You will be using the dataset “Infectious\_Diseases\_Google.csv” in this homework. The description for each of the variables is provided in the document “Data Description.pdf”. You need to show the code and the output for each of these questions.

Note: You need to share the .rmd file also in the submission. The file should have the naming convention “HW1\_<FirstName\_LastName>.docx”. Otherwise, you will not get any credit. Please answer the questions by pasting the code and output below each question in this file. Your need to use the “Knit to Word” tool in RStudio.

1. Read the dataset into a variable “mydata”. Remove all the rows with missing data on one or more of the variables. Provide a summary of the dataset. For the month of August in the year 2017, for the state of Pennsylvania, which disease had the highest value of “count”? Similarly, which month in the year 2016 had the lowest value of “hits” for the state of Georgia (You need the name of the disease also here)? (2 Points)

mydata<-read.csv("Infectious\_Diseases\_Google.csv",

header=TRUE, sep=",")

mydata[!complete.cases(mydata),]

mydata2<-na.omit(mydata)

summary(mydata2)

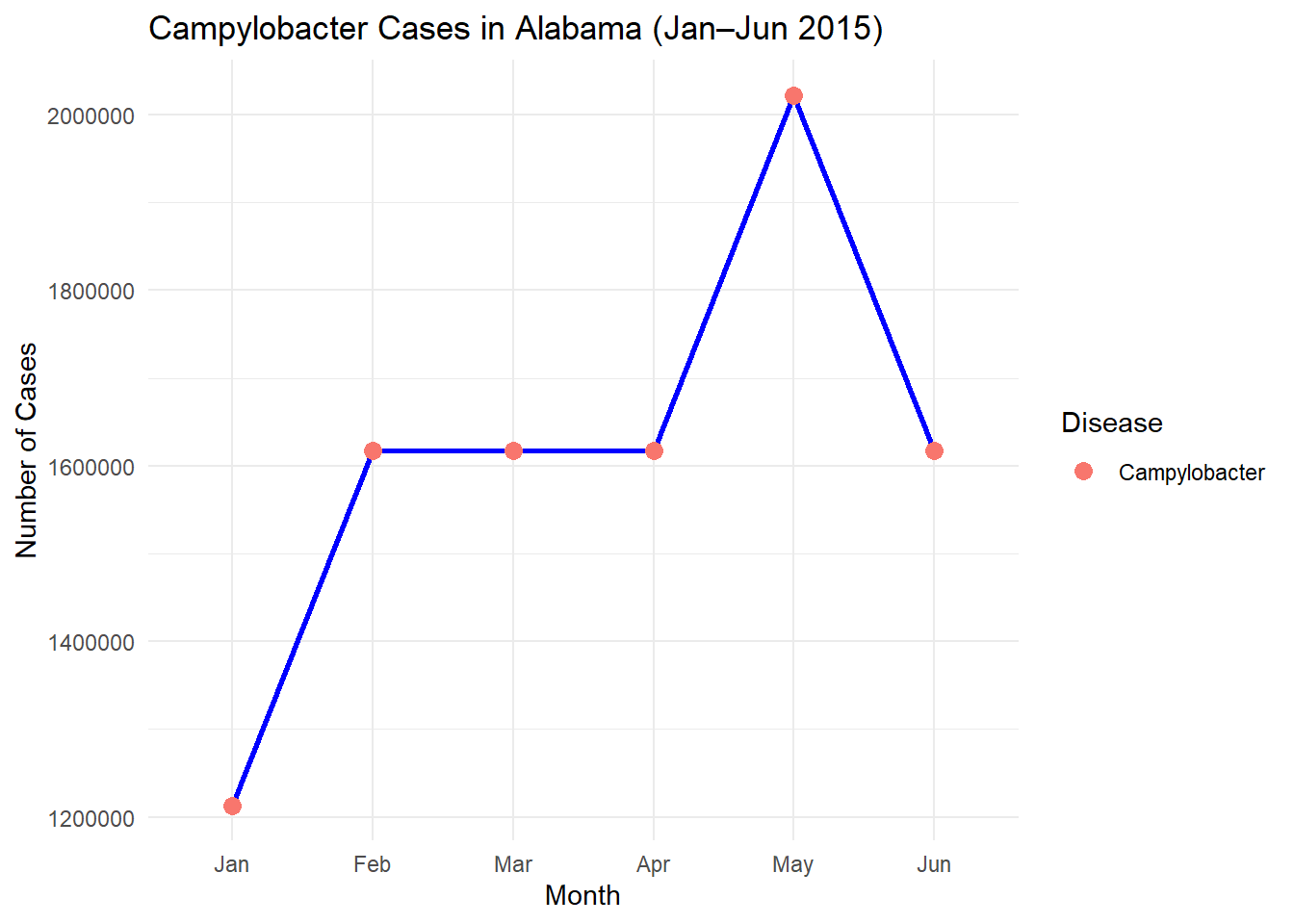
**August 2017 – Highest Count: Chlamydia with 886**

**December 2016 for state of Georgia: Campylobacter**

sql.aggregate.2017 <-sqldf("Select Year, Month, Disease, count, state from mydata2 where year= 2017 and month= 8 and state= 'PENNSYLVANIA' order by count desc limit 1")

sql.aggregate.2016 <-sqldf("Select month, year, disease, hits, state from mydata2 where year= 2016 and state= 'GEORGIA' order by hits asc limit 1")

1. Generate a visualization that shows the number of cases of the disease “campylobacter” in the state “Alabama” for the year 2015 across all the first six months. You need to pick the appropriate visualization, label the axes and have a title. There also needs to be a “legend” on the visualization if it is appropriate. What trends do you notice? Compare this visualization with another visualization for the next six months and provide the trends you notice. (2 Points)
2. **library**(ggplot2)
3. **library**(sqldf)
4. sql.dataframe <- sqldf("select month, SUM(month\_count) as Cases
5. from mydata2
6. where year = 2015
7. and month in ('1', '2', '3', '4', '5', '6')
8. and disease = 'campylobacter'
9. and state = 'ALABAMA'
10. group by month
11. ")
12. *# Make sure month is ordered correctly (Jan–Jun)*
13. sql.dataframe$month <- factor(sql.dataframe$month,
14. levels = c("1","2","3","4","5","6"),
15. labels = c("Jan","Feb","Mar","Apr","May","Jun"))
16. *# Plot 1*
17. ggplot(sql.dataframe, aes(x = month, y = Cases, group = 1, color = "Campylobacter")) +
18. geom\_line(color = "Blue", linewidth = 1) +
19. geom\_point(size = 3) +
20. labs(title = "Campylobacter Cases in Alabama (Jan–Jun 2015)",
21. x = "Month",
22. y = "Number of Cases",
23. color = "Disease") +
24. theme\_minimal()



**library**(ggplot2)

sql.dataframe2 <- sqldf("select month, SUM(month\_count) as Cases

from mydata2

where year = 2015

and month in ('7', '8', '9', '10', '11', '12')

and disease = 'campylobacter'

and state = 'ALABAMA'

group by month

")

*# Make sure month is ordered correctly (Jul–Dec)*

sql.dataframe2$month <- factor(sql.dataframe2$month,

levels = c("7","8","9","10","11","12"),

labels = c("Jul","Aug","Sept","Oct","Nov","Dec"))

ggplot(sql.dataframe2, aes(x = month, y = Cases, group = 1, color = "Campylobacter")) +

geom\_line(color = "Green", linewidth = 1) +

geom\_point(size = 3) +

labs(title = "Campylobacter Cases in Alabama (Jul–Dec 2015)",

x = "Month",

y = "Number of Cases",

color = "Disease") +

theme\_minimal()

A graph of a disease

AI-generated content may be incorrect.

**Trends I noticed were that during the first 6 months, the number of cases continued to rise from the beginning of the year. During the second half of the year, the number of cases would increase for one month, then drop back down for the next and finally plateaued from November to December.**

1. Building on the visualizations in the previous question, generate a visualization that compares the cases of the same disease for the year 2016 (across all the 12 months) for the neighboring states of “Alabama”. You need to investigate Geographic map of the United States to determine which states share a border with Alabama. What trends do you notice? (2 Points)

**library**(sqldf)

sql.dataframe3 <- sqldf("select state, month, SUM(month\_count) as Cases

from mydata2

where year = 2016

and month between '1' and '12'

and disease = 'campylobacter'

and state in ('TENNESSEE', 'GEORGIA', 'MISSISSIPPI', 'FLORIDA')

group by state, month

order by state, month

")

*# Month order*

sql.dataframe3$month <- factor(sql.dataframe3$month,

levels = c("1","2","3","4","5","6","7","8","9","10","11","12"),

labels = c("Jan","Feb","Mar","Apr","May","Jun","Jul","Aug","Sept","Oct","Nov","Dec"))

*# Plot*

**library**(ggplot2)

ggplot(sql.dataframe3, aes(x = factor(month), y = Cases, fill = state)) +

geom\_bar(stat = "identity", position = position\_dodge(width = 0.8)) +

scale\_x\_discrete(labels = month.abb[1:12]) +

labs(

title = "Campylobacter Cases in Neighboring States of Alabama (2016)",

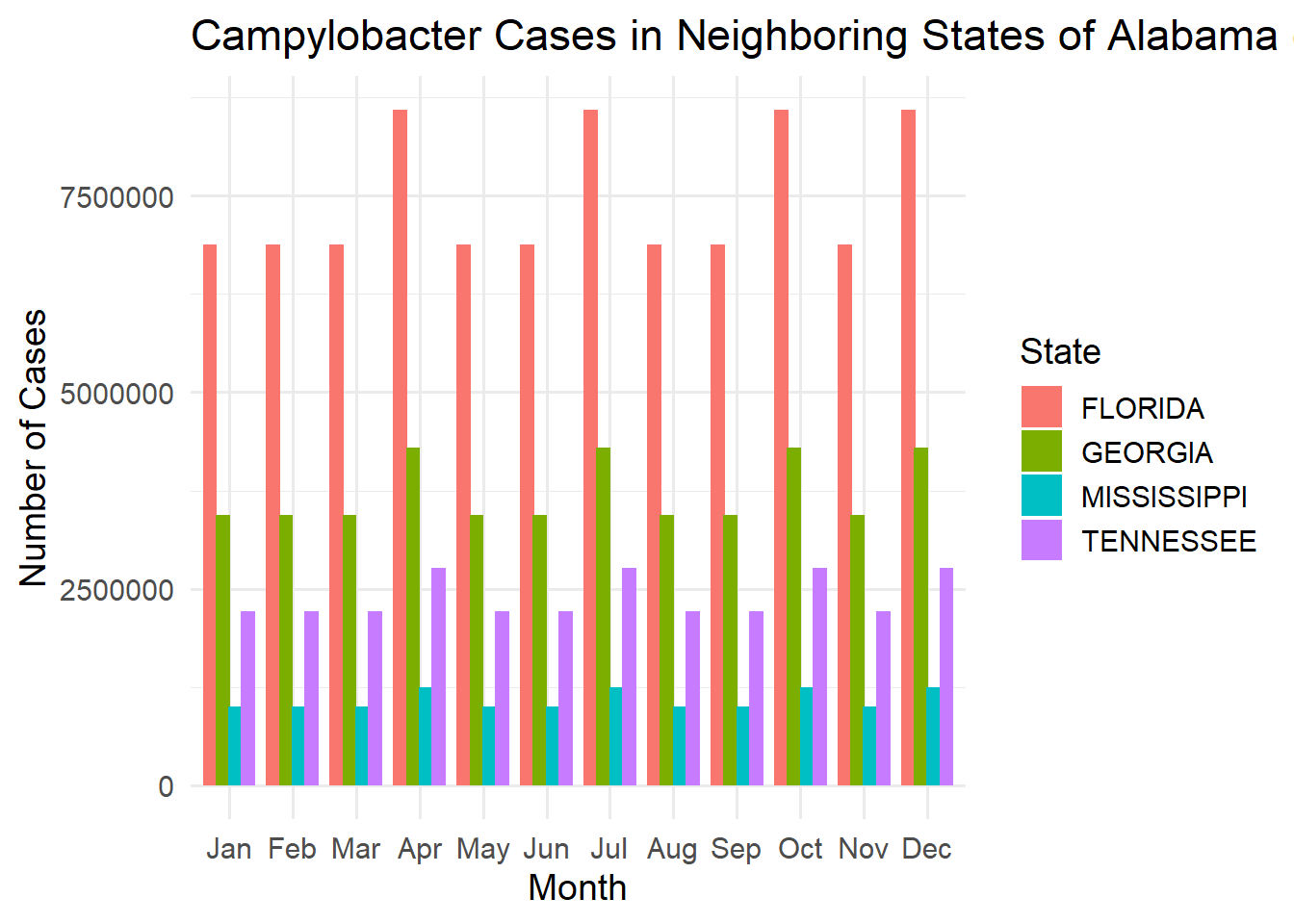
x = "Month",

y = "Number of Cases",

fill = "State"

) +

theme\_minimal(base\_size = 14)



**Looking at the bar chart, the trend that stands out right away is that the state of Florida had the highest number of cases throughout the year 2016, by far. Looking deeper into the states individually, it shows that the states with the larger population count tend to have the higher number of cases. With Mississippi having the lowest population count, they also had the least number of cases, which remained under 2,500,000 per month, for the entire year.**

1. Generate a new dataset that aggregates “mydata” at the year level instead of month level. The new data set should contain the following columns – Year, Disease, Total Cases in the year, Average Cases Per Month, Median Cases Per Month, Average Hits per Month. Generate an appropriate visualization that shows the trends for “Median Cases Per Month” for the disease “Salmonella” across all the years in the dataset (2 Points)
2. *#Question 4*
3. **library**(sqldf)
4. *# Step 1: run SQL for sum, avg, and hits*
5. sql.dataframe4 <- sqldf("SELECT year, disease,
6. SUM(month\_count) AS total\_cases\_year,
7. AVG(month\_count) AS avg\_cases\_per\_month,
8. AVG(hits) AS avg\_hits\_per\_month
9. FROM mydata2
10. GROUP BY year, disease
11. ORDER BY year, disease
12. ")
13. *# Step 2: add the median*
14. median <- aggregate(month\_count ~ year + disease, data = mydata2, FUN = median)
15. *# Step 3: merge results*
16. sql.dataframe4 <- merge(sql.dataframe4, median,
17. by = c("year", "disease"))
18. *# Rename for clarity*
19. names(sql.dataframe4)[names(sql.dataframe4) == "month\_count"] <- "median\_cases\_per\_month"
20. salmonella\_df <- subset(sql.dataframe4, disease == "Salmonella")
21. **library**(ggplot2)
22. *# Normalize disease values (trim + lowercase)*
23. sql.dataframe4$disease\_clean <- tolower(trimws(as.character(sql.dataframe4$disease)))
24. *#Try an exact match first, then a contains match*
25. salmonella\_df <- subset(sql.dataframe4, disease\_clean == "salmonella")
26. *#Plot trend of median cases per month across years*
27. ggplot(salmonella\_df, aes(x = year, y = median\_cases\_per\_month)) +
28. geom\_line(linewidth = 1.2) +
29. geom\_point(size = 2) +
30. geom\_text(aes(label = median\_cases\_per\_month), vjust = -0.6, size = 3) +
31. scale\_x\_continuous(breaks = unique(salmonella\_df$year)) +
32. labs(
33. title = "Median Cases Per Month for Salmonella",
34. x = "Year",
35. y = "Median Cases Per Month"
36. ) +
37. theme\_minimal(base\_size = 14)

A graph showing the number of cases per month

AI-generated content may be incorrect.

(5) Generate a pie chart for the total number of cases of the disease “chlamydia” for the year “2016” using the new dataset. Each section of the pie should represent a particular state. Based on this, what are top three states for this disease? (2 Points)

*#Question 5*

**library**(sqldf)

sql.dataframe5 <- sqldf("select year, state, disease, Sum(month\_count) as annual\_total\_cases

from mydata2

where year = 2016

and disease = 'chlamydia'

group by state

")

**library**(ggplot2)

*# Use sql.dataframe5 (all states included)*

ggplot(sql.dataframe5, aes(x = "", y = annual\_total\_cases, fill = factor(state))) +

geom\_bar(width = 1, stat = "identity") +

coord\_polar(theta = "y") + *# make it a pie chart*

theme(axis.line = element\_blank(),

axis.text = element\_blank(),

axis.ticks = element\_blank(),

plot.title = element\_text(hjust = 0.5)) +

labs(fill = "State",

x = NULL,

y = NULL,

title = "Chlamydia Cases by State (2016)",

caption = "Source: mydata2")

A colorful chart with names

AI-generated content may be incorrect.

**I could not get the percentages to show clearly on the pie chart but once you aggregate the data and filter it, California, Texas and Florida are the top 3 states for this disease.**