MULTIMEDIA UNIVERSITY®

FACULTY OF COMPUTING AND INFORMATICS (FCI)
MULTIMEDIA UNIVERSITY
CYBERJAYA.

TDS2101 - INTRODUCTION TO DATA SCIENCE
SESSION - TT02

Lecturer Name: John See Su Yang

Topic : Infectious Disease (Health)

PART B

Group Members	**Student ID**
Sarah Batrisyia Binti Ahmad Salman	1181301724
Amiera Nur Hafidzah Binti Muzaffar	1181302678

Table of Contents Page

Introduction

In this assignment we've decided to analyze the data regarding Infectious Diseases that has spread in California in the year 2005 until 2014. Infectious diseases are caused by microorganisms which are called pathogens that can disrupt the one's body normal process and stimulate the immune system. The reason why we chose this topic is that we want to know what are the most disease cases that happened in the year 2005 until 2014. We limit the area which is in a county in California, USA.

Our criteria for datasets that we would want to choose are the dataset that have type or name of the disease so that we can know what are the most diseases for us to analyze the data.

Therefore, we choose the datasets "Infectious Disease 2001-2014" from "data.world" because it fits our criteria to do data wrangling for this assignment and we limit the year from 2005 until 2014. Inside this data contains :

1. **Disease** - The name of the disease reported for the patient.
2. **County** - The county in which the case resided when they were diagnosed and/or where they are currently receiving care; in most cases this will be the county that reported the case.
3. **Year** - Year is derived from the estimated illness onset date. We defined the estimated illness onset date for each case as the date closest to the time when symptoms first appeared.
4. **Sex** - The patient's biological sex at birth.
5. **Count** - The number of occurrences of each disease that meet the surveillance definition and/or inclusion criteria specific to that disease for that County, Year, Sex strata.

6. **Population** - The estimated population size (rounded to the nearest integer) for each County, Year, Sex strata.
7. **Rate** - The rate of disease per 100,000 population for the corresponding County, Year, Sex strata using the standard calculation (Count *100,000/Population)
8. **CI.upper** - The lower bound of the 95% confidence interval for the calculated rate.
9. **CI.lower** - The upper bound of the 95% confidence interval for the calculated rate.

Data Cleaning Method

Since the data that we have found has a huge amount of data (originally has 141777 rows and 10 columns), we have to cut down the data to a suitable amount. The original data contain the data from 2001 - 2014. With the advice from the lecturer, we decided to choose the data from 2005 - 2014. After cutting down the years, we manage to get a data contains 104961 rows and 10 columns.

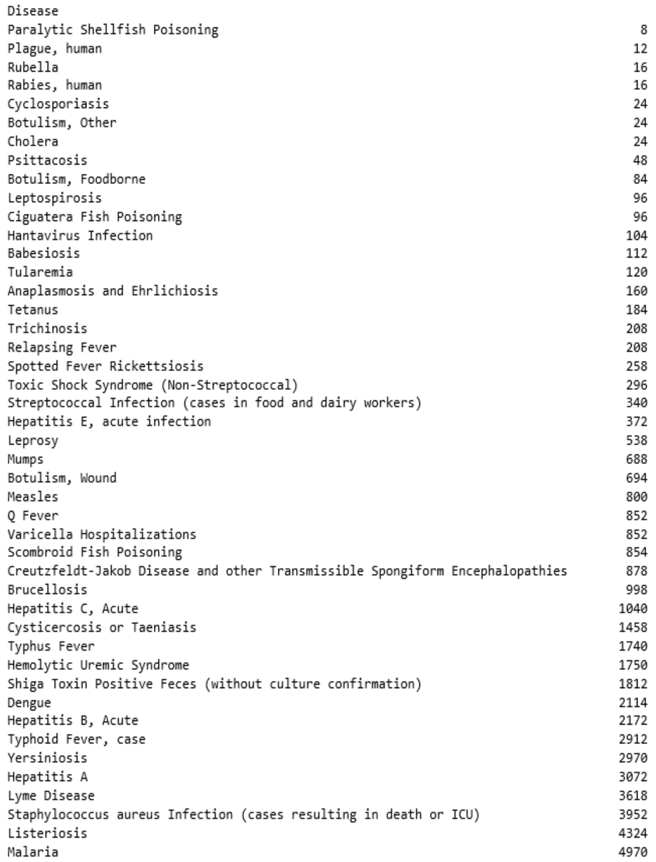
After that we remove the column labeled 'Unstable' because from what we observed, we will not be using that column in this assignment.

Lastly, we removed all the null values inside the data so that there will be no wrong calculation throughout the data processing stage.

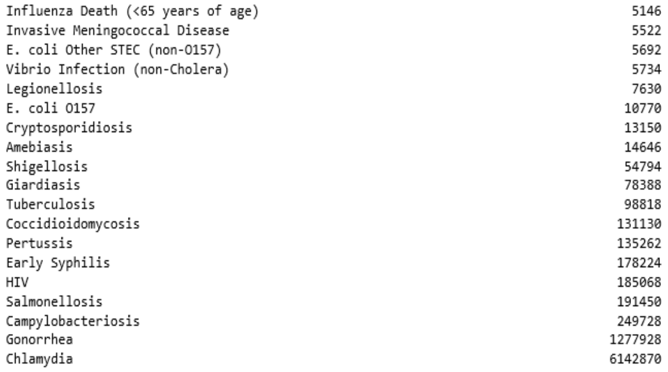
Overall, we manage to clean our data to ensure that the data that we are using throughout this assignment is correct, consistent and usable.

Question 1

a. *What is the total rate of each disease that has occur throughout 2005-2014 in California?*



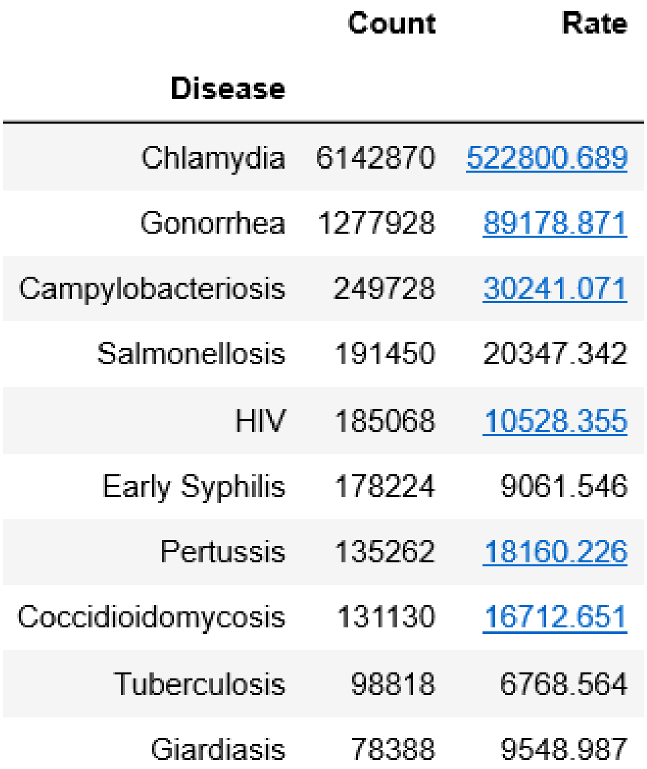
```
Disease
Paralytic Shellfish Poisoning                                                    8
Plague, human                                                                   12
Rubella                                                                         16
Rabies, human                                                                   16
Cyclosporiasis                                                                  24
Botulism, Other                                                                 24
Cholera                                                                         24
Psittacosis                                                                     48
Botulism, Foodborne                                                             84
Leptospirosis                                                                   96
Ciguatera Fish Poisoning                                                        96
Hantavirus Infection                                                           104
Babesiosis                                                                     112
Tularemia                                                                      120
Anaplasmosis and Ehrlichiosis                                                  160
Tetanus                                                                        184
Trichinosis                                                                    208
Relapsing Fever                                                                208
Spotted Fever Rickettsiosis                                                    258
Toxic Shock Syndrome (Non-Streptococcal)                                       296
Streptococcal Infection (cases in food and dairy workers)                      340
Hepatitis E, acute infection                                                   372
Leprosy                                                                        538
Mumps                                                                          688
Botulism, Wound                                                                694
Measles                                                                        800
Q Fever                                                                        852
Varicella Hospitalizations                                                     852
Scombroid Fish Poisoning                                                       854
Creutzfeldt-Jakob Disease and other Transmissible Spongiform Encephalopathies  878
Brucellosis                                                                    998
Hepatitis C, Acute                                                            1040
Cysticercosis or Taeniasis                                                    1458
Typhus Fever                                                                  1740
Hemolytic Uremic Syndrome                                                     1750
Shiga Toxin Positive Feces (without culture confirmation)                     1812
Dengue                                                                        2114
Hepatitis B, Acute                                                            2172
Typhoid Fever, case                                                           2912
Yersiniosis                                                                   2970
Hepatitis A                                                                   3072
Lyme Disease                                                                  3618
Staphylococcus aureus Infection (cases resulting in death or ICU)             3952
Listeriosis                                                                   4324
Malaria                                                                       4970
```



```
Influenza Death (<65 years of age)                                            5146
Invasive Meningococcal Disease                                                5522
E. coli Other STEC (non-O157)                                                 5692
Vibrio Infection (non-Cholera)                                                5734
Legionellosis                                                                 7630
E. coli O157                                                                 10770
Cryptosporidiosis                                                            13150
Amebiasis                                                                    14646
Shigellosis                                                                  54794
Giardiasis                                                                   78388
Tuberculosis                                                                 98818
Coccidioidomycosis                                                          131130
Pertussis                                                                   135262
Early Syphilis                                                              178224
HIV                                                                         185068
Salmonellosis                                                               191450
Campylobacteriosis                                                          249728
Gonorrhea                                                                  1277928
Chlamydia                                                                  6142870
```

The figure above shows the total rates of each disease from the year 2005 until 2014 in California. We can conclude that the lowest disease rate is Paralytic Shellfish Poisoning which has 8 total rates and the highest disease rate is Chlamydia which has 614,2878 total rates.

b. *What is the top 10 disease that has the highest count and rate from 2005 - 2014?*



	Count	Rate
Disease		
Chlamydia	6142870	522800.689
Gonorrhea	1277928	89178.871
Campylobacteriosis	249728	30241.071
Salmonellosis	191450	20347.342
HIV	185068	10528.355
Early Syphilis	178224	9061.546
Pertussis	135262	18160.226
Coccidioidomycosis	131130	16712.651
Tuberculosis	98818	6768.564
Giardiasis	78388	9548.987

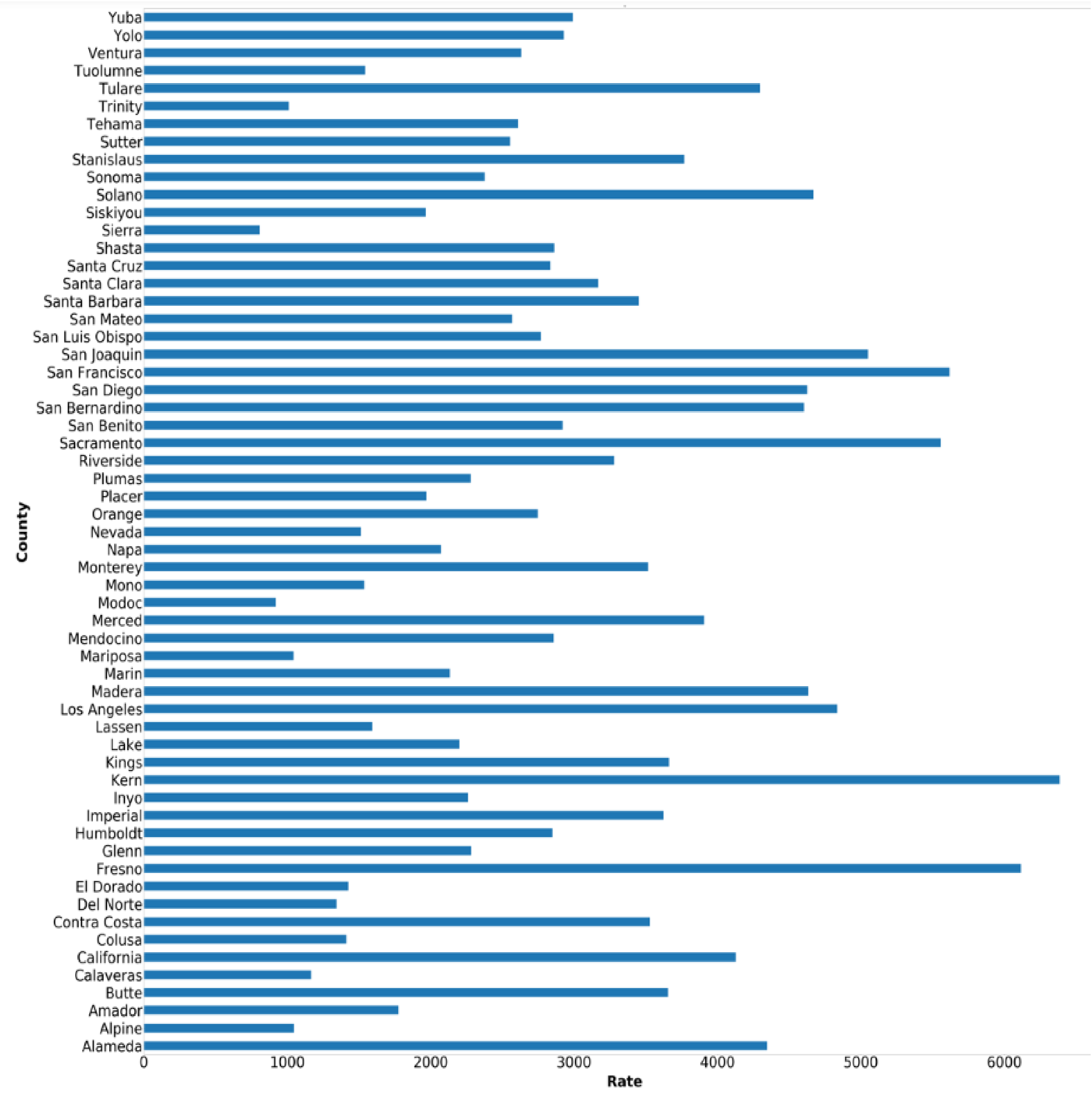
The figure above shows the top 10 highest count followed by rates. The highest infectious disease in California is Chlamydia. We cut it down to the top 10 to make it easier for us to see the number comparisons between each disease.

Chlamydia is a disease that can infect female and male. It is caused by bacteria called Chlamydia trachomatis. Women can get chlamydia in the cervix, rectum, or throat. Men can get chlamydia in the urethra (inside the penis), rectum, or throat.

Therefore, we choose Chlamydia to do data wrangling, and analyze more about this disease that has the highest count and rate in California.

Question 2

a. *Which county has the most disease rates because of X disease?*

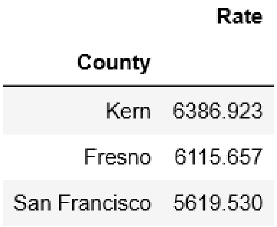


The county that has the most X disease rates according to the figure above is Kern with over 6000 rates of X disease. There were 6,348 chlamydia cases, 42 percent higher than the state average.

We choose to do a bar graph because it can summarize data set in a visual form and we can make sure the calculation is done correctly by observing the visual.

From what we gain in this bar graph we will be proceeding to analyze the county with the highest rate and count.

b. *What is the top 3 county with the highest rate of Chlamydia Disease?*

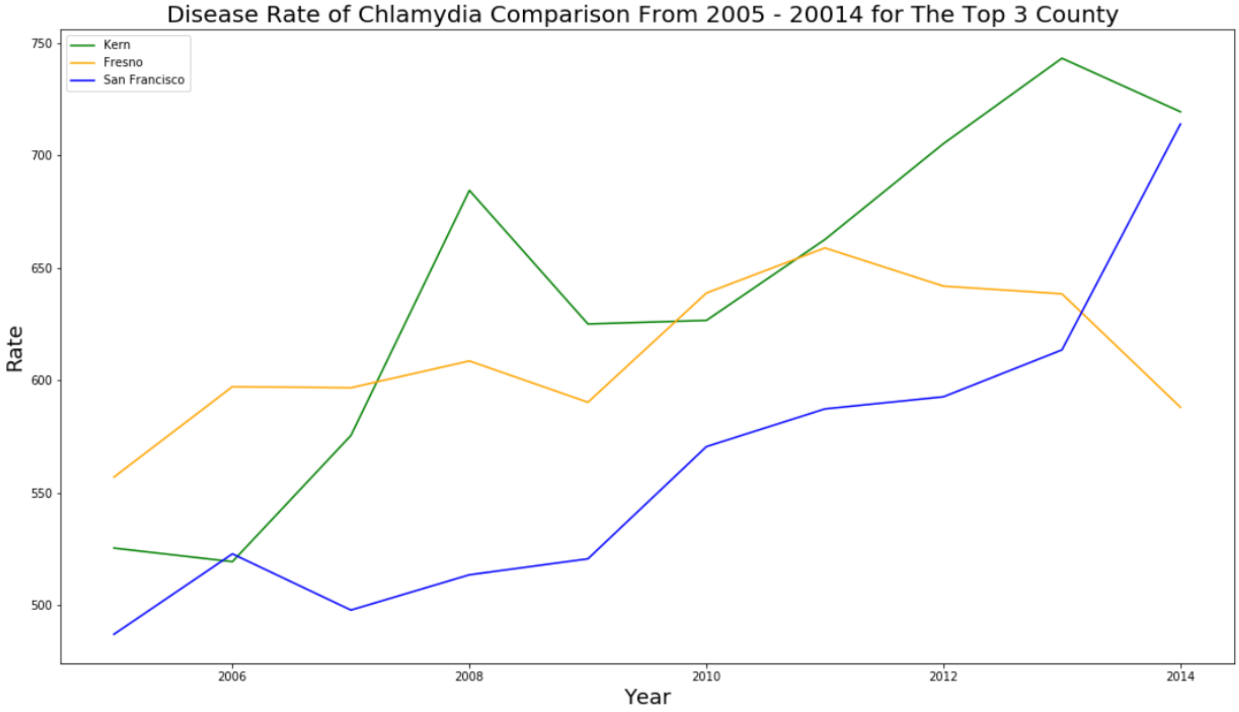


	Rate
County	
Kern	6386.923
Fresno	6115.657
San Francisco	5619.530

The counties that have the highest rate of Chlamydia disease are Kern with 6386.923 followed by Fresno which is 6115.657 and San Francisco which have 5619.530 rates of Chlamydia disease.

We cut it down to top 3 so that we can see the differences of the disease rate between the three counties and analyse it further.

c. Compare the top 3 county with Chlamydia throughout 2005 - 2014

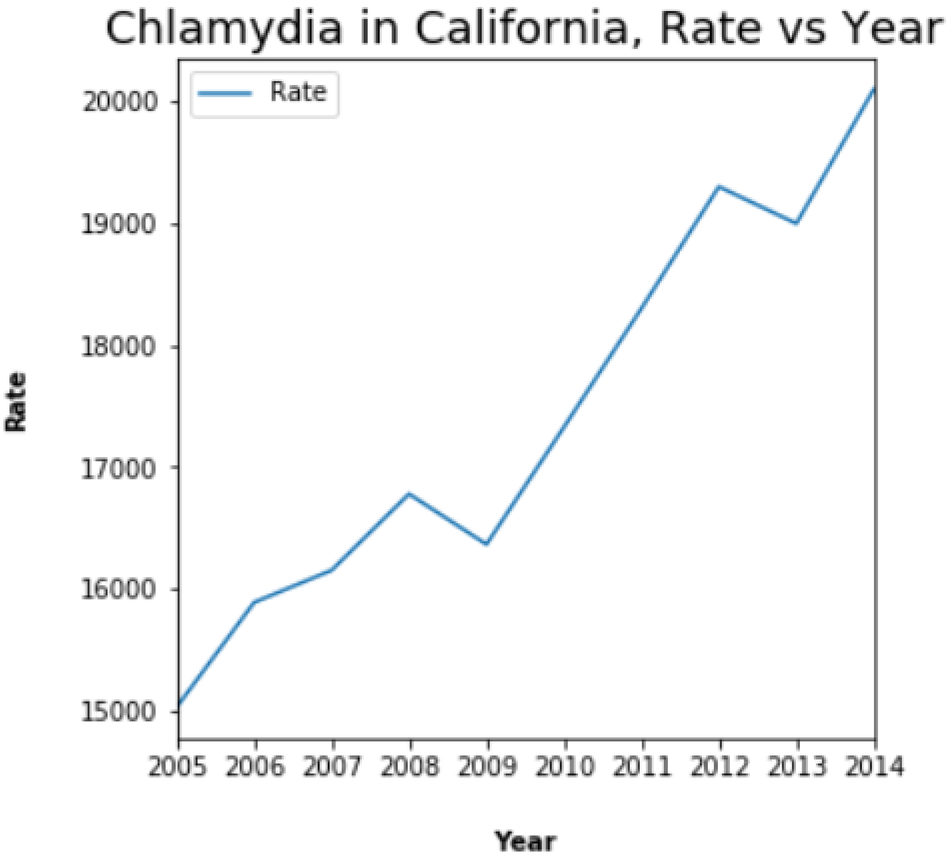


From the figure above we can conclude that Fresno has about a constant rate of Chlamydia disease throughout the year 2005 until 2014 compared to Kern and San Francisco. We can see clearly in 2009, the graph for all the three counties decreased. One of the reasons that there is a decline during that time is because they spread awareness to the people in California to do prevention steps. They stated in the text during 2009, " *Two recent studies reported trends among women 15–24 years of age attending Infertility Prevention Project clinics*"[1].

We used a line graph to analyse this data because it gives a good visualization for a data throughout 2005 - 2014. Plus, we can easily observe if there are any changes during the years.

Question 3

Can Chlamydia disease be decreasing in the future? (Prediction)

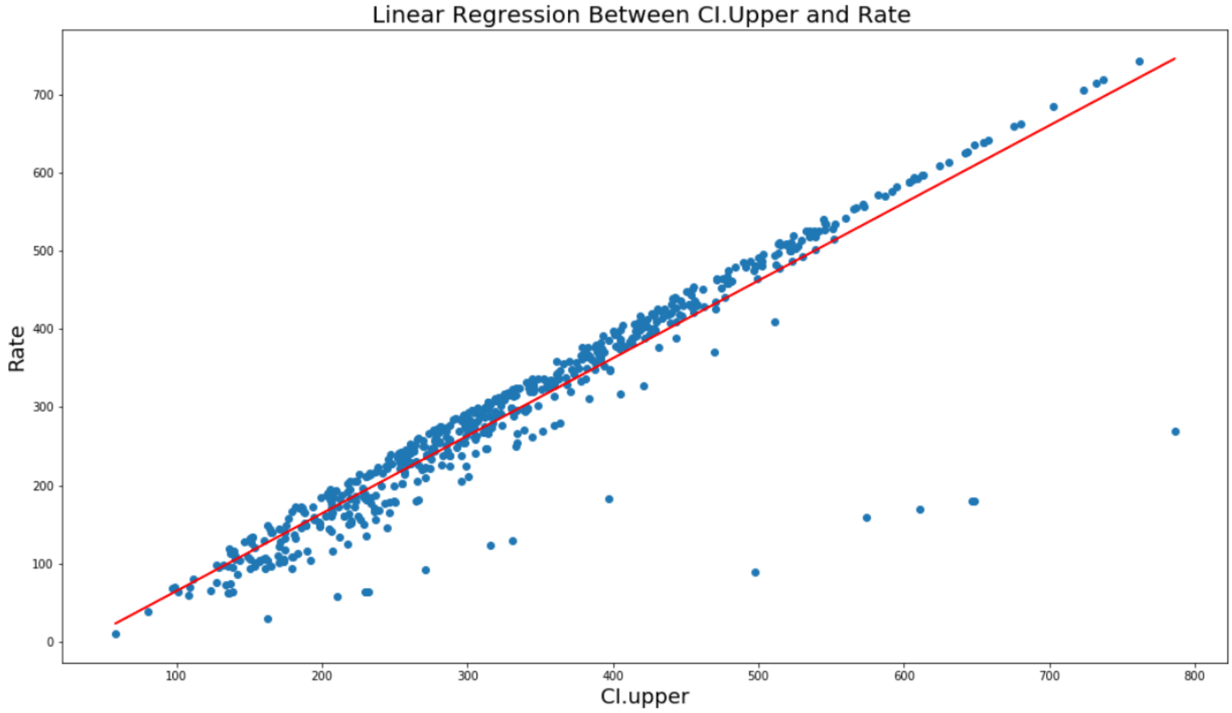


From the figure above shows a graph that there will be an increase in Chlamydia disease in future. We choose the line graph so that we can easily see the overall rates of Chlamydia disease throughout the year.

The graph is decreasing due to the influenza A(H1N1) pandemic outbreak. According to the news online from cdc.gov, The first cases of 2009 H1N1 pandemic influenza were reported by CDC on April 21, 2009. Therefore, most people do not want to make close contact with people sexually during that year.

Question 4

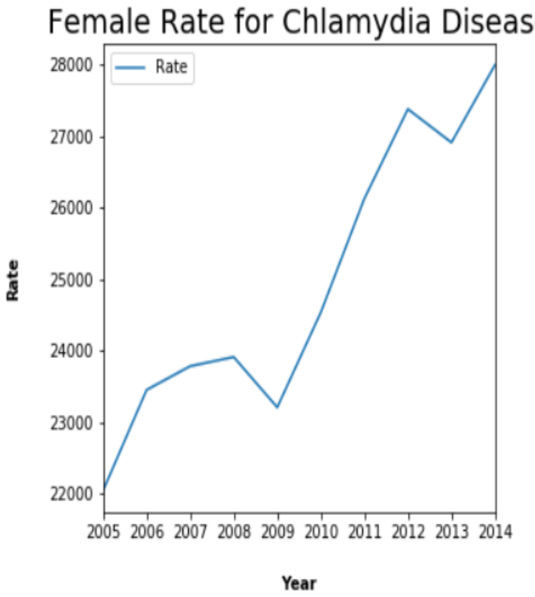
What is the relationship between CI.Upper and Rate for the Chlamydia disease?

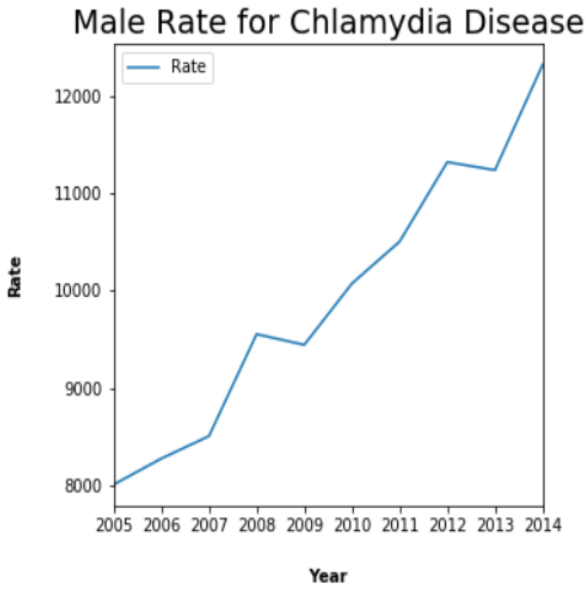


The figure above shows graph of Cl.Upper and Rate are nearly close to each other. Therefore, we can predict that Cl.Upper will be increasing in value followed by Rate. CI.Upper has the upper 95% of confidence for an individual that has Chlamydia disease. Therefore we can see in the linear regression graph that CI.upper and Rate have a linear relationship because most of the data tend to fit with the best fit line. We can expect that the Rate would increase when the CI.upper increase.

Question 5

Compare the rate of having Chlamydia disease between male and female.





According to the both graph above, female have the highest rate of having Chlamydia disease compared to male.

From the female graph, the rates of the Chlamydia disease increased drastically from the year of 2009 until 2012.

From the male graph, the rates of Chlamydia disease will always increase but the the number of rates are still lesser than the female number of Chlamydia rates.

According to Napa County, California Health and Human Services Agency, Public Health Division, Chlamydia infection is usually common in young females due to it becoming asymptomatic in male.

Conclusions

Throughout the assignment, we have learnt a lot of things individually and as a group. We learnt the values to work together despite disputes with one another. We also gain a lot of knowledge on how to analyze the data, learn a little about Chlamydia disease and we also learn what type of graph should be used for the data to make it understandable.

After we are done with analyzing the data for this assignment, we can conclude that Chlamydia has the highest rate of infectious disease from the year 2005 until 2006. We can also see that Kern is the county that has the highest count of an individual that has Chlamydia disease followed by Fresno and San Francisco.

Overall, we can conclude that the rate of a female to be infected with Chlamydia disease is higher than man. The reason behind this biologically is because the women's sex organ are more exposed and vulnerable to sexually transmitted diseases than male's sex organ.

Some of the challenges that we have encountered during processing the data is to try and combine the line graph for male and female. This is because in the data that we found, the sex column contains total, male and female. Therefore it makes it hard for us to take out the value and combine it. We tried to combine it but the graph ended up gave the wrong output. We were hoping to do more analysis for the female and male, but didn't manage to do so.

Citations

1. https://www.cdc.gov/std/chlamydia/stdfact-chlamydia.htm
2. https://medlineplus.gov/chlamydiainfections.html
3. http://worldpopulationreview.com/us-counties/ca/
4. https://www.countyofnapa.org/DocumentCenter/View/1604/2011-2013-Napa-County-STD-Report-PDF
5. https://www.medicalnewstoday.com/articles/8181
6. https://www.cdc.gov/mmwr/preview/mmwrhtml/mm5851a3.htm
7. https://academic.oup.com/jid/article/207/1/30/875048 [1]