Census Income Project

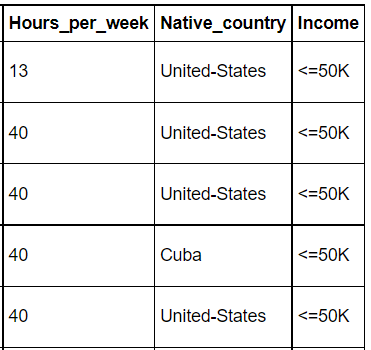
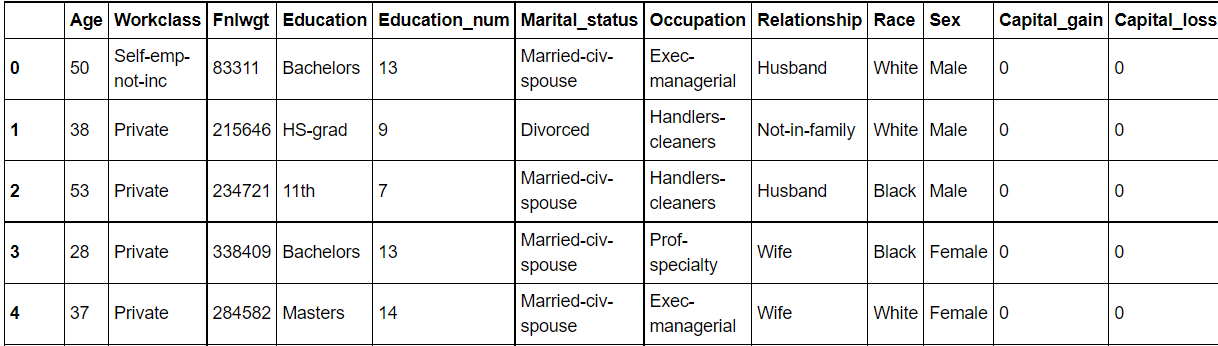
## Problem Definition

This data was extracted from the 1994 Census bureau database. A set of reasonably clean records was extracted. *The prediction task is to determine whether a person makes over $50K a year*.

## Data Analysis

### Data Import:

* The dataset is in csv format, we shall import the dataset using ‘read\_csv’ function.
* Once the dataset is imported and converted into a data frame, store the data frame and print it, to analyse the data points in rows and columns.

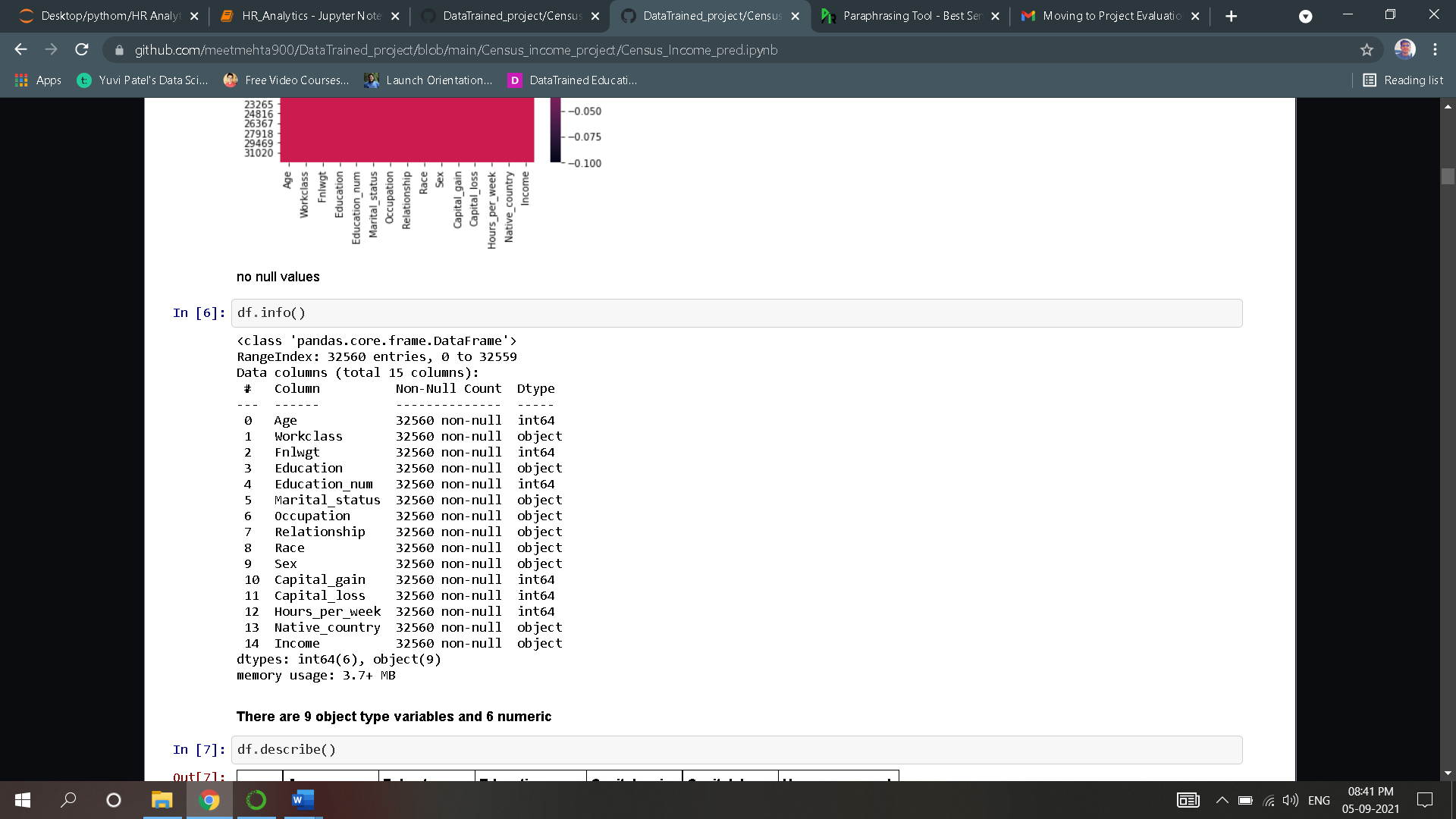


* Unfortunately, due to pandas limitation not all rows and columns are displayed.
* Our project’s objective is to predict Income category of an individual using their personal information provided, hence column “Income” is our target variable and our input features are Age, Workclass, Fnlwgt, Education, Education\_num, Marital\_status, Occupation, Relationship, Race, Sex, Capital\_gain, Capital\_loss, Hours\_per\_week, Native\_country, Income.

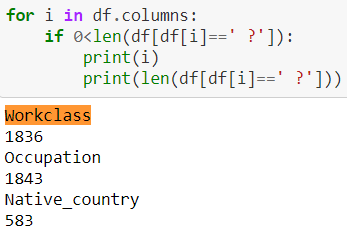
|  |  |  |
| --- | --- | --- |
| **Column Name** | **Type** | **Description** |
| age | Continuous | The age of the individual |
| workclass | Categorical | The type of employer the individual has (government, military, private, etc.). |
| fnlwgt | Continuous | The number of people the census takers believe that observation represents (sample weight). This variable will not be used. |
| education | Categorical | The highest level of education achieved for that individual. |
| education\_num | Continuous | The highest level of education in numerical form. |
| marital\_status | Categorical | Marital status of the individual. |
| occupation | Categorical | The occupation of the individual. |
| relationship | Categorical | Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. |
| race | Categorical | White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black. |
| gender | Categorical | Female, Male. |
| capital\_gain | Continuous | Capital gains recorded. |
| capital\_loss | Continuous | Capital Losses recorded. |
| hours\_per\_week | Continuous | Hours worked per week. |
| native\_country | Categorical | Country of origin of the individual. |
| income | Categorical | ">50K" or "<=50K", meaning whether the person makes more than \$50,000 annually. |

2.2 Data Analysis:

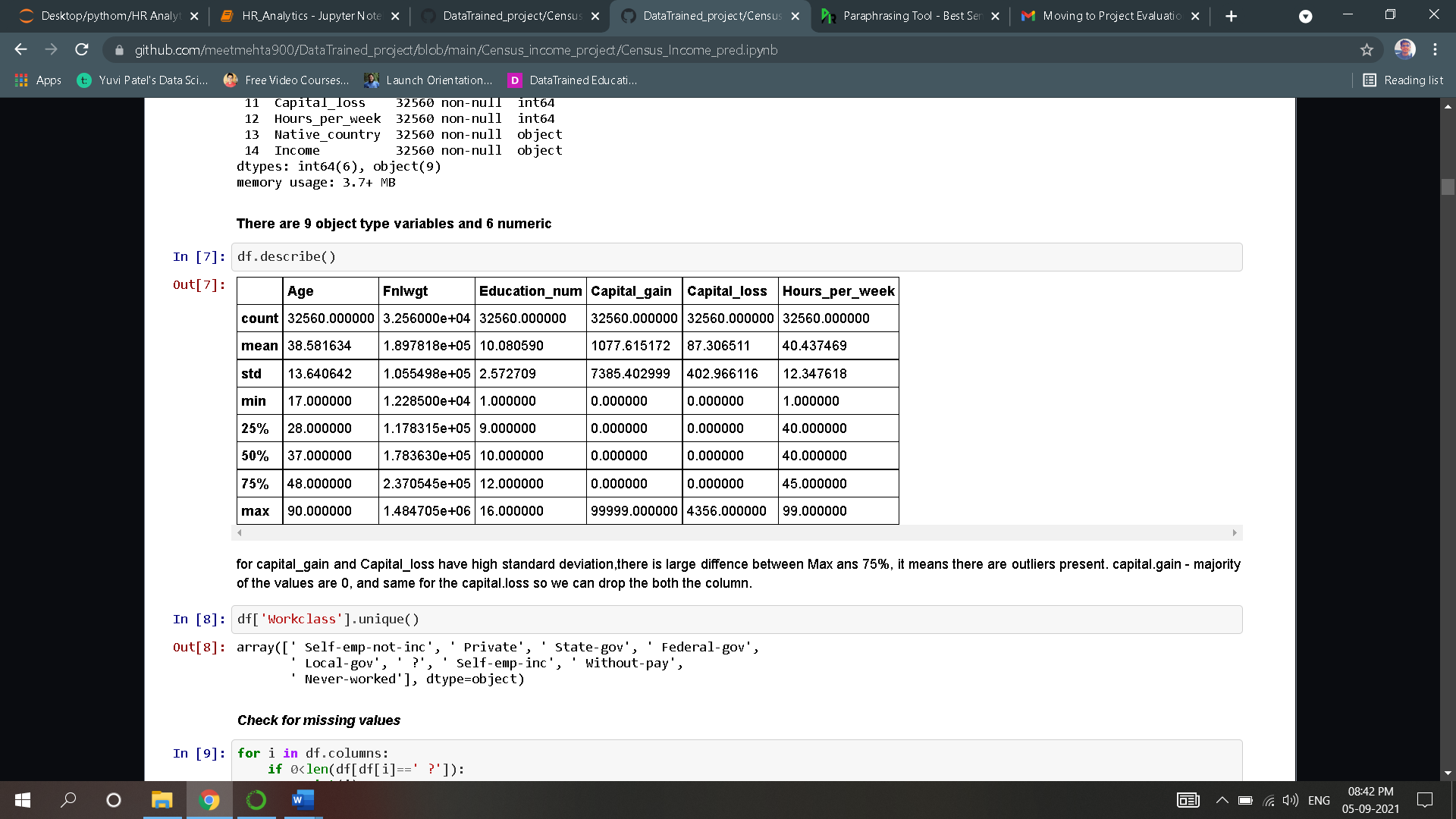
* With the help of pandas inbuilt functions, we can deduce following key points:
  1. Number of rows in the dataset is 32560.
  2. Number of columns in the dataset is 15.



* After executing the ‘info’ command we obtain the above output and following is our observation:
* Dataset contains any null values: False
* Check if the dataset has null values in the form of whitespaces or in any other form.
* Null values were found in the form of ‘?’.

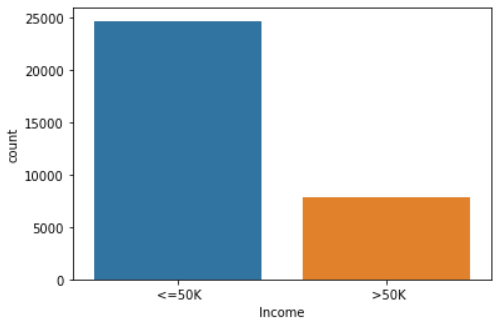
****

* From above we can deduce that columns such as workclass, occupation & native\_country contain null values.
* It is important that we fill these fill these null values, for that first we need to replace these rows containing ‘?’ with null values.
* As it is a categorical value we need to fill these null values w.r.t the mode value of that specific column.
* In this manner we successfully filled all null values.
* For further analyses of the dataset, we require a Statistical Summary of each column such as mean value, median value, max value, min value, standard deviation value of non-categorical data.
* Describing the function of pandas can provide us with a statistical summary along with the count of non-null rows, lower percentile, upper percentiles.
* In statistics, a percentile is a score *at or below which* a given percentage falls.
* For example, the 50th percentile is the score below which 50% of the scores in the distribution may be found.



### 2.3 Univariate Analysis:

* To perform univariate analysis, we need to check the following:
  + Class imbalancement
  + Count Plot of every categorical column.
  + Boxen plot of non-categorical column.
* Check for class imbalancement with the help of countplot of target variable “Income”.



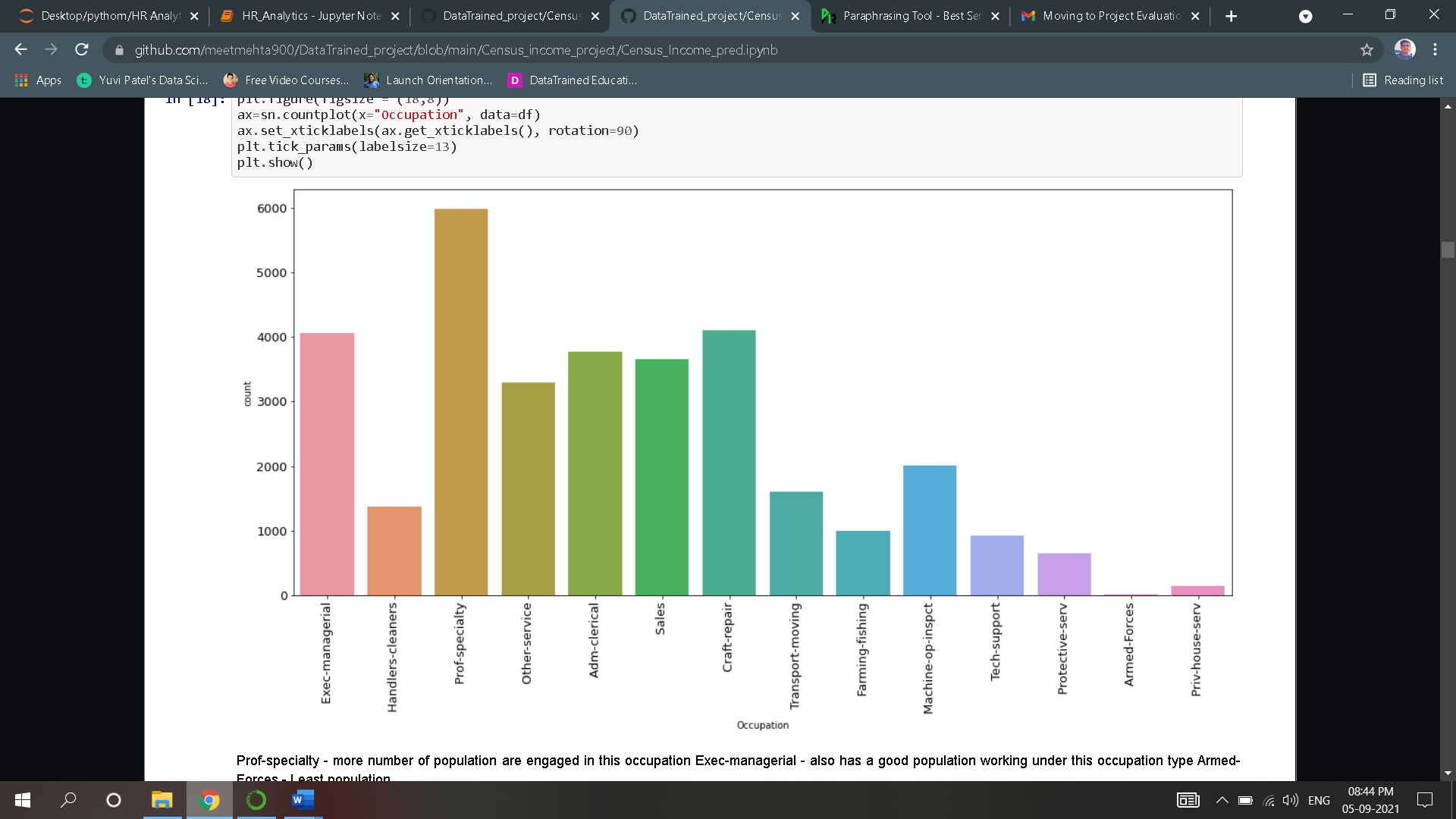
#### Observation:

* The highest number of the population are earning less than 50k "<=50K".
* Very few individuals are earning ">50K".
* Our label column is Imbalanced and we would need to balance it by using "Under sampling/Over Sampling method".
* Now plot Boxen plot of column Age.



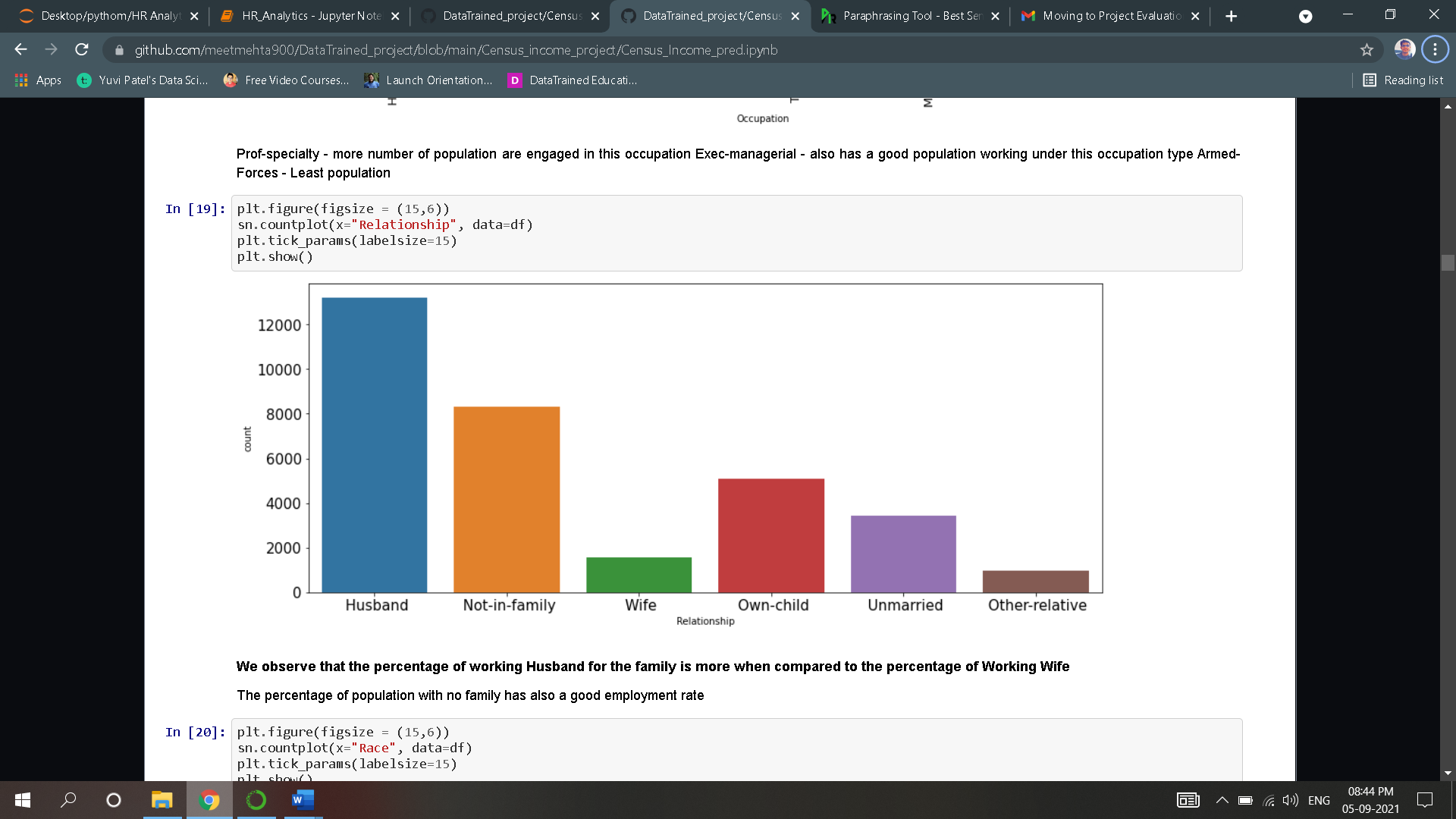
#### Observation:[¶](https://render.githubusercontent.com/view/ipynb?color_mode=auto&commit=d8718b63e1e962d94dfe5400c333f17ae45355c7&enc_url=68747470733a2f2f7261772e67697468756275736572636f6e74656e742e636f6d2f506f6f6e616d52616a70757431362f44617461547261696e65645f4576616c756174696f6e5f50726f6a656374732f643837313862363365316539363264393464666535343030633333336631376165343533353563372f43656e737573253230496e636f6d6525323050726564696374696f6e732e6970796e62&nwo=PoonamRajput16%2FDataTrained_Evaluation_Projects&path=Census+Income+Predictions.ipynb&repository_id=390734498&repository_type=Repository#Observation:)

* + Highest number of the population works around the age range of 25 - 50.
  + After this age group, we see reduction in the population working.
* Count Plot for column occupation.



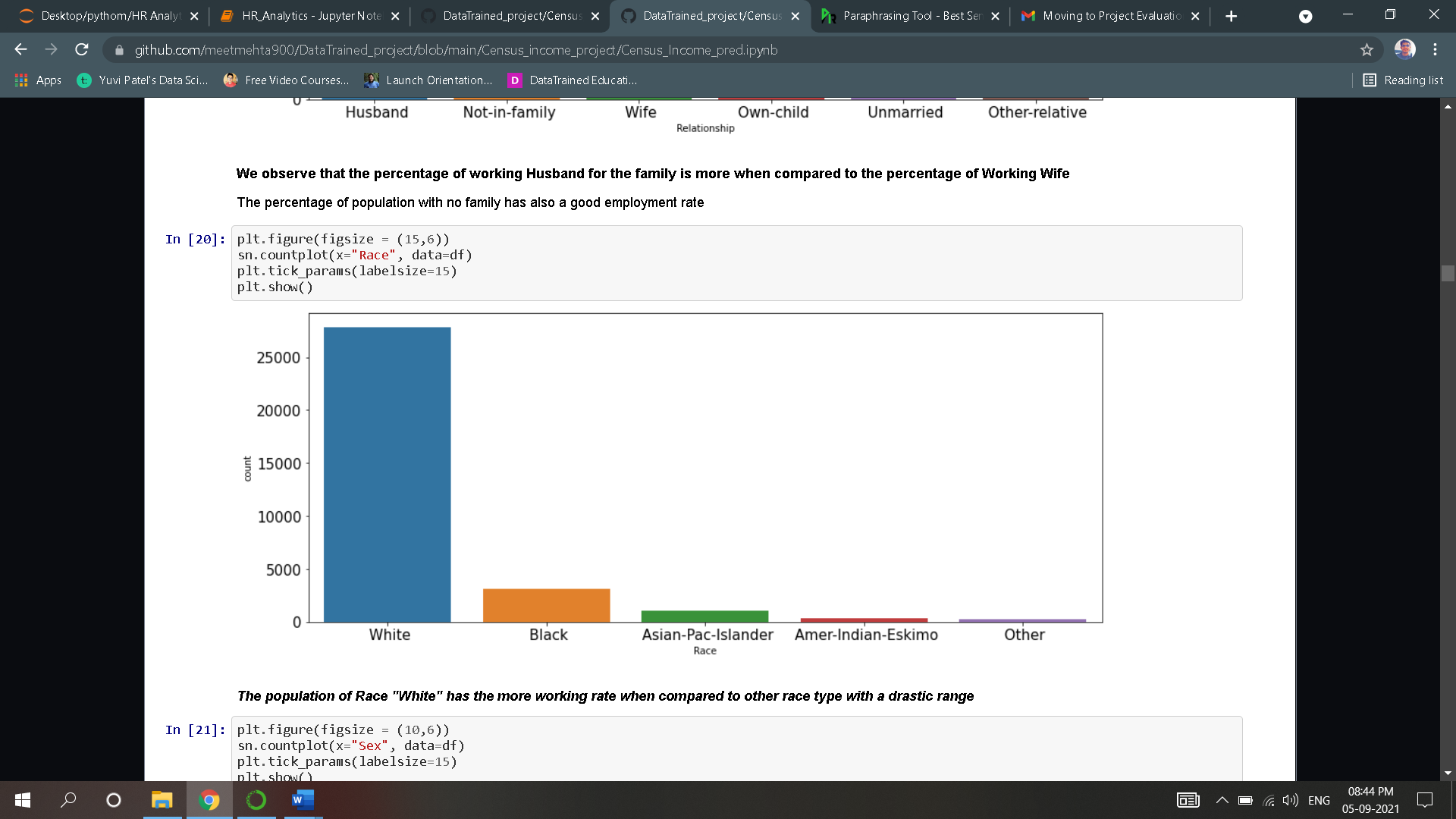
#### Observation:

* + Prof-specialty - more number of population are engaged in this occupation
  + Exec-managerial - also has a good population working under this occupation type
  + Armed-Forces - Least population
* Count plot for column Relationship type

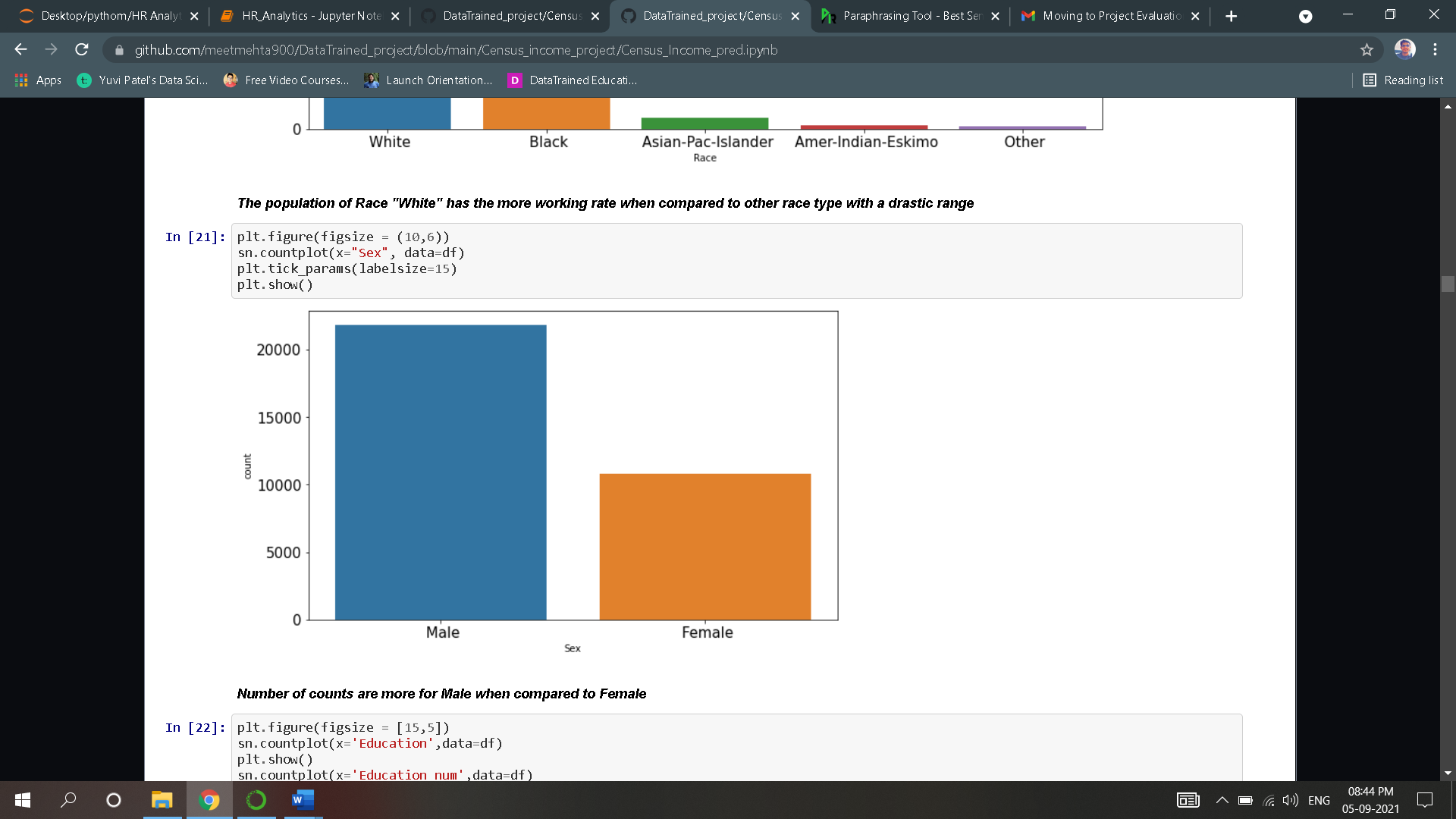


#### Observation:

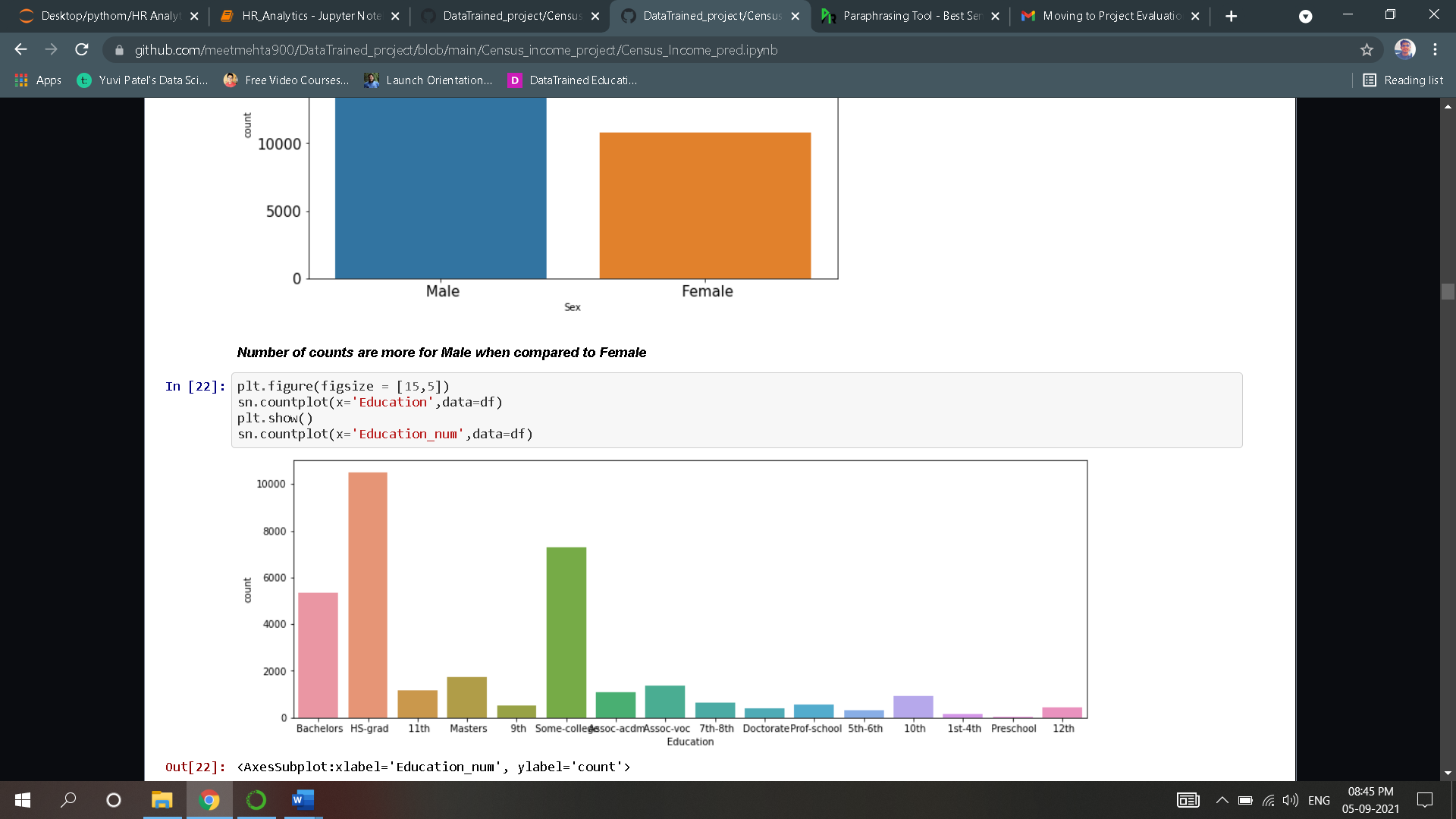
* We observe that the percentage of working Husband for the family is more when compared to the percentage of Working Wife
* The percentage of population with no family has also a good employment rate
* Count plot for column Race.

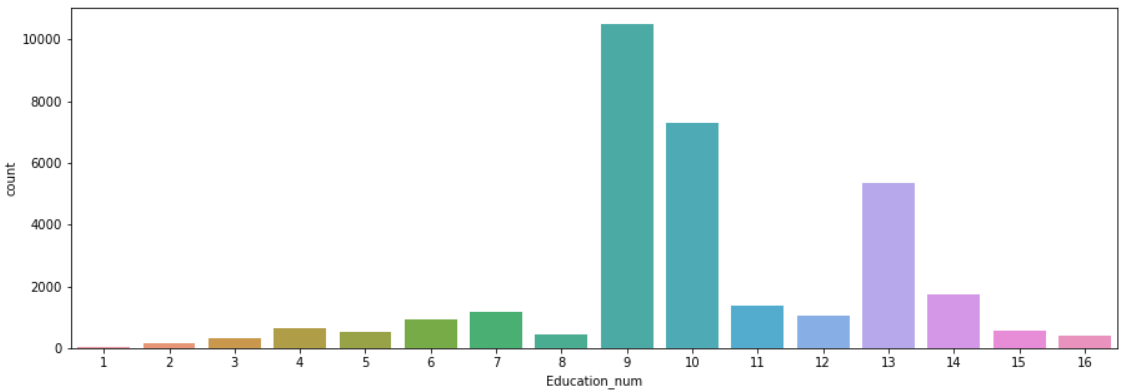


* Observation:
  + The population of Race "White" has the more working rate when compared to other race type with a drastic range
* Count plot for column Gender



* Observation:
* As per 'Sex ratio' - Number of counts are more for Female when compared to Female
* Count plot for column Education & Education\_Num

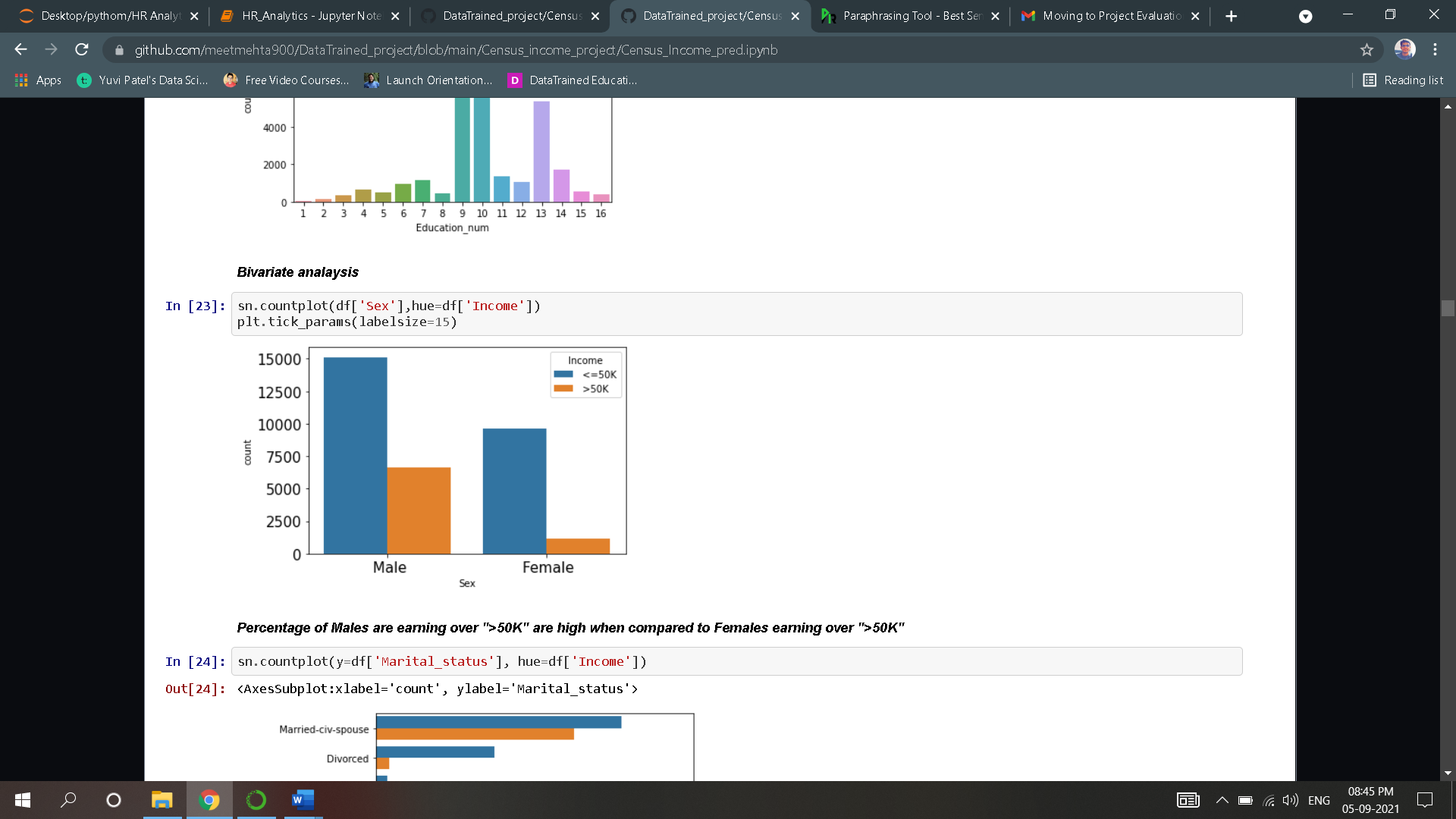


****

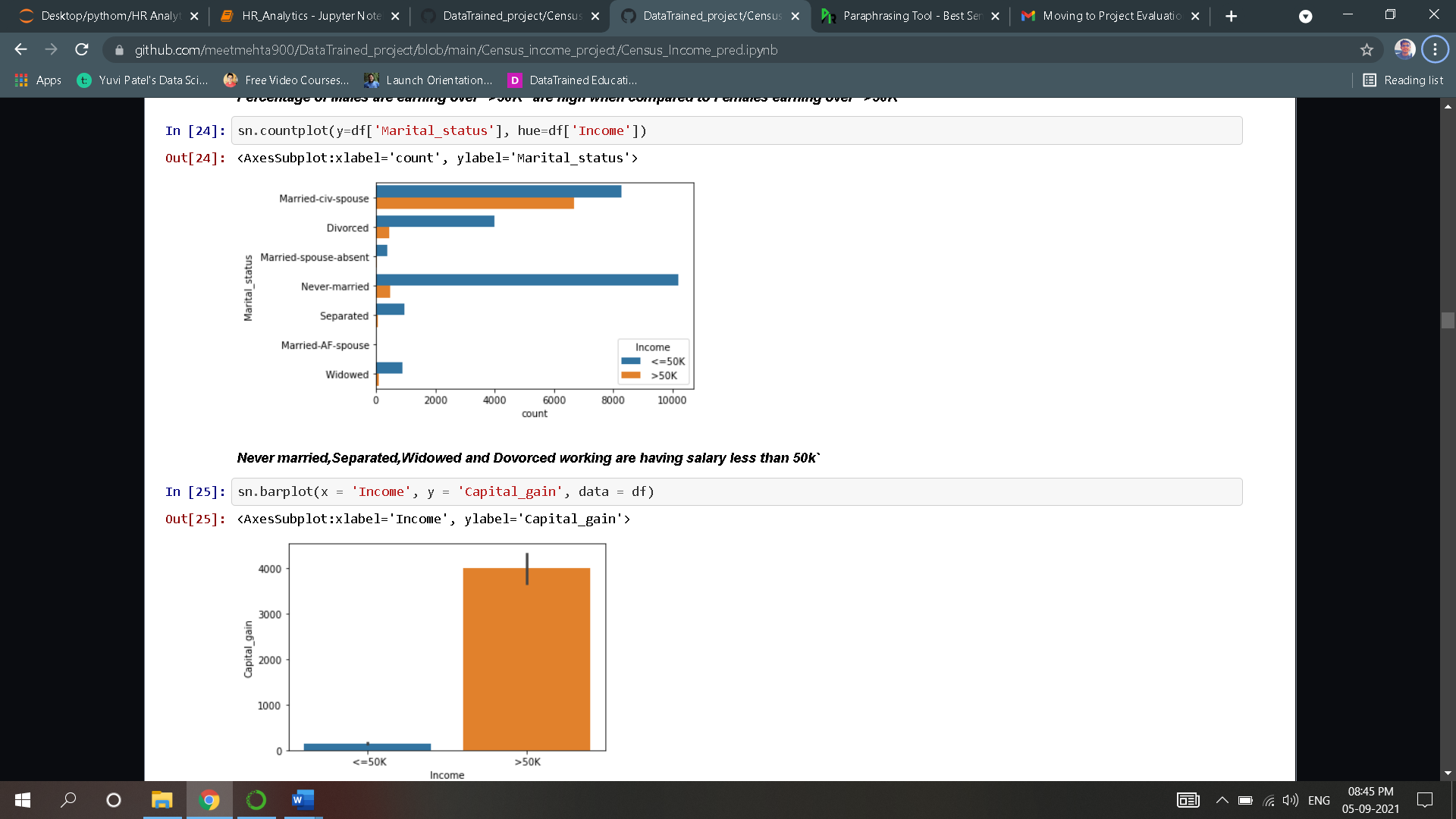
* Observation:
  + By comparing both the columns: "Education\_num" and "Education"
  + We see that data counts are one and the same for both of these column data
  + We would just need to name columns unique data according to their education grade and name the range "1 to 16"
  + Then, our data for both columns will look exactly same
  + Hence, we can drop any one column, let's drop the column: "Education\_num" later.

## 2.4 Bivariate Analysis:

* Count Plot of Sex w.r.t Income column.

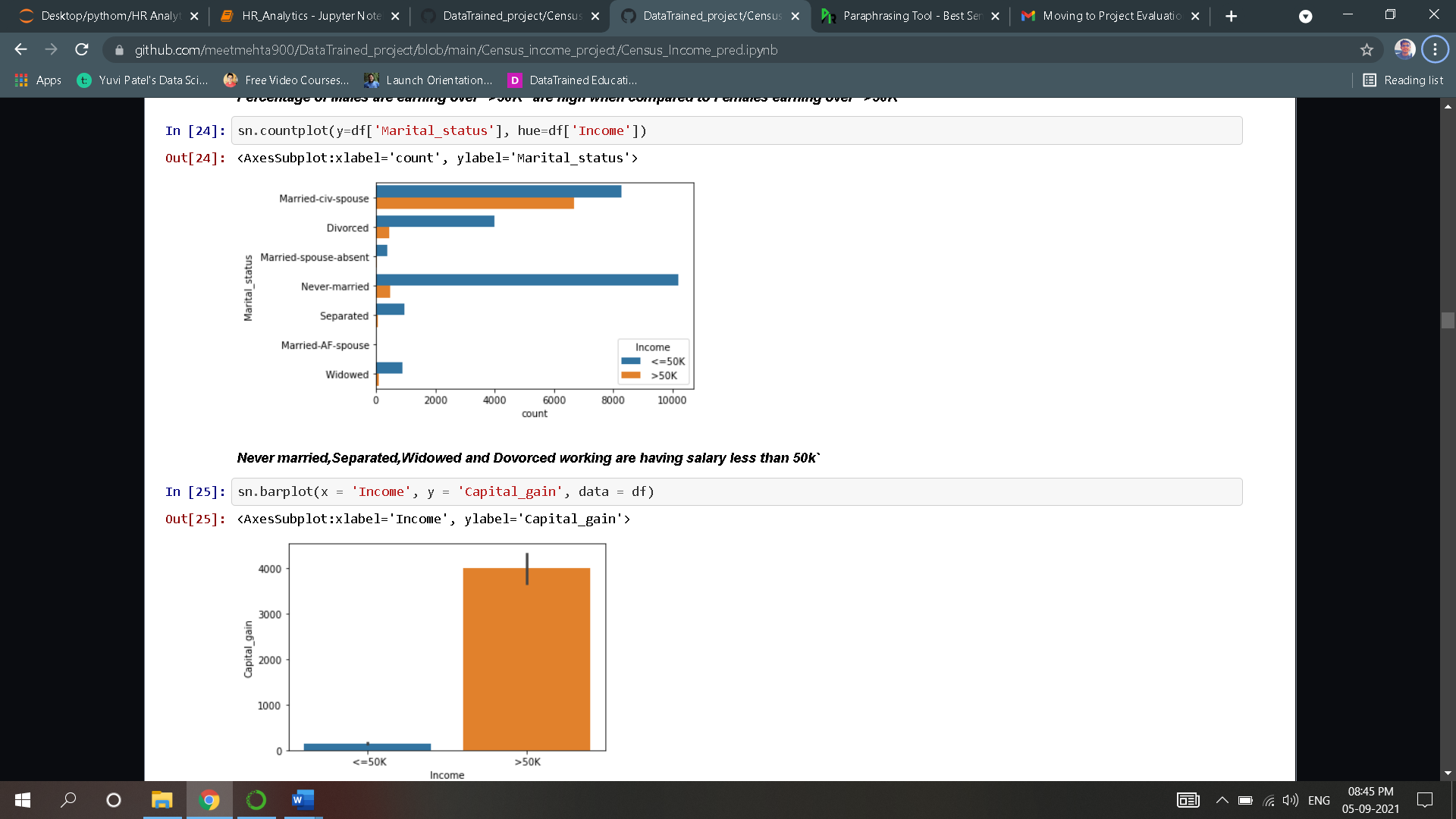


* Observation:
  + Percentage of Males are earning over ">50K" are high when compared to Females earning over ">50K"
* Count Plot of Marital\_Status w.r.t Income column

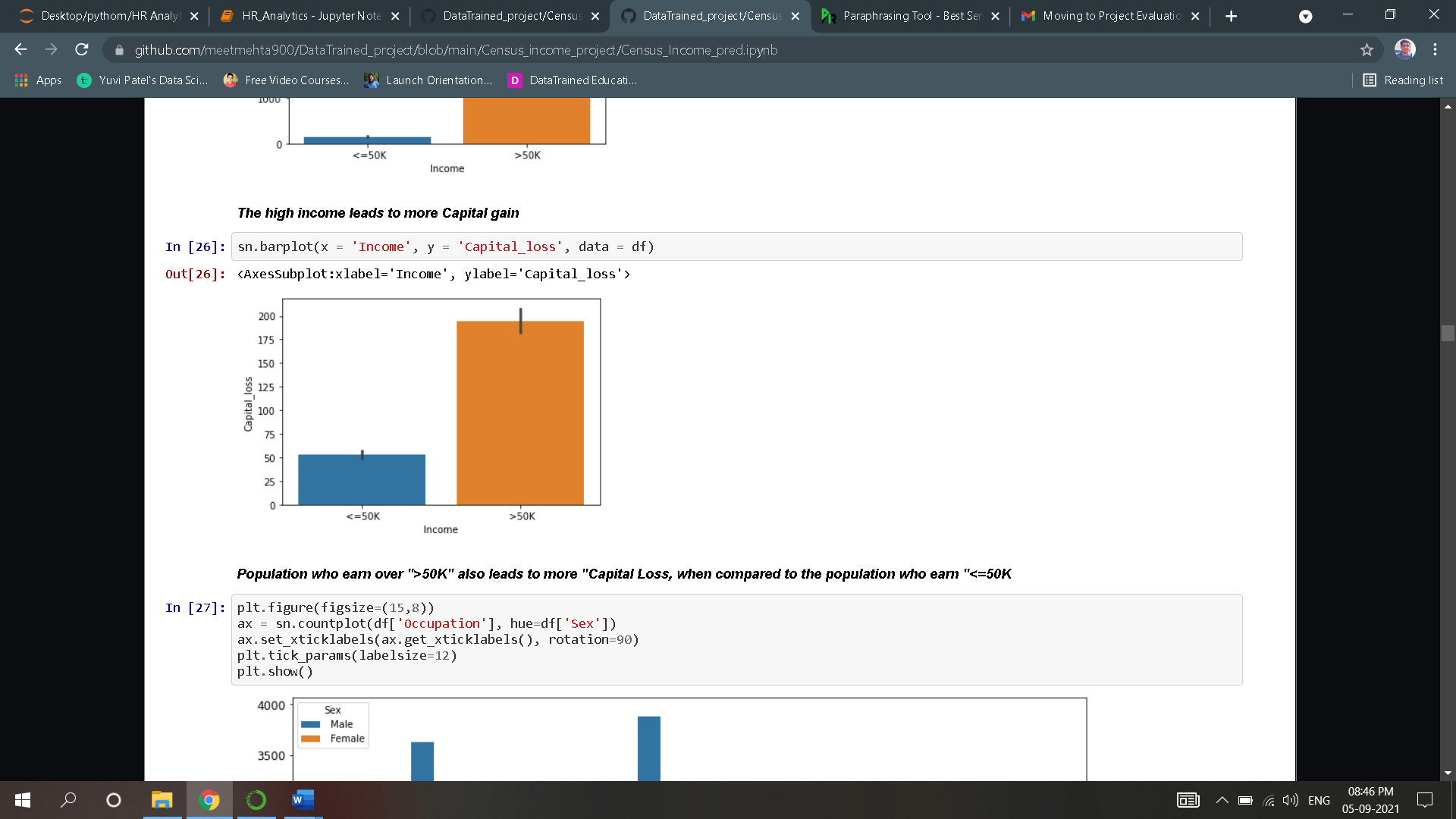


### Observation:

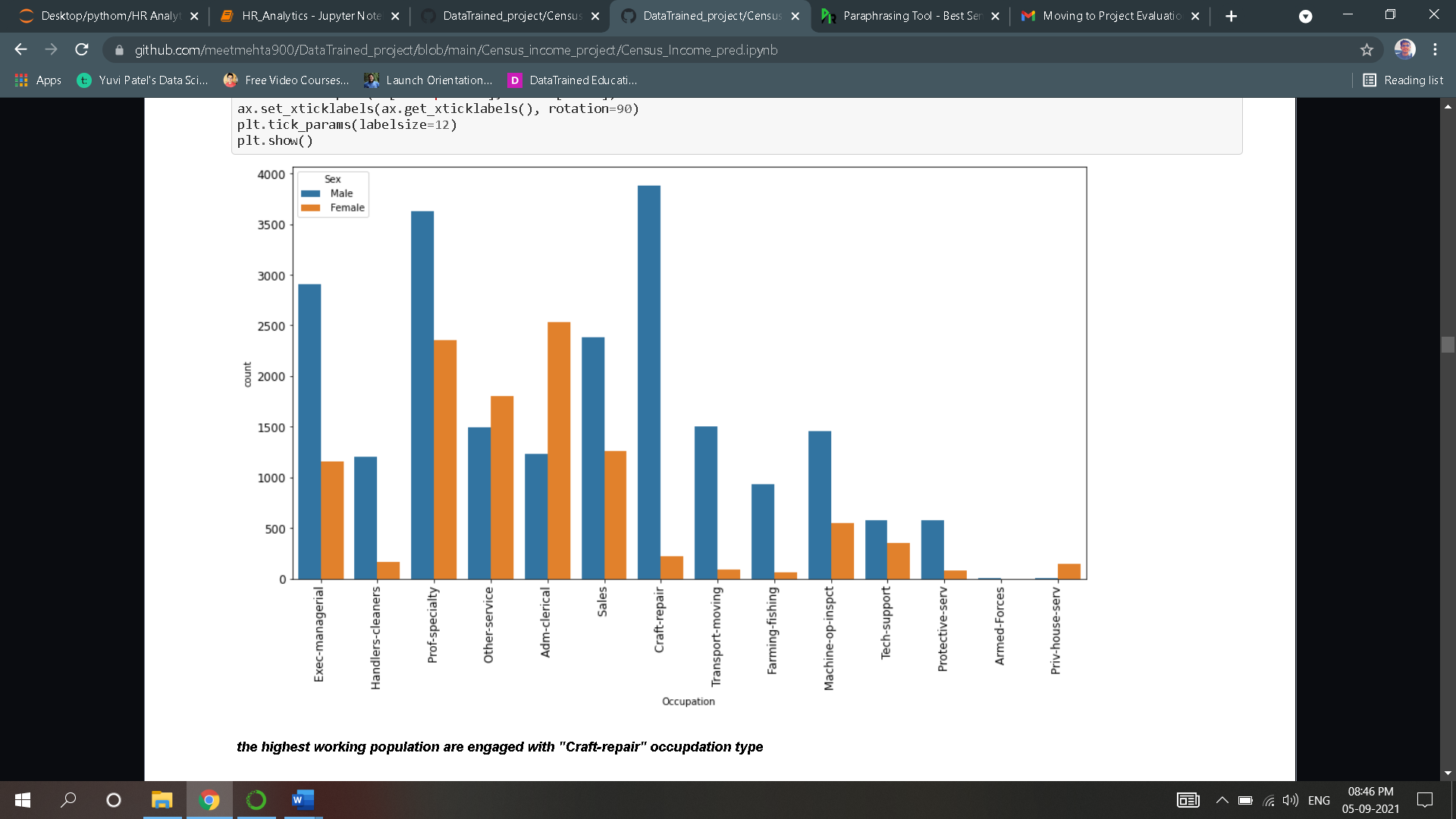
* + Married-civ-spouse - works with an income over >50K
  + Never-married" - more population who are not married, do with less income <=50 K
* Count plot of Column Capital\_gain w.r.t. target column Income.



* Observation:
  + Population who earn over 50K leads to more Capital gain, when compared to the population who earn less than 50K.
  + The higher the income the more is the Capital gain.
* Count plot of Column Capital\_loss w.r.t. target column Income.



* Observation:
  + Population who earn over 50K also leads to more Capital Loss, when compared to the population who earn less than 50K.
  + The high income leads to more Capital loss
* Count plot of Column occupation w.r.t. target column Income.



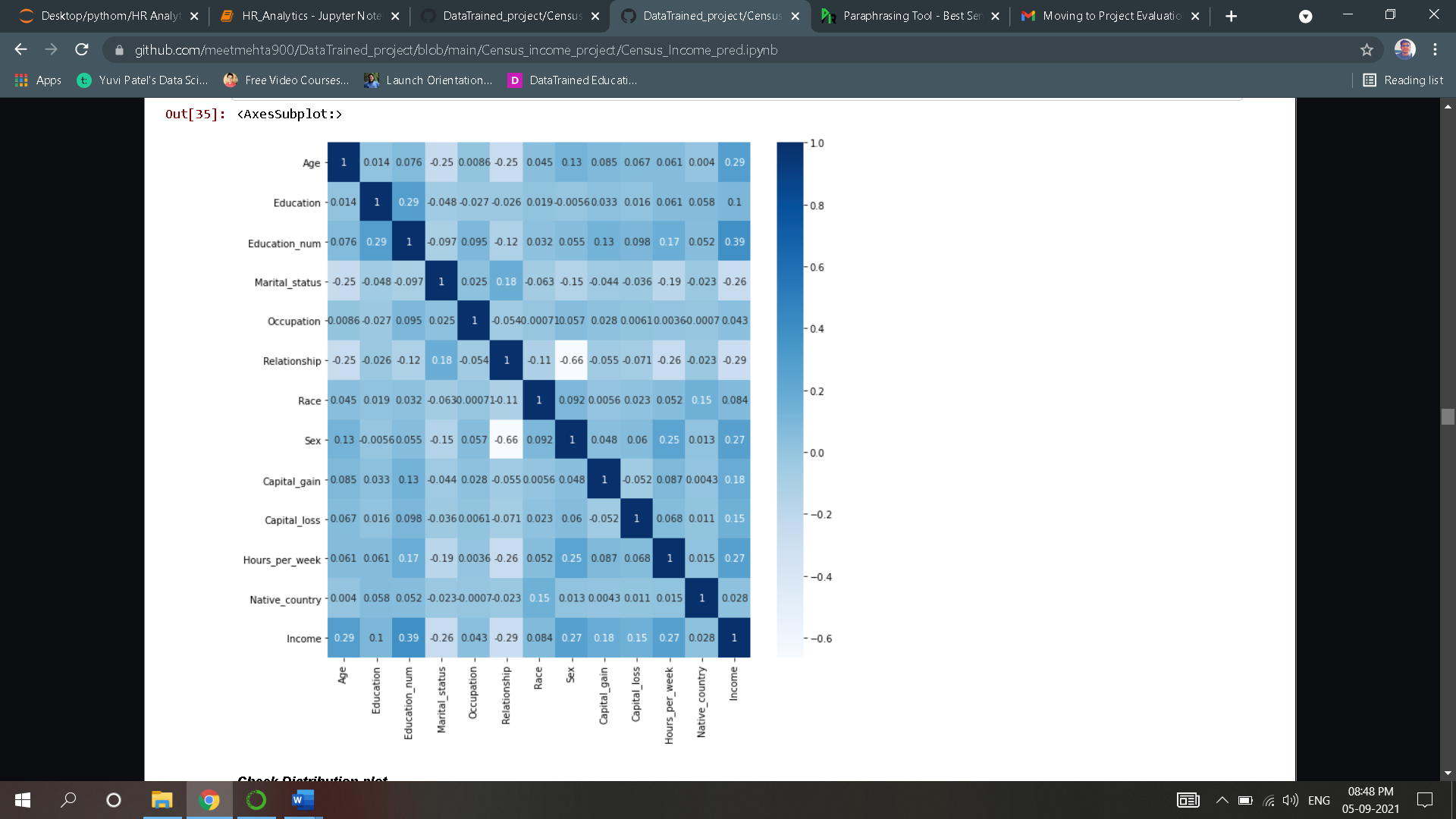
* Observation:
  + In most of the occupation type there are more number of Male population working when compared to Female population
  + In our dataset, the highest working population are engaged with "Craft-repair" occupation type.

## 2.5 Multivariate Analysis:

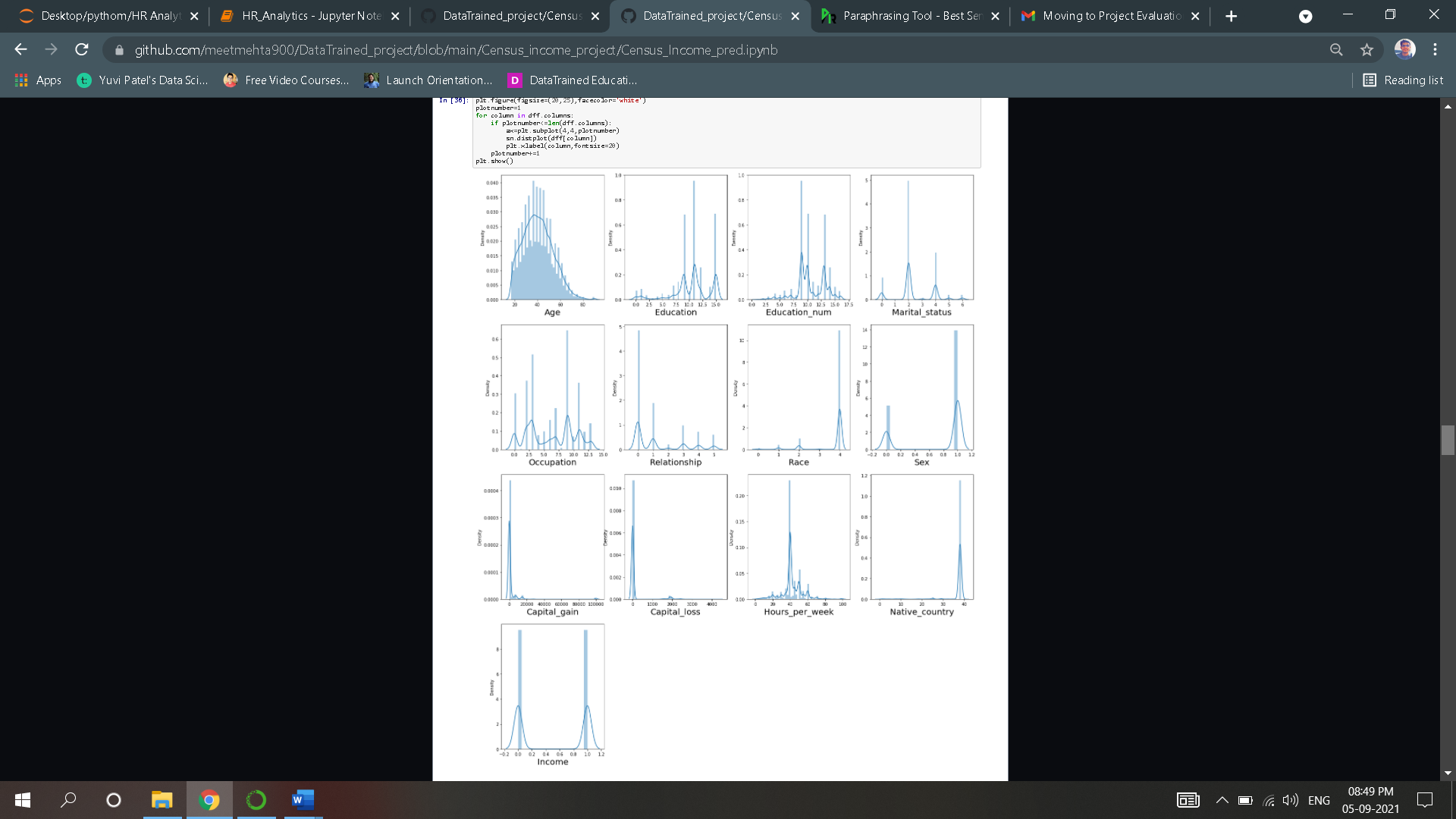
* Display pairplot of each column w.r.t column Income.



* To complete our analysis, we need to convert categorical object dtype into numeric dtype.
* Conversion of categorical object dtype to numeric dtype is simplified using label encoder.
* From univariate analysis we found that we have an imbalance dataset.
* Balancing the dataset is important otherwise our model is going to get biased.
* Using a resample function we can balance back or dataset.
* Check for correlation of each column w.r.t our target variable by plotting a heatmap.



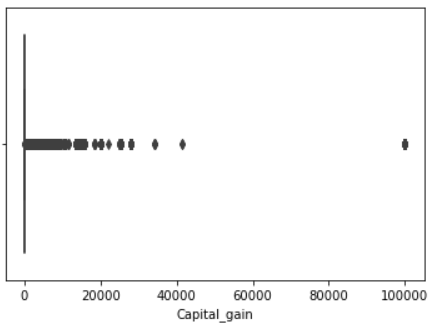
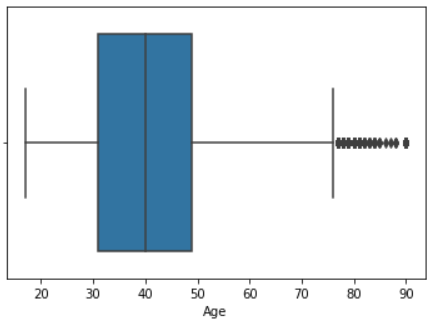
* Observation:
  + 'Fnlwgt' and 'Capital\_gain' has least correlation with value "0.000437"
  + 'Education\_num' has more correlation in our dataset with 'Hours\_per\_week'and 'Capital\_gain'.
* Check for the distribution plot of each column for the purpose of analysing the dataset.

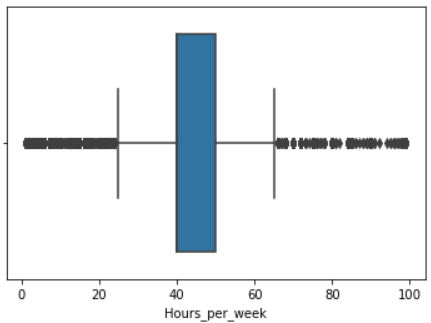
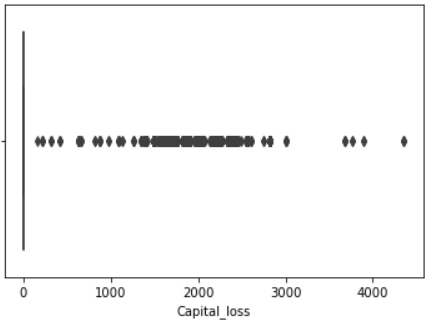


* Columns having normal distribution plot: native\_country (categorical data), capital\_loss (positively skewed), capital\_gain (positively skewed), Age
* Rest all columns have bimodal distribution plot

### We must also check for outliers using boxplot in our dataset as it can hamper our model performance.

### Outliers of only non-categorical data should be checked.





* By observing the box plot of non categorical columns, we can conclude that outliers are available in all of the non categorical columns.
* We need to remove those outliers, otherwise it will hamper our model’s performance.
* There are 2 ways to remove outliers:

o Z-score method

o IQR method

* Try the Z-score method and recheck the columns for any outliers’ present.
* After executing the Z-score method to remove outliers, check the dataset value and data loss.
* Original dataset was 49438.
* After the z-score, the dataset shape changed to 41869.
* There seems to be data loss of more than 15%, hence we cannot remove any outliers and use the original dataset.

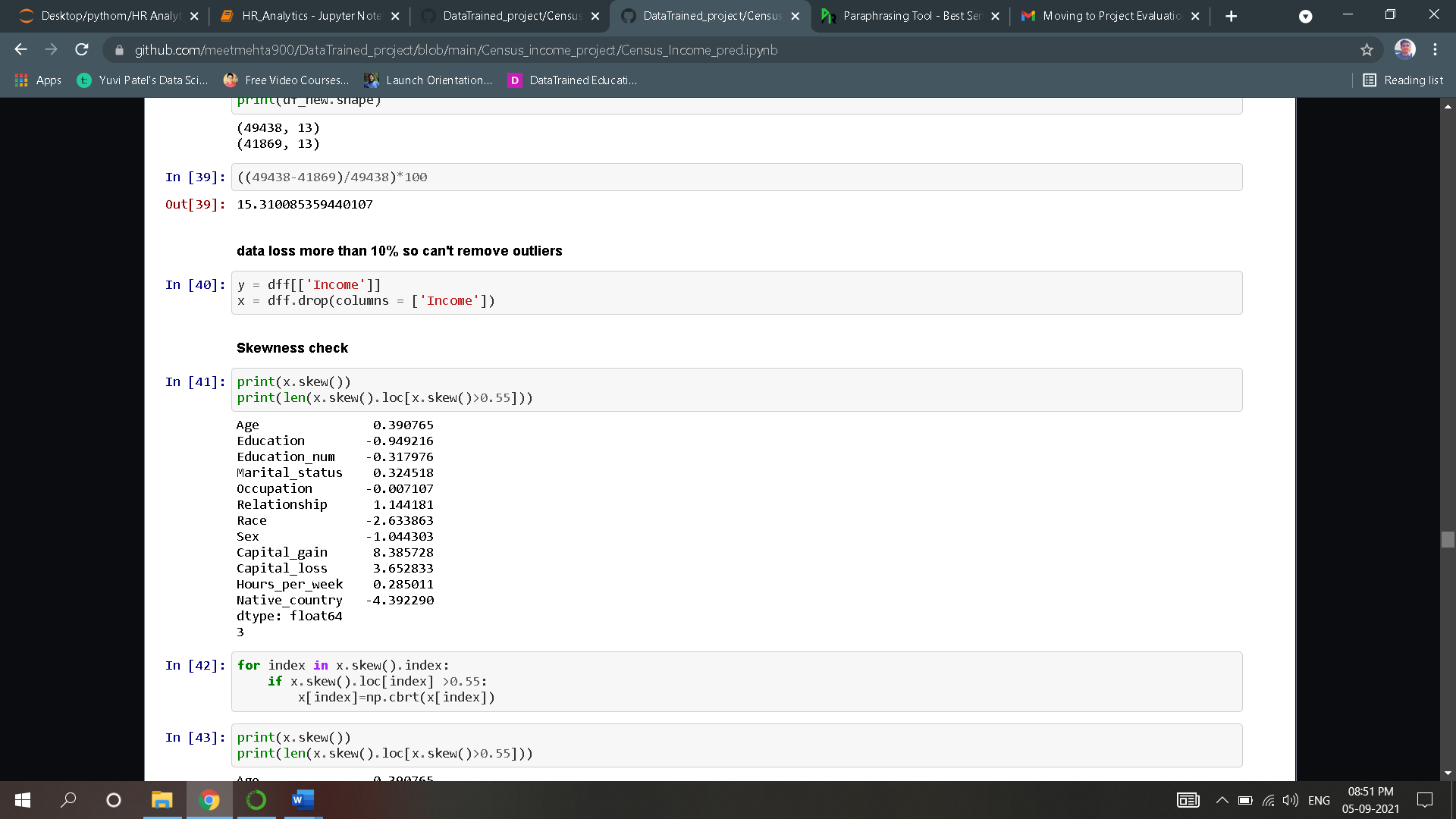
## 3. EDA Concluding Remark

* Imbalanced dataset was balanced.
* Outliers were detected in the dataset, only after further analyses we understood that removal of these outliers is causing data loss more than 10%, hence we cannot remove any outliers.

## 4.Preprocessing Pipeline

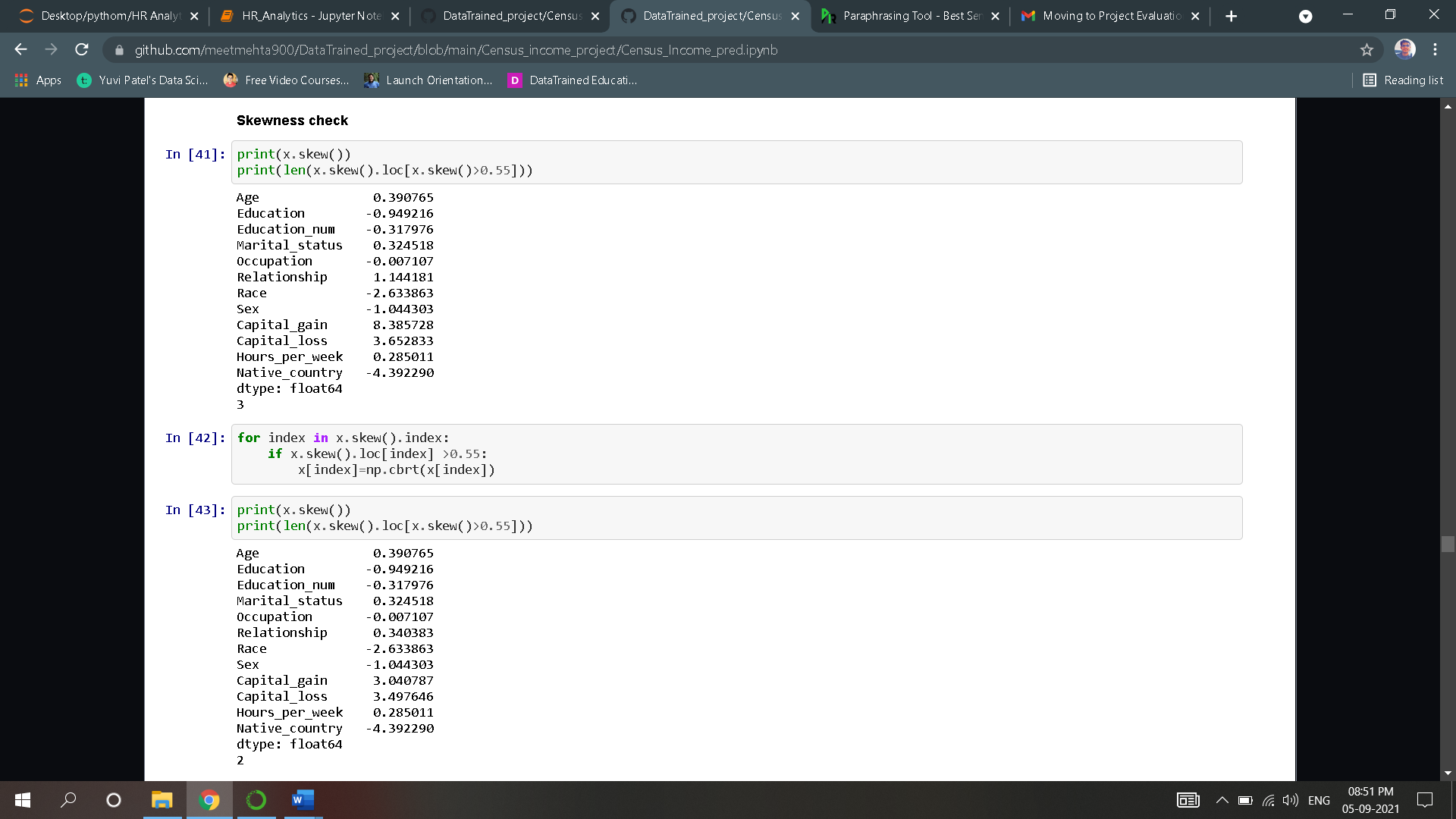
**·** To train the model we need to split target variables and input features.

· To avoid biasing issues, we need to check for biasing using the skew method of input features and remove if required.

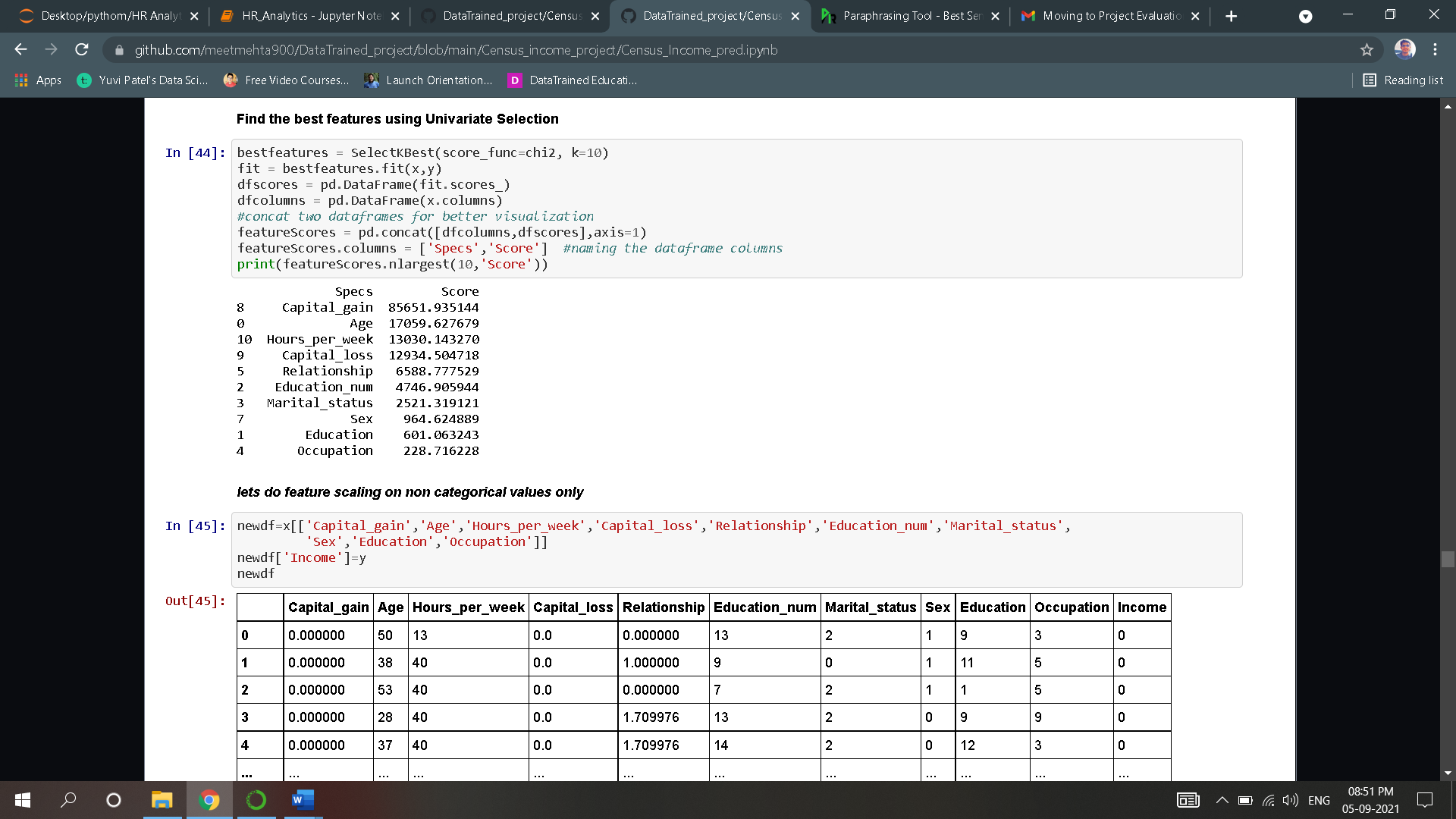


**·** Skewness is present if the skew value of a column is more than 0.55.

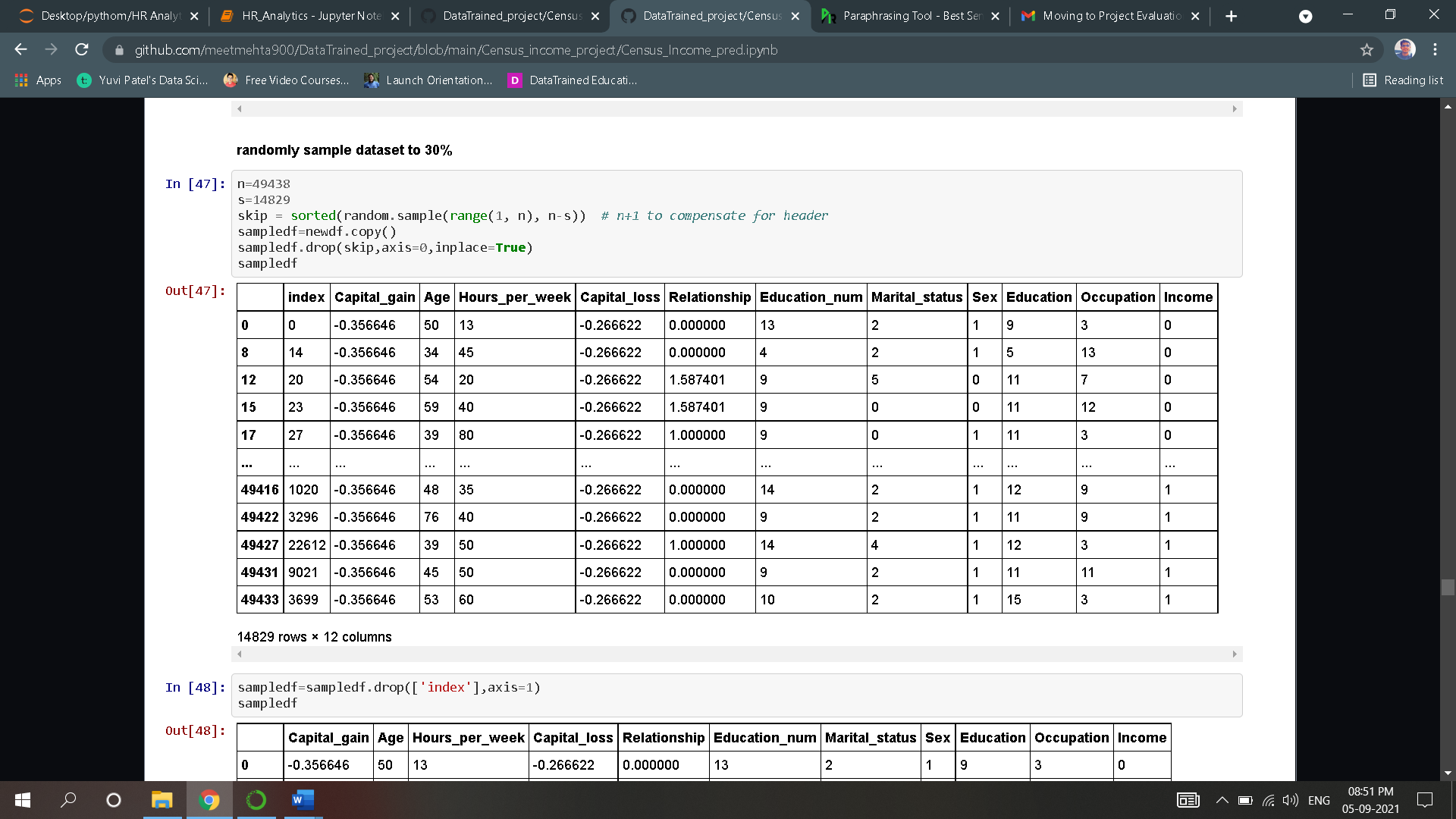
· Remove biasing of specific columns by replacing its data points by cube root of its data points.



* To improve on the model accuracy, we need to find the best features using Univariate Selection



* From above we can observe the columns w.r.t its score regarding the column’s importance for model training. By evaluating columns according to its score, we can obtain the best features required to train the model.
* According to the score we will pick Capital\_gain','Age','Hours\_per\_week','Capital\_loss','Relationship','Education\_num','Marital\_status', 'Sex','Education','Occupation' · features to train the model.
* Perform feature scaling using standard scaler algorithm on non-categorical columns only.
* The objective of performing feature Scaling is to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing.



* Instead of finding the best models using the whole dataset which will be very computationally heavy and would take a lot of time to find best hyperparameters, it will be better to take 30% randomly sampled dataset. using this randomly sampled dataset build and find best model and then its best hyperparameters. and then build that model on the whole dataset.
* Randomly sample dataset to 30% and use that dataset to build the model.

## 5. Building Machine Learning Models

### Now we will start with the model selection and fine-tuning process.

### First, we need to find the most optimum model.

### We shall evaluate the model on roc auc score & cv score.

### The types of models through which we need to iterate are:

### LogisticRegression

### DecisionTreeClassifier

### KNeighborsClassifier

### RandomForestClassifier

### SVC

### RidgeClassifier

### BaggingClassifier

### GradientBoostingClassifier

### SGDClassifier

### LGBMClassifier

### XGBClassifier

### ExtraTreesClassifier

### AdaBoostClassifier

### QuadraticDiscriminantAnalysis

### CalibratedClassifierCV

### LinearSVC

### NuSVC

### LinearDiscriminantAnalysis

### RidgeClassifierCV

### GaussianNB

### BernoulliNB

### PassiveAggressiveClassifier

### Perceptron

### DummyClassifier

### · After iterating through all the above algorithms, we obtained the following roc auc accuracy, cv score and the difference between roc auc score and cv score.

### · Display the top 5 models metrics.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Cross-Val Score | ROC\_AUC Score | Difference |
| LGBM Classifier | 93.17% | 85.42% | 7.7% |
| XGB Classifier | 93.10% | 85.48% | 7.6% |
| RandomForest Classifier | 92.32% | 86.36% | 5.95% |
| Gradient Boosting Classifier | 92.28% | 84.81% | 7.5% |
| Extra Trees Classifier | 90.67% | 84.91% | 5.75% |

### From above we can conclude that Extra Trees Classifier is the best model without any issues of underfitting or overfitting.

### Perform fine tuning on Extra Trees Classifier model and find the best parameters to be used for the model by using GridsearchCV algorithm.

### For grid search use of the following parameters for extra trees classifier:

### 'criterion': ['gini', 'entropy']

### 'max\_features': ['auto', 'sqrt', 'log2']

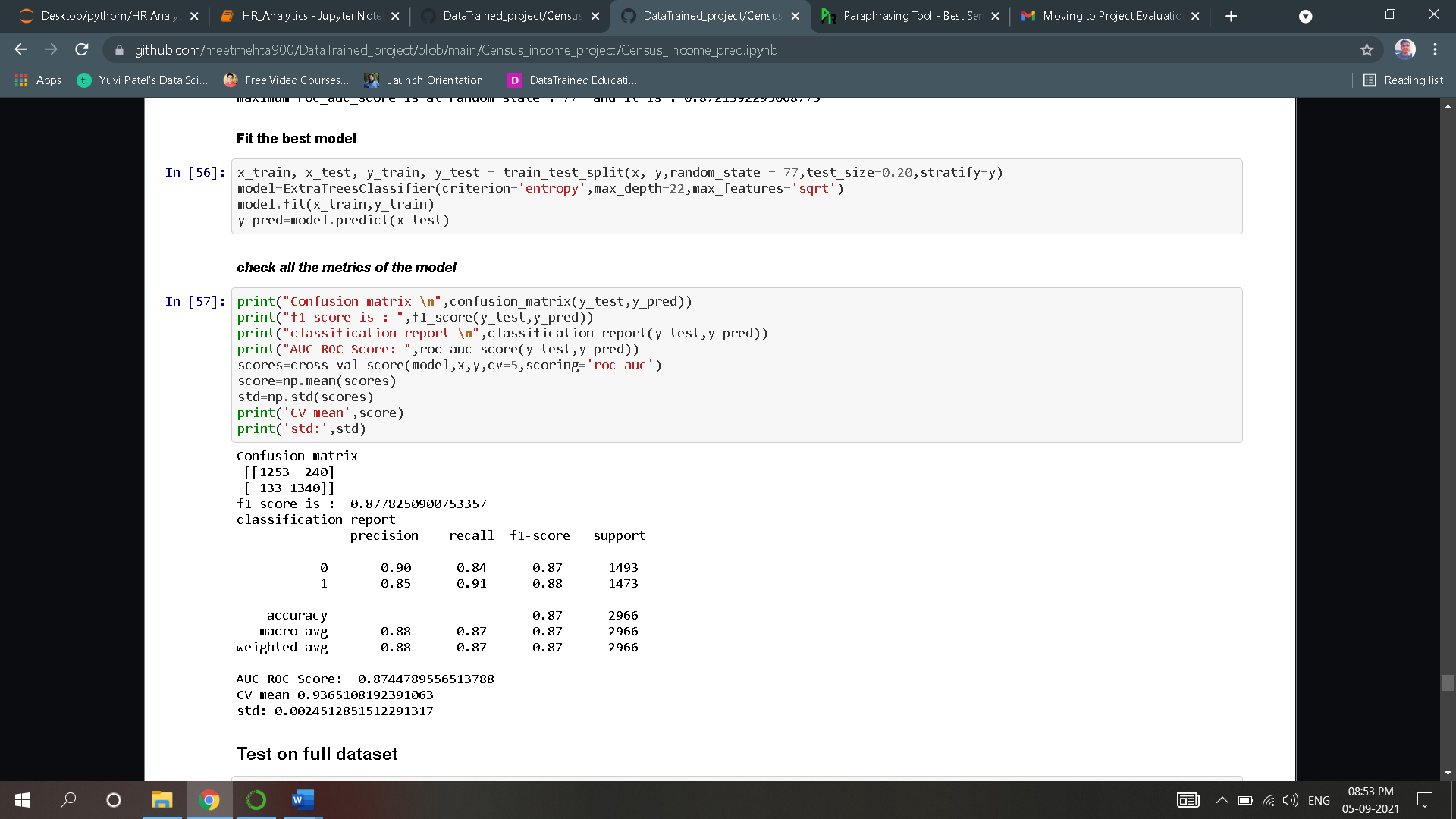
### 'max\_depth':list(range(3,36))

### **Best parameters for Extra Trees Classifier obtained by executing Grid Search are:**

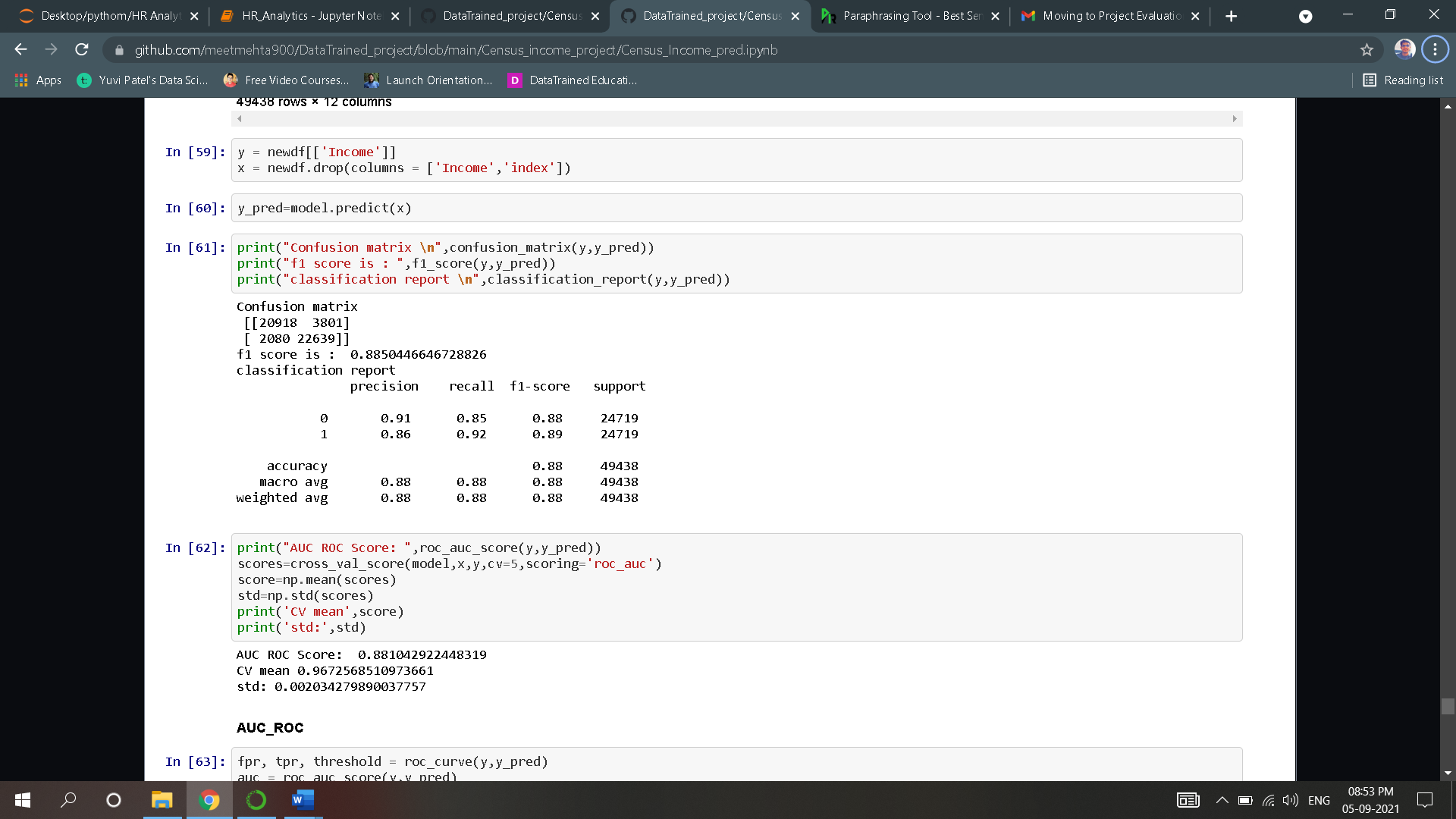
* 'criterion': entropy
* 'max\_features': 'sqrt'
* 'max\_depth ': 22
* Now we need to find the most optimum random state in the train test split for Extra Trees Classifier model to get the best score, in this case the best random state is 77.

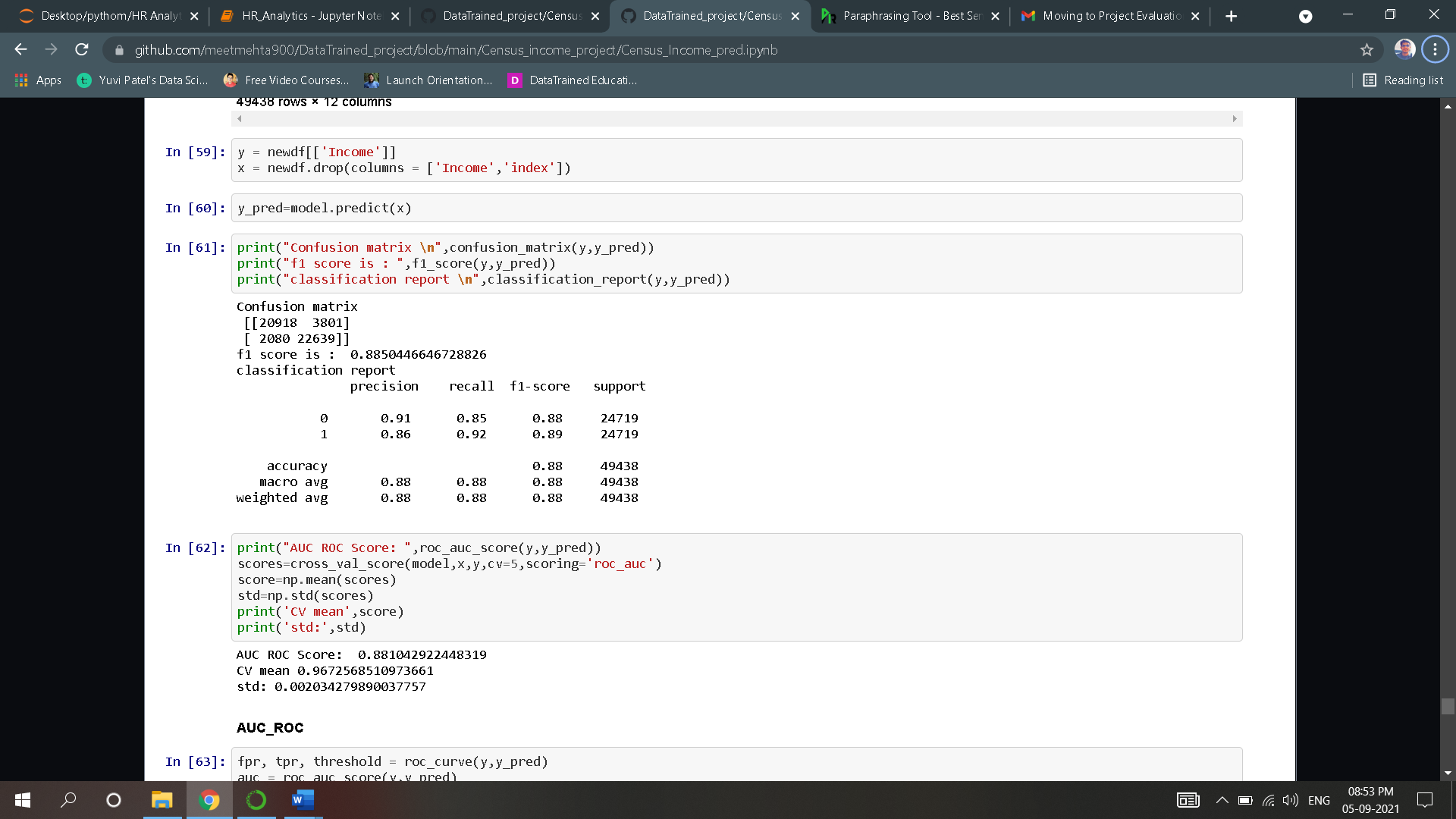
### We are going to split the dataset by keeping 20% for testing and 80% for training.

* Using the model extra trees classifier along with the best parameters obtained from grid search, fit the model w.r.t the train dataset.
* Evaluate the model based on sampled dataset and obtain all the metrics of the currently trained model:

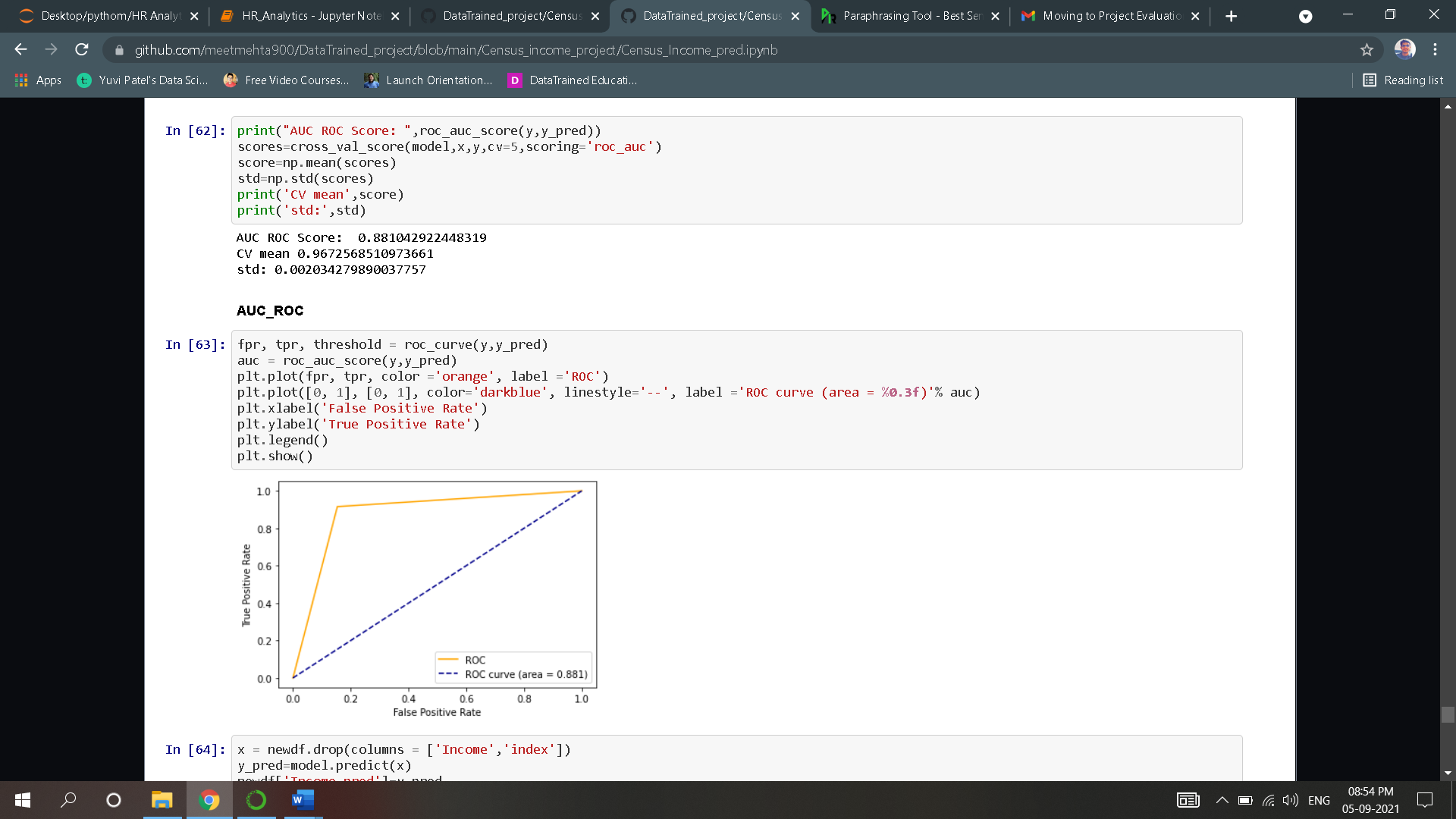


### Now Evaluate the model based on sampled dataset and obtain all the metrics of the currently trained model i.e.on full dataset

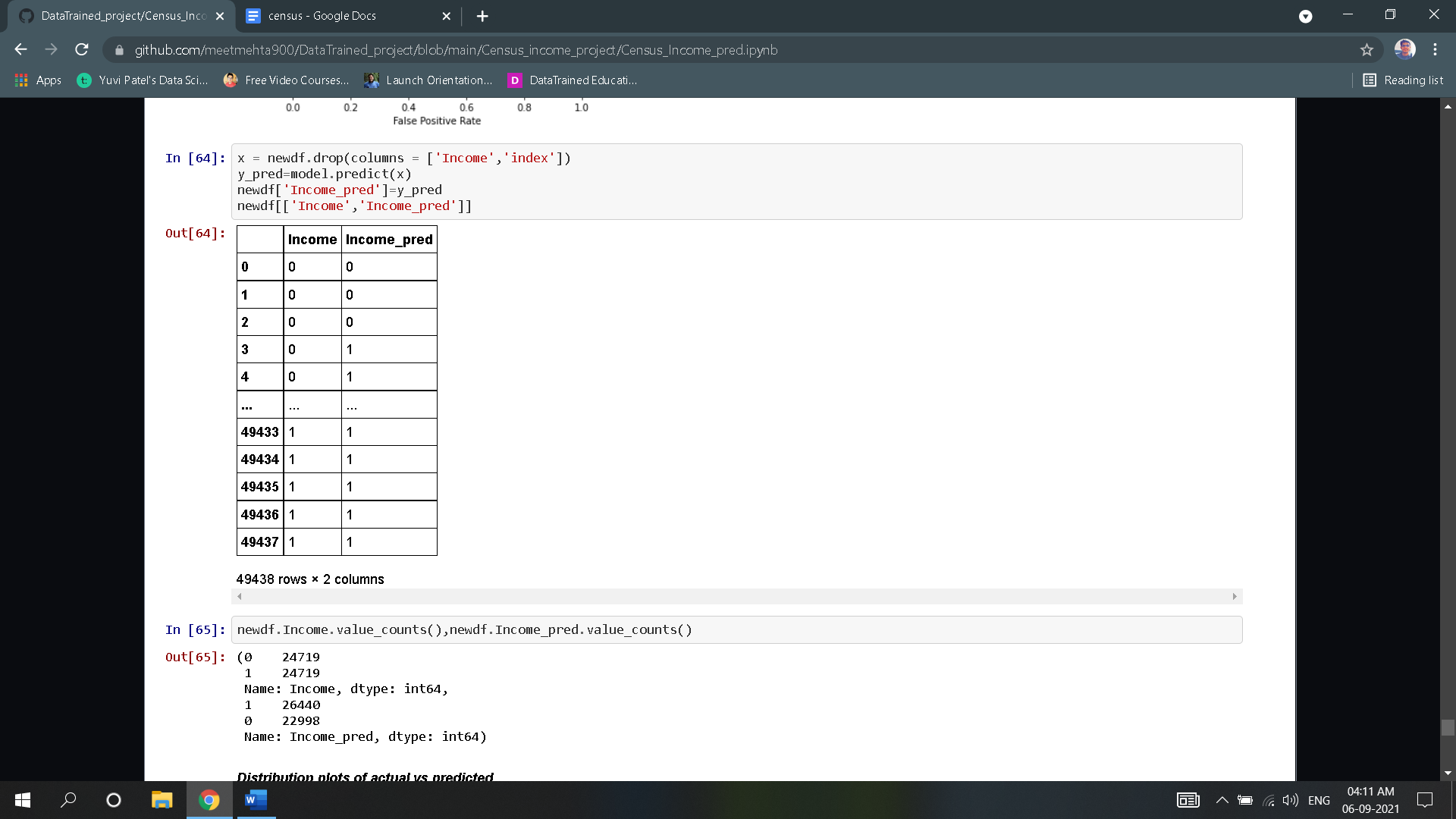




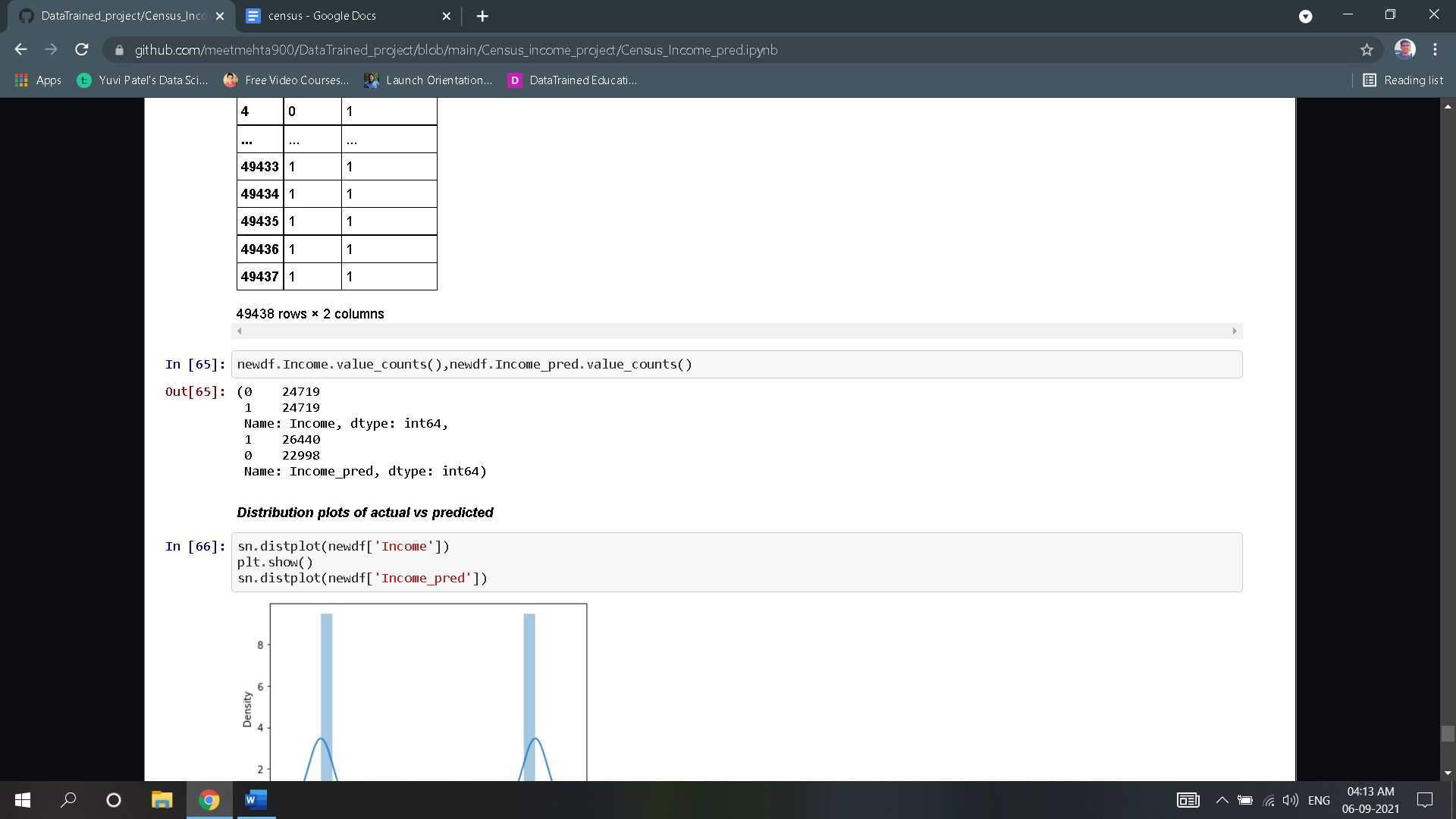
* As observed, the accuracy of the model varies from 88.5% to 88.1%.
* Plot the ROC curve of the model to visualize the accuracy of the model.



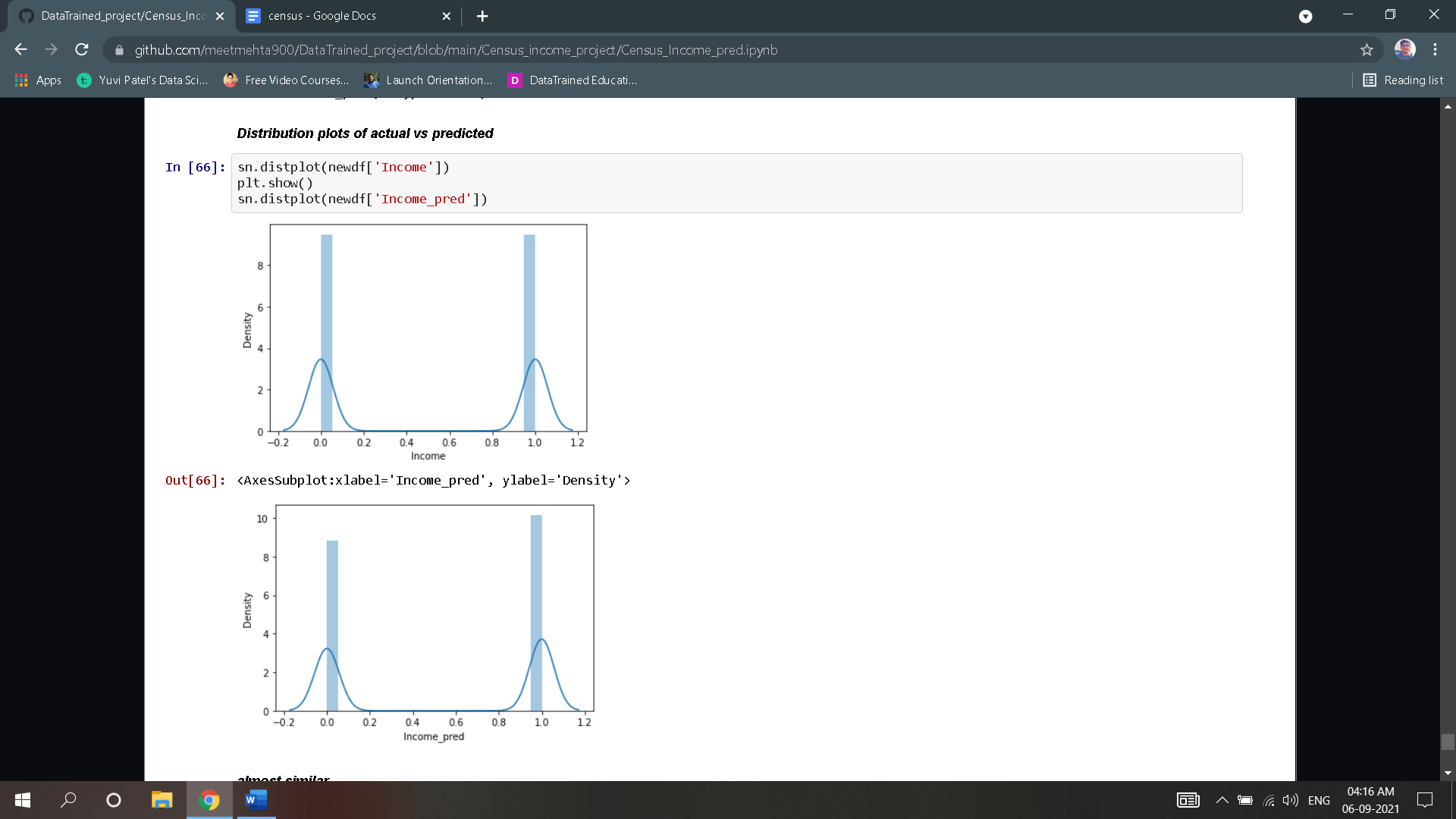
* Put target variable ‘Income’ column & predicted “Income-pred” column side by side to observe the difference between the actual and predicted data points

****

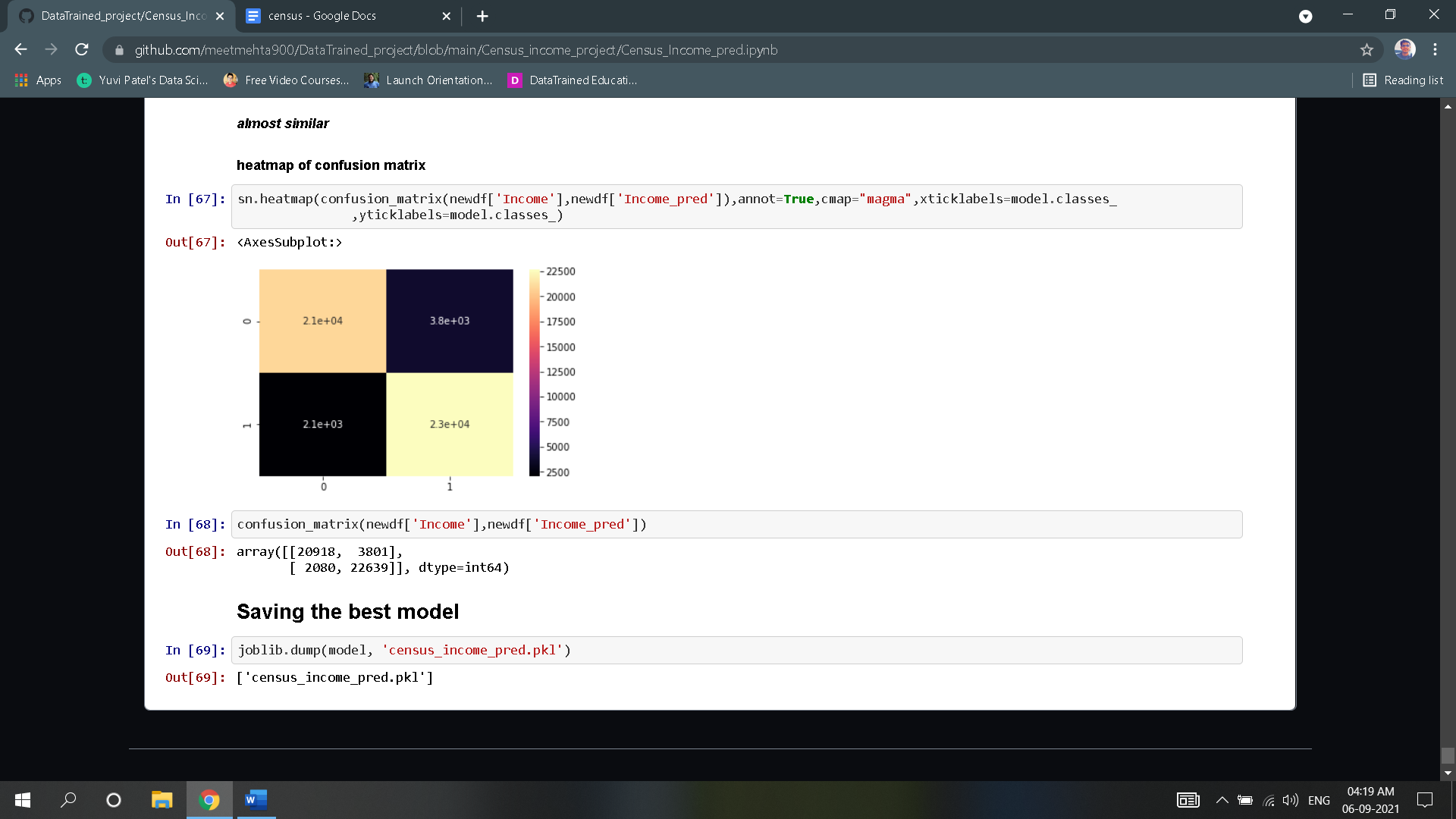
* Let's get the values of actual & predicted.



* Check Distribution plot of both the columns



* They plot look extremely similar and density has also not that much changed, so plot more graphs to evaluate model more accurately
* Display the heatmap of confusion matrix



* Type I error:3801
* Type II error:2080
* Finally, we save the model using the **“joblib**” library which can be reused for further prediction.

#### · 6. Concluding remarks

We were able to create a model that was accurate to 88 % to predict whether a person makes over $50K a year in advance based on selected key features. We used the Extra Tree classifier model for high precision so we got low standard deviation error and no problems with overfitting or mismatching. The difference between the actual and the predicted value was obtained. By further visualizing the data points, we can conclude that the model can accurately predict whether a person makes over $50K a year. The saved model can be loaded and reused for prediction. The accuracy of the model can be increased by providing more training data.

### **Click the link below to go through the jupyter notebook:**

<https://github.com/meetmehta900/DataTrained_project/blob/main/Census_income_project/Census_Income_pred.ipynb>