Food Borne Pathogens

Anthony Girard, Timothy Hedspeth, Yutong Li

2022-10-05

Please note that the text in this file is not up to date, but the code is what we use for the EDA for our project, to just see the code please see the EDA R file

Data Tools

In this analysis we consider longitudinal data from 2013-2022 regarding 3 food borne illnesses, Salmonella, E.Coli, and Camplyobacter. We have 3 different data sets which contain 25 variables for each of the illnesses individually, we note that the Salmonella data set is the largest with 93,763 observations, Camplyobacter is the second largest with 44,949 observations and E. Coli with 34,805 observations. In order to increase the efficiency of the exploratory analysis we load these data sets separately so computation can be done on individual data sets.

The data was extracted from ncbi and in our extraction process we conditioned on the *Collection.date* variable in the data collection process to extract information that was collected from 2013 to 2022. The objective of this analysis, at least at baseline is to focus on food borne illness in the United States. Aside from conditioning on *Collection.date* the only other condition we imposed was on the illness type, the three of which are discussed above.

In initial analysis we noted some preprocessing was required. As noted in the objectives in the **Introduction** we are interested in examining data regarding the United States the first R scripts (*illness-name*_LocationClean.R) are used to subset the three illnesses to only include only locations that are in the United States using the %**like**% command, as the extracted data contained information from many countries, and these subsetted data sets are extracted. After running this script we noted an extreme level of heterogeneity($>$ 6,000 levels in the Salmonella data) in the *Isolation.Source* variable, which required preprocessing prior to conducting an exploratory analysis. The second R script (source_group.R) uses the **tolower()** command to make the levels all lowercase and therefore easier to work with. With the data in a more manageable format we used the %**like**% command to find specified levels and place them into more general groups for our analysis, e.g. grouping levels that contain apple into a produce category. After the data is extracted from the secnd R script, we noted that the *Collection.date* variable, the primary time metric of interest, had different reporting practices, which we correct for in our third R script (Clean_dates.R). This script uses the **tidyvserse**, **anytime** and **lubridate** packages to create data sets with the different different date formats (full date, year and month, just year) and decomposes this *Collection.date* into 3 new variables *year*, *month*, and *day*, so that it is easier to work with this date information. These cleaned data sets and R scripts the ones used for this analysis, and are publicly available on our Github page. As we further refine our methods for analysis we will update this section with other relevant packages and their purpose.

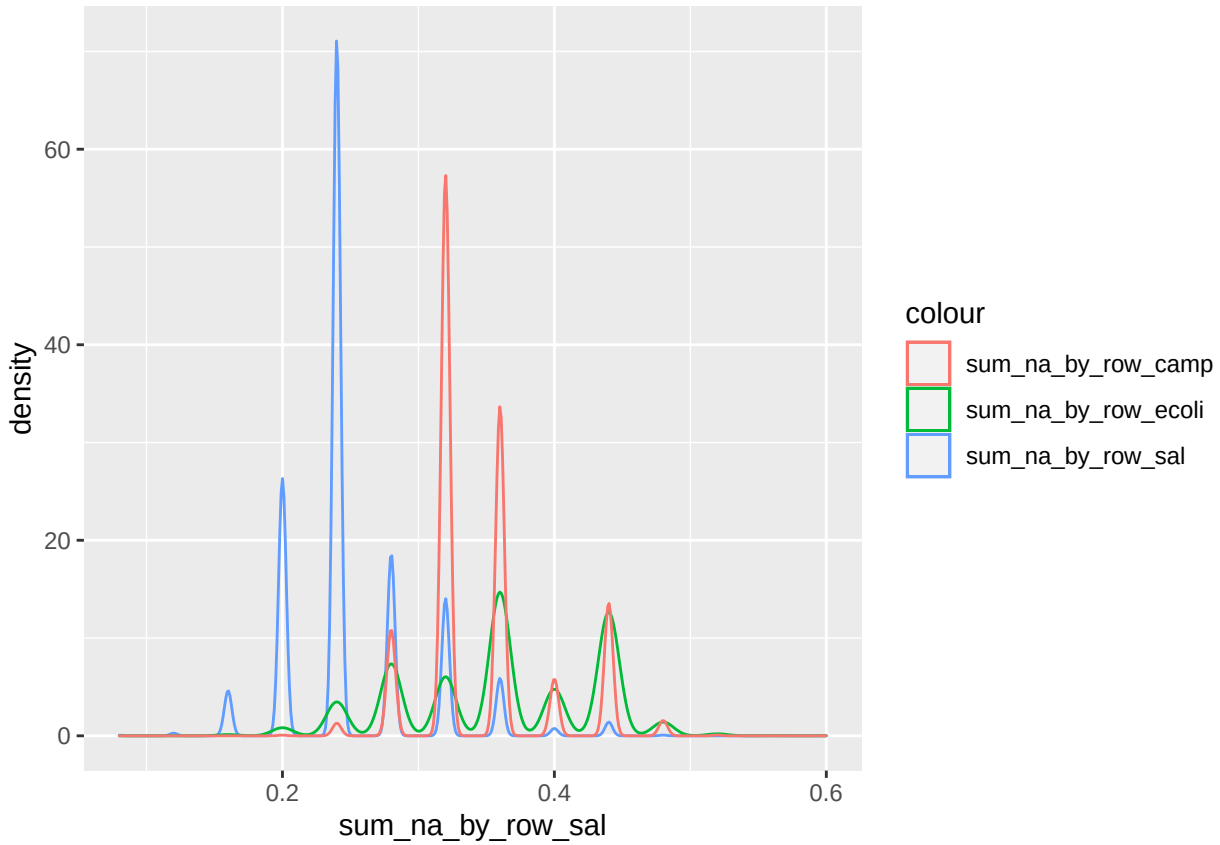
Exploratory Data Analysis

Missing Data

The NCBI collects data from surveillance and research efforts that are currently ongoing that look at a multitude of sources such as food, patients, production, etc. After data is submitted it is clustered to related pathogens, allowing for people to look for closely related pathogens. Given that the data is being uploaded from multiple sources with what we presume to be free text fields in some columns eg. Isolation source there is a lot of heterogeneity in reporting and in data quality. Data quality aside we are interested in the missingness, which we explore in table and figure 1 below.

Table 1: Percent missing for each illness dataset

	Salmonella	E. Coli	Campylobacter
Serovar	0.22	0.87	1.00
Host.disease	0.99	0.90	1.00
Isolation.source	0.11	0.15	0.04
Lat.Lon	0.98	0.81	1.00
Source.type	1.00	1.00	1.00
SNP.cluster	0.06	0.47	0.15
Min.same	0.06	0.53	0.16
Min.diff	0.17	0.84	0.37
Assembly	0.11	0.07	0.01
Outbreak	1.00	1.00	1.00
day	0.85	0.67	0.98
month	0.69	0.54	0.84



We note from table 1 that there are some variables that are missing in great quantities, such as *Serovar*, *Host_disease*, *Lat.Lon*, *Source_type*, *Computed.types*, and *outbreak* given the extreme missingness for the variables in most of our illness types we decide to remove them from our analysis. On the contrary we have no missingness in the *organism group*, *Isolate*, *Create.date*, *Collection.date*, *Location*, *Biosample*, and the variables we created in the preprocessing *Isolation.source.category* and *year*. Though we observe that there are extremes there are variables with differing levels of missingness, which is likely due to the fact that this data is reported by the researcher. The *Strain.name*,*Isolation.Source*, and *Isolation.type* is missing in very small quantity which could be due to an oversight by the researcher when submitting their data. It appears that *SNP.cluster* is missing for almost half of the E.coli data, and has some missingness in the other illnesses, but not nearly to this extent. This could be due to issues pr delays in processing the pathogens related to E. Coli. It is also apparent that the month and year variable that we created for this analysis have a lot of missingness, which can be explained by the issues in *Collection.date* that we observed. Generally a lot of variables are dependent upon reporting practices of the people submitting to the pathogen detection database, which can explain why some of these variables have a lot of missing data.

Variables of interest

Prior to proceeding to further analysis we will further discuss the variables aviable and if they will remain in our analysis. Recall that the goal of this project is to predict outbreaks of these food borne illness and explore their seasonality. This means that our main interest lays in the *Collection.Date* and *Location* as we want to observe if there are temporal or spatial trends regarding the illnesses. On top of this we consider that information regarding *strain*, *Isolation.Source* and *Isolation.type* allow for us to look for potential causes of the outbreak. Also looking for similarities between strains we can look at the *SNP.Cluster* or the *Min.Same* and *Min.Diff* variables, and since there is approximately half of the data missing in the *SNP.Cluster* we will assess relative closeness with the *Min.Same* and *Min.diff* variables. Given that we do not specifically look at antibiotic resistance we do not look at *AMR.Genotypes* and we do not stand to gain a lot from looking at *X.Organism.group* or *Assembly* so we do not consider these. We will use the variables of interest to see if we

can extract any noticeable patterns in the illnesses.

Trends over Time and Location

Given that we are interested in outbreaks we find it most pertinent to first understand where and when cases are poping up. This will, hopefully, afford us the opportunity to see historical trends in the data and if cases are more concentrated in some regions of the United States, which will help inform our decisions regarding modeling in future work.

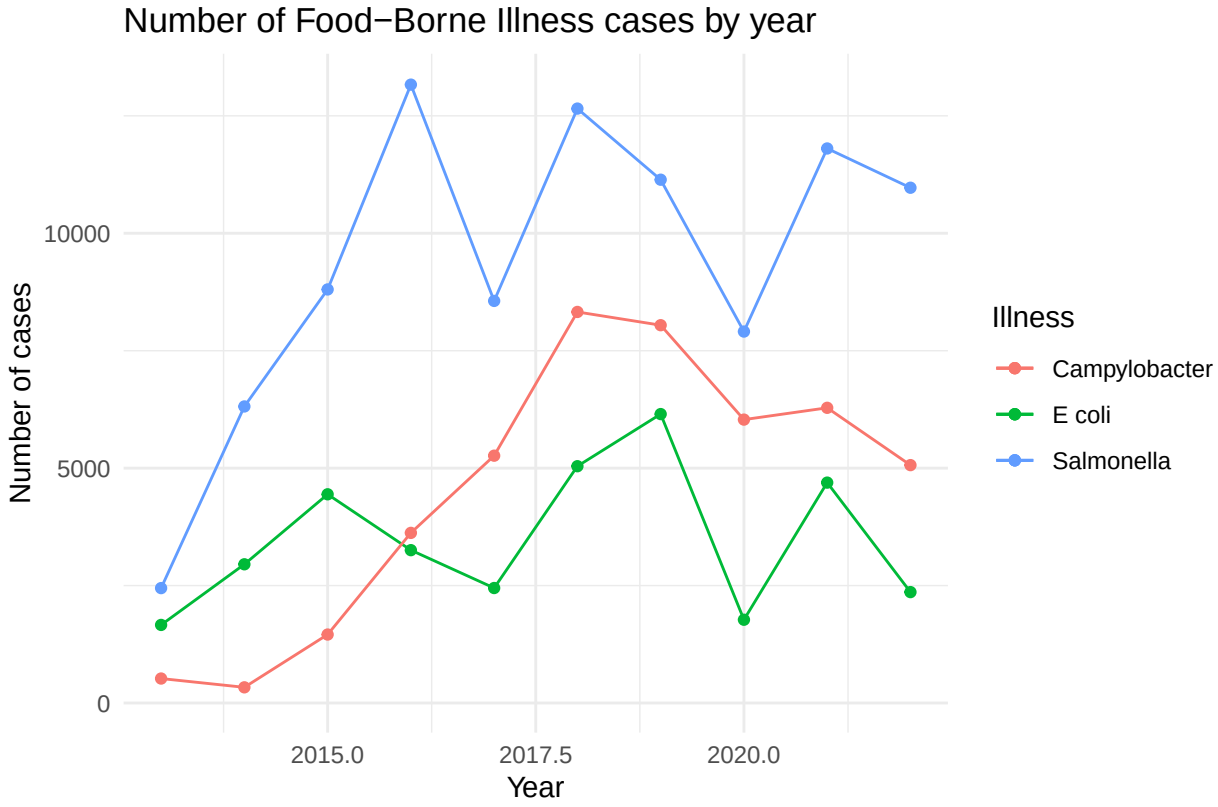


Figure 1: Number of cases per year, all illnesses

From Figure 2 we note that there are more cases of Salmonella than any other food borne illnesses in any given year. The number of cases of campylobacter is the smallest at onset, but the number of cases of camylobacter passes E.Coli in 2016 and remains larger than the number of E. Coli cases for remaining years. We observe that for all of the illnesses that the number of cases decrease in 2020, which we note is the year the Covid-19 pandemic started, though we cannot confirm a causal relationship with the provided data we can assume that this is due to less reporting or less going out to eat and getting exposure to food that could make people ill, though this is conjecture and we cannot further comment on this.

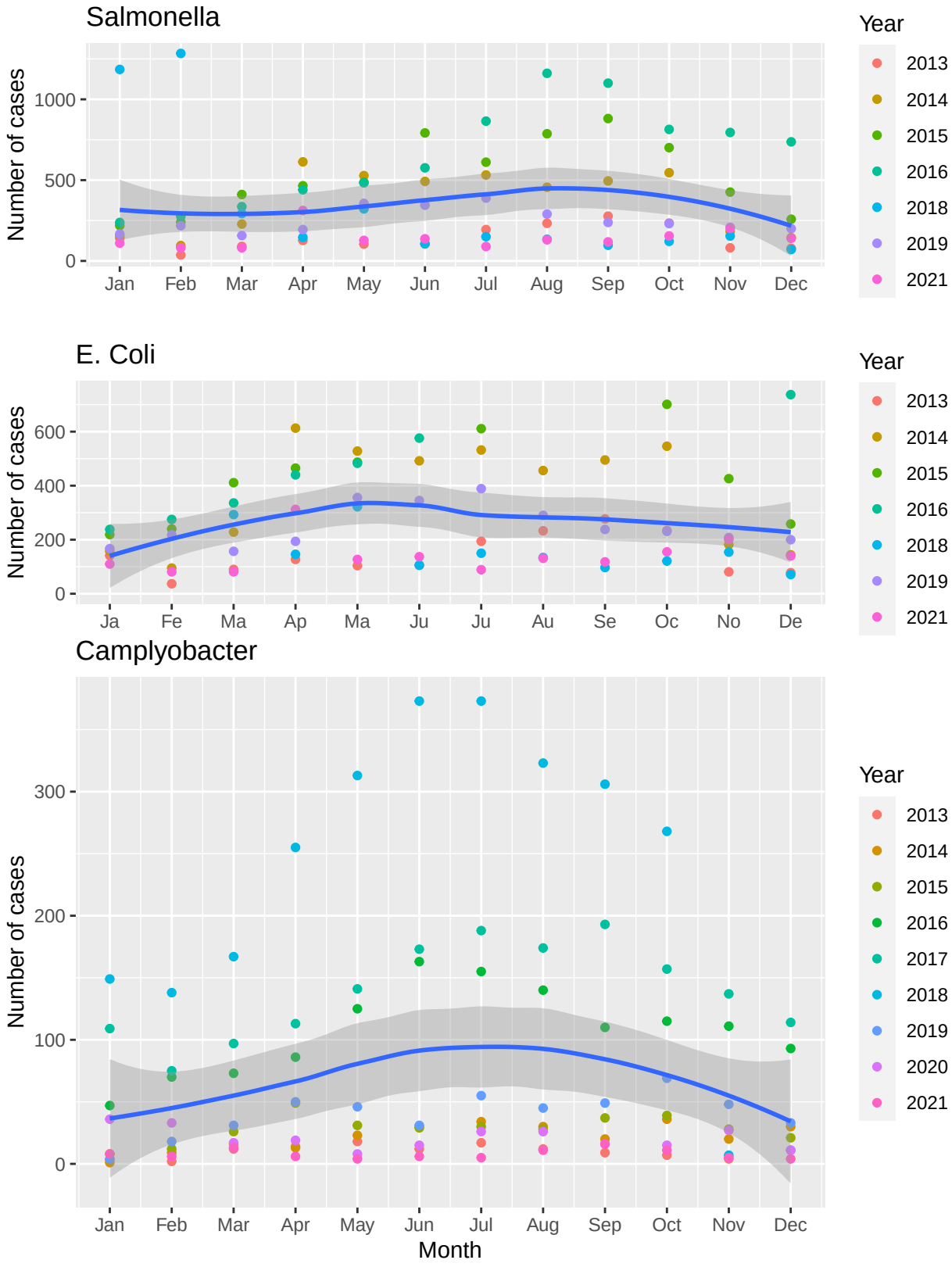


Figure 2: Number of cases per month, by Illness type

We further expand upon our analysis of number of cases over by looking at the number of cases per month

for each given illness. For the illnesses, there are no month level data that is observed for any of the illnesses, and there is no month level data for the year 2020 for the Salmonella and E. Coli data. Looking at figure X, it can be observed from the smoothed regression line fit to the data that there are some potential trends in the seasonality, noteably the number of slamonella cases looks to increase in early summer before hitting a peak around august and september before falling in the fall months. E. Coli on the other hand deems to start increasing in the winter months and peak around May/June before generally starting to decrease. Out of all the illnesses Campylobacter shows the most distinict trend in seasonality, with the number of cases increasing from the winter and spring to a peak in the summer around July and starting to decrease in the later summer/fall into the winter. We note that as mentioned in the EDA there is a high level of missingness in the month of the collection date, which limits our analysis, though we do see

Table 2: Percent of each years total cases by region of US by Illness

year	USA, General	Midwest	Northeast	South	West
Salomella					
2013	0.13	0.19	0.13	0.32	0.24
2014	0.19	0.14	0.09	0.29	0.29
2015	0.31	0.22	0.1	0.23	0.14
2016	0.48	0.18	0.04	0.21	0.09
2017	0.08	0.27	0.07	0.45	0.13
2018	0.18	0.23	0.09	0.35	0.14
2019	0.03	0.31	0.11	0.41	0.14
2020	0.13	0.27	0.1	0.36	0.13
2021	0.27	0.23	0.08	0.31	0.11
2022	0.37	0.15	0.06	0.37	0.06
E. Coli					
2013	0.32	0.15	0.07	0.17	0.29
2014	0.27	0.23	0.07	0.3	0.13
2015	0.46	0.04	0.11	0.32	0.07
2016	0.3	0.08	0.11	0.16	0.34
2017	0.25	0.23	0.14	0.27	0.11
2018	0.25	0.25	0.12	0.23	0.15
2019	0.07	0.25	0.09	0.37	0.22
2020	0.14	0.25	0.13	0.35	0.12
2021	0.31	0.3	0.07	0.22	0.1
2022	0.57	0.12	0.11	0.17	0.03
Campylobacter					
2013	0.12	0.29	0.11	0.18	0.3
2014	0.52	0.16	0.02	0.04	0.26
2015	0.13	0.19	0.11	0.29	0.28
2016	0.29	0.19	0.09	0.24	0.2
2017	0.29	0.2	0.09	0.25	0.17
2018	0.34	0.16	0.07	0.33	0.11
2019	0.03	0.2	0.09	0.54	0.14
2020	0.02	0.16	0.09	0.61	0.12
2021	0.09	0.15	0.09	0.56	0.11
2022	0.14	0.15	0.1	0.52	0.09

From table X above we note that for all the illnesses that the percent of total cases by region is generally variable, over time. Though we can see that for Salmonella and Camylobacter that the Southern region of

the United States generally has the highest percent of cases in any given year, especially in more recent years. While the Northeast generally has the lowest percentage of cases across all of the illnesses. In E.Coli it appears that in earlier years there is generally a higher percent of the cases in the Western Region of the United States, though in later years it appears that the percent of cases in the US lowers for the West, and starts to increase for the Southern US. Thus it appears that when looking into outbreak prevention it may be beneficial to consider focusing on the South. Though we do concede that these are raw percentages that fail to account for population density in these regions.

Sources of outbreaks

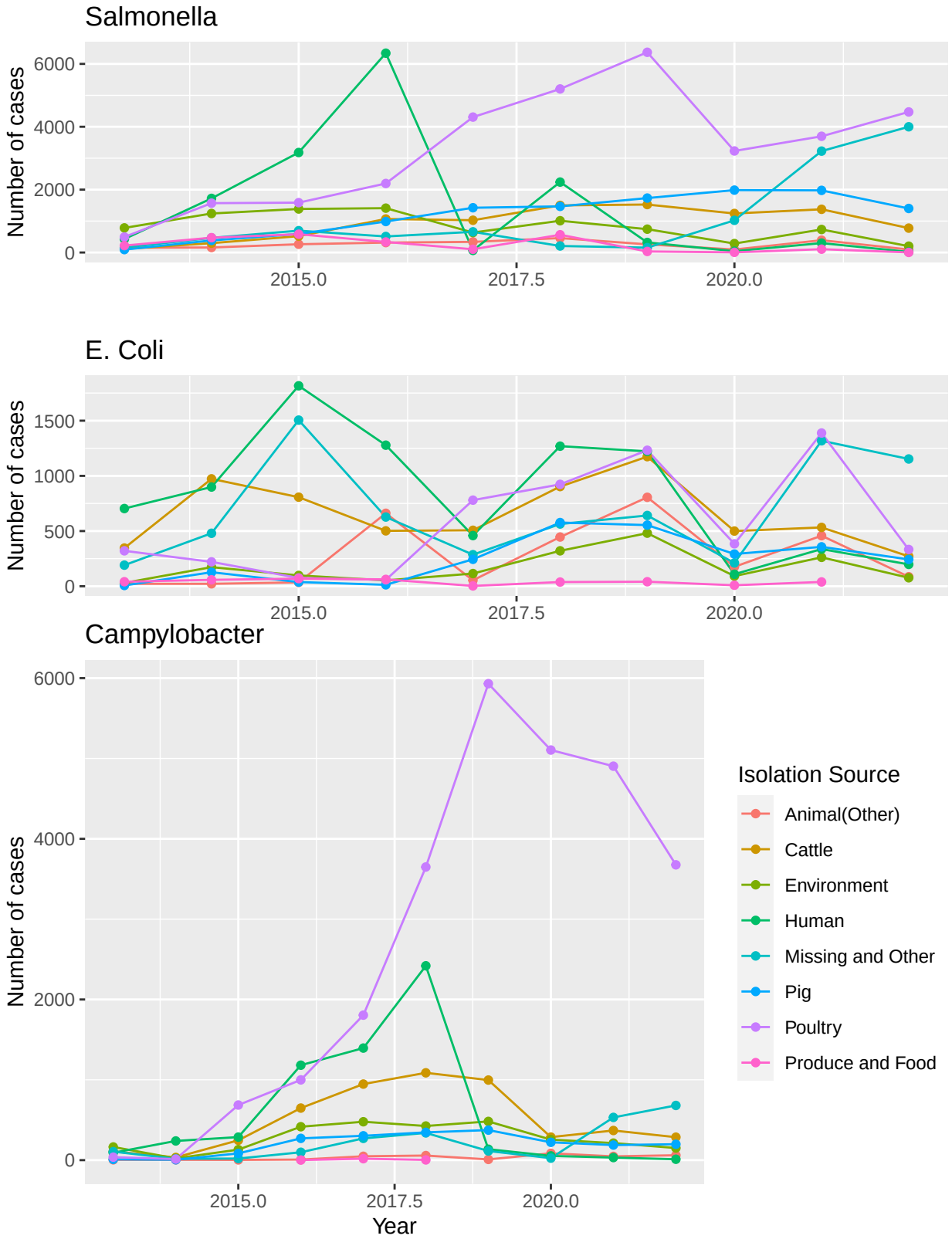


Figure 3: Number of cases by Isolation Source

We note that looking at the trends over time the Isolation source that generally for all 3 of the illnesses that in later years poultry is one of (if not the biggest) causes of the illnesses we have included. Though we do note that the number of Isolation from humans is quite large for all illnesses as well. Other sources such as produce and pig based sources seem to be relatively low across all 3 illness, which could indicate that in our modeling that we may want to focus on sources that have higher incidence for our illnesses over time as it may be easier to capture an outbreak this way.

Table 3: Percentanges regarding Closely related SNPs by Illness

Minimum Same	Percent
Salmonella	
Missing	0.06
Closely related (within 7 SNPs)	0.67
Not closely related	0.27
Ecoli	
Missing	0.53
Closely related (within 7 SNPs)	0.29
Not closely related	0.18
Campylobacter	
Missing	0.16
Closely related (within 7 SNPs)	0.6
Not closely related	0.24

We recall from the variable descriptions that this variable tells us the closeness to other strains that were isolated in the same source (clinical or environmental). As we recall from our discussion with Subject matter expert Dr. Julian if isolates are within 7 SNPs they are closely related to each other (cite lecture). Since we want to look at outbreaks we assume that these would be derived of isolates that are close to others genetically and if they are popping up in similar regions to see if there are any regions/times that pop up in the data. Given the level of missingness in Date information that we will consider the trends over month, though it would be more beneficial to look at trends by week the reporting quality on *Collection.date* is very poor, and as we can observe, data pertaining to the day is mostly missing, so we will look at specimens that were collected by month. Given that the data is longitudinal in nature over 9 years with many regions, so we want to look at very closely related strains to see if there are concentrations of these cases in a region. Due to the large number of combinations of months years and regions we look for instances where there were more than 100 cases of an illness that are on average the *Min.same* is less than 3 SNPs from the others.

We note that the data regarding Salmonella has the most instances (13) where over 100 individuals in a region contract similar strains of the illness in a month, while there are also some instances of concerning outbreaks with the E. Coli (7), we observe that when we look across the years and months for camylobacter none of these instances meet our threshold for inclusion as a large outbreak with the month having the most cases of very cloesly related campylobacter across our studies is Spetemeber 2018 when there were 43 cloesly related cases of campylobacter in the Southern United States. Looking at table X, we observe that for samonella that in 2015 the 3 summer months we had multiple cases of highly related strains in the Southern United States, which are likely linked, simiary we observe that in March and febuaray of 2018 there is a large amount of closely related cases in the Midwest. No appearnt trends pertaining to a certain region stick out in the E. coli data.

```
knitr::opts_chunk$set(echo = FALSE)

#~~~~~~~~~~~~#
## Packages ##
#~~~~~~~~~~~~#
```

Table 4: Average (Standard Deviation) Min.same for closely related strains when more than 100 cases occur in a region, Salmonella

Year	Month	Number of cases	Midwest	South	West	Northeast
2013	Oct	119	1.2 (2.81)			
2014	Jun	117	1.4 (2.15)			
2014	Oct	254	2.83 (5.53)			
2015	Aug	240		2.58 (4.78)		
2015	Jul	118		2.14 (5.17)		
2015	Jun	247		2.41 (5.2)		
2015	Nov	139	2.41 (4.48)			
2018	Feb	111	2.07 (5.92)			
2018	Mar	154	1.49 (5.44)			
2018	May	176			1.63 (4.83)	
2019	Jun	183				2.64 (3.89)
2021	Apr	219	0.85 (2.74)			
2021	Nov	113		2.56 (5.7)		

Table 5: Average (Standard Deviation) Min.same for closely related strains when more than 100 cases occur in a region, E. Coli

Year	Month	Number of cases	West	Midwest	South	USA
2013	Nov	130	1.71 (2.52)			
2014	Nov	117		1.03 (1.85)		
2015	Jun	310			1.54 (4.8)	
2015	Sep	133			2.28 (6.53)	
2018	Aug	114				0.83 (1.52)
2018	May	139				1.59 (1.65)
2019	Apr	198			1.78 (4.89)	

```
library(lubridate) # working with dates
library(tidyverse) # working with data
library(kableExtra) # Make nicer plots
library(ggpubr) # Arrange GGplots

#~~~~~~~~#
## Data ##
#~~~~~~~~#

# Set the working directory and read in the preprocessed data
setwd("~/Desktop/Semester_3/Practical/Final")
salmonella <- read.csv("final_salmonella_date.csv")
ecoli <- read.csv("final_ecoli_date.csv")
campylobacter <- read.csv("final_Campylobacter_date.csv")

# Suppress summaries info when knitting to a pdf
options(dplyr.summarise.inform = FALSE)
library(dplyr, warn.conflicts = FALSE)
options(tidyverse.quiet = TRUE)
```

```r
#~~~~~~~~~~~~~~~~#
## Missing Data ##
#~~~~~~~~~~~~~~~~#

# The first objective of this EDA is to explore the missing patterns in our
# data at the surface level


# Replace Missing strings with NA
salmonella[salmonella == ""] <- NA
ecoli[ecoli == ""] <- NA
campylobacter[campylobacter == ""] <- NA


## Missingness by column ##

# Salmonella
na_by_cols_sal <- as.data.frame(apply(salmonella,2, is.na))
sum_na_by_col_sal <- apply(na_by_cols_sal,2,sum)/nrow(na_by_cols_sal)
sum_na_by_col_sal <- as.data.frame(round(sum_na_by_col_sal,2))

# E. Coli
na_by_cols_ecoli <- as.data.frame(apply(ecoli,2, is.na))
sum_na_by_col_ecoli <- apply(na_by_cols_ecoli,2,sum)/nrow(na_by_cols_ecoli)
sum_na_by_col_ecoli <- as.data.frame(round(sum_na_by_col_ecoli,2))

# Camplyobaceter
na_by_cols_camp <- as.data.frame(apply(campylobacter,2, is.na))
sum_na_by_col_camp <- apply(na_by_cols_camp,2,sum)/nrow(na_by_cols_camp)
sum_na_by_col_camp <- as.data.frame(round(sum_na_by_col_camp,2))

# Get the data together
Missing_by_col <- cbind(sum_na_by_col_sal, sum_na_by_col_ecoli, sum_na_by_col_camp)
names(Missing_by_col) <- c("x", "y", "z") # Create a dummy variable for filtering
Missing_by_col <- Missing_by_col %>% filter(x != 0 &
                                            y != 0 &
                                            z != 0   ) # Filter out all variables with no missing data
names(Missing_by_col) <- c("Salmonella", "E. Coli", "Campylobacter")

# Print the data frame
Missing_by_col  %>%
  kbl(caption = "Percent missing for each illness dataset", booktabs=T,    escape=F, align = "c") %>%
  kable_styling(full_width = FALSE, latex_options = c('hold_position'))



## Missingness by row

# Using the commented out code below we check to see if any of the data is
# complete for a given observation

#sum(complete.cases(salmonella))
```

```r
#sum(complete.cases(ecoli))
#sum(complete.cases(campylobacter))

# There are no complete cases in the data set

# Salmonella
na_by_row_sal <- as.data.frame(apply(salmonella,1, is.na))
sum_na_by_row_sal <- apply(na_by_row_sal,2,sum)/nrow(na_by_row_sal)
salmonella_missing <- as.data.frame(sum_na_by_row_sal)

# E. Coli
na_by_row_ecoli <- as.data.frame(apply(ecoli,1, is.na))
sum_na_by_row_ecoli <- apply(na_by_row_ecoli,2,sum)/nrow(na_by_row_ecoli)
ecoli_missing <- as.data.frame(sum_na_by_row_ecoli)

# Campylobacter
na_by_row_camp <- as.data.frame(apply(campylobacter,1, is.na))
sum_na_by_row_camp <- apply(na_by_row_camp,2,sum)/nrow(na_by_row_camp)
campylobacter_missing <- as.data.frame(sum_na_by_row_camp)

# Make a plot of missingness by observation, This did not make it into the final EDA
ggplot(data = salmonella_missing, aes(x = sum_na_by_row_sal, color = "sum_na_by_row_sal")) +
  geom_density(alpha = .9) +
  geom_density(data = ecoli_missing, aes(x = sum_na_by_row_ecoli,
                                         color = "sum_na_by_row_ecoli"), alpha = .9) +
  geom_density(data = campylobacter_missing, aes(x = sum_na_by_row_camp,
                                                 color = "sum_na_by_row_camp"), alpha = .9)


## Delete variables with most of the values missing

salmonella <- salmonella[,-which(names(salmonella) %in% c("Serovar", "Host.Disease", "Lat.Lon", "Source
ecoli <- ecoli[,-which(names(ecoli) %in% c("Serovar", "Host.Disease", "Lat.Lon", "Source.type", "Outbrea
campylobacter <- campylobacter[,-which(names(campylobacter) %in% c("Serovar", "Host.Disease", "Lat.Lon"


# Let us consider data by year in general



# How many Samonella cases per year?
Num_per_year_salmonella <- salmonella %>%
                           group_by(year) %>%
                           summarize(num = n())

# How many ecoli cases per year?
Num_per_year_ecoli <- ecoli %>%
                      group_by(year) %>%
                      summarize(num = n())

# How many ecoli cases per year?
Num_per_year_campylobacter <- campylobacter %>%
                              group_by(year) %>%
```

```
                              summarize(num = n())

# Lets look at the trend over time, and
figure1 <- ggplot(data = Num_per_year_salmonella, aes(x = year, y = num, color = "Salmonella")) +
  geom_point() +
  geom_line() +
  # E coli
  geom_point(data = Num_per_year_ecoli, aes(x = year, y = num, color = "E coli")) +
  geom_line(data = Num_per_year_ecoli, aes(x = year, y = num, color = "E coli")) +
  # campylobacter
  geom_point(data = Num_per_year_campylobacter, aes(x = year, y = num, color = "Campylobacter")) +
  geom_line(data = Num_per_year_campylobacter, aes(x = year, y = num, color = "Campylobacter")) +
  theme_minimal() +
  labs(title= "Number of Food-Borne Illness cases by year",
       x = "Year",
       y = "Number of cases")  +
  scale_color_discrete(name = "Illness")

# Print the figure out
annotate_figure(figure1,
                bottom = text_grob("Figure 1: Number of cases per year, all illnesses", color = "black


## Monthly cases over all years ##

# Salmonella

# Cases per month
Num_per_month_sal <- salmonella %>%
                     group_by(year, month) %>%
                     summarize(num =  n())
Num_per_month_sal <- Num_per_month_sal %>% filter(!is.na(month))

# Plot the number of cases per month
plot_by_month_sal <- ggplot(data = Num_per_month_sal, aes(x = as.numeric(month), y = num)) +
                      geom_point(aes(color = as.factor(year))) +
                      scale_x_continuous(breaks=seq(1,12,1), labels=c("Jan", "Feb",
                                                                      "Mar", "Apr", "May",
                                                                      "Jun", "Jul", "Aug",
                                                                      "Sep", "Oct",
                                                                      "Nov", "Dec"))+
                      geom_smooth() +
                      labs(title= "Salmonella",
                           x = "",
                           y = "Number of cases")  +
                      scale_color_discrete(name = "Year")



# E. Coli
Num_per_month_ecoli <- ecoli %>%
                     group_by(year, month) %>%
                     summarize(num =  n())
```

```r
Num_per_month_ecoli <- Num_per_month_ecoli %>% filter(!is.na(month))

# Plot the number of cases per month
plot_by_month_eco <- ggplot(data = Num_per_month_sal, aes(x = as.numeric(month), y = num)) +
                      geom_point(aes(color = as.factor(year))) +
                      scale_x_continuous(breaks=seq(1,12,1), labels=c("Ja", "Fe",
                                                                      "Ma", "Ap", "Ma",
                                                                      "Ju", "Ju", "Au",
                                                                      "Se", "Oc",
                                                                      "No", "De"))+
                      geom_smooth() +
                      labs(title= "E. Coli",
                           x = "",
                           y = "Number of cases")  +
                      scale_color_discrete(name = "Year") +
                      ylim(0,750)



# Camplyobacter
Num_per_month_camp <- campylobacter %>%
                     group_by(year, month) %>%
                     summarize(num =  n())
Num_per_month_camp <- Num_per_month_camp %>% filter(!is.na(month))

# Plot
plot_by_month_cal <- ggplot(data = Num_per_month_camp, aes(x = as.numeric(month), y = num)) +
                      geom_point(aes(color = as.factor(year))) +
                      scale_x_continuous(breaks=seq(1,12,1), labels=c("Jan", "Feb",
                                                "Mar", "Apr", "May",
                                                "Jun", "Jul", "Aug",
                                               "Sep", "Oct",
                                               "Nov", "Dec"))+
                      geom_smooth() +
                      labs(title= "Camplyobacter",
                           x = "Month",
                           y = "Number of cases")  +
                      scale_color_discrete(name = "Year")


# Arrange the plots and annotate them
figure2 <- ggarrange(plot_by_month_sal, plot_by_month_eco, nrow =2 )
figure2
annotate_figure(plot_by_month_cal,
                bottom = text_grob("Figure 2: Number of cases per month, by Illness type", color = "bla



# We can also look to trends over regions in the US


## Define the regions in all the data sets
```

```
salmonella <- salmonella %>% mutate(region = case_when(Location %in% c("ME", "VT", "NH",
                                                                       "MA", "CT", "RI",
                                                                       "NY", "NJ", "PA",
                                                                       "Northeast") ~ "Northeast",
                                                       Location %in% c("DE", "MD", "DC",
                                                                       "WV", "VA", "NC",
                                                                       "SC", "GA", "FL",
                                                                       "AL", "MS", "LA",
                                                                       "TX", "OK", "AR",
                                                                       "TN", "KY",
                                                                       "South") ~ "South",
                                                       Location %in% c("OH", "MI", "IN",
                                                                       "IL", "WI", "MN",
                                                                       "IA", "MO", "KS",
                                                                       "NE", "SD", "ND",
                                                                       "Midwest") ~ "Midwest",
                                                       Location %in% c("NM", "CO", "WY",
                                                                       "MT", "ID", "UT",
                                                                       "AZ", "NV", "WA",
                                                                       "OR", "CA", "HI",
                                                                       "AK", "West",
                                                                       "Western Region") ~ "West",
                                                       Location %in% c("USA", "PR", "GU") ~ "USA, Genera


ecoli <- ecoli %>% mutate(region = case_when(Location %in% c("ME", "VT", "NH",
                                                             "MA", "CT", "RI",
                                                             "NY", "NJ", "PA",
                                                             "Northeast") ~ "Northeast",
                                                 Location %in% c("DE", "MD", "DC",
                                                                 "WV", "VA", "NC",
                                                                 "SC", "GA", "FL",
                                                                 "AL", "MS", "LA",
                                                                 "TX", "OK", "AR",
                                                                 "TN", "KY",
                                                                 "South") ~ "South",
                                                 Location %in% c("OH", "MI", "IN",
                                                                 "IL", "WI", "MN",
                                                                 "IA", "MO", "KS",
                                                                 "NE", "SD", "ND",
                                                                 "Midwest") ~ "Midwest",
                                                 Location %in% c("NM", "CO", "WY",
                                                                 "MT", "ID", "UT",
                                                                 "AZ", "NV", "WA",
                                                                 "OR", "CA", "HI",
                                                                 "AK", "West",
                                                                 "Western Region") ~ "West",
                                                 Location %in% c("USA", "PR", "GU") ~ "USA, Genera

campylobacter <- campylobacter %>% mutate(region = case_when(Location %in% c("ME", "VT", "NH",
                                                                             "MA", "CT", "RI",
                                                                             "NY", "NJ", "PA",
                                                                             "Northeast") ~ "Northeast",
```

```
                                                    Location %in% c("DE", "MD", "DC",
                                                                    "WV", "VA", "NC",
                                                                    "SC", "GA", "FL",
                                                                    "AL", "MS", "LA",
                                                                    "TX", "OK", "AR",
                                                                    "TN", "KY",
                                                                    "South") ~ "South",
                                                    Location %in% c("OH", "MI", "IN",
                                                                    "IL", "WI", "MN",
                                                                    "IA", "MO", "KS",
                                                                    "NE", "SD", "ND",
                                                                    "Midwest") ~ "Midwest",
                                                    Location %in% c("NM", "CO", "WY",
                                                                    "MT", "ID", "UT",
                                                                    "AZ", "NV", "WA",
                                                                    "OR", "CA", "HI",
                                                                    "AK", "West",
                                                                    "Western Region") ~ "West",
                                                    Location %in% c("USA", "PR", "GU") ~ "USA, Genera

# Now we will try to look at the proportion of cases in each region in each  year

# Salmonella
cases_by_region_sal <- salmonella %>% filter(!is.na(year)) %>%
                                      group_by(region, year) %>%
                                      summarize(num = n()) %>%
                                      ungroup() %>%
                                      group_by(year) %>%
                                      summarize(region = region, percent = round(num/sum(num),2)) %>%
                                      pivot_wider(names_from = region, values_from = percent)



# E. Coli

cases_by_region_ecoli <- ecoli %>% filter(!is.na(year)) %>%
                                   group_by(region, year) %>%
                                   summarize(num = n()) %>%
                                   ungroup() %>%
                                   group_by(year) %>%
                                   summarize(region = region, percent = round(num/sum(num),2)) %>%
                                   pivot_wider(names_from = region, values_from = percent)



# campylobacter

cases_by_region_camp <- campylobacter %>% filter(!is.na(year)) %>%
                                          group_by(region, year) %>%
                                          summarize(num = n()) %>%
                                          ungroup() %>%
                                          group_by(year) %>%
                                          summarize(region = region, percent = round(num/sum(num),2)) %
```

```r
## Get the data into a table
cases_region <- as.data.frame(rbind(cases_by_region_sal, cases_by_region_ecoli, cases_by_region_camp))
cases_region[31:33, ] <- ""
cases_region <- apply(cases_region, 2, as.character)
cases_region[31, 1] <- "Salomella"
cases_region[32, 1] <- "E. Coli"
cases_region[33, 1] <- "Campylobacter"

# Rearrange the data frame
cases_region <- cases_region[c(31,1:10,32,11:20,33,21:30), c(1,5,2,3,4,6)]

# Print the data to a kable table
cases_region %>%
  kbl(caption = "Percent of each years total cases by region of US by Illness", booktabs=T,    escape=F
  kable_styling(full_width = FALSE, latex_options = c('hold_position'))


## We now want to look at our Isolation sources,  for each of the Illnesses by year

# Salmonella #

# Get the data for the data
plotdata_sal <- salmonella %>% group_by(year, Isolation.source.category) %>%
                               summarize(num = n())

# Create a plot
plot5 <- ggplot(data = plotdata_sal, aes(x = year, y = num, color = Isolation.source.category)) +
          geom_line() +
          geom_point() +
          theme(legend.position = "none") +
          labs(title= "Salmonella",
                x = "",
                y = "Number of cases")


# E. Coli #

# Get the data
plotdata_ecoli <- ecoli %>% group_by(year, Isolation.source.category) %>%
                            summarize(num = n())

# Plot the data
plot6 <- ggplot(data = plotdata_ecoli, aes(x = year, y = num, color = Isolation.source.category)) +
          geom_line() +
          geom_point() +
          theme(legend.position = "none") +
          labs(title= "E. Coli",
                 x = "",
                 y = "Number of cases")  +
          scale_color_discrete(name = "Illness source")


# Campylobacter #
```

```r
# Get the data for the plot
plotdata_camp <- campylobacter %>% group_by(year, Isolation.source.category) %>%
                                   summarize(num = n())

# Create the plot
plot7 <- ggplot(data = plotdata_camp, aes(x = year, y = num, color = Isolation.source.category)) +
          geom_line() +
          geom_point() +
          labs(title= "Campylobacter",
                          x = "Year",
                          y = "Number of cases") +
          scale_color_discrete(name = "Isolation Source")

# Arrange and annotate the plots
figure3 <- ggarrange(plot5, plot6, nrow =2)
figure3
annotate_figure(plot7,
                bottom = text_grob("Figure 3: Number of cases by Isolation Source", color = "black",


# Lets look at Min same as a metric for similarity and defining an outbreak
# Generally all of the codes below are to find the percent of missing data, % which
# are closely related (within 7 SNPs) and those that are distant


##  Salmonella ##

# Get percents for each of the 3 categories, missing, closely related, not closely related
Missing_min_same <- length(which(is.na(salmonella$Min.same)))/nrow(salmonella)
close_min_same <- length(which(salmonella$Min.same <= 7))/nrow(salmonella)
far_min_same <- length(which(salmonella$Min.same > 7))/nrow(salmonella)

# Make the data frame and reaarange
min_same_salmonella <- data.frame()
min_same_salmonella[1,1] <- "Missing"
min_same_salmonella[1,2] <- round(Missing_min_same,2)
min_same_salmonella[2,1] <- "Closely related (within 7 SNPs)"
min_same_salmonella[2,2] <- round(close_min_same,2)
min_same_salmonella[3,1] <- "Not closely related"
min_same_salmonella[3,2] <- round(far_min_same,2)
names(min_same_salmonella) <- c("Minimum Same", "Percent")


# E. Coli ##

# Get percents for each of the 3 categories, missing, closely related, not closely related
Missing_min_same <- length(which(is.na(ecoli$Min.same)))/nrow(ecoli)
close_min_same <- length(which(ecoli$Min.same <= 7))/nrow(ecoli)
far_min_same <- length(which(ecoli$Min.same > 7))/nrow(ecoli)

# Make the data frame and reaarange
min_same_ecoli <- data.frame()
min_same_ecoli[1,1] <- "Missing"
```

```r
min_same_ecoli[1,2] <- round(Missing_min_same,2)
min_same_ecoli[2,1] <- "Closely related (within 7 SNPs)"
min_same_ecoli[2,2] <- round(close_min_same,2)
min_same_ecoli[3,1] <- "Not closely related"
min_same_ecoli[3,2] <- round(far_min_same,2)
names(min_same_ecoli) <- c("Minimum Same", "Percent")



## Campylobacter ##

# Get percents for each of the 3 categories, missing, closely related, not closely related
Missing_min_same <- length(which(is.na(campylobacter$Min.same)))/nrow(campylobacter)
close_min_same <- length(which(campylobacter$Min.same <= 7))/nrow(campylobacter)
far_min_same <- length(which(campylobacter$Min.same > 7))/nrow(campylobacter)

# Make the data frame and reaarange
min_same_campylobacter <- data.frame()
min_same_campylobacter[1,1] <- "Missing"
min_same_campylobacter[1,2] <- round(Missing_min_same,2)
min_same_campylobacter[2,1] <- "Closely related (within 7 SNPs)"
min_same_campylobacter[2,2] <- round(close_min_same,2)
min_same_campylobacter[3,1] <- "Not closely related"
min_same_campylobacter[3,2] <- round(far_min_same,2)
names(min_same_campylobacter) <- c("Minimum Same", "Percent")

## Create a data frame with all the data in a nice format, which requires rearranging

Minimum_same_illnesses <- as.data.frame(rbind(min_same_salmonella, min_same_ecoli, min_same_campylobacte
Minimum_same_illnesses[10:12,] <- ""
Minimum_same_illnesses <- apply(Minimum_same_illnesses, 2, as.character)
Minimum_same_illnesses[10,1] <- "Salmonella"
Minimum_same_illnesses[11,1] <- "Ecoli"
Minimum_same_illnesses[12,1] <- "Campylobacter"
Minimum_same_illnesses <- Minimum_same_illnesses[c(10,1:3,11,4:6,12,7:9),]

# print  to a data frame
Minimum_same_illnesses %>%
  kbl(caption = "Percentanges regarding Closely related SNPs by Illness", booktabs=T,    escape=F, alig
  kable_styling(full_width = FALSE, latex_options = c('hold_position'))



## We can now look at what we define to be outbreaks, in which we find there are more than 100 people
## in a given month that have a closely related strain of an illness

# Get the month name into a readable form, we only need this for Salmonella and E. Coli as the
# results are not worthy of a table for Campylobacter

# Change the months #
salmonella <- salmonella %>% mutate(month1 = case_when(month ==1 ~ "Jan",
                                                       month ==2 ~ "Feb",
```

```
                                                    month ==3 ~ "Mar",
                                                    month ==4 ~ "Apr",
                                                    month ==5 ~ "May",
                                                    month ==6 ~ "Jun",
                                                    month ==7 ~ "Jul",
                                                    month ==8 ~ "Aug",
                                                    month ==9 ~ "Sep",
                                                    month ==10 ~ "Oct",
                                                    month ==11 ~ "Nov",
                                                    month ==12 ~ "Dec"))

ecoli <- ecoli %>% mutate(month1 = case_when(month ==1 ~ "Jan",
                                             month ==2 ~ "Feb",
                                             month ==3 ~ "Mar",
                                             month ==4 ~ "Apr",
                                             month ==5 ~ "May",
                                             month ==6 ~ "Jun",
                                             month ==7 ~ "Jul",
                                             month ==8 ~ "Aug",
                                             month ==9 ~ "Sep",
                                             month ==10 ~ "Oct",
                                             month ==11 ~ "Nov",
                                             month ==12 ~ "Dec"))



# Salmonella "outbreaks" #

# Look at the number of cases in a region in a given month/year and find how closely related they
# are on average
sal_subset <- salmonella %>%
              filter(!is.na(month1)) %>%
              group_by(year, month1, region) %>%
              summarize(mean1 = round(mean(Min.same, na.rm = T),2),
                        sd1 = round(sd(Min.same, na.rm = T),2),
                        num = n())  %>%
              ungroup() %>%
          # Substantial number of cases that are very closely related
              filter(mean1 < 3 & num > 100) %>%
              pivot_wider(names_from = region, values_from = c(mean1, sd1))

# Get the data into a data frame, copy and paste the mean and sd in  one column
sal_subset <- as.data.frame(sal_subset)
sal_subset[,4] <- paste(sal_subset[,4], "(", sal_subset[,8], ")")
sal_subset[,5] <- paste(sal_subset[,5], "(", sal_subset[,9], ")")
sal_subset[,6] <- paste(sal_subset[,6], "(", sal_subset[,10], ")")
sal_subset[,7] <- paste(sal_subset[,7], "(", sal_subset[,11], ")")
sal_subset <- sal_subset[,-c(8:11)]
sal_subset <- apply(sal_subset,2,as.character)
sal_subset <- as.data.frame(sal_subset)

# Get rid of NA  (NA) from cells in the data frame
sal_subset$mean1_Northeast <- ifelse(nchar(sal_subset$mean1_Northeast) == 9,
```

```
                                    mean1_Northeast <- "",
                                    mean1_Northeast <- sal_subset$mean1_Northeast)
sal_subset$mean1_West <- ifelse(nchar(sal_subset$mean1_West) == 9,
                                    mean1_West <- "",
                                    mean1_West <- sal_subset$mean1_West)
sal_subset$mean1_South <- ifelse(nchar(sal_subset$mean1_South) == 9,
                                    mean1_South <- "",
                                    mean1_South <- sal_subset$mean1_South)
sal_subset$mean1_Midwest <- ifelse(nchar(sal_subset$mean1_Midwest) == 9,
                                    mean1_Midwest <- "",
                                    mean1_Midwest <- sal_subset$mean1_Midwest)

names(sal_subset) <- c("Year", "Month", "Number of cases", "Midwest", "South", "West", "Northeast")

# Print to a pretty data frame
sal_subset %>%
  kbl(caption = "Average (Standard Deviation) Min.same for closely related strains when more than 100 ca
  kable_styling(full_width = FALSE, latex_options = c('hold_position'))




# Ecoli  "outbreaks"

# Look at the number of cases in a region in a given month/year and find how closely related they
# are on average
eco_subset <- ecoli %>% filter(!is.na(month1)) %>%
                    group_by(year, month1, region) %>%
                    summarize(mean1 = round(mean(Min.same, na.rm = T),2),
                              sd1 = round(sd(Min.same, na.rm = T),2),
                            num = n())  %>%
                ungroup() %>%
                # Substantial number of cases that are very closely related
                filter(mean1 < 3 & num > 100) %>%
                pivot_wider(names_from = region, values_from = c(mean1, sd1))

# Get the data into a data frame, copy and paste the mean and sd in  one column
eco_subset <- as.data.frame(eco_subset)
eco_subset[,4] <- paste(eco_subset[,4], "(", eco_subset[,8], ")")
eco_subset[,5] <- paste(eco_subset[,5], "(", eco_subset[,9], ")")
eco_subset[,6] <- paste(eco_subset[,6], "(", eco_subset[,10], ")")
eco_subset[,7] <- paste(eco_subset[,7], "(", eco_subset[,11], ")")
eco_subset <- eco_subset[,-c(8:11)]
eco_subset <- apply(eco_subset,2,as.character)

eco_subset <- as.data.frame(eco_subset)
names(eco_subset) <- c("Year", "Month", "Number of cases", "West", "Midwest", "South", "USA")


# Get rid of NA  (NA) from cells in the data frame
eco_subset$USA <- ifelse(nchar(eco_subset$USA) == 9,
                            USA <- "",
```

```
                          USA <- eco_subset$USA)
eco_subset$West <- ifelse(nchar(eco_subset$West) == 9,
                                    West <- "",
                                    West <- eco_subset$West)
eco_subset$South <- ifelse(nchar(eco_subset$South) == 9,
                                    South <- "",
                                    South <- eco_subset$South)
eco_subset$Midwest <- ifelse(nchar(eco_subset$Midwest) == 9,
                                    Midwest <- "",
                                    Midwest <- eco_subset$Midwest)

# Print to a pretty data frame
eco_subset %>%
  kbl(caption = "Average (Standard Deviation) Min.same for closely related strains when more than 100 ca
  kable_styling(full_width = FALSE, latex_options = c('hold_position'))


# Campylobacter

# Look at the number of cases in a region in a given month/year and find how closely related they
# are on average
camp_subset <- campylobacter %>%
               filter(!is.na(month)) %>%
               group_by(year, month, region) %>%
               summarize(mean1 = round(mean(Min.same, na.rm = T),2),
                         sd1 = round(sd(Min.same, na.rm = T),2),
                         num = n())  %>%
               ungroup() %>%
            # Substantial number of cases that are very closely related
               filter(mean1 < 3 & num == 43) %>%
               pivot_wider(names_from = region, values_from = c(mean1, sd1))

## Nothing by our threshold, just the max that we have changed in the summarization above
```