ETX2250/ETF5922 Data Visualisation and Analytics

Assignment 1: Visualisation – *The Arts*

Submission instructions

This assignment comprises 15% of the assessment in ETX2250 and ETF5922.

Your assignment submission will consist of a pdf document. The document must include both your code and graphical and other output as well as paragraph answers that provide description and discussion of the output. The presentation of graphs is important. In particular, headings and labels should be provided.

In your completed assignment, graphs should be easy to read, with appropriate use of headings, colour, formatting and text.

The pdf document must be submitted as a hard copy in class, and also must be uploaded to Moodle.

The pdf document should be titled <studentid>*_A1.pdf*

(Note: No spaces, No Names, No other characters, No extra characters)

The assignment is due before class on Friday 25 January 2019 (9.00 am), and a hard copy of the pdf document should be submitted in class, securely stapled in the top left corner and with a signed COVER SHEET on top.

If for some reason you are unable to submit the assignment personally, there will be an assignment box available on level 5 of Building H to place it there **before** the due time.

The cover sheet is available at URL:

https://www.monash.edu/__data/assets/word_doc/0004/903379/assignment-cover-sheet-fbe-1.doc you must supply the unit code and further details.

Upload your assignment to Moodle as follows:

- Go to the *Assignments* section
- Click on *Submission of Assignment 1: Visualisation*
- Click on *Add submission*
- Drag and drop the file to submit it
- *Save changes*

To confirm that your upload was successful, go to the *Assignments* section, and click on the *Assignment 1: Visualisation* link. The uploaded filename will be shown.

Retain your marked assignment until after the publication of the final results for this unit.

Assignment 1: Introduction

This assignment relates to the Visualisation component of the unit. We ask you to report on two different data sets.

Part A: Painters

The first is the data set *painters* relating to various qualities of the artwork of some famous classical painters.

Part B: Movies

The second part of the assignment is unrelated to the first. You will investigate certain aspects of a data set about movies.

Assignment 1 Part A (10 Points)

The data set painters.csv is available on the Moodle site. This data consists of scores assigned to 54 classical painters on the basis of the four attributes:

- Composition
- Drawing
- Colour
- Expression

Thus there are 216 scores altogether. The names of the artists are given as row label. In addition there is a categorical variable: School , taking values A, B, C, D, E, F, G, H (see the details in the library MASS and type ?painters in a cell).

a) Data Wrangling (1 Point)

Some of the graphics you will be asked to produce relating to the painters data set are more conveniently produced using long form data. So prepare by producing a version of this data in long form. Also provide the head and tail of the resulting data frame.

b) Produce each of the following graphs. (2 Points)

(i) To determine the distribution of scores over the possible range, provide a histogram of all 216 scores. Use a binwidth of 1 so that it is clear exactly how many of each score have been assigned. **(0.5 Points)**

(ii) Provide four histograms showing the distribution of scores in each of the four Attributes. **(0.5 Points)**

(iii) Provide boxplots of the scores, one for each of the four Attributes. **(0.5 Points)**

(iv) Provide a scatterplot matrix for the scores across all attributes (i.e. the scatterplot matrix includes scatterplots for each pair of attributes.) **(0.5 Points)**

(c) Summarise what you learn from the plots in (b) about the distributions of scores. (3 Points)¶

(d) Suppose we now would like to also distinguish scores between the different schools (A to H) (1.5 Points)

(i) Apply a facet wrap to the graph produced in (b) (iii) to obtain thirty-two boxplots, four for each School, corresponding to the four Attributes. **(0.5 Points)**

(ii) Provide a graph of the scores for each School, coloured by Attributes. (e.g. schools A to H along x axis, and in a line above a given school, points representing the scores for painters in that school, coloured by attribute.) **(0.5 Points)**

(iii) Provide a facet plot where facets correspond to schools, and for which in each facet, attributes are plotted on the x axis and scores on the y axis, with dots coloured by attribute and appropriate use of jittering. **(0.5 Points)**

(e) For the various plots you have provided in part (d), discuss the advantages and disadvantages of using them to represent the data about scores by schools. (1 Point)

(f) Choose your preferred plot from part (d). Choose three Schools to focus on, and discuss what your chosen plot indicates about the attributes of these schools. (1.5 Points)

Assignment 1 Part B (4 Points)

There is a package ggplot2movies containing the data set movies.

(a) Create a histogram for the variable length in the data set movies, limiting the range of the x axis so that the shape can be seen. Comment on the shape. (1 Points)

b) (1 Point)¶

We are going to focus on certain feature length movies. Use the subset commands on the data frame movies to create a data frame entitled movies.df satisfying the following:

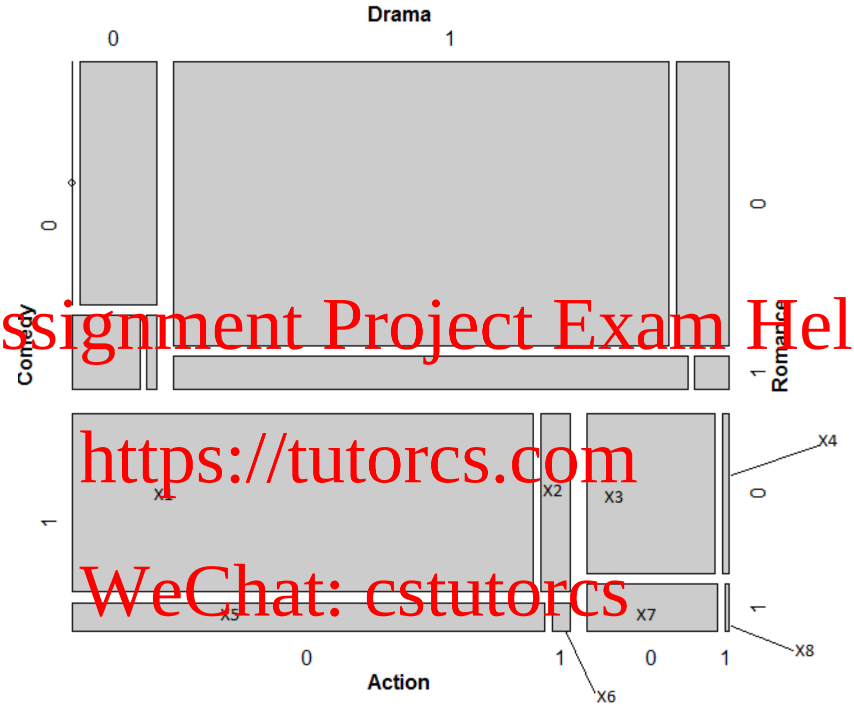
- It includes only the variables title, year, length, rating, votes, Action, Comedy, Drama, and Romance.
- It includes only those movies that are at least 60 minutes and at most 180 minutes long.

- It includes only those movies that are listed as being in at least one of the four categories Action, Comedy, Drama, and Romance

c) (1 Point)

(i) Construct a mosaic plot for the four binary variables Action, Comedy, Drama, and Romance. Choose the order of the variables so that the mosaic plot splits simply into a section that is comedy and a section that is not comedy. **(0.25 Points)**

(ii) One possible plot that you might obtain is:



For each segment of the mosaic plot labelled X1 through X8, state which of the classifications Comedy, Drama, Romance, and Action it belongs to. (Of course, in many cases, it will belong to more than one. You should state all categories it belongs to.) **(0.25 Points)**

(iii) Given that a movie is a Comedy, which is it most likely to also be: Action, Drama or Romance? Which is it least likely to also be? Explain your answers by referring to the mosaic plot above. **(0.5 Points)**

d) (1 Point)

Using the movies in the movies.df data frame, construct a scatter plot for rating vs votes. Discuss what conclusions can be drawn from this scatter plot about these two variables and their relationship.

Overall presentation: 1 point