Lab 2: Data Visualization

Math 141

Insert Your Name Here

Due: Tuesday, February 10th at 11:59pm on Gradescope

In this lab, you will practice

1. Decomposing graphs by identifying their `geom`s and determining how the variables map to the `aes`thetics of those `geom`s.
2. Reflecting on the editorial choices of a graphic and how those choices impact the messages conveyed by the graphic.
3. Creating and interpreting `ggplot2` graphs.
4. Producing and interpreting summary statistics.

Add any collaborators here!

I collaborated with…

Remember that you are not allowed to use any generative AI models (like ChatGPT) on your lab.

Problems

- For this lab, you don't need to worry about labels and a title for your plots **unless a question specifies otherwise**.
- Run the following chunk to load the necessary packages.

```r
# Load the necessary packages
library(lubridate)
library(tidyverse)  # Includes ggplot2
```

Problem 1: Visualizing somerville with barplots

For Problems 1-4, we will use the somerville dataset that you started exploring in Lab 1. This version of the dataset in the somerville_clean.csv file has been cleaned, with the 999 values replaced with NA. The atypically large values for age have also been converted to NA.

Starting in 2011, Somerville city officials decided to start measuring residents' overall happiness and satisfaction with many aspects of life in Somerville, MA; their goal was to become the first city in the United States to systematically track people's happiness. The Happiness Survey is an initiative of the Mayor's Office of Analytics and Innovation (SomerStat) and has the goal of supporting the local government's understanding of what makes Somerville a great place to live.

The happiness survey is sent out every two years to a random sample of Somerville households, with the instruction that each survey should be filled out by one person. The survey asks residents to rate their personal happiness, wellbeing, and satisfaction with City services. This combined dataset includes the survey responses from 2011 to 2019.

Run the code in the following chunk to load the cleaned version of the somerville data.

```r
somerville <- read.csv("data/somerville_clean.csv")
```

Variables in the somerville data:

- year: year that response was collected
- happy: happiness rating, on a scale from 1 to 10 with 10 being most happy
- satisfied_life: rating for satisfaction with life in general, on a scale from 1 to 10 with 10 being most satisfied
- satisfied_somerville: rating for satisfaction with Somerville as a place to live, on a scale from 1 to 10 with 10 being most satisfied
- age: age in years
- children_home: whether respondent has children under 18 living at home
- ward: neighborhood ward the respondent lives in
- precinct: neighborhood precinct the respondent lives in
- student: whether respondent is a student
- convenient_transport: rating for how convenient it is to get to where you want to go, on a scale from 1 to 10 with 10 being most satisfied
- direction_somerville: whether respondent feels the City is headed in the right direction or is on the wrong track
- housing: whether respondent rents or owns their home
- primary_transport: primary form of transportation
- lived_here: years respondent has lived in Somerville

a. Let's start by getting a sense of how people get around Somerville. Fill in the blanks (`---`) to create a bar graph of the variable `primary_transport`. Make sure to put the completed code into an R chunk.

```
ggplot(data = ---, mapping = aes(x = ---)) +
  geom_---()
```

```
# Create bar graph
```

b. Reorder the bars of your barplot from (a) by the number of responses in a group. *Hint: Try using either `reorder()` or `forcats::fct_infreq()`. Use `?function_name` to read the documentation about each one.*

```
# Create reordered bar graph
```

c. Create two bar plots of the variables `primary_transport` and `direction_somerville`:

- For the first one, display the **counts** of each category of `direction_somerville` within `primary_transport` groups (bars).
- For the second one, display the **proportions** of each category of `direction_somerville` within `primary_transport` groups (bars).

Compare and contrast the information provided in the two plots.

```
# Bar plot: Counts


# Bar plot: Proportions
```

d. Use `count()`, `group_by`, and `mutate()` to reproduce the values shown from your plots in (c) as a contingency table and conditional proportion table.

```
# Contingency table


# Conditional proportions table
```

e. Suppose that we don't like the default order of the categories of `direction_somerville`. It can be helpful to store categorical variables using the `factor` data type - this allows us to set the possible categories and their order explicitly. Run the code below to manually reorder the categories by turning `direction_somerville` into a factor (as well as remove observations that have missing values of `direction_somerville`). What do you think the default ordering of categories was? Recreate the bar graph of proportions from (c) with this new ordering for `direction_somerville` in the chunk below.

```r
# Remove missing values of `direction_somerville`
somerville <- somerville %>%
  filter(!is.na(direction_somerville))

# Reorder the categories of population_trend
somerville$direction_somerville <- factor(somerville$direction_somerville,
                                          levels = c("Wrong track",
                                                     "Not sure",
                                                     "Right direction"))

# Recreate the plot below
```

f. The label for "Multiple Modes" of transportation is long and sometimes spills over onto other labels. Use `mutate()` and `case_when()` to overwrite the `primary_transport` variable so that "Multiple Modes" is changed to just "Multiple," and other categories are left the same. Then, recreate the plot from (e).

g. Draw two conclusions from your graph in (f).

Problem 2: Visualizing `somerville` with scatterplots

a. Create a scatterplot that shows the relationship between `satisfied_somerville` and `convenient_transport`; on the x-axis, put the variable that you think may *explain* the behavior of the other variable.

b. It is difficult to detect a meaningful pattern in the scatterplot from part (a); since the responses were given as whole numbers on a 1 - 10 scale, many of the responses are overlaid on top of each other. Recreate the plot from part (a) by using `geom_jitter()` rather than the `geom` used previously and by making the `geom` somewhat transparent.

c. Use `group_by()` and `summarize()` to calculate the **mean** value of `satisfied_somerville` for each level of `convenient_transport`. How do these values compare with what you observe in the graph?

d. The geom `geom_smooth(method = "lm", se = FALSE)` adds a line that tries to capture the general trend. Add this layer to the scatterplot in part (b) and describe the observed relationship in the context of the data.

e. Explore whether the relationship between the two quantitative variables from part (d) hold within each category of response for `direction_somerville`. Describe what you see. Does the observed association from (d) seem to be consistent within each group?

f. The "Not sure" category can seem ill-defined, and we may wish to directly compare those who believe Somerville is on the wrong track with those who believe it is on the right track without the distraction of the "not sure"s. Use `filter()` to remove those who aren't sure about the direction of Somerville, and save the result in a new data frame called `somerville2`. Then, recreate the graph from part (e) using `somerville2`.

Problem 3: Visualizing `somerville` with histograms

a. Create a histogram of the `age` variable. You may need to play around with the number of bins to get a clear idea of the shape. Compute the mean, median, variance, and standard deviation of `age.`

b. Create a histogram of the `lived_here` variable. Describe the distribution in the context of the data. Compute the mean, median, variance, and standard deviation of `lived_here.` Why are the mean and median so different?

Problem 4: Visualizing `somerville` with boxplots vs histograms

a. Create a boxplot that shows how the distribution of scores for `satisfied_somerville` differs by response category for `direction_somerville`.

b. Create a **faceted** histogram plot that shows how the distribution of scores for `satisfied_somerville` differs by response category for `direction_somerville`. *Hint: You'll want to mess with the* `bins` *size to get a good sense of the distribution.*

c. Compare and contrast what you learned from the boxplot in (a) and the histogram plot in (b).

d. Which response category for `direction_somerville` tends to have the highest satisfaction score? Does your answer imply that people who think Somerville is headed in the right direction are all very satisfied with Somerville? Use the boxplot in (a) and/or the histograms in (b) in your justification.

Problem 5: MWRA Visualization

The Massachusetts Water Resources Authority (MWRA) graph tracks the presence of COVID-19 in the Boston-area wastewater. Run the following `R` chunk to load the data.

```r
# Load the data
wastewater <- read.csv("data/wastewater.csv")

# Make date a date type (this helps date labels automatically look better)
wastewater$Sample_date <- as.Date(wastewater$Sample_date)
```

a. Go to https://www.mwra.com/biobot/biobotdata.htm and look at the first graph. What are the geoms? For each geom, discuss what variables are being mapped to the aesthetics of that geom.

Note: The graph doesn't indicate this but the line segments incorporate interval estimates (which we will learn in detail later in the semester) of the daily RNA copies per mL. For now, when discussing these interval estimates, feel free to refer to the *lower bound of the interval estimate* and the *upper bound of the interval estimate*.

b. Create a scatterplot of the copies/mL over time. For now, don't include the moving average.

```
# Hint: Before jumping into graphing, insert code to inspect the data
# and to find your variable names.

# Create scatterplot
```

c. Modify your scatterplot to deal with the overplotting and to color your points based on the station. Similar to MWRA, let's use green and orange to color the points.

```
# Modify scatterplot
```

d. Now add the seven day moving average to your scatterplot. (We will add in the uncertainty measures, the vertical line segments, later in the semester.)

What is tricky here is that you will need to **change** what variable is mapped to `y` for your new `geom`. We have provided some partially completed sample code below to help you get started. Make sure to put that code in an R chunk below.

```
# Add moving average to your graph
___insert ggplot code from part b___ +
  geom_---(mapping = aes(y = ---))
```

e. The **order** of the layers in a `ggplot` matter. Make a new graph that swaps the order of your two `geoms` in (d).

f. Now give your graph from (e) nice labels and a title.

g. Reflect on the graph you have just created. How well do the **geom**s represent the data? Do you prefer the **point**, **line**, or both geoms? Why?