DAB501 - Project #1 - Data Visualization

The audience for this assignment is another budding Data Analyst/Data Scientist, like yourself. Your visualizations will help them understand the data set in question without them having to do the work themselves. They may be required to reproduce or extend your work later, so you need to provide them with not only the final visualizations but also the code (plus any comments/explanations of what you did) that will allow them to do so. This person could very well be you. For example, the company you are working for has acquired a new data set and your boss asks you to take a look at it to see if it will be useful for the business. You've started to dig into the data, made a few plots, when you are called away to work on a more pressing problem. Three months later you need to pick up where you left off! Good thing you have everything documented so clearly.

For this assignment we are not interested in a complete exploration of a given data set (that's what the final project is about). Your task is to clearly visualize subsets of the data to help the reader better understand the nature of the variables presented.

As noted by Selva Prabhakaran, an effective visualization has the following characteristics:
- It conveys the right information without distorting facts;
- It is simple but elegant. It should not force you to think much in order to understand it;
- The physical appearance of the visualization supports the communication of information rather than hinders it;
- It is not overloaded with information.

What you need to do
- Create 8 plots using the **ggplot2** package for R:
  - two plots displaying the distribution of a single continuous variable
  - two plots displaying information about a single categorical variable
  - one plot displaying information about both a continuous variable and a categorical variable
  - two plots should display information that shows a relationship between two variables
  - one plot should use faceting and display information about 4 variables
- Make sure each plot has (**all done within ggplot2**):
  - A title
  - Properly labeled axes
  - Properly labeled legend (if appropriate)
  - A caption
- Make sure that each plot (with title, labels, and caption) is a stand-alone entity. This means that someone viewing your plot does not need any other information to properly understand the message you are trying to convey to them.

- Briefly note the contributions of each team member.

What to hand in

- An **HTML** document (output from an R Notebook) that has a structure that is logical and easy to follow and contains:
    - A statement noting that the project represents your own work and that you have adhered to St. Clair College's **Academic Integrity policies** in completing this project
    - the versions of R (use *R.version.string* in console) and Rstudio (use Help/About Rstudio from top menu) that you used
    - a list of R packages used and their versions (use *packageVersion()*)
    - Documentation of the data sets:
        - attribution of the owner/creator of the data
        - links to the data
        - summary information about the data (you may have to edit the original documentation)
    - Documentation of any changes made to the original data sets, if applicable
    - All R code necessary to process the data sets and produce all the visualizations; your R code should be surrounded by well-written explanatory text
    - 8 plots that satisfy the requirements noted above
    - References to any external sources used and how they were used
        - note any books, blogs, websites, etc that you referred to in completing this project
    - Your answers to the following questions:
        - In what ways do you think data visualization is important to understanding a data set?
        - In what ways do you think data visualization is important to communicating important aspects of a data set?
        - What role does your integrity as an analyst play when creating a data visualization for communicating results to others?
        - How many variables do you think you can successfully represent in a visualization? What happens when you exceed this number?

Resources

Resources that you may find helpful:
- [ggplot2 cheat sheet](#)
- [R Graphics Cookbook](#)
- [Top 50 ggplot2 Visualizations](#)
    - for inspiration