

STATS 3DS3 Winter 2023

ASSIGNMENT 1

Submit through Avenue to Learn.

Due before 11pm on Friday, January 27th.

Your assignment must conform to the Assignment Standards listed below.

Assignments submitted up to 24 hours late will incur a 30% penalty.

Assignments submitted more than 24 hours late will receive a zero grade.

Answer all 3 questions. Not all questions carry equal marks.

All graphs must be labelled (including axes).

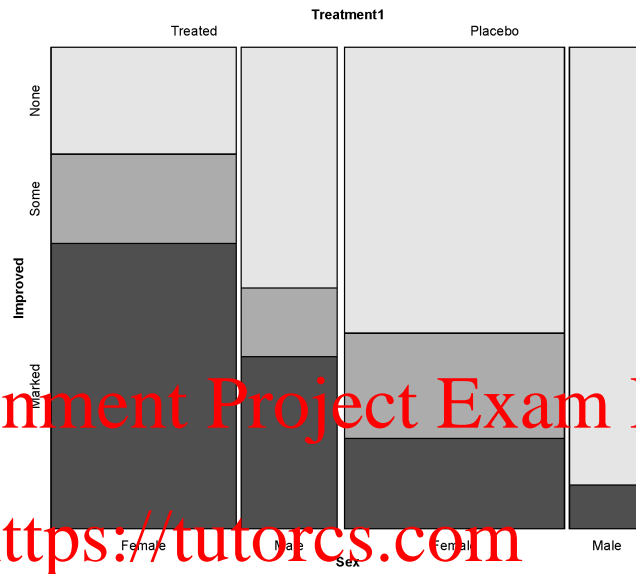
1. (10 MARKS) Go to the following website, “UCI Machine Learning Repository” and select a dataset that is suitable for visual data analysis. Do not use a dataset that has been used in class or in your lab.

- (a) Report the name of your dataset, the URL that you downloaded it from, and the date and time of downloading.
- (b) Produce a table of summary statistics for your dataset including variable names, number of observations, etc.
- (c) Using your dataset, produce 3 graphs from the list below:
 - i. a simple grid, a.k.a. “bar chart”
 - ii. a pairs plot
 - iii. a scatter plot
 - iv. a parallel co-ordinates plot
 - v. a box plot
 - vi. a violin plot
- (d) Create a “summary panel” (see Lecture 2 slide 8), displaying 2 of your graphs in one image.

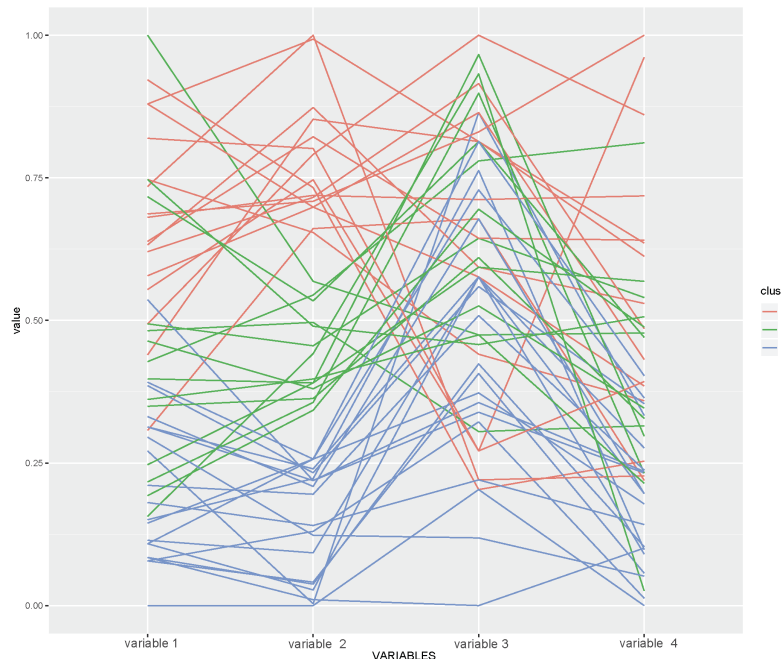
2. (4 MARKS)

- (a) Consider the Double Decker plot below; it displays 3 different levels of improvement (None, Some or Marked) that a patient can experience after receiving 1 of 2 medical Treatments (Placebo or Treated). Note: “improvement” is labeled as “Improved” on the graph.

- i. Which **sex** in the Treated group showed the best level of improvement? (i.e. which **sex** benefited most from the treatment?)
- ii. Are there more **males**, or more **females**, in the study group?
- iii. For male patients in the Treated group, what was the least reported level of improvement?



- (b) Consider the Parallel Co-ordinates plot below; it displays 3 different types of response, with names clus 1, clus 2 and clus 3.
- i. Name the best predictor variable for separating out the responses.



3. **(6 MARKS)** Find a text document that you think would be interesting for Word Cloud analysis;

- (a) Reference where you obtained the original document and state the word count.
- (b) Prepare your document e.g. remove stop words, etc.
- (c) Produce a Word Cloud.
- (d) Write a summary paragraph about your Word Cloud.

Note: copy and paste the uncleaned text used in your Word Cloud underneath your answers at the bottom of the document.

Assignment Standards

- L^AT_EX is strongly recommended but not strictly required. The use of Markdown in R studio is also recommended.
 - Submit your assignment as one **.pdf document**. Answer all the questions and then copy and paste the uncleaned text used in your Word Cloud, underneath your answers at the bottom of the document. **All R code should be included and organized either at the end of the assignment or inline (if using R Markdown).**
 - Eleven-point font (times or similar) must be used with 1.5 line spacing and margins of at least 1 inch all around.
 - Do not include a title page. The title and your name should be printed at the top of the first page.
 - Various tools, including publicly available internet tools, may be used by the instructor to check the originality of submitted work.
-