Faculty of Information Technology

Semester 1, 2024

FIT5145: Foundations of Data Science

Assignment 2: Description

Due Date: Sunday, Week 8 (April 28, 2024), 11:55 PM

The aim of this assignment is to investigate and visualise data using **R**. It will test your ability to:

1. Read data files and extract related data from those files;
2. Clean and process data into the required formats;
3. Perform exploratory data analysis and visualisation;
4. Use basic tools for managing and processing data; and
5. Communicate your findings in your report.

Assessment Details:

- Assessment Type: Individual Assignment
- Total marks: 15%
- Due Date: Sunday, Week 8 (April 28, 2024), 11:55 PM. Please note that submissions will not be accepted after 5 May 2024 (i.e., 7 days after the due date).

Submission Details:

You will need to submit two separate files (**PDF report and RMD file**).

1. A **report in PDF** containing your ***(a) code, (b) answer, and (c) explanation*** used to answer each question.

   (a) code: Please directly convert the RMD file including your codes into the PDF file (Note: Please Knit RMD to HTML and print the HTML as pdf). If you want to use Word or other word processing software to format your submission, please **copy** your codes from the RMD file and **paste** into Word (Please do **NOT** take the screenshots of your code).

   (b) answer: Please make sure to include **screenshots/images of the code outputs and written answers** for each question.

   (c) explanation: Please explain how you answered each question (i.e.: explaining your codes or summarising your work for each question).

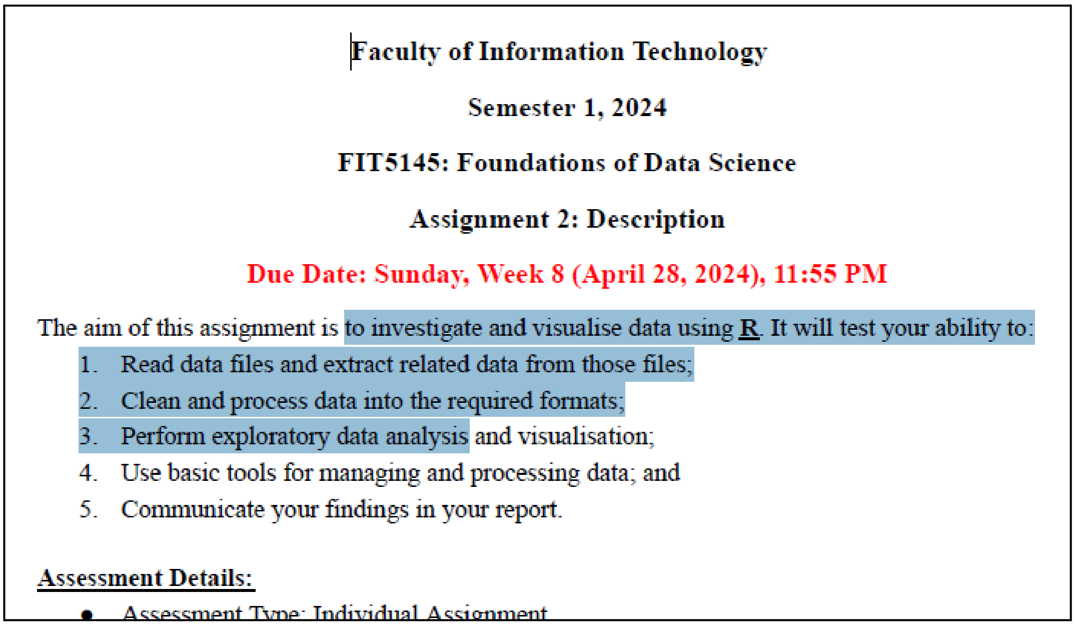
   Marks will be assigned to reports based on their correctness and clarity. For example, higher marks will be awarded to reports that include graphs with appropriately labelled axes and sufficient comments for the code.

2. The **RMarkdown** file: Please submit the RMarkdown file that contains your R codes. Your file should contain all the codes, and proper comments. If your work uses new libraries that

have not been introduced in classes, please include instructions on how to get these libraries installed.

Notes:

1. Whenever a question asks for a certain value, your code should produce the value. For example, when a question asks for the number of rows contained in a table, your code should print out the number of the rows. Extraction of the answer manually by eye-examination will not earn any marks.
2. Assignment should be submitted in two files (PDF report and RMD file):
   a. Failing to submit one of the two files will result in losing 20% of the total mark of this assignment.
   b. An RMD file that generates errors when running will not be considered.
3. **Please do NOT zip your submission files**. Zip file submission will have a penalty of 20% of the total mark of the assignment.
4. **Please make sure that you can select and highlight texts in your PDF,** as shown below then the turnitin score can be generated properly for your PDF file (we just need the Turnitin score for the PDF file, not the RMD file).



Faculty of Information Technology

Semester 1, 2024

FIT5145: Foundations of Data Science

Assignment 2: Description

Due Date: Sunday, Week 8 (April 28, 2024), 11:55 PM

The aim of this assignment is to investigate and visualise data using **R**. It will test your ability to:

1. Read data files and extract related data from those files;
2. Clean and process data into the required formats;
3. Perform exploratory data analysis and visualisation;
4. Use basic tools for managing and processing data; and
5. Communicate your findings in your report.

Assessment Details:

- Assessment Type: Individual Assignment

Assignment Task:

The dataset for this assignment is in the file *"ireland_news.csv"* and it can be accessed from the Assessments page on Moodle.

The data used in this task contains the news articles published by various news providers in Ireland from 1996 to 2021. It provides a long term birds eye view of the happenings in Europe. Therefore, analysing this dataset with news text is important for extracting insights such as trends, key events and other interesting points in relation to news, news topics and news providers.

Each row in the data file details information about a single news article, including the date of publication, its headline category (topic), the article's headline text, and the corresponding news provider. For instance, a sample row in the data is shown below, detailing the article's category as "news", publication date on 14/11/2009 by Irish Times, and headline text: "Power supply lost to Dublin airport".

pulish_date	**headline_category**	**headline_text**	**news_provider**
Saturday, 14th of November, 2009	news	Power supply lost to Dublin airport	Irish Times
...	...	...	...

In this assignment, you are supposed to perform data analysis to gain a better understanding of the published news articles of different headline categories and news providers in Ireland over different periods of time in the past years.

1. What are the earliest and latest articles from Irish Independent, irrespective of headline category? Please also sort the data according to the column *publish_date* in an ascending manner and display the last 5 records of the data.

2. How many unique *headline_category* values are there in the data file? Please consider variations (e.g.: capitalisation, potential inconsistencies) of the *headline_category* values, when counting them.

   How many news category articles contain either the keyword, "Ireland", "Irish", "US", or "USA" along with year digits from 2000 to 2024 in *headline_text*? For example, you need to search and count articles containing both "Ireland" and the year digits, or containing both "Irish" and the year digits, and so on.

3. Please display the top 10 headline categories with the largest number of articles published on Monday throughout the years. Then, draw a chart showing the total number of articles for the top 10 headline categories (as identified previously) for each year. What can you observe? Please discuss the chart and your findings.

4. Compute the total number of articles for each headline category and news provider. Then, use a single R function/command to display the statistical information, i.e., Min, Max, and Mean,

of the total number of articles (as computed previously) for each news provider. Note: You can use multiple functions/commands to get the desired pre-processed data table, but when you compute and display the statistical information, you need to use a single R function/command.

5. Please compute the total number of articles for each headline category, news provider, and the day of the week. Then, compute the average number of articles for each news provider and the day of the week, based on the total number of articles computed previously.

   After that, please display the day of the week with the highest average number of articles for each provider. The output data should be structured in the following format.

News provider	**The day of week** (with the highest average number of articles)	**The highest average number of articles**
Irish Examiner	Mon	843
Irish TImes	Wed	1023
...	...	...

6. Please select the 2019 and 2020 data and categorise the data based on the period it falls in. To do this, add a new column named "Period" and fill it with the suitable period names based on the *publish_date* value. The *publish_date* should be assigned the following period names:

publish_date	**Value to be inserted in Period column**
1/1/19 - 31/3/19	Period 1
1/4/19 - 30/6/19	Period 2
1/7/19 - 30/9/19	Period 3
1/10/19 - 31/12/19	Period 4
1/1/20 - 31/3/20	Period 5
1/4/20- 30/6/20	Period 6
1/7/20 - 30/9/20	Period 7
1/10/20 - 31/12/20	Period 8

   After that, write the code to compute the total number of articles by period and headline category for the top 10 headline categories, then generate a boxplot showing the total number of articles for each of the periods.

7. Please sample 1% of the data, conduct the text pre-processing for the values of the *headline_text* column in the sampled data, and display a portion (the first few columns and rows) of a document-term matrix generated. Then, draw a plot showing the top 10 most

frequent words where the x-axis represents the frequency of words and the y-axis represents the words themselves. Additionally, generate a word cloud.

8. Now using the original dataset, please draw any chart or create any summary table which hasn’t been investigated in the previous questions and discuss them. Marks will be given based on depth of investigation (What analysis and discussion have been done) and reasoning (How insights and conclusions have been drawn).