**Census Income Project**



Submitted by

Harshitha Valluri

**Problem Definition:**

This data was extracted from the [1994 Census bureau database](http://www.census.gov/en.html) by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1) && (HRSWK>0)). The prediction task is to determine whether a person makes over $50K a year.

**Description of fnlwgt (final weight)**

The weights on the Current Population Survey (CPS) files are controlled to independent estimates of the civilian non-institutional population of the US. These are prepared monthly for us by Population Division here at the Census Bureau. We use 3 sets of controls. These are:

-A single cell estimates of the population 16+ for each state

-Controls for Hispanic Origin by age and sex

-Controls by Race, age and sex

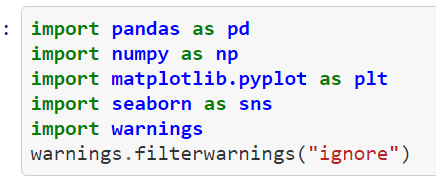
We use all three sets of controls in our weighting program and "rake" through them 6 times so that by the end we come back to all the controls we used. The term estimate refers to population totals derived from CPS by creating "weighted tallies" of any specified socio-economic characteristics of the population. People with similar demographic characteristics should have similar weights. There is one important caveat to remember about this statement- that is that since the CPS sample is actually a collection of 51 state samples, each with its own probability of selection, the statement only applies within state.

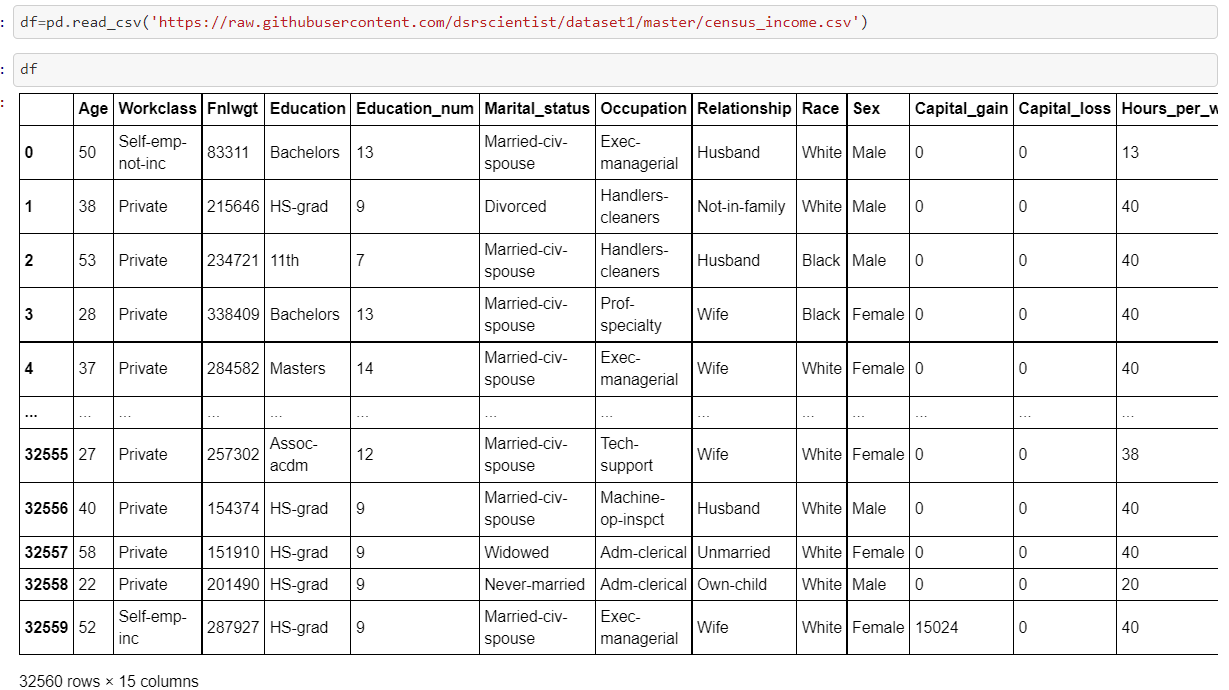
**Dataset:**

<https://raw.githubusercontent.com/dsrscientist/dataset1/master/census_income.csv>

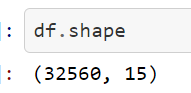
**Data Analysis:**

* Reading the CSV file and displaying the data



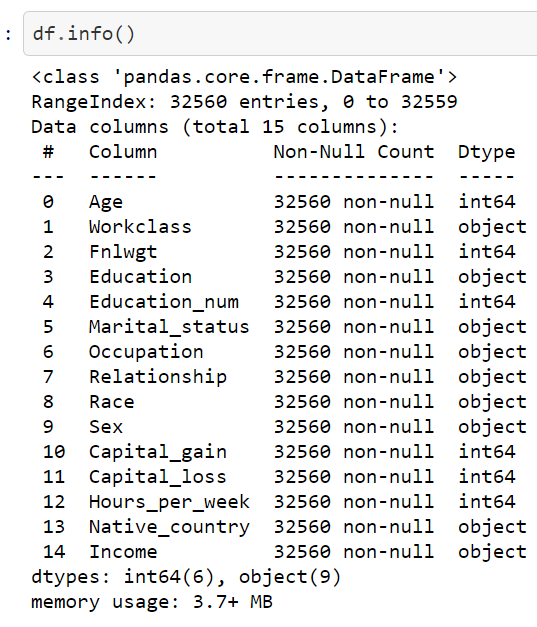


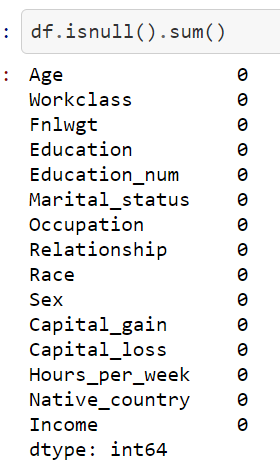
* Data Inspection



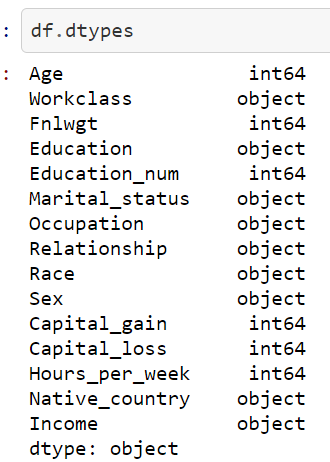
There are 32560 rows and 15 columns

Getting information about the data like null values, data types, columns in the dataset-



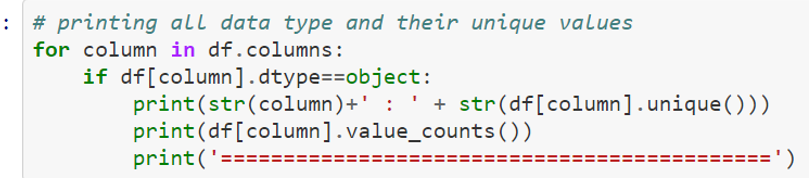


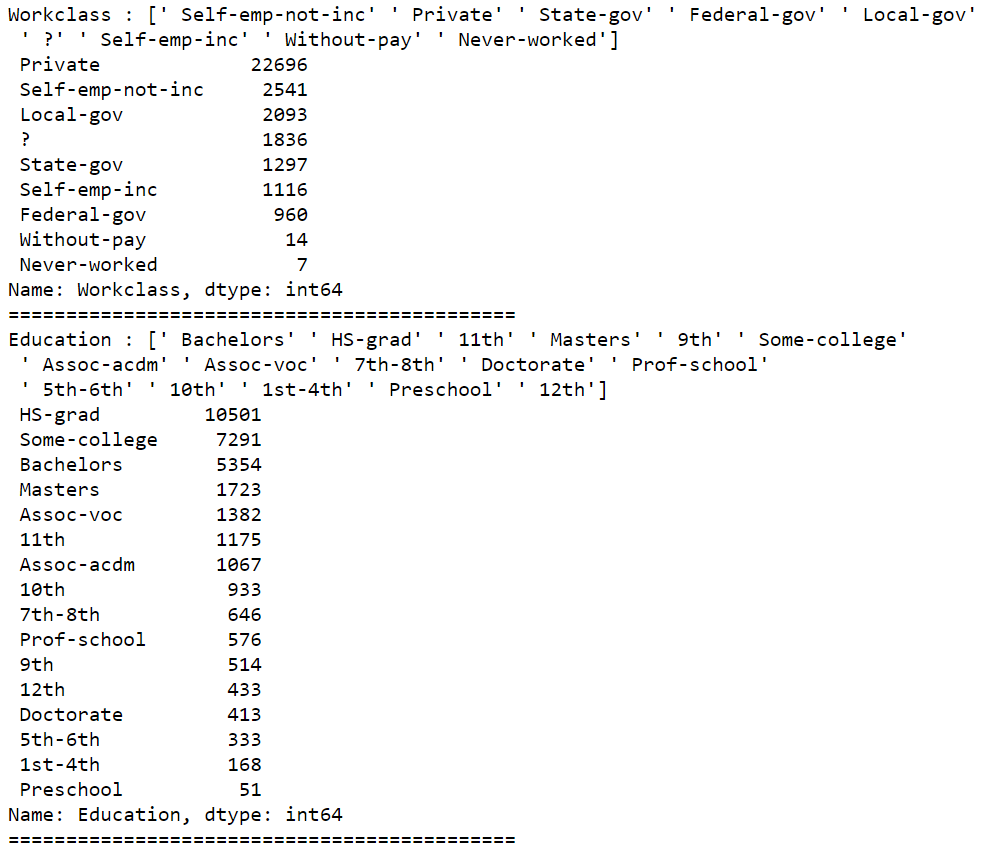
There are no null values in this dataset

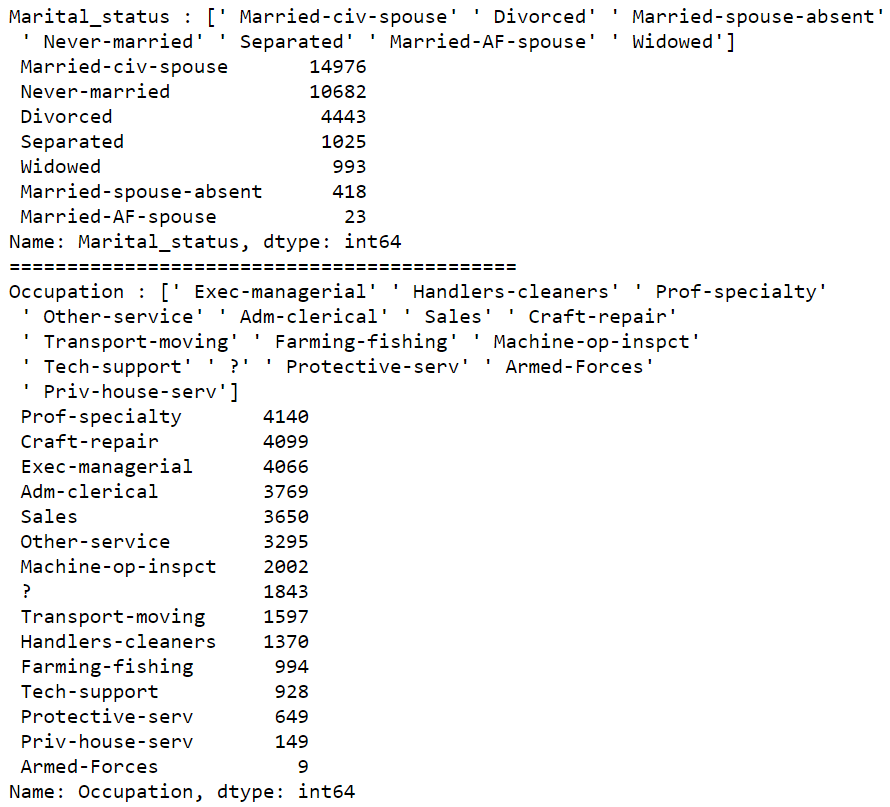


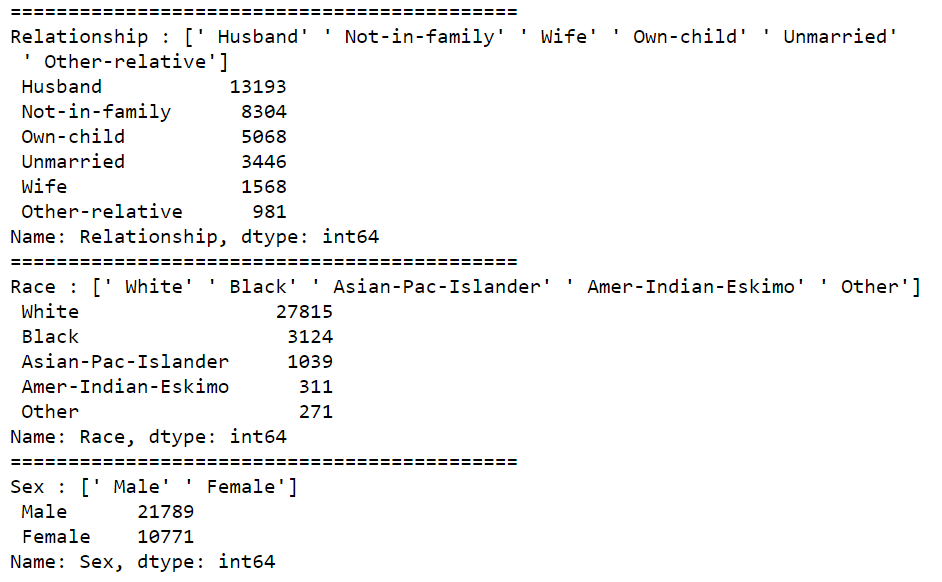
Shows the datatypes of columns

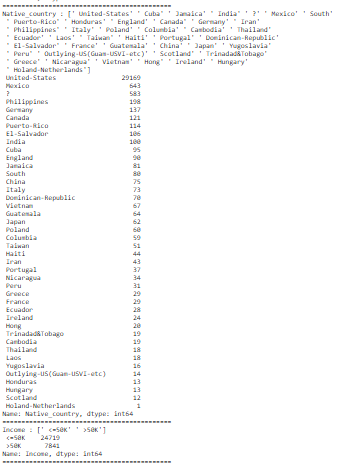
Showing all the data types and their unique values-

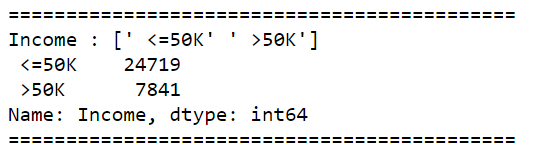




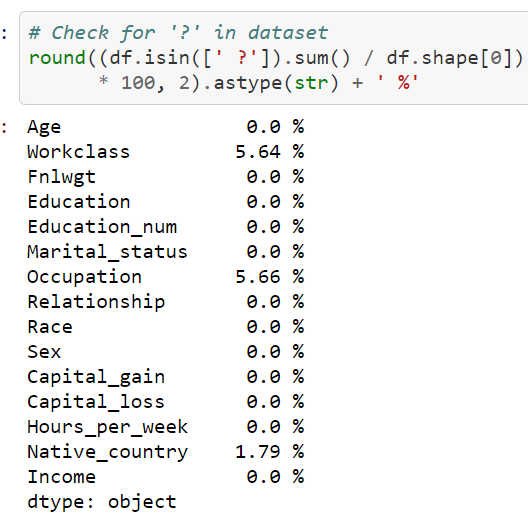






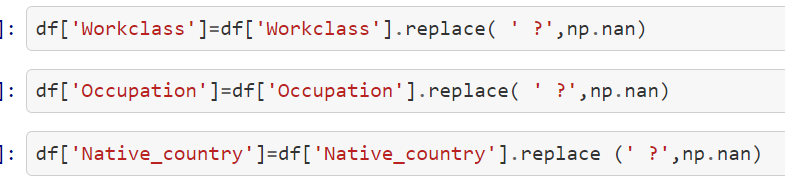


Checking for ‘?’ in the dataset,

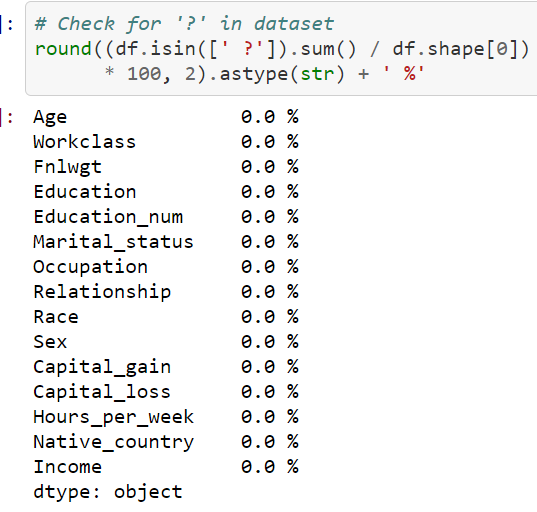


There is no null value in data but it contains missing values in the form of '?' which needs to be pre-processed.

Replacing the values containing ‘?’ with null values

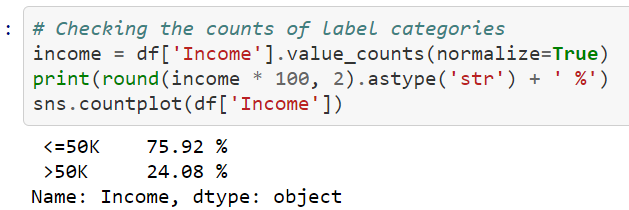


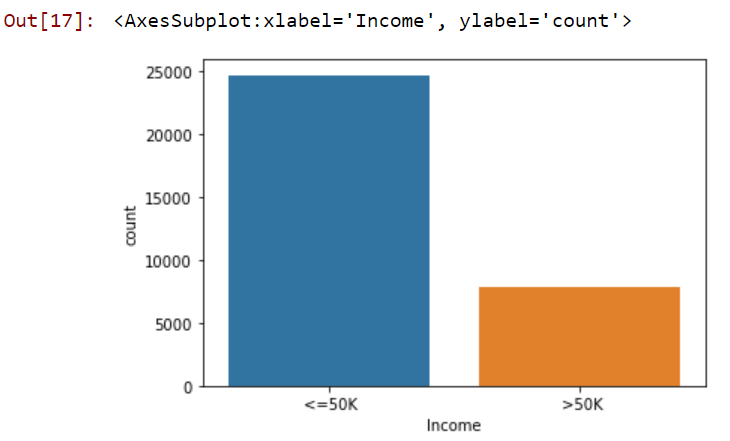
Checking if there are any ‘?’ values



We can see that there are no values containing ‘?’

Checking the count of label categories now,

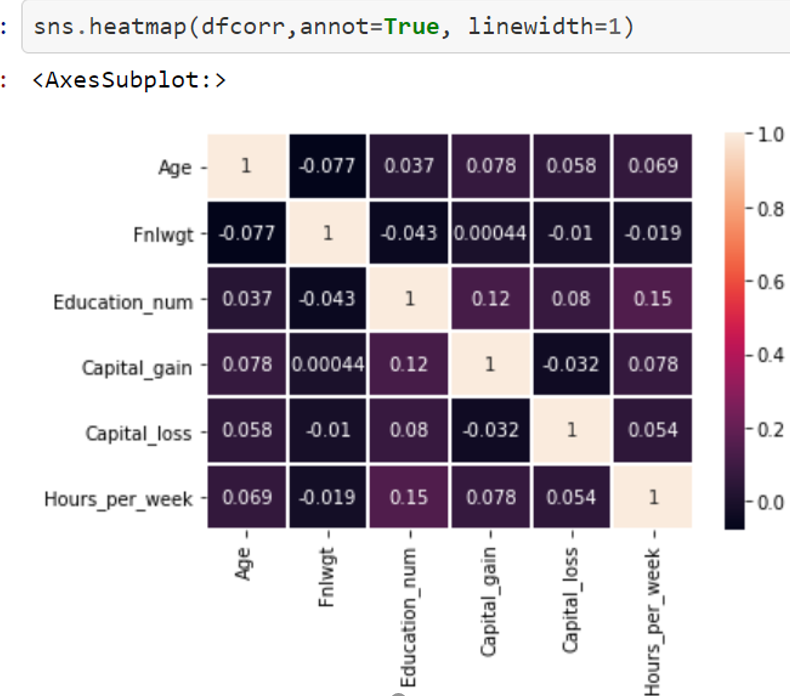




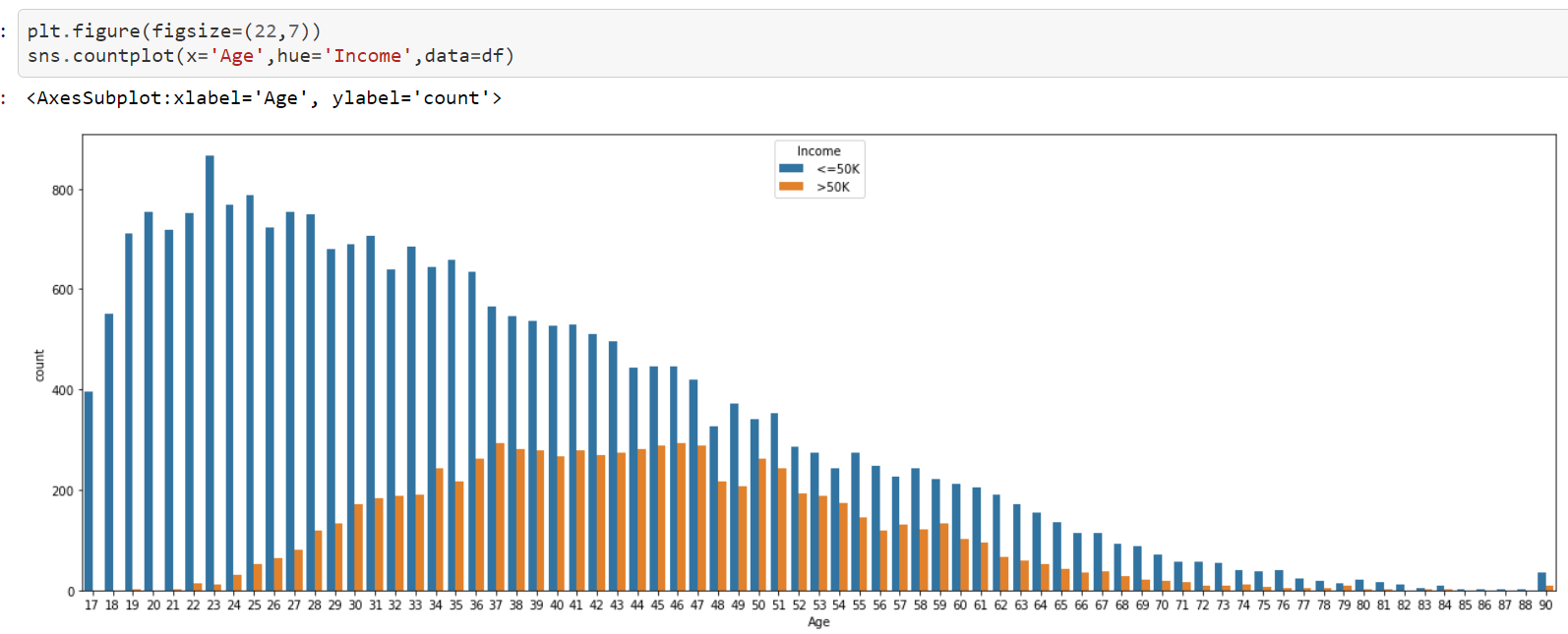
We can see that the data is imbalanced, it needs to be balanced

**Exploratory Data Analysis:**

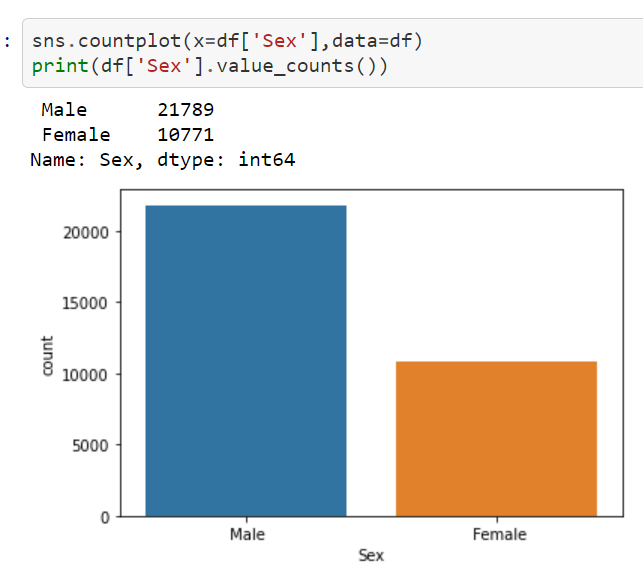
Checking the correlation between the variables

****

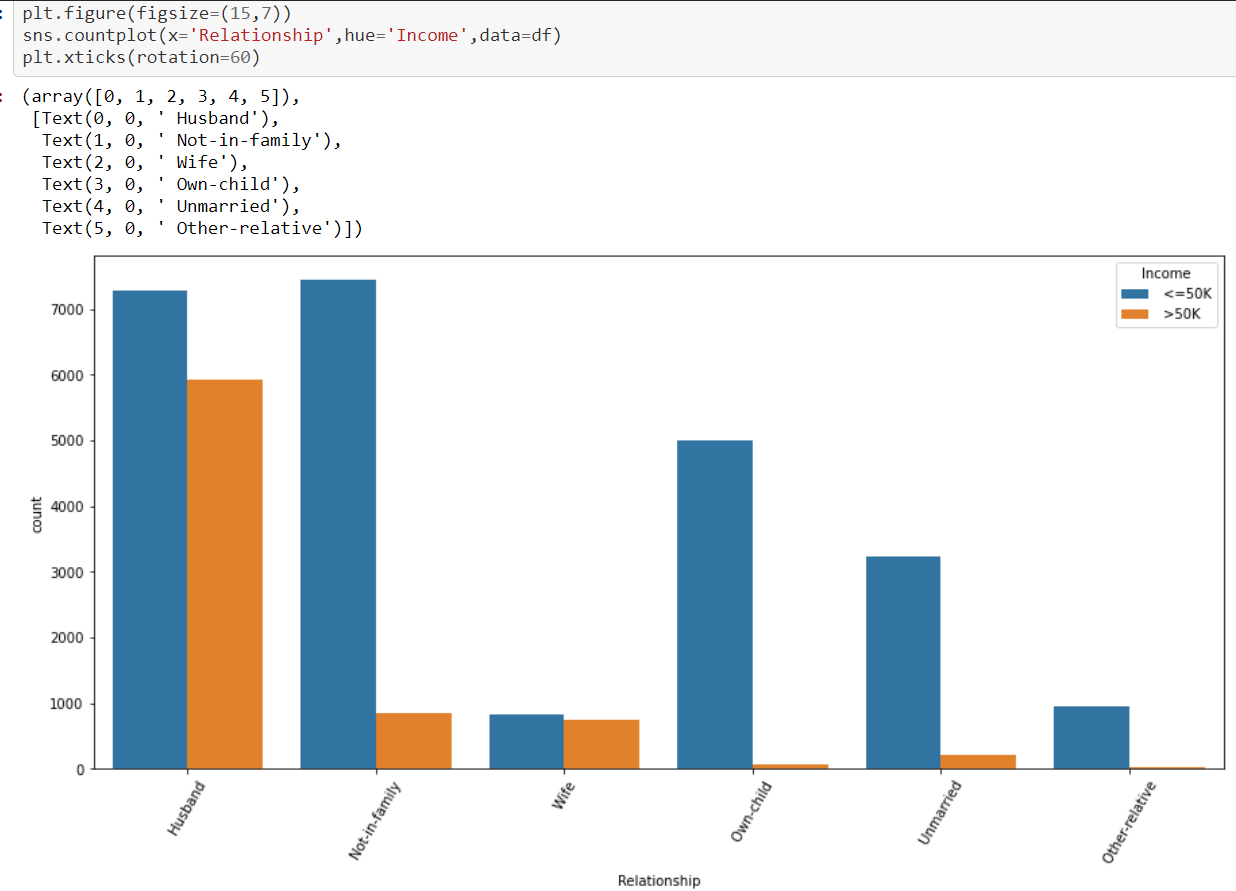
Checking the income of people with respect to age to check which category earns who earn > or < than 50k,



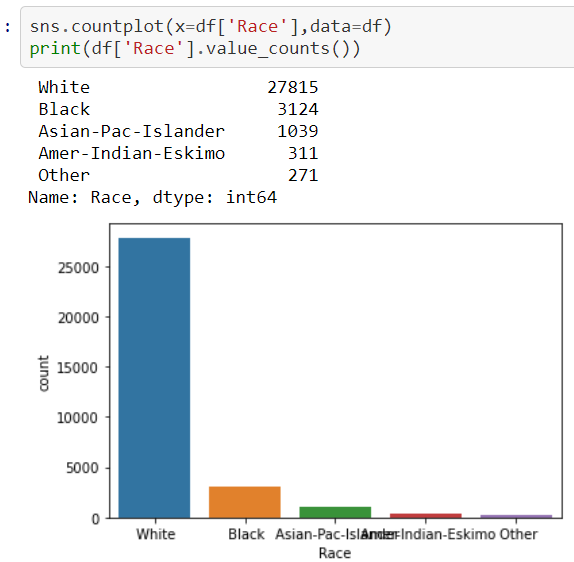
Showing the number of males and females in the dataset



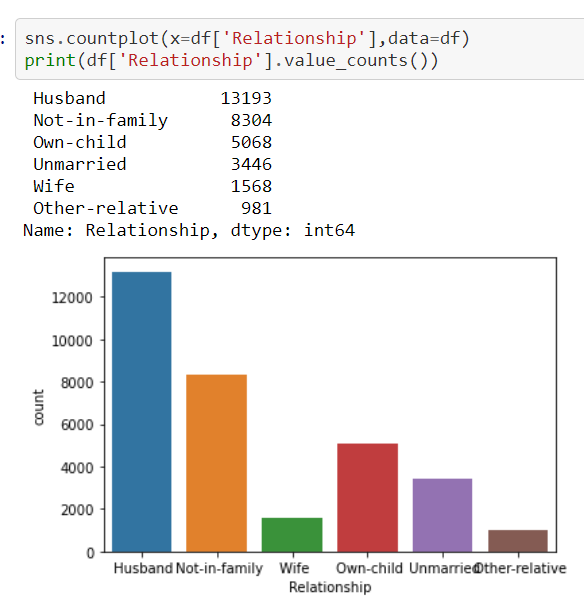
Now let’s check the income of people with respect to their relationship status,

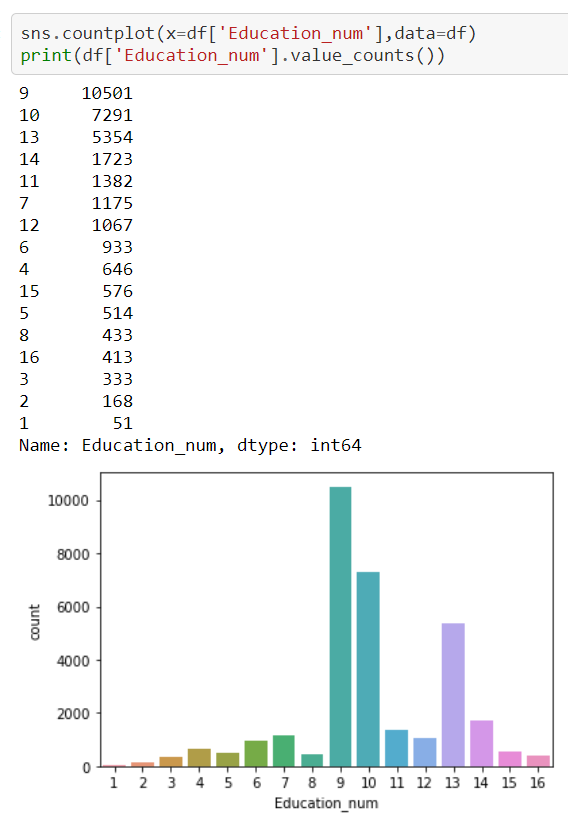


Checking the race of people in the dataset,

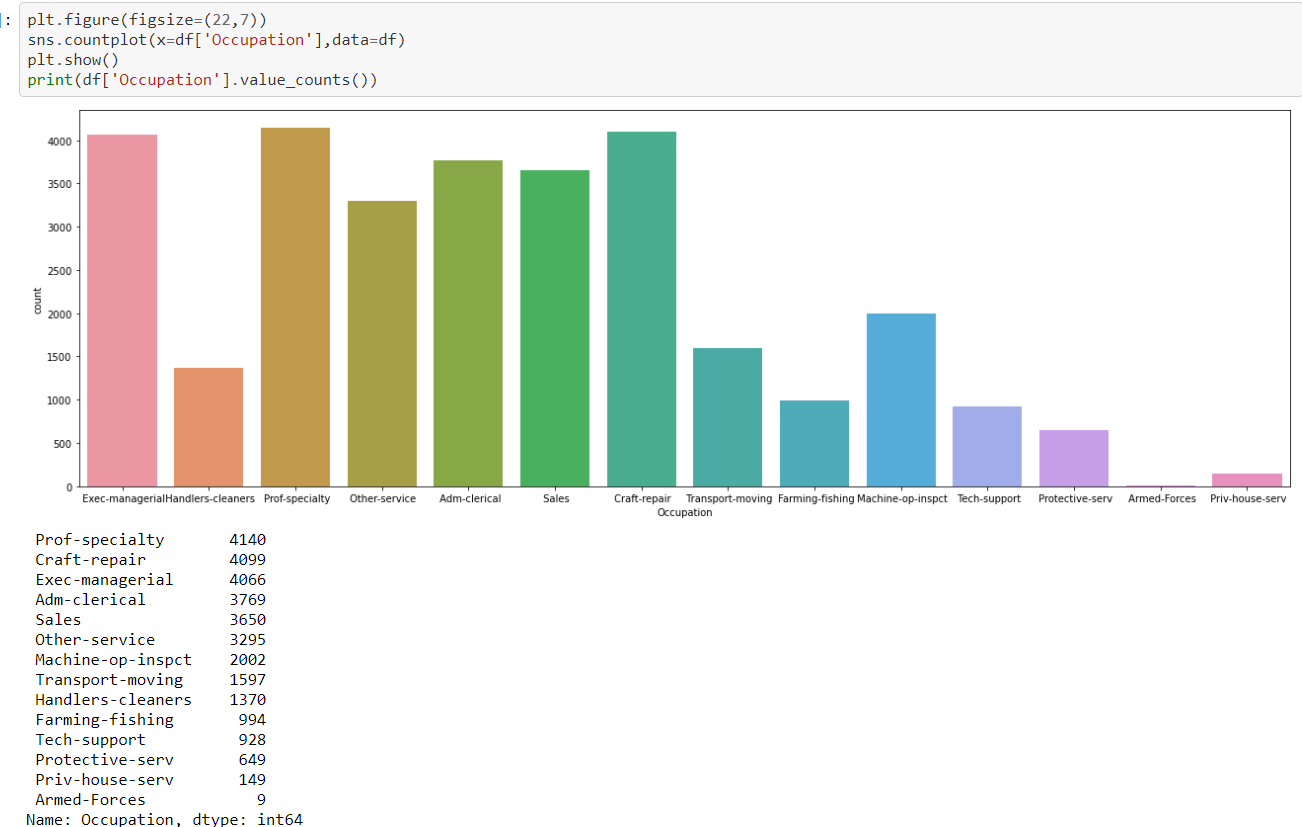


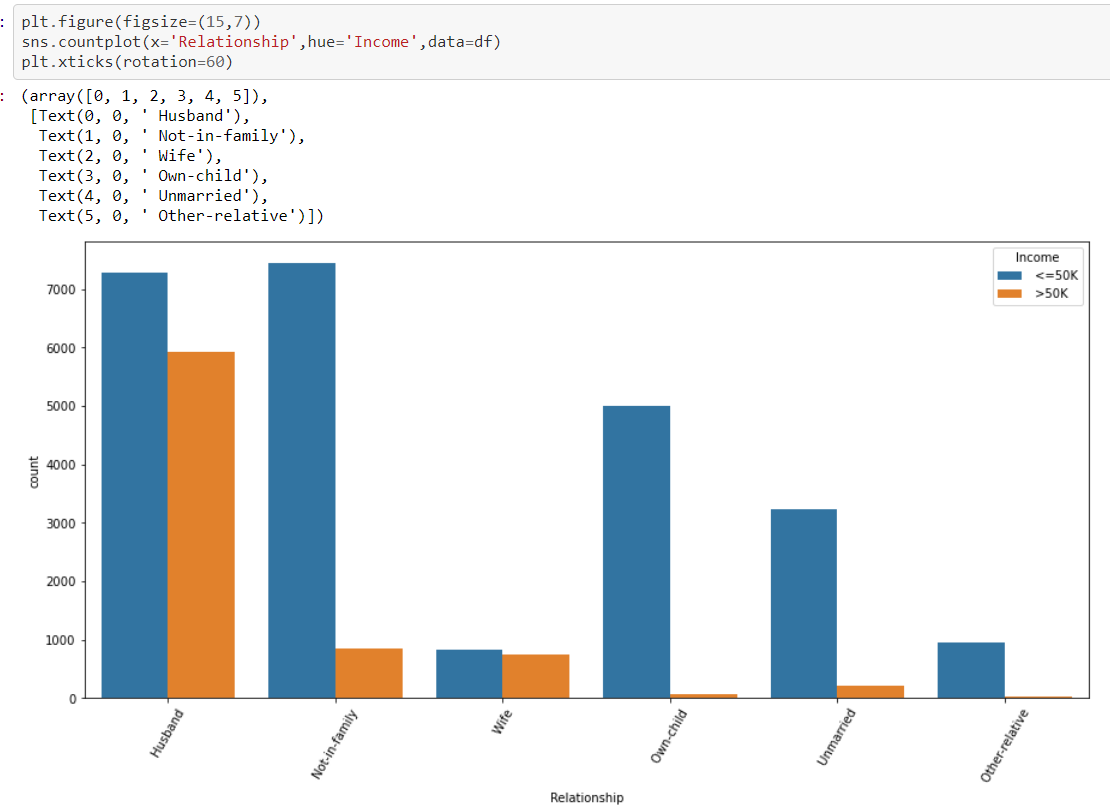
It looks like white people have the most count with 27815 followed by black who are 3124, Asian-Pac-Islander who are 1039, Amer-Indian-Eskimo who are 311 and finally others with 271 in count





Shows the occupation of people and their count. Prof- Speciality have the most count with 4140 whereas there are 9 people who are in Armed-Forces

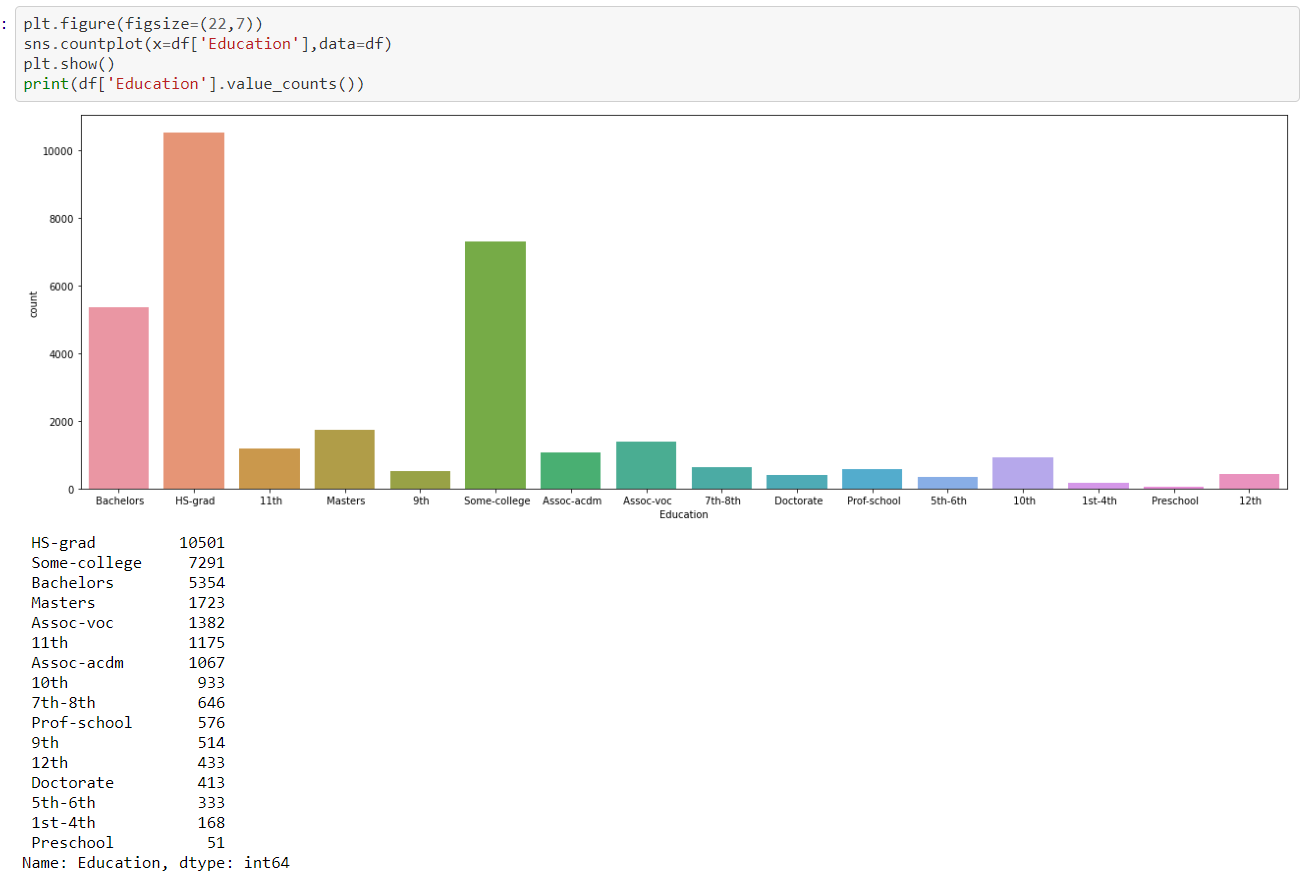




Only in Husband we can see that almost 6000 people have more than 50k income

All in other relations, we have most of the people less than 50k income

Shows what is the highest qualification of people here,



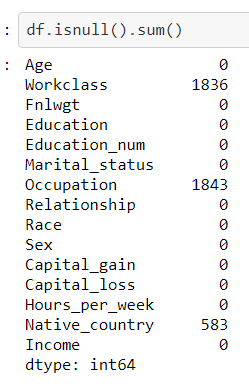
-HS-grad Education is having high impact on income

-People who have done HS-grad have less than 50k

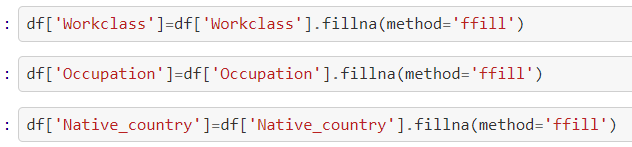
-People who have passed only till 11th, amongst those very few of them have more than 50k

-Doctorate and Prof-school most of the people have more than 50k and in all other categories, most of people have less than 50k

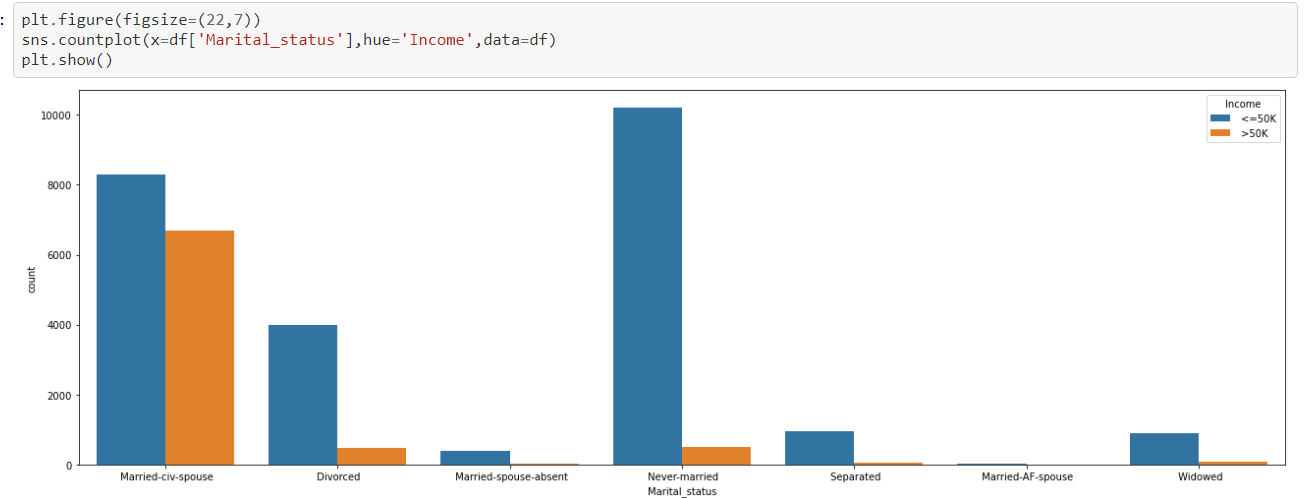
Showing the null values if any in dataset,



We can see that there are null values present in Workclass, Occupation and Native-country column. Removing null values-



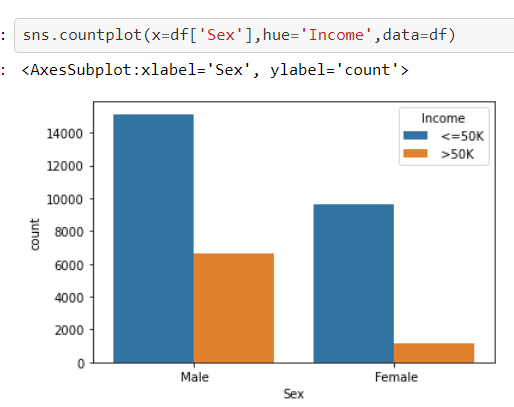
Showing income of people with respect to their marital status,



-People under married-civ-spouse category have most incomes above 50000 compared to the rest

-People under Never-married category have most incomes below 50000 compared to the rest

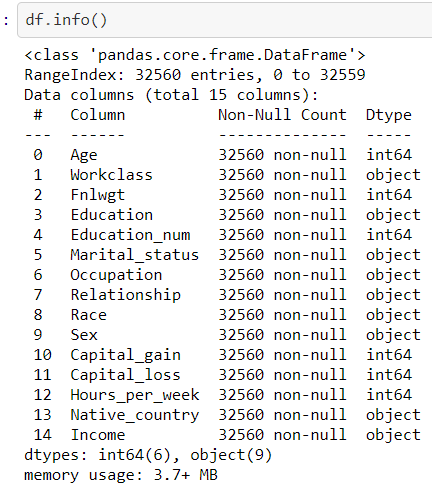
Showing the sex with respect to income.



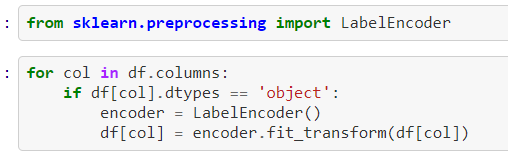
We can see that the count of males is more having income<=50k and also having income >50k

**Pre-processing Pipeline:**

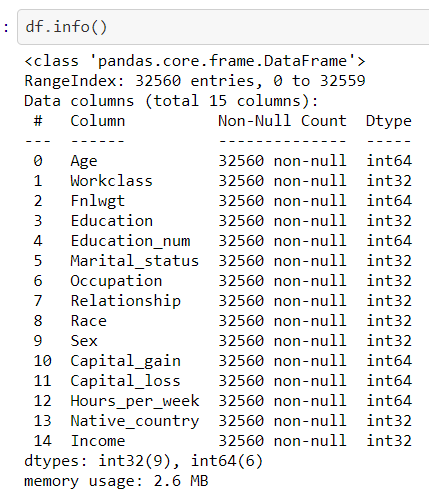
Using label encoder for converting categorical to numerical values







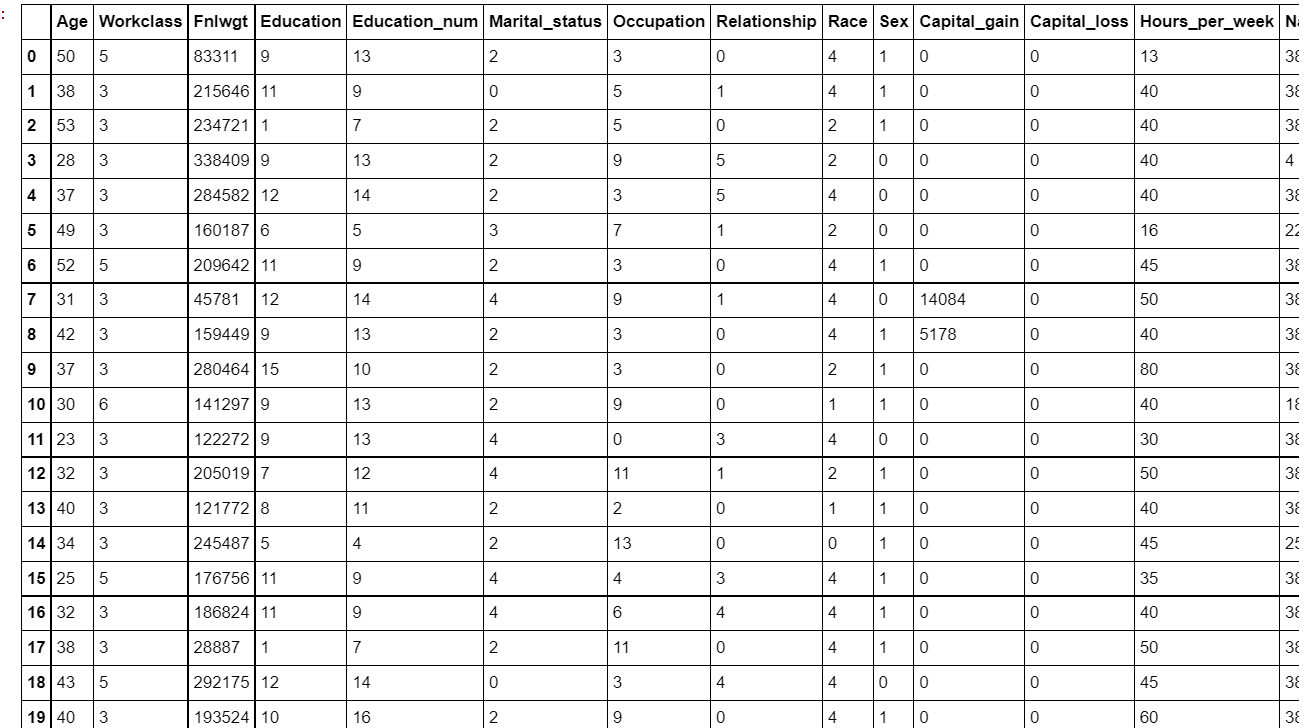
Now let’s check if the variables have changed or not,



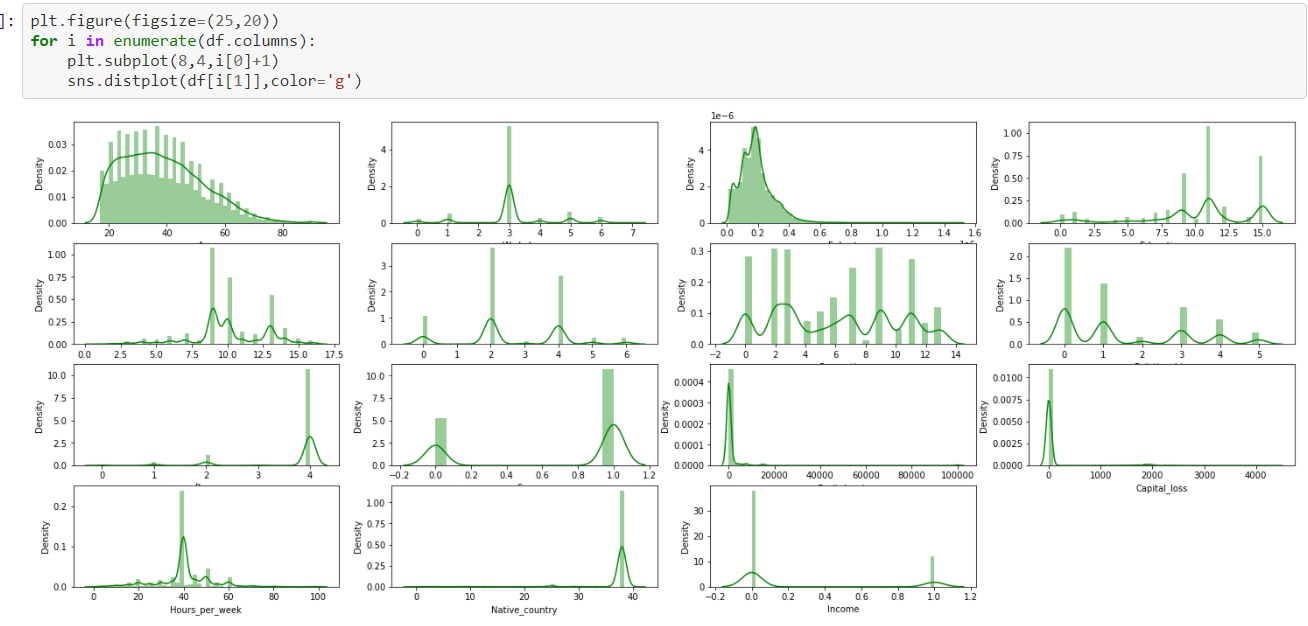
We can see that the values have changed to int datatype

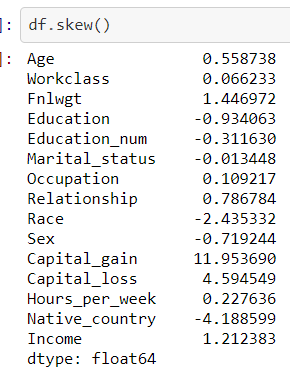
Displaying the data-





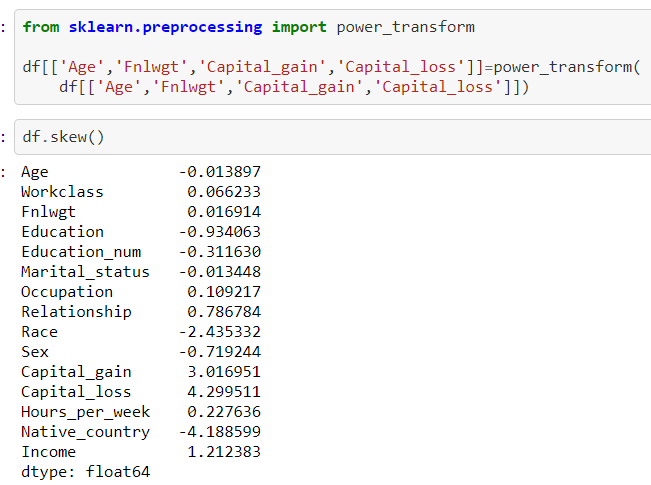
Skewness Handling:



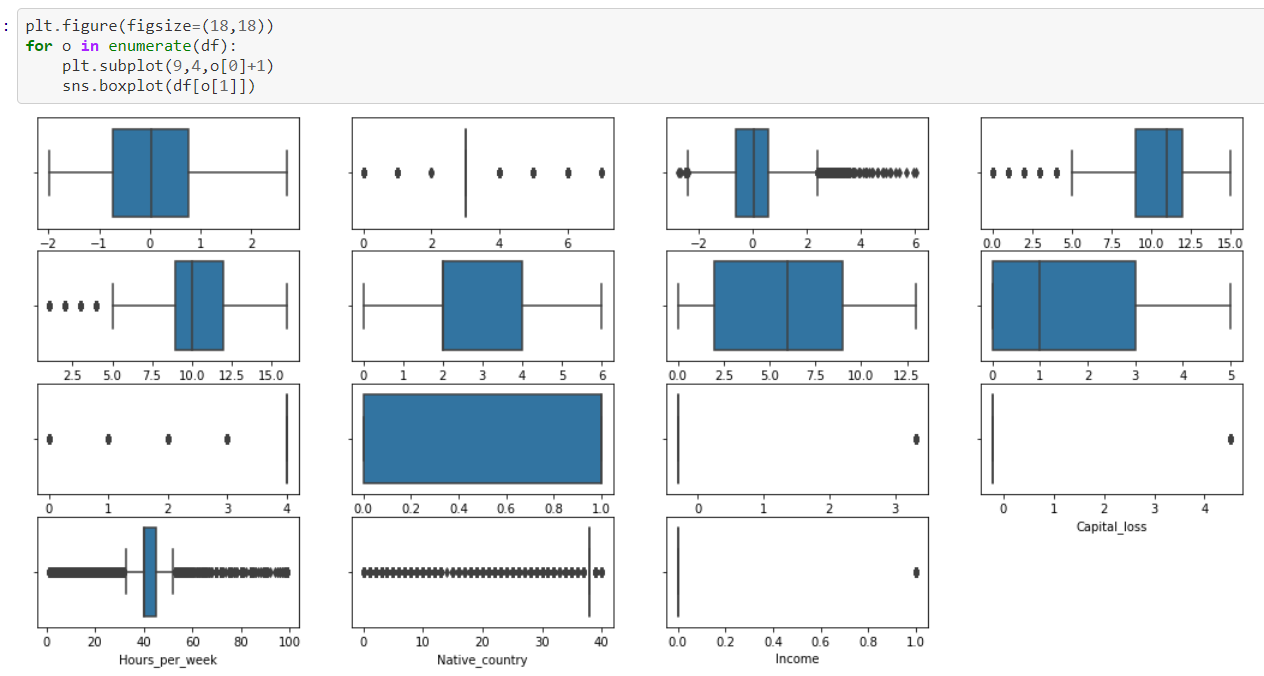


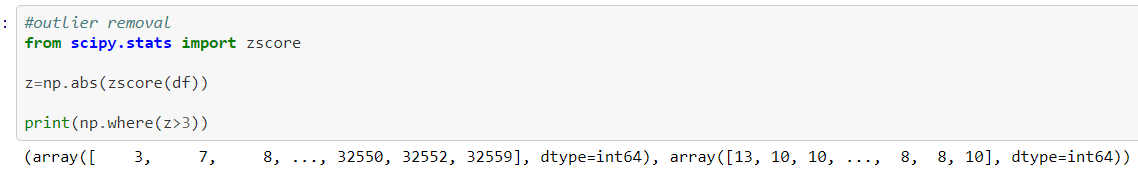
Skewness more than+/- 0.5 will we treated

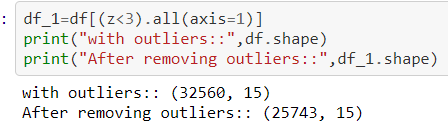
We can see here most of the columns are skewed

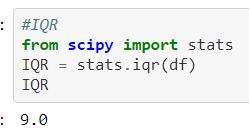


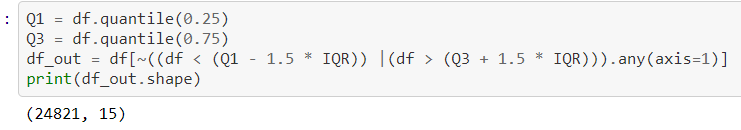
Outliers:



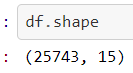




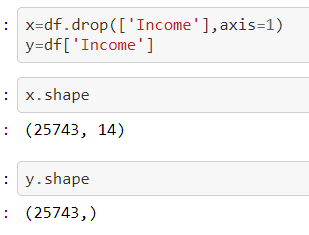




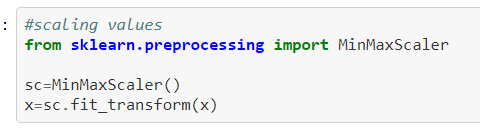


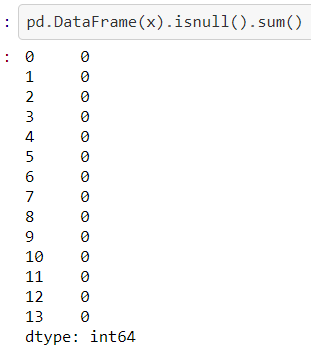


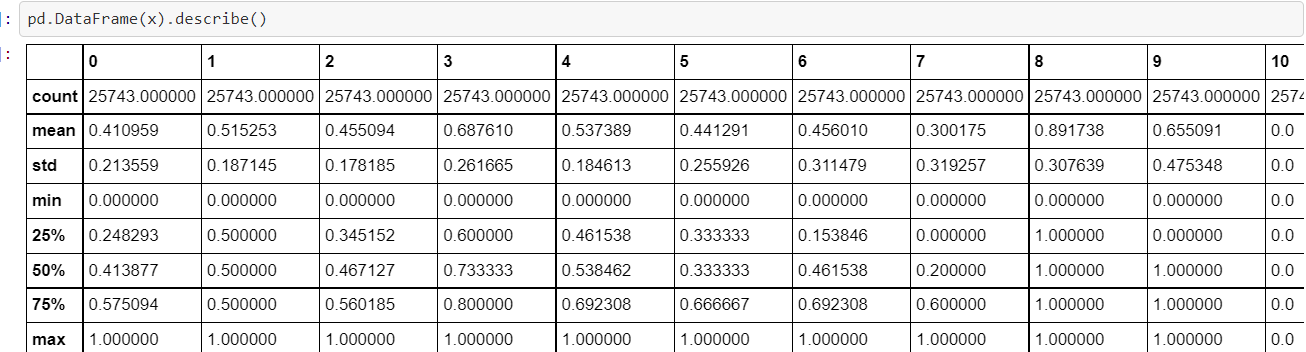
Splitting into Dependent and independent variable:



Scaling the values,

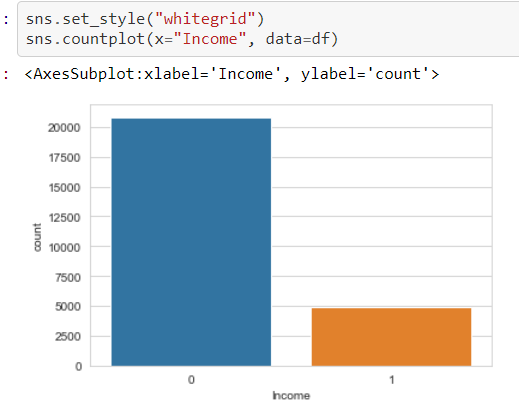


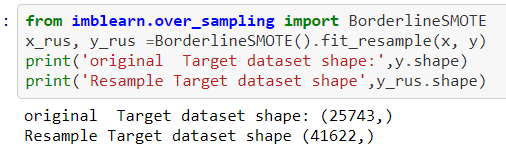


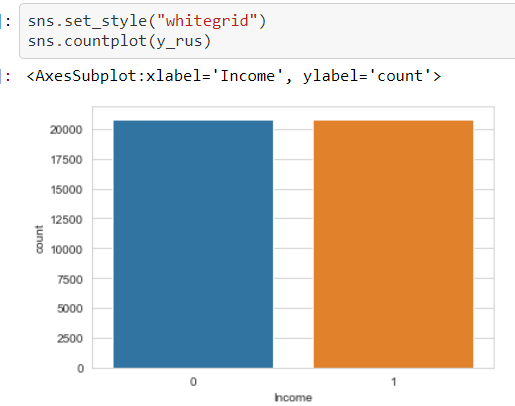


Imbalanced data:

Using OverSamling\_BorderlineSMOTE because there is less data in attrition for yes

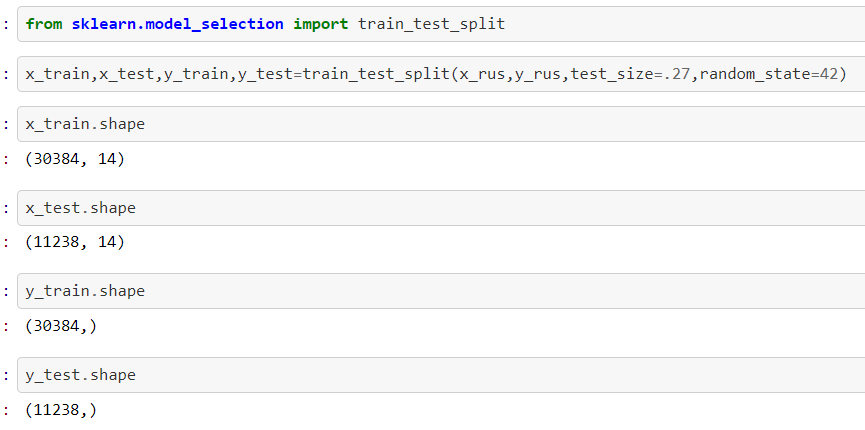




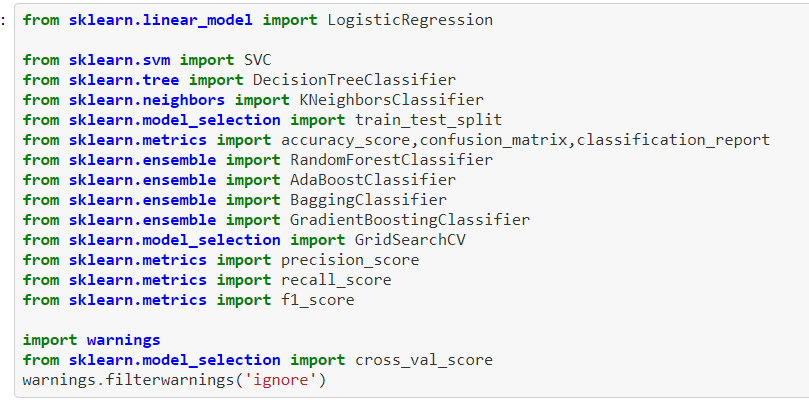


We can see that the data is balanced now.

Splitting the dataset

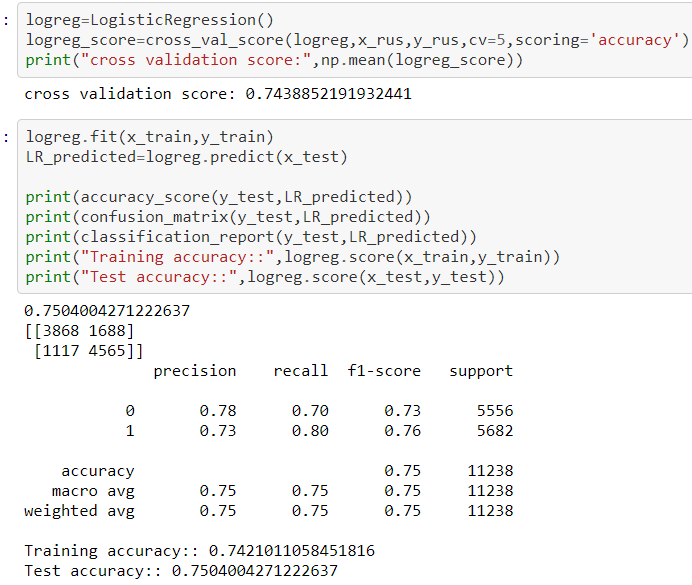


**Building Machine Learning Models:**

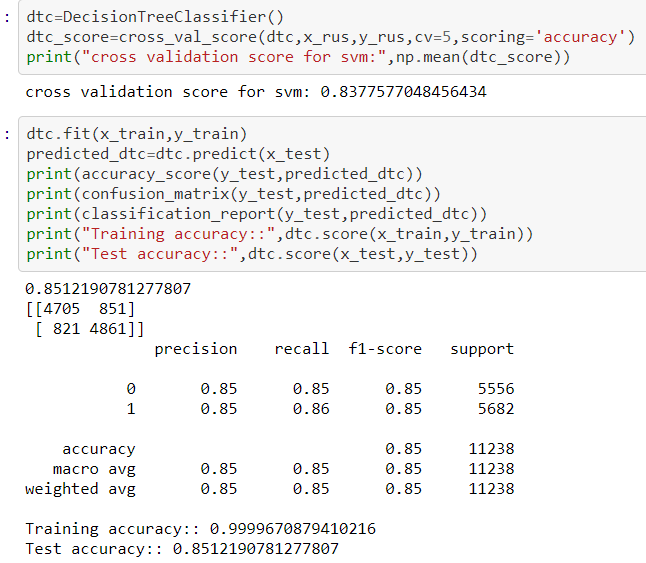


We’ll find out which ML model shows the best accuracy and cross check it doing cross validation and the model which has the least difference between the accuracy and cross validation will be chosen. Hyper parameter tuning will be done further.

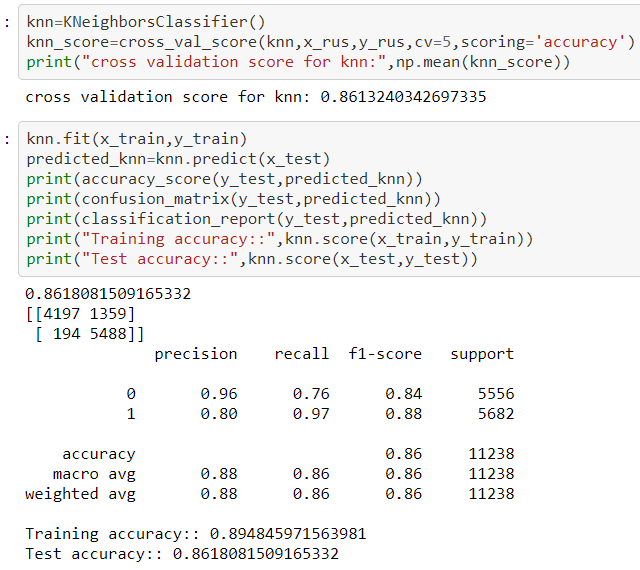
Logistic regression:



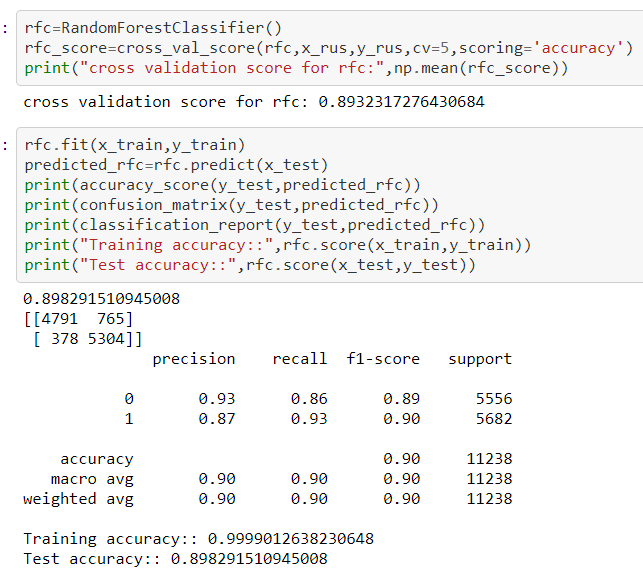
Decision Tree Classifier:



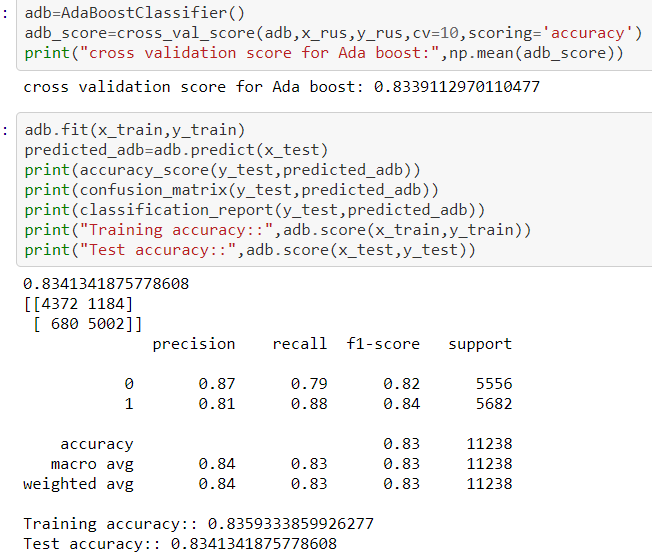
K Neighbour Classifier:



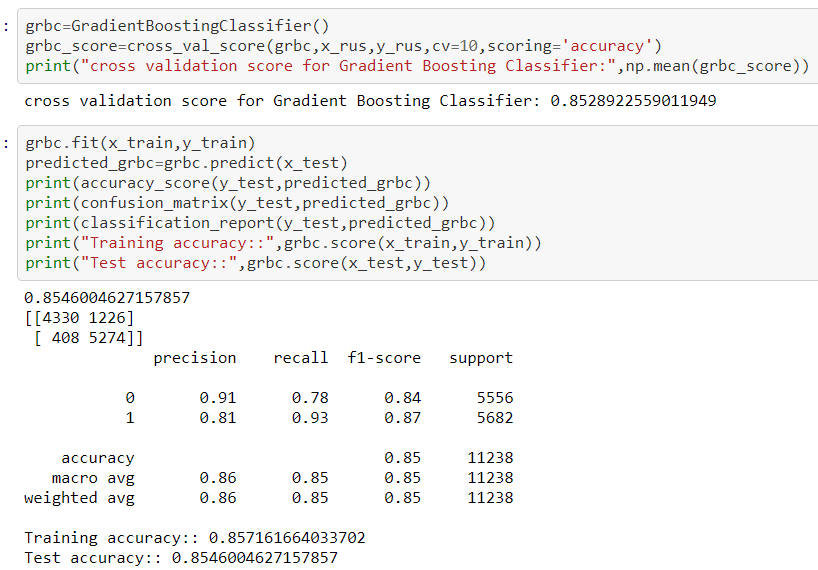
Random Forest Classifier:



ADA Boost Classifier:

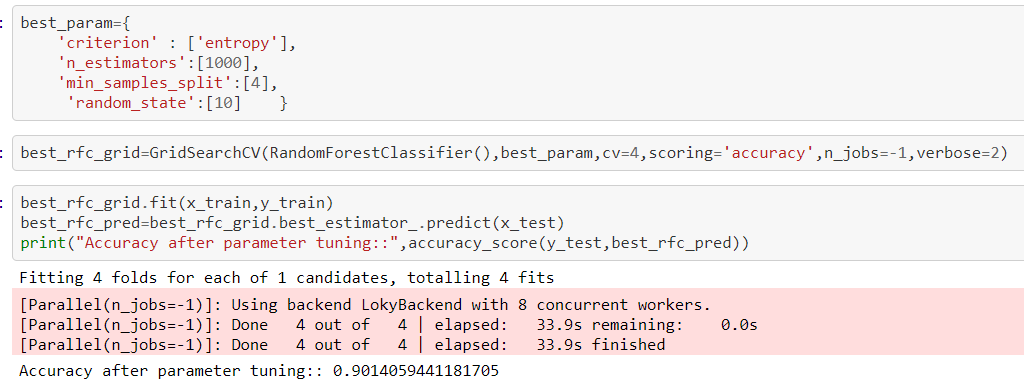


Gradient Boost Classifier:

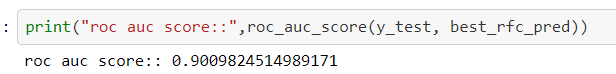


Random Forest Classifier is our best model

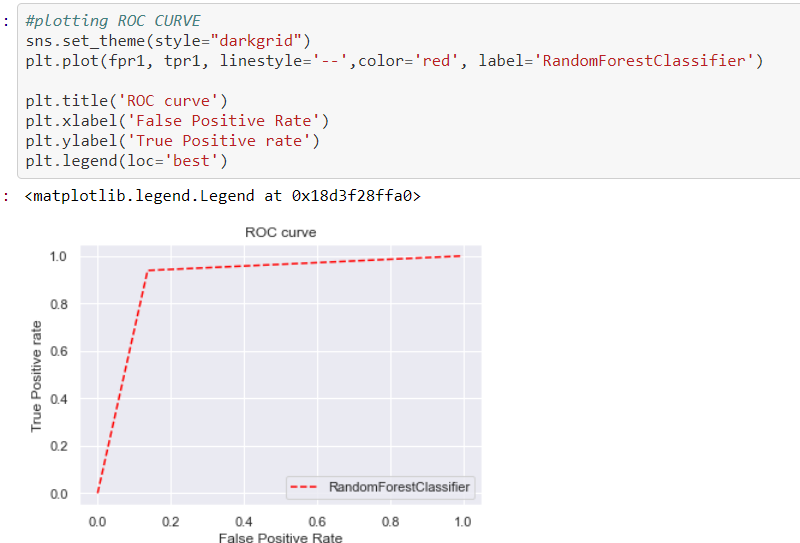
Hyper parameter Tuning











**Concluding Remarks:**

Feedback gathered during consultations has led to the successful integration of new questions in the past. Many factors are thoughtfully deliberated, however, before a new question is added or an existing one is modified.

As of now, I don’t find any such limitation in this project but apart from that, more learning, analysing & exploration is needed in Data science field that I am continuing. Because, **The Journey to Data Science is a Marathon, not a Sprint**