COMP6037: Foundations of Data Analytics

Coursework 2: Semester 1, 2024-25

School of Engineering, Computing and Mathematics
Oxford Brookes University

This is individual coursework, so an independent submission is expected from each student.

This assignment is 60% of your final grade for the module.

Learning Outcomes Assessed

The learning outcomes assessed in this assignment are:

- Critically analyse data visualisation approaches with respect to human sensory modalities
- Create appropriate visualisations for temporal, dynamic, and high dimensionality data
- Devise methodologies for data interaction to facilitate exploratory data analysis

Introduction

In this coursework you will select a big data set and produce a visualisation to communicate and explore features in this data. It will be an interactive HTML file. You will write a report, an R script and submit these with your data and the visualisation.

You are free to choose any data set which is publicly accessible and you will have the opportunity to discuss your choice (of data set) in the consolidation weeks.

You must import the data into an R session to perform exploratory data analysis and develop the visualisation. You may access and transform the data in order to import it, however this is not assessed and is out of the scope of this module. It is therefore recommended that you find a source which holds the data as static csv files.

You should be able to represent the data you are using in a single dataframe (variable with class "data.frame") in R. As part of your submission you should export the dataframe to a csv file with filename data.csv using the write_csv function from readr package.

e.g. if the dataframe has the variable name df you could export the file with code:

```
# load the readr package

library(readr)

# write the dataframe to a csv file

write_csv(df, "data.csv")
```

Part 1: Context and EDA

Once you have chosen and imported your data set, you will need to perform exploratory data analysis (EDA) on the data. You should then write a short report which should:
- describe the source and context of the data (100 - 200 words, max 1 figure),
    - this can include an image that helps understand the data, you do not need to have created this, but it should be referenced.
    - this should provide links to the publicly accessible source of the data.
- provide some summary statistics of insight on the selected variables (columns) in the data. You do not need to provide summary statistics on every variable and may group by some categorical variables to report on others. (200 - 300 words, 1 - 2 tables, max 1 figure).
    - you should report on the size of the data (number of rows and columns, size of the exported csv data.csv)
    - you should report on the datatype (class) of each column in the data
    - you can report on missing data and if useful can include a viz_miss plot from visdat package.

Part 2: Design

- Discuss an insight (trend or behaviour of interest) in the data that you will visualize in for your final submission (300 - 500 words, max 3 figures). This does not have to be modelled or analysed — it is sufficient to describe what you would like the audience to comprehend from a visualisation. The report should include:

- the purpose of visualising this insight
- the prospective audience for the visualisation
- a description of your planned visualization — an interactive visualization (created with the plotly)
- you may include up to (3 static figures) to help describe the insight you would like to communicate and your design

Part 3: Final Visualisations

You should now create the final visualisation, which should be reproduceable in an R script that imports the data.csv file created in part 1. There should be a single well commented script that creates the visualisation with a call to the function htmlwidgets::saveWidget() to save the visualisation.

Visualisation

This visualisation should be a self contained HTML file with filename visualisation_2.html.

Visualisation Commentary

In your report you should include 100 - 200 words for the final visualisations describing decisions you made about the design and consequent code you wrote. This can include considerations of:

- Accessibility
- Filesize
- Audience Comprehension
- Constraints of the brief

Submission Details

The assignment submission must be uploaded to Moodle by **16 December 2024, 1:00 PM (Week 13).**

A template word document (COMP7037 Data Visualisation Assignment Report Template.docx) can be used to write your report, which should then be saved as a pdf for uploading.

Marking Scheme

Part 1: Context and EDA (20 marks)

- Section Title "Part 1: Context and EDA", appropriate formatting, figure and table captioning and keeping to the word counts.   (5 marks)
- Context and Source description   (5 marks)
- Summary statistics and data set description   (10 marks)

Part 2: Design (20 marks)

- Section Title "Part 2: Design", appropriate formatting, figure and table captioning and keeping to the word counts.   (5 marks)
- Description of purpose   (5 marks)
- Description of audience   (5 marks)
- Design plan of the visualisations   (5 marks)

Part 3: Final Visualisation – demo (20 marks)

- A submission of the files: (2 marks)
  - report.pdf
  - script.R
  - data.csv
  - visualisation_2.html

- Visual appeal and effectiveness of Visualisation   (10 marks)
- Commentary on Visualisation   (5 marks)
- The script in the file script.R that creates the final visualisation from the file data.csv.   (3 marks)

TOTAL 60 Marks (60% of the module total marks)

Referencing

You are required to cite the work of others used in your assignment and include a list of references (use the university recommended referencing style) and must avoid plagiarism and collusion.

More information can be found at https://www.brookes.ac.uk/library/library-services/information-skills/citing-references-in-your-work-and-plagiarism/.

Feedback

Assessment feedback will be provided after the exam committee.