Census Income Project

Introduction

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:

1. A single cell estimate of the population 16+ for each state.
2. Controls for Hispanic Origin by age and sex.
3. 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 a collection of 51 state samples, each with its probability of selection, the statement only applies within the state.

**The below is the list of columns provided to predict the Census Income Project:-**

**Age:** shows the age of the person

**Workclass:** shows the working class of the person

**Fnlwgt:** shows the Fnlwgt value

**Education:** shows the education of the person

**Education\_num:** shows the Education number

**Marital\_status:** shows the marital status of the person

**Occupation:** This Shows the occupation of a person

**Relationship:** shows the relationship type

**Race:** Shows the race of the person

**Sex:** shows the Gender of the person

**Capital\_gain:** Shows the capital gain

**Capital\_loss:** This shows the capital loss

**Hours\_per\_week:** This shows the hours invested by a person per week

**Native\_country:** Shows the native country of the person

**Income:** shows the income of the person

**Process Steps**

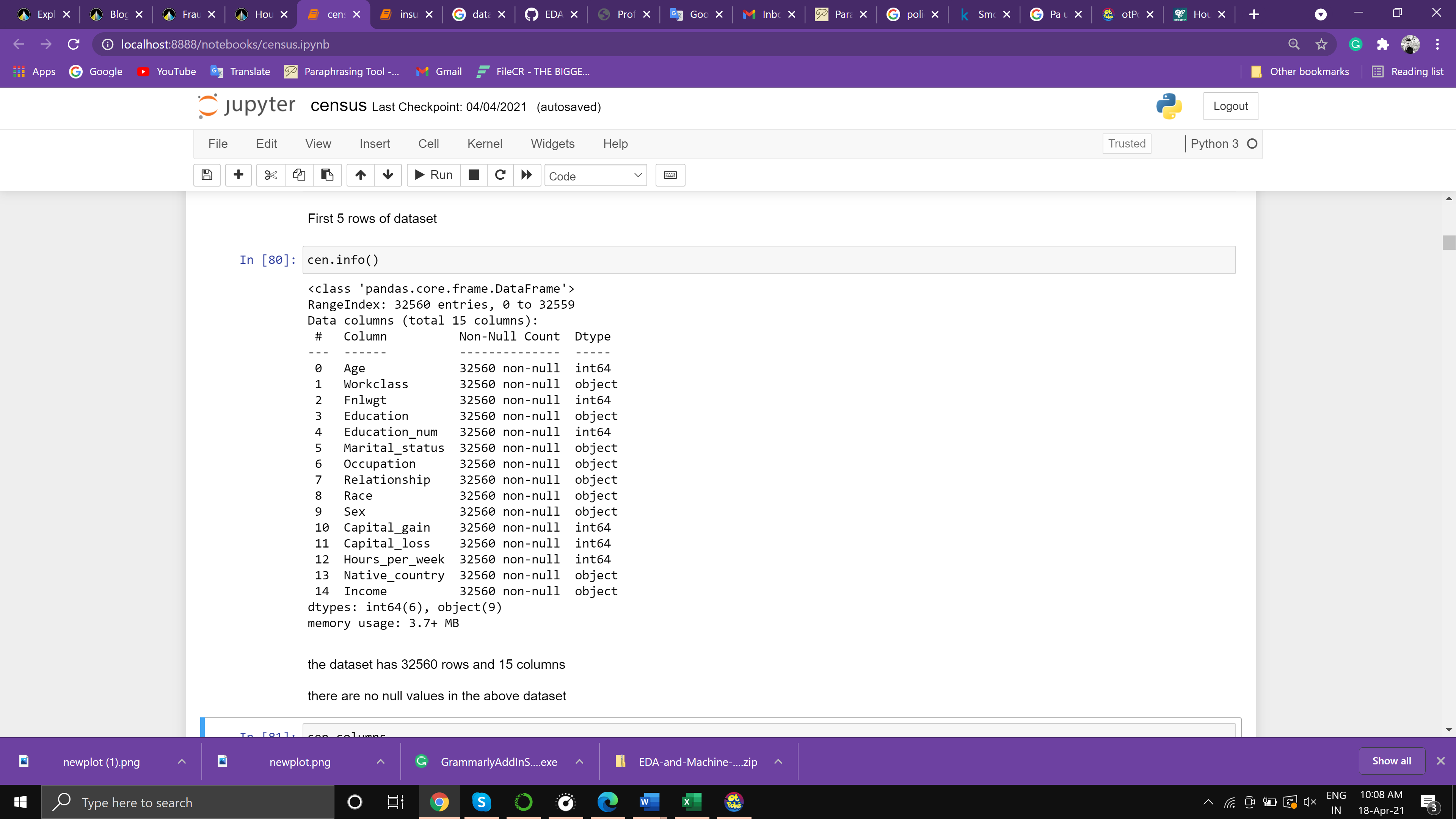
The entire process of machine learning can be divided into 4 main steps to get the desired prediction.

* Getting Data – Data can be collected from the source in this step for exploratory data analysis and visualization for understanding the current/historic data and determine the next step.
* Data Pre-Processing – Pre-Processing aka data wrangling is the technique of cleaning and transforming the raw data (which can be incomplete and inconsistent) to a proper format to be used for modeling.
* Model Evaluation – Different machine learning algorithms can be used and evaluated in this step to measure the accuracy and other confusion metrics.
* Prediction – Depending on the choice of the best model, prediction is done.

**Exploratory data analysis**

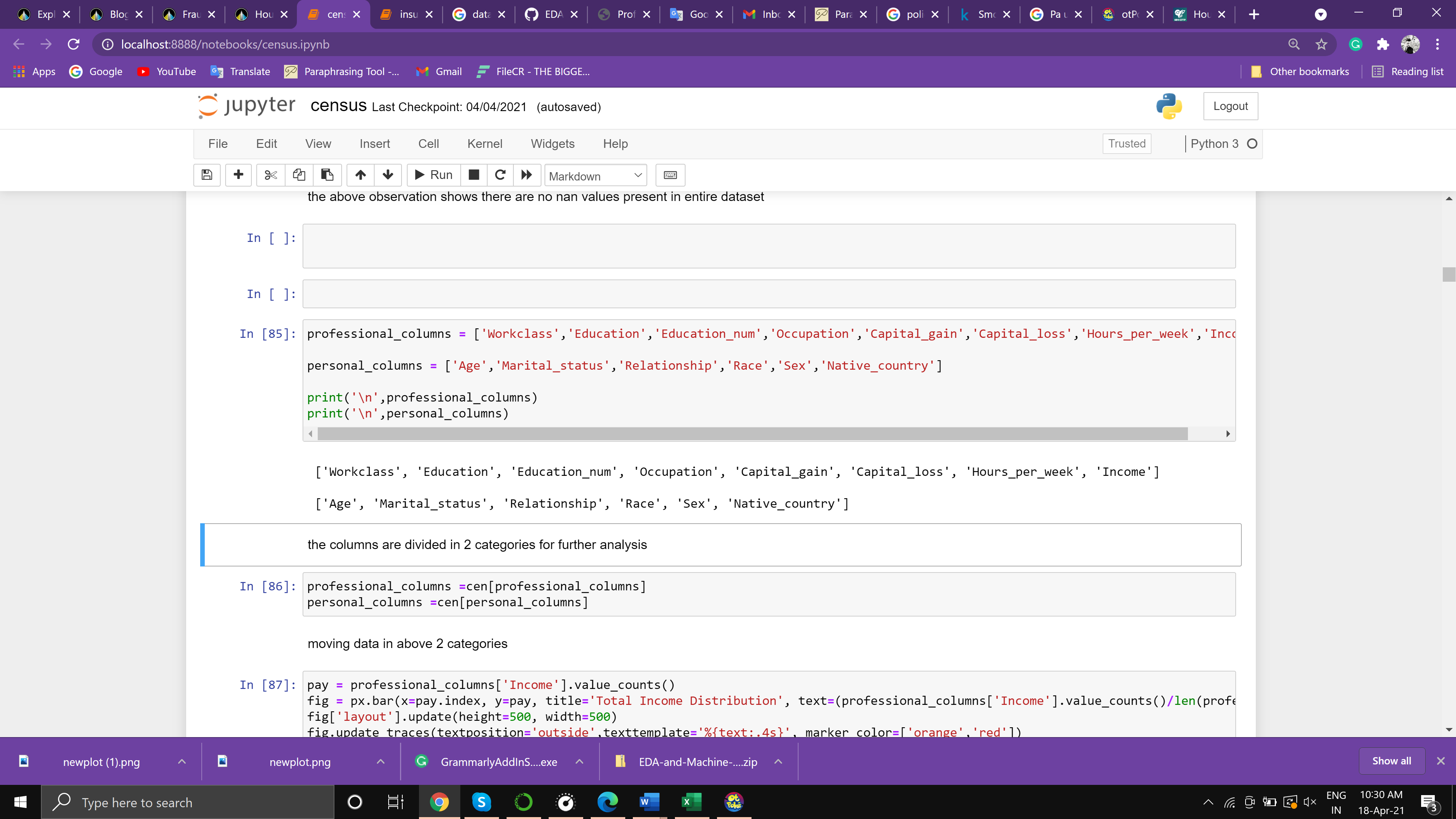
Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations.

As the data has been imported, the information about the data has been checked



There are no null values in the dataset, there are int and object values present in the dataset.

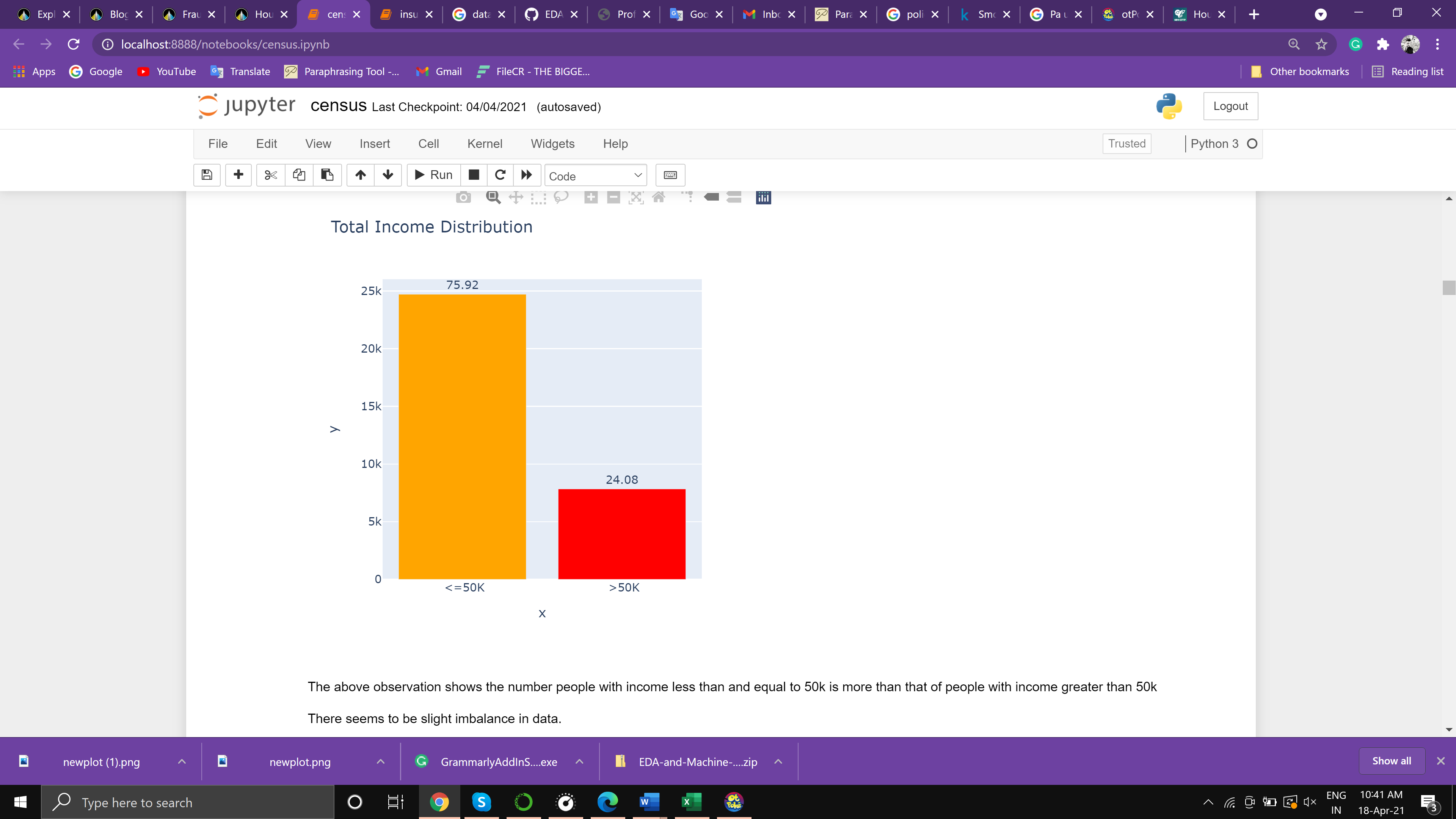
Dropping fnlwgt columns as it's not important for future analysis. also, replacing ‘?‘ values with the unknown in the dataset



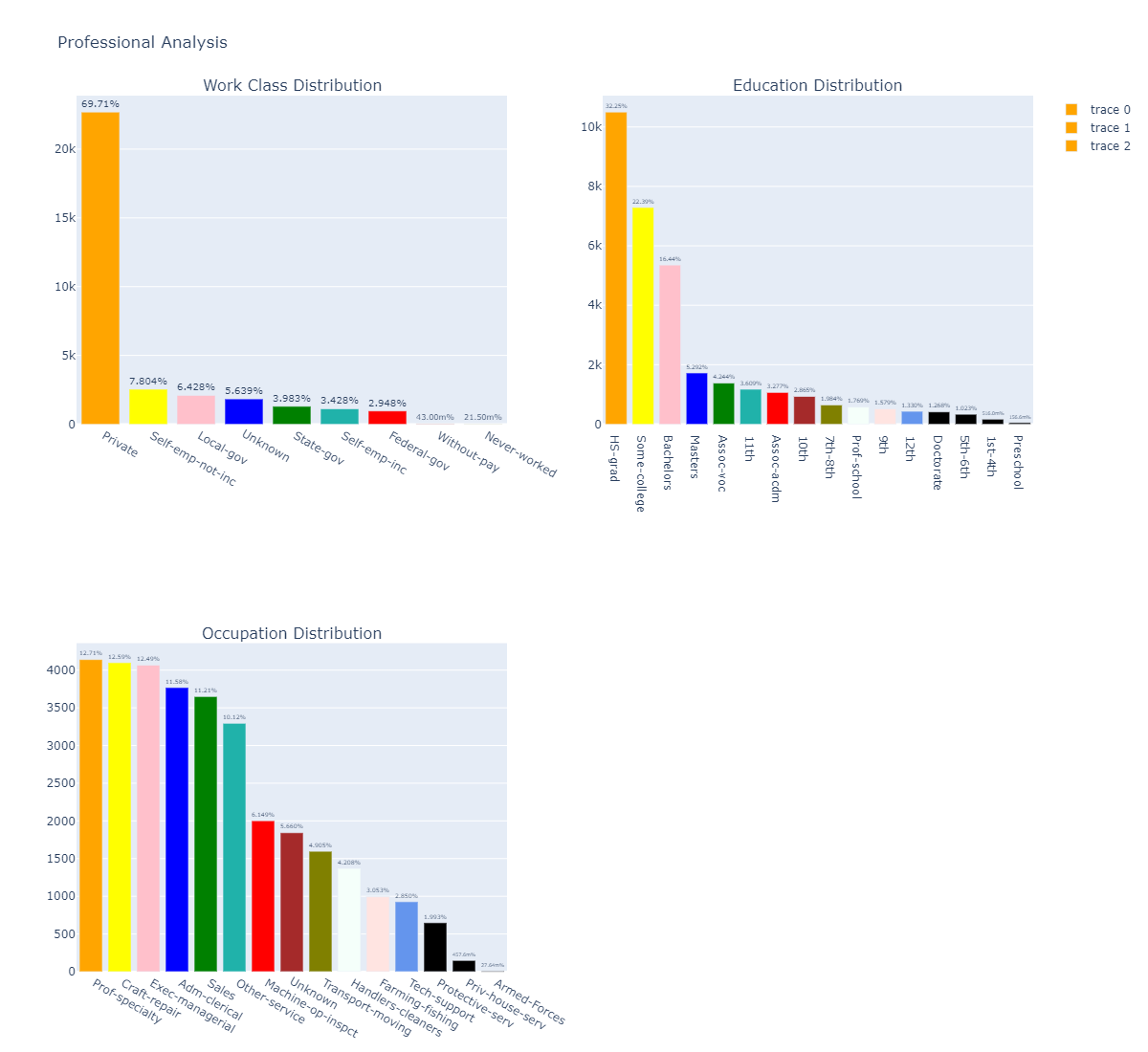
For further EDA the data has been divided into 2 categories

Professional column: Workclass, Education, Education\_num, Occupation, Capital\_gain, Capital\_loss, Hours\_per\_week, Income

Personal column: Age, Marital\_status, Relationship, Race, Sex, Native\_country



The above figure shows the income distribution, the number of people with income less than and equal to 50k is more than that of people with income greater than 50k

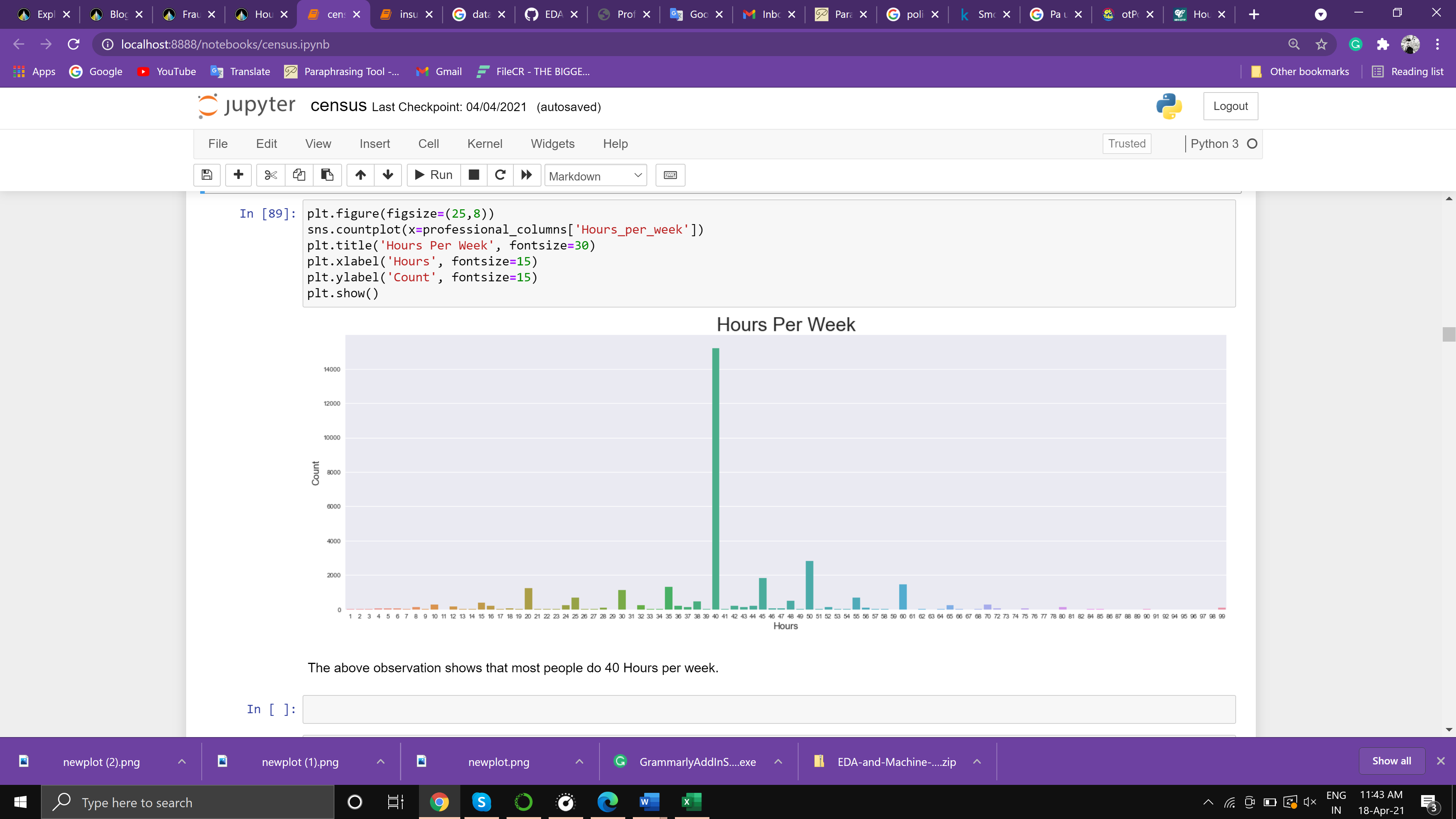


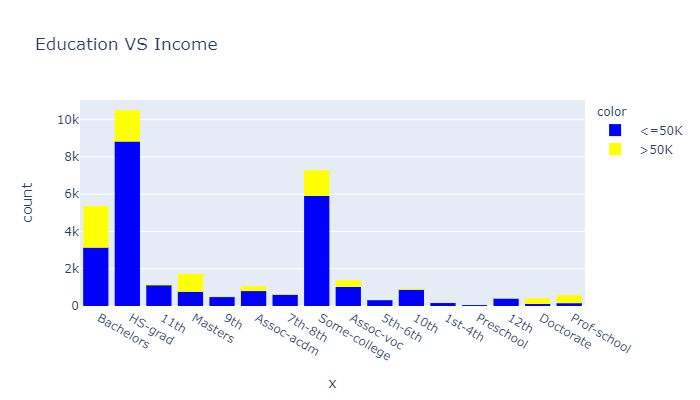
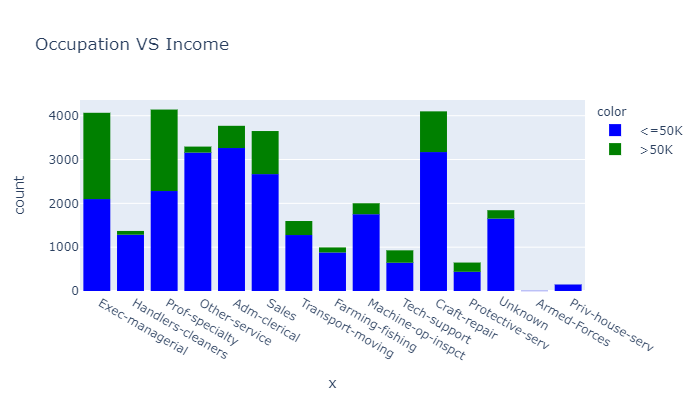
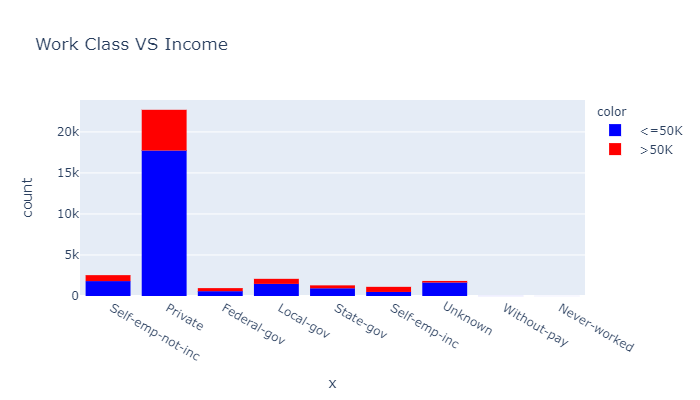
The above observation shows the percentage distribution of professional columns like Workclass, education, and occupation column

According to the observation around 69.7% of people are working privately.

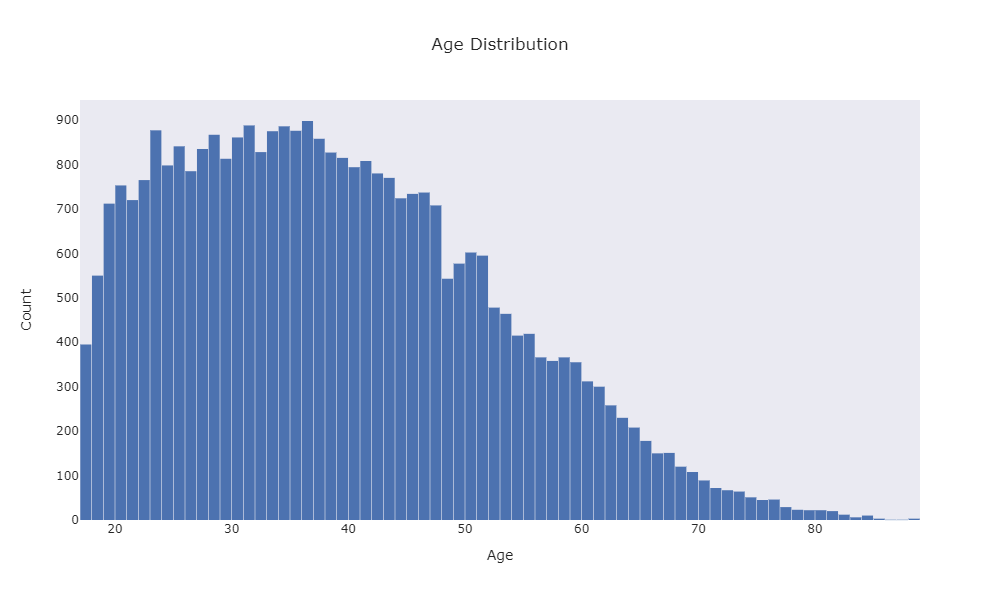
Around 32.2% were high school graduates and 22.39% are college graduates.

Prof-specialty, Craft-repair, and Exec-management occupations are higher in count than other occupations.



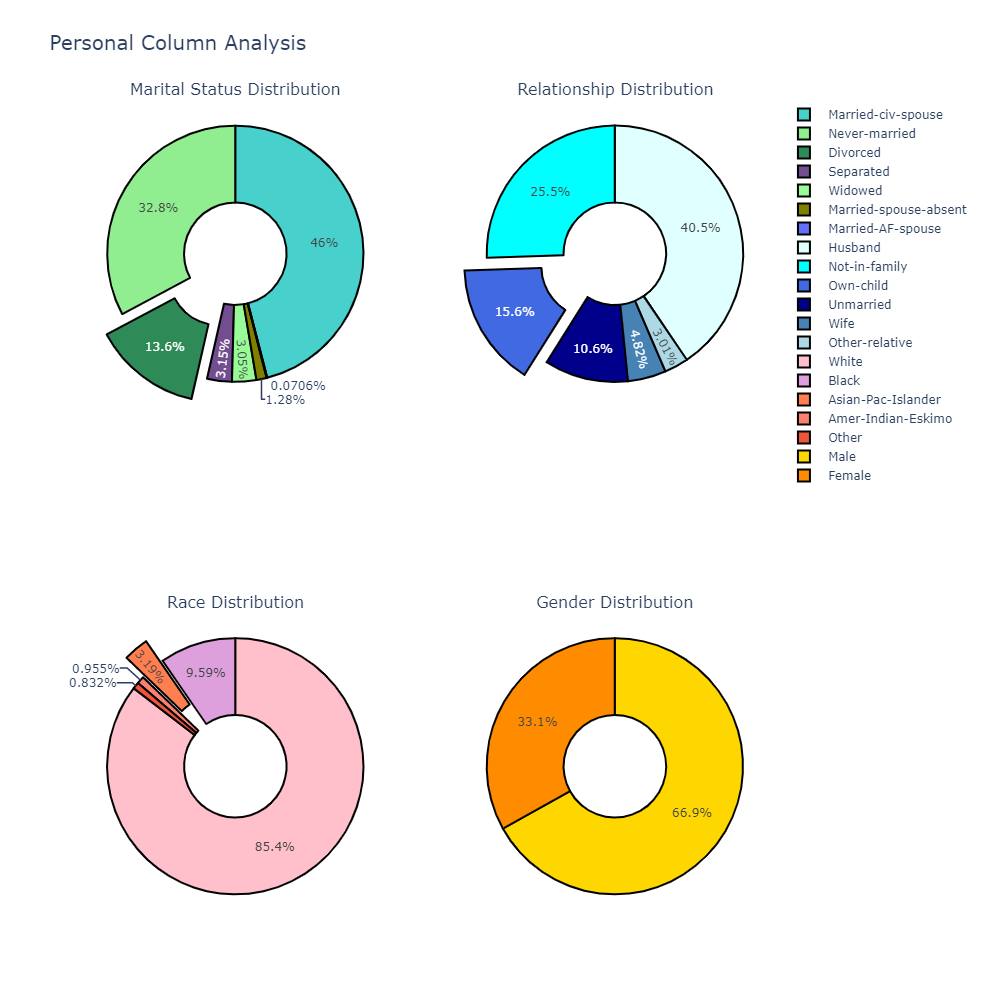
The above figure shows the bar graph of hours per week by a person, and it's been observed that the maximum number of people work 40 hours a week

The above 3 figures show the relation of income column with WorkClass, Education, and Occupation

People doing private jobs have a higher rate of earning >=50k as well as <50k. Also, the rate of earning >=50k is higher in Exec-managerial, Prof-specialty occupation. Bachelors degree holders have a higher chance of earning >=50k. Masters, Doctorate degrees have a lower total count but their rate of earning >=50k is a lot higher.

The above plot shows the age distribution of the working class

There is a lot of variation in age while as per the above observation, the max working age group is between 20 to 42.

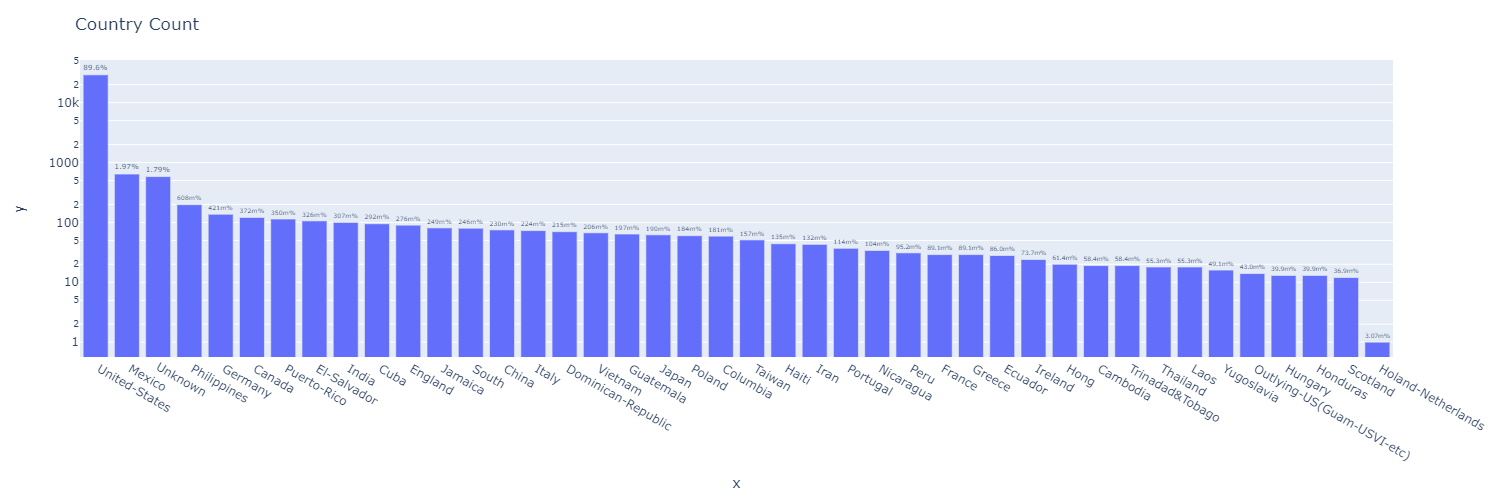


The above figure shows the percentage distribution of personal columns Marital status, Relationship, Race, Sex

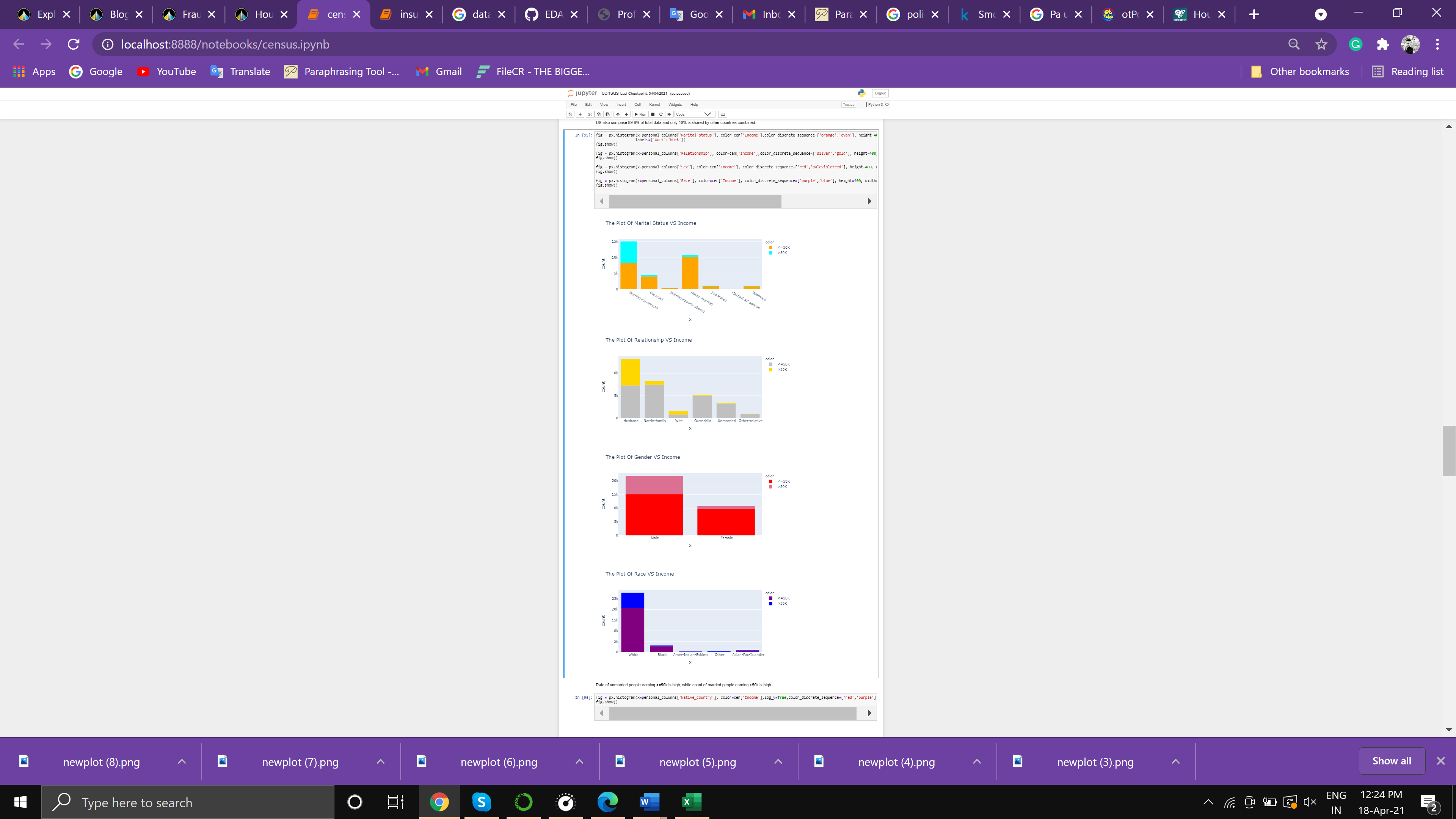
46% of people are married and 32.8% are unmarried.

The count of the working husband is a lot higher than the wife.

The count of the White race is higher than other races also Male count is double that of female.



The above figure shows that the US also comprises 89.6% of total data and only 10% is shared by other countries combined.



The rate of unmarried people earning >=50k is high,

While the count of the husband earning >50k is high.

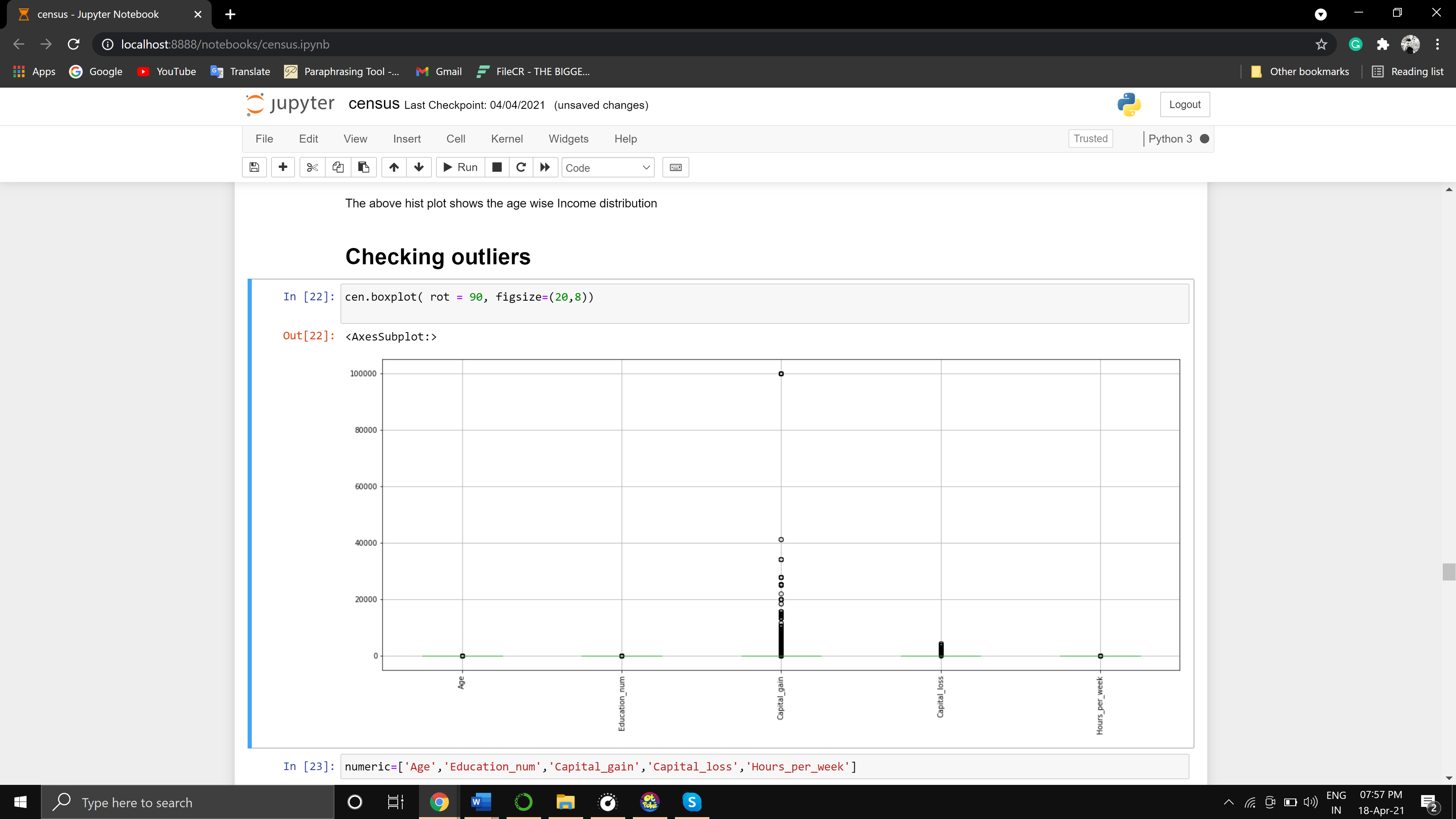
Male count of earning is more than female count

The white race dominates all other races in income

**Outlier Analysis**

Outlier Analysis is a process that involves identifying the anomalous observation in the dataset.” ... In some instances like fraud detection, the outlier indicates a fraudulent activity. Outlier Analysis is a data mining task which is referred to as an “outlier mining”.

After performing EDA we are checking for outliers and the maximum number of outliers were present in the Capital\_gain column as we can see in the figure below.



The outliers and detected and removed using the Z-score method which caused deletion of some rows from the dataset

* shape before removing outliers (32560, 14)
* shape after removing outliers (30150, 14)

also, skewness was present in the following columns.

* Capital\_gain 4.940872
* Capital\_loss 29.821934

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or [normal distribution](https://www.investopedia.com/terms/n/normaldistribution.asp), in a set of data. If the curve is shifted to the left or the right, it is said to be skewed. Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution. A normal distribution has a skew of zero, while a [lognormal distribution](https://www.investopedia.com/articles/investing/102014/lognormal-and-normal-distribution.asp),

The skewness was removed using the power transform method to bring the data into a normalized curve.

After a bit of data cleaning the categorical/ object columns were converted to numeric form. label encoder was used to convert the categorical columns into numeric form, which can be used to perform ML further.

Below columns were label encoded

* Workclass
* Education
* Occupation
* Race
* Sex
* Native\_country
* Marital\_status
* Relationship
* Income

Scaling Dataset

scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step

Before moving towards modeling, we have divided data in x which includes all variables, and in y which include label/target

Also, we scaled the data using the standard scalar method before splitting the data

That data has been split into x train, y train, x test, y test for fitting data into the model and performing ML

Model Evaluation

Model selection and evaluation is a critical step of any machine learning project as identifying the pattern and applying the correct algorithm is not a very easy process. Machine Learning provides multiple models to generalize it to the unseen data from the same population and measuring the performance. Along with meeting the business objective, the model should take care of accuracy, execution time, complexity, and scalability as well to be considered as the best model. Sometimes the size of the training data set and numbers of predictor features can be decision-making criteria for model selection. Cross-Validation Score is the performance metric for this project.

As this is a classification project, classifier algorithms can be used to predict results. we are using for loop to run 4 different algorithms i.e

* AdaBoost
* GradientBoosting
* Random forest
* Logistic Regression

The performance will be measured based on a cross-validation score

AdaBoost:

Training Score: 0.842910

Testing Score: 0.840299

Cross val score is : 0.8370019011771936

The difference between accuracy score and cross-validation score is: 0.003296606285492887

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

GradientBoosting:

Training Score: 0.853980

Testing Score: 0.840796

Cross Val score is: 0.8383240834981661

The difference between accuracy score and cross-validation score is: 0.002471936402331343

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

Random forest:

Training Score: 0.977363

Testing Score: 0.831012

Cross Val score is: 0.8232231674764273

The difference between accuracy score and cross-validation score is: 0.0077884411471216275

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

Logistic Regression:

Training Score: 0.804809

Testing Score: 0.801658

Cross Val score is: 0.8011531887926566

The difference between accuracy score and cross-validation score is: 0.0005051860000465647

As we can observe above, that 4 four different algorithms are been used out of which GradientBoost performs better than all

Cross Val score is: 0.8383240834981661

As Ada boost performs slightly less than the Gradient boost so we consider gradient boost for hyper tuning.

Hyperparameter tuning

In machine learning, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process.

It was observed performance of the Gradient boost has been slightly lowered after tuning. so we decided to save the model without tuning.

Saving model

Model progress can be saved during and after training. This means a model can resume where it left off and avoid long training times. Saving also means you can share your model and others can recreate your work.

The model is been saved with Object of Gradient boost after fitting with train data, to reuse it in further prediction.

Conclusion

The whole project was classification-based, there are numerical as well as classification data present in this dataset. There are no null values present here but some unknown

data is present which was been treated later for ease of cleaning and EDA the data was divided into 2 categories i.e numerical and categorical

Numerical: the numerical data was being checked for outliers and skewness to perform better and accurately

Categorical data: that was been modified with values which were unknown and After performing EDA data has been converted into numerical data to perform. After we analyzed the dataset, we were able to conclude. The number of people with income less than and equal to 50k is more than that of people with income greater than 50k,69.7% of people are working privately. Around 32.2% were high school graduates and 22.39% are college graduates. Prof-specialty, Craft-repair, and Exec-management occupations are higher in count than other occupations. The maximum number of people work 40 hours a week. Private jobs have a higher rate of earning >=50k as well as <50k. Also, the rate of earning >=50k is higher in Exec-managerial, Prof-specialty occupation. Bachelors degree holders have a higher chance of earning >=50k. Masters, Doctorate degrees have a lower total count but their rate of earning >=50k is a lot higher. The max working age group is between 20 to 42,46% of people are married and 32.8% are unmarried. The count of the working husband is a lot higher than the wife. The count of the White race is higher than other races also Male count is double that of female.US also comprises 89.6% of the total data and only 10% is shared by other countries. The rate of unmarried people earning >=50k is high, While the count of the husband earning >50k is high. The male count of earning is more than the female count. The white race dominates all other races in income. The Gradient boost model has the potential to perfectly predict the Income