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 occurences 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/")
install.packages("reshape2",repos = "https://cloud.r-project.org/")
devtools::install_github("dgrtwo/gganimate",repos = "https://cloud.r-project.org/")
```

```
## Downloading GitHub repo dgrtwo/gganimate@master
```

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

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

```
library(tidyverse)
library(ggridges)
library(tseries)
library(cluster)
library(maps)
library(mapdata)
library(ggmap)
library(reshape2)
library(gganimate)
library(nnet)
library(ROCR)
library(caret)
```

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.

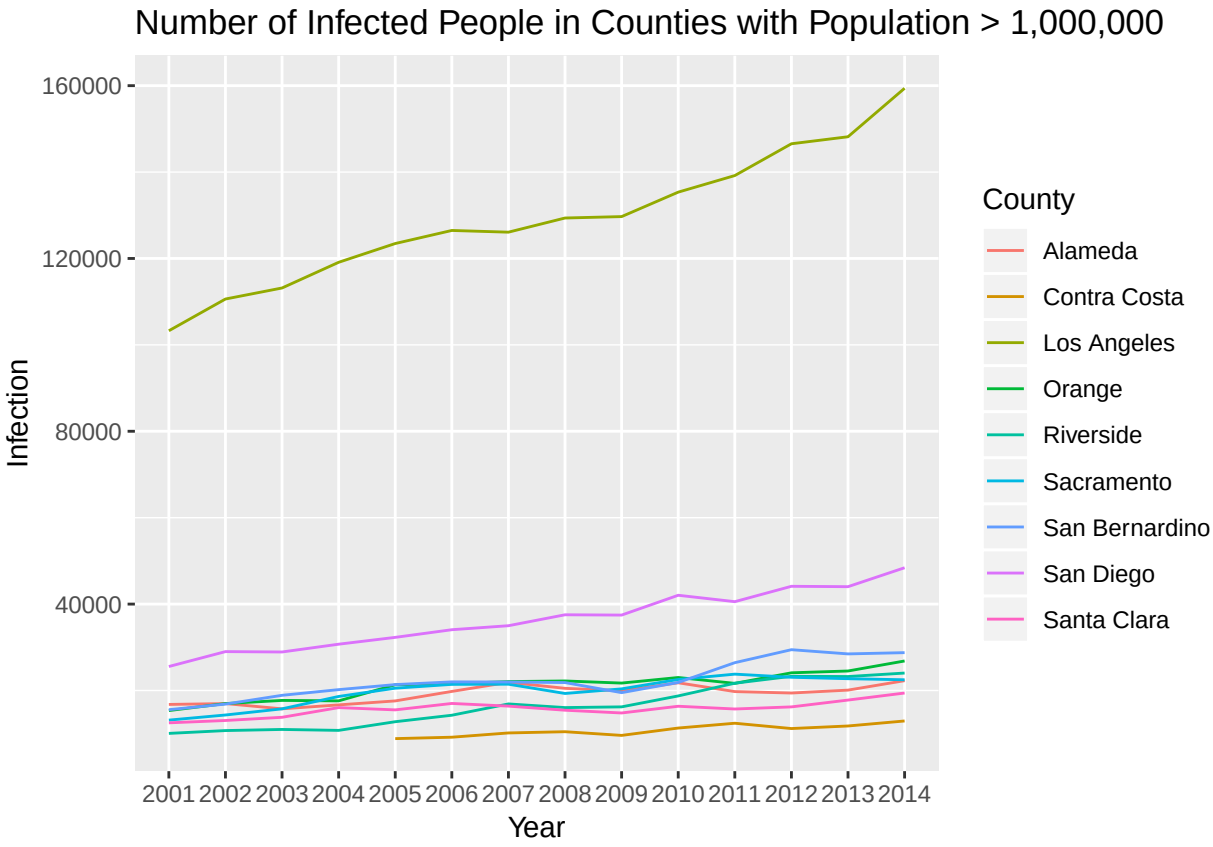
```
setwd("/Users/DanielsMac/Desktop/SBHacks")
#load data
health_data <- read.csv("rows.csv")
head(health_data)
```

```
##     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
```

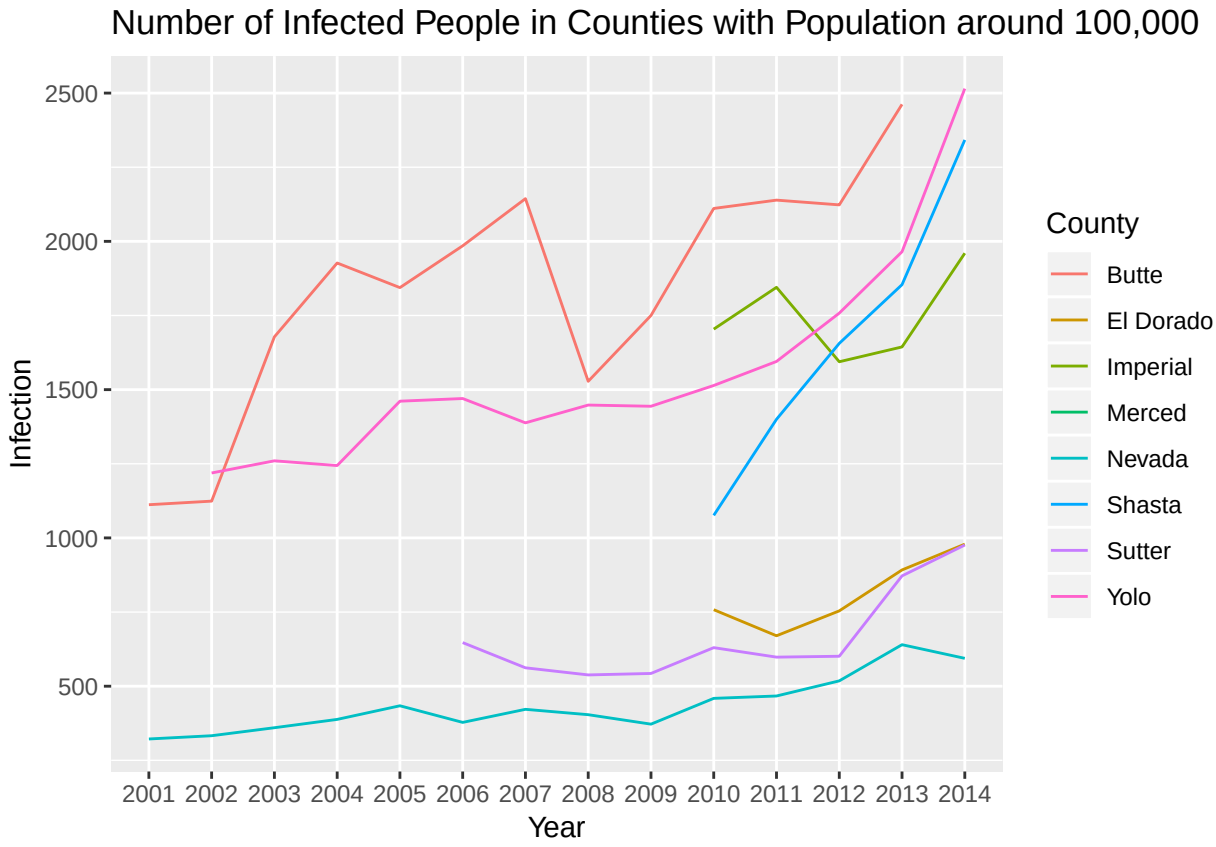
```
## 3    2.355
## 4    1.389
## 5    0.972
## 6    1.805
```

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 variablity, but overall there is an upward trend to the number of incidents of infectious diseases over time.

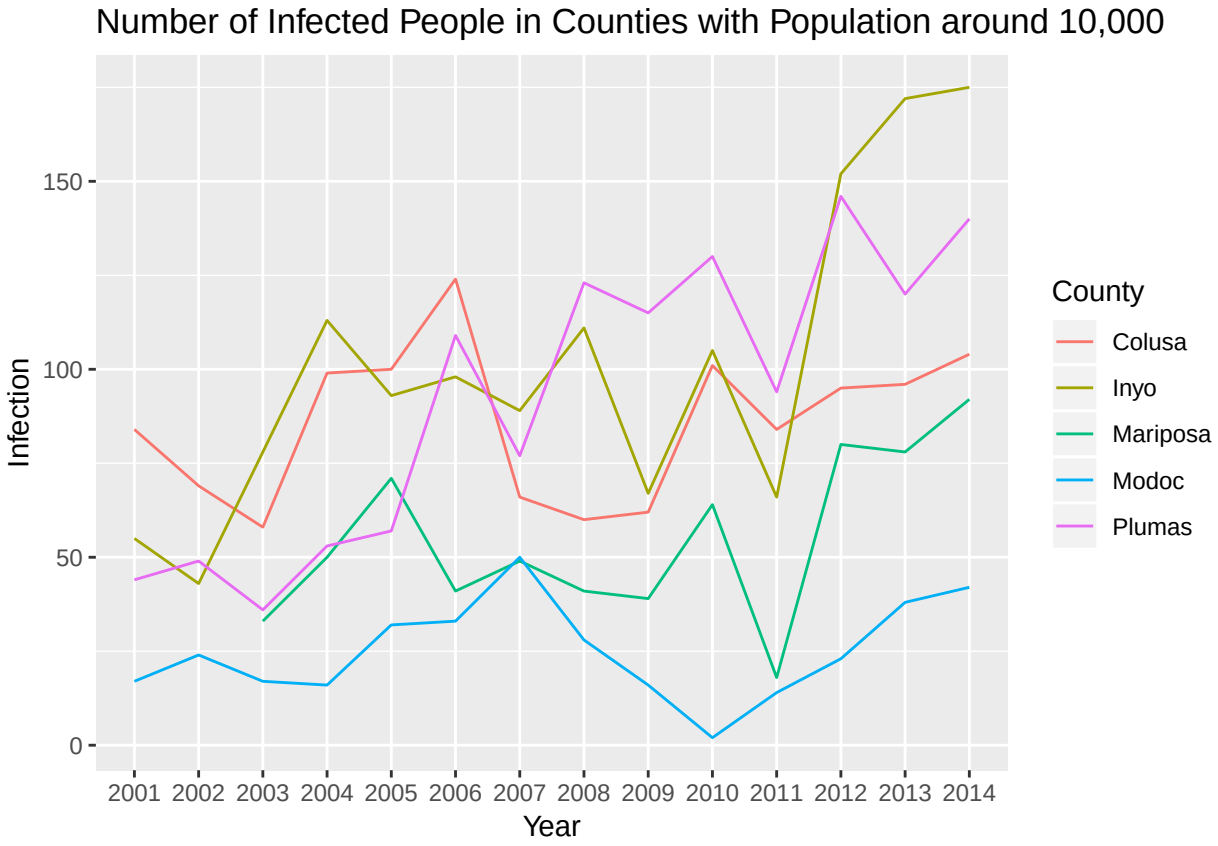
```
#Get list of years present in data set
years <- c(min(health_data$Year):max(health_data$Year))
years.char <- as.character(years)

#Plots to unerstand 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)
```



```r
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)
```




```r
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)
```



Relationsips between Variables 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 classificaion. - 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 parasitic = disease is potentially lethal and caused by parasites (fungus included) - Non-Lethal Parastic = 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")

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-Le
```

In total, there were 65 different diseases in the dataset. To reduce dimmensions of the classification problem in later parts of this analysis I hand-classified each disease into a category based on the causing factor of the disease. For example, Chalmydia is classified as STD, and Babesiosis a disease caused from tick bites is classified as Parasitic. Beyond this, the diseases were classified into whether or not they are lethal. My basis of this was if the CDC of WHO said normal recovery involved little to no treatment, or if lethal cases were very rare then I stated the disease was non-lethal. Similarly, if the CDC or WHO stated treatment was needed to prevent death or severe bodily harm, I classified the disease as lethal. After this, I got the following categories:

	Classification
1	Lethal Viral
2	Non-Lethal Viral
3	Lethal Bacterial
4	Non-Lethal Bacterial
5	Lethal Parasitic (fungus included)
6	Non-Lethal Parastic
7	STD
8	Vaccine-Preventable (person who contracted disease of this category did not have a working vaccine)
9	Genetic

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 q0 rows of the data we will use for exploratory plots")
```

```
## [1] "First q0 rows of the data we will use for exploratory plots"
```

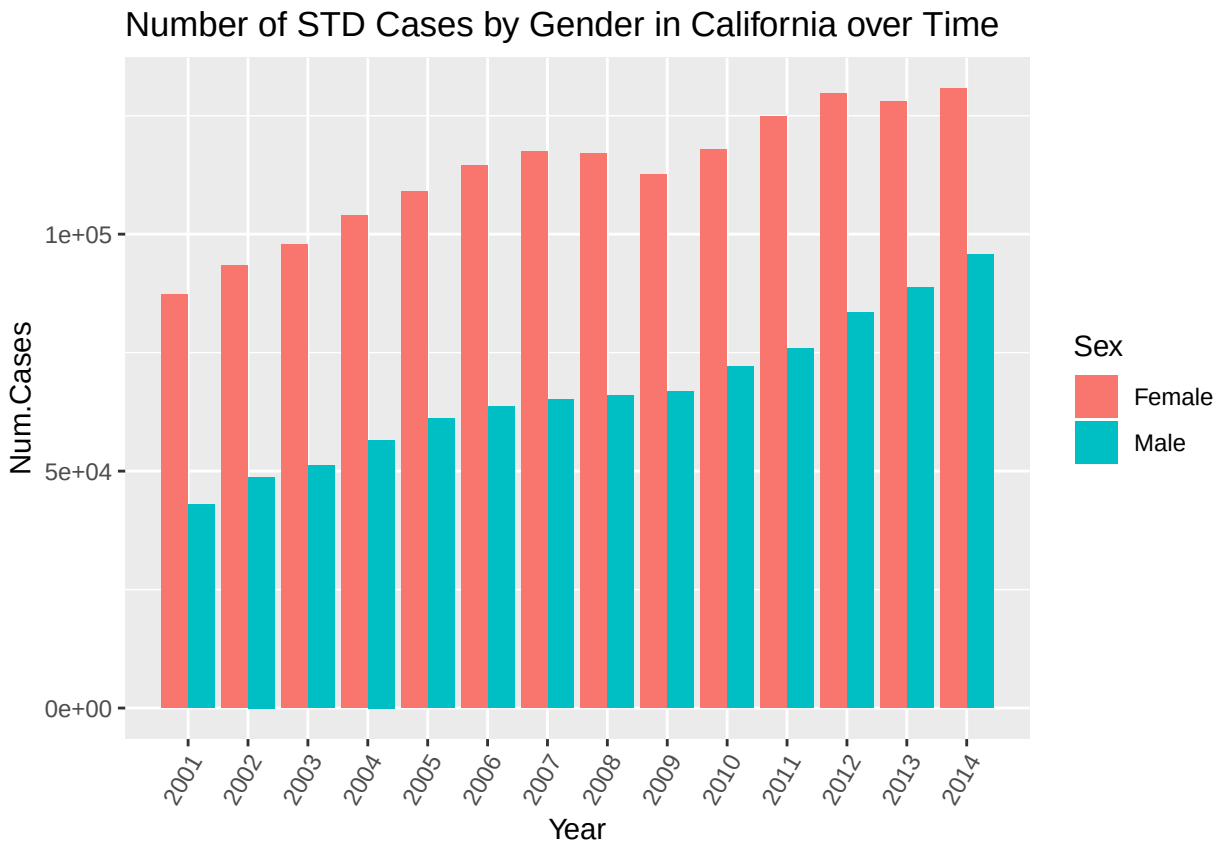
```
head(health_data2,10)
```

```
## # 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 bettter 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.*

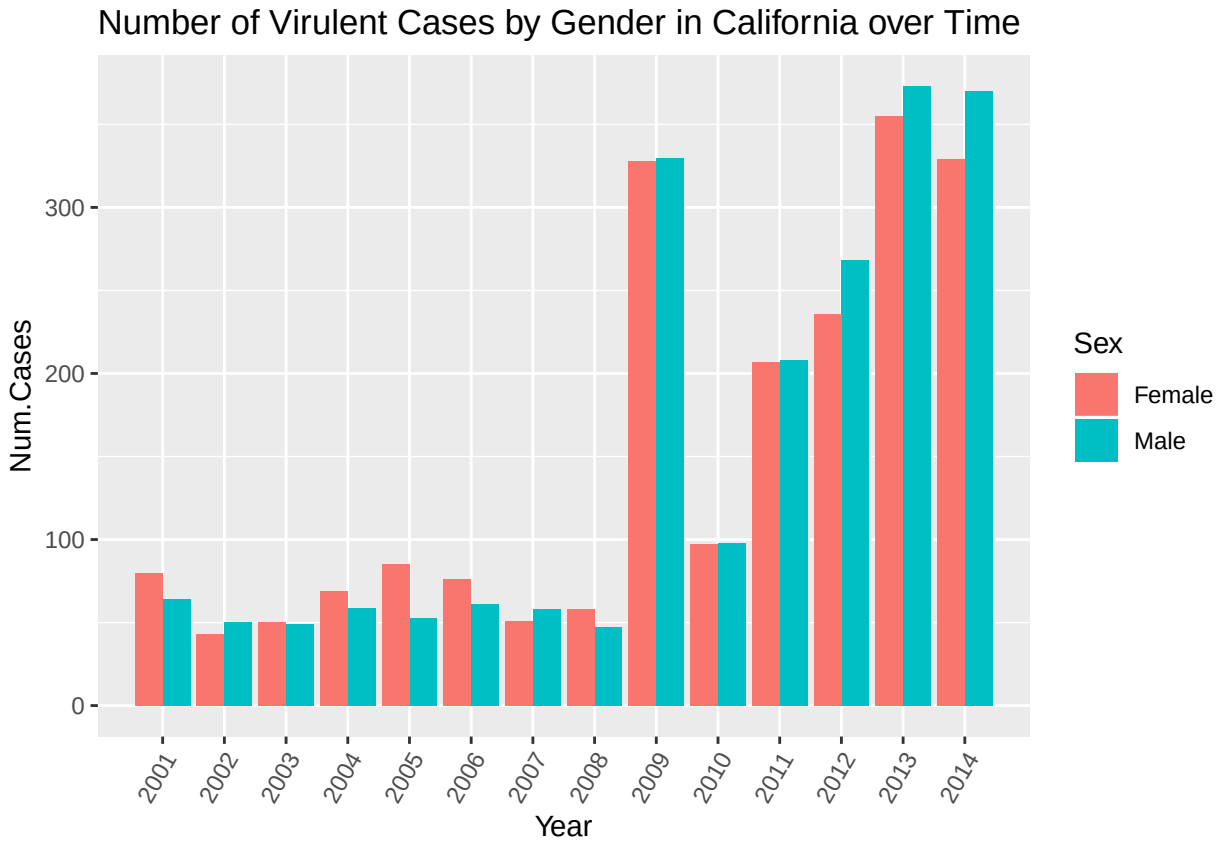
```
#Gender vs STD
health_data2%>%
  filter(Classification == 7)%>%
  ggplot(aes(x = Year, y = Num.Cases))+
  geom_bar(aes(fill = Sex), stat = "identity",position = "dodge")+
  scale_x_discrete("Year", years, years.char, years)+
  theme(axis.text.x = element_text(angle = 60, hjust = 1))+
  ggtitle("Number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



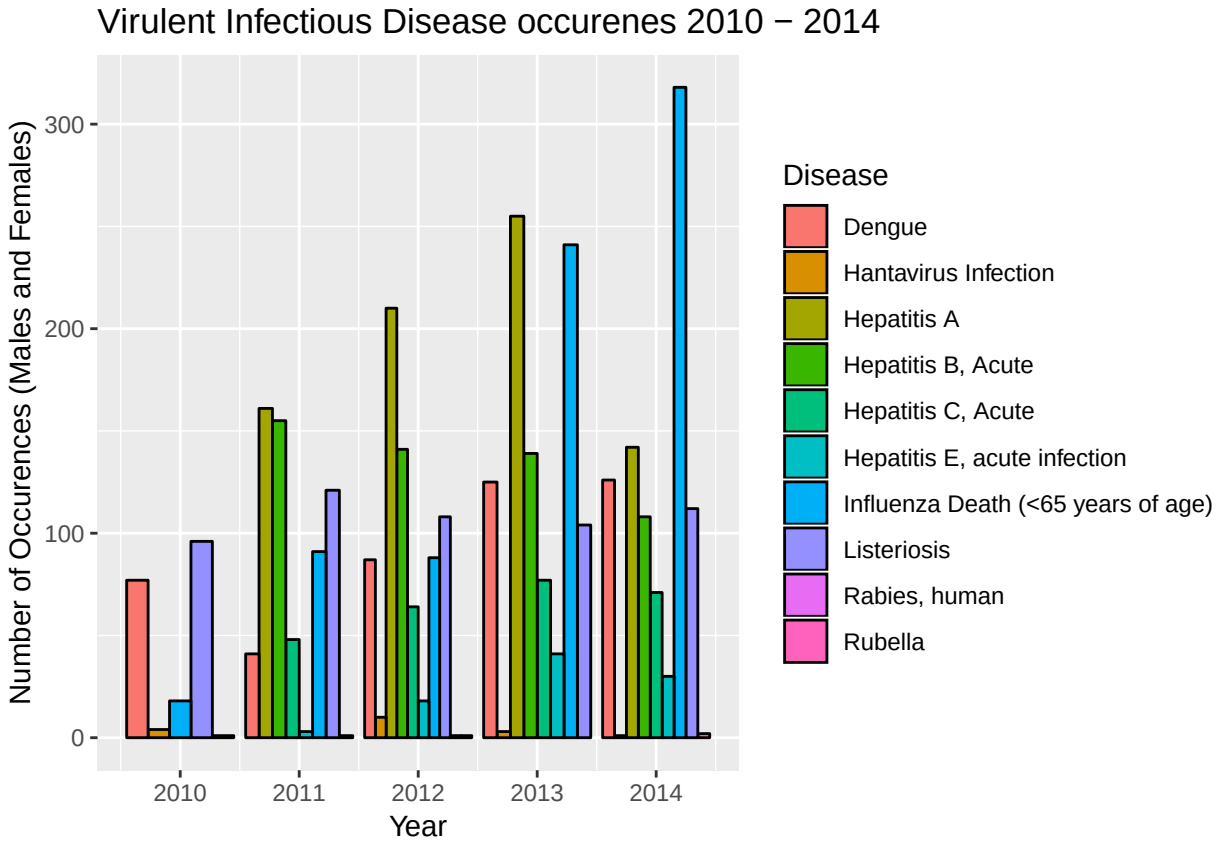
So, we can infer two things from this graph. First, we can sumize that within this data set, females have a higher rate of STD infection than do males. Second, we can observe that STD's are the most frequent class of infectious disease reported to the The California Department of Public Health (CDPH).

```
#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")
```

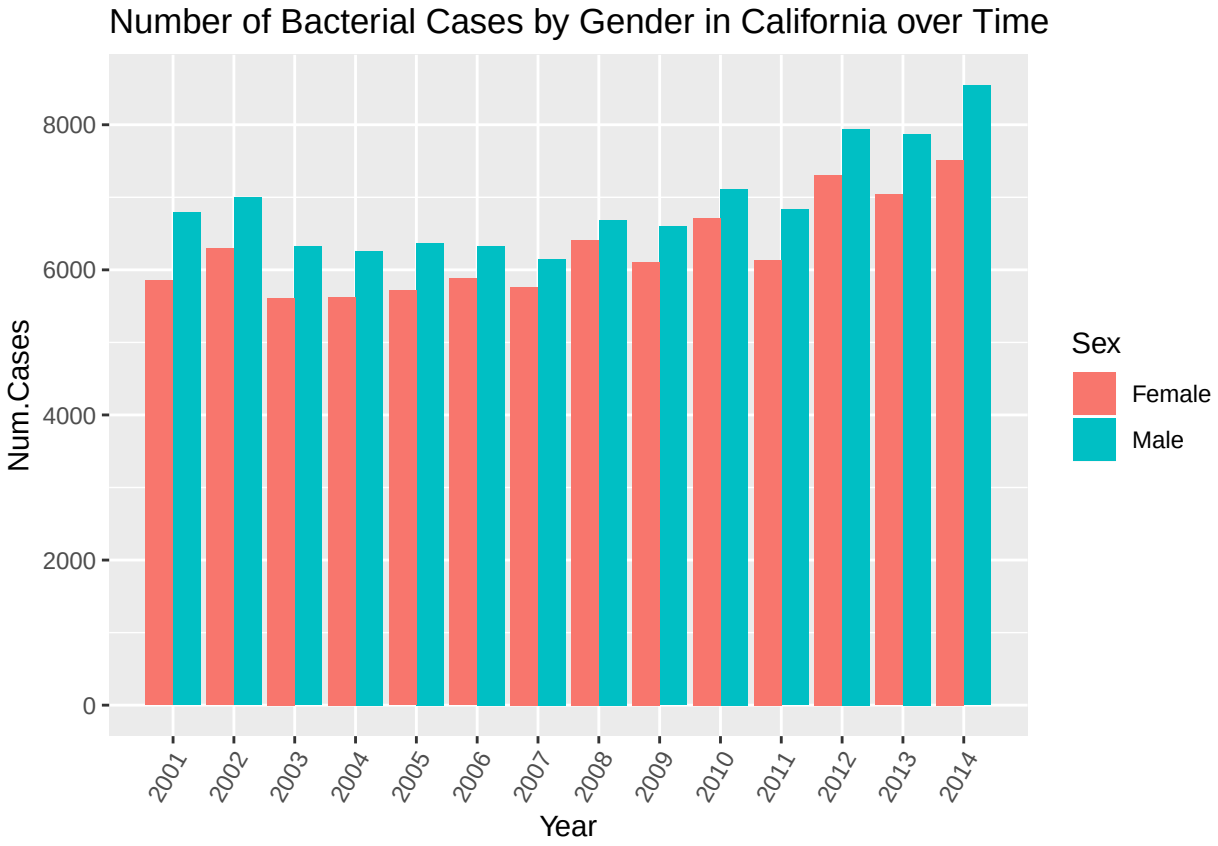



```
virusGrowth <- health_data%>%
  filter(Classification == 1 | Classification == 2)%>%
  filter(Year > 2009)%>%
  #mutate(Year = as.factor(Year))%>%
  select(Year, Disease,Count)%>%
  group_by(Disease, Year)%>%
  mutate(number_Cases = sum(Count))%>%
  filter(number_Cases > 0)%>%
  select(Year,Disease, 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
```



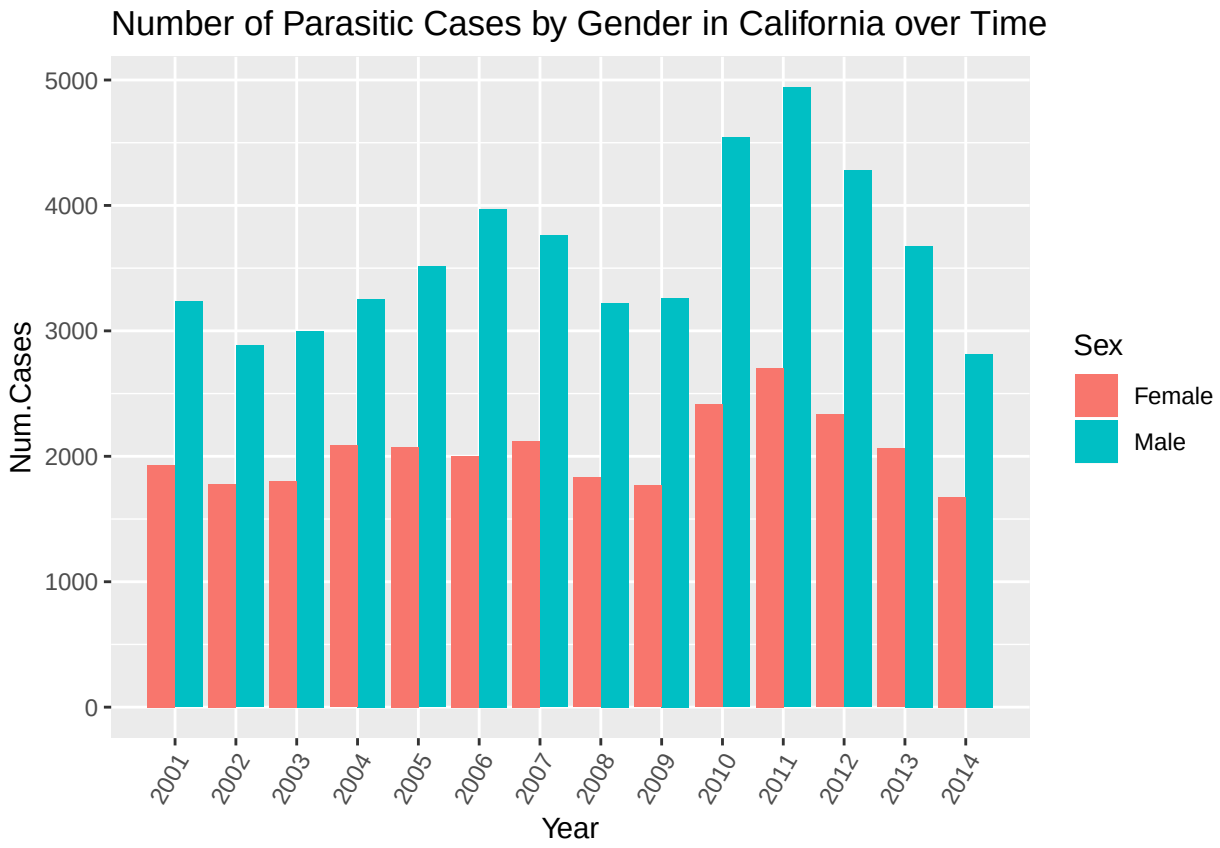
We can infer several things from these plots. First, we observe that virulent cases do not have a strong relation to gender, that is to say there does not exist a significant difference between the number of males and females who became infected with infectious virulent diseases. Seond, we can take note 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")
```

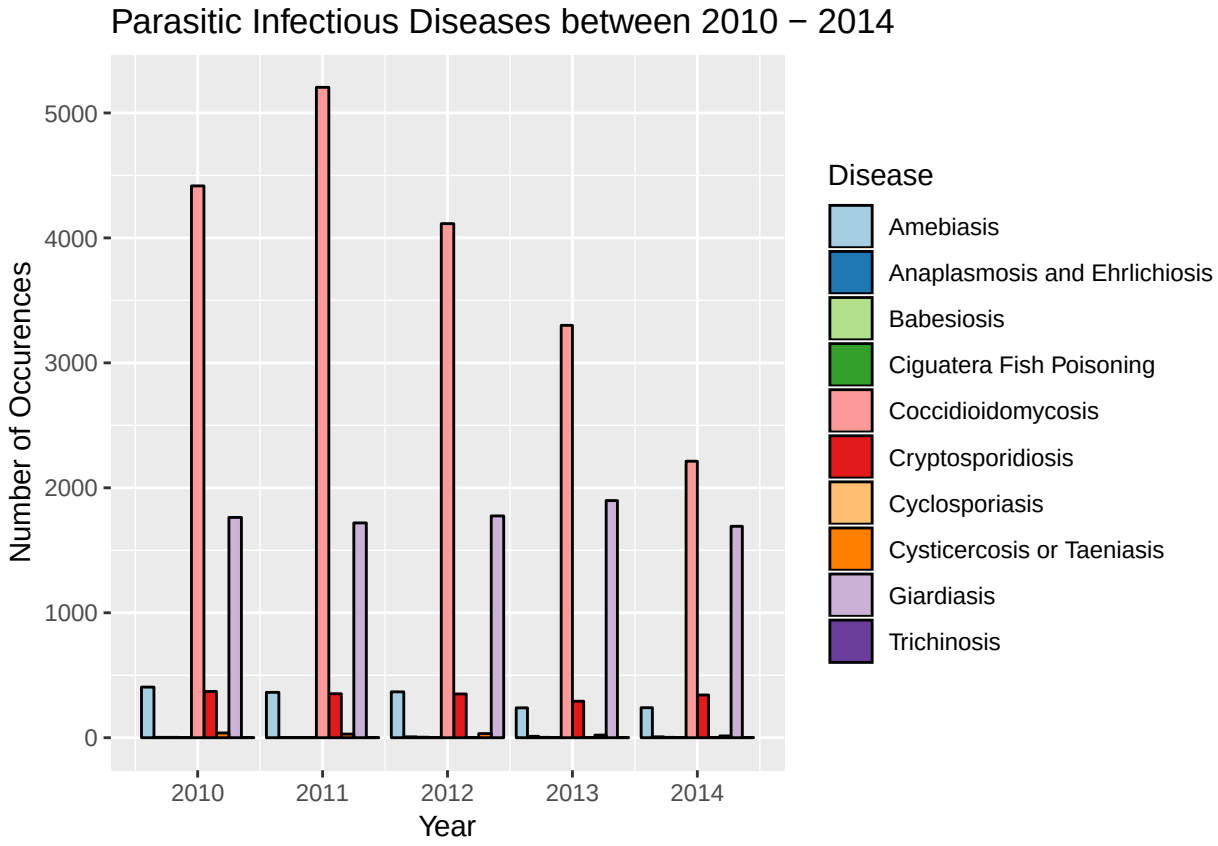


We can infer from this plot that bacterial cases do not have a strong relation to gender, that is to say there does not exist a significant difference between the number of males and females who became infected with an infectious bacterial disease.

```
#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")
```

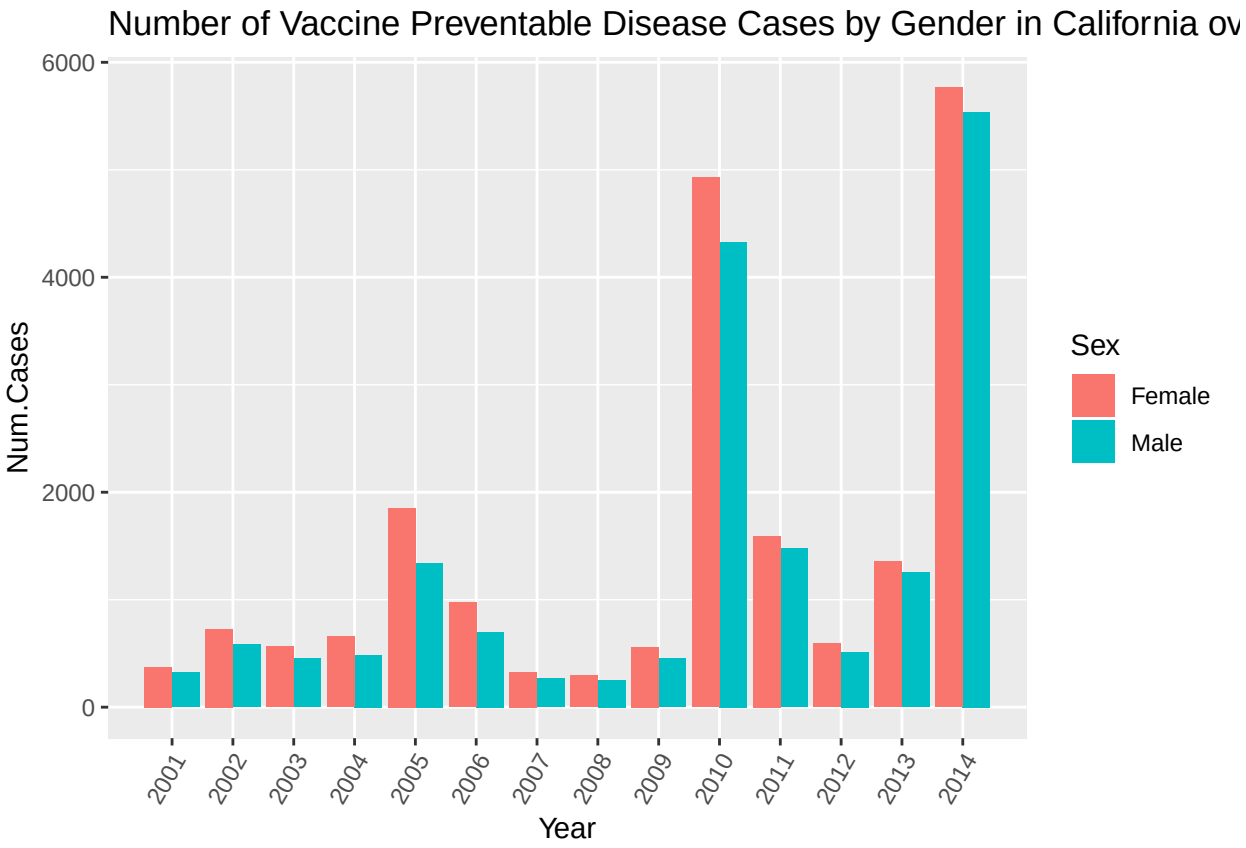



```
paraGrowth <- health_data%>%
  filter(Classification == 5 | Classification == 6)%>%
  filter(Year > 2009)%>%
  #mutate(Year = as.factor(Year))%>%
  select(Year, Disease,Count)%>%
  group_by(Disease, Year)%>%
  mutate(number_Cases = sum(Count))%>%
  filter(number_Cases > 0)%>%
  select(Year,Disease, 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
```



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 parasiric diseases. Seond, 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 inafforable to many people and the * anti-vaxx* movement remains, the whooping cough will contniue to cycle with an outbreak every 3-5 years.

Machine Learning

Multinomial Logistic Regression

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,

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

$$d_i = \text{ disease classification of person-i, where}$$

$$i = \text{Lethal Virulent, Non-Lethal Virulent, Lethal Bacterial, ... , Genetic}$$

$$t_{ij} = \text{number of cases of diease-i in year-j}$$

First we will seperate 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.

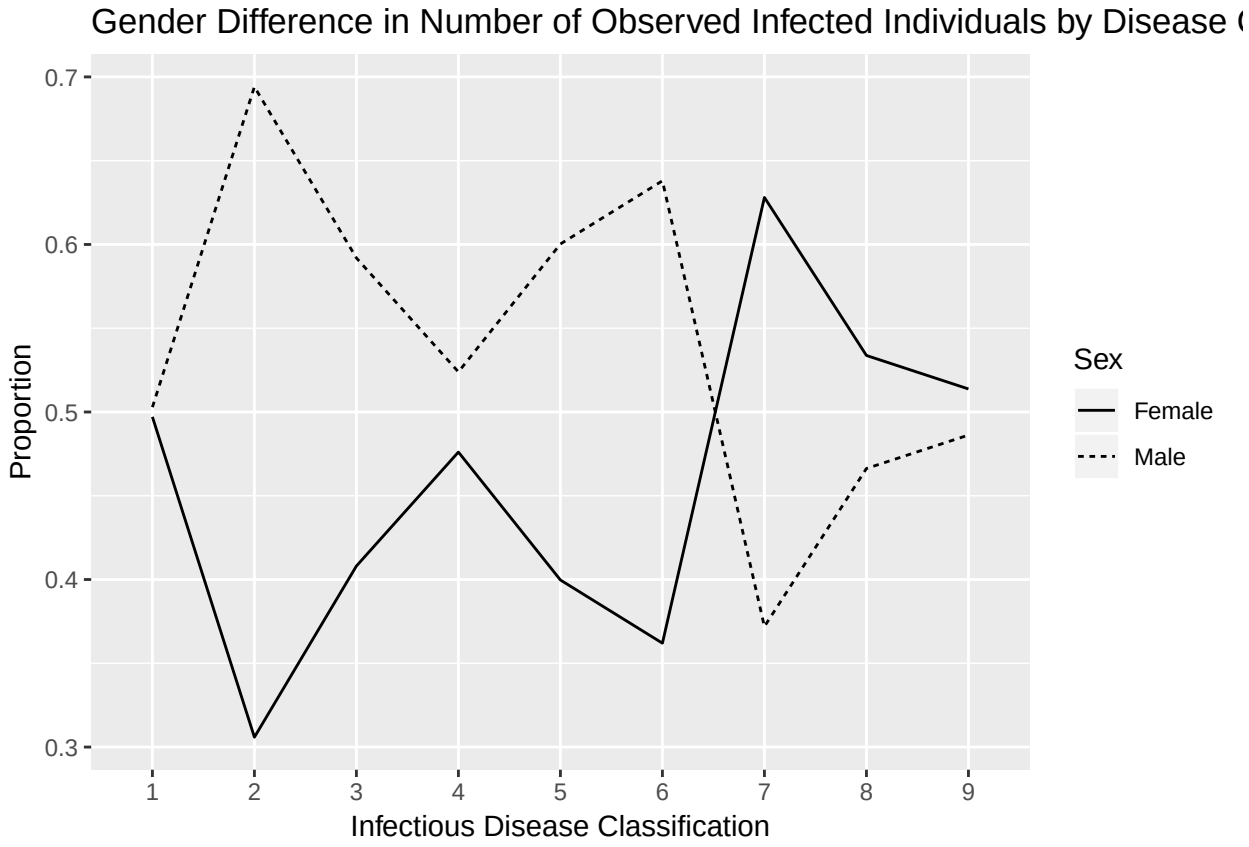
```
#Examine Data

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()%>%
  mutate(Sex = droplevels(Sex, "Total"))

#Rename classes to reflect what the classifications mean:
levels(health_ML$Classification) <- c("Lethal Viral", "Non-Lethal Viral", "Lethal Bacterial", "Non-Letha
#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 reproducability
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,]
```

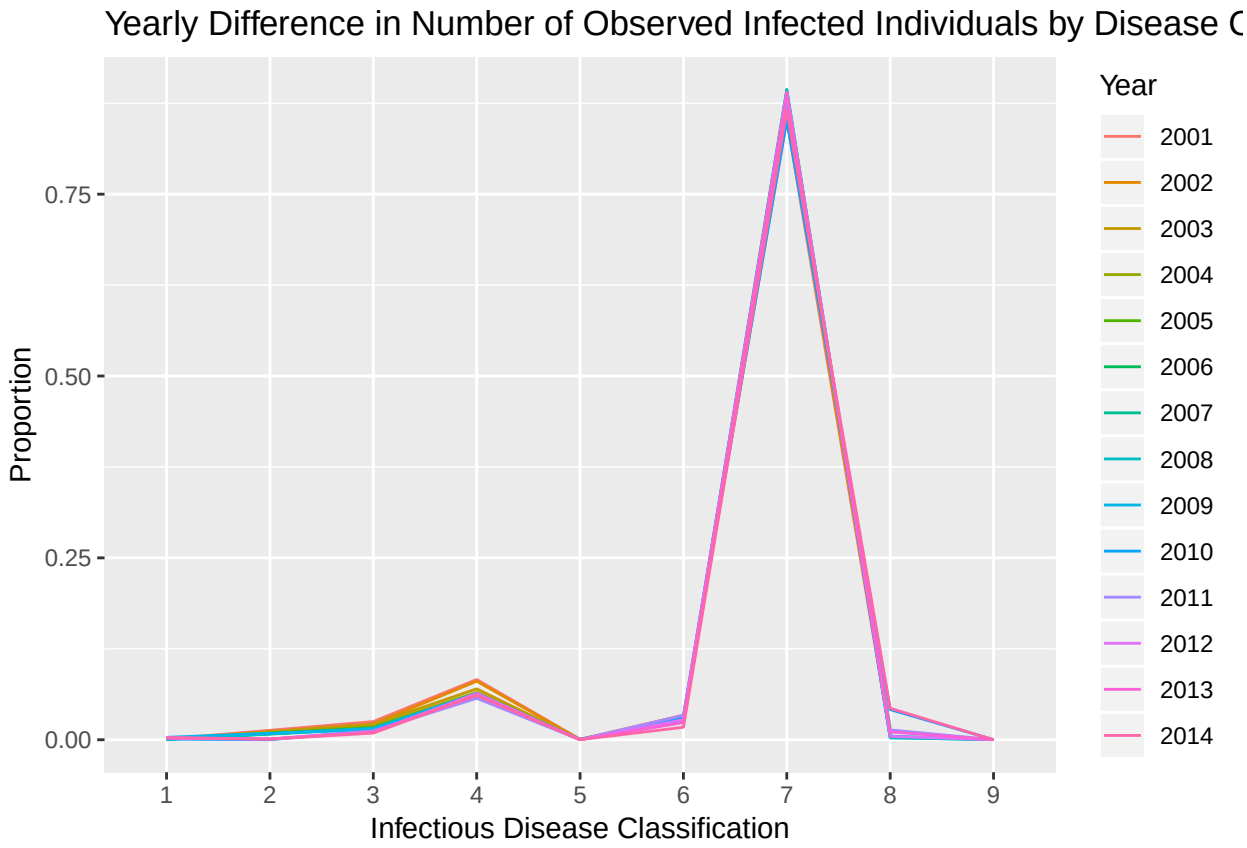
** Gaining Intuition on Model 1.**

```
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")
```



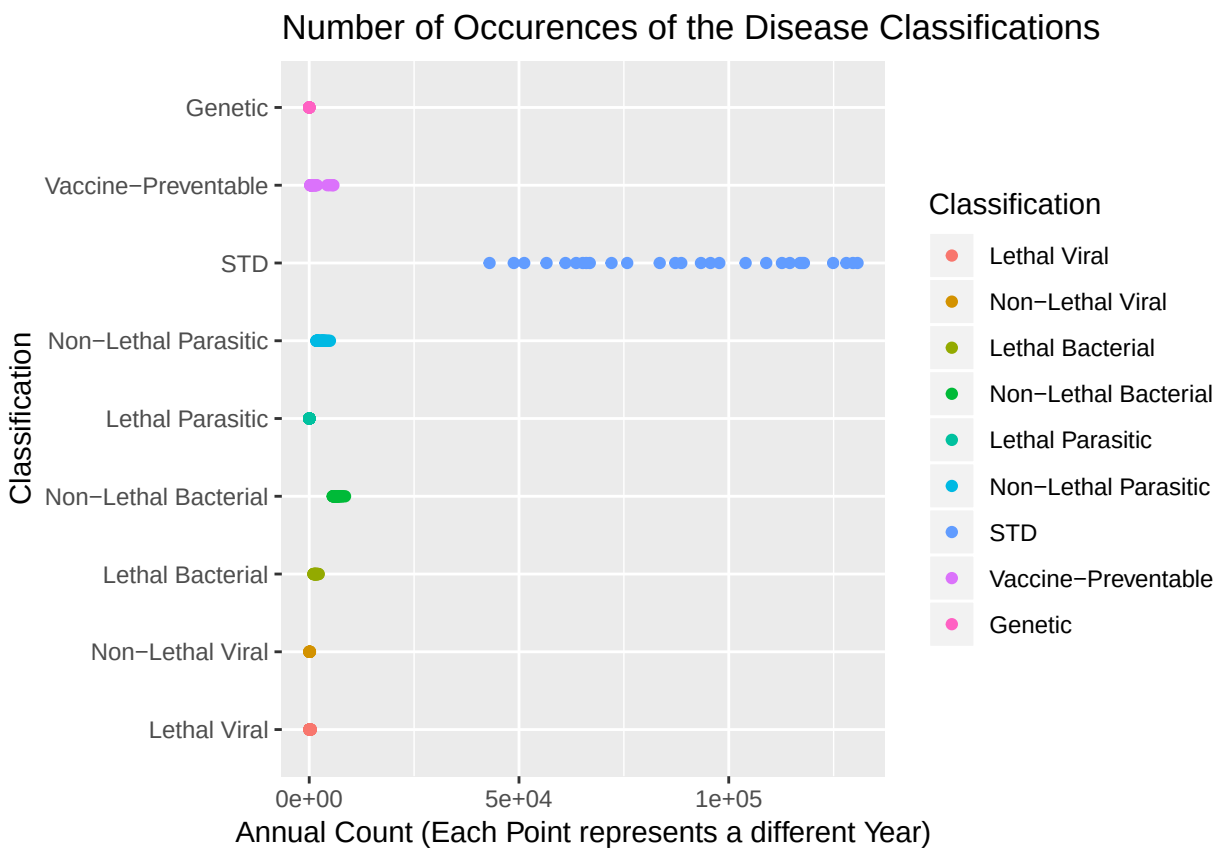
We observe that there does exist difference in the number of instances of a certain classs of infectious diseases for males and females.

```
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()
```

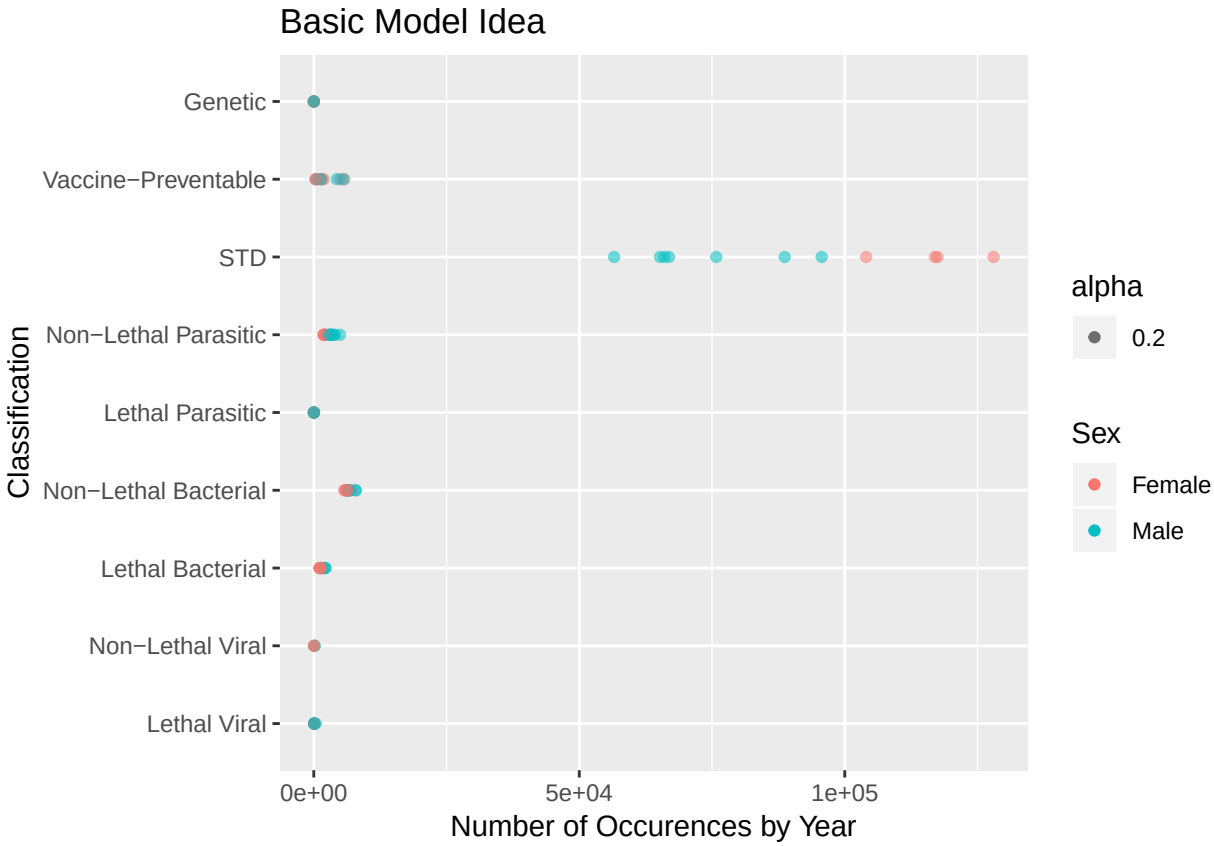



```
#Examine what proportion of the data is an STD
is.STD <- paste(round(table(health_ML$Classification == "STD")/length(health_ML$Classification),2)*100,
names(is.STD) <- c("Percentage STD", "Percentage of Non-STD")
```

So, we do observe that the number of occurences of a certain class of infectious diseases does correlate to the number of occurences 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 indiviuals 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)
```

```
## # 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 unneccessary predictor va
modelAIC <- step(multiLogitModelAll)
```

```
## Start:  AIC=186.08
## Sex ~ Year + Classification + Year.Count
##
## trying - Year
## # weights:  11 (10 variable)
## initial  value 92.188575
```

```
## iter  10 value 82.279611
## final  value 82.055801
## converged
## trying - Classification
## # 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
## # 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
## # 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
## # weights:  3 (2 variable)
## initial  value 92.188575
## final  value 90.301867
## converged
## trying - Year.Count
## # 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
```

```r
#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.

```r
#Predict the gender of individuals in our test set the Multinomial Logistic Regression Model
predictedSex <- predict(modelAIC, health_test[,1:3], type = "class")

#Examining Model Performance:
confusionMatrix(predictedSex, health_test$Sex)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Female Male
##     Female     23   28
##     Male       22   16
##
##                Accuracy : 0.4382
##                  95% CI : (0.3332, 0.5475)
##     No Information Rate : 0.5056
##     P-Value [Acc > NIR] : 0.9160
##
##                   Kappa : -0.1254
##  Mcnemar's Test P-Value : 0.4795
##
##             Sensitivity : 0.5111
##             Specificity : 0.3636
##          Pos Pred Value : 0.4510
##          Neg Pred Value : 0.4211
##              Prevalence : 0.5056
##          Detection Rate : 0.2584
##    Detection Prevalence : 0.5730
##       Balanced Accuracy : 0.4374
##
```

```
##        'Positive' Class : Female
##
```

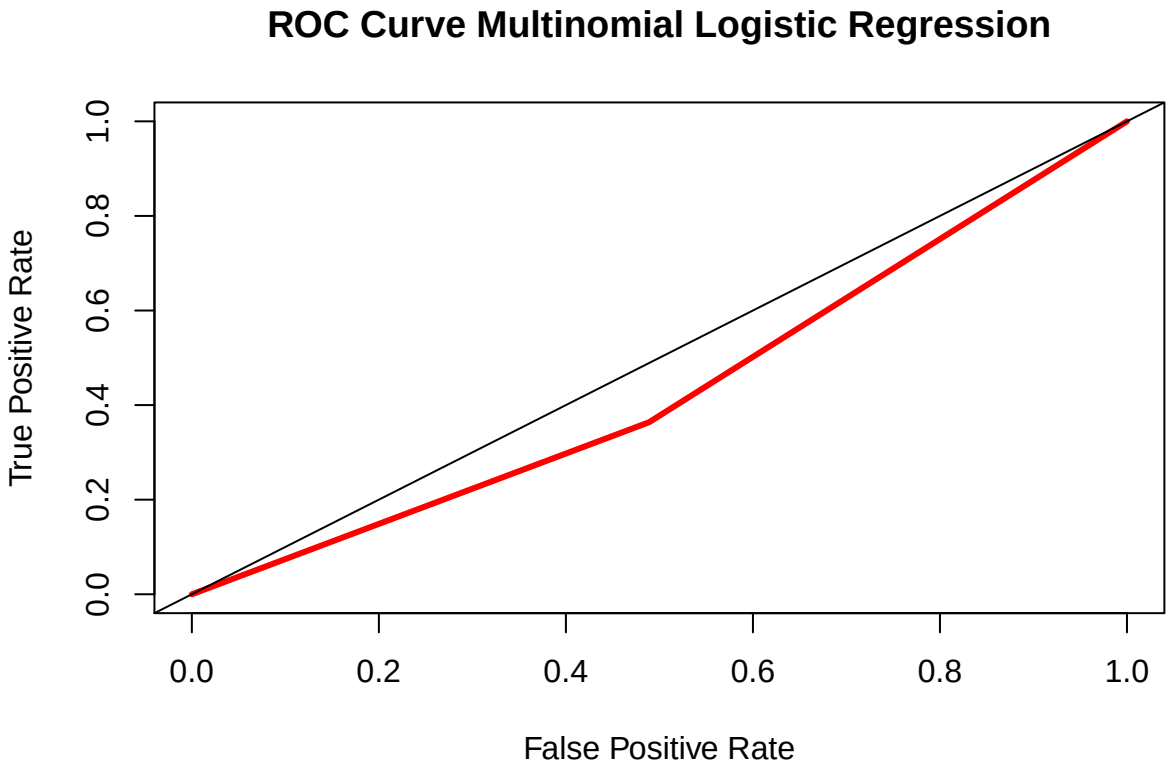
A confusion matrix is a tool that will allow us to better quantify how well our model actually classifies. There does not exist any inherent postive or negative class for gender. Unforunately, the confusion the utility of the confusion matrix is usually in the context classifying the status of af event 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 positive this is called a false negative.

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

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 a 95% Confidence Interval for the accuracy of this model is between 39.75% and 61.33%.

Alternitvely, we can examine the Area Under Curve(AUC) of the ROC. ROC is the Receiver Operating Characteristics of a model, ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with that are and those that are female. Ideally, a good model's ROC curve hugs the top left of the plot indicating the model recognizes true positives and not false positives. Our ROC curve is,

```
preds <- predict(modelAIC,health_test[c("Year","Classification","Year.Count")],type = "class")
actual <- droplevels(health_test$Sex, exclude = "Total")

perf.multinomial <- performance(prediction(predictions = as.numeric(preds), as.numeric(actual)), measur

#ROC tree
plot(perf.multinomial, col=2, lwd=3, main="ROC Curve Multinomial Logistic Regression", xlab = "False Po
abline(0,1)
```



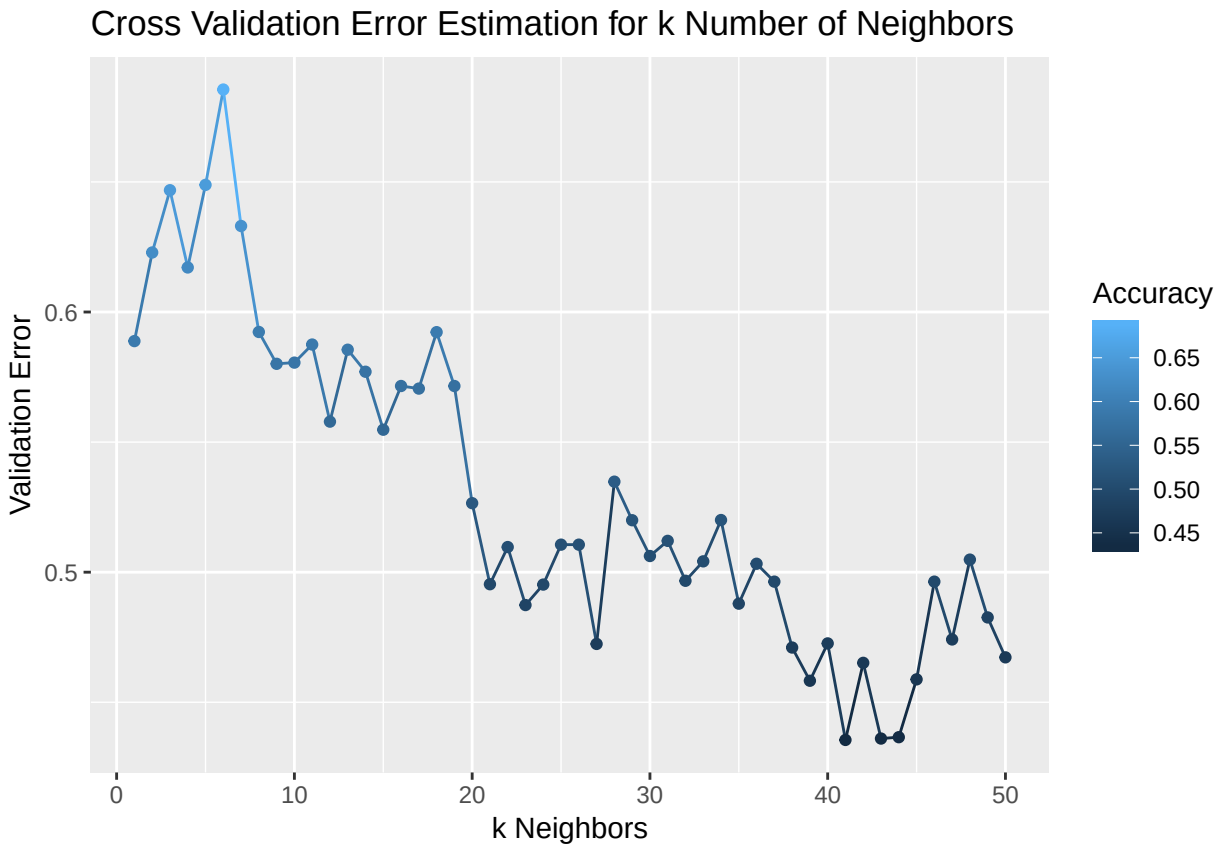
These results make sense since we have a much higher number of observations of females with STD's than males with STD's, and given how common observations of STD's are in our data, it makes sense that the model has a bias towards predicting Sex as female which drives down the accuracy. This explination is supported by evidence the coefficient associated with disease classification STD has a maximum-likelihood coefficient orders greater than other coefficients. Furthermore, the `caret` package reaffirms this by stating that `'Positive' Class : Female` which means that our model is positively biased towards predicting an infected person's gender as female since females have a higher rate of STDs *within this dataset*, and STD's account for a large plurality of all observed classes of infectious diseases in California.

KNN using 9-fold cross-validation:

The goal of this section is to test whether other methods of learning will achieve better results. I will be using a knn algorithm and 9-fold cross validation to train the model using the `caret` package.

```
#Set seed for reproducability
set.seed(3)

#Instantiate Cross Validation Fold ID's
trControl <- trainControl(method  = "cv",
                          number  = 9)
#Train model
fit <- train(Sex ~ .,
             method     = "knn",
             tuneGrid   = expand.grid(k = 1:50),
             trControl  = trControl,
             metric     = "Accuracy",
             data       = health_train)

#Plot errors vs k
ggplot(fit$results,aes(x = k, y = Accuracy))+
```

```
  geom_line(aes(colour = Accuracy))+
  geom_point(aes(colour = Accuracy))+
  xlab("k Neighbors")+
  ylab("Validation Error")+
  ggtitle("Cross Validation Error Estimation for k Number of Neighbors")
```




```
#Select best k
best_k = fit$results$k[which(fit$results$Accuracy == max(fit$results$Accuracy))]
```

Since, the final value used for the model was k = 6. Then, we can now build a KNN classifier with k = 6, and then again check the testing error to see if KNN outperforms the Multinomial Logistic Regression Classifier.

```
#Predict gender labels in test data
knnPredict <- predict(fit,newdata = health_test[c("Year", "Classification","Year.Count")])

#Note I now realize there is an existing function in R to create the confusion matrix, so I will that h
x <- confusionMatrix(knnPredict, health_test$Sex)


score <- round(x$overall["Accuracy"]*100,2)
lower_bound_accuracy <- round(x$overall["AccuracyLower"]*100,2)
upper_bound_accuracy <- round(x$overall["AccuracyUpper"]*100,2)

x
```

```
## Confusion Matrix and Statistics
##
```

```
##           Reference
## Prediction Female Male
##     Female     29   14
##     Male       16   30
##
##                Accuracy : 0.6629
##                  95% CI : (0.5549, 0.7597)
##     No Information Rate : 0.5056
##     P-Value [Acc > NIR] : 0.001956
##
##                   Kappa : 0.3261
##  Mcnemar's Test P-Value : 0.855132
##
##             Sensitivity : 0.6444
##             Specificity : 0.6818
##          Pos Pred Value : 0.6744
##          Neg Pred Value : 0.6522
##              Prevalence : 0.5056
##          Detection Rate : 0.3258
##    Detection Prevalence : 0.4831
##       Balanced Accuracy : 0.6631
##
##        'Positive' Class : Female
##
```

So, the KNN classifier recieved an accuracy score of 66.29% and a 95% Confidence Interval for the accuracy of the KNN model is that the accuracy lie between 55.49% and 75.97%. Furthermore, we again see the same positive bias towards predicting the gender of an infected person as female.

Thus, the KNN classifier did outperform the Multinomial Logistic Regression Model. I would assume that this is because due to the sparse nature of this data, the more granular KNN classifier is able to better predict gender by looking at the gender of 6 of its neighbors.

Next Steps:

- Employ Regularization to try to reduce the bias in the dataset towards certain categories.
- Supplement Data with Different Predictors.
- Get more data.
  - After the proccessing 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 a good enough of a fit with the amount of data I am working with.
- Dimmensionality Reduction.
  - Due to the sparse nature of this dataset, after rolling up the county-level data into state data information was likely lost in the proccess. A Principal Component Analysis may help this problem.
- Include more features.
  - This dataset has county locations per case, as well as time, so using these more granular features may increase model acccuracy by giving us more data to work with. The trade-off for choosing to do this would be that the model would be less interpretable, which is why in this analysis I chose to aggregate the data into statewide and annual levels.

Thank you for your time. Daniel Fields