HR Analytics Project



1. Problem Definition.

2. Data Analysis.

3. EDA Concluding Remark.

4. Pre-Processing Pipeline.

5. Building Machine Learning Models.

6. Concluding Remarks.

**Presented by-**

**Shailja Tiwari**

**Data Trained**

**Batch-1840**

1. Problem Definition.

**Theoretical overview:**

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:**

* HR Analytics is also known as people 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 is the departure of employees from the organization for any reason (voluntary or involuntary), including resignation, termination, death or retirement.
* 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.
* How does Attrition affect companies? and how does HR Analytics help in analysing attrition? We will discuss the first question here and for the second question, we will write the code and try to understand the process step by step.

**Attrition affecting Companies:**

* A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork, and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time. This is especially concerning if your business is customer-facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if you constantly have new workers.

**Data Source:**

* <https://github.com/dsrscientist/IBM_HR_Attrition_Rate_Analytics>

**Problem Statement:**

We have to analyse the given data and apply all the domain technical knowledge in finding the factors that affect the employee attrition rate in the various organisation.

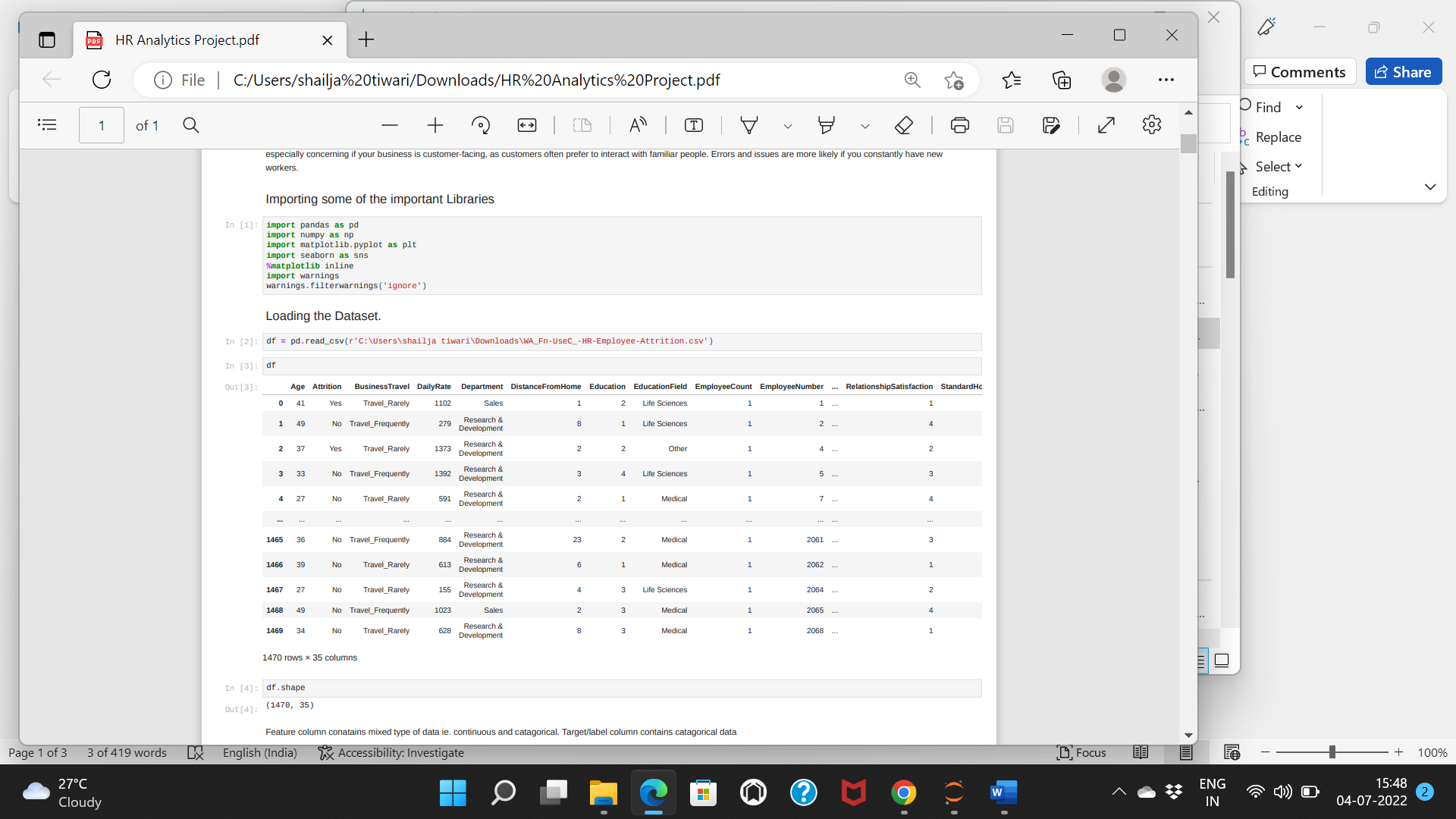
2. Data Analysis.

Data analysis refers to the process of inspecting, cleansing, transforming, and modelling data with the aim of discovering useful insights and information from the data which can be used for business developments and easy representation of the raw data to the non tech group and decision makers of any organization.

In data analytics and data science, there are four main types of analysis:

* Descriptive.
* Diagnostic.
* predictive
* prescriptive.

Various steps are involved in the data analysis process. These steps include importing all the important libraries, loading the dataset in the form of DataFrame by the help of Pandas library present in Python.



* So in our dataset we have total 1470 number of rows and 35 columns are present.

df.shape

(1470, 35)

**Dataframe Description:**

Dataset is divided in two parts input (Independent/Feature) variables and output(dependent/Label/Target) variables.

**Feature Columns:**

* Age
* BusinessTravel
* DailyRate
* Department
* DistanceFromHome
* Education
* EducationField
* EmployeeCount
* EmployeeNumber
* EnvironmentSatisfaction
* Gender
* HourlyRate
* JobInvolvement
* JobLevel
* JobRole
* JobSatisfaction
* MaritalStatus
* MonthlyIncome
* MonthlyRate
* NumCompaniesWorked
* Over18
* OverTime
* PercentSalaryHike
* PerformanceRating
* RelationshipSatisfaction
* StandardHours
* StockOptionLevel
* TotalWorkingYears
* TrainingTimesLastYear
* WorkLifeBalance
* YearsAtCompany
* YearsInCurrentRole
* YearsSinceLastPromotion
* YearsWithCurrManager

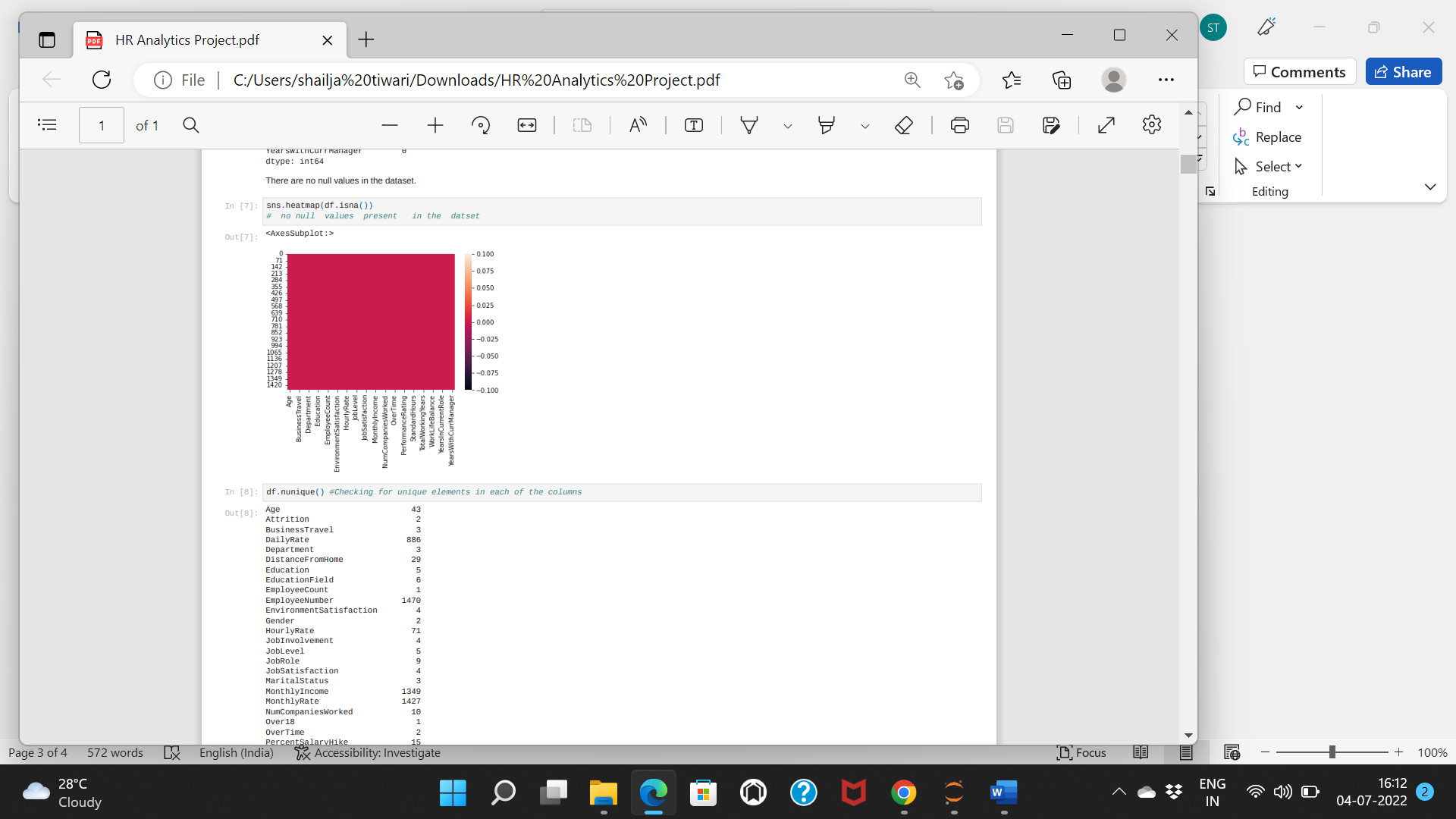
**Label column**:

* Attrition
* Feature columns also called as input or independent column has mixed type of data i.e continuous and categorical.
* Label/target column also called as output or dependent variable has numerical data.

Checking for the null or missing values in the dataset for proper model building and prediction the data set must contain non null values.

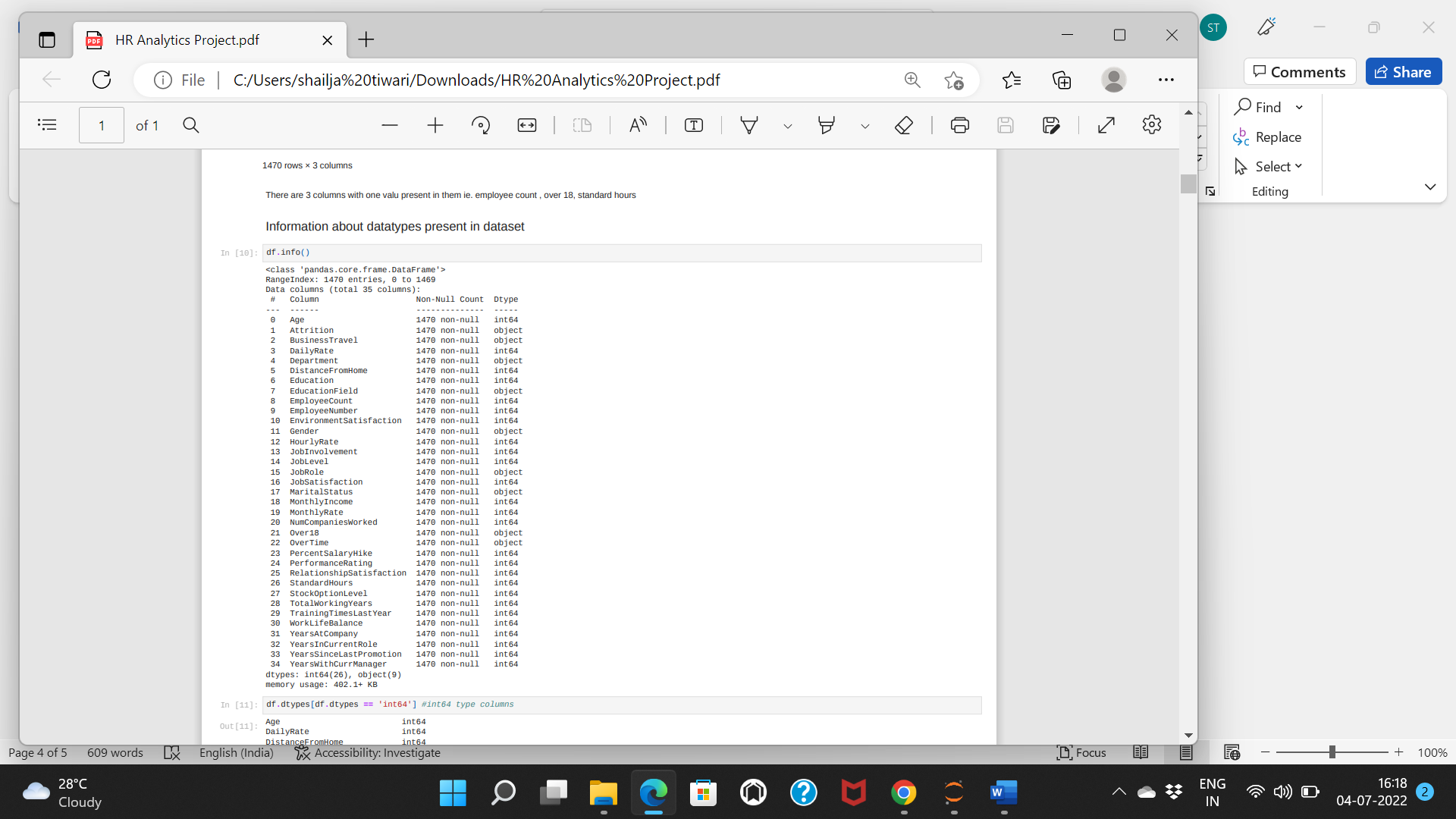
We will visualise the same with the help of heatmap and code.

df.isnull().sum()



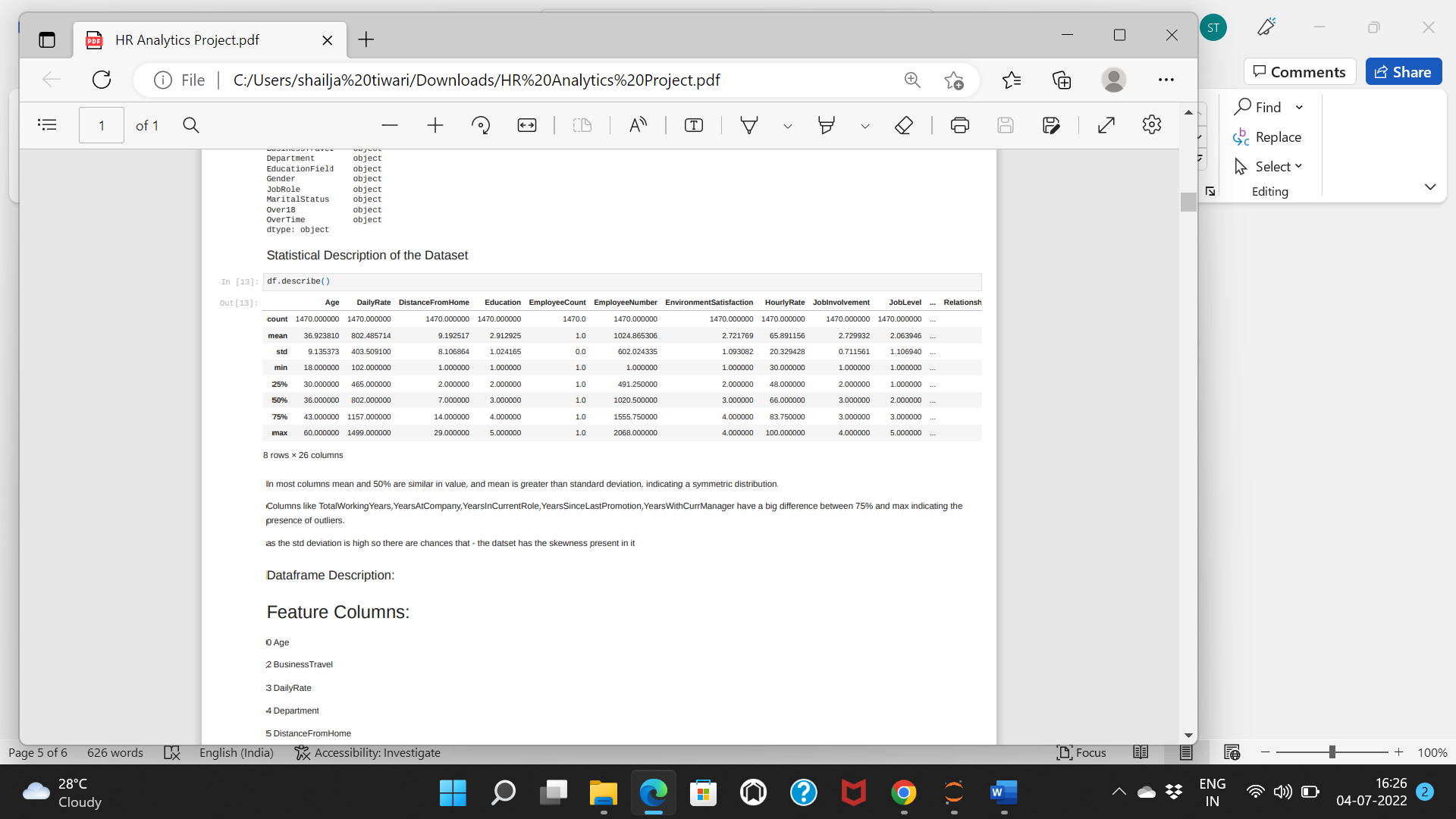
Getting the information about datatypes present in dataset.

df.info()



Establishing the Statistical description of the dataset.

df.describe()



* In most columns mean and 50% are similar in value, and mean is greater than standard deviation, indicating a symmetric distribution.
* TotalWorkingYears,YearsAtCompany,YearsInCurrentRole,YearsSinceLastPromotion,YearsWithCurrManager have a big difference between 75% and max indicating the presence of outliers.
* Due to higher standard deviation skewness may be present in the dataset.

# 3. Exploratory Data Analysis (EDA).

EDA is the method of analysing data using visual methods.  It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

**TYPES OF EXPLORATORY DATA ANALYSIS:**

* Univariate Analysis
* Bivariate Analysis
* Multivariate Analysis

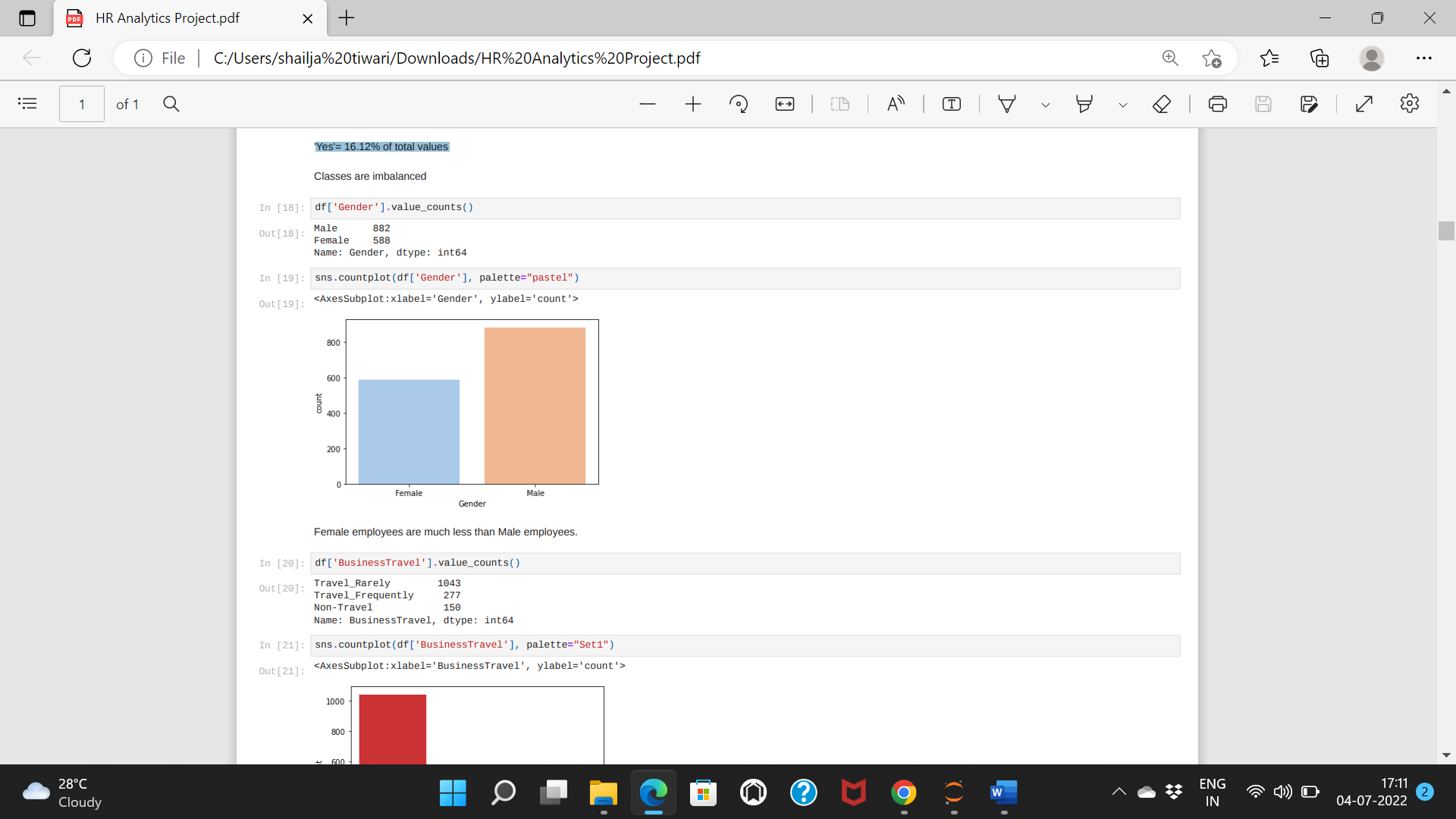
# 

# 

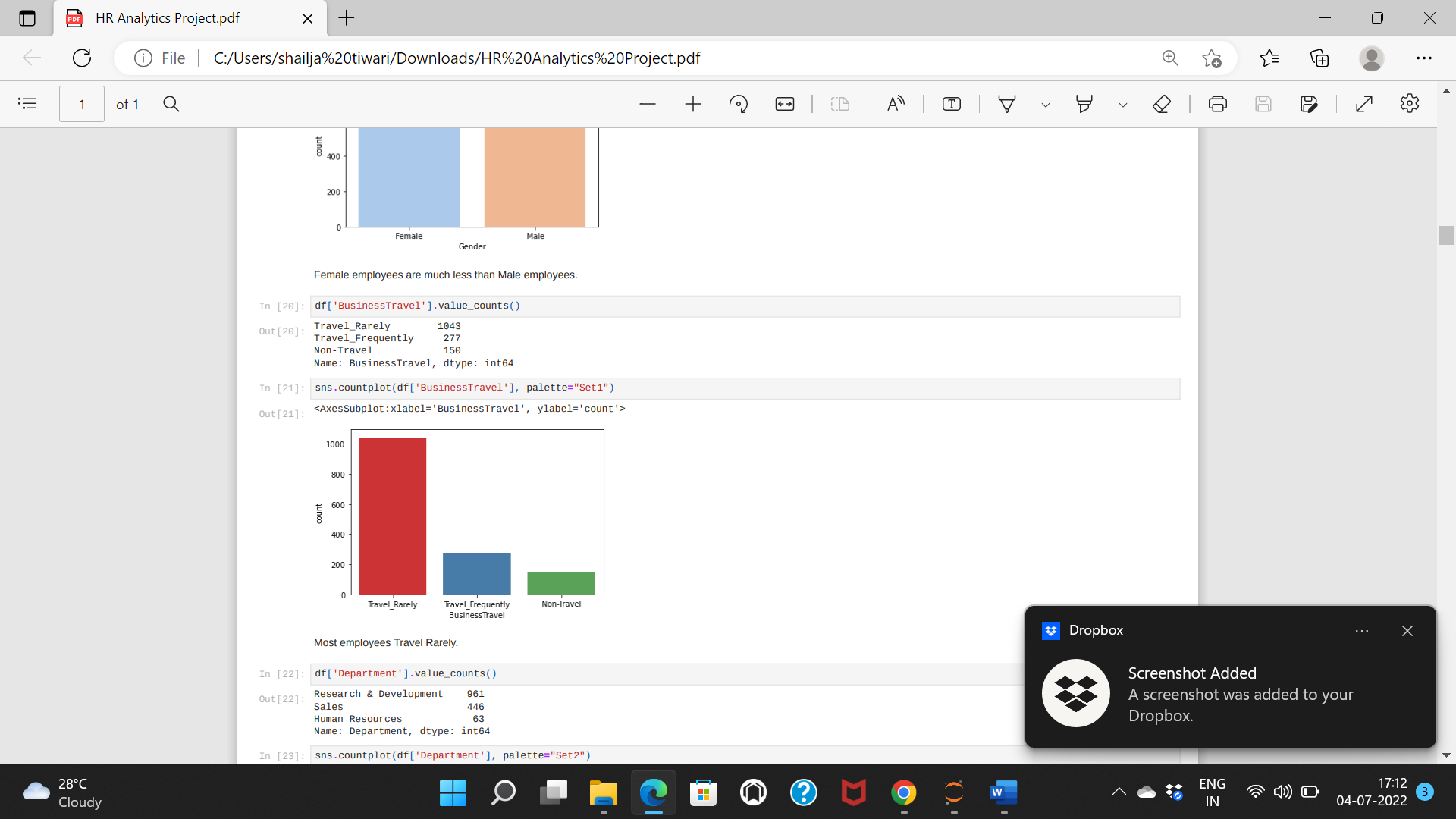
Class

'No' = 83.88% of total values

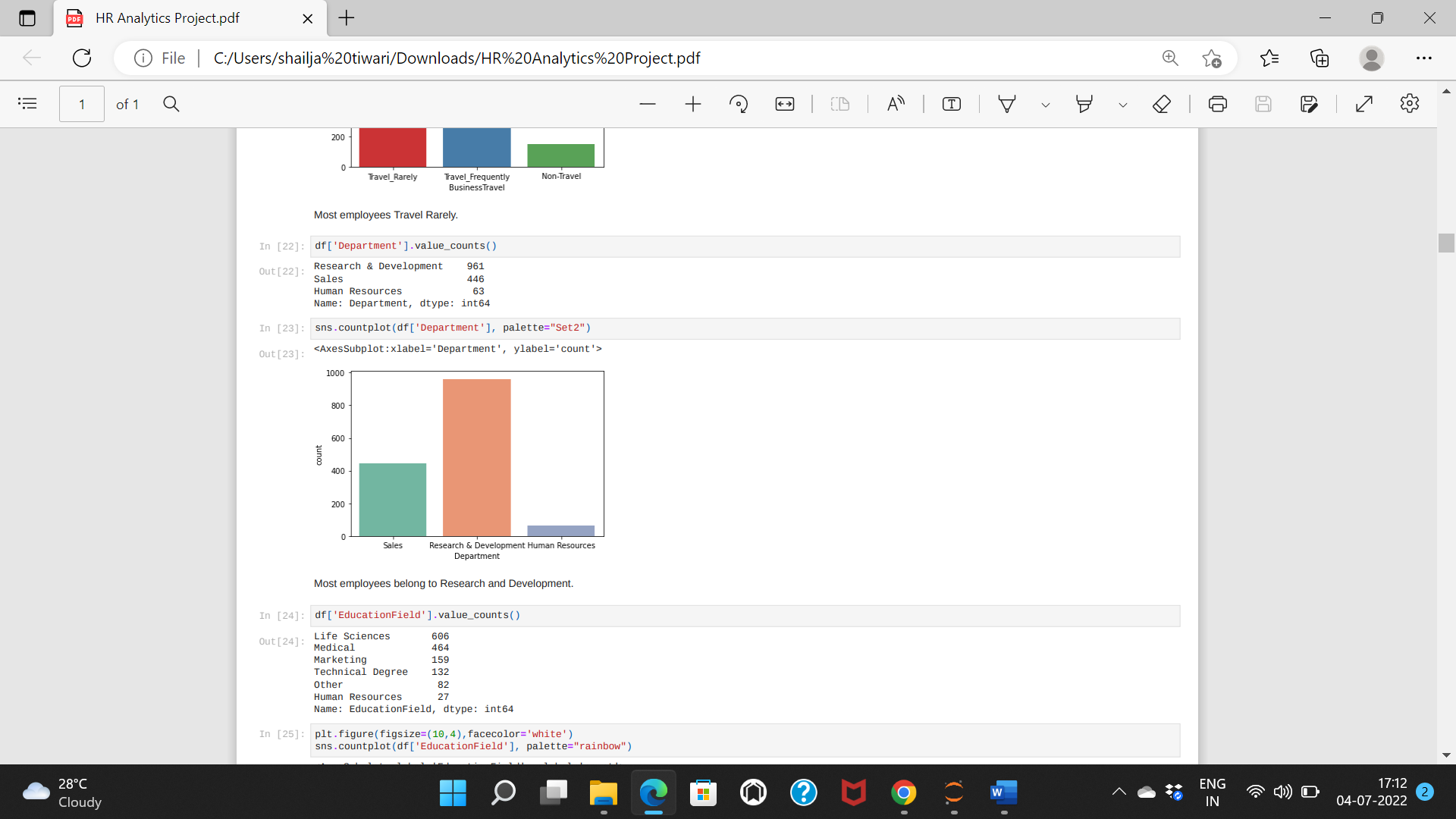
'Yes'= 16.12% of total values

****

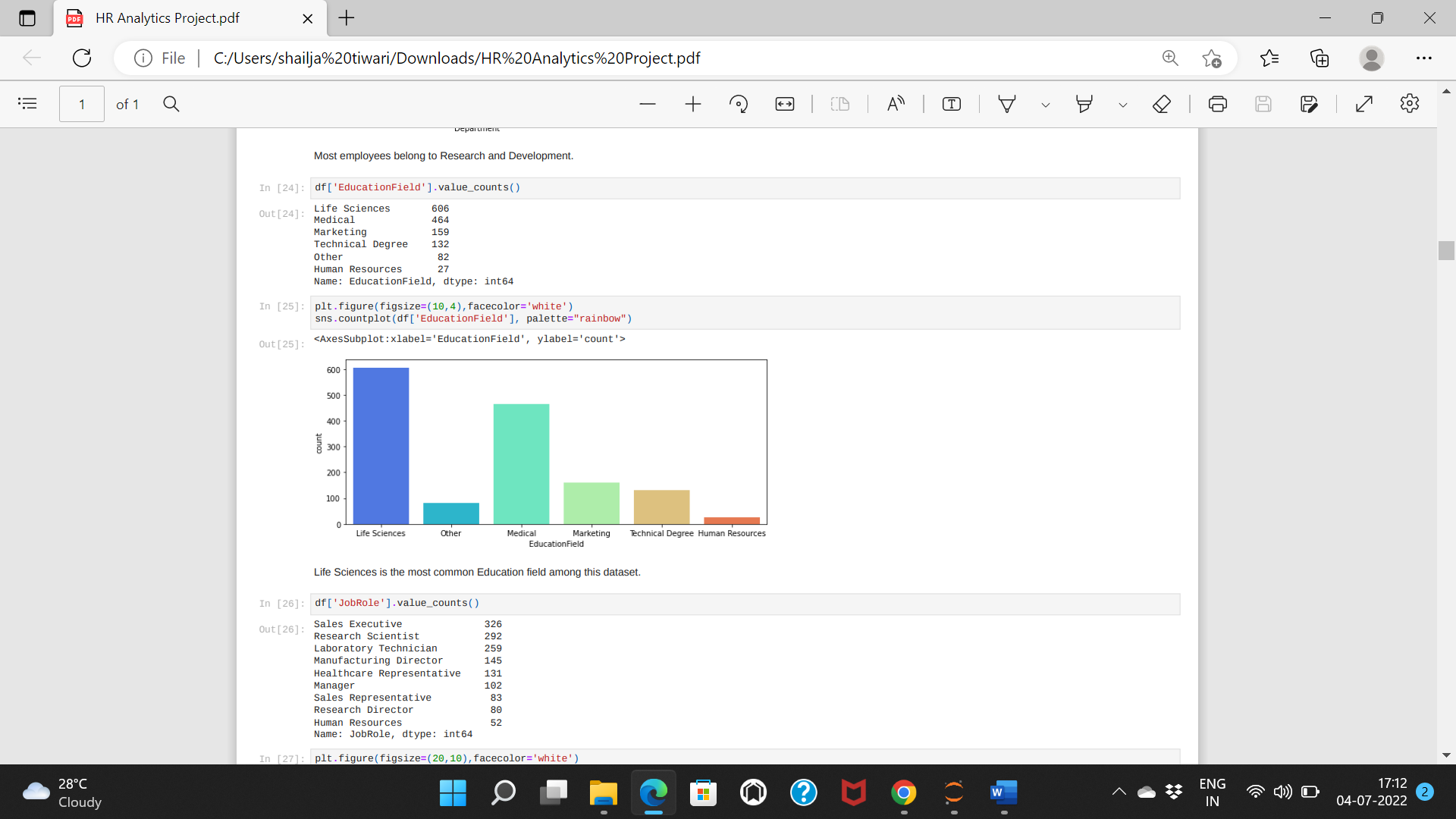
Female employees are less than Male employees



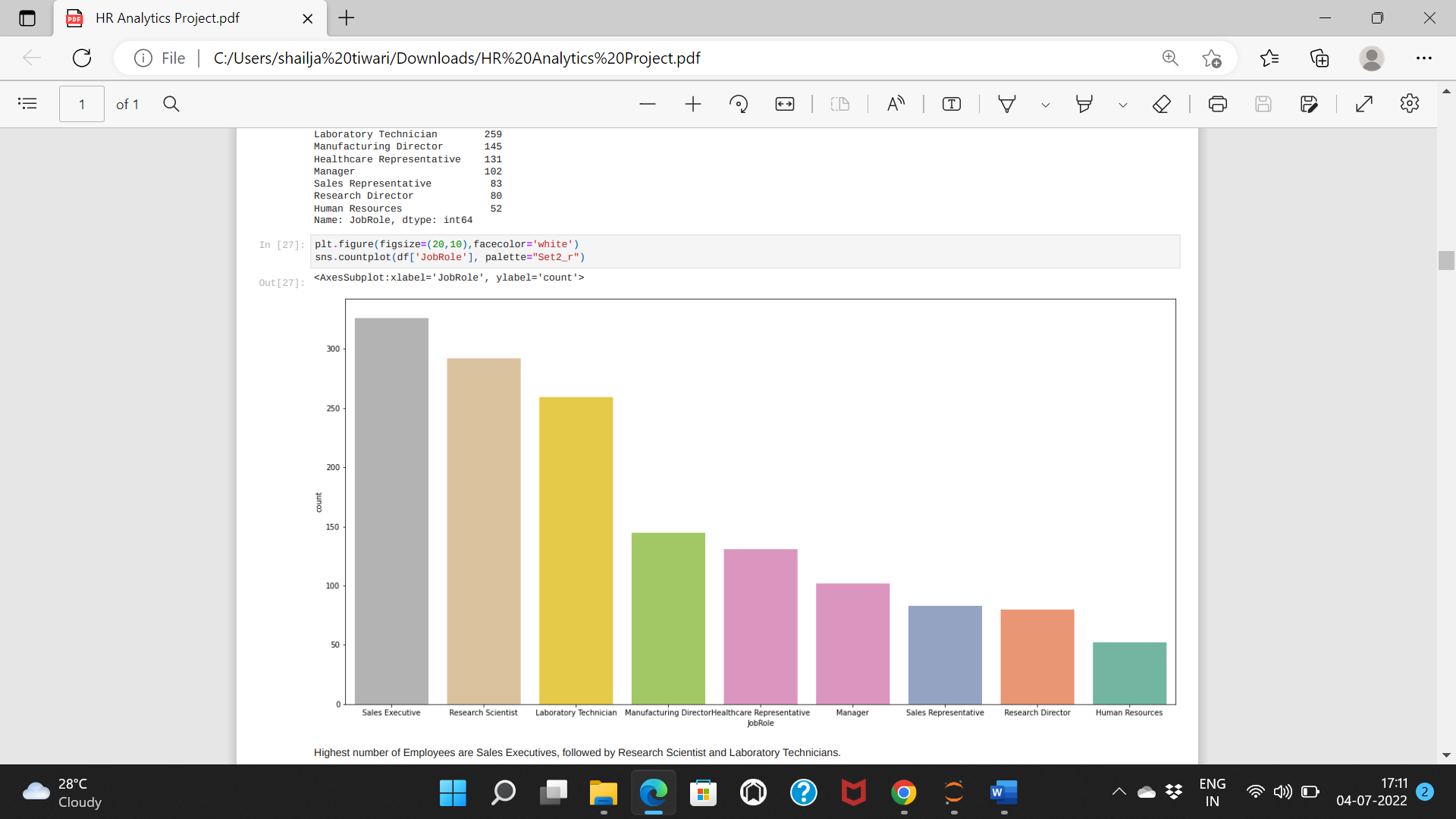
Most employees Travel Rarely



Most employees belong to Research and Development.



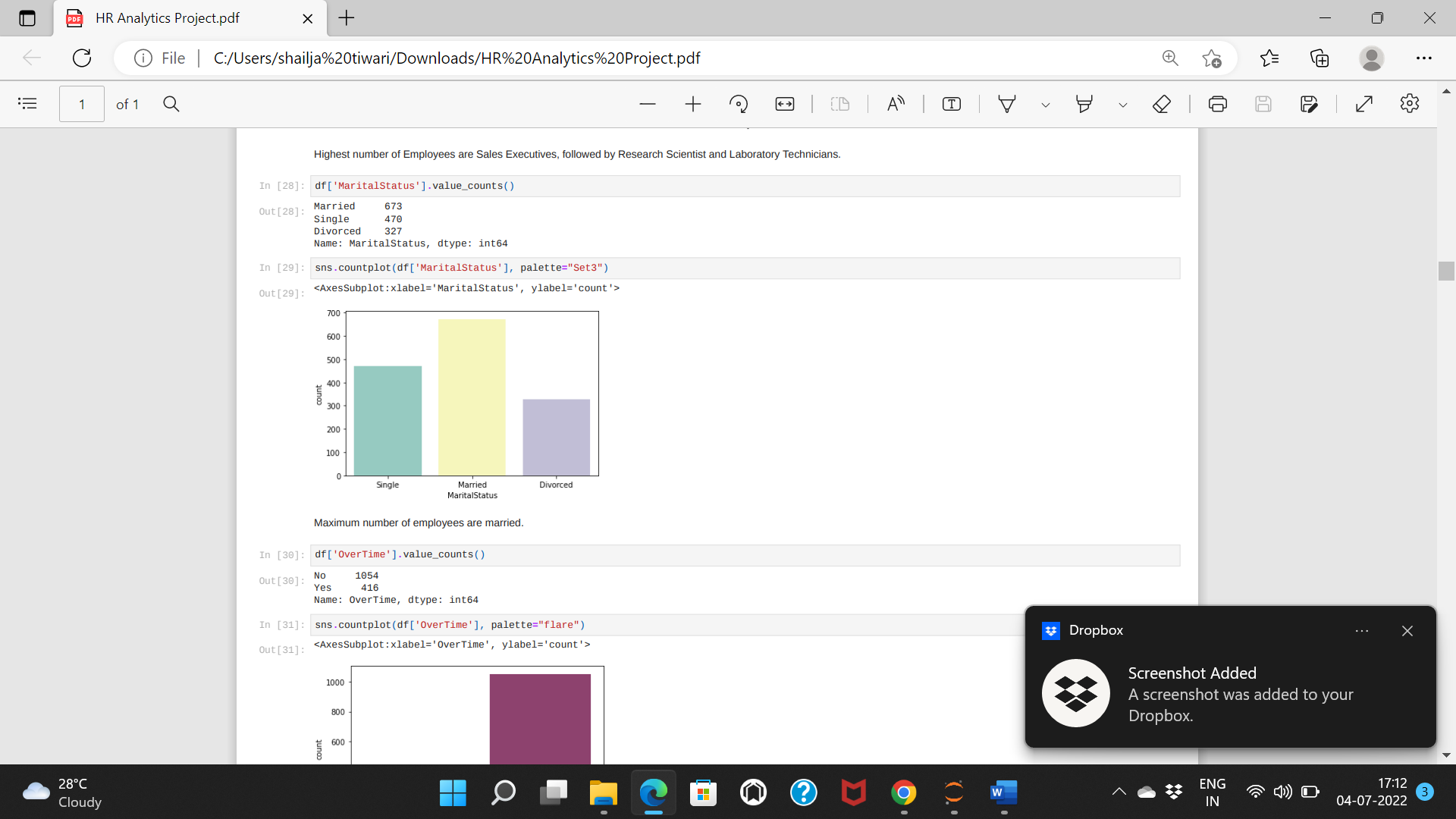
Life Sciences is the most common Education field which is found in this dataset.



Highest number of Employees are Sales Executives, followed by Research Scientist and Laboratory Technicians



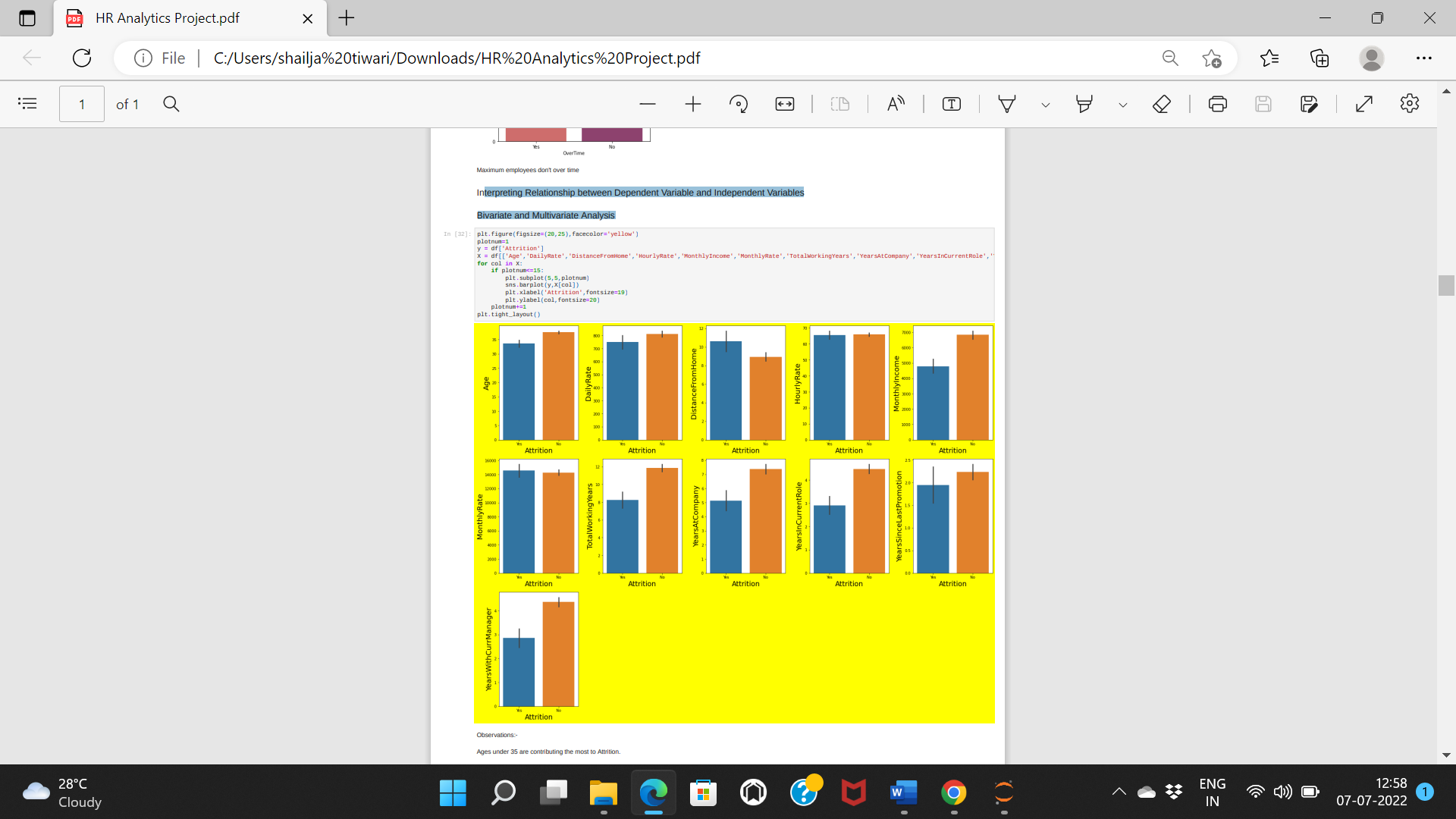
Maximum employees don't over time.

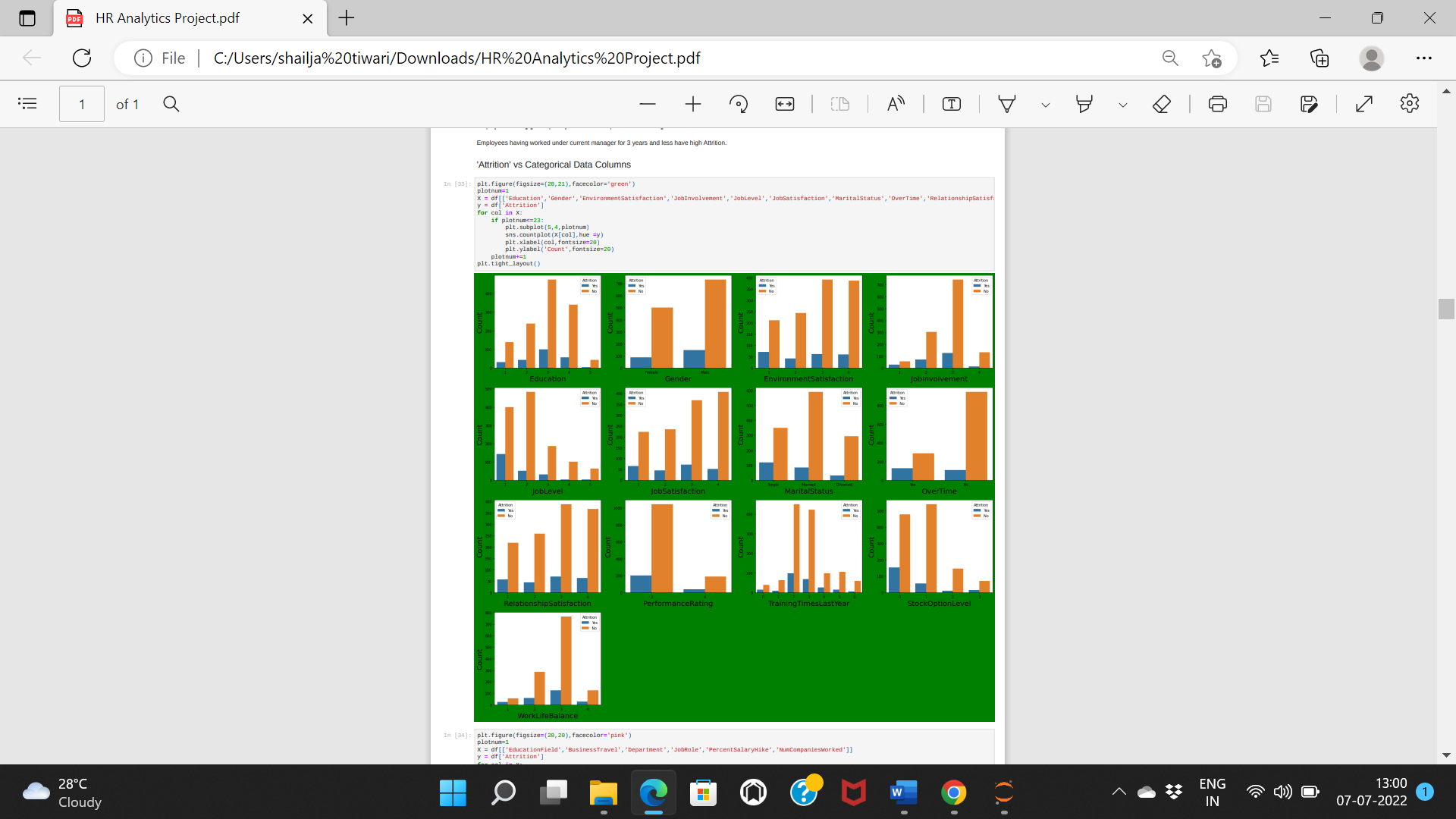


Maximum number of employees are married.

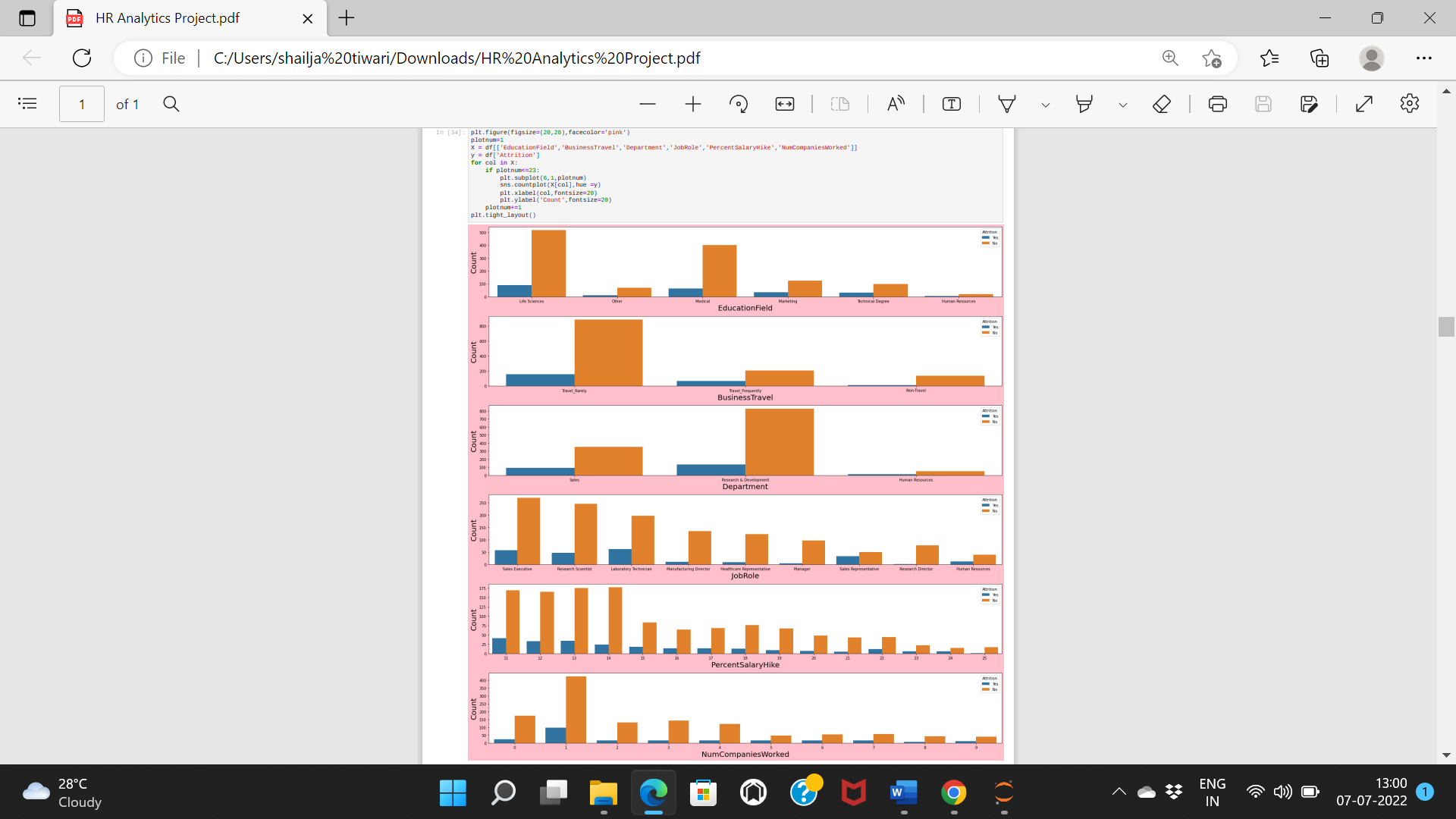
**Interpreting Relationship between Dependent Variable and Independent Variables**

* **Bivariate and Multivariate Analysis**





* Employees with Education '3' and '4' are retained the most.
* Employees with Education '3' mostly leave.
* Employees with EducationFields Life Sciences and Medical are retained the most.
* Employees with EducationFields Life Sciences mostly leave
* The number of male employees retained is a bit greater than the number of female employees retained.
* Male employees also contribute the most to Attrition.
* Employees with greater Environment Satisfaction are retained. While Employees with lowest Environment Satisfaction contribute the most to Attrition.
* Employees with Job Involvement of 3 are retained the most.



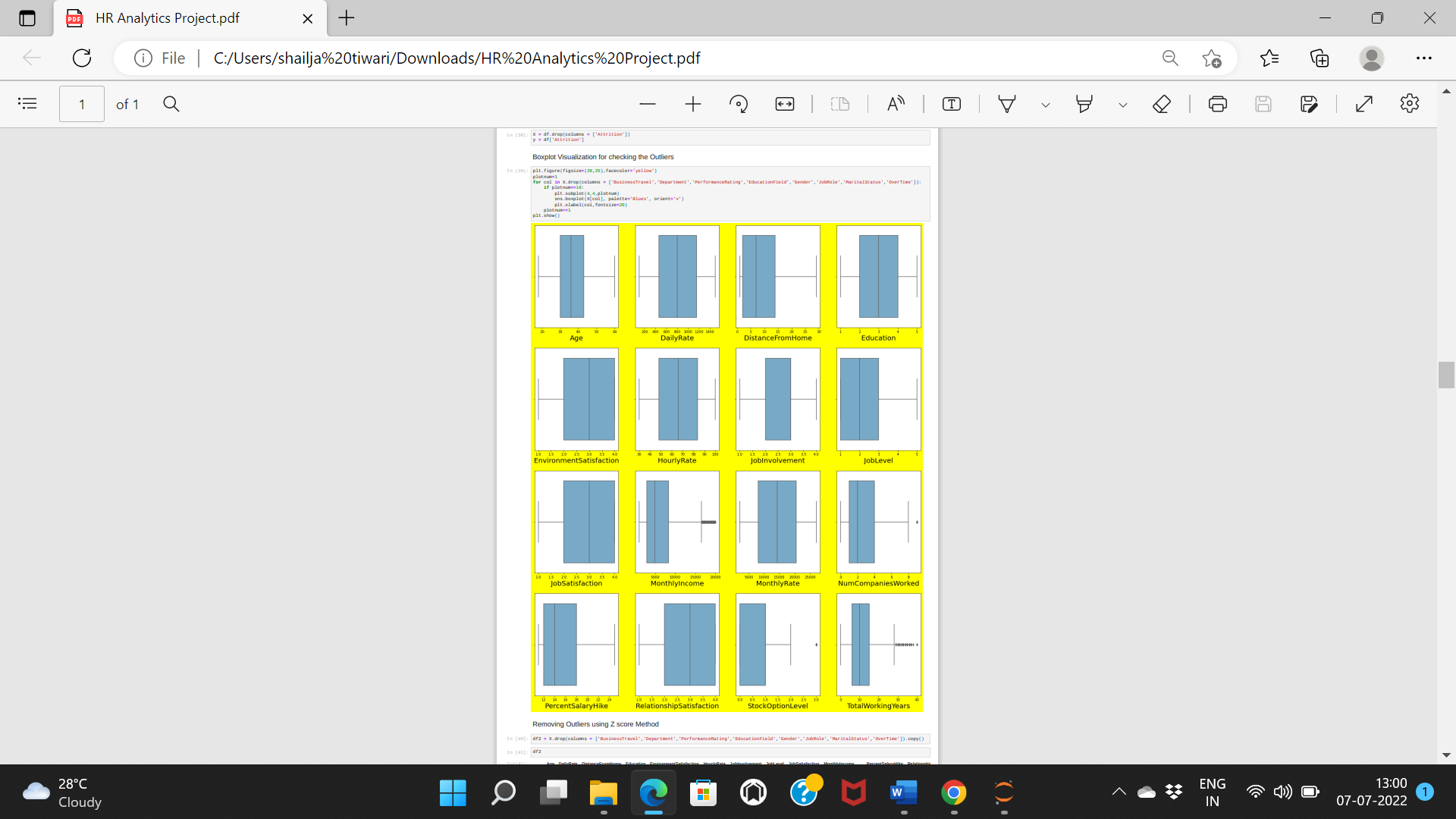
* Employees who Travel rarely are retained the most.
* Employees belonging to Reseach and Development are retained the most followed by Sales.
* While Employees belonging to Reseach and Development also contribute the most to Attrition.

4. Pre-Processing Pipeline.



**Outliers visualisation and removal**

**Use of Boxplot for visualising and checking outliers**



An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. During the model building and for precise and accurate predictions these normal values are removed.

Outliers are stragglers extremely high or extremely low values. For example, if you were measuring children's nose length, your average value might be thrown off if Pinocchio was in the class.

**1. Removing Outliers using Interquartile Range or IQR**

The IQR describes the middle 50% of values when ordered from lowest to highest. The IQR is then the difference between Third quartile and First quartile.

IQR=Q3-Q1

**2. Removing Outliers using Z score method**

* A Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values.
* Z-score is measured in terms of [standard deviations](https://www.investopedia.com/terms/s/standarddeviation.asp) from the mean.
* If a Z-score is 0, it indicates that the data point's score is identical to the mean score.
* A Z-score of 1.0 would indicate a value that is one standard deviation from the mean.
* Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

from scipy.stats import zscore zscor = zscore(df2) z\_score\_abs = np.abs(zscor)

df3 = df2[(z\_score\_abs < 3).all(axis=1)]

#taking 3 as threshold value

**Checking for skewness in the data set**

* The skewness is a measure of symmetry or asymmetry of data distribution, and kurtosis measures whether data is heavy-tailed or light-tailed in a normal distribution.
* Data can be positive-skewed (data-pushed towards the right side) or negative-skewed (data-pushed towards the left side).
* It is the asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right.

**Effect of skewness on the dataset**

* Skewed data degrades the model's ability (especially regression based models) to describe typical cases as it has to deal with rare cases on extreme values.

**To check for skew in data:**

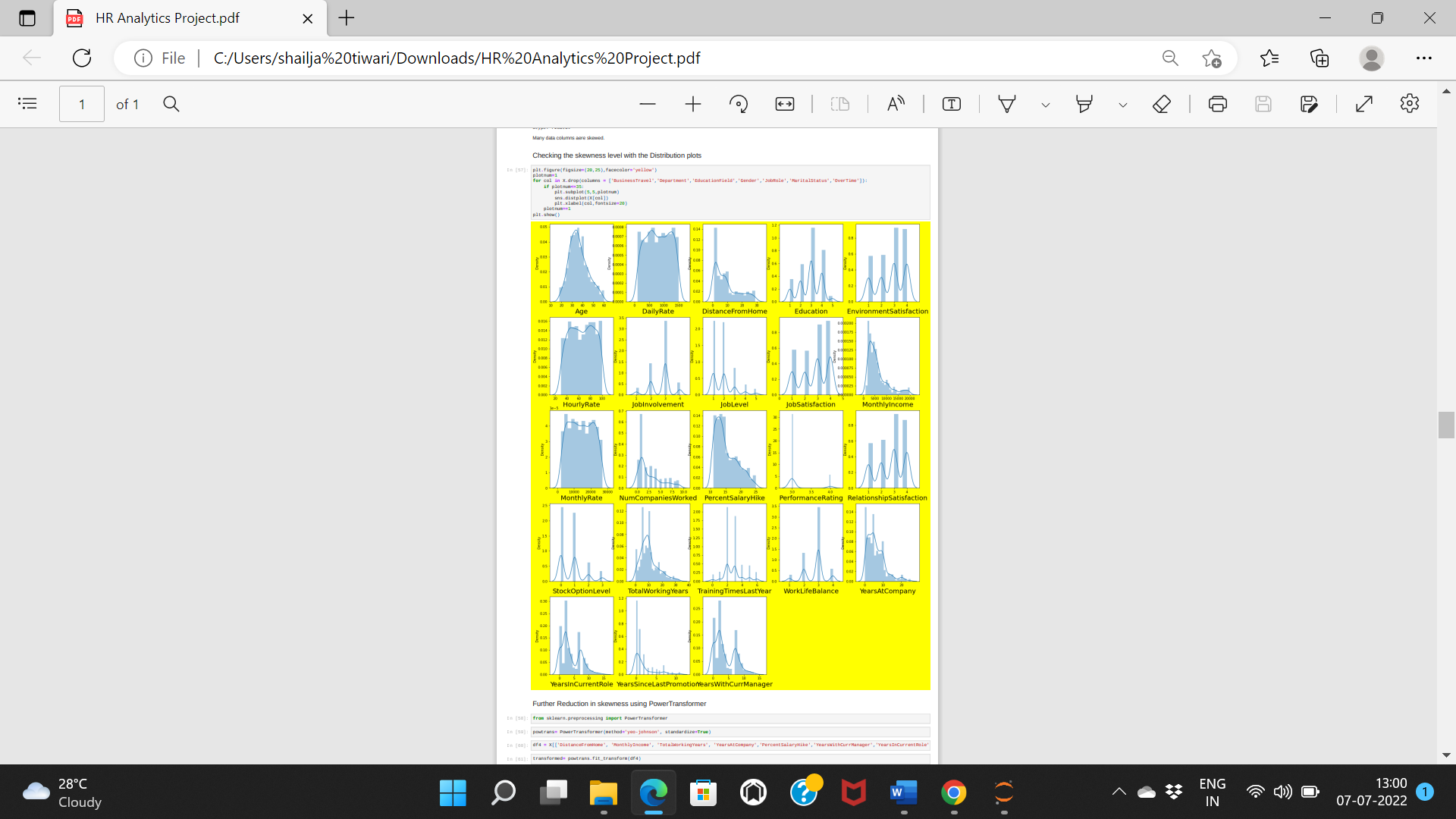
* df.skew().sort\_values(ascending=False)

**Reducing skewness in the dataset and for receiving a normal distribution**

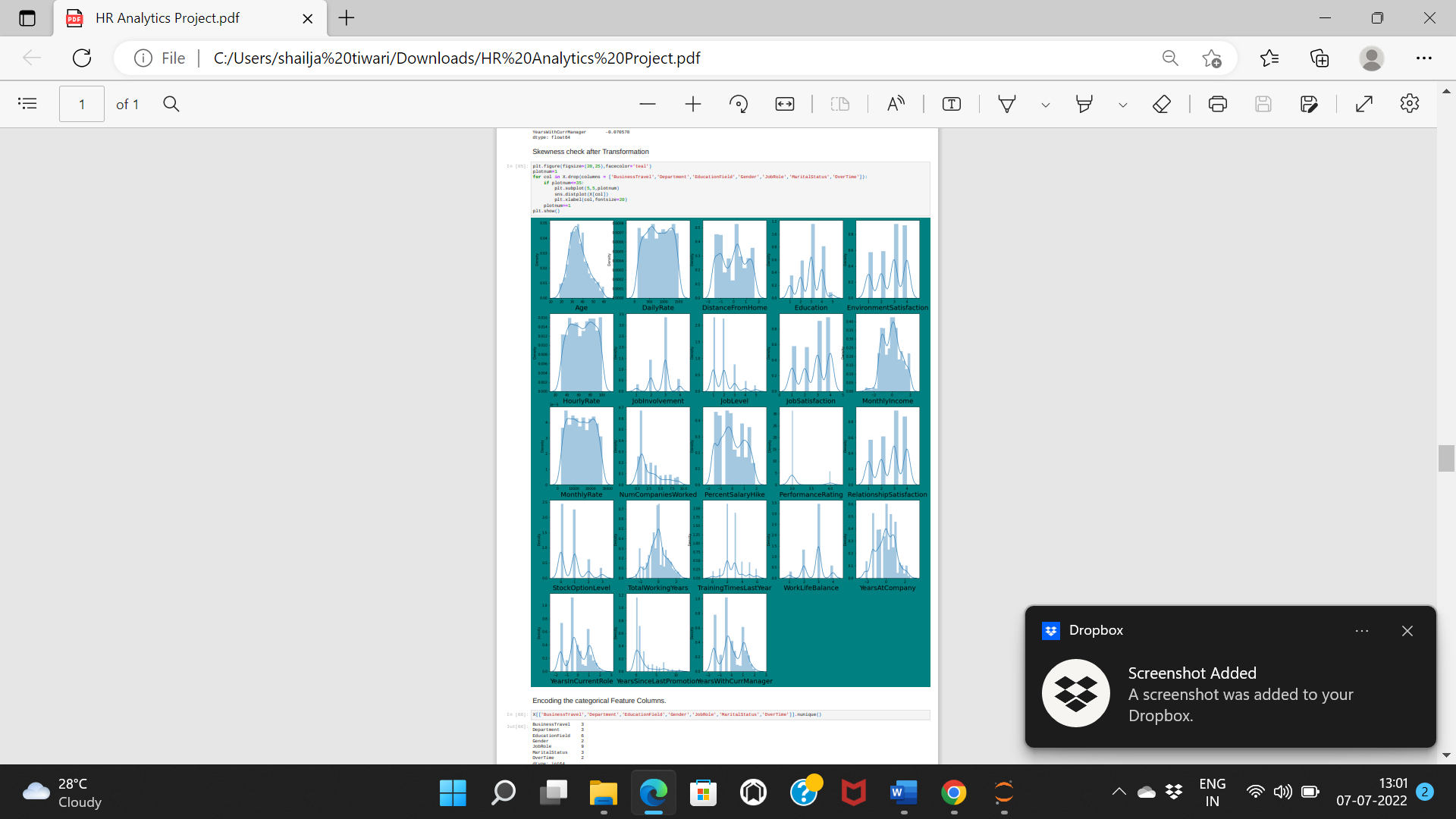
**Using Power transformer technique**

* from sklearn.preprocessing import PowerTransformer
* powtrans= PowerTransformer(method='yeo-johnson', standardize=True)
* df4 = X[['DistanceFromHome', 'MonthlyIncome', 'TotalWorkingYears', 'YearsAtCompany','PercentSalaryHike','YearsWithCurrManager','YearsInCurrentRole' ]]
* transformed= powtrans.fit\_transform(df4)
* transformed = pd.DataFrame(transformed, columns=df4.columns)
* transformed.skew()

**Visualisation using distribution plot.**



**Normal distribution of skewed column**

****

**Encoding the dataset using Label Encoder**

**Label Encoding or Ordinal Encoding**

* We use this categorical data encoding technique when the categorical feature is ordinal. In this case, retaining the order is important. Hence encoding should reflect the sequence.
* In Label encoding, each label is converted into an integer value. We will create a variable that contains the categories representing the education qualification of a person. X[['BusinessTravel','Department','EducationField','Gender','JobRole','MaritalStatus','OverTime']].nunique()
* df\_train\_transformed = encoder.fit\_transform(train\_df)

**Encoding using get\_dummies()**

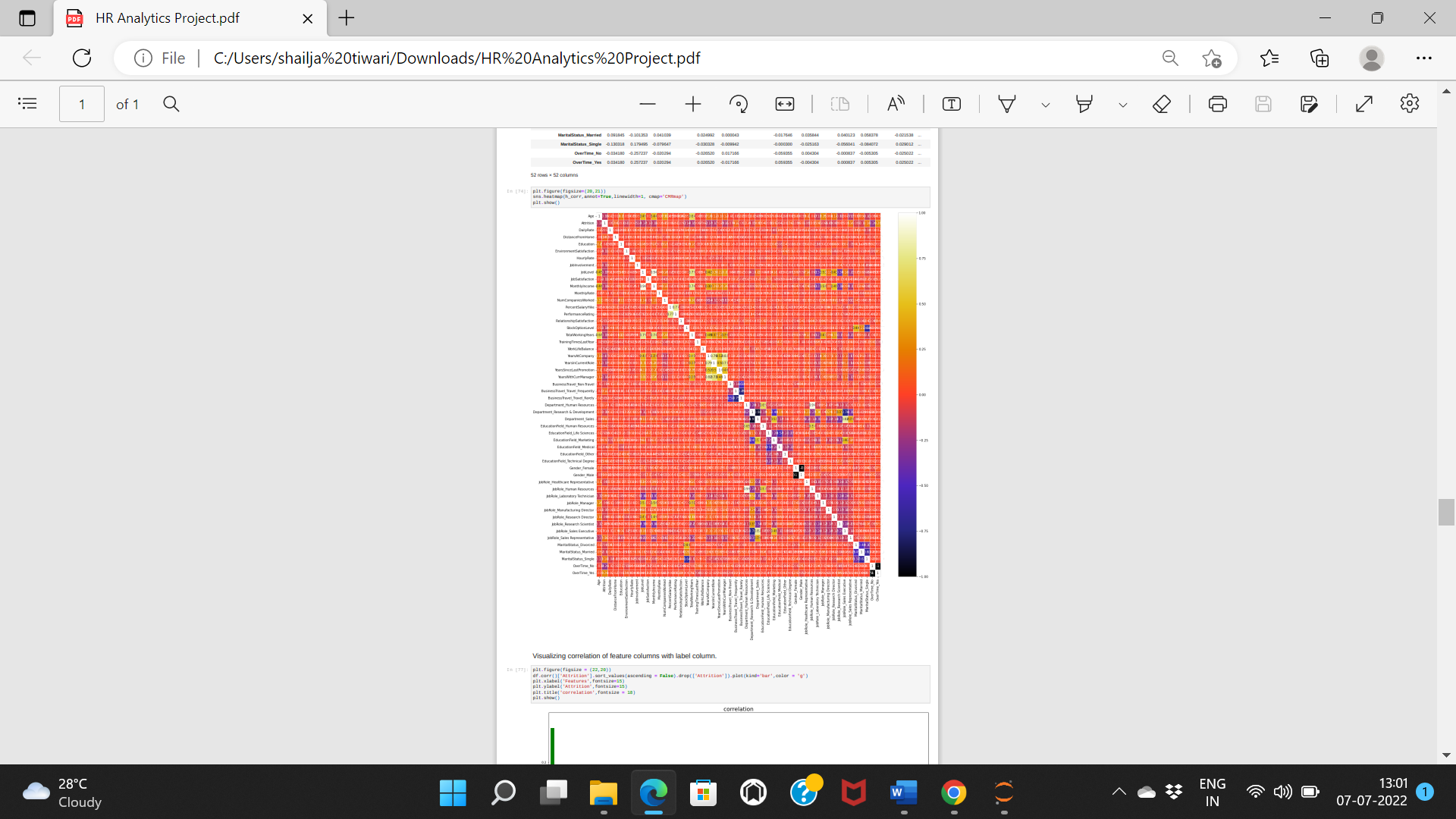
* dumm=pd.get\_dummies(X[['BusinessTravel','Department','EducationField','Gender','JobRole','MaritalStatus','OverTime']],drop\_first = False)
* dumm

**correlation matrix**

* Correlation is an indication about the changes between two variables.
* A matrix is an array of numbers arranged in rows and columns.
* A correlation matrix is simply a table showing the correlation coefficients between variables. Here, the variables are represented in the first row, and in the first column

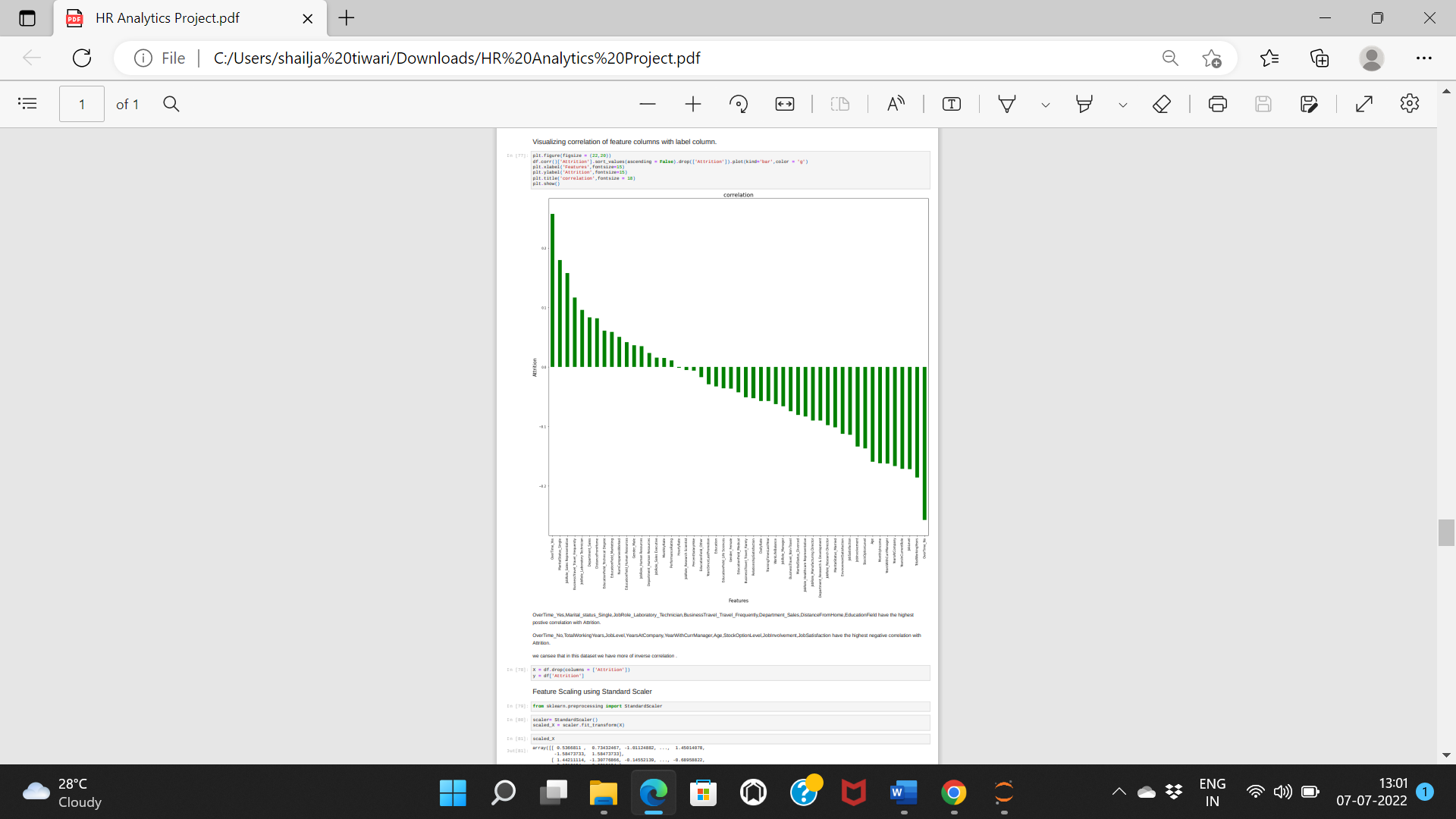
**Establishing the correlation between various columns**

* h\_corr = df.corr()
* h\_corr



**Visualizing correlation of feature columns with label column.**

* plt.figure(figsize = (22,20))
* df.corr()['Attrition'].sort\_values(ascending = False).drop(['Attrition']).plot(kind='bar',color = 'g')
* plt.xlabel('Features',fontsize=15) plt.ylabel('Attrition',fontsize=15) plt.title('correlation',fontsize = 18)
* plt.show()



**Feature Scaling using Standard Scaler**

* Feature Scaling is a technique of bringing down the values of all the independent features of our dataset on the same scale. Feature selection helps to do calculations in algorithms very quickly. It is the important stage of data pre-processing.

**Feature Scaling using Standard Scaler**

Standard Scaler removes the mean and scales the data to the unit variance. Standard Scaler comes into play when the characteristics of the input dataset differ greatly between their ranges, or simply when they are measured in different units of measure.

* from sklearn.preprocessing import StandardScaler
* scaler= StandardScaler()
* scaled\_X = scaler.fit\_transform(X)
* scaled\_X

**Multicollinearity**

Multicollinearity is a well-known challenge in multiple regression. The term refers to the high correlation between two or more explanatory variables, i.e. predictors.

According to [Graham’s study](http://brianmcgill.org/614/reading/graham2003.pdf?ref=hackernoon.com), multicollinearity in multiple regression leads to:

* Inaccurate parameter estimates,
* Decreased power
* Exclusion of significant predictors

**Dealing with Multicollinearity**

* You can deal with multicollinearity by
* dropping highly correlated variables
* extracting new features with Principal Component Analysis (PCA)

**Multicollinearity using Variance Inflation Factor**

* from statsmodels.stats.outliers\_influence import variance\_inflation\_factor
* vif = pd.DataFrame()
* vif["Features"] = X.columns vif['vif'] = [variance\_inflation\_factor(scaled\_X,i) for i in range(scaled\_X.shape[1])]
* vif

**Selecting best features based on their scores and dropping a highly colinear columns.**

* **X\_best = X.drop(columns=['MonthlyIncome','YearsSinceLastPromotion','Department\_Human Resources','EducationField\_Other','JobRole\_Sales Executive',' MonthlyIncome]**
* **x\_best**

5. Building Machine Learning Models.

**Machine Learning**

A machine learning model is defined as a mathematical representation of the output of the training process.

Machine learning is the study of different algorithms that can improve automatically through experience & old data and build the model.

A machine learning model is similar to computer software designed to recognize patterns or behaviours based on previous experience or data.

The learning algorithm discovers patterns within the training data, and it outputs an ML model which captures these patterns and makes predictions on new data.

**Classification of Machine Learning Models:**

Based on different business goals and data sets, there are three learning models for algorithms. Each machine learning algorithm settles into one of the three models:

* Supervised Learning
* Unsupervised Learning
* Reinforcement Learning

**Supervised Learning is further divided into two categories:**

* Classification
* Regression

**Unsupervised Learning is also divided into below categories:**

* Clustering
* Association Rule
* Dimensionality Reduction

**Since the Dataset is the classification problem as the Label data is non-numerical data Classification Model Building**

### **Classification**

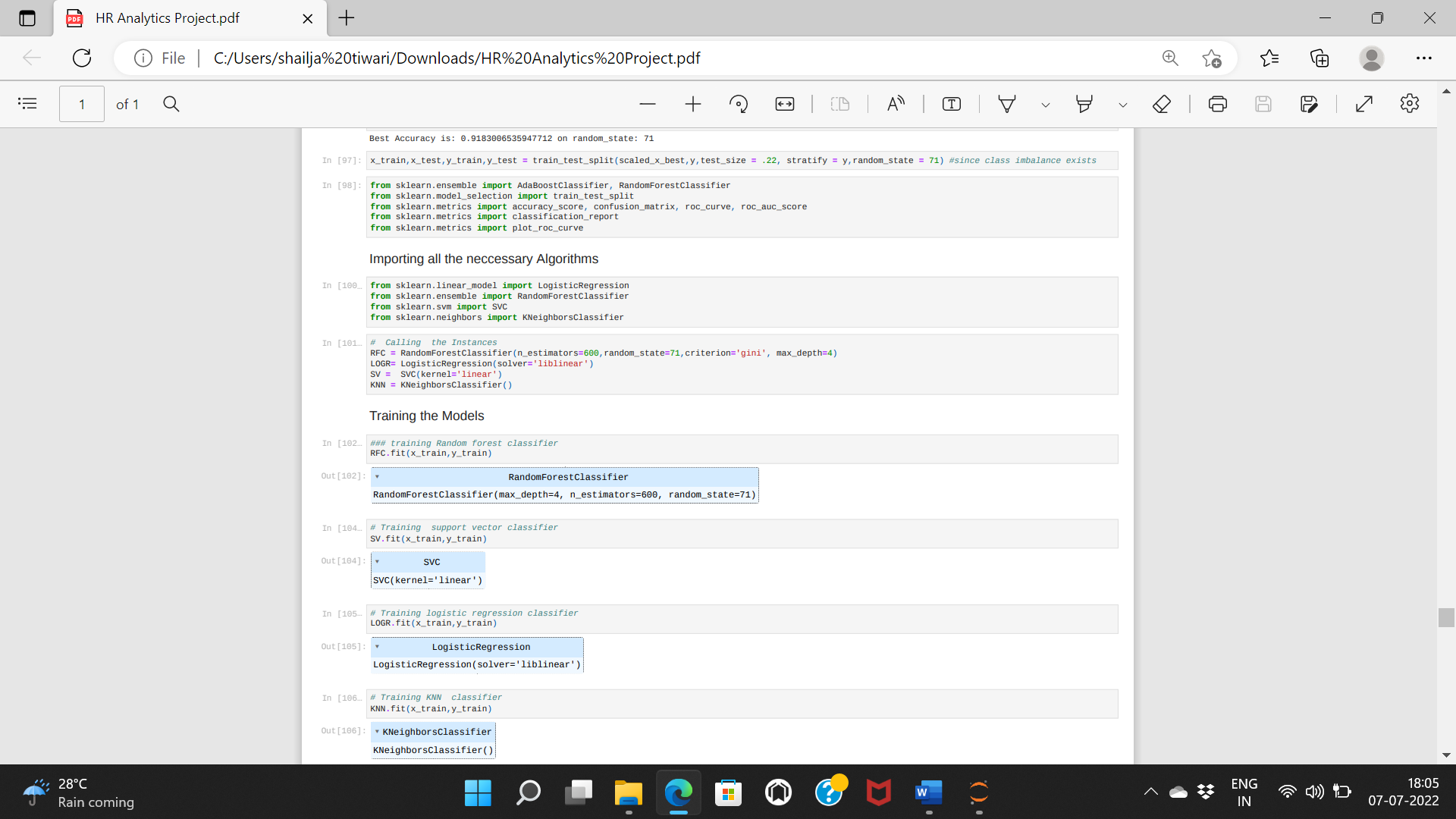
Classification models are the second type of Supervised Learning techniques, which are used to generate conclusions from observed values in the categorical form. For example, the classification model can identify if the email is spam or not; a buyer will purchase the product or not, etc. Classification algorithms are used to predict two classes and categorize the output into different groups.

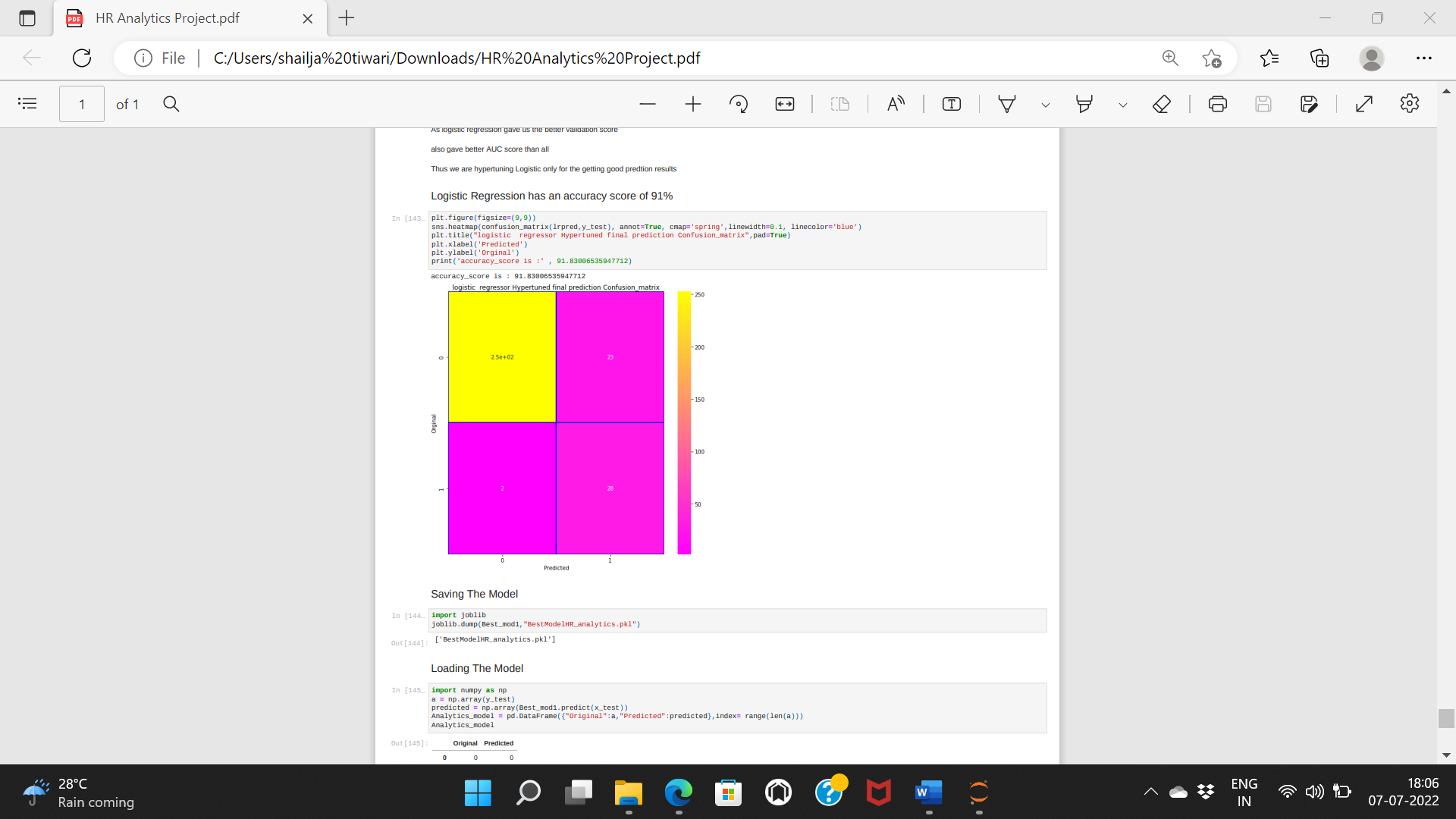
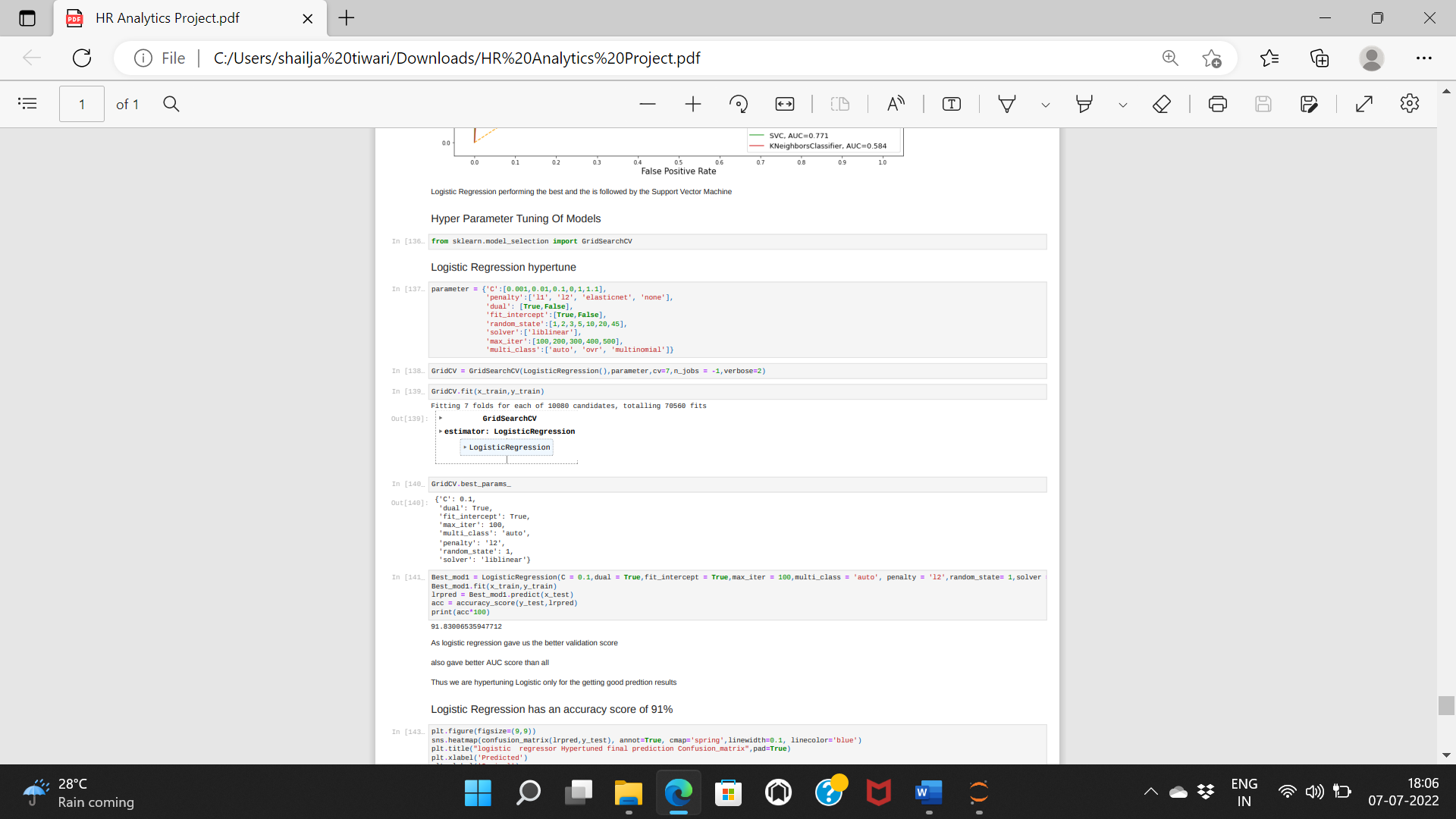
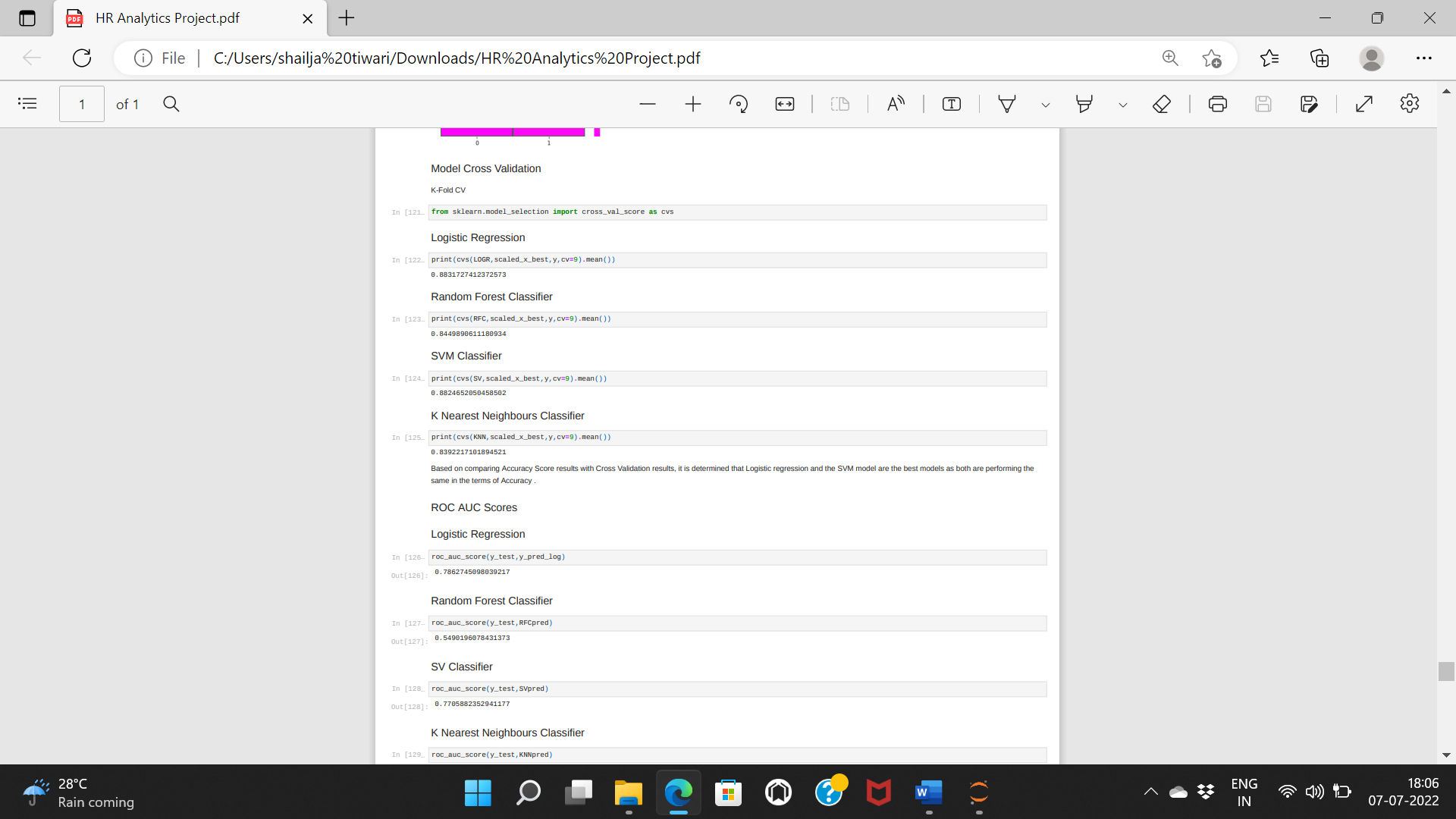
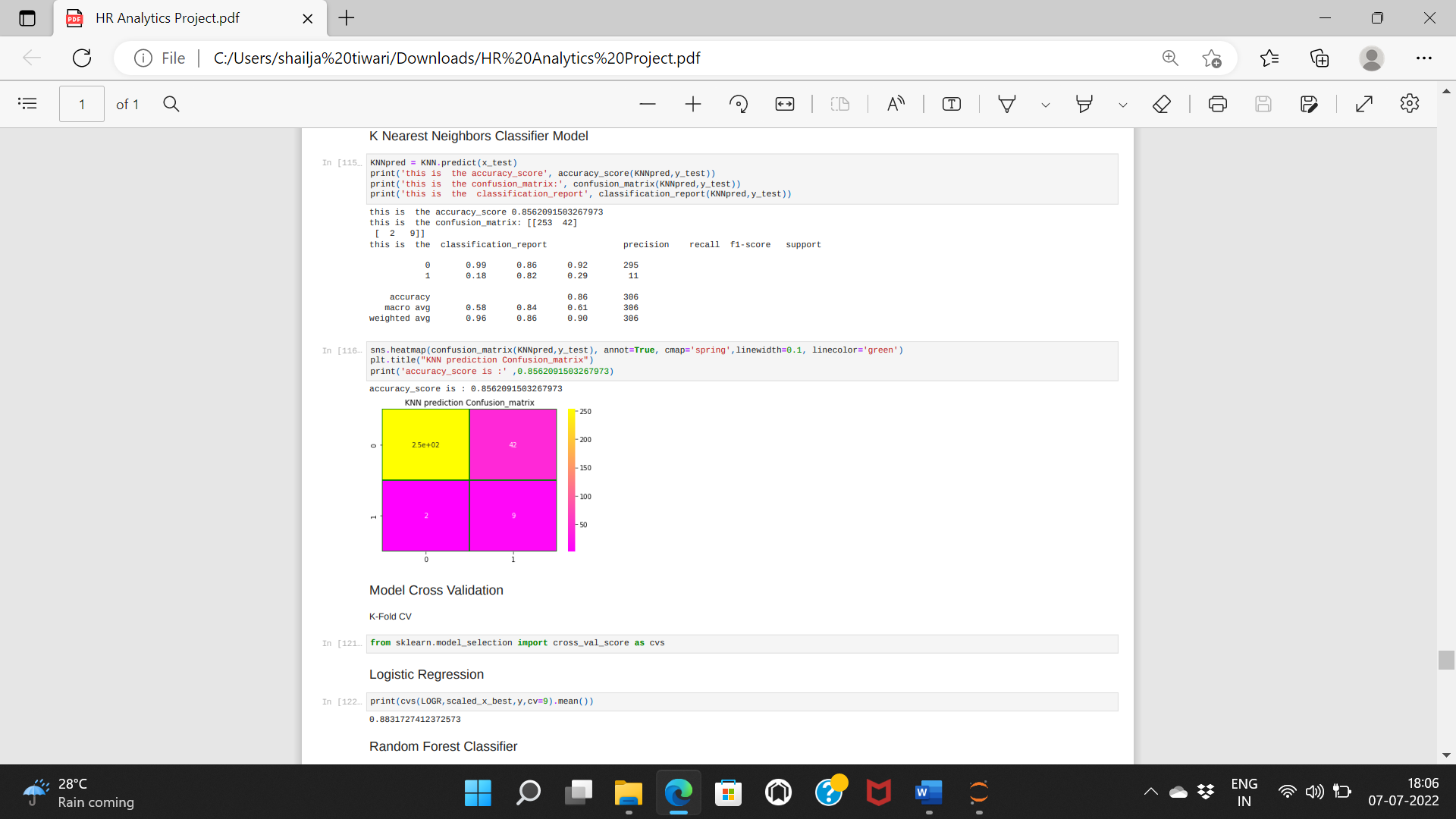
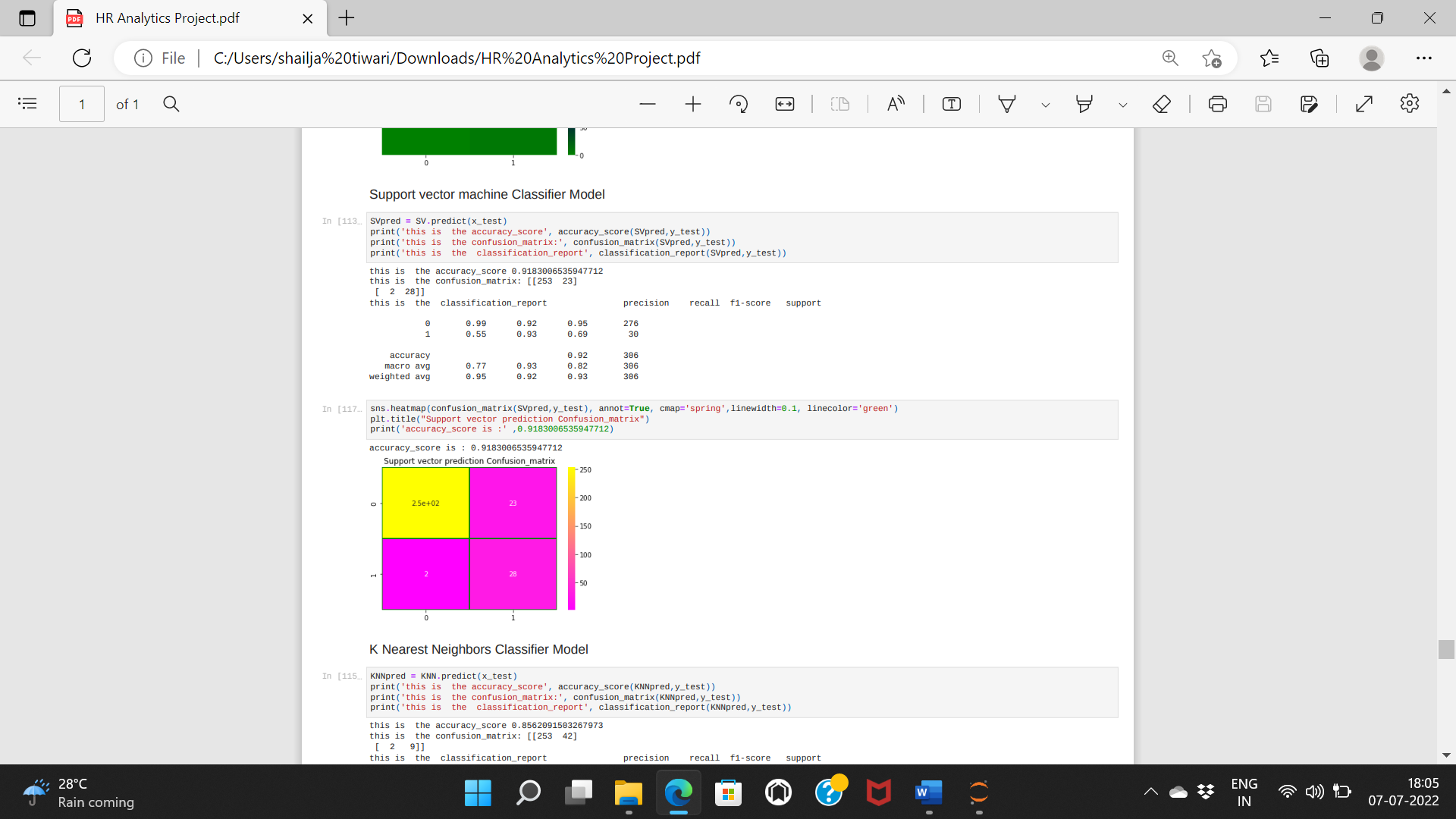
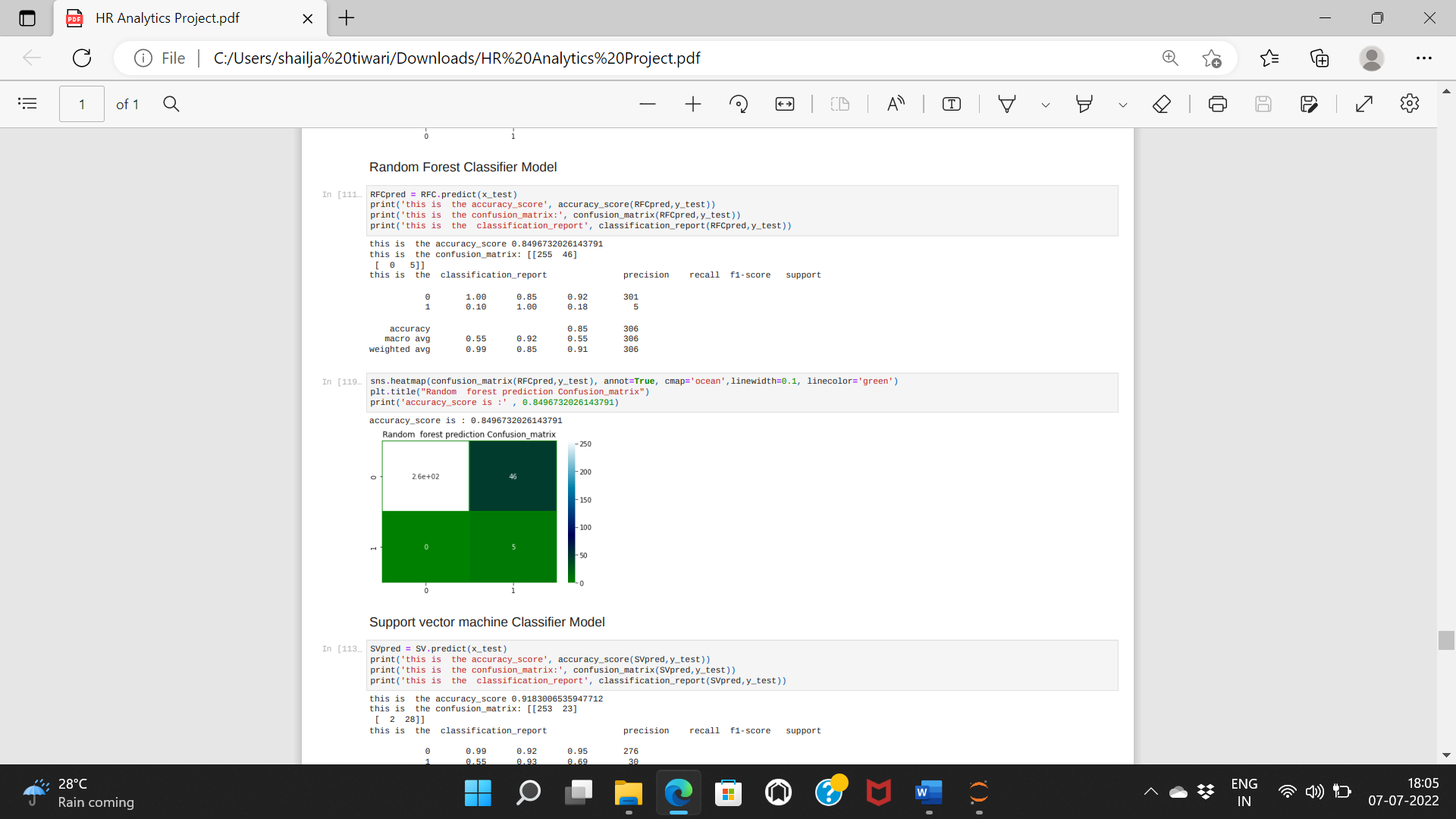
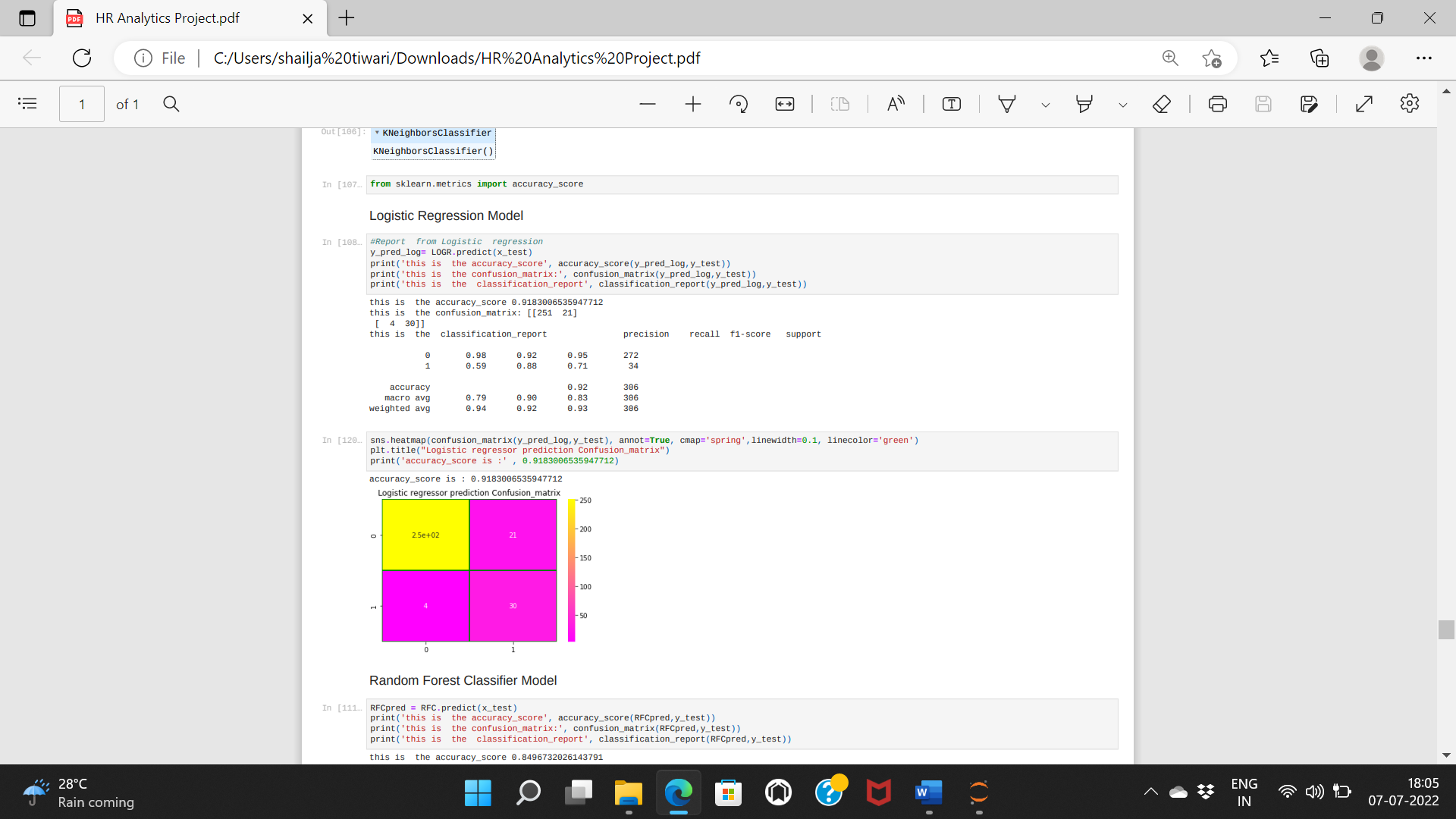
**Classification Model Building**

* from sklearn.model\_selection import train\_test\_split
* from sklearn.metrics import accuracy\_score

### **Finding the best random state**

* from sklearn.linear\_model import LogisticRegression
* maxAcc = 0
* maxRS=0
* for i in range(1,100):
* x\_train,x\_test,y\_train,y\_test = train\_test\_split(scaled\_x\_best,y,test\_size = .22, random\_state = i,stratify = y)
* modRF = LogisticRegression()
* modRF.fit(x\_train,y\_train)
* pred = modRF.predict(x\_test)
* acc = accuracy\_score(y\_test,pred)
* if acc>maxAcc:
* maxAcc=acc
* maxRS=i
* print(f"Best Accuracy is: {maxAcc} on random\_state: {maxRS}")

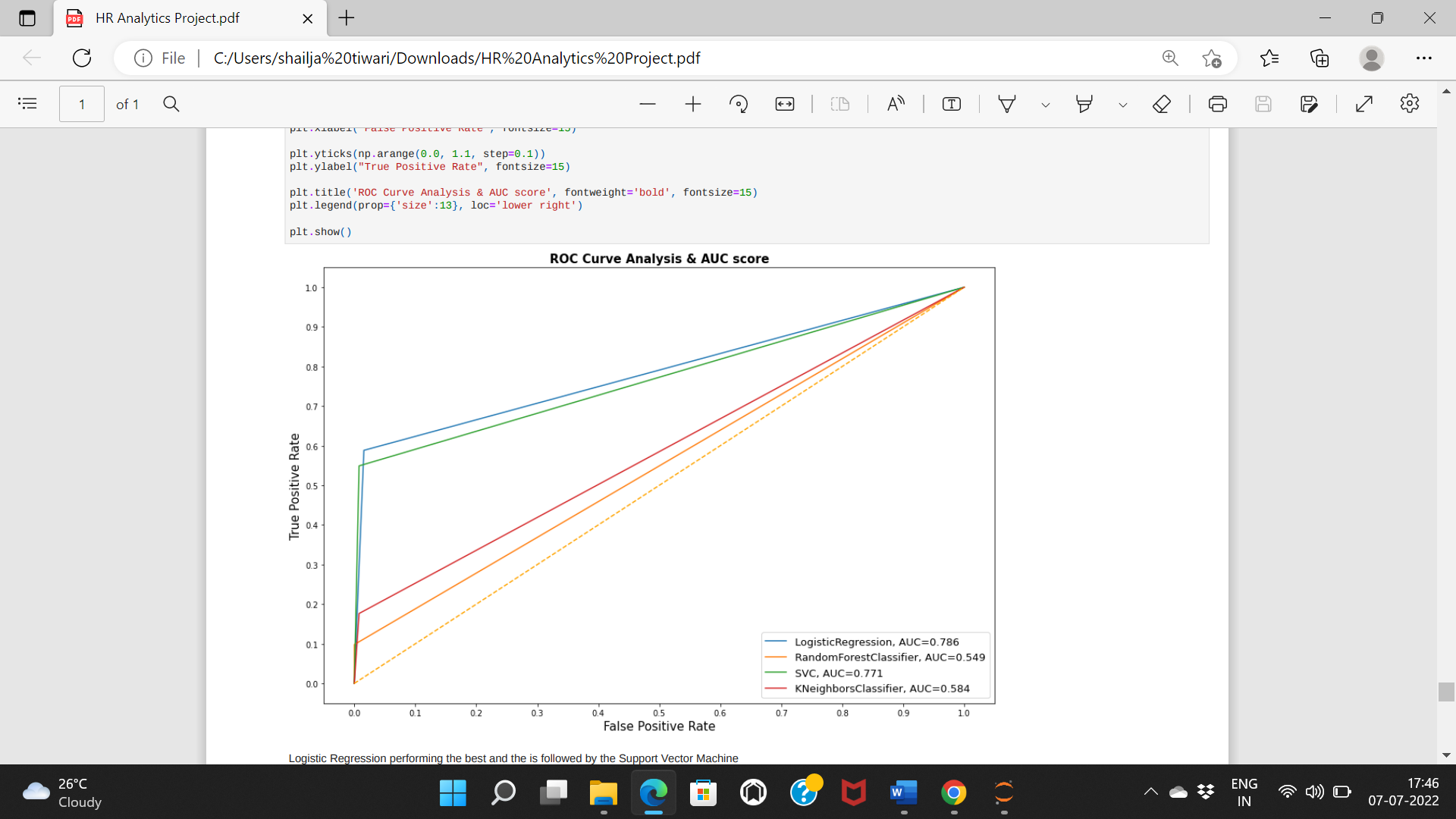
****

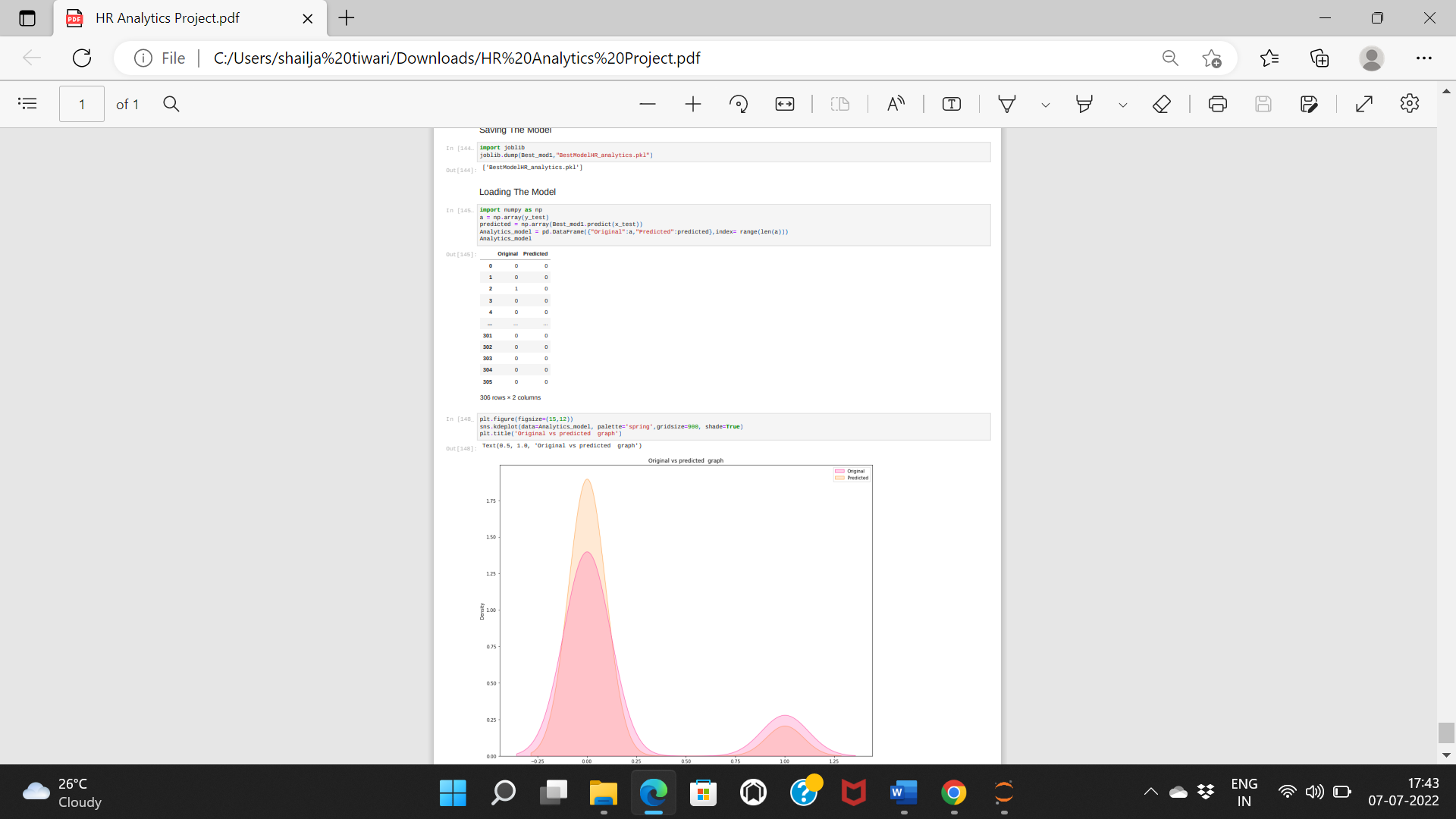
****

**Best Accuracy is: 0.9183006535947712 on random\_state: 71**

As logistic regression gave us the better validation score also gave better AUC score than all Thus we are hypertuning Logistic only for the getting good prediction results

Logistic Regression has an accuracy score of 91%





6. Concluding Remarks/Key Findings.

* Feature column contains mixed type of data i.e. continuous and categorical. Target/label column contains categorical data
* There are no null values in the dataset
* Target column have 2 unique values "Yes", "No"
* There are 3 columns with one value present in them i.e. employee count, over 18, standard hours
* In most columns mean and 50% are similar in value, and mean is greater than standard deviation, indicating a symmetric distribution.
* TotalWorkingYears,YearsAtCompany,YearsInCurrentRole,YearsSinceLastPromotion,YearsWithCurrManager have a big difference between 75% and max indicating the presence of outliers. as the std deviation is high so there are chances that - the dataset has the skewness present in it
* 2 unique categorical values in the Label column / target variable.
* Class- 'No' : Has 1233 values 'Yes' : Has 237 values
* There is class imbalance in the Target Data
* Class 'No' = 83.88% of total values 'Yes'= 16.12% of total values
* Classes are imbalanced
* Female employees are much less than Male employees
* Most employees Travel Rarely.
* Most employees belong to Research and Development.
* Life Sciences is the most common Education field among this dataset.
* Highest number of Employees are Sales Executives, followed by Research Scientist and Laboratory Technicians.
* Maximum number of employees are married.
* Maximum employees don't over time
* Ages under 35 are contributing the most to Attrition.
* High number of people who are 35 and above have been retained.
* In terms of DailyRate, there isn't much of a difference in number between those who left and those who did not, however attrition is higher amongst those with lower DailyRate
* Distance from home greater than 10 is a deciding factor for Attrition
* HourlyRate doesn't seem to contribute to Attrition.
* Those employees whose MonthlyIncome is below 5000 are contributing to Attrition
* MonthlyRate doesn't have a big effect on Attrition
* Employees having Total working years under 9 are mostly contributing to Attrition
* Employees having worked 6 years at the company or less have more impact on Attrition.
* Employees having gone up to 2 years since last promotion have higher Attrition.
* Employees having worked under current manager for 3 years and less have high Attrition.
* Employees with Education '3' and '4' are retained the most.
* Employees with Education '3' mostly leave.
* Employees with EducationFields Life Sciences and Medical are retained the most.
* Employees with EducationFields Life Sciences mostly leave.
* Employees who Travel rarely are retained the most.
* Employees belonging to Research and Development are retained the most followed by Sales. While Employees belonging to Research and Development also contribute the most to Attrition.
* The number of male employees retained is a bit greater than the number of female employees retained. Male employees also contribute the most to Attrition.
* Employees with greater Environment Satisfaction are retained. While Employees with lowest Environment Satisfaction contribute the most to Attrition.
* Employees with Job Involvement of 3 are retained the most.
* Employees with Job level 3 and 2 are retained the most. While employees with job Level 1 contribute the most to Attrition.
* Sales Executives,Research Scientists, Laboratory Technicians are amongst the most retained employees. Sales Executives and Laboratory Technicians also contribute more to attrition than Research Scientists.
* Employees with Highest Job Satisfaction of 4 and 3 are amongst the most retained employees.
* More Married Employees are retained than single and divorced.
* Employees who don't do overtime are retained the most.
* Employees who have worked in 1 and 0 companies before are retained more than those with work history in multiple companies. While employees who have worked in 1 company before are the most to leave.
* Employees with lowest Salary percentage hike contribute most to Attrition.
* Employees with Highest Relationship Satisfaction of 3 and 4 are those who are retained the most.
* Employees with Performance Rating 3 are retained a lot more than those with Performance Rating 4
* Employees with Training Times of 2 and 3 Last Year are retained the most.
* Employees with StockOption level 1 are retained the most. While Employees with stock optoion level 0 contribute most to Attrition.
* Employees with worklife balance of 3 are retained the most.
* OverTime\_Yes,Marital\_status\_Single,JobRole\_Laboratory\_Technician,BusinessTravel\_Travel\_Frequently,Department\_Sales,DistanceFromHome,EducationField have the highest positive correlation with Attrition.
* OverTime\_No,TotalWorkingYears,JobLevel,YearsAtCompany,YearWithCurrManager,Age,StockOptionLevel,JobInvolvement,JobSatisfaction have the highest negative correlation with Attrition.
* we can see that in this dataset we have more of inverse correlation .
* Based on comparing Accuracy Score results with Cross Validation results, it is determined that Logistic regression and the SVM model are the best models as both are performing the same in the terms of Accuracy .
* Logistic Regression performing the best and the is followed by the Support Vector Machine
* As logistic regression gave us the better validation score
* also gave better AUC score than all
* Thus, we are hypertuning Logistic only for the getting good prediction results

### Logistic Regression has an accuracy score of 91%.