Class Assignment 1

Nahid Ferdous

Problem 1

Use the TTU graduate student exit survey data.

```
grad <- read.csv("https://raw.githubusercontent.com/asheikhz2/TTU_Zadeh/main/pgs.csv")
```

a) Create a new data-frame for three variables: “GenRating”, “DeptStaff”, “Housing”.

```
mydata = grad[, c("GenRating", "DeptStaff", "Housing")]

head(mydata)
```

```
##   GenRating DeptStaff Housing
## 1         3         4       4
## 2         5         4       3
## 3         4         3       4
## 4         2         4       2
## 5         5         4      NA
## 6         3         2       4
```

b) There are some missing values in this data. Find a correlation matrix for the data of part (a). If there are NAs (missing values) in your data, estimate the correlation matrix by all three following methods: (1) Complete Case Analysis, (2) MLE, (3) Median insertion.

```
# Complete Case Analysis
mydata_na_omit <- na.omit(mydata)
corr_na_omit <- cor(mydata_na_omit)
print("Complete Case Analysis")
```

```
## [1] "Complete Case Analysis"
```

```
corr_na_omit
```

```
##           GenRating DeptStaff   Housing
## GenRating 1.0000000 0.3479520 0.2162194
## DeptStaff 0.3479520 1.0000000 0.1547242
## Housing   0.2162194 0.1547242 1.0000000
```

```r
# Maximum Likelihood Estimation
library(mvnmle)
mydata_MLE_fit <- mlest(mydata)
mydata_MLE_cov <- mydata_MLE_fit$sigmahat
mydata_MLE_corr <- cov2cor(mydata_MLE_cov)
cat("\n\n")
```

```r
print("Maximum Likelihood Estimation")
```

```
## [1] "Maximum Likelihood Estimation"
```

```r
mydata_MLE_corr
```

```
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.3548599 0.2169721
## [2,] 0.3548599 1.0000000 0.1556537
## [3,] 0.2169721 0.1556537 1.0000000
```

```r
# Median insertion

mydata_Median_insertion <- mydata

for (c in 1: ncol(mydata_Median_insertion)){
  NaN_bool <- is.na(mydata_Median_insertion[,c])
  NaN_index <- which(NaN_bool)
  mydata_Median_insertion[NaN_index,c] <- median(mydata_Median_insertion[,c], na.rm = TRUE)
}
cat("\n\n")
```

```r
print("Median insertion")
```

```
## [1] "Median insertion"
```

```r
mydata_Median_insertion_corr = cor(mydata_Median_insertion)
mydata_Median_insertion_corr
```

```
##           GenRating DeptStaff   Housing
## GenRating 1.0000000 0.3504238 0.2135113
## DeptStaff 0.3504238 1.0000000 0.1450470
## Housing   0.2135113 0.1450470 1.0000000
```

Problem 2

Read the crime data set.

```r
crime <- read.csv("https://raw.githubusercontent.com/asheikhz2/TTU_Zadeh/main/crime.csv",
row.names = "STATE")
head(crime)
```

```
##            MURDER RAPE ROBBERY ASSAULT BURGLARY LARCENY  AUTO
## ALABAMA      14.2 25.2    96.8   278.3   1135.5  1881.9 280.7
## ALASKA       10.8 51.6    96.8   284.0   1331.7  3369.8 753.3
## ARIZONA       9.5 34.2   138.2   312.3   2346.1  4467.4 439.5
## ARKANSAS      8.8 27.6    83.2   203.4    972.6  1862.1 183.4
## CALIFORNIA   11.5 49.4   287.0   358.0   2139.4  3499.8 663.5
## COLORADO      6.3 42.0   170.7   292.9   1935.2  3903.2 477.1
```

a) Find the correlation matrix of the data.

```r
crime_cor <- cor(crime)
crime_cor
```

```
##              MURDER      RAPE   ROBBERY   ASSAULT  BURGLARY   LARCENY
## MURDER   1.00000000 0.6012205 0.4837076 0.6485505 0.3858168 0.1019198
## RAPE     0.60122047 1.0000000 0.5918793 0.7402595 0.7121301 0.6139882
## ROBBERY  0.48370757 0.5918793 1.0000000 0.5570782 0.6372420 0.4467399
## ASSAULT  0.64855048 0.7402595 0.5570782 1.0000000 0.6229085 0.4043633
## BURGLARY 0.38581683 0.7121301 0.6372420 0.6229085 1.0000000 0.7921210
## LARCENY  0.10191983 0.6139882 0.4467399 0.4043633 0.7921210 1.0000000
## AUTO     0.06881448 0.3489015 0.5906795 0.2758426 0.5579533 0.4441799
##                AUTO
## MURDER   0.06881448
## RAPE     0.34890153
## ROBBERY  0.59067951
## ASSAULT  0.27584265
## BURGLARY 0.55795326
## LARCENY  0.44417992
## AUTO     1.00000000
```

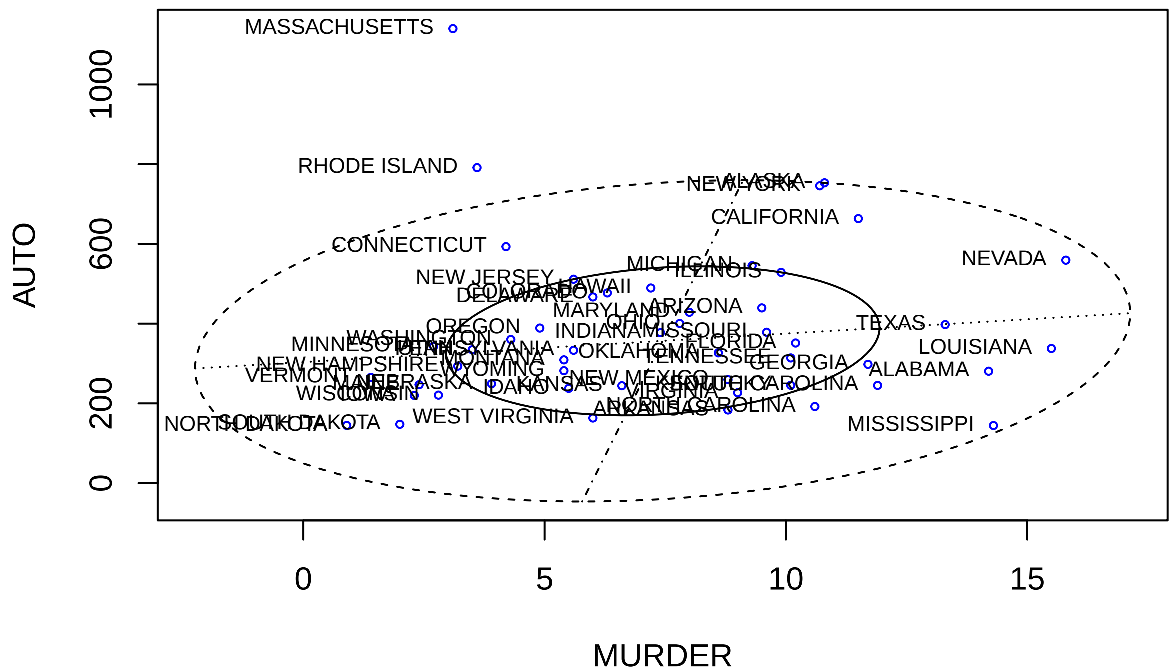
b) Create a bi-variate boxplot for “MURDER” vs “AUTO”.

```r
library(MVA)
```

```
## Loading required package: HSAUR2
```

```
## Loading required package: tools
```

```r
MURDER_AUTO_data= crime[,c("MURDER", "AUTO")]
bvbox(MURDER_AUTO_data, xlab = "MURDER", ylab = "AUTO", col = "blue", cex = 0.5)
text(MURDER_AUTO_data, labels = row.names(crime), cex = 0.7, pos = 2)
```

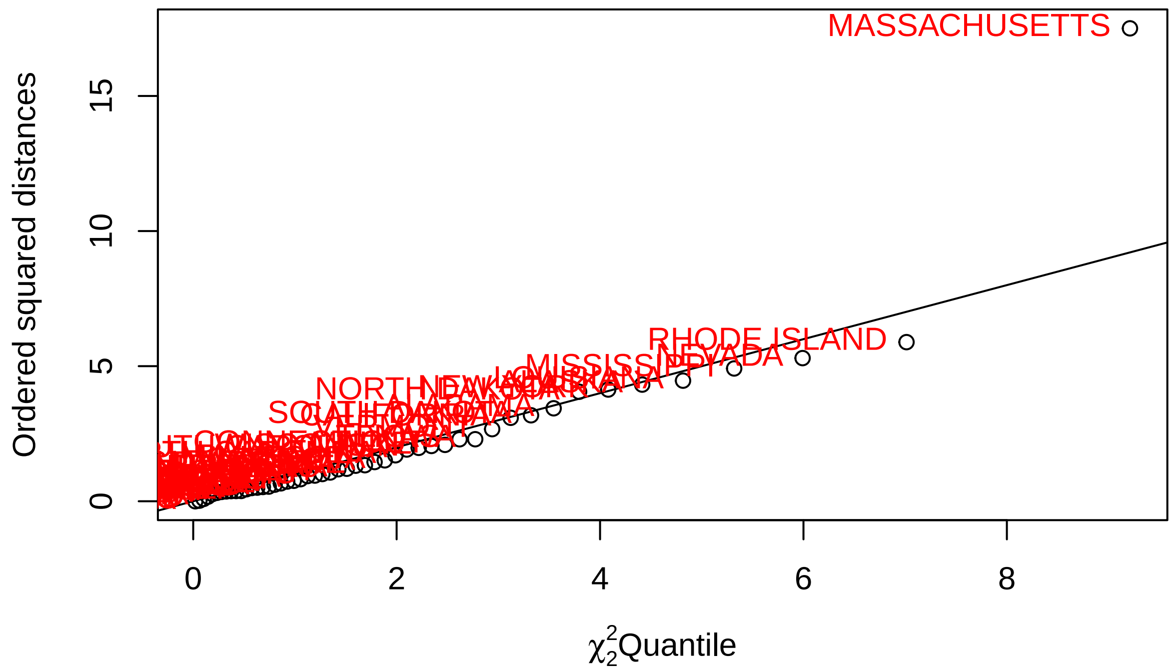


c) Create Chi-sq plot for mahalanobis distances (Lecture 8).

```
x <- MURDER_AUTO_data
xbar <- colMeans(x)
s <- cov(x)
# mahalanobis distances
d2 <- mahalanobis(x,xbar, s)
number_of_variables <- ncol(x)
position <- (1:nrow(x) -.5)/nrow(x)
quantiles <- qchisq(position, df = number_of_variables)

# chi_sq plot
plot(quantiles, sort(d2),
     xlab = expression(paste(chi[2]^2, "Quantile" )),
     ylab = "Ordered squared distances")
abline(a= 0, b =1)

text(quantiles, sort(d2), labels = names(sort(d2)), col = "red", pos = 2)
```



d) Given what you see in part b and c, identify outliers. Remove those outliers, then find the correlation matrix. Compare the correlation matrices before and after removing outliers.

```r
# In Part B, we identified two outliers: 'RHODE ISLAND' and 'MASSACHUSETTS'. However, in Part C, only '

print("Original correlation matrix")
```

```
## [1] "Original correlation matrix"
```

```r
crime_cor
```

```
##              MURDER      RAPE   ROBBERY   ASSAULT  BURGLARY   LARCENY
## MURDER   1.00000000 0.6012205 0.4837076 0.6485505 0.3858168 0.1019198
## RAPE     0.60122047 1.0000000 0.5918793 0.7402595 0.7121301 0.6139882
## ROBBERY  0.48370757 0.5918793 1.0000000 0.5570782 0.6372420 0.4467399
## ASSAULT  0.64855048 0.7402595 0.5570782 1.0000000 0.6229085 0.4043633
## BURGLARY 0.38581683 0.7121301 0.6372420 0.6229085 1.0000000 0.7921210
## LARCENY  0.10191983 0.6139882 0.4467399 0.4043633 0.7921210 1.0000000
## AUTO     0.06881448 0.3489015 0.5906795 0.2758426 0.5579533 0.4441799
##                AUTO
## MURDER   0.06881448
## RAPE     0.34890153
## ROBBERY  0.59067951
## ASSAULT  0.27584265
```

```
## BURGLARY 0.55795326
## LARCENY  0.44417992
## AUTO     1.00000000
```

```r
cat("\n\n")
```

```r
# first consider "RHODE ISLAND" and "MASSACHUSETTS" our outliers
two_outliers <- c("RHODE ISLAND", "MASSACHUSETTS")
two_outliers_indexs <- match(two_outliers, row.names(crime))
two_outliers_clean <- crime[-two_outliers_indexs,]
two_outliers_clean_corr <- cor(two_outliers_clean)
print("Two Outliers removed - RHODE ISLAND and  MASSACHUSETTS (correlation matrix)")
```

```
## [1] "Two Outliers removed - RHODE ISLAND and  MASSACHUSETTS (correlation matrix)"
```

```r
two_outliers_clean_corr
```

```
##              MURDER      RAPE   ROBBERY   ASSAULT  BURGLARY    LARCENY
## MURDER   1.00000000 0.5880901 0.5011673 0.6677069 0.4212009 0.09781218
## RAPE     0.58809015 1.0000000 0.6012715 0.7575550 0.7531904 0.63289373
## ROBBERY  0.50116735 0.6012715 1.0000000 0.5568803 0.6418124 0.45744886
## ASSAULT  0.66770692 0.7575550 0.5568803 1.0000000 0.6253039 0.40840666
## BURGLARY 0.42120085 0.7531904 0.6418124 0.6253039 1.0000000 0.80254054
## LARCENY  0.09781218 0.6328937 0.4574489 0.4084067 0.8025405 1.00000000
## AUTO     0.28162764 0.6119649 0.7535024 0.3484999 0.6515108 0.62833583
##               AUTO
## MURDER   0.2816276
## RAPE     0.6119649
## ROBBERY  0.7535024
## ASSAULT  0.3484999
## BURGLARY 0.6515108
## LARCENY  0.6283358
## AUTO     1.0000000
```

```r
cat("\n\n")
```

```r
# consider "MASSACHUSETTS" our outlier
one_outliers <- c("MASSACHUSETTS")
one_outliers_indexs <- match(one_outliers, row.names(crime))
one_outliers_clean <- crime[-one_outliers_indexs,]
one_outliers_clean_corr <- cor(one_outliers_clean)
print("One Outlier removed - MASSACHUSETTS (correlation matrix)")
```

```
## [1] "One Outlier removed - MASSACHUSETTS (correlation matrix)"
```

```r
one_outliers_clean_corr
```

```
##              MURDER      RAPE   ROBBERY   ASSAULT  BURGLARY    LARCENY
## MURDER   1.00000000 0.5997231 0.5036328 0.6623284 0.4054684 0.09176411
## RAPE     0.59972306 1.0000000 0.5996778 0.7441413 0.7213350 0.61216041
## ROBBERY  0.50363285 0.5996778 1.0000000 0.5566735 0.6351052 0.45439246
```

```
## ASSAULT  0.66232838 0.7441413 0.5566735 1.0000000 0.6228361 0.40767337
## BURGLARY 0.40546837 0.7213350 0.6351052 0.6228361 1.0000000 0.80249442
## LARCENY  0.09176411 0.6121604 0.4543925 0.4076734 0.8024944 1.00000000
## AUTO     0.19848837 0.4711101 0.6692457 0.3153578 0.6250350 0.59118744
##               AUTO
## MURDER   0.1984884
## RAPE     0.4711101
## ROBBERY  0.6692457
## ASSAULT  0.3153578
## BURGLARY 0.6250350
## LARCENY  0.5911874
## AUTO     1.0000000
```

```r
cat("\n\n")
```

```r
# Lets calculate Mean squared error
print("Original vs two outliers clean data ")
```

```
## [1] "Original vs two outliers clean data "
```

```r
MSE_two_outliers_clean <- mean((two_outliers_clean_corr-crime_cor)^2)
print("MSE- Original vs Two_outliers_clean data")
```

```
## [1] "MSE- Original vs Two_outliers_clean data"
```

```r
MSE_two_outliers_clean
```

```
## [1] 0.007908516
```

```r
cat("\n\n")
```

```r
print("Original vs one outlier clean data ")
```

```
## [1] "Original vs one outlier clean data "
```

```r
MSE_one_outliers_clean <- mean((one_outliers_clean_corr-crime_cor)^2)
print("MSE- Original vs One_outliers_clean data")
```

```
## [1] "MSE- Original vs One_outliers_clean data"
```

```r
MSE_one_outliers_clean
```

```
## [1] 0.0027355
```

Removing only 'MASSACHUSETTS' as an outlier results in a significantly lower MSE, indicating a better fit for the data compared to the model that also excludes 'RHODE ISLAND.' Therefore, 'RHODE ISLAND' may not be a true statistical outlier.