**Machine Learning Project**

**Census Income Prediction**

Submitted by- Manish Kumar Mohanty

Batch- 1836

1. **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*.***

We will today have a look on 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, but it would surely help us to analyse what role different features play in predicting the income of an individual.

**Data Source:**

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

2.**Data analysis**

The dataset provided to us contains 32560 rows, and 14 different independent features. We aim to predict if a person earns more than 50k$ per year or not. Since the data predicts 2 values (>50K or <=50K), this clearly is a classification problem, and we will train the classification models to predict the desired outputs.

Mentioned below are the details of the features provided to us, which we will be feeding to our classification model to train it.

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:

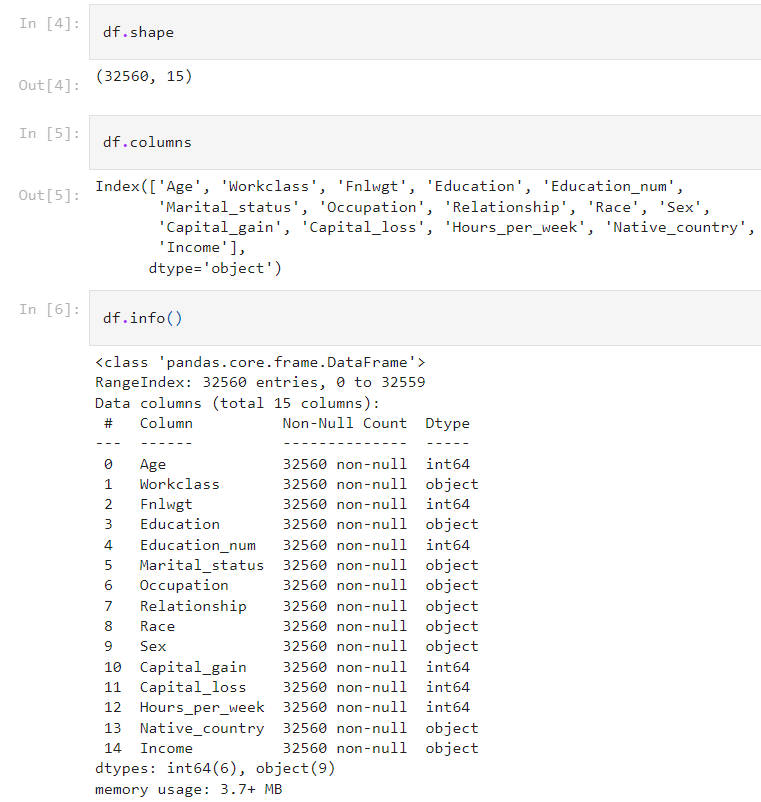
Income — The target variable, which predicts if the income is higher or lower than 50K$.

# Importing the required libraries

# 

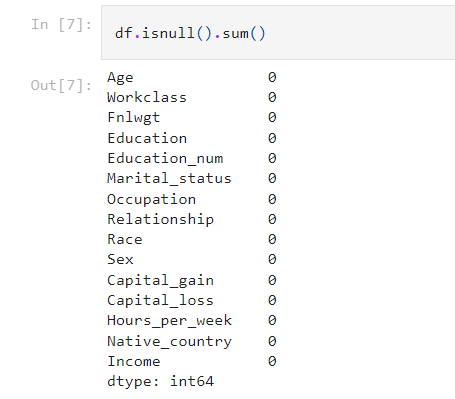
**2.EDA Concluding Remarks**

The first step that we do is to check the information about our data. We see the results shown in the image below:

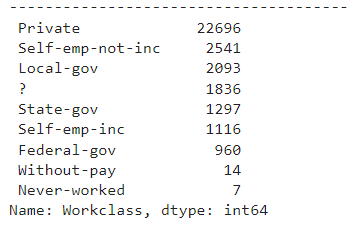


We see that we have a mixture of categorical and numeric columns. We have 6 integer columns and 9 object type columns. We observe that the count of entries is 32560 for all columns, hence no NaN values are present in our dataset.

We confirm this assumption using data.isnull().sum() command –

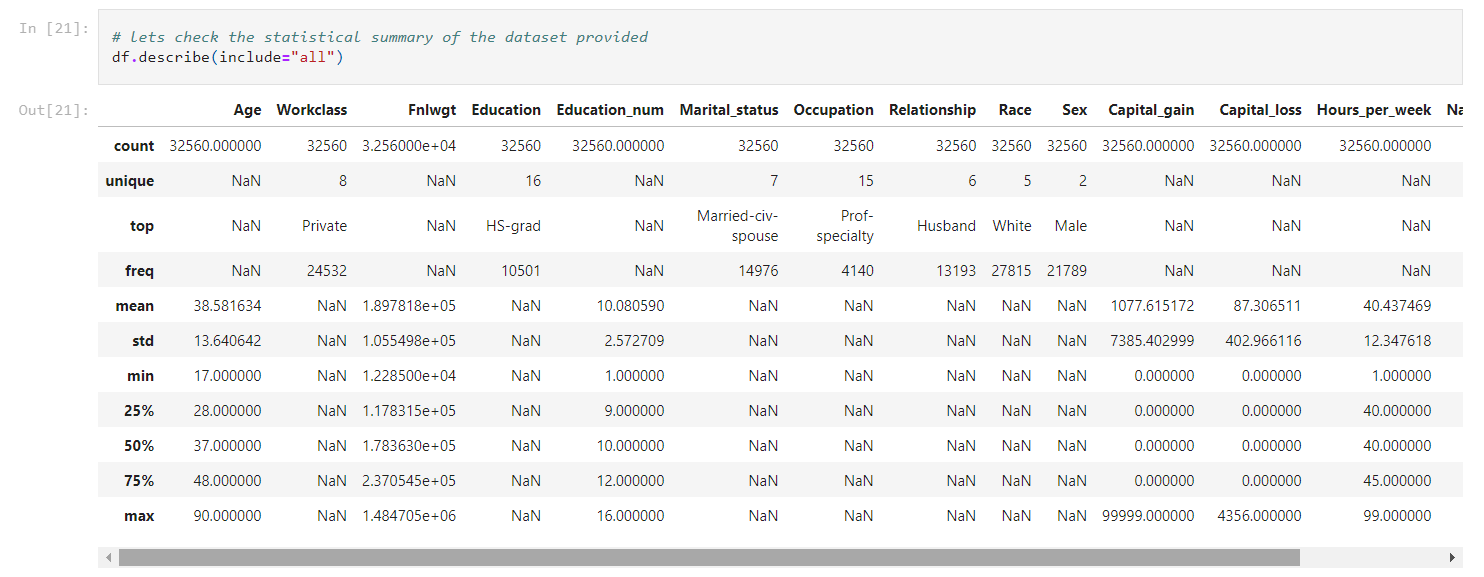


But, while having a close look at our dataset, we observe some of the values as ‘?’, which represent missing values.



Hence, we deduce that there are some values in our data set which need to be treated.

We also check the numerical statistics of our data using df.describe() command –



Following observations are made in this step –

- The age column has a range of 17 to 90.

- The fnlwgt column has a minimum value of 12285 and maximum value of 1484705

- The education number has a range of 1 to 16

- The capital gain starts from 0 and ends at 99999

- The capital loss starts at 0 and ends at 4356

- Hours per week range between 1–99.

- There are outliers expected in Capital gain column as the values till 75% are 0. Same is the case with capital loss as well.

- The fnlwgt column also has a huge difference between 75% values and the max value. There is a chance of getting outliers here.

**Further, we have a look at our dataset and explore the data in various columns, one by one –**

**Univariate Analysis:**

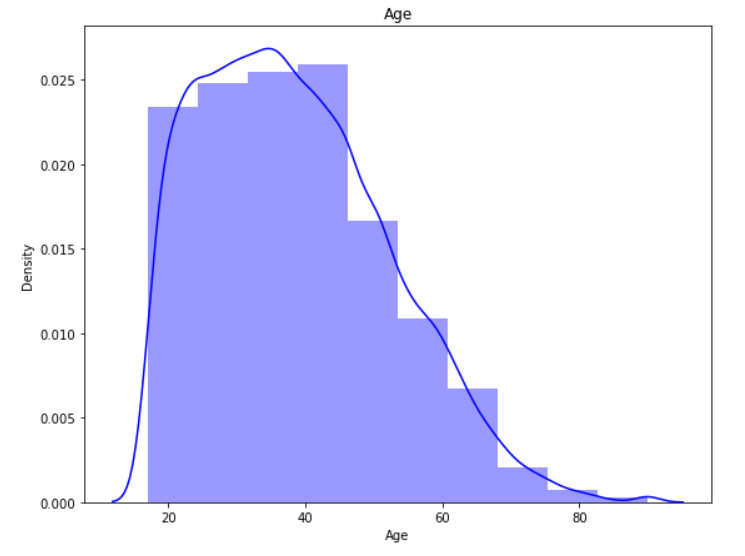
**Income column**

The income column is our target variable with 2 values — ‘<=50K’ and ‘>50K’. The count of these values is 24719 and 7841 respectively, suggesting that people with income higher than 50K are significantly less, and our data set is imbalanced considering the target variable.

**Age column**

The data in age column has a minimum value 17 and max value 90.

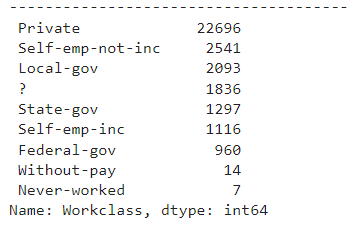
We create a distribution plot for the age column –

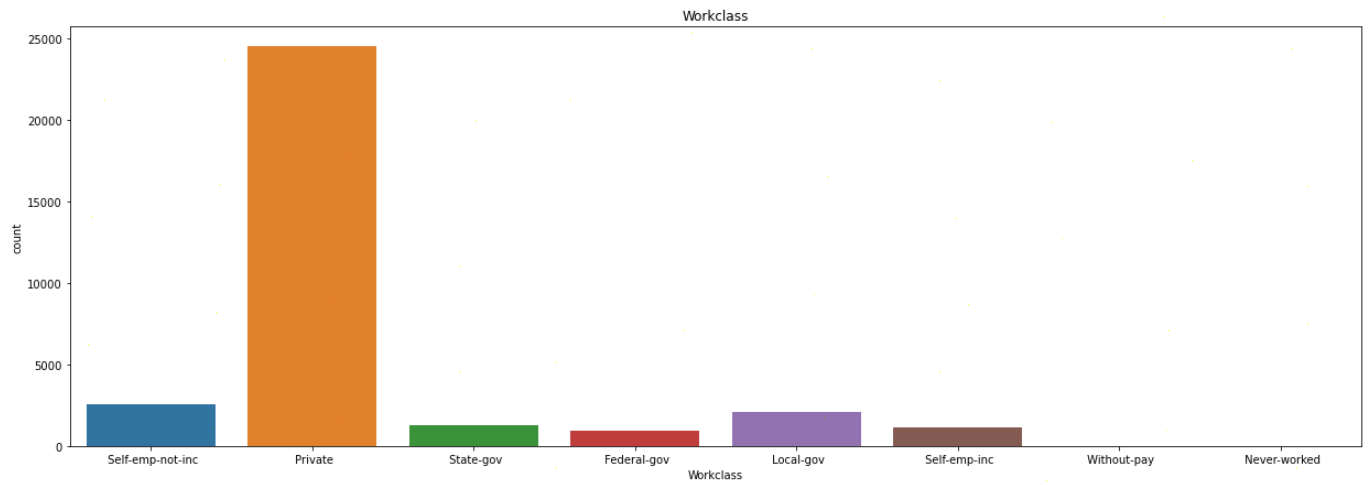


We observe that our data is has right skewness, with majority of the ages falling in the 20–50. The count keeps on decreasing as the age increases.

We also observe that we do not have any null values in the age column.

**Workclass column**

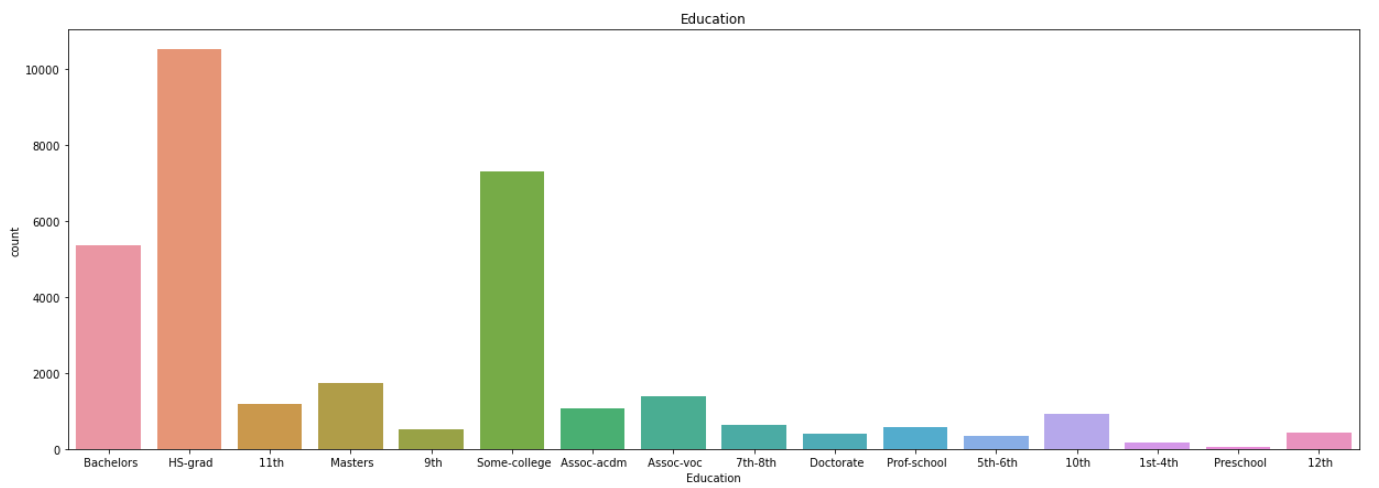
While checking the unique values for workclass, we see that we have 7 different types of values, along with some missing values represented by ‘?’. The count of null values is 1836, which is around 5% of the data.



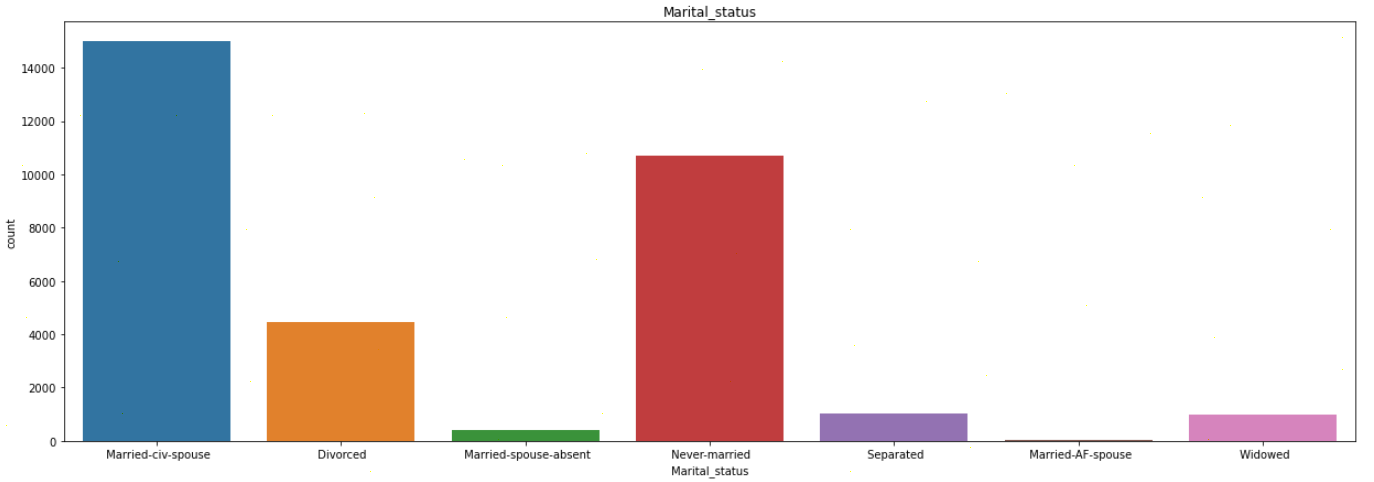
Here we observe that majority of the people belong to ‘Private’ sector workclass.

**Education column**

The ‘Education’ column has 16 different categories available. Majority of these categories belong to ‘School’ type (different classes are divided into multiple categories)



We observe no missing values in this column, and also find out that majority of the people have education level as ‘HS-grad’, followed by ‘Some-college’ and ‘Bachelors’.

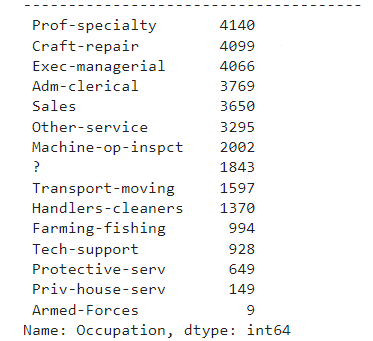
**Marital\_Status column**

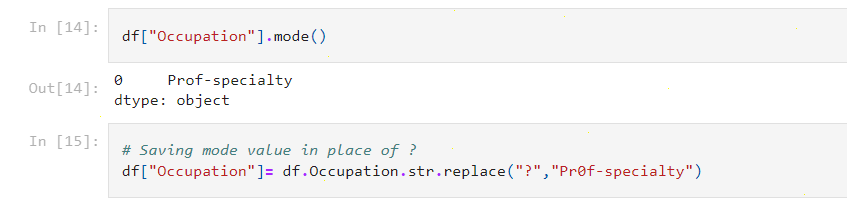
The ‘Marital\_Status’ column has 7 different categories available, and has no missing values.

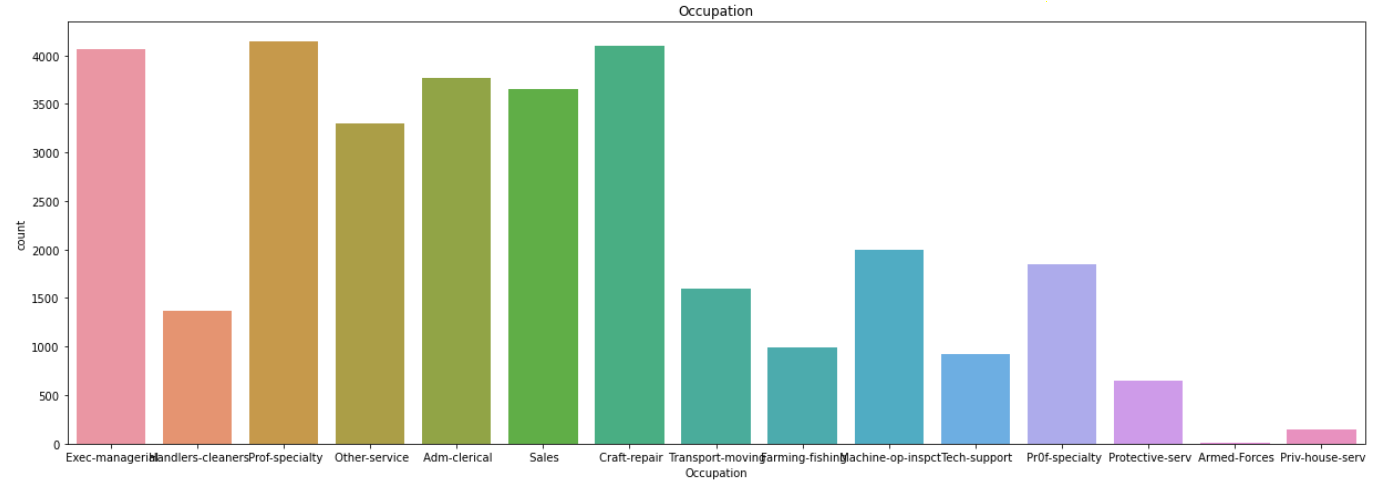
Majority of the people have ‘Marital\_Status’ as ‘Married-civ-spouse’, and least have ‘Married-AF-spouse’. Count of ‘Never-married’ is also quite high.

**Occupation column**

The occupation column contains 14 different categories, and have missing values represented by ‘?’ (which we have already observed, and replaced it with the mode value).



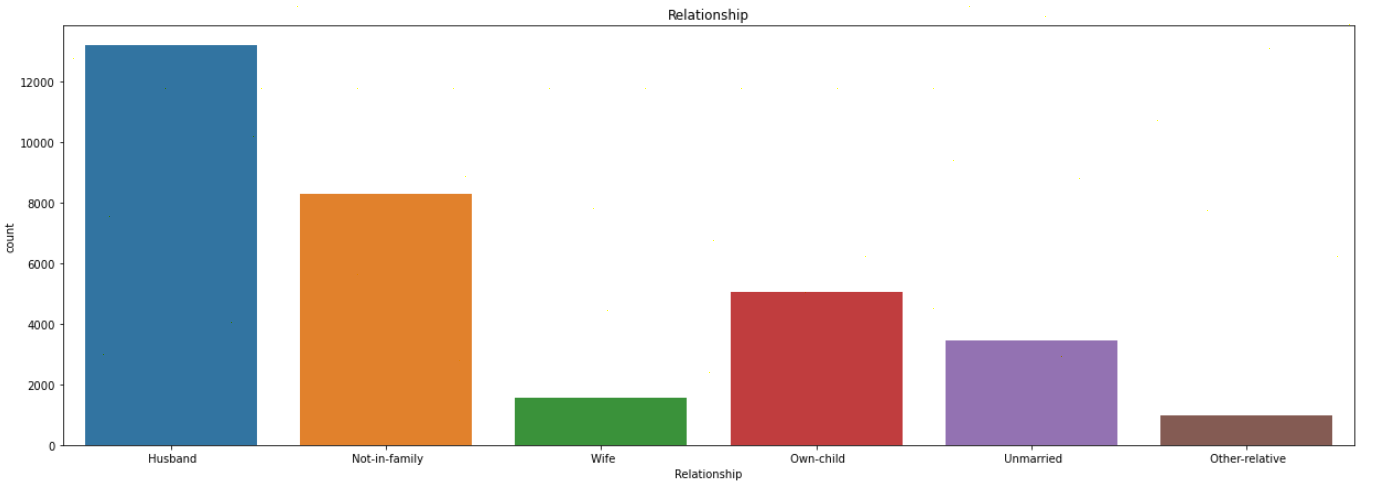




In column Occupation most of the employees are into prof-speciality, executive manager, craft repair.

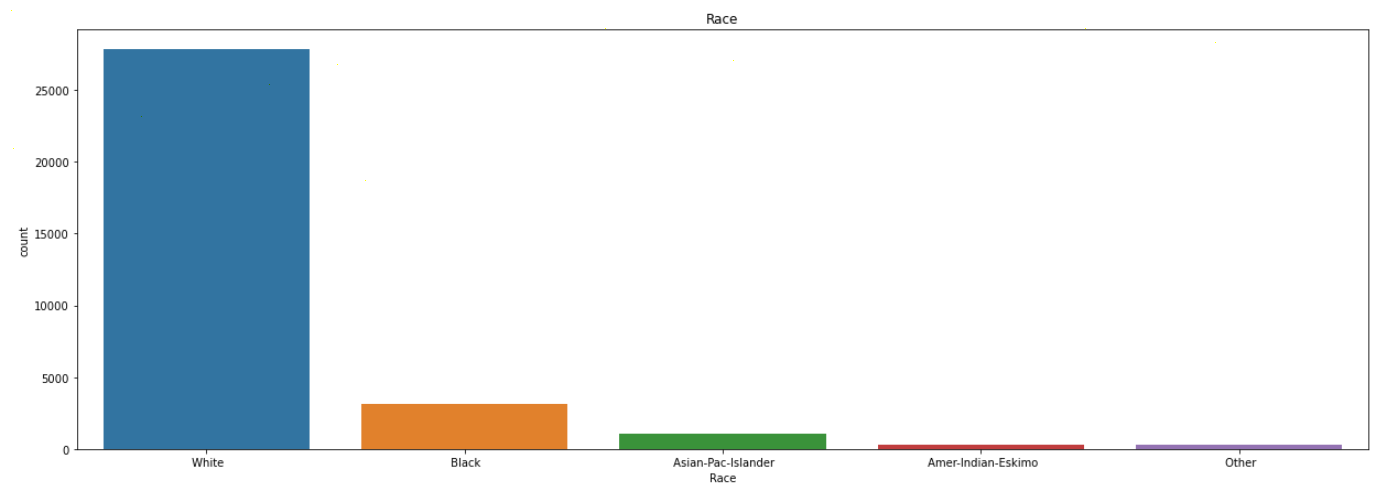
**Relationship column**

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.

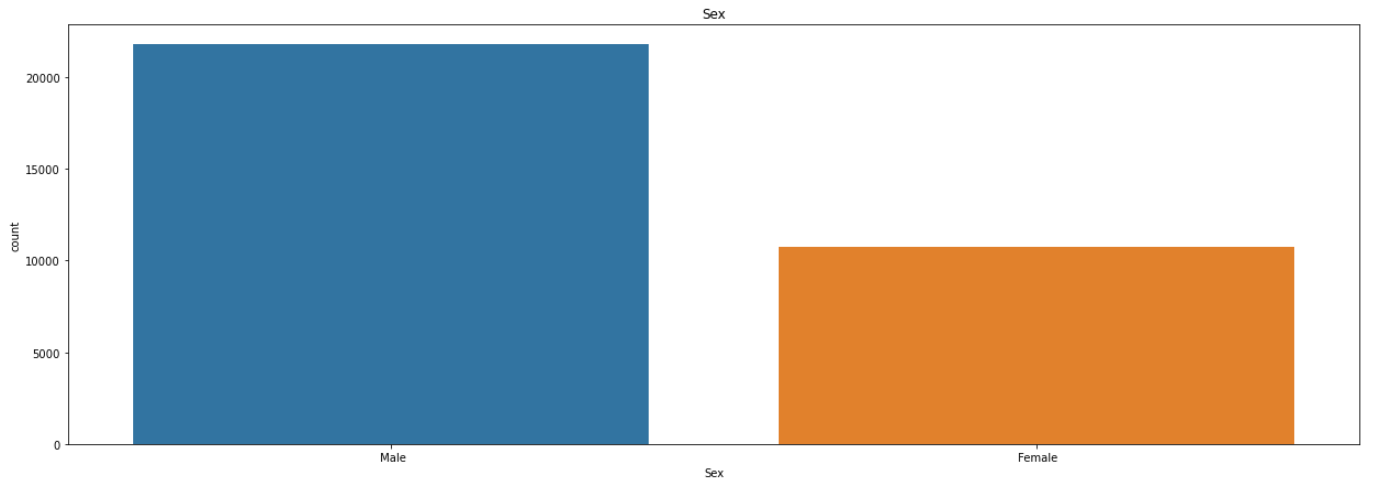


# Race column

The Race column has 5 different categories, and no missing data. Highest number of people have race as ‘White’ (significantly high numbers).

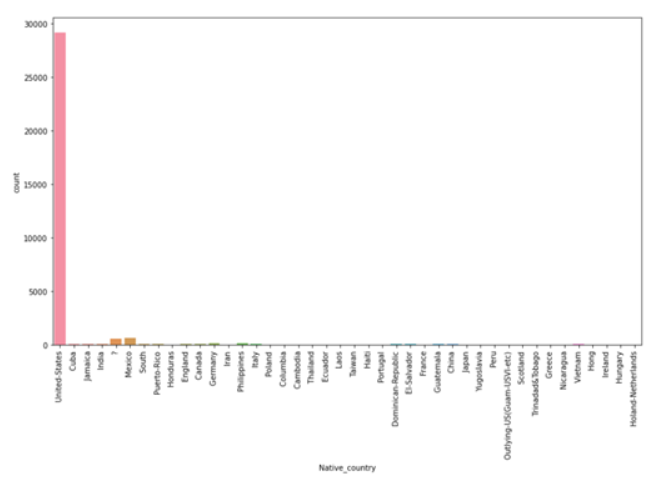


# Sex column

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.

**Native\_country column**

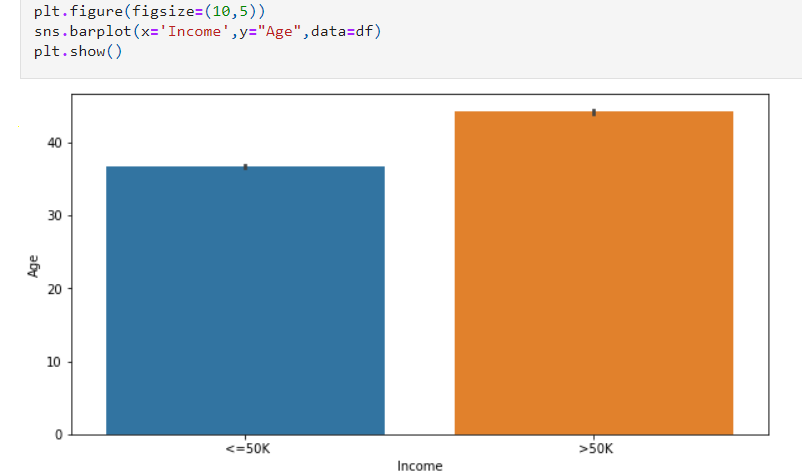
The Native\_country column contains the highest count set to ‘United-States’, and rest of the rows contain quite few numbers (highest count after US is 643). We also have 583 missing values in this column, which we replaced it with mode value.



# Bivariate Analysis:

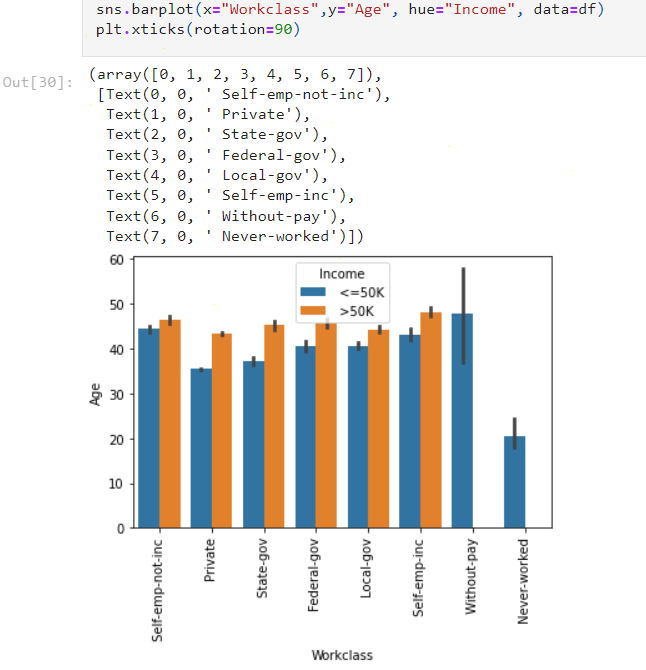
Using this we can analyse 2 columns together; we can check correlation between two columns.

**Correlation between age and income column**



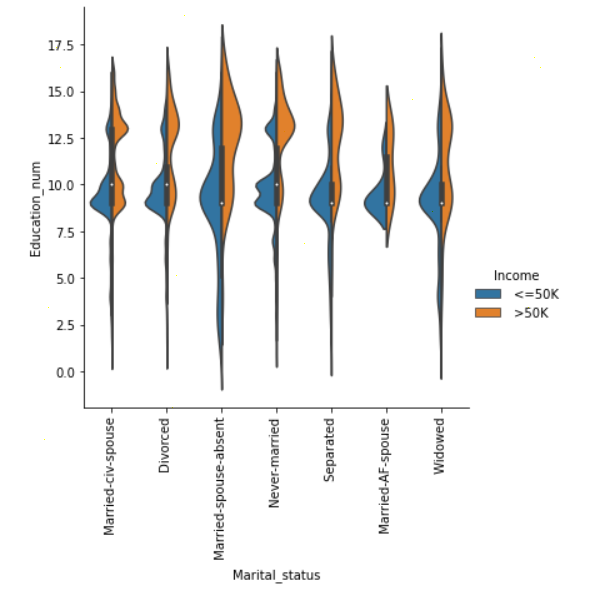
We can clearly observe that more the age and experience of employee more their salaries are.

**Workclass and Age**



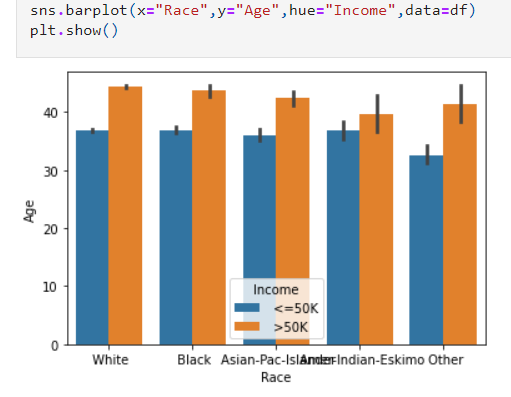
We can observe that clearly people in each field have a chance of getting income >50k except people who never worked or worked without pay.

**Marital\_status and Education\_num**

****

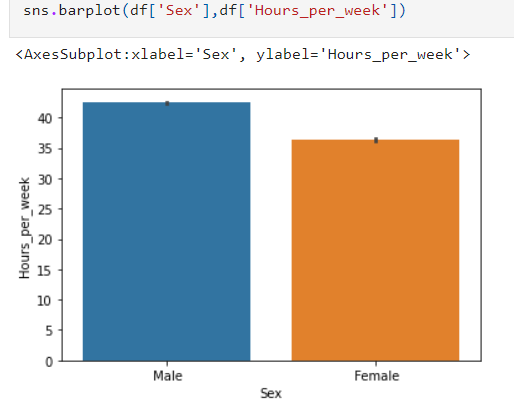
Showing correlation of marital\_status and Education\_num this plot is showing people who are married but spouse is not present has high education number.

**Age and race**

****

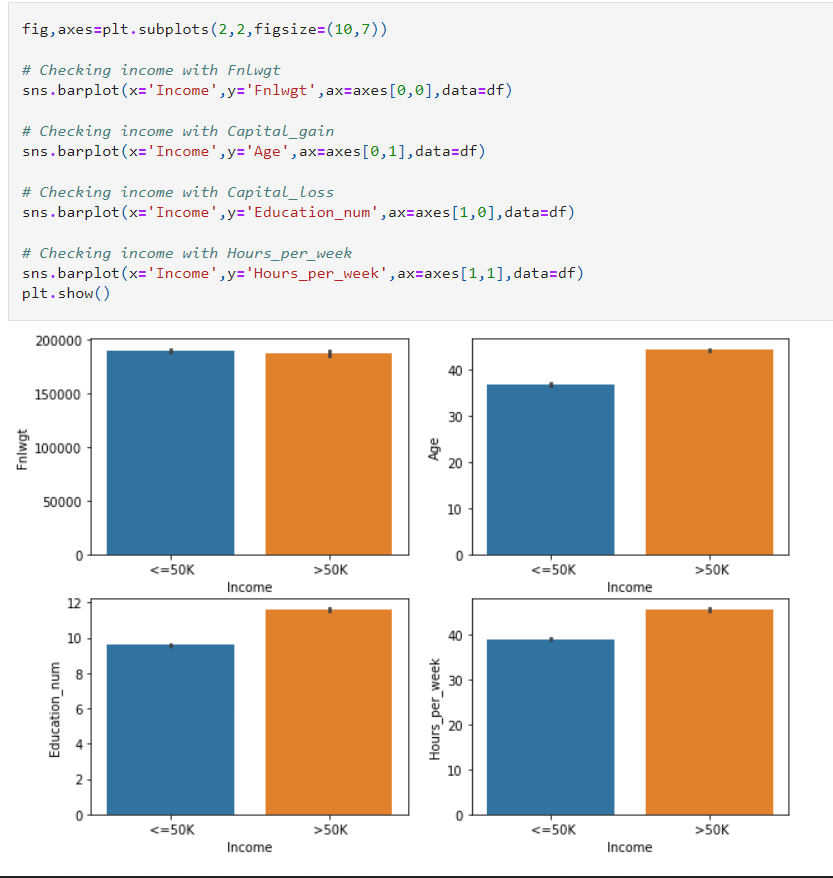
We can clearly observe that white people are getting Higher Incomes in the US as compared to any other category.

**Hours\_per\_week with column Age**

****

Here we can clearly see that males are working more hours per week as compared to females.

We further check how the income gets impacted due to the features Fnlwgt, Capital\_gain, Capital\_loss, Hours\_per\_week :



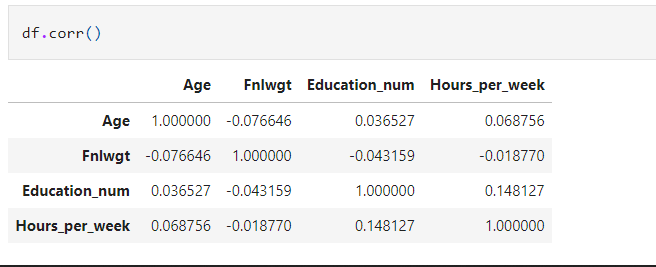
With age chances of getting >50k are more as u will also gain experience by that time.

Higher the Education\_num, higher are your chances of getting >50k.

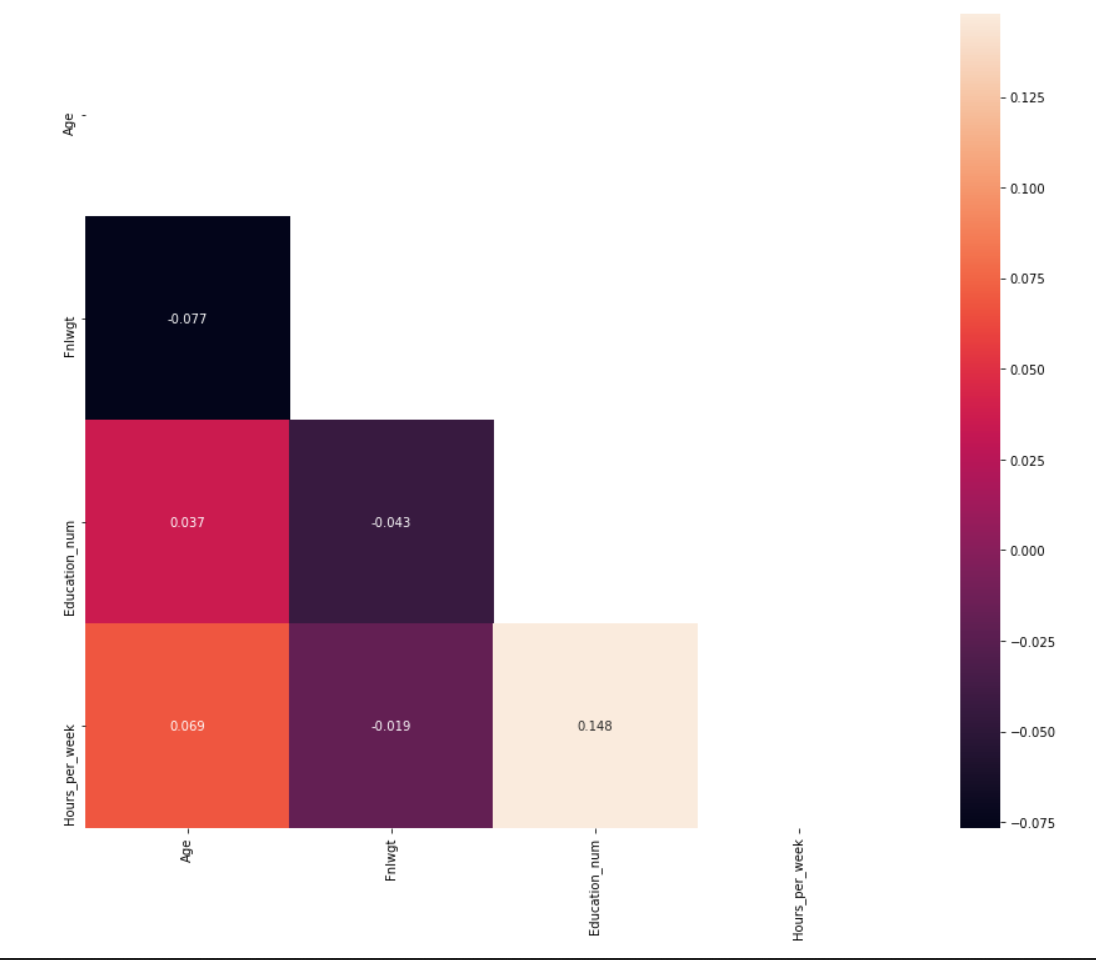
More you work, more income you get, more hours per week you work, more are the chances of getting income more than 50k.

**Correlation of all columns in the dataframe**

Here, we use data visualization technique which contains values representing various shades of the same colour for each value to be plotted. Usually, the darker shades in the map represent higher values than the lighter shade. For a very different value a completely different colour can also be used.

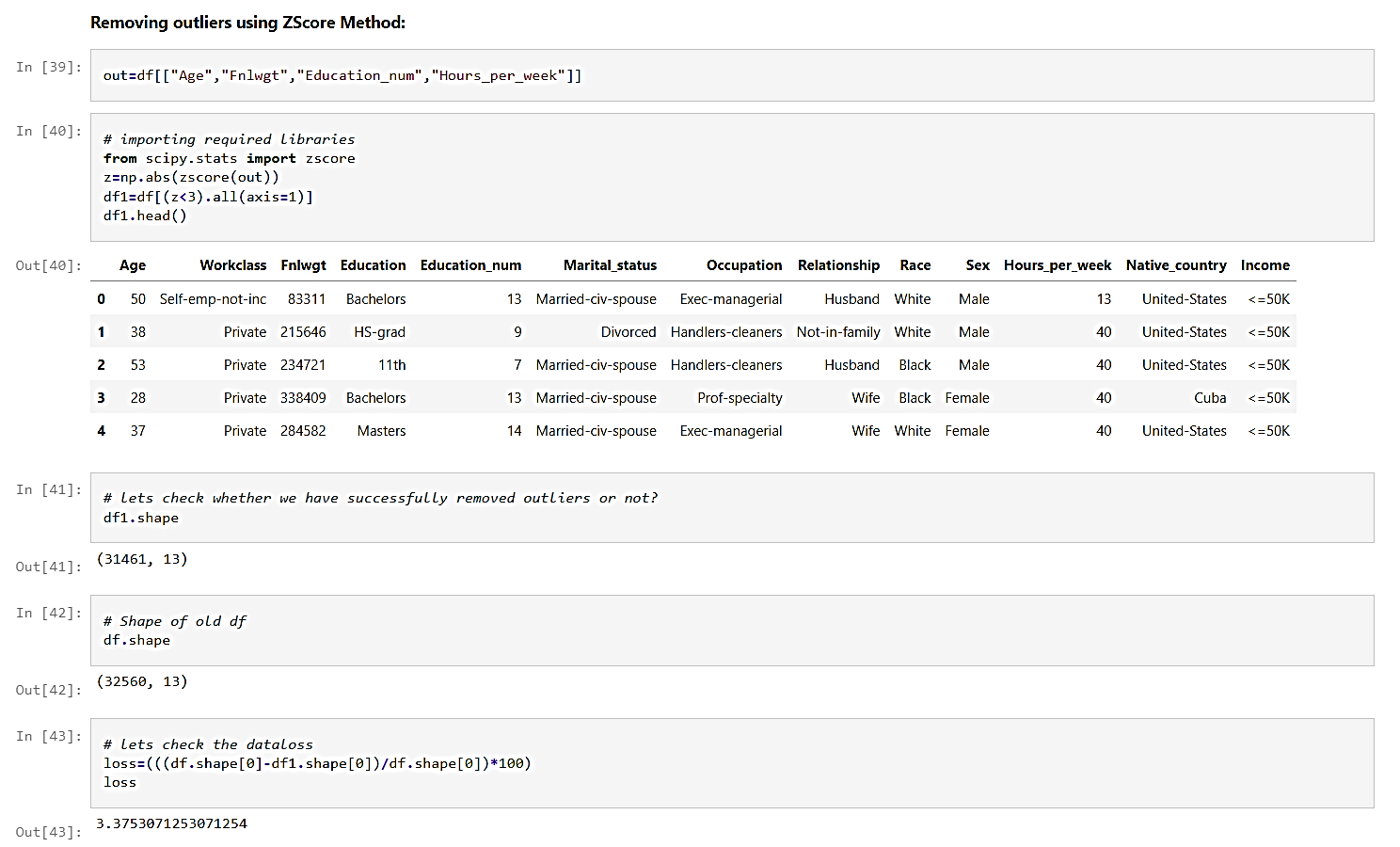


**Heatmap**



**4.Pre-processing Pipeline**

**Outlier Removal**

**Outliers**are data points that are distant from other similar points. They may be due to variability in the measurement or may indicate experimental errors. If possible, outliers should be excluded from the data set. There are two methods to remove outliers: Zscore and IQR. Here we are using Zscore method for outlier removal as there are few columns having outliers.

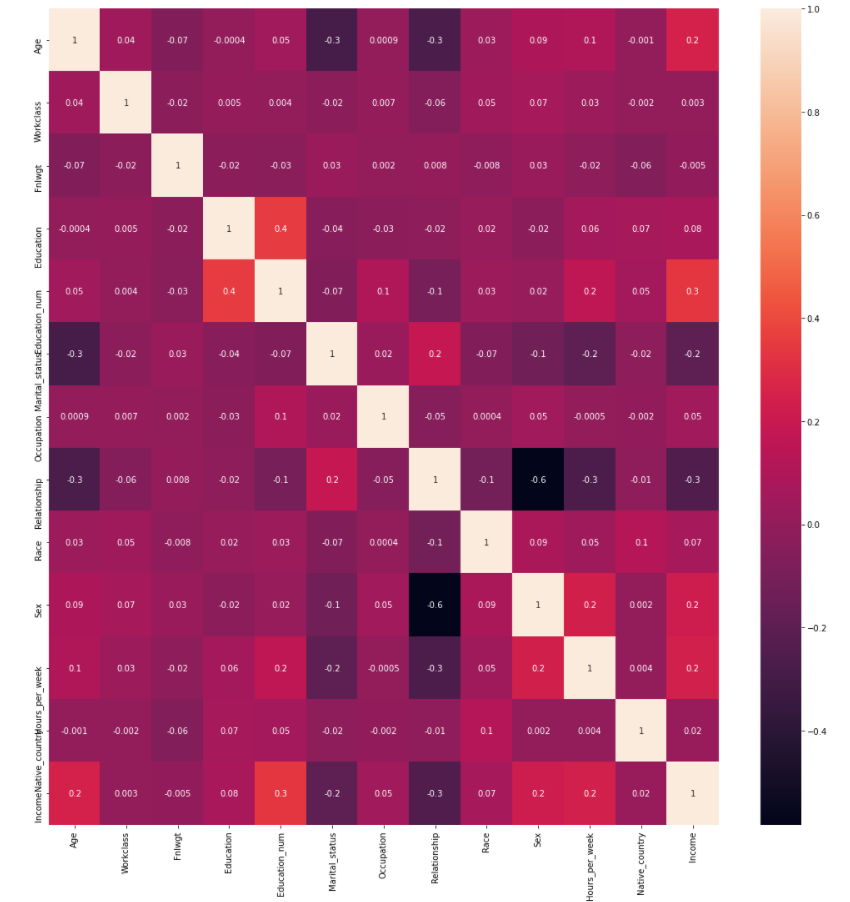
**Checking skewness and removing it from our data**

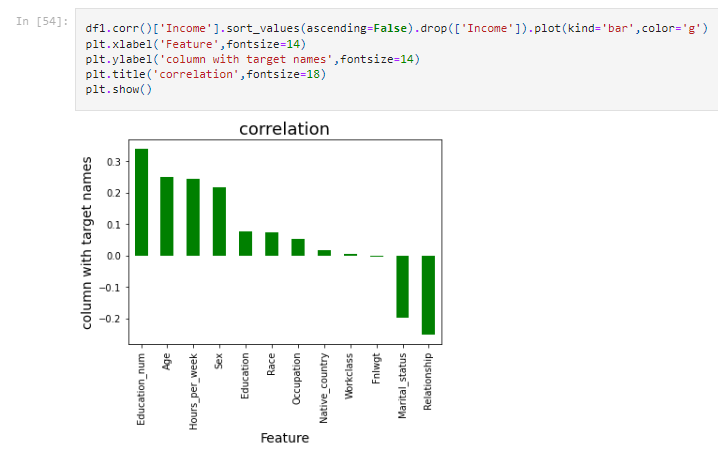
****

**Now we will** convert the labels into numeric form so as to convert it into the machine-readable form using label Encoder. Machine learning algorithms can then decide in a better way on how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

**Label Encoding:**

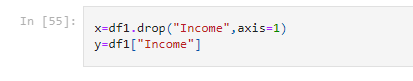


**Check correlation in all columns using heatmap**

Here in the heatmap we can observe that only age and education no. have some positive correlation with our target column Income. Even hours\_per\_week, sex and native\_country also relate positively with the target column income. Relationship and marital\_status have negative correlation with target column.

Here we can see more clearly that marital\_status and relationship are negatively correlated with the target column.

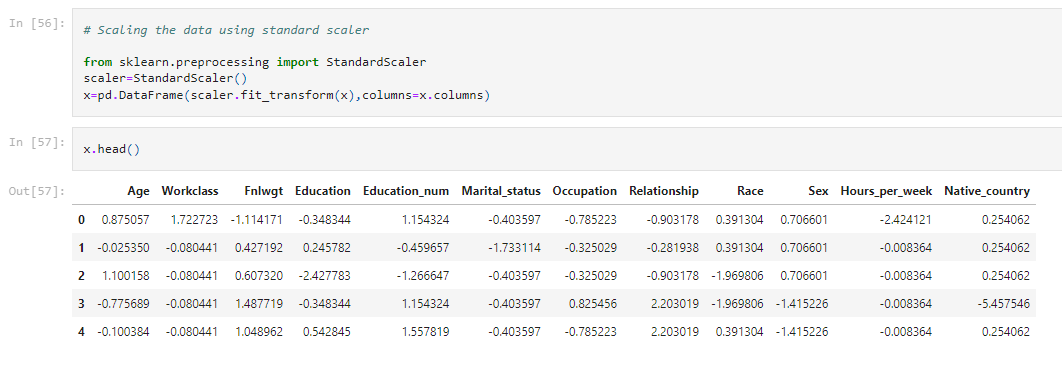
**Separating x and y for training and testing purpose**

****

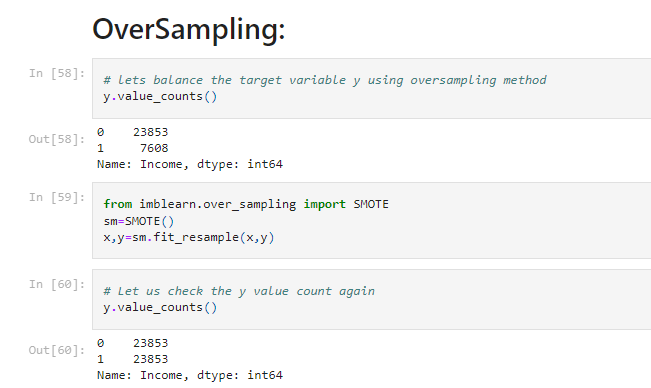
**Scaling the data**

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 –

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

**Balancing our imbalanced data**

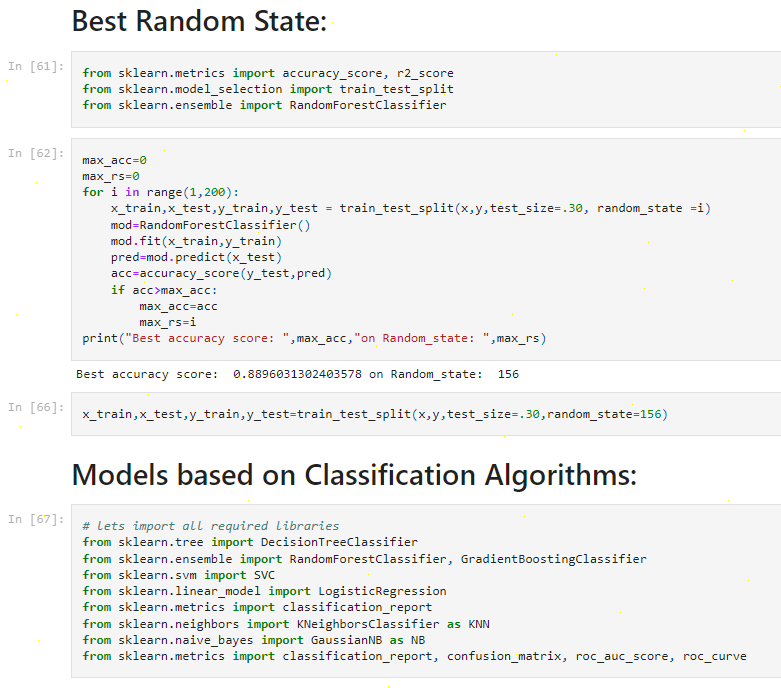


Here we can see that we have successfully balanced the data using oversampling.

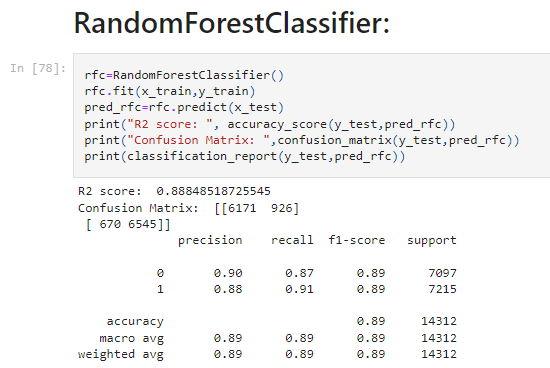
**5.** **Building Machine Learning Models**

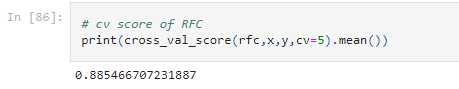
**Fitting data into classification models**

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 –

****

After performing train\_test\_split we train several machine learning models i.e., Logistic Regression, Random Forest Classifier, Decision Tree Classifier, Support Vector Classifier, K-Neighbour Classifier, Gradient Boosting Classifier and compare their results.

Out of all, we select Random Forest Classifier as our final model with accuracy 88.9% and its cross-validation score 88.54%.



**Hyper parameter tuning on best classifier model**

Hyper parameter optimization in machine learning intends to find the hyper parameters of a given machine learning algorithm which delivers the best performance as measured on a validation set.

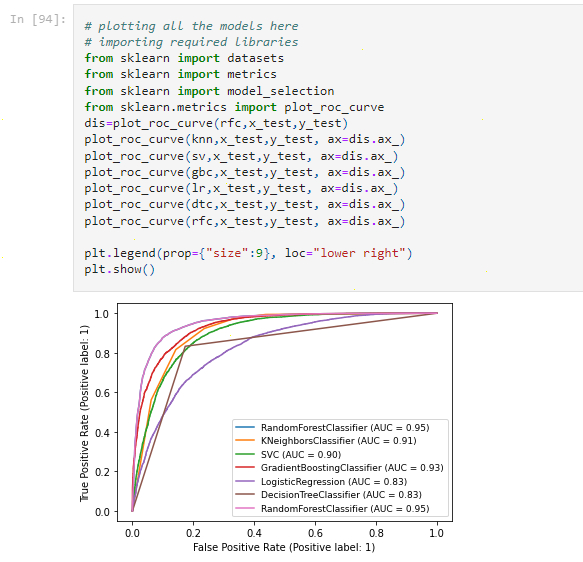
Hyper parameterscannot be directly learned from the regular training process. They are usually fixed before the actual training process begins. These parameters express important properties of the model such as its complexity or how fast it should learn. There are many hyper parameters as if now we are using GridSearchCV.

****In GridSearchCV approach, machine learning model is evaluated for a range of hyperparameter values. This approach is called GridSearchCV, because it searches for best set of hyperparameters from a grid of hyperparameters values.

The accuracy after using hyper parameter tuning is 88.89% for our model which is quite good.

**ROC AUC Curve:**

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.



The AUC model for our RandomForestClassifier model is 95%.

6. **Concluding Remarks**

The project aims to increase the awareness about how the income factor actually has an impact not only on the personal lives of people, but also an impact on the nation and its betterment. We have had a look on the data extracted from the 1994 Census bureau database, and tried 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, but it would surely help us to analyze what role different features play in predicting the income of an individual.

We moved step by step, analyzing, cleaning and modelling 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.

The best and final fitted model was RandomForestClassifier with R2 score of 88.9% and cross-validation score of 88.54%. The model performed excellent. The model’s R2 score and ROC AUC scores were the highest amongst the other models.

**There is still room for improvement, like doing a more extensive feature engineering, by comparing and plotting the features against each other and identifying and removing the noisy features.**