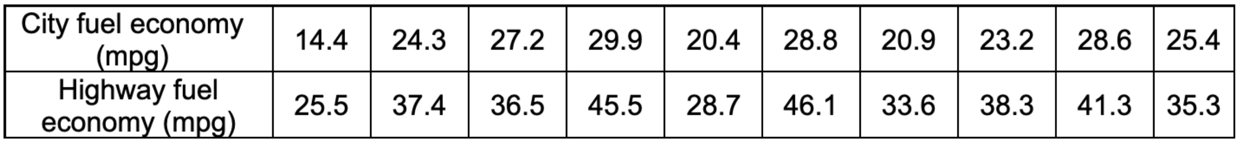
Assignment #2
October 12, 2022

This assignment is intended to test your understanding of the analysis and visualization of two variable data. Assignments should be submitted as a digitally generated LaTeX document (questions 9-11 can be done by hand in class). The datasets can be found on the jupyter server and listed below:

- pokemon.csv: Information on all generations of Pokémon.
- lego.csv: Information on every lego set ever released
- heart.csv: Information on heart disease on a large number of people
- ted.csv: Information on ted talks
- youtube.csv: Information on yotube videos

1. Using the heart disease dataset create a barchart showing the number of people who have heart disease and who do not.
   (a) Using the chart you created estimate the probability someone has heart disease.
   (b) If you were told that a person was 18 years old and wanted to estimate the probability they have heart disease, would you use your estimated probability from part (a)? Why or why not?
   (c) Estimate the probability someone has Heart disease given they are 18 years old.
2. Use a relative barchart to compare the proportion of heart disease across people who smoke. What can you infer from your plot?
3. Create a contingency table showing Pokémon type across the variable legendary status. Are the two variables independent?
4. Which type of Pokémon is the most likely to be legendary?
5. Fueleconomy.gov gives the city and highway fuel economy for all makes and models of vehicles back to 1984. The scatterplot displays the city and highway fuel economy (mpg) for a random sample of ten 2021 vehicles.



City fuel economy (mpg)	14.4	24.3	27.2	29.9	20.4	28.8	20.9	23.2	28.6	25.4
Highway fuel economy (mpg)	25.5	37.4	36.5	45.5	28.7	46.1	33.6	38.3	41.3	35.3

   (a) Calculate and interpret the correlation between city fuel economy and highway fuel economy of vehicles.
   (b) If fuel economy was measured in feet per gallon, how would the value of the correlation be affected?
   (c) The Rolls-Royce Ghost EWB gets 14.4 city mpg and 25.5 highway mpg. What affect does this point have on the correlation. Explain.

6. Find the LSRL for the cost of a lego set vs. the number of pieces it has. Is a linear model appropriate here?

7. Find the LSRL for the attack of a Pokémon vs. the defense of a Pokémon. Is a linear model appropriate here?

8. Create a linear model describing the relationship between the number of likes and views for ted talks. Calculate and interpret the coefficient of determination for your model.

9. Create a linear model showing the number of number of likes a youtube video gets vs. the number of comments it has. Is a linear model appropriate here?

10. Consider bivariate quantitative data with variables X and Y. Prove that the correlation coefficient r is the same when X is transformed as $X' = aX + c$ for scalars a and c.