PM566 Final Project

Qiushi Peng

This is my PM566 Final Project website.

Introduction

The dataset, called *Infectious Diseases by Disease, County, Year, and Sex*, is downloaded from California Health and Human Services Open Data Portal.

This dataset contains case counts and rates by disease, county, year, and sex for selected infectious diseases that met the surveillance case definition in California. There are 9 columns in the dataset: *Disease*, *County*, *Year*, *Sex*, *Cases*, *Population*, *Rate*, *Lower_95___CI*, and *Upper_95___CI*. There are 167,974 rows. The data represent cases with an estimated illness onset date from 2001 through the last year indicated from California Confidential Morbidity Reports and/or Laboratory Reports. Data captured represent reportable case counts as of the date indicated in the "Temporal Coverage" section below, so the data presented may differ from previous publications due to delays inherent to case reporting, laboratory reporting, and epidemiologic investigation.

After looking at the whole dataset, we formed two questions: We would like to know the infectious diseases with the highest prevalence, and in which year the diseases had a highest infection rate. In that year, did the diseases spread evenly across the whole state? Is there a significant difference in infection rates between males and females?

Methods

Data acquisition

Infectious-disease dataset was downloaded from "https://data.chhs.ca.gov/dataset/infectious-disease".

Geographical dataset was downloaded from "https://public.opendatasoft.com/explore/dataset/us-county-boundaries/export/?disjunctive.statefp&disjunctive.countyfp&disjunctive.name&disjunctive.namelsad&disjunctive.stusab&disjunctive.state_name&refine.stusab=CA".

Data cleaning and wrangling

1. Merge *Infectious-disease dataset* and *Geographical dataset*.
2. The dataset has 6 columns. Among them, columns "Cases" and "Rate" have several missing values because of "Scoring Criteria" prevent them from being publicated. Thus, we can remove them.
3. Remove NA rows.
4. The data type of column "Rate" is *chr*, which we do not want it to be. Thus, we change the data type to *num*.

5. The “County” column includes rows called “California”, which is the state not a county, so we delete them. We saved the aggregate “California” data into a new variable “Cal”.

Libraries used

We used several R libraries: *data.table*, *tidyverse*, *dplyr*, *plotly*, *DT*, *knitr*

Results

Figures and Table

Figure 1

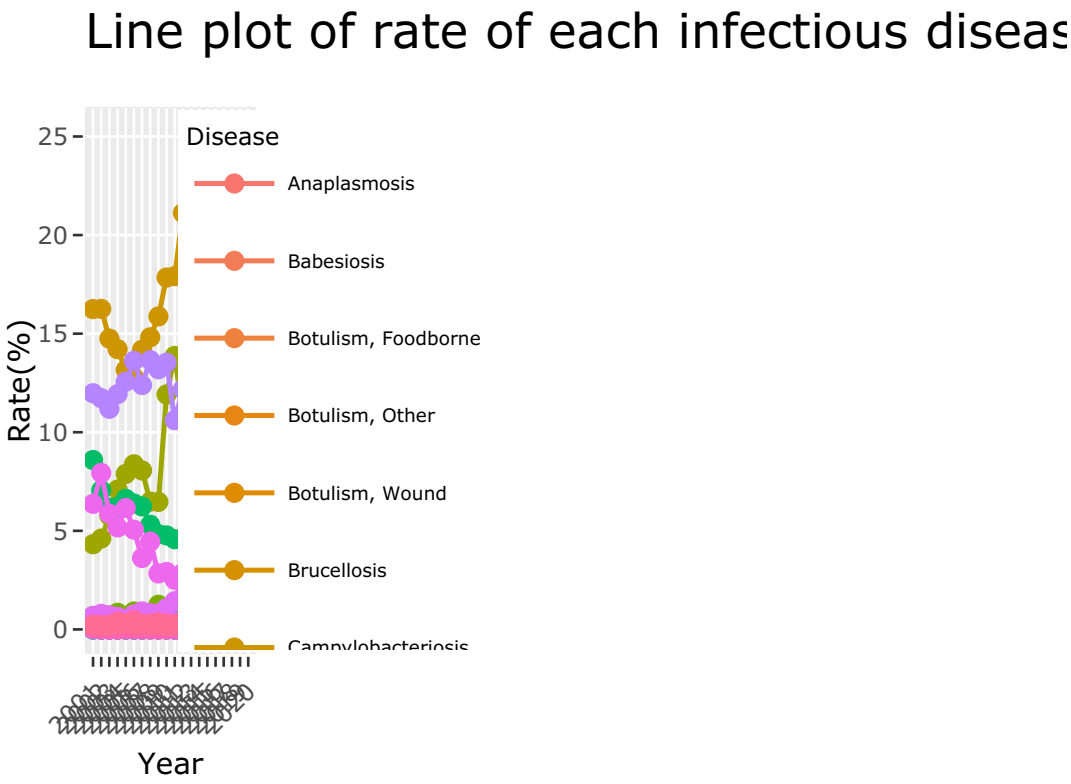

Figure 2

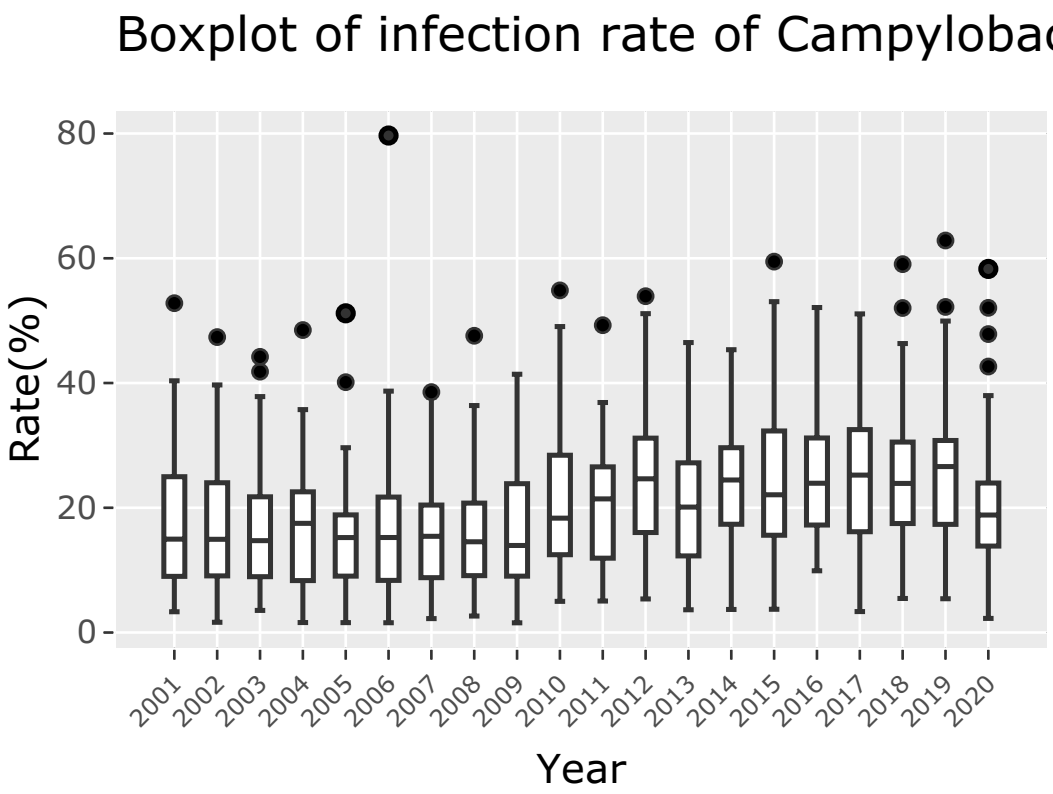
Boxplot of infection rate of Campylobac
Rate(%)
0
20
40
60
80
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year

Figure 3

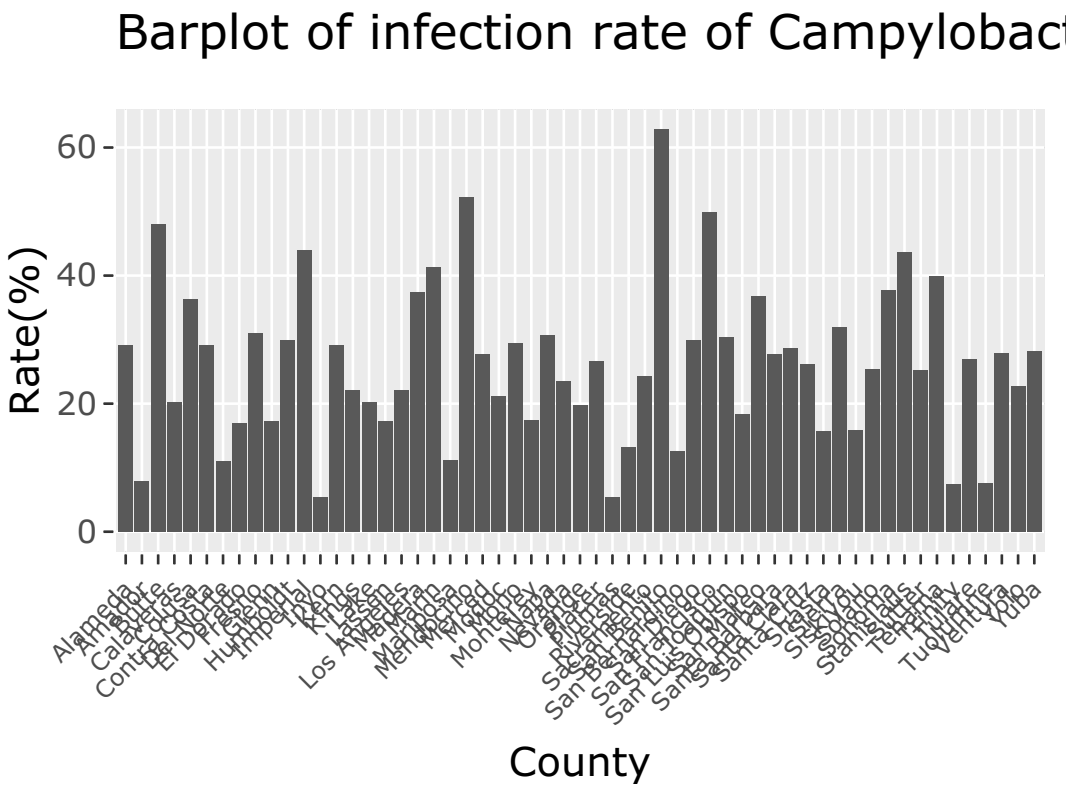
Barplot of infection rate of Campylobact
Rate(%)
0
20
40
60
County

Figure 4

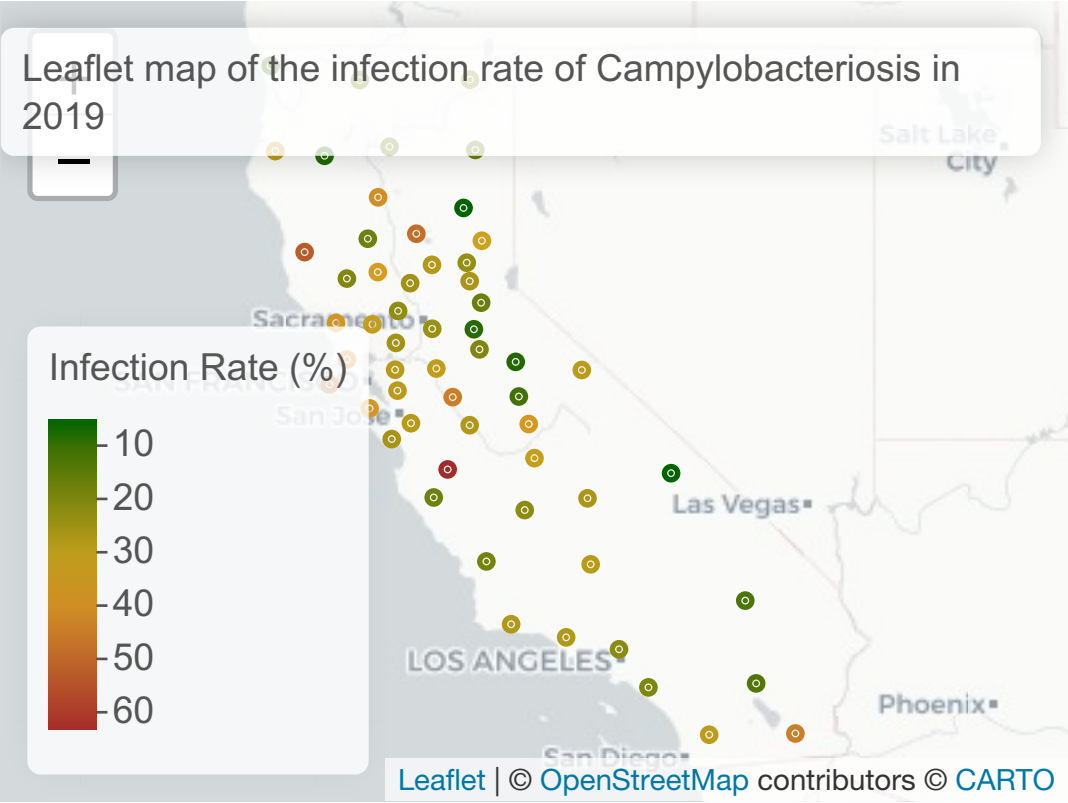

Table 1

Sex Cases Population Rate

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  contingency_table
## X-squared = 0.00027594, df = 1, p-value = 0.9867
```

Conclusion and Summary

1. We can see that *Campylobacteriosis*, *Salmonellosis*, *Giardiasis*, *Shigellosis*, and *Coccidioidomycosis* always have a higher infection rate from 2001 to 2020 than other infectious diseases. *Shiga toxin-producing E. coli (STEC) without HUS* infection rate has increased significantly a lot since 2011 (Figure 1).
2. *San Benito* had a very high infection rate of *Campylobacteriosis* in 2019 (Figure 2, 3), which is more than 60%.
3. There is not a significant difference between male and female for the infection rate of Campylobacteriosis in San Benito in 2019 ($p = 0.9867$, Table 1).
4. Counties around San Francisco had a higher infection rate of Campylobacteriosis in 2019. Inland area had relatively lower infection rate of Campylobacteriosis (Figure 4).

Reference

1. California Department of Public Health, Center for Infectious Diseases, Infectious Diseases Branch, Surveillance and Statistics Section, 2001-2020. Infectious-Diseases-by-Disease-County-Year-Sex.csv