As you have probably heard us say many times already, data analytics can only be successfully learned by hands on experience — we have to learn by doing. Thus, we would like to encourage you to get a dataset that you are interested in and start exploring it using all of the tools we have been talking about in class. We hope you will pick data you are inherently interested in so that you have fun playing with and understanding the data. We have found that you are more likely to remember and learn these tools when they help you do something meaningful that you are interested in.

**Instructions**

1. Find a dataset that you would like to understand, experiment on, and work with that contains the following:
   * at least two numeric features/columns,
   * at least two categorical columns,
   * (not required) one date/ time column, and,
   * a minimum of five hundred observations.
2. The dataset may come from a public source or your workplace, but make sure it is something that you are inherently interested in working on. If the dataset comes from your workplace, make sure no identifiable, restricted, or sensitive information is included in the deliverables. That is, we do not ask you to share the data, but we do ask you to share views of the data in your submission file. It is your responsibility to make sure you do not show us anything that is problematic or sensitive. A few sources for finding a data set of your interest are provided at the bottom of this page.
3. Read in the data.
4. Make sure all columns are set to correct data types. That is, for example, if a column contains dates, it is set as "date" type, and if a column contains numbers, it is set as "numeric" type.
5. Handle missing values appropriately by either deleting them or imputing them (replacing them in some way that makes sense).
6. Do any other ETL and clean-up tasks as desired.
7. Show a printout of a view from the data (e.g., use the `head()` function in RStudio or a screen shot of table from Power BI to show a few rows).
8. Print the data types of each column (e.g., use the `str()` function in RSudio or a screen shot of Power BI to show the data type).
9. Show univariate distributions for any two numeric variables. You may use boxplots or histograms.
10. Show distribution for the categories for at least two categorical variables. You may use bar plots.
11. Show the relationship between two numerical variables. You may use a scatter plot for this purpose.
12. Do at least one other exploratory analysis of your choosing (again, we hope this is data that you are interested in, so you might have many additional analyses here).
13. Write a summary of the insights you found from the visualizations. The suggested length is between 100 and 350 words.

**Submission\***

1. Data description file in a .txt .rtf, .doc(x), or .pdf format. Describe the data, where it came from, what it is used for, and give a brief description of each column.
2. You may do this exercise in R or Power BI.
   1. If using R, create an RMD (all code, charts, and output) file and knit it to an .html file. Then compress the .html file as a .zip or .rar file for submission.
   2. If using Power BI, copy and paste all output and/or charts in a pdf file and submit the pdf file for submission.

**A few suggested data sources**

* https://www.pewresearch.org/download-datasets/
* https://data.cityofchicago.org/
* https://opendata.cityofnewyork.us/
* https://datasf.org/opendata/
* https://www.kaggle.com/datasets