E401/M518: Problem Set 1a

Introduction to R & Data Visualization

Fall 2023

Due: September 19 2023

Please work on the following questions and hand in your solutions in groups of at most 3 students.

Part 1: R questions

Question 1: First Steps with R

Work through the R Tutorial uploaded on Canvas answering all the questions. You should have done this during week 1 anyway. Since a good understanding of basic R is essential for the rest of the semester, I ask you to explicitly hand in your answers to Part 2 (Question 2), Part 5, and Part 6.

Question 2: Data Visualization

The following questions use the `mpg` data set that comes with the `tidyverse` library.

Question 2.1: Visualization Basics

1. Run `ggplot(data = mpg)`. What do you see and why?
2. What does the `drv` variable describe? Read the help for `?mpg` to find out.
3. What happens if you make a scatterplot of `class` vs `drv`? Why is the plot not useful?

Question 2.2: Aesthetic Mappings

1. What's gone wrong with this code? Why are the points not blue?

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy, color = "blue"))
```

2. Which variables in `mpg` are categorical? Which variables are continuous? (Hint: type `?mpg` to read the documentation for the dataset). How can you see this information when you run `mpg`?

3. Map a continuous variable to `color`, `size`, and `shape`. How do these aesthetics behave differently for categorical vs. continuous variables?

4. What happens if you map the same variable to multiple aesthetics?

Question 2.3: Facets

1. What happens if you facet on a continuous variable?

2. What do the empty cells in plot with `facet_grid(drv ~ cyl)` mean? How do they relate to this plot?

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = drv, y = cyl))

ggplot(data = mpg) +
    geom_point(mapping = aes(x = drv, y = cyl)) +
    facet_grid(drv ~ cyl)
```

3. Take the following faceted plot:

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(~ class, nrow = 2)
```

What are the advantages to using faceting instead of the colour aesthetic? What are the disadvantages? How might the balance change if you had a larger dataset?

Question 2.4: Geometric Objects

1. What geom would you use to draw a line chart? A boxplot? A histogram? An area chart?

2. Will these two graphs look different? Why/why not?

```
ggplot(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_point() +
  geom_smooth()

ggplot() +
  geom_point(data = mpg, mapping = aes(x = displ, y = hwy)) +
  geom_smooth(data = mpg, mapping = aes(x = displ, y = hwy))
```

Part 2: Your project

Question 3

Get started on your project! Think about an empirical question that you find exciting and start looking for available data to answer this question. It is very likely that many questions/project ideas will fail eventually, so it's a good idea to get started by just brainstorming to get several ideas. Then start to investigate the 2 or 3 ideas that you like most in more detail.

If you cannot come up with a specific question, you can also start the other way round: start browsing the Internet for interesting data sets. Check out what is available, what kind of information the data contains and how these data could be used for an interesting analysis.

Write up to one page about this process and what you think the most promising idea for your project is and why.Finally, use what you learned about data visualization to create 2 or 3 graphs that visualize an interesting aspect of the data set(s) that you were exploring. These graphs don't have to be perfect and they don't commit you to a specific project, but I would like to see you get started on something.