

## Census Income Project.

Problem Statement: -

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 or less then*.**

# 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 estimate 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.****

Here’s a step-by-step outline of the project:

1. Importing necessary libraries.
2. Importing dataset.
3. Exploratory Data Analysis (EDA).
4. Data Pre-processing & Feature Engineering.
5. Model building and saving.

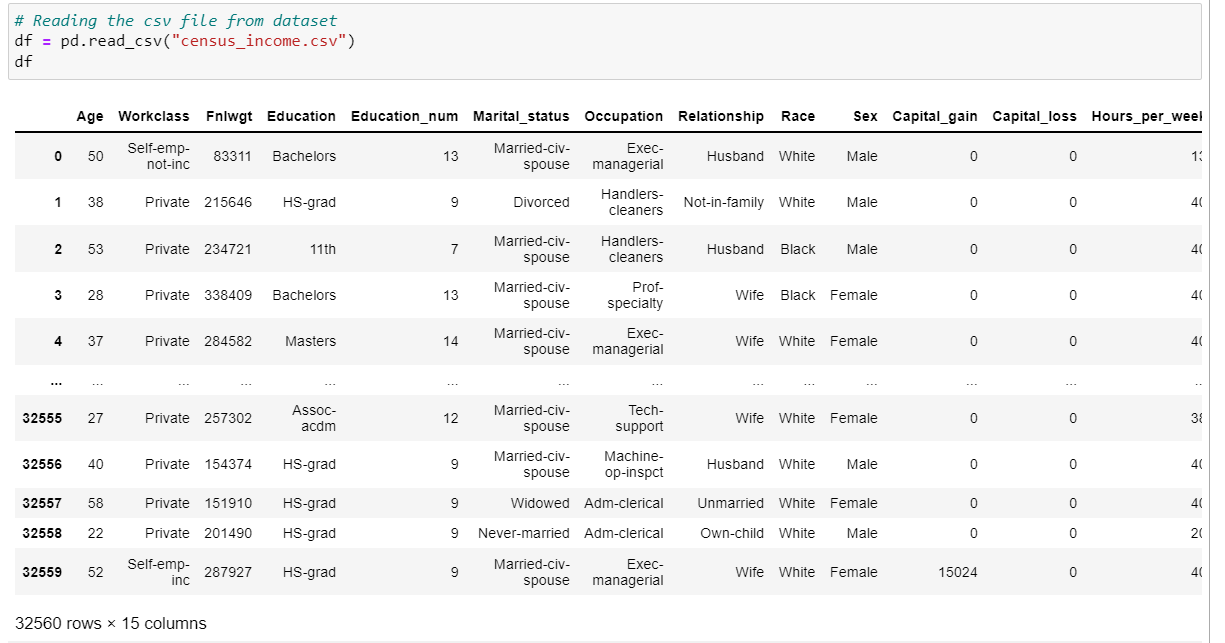
-------------------------------------------------------------------------------------------------------------------------

Now Let’s Import Some Basic and important Library: -

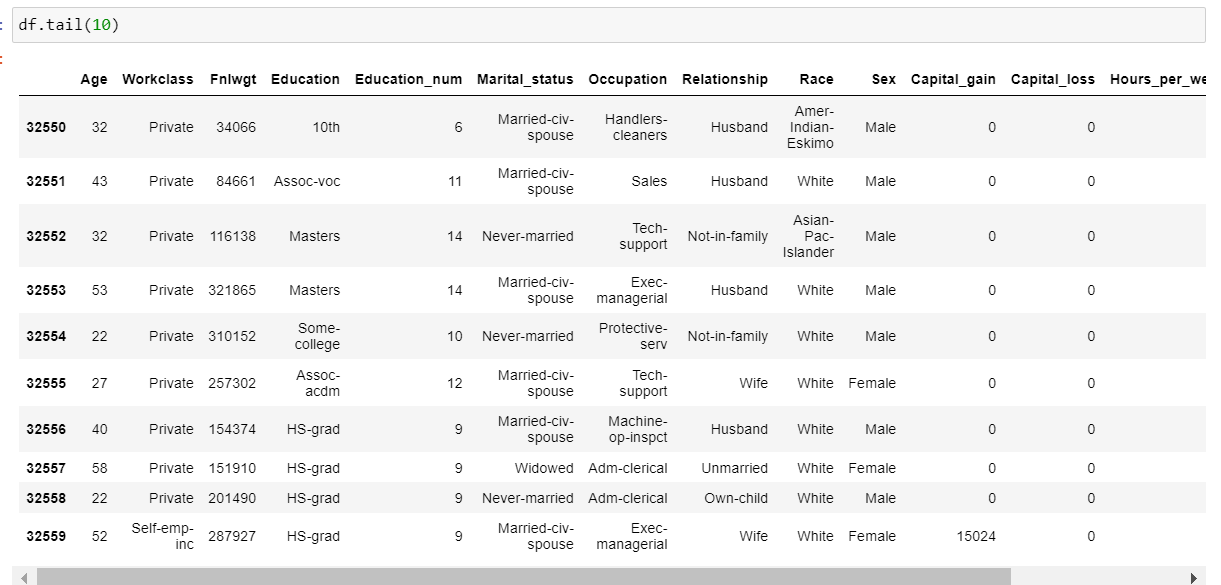


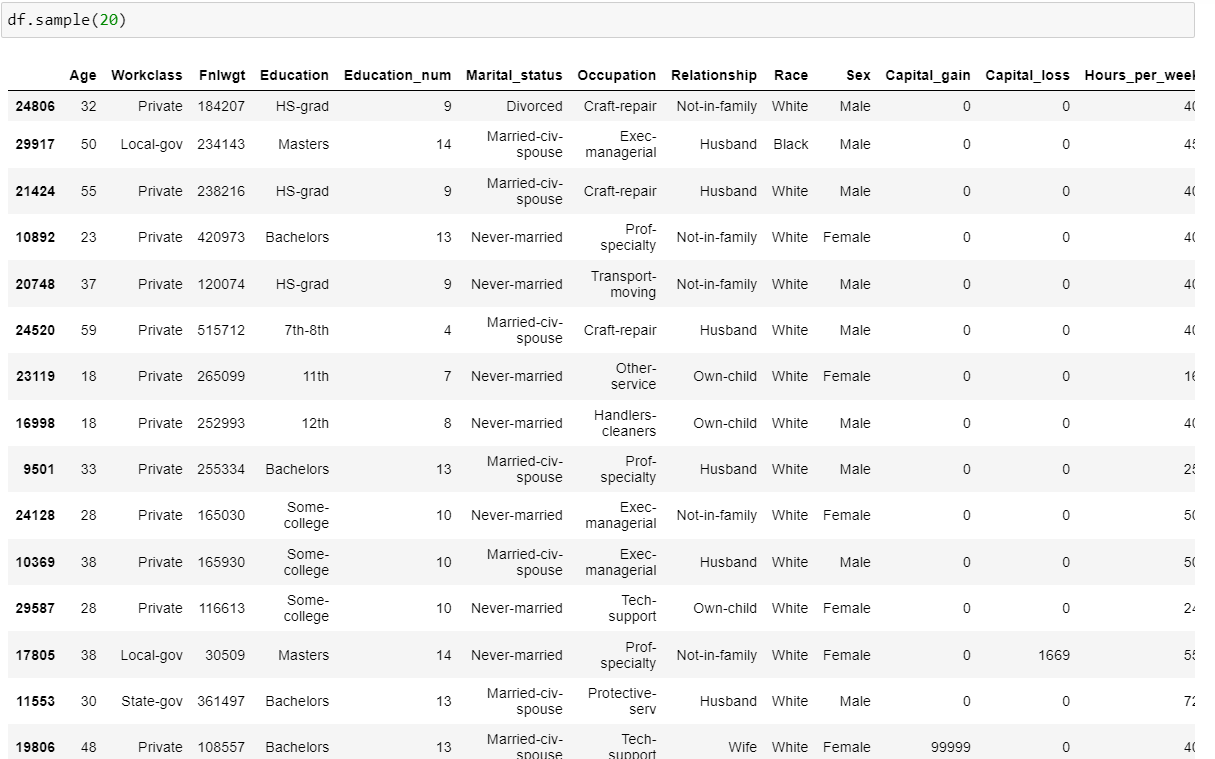
**Here I have imported Some the basic and import Libraries to get started.**

## Let’s Import the DATASET: -



## This dataset contain 32560 rows and 15 COLUMNS, THE dataset contains the details of the annual income of the persons. Here the INCOME COLUMN is the target variable

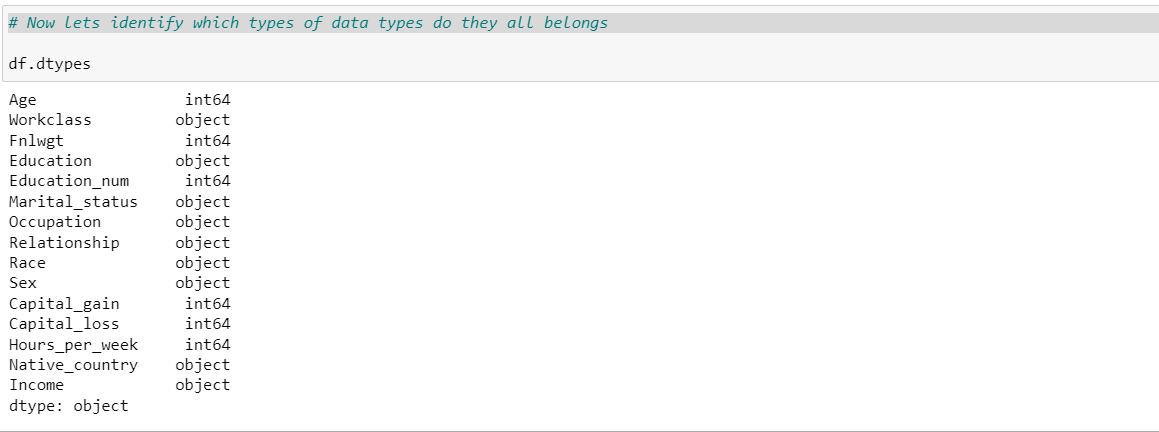


****

I have done tail and sample method to see the last 10 and 20 rows of the dataset. Here I have shuffled the dataset. we can see that there are some columns which are having’?’ as a Null Values. The dataset contains both numerical and categorical columns. We need to predict whether the income of the person is over $50k per year or not. Income variable seems to have seemed to be having two classes so it will be termed to be a "Classification Problem".

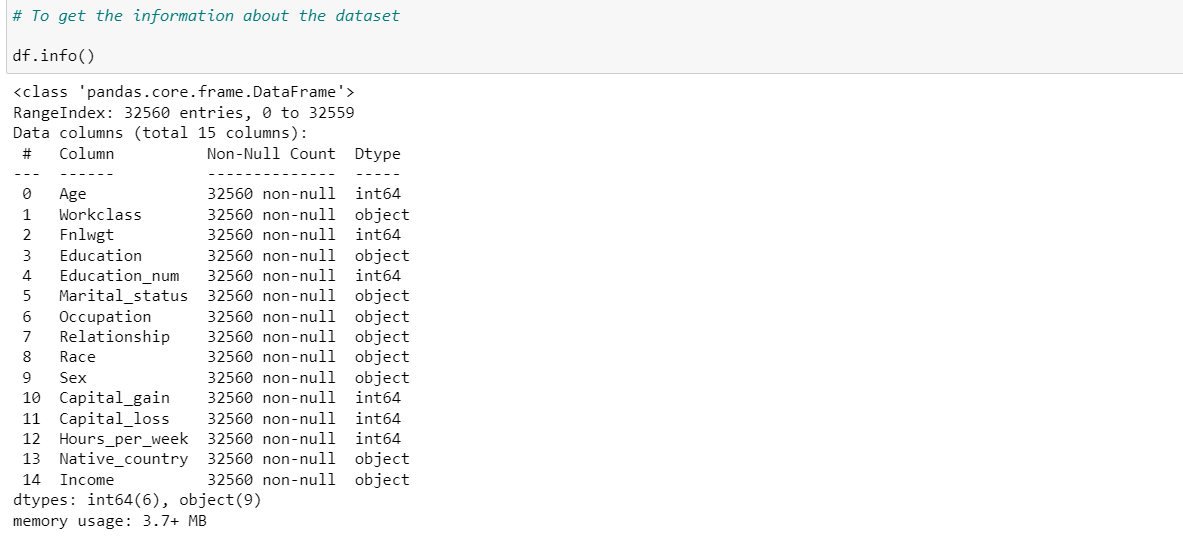
### And in this dataset, we need to see the value counts of many columns also.

Now let’s identify which types of data types do they all belongs

****

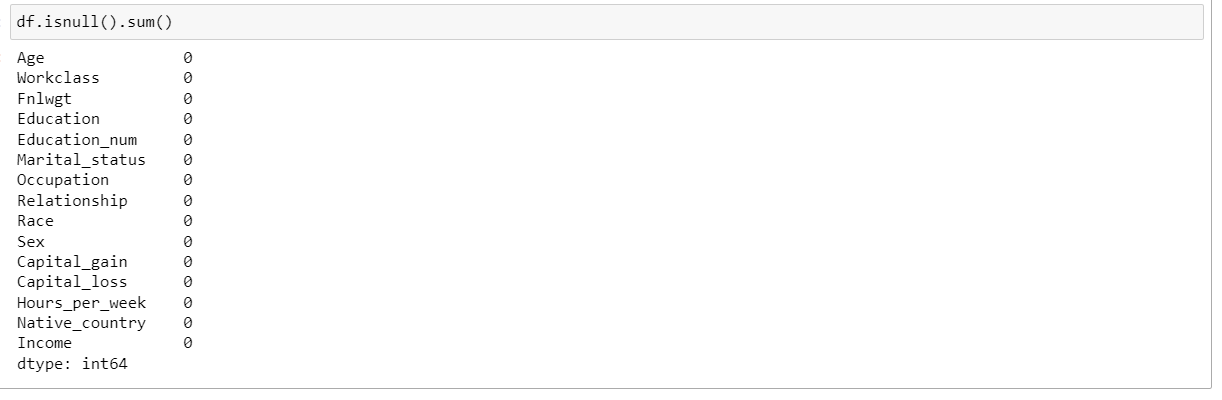
## We can see that our target variable is Object type which make it a classification PROBLEM. Rest all the input features are in both int64 and object. SO, for all the objects columns we will use Encoding techniques.

## EDA

****

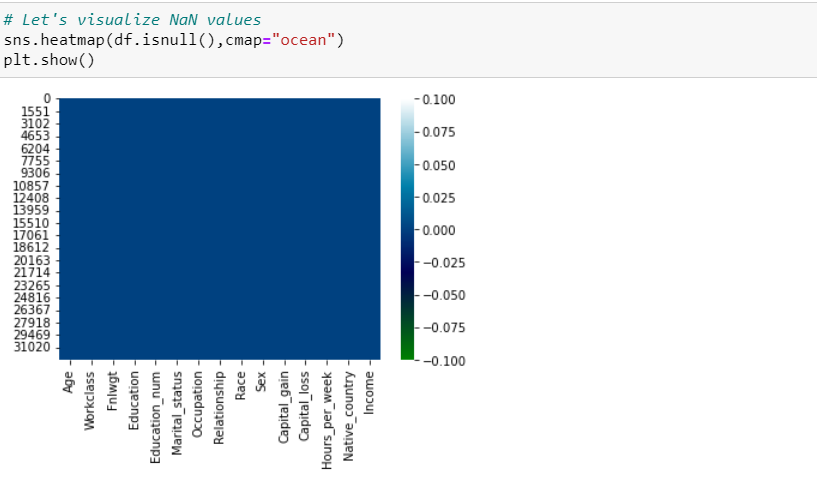
## Here we get the Information regarding Column, On-Null Count (columns having NaN Values) , Dtypes and Memory usage of the dataset ..We can notice one thing too that here it is showing that there are Columns having Null values ,which is present in the dataset as “?” .. These are Null Values and should be replaced according.

## NOW LET’S CHECK THE NULL VALUES :-



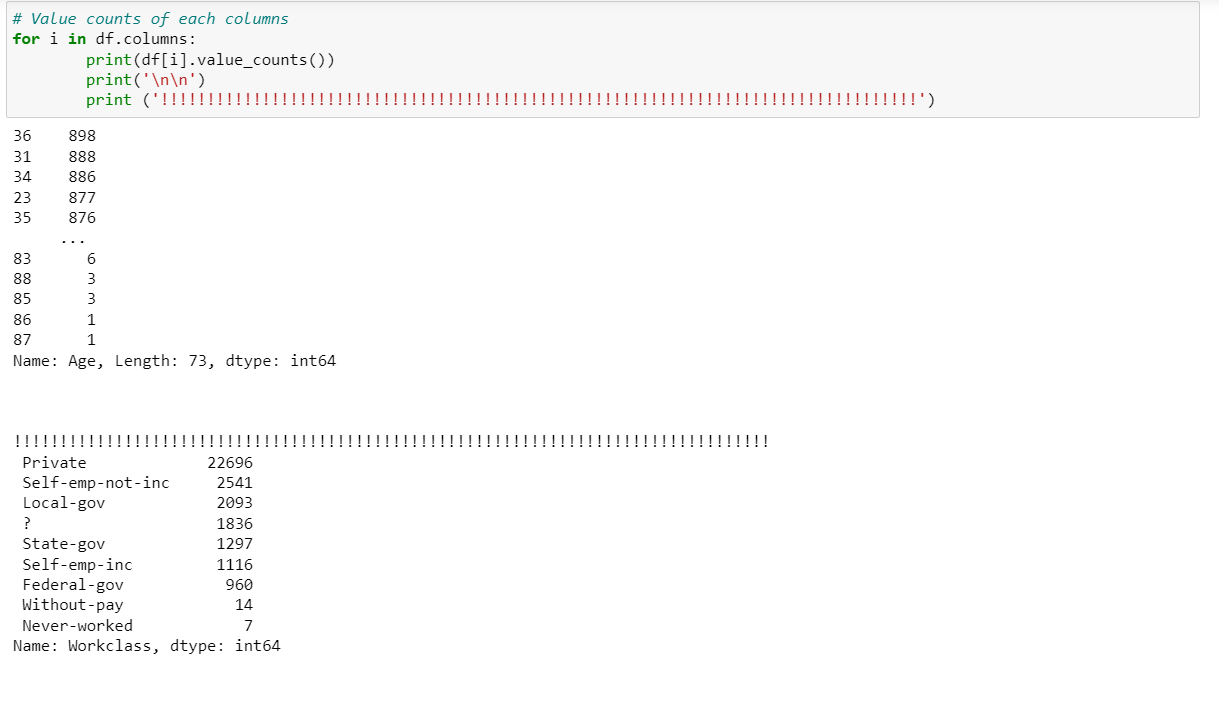
### Here it is showing that there are 0 null values in all the columns and we know Null values are present as ‘?’ that why it is showing NaN Values as 0. We will replace it later.

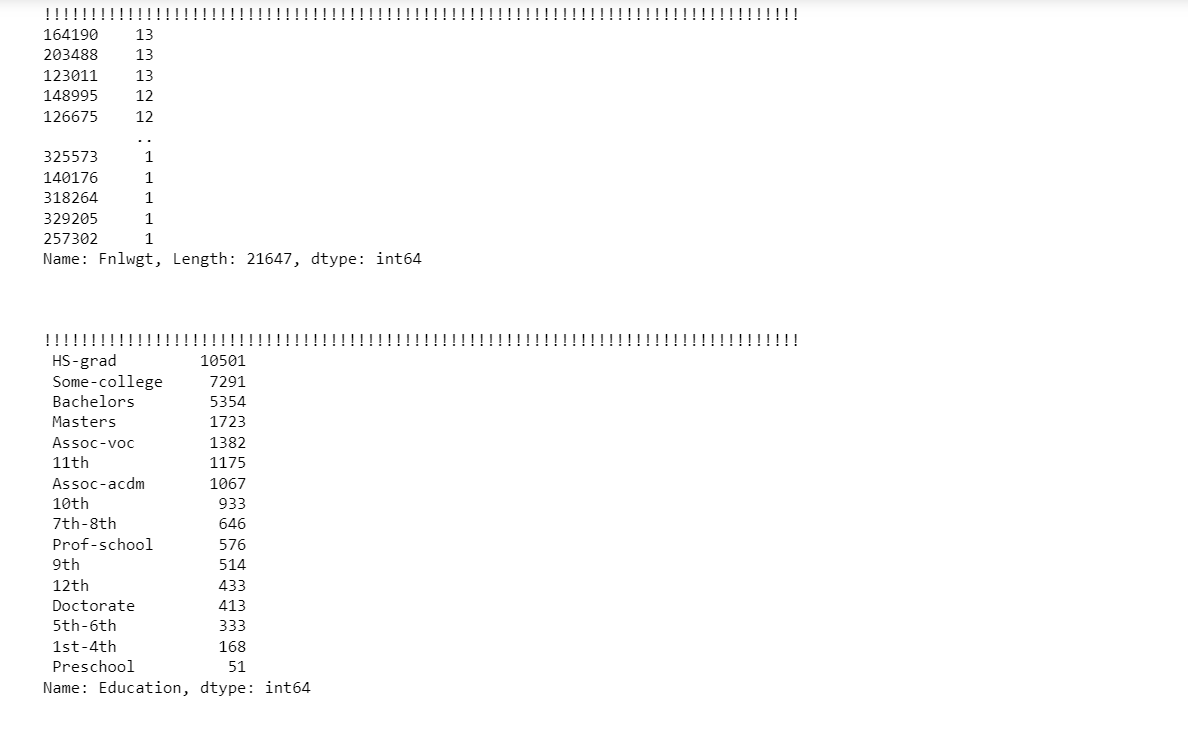
Let's visualize NaN values :-

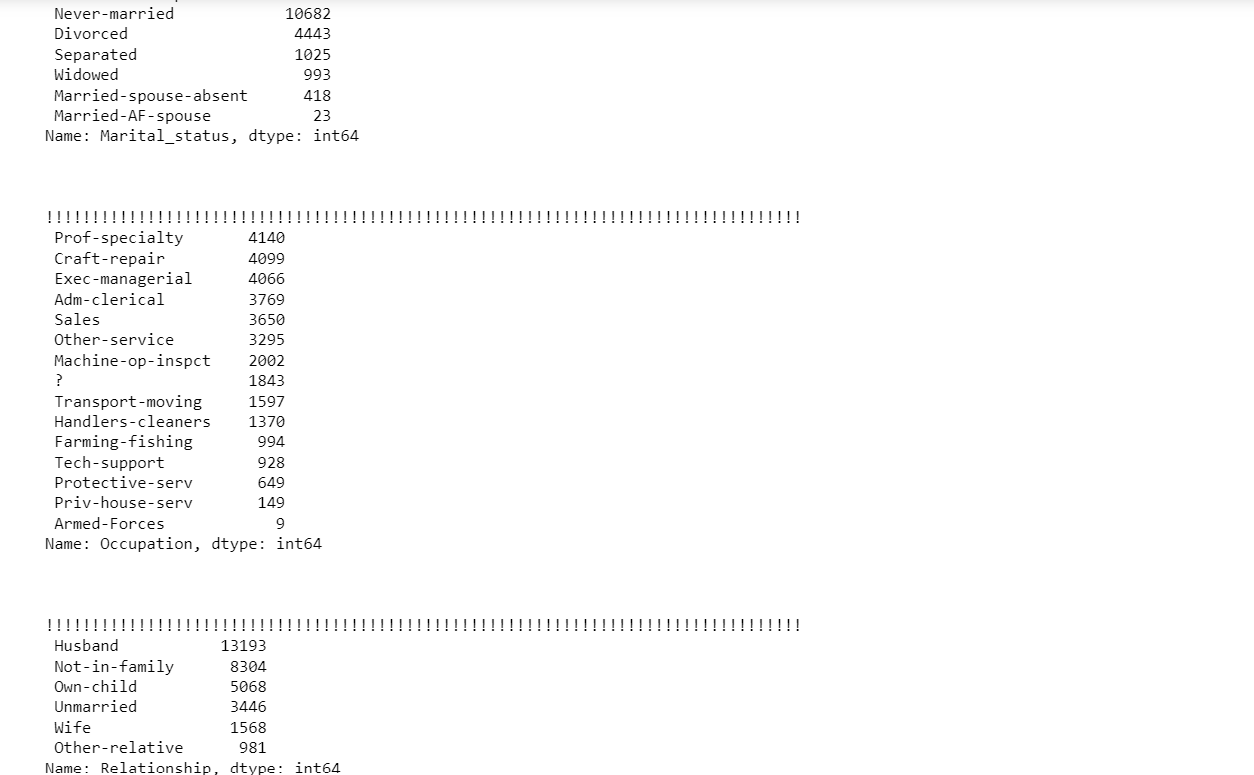


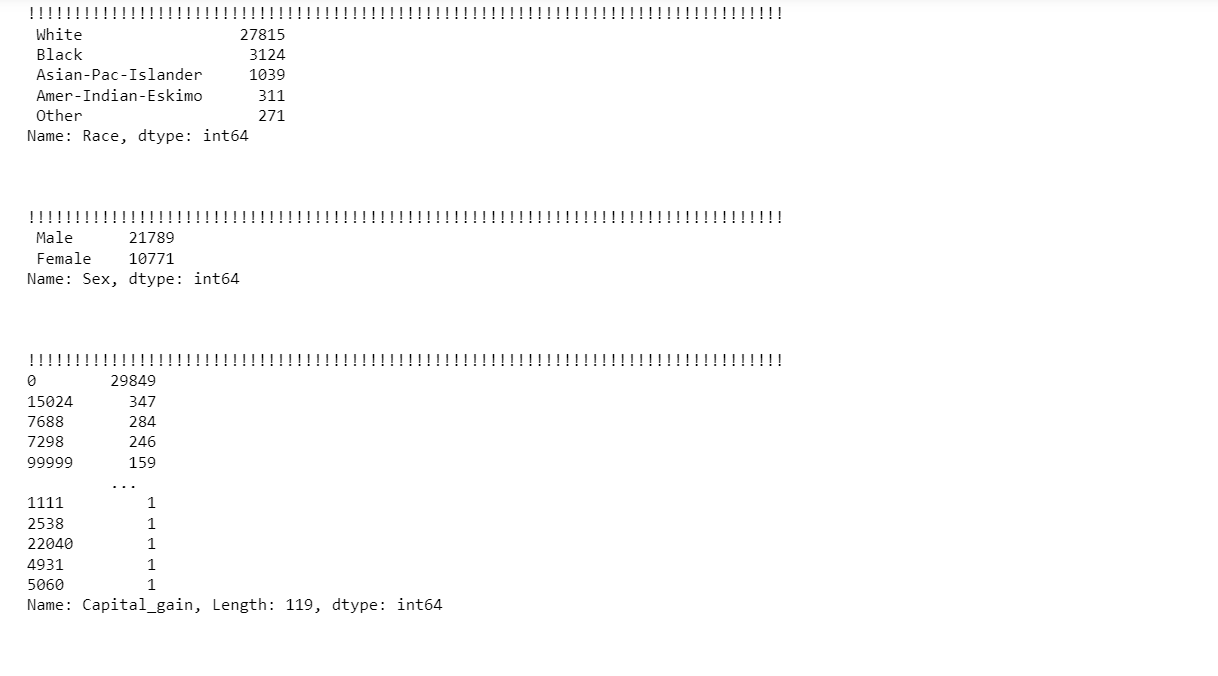
No NULL values are present in the dataset as per Visualization.

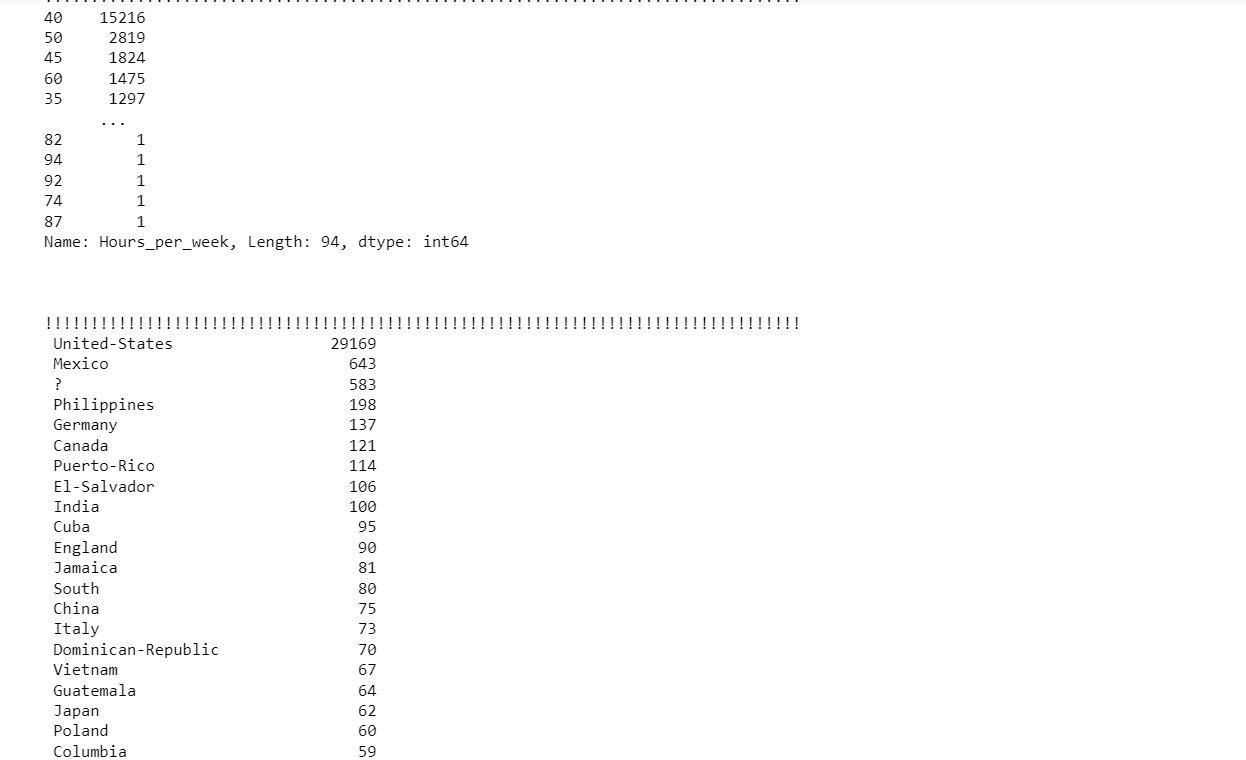
### Now Let’s see the value counts of the columns, so that we can check which column is having? (NaN)

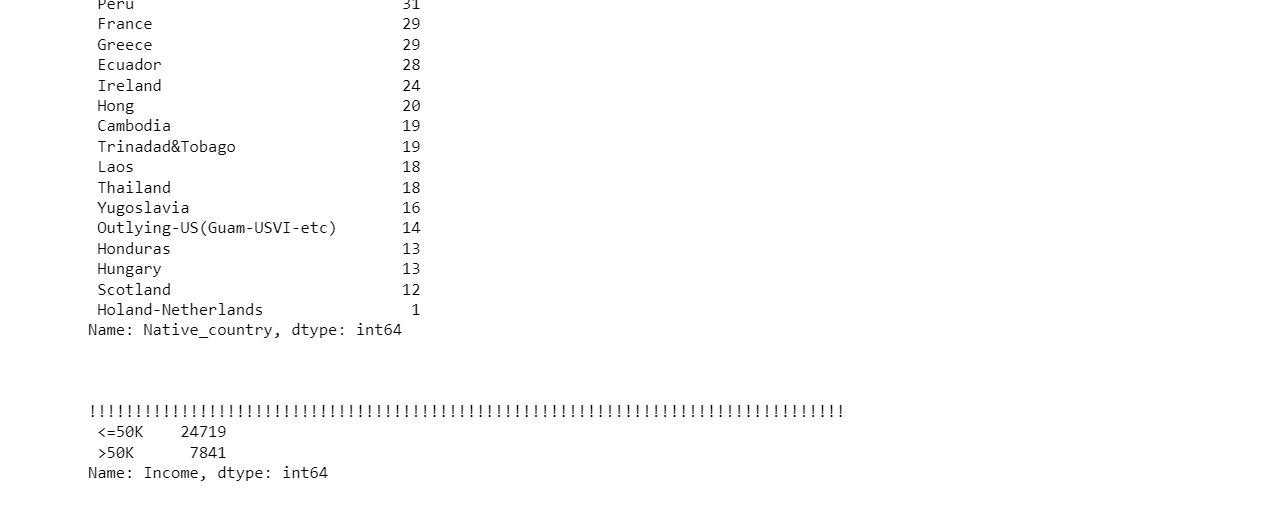






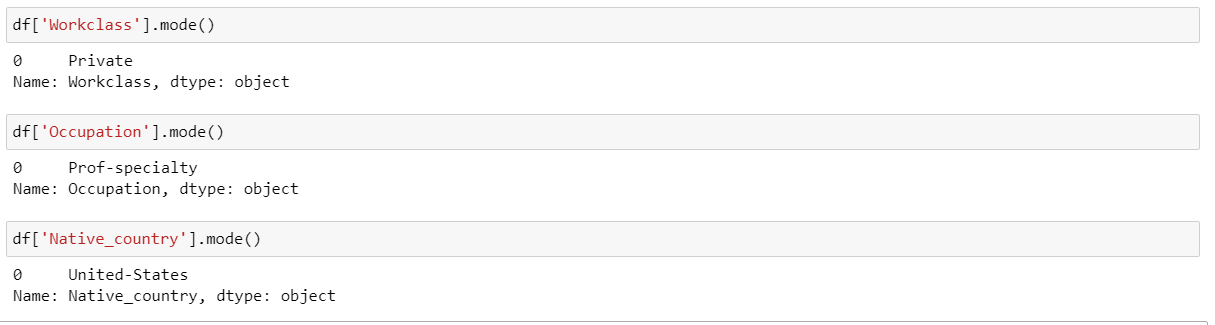




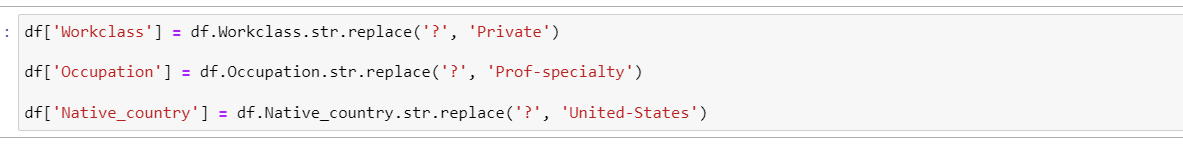


## Workclass, Occupation , Native\_country columns have NaN values(?) and need to be replaced

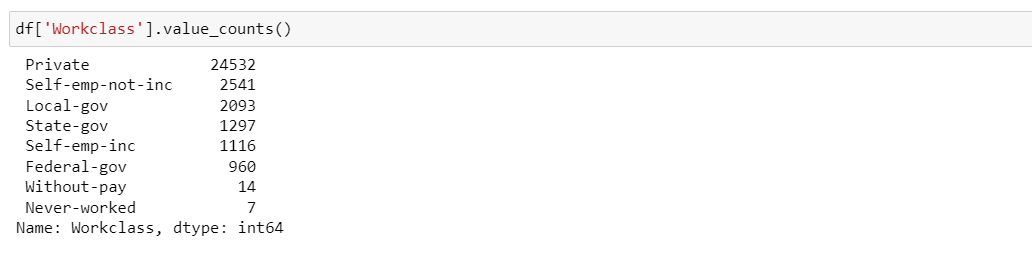
## We will replace the NaN Values with the most frequent as all the columns are CATEGORICAL.

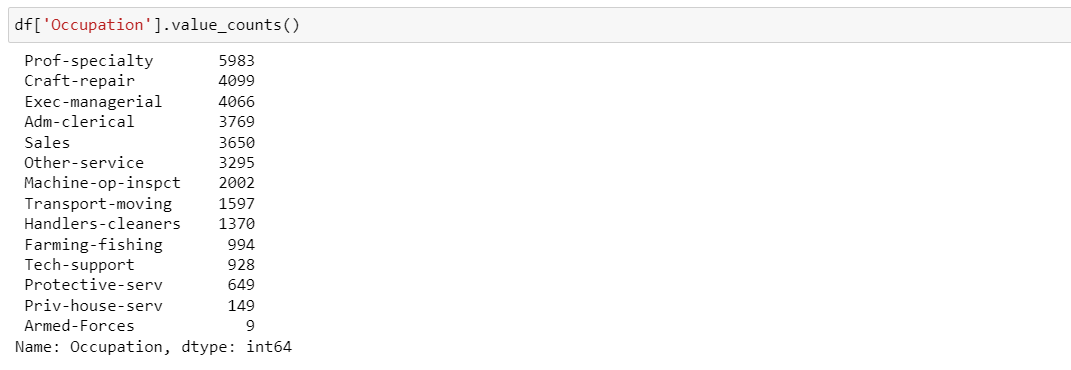


## Now Let’s Replace THIS “?” with most Frequent in the columns

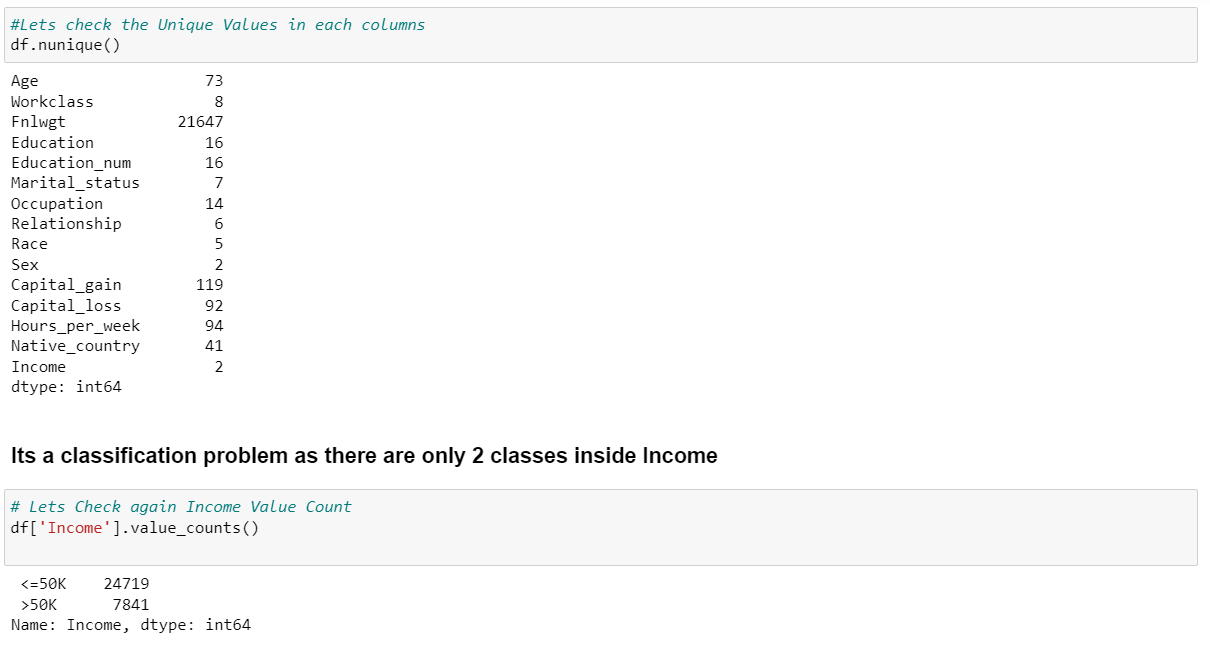


Let’s Check Once whether “?” have been removed from the column or not.

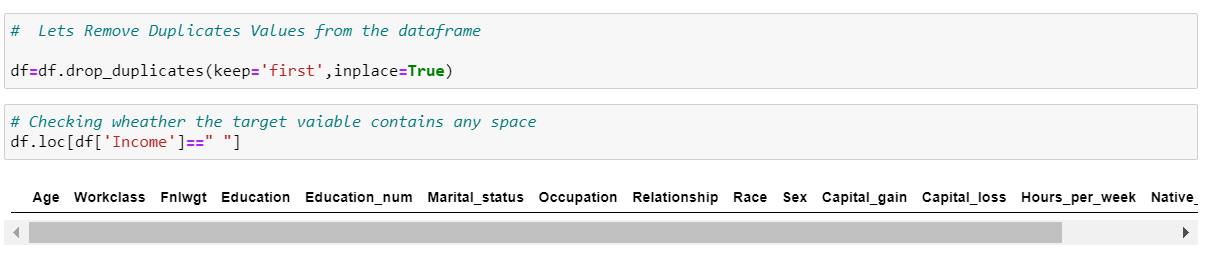




From some of the columns above we can see that “?” have been removed from the dataset.

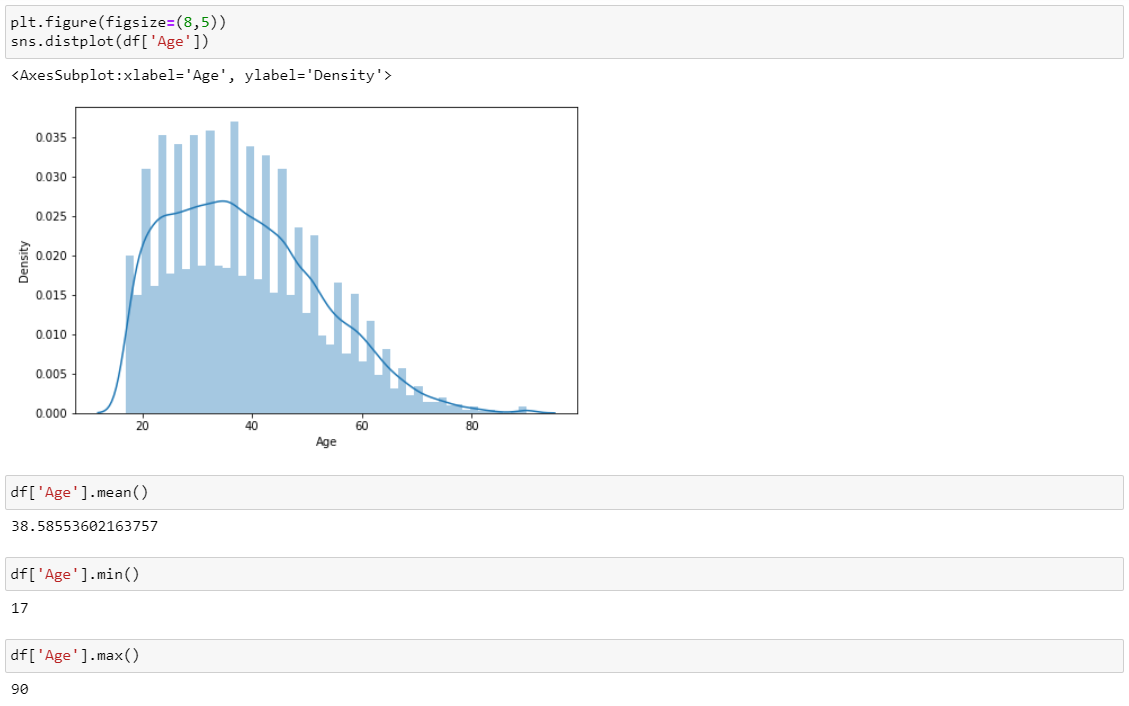


## 1st of all we will balance the dataset as it is imbalanced by Using SMOTE. And as i concluded earlier it have 2 classes which is more than 50k and less than 50k

****

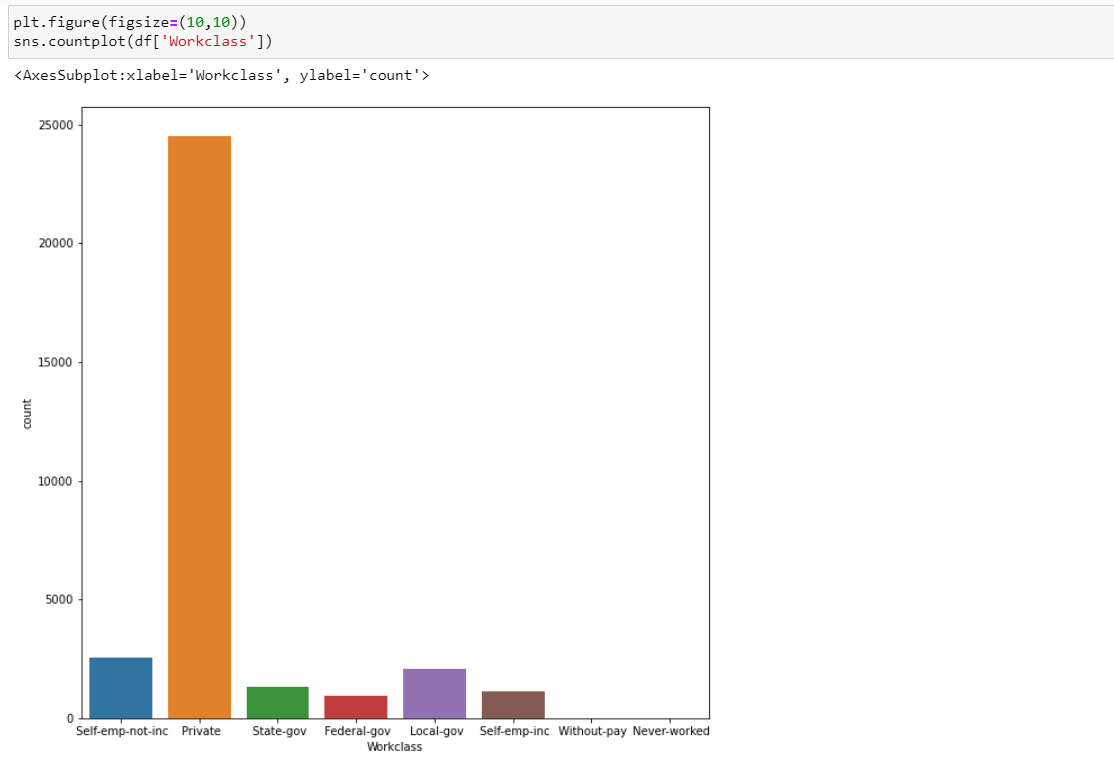
**Here I have dropped the duplicate values and also checked whether it has any space present in the dataset.**

## A. Univariate Analysis

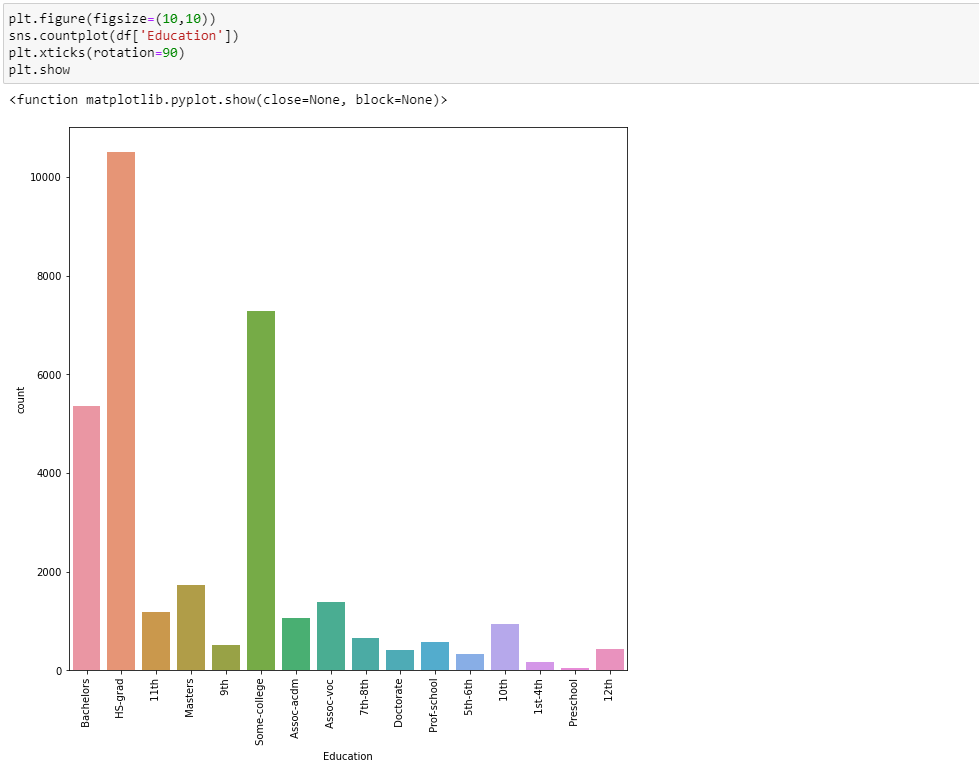
****

## here we can conclude that the average aged people are from 38-39 year and we can acknowledge the same through the graph above .

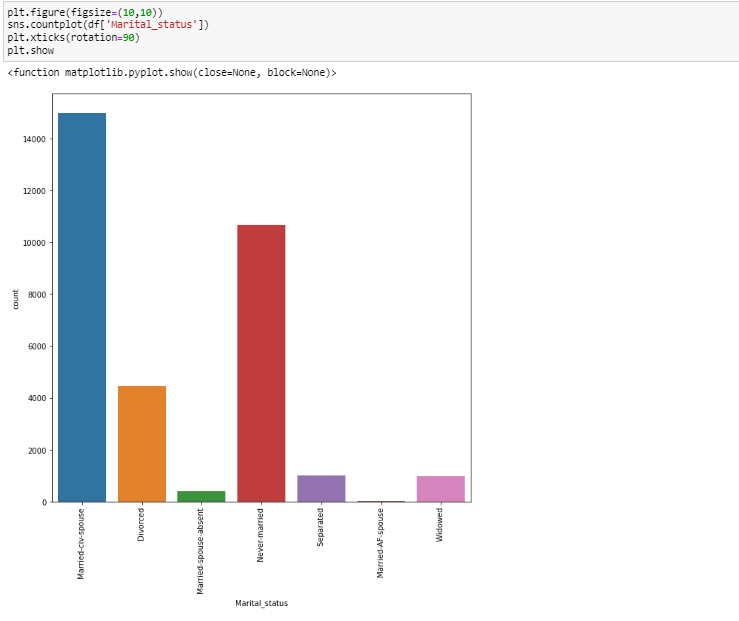
### The minimum age of the people is 17 and max age people is 90.



### The maximum people contributing to the census belongs from private sector and the least are from Without-pay and Never-Worked group.



## The maximum number of people who are contributing to Census are from hs-grad , some-college and Bachelors And the least are from Pre-School .

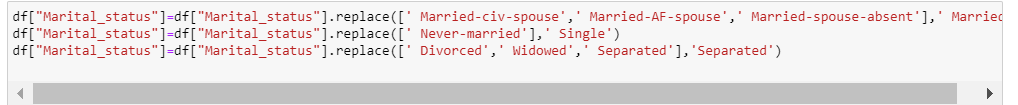


## Married­civ­spouse corresponds to a civilian SPOUSE.

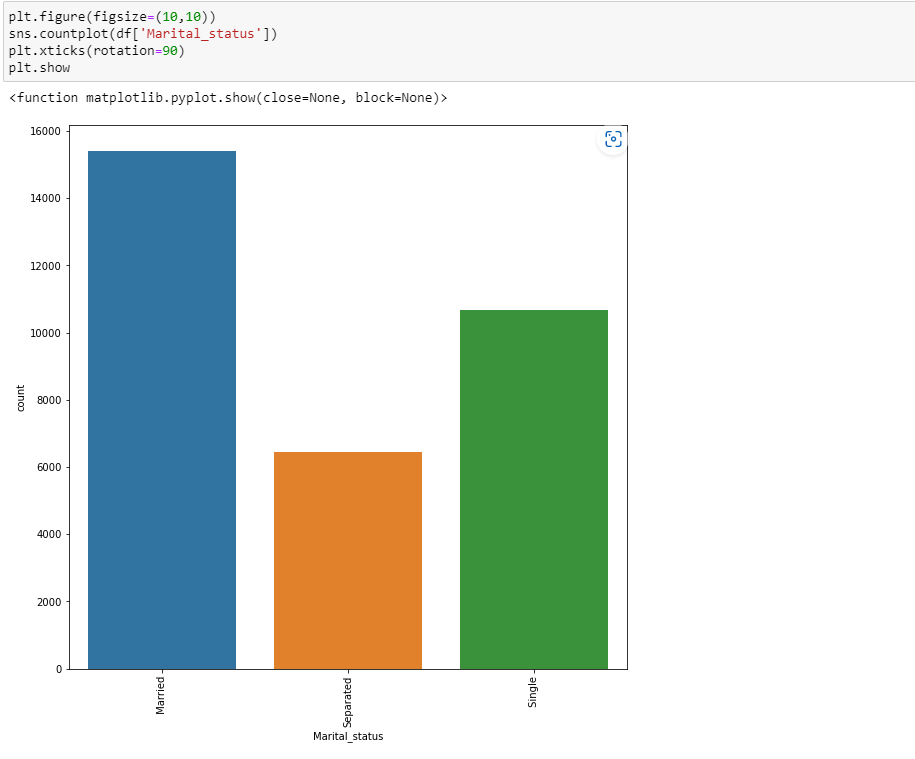
## while Married­AF­spouse is a spouse in the Armed Forces.

## Married-spouse-absent -> Married people living apart because either the husband or wife was employed and living at a considerable distance from home, was serving away from home in the Armed FORCES.

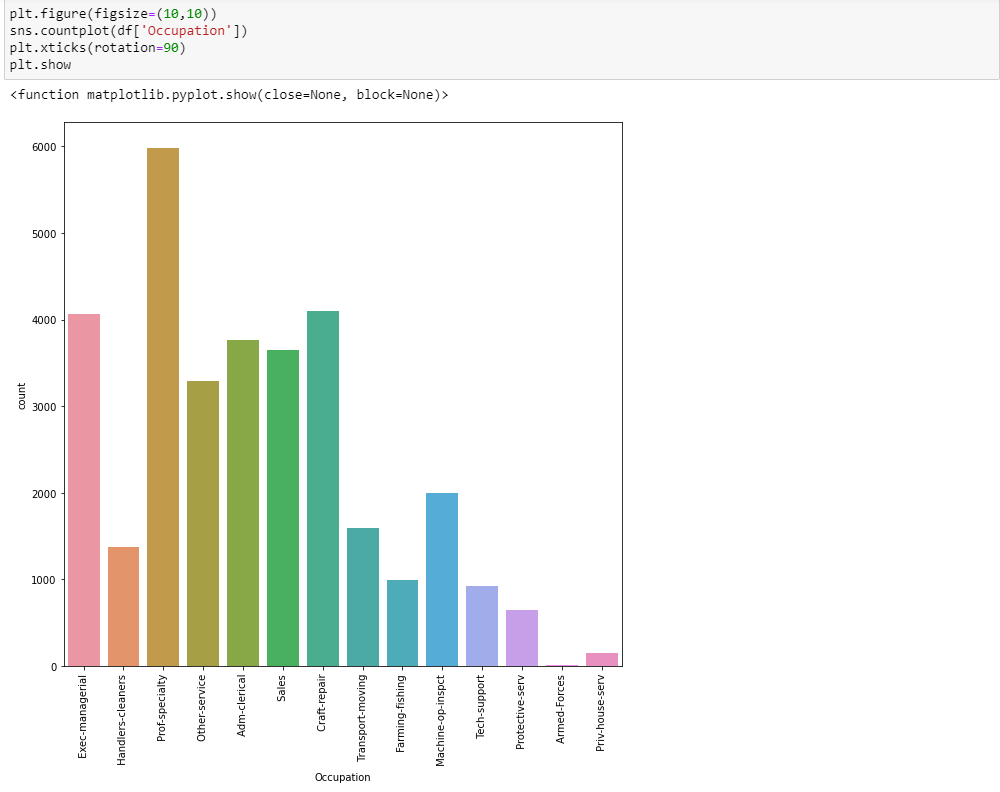
### Here we can group them easily in married, separated and single classes inside marital status .



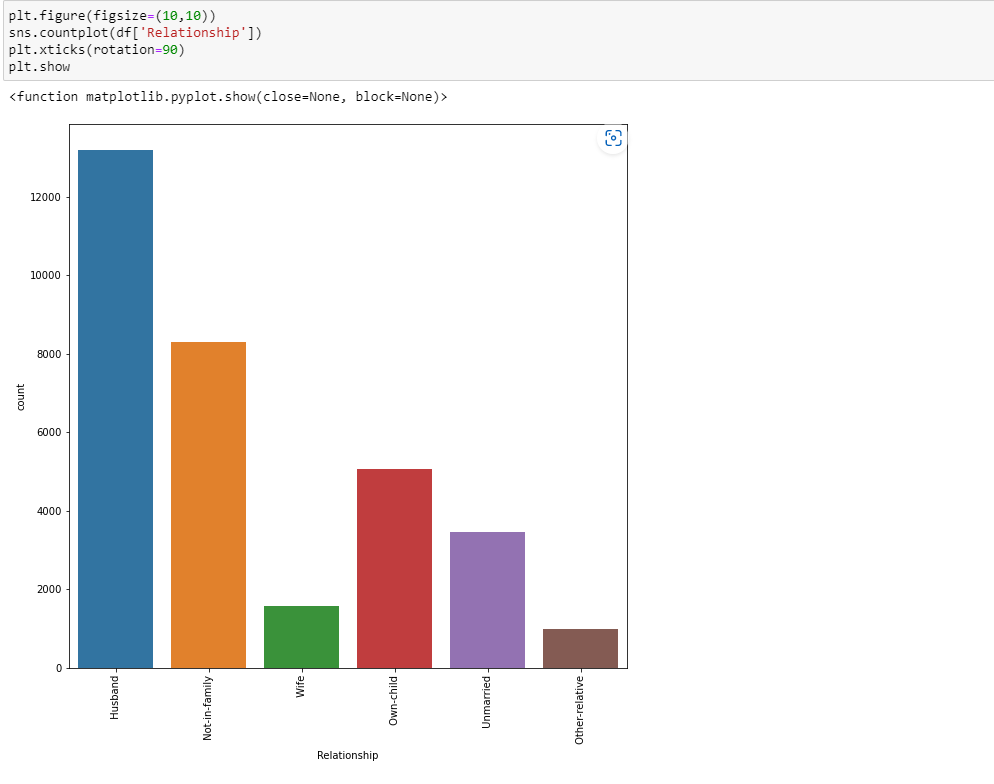
Plotting Again . . .



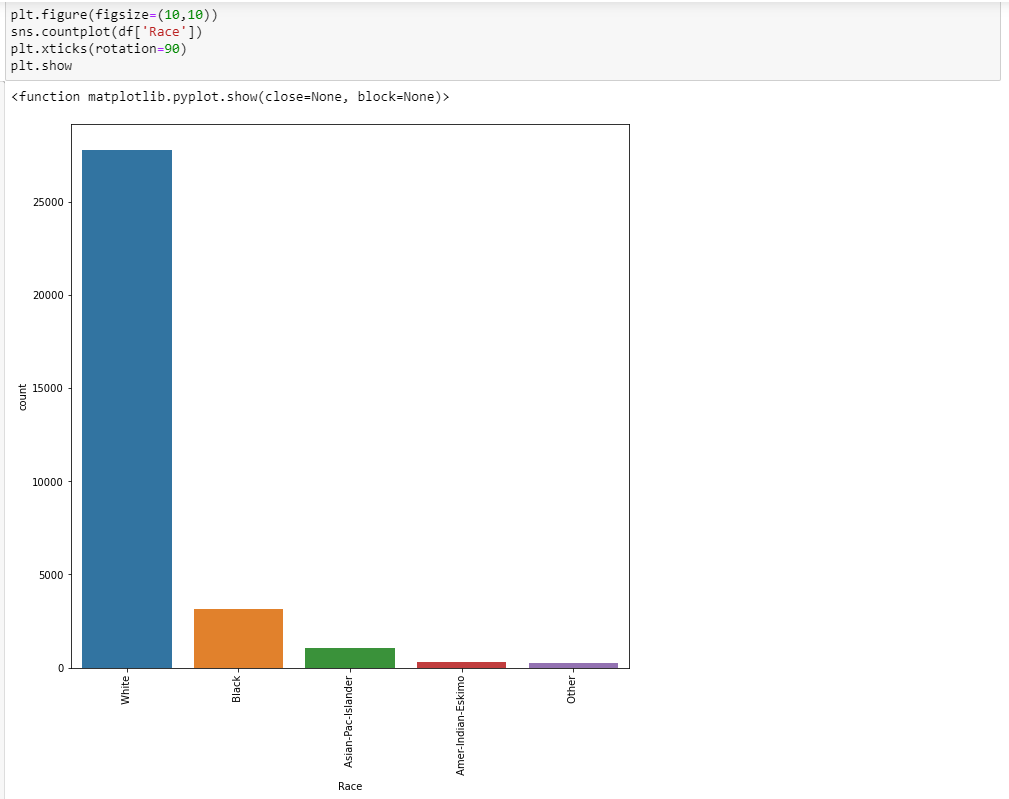
### So here we can find out that the married people are contributing more to the census income followed Single and the least is separated.



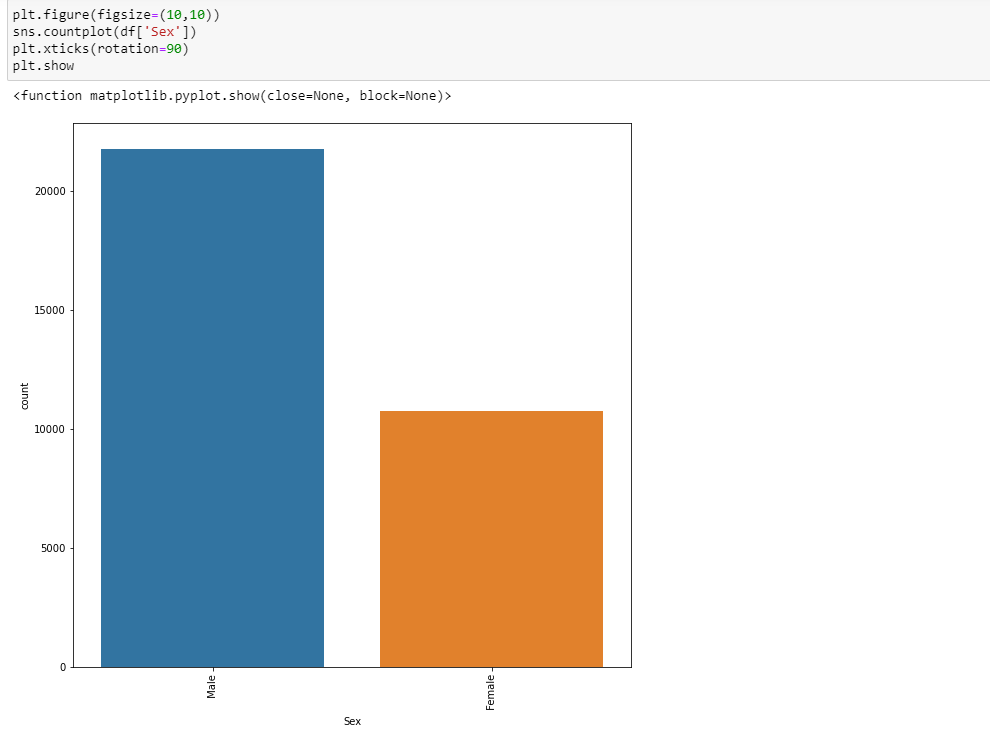
### Here we can see that the maximum contribution to census income is given by Prof-Speciality and the least is given by armed forces.



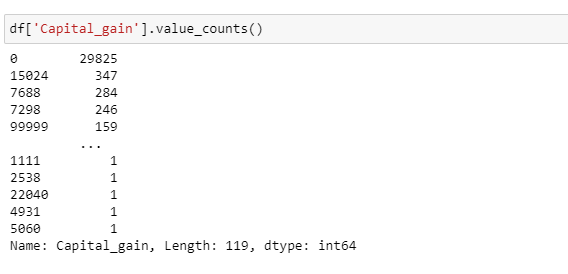
## The count is high and maximum in Husband and the least is in other RELATIVE.

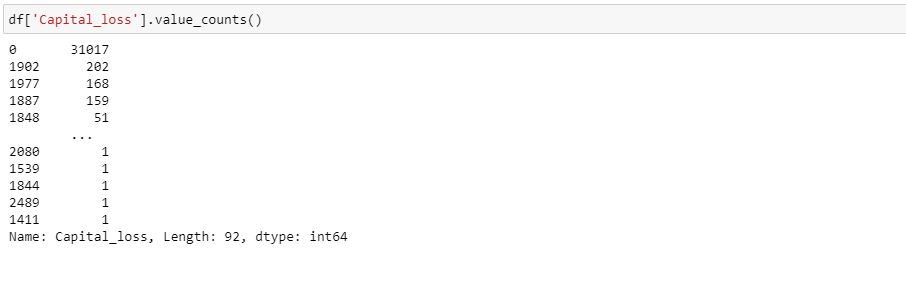


## The Counts of White in the race are maximum and Other are LESS.

****

### The Count or contribution of male is more in income census as compared to Female.



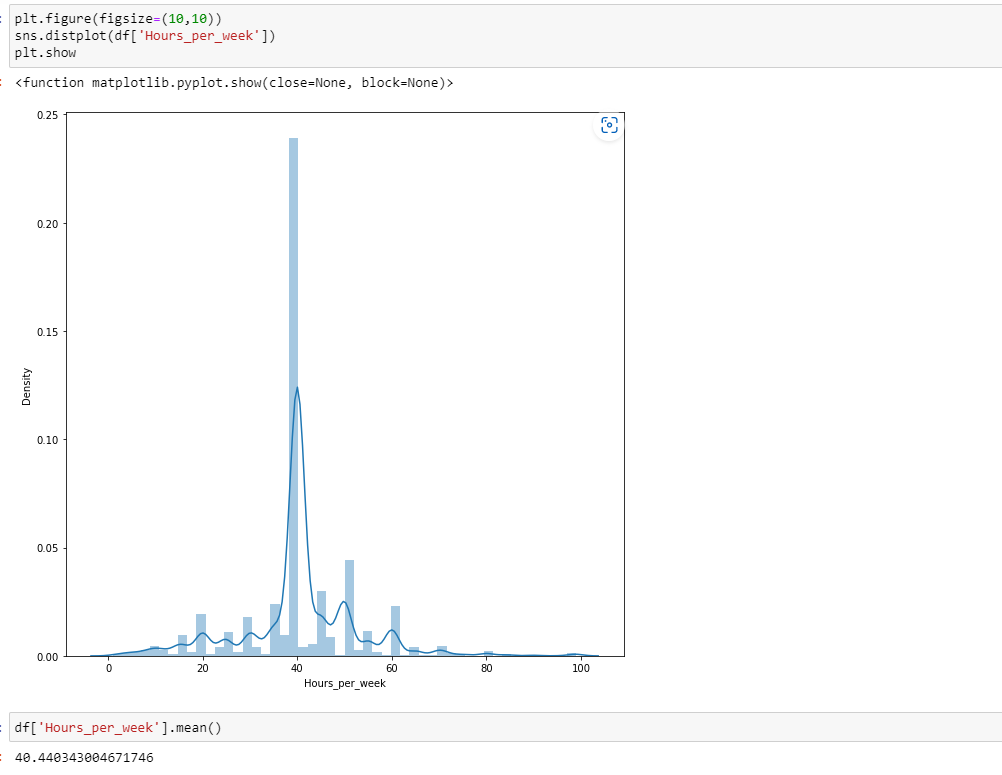


## We can see that In Capital gain column the maximum counts are of 0 which 29825 in count which seem to be covering more than 80% of data.so will drop it columns it .

## Here in capital loss too we can see the same thing ,0 in the columns are more so will drop this COLUMNS.

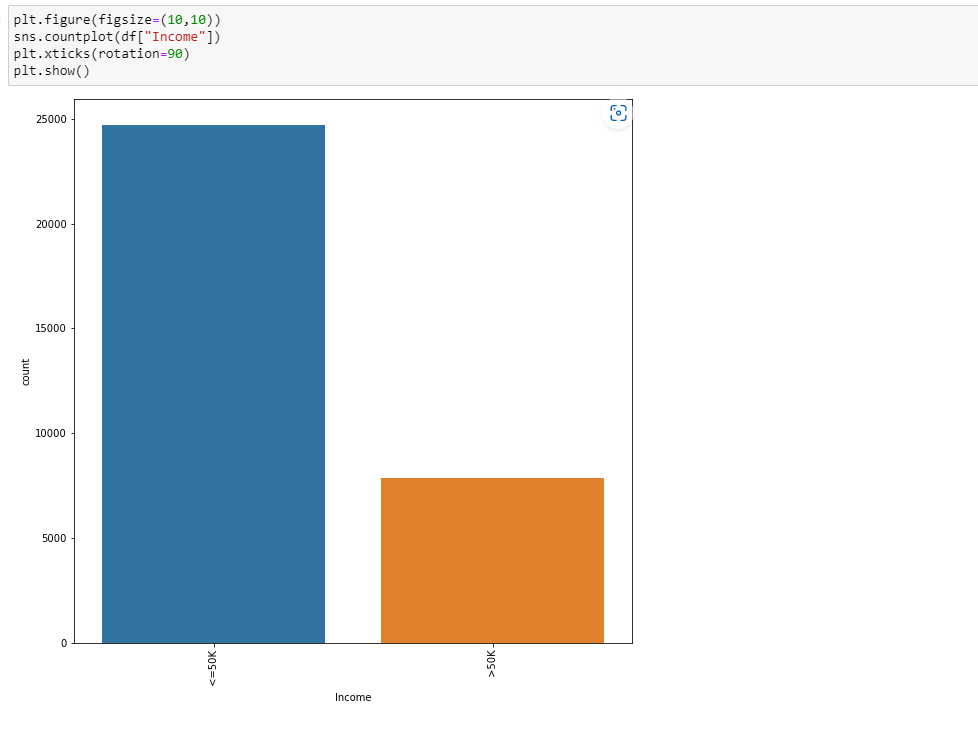


I have dropped these two columns because they having max number of values in hundred.



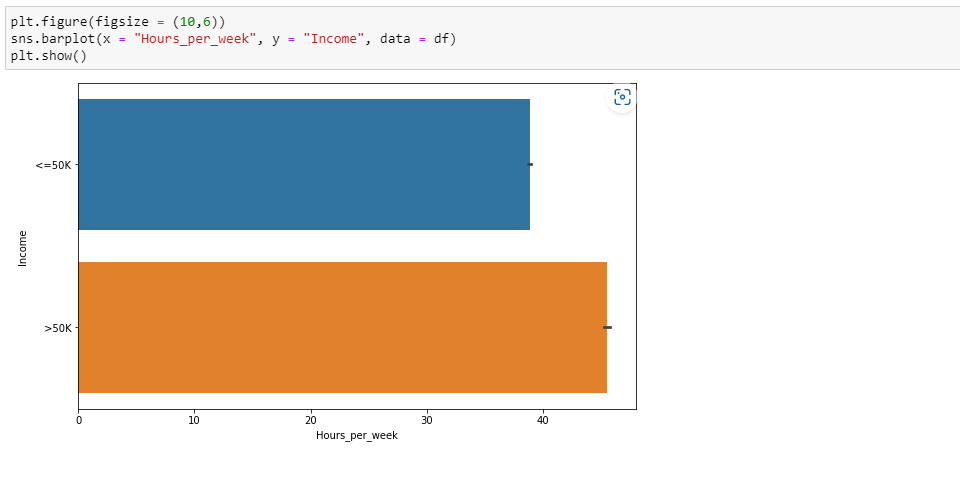
## Yes we can clearly that the average hours per week is 40 hrs ,which is contributing to census income

..

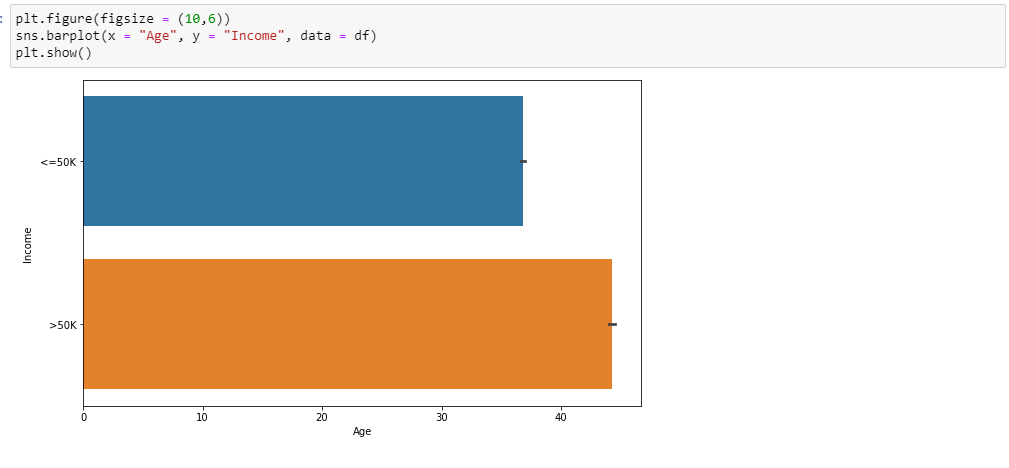


## This is Our Target VARIABLE. The count of people are more who are paid more than 50k in a YEAR.

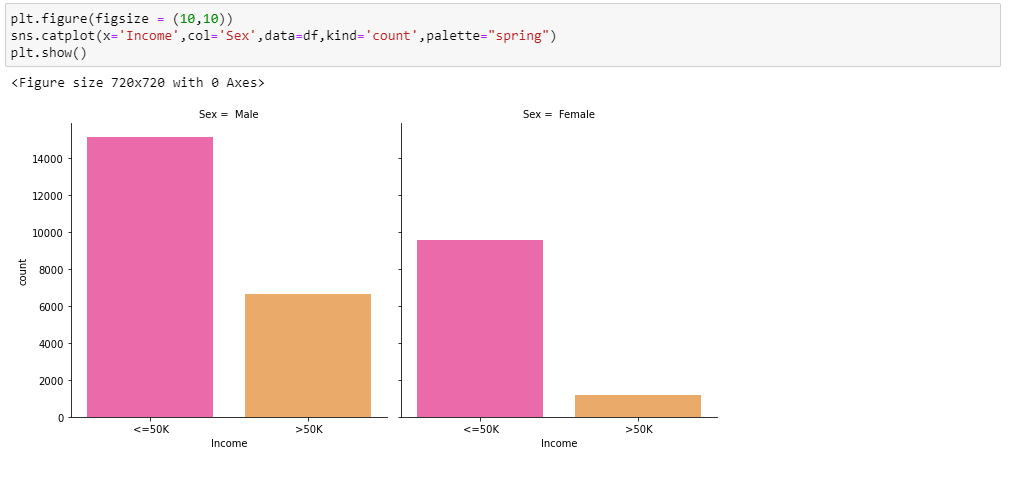
## B. Bivariate Analysis



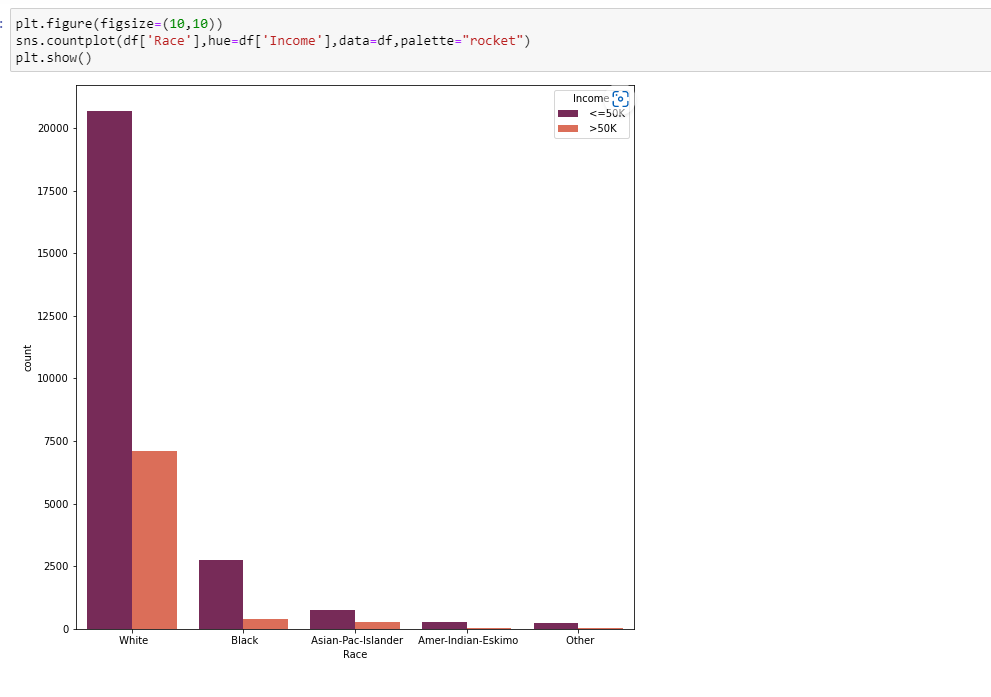
## People who are working for more hours in a week are paid more as compared to the people who are working for less hours in a WEEKS.



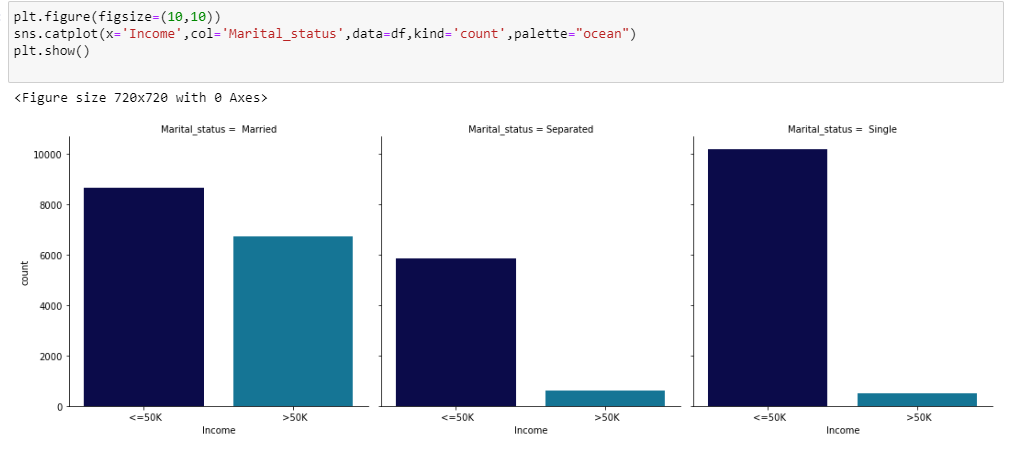
## Not the exact conclusion we can say but to one extended we can say that as the age of the people increase their salary also increases this may be because of experience and Nowadays people are paid more irrespective of their age but on basic of their education and knowledge and we can clearly see that on the GRAPH, AS people who are earning above 50k are falling in age group of 0 to 40+ and people who paid less than 50k fall in the age group which above the age till 38



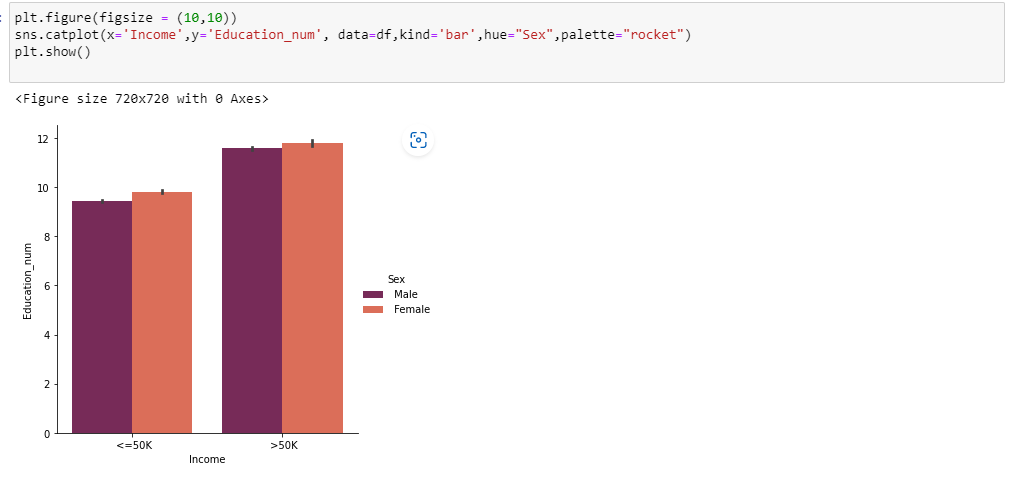
### More Male are paid income more than 50k than females and Female are paid less who are paid above 50k .



# Here we can that White Race People are paid more as compared TO OTHER and are more in number in case of EMPLOYMENT.

****

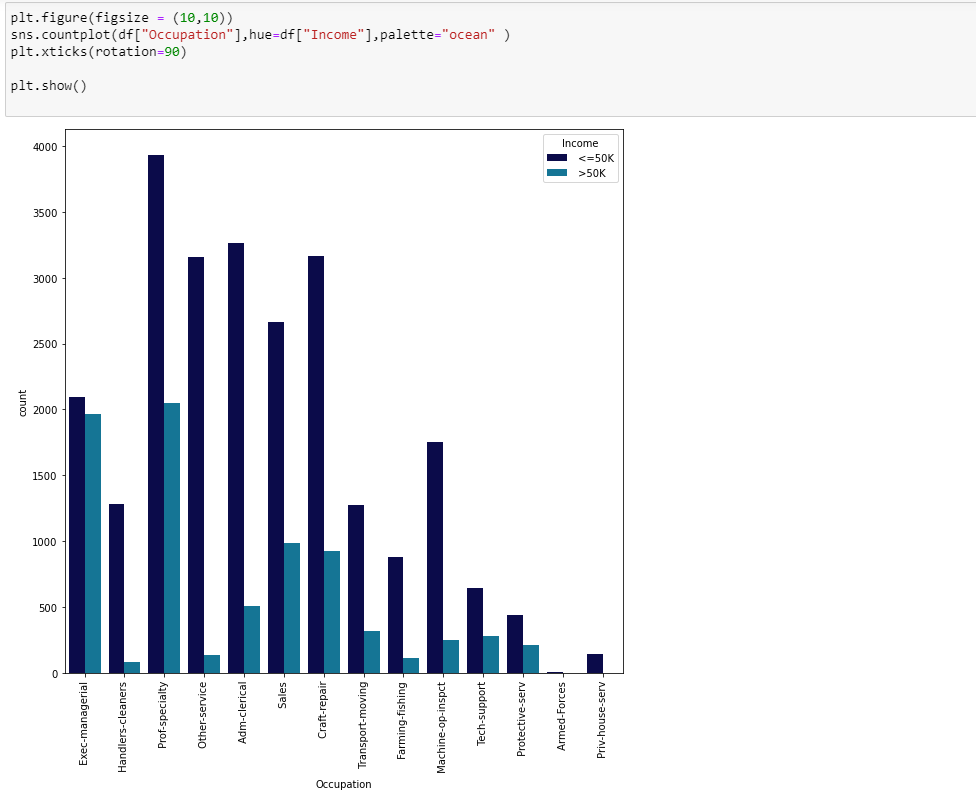
## Here i can see that singles are having maximum income followed by Married people and least income is with Separated.



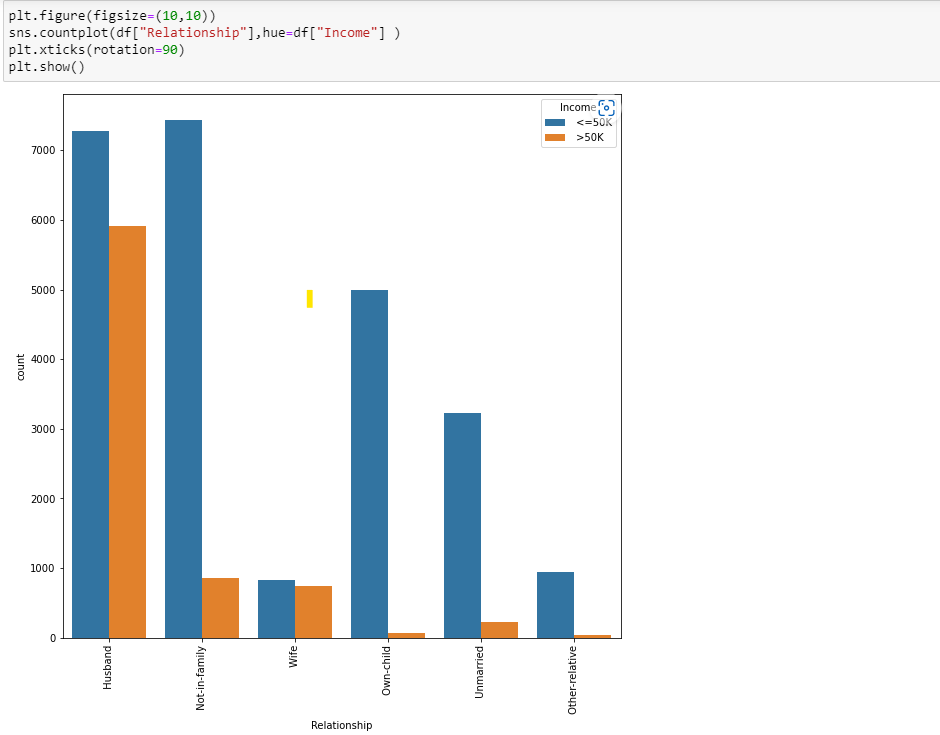
## The Income is more than 50K for the people having high education number. Here both gender have the income of more than 50K.

****

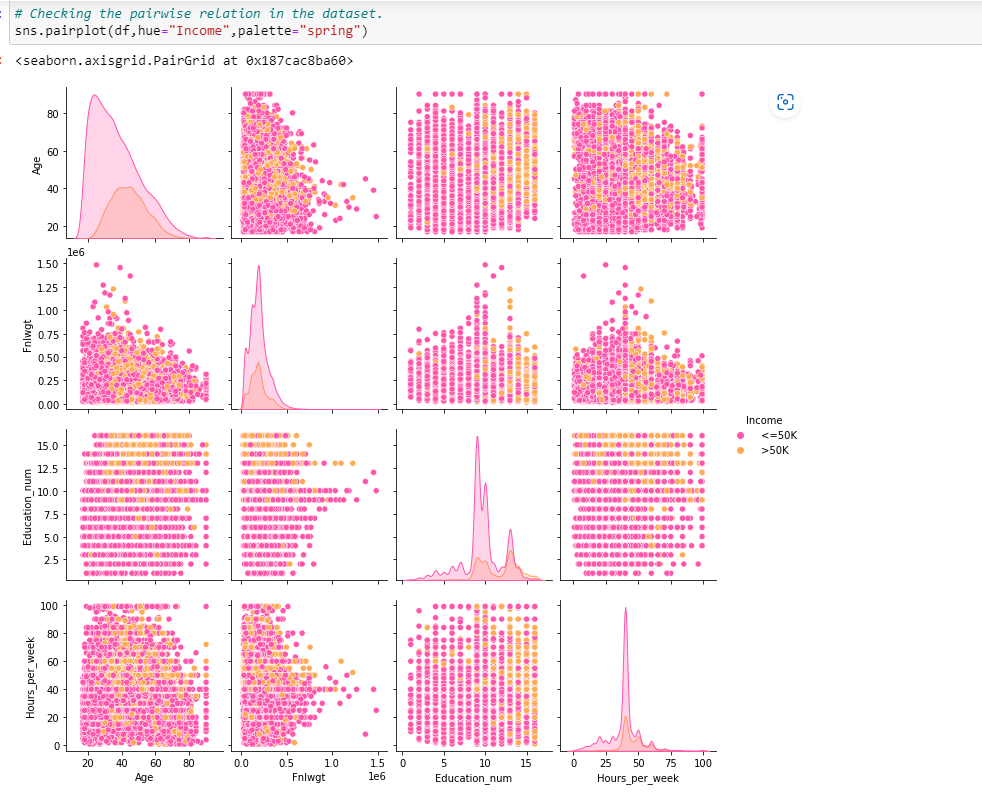
## Here we can see that maximum income is paid to bachelors and the maximum people who are paid less than 50k belongs from HS-grad



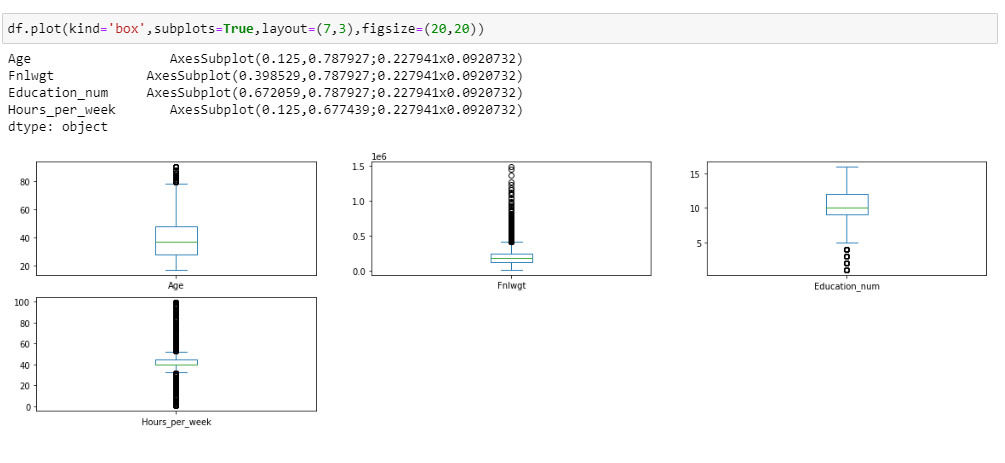
## The maximum income is paid to Prof-speciality and Exec-managerial and the least is paid to Armed forces and Priv House Servant



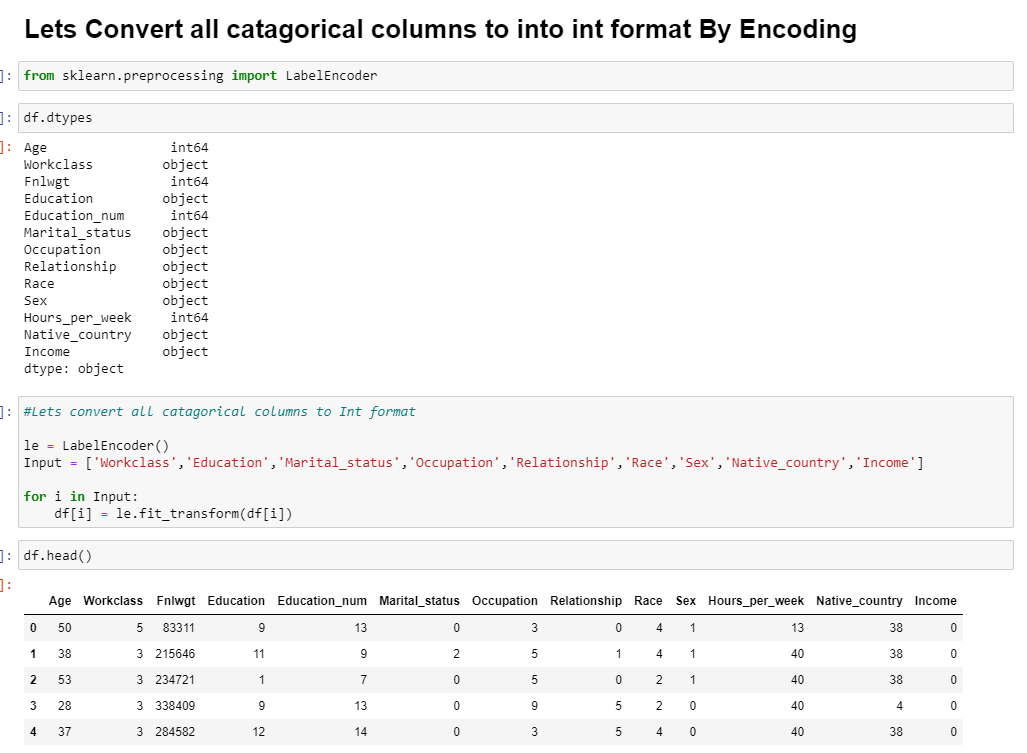
### In Relationship Columns Husband are paid more income as compared to all other relationships



## This pair plot gives the pairwise relation between the columns which is plotted on the basis of target variable "Income". Here we can observe the relation between the features and label. We can see a LOT OF skewness is present in the Numerical COLUMNS.

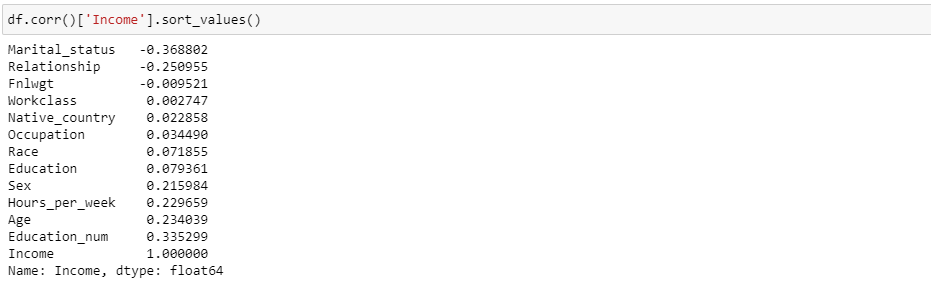


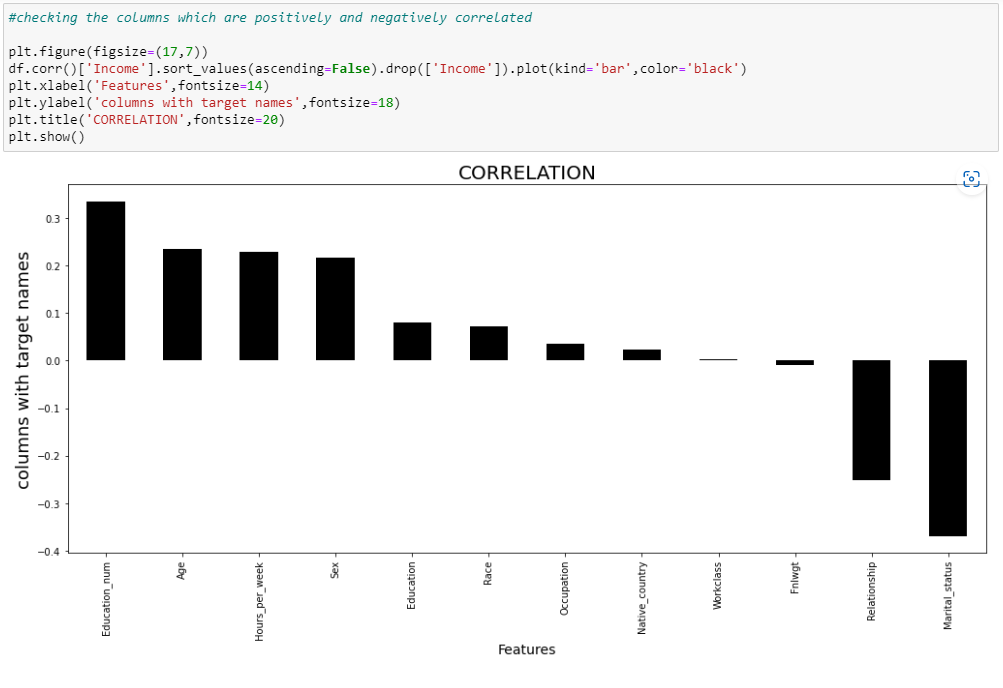
## All the Numerical columns are having Outliers , we will remove it in the later steps.



Here I have converted all the categorical columns into Numerical columns in the dataset by label encoder.

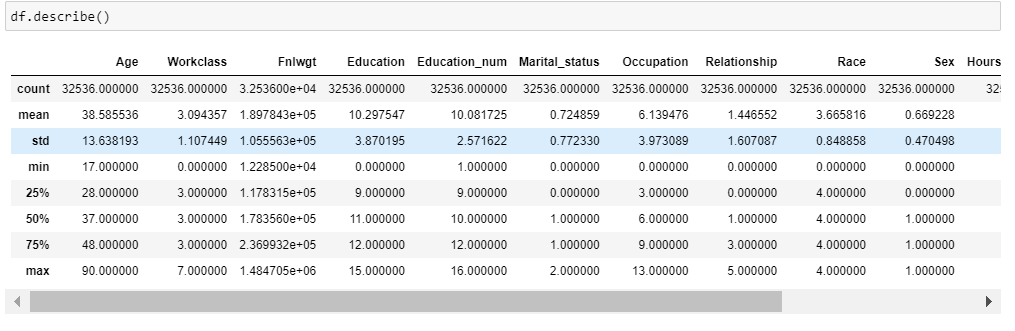
## CORRELATION

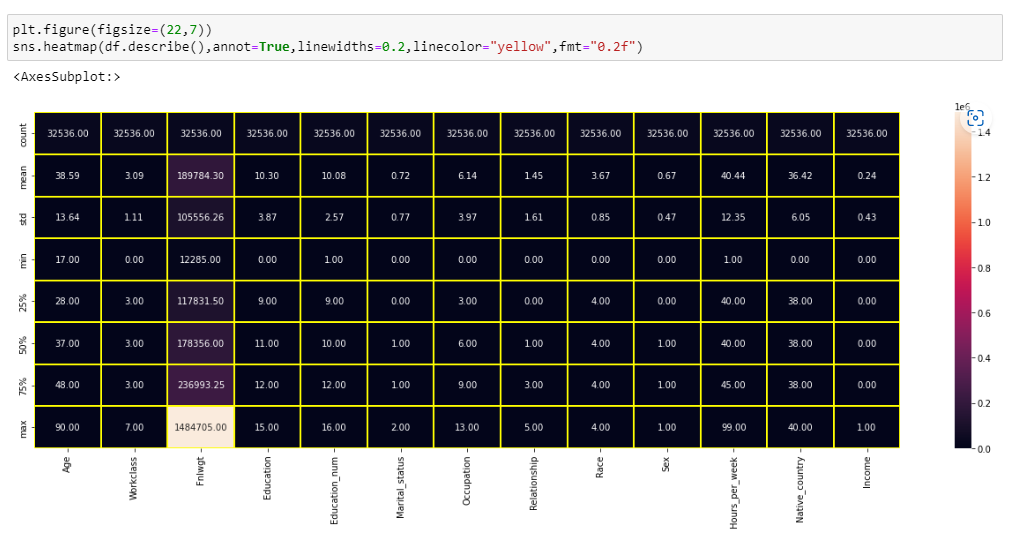




## Here we can see the least correlated columns are-> Workclass and Fnlwgt and the max postive and Negative correlated columns as Education\_num and Marital Status.

## DESCRIPTIVE STATISTICS





## Outcome of describe() of dataset :-

**from the above plotting we are determining mean,standard deviation,minimum and maximum value of each column**

**Total number rows and column in dataset:-**

**Rows:- 32536**

**columns:- 13**

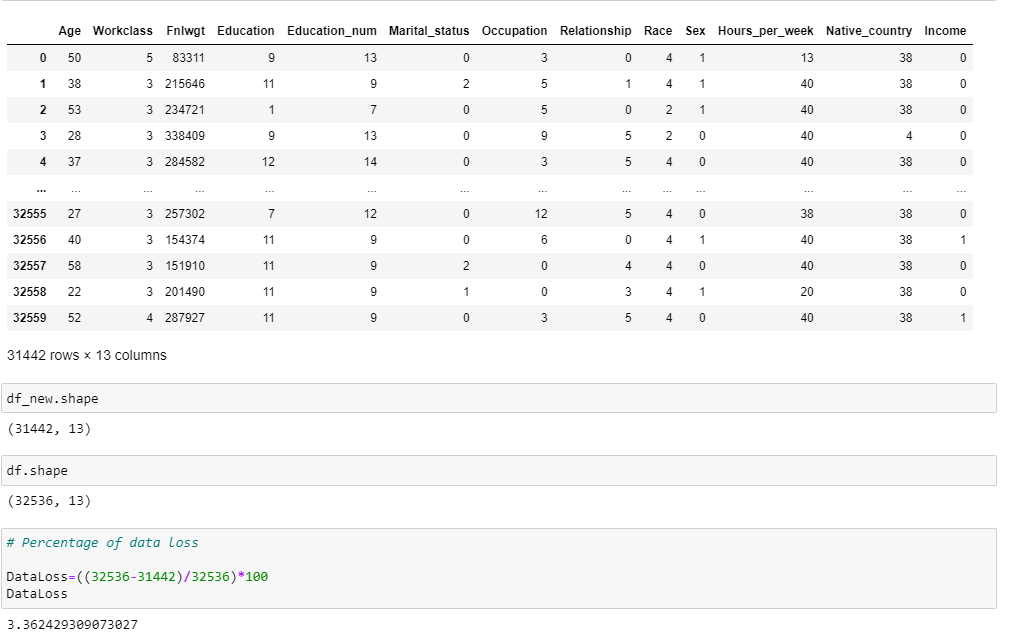
## Now Let’s Remove Outliers Now

# ZSCORE METHOD

****

Here I am Removing Outliers from All the Numerical columns in the dataset.

New Dataframe created after Removing Outliers



## 3.362429309073027% is the data loosed during removal of Outliers by ZScore Method

IQR METHOD

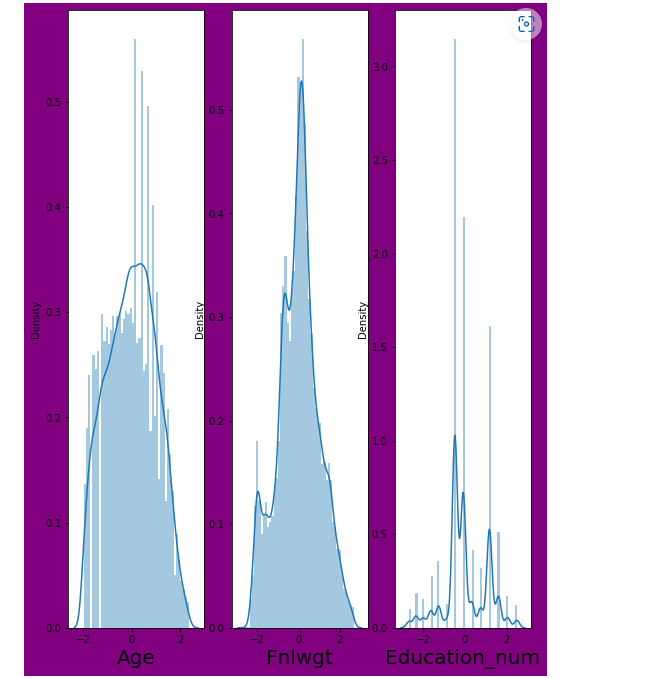


## 72.03098106712564% is the data loosed during removal of Outliers by IQR Method

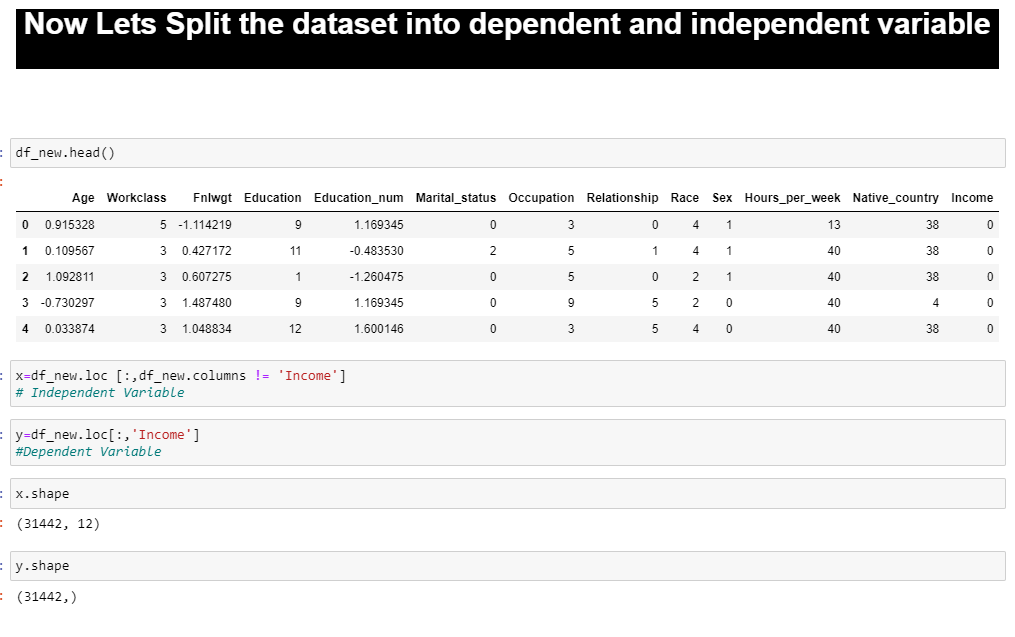
# SO HERE WE WILL GO WITH ZSCORE METHOD BECAUSE IT HAS LESS DATALOSS AS COMPARED TO IQR

## Now LET’S Remove Skewness

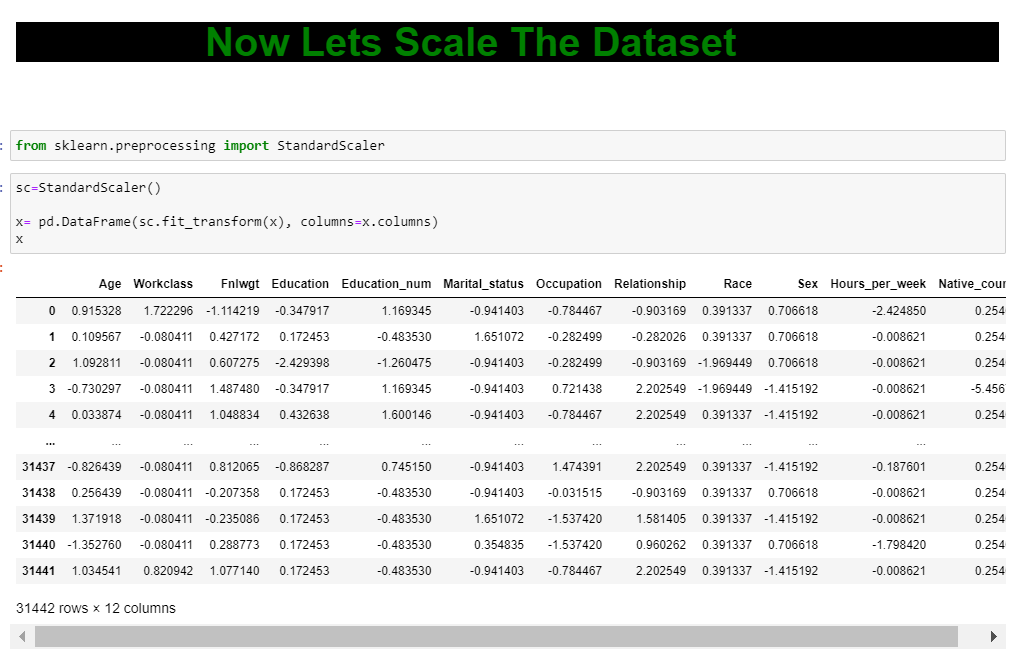




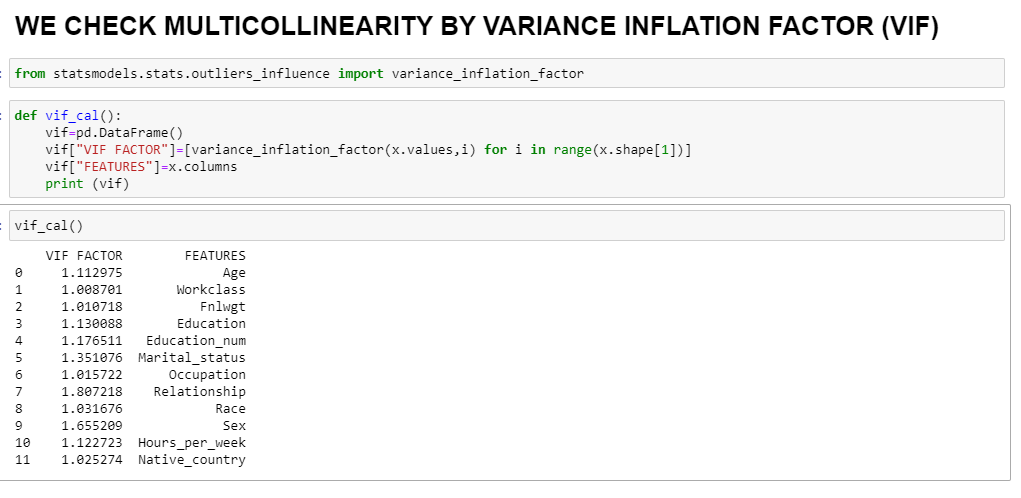
Here I can see that my skewness has been removed from all numerical columns by power transform method.



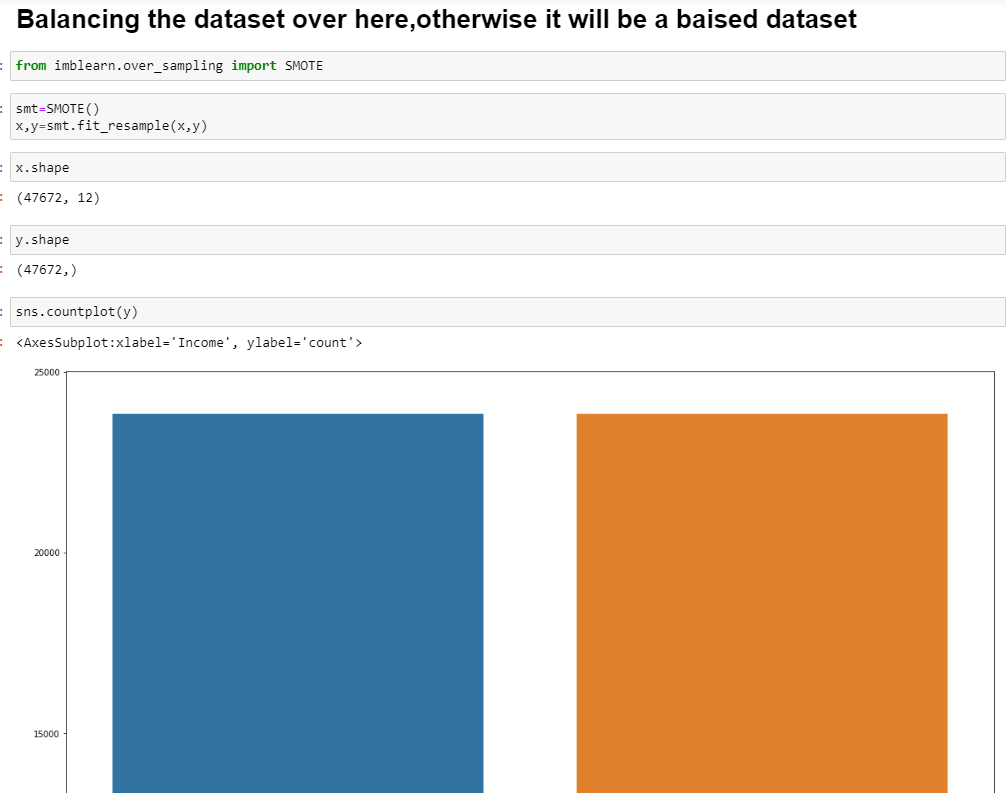
Here I have splitted my x(input) and y (variable) from the dataset.



I have scaled my dataset using Standard scaler.

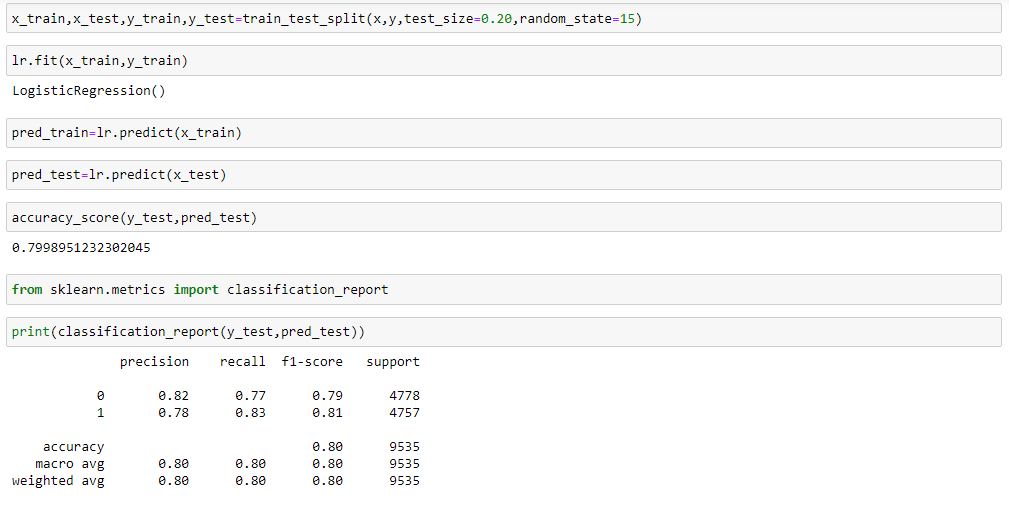


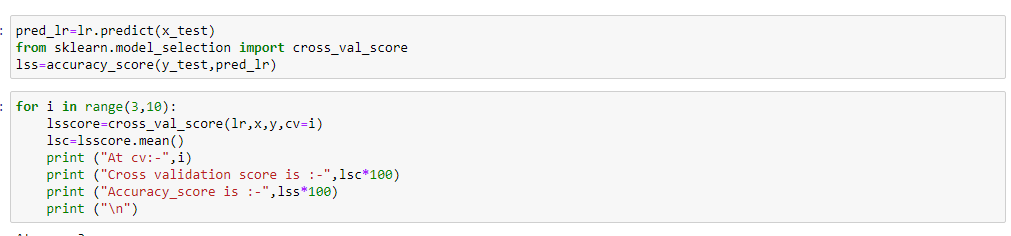
## No Multicollinearity is present in the dataset...Here we can observe all the columns have VIF less than 10 which means we are free from multicollinearity problem. Let's move to building machine learning models....

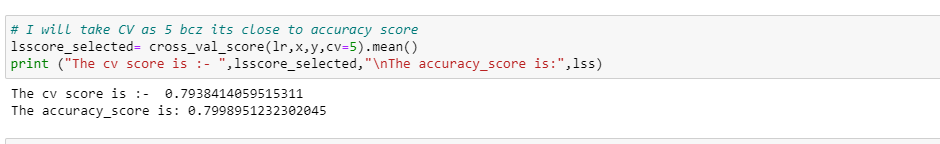
****

**Here I have used SMOTE Technique to balance my imbalanced dataset.**

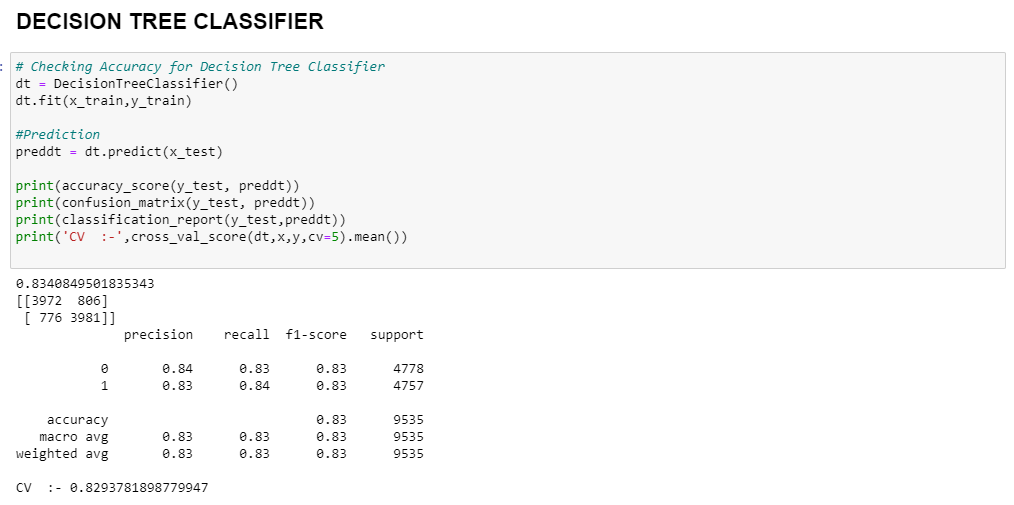
****

****

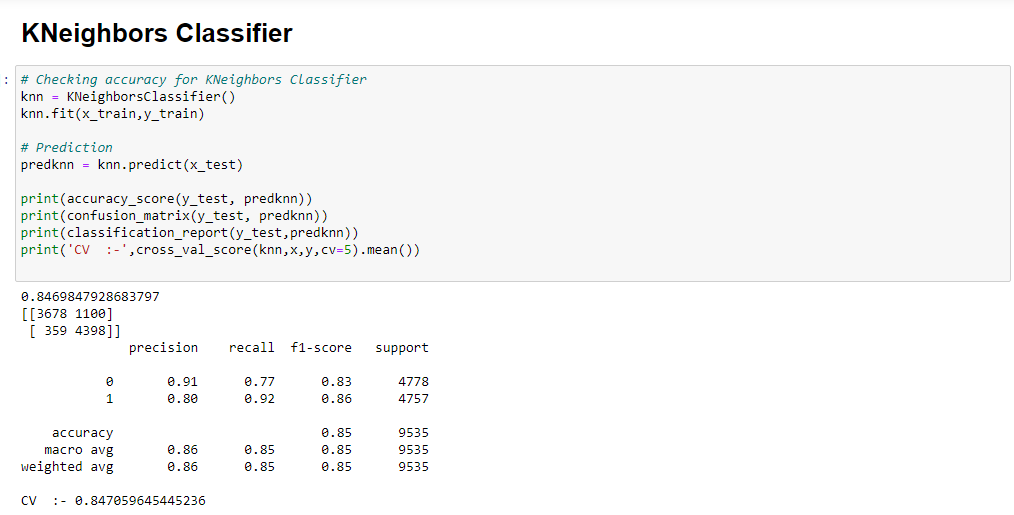
****

****

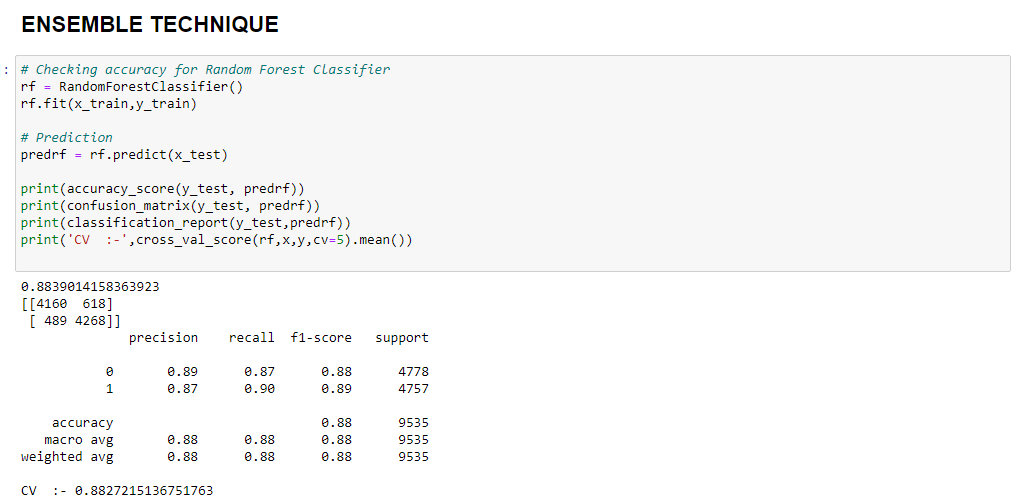
My Linear Regression model is Performing almost 80% in case of both accuracy score and cross validation score.



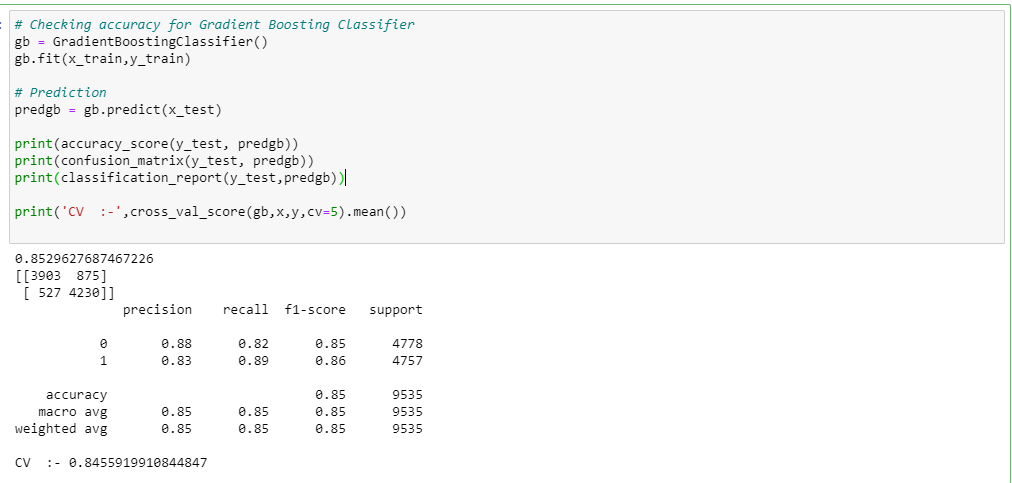
My Decision tree classifier is giving accuracy as 83% and CV as 83% .



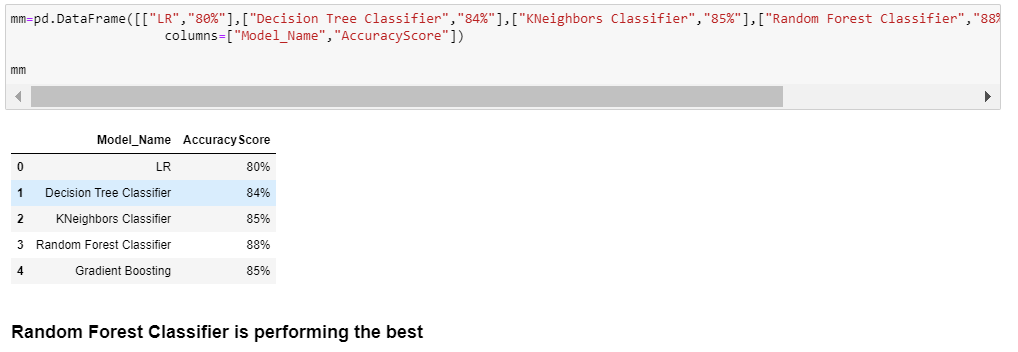
My kNeighbor classifier is also working very good and Here my accuracy and Cv is both giving 85%.



Here my random forest classifier is working very well and it’s giving accuracy and Cv as 85% both ..



My Gradient boosting is giving me 85% as Accuracy score and 85% as Cv score but still Random forest is performing well .

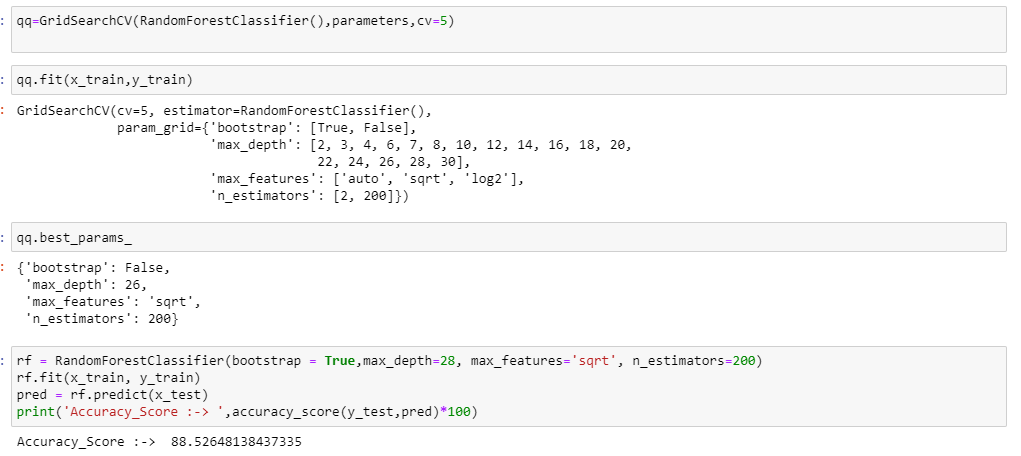


Making a dataframe to see which model is performing well. here we can see that random forest is the best performer so far .

I have taken random forest in hyper parameter tuning



I have given parameters for rfc.



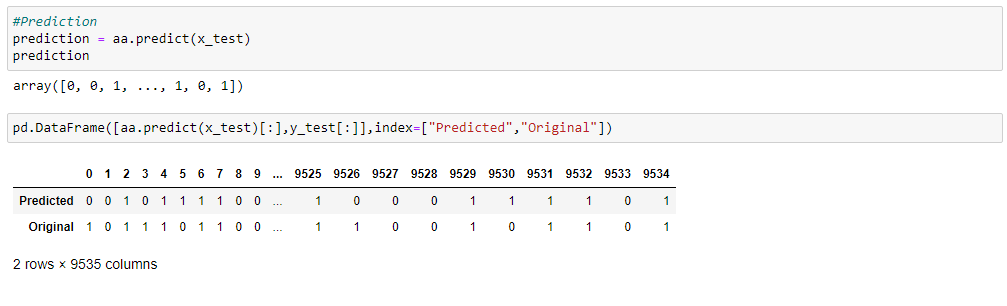
Here I have fitted the model and have got the best parameter to apply on last model .

At last we can can see that my accuracy have been Improved by some percent’s.

Now Let’s save the final Model . .



Here I have saved my final model.



I have made a prediction on my final dataset and it seems to be performing well.

# Concluding Remarks :->

We started our project with importing various libraries and dataset. Observing the many important points like problem type and how many columns contains int, float and object values. As per statistic observations we found huge variations among the features and we have used standard scaler to scale the variables. Besides this, we have identified there are 2 columns where instead the NaN it is in “?” so we deleted such rows. During this process we used seaborn and matplotlib to do the visualizations and converted categorical features into numeric using label encoder Afterwards we started training different different machine learning models, picked one of them which is Random Forest classifier and applied cross validation on it and we tried to tune model using hyperparameter tuning.

To conclude, there are many other ways also to improve the model accuracy like doing a more extensive feature engineering, by comparing and plotting the features against each other and identifying and removing the noisy features, along with resampling the data in case of imbalance or more extensive hyperparameter tuning on several machine learning models .. . ..