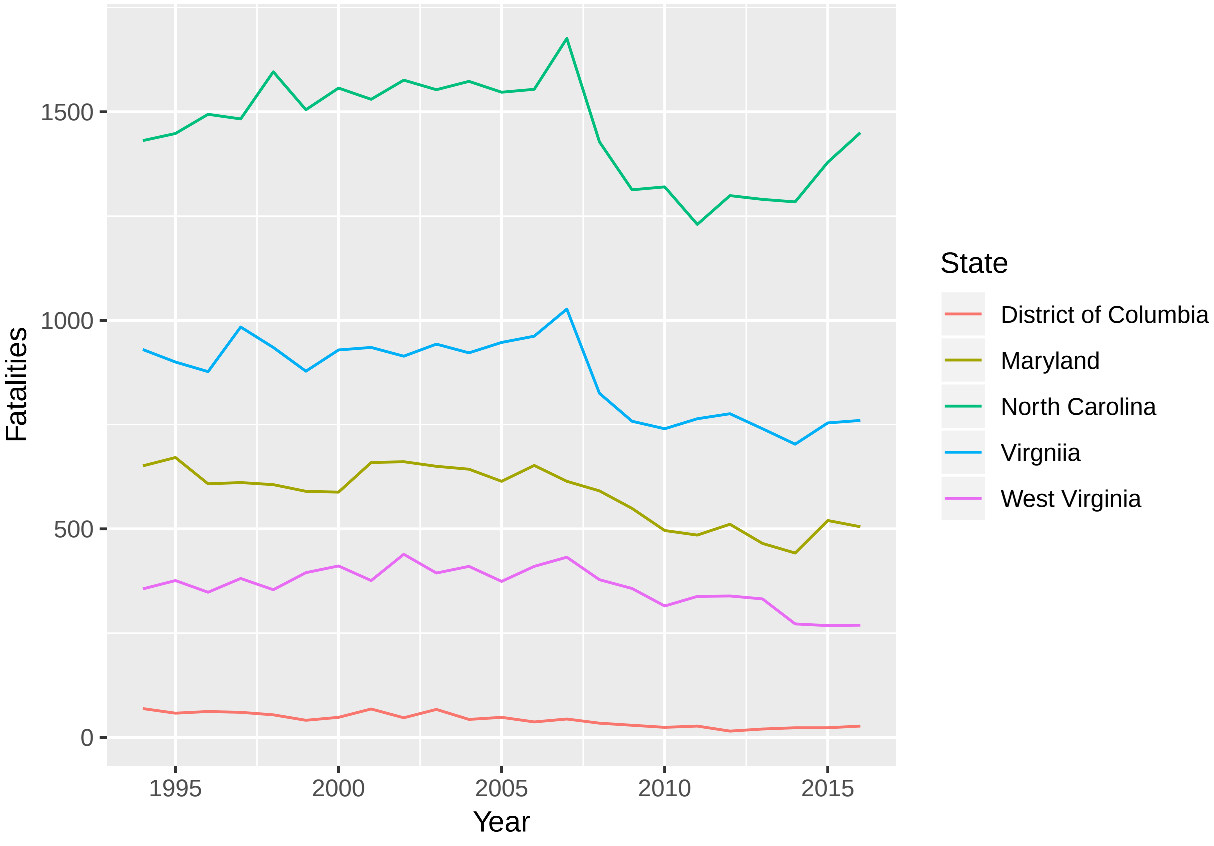
Homework 4

Max Ryoo

Problem 1

A

```
library(ggplot2)
setwd("/Users/maxryoo/Documents/Fall 2018/STAT3080/HW4/")
fatalities <- read.csv("fatalities.csv")
lsmooth <- ggplot(fatalities, aes(x=Year, y=Fatalities, color=State)) + geom_line()
print(lsmooth)
```



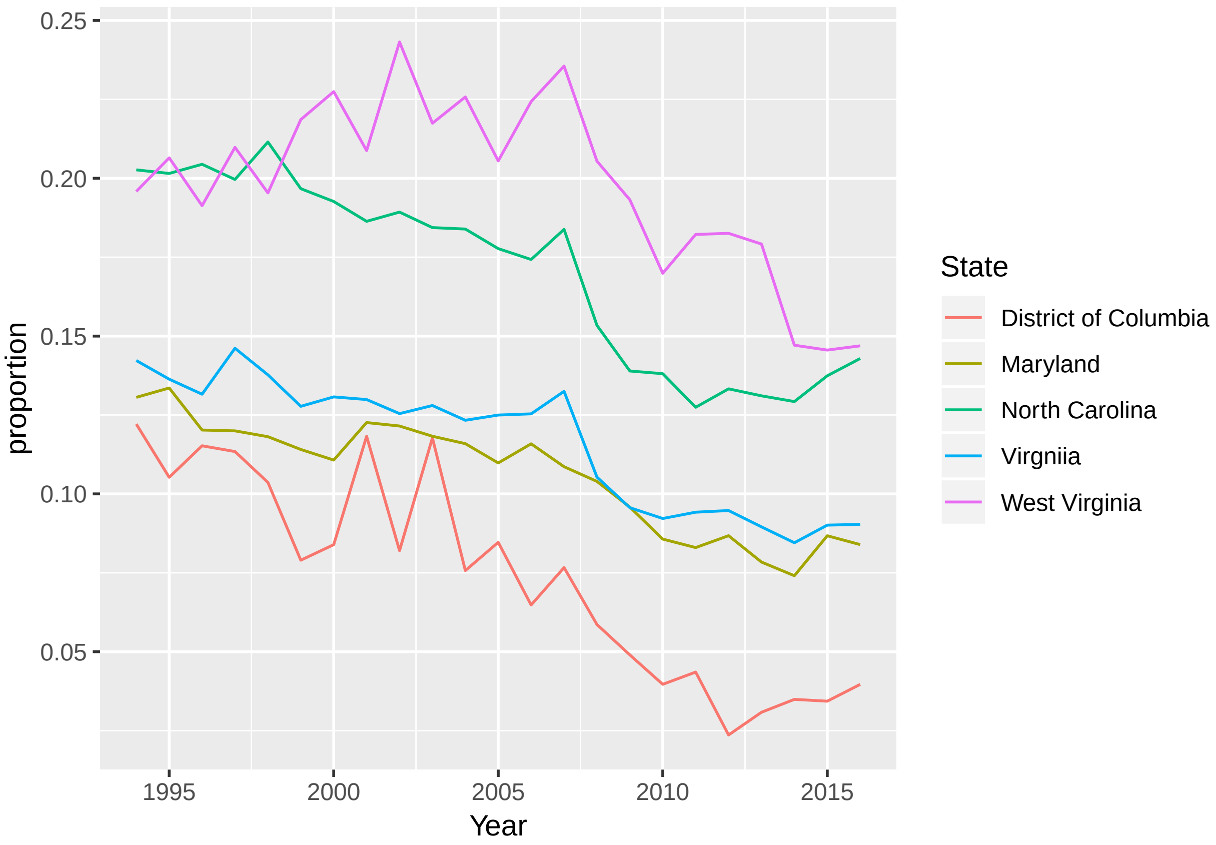
I

first read in teh csv file and set x as year y as fatalities and made a color distinction by state. I then made a smooth line for all the colors and plotted them in the plane. I then placed it in lsmooth.

B

```
fatalitiesp <- fatalities
```

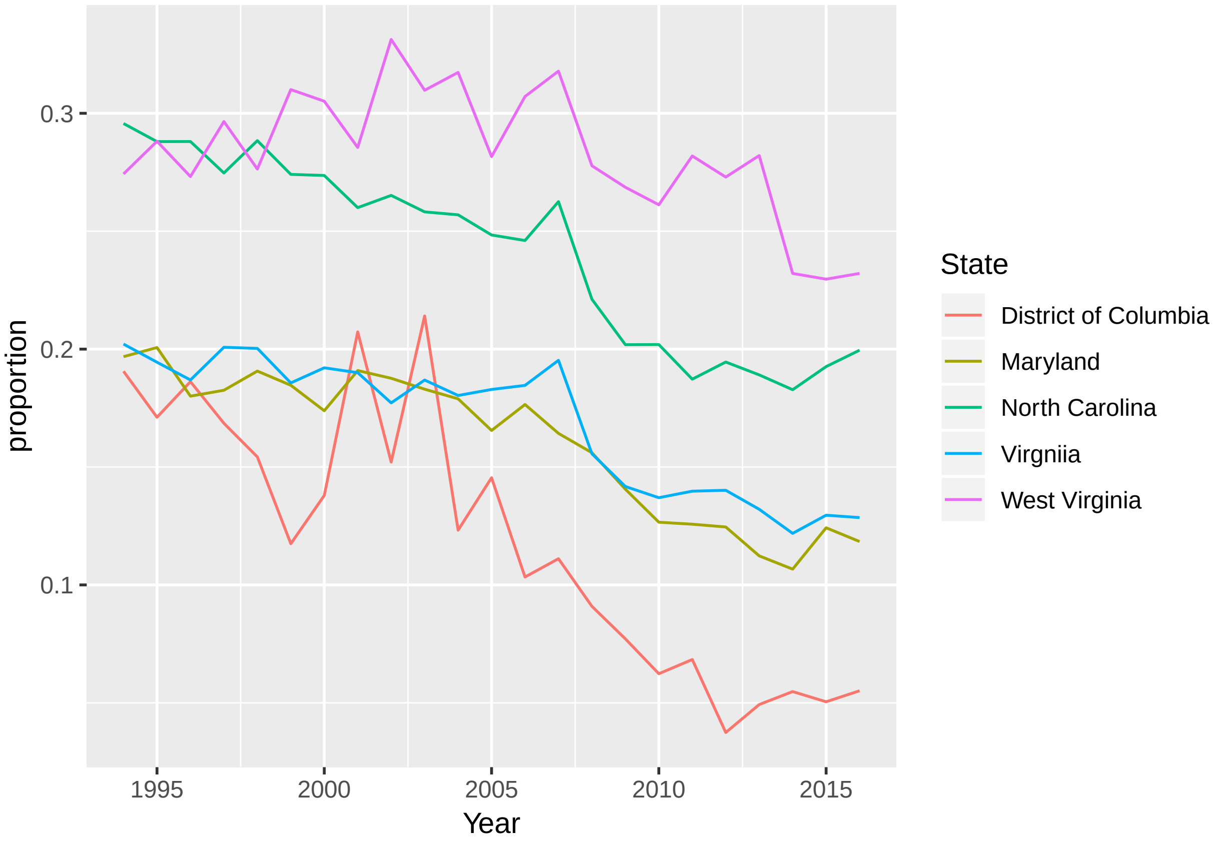
```
fatalitiesp$proportion <-
  fatalities$Fatalities / fatalities$Resident.Population
lsmoothp <- ggplot(fatalitiesp, aes(x=Year, y=proportion, color=State)) + geom_line()
print(lsmoothp)
```



I made a new dataframe called fatalitiesp where there was a new column of proportions made by dividing fatalities by population. I then set that dataframe as the data to graph. I set y as proportion and made a line.

C

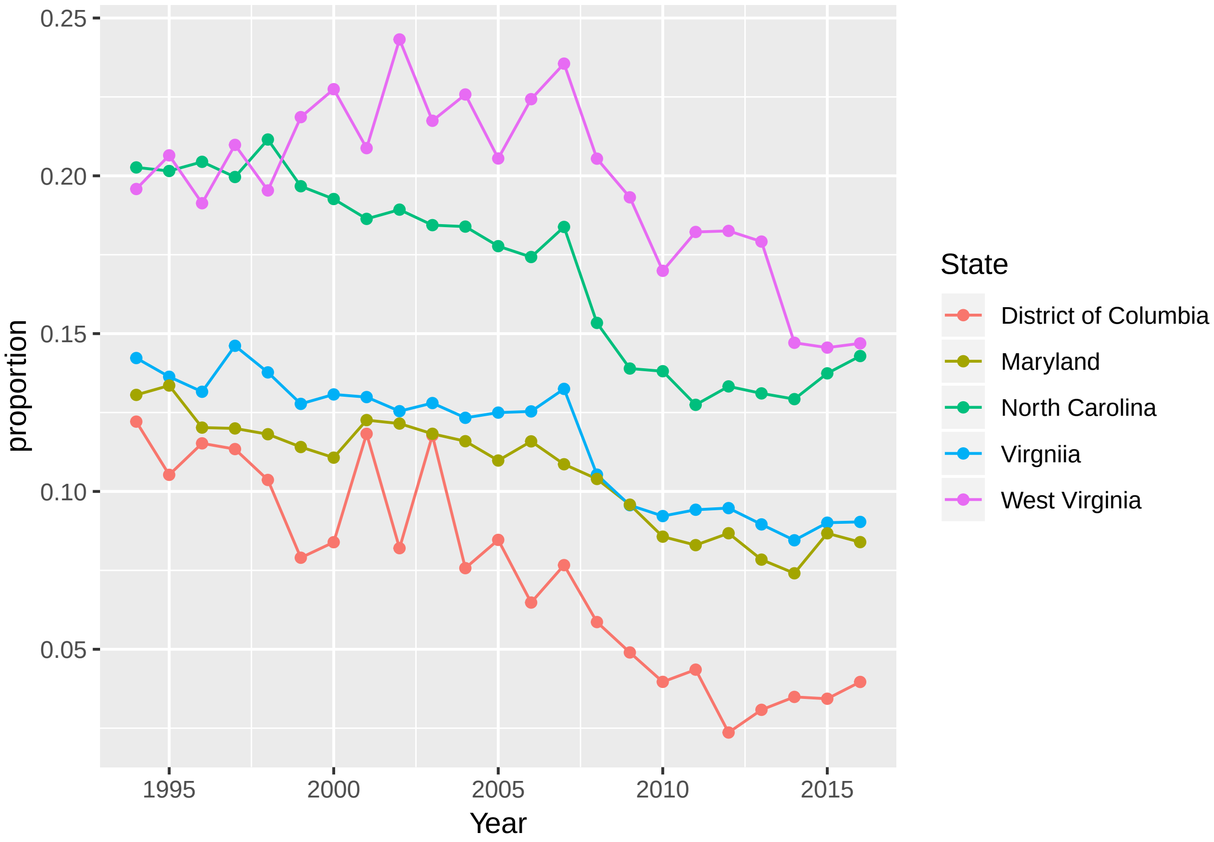
```
fatalitieslp <- fatalities
fatalitieslp$proportion <-
  fatalities$Fatalities / fatalities$Licensed.Drivers
lsmoothlp <- ggplot(fatalitieslp, aes(x=Year, y=proportion, color=State)) + geom_line()
print(lsmoothlp)
```



I made a new dataframe called fatalitieslp where there was a new column of proprtions made by dividing the fatalities by licensed drivers. I then set that dataframe as the data to grpah. I set y as proportion and made a line connecting the points.

D

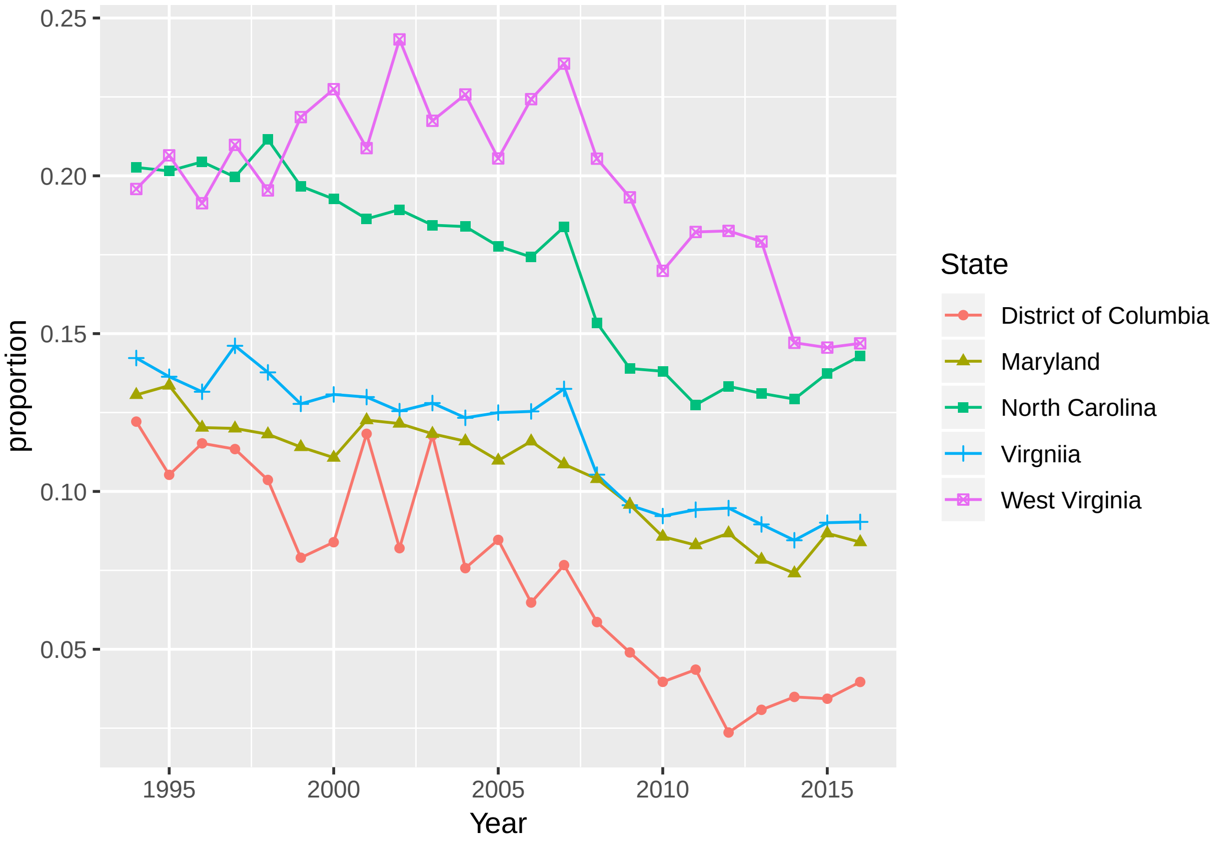
```
pointslsmoothp1 <- lsmoothp + geom_point()
print(pointslsmoothp1)
```



I added points to the graph made it part b. The points are displayed as black to made it easy to read

E

```
pointslsmoothp2 <- lsmoothp + geom_point(aes(shape = State))
print(pointslsmoothp2)
```

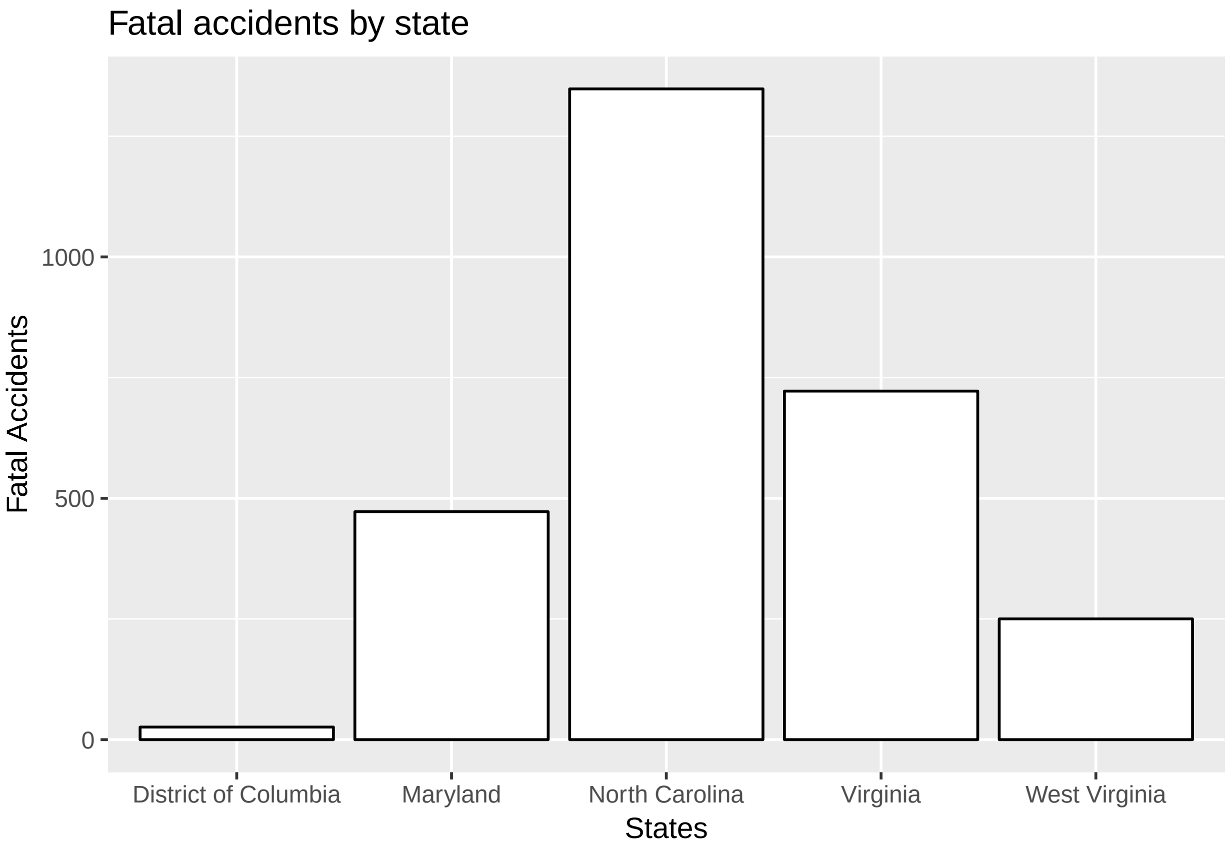


I added a parameter for different shapes in the aes of geom_point. Each state has different points for representation.

2

A

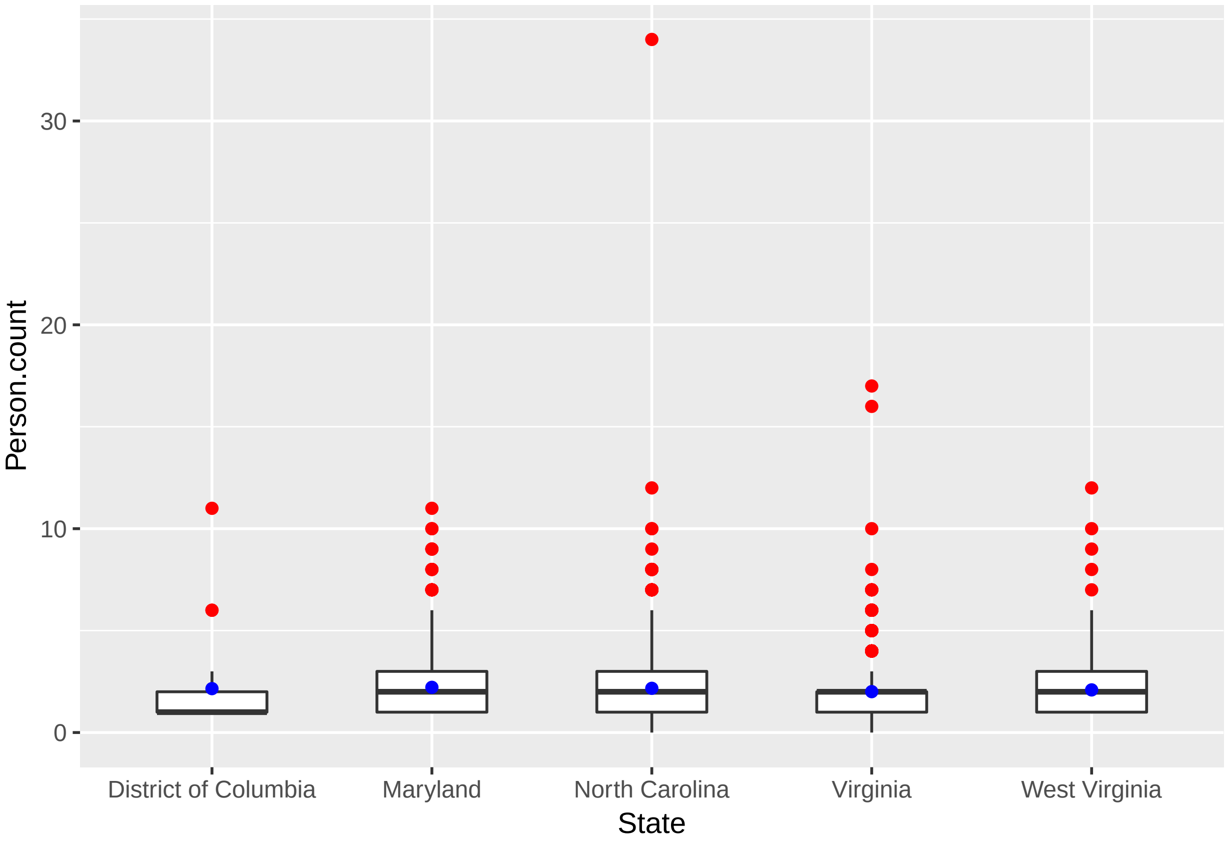
```
fatalacc <- read.csv("fatal accidents.csv")
bystate <- table(fatalacc$State)
fatalstate <- as.data.frame(bystate)
barbystate <- ggplot(fatalstate, aes(x=Var1, y=Freq)) +
  geom_bar(stat="identity", fill="white", colour="black")
print(barbystate + labs(title="Fatal accidents by state",
                    x="States", y="Fatal Accidents"))
```



I first read in the csv containing the data. I had to count the number of entires for each state which I did by using the table function for the state column in the original read. I then made the frequency table into a dataframe, which i made a box plot and gave respective names.

B

```
bxperson <- ggplot(fatalacc, aes(x=State, y=Person.count))
mean_out_bx <- bxperson +
  geom_boxplot(width=0.5, outlier.size=2, outlier.shape=16,
               outlier.colour = "red") +
  stat_summary(fun.y="mean", geom="point",
               shape=16, size=2, color="blue")
print(mean_out_bx)
```



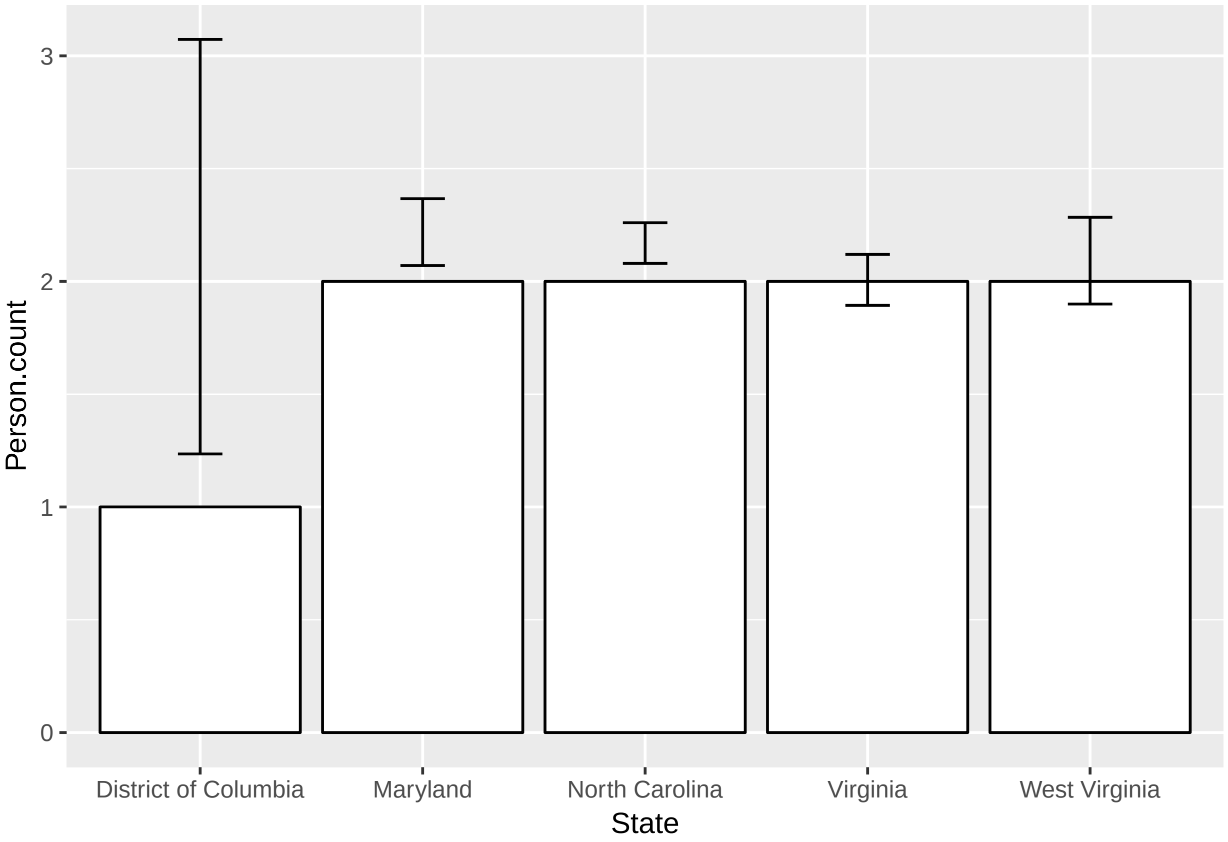
I first made a ggplot of x and y being state and person.count respectivly I then added the boxplot with the geom_boxplot with a width of 0.5. I added outliers and mean for which I made a clear color distinction. I used teh stat_summary for the mean and I used geom_boxplot to show outliers of the box plot.

C

The reason that the District of Columbia is more influenced by outliers is the the fact that the sample size is a lot smaller than north carolina as shown by the bar graph in part A question 2. Becuase there are less point, when calculating the mean the outliers in D.C will hold more weight.

D

```
barmedian <- ggplot(fatalacc,
                    aes(x=State, y=Person.count)) +
  stat_summary(fun.y=median, geom="bar",
               position="dodge",fill="white", colour="black")
barmedianer <- barmedian+
  stat_summary(fun.data=mean_cl_normal,
               geom="errorbar", width=0.2)
print(barmedianer)
```

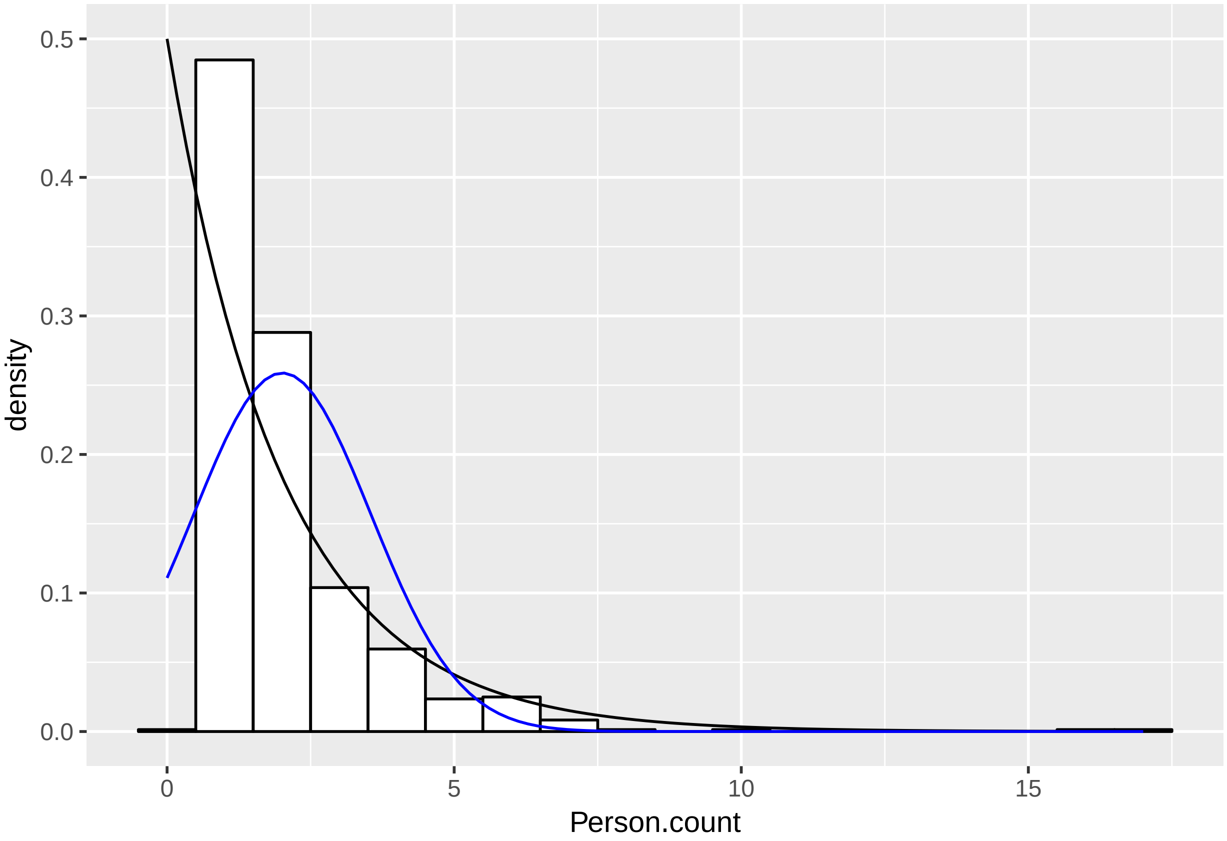


E

The plot in part B provides a better statistically summary since there are more details provided. There are outlier shown and mean of each state. We can see how influential the outlier points are for the data, while for the graph in Part D is not as easy to visualize.

F

```
fatalaccva <- fatalacc[which(
  fatalacc$State == "Virginia"),]
x=seq(0,15,5)
density <- dnorm(x,
                 mean(fatalaccva$Person.count)
                 ,sd(fatalaccva$Person.count))
histvirginia <- ggplot(fatalaccva,
                       aes(x=Person.count)) +
  geom_histogram(binwidth = 1, fill="white",
                 colour="black", aes(y=..density..)) +
  stat_function(fun=dchisq, args=list(df=2)) +
  stat_function(fun=dnorm, colour="blue",
                args=list(mean=mean(fatalaccva$Person.count),
                          sd=sd(fatalaccva$Person.count)))
print(histvirginia)
```



References

1. <https://stackoverflow.com/questions/1923273/ counting-the-number-of-elements-with-the-values-of-x-in-a-vector>