Data Analysis Skills – Group Project 1

Objective: Create a group poster summarising an analysis of data of each group's choosing with two members of each group presenting the poster in a live online viva.

Deadlines:

- File submissions: **17:00 Friday 2nd July**
- Live online viva: **10:00-12:00 Monday 5th July and 08:20-11:00 Tuesday 6th July**

Contribution to the final grade: 25% (See Marking Scheme below)

Overview: Each group is tasked with sourcing and analyzing a data set and preparing a poster to communicate the results of the analysis. The poster will be presented to two members of staff by two members of each group in a live online viva (i.e. presentation of poster followed by answering some questions).

Group Poster

There are three stages to creating the poster:

Stage 1. Source an interesting data set

In this digital information age there is no shortage of publicly available data available for analysis. You are required to find an interesting data set that you will then analyse and report on as a group. You can decide what 'interesting' means, but here are a few requirements:

- The data must be relevant to one or two research questions of interest which can be explored using the methods covered in Data Analysis Skills up to and including Week 8 "Inference for Model Parameters and Model Selection".
- The data must be uploaded into R and submitted in "tidy" format for assessment as a .csv file (see "Submission Instructions" below).
- The data should **not** have more than 500 observations. You may choose to subset observations from a larger data set, either randomly or by certain (justifiable) criteria.
- The data must **not** be included in any MSc course material (including Intro to R Programming, Regression Models, Statistical Inference and Advanced Predictive Models) nor can it be part of an R package (e.g. moderndive, fivethirtyeight, etc.).
- These sites list a number of data sets that you may find interesting (although you are not limited to these sites):
    - TidyTuesday
    - NHS Scotland Open Data
    - Open access to Scotland's official statistics
    - UK Gov Data
    - Awesome public datasets
    - Kaggle datasets
    - PRISM Data Archive Project
    - Harvard Dataverse
    - Youth Risk Behavior Surveillance System (YRBSS)
    - Bikeshare data portal

Stage 2. Analyse an interesting data set

Having posed the research question(s) in Stage 1, proceed to analyse the data with a view to answering the research questions. The analysis could include:

- Illuminating visualizations of the data;
- Suitable numerical summaries displayed in an informative way;
- Appropriate linear models fitted to the data (and assumptions checked);
- Relevant confidence intervals for estimates and model parameters and model selection.

All of this analysis must be conducted using tidyverse functions in R and contained in an RMarkdown file (.Rmd) with appropriate comments throughout that explain what the code is doing. Comments justifying the choice of the data and stating the research questions must also be included in the .Rmd file. Please note that it is not required to produce a document from the .Rmd file.

No graphs or summaries or any other output used in the poster can be 'copied and pasted' from another source, but all of the analysis must be *reproducible* from the R Markdown file.

Stage 3. Design and produce a poster

Summarise and present the work in Stages 1 and 2 in a poster such that:

- The poster is one A4 Powerpoint page and can be landscape or portrait. (Obviously it would be enlarged in real life, so when creating it imagine it will be magnified/printed/viewed in A1 size).
- The 'target audience' of the poster are your fellow students on the MSc programme, i.e. you can assume some knowledge of statistical models and inference.
- The main text should be minimum 11pt font.
- The poster should include a meaningful title and the names of group members.
- The poster should also include at least one table or graph appropriately labeled.
- R code should **not** be included in the poster.
- Save the poster as a pdf with the file name **Group_##_poster.pdf**.

Submission Instructions

You must decide on one member of each group to be responsible for submitting:

- **one poster** for each group in the file **Group_*NUMBER*.pdf**,
- the file **Group_*NUMBER*.Rmd** containing the analysis (NB: you are **NOT** required to produce the poster or any other document using RMarkdown)**,**
- the file **Group_*NUMBER*.csv** containing the data explored in the analysis/poster which can be read using the **Group_*NUMBER*.Rmd** file to reproduce the analysis.

Each of these files must be submitted using the respective upload links in the "Week 6: Group Project 1" section on the Data Analysis Skills Moodle page.

The deadline for submission of the files is **17:00 Friday 2nd July**.

Notes on evaluation of individual contributions

In addition to submitting one .pdf poster (and its corresponding files) per group**, each person** must complete a *Group Project 1 Contribution Evaluation* on Moodle. This will give you the opportunity to evaluate how well you feel you and your fellow group members worked together. The evaluations by all group members may be used to assign different grades to individuals within the same group if there is evidence that individual members didn't contribute significantly.

The form will ask you to evaluate yourself and each of the group members on each of the following criteria:

- Collaboration
  - Listened to, valued, and supported the efforts and opinions of others.
  - Tried to keep people working well together.
- Preparedness
  - Had agreed work prepared to a sufficient standard for group meetings.
- Effort
  - Participated in and contributed meaningfully to group discussions.
  - Submitted high quality work.
  - Was interested and enthusiastic.
- Contribution
  - An active member of the group.
  - Took their fair share of the workload.
  - May consist of, but is not limited to:
    - contributing to finding and selecting a suitable data set
    - writing R code to answers questions of interest
    - interpreting the output obtained from the analysis
    - preparing and writing the final report in R Markdown

You will indicate whether you *Strongly Agree / Agree / Disagree / Strongly Disagree* that you and each of the group members demonstrated each of the qualities described above during the group project.

You will also be asked to:

- Comment on what you think your contribution to the group project was.
- Comment on how well you think your group worked together.
- Say if there is anything you would do differently in future group work tasks.

Note on Declaration of Originality

Together with the *Group Project Contribution Evaluation*, **each person** must make a *Declaration of Originality*. These forms will be included together in a Moodle form alongside the .pdf /.Rmd /.csv upload links. Although only one person should submit the .pdf/.Rmd/.csv files on behalf of the group, **each member of the group** must complete the Moodle *Declaration of Originality* form before **17:00 Friday 2nd July**.

Live Online Viva

- The viva will involve **2 members of each group** giving a live oral **overview** of the poster and a short description of how the analysis could be **extended**.
- Both group members must speak within **4 minutes** (i.e. approximately 2 minutes each). The overall time limit will be strictly enforced.
- Following the overview of the poster there will be 3-4 minutes of **follow-up questions** by the staff, with each of the two group members required to answer at least one question.
- The 2 members of staff will grade individually, with grades being averaged at the component level (see marking scheme below) before being added to give a final mark.
- The 2 group members who attend the viva **will not** deliver the recorded video presentation which will be required for the second group project in Weeks 9-11 (i.e. other group members will do this).

Live online vivas will take place **10:00-12:00 Monday 5th July** and **08:20-11:00 Tuesday 6th July** via MSTeams.

A scheduler will be available on Moodle which each group will use to select a time slot for their viva by **17:00 Thursday 1st July**. After this deadline unallocated slots will be allocated by members of staff.

At each selected/allocated time, a video meeting will be made to each group channel on MSTeams **but only the two groups members allocated to the viva** will join the meeting. These two group members must be logged into MSTeams before the viva is due to take place and have their poster .pdf file open and ready to show via 'share screen'. Each group member must have their audio and video switched on throughout the viva.

Marking Scheme

The marking scheme for this assessment is split into 3 sections: Analysis, Poster Design & Content and Poster Viva. This assessment will be marked out of 60.

Analysis [20 Marks]

The R code and comments in the .Rmd file will be assessed. Specifically:

- Justification of data and research questions (e.g. relevance and suitability) [6 marks]
- Relevant exploration of the data set using multiple exploratory techniques has been conducted and any possible patterns/anomalies have been identified. [4 marks]
- Appropriate statistical methods have been correctly applied. [6 marks]
- R code is tidy, reproducible and well commented [4 marks]

Poster Design & Content [25 Marks]

- Informative title, author information and referencing. [2 marks]
- Clarity of explanation and lack of typos. [2 marks]
- Visual layout and use of space. [2 marks]
- Appropriate amount of text and size [2 marks]
- Inclusion of appropriate plots/tables and appropriate sizing [2 marks]
- Appropriate captioning and structure of poster. [2 marks]
- Appropriate colour scheme. [2 marks]
- Introduction to the problem to be tackled [2 marks]
- Description of the data, modelling, analysis and results [5 marks]
- Statement of conclusions [2 marks]
- Logical flow of the material [2 marks]

Poster Viva [15 Marks]

- Summary of analysis [6 marks]
- Response to questions [5 marks]
- Extension/Further work [4 marks]