01 Course Overview

[Autogenerated] Hi, My name is Guillermo Fernandez. Welcome to my course, Espirito Re data analyses with Python am Morita, Scientist on Windsor. With you my own experience in this court, we're going to learn how to perform a complete exploratory data analysis also known as D A from into end by understanding the in Bolton Next went way on how to apply them by using python show of a signal. It's off. Python will ease your way. Some of the major topics that we will cover include being aware off the data analyses and 88 by plane, as well as its main goals on benefits apply in only varied on multivariate techniques to obtain insight from Giaretta, performing feature selection and feature engineering and also understanding how to communicate your findings based on your data. By the end of this course, you'll know which are the best 80 techniques to check your hypotheses and make better decisions based on data. I hope you'll join me on this journey to lane area with exploratory data analysis with Python course at Parasite

Understanding the Goals and Benefits of Exploratory Data Analysis (EDA)

01 Overview

[Autogenerated] Hi. Welcome to this model about understanding the gold. Some benefits off performing an exploratory data analysis. We will define what idiot is understand idiot gold, some benefits and give a quick, real example overview on know and understand the steps involved in the D A workflow.

02 What Is EDA

[Autogenerated] what is sport relate? Analysis, which from now on we will refer to it as idiot. Nowadays they say so much every available that its size and complexity desolate as discover its structure directly. Well, there doesn't seem to exist as short, unique definition that satisfies everything that involves a d A. Instead, we can say the following is an approach not a set off techniques, but an attitude or philosophy about how worried analysis should be carried out. If it is an approach, you might begin with certain assumptions. And if it is an attitude, it will involve an active mind trained to achieve search and goal, like evaluating. Our hypothesis is actively incisive rather than passively descriptive. So you don't stop at this craving data. You go a step farther to get no, let's and explain it our willingness to look for those things that we believe are not there as well as those we believe to be there. This is taking our sometimes on defy believes this way. We forced her discovery. We will want to discover unknown things in data even maybe top secret. A. D. A. Has several gold's here are them englands. Any other one that you can think of will be somehow related to this. E. D. A. Lets you understand your data and improvement. You're no let's about certain domain. If you want to solve some problems or shirts for the answer to a question, you need to understand your data and get knowledge by finding instructor in it. Draw valid conclusions. Diello may not be enough to construct completely precise conclusions, but for sure it will help to support them. 18. Decision making and planning. It helps to base your resistance in data or better said in the inside you get from it. This helps you get in right off from believes, in case you have them. Hope in Khost formalities in real life scenarios, this will not be usually easy, but D A will truly help you Discover were to look in order to spot the coast. Often effect, for example, do rainy days increased the cell of gasoline. This way, Edie A. Serves the partners off generating and confirming hypothesis by building intuition and gave in inside about data. What should you begin? The first thing you need to do is formulate a hypothesis. If you begin an idiot with no pulse, she's at all. You will probably end lost during the process. It's all off this enough to solve our data problem. Joan Wilder, Turkey, considered a reference in the D A. Said an approximate answer to direct problem is worth a good deal more than an exact answer. Tow an approximate problem when facing a data parole. You might end in a situation in which do did instruct all the bullets or certainty that you desired. In real life, this can happen. But don't get this courage. You formulate the correct question. Having an approximate answer to it is not as bad as it might look like. Something that will help is being creative. You build this kind of creativity by practicing and gaining knowledge. Let me briefly illustrate a real example where idea helped to shop a problem during World War two. Alan Turing Build a machine to the Creep, the Nazi's Enigma code Enigma was a complex miss. Its encryption machine and the \_\_\_\_\_ change its encryption key every day, so it seemed impossible to crack the coat by performing an area. Alan Turing and his team noticed that certain soldiers follow apartment for saying goodbye at the end of transmissions this way by manually decrypting this few words. Our tent. They could get them, keep some key for the day, introduce it in the decryption machine on finally cracked the \_\_\_\_ scope. This is a very good thing sample in which we see how to proceed, to perform an idiot unit to formulate an initial hypothesis and then try to take assumptions, discover structure in data, understand on learn. As you can see the D and rely just in maths they injected, they roll it into the problem. This is the D A attitude or philosophy.

03 The EDA Workflow

[Autogenerated] in this video, we will talk about the steps involved in any d. A workflow. Where does Eadie a bolo? It is embedded within the data analyses workflow. First, we need to have my name and scope. It refers to study a film based on some hypothesis. The aim and scope will define the features on the terminate expected to describe and explain the problem to be solved. Date. Elaboration. One's domain and features are chosen. We should only keep those that can be observed with reliability. Also, the acceptable values for valuables must be known. In order to validate data input, they take election in this face, the proceedings for collecting data and its story should be defined. Will there be automatic methods by taking service, for example, what will be the former in which data will be saved? The Winnie to model a database data input. Here we decide how to enter data in our storage system so it will be easily accessible and follow acceptable data formats to be consumed by a data management system, later managing it in place, creating a dating, extracting Agnes aggregating data Mon. Others Usually they did us in exist in the phone. We went for our buy in data analysis methods. In sand resources, this step is included a spot of the e T A workflow. We reach the area. We start without one dimensional exploration. Here we describe information by one variable at a time. Techniques will differ according to the tape of variable. If it's categorical, quantitative or a tame siris two dimensional expiration, its goal is to study the relationship or dependency between two variables. We can do this also with n variables representing simultaneously the relationships between all of them without any modeling factor analysis, they mention Ality Reduction Commission to play. Its objective is to represent data observations and features in idiomatic space, whose dimensionally says the smallest possible. This means the reduction off the initial variables to a smaller number off what is called factors. Principal component analysis is an example of this. This can help to sport, which are the most important features off our data data clustering Its goal is to reconnect classes or clusters off observations off variables here. Article clustering is one of these methods. After performing our e. T. A. We go to data synthesize in It consists in summaries in your results as explainable ous possible, making them ready for the communication step they take communicating here Do percent results in such a way that decisions can be made? The goal is to present valid conclusions based on understandable graphics. This enables decision making in your company or results Tim on finally their decision communication. You get to make decisions based on the inside obtained from your data on Communicate this decision to the stakeholders as a final remark. There exist other approaches about from media. These are classical data analyses. Bias in data analysis on dhe summary data analysis The differing the sequence off the analysis off course all start with the problem to solve. Then look at the data. Next we find that in city A and buy you see in that analysis, you guess a model the leader should follow on maker later analysis while in 88 you take the opposite approach by first analyzing the data before deciding which model is more appropriate. Finally, off course their conclusions. Incendiary data analysis on America Reduction off historical data is mate. This alone is a passive analysis of data Diego's father and tries to get no leads in an active way. Looking to our future application to end, we must say that in practice, a mixture of these approaches is usually applied as they complement each other.

Determining When and Why to Use Univariate Analysis

01 Overview

[Autogenerated] Hi. Welcome to these model about unit berate analyses He's more do on the next one will be like a puzzle. We will cover its parts on aims to later combined them. We will walk through the available unit very techniques. So you get used to them, understand when and why to use them on also understand the inside we can get from its technique. Finally, we will perform some of these techniques in python.

02 Tracing A Knowledge Map

[Autogenerated] let's begin by tracing and all its map. You know, very internalised he's has a scold. So similarly sub Serb ations to characterize data it's observations has some features or variables that define it within a very tight analysis. We want to analyze feature by feature in order to characterize this crepe and understand all the observations. This is all our. Later we have basically two ways of doing this. It did by numerical methods or graphical methods to represent the data information on enable her Understanding America and visual techniques both tried to represent the same information, but in different languages. We usually use both for a better understanding. Which kind of variables can we find? Basically three types off tables quantitative, more related with percents counts numbers. They have a measure on a numeric scale like height speed number off produce goods in a factory qualitative, categorical or nominal. More related with a description is a variable that only takes affinity. Number off this thing values, which are called categories defined by labels. Examples off. This can be the quality off some bottles of wine. Some other variables, like ready days, any day. Finally, we have chronological variables the air quantitative variables that are ordered by time. For example, the price of our stock in the stock market, our brain activity during one day. Any quantity of Barry Well, that makes sense to study with time. Idiot techniques can be thought off like a parcel or map in which its part has its goal. This is a common reference map for quantitative, qualitative and chronological variables in both univerity on multivariate analyses. If you want to show a distribution when you want to know your later local concentration, dispersion and safe, we will see later. What these men's comparison when you want to answer questions like How much is something greater than all the thing or when sand event happens compared to another one relationship when you want to look at how some variables behave together, this is related with multivariate analyses. Composition when you want to. So proportions between variables somehow is like a particular way of comparison. Take this map as a base line. I strongly recommend to look at the resource is I put in. This is like as they describe a very complete extent and map to Cooper. Almost every situation it will be very 10 consuming to cover all of them in the scores. So we will go through the main ones from here. You will perfectly understand any new chart as the underlying theory states the same.

03 Characterizing Data

[Autogenerated] let's summarize some basic concepts about how to characterize your later. For that. Let me introduce to you George or digital. Who was Scottish? Is statistician who stated some conditions for measurements in order to rigorously characterize a data set, Um, instruments would be independent off those server. This means that you and I should be able to compute the same number, applying the same formula to the same day to sit the bent on old values off Siri's. We would not discuss any observation off her ladies that the tape of value must have concrete meaning. It should be easy to compute and not sensitive to random processes. So there are what we call Messrs off Central Tendency. And this person these missiles help us answer some questions about our data like it's local concentration. For that, we have measures like they mean the median them out quintals. We will go through them in a minute, measures off. This person desires standard deviation variants. We have SK unis on DK you chose is giving a data set too easily will compute what is called the arithmetic mean. Imagine the whole world population height. Follow these distribution the X axis describes the hate value on the Y axis. How many people do have that except Kate? So you take the first person and see that hears he's very tall and force around here, then do the same for that expression. This time he was, he wasa little one. Keep on doing this. Still, you count the last person on Earth. If you sum up all the people hates on divide by the number off people, you will have the arithmetic mean off your data set. What's the median will give you big your observations according to a future value in an ascending or descending order. The median easy Bali, for which the number off observations, smaller and grated on it, are equal. It divides D frequency. He's to Graham into two equal surfaces. What does these men it would take again? The world population hate as our feature on half of the population is under 5.6 feet and the other half is taller than these value. Then it turns out to be the median off tuition mathematically expressed. If we have to Pete plus one observations people as well. Observation is the median on it have to \_\_\_ observations. The P observation is 1,000,000. What's the moat? The mode is the value over the maximum off the frequency distribution. We get the point of maximum frequency by looking at the Y axis. Supposed out 5.4 feet is the most frequent height among the world population. Then it would be the moat. If our distribution is unit bottle and symmetrical like this one, the following relation can be used to compute them off. Instead of counting the most recon value for our feature, Quintiles one tells our values of the features that divide our distribution into equal parts. These means equal areas. So, for example, if we're taking about choir tiles were dividing up distribution into four equal areas. So we have the first court trial, the second, the third. We only have three, but they divide our distribution in four equal areas. The Siles is the same concept, but instead of four areas, we have 10. So the 1st 1 defines the 1st 10% the 2nd 3rd and so on. Then we have measures off this person. We also want to know how this person is already down. But this purse from what From the central value. This is the mean. We can compute what is called the standard deviation. Following this formula, it gives a mystery bit off how far they days away from them in value. We usually work with the barriers to that is simply the school body off the standard deviation. In our example, the top distribution has a lower barriers that the bottom one it has less. This person in our latest says we can also have what is called out liars, and these held liars can affect our measures off. Central tendency on outlier is defined as an observation that has some extreme value. You're really out. What is our child? So we have Cuban and coo three. The distance between them is known as the \_\_\_\_\_\_\_\_\_\_\_ Trial Rains, so a note liar is defined as those officer racers. That's our at least 1.5 five times the inter cartel range of off from either Q one or Christie. If an observation is fear more times far from these distance, we say it is an extreme outlier. Depending on our data problem, out layers could be good or could be bad. For example, if we are analyzing a manufacturing process in a factory. Maybe there's some machine that he's broken and makes good with anomalies. If we're studying our production standards, maybe we should get rid of these anomalies while making our analysis instead. If we are performing fraud detection in a finance problem, for example, a normal behaviors are important to take into consideration. So this time we should keep out layers. In the analysis. SK unis are keratosis que nous tells us the degree off a symmetry off a distribution. It follows this formula three times the mean miners, the median divided by the standard deviation. If it's to the left, we say it's never people. It's cute. If it's to the right, we say it's positively skewed. Encore Tosses, it tells, has the degree of bigness off a distribution, usually taking relative to a normal distribution. It follows these formulas. If it's about three, it's extremely picked. If it equals three is normally big and if it is below three, is somewhat flattered. Indeed, they are called left a coated mess. Aquatic unpractical, tick

04 Demo Characterizing Data

[Autogenerated] in the CMO, we will learn different ways to compute missiles off central tendency on this person With python, we will be using by some packages, study sticks, pandas none by side by stats. And we will also put our first graph with Seaborn. Let me present you the D mortals. You probably know about Jupiter notebooks for running scrapes by code blocks. Google has made available a free service for running. You put their notebooks called CA Laboratory, with a pre configured environment that once on the clout, this means that most of the common libraries we will need to use are already available. So this is the tool that will be using on our the most about visualization libraries. We will be using that plot lip with this. Python's most well known brought in library Seaborn is Harada Visualization Library, based on mattered lip, providing a high level in their face, and it tries to make things easier to goat, although sometimes we will need to use my \_\_\_\_\_\_ leap if we want to customize some parameters in a low level way. First we'll import. They needed libraries like pandas numb by mud, put Lip and Seaborn, we begin by looking at the features, we're going to read a data set about forest fires. We get this data set from the machine learning repositories off you see a after having a quick look to the data, we're going to compute the local concentration measurements. Let's go for data temperature on. We're going to compute them in by three different ways with pandas, with statistics on with number, we have three alternatives. A one, a two, a three, and we see that all the values are the same. Now let's compute the moat number. He doesn't have a way to compute the moat, so we're going to use the stats library. As you can see, the motor function from this Texas library gives a never a sage respects only one value to be remote, but it found two equal values in a similar way. We compute the median, having a 19.3 value. Let's go for the dispersing values. First, we're going to compute the standard deviation. As you see, values are almost the same, although they are slight differences. And if we compute the variance, we can take that he's the squirt value of the standard deviation What about the shape? We're going to compute this que nous with panda sensei pie. We get more or less a similar value. And what about coach horses? Well, we have two different factions on Banda Sensei Pie, and we have again similar values. But if you don't want to compute these separately, you have ah, function called describe that outputs our table with the most common measurements computed for you like they mean the standard deviation, the meaning max values on the cartels. Now we're going to do our first be civilization. We import their seaboard library said the color coats on. We're going to plot the distribution for the temperature value off our data that we're also going to use my \_\_\_\_\_\_ lip as we're going to plot. Vertical lines to visualize were demean the median on the quarters are he had you see the meat and the 1,000,000 in blue and red color. If we want to plot particular lines off the quarters, we can do so. We can also said, I'm the supplies on upper a liar limit on lower outlier limit

05 Univariate Distribution Plots

[Autogenerated] Let's begin with only bury it. Distribution pots. Let's divide our data set into intervals off observations on Let's call them classes. This way we comport. Ah, hissed a gram. It is a set of rectangles on a horizontal axis with centers at the Classmate Point on lends equal to the class in their bus ices. It's your dangle is called a bin. For example. These being could represent the classic off people with height between Fight pull three on 5.5 Fitz being the classmate 0.5 point four with 700,000,000 people. For example, If we're taking the whole world population, these other bean, we have people between 5.5 on 5.7 fit being the classmate 0.5 point six with maybe around 900,000,000 people. As you can see, it's rock tunnel has the same with Alan the horizontal axis. Although they don't need to, they are usually the same but can have different heights on a vertical axis. This way, it rectangle area is proportional to the class frequency. This is because it is a frequency distribution. Ah, hey! So Graham, let us know what is the same off have a distribution where they mean approximately is on if it has low or high dispersion. This lets us introduced a frequency. Polygon is Alan Graf, off class frequency Plotted against classmate points. It basically gives the same information that in the history but plotted in other style group data. He generally destroyed much of the regional detail off the data but highlights the main features contained in both Hissed a grandma frequency polygon. If we were able to choose infinite Ben's, it will approximate as Most Corp Next, we have a density plot. Let's has visualize the distribution off our data over a continuous interval on the vertical axis instead of an absolute frequency, it'll says what is called the probability density function. What is the probably density faction is the probability unit on the horizontal axis on its properties that the area under the curve should some upto one. This means that if we buy the girls in many little rectangles like this on, we multiply its little with by its height, which is the value off the probability density function on. We keep doing this until we have all the area under the curve covert. It should equal one. What does he do? Useful for it Better tells this tape off the distribution that history because it is not affected by the number of beans, you'll really know the main different types of distributions. You can find a \_\_\_\_ quick reminder. These are unique model symmetric, normal distribution by model distribution right and left skewed distribution on high and low pick distribution. We can also put what is called a buck's book is at this person graph. It visually solves the summary off central tendency misters on dispersion. The central box does has worked quite. Tiles are then the so called whiskers. At the end of the whiskers, we find the minimum maximum values off the cartels, miners or plus the inter Quilty range finally a separate points above and below. This limits the pliers. If they are too far, they're called extreme of players. This plot serves the purpose off. Putting different characteristics on the same graph gives a quick vision of symmetry. How close median amine are and you get a quick general idea off. Discretion also is useful for spotting out liars. Similar trees we find the by allowing plot follows the same ideas than box plot, but instead So the probability density plot making us remember the Sape off a violin. So we have these names city on. We mirror it on this. We have the violin plot. The strength off a violin plot is that it sourced the full distribution. This is useful when they dies multimodal because in this case, a box, but would be enough. Imagine this box plot. You can try guessing which distribution it relates to, but indeed it can relate to many different ones. You might think it corresponds to something like this, but in this example, the institution for which this box plot was generated. Whilst this one Next we have the strict blood inviting. You have this kind of distribution. A \_\_\_\_\_ \_\_\_\_ is a part of distorted values alone. One access something like this. But with all these points overlap, we say nothing. So we are what is called a bit off deter. This way the vertical axis doesn't represent any value at all. It is usually preferred. Over I saw were emergency deployed when we have a small data set. If not, we will have too many points on. They would be difficult to defer. Assurance brought is almost the same as a straight. But but observations are adjusted to avoid overlapping, providing a better visualization again, it is better recommended for small data sets on the vertical axis doesn't represent any value.

06 Demo Univariate Distribution Plots

[Autogenerated] in this, the more you will learn how to plot, you know, very distribution chats with python. We will be using python packages, math, not Lip and Seaborn, for Do It is you will also learn how to customize some simple, rather aesthetics like the color. Let's begin. By important, they needed libraries like Seaborn Mattress sleep number. I will set a random sit on defined. I mean on a variance off zero and one with numb By random, we can create a normal distribution. And with this plot, we complete the Zeta half iced graham with her then city block. By default, we can play with the fashion parameters like core its color on laying. With this way, we can see the history Graham in a clear way. We can also define the number of beans you see if we put 10 on dhe if we put 50 were then also said de que ta two false. So instead of having the density function, we have the pure hissed a gram with no density of action, and we also can put a shade related to the density fashion. If we only want to presidency. The faction is the key D A plot for doing this again. We can put the color we want undefined shape again. Pay attention to this because if we have only they surround. We have absolute values. But if we plot the density faction what here is values for the density function, not absolute values for the history. Graham, what about the box plot is a simple Is this just a function with the same name on here? We can see the layers. We can also play by defining some color. And this figure object has elements named artists. And we can set the color off each one. For example, in this for rope with red green on blue patterns on an Alfa for transparency here we said also all the line, some point torrent. But there was a little back here. I only wanted the elements. Not all of the plot. Now it looks better. Myelin plot as simple as this again. A function with the same name. As you see, we have the same information as inbox plot. But we also see the shape off the distribution. Although we don't see outliers so clear, we gave a bit off Corrigan. If we plot a violin plot with matter clip. We easily can decide if we want it in an horizontal way or vertical away. And again we have its parts called bodies and we can set its core again. As in the case off the seaborne functions, we can decide to show they mean don't worry about all these parameters. Is as a matter off getting usedto this functions and read a the commendation there plenty of parameters. The important thing here is your stick with the idea on the function that floats your data in a simple way. We have also a strip plot function for these We're going to load one off the sea. Born later, sets about tips on We need to define the future we want to plot. We can also define the jitter. If we put a cheater off zero, we see nothing. Everything is overlapped. If we put one, we see things a bit more clear. But remember that the vertical axis doesn't represent any value is just a matter of Peter. Finally, we have also this run put in this case. We don't need to define any Jeeter. The plot itself orders the point so they don't overlap. As you see, these are perfect correspondence between the horizontal axis off both plots as we are brought in the same data.

07 Univariate Comparison Plots

[Autogenerated] What about comparison bullets? Let's begin with the bio diagram. It presents categorical data with rectangular bars. The with must be the same for every rectangle one axis represents. The categories on the other access their value. There is no established order in which to put the different categories. For example, this could be investments on science and technology during last year. For the following countries, these diagram helps as camper the value off a numeric variable among different categorical variables off our data set. As you see, it is totally different from a history. Graham, keep this in mind. What's a line chart? It is a two dimensional graph made off off lanes that connect different points. This points represent the value off some variable. For example, this could be the amount of sunlight in London by hours, you could say Hey, wait a minute. This is the memorial about. You never ate analysis. Where do we have some light on time? You're right. But we're not cheating. We are representing a chronological variable. This is time dependent. It is somehow halfway between universe on multivariate analysis. This candle graph helps us to understand how your body will revolves. overtime based on the same area. We have the rent IRT. It can be thought off like a type of line chart. But instead of showing the evolution off a valuable with time as a general idea, it is usually related with some kindof process, giving some measure off central tendency like they mean or the median. As a reference. This kind of graph helps detect sifts in location and scale. Also out, liars can be easily detected. Runts That's our allies to find anomalies that makes you just sits in a process over time on also factors influence in its variability. Here, for example, process a approaches we, and we can compare them with the previous processes mean What's the spark line? It was introduced by Edward, 50 in 1983. It represents the general shape on variation off Sam value over time, this could be stock market prices, humidity temperature. It encourages an extreme compaction of visual information usually represent with her point. At the end, the actual value on usually know also the initial value of the graph. It is used when you want to know if a value far variable is going up or down since reported without very quick look at it, representing a general tendency or shape off inspiration. Lack put electrodes represents a chronology. Calvary will on the horizontal axis, elect representation, off points and in the vertical axis. The observation that happened immediately after, for example, air pollution values AT T minus one against the values one hour later AT T. It helps to spot outliers easily in time. Siri's also let us find linear patterns that can suggest that date is also correlated, and we can also take it. There is randomness like this. Next. This is Gloria Tart. It looks like us by the Web or Rader. It's reduced, corresponds to a point in a time. Variable, for example, months and Alan. It's one will represent the values off our variable off interest. For example, rainy days during that month. It's like if we get Alliance erred on bended till it has a circular shape. This kind of chart is also called sundry plot. It helps us to compare the value of a variable among several categorical variables or the evolution in time off one variable. We also have Cata grams, he said. Geographical comparision off a variable this is a variable. The has a geographic relation, and for these reasons it involves a map. It gives us a sense off the value of her variable according to a set off locations represented, usually by different colors or patterns.

08 Demo Univariate Comparison Plots

[Autogenerated] in this demo, we are going to learn how to brought you Never. It combines on SATs with by phone again. We're going to use several python packages like Matt brought Lip Seaborn Pandas, do you blunders and your blood. We will also learn how to customize some simple graphics that x like color. As you swell, we begin importing our needed libraries. We're going to load the deep state of that. And we're going to use the about plot function from Stephen to plot the total bill against the group size off the people that goes toe lance or dinner or whatever on the horizontal axis will specify the class in these cases, the grip size on on the vertical axis. The amount we want to put in this case is the total bill, which specifies the data he steps on. We put some color and there we have. It seems that the bigger the group, the higher the total bill. Okay, this makes sense. We can do the same with my bra clip. We're going to make up some values here. Divine some labels. Would you find a new order away with some pie on with plot bar we specify the position to which we want to represent the values and we can set red, green, blue color on transparency parameter. You can also define the its color on the line with specified there ticks for representing also the labels name. Another way to do this is through their \_\_\_\_\_\_\_ that a friend faction. Now let's create a data frame with sunlight values on time. This is how it looks on with lying plot from C one. We just need to define the horizontal and vertical axis to represent these data. There you will. We have sunlight. Let's use the time off day by hours from 6 to 10. If we just specified the data. The line plot is also represented. But as we didn't find any horizontal axis, it has plots a sequential Siri's from 0 to 4 instead, off the time variable we want to plot. Now we're going to have to already frame values off wind spit. Now, our data framing looks like this. And for putting both variables, we can define data as a least off these two columns. This is a quick way off floating around that, for example, in this case, is sunlight and wind spit, but it could be signed light over different days or manufacturing processes. But by defining late at his way, we don't have the time variable. So we need to change form our data frame with the multi faction, so sunlight and wind speeds are melted into one column in our data frame. Looks like this. If we now plot again our line plot on define the re cental access s time on the vertical axis as the value on we had a new permit er called Hugh equal to our variable column. We have both barrels represented on the same chart, and we also cuffed the time variable again. This is a simple way for putting a rented if necessary. We can do the same with the math lutely library, but here instead we have a bit of freedom to define some parameters so we can make up some data. Bloodied as a line chart, I used to feel between faction defined the feeling limits. We want to feel the area below the function with some transparency on the color, which is it's OK. We have now are implode with a shade, but we can customize these even more. We can put a point at the end of our data. This way we have turns for Ally entered into a sparkling Next we want to put electrodes So we're going to use thunders floating function toe peacefully represent these data. As you can see, these data is thought as a temple, a serious and we can define the number off lacks between its representation We have 54 and to LAX after 65. Next, we're going to put a circular area, Todd. So we defined, for example, some subjects on dhe. The restaurant's related to them on. We define a traitor for him with one student named Young. With these grates on the subjects, we're going to convert these grates to a list and copy the first element as the last one. We all show t to compute the angles. To wit, it's great is going to be represented. These there are angles in radiance. We find our put in polar format disease represented by uncles On our radius measurement. We define the ticks and put the angles best was the values We're going to put some color on transparency. There it is for putting Carta Grams. We're going to rely in an external library called Dupe. Unless we can install it in. Are you paternal book? As in my case, you will probably need to install some other dependencies. Once the installation is finished, we import you pandas and your plot, we're goingto load one off The default data sets from Japan is called natural Earth Lowers on compute the GDP per person in a population. This is a quick look to the later we have to population the GDP on. We have computed the GDP per person these way By using dear plot Coppola function, we can plot our data frame values, giving us you the GDP per person We have just computed we can set the figure size on the palette color in this case screens. Here we see indifferent states of green the GDP per person related to its country in the world having more wealth. The darkest countries

09 Univariate Composition Plots

[Autogenerated] composition plots are similar to competition plots in the sense that they serve for comparing variable magnitudes. But in a way in which you visually join areas representing those magnitudes, Here's the biter. It's a circle with its area divided in different portions. That's it's off member percent in a proportion off the total area. It should help to ease the G civilisation off different proportions. But we, as humans, are not very good at interpreting circle or areas with precision. Alice the distinction is a very simple one, like the one in this slip. For this reason, they're not recommended to use on less necessary. And only two items are sewn, and you only want to show that one is greater than the other waffle char. Sometimes they're called the square picture, although awful presents the advantage off. So in equally sized atomic square areas, you're not supposed to manually count it squared off course. But you can get a sense off which group with same color is bigger, smaller than the other one. For example, here we represent the MBA titles again. It helps you find proportions between variables here in a blink of an eye, you can spot the two teams with more titles in the trim up. The viol quantities are represented by rectangle arias proportional to them. It helps us to get an idea about proportions, but we don't know the exact value related to its area unless we brought these value, too. You know what a full chart is? A chart that helps understand the cumulative effect off positive and negative values introduce sequentially. It is also known as Flying Breaks Chart or Mario Chart for being alike. The platforms from the game in finance is referred to as Britt. It helps us to see the accumulation effects to certain quantity in this craft, for example, these numbers could corresponds to sales credits later purchases. Finally, usual notice that in this chart, data is represented sequentially.

10 Demo Univariate Composition Plots

[Autogenerated] in this, the more we will learn how to plot, you know, buried composition sides with python you seem packages. Mudblood clip Seaborn, pandas by waffle and square. If I, We will also learn again how to customize some simple graphic aesthetics like the color. Let's import her libraries. As usual, Seaborn mattered. Leap and number. You want to make out some later on dhe, Use the Pie Faction from Matt Clip. We can specify also the start angle off course. The color and the words props off the external lines lying with and transparency. Nowhere went to install pie waffle After the installation is complete, we worked import by waffle. We make out some later. For example, the number off rings some, eh? NBA teams half a tip. Now we said the figure properties figure class equals waffle. We want six rows. Undervalues is going to be all data hopes Waffle is no defined. Instead, import waffle. We need to do from my waffle import waffle. We can also said a lot of for the parameters like the title colors the labels off the greater items Celtics, Lakers, Bulls on Also brilliant. Now this is the output and this is how data items Look. Now we're going to install square. If I once it is installed, we can do import square. If I again we're goingto make up some data on some labels from a to eat ondas with square. If I That plot on defining the sizes parameter equals data on the labels as labels, we have our dream of plot off course we can set the cause. We want green, gray, blue, red and purple. We can also said transparency. Now get going toe make out some data for the waterfall chart. This is not a simple teacher to blot under so direct function to do this. So we're going to set some categories on some quantities on Define a great a frame with these values like we win some money in the lottery. But then we have to pay a credit. We're going to use the accumulated some property deceived on feel non values. This way we have a second column that is the same as the 1st 1 but shifted and with the created some off the quantities, we're going to save these aggravated some in a separate burial. We're going to plot our little friend. Values specify in the kind of plot being bar for being stacked. And we also have to define the bottom values. So he gets these wonderful shape. We will define the horizontal lines on. We're going to do this by repeating values three times sifting on placing, then every three positions. We will also need to know the total amount, the final sum in order to plot it. So we brought kind but brought stacked equals true. The bottom values off reverends, they're going to be the accumulated some Andi are orange color. We did find some horizontal lines tojoin the bars were going to plot in black color on Dhe. This is our final waterfall shirt.

11 Univariate Analysis Tests

[Autogenerated] when performing on honey period analysis. Some numerical techniques prove to be useful. We will perform a very brief introduction to the two most common ones. Hypothesis. Testing. A hypothesis isn't educated. Guess about something in the world around you. It would be testable either by experiment or observation. Hypothesis. Testing is the use off statistics to determine the probability that, given hypothesis is true, the usual process consists off four steps. First, you formulate what is called the null hypothesis disease. They accepted fact. Then you stayed on. Alternate hypothesis related to the fact happening might sense. You also state what is called the rejection region, and finally you test if their Selves scenario is statistically significant. I just want to stick with the main concept so it will present an example a simple as possible. You might think we have a process that behaves related to a normal distribution like this. Let's imagine we're manufacturing three reprinted parts, all expected to be around the same size one day. We see that were manufactured 30 pieces with that mean size or 15 but we think they should have a mean off 10. Maybe something happened to the printing conditions that provoked this change. Or maybe not and they were just obtained by tense. How can we know? Well, we will state are no hypotheses off having a mean off size 10 and our alternate hypothesis off having a mean greater than 10 which state what is called our elf, a level which defines the limit off the so called reaction area. Next we compute what is called the set score, which is the value after normalizing the distribution toe, a normal distribution centered on zero with a variance of one. After normalizing our distribution there, 10 becomes a zero. The limit that defines our reaction area is one point 65 on the 15 times to 3.91. As you see, 15 or 3.91 is within the reaction area. This means that if the new lie prophecies was true off serving a mean off 15 as probably the off less than 5%. So we reject the null hypothesis. This is we reject that the observed 30 pieces were manufactured under the same conditions as the one we know that have mean size off 10. These 3.91 value is related to this blue area, and this area is what is known as the P value. The P value is the probability off unobserved or more extreme value to happen. Assuming that the newly hypothesis is true, if the P value is lower than our alpha level, we reject the no hypothesis. Next, we will give her briefing detection toe on over and over or analysis. Off variance is a statistical test for it. There's unity of means by analysis off group balances. This means journalist sample means based on the sample. Variance is that's the reason for the name and over. Imagine. You have a group off the desserts and you want to perform hypothesis. Test over there means, but to compute the mean off it's one is expensive because they all have a lot of later. So you take a sample from its data set and compute the means. But now you realize that performing eye prosthesis test over all the possible combinations off to data sets will be very time consuming. When you go, it's just to know if there is any that behaves in a different way. For sure, being this, we have the a noble method now we can define our no hypothesis as all the means are equal on the alternate hypotheses, as at least one is different from the others, the main This advantage here is that just with the nova method, you will not know we to mean is different. And for doing this you will need all the methods. Finally, you should keep in mind that for applying hypothesis test on an Obama methods, you will also need to test some assumptions like normality. Later sets must behave with a normal distribution. Almost did. Estes ity Marion's of data sets should be homogeneous on independent observations. This means that data sets must be independent from each other.

12 Demo Univariate Analysis Tests

[Autogenerated] in this demo, we will How to perform a quick hypothesis on a no protests for these will be used by from packets site by. We will begin by important plantas on Saipan libraries. Next, we're going to import Iris that a set that is, have every well known data's hurt about flowers from the U. C. I depository. These data set looks like this. It has some features or columns like the zeppelin principle with but the length petal with on the class, we're going to save this. It'll psych class in a separate burial were said they next. So it starts counting on See Reagan. We're going to do the same with the bicycle class on it looks like this. Now we have two different eight frames. With these two classes on, we are going to perform a hypothesis test the stairs, remember, is going to compare or to have has no hypotheses that both they mean over the future, which is in this case, simple with is the same in both classes. As you see, we have happy value that is nearly cereal. These means that we reject the null hypothesis. So the separate with off these two classes cannot be said to belong to the same distribution on with the function f one way from Cyprus that's going to perform on an over test. Remember that with a Nova test, our new hypothesis is that all means are equal. We're going to choose again. The simple with feature off the three different classes Class Sentosa class basic or on class Bill Jessica As you can see the people you is again, nearly zero on this again means that we reject the little hypotheses so at least one of the classes has possible with that is different from the other glasses.

Determining When and Why Multivariate Analysis

01 Overview

[Autogenerated] Hi. Welcome to this model about multivariate analysis. In this model, we will walk through there by level Motive Area techniques understand, when it's hit more appropriate to use them on way also understand the insight we can get from its technique. This way we know which questions we can ask, and finally we will perform some off these techniques in python.

02 Finding Relationships In Data

[Autogenerated] in multivariate analyses the recently to study and find relationships in data in this model will introduce some techniques at high level to find and study these relationships. We perform multivariate analyses in order to show the statistical dependence between variables were impossible to sum up the relations by a law operation or a statistical dependence on to characterize them by a numerical coefficient independent off the units off. Measure off the variables. A lot of the ways off studying and computed relations in data. I will have it quickly because they are the kind of ideas do you really learn when you put them into practice? Ask yourself questions and try to answer them to see the results. One of the's techniques is the Koreans is semester off doing probability. This means that computed over two identical variables, it will be equal to the variance, its source tendency in a linear relationship, either positive or negative, and it has units. This means that it is not easy to interpret the strength off the relationship because it is not normalized. The convergence intuition is as follows is the measure off the Gen Probability off to random variables. As you see is respected value. This is them in off the multiplication off the differences between its burial with its own mean, basically, you take the first point. I multiplied this to distances, then a second point, and so on, adding all the computations on dividing by the number of points. This way. A positive co variance means that the two aerials move in the same direction on a negative co variants that are inversely related. But as it is not a standard measurement, you can't tell the degree of relationship for solving this. We have the correlation. It is useful because it can indicate a predictive relationship. It owns hotels in strength, on direction. Off relationship, either negative or positive on commonly refers to a linear relationship. One off the most common used correlations is the peace one correlation. This is its formula. Is the Koreans divided by your valuables? Standard deviations Also, they noted, by the Greek letter row Correlation is about sewing that when something happens, some other thing happens, but is not about. If something happens, then some other thing must happen. It's a bit tricky, but this is a key area about correlation. This concept is commonly known as Correlation does not imply causation. If the change in one variable is accompanied by a change in the other than the variables are said to be correlated. Correlation is only appropriate for examining the relationship between meaningful, quantifiable data rather than categorical data such as gender color. Correlation can be between one and miners one, while barbarians has no defiant reigns. Remember this concept. It is said that to Ariel Sharon abandoned if the occurrence of one does not affect the brutality off occurance off the other. But be aware of something very important. If the correlation off two variables is zero, like in this example, this doesn't mean that they are independent. So you should keep in mind that if two body will start independent, the correlation is cereal. But a correlation of Shiro does not imply that to Barry Walsh are independent. A correlation value cannot replace visual examination of the data. You need to look at the distribution. Also, I'm the most important one. The presence off correlation is not enough to infer Cochet, Shin

03 Demo Finding Relationships In Data

[Autogenerated] in these, the more we learn how to compute co variants on Correlation with python by using baggage is lump. I am pandas. First, we're going to import Nan Pi Pandas civil on mattered leap libraries. We are going to load it Deep state is that now we're going to compute the Kobe Arians First, we're going to start with the cost function from number on. We're going to select columns, total bill and tips. So we obtain a matrix. The diagonal is the Koreans off its variable with itself on the other values. The 8.32 is the co variance between the total bill on the tip Will you have the cough function from pandas data frame that provides a visualization off The co variances between all the new married party was off the data frame. Not going to compute the correlation remembered article relation. I have values between minds one on one, while Kocharian's have no limited range again, we're going to use numb pie cork 0/2 function and we have her matrix. The diagonal. Is there a correlation off its burial with itself on the correlation between total Leland tip is 0.67. If we plot a scatter plot, you can see that we would be able to draw a line were present in these she'll 0.67 value course. We also have dependents core function that's so sick relation between all the numeric variables off the later frame. We're going to start these values in a variable, and we're going to draw a hitman for base wise in this palace. He, as you can see it's column, are row with variable as a label on a color, representing the correlation between two off them.

04 Multivariate Distribution Plots

[Autogenerated] it's beginning with multi berry distribution plots. First we have this cutter plot. It is the analysis off to quantitative variables also called this passion plot, because it highlights the expression off the cloud of points. For example, this could represent the same results off some students against the amount off hours study. It lets us be swell, you detect either could be a relationship between two variables. In this example, there seems to be a positive correlation between both variables. We can represent this catapult with these storms off both burials at the sites this history grams are called marginal. Distributions on together represent what is called the joint distribution off the variables. This way we can get a complete description off both variables distributions. If we represent the John distribution off this two aerials as a three d historian, we get what is called a steer a gram or be buried history. Graham. We can also put a surface area. For example, if you pick on Lee, the top values off a bribery. It hissed a gram and smooth transitions between these values. You will get a surface area area tarts as you squint the summation off. Quantitative data is to be communicated rather than the BBD ll data values. This lets us introduced with our level curves the a two dimensional representation off a three dimensional surface. If you get a surface with a serious off parallel, Blaine's equally this thing between them. You will get its level carves related to the vertical axis you choose. It lets us visually analyzed the steepness off the surfers. We can apply this idea to the joint the solution off two barrels and get a two dimensional 80 after distribution. In this case, the difference with this theory a gram is that it so's the density faction instead of absolute values. In this example, this point appears to be in High Slope area, and they saw the one in a low slope area we can also represented along without stayed in a more clear way. Having always in mind that it's curved is equally spaced, particularly from its neighbor carbs. This one could represent the height off 0.5 These ones here appoint one has all we also have box plots as in unit great analysis box plots graphically sold the sabri off central tendency, measures and discretion in multivariate analysis, Hopeless compared them on different classes. Off our data set, for example, the horizontal axes guru percent the classes weekdays and weakens the vertical axis. The number of people falling in love during weekdays or weekends on the color. If it happens on Sonny or any countries you can also represent by Ellen, plots seem as before with books plots but now, with the extra information off the distribution shape among classes.

05 Demo Multivariate Distribution Plots

[Autogenerated] in these \_\_\_\_\_. We are going to learn how to plot multivariate comparison shots with python. You seen the packages, Matt? Put leap and see Boram as usual there. How to customize some simple graphic aesthetics As usual, we import, say born mature clip pandas and numb pie as well as the tips data set with it. We're going to use scatter plot function from Sieber. Ah, horizontal axis is going to be the total bill or vertical axis. The tip on golf course, the data, the tips we put all rich color I'm gonna soul to set our bigger size. Next, we're going to use the joint blood function. This way, we're going to visualize the joint extrusion about the total bill on the tips. Here we have not only the scatter plot, but the hist a grams off both variables. We can also set different kinds off representation like regression. So we see a line that tries to feet the data on Also, we see the density distribution on top off the history grams. We also can set a kind off X. So we see the density off the data on exact figures. They also exist. They conceded institution in which we see a level course kind of plot. And instead of the History grams, we now have the density faction off its variable welcome going to use in peel tool kits. Library toe brought a three D diagram. We said our figure with Matt got lip. We define the access properties we're going to specify. It's a three D projection. We use a function called Plot Tree. Scharf is going to a plot. Ah, three d starfish. And here we have our float. As you see, maybe it's not a good choice for representing our data as we can see nothing but at least serves as an example of how to blow to this kind of figure. Be the height off the chart related with the group Shay's that went toe lunch or dinner. If we change the sign, it's somehow equivalent to rotating detect, but we still see nothing interesting. So as I said, for these particular case, this is not a good choice off blood. Now get going toe use the plot par three D function. Using the same data as you can see is somehow similar to this her face plot. But in this case we see the bar's not only the top surface off the data points in this case, it doesn't seem to be a good choice again, for the swell is in these data, as is everything a bit cluttered. Now we're going to be swell eyes with books plot the deeps left by smokers and non smokers. Ah, horizontal axis is going to be the time on our vertical axis the tips and the hue if the group uars a \_\_\_\_\_\_ or non \_\_\_\_\_\_ one. So here you see that durian land, it seems that smokers give a little more tip than non smokers. But during dinner time they seem to be more equal. Here. You see, there are people that are specified us out liars that leave a huge amount of the We can use this in data, but instead here's a violin plot. Now, as you know, we also see the distribution. So, for example, during lands time, non smokers have far more skewed distribution. Under in dinner is the smokers, which have a more skewed distribution

06 Multivariate Comparison Plots

[Autogenerated] let's begin with multi varied comparison bloods as you'll really know about that percents. Categorical data with rectangular bars on the with must be the same for every rich jungle. Apply to multivariate analyses it let us compare quantitative variables among classes. According to the example, there seems to be more people falling in love during weekends in seven countries. Height map is hap lot, easily composed by small scores that represent the value off some variable related to other, too, with different colors within a range of shade on a bright, dark scale. It helps us to see not only the abolishes the air, but if they might exist, I'm kind of tendency or value concentration. In this example, we have the number of flights related to dears on months. We can see very quickly that the in summertime, the recent increase in the flight number, as well as if we compare earlier years with past ones in a neighborhood analysis whistle, how we could reverse and the quantity of our variable related to its geographical localization. In multivariate analysis, we see that it's possible to represent more than one burial. In this example, we can see the population size off its country in Africa, given by their size off the country's safe and at the same time we're percent its wealth, with color being the brighter colors related to richer countries.

07 Demo Multivariate Comparison Plots

[Autogenerated] in the Zemo. We're going to learn how to plot multi variant comparison sites with python. You've seen Marta? Totally. See Warm. Found this. D'oh! Pandas and tear blood. We will also play with Semistate experimenters. As usual, we import Siebel not totally pandas and non pay libraries. And again, we're going to load. The deep state has said from Seaborn, As you already know, the data frame looks like this. We're going to say the figure size armed with that blood to see in the future to be displayed on the horizontal axis on with Why the tip on the vertical axis we're going toe. Also, choose the hue that is \_\_\_. And of course, the data tips. This plots kind off strip plot. Ingrid's, you can see the tip of the group was a \_\_\_\_\_\_. No. And if it there were male or female. But if we specify the kind off the cut plot writing bar kind equals bar. We get about proteins that here it seems that known \_\_\_\_\_\_ males give her your tips. Now we're going to load another data set from Seaborn about flights. This is how it looks. You have the deer the month on the passengers during that period. We're going to use a function called people, and we're going to people over a month and year, and this is how it looks. Now. We have instead of our data frame for matrix with the deer, the month on the number of passengers again. But in another format, this format is going to be useful for plotting a heat map. So with the hitman function from cheaper, we just choose the flights. You can see a number of passengers distributed alone months over years as the new neighbor, it analyses were goingto use again. D'oh! Pandas and Geo plot. As this is a different no book, we need to install them again. No, the insulation is finished. We're going to use the same day to say that's before the natural else. Lois on. We're going to compute again the GDP per person as you remember this, the plot with it in univerity analysis and now with the cattle Graham function, we're going to specify which continent scale related to the population size on issue, the GDP per person. We're going also to set the limits and figure size, and, as you can see we are visualizing three body will stuttering the population size rated to this size off the country shape on the GDP per person related to the color assigned to its country.

08 Multivariate Relationship Plots

[Autogenerated] What about multi varied relationship puts? We find again the scatter plot to order. You know this is because it not only gives an idea about the body of yours distribution, but also about the relationship. We can go one step farther on instead of represented. Equally sized points blocked them with their size related to some quantity. This way we have what is called a bubble blood, for example. Here it's bubble size is related to the number of students related to the bubble under consideration.

09 Demo Multivariate Relationship Plots

[Autogenerated] In these demo, you will learn how to plot a multivariate comparison sets with python. You seem python packages, but to sleep and Seabourn as usual and, as always, lead to customize some simple graphics that X like the color. So we begin by important our libraries on dear tips. They decide again, This is how it looks. And now we're going to use this scatter plot function from she born, and we get applaud like this, visualizing the total well against the tips. Next, we're going to use the function. 11 plot are horizontal axis is going to be the total build on our vertical axis. The tip off course the date of the tapes were going to define some parameters like s for size on. We're goingto make its point off this scatter plot proportional today Tips, size, feature or column. This way, we transform a scatter plot into a bubble plot. So it's point is referred to the size of people that went to Lance Dinner the total well off that meal on the tip they left. Now they're going to define three separate variables again. The total bill. The tip. We're goingto make a victor a slot as the total bill variable on By using the scatter function from that put lip, we're going to define our horizontal vertical axis on this size related to each point off this catapult we're going to define on Al fallible on the color. This way, we have a bubble bloat implemented with matter clip.

10 Multivariate Composition Plots

[Autogenerated] let's continue with composition parts. First, we have the stacked area. It is a composition off different, Serious added. It's one on top, the others. This way, it gives a graphical representation off the extent of contribution off each one. If related to time, we can analyze the proportional contribution evolution. This could be the number of for aggravated sails through in these years, off rock in the pop music, for example, if needed, we can represent it centered on the Ori simple access in case some reference value is important for us. Next we have this tack bars. It is similar to a stack area chart, but instead of serious, we represent the contribution of different classes we can represent again. The number of Federated sails off rock in the ample music, but instead off years, we can represent its relation with its age group. Finally, we have the tornado chart is a part that composition, visually represented in an orderly manner related to the numerical burial we went to study. This way we can compare the value related to two classes. I'm on all the classes, for example, here we have the month off hours off, slip off men and woman according to different months

11 Demo Multivariate Composition Plots

[Autogenerated] in this, the more we will focus on multivariate compositions That's using our well known libraries. Matt, Brute leap and sip on on. We're going also to set our tad schooler. We're going to use Seaborn, Matt Bradley, pandas and number. We're going to define some variables here, and we're going to use this textbook function from Matt. Let slip. This is going to represent as if we have three different variables, ABC. With the previously defined values, we can define a different baseline instead off in Ciro, a symmetric one. So you get something like this on there. So they're available, like waited, twinkle, and you get something like this. It all depends on the message you want to communicate. This way, you choose the most convenient baseline. Now we're going to use this in variables that we have defined. But we're going to put a bar put with matter that lip. But we're going to blow at it by parts. So it is a composition plot. The way in which we do this is by defining the vertical axis as separate parts. Did you see that Davis seemed to be overlapped? This is because we didn't define the bottom off its back lot. So if we do so you know, we have what? We pretended to have a stacked by plot with my split lip. We're going to load. Ah, well known tape status Head from Seaborn. We're going to set the tips given by men as a negative value. So we east our work in order to plot a tornado that we define our group as their \_\_\_ on the order off parse as the day with colors blue and current. You see that the ordered off parties for either Thursday, Saturday and Sunday. So our tornado chat is going to big this order as reference. We said our figures, eyes, we're going to relate colors on the group called That Is \_\_\_ feature from the tips later set. So this way we have male and female, represented with different colors on. We're going to use the seaborne about a plot function to put men and woman in a separate way, our horizontal axes going to be this heist or vertical axis today on our data, the one related to either men or woman. This is the reason for the four loop and we're going to set the order. As the days we find the ordered off bars on the color, as they call us, we'll find Blue undercurrent. We're going to label by a group Blood. The legend hopes some little back here, Let's fixed it. And here we have our tornado chart ordered by days. And so in mail and family information off our date, is it?

Feature Engineering and Feature Selection

01 Overview

[Autogenerated] welcome to these model about the basics off featuring in \_\_\_\_\_ and selection. They are a useful skill for continuing our E. T. A. On for preparing data for predictive models. You will learned how to extract features with the help of dimensionality reduction techniques. Understand what it's factor analysis as well as the most common cluster in techniques. Finally, we will perform feature selection and fit your engineering methods.

02 Extracting Features

[Autogenerated] to begin with, we can think of future extraction as an automatic construction off new features from road later, The most well known method is principal component analyses, or PC A SM. I thought for compressing data into something that captures the essence off the original data, trying to reel its internal structure. It is a linear transformation algorithm that represents compress original data into new space. This transformation finds directions off maximum variants that are motel you're so gonna this way is a technique that is useful to try to reduce dimensionality in case we have to deal with a lot of features. And this it can also help for feature selection and a separate step for data cluster in. Imagine. You have these data points that represent height versus wait. As we said, PC a finalist. The directions in which they ate at variance is maximum. So these are first principle component, as this is the direction off maximum Brian's of her data. Then it finds the second direction of maximum variance with the constraint of being also gonna toe the first principle component. This principle components can be considered as a new variables or features, since they are a linear combination off per weiss and correlated robe burials related to its principal component. We have what is called an egg in value that represent a proportion of the variance explained by the component. You compute the proportion to read some. In our case, the sam is 1.1, so you divide its again value by this total. Then, if you had these proportions, a second relative, some you can know how much variance is explained by the first components. For example, here, the 1st 2 components explained the 80% of barriers. How many components should we chose? It depends on the domain. If we pick all of them, we can reconstruct 100% off original later, but end with the same number of features as their you know later without achieving any dimensionality reduction. As a rule of thumb, you normally pick components until 95% off variance is explained, but in practice it depends from the problem. Domain P. C. A. Is sensitive to their relative scaling off the renal variables, so feature scaling or mean normalization are usually needed. Keep in mind that it only captures linear correlations, although there exist no linear adaptations. It explains the burying in data, and it's closely related to factor analogies. But Liz domain specific. Finally, although they're out of scope off this course, I just want to mention two of their methods that are commonly years as non linear techniques for dimensionality reduction. T distributed stochastic, neighbor embedding and self organized maps. Being the last based on neural networks SP, she ate both help in future structure on the separatist step for future selection or clustering.

03 Demo Principal Component Analysis

[Autogenerated] In this demo, you will learn how to perform a principle components analysis with python By using the packets sake it learn in this notebook, I'm going to show you how to compute principle components analysis Might using psychic learned libraries were going to import from people says in their standard scaler function I'm from so I could learn decomposition were going to import P c A. We're going to load. Ah, well known tips data set from Seaborn and we're going to put a scatter plot with equalized access. This means that both access are going toe ruby. Represent it under the same proportions. We can do this with metal function access, given my permit here, the world equal. So you see this predominant direction in which we expect to discover the first principle component. We're going to make an away with the total bill and tapes features from the tips. They test that I'm going to apply the standard scaler function so we can compute r p c A. This way, we are sure that the directions found by the PC A are not affected by the feature variances. Now we computer pc A. We choose to have two components. We execute the feet function and we're going to print the PC A components. This is the vectors that percent these pc A directions. They explained variants. Andi, explain various radio. So he had you see the coordinates off the principal components. They explained variance on the radio. This means that the first principle component explains around 84% off the variance on the 2nd 1 around 16%. Next, we're going to visualize this soup principle components over our data set. We put our scatter plot like before, but with the scale array we defined, we're going to pick the explain variance as the vector length on the components as the vector coordinates. We get the access current state on dhe. We're going to make annotations at these coordinates. I know in the PC. Amen. On the victor we just defined. We're going also to describe or define the properties for the arrows that are going to be plotted. We're going to state that the axis are equalized as before, so now we can see the directions that represent the two principal components. The larger is the 1st 1 on the shorter is a 2nd 1 As you can see, they are also canal

04 Factor Analysis

[Autogenerated] well. What is factor Analysis is a method to model or search of sir burials in terms of a smaller number of influential underlay in an observable factors or latent variables. The factors typically are viewed as broad concepts or ideas that may describe an absurd phenomenon. It might be associated with multiple of sir variables with common patterns. Off responses is similar to principal components as it explains a certain variance in the officer variables, although more elaborate and usually requires subjective judgments. Its aim. It's to extract maximum variance to rivers and geometrically the information in a data set in a low dimensional Euclidean space. Another goal is to help interpreting data identify in influential features and highlighting relations among observations. This way, the service burial way can be represented as a linear combination off parameters Beata multiplied by each factor. An example could be for music being observed variables, the rhythm, the amount of noise on the instruments being played on with that established the factor off music quality. It takes also some assumptions, for example, that there are no use in the date. Is it that the Raiders that size is rated on the number of factors valuables, not percent perfect multi culinary ity, and it doesn't require almost domesticity between the variables. Mainly, there are two types exploratory, which assumes any absurd variable disassociated with any factor on confirmatory with assumes. It's factor is associated with certain subset off of said variables. The steps for performing a factor Analyses are factor extraction Ingrid's. It uses variance partitioning methods on factor rotation that tries to touch foreign factors into uncorrelated factors. For better interpretation. S. M P c. A. In factor analyses. You also have a gin values on board, he said. In the number of factors to keep, it is commonly used what is called as creep lot, representing the pagan values in the sending order. Keeping those of off one or were the curve makes an elbow, which can be something subjective. So what are the difference between PC? A. Nonfactor? Analyses. Well, they might seem very similar, but indeed they are quite different. PC explains. Maximum amount of variance on factor analysis indeed, explains Kocharian's PC A components are orthogonal while in factor analysis is desired but not needed. PC is Helene a commission off off Sir burials and a factor analysis is a linear combination off unobserved variables. PCH is an interpreter. Ble a factor analysis is interpret able PCs observational and factor analysis is a model in technique.

05 Demo Factor Analysis

[Autogenerated] in the limo, you will perform a factory releases in python using packages Psychic learn on factor analyzer in this team. Or we're going to install the factor analyzer packet just to show that we don't only need to rely on the default packages off Python. So we're going to show how to use this one. Also, after the installation is finished, we import our libraries, including the factor a nicer one. We're going to load the deep state of said again from Seaborn, and we're going to do this factor analyses first. With this, I could learn like it will define the X ray with the total bill. Tip on size features As psychic learn documentation States Factor analysis performs a maximum likelihood estimate off their so called loading matrix. The transformation off the latent burials today observed ones. This is done with the factor analyses function. So we define the number of components or factors that we want to find. Remember that this number cannot be greater than the number of features. In our case, these maximum is going to be three. So let's try this number. We run the feet function and we can print they're loadings or wait that this function has given to its future. So as you can see the first factor as a load in off 8.8 on total Bill Sirrah 0.94 on the tip on 0.57 on this ice on factor number one has some other loadings or weights on as you see factor number two as Cyril Odin's. This might mean that Ah, third factor is needed so we might just keep the 1st 2 Let's do these same banalities with a factor allies her pocket. It enables to perform our but let's and Keyssar mayor all contests. But let's test off s free city text whether or not they have certain variables. Inter correlate using the officer correlation matrix against the identity matrix. If the test is found a statistically known significant, you should not employ a factor analysis. In our case, we see the \_\_ value is extremely close to zero. So we stayed that we can use under these tests a factor Analysis. The case of Mayor all King test measures the suitability off data for factor analyses. It determines the adequacy for it's absurd. Variable on for the complete model It also estimates the proportion of Barron's among all the absurd variables. Its values reigns between zero and one on a value less than 0.6 is considered another quick. In our case, we have 0.68 so we are quite close to 0.6, but we're still able to perform a factor analysis. Let's choose the number of factors with a factor. Analyzer class. We can create an object and run the analyze function without data frame on defining a maximum off three factors as factors are indexed from zero. These two represents three factors on rotation. None, because we're going to take the original data as it is, we compute the again values, and we can see that the again value for the first factor is very large comfort to the other two. This gets a hint about how many factors should we keep If we plot as creep lot, the rule of thumb tells has to keep on Lee. The wagon values greater than one, so in this case to be re grows, which would only keep the first factor. But instead we're going to decide to keep the three off them to see what Russell's do we have with the same faction is before the analyzed one and two seem to have three factors performing there. Barry Max Rotation. For more details, I recommend to read their packets in the communication. We can see the factor analysis loadings. So as you can check, these packets gives different values from the output from psychic learn packets. But loadings for factor three are quite small compared to the other two. So although the numerical results are different, the conceptual output is the same. It seems that we should keep the first or the 1st 2 factors as the 3rd 1 doesn't add too much information because it has very low loadings. You can think about loadings like kind of weights. We can also compute the barriers off its factor. The proportion off the explained variants on the community of explain barriers as you can see factors one, and to explain more or less the same amount of variance. Bad factor three explains almost none off the variance. So we should discuss this third factor

06 Clustering

[Autogenerated] welcome to clustering. It is a great exploration technique. 40 writing a multi very data set into Let's call them natural groups to explore whether previously undefined clusters exist in the data set. It is used when we believe that simple units come from unknown number off distinct sub populations. Its goal is to describe those populations with the observed data there exist here article methods in which we have argumentative, where bigger clusters are built by merging smaller ones on device, which takes the alternate approach. It begins with one big cluster and split it into smaller ones. No hierarchical methods in wits. A number of clusters is initially said based on the first guest or randomly and ate ice party in our accordingly. Then they the points are 80 relatively moved into different clusters until no rear segment gives an improvement. I finally there what is called model based that uses a mixture model to specified it, and satisfaction of variables for building clusters are go to village rely on what is called association or similarity between the observations. There exist several, and two of the most common ones are Euclidean distance and my head and distance. Having these two points, Euclidean distance will be computed as this, while my head and distance as this talking about here, article clustering will focus on the aglow merited approach most commonly used. Then the device approach Lemon ADF Argumentative methods is that the algorithm begins with small clusters and combines them until it reaches the desired number of clusters as a threshold. Here, article clustering gives what is called a dent program or tree diagram. It sold. The hierarchical relationship between observations and it's cold is to work out the best way to allocate those observations into clusters. You can see how many clusters you will have if you define certain distance. For example, if we cut at this distance, we will find four clusters. But although a dental gramps suggests the right number of clusters, you cannot tell how many there should be how that K means work. First, you decide a number of clusters and start by setting as many points called central. It's as clusters, usually randomly. You scan all observations and assigned them to the nearest cluster. Then you move the sentry, the center off its cluster on. Repeat this process, reassigning observations to clusters until no more reassignments are made. One advantage of these method is that we don't have to calculate the distance measures between Alpers off subjects. Therefore, these procedures seems much more efficient or practical. When you have a very large data sets for deciding the number of clusters, there are several methods. One of the most common is what is called the elbow method. You compute the clustering algorithm with different number of clusters and represent them against thematic called the Within cluster some off square, although depending on the case it can be and the walls you normally tools the number in which event happens in this case, four clusters will be the number.

07 Demo Clustering

[Autogenerated] in these dim or we will perform K means on hierarchical clustering techniques in python by using packages cycled, learned on site by. We're going to import \_\_\_\_ broadly, pandas and number on, of course, secular. We're going to import also Seipei, Spatial Distance and Psychic Learn Cluster. We're going to load the iris night, I said. We're going to separate it in features on target for simplicity. We are just going to keep the 1st 2 features and we're going to plot how they look like our caller is going to be the target. This is how it looks according toe CEPAL, length and supple with First, we're going to perform a K means clustering. We're going to initialize. I mean, this thing's least on a burial to set the maximum number of clusters. In our case, we're going to try 10. Then we're going to do for loop to test. How does he look like to have 123 tilting different clusters for these? We're going to use the function K means on defined the parameters number of clusters on the random state we run the feet function on. We're going to upend the mean distance off its observation to the cluster sent right. This will let us put our scree plot. This way we can see the mean distance depending on the number of clusters. On this way. Applying the so called elbow method to the right number of clusters. As you see here, it seems that between two and four is the right number of clusters show we're going to pick three. We run the k means function again with the selected clusters. In this case three. As we said, we were going to plot how this cluster in looks and compare it to the real data. As you can see, it does a pretty good job. It doesn't have 100% accuracy, but it makes a quite reasonable clustering. No go, go to perform. Ah, hierarchical clustering forties were going to import from sight by the pentagram on link it By executing the linkage function on our radar set, we can then visualize how these here article clustering looks without dental graham, we can establish the level of detail. For example, these people ometer refers to the number of observations and we are going also to put on a recent online at distance equal 10. As you see here, you have your dental graham on just by looking at it. You don't know how many clusters is the current one, but you can state a distance from which to select, which is the adequate number of clusters. If you don't want so much detail, you can change Pete to, for example, 50 and instead off 100 off survey Shin's you'll have asset off 50 Group off serve Asians.

08 Selecting Features

[Autogenerated] let's begin in selecting pictures as Peter nor Big One set more data bits, clever algorithms, but better lay that beats more data. The girls off selective features are identify important features, remove irrelevant and redundant features and improve interpret ability and predictive model performance. Future selection techniques mainly eliminates that so called course off dimensionality. It enables algorithms to train faster. It also reduces the complexity off a model and makes it easier to interpret. It also improves accuracy off a model if the rights upset \_\_\_\_\_\_ ocean and it reduces over fitting. This means that the algorithm performs better when dealing with unseen data. There are three main types off mothers We will forgive. Us unfilled their methods, which are no based on models, although there exists also with other methods, which are based on models and embedded methods that are based on models and tries to combine filter and rubber methods. Filter methods use evaluation criteria from the intrinsic connections between features to score a features upset. They are independent to the type of predictive model. On some of the available techniques are the stands, metrics, correlation, multiple information which quantifies in some way the amount of information about one feature through other feature cheese were, which is a statistical test for no negative features. The intuition behind is that features that they're in abundant to the target are uninformative in formation. Gain muscles. How much information officer gives about the class we want to predict variants threshold. This approach removes all features. Who's variants does not meet some kind of threshold? One thing that should be kept in mind is that filter My thoughts do not remove multiple immunity, so you must deal with nautical linearity as well. Before training models for your data weapon methods used the learning model as a black books to school shop sets of variables that go into the performance. Some \_\_\_\_. What are Step forward here to begin with one future drink. That's the model at Chandler Future and see if it improves performance. You're still doing this until you see no improvement. Step backwards the same US forward. But you begin with all the features on remove one at a time by directional Certs. It is the two previous approaches at the same time and stops one. Both find the same future combination. Breathe first shirts. It enumerates all possible future combinations and random shirt. First, you randomly generate a subset of features and then perform all the algorithms on that sub set off course. There are a lot more, but these are the main ones imbedded and makes the future \_\_\_ Antagonism is integrated as part of the learning algorithm, having its own built in future selection. For example, we can use victory importance in decision tree based algorithms as well as lasso and rich regularization techniques.

09 Demo Selecting Features

[Autogenerated] in this demo, we will perform some of the most common fleeter methods for selecting features. You seem packages psychic alone on side by. We begin by importing mad plot lip anger going to load the virus That said just us Ah reminder to better understand these data set. Although we have, you see it in the past. This is a picture identifying what is the petal on with this? The simple. So we load these data set its features. Are they separate length, simple with but the length on battle with In this same order, we're going to visualize them. And as you can see, we have four columns for the features on one for the target. We're going to try the variance threshold method from secular and we are going to import variance. There's whole function and we're going to try. Ah, threshold off 0.3. We ran the feet transformed function on the features on. We're going just to show the 1st 10 rows. And as you see, the second column has disappeared because these variants is lowered on 0.3, this function is typically used for dropping those values that are constant or half Ah, nearly Cyril Barry ins. Next, we're going to explore the correlated features were going to use from Cyprus. That's the person are correlation. Remember that this type of correlation just identifies linear relationships. We generate a random, normal distribution with zero mean on one standard deviation on we're going toe generate a 2nd 1 adding another normal distribution on dhe. Finally, 1/3 1 But this time we're going to add a normal distribution with different violence. If we plot the three variables in a scatter plot, we basically say that between X one and X two, there seems to be some kind of correlation while between X one and X ray, there is very low or almost none correlation. If we apply the piece on our function between X one and X Ooh, on also between X one On extreme, we can see that what we suspect it becomes really between X one and next to Theresa Sirrah 10.68 Correlation. Andi. It's P value is almost zero, so it's quite significantly, while between X one on extremely reserved, very low correlation on the P value is not statistically significant. If if we use the same technique with the Irish status that we can see that the first feature with the 2nd 1 has almost no correlation. But with the third and the fourth feature, he has, ah, large correlation. Now they're going to import from psychic learn. The mutual information function in this case is going to be most willing for classes because we're doing a classification problem. There is also the alternative off doing are regressing one if we apply it to the features on the target. This is capital X on way. We can see that the last two features contain a lot off information about the target, while the 1st 2 not so much. So this is a hint to think that the last two features are more important for predicting the target on ditty square test. These kind of test is going to tell us which features are or not informally about the target we import Select K Best anti \_\_\_\_\_\_. Yes, that's a reminder we're going to bring the first turn, rose off her features, and we are going to define a new variable. X knew I'm going to twos with this select K best function and sea squirt two feet the three most informative features for obtaining the target on. We're going to print again the first and rode off this new barrel. And as you see the second column has disappeared. This means that is the less informally about the target, according to the T Square thirst.

10 Engineering Features

[Autogenerated] featuring scenery is key and company find us the process of transforming later into features that bit represent the underlying problem. Toe the predictive models, resulting in improved model accuracy on and seen data. Is it also worth to mention and you nd words coming up with features is difficult. Tomkins Human requires expert. No bullets. Applied machine learning is basically feature engineer in. Ideally, you make it at the beginning, but you might have the knoll. It's to do so after you perform a complete a d. A. So there is no straight rule about when to do it. Feature in Unity is a representation problem, and you have to be a were that feature in January in on Future Selection are not mutually exclusive. They're kind off two sides of the same going the goats off engineering features are getting the most out of your data for predictive modeling on data interpretation, improver of demise. Predictive model results Find the best representation off your data to learn a solution to our problem. The benefits of doing these are flexibility, but your features might lead. You use legs complex models that are faster to run, easy to understand and to maintain simpler models. You don't need to work a skirt to pick the best models, and most optimized parameters better results with what features. You are closer to the underlying problem on best representation off your data, they suggested by blame for performing feature Indian you're in that you can find on the Internet follows this sequence. First, your enstrom orders the features, then you decide which features to create. Next, you create those features, and you take how the features work with your model. And then you have to look if on improvement off the features is needed and go back to the beginning to curate more features until the work is done, some off The high level techniques are imputation, dealing with missing values, handling a pliers, something really know. We should decide whether to keep or remove them by name. It is related to grouping later into classes. Look, transform. It helps handle eh Skewed data. One holding coding. It transformed category. Gil data into Callen flacks off zeros and ones group in operations for aggregating information about observations, feature split, for example, a complete name string to a separate column for name on other for surname. Scaling is for mating data under a common range. Extracting date Mrs Wright to format in date in a more useful way. Unfettered construction. This is kind of an art, and it involves injecting domain knowledge.

11 Demo Engineering Features

[Autogenerated] performed some of the most common methods for engineering features by using packages, pandas numb by psychic learn and date them as usual. We begin by importing our libraries number. I am bonus. We're going to load the diabetes data set from psychic Learn, these data said, is already a curated, but we're going to use it as a baseline for introducing some non value dropping filtering. You will defining as a data frame with the future names as columns on their target variable. This is how it looks. You have AIDS, \_\_\_, body mass index, etcetera. We're going to check how many values are none values. Listen, it seems that there's no one. There are Siro, so we're goingto introduce when at Row three, with Numb by who introduced it. Now, if you compute, they mean off. The non values is not zero anymore. So going to demonstrate how imputation is made as an example, we take that there exists a column with Nan values. Mrs eight as well, really knew. On first we're going to define athletes. Hold on, give a demonstration about how to drop missing values so we filter by the columns with missing values under their defined First. Hold on. As you see the eights, Callum has disappeared. This is normally done when some cologne has a lot number off missing values, we can ah drop off rose instead of columns. The filter is quite similar, but you have to define the access. In this case, access equals one. How does he look? You see, we keep the cologne eight, but roll three has disappeared. This is long when you need to basins without off missing values. Another technique is imputing these values. The easiest one is to feel those with zero off course. We need to think if for our specific problem this has sense. If not, Nick's like feeling missing values with medians off the columns with the field and a function. We can do this on dhe. This is how it looks. That's the median off the column. Second learn has also a simple in pewter function. You can specific, amiss in value on with strategy to apply impute by medium by mean is very similar as defining our model in psychic learn because you run ah fit function. We're going to take What's the eight feature mean? The value is imputed as you see the meaning is 0.2 on That's the same value imputed replacing the non value next something you'll really know. It's about handling out players. Here we have two police thunder deviation policy in which we take them in and we compute the mean plus three times the standard deviation on my nose three times the standard deviation. Any value outside these limits is considered an outlier, depending on the problem, if we need to keep them, if they are important for getting insight or if they are just nice and we should get rid of them after defining these limits, we just compute this condition in case we want to drop them as you see the third row. So that observation was categorized as another player. Then we have percentile day aviation policy. Remember that we need to compute the Quintiles, their first under third as upper and lower limits. Sir. Inter quarter range Andi, as before, we compute the condition for dropping these off players again. If this drop has sense, if not, we should keep them again. As you can see that the probe has been dropped, been in fixing data within certain group ranges. This technique helps preventing over fitting but its sacrifices performance, so use it with caution for these. We're going to use pandas function, cut this faction, let different limits for the Rangers we want to create This way we transform continues data in tow, categorical one like low, mid, low, mid, high high values that don't fall within these categories are assigned next. Look, transform What is hideous? Well, for it helps to handle skewed data. But the original data that you applied look transform to must have only positive values. If not, there's a little trick with his one. When you perform this transformation, for example, we're going to define on exponential distribution. We're going to put its history, Graham, and it looks like this. Now we're going to perform. They look transform Andi, as we said, we're going to. At one. There is some negative values. This is not the case here, but we apply it for the demonstration. Now, if we plot their transformed in data heist a gram, it looks somewhat normal. It has a cat at zero, but as you see, it's similar to a normal shape. Next is one holding coding. What is it useful for? Is useful for mapping categorical data to ones and zeros. For example, here we have several teams that belonged to several cities the number of games, and we also have the M V P player from It's Tim. If we perform a one holding going by using, they get them its function and applying it to the city column. Where we get is this. This may be useful for algorithms that need this time off. Input for being able to train there exists also group operations with the group. By function, for example, we can go by city, the number of games and tourism. So this way we have that 15 games were played in Hong Kong 41 the same for room. We can also apply toe all the valuable like the team. So one team is from Helsinki to from Hong Kong on three from Rome. Now get going toe load the from sea borne on as well, really know perform at people to transformation This way. We changed the data frame toe a matrix in which we have my rows and columns. The year on the month on the content off the matrix is the number of passengers that matter and roll. We also have feature split. This is normally useful when we have strange features like names off people, cities we can define a spirit marker on with the map function. We're going to apply it to every data frame. This way we're going to pick on Lee the name on Do not certainly like this Scaling is another operation we can do is useful force home algorithms in order to be trained and can ease also are comprehends, You know, Miss Wood Time data, for example. We're going to perform normalization with this I could learn Mean Max Caylor function. We make up send data. This example is taken from the psychic learn documentation We apply these mean max to these days. As you see it fits it toe are serial one range. Then we have the standardization. This transformation fits the later into a normal distribution. Another thing we can do is struck date Date is especially as you have to refine its format. For example, here we define at a different with these strain dates and now we're going to transform them to ah, daytime format with a day, month and year order. This is how it looks. This makes our work is here because now we can just access its field. For example, if you want to know that year, we can do this with ditty that year. We can do the same with month on with Dae So having a date, good idea to ease our work. You shall have all the valuables like they they name in this case Sunday on Saturday.

Presenting Your EDA to Others

01 Overview

[Autogenerated] welcome to presenting your e t a two oldest. In this model, you will become our off the deuce on don't in data visualization, you will also think about how to target your audience and learn how to structure your idiot report.

02 Dos And Donts

[Autogenerated] Let's talk about the deuce and loans. According to Alberto Cairo, a great expert on data visualization. A general rule we should follow is to show on don't tell, but he also gives some advice. These rules might be misguided if taking to literally on at face value. Remember that E. T. A. Should enable exploration, analysis, discovery and communication aside. \_\_\_\_\_\_\_ advice. Be truthful on always based on real data. This seems quite obvious. Medusa. Keep an eye on this because you might be lying without even knowing me. Accurate. Avoid, um, be witty East interpretation to your audience. Don't make them work. Trained to decipher a tart or some computation finally be elegant and aesthetically pleasing. This is not about doing pretty chats. It's about better understanding how not to lie. Don't true, Kate. If you took read your chat, you will communicate a very different message by manipulating people's perception in this first graf. Carby since March. Cheaper than a But if you don't truncate the tart, the real difference so sharp, better and you can see that they are not so different in price. By using a run scale, this heap of her plot can dramatically taint, and that's the message that you communicate. Imagine this is the increase of wealth in the world from years 1900 to 2000 as the horizontal axis is not correctly scaled, it seems that wealth has increased more rapidly between 1920 on 2000. But this is our share price when we apply the correct scale that the message is exactly the opposite about proportions. Some projections, unless it is absolutely necessary. Police. The new Spite shirts don't use three D graphs because humans are not made to interpret areas accurately. Think, for example, in optical illusions. They are based on this fact and never, ever used double access because they give the false message off, comparing two graphs that can be compared because they are expressed on different scales. A very important thing is context. If you look at these two graphs from 2014 you might think that Ukraine was a divided country. People who spoke to Russian booted one political party on people who don't vote it the other political party. But indeed, these charts are lying because they don't splitting the diversity of its territory. They just simplify them assets by stating on all or nothing logic, while in reality there exists a lot of grace in between. And added to this, they don't show any relation between Russian speakers on the party to which they gave their vote, too. This is a well known picture about bullet impacts on planes that returned safely from war, so it was going to be decided to reinforce the parts with more impacts. But likely someone notice here there was a missing information by us. Wouldn't it be better to analyze those planes that actually didn't return to discover on dreamforce this sensitive areas? What about shooting color? As a rule of thumb, you can think that the less information, the better, but without loosen, explain ability. Albert using too many different colors and Seitz, if not necessary, They maybe start on distract you to information on sensory ALS saturation. Furthermore, different colors in different cultures, half different meaning, so you might even in use some kind of pre interpretation to your audience. The advice is to use neutral colors unless your mess it's needs to spot something out. Remember that the goal is always to support the street. You are communicating. You must also consider Cole blindness. As you probably know, there exists three types off color blindness red, green, blue, yellow on complete. So here, in order to be respectful, used color palettes adapted for colorblindness also rely on different shapes. If possible. There are several tools for checking your visualizations. One of name is Color Oracle and Opens or so on. Five. Fox has also a plug in for these as a final remark. After being aware awful Dave off What to choose. How to show It all depends on the methods to communicate. You might try different plots and techniques until you're capable of communicating your message the right way by being truthful, accurate, elegant and in an easy manner.

03 Demo Color Blind Test

[Autogenerated] This is going to be a quick demo in which you will learn how to perform a color blind test by using the open source software color Oracle First, we're going to go to Ko Oracle website on now. Loaded. The installation is extremely simple. Do yes. Have to download extract on double click. Once we have it installed, we can check how it works with an example like this. Do yes have to press on dhe selected type off color blindness. You want to check? So here we see the Otay Topia Proton topia treating Topia on Grace Girl. If with any of thes visual transformation, there are two colors that are not distinctive A ll Then you should change your visualization to be respectful with color blind people.

04 Targeting Your Audience

[Autogenerated] pore targeting your audience Remembered are communicating is all about understanding. Like in all the life situations, there are high level general rules that are suggested to follow. Although it's situation, it's a specific approach on different needs to be fulfilled as there doesn't seem to exist. Ah, well documented procedure for days. I will get these video based on my own experience as a general idea. Although there are plenty most of the cases you will culture three types off audience technical people. Ideally, they're willing to be able to replicate it, so they want to know how it's made while you decide these for that business people, they are obsessed with utility disease. Hole is good for business. How much impact enabled This is him making on then sales people. They want to know how it is useful for my customer. How can I ever think my offering on attract new customers off course in real life? The recent overlap and independently off the profile? Everyone should be interested in answering these questions. What about the technical audience? Well, my recommendations are confirmed. Ideas disagree before you ugly search for weaknesses in reasonings and exploit until the truth comes out. Ask for wise. Don't give anything for granted. Explain every detail when possible. This has your data sources that acknowledges you used how you just them that performance, the conclusions you made on the arguments analyses that support them. A straight truthful argue based on data and be humble. It's not a knowledge competition. It's about obtaining the best outcome as a team business audience. If you need to explain something that is high dimensional, explain alot dimension equivalent as introduction to speak in a simple, unclear language to make sure everyone understands but don't oversimplify, avoid wrong conclusions. Loan. Let anyone conclude anything that is not supported by your A D A. Especially when it comes from someone with authority on. People are afraid to tell him or her about being run. Prepare counter fax forties estimate. Business impact off conclusions were impossible on \_\_\_\_\_\_. Respectful. It is not about being the smartest in the room is about making everyone the smartest in the room. Most of the advice for communicating to business people apply, but furthermore, go right to the point. Be time efficient saves. People need to have clear the main ideas that will support the spits. These doesn't mean just explain just two conclusions, but you should not entertain in details, show capabilities, what can be don't or decided with the insight. So specific examples, if possible off how inside was applied on previous similar problems forced her counter facts. These will help to prepare people to feel argumentation is with a customer or potential customer on be clear, focus on the main message that everyone should remember after the meeting.

05 Structuring Your Report

[Autogenerated] for structuring your report. You should first set a hypothesis and explain its meaning and the reasoning behind. Then you take into account the initial conditions and constraints. It's all about understanding the problem you face. Next. You must develop your reasoning unjustified that exploratory decisions you make. You should also mention there things you discovered while performing your e t a and explain. How did they impact or change the initial approach with a level you need to argument the inside. This is the knoll. It's acquired thanks to the perform. Edie A. Finally, you graduate conclusions by taking their initial hypothesis and so impossible applications on impact off the generated no bullets. And remember the kiss principle. Keep it simple and \_\_\_\_\_\_. Become size and accurate. Don't add extra information if it is not needed. Join explanation Blocks with Weiss and house on Dhe. Be clear. Don't be Abby was

Practicing Data Analysis with Python

01 Overview

[Autogenerated] Welcome to these model about practicing respiratory data analyses with python. In this model we will practice the choir area in all its toe. A real problem. These real problem is going to be a past Kagel competition as it is a complex problem and in that explanation will take too long. This is why this is going to be a brief introduction to Dady is related to course content to get the initial steps toe What's inside? Certs? Andi as it is a competition. I encouraged you to learn from others. Look at the Colonel's on Dhe Sir, your ideas.

02 Understand The Problem

[Autogenerated] as always, the first thing to do when trying to perform an e t a is to understand very well the problem. We're going to use the NFL Punt Analytics competition from Kagel. Andi, if you sharp the Web, it tells you that the talent is to reduce the occurrence of concussions by proportion, Specific role modifications on DAT. The evaluation is going to take into account the efficacy and the game integrity. Great. This this man Well, for efficacy, they tell that they need you to demonstrate understanding that the analysts his do you exposed is easy to understand on that your analyses can be reproduced and for game integrity. What they mean is that the proposal needs to be actionable. This is it can be applied on DDE that it doesn't introduce new risks to playerssafety. So this is the goal on We can think off it as the main hypothesis that by modifying rolls off the game, we can reduce the occurance off concussions during point place. If you're like me that don't know what it's all about, the NFL, maybe you're asking what's upon play well upon play looks like this. It involves the \_\_\_\_\_\_\_ team lining up at the line off what is called scrimmage with the kicker they're receiving team lame, zapped with one or two players and feel to catch the ball. The player with the roll off center makes Alan snap to the kicker, who then drops the ball on. Dhe kicks it before it hits the ground. The player who got says the ball is then entitled to attempt to advance the ball. Although we have a specific goal that the organizer's off, the Talyn's have explained, we can set also some initial hypothesis and try to demonstrate them why we perform our e t A. For example, we can think that faster players are more prone to concussions or shorter players. We can also think that concussion succour at the end of games because players might be more motivated. Or maybe we can think that there is no need to change the rules of the game. These on any of the hypotheses we can think off should be supported by our 88 Surfing the Web off. Kagel again see the talent description on data. The data has had quite complete description, with separate files on dhe. The description off its feature on even they are fights with narrative of its play. You have also be reviews is our very complete data set behind you can click any file on DSI. The column names on their data types. Also, you have a quick view off some idiot graphs, like buy plots or hissed a grams. This is quite useful. To start in the area is to fully understand these data to perform your idiot.

03 Join The Competition

[Autogenerated] for practicing our ta No, let's We're going toe. Join the NFL Cos Whom from Kagel. After joining you can click on kernels you can choose which in script on notebook. We're going to choose a notebook. Under it is our new colonel just created.

04 Taking A Look At The Data

[Autogenerated] the nice teen off using a colonel is that it already has some libraries installed on. Do you have access to the data? For example, if we bring the input derogatory content, you see all the files off the data set. We bought our month long sleep on Seaborn libraries on DDE. At first, we're going to take a look to the later this is the BDO review data set, usually the season years again. Key. Also, the idea for the player the activity arrive. Who was the primary impact type? Wasit. Ah, Helmet to helmet. Was it friendly fire? This is due to our colleague due to our team partner. You also have the game data. So here we have the season year disease. Some type other in for, like, the game surfers. Was it grass was hit, artificial or turf? How was the weather like doing the game the start time to these later said, We're going to add a column being true off us if there was a concussion or not. So we filter by the \_\_\_\_\_, and as you can see, we successfully at it. This new column. You also have data about the player role again we're going to attack along with True Father's about If a concussion happened or not, is we have this issue in the game key. They play I d. The roll off certain player on dhe. If there was a concussion or not in the play, we also have the play in four. Again. We're going to add the concussion column. You have information about the game date, the week, the game clock, the just line, the quarter, etcetera. You have the play description. I'm from it. We're going to extract the distance disease, the number of charts So we do a stream split. We define that we're going to shirt for information. After the punts, work would find a new column called Bunt, this time from the speed it string we're going to see tracked. The number of guards on those missing values are going to be filled with zero on. Dwyer went to transform it toe a numeric feature. As you see, we have here 52 52 51 51. It makes sense seems to eat on correctly. Next, we're going to load some off the regular system data sets from weeks 1 to 17 we opened the three of them on with They can cut function from pandas were goingto make these 31 data frame congratulating one after another. The first time you do this might take several minutes. We're going also to what the concussion feature. As in the other data sets. This is how it looks. So we have the coordinates off its play on it's game. And if there was a concussion or not, I don't know what an event is. So I just do a set off this to take the unique variables in which there are also missing values like you see here with an end on Okay, they are every kind of action that can be performed in a game like Gran Shift, etcetera.

05 Practicing Univariate Analysis

[Autogenerated] Now they're going to perform our unit. Very analysis. This is a complex problem on dhe. I recommend you to take several days or even weeks just by performance and little transformations and trying new things and new ideas. But for the sake of demonstration, we're going to use some off the possible techniques to give a general idea as an advantage. You can also look to other people colonels and get some crazy or creative ideas. This way, you can also learn from them. Okay, I'm curious about how they kicked the instance. Distribution is, as you see, is quite similar to a normal distribution, although what 40 yards there seems to be a pick on. Maybe it's a clue to do some research on that value. We can also think about what's the number of concussions per quarter. Maybe players are more motivated to what's the end of the game and more concussions happen, Who knows? So we're going to do a bar plot, Andi, as we are just beginning to ado Annie Barrett analyses, we pick their place in which a concussion happened. So it seems that in quarters to 13 most off the concussions happen but it's too early to make any conclusion. We need to compare this with the total amount off place, but this is work for the multivariate analysis step. What about the start time? Can it be influential in the number of concussions that happen? Well, let's speak the game. Start our by doing a stream split off the start time feature. You know, the to escape the hour. In this case, we don't want the minutes through here. We have to start time on. This is how it looks. Our transformation with the game start our So we brought again about put as you see, these data sets half a lot off categorical data. So a good decision for representing data unplugging it are Bart. It's on any comparison, plot or even composition. It depends on what you're searching for. So we brought our bar put on here. We see that at land on dinner times. There seems to be the most number off concussions. But again, this is just by picking those place in which a concussion happen, we need to compare it with the whole number off games played at that hour. But it might be something to look at as we are. Human beings don't perform the same at different hours off the day. Maybe if I play in a different hours, we reduce the number of concussions that happen. We are also clear about what's the impact type on activities. Really an impact Again, we can see this with a simple bar. Put picking the primary impact. Type off the video review data set we're going to do to support Matrix. And as you can see, it seems that helmet to body and helmet to helmet on. Also tackling on blocked are the primary impact type on Bree activities arrived in which most of concussions happen on what about they gain surface. Maybe some surfaces make players more prone to slip on. DDE have concussions. I'm going to pick the unique values for the turf surface, as you can see there at this one's artificial grass filter there even misspelled ones. Or you can find natural grass retaining different forms. For example, we can customize some kind of transformation toe aggregate these different surface names, for example, everything that has grassed. I'm going toe replace hit with grass so natural grass grass alone are going to be transformed into just grass. We define a dictionary with these transformations on a loop for taking each observation tower of value on according toe some conditions. If the stream is within this value, I'm going to change it to the value they find at the Rep Dictionary. Show everything, for example, every observation that contains net within its dream. It's going to be transformed to grass if they're missing values. I'm goingto keep them as missing values with numb by that man. We know that the length is \_\_\_ We're going to check it by computing the length off the clean tar for variable on it seems right. Okay, so I'm going to what, these clean, valuable as an extra feature. I could substitute Terry No. One, but for the moment, I prefer to know to do so with row of our float and we see that it seems that when the surface off the game is grass or turf, more concussions occur. But again, this is not enough. We need to see the proportion off the games plate in grass turf, et cetera, to get an idea. If these values represent significant differences or not within the regular system, Data said. We have the coordinates off its play so we can ask ourselves, Were concussions happen for replacing these disillusion we can use at first? For example, books plot and see how it looks. We pick only those observations in which a concussion happens, and this is how it looks. It doesn't seem to tell us, Mats, although we see here some kindof outliers, maybe it's better idea to try with our bio in plot to check the distribution more in detail. As you see it. A is not a thought plan. You test. What do you think? It's going to be the best technique, But if you see it doesn't give you the insect do set for, you have to sweet and training things. In this case, the violin plot gives us a bit more information. It seems like there are two kinds off stones where the confessor's happened the most. I am also interested in knowing which player role suffers the most. This is which one has the most. Number of concussions may be attentive, and rolls could partially apply to this kind of players were going to import score if I as we're goingto visualize these as a tree map on dhe. We're going to replace their true false values with one and zero so we can aggregate them. We're going to group by the roll off the player defining as index equals falls. So instead of her serious, we maintain a panda state. A frame format for the variable and we're going to do are some over the concussion feature after grouping by Roll Girl. Also going to short these values in the standing order. This way, we said the ascending burial as false and this is how it looks. We have to roll on the number of concussions. Parole. Now we find our color map. I'm not going to give a lot of detail here because it's not the important thing, as you can check it later. But we did find this color map for putting our trim up with score. If I'd that plot being the size the another of concussions role on the label off course, the role we're going to set an Alfa off syrup on five Onda Scholar. The previously defined Carlos map. We're not going to show the Xs on this is how it looks, the darker on the bigger the more number of concussions that a player role suffers

06 Practicing Multivariate Analysis

[Autogenerated] Now let's begin multivariate analysis. First, we're going to check the proportion off concussions during quarters. So we're going to do about plot for each quarter being the hue, the concussion. And, as you see second on third quarters are the one that half the most. But at the same time, there are a lot of place that don't have a concussion. So we need to look at this regarding the proportion. For example, it looks like the proportion in the third quarter is the highest. So maybe we should look for what is different in this third quarter. Maybe it's their players attitude. They are more motivated. Who knows when it to continue attracting more knowledge on dhe. How many concussions were due to friendly fire were there, then also compare on cities with our bar plot. And as you see most of them are not friendly fire. But when tackling, I'm being blocked. Some concussions do tow team partners happen, so maybe that's another thing to look at. More in detail. On looking again at the start, our we see that a lot of games begin at lunchtime show the proportion off games with concussion against those that don't have any concussion is very small, the proportion off those that begin at 14 new clock. Maybe we should take a look at them on Dhe. Also, for those that begin at 19 in the evening, we're going also to look about the proportion of games were guarding its surfer style if they have grass turf, etcetera. So he had to see that the games with grass are most off them, so the proportion is quite similar to the others, at least visually. Let's look at it numerically, so we find the surface proportion by grouping by the client off variable we made before Andi concussion on, We're going to count them regarding the game key. This is the different games, so this is how it looks. You have the games with no concussions and the games with concussions. With these later, we're going to compute the proportion again by doing a group by over these results on a plane, eat to every group variable and as you can see the game seeing grass half the higher proportion off concussion. But it's quite near to the others, so it doesn't seem that the surface has a clear impact as we have the coordinates for its came on its play. Maybe it's interesting to plot. Ah, heat map off. How many concussions happen at its part off the field. So, for example, for our ADA said, Well, yes, pick the values off the game keys that have concussions were going to pick one for performing this example. Therefore, 73 And we're going to pick Yes, the columns off the coordinates on if there were, if there happened a concussion or not, as the coordinates are float variables with the symbols. We are just interested in the yard by the art show. We're going to transform them. Tow \_\_\_\_\_\_\_\_\_\_\_ on dhe. Also as a categorical, valuable. We're going to duties with X and weigh on. This is how it looks on for making our aggregation easier. We're going to replace their true falls off the concussion column by ones and zeroes this way by making a group by the X and Y coordinates as index equals falls to keep the data foreign format, we do over the concussion column as some operation. So this is how it looks. We have the number of concussions that happen at it field coordinate going toe perform people to transformation. So we have this matrix format. We're going to feel the missing values has zero. This is how it looks. So now we're going to use this. He warned Hitman function on here we have something like the field on with bright color the points off the field in which most off the concussions happen. But as you see when between Bert the Y access to have our realistic view, this is the origins off X and way should match. So here you see the number of concussions that happened at its point off the game. You can clearly see that there are certain areas in which most off the concussions happen. It would be interesting to compare if these areas are the same among all the games. Remember that this is just the example for one game

07 Report Examples

[Autogenerated] Now we're going to look at some examples off how good report looks like. As you see, this is very well ordered. One on it looks at several points off the game. It has an introduction before we start explaining the problem. What is the initial hypothesis? And then it performs the date exploration on what is nice about this report is that it has what we learned sanction. Describing some concepts like player desperation, they use a helmet deceleration time on. It also proposes some rule changes based on the exploratory data. Analysis on all their nice One is this one. It gives an introduction in waits. It describes the problem on, also introduces the proposed changes and quantifies the potential reduction off concussions. This also a very good way to start a report that he has the methodology, the analyses, the concussion. It also has a background section, and that helps to explain what he's going to introduce at the e. T. A. As you can see in the methodology section, it Ray's tow explore the problem by understanding the situation on the problem at first, then identifying the correlations and cautious off concussion by performing numerical analysis and in this case, machine learning the competition didn't ask for performing machine learning algorithms, but it could be also a good approach. And with their assaults off these numerical analysis and machine learning, he develops some proposed changes. As you can see, the main 80 years behind about how to structure to report, as seen in these two examples but with different details on approach is that help to do so? Like, for example, these logic tree, then it tries to identify based on the evidence off the game on has also references to papers. It has some nice visualizations like this on these are the one. This is Han Sang Ki diagram, in which he relates the concussion rolls on events or activities in the game. If you can, animations might be quite useful, like in this case, you can see the movements off its player. Finally, they can classrooms and some extra later, a SZ appendix. As a final remark, I've encouraged you to explore other catalyst reports on dhe. Learn different approaches about how to communicate your data to your audience