PPOL 563 - Problem Set 2

Alex Lundry

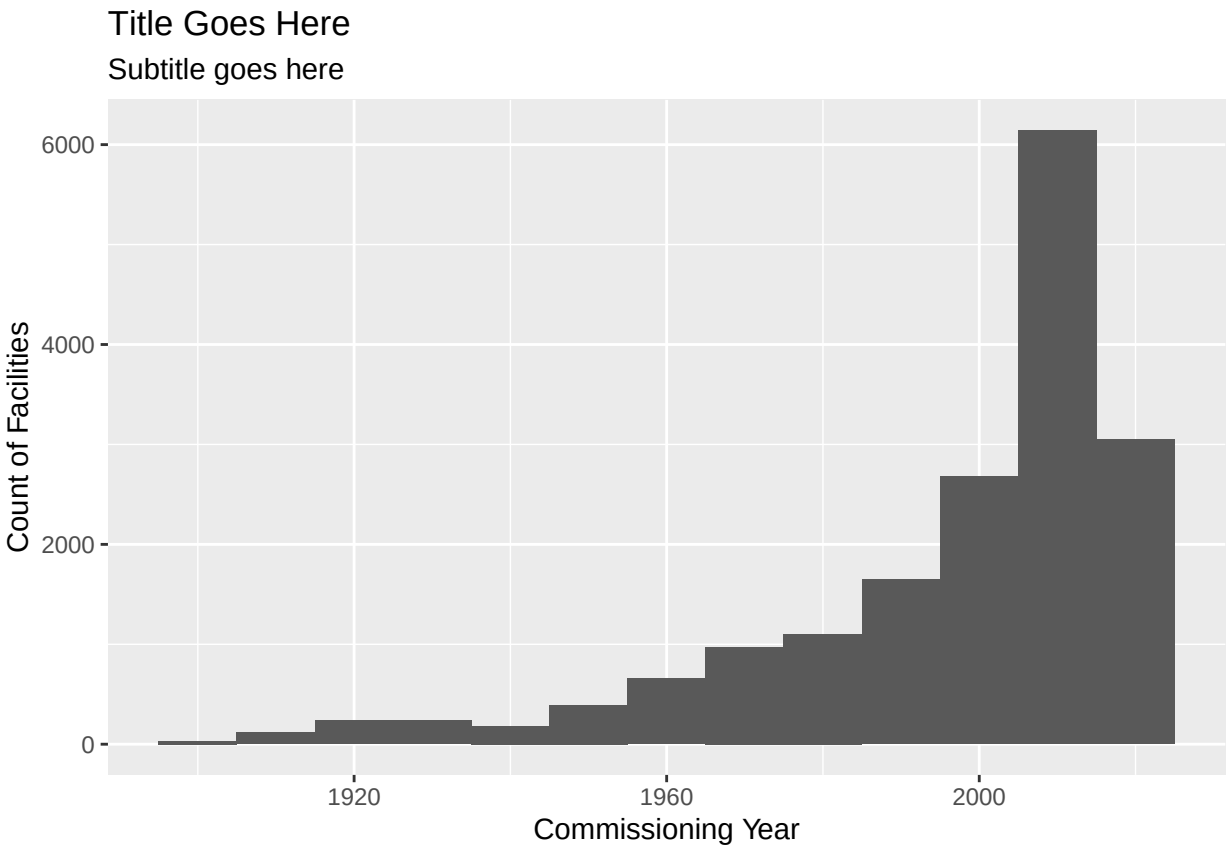
2023-02-14

Problem Set 2 - PPOL 563

The first question requires 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: 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.


Title Goes Here
Subtitle goes here
6000
4000
2000
0
Count of Facilities
1920
1960
2000
Commissioning Year

For the next question of the problem set, you will be required to make your own visualization 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 2: 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.