Food Borne Pathogens

Anthony Girard, Timothy Hedspeth, Yutong Li

2022-10-05

Introduction

Motivating question

Have you ever woken up in the morning and saw a news headline, like "E Coli. Outbreak at Chipolte leads to national recall"? Though it is easy to shrug this off and move on, but in the United States we note that every year there are an estimated 37 million cases of food borne illness.

Literature review

Data Tools

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x lubridate::as.difftime() masks base::as.difftime()
## x lubridate::date()        masks base::date()
## x dplyr::filter()          masks stats::filter()
## x lubridate::intersect()   masks base::intersect()
## x dplyr::lag()             masks stats::lag()
## x lubridate::setdiff()     masks base::setdiff()
## x lubridate::union()       masks base::union()

##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

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

Table 1: Variables with some missing data

	Salmonella	E. Coli	Campylobacter
Strain	0.0079	0.0738	0.0010
Serovar	0.2246	0.8652	1.0000
Host.disease	0.9865	0.8957	0.9990
Isolation.source	0.1085	0.1511	0.0408
Isolation.type	0.0009	0.0683	0.0007
Lat.Lon	0.9783	0.8139	0.9982
Source.type	0.9995	1.0000	1.0000
SNP.cluster	0.0563	0.4744	0.1468
Min.same	0.0648	0.5307	0.1584
Min.diff	0.1689	0.8448	0.3680
Assembly	0.1130	0.0679	0.0072
Outbreak	0.9993	0.9992	1.0000
AMR.genotypes	0.0121	0.0001	0.0980
Computed.types	0.0010	1.0000	1.0000

the efficiency of the exploratory analysis we load these data sets separately so computation can be done on individual data sets.

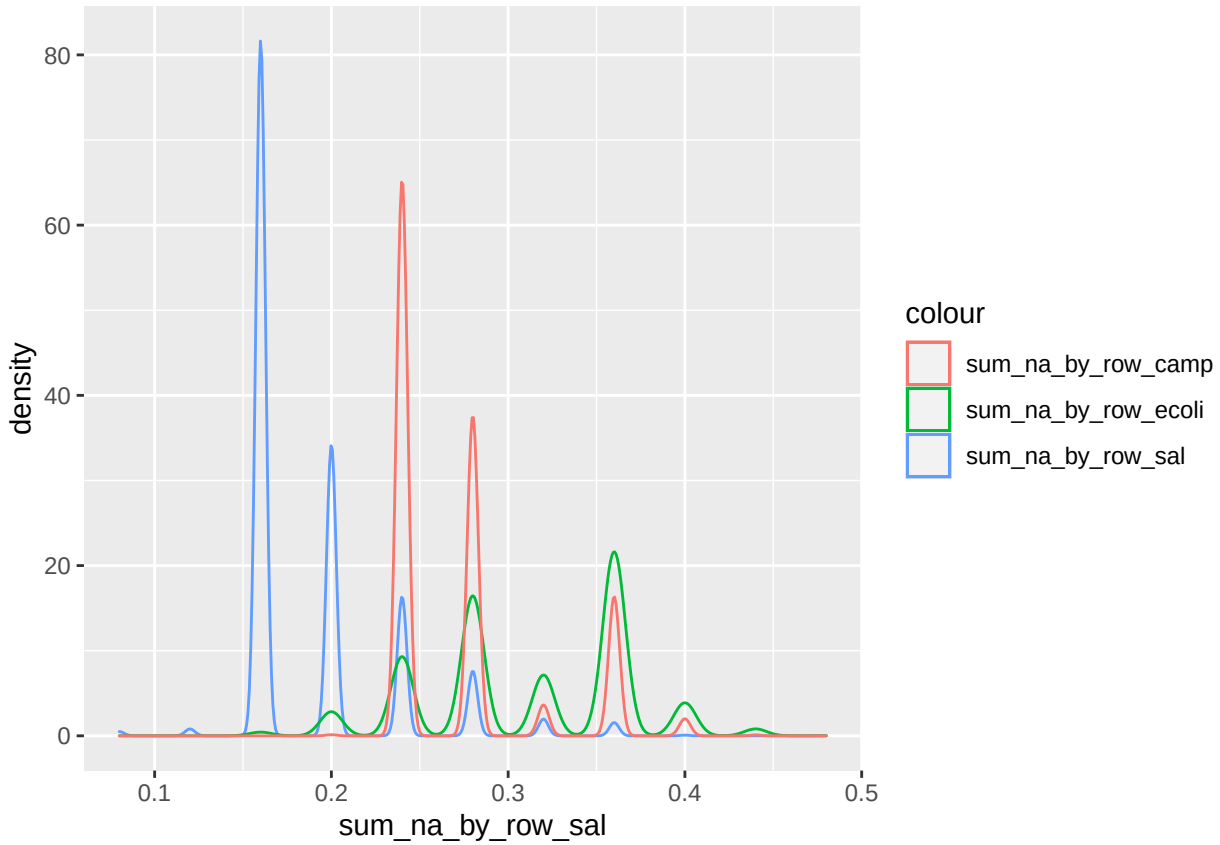
The data was extracted from ncbi and in our extraction process we conditioned on the *Collection.date* variable to extract information from. In initial analysis we noted that there was an extreme level of heterogeneity($>$ 6,000 levels in the Salmonella data) in the *Isolation.Source* variable, which required preprocessing prior to conducting an exploratory analysis. The 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.

Description of cleaning for USA and Date.

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.



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 variable we created in the preprocessing *Isolation.source.category*. 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.

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 approximatley 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.

```
## [1] 0
```

```
## character(0)
```

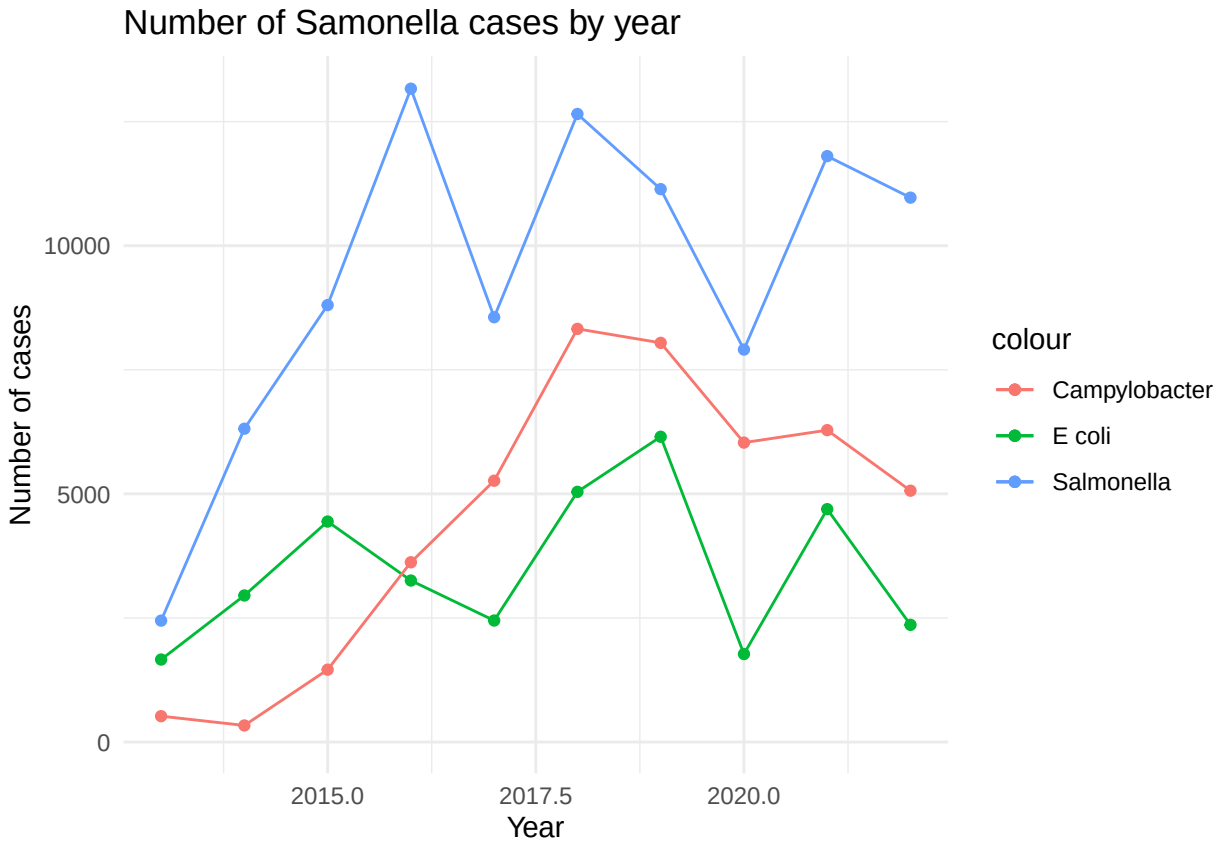
For our analysis there are a few variables that are of interest to our analysis. The first of which is the **Collection.date** variable, indicating when the sample was collected. **AMR_genotypes** is the Antimicrobial resistance genes, where a missing value indicates that there is no antimicrobial resistance detected.

AST_phenotypes tells us the phenotype based on testing, **host** is the host species, **host_disease** could be missing, **isolate_identifiers** a list alternative isolates, **isolation_source** describes the physical, environmental and/or local geographical source of the biological sample from which the sample was derived, if provided by the submitter, **epi_type** can be either clinical or environmental, **k-mer** is the organism work, **Min-same** is a measure of closeness to other isolates of the same type, **outbreak** if there is an unusal number of cases in an area, **erd_group** is the SNP cluster, **source_type** where the data comes from.

Trends over time

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

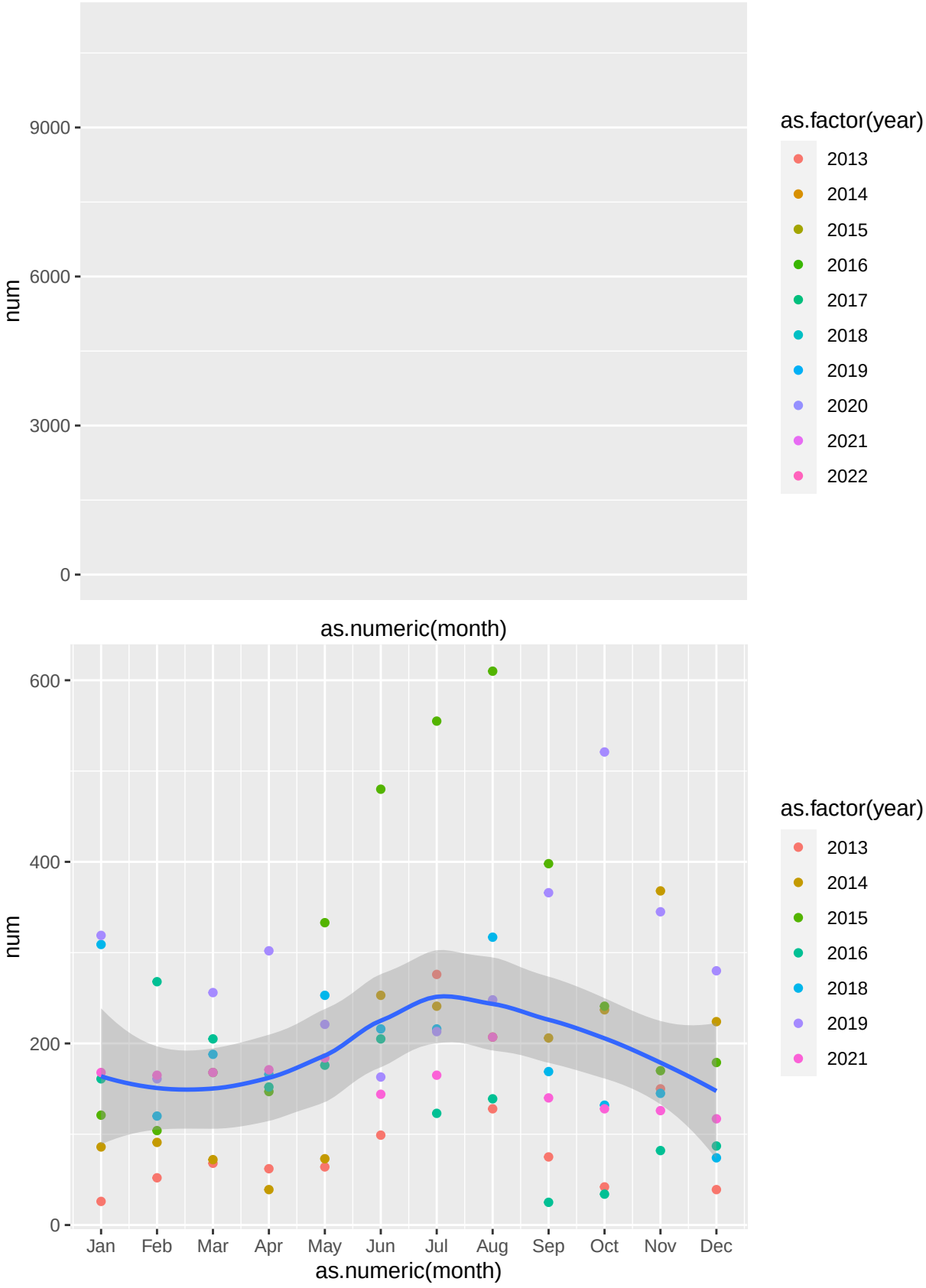
```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```

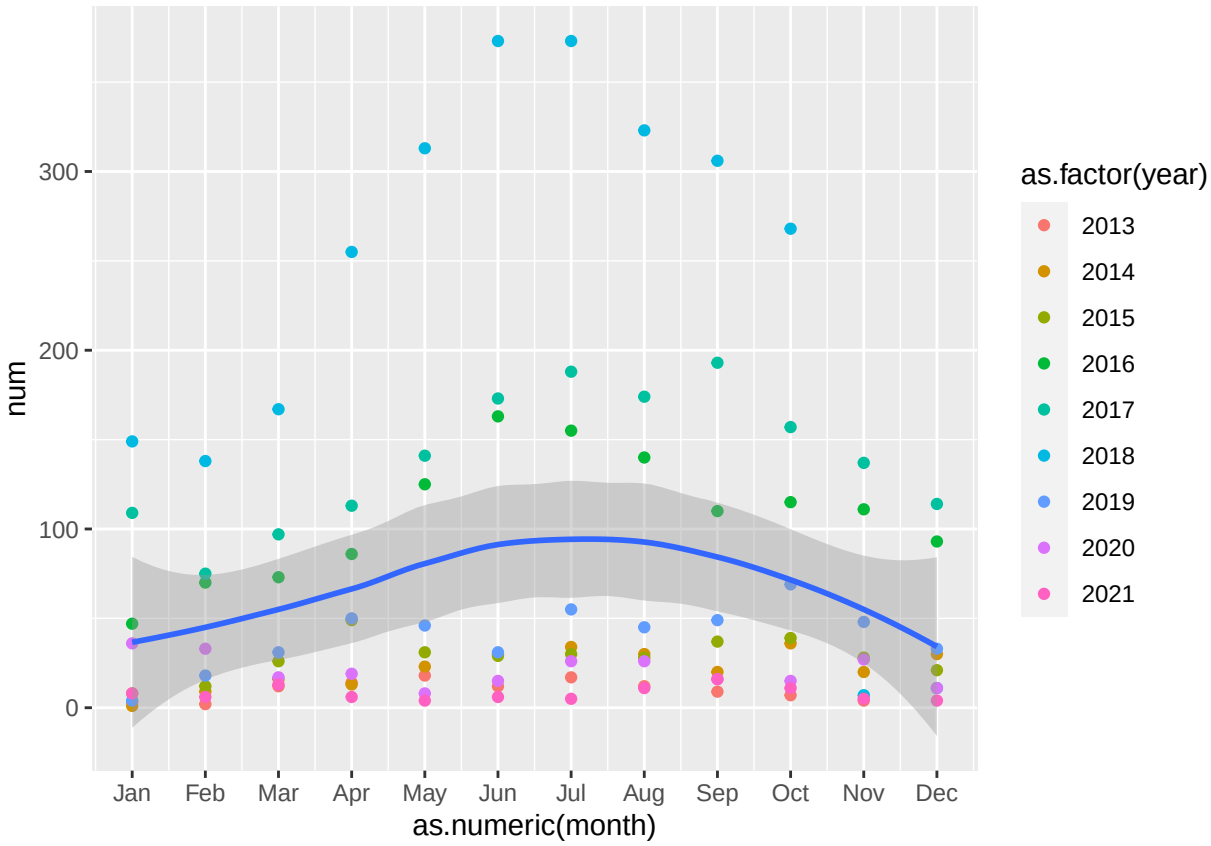



```
## Warning: Removed 94 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 94 rows containing missing values (geom_point).
```

```
## Warning: Position guide is perpendicular to the intended axis. Did you mean to
## specify a different guide `position`?
```

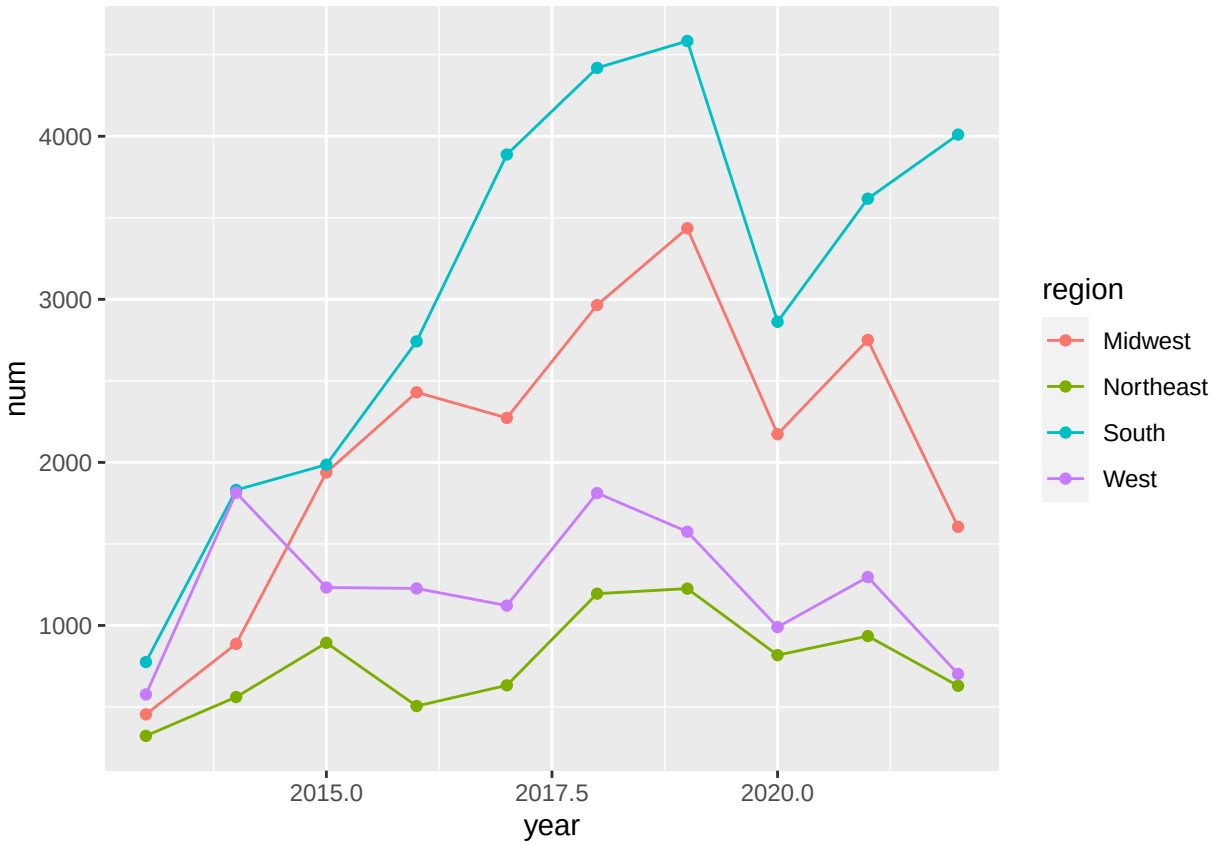

9000
6000
3000
0
num
as.numeric(month)
as.factor(year)
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
600
400
200
0
num
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
as.numeric(month)
as.factor(year)
2013
2014
2015
2016
2018
2019
2021



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 relaltionship 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.

We also examine and note the trends in seasonality

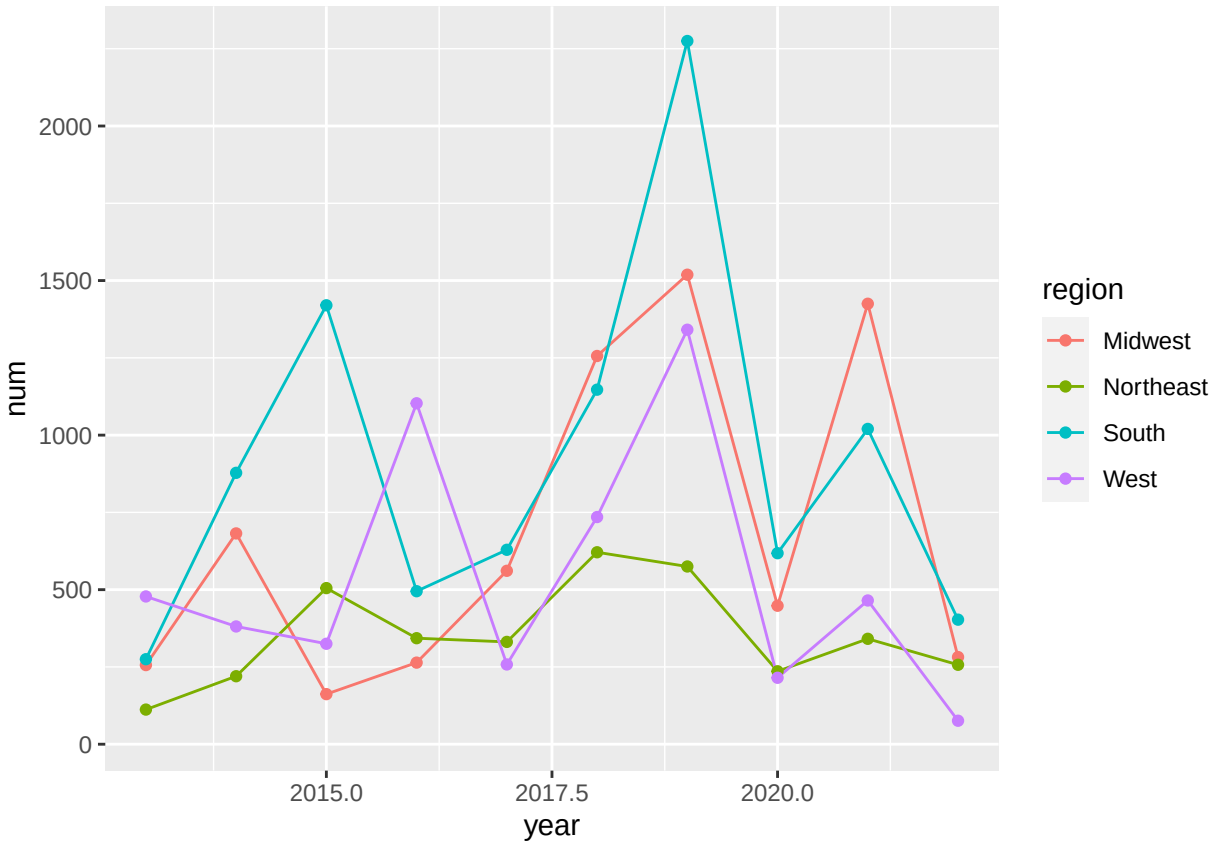
```
## `summarise()` has grouped output by 'region'. You can override using the
## `.groups` argument.
```



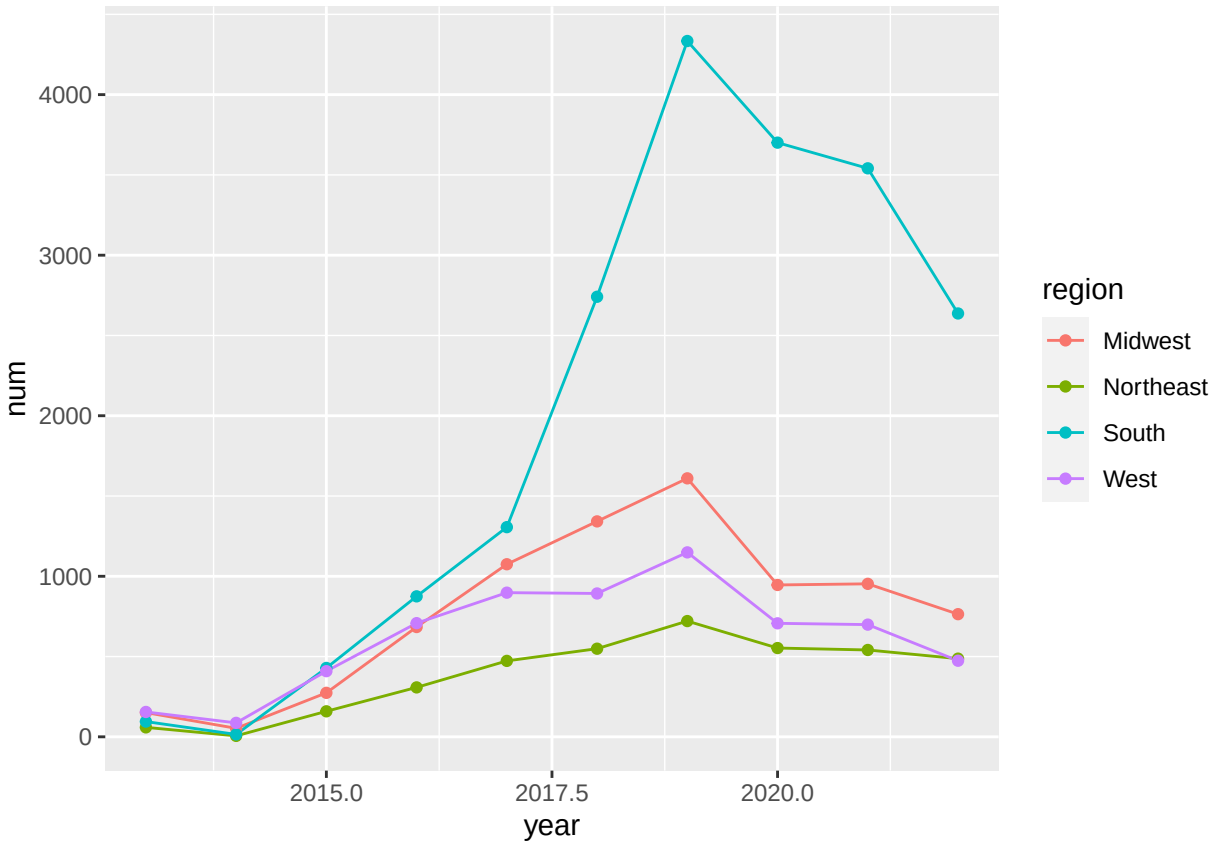

```
## `summarise()` has grouped output by 'region'. You can override using the
## `.groups` argument.
```

```
## Warning: Removed 1 row(s) containing missing values (geom_path).
```

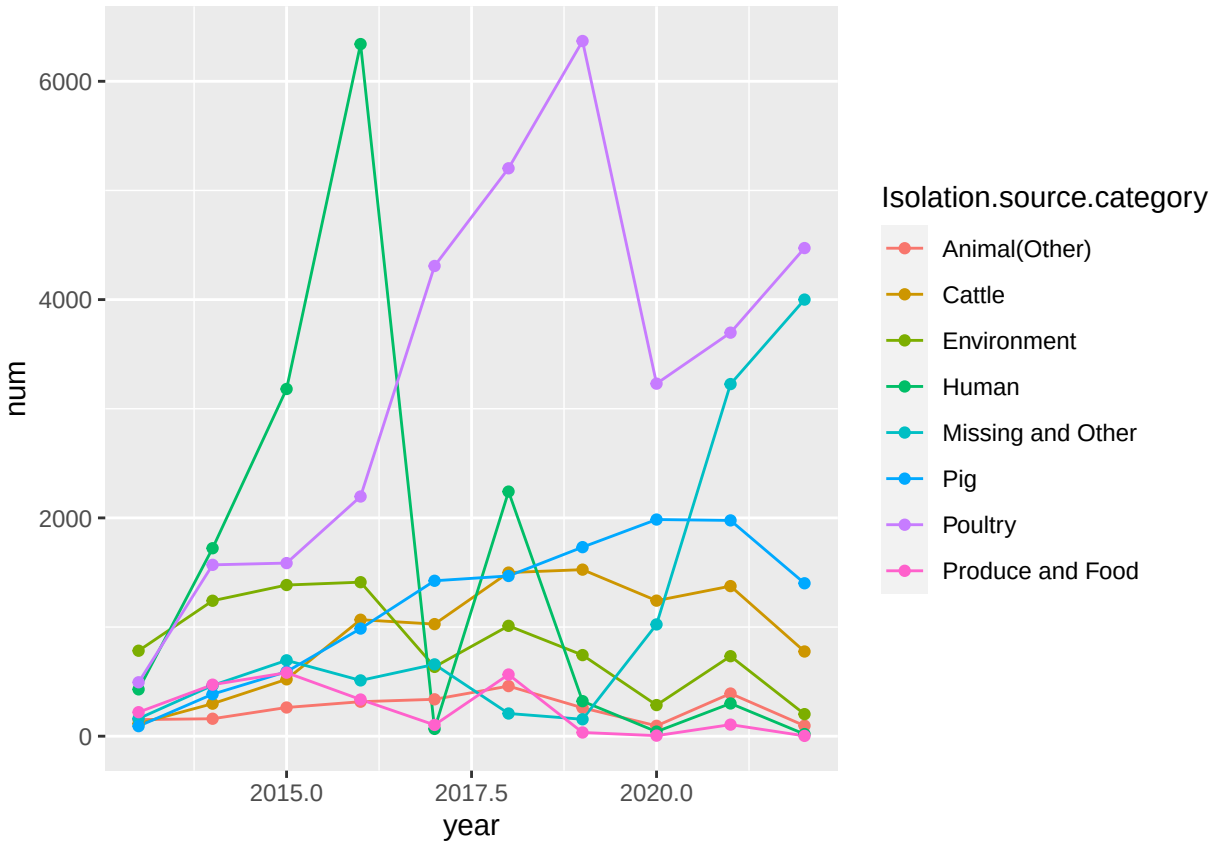
```
## Warning: Removed 1 rows containing missing values (geom_point).
```




```
## `summarise()` has grouped output by 'region'. You can override using the
## `.groups` argument.
```



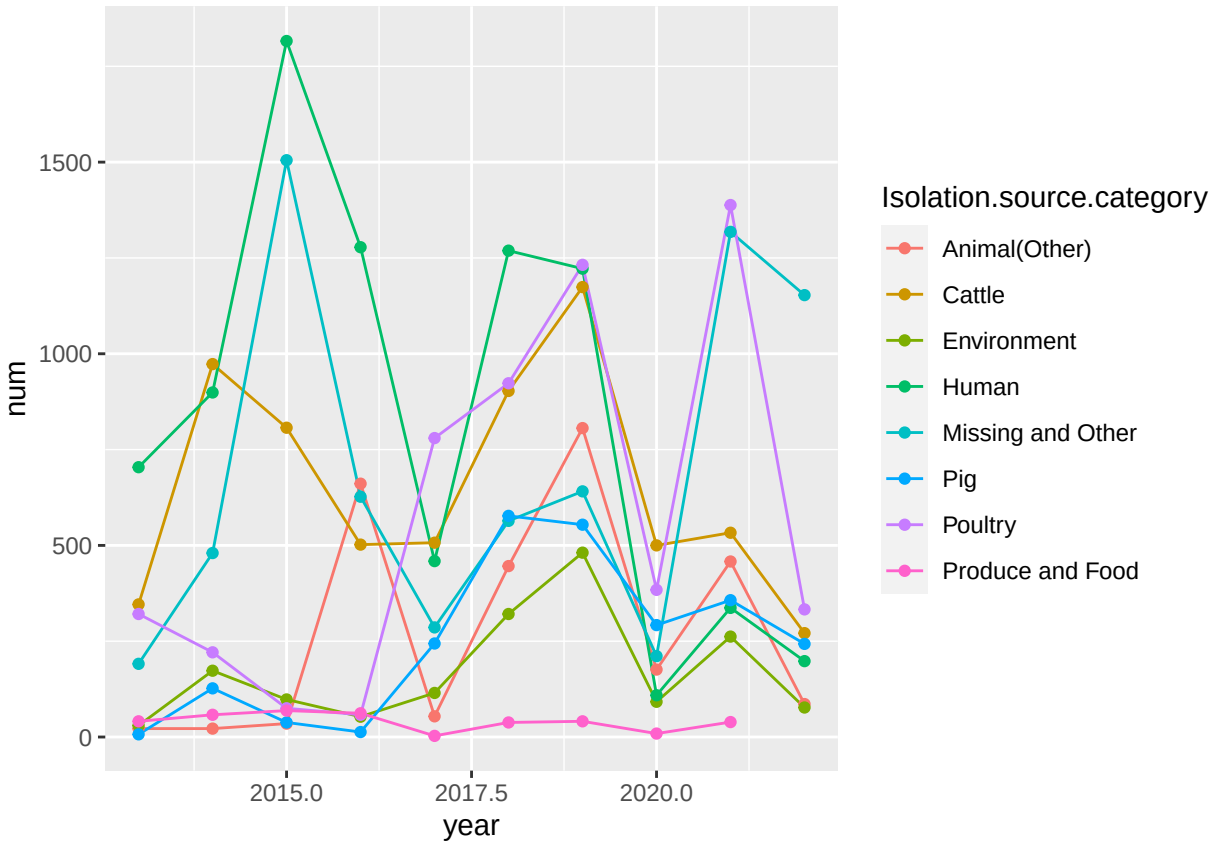

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```



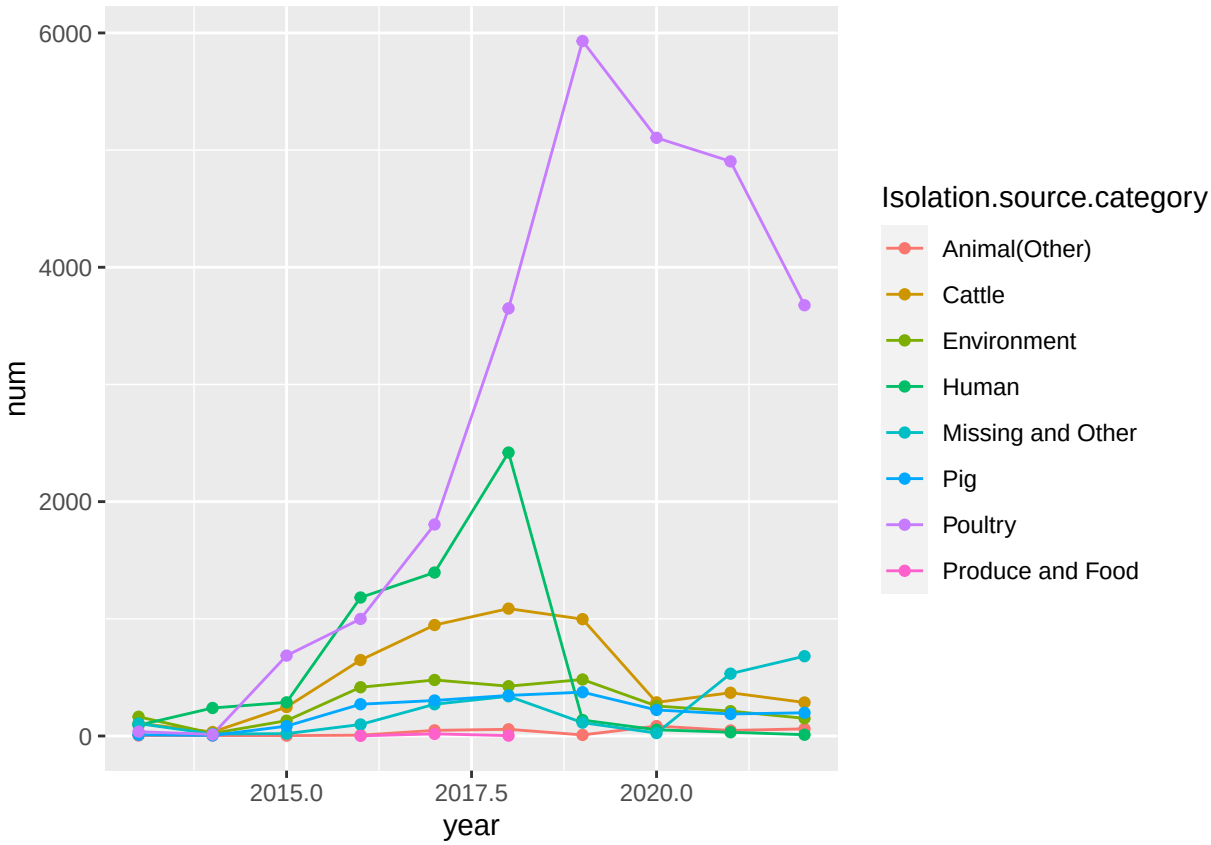

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

```
## Warning: Removed 3 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 3 rows containing missing values (geom_point).
```




```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```



We assume that if there is a missing value in the min difference column this will mean that there is no relation to other strains.

Minimum Same	Percent
Missing	0.06
Closely related (within 7 SNPs)	0.67
Not closely related	0.27

Minimum Same	Percent
Missing	0.53
Closely related (within 7 SNPs)	0.29
Not closely related	0.18

Minimum Same	Percent
Missing	0.16
Closely related (within 7 SNPs)	0.60
Not closely related	0.24

We observe that the min same is the same. Are there temporal trends by region? We will seek to see if any given

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

#~~~~~~~~~~~~#
## Packages ##
#~~~~~~~~~~~~#

library(lubridate) # working with dates
```

```r
library(tidyverse) # working with data
library(kableExtra)
library(anytime)


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

setwd("~/Desktop/Semester_3/Practical/Final")
salmonella <- read.csv("final_salmonella.csv")
ecoli <- read.csv("final_ecoli.csv")
campylobacter <- read.csv("final_Campylobacter.csv")




#~~~~~~~~~~~~~~~~#
## 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,4))

# 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,4))

# 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,4))

# 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   )
names(Missing_by_col) <- c("Salmonella", "E. Coli", "Campylobacter")

# Print the data frame
Missing_by_col %>% kbl(caption = "Variables with some missing data")



## 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))
#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
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"



#~~~~~~~~~~~~~~~~~~~~~~~~~#
## Baseline understanding ##
#~~~~~~~~~~~~~~~~~~~~~~~~~#
```

```r
# We note that some of the variables may not be that informative to our analysis
# so we explore to see if the columns tell us anything. This will allow for us
# to determine which variables to include in this analysis


## X.Organism.group

sal_group <- levels(as.factor(salmonella$X.Organism.group))
ecoli_grop <- levels(as.factor(ecoli$X.Organism.group))
camp_group <-  levels(as.factor(campylobacter$X.Organism.group))

## Strain
sal_strain <- length(levels(as.factor(salmonella$Strain)))
ecoli_strain <- length(levels(as.factor(ecoli$Strain)))
camp_strain <- length(levels(as.factor(campylobacter$Strain)))

## Isolate identifiers

# This is more complex than others


##

length(levels(as.factor(salmonella$Serovar)))


levels(as.factor(salmonella$Outbreak))




# We have to consider the date, given that the Colelction date is very bad we will
# have to clean it ourself

# Start with dates that have complete information
sal_complete_date <- salmonella %>% filter(nchar(Collection.date) > 7)
sal_complete_date$Collection.date <- as.Date(sal_complete_date$Collection.date)
sal_complete_date$year <- year(sal_complete_date$Collection.date)
sal_complete_date$month <- month(sal_complete_date$Collection.date)
sal_complete_date$day <- day(sal_complete_date$Collection.date)

# Some dates have year and month lets get this info
sal_only_month <- salmonella %>% filter(nchar(Collection.date) == 7)
sal_only_month$Collection.date <- anydate(sal_only_month$Collection.date)
sal_only_month$year <- year(sal_only_month$Collection.date)
sal_only_month$month <- month(sal_only_month$Collection.date)
sal_only_month$day <- NA

# Some dates have only the year, so we must account for this
sal_only_year <- salmonella %>% filter(nchar(Collection.date) < 7)
sal_only_year$Collection.date <- anydate(sal_only_year$Collection.date)
sal_only_year$year <- year(sal_only_year$Collection.date)
sal_only_year$month <- NA
```

```r
sal_only_year$day <- NA

salmonella <- rbind(sal_complete_date, sal_only_month, sal_only_year)

## Ecoli
# Start with dates that have complete information
ecoli_complete_date <- ecoli %>% filter(nchar(Collection.date) > 7)
ecoli_complete_date$Collection.date <- as.Date(ecoli_complete_date$Collection.date)
ecoli_complete_date$year <- year(ecoli_complete_date$Collection.date)
ecoli_complete_date$month <- month(ecoli_complete_date$Collection.date)
ecoli_complete_date$day <- day(ecoli_complete_date$Collection.date)

# Some dates have year and month lets get this info
ecoli_only_month <- ecoli %>% filter(nchar(Collection.date) == 7)
ecoli_only_month$Collection.date <- anydate(ecoli_only_month$Collection.date)
ecoli_only_month$year <- year(ecoli_only_month$Collection.date)
ecoli_only_month$month <- month(ecoli_only_month$Collection.date)
ecoli_only_month$day <- NA

# Some dates have only the year, so we must account for this
ecoli_only_year <- ecoli %>% filter(nchar(Collection.date) < 7)
ecoli_only_year$Collection.date <- anydate(ecoli_only_year$Collection.date)
ecoli_only_year$year <- year(ecoli_only_year$Collection.date)
ecoli_only_year$month <- NA
ecoli_only_year$day <- NA

ecoli <- rbind(ecoli_complete_date, ecoli_only_month, ecoli_only_year)


## Camploybacter

# Start with dates that have complete information
camp_complete_date <- campylobacter %>% filter(nchar(Collection.date) > 7)
camp_complete_date$Collection.date <- as.Date(camp_complete_date$Collection.date)
camp_complete_date$year <- year(camp_complete_date$Collection.date)
camp_complete_date$month <- month(camp_complete_date$Collection.date)
camp_complete_date$day <- day(camp_complete_date$Collection.date)

# Some dates have year and month lets get this info
camp_only_month <- campylobacter %>% filter(nchar(Collection.date) == 7)
camp_only_month$Collection.date <- anydate(camp_only_month$Collection.date)
camp_only_month$year <- year(camp_only_month$Collection.date)
camp_only_month$month <- month(camp_only_month$Collection.date)
camp_only_month$day <- NA

# Some dates have only the year, so we must account for this
camp_only_year <- campylobacter %>% filter(nchar(Collection.date) < 7)
camp_only_year$Collection.date <- anydate(camp_only_year$Collection.date)
camp_only_year$year <- year(camp_only_year$Collection.date)
camp_only_year$month <- NA
camp_only_year$day <- NA

campylobacter <- rbind(camp_complete_date, camp_only_month, camp_only_year)
```

```r
# Let us consider data by year



# 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
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 Samonella cases by year",
       x = "Year",
       y = "Number of cases")




# Number of cases per month for salmonella
Num_per_month_sal <- salmonella %>%
                     group_by(year, month) %>%
                     summarize(num =  n())
Num_per_month_sal$month <- factor(Num_per_month_sal$month, levels = c("Jan", "Feb",
                                                                      "Mar", "Apr", "May",
                                                                     "Jun", "Jul", "Aug",
                                                                     "Sep", "Oct",
                                                                     "Nov", "Dec"))
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()


# E. Coli
Num_per_month_ecoli <- ecoli %>%
                     group_by(year, month) %>%
                     summarize(num =  n())
Num_per_month_ecoli <- Num_per_month_ecoli %>% filter(!is.na(month))
#Num_per_month_ecoli$month <- factor(Num_per_month_ecoli$month, levels = c("Jan", "Feb",
     #                                                                      "Mar", "Apr", "May",
  #                                                                        "Jun", "Jul", "Aug",
     #                                                                      "Sep", "Oct",
     #                                                                      "Nov", "Dec"))
ggplot(data = Num_per_month_ecoli, 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()


# 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))
#Num_per_month_ecoli$month <- factor(Num_per_month_ecoli$month, levels = c("Jan", "Feb",
     #                                                                      "Mar", "Apr", "May",
  #                                                                        "Jun", "Jul", "Aug",
     #                                                                      "Sep", "Oct",
     #                                                                      "Nov", "Dec"))
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()




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

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


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

campylobacter <- campylobacter %>% mutate(region = case_when(Location %in% c("ME", "VT", "NH",
                                                                             "MA", "CT", "RI",
                                                                             "NY", "NJ", "PA") ~ "Northeast",
                                                             Location %in% c("DE", "MD", "DC",
                                                                             "WV", "VA", "NC",
                                                                             "SC", "GA", "FL",
                                                                             "AL", "MS", "LA",
                                                                             "TX", "OK", "AR",
                                                                             "TN", "KY") ~ "South",
                                                             Location %in% c("OH", "MI", "IN",
                                                                             "IL", "WI", "MN",
```

```
                                                                    "IA", "MO", "KS",
                                                                    "NE", "SD", "ND") ~ "Midwest",
                                                        Location %in% c("NM", "CO", "WY",
                                                                    "MT", "ID", "UT",
                                                                    "AZ", "NV", "WA",
                                                                    "OR", "CA", "HI",
                                                                    "AK") ~ "West",
                                                        Location == "USA" ~ "USA, General"))


# Salmomella

regions_sal <- salmonella %>% filter(region != "USA, General")
cases_by_region_sal <- regions_sal %>% group_by(region, year) %>%
                               summarize(num = n())

ggplot(data = cases_by_region_sal, aes(x = year, y = num, color = region)) +
  geom_line() +
  geom_point()

# E. Coli

regions_ecoli <- ecoli %>% filter(region != "USA, General")
cases_by_region_ecoli <- regions_ecoli %>% group_by(region, year) %>%
                                                    summarize(num = n())

ggplot(data = cases_by_region_ecoli, aes(x = year, y = num, color = region)) +
  geom_line() +
  geom_point()


# campylobacter

regions_camp <- campylobacter %>% filter(region != "USA, General")
cases_by_region_camp <- regions_camp %>% group_by(region, year) %>%
                               summarize(num = n())

ggplot(data = cases_by_region_camp, aes(x = year, y = num, color = region)) +
  geom_line() +
  geom_point()
## Lets start doing some eda with some of the categories we have defiend
plotdata_sal <- salmonella %>% group_by(year, Isolation.source.category) %>% summarize(num = n())

ggplot(data = plotdata_sal, aes(x = year, y = num, color = Isolation.source.category)) +
  geom_line() +
  geom_point()


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

ggplot(data = plotdata_ecoli, aes(x = year, y = num, color = Isolation.source.category)) +
  geom_line() +
```

```
  geom_point()

# Camp
plotdata_camp <- campylobacter %>% group_by(year, Isolation.source.category) %>% summarize(num = n())

ggplot(data = plotdata_camp, aes(x = year, y = num, color = Isolation.source.category)) +
  geom_line() +
  geom_point()
# Lets look at Min differences


# Salmonella
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)

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")

min_same_salmonella %>% kbl()

#
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)

min_same_ecoli <- data.frame()
min_same_ecoli[1,1] <- "Missing"
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")

min_same_ecoli %>% kbl()

#
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)

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")

min_same_campylobacter %>% kbl()

# Month and year

#sal_subset$week <- week(sal_subset$Create.date)
#x <- sal_subset %>%  group_by(year, week) %>%
 #                summarize(meaa = mean(Min.same, na.rm = T),
  #                         num_mn = sum(is.na(Min.same)),
   #                        num = n())  %>%
    #             ungroup() %>%
     #            arrange(year, week)


#x <- as.data.frame(x)

#x$try <- 0

#x <- x %>% mutate(weeks =  case_when(year == 2013 ~ week,
 #                                   year == 2014 ~ week+52,
  #                                  year == 2015 ~ week+(2*52),
   #                                 year == 2016 ~ week+(3*52),
    #                                year == 2017 ~ week+(4*52),
     #                               year == 2018 ~ week+(5*52),
     #                               year == 2019 ~ week+(6*52),
      #                              year == 2020 ~ week+(7*52),
       #                             year == 2021 ~ week+(8*52),
        #                            year == 2022 ~ week+(9*52)))

#x <- as.data.frame(x)
#for(i in 2:nrow(x)){
 # x$try[i] <- x$weeks[i]-x$weeks[i-1]
#}

# We want to look at outbreaks close in time
#x %>% filter(try < 2 & meaa  < 7)
```