**Census Income prediction**

The economic well-being of a Nation is highly driven by the income of the residents. Income of person is a part of GDP (Gross domestic product) of country. Census income is a constituent of national income of the country Other constituent that are also involved are,

There are so many decisions which are dependent on the census income like,

1) Investment of other countries

2) If the country has people whose local labor are cheap the manufacturing plants for big companies can be set at that country

3) We can also see the country employment classes

4) Census data is the backbone of the democratic system of government, highly affecting the economic sectors. Census-related figures are used to distribute the federal funding by the government into different states and localities

5) Not only the above, the census data is also used for post census population estimates and projections, economic and social science research, and many other such applications. Hence, the importance of this data and its correct predictions is very clear to us

6) Data has always been the backbone of many important decisions. When an assumption is backed up by facts and numbers, the chances of incorrectness and bad decisions decrease.

### **Constituents of GDP**

* Wages and salaries
* Rent
* Interest
* Undistributed profits
* Mixed-income
* Direct taxes
* Dividend
* Depreciation

**Problem statement** - The prediction task is to determine whether a person makes over $50K a year.

The above introduction had an aim to increase the awareness about how the income factor has an impact not only on the personal lives of people but also an impact on the nation and its betterment. We will today have a look at the data extracted from the 1994 Census bureau database, and try to find insights about how different features have an impact on the income of an individual. Though the data is quite old, and the insights drawn cannot be directly used for derivation in the modern world, it would surely help us to analyze what role different features play in predicting the income of an individual.

**Data** - This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). A set of reasonably clean records was extracted using the following conditions: ((AGE>16) & (AGI>100) & (FNLWGT>1).

AGI - Adjusted Gross Income

FNLWGT – Final Weight

**Description of 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 (pertaining to Spain) 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.

Features in dataset :-

1. Age — The age of an individual, this ranges from 17 to 90.

2. Workclass — The class of work to which an individual belongs.

3. Fnlwgt — The weight assigned to the combination of features (an estimate of how many people belong to this set of combination)

4. Education — Highest level of education

5. Education\_num — Number of years for which education was taken

6. Marital\_Status — Represents the category assigned on the basis of marriage status of a person

7. Occupation — Profession of a person

8. Relationship — Relation of the person in his family

9. Race — Origin background of a person

10. Sex — Gender of a person

11. Capital\_gain — Capital gained by a person

12. Capital\_loss — Loss of capital for a person

13. Hours\_per\_week — Number of hours for which an individual works per week

14. Native\_Country — Country to which a person belongs

**Output:**

Income — The target/label variable, which predicts if the income is higher or lower than 50K$.

**Contents of the article:**

The following information/steps will be covered further in the article –

* Data cleaning and analysis of data
* Encoding the data — Label Encoder
* Exploratory data analysis
* Data modeling
* Outlier detection and skewness treatment
* Scaling the data — Standard scaler
* VIF - Checking multicollinearity
* Fitting the machine learning models
* Cross-validation of the selected model
* Model hyper tuning
* AUC-ROC curve
* Saving the final model and prediction using saved model

**Data analysis:**

**1) Checking Numeric data type or object data type**

Age int64

Workclass object

Fnlwgt int64

Education object

Education\_num int64

Marital\_status object

Occupation object

Relationship object

Race object

Sex object

Capital\_gain int64

Capital\_loss int64

Hours\_per\_week int64

Native\_country object

Income object

**2) Null value or missing value checking**

Age 0

Workclass 0

Fnlwgt 0

Education 0

Education\_num 0

Marital\_status 0

Occupation 0

Relationship 0

Race 0

Sex 0

Capital\_gain 0

Capital\_loss 0

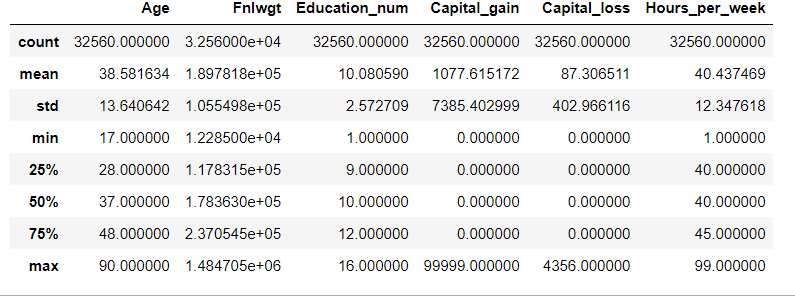
Hours\_per\_week 0

Native\_country 0

Income 0

No null value in data set and we will again cross check values, are there any other data which is also missing in further steps

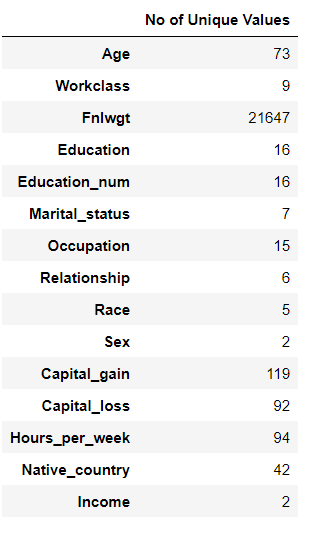
**3) Statistical analysis of int and float columns**



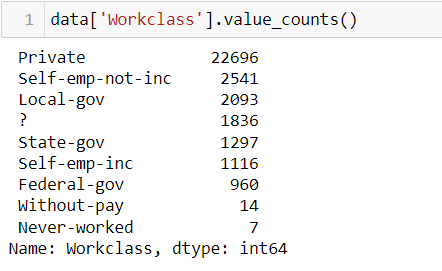
**Observation:**

1. capital\_gain have some irregular values
2. capital\_loss have some irregular values
3. Hours\_per\_week have same 25% value and 50% value so some error in data in further steps we will correct it

**4) Unique value in each column**



**5) Insights from feature column (wrong data)**

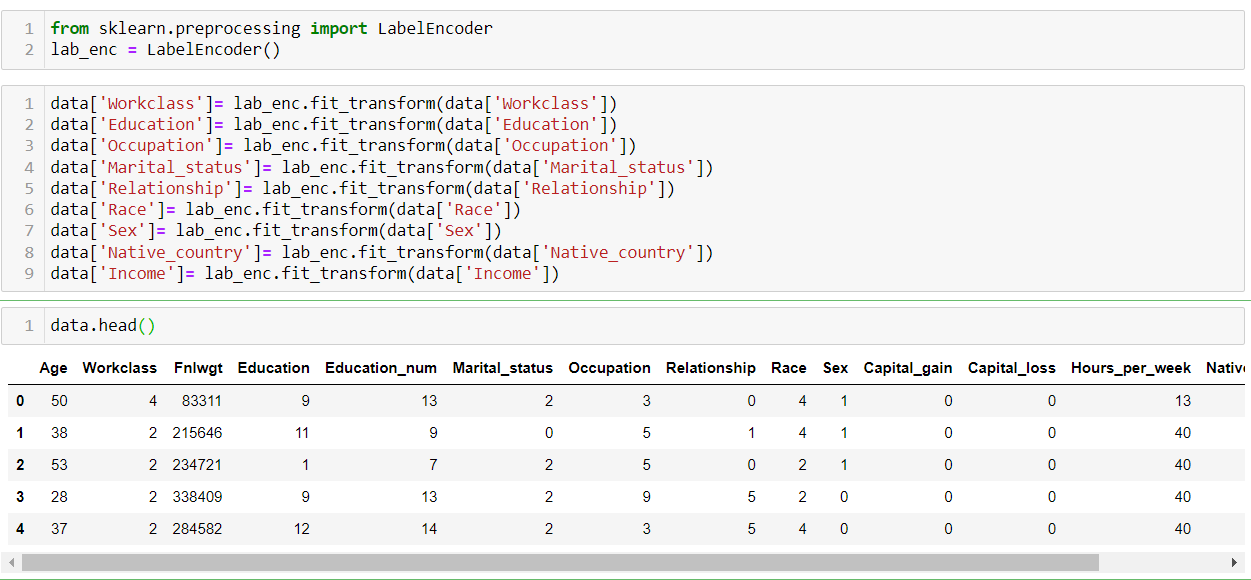


here we can see there is some missing values which is denoted by "?" and value\_count is 1836 which should be replaced by some meaningful data or can be dropped (workclass and native country column having missing data)

we have sufficient **data** better to drop the “?” row.

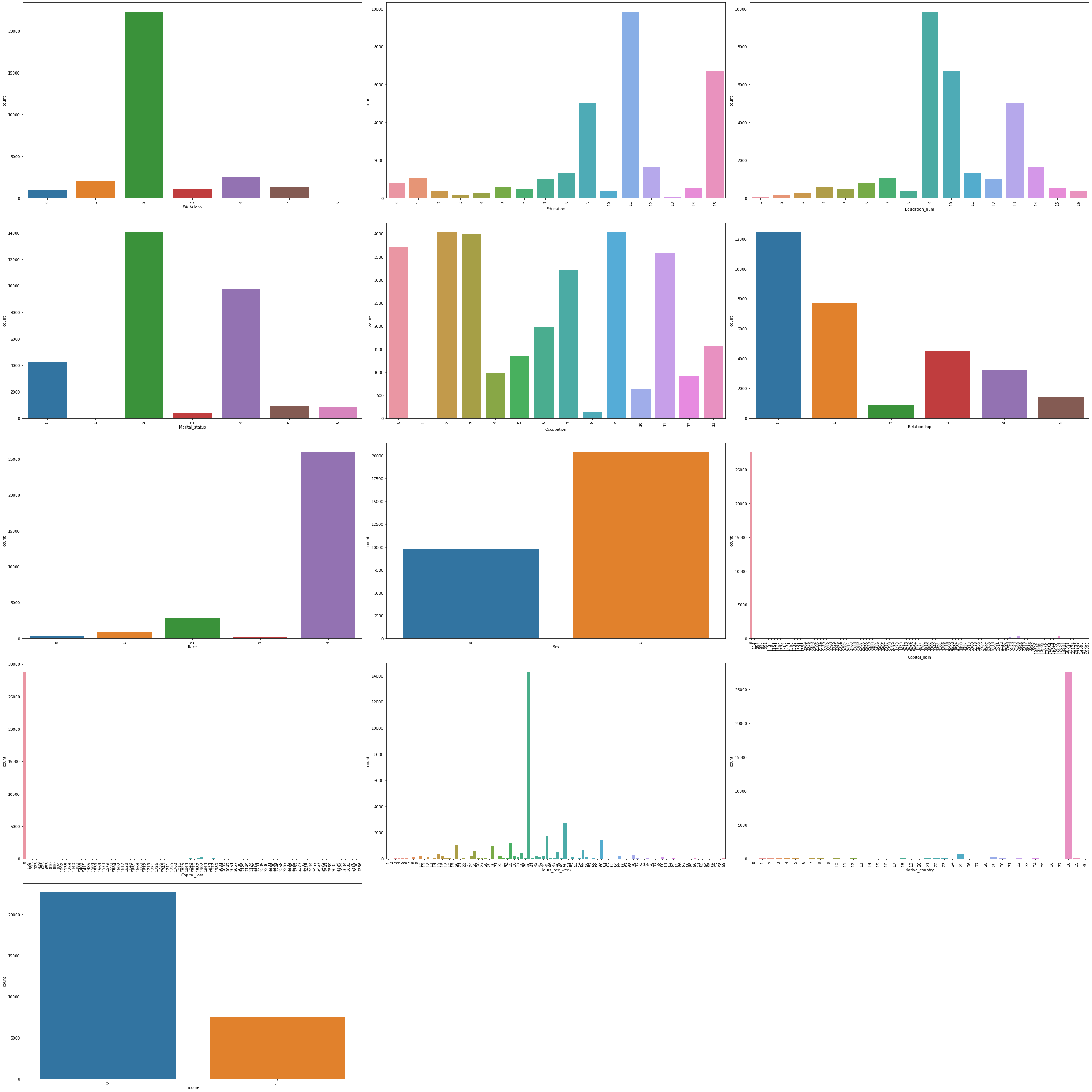
**Data encoding (Label encoder):**

Encode target labels with value between 0 and n\_classes-1.



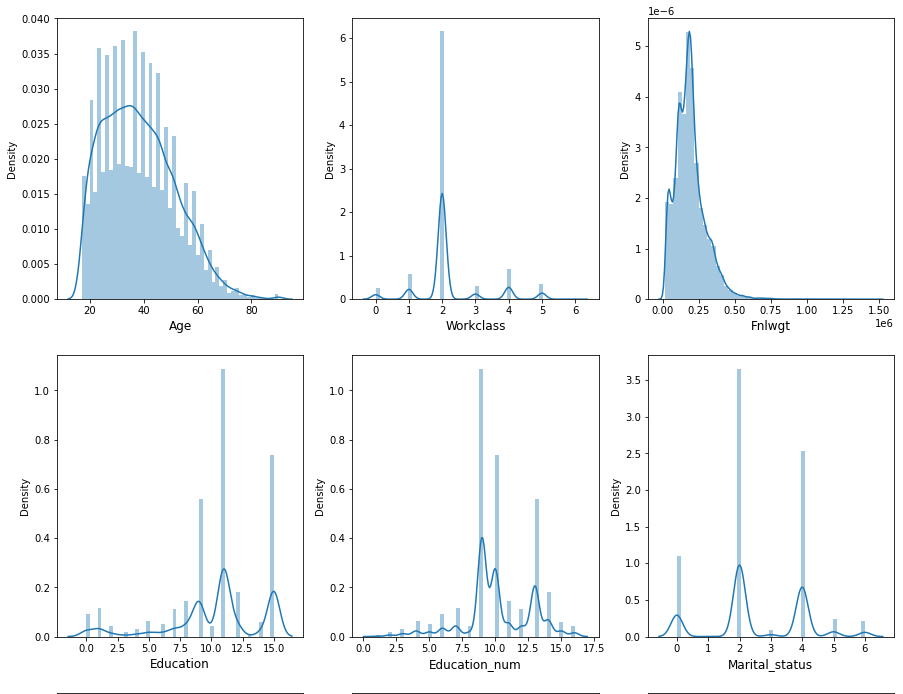
**EDA (Exploratory data analysis):**

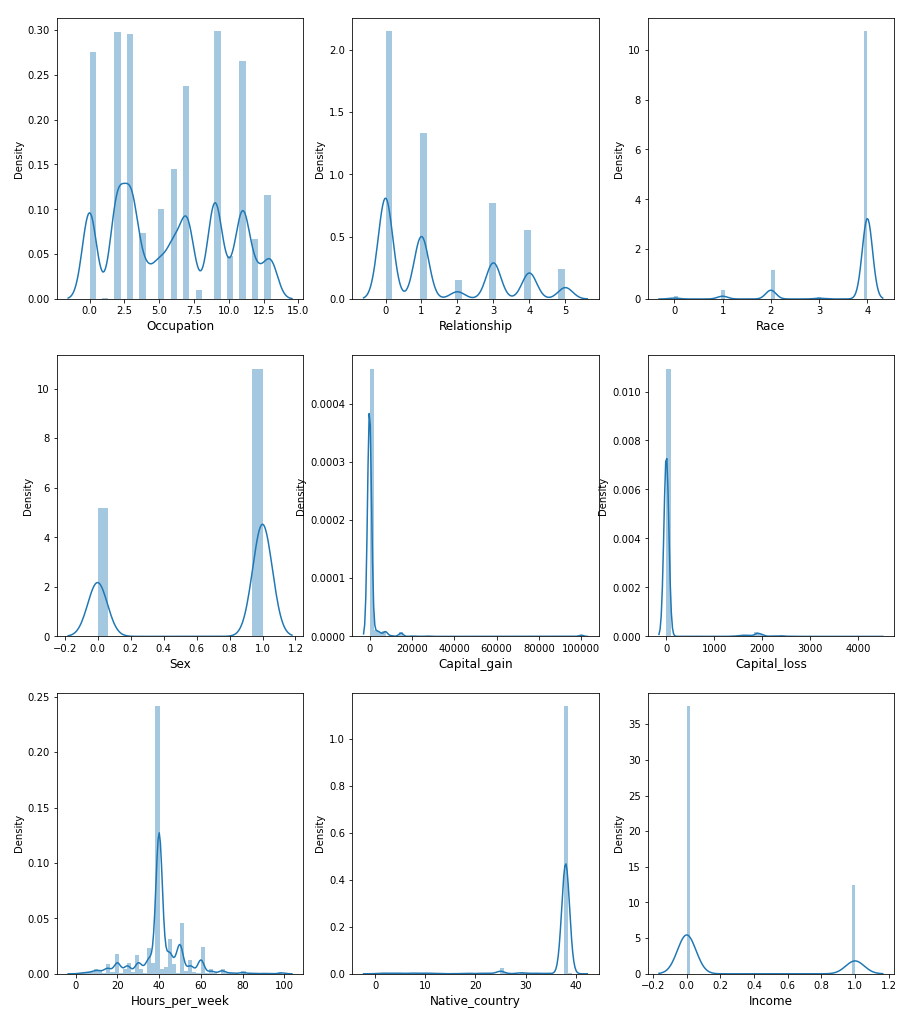
* Univariate analysis



* While checking the unique values for workclass, we see that we have 6 different types of values,
* The ‘Education’ column has 16 different categories available. Majority of these categories belong to ‘School’ type (different classes are divided into multiple categories)
* The education number is the number of years for which a person received education. This is an ordinal column, which contains 16 different values. When we check the division of ‘Education\_num’ column, we observe that the count of ‘Education’ column and ‘Education\_num’ is exactly same! Which means, the ‘Education\_num’ column is providing same information as ‘Education’ column, but in a numeric manner!
* The ‘Marital\_Status’ column has 7 different categories available, and has no missing values.
* Majority of the people have ‘Marital\_Status’ as ‘Married-civ-spouse’, and least have ‘Married-AF-spouse’. Count of ‘Never-married’ is also quite high.
* The occupation column contains 14 different categories, and have missing values represented by ‘?’ (which we have already observed, and combined with ‘Workclass’ column).
* The relationship column contains 6 different types of values, with highest number set for ‘Husband’ and lowest for ‘Other-relative’. The column does not have any missing value.
* The Race column has 5 different categories, and no missing data. Highest number of people have race as ‘White’ (significantly high numbers).
* The ‘Sex’ column has 2 categories — Male and Female, where number of males are almost double to number of females. Missing values are not found in this column.

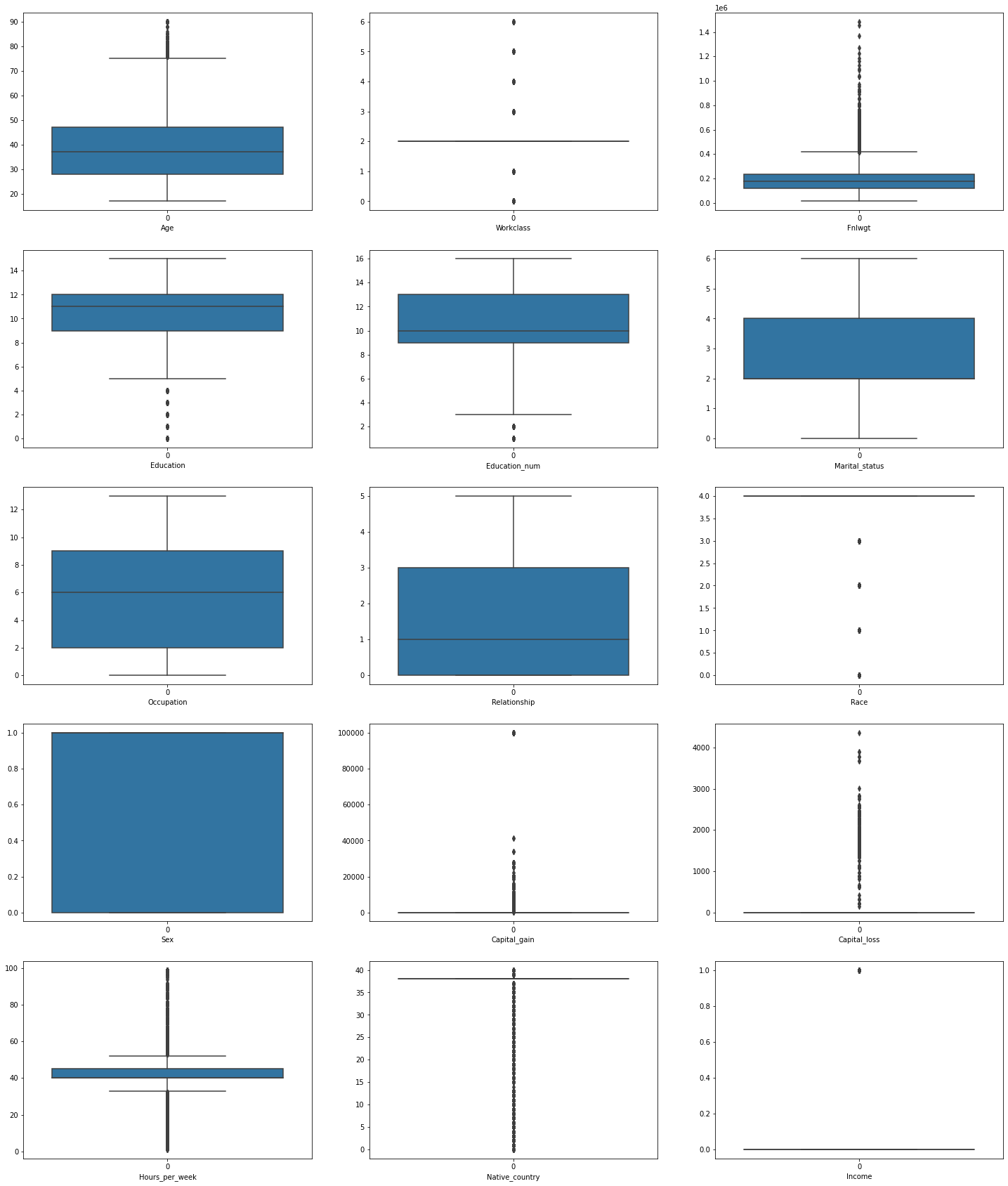
**Distribution and bar plot:**



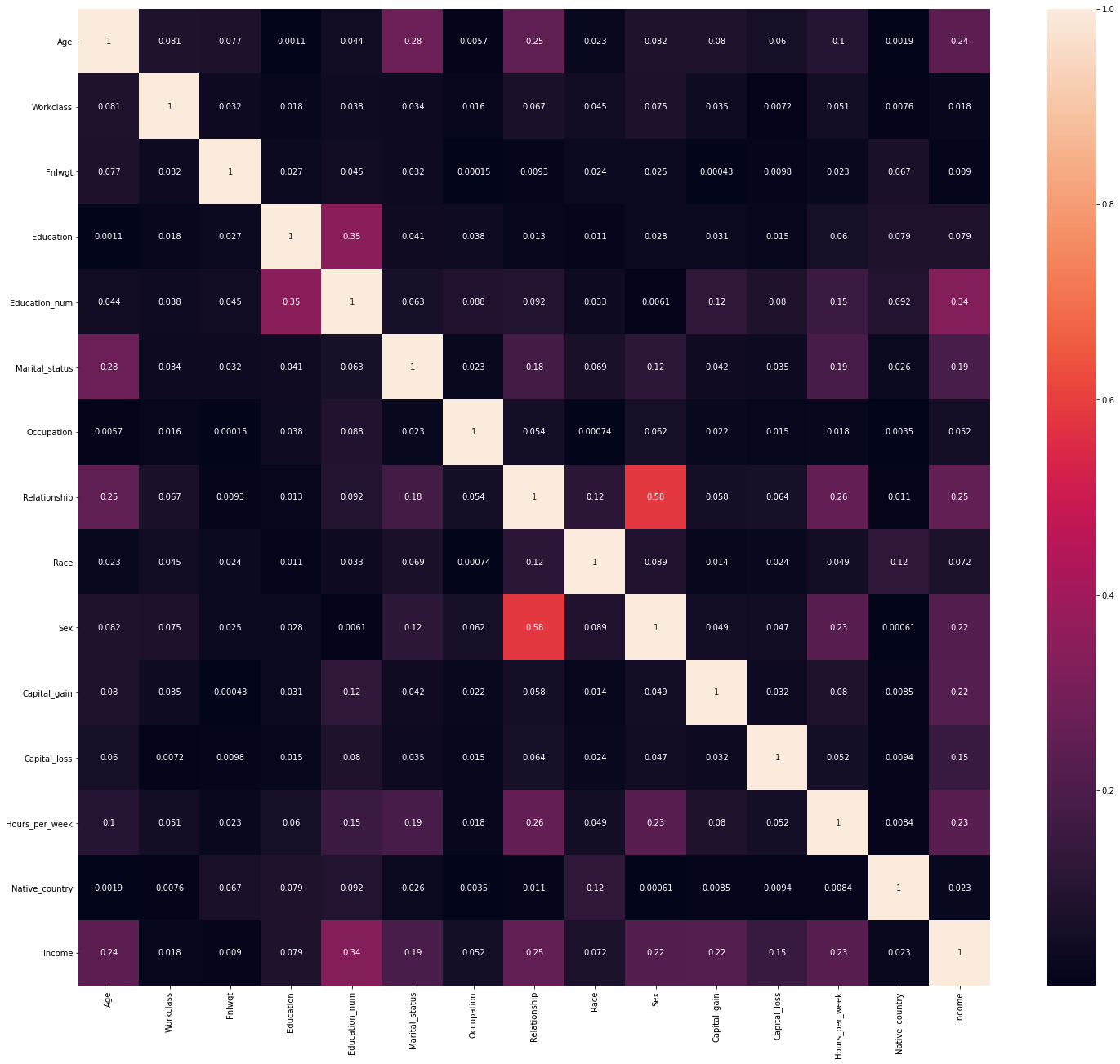
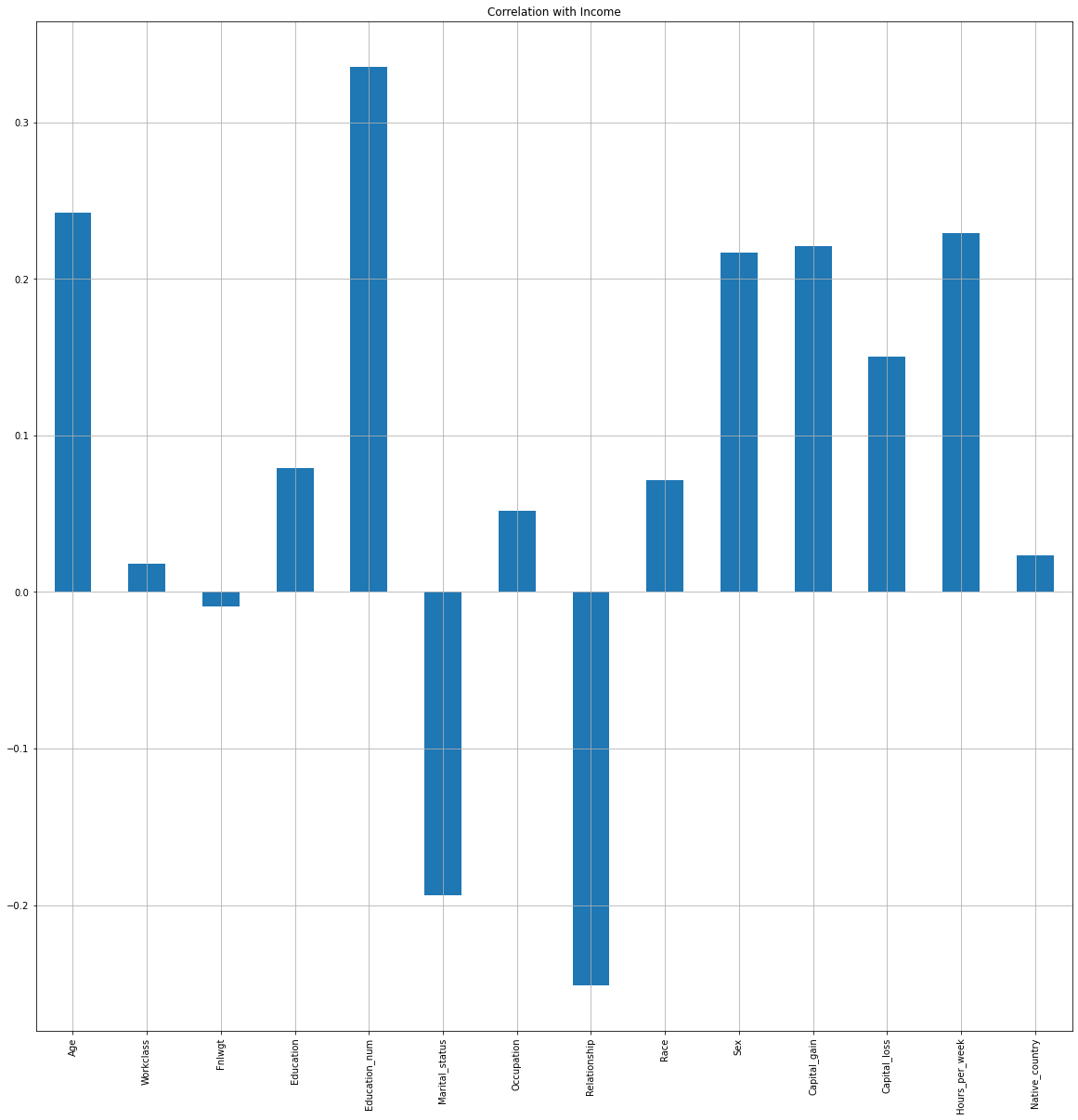


* Age and Fnlwgt are the features which are continuous in nature and are right skewed so skewness should be removed
* The hours per week column have values scattered over a range of 1–99. The column does not have any missing values. Majority of the values have data near 40 hours and hence a high peak can be observed in the below distribution plot

**OUTLIERS:**

Fnlwgt and Age are columns which are having outliers all other are categorical columns.

Data correlation – is also checked with Corr with method and plotting bar plot and heatmap for it,



Some columns are having negative and some are having positive relationship

We are also checking multicollinearity with heatmap and will verify in further steps by vif value,

**Data Normalization, Scaling and multicollinearity:**

* **Outliers removed - IQR**
* **Skewness Treatment- power transformer**
* **Vif value for multicollinearity**

**Outliers removing by Inter quantile range**

Outliers detection formula

higher side - Q3+(1.5\*IQR)

lower side - Q3 -(1.5\*IQR)

IQR - Interquartile range defines the difference between the third and the first quartile. Quartiles are the partitioned values that divide the whole series into 4 equal parts. So, there are 3 quartiles. First Quartile is denoted by Q1 known as the lower quartile, the second Quartile is denoted by Q2 and the third Quartile is denoted by Q3 known as the upper quartile. Therefore, the interquartile range is equal to the upper quartile minus lower quartile.

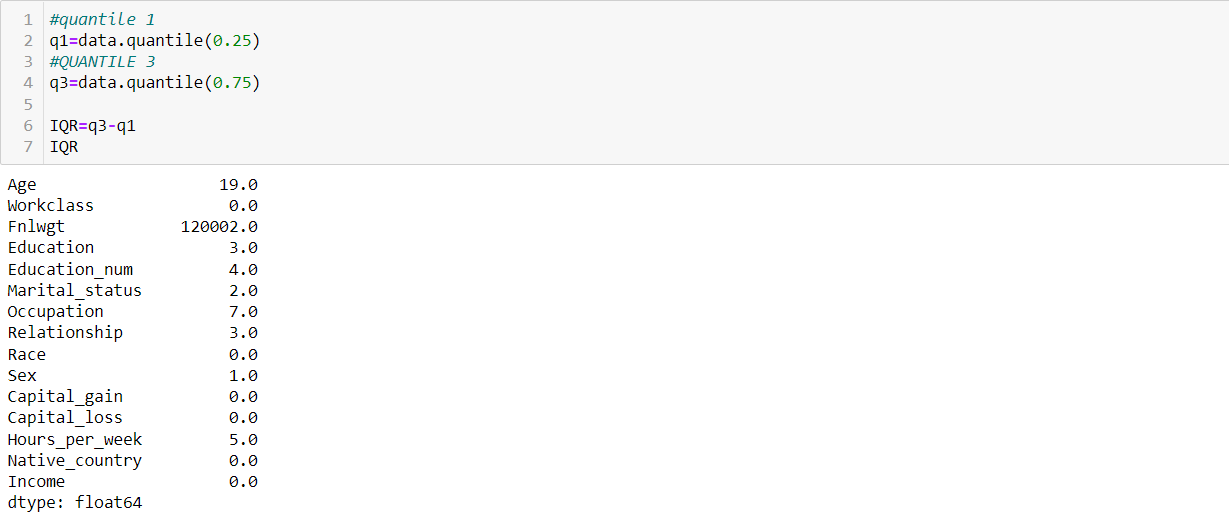
Interquartile Range Formula

The difference between the upper and lower quartile is known as the interquartile range. The formula for the interquartile range is given below

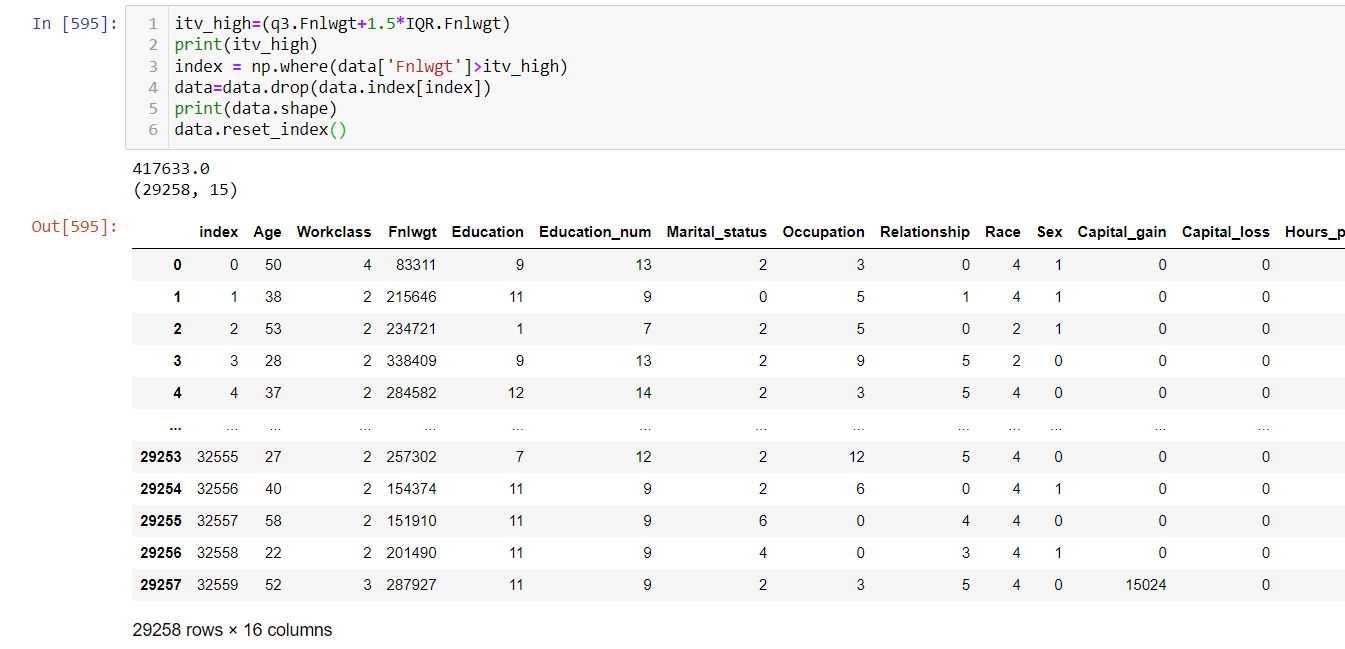
**Interquartile range = Upper Quartile – Lower Quartile = Q­3 – Q­1**

where Q1 is the first quartile and Q3 is the third quartile of the series.

The below figure shows the occurrence of median and interquartile range for the data set.



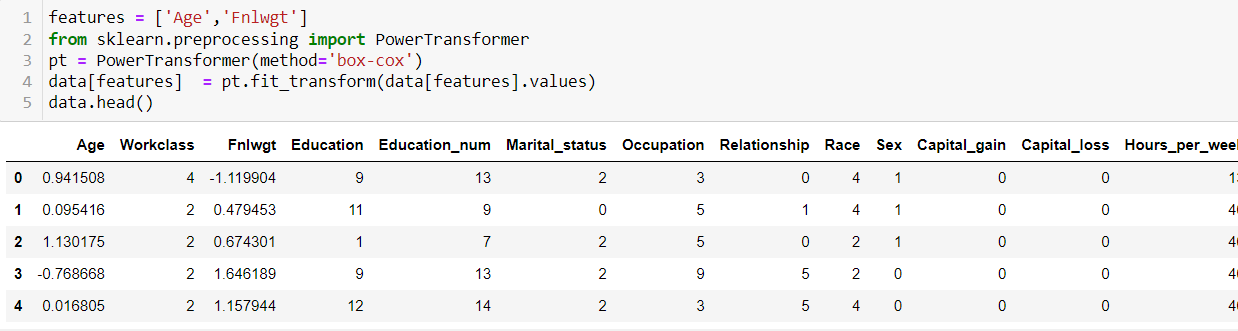
IQR is applied on Fnlwgt and Age



Same is applied on Age Feature as it also having outliers

**Skewness removal - power transformer(box-cox)**

We now proceed with treating skewness in our data, which allows us to fit our data in a symmetric distribution, which further allows our model to learn better.



We can see Fnlwgt feature that power transformer does change the value of Fnlwgt so that curve of distribution should be **Gaussian-like probability distribution.**

# **Scaling the data and checking multicollinearity**

The next step is to bring the data to a common scale, since there are certain columns with very small values and some columns with high values. This process is important as values on a similar scale allow the model to learn better.

We use standard scaler for this process –

***‘****Standard Scaler follows Standard Normal Distribution (SND). Therefore, it* ***makes mean = 0*** *and scales the data to* ***unit variance****’*

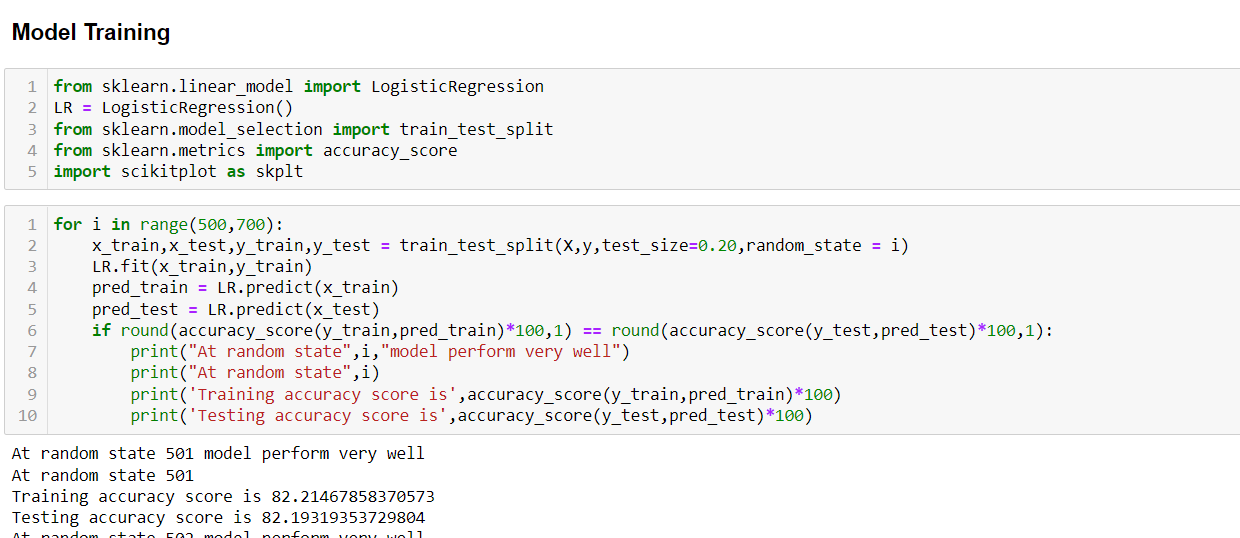


* Vif<5 so no multicollinearity problem,

**Fitting the machine learning model:**

We now proceed to the main step of our machine learning, fitting the model and predicting the outputs. We fit the data into multiple classification models to compare the performance of all models and select the best model –

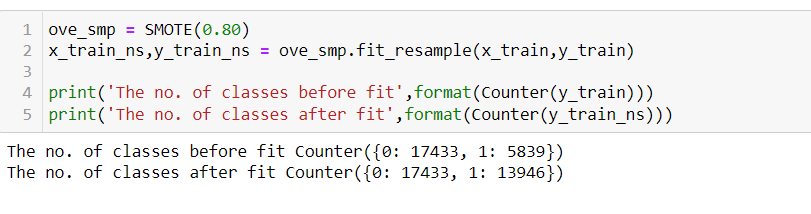
First, we have to select best random state value so that there should not be overfitting and underfitting of data



* Random state value is 543

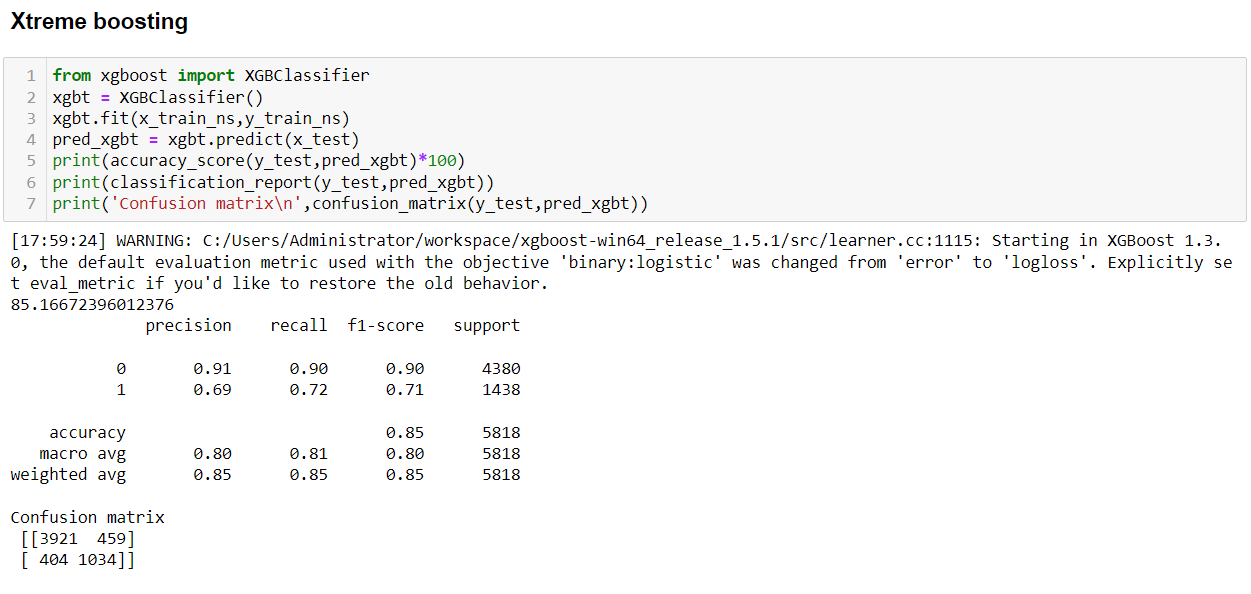
Checking Imbalancing of Target column as we can see Training outcome - 0:17433 and 1:5839if we train this model there is chance of bad recall and precision so,

We balance the dataset by SMOTE,



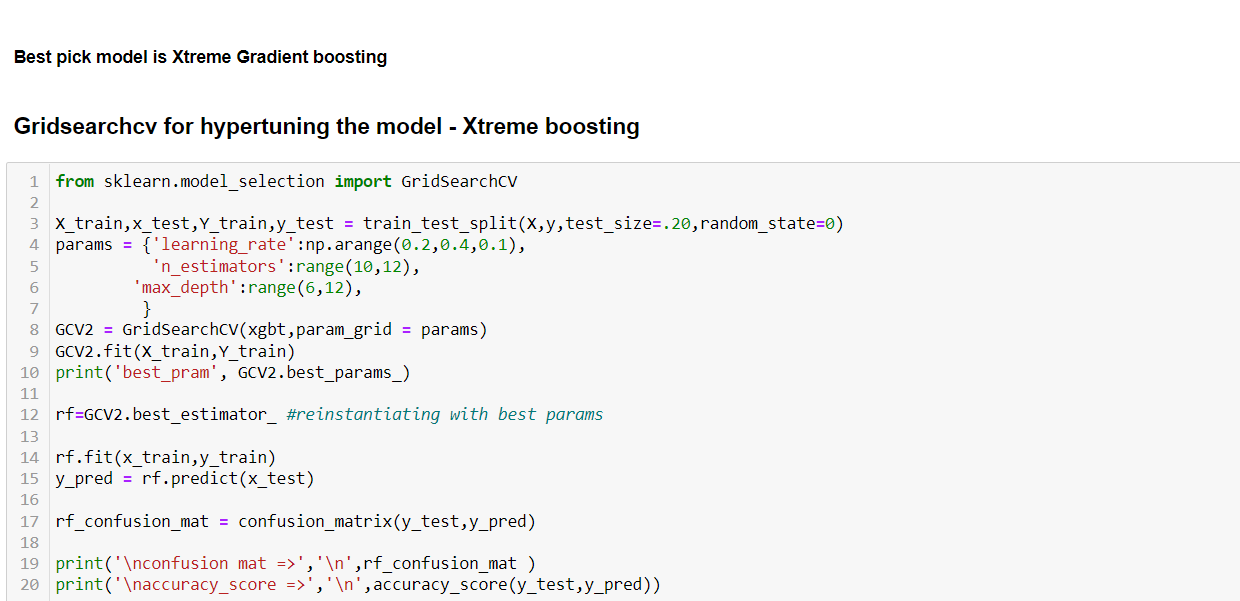
Now we are all good to train our model,

We are achieving the best result by Xtreme Gradient boosting – 85.166%



We can see,

* True positive value as – 3921
* True negative value as – 1034
* False Positive value as – 459
* False negative value as – 404

confusion mat => [[4150 207] [ 445 1016]]accuracy score => 0.8879339979374355

We have increased our accuracy by Hyper tuning the model

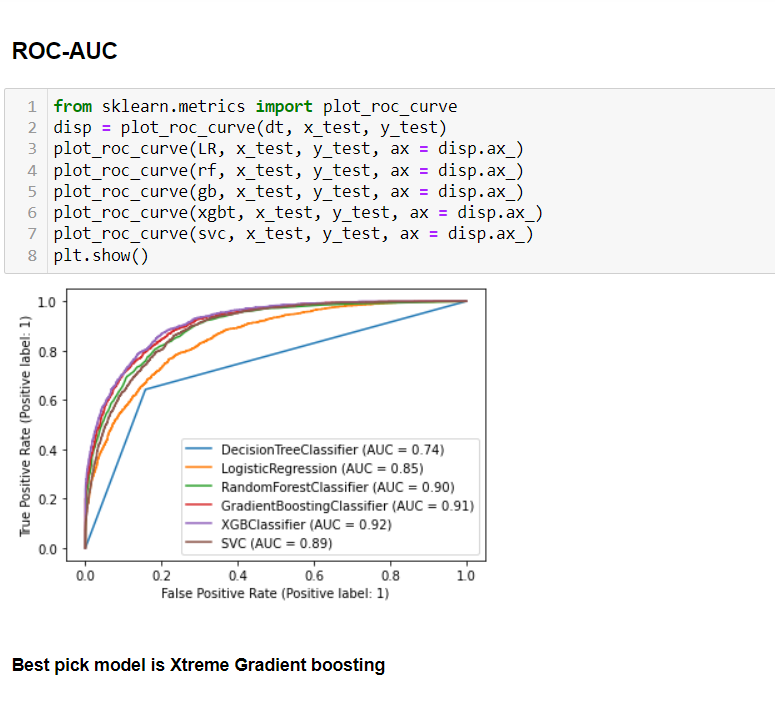
**Cross validation-**

The **goal of cross**-**validation** is to test the model’s ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset, for instance from a real problem).

Cross validation score of xtreme gradient boosting model is 86.50739085596426

**Roc-Auc -**

AUC — ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0’s as 0’s and 1’s as 1’s.



**Conclusion -**

Now we save the model,



1) we have model accuracy of 88% after hyper tuning the model by gridsearchcv

2) we have applied SMOTE for imbalancing of dataset

3) we have also plotted roc-auc to see area under score

4) xtreme boosting is giving best results

This marks the end of our process; we have successfully trained our model to predict the income of a person, with an accuracy of ~88%.

We moved step by step, analyzing, cleaning and modeling the data, and applied various machine learning models to achieve the desired predictions. We also tuned the model to improve the accuracy, and were able to achieve a model with quite a good accuracy.