Alyssa Bockman and Thalia Koutsougeras
Milestone 1
Introduction To Data Science (CMPS3160)

Currently, we are looking at two main datasets. One is focusing on distribution of diseases across the 50 states over the last 100 years, and the other focuses on food access in counties in Louisiana and includes categories and descriptors like access to a vehicle, income level, age, race and receival of SNAP benefits. The reasoning behind these choices is that Alyssa is interested in public health, thus leading to the selection of the diseases dataset. Thalia is interested in investigating food access disparities in Louisiana because she witnesses this issue first hand as someone from New Orleans and has enjoyed volunteering for Second Harvest (as has Alyssa).

Our collaboration plan is to meet at least once a week over zoom on Wednesday nights. We met in person twice to work on Milestone 1. We already have each other's phone numbers, and we have set up our private Github repository where we uploaded our preliminary notebook files.

For the disease distribution dataset, we have successfully read the csv file into a dataframe, performed some light manipulation including removing the "NaN" values and mapping each disease to a specific color, and plotted the prevalence of disease vs the year on a scatter plot. This revealed a detailed (although cluttered) graph depicting the peaks and valleys of each disease's prevalence over the years, and we think it would be interesting to complete some research on things like if and when vaccines were developed for each disease to see the effects on the disease prevalence. One interesting stat we found was how measles was extremely prevalent before 1970, and how some diseases like smallpox became virtually nonexistent before 1960, which matches what we know to be true! Additionally, we would like to examine disease prevalence in different regions of the country to try to distinguish patterns such as how hotter vs. colder and wetter vs. drier climates affect the spread of the disease. If possible, it would be interesting to find a dataset that shows average high and low temperatures across each state over the last hundred or so years. We have also discussed finding data on population size of each state and comparing the number of people with the number infected to get an accurate view of transmission. From there we could identify which states feature the highest and lowest transmission of any given disease which could open up some other questions such as effects of population density as well as how rates of vaccination compare to transmission. With infectious diseases there are so many questions that can be asked to try to identify some kind of measurable pattern

For the food access dataset, we have slightly modified the original datasheet by extracting only the data for Louisiana into another sheet on the excel file because 75,000 rows of data yields issues when trying to load the data into a dataframe. We successfully loaded this sheet converted as a csv into a dataframe. This dataframe had 147 variables, so we narrowed down the dataframe into particular columns that we were interested in to make the dataset more workable.

We began graphing some scatter plots to get a better sense of the data. We graphed housing units receiving SNAP vs. median family income, which yielded a negative correlation, as expected! We hope to answer the question of what factors influence disparities in food access in Louisiana, such as whether or not race, age, vehicle access, distance from a grocery store, etc.