Lyme_Disease_Analysis

Nikhil Eatalpacka

2024-05-07

```
install.packages("reshape2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
## (as 'lib' is unspecified)
```

```
library(ggplot2)
library(reshape2)
library(dplyr)
```

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

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

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

```
library(readxl)
```

```
data <- read_excel("Lyme disease case by state or locality.xlsx")
# Print the first few rows of the dataset to verify the import
head(data)
```

```
## # A tibble: 6 x 16
##   State    `2010` `2011` `2012` `2013` `2014` `2015` `2016` `2017` `2018` `2019`
##   <chr>     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1 Connect~   3896   4156   3068   3039   2657   2925   2360   2541   1748   2051
## 2 Maine       908    970    751   1006   1111   1373   1401   1201   1487   1850
## 3 Maryland   2218   2024   1617   1351   1651   1197   1373   1728   1866   1891
## 4 Massach~   4582   5256   3263   2476   5138   5290   5304   4224    198    410
## 5 New Ham~   1601   1415   1339   1299   1450   1687    724    529    891   1381
## 6 New Jer~   3485   4973   3712   4262   3576   3766   3286   4855   4350   5092
## # i 5 more variables: `2020` <dbl>, `2021` <dbl>, `2022` <dbl>, `2023` <dbl>,
## #   `2024` <dbl>
```

```
# Melt the data
melted_data <- melt(data, id.vars = "State", variable.name = "Year", value.name = "Cases")
# Print the first few rows of the melted data to verify
head(melted_data)
```

```
##          State Year Cases
## 1   Connecticut 2010  3896
## 2        Maine 2010   908
```

```
## 3      Maryland 2010  2218
## 4 Massachusetts 2010  4582
## 5 New Hampshire 2010  1601
## 6    New Jersey 2010  3485
```

```
# Calculate the average number of cases for each state
data <- data %>%
  mutate(Average = rowMeans(.[,2:ncol(.)], na.rm = TRUE))

# Order the data by average cases in descending order
data <- data[order(-data$Average),]
summary(data)
```

```
##     State                2010           2011           2012           2013
##  Length:10          Min.   : 404   Min.   : 408   Min.   : 356   Min.   : 623
##  Class :character   1st Qu.:1100   1st Qu.:1081   1st Qu.:1268   1st Qu.:1092
##  Mode  :character   Median :2852   Median :3090   Median :2342   Median :1914
##                     Mean   :2964   Mean   :3148   Mean   :2258   Mean   :2493
##                     3rd Qu.:3876   3rd Qu.:5185   3rd Qu.:3384   3rd Qu.:3956
##                     Max.   :7794   Max.   :5722   Max.   :3805   Max.   :5362
##       2014           2015           2016           2017           2018
##  Min.   : 522   Min.   : 893   Min.   : 599   Min.   : 529   Min.   :  198
##  1st Qu.:1196   1st Qu.:1324   1st Qu.:1353   1st Qu.:1286   1st Qu.: 1006
##  Median :2154   Median :2306   Median :1880   Median :2134   Median : 1618
##  Mean   :2525   Mean   :2881   Mean   :2762   Mean   :3069   Mean   : 2798
##  3rd Qu.:3432   3rd Qu.:4403   3rd Qu.:3624   3rd Qu.:4292   3rd Qu.: 3378
##  Max.   :5138   Max.   :5758   Max.   :7487   Max.   :9048   Max.   :11443
##       2019           2020           2021           2022           2023
##  Min.   :  410   Min.   :   16   Min.   :   7   Min.   :  59   Min.   :  27.0
##  1st Qu.: 1450   1st Qu.: 1200   1st Qu.:1208   1st Qu.: 363   1st Qu.: 433.0
##  Median : 1870   Median : 1416   Median :1564   Median : 728   Median : 810.5
##  Mean   : 3248   Mean   : 2565   Mean   :2566   Mean   :1174   Mean   :1363.5
##  3rd Qu.: 4332   3rd Qu.: 3193   3rd Qu.:3256   3rd Qu.:1962   3rd Qu.:2552.5
##  Max.   :11900   Max.   :10208   Max.   :8998   Max.   :3334   Max.   :3518.0
##       2024          Average
##  Min.   : 1085   Min.   : 632.9
##  1st Qu.: 1558   1st Qu.:1219.6
##  Median : 2344   Median :1940.6
##  Mean   : 4667   Mean   :2698.7
##  3rd Qu.: 5686   3rd Qu.:3737.0
##  Max.   :16798   Max.   :6881.9
```

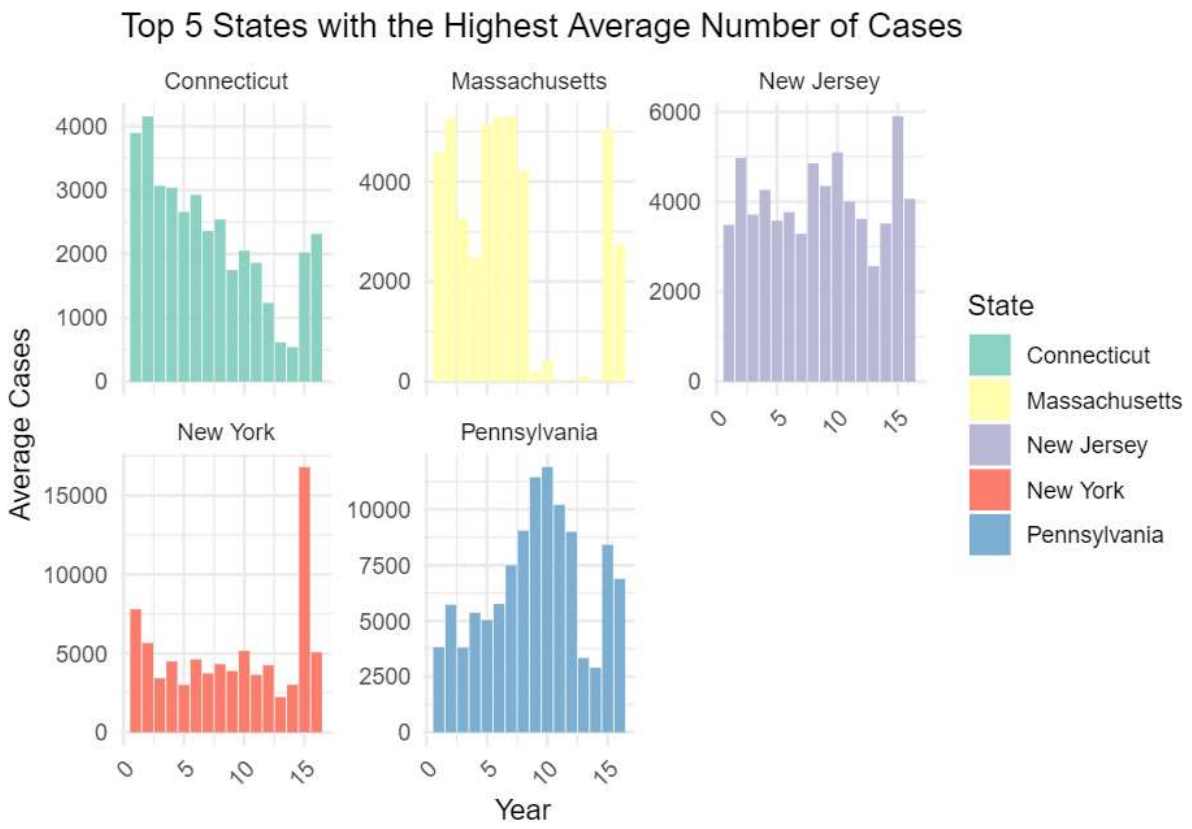
```
# Select the top 5 states with the highest average number of cases
top_5_states <- head(data$State, 5)


# Filter the data for the top 5 states
top_5_data <- data[data$State %in% top_5_states,]
summary(top_5_data)
```

```
##     State                2010           2011           2012           2013
##  Length:5           Min.   :3485   Min.   :4156   Min.   :3068   Min.   :2476
##  Class :character   1st Qu.:3818   1st Qu.:4973   1st Qu.:3263   1st Qu.:3039
##  Mode  :character   Median :3896   Median :5256   Median :3425   Median :4262
##                     Mean   :4715   Mean   :5152   Mean   :3455   Mean   :3926
##                     3rd Qu.:4582   3rd Qu.:5651   3rd Qu.:3712   3rd Qu.:4490
```

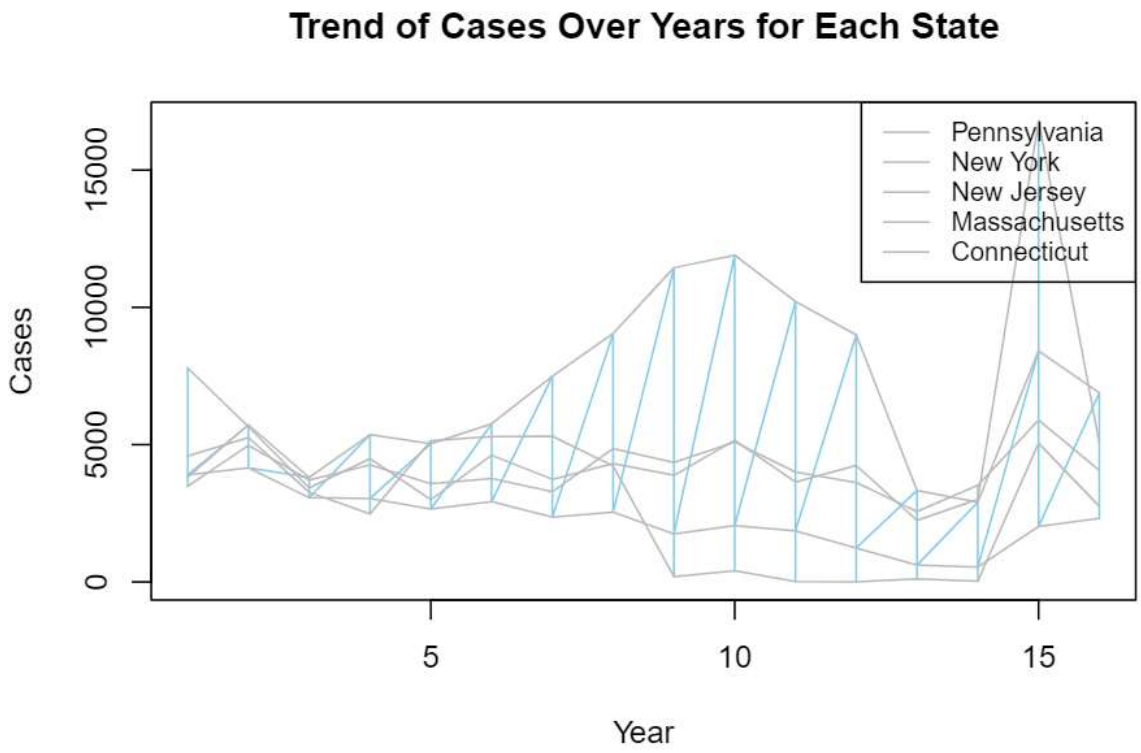
```
##                   Max.   :7794   Max.   :5722   Max.   :3805   Max.   :5362
##       2014            2015           2016           2017           2018
##  Min.   :2657   Min.   :2925   Min.   :2360   Min.   :2541   Min.   :  198
##  1st Qu.:2998   1st Qu.:3766   1st Qu.:3286   1st Qu.:4224   1st Qu.: 1748
##  Median :3576   Median :4615   Median :3736   Median :4314   Median : 3882
##  Mean   :3880   Mean   :4471   Mean   :4435   Mean   :4996   Mean   : 4324
##  3rd Qu.:5033   3rd Qu.:5290   3rd Qu.:5304   3rd Qu.:4855   3rd Qu.: 4350
##  Max.   :5138   Max.   :5758   Max.   :7487   Max.   :9048   Max.   :11443
##       2019            2020           2021           2022           2023
##  Min.   :  410   Min.   :   16   Min.   :   7   Min.   : 107   Min.   :  27
##  1st Qu.: 2051   1st Qu.: 1859   1st Qu.:1233   1st Qu.: 614   1st Qu.: 541
##  Median : 5092   Median : 3638   Median :3619   Median :2240   Median :2900
##  Mean   : 4922   Mean   : 3944   Mean   :3620   Mean   :1772   Mean   :1998
##  3rd Qu.: 5155   3rd Qu.: 4000   3rd Qu.:4243   3rd Qu.:2566   3rd Qu.:3006
##  Max.   :11900   Max.   :10208   Max.   :8998   Max.   :3334   Max.   :3518
##       2024          Average
##  Min.   : 2022   Min.   :2314
##  1st Qu.: 5052   1st Qu.:2757
##  Median : 5897   Median :4064
##  Mean   : 7636   Mean   :4216
##  3rd Qu.: 8413   3rd Qu.:5066
##  Max.   :16798   Max.   :6882
```

```r
# Melt the data to long format for easier plotting
melted_data <- melt(top_5_data, id.vars = "State", variable.name = "Year", value.name = "Cases")

# Convert Year column to numeric
melted_data$Year <- as.numeric(melted_data$Year)
```

```r
# Plot the bar plot using ggplot2
ggplot(melted_data, aes(x = Year, y = Cases, fill = State)) +
  geom_bar(stat = "identity") +
  labs(title = "Top 5 States with the Highest Average Number of Cases",
       x = "Year", y = "Average Cases", fill = "State") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set3") +
  facet_wrap(~ State, scales = "free_y") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
# Plot the line plot
plot(melted_data$Year, melted_data$Cases, type = "l", col = "skyblue",
     main = "Trend of Cases Over Years for Each State",
     xlab = "Year", ylab = "Cases")
# Add lines for each state
# Add lines for each state
for (state in unique(melted_data$State)) {
  lines(melted_data$Year[melted_data$State == state], melted_data$Cases[melted_data$State == state], ty
}

# Add legend
legend("topright", legend = unique(melted_data$State), col = rep("gray", length(unique(melted_data$Stat
```




```
# Descriptive Statistics
# Calculate mean, median, and standard deviation of Lyme disease cases across all states
mean_cases <- mean(unlist(data[, -1]))
median_cases <- median(unlist(data[, -1]))
sd_cases <- sd(unlist(data[, -1]))

cat("Descriptive Statistics for Lyme Disease Cases:\n")
```

```
## Descriptive Statistics for Lyme Disease Cases:
```

```
cat("Mean:", mean_cases, "\n")
```

```
## Mean: 2698.72
```

```
cat("Median:", median_cases, "\n")
```

```
## Median: 1719
```

```
cat("Standard Deviation:", sd_cases, "\n")
```

```
## Standard Deviation: 2514.479
```

```
# Frequentist Analyses
# Check for normality using Shapiro-Wilk test for Lyme disease cases
shapiro_test <- shapiro.test(unlist(data[, -1]))
cat("\nShapiro-Wilk Test for Normality:\n")
```

```
##
## Shapiro-Wilk Test for Normality:
```

```
print(shapiro_test)
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  unlist(data[, -1])
## W = 0.79839, p-value = 1.405e-13
```

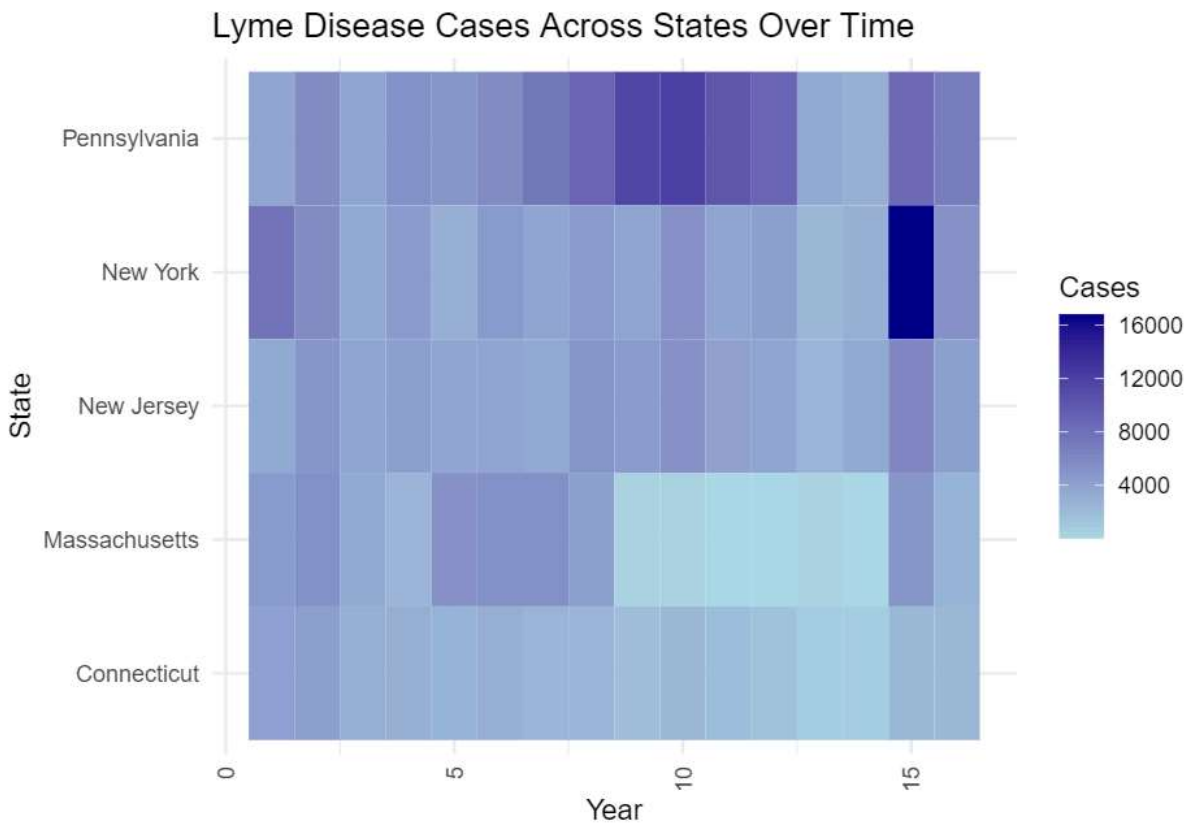
```
# As the data may not be normally distributed, perform non-parametric analysis
# Kruskal-Wallis test to compare Lyme disease cases across different years
kruskal_test <- kruskal.test(as.vector(t(data[, -1])) ~ rep(names(data)[-1], each = nrow(data)))
cat("\nKruskal-Wallis Test for Comparing Lyme Disease Cases Across Years:\n")
```

```
##
## Kruskal-Wallis Test for Comparing Lyme Disease Cases Across Years:
```

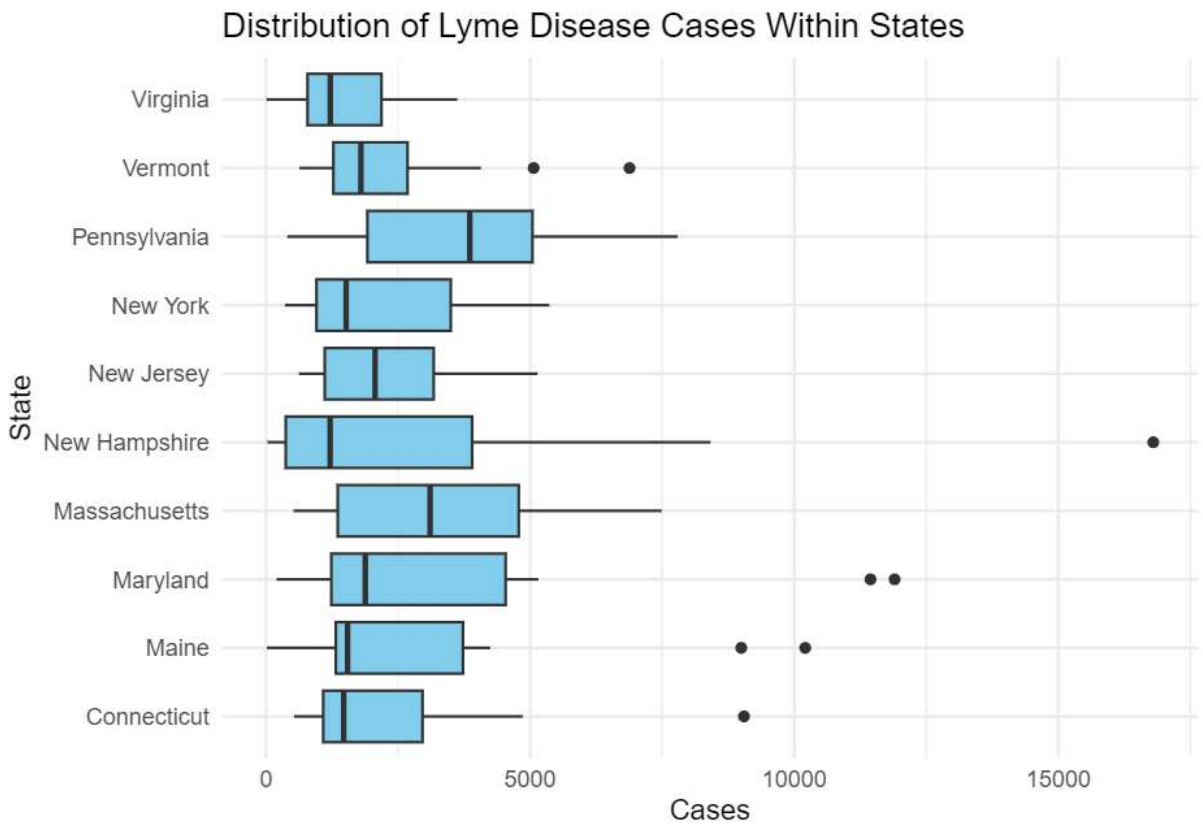
```
print(kruskal_test)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  as.vector(t(data[, -1])) by rep(names(data)[-1], each = nrow(data))
## Kruskal-Wallis chi-squared = 110.35, df = 15, p-value < 2.2e-16
```

```
# Heatmap showing distribution of Lyme disease cases across states over time
heatmap_data <- acast(melted_data, State ~ Year, value.var = "Cases")
heatmap_plot <- ggplot(data = melted_data, aes(x = Year, y = State, fill = Cases)) +
  geom_tile() +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(title = "Lyme Disease Cases Across States Over Time",
       x = "Year", y = "State", fill = "Cases") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
print(heatmap_plot)
```




```
# Boxplot for distribution of Lyme disease cases within states
boxplot_data <- data.frame(State = rep(data$State, each = ncol(data) - 1),
                           Cases = c(unlist(data[, -1])))
boxplot_plot <- ggplot(boxplot_data, aes(x = State, y = Cases)) +
  geom_boxplot(fill = "skyblue") +
  coord_flip() +
  labs(title = "Distribution of Lyme Disease Cases Within States",
       x = "State", y = "Cases") +
  theme_minimal()
print(boxplot_plot)
```


Distribution of Lyme Disease Cases Within States
Virginia
Vermont
Pennsylvania
New York
New Jersey
New Hampshire
Massachusetts
Maryland
Maine
Connecticut
State
0
5000
10000
15000
Cases