Analysis of STD Infection Rates in America: A data Visualization Approach

David Apamo

2024-11-13

Introduction

Sexually transmitted diseases (STDs) remain a significant public health concern in the United States, with millions of new cases being reported each year. According to CDC 2022 report on STI surveillance, "*more than 2.5 million cases of syphilis, gonorrhea and chlamydia were reported in the United States. The most alarming concerns center around syphilis and congenital syphilis epidemics, signaling an urgent need for swift innovation and collaboration from all STI prevention partners.*" These infections not only pose immediate health risks, but also have long-term consequences, including chronic pain, infertility in women, stillbirths, low-birth weights and prematurity, neonatal deaths and increased risk of HIV acquisition (WHO, 2023). Understanding the trends, patterns and risk factors associated with STDs is essential for the formulation of effective prevention and control measures.

```
# Load the required packages
suppressMessages({
  library(tidyverse)
  library(maps)
  library(usmap)
})

# Import the STD Cases data set
STD_Cases <- read_csv("STD Cases.csv")

## Rows: 42680 Columns: 10
## ── Column specification
──────────────────────────────────────────────────────
## Delimiter: ","
## chr (5): Disease, State, Gender, Age, Age Code
## dbl (5): Disease Code, Year, STD Cases, Population, Rate per 100K
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

The data has 42,680 observations and 10 variables. The variables disease, state, gender, age and age code are characters, while the rest of the variables are of numeric type (double).

```r
# View the first few observations
head(STD_Cases)
```

```
## # A tibble: 6 × 10
##   Disease   `Disease Code` State    Year Gender Age       `Age Code` `STD
Cases`
##   <chr>              <dbl> <chr>   <dbl> <chr>  <chr>     <chr>
<dbl>
## 1 Chlamydia            274 Alabama  1996 Male   0-14 yea… 0-14
25
## 2 Chlamydia            274 Alabama  1996 Male   15-19 ye… 15-19
164
## 3 Chlamydia            274 Alabama  1996 Male   20-24 ye… 20-24
193
## 4 Chlamydia            274 Alabama  1996 Male   25-29 ye… 25-29
88
## 5 Chlamydia            274 Alabama  1996 Male   30-34 ye… 30-34
55
## 6 Chlamydia            274 Alabama  1996 Male   35-39 ye… 35-39
30
## # ℹ 2 more variables: Population <dbl>, `Rate per 100K` <dbl>
```

Data Cleaning and Preprocessing

I'll clean the data by first cleaning variable names, converting character variables to factors and addressing data quality issues like missing data, duplicated observations, white spaces and inconsistencies.

```r
# Clean variable names
STD_Cases <- janitor::clean_names(STD_Cases)

# Check for missing values
map_dbl(STD_Cases, ~sum(is.na(.)))
```

```
##       disease  disease_code         state          year        gender
##           150           150           150           150           150
##           age      age_code     std_cases    population rate_per_100k
##           150           150           150          7197          7197
```

There are several cases with missing values in the data. The variables disease, disease code, state, year, gender, age, age code and STD cases have 150 missing values each, while population and rate per 100k have 7,197 missing values each. Since the data is very large, I will drop the observations with missing values. I still won't loose much information.

```r
# Check for duplicated observations
sum(duplicated(STD_Cases))
```

```
## [1] 149
```

There are 149 duplicated observations in the data. These mainly consist of the rows which had missing values for all the variables.

```
# Drop the observations with missing values
STD_Cases <- na.omit(STD_Cases)

# Drop the duplicated observations
STD_Cases <- distinct(STD_Cases)
```

After dropping NAs, no duplicates remained in the data. The duplicated observations were definitely rows which had missing values for all the variables.

```
# Convert the character variables disease, state, gender, age and age_code to
factors

# Specify the columns to factor
cols_to_factor <- c("disease", "disease_code", "state", "gender", "age",
"age_code")
# Convert the specified columns to factors
STD_Cases <- STD_Cases %>% mutate_at(.vars = cols_to_factor, .funs = factor)

# Generate summary statistics for all the variables
summary(STD_Cases)
```

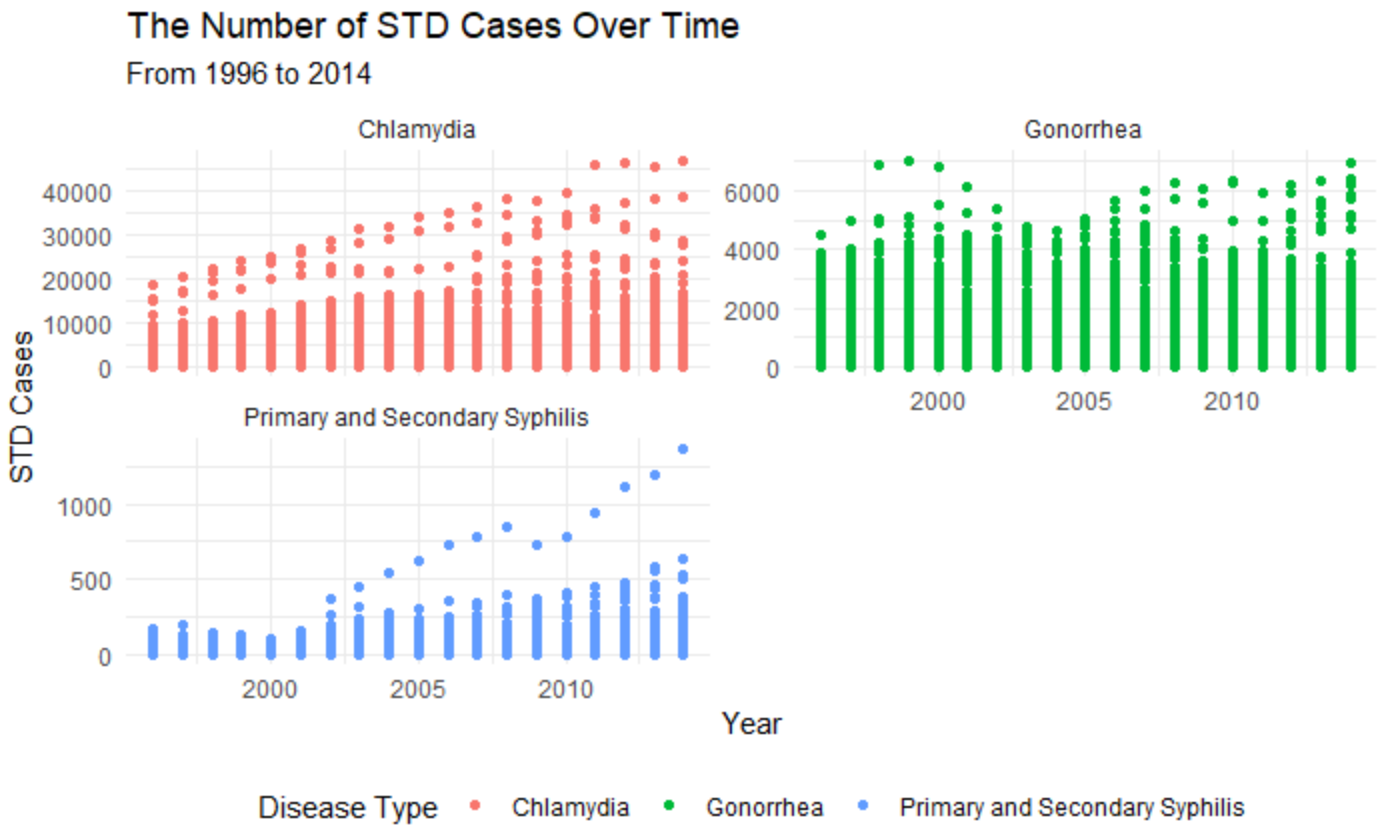
```
##                              disease      disease_code            state
##  Chlamydia                       :13512   274:13512    Louisiana     :  784
##  Gonorrhea                       :13233   280:13233    Texas         :  783
##  Primary and Secondary Syphilis: 8738   310: 8738    Florida       :  779
##                                                      California    :  776
##                                                      Tennessee     :  776
##                                                      North Carolina:  774
##                                                      (Other)       :30811
##       year          gender                age         age_code
##  Min.   :1996   Female:17351   0-14 years :3867   0-14 :3867
##  1st Qu.:2000   Male  :18132   15-19 years:5127   15-19:5127
##  Median :2005                  20-24 years:5336   20-24:5336
##  Mean   :2005                  25-29 years:5287   25-29:5287
##  3rd Qu.:2010                  30-34 years:5271   30-34:5271
##  Max.   :2014                  35-39 years:5240   35-39:5240
##                                40+ years  :5355   40+  :5355
##    std_cases          population      rate_per_100k
##  Min.   :    1.0   Min.   :  12937   Min.   :   0.02
##  1st Qu.:   16.0   1st Qu.:  94181   1st Qu.:   6.95
##  Median :  109.0   Median : 196628   Median :  52.89
##  Mean   :  716.9   Mean   : 438979   Mean   : 325.66
##  3rd Qu.:  519.0   3rd Qu.: 432225   3rd Qu.: 302.95
##  Max.   :46885.0   Max.   :8880836   Max.   :9078.95
##
```

- The data contains three types of STDs, with Chlamydia being the most frequent STD, followed by Gonorrhea then Primary and Secondary Syphilis (13,512, 13,233 and 8,738 cases respectively).
- Males were mostly affected by STD than females. (18,132 compared to 17, 351)
- The age group 40+ years were the mostly affected, followed by age group 20-24 years, then 25-29 years, 30-34 years, 35-39 years, 15-19 years and 0-14 years respectively.
- The mean and median number of STD cases was 716.9 and 109.0 respectively. The difference between the mean and the median is large, implying that the number of STD cases has a skewed distribution.
- The mean and median population was 438,979 and 196,628 respectively.
- The highest recorded STD rate per 100k population was 9078.95

Visualizing the data

```
## Visualize the number of STD cases over time

# define grid and aesthetic mappings
ggplot(STD_Cases, aes(x = year, y = std_cases, colour = disease)) +
  # add a scatter plot layer with data points
  geom_point() +
  # facet the grid by disease type
  facet_wrap(~disease, scales = "free_y", nrow = 2) +
  # add title and label the axes
  labs(title = "The Number of STD Cases Over Time",
       subtitle = "From 1996 to 2014",
       x = "Year", y = "STD Cases",
       color = "Disease Type") +
  # add minimal theme
  theme_minimal() +
  # place the legend at the bottom
  theme(legend.position = "bottom")
```



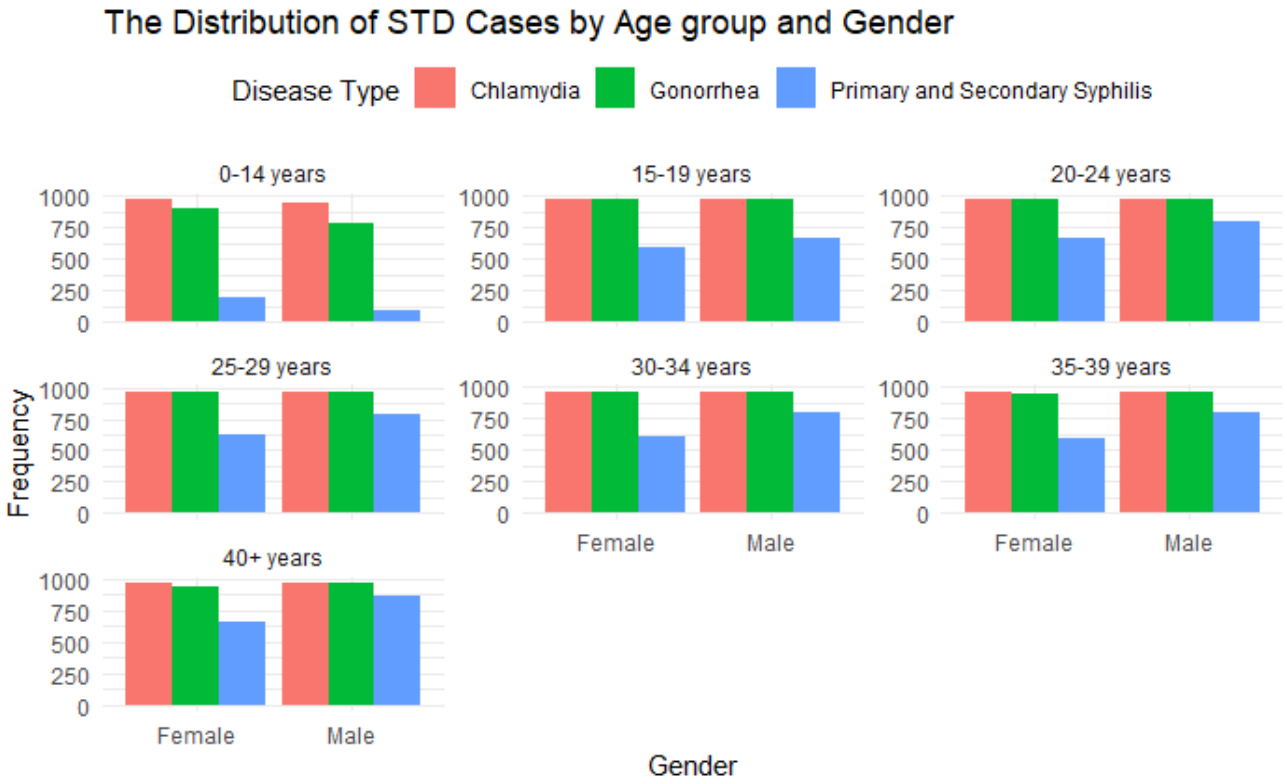
Generally, Chlamydia and Syphilis cases increases with time for various US states. Gonorrhea on the other hand, doesn't show much increase over time, but has high number of cases all through.

The increasing cases of Chlamydia and Syphilis can be attributed to enhanced screening and diagnostic efforts, and more cases are being detected. Public health campaigns and awareness about the risks associated with STD, has also played a significant role in encouraging people to go for testing, hence more cases are being reported. Also, many cases of Chlamydia are asymptomatic and individuals may unknowingly spread the infection. The surging cases of syphilis are attributed to the rising number of unprotected sex among men who have sex with fellow men (CDC, 2018 and Belluz, 2019).

```
## visualize the distribution of STD cases by age group and gender

# define grid and aesthetic mappings
ggplot(STD_Cases, aes(x = gender, fill = disease)) +
  # add a bar plot layer and place the bars side by side
  geom_bar(position = "dodge") +
  # facet the grid by age group
  facet_wrap(~age, scales = "free_y") +
  # add title and label the axes
  labs(title = "The Distribution of STD Cases by Age group and Gender",
       x = "Gender", y = "Frequency",
```

```
      fill = "Disease Type") +
  # add a minimal theme
  theme_minimal() +
  # place the legend at the top
  theme(legend.position = "top")
```



- For age group 0-14, Syphilis was more common in females than males. This is different for the other age groups where Syphilis was more common in males than in females.
- Gonorrhea and Chlamydia were more common in all the age groups, both for males and females.
- For all age groups, except age group 0-14, males and females had nearly the same number of cases for chlamydia and gonorrhea.
- Syphilis was mostly common in males aged 40 and above, but was also common in age groups 20-24, 25-29, 30-34 and 35-39.

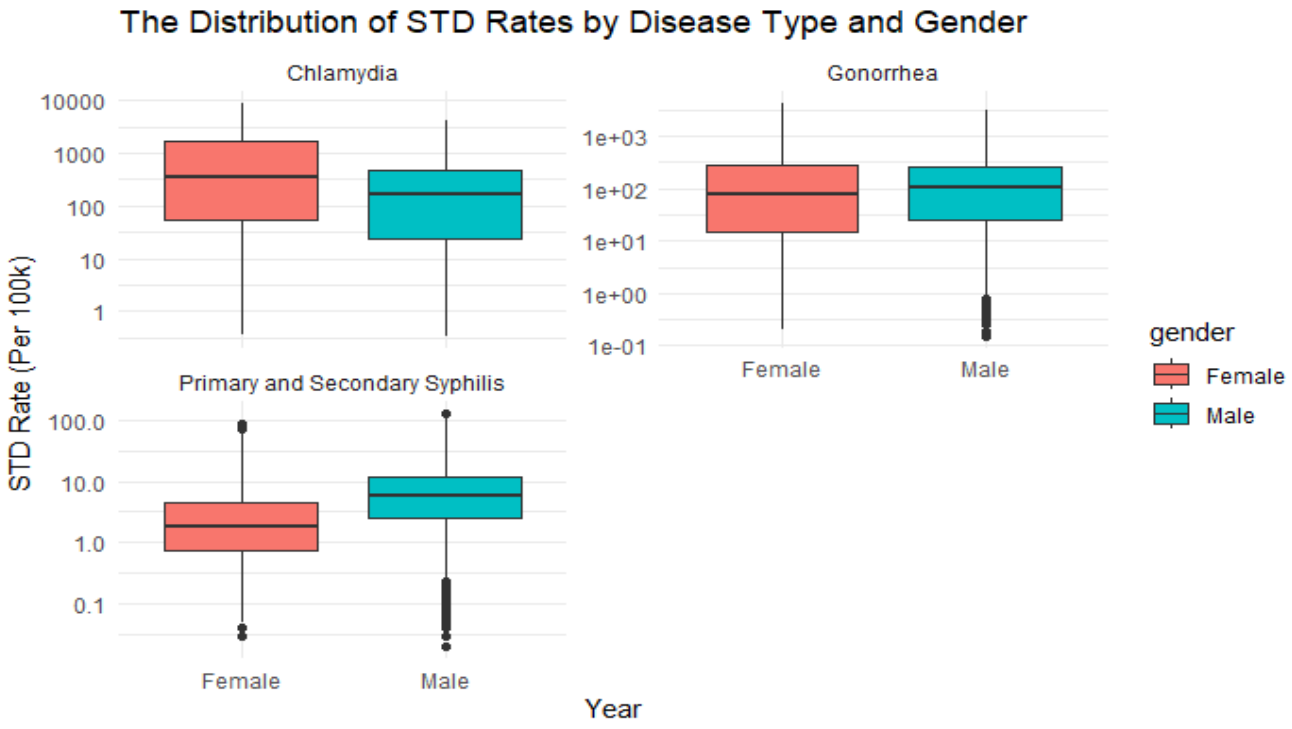
The few cases of syphilis in children aged 0-14 years are mainly attributed to mother-to-child transmission. According to (CDC, 2018), "*more women in America are getting syphilis and they're passing it to their babies. Congenital syphilis is associated with serious health consequences, like stillbirths and neonatal deaths.*" There has been a rise in the number of newborn deaths linked to congenital syphilis.

The high number of syphilis cases among males is mainly attributed to high-risk sexual behavior, such as men who have unprotected sex with other men, and also having multiple sexual partners (Belluz, 2019).

The surge in Chlamydia and Gonorrhea cases is attributed to the asymptomatic nature of many Chlamydia cases and the antibiotic resistance in gonorrhea. Improved healthcare systems and advancement in technology has also led to the detection and reporting of more STD cases (CDC, 2018 and Belluz, 2019).

```
## Compare the distribution of STD rate by disease type and gender

# define grid and aesthetic mapping
ggplot(STD_Cases, aes(x = gender, y = rate_per_100k, fill = gender)) +
  # add a box plot layer
  geom_boxplot() +
  # use a log scale on the y-axis for easy visibility
  scale_y_log10() +
  # facet the grid by disease type and use free scale on the y-axis
  facet_wrap(~disease, scales = "free_y", nrow = 2) +
  # add title and label the axes
  labs( title = "The Distribution of STD Rates by Disease Type and Gender",
        x = "Year", y = "STD Rate (Per 100k)") +
  # add minimal theme
  theme_minimal()
```



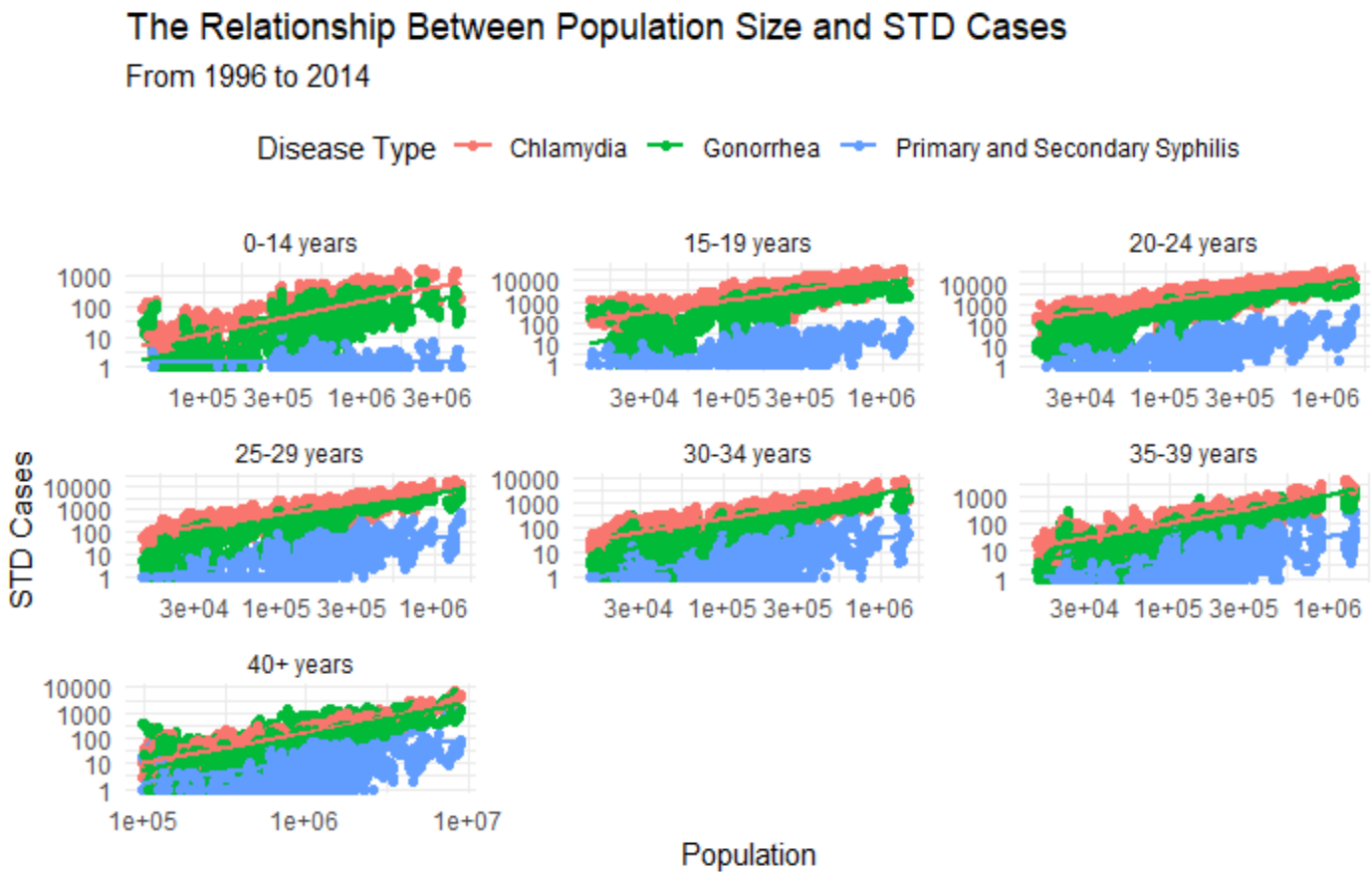
- Chlamydia was highly prevalent in females than males, while Syphilis was highly prevalent in males than females.
- Gonorrhea was also slightly more prevalent in females than in males.
- Some states had extremely low prevalence for syphilis and gonorrhea in males.

The high prevalence of Chlamydia in females is attributed to the increased number of females going for screening. According to (CDC, 2018), young women have been targeted for routine Chlamydia screening due to asymptomatic nature of the disease, posting a risk of infertility among women. So more testing has led to an increased number of cases being detected. The high prevalence for gonorrhea can be attributed to emergence of antibiotic-resistant strains of *Neisseria gonorrhoeae* that makes treatment of the disease more difficult and can lead to higher transmission rates. The high prevalence of syphilis in males is attributed to high-risk behaviors, such as men having unprotected sex with other men and also people who inject drugs.

```
## visualize the relationship between Population size and number of STD cases

# define grid and aesthetic mappings
ggplot(STD_Cases, aes(x = population, y = std_cases, colour = disease)) +
  # add a scatter plot layer with points
  geom_point() +
  # add a smoothing layer with a linear trend
  geom_smooth(method = lm, se = FALSE) +
  # use log scale on both axes for easy visibility
  scale_y_log10() + scale_x_log10() +
  # facet the grid by age group
  facet_wrap(~ age, scales = "free") +
  # add title and label the axes
  labs(title = "The Relationship Between Population Size and STD Cases",
       subtitle = "From 1996 to 2014",
       x = "Population", y = "STD Cases",
       color = "Disease Type") +
  # add a minimal theme
  theme_minimal() +
  # place the legend at the top
  theme(legend.position = "top")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



There's a linear relationship between Population size and chlamydia and gonorrhea cases. Chlamydia and gonorrhea cases tend to increase as population increases. Syphilis on the other hand, has a weak linear relationship with population.

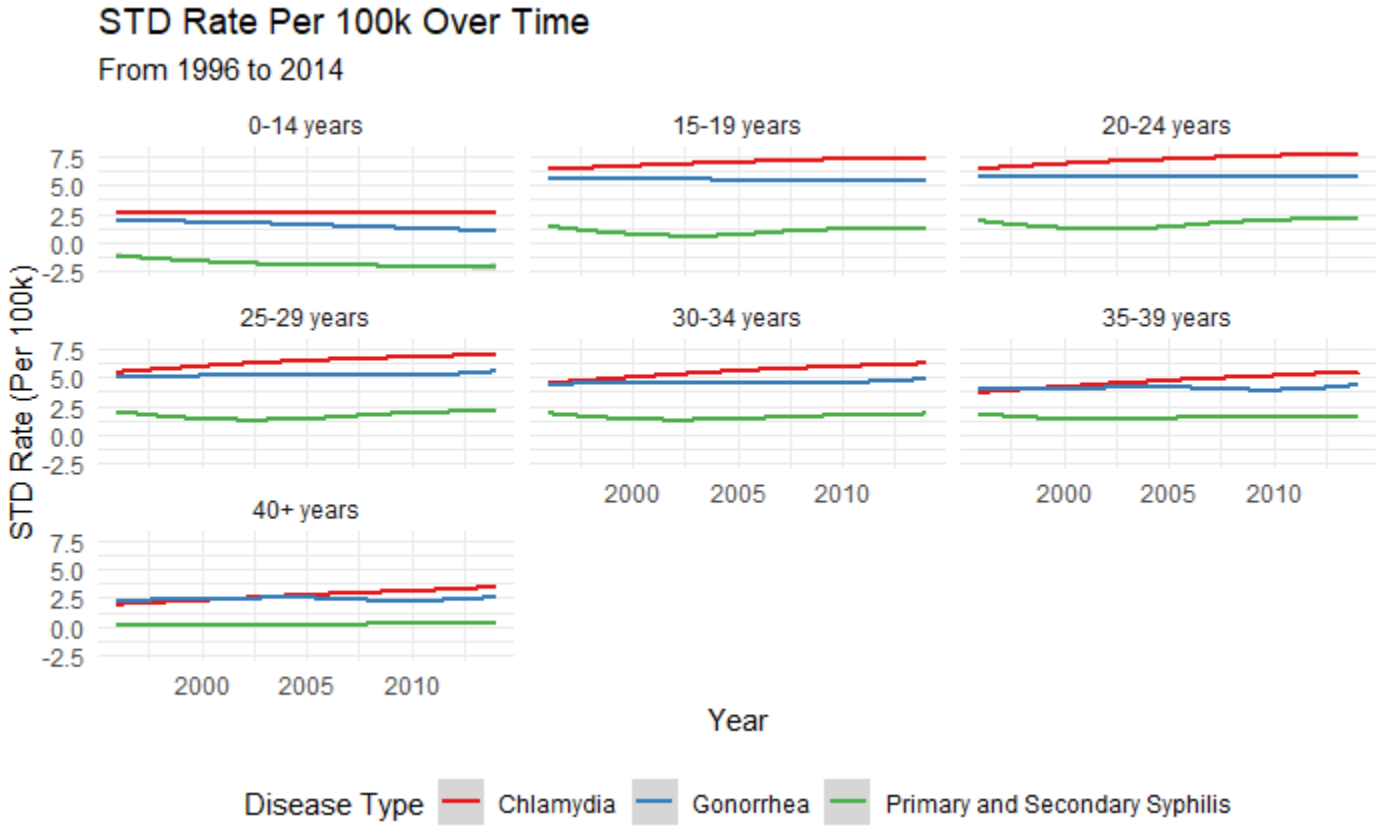
The positive correlation between population and STD cases is attributed to the fact that higher population, especially in urban areas, can facilitate the spread of STD due to closer social interactions and potential growth in sexual networks. Urban areas also have better access to testing facilities, and this leads to more cases being reported. The high population mostly comprise of the youthful age groups who are more likely to engage in risky sexual behaviors, thus increasing the spread of STD (Belluz, 2019). Syphilis has multiple stages of development, and it can be more challenging to diagnose and report accurately, especially in its latent stages. Therefore, the disease may not be consistently reported across different states or regions, leading to weaker correlations with population size. The stigma associated with syphilis and the differences in public health campaigns and screening efforts across states, can also affect reporting rates (CDC, 2018).

```
## visualize STD rate per 100k over time

# define grid and aesthetic mappings
ggplot(STD_Cases, aes(x = year, y = log(rate_per_100k), colour = disease)) +
  # add a smoothing layer to visualize the trend
```

```
  geom_smooth() +
  # facet the grid by age group
  facet_wrap(~age) +
  # add title, subtitle and label the axes
  labs(title = "STD Rate Per 100k Over Time",
       subtitle = "From 1996 to 2014",
       x = "Year", y = "STD Rate (Per 100k)",
       color = "Disease Type") +
  # add a minimal theme
  theme_minimal() +
  # place the legend at the bottom
  theme(legend.position = "bottom") +
  # choose a color palette
  scale_color_brewer(palette = "Set1")
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



- For age group 0-14, the rate of chlamydia/100k population remained constant over time, while the rates of gonorrhea and syphilis decreased over time.
- For the remaining age groups, the rate of chlamydia/100k population increased steadily with time from 1996 through to 2014.
- The prevalence for gonorrhea was roughly constant for age groups 15-19 and 20-24. For age groups, 25-29 and 30-34, the prevalence for gonorrhea increased

slightly with time, while for age groups 35-39 and 40+ years, the prevalence for gonorrhea increased steadily up to around 2006, decreased slightly towards 2010 then increased again.

- For age group 40+ years, the prevalence for syphilis was roughly constant up to around 2008, then rose steadily up-to 2014. For age groups 15-19, 20-24, 25-29, 30-34 and 25-39 the prevalence for syphilis decreased first for some years then increased up-to 2014.

The constant rate of Chlamydia and the decreasing trends in gonorrhea and syphilis among children aged 0-14 years, is attributed to low rates of sexual activity within this age group, posting a lower risk of sexual transmission (CDC, 2018). The decreasing trends among age group 0-14, are also due to improved prevention measures, such as better maternal healthcare and prenatal screening, and effective public health interventions and awareness campaigns targeted at preventing mother-to-child transmission, potentially reducing the risk of stillbirths and neonatal deaths (CDC, 2018 and WHO 2023).

Among other age groups, the increasing trends are attributed to several factors, such as high risk behavior (more so among the youthful age groups), improved testing and screening techniques and better access to healthcare, leading to detection of more cases. Another factor contributing to the high prevalence for STD is the rise in dating apps, which has made sex more readily available and anonymous. This has derailed the tracking of STD outbreaks by health practitioners (Belluz, 2019).

Creating Maps

Based on the previous visualizations, the population, STD cases and prevalence (rate_per_100k) were all high in 2014. Therefore, I will map the STD cases and prevalence for the year 2014. I'll create three maps; the first one to show states with few (less than 100) number of STD cases, the second one to show states with high number of STD cases (more than 2000) and the third one to show countries with high prevalence for STD (rate_per_100k greater than 4000).

When creating the maps, I'll use the plot_usmap() function from "usmap" package. The function is easy to use, as it takes few arguments and makes it easier to label each state on the map.

```
# Load the state population data
data("statepop")
# View the state population data
view(statepop)
```

The "statepop" data contains the population estimates of the US states for the year 2022. The data has 52 observations and 4 variables. This data has both the state names and abbreviations, which will help me label the states on my maps for easy readability.

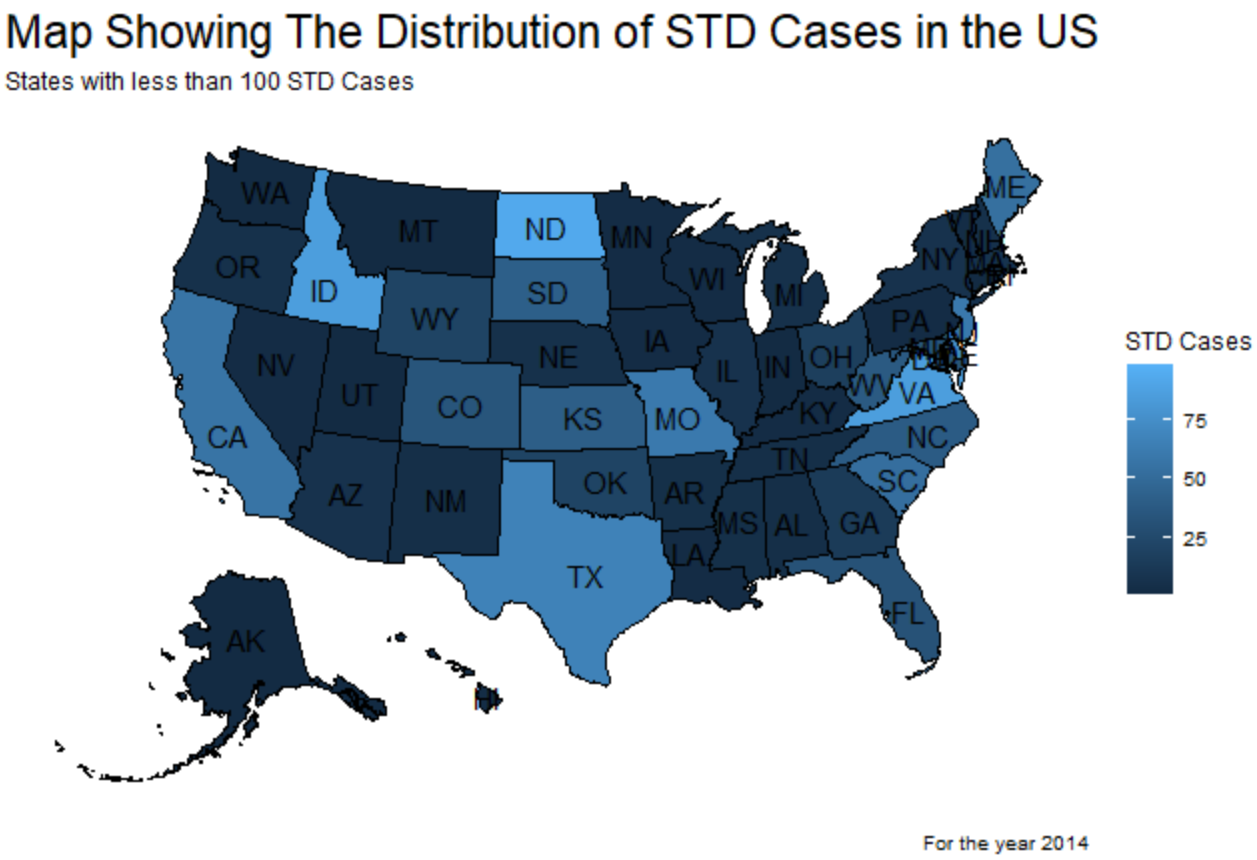
The STD Cases data has the state names in the state column, while statepop data has the state names in column named full. I will therefore join the two data sets by state name using right_join() function from dplyr package.

```
# Join the US State Population data with STD_Cases data set
Joined_data <- STD_Cases %>% right_join(statepop,
                                        # join by state name
                                        by = c("state" = "full"))

## Map states with few number STD cases, for the year 2014

# First filter the data for 2014, where the number of STD cases is less than
100
df1 <- Joined_data %>% filter(year == 2014 & std_cases < 100)

# plot the US map, fill by values of std_cases and label each state with its
abbreviation
plot_usmap(data = df1, values = "std_cases", labels = "TRUE") +
  # add a title, subtitle and a caption
  labs(title = "Map Showing The Distribution of STD Cases in the US",
       subtitle = "States with less than 100 STD Cases",
       caption = "For the year 2014",
       fill = "STD Cases") +
  # increase title font size and place the legend on the right side of the
plot
  theme(plot.title = element_text(size = 17),
        legend.position = "right")
```



Various states had less than 100 STD cases for certain age groups in the year 2014. The states include;

- Montana, Washington, Oregon, Nevada, Utah, Arizona, New Mexico, Minnesota, Wisconsin, Iowa, Arkansas, Louisiana, Mississippi, Kentucky, New York and Pennsylvania.

However, some of these states have less population and so the prevalence for STD might still be high, despite them having few number of STD cases within certain age groups.

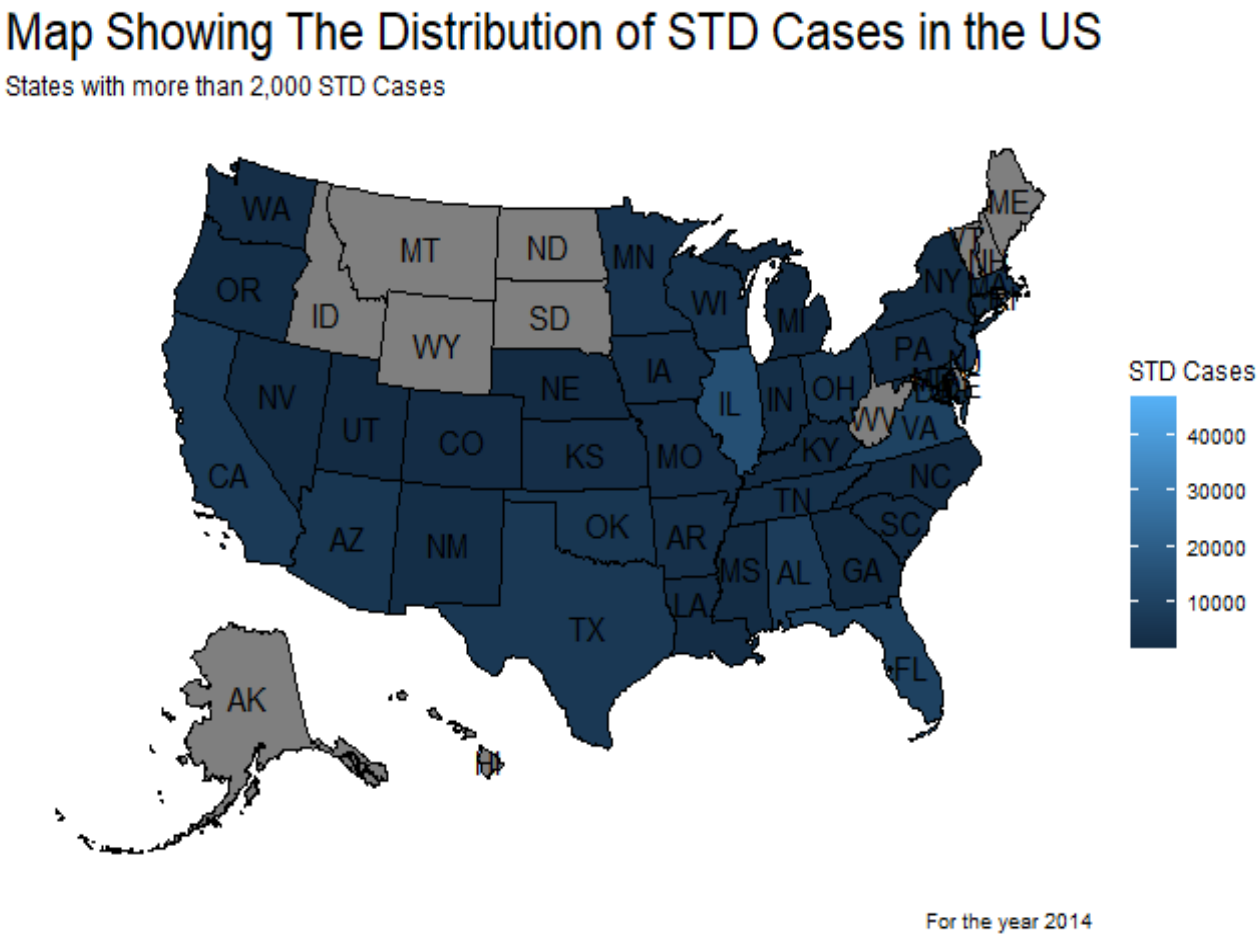
States like Montana, North Dakota, New Mexico and Utah have low population densities, and there are fewer individuals at risk of contracting STDs. Some of these states with fewer STD cases have robust healthcare systems, helping in effective management and prevention of STDs.

```
## Map states with high number of STD cases, for the year 2014

# Filter the data for 2014, where the number of STD cases is greater than
2000
df2 <- Joined_data %>% filter(year == 2014 & std_cases > 2000)

# plot the US map, fill by values of std_cases and label each state with its
abbreviation
```

```
plot_usmap(data = df2, values = "std_cases", labels = "TRUE") +
  # add a title, subtitle and a caption
  labs(title = "Map Showing The Distribution of STD Cases in the US",
       subtitle = "States with more than 2,000 STD Cases",
       caption = "For the year 2014", fill = "STD Cases") +
  # increase title font size and place the legend on the right side of the
plot
  theme(plot.title = element_text(size = 17),
        legend.position = "right")
```



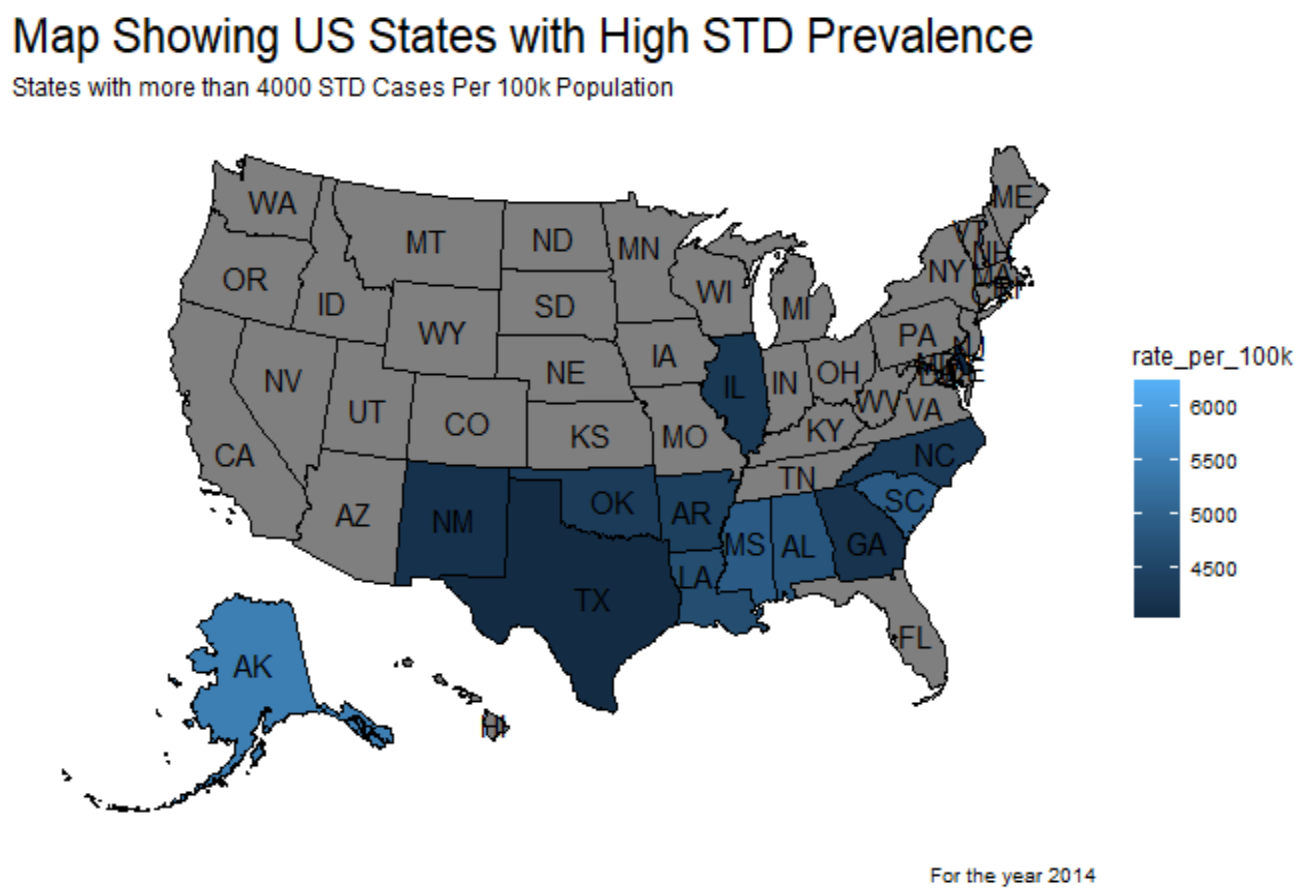
Illinois, Alabama, Virginia and Florida had the highest number of STD cases. California, Arizona, Texas and Oklahoma also had high number of STD cases. The age groups dominating these cases are age group 15-19, 20-24, 25-29 and 30-34.

These states with high number of STD cases are generally well-developed and consist of many urban areas. They have high population densities, leading to increased social and sexual networks. They also have more robust healthcare systems and better access to healthcare and testing facilities, leading to increased reporting of more STD cases.

```
## Map states with high STD prevalence, for the year 2014
```

```r
# Filter the data for 2014, where rate/100k is greater than 4000
df3 <- Joined_data %>% filter(year == 2014 & rate_per_100k > 4000)

# plot the US map, fill by rate_per_100k and label each state with its
abbreviation
plot_usmap(data = df3, values = "rate_per_100k", labels = "TRUE") +
  # add a title, subtitle and a caption
  labs(title = "Map Showing US States with High STD Prevalence",
       subtitle = "States with more than 4000 STD Cases Per 100k Population",
       caption = "For the year 2014") +
  # increase title font size and place the legend on the right side of the
plot
  theme(plot.title = element_text(size = 17),
        legend.position = "right")
```



The sates with the highest prevalence for STD in 2014 were;

- Alaska, Mississippi, South Carolina, Alabama and Louisiana.
- New Mexico, Texas, Oklahoma, Arkansas, Illinois, Georgia and North Carolina also had high prevalence for STD.
- This high prevalence was precisely for Chlamydia, and the age groups dominating these cases are age group 15-19 and 20-24.

These are developed states with high population densities. High population densities imply closer social interactions and potential increase in the number of sexual partners, thereby facilitating the spread of STDs (Belluz, 2019). Other factors contributing to the high prevalence are high-risk behavior, such as unprotected sex among gays and men who have sex with other men, robust healthcare systems and better access to healthcare and testing facilities, leading to detection and reporting of more cases. These states need targeted interventions in order to contain the spread of STDs.

References

Belluz, J. (2019, October 10). 5 reasons why 3 STDs are roaring back in America: The CDC found spikes in cases of syphilis, gonorrhea, and chlamydia — again. Vox. https://www.vox.com/science-and-health/2017/9/27/16371142/stds-syphilis-gonorrhea-chlamydia

Centers for Disease Control and Prevention. (2023). Sexually Transmitted Infections Surveillance, 2022. U.S. Department of Health and Human Services. https://www.cdc.gov/std/statistics/2022/default.htm

World Health Organization. (2023, July 10). Sexually transmitted infections (STIs). https://www.who.int/news-room/fact-sheets/detail/sexually-transmitted-infections-(stis)