Assignment #2 - Introduction to Data Analysis

Sheetal Gangakhedkar

May 3, 2017

## Questions

**1. Create two matrices from the given set of numbers X1={2,3,7,1,6,2,3,5,1} and x2={3,2,9,0,7,8,5,8,2}**

# create matrix X1 and X2 with given data  
X1 = matrix(c(2,3,7,1,6,2,3,5,1), nrow=3, ncol=3, byrow=FALSE)  
X2 = matrix(c(3,2,9,0,7,8,5,8,2), nrow=3, ncol=3, byrow=FALSE)

**2. Find the matrix product**

# multiply matrix X1 with X2  
prodX1X2 = X1 %\*% X2  
prodX1X2

## [,1] [,2] [,3]  
## [1,] 35 31 24  
## [2,] 66 82 73  
## [3,] 34 22 53

**3. Find the inverse of the matrix and prove that it is correct.**

# solve to get the inverse of matrix  
iX1 = solve(X1)  
class(iX1) == "matrix"

## [1] TRUE

iX1

## [,1] [,2] [,3]  
## [1,] 0.04761905 -0.05952381 0.15476190  
## [2,] -0.38095238 0.22619048 0.01190476  
## [3,] 0.42857143 -0.03571429 -0.10714286

# prove that X1 multiplied with its inverse returns an identity matrix  
identityX1 = iX1 %\*% X1  
identityX1

## [,1] [,2] [,3]  
## [1,] 1.000000e+00 1.110223e-16 -2.775558e-17  
## [2,] 1.387779e-17 1.000000e+00 2.550044e-16  
## [3,] 0.000000e+00 0.000000e+00 1.000000e+00

isSymmetric(identityX1)

## [1] TRUE

# solve to get the inverse of matrix  
iX2 = solve(X2)  
class(iX2) == "matrix"

## [1] TRUE

iX2

## [,1] [,2] [,3]  
## [1,] 0.1298701 -0.10389610 0.09090909  
## [2,] -0.1766234 0.10129870 0.03636364  
## [3,] 0.1220779 0.06233766 -0.05454545

# prove that X1 multiplied with its inverse returns an identity matrix  
identityX2 = iX2 %\*% X2  
identityX2

## [,1] [,2] [,3]  
## [1,] 1.000000e+00 -1.110223e-16 -8.326673e-17  
## [2,] 1.665335e-16 1.000000e+00 0.000000e+00  
## [3,] -5.551115e-17 -5.551115e-17 1.000000e+00

isSymmetric(identityX2)

## [1] TRUE

**4. Load the Animals data set from the "MASS" package**

data(package="MASS", "Animals")  
head(Animals)

## body brain  
## Mountain beaver 1.35 8.1  
## Cow 465.00 423.0  
## Grey wolf 36.33 119.5  
## Goat 27.66 115.0  
## Guinea pig 1.04 5.5  
## Dipliodocus 11700.00 50.0

**5. Find the correlation coefficient of brain and body in this data set and comment on the relation between them**

# convert animal body weight into same units 'grams' as the brain weight, and then get correlation coeff  
# help(package="MASS", "Animals")  
cor(Animals$body\*1000, Animals$brain, method="pearson")

## [1] -0.005341163

It appears the Animal body weight and brain weight have a very low negative relationship and almost no correlation, using the default "pearson" method.

**6. Load USArrests data set. Comment on the distribution of the variables. Is there any relation between UrbanPop and three crimes.**

data(USArrests)  
head(USArrests)

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

dim(USArrests)

## [1] 50 4

***Understanding the distribution of variables in the USArrests data set.***

summary(USArrests)

## Murder Assault UrbanPop Rape   
## Min. : 0.800 Min. : 45.0 Min. :32.00 Min. : 7.30   
## 1st Qu.: 4.075 1st Qu.:109.0 1st Qu.:54.50 1st Qu.:15.07   
## Median : 7.250 Median :159.0 Median :66.00 Median :20.10   
## Mean : 7.788 Mean :170.8 Mean :65.54 Mean :21.23   
## 3rd Qu.:11.250 3rd Qu.:249.0 3rd Qu.:77.75 3rd Qu.:26.18   
## Max. :17.400 Max. :337.0 Max. :91.00 Max. :46.00

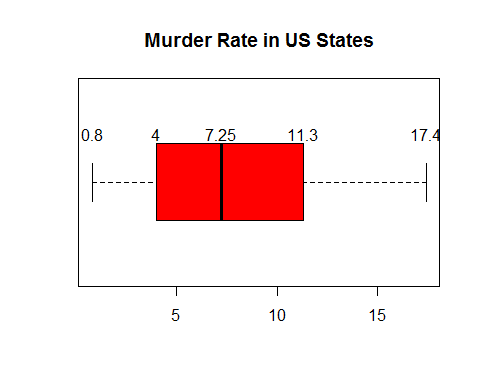
We see that Murder rate is centered at 7.25, spread is 16.6 and interquartile range is 7.175

We see that Assault rate is centered at 159.0, spread is 292.0 and interquartile range is 140.0

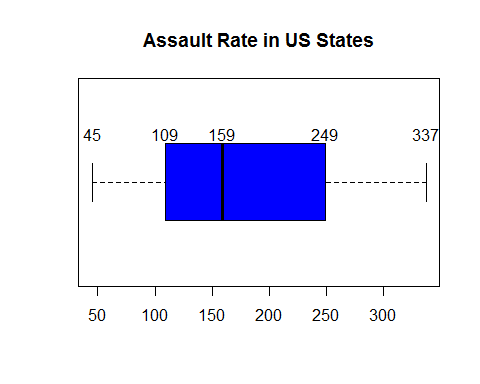
We see that Urban Population rate is centered at 66.0, spread is 59 and interquartile range is 23.25

We see that Rape rate is centered at 20.10, spread is 38.7 and interquartile range is 11.11

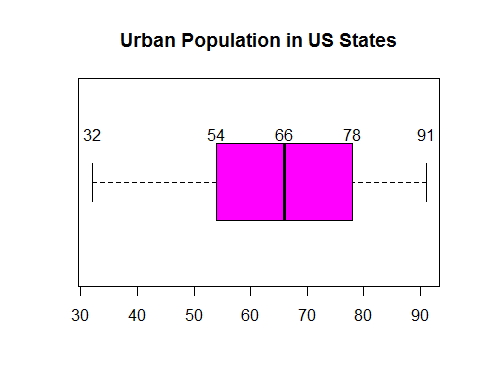
boxplot(USArrests$Murder, horizontal=TRUE, col=2, main="Murder Rate in US States")  
text(x=fivenum(USArrests$Murder), labels =fivenum(USArrests$Murder), y=1.25)



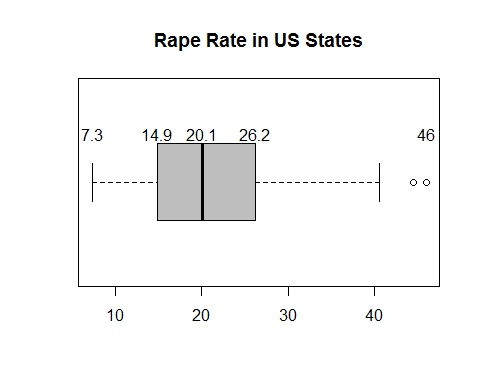
boxplot(USArrests$Assault, horizontal=TRUE, col=4, main="Assault Rate in US States")  
text(x=fivenum(USArrests$Assault), labels =fivenum(USArrests$Assault), y=1.25)



boxplot(USArrests$UrbanPop, horizontal=TRUE, col=6, main="Urban Population in US States")  
text(x=fivenum(USArrests$UrbanPop), labels =fivenum(USArrests$UrbanPop), y=1.25)



boxplot(USArrests$Rape, horizontal=TRUE, col=8, main="Rape Rate in US States")  
text(x=fivenum(USArrests$Rape), labels=fivenum(USArrests$Rape), y=1.25)



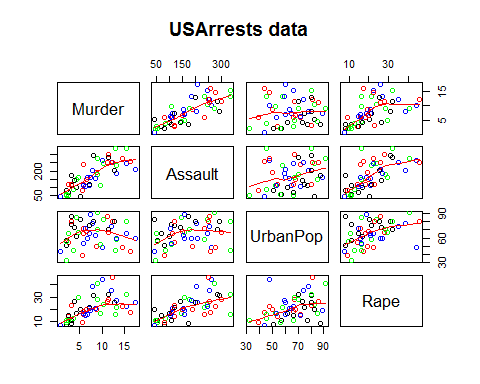
***Standard deviation of USArrests data for Murder, Assault, Urban Population and Rape rates data***

sapply(USArrests[,1:4], sd)

## Murder Assault UrbanPop Rape   
## 4.355510 83.337661 14.474763 9.366385

***Distribution plot of USArrests data set***

pairs(USArrests, panel = panel.smooth, main = "USArrests data", col=c("green","blue","black","red"))



***Correlation of data columns in USArrests data set***

cor(USArrests[,1:4])

## Murder Assault UrbanPop Rape  
## Murder 1.00000000 0.8018733 0.06957262 0.5635788  
## Assault 0.80187331 1.0000000 0.25887170 0.6652412  
## UrbanPop 0.06957262 0.2588717 1.00000000 0.4113412  
## Rape 0.56357883 0.6652412 0.41134124 1.0000000

Correlation data shows that Murder and Assault rates are positively and highly correlated, followed medium level correlation between Assault and Rape rates. The least related is the Urban Polulation and Murder rates data.

cor(USArrests$UrbanPop, USArrests$Murder)

## [1] 0.06957262

cor(USArrests$UrbanPop, USArrests$Assault)

## [1] 0.2588717

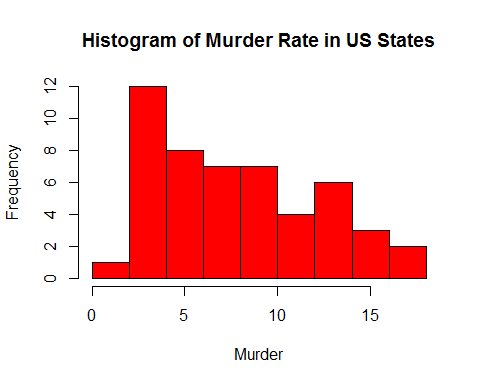
cor(USArrests$UrbanPop, USArrests$Rape)

## [1] 0.4113412

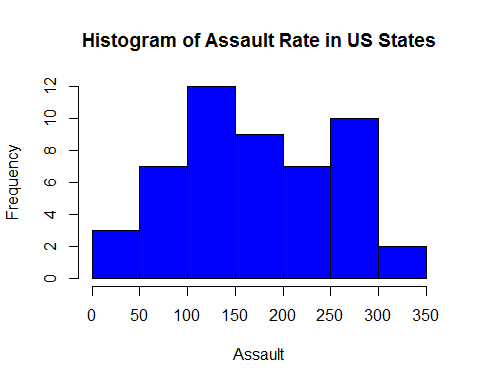
Urban Population and Murder rates have no correlation with a positive but very weak or no relationship. Urban Population and Assault rates have a low correlation with a positive relationship. Urban Population and Rape rates have a medium correlation with a positive relationship.

***Histogram Distribution of variables in the USArrests data set.***

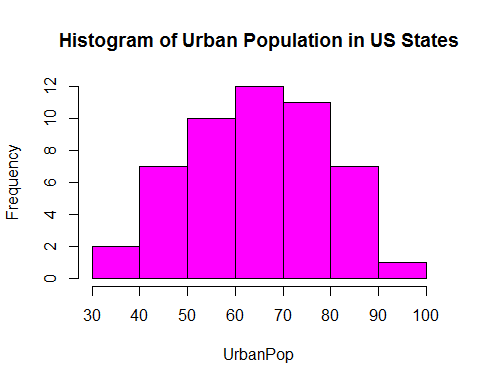
# histogram of Murder arrests per 100k popluation in 50 US states  
hist(USArrests[,1], col=2, main="Histogram of Murder Rate in US States", xlab=colnames(USArrests)[1])



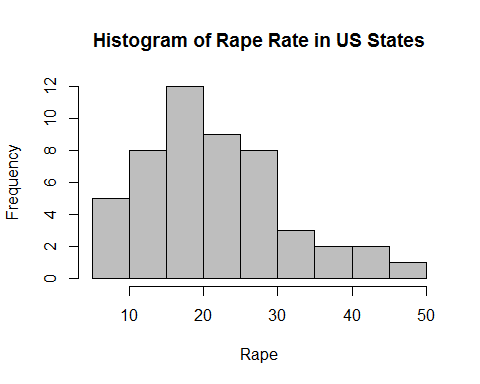
# histogram of Assault arrests per 100k popluation in 50 US states  
hist(USArrests[,2], col=4, main="Histogram of Assault Rate in US States", xlab=colnames(USArrests)[2])



# histogram of Urban population per 100k popluation in 50 US states  
hist(USArrests[,3], col=6, main="Histogram of Urban Population in US States", xlab=colnames(USArrests)[3])



# histogram of Rape arrests per 100k popluation in 50 US states  
hist(USArrests[,4], col=8, main="Histogram of Rape Rate in US States", xlab=colnames(USArrests)[4])



**7. Which states has most and least assault, murder, and rape arrests.**

# states with least Murder arrests  
USArrests[which(USArrests$Murder==min(USArrests$Murder)),]

## Murder Assault UrbanPop Rape  
## North Dakota 0.8 45 44 7.3

row.names(USArrests[which(USArrests$Murder==min(USArrests$Murder)),])

## [1] "North Dakota"

# states with least Assault arrests  
USArrests[which(USArrests$Assault==min(USArrests$Assault)),]

## Murder Assault UrbanPop Rape  
## North Dakota 0.8 45 44 7.3

row.names(USArrests[which(USArrests$Assault==min(USArrests$Assault)),])

## [1] "North Dakota"

# states with least Rape arrests  
USArrests[which(USArrests$Rape==min(USArrests$Rape)),]

## Murder Assault UrbanPop Rape  
## North Dakota 0.8 45 44 7.3

row.names(USArrests[which(USArrests$Rape==min(USArrests$Rape)),])

## [1] "North Dakota"

# states with most Murder arrests  
USArrests[which(USArrests$Murder==max(USArrests$Murder)),]

## Murder Assault UrbanPop Rape  
## Georgia 17.4 211 60 25.8

row.names(USArrests[which(USArrests$Murder==max(USArrests$Murder)),])

## [1] "Georgia"

# states with most Assault arrests  
USArrests[which(USArrests$Assault==max(USArrests$Assault)),]

## Murder Assault UrbanPop Rape  
## North Carolina 13 337 45 16.1

row.names(USArrests[which(USArrests$Assault==max(USArrests$Assault)),])

## [1] "North Carolina"

# states with most Rape arrests  
USArrests[which(USArrests$Rape==max(USArrests$Rape)),]

## Murder Assault UrbanPop Rape  
## Nevada 12.2 252 81 46

row.names(USArrests[which(USArrests$Rape==max(USArrests$Rape)),])

## [1] "Nevada"

**8. List the states which have assault arrests more than median of the country.**

median(USArrests$Assault)

## [1] 159

USArrests[which(USArrests$Assault>median(USArrests$Assault)),]

## Murder Assault UrbanPop Rape  
## Alabama 13.2 236 58 21.2  
## Alaska 10.0 263 48 44.5  
## Arizona 8.1 294 80 31.0  
## Arkansas 8.8 190 50 19.5  
## California 9.0 276 91 40.6  
## Colorado 7.9 204 78 38.7  
## Delaware 5.9 238 72 15.8  
## Florida 15.4 335 80 31.9  
## Georgia 17.4 211 60 25.8  
## Illinois 10.4 249 83 24.0  
## Louisiana 15.4 249 66 22.2  
## Maryland 11.3 300 67 27.8  
## Michigan 12.1 255 74 35.1  
## Mississippi 16.1 259 44 17.1  
## Missouri 9.0 178 70 28.2  
## Nevada 12.2 252 81 46.0  
## New Mexico 11.4 285 70 32.1  
## New York 11.1 254 86 26.1  
## North Carolina 13.0 337 45 16.1  
## Rhode Island 3.4 174 87 8.3  
## South Carolina 14.4 279 48 22.5  
## Tennessee 13.2 188 59 26.9  
## Texas 12.7 201 80 25.5  
## Wyoming 6.8 161 60 15.6

row.names(USArrests[which(USArrests$Assault>median(USArrests$Assault)),])

## [1] "Alabama" "Alaska" "Arizona" "Arkansas"   
## [5] "California" "Colorado" "Delaware" "Florida"   
## [9] "Georgia" "Illinois" "Louisiana" "Maryland"   
## [13] "Michigan" "Mississippi" "Missouri" "Nevada"   
## [17] "New Mexico" "New York" "North Carolina" "Rhode Island"   
## [21] "South Carolina" "Tennessee" "Texas" "Wyoming"

**9. Which states are in the bottom 25% of murder**

Bottom 25% means in this case, states having murder rates above 75% or 3rd Quartile murder rate

# 3rd Quantile for murder rates  
quartile3 = quantile(USArrests$Murder)[4]  
quartile3

## 75%   
## 11.25

USArrests[which(USArrests$Murder>=quartile3),]

## Murder Assault UrbanPop Rape  
## Alabama 13.2 236 58 21.2  
## Florida 15.4 335 80 31.9  
## Georgia 17.4 211 60 25.8  
## Louisiana 15.4 249 66 22.2  
## Maryland 11.3 300 67 27.8  
## Michigan 12.1 255 74 35.1  
## Mississippi 16.1 259 44 17.1  
## Nevada 12.2 252 81 46.0  
## New Mexico 11.4 285 70 32.1  
## North Carolina 13.0 337 45 16.1  
## South Carolina 14.4 279 48 22.5  
## Tennessee 13.2 188 59 26.9  
## Texas 12.7 201 80 25.5

row.names(USArrests[which(USArrests$Murder>=quartile3),])

## [1] "Alabama" "Florida" "Georgia" "Louisiana"   
## [5] "Maryland" "Michigan" "Mississippi" "Nevada"   
## [9] "New Mexico" "North Carolina" "South Carolina" "Tennessee"   
## [13] "Texas"

**10. Which states are in the top 25% of the murder.**

Top 25% means in this case, states having murder rates below 25% or 1st Quartile murder rate

# 1st Quantile for murder rates  
quartile1 = quantile(USArrests$Murder)[2]  
quartile1

## 25%   
## 4.075

USArrests[which(USArrests$Murder<=quartile1),]

## Murder Assault UrbanPop Rape  
## Connecticut 3.3 110 77 11.1  
## Idaho 2.6 120 54 14.2  
## Iowa 2.2 56 57 11.3  
## Maine 2.1 83 51 7.8  
## Minnesota 2.7 72 66 14.9  
## New Hampshire 2.1 57 56 9.5  
## North Dakota 0.8 45 44 7.3  
## Rhode Island 3.4 174 87 8.3  
## South Dakota 3.8 86 45 12.8  
## Utah 3.2 120 80 22.9  
## Vermont 2.2 48 32 11.2  
## Washington 4.0 145 73 26.2  
## Wisconsin 2.6 53 66 10.8

row.names(USArrests[which(USArrests$Murder<=quartile1),])

## [1] "Connecticut" "Idaho" "Iowa" "Maine"   
## [5] "Minnesota" "New Hampshire" "North Dakota" "Rhode Island"   
## [9] "South Dakota" "Utah" "Vermont" "Washington"   
## [13] "Wisconsin"