Assignment 03

DSA2101 AY 22/23 Sem I

Material Covered

The main learning outcomes for this assignment are:

1. To be able to ask questions/formulate hypothesis about a data set, and then to be able to make a plot that is aimed at answering said question.
2. To be able to summarise the insights from a chart/plot.
3. To demonstrate proficiency in ggplot2
4. Practice with creating documents using Rmarkdown.

Tasks

The github repository Tidy Tuesday contains a collection of datasets that you can use for practicing with data exploration. A new dataset has been contributed each week since April 2018.

If you scroll down from the page linked above, you will find a table listing each dataset in each year, along with a reference article. Your task is to **pick one of the datasets**, and **create two graphs from it** using ggplot.

Put your code into a Rmd document, along with the following text sections:

1. Introduction to the dataset
2. Plot 1: discuss the insights you glean from it, and your design choices. For instance, you may highlight why you settled on this geom, or this colour palette, or this particular binning of a continuous variable. You could also discuss who this plot would be appropriate for (e.g. a layman or a technical audience).
3. Plot 2: same as for Plot 1.

Each plot should depict at least 3 variables, and the two plots should differ in at least two of the variables they present.

Please try to keep your knitted pdf document to 2 pages, but this is just a guide. There is no penalty if it is longer. If you have problems setting up the pdf generation on your computer, you can simply submit the generated html.

It is possible to find contributed code for each dataset on the internet. In fact, I encourage you to look through what has been done. However, it is important to then sit down and write code on your own; without that it is difficult to improve and to cultivate the skill of data exploration.

Notes.

1. This assignment will be graded out of 10 marks. It contributes 10 percent of your final grade.
2. You must submit the following items to Canvas before the deadline Nov 11 2022, 2359hrs:
    1. the dataset you used.
    2. the Rmd file you wrote and that should knit.
    3. the knitted pdf or html version of the Rmd file.
3. You will be graded on:

- Individual effort. **This is an individual assignment**, so we will use the Rmd files to check for similarities.
- Demonstrating awareness of what constitutes a good graph. Try to incorporate as much of the principles from topic 04 as possible.
- Content. Please make the effort to include meaningful insights from the plots you have made.
- Proper formatting of your knitted document. Rmd documents can utilise several templates, e.g. tufte. Do explore and experiment!

4. The primary objective of this assignment is to give you an opportunity to explore a dataset. We will not be penalising for inefficient code. In fact, you can hide all the code chunks from the final pdf/html.
5. You can use any packages you like, but do keep in mind that this is a data exploration exercise, and not a prediction task. Do try to use the methods and ideas we cover in this class.