**HR Analytics Case Study**

1. Problem Definition

Every year a lot of companies hire a number of employees. The companies invest time and money in training those employees, not just this but there are training programs within the companies for their existing employees as well. The aim of these programs is to increase the effectiveness of their employees. But where HR Analytics fit in this? and is it just about improving the performance of employees?

**HR Analytics**

Human resource analytics (HR analytics) is a subset of analytics that involves applying analytic techniques to an organization's human resource department in the hopes of enhancing employee performance and thereby increasing return on investment. HR analytics is more than just collecting data on employee productivity. Rather, it seeks to provide information about each step.

**Attrition in HR**

Attracting human resources means the gradual loss of employees' overtime work. In general, companies face the problem of relatively high staff turnover. HR professionals often take the lead in developing a company's compensation programs, work culture, and incentive systems that help organizations retain the best employees.

How does Attrition affect companies? and how does HR Analytics help in analysing attrition?

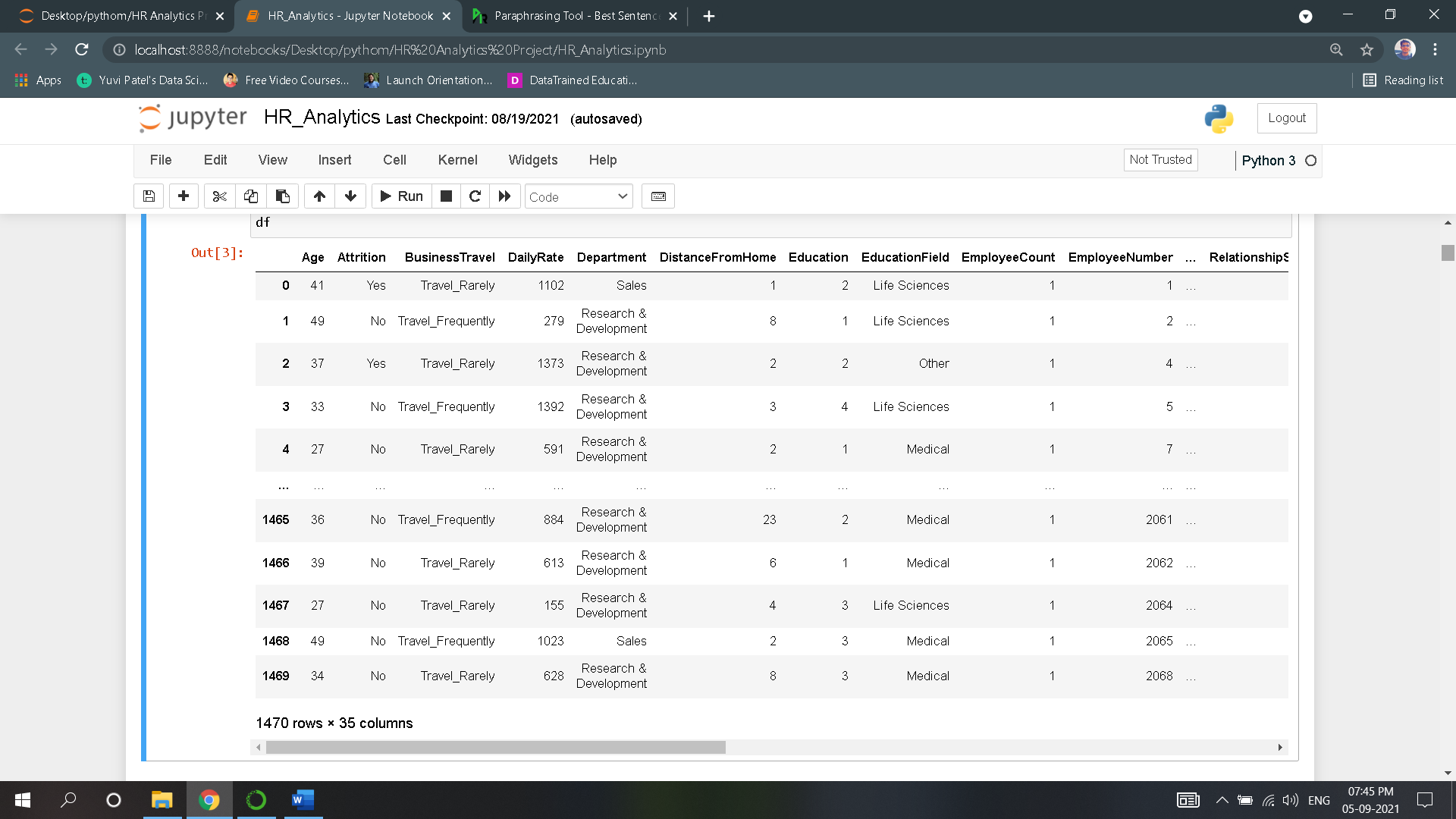
**Attrition affecting Companies**

The main concern with leaving many employees is the cost to the organization. Publishing, hiring, paperwork and training new hires are some of the most common costs of losing and replacing employees. In addition, frequent membership changes prevent your organization from increasing its collective knowledge and experience over time. This is especially important if your business is customer-centric, as customers often enjoy interacting with people they know. There will be more errors and crashes if you constantly have new workers.

2. Data Analysis

2.1 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 datapoints in rows and columns.



* Our requirement is to predict Attrition with the help of several different attributes.
* Each column in the Data frame is as follows:

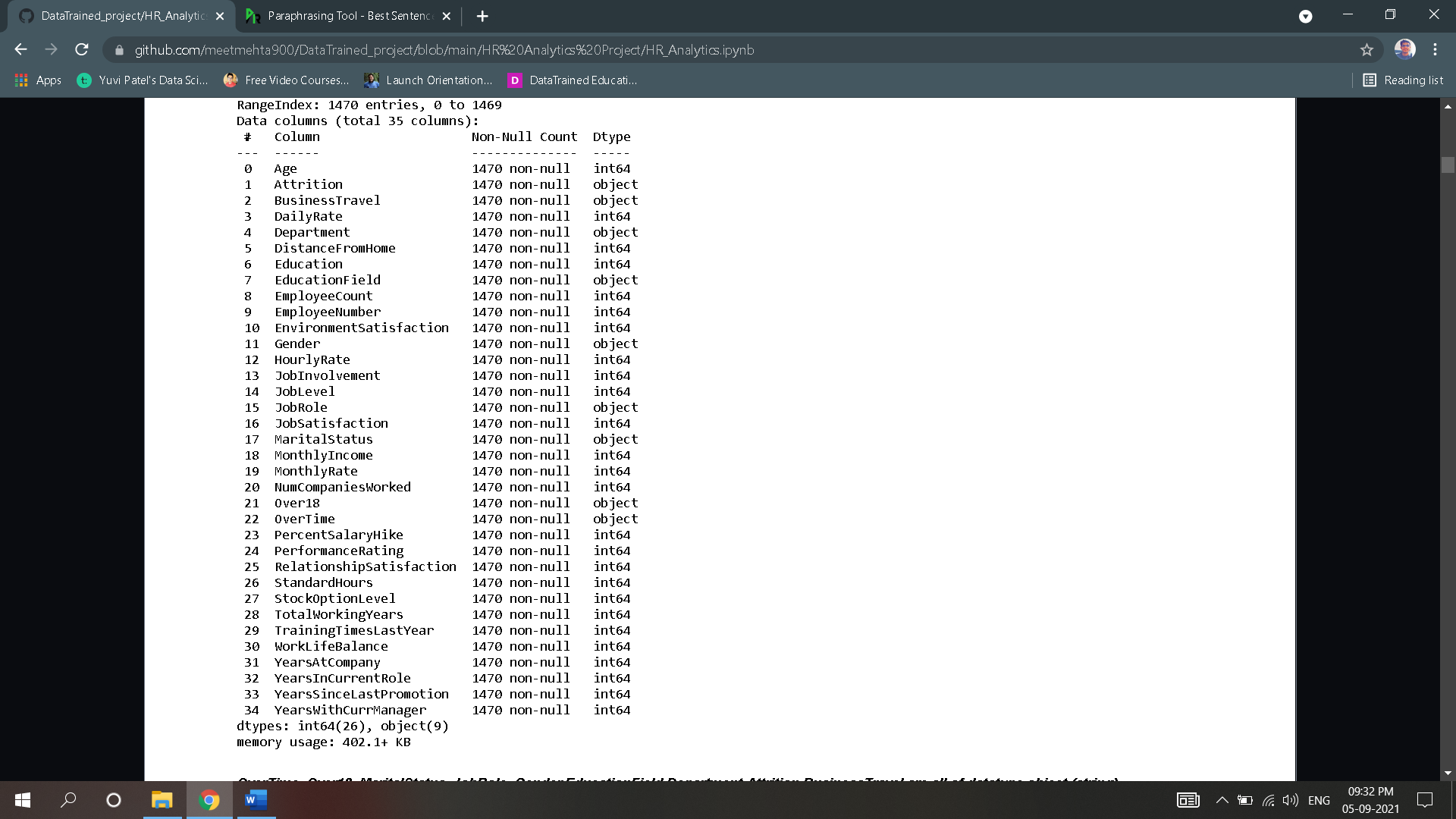
Output/Target Variable:

* Attrition: whether employees get fired or not

Input/Independent Variable:

* Age
* Business Travel
* Daily Rate
* Department
* Distance from Home
* Education
* Education Field
* Employee Count
* Employee Number
* Environment Satisfaction
* Gender
* Hourly Rate
* Job Involvement
* Job Level
* Job Role
* Job Satisfaction
* Marital Status
* Monthly Income
* Monthly Rate
* No. of Companies Worked
* Over18
* Overtime,
* Percent Salary Hike
* Performance Rating
* Relationship Satisfaction
* Standard Hours
* Stock Option Level
* Total Working Years
* Training Times Last Year
* Work-Life Balance
* Years at Company
* Years in Current Role
* Years Since Last Promotion
* Years with Current Manager

2.2 Data Analysis



* After executing the ‘info’ command we obtain the above output and following is our observation:
* Number of rows in dataset are:1470
* Number of columns in dataset are:35
* Dataset does not contain any null values.
* There are 26 numerical columns and 9 categorical columns which includes target variable.

2.2.1. Data Preparation & Cleaning

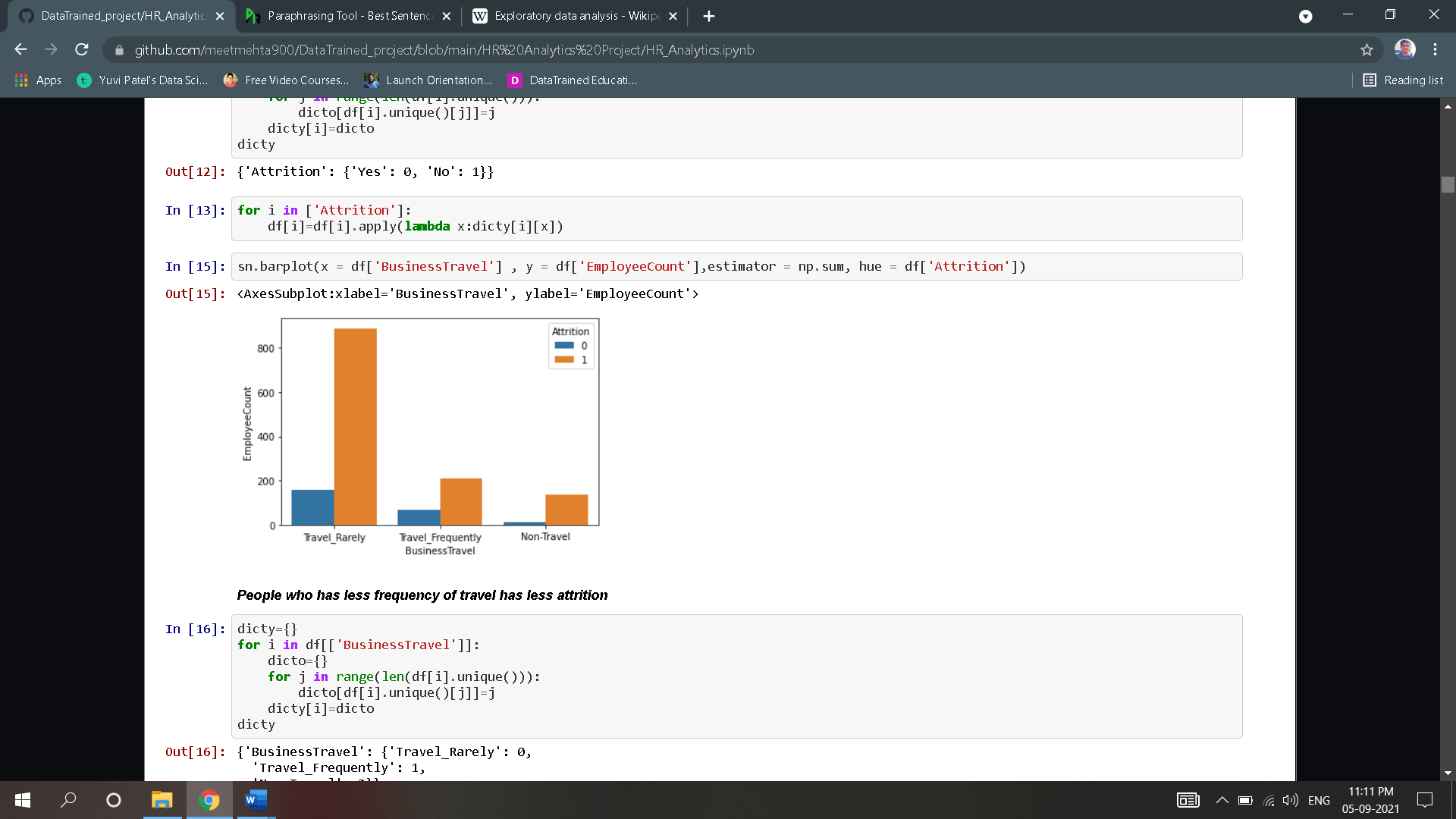
* Converting Categorical Columns into Numeric type as model which are about to build does not accept categorical values.
* For we need to know unique values of each categorical columns.
* {'Attrition': {'Yes': 0, 'No': 1}}
* {'BusinessTravel': {'Travel\_Rarely': 0,'Travel\_Frequently': 1,'Non-Travel': 2}}
* {'Department': {'Sales': 0, 'Research & Development': 1, 'Human Resources': 2}}
* {'EducationField': {'Life Sciences': 0,'Other': 1,'Medical': 2,'Marketing': 3,'Technical Degree': 4,'Human Resources': 5}}
* {'Gender': {'Female': 0, 'Male': 1}}
* {'JobRole': {'Sales Executive': 0,'Research Scientist': 1,'Laboratory Technician': 2,'Manufacturing Director': 3,'Healthcare Representative': 4,'Manager': 5, 'Sales Representative': 6,'Research Director': 7,'Human Resources': 8}}
* {'MaritalStatus': {'Single': 0, 'Married': 1, 'Divorced': 2}}
* {'Over18': {'Y': 0}}
* {'OverTime': {'Yes': 0, 'No': 1}}
* Now this Unique value will be converted to numerical type by using **Label encoder.**
* Now that all columns are converted to numeric type, we can have brief **Statistical Summary** containing important metrics such as Mean, Standard deviation, Minimum value, Maximum value of each column are obtained using describe function.
* These metrics gives us information about the possibility of outliers present in column, whether all values are balanced or not.



* Upon the above summary, there is huge variance in quartile range of many columns and standard is high comparatively which means outliers are present in columns which needs to taken care of in further analysis.

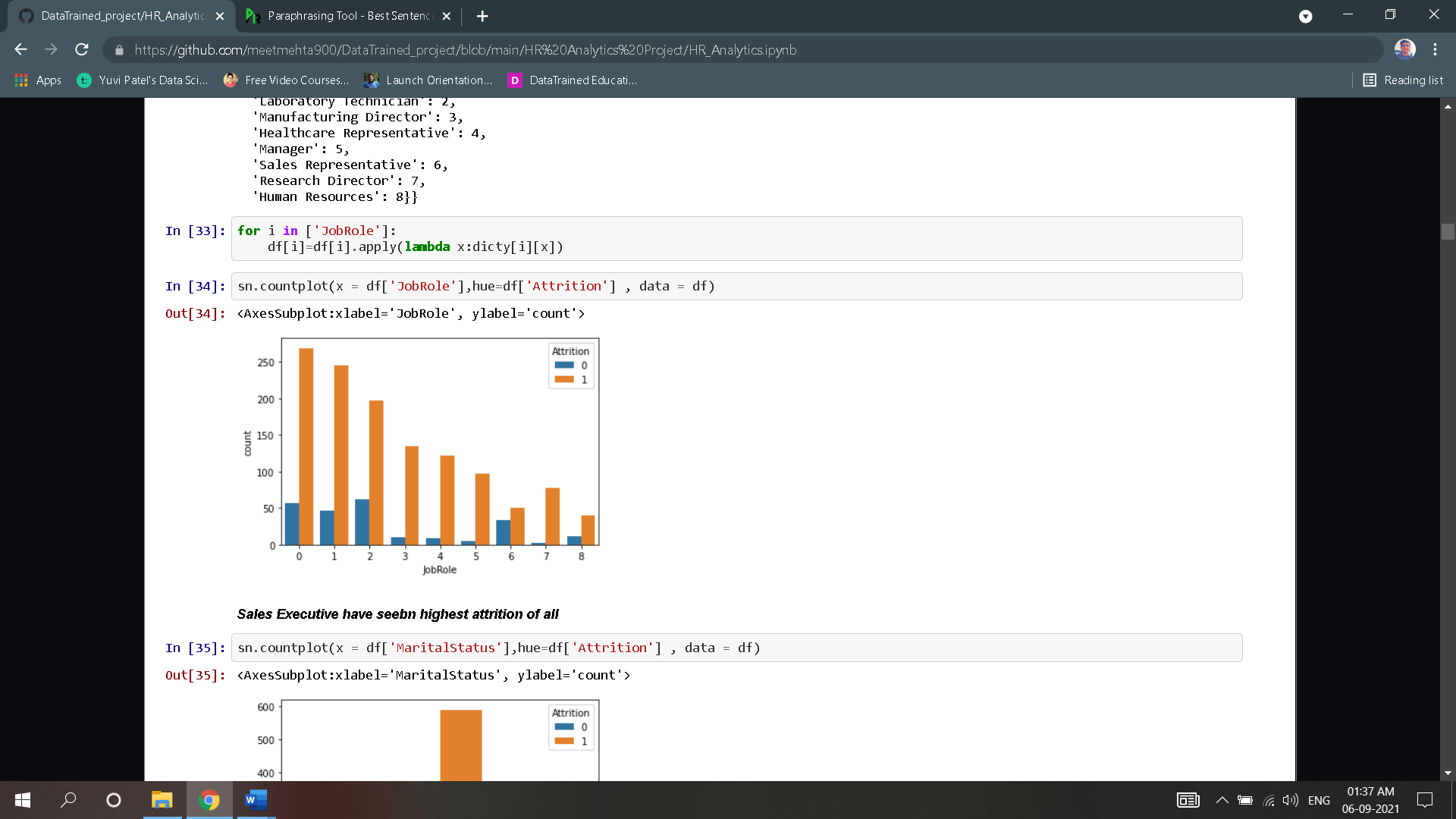
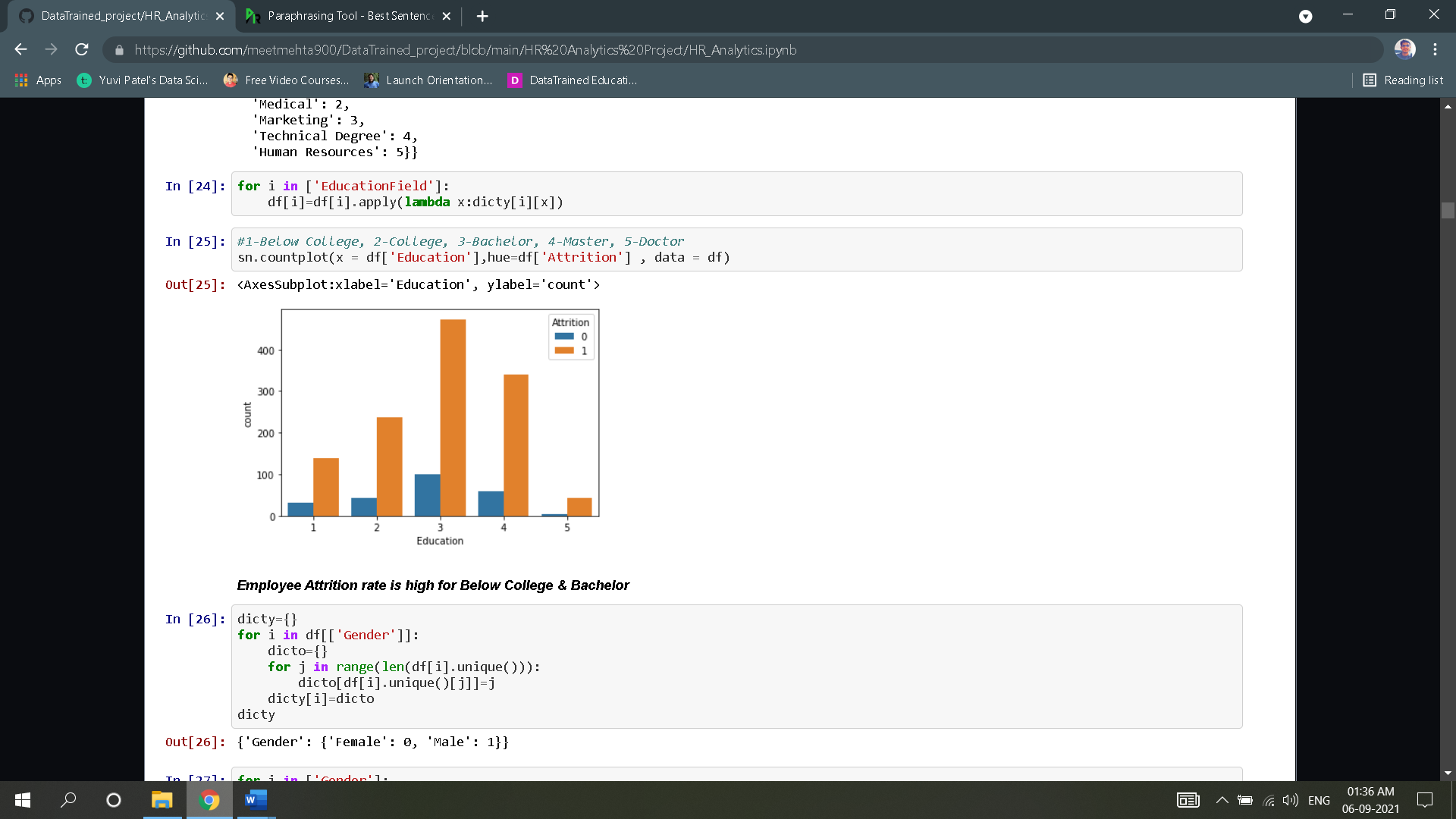
3.Exploratory Data Analysis

* Exploratory data analysis is an approach of analysing [data sets](https://en.wikipedia.org/wiki/Data_set) to summarize their main characteristics, often using [statistical graphics](https://en.wikipedia.org/wiki/Statistical_graphics) and other [data visualization](https://en.wikipedia.org/wiki/Data_visualization) methods.

##### Upon above plots, we can see that, People who has less frequency of travel has less attrition

##### Sales and R&D department has highest attrition rate.



##### Upon above plots, we can see *that Employee Attrition rate is high for Below College & Bachelor*

##### Sales Executive have seen highest attrition of all.

##### 

##### Attrition is higher in people who are married.

##### Employees who are not doing overtime has higher attrition.

##### 

##### Employees with last promotion of more than 3 years has higher chances of attrition.

##### Employees lasting less than 2 years with current manager has low chance of attrition.

##### 

##### Employees working in more than 1 company has high attrition rate

##### As the Salary hike percent increases, the attrition rate is also increased.

##### 

##### Employees with Stock Option level 1 has higher attrition rate.

##### Employees with Total working years of more than 7 years has higher chances of attrition.

##### Let’s Check the Distribution plot of the dataset:

##### 

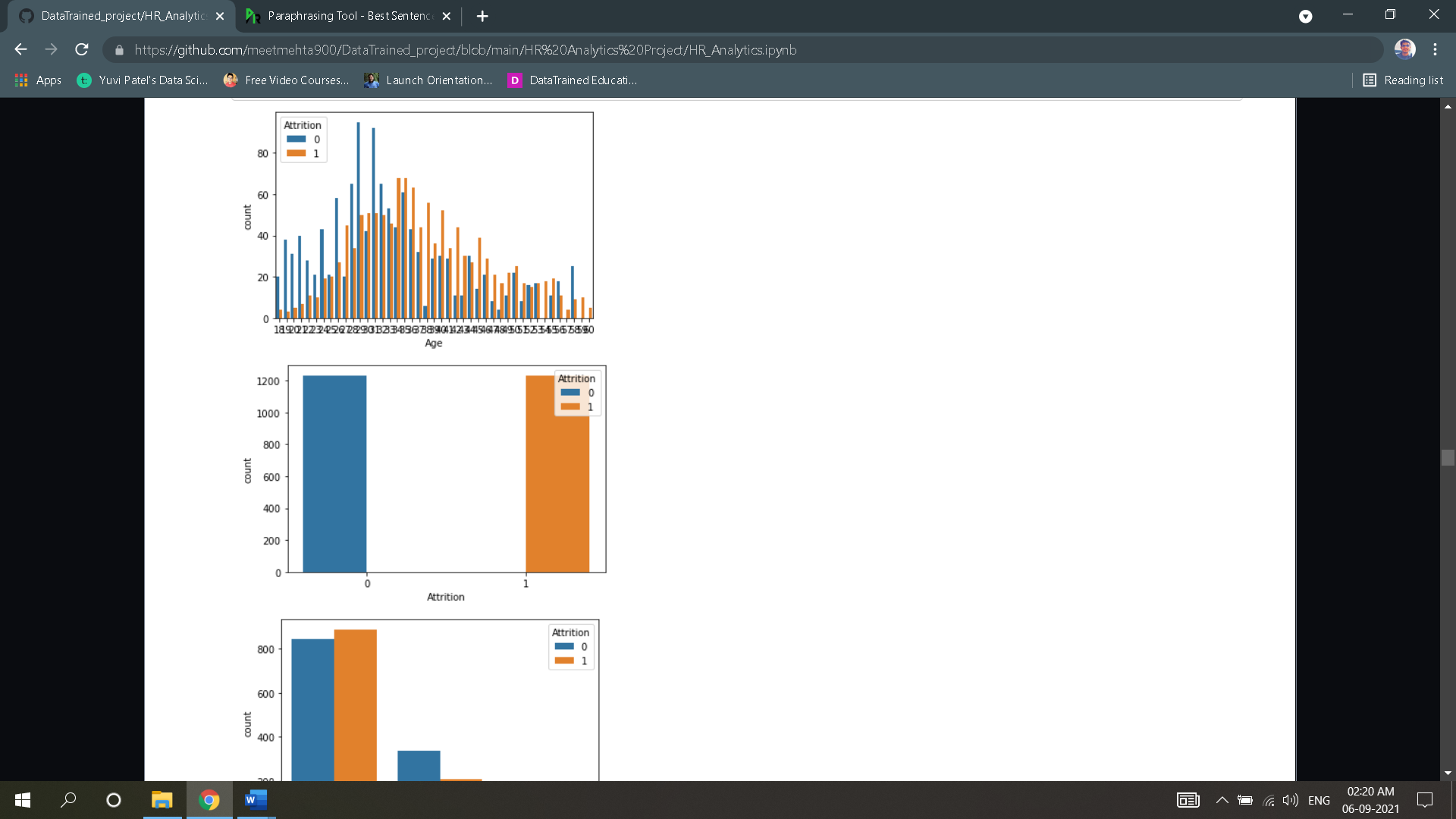
##### 

* **Monthly rate, Hourly rate & Daily rate have rectangular plot.**
* **YearsWithCurrManager, YearsInCurrentRole,WokLifeBalance,TrainingTimesLastYear, StockOptionLevel,RelationshipSatisfaction,PerformanceRating,OverTime,JobSatisfaction,MaritalStatus,JobLevel,JobInvolvement,Gender,EnvironmentSatisfaction,EducationField, Education, Department, Business Travel, Attrition have bimodal plot.**
* **YearsSinceLastPromotion (Positively skewed), YearsAtCompany (Positively skewed),TotalWorkingYears (Positively skewed), PercentSalaryHike (Positively skewed), NumCompaniesWorked (Positively skewed), MonthlyIncome (Positively skewed), JobRole (Positively skewed), DistanceFromHome (Positively skewed), Age.**
* **As this is a classification problem we also need to check if there is data imbalance.**
* Data imbalance is generally caused when one class datapoints are high compared to that of the other class.
* For example, in this project if Attrition value counts of yes/No is higher than the other, then there is data imbalance problem.

Evaluate the value counts of Attrition datapoints

##### 

* Number of Attrition ‘no’ datapoints are greater yes datapoints.
* Thus, we can deduce that we have an unbalanced dataset.
* Balance the dataset using resample function and plot bar graph.



**Upon this Analysis, we shall drop columns accordingly:**

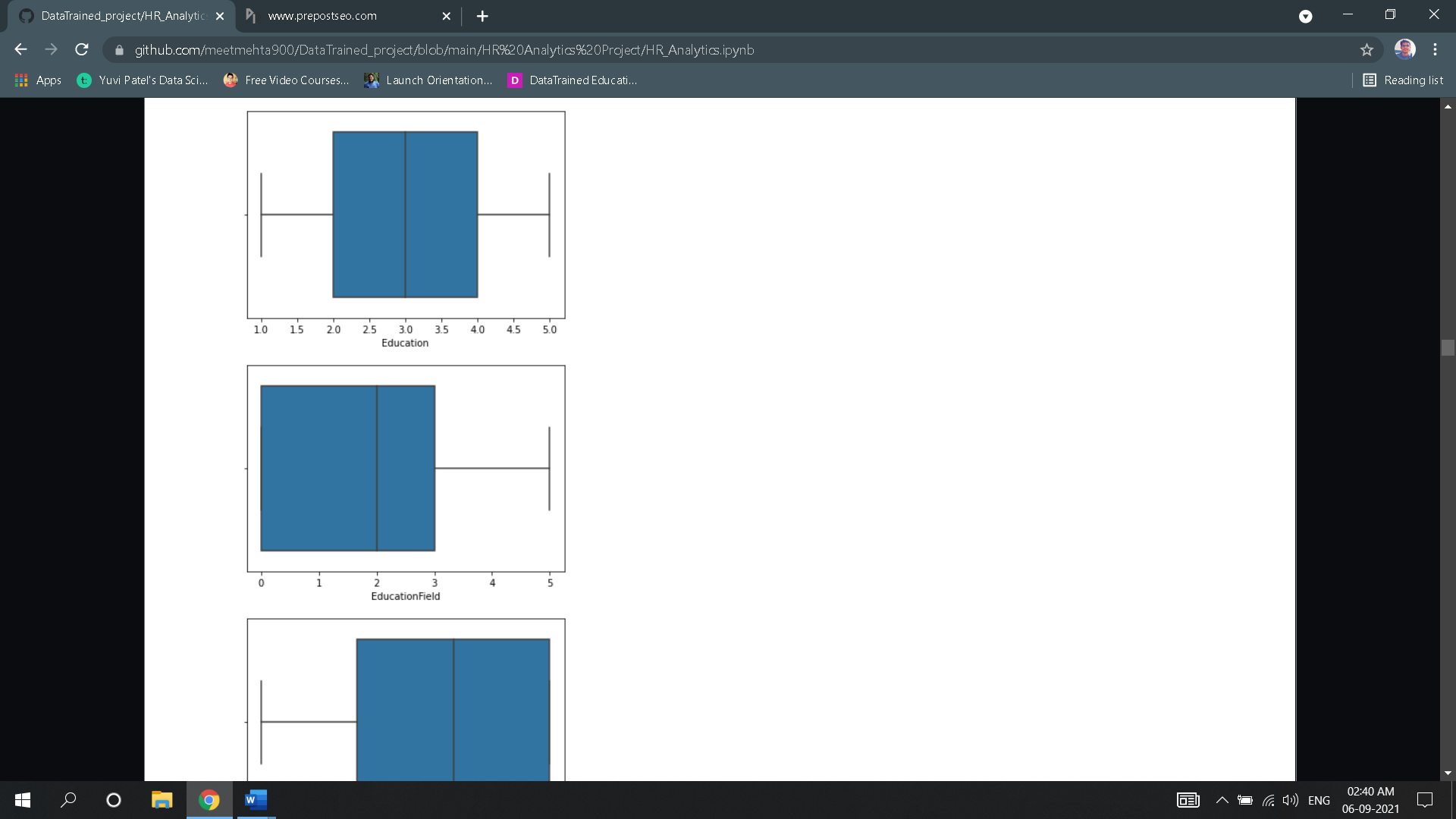
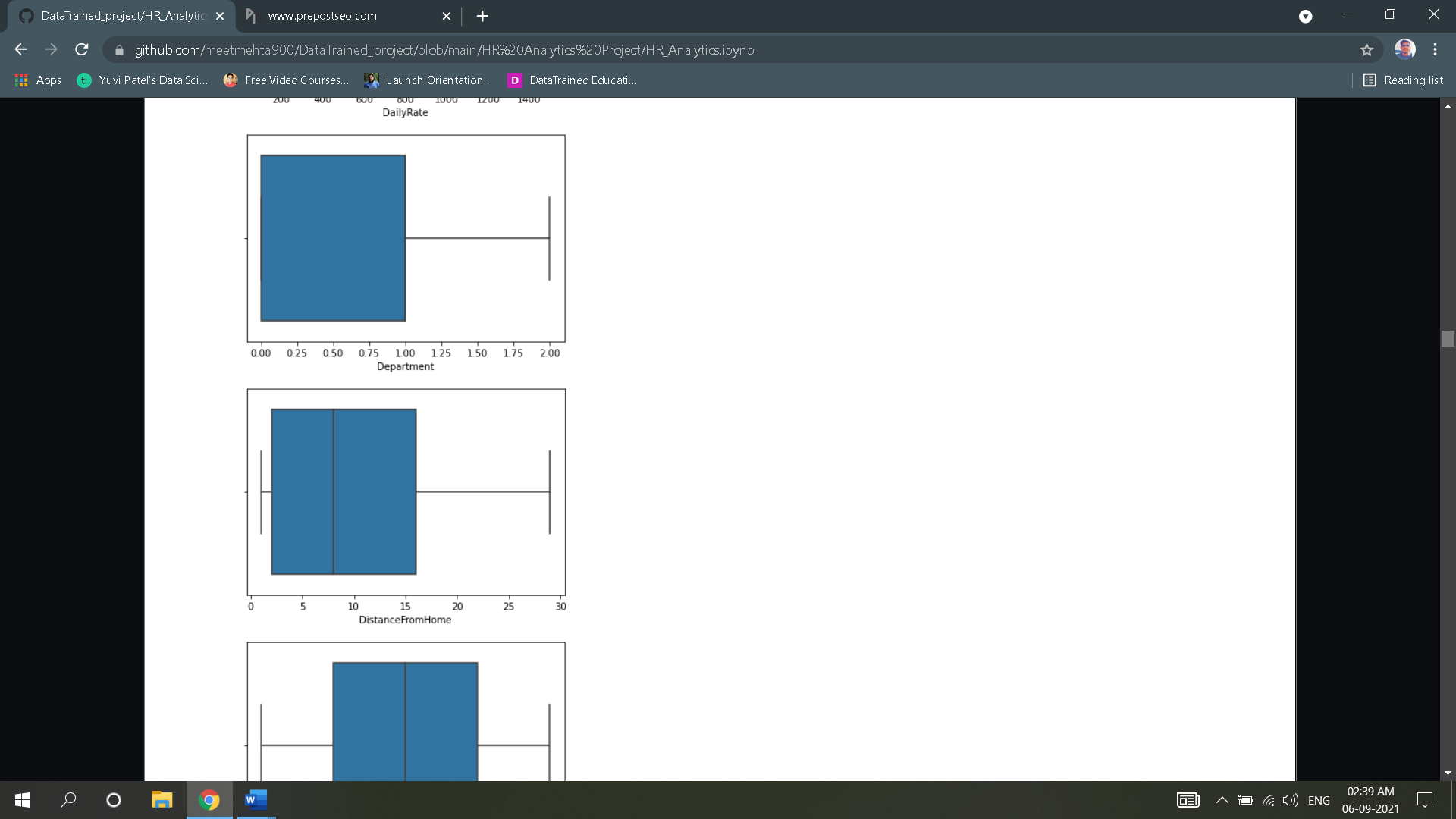
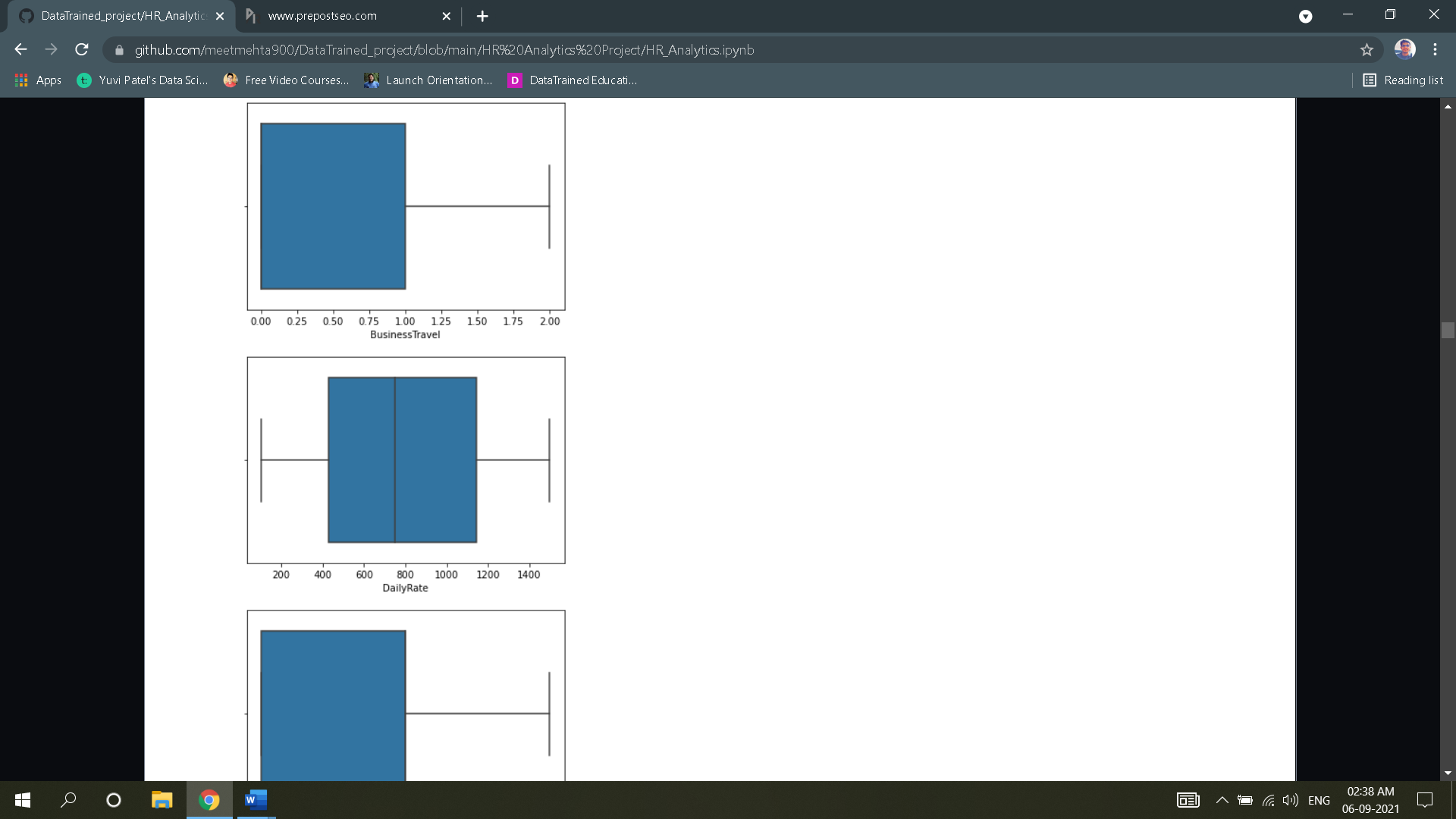
##### Over18*: it contains no datapoints except for 0(y)*

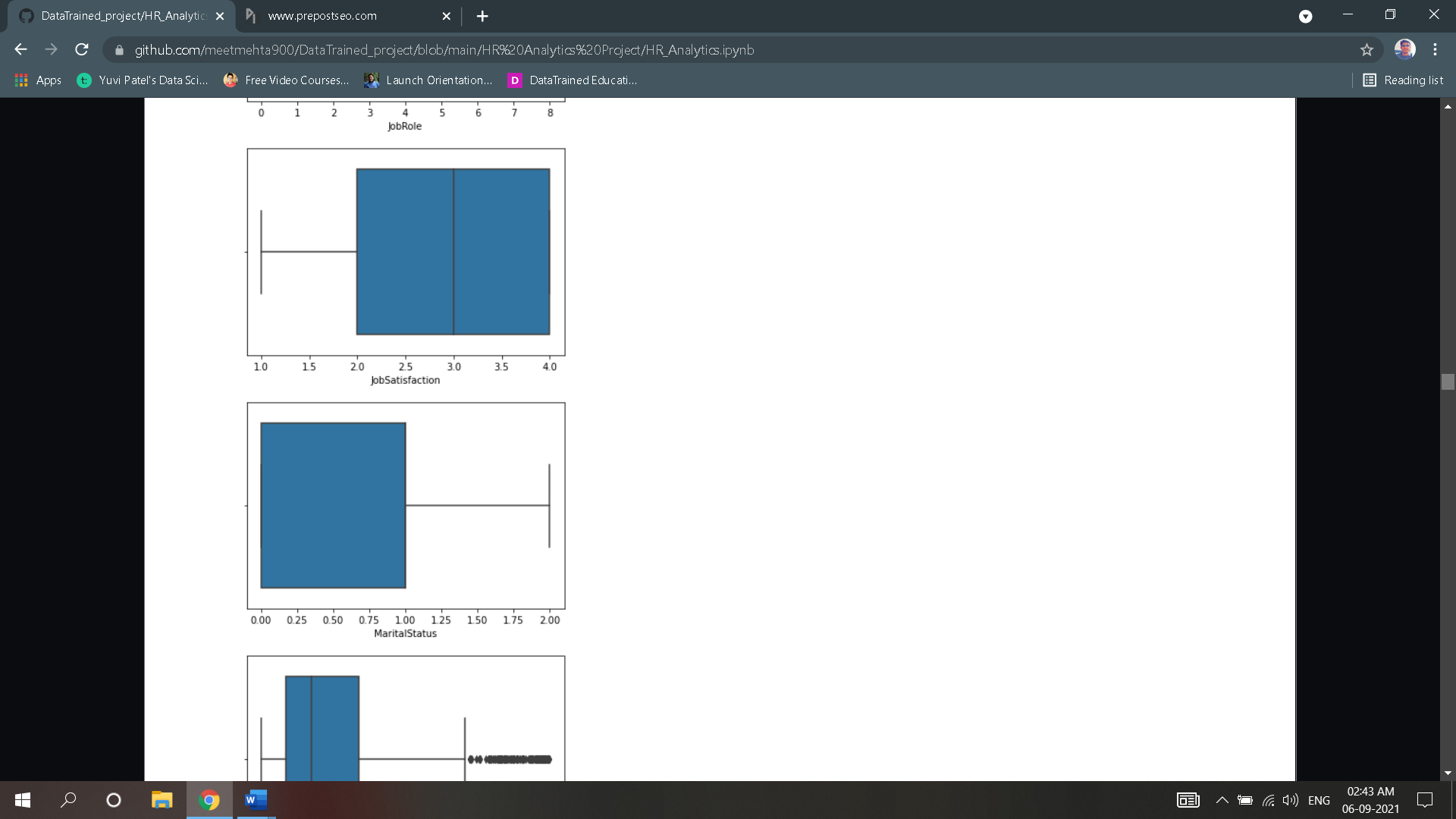
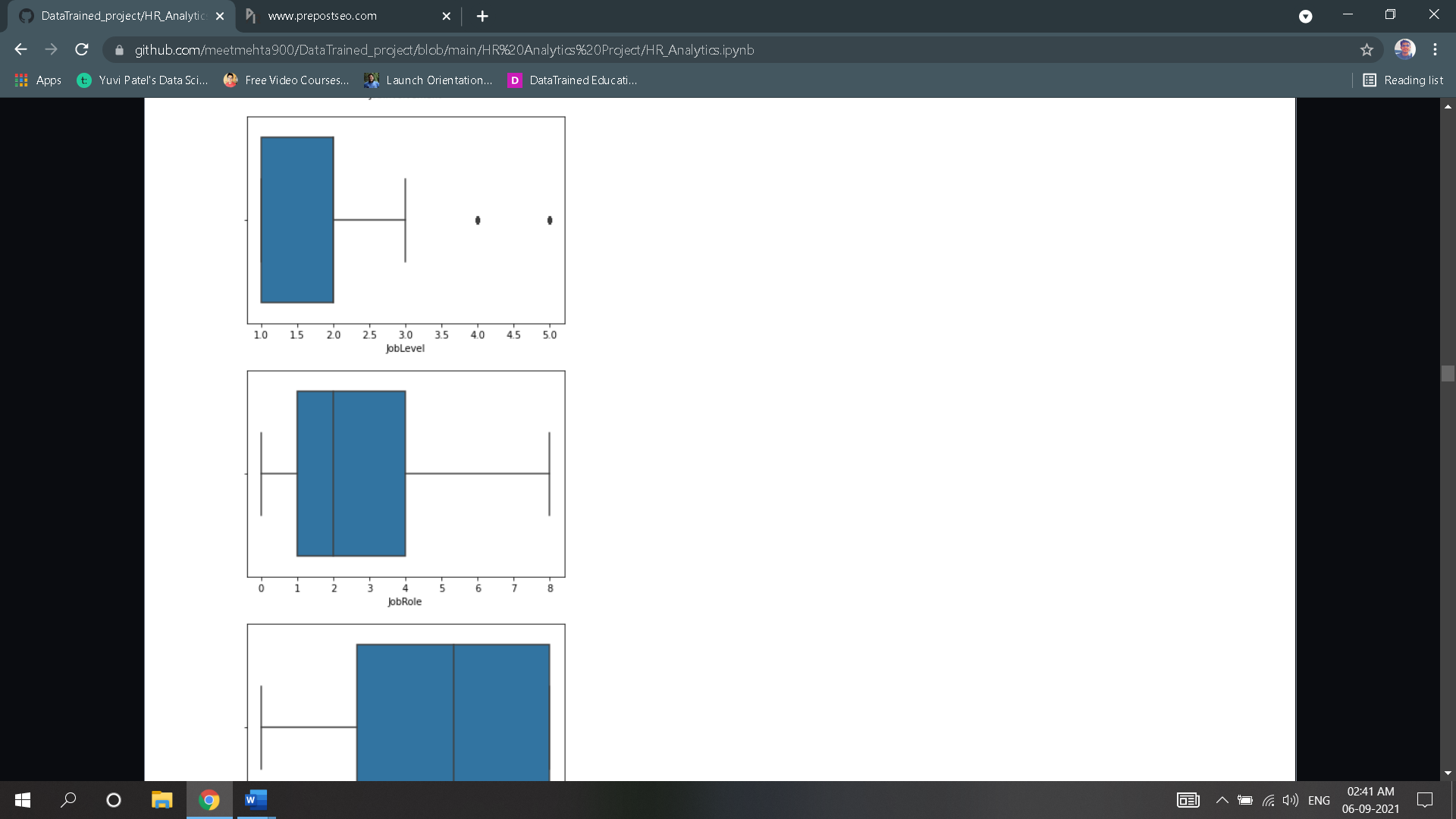
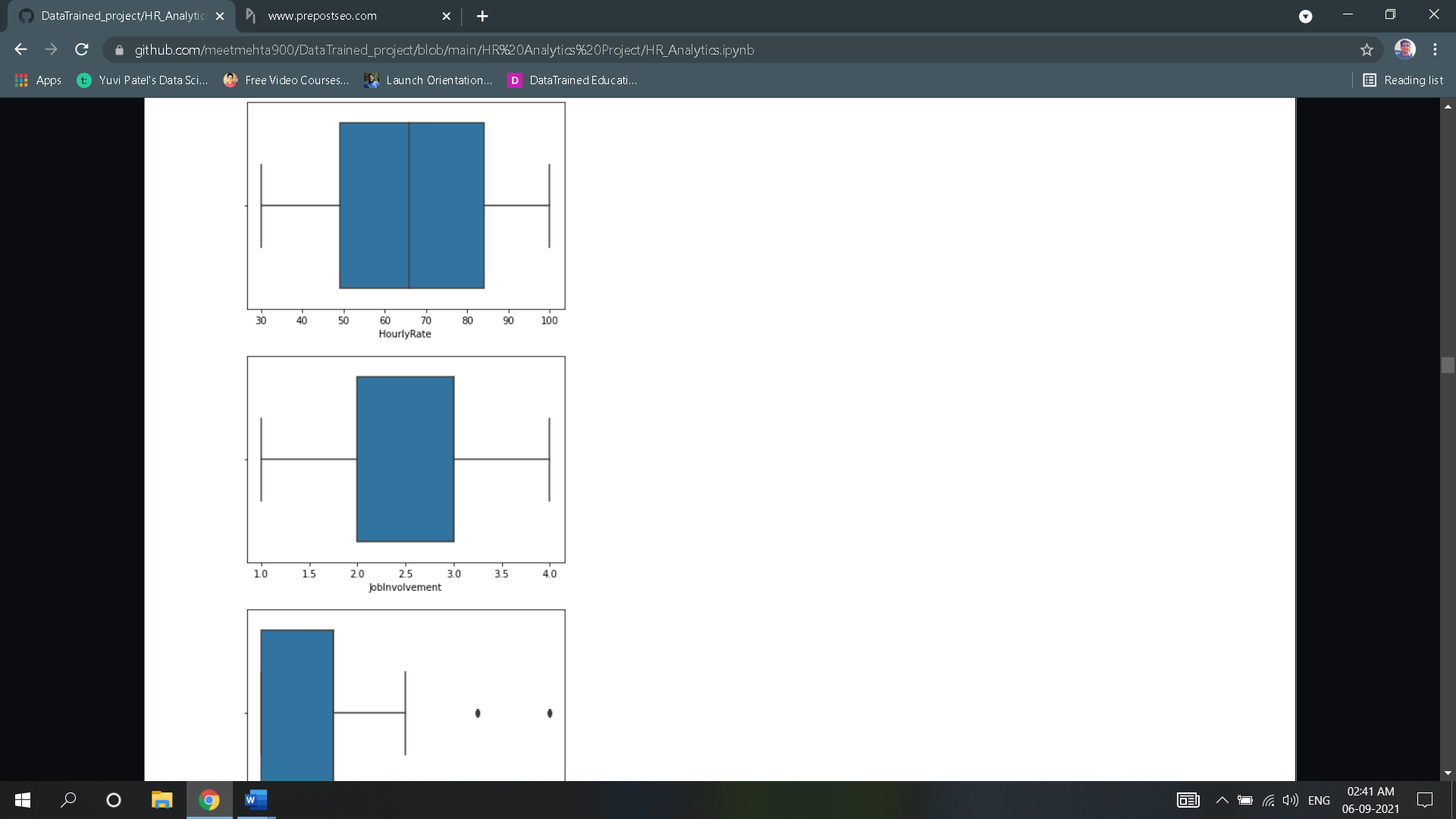
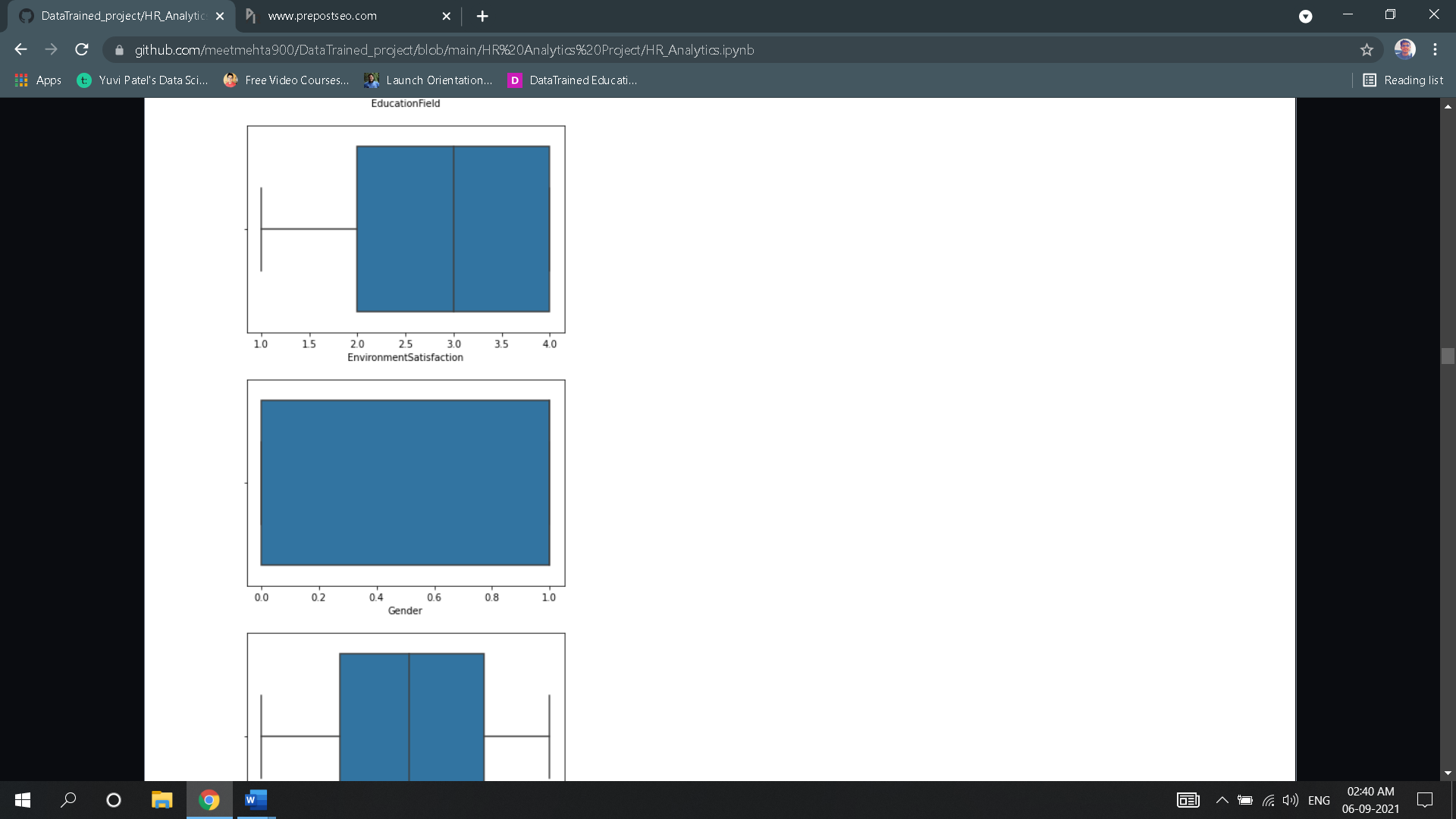
##### Employee Count: it has only one unique value

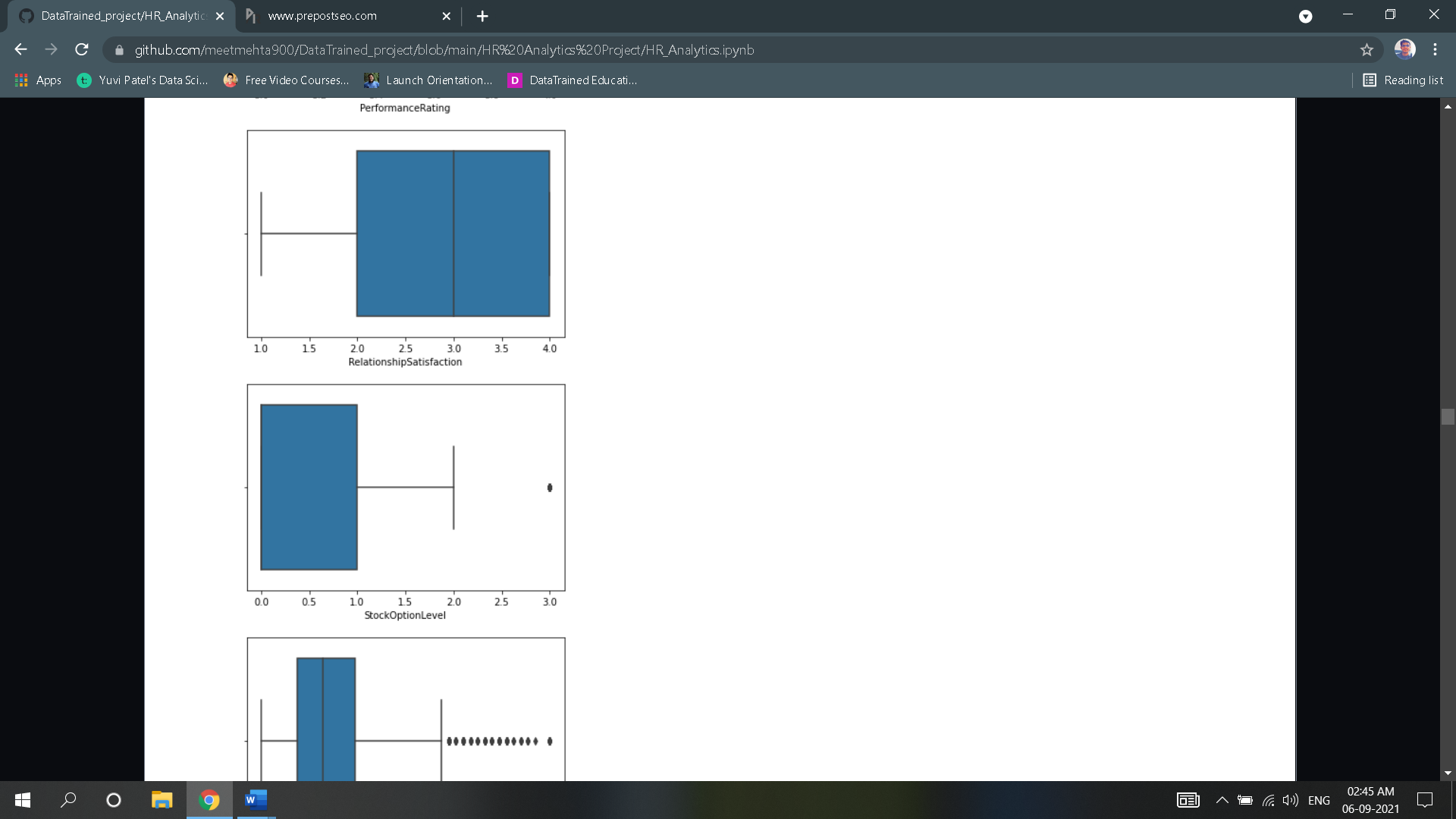
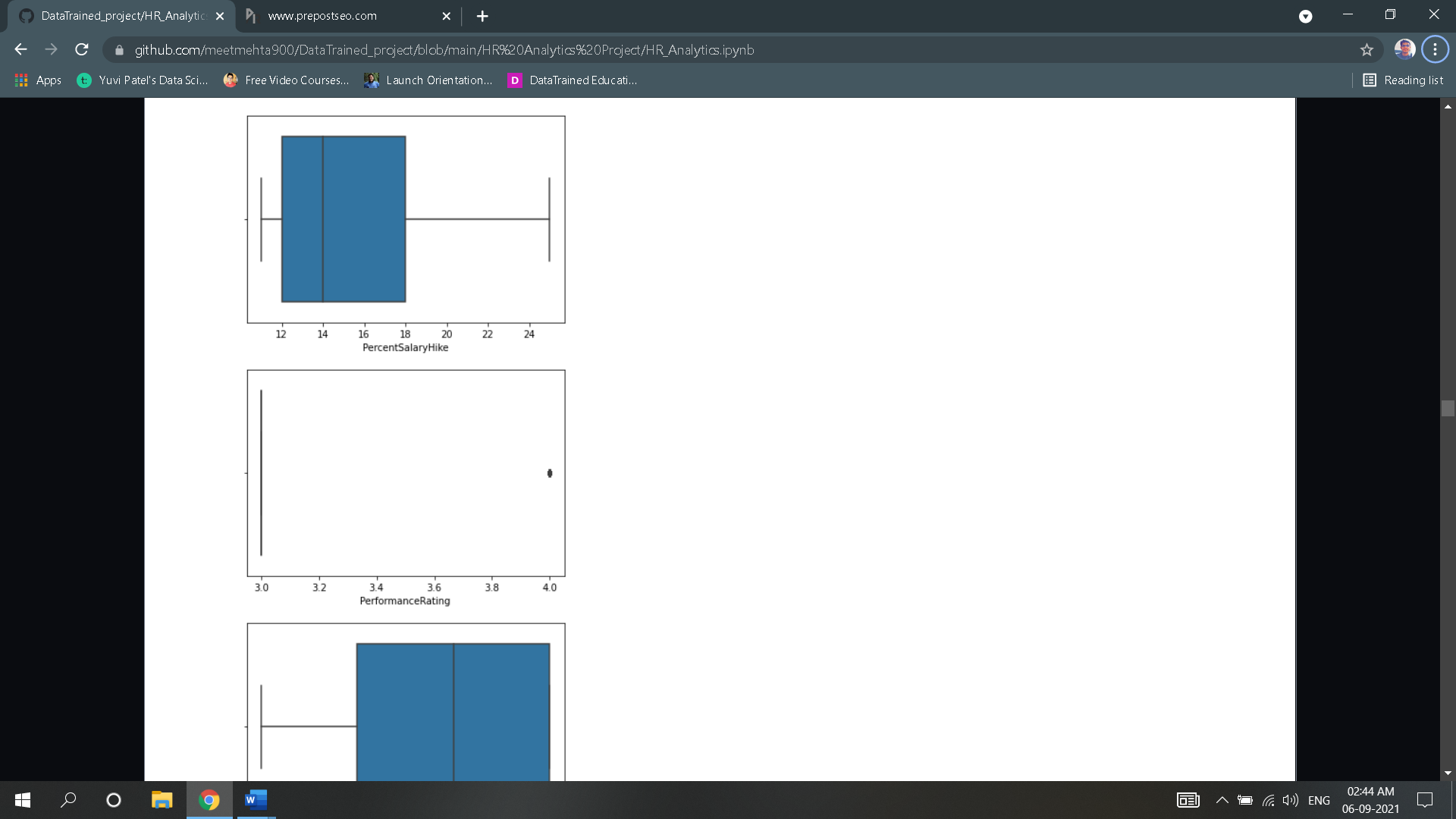
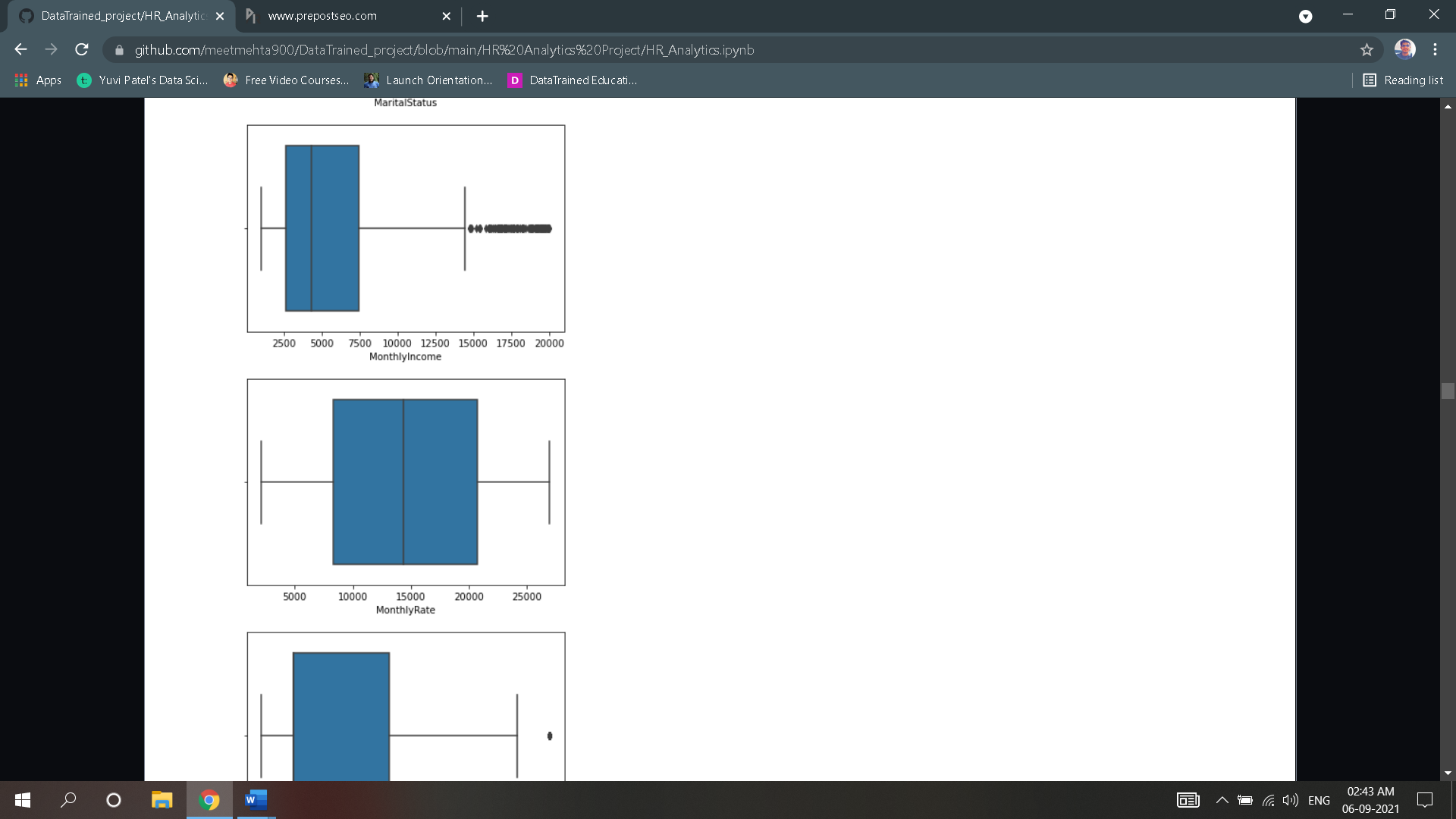
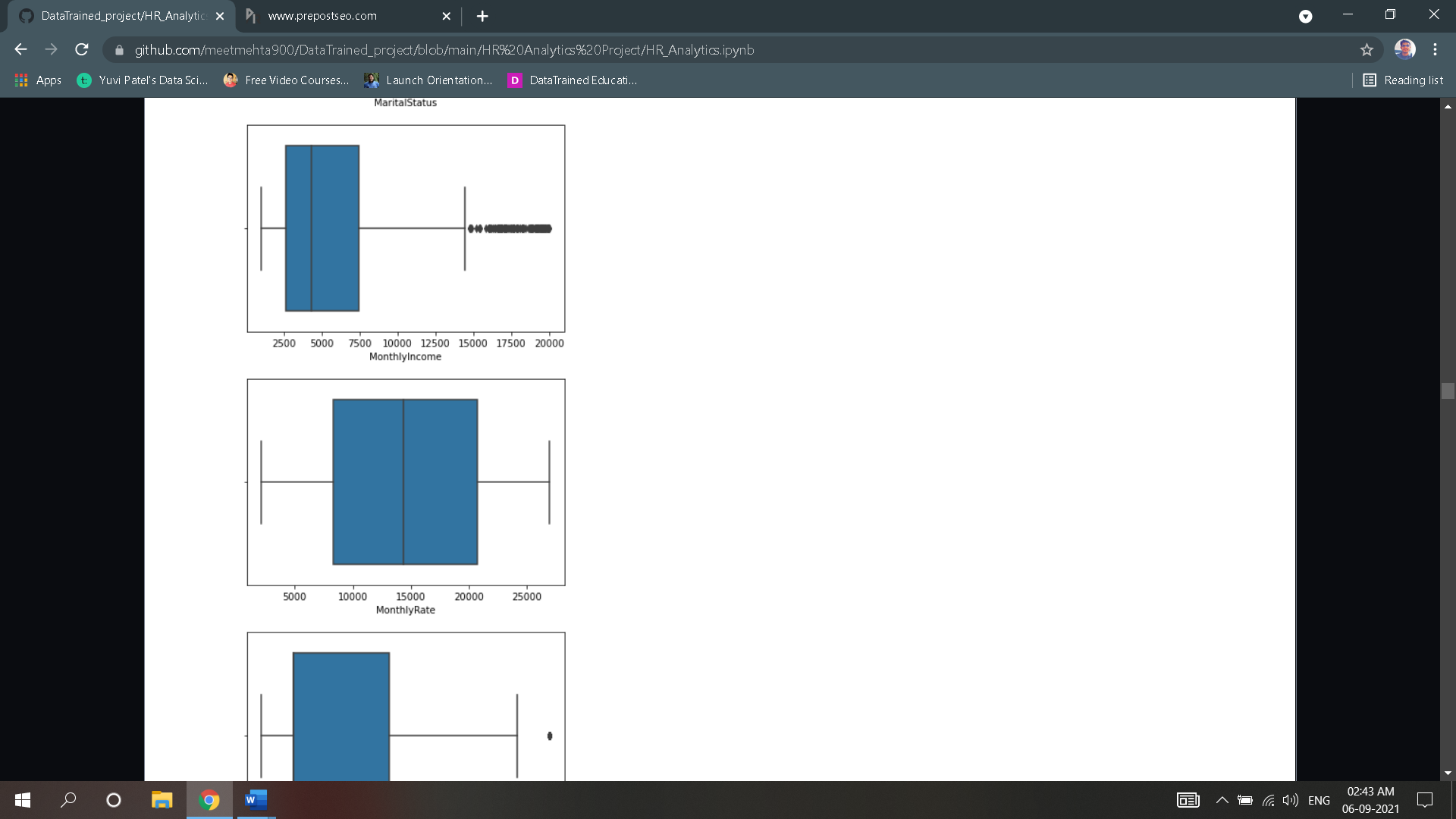
##### Standard hours: it has same hours data

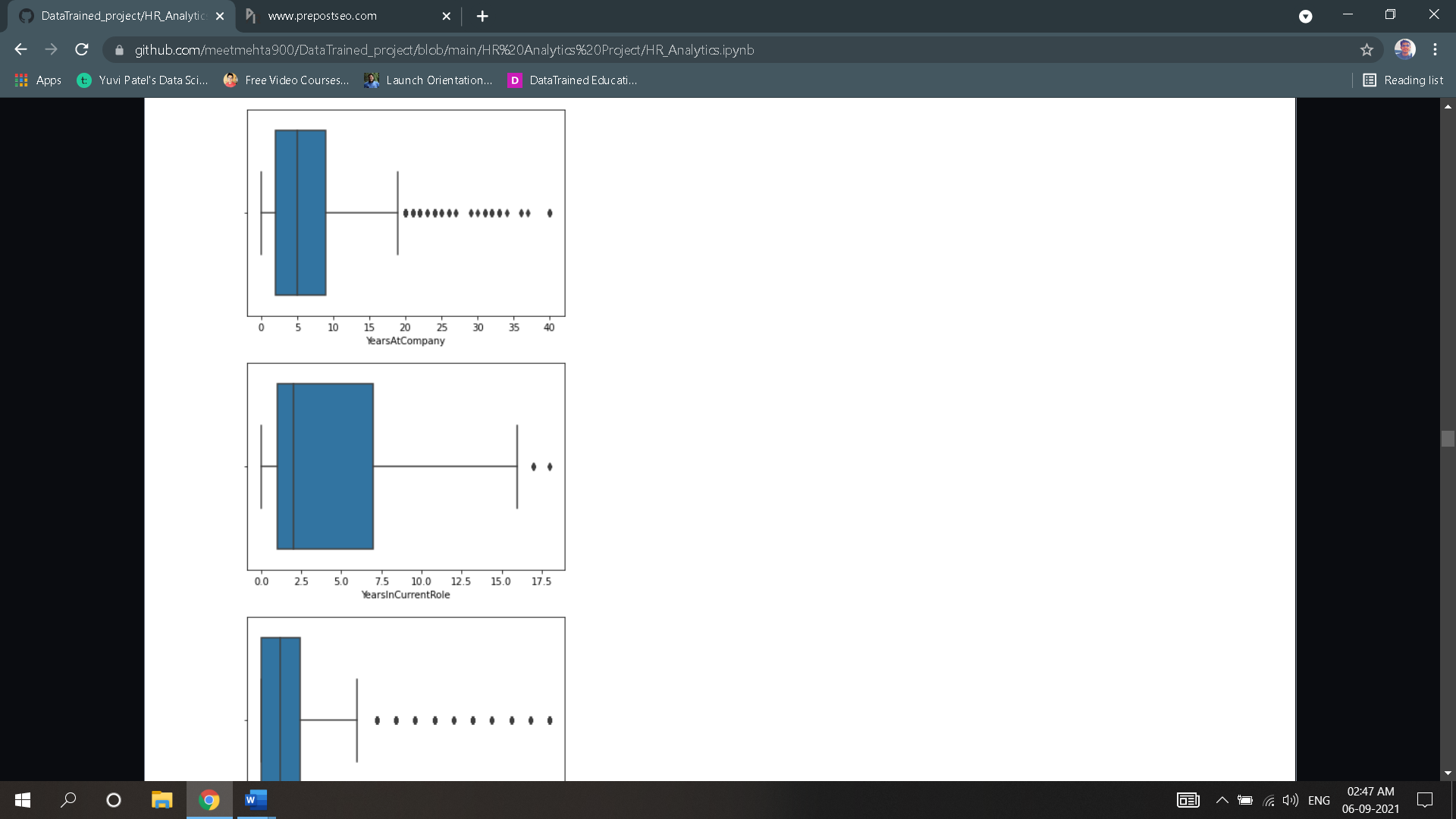
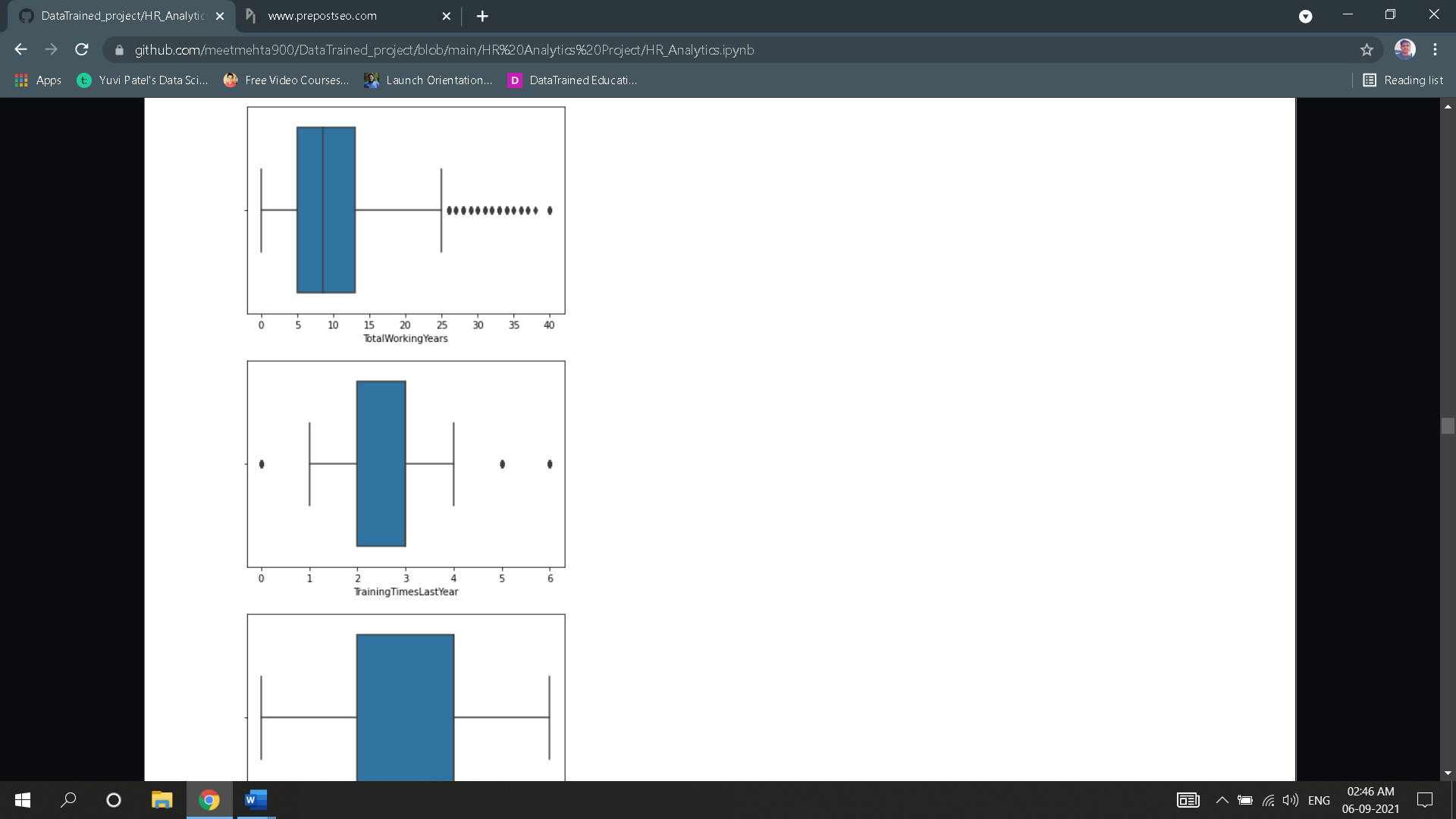
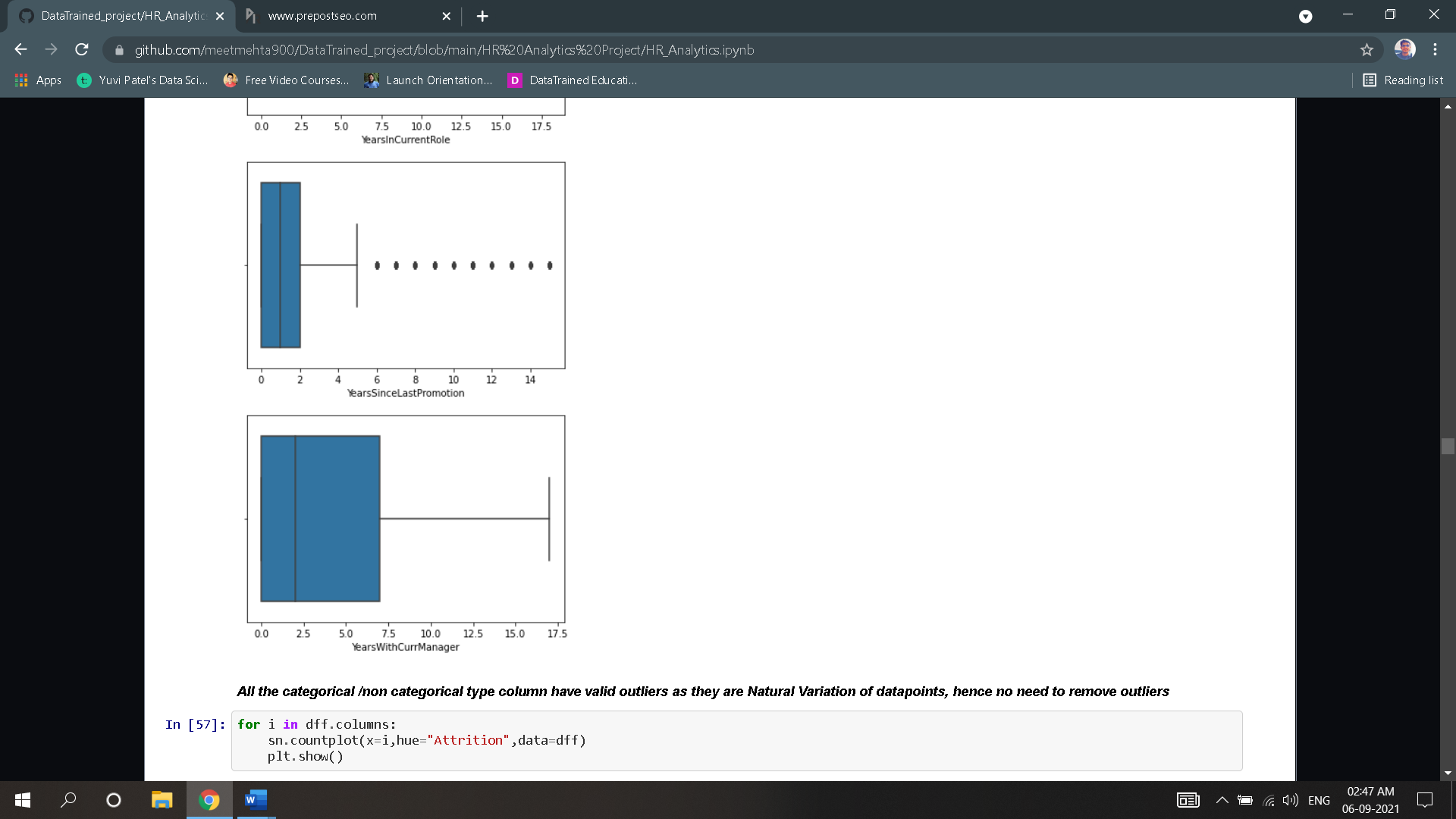
### We must also check for outliers using boxplot in our dataset as it can hamper our model performance.

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







* All the categorical /non categorical type column have valid outliers as they are Natural Variation of datapoints, hence no need to remove outliers.
* Next Step is to check and analyse the correlation of input features w.r.t output column.

##### 

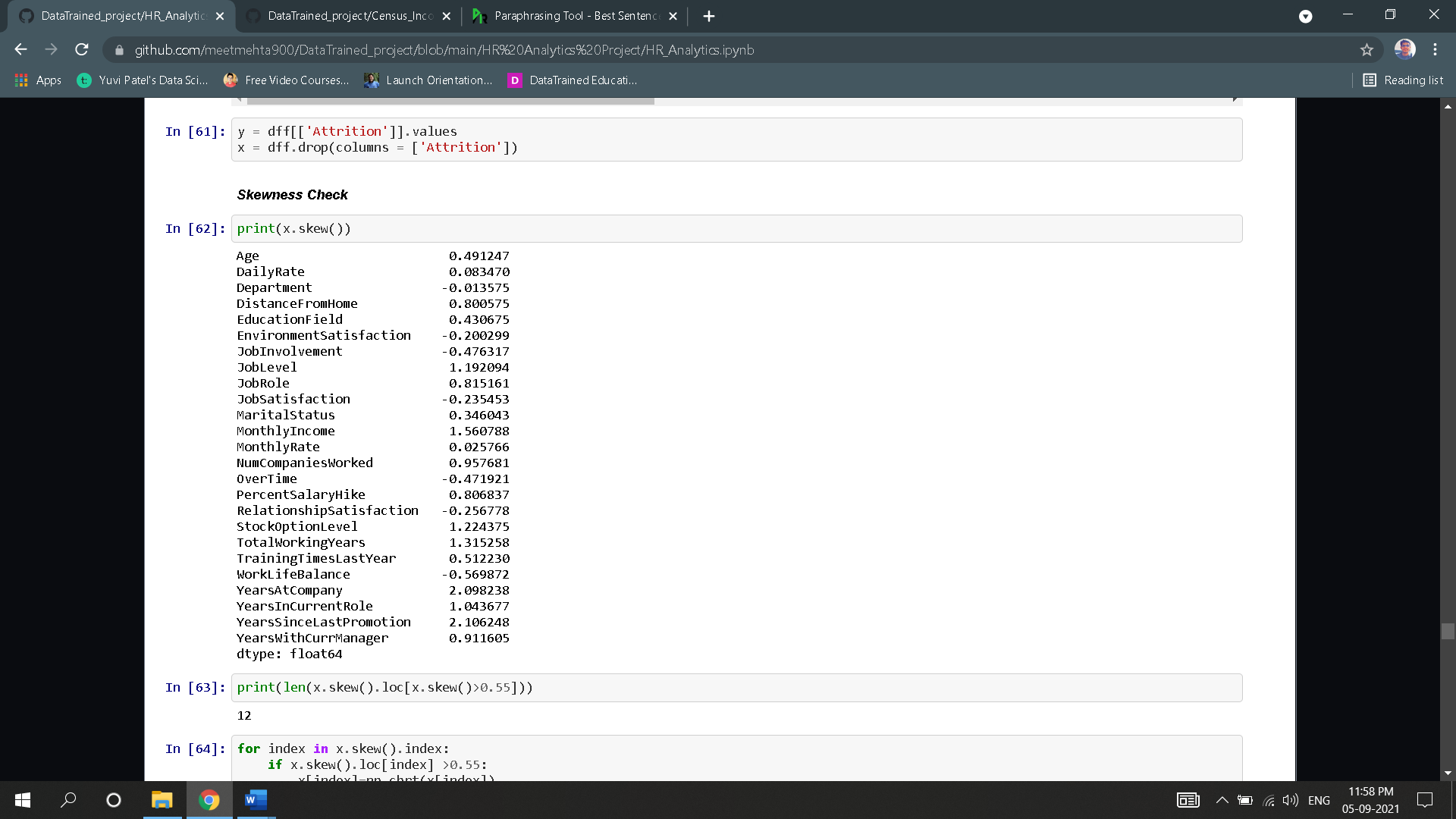
* We can deduce which columns are highly correlated to the target column and which columns are least correlated. We perform these analyses to determine which columns can dropped based on their correlation w.r.t target variable.
* Visualising correlation of the entire dataset using Heatmap.

##### 

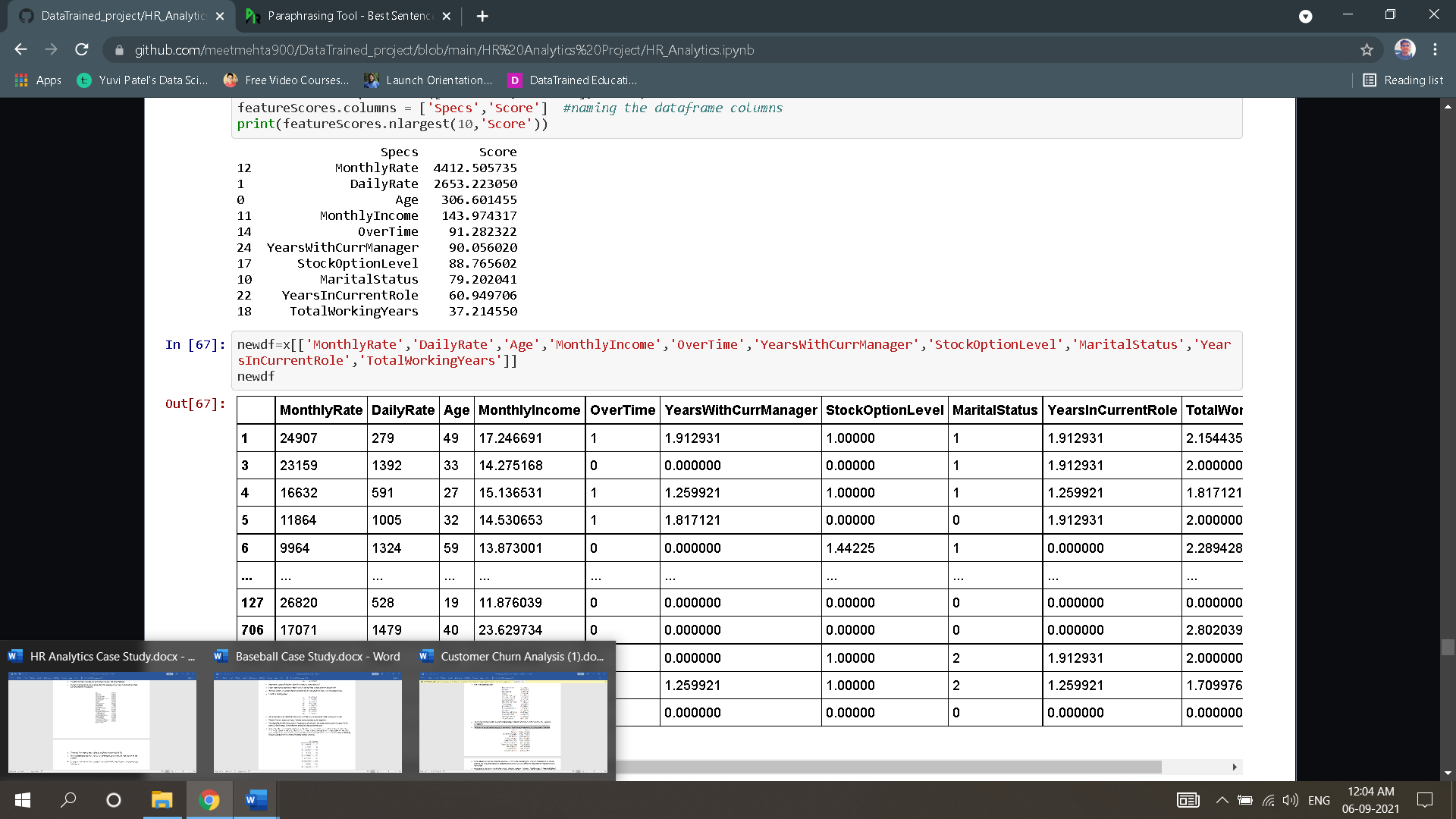
##### According this heatmap and df.corr() we can drop columns with least correlation i.e.Performance rating, Gender, education, Business Travel & hourly rate.

4.Pre-Processing Pipeline

* To train the model we need to split target variable and input features.
* To avoid biasing issue, we need to check for biasing using skew method of input features and remove if required.



* Skewness is present if skew value of a column is more than 0.55.
* Remove biasing of specific column 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 'MonthlyRate','DailyRate','Age','MonthlyIncome','OverTime','YearsWithCurrManager','StockOptionLevel','MaritalStatus','YearsInCurrentRole','TotalWorkingYears' 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.

5. Building Machine Learning Models

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

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

### We shall evaluate 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** | **CV Score** | **Roc Auc score** | **Difference** |
| **Extra Trees Classifier** | **98.5830%** | **98.5829%** | **0.3%** |
| **Random Forest Classifier** | 98.80**%** | 95.74**%** | **3%** |
| **XGBClassifier** | **97.97%** | **93.92%** | **4%** |
| **LGBMClassifier** | **97.74%** | **94.12%** | **3.6%** |
| **Bagging Classifier** | **97.72%** | **93.31%** | **4.4%** |

### 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': [2,8,16,32,50]
* class\_weight':['balanced', 'balanced\_subsample']

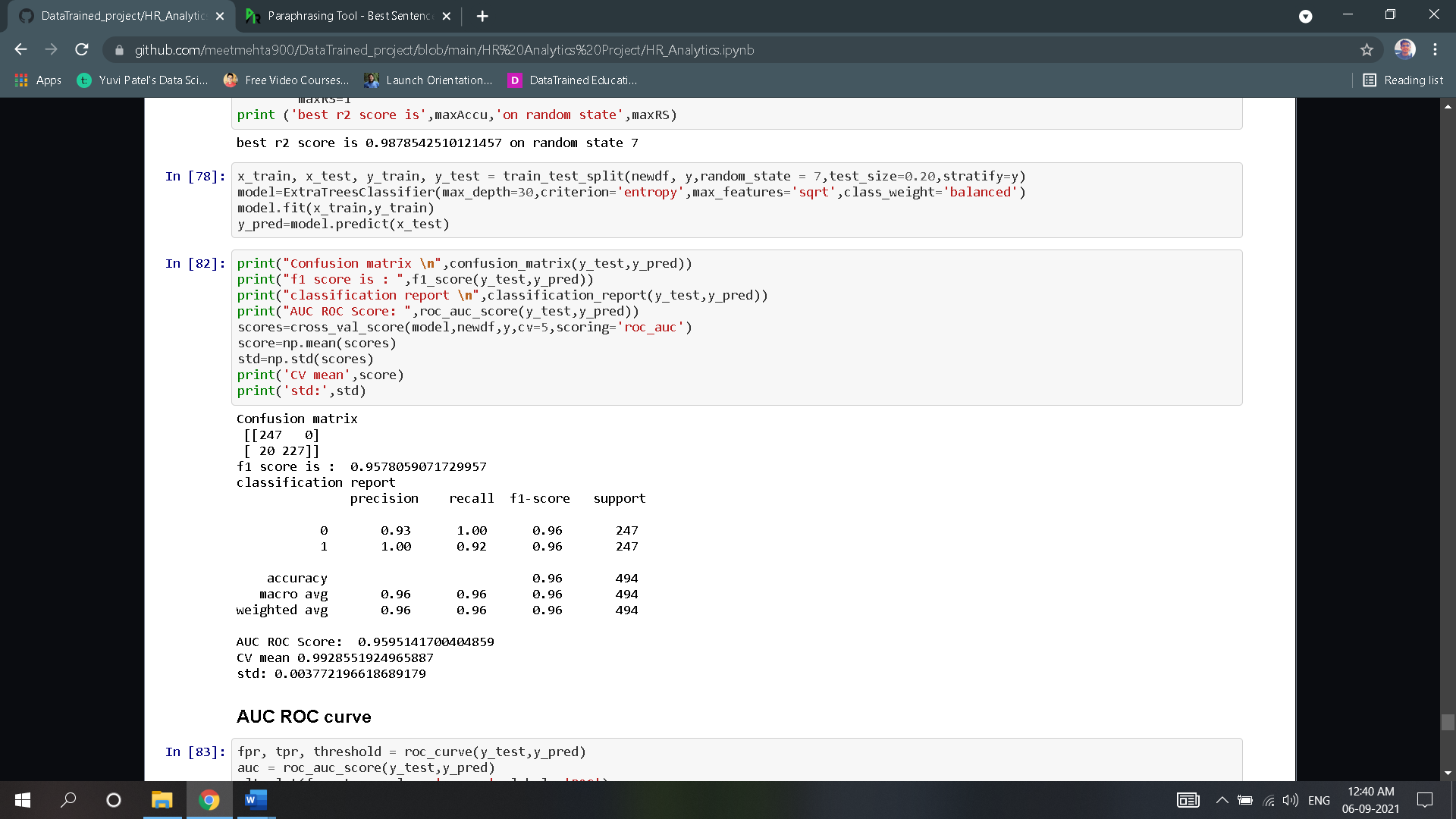
### Best parameters for Extra Trees Classifier obtained by executing Grid Search are:

* 'criterion': entropy
* 'max\_features': 'sqrt'
* 'max\_depth ': 32
* class\_wieght: balanced

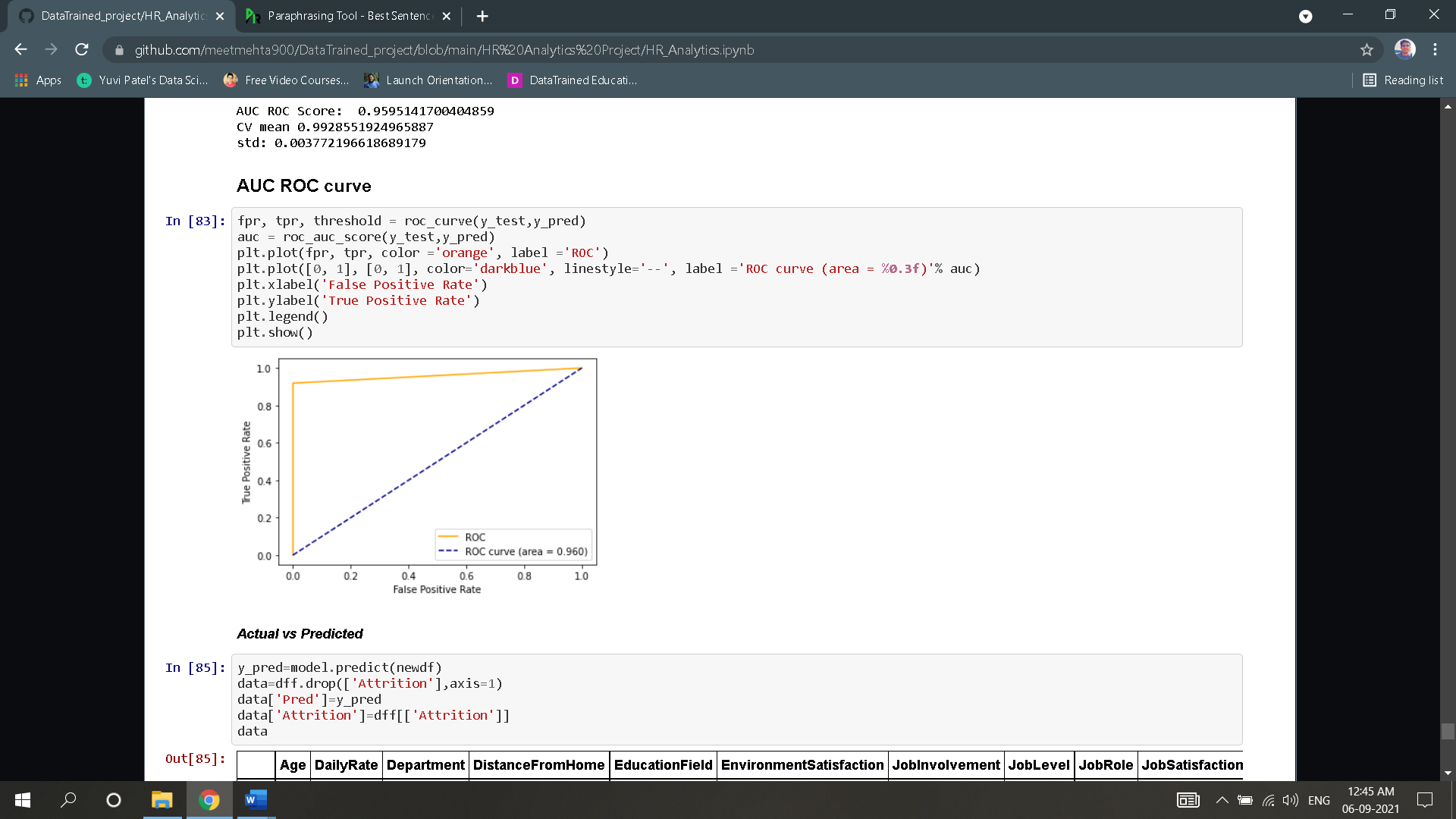
### Now we need to find the most optimum random state in train test split for Extra Trees Classifier model to get best score, in this case best random state is 7.

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

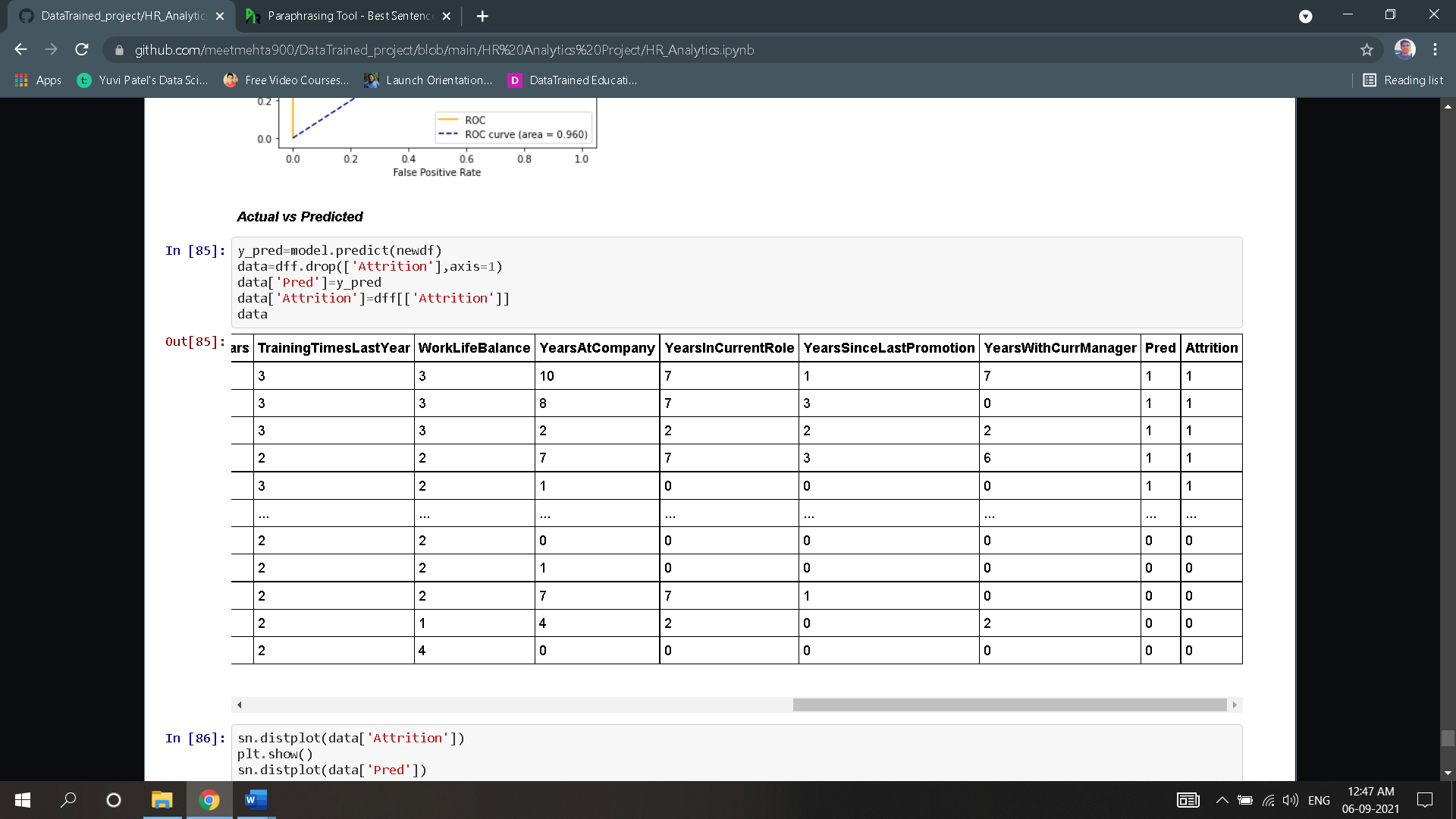
* Using the model extra trees classifier along with the best parameters obtained from grid search, fit the model w.r.t the train dataset.
* Evaluate the model based on test dataset and obtain all the metrics of the currently trained model:



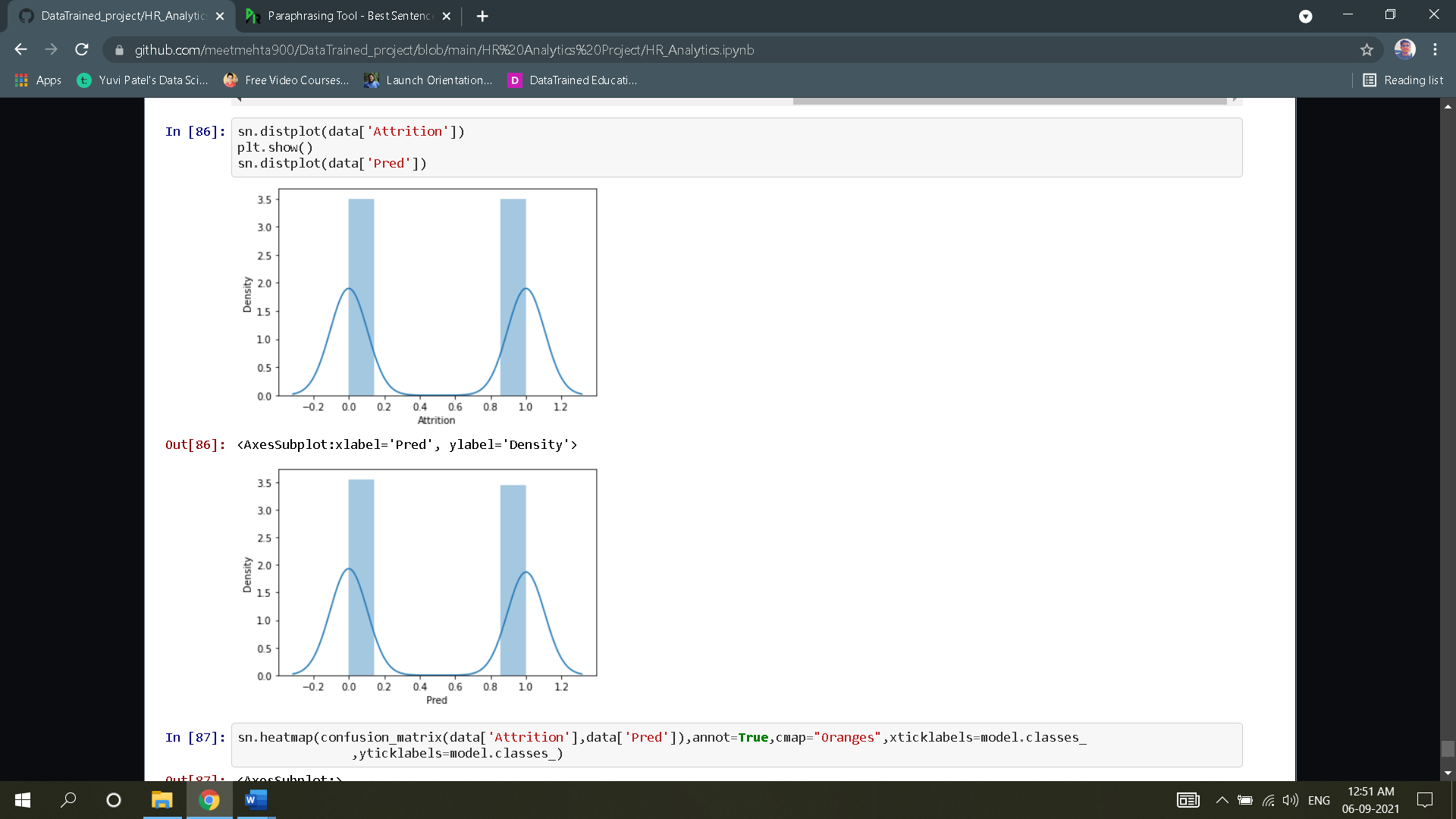
* As observed, the accuracy of the model varies from 95.57% to 95.95%.
* Plot roc curve of the model to visualize the accuracy of the model.



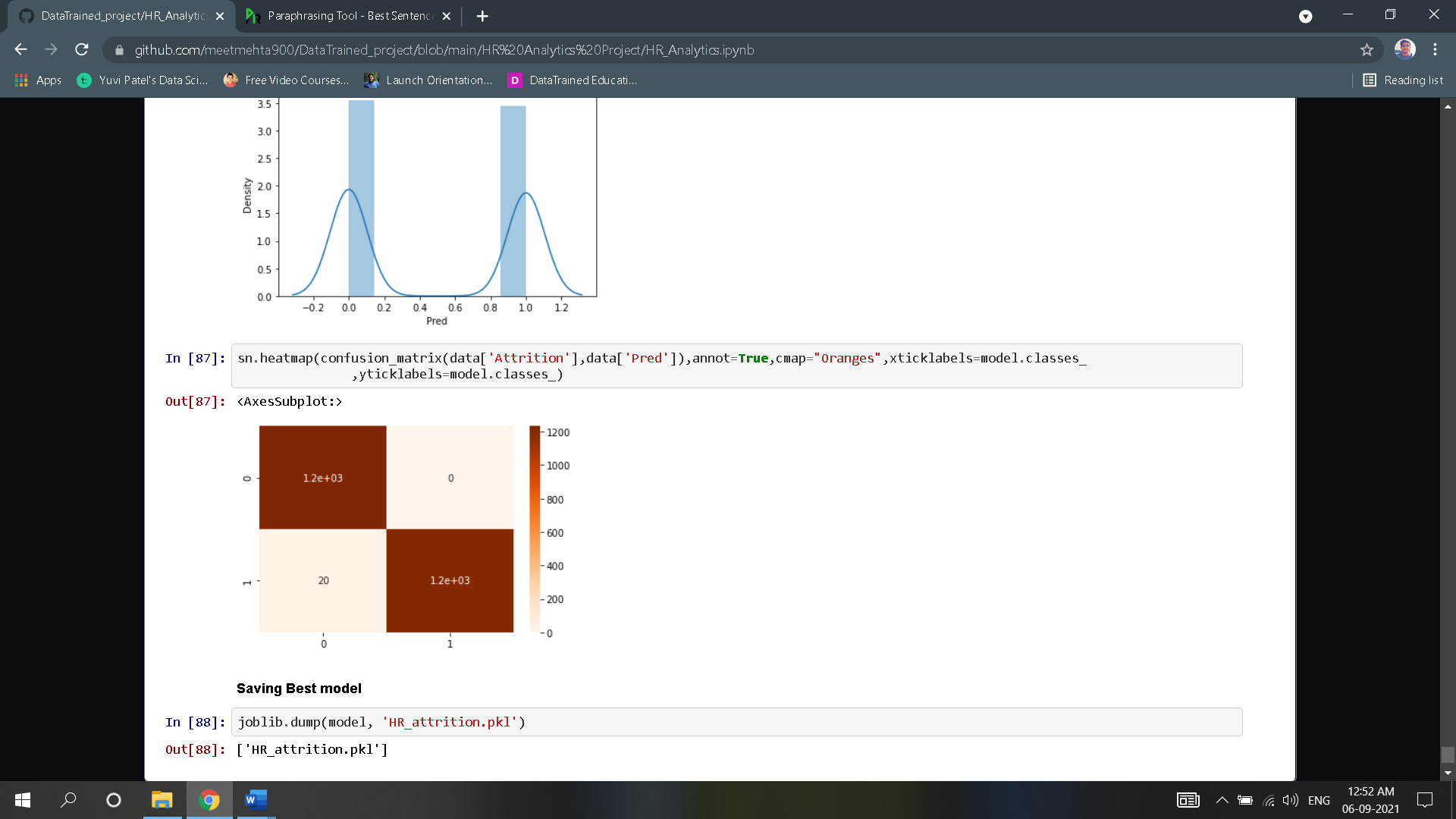
### Put target variable ‘Attrition’ column & predicted “Pred” column side by side to observe the difference between the actual and predicted datapoints.



* Unfortunately, we cannot observe the difference between the actual and predicted because the dataset is too large for us to go through the values, hence plot appropriate graphs to check the similarity between actual and predicted values.
* First 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:0**
* **Type II error:20**
* This means the model we build has only guessed 20 False Negative Values.
* Finally, we save the model using **“joblib**” library which can be reused for further prediction.
* 6. Concluding remarks

### We were able to create a model that was accurate to 96 % to predict the rate of attrition 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. By further visualizing the data points, we can conclude that the model can accurately predict the rate of attrition. The saved model can be loaded and reused to predict the rate of Attrition. 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/HR%20Analytics%20Project/HR_Analytics.ipynb>