**Adult Salary Prediction**

1. Problem Definition

In this project, We aim to Predict whether income exceeds 50K/yr

based on census data. The data has been downloaded from the UCI Repository website ([Adult](https://archive.ics.uci.edu/ml/datasets/Census+Income)). We implemented the Artificial Neural Network (ANN) on Python to solve this problem.

The data contains the following culumns:

The dataset provides 14 input variables that are a mixture of categorical, ordinal, and numerical data types. The complete list of variables is as follows:

* Age.
* Workclass.
* Final Weight.
* Education.
* Education Number of Years.
* Marital-status.
* Occupation.
* Relationship.
* Race.
* Sex.
* Capital-gain.
* Capital-loss.
* Hours-per-week.
* Native-country.

The dataset contains missing values that are marked with a question mark character (?).

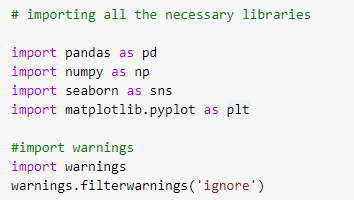
There are a total of 48,842 rows of data, and 3,620 with missing values, leaving 45,222 complete rows.

There are two class values ‘*>50K*‘ and ‘*<=50K*‘, meaning it is a binary classification task. The classes are imbalanced, with a skew toward the ‘*<=50K*‘ class label.

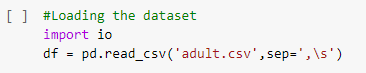
* **‘>50K’**: majority class, approximately 25%.
* **‘<=50K’**: minority class, approximately 75%.

Our task is to analyze the dataset and predict whether the income of an adult will exceed 50k per year or not by developing a supervised machine learning model.

* Importing required Libraries:

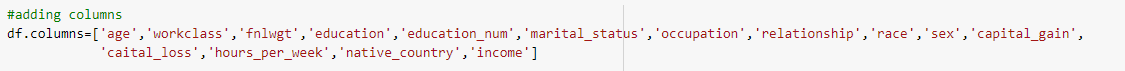


* Loading the dataset:



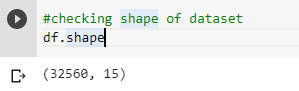
Now, if you follow closely this one line of code above may have intrigued you. This one line of code explains a lot about Python’s simple and natural language syntax. As we want to read/load a CSV, we just use the **read\_csv()** function provided by the Pandas library which we already imported and aliased it as **pd**. So, now all we need to do is to call the appropriate function from the library i.e. **pd.read\_csv()**. This shows us how powerful the Python libraries are that can load a CSV file in a single line.

1. Data Analysis:



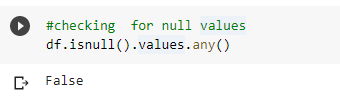
As we see there are no columns name so we add column according to problem statement.

* Checking the  shape of dataset:



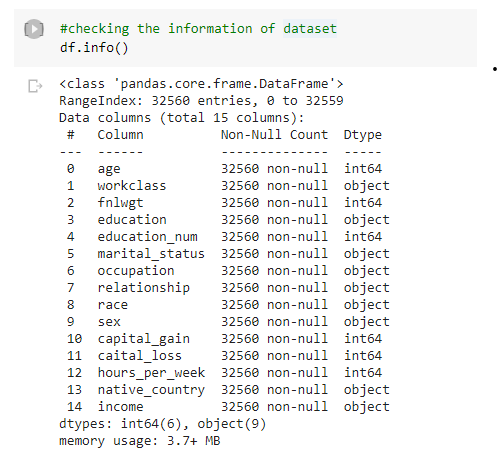
There are 32560 rows and 15 columns are present in our dataset.

* Checking  for null values:



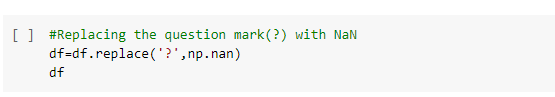
We see there are no null values but there is '?' present in our dataset which is treated as null value.

* checking the information of dataset:

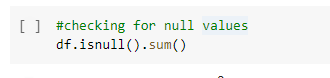


There are no missing values in our dataset but '?' is present so we will encode this as Nan value.

* Replacing the question mark(?) with NaN:

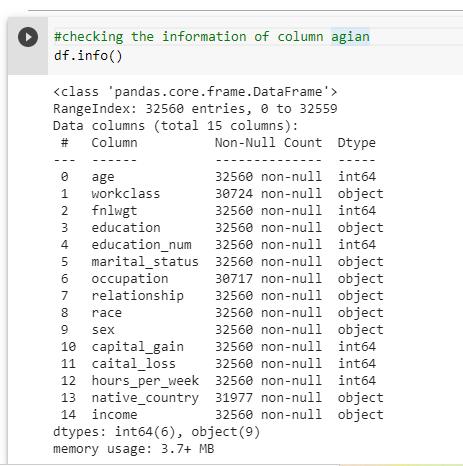


* checking for null values again:



Now it tells that nan values are present.so we check null values for particular column.

* checking for null values again for particular columns:

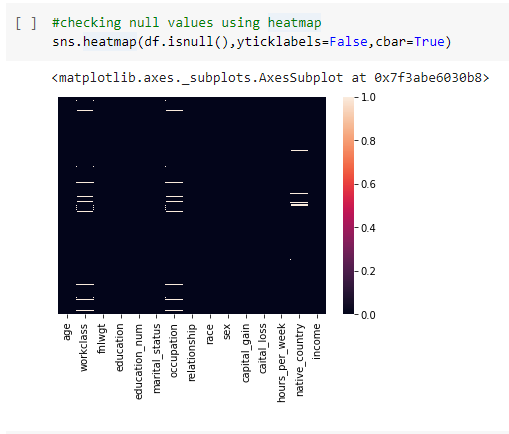


There are 1836 null values are present in workclass column

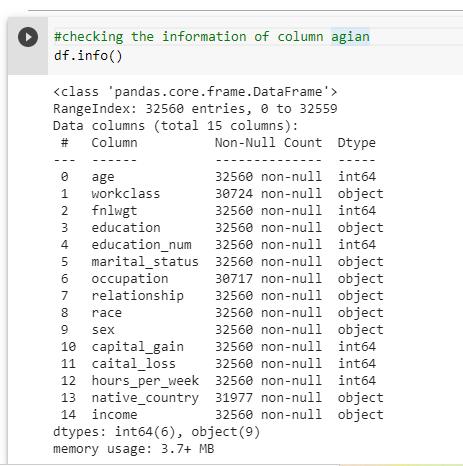
1843 null values are present in occupation column.

583 null values present in native\_country column.

* checking null values using heatmap



* checking the information of column again



Workclass, occupation , native\_country contain missing values.

All of these varibles are categorical datatype .So, we impute the missing values with most frequent value i.e. mode.

* Imputing the missing value with mode
* now again check for missing values

No null values are present in our dataset

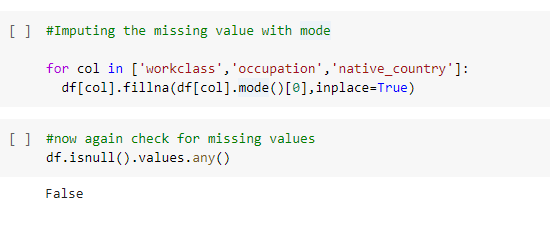
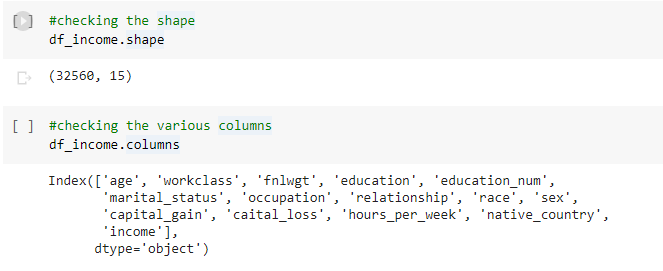
* Now Our data is ready for Visualisation without any Missing values.
* checking the shape
* checking the various columns
* checking the datatypes of columns

**Numeric Features**= age , fnlwgt, education\_num, capital\_gain, capita\_loss, hours per week.

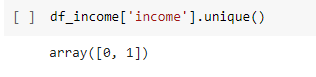
**Categorical Features**= workclass, education, marital\_status, occupation , relationship, race , sex, native\_country, income.

## ****For Analyzing The Data with Target i.e. income we have to change income into numeric type****

* Changing the datatype of the target column  by applying Label encoder

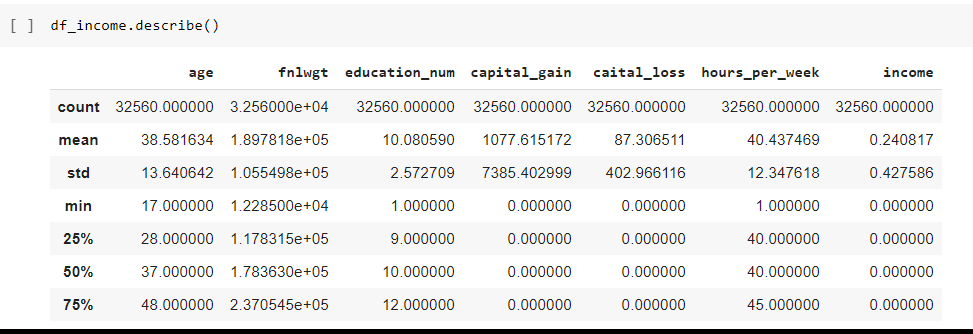
 

* **<=50k = 0**
* **>50k = 1**



So, here we are finally at our all numerical notations dataset. which can be now fed into any ML predictive modeling algorithm. Before we try to do that we should visualize the data and look for any correlations we can derive between the variables of our dataset.

## ****Summary Statistics****



We observe that:-

1- Minimum age is 17 years and Maximum age is 90 years.

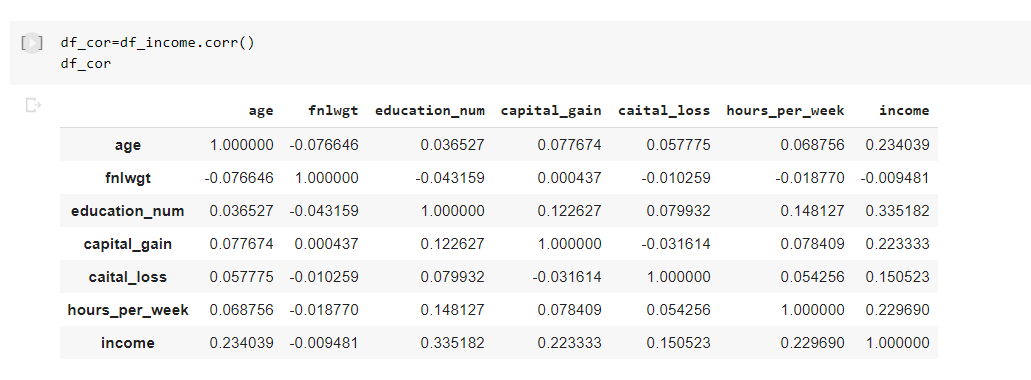
2-Minimum education\_num is 1 and maximum is 16.

3- Minimum hours per week is 1 and maximum is 99.

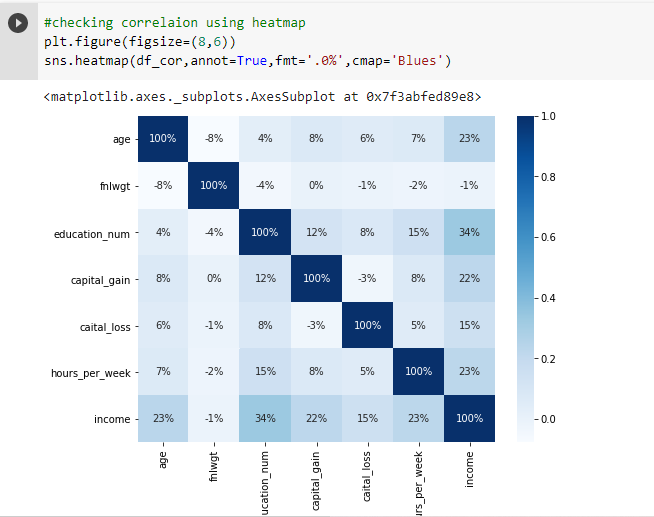
4- In age ,fnlwgt, education\_num, capital\_gain,capital\_loss attributes mean is greater than the median therefore in these attrtbutes data is right skewed.

5- difference b/w 75% and max is higher in age, fnlwgt, capital\_gain,capital\_loss,and hours\_per\_week therefore outliers present in these column which we have to remove it.

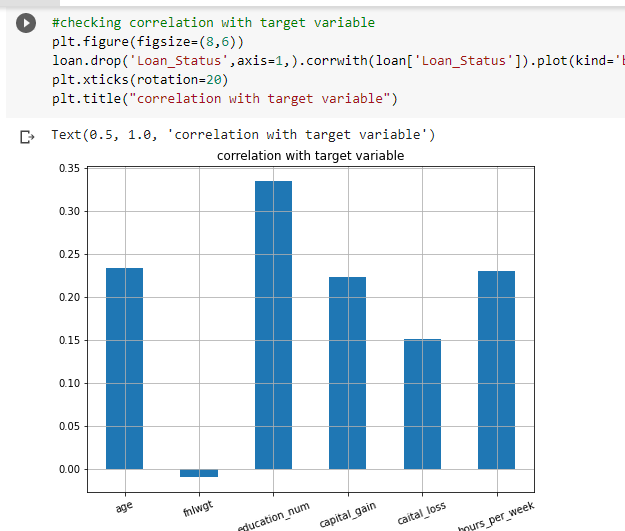
### ****Checking the Correlation****



Checking correlation using heatmap:



Checking correlation with Target variable:



1-fnlwgt is only negatively correlated with target variable income.

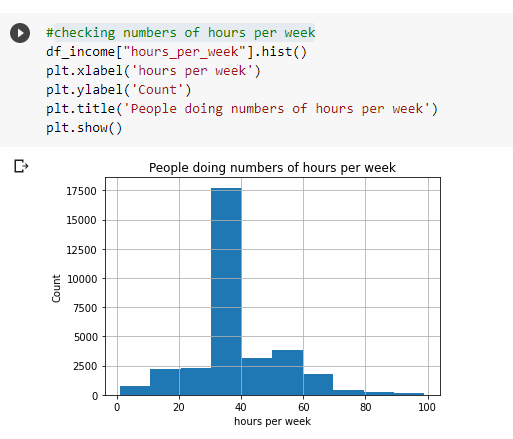
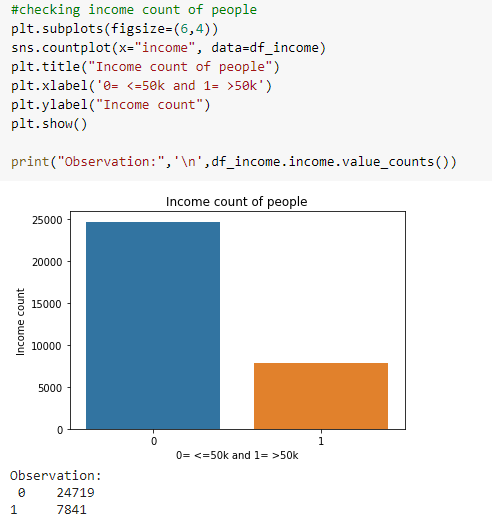
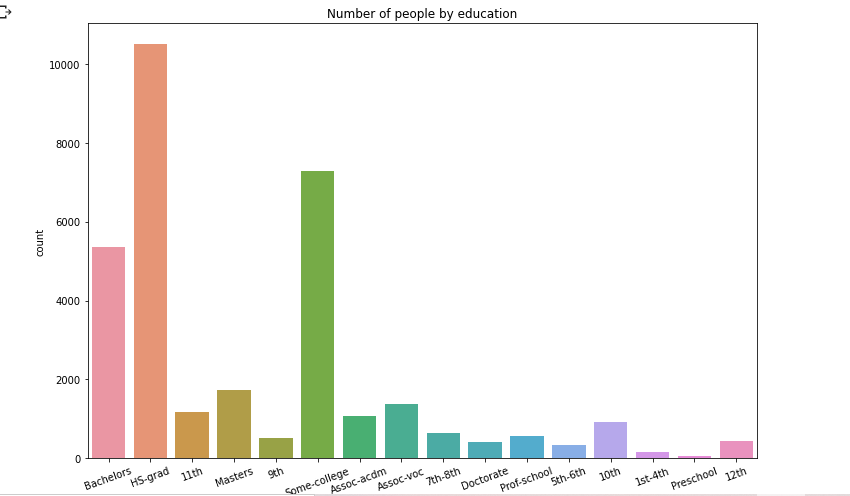
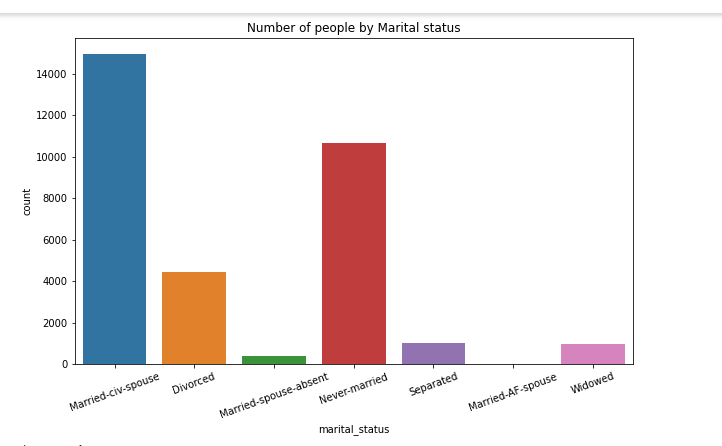
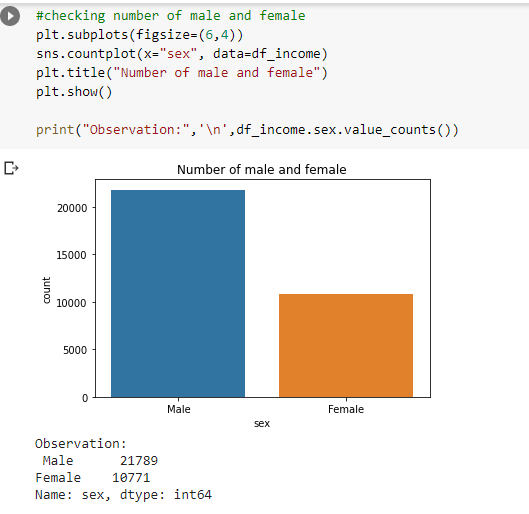
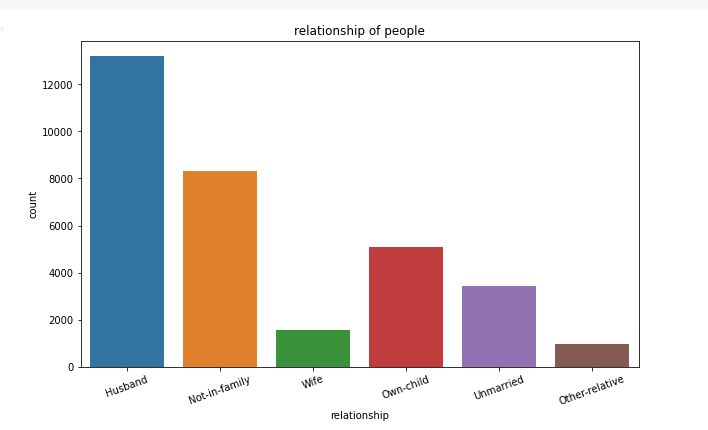
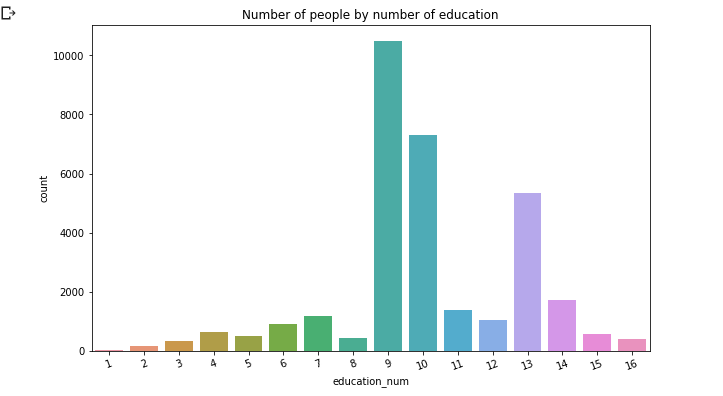
2-education\_num is highly correlated with target variable among all input variables.

## ****Exploratory Data Analysis(EDA):****

* our feature varible of interest is income.
* it refers to the monthly income of the people.
* 1 stands for greater than 50k and 0 stands for less than or equal to 50k.

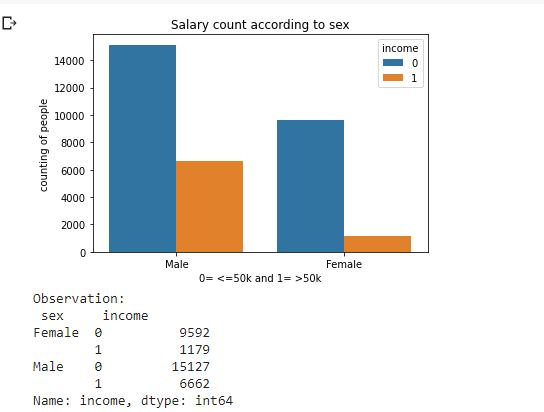
### ****Univariate Analysis****

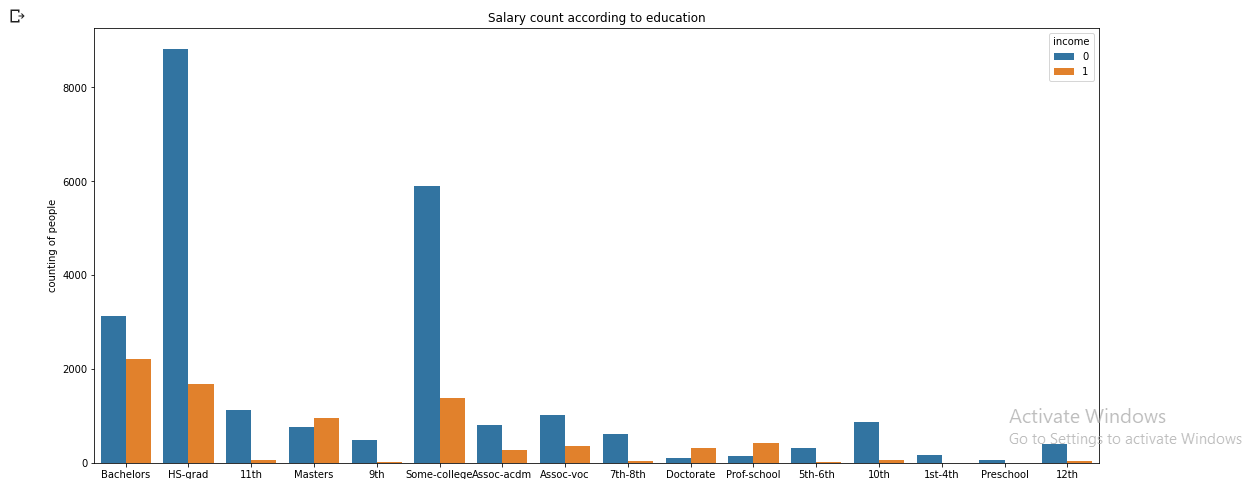
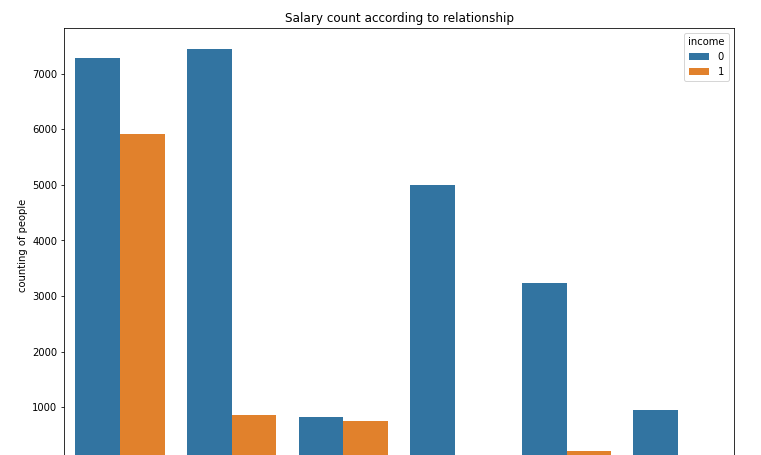
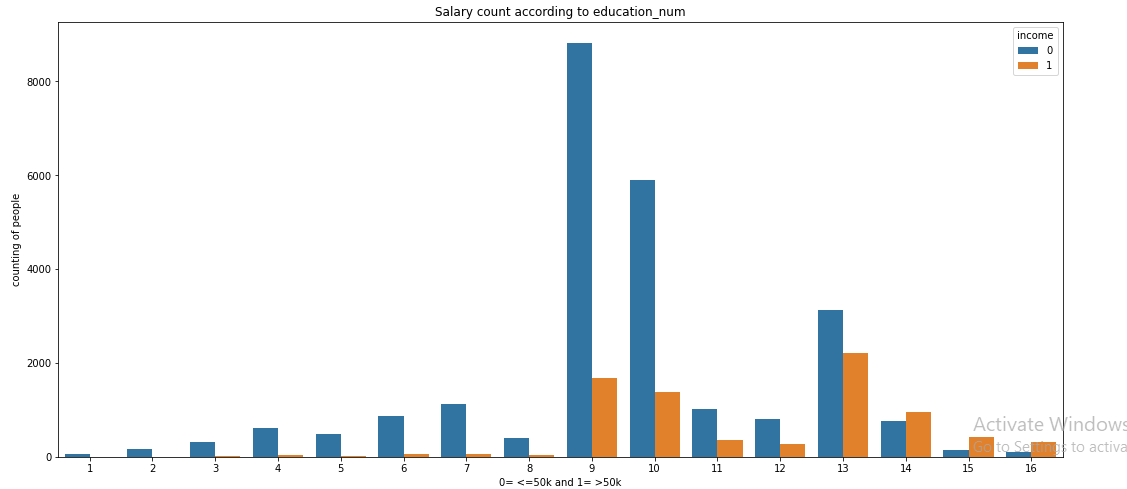
* checking income count of people, number of male and female, Marital status of people, relationship status of people ,Checking the number of people by education,  Checking the number of people by education number, Checking the number of people by Native country, checking numbers of hours per week



By analysing above we can say that Approx. 17500 people doing 30 to 40 hours\_per\_week.

Biverate Analysis:

Visualising Sex vs Income, Education vs Income, Education\_num vs Income, Relationship vs Income, orkclass vs Income



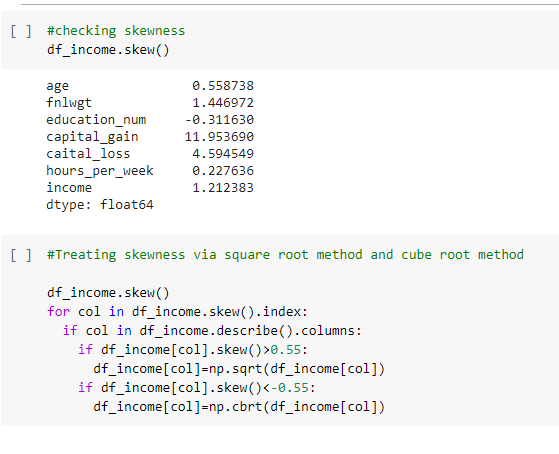
By analysing above we can say that

1. Adults with an educational background of Prof-school (7) and Doctorate (10) will have a better income and it is likely possible that their income is higher than 50K.
2. Our data suggest that people with occupation Prof-specialty (5) and Exec-managerial (7) will have a better chance of earning an income of more than 50K.
3. The gender bar chart provides us some useful insight into the data that Men (0) are more likely to have a higher income.
4. relationship chart shows us that wife (1) and husband (4) has a higher income. A married couple would most likely earn >50K.
5. As per the data, an Asian-Pac-Islander (1) or a white (3) have more chances of earning more than 50K.
6. Self-emp-in (0), Federal-gov(2) workclass groups have a higher chance of earning more than 50K.

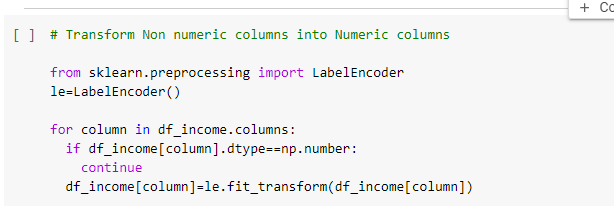
We can see that all of our numerical attributes have some influence on our target variable income. So, we should create an ML model by feeding all of our numerical data as input to it.

## ****Checking Skewness****

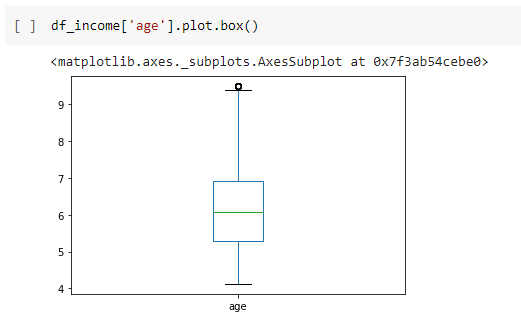
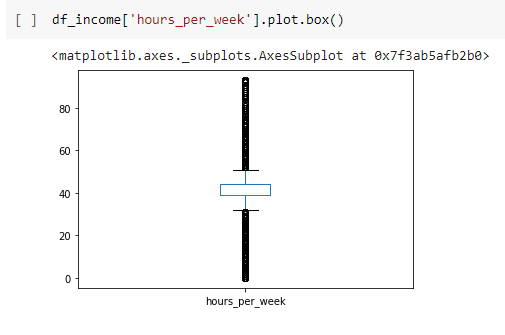
* Treating skewness via square root method and cube root method

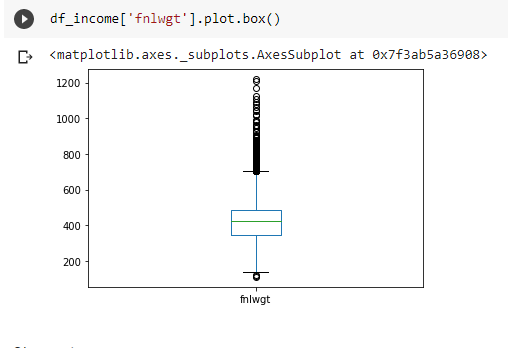


### ****converting the categorical data into numeric variables****



**Plotting Outlier**



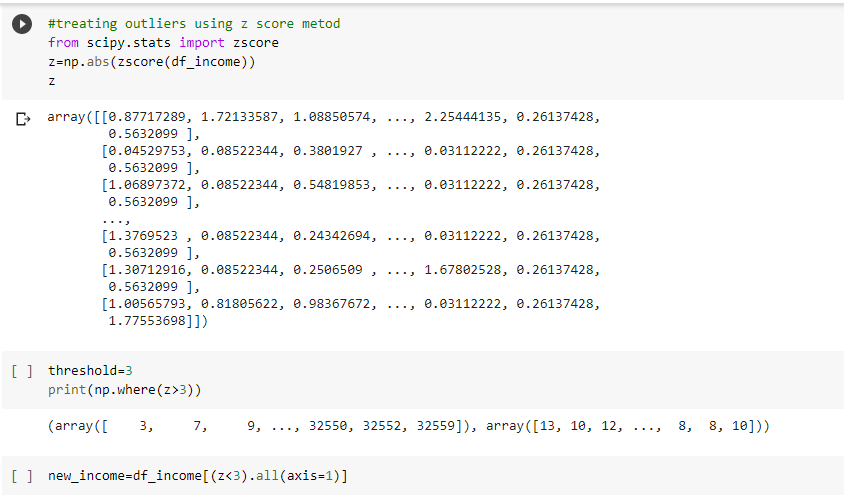


From above we can say that many outliers are present in cloumns.

An **outlier** is a data point in a data set that is distant from all other observations. A data point that lies outside the overall distribution of the data set

Now that we know outliers can either be a mistake or just variance, how would you decide if they are important or not. Well, it is simple if they are the result of a mistake, then we can ignore them, but if it is just a variance in the data, we would need think a bit further.

### ****Removing Outliers****



From above image we can clear see that there are number of black dots in most of the column which are referring to the outliers, so it means most of the data are outside the distribution.

So now we detect the outliers now the second step is to remove the outliers, there are different way to remove the outliers that are find the IQR, zscore values.

I am using both zscore value then I again check if there are some of the outliers then I will remove it by replacing the outliers with the mean value of that column.

So, I first find the zscore value and then I decide to make one threshold value as 3 which is standard of industry recommend value and then I remove all the outliers which zscore value is greater than 3.

**Pre Processing Pipeline:**

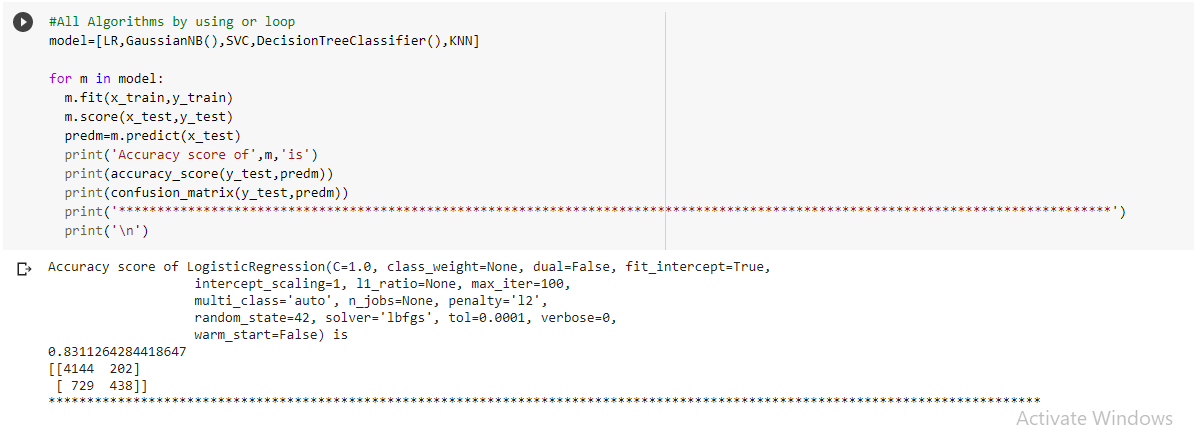
**Separating data variable and independent variable.**

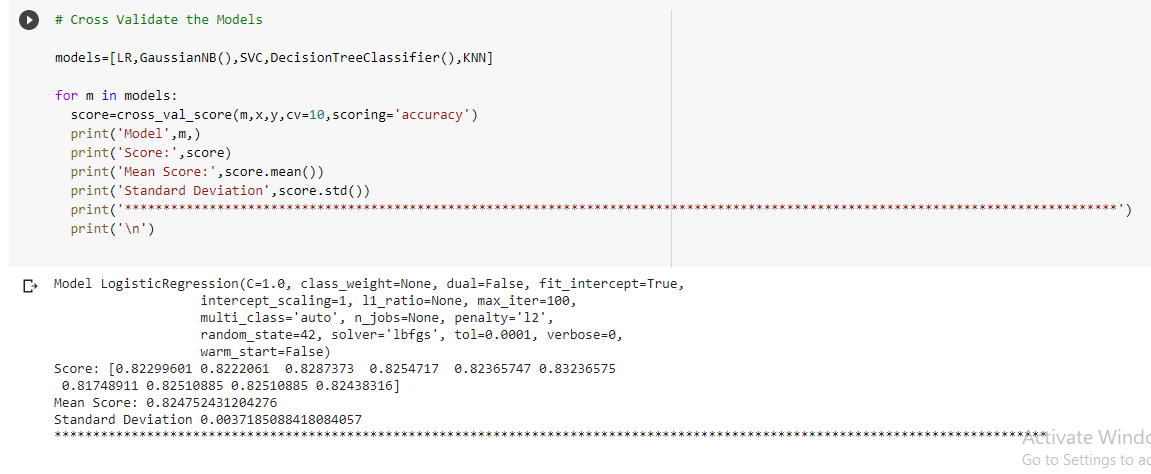
**Drop and Standard Scaler:**



**Building Machine Learning Models:**

All Algorithms by using or loop :



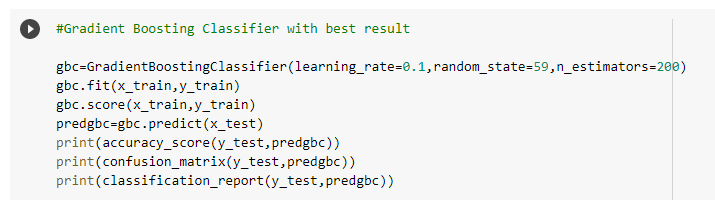
* Cross Validate the Models
* 

## ****Using Ensemble Technique to Boost our score****

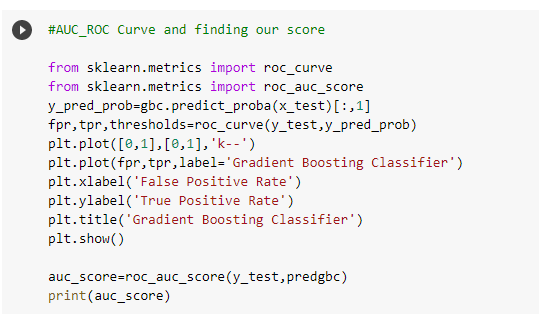
* All Algorithm by using for loop

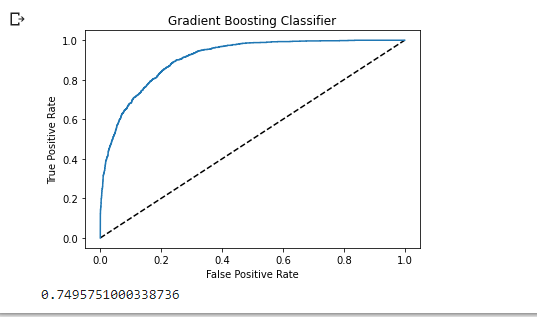


Gradient Boosting Classifier gives best score and its is not suffering from underfitting and overfitting

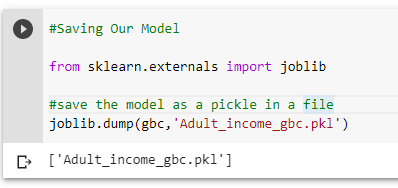


**So we choose Gradient Boosting Classifier as our best model.**

* AUC\_ROC Curve and finding our score 



* Saving Our Model



**Conclusion**

Finally! we successfully created a classification Machine Learning prediction model using Python and its powerful libraries which predicts whether a given adult’s income will be >50K or not.