Follow the given github link of connecting my project

<https://github.com/asiwach08/ML-Projects/blob/main/Project_2-HR_Analytics.ipynb>

**Application of Machine Learning in HR attrition :-**

Employee attrition refers to the percentage of workers who leave an organization and are replaced by new employees. A high rate of attrition in an organization leads to increased recruitment, hiring and training costs. Not only it is costly, but qualified and competent replacements are hard to find. In most industries, the top 20% of people produce about 50% of the output.

Attrition rate is measured as a percentage of the employees that left a business for any reason over a specific period. This periodic measure determines how successfully a company is retaining their employees.

if your company is doing poorly, a high attrition rate may simply be a result of a strategy of cost-cutting through downsizing. Generally, however, a low attrition rate means that your company does well at keeping your employees happy and your workplace running smoothly.

Every data has a lot of hidden information related to the topic the dataset set is designes for .This hidden information was required to be investigated to find out the hidden patterns which can be analysed. These patterns can be helpful in making decisions on the procedure,removal of any ambiguity and also in getting key business insights.To solve all these questions exploratory data analysis was introduced.

This use case takes HR data and uses machine learning models to predict what employees will be more likely to leave given some attributes. Such model would help an organization predict employee attrition and define a strategy to reduce such costly problem.

**Introduction:-**

Exploratory data analysis(EDA) is all about getting and overall understanding of data .It is mainly done to find it’s properties ,patterns and visualizations .It helps us to assure that our data is correct and ready to use for machine learning algorithms. In this blog we are using python(Jupyter Notebook) as our programming language for the analysis purpose.Python has a wide variety of libraries like pandas ,seaborn, numpy, matplotlib which can be used for this purpose.We are using the machine learning dataset where we are going to find the attrition of employees.

Exploratory data analysis is a preprocessing step to explore the data. There are numerous method and steps in performing EDA including visulalization,wrangling ,cleaning ,etc, however most of them are specific ,focusing on either visualization or distribution and are incomplete.Therefore ,here ,I will take you through step-by-step to understand ,explore and extract the information from the data to answer the question or assumptions. There are no structured steps or methods to follow ,however this project will provide insights on EDA.

**Brief introduction to the used libraries:-**

* *Numpy -* NumPy is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very useful for fundamental scientific computations in Machine Learning. It is particularly useful for linear algebra, Fourier transform, and random number capabilities.
* *Pandas -* Pandas is a popular Python library for data analysis. It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation. It provides high-level data structures and wide variety tools for data analysis. It provides many inbuilt methods for groping, combining and filtering data.
* *Matplotlib -* Matpoltlib is a very popular Python library for data visualization. Like Pandas, it is not directly related to Machine Learning. It particularly comes in handy when a programmer wants to visualize the patterns in the data. It is a 2D plotting library used for creating 2D graphs and plots. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, font properties, formatting axes, etc. It provides various kinds of graphs and plots for data visualization, viz., histogram, error charts, bar chats, etc.
* *Seaborn -* Seaborn is a Python data visualization library based on [matplotlib](https://matplotlib.org/). It provides a high-level interface for drawing attractive and informative statistical graphics.

# **Understanding the Attrition in HR:-**

**HR Analytics**

Human resource analytics (HR analytics) is an area in the field of analytics that refers to applying analytic processes to the human resource department of an organization in the hope of improving employee performance and therefore getting a better return on investment. HR analytics does not just deal with gathering data on employee efficiency. Instead, it aims to provide insight into each process by gathering data and then using it to make relevant decisions about how to improve these processes.

**Attrition in HR**

Attrition in human resources refers to the gradual loss of employees overtime. In general, relatively high attrition is problematic for companies. HR professionals often assume a leadership role in designing company compensation programs, work culture, and motivation systems that help the organization retain top employees.

**Attributes information:-**

Age int64

Attrition object

BusinessTravel object

DailyRate int64

Department object

DistanceFromHome int64

Education int64

EducationField object

EmployeeCount int64

EmployeeNumber int64

EnvironmentSatisfaction int64

Gender object

HourlyRate int64

JobInvolvement int64

JobLevel int64

JobRole object

JobSatisfaction int64

MaritalStatus object

MonthlyIncome int64

MonthlyRate int64

NumCompaniesWorked int64

Over18 object

OverTime object

PercentSalaryHike int64

PerformanceRating int64

RelationshipSatisfaction int64

StandardHours int64

StockOptionLevel int64

TotalWorkingYears int64

TrainingTimesLastYear int64

WorkLifeBalance int64

YearsAtCompany int64

YearsInCurrentRole int64

YearsSinceLastPromotion int64

YearsWithCurrManager int64

**Problem Statement:**

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?

**The Outline of EDA is as follows:-**

**1. Import the data**

**\*** Importing the data using libraries

\* Loading the data

\* checking head and tails of data

**2. Checking the data**

\* includes -

\* checking number of rows and columns

\* checking data types

\* checking the information of features

\* checking null values and duplicates values

\* statistics summary

**3. Visualization**

**\*** target data and gender data distribution

\* bar plot of gender based on attrition

\* boxplot of marital status with attrition

\* factorplot on jobrole

\* boxplot for checking outliers

\* histogram for skewness in data

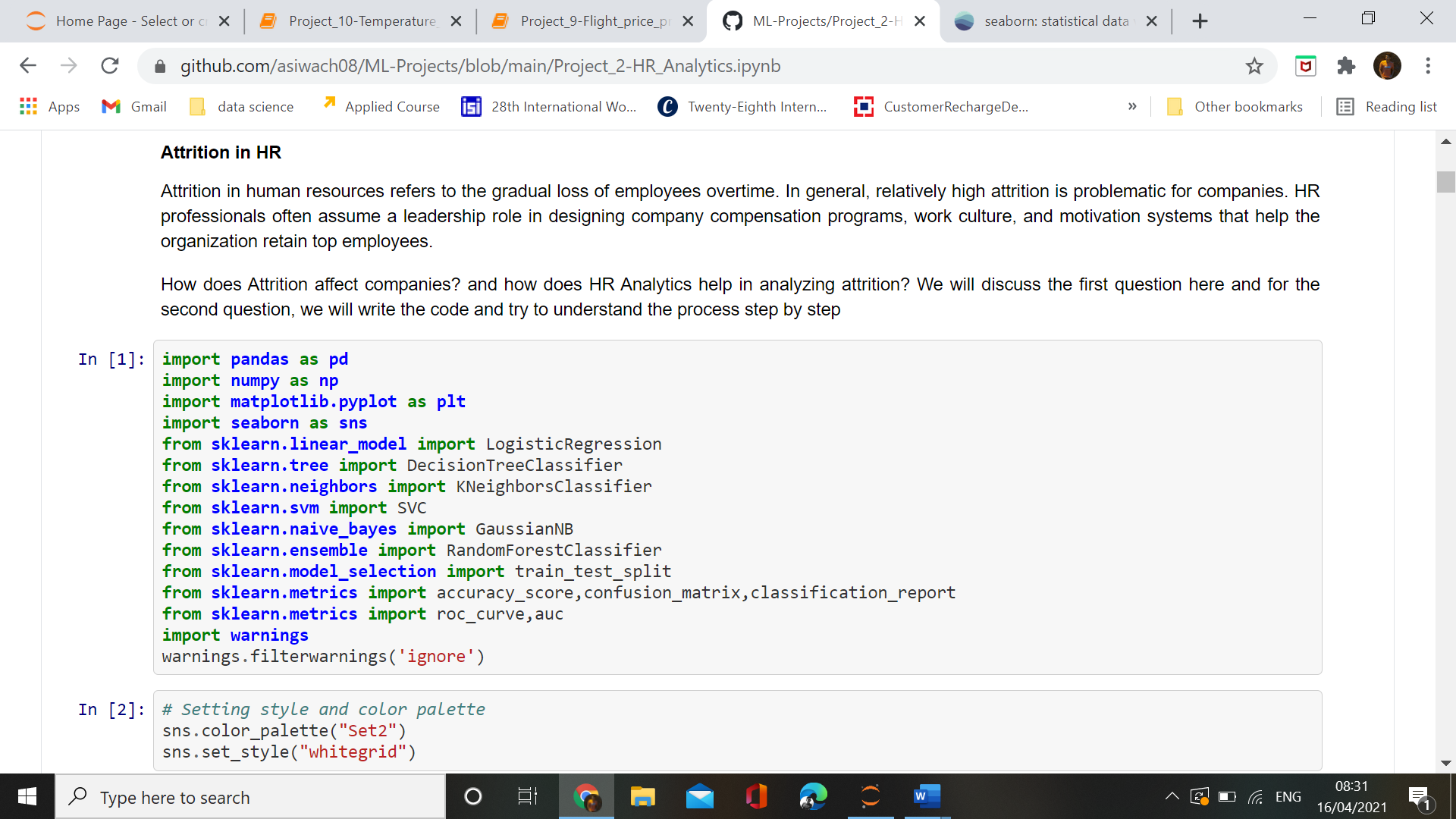
\* heatmap correlation graph for correlation in features

**4. Splitting train and test data and model prepration**

**5. Conclusion**

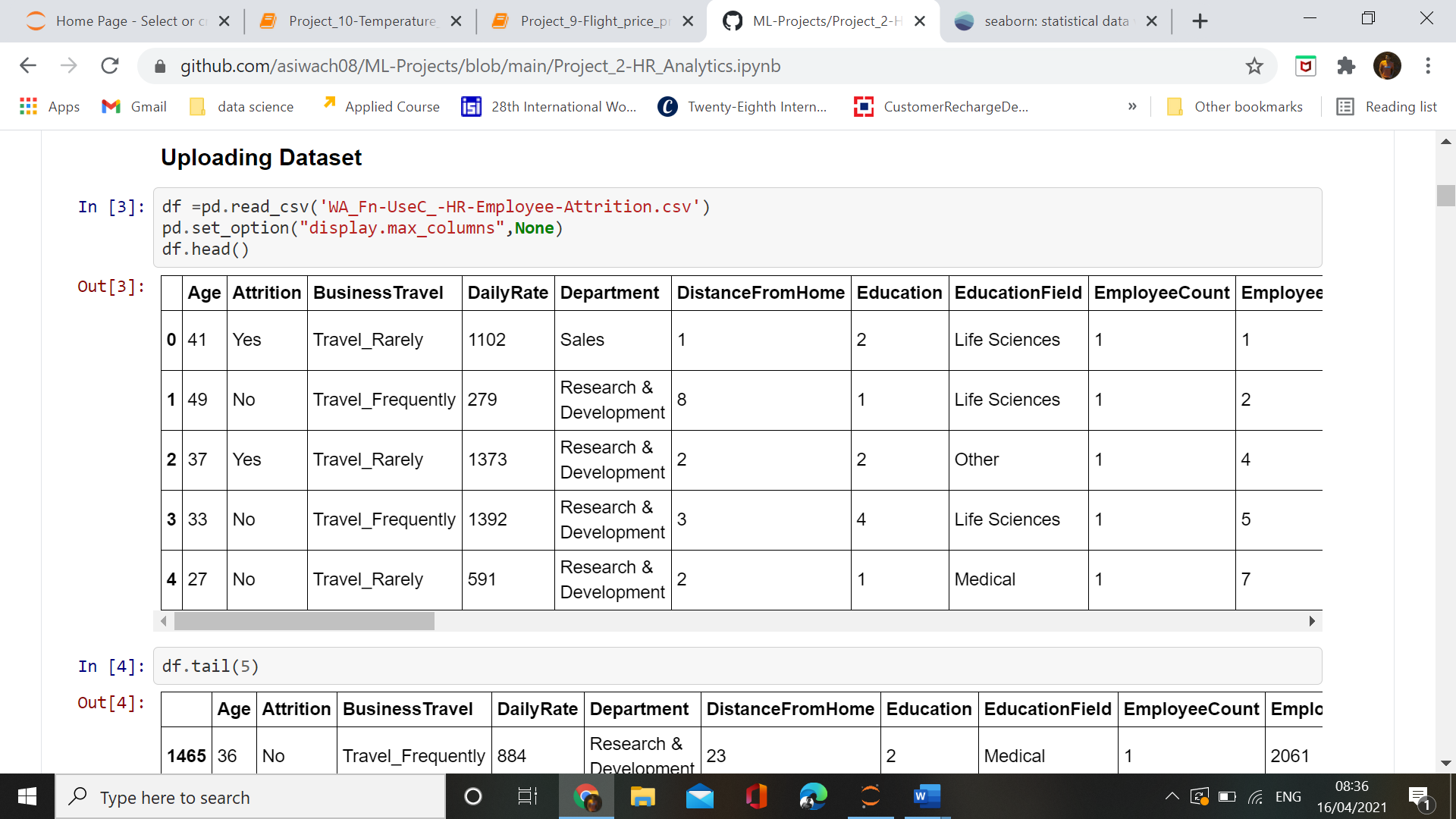
Let’s get started…!!!

**Importing required libraries**



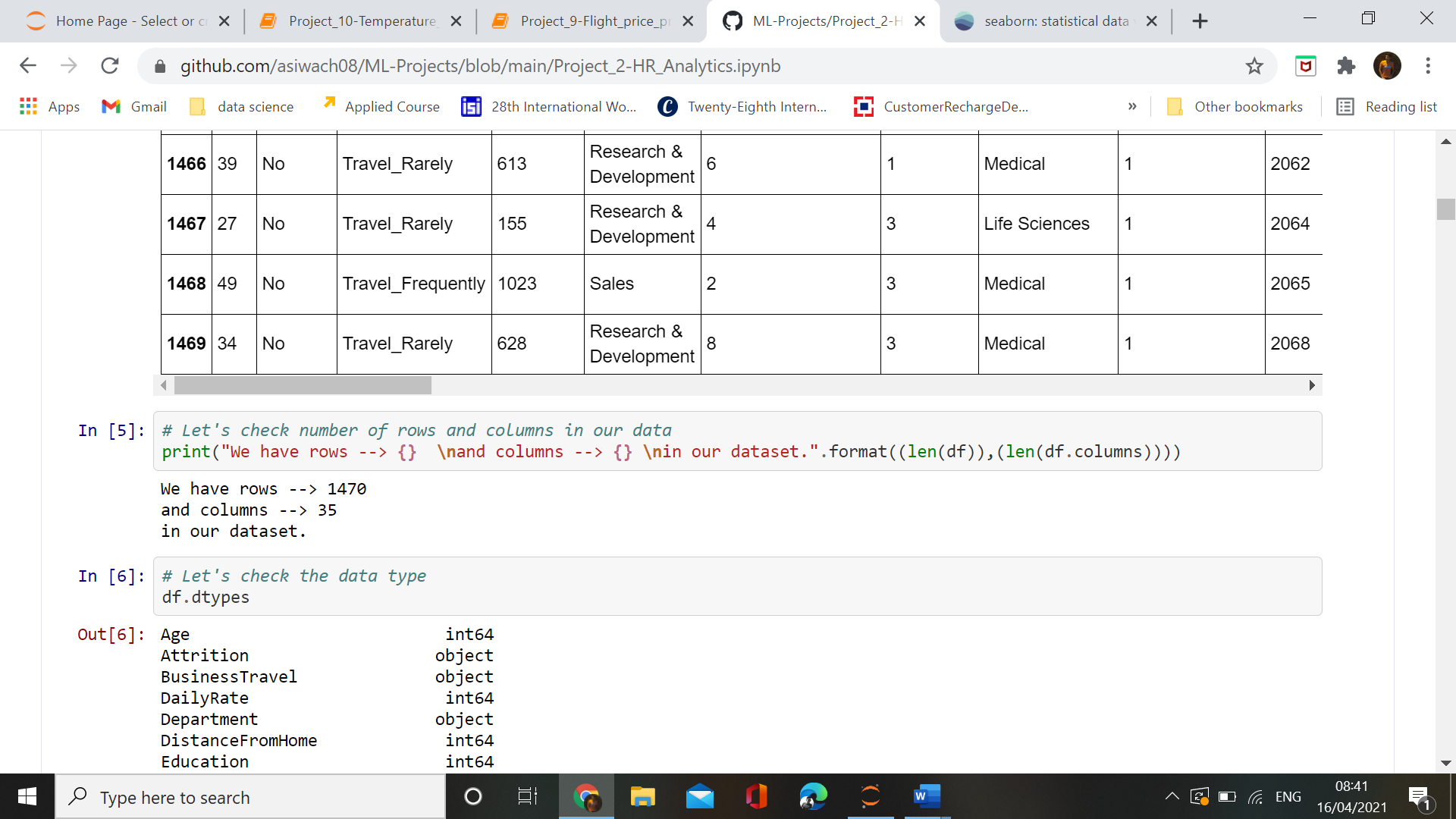
All libraries are imported

**Uploading dataset and checking head and rows**



We have loaded the dataset, as dataset contains more features so we have used display maximum columns.

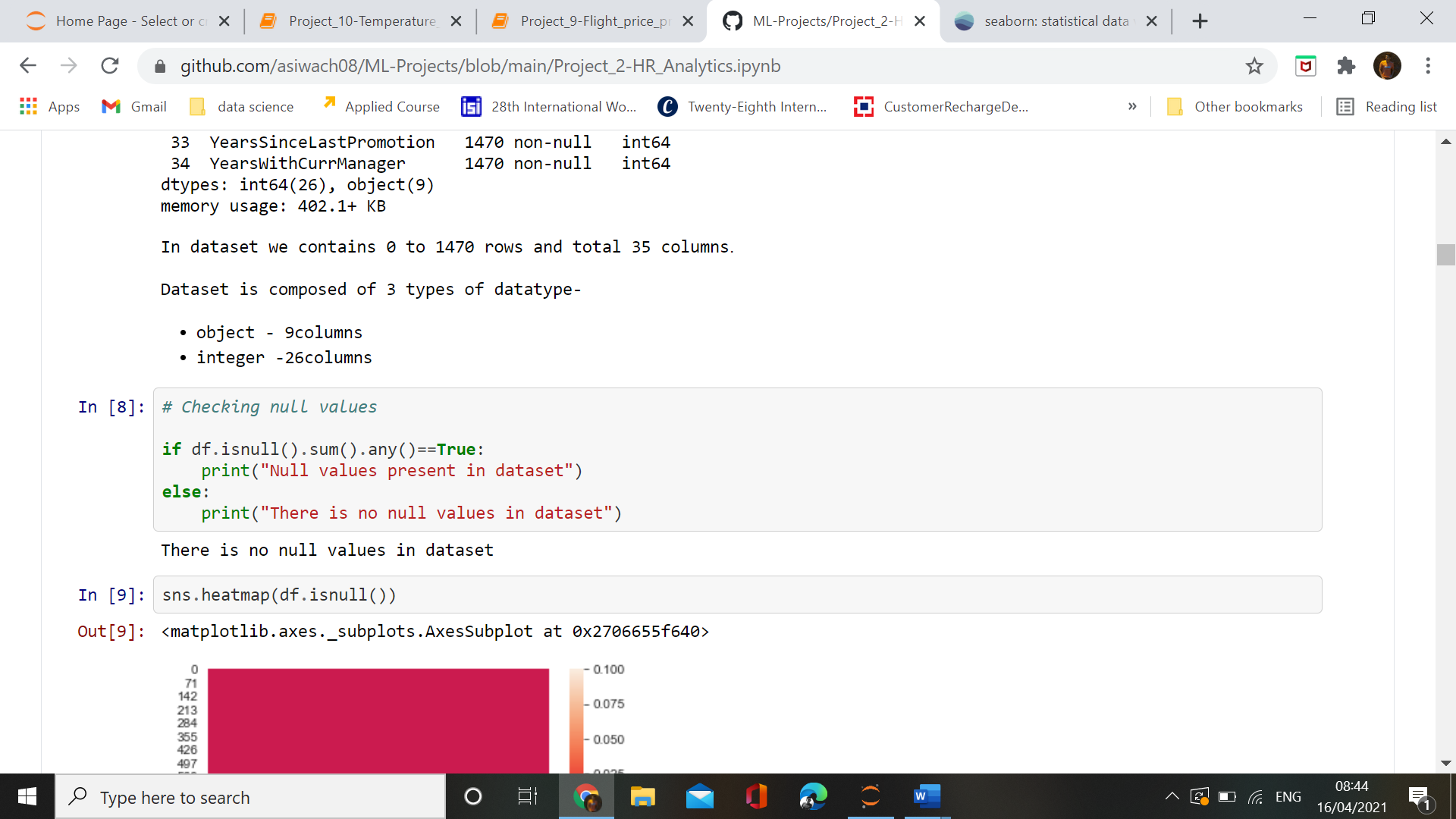
**Checking rows and columns**

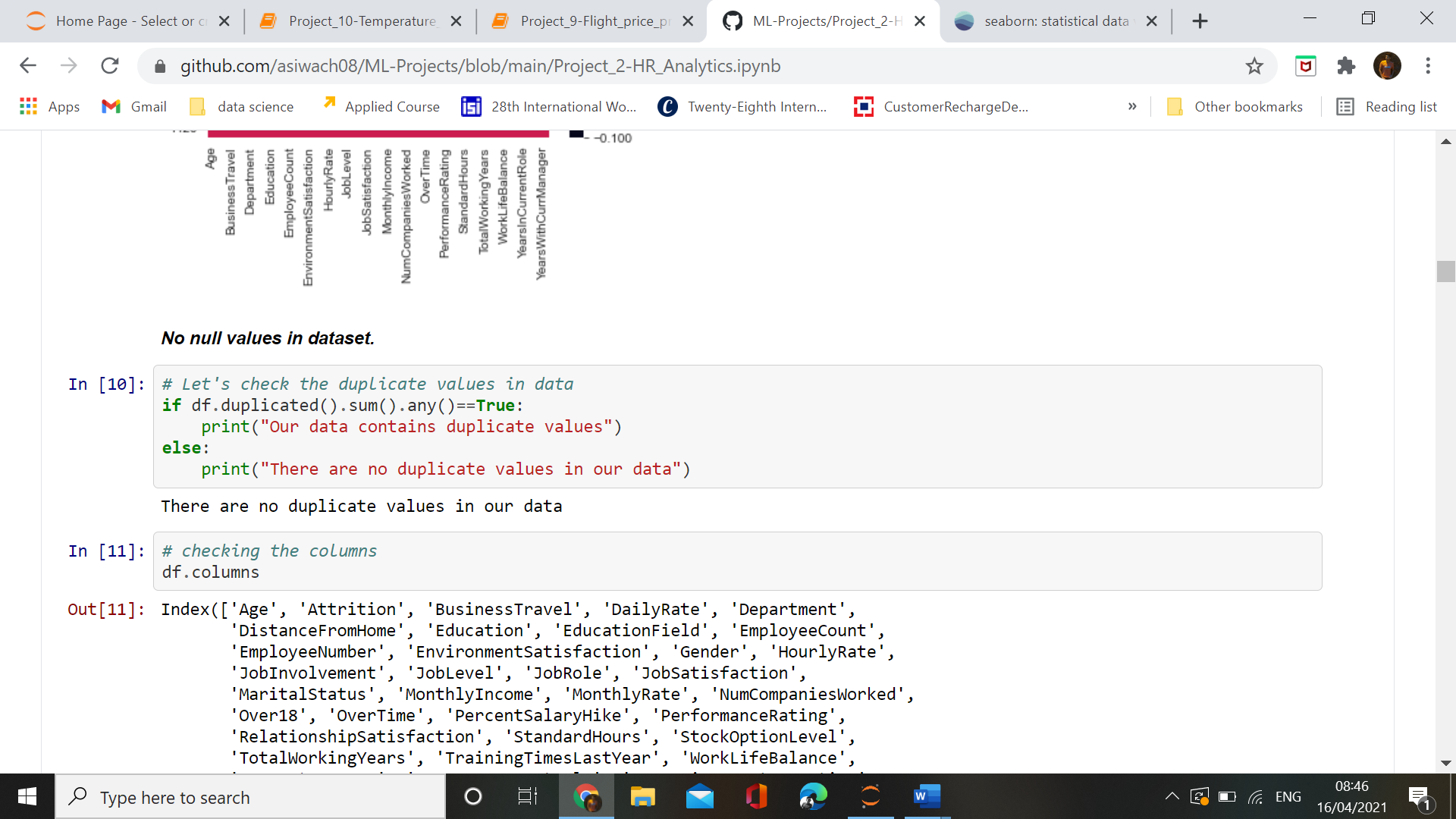


In our dataset ,we have 1470 rows and 35 columns.

**Checking null values and duplicates values**

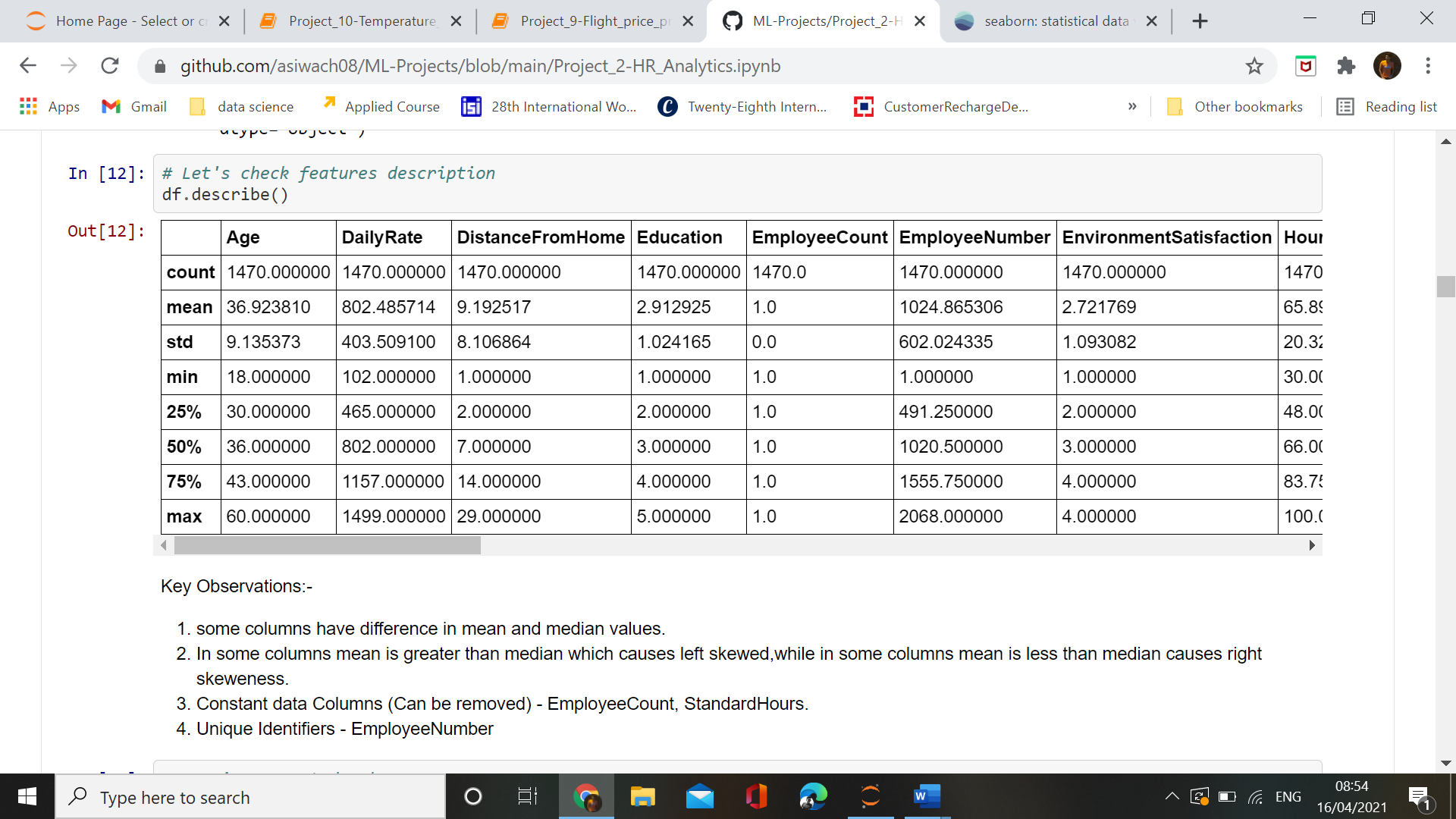
We need to check the presence of the missing values and need to replace them with mean, median and mode of data accordingly. Sometimes we have null values in the form of 0 , so we need to convert them to NaN and then replace them accordingly. The missing values can be removed also but it should be less than 5 percent of the whole dataset.and visualize the missing values using Heatmap. The missing values are represented by the horizontal lines. This map provides an informative way of visualizing the missing values located in each column, and to see whether there is any correlation between missing values of different columns.





There are no null values and no duplicates values in our dataset, it means our dataset need less treatment before sending model selection.

**Statistical summary**



Key Observations:-

1. some columns have difference in mean and median values.

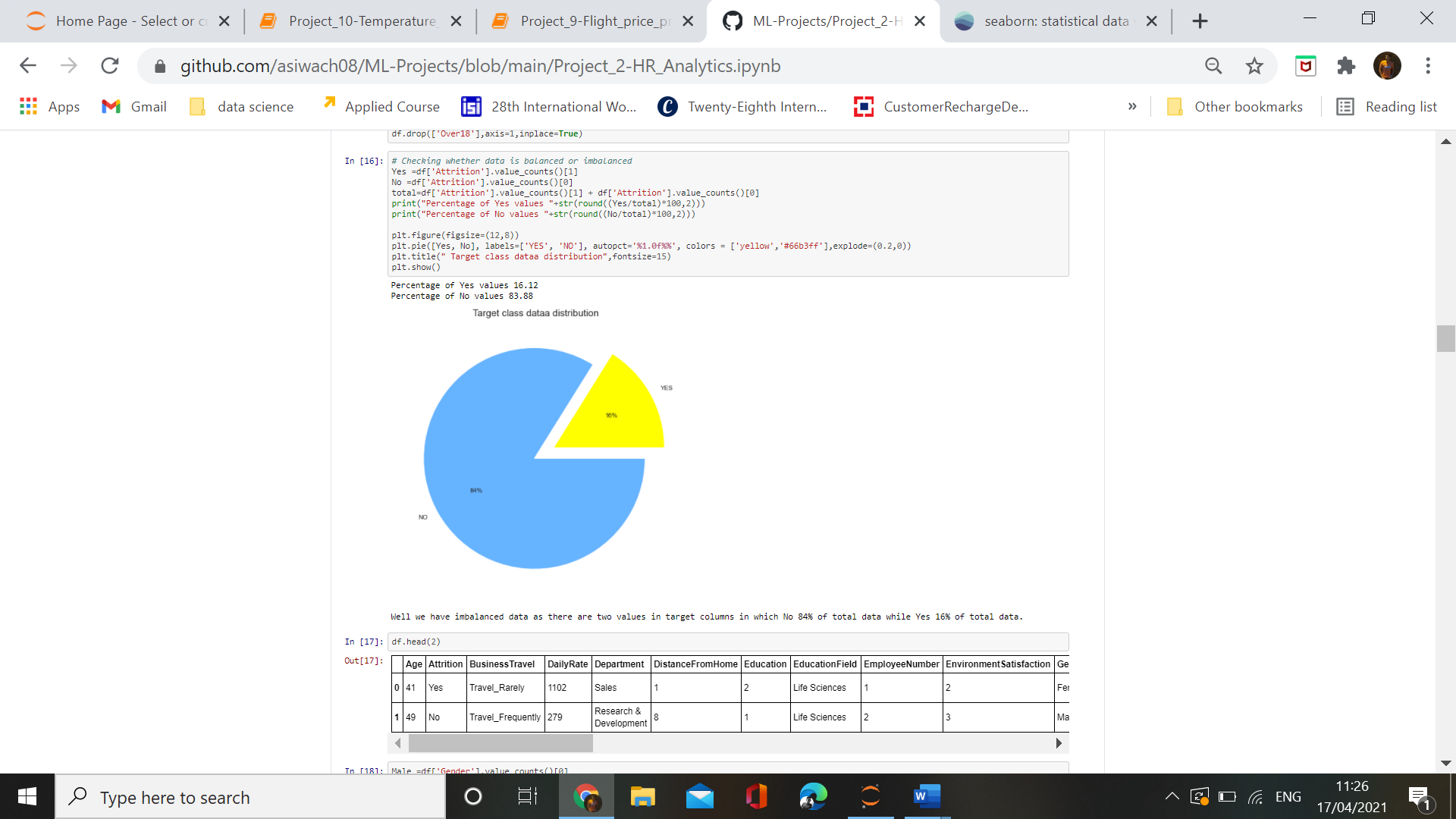
2. In some columns mean is greater than median which causes left skewed,while in some columns mean is less than median causes right skeweness.

3. Constant data Columns (Can be removed) - EmployeeCount, StandardHours.

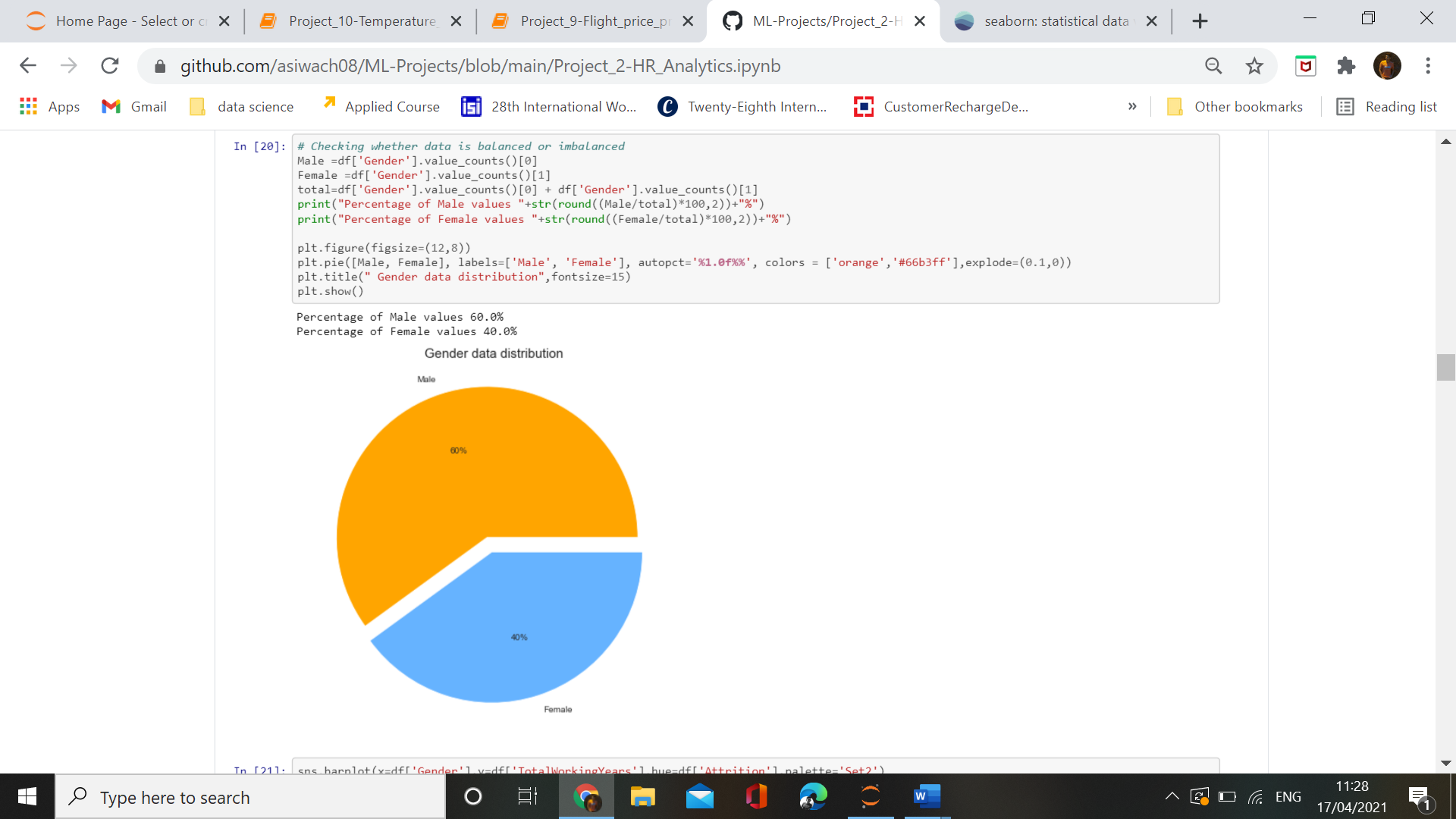
4. Unique Identifiers – EmployeeNumber

**3. Visualization**

Target data and gender data distribution

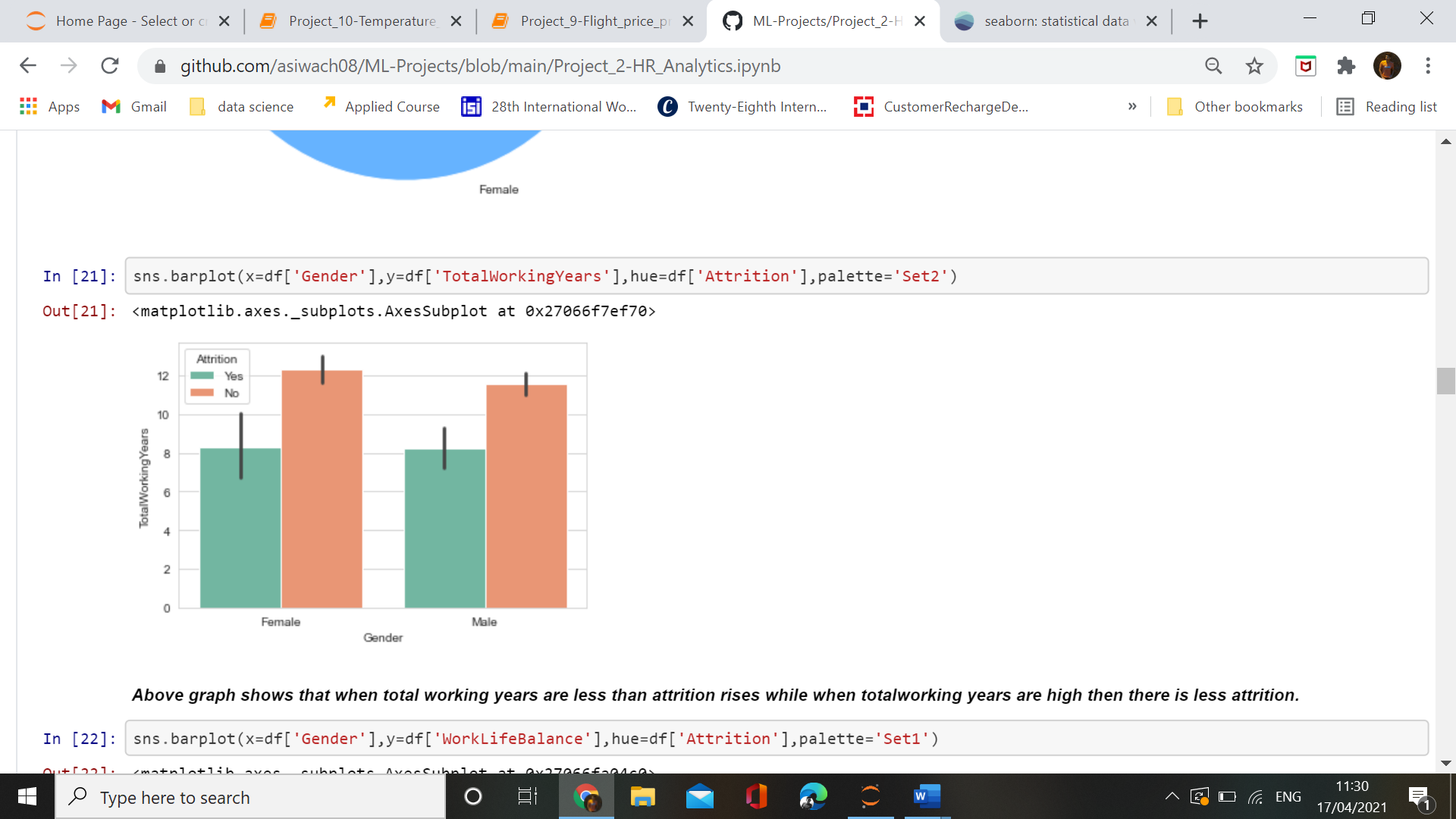


In HR Analytics dataset ,we are having an imbalanced target feature. The target feature of our dataset contains 2 values. 1) Yes, and 2) No. Percentage of yes values is only 16% while percentage of no values is 84%. This imbalance in target feature will not give us accurate results for prediction.



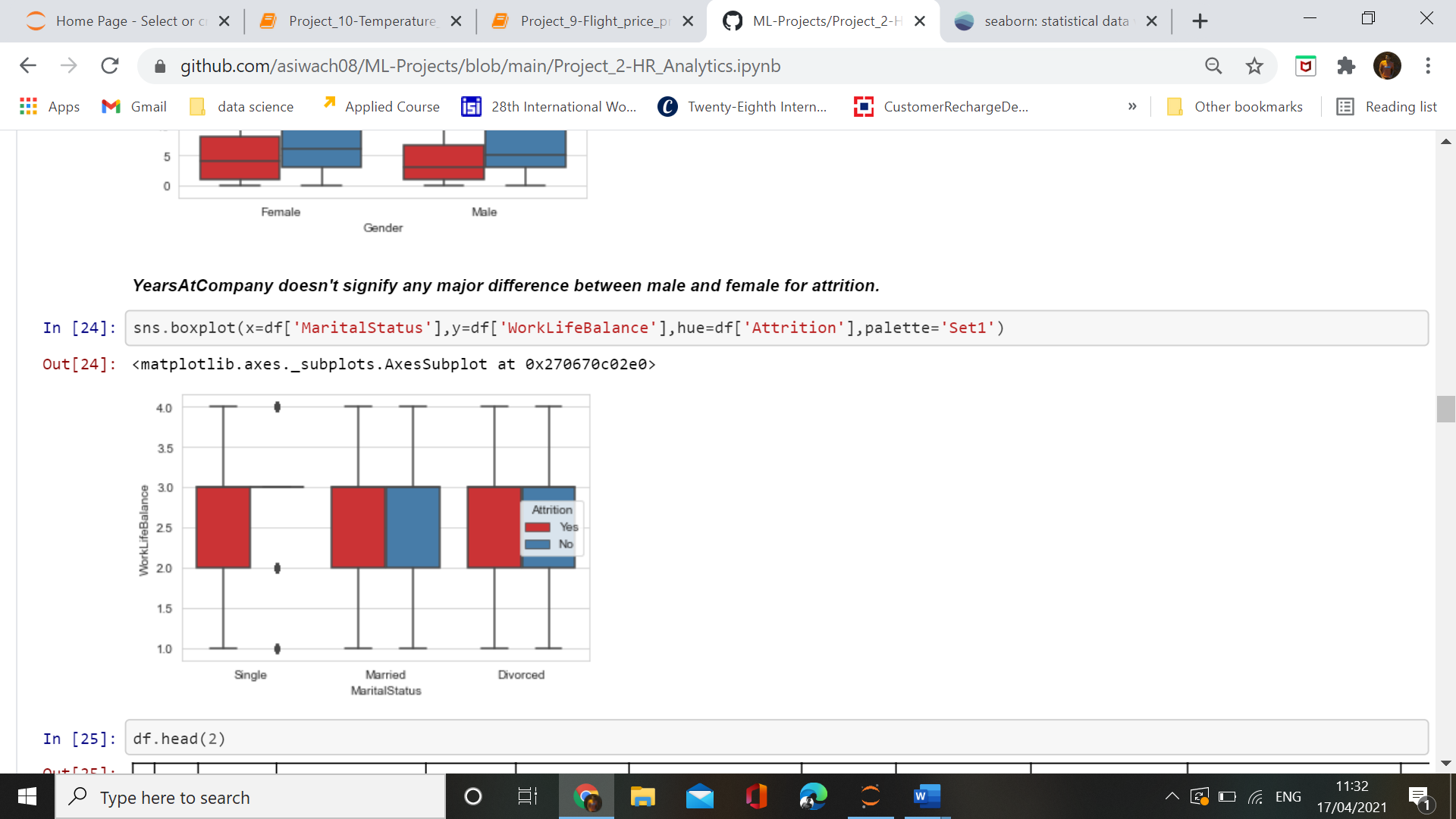
Gender feature of dataset contains percentage of Males -60% while percentage of Females – 40% of total data .So our dataset contains almost balanced data based on the gender.

**Barplot of gender with respect to attrition**



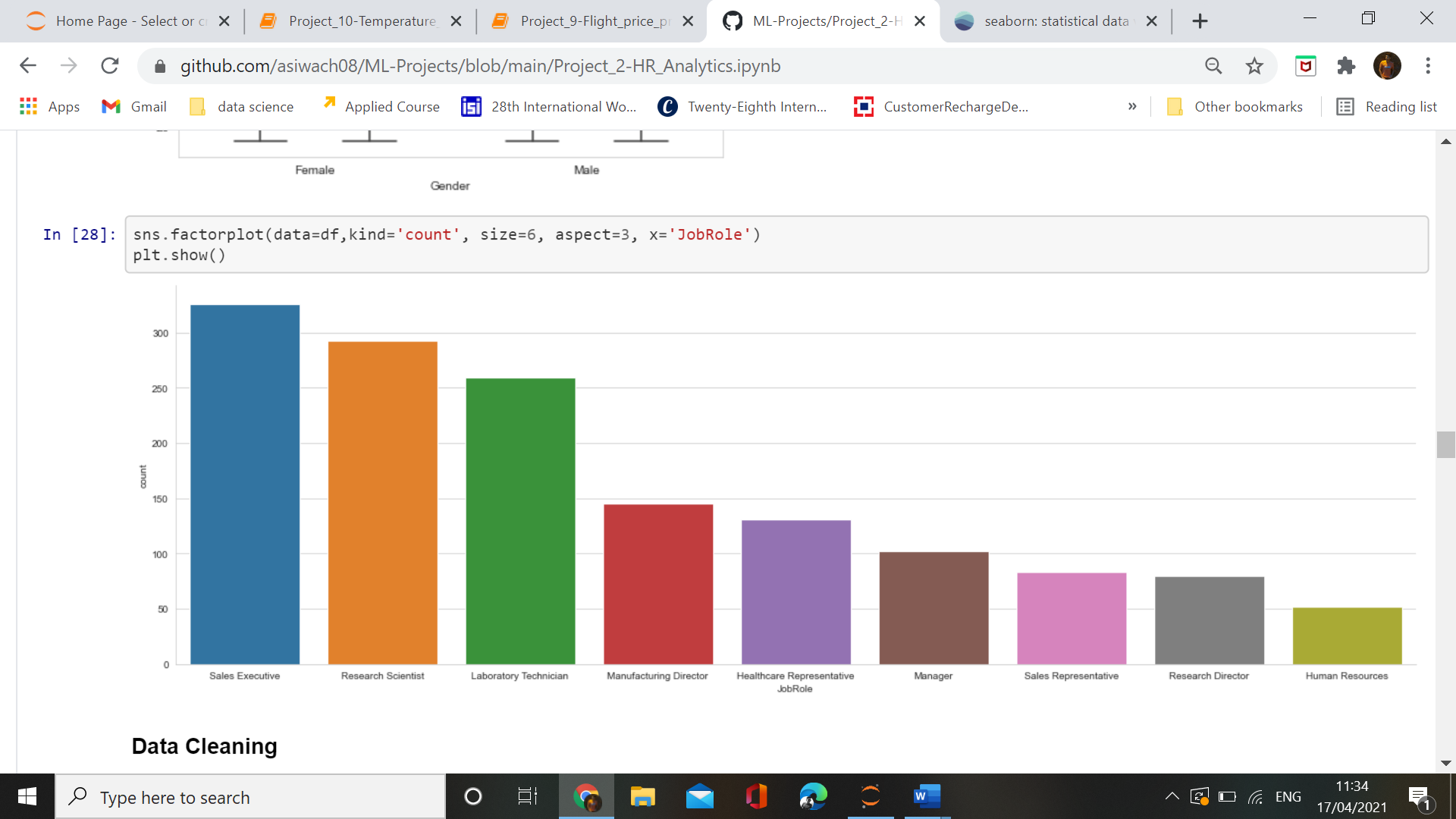
Barplot of gender with totalWorkingYears shows that there is no relation of attrition in gender based on the totalWorkingYears in a company.

**Boxplot of marital status**



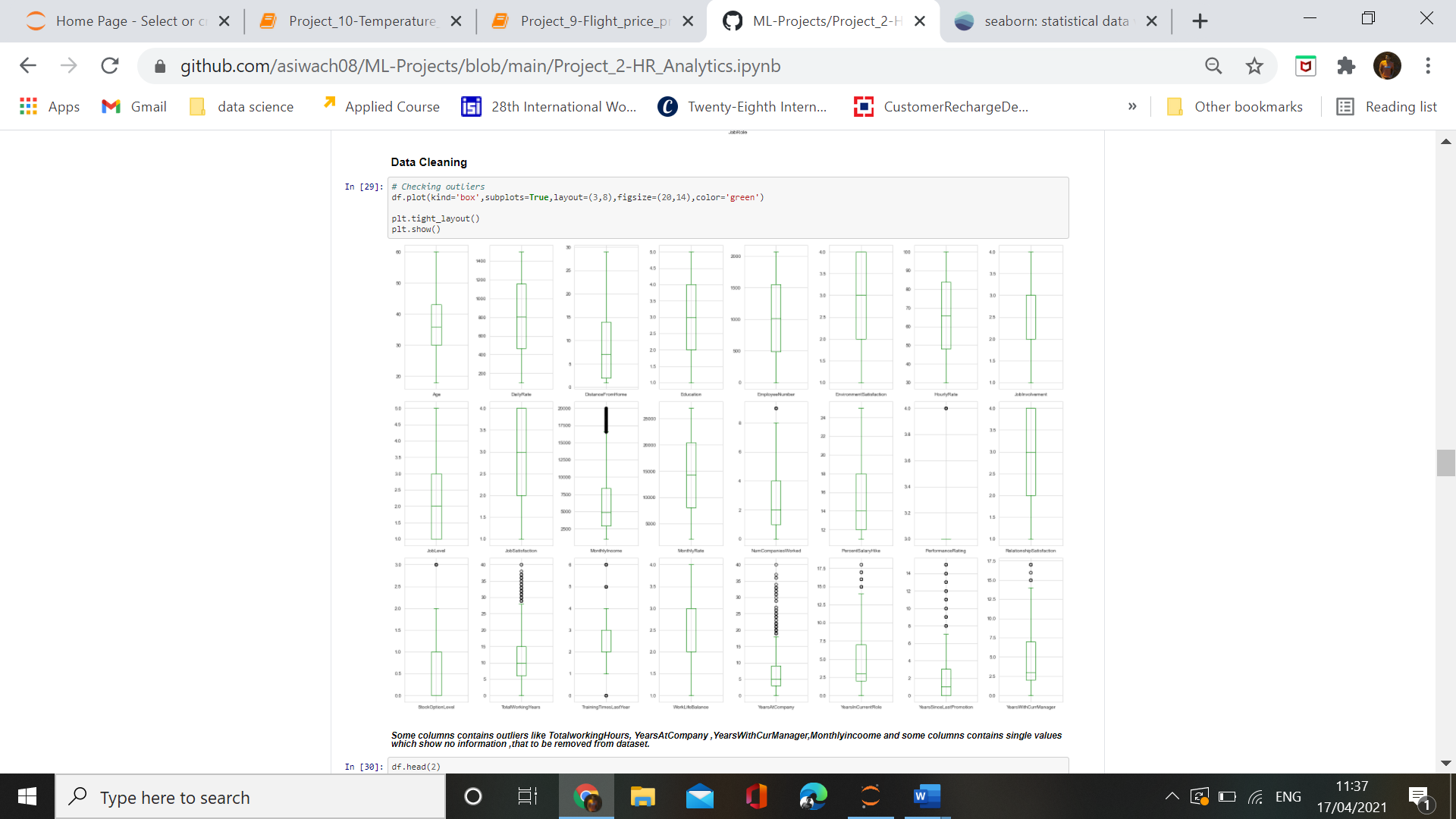
Boxplot on maritalStatus and WorkLifeBAlance suggest that based on the marital status and worklife there is no relation for the aattrition of employees.

**Factorplot on Jobrole**



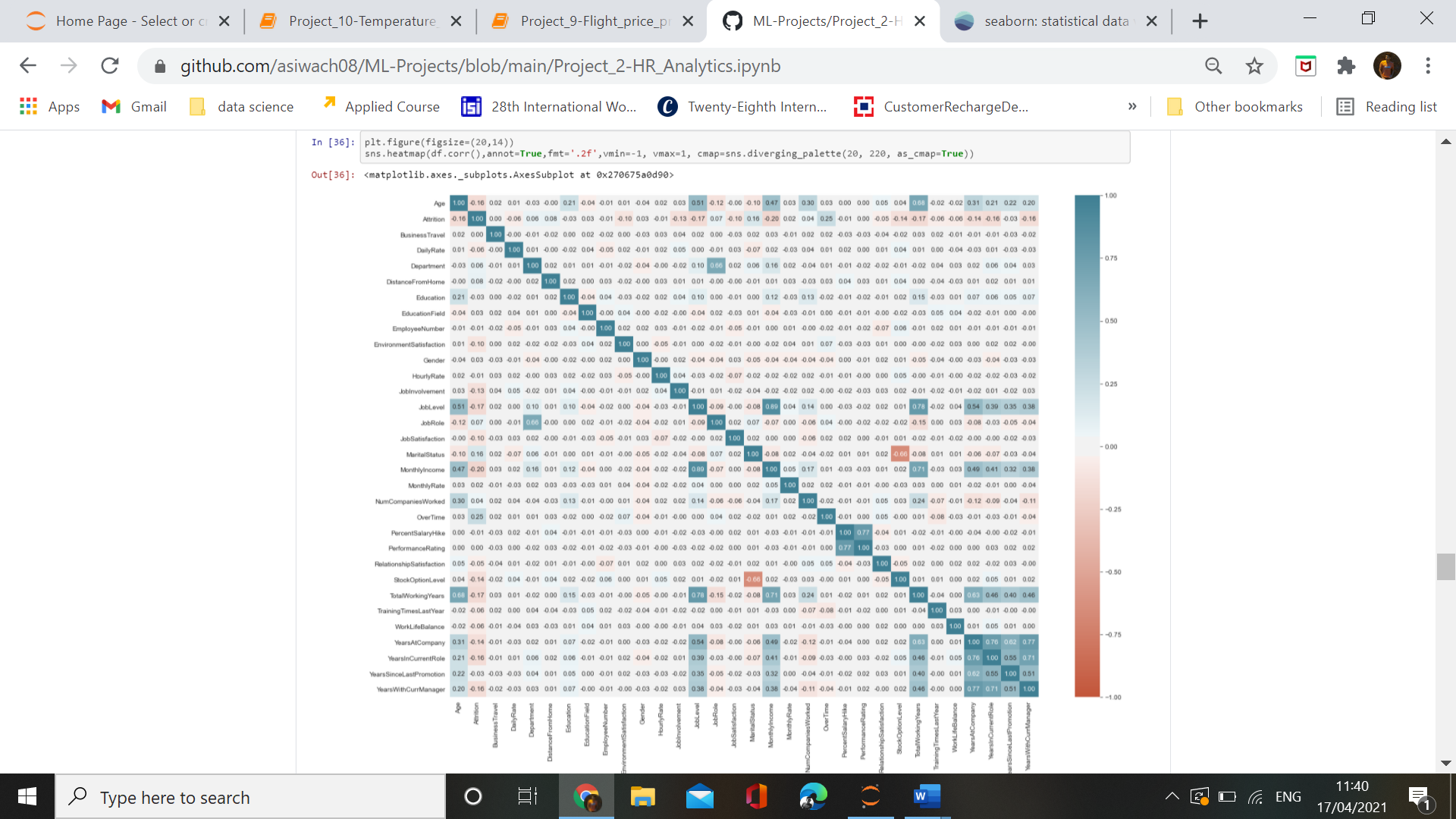
Factorplot on Jobrole helps in understanding the number of employees in different departments. We can see that Sales Executives are more in number while Human Resources are least in number.

Outliers



Using the boxplot ,we find out that outliers are present in many columns like TotalWorkingHours, YearsAtCompany,YearsWithCurManager ,MonthlyIncome and some other columns . These outliers will create variance in our model accuracy .We will deal with outliers if necessary in further steps ,but due to removal of oultliers if data loss is more than 5% we will not remove outliers.

Heatmap



Key Observations :-

From the correlation graph we can see that in some of the independent columns ,`multicolinearity` is present.

1. Like` MonthlyIncome` and `Joblevel` share a strong correlation of 89% dependence between them.

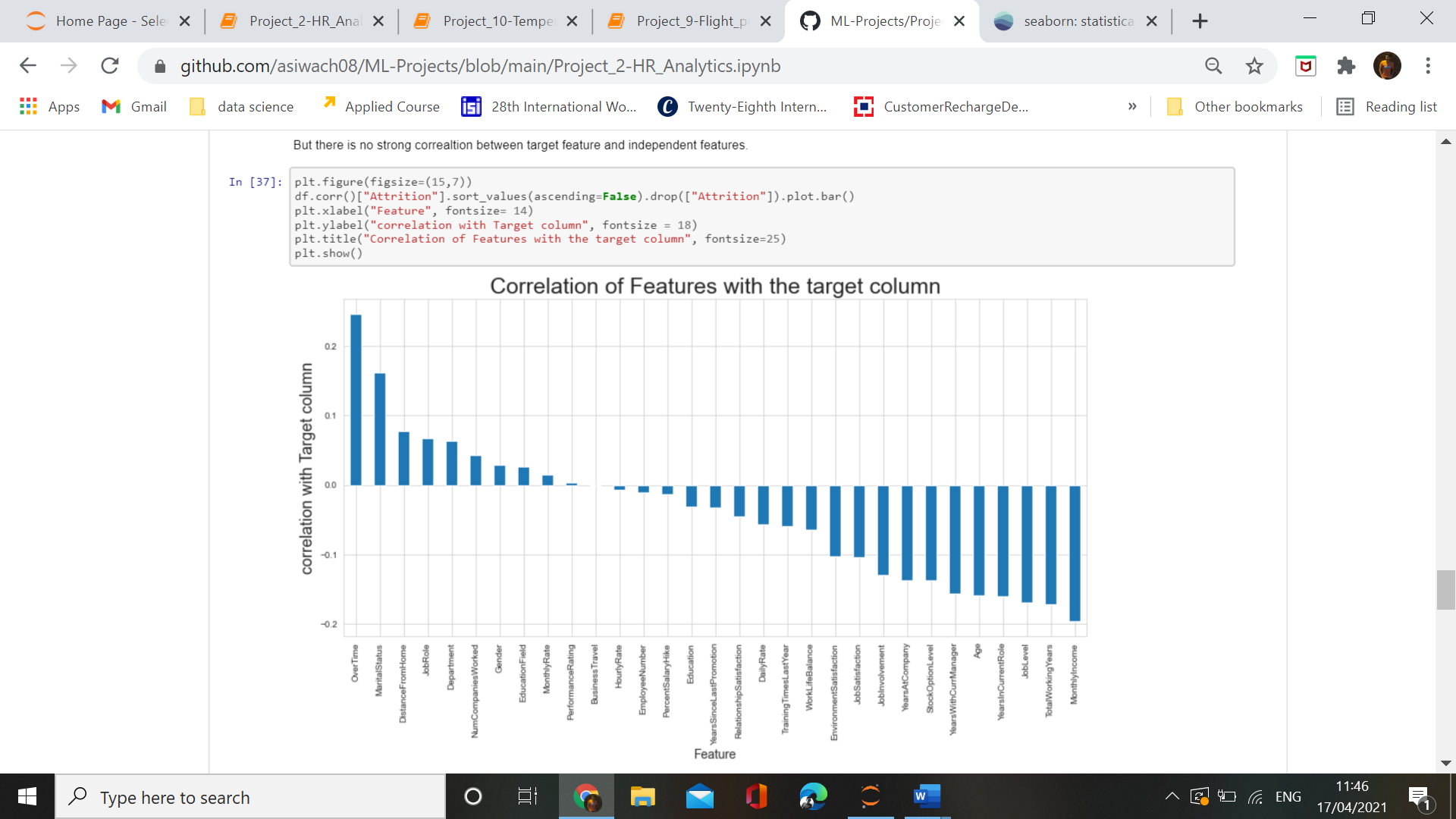
2. `TotalWorkingYear` and `JobLevel` also shares a strong correlation of more than 75% dependence on each other.

3. `TotalWorkingYear` also have correaltion with `Age` which shares a correaltion of 68% dependence on each other.

4. `Marital\_Status` are negatively related to `stockoptionlevel` with a dependence of 66% .

But there is no strong correaltion between target feature and independent features.

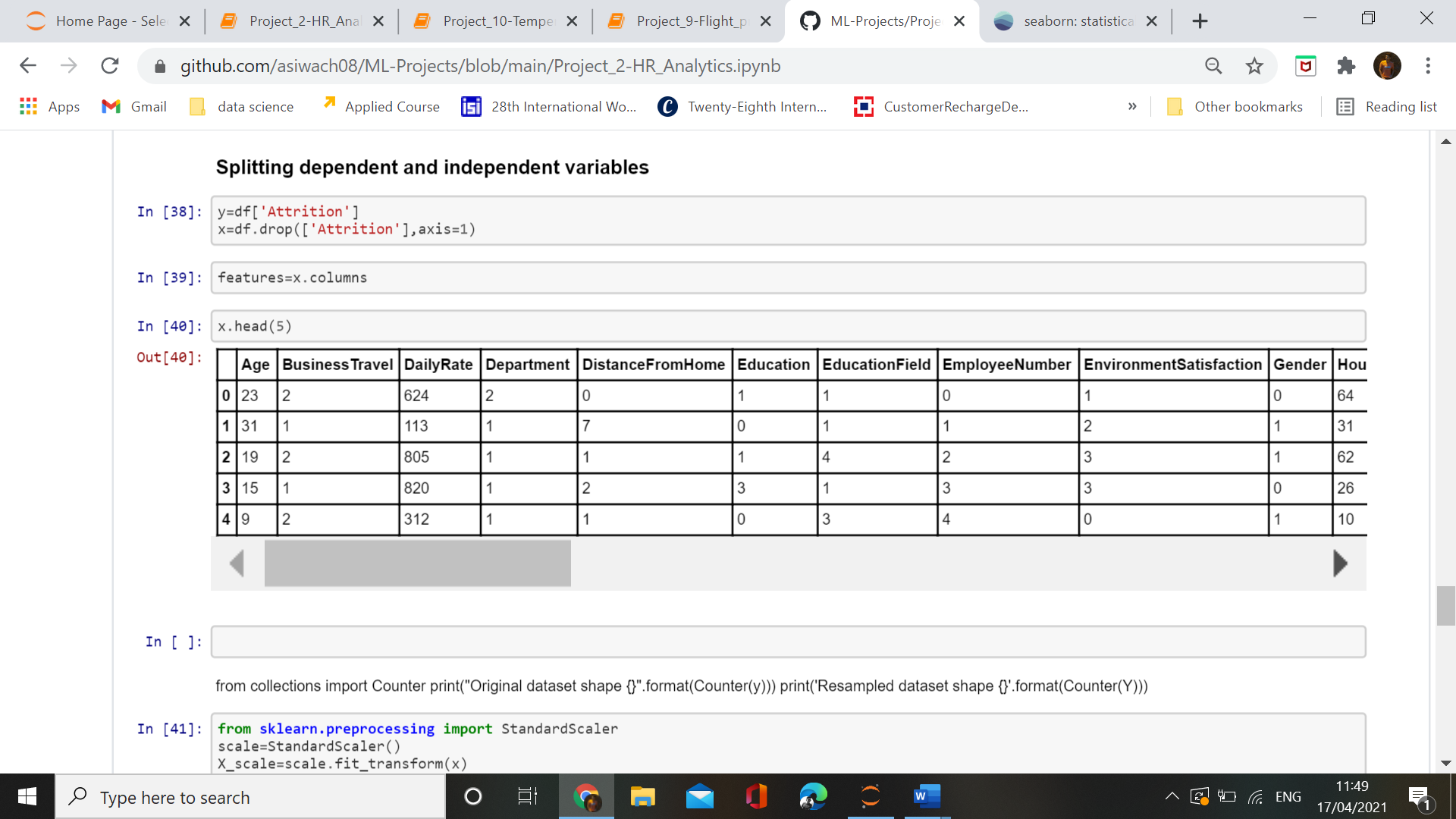
Correlation Graph



The correlation graph shows that most of the features are negatively correalted with the target variable and those who are correlated with the target variable are not so strongly correlated. This shows that there is no strong effect of any independent feature on the target feature. While features like MonthlyIncome is having negative correlation with the attrition of employee which means that as the monthlyIncome of employee increases ,changes of their attrition will decrease. Overtime feature is possessing positive correlation degree high than others with target feature which suggest that as the overtime period increases, employees seeks towards the attrition.

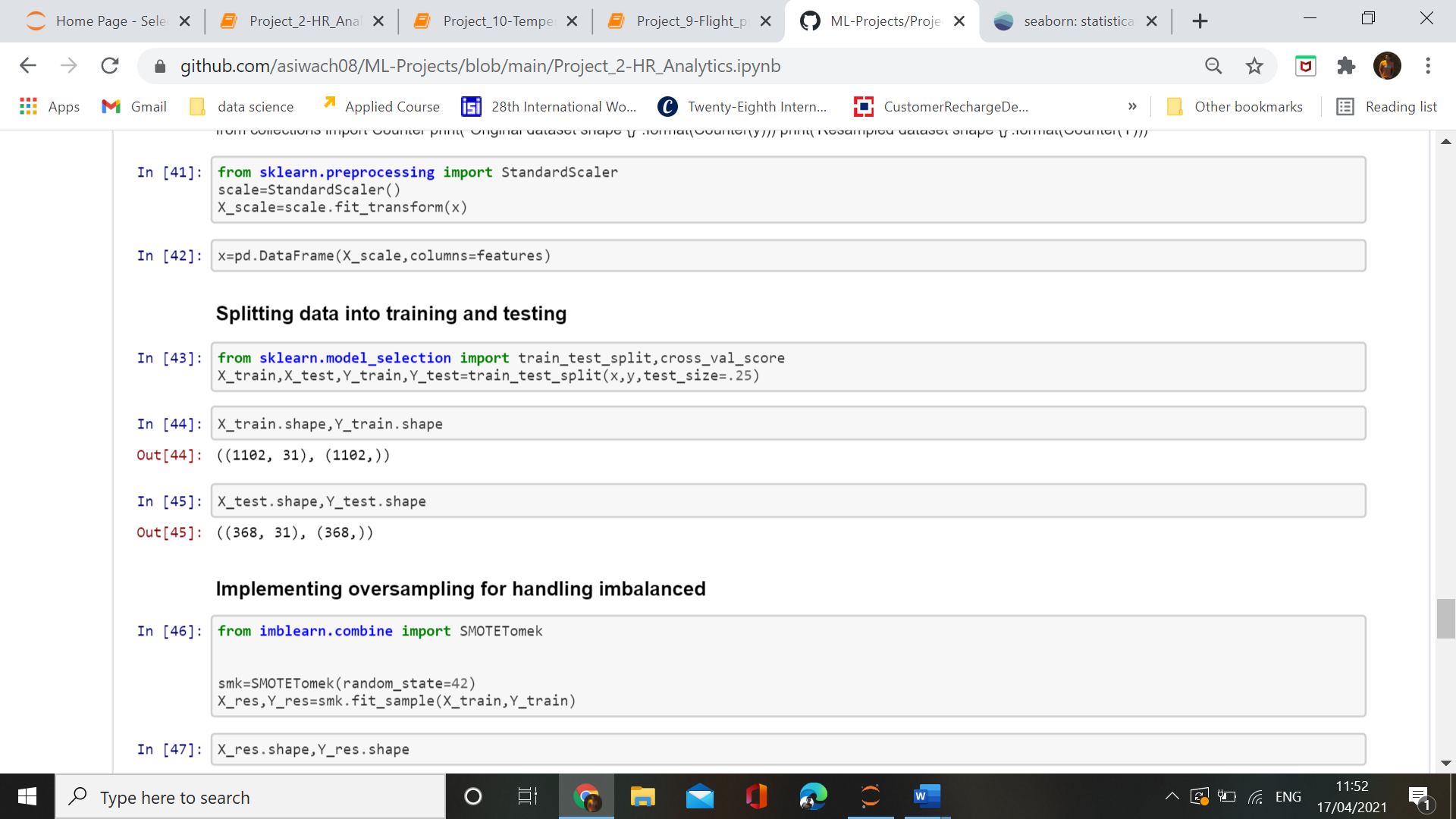
**4. Building Machine Learning Model**

Splitting data into dependent and independent features



We have splitted dataset ,and y is our dependent feature and x is our independent features

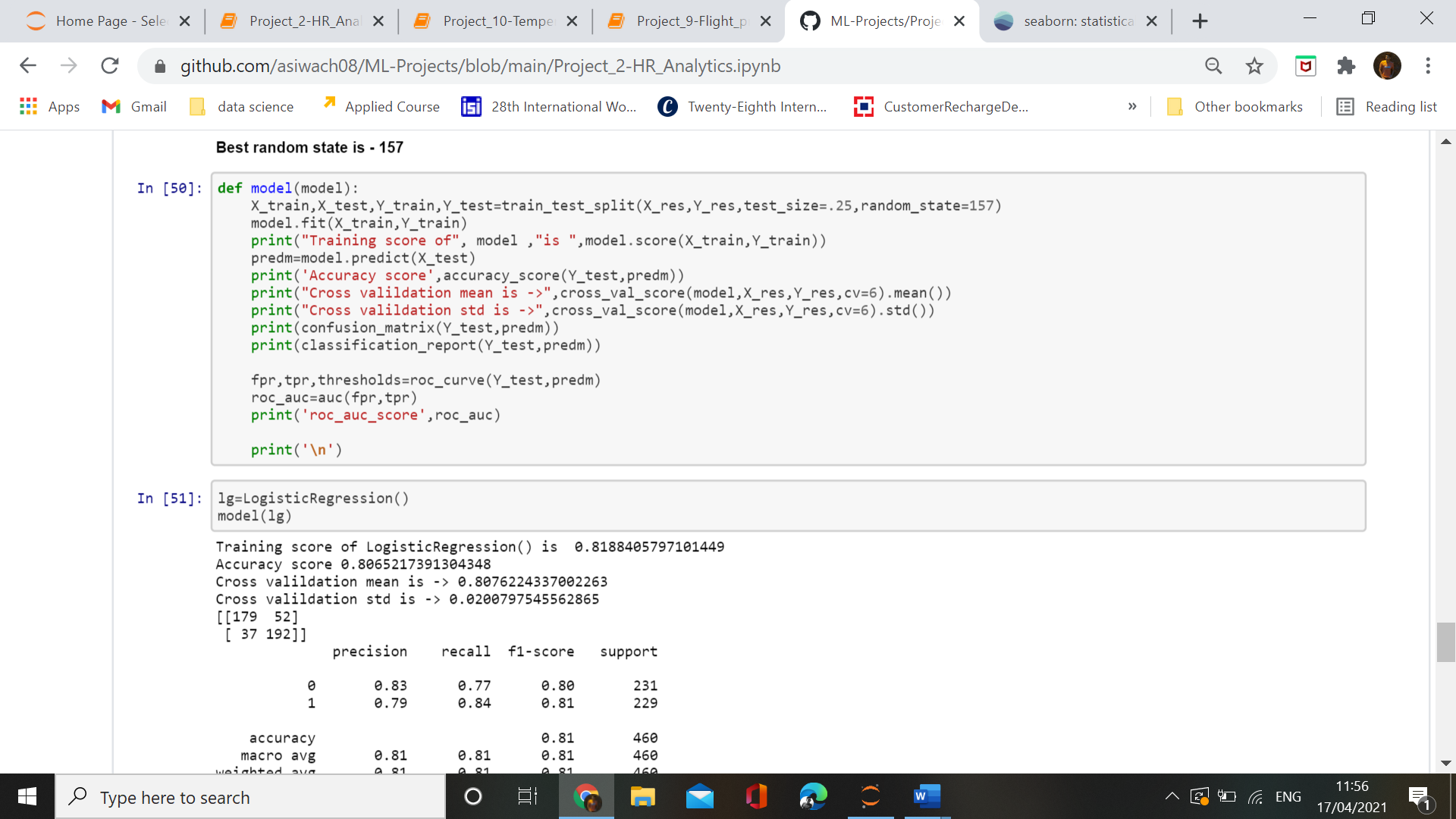
In our dataset we are using StandardScaler for scaling our dataset so that model will not be biased. As if some features contains higher values than other features ,in this condition our model will start giving more preference to those columns which have higher values i.e why we are using standardsclaer method for scaling our data.



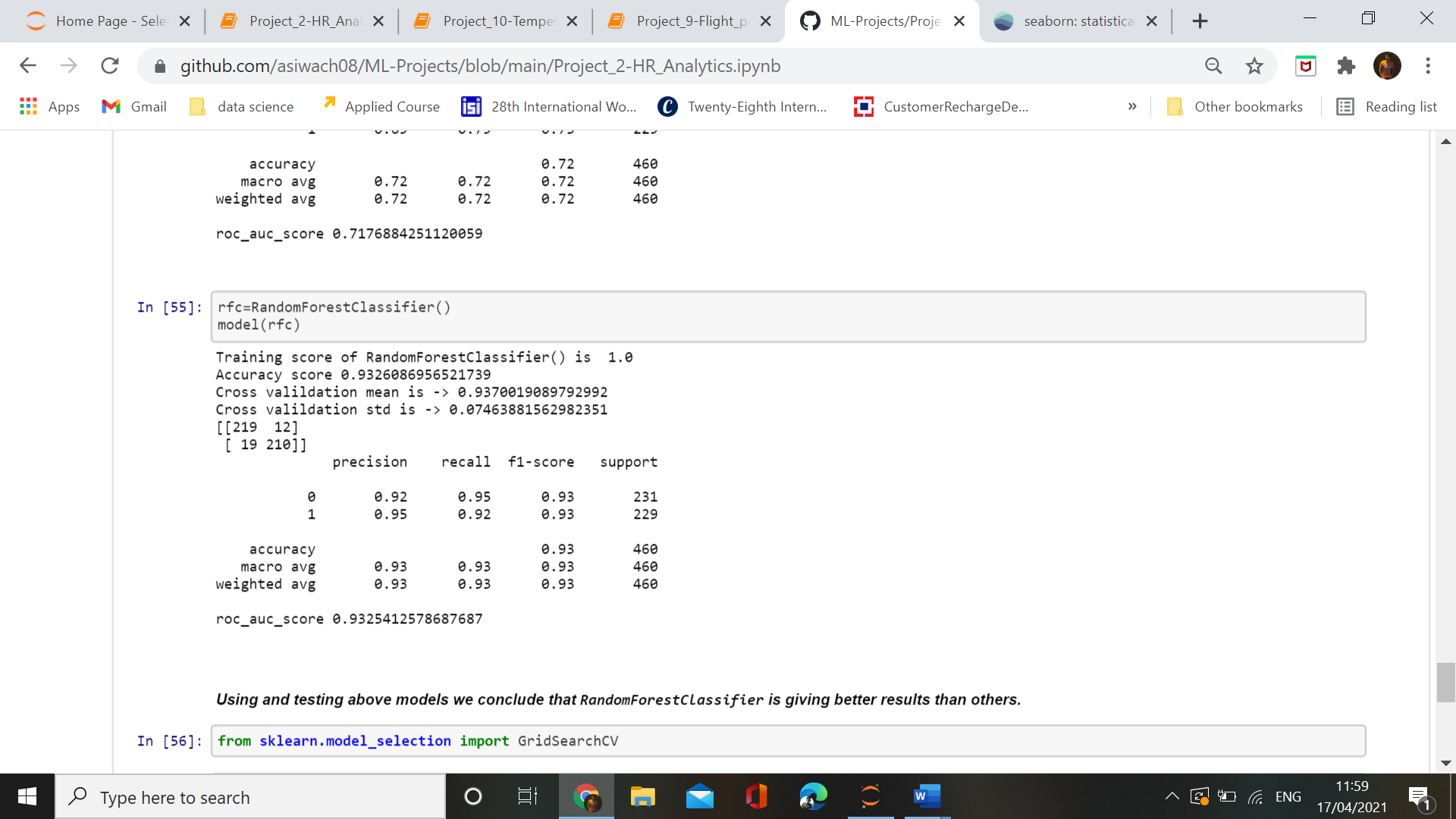
Using the train\_test\_split method we splitted the data into training purpose and testing purpose. Using training data we are going to train our model while using the test data we will test whether our mode is model is behaving ideal and giving good results or getting biased.

As our dataset was imbalanced ,so we are using oversampling technique for balancing of target feature.Oversampling method will automatically increase the number of samples which are less in number as compared to other class. In our dataset ,we have NO percentage very less than Yes percentage, so oversampling method will increase the percentage of No samples in our dataset by simply increasing the present values.

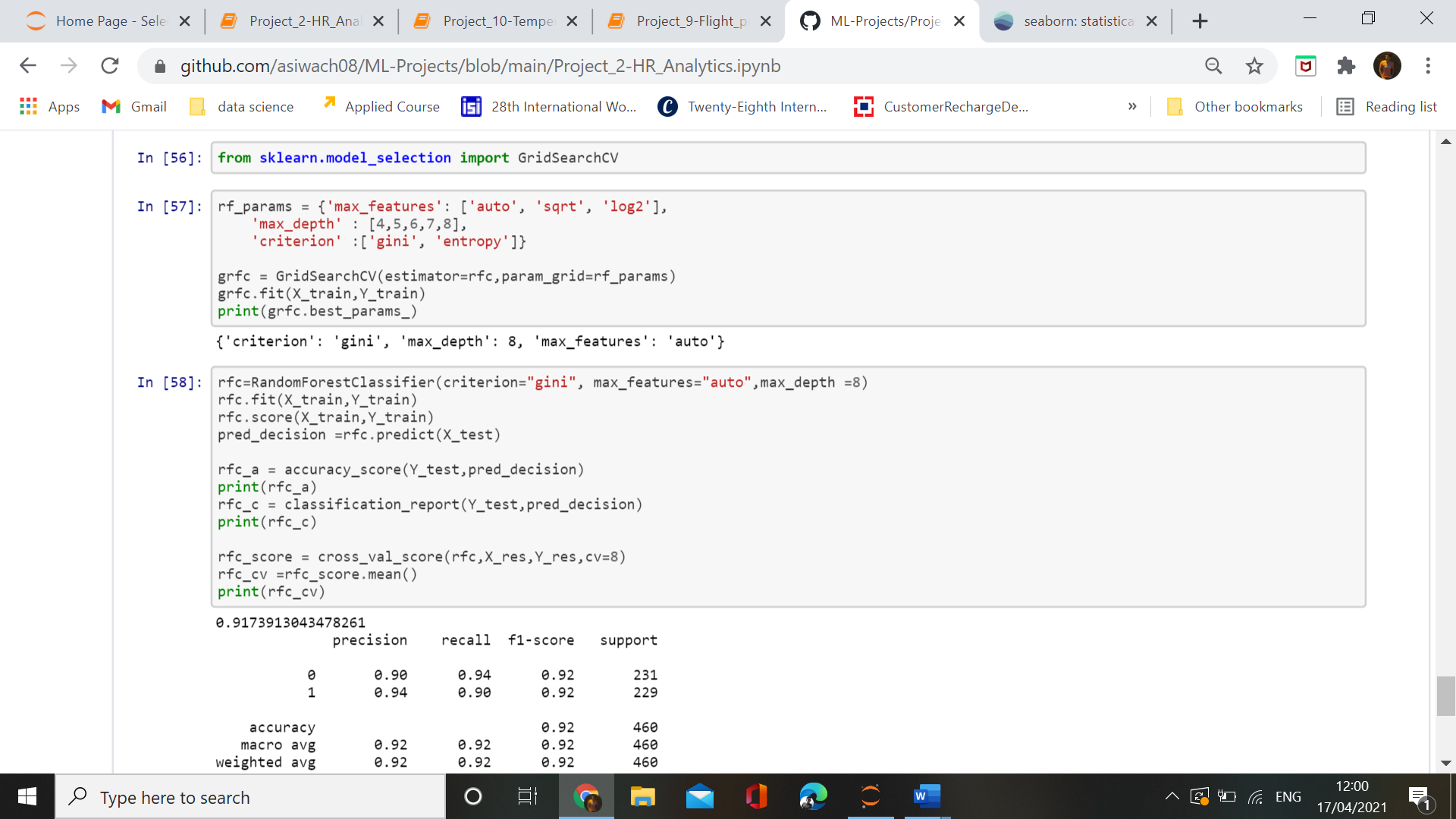
Applying method for finding the best accuracy of models.



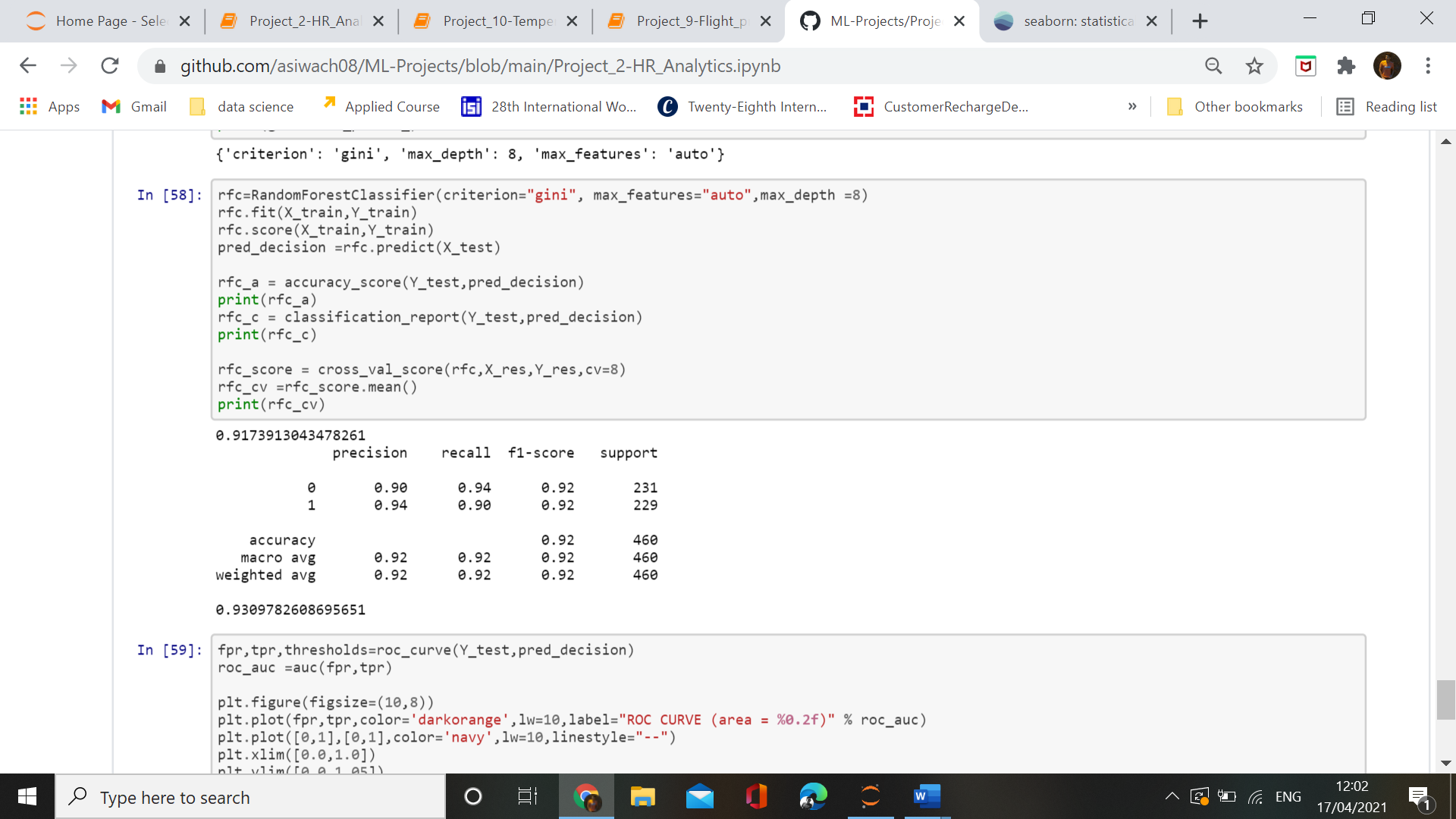
By testing and training different models ,we find out that RandomForestClassifier is giving better results than others.

  
It is giving us accuracy of 93% with f1-score 93%. We need to check f1\_score when the dataset is imbalanced , but in our case we used oversampling for balancing the dataset that why f1\_score is same for both the class of target feature.

Importing a model selection method called GridSearchCV for enhancing model accuracy by finding best parameters for the best algorithm.



Now testing the best parameters on best algorithm and finding the best accuracy and all other metrics result.



By using the best parameter find out by GridSearchCV we are getting accuracy\_score and cross\_validation\_score with minimum difference which means that there is no underfitting or overfitting in our model.



**Conclusion:-**

 This step-by-step blog article is just an example of what advanced analytics can do for your business, and of how easy is to do it with the proper tool using EDA and machine learing algorithms to predict the future cases of attrition of employees in a company.