Python for Data Science
Prof. Ragunathan Rengasamy
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture – 18
Exploratory data analysis

Welcome to the lecture on Exploratory Data Analysis.

(Refer Slide Time: 00:18)
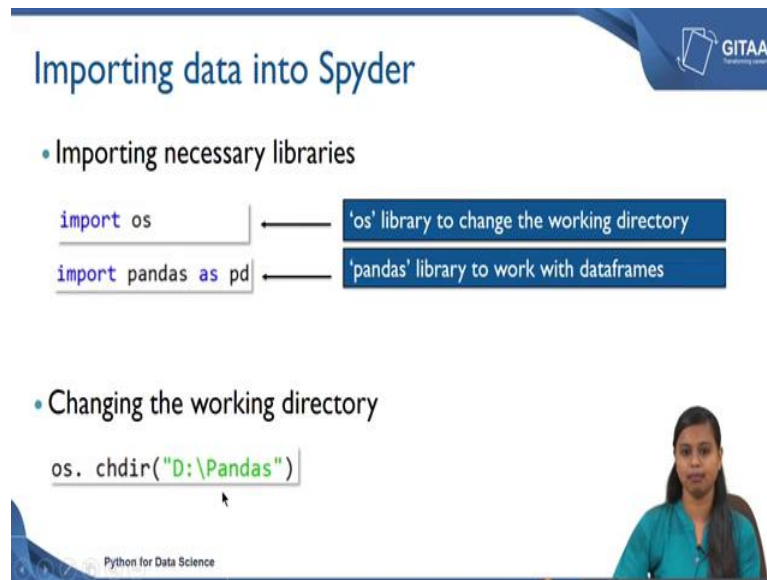

So, in this lecture we are going to explore more on the data that we were working on using frequency tables two-way tables. Followed by that, we are also going to look at how to get the joint probability out of two-way tables. We will also be looking at how to get marginal probability and conditional probability using two-way tables. So, we will see in detail about each of the topic listed here. At last we are also going to look at a measure called correlation because, all of the points which are listed above are to interpret or to check the relationship between categorical variables.

But we will also have numerical variables in our data frame, in which case we will also be looking at. So, in that case if you want to check the relationship between two numerical variables, there is a measure called correlation that is what we are going to see in this lecture. We will also be seeing in detail about what correlation measure is about.

(Refer Slide Time: 01:16)

So, before exploring on the data, we need to import the data into spyder to work on that. So, prior to importing data, we need to import the necessary libraries that are acquired for importing any data into spyder. So, let us do that first. First we are importing the os library we use os library to change the working directory. Next we will also be working with data frames because once we read any data into spyder that becomes a data frame. So, to deal with data frames we need to load the library called pandas.

And we have imported it as pd because pd is just analyzed to the library called pandas. So, now, we have imported the necessary libraries to change the working directory using the command os.chdir. chdir chance for changing directory and inside the command you can just give the path from which you are going to access the data from.

(Refer Slide Time: 02:09)

Now, let us import the data into spyder. So, we have a dataset called Toyota.csv which are nothing but the details of the cars. The details of the cars having captured in terms of various attributes like price, age, kilometre, FuelType you can look at the snipped below here. So, since this is csv data we need to use read_csv command to read in csv files and that is from the panda's library.

So, that is where we have used pd.read_csv. And inside which we have just given the filename and we have set index_col as 0; just to make sure that we are setting the first column as our index column. Since we have already worked with this Toyota data, we know that there are some missing values that are in the form of question marks. This question marks these question marks does not convey any message from the data or about the data.
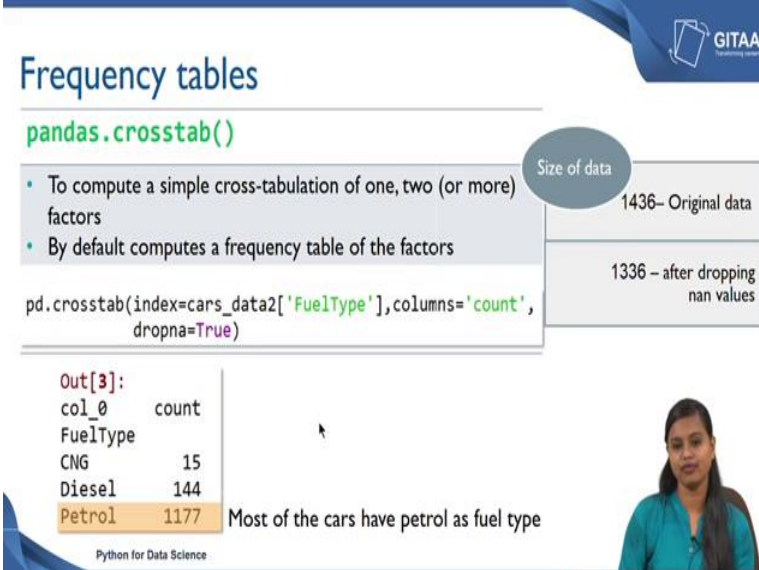
So, in that case we will be considering this as missing value, but here I have given it under na_values because I am going to consider these special strings with the default none values of python. Because, python offers several functions which will allow us to deal with the default nan values.

In this case if I want to perform any operations by considering these missing values it will be a tedious process. Rather I can just consider these strings as default nan values which also the representation of not available values. While reading itself, we are considering all the question marks as nan value.

So, that we can perform all the operations that are related to missing values; and the snipped given below will give you an idea about how the data will look like. Now we have read the data into spyder we are going to create a copy of the original data, so that whatever operations or modifications we are going to do on the data frame will not be reflected in the same data set itself. Rather we can just have a copy of it so that we can cross verify with the original data at any point of analysis.

So, in that case we are going to create a copy using the dot copy function proceeded with the data frame name. So, here cars_data is a data frame name and using .copy I am creating a copy of data and I am saving it into a new object called cars_data 2. So, cars_data two becomes my new data frame which was copied from the original data frame. Now we can use this data frame to do the further analysis or to do operations so that the original data will not be modified.

(Refer Slide Time: 04:50)

So, now we are going to look how we are going to create a frequency table before creating we need to know what frequency table is about. We have multiple variables in our data frame and if you want to understand the data more, you basically want to check the relationship that exist between the variables. But we cannot just check the relationship between all the variables we can do one by one.

For example, we can check the relationship between any categorical variables using cross tabulation or if you want to do univariate analysis on a categorical variable you can

also create a frequency table. So, that you will know what is the frequency of each categories that are available under a variable. Now, let us see how to create a frequency table? So, crosstab is the function that comes from the panda's library, which is used to compute a simple cross tabulation using one two or more variables by default it creates a frequency table of the factors. So, now, let us see how to create a frequency table. So, I have used pd.crosstab that is the function that is used to create a frequency table.

And as an input to the function I have used it as index is equal to cars 2 of FuelType; that means, that I am interested in getting the frequencies of the categories that are available under the variable FuelType. And we know that that is from the data frame cars data 2 and the variable of interest should be given under the parameter called index.

We also need to have the corresponding frequencies of it. So, basically it give you the count for each categories of FuelType. And since we know that we have so many missing values in our data frame, we do not want to consider that while we are interpreting from the frequency table. In that case you can drop all those missing values by setting dropna is equal to true.

By setting that you will get rid of all the records wherever there are missing values. So, you will get the frequencies for each categories for the records where there are no missing values. By setting dropna is equal to true, if you look at the original data size and the data size after dropping nan values you will get an idea about how many records we have lost.

So, the original data size is about 1436 rows and after dropping the missing values we are left out with only 1336 rows because there were 100 records where the FuelType were missing that is why we are left out with only 1336 records. Now, we are going to create the frequency table by considering only 1336 records.

So, now let us see how the output would look like. So, if you see here you have the variable here and you have the corresponding categories under the variable FuelType. So, basically there are three categories under the variable FuelType, CNG, petrol and diesel. And you can also look at the corresponding frequencies of each of it. So, it is very evident from the output, there are only fifteen cars whose FuelType is of CNG. And there are 144 records or 144 cars have the FuelType as diesel. And if you see here petrol

has the frequency as 1177. So, in this case most of the cars have petrol type as FuelType because there are only few cars whose FuelType is are of CNG and diesel.

So, in this case you will have an idea about though there are so many categories that are available under FuelType most of the cars FuelType are of petrol.

(Refer Slide Time: 08:31)

So, in the previous example we have considered only one categorical variable just to get the frequency distribution of each categories. Now we can also have one more variable to check the relationship between those two categorical variables. If you want to check the relation between two categorical variables you can go for two-way tables.

And here we are going to use the same function to create a two-way table and in this example we are going to look at the frequency distribution of gearbox types with respect to difference FuelTypes of the car. In the previous example you have just looked at the frequency distribution of FuelType, but here we are going to look at the frequency distribution of gearbox with respect to the different FuelTypes of the car.

So, let us see how to do that and if you want to know more about gearbox types that are available for the data set that we have here is a snippet. So, gearbox type have been represented using a variable called Automatic, it has two values 0s and 1s and 0 is the representation of the car having a Manual gearbox and one is representation of the car

having an Automatic gearbox. Now let us see how to create a two-way table using crosstab function.

So, the function remains the same that is pd.crosstab, the first parameter that goes into the function is index under that I have just specified the variable as Automatic. So, under the index variable I have specified the variable as Automatic and under the columns you can specify one more variable that is FuelType because we want to look at the frequency distribution of gearbox type with respect to different FuelTypes of the car.

And I am also going to remove all the missing values from the data frame. So, by setting na is equal to True you mean that you it will consider the two-way table will be created by considering only the rows where both Automatic and FuelType are found. That is there should be no missing values in both Automatic and FuelType in that case only it will create a frequency table.
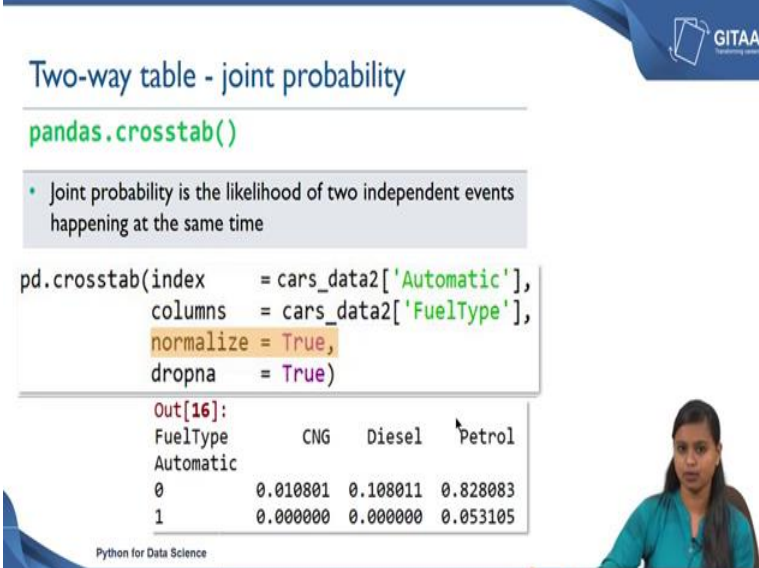
Because, if you have a missing value under a Automatic column and if you have a field value under the FuelType column it will not be able to get the count of it. Rather it will just over that particular row itself and you will be left out with the rows where there are no missing values in both Automatic and FuelType column.

So, now let us look at the output to check the relationship between those two variables. So, since I have just since I have given Automatic under the index argument and FuelType under the columns argument the output is a representation where rows correspond to Automatically and columns corresponds to FuelType. So, from the output it is very evident that. So, now, let us see how to interpret the output. So, if you see 15 here so, you can see the.

So, you can see the value 15 here; that means, that there are only 15 records where the FuelTypes of CNG and the car has a manual gearbox. And there are only 144 records where the FuelType of the car is Diesel and also the car is having a manual gearbox. And interesting and the interesting thing about output is you can see some 0s here; that means, that the cars with Automatic gearbox does not have the FuelType of either CNG or Diesel. If the car is of Automatic gearbox then it has only the FuelType as Petrol. So, low cars have the FuelType as CNG or Diesel given the gearbox is of Automatic.

So, this is very interesting about the relationship between Automatic and FuelType. So, now, this gives us the idea about what is the relationship that exist between the Automatic and FuelType variables.

(Refer Slide Time: 12:25)

So, we have looked at the output in terms of numbers. There is also a way were you can convert the table values in terms of proportion and that is what we mean by joint probability. By converting the table values from numbers to proportion you will get a joint probability values you will get the joint probability values that is also using the same function crosstab. Let see how to do that, we are going to use the same function crosstab to arrive at the joint probability values.

What do you mean by joint probability first? Joint probability is the likelihood of two events if joint probability is the likelihood of two independent events happening at the same time. So, if you have two independent events happening at the same time what is the probability of it, that is what the joint probability give you let see how to do that.

The all the other quote remains the same, but you just need to add one more parameter called normalize is equal to true. If you set normalize is equal to True you are basically converting all the table values from numbers to proportion that is what normalize means. Now let us see how the output will look like. So, you have the same table here, but the values have been converted from numbers two proportions.

You can interpret the output like, the joint probability of the car having a manual gearbox and having the FuelType of CNG is only 0.01. But if you look at a value here that is 0.82 that represents that the joint probability of the car having a manual gearbox and the FuelType is also petrol the probability is really high there.

And if you see here there is no probability that you will get a car with an Automatic gearbox as well as with the FuelType CNG or diesel, but all these are from the data that we have read now. So, all these are interpretation that we have made are based on the data that we have now, there can be cases where the interpretations can be different with respect to different sets of records.

(Refer Slide Time: 14:30)

Now, we going to look at how to get the marginal probability using the two-way table. We are going to use a same function but by just tweaking or by just adding one more parameters we will be able to arrive at different types of probability values. So, here we are going to look at marginal probability. So, marginal probability is the probability of the occurrence of the single event, it will consider only the occurrence of a single event alone.
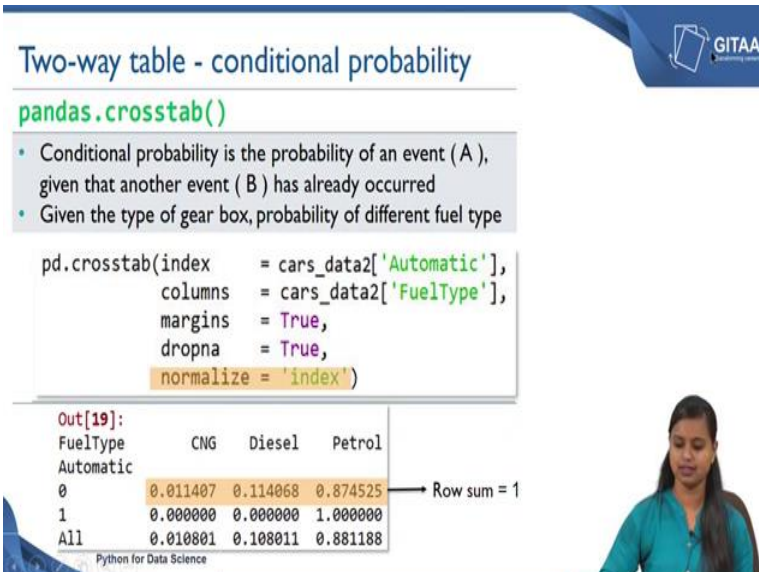
So, here is the code for that, we have used the same pd.crosstab function. So, here index and columns parameters remains the same and we have used dropna is equal to True and normalize is equal to True because we want all the table values in terms of proportions or the probability values. And I have also set margins is equal to True in order to get the

marginal probability value by setting margin is equal to True. You are basically going to get the rows sums and the column sums for your table values.

Let us see how the output will look. So, here is the output in the previous example you would have got till here you did not get the rows sum and the column sum of it. But by setting margins is equal to True you will get the row sums and the columns as well in the name of all. What is we mean by marginal probability the highlighted values are nothing, but the marginal probability values and how do interpret these values.

So, if you take the first value that is 0.946895. So, now, how can we interpret from the value 0.94 because the 0.94 value is nothing but the probability of cars having manual gearbox when the FuelType is of either CNG or Diesel or Petrol. You can infer the 0.88 value as the probability of the car having a FuelType as Petrol and when the gearbox type is of either Automatic or manual. So, this is what you can get and if you sum up everything the total probability value will be 1.

(Refer Slide Time: 16:44)

Now let us move on to get the conditional probability using the two-way table. So, here also we are going to use a same function that is panda's.crosstab and let us see what conditional probability is about. So, conditional probability is the probability of an event A, given that another event B has already occurred.

For example, if you want to get the probability of an event A; by considering another event has already occurred then you call it as a conditional probability. And now what is the example that we are going to look using conditional probability is that, given the type of gearbox what is the probability of different FuelType?

So, let us see how to get that, but if you see here the first four parameters remains the same, we have just tweaked the normalised parameter from we have just tweaked the normalized parameter we have initially said that as true, but we have changing it to index just to get the conditional probability values.

So, now we are going to look at the output to get the inferences. So, if you see here this is a cross tabulation of Automatic and FuelType variable and all the values are in terms of probability values. Since we have set normalize is equal to index you will get the row sum as one because that is what we mean by the conditional probability. So, given the gearbox types is of manual the probability of getting a CNG FuelType is 0.01 and the probability of getting diesel FuelType is 0.11 and the probability of getting FuelType as petrol is really high when compared to the other FuelTypes.

So, this gives you an idea about for any manual gear box petrol can be the FuelType because at the max we are getting the probability is for petrol. So, there is a really high probability value that you can get. So, from the high probability value you can say that so, from high probability value of 0.87 you can say that for any car which are of manual gearbox the probability is really high for having a petrol type for having a FuelType as petrol.

And similarly you can see here there is no property that you can get because there is no probability value that you can get for CNG and diesel. Because the probability is 0 and the probability is 1 for petrol because for all the Automatic gearbox cars the FuelType is only petrol. This we know that from the previous examples as well. So, this is how we get the conditional probability. Here we have got the rows sum to 1; we can also get the column sum to 1 in that case you will be looking at the cross table in terms of given the type of fuel being used for the car. So, given the type of fuel, you will get the probability of different gearbox types.

(Refer Slide Time: 19:42)

So, let us see how to get that. So, we are going to use a crosstab function to arrive at a conditional probability. So, I have initially set normalize is equal to index, but here I am changing it to column just to get the column sums as 1. But all other parameters remains the same.

Now let us try interpret from the output. So, given though FuelType of the car is CNG what is a probability of the car having a manual gearbox? It is 1. So, in this case there is 0 because we know that there is no Automatic gearbox which are of CNG or diesel FuelType. But the probability of getting a car given that the FuelType is petrol and the car has also manual gearbox is 0.93.

So, now we have seen how to get the conditional probability by considering the two variables because, we have also set the normalized parameter as columns and we have also seen how to set normalize is equal to index. And we have also seen how to interpret those results.

Next we are going to look at correlation because, till now we have been looking at to check the relationship between two categorical variables using cross tabulation. Now, we are going to look at how to check the relationship between two numerical variables used the measure called correlation. And what is correlation? Correlation is just to check the strength of association between two numerical variables and it need not be always numerical variables, but in this case or in this lecture we are going to look at the correlation for numerical variables.
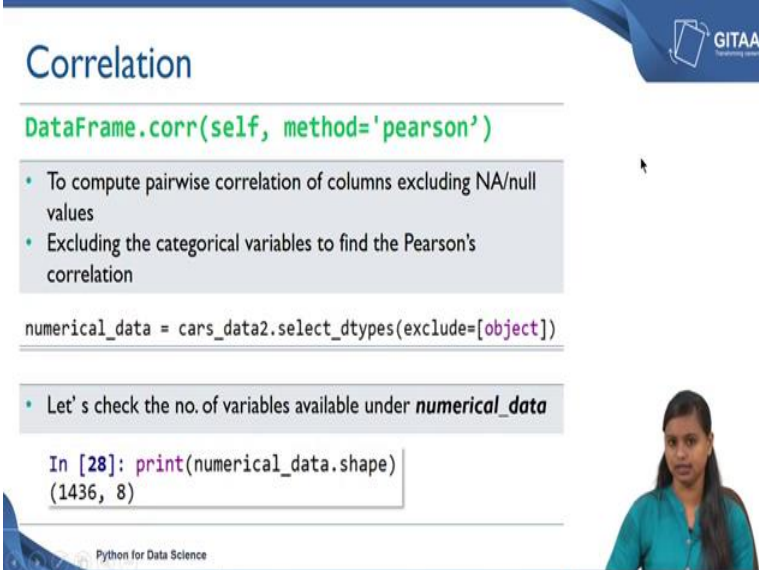
We are going to look at a visual representation of correlation using scatter plot. So, if you see here the first plot says positive trend because, as one variable increases the other variable is also increasing. For example, as your weight of the car increases the price might go up, in that case you call it is as a positive trend when you see the points are scattered. In this case and there is another plot that you can see here is little or no correlation.

Because there is no pattern that you can find out from this scatter plot here rather all the points have been scattered all over in that case you can say that there is no correlation between those two variables at all. Theoretically or numerically if you want to interpret from the correlation measure then the correlation values bounded between -1 and +1. And closer to one in either ns is the represent the higher correlation it can be either a negative or positive sign.

Because, the correlation can be there can be high correlation negatively and there can be high positively. If you want to interpret from the correlation measured in terms of numbers then the correlation value will be bounded between - 1 to + 1. 0 represents there is no correlation at all between any two numerical variables. And closer to 1 represents there is a strong correlation between two variables positively. Theoretically, above 0.7 we can say there is a fair correlation between two numerical variables.

If you can take it to the other side of it 0 to - 1 then closer to - 1 will give you high negative correlation like this whenever the age of the car increases the price will always decrease. Because for the newer aged car the price will always be really high, in that case there can be a strong negative relationship between those two variables so the value will be closer to - 1. So, now we have got an idea about what correlation measures is about and how we can interpret visually and how we can interpret numerically.
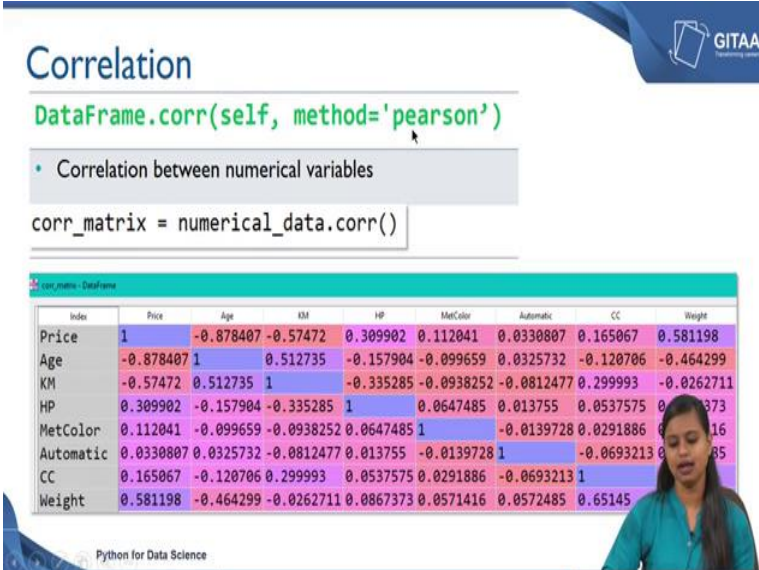
(Refer Slide Time: 23:42)

So, now we are going to see how to get the correlation using python, corr is the function that is used to calculate correlation between any variables that you can use that for a data frame. Because, by using the.corr function we are going to compute the pairwise correlation of columns by excluding all the null values here. Because, we are not just going to consider only two variables rather we are going to consider all the variables at a time and the function computes the pairwise correlation. We will see what pairwise correlation is when we get the output.

But this function is used to get the pairwise correlation of columns and by default it excludes all the missing values from the data frame and then it calculates the correlation value. And the method I have specified here is Pearson's because by default it calculates the Pearson correlation. And Pearson correlation is also used for to check the strength of association between two numerical variables.

If you have ordinal variables then you can go for other measures as Kendall rank correlation and Spearman rank correlation. So, in that case we need to exclude the categorical variables to find the Pearson correlation. Now let us see how to exclude those variables which are of categorical data type. So, here cars_data two is a data frame that I am working on from the data frame I am going to select only the columns which are of numerical data type.

Since I am just going to exclude only categorical variables I have given object under exclude. I have saved that to a new data frame called numerical data. So, let us see what would be the dimension of it by checking what is the number of variables that are available under numerical data; if you see if you print and see the shape of it you can look you can see that there are 1436 observation with eight variables. Initially we had 10, we have we are left out with only 8 variables now which are of numerical data type.

(Refer Slide Time: 25:45)

So now, let us see how to create the correlation matrix. So, we are going to look at the correlation between the numerical variables using the command.corr. And the data frame

that I am applying it to is the numerical data and I am saving this input to an output called corr matrix.

Now, we are going to calculate the correlation from the data frame that we have now using the function.corr here the interest is to find the correlation between the numerical variables. And the data frame that we are appling it here is numerical data because that is where we have all the columns as numbers.
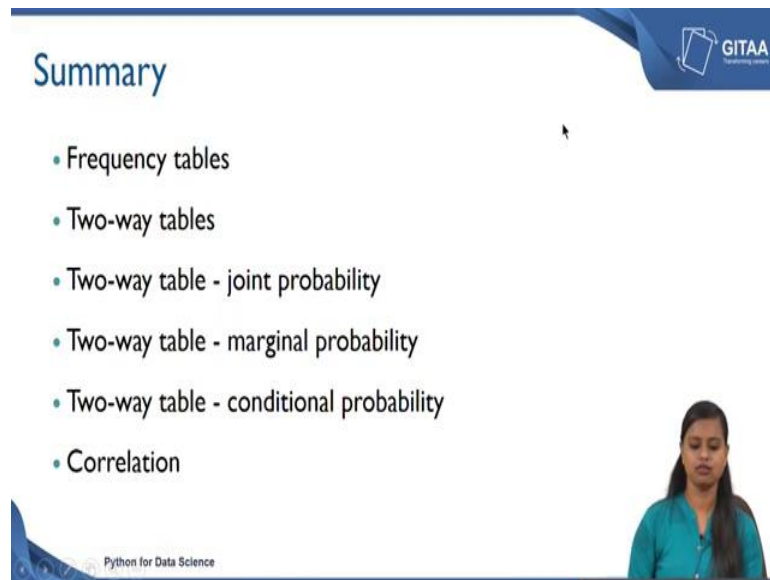
And we are also saving the output to an object called corr_matrix so that will have the output of the correlation matrix. So, now, let us look at the output and try to interpret the correlation values. So, this is a snippet that is taken from the variable explorer. And if you see here the index represents the variable names the variables in separate columns.

So, if you look at the principal diagonal which are marked in purple all are of one. Because the correlation between price and price would be 1; because the relationship is being checked against that we rebuilt itself that is why you are getting the value as 1 but if you see the value - 0.87 that represents that, there is a negative correlation between price and age. Since there is a negative symbol over there and the correlation is above 0.5; that means, that the correlation there is a strong negative correlation between age and price.

Whenever the age of the car increases, the price is decreasing that applies same to the kilometre though the correlation value is slightly lesser than kilometre it has 0.5, which is 0.57 which is equivalent to 0.6. It also have a fair negative correlation between kilometre and price whenever the car has travelled a lot the price will automatically go down. For a newer car and for the cars which have travelled really less in that case the price is always really high.

And if you look at 0.58 as the weight of the car increases the price is also increasing that is why there is a positive correlation value here that is 0.58. You can also look at the relationship between kilometre and age. There is a fair relationship positive relationship between kilometre and age because the values is 0.5 as the age of the car increases the kilometre is already increased. Similarly you can interpret from different values of different variables and here we have used the Pearson correlation. If you want to look out for other correlation measures you can specify that under the method argument.

(Refer Slide Time: 28:35)

So, now we have come to end of the session. So, let summarize whatever we have seen it in this lecture. We have started with creating frequency tables to check what is the frequency of each categories in the categorical variable. And then we were interested in looking at the relationship between two categorical variables using a two-way tables. And then we have also seen how to convert the two-way table into joint probabilities, marginal probabilities and conditional probability. And we have also seen how to check the relationship between two numerical variables using a measure called correlation.

Thank you.