Exploring Data Part 1

Ian Vert

6/18/2021

This Data is from : https://data.chhs.ca.gov/dataset/infectious-disease

This Dataset is Infectious Diseases by disease, County Year, and Sex for the State of California.

```
library(tidyr)
library(stringr)
library(tibble)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
library(ggplot2)
```

```
new_df <- read.csv('infectious-diseases-by-county-year-and-sex.csv')
head(new_df)
```

```
##     Disease  County Year    Sex Cases Population   Rate Lower_95__CI
## 1 Amebiasis Alameda 2001 Female     7     746596 0.938*        0.377
## 2 Amebiasis Alameda 2001   Male     9     718968 1.252*        0.572
## 3 Amebiasis Alameda 2001  Total    16    1465564 1.092*        0.624
## 4 Amebiasis Alameda 2002 Female     4     747987 0.535*        0.146
## 5 Amebiasis Alameda 2002   Male     5     720481 0.694*        0.225
## 6 Amebiasis Alameda 2002  Total     9    1468468 0.613*        0.280
##   Upper_95__CI
## 1        1.932
## 2        2.376
## 3        1.773
## 4        1.369
## 5        1.620
## 6        1.163
```

I have a DataFrame of 164,433 rows

Cleaning the Data

The Rate Column can have multiple different entries which would make it difficult to analyze.

First I will remove any rows that have a '-' in the Rate column as this indicases a Zero case counts according to the documentation.

```
new_df$Rate[new_df$Rate == '-'] <- NA

disease <- new_df %>% drop_na()
```

I now have a DataFrame of 33,984 rows

Now I need to remove any rows that have a '*' at the end of the Rate, because according to the documentation, 'indicates an unstable relative standard error wherein the relative standard error was 23 percent or more of the incidence rate estimate–a threshold recommended by the National Center for Health Statistics.'

```
new_vc <- vector()
for (x in disease$Rate) {
  if (str_sub(x, -1,-1) == '*'){
    x <- NA}
    new_vc <- c(new_vc,x)
  }
```

Here I created a subset of the Dataset without the rate column then added new_vc from the code chunk above

```
disease_sub <- subset(disease, select = -Rate)


disease_sub$Rate <- new_vc


disease_data <- disease_sub %>% drop_na()
```

Now without the rows with a Rate with a '*' we now have a DataFrame of 9,202 rows.

```
disease_data$Rate <- as.numeric(disease_data$Rate)
glimpse(disease_data)
```

```
## Rows: 9,202
## Columns: 9
## $ Disease      <chr> "Amebiasis", "Amebiasis", "Amebiasis", "Amebiasis", "Ameb~
## $ County       <chr> "Alameda", "Alameda", "Alameda", "Alameda", "Alameda", "A~
## $ Year         <int> 2010, 2010, 2011, 2012, 2012, 2015, 2015, 2015, 2016, 201~
## $ Sex          <chr> "Male", "Total", "Total", "Male", "Total", "Female", "Mal~
## $ Cases        <int> 20, 24, 21, 19, 27, 28, 37, 65, 33, 29, 62, 24, 32, 56, 2~
## $ Population   <int> 740574, 1510272, 1534536, 762985, 1557085, 825506, 796209~
## $ Lower_95__CI <dbl> 1.650, 1.018, 0.847, 1.499, 1.143, 2.254, 3.272, 3.094, 2~
## $ Upper_95__CI <dbl> 4.171, 2.364, 2.092, 3.889, 2.523, 4.902, 6.405, 5.109, 5~
## $ Rate         <dbl> 2.701, 1.589, 1.368, 2.490, 1.734, 3.392, 4.647, 4.009, 3~
```

Now that the dataset is ready for analysis I will change it to the Tibble.

```
clean_disease_data <- as_tibble(disease_data)


invisible(write.csv(clean_disease_data, 'clean_disease_data.csv'))

head(clean_disease_data)
```

```
## # A tibble: 6 x 9
##   Disease   County   Year Sex   Cases Population Lower_95__CI Upper_95__CI  Rate
```

```
##   <chr>     <chr>   <int> <chr> <int>      <int>        <dbl>        <dbl> <dbl>
## 1 Amebiasis Alameda  2010 Male     20     740574        1.65         4.17  2.70
## 2 Amebiasis Alameda  2010 Total    24    1510272        1.02         2.36  1.59
## 3 Amebiasis Alameda  2011 Total    21    1534536        0.847        2.09  1.37
## 4 Amebiasis Alameda  2012 Male     19     762985        1.50         3.89  2.49
## 5 Amebiasis Alameda  2012 Total    27    1557085        1.14         2.52  1.73
## 6 Amebiasis Alameda  2015 Fema~    28     825506        2.25         4.90  3.39
```

For each Disease and year the 'Sex' column has 'Male', 'Female', and 'Total', Currently I am only focused on Total number of cases. I due plan on analyzing Sex specific data later.

```
total_clean_disease_data <- clean_disease_data %>% filter(Sex == 'Total')

head(total_clean_disease_data)
```

```
## # A tibble: 6 x 9
##   Disease   County   Year Sex   Cases Population Lower_95__CI Upper_95__CI  Rate
##   <chr>     <chr>   <int> <chr> <int>      <int>        <dbl>        <dbl> <dbl>
## 1 Amebiasis Alameda  2010 Total    24    1510272        1.02         2.36  1.59
## 2 Amebiasis Alameda  2011 Total    21    1534536        0.847        2.09  1.37
## 3 Amebiasis Alameda  2012 Total    27    1557085        1.14         2.52  1.73
## 4 Amebiasis Alameda  2015 Total    65    1621520        3.09         5.11  4.01
## 5 Amebiasis Alameda  2016 Total    62    1637792        2.90         4.85  3.79
## 6 Amebiasis Alameda  2017 Total    56    1651559        2.56         4.40  3.39
```

Key question: What are the top 6 Diseases with the highest mean number of cases?

```
avg_total_clean_disease_data <- total_clean_disease_data %>% group_by(Disease) %>% mutate(mean(Cases))

head(avg_total_clean_disease_data)
```

```
## # A tibble: 6 x 10
## # Groups:   Disease [1]
##   Disease   County   Year Sex   Cases Population Lower_95__CI Upper_95__CI  Rate
##   <chr>     <chr>   <int> <chr> <int>      <int>        <dbl>        <dbl> <dbl>
## 1 Amebiasis Alameda  2010 Total    24    1510272        1.02         2.36  1.59
## 2 Amebiasis Alameda  2011 Total    21    1534536        0.847        2.09  1.37
## 3 Amebiasis Alameda  2012 Total    27    1557085        1.14         2.52  1.73
## 4 Amebiasis Alameda  2015 Total    65    1621520        3.09         5.11  4.01
## 5 Amebiasis Alameda  2016 Total    62    1637792        2.90         4.85  3.79
## 6 Amebiasis Alameda  2017 Total    56    1651559        2.56         4.40  3.39
## # ... with 1 more variable: mean(Cases) <dbl>
```

```
unique(avg_total_clean_disease_data %>% select(Disease, `mean(Cases)`)) %>% arrange(desc(`mean(Cases)`))
```

```
## # A tibble: 36 x 2
## # Groups:   Disease [36]
##    Disease                                      `mean(Cases)`
##    <chr>                                                <dbl>
##  1 Coccidioidomycosis                                    390.
##  2 Campylobacteriosis                                    287.
##  3 Salmonellosis                                         223.
##  4 Shigellosis                                           188.
##  5 Giardiasis                                            148.
##  6 Amebiasis                                             131.
```

```
##  7 Shiga toxin-producing E. coli (STEC) without HUS         109.
##  8 Legionellosis                                            107.
##  9 Cryptosporidiosis                                         80.0
## 10 Vibrio Infection (non-Cholera)                            79.7
## # ... with 26 more rows
```

The top 6 Diseases with the highest mean number of cases are Coccidioidomycosis, Campylobacteriosis, Salmonellosis, Shigellosis, Giardiasis, Amebiasis

Key Question: How do the cases of these 6 Diseases change over time?

```
total_top_six <- total_clean_disease_data %>% filter(Disease == 'Campylobacteriosis' |
                                                     Disease == 'Coccidioidomycosis'|
                                                     Disease == 'Salmonellosis' |
                                                     Disease == 'Giardiasis' |
                                                     Disease == 'Amebiasis')

head(total_top_six)
```

```
## # A tibble: 6 x 9
##   Disease   County   Year Sex   Cases Population Lower_95__CI Upper_95__CI  Rate
##   <chr>     <chr>   <int> <chr> <int>      <int>        <dbl>        <dbl> <dbl>
## 1 Amebiasis Alameda  2010 Total    24    1510272        1.02          2.36  1.59
## 2 Amebiasis Alameda  2011 Total    21    1534536        0.847         2.09  1.37
## 3 Amebiasis Alameda  2012 Total    27    1557085        1.14          2.52  1.73
## 4 Amebiasis Alameda  2015 Total    65    1621520        3.09          5.11  4.01
## 5 Amebiasis Alameda  2016 Total    62    1637792        2.90          4.85  3.79
## 6 Amebiasis Alameda  2017 Total    56    1651559        2.56          4.40  3.39
```

```
sum_total_clean_disease_data <- unique(total_top_six %>%
              group_by(Year, Disease) %>%
              mutate(sum(Cases)) %>%
              select(Disease,Year, `sum(Cases)`) %>%
              arrange((Year)))

head(sum_total_clean_disease_data,10)
```

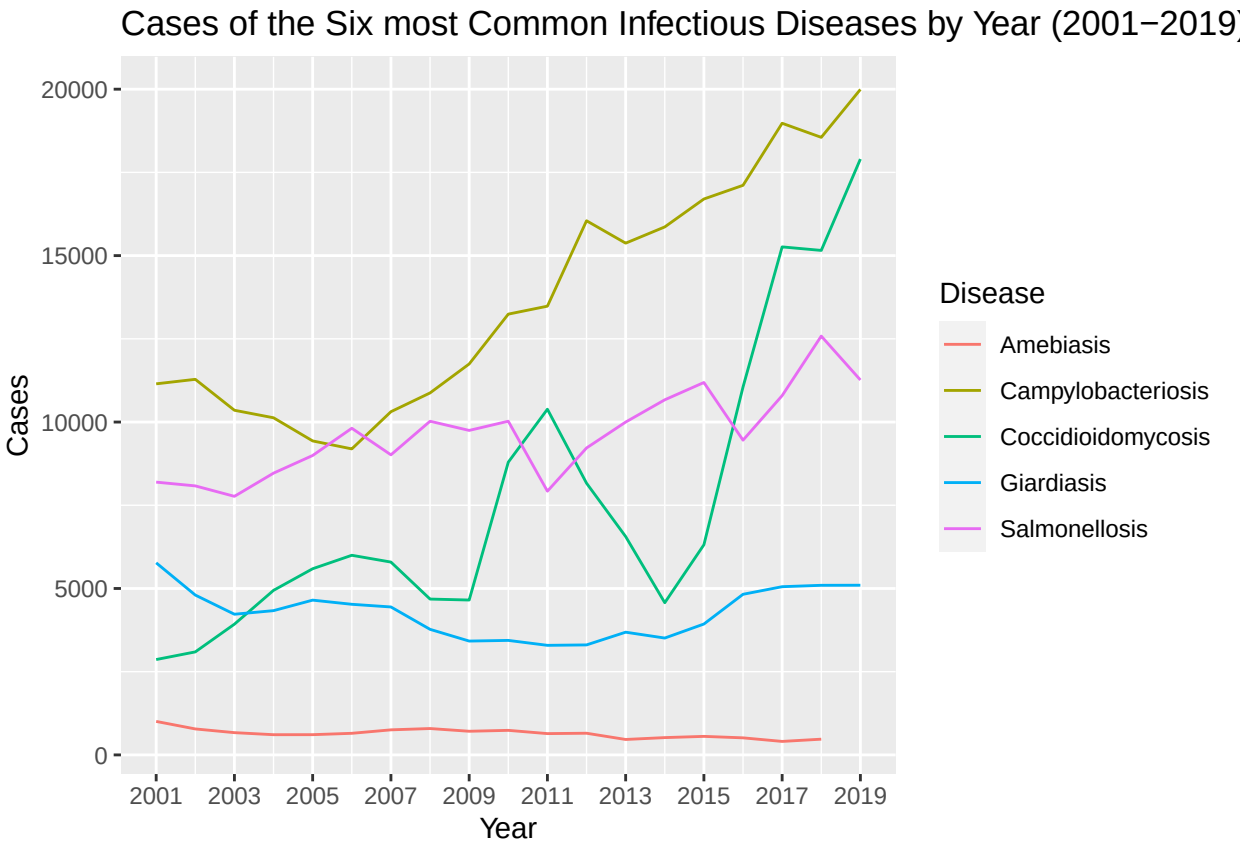
```
## # A tibble: 10 x 3
## # Groups:   Year, Disease [10]
##    Disease             Year `sum(Cases)`
##    <chr>              <int>        <int>
##  1 Amebiasis           2001         1006
##  2 Campylobacteriosis  2001        11148
##  3 Coccidioidomycosis  2001         2865
##  4 Giardiasis          2001         5771
##  5 Salmonellosis       2001         8195
##  6 Amebiasis           2002          780
##  7 Campylobacteriosis  2002        11283
##  8 Coccidioidomycosis  2002         3095
##  9 Giardiasis          2002         4802
## 10 Salmonellosis       2002         8082
```

```
p <- sum_total_clean_disease_data %>% ggplot(mapping = aes(x = Year ,
                                                           y = `sum(Cases)`,
                                                           color = Disease)) +
```

```
  geom_line(aes(group = Disease)) +

  scale_x_continuous(breaks = round(seq(min(sum_total_clean_disease_data$Year),
                                        max(sum_total_clean_disease_data$Year),
                                        by = 2))) +
  labs(x = 'Year',
       y = 'Cases',
       title = 'Cases of the Six most Common Infectious Diseases by Year (2001-2019)')

p
```



Interesting Notes from the graph:

Campylobacteriosis has increased steadily since 2006 and Coccidioidomycosis increased steadily since 2014

Coccidioidomycosis had a spike in cases from 2009 to 2011 and decreased from 2011 to 2014

While Amebiasis and Giardiasis are both fairly stable in cases while Salmonellosis is sporadic