# PPOL 563 - Problem Set 1

Alex Lundry

2022-09-28

## Problem Set 1 - PPOL 563

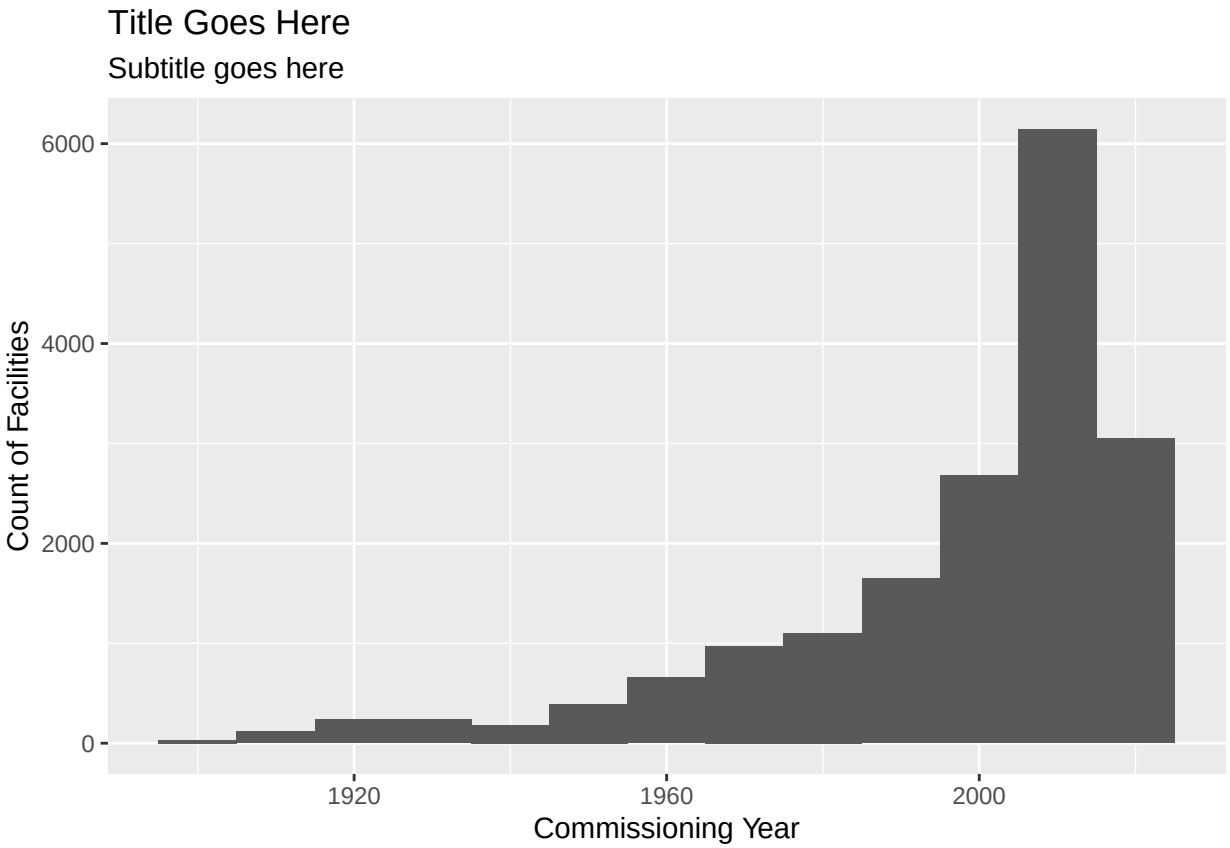
The first set of questions require you to work with the **Global Power Plant Database** contained in this repository. You can also find associated documentation for the dataset, including a data dictionary, a README file and release notes. The data and all associated materials can be found here.

```
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE)

library(tidyverse)
library(countrycode)
library(ggrepel)
library(scales)

# Data originally downloaded from:
# https://datasets.wri.org/dataset/globalpowerplantdatabase
d1 <- read_csv("global_power_plant_database_v_1_3/global_power_plant_database.csv")

# Pre-add continent for problem set
d1$continent <- countrycode(sourcevar = d1[, "country"] %>% pull(),
                            # pull out the country column from orig data as a vector
                            origin = 'iso3c', # naming convention of the orig data
                            destination = "continent") # name of new var to create

# Add continent for missing data
# Add a renewable energy flag
d1 <- d1 %>%
   mutate(continent = case_when(country_long == "Antarctica" ~ "Antarctica",
                                country_long == "Kosovo" ~ "Europe",
                                TRUE ~ continent),
          renewable = ifelse(primary_fuel %in% c("Solar", "Hydro", "Wind",
                                                 "Biomass", "Geothermal", "Wave and Tidal"),
                             "Renewable Energy", "Non-Renewable Energy"))
```

**Question 1**: This database is not a complete representation of all power plants. In fact, it only covers about 30% of all Solar energy produced worldwide. Please write a brief paragraph you would include in an email to colleagues describing the scope and source of the limitations of this data.

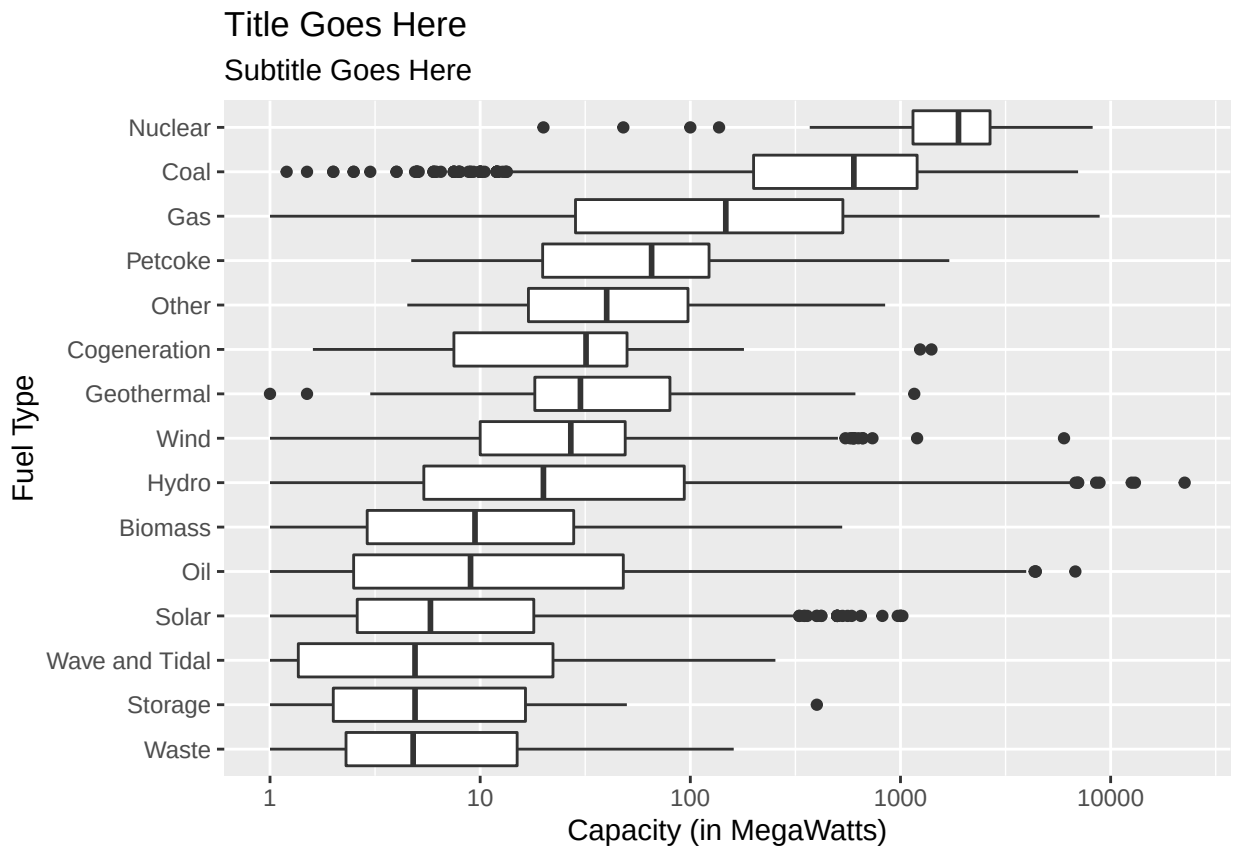
**Question 2**: Reproduce the histogram plot below, in which we are visualizing `commissioning_year` with a bin width of 10. Please insert a meaningful title and subtitle.

---



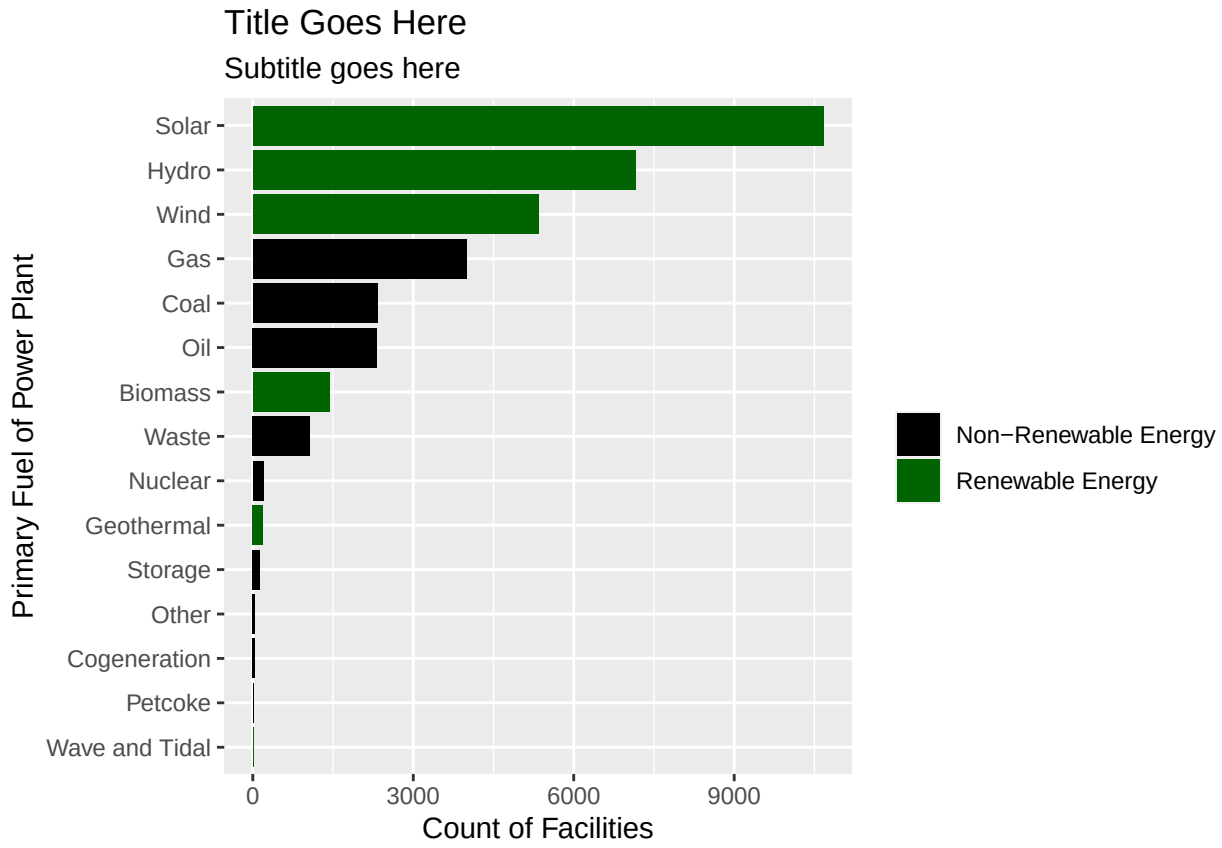
**Question 3**: Reproduce the boxplot below, in which we are visualizing `capacity_mw`, grouped by `primary_fuel` (which is sorted by the median of `capacity_mw`). Take note of the logged X axis, and please insert a meaningful title and subtitle.

---



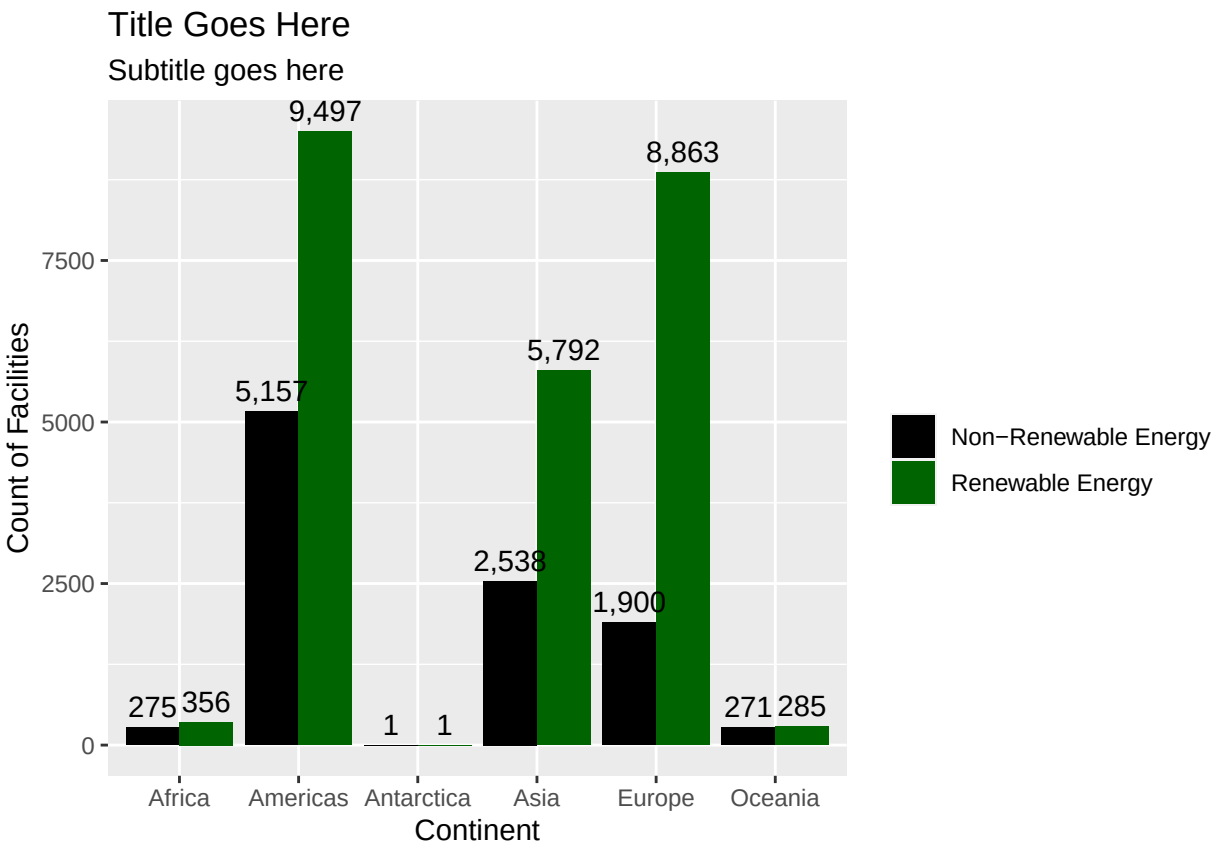
**Question 4**: Reproduce the barchart below, visualizing `primary_fuel` and colored by `renewable`. Please insert a meaningful title and subtitle.

---



**Question 5**: Reproduce the grouped barchart below, visualizing counts of the number of facilities in each `continent`, colored by the `renewable` variable. Notice the location of the text labels on each bar - this is tricky to do, so you may have to do some googling. Inserting your own meaningful title and subtitle.

---

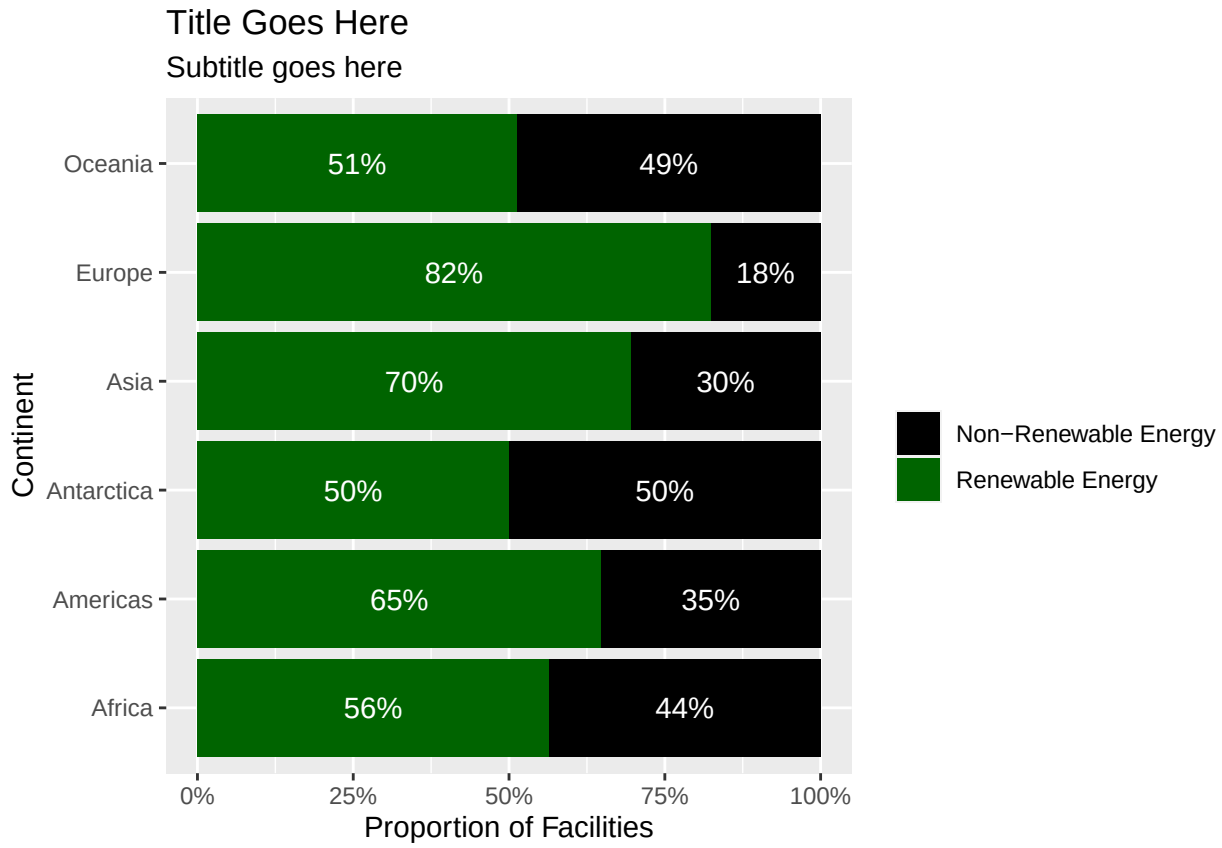


**Question 6**: Reproduce the 100% stacked barchart below. This is a visualiation of a count of facilities by `continent` grouped by the `renewable` variable, and then computed as proportions. Take note of the text labels, which are:

- rounded to a whole number
- text is white
- % suffix
- in the middle of their respective colored bar

Please insert your own meaningful title and subtitle.

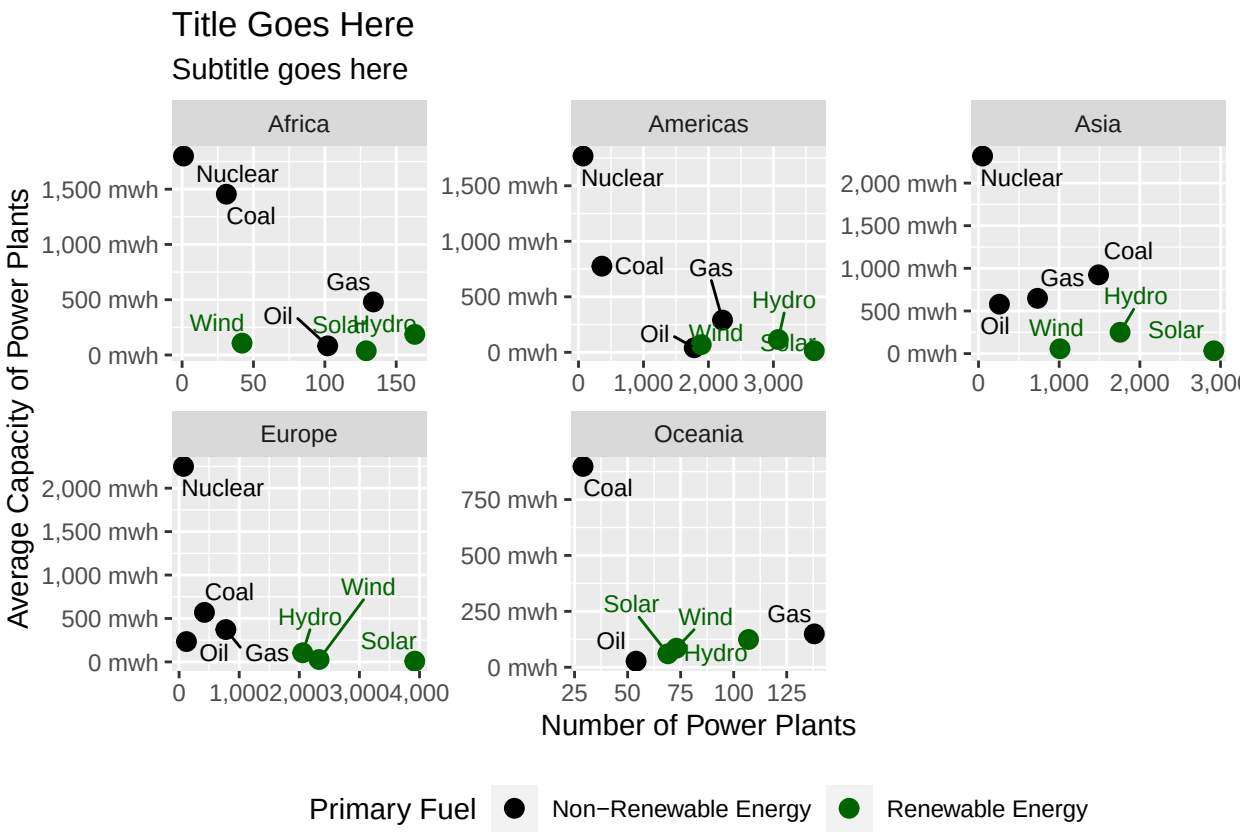
---



**Question 7**: Reproduce the scatterplot below, inserting your own meaningful title and subtitle. Key details to reproduce:

- Antarctica is filtered out
- We are only including `primary_fuel` == Solar, Hydro, Wind, Gas, Coal, Oil, Nuclear
- The x axis is the number of facilities for each continent/renewable/primary_fuel combo
- The y axis is a calculated variable of the *mean* of `capacity_mw`
- The labels are repelled
- The x and y axis numeric labels are formatted with commas
- The y axis has a suffix of *"mwh"*
- The legend is on the bottom of the axis

---



For the final three questions of the problem set, you will be required to make your own visualizations without any reference viz to guide you. You may choose to use either of the following data sets, which are included in the Problem Set repository:

- `chicago_schools.csv`, a file of progress report cards for Chicago schools from 2011-2012. You can find the documentation for this dataset here. The provided data is adjusted so that the variable names are easier to work with in R.
- `gun_background_checks`, a file of Firearm Background Check data collected by the FBI. You can find details on the data here

---

**Question 8**: Create a bar chart of any element from your chosen dataset. The plot should either be colored, grouped or faceted to show how the key variable varies by some other categorical variable. It may be any orientation and it may be grouped, stacked, or 100% stacked.

> Your plot should have a legend (if necessary), descriptive title and subtitle, should *not* use default `ggplot2` colors or the default `ggplot2` theme, and all plot elements should be human readable (no overlapping text, no acronyms unless they are defined, no underscores). Axis scales should make sense and be rounded to 2 digits or less (if applicable).

**BONUS: (not required)** - change the font of the title, subtitle, and caption *without* removing other features of the theme.

---

**Question 9**: Create a boxplot of any numeric variable from your selected data, segmented by some relevant categorical data point. Be sure the boxplots are meaningfully sorted by some key statistic of the selected numeric variable.

> Your plot should have a legend (if necessary), descriptive title and subtitle, should *not* use default `ggplot2` colors or the default `ggplot2` theme, and all plot elements should be human readable (no overlapping text, no acronyms unless they are defined, no underscores). Axis scales should make sense and be rounded to 2 digits or less (if applicable).

**BONUS: (not required)** - color the boxplots by the grouped median of your numeric variable.

---

**Question 10**: Create a scatterplot using two numeric variables from the data. Use a categorical variable to highlight *one* category from the data. (For example, the `safety_icon` variable in the `chicago_schools.csv` data has six categories, but your visualization should only highlight one category with a color).

> Your plot should have a legend (if necessary), descriptive title and subtitle, should *not* use default `ggplot2` colors or the default `ggplot2` theme, and all plot elements should be human readable (no overlapping text, no acronyms unless they are defined, no underscores). Axis scales should make sense and be rounded to 2 digits or less (if applicable).

**BONUS: (not required)** - add a faint dotted line that represents the average for X axis variable. Do the same for the Y axis variable.