Data Visualization Suburb Report

Nate Kaduk
Professor Boyd-Swan

Part 1:
All variables have 77 observations. The summary statistics for five variables is given below.
Table 1.1. Five Variable (Plus Top 20) Summary Statistics

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
SidewalkPct	77	55	37	0	16	91	100
PropTax	77	2.4	0.57	1.4	2.1	2.7	4
Pov	77	8.9	6.8	1.6	4.2	11	42
Police	77	2.5	2.4	1.1	1.6	2.4	21
CommServ	77	0.01	0.0029	0.002	0.009	0.012	0.015
Top20	77	0.26	0.44	0	0	1	1

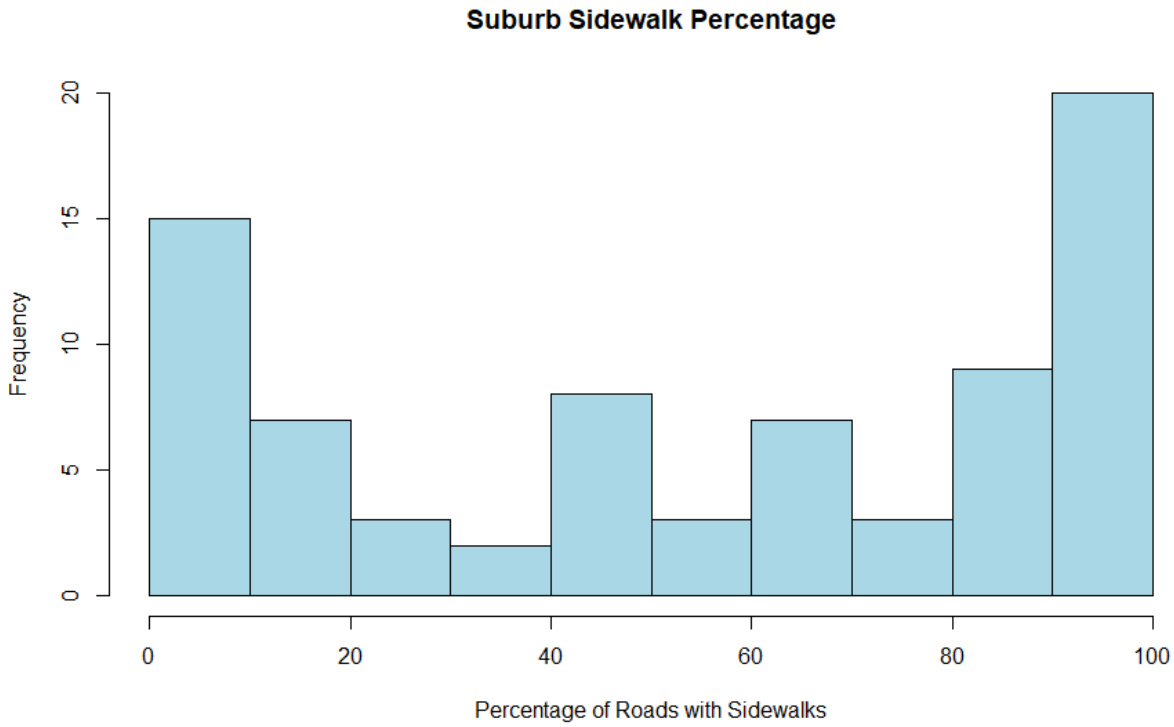
For sidewalk percentage, the mean percentage of roads with sidewalks is 55%. The standard deviation is 37%. One big culprit in the high standard deviation is the high number of suburbs at both ends of the percentage spectrum (which will be shown with the histograms). There are a number of suburbs with fewer than 10% of roads having sidewalks as well as an even larger number of suburbs with more than 90% of roads having sidewalks. Property tax has a mean of 2.4% and a standard deviation of 0.57.

The poverty rate has a mean of 8.9 (i.e. 8.9% of a suburb's population is in poverty on average), and a standard deviation of 6.8. The smallest poverty rate is 1.6. Considering that the 75th percentile only has a poverty rate of 11, it is shocking that a poverty rate of 42 is the maximum of all the suburbs.

The number of police per 1000 people is 2.5 on average. The outstanding statistic for this variable is standard deviation, which almost equals the mean. The 75th percentile of suburbs have 2.4 police per 1000 people, but the suburb with the most police has 21 police per 1000 people.

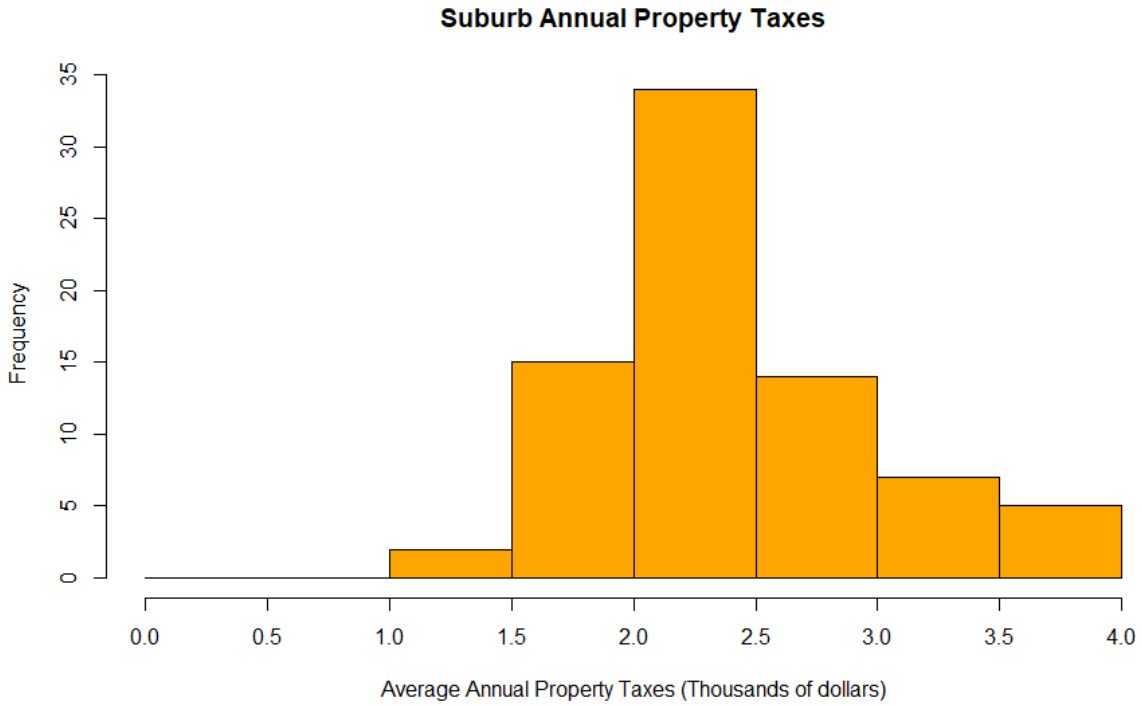
Community services per 1000 people has a mean of 0.01 and an extremely small standard deviation of 0.0029.

Figure 1.1. Suburb Sidewalk Percentage



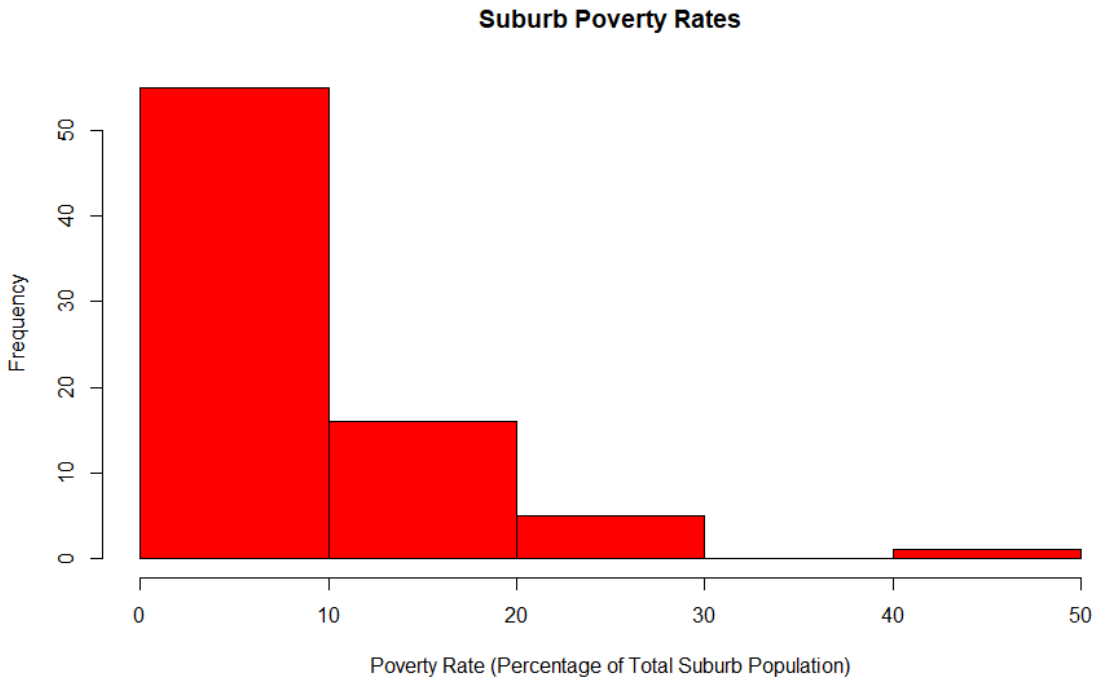
A massive chunk of suburbs have sidewalks either on most of their roads, or virtually none of their roads. This difference helps contribute to the high standard deviation specified earlier. In general, there are more suburbs with more than 80% of their roads having sidewalks than there are suburbs with less than 20% of their roads having sidewalks. Still, the number of suburbs at these extremes is very large.

Figure 1.2. Suburb Annual Property Taxes



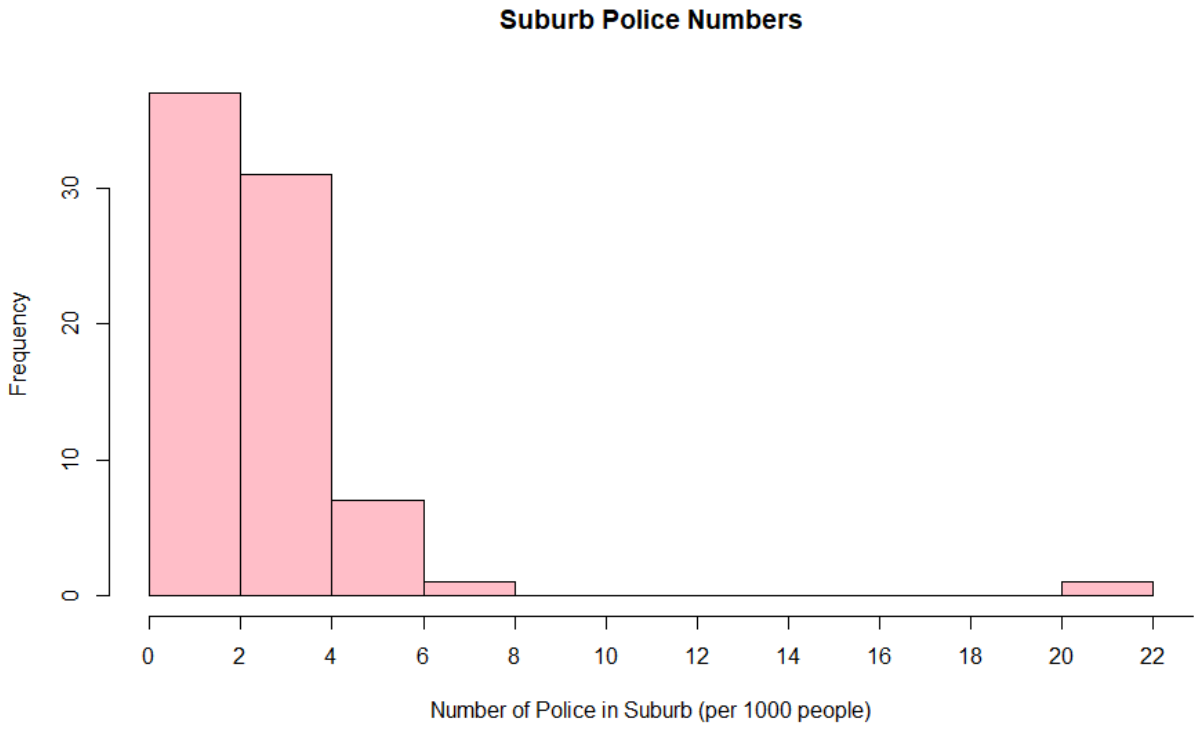
The vast majority of property taxes lies between $2,000 and $2,500. There are more suburbs with annual property taxes above this mark than those below it.

Figure 1.3. Suburb Poverty Rates



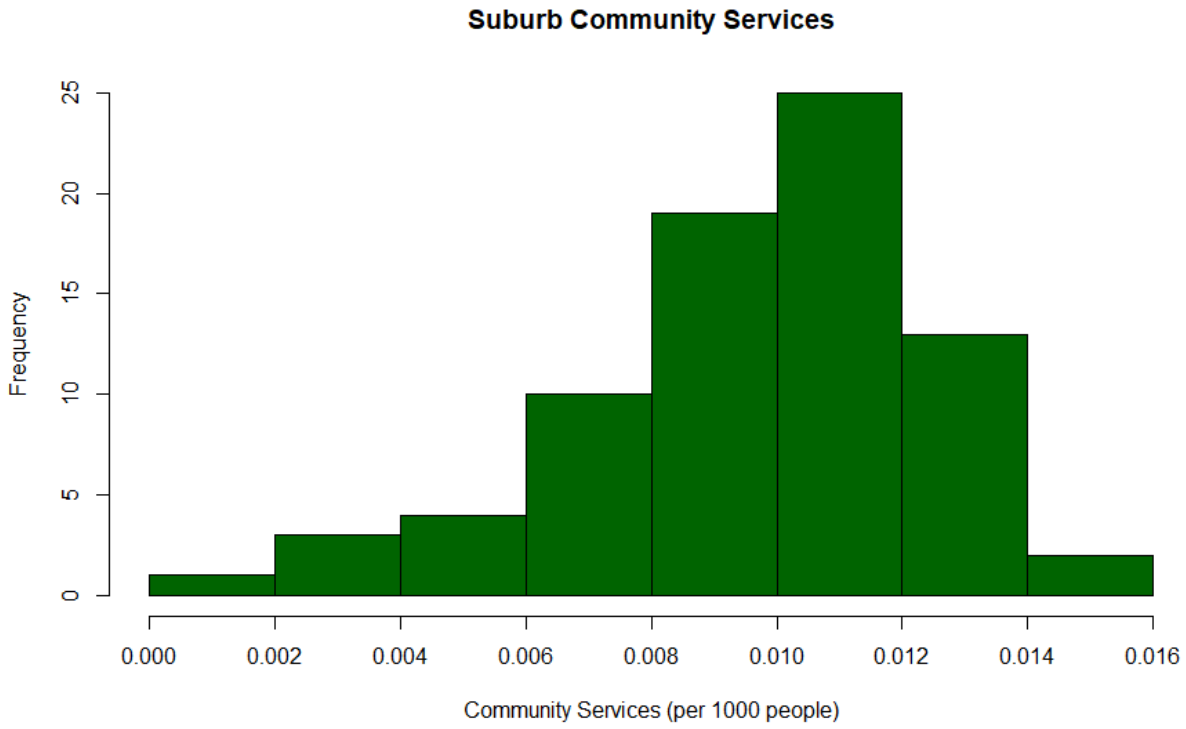
The vast majority of suburbs have a poverty rate below 20%. However, there are a few impoverished suburbs with 20-30% poverty rates. There is also a really poor suburb with a poverty rate of 42%.

Figure 1.4. Suburb Police Numbers



Most suburbs have between 0-8 police per 1000 people. There is one suburb with 21 police per 1000 people.

Figure 1.5. Suburb Community Services



In all suburbs, the amount of community services per 1000 people is small. In the suburb with the highest number of community services, there are only 0.015 services per 1000 people. By comparison, the suburb with the lowest amount of community services only has 0.002 community services per 1000 people.

Part 2:

Table 2.1. Top 20 Summary Statistics Comparison

Summary Statistics

Top20		0			1	
Variable	N	Mean	SD	N	Mean	SD
SidewalkPct	57	56	37	20	52	36
PropTax	57	2.5	0.61	20	2.2	0.33
Pov	57	11	7.3	20	4.3	1.3
Police	57	2.5	2.7	20	2.6	1.2
CommServ	57	0.0094	0.0029	20	0.012	0.0013

Between the suburbs in the top 20 and the suburbs not in the top 20, the biggest difference between the five chosen variables is in the poverty rate. The suburbs not in the top 20 have a mean poverty rate of 11% whereas those in the top 20 only have a mean poverty rate of 4.3%. The sidewalk percentage and annual property taxes are roughly the same between top 20 and

non-top 20 suburbs. The number of community services per 1000 people is small for both kinds of suburbs, but the standard deviation for non-top 20 suburbs is fairly higher. The number of police per 1000 people is the most interesting. Both top 20 and non-top 20 suburbs have roughly the same average number of police. However, the non-top 20 suburbs have a much higher standard deviation when it comes to the number of police.

Figure 2.1.a. Suburb Sidewalk Percentage (Top 20 Suburbs)

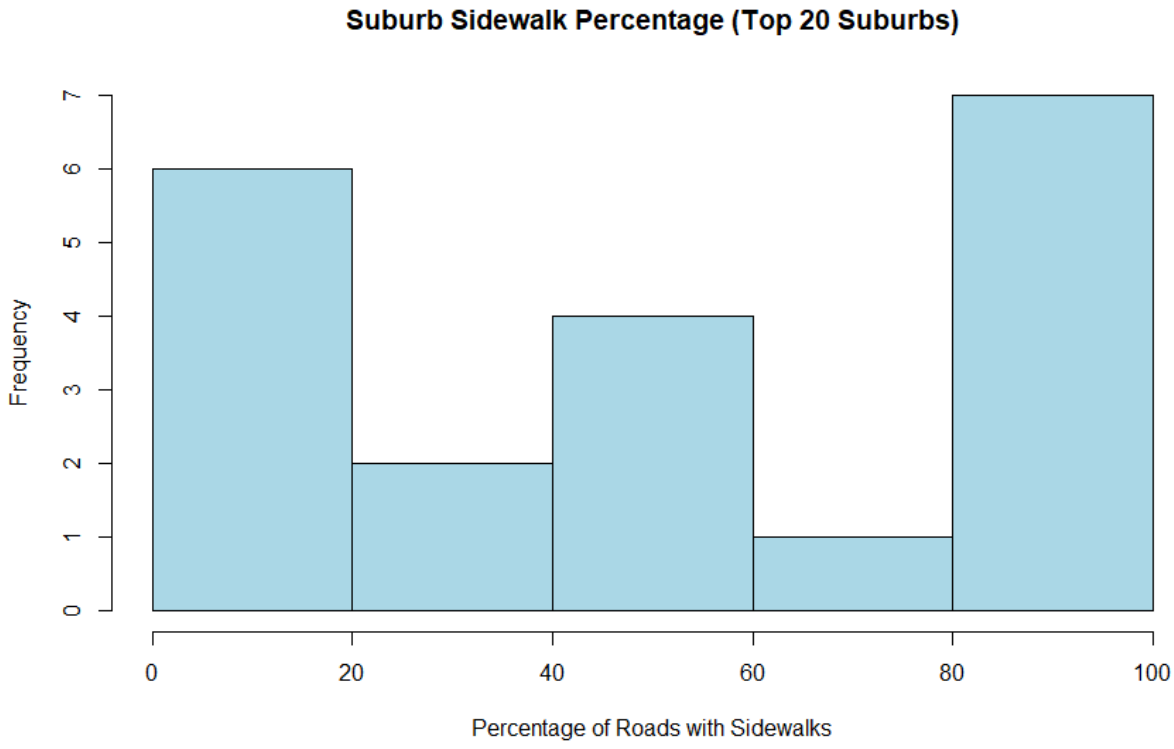
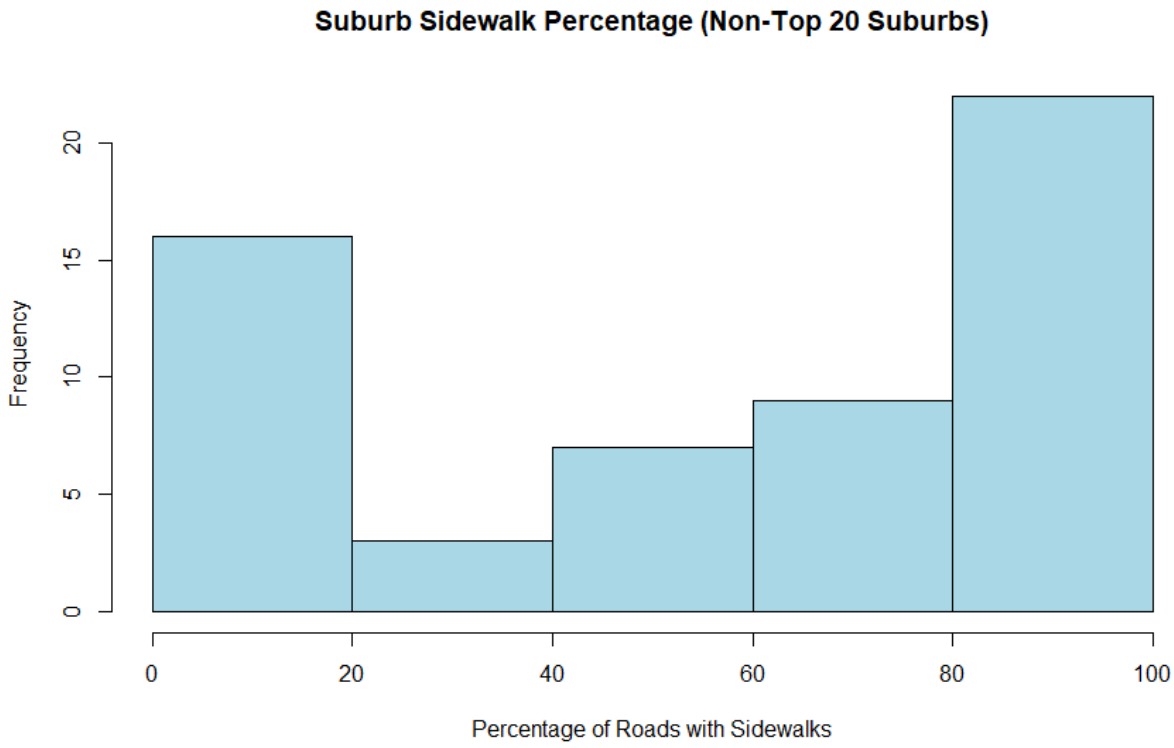


Figure 2.1.b. Suburb Sidewalk Percentage (Non-Top 20 Suburbs)



Both top 20 and non-top twenty suburbs have the vast majority of their roads either all with sidewalks or close to none with sidewalks. With this being said, a higher portion of top 20 suburbs have sidewalks with between 40-60% of their roads.

Figure 2.2.a. Suburb Annual Property Taxes (Top 20 Suburbs)

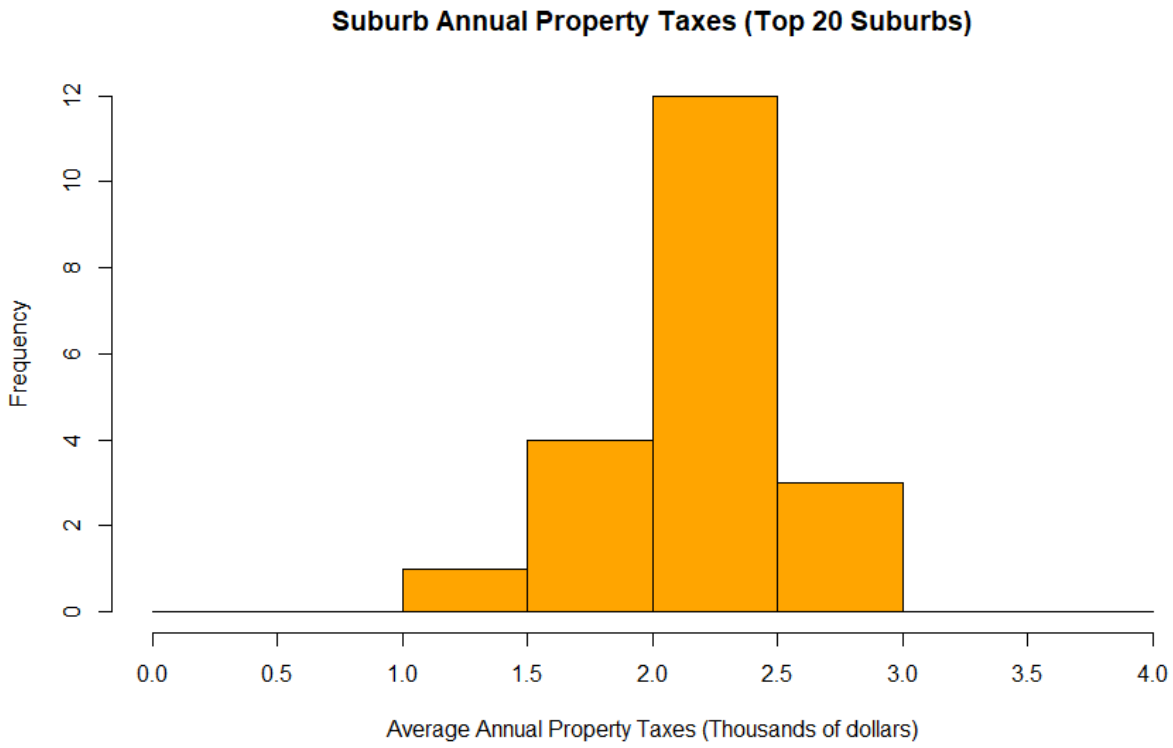
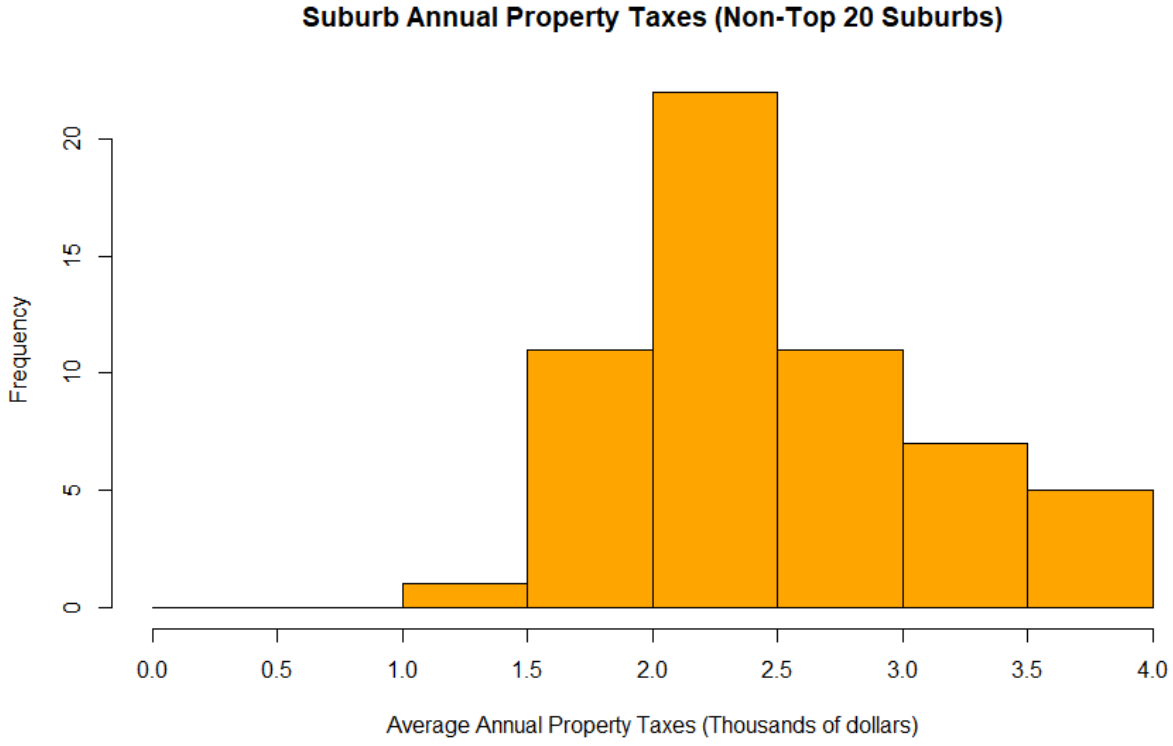


Figure 2.2.b. Suburb Annual Property Taxes (Non-Top 20 Suburbs)



The majority of suburbs have annual property taxes ranging between $2,000-$2,500. The suburbs that are not in the top 20 have suburbs with higher values than this between $3,000 and $4,000.

Figure 2.3.a. Suburb Poverty Rates (Top 20 Suburbs)

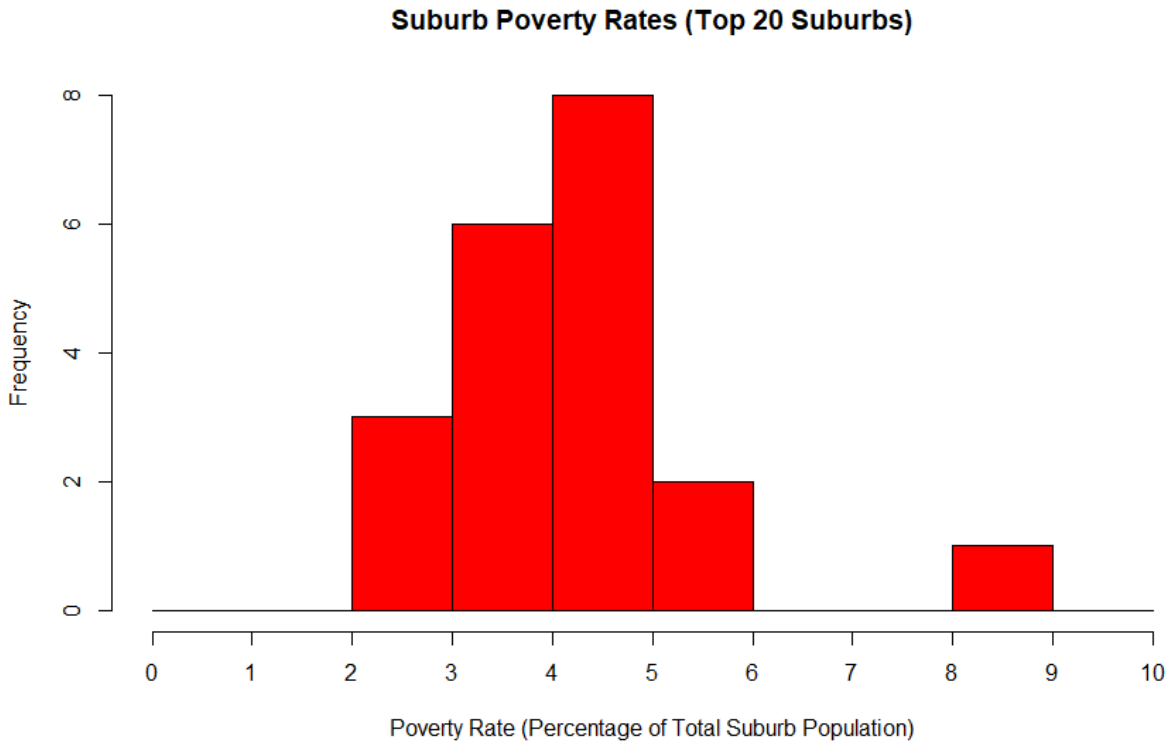
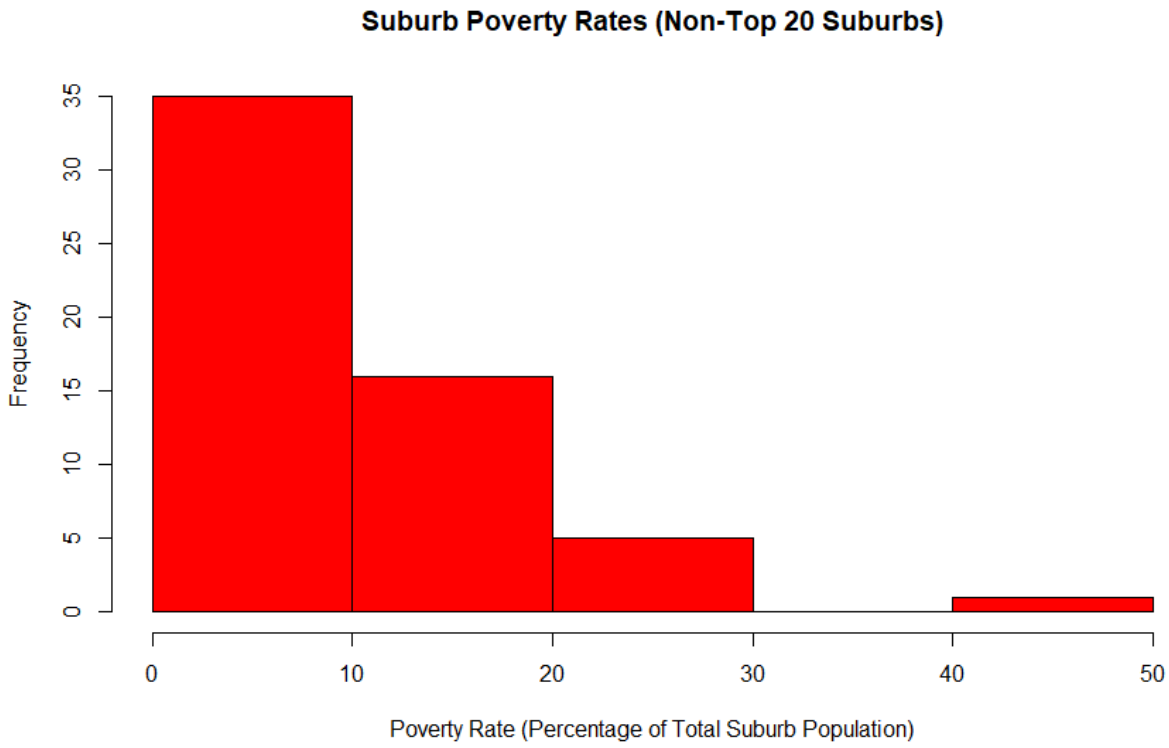


Figure 2.3.b. Suburb Poverty Rates (Non-Top 20 Suburbs)



The suburb poverty rate is much higher for non-top 20 suburbs than it is for top-20 suburbs. All observations of the poverty rate in the top 20 suburbs are below 9%. More than half of the suburbs in non-top 20 suburbs have a poverty rate over 10%.

Figure 2.4.a. Suburb Police Numbers (Top 20 Suburbs)

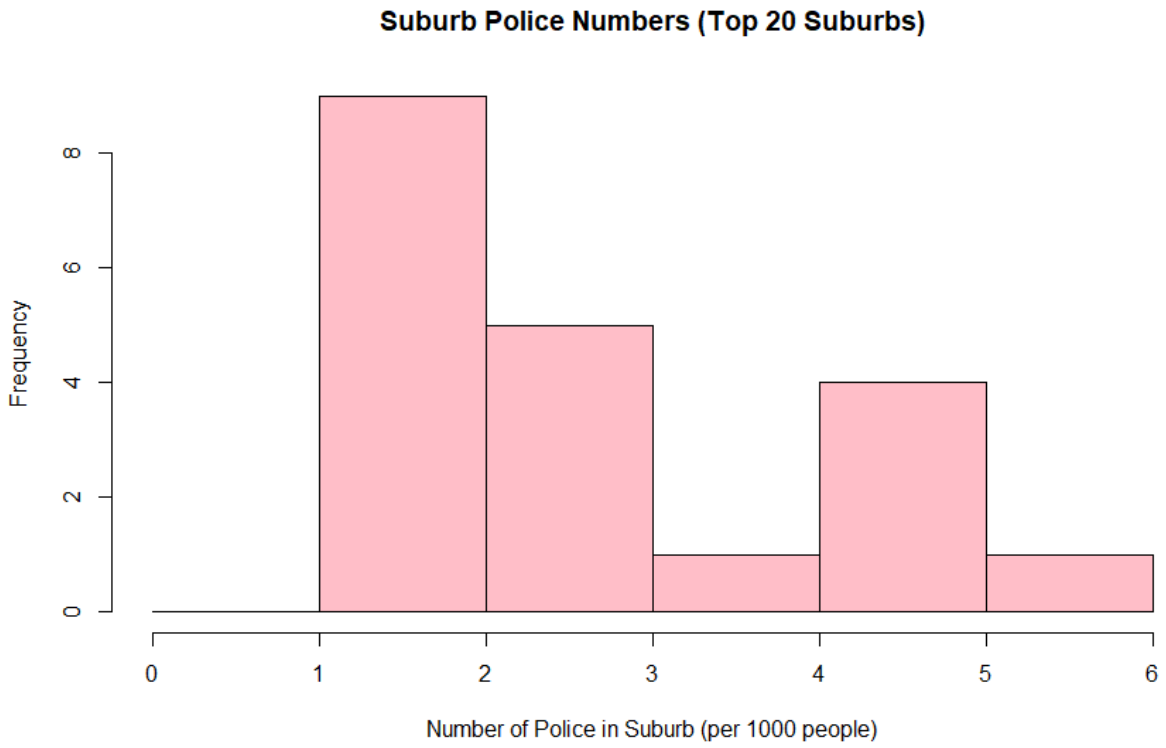
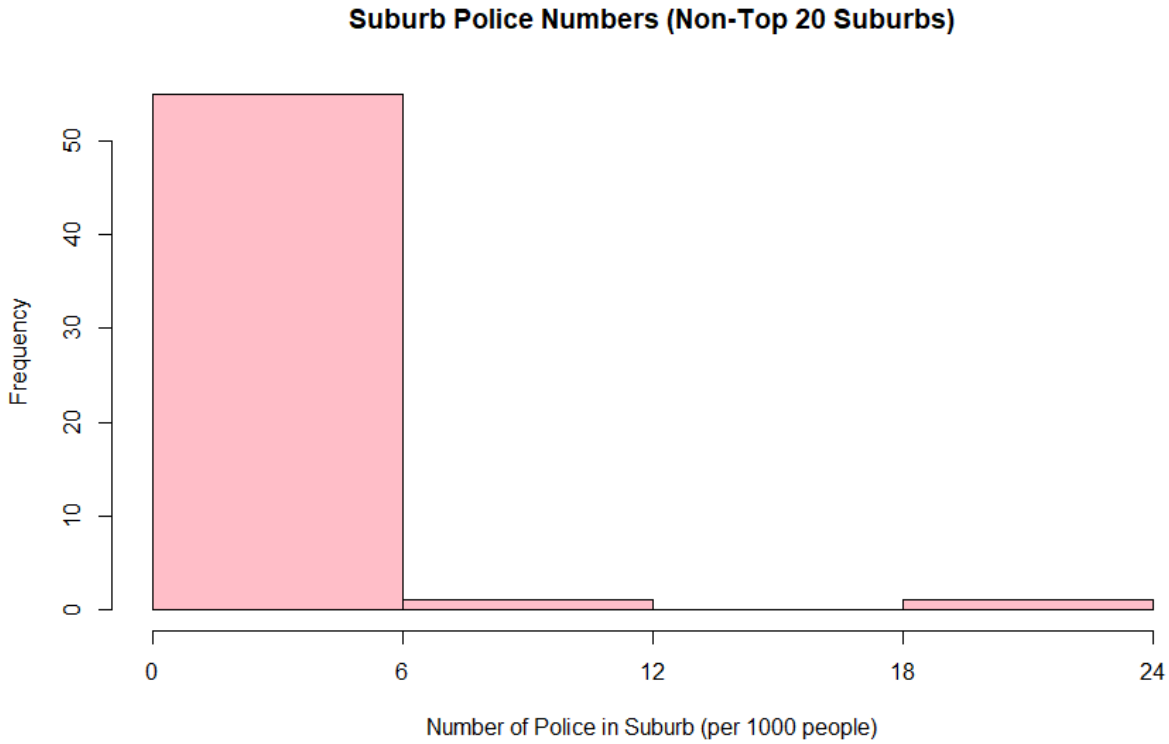


Figure 2.4.b. Suburb Poverty Rates (Non-Top 20 Suburbs)



All top 20 suburbs have less than or equal to 6 police per 1000 citizens. The vast majority of non-top 20 suburbs also have less than or equal to 6 police officers per 1000 citizens. However, there are two non-top 20 suburbs with more than 6 police officers per 1000 people, with one suburb having 21 police officers per 1000 people.

Figure 2.5.a. Suburb Community Services (Top 20 Suburbs)

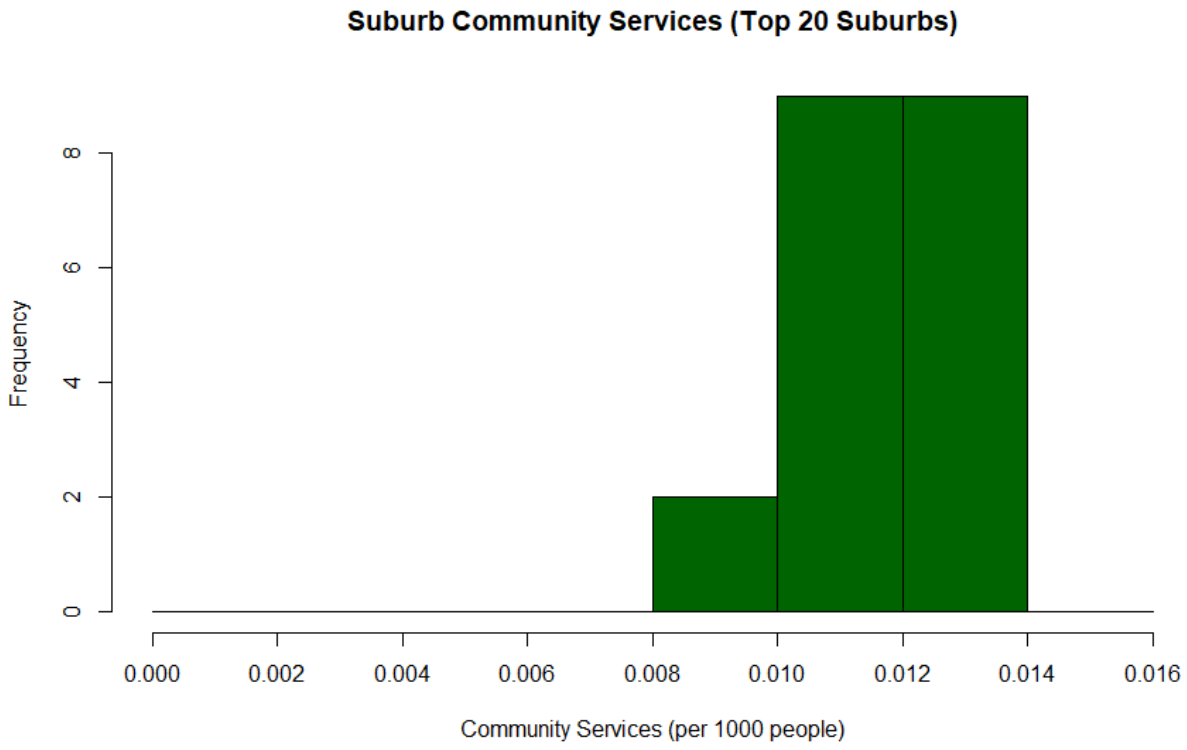
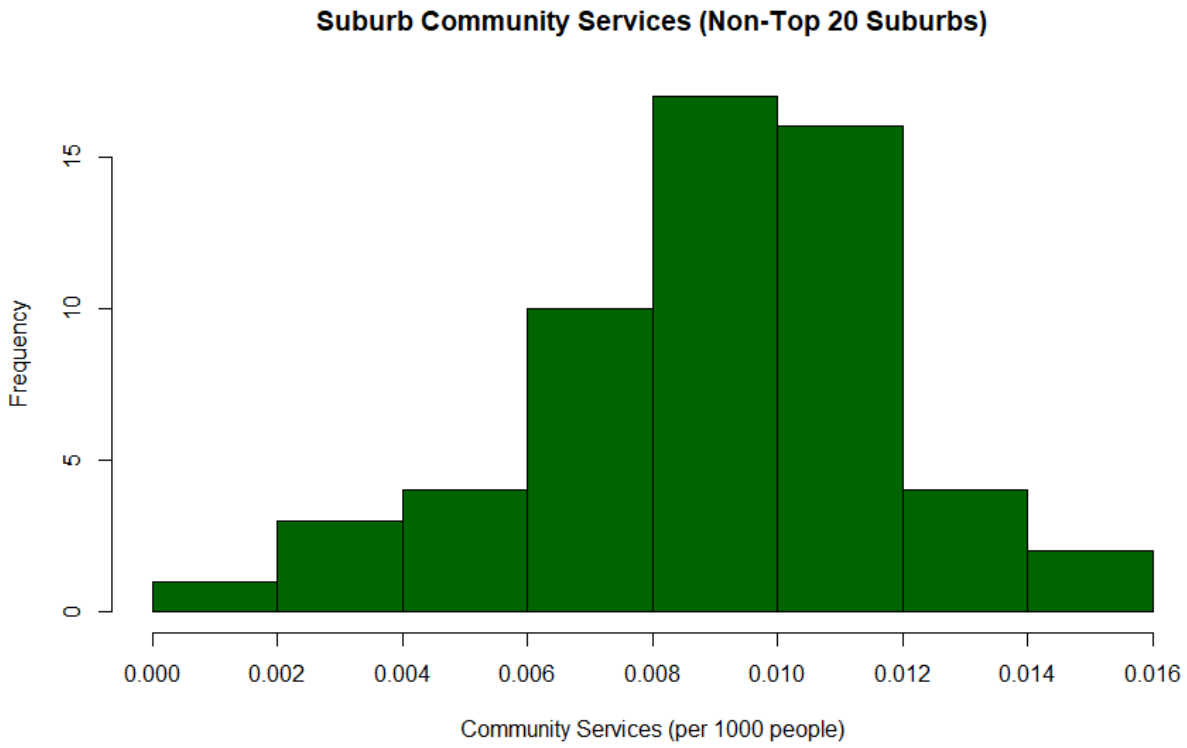


Figure 2.5.b. Suburb Community Services (Non-Top 20 Suburbs)

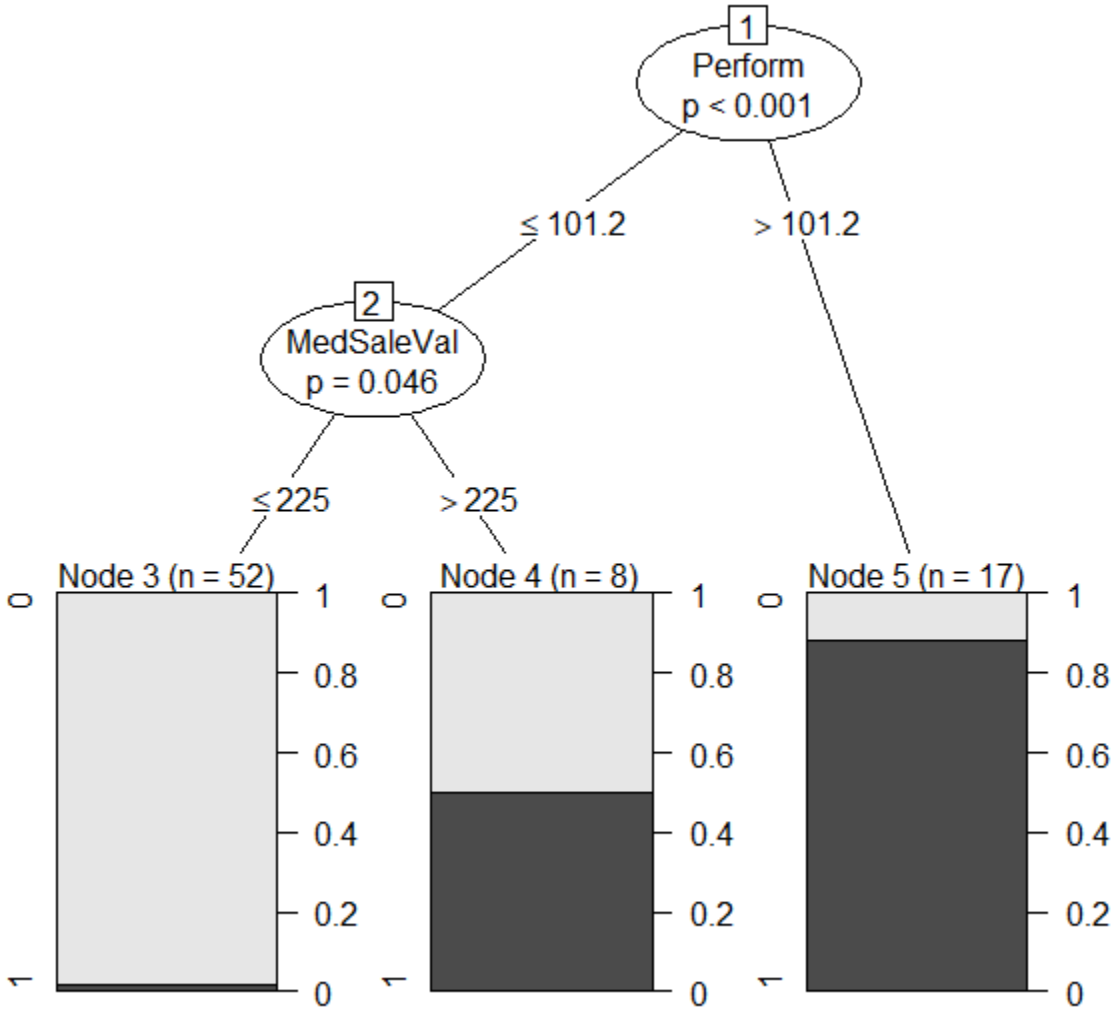


There is not a single top-20 suburb with community services per 1000 people outside the range of 0.008-0.014. By comparison, the majority of non-top 20 suburbs are also within this range. However, there exists a countable number of suburbs with less community services per 1000 people and more community services per 1000 people.

Part 3:

a.

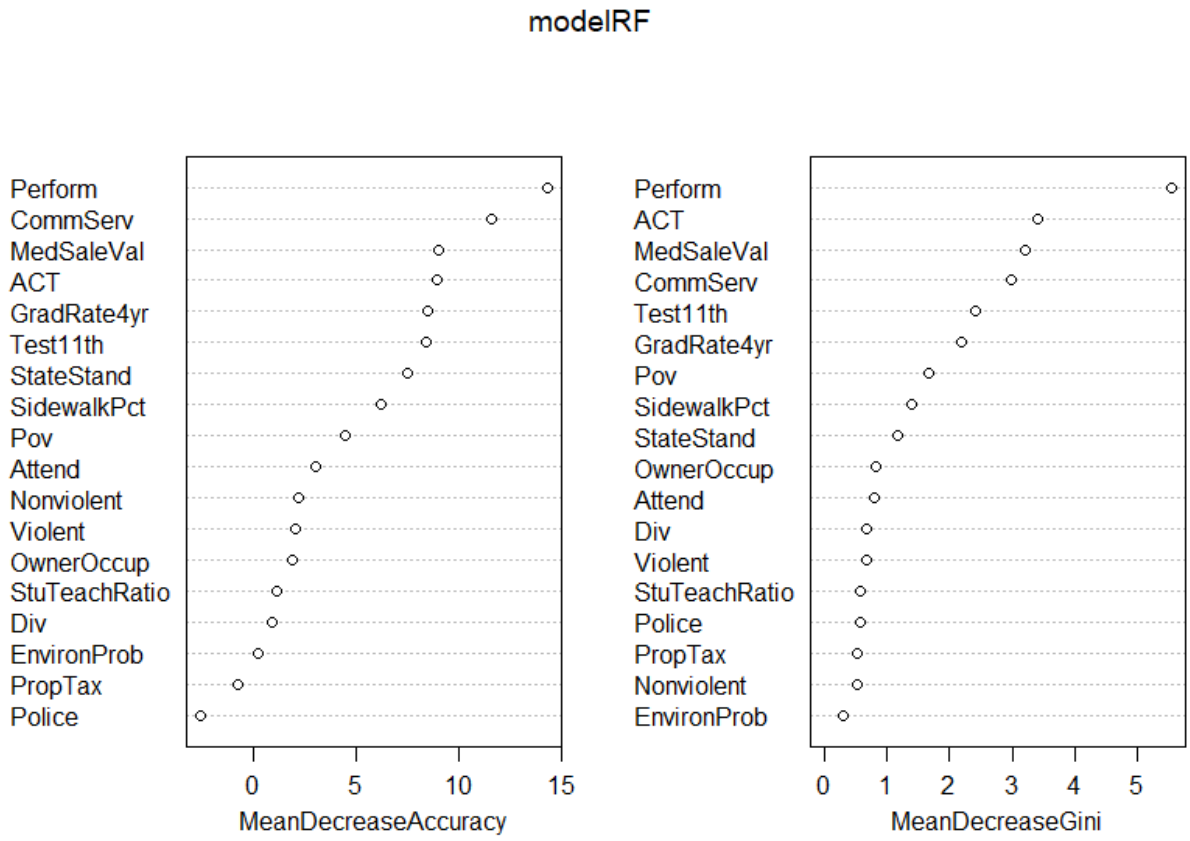
Figure 3.1. Top 20 Suburbs Decision Tree



The Perform variable is a very good indicator if the suburb is in the top 20. If the performance value is below or equal to 101.2, the median sale value is the next indicator. If the median sale value is greater than 225, the indication that a suburb is in the top twenty is about 50%. However, if the performance is lower than/equal to 101.2 and their median sale value is less than 225, the chance that the suburb is not in the top 20 is nearly 100%.

b.

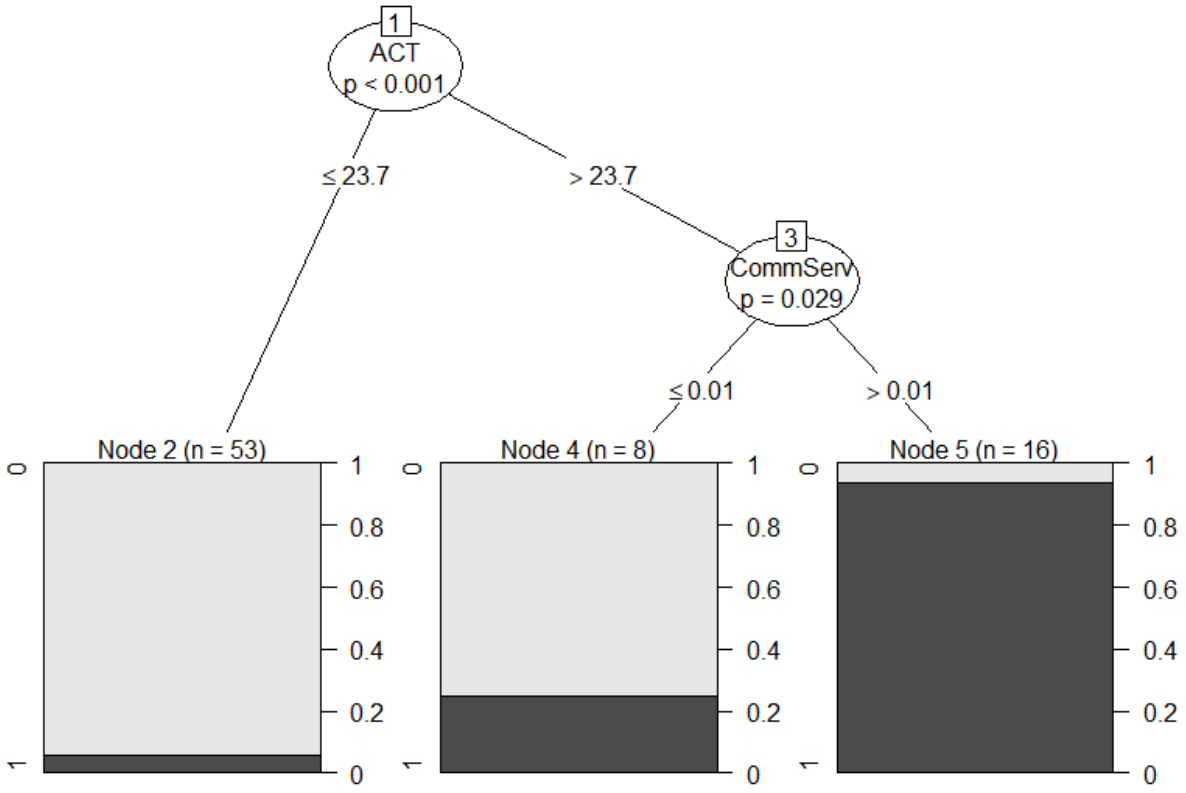
Figure 3.2. Top 20 Suburbs Random Forest



The Ohio department of education performing index is the best indicator of top 20 school performance when it comes to single variables or combined variables. The next best independent factor is community services per 1000 people. However, when being compared to everything else, community services aren't as important as they are individually. Both the median sale of a person's home and the average ACT score in a neighborhood are good factors for determining if a neighborhood is one of the best 20. These factors work both as individual units and in conjunction will all of the other variables. The number of environmental problems per 1000n people is not a good indicator of neighborhood success, whether it is considered individually or with all of the other variables. The same is true for the number of police per 1000 people or the average annual property taxes.

c.

Figure 3.3. Top 20 Suburbs Decision Tree (Without MedSaleVal or Perform)



After removing the factors of the Ohio department education index and the median house values when sold, the most important factors to determine if a neighborhood is one of the top 20 are ACT scores and the number of community services per 1000 people. If the mean ACT score in a neighborhood is less than or equal to 23.7, chances are extremely high that the neighborhood is not one of the best 20. For neighborhoods with an average ACT score above this amount, the number of community services per 1000 people is the best indicator. If the number of community services is greater than 0.01 (and it has above a 23.7 ACT score mean), then the chances of that neighborhood being in the top 20 is extremely close to 1. If a neighborhood is at this limit or below when it comes to community services, the chances of it being a top 20 neighborhood are low, but not as much compared to the neighborhoods with a mean ACT score of 23.7 or less.

Part 4:

a.

I.

After clustering, there are 34 data points in the first cluster and then 43 data points in the second. This can be observed by the table in part II.

II.

Table 4.1. Two Cluster Summary Statistics

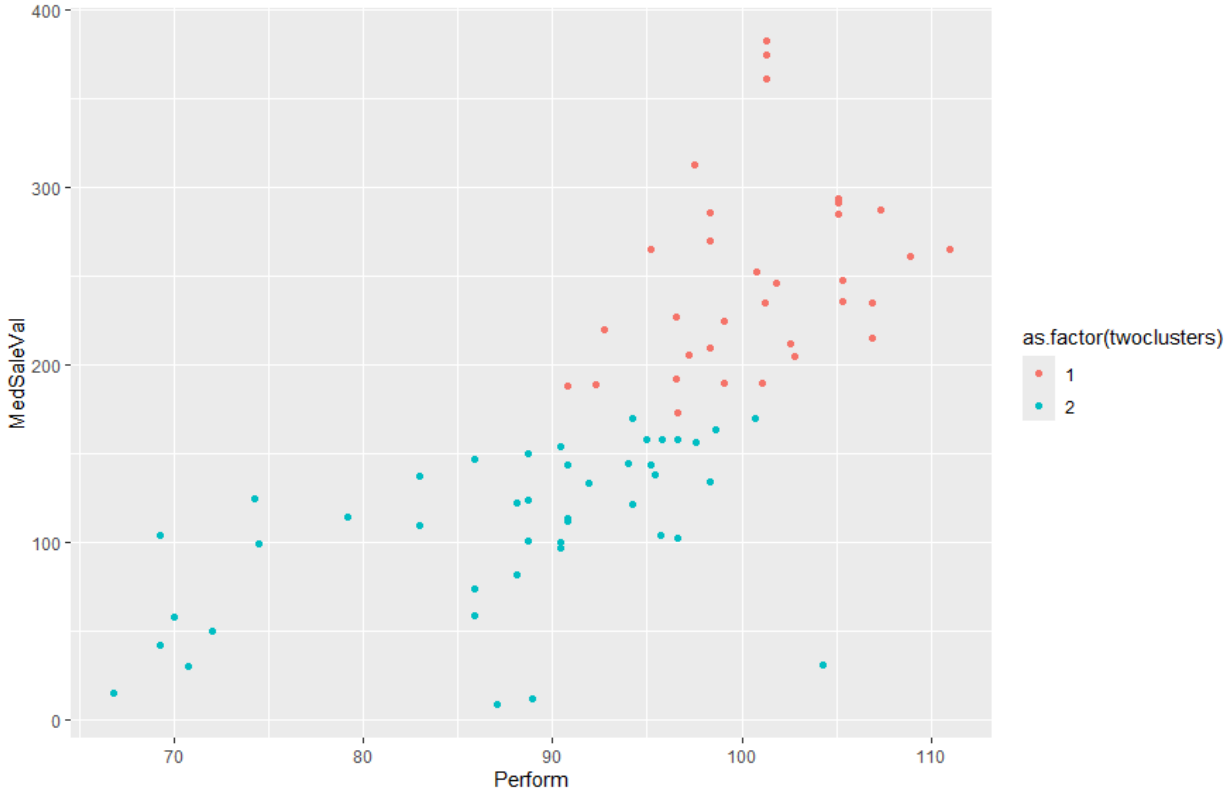
Summary Statistics

twoclusters	1			2		
Variable	N	Mean	SD	N	Mean	SD
MedSaleVal	34	250	53	43	109	46
OwnerOccup	34	85	8.7	43	70	12
PropTax	34	2.2	0.45	43	2.6	0.6
EnvironProb	34	0.0019	0.0021	43	0.0042	0.0049
CommServ	34	0.01	0.0034	43	0.01	0.0025
SidewalkPct	34	36	36	43	70	30
Pov	34	4.4	1.5	43	12	7.3
Div	34	12	9.3	43	28	27
Violent	34	2.6	1.9	43	8.1	11
Nonviolent	34	13	10	43	25	36
Police	34	2.5	1.3	43	2.5	3
StuTeachRatio	34	16	2.2	43	16	1.7
ACT	34	24	1.3	43	21	2.3
StateStand	34	31	1.8	43	20	11
Test11th	34	488	8.8	43	454	32
GradRate4yr	34	95	3.2	43	87	8.4
Attend	34	96	0.5	43	93	6.4
Perform	34	101	4.8	43	88	9.6

The mean median home sale price is very different between clusters. Same is true for variables such as environmental problems per 1000 people(EnvironProb), poverty rate, performance, and violent/nonviolent crimes. Other variables such as community services per 1000 people are roughly the same between clusters. One interesting variable is grade 11 test scores. The mean of this variable between the two clusters is very similar, but there is a massive difference between each standard deviation. The first cluster has a standard deviation of 8.8 whereas the second's standard deviation is 32.

III.

Figure 4.1. Two Clusters Graph



The first cluster has expensive homes in their neighborhood. The exact median value of each home fluctuates, but each suburb in this cluster has more expensive house prices than cluster 2. The education performance index is also much higher for this cluster, as all suburbs in this cluster are on the higher end when it comes to performance.

By contrast, the second cluster represents neighborhoods with a lower median household price. Their performance is scattered, but the vast majority of suburbs in this cluster have below a 100 in the Ohio performance index.

b.

I.

Clusters 1,2,3,4, and 5 have 14, 11, 15, 24, and 13 data points respectively.

II.

Table 4.2. Five Cluster Summary Statistics

Summary Statistics

fiveclusters	1			2			3			4			5		
Variable	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD
MedSaleVal	14	62	38	11	239	21	15	117	30	24	168	31	13	298	48
OwnerOccup	14	59	9.8	11	84	8.3	15	74	11	24	78	9.5	13	89	6.6
PropTax	14	2.9	0.67	11	2.4	0.6	15	2.7	0.5	24	2.2	0.36	13	2.1	0.34
EnvironProb	14	0.0065	0.0057	11	0.003	0.0028	15	0.0037	0.0051	24	0.002	0.0022	13	0.0014	0.0015
CommServ	14	0.011	0.0026	11	0.013	0.0011	15	0.011	0.0022	24	0.0087	0.0028	13	0.0094	0.0035
SidewalkPct	14	88	13	11	78	26	15	84	21	24	30	23	13	12	17
Pov	14	19	8.8	11	4.6	1.6	15	10	4.2	24	7.1	3.3	13	3.8	0.64
Div	14	56	25	11	16	13	15	13	5.9	24	14	19	13	8.9	6.6
Violent	14	10	7.3	11	2.7	1.4	15	9.4	18	24	3.6	2.2	13	2.4	2.2
Nonviolent	14	27	5	11	15	15	15	33	60	24	14	6.6	13	11	7.9
Police	14	2.2	0.65	11	2.4	1	15	3.2	5	24	2.1	1.2	13	2.9	1.4
StuTeachRatio	14	15	1.7	11	16	2.3	15	15	1.8	24	16	1.5	13	16	2.6
ACT	14	18	1.7	11	25	0.89	15	22	1.5	24	22	1.3	13	25	1.2
StateStand	14	7.9	7	11	32	1.2	15	25	6.9	24	29	4.5	13	31	1.1
Test11th	14	420	30	11	486	10	15	468	17	24	478	12	13	492	6.3
GradRate4yr	14	78	8.7	11	96	3.5	15	90	5.1	24	92	3.8	13	97	2.5
Attend	14	91	7.7	11	96	0.59	15	93	7.7	24	95	0.58	13	96	0.33

Cluster 1 has an extremely low mean house value in their suburbs. Clusters 5 and 2 have extremely high mean house values in their suburbs.

There are very few sidewalks in the suburbs of clusters 4 and 5.

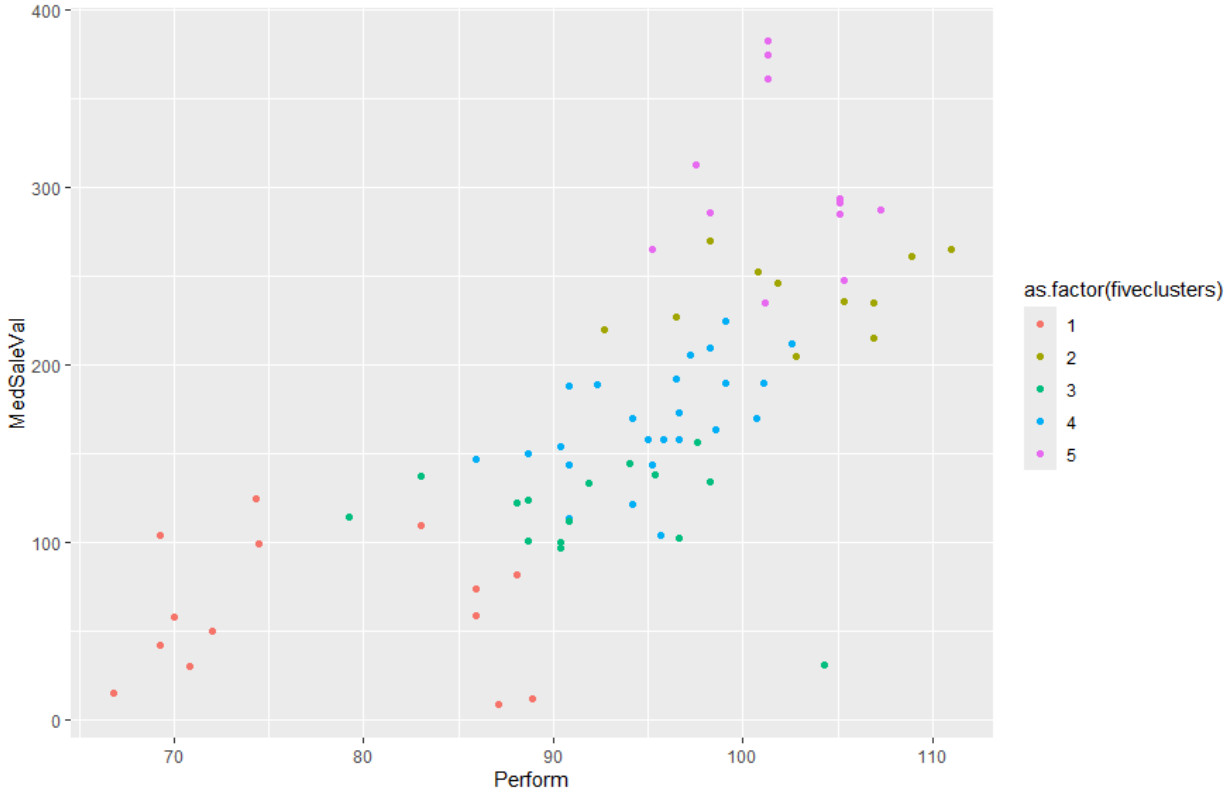
Cluster 1 has high diversity compared to the other groups.

Cluster 1 has a large amount of environmental problems. They also have low average for 4-year graduation rates and ACT scores.

On average, the clusters with the most violent crimes are clusters 1 and 3. Cluster 3 has the most police officers per 1000 people.

III.

Figure 4.2. Five Clusters Graph



Clusters 1, 3, 4, 2, and 5 all have increasing median home values as well as increasing performance scores by the Ohio Department of Education. Each cluster is somewhat mingled with the previous/next one. For example, the bottom portion of cluster 4 is mingled with cluster three, while the top portion of cluster 4 is mingled with cluster 2. Cluster 1 is the most spread out when it comes to performance. Clusters 5, 4, and 1 are very spread out when it comes to the median value of homes in the suburb.

Code:

```
suburbData<-read.csv("C:/Users/natek/Downloads/suburbs_ranking_2016.csv")
head(suburbData)

smallerData<- suburbData[c("SidewalkPct", "PropTax","Pov","Police","CommServ", "Top20")]
head(smallerData)
#Part 1
library(vtable)
testData<-summary(smallerData)
st(smallerData)
library(ggplot2)

ggplot(smallerData, aes(smallerData$SidewalkPct))+geom_histogram(fill="red")

hist(smallerData$SidewalkPct, xlab = "Percentage of Roads with Sidewalks", main="Suburb
Sidewalk Percentage", c="lightblue")
breaksPT<-seq(0, 4, by=0.5)
hist(smallerData$PropTax, breaks=breaksPT, xlab = "Average Annual Property Taxes
(Thousands of dollars)", main="Suburb Annual Property Taxes", c="orange",xaxt="n")
axis(side=1, at =seq(0,4, by=0.5))


hist(suburbData$Pov, breaks = 5, xlab = "Poverty Rate (Percentage of Total Suburb
Population)", main="Suburb Poverty Rates", c="red")

hist(suburbData$Police, xlab = "Number of Police in Suburb (per 1000 people)", main="Suburb
Police Numbers", c="pink",xaxt="n")
axis(side=1, at =seq(0,24, by=2))

breaksCS <- seq(0, 0.016, by = 0.002)
hist(smallerData$CommServ, breaks=breaksCS, xlab = "Community Services (per 1000
people)", main="Suburb Community Services", c="darkgreen",xaxt="n")
axis(side=1, at =seq(0,0.016,0.002))

#Part 2
library(doBy)
NumberOfObservations <- function(x)return(length(x))

summaryBy(SidewalkPct+PropTax+StuTeachRatio+ACT+Attend~Top20, FUN=c(mean,sd,
NumberOfObservations), data=suburbData)

st(smallerData,group='Top20')
```

```
smallerData1 <- smallerData[smallerData['Top20']==1,]
smallerData0 <- smallerData[smallerData['Top20']==0,]

#Histograms for top 20 schools
hist(smallerData1$SidewalkPct, xlab = "Percentage of Roads with Sidewalks", main="Suburb
Sidewalk Percentage (Top 20 Suburbs)", c="lightblue")

breaksPT<-seq(0, 4, by=0.5)
hist(smallerData1$PropTax, breaks=breaksPT, xlab = "Average Annual Property Taxes
(Thousands of dollars)", main="Suburb Annual Property Taxes (Top 20 Suburbs)",
c="orange",xaxt="n")
axis(side=1, at =seq(0,4, by=0.5))

breaksPov <- seq(0, 10, by=1)
hist(smallerData1$Pov, breaks = breaksPov, xlab = "Poverty Rate (Percentage of Total Suburb
Population)", main="Suburb Poverty Rates (Top 20 Suburbs)", c="red", xaxt="n")
axis(side=1, at=seq(0,10,by=1))

breaksPolice <- seq(0,6, by=1)
hist(smallerData1$Police, breaks=breaksPolice, xlab = "Number of Police in Suburb (per 1000
people)", main="Suburb Police Numbers (Top 20 Suburbs)", c="pink",xaxt="n")
axis(side=1, at =seq(0,6, by=1))

breaksCS <- seq(0, 0.016, by = 0.002)
hist(smallerData1$CommServ, breaks=breaksCS, xlab = "Community Services (per 1000
people)", main="Suburb Community Services (Top 20 Suburbs)", c="darkgreen",xaxt="n")
axis(side=1, at =seq(0,0.016,0.002))


#Histograms for non-top 20 schools

hist(smallerData0$SidewalkPct, xlab = "Percentage of Roads with Sidewalks", main="Suburb
Sidewalk Percentage (Non-Top 20 Suburbs)", c="lightblue")

breaksPT<-seq(0, 4, by=0.5)
hist(smallerData0$PropTax, breaks=breaksPT, xlab = "Average Annual Property Taxes
(Thousands of dollars)", main="Suburb Annual Property Taxes (Non-Top 20 Suburbs)",
c="orange",xaxt="n")
axis(side=1, at =seq(0,4, by=0.5))


hist(smallerData0$Pov, breaks = 5, xlab = "Poverty Rate (Percentage of Total Suburb
Population)", main="Suburb Poverty Rates (Non-Top 20 Suburbs)", c="red")
```

```
breaksPolice <- seq(0, 24, by=6)
hist(smallerData0$Police, breaks=breaksPolice,xlab = "Number of Police in Suburb (per 1000
people)", main="Suburb Police Numbers (Non-Top 20 Suburbs)", c="pink",xaxt="n")
axis(side=1, at =seq(0,24, by=6))

breaksCS <- seq(0, 0.016, by = 0.002)
hist(smallerData0$CommServ, breaks=breaksCS, xlab = "Community Services (per 1000
people)", main="Suburb Community Services (Non-Top 20 Suburbs)", c="darkgreen",xaxt="n")
axis(side=1, at =seq(0,0.016,0.002))


#Part 3
#a
library(party)

excludingData <- suburbData[,-match(c("OverallRank",
"SchoolRank","SafetyRank","community", "suburbid"), names(suburbData))]
head(excludingData)

model.ctree <- ctree(as.factor(Top20)~., data=excludingData)
print(model.ctree)
plot(model.ctree)

#b

library(randomForest)
modelRF <- randomForest(as.factor(Top20)~., data = excludingData, importance=T, ntree=)
importance(modelRF)
varImpPlot(modelRF)
#c

part3Data = excludingData[,-match(c("MedSaleVal", "Perform"),names(excludingData))]
part3Model <- ctree(as.factor(Top20)~., data=part3Data)

print(part3Model)
plot(part3Model)

attach(suburbData)
subTable <- table(suburbData$Top20, suburbData$StateStand)
subTable

library(psych)
describe(c(suburbData$ACT,suburbData$Top20))
```

```
#Part 4
clusterableData <- suburbData[,7:24]
head(clusterableData)
set.seed(1234)

#Cluster data
clusterKMeans <- kmeans(clusterableData, 2)
#Find number of suburbs in each group
length(clusterKMeans$cluster[clusterKMeans$cluster==1])
length(clusterKMeans$cluster[clusterKMeans$cluster==2])

suburbData$twoclusters<- clusterKMeans$cluster
#Redefine clusterableData so that I can use the group parameter of the st function.
clusterableData <- suburbData[,7:25]


st(clusterableData, group="twoclusters")

library(ggplot2)


ggplot(suburbData, aes(x=Perform, y=MedSaleVal))
+geom_point(aes(color=as.factor(twoclusters))) #+geom_mark_ellipse(data = suburbData,
aes(color = as.factor(twoclusters)))

kMeans5 <- kmeans(clusterableData, 5)
suburbData$fiveclusters <- kMeans5$cluster
clusterableData <- suburbData[,7:26]

st(clusterableData[,-match(c('twoclusters'), names(clusterableData))], group="fiveclusters")

ggplot(suburbData, aes(x=Perform, y=MedSaleVal, color=as.factor(fiveclusters))) +geom_point()
```