Homework - Week 8

Write your name here

2025-03-12

Preface

The goal of this assignment is to help you gain more familiarity with using **ggplot** to visualize proportions and distributions. In this homework we provide less scaffolding and more open-ended questions. As always, please come to office hours and reach out to your teaching staff if you have any questions.

Data

We will work with data on NYC Yellow cab trip data in January 2021 from TCL. We'll start by importing these data and assigning them to the name `trips`, and creating another data frame assigned to the name `speeds`.

```
trips <- read_csv("taxi-trips.csv")

speeds <- trips |>
  mutate(
    duration = dropoff_datetime - pickup_datetime,
    hour = hour(pickup_datetime)
  ) |>
  filter(duration > 0 & trip_distance > 0) |>
  select(pickup_datetime, dropoff_datetime, hour, trip_distance, duration) |>
  mutate(speed = trip_distance / (as.numeric(duration)/60/60)) |>
  filter(speed < 55) # filter out 55+ miles per hour
```

1. We'll start by making some simple visualizations of the distribution of the trip distance. First, make a histogram. Make the first bar start right at zero, customize the bins or binwidth to suit your tastes, and use a named color to delineate between bars.

2. Make a density plot of the distribution of the trip distance.

3. Use the data frame `speeds` to plot the density of trip speeds, faceting by the hour during which the trip began. Use `geom_density`'s `fill` argument to fill the densities with a named color or hex code of your choice. Arrange the facets so they are in 6 rows and 4 columns. What, if any, conclusions can you draw from the resulting plot?

...

4. Make an overlapping density plot of the distribution of speeds. Fill by hour, treating hour as a categorical variable by encoding it as a factor (see `?as_factor()`). Adjust the transparency to make it more readable.

5. Reproduce the plot above as a ridgeline plot instead of an overlapping density plot. Fill all the densities with a single named color of your choice, and use `color` to make the density lines themselves white. What, if any, conclusions can you draw from the resulting plot?

...

Comparing your plots from questions 3, 4, and 5, which one do you think is the most effective visualization? Why?

...

6. How much do riders tip? Compute the tip percentage using `tip_amount` as a share of `fare_amount`, converted to percentage points. Remove outliers where tips are less than 0% or more than 100% so we can focus on the main part of the distribution. Create a histogram of tip percentage. Is the distribution unimodal or bimodal? Explain your answer.

...

7. What fraction of NYC cab rides include a tip? Use the `trips` data to classify trips according to whether they included a tip amount of zero, or greater than zero. Remove any observations with negative tip amounts. Finally, make a side-by-side bar plot of the proportion of rides that included a tip and did not include a tip. Are side-by-side bars a good choice for this particular data visualization?

Side-by-side bars are ...

8. Now use the data to make a stacked bar plot of the proportion of rides that included a tip and did not include a tip. Are stacked bars a good choice for this particular data visualization?

Stacked bars are ...

9. Finally, use the data to make a pie chart of the proportion of rides that included a tip and did not include a tip. Is a pie chart a good choice for this particular data visualization?

A pie chart is ...