Week\_8\_EDA

Christopher Richards

May 1, 2019

Introduction

This file is my attempt at answering the questions in this week’s assignment.

Three datasets were selected for analysis. The data sets are:  
New York Air Quality Measurements  
Salaries for Professors  
Violent Crime Rates by U.S. States

New York Air Quality Data set

1. Formulate your question

#### Question 1: Is solar radiation affected by ozone density?

1. Read in your data

* I examined the data set in Notepad++ and noticed a number of NAs in the Ozone and Solar.R series. Since these are the values of interest I have decided to omit them.  
  The data set is read in, assigned and viewed.

d\_air <- read.csv("airquality.csv")

1. Check the packaging

* The structure of the data set is examined. First by row count, then by a count of columns.

nrow(d\_air)

## [1] 153

ncol(d\_air)

## [1] 7

1. Run str()

str(d\_air)

## 'data.frame': 153 obs. of 7 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Ozone : int 41 36 12 18 NA 28 23 19 8 NA ...  
## $ Solar.R: int 190 118 149 313 NA NA 299 99 19 194 ...  
## $ Wind : num 7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...  
## $ Temp : int 67 72 74 62 56 66 65 59 61 69 ...  
## $ Month : int 5 5 5 5 5 5 5 5 5 5 ...  
## $ Day : int 1 2 3 4 5 6 7 8 9 10 ...

1. Look at the top and the bottom of your data

head(d\_air)

## X Ozone Solar.R Wind Temp Month Day  
## 1 1 41 190 7.4 67 5 1  
## 2 2 36 118 8.0 72 5 2  
## 3 3 12 149 12.6 74 5 3  
## 4 4 18 313 11.5 62 5 4  
## 5 5 NA NA 14.3 56 5 5  
## 6 6 28 NA 14.9 66 5 6

tail(d\_air)

## X Ozone Solar.R Wind Temp Month Day  
## 148 148 14 20 16.6 63 9 25  
## 149 149 30 193 6.9 70 9 26  
## 150 150 NA 145 13.2 77 9 27  
## 151 151 14 191 14.3 75 9 28  
## 152 152 18 131 8.0 76 9 29  
## 153 153 20 223 11.5 68 9 30

6.Check for missing values

summary(d\_air)

## X Ozone Solar.R Wind   
## Min. : 1 Min. : 1.00 Min. : 7.0 Min. : 1.700   
## 1st Qu.: 39 1st Qu.: 18.00 1st Qu.:115.8 1st Qu.: 7.400   
## Median : 77 Median : 31.50 Median :205.0 Median : 9.700   
## Mean : 77 Mean : 42.13 Mean :185.9 Mean : 9.958   
## 3rd Qu.:115 3rd Qu.: 63.25 3rd Qu.:258.8 3rd Qu.:11.500   
## Max. :153 Max. :168.00 Max. :334.0 Max. :20.700   
## NA's :37 NA's :7   
## Temp Month Day   
## Min. :56.00 Min. :5.000 Min. : 1.0   
## 1st Qu.:72.00 1st Qu.:6.000 1st Qu.: 8.0   
## Median :79.00 Median :7.000 Median :16.0   
## Mean :77.88 Mean :6.993 Mean :15.8   
## 3rd Qu.:85.00 3rd Qu.:8.000 3rd Qu.:23.0   
## Max. :97.00 Max. :9.000 Max. :31.0   
##

As noted after visually inspecting the data, there are Na’s present in the Ozone and Solar.R. I will remove them.

d\_air2 <- d\_air[complete.cases(d\_air),]  
summary(d\_air2)

## X Ozone Solar.R Wind   
## Min. : 1.00 Min. : 1.0 Min. : 7.0 Min. : 2.30   
## 1st Qu.: 45.50 1st Qu.: 18.0 1st Qu.:113.5 1st Qu.: 7.40   
## Median : 89.00 Median : 31.0 Median :207.0 Median : 9.70   
## Mean : 83.95 Mean : 42.1 Mean :184.8 Mean : 9.94   
## 3rd Qu.:124.50 3rd Qu.: 62.0 3rd Qu.:255.5 3rd Qu.:11.50   
## Max. :153.00 Max. :168.0 Max. :334.0 Max. :20.70   
## Temp Month Day   
## Min. :57.00 Min. :5.000 Min. : 1.00   
## 1st Qu.:71.00 1st Qu.:6.000 1st Qu.: 9.00   
## Median :79.00 Median :7.000 Median :16.00   
## Mean :77.79 Mean :7.216 Mean :15.95   
## 3rd Qu.:84.50 3rd Qu.:9.000 3rd Qu.:22.50   
## Max. :97.00 Max. :9.000 Max. :31.00

No Na’s are found.

1. Select relevant columns

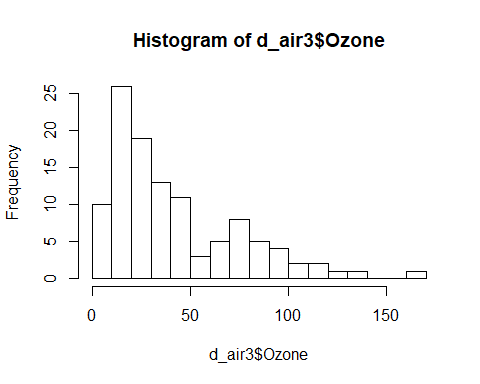
d\_air3 <- d\_air2 %>% select(Ozone, Solar.R)  
summary(d\_air3)

## Ozone Solar.R   
## Min. : 1.0 Min. : 7.0   
## 1st Qu.: 18.0 1st Qu.:113.5   
## Median : 31.0 Median :207.0   
## Mean : 42.1 Mean :184.8   
## 3rd Qu.: 62.0 3rd Qu.:255.5   
## Max. :168.0 Max. :334.0

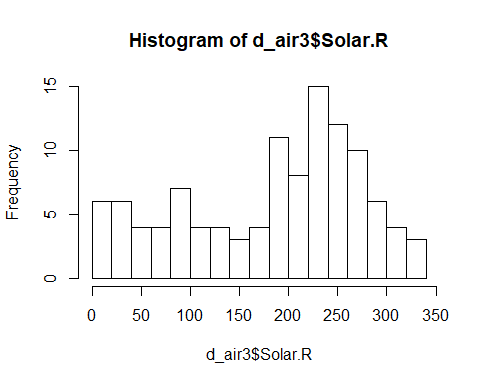
1. Visualize the data

* Peng’s checklist doesn’t include this step here, but, I like to see the data earlier in the process.  
  I’ll perform some simple plots to see the distribution of the data.

hist(d\_air3$Ozone, breaks = 20)



hist(d\_air3$Solar.R, breaks = 20)



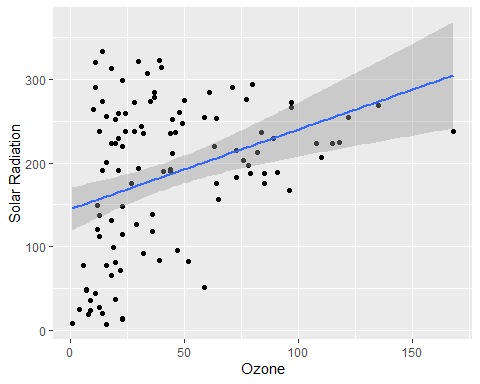
The distribution of Ozone readings is heavily skewed with the majority of points at the low end of the scale. Additionaly, there appears to be outliers above 150 parts per billion.

The distribution of solar radiation readings is less skewed although the highest frequency occurs around 240 Angstroms.

1. Try the simple solution

* There are two simple approaches to answer the question of whether solar radiation is affected by ozone density. One would be a simple scatter plot with a linear regression line to visualize any relationship between the two features. The other solution would be to use a correlation plot to generate numeric values detailing the relationship. I will use both here.

gscat <- ggplot(d\_air3, aes(x=d\_air3$Ozone, y=d\_air3$Solar.R)) +  
 geom\_point() +  
 stat\_smooth(method=lm) +  
 xlab("Ozone") +  
 ylab("Solar Radiation")  
gscat



Here the regression line is trending upwards at an upward rate indicating a positive correlation between solar radiation and ozone. As the density of ozone increases there is a measured increase in solar radiation measured.

Let’s use a correlation matrix next.

cor(d\_air3)

## Ozone Solar.R  
## Ozone 1.0000000 0.3483417  
## Solar.R 0.3483417 1.0000000

As we can see the correlation between the two is positive at .348 which confirms the interpretation of the regression model above.

#### Question 2: Solar Radiation vs Month. Cluster Analysis

In this section, I’ll perform a cluster analysis of solar radiation by month.

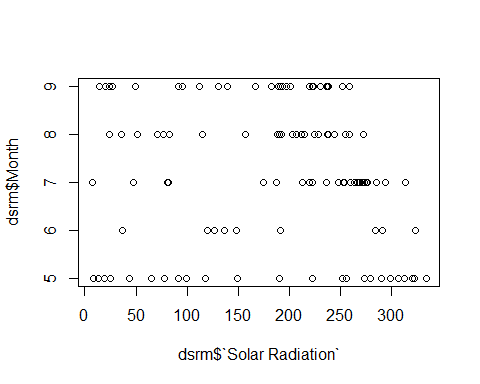
I’ll create a dataframe of the desired features, solar radiationa nd month. The distance between points in the dataframe will be calculated and passed to the clustering algorithim.

dsrm <- data.frame(y=d\_air2$Solar.R, x=d\_air2$Month)  
names(dsrm) <- c("Solar Radiation","Month")  
dsrm

## Solar Radiation Month  
## 1 190 5  
## 2 118 5  
## 3 149 5  
## 4 313 5  
## 5 299 5  
## 6 99 5  
## 7 19 5  
## 8 256 5  
## 9 290 5  
## 10 274 5  
## 11 65 5  
## 12 334 5  
## 13 307 5  
## 14 78 5  
## 15 322 5  
## 16 44 5  
## 17 8 5  
## 18 320 5  
## 19 25 5  
## 20 92 5  
## 21 13 5  
## 22 252 5  
## 23 223 5  
## 24 279 5  
## 25 127 6  
## 26 291 6  
## 27 323 6  
## 28 148 6  
## 29 191 6  
## 30 284 6  
## 31 37 6  
## 32 120 6  
## 33 137 6  
## 34 269 7  
## 35 248 7  
## 36 236 7  
## 37 175 7  
## 38 314 7  
## 39 276 7  
## 40 267 7  
## 41 272 7  
## 42 175 7  
## 43 264 7  
## 44 175 7  
## 45 48 7  
## 46 260 7  
## 47 274 7  
## 48 285 7  
## 49 187 7  
## 50 220 7  
## 51 7 7  
## 52 294 7  
## 53 223 7  
## 54 81 7  
## 55 82 7  
## 56 213 7  
## 57 275 7  
## 58 253 7  
## 59 254 7  
## 60 83 8  
## 61 24 8  
## 62 77 8  
## 63 255 8  
## 64 229 8  
## 65 207 8  
## 66 192 8  
## 67 273 8  
## 68 157 8  
## 69 71 8  
## 70 51 8  
## 71 115 8  
## 72 244 8  
## 73 190 8  
## 74 259 8  
## 75 36 8  
## 76 212 8  
## 77 238 8  
## 78 215 8  
## 79 203 8  
## 80 225 8  
## 81 237 8  
## 82 188 8  
## 83 167 9  
## 84 197 9  
## 85 183 9  
## 86 189 9  
## 87 95 9  
## 88 92 9  
## 89 252 9  
## 90 220 9  
## 91 230 9  
## 92 259 9  
## 93 236 9  
## 94 259 9  
## 95 238 9  
## 96 24 9  
## 97 112 9  
## 98 237 9  
## 99 224 9  
## 100 27 9  
## 101 238 9  
## 102 201 9  
## 103 238 9  
## 104 14 9  
## 105 139 9  
## 106 49 9  
## 107 20 9  
## 108 193 9  
## 109 191 9  
## 110 131 9  
## 111 223 9

Scatterplot of months vs solar radiation

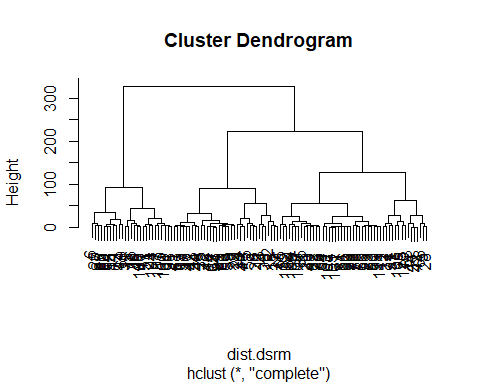
plot(dsrm$`Solar Radiation`, dsrm$Month)



dist.dsrm <- dist(dsrm)

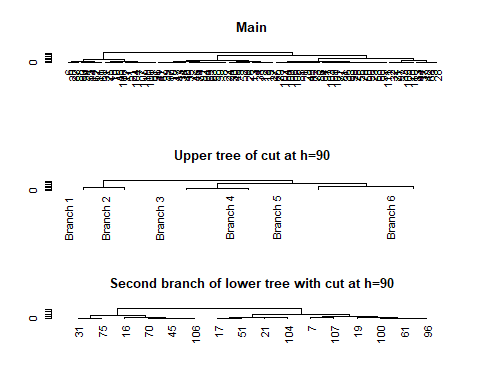
dsrm.clust <- hclust(dist.dsrm)

plot(dsrm.clust)



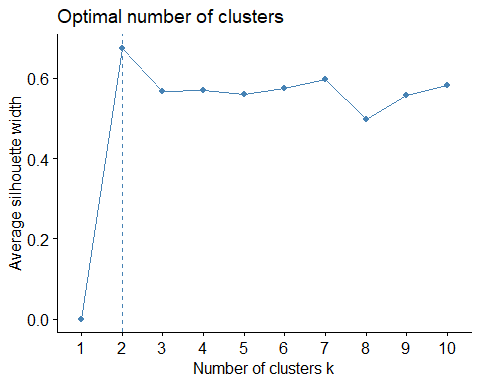
The algorithim has clustered the data however the amount of data points and resulting clusters makes this difficult to read. First, the clustering object will be converted to a dendrogram. Next, I’ll cut the dendrogram at various points to better visualize the results.

hcd <- as.dendrogram(dsrm.clust)  
par(mfrow=c(3,1))  
  
plot(hcd, main="Main")  
plot(cut(x=hcd, h=90)$upper,   
 main="Upper tree of cut at h=90")  
plot(cut(hcd, h=90)$lower[[2]],   
 main="Second branch of lower tree with cut at h=90")



This result isn’t very intuitive to me. I’ll try using kmeans to cluster the data. ### Determine Optimum Clusters #### Average Sihouette Method

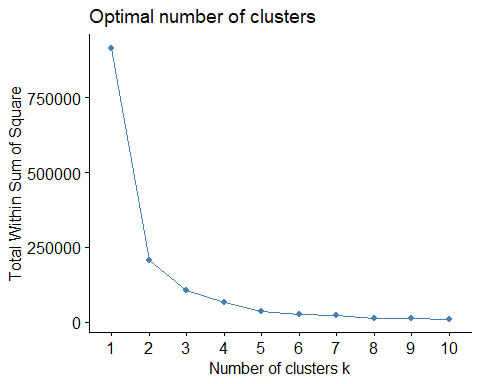
fviz\_nbclust(dsrm, FUNcluster = kmeans, method = "silhouette")



#### Elbow Chart

The best number of clusters will be at the point in the line where it bends to form an “elbow”.

fviz\_nbclust(dsrm, kmeans, method = "wss")



Both methods result in two as the optimal number of clusters. I’ll create a kmeans model of two clusters.

#### KMeans Cluster

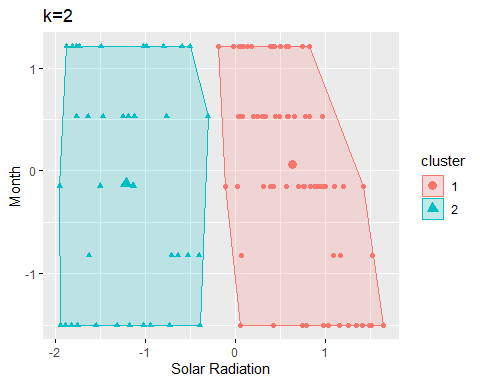
dsrm.kmeans <- kmeans(dsrm, centers = 2)  
dsrm.kmeans

## K-means clustering with 2 clusters of sizes 73, 38  
##   
## Cluster means:  
## Solar Radiation Month  
## 1 242.45205 7.315068  
## 2 74.05263 7.026316  
##   
## Clustering vector:  
## [1] 1 2 2 1 1 2 2 1 1 1 2 1 1 2 1 2 2 1 2 2 2 1 1 1 2 1 1 2 1 1 2 2 2 1 1  
## [36] 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 2 1 1 2 2 1 1 1 1 2 2 2 1 1 1 1 1 2 2 2  
## [71] 2 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 1 1 1 2 2 1 1 2 1 1 1 2 2  
## [106] 2 2 1 1 2 1  
##   
## Within cluster sum of squares by cluster:  
## [1] 126695.84 78798.87  
## (between\_SS / total\_SS = 77.5 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss"   
## [5] "tot.withinss" "betweenss" "size" "iter"   
## [9] "ifault"

The resulting observations in each cluster are unbalanced with almost twice as many points in cluster 1 as in cluster 2. Cluster 1 means are vastly larger then the means in Cluster 2. This may be a result in the difference in numbers of points in each clusters. More likely, though, the points in cluster 1 may simply be distributed across a larger area.

I’ll plot the clusters to visualize the results.

dsrm.p <- fviz\_cluster(dsrm.kmeans, geom = "point", data = dsrm) + ggtitle('k=2')  
dsrm.p



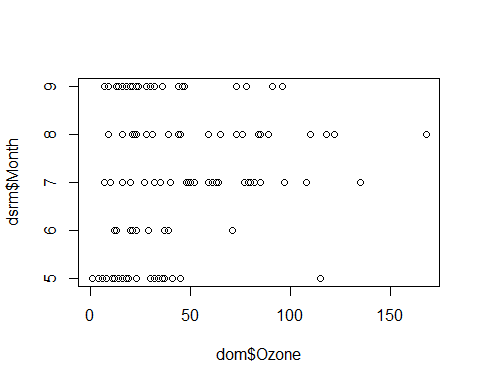
#### Question 3: Ozone vs Month. Cluster Analysis

dom <- data.frame(y=d\_air2$Ozone, x=d\_air2$Month)  
names(dom) <- c("Ozone","Month")  
dom

## Ozone Month  
## 1 41 5  
## 2 36 5  
## 3 12 5  
## 4 18 5  
## 5 23 5  
## 6 19 5  
## 7 8 5  
## 8 16 5  
## 9 11 5  
## 10 14 5  
## 11 18 5  
## 12 14 5  
## 13 34 5  
## 14 6 5  
## 15 30 5  
## 16 11 5  
## 17 1 5  
## 18 11 5  
## 19 4 5  
## 20 32 5  
## 21 23 5  
## 22 45 5  
## 23 115 5  
## 24 37 5  
## 25 29 6  
## 26 71 6  
## 27 39 6  
## 28 23 6  
## 29 21 6  
## 30 37 6  
## 31 20 6  
## 32 12 6  
## 33 13 6  
## 34 135 7  
## 35 49 7  
## 36 32 7  
## 37 64 7  
## 38 40 7  
## 39 77 7  
## 40 97 7  
## 41 97 7  
## 42 85 7  
## 43 10 7  
## 44 27 7  
## 45 7 7  
## 46 48 7  
## 47 35 7  
## 48 61 7  
## 49 79 7  
## 50 63 7  
## 51 16 7  
## 52 80 7  
## 53 108 7  
## 54 20 7  
## 55 52 7  
## 56 82 7  
## 57 50 7  
## 58 64 7  
## 59 59 7  
## 60 39 8  
## 61 9 8  
## 62 16 8  
## 63 122 8  
## 64 89 8  
## 65 110 8  
## 66 44 8  
## 67 28 8  
## 68 65 8  
## 69 22 8  
## 70 59 8  
## 71 23 8  
## 72 31 8  
## 73 44 8  
## 74 21 8  
## 75 9 8  
## 76 45 8  
## 77 168 8  
## 78 73 8  
## 79 76 8  
## 80 118 8  
## 81 84 8  
## 82 85 8  
## 83 96 9  
## 84 78 9  
## 85 73 9  
## 86 91 9  
## 87 47 9  
## 88 32 9  
## 89 20 9  
## 90 23 9  
## 91 21 9  
## 92 24 9  
## 93 44 9  
## 94 21 9  
## 95 28 9  
## 96 9 9  
## 97 13 9  
## 98 46 9  
## 99 18 9  
## 100 13 9  
## 101 24 9  
## 102 16 9  
## 103 13 9  
## 104 23 9  
## 105 36 9  
## 106 7 9  
## 107 14 9  
## 108 30 9  
## 109 14 9  
## 110 18 9  
## 111 20 9

Scatterplot of months vs ozone

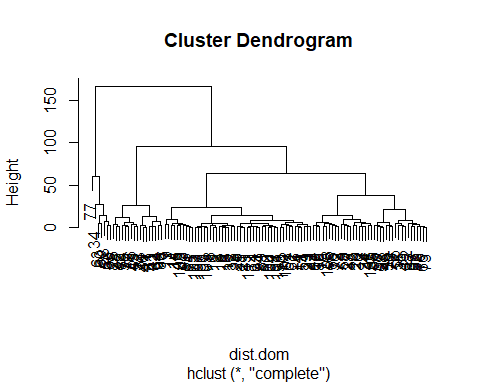
plot(dom$`Ozone`, dsrm$Month)



dist.dom <- dist(dom)

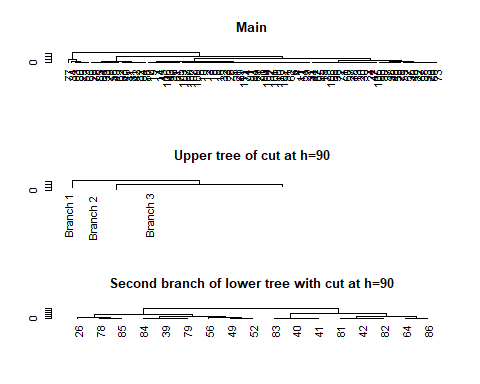
dom.clust <- hclust(dist.dom)

plot(dom.clust)



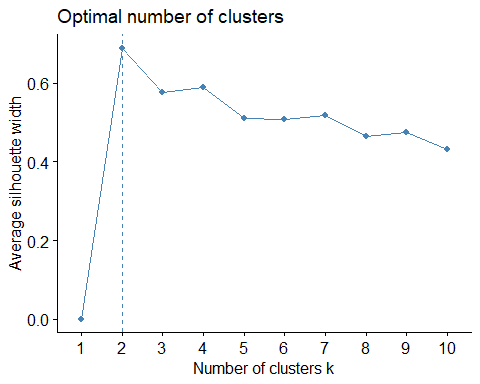
The algorithim has clustered the data however the amount of data points and resulting clusters makes this difficult to read. First, the clustering object will be converted to a dendrogram. Next, I’ll cut the dendrogram at various points to better visualize the results.

hcd <- as.dendrogram(dom.clust)  
par(mfrow=c(3,1))  
  
plot(hcd, main="Main")  
plot(cut(x=hcd, h=90)$upper,   
 main="Upper tree of cut at h=90")  
plot(cut(hcd, h=90)$lower[[2]],   
 main="Second branch of lower tree with cut at h=90")



This result isn’t very intuitive to me. I’ll try using kmeans to cluster the data. ### Determine Optimum Clusters #### Average Sihouette Method

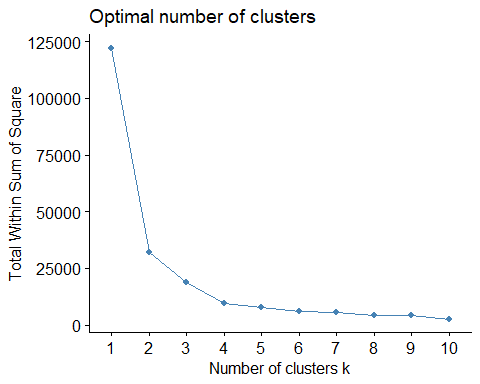
fviz\_nbclust(dom, FUNcluster = kmeans, method = "silhouette")



#### Elbow Chart

The best number of clusters will be at the point in the line where it bends to form an “elbow”.

fviz\_nbclust(dom, kmeans, method = "wss")



Both methods result in two as the optimal number of clusters. I’ll create a kmeans model of two clusters.

#### KMeans Cluster

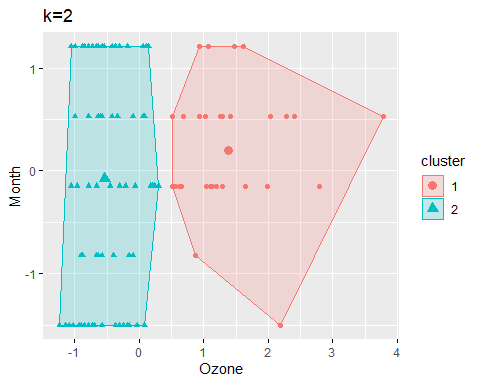
dom.kmeans <- kmeans(dom, centers = 2)  
dom.kmeans

## K-means clustering with 2 clusters of sizes 31, 80  
##   
## Cluster means:  
## Ozone Month  
## 1 87.87097 7.516129  
## 2 24.36250 7.100000  
##   
## Clustering vector:  
## [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 1 2 2 2 2 2 2 2 1 2  
## [36] 2 1 2 1 1 1 1 2 2 2 2 2 1 1 1 2 1 1 2 2 1 2 1 1 2 2 2 1 1 1 2 2 1 2 1  
## [71] 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
## [106] 2 2 2 2 2 2  
##   
## Within cluster sum of squares by cluster:  
## [1] 18573.23 13349.69  
## (between\_SS / total\_SS = 73.8 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss"   
## [5] "tot.withinss" "betweenss" "size" "iter"   
## [9] "ifault"

The resulting observations in each cluster are unbalanced again, although not as badly as the Solar Radiation vs Month clusters. This may be a result in the difference in numbers of points in each clusters. More likely, though, the points in cluster 1 may simply be distributed across a larger area.

I’ll plot the clusters to visualize the results.

dom.p <- fviz\_cluster(dom.kmeans, geom = "point", data = dom) + ggtitle('k=2')  
dom.p



As we can see from the table of cluster means and the cluster plot, Cluster 1 encompasses a larger area then Cluster 2. Clearly there are several points in Cluster 1 which account for this dispersion.

Salaries for Professors Data set

1. Formulate your question

#### Question 1: Do males achieve higher rank earlier in their professional careers then females?

1. Read in your data

d.sal <- read.csv("Salaries.csv")

1. Check the packaging

* The structure of the data set is examined. First by row count, then by a count of columns.

nrow(d.sal)

## [1] 397

ncol(d.sal)

## [1] 7

1. Run str()

str(d.sal)

## 'data.frame': 397 obs. of 7 variables:  
## $ X : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ rank : Factor w/ 3 levels "AssocProf","AsstProf",..: 3 3 2 3 3 1 3 3 3 3 ...  
## $ discipline : Factor w/ 2 levels "A","B": 2 2 2 2 2 2 2 2 2 2 ...  
## $ yrs.since.phd: int 19 20 4 45 40 6 30 45 21 18 ...  
## $ yrs.service : int 18 16 3 39 41 6 23 45 20 18 ...  
## $ sex : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 1 ...  
## $ salary : int 139750 173200 79750 115000 141500 97000 175000 147765 119250 129000 ...

1. Look at the top and the bottom of your data

head(d.sal)

## X rank discipline yrs.since.phd yrs.service sex salary  
## 1 1 Prof B 19 18 Male 139750  
## 2 2 Prof B 20 16 Male 173200  
## 3 3 AsstProf B 4 3 Male 79750  
## 4 4 Prof B 45 39 Male 115000  
## 5 5 Prof B 40 41 Male 141500  
## 6 6 AssocProf B 6 6 Male 97000

tail(d.sal)

## X rank discipline yrs.since.phd yrs.service sex salary  
## 392 392 Prof A 30 19 Male 151292  
## 393 393 Prof A 33 30 Male 103106  
## 394 394 Prof A 31 19 Male 150564  
## 395 395 Prof A 42 25 Male 101738  
## 396 396 Prof A 25 15 Male 95329  
## 397 397 AsstProf A 8 4 Male 81035

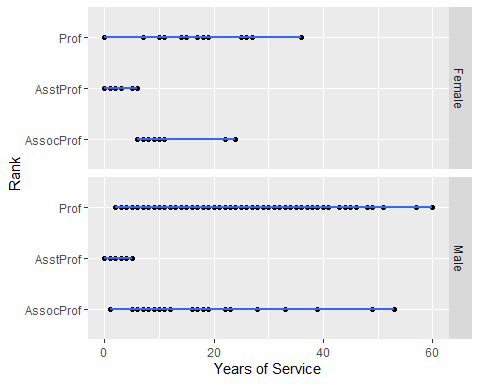
6.Check for missing values

summary(d.sal)

## X rank discipline yrs.since.phd yrs.service   
## Min. : 1 AssocProf: 64 A:181 Min. : 1.00 Min. : 0.00   
## 1st Qu.:100 AsstProf : 67 B:216 1st Qu.:12.00 1st Qu.: 7.00   
## Median :199 Prof :266 Median :21.00 Median :16.00   
## Mean :199 Mean :22.31 Mean :17.61   
## 3rd Qu.:298 3rd Qu.:32.00 3rd Qu.:27.00   
## Max. :397 Max. :56.00 Max. :60.00   
## sex salary   
## Female: 39 Min. : 57800   
## Male :358 1st Qu.: 91000   
## Median :107300   
## Mean :113706   
## 3rd Qu.:134185   
## Max. :231545

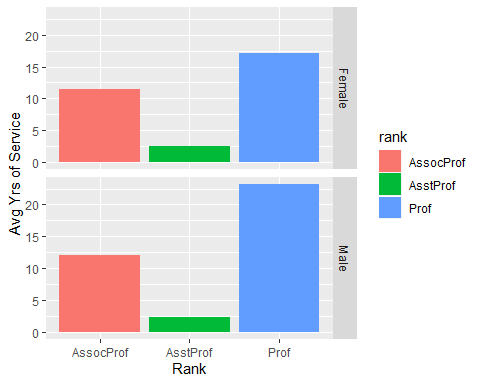
Plot service yrs vs rank, grouped by sex

gsal <- ggplot(data=d.sal, aes(x=d.sal$yrs.service, y=d.sal$rank)) +   
 geom\_point() +  
 facet\_grid(sex~.) +  
 geom\_smooth(aes(line = d.sal$sex), method = lm) +  
 xlab("Years of Service") +  
 ylab("Rank")  
gsal



Try a barchart: x is rank, y is avg yrs, bars by sex

ggplot(d.sal, aes(x=factor(d.sal$rank),  
 y=d.sal$yrs.service,  
 fill = rank)) +  
 stat\_summary(fun.y="mean", geom="bar") +  
 xlab("Rank") +  
 ylab("Avg Yrs of Service") +  
 facet\_grid(sex~.)



In order to answer this question, I plotted the average years of service by sex for each rank.  
For Associate Professors the average years of service for each sex is fairly even with females having slightly fewer years of service. Associate Professors are essentially even. Full Professors show the most deviation between the sexes. Males have over 22 years of service on average while females have fewer years. Females have about 17 years of service.

Based on the information available, I would say that our hypothesis, do males achieve higher rank earlier in their careers then females, to be false. If it were true I would have expected males to have a lower average years of service for full Professor positions. This chart shows the opposite of that.

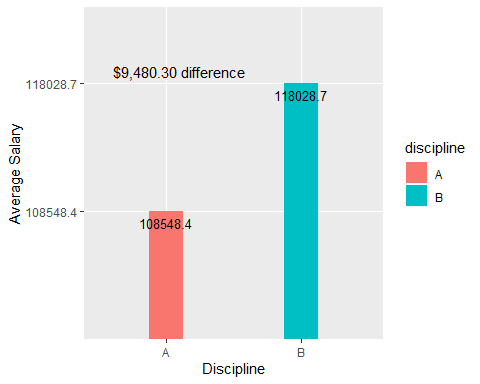
1. Formulate your question

#### Question 2: Does discipline make a difference in salary?

* In this section, I will examine whether the choice of discipline makes a difference in salaries for professors.  
  The provided data for the “discipline” feature is limited. The levels are “A”, theoretical disciplines, and “B”, applied disciplines.
* Since our data is already read in I will proceed with a simple solution. To answer the question I will find the average salary by discipline. All data will be considered in the calculation

mean.sal.disc <- d.sal %>%  
 group\_by(discipline) %>%  
 summarise(mean = mean(salary)) %>%  
 mutate\_if(is.numeric, format, 0)

ggplot(mean.sal.disc, aes(x = discipline, y = mean, fill = discipline)) +  
 geom\_bar(stat = "identity", width = .25) +   
 geom\_text(aes(label=mean), vjust=1.5, color="black", size=3.5) +  
 xlab("Discipline") +  
 ylab("Average Salary") +  
 annotate("text", x = 1.1, y=2.1, label = "$9,480.30 difference")



The average salary for applied disciplines, “B”, is higher by $9480.30 then for theoretical, “A”, disciplines. This simple approach tells us that, in general, the discipline of the professor makes a difference in salary.

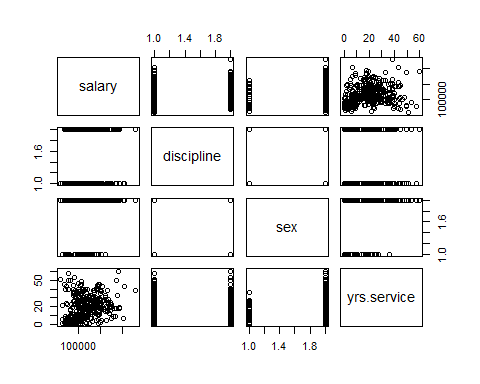
1. Formulate your question

#### Question 3: Which factor (discipline, sex, years of service) is more important in determining salary?

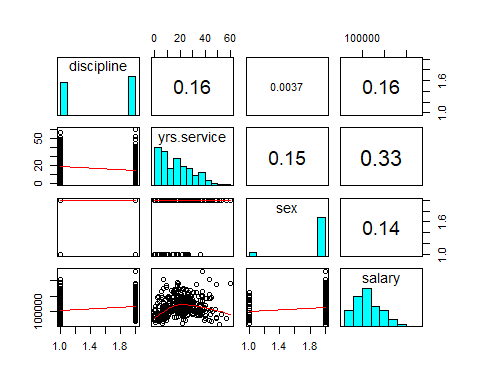
* In this section, I will examine which factor has the choice of discipline makes a difference in salaries for professors.

corr matrix  
encode discipline, sex

pairs(~ salary + discipline + sex + yrs.service, data = d.sal)



## put histograms on the diagonal  
panel.hist <- function(x, ...)  
{  
 usr <- par("usr"); on.exit(par(usr))  
 par(usr = c(usr[1:2], 0, 1.5) )  
 h <- hist(x, plot = FALSE)  
 breaks <- h$breaks; nB <- length(breaks)  
 y <- h$counts; y <- y/max(y)  
 rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)  
}  
## put (absolute) correlations on the upper panels,  
## with size proportional to the correlations.  
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)  
{  
 usr <- par("usr"); on.exit(par(usr))  
 par(usr = c(0, 1, 0, 1))  
 r <- abs(cor(x, y))  
 txt <- format(c(r, 0.123456789), digits = digits)[1]  
 txt <- paste0(prefix, txt)  
 if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)  
 # text(0.5, 0.5, txt, cex = cex.cor \* r) # This is the original from ?pairs()  
 text(0.5, 0.5, txt, cex = cex.cor \* (1 + r) / 2) # This is modified to soften font size changes  
}  
pairs(d.sal[, c(3,5,6,7)], upper.panel = panel.cor,  
 diag.panel = panel.hist,  
 lower.panel = panel.smooth)

 According to the correlation chart, Years of Service is the most important factor when determining salary. The Years of Service correlation is .33, about twice as much as the other factors.

Violent Crime Rates by State Data set

1. Formulate your question

#### Question 1: Which region is the most “murderous”?

1. Read in your data

d.crime <- read.csv("USArrests.csv")

1. Check the packaging

* The structure of the data set is examined. First by row count, then by a count of columns.

nrow(d.crime)

## [1] 50

ncol(d.crime)

## [1] 5

1. Run str()

str(d.crime)

## 'data.frame': 50 obs. of 5 variables:  
## $ X : Factor w/ 50 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...  
## $ Murder : num 13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...  
## $ Assault : int 236 263 294 190 276 204 110 238 335 211 ...  
## $ UrbanPop: int 58 48 80 50 91 78 77 72 80 60 ...  
## $ Rape : num 21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...

1. Look at the top and the bottom of your data

head(d.crime)

## X Murder Assault UrbanPop Rape  
## 1 Alabama 13.2 236 58 21.2  
## 2 Alaska 10.0 263 48 44.5  
## 3 Arizona 8.1 294 80 31.0  
## 4 Arkansas 8.8 190 50 19.5  
## 5 California 9.0 276 91 40.6  
## 6 Colorado 7.9 204 78 38.7

tail(d.crime)

## X Murder Assault UrbanPop Rape  
## 45 Vermont 2.2 48 32 11.2  
## 46 Virginia 8.5 156 63 20.7  
## 47 Washington 4.0 145 73 26.2  
## 48 West Virginia 5.7 81 39 9.3  
## 49 Wisconsin 2.6 53 66 10.8  
## 50 Wyoming 6.8 161 60 15.6

6.Check for missing values

summary(d.crime)

## X Murder Assault UrbanPop   
## Alabama : 1 Min. : 0.800 Min. : 45.0 Min. :32.00   
## Alaska : 1 1st Qu.: 4.075 1st Qu.:109.0 1st Qu.:54.50   
## Arizona : 1 Median : 7.250 Median :159.0 Median :66.00   
## Arkansas : 1 Mean : 7.788 Mean :170.8 Mean :65.54   
## California: 1 3rd Qu.:11.250 3rd Qu.:249.0 3rd Qu.:77.75   
## Colorado : 1 Max. :17.400 Max. :337.0 Max. :91.00   
## (Other) :44   
## Rape   
## Min. : 7.30   
## 1st Qu.:15.07   
## Median :20.10   
## Mean :21.23   
## 3rd Qu.:26.18   
## Max. :46.00   
##

Data looks clean. No Na’s are found.

1. Select relevant columns

* I’ll create a new dataframe of State and Murder observations. State names are in the “X” column. I’ll rename it to “State”.

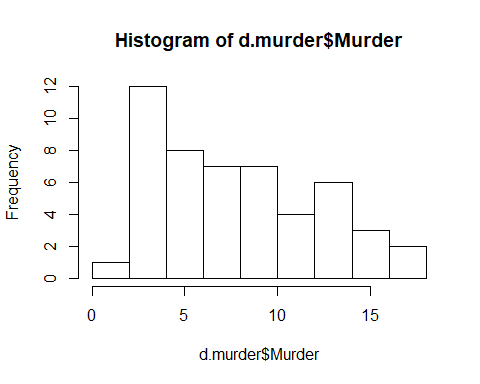
colnames(d.crime)[1] <- "State"  
d.murder <- d.crime %>% select(State, Murder)  
summary(d.murder)

## State Murder   
## Alabama : 1 Min. : 0.800   
## Alaska : 1 1st Qu.: 4.075   
## Arizona : 1 Median : 7.250   
## Arkansas : 1 Mean : 7.788   
## California: 1 3rd Qu.:11.250   
## Colorado : 1 Max. :17.400   
## (Other) :44

1. Visualize the data

* Murder Histogram

hist(d.murder$Murder)



1. Try the simple solution.

* I’m not sure how to do a cluster analysis for this data set. The State feature is a factor that will not lend itself to clustering using kmeans, etc since it is unable to calculate a distance from a factor. I could encode the states but the resulting encoded data would indicate an ordered relationship which is misleading.

Additionally, the question asks for finding the region with the most murders but there is no “region” variable in the data. I will add a “region” feature using the data found at <https://www.50states.com/city/regions.htm>.

I’ve created a list of states and their associated region. I’ll load that now.

regions <- read.csv("us\_states.csv")  
head(regions)

## state region  
## 1 Alabama Deep South  
## 2 Alaska Pacific West  
## 3 Arizona Southwest  
## 4 Arkansas South Central  
## 5 California Pacific West  
## 6 Colorado Southwest

Add the region column to the d.murder dataframe.

d.murderreg <- d.murder %>% mutate(region = regions$region)  
head(d.murderreg)

## State Murder region  
## 1 Alabama 13.2 Deep South  
## 2 Alaska 10.0 Pacific West  
## 3 Arizona 8.1 Southwest  
## 4 Arkansas 8.8 South Central  
## 5 California 9.0 Pacific West  
## 6 Colorado 7.9 Southwest

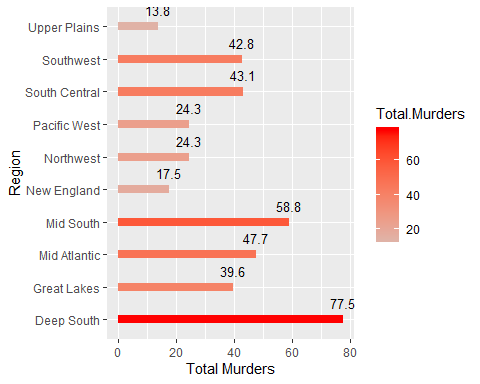
I’ll group the data by region then determine the total murder for each region.

mrdrgn <- d.murderreg %>%  
 group\_by(region) %>%  
 summarise("Total.Murders" = sum(Murder)) %>%  
 data.frame() %>%   
 arrange(desc(Total.Murders))  
mrdrgn

## region Total.Murders  
## 1 Deep South 77.5  
## 2 Mid South 58.8  
## 3 Mid Atlantic 47.7  
## 4 South Central 43.1  
## 5 Southwest 42.8  
## 6 Great Lakes 39.6  
## 7 Northwest 24.3  
## 8 Pacific West 24.3  
## 9 New England 17.5  
## 10 Upper Plains 13.8

I’ll show the result in a barplot

ggplot(mrdrgn, aes(x = region, y = Total.Murders, fill = Total.Murders)) +  
 geom\_bar(stat = "identity", width = .25 ) +   
 geom\_text(aes(label=Total.Murders), vjust=-1, color="black", size=3.5) +  
 coord\_flip() +  
 scale\_fill\_gradient2(low='green', mid='snow3', high='red') +  
 xlab("Region") +  
 ylab("Total Murders") #+



# annotate("text", x = 1, y=40, label = "Most Murderous Region is the Deep South")

The most murderous region in the United States is the Deep South.

#### Question 2: Which region is rape most likely to occur?

1. Select relevant columns

* I’ll create a new dataframe of State and Murder observations. State names are in the “X” column. I’ll rename it to “State”.

d.rape <- d.crime %>% select(State, Rape)  
summary(d.rape)

## State Rape   
## Alabama : 1 Min. : 7.30   
## Alaska : 1 1st Qu.:15.07   
## Arizona : 1 Median :20.10   
## Arkansas : 1 Mean :21.23   
## California: 1 3rd Qu.:26.18   
## Colorado : 1 Max. :46.00   
## (Other) :44

Add Region column.

d.rapereg <- d.rape %>% mutate(region = regions$region)  
head(d.rapereg)

## State Rape region  
## 1 Alabama 21.2 Deep South  
## 2 Alaska 44.5 Pacific West  
## 3 Arizona 31.0 Southwest  
## 4 Arkansas 19.5 South Central  
## 5 California 40.6 Pacific West  
## 6 Colorado 38.7 Southwest

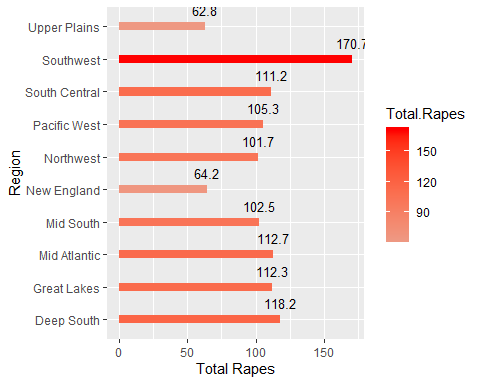
I’ll group the data by region then determine the total murder for each region.

rprgn <- d.rapereg %>%  
 group\_by(region) %>%  
 summarise("Total.Rapes" = sum(Rape)) %>%  
 data.frame() %>%   
 arrange(desc(Total.Rapes))  
rprgn

## region Total.Rapes  
## 1 Southwest 170.7  
## 2 Deep South 118.2  
## 3 Mid Atlantic 112.7  
## 4 Great Lakes 112.3  
## 5 South Central 111.2  
## 6 Pacific West 105.3  
## 7 Mid South 102.5  
## 8 Northwest 101.7  
## 9 New England 64.2  
## 10 Upper Plains 62.8

I’ll show the result in a barplot

ggplot(rprgn, aes(x = region, y = Total.Rapes, fill = Total.Rapes)) +  
 geom\_bar(stat = "identity", width = .25 ) +   
 geom\_text(aes(label=Total.Rapes), vjust=-1, color="black", size=3.5) +  
 coord\_flip() +  
 scale\_fill\_gradient2(low='green', mid='snow3', high='red') +  
 xlab("Region") +  
 ylab("Total Rapes") #+



# annotate("text", x = 1, y=40, label = "Most Murderous Region is the Deep South")

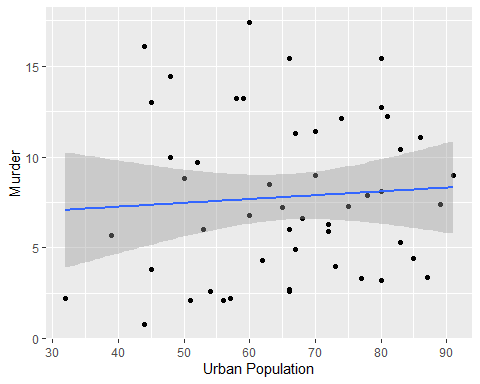
The region a woman is most likely to be raped in is the Southwest with 170.7 rapes recorded.

#### Question 3. How does Population Desity affect all types of crimes?

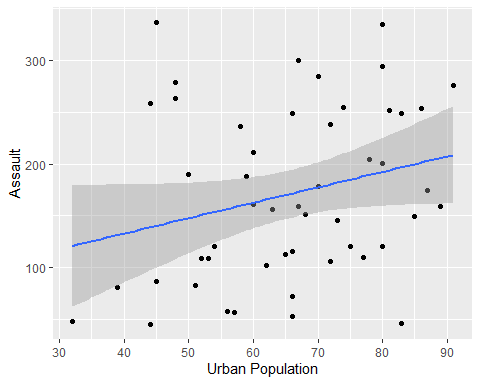
1. Try the simple solution

* In this case, I will plot scatterplots with regression lines of all types of crime against Population Density. ```

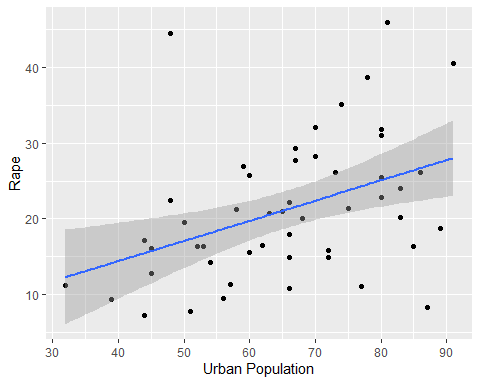
mrdr.pop <- ggplot(d.crime, aes(x=d.crime$UrbanPop, y=d.crime$Murder)) +  
 geom\_point() +  
 geom\_smooth(method = lm) +  
 xlab("Urban Population") +  
 ylab("Murder")  
mrdr.pop



aslt.pop <- ggplot(d.crime, aes(x=d.crime$UrbanPop, y=d.crime$Assault)) +  
 geom\_point() +  
 geom\_smooth(method = lm) +  
 xlab("Urban Population") +  
 ylab("Assault")  
aslt.pop



rape.pop <- ggplot(d.crime, aes(x=d.crime$UrbanPop, y=d.crime$Rape)) +  
 geom\_point() +  
 geom\_smooth(method = lm) +  
 xlab("Urban Population") +  
 ylab("Rape")  
rape.pop

 It appears from the regression lines of the plots that all three categories of crime are positively affected by an increase in urban population density. Judging by the slope of the lines, assault appears to be most affected as it increases the most when urban population increases. Arrests for rape rises, as well, although not as rapidly. Murder is the least affected, although it does show a small increase.

Conclusion

This assignment was very challenging. I feel I learned a lot by working through the analyses as well as I could. Unfortunately, it’s not clear if my approaches to the questions were appropriate. I found this assignement frustrating because of that uncertainty and would have benefited from more direction in approaching the questions.