**CAPSTONE PROJECT 2:**

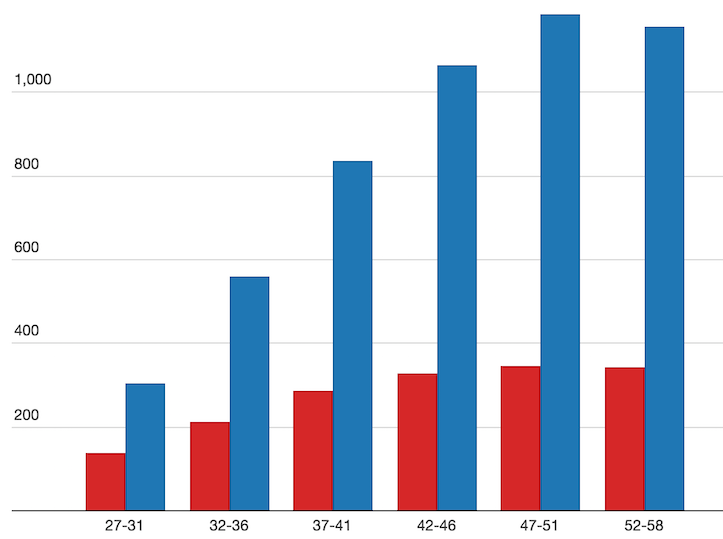
**FINAL REPORT**

**INCOME CLASSIFICATION**

**12/8/2019**

**Mehmet KETENCI**

**KANSAS CITY, MO**



**PROBLEM**: Our goal is to be able to classify given data with respect to certain income level.

**CLIENTS AND WHY DO THEY CARE**: Result of this dataset has crucial importance for those industries below in determining tendency of group of people both high income level with low income level.

- **Marketing and e-marketing companies**:Companies may offer their different products with respect to different income levels.

- **Health, car, home and life insurance companies**: Data will be produced by statistical inferences and Eda with Machine Learning is beneficial for this industry due to mapping right customer and right insurance type and amount.

- **Investment industry**:This industry can aim right income class with respect to customers` different features.

- **Loan companies and banks**:Classified data can be used for this industry to match correct loan or credit amount with targeted people.

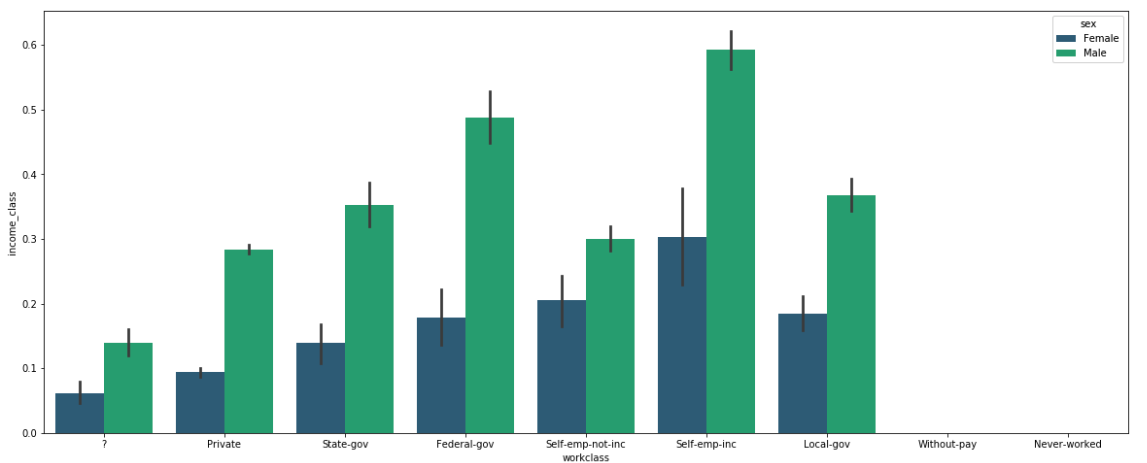
- **Travel agencies**: People have different vacation habits,some people may want to go to the seaside while some enjoys spending time mountains and some like museums or historical places.This industry may offer right travel options to correct people.

**WHAT WILL CLIENTS DECIDE AFTER MY PROJECT:** They can easily determine their target mass with simplified numbers or values and prominent and well-organized visuals.

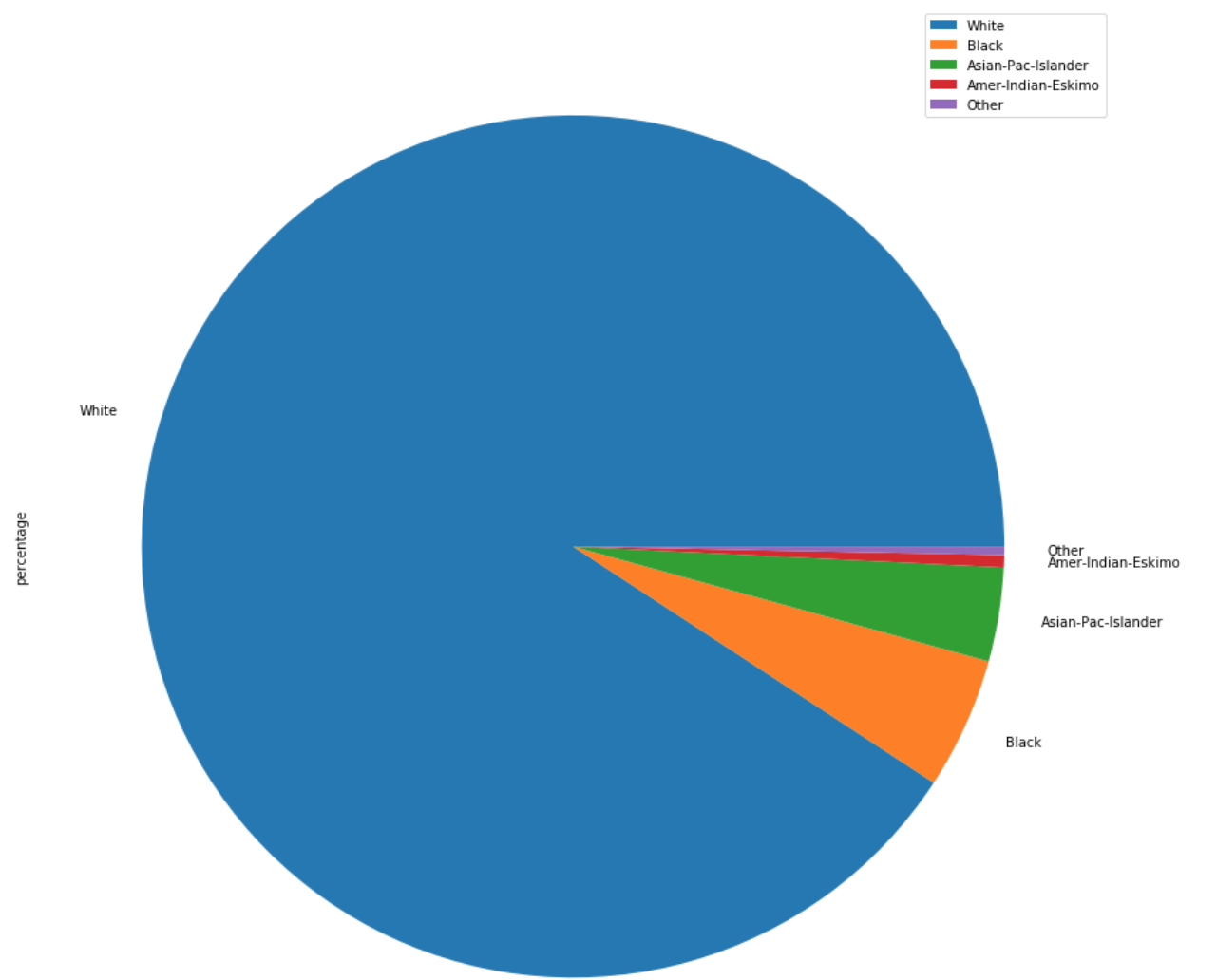
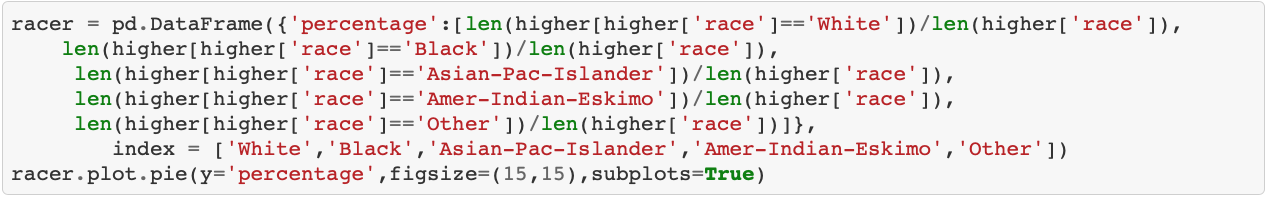
**HOW WILL I ACQUIRE THE DATA?**

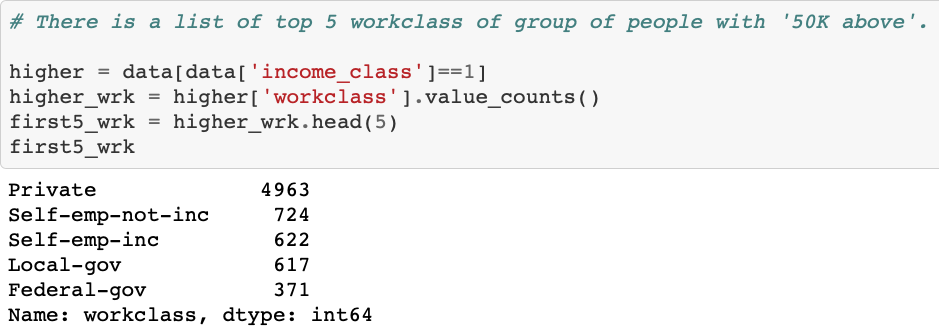
Data itself was obtained on <http://archive.ics.uci.edu/ml/datasets/Adult>

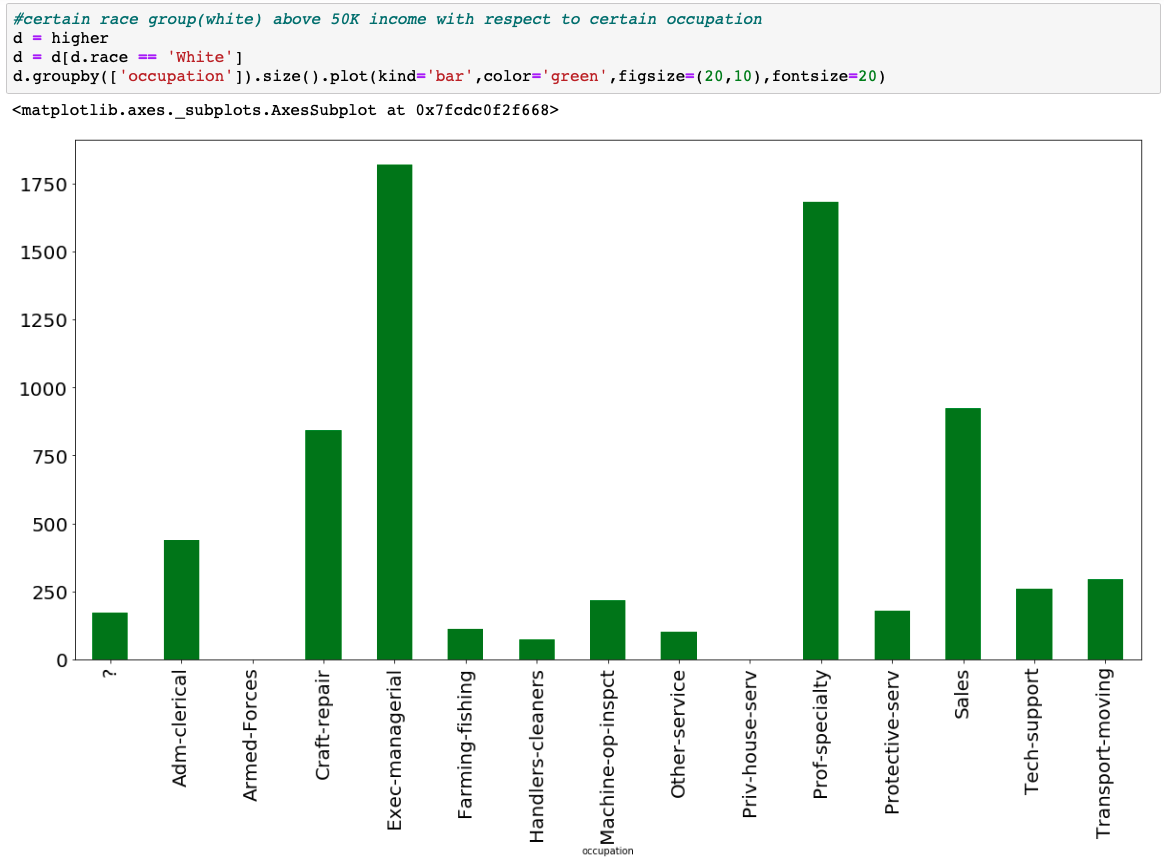
**DATA WRANGLING AND EDA PART**



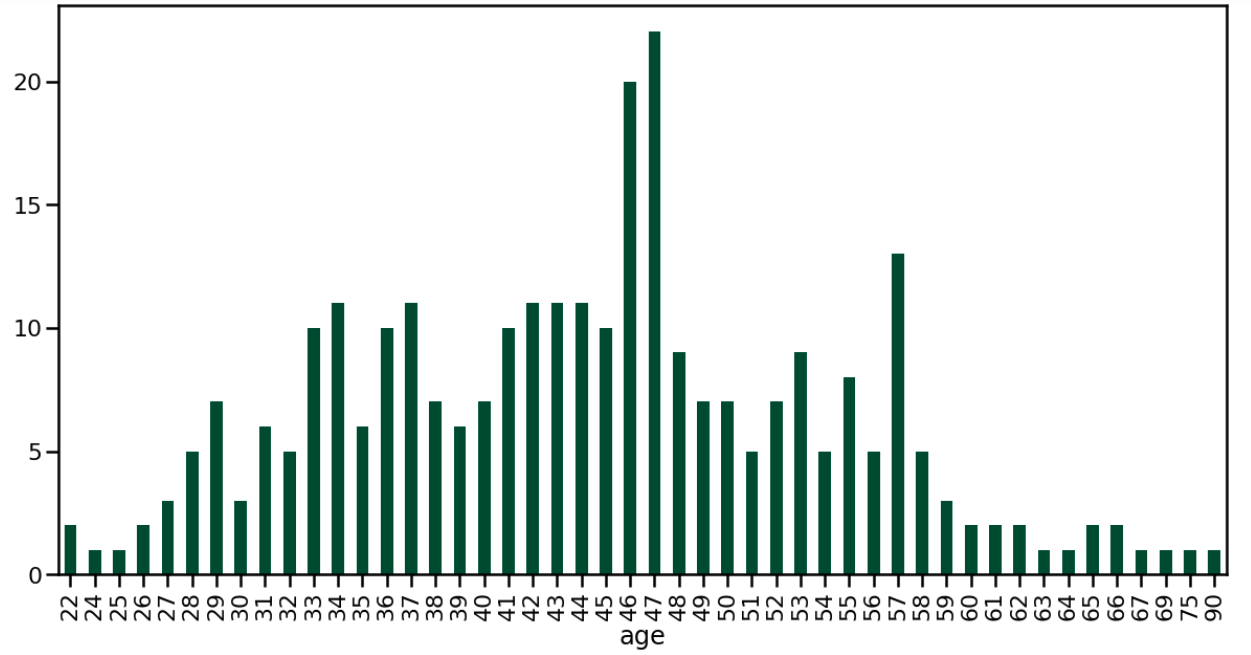
This graph shows the income distribution of groups that depend on a specific work class and gender.



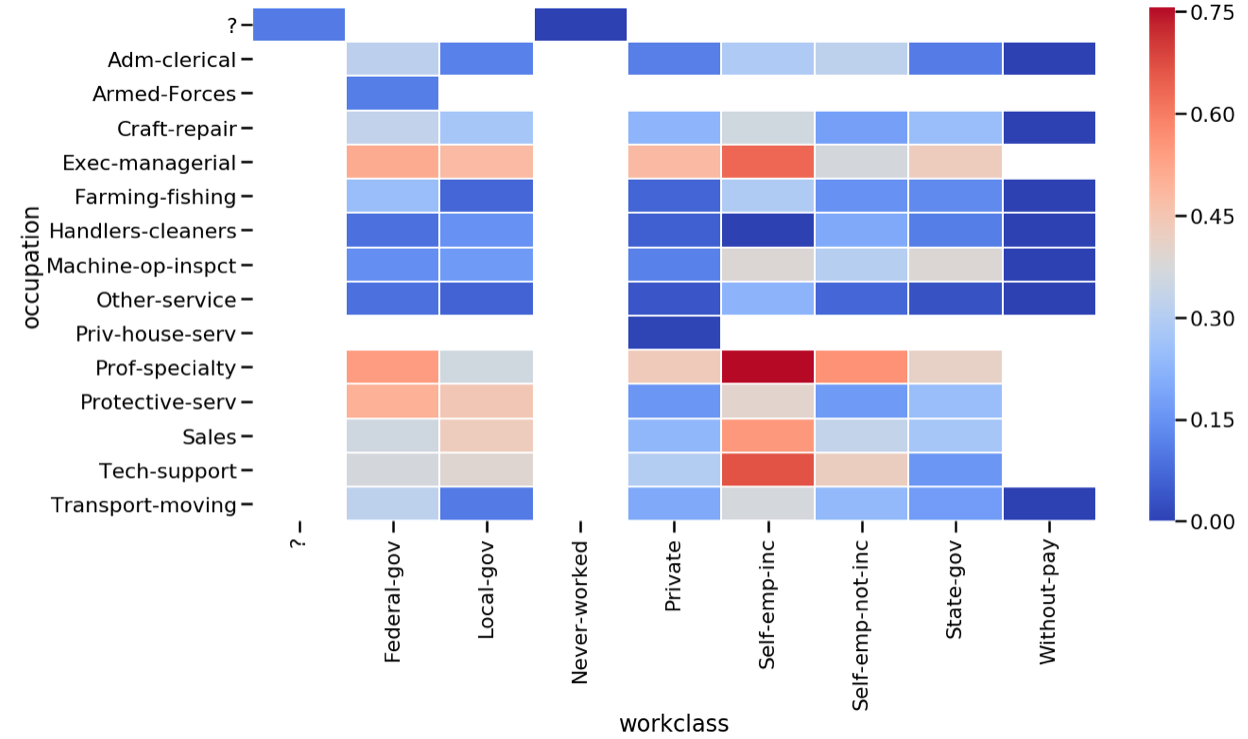


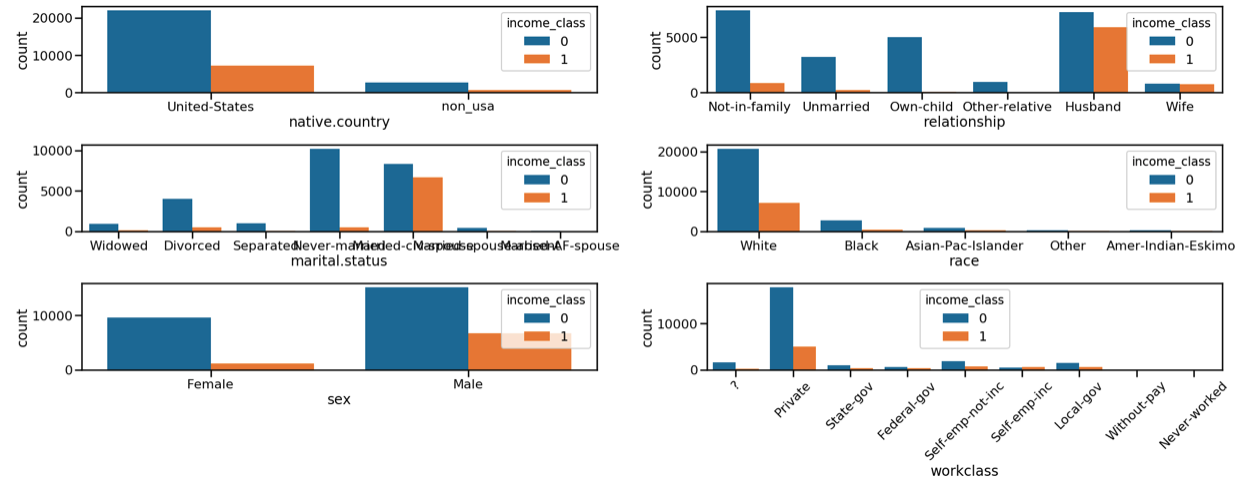


This graph shows the professional distribution of the white population with more than 50 thousand income. The top 5 consists of Exec-managerial,Prof-speciality,Sales,Craft-repair and Adm-clerical.

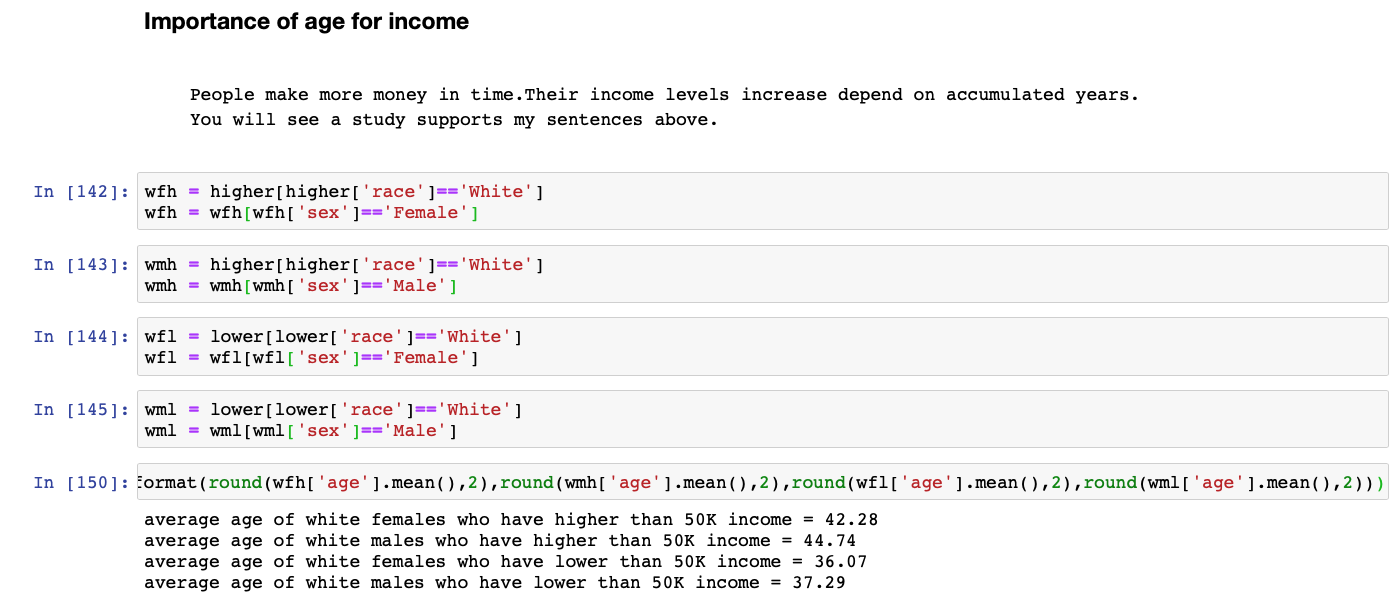
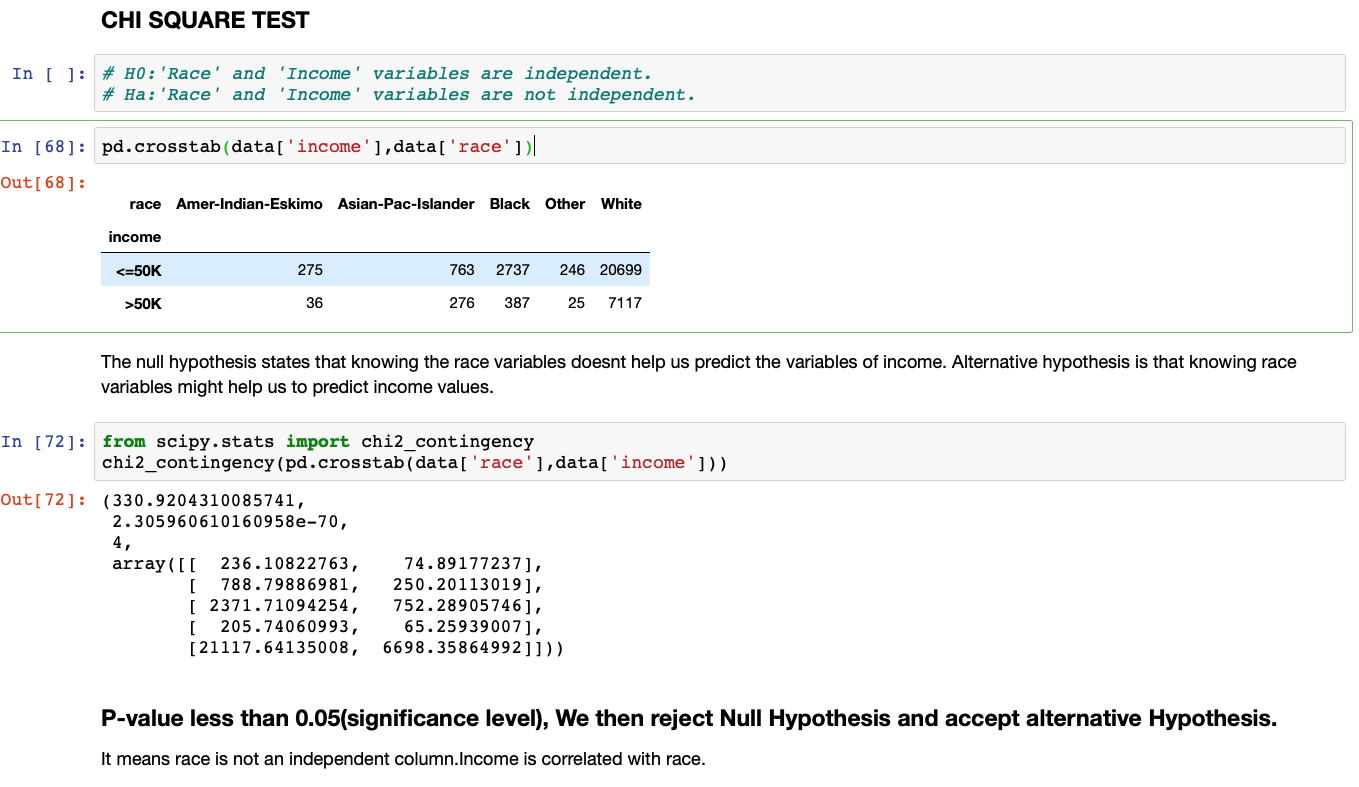


This graph shows the age distribution of the high-income group of black men. According to the graph, the surplus of people aged 46 and 47 is in the foreground.

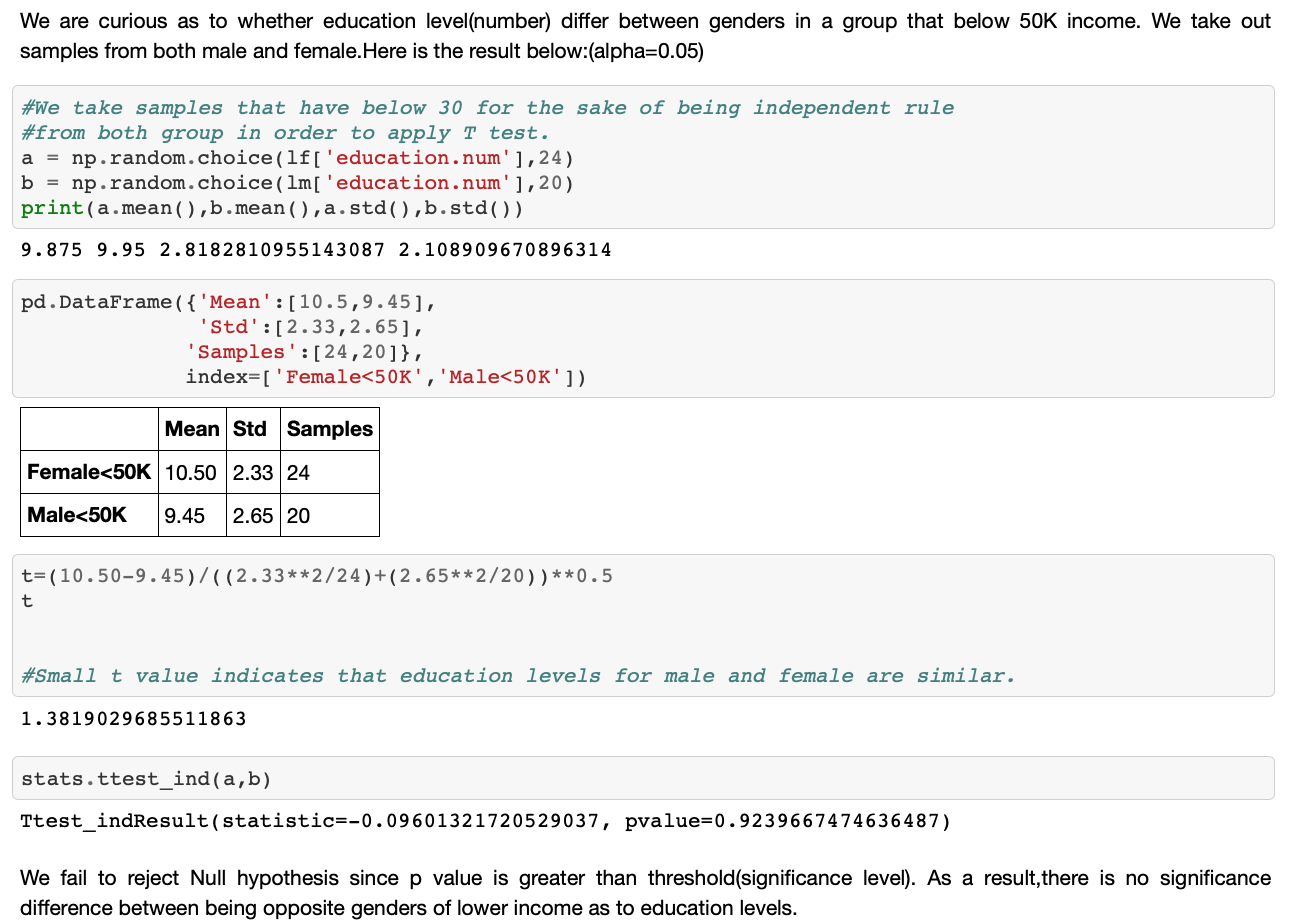
Here we see the heatmap graph according to the income averages of the people working in a certain occupational group and sector and according to this graph (Prof-specialty, Self-emp-inc), (Exec-managerial, Self-emp-inc), (Tech-support, Self-emp) -inc), (Sales, Self-emp-inc) and () combinations have a high average income.



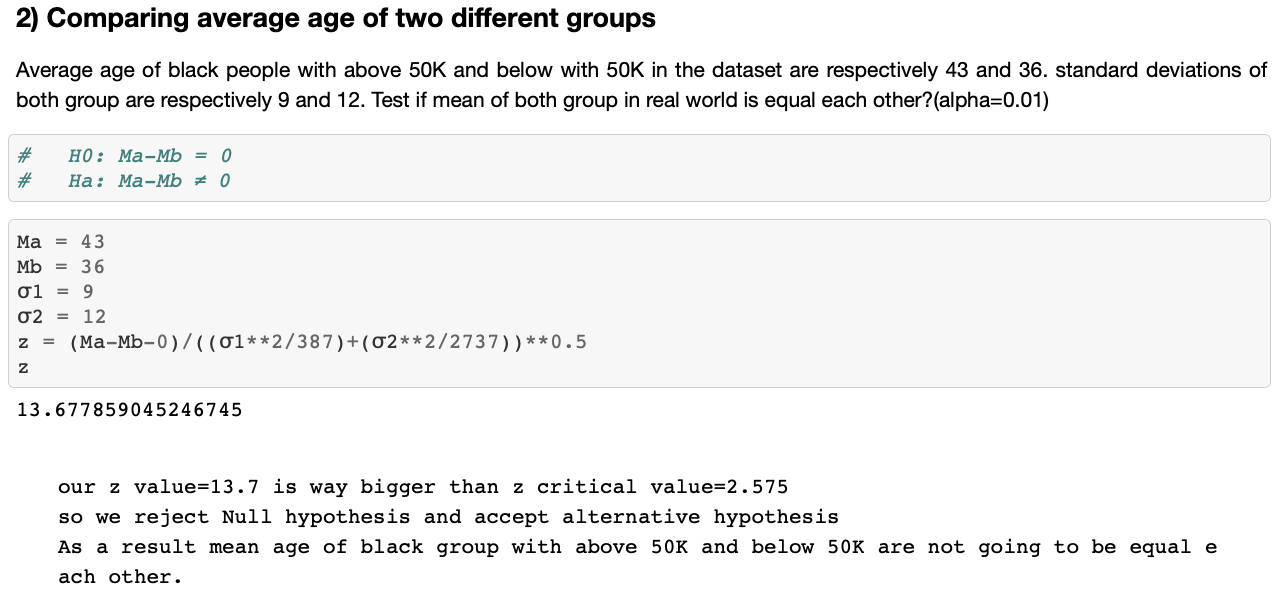
Here we also see various comparisons based on income classes.

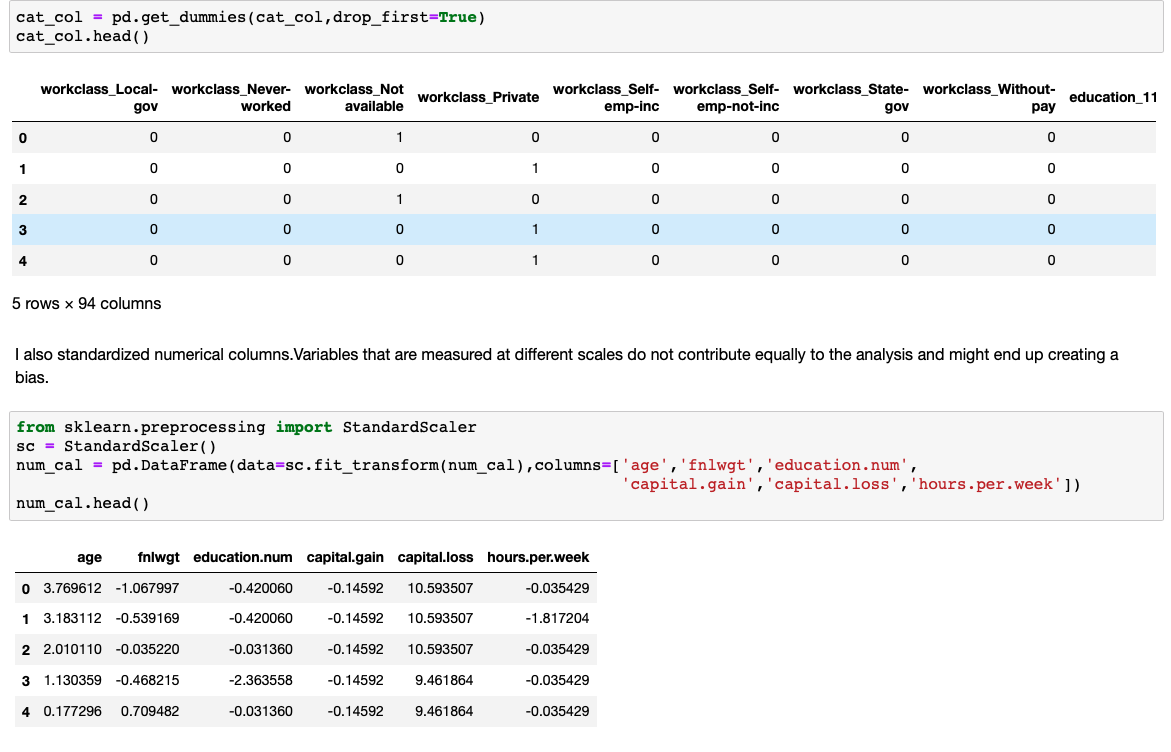
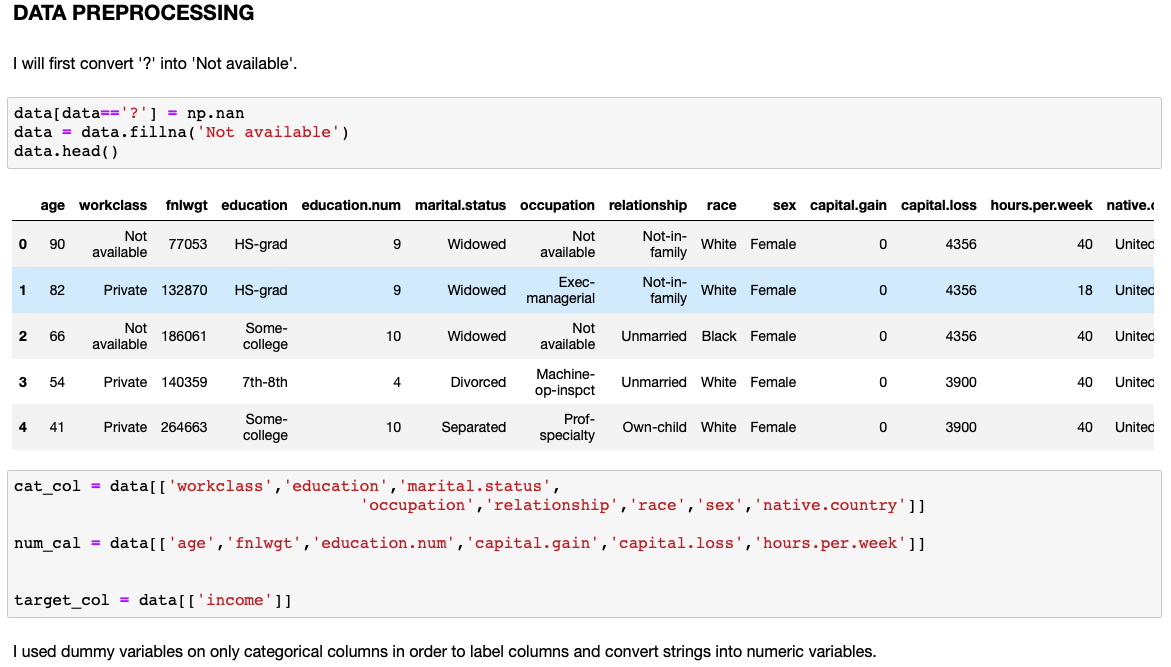
**STATISTICAL PART**

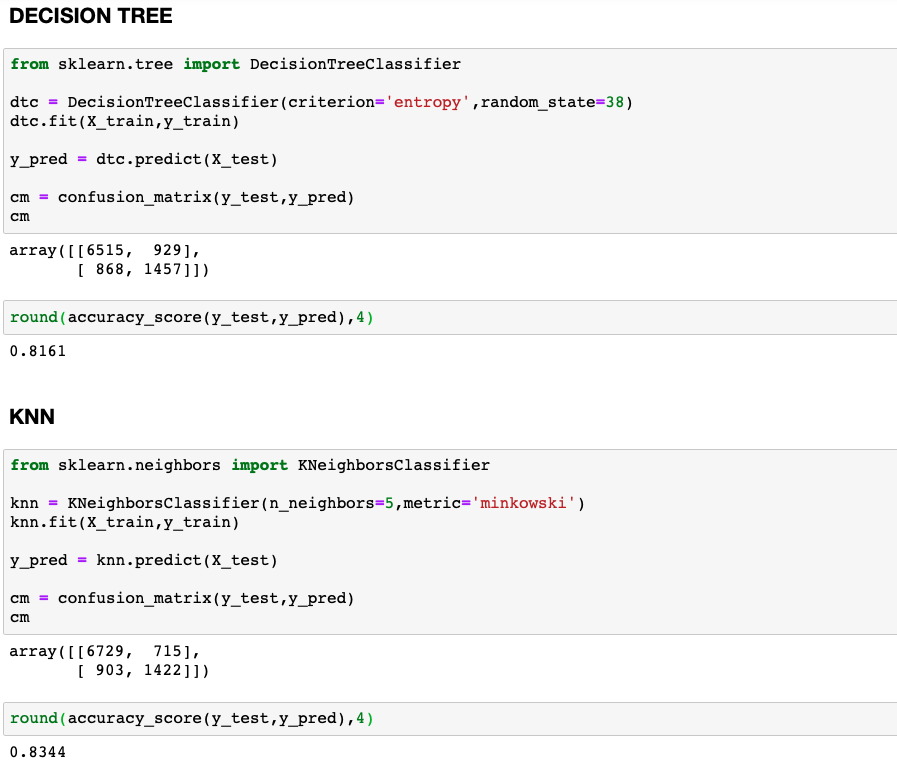
**T-TEST**



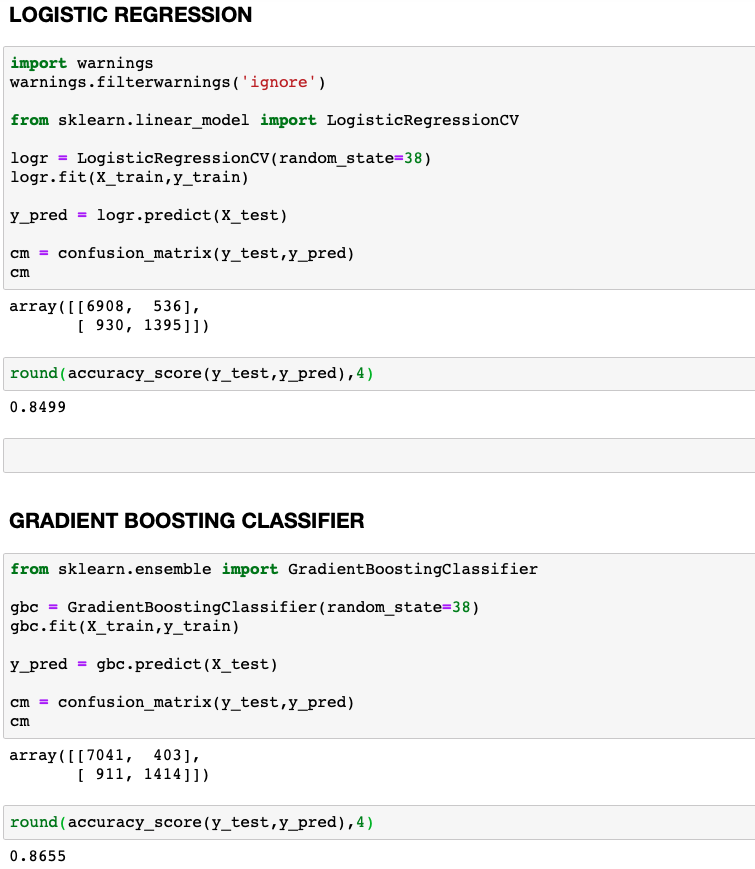
**Z-TEST**

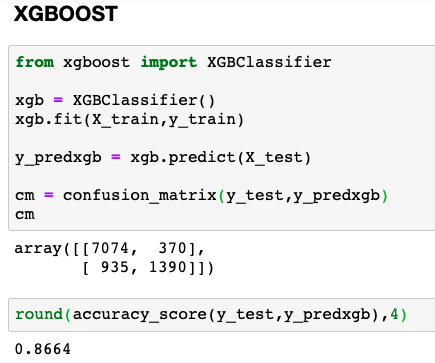
****

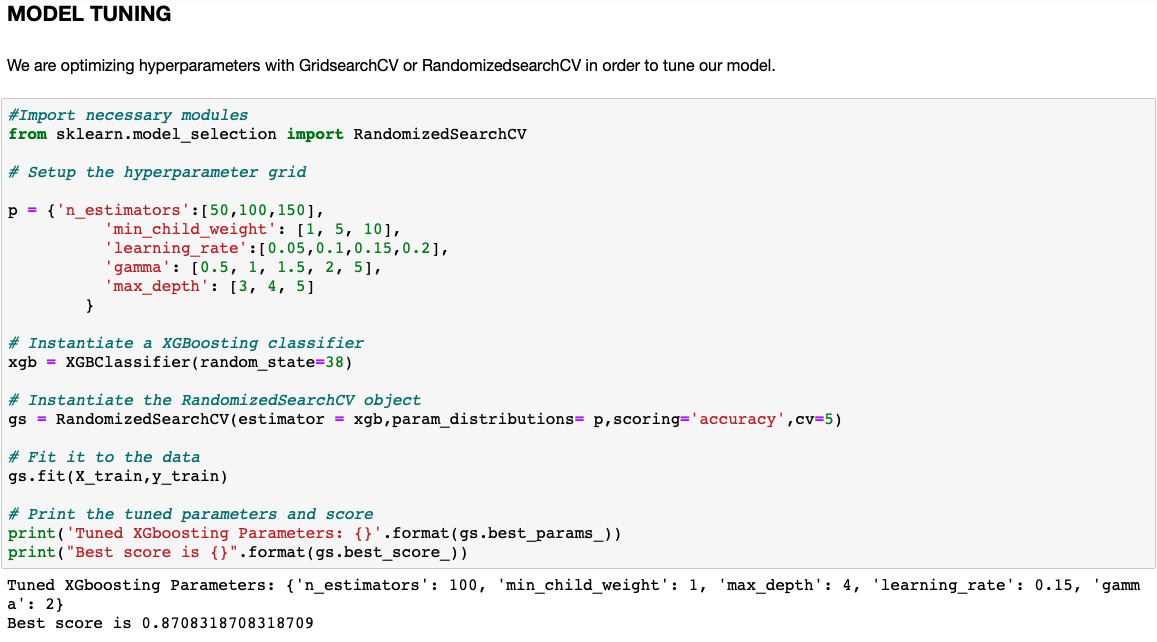
**MACHINE LEARNING PART**

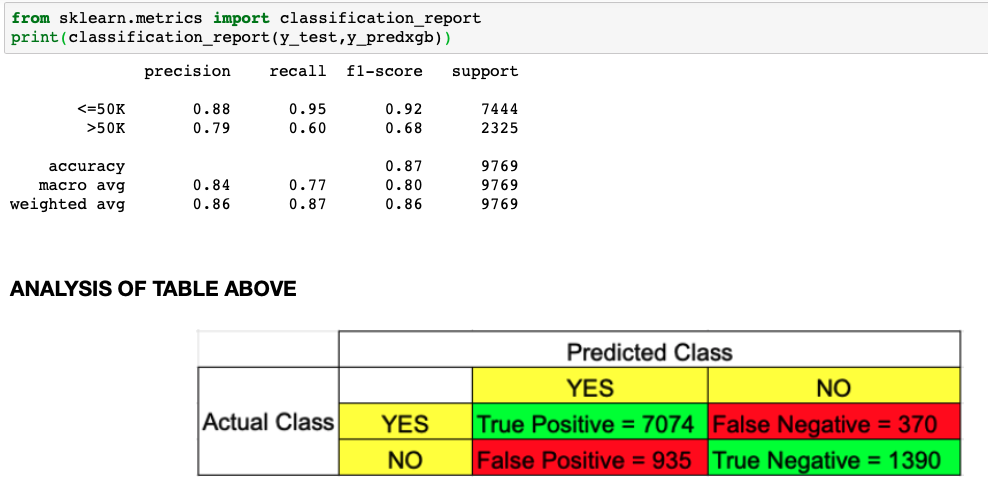










The first metric we look as a data scientist is accuracy = 87%. It looks awesome but it is insufficient 

if we check merely accuracy.There are other concepts to analyze our dataset being health enough as well.

Let's dive into it.

Precision: Precision is a good measure to determine, when the costs of False Positive is high. Since our false positive value,precision can be good interpreter in here.It is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of people that labeled as below 50K, how many of them actually below 50K? or of labeled as above 50K, how many of them actually above 50K?

In here almost 90% of correct prediction on <=50K and almost 80% correct prediction on >50K.These high precision make our model more reliable.

Recall: Recall actually calculates how many of the Actual Positives our model capture through labeling it as Positive (True Positive). Applying the same understanding, we know that Recall shall be the model metric we use to select our best model when there is a high cost associated with False Negative.Especially predicting 95% correct on people below 50K is incredible estimation while other group needs to be worked much on it.

F1-Score: It is required when we seek a balance between Precision and Recall.It is a harmonic mean of both precision and recall.F1-score for both groups, especially people who have income below 50K is very high.

Overall with %87 accuracy, we built very good model by ML algorithms.

