**Section - b : 25 Marks**

**Instructions :**

1. **This section is openbook**
2. **Write the answer next to the question in this word document.**
3. **Submit this word document and the R file in a zip folder**

Attribute Information:

Input variables:

# bank client data:

1 - age (numeric)

2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')

3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)

4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')

5 - default: has credit in default? (categorical: 'no','yes','unknown')

6 - housing: has housing loan? (categorical: 'no','yes','unknown')

7 - loan: has personal loan? (categorical: 'no','yes','unknown')

# related with the last contact of the current campaign:

8 - contact: contact communication type (categorical: 'cellular','telephone')

9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')

10 - day\_of\_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')

11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.

# other attributes:

12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14 - previous: number of contacts performed before this campaign and for this client (numeric)

Output variable (desired target):

15 - y - has the client subscribed a term deposit? (binary: 'yes','no')

**Logistic Regression and Trees (Classification Problems) : 6 Marks**

**Q1. Answer the following questions from the dataset “bank-full.csv”**

Read the dataset and split into test and training sets and before splitting set the seed to 1000 and 60% should go into training set.

1. Build a logistic regression model(model1) for predicting “y” with the help of the variables “age”, “balance”, ”campaign” and “duration”. Build another regression model (model2) with above mentioned attributes excluding “campaign”. Specify the AIC value in both the models and mention which is the best model among both.

**Answer : AIC model1: 16262; AIC model2: 16407; Best model is model1, because it has lower AIC**

2. Compute the values of Sensitivity, Specificity for above model (with campaign).

**Answer : Sensitivity: 0.1645131; Specificity: 0.9854715**

3. Make predictions on the test set and Compute the AUC of the “model1”

**Answer : AUC model1 value = 0.8179142**

4. Build a CART model for predicting “y” with the help of the variables “age”, “balance” and “duration”. Plot it and mention the number of splits you see in the plot.

**Answer : Number of splits: 2**

**5.** Make predictions on test data using the model created in above Problem 2 and compute the value of AUC.  **0.8097155**

6. What proportion of the customers are “Married” and have “technician” job. **Answer : proportion = 0.08962421**

**Text Analytics and Clustering**

**Q2. Answer the below questions from the dataset “Movies.txt”**

Load the data into R and assign the following variables as the column names in the same order.

"ID","Title","ReleaseDate","VideoReleaseDate","IMDB","Unknown","Action","Adventure","Animation","Childrens","Comedy","Crime","Documentary","Drama","Fantasy","FilmNoir","Horror","Musical","Mystery","Romance","SciFi","Thriller","War","Western"

1) Eliminate the first four variables from the dataframe. What is the number of movies which belong to both “action” and “horror” category. **1Mark**

**Answer : 13**

2) Build a hierarchical clustering model with the Euclidean distances. Plot the dendogram. What is the number of clusters at a height of 150.  **1Mark**

**Answer : 4**

3) Split the above model into 7 clusters. What are the clusters with maximum and minimum number of observations.  **1Mark**

**Answer : Max: 2 and Min: 7**

4) What is the number of Adventure category movies in Cluster 1 of the above model.  **1Mark**

**Answer : 56**

5) Which is the cluster with highest number of movies belonging to “Children” category.  **1Mark**

**Answer : Cluster 1**

6) Which is/are the clusters with least number of movies belonging to “Fantasy” category. **1Mark**

**Answer : Cluster 3,4,5,67 with 0**

7) Build a K-means clustering model with seed value 1000 and same number of clusters. Mention the clusters which have the highest and least number of observations.

**0.5Mark**

**Answer : Max: 6 and Min: 1**

8) Which Hierarchical Cluster best corresponds to K-Means Cluster 6? **0.5**

**Answer : Cluster 3**

9) Which Hierarchical Cluster best corresponds to K-Means Cluster 4? **0.5**

**Answer : Cluster 4**

10) Which Hierarchical Cluster best corresponds to K-Means Cluster 3? **0.5**

**Answer : Cluster 2**

11) Which K-means cluster has got more number of movies belonging to “Action” genre.**0.5**

**Answer : Cluster 3**

12) Which K-means cluster has got more number of movies belonging to “War” genre. **0.5**

**Answer : Cluster 6**

**Text Analytics : 5 Marks**

**Q3.** **Answer the below questions using the dataset “energy\_readings.csv”**

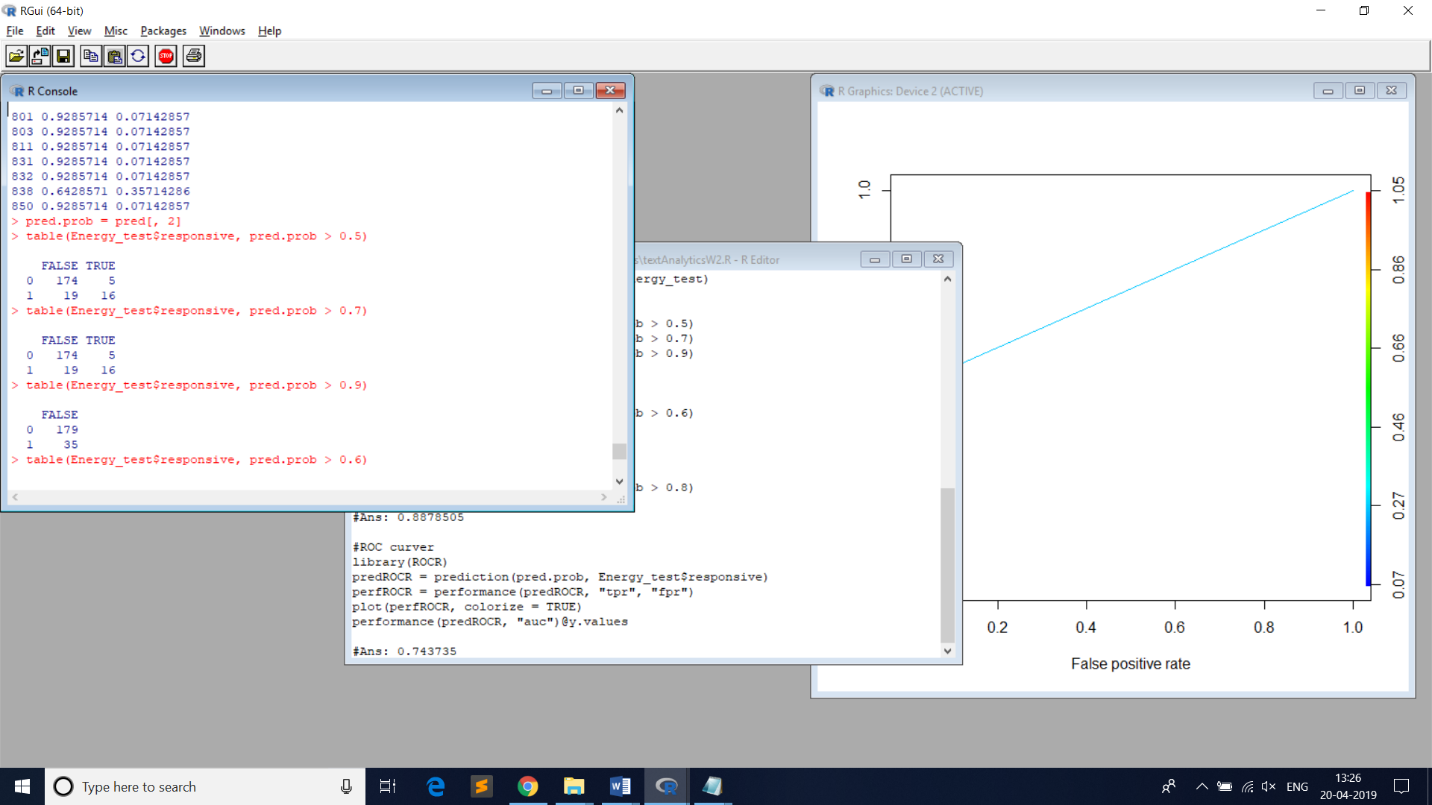
1) What is the number of observations in the dataset. What is the proportion of emails that are responsive in the dataset.

**Answer : 0.1625731**

2) Convert all alphabets into lowercase, remove punctuations, eliminate stop words and go for stemDocument and also remove sparse terms.

Build a CART model(classification) with seed value 1500 and train the model with 75% of the observations and plot the model.

3) Make predictions on the test set and mention the proportions of responses with value more than (i)0.5 (ii)0.7 (iii)0.9

**Answer : **

4) What is the accuracy of the model with predicted response of the test set (i)>0.6 (ii)0.8

**Answer : For both answer is 0.8878505**

5) Plot the ROC curve for the model and computer the value of AUC.

**Answer : 0.743735**

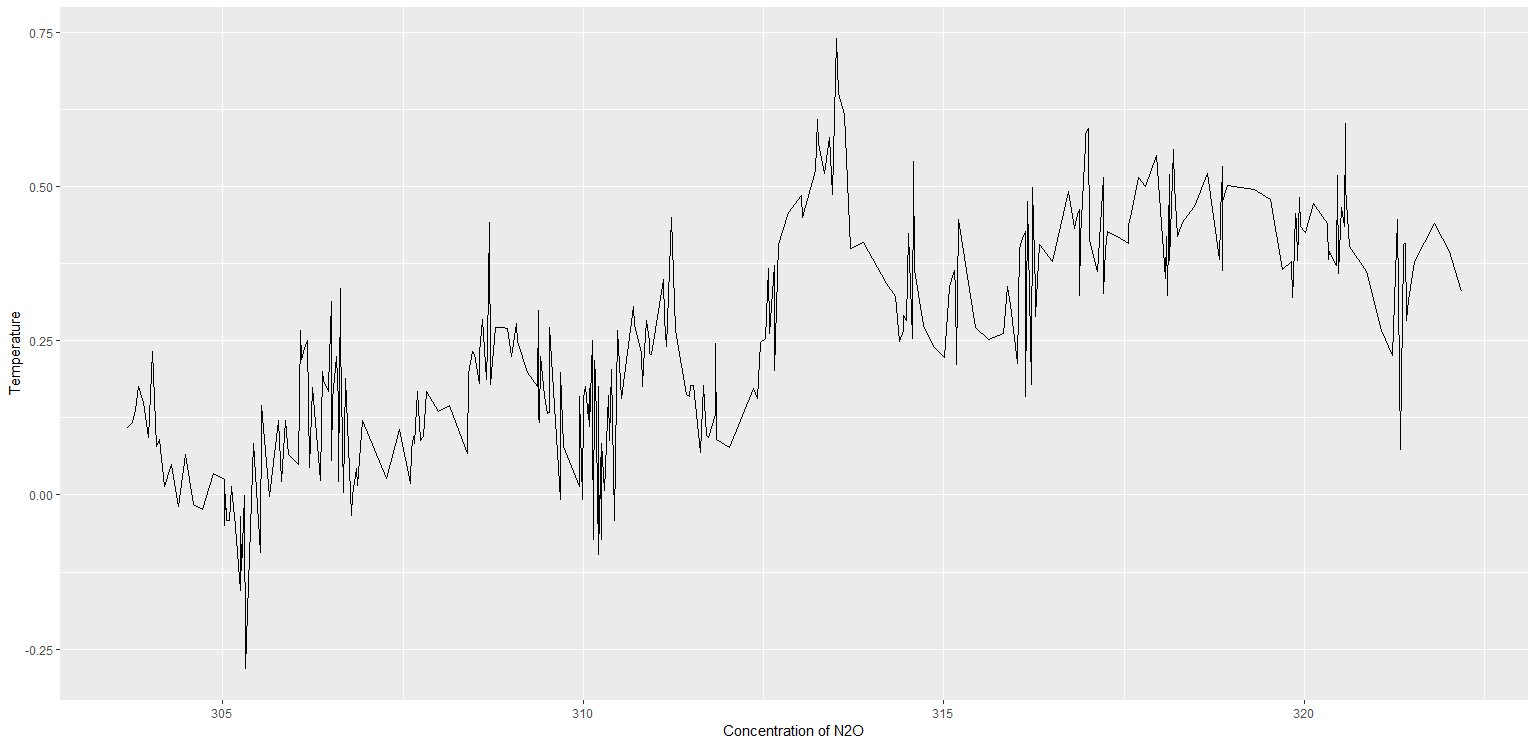
**Visualization : 5Marks**

**Q4. Answer the following questions from climate\_change.csv dataset**

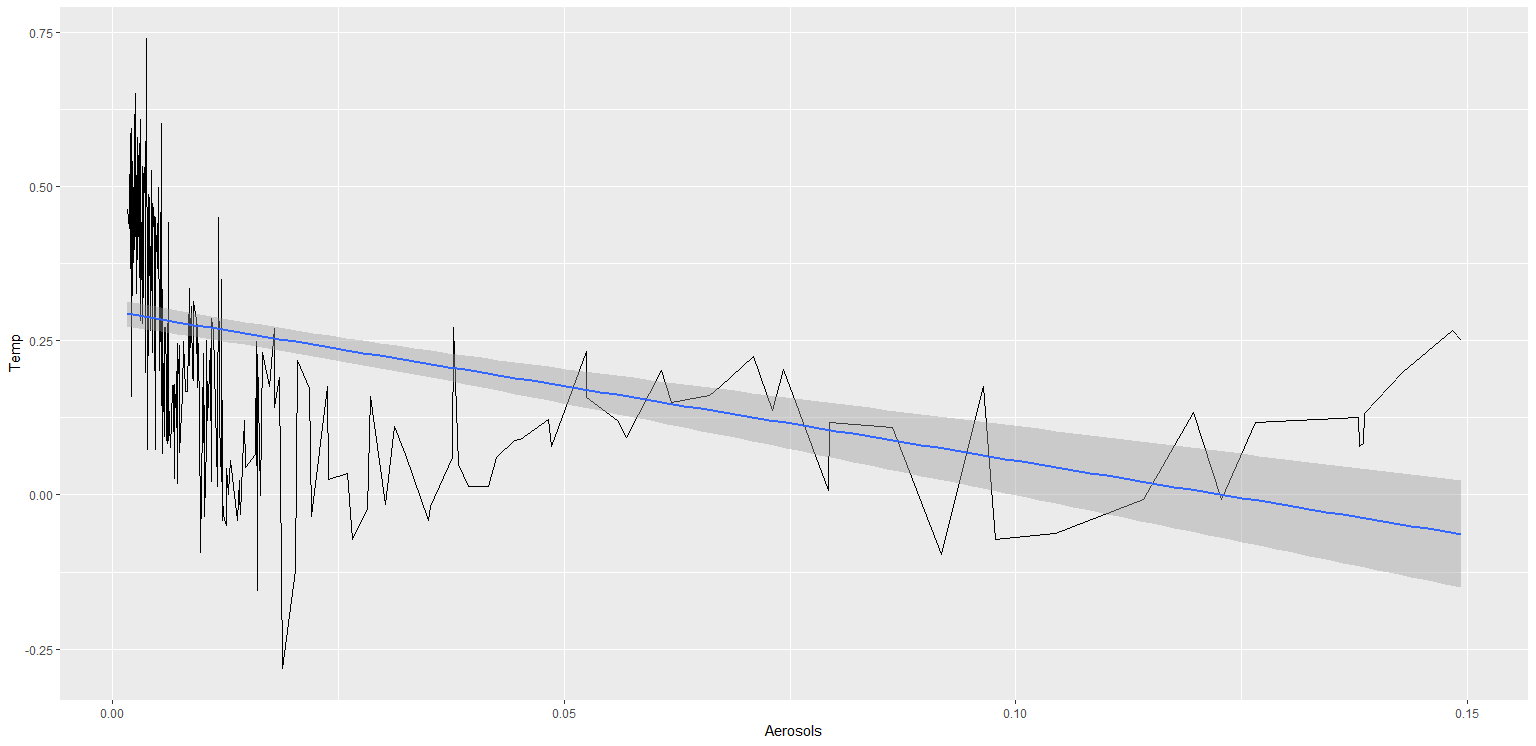
1) Load the data into R and find out the number of observations and the number of unique years.

**Answer : Observations = 308 and Unique Years = 26**

2) Plot the variables “N2O” and “Temp” on X and Y axes respectively and make it a line plot. Name the axes as “Concentration of N2O” and “Temperature” respectively.

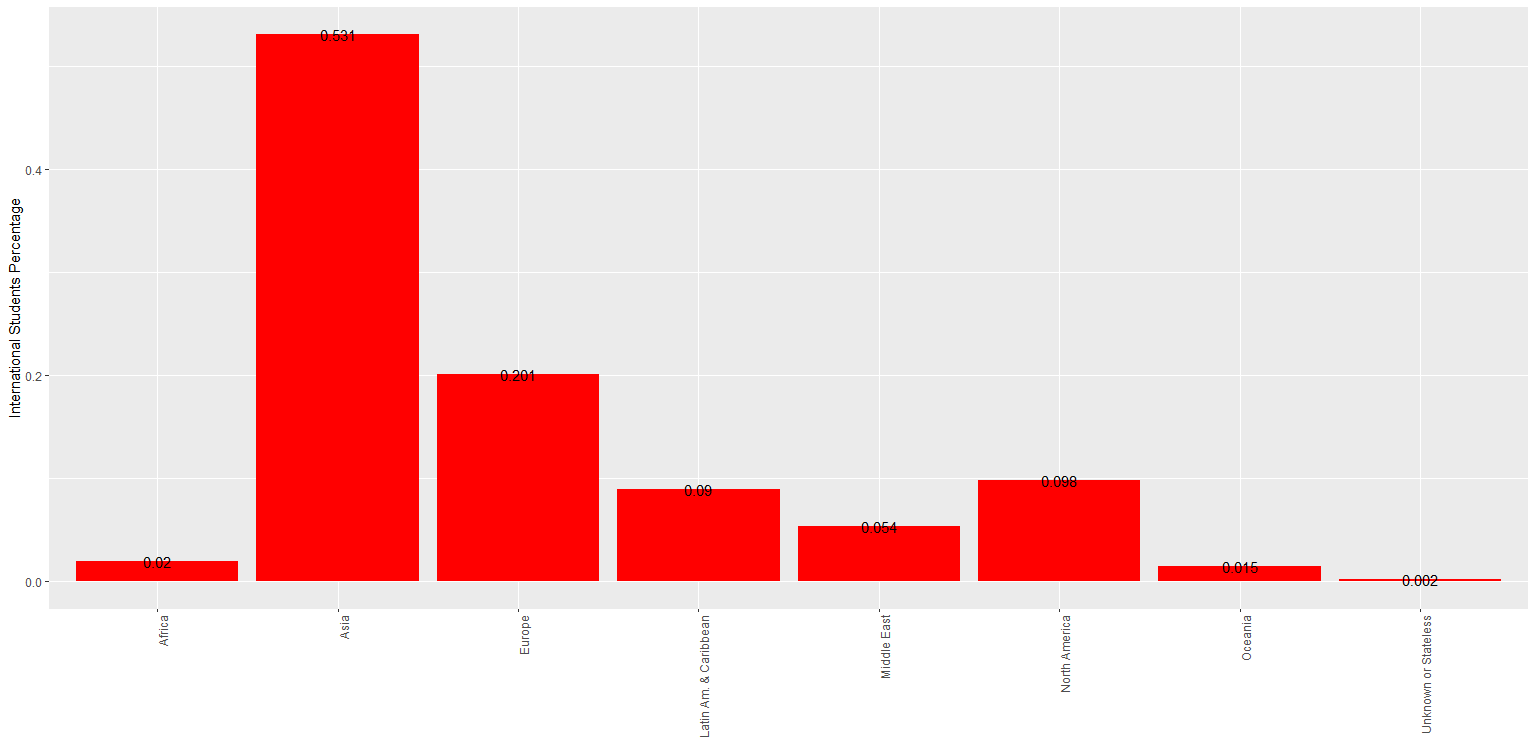
**Image :**

3) Build a linear regression model to predict “Temp” over “Aerosols” and plot the linear equation using ggplot2.(Go for a line graph) Also plot the regression line.

**Image:**

**Answer the following questions from “intl.csv”**

4) Plot the bar chart with region on X-axis and Percentage of International students on Y axis. Keep the stat as “identity” and fill in the bars with red color and add label “International Students Percentage” on y axis and the element text angle of 90 and horizontal justification of 1.

**Image:** 

5) Plot the pie chart with the regions as the labels.

**Image:** 