Problem Set 1 Spring 2022

Note: Grading is based both on your graphs and verbal explanations. Follow all best practices as discussed in class, including choosing appropriate parameters for all graphs. *Do not expect the assignment questions to spell out precisely how the graphs should be drawn. Sometimes guidance will be provided, but the absense of guidance does not mean that all choices are ok.*

Read *Graphical Data Analysis with R*, Ch. 3

1. SATs

[7 points]

Data: *StudentSurvey* in **Lock5withR** package (Remember to add a proper title and labels to every plot.)

a) Draw multiple horizontal boxplots of `SAT`, by `Year`. What do you observe? (Hint:You can remove all blank and NAs)

b) Draw a grouped bar chart of average `Exercise` by `Award` filled with `Year`. (You can ignore NAs.)

c) Draw a percentage stacked barchart (each bar = 100%) of average `Exercise` by `Award` filled with `Year`. Compare to the plot in b), which one do you prefer and why?

2. Bad Drivers

[7 points]

Data: *bad_drivers* in **fivethirtyeight** package

a) Draw two histograms–one with base R and the other with **ggplot2**–of the variable representing the Percentage of drivers involved in fatal collisions who were alcohol-impaired without setting any parameters. What is the default method each uses to determine the number of bins? (For base R, show the calculation.) Which do you think is a better choice for this dataset and why?

b) Draw two histograms of the `perc_alcohol` variable with boundaries at multiples of 5, one right closed and one right open. Every boundary should be labeled (15, 20, 25, etc.)

c) Adjust parameters–the same for both–so that the right open and right closed versions become identical. Explain your strategy.

3.Titanic Survival

[8 points]

Data: *TitanicSurvival* in **carData** package

a) Use QQ (quantile-quantile) plots with theoretical normal lines to compare `age` of **passengers who did not survive from Titanic** for the three different levels of `passengerClass`. What are some findings and for which class does the distribution of the `age` variable appear to be closest to a normal distribution?

b) Draw density histograms with density curves and theoretical normal curves overlaid of `age` for the three passenger classes.

c) Use a statistical method of your choice, such as the Shapiro-Wilk test, to determine which `age` distribution is closest to a normal distribution.

d) Did all of the methods for testing for normality (a, b, and c) produce the same results? Briefly explain.

4. Birds

[8 points]

Data: *birds* in **openintro** package

a) Use appropriate techniques to describe the distribution of the `speed` variable noting interesting features.

b) Create horizontal boxplots of `speed`, one for each level of `time_of_day`.

c) Create ridgeline plots for the same data as in b)

d) Compare the boxplot plots and the ridgeline plots.