

An Investigation of California's Infectious Diseases

Daniel Fields

1/11/2019

SB Hacks 2019 Project

Goal: Understand the relationship between gender, disease class, and the number of occurrences of a disease in a given contracting different types of diseases.

First we will install some packages to aid in our analysis.

```
install.packages(c("maps", "mapdata"), repos = "https://cloud.r-project.org/")
```

```
##  
## The downloaded binary packages are in  
## /var/folders/35/jq1trs4930qghj188nd9pvvw0000gn/T//RtmpZ6rhPB/downloaded_packages
```

```
install.packages("reshape2", repos = "https://cloud.r-project.org/")
```

```
##  
## The downloaded binary packages are in  
## /var/folders/35/jq1trs4930qghj188nd9pvvw0000gn/T//RtmpZ6rhPB/downloaded_packages
```

```
devtools::install_github("dgrtwo/gganimate", repos = "https://cloud.r-project.org/")
```

```
## Skipping install of 'gganimate' from a github remote, the SHA1 (e16365b4) has not changed since last  
## Use `force = TRUE` to force installation
```

Next, we will load in the packages to our workspace so that we may use the functions within them.

```
library(tidyverse)
```

```
## -- Attaching packages -----  
## √ ggplot2 3.1.0      √ purrr  0.2.5  
## √ tibble  1.4.2      √ dplyr  0.7.5  
## √ tidyr   0.8.1      √ stringr 1.3.1  
## √ readr   1.1.1      √ forcats 0.3.0
```

```
## -- Conflicts ----- t  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag()    masks stats::lag()
```

```
library(ggribes)
```

```
##  
## Attaching package: 'ggribes'  
## The following object is masked from 'package:ggplot2':  
##  
## scale_discrete_manual
```

```
library(tseries)  
library(cluster)  
library(maps)
```

```
##  
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:cluster':
##
## votes.repub
```

```
## The following object is masked from 'package:purrr':
##
## map
```

```
library(mapdata)
library(ggmap)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
## smiths
```

```
library(gganimate)
library(nnet)
```

Next, we will load in the data. This data comes from The California Department of Public Health (CDPH). This data is formally titled: “Infectious Disease Data among California Residents by Disease, County, Sex, and Year, 2001-2014”. The raw data is shown below.

```
setwd("/Users/DanielsMac/Desktop/SBHacks")
#load data
health_data <- read.csv("rows.csv")
head(health_data)
```

```
## Disease County Year Sex Count Population Rate CI.lower
## 1 Amebiasis California 2001 Total 571 34514777 1.654 1.521
## 2 Amebiasis California 2001 Female 176 17340743 1.015 0.871
## 3 Amebiasis California 2001 Male 365 17174034 2.125 1.913
## 4 Amebiasis California 2002 Total 442 34940334 1.265 1.150
## 5 Amebiasis California 2002 Female 145 17555714 0.826 0.697
## 6 Amebiasis California 2002 Male 279 17384620 1.605 1.422
## CI.upper Unstable
## 1 1.796
## 2 1.176
## 3 2.355
## 4 1.389
## 5 0.972
## 6 1.805
```

Below we have some plots that analyze how counties of different sizes total infected persons vary by population. What we see is that smaller counties tend to have more variability, but overall there is an upward trend to the number of incidents of infectious diseases over time.

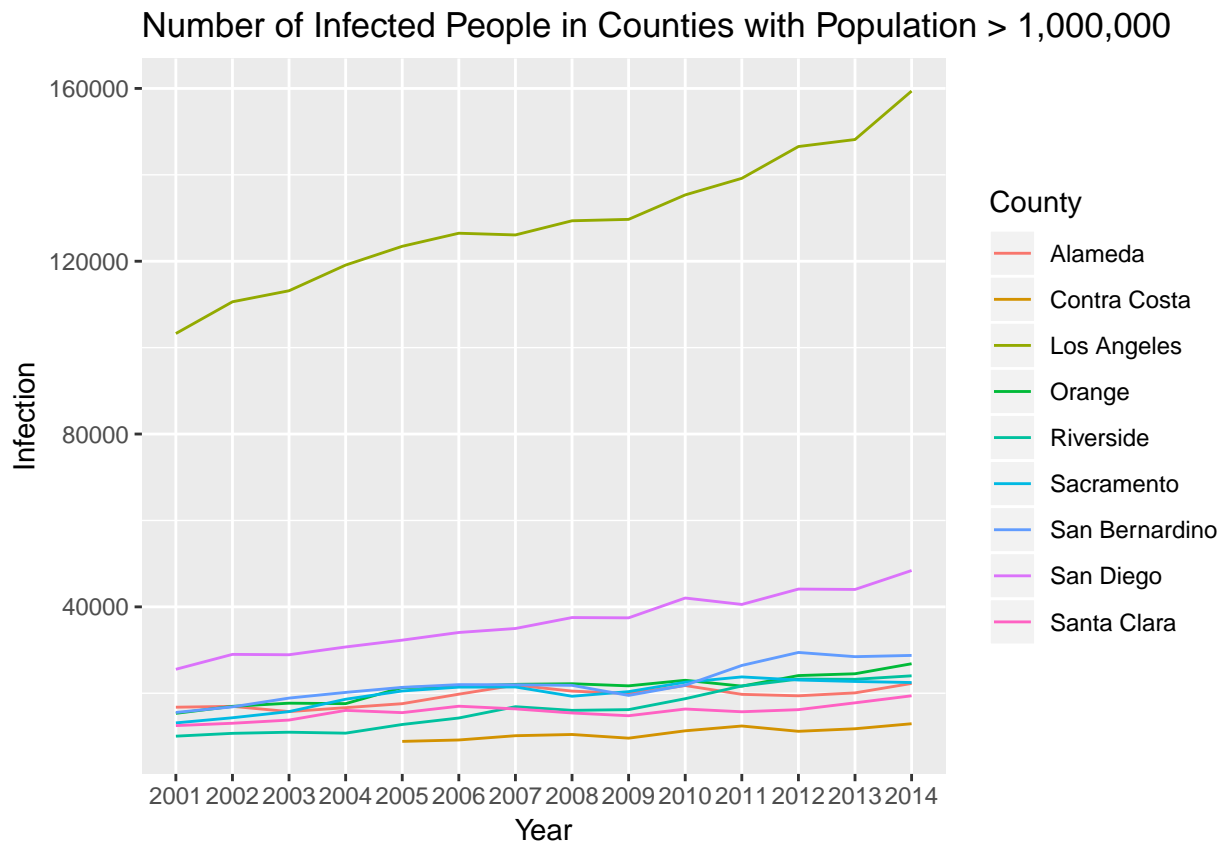
```
#Get list of years present in data set
years <- c(min(health_data$Year):max(health_data$Year))
years.char <- as.character(years)

#Plots to understand how many people get infected by all types of diseases
health_data%>%
  group_by(County,Year)%>%
  mutate(Infection = sum(Count))%>%
```

```

filter(County != "California")%>%
filter(Population >= 1000000)%>%
select(County, Year, Infection)%>%
unique()%>%
arrange(Infection)%>%
ggplot(aes(x = Year, y = Infection, colour = County))+
geom_line()+
ggtitle("Number of Infected People in Counties with Population > 1,000,000")+
scale_x_discrete("Year", years, years.char, years)

```

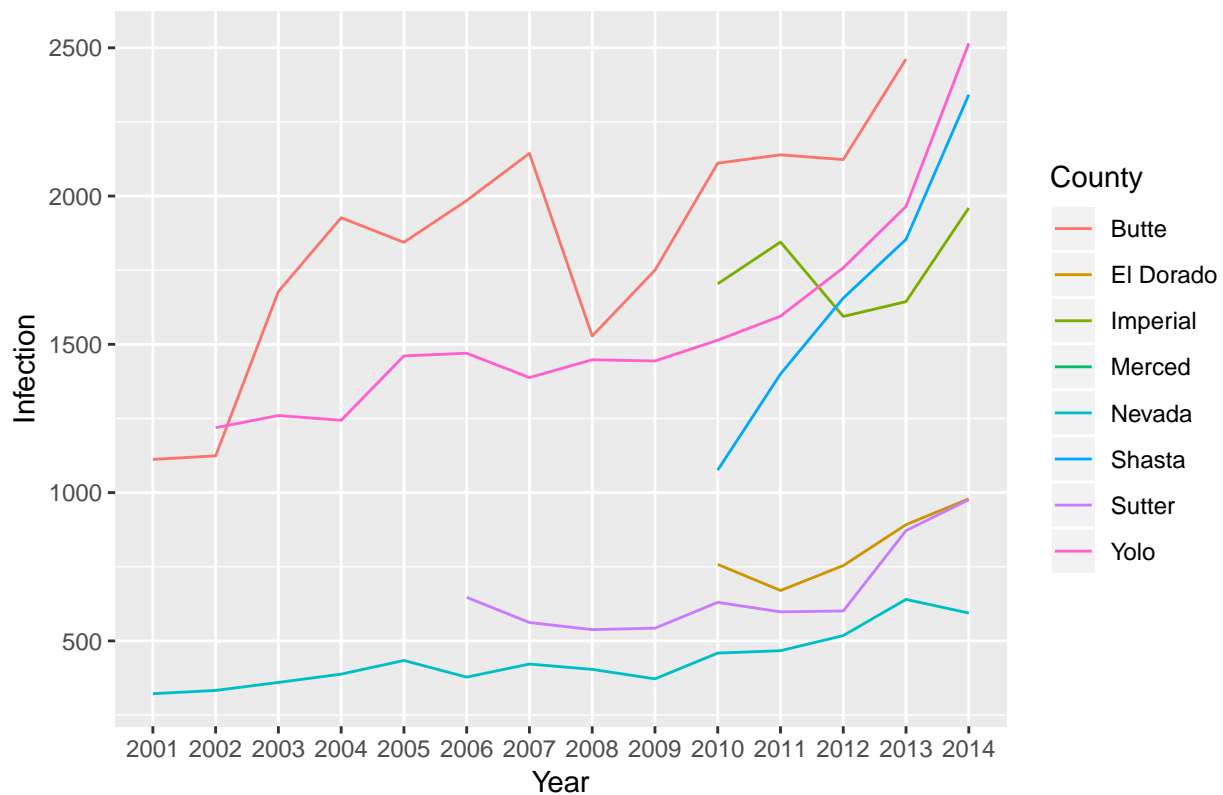


```

health_data%>%
  group_by(County, Year)%>%
  mutate(Infection = sum(Count))%>%
  filter(County != "California")%>%
  filter(Population > 90000 & Population < 110000)%>%
  select(County, Year, Infection)%>%
  unique()%>%
  arrange(Infection)%>%
  ggplot(aes(x = Year, y = Infection, colour = County))+
  geom_line()+
  ggtitle("Number of Infected People in Counties with Population around 100,000")+
  scale_x_discrete("Year", years, years.char, years)

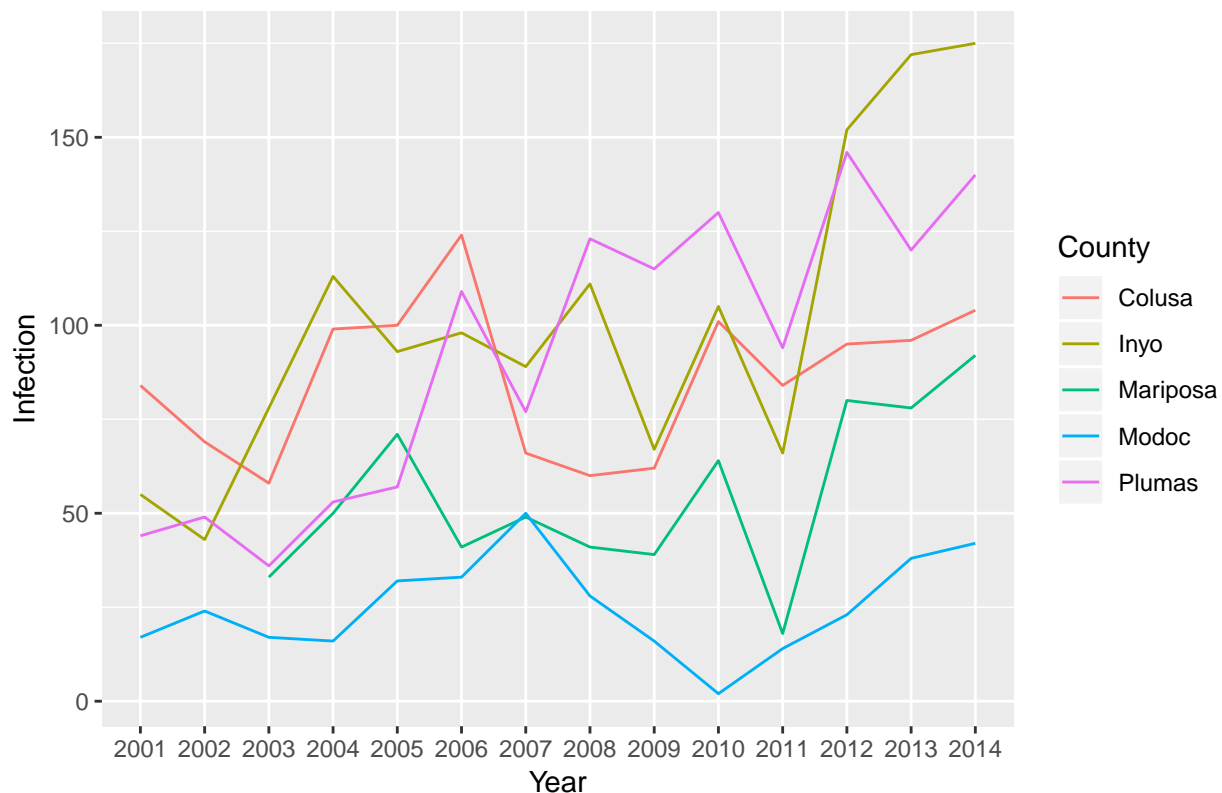
```

Number of Infected People in Counties with Population around 100,000



```
health_data%>%
  group_by(County, Year)%>%
  mutate(Infection = sum(Count))%>%
  filter(County != "California")%>%
  filter(Population > 9000 & Population < 11000)%>%
  select(County, Year, Infection)%>%
  unique()%>%
  arrange(Infection)%>%
  ggplot(aes(x = Year, y = Infection, colour = County))+
  geom_line()+
  ggtitle("Number of Infected People in Counties with Population around 10,000")+
  scale_x_discrete("Year", years, years.char, years)
```

Number of Infected People in Counties with Population around 10,000



So, in the original data, there are 64 different infectious diseases. To bring down the scale of our analysis, I decided to classify each disease into a classification. - Lethal Viral = disease is potentially lethal and caused by a virus - Non-Lethal Viral = disease is caused by a virus - Lethal Bacterial = disease is potentially lethal and caused by a bacteria - Non-Lethal Bacterial = disease is caused by a bacteria - Lethal Parasitic = disease is potentially lethal and caused by parasites (fungus included) - Non-Lethal Parasitic = disease is caused by parasites (fungus included) - STD - Sexually Transmitted Disease - Vaccine-Preventable (person who contracted disease of this category did not have a working vaccine) - Genetic - disease acts on protein synthesis.

```
#Load in the disease classifications.
setwd("/Users/DanielsMac/Desktop/SBHacks")
classes <- read.csv("classes.csv")
classes
```

```
##
## 1 Disease
## 2 Amebiasis
## 3 Anaplasmosis and Ehrlichiosis
## 4 Babesiosis
## 5 Botulism, Foodborne
## 6 Botulism, Other
## 7 Botulism, Wound
## 8 Brucellosis
## 9 Campylobacteriosis
## 10 Chlamydia
## 11 Cholera
## 12 Ciguatera Fish Poisoning
## 13 Coccidioidomycosis
## 14 Creutzfeldt-Jakob Disease and other Transmissible Spongiform Encephalopathies
```

## 14	Cryptosporidiosis
## 15	Cyclosporiasis
## 16	Cysticercosis or Taeniasis
## 17	Dengue
## 18	Diphtheria
## 19	E. coli 0157
## 20	E. coli Other STEC (non-0157)
## 21	Early Syphilis
## 22	Giardiasis
## 23	Gonorrhea
## 24	Hantavirus Infection
## 25	Hemolytic Uremic Syndrome
## 26	Hepatitis A
## 27	Hepatitis B, Acute
## 28	Hepatitis C, Acute
## 29	Hepatitis E, acute infection
## 30	HIV
## 31	Influenza Death (<65 years of age)
## 32	Invasive Meningococcal Disease
## 33	Legionellosis
## 34	Leprosy
## 35	Leptospirosis
## 36	Listeriosis
## 37	Lyme Disease
## 38	Malaria
## 39	Measles
## 40	Mumps
## 41	Paralytic Shellfish Poisoning
## 42	Pertussis
## 43	Plague, human
## 44	Psittacosis
## 45	Q Fever
## 46	Rabies, human
## 47	Relapsing Fever
## 48	Rubella
## 49	Salmonellosis
## 50	Scombroid Fish Poisoning
## 51	Shiga Toxin Positive Feces (without culture confirmation)
## 52	Shigellosis
## 53	Spotted Fever Rickettsiosis
## 54	Staphylococcus aureus Infection (cases resulting in death or ICU)
## 55	Streptococcal Infection (cases in food and dairy workers)
## 56	Tetanus
## 57	Toxic Shock Syndrome (Non-Streptococcal)
## 58	Trichinosis
## 59	Tuberculosis
## 60	Tularemia
## 61	Typhoid Fever, case
## 62	Typhus Fever
## 63	Varicella Hospitalizations
## 64	Vibrio Infection (non-Cholera)
## 65	Yersiniosis
##	Classification
## 1	6

## 2	5
## 3	5
## 4	3
## 5	3
## 6	3
## 7	4
## 8	4
## 9	7
## 10	3
## 11	6
## 12	6
## 13	9
## 14	6
## 15	6
## 16	5
## 17	1
## 18	8
## 19	4
## 20	4
## 21	7
## 22	6
## 23	7
## 24	1
## 25	3
## 26	1
## 27	2
## 28	2
## 29	2
## 30	7
## 31	1
## 32	3
## 33	4
## 34	4
## 35	3
## 36	1
## 37	3
## 38	4
## 39	8
## 40	8
## 41	3
## 42	8
## 43	3
## 44	4
## 45	4
## 46	1
## 47	4
## 48	2
## 49	4
## 50	4
## 51	4
## 52	4
## 53	4
## 54	3
## 55	4

```
## 56      8
## 57      3
## 58      6
## 59      3
## 60      3
## 61      4
## 62      4
## 63      8
## 64      4
## 65      4
```

```
health_data <- health_data%>%
  mutate(Classification = 0)

#Adding classifications to the individual diseases
for(i in 1:length(health_data$Disease)){
  for(j in 1:length(classes$Disease)){
    if( health_data$Disease[i] == classes$Disease[j] ){
      health_data$Classification[i] <- classes$Classification[j]
    }
  }
}

#remove rows that hold summary data
health_data <- health_data%>%
  mutate(Classification = as.factor(Classification))%>%
  filter(County != "California")%>% #Ignore State Totals
  filter(Sex != "Total")#Ignore Gender Totals

class.print <- classes%>%
  mutate(Classification = as.factor(Classification))
levels(class.print$Classification) <- c("Lethal Viral", "Non-Lethal Viral", "Lethal Bacterial", "Non-Lethal Bacterial")
print("All of the different diseases present in our data and their subsequent classifications are:")

## [1] "All of the different diseases present in our data and their subsequent classifications are:"
class.print
```

```
##
## 1      Disease
## 2      Amebiasis
## 3      Anaplasmosis and Ehrlichiosis
## 4      Babesiosis
## 5      Botulism, Foodborne
## 6      Botulism, Other
## 7      Botulism, Wound
## 8      Brucellosis
## 9      Campylobacteriosis
## 10     Chlamydia
## 11     Cholera
## 12     Ciguatera Fish Poisoning
## 13     Coccidioidomycosis
## 14     Creutzfeldt-Jakob Disease and other Transmissible Spongiform Encephalopathies
## 15     Cryptosporidiosis
## 16     Cyclosporiasis
## 17     Cysticercosis or Taeniasis
```


## 17	Dengue
## 18	Diphtheria
## 19	E. coli 0157
## 20	E. coli Other STEC (non-0157)
## 21	Early Syphilis
## 22	Giardiasis
## 23	Gonorrhea
## 24	Hantavirus Infection
## 25	Hemolytic Uremic Syndrome
## 26	Hepatitis A
## 27	Hepatitis B, Acute
## 28	Hepatitis C, Acute
## 29	Hepatitis E, acute infection
## 30	HIV
## 31	Influenza Death (<65 years of age)
## 32	Invasive Meningococcal Disease
## 33	Legionellosis
## 34	Leprosy
## 35	Leptospirosis
## 36	Listeriosis
## 37	Lyme Disease
## 38	Malaria
## 39	Measles
## 40	Mumps
## 41	Paralytic Shellfish Poisoning
## 42	Pertussis
## 43	Plague, human
## 44	Psittacosis
## 45	Q Fever
## 46	Rabies, human
## 47	Relapsing Fever
## 48	Rubella
## 49	Salmonellosis
## 50	Scombroid Fish Poisoning
## 51	Shiga Toxin Positive Feces (without culture confirmation)
## 52	Shigellosis
## 53	Spotted Fever Rickettsiosis
## 54	Staphylococcus aureus Infection (cases resulting in death or ICU)
## 55	Streptococcal Infection (cases in food and dairy workers)
## 56	Tetanus
## 57	Toxic Shock Syndrome (Non-Streptococcal)
## 58	Trichinosis
## 59	Tuberculosis
## 60	Tularemia
## 61	Typhoid Fever, case
## 62	Typhus Fever
## 63	Varicella Hospitalizations
## 64	Vibrio Infection (non-Cholera)
## 65	Versiniosis
##	Classification
## 1	Non-Lethal Parasitic
## 2	Lethal Parasitic
## 3	Lethal Parasitic
## 4	Lethal Bacterial

```

## 5      Lethal Bacterial
## 6      Lethal Bacterial
## 7  Non-Lethal Bacterial
## 8  Non-Lethal Bacterial
## 9              STD
## 10     Lethal Bacterial
## 11 Non-Lethal Parasitic
## 12 Non-Lethal Parasitic
## 13              Genetic
## 14 Non-Lethal Parasitic
## 15 Non-Lethal Parasitic
## 16     Lethal Parasitic
## 17     Lethal Viral
## 18 Vaccine-Preventable
## 19 Non-Lethal Bacterial
## 20 Non-Lethal Bacterial
## 21              STD
## 22 Non-Lethal Parasitic
## 23              STD
## 24     Lethal Viral
## 25     Lethal Bacterial
## 26     Lethal Viral
## 27     Non-Lethal Viral
## 28     Non-Lethal Viral
## 29     Non-Lethal Viral
## 30              STD
## 31     Lethal Viral
## 32     Lethal Bacterial
## 33 Non-Lethal Bacterial
## 34 Non-Lethal Bacterial
## 35     Lethal Bacterial
## 36     Lethal Viral
## 37     Lethal Bacterial
## 38 Non-Lethal Bacterial
## 39 Vaccine-Preventable
## 40 Vaccine-Preventable
## 41     Lethal Bacterial
## 42 Vaccine-Preventable
## 43     Lethal Bacterial
## 44 Non-Lethal Bacterial
## 45 Non-Lethal Bacterial
## 46     Lethal Viral
## 47 Non-Lethal Bacterial
## 48     Non-Lethal Viral
## 49 Non-Lethal Bacterial
## 50 Non-Lethal Bacterial
## 51 Non-Lethal Bacterial
## 52 Non-Lethal Bacterial
## 53 Non-Lethal Bacterial
## 54     Lethal Bacterial
## 55 Non-Lethal Bacterial
## 56 Vaccine-Preventable
## 57     Lethal Bacterial
## 58 Non-Lethal Parasitic

```

```
## 59      Lethal Bacterial
## 60      Lethal Bacterial
## 61 Non-Lethal Bacterial
## 62 Non-Lethal Bacterial
## 63 Vaccine-Preventable
## 64 Non-Lethal Bacterial
## 65 Non-Lethal Bacterial
```

Now we will examine the relationship between different infectious disease features. To accomplish this, I decided to group the data by the predictors we are interested in (Number of Cases per Year, Gender, Disease Classification) and then creating the subsequent plots.

```
health_data2 <- health_data%>%
  mutate(Classification = as.factor(Classification))%>%
  filter(Sex != "Total")%>%
  group_by(Sex, Classification, Year)%>%
  mutate(Num.Cases = sum(Count))%>%
  select(Sex, Num.Cases, Year, Classification)%>%
  unique()
print("First 10 rows of the data we will use for exploratory plots")
```

```
## [1] "First 10 rows of the data we will use for exploratory plots"
```

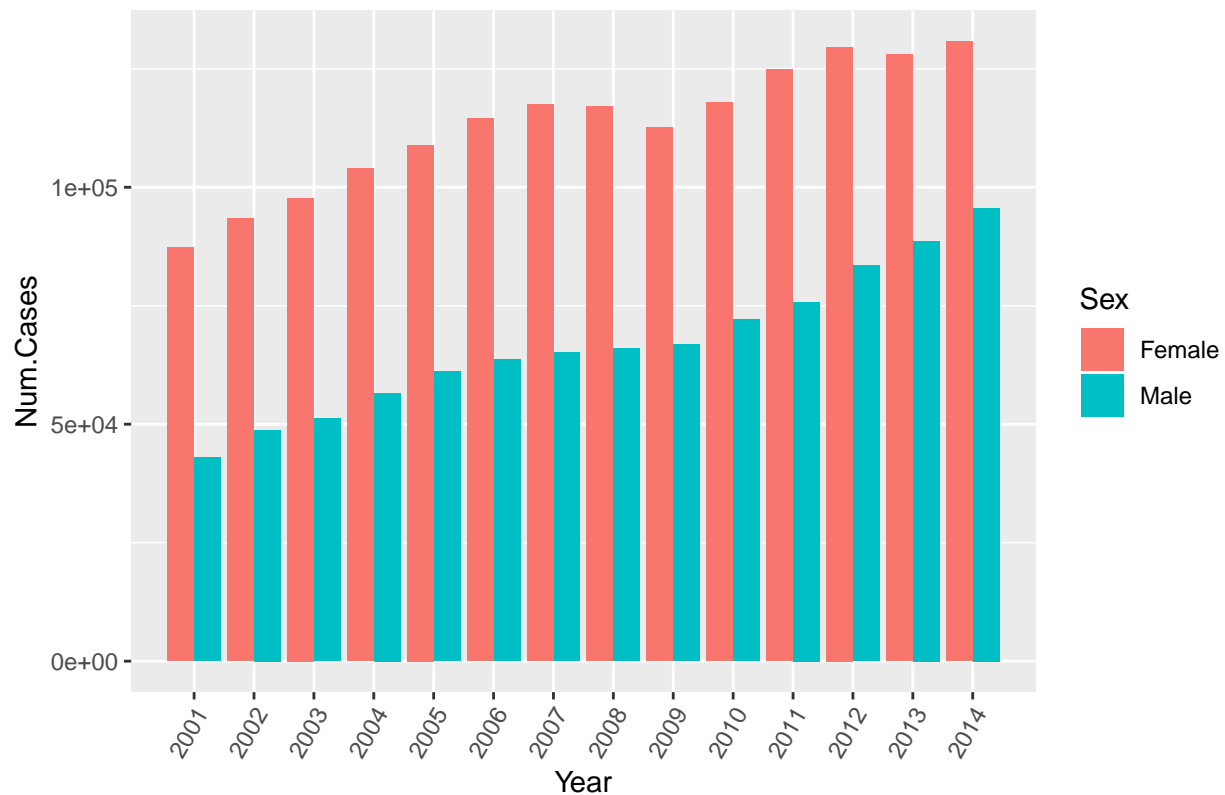
```
head(health_data2, 10)
```

```
## # A tibble: 10 x 4
## # Groups:   Sex, Classification, Year [10]
##   Sex      Num.Cases  Year Classification
##   <fct>      <int> <int> <fct>
## 1 Female      1924  2001 6
## 2 Male       3234  2001 6
## 3 Female      1774  2002 6
## 4 Male       2883  2002 6
## 5 Female      1804  2003 6
## 6 Male       2999  2003 6
## 7 Female      2086  2004 6
## 8 Male       3253  2004 6
## 9 Female      2072  2005 6
## 10 Male       3515  2005 6
```

These plots allow us to better understand the relationship between gender and different types of diseases. *Note: when comparing types of diseases I am not looking for difference between lethal and non-lethal within the same type of disease cause.*

```
#Gender vs STD
health_data2%>%
  filter(Classification == 7)%>%
  ggplot(aes(x = Year, y = Num.Cases))+
  geom_bar(aes(fill = Sex), stat = "identity", position = "dodge")+
  scale_x_discrete("Year", years, years.char, years)+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))+
  ggtitle("Number of STD Cases by Gender in California over Time")+
  xlab("Number of Observed Cases")
```

Number of STD Cases by Gender in California over Time



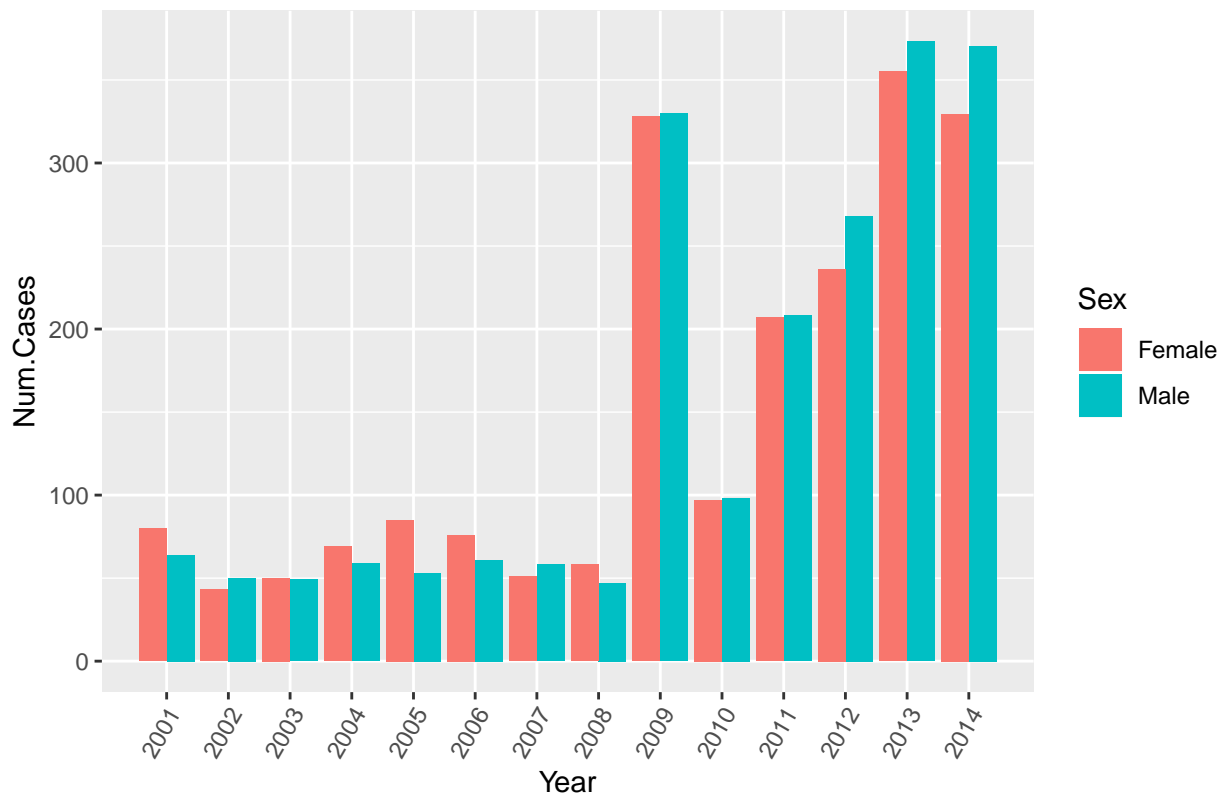
So, we can infer two things from this graph. First, we can summarize that within this data set, females have a higher rate of STD infection than do males. Second, we can observe that STD's are the most frequent class of infectious disease reported to the The California Department of Public Health (CDPH).

#Gender vs Virus

health_data2%>%

```
filter(Classification == 1 | Classification == 2)%>%
ggplot(aes(x = Year, y = Num.Cases))+
geom_bar(aes(fill = Sex), stat = "identity", position = "dodge")+
scale_x_discrete("Year", years, years.char, years)+
theme(axis.text.x = element_text(angle = 60, hjust = 1))+
ggtitle("Number of Virulent Cases by Gender in California over Time")+
xlab("Number of Observed Cases")
```

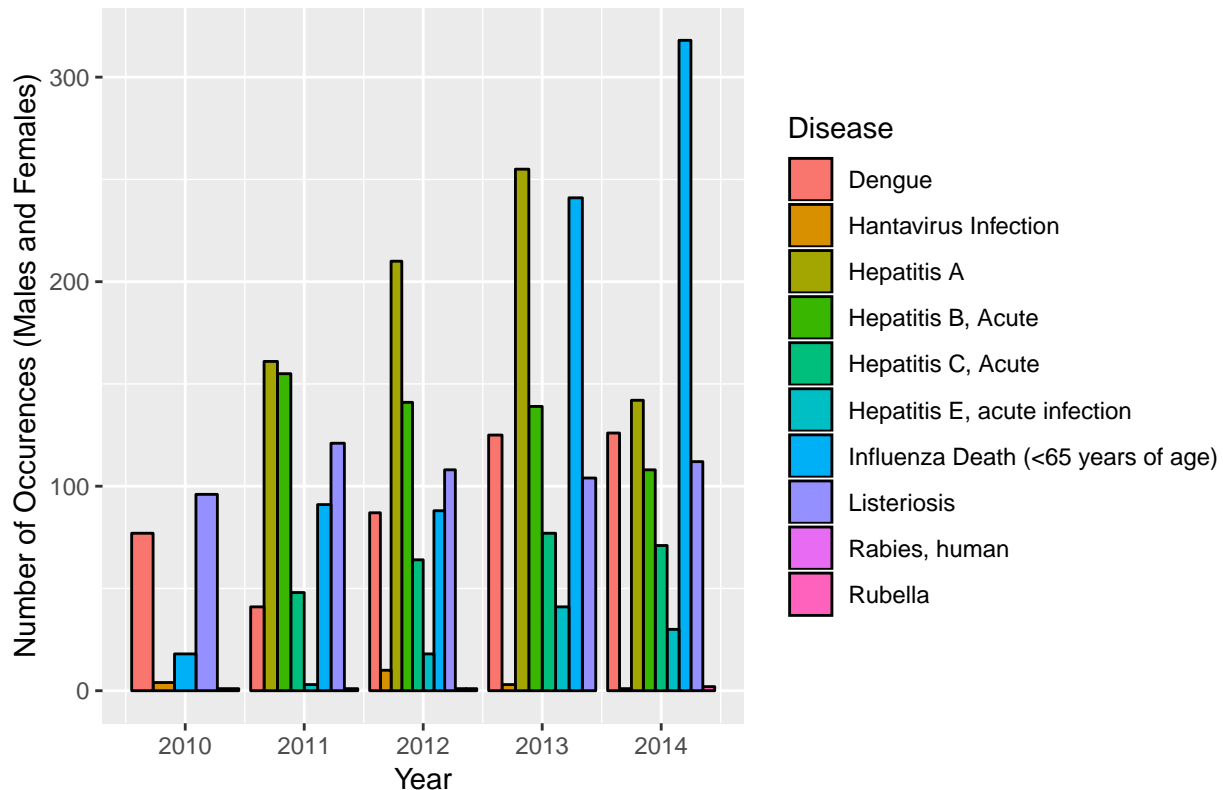
Number of Virulent Cases by Gender in California over Time



```
virusGrowth <- health_data%>%
  filter(Classification == 1 | Classification == 2)%>%
  filter(Year > 2009)%>%
  #mutate(Year = as.factor(Year))%>%
  select(Year, Disease, Count)%>%
  group_by(Disease, Year)%>%
  mutate(number_Cases = sum(Count))%>%
  filter(number_Cases > 0)%>%
  select(Year, Disease, number_Cases)%>%
  unique()%>%
  ggplot(aes(x = Year, y = number_Cases, fill = Disease))+
  geom_bar(stat = "identity", position = "dodge", colour = "black")+
  ggtitle("Virulent Infectious Disease occurenes 2010 - 2014")+
  ylab("Number of Occurences (Males and Females)")

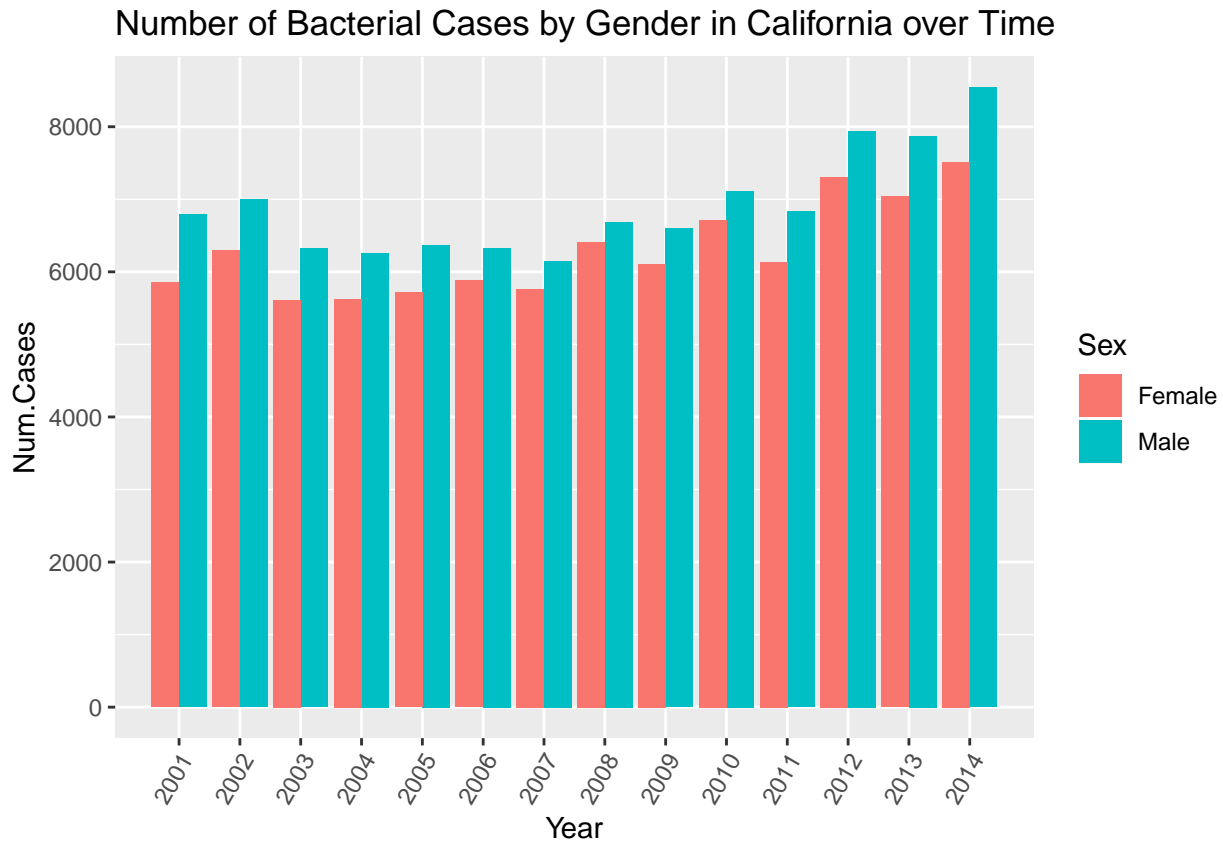
virusGrowth
```

Virulent Infectious Disease occurenes 2010 – 2014



We can infer several things from these plots. First, we observe that virulent cases do not have a strong relation to gender, that is to say there does not exist a significant difference between the number of males and females who became infected with infectious virulent diseases. Seond, we can take note of the large spike in virulent cases in 2009 reported to the The California Department of Public Health (CDPH). This is mainly caused by the outbreak of the H1N1 Flu, also known as the Swine Flu in 2009. Lastly, We can also infer that the spike in virulent cases between 2010 and 2014 is caused mainly by increased cases of Influenza Death (<65 years of age), Listeriosis, Dengue, Hepatitis A, Hepatitis B Acute, and Hepatitis C Acute.

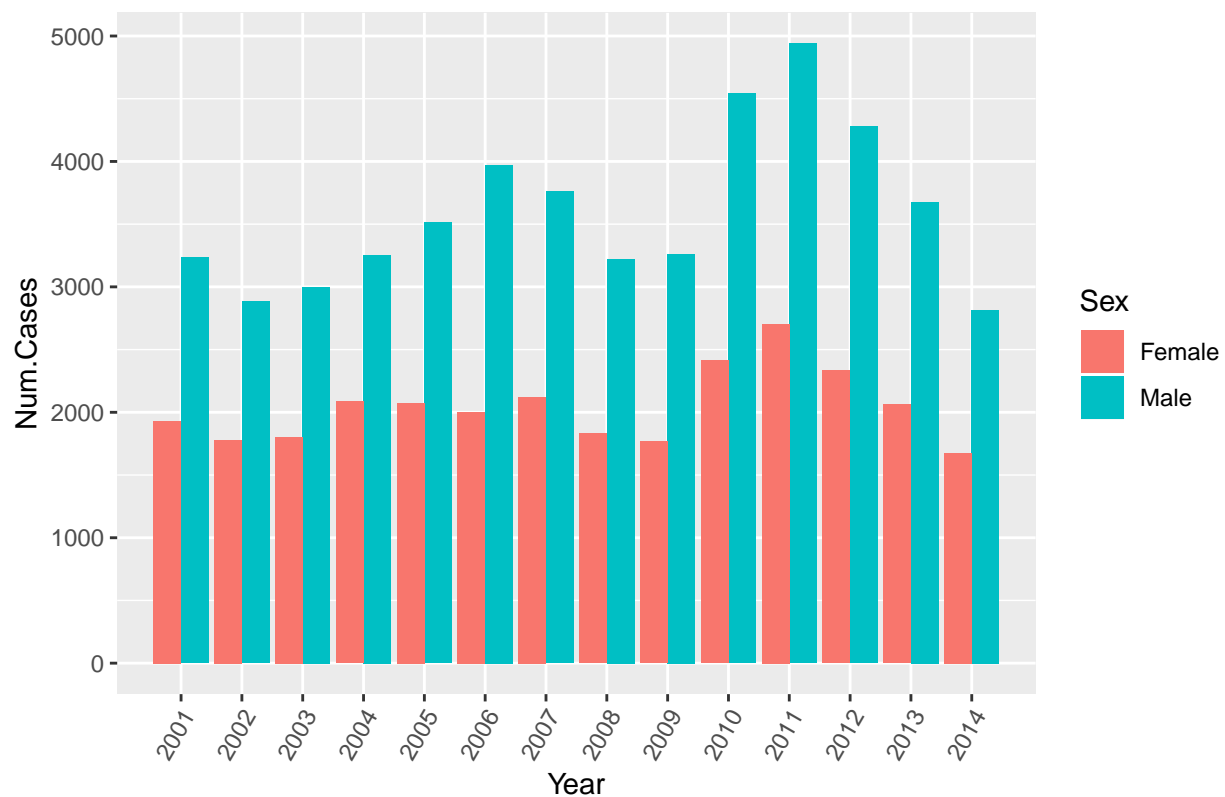
```
#Gender vs Bacteria
health_data2%>%
  filter(Classification == 3 | Classification == 4)%>%
  ggplot(aes(x = Year, y = Num.Cases))+
  geom_bar(aes(fill = Sex), stat = "identity", position = "dodge")+
  scale_x_discrete("Year", years, years.char, years)+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))+
  ggtitle("Number of Bacterial Cases by Gender in California over Time")+
  xlab("Number of Observed Cases")
```



We can infer from this plot that bacterial cases do not have a strong relation to gender, that is to say there does not exist a significant difference between the number of males and females who became infected with an infectious bacterial disease.

```
#Gender vs Parasite
health_data2%>%
  filter(Classification == 5 | Classification == 6)%>%
  ggplot(aes(x = Year, y = Num.Cases))+
  geom_bar(aes(fill = Sex), stat = "identity", position = "dodge")+
  scale_x_discrete("Year", years, years.char, years)+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))+
  ggtitle("Number of Parasitic Cases by Gender in California over Time")+
  xlab("Number of Observed Cases")
```

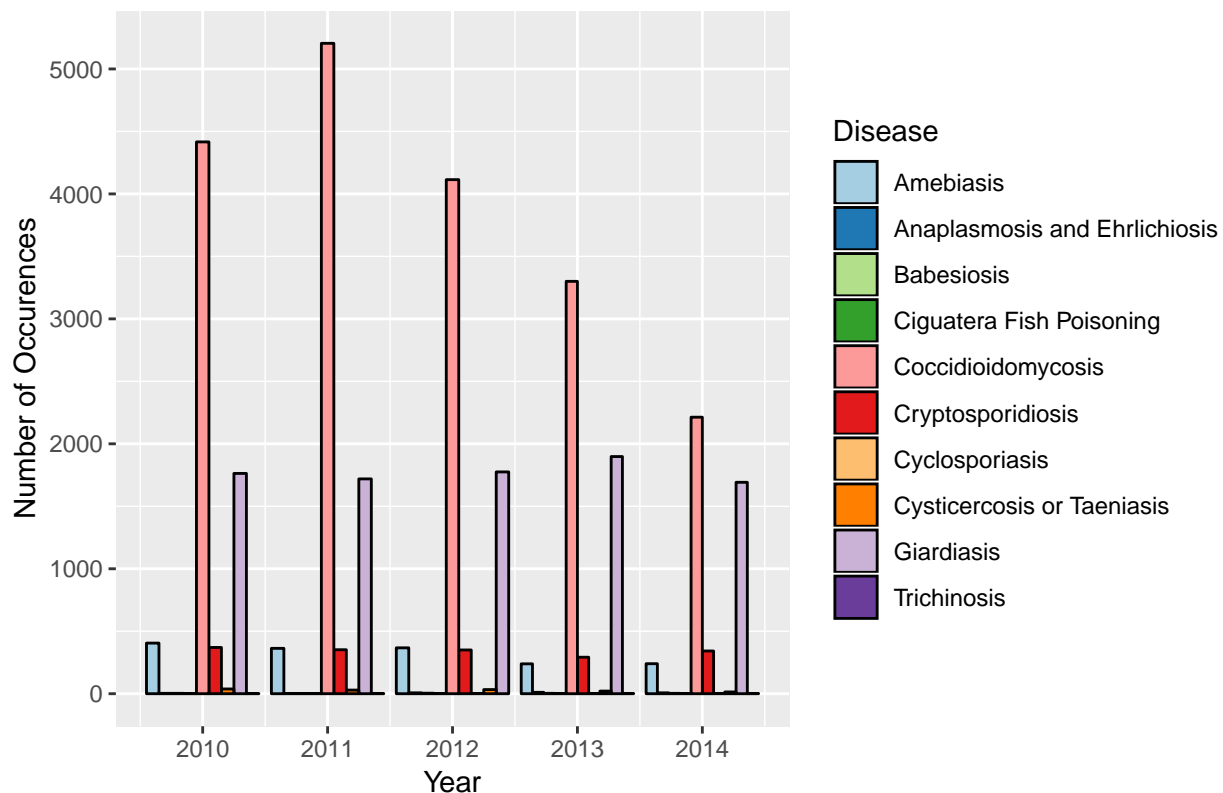
Number of Parasitic Cases by Gender in California over Time



```
paraGrowth <- health_data%>%
  filter(Classification == 5 | Classification == 6)%>%
  filter(Year > 2009)%>%
  #mutate(Year = as.factor(Year))%>%
  select(Year, Disease, Count)%>%
  group_by(Disease, Year)%>%
  mutate(number_Cases = sum(Count))%>%
  filter(number_Cases > 0)%>%
  select(Year, Disease, number_Cases)%>%
  unique()%>%
  ggplot(aes(x = Year, y = number_Cases, fill = Disease))+
  geom_bar(stat = "identity", position = "dodge", colour = "black")+
  ggtitle("Parasitic Infectious Diseases between 2010 - 2014")+
  scale_fill_brewer(palette="Paired")+
  ylab("Number of Occurences")

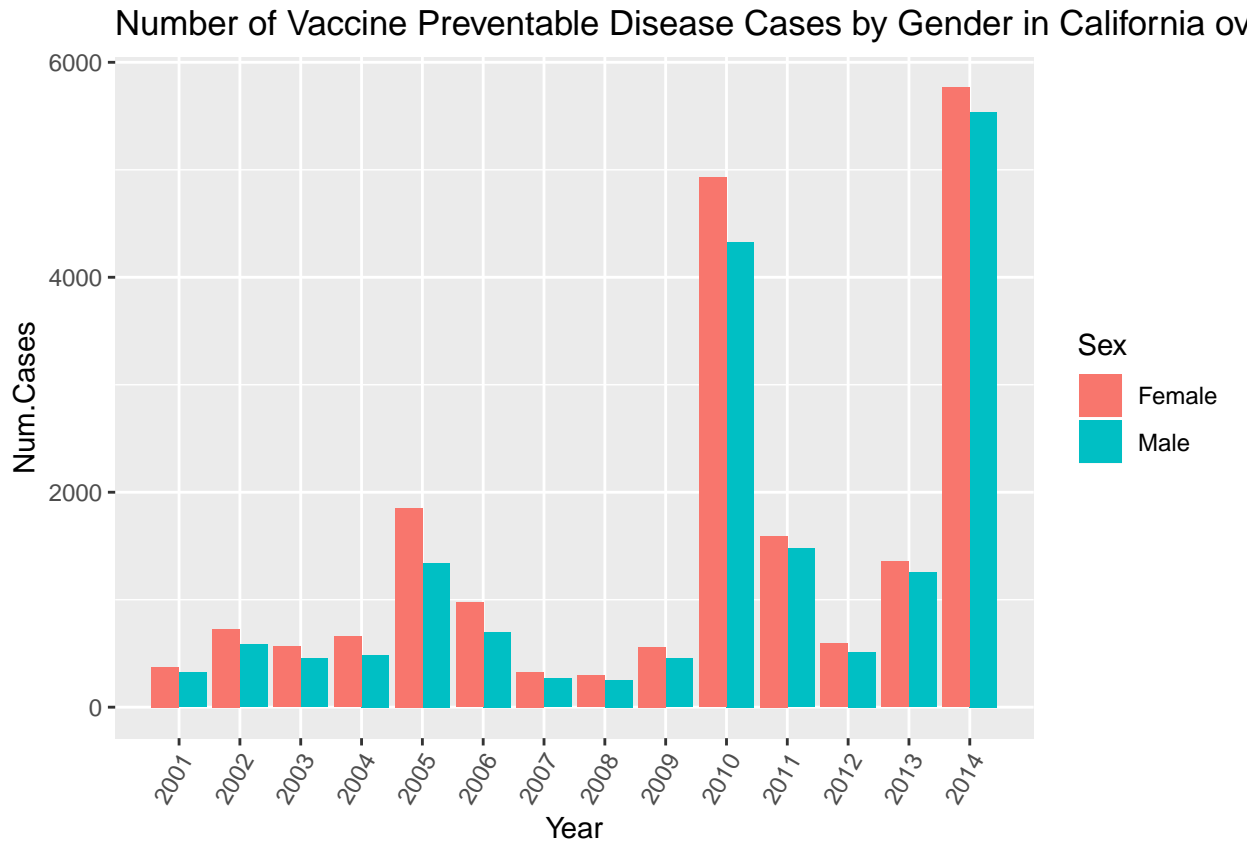
paraGrowth
```


Parasitic Infectious Diseases between 2010 – 2014



We can infer several things from these plots on parasites. First, we observe that parasitic cases do have a relation to gender, that is to say there does exist a significant difference between the number of males and females who became infected with infectious parasitic diseases. Second, we can take note of the growth in parasitic cases between 2010 and 2014 reported to the The California Department of Public Health (CDPH). This is mainly caused by bacteria delivered from parasites coccidioidomycosis, also called Valley Fever, an infection caused by the fungus *Coccidioides*. According to the CDC, *Coccidioides* lives in dust and soil in some areas in the southwestern United States, Mexico, and South America. In the United States, *Coccidioides* lives in Arizona, California, Nevada, New Mexico, Texas, and Utah. So, I would speculate that the California Wildfires are a factor in the increased cases of Valley Fever. [https://www.cdc.gov/fungal/diseases/coccidioidomycosis/index.html]. Second main cause of growth is by the parasite giardiasis which causes Giardia. According to the Mayo Clinic, Giardia infection is an intestinal infection marked by abdominal cramps, bloating, nausea and bouts of watery diarrhea. Giardia infection is caused by a microscopic parasite that is found worldwide, especially in areas with poor sanitation and unsafe water. [https://www.mayoclinic.org/diseases-conditions/giardia-infection/symptoms-causes/syc-20372786]. I would speculate, it could be the case that climate change contributing to larger storms, which can cause floods and large surf conditions, could be a possible cause for the increases in cases of Giardia.

```
#Gender vs Vaccinated Diseases
health_data2 %>%
  filter(Classification == 8) %>%
  ggplot(aes(x = Year, y = Num.Cases)) +
  geom_bar(aes(fill = Sex), stat = "identity", position = "dodge") +
  scale_x_discrete("Year", years, years.char, years) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1)) +
  ggtitle("Number of Vaccine Preventable Disease Cases by Gender in California over Time") +
  xlab("Number of Observed Cases")
```



This plot shows there does not exist a significant difference between the number of females and males who become infected with infectious diseases for which a vaccine exists. Both of the large spikes are caused by outbreaks of whooping cough in California. In 2010, a whooping cough outbreak in California sickened 9,120 people, more than in any year since 1947.[<https://www.npr.org/sections/health-shots/2013/09/25/226147147/vaccine-refusals-fueled-californias-whooping-cough-epidemic>] In the paper, “*Nonmedical Vaccine Exemptions and Pertussis in California, 2010*” written by Jessica E. Atwell, Josh Van Otterloo, Jennifer Zipprich, Kathleen Winter, Kathleen Harriman, Daniel A. Salmon, Neal A. Halsey, Saad B. Omer of Johns Hopkins Bloomberg School of Public Health found that people who lived in areas with high rates of personal belief exemptions were 2 1/2 times more likely to live in a place with lots of pertussis cases.[<http://pediatrics.aappublications.org/content/132/4/624.abstract>]. Which is to say this outbreak can be explained by people not wanting to vaccinate their children. Again in 2014, the California Department of Public Health (CDPH) declared that a pertussis epidemic was occurring in the stated that the incidence of pertussis in the United States is cyclical, with peaks every 3–5 years, as the number of susceptible persons in the population increases. Additionally, they report that [the] “CDPH is working with local public health departments as well as prenatal and pediatric health care providers, with the primary goal of encouraging vaccination of pregnant women and infants. In addition, CDPH is providing free Tdap to local health departments and community health centers to support vaccination of uninsured and underinsured pregnant women and is working to identify and mitigate barriers to Tdap vaccination for pregnant women.”[<https://www.cdc.gov/mmWr/preview/mmwrhtml/mm6348a2.htm>] So, one could speculate as healthcare costs continue to be inaffordable to many people and the * anti-vaxx* movement remains, the whooping cough will continue to cycle with an outbreak every 3-5 years.

Machine Learning

In this portion of the analysis, we wish to Predict the gender of an infected person disease based on the disease class they have become infected with and the number of people who contracted the disease of class i

in year j using a Multinomial Logistic Regression. More formally,

$$Y_i = \text{Gender of person } i$$

d_i = disease classification of person-i, where

i = Lethal Virulent, Non-Lethal Virulent, Lethal Bacterial, ... , Genetic

t_{ij} = number of cases of disease-i in year-j

First we will separate the data into the train and test splits, so that we can evaluate our model after we create it.

I am deciding to use a 60:40 ratio to split the data. So then, 60% of the observations will be used to train the model and 40% of the observations will be used to test the model.

#Examine Data

```
health_ML <- health_data%>%
  group_by(Year,Classification,Sex)%>%
  mutate(Year.Count = sum(Count))%>%
  select(Year,Classification,Year.Count,Sex)%>%
  filter(Sex != "Total")%>%
  unique()%>%
  ungroup()

#Rename classes to reflect what the classifications mean:
levels(health_ML$Classification) <- c("Lethal Viral", "Non-Lethal Viral", "Lethal Bacterial", "Non-Lethal Bacterial")
#Examine Data that will be used for prediction:
head(health_ML)
```

```
## # A tibble: 6 x 4
##   Year Classification      Year.Count Sex
##   <int> <fct>              <int> <fct>
## 1  2001 Non-Lethal Parasitic      1924 Female
## 2  2001 Non-Lethal Parasitic      3234 Male
## 3  2002 Non-Lethal Parasitic      1774 Female
## 4  2002 Non-Lethal Parasitic      2883 Male
## 5  2003 Non-Lethal Parasitic      1804 Female
## 6  2003 Non-Lethal Parasitic      2999 Male
```

#We only need the predictors

#Establish cut-off point

```
cutoff <- round(0.6*dim(health_ML)[1],0)
```

#Establish training and testing sets by randomly sampling from the data with the above ratios for test

#Set seed for reproducibility

```
set.seed(3)
```

```
training_indexes <- sample.int(nrow(health_ML), size = cutoff, replace = FALSE, prob = NULL)
```

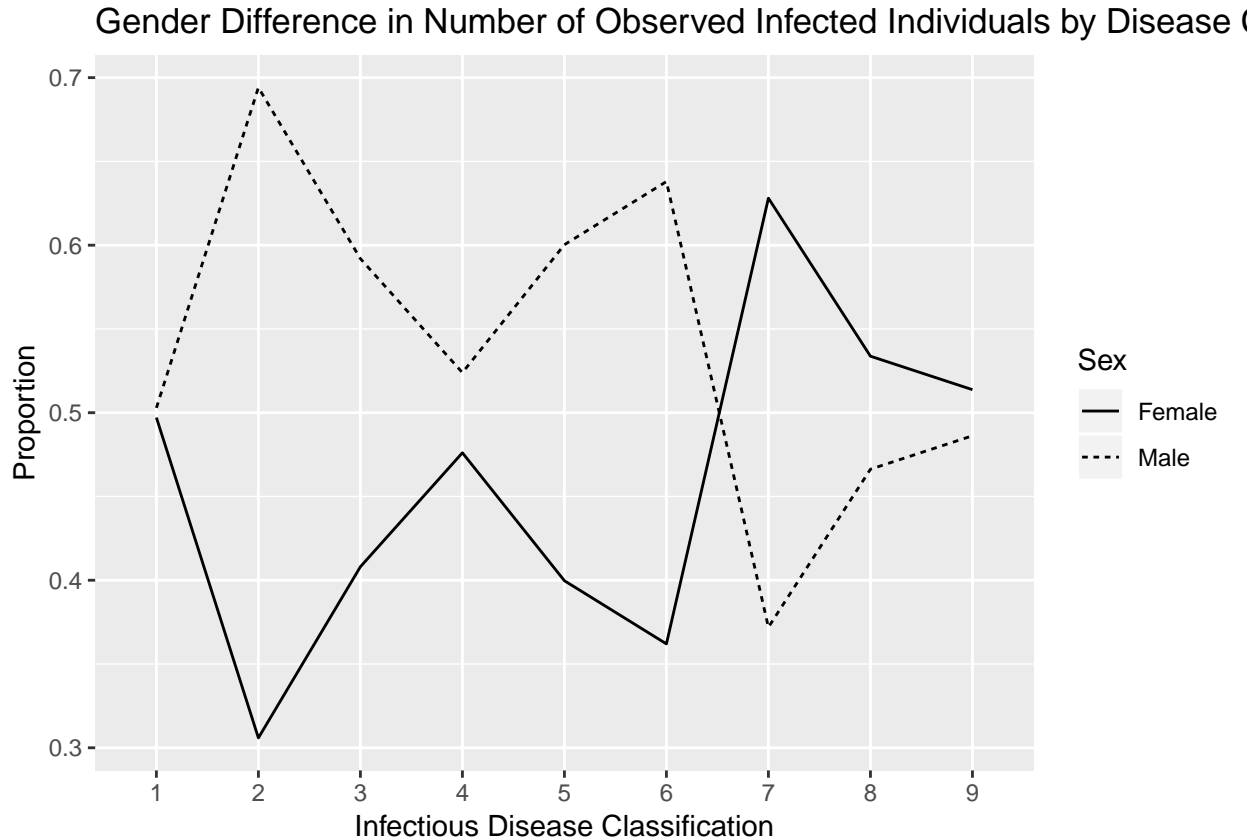
```
health_train <- health_ML[training_indexes,]
```

```
health_test <- health_ML[-training_indexes,]
```

**** Gaining Intuition on Model 1.****

```
genderProportions <- group_by(health_data,Classification,Sex)%>%
  summarise(count = sum(Count))%>%
  group_by(Classification)%>%
  mutate(classTotals = sum(count),proportion = count/classTotals)
```

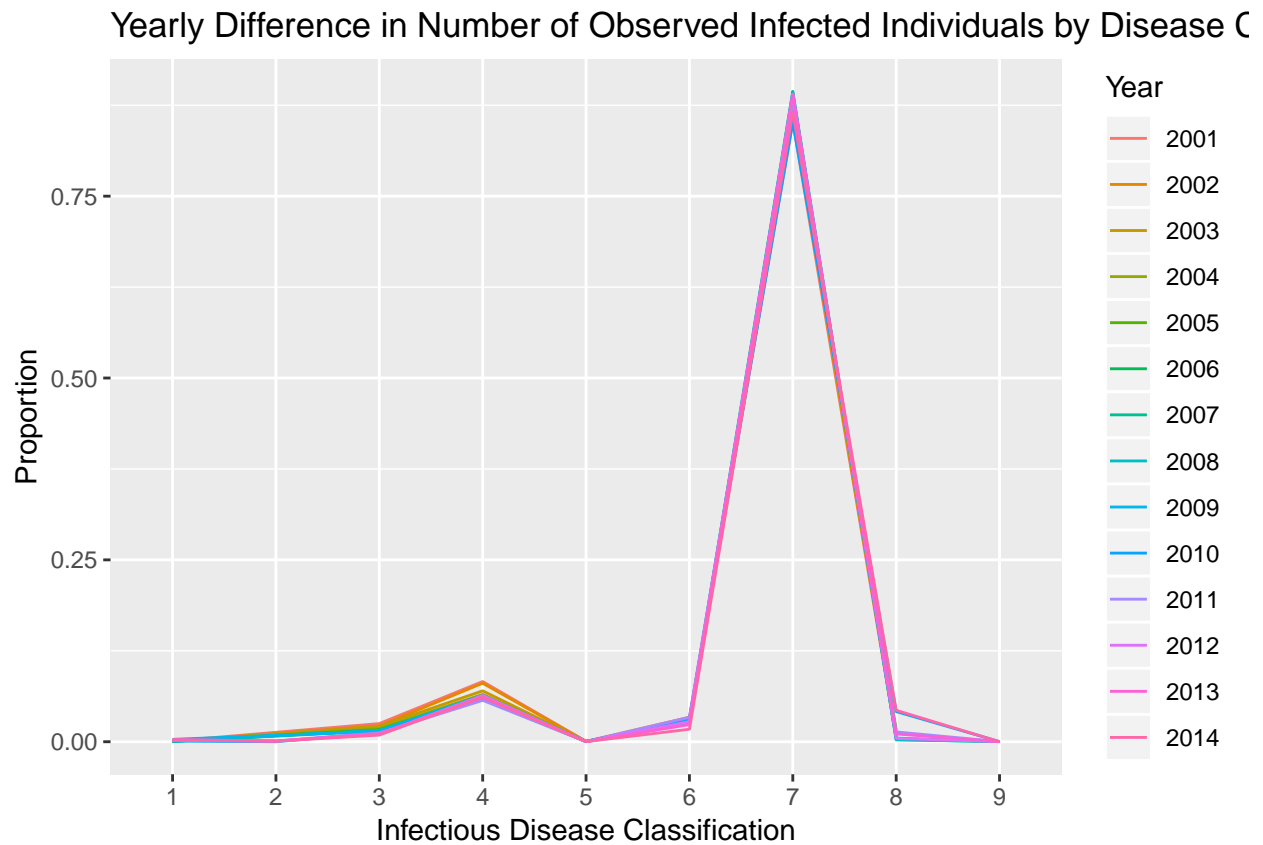
```
ggplot(genderProportions, aes(x = Classification, y = proportion, group = Sex, linetype = Sex))+
  geom_line()+
  xlab("Infectious Disease Classification")+
  ylab("Proportion")+
  ggtitle("Gender Difference in Number of Observed Infected Individuals by Disease Class")
```



We observe that there does exist difference in the number of instances of a certain class of infectious diseases for males and females.

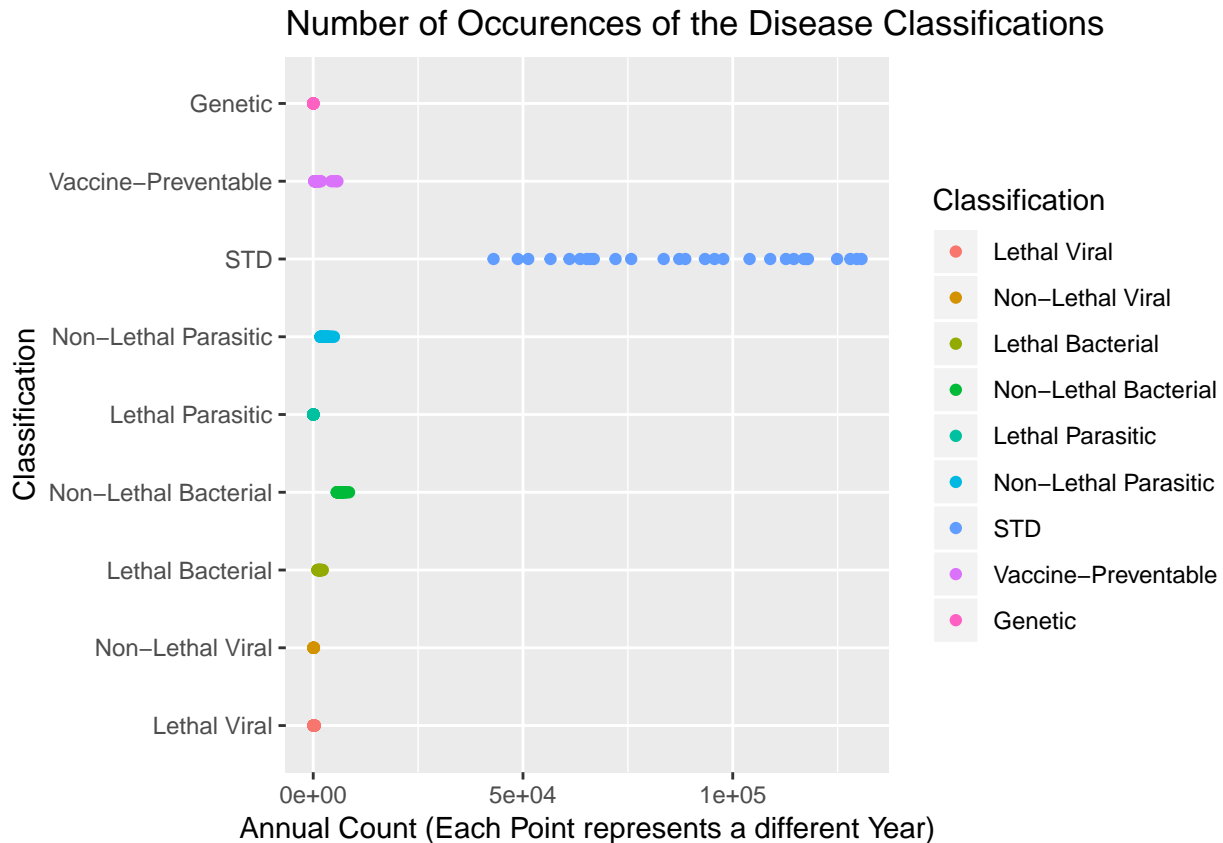
```
timeProportions <- health_data%>%
  mutate(Year = as.factor(Year))%>%
  group_by(Classification, Year)%>%
  summarise(count = sum(Count))%>%
  group_by(Year)%>%
  mutate(timeTotals = sum(count), proportion = count/timeTotals)

ggplot(timeProportions, aes(x = Classification, y = proportion, group = Year, colour = Year))+
  geom_line()+
  xlab("Infectious Disease Classification")+
  ylab("Proportion")+
  ggtitle("Yearly Difference in Number of Observed Infected Individuals by Disease Class")
```



We observe that there does not exist a difference in the number of instances of a certain infectious disease by year.

```
health_ML%>%
  ggplot(aes(x = Classification, y = Year.Count, colour = Classification))+
  geom_point()+
  ggtitle("Number of Occurences of the Disease Classifications")+
  ylab("Annual Count (Each Point represents a different Year))+
  coord_flip()
```

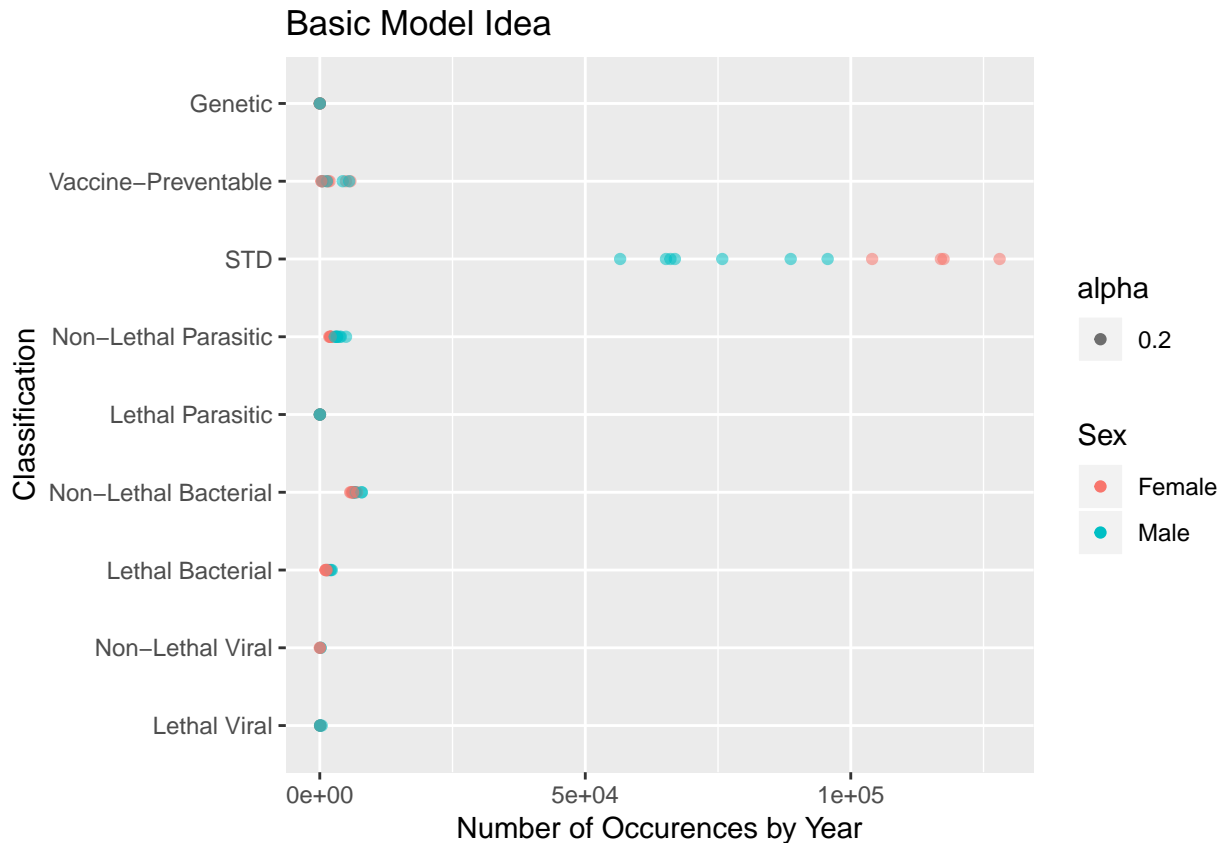


```
#Examine what proportion of the data is an STD
is.STD <- paste(round(table(health_ML$Classification == "STD")/length(health_ML$Classification),2)*100,
names(is.STD) <- c("Percentage STD", "Percentage of Non-STD")
```

So, we do observe that the number of occurrences of a certain class of infectious diseases does correlate to the number of occurrences of infection. However, we do observe a large number of the classifications are classified as STD's. In fact, 87 %, 13 % is what we have, with about 87% of all data points being of classification STD.

Next we will predict whether a given individual is a male or a female as a function of the class of disease the person has contracted and the number of people who had that disease that year.

```
ggplot(health_test, aes(x = Year.Count, y = Classification, colour = Sex))+
  geom_point(aes(alpha = 0.2))+
  xlab("Number of Occurences by Year")+
  ggtitle("Basic Model Idea")
```



So, we see that for STD's we would expect that the model will classify the gender of infected individuals well and for diseases with less observations we would expect the model to have less accuracy. And since 87% of the values have classification STD, and STD's tend to be more frequently observed in females, it is likely our model will be biased towards predicting the gender of an unobserved person as female.

Modeling

#Create Model:

```
multiLogitModelAll <- multinom(Sex ~ Year + Classification + Year.Count, health_train)
```

```
## Warning in multinom(Sex ~ Year + Classification + Year.Count,
## health_train): group 'Total' is empty
```

```
## # weights: 12 (11 variable)
```

```
## initial value 92.188575
```

```
## iter 10 value 82.197565
```

```
## iter 20 value 82.056249
```

```
## iter 30 value 82.054275
```

```
## iter 40 value 82.053735
```

```
## final value 82.039055
```

```
## converged
```

#Use Akaike Information Criterion (AIC) to select a model. This will penalize unnecessary predictor variables

```
modelAIC <- step(multiLogitModelAll)
```

```
## Start: AIC=186.08
```

```
## Sex ~ Year + Classification + Year.Count
```

```
##
```

```
## trying - Year
```

```

## Warning in multinom(formula = Sex ~ Classification + Year.Count, data =
## health_train, : group 'Total' is empty

## # weights:  11 (10 variable)
## initial  value 92.188575
## iter   10 value 82.279611
## final   value 82.055801
## converged
## trying - Classification

## Warning in multinom(formula = Sex ~ Year + Year.Count, data =
## health_train, : group 'Total' is empty

## # weights:  4 (3 variable)
## initial  value 92.188575
## iter   10 value 90.302139
## iter   20 value 90.295723
## iter   30 value 90.274851
## iter   40 value 90.252727
## iter   50 value 90.216433
## iter   60 value 90.180873
## iter   60 value 90.180873
## iter   60 value 90.180873
## final   value 90.180873
## converged
## trying - Year.Count

## Warning in multinom(formula = Sex ~ Year + Classification, data =
## health_train, : group 'Total' is empty

## # weights:  11 (10 variable)
## initial  value 92.188575
## iter   10 value 90.243998
## iter   20 value 90.221099
## iter   30 value 90.121533
## iter   40 value 90.120748
## final   value 90.113551
## converged
##
##           Df      AIC
## - Year      10 184.1116
## <none>      11 186.0781
## - Classification  3 186.3617
## - Year.Count    10 200.2271

## Warning in multinom(formula = Sex ~ Classification + Year.Count, data =
## health_train): group 'Total' is empty

## # weights:  11 (10 variable)
## initial  value 92.188575
## iter   10 value 82.279611
## final   value 82.055801
## converged
##
## Step:  AIC=184.11
## Sex ~ Classification + Year.Count
##
## trying - Classification

```



```
## Warning in multinom(formula = Sex ~ Year.Count, data = health_train, trace
## = trace): group 'Total' is empty

## # weights:  3 (2 variable)
## initial  value 92.188575
## final    value 90.301867
## converged
## trying - Year.Count

## Warning in multinom(formula = Sex ~ Classification, data = health_train, :
## group 'Total' is empty

## # weights:  10 (9 variable)
## initial  value 92.188575
## iter 10 value 90.244943
## final    value 90.244473
## converged
##
##              Df      AIC
## <none>         10 184.1116
## - Classification  2 184.6037
## - Year.Count     9 198.4889
```

#Examine the Model's Coefficients

```
summary(modelAIC)
```

```
## Call:
## multinom(formula = Sex ~ Classification + Year.Count, data = health_train)
##
## Coefficients:
##
##              Values      Std. Err.
## (Intercept)    -0.2640544496 1.852400e-10
## ClassificationNon-Lethal Viral    0.2774674055 1.975710e-13
## ClassificationLethal Bacterial    0.8494293876 9.253841e-12
## ClassificationNon-Lethal Bacterial 1.0852513779 3.081728e-11
## ClassificationLethal Parasitic    0.2677838163 1.831272e-13
## ClassificationNon-Lethal Parasitic 0.3817726132 1.354341e-11
## ClassificationSTD    12.5848294633 1.265637e-10
## ClassificationVaccine-Preventable 0.8681123575 3.442965e-12
## ClassificationGenetic    0.9595100832 4.527455e-14
## Year.Count    -0.0001470738 1.130494e-05
##
## Residual Deviance: 164.1116
## AIC: 184.1116
```

Examining the Performance of our Multinomial Logistic Regression Model.

#Split test set into true values and predictors which we will use on our model to evaluate its performance

```
health_test.X <- health_test%>%
  select(Classification, Year.Count)
health_test.Y <- health_test%>%
  select(Sex)
```

#Predict the gender of individuals in our test set the Multinomial Logistic Regression Model

```
predictedSex <- predict(modelAIC, health_test)
```

#Examining Model Performance:

```

confusion_matrix <- matrix(nrow = 2, ncol = 2,0)
colnames(confusion_matrix) <- c("Observed Male", "Observed Female")
rownames(confusion_matrix) <- c("Precited Male", "Predicted Female")

for(i in predictedSex){
  for(j in health_test.Y$Sex){
    if(i == "Male" & j == "Male"){
      confusion_matrix[1,1] <- confusion_matrix[1,1] + 1
    }
    if(i == "Female" & j == "Male"){
      confusion_matrix[2,1] <- confusion_matrix[2,1] + 1
    }
    if(i == "Male" & j == "Female"){
      confusion_matrix[1,2] <- confusion_matrix[1,2] + 1
    }
    if(i == "Female" & j == "Female"){
      confusion_matrix[2,2] <- confusion_matrix[2,2] + 1
    }
  }
}

#Plot results of prediction
data.frame(Predicted = predictedSex, Observed = health_test.Y$Sex)%>%
  mutate()

```

```

##      Predicted Observed
## 1      Female   Female
## 2      Female   Female
## 3      Female    Male
## 4      Female   Female
## 5      Female    Male
## 6      Female   Female
## 7      Female    Male
## 8      Female   Female
## 9      Female    Male
## 10     Female    Male
## 11     Female    Male
## 12     Female    Male
## 13     Female   Female
## 14     Female    Male
## 15     Female   Female
## 16     Female    Male
## 17     Female   Female
## 18      Male    Female
## 19      Male    Male
## 20      Male    Male
## 21      Male   Female
## 22      Male    Male
## 23      Male   Female
## 24      Male   Female
## 25      Male    Male
## 26      Male    Male
## 27      Male   Female
## 28      Male   Female

```

## 29	Male	Male
## 30	Male	Male
## 31	Male	Female
## 32	Male	Female
## 33	Male	Female
## 34	Female	Male
## 35	Female	Female
## 36	Female	Male
## 37	Female	Female
## 38	Female	Male
## 39	Female	Female
## 40	Female	Male
## 41	Female	Female
## 42	Female	Male
## 43	Female	Female
## 44	Female	Male
## 45	Female	Male
## 46	Female	Female
## 47	Female	Male
## 48	Female	Female
## 49	Female	Female
## 50	Male	Male
## 51	Female	Female
## 52	Male	Male
## 53	Female	Female
## 54	Male	Male
## 55	Male	Male
## 56	Male	Male
## 57	Female	Female
## 58	Female	Male
## 59	Female	Male
## 60	Female	Male
## 61	Female	Female
## 62	Female	Male
## 63	Male	Female
## 64	Male	Female
## 65	Male	Female
## 66	Male	Female
## 67	Male	Male
## 68	Male	Female
## 69	Male	Male
## 70	Female	Male
## 71	Female	Male
## 72	Female	Female
## 73	Female	Male
## 74	Female	Female
## 75	Female	Male
## 76	Male	Female
## 77	Male	Female
## 78	Male	Female
## 79	Female	Female
## 80	Female	Male
## 81	Male	Female
## 82	Male	Male

```
## 83      Male   Female
## 84      Male     Male
## 85      Male   Female
## 86   Female     Male
## 87   Female     Male
## 88      Male   Female
## 89      Male   Female
```

```
#Display Confusion Matrix
confusion_matrix
```

```
##              Observed Male Observed Female
## Predicted Male           1672           1710
## Predicted Female        2244           2295
```

```
#calculate performance metrics
```

```
recall <- confusion_matrix[1,1]/(confusion_matrix[1,1]+confusion_matrix[1,2])
precision <- confusion_matrix[1,1]/(confusion_matrix[1,1]+confusion_matrix[2,1])
accuracy <- (confusion_matrix[1,1] + confusion_matrix[2,2])/(confusion_matrix[1,1]+confusion_matrix[1,2]
```

A confusion matrix is a tool that will allow us to better quantify how well our model actually classifies. There does not exist any inherent positive or negative class for gender. Unfortunately, the confusion matrix is usually in the context of classifying the status of an event happening, in which case it is useful to think of the event you are classifying as positive (the event did happen) versus negative (the event did not happen). - If one predicts positive and the true value is a positive this is called a true positive. - If one predicts positive and the true value is a negative this is called a false positive. - If one predicts negative and the true value is a positive this is called a true negative. - If one predicts negative and the true value is a negative this is called a false negative.

From the confusion matrix, one can create metrics that gauge the performance of the model, mainly Precision, Recall, and Accuracy.

$$\text{Precision} = \frac{\text{True Positive Predictions}}{\text{Observed Results}}$$

$$\text{Recall} = \frac{\text{True Positive Predictions}}{\text{Predicted Results}}$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative Predictions}}{\text{Total}}$$

Precision means the percentage of your results which are relevant. On the other hand, recall refers to the percentage of total relevant results correctly classified by your algorithm. [https://towardsdatascience.com/precision-vs-recall-386cf9f89488] And Accuracy measures how well our model predicts values. Here we have that: * Recall = 0.494382 * Precision = 0.4269663 * Accuracy = 0.5008206

These results make sense since we have a much higher number of observations of females with STD's than males with STD's, and given how common observations of STD's are in our data, it makes sense that the model has a bias towards predicting Sex as female which drives down the accuracy.

Next Steps:

Due to the low accuracy score, steps should be taken in the future to make the model perform better.

1. Get more data. After the processing and grouping we were left with only 222 values, and after splitting the data into train and test sets, the model is likely not able to make enough of a fit with the amount of data I am working with.
2. Implement a model that has a higher ability to fit data with less observations, possibly K-Nearest Neighbors or a Hierarchical Clustering Tree.

3. Build a model that aims to predict the class of disease based on gender and number of occurrences. Possibly, a k-means clustering analysis or a bayesian hierarchal model in which we specify prior parameters about the distribution of different classes of diseases.