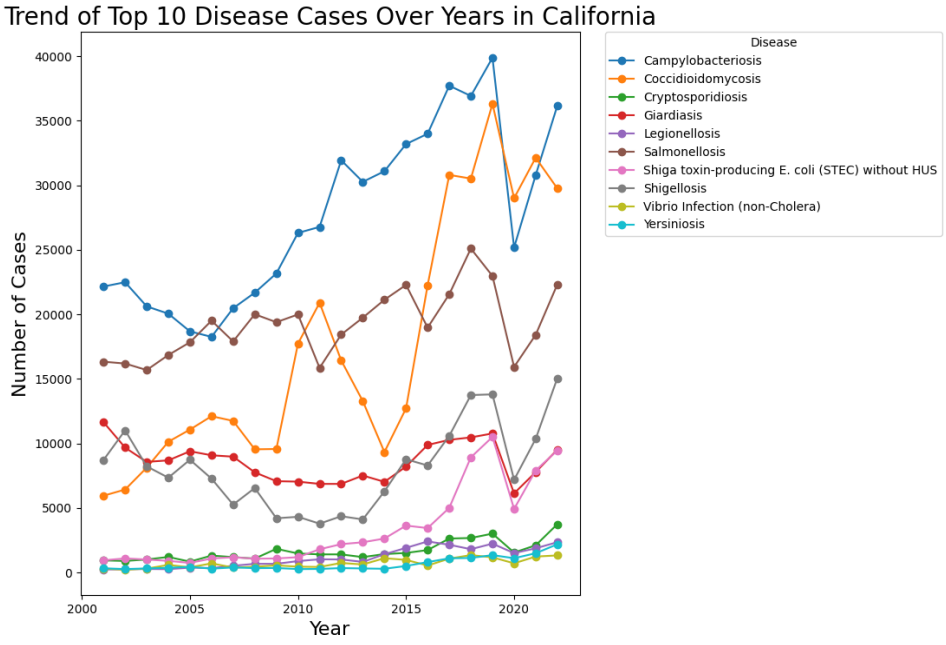
Project 1 – Data Analysis and Visualization

Data source used: https://catalog.data.gov/dataset/infectious-diseases-by-disease-county-year-and-sex-6e856

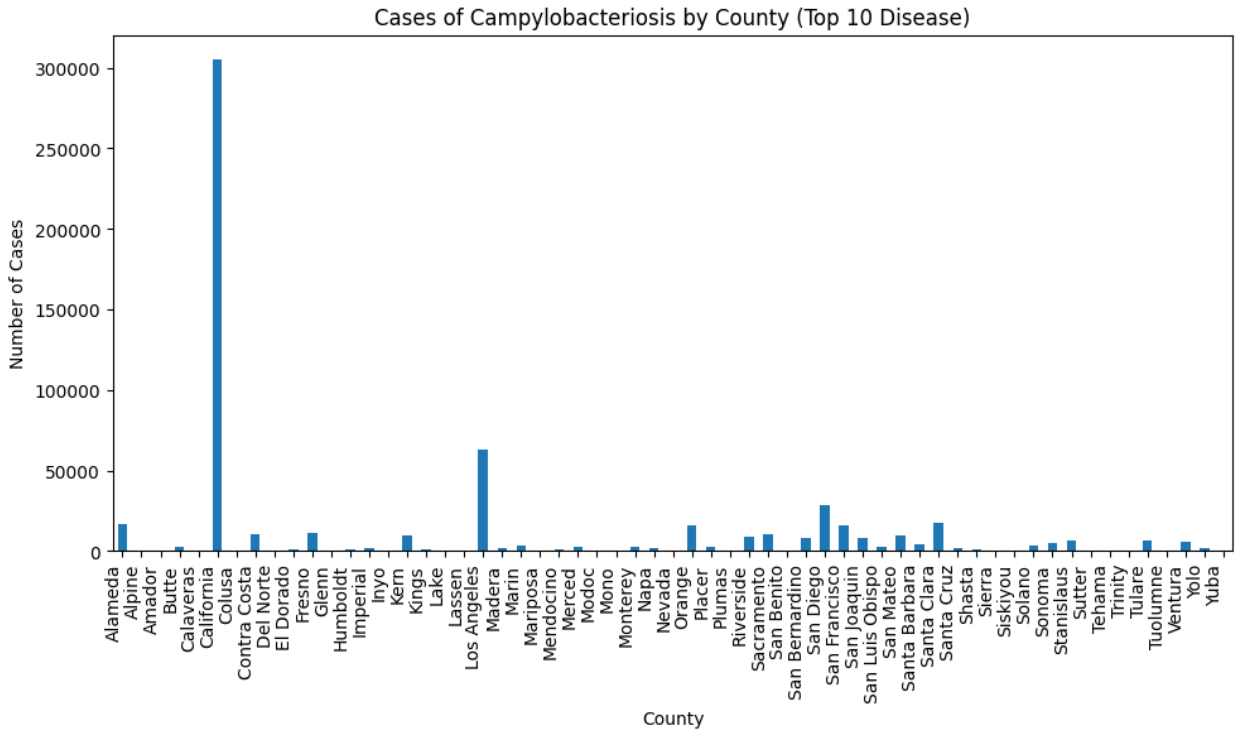
This is data focuses on infectious diseases in California from 2001 to 2023. The data is grouped by disease, county, number of cases (female, male, total), year, population, rate, lower 95 CI, upper 95 CI

The three angles chosen to represent this data are listed below:

1. **Focusing on the top 10 most prevalent diseases, show case numbers for each disease over years (2001 to 2023)**
    - **Why does this angle make sense?**
        - It is important to easily visualize disease trends to understand whether the diseases are becoming more or less prevalent which is beneficial for physicians, scientists, and public health experts
    - **How did I accomplish this task?**
        - To accomplish this task, I used pandas and matplotlib in Python. First, I had to load the data set into a pandas data frame. Next, I had to calculate total number of cases for each disease to then determine the top ten diseases based on case numbers. Lastly, I had to group by year and disease and assign it to a value which I use for the lines on the line graph, which I use matplotlib to create. I had to play around a lot with the graph/key margins and spacing because the key was cut off significantly.



2. **Show total case numbers of a specific top 10 disease in each county**
    - **Why does this angle make sense?**
        - It is important for physicians, scientists, and public health experts to visualize disease case numbers by county because this will allow them to see important trends. For example, maybe there is an environmental or demographic factor that contributes to a particular disease being more prevalent in a particular county
    - **How did I accomplish this task?**
        - To accomplish this task, I used pandas and matplotlib in Python. First, I had to load the data set into a pandas data frame. Next, I had to calculate total number of cases for each disease to then determine the top ten diseases based on case numbers. I had to implement code to ensure the selected disease is in the top ten diseases. Lastly, I had to filter through the diseases for the specific disease and group by the county and case number. This is assigned to a value which I use for the bar graph, which I use matplotlib to create. I also had to play around with the graph for this one because the labels for the counties were overlapping.



3. **Focusing on the top 10 most prevalent diseases, show total number of cases for female, male, and both genders**
    - **Why does this angle make sense?**
        - It is important for physicians, scientists, and public health experts to see trends in infectious diseases amongst gender. This may provide insight on how these diseases are contracted and development of strategies for minimizing disease spread.
    - **How did I accomplish this task?**
        - To accomplish this task, I used pandas and matplotlib in Python. First, I had to load the data set into a pandas data frame. Next, I had to calculate total number of cases for each disease to then determine the top ten diseases based on case numbers. I grouped by disease/gender and assigned that to a value, which I use for the bar graph, which I use matplotlib to create.

Project 1 – Data Analysis and Visualization