Portfolio Problem 1: Data Visualization

Data Overview

For my research, I often work with policing data and I collaborate with community organizations to better understand how their communities are being policed. These datasets can be complex and very challenging to work with. Policing datasets are often good examples of messy data in the real world. For this portfolio problem, I've downloaded a dataset for Police Use of Force from the Minneapolis Open Data Initiative. Here is a link to the website. This dataset includes 39,718 records.

To access the data, run the following code chunk. It runs off the page, but you should be able to copy/paste it.

```
MN_puof <- readr::read_csv("https://www.math.carleton.edu/ckelling/data_science/MN_Police_Use_of_Force.c
```

You can see the "codebook" at the link above. One challenge of doing research on policing data is that the data documentation is often very poor. You will see that the variables listed at the website do not include written out definitions, but the column names are pretty informative. You will want to consider any necessary caveats at the end of this assignment.

You can filter the dataset to just certain neighborhoods using the code below. You can change the neighborhoods of interest using the `Neighborhood` vector.

```
library(dplyr)
MN_puof_subset <- filter(MN_puof, Neighborhood %in% c("West Calhoun", "Beltrami", "Linden Hills"))
```

Note: Studying policing can be emotionally and mentally challenging! My goal is for you to see how we can use statistics and data science to explore a topic of urgent societal importance. However, I encourage you to take time and space to care for yourself if this is a difficult topic for you.

Assignment

Choose 2-5 neighborhoods in the `MN_puof` data set and create a visualization that compares and contrasts the incident-level features in these neighborhoods (Problem, Is911Call, Sex, etc.). Your goal is to tell us something interesting using a well-crafted, thoughtfully-prepared data visualization. One-two graphics should suffice, but you may include more if you choose (do not include more than 4). Once you have designed and created your graphic(s), you should write a short blog post describing your findings. (Note: your graphics do not need to include all of the incident features; the key is to find interesting comparisons.)

Your blog post should be short— I envision an introductory paragraph that explains your findings and provides some context to your data, the data graphic(s), and then a caption-like paragraph providing more detail about what to look for in the data graphic and how to interpret it. That's it. I'm looking for something that is insightful and well-crafted. As always, you should spend some time thinking about context, scale, color, etc.

You should write your blog post in R Markdown and create your data graphic using `ggplot2`. To submit your work, push both your R Markdown (.Rmd) file and knitted PDF document to GitHub. You should then link the submission to your Gradescope assignment. Do not forget to give your post an informative title!

Data consideration: Be careful about using the spatial locations in the data (`X`, `Y`) if you choose to use them! The Minneapolis Police Department spatially privatizes their policing data. You do not need to create map for this assignment, but if you do, consider using *jittered locations* as well as any necessary caveats in your blog post.

Here are some examples of articles that are similar in spirit to yours. Most of these are much longer than yours will be, and may contain multiple graphics, but the idea is similar: use a good data graphic to tell us something we don't already know.

- How to Tell Someone's Age When All You Know Is Her Name
- A Better Way to Find the Best Flights and Avoid the Worst Airports
- NYC Taxis and Ubers
- Data on people who went to the ER for wall-punching

Assessment

Below, I've included some items that I will consider for assessment of your first portfolio problem.

Reproducibility

- All necessary components of the problem were pushed to GitHub and linked to Gradescope
- The graphic and blog post is contained in a single .Rmd file
- The .Rmd file contains all necessary code to reproduce the graphic(s)
- The .Rmd file does not contain any unnecessary code
- The .Rmd file does not produce any unnecessary content (package loading messages, warnings, etc.)

Data Visualization

- The graphics are correctly constructed and appropriate for the task at hand
- The graphics are interpretable without reading the text of the blog post
- Aesthetics (color, shape, etc.) faithfully reflect the representation of the underlying data
- The graphics have readable axis labels, units, and legends (where appropriate)
- The graphic lists the source(s) of the data
- The graphic contains no color, symbolism, or text that is irrelevant to the question it seeks to answer or argument that it seeks to make.

Communication

- An informative title is included for the blog post
- The blog post contains very few grammatical/spelling mistakes
- The blog post clearly introduces the data set in your own words (it cannot be copied from the prompt)
- The blog post clearly summarizes the major takeaways from your data visualization substantively
- The blog post includes any necessary caveats, given the data
- The caption paragraph helps the reader interpret the graphic(s)