Day 2 Mini Problem Set for All Things Data

Your Name

Date

Introduction

Try to solve each problem using day two's material and some minimal Googling. If you get stuck, please come to office hours which are 2-5pm in Sage 8A.

Remember to consider writing pseudo code prior to trying to code your answer to the problem!

This problem set should take 1-2 hours. If you feel like you're struggling through getting things to click, don't get discouraged. Struggling through a problem is the best way to learn it! (If it starts to get discouraging, come to office hours!)

This pset will apply the concepts of data manipulation using `dplyr` and `tidyr`, and creating basic visualizations using `ggplot2`.

Data Manipulation Problems

Problem 1: Install and Load Necessary Packages

Install and load `dplyr`, `tidyr`, and `ggplot2` if you haven't already.

```
# Insert code here
```

Problem 2: Data Filtering

Load the `mpg` dataset from `ggplot2` by running `data(mpg)`. Make a new dataset called `toyota_cars` that's filter for cars manufactured by "toyota".

```
# Loading the mpg dataset
data(mpg)

# make new dataset
toyota_cars <- mpg %>%
  filter(INSERT_CODE_HERE)
```

Problem 3: Data Transformation

Use `mutate()` to create a new column `hwy_km` in `toyota_cars` converting highway miles per gallon (`hwy`) to kilometers per liter (1 mile = 1.60934 km, 1 gallon = 3.78541 liters). This is a basic unit transformation.

hwy_km_per_liter = hwy * 1.6 / 3.79

```
# make new column
toyota_cars <- toyota_cars %>%
  mutate(hwy_km = INSERT_UNIT_CONVERSION_HERE)
```

Problem 4: Summarizing Data

Group the `toyota_cars` dataset by the variable `class` and summarize the average highway miles per gallon for each class.

```
# Data manipulation
summary <- toyota_cars %>%
  # use the group_by() function, where the input is class, then a pipe
  # use summarize() funciton, where you create a new variable and assign it the mean hwy)

# print the new dataframe
summary
```

Data Visualization Problems

Problem 5: Basic Histogram

Create a histogram of the `hwy` variable for the `toyota_cars`.

```
# basic histogram
ggplot(INSERT_DATA_SET, aes(x = INSERT_X_VARIABLE)) +
  # add the function that makes a histogram
```

Problem 6: Scatter Plot

Create a scatter plot with `displ` on the x-axis and `hwy` on the y-axis. Color the points by `class`.

```
# basic scatter plot
ggplot(INSERT_DATA_SET, aes(x = INSERT_X_VAR,
                            y = INSERT_Y_VAR,
                            color = INSERT_COLOR_VAR)) +
  # add function that makes a scatter plot
```

Problem 7: Best Data Visualization Practice

Choose one of the plots above to apply best data visualization practices to. Specifically:

- Write clear labels and titles
- Make it as simple as possible while not becoming reductive
- Make sure all parts of graph are legible
- Consider the colors, if using them

```
# Insert code here
```

Problem 8: Save Your Plot

Save one of the plots you created to your project directory as a PNG file.

```
# Insert code here
```

Problem 9: Best File Practice

Consider a research project (one of your own or one you made up). Describe the file structure you would use for your project, including what your raw data may look like and what your cleaned data may look like. List the name of the scripts you'd write, and what each would do.

Your answer here. No need to code for this question, but you can write pseudo code if you want to.