PYTHON DATA SCIENCE

HOW TO LEARN STEP BY STEP PROGRAMMING, DATA ANALYTICS AND CODING ESSENTIALS TOOLS.

TONY HACKING

Chapter 1: The Basics of Data Science

Data science is a role that is taking up a lot of space for many businesses. There is a wealth of information out there that they are able to use for their own advantage, but they just need to know where to gather it, and how to analyze all of that data for their own needs. Sometimes, this is going to be a process that takes a lot of time and effort and can be hard to keep up with and ensure that we are doing it in the right manner.

Data science is the process of gathering, organizing and cleaning, analyzing, and then visualizing data so that we can use that information to make smart business decisions. It is becoming more and more important to a lot of businesses, and it is likely that this will take over as one of the main forms of making big decisions in the future. With that in mind, let's take some time to look more in-depth at data science and how businesses are using it for their own needs.

Why Is Data Science So Important?

The first thing that we need to take a look at here is the idea of why data science is so important and helpful for our needs. In a traditional manner, and in the past, the data that businesses were supplied with would be structured and smaller in size. This would make it so much easier to go through the data and see what was there, and often the business intelligence tools were all that was needed to see what was available and what business decisions needed to be made.

However, the data that we are able to take a look at today is so much different. Unlike the traditional data systems, which was mostly structured, it is common for the data that companies collect today to be unstructured, or at the very least, semi-structured. This is going to make it more difficult to sort through and understand, and that is why the process of data science has expanded into what it is today.

The unstructured data that companies are collecting today is going to be produced from distinctive sources like text files, financial logs, multimedia forms, devices, and sensors. The straightforward business intelligence tools that were so popular in the past are not able to handle the job, and businesses are turning to more complex and advanced kinds of tools and algorithms to get some of the analysis done and to help with getting all of the meaningful information and insights out of that data.

But this is not the only reason why businesses are finding the process of data science to be so popular and helpful to what they are doing. Data science is able to help the company recognizes the definite requirements of their customers based on the actual data that is out there. Data science is able to help us handle some things like predictive analysis to figure out what is the most likely outcome of a decision. And it can help us figure out how to make the right decisions that will push our business into the future and beat out all of the competition that is out there.

Many businesses, no matter what kind of industry they conduct business in, will find that working with data science is one of the best options for them. Data science is able to help them to really learn about their industry, and even gain a leg up on the competition. Many of the companies out there are going to already collect a lot of data and information about things like the competition, the industry, and their customers, and data science is going to help them to actually see what insights and information are inside of that data and use it for their advantage

There are many times when bringing out data science is going to be beneficial, and it will be able to propel your business forward more than anything else is able to do. When we are able to focus on the data and the process of analyzing it and seeing what good insights and predictions are inside, we will be able to make accurate decisions that will help us to make a big difference. Companies who have been able to implement a successful data science project from beginning to end are the ones who are doing the best overall in their respective industries.

More about Data Science

With the above in mind, it is time to not only look at some of the benefits that come with data science, but we also need to take a look at some more information about what data science is all about, and how a business is able to use data science for their own needs.

The use of data science is becoming ever more prominent in many businesses and in a lot of different industries as well. But that doesn’t really explain to us what data science is all about? We may want to figure out what it takes to do one of this analysis, how tools we need to make the predictions, and so much more. These are all questions that we are able to answer in this part of the guidebook to make things a bit easier.

First, we need to see what data science is all about. Data science is going to be a blend of algorithms, tools, and principles that come with machine learning. And all of these different things come together with the intention of discovering covert patterns from the raw data. How is this going to be so much different than what we have seen statisticians doing for years now? The answer that we are going to get for this one is going to lie in the difference between predicting and explaining something.

A data analyst is usually going to be that person who will explain what is taking place by preparing the history that comes with the data. But then a data scientist is a bit different because they are not only working through some of that exploratory analysis from above, but they are going to use a lot of different algorithms that are advanced in machine learning to help them figure out how likely a certain event in the future is. A data scientist is able to take that data and look at is from a variety of angles, and hopefully, see it from a few angles that were not even explored in the past.

There are actually a few different types of analysis and analytics that we are able to work in order to make some smart decisions and use the data that we have to our advantage.

These are going to include the following:

The predictive causal analytics. This is the one we would want to use anytime that we have a model that can predict how likely an event will happen in the future. So, if you would like to provide someone with a loan on credit, you would want to use this kind of analytics to figure out how likely it is that the customer would make their payments in the future. We are able to build up a model that can perform these analytics based on how well that customer has made their payments in the past.

Then we can move on to a method that is known as prescriptive analytics. This is the one that we use if we want to work with a model that has the intelligence and the information that is needed to make its own verdict, and you still get the capacity to revise the model with your own dynamic criterions. This is a newer field and there are still some kinks that are being worked out with it, but it can often help us with getting advice and knowing what course of action to take out of several options that are available.

The third option that we are going to spend some time focusing on here is how

The third option that we are going to spend some time focusing on here is how we can work with machine learning to help us make some good predictions. If you are working with something like transactional data in the financial world, and you would like to build up a new model to determine the future trend that maybe there, then you will find that a few of the algorithms that work with machine learning will be the best.

This is going to be a version of supervised learning because we have to provide the system with a bunch of examples so that it can learn how to behave over time. A good example of how this one will work is when we create a model or algorithm that can detect fraud, based on historical records of purchases that ended up being fraudulent.

Finally, the last part that can be included in this is using machine learning to help us discover some new patterns in our data. If you are going into the data and you are not sure what parameters you will work with to help make some predictions, then it is time to dig in and find out what patterns are present in your batch of data that you can then employ to make some predictions that are pretty meaningful. This is going to be the same as an unsupervised learning model because the program has to go through and see what is found in the data, without any options from you or parameters, and figure out what is there that you can actually use.

The Importance of the Data Science Lifecycle

The next thing that we will take a glance at is the data science lifecycle. This is an important step because it helps us to really see why this process is so important, and why so many businesses prefer to work with this process and make this a part of their whole process. Let us take a look at some of the stages and steps that come with the data lifecycle to help you get started.

The first phase of this lifecycle is going to be the discovery. Before we get started on any data science project, we need to understand all of the different parts of the data. This is going to be the place where we will understand our various specifications, requirements, and priorities, as well as the budget that is required before we get started. This is basically the stage where we will focus on gathering up the information that we need in order to handle the model and algorithm that we want to work with during this time.
Now, to make this work, we need to acquire the qualification to ask the right questions. This ensures that we can take all of the information and actually find

something useful inside all of it. Here we have to evaluate if you currently have the necessary assets present in conditions of the people to handle the model, the technology that the model needs, the time, and the right kind of data to support the project and all of the work that you want to do. In this phase, we need to make sure that we frame the business problem and come up with the initial hypotheses that we want to work with.

It is hard to prepare all of the models that we want in data science if we do not have the data in front of us. This is the stage where we both go out and find the data that we need or sort through all of the data that we already have to see what is inside. This process may take some time, but finding the right data and making sure that it is high quality is one of the most important aspects of this process.

The second phase that we need to take a look at, once we have collected all of the data that we need is time to work with data preparation. In this phase, you will need to work with an analytical sandbox in which we are able to accomplish analysis for the whole span of the project. This helps us to make sure that the data is going fit into our model and be ready to go and will work in the model that we want.

Because most of the data that we want to collect will be found from a variety of sources and will be really unstructured. This means that all of the data will be in different formats, different quantities, and they will have to make sure that it is ready to work well.
If the data stays in different formats or has a bunch of duplicates or missing information, then it is going to really mess up with the algorithm that we are working with. Taking some time to prepare the data and clean it off, and making sure that it makes sense for what we want to create with the algorithm can help us to get the best results.

In this stage, we also need to make sure that we are performing ETLT, or extract transform, load, and transform, to make sure that we get the data into our sandbox. You are able to work with the R language, even though Python is one of the preferred coding languages to adopt with this. This is going to aid you to spot some of the outliers and establish a relationship between the variables. Once you have been able to clean and then prepare the data, it is time to do some preliminary analytics on it. This will ensure that the data is ready to go, we can pick the right algorithms to see results, and that the insights that we need will show up.

The third phase that we need to explore here is the idea of model outlining. Here

The third phase that we need to explore here is the idea of model outlining. Here, you will be able to regulate some of the means and styles in order to draw the relationships that are present between the variables. These are important relationships that we want to focus on because they will see the foundation for the algorithms, which we can enforce in the next chapter. We want to work with the EDA, or Exploratory Data Analytics, using a lot of different statistical formulas and visualization tools. There are a number of tools that we are able to work with here when we are ready to get started. Some of the best ones to work with will include:

1. R. This is going to be a great programming language to go with because it provides us with a complete set of modeling capabilities and can provide us with the environment that we need to build our own interpretive models.

2. Python: As we will discuss in this guidebook, the Python language is another great choice to go with when we want to work on data science and some of the machine learning algorithms that we need for this. There are a lot of great libraries and extensions with Python that can help us to get some of our work done.

3. SQL Analysis services: These are helpful because they help us perform in-database analytics that is common for helping with data mining functions and some of the basic predictive models that we need.

4. SAS/ACCESS: This is something that we can use to help approach data from Hadoop and it is utilized to create sustainable and reusable model flow drafts

Although there are a lot of different tools that we are able to use when it comes to working on our data science projects, R is often one of the most commonly used tools that we can focus on. Once we have been able to use these tools and we have some great insights into the nature of the data, and we know which of the various algorithms we would like to use, it is time to move on to applying these algorithms and then build up a new model.

Now it is time to move on with model building. When we get to this phase, we are going to develop a set of data that we want to use in order to train and test

our work. This will ensure that the model when we are ready to use it will be able to go through the data and make some accurate predictions and provide us with the insights that we need.

When we do this process, we are going to consider whether the tools that we already have are going to be enough to get it all done.

The models take time and money and computer resources, and it is often best if we are able to find someone who can handle running them and doing the coding that is needed. If something is lacking out of these, then it is going to cause some issues in the results that you will see. You may need to add in a more robust environment, including processing that is parallel and fast.

A business is able to use a variety of learning techniques to help you get this all done. Some of the techniques that we are able to work with will include clustering, classification, and association to help get that model done. We are then able to use a lot of different tools to build up the models that we would like to use including Matlab, Statistica, Alpine Miner, WEKA and more.

The fifth phase that we are going to take to get done with data science is operationalized. When we enter this phase, we are going to deliver some final reports, briefings, code, and technical documents.

This is where we are going to be able to make the model work, and see which methods we should take based on that data.

After we have been able to go through and make the model and the algorithm work the way that we want, and it has been trained and tested properly. This makes it easier for us to go through and figure out what predictions and insights are in that data. It is now time for us to really get an idea of what business decisions we should take in the future. Many companies like to spend time looking through their data to help them make smart and information-driven decisions. Whether it is about customer service, which products to offer, or something else, the process of data science will be able to handle that. And this is the step that can help us out.

In addition, sometimes the company will choose to work with a pilot program because this can help them to see how their decision will behave before they try to implement it in the whole company. This provides them with a clear picture of the performance and some of the other associated restraints that can come up

the performance and some of the other associated results that can come up with that program on a smaller scale and then can consider whether we want to take that implementation before deploying it fully in the company.

And then we can reach the final step of this process, which is where we are going to communicate our results. Now, when we get to this stage, we will find that it is critical to appraise whether or not you were able to achieve your goal, the goal that you came up with and decided on in the first phase.
We want to see whether the goals are met, and if we actually have seen some of the results that we want, or if we need to see what other insights and processes are going to be in the data for us to utilize.

In this final phase, we are going to make sure that we can identify all of the key findings that are in the process, and then we need to also be able to communicate the information to the stakeholders. This helps us to determine whether the results that we got out of that model or out of this project will be a failure or success. And we figure this out based on the criteria that were developed during the first phase.

During this time, we have to make sure that we use any means that are necessary to handle and show off the information. This means that visualizations are often going to be the kind that we need to bring out and use for our needs here. This helps us to see some of the graphs and charts to make it easier to notice what insights and information were found in that data. For most people, this makes it easier to see the complex relationships that show up in the data that we look through, and it can make life easier in most cases when it is time to make our own predictions and decisions based on that data.

Data science is a really great process to help us learn more about our customers and the industry that we are going to work with.

This helps us to really gain an advantage over the competition and will ensure that we can get some of the best results possible.

Companies who have embraced the idea of data science and are happy to use them to make sense of all the data coming into them, and to help them really learn how to make themselves unique from others in the same industry are really seeing results.

Chapter 2: Using the Python Language

Working with data science and some of the different parts that come with this can be an exciting time. There is so much data that is held in your storage, and it is likely that you have been anxiously awaiting a time to look through it and see how that data is able to work for your needs. But once you take a look at a large amount of data that is available, it is likely that you will feel overwhelmed and like it will be almost impossible for you to get through it all. With the help of the Python language, you will find that it is actually pretty easy to get started with coding and creating your own models and algorithms to go along with your data science, and you will be able to read through all of that information in no time.

There are a lot of reasons why so many programmers are interested in using Python to help with all of their data science needs. Yes, there are a lot of other coding languages that you can use with data science and machine learning, but Python is one of the best out there. Some of the benefits of working with the Python coding language, and all it has to offer includes:

1. It is easy for a beginner to learn how to work with.

2. It has all of the power that you need to handle complex data science algorithms and models.

3. Lots of extensions and libraries that can handle machine learning and get the work done.
4. It is compatible with other coding languages so you can combine them together to make the best models possible.

5. A big community to ensure that you can really learn how to work with this language, ask the questions that are needed, and so much more in a short amount of time.

These are just a few benefits that come with using the Python language, and we can quickly see why so many people would want to add this into their knowledge base and make sure that it can be used with data science and so much more. With that in mind, let’s take a look at how to download this coding language and how to get it set up and ready to go on our system in no time.

Which Version of Python to Work With?

Because Python is a type of interpreted language, it is going to provide us with a lot of benefits over some of the other programming languages out there. One of the advantages that you will notice is that it is able to grow and make some big changes as all of your computing needs to evolve and make changes as well.

Just like some of the applications on your desktop, the fact that the Python language is going to continually develop makes it easier for us to add in some new features and it will not take long for us to see refinements added into Python, which make it easier to use.

Throughout the years though, we will find that there are actually quite a few versions of this language that have been released for programmers. Each one of these is going to be a bit different in the benefits and features that it is able to provide compared to the version that came out before it. Some of the options that you can choose when you are ready to work with the Python language will include the following:

Python 2.X

This version was released in 2000 and has gone through several improvements since then. The 2.7 version was released in 2010 and is the most current version of this branch right now. Even though there aren't any more plans for the future development of this version, there are still some reasons that it is the one you would want to go with.

First, if you have any familiarity with 2.X or it is already installed on your computer, it is an easy version to work with and can save some time and hassle. It can do a lot of the coding that you want and it has many of the features that you will need, without a lot of extras that could slow down the system.

You may also have a few programming needs in your organization that would do better with older technology, then the 2.X is the best option. For example, if your organization has policies that discourage or ban the installation of unapproved software from outside sources, then this version may be the best one for you to choose. Many times this version of Python was already installed on the system so this is the one that you would want to use.

In addition, there are many third-party libraries and packages that are used to help extend what capabilities the 2.X version can handle, and some of them are

not present in the newer 3.X version. If you want to work with a specific library for your application, you may find that it is only available in the 2.X version. You would need to download this version to get it to work for you.

If you do decide that this version is the best one, you should still take a look at the 3.X version. There are some differences in the best programming practices of each and you may even be able to make slight modifications to your code on the 2.X version and get it to work on the newer version.

Python 3.X

Many of those who decide to work with the Python coding language are going to choose one of the versions of Python 3 to help them get started. The original Python 3 was released in 2008, and it provided programmers with a lot of big improvements to the system since that time. There are a few different versions of this one available, so checking out the www.python.org website to see which one is the newest when you are ready to start programming is often a good idea.

As one of the most current versions, it is the one that is going to be suggested the most as we work to develop on this language and see more results over time. You can work with some of the older versions of Python if you would like. Just be aware that Python 3 is going to be the one that is the most current, the one that is being developed the most, and the one that most programming is going to be done with.

Installing the Python Language

Now that we know a bit more about the Python language and the different options that you have when it is time to work with choosing what to program with, it is time to go through and install the Python language for your needs. We are going to spend our time looking at how to download Python from the www.python.org website.

There are other options that you can choose with, and some of them offer some different features and more that may suit your needs better. But for this one, we will know that all of the additions and files that are needed to get the Python programming up and running will be present, and we won't have to go searching for them later on. This is not something that is guaranteed with some of the other options. With this in mind, let's dive into how we can install the Python coding language on Windows, Mac, and the Linux operating systems.

Mac OS X

The first operating system that we are going to explore here is the Mac OS X. if you have a computer that has this kind of operating system already on it, then you will find that the Python 2 is going to be there. Depending on how old your computer is and how long you have had it, the version of Python 2 that is on your system can vary.

If you would like to check which exact version is present on your computer, then you can type in the code below to your command prompt:

python – V

This process is going to show us the version of Python that is on our operating system so there should be some number that comes up such as Python 2.2. You can also choose that you want to install the Python 3 to your system rather than having to work with the Python 2 version if you would like. To check whether or not there is already a version of Python 3 on your computer, just to make sure, you need to make sure that the terminal app is open and then type in the prompt below:

Python3 – V

Keep in mind with this one that the default of Mac computers is that the Python 3 is not going to be on the system and you will have to take the steps that are necessary to get it on there. If you would like to install the Python 3 program, you will be able to visit that website we talked about before and use some of the installers that are there. This will provide you with an option that is compatible with the Mac operating system and will ensure that all of the right files and folders will be present on the system when you need them.

Being able to run the IDLE and the Python shell is going to be dependent on which version you choose and some of your own personal preferences. You can employ the subsequent commands to help you start the shell and IDLE applications:

- For Python 2.X just type in “Idle”
- For Python 3.X, just type in “idle3”

As we mentioned before, when you download and install Python 3, you are

going to install IDLE as a standard application in the Applications folder.

In order to start this program from your desktop, you simply need to open up that folder and then double click on the IDLE application.

Windows System

The next operating system that we are going to spend some time here is the Windows operating system. It is easy to add on the Python language when you are working with this operating system, just keep in mind it may take a few extra steps. Microsoft has its own programming language, and that language is installed with this operating system, rather than the Python language.

So you will not find Python present on the computer unless someone actually went through and installed it for you.

This doesn't mean that Python won't work on this kind of system. It simply means that you have to go through some steps to get it all done. Once the Python language is present on your computer, it will work just fine and you can complete all of the codings that you would like to do.

With this in mind, the steps that you will need to take in order to get the Python program set up on your computer and ready to go include:

1. To set this up, you need to go to the authorize Python download page and take the Windows installer. You can select to do the latest version of Python 3, or go with another option. By default, the installer is going to provide you with the 32-bit version of Python, but you can choose to switch this to the 64-bit version if you wish. The 32-bit is often best to make sure that there are not any compatibility issues with the older packages, but you can experiment if you wish.

2. Now right-click on the installer and select "Run as Administrator." There are going to be two options to choose from. You will want to pick out "Customize Installation"

3. On the following screen, make sure the entire of the boxes below "Optional Features" is clicked and then click to move on.

4. While under Advanced Options" you should pick out the location where you want Python to be installed. Click on Install. Give it some time to finish and then close the installer.

5. Next, set the PATH variable for the system so that it includes directories that will include packages and other components that you will need later. To do this use the following instructions:
    a. Open up the Control Panel. Do this by clicking on the taskbar and typing in Control Panel. Click on the icon.
    b. Inside the Control Panel, look up the Environment. Then pick Edit the System Environment Variables. From here, you can click on the button for Environment Variables.
    c. Go to the section for User Variables. You can either edit the PATH variable that is there, or you can create one.
    d. If there is not a variable for PATH on the system, then create one by clicking on New. Make the name for the PATH variable and add it to the directories that you want. Click on close all the control Panel dialogs and move on.

6. Now you can open up your command prompt. Do this by clicking on Start Menu, then Windows System, and then Command Prompt. Write in "python." This is going to load up the Python interpreter for you.

From this point, we should have a program of Python on our computer and read to work with the Windows operating system. You can choose to open up the interpreter that comes with Python and everything else, including all of the necessary files, and then start writing out any of the code that you need.

Linux Operating System

And finally, we need to take notice of how we can install the Python coding language on our Linux operating system. While Linux may not be used as often as the other two operating systems, it definitely has made a name for itself and is a great program to work with when you need to do coding and more. Python is going to work on this system very well, and it is one of the best options that you can focus on to get some good results.

The first step that we need to take when we get to this part is checking to see whether our operating system has Python 3 already present. Some of these do, and some of them do not, so it is always worth it to open up the command line and type in the following code to check before we do the rest of the work that is needed:

```
$ python3 - - version
```

If you are on Ubuntu 16.10 or latest, then it is a simple process to install Python 3.6. you just need to use the following commands:

```
$ sudo apt-get update
$ sudo apt-get install Python3.6
```

If you are relying on an older version of Ubuntu or another version, then you may want to work with the deadsnakes PPA, or another tool, to help you download the Python 3.6 version. The code that you need to do this includes:

```
$ sudo apt-get install software-properties-common
$ sudo add-apt repository ppa:deadsnakes/ppa
# sudo apt-get update
$ sudo apt-get install python3.6
```

The good thing to consider here is that if you are working with some other kind of distribution of Linux, it is pretty likely that there is going to be some version of the Python 3 library already installed on your system. If not, then you will be able to use the steps above to make sure that you get this coding language started on your computer and ready to go the way that you would like.

With the steps above done, you can now see that the Python coding language is ready to go and you will be able to get the results you want with some playing around on the system.
Take your time to learn more about the system and how it works, and then get ready to see how we can use this with all of our data science projects.

Chapter 3: Some of the Basic Fundamentals of Python

While we are on the topic of the Python language, we need to spend some time looking at some of the basics that come with this language. The Python language is relatively easy to learn about and work with, so that should be welcome news to those who are brand new into coding and have not been able to work with it in the past.
With this in mind, let's dive in a bit and learn more about how to work with the Python language.

The Python Keywords

First, we are going to start with the Python keywords. These keywords are reserved to tell the compiler command. You do not want to use them anywhere else in the code, and you must make sure that you use them properly. If you try to use the keyword in the wrong place of the code, it will result in an error.

These keywords are there to give commands to the compiler so that it knows how it should react to your code. They are important to the code and to the compiler so make sure to only use them where they are needed and as a command to the compiler

How to Name an Identifier

The next thing that we need to take a look at when it comes to our own coding is how to name the identifiers. These identifiers are important in the code and there are actually quite a few of these identifiers that we need to pay attention to when we work on this kind of coding language. You will find that they come in different names, and you may see them as things like classes, entities, functions, and variables to name a few when you are ready to name one of your identifiers, you can use the same information, as well as the same rules, when you name each of them. This can make it a bit easier to remember rather than having different rules for each type.

Now, the first rule that we have to focus on when it comes to these identifiers is that there are a lot of options and you are pretty free here. You are able to work

that there are a lot of options and you are pretty free here. You are able to work with all kinds of letters, whether they are uppercase or lowercase or a combination of the two. Numbers and the underscore symbol are also allowed as well. You can combine all of these characters together as well.

One thing to remember when we are going to name these identifiers is that you are not able to start the name of one of them with a number, and you never want to have a space between the words that you are writing. So something like 4birds or four birds would not be allowed, but writing out four birds or four_birds would be just fine.

While it is not necessarily one of the rules that you have to follow, something to consider when you are naming these is to go with a name that you will be able to remember.
This makes it easier to find that identifier and call it back out later on if you would like to use it in the code. It just makes things easier when it makes sense for the part of the code that you are in, and that it is something you can remember later.

The Python Statements

We can also spend a moment on the Python statements. These are pretty simple to work with, but it is still a good idea to spend some time looking at them and seeing how they all fit together to do the work that you would like.

So, to explore this further, the statements are going to be the strings of code that you write out and that you would like the compiler to show up on your screen. When you give the instructions over to the compiler for what you want the statements to say, it is going to put that information up on the screen. As long as you write them out properly, then the compiler will make sure that the message you are asking for will be on the screen and ready to go on time. it is as simple as that!

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.

Bringing In the Python Variables

The next thing on the list that we can explore is the Python variables. These are another important part of the code that we need to spend some time on, mainly because they show up in the code so much and are so common, that you will see them quite a bit.

These variables are going to show up in your code in order to help store and hold onto the values that are important to helping your code function in the right manner. This helps everything to stay as nice and organized as you would like.

You can easily add in some of the values to the right variable as you would like, and it only takes adding in the equal sign between both of them.

The Operators

Operators are pretty simple parts of your code, but you should still know how they work. You will find that there are actually a few different types of them that work well. For example, the arithmetic functions are great for helping you to add, divide, subtract, and multiply different parts of the code together. There are assignment operators that will assign a specific value to your variable so that the compiler knows how to treat this. There are also comparison operators that will allow you to look at a few different pieces of code and then determine if they are similar or not and how the computer should react based on that information.

The Python Functions

Another topic that we need to take a moment to explore with this language is known as the Python functions. These functions can basically be a set of expressions, and sometimes they are given the name of statements, that are capable of being named or being kept anonymous depending on what you would like to see done with your code. These are going to be some of the very first-in-class objects for the code, which means that you are not going to spend a lot of time worrying about the restrictions that come with these.

When working on these Python functions, you will be able to use them in a manner that is similar to other values, including values like strings and numbers, and they will have other attributes that we are able to pull out and use in any manner that we would like. The good news with the functions is that they are pretty diversified to work with, and there are a lot of different attributes that we are able to use in order to create and then bring out these functions. Some of the choices that we can have with these kinds of functions include some of the following:

- __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.

Now, keep in mind here that there are going to be a few different things that we are able to do with these functions, such as bringing them out to pass an argument from one function to another, as needed.

Any function that is able to take on a new one as the argument is going to be the

higher-order function in our code.

The Python Classes

No discussion on the basics of Python will be complete without a discussion about the classes. These are going to be all about how the code in Python is organized, and how we can make sure that the parts come together and do what we would like in the end. These classes are going to be simple containers in the code that can hold onto the objects that we want to hold onto in our code. We have to make sure that the naming of the classes is done right and then put the objects in them the right way, but this can ensure that the classes are going to work the way that we want.

The neat thing about these classes is that they can hold onto and store anything that we would like. But keep in mind that the objects that are in one class make sense that they go with one another, and won't confuse others as to why they show up together. The items don't have to turn out to be identical at all, but they do need to make sense with each other. You could have a class of food, vehicles, or colors if you would like, but they do need to make sense of how they go together.

Classes are going to be very important when it is time to write out one of the codes that we would like to use. These classes are responsible for holding onto the various objects that we would like to work with. They can also ensure that it is easier to bring out all of the different parts of our code when we need them to execute and behave the way that we would like them to.

As you can see, there are a lot of different parts that will come together to write out a good code in the Python language.
They are sometimes a bit hard to get started on, and it may seem like a lot of information. But in reality, all of this goes together and makes the coding easier than ever before. Make sure to learn these parts and understand how they go with one another so that you can start writing out some of your very own codes in no time.

Chapter 4: The Python Data Types

The next thing that we need to take a look at is the Python data types. Each value in Python has a type of data.

Since entirety is an object in Python programming, data types are going to be like classes and variables are going to be the instance, which is also known as objects, of these classes. There are a lot of different types of data in Python. Some of the crucial data types that we are able to work with includes:

Python numbers

The first option that we are able to work on Python data includes the Python numbers. These are going to include things like complex numbers, floating-point numbers, and even integers. They are going to be defined as complex, float, and int classes in Python. For example, we are able to work with the type() function to identify which category a value or a variable affiliated with to, and then the isinstance() function to audit if an object exists to a distinct class.

When we work with integers can be of any length, it is going to only find limitations in how much memory you have available on your computer. Then there is the floating-point number.

This is going to be accurate up to 15 decimal places, though you can definitely go with a smaller amount as well.

The floating points are going to be separated by a decimal point. 1 is going to be an integer, and 10 will be a floating-point number.

And finally, we have complex numbers. These are going to be the numbers that we will want to write out as x + y, where x is going to be the real point, and then they are going to be the imaginary part.
We need to have these two put together in order to make up the complexity that we need with this kind of number.

Python lists

The next type of data that will show up in the Python language. The Python list is going to be a regulated series of items. It is going to be one of the data types

that are used the most in Python, and it is exceedingly responsive.

All of the items that will show up on the list can be similar, but this is not a requirement. You are able to work with a lot of different items on your list, without them being the same type, to make it easier to work with.

Being able to declare a list is going to be a straightforward option that we are able to work with. The items are going to be separated out by commas and then we just need to include them inside some brackets like this: [ ] we can also employ the slicing operator to help us obtain out a piece or a selection of items out of that list.

The index starts at 0 in Python.

And we have to remember while working on these that lists are going to be mutable.

What this means is that the value of the elements that are on your list can be altered in order to meet your own needs overall.

Python Tuple

We can also work with something that is known as a Python Tuple. The Tuple is going to be an ordered series of components that is the duplicate as a list, and it is sometimes hard to see how these are going to be similar and how they are going to be different.

The gigantic diverse that we are going to see with a Tuple and a list is that the tuples are going to be immutable.

Tuples, once you create them, are not modifiable.

Tuples are applied to write-protect data, and we are generally quicker than a list, as they cannot shift actively. It is going to be determined with parentheses () where the items are also going to be separated out by a comma as we see with the lists.

We can then use the slicing operator to help us wring some of the components that we want to use, but we still are not able to change the value while we are working with the code or the program.

Python Strings

Python strings are also important as well. The string is going to be a sequence that will include some Unicode characters.

We can work with either a single quote or a double quote to show off our strings, but we need to make sure that the type of quote that we use at the beginning is the one that we finish it off with, or we will cause some confusion with the compiler.

We can even work with multi-line strings with the help of a triple quite.

Like what we are going to see when we use the tuple or the list that we talked about above, the slicing operator is something that we are able to use with our string as well. And just like with what we see in the tuples, we will find that the string is going to be immutable.

Python Set

Next on the list is going to be the Python set. The set is going to be an option from Python that will include an unordered collection of items that are unique. The set is going to be defined by values that we can separate with a comma in braces. The elements in the batch are not going to be ordered, so we can use them in any manner that we would like.

We have the option to perform this set of operations at the same time as a union or have an intersection on two sets.
The sets that we work with are going to be unique values and they will make sure that we eliminate the duplicates. Since the set is going to be an unordered compilation. Cataloged has no aim.

Therefore the slicing operator is not going to work for this kind of option.

Python Dictionary

And the final type of Python data that we are going to take a look at is known as the Python dictionary. This is going to be an unordered collection of key-value pairs that we are able to work with. It is generally going to be used when we are working with a very large amount of data. The dictionary can be optimized in such a way that they do a great job of retrieving our data. We have to know the

key to retrieve the value ahead of time to make these work.

When we are working with the Python language, a dictionary is going to be decided inside braces, with every component being a combination in the form of key: value. The key and the value can be any type that you would like based on the kind of code that you would like to write. We can also use the key to help us retrieve the respective value that we need. But we are not able to turn this around and work it in that manner at all.

Working with the different types of data is going to be so important for all of the work that you can do in a Python coding, and can help you out when it is time to work with data science.

Take a look at the different types of data that are available with the Python language, and see how great this can be to any of the codes and algorithms that you want to write into your data science project overall.

Chapter 5: Completing a Data Analysis with Python

It is now time for the big event. We are going to spend some time taking a look at how we can complete data analysis and make it work for our business. Data analysis is going to be mainly the part of data science where we take our data and actually analyze what is found inside of it. If we can do this successfully, we are going to learn some insights and actually figure out how to make predictions that are right for our business and for the future of our company.

Often the problem for most businesses when it comes to data science is not the lack of data. There are so many locations where we are able to collect this data today, and it is often so overwhelming how much we are able to collect and use for our needs. But the problem is that there is just too much information, and now we need to be able to go through it all and figure out how we can use it, and what predictions and insights are found inside of that data as well. This is where data analysis is going to come into play.

With the right process of data analysis and the right tools, we can take that overabundance of data and then actually sort through it and learn what is inside. This helps us to come up with some clear decision points that we can trust and will make it easier overall to decide where to take our business in the future.

The good news here is that there are a few simple steps that we can work with in order to really complete our data analysis and to make this process as smooth as possible. Some of these steps include:

Define the Questions

The first thing that we need to work on when it comes to completing a data analysis is to define the questions that we would like to see answered. In a data analysis for a business or an organization, we need to begin with the right kinds of questions. This will help us to either search for the data that we want, or do a query of the data we already have to answer what we need to know. When picking out a question that we would like to use, make sure that we are working with one that is measurable, clear, and concise. Design the questions so that they will either qualify or disqualify, a potential solution to the problem that you want to work with.

For example, we want to make sure that we start out the whole process with a problem that is clearly defined. For our example, we are going to look at a

problem that is clearly defined. For our example, we are going to look at a government contractor who is dealing with rising costs. Due to these costs, they are struggling with submitting competitive contract proposals for the jobs.

One of the many questions that the contractor could look at to solve this problem could include whether or not they can reduce their staff without compromising the quality.

Set Priorities that are clear

The next step is to make sure that the measurement priorities that you are relying on are going to be as clear and concise as possible. There are two ways that we are able to do this. The first step is to decide what we would like to measure. And then the second step is to decide how we need to measure that item in the first place.
So, let's start out with the idea of what we would like to measure. Going back to the contractor from before, we need to consider the type of data that is needed in order to answer the key question, which in this case is "can the company reduce its staff without compromising quality?" To figure out this one, we would require to familiarize the number and the cost of the staff that we are working with right now, and the ratio of time they employ on functions that are crucial to the business.

When we answer this question, it is likely that we would have to go through and answer a lot of sub-questions as well to get the true answer. This can help us figure out whether the staff is being utilized in the proper manner and how we could improve the work they are doing.

Finally, when it comes to the decision on what to measure, we also need to make sure that we are taking into account any of the objections, the ones that are reasonable at least, that the stakeholders might have.
For example, if you do end up reducing the staff, how would you respond if there was a sudden surge in the demand from customers?

From here, we need to be able to decide how we want to measure. Thinking about how to measure the data is going to be an important step as well, especially right before we go through and collect the data that we need. This is because the measuring process that we choose is either going to discredit or back

up the analysis that we are doing at a later time. The key questions that we may want to ask during this time to help us stay on track and get the results that we would like to include:

1. What is the time frame that we are looking at here?
2. What is the unit of measure that we want to work with?
3. What kinds of factors should we include in this as well?

Collecting Our Data

At this point, we need to take some time to go through and collect the data that can help to make the analysis work for us.

When we have the clearly defined question that we want to work with, and we have the priorities of measurement set up, it is now time to acquire the data.

As we collect and work to organize the data, we have to keep a few important things in mind while doing this, and these will include:

1. Before we try to collect any of the new data, we have to regulate what information could be accumulated from the existing sources or databases that we have on hand. These are often the easiest to find and collect, and can sometimes have enough information to help us get started. Working with these first can save time and money and still get the work all done for us.
2. Determine the best file storing and naming system that we want to use, before collecting the data, in order to help all of the people who are working on this analysis to collaborating with one another. This is a good process to work on right away because it can save some time and will prevent the team members from wasting time by collecting the same information more than once.
3. If you do end up needing to gather information through observation or interviews, then it is important to develop the interview template

that you want to use ahead of time. This can save time, helps everyone to stay on course, and will ensure that there is some level of consistency in the results that you are going to get out of this process.

4. We also need to make sure that we are able to keep all of the data that is collected organized in a log with collection dates. You can even go through all of this and add in the source notes that are needed as we go, including any of the data normalizations that we decided to form. This is going to be a great way for us to take our conclusions and validate them later on if this is needed.

Analyze the Data

Once we have been able to go through some of the steps above, it is time to get down to business and actually analyze some of the data that we are working with.

This can take some time to accomplish, but it is a great process that can help us to get started. After we have been able to collect the right data to help us answer our questions from the first step, it is time for us to go through a deeper data analysis so that we can actually see what predictions and insights are found in the data.

We want to make sure that we begin the process of manipulating the data in a number of distinct manners, such as laying it out and finding some of the correlations, or by creating a pivot table with the help of MS Excel.

A pivot table is a great option to go with because it can help us to sort and filter the data by different types of variables, and can even help when it is time to calculate out the standard deviation, minimum, maximum, and mean of the data that we are working with.

As we work through the process of manipulating our data, it may be possible that you already have the exact type of data that is needed. But often, you may need to go through and make some changes to this to help make it work the way that you want.

Often you will need to go through and revise the original question that you have as you learn more about the options, or you will need to go through and collect some more data to finish answering that question.

No matter which one is right for you, this initial analysis of variations, outliers, trends, and correlations are going to make it easier to focus the data analysis that you are doing on better answering your question and any of the objections that others could have against the work that you are doing.
As you are going through this step, you will find that adding in some of the Python libraries and the algorithms that come with this language can be extremely useful.

Knowing how to handle the information and creating some of your own models along the way will make it easier to see what insights and information are hidden in all of that data, and can ensure that you actually can get the answers that you want to many common business problems.

Keep in mind that there are a lot of models that you are able to use in a data analysis in order to figure out what information is there, and the one that you go through will depend on the type and amount of data available and the answers that you would like to get out of the process.
There are many options including neural networks, decision trees, clustering, linear classifications and regressions, and more that we can choose to work through our data and use the Python language as well

How to Interpret the Results

After you have spent some time working with Python and creating your own models to analyze your data and possibly conduct some more research, it is time to interpret the results that you get.
As you go through and interpret the analysis that you are doing, keep in mind that you have to really pay attention to the information and see what it is providing to you.

Keep in mind with this one though that no matter how much time you spend on this, or how much data you do end up collecting, there is always a chance that something could interfere with the results that you get.

As we are going through the analysis and working on interpreting the results that we get, there are a few things that we need to keep in mind to make it easier and to ensure that we come up with the right interpretation of our data.

The three key questions that you can ask yourself while interpreting the results of the data will include:

1. Is the data able to help us to answer the original question that we have and if so, how?

2. Does the data help us to fight off and defend against some of the objections that are out there about our plan, and if so, how?

3. Are there going to be any limitations that could be found with the conclusions you are using or any kind of angle that you haven't been able to consider yet?

If the interpretation that you have for the data is able to hold up under all of these questions and considerations, then it is likely that the conclusion that you are working with is a productive one and one that you are able to work with as you move forward through your work.

The only step that is remaining to go through is to use all of the results that come with your process of data analysis is to decide the best course of action that you can take from this point.

That is one of the beauties that come with data analysis and all of the work that we did through the rest of our prepared steps. We are using this data to help us make good decisions and predictions about where to take our business into the future.

When the process is done in the right manner, and we are able to explore it with the right model and algorithm overall, we can find that it is one of the best ways to beat out the competition in any manner that we can.

By taking the time to follow these five steps and complete the process of data analysis, the business is able to make some of the best business decisions for their needs.

And this is even better because all of the choices that we have are going to be backed up by data that has been collected and analyzed in the proper manner.

With practice, this process is going to get faster and more accurate. This means that we are able to make better and more informed decisions, ones that are going

to help our organization as effective as possible.

Chapter 6: Popular Python Libraries for Data Science

Now that we have had some time to look at the process of data science and data analysis, it is time to take a look at some of the best libraries that work with Python and can help us get our data science work done.
Python is one of the best coding languages out there and can really help us to make use of machine learning and other algorithms that finish up our analysis. And when we combine it with some of the best data science libraries that are out there, we will find that there is so much that we can do with the data we collect. With that in mind, it is time to take a look at some of the best Python data science and data analysis libraries that can help us get our work done.

NumPy

The first library that we are going to take a look at is known as NumPy. This is going to be one of the principal packages that come with the ability to work on scientific applications. It is going to be used to help out when we need to process large multidimensional arrays and matrices, and there is also an extensive collection of high-level mathematical functions and implemented methods that can make it possible to perform a large variety of operations out of the objects that come with.

There are a lot of improvements that have been seen in this kind of library. In addition to the bugs that have been fixed and some of the compatibility issues, the crucial changes are going to regard the possibilities that come with this library, such as the format of printing in this library. There are also a few functions now that are able to handle files of any encoding that is available in Python.

SciPy

Another one of the core libraries that are used in scientific computing with Python is known as SciPy. This one is going to be based on a lot of the parts of NumPy so it is often best to download both of these at the same time to utilize them well. SciPy is basically going to be a library that is able to extend the capabilities of NumPy and all that it can do.

The main structure that comes with SciPy is going to be a multidimensional

The main structure that comes with SciPy is going to be a multidimensional array, implemented by NumPy of course. The package is going to contain tools that can help us out when it was time to solve linear algebra, probability theory, and integral calculus to name a few of the complex tasks that are possible with this kind of library.

Recently, SciPy has been going through a lot of major build improvements and these come mostly in the form of continuous integration into some of the different operating systems available. It also comes with new functions and methods, and it is going to have an updated optimizer. Also, there were a lot of functions, including ones known as LAPACK and BLAS that were wrapped in the process.

Pandas

If you would like to do some of the more complex tasks that are found in data science and machine learning with Python, then the Pandas library is definitely a choice that you need to make. This library is able to handle all of the different parts of data science, and it is going to include algorithms to get it all done for you. There are some drawbacks that can come with it sometimes, but overall, Pandas is the one that you need.

Pandas are going to be the library that comes with Python that can provide us with some higher-level data structures, and a lot of tools that help out with the analysis. The great feature that comes with this particular package is that it has the ability to translate some of the complex operations with data into just a few commands instead.

Pandas can contain a lot of built-in methods for grouping, filtering, and combining data, as well as some of the time-series functionality. All of this is able to be done with very fast speed indicators that can impress anyone who is trying to get ahead with this.

There have actually been a few releases that are rather new with the Pandas library, and there are a ton of new features that come with it, along with bug fixes, enhancements, and API changes. The improvements are going to help out the ability that pandas has for grouping and sorting data, providing us with a more suitable output for the apply method, and some support in working on the custom operations that are needed.

Matplotlib

Matplotlib

The next library that we need to spend some time looking at is known as Matplotlib. This is a great library to bring out when it is time to work with visualizing the data that we have and making more sense out of the other steps we have spent time on. Often visualization is one of the best things that you can work on when it comes to your work, because it helps us to make more sense out of the various complex relationships that are found in our data, so taking some time to use it and add it to your system, and getting it set up to work with Python, can be a great option.

Matplotlib is going to be considered a low-level library that can work on 2D graphs and diagrams of your data. With the help of this library, you are able to build up a lot of diverse charts, from histograms and scatterplots and so much more. In addition, many of the other popular plotting libraries that are out there and can work with data science and data analysis are going to work along with this library, so it is one that we should spend some time with.

Of course, as this library has been around for some time, we can see that there are a lot of style changes when it comes to the colors, sizes, fonts, legends, and more that are available with this library and the visuals that it can provide.

And the types of visuals that are available with this kind of library are impressive and growing all of the time.

Scikit-Learn

If you would like to add in a bit of machine learning to the mix, then the Scikit-Learn library is one of the best ones for you to consider. Machine learning is an important part of the data analysis process because it can be combined together with the Python language to form the necessary models and algorithms to analyze your data. There are a lot of times when we need to bring some machine learning into the mix, and this library is a great option to help with it.

This is a Python library that is going to be based on the NumPy and SciPy libraries that we talked about before, so it is worth our time to download both of those and make sure they are ready to go if we want to implement some of the machine learning in this library. It is also one of the best libraries from Python that is able to work with data and get things organized and analyzed the way that we would like.

we would like.

The Scikit-Learn library is going to be able to handle a lot of the different things that you will want to do when analyzing your data. It is going to provide us with a lot of the algorithms that are needed for many tasks of data mining and machine learning such as model selection, dimensionality reduction, classification, regression, and clustering.
There have been a lot of enhancements over the years to this library to make it better at the job that it does.

The cross-validation has been modified quite a bit, which is going to provide us with the ability to work with more than one metric at a time.

There are also a few other training methods, including the logistic regression and nearest neighbors, that have seen some great improvements over time as well.

TensorFlow

It is possible that you will want to spend some time working with deep learning and other similar methods when it comes to data analysis, and if this is your goal, then adding in the TensorFlow library is the best option to help you get this done.

TensorFlow is going to be one of the popular frameworks that you can deal with when it comes to machine learning and deep learning. It was also a library that the Google Brain Team created, so we know there are a lot of features and more that come with this library and that we can make use of as well.

TensorFlow is going to provide us with a lot of abilities to work with artificial neural networks when we have more than one data set. Among the most popular, TensorFlow applications are going to include things like speech recognition and object identification to name a few. There are going to be a few different layer helpers that work on top of regular TensorFlow that can make it easier to work with.

This library is going to be quick in some of the new releases, introducing a lot of new features along the way.

Among some of the latest is going to be fixed in some of the vulnerabilities for the security of this library, and more improved integration between GPU and TensorFlow to make things easier.

PyTorch

The next library that we can talk about when it comes to working with data science. PyTorch is going to be a large framework that will allow you to get some of the computations of the tensor to perform with the GPU acceleration, create some computational graphs that are dynamic, and automatically calculate the gradients. Above this, PyTorch is going to offer us a really rich API for solving applications that are related back to neural networks.

The library is going to be based on Torch, which is going to be a deep learning library that will usually be implemented in the C language, and then it wraps around with Lua The Python API was introduced back in 2017 and from that point on, the framework is gaining a lot of popularity and attracting an increasing amount of data scientists.

Keras

The final Python library that works well for data science is going to be the Keras library. This one is going to be another high-level library that can help us work with neural networks, and will run on top of some of the other libraries that are out there including Theano, TensorFlow and a few other new releases. It is also possible to work the MxNet and CNTK as the backend that will be important here. It is going to be useful when you want to handle a lot of specific tasks and can reduce the amount of monotonous code that you will need to write. However, some of the more complicated types of coding that you will write out will not work as well with this option.

Working with the Python library can be a great way to finish up a bunch of the tasks that you would like, and will ensure the algorithm that you work within the data science process will be done the way that you would like.
Take a look at a few of these libraries and see which one is likely to provide you with some of the results that you need to get the job done

Chapter 7: The Importance of Data Visualization

The next topic that we need to spend some time looking through is the idea of data visualization.

This is a unique part of our data science journey, and it is so important that we spend some time and effort looking through it and understanding what this process is all about. Data visualization is so important when it comes to our data analysis. It can take all of the complex relationships that we have been focusing on in our analysis and puts them in a graph format, or at least in another visual format that is easier to look through.

Sometimes, looking at all of those numbers and figures can be boring and really hard to concentrate on. It can take a long time for us to figure out what relationships are present, and which ones are something that we should ignore. But when we put the information into some kind of graph form, such as a graph, a chart, or something similar, then we will be able to easily see some of the complex relationships that show up, and the information will make more sense overall.

Many of those who are in charge of making decisions based on that data and on the analysis that you have worked on will appreciate having a graph or another tool in place to help them out.

Having the rest of the information in place as well can make a difference and can back up what you are saying, but having that graph in place is one of the best ways to ensure that they are able to understand the data and the insights that you have found.

To make it simple, data visualization is going to be the presentation of data that shows up in a graphical or a pictorial format of some kind or another. It is going to enable those who make the big decisions for a business to see the analytics presented in a more visual manner so that they can really grasp some of the more difficult concepts or find some new patterns out of the data that they would never have known in any other manner.

There are a lot of different options that you are able to work with when it comes to data visualization, and having it organized and ready to go the way that you like, using the right tools along the way, can make life easier. With an interactive type of visual, for example, you will be able to take this concept a bit further and

use technology to drill down the information, while interactively changing what data is shown and how it is processed for you.

A Look at the History

As we can imagine, the process of visualization, and using pictures to help us understand the data in front of us is something that has been around for a long time. Whether we look at the pictures that show up in our books or even maps and graphs that were found in the 17th century and before, we have been using images and more to help us make sense of the world around us and all of the data that we have to sort through can really be understood with some of these visuals as well.

However, it is really a big boost in technology that has helped to make data visualization something that is as popular as it is today. For example, computers are really making it possible for us to process a large amount of data, and we are able to do this at faster speeds than ever before. Today, the data visualization and all that comes with it is an industry or a field that is rapidly evolving. Add to it that this is now something that needs a nice blend of science and art and that it can go a long way in helping us to work with our own data analysis, and it is no wonder that these visuals are as popular as they seem.

Why Is Data Visualization So Important?

The next thing that we need to take a look at here is why data visualization is so important to us. The reason that data visualization is something that we want to spend our time and energy on is because of the way that someone is able to process information. It is hard to gather all of the important insights and more on a process when we have to just read it off a table or a piece of paper. Sure the information is all right there, but sometimes it is still hard to form the conclusions and actually see what we are doing when it is just in text format for us.

For most people, being able to look at a chart or a graph or some other kind of visual can make things a little bit easier.

Because of the way that our brains work and process the information that we see, using graphs and charts to visualize a large amount of complex data is going to be so much easier compared to pouring over some reports or spreadsheets.

When we work with data visualization, we will find that it is a quick and easy way to convey a lot of hard and challenging concepts, usually in a manner that is more universal. And we are able to experiment with the different scenarios by using an interactive visual that can make some slight adjustments when we need it the most.

This is just the beginning of what data visualization is able to do for us though, and it is likely that we will find more and more uses for this as time goes on. Some of the other ways that data visualization will be able to help us out will include:

1. Identify areas that will need the most attention when it comes to improvements and attention.

2. Help us to figure out which of our products we should place where.

3. It can clarify which factors are the most likely to influence the behavior of a customer.

4. It can make it easier to tell and make predictions about our sales volumes, whether these volumes are going to be higher or lower at a specific time period.

The process of data visualization is going to help us change up the way that we can work with the data that we are using. Data analysis is supposed to respond to any issues that are found in the company in a faster manner than ever before. And they need to be able to dig through and find more insights as well, look at data in a different manner, and learn how to be more imaginative and creative in the process. This is exactly something that data visualization is able to help us out with.

How Can We Use Data Visualization?

The next thing that we need to take some time to look at is how companies throughout many industries are able to use data visualization for their own needs. No matter the size of the company or what kind of industry they are in, it is possible to use some of the basics of data visualization in order to help make more sense of the data at hand. And there are a variety of ways that this data

visualization will be able to help you succeed

The first benefit that we can look at is the fact that these visuals are going to be a great way for us to comprehend the information that we see in a faster fashion. When we are able to use a graphical representation of all that data on our business, rather than reading through charts and spreadsheets, we will be able to see these large amounts of data in a clear and cohesive manner.

It is much easier to go through all of that information and see what is found inside, rather than having to try and guess and draw the conclusions on our own.

And since it is often much faster for us to analyze this kind of information in a graphical format, rather than analyzing it on a spreadsheet, it becomes easier for us to understand what is there. When we are able to do it in this manner, it is so much easier for a business to address problems or answer some of their big questions in a timely manner so that things are fixed without issue or without having to worry about more damage.

The second benefit that comes with using data visuals to help out with the process of data science is that they can really make it easy to pinpoint some of the emerging trends that we need to focus on. This information is within the data, and we are going to be able to find them even if we just read through the spreadsheets and the documents.

But this takes a lot of time, can be boring, and often it is hard for us to really see these correlations and relationships, and we may miss out on some of the more important information that we need.

Using the idea of these visuals to handle the data, and to discover trends, whether this is the trends just in the individual business or in the market as a whole, can really help to ensure that your business gains some big advantages over others in your competition base. And of course, any time that you are able to beat out the competition, it is going to positively affect your bottom line. When you use the right visual to help you get the work done, it is much easier to spot some of the outliers that are present, the ones that are more likely to affect the quality of your product, the customer churn, or other factors that will change your business. In addition, it is going to help you to address issues before they are able to turn into much bigger problems that you have to work with.

Next on the list is that these visuals are going to be able to help you identify

some relationships and patterns that are found in all of that data that you are using. Even with extensive amounts of data that is complicated, we can find that the information starts to make more sense when it is presented in a graphic format, rather than in just a spreadsheet or another format.

With the visuals, it becomes so much easier for a business to recognize some of the different parameters that are there and how these are highly correlated with one another. Some of the correlations that we are able to see within our data are going to be pretty obvious, but there are others that won't be as obvious. When we use these visuals to help us find and know about these relationships, it is going to make it much easier for our business to really focus on the areas that are the most likely to influence some of our most important goals.

We may also find that working with these visuals can help us to find some of the outliers in the information that is there. Sometimes these outliers mean nothing. If you are looking at the charts and graphs and find just a few random outliers that don't seem to connect with each other, it is best to cut these out of the system and not worry about them.

But there are times when these outliers are going to be important and we should pay more attention to them.
If you are looking at some of the visuals that you have and you notice that there are a substantial amount of them that fall in the same area, then you will need to pay closer attention. This could be an area that you can focus your attention on to reach more customers, a problem that could grow into a major challenge if you are not careful, or something else that you need to pay some attention to.

These visuals can also help us to learn more about our customers. We can use them to figure out where our customers are, what kinds of products our customers would be the happiest with, how we can provide better services to our customers, and more. Many companies decide to work with data visualization to help them learn more about their customers and to ensure that they can really stand out from the crowd with the work they do.

And finally, we need to take a look at how these visuals are a great way to communicate a story to someone else. Once your business has had the time to uncover some new insights from visual analytics, the next step here is to communicate some of those insights to others. It isn't going to do you much good to come up with all of those insights, and then not actually show them to the people responsible for key decisions in the business.

Now, we could just hand these individuals, the ones who make some of the big decisions, the spreadsheets and some of the reports that we have. And they will probably be able to learn a lot of information from that. But this is not always the best way to do things.

Instead, we need to make sure that we set things up with the help of a visual, ensuring that these individuals who make the big decisions can look it over and see some of the key relationships and information at a glance.

Using graphs, charts, and some of the other visuals that are really impactful as a representation of our data is going to be so important in this step because it is going to be engaging and can help us to get our message across to others in a faster manner than before.

As we can see, there are a lot of benefits that come in when we talk about data visualizations and all of the things that we are able to do with them. Being able to figure out the best kind of visualization that works for your needs, and ensuring that you can actually turn that data into a graph or chart or another visualization is going to be so important when it is time to work with your data analysis.

We can certainly do the analysis without data visualization. But when it comes to showcasing the findings in an attractive and easy to understand format, nothing is going to be better than data visualization.

How to Lay the Groundwork

Before we try to implement in a brand new technology of any sort, there are going to be a few types of steps that we need to take and go through to see the results.
Not only is it important for us to have a nice solid grasp on the data that we want to use which is something that should happen during the data analysis part of the process, we also need to understand three other important things including the needs of the company, the goals of the company, and the audience of your company.

Some of the things that have to happen before you can prepare and organize all of the data that you have and complete this kind of data visualization will include:

1. Understand the data that we need to visualize in the first place. This means that we need to know how much cardinality is present in the data, meaning how much uniqueness is going to show up in the columns, and we need to know the size of the data. Some of the algorithms that we will use to work on the data visualization are not going to do as well when it comes to very large sets of data.

2. Determine what you would like to visualize, and the kind of information that we want to be able to communicate with this. This will make it a bit easier to figure out which type of visual we want to be able to go within this process.

3. Know the audience that we are working with and understand how they are going to process the visual information that you want to show off. Management may need a different visual than a team. Those in manufacturing may need a different visual than someone in a more creative role. Being able to make the visual fit to the audience so that they can actually utilize the information is going to be a critical step.

4. Use a visual that is able to convey the information in the form that is not only the best but also the simplest, for your audience. There are a lot of cool visuals out there that you can work with, and they can offer a lot of different ways to show your data. But if the visual is too difficult to understand, it is not going to make anyone happy. Put away some of the neat gadgets and find the best way to showcase that information that makes the most sense to your audience.

Once you have been able to go through and answer all of the initial questions that we had about the data type that we would like to work with, and you know what kind of audience is going to be there to consume the information, it is time for us to make some preparations for the amount of data that we plan to work within this process

Keep in mind here that big data is great for many businesses and is often necessary to make data science work. But it is also going to bring in a few new challenges to the visualization that we are doing. Large volumes, varying

velocities, and different varieties are all going to be taken into account with this one.

Plus, data is often going to be generated at a rate that is much faster than it can be managed and analyzed so we have to figure out the best way to deal with this problem.

There are factors that we need to consider in this process as well, including the cardinality of the columns that we want to be able to work with.

We have to be aware of whether there is a high level of cardinality in the process or a low level. If we are dealing with high cardinality, this is a sign that we are going to have a lot of unique values in our data. A good example of this would include bank account numbers since each individual would have a unique account number.

Then it is possible that your data is going to have a low cardinality. This means that the column of data that you are working with will come with a large percentage of repeat values. This is something that we may notice when it comes to the gender column on our system. The algorithm is going to handle the amount of cardinality, whether it is high or low, in a different manner, so we always have to take this into some consideration when we do our work.

Different Types of Data Visualization Available

As you go through and start to work on adding some visualizations to your own project, you will quickly notice that there are a lot of choices that you are able to make. And all of them can work well depending on the kind of data that you are working with, and the way that you would like to present it.
Sometimes the question that you are asking out of the data will help to determine which type of visualization is going to be the best for your own needs.

There are options like bar graphs, line graphs, histograms, pie charts, and more that can all show the information if you are trying to separate information into groups and see where your customers lie or which decision is the best for you, something like a scatterplot could be the best option to work with.
There are a lot of options when it comes to working with visuals, and we have to just figure out which one is the best for our needs.

When you are first exploring some of the new data that you have collected, for example, you may find that something like an auto-chart is the best option for

example, you may find that something like an auto-chart is the best option for your needs. This is because they can give you kind of a quick view into a large amount of data, in a way that other options just are not able to do. It may not be the final step that you take, but it is going to make a difference in how well you are able to understand the data in the beginning, and can lead you on the right path to picking out models and algorithms to work with later on.

This kind of data exploration capability is going to be helpful, even to those who are more experienced in machine learning, data science, and statistics as they seek to speed up the process of analytics because it is going to eliminate some of the repeated samplings that has to happen for each of the models that you are working on overall.

None of the visual options are necessarily going to be bad ones, we just have to learn which one is the best option for the data we have, and for the uses that we want to do with the data. Each set of data is going to lend itself well to one type of visual or another, and having a good understanding of what you are expecting out of this data, and what your data contains in the end, can help us to figure out which visual we are most interested in.

Data visualization is definitely a part of data science that we do not want to forget about.
Being able to make this work for our needs, and understanding some of the process that comes with it, as well as why we actually need to work with a visual overall, can be important. Make sure to figure out which visual is going to be the best for your needs to ensure that you will get the best way to understand the complex relationships in your data in no time.

Chapter 8: The Data Science Pipeline

Now that we have had some time to look through the different parts that come with data science, and all that it entails, it is time for us to dive a bit more into the data science pipeline.

There are a lot of different parts that come into play when we are working with data science, and making sure that they all line up with one another well, and that they are able to complete some of the tasks in the way that we would like can be a challenge. But when it is all said and done, it is going to be one of the best ways for us to see the results that you want.

In this chapter, we are going to explore a bit more about the data science pipeline and what all comes with this overall. We will look at how to work with speeding up the machine learning algorithms that we want to use, how to use different parts to create the visualizations that we would like, and so much more.

So, let's dive in and see what all we can learn when it comes to working with data science and more!

Binary classification

The first thing that we need to take a look at is known as binary classification. This is going to be the task of classifying the elements of a given set that you have into two groups (predicting which group each one is going to belong to), on the basis of the rule of classification. Contexts that require a decision as to whether or not an item has some kind of property that is qualitative, some specified characteristics, or a typical binary classification can include some of the following examples:

1. Medical testing: This will need a binary classification to help us determine whether a patient has a disease or not. The property of classification that we are going to focus on is the presence of the disease.

2. A pass or fail test method or quality control in the manufacturing factory could include the same idea as deciding if a specification has or hasn't been met at all. This kind of classification is going to be

called the go/no go classification.

3. Information retrieval: This one is going to be mostly deciding whether a page or an article should be inside the result set of a search or not. In this case, the classification property is going to be how relevant the article is or how useful the user is going to find that article.

Binary classification is going to be the dichotomization applied to practical purposes, and in many of these problems, the two groups that you have are not going to turn out to be symmetric.

Rather than overall accuracy, the relative proportion of different types of errors will be something that interests us.

For example, in medical testing, a false positive, which means that we detect a disease when it is not actually present, is going to be considered different from what is called a false negative, or not detecting a disease when it is actually present in the patient.

We also need to take some time to explore the statistical binary classifications and how these affect the work that we are doing. This is going to be one of the problems that can be studied through machine learning, and in specific, it is a type of supervised machine learning. This means that it is able to learn based on lots of examples, with the corresponding output coming with the input as it learns.

This method of machine learning will have all of the categories predefined and will be used when we want to work with categorizing new probabilities observations into our categories.

When we are only working with two known categories for our problem, then we are going to work with what is known as a statistical binary classification. Some of the methods that are often used for this kind of binary classification are going to include the following:

1. Probit model

2. Logistic regression

3. Neural networks
4. Support vector machines
5. Bayesian networks
6. Random forests
7. Decision trees

Each of the classifiers is going to be best in just their own select domain, and this is going to be based on the number of observations that we are able to find, the dimensionality that comes with the feature vector, the amount of noise that is in the data, and a variety of other factors. For example, random forests are going to perform better than the SVM classifiers when we work with 3D point clouds.

There are going to be a lot of different metrics that we may use in order to measure how well our classifier or predictor is going to perform. There are also different fields that will end up with different preferences for the specific metrics due to the different goals that we have.

For example, in the field of medicine sensitivity and specificity are the parts that are often used, but when we are working with something like information retrieval, we are going to rely on recall and precision instead.

Something that is important to keep apart in the metrics is that they are going to be independent on the prevalence, or how often each of the categories is going to occur in our population, and the metrics that depend on the prevalence.

Both types are going to be useful, but they are going to come in with very different properties in the process.

Given a classification on a specific set of data that we are going to work with, there can be four combinations of the actual data category and the assigned category there can be true positives, true negatives, false positives, and false negatives.

We are able to put these into a contingency table that goes 2 by 2 with the columns corresponding back to the actual value and the rows will correspond back with the classification value.

In addition to what we did above, there are going to be a total of eight basic ratios that one is able to compute from this specific table, which is going to come to us in complementary pairs of four, each one summing to 1. These are something that you are able to obtain by dividing each of the four numbers by the sum of its row or column, which is going to provide you with eight numbers.

We are able to refer to these in the generic by calling them the true positive row ratio or the or the false negative column ratio, though there are some terms that are considered more conventional for how we would like to deal with these

As we work through this, we may find that there are going to be two pairs of column ratios and two pairs of row ratios and one is able to summarize these with four numbers because it is easy to do so by choosing one ratio out of each pair.
The other four numbers that we are looking at can be known as the complements in this process.

When we are working with something like diagnostic testing, the main ratios are going to be the ones that are used for the true column ratios, which means the True Negative Rate and the True Positive Rate, when they are known as the parts for sensitivity and specificity.

In the retrieval of information, the main ratios that we need to focus on are going to be the true positive ratios, both the rows and the column, where they are known more as the precision and the recall that we talked about before.

One is able to take the ratios of a complementary pair of ratios, which is going to give them four of the likelihood ratios we are going to get from this. These are going to include the two-column ratio of ratios. This is something that is done primarily for the column, or the condition, ratios, yielding for us the likelihood ratios in diagnostic testing.

Of course, there are some other metrics that we are able to work with as well. The most simple of this is going to be the accuracy or the Fraction Correct. This one is nice because it can measure out the fraction of all instances that are correctly categorized: the complement is going to be the Fraction Incorrect.

The F-score is able to combine the precision and the recall into one number through the choice of weight, most simply that it equals weighing as a way of balancing out the F-score This, of course, is just one of the many options that you are able to work with when it comes to working through some of the things

you are able to work with when it comes to working through some of the things that you would like to do with binary classification.

PCA

We also need to spend some time looking at the beauty of PCA, or Principal Component Analysis, and how this is going to work for your needs. This one is one of the best ways that you can speed up some of the analysis because it works to fit the machine learning algorithm to the data, mainly because it can change up the optimization that comes with this algorithm. Logistic regression is one of the methods that you are able to use for this, but PCA is going to be more common, and more efficient at the work that it can do.

If you are working on your data analysis and you have chosen a specific algorithm from Python in order to handle the data, but you find that this chosen algorithm is way too slow, usually because the input dimension that you work with is going to be too high for that algorithm to handle, then you will find that a reasonable choice to fixing this issue will be using PCA to speed it all up.

Now, the option of speeding up your algorithm is just one of the choices that you can do with PCA, and it is actually really helpful at working on another process. If you are using PCA, it is likely that it is to help out with speeding up some of the computational power of your algorithms but it is also a great option to work with when it comes to data visualization.

In the next section we will take a look at how the PCA is going to work well for helping out with the visualizations that you would like to focus on, but we are first going to just take a look at some of the basic steps that come with PCA and how it can work to take any of the algorithms that you are doing in Python and speed them up:

1. The first thing to do here is to normalize some of the original features that come with your algorithm. This means that we want to make sure they mean it is removed from each feature.

2. Now we can work on the covariance matrix of our normalized data. This is going to be an asymmetric matrix that is n by n, where n is going to be the number of original features, and then the element in row I and column j is going to be the covariance that shows up

between these columns in the set of data.

3. Next, we can focus on calculating the eigenvectors and the eigenvalues of our matrix. These eigenvectors need to be unit eigenvectors. This means that their lengths must end up being one. This step is going to be one of the most intricate, and it is likely that your chosen software package is going to do this automatically.

4. Choose the k eigenvectors with the highest eigenvalues.

5. Now it is time for us to compute the final k features, and this is going to be the one that is associated with the k highest eigenvalues. For each one of these, we need to multiply the set of a data matrix, by the eigenvector that is associated back with it. Here we are going to make the assumption that the eigenvector is just one column and n rows (n is going to be the number of the original variables), and then the data set matrix has n columns and m rows (m is going to be how many observations that we have) Thus, the resulting final features have m rows and one column and it is going to provide us with the values of the new features, which are computed at each of the observations.

6. You can then go back through here and add back in the mean that we had removed in that first step before.

7. Now we need to look at something else. Here we look at the proportion of the variance that each eigenvector represents that can be calculated by dividing the eigenvalue corresponding to that eigenvector by the sum of all the eigenvalues.

Now, there are going to be a few things that we need to take a look at here before we jump right in. first, the original features may turn out to be highly correlated. If this happens, then whatever solution you are able to get you to get is not going to be that stable. Also, the new features are going to be linear combinations of some of the original features and this means that they can lack interpretation.

The data doesn’t need to be multinormal, except if you are using this technique

to help you with predictive modeling with the help of normal models to compute the confidence intervals.

PCA for Data Visualization

Now that we have taken a bit of time to look at the PCA and how we can handle just the regular algorithm with this one, it is time to take a look at how PCA can be used for some of the data visualizations that you want to focus on. For a lot of the applications that come with machine learning, it is going to be really helpful to visualize the data you are working with. Going through and visualizing 2 or 3D data is not all that challenging.
However, even the Iris data set is going to be used in this part of the tutorial is 4 dimensional. We are able to use the idea of the PCA in order to reduce that 4-dimensional data into 2 or 3 dimensions so that you can plot and hopefully gain a better understanding of the data that you are presented with.

The first step that we need to work with is loading up our Iris dataset. This is going to be one of the options that are available through Scikit-Learn and while it can help out with some of the visualizations that we need to work with, it is not going to require us to go through and download any of the files that come from external websites, making it safer to work with overall. The code that you need in order to make this dataset load on your computer includes the following:

```
import pandas as pd
URL = " https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
"

# load dataset into Pandas DataFrame
df = pd.read_csv(url, names=['sepal length','sepal width','petal length','petal
width','target'])
```

The next step that we need to work with here is standardizing the data that we have. PCA is going to be affected a lot by the scale of the data or the project that you are working with, so we need to make sure that we have taken the time to calibrate the features in your data before you try to use this algorithm at all. Use the StandardScaler to help us get this done to help standardize the set of data features into a unit scale.
This means that we want to make a scale where the mean is 0 and the variance is 1. This is going to be one of the requirements that come with the optimal performance of many of the machine learning algorithms that you want to work

with.
If you would like to see some of the negative effects that can happen on all of this if you do not scale the data that you have, then the Scikit-Learn library is going to have a section present that can show you some of the different things that happen when you are not able to standardize the data before you start.

After this is done, it is time to move onto the PCA projection to 2D. The original data that we are working with is going to come with four columns.

These are going to include the sepal length, the sepal width, the petal length, and the petal width. In this part of the code, we are going to see that there is a projection of the authentic data, which is 4 dimensional into 2 dimensions.

One thing to note with this one is that after the reduction in dimensionality, there usually is not going to be a particular meaning that is assigned to each of your principal components. The new components that you are able to get out of all of this will just be the two main dimensions of the variation.

From there, we are going to be able to work on being able to visualize our 2D projection. This section will be about plotting something out on a graph so we can see the classes. We are going to work with the code below to help us get started, and we should end up with a graph that will separate the different classes from one another. This is a great way to ensure that we can see what is going on in the process, and we can best figure out what steps we want to take next. The code that we need to use to make this happen includes:

```
fig = plt.figure(fig size = (8,8))
ax = fig.add_subplot(1,1,1)
ax.set_xlabel('Principal Component 1', fontsize = 15)
ax.set_ylabel('Principal Component 2', fontsize = 15)
ax.set_title('2 component PCA', fontsize = 20)targets = ['Iris-setosa', 'Iris-
versicolor', 'Iris-virginica']
colors = ['r', 'g', 'b']
for target, color in zip(targets,colors):
   indicesToKeep = finalDf['target'] == target
   ax.scatter(finalDf.loc[indicesToKeep, 'principal component 1']
            , finalDf.loc[indicesToKeep, 'principal component 2']
            , c = color
            , s = 50)
ax.legend(targets)
```

ax.grid()

There is one final thing that we need to take a look at in this section and this is the explained variance. This is the part that will tell us how much information, or variance, we can attribute to each of the principal components. This is going to be important because while we are able to convert a 4D into a 2D space, we are still going to lose a bit of our variance or the information when we decide to do this process.

However, when we choose to use the attributes of explained_variance_ratio_, we are going to see that the first principal component that we are dealing with is going to contain 72.77 percent of the variance that is there, and then the second principal component is going to contain 23.03 percent of the variance. This means that when we combine together the two components that we just talked about above, we can find that they will contain about 95.80 percent of the information that we need.

PCA to Speed-Up Machine Learning Algorithms

One of the most important ways that we are able to work with PCA is to help us speed up the algorithm that we are working within machine learning. Using the IRIS dataset would be illogical in the example that we are doing here because there are only 150 rows and four feature columns. This is a pretty small data set and is not going to make that much of a difference overall. Any of the Python algorithms that you are using in machine learning are going to be able to handle this, so it doesn't make much sense to try and speed it up.

However, we will find that the MNIST database is going to be a much better option. This database is going to come with handwritten digits that are going to work better for this because it comes with 784 feature columns and a training set that has more than 60000 examples.

And the test set adds more to this with a total of 10,000 examples. This is something where the algorithm could use a speed up, and that is what we are going to focus on here.

So, the first step to help speed up our own machine learning algorithms with PCA will be to download the database and then load up the data that you want to work with. We are also able to add in the data_home parameter to the fetch_mldata to change where we are downloading this data from. The code that we need to use to make this happen includes:

we need to use to make this happen includes:

from sklearn.datasets import fetch_mldata
mnist = fetch_mldata('MNIST original')

The images that you are going to download with this code are going to be found in the mnist.data and they will have a shape that says (70000, 784). This means that there are going to be 70,000 images that have 784 features or dimensions. The labels, which are going to be the integers of 0 to 9, are going to be contained in the part that says mnist.target. The features are going to be 784 dimensional (and they are 28 by 28 images), and the labels are going to just be numbers from 0 to 9.
Next on the list is to split the data up so that we can have one set for our training purposes, and one for our test sets. Frequently, the train test divvying is going to include about 80 percent for training and then another 20 percent that is used for the testing.

In this case, we are going to work with about $6/7^{th}$ of the data to be part of the training set, and $1/7^{th}$ of the data to be the test set. The code that we are able to use to make this happen is below:

from sklearn.model_selection import train_test_split

 # test_size: what proportion of original data is used for test set
train_img, test_img, train_lbl, test_lbl = train_test_split(mnist.data, mnist.target, test_size=1/7.0, random_state=0)

The final step that we need to do before it is time to bring in and apply the PCA is to standardize the data that we are working with. The text in this paragraph is going to be pretty much the same as what we talked about in the last section. The PCA is going to find that it changes a lot based on the scale that we are working with, so we need to make sure that we go through and scale the features in the data before we try to apply the PCA. We also want to work with the mean being 0 and the variance being 1.

This is going to be a big requirement when it comes to the ideal effectiveness of many machine-learning algorithms that we want to focus on.

Take your time making the data to scale and standardizing some of the data as well. We want to ensure that the data is going to work well for the process that

well. We want to ensure that the data is going to work well for the process that we are doing, and that we are not missing out on some of the information that we want in the process, or that things are not going to be missed in this process either.

Next on the list is for us to go through, import, and apply the PCA. Notice that the code we are going to have in a moment has .95 for the number of components that will be in the parameter. This means that the library of Scikit-Learn will choose the minimum number of principal components such that 95 percent of the variance will be retained. We will also look at how we can fit the PCA on the training set in our code below:

```
from sklearn.decomposition import PCA
# Make an instance of the Model
pca = PCA(.95)

pca.fit(train_img)
```

We can finish this to work with applying the mapping or transform, to both the training set and the test set. We then can work with adding in the logistic regression to the data that we have transformed in order to make sure that we can get the algorithm to work the way that we want, and then add in the PCA. The steps that we are able to work with here are going to include:

1. Import the model that you are planning to use here. In Scikit-Learn, all of the machine learning models that you want to use will be implemented as a type of Python class.

2. Make an instance of the Model.

3. Prepare the model on the data that you have. Make sure that the model is storing and remembering all of the information that it is storing from the data. The model is learning the relationship between the digits and the labels that are present.

4. From here, we need to go through and predict the labels of our new data, which will be the new images in these cases. We can use the information that the model was able to learn during our training

process.

5. Measure the performance of the model. While efficacy is not consistently going to be the best metric for us to use when it comes to machine learning algorithms because options like ROC curve and precision would often be better, it is going to be used here to keep things simple. Make sure to measure out the performance of the model to see how we are able to make this work for our needs.

From here, we need to do the timing fitting with the logistic relapse after PCA. The entire idea of this part of the academic is to help us see how we can use PCA in order to speed up how well these machine learning algorithms are able to fit into the model that you are creating. The table that we have below is going to show us how long it is going to take for us to fit the logistic regression onto our system with the help of PCA, keeping in mind that we are retaining different amounts of variance each time that we do this:

Variance Retained	Number of Components	Time (seconds)	Accuracy
1.00	784	48.94	0.9158
0.99	541	34.69	0.9169
0.95	330	13.89	0.9200
0.90	236	10.56	0.9168
0.85	184	8.85	0.9156

The covariance of the matrix

The next thing that we need to take a look at here is known as the covariance matrix. When we are looking at statistics and probability theory on our data, we will find that this matrix is going to be an element in the i, j position is the covariance between the i-th and the j-th element of a random vector.

A random vector is going to be one of the variables that we have that are random and have a bunch of different dimensions that we need to focus on. Another thing to consider with this is that each of the elements out of that vector is going to turn into a scalar of the random variable.

Each of the elements will either come with a finite number of empirical values that we are able to actually observer, or we can find that it comes with a limited or limitless number of values that could potentially meet up with it. The potential values of this process are going to be specified with a theoretical joint probability distribution.

When we look at this in an intuitive manner, the covariance matrix concludes the notion of variance to multiple dimensions. We can look at an example of this one to see how it will work for us. For example, the variation that can come in a collection of random points in a 2D space cannot be characterized, no matter how hard we try, but a single number.
Something would be missed out in the process. In addition, the variances in the x and y directions contain all of the information that is needed.

So, what this means for us in data science is that we have to make sure we have more than one number in place to describe our data. A 2 by 2 matric is actually necessary when it comes to fully characterize the two-dimensional variation that we are looking for.

Because of the covariance, that we see with the i-th random variable with itself is simply that the random variable's variance, each element on the principal diagonal of the covariance matrix is going to be the same variance as one of the random variables that we see with this.

Because of the covariance of the i-th random variable with the j-th one is the same thing as the covariance of the j-t random variable with the i-th random variable, every covariance matrix is going to be symmetric in some manner. Also, each covariance matrix is going to be semi-definite in the process.

This variance is going to be important when we work with our data pipeline because it ensures that we know how accurate the whole process is going to be. The greater the variance that shows up in the process, the less we can trust the model that we are working with, and the more that we need to take a look into it and see what is available and what changes we can make. The smaller the variance that shows up, the more we can rely on the results that we get out of that process, and the better it is for everyone who is relying on that model.

Understanding how data science works can be a tough thing because there are so many variables and different parts that need to come into play before it works the way that we would like.

Making these models work, no matter what kind of process we are working with or what kind of data we want to focus on can be important when it is time to work with these to make some informed data decisions overall.

Chapter 9: A Practical Example of Working with Python Data Science

Now that we have spent some time taking a look at data science and the Python language, it is time to work with an example of how we can bring this all together and work on our own project with data science.

In this one, we are going to work to anticipate the circumstance of diabetes in patients and then take the right measures ahead of time to help prevent this issue. In this use case, we are going to spend our time figuring out the occurrence of diabetes, with the help of an entire lifecycle that we talked about earlier in this guidebook. Some of the steps that we need to work with will be below:

First, we need to make sure that we can compile the data found on the medical history that comes with the patient.

This is going to be the research that will create and test out our model and can help us later if we need to submit some information on a new patient.

We are going to use the sample data that is available below to help us create this kind of model:

;npreg;glu;bp;skin;bmi;ped;age,income
1;6;148;72;35;33.6;0.627;50
2;1;85;66;29;26.6;0.351;31
3;1;89;80;23;28.1;0.167;21
4;3;78;50;32;31;0.248;26
5;2;197;70;45;30.5;0.158;53
6;5;166;72;19;25.8;0.587;51
7;0;118;84;47;45.8;0.551;31
8;1;103;30;38;43.3;0.183;33
9;3;126;88;41;39.3;0.704;27
10;9;119;80;35;29;0.263;29
11;1;97;66;15;23.2;0.487;22
12;5;109;75;26;36;0.546;60
13;3;88;58;11;24.8;0.267;22
14;10;122;78;31;27.6;0.512;45
15;4;97;60;33;24;0.966;33
16;9;102;76;37;32.9;0.665;46
17;2;90;68;42;38.2;0.503;27
18;4;111;72;47;37.1;1.39;56
19;3;180;64;25;34;0.271;26
20;7;106;92;18;39;0.235;48
21;9;171;110;24;45.4;0.721;54

There are going to be a lot of attributes that we need to work within the data set above. These can include:

1. Npreg: This is the number of times the patient has been pregnant.

2. Glucose: This is going to include the plasma glucose concentration.

3. Bp: This is the blood pressure of the patient.

4. Skin: This is going to be the tricep skinfold thickness.

5. Bmi: This is going to include the body mass index of the patient.
6. Ped: This is going to be the diabetes pedigree function of the patient.
7. Age: This is going to include the age of the patient.
8. Income: This takes a look at the income of the patient.

From here, we are able to move onto the second step, or the second phase, of the data lifecycle that we talked about before. Now that we have access to all of that data, it is time to clean and prepare it to work for our data analysis. This data is going to come with a lot of inconsistencies in them, including columns that are blank, missing values, abrupt values, and data that is in the wrong format. This means that we need to be able to clean the data and make it easier to run through our algorithm.

If we take a look at the data that we have above, we will find that there a lot of inconsistencies that come in the data that we have. There are parts in the pregnancy where the number 1, is written out as one, and then there are values of blood pressure that are up at 6600, which is not even possible for humans.

The income column is also going to be blank, mostly because it really doesn't matter when it comes to predicting diabetes in the first place and most doctors and their offices are not going to ask how much the patient makes in a year.

Because of all this information and these issues that show up in it, we will need to take some time to clean and preprocess the data before we can add it to our models or algorithms. We can clean the data by getting rid of the outliers, filling up the null values, and making sure that we normalize the data type. This is the second phase of preprocessing the data. We are then able to get clean data, which we can see with the chart below:

	npreg	glu	bp	skin	bmi	ped	age
1	6	148	72	35	33.6	0.627	50
2	1	85	66	29	26.6	0.351	31
3	1	89	80	23	28.1	0.167	21
4	3	78	50	32	31	0.248	26
5	2	197	70	45	30.5	0.158	53
6	5	166	72	19	25.8	0.587	51
7	0	118	84	47	45.8	0.551	31
8	1	103	30	38	43.3	0.183	33
9	3	126	88	41	39.3	0.704	27
10	9	119	80	35	29	0.263	29
11	1	97	66	15	23.2	0.487	22
12	5	109	75	26	36	0.546	60
13	3	88	58	11	24.8	0.267	22
14	10	122	78	31	27.6	0.512	45
15	4	97	60	33	24	0.966	33
16	9	102	76	37	32.9	0.665	46
17	2	90	68	42	38.2	0.503	27
18	4	111	72	47	37.1	1.39	56
19	3	180	64	25	34	0.271	26
20	7	106	92	18	39	0.235	48
21	9	171	110	24	45.4	0.721	54

Now we want to take some time to do an analysis for the third phase of this process. First, we need to be able to load all of this data into our analytical sandbox, and then apply a lot of statistical functions to it. For example, the R programming language can have a function that will describe things and give us unique or missing values. There are also a number of things that we can do with the help of the Python coding language, depending on which libraries we want to work with and how we want to handle the data.

As we go through this and mess around with the data a bit, we will be able to implement a few of the various visualization techniques including line graphs, histograms, and box plots. This makes it a lot easier to get a good idea of how data is distributed and what we can learn from this data because of how it is distributed.

From here, we are going to move on to the fourth phase. Going based on the insights that we got from the other step, the best fit for this kind of problem is going to be one of the decision trees. There are other algorithms that we are able to work with, but the decision tree is going to be the one that we will focus on

here.

Since we already have a few major attributes that we are able to use for the analysis, including the BMI and npreg, so we will employ managed learning techniques in order to assemble up the model that we have.

Further, we are going to use the decision tree because it is able to take all of the different attributes that we have into developing in one go, like the ones which have a straightforward relationship as well as those which have a non-linear relationship.

In our case, we are going to work with a relationship that is more linear between the age and the npreg, but then a relationship that is more nonlinear between the ped and the npreg.

And finally, the decision tree model can also become very robust because it allows us a chance to employ the unique combinations of characteristics to make different trees, and then eventually carry out the one with the ultimate amount of effects that we need to look through. Here, the most important parameter that we want to work with is the parameter of glucose.

So this is going to be the root node. Now, the current node and its value decided the next crucial specification to be appropriated. It goes on till we are going to get a result and this is going to be either neg or pas. Pos means that the person has a tendency to having diabetes, and neg means that they do not have a tendency of having diabetes.

Then we can move to the fifth step. In this phase, we are going to work on running a meager pilot examine in order to check whether or not our results are suitable for what we need and if we would like to enhance this to a larger scale. We can also use this to look for any constraints of the performance that maybe there. If the results that we see are not that accurate, then it is time to change up some things and figure out the best way to re-plan and rebuild our model.

Once we have been able to execute the project in a successful manner, we are going to share the output so that we can work on full deployment. We can then work with visualizations and more in order to figure out how to make the best decision and which course of action is going to be the best for our needs. Once everyone is on board, it is a lot easier to implement this throughout the whole company and see the great results in the process.

Being a data scientist means that you need to be able to go through all of these steps and get them to work, based on the data, as easily and smoothly as possible for the company. Following the steps above can help us to figure out whether a patient, based on a few factors, will suffer from diabetes or not.
And it can definitely be used in a lot of other formats as well. A data scientist has to be able to go through all of this information and determine how to use that data to get the best results.

Conclusion

Thank you for making it through to the end of *Python Data Science* , 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 implementing data science into your own business and seeing what results are available with this. Data science is taking over the business world, and many companies, no matter what industry they are in, have found that this kind of process is exactly what they need to not only collect the data they have but also to clean it and perform an analysis to find the insights and predictions that are inside. When data science is used in the proper manner, and we add in some Python to help create the models and more that is needed, we are going to be able to find the best way to make business decisions that improve our standing in the industry.

There are a lot of different parts that come with data science, and being able to put them all together can really help us to do better with helping our customers, finding new products to bring to market, and more. And with the help of this guidebook, we can hopefully find the best ways to beat out the competition and see the results that will work for us. It takes some time, and a good data analysis with the right algorithms from Python, but it can be one of the best ways to make some smart and sound decisions for your business.

The process of Python data science is not an easy one, and learning how to make this work for your needs, and to put all of the parts together can make a big difference in the way that you run your business, and how much success you will see when it comes to your business growing in the future. When you are ready to learn more about working with Python data science and how to make this work for your business, make sure to check out this guidebook to get started.

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