# A picture containing text Description automatically generated

**J.E. CAIRNES SCHOOL OF BUSINESS & ECONOMICS**

**INDIVIDUAL ASSIGNMENT COVER PAGE**

**Title: Individual Assignment 2**

**Module Name and Code: Applied Customer Analytics MS5108**

**Student Name and ID: Dilip Venkatesan Sankar 22225743**

In submitting this assignment, I am aware that it is my responsibility to adhere to the submission guidelines. Please tick (double click… or Yes/No) for the following:

|  |  |  |
| --- | --- | --- |
|  | **Yes** | **No** |
| I am aware of what the UOG plagiarism policy entails. |  |  |
| I have named the assignment file (MS Word docx , doc, or jar ) to contain my student ID and the module code and assignment number (e.g., 1187404\_MS220\_A1.docx or 1187404\_MS220\_A1.jar). |  |  |

**Declaration for this Assignment Submission*:***

*In submitting this work, I confirm that it is entirely my own. I acknowledge that I may be invited to interview if there is any concern in relation to the integrity, and I am aware that any breach will be subject to the University’s Procedures for dealing with plagiarism (*[*http://www.universityofgalway.ie/plagiarism*](http://www.universityofgalway.ie/plagiarism) *).*

## **Question1**

**Dataset:** Latest Buoy reports for M6

**Description:** The dataset contains 26 hours of observations collected by the buoy in Ireland which is vital for Ireland’s weather forecast. It contains information about the sea level pressure, wind direction, wind speed, dew point, wave height & period, sea temperature, and air temperature.

**Source dataset:** [Latest Buoy reports for M6 - Datasets - data.gov.ie](https://data.gov.ie/dataset/latest-buoy-reports-for-m6?package_type=dataset)

**Part A: Histogram R base graphics**

The histogram is plotted for the numerical data column called sea pressure and below is the output,

**Chart, histogram

Description automatically generated**

**Figure 1 Histogram for Sea level pressure**

**Part B: Histogram GGPLOT2**

**Chart, histogram

Description automatically generated**

**Figure 2 Histogram for Sea level pressure using GGPLOT2**

**Observation:** The x-axis contains the range of values divided into bins and the y-axis shows the frequency or proportion of data points. The histogram does not have a bell-shaped curve, so it’s not symmetrically distributed. From the output, we can see that in the last 26 hrs, the sea level pressure was within the range of 1029 and 1034 where is frequency is more at 1032 and 1033 compared to other points.

**Part C: Scatterplot R base graphics**

The scatterplot shows the graphical representation of two numerical data where the below plot contains x-axis denoting the independent variable (direction) and y-axis denoting the dependent variable (speed). The final scatterplot with car library adds box plots in the margins, a non-parametric regression smooth,

smoothed conditional spread, outlier identification, and a regression line.

Based on the output we can see there is a negative correlation and shows weak clustering and relationship,

Chart, line chart

Description automatically generatedChart, scatter chart

Description automatically generated

**Figure 3: Basis scatterplot**

**Figure 4: Enhanced scatterplot with regression line**

**Part D: Scatterplot using GGPLOT2**

**Chart, scatter chart

Description automatically generated**

**Figure 5: Scattplot using GGPLOT2 with Sea Temperature**

**Observation:** The goem\_point and geom\_smooth was used to add the fitted points and from the plots, we can see the points are not in a straight line which indicated non-linearity, and also we have a couple of outliers. So, when the speed is within a range of 15-20, the wind is traveling in a direction of 110 degrees and above, and for the last 26 hrs, the sea temperature is 11.1 degree Celsius in most cases.

## **Question2**

**Part A: Time series plot using R base graphics**

Chart, line chart

Description automatically generated

**Figure 6: Basic Time series plot**

The time series data set contains data about customer spending on different platforms. From the plot, we can see the spendings were almost linear for the first couple of months till July, with Cisco recording the maximum spending from January to July but in August there was a fall/dip in spending of all companies, with SAP at the lowest and again in December the spending went to peak with Amazon recording the highest spending compared to other companies.

**Part B: Adding functionalities to the Time series plot**

The geom\_line and geom\_point using ggplot are used to add more functionality to the plot in order to analyze the trend more professionally. Since in the previous graph, the y-axis was exponential, we converted the axis into a log scale for easy interpretation,

Chart, line chart

Description automatically generated

**Figure 7: Time series plot using GGPLOT2**

From the above, we can see spendings in August were very less for SAP and Amazon which is highlighted in “Violet” and “Pink” dashed lines respectively. But Amazon reached peak spending in December whereas SAP did not progress well. More Analysis can be performed using dygraphs as it is more interactive and enables filtering options. The below plot shows the spending for July-August month specifically,

Chart, line chart

Description automatically generated

**Figure 8: Advanced time series plot using DyGraph**

## **Question3**

**Preprocessing dataset**

The dataset contains 4757 tweets from different sources, and we need to perform preprocessing steps and clean up the dataset before performing any analysis,

1. The dataset is imported as CSV and converted into a corpus

2. Removing duplicates, NA values and converting all text to lowercase

3. Removing digits, hashtags, @ symbol, rt, special characters, blanks, whitespace

4. Removing stop words and punctuation.

**Part A: Frequency plot of top25 words**

Once the dataset is converted into a document term matrix, we use textstat\_frequency for filtering the top 25 words and below is the output,

Chart

Description automatically generated

**Figure 9: Frequency plot for Top 25 words used in Tweets**

Based on the output we can see the top 25 words frequently used by Mr.XYZ in tweets and from which **‘great’** has the highest count in usage.

**Part B: Word Cloud of 40 common words**

The word cloud is created and from the output, we can see the 40 most commonly used words. We can observe a variation in font size from large fonts to smaller fonts representing the usage based on size.

**Text

Description automatically generated with medium confidence**

**Figure 10: Wordcloud for 40 Common Words**

**Part C: Social media usage during breakfast time (6-10 AM)**

The below plot shows the usage of social media by Mr.XYZ during breakfast time 6-10 AM,

Chart, histogram

Description automatically generated

**Figure 11: Geom Bar for Social Media usage**

The plot shows that Mr.XYZ is not very active during early morning hours but the usage gradually increased after 9 AM and the usage reached a maximum after 10 AM, and then gradually decreased after 2 PM where the usage was fairly close from 3-11 pm.

**Part D: Usage per month**

The monthly usage of Mr.XYZ can we obtained by plotting a bar plot,

**Chart, bar chart, histogram

Description automatically generated**

**Figure 12: Bar graph for Month wise Social media usage**

We can observe that social media usage gradually increased after April and reached a maximum during June, July, and August but usage was very less from September to December and was lowest during January and December.

**Part E: Top-15 words for source=’iPhone’ and source=’Media Studio**

After filtering and removing stop words from the dataset, the unnest function is used to make one row for each word. And the GGPLOT is used to plot the top 15 words used in both the source and below are the outputs and we can infer that even though the usage of iPhone for tweeting is greater than Media Studio, the word **‘great’** is the most frequently used word in the tweets from both sources. Most of the tweets are from iPhone and very few from Media studio.

**Figure 13: Top 15 words used from Iphone and Media Studio**

Chart

Description automatically generated

Chart, funnel chart

Description automatically generated

There are also some common words like **‘great’, ‘president’, ’country’, and ‘people’**, and adding to that we can also see words like **‘Trump’, ‘Honor’, ‘Democrats’, and ‘America’** which show a pattern and the tweets might be regarding the election and it can be confirmed that it’s from the same person.

**Part F: Six words not used in the last six months of the data but were frequently used in the first six months**

We are creating two functions called **last\_six** which contains words tweeted in the last six months and **first\_six** which contains words tweeted in the first six months,

We use a full join to map each word with its matching component and use filtering based on NA to figure out words not used in the last six months of the data but were frequently used in the first six months,

Table

Description automatically generated

From the output, we can see the top 6 used in the first six months but not the last six months which are **‘obamacare’, ‘americafirst’, ‘icymi’, ’wh’, ‘notice’, and ,‘premiums’**.