The purpose of this assignment is to serve as a practice run for your individual projects. For this assignment:

Find a public data set in a field of your interest. The data set must have:

* at least 50 observations and 3 complete (no NAs) numeric variables.
* two categorical variables (or two additional numeric variables that you can convert to categories).
* in total, at least 50 observations with 5 filled data points each for a total of 250 non-missing "cells."

Of course, you can have more than 50 observations and 5 variables, and it is highly recommended that you find a dataset of 10+ variables and 100+ observations if you were to use this same dataset for the individual project.

Options for working with messy data

* You can find a data set that has more than 50 observations and delete rows with missing variables, as long as you end of with at least 50 observations with complete data.
* You can find data set a lot of variables and ignore variables with a lot of NAs, as long as you end with 5 variables (3 numeric, 2 categorical) for 100 observations.
* You cannot download a data set that has been created for statistics education purposes (imagine a scenario where you're responsible for your work data processing and cleaning).
* You can compile a data set of your own by pulling information from the internet/books, but please document exactly how you compiled your data set.

Instructions

In a Word Document:

1. **In your own words**, describe the data and how you obtained the data.
2. How did you address cleaning the data? What did you do for missing data (if any)? What rows and columns did you delete? Provide the R code, if any. If you did the cleaning in Excel, describe your step-by-step approach.
3. Use summary() to generate summary statistics of your three numeric variables (e.g. var1, var2, var3) by one of your categorical variables (e.g. summary statistics for "region 1" versus "region 2" versus "region 3"). Use boxplot() to visually show the summaries.
   1. Provide the R code
   2. Followed by graphs
   3. Followed by a brief description of what you see (in complete sentences).
4. Generate a histogram with an overlaid density plot of your three numeric variables by one of your categorical variables.
   1. Provide the R code
   2. Followed by graphs
   3. Followed by a brief description of what you see (in complete sentences). You can compare the histograms/density plots of the categories.
5. Create a bar plot showing the count of one of your categorical variables.
   1. Provide the R code
   2. Followed by graphs
   3. Followed by a brief description of any patterns that you might see.
6. Create a bar plot showing the by-group (category) average of one of your numeric variables:
   1. Provide the R code
   2. Followed by graphs
   3. Followed by a brief description of any patterns that you might see.
7. Create an interesting pie chart.
   1. Provide the R code
   2. Followed by graphs
   3. Followed by a brief description of any patterns that you might see.

**Upload the word document with your answers and your cleaned dataset.**

Some possible websites for datasets

* [Centers for Disease Control (Links to an external site.)](https://data.cdc.gov/browse)
* [UN Data (Links to an external site.)](http://data.un.org/Explorer.aspx)
* [Dataverse (Links to an external site.)](https://dataverse.org/)
* [The National Center for Education Statistics (Links to an external site.)](https://nces.ed.gov/)
* [The National Center for Health Statistics](https://www.cdc.gov/nchs/index.htm)