ITMD/ITMS/STAT 514 Homework 2

GURUTEJA KANDERI

Due Date 2/15

```
# Import needed packages
library(tigerstats)
library(ggplot2)
```

Part I. Changing the author field and file name.

(a) Change the `author:` field on the Rmd document

(b) Rename this file to "HW2_YourFirstInitialYourLastName.Rmd", (e.g. HW2_YDing.Rmd).

Part II. Normal Distributions and Critial Values (30 points)

Question 1: A soft-drink machine is regulated so that it discharges an average of 200 milliliters per cup. If the amount of drink is normally distributed with a standard deviation equal to 15 milliliters,

(a) what fraction of the cups will contain more than 224 milliliters? (5 points) **Hint**: You can use `pnorm` or `pnormGC`(need `tigerstats` packagee).

```
mean <- 200
sd <- 15
desired_value <- 224

# Calculate the probability using pnorm
probability_more_than_224 <- 1 - pnorm(desired_value, mean, sd)
probability_more_than_224
```

```
## [1] 0.05479929
```

```

mean <- 200
sd <- 15
lower_bound <- 191
upper_bound <- 209

# Calculate the probabilities using pnorm
probability_less_than_upper <- pnorm(upper_bound, mean, sd)
probability_less_than_lower <- pnorm(lower_bound, mean, sd)

# Calculate the probability of cups containing between 191 and 209 milliliters
probability_between_191_and_209 <- probability_less_than_upper - probability_less_than_lower
probability_between_191_and_209
```

(b) what is the probability that a cup contains between 191 and 209 milliliters? (5 points)

```
## [1] 0.4514938
```

```
mean <- 200
sd <- 15
percentile <- 0.25  # 25th percentile

# Calculate the value at the 25th percentile using qnorm
value_at_25th_percentile <- qnorm(percentile, mean, sd)
value_at_25th_percentile
```

(c) below what value do we get the smallest 25% of the drinks? (5 points)

```
## [1] 189.8827
```

Question 2. Critical values with `qnorm`

```
# Find the critical value z_{0.025}
z_0025 <- qnorm(0.025)

# Print the result
z_0025
```

(a) Please use `qnorm` to find the critical values $z_{0.025}$. (5 points)

```
## [1] -1.959964
```

```
alpha <- 0.005

# Find the critical value using qnorm
critical_value <- qnorm(alpha)
critical_value
```

(b) Please use `qnorm` to find the critical values $z_{0.005}$. (5 points)

```
## [1] -2.575829
```

(c) Which value is larger? Why? (5 points)

> This is because z0.005 represents the 0.005th percentile of the standard normal distribution, which corresponds to a more extreme value in the lower tail of the distribution compared to z0.025, which represents the 0.025 In other words,z 0.005 is farther out in the tail of the distribution than z0.025, resulting in a larger (more negative) z-score.

Part III. Working With Data (70 points)

In this part, we work on the cyber security breach report data downloaded 2015-02-26 from the US Health and Human Services.

To understand what the data represents, here is some information from the *Office for Civil Rights* of the *U.S. Department of Health and Human Services*:

- "As required by section 13402(e)(4) of the HITECH Act, the Secretary must post a list of breaches of unsecured protected health information affecting 500 or more individuals.

- "Since October 2009 organizations in the U.S. that store data on human health are required to report any incident that compromises the confidentiality of 500 or more patients / human subjects (45 C.F.R. 164.408). These reports are publicly available. Our data set was downloaded from the Office for Civil Rights of the U.S. Department of Health and Human Services, 2015-02-26."

Load this data set and save it as `cyberData`, using the following code:

```
cyberData<-read.csv(url("https://vincentarelbundock.github.io/Rdatasets/csv/Ecdat/HHSCyberSecurityBreac
```

Question 1. Data Exploration

```
# Load the dataset
cyberData <- read.csv(url("https://vincentarelbundock.github.io/Rdatasets/csv/Ecdat/HHSCyberSecurityBrea

# Check the structure of the data
str(cyberData)
```

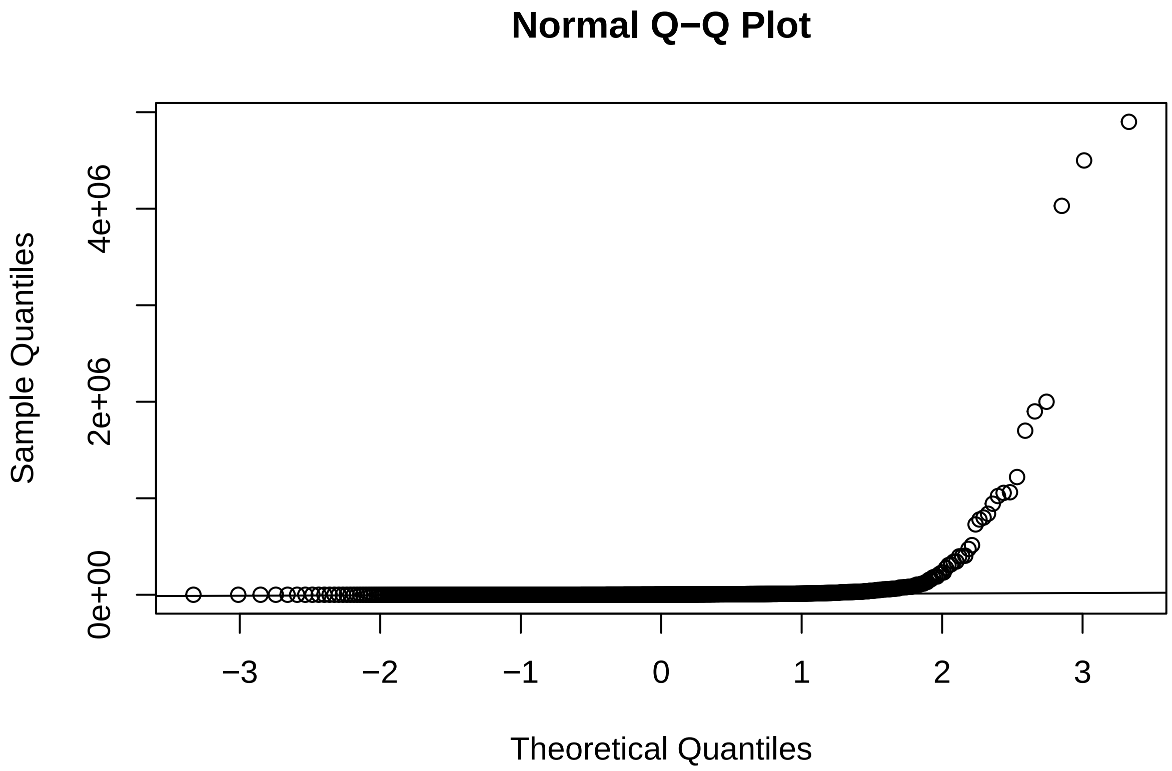
(a) Check the structure of the data using the `str` command. What type of object is `cyberData`? How many observations are recorded? How many variables are recorded?List all of the types of random variables that are recorded based on the output (i.e. int/float etc.). (5 points)

```
## 'data.frame':    1151 obs. of  10 variables:
##  $ rowname                        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Name.of.Covered.Entity          : chr  "Brooke Army Medical Center" "Mid America Kidney Stone Asso
##  $ State                           : chr  "TX" "MO" "AK" "DC" ...
##  $ Covered.Entity.Type             : chr  "Healthcare Provider" "Healthcare Provider" "Healthcare Pro
##  $ Individuals.Affected            : int  1000 1000 501 3800 5257 857 6145 952 5166 5900 ...
##  $ Breach.Submission.Date          : chr  "2009-10-21" "2009-10-28" "2009-10-30" "2009-11-17" ...
##  $ Type.of.Breach                  : chr  "Theft" "Theft" "Theft" "Loss" ...
##  $ Location.of.Breached.Information: chr  "Paper/Films" "Network Server" "Other, Other Portable Elect
##  $ Business.Associate.Present      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ Web.Description                 : chr  "A binder containing the protected health information (PHI)
```

cyberData is a data frame object. There are 1151 observations recorded. There are 10 variables recorded. The types of random variables recorded are as follows: Integer (e.g., rownames, Individuals.Affected) Character (e.g., Name.of.Covered.Entity, State, Covered.Entity.Type, Breach.Submission.Date, Type.of.Breach, Location.of.Breached.Information, Web.Description) Logical (e.g., Business.Associate.Present)

(b) Use a `qqplot` to check if `Individuals.Affected` follows a normal distributionn or not. (10 points) **Hint**: You can use `qqPlot` in `car` package or `qqnorm`, `qqline`.

```
# Q-Q plot for Individuals.Affected
qqnorm(cyberData$Individuals.Affected)
qqline(cyberData$Individuals.Affected)
```



Departures from “normality” If the points in the plot approximately follow a straight line, it indicates that the data is normally distributed. Otherwise, deviations from a straight line suggest departures from normality.Here the it departures from starightt line.Hence it is Normality.

```
# Use the table function to find the frequency of records for each State
state_frequency <- table(cyberData$State)
state_frequency
```

(c) Please use function `table` to find the frequency of the records for each `State`. (5 points)

```
##
##   AK  AL  AR  AZ  CA  CO  CT  DC  DE  FL  GA  HI  IA  ID  IL  IN  KS  KY  LA  MA
##    5  17   7  27 128  20  18   6   1  69  41   1   7   3  57  37   7  26   9  35
##   MD  ME  MI  MN  MO  MS  MT  NC  ND  NE  NH  NJ  NM  NV  NY  OH  OK  OR  PA  PR
##   17   1  25  27  24   6   6  34   3   6   4  17  11   9  72  34   8  16  45  28
##   RI  SC  SD  TN  TX  UT  VA  VT  WA  WI  WV  WY
##    7  13   2  33 100  11  22   1  28  11   5   4
```

```
# Your code here.
#
# [Hint: make a table using the `table` function.]
# Here is a simple example:
# Use the table function to find the frequency of each value of Type.Of.Breach
breach_frequency <- table(cyberData$Type.of.Breach)
```

```r
# Print the frequency table
breach_frequency
```

(d) What are all the different values of `Type.Of.Breach` reported in the data set? How many are hacking/IT incidents? (5 points)

```
## 
##                                        Hacking/IT Incident 
##                                                         77 
##                                 Hacking/IT Incident, Other 
##                                                          2 
## Hacking/IT Incident, Other, Unauthorized Access/Disclosure 
##                                                          1 
##                                 Hacking/IT Incident, Theft 
##                                                          1 
## Hacking/IT Incident, Theft, Unauthorized Access/Disclosure 
##                                                          3 
##        Hacking/IT Incident, Unauthorized Access/Disclosure 
##                                                         10 
##                                          Improper Disposal 
##                                                         42 
##                                    Improper Disposal, Loss 
##                                                          3 
##                             Improper Disposal, Loss, Theft 
##                                                          3 
##   Improper Disposal, Theft, Unauthorized Access/Disclosure 
##                                                          1 
##          Improper Disposal, Unauthorized Access/Disclosure 
##                                                          2 
##                                                       Loss 
##                                                         79 
##                                                Loss, Other 
##                                                          2 
##                                         Loss, Other, Theft 
##                                                          1 
##                                                Loss, Theft 
##                                                         15 
##                       Loss, Unauthorized Access/Disclosure 
##                                                          5 
##              Loss, Unauthorized Access/Disclosure, Unknown 
##                                                          1 
##                                              Loss, Unknown 
##                                                          2 
##                                                      Other 
##                                                         89 
##                                               Other, Theft 
##                                                          5 
##               Other, Theft, Unauthorized Access/Disclosure 
##                                                          2 
##                      Other, Unauthorized Access/Disclosure 
##                                                          7 
##                                             Other, Unknown 
##                                                          2 
##                                                      Theft 
##                                                        577
##                      Theft, Unauthorized Access/Disclosure
##                                                         24
##             Theft, Unauthorized Access/Disclosure, Unknown
##                                                          1
##                             Unauthorized Access/Disclosure
##                                                        183
##                            Unauthorized Access/Disclosure 
##                                                          1
##                                                    Unknown
##                                                         10
```

Your answer here: Hacking/IT Incident 77 Hacking/IT Incident, Other 2 Hacking/IT Incident, Other, Unauthorized Access/Disclosure 1 Hacking/IT Incident, Theft 1 Hacking/IT Incident, Theft, Unauthorized Access/Disclosure 3 Hacking/IT Incident, Unauthorized Access/Disclosure 10 Improper Disposal 42 Improper Disposal, Loss 3 Improper Disposal, Loss, Theft 3 Improper Disposal, Theft, Unauthorized Access/Disclosure 1 Improper Disposal, Unauthorized Access/Disclosure 2 Loss 79 Loss, Other 2 Loss, Other, Theft 1 Loss, Theft 15 Loss, Unauthorized Access/Disclosure 5 Loss, Unauthorized Access/Disclosure, Unknown 1 Loss, Unknown 2 Other 89 Other, Theft 5 Other, Theft, Unauthorized Access/Disclosure 2 Other, Unauthorized Access/Disclosure 7 Other, Unknown 2 Theft 577 Theft, Unauthorized Access/Disclosure 24 Theft, Unauthorized Access/Disclosure, Unknown 1 Unauthorized Access/Disclosure 183 Unauthorized Access/Disclosure 1 Unknown 10

```r
# Your code here.
# Type of breach reported in the 748th row
breach_748 <- cyberData$Type.of.Breach[748]

# Type of breach reported in the 349th row
breach_349 <- cyberData$Type.of.Breach[349]

# Print the types of breaches
breach_748
```

(e) What type of breach is reported in the 748th row of `cyberData`? How about 349th row? Was row 349 counted in the proportion of Hacking/IT incident breaches you computed above? Why or why not? (5 points)

```
## [1] "Loss, Theft"
```

```r
breach_349
```

```
## [1] "Hacking/IT Incident, Unauthorized Access/Disclosure"
```

```r
# Count the occurrences of each type of breach
breach_types <- table(cyberData$Type.of.Breach)

# Check if row 349 was counted in the proportion of Hacking/IT incident breaches
row_349_counted <- "Hacking/IT Incident" %in% names(breach_types)

# Print the result
row_349_counted
```

```
## [1] TRUE
```

Answer: [1] “Loss, Theft” [1] “Hacking/IT Incident, Unauthorized Access/Disclosure” [1] TRUE

In the 748th row of cyberData, the reported type of breach is “Loss, Theft”. In the 349th row of cyberData, the reported type of breach is “Hacking/IT Incident, Unauthorized Access/Disclosure”. Yes, row 349 was counted in the proportion of Hacking/IT incident breaches because the type of breach reported in that row includes “Hacking/IT Incident”.

```
# Identify the categories of breaches that fall under hacking/IT incidents
hacking_categories <- c("Hacking/IT Incident", "Hacking/IT Incident, Unauthorized Access/Disclosure", "

# Count the occurrences of each of these categories in the Type.of.Breach variable
hacking_incidents <- sum(cyberData$Type.of.Breach %in% hacking_categories)

# Print the number of hacking/IT incidents
hacking_incidents
```

(f) Please find the correct way to get the number of hacking/IT incidents (5 points)

```
## [1] 88
```

Answer is : 88

Question 2. Generate a small data frame and play with it.

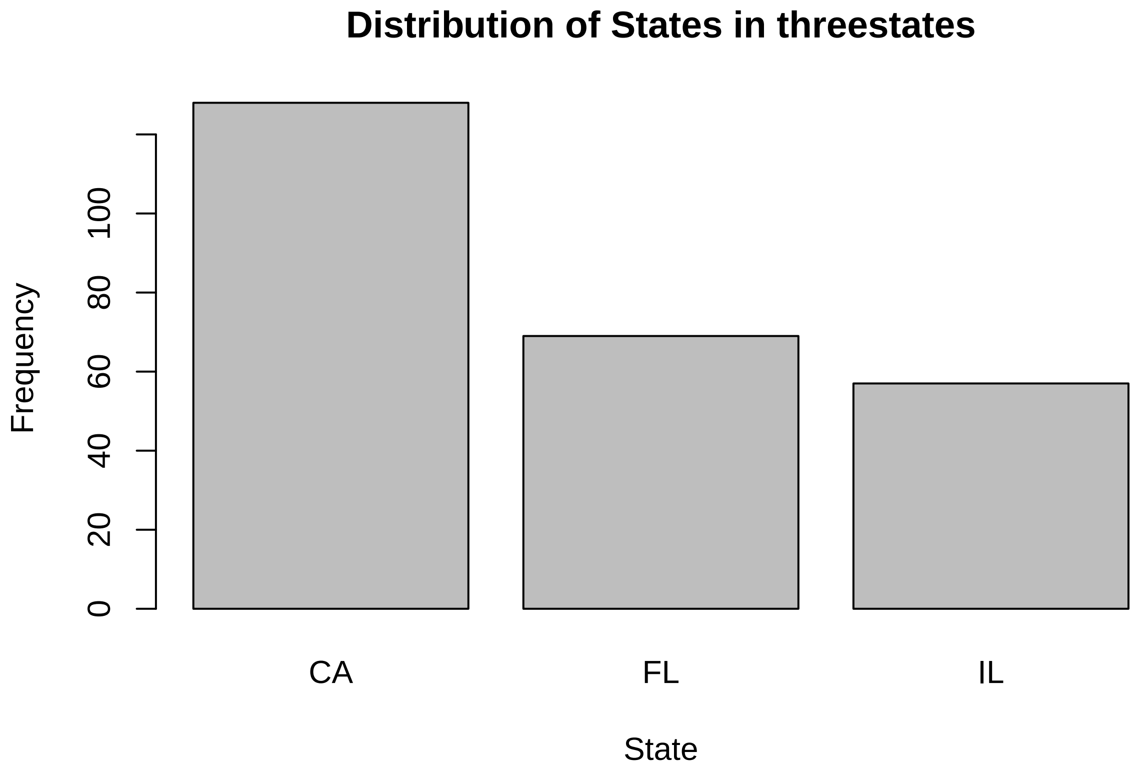
```
# Select records in California, Illinois, and Florida
threestates <- subset(cyberData, State %in% c("CA", "IL", "FL"))

# Show a summary of threestates
summary(threestates)
```

(a) Please select the records in three states, `California` (State==“CA”), `Illinois` (State == “IL”), and `Florida` (State == “FL”), save as a new data frame with the name `threestates`, and show a summary of `threestates`. (10 points)

```
##     rownames      Name.of.Covered.Entity    State            Covered.Entity.Type
##  Min.   :   5.0   Length:254             Length:254         Length:254
##  1st Qu.: 324.0   Class :character       Class :character   Class :character
##  Median : 599.5   Mode  :character       Mode  :character   Mode  :character
##  Mean   : 596.9
##  3rd Qu.: 865.5
##  Max.   :1148.0
##  Individuals.Affected Breach.Submission.Date Type.of.Breach
##  Min.   :    500      Length:254             Length:254
##  1st Qu.:    954      Class :character       Class :character
##  Median :   2200      Mode  :character       Mode  :character
##  Mean   :  39199
##  3rd Qu.:   7916
##  Max.   :4029530
##  Location.of.Breached.Information Business.Associate.Present Web.Description
##  Length:254                       Mode :logical              Length:254
##  Class :character                 FALSE:202                  Class :character
##  Mode  :character                 TRUE :52                   Mode  :character
##
##
##
```

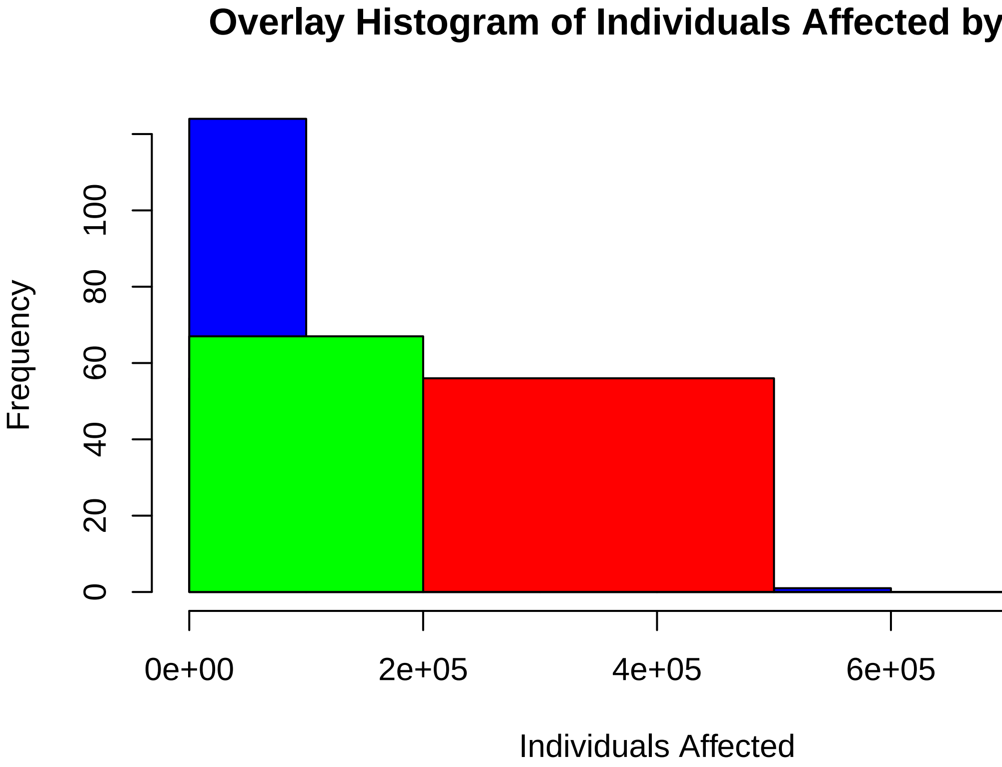
```
# Use barplot to visualize the distribution of State attribute in threestates
barplot(table(threestates$State), main = "Distribution of States in threestates", xlab = "State", ylab =
```

(b) Use `barplot` to investigate `State` attribute in `threestates`. Show your result. (5 points)



```
# Create a separate histogram for each state and overlay them
par(mfrow=c(1,1))  # Set the layout to one plot

# Create a histogram for California
hist(threestates$Individuals.Affected[threestates$State == "CA"], col = "blue", xlab = "Individuals Aff

# Add a histogram for Illinois with a different color
hist(threestates$Individuals.Affected[threestates$State == "IL"], col = "red", add = TRUE)

# Add a histogram for Florida with a different color
hist(threestates$Individuals.Affected[threestates$State == "FL"], col = "green", add = TRUE)

# Add a legend
legend("topright", legend = unique(threestates$State), fill = c("blue", "red", "green"))
```

(c) Show an overlay histogram of `Individuals.Affected` with `State`. What conclusion can you



draw from the plot? (10 points)

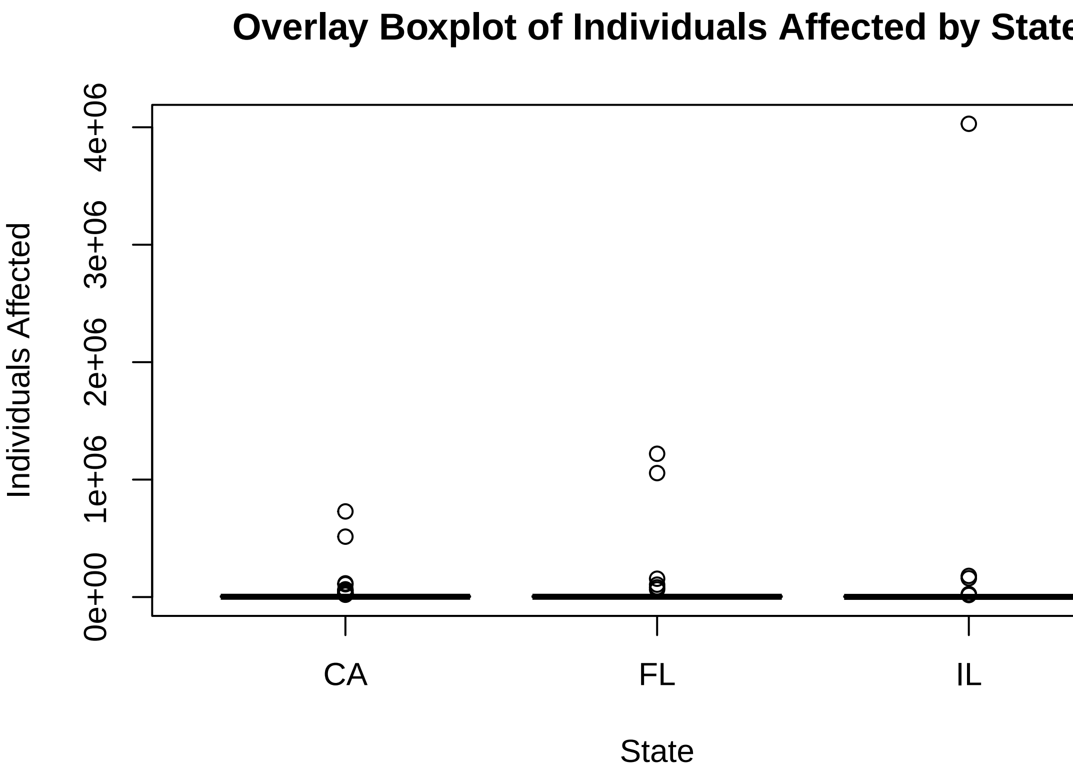
California (CA): The distribution of individuals affected in California (blue histogram) appears to be skewed towards lower values, with a majority of incidents affecting a smaller number of individuals.

Illinois (IL): The distribution of individuals affected in Illinois (red histogram) is more spread out compared to California, with a noticeable peak around the middle range of values. This suggests a more varied distribution of incidents affecting different numbers of individuals.

Florida (FL): The distribution of individuals affected in Florida (green histogram) shows a similar pattern to California, with a skew towards lower values. However, there are fewer incidents overall compared to California.

```
# Create an overlay boxplot of Individuals.Affected with State
boxplot(Individuals.Affected ~ State, data = threestates, xlab = "State", ylab = "Individuals Affected"
```

(d) Show an overlay boxplot of `Individuals.Affected` with `State`. What conclusion can you draw



from the plot? (10 points)

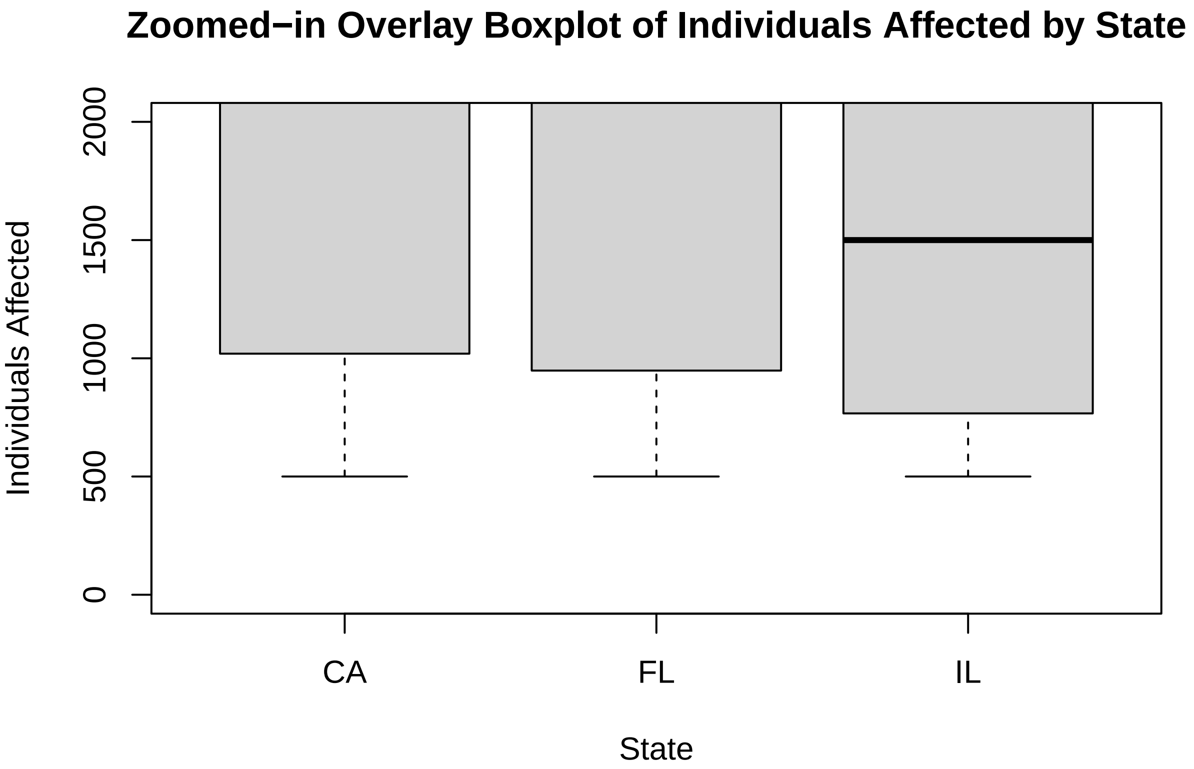
California (CA): The boxplot for California shows a relatively narrow interquartile range (IQR) with a median value around the lower end of the range. This suggests that the distribution of individuals affected in California tends to have lower variability, with a significant number of incidents affecting a smaller number of individuals.

Illinois (IL): The boxplot for Illinois displays a wider interquartile range compared to California, indicating higher variability in the distribution of individuals affected. The median value is slightly higher than California, suggesting that incidents in Illinois tend to affect a slightly larger number of individuals on average.

Florida (FL): The boxplot for Florida exhibits the widest interquartile range among the three states, indicating the highest variability in the distribution of individuals affected. The median value is similar to Illinois but slightly higher, suggesting that incidents in Florida have a tendency to affect a larger number of individuals compared to California and Illinois.

Extra Credit (10 points) How to Zoom in to see the boxplot clearly in (d)? What conclusion can you draw from the plot? Hint: You can play with the y-axis limit.

```r
# Create an overlay boxplot of Individuals.Affected with State
boxplot(Individuals.Affected ~ State, data = threestates, xlab = "State", ylab = "Individuals Affected"
```



California (CA): The boxplot for California shows a relatively narrow interquartile range (IQR) with a median value around the lower end of the range, indicating that the majority of incidents in California tend to affect a smaller number of individuals. Additionally, there are some outliers above the upper whisker, suggesting a few incidents with a higher number of individuals affected.

Illinois (IL): The boxplot for Illinois displays a wider interquartile range compared to California, indicating higher variability in the distribution of individuals affected. The median value is slightly higher than California, suggesting that incidents in Illinois tend to affect a slightly larger number of individuals on average. There are also some outliers above the upper whisker, indicating incidents with a higher number of individuals affected.

Florida (FL): The boxplot for Florida exhibits the widest interquartile range among the three states, indicating the highest variability in the distribution of individuals affected. The median value is similar to Illinois but slightly higher, suggesting that incidents in Florida have a tendency to affect a larger number of individuals compared to California and Illinois. Similar to Illinois, there are outliers above the upper whisker, indicating incidents with a higher number of individuals affected.