to work with this kind of data. It is neat and organized and can be really easy to build a model around and get some good results in the process. Unfortunately, it is not always possible to collect structured data. It takes a long time and can get expensive. Many companies, if they can, will get some structured data, and then combine it together with unstructured data to meet their needs.

The second type of data that we can explore is Unstructured Data. This is a big type of data that is growing more and more, and it is being used by companies in order to leverage new and emerging data sources. These new sources of data can come from a lot of different sources, including mobile applications, the Internet of Things, social media, and more.

Since there is so much diversity that shows up with the sources of unstructured data, it can cause businesses more trouble to manage it than they would have with structured data. Because of this, companies are facing some challenges with their data in manners that they just weren’t before, with a lot

of creativity coming into the mix to pull out the needed insights and data to analyze.

The growth of data, and the maturation of data lakes, and other platforms are all a direct result of companies gathering more unstructured data than ever before. The traditional environments for data warehouses are not able to keep up with the new types of data that companies want to take some time to analyze. Because of this, there are now more storage places than ever to hold onto all of that unstructured data and make it work for you.

While the unstructured data is often harder to organize and read through, it can still provide us with some unique opportunities to get ahead. There is more information that we can gather when we talk about unstructured data. It is often less expensive to collect and store. And often, this information is going to bring out a lot of new predictions and insights that can keep us ahead of the game.

Collecting the Data

When we are the best planning on how we should collect the needed data, it is important that we are aware of many best practices and practical considerations for addressing any of the logistical challenges that a company can face when they reach this part of the process. Implementing a plan that helps you to collect data requires attention on a lot of matters, including:

1. Making sure that you can get the senior leadership, as well as some of the key stakeholders, to agree to this process. This group can include a lot of different people including customers, tenants, employees, union representatives, management committees, and the boards of directors for the company.
2. Establishing what will be the steering committee, or selecting someone to be accountable and consulted for all of the big decisions about this process. This may include some things like the design, communication management, logistics, finances, and coordination of all the parts.
3. Determining who is going to be responsible for collecting the data. This could be employees who are trained for the job, or some experts to handle this.

4. Identifying the technology, people, resources, and logistics that are needed to develop and implement initiatives for data collecting.
5. Anticipating and addressing concerns of key stakeholders and any of the questions that come about with the project.
6. Designing a strategy for consultation and communication that is going to explain the initiative of data collecting and encourages the highest possible participation rate.
7. Protecting the personal information and privacy of those who you collect the data from, and using carefully controlled procedures when it comes to collecting, storing, and then accessing the data. All the confidentiality and dignity of the other person needs to be respected at all times.
8. Minimizing the impact and even the inconvenience for those who are affected in the workplace or the service environment. This can go even as far as choosing the best time to collect the data.
9. Aiming to have as much flexibility as needed to allow for some changes without a ton of expenses or inconvenience for anyone.
10. Consider working with a pilot phase or a test period that will allow you to improve and even modify some of your methods of data collecting, any time that these changes are needed.

How Long Should I Collect Data For?

One question that a lot of data scientists are going to have is how long they should collect the data. They want to make sure that they are collecting the data for long enough to gather enough and learn some valuable insights from that information. But they don’t want to drag it on for too long, wasting time and money and not learning anything in the process.

The amount of time you will spend collecting the data is going to vary based on what your goals are, and what your business problem is all about. You want to collect the data for as long as needed. When you feel like you have gathered the amount of information that is needed to answer your business question or the challenge your business is facing, then it is time to stop

working with this data collecting altogether.

The first step that we need to follow when it is time to work on the process of Data Science is working with collecting and then manipulating our data. This process does take some time, as most businesses want to make sure that they can gather a lot of information in a short amount of time. But this is the data we will use to create our models and organize things as much as possible. With high-quality data and lots of it, we can create the models that we need to really find the predictions and the insights that our business needs.

Chapter 6: Data Cleaning and Preparation

The next topic that we need to take a look at in our process of Data Science is known as data cleaning and preparation. During the course of doing our own data analysis and modeling, a lot of time is going to be spent on preparing the data before it even enters into the model that we want to use. The process of data preparation is going to include a lot of different tasks, including loading, cleaning, transforming, and rearranging the data. These tasks are so important and take up so much of our time, an analyst it is likely going to spend at least 80 percent of their time on this.

Sometimes the way that we see the data stored in a database or a file is not going to provide us with the right format when we work with a particular task. Many researchers find that it is easier to do ad hoc processing of the data, taking it from one form to another working with some programming language. The most common programming languages to use to make this happen include Perl, R, Python, or Java.

The good news here though is that the Pandas library that we talked about before, along with the features it gets from Python, can provide us with everything that we need. It has the right tools that are fast, flexible, and high-level that will enable us to get the data manipulated into the form that is most needed at that time. There are a few steps that we are able to work with, in order to clean the data and get it all prepared, and these include:

What Is Data Preparation?

Let’s suppose that you are going through some of the log files of a website and analyzing these, hoping to find out which IP out of all the options the spammers are coming from. Or you can use this to figure out which demographic on the website is leading to more sales. To answer these questions or more, an analysis has to be performed on the data with two important columns. These are going to include the number of hits that have been made to the website, and the IP address of the hit.

As we can imagine here, the log files that you are analyzing are not going to be structured, and they could contain a lot of textual information that is unstructured. To keep this simple, preparing the log file to extract the data in the format that you require in order to analyze it can be the process known as

data preparation.

Data preparation is a big part of the whole Data Science process. According to CrowdFlower, which is a provider of data enrichment platforms that data scientists can work with, it is seen that out of 80 data scientists, they will spend their day in the following:

1. 60 percent of their time is spent on organizing and then cleaning the data they have collected.
2. 19 percent is spent on collecting the sets of data that they want to use.
3. 9 percent is used to mine the data that they have collected and prepared in order to draw the necessary patterns.
4. 3 percent of their time will be spent doing any of the necessary training for the sets of data.
5. 4 percent of the time is going to be spent trying to refine the algorithms that were created and working on getting them better at their jobs.
6. 5 percent of the time is spent on some of the other tasks that are needed for this job.

As we can see from the statistics of the survey above, it helps us to see that most of the time for that data scientist is spent in preparing the data, which means they have to spend a good deal of time organizing, cleaning, and collecting, before they are even able to start on the process of analyzing the data. There are a few valuable tasks of Data Science like data visualization and data exploration, but the least enjoyable process of Data Science is going to be the data preparation.

The amount of time that you actually will spend on preparing the data for a specific problem with the analysis is going to depend on the health of the data directly. If there are a lot of errors, missing parts, and duplicate values, then this is a process that will take a lot longer. But if the data is well-organized and doesn’t need a lot of fixing, then the data preparation process is not going to take that long at all.

Why Do I Need Data Preparation?

One question that a lot of people have when it is time to work on the process of data preparation is why they need to do it in the first place. It may seem to someone who is just getting started in this field that collecting the data and getting it all as organized as possible would be the best steps to take, and then they can go on to making their own model. But there are a few different reasons why data preparation will be so important to this process and they will include the following:

1. The set of data that you are working with could contain a few discrepancies in the codes or the names that you are using.
2. The set of data that you are working with could contain a lot of outliers or some errors that mess with the results.
3. The set of data that you are working with will lack your attributes of interest to help with the analysis.
4. The set of data that you want to explore is not going to be qualitative, but it is going to be quantitative. These are not the same things, and often having more quality is going to be the most important.

Each of these things has the potential to really mess up the model that you are working on and could get you results or predictions that are not as accurate as you would like. Taking the time to prepare your data and get it clean and ready to go can solve this issue, and will ensure that your data is going to be more than ready to use in no time.

What Are the Steps for Data Preparation?

At this point, we need to take some time to look at some of the steps that are needed to handle the data preparation for data mining. The first step is to clean the data. This is one of the first and most important steps to handling the data and getting it prepared. We need to go through and correct any of the data that is inconsistent by filling out some of the values that are missing and then smoothing out the outliers and any data that is making a lot of noise and influencing the analysis in a negative manner.

There is the possibility that we end up with many rows in our set of data that do not have a value for the attributes of interest, or they could be inconsistent data that is there as well. In some cases, there are records that have been

duplicated or some other random error that shows up. We need to tackle all of these issues with the data quality as quickly as possible in order to get a model at the end that provides us with an honest and reliable prediction.

There are a few methods that we can use to handle some of the missing values. The method that is chosen is going to be dependent on the requirement either by ignoring the tuple or filling in some of the missing values with the mean value of the attribute. This can be done with the help of the global constant or with some of the other Python Machine Learning techniques including the Bayesian formulae or a decision tree.

We can also take some time to tackle the noisy data when needed. It is possible to handle this in a manual manner. Or there are several techniques of clustering or regression that can help us to handle this as well. You have to choose the one that is needed based on the data that you have.

The second step that we need to focus on here is going to be known as data integration. This step is going to involve a few things like integrating the schema, resolving some of the conflicts of the data if any shows up, and even handling any of the redundancies that show up in the data that you are using.

Next on the list is going to be the idea of data transformation. This step is going to be important because it will take the time to handle some of the noise that is found in your data. This step is going to help us to take out that noise from the data so it will not cause the analysis you have to go wrong. We can also see the steps of normalization, aggregation, and generalization showing up in this step as well.

We can then move on to the fourth step, which is going to be all about reducing the data. The data warehouse that you are using might be able to contain petabytes of data, and running an analysis on this complete set of data could take up a lot of time and may not be necessary for the goals that you want to get in the end with your model.

In this step, it is the responsibility of the Data Science to obtain a reduced representation of their set of data. We want this set to be smaller in size than some of the others, but inclusive enough that it will provide us with some of the same analysis outcomes that we want. This can be hard when we have a very large set of data, but there are a few reduction strategies for the data that we can apply. Some of these are going to include the numerosity reduction,

aggregation, data cube, and dimensionality reduction, and more, based on the requirements that you have.

And finally, the fifth step of this is going to be known as data discretization. The set of data that you are working with will contain three types of attributes. These three attributes are going to include continuous, nominal, and ordinal. Some of the algorithms that you will choose to work with only handle the attributes that are categorical.

This step of data discretization can help someone in Data Science divide continuous attributes into intervals, and can also help reduce the size of the data. This helps us to prepare it for analysis. Take your time with this one to make sure that it all matches up and does some of the things that you are expecting.

Many of the methods and the techniques that you are able to use with this part of the process are going to be strong and can get a lot of the work with you. But even with all of these tools, it is still considered an area of research, one that many scientists are going to explore more and hopefully come up with some new strategies and techniques that you can use to get it done.

Handling the Missing Data

It is common for data to become missing in many applications of data analysis. One of the goals of working with the Pandas Library here is that we want to make working with some of this missing data as easy and as painless as possible. For example, all of the descriptive statistics that happen on the objects of Pandas exclude the missing data by default.

The way that this data is going to be represented in Pandas is going to have some problems, but it can be really useful for many of the users who decide to go with this kind of library. For some of the numeric data that we may have to work with, the Pandas library is going to work with a floating-point value that is known as NaN, or not a number, to represent the data that is missing inside of our set of data.

In the Pandas Library, we have adopted a convention that is used in the programming language of R in order to refer to the missing data. This missing data is going to show up as NA, which means not available right now. In the applications of statistics, NA data can either be data that doesn’t exist at all, or that exists, but we are not going to be able to observe through

problems with collecting the data. When cleaning up the data to be analyzed, it is often important to do some of the analysis on the missing data itself to help identify the collection of the data and any problems or potential biases in the data that has been caused by the missing data.

There are also times when the data is going to have duplicates. When you get information online or from other sets of data, it is possible that some of the results will be duplicated. If this happens often, then there is going to be a mess with the insights and predictions that you get. The data is going to lean towards the duplicates, and it will not work the way that you would like. There are ways that you can work with the Pandas library in order to really improve this and make sure that the duplicates are eliminated or are at least limited at least a little bit.

There is so much that we are able to do when it comes to working with data preparation in order to complete the process of data mining and getting the results that we want in no time with our analysis. Make sure to take some time on this part, as it can really make or break the system that we are trying to create. If you do spend enough time on it, and ensure that the data is as organized and clean as possible, you are going to be happy with the results and ready to take on the rest of the process.

Chapter 7: What is Data Wrangling?

The next topic that we need to spend some time on is known as data wrangling. This is basically the process where we are able to clean, and then unify, the mess and complex sets of data that we have, in order to make them easier to access and analyze when we would like. This may seem like part of the boring stuff when it comes to our Data Science proves, but it is going to be so important to the final results, so we need to spend some time on seeing how this works.

With all of the vast amounts of data that are present in the world right now, and with all of the sources of that data growing at a rapid rate and always expanding, it is getting more and more essential for these large amounts of available data to get organized and ready to go before you try to accomplish any analysis. If you just leave the data in the messy form from before, then it is not going to provide you with an accurate analysis in the end, and you will be disappointed by the results.

Now, the process of data wrangling is typically going to include a few steps. We may find that we need to manually convert or map out data from one raw form into another format. The reason that this is done in the first place is that it allows us to have a more convenient consumption for the company who wants to use that data.

What Is Data Wrangling?

When you work with your own project in Data Science, there are going to be times when you gather a lot of data and it is incomplete or messy. This is pretty normal considering all of the types of data you have to collect from a variety of sources overall. The raw data that we are going to gather from all of those different sources is often going to be hard to use in the beginning. And this is why we need to spend some time cleaning it. Without the data being cleaned properly, it will not work with the analytical algorithm that we want to create.

Our algorithm is going to be an important part of this process as well. It is able to take all of the data you collect over time and will turn it into some good insights and predictions that can then help to propel your business into

the future with success. But if you are feeding the analytical data, a lot of information that is unorganized or doesn't make sense for your goals, then you are going to end up with a mess. To ensure that the algorithm works the way that you want, you need to make sure that you clean it first, and this is the process that we can call data wrangling.

If you, as the programmer would like to create your efficient ETL pipeline, which is going to include extract, transform and load, or if you would like to create some great looking data visualizations of your work when you are done, then just get prepared now for the data wrangling.

Like most data scientists, data analysts, and statisticians will admit, most of the time that they spend implementing an analysis is going to be devoted to cleaning or wrangling up the data on its own, rather than in actually coding or running the model or algorithm that they want to use with the data. According to the O'Reilly 2016 Data Science Salary Survey, almost 70 percent of data scientists will spend a big portion of their time dealing with a basic analysis of exploratory data, and then 53 percent will spend their time on the process of cleaning their data before using in an algorithm.

Data wrangling, as we can see here, is going to be an essential part of the Data Science process. And if you are able to gain some skills in data wrangling, and become more proficient with it, you will soon find that you are one of those people who can be trusted and relied on when it comes to some of the cutting-edge Data Science work.

Data Wrangling with Pandas

Another topic that we can discuss in this chapter is the idea of data wrangling with Pandas. Pandas is seen as one of the most popular libraries in Python for Data Science, and specifically to help with data wrangling. Pandas is able to help us to learn a variety of techniques that work well with data wrangling, and when these come together to help us deal with some of the data formats that are the most common out there, along with some of their transformations.

We have already spent a good deal of time talking about what the Pandas

library is all about. And when it comes to Data Science, Pandas can definitely step in and help get a ton of the work done. With that said, it is especially good at helping us to get a lot of the data wrangling process that we want doing as well. There may be a few other libraries out there that can do the job, but none are going to be as efficient or as great to work with as the Pandas library.

Pandas will have all of the functions and the tools that you need to really make your project stand out, and to ensure that we are going to see some great results in the process of data wrangling as well. So, when you are ready to work with data wrangling, make sure to download the Pandas library, and any of the other extensions that it needs.

Our Goals with Data Wrangling

When it comes to data wrangling, most data scientists are going to have a few goals that they would like to meet in order to get the best results. Some of the main goals that can come up with data wrangling, and should be high on the list of priorities, include:

1. Reveal a deep intelligence inside of the data that you are working with. This is often going to be accomplished by gathering data from multiple sources.
2. Provides us with accurate and actionable data and then puts it in the hands of n analyst for the business, in a timely manner so they can see what is there.
3. Reduce the time that is spent collecting, and even organizing some of the really unruly data, before it can be analyzed and utilized by that business.
4. Enables the data scientists, and any other analyst to focus on the analysis of the data, rather than just the process of wrangling.
5. Drives better skills for making decisions by senior leaders in that company.

The Key Steps with Data Wrangling

Just like with some of the other processes that we have discussed in this

guidebook, there are a few key steps that need to come into play when it comes to data wrangling. There are three main steps that we can focus on for now, but depending on the goals you have and the data that you are trying to handle, there could be a few more that get added in as well. The three key steps that we are going to focus on here, though, will include data acquisition, joining data, and data cleansing.

First on the list is Data Acquisition. How are you meant to organize and get the data ready for your model if you don't even have the data in the first place? In this part of the process, our goal is to first identify and then obtain access to the data that is in your preferred sources so that you can use it as you need in the model.

The second step is going to be where we join together the data. You have already been able to gather in the data that you want to use from a variety of sources and even did a bit of editing in the process. Now it is time for us to combine together the edited data for further use and more analysis in the process.

And then we can end up with the process that is known as data cleansing. Remember that we talked about this a bit in the last chapter, but it is still important as we work with the process of data wrangling. In the data cleansing process, we need to redesign the data into a format that is functional and usable, and then remove or correct any of the data that we consider as something bad.

What to Expect with Data Wrangling?

The process of data wrangling can be pretty complex, and we need to take some time to get through all of it and make sure that we have things in the right order. When people first get into the process of data wrangling, they are often surprised that there are a number of steps, but each of these is going to be important to ensure that we can see the results that we want.

To keep things simple for now, we are going to recognize that the data wrangling process is going to contain six iterative steps. These are going to include the following:

- The process of Discovering. Before you are able to dive into the data and the analysis that you want to do too deeply, we first need

to gain a better understanding of what might be found in the data. This information is going to give you more guidance on how you would like to analyze the data. How you wrangle your customer data, as an example, maybe informed by where they are located, what the customer decided to buy, and what promotions they were sent and then used.

- The second iterative step that comes with the data wrangling process is going to be Structuring. This means that we need to organize the data. This is a necessary process because the raw data that we have collected may be useful, but it does come to us in a variety of shapes and sizes. A single column may actually turn into a few rows to make the analysis a bit easier to work within the end. One column can sometimes become two. No matter how we change up some of the work, remember that the movement of our data is necessary in order to allow our analysis and computation to become so much easier than before.
- Then we can go on to the process of Cleaning. We are not able to take that data and then just throw it at the model or the algorithm that we want to work with. We do not want to allow all of those outliers and errors into the data because they are likely to skew some of our data and ruin the results that we are going to get. This is why we want to clean off the data.

There are a number of things that are going to spend our time cleaning when it comes to the data in this step. We can get rid of some of the noise and the outliers we can take some of the null values and change this around to make them worth something. Sometimes it is as simple as adding in the standard format, changing the missing values, or handling some of the duplicates that show up in the data. The point of doing this, though, is to increase the quality of the data that you have, no matter what source you were able to find it from.

- Next on the list is the process of Enriching the Data. Here we are going to take stock of the data that we are working with, and then we can strategize about how some other additional data might be able to augment it out. This is going to be a stage of questions to make sure that it works, so get ready to put on your thinking cap.

Some of the questions that you may want to ask during this step could include things like: What new types of data can I derive from what I already have? What other information would better inform my decision making about this current data? This is the part where we will fill in some of the holes that may have found their way into the data, and then find the supplementation that is needed to make that data pop out.

From here, we can move on to the step of validation. The validation rules that we are going to work with this step in the Data Science process are going to be repetitive programming sequences. The point of working with these is that we want to check out and verify the consistency, quality, and security of our data to make sure that it is going to do the work that we want.

There are a lot of examples that come with the validation stage. But this can include something like ensuring the uniform distribution of attributes that should be distributed in a normal way, such as birth dates. It can also be used as a way to confirm the accuracy of fields through a check across the data.

- And the last stage is going to be Publishing. Analysts are going to be able to prepare the wrangled data to use downstream, whether by a software or a particular user. This one also needs us to go through and document any of the special steps that were taken or the logic that we used to wrangle this data. Those who have spent some time wrangling data understand that implementation of the insights is going to rely upon the ease with which we are able to get others the information, and how easy it is for these others to access and utilize the data at hand.

Data wrangling is an important part of our process and ensures that we are able to get the best results with any process that we undertake. We need to remember that this is going to help us to get ahead with many of the aspects of our Data Science project, and without the proper steps being taken, we are going to be disappointed in what we see as the result in the end. Make sure to understand what data wrangling is all about, and why it is so important so that it can be used to help with your Data Science project.

Chapter 8: Taking Our Results and Plotting Them to Visualize What We Learned

The next topic that we need to take some time to learn about is known as Data Visualization. The point of this one is to take all of that data we have been collecting, and then learn how to plot it or turn it into some other form of graphical representation. This can then be presented, along with the other information you have on the findings to key decision-makers. Graphs and charts are much easier to read through and understand compared to large blocks of text, so this is another addition to your work that can ensure the key decision-makers actually understand what you did and what predictions and insights were found inside.

Data visualization is simply going to be the presentation of data in a graphical or pictorial format. It is helpful because it will enable key decision-makers to see analytics presented in a visual form. This makes it easier to grasp some difficult concepts quickly or even identify some brand-new patterns. With an interactive form of visualization, it is easier to take this concept even further by using a variety of technologies to drill down into charts and graphs for more detail, interactively changing what data you are going to see, and how this data can be processed.

The History of Data Visualization

The ideas that come with data visualization are going to be important to ensure that we can get the most out of what we see with our project. The concept of using any kind of picture or visual to help understand large amounts of data is something that has been around for centuries. Look back to some of the early maps and graphs and even to the invention of the pie chart in the early 1800s. Several decades later, one of the most cited examples of one of these graphs occurred when Charles Minard was able to map the invasion of Russia by Napoleon.

This graph was truly amazing and helped to show us a lot of complex information in one picture. For example, this map was able to depict the size of the army throughout the invasion, as well as the path that Napoleon took in retreat from Moscow. And this information was also tied back to the temperature and time scales as well to make it even easier to understand the

event and what was happening at the different stages.

It is technology however that is the thing that really added some fire to the process of data visualization. Computers have made it easier than ever to process a large amount of data at speeds that are faster than ever. Because of this, data visualization today has become a rapidly evolving blend of science, and it is going to be even bigger changes in the landscape of many businesses over the next few years.

Why Is This Data Visualization So Important?

The next question that we need to take a look at here is why this data visualization is so important. Due to the way that the human brain is able to process information, using graphs and charts and other options to visualize a large amount of data that is more complex, proves to be easier than pouring over reports and spreadsheets. We can learn a lot more from a chart than we are from ten pages of data and at a much faster rate.

Data visualization is helpful because it is quick and it is one of the easiest ways to convey concepts in an universal manner. And you can also move some of the information around and do some experimenting to make slight adjustments and see how that influences the data that you have. In addition, data visualization is also able to help with:

1. Identifying areas that need more attention or some improvement within them.
2. Clarify out which factors are the most likely to influence the behavior of the customer.
3. It can help us to understand which products we can place in different locations.
4. It is a great way to predict sales volumes.

How is Data Visualization Being Used?

The next thing that we need to take a look at here is how data visualization is actually being used. No matter the size of the industry, all types of businesses have found the value in using data visualization, and some of the tools that come with this, to help them make some more sense out of all that data they

collected. There are a number of benefits to using this kind of tool in your Data Science project, and the way that you can use it will depend on what your overall goals are. Some of the ways that we can use data visualization to help make sense of our data will include:

It can help us to comprehend data and information faster. By using representations of our information in a graphic form, businesses have a better way to see large amounts of data in a clear and cohesive manner. They are then able to draw some conclusions based on that information. And since using this is going to be faster when it comes to analyzing the information, especially when compared to reading the information of a spreadsheet or another format, businesses are better able to address any problems or answer the needed questions in a timelier manner.

Another reason that a company would want to go with data visualization as part of their Data Science project is that it can help them to pinpoint some emerging trends. Using this tool to help discover new trends, both in the market and in the business itself, can give any business the edge that they need over their competition. And when we can beat out the competition, this helps to positively influence our bottom line.

A data visualization makes it so much easier to make this happen. With the data visualization, companies can find that it is much easier to spot any of the noise in the data, or the outliers that could affect the quality of the product or even the customer churn. Then they can address any of the issues that could affect them, long before these issues become bigger problems.

We can also see that this data visualization is going to help a business to identify relationships and patterns in their data. Even extensive amounts of complicated data can start making sense to us when we take it out of the words and turn it into a graph or a chart. Businesses that uses this method will start to recognize what parameters are there and how they can be highly correlated to one another at the same time.

Some of the correlations that show up in your data are going to be more obvious, and you will know why they are set up that way in the first place. But then there are some correlations that won’t be as obvious. Being able to identify these relationships is a key to ensuring that organizations focus on the areas that are able to have some control over their most important goals, rather than wasting time and money in the process.

And finally, another way that we are able to see Data Science at work is when it communicates a story to others once a business has been able to uncover some of the new insights that are there through the process of data analytics, they then need to go through and communicate these insights to others, usually the key decision-makers or the stakeholders of that company.

With the help of graphs, charts, and other representations that can be seen as visually impactful, we are able to engage the other person, make the complex data easier to understand and ensure that the messages get across as quickly as possible. Those who have to use this information to make key business decisions are going to enjoy that they can just look through a graph and see the information that they want in one form.

Laying the Groundwork for Data Visualization

The next thing that we need to take a look at is some of the groundwork that needs to come into play before we are able to work with our Data Visualizations. Before we can implement any kind of new technology, there are a few steps that need to be taken. Not only do we need to have a good grasp on the data that we are using, but we also need to understand our own goals, the needs that our business has, and the audience overall. Having all of these in place and understanding how we want the data to help us can be important to see the results that we would like.

There are a number of steps that we need to do before we can work with data visualization and get it to work the way that we want. Preparing the business for the technology that helps with data visualization requires that we first do the following:

1. We need to understand the data that we want to visualize. It is impossible to pick out the type of visualization that we want to work with if we know nothing about the data that will make up that visualization. We need to know the size of the data, and the cardinality, which includes how unique the data values are in a column.
2. We need to determine what we would like to visualize, and what

kind of information is the most important to communicate in this process.

3. We need to know our audience well and understand how this audience is going to be able to visualize that information.
4. We need to use a visual to convey the information in the best, and often the simplest form, to the audience.

Once we have been able to answer the first four questions that we have above, including about the type of data we have and the audience who is most likely to consume this information, it is time to do some preparation. This needs to be done on the amount of data that you plan to work on for this visualization. As many businesses are finding out, big data can bring on some new challenges when we work with visualization.

When we factor in all of the large volumes, the different varieties, and some of the varying velocities of the big data, it is going to change the way that the visualization is able to work overall. Add to this that data can be generated at a much faster rate than ever before, and it is sometimes hard to manage or analyze it without a good method in place.

There are a lot of factors that we need to consider when it is time to create our visualization for our data. One of these includes the cardinality that shows up in the columns that we want to visualize. High cardinality means that there is going to be a large percentage of unique values. This could include something like bank account numbers since we want to have each of the items unique from the others.

Then it is possible that the cardinality is going to be lower. Low cardinality just means that the data in the column will contain a large percentage of values that are able to repeat again and again. This might be in the column that includes the gender of the customer.

Which Visual is the Right One for My Project?

One of the biggest challenges that we are going to face in this part of Data Science is figuring out which of the visuals are the best to represent all of that data and information we have been working on. There are a lot of choices when it comes to a good visual that will get your point across to the users, but we want to make sure that we are going with the one that will represent our

data in the proper manner without any worries about errors or misrepresenting our work.

When you first start to explore a new set of data, auto-charts are sometimes a useful tool to work with simply because they can give you a quick view of large amounts of data. This capability of data exploration is helpful, even to statisticians who are more experienced because it helps them to speed up the lifecycle process of their analysis. This happens because the auto-chart is able to eliminate the need for repeated sampling to determine what kind of data is going to be the best for each model that you create.

There are also a lot of other types of visualizations that we are able to rely on based on what kind of data we explore and how we would like to view it. Some of the best types of data visualizations that we can do with the Python language will include:

1. **The scatterplot:** This is the visualization that is used to help us find any relationship that is present in bivariate data. It is going to be used in many cases to help us take two continuous variables and find the correlation between them.
2. **Histogram:** The histogram is going to help us to see the distribution that is there between a variable that is continuous. It is good for discovering the frequency distribution for a single variable when we are completing an analysis that is univariate.
3. **Bar chart:** This is the type of visualization that can be used to represent categorical data with either horizontal or vertical bars. It is a general type of plate that makes it easier to aggregate some of our categorical data based on a function. The default of this is going to be the mean, but you can choose to have the function as something else.
4. **Pie chart:** This graph is going to be an example of a plot that is used to represent the portion of each category when we have categorical data. The whole pie is going to get divided up into slices, which are going to be the same as the number of categories.

Working with data visualization is going to be very important to the results that we are looking for with our Data Science project. Without the right visualization in process, it is really hard to see the data act the way that we

want, and we may spend hours poring over spreadsheets and more, trying to guess what information is hidden inside. With the visualization, no matter what kind you use as long as it works for your data, this is no longer a big issue. We are able to focus on what is in the chart or graph, and see exactly what trends, insights, and predictions, are found in our data right away.

Chapter 9: Data Aggregation and Group Operations

Taking the time to categorize our set of data, and giving a function to each of the different groups that we have, whether it is transformation or aggregation, is often going to be a critical part of the workflow for data analysis. After we take the time to load, merge, and prepare a set of data, it is then time to compute some more information, such as the group statistics or the pivot tables. This is done to help with reporting or with visualizations of that data.

There are a few options that we are able to work with here to get this process done. But Pandas is one of the best because it provides us with an interface that is flexible. We can use this interface to slice, dice, and then summarize some of the sets of data we have in a more natural manner.

One reason that we see a lot of popularity for SQL and relational databases of all kinds, is because we can use them to ease up the process which joins, filters, transforms, and aggregates the data that we have. However, some of the query languages, including SQL, that we want to use are going to be more constrained in the kinds of group operations that we are able to perform right with them.

As we are going to see with some of the expressiveness that happens with the Pandas library, and with Python, in general, we can perform a lot of operations that are more complex. This is done by simply utilizing any function that is able to accept an array from NumPy or an object from Pandas.

Each of the grouping keys that you want to work with can end up taking a variety of forms. And we can see that the keys don’t have to all come in as the same type. Some of the forms that these grouping keys can come in for us to work on includes:

1. An array or a list that is the same length as the axis that we want to group.
2. A value that is going to indicate the name of the column in a DataFrame.
3. A Dict or a Series that is going to give the correspondence

between the values of the axis that is being grouped here, and the group names you have.

4. A function that can then be invoked on the axis index, or on some of the individual labels in the index.

Note that the last three methods of this are going to be a type of shortcut that helps us to produce an array of values to be used when splitting up the object. This can seem a bit abstract right now, but don't let this bother you. It will all make more sense as we go through the steps and learn more about how all of this is meant to work. With this in mind, it is time to talk more about data aggregation and how we are able to make this work for our needs.

What Is Data Aggregation?

Data Aggregation is going to be any kind of process in which information can be gathered and then expressed in the form of a summary, usually for the purpose of analysis. One of the common purposes that come with aggregation is to help us get some more information about a particular topic or a group, based on a lot of variables like profession, income, and age.

The information about these groups is often going to be used in order to personalize a website, allowing them to choose what content and advertising that is likely to appeal to an individual who belongs to one or more groups where the data was originally collected from. Let's take a look at how this works.

We can work with a site that is responsible for selling music CDs. They could use the ideas of data aggregation in order to advertise specific types of CD's based on the age of the user, and the data aggregate that is collected for others in that age group. The OLAP, or Online Analytic Processing, is a simple option with data aggregation in which the market is going to use mechanisms for online reporting to help the business process through all of this information.

Data Aggregation can be a lot of different things as well. For example, it could be more user-based than some of the other programs that we may have seen in the past. Personal Data Aggregation Services are popular, and they will offer any user a single point for collection of their personal information from a host of other websites that we want to work with.

In these systems, the customer is going to work with a single master PIN, or personal identification number, which allows them the access they need to various accounts. This could include things like music clubs, book clubs, airlines, financial institutions, and so on. Performing this type of Data Aggregation can take some time and will be a more complex system to put in, but we will see that it comes under the title of screen scraping.

This is just one example of how we are able to work through the process of data aggregation. It is one of the best methods to help companies really gain the knowledge and the power that they need based on the users they have at the time. It often works well with Pandas, Python, and even databases because it can collect a lot of the information that is found in those, and then recommends options to our customers or our users, based on where they fit in with the rest of the information.

Yes, there are always going to be some outliers to the information, and times when the information is not going to apply to a person, no matter where they fit in the database or how good the data aggregation algorithms are. But it will be able to increase the likelihood that you will reach the customers and the users you want, providing them with the information and the content that they need, based on their own features and how they will react compared to other similar customers.

Chapter 10: What Is the Time Series?

Another topic that we need to spend some time discussing in this guidebook that can help us learn more about the process of Data Science and what it all entails is the idea of a Time Series. This is a bit different than some of the other topics that we have had some time to explore so far, but it is still an important part of the process that we need to be able to look through.

The first question here is what a Time Series is all about. A Time Series is going to be a sequence of numerical points that happen in successive order. When we talk about investing, for example, a time series is able to track the movement of any points of data that you choose, such as the price of a security over a specified time period with the points of data recorded at intervals that are regular, such as once a day.

When it comes to Data Science, we are able to work with this in a slightly different way. But it is still the same idea. We will find the object that we want to track, such as customer satisfaction, and then follow these at regular intervals to see what is happening with the satisfaction of our customers and whether the new improvements we are working with are going to really help provide us with a higher amount of customer satisfaction overall.

There isn't going to be a minimum or even a maximum amount of time that we need to include with our Time Series. This allows the business to gather and collect their data in a manner that will provide the information they need. If they find that they can gather the information in a shorter amount of time, then the business can stop doing some of the analysis a bit earlier. If they feel that this is something that they should monitor and watch for a longer period of time to get accurate results, then this is what they can do.

This is one of the nice things that come with the Time Series. We are able to make some customizations to it, especially when it comes to how long we want the intervals for testing to be, and how long we wish to keep the time series going for. This customization is going to ensure that we can get it to work the right way that we need for our project.

Understanding the Time Series

Let's dive a bit more into what the time series is all about. A Time Series can be taken on any kind of variable that we want, as long as this variable

changes over time. When we see this in investing, for example, it is common to work with the time series in order to track the price of a security over time. With your business, you may use it to track the level of customer satisfaction that you have, or how well the market is accepting your new product.

We are able to track the time series over the short term, such as the price of something over the hour, or even the full course of a business day. And then there is the possibility that we want to track the variable over the long term, such as looking at the price of the security when you get to the close on the last day of each month over two or three years.

We can also perform what is known as a time series analysis. This can be a useful tool to use because it helps us to really track and look through how a given economic variable, security, or asset is going to change over time. We can also use this when it is time to examine how the changes associated with the point of data we chose compare to shifts in other variables over a similar period of time.

For example, let's say that we want to go back to the idea of investing for a bit and analyze the Time Series of a closing stock daily with the prices, looking at the stock for a year. You would first need to obtain a list of the prices for that stock at closing for each day in the past year, and then get these listed out in chronological order. This would be the time series for the closing price of one year for the stock. We can use this to figure out what the peaks and troughs are with that stock, so we can make some better predictions of how it will behave in the future.

You can also use this to figure out your profits and how they vary from one part of the year to another. You could do this same idea, monitoring the monthly gross profit each month for three years and write these out. With this information in place, we are able to see that the price goes up at some points of the year, and down at other points of the year. This makes it easier to figure out when to extend hours, how many employees to have on hand at different parts of the year, and more.

Another neat thing that we are able to do here when it comes to working with the time series is to use it in some of our forecasting endeavors. Time Series Forecasting is helpful because it is going to use information regarding historical values and associated patterns to make it easier to predict the future activity of the variable you are monitoring.

This could include a lot of different parts, but may relate to trending analysis, issues with seasonality (is one season of the year naturally a slower period for you?), and a cyclical fluctuation analysis. As with all of the forecasting methods, success is not going to be guaranteed, but it can help with planning and ensuring that you can best serve your customers, create high-quality products, and get the most profits possible.

Working with the Time Series is a great way to enhance the knowledge you have about a particular topic and can help you determine whether it is something that you need to pursue or not. It can even make it easier to plan out some of the work that you need to do with your business, adding so much more to your business that other competitors are not able to beat out.

Chapter 11: What Is Machine learning and How It Fits with Data Science

The next topic that we need to take a look at here is Machine learning and how it can come into play when we work with Data Science and all of the neat things that we are able to do with this topic. Machine Learning can definitely be an important part of the Data Science process, as long as we use it in the proper manner.

Remember, as we go through this process, that part of Data Science is working on data analysis. This helps us to take a lot of the data we have collected along the way, and then actually see the insights and the predictions that are inside of it. To make this happen, we need to be able to create our own models, models that are able to sort through all of the data, find the hidden patterns, and provide us with our insights.

To help us make these models, and to make sure that they actually work the way that we want, we need to have a variety of good algorithms in place, and this is where Machine Learning is going to come into play quite a bit. You will find that with the help of Machine Learning, and the variety of algorithms that are present in Machine Learning, we can create models that are able to go through any kind of data we have, whether it is big or small, and provide us with the answers that we need here.

Machine Learning is basically a process that we can use in order to make the system or the machine we are working with think in a manner that humans do. This allows the algorithm to really go through and find hidden patterns in the same manner that a human would be able to do, but it can do it much faster and more efficiently than any human could do manually.

Think about how hard this would be to do manually for any human, or even for a group of people who are trying to get through all of that data. It could take them years to get through all of that data and find the insights that they need. And with how fast data is being generated and collected, those predictions and insights would be worthless by the time we got to that point anyway.

Machine Learning can make this process so much easier. It allows us to have a way to think through the data and find the hidden patterns and insights that

are inside for our needs. With the right Machine Learning algorithm, we are able to really learn how the process works, and all of the steps that are necessary to make this happen for us. With this in mind, it is time to take a closer look at Machine Learning, and all of the parts that we need to know to make this work for our needs.

What is Machine learning?

The first thing that we need to take a look at here is the basics of Machine Learning. Machine Learning is going to be one of the applications of artificial intelligence that can provide a system with the ability to learn, all on its own, without the help of a programmer telling the system what to do. The system can even take this a bit further and can work to improve based on its own experience, and none of this is done with the system being explicitly programmed in the process. The idea of Machine Learning is going to be done with a focus on the development of programs on the computer that is able to access any data you have, and can then use that presented data to learn something new, and how you would like it to behave.

There are going to be a few different applications that we can look at when it comes to using Machine Learning. As we start to explore more about what Machine learning is able to do, you may notice that over the years, it has been able to change and develop into something that programmers are going to enjoy working with more than ever. When you want to make your machine or system do a lot of the work on its own, without you having to step in and program every step, then Machine Learning is the right option for you.

When it comes to the world of technology, we will find that Machine Learning is pretty unique and can add to a level of fun to the coding that we do. There are already a lot of companies, in a variety of industries (which we will talk about in a bit), that will use Machine learning and are already receiving a ton of benefits from it.

There are a lot of different applications when it comes to using Machine Learning, and it is amazing what all we can do with this kind of artificial intelligence. Some of the best methods that we are able to follow and focus our time on when it comes to Machine Learning include:

1. **Research on statistics:** Machine Learning is already making some headway when it comes to the world of IT. You will find

that Machine Learning is able to help you go through a ton of complex data, looking for the large and important patterns that are in the data. Some of the different applications of Machine Learning under this category will include things like spam filtering, credit cards, and search engines.

2. **An analysis of Big Data:** There are a lot of companies who have spent time collecting what is known as Big Data, and now they have to find a way to sort through and learn from that data in a short amount of time. These companies are able to use this data to learn more about how money is spent by the customers, and even to help them make important decisions about the future. If we had someone go through and manually do the work, it would take much too long. But with Machine Learning, we are able to get it all done. Options like the medical field, election campaigns, and even retail stores have started to turn to Machine Learning to gain some of these benefits.
3. **The financial world:** There are many financial companies that have been able to rely on Machine Learning. Stock trading online, for example, will rely on this kind of work, and we will find that Machine Learning can help with fraud detection, loan approvals, and more.

To help us get going with this one, and to understand how we are able to receive the value that we want out of Machine Learning, we have to make sure that we pair the best algorithms with the right processes and tools. If you are using the wrong kind of algorithm to sort through this data, you are basically going to get a lot of inaccurate information, and the results will not give you the help that you need. Working with the right algorithm the whole time will make a big difference.

The really cool thing that we will see with this one is that there are a lot of Machine learning algorithms that we can choose from at this point to work on your model. Each of these works in a different manner than the others, but this ensures that you are able to handle any kind of problem that comes along with your own project. With this in mind though, you will notice that some of the different algorithms that are available, and really great at doing your job,

and finding the insights that you want, include random forests, neural networks, clustering, support vector machines, and more.

As we are working on some of the models that we want to produce, we will also notice that there are a ton of tools and other processes that are available for us to work with. We need to make sure that we pick the right one to ensure that the algorithm, and the model that you are working with, will perform the way that you would like. The different tools that are available with Machine learning will include:

1. Comprehensive management and data quality.
2. Automated ensemble evaluation of the model to help see where the best performers will show up.
3. GUIs for helping to build up the models that you want along with the process flows being built up as well.
4. Easy deployment of this so that you can get results that are reliable and repeatable in a quick manner.
5. Interactive exploration of the data and even some visualizations that help us to view the information easier.
6. A platform that is integrated and end to end to help with the automation of some of the data to decision process that you would like to follow.
7. A tool to compare the different models of Machine learning to help us identify the best one to use in a quick and efficient manner.

The Benefits of Machine learning

We also need to take some time to look at a few of the benefits that come with Machine learning. There are a lot of reasons why we would want to choose to go with Machine Learning to help our Data Science Project. It is impossible to create some good algorithms or models that can accurately make predictions out of the data you send through it. And there are a lot of other benefits that can come with this as well. Some of the best benefits that we can see when we decide to work with Machine learning include:

1. **Marketing products are easier:** When you are able to reach your customers right where they are looking for you, online and on social media, it can increase the sales. You can use Machine Learning to figure out what your target audience is going to respond to, and you can make sure that the products you are releasing work for what the customer wants as well.
2. **Machine Learning can help with accurate medical predictions:** The medical field is always busy, and it is believed that a lot of the current job openings are going to be left unfilled. Even a regular doctor, with no specialties, will need to go through and deal with lots of patients throughout the day. Keeping up with all of this can be a hassle. But with the help of Machine learning, we can create a model that is able to look at images, and recognize when something is wrong or not. This can save doctors a lot of time and hassle and can make them more efficient at their jobs.

This is just one area where Machine learning will be able to help out with the medical field. It can assist with surgeries, helps with taking notes for a doctor, looking for things in x-rays and other imaging, and even helping with front desk operations.

3. **Can make data entry easier:** There are times when we need to make sure that all the information is entered into a database in an efficient and quick manner. If there is a ton of data to sort through, and we are short on time, this can seem like a task that is impossible. But with Machine Learning and the tools that come with it, we are able to get it all done in no time.
4. **Helps with spam detection:** Thanks to some of the learning processes that come with Machine Learning, we find that this can be used to prevent spam. Most of the major email servers right now will use some form of Machine learning in order to handle spam and keep it away from your regular inbox.
5. **Can improve the financial world:** Machine Learning is able to come in and work with a lot of the different tasks of the financial world. It helps with detecting fraud, offering new products to customers, approving loans, and so much more.

6. **Can make manufacturing more efficient:** Those in the manufacturing world are able to use Machine Learning to help them be more efficient and better at their job. It can figure out when things are going to be slowing the process down and needs to be fixed, it will look at when a piece of a machine is likely to die out, and so much more.
7. **It provides us with a better understanding of the customer:** All companies want to know as much about their customers as possible, ensuring that they can really learn how to market to these individuals, what products to offer, and which methods they can take to make the customer as happy as possible.

Supervised Machine learning

The first type of Machine learning algorithm that we will take a look at is known as supervised Machine learning. This Machine learning type is the kind where someone is going to train the system, and the way that they do this is by making sure to provide input, with the corresponding output, to the system so it knows the right answers. You also have to take the time to furnish the feedback into the system, based on whether the system or the machine was accurate in the predictions that it made.

This is a process that takes time, and you need to have a good deal of data present in order to make it happen. The trainer has to show a bunch of different and diverse examples to that system, and then also show the system the output, or the corresponding answers, so that it can learn how it is supposed to behave in the process as well.

After the completion of the training, the algorithm will need to apply what it learned from the data earlier on to make the best predictions. The concept that comes with supervised learning can be seen to be similar to learning under the supervision of a teacher to their students. The teacher is going to give a lesson to the students with some examples, and then the student is going to derive the new rules and knowledge from these examples. They can then take the knowledge and apply it to different situations, even if they don’t match up directly to the examples that the teacher gives.

In supervised Machine learning, we are going to spend our time looking at how to train, and then test the method we are working with. We need to have

lots of data to present to the model, and it needs to be labeled or structured in format. This can take some time to accomplish, and often we need to go through quite a few iterations of the training and testing before we are able to find the results that we need. When the model is properly trained, we will be able to handle the data that you present to it later, and then we can receive the predictions and more that we need.

Unsupervised Machine learning

Now we can move on to the idea of Unsupervised Machine learning and see how this one is going to work compared to supervised learning. With unsupervised Machine learning, we will find that there are going to be a big difference when it is compared to the other methods but can train the system how to behave, without all of the examples and labeled data along the way.

We will find that with unsupervised learning, the model is not going to be provided with the output for it to be taught how to behave. This is because the goal of this kind of learning is that we want the machine to learn what is there, based on the unknown input. The machine is able to learn how to do this all on its own, rather than having the programmer come in and do all of the work on it.

This process, or this approach, is something that is known as deep learning. Deep learning is a form of Machine learning that basically lets the computer learn on its own, rather than having to be trained the whole time. This kind of deep learning is going to be an iterative approach because it is able to review all of the data it is holding onto, and then figure out the conclusions that it wants to make from there.

What this does is make an unsupervised learning approach more suitable for use in a variety of processing tasks, which can be more complex than what you would do with supervised learning algorithms. This means that the learning algorithms that are unsupervised are going to learn just from examples, without getting any responses to it. The algorithm will strive to find the patterns that come from those examples all on its own, rather than being told the answers.

Many of the recommender types of systems that you encounter, such as when you are purchasing something online, are going to work with the help of an unsupervised learning algorithm. In this kind of case, the algorithm is going

to derive what to suggest to you to purchase based on what you went through and purchased before. The algorithm then has to estimate the customers you resemble the most based on your purchases and then will provide you with some good recommendations from there.

Reinforcement Machine learning

The third type of Machine learning method that we need to take a look at here is going to be known as reinforcement Machine learning. This algorithm type is newer than the other two, and it is going to be the one that we work with any time that the presented algorithm has examples, but these examples are not going to have any labels on them at all.

This is often the algorithm that is going to happen in a way that looks similar to unsupervised Machine learning. And sometimes, if we are not familiar with this type of learning, it may seem like the two categories are going to be the same thing. But there are some differences, mainly that it works with trial and error compared to the other option. The reinforcement Machine learning method is going to rely more on positive feedback, and sometimes negative feedback. This is going to be based on the solution that is proposed by our algorithm, and whether it all ends up matching up to what the programmer is looking for.

With reinforcement Machine learning, we will find that it is used with any kind of application that has the algorithm to make some decisions. These decisions will face some kind of consequence, either positive or negative, based on what the decision was and how it relates to your conditions.

Errors are fine in this because they are going to become useful in the learning process when they are associated with a penalty, such as loss of time, cost, and pain. In the process of reinforced learning, some actions are going to be more likely to succeed while others are less likely to succeed.

Machine Learning processes are going to be similar to what we see with predictive modeling and data mining. In both cases, patterns are then going to be adjusted inside the program accordingly. A good example of Machine learning is the recommender system. If you purchase an item online, you will then see an ad that is going to be related to that item.

There are some people who see reinforcement learning as the same thing as unsupervised learning because they are so similar, but it is important to

understand that they are different. First, the input that is given to these algorithms will need to have some mechanisms for feedback. You can set these up to be either negative or positive based on the algorithm that you decide to write out.

So, if you decide to work with this method of Machine learning, you are going to work with a Machine learning Technique that is similar to trial and error. Think about when you are doing something with a younger child. When they go through and perform an action that you don't want them to repeat, you will stop by letting them know the action is not something they are allowed to do. You explain that they need to stop, or they will end up in a time out or another action that you want here.

There is so much that we will be able to do when it comes to working with Machine learning. There are a lot of benefits that we can look for when it is time to explore Machine learning, and we can look at the three different types of Machine learning. Each of these three types, including supervised, unsupervised, and reinforcement Machine learning, can help us to get the results that we want as we work with Data Science and to really see the insights and the predictions that are needed out of our data.

Machine Learning is a buzzword that is creating a stir in many businesses, no matter what industry they fall in. Understanding how this process works, and how you are able to use it with your own Data Science project, can help us to get the insights and predictions that we are looking for.

Chapter 12: Other Libraries That Can Help with Python

We have already spent a good deal of time in this guidebook exploring the different types of coding that we can do, and even a few of the options with Python Libraries that help out with Data Science. This can make it easier than ever to take some of the ease of use and the power of Python, and put it all to good use for those Data Science models. The libraries that we have already spent some time discussing in this chapter are going to be useful because they can handle a lot of the various tasks and more that you want to do with Python.

With that said, we also need to take some time to look at a few of the other libraries that we can add to the mix, and that will help us to get things done. There are a lot of Machine learning, Deep Learning, and Data Science libraries out there that can help us with each step of the process of our data analysis. Some of the additional libraries that a programmer can consider using along with the Python Language in Data Science includes:

IPython

The first library that we are going to take a look at is known as IPython. This is going to be an interactive shell that works with the Python programming language and is able to offer an enhanced introspection, lab completion, and additional syntax for the shell that we are not able to get with some of the other options.

This brings up the idea of why you would want to work with IPython in the first place. There is already a basic shell that comes with the traditional Python library and will help us to get a lot of the coding and work did that we need. But in some cases, the shell that you are able to get with Python is going to seem a little bit basic, and many programmers find that they want something a little bit more. And IPython is able to provide that to us.

If you find that the default shell that comes with the Python Language is a bit too basic to help you out, then the alternative of using IPython is a great one to work with. You will get all of the same features and functions that are available with the Basic interpreter, but it also provides us with a lot of extras

that make some of our coding a bit easier. You may find that it comes with things like a help function, editing that is more advanced, line numbers, and more.

When it is time to do some of the models and algorithms that you want to accomplish in Data Science, you may find that working with the basic shell of Python is just not going to be enough to get it all done. With the help of IPython and all of the neat things that it is able to provide, we will be able to take some of our codings to the next level and really see some results in no time.

Jupyter

Jupyter is the next library that is on our list that we can explore a little bit here. The Jupyter Notebook is a unique addition to Python and Data Science and it is able to combine live code, visualizations, graphics, and text into some notebooks that are shareable and run well in the web browser of your choice.

The idea with this one is that at some point, we all need to be able to show off our work. Most of the programming work that is done out there will either be shared as the code that is raw still or as a compiled executable that others can try out. The source of the code is going to provide us with the complete information that we need, but it is going to be done in a manner that “tells” the other person things, rather than “shows” anything. Then we have the executable code that will show us exactly what that kind of software does. But even when it is shipped out with some of the source code, it is still difficult to figure out exactly how all of this is supposed to work.

Imagine how frustrating this could be? And imagine how it would feel if you were able to view the code and then still execute it in the same user interface. This could allow you to make the changes that you need to the code, and then still see the results of any changes you made, instantly in real-time? If this is something that interests you at all, then the Jupyter Notebook is the right option for you.

This Notebook was created with the idea that it needed to be easier for a coder to show their programming work, and to let others join in and make suggestions or even changes. This Notebook is going to allow us to take one interactive document, which is known as a notebook, and combine together

the code, multimedia, comments, and visualizations in one place. In addition to this, we can take that interactive document and share it with others, re-use it as needed, and even re-work some of the parts that are needed.

And hosting this kind of Notebook is pretty easy as well. Since the Jupyter Notebook is going to run via a web browser, the notebook on its own can be hosted on either a remote server or even on the local machine that you are using. You can choose to put this where it is the most convenient for your needs.

There are a lot of benefits that come with the Jupyter Notebook, which is why it is one of the most popular interfaces out there for programmers to work with. This was originally developed for various applications of Data Science that were written in Julia, R, and Python. In addition, it can make its way into a lot of other projects as well.

To start, this Notebook is going to be great when you want to focus on data visualization. Most of those who gain exposure to Jupyter will do so while working on their data visualization. Jupyter Notebook is a great option that lets the author create their own visualizations, while also sharing these and allowing for some interactive changes to the shared code and the set of data.

Another benefit is that this allows for some code sharing. There are several cloud services, including Pastebin and GitHub, that will provide us with ways to share the code. But while these do offer that service, we have to remember that these are not going to be interactive. When we add in this Notebook, we are able to view the code, execute it, and then display the results directly to the web browser that we are using.

It is also possible to work with some live interaction in your code. The Notebook that we are taking a look at here is not going to be static. It is designed to be edited and re-run incrementally in real-time. With the feedback that is provided directly in the browser, we know that this can all get done in no time. Notebooks are able to embed some of the controls from the user, which can then be used as the sources of input in the code.

And finally, another benefit that a programmer may like about the Jupyter Notebook is that it can help document code samples. If you have a piece of code and your goal is to explain how it works going one line at a time, with some feedback that is live along the way, you may want to use this Notebook.

The best thing with this one is that while you do this process, the code will maintain its functionality. You can just add along in an interactive manner with the explanation, basically being able to show and tell all at the same time.

It is also important to take a look at some of the components that come with the Jupyter Notebook. This notebook is going to require a few different ingredients in order to make sure that it can behave in the manner that we want. Each of these ingredients also has to be organized out, so they are in their own discrete blocks. Some of the components that have to be in place for the Jupyter Notebook to work include:

1. **The HTML and Text**: Plain text, or text annotated in the Markdown syntax to help us generate some HTML, can be inserted inside of your document, no matter which point you would like it to be at. CSS styling is also something that we can include inline, or we can add it to the template that is used to help generate the notebook.
2. **The code and the output**: The code that we will see with this kind of notebook is often going to be Python, but it does work with a few other languages if you would choose. The results of the code that you choose to execute are going to show up right after the blocks of code. And these blocks of code can be executed, as well as re-executed, in any order that you want, and as often as you would like.
3. **Visuals**: Things like charts and graphs can be generated from code by way of modules, including Bokeh, Plotly, or Matplotlib. Like the output, these are going to appear inline, right next to whatever code generates them. However, the code can be changed up in order to write these out as external files any time it is needed.
4. **Multimedia**: Because this Notebook is going to be built on web technology, it is able to display all types of multimedia that you want, and will support them in the web page. You can include them in a notebook as the elements of HTML, or you can decide to generate them programmatically using IPython.A display module.

5. **Data:** The data that you have for this notebook is going to be provided in a separate file alongside the .ipynb file. This is going to constitute a Jupyter Notebook, or it can be imported by doing the process programmatically. For example, it can do this by including some code into the notebook and telling it to download the data from an internet repository that is free or access it by a connection to another database.

Scikit-Learn

Another library that we can focus on when it is time to do some work with Python and Data Science is known as Scikit-Learn. This is a key library with the programming language of Python that is used to help out with a variety of Machine learning projects. For the most part, this library is going to be focused on a lot of tools that work on Machine learning projects, including statistical, mathematical, and general-purpose algorithms. These algorithms can all work together to help us form the basis of many Machine learning Technologies that we want to do.

Because this is a free tool that has a tremendous amount of power behind it, it makes sense that this is one of the libraries that we need to focus on in Machine learning and Data Science for that matter. Scikit-Learn can be helpful with the development of many algorithms that work with Machine learning, and many of the other technologies that you may want to create.

Some of the key parts and elements that come with Scikit-Learn make it particularly helpful when we work in Machine learning. Some of the algorithms that it is able to work with, including clustering, regression, and classification, can help with almost any kind of model we want to create for data analysis. For example, this library is able to support work with random forests, where the individual tree nodes hold onto information that can combine with the architecture of other trees, achieving a look like a forest.

But the random forest is not the only thing that we are able to do with this kind of library. In addition to this algorithm, Scikit-Learn can be a great option to help out with things like gradient boosting, vector machines, and some of the other elements of Machine learning that can be seen as key to achieving the results that you want. As a resource that is able to gather up a lot of algorithms and make them easier to use, this is definitely one of the

libraries that we want to spend our time on.

TensorFlow

The final library that we are going to take some time to explore in this guidebook is known as the TensorFlow library. This library is currently one of the best-known deep learning libraries in the world. It was developed by Google to help with one of their projects and is designed to be used in Machine learning for things like recommendation systems, image captioning, translation, and search engine results.

The architecture that comes with TensorFlow is going to come in with three parts that you can use, and these will include:

1. The ability to take the data you have and preprocess it.
2. It helps you to build up the model that you want to use, adding in the right algorithms to make this happen.
3. It can help you to train and estimate the model.

The reason that we call this library TensorFlow is that it is going to take input in as an array that is multi-dimensional, which is also called a tensor. You are able to construct a type of flowchart of the operations that you use, which is called a graph, that all comes together to perform on your input. With this one, the input is going to head in one end, and then will flow through the system, with many operations in the process, until it comes out on the other side as the output that you will rely on.

TensorFlow is also able to meet some hardware and software requirements. These are going to include a few classifications that help the program run smoothly. First, there is the development phase. This is going to be part of the program where you work to train the model. Training is going to be done on a laptop or a desktop in most cases.

Then we can move on to the phase that is known as the run phase, r the inference phase. When all of the training is done, the TensorFlow program is going to be available to run on any operating system or platform that you would like to use. This means that the programmer will be able to take their model and run it on some of the following:

1. A desktop that is able to use the Linux, macOS, or Windows operating systems.
2. The cloud as a web service.
3. A mobile device including Android and iOS.

You can also train the model that you are doing on more than one machine. This allows it to run on a different machine as well, as soon as the model is completely trained.

To add to some of this, we are able to train the TensorFlow model to work with CPUs and GPUs. GPUs, until late 2010, was only used for video games. But since it was discovered that the GPU is also a good one to work with when handling algebra and matrix operations, and can handle these kinds of calculations in a quick manner, it is becoming more prevalent to see these shows up in the various models that are designed.

All of these libraries can be great additions to help us get some of the work that we want to be done in a short amount of time. It is not always as easy as it may seem in the beginning, but working with some of the right codings, and deciding what kind of library we want to work with when handling a Data Science or deep learning project, can be the number one key to ensuring that we will actually get this set up and ready to go and see results with our work.

Chapter 13: Practical Examples of Python Data Science

Now that we have spent some time looking at Data Science and how Python data can be used with it, it is time for us to work on the next part that comes with this. We are going to look closer at some of the practical examples that we can do when it comes to writing out codes to create the models you need in Data Science. There are a lot of different options that we are able to work with here, and you can choose the one that is the best for your needs and will help you to create the model that helps us sort through the data we want. Some of the best examples of how you can complete your own Python Data Science model will include:

K-Means Clustering

The first example that we are going to take a look at when it comes to Python Data Science is the K-means clustering. Clustering is a simple but common process that helps with a lot of data analytics projects with Machine learning, and it can help us to take all of the data points that we want to work with and then divides them up into groups. The points that are in the same cluster are going to share a lot of similarities to one another, and the ones that are in other clusters are going to not be as similar to the others.

For example, you may decide to go through the data points that we want to look at, and then separate out the customers into females and males. With this one, we are going to end up with two clusters, and when the algorithm is able to sort through all of the data it has, they will get all of the points in your data to fit into one out of the two clusters.

That was a pretty simple option to work with, but we can go through and sort through the data in a different manner. You could create a model where you want to learn more about your customers and see who is the most likely to purchase a certain product, and which group they are likely to fall into. This is going to help us with some of the marketing and the sales that you would like to work with. In this case, maybe you would need five clusters so that you can get the best idea of the customers you are working with, and even some of the outliers to see if there is a new customer base you can reach through.

The idea that we will see here is that any of the objects or the data points that are in the same cluster are going to be the ones that are related to one another closely. And if something shows up in a different cluster, they won't really share similarities with one another. The amount of similarity that comes with this is going to be important because it is going to help us learn the main metric that we are able to use to help us see how strong the relationship becomes between two or more objects that we want to look over.

If you want to work with k-means clustering, it may sound complicated and you are worried that it is going to not work for the project that you want. But this code is not meant to be too difficult, and when it is done with the Python language, it can really make a strong model that works well with Data Science. An example of the kind of code that you will want to write out when it comes to the k-means clustering algorithm will include:

- *import numpy as np*
- *import matplotlib.pyplot as plt*

```
def d(u, v):
    diff = u - v
    return diff.dot(diff)


def cost(X, R, M):
    cost = 0
    for k in xrange(len(M)):
        for n in xrange(len(X)):
            cost += R[n,k]*d(M[k], X[n])
    return cost

def plot_k_means(X, K, max_iter=20, beta=1.0):
```

```
N, D = X.shape
M = np.zeros((K, D))
R = np.ones((N, K)) / K

# initialize M to random
for k in xrange(K):
    M[k] = X[np.random.choice(N)]

grid_width = 5
grid_height = max_iter / grid_width
random_colors = np.random.random((K, 3))
plt.figure()

costs = np.zeros(max_iter)
for i in xrange(max_iter):
    # moved the plot inside the for loop
    colors = R.dot(random_colors)
    plt.subplot(grid_width, grid_height, i+1)
    plt.scatter(X[:,0], X[:,1], c=colors)

    # step 1: determine assignments / resposibilities
    # is this inefficient?
    for k in xrange(K):
        for n in xrange(N):
            R[n,k] = np.exp(-beta*d(M[k], X[n])) / np.sum( np.exp(-beta*d(M[j], X[n])) for j in xrange(K) )

    # step 2: recalculate means
```

```
        for k in xrange(K):
            M[k] = R[:,k].dot(X) / R[:,k].sum()

        costs[i] = cost(X, R, M)
        if i > 0:
            if np.abs(costs[i] - costs[i-1]) < 10e-5:
                break

    plt.show()


def main():
    # assume 3 means
    D = 2 # so we can visualize it more easily
    s = 4 # separation so we can control how far apart the means are
    mu1 = np.array([0, 0])
    mu2 = np.array([s, s])
    mu3 = np.array([0, s])

    N = 900 # number of samples
    X = np.zeros((N, D))
    X[:300, :] = np.random.randn(300, D) + mu1
    X[300:600, :] = np.random.randn(300, D) + mu2
    X[600:, :] = np.random.randn(300, D) + mu3

    # what does it look like without clustering?
    plt.scatter(X[:,0], X[:,1])
    plt.show()
```

```
    K = 3 # luckily, we already know this
    plot_k_means(X, K)

    # K = 5 # what happens if we choose a "bad" K?
    # plot_k_means(X, K, max_iter=30)

    # K = 5 # what happens if we change beta?
    # plot_k_means(X, K, max_iter=30, beta=0.3)

if __name__ == '__main__':
    main()
```

As you take a look through some of that code, it may seem like a lot to handle, and you may be worried that you won't be able to make this work for your needs at all. But remember that in the beginning, we are just pulling a lot of libraries out, and then we simply ask them to take the various points of data that we are working with and turn them into a form that is easier to understand after reading through.

Being able to create these clusters, and then reading through the data to determine where each of the points of data should fall, and which cluster they fit into, is a challenge, but this is where the K-Means clustering algorithm is going to come into play. With the code above and a little bit of practice, you will be able to make all of this come together for you.

Neural Networks

Now that we have had some time to discuss what the K-Means clustering algorithm is all about, it is time for us to move on to the second example that we are going to work with is known as a neural network. When it comes to working with neural networks, we will see that they are really powerful codes, and it can take us some time to learn the best way to handle them.

Once you learn how to handle the neural networks, we will find that they will help us to handle the codes, and this shows us why so many programmers like to add these into their data analysis. But let's explore it a bit more to

learn how it can work with artificial intelligence, why it fits in with Machine learning, and how we can use this to create the best models for our Data Science project.

These neural networks are going to be set up in a manner that they can teach our system how to think similar to the human mind. Through learning and remembering some of the things it has learned and done in the past, the neural network is able to become ‘smarter” and can make better decisions in the future. The more work that it does, the faster and more efficient it will get at doing these tasks. You will quickly find that when you create a model that relies on neural networks, it will be really good at its job.

Neural networks are useful simply because they are able to work in a manner that is similar to what we see with the human brain. It is done through our chosen system or machine; however, to get it to work in this manner, we will need to make sure that we rely on the right procedure to get it done. If we end up picking out bad information or information that is low-quality, then it is going to train the algorithm in the wrong way, and it won’t give you the results that you want.

Always remember with the neural networks, the higher the quality of data that you can feed into the machine, the better. This helps us to make sure that the model is going to work the way that we want. With this higher quality information, we will ensure that the neural network will give the right predictions and insights. And in the process, we will have the system learn what it needs to do in an accurate and faster method than before.

Keep in mind with this one that when creating our Neural Network, there are going to be many parts that have to come together. But for now, we are going to take a detour and focus on the code that is needed to get the neural network started and ready to go. While the neural network may seem difficult to work with, but there is a simple code that we can work with to get the basic neural network ready to go. The code that we need to work with, using the Python coding language along the way, to go through our data and make sure that the model learns along the way includes:

- *import torch*
- *import torch.nn as nn*
- *import torch.nn.functional as F*

```
class Net(nn.Module):

    def __init__(self):
        super(Net, self).__init__()
        # 1 input image channel, 6 output channels, 3x3 square convolution
        # kernel
        self.conv1 = nn.Conv2d(1, 6, 3)
        self.conv2 = nn.Conv2d(6, 16, 3)
        # an affine operation: y = Wx + b
        self.fc1 = nn.Linear(16 * 6 * 6, 120)  # 6*6 from image dimension
        self.fc2 = nn.Linear(120, 84)
        self.fc3 = nn.Linear(84, 10)

    def forward(self, x):
        # Max pooling over a (2, 2) window
        x = F.max_pool2d(F.relu(self.conv1(x)), (2, 2))
        # If the size is a square you can only specify a single number
        x = F.max_pool2d(F.relu(self.conv2(x)), 2)
        x = x.view(-1, self.num_flat_features(x))
        x = F.relu(self.fc1(x))
        x = F.relu(self.fc2(x))
        x = self.fc3(x)
        return x

    def num_flat_features(self, x):
        size = x.size()[1:]  # all dimensions except the batch dimension
        num_features = 1
```

```
    for s in size:
        num_features *= s
    return num_features


net = Net()
print(net)
```

Many programmers worry that they will not be able to work with neural networks because they feel that these networks are going to be too difficult for them to handle. These are more advanced than what we will see with some of the other forms of coding, and some of the other Machine learning algorithms that you want to work with. But with some of the work that we did with the coding above, neural networks are not going to be so bad, but the tasks that they can take on, and the way they work, can improve the model that you are writing, and what you can do when you bring Python into your Data Science project.

Conclusion

Thank you for making it through to the end of *Data Science Python*, let's hope it was informative and able to provide you with all of the tools you need to achieve your goals whatever they may be.

The next step is to start putting the information and examples that we talked about in this guidebook to good use. There is a lot of information inside all that data that we have been collecting for some time now. But all of that data is worthless if we are not able to analyze it and find out what predictions and insights are in there. This is part of what the process of Data Science is all about, and when it is combined together with the Python language, we are going to see some amazing results in the process as well.

This guidebook took some time to explore more about Data Science and what it all entails. This is an in-depth and complex process, one that often includes more steps than what data scientists were aware of when they first get started. But if a business wants to be able to actually learn the insights that are in their data, and they want to gain that competitive edge in so many ways, they need to be willing to take on these steps of Data Science, and make it work for their needs.

This guidebook went through all of the steps that you need to know in order to get started with Data Science and some of the basic parts of the Python code. We can then put all of this together to create the right analytical algorithm that, once it is trained properly and tested with the right kinds of data, will work to make predictions, provide information, and even show us insights that were never possible before. And all that you need to do to get this information is to use the steps that we outline and discuss in this guidebook.

There are so many great ways that you can use the data you have been collecting for some time now, and being able to complete the process of data visualization will ensure that you get it all done. When you are ready to get started with Python Data Science, make sure to check out this guidebook to learn how.

Finally, if you found this book useful in any way, a review on Amazon is always appreciated!