hello everyone hello everyone uh let's wait for few more minutes just two more minutes and then we'll quickly get started before that uh watch out my channel intro [Music] okay so uh welcome guys to the stream as you all know that today's class is going to be on exploratory data analysis uh again this is a continuation class uh this was basically exploratory data analysis has nothing to do with power bi or python or any of these tools it's just a normal statistical concept okay and as you all are attending my python sorry my power bi classes uh we have already finished three classes as you all know and uh just before the live dashboard there was a topic called as eda and i i just didn't wanted to skip that part and because of that and already we had announced the four days classes right so i couldn't fit ada in the program because of time constraints and because of that we are conducting this session so in this class i will be talking about eda we'll be talking about uh exploratory data analysis what exactly does this mean and uh we'll also be doing some python analysis like eda in python and will also be side by side we'll also be doing eda in power being see the whole idea about this session is uh just understanding the concepts of eda what exactly is eda that's that's more important how you do it is not that important because some people are confident on python some people are confident on you know power bi some people and confident on other tools so you can just do it whenever you want to that's that's not a problem okay so give me a second probably i'll try to switch my uh presentation to another file just give me one second everybody is able to hear me right do you plan to teach data warehousing concepts no data warehousing concepts are not required for data analysts they are required for data warehouse experts so i'll i don't cover that part uh give me ah yeah i got it i got it any other questions before we get started i'm just trying to get that presentation and i'll quickly share because i was earlier presenting uh this uh slides on google slide and on google slide i cannot do the annotation so that was a problem so i'm just switching back to uh powerpoint and should be yes everything is all right let me know if you are able to see my screen just in two seconds one two three you are able to see so great so we'll quickly get started oh sorry this presentation mode i will just move on to the whole screen okay everything looks fine let's get started 20 people and uh let's just enjoy the session everything is all right live session on exploratory data analysis spelling mistakes no spelling mistakes right else people will tell me oh you doesn't you don't know the spelling of this word everything is all right okay good thank you so welcome back we'll be talking about exploratory data analysis now before getting into the eda part we will try to understand what is eda which tools do we use to perform eta because eda is by far the most important concept for a data analyst for data scientist as well and trust me if you watch this particular session end to end there are two more ada lectures on my session on my youtube channel as well that also you can follow so nothing to worry about this is nothing but a repetitive session but uh it is fun to watch a session live right rather than watching it in 2x mode anyways so what is ed eda deals with the process of performing initial investigations on data with the help of summary statistics and graphical representations in simple terms eda see eda is just a fancy term and statisticians like to call it eda in layman's term it is nothing but data analysis in layman's term it is nothing but inside gathering we have to find the information from the data that's it similarly there are many more fancier words as well in statistics statisticians like to call that way if you just don't call that word they will be offended another example is average one of my previous company where i joined another company and not named the company there my manager asked me what is the difference between mean and average and i was like what the hell the both are same then he was like no mean is the statistical definition of average and i was like seriously i mean all you understand is they both do the same work right so anyways so fun let's keep the fun side apart now these are the four different items or four different sub categories of topics that we will be studying that eda is all about discovering patterns spotting anomalies what is anomaly and then testing hypothesis and checking assumptions what is anomaly so many topics will be covered in this video it will be short one hour or max two hours i don't know i don't know the exact time that we'll take for this session i am planning for one hour because i have some other work to do after this but if it gets extended to two hours i cannot help so anomalies are nothing but simple terms those data points that are far away from the other data points in your data set abnormalities yes somebody responded in the chat box okay outliers let's say you are analyzing a data something like this these points probably are some anomalies that are easily detected by a human eye why this is an anomaly because this is far away from the rest of the data points but there are many more anomalies in data that is not identifiable by human eyes okay so in that case we have to perform some anomaly detection techniques to identify those anomalies what do we do with anomalies is a separate story for eda purpose we don't get rid of anomalies after eda when you get into machine learning when you are working on multiple algorithms in those scenarios anomalies will be very much important because there are certain algorithms that are able to deal with anomalies and there are certain algorithms that don't deal with anomalies in that case you have to get rid of anomalies and then do your predictive analytics part okay so these are the four different steps in eda outline of performing eda why are we performing ada see try to understand this question that people who are going to join as a data analyst some other day you will have to perform ebay some or the other day guaranteed so whenever you get a task of performing eda on a specific data always these are the few questions that should always come to your mind what questions are you trying to solve what is the business problem what are we trying to solve if your manager simply gives you hey this is the data and oid like what the hell is that what is the data what is the use case you have to ask him what is the problem that we are trying to solve okay either you ask him or try to understand it by yourself maybe if you have domain knowledge you will be able to understand oh this is the customer data probably we are doing this this and this right so if you don't understand the problem statement ask what kind of data do you have and how do you treat different types of data what's missing from the data and how to deal with it this is one of the most important topics in ada how to handle missing values okay where are the outliers where are the outliers and why should you care about them outliers are very important because they are far away from the rest of the data points so identifying an outlier and dealing with it is very important how can you add change or remove features what does this mean now just imagine you have customer id you have age you have tenure you have gender you have geography let's say you want to add a new column called as age bins now this is nothing but your addition of a feature that means you are creating a new feature let's say customer id is not important so you remove those features right so how can you add remove and change features to get more of your data this is your overall outline of performing ada okay steps involved in ed these are the major steps that are involved in ed the first step is always data sourcing okay so data sourcing i i usually give the same story that data sourcing depends from companies to company now if i distribute or divide the companies in three categories let's say one is mnc another one is a mid-level company mid-level company another one is a startup now in mncs or definitely they have been capturing data since quite a long time right mncs means multinational companies they might be having data a lot of data so in that case you have to ask a question that am i going to use our existing data or are we doing some research on some public data so that once you answer that question then you will be able to understand from where do you get the data from most likely they might already be having a lot of data in their databases so you have to use that same with mid-level companies as well now coming on to startups now startups has two types of environment one startup is having a client and another startup is very new they don't have a client now again if you have a client that means you should always ask data to the client again in in this scenario your client could also tell oh i don't have data can we just work on some public data sets or can you just try to scrape data from online or from some website or from some some of my competitors website in that case as a data analyst you should always be focused on the data scripting part as well so data scraping is also important for data analyst it's not that simple you just learn power bi and find a job it's not like that right right now with the current trend in the market so many people are entering the field of data analytics data science so you always have to up skill okay so if your client has data take his data if not take the public data set and in this scenario where you don't have client it's always public data set so in these two scenarios you always need data scraping concepts of data scripting knowledge as well so data sourcing is the first topic first most important topic in ada after you have got the data let's say this is your data that you have got could be from a database could be a flat file could be anything now what now the next step is data cleaning inside cleaning data there are lot more topics handling missing values handling invalid data type handling any issues with the mic i'm already using a i'm using a razer mic others can can we can you tell me is the voice not clear maybe hers you should uh join back the stream could be an issue with your end because i am already using razer mic which costs

8 000 rupees okay just just kidding let me don't take it seriously okay just rejoin the stream it will be okay so data cleaning is very important handling missing values handling invalid data types getting rid of outliers and lot of topics are going to be studied in this part okay after your data cleaning is done that's where your final data comes into picture final data means your transform data people also call it as transform data you call it as raw data raw data okay now again there is a concept here people also are very confused data analysts are being asked about etl and elt and all those things what does that mean etl simply means extract transform and load elt means extract load and transform again these are two different topics that you might need to know about but not in depth okay so i will not cover etl elt all these things in this class else it will be a punch of everything so we will not cover that now just imagine from raw data you are converting into a transform data and this transform data is ready for your analysis okay now now you have your transform data in your data our first task is to identify what are the categorical features and what are the numerical features once you know the categorical features there are multiple analysis that you need to do so categorical analysis includes univariate analysis univariate write down the names i will explain you later bivariate analysis and multivariate analysis multivariate these are the three different analysis that we do in categorical features in numerical features we do correlation we plot some scatter plots and check some values we plot some dist plots to check how the you know distribution is among these numerical objects and a lot of things after that there are certain scenarios where we also derive metrics derived matrix basically means if you want to yeah we can cover etl some other day i will have another live session on 18 maybe next week maybe next week wednesday or something we can figure this out okay so derive matrix is nothing but you are extracting some features from your existing features that's it okay so this is all about ed now even if you don't memorize it you will still be able to remember i mean remember all these things right first is data sourcing getting the data second is data cleaning third is your analysis and fourth is your derived matrix inside analysis there will be categorical analysis there will be numerical analysis okay now comes the data cleaning part so we'll quickly go through the theoretical concepts maybe in 15 20 minutes and then we will jump into the practical part okay so practical i have the python code ready so i will run it in python in parallel i will also show it in power b okay so it's a mix of concepts basically it's a session of ada in python and power bi both okay and it will be little bit fast so uh if you want a detailed session you can watch out the live recordings okay which i already have on the channel data cleaning what does data cleaning mean handling missing values okay now what does handling missing values mean handling missing values means handling missing values right let's say you have missing values you have to handle it now there are some logics that we use for example you have data okay some columns are here now just imagine this particular column 80 percent of the data points let's say 61 30 so if more than 75 percent or 80 percent of the data is null why do we need this column right so logically what we do is we delete the rows of columns that has more than 75 percent missing data okay again it's not a very hard and fast rule to do this okay i will explain you a scenario where you need to analyze that and then delete it i will talk about it the second approach is replacing the missing values with mean median and mode so what is mean if i take few numbers 5 15 20 7 9 10 what will be the mean the mean will be 5 plus 15 20 plus 20 40 47 56 68 68 divided by 6 it will be 10 point something okay that is going to be the mean now if you have a missing value in that case you are going to impute it using the mean value if you do it using median what is median 5 15 20 7 9 to n what will be the median here to find the median median is basically your middle value to find the median you have to sort it out first 5 7 9 12 15 20. now what is the middle value the middle value is 9 and 12 now 9 plus 12 is 21 divided by 2 is 10.5 so if you want to impute it using median value you can impute it using median value so these are few techniques to replace values as well if you have a categorical feature let's say you have gender you have female female null null male female so in this case three are female and two are sorry one is male so here the mood will be clearly female so some people also impute it as female some people also do this thing let's say female female null null male now here the distribution is three female and one male that means 75 percent female and 25 percent male some people also do this way they impute three records as female and one as male just to maintain this ratio okay so there is no hard and fast rule in this particular step what we do is we do random hit and trial so in eda you can basically fill it with anything but these kind of

techniques are very important when you deal with your predictive analytics part in predictive analytics let's say you are creating you are creating a ai model uh maybe a classification model in that case what we do is we create model here let's say we impute it as mean and then after creating model we check the accuracy maybe after the accuracy is low i will again come back to this step and impute using median again check the accuracy so in in your predictive analytics part a lot of hidden trial is performed okay so you can use any of them the third technique is algorithm imputation as i already told you that in machine learning there are many algorithms that don't require you to deal with the scenarios they automatically deal with it so null values imputation is also not required in some scenarios people also do some predicting the missing values which means let me just clear it out which basically means what is this scenario predicting missing values now just imagine you have weather data or not whether temperature 25 26 null 27 null 31 27 29 now let's let's say you have four missing values okay so if i take the mean technique i know i have to calculate the mean and then impute it right this technique is basically nothing but let's say you take two values and do a forecasting so there are some time series algorithms which are used for forecasting now you forecast this value just imagine your forecasted value is 25.5 now you take all these values and forecast this then you take all these values and forecast this then you take all these values and forecast this so this is a very tedious task it takes a lot of time in imputing the missing values so normally we don't do this but it is good for explanation if you are going for interviews if somebody asks you then you can probably explain these scenarios okay snowflake is not covered i will i'm not going to take snowflake gateways yes i will cover in the in the next sunday's class 15 20 minutes i am going to cover about scheduled refresh and gateways and then i will jump into the dashboard so i already covered this part types of data which is qualitative and quantitative again fancy names don't get confused qualitative data means a variable to describe the quality of the population simply qualitative means categorical data quantitative means numerical data fancy terms that's it okay again in categorical data there are multiple options like nominal and ordinal what is nominal and what is ordinal do we have an example yes we have an example here nominal data is nothing but let's say gender male or female location this or that mumbai or delhi or kolkata that's a nominal data ordinal data means a categorical feature which is in order let's say economic status status low medium or high uh maybe gradation student grades a b c d e all these are ordinal data okay in quantitative there are discrete and continuous so number of students in a class and height of students all these are different types of data derived matrix is something which is being derived or extracted from your existing features create a new variable from existing variables to get insightful information from the data the first technique is feature winning what is feature winning age to age bin let's say you have this numerical feature and you want to convert it into bin 0 to 20 0 to 20 let's say you have 30 00 to 20 this is 2240 something like that that means you are extracting this bin from this particular feature another concept is feature encoding now what is feature encoding feature encoding is simply nothing but in machine learning uh logically what happens is when we have categorical data ultimately when we pass our data to our ai models in that case ai models are simply nothing but some programs in our computer machines right and computers don't understand the categorical data okay so computers don't understand categorical data so it always required to convert categorically into numerical now one quick example will be let's say you have gender male female male female female simply how can i convert it into a numbers i will give male as one female as zero or vice versa this is one technique another technique is so this technique is called as label encoding another technique is one hot encoding let's say you have location you have delhi mumbai kolkata so what we do is we create three columns location delhi location mumbai location kolkata so wherever you see delhi one zero zero mumbai zero one zero kolkata zero zero one delhi one zero zero mumbai zero one zero now this data has been converted from a categorical data to numerical data this concept is called as one hot encoding so there are multiple encoding techniques one hot encoding uh there is uh label encoding there is dummy encoding there is hash encoding i have a detailed video on encoding on my channel in case you are interested just save it maybe just write it down search data encoding and my name you will be able to get the video okay i'm not that famous on youtube so whichever topic you search search with my name you will get the videos okay and the next one is from domain knowledge so let's say sometimes using our domain knowledge also we create multiple features yes dummy variable that's what i told why not encoding and dummy encoding right so many people might not be knowing about this concept called as feature encoding just go through the videos and next one is calculated from data so it could be multiple scenarios where you might have to derive some features from our existing features that's also a very important part so repetitive slide i will just skip it so here handling missing values i have already taught feature scaling technique now feature scaling is a very important technique simply what is feature scaling i will try to explain in a layman's point of view okay and feature scaling is very important when you are feeding your data to your ai models for predictive analytics it is not required for et okay for eda feature scaling is not required but this concept is very i this concept is very required for ai models okay so what is feature skill scaling just imagine when you are in my class you have to imagine so there is a imaginary world and i keep talking about this thing imagine imagine imagine so have customer id it's not customer id i will just change it uh let's say i have i have i have i have i have just name okay some name height and weight okay there are some there are some people and there are some height let's say 172 171 180 165 170 and weight is 85 90 79 80 100 okay now simple question what is the units here is this centimeters meters or feet can somebody tell me in the chat box write down in the chat box okay great i'm not able to see the chat box but keep writing what is the unit for height and what is the unit for weight the unit for height is very simple it is centimeters the unit for weight is kgs how are you able to tell this simple answer is because you know this from quite a long time right you know these things since a lot of since you're since maybe since quite a long time but how will computers know about it if you pass this data to the computer computer will understand this is nothing but 172.85 what it will interpret so computers don't understand the units they don't understand they only understand the magnitude so normally what happens is computers treat they treat numbers equally so what they do is they will give some priorities to this number and give some low priorities to this number there could be some scenarios where they could give high priorities to this number no priorities to this feature in that case your entire predictive modeling part will be messed up right in order to not do that you don't want to create a biased model so feature scaling is important now if i scale it down what will be the output i'm just writing it here height and weight scaling means you are scaling down to a certain number let's say i want to scale this to a range of 0 to 1 i also want to scale this to a range of 0 to 1. what can i do i can simply divide all the numbers by the maximum number so 172 divided by 180 maybe 0.95 i'm not maybe we can do it divided by 180 0.95 171 divided by 180 again 0.95 is 1 165 divided by 180 0.91 170 divided by 180 0.94 now all the numbers are within the same range same with weight i will do 85 divided by 100 0.85 0.9 0.79 0.81 now if you give this numbers to computers they will treat them equally see in feature importance wise feature importance wise this number was the highest here it is still the highest here this number was the highest for weight this number is still the highest for weight so feature importance is remaining the same but the numbers are scaled down so this is a very important concept that we do in our predictive modeling part but try to understand this way that there are two techniques one is standardization and another technique is called as normalization standardization is simply converting the data into x minus mean divided by standard deviation simply what you can do here is try to open your excel sheets okay try to do this exercise after this call try to write down few data points 170 to 171 180 165 and then 170 let's say this is your height right okay now calculate your mean what will be your mean let's say mean is nothing but mean mean is nothing but average of these four numbers that's your mean what will be your standard deviation so you can pick any in any tone yeah i'm just picking excel so what will be standard deviation let's say standard deviation of population so on this now what will be the standardized value x minus mean divided by sigma so this equals to let me just sorry create the mean column here which is 171.6 171.6 17.6 1.6 okay this will be height minus mean what will be the value of this is minus this 0.4 i'll just extract it now i will do this divided by 4.84 this divided by 4.84 now you can see all these numbers are scaled down and this is the technique which is used here standardization similarly in normalization what you need to do you need to find the minimum value and the maximum value let's say i will just copy paste it for another sheet now i don't need mean i don't need all these things here right so here what i will do is i need the minimum value and the maximum value so here i will do this minus mean what is the minimum value 165 divided by max minus mean if i do this sorry something is wrong 0.67 similarly if you do this for all these values it will again be scaled down and this is what i told you this concept is called as normalization okay so these are the two concepts that is very much important if somebody asks you what is standardization and normalization which is again very important question in interviews you have to memorize this formula standardization is nothing but x minus mu by sigma and normalization is nothing but x minus x min divided by x max minus x min okay our next topic is outlier treatment outliers are most extreme values in the data it is an abnormal observation that deviates from the norm outliers do not fit in the normal behavior of the data detect outliers using following methods so we detect outliers using multiple methods okay and using these methods we can identify outlets sometimes we have to remove the outliers but as i told you in eda it's not required okay i i haven't done outlier removal in ads so far it's not required because outlier detection and removal is more important for your ai modeling part not for eda okay but you can do it just in case you want to show some more insights you can do it now how to perform these steps again i do have a video on this probably you don't remember maybe i'll just try to open [Music] i don't remember yes outlet detection is already there using standard deviation box plot and outliers probably you can check out that video and then you can go through it okay so i'll not cover this because it's already covered because we have a lot to cover right now handle invalid value so there will be certain scenarios where you are not having data in a proper form proper format right there could be some encoding issue that could be some incorrect data types sometimes data date is in this format and you need to convert it into this format so there could be certain scenarios like this where you could all you should also do these steps right you should convert it into a proper format so that you can use it in the future now analysis eta is evolving around these four concepts so these are the four concepts that it is evolving around one is univariate analysis bivariate analysis correlation and outliers yeah videos are on my channel okay i think mostly ed is covered on my channel already if you want to have sub topics just search like outliers you search outlier satyajit you will be getting some videos outlier correlation you will be getting some videos or something like that i don't remember but there are some videos okay so ada is evolving around these four concepts univariate bivariate correlation and outliers now use cases of eda eda as i told you in data science right now also right now also at least 60 70 percent of the companies are only dependent on eda only 30 to 40 percent companies are working on the predictive modeling part and among this 30 to 40 percent there could be 10 to 20 percent who would be working on computer vision and very high-end nlp concepts still there are many companies who are hiring more and more research analysts who would be more interested in performing ada because ed is a very strong concept okay and use cases are numeric customer churn analysis cancer data analysis fraud data analytics banking related analytics google analytics so you just pick your domain try to search data analytics use cases in my domain like my use cases in marketing there will be lot of lot of things okay funnel analysis lot of things are there so in this class we will be talking about subscriber channel okay customer churn analysis so many people if you are a repetitive subscriber you might be knowing that i have a lot of videos related to this particular use case so i will be repeating the same and in this class we will be learning about subscriber churn analysis that means i will talk about these things i will talk about the data i'll talk about the business understanding okay business understanding i will talk about why that means why are we solving this problem and i will also talking about the how part how are we solving it okay first of all data is kaggle data okay but the the python notebook that i will share with you which i have already shared on my channel with many other people uh that is already on my github repository as well it's a kaggle data but the eda file that i am going to give you you can simply use that architecture or skeleton date skeleton file for any kind of eda in the entire world okay very few things you have to change so my file is well structured and it has everything line by line everything is explained in that particular file so there is nothing problem when you go through that file it's a python file but as i told you that it's not going to be very difficult for you to understand those lines and you are going to simply take that code and apply for your own specific eda use case okay so you don't have to write the quotes from scratch so don't get frightened by seeing python notebooks okay second one is business understanding what is the business understanding about this particular problem now i usually pick this project because i out of my ten and half years of experience seven years come from telecom domain so this is one of my classic projects that i have created end to end i can bet people who have this project on their channel or on their course i'm pretty much sure they might not be knowing the end-to-end use case but i have already created this okay people think that it's very simple thing just take it prepare a classification model done it does not happen that way subscriber churn is very complex okay but we will not be focused on the ai part we will be focused on the ada part we will try to analyze the data first okay but before understanding the data we will talk about the business understanding okay what is the use case all about what are we trying to solve and all those things why this use case and how okay so i have some beautiful visualizations which people will like it as i told you that people think subscriber churn is simply okay before before getting into this let me explain you this use case once more churn means somebody who leaves a company okay subscriber means subscriber you know that i will give examples telecom let's say today you all are using vodafone okay and just imagine you are using 199 rupees plan you are getting unlimited data unlimited calls blah blah blah okay now airtel launches a new new plan 99 rupees everything is unlimited there will be 100 there will be so many people who will churn this like they will leave vodafone and they will switch to airtel they will do the portability and they will quickly jump in now this scenario where customers are moving from vodafone to airtel is basically called as customer churning and these customers are a loss to vodafone company it's a gain to airtel company but lost to vodafone now imagine you are working as a data scientist in vodafone you will be given this particular data and you will be given can you analyze what is the exact reason why people are living another use case of churning is banking see just pick any industry churning is a common use case okay you can simply relate with the with your own use with your own domain how banking is related to churning simple today you are using kotak it is giving you fd six percent on two lakhs and above for an example now you are exploring hdfc ex hdfc is providing you seven percent ft on two lakhs or else hdfc is giving you few more features so there will be certain amount of people who will be churning from kotak and joining sdfc right this is nothing but the customer churn another use case a very one of the use cases which you will be able to relate that i know many people are into gaming right many people play games even i play games if you open my desktop you will see a lot of games even after this call is over i will play some games so in gaming people tend to install and uninstall apps right now if you install an application x after few days you get bored and you jump into application y yes i play valorent if you are installing if you are playing x and you are pissed off and you switch to y that means you are churning from x what is the reason that there could be multiple reasons let's say in game x you think that the game is too too easy and you are pissed off what the hell is this and then you are pissed off and you drop and you join another game or else you are stuck at one particular level for ages and you are pissed off oh i am not able to move my rank this is not a good game then you switch to another game so any industry you talk about there will definitely be a case like this hr analytics in hr domain employees leave companies right why do employees leave the companies now that's where hr's needs to do our eda on their candidate's profile data right so any use any industry you go there will be an use case related to churning and that is one of the reasons why i picked this use case because it is applicable for all the industries okay now these presentations are very much related to telecom so try to have some focus here because in telecom also churning is not simply moving from one network to another network that is just one type of churning that is just one type of churning there are multiple types of churning first is tariff plan churn for an example you are on a 999 plan and you move to a 4.99 plan now you will ask me why is is it called as a churn simply because earlier there is a 500 rupees reduction from one user for one month that means six thousand indian rupees of loss for one customer just imagine one million customers you can just imagine how much cost it is right sorry not 500 yeah 500 right so this is also a type of churn we have to understand why people are moving from this plant to this plant okay even if we are providing everything here here there is a limited uh limited let's say here your data is unlimited here your data is 2gb why people are moving that you need to analyze another type of churn is service chart example weekly monthly subscription let's say you are you are on prepaid you have some subscriptions caller teon etc etc and then you cancel out some plans this is also a type of churning another churning is production example post paid to prepaid why this is called as a churn imagine in post paid it's a regular income to the company if you are on 9.99 plan 100 guarantee that your bill will at least be 9.99 on top of that if you are doing some extra international calling and all those things it will definitely be greater than that it cannot be lesser than that that means there is a guaranteed 12 000 income from a particular customer if this customer moves to prepaid well today he is using 9.99 that's fine but after three months he could move to 399 he could move to 299 so there is no guarantee of this income this is also a type of church and another type of churn is usage we are taking a lot of time explaining these concepts i'll try to quickly wind it up in 5-10 minutes and then we'll jump into the code part okay decision cycle of a subscriber changes as per needs and our experiences now this is the decision cycle of a subscriber i am a mobile customer and i haven't thought about churning why because it's too complex honestly speaking this type of customer is me okay because i find portability and all those things the process is too too complex for me it's too irritating for me to i'm feeling too lazy for this activity so these kind of customers are inner subscribers the next one is i am a customer i haven't thought about churning because my operator is the best this is unconditionally loyal customer and my father falls in this category he still uses bsnl i mean he still uses bsn since past 400 years so he is an unconditionally loyal customer these kind of customers are never going to churn so the companies can bet on them definitely they will not leave the company another type of customer is i am a customer i haven't thought about churning because i'm logged in a contract now this concept is not available in india but outside india there are contracts for example i am right now in hong kong in hong kong we do have contracts our mobile networks are on contract our wi-fi are on contract when we sign a contract we are we have signed for 12 months or 24 months that means even if i face a problem they don't give a about it you have to either pay the whole amount or you have to live with it simple okay so they are locked in subscribers coming on to the dotted line below the dotted line i am a customer i have thought about churning and i am not locked by any contract but i have somehow decided to stay these customers are conditionally lawyer again you cannot touch them they are conditionally loyal people who have decided to live because they have found a better offer they are conditional churners people whose needs have changed they are lifestyle migrators they are never happy customers they switch companies every three months every four months they switch companies okay because i am not satisfied unsatisfied churner now this is where our company let's say i am working for vodafone if i am able to identify customers who are falling in this category or this category i will be able to retain them but not this category lifestyle migrators are very difficult to retain okay this is the entire decision cycle of a subscriber so these are the four different segments this is a very high level overview of data science led approach to manager it all starts with capturing and analyzing this is where data analyst comes into picture the second part is reporting and predicting in the reporting part also data analyst or the data visualization expert comes into picture predictive part machine learning or data science team and here also your predictive analytics here also data science team comes into picture okay so that's it i'll quickly jump into the practical code i already have the churn ada uh file with me we'll quickly take a five minutes break we will be back at my time 9 30 7 p.m for you we'll be back at 7 00 pm ist okay give me five minutes and quickly have some water and i will be back in case you have any questions write down in the chat box i will be answering the questions at the end of the session and okay give me four minutes what five minutes oh i am back but as we have committed let's give two more minutes break in case somebody has gone for water break or something like that and in the meanwhile let me pick the questions so good evening good evening okay i already answered do you plan to take data various in concepts no i'm not take standardization may reduce certain number of outliers no sir it will be greater great help if you could please schedule schedule a session on atm i'll have to figure this out because let's see next wednesday maybe i'm not sure okay not much question snowflake will not be covered etl yes i'll try to cover etl someday at least i will i will not be able to show etl because etl is just a concept which is you know there are various etl tools okay i will not be able to explain about those tools because i don't have those tools handy on my system but i can take a one hour session on etl where i can explain how theoretically how things work which can help you out in you know answering your interview questions that i can do but practicals on a ta is not possible atl is not possible if that is fine then we can go through a theoretical concepts or etl and we can do this can you please share the link of the github it's very simple remember my github repository as you are part of my family right now pik 1989 1989 is my birth year p i k is what i don't know uh this is my repository simply go here search for churn you will get a lot of repositories okay go for ml project custom ml project churn prediction okay and here you will find this ada file you will file you will find the model building part you will also find this what is this okay you will also find this eda it's a very old eda pdf file and you will also find some you know some images for the ppt that i shared okay that you can do hmm and here also there is a detailed session on ada you can go through this edf video as well and once you are into your data science journey you want to learn about machine learning model building and all then you can watch out this video which is one of my top videos on my channel so you can watch out that you can see see i'm famous 24 000 views okay so here also i've explained about the use case and then i have done the project deployment and everything so i have created your everything has been done here end-to-end okay now we have already reached 7 p.m let's get started with the edm process i will quickly run through this code i will not take much time because detailed line by line code it will be very difficult for me to go through initially i am importing all these libraries and i am extracting the data okay so i'm reading the data reading it into a pandas data frame okay once i have done that i'm doing data dot head in parallel we will also open power bi desktop so that we can also do these things in parallel in power bi initially i planned for one hour session but it's already one hour so let's see tell you coach this one i think this one this one here ah this one this one yes i'll load it guys if you are liking the stream please like the video support the video anyway it's free of cost doesn't take much of your money from the pocket like it and uh [Music] that's it subscribe the channel if you haven't subscribed okay so i have already taken the data in parallel i will also show you what i am doing in python okay so here i am checking how the data looks like okay so you can do the same in power bi by going to the power query editor and you can check your data right or this is what the data looks like next step is check various attributes of data like shape how many columns how many data types all these things these are very basic concepts to check you can simply check it using your record count you can simply go to your home screen and you can simply create a new card and check oh my total record count is count is four zero four seven zero four three okay all these things you can easily do here i am doing column dot values here i am doing the data types i am just checking all the data types you can see some of them are objects some of them are integer all these things here i am doing some description like i am just describing some features to understand what these features are and what are the statistical information from this particular columns even though my data frame has so many columns in describe why am i only seeing three columns because describe function only works for numerical attributes so only those attributes that are integer or float for them only describe will give you an output for others which are object it treats object as a categorical feature so it does not return you the output in the description okay so again this will be a little bit manual process in case you want to do the same thing in power bi you simply have to you know let's say monthly countries simply have to take your monthly charges and compare it manually you have to do all these things so what is the average what is the minimum value what is the maximum value this you have to do minimum value maximum value what is the standard deviation what is the variance what is the median all these things okay so similarly you can create multiple cards for monthly charges for 10 year and do the same thing okay in parallel i'm just showing you how you need to do here and there and i will give you both the options you can use both options in power python or powerpi but i personally prefer python senior citizen is actually a categorical feature hence these 25 percentile 50th percentile or 75th percentile distribution is not proper okay 75 percent customers have 10 year or less than 55 months average monthly charges are 64 usd whereas 25 percent customers pay more than 90 dollars a month that's the inside that we got from this particular thing i'll move on here what i'm doing is i am trying to plot my target variable what is my target variable churn so i am plotting what is the distribution of chunk so how can i do it here simply i can get my churn column here access values done and here i will give data labels done 5.2 k 1.9 k it's quite visible that this data is imbalanced in nature so more than 79 or 80 percent people are not churned or active and 20 25 people have left the company that's what it is telling here moving on here it is showing the exact percentage okay you can also do the exact percentage i haven't tried it though but i think somewhere it should be possible i'll check on this i don't remember how to plot the exact percentage but this is how it is okay or else you can calculate it manually right so data is highly imbalanced ratio is 73 is to 23 so we analyze the data with other features while taking the target value separately here i am doing data dot info info basically gives you an overall idea about the data what are the different columns that are available what are the different types of like data points that you have do you have non-null or do you have null characters and something like that this piece of code will show you percentage of missing values for each column and as i told you repeating again that do not get stressed out by looking at the code you can simply use the same code for your own ed you just have to change few things here instead of this you have to change it to your own data frame here you want to change it to your own data from that's it rest everything remains the same so you take the code use the skeleton that's it missing data initial intuition here is we don't have any missing data general thumb rules for features with less missing data can use regression to predict the missing values or fill the mean of the values present depending on the feature for features with high number of missing values it is better to drop now imagine there is a scenario there is a scenario car type there is a column called as car type now car type is for some people it is suv null non xuv non suv imagine you have 1 2 3 four five six seven eight nine ten out of ten records only three records are available seventy percent missing values so what should we do should we remove it or should we just populate it using our median value so in this case the general thumb rule says that you have to remove this but you should always analyze the column and see whether this column is dependent on another column or not for example based on your domain knowledge you get to know that this column is dependent on another column called as iscar yes no no yes no no yes no no no and after analyzing this one you get to know oh wherever i have no there it is none it's quite obvious if the customer does not have car then the car type will be null right because it does not have the car itself so a better way to solve this problem will be imputing using something called as no car no car no car no car no car this could be the best solution here okay rather than just populating using the median value what is the median here two suv and one xuv median is xuv right if you would have imputed media xuv that could have been wrong okay so here yes good topic that you mentioned is that whatever we did here like missing values and all those things how to perform this in your power bi you simply have to go to your transform data and you have to do your data profiling column quality will give you error percentage and empty percentage right so this is what you need to do in your power bi coming back to eba what is ada then comes your data cleaning part which is the most important thing here we have to clean the data so i have just copy pasted the data and created another data frame here i am converting my total charges to numerical object why because total charges is an object so here am converting it into a numerical data so same you can see if you go to your data when you have clicked on column quality you can see here total charges here you are able to identify because power bi automatically has identified total charges as a number but pandas is not doing a good job pandas identified it as an object sorry this one so in pandas you have to convert the data into two numerical data and then you have to deal with it whether you want to delete these rows or you want to impute it you have to do it so here in this case as it's just 11 records 11 records out of 7043 which is around 200 which is around 0.15 so what i have done is i have blindly remove these records so here if you want to impute these records you can use this replace technique or you can remove the empty datasets or remove the remove the empty records you can do that uh let's say you want to replace it first null with zero and now let's say you want to select this zero how many records do we have these records now if you want to delete you can delete these records up to you it's up to you whether you want to delete it or you want to impute it but both the ways are possible here right moving ahead here i am converting my tenure into yes notebook is available just go to my github repository you will get it okay here what i am doing is i am converting the tenure into ten year group so how to do this in power bi i had already explained people who are joining my live sessions they know this right you can simply create a conditional column then your wins and then you can convert it here if any r is greater than 61 or greater than equal to 61 then 61 to 72 and safe tenure is this to this so this is you possible using conditional column if you don't know about it go through the day two videos which is 27th march video you will be able to understand it here i am removing some columns like customer id and tenure why am i removing 10 year simple because now i have 10 year group so i don't need the original column i can delete it same with customer id customer id is not required so how to do this in power bi right click remove done so both the ways you can do power bi or tableau sorry or python till here you have already cleaned the data okay after cleaning comes your data exploration in data exploration the first topic is your categorical data analysis in categorical data analysis one is univariate one is bivariate one is multivariate now simple explanation is univariate is nothing but analysis of one column let's say you are analyzing gender male female bivariate means you are analyzing gender male female along with that you are also analyzing some other column let's say gender with location let's say there are mumbai and delhi main mumbai main delhi something like that so male mumbai and delhi mumbai and delhi female mumbai and delhi so the inside here will be there are more people who are male they are staying in mumbai more females are staying in delhi here where you are analyzing two columns that's called as bivariate when you analyze multiple columns that becomes multivariate how to do it in python i have simply written down three line of code where i am using an iterator i am basically using enumerate here i am dropping these three columns and taking all the other columns as a predictor and as a data and here are the insights so for churn you can see i'm this is the gender female and male female and men senior citizen you can see senior citizen is basically giving you better insights so let's try to analyze it using our power bi we will have senior citizen in my access let me just first convert the senior citizen yeah code will be updated after the class is over but if you need it in the class itself you can also go through my github where is the so let's say i will have my senior citizen and and then i will have my [Music] i think i know the problem why because senior citizen is a numerical attribute i have to change it to categorical that is a problem maybe i will pick another one i know another important feature which is your where we have contract we will use contract okay contract we'll use it as an axis i'll use it as values and you can see this is my contract okay you can see that for month to month let's say i want to give legend some colors data labels colors no is maybe in red and yes in [Music] blue okay now it makes little bit clearer so you can see that what does the data tell me here that for my month to month customers how many customers are there probably around 3.9 k how many two yearly customers so two year contract customers around 1.6 k okay blue is very minimal it's not visible at all maybe if i can expand it and then one year which is one year contract which is around 1.5 k now here you can see people who have left the organization churn as yes you can see for month to month customers that churn ratio is very high right for two yearly contracts what is the current ratio just imagine they are in hundreds or two hundreds so maybe 100 divided by 1700 the ratio is very low 100 divided by 1700 multiplied 100 which is 5.88 around that one year how many let's say 200 divided by 1500 multiplied with 100 which is around 13 so five percent people have left who are having two year contracts one year contract 13 people have left but month to month how many 1700 divided by 2900 multiplied 100 around 43 percent have left the company so this is an insight to you right more monthly customers people are churning so you can also do this way in your python way also do it in your power bi way also okay similarly here what i'm doing is i'm converting the target variable to 1 and 0 that means yes is 1 and no is 0 and then i am dividing the data here i am performing some get dummies operation basically to convert the categorical feature to numerical feature here i am doing some relationship analysis between monthly charges and total charges total charges increase as monthly charges increase which is as expected right here i am doing churn by monthly charges and total charges if you look at this graph this graph is basically monthly charges and total charges with respect to churn or not you can do that using a line chart also let's say you want to perform a similar kind of thing you can do using line charts so what you need to do is you need to add monthly charges i will remove this legend okay this is the yeah you can see that for monthly targets there is a spike here and if you add the total charges maybe i will also add churn as a legend you can see this churn no for churn no you can see there is a spike here right so this is the same thing here there is a spike here in the red category red basically means your which one rate basically means your monthly charges so churn is high when monthly charges are high similarly if you do it for total charges this is how it looks like if you remove the total charges this is how it looks like so in total charges surprising insight is that higher the churn at lower total charges however if we combine the insides of three parameters that is 10-year monthly charges and total charges then the picture is bit clear higher monthly charge at lower 10-year results into lower total charge hence all these three factors that is higher monthly charge lower tenure and lower total charges are linked to hygiene so this is what the inside that we got similarly i have done some correlation build some correlation matrix here so what is the correlation matrix correlation matrix is simply again lot of statistics is required here probably you can go through one of my videos i have go to my channel you know my name you go to my channel go to the playlist there will be a lot of playlist search for business statistics try to go through each of the videos there are five videos it will hardly take you one and half hour go through all of them because those five videos are meant to be meant to be uh what do you say it is required for a data analyst okay so go through that in one of the videos i have explained correlation in depth but in short correlation is simply nothing but the relationship between multiple features let's say you have height and you have weight now if you analyze height and weight if your data points are like this which means when your height is increasing your weight is increasing that means height and weight are positively correlated so your correlation value will be somewhere around 1 definitely greater than 0 okay if your data points are like this which means the data points are decreasing when height is increasing your weight is decreasing this means negatively correlated where your values of correlation will be very closer to minus one and definitely less than zero and imagine a scenario where your data points are randomly scattered there is no relationship you don't get it whether increasing one column is impacting the increasing in another or not this is neutral that means those two parameters are not related to each other here your correlation value will be somewhere around zero could be around minus point two two point two somewhere around that so here i am performing the correlation this graph is simply the correlation with respect to churn similarly there are some insights here you can go through it and this is the correlation matrix i'm doing on my entire data frame dot core i'm just calling the dot core function and i'm plotting it using the heat map and using you can see this is the heat map correlation okay so correlation is very important i will also help you get some insights from this because without insights does not make sense so you can see i will try to find purple ones okay you can see you can see this box that means monthly charges and streaming movies yes so people who are streaming movies yes and monthly charges they are basically interrelated to each other okay similarly in power bi i haven't built a heat map honestly speaking but there should be definitely a heat map here and we are i think we got it i think this one i haven't used heat map in power bi but this should work let's just try it out heat map here it is asking for category and y variable okay let's say i want to do a heat map on monthly charges and total charges and turn okay no idea how to read this probably i will just get back to you soon i have no idea on how to create a heat map on power b i will i will let you know this is open open question open open i am not getting the word this is a open-ended query or whatever it is i will get back to you okay probably on whatsapp groups or something i'll get back to you in this and then i am doing some bivariate analysis by variate also you know that we analyze multiple features as in two features so here i am converting it into two different data sets one is churn data set and one is a nonchant data set here i have defined a function uniplot function again you have to simply copy paste this function you don't need to understand this once you copy paste it you simply have to call this plot on your data frame you have to pass a particular column and then you have to pass a hue on gender that's it so this is nothing but your partner versus gender graph this is bivariate analysis partner versus gender graph for your joint subscribers for non-church subscribers and this way you have to move ahead go through it it's going to be dead simple very simple to understand this particular notebook go through it i'll also apply update the code after the session is over and the conclusion here is these are some of the quick insights from the exercise okay electronic check medium are the highest earners contract type monthly customers are more likely to churn because of no contract terms as they are free to go customers no online security no tech support category are high churners non-senior citizens are high jobs there could be many more such insights so take this as an assignment and try to get more insights so that is all about this particular session on eda i had never done this video on a comparison between python and power bi so i hope you liked it and that's all about this particular session we can take some questions uh let me switch to our stream here and then we can probably take few questions and that's it guys is this the same ada required for yeah eda is eda same idea yes does not matter if you do eda using python or power b but ada is eda exploratory data analysis it has data analysis in it right so it's required for data analyst i'll take some questions guys uh in case you have any questions let me know or else maybe you can utilize this session in asking something related to your career or something like that i will also be able to answer that uh you can start answering you can start asking how many days class will be going for ada this is just a single class okay so this was embedded as part of the free live power b classes we already had three classes on power ba and our next class is on 10th of april which is on live dashboard and eda was a little bit prerequisite to it so i had to take this class okay and henceforth we will also be doing a lot of uh live sessions so don't worry about all these things i'm totally converting my channel to live classes everything will be live no only live classes the only way that you can support me is by buying some of my udemy courses or something like that which i will notify you on channel or on whatsapp groups but other than that everything is going to be free so down the line in next six months i'm going to teach everything sql machine learning data science deep learning lot of things will be done so questions questions questions something that is not as for expected yeah abnormalities yes correct i think it was already covered what exactly is data cleaning it's covered in the session uh you are audible okay thank you when you teach snowflake i think already answered on snowflake i'm not a snowflake expert i haven't used snowflake at all so i don't know when i can i don't know about snowflake i haven't used snowflake but i will teach gateways on this sunday probably for 15 20 minutes not more than that because our class is dedicated for live dashboard thank you deepti you are becoming famous thank you thank you i'm not famous on youtube i need to be famous anyways uh questions questions let me scroll down are these videos available yes it's going to be available forever 18 classes has committed i'll take one session it will be a theoretical concept session on maybe next wednesday or something i don't know provide wednesday i'll take i'll drive ticket link i will type out the link here github dot com slash speak 1989 that's my github and simply you can find it out okay i'm just mentioning my github repository to duplicate the how to delete duplicates from large data sets it's very simple in power bi if you want to do it you can do it but it's better if you do it in the source data itself just search it out it depends on which tool are you using if you are using python just search how to delete duplicates you will get different codes on stack overflow you can use that it's very simple is it enough that sql and power bi for data analytics no i mean good enough for 40 50 percent of the companies like maybe if you are working if you are trying for mnc companies or some generic companies that's fine but if you are looking out for ai focused companies or you know companies who are very much focused on data science and deep learning activities or data analytics activities in those scenarios power behind sql will not be enough you also might need to know about a lot of things so recommendation will be learning more and more things in parallel data scraping and excel and all these things i am from ghana and i really like this session thank you isaac if i spend out your name correctly so for refresher in data science they expect cloud deployment knowledge any resource for learning cloud cloud is not required if you are getting into data science domain but if you want to for just practicing and preparing for your profile what you can do is you can start deploying your machine learning and deep learning models on cloud for that yes there is something for you as i told you there is a platform here you can simply go oh i'm not sharing my screen i'll share it you can simply go to this platform and you will find this link machine learning model deployment free course take it and it's free because it's available on youtube as well if you want to go through my youtube channel simply go to my youtube channel and sorry sorry sorry sorry i'm watching my own video go to playlists and there will be machine learning of it uh dedicated employment deployment huh this one learning model deployment this is more than enough for a data scientist you start deploying models start learning about flask that's it you don't need end-to-end cloud expertise okay next question i think i already answered how to delete duplicates a very informative session thank you when can we expect sql classes in may definitely in may uh so i will be teaching that in may because i'm going back to india in april 14th uh so once i am back i will be taking the classes on sql okay what is dax i think psy you should watch my previous video on power bi part 3 go to my channel videos the video that was published on 3rd of april sorry on 27th of march so go through that video i have explained that in details okay data analysis expressions just five more minutes and then i will wind up the session in case you have any questions related to any off topic question also that's fine you can ask me and then we'll wind up okay just three more minutes tick tick tick tick the clock is ticking guys no questions okay so thank you everyone for joining the session it was really nice at least for me how nice it was for you you can let me know in the comment section below after the video is published because the video as it is a live session it will take some time to be visible on the channel so it's a request as you already are watching 40 people are already watching right now it's a humble request to comment down in the video once you see the video tomorrow okay comment down something if it is useful coming down if it is shitty comment down whatever it is coming down okay because comments and likes is what youtube needs based on that my video will be recommended to other users and without that it's very difficult right so comments and likes are very important and rest i know all of you will not be able to share the video because of laziness i am completely okay with that but there will be some people dedicated people who will be definitely helping me out in sharing the video please share and subscribe my channel as well and we'll see you in the next video or the next session on 10th of april which is going to be a live dashboard creation on power bi how to join a project under your leadership if you are on whatsapp group probably you can reach out to me personally i will try to give you some ways but publicly i don't usually tell this because a lot of people will reach out to me okay so you can reach out to me on whatsapp okay so that's it that's it guys see you bye bye have a good wednesday and see you all on next sunday what do companies look out for when hiring data scientists it depends from company to company many companies look out for multiple roles like i have applied for multiple roles they sometimes if the company is too much focused on nlp they will be asking or having nlp related criterias in the in the in the job description in that case you need to know about machine learning you need to know about nlp you need to know about the deployment part so it depends from company to company but whatever it is they mention everything in the gta itself so what list that you need to have is going to be machine learning and definitely deployment part flask and all those things is also very required eda concepts of statistics concepts of data analytics all these things are also required okay see you guys and uh see you on monday bye