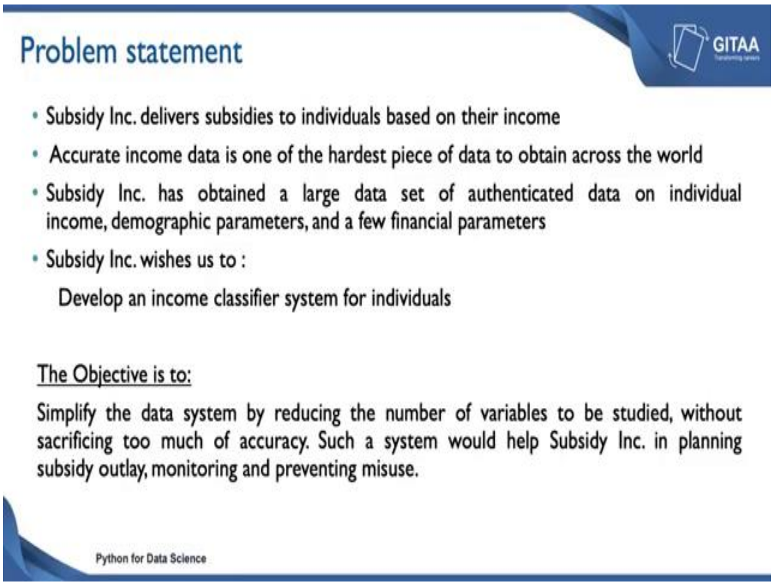
Python for Data Science
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture - 40
Classifying Personal Income

Hello all, welcome to the lecture on the case study. So, in this lecture, we are going to look at a case study called Classifying Personal Income, where we are going to see how to solve that problem using python.

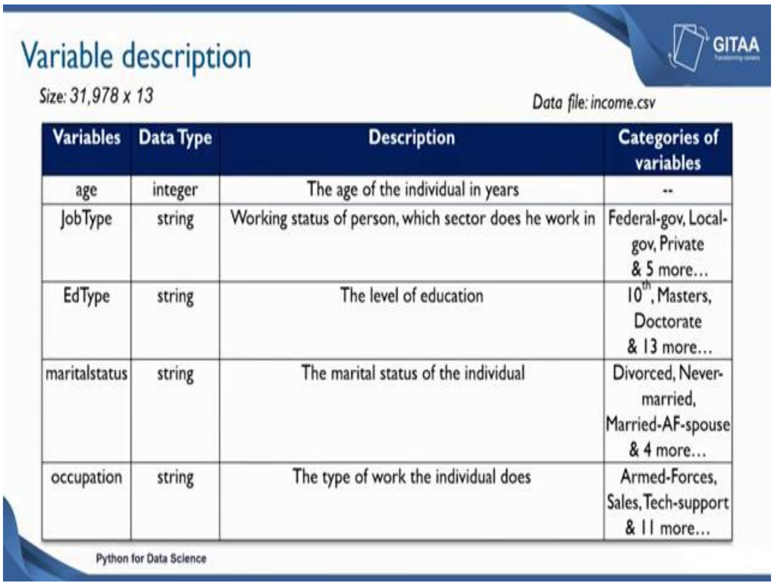
(Refer Slide Time: 00:27)



So, let us look at the problem statement of the case study that we are going to solve. Subsidy inc. is a company which delivers subsidies to individuals based on their income. And accurate income data is one of the hardest piece of data to obtain across the world. So, they have obtain the large set of data on individual income, demographic parameters, and based on few financial parameters. So, they wish us to develop an income classify system for individuals.

The main objective of the case study is to simplify the data system by reducing the number of variables to be studied without sacrificing too much of accuracy, so that the data collection part would be easier. If we have very few parameters that needs to be measure, then the data collection part will be easier. So, that such a system would help subsidy income in planning subsidy outlay monitoring and preventing misuse. Let us quickly look at the variables that are available in the case study that we are going to solve.

(Refer Slide Time: 01:34)

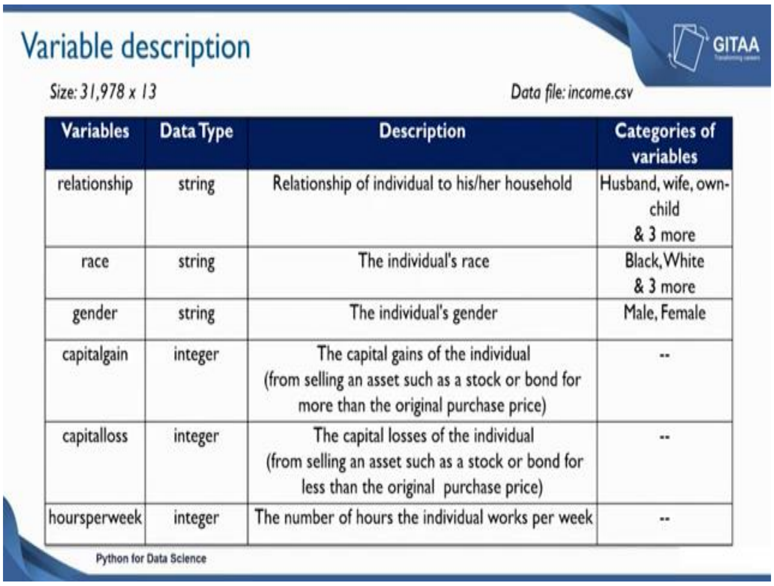


Variables	Data Type	Description	Categories of variables
age	integer	The age of the individual in years	--
JobType	string	Working status of person, which sector does he work in	Federal-gov, Local-gov, Private & 5 more…
EdType	string	The level of education	10^{th}, Masters, Doctorate & 13 more…
maritalstatus	string	The marital status of the individual	Divorced, Never-married, Married-AF-spouse & 4 more…
occupation	string	The type of work the individual does	Armed-Forces, Sales, Tech-support & 11 more…

The first variable is age which represents the age of the individual in years. Next one is a JobType which represents the working status of a person which sector does he work in like federal government local government and so on and so forth. Next is the educational type the level of education like 10^{th}, masters, doctorate and so on. Next one is the maritalstatus the maritalstatus of the individual, whether a person is married the worst or never married and so on.

Next one is the occupation the type of work the individual does whether the individual is working in an armed forces or sales or technical support and you have eleven more categories like that.

(Refer Slide Time: 02:20)

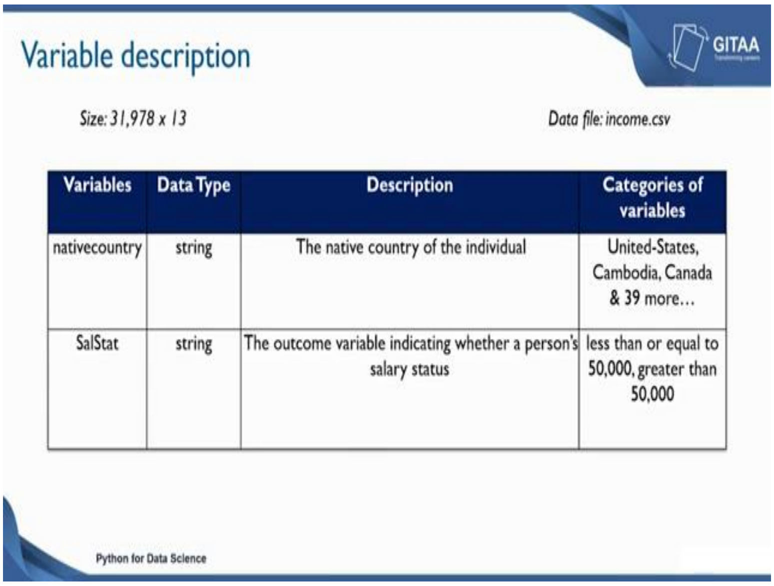


Variables	Data Type	Description	Categories of variables
relationship	string	Relationship of individual to his/her household	Husband, wife, own-child & 3 more
race	string	The individual's race	Black, White & 3 more
gender	string	The individual's gender	Male, Female
capitalgain	integer	The capital gains of the individual (from selling an asset such as a stock or bond for more than the original purchase price)	--
capitalloss	integer	The capital losses of the individual (from selling an asset such as a stock or bond for less than the original purchase price)	--
hoursperweek	integer	The number of hours the individual works per week	--

Next one is relationship of the individual to his or her household. The next one is the race; it is being represented by black, white and 3 more. Next one is the gender the individual's gender being represented as male and female. The next one is the capitalgain; the capital gains of the individual basically from selling an asset such a stock or bond for more than the purchase price.

Next one is the capitalloss the converse of that, the capital losses of the individual from selling an asset such as a stock or bond for less than the original purchase price. Next one is the hoursperweek, the number of hours that the individual works per week in a company.

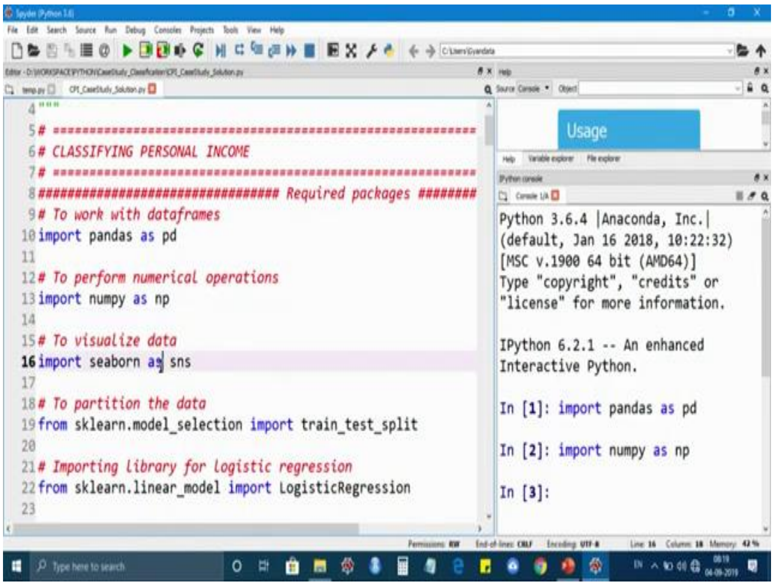
(Refer Slide Time: 03:03)



And we have nativecountry of the individual being represented with like categories like United States, with countries like United States, Cambodia, Canada and so on and so forth. Next one is the last one is the salary status which is the outcome variable. The outcome variable indicating whether a person salary status, whether it is a less than or equal to 50,000.

So, it has basically two categories less than or equal to 50,000 and greater than 50,000. So, this is the overall idea about the problem and the description of the variables. Now, we are going to go back to spyder and we are going to see how to solve this problem.

(Refer Slide Time: 03:45)

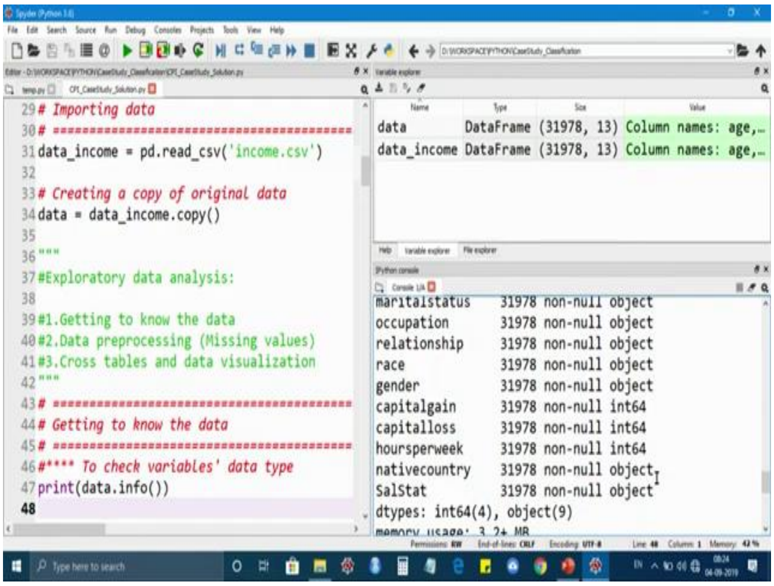


So, now, let us start by importing required packages. To work with data frames, we are going to import pandas as pd. And to perform any numerical operations on it, we are going to import numpy as np. Followed by that, we will be visualizing the data using the seaborn library. So, we are going to import seaborn as sns. Followed by that, we are going to import train test split from the package called sklearn and the model_selection is a sub package under sklearn.

So, we are going to import train_test_split from sklearn package. After that we are going to import LogisticRegression from sklearn linear_model. And to look for performance matrix, we are going to import accuracy and confusion matrix from sklearn matrix. So, now, we have imported the required packages. Now, let us import the data into spyder. So, income is the filename and it is in csv format.

So, now let us import the data into spyder. Income is the filename and it is in csv format, we going to read that and we have storing it onto a data frame call data_income. So, once we read the data, we can see it under the variable explorer.

(Refer Slide Time: 05:24)



So, the data underscore income is of data frame and the size of the data frame is 31978 observations with 13 columns. After reading data, we are going to create a copy of the original data, so that original data frame will not be touched. And further analysis will be made on the copy data frame. So, we are going to create a copy using '.copy' function and we are storing it on to a data frame called data new data frame now. So, now, we have read the data.

Now, it is time to explore the data to understand the data even more better. So, under exploratory data analysis, we are going to broadly look into three topics the first one is getting to know the data, where we basically see what type of; what type of variables that we are going to deal with. The next thing that we are going to see here is data pre processing, where are we going to deal with missing values like how to identify the missing values and how are we are going to deal with that.

And after that we are going to understand the data through visualization and cross tables, because we will be looking into checking the relationship between the variables using cross tables and data visualization. So, basically once we write the data we would be interested in getting the data type of each variable, so that it gives an idea about what type of data that we are going to deal with. So, '.info' command on a data frame gives you the data type of each variables.

But if you check here all the variables are read with expected data type, because age is read as integer, JobType, from JobType to gender, it has been read as object because those have only categories to with and capitalloss, capitalgain, hoursperweek have been read as integer, because those variables have numerical variables. And the SalStat nativecountry has been read with object data type, because we know that their exist categories under that variable. So, it has been read as object.

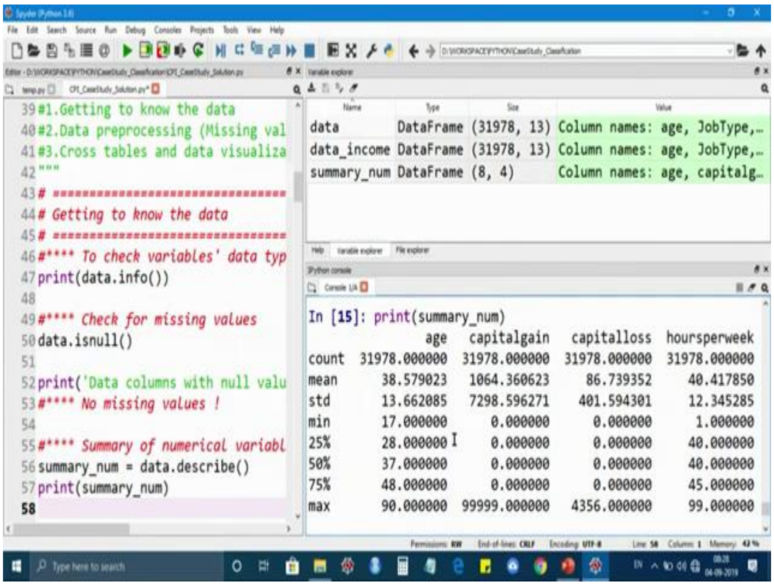
So, now, we have seen what are the data types of each variables and we know that all the variables are read with the proper unexpected data type. So, now, we will check whether there are any missing values in the data frame using a isnull function. So, let us check whether there are any missing values. So, basically a isnull returns Boolean values that is true and false true indicates the missing data and the false indicates there are no missing data in a particular cell.

But if you see the output, it will be difficult to skim through the whole output. So, let us take the sum of missing values in each column, so that we will get an idea about how many values or how many cells are missing under each variable. So, this can be done by adding dot sum to the existing command. So, let us see how to do that. So, this is just the print statement, but the actual command is here data.isnull.sum.

So, if we execute this line, you will get an output which says no columns have missing values, because the corresponding values of each variable are 0. So, it represent that there are no missing values under any of the variables. So, now we have checked whether there any missing values in a data frame or not. Now, it is time to understand the data by getting the descriptive statistics out of it.

In our data frame, both numerical and categorical variables are there. So, the descriptive statistic can give you great insights into the shape of each variable, so that describe function can be used to get the basic descriptive statistics out of data. So, I am going to use the dot described function on a data frame data. And I am storing that output to an object call summary_num, because by default the describe function gives you eight statistical properties of numerical variables.

(Refer Slide Time: 09:47)



The first one being count mean standard deviation minimum 25 percent, 50 percent, 75 percent and the maximum value. Basically the count their count represent the count of observations under other particular variable that is age. When we look at the mean of age, the average age of the individual is turned out to be 39 years. And the standard deviation is around 14 and the minimum age of the individual is 17 years; the next is 25 percent that is 25 percent of the individuals age is less than 28 years.

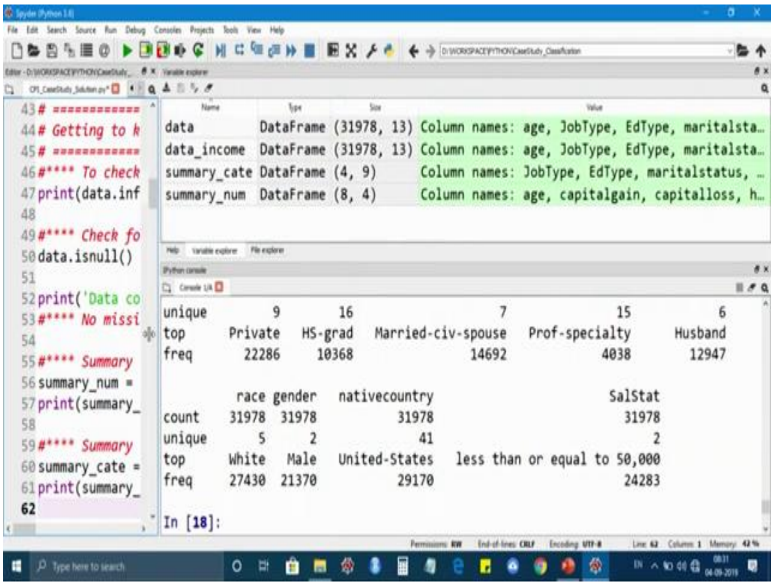
The next is 50 percent; the median age is 37 years here. Next is 75 percent, 75 percent of the individual age is less than 48 years, at last you have maximum value we know that at the max the age of the individual is 90 years. Similarly, if you look at the capitalgain, it is a profit from the sale of property or an investment, so whether the sale price exceeds the purchase price.

So, if you look at the capitalgain, it is a profit from the sale of property or an investment, where the sale price exceeds the purchase price then only we call it as a capitalgain. If you see on an average the capitalgain of the individual is 1069 and the standard deviation is really high that is 7298. And if you look at the maximum value, it is 99999, but the minimum value is 0.

So, in this case, it turned out that only 25 percentage of the capitalgain is greater than 0 that is why you have nothing under from minimum to 75 percent age, because it is very evident that very few people will be investing in stocks or any other investment, so that they get some profit out of it. So, conversely a capitalloss arises, if the proceeds from the sale of a capital or a asset are less than the purchase price, it is turned out that only 25 percent of the capitalloss is greater than 0, though the maximum values 4356, the minimum value is still 0.

So, similarly you can look at the spread of the hoursperweek, like on an average individual spends 40 hoursperweek in a company. Similarly, you can look at the other statistics. And we also have categorical variables in our data frame. On that note, we would be interested and getting the frequencies of each categories under a variable. So, the same command is being used by setting include is equal to o, o represents object. So, let us see what the output gives us. And we are storing that onto a object call summary_cate that represents summary for categorical variables.

(Refer Slide Time: 12:59)



So, basically when you set when you set include is equal to o, it gives you four measures starting from count, count basically gives you how many observations are considered while giving you the summary and then unique, unique represents, how many unique categories are available under that particular variable. For example, under JobType, we have nine unique categories. And next is the top which gives you the model value of the variable that is the most frequently occurring category.

So, for JobType the most frequently occurring category is private. And you can also get the corresponding frequency. Here 22286 observations correspond to private category.

Similarly, you can look at the frequencies of the categories that are available under other variables. But from this output, we now just know how many unique values are there, but we would be interested in getting the unique categories under a variable, so that it gives us an idea what are the categories that are there in a particular variable.

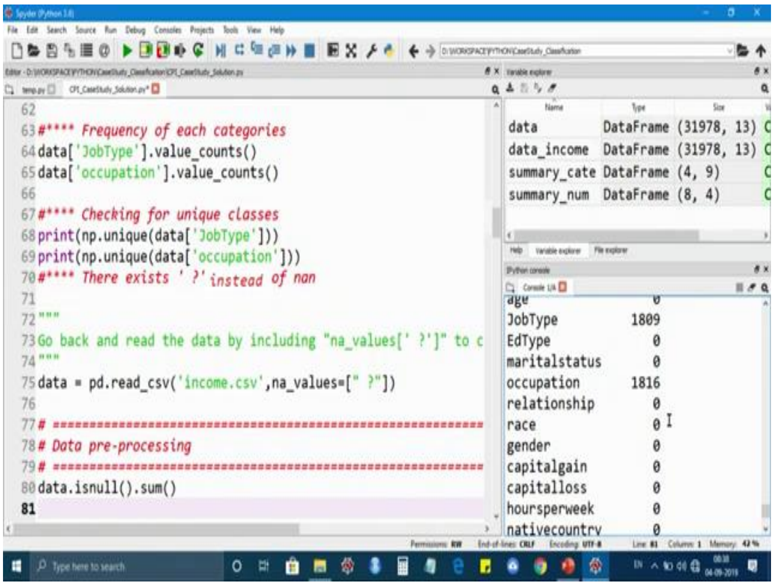
So, for that case, in that case, the value_counts gives us the better representation. So, let us see how to use the value underscore function. So, I am just going to take the first variable that is JobType. I am going to get the frequencies of each categories that are available under JobType. So, basically this gives us the better representation by listing all the available categories under frequencies. If you look at here, it basically gives you the frequencies of each categories under JobType. You can also see there exist question mark instead of nun.

So, there exist a question mark instead of nan, because by default python only reads the blank cells as nan not other special characters. In this case, we have a special character, but it is not being read as nan, but we have encounter that there are some missing values in the form of question marks. Similarly, we can also get the frequencies of categories using the value_counts for other variables, but it turned out that occupation also has question marks here. But when we looked at the data, there were only question marks under two variables that is occupation and JobType.

So, to basically see how exactly the special characters are, we can use the unique function. So, let us check the unique classes of JobType. So, basically unique function is from numpy library and inside the function as specify the JobType from data frame data. So, if you see that there exists the special character which is question mark there is the space before the question mark, this is how exactly the levels of the JobTypeare being entered.

So, if you see here before the starting letter of the every level, there is a space in front of it. So, similarly we have a space in front of question marks also, this is how exactly the special character is been entered here. So, now, we know that there are some special characters which is just the representation of missing values. So, what we are going to do here is, we are going to go back and read the data by including the na values to consider the special character as nan. By adding na_values with a list of a string values we are going to consider this special character as nan values.

(Refer Slide Time: 16:59)

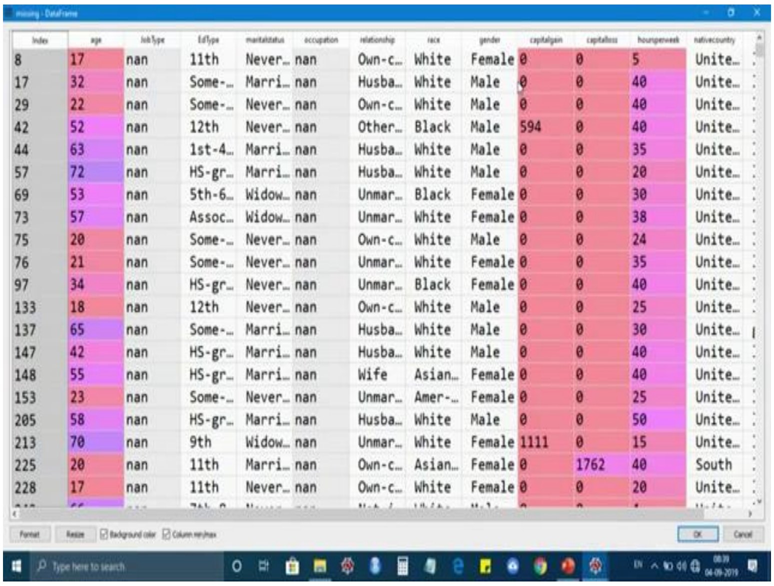


So, once we have executed this data, you will be able to see the data frame size is 31978 observation with 13 variables. Now, we have seen how to consider other special characters as nan values. Now, it is time to pre process the data. Basically now we know that there are some missing values in a data frame to deal with the missing data, we need to first identify the missing data and then we can examine the missing value pattern.

So, by using the same isnull function, we found out that 1809 cells are missing under JobType and 1816 cells are missing under occupation and no other variables have missing values except JobType and occupation. So, before deciding on how to deal with the missing data, let us see in a particular row either one of the column is missing or both the column values are missing. For that let us subset the rows at least one column is missing in a row.

So, here we going to subset the rows by giving .any_axis is equal to 1, so that it considered at least one missing column in a particular row. Now, we have subset the rows with missing values. If you look at the size of the data frame, it is 1816. It is with 1816 rows with 13 variables.

(Refer Slide Time: 18:32)

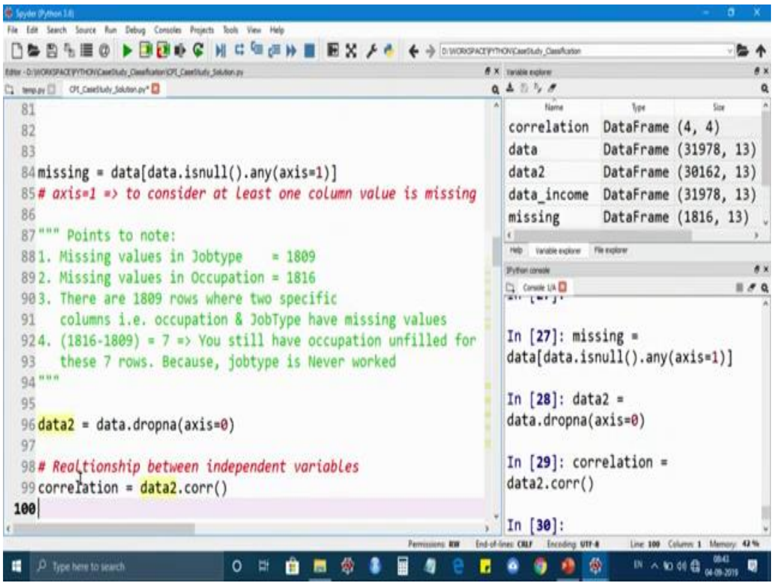


Index	age	JobType	EdType	maritalstatus	occupation	relationship	race	gender	capitalgain	capitalloss	hoursperweek	nativecountry
8	17	nan	11th	Never…	nan	Own-c…	White	Female	0	0	5	Unite…
17	32	nan	Some-…	Marri…	nan	Husba…	White	Male	0	0	40	Unite…
29	22	nan	Some-…	Never…	nan	Own-c…	White	Male	0	0	40	Unite…
42	52	nan	12th	Never…	nan	Other…	Black	Male	594	0	40	Unite…
44	63	nan	1st-4…	Marri…	nan	Husba…	White	Male	0	0	35	Unite…
57	72	nan	HS-gr…	Marri…	nan	Husba…	White	Male	0	0	20	Unite…
69	53	nan	5th-6…	Widow…	nan	Unmar…	Black	Female	0	0	30	Unite…
73	57	nan	Assoc…	Widow…	nan	Unmar…	White	Female	0	0	38	Unite…
75	20	nan	Some-…	Never…	nan	Own-c…	White	Male	0	0	24	Unite…
76	21	nan	Some-…	Never…	nan	Unmar…	White	Female	0	0	35	Unite…
97	34	nan	HS-gr…	Never…	nan	Unmar…	Black	Female	0	0	40	Unite…
133	18	nan	12th	Never…	nan	Own-c…	White	Male	0	0	25	Unite…
137	65	nan	Some-…	Marri…	nan	Husba…	White	Male	0	0	30	Unite…
147	42	nan	HS-gr…	Marri…	nan	Husba…	White	Male	0	0	40	Unite…
148	55	nan	HS-gr…	Marri…	nan	Wife	Asian…	Female	0	0	40	Unite…
153	23	nan	Some-…	Never…	nan	Unmar…	Amer-…	Female	0	0	25	Unite…
205	58	nan	HS-gr…	Marri…	nan	Husba…	White	Male	0	0	50	Unite…
213	70	nan	9th	Widow…	nan	Unmar…	White	Female	1111	0	15	Unite…
225	20	nan	11th	Marri…	nan	Own-c…	Asian…	Female	0	1762	40	South
228	17	nan	11th	Never…	nan	Own-c…	White	Female	0	0	20	Unite…

So, now let us open the missing data frame and inspect the missing data. So, it is very clear that whenever JobType is missing, the occupation is also missing, but the total number of missing values is 1816 rows. So, there is a special category called never worked and the corresponding value of the occupation is nan. Since the JobType is never worked.

So, if you saw the JobType, so if you saw the JobType there is a category called never worked. If the individual has never worked, then occupation could not be fill that is the reason you have nan values under occupation. So, now let us go back to the script in window and see and what we going to do with it.

(Refer Slide Time: 19:27)



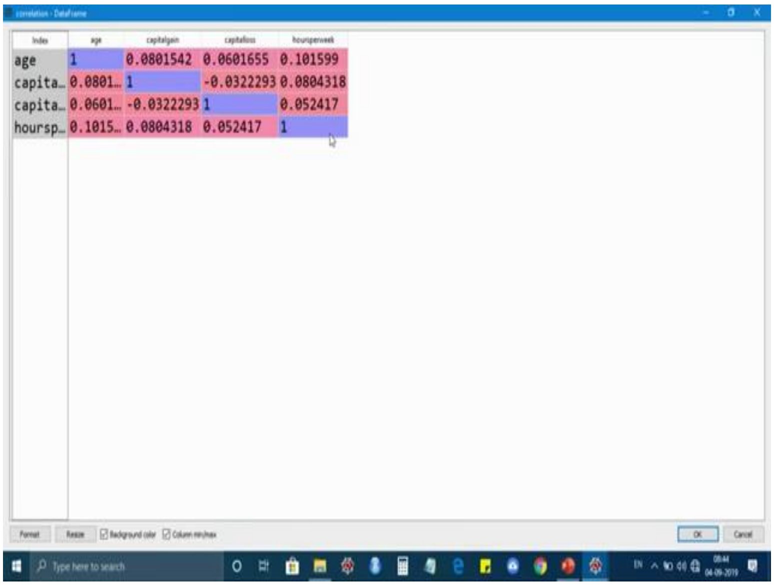
So, let us quickly note some points that we have got from the missing data. So, missing values in JobType is equal 1809 rows. So, there are 1816 rows that are missing under occupation and there are 1809 rows where two specific column that is occupation and JobType have missing values. So, there is the difference between 1816 and 1809 that is 7 rows. You still have occupation and fill for those seven rows because JobType is never worked that is why the total number of missing values is 1816.

Now, in this case we can delete the cases containing the missing data or replace the missing values with reasonable alternative data values. If the data are not missing at random, we have to model the mechanisms that produce missing values as well as the relationship of interest which is very complex in these contexts. So, here we are going to remove all the rows with missing values in consider only the complete cases alone. So, in that case, we are going to use a command called dropna. And by setting axis is equal to 0, we are dropping out all the rows wherever there are missing values.

So, if you look at the size of the data to data frame that is 30162 rows with 13 variables; that means, that we have gotten read of 1816 rows wherever they were missing values, why because we could not find out the mechanism that goes behind which produce the missing values and we do not know the relationship of interest as well. So, we are going in this session, we are going to remove all the missing values and we are going to continue with the further analysis.

So, with complete cases let us look at the relationship between numerical variables that is between independent variables using the correlation measure. So, you can get the correlation, you can get the pair wise correlation using the .corr function using the corr function. So, and I save my output to a variable called correlation onto an object check called correlation.

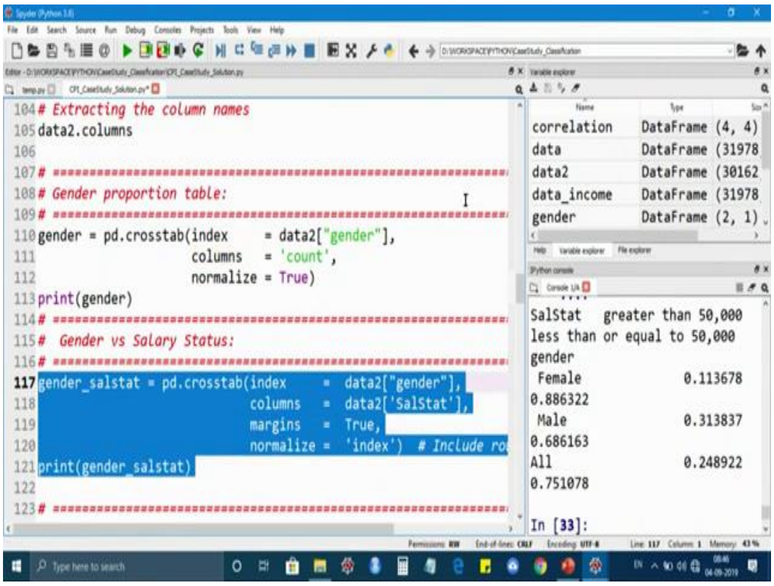
(Refer Slide Time: 21:34)



So, let us look at the correlation values. So, if you look at the correlation values, none of the variables correlation, so none of the values are nearer to 1; most of the values are nearer to 0. It represents that none of the variables are correlated with each other because the correlation values lies from -1 to +1. And if it is closer to 1, we say that there is a strong relationship between two variables; and if it is closer to 0, then we say that there is a there are no a little correlation between variables.

So, in our case, none of the variables are correlated with each other. So, now, we will consider the categories categorical variables to look at the relationship offered.

(Refer Slide Time: 22:36)

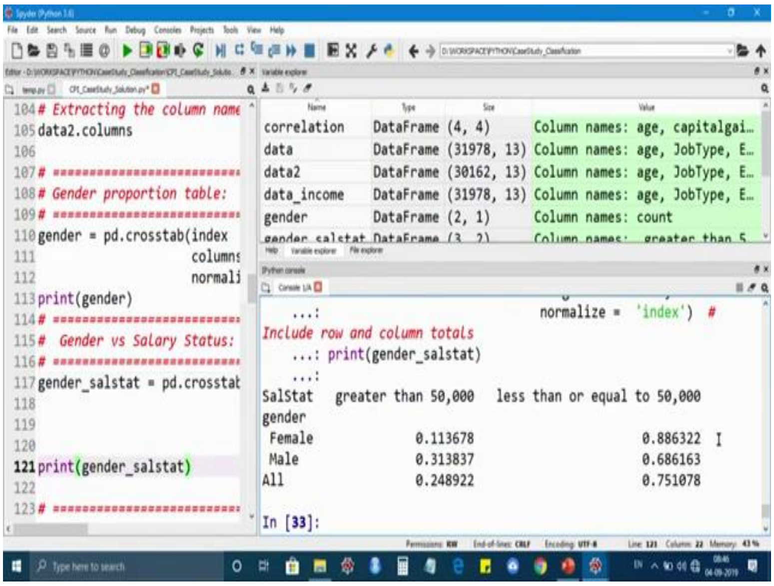


So, now we will look at the gender proportion using the cross table function. So, here I am accessing the columns of data2 basically it gives you the column names of each variable, so that you can use the column names in the further analysis. So, here we are going to look at the gender proportion using the cross table function. And by setting normalise=true, you will get the proportion table.

And you will give the variable of interest under the index column. And let us look at the output. Now, if you see here male are high in frequency that is 67 percent corresponds to male and only 33 percent corresponds to female. So, now we have got an idea about what is the proportion of gender that we have. Now, it is time to check how salary status varies across the gender.

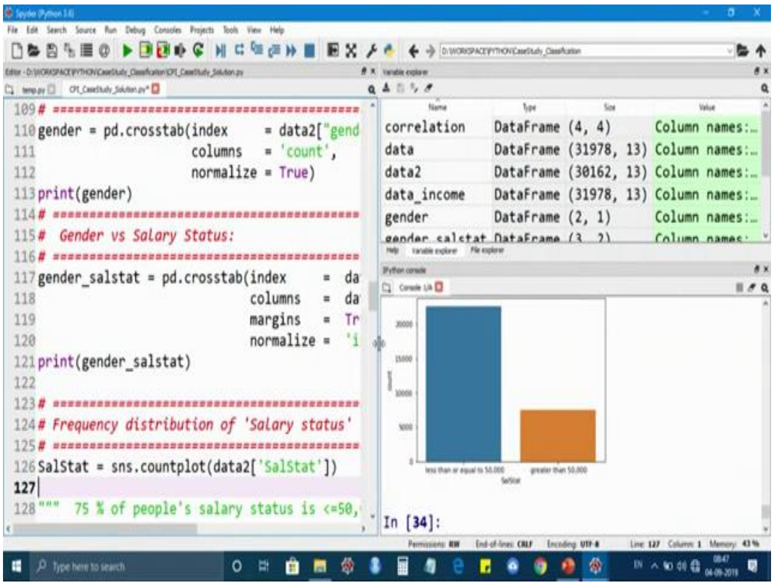
So, we are going to look at the two-way table where we are going to check the relationship between gender and salary status. And in a row I want gender, and in columns I want; and in rows I want gender, and in columns I want salary status. And by setting normalised is equal to index, I am going to get the row proportion equals to 1.

(Refer Slide Time: 23:57)



So, from the output, it is very clear that only 11 percent of the female earn greater than 50000 US dollars and 89 percentage earn less than 50000 US dollars, but men are more likely to earn more than when compared with women. On classification problems, you need to know how balance the class values are. For example, you need to know the proportion of your output variable. So, let us visually look at the frequency distribution of the salary status using the count plot function from seaborn and the variable of interest is SalStat, because we are going to look at the bai plot of salary status.

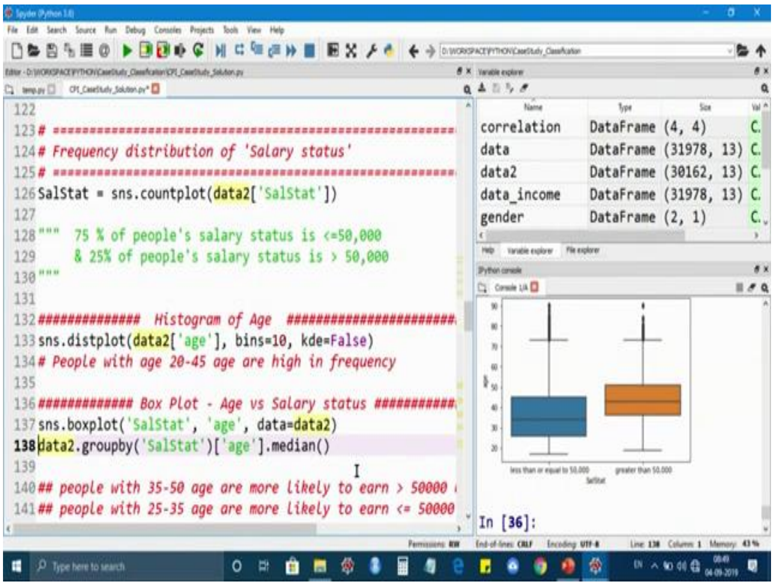
(Refer Slide Time: 24:40)



It is very clear that around 75 percentage of the observation corresponds to less than or equal to 50000 and only 25 percent corresponds to greater than 50000. So, now, let us plot the histogram to check the underlined frequency distribution of the age variable. So, I am using I am going to plot histogram using this plot from the sea born library and for the age variable. And by setting bins is equal to 10, I will have only ten bins to interpret from and I am setting kde is equal to false, so that I will have the frequency is on the y-axis.

So, from the histogram, it is very clear that people with age 20 to 45 are high in frequency and we also have the records for the other ages, but between 20 to 45, the frequency is high. Similarly, we can generate cross tables and plots to understand the relationship between other variables. So, if you want to do a bivariate analysis to check how age is affecting the salary status, then we can do a box plot to check how salary is wearing across age.

(Refer Slide Time: 26:05)

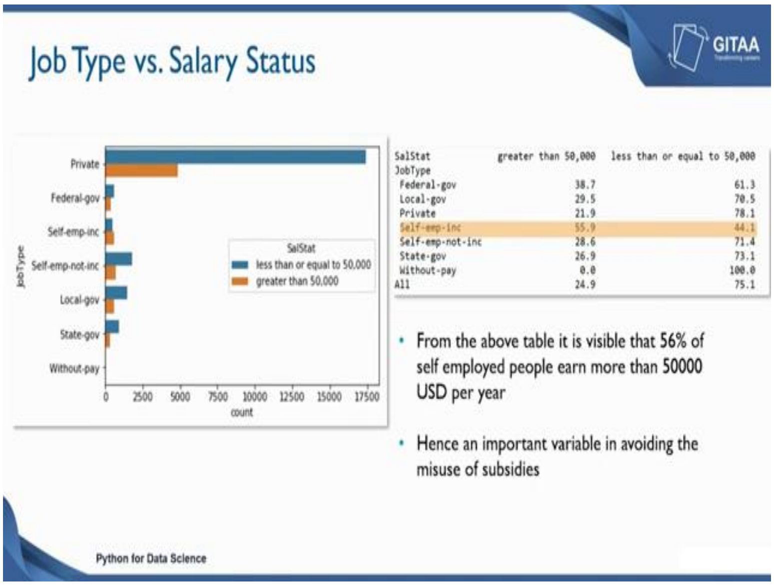


So, people with 32, so people with 35 to 50 age or more likely to earn more than 50000 US dollars, but people with 25 to 35 age are more likely to earn less than 50000 dollars. So, similarly you can generate the cross tables and plots to understand the relationship between other variables.

(Refer Slide Time: 26:27)

Using python we will look at the relationship between the variables using cross tables and a visualization. So, now, I am going to quickly take you through the exploratory data analysis through slides, where I am going to show you the relationship between variables using the same commands that we have used in the python.

(Refer Slide Time: 26:50)

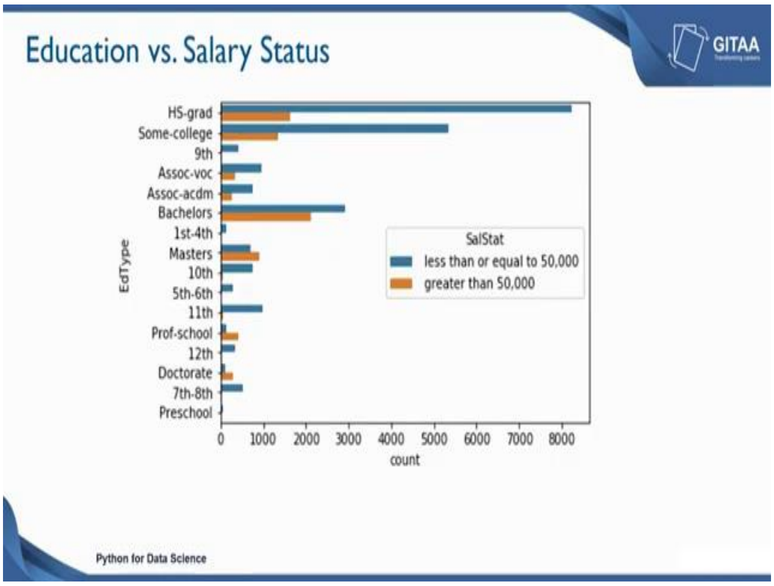


So, first we are going to look at the JobType versus salary status. So, first we are going to look at the relationship between JobType and SalStat. So, here is the bar plot which gives you an idea about what is the count, what is the frequency of each category it is available under a JobType. So, for example, if you see here most of the individuals work in a private organisation and very few work in the state government or federal government organisation.

This is the frequency distribution of the JobType. But if you want to see how salary is wearing across the JobType, then we can look at the cross table. And I have highlighted only a single row which will give you an idea about what is the percentage splits for an individual. If he is if he is self employed then what is the percentage spilt that, he will earn greater than 50000 or less than or equal to 50000. For the greater than 50000 it turned out to be 56 percent and for the less than or equal to 50000 it is 44 percent.

And for the rest of the JobType more or less 70 more or less 70 percentage of the individuals earned less than or equal to 50 and the rest of the percentage belongs to greater than 50. Only for the self employed people there is an equal proportion more or less equal proportion for both the categories. And on the whole from the above table, it is visible that one whole if you look at the JobType, it will be an important variable in avoiding the misuse of subsidies.

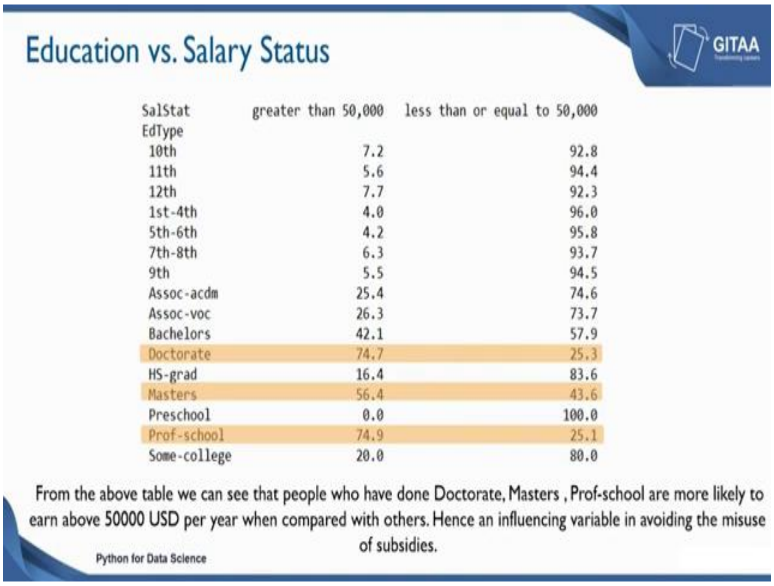
(Refer Slide Time: 28:23)



So, next we will look at a variable call education. Let us just look at the frequency distribution of education. Education basically has too many levels to it starting from preschool to doctorate. And if you look at the frequencies of each categories only the higher secondary graduation has the highest frequency. People most of the people have finished the higher secondary graduation, and followed by some college.

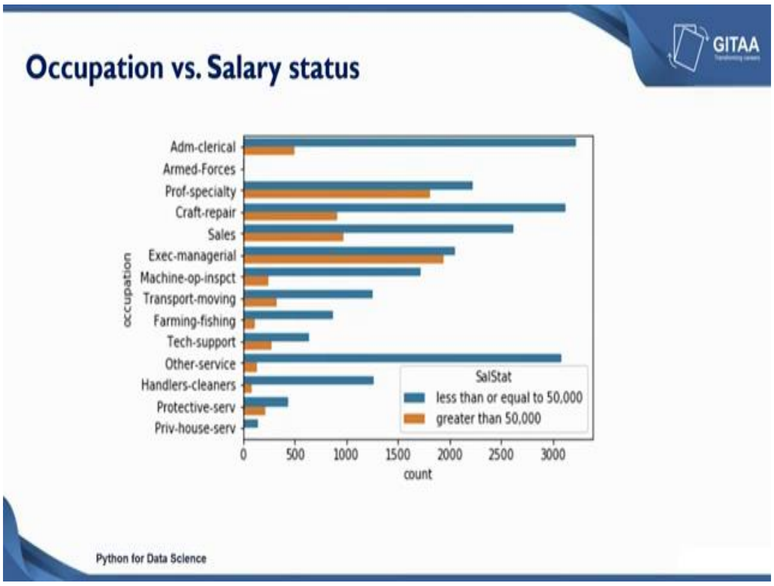
And if you see here the blue bar is the representative of less than or equal to 50 and greater than 50. So, now let us look at the relationship that exists between the salary status in education to see how the salary state is varying across the education type.

(Refer Slide Time: 29:10)



So, we are going to use the cross table for that. So, from the table it is, so from the table we can see that people who have done doctorate, masters and professional school are more likely to earn above 50000 US dollars per year when compared with the other education type. So, that it and hence it can be an influencing variable in avoiding the misuse of the subsidies.

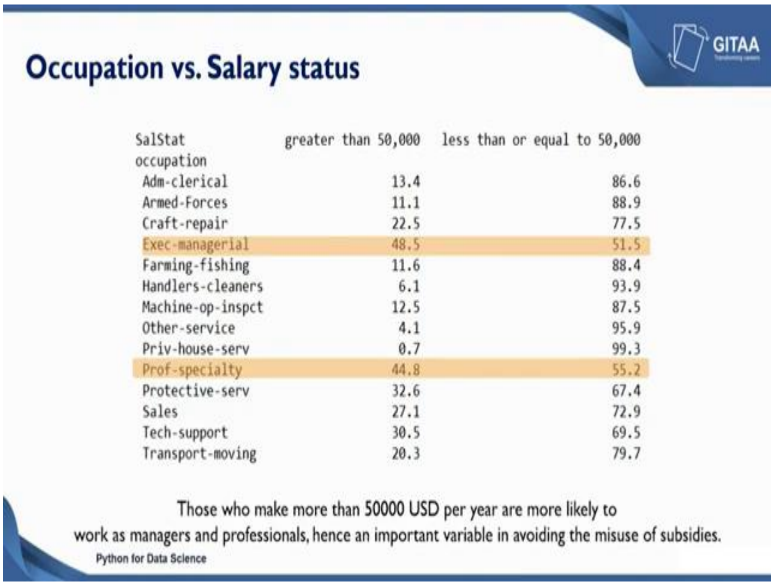
(Refer Slide Time: 29:36)



Next we will look at a variable called occupation. This is the bar diagram of the occupation variable. And if you look at the categories under the occupation, it has so many to it starting from administrative, clerical and it has so many categories like armed forces, farm, fishing, technical support and other services.

If you try to interpret how salary status is dependent on occupation, then there is a clear demarcation because there is no equal length of bars for both the colours like for less than or equal to 50 or greater than 50 for each and every level of the occupation the salary status, the proportion or the frequency of salary status is differing. So, let us get an a clear idea using the cross table.

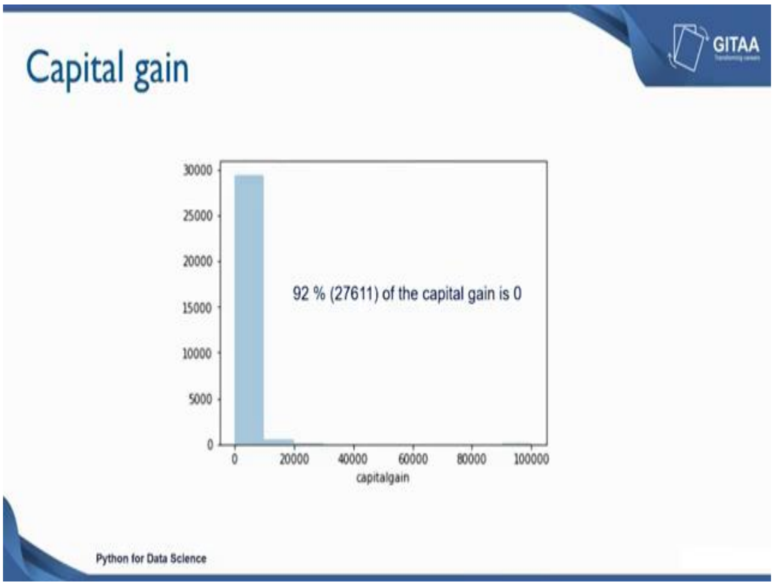
(Refer Slide Time: 30:21)



So, here is the cross table for occupation versus salary, so that we will get an idea whether the salary status is dependent on occupation or not. So, from the output, I have highlighted two levels here, one is executive managerial and the other one is. So, I have highlighted two levels; one is executive managerial and the other is prof specialty.

And those who make more than 50000 US dollars per year are more likely to work as a managers and professionals. Hence it can be an important variable in avoiding the misuse of subsides. So, let us see how this variable is impacting the salary status when we build the model.

(Refer Slide Time: 31:06)

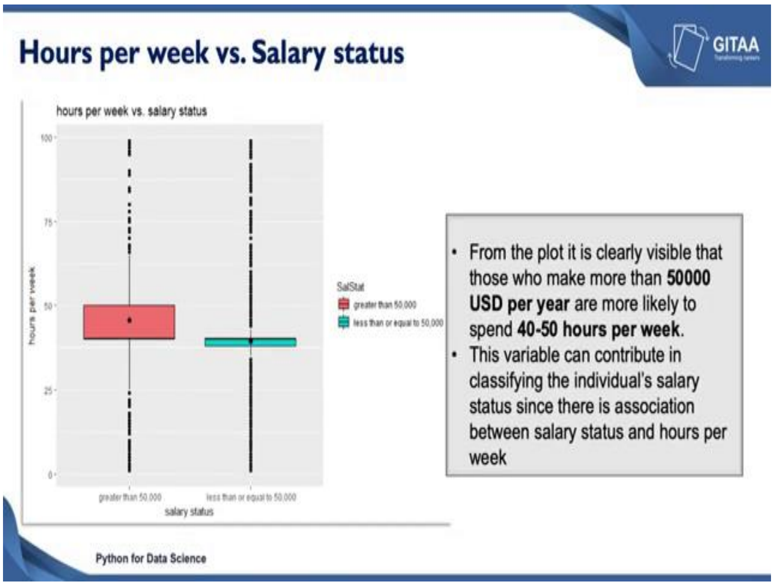


Capitalgain is one of the variable that we have from the data frame which is an important variable. Because if the capitalgain of the individual is high, then the salary status could be high. So, this could be one of the important variables to classify the individuals salary status and the plot shows the frequency distribution of the capitalgain variable.

Though the capitalgain is ranging from 0 to 1 lakh, though the capitalgain is ranging from 0 to 1 lakh, but there are more observations in between the bins 0 to 20000, so that means, 92 percentage of the capitalgain is 0 which corresponds to 27611 observation, and only 8 percentage of the people have gained something from selling their asset or a gain profit out of their investment. So, this could be one of the important variable to classify the individual salary status.

So, let us see whether capitalgain is an important variable when we build the model capitalloss values are also ranging from 0 to 4000, but 95 percentage of the capitalloss is 0. So, in our records either we have the records for capital gain or loss because either they will loosed or gain from the investment. And if you see 28721 individual's capitalloss is 0, so either they would have not invested or the capitalloss is 0 for them or the capitalloss is still 0 that could be the reason because they have not invested or they have not loss anything from their investment. So, let us see whether this will be an important variable when we try to; when we try to include this in a model.

(Refer Slide Time: 32:58)



The next one is the hours spent variable. So, the plot shows the relationship between the salary status and hour spent using the box plot. On the x-axis, I have the salary status; on the y-axis I have the hours per. And on the x-axis the salary status is represented as greater than 50 and less than or equal to 50. And the plot shows the relationship between that the relationship that exists between the hours spend and the salary status.

Here we are going to check how the salary status is vary with respect to the hour spend by the individual in a company. And if you see here from the plot it is clearly visible that those who make more than 50000 US dollars per year or more likely to spend between 40 to 50 hoursperweek on an average. And others variable can contribute in classifying the individual’s salary status since there is an association between salary status and hour per week.

Because when you consider the less than or equal to 50000 category, the median is very low when compare to the greater than 50000 category. And the minimum hours spent and the maximum hours spent by the individual is also very low that could also be the reason that is why they are earning less than 50000 US dollars per year. But all these interpretation that we have made are from the data that we have, it might decorate with respect to different sense of data.