School of Computing, Engineering and Built Environment

Data Visualisation

Module Code: MMI226820

Coursework 2

Issue date: 26th March 2024

This coursework comprises 70% of the overall mark for the module.

Attention is drawn to the university regulations on plagiarism. Whilst discussion of the coursework between individual students is encouraged, the actual work has to be undertaken individually. Collusion may result in a zero mark being recorded for the coursework for all concerned and may result in further action being taken.

Exploratory Data Analysis using Data Visualisation

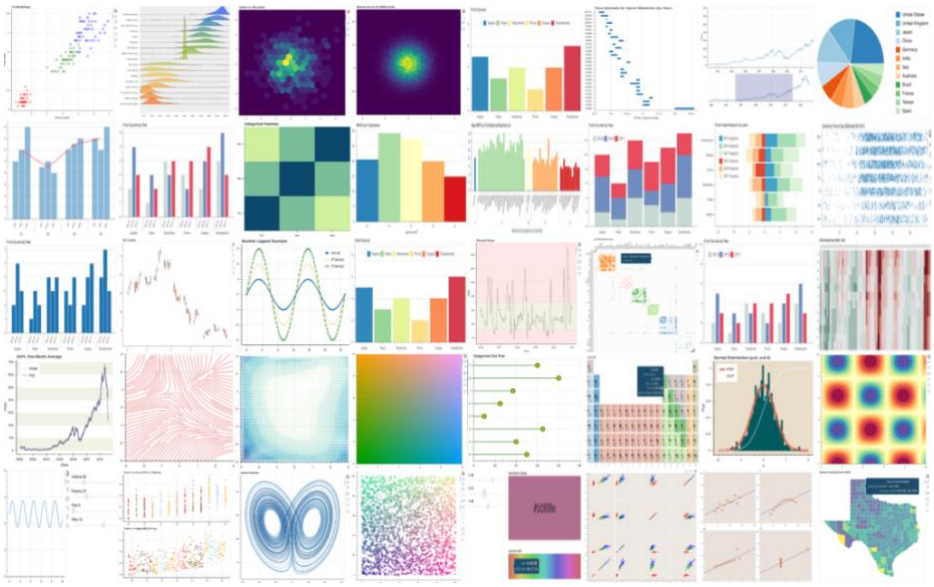

Figure 1. Examples of types of Data Visualisation charts in R.

1. Introduction

The goal of this coursework is for you to demonstrate proficiency in the techniques we have covered in this class (and beyond if you like) and apply them to a novel dataset in a meaningful way. In other words, you get to show off all the tools and techniques you learned by creating beautiful, truthful, narrative visualisations!

For this coursework, you will take a dataset, explore it, and tell a story about it using at least three different types of visualisations (or graphs).

In this coursework you will use R Studio to generate a report created using R Markdown.

2. Dataset

You need to choose **one** of the following**:**

i. **Retail Insights: A Comprehensive Sales Dataset.** This dataset has 5000 entries and 24 columns. It is a synthetic representation of a company's sales data. The information includes various details related to sales transactions, such as order details, customer information, product details, pricing, and shipping details.
More information on this dataset can be found here:
https://www.kaggle.com/datasets/rajneesh231/retail-insights-a-comprehensive-sales-dataset.

Data file:

Retail insights _sales data.csv

ii. **Apple Quality Analysis Dataset.** This dataset consists of various apple fruit qualities by including information about different aspects of the fruits. Information like fruit ID, size, weight, juiciness, maturity, acidity, crunchiness, sweetness, and quality are all included in the dataset. More information on this dataset can be found here: https://www.kaggle.com/datasets/tejpal123/apple-quality-analysis-dataset

Data file:

Apple _Quality data.csv

iii. **Restaurants Revenue Dataset.** This dataset is an extensive set of synthetic data on monthly sales for several imaginary eateries. This dataset that was produced for educational and demonstrative purposes. More information on this dataset can be found here: https://www.kaggle.com/datasets/mrsimple07/restaurants-revenue-prediction

Data file:

Restaurants revenue _ prediction data.csv

iv. **Top 5000 Albums - Spotify features.** This is an amazing dataset made by Michael Bryant for people trying to improve their EDA skills. More information on this dataset can be found here: Top 5000 Albums of All Time - Spotify features (kaggle.com)

Data file:

Top 5000 Albums _ Spotify features.csv

3. Instructions

Here's what you will need to do:

a. **Download** a dataset and explore it. Most of the datasets will have nice categorical variables that you can use for grouping and summarizing, and some will have time components too, so you can look at trends. Your past lab exercises will come in handy here.

b. Use R for **pre-processing** the data, **exploring** the data using data visualisation and basic statistics (descriptive statistics etc.), detailing and appropriately documenting any information and insights you can extract from the data.

   The goal is not to do an exhaustive data analysis i.e., do not calculate every statistic and procedure you have learned for every variable, but rather let me know that you are proficient at *asking meaningful questions* and *answering them with results of data*

visualisation and analysis, that you are *proficient in using R*, and that you are *proficient at interpreting and presenting the results*. Focus on methods that help you begin to answer your research questions. You do not have to apply every data visualisation we learned.

c. **Find** a story in the data. Explore that story and make sure it's true and insightful.

d. **Create** multiple graphs to tell the story. You can make as many graphs as you want, but you must use at least **three** different chart types (don't just make three scatterplots or three maps).

e. Write a report of your work using **R Markdown** (see below for details of this report).

4. Report

Write a report using **R Markdown** to introduce, frame, and describe your story and figures. This document should provide a discussion and demonstration of the steps you have undertaken to perform exploratory data analysis using data visualisation detailing the reasoning for undertaking each step. Please demonstrate the complete workflow and include all your code!

The report should include the following:

- Background information and summary of the dataset
- Research Questions (aim for 2 or 3 questions, but they may be related)
- Characteristics of the variables of interest
- Description of the Exploratory Data Analysis techniques used
- At least 3 different types of data visualisations that address the research questions
- Explanation/justification, description and code for each individual data visualisation
- Conclusion and Insights from the analysis, and how these might help answer the research questions.

Neatness, coherency, and clarity will count. All analyses must be done in RStudio, using R. There is no limit on what tools or packages you may use, as long as you predominantly use the packages we learned in class (tidyverse and ggplot2).

5. Final deliverables

The Coursework submission should be in the format of a **R Markdown document** and a generated **report**. These two files must be submitted as a single pdf document. Specifically, the R Markdown document should be appended to the report as an appendix.

Please follow the instructions below:

a. To transform your **R markdown** file into a report, click the **"Knit**" icon that appears above your file in the script editor. If knitting into a pdf document does not work,

please knit as a Word document instead (Figure 2). Afterwards you can export/save the word document as a pdf document.

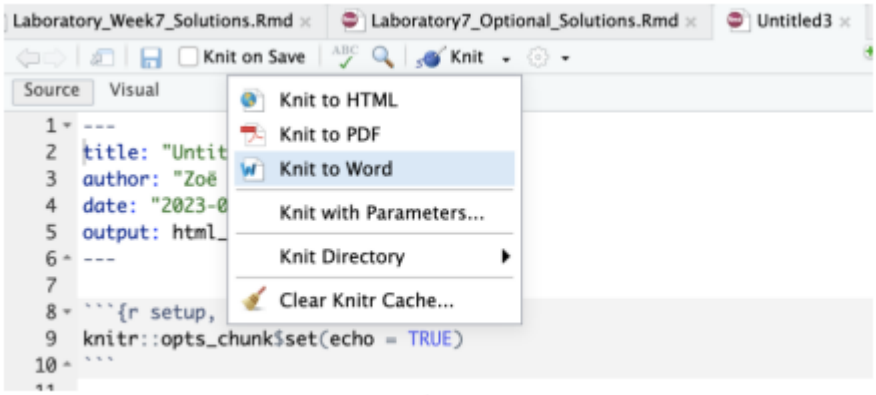

Figure 2. Knitting the R Markdown file into a Word document.

b. From the R markdown file (*.Rmd), create a pdf document (*.pdf) using one of these two methods:

1. Select File -> Print. This will give you the option to save the file as a pdf document.
2. Run the following code from the R Console:

```
knitr::stitch('filename.Rmd') # replace filename with your Rmd document
```

- Go to an online free pdf merger tool, for example https://smallpdf.com/merge-pdf.
- Select the report and the R Markdown document, both in pdf format. Next, merge the pdf's and download the resulting file. This is the file that you can then submit using Turnitin.

Coursework reports should be submitted to GCULearn via Turnitin no later than **Wednesday, 24th of April 2024 23.59**

Please note that only one submission can be made so please make sure you are happy with your work before you submit.

6. Marking criteria

Coursework 2 is worth 70% of the Data Visualisation module assessment. The Coursework will be awarded a mark out of 100 which will be weighted at 70% within the overall module mark. A separate mark is given for each element in the marking scheme. A rubric for Coursework 2 is provided in Table 1 (next page).

Table 1. Rubric Coursework 2.

Criteria	Insufficient	Substandard	Acceptable	Good	Excellent
General introduction and background of the dataset [/5%]	Not shown or not sufficient (eg. no Introduction provided, source of data not mentioned, dataset not described).	Limited introduction and background of the dataset (eg. no context provided, dataset mentioned but no further detail provided).	General introduction and background to data is provided with several omissions and/or explanations are unclear. The source of the data is provided.	Introduction and background for the data is mostly complete and contains clear, relevant information. Some information is missing or not cohesive.	Cohesive and detailed, well-written introduction and background to the data, with appropriate diagram(s), images, table(s) and/or appropriate references. Introduction places the purpose of the work in context.
Research Questions [/5%]	Not shown or not sufficient.	Research questions defined but not explained and/or questions are unclear, irrelevant or incorrect.	Research questions defined and relevant, but with limited explanation.	Research questions defined, relevant and explained well with some further clarity required.	Relevant and well-written Research questions; clearly defined with detailed explanation as to why chosen.
Characteristics of the variables of interest [/10%]	Not shown or not sufficient.	Variables are listed but several key characteristics are missing eg. description, type, distributions with visualisation and explanation, summary statistics, missing values and outliers).	Variables are listed with some description and characteristics of variables provided, though with some omissions/errors.	Adequate description of variables with sufficient detail on the characteristics of variables, with some further clarity required.	Clear description of variables with detailed description of characteristics, aided by clear and appropriate visualisations and summary statistics.
Data cleaning/wrangling process [/10%]	Not shown or not sufficient.	Evidence of minimal data cleaning/ wrangling but not further detailed or justified in the text and/or containing errors. Missing values not considered.	Data cleaning/wrangling process defined and detailed in the text with some omissions. Missing values are considered.	Data cleaning/wrangling process defined and detailed in the text, with some further clarity required.	Data cleaning/wrangling process clearly defined, detailed and justified in the text Clear justification of how any missing values were handled.
Choice of data visualisations and rationale [/20%]	No data visualisations or only one type of visualisation shown. No rationale provided.	Only 2 types of data visualisations shown, or choice of data visualisations mostly not appropriate to the data. No justification provided on why the visualisations were chosen.	Choice of data visualisations (at least 3 types) for the most part appropriate to the data, with limited or incorrect justification of the choices made.	Appropriate choice of data visualisations (3 types), with justification of data visualisations provided. Some further clarity/justification required.	Excellent choice of data visualisations (3 types), with clear and detailed justification of data visualisations provided. Originality regarding choice of data visualisations is demonstrated.
Presentation of data visualisations [/25%]	Not shown or not sufficient.	Visualisations contain multiple errors, and the visualisations are not properly formatted (eg. incorrect/illegible/no axis labelling etc.).	Adequate data visualisation with errors or formatting issues (colour use, axis labelling, figure captions etc.). Not much formatting beyond the basic requirements of a plot.	Good data visualisations and clear evidence of formatting, with some further improvements or clarity required.	Compelling and well-formatted data visualisations including plot title, axis labels, appropriate use of colour, annotation, etc. The visualisations tell a clear story. Creativity is demonstrated.
Conclusions [5%]	Not shown or not sufficient (eg. no Conclusion section is provided, no conclusions drawn).	Limited conclusions shown, or the conclusions do not follow from the visualisations.	General conclusion provided, with some omissions or not linked to research questions or not following from the data visualisations.	Detailed conclusion provided, following from the data visualisations and linking to the research questions, but with some further detail/clarity required.	Well-written and detailed conclusion provided. Interpretation and discussion of produced results, linked to research questions. Implications and/or recommendations for further analysis discussed. Novel/creative insights.
R Code [/10%]	Not shown or not sufficient (eg. ggplot2 not used, code does not work, and tidyverse coding style not followed).	Limited or no commenting of code and other errors in code and coding style.	Code works for the most part, tidyverse coding style is somewhat followed. Code is mostly commented.	Code works well and tidyverse coding style is mostly followed.	Clearly formatted and commented code, tidyverse coding style is followed (%>%, consistent commenting, ggplot layers on new lines etc.). Code is reproducible.
Presentation of information [/10%]	Not sufficient. Poor formatting, style and organisation of report with no evidence of formatting. Narrative is unclear due to many awkward sentences or spelling mistakes; or no evidence that report was created using R Markdown.	Some attempt at formatting the report is evident, but with several errors or omissions in the presentation of information (eg. the figure numbering, section headings etc.). Narrative is hard to follow, often due to grammar and spelling mistakes.	Information presented for most of the material, with some omissions; formatting mostly fine.	Information presented in all cases with some improvements in clarity required. Well-formatted report.	All information clearly presented. Well-formatted report.