Homework 3

Insert Name

Math 141, Week 3

Due: 11:59pm, Friday February 11

Instructions

Work through the problems below and submit this document as a knitted .pdf to the Math 141 S22 Wells Lecture gradescope page.

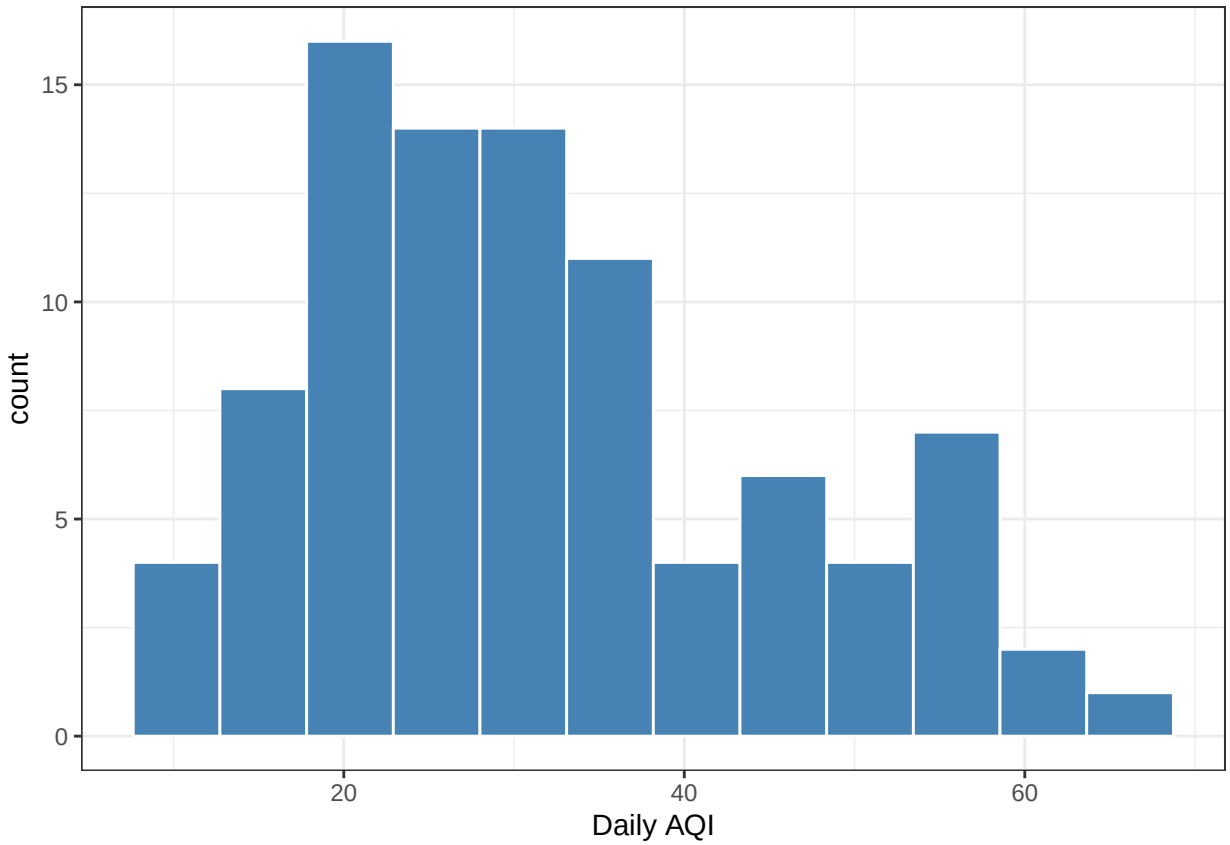
For each problem, put your solution between the bars of red stars.

Acknowledgements

If you work with a classmate, please write a note acknowledging this.

Exercise 1

Daily air quality is measured by the air quality index (AQI) reported by the Environmental Protection Agency. This index reports the pollution level and what associated health effects might be a concern. The index is calculated for five major air pollutants regulated by the Clean Air Act and takes values from 0 to 300, where a higher value indicates lower air quality. AQI was reported for a sample of 91 days in 2011 in Durham, NC. The histogram below shows the distribution of the AQI values on these days.



a. Estimate the median AQI value of this sample based on the graph.

b. Would you expect the mean AQI value of this sample to be higher or lower than the median? Explain your reasoning.

c. Estimate Q1, Q3, and IQR for the distribution based on the graph.

d. Would any of the days in this sample be considered to have an unusually low or high AQI? Explain your reasoning.

Exercise 2

It is time for each of you to create your own version of the Anne Hathaway graph from the first week of class. For an actor of your choice (other than Anne Hathaway), create a scatterplot of Rotten Tomatoes rating and Box office gross, with points colored according to another categorical variable of your choice

a. Which actor will you be graphing? Beyond Rotten Tomatoes rating and Box office gross, what is your third variable? Note: It should be a variable you can collect online or can be defined by the user.

b. Create the data frame for your actor and include at least 10 movies. To show how to create your own data frame in R, an example has been included below containing 4 Ryan Gosling movies. For reference, Rotten Tomatoes ratings and Box Office can be found here.

c. Construct your graph. Choose a color palette that is not the default `ggplot2` palette.

d. Let's add a bit more context to our graph. If you haven't already, add nice labels, a title, and a caption with the data source. Additionally, label the movies in the graph. Here's some code to get you started. Make sure to replace all lines that say *insert _____* with the appropriate layers, and then change `eval = FALSE` to `eval = TRUE`.

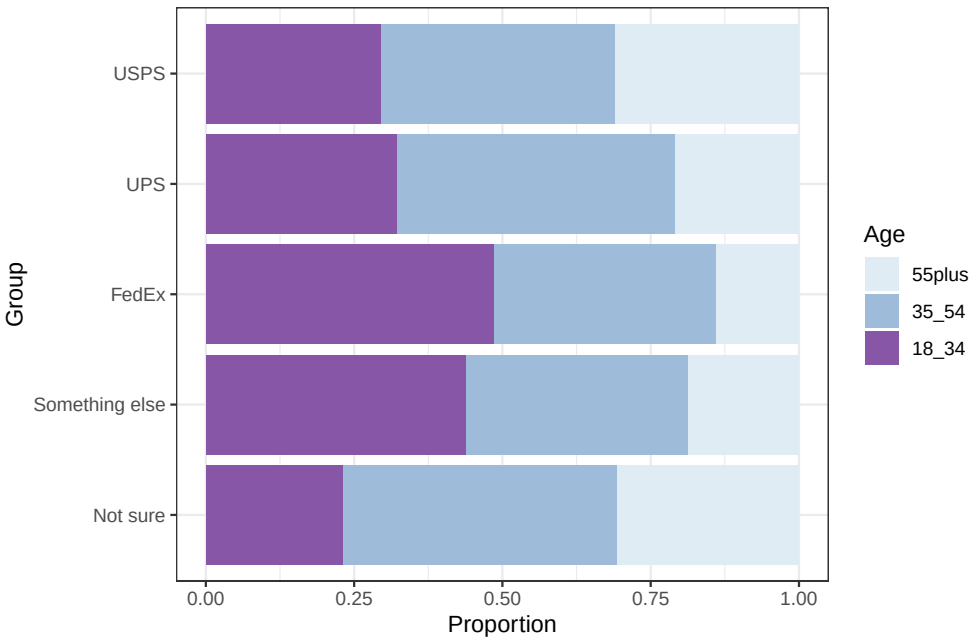
e. Draw some conclusions about the relationships between box office gross, ratings, and your third variable. What does this tell us about your actor?

Exercise 3

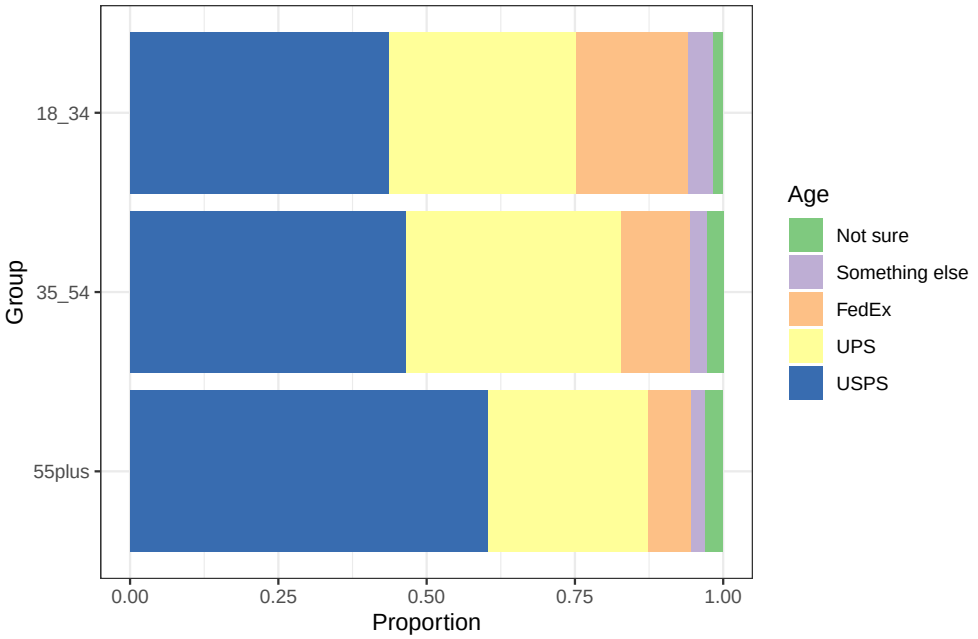
A local news survey asked 500 randomly sampled Los Angeles residents which shipping carrier they prefer to use for shipping holiday gifts. Data on the age group of each respondent, along with shipping preference, is contained in the `shipping` data frame, which can be loaded with the following code (don't worry about interpreting what each line of code means)

a. Consider the two graphics created from the data by the code below. Which graph would you use to understand the shipping choices of people of different ages? Which graph would you use to understand the age distribution across different types of shipping choices?

Graph 1



Graph 2



b. Use the `table` and `prop.table` functions to create a contingency table representing the same information in Graph 1.

c. Use the `table` and `prop.table` functions to create a contingency table representing the same information

in Graph 2.

d. A new shipping company would like to market to people over the age of 55. Who will be their biggest competitor?

d. FedEx would like to reach out to grow their market share to balance the age demographics of FedEx users. To what age group should FedEx market?