# Problemset 1

Bargetto Cristina 885847, Iavarone Marika 886338, Scanu Anna 1012903

## Exercise 1

Consider the air pollution data which consists of 7 measurements recorded at n = 41 cities in the United States. The variables are:

- SO2: Sulphur dioxide content in micrograms per cubic meter
- Neg.Temp: Average annual temperature in Fo (negative values)
- Manuf: Number of manufacturing enterprises employing 20 or more workers
- Pop: Population size (1970 census) in thousands
- Wind: Average annual wind speed in miles per hour
- Precip: Average annual precipitation in inches
- Days: Average number of days with precipitation per year

Table 1: Table continues below

| SO2 | Neg.Temp | Manuf | Pop |
|---|---|---|---|
| Min. : 8.00 | Min. :-75.50 | Min. : 35.0 | Min. : 71.0 |
| 1st Qu.: 13.00 | 1st Qu.:-59.30 | 1st Qu.: 181.0 | 1st Qu.: 299.0 |
| Median : 26.00 | Median :-54.60 | Median : 347.0 | Median : 515.0 |
| Mean : 30.05 | Mean :-55.76 | Mean : 463.1 | Mean : 608.6 |
| 3rd Qu.: 35.00 | 3rd Qu.:-50.60 | 3rd Qu.: 462.0 | 3rd Qu.: 717.0 |
| Max. :110.00 | Max. :-43.50 | Max. :3344.0 | Max. :3369.0 |

| Wind | Precip | Days |
|---|---|---|
| Min. : 6.000 | Min. : 7.05 | Min. : 36.0 |
| 1st Qu.: 8.700 | 1st Qu.:30.96 | 1st Qu.:103.0 |
| Median : 9.300 | Median :38.74 | Median :115.0 |
| Mean : 9.444 | Mean :36.77 | Mean :113.9 |
| 3rd Qu.:10.600 | 3rd Qu.:43.11 | 3rd Qu.:128.0 |
| Max. :12.700 | Max. :59.80 | Max. :166.0 |

We will ignore the S02 variable studying only the remaining six.

### 1.1

First we compute the sample mean of the variables.

| Neg.Temp | Manuf | Pop | Wind | Precip | Days |
|---|---|---|---|---|---|
| -55.76 | 463.1 | 608.6 | 9.444 | 36.77 | 113.9 |

The sample covariance matrix is

| | Neg.Temp | Manuf | Pop | Wind | Precip | Days |
|---|---|---|---|---|---|---|
| **Neg.Temp** | 52.24 | 774 | 262.4 | 3.611 | -32.86 | 82.43 |
| **Manuf** | 774 | 317503 | 311719 | 191.5 | -215 | 1969 |
| **Pop** | 262.4 | 311719 | 335372 | 175.9 | -178.1 | 646 |
| **Wind** | 3.611 | 191.5 | 175.9 | 2.041 | -0.219 | 6.214 |
| **Precip** | -32.86 | -215 | -178.1 | -0.219 | 138.6 | 154.8 |
| **Days** | 82.43 | 1969 | 646 | 6.214 | 154.8 | 702.6 |

Then we compute the sample correlation matrix **R**.

| | Neg.Temp | Manuf | Pop | Wind | Precip | Days |
|---|---|---|---|---|---|---|
| **Neg.Temp** | 1 | 0.19 | 0.063 | 0.35 | -0.386 | 0.43 |
| **Manuf** | 0.19 | 1 | 0.955 | 0.238 | -0.032 | 0.132 |
| **Pop** | 0.063 | 0.955 | 1 | 0.213 | -0.026 | 0.042 |
| **Wind** | 0.35 | 0.238 | 0.213 | 1 | -0.013 | 0.164 |
| **Precip** | -0.386 | -0.032 | -0.026 | -0.013 | 1 | 0.496 |
| **Days** | 0.43 | 0.132 | 0.042 | 0.164 | 0.496 | 1 |

We can see that Manuf and Pop are highly correlated since the correlation coefficient is larger than 0.5.

In fact sorting in decreasing order the correlation coefficients we see that the only value above 0.5 is the correlation between Manuf and Pop (2nd and 3rd variables).

```
##  [1]  0.95526935  0.49609671  0.43024212  0.34973963  0.23794683  0.21264375
##  [7]  0.19004216  0.16410559  0.13182930  0.06267813  0.04208319 -0.01299438
## [13] -0.02611873 -0.03241688 -0.38625342
```

If you need to find which is the position of the highest correlation you can use

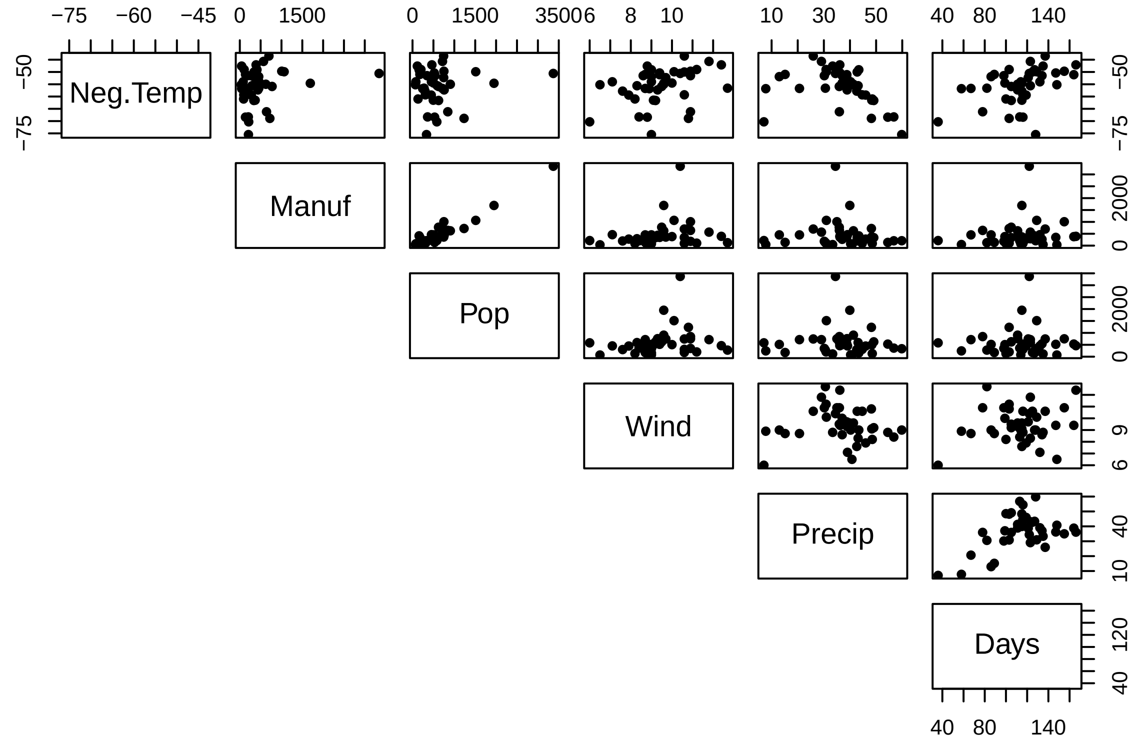
```
order.cor<-order(R,decreasing =T)[1:15]
order.cor
```

```
##  [1]  9 30  6  4 10 16  2 24 12  3 18 23 17 11  5
```

where these positions are associated to the matrix as follows

```
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    7   13   19   25   31
## [2,]    2    8   14   20   26   32
## [3,]    3    9   15   21   27   33
## [4,]    4   10   16   22   28   34
## [5,]    5   11   17   23   29   35
## [6,]    6   12   18   24   30   36
```

Plotting the correlation between two variables we see that for example Pop and Manuf are highly correlated since the plot seems to form an increasig line. On the other hand among the pairs of negatively correlated variables, Neg.Temp and Precip seems to form a decreasing line, in fact the correlation coefficient is -0.39 which is the lowest one.



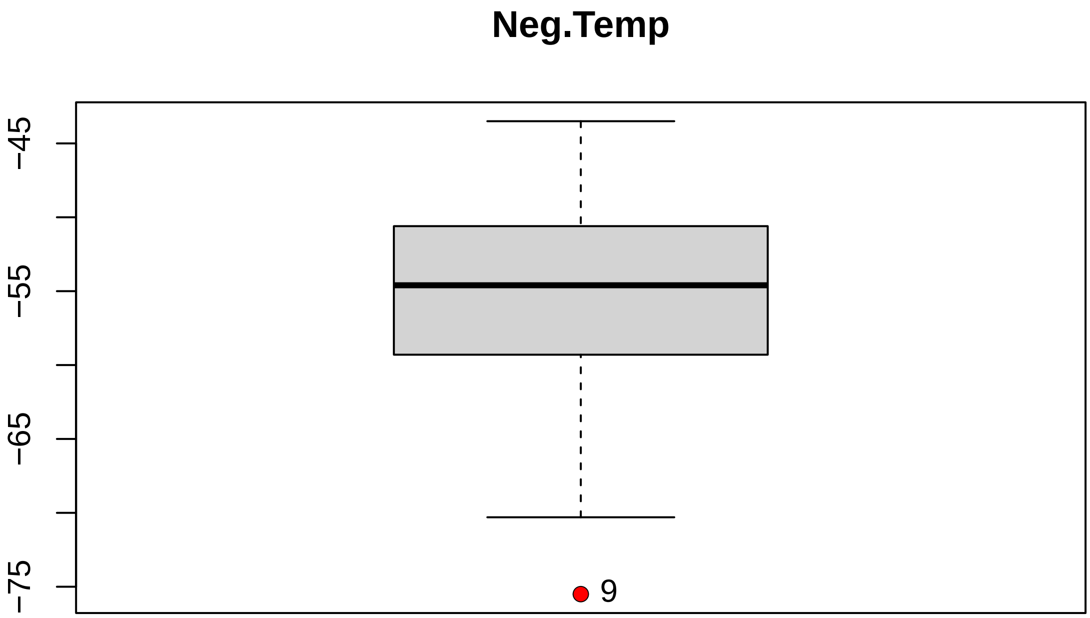

If we also want to see the correlation between variables in decreasing order we have

```
##       row      col
##  [1,] "Pop"    "Manuf"
##  [2,] "Days"   "Precip"
##  [3,] "Days"   "Neg.Temp"
##  [4,] "Wind"   "Neg.Temp"
##  [5,] "Wind"   "Manuf"
##  [6,] "Wind"   "Pop"
##  [7,] "Manuf"  "Neg.Temp"
##  [8,] "Days"   "Wind"
##  [9,] "Days"   "Manuf"
## [10,] "Pop"    "Neg.Temp"
## [11,] "Days"   "Pop"
## [12,] "Precip" "Wind"
## [13,] "Precip" "Pop"
## [14,] "Precip" "Manuf"
## [15,] "Precip" "Neg.Temp"
```

The results of this analysis does not surprise us, since cities with an high population have also many manufacturing enterprises. Also for Days and Precip it makes sense having a correlation coefficient of almost 0.5.

### 1.2

In order to identify the outliers we make use of boxplot representation of each variable:

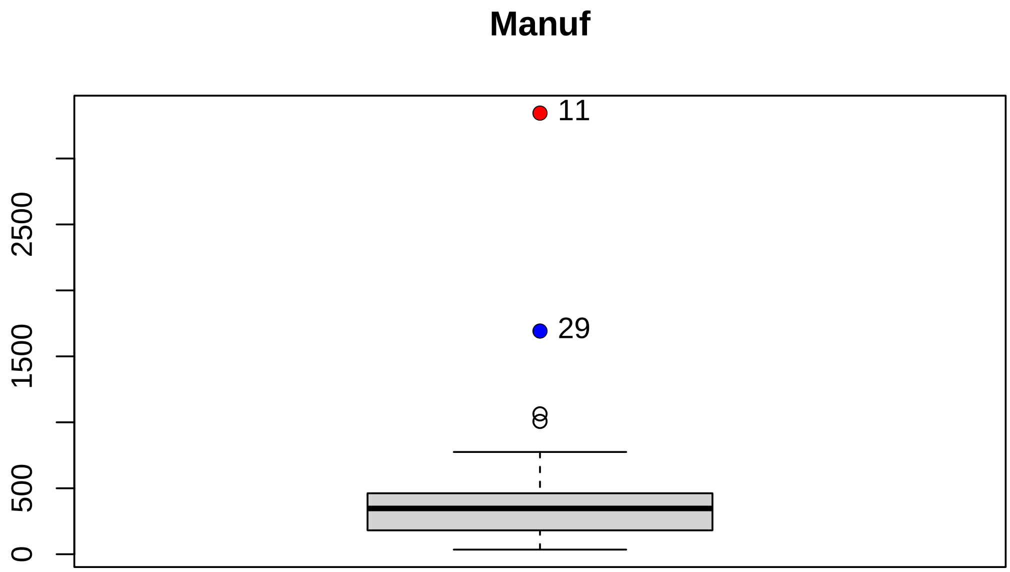



Using the following commands we get that the value of the outlier with respect to Neg.Temp. is -75.5 and it corresponds to the observation 9.

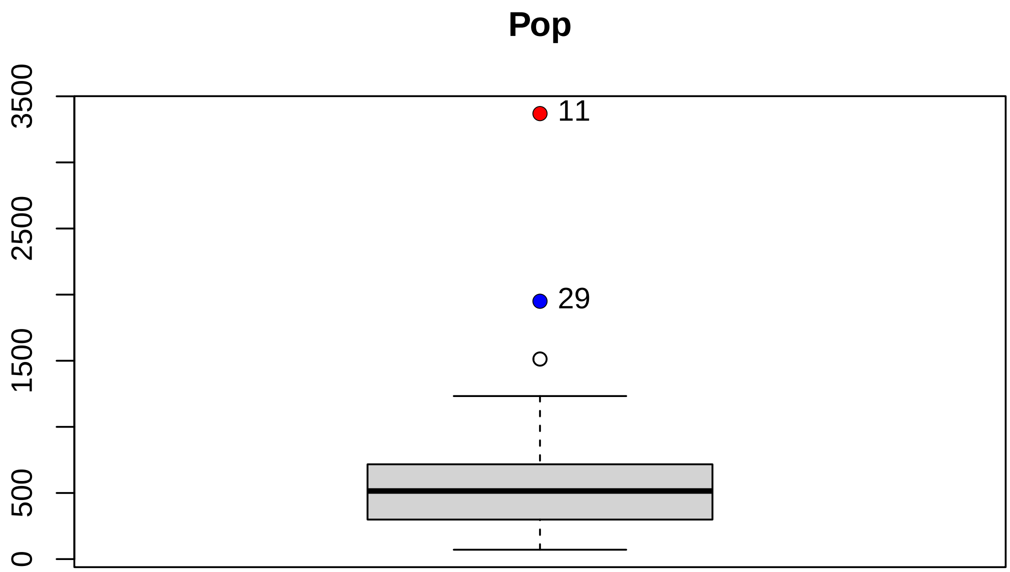
```
boxplot(usair$Neg.Temp, main="Neg.Temp")
boxplot(usair$Neg.Temp)$out
which(usair$Neg.Temp==-75.5) # obs.9
```

In the same way as we did for Neg.Temp. we find the values of the outliers with respect to the remaining variables and their corresponding observations.

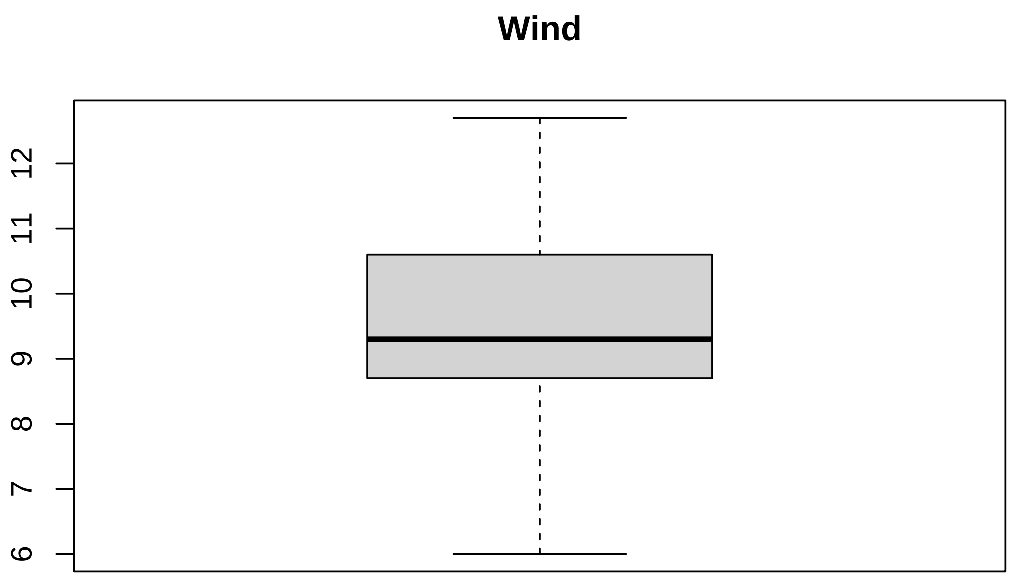
**Note:** When we will find more than two outliers we will choose the ones that are the farthest from the mean value.



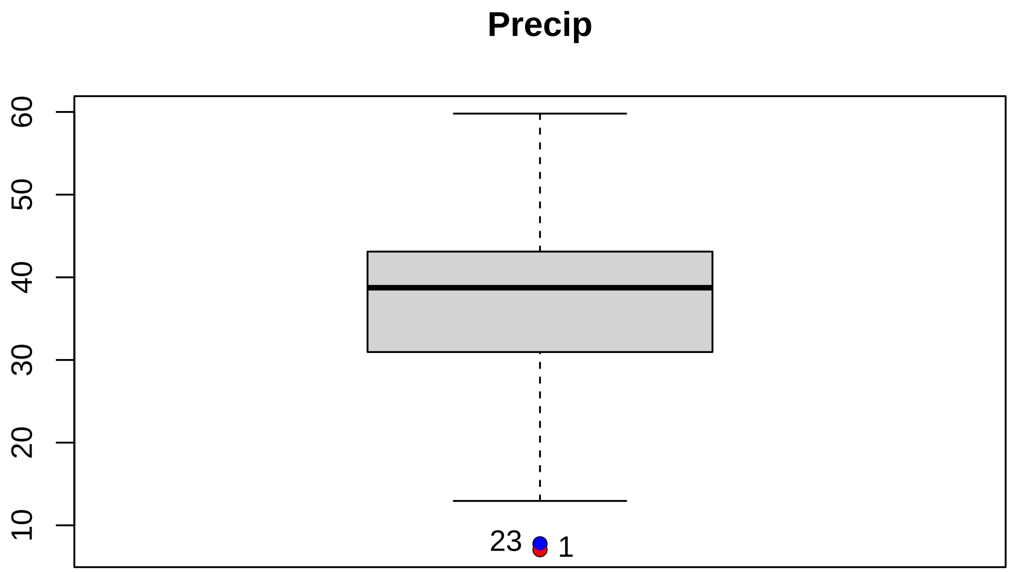

The values of the outliers with respect to Manuf are 3344 and 1629 and they correspond to observations 11 and 29.



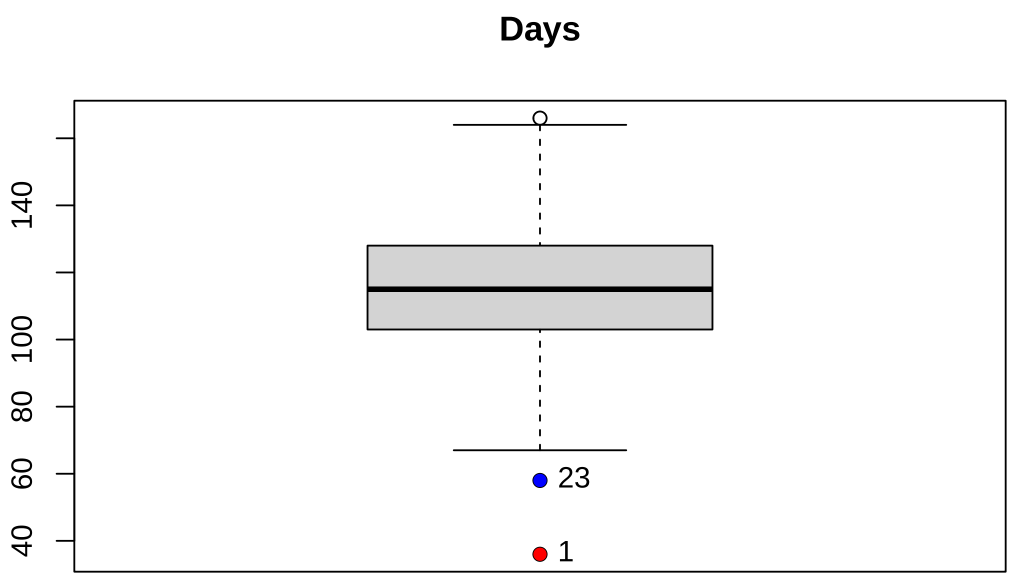

The values of the outliers with respect to Pop are 3369 and 1950 and they correspond to observations 11 and 29.




For Wind we don't have any outliers.




The values of the outliers with respect to Precip are 7.05 and 7.77 and they correspond to observations 1 and 23.

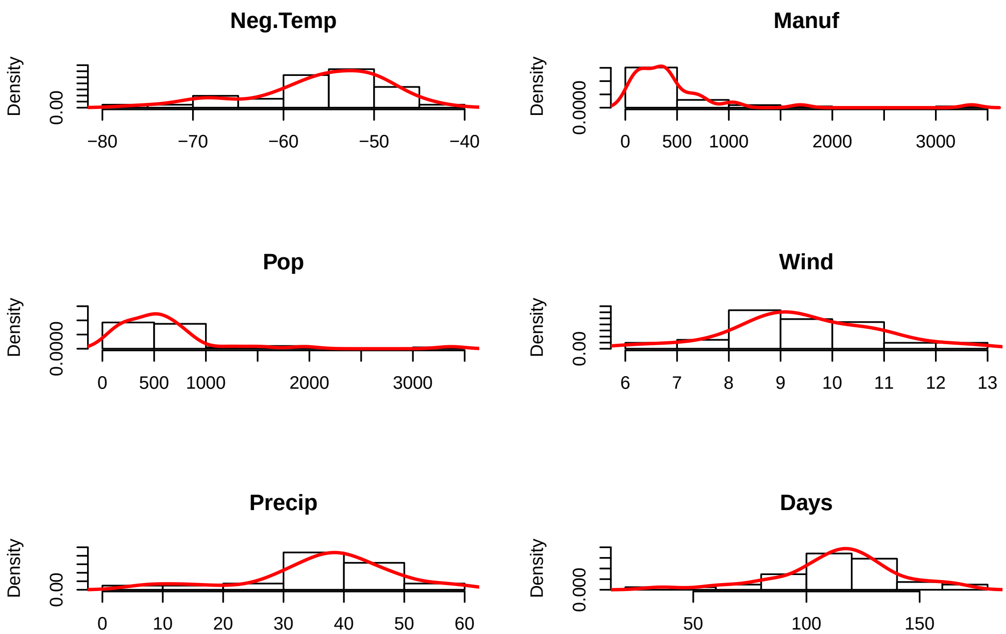



The values of the outliers with respect to Days are 36 and 58 and they correspond to observations 1 and 23.

### 1.3

To check normality of the sample for each variable, we first plot the density.

```
par(mfrow=c(3,2))
limits<-list(c(0,0.07), c(0,0.0016), c(0,0.0015), c(0,0.35),c(0,0.05), c(0,0.02))
for (j in 1:6)
{
  x<-usair[,j]
  hist(x,probability = T, col="white", ylim=limits[[j]], xlab = "", main = names(usair)[j])
  lines(density(x), lwd=2, col="red")
}
```

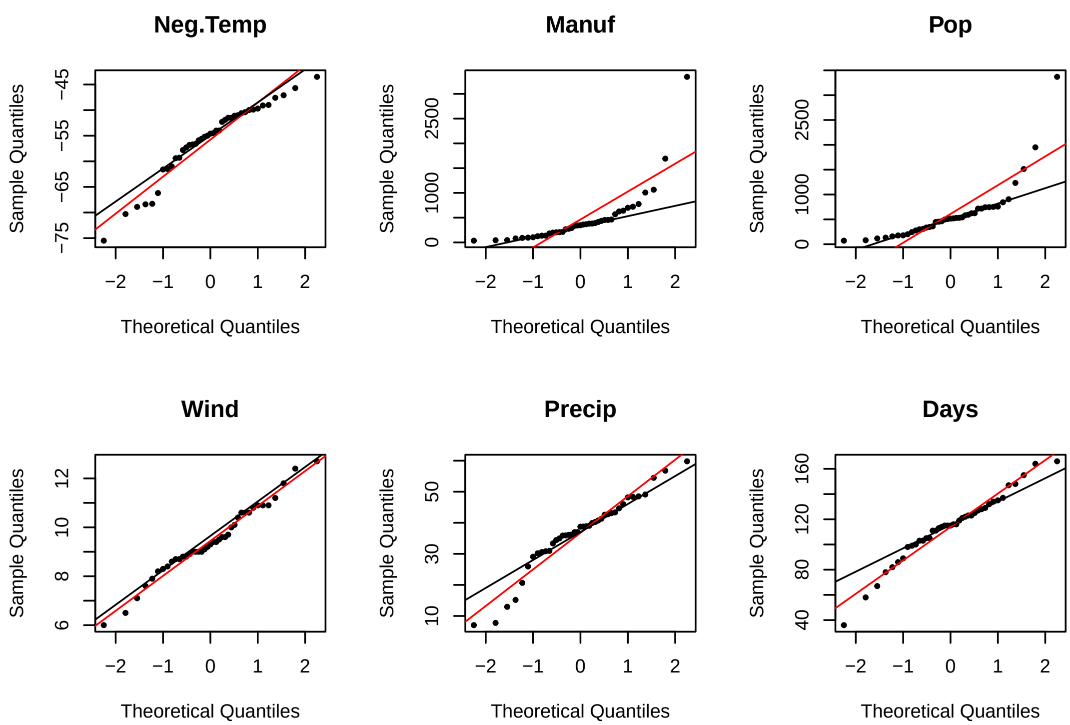



From the plot it seems that they all have a bell shape and we can notice that the most correlated variables have also similar density plots (e.g. Manuf and Pop, Days and Precip).

Next, we construct the normal Q-Q plot for each variable.

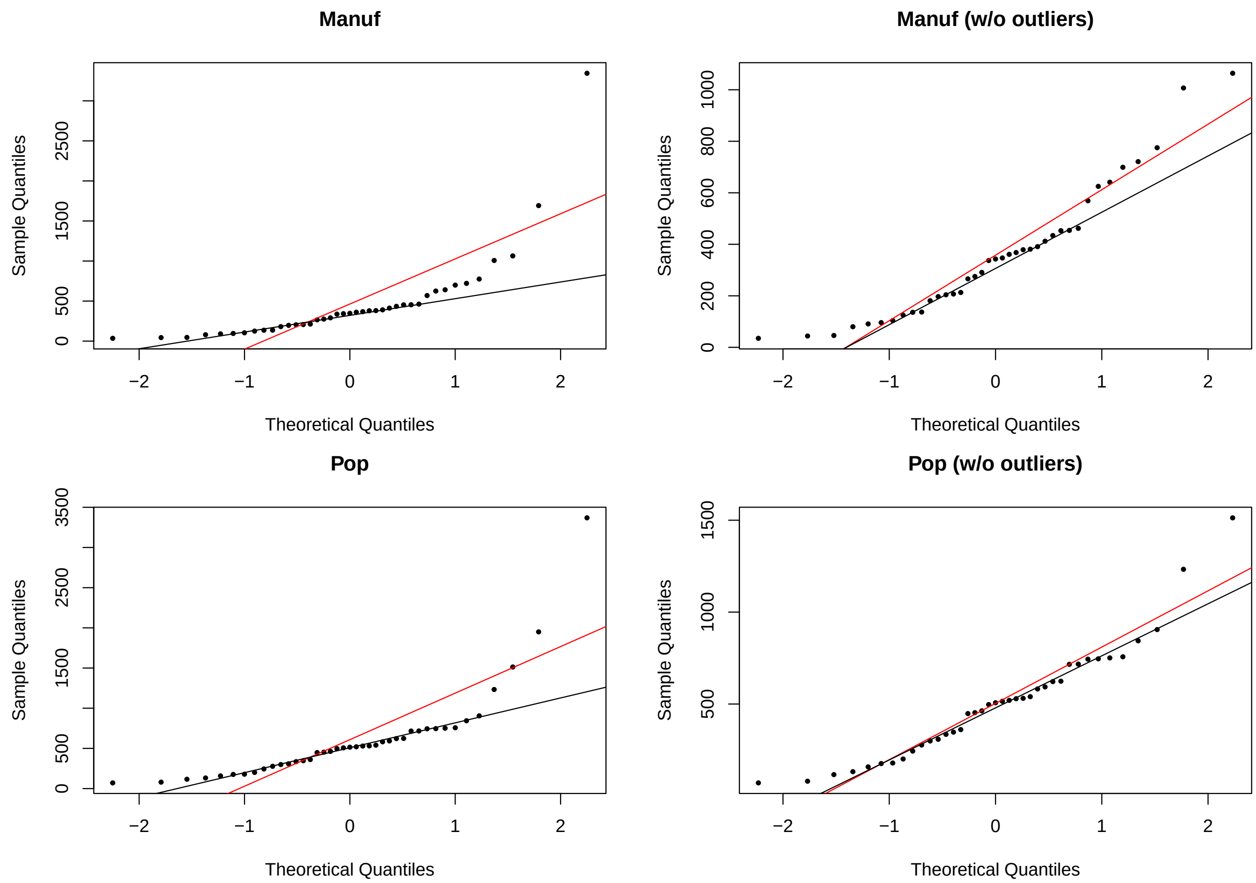
We plot the theoretical quantile of the standard normal distribution against the sample quantiles (i.e. the red line, where the intercept corresponds to the sample mean and the slope corresponds to the standard deviation).

**Note**: Since we want to see if normality is satisfied we have to check whether empirical quantiles are equal to the theoretical ones. To do so we compare the line formed by the points (i.e.black line) and the red one. Indeed, if normality is satisfied the two lines should be equal.



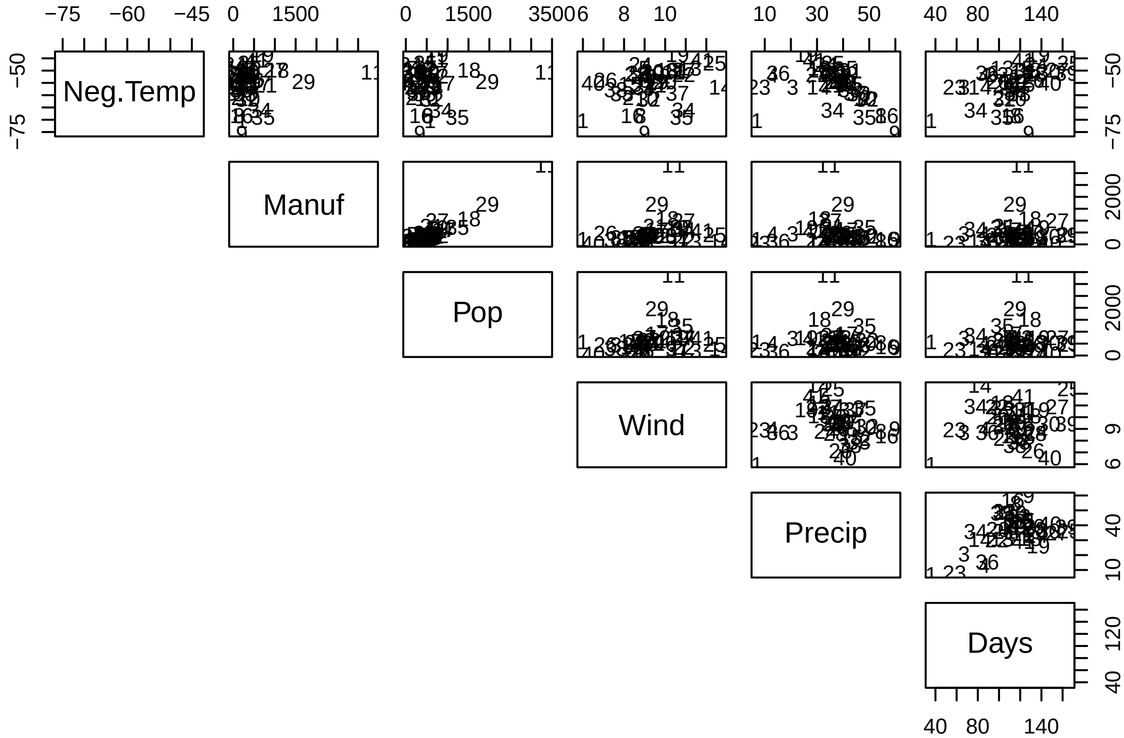
For example from these plot we can see that red and the black lines in Wind and Neg. Temp are very close, in fact their density is similar to a Gaussian. On the countrary the lines in Manuf and Pop tend to step away from each other. For what concerns the variable Days, according to the previous density plot, it seems the most bell shaped variable but from the qq plot the two lines look distant. This can be explained by the fact that the black line is an approximation of the position of the points in the graph but in this case it does not represent well the points, since if we compare directly the points and the red line, the variable Days seems to be distributed as a Gaussian.

**Note**: In Manuf and Pop the red line is considerably influenced by the most isolated points. So we expect that, leaving out the outliers identified in the previous point, the Q-Q plot shows a more Gaussian trend.



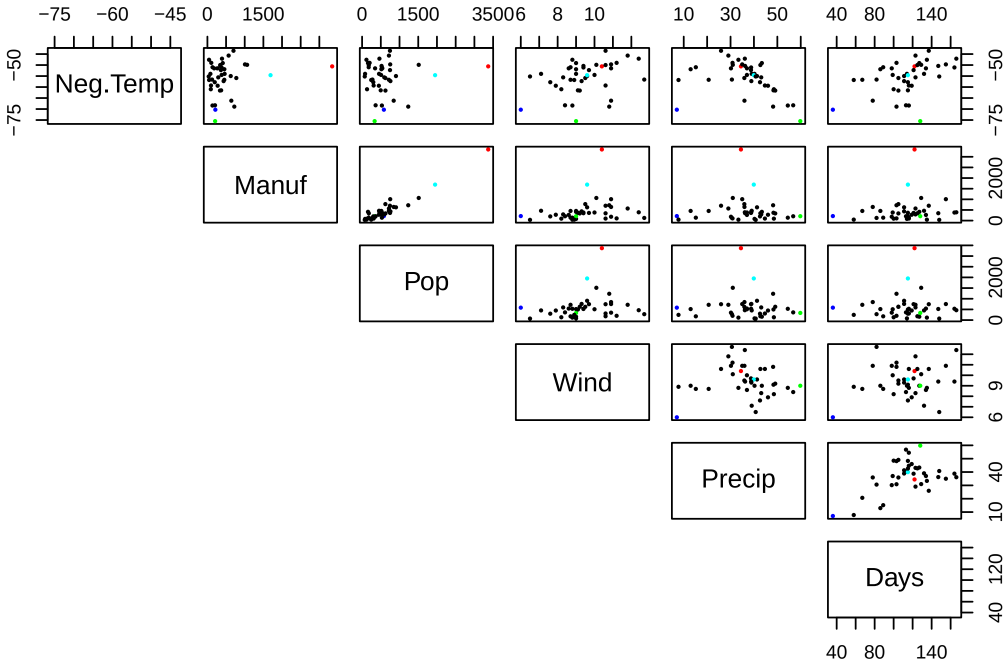
## 1.4

Let's now make the scatter plots to find the indeces of the observations that might be considered unusual values.




We associate different colors to the observations we have visualized.

```
col.index<-rep("black",41)
col.index[11]<-"red"; col.index[1]<-"blue" ; col.index[9]<-"green"
col.index[29]<- "cyan"

pairs(usair,cex=1/2, pch=16, col=col.index,lower.panel=NULL)
```




As a result we have that the observations 1, 11, 29, 9 might be outliers, as we saw in (1.2). On the other hand observation 23 can not be detected from scatter plots although we have identified it as unusual value for the variables Precip and Days.

## 1.5

In the following we will construct the chi-square Q-Q plot of the squared Mahalanobis distances (which we can summarize as follows)

Table 6: Table continues below

| Phoenix | Little Rock | San Francisco | Denver | Hartford | Wilmington |
|---|---|---|---|---|---|
| 19.21 | 4.86 | 4.14 | 4.65 | 7.75 | 2.04 |

Table 7: Table continues below

| Washington | Jacksonville | Miami | Atlanta | Chicago | Indianapolis |
|---|---|---|---|---|---|
| 1.28 | 5.01 | 14.26 | 1.36 | 26.89 | 4.09 |

Table 8: Table continues below

| Des Moines | Wichita | Louisville | New Orleans | Baltimore | Detroit |
|---|---|---|---|---|---|
| 4.09 | 9.06 | 2.98 | 4.65 | 2.16 | 7.22 |

Table 9: Table continues below

| Minneapolis-St. Paul | Kansas City | St. Louis | Omaha | Alburquerque | Albany |
|---|---|---|---|---|---|
| 3.51 | 1.75 | 4.53 | 2.84 | 7.91 | 2.97 |

Table 10: Table continues below

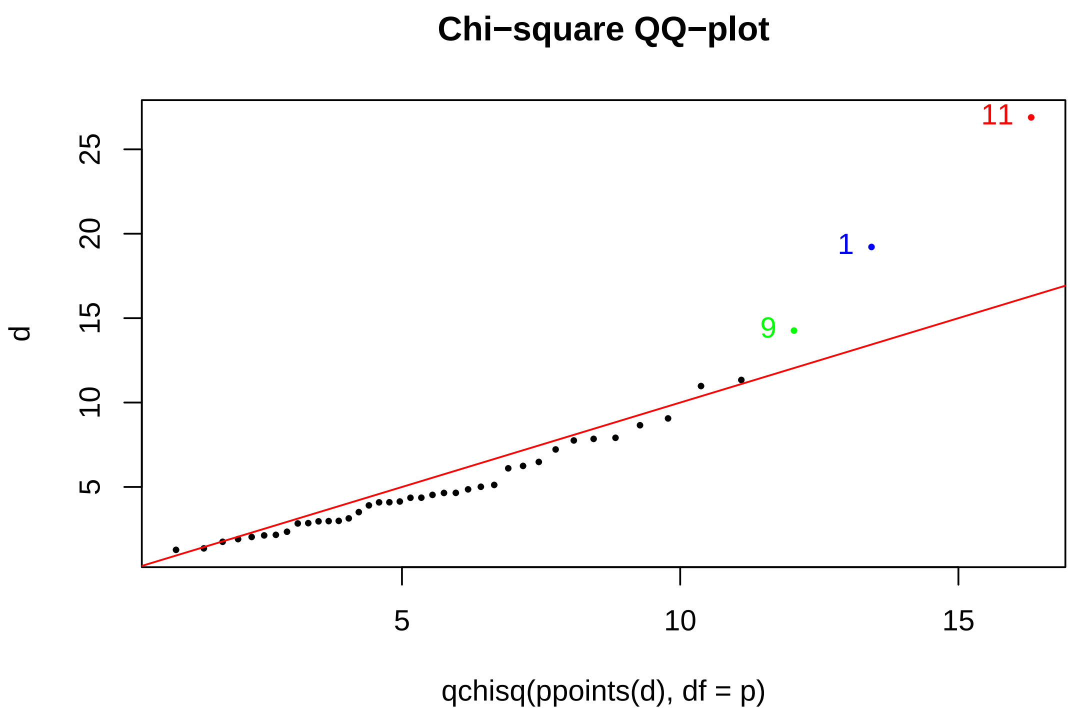
| Buffalo | Cincinnati | Cleveland | Columbus | Philadelphia | Pittsburgh |
|---|---|---|---|---|---|
| 10.98 | 4.36 | 11.34 | 3.14 | 6.25 | 2.99 |

Table 11: Table continues below

| Providence | Memphis | Nashville | Dallas | Houston | Salt Lake City | Norfolk |
|---|---|---|---|---|---|---|
| 5.12 | 2.86 | 1.91 | 6.1 | 8.66 | 4.36 | 2.13 |

| Richmond | Seattle | Charleston | Milwaukee |
|---|---|---|---|
| 2.35 | 6.48 | 7.85 | 3.91 |

The chi-squared Q-Q plot of squared Mahalanobis distance in general can be used to show whether the observations may have a multivariate normal distribution.




From the plot we see that the Mahalanobis distances have an approximate Chi-square distribution, since the points form a straight line close to the red one. Hence the observed data have multivariate normal distribution.

Furthermore on the upper right side of the plot, we have the points with Mahalanobis distance grater than the Chi-square quantile value. They correspond to the observations we identified as outliers.

## 1.6

As far as it concerns multivariate outliers, first we sort the observations in decreasing order with respect to their Mahalonibis distance.

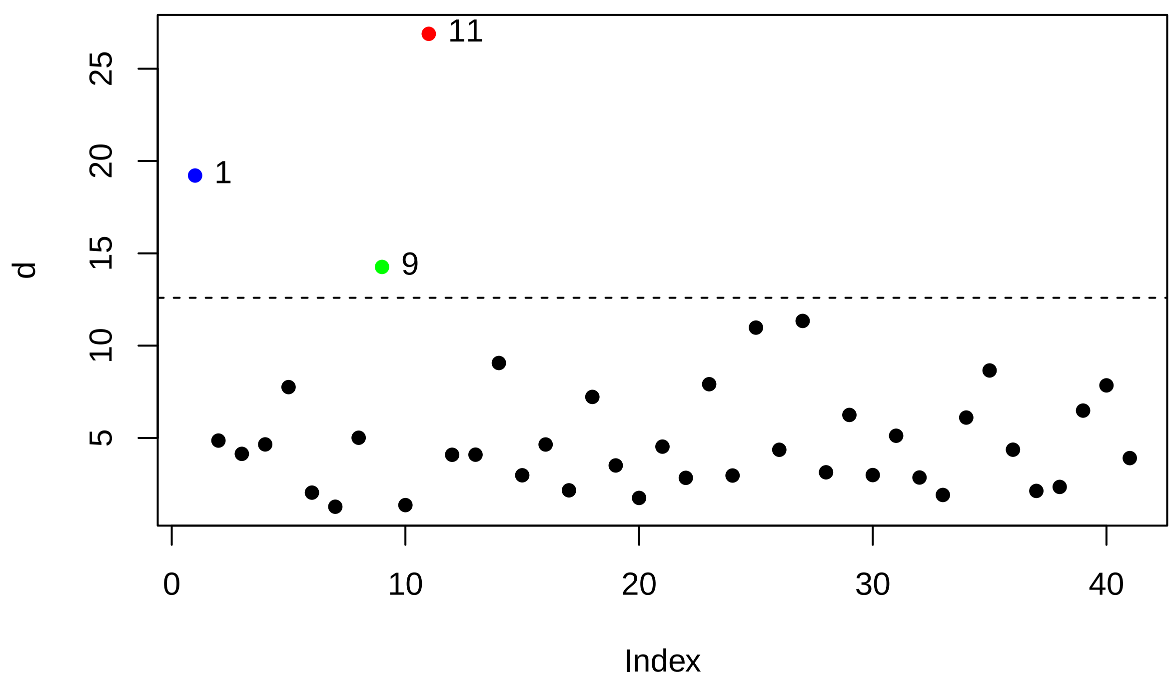
```
p<-dim(usair)[2]
d<-mahalanobis(X,center=bar.x,cov=S)
round(d,2)
order(d,decreasing=T)
```

```
##  [1] 11  1  9 27 25 14 35 23 40  5 18 39 29 34 31  8  2  4 16 21 36 26  3 13 12
## [26] 41 19 28 30 15 24 32 22 38 17 37  6 33 20 10  7
```

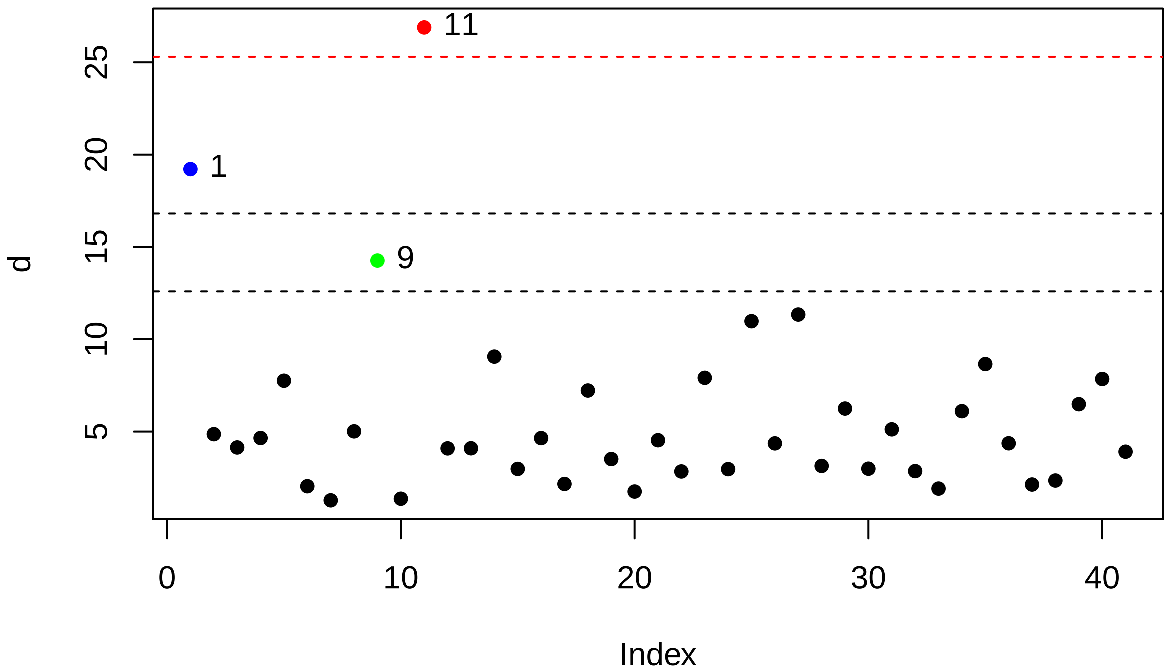
So observations 11, 1, 9 could be considered as possible multivariate outliers. These three observations correspond to the univariate outliers found in (1.2).

On the other hand, the observation 23, that we have consider as univariate outlier, has a small Mahalnobis distance with respect to other observations. So we can not consider it as a multivariate outlier. In fact as we saw in point (1.4), it couldn't be detected from the scatter plots.

According to the upper quantile 0.95, we obtain that all three observations could be possible multivariate outliers.




However 0.95 is not significantly accurate, hence we can try with other arbitrary values. With a level of significance of 0.99 we can consider the observation 1 and observation 11 as possible multivariate outliers but with a level of significance of 0.9997 we have that only the observation 11 is left out,hence we are confident that it is a multivariate outlier.


11
1
9
d
5
10
15
20
25
0
10
20
30
40
Index

### Exercise 2

Consider $X = (X_1, X_2, X_3)$ distributed as $N_3(\mu, \mathbf{\Sigma})$ with

$\mu = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$ and $\mathbf{\Sigma}= \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix}$

### 2.1

We want to find $\rho$ such that the first two population principal components of X, PC1 and PC2, account for more than 80% of the total variation of X.

In order to do that we first need to compute the determinant of $\mathbf{\Sigma} - \lambda I$:

$det(\ \mathbf{\Sigma} - \lambda I) = (1-\lambda)((1-\lambda)^2 - 2\rho^2)$

imposing it equal to 0 and solving the equation with respect to $\lambda$, we get the eigenvalues:

$\lambda_1 = 1 + \sqrt{2}\rho$

$\lambda_2 = 1$

$\lambda_3 = 1 - \sqrt{2}\rho$

By assumption we have that $-\sqrt{2}/2 < \rho < \sqrt{2}/2$ and we want to consider only PC1 and PC2. So if we consider $\rho > 0$ (i.e. $0 < \rho < \sqrt{2}/2$) we have that $\lambda_1 > \lambda_2 > \lambda_3$ and $\lambda_j$ corresponds to the variance of the $j^{th}$ principal component. Hence, sice we need the proportion of variance of the first two principal components to be grater than 80%, we have to solve the following inequality with respect to $\rho$:

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} > 0.8$$

and so

$$\frac{2 + \sqrt{2}\rho}{3} > \frac{4}{5}$$

which is equivalent to (considering the assumption on $\rho$)

$$\sqrt{2}/5 < \rho < \sqrt{2}/2$$

On the other hand, if we consider $\rho < 0$ (i.e. $-\sqrt{2}/2 < \rho < 0$) we have that $\lambda_1 > \lambda_2 > \lambda_3$ if and only if we impose

$\lambda_1 = 1 - \sqrt{2}\rho$

$\lambda_2 = 1$

$\lambda_3 = 1 + \sqrt{2}\rho$

and in this case we get:

$$-\sqrt{2}/2 < \rho < -\sqrt{2}/5$$

So we get the range

$$\sqrt{2}/5 < |\rho| < \sqrt{2}/2$$

## 2.2

In order to give an interpretation to PC1 and PC2 in terms of the original variables we need to compute the unit norm eigenvectors $e_j$ of $\mathbf{\Sigma}$, since $a_j = (a_{j1}, a_{j2}, a_{j3})$ are equivalent to the eigenvectors $e_j$ for $j = 1, 2, 3$.

Consider the case $\rho > 0$.

We want to solve the system $\mathbf{\Sigma}\ e_j = \lambda_j e_j$ for $j = 1, 2, 3$ where

$\lambda_1 = 1 + \sqrt{2}\rho$

$\lambda_2 = 1$

$\lambda_3 = 1 - \sqrt{2}\rho$

and considering the fact that $e_j$'s have to be unit norm vectors we obtain

$$e_1 = \begin{pmatrix} 1/2 \\ \sqrt{2}/2 \\ 1/2 \end{pmatrix},\ e_2 = \begin{pmatrix} \sqrt{2}/2 \\ 0 \\ -\sqrt{2}/2 \end{pmatrix},\ e_3 = \begin{pmatrix} 1/2 \\ -\sqrt{2}/2 \\ 1/2 \end{pmatrix}$$

so we have found the linear combination of the variables that gives us the principal components, since the loadings coincide with the components of the eigenvectors

$$Z_1 = a_1^t X = \frac{1}{2}X_1 + \frac{\sqrt{2}}{2}X_2 + \frac{1}{2}X_3$$

$$Z_2 = a_2^t X = \frac{\sqrt{2}}{2}X_1 - \frac{\sqrt{2}}{2}X_3$$

We have that the correlation between the the $j^{th}$ principal component and the $k^{th}$ variable is defined as

$$r_{Z_j X_k} = \frac{a_{jk}\sqrt{\lambda_j}}{\sqrt{s_{kk}}}$$

and it measures the univariate contribution of an individual $X_k$ to a principal component $Z_j$. But in order to analize the relation between $Z_j$ and all the variables we must consider the loadings $a_{jk}$.

We have that the variable which mostly contributes to $Z_1$ is $X_2$ since $|\sqrt{2}/2| > |1/2|$. While for $Z_2$ we have that $X_1$ and $X_3$ equally contribute and $X_2$ has not contribution to $Z_2$.

To interpret the principal components we have to take the signs of the loadings into account:

- in $Z_1$ all the loadings are positive;
- in $Z_2$ we have a contrast between $X_1$ and $X_3$ since their loadings have opposite sign.

On the other hand, if we consider $\rho < 0$ we have the eigenvalues $\lambda_1 > \lambda_2 > \lambda_3$

$\lambda_1 = 1 - \sqrt{2}\rho$

$\lambda_2 = 1$

$\lambda_3 = 1 + \sqrt{2}\rho$

and the eigenvectors are

$$e_1 = \begin{pmatrix} 1/2 \\ -\sqrt{2}/2 \\ 1/2 \end{pmatrix},\ e_2 = \begin{pmatrix} \sqrt{2}/2 \\ 0 \\ -\sqrt{2}/2 \end{pmatrix},\ e_3 = \begin{pmatrix} 1/2 \\ \sqrt{2}/2 \\ 1/2 \end{pmatrix}$$

so the first two principal components are

$$Z_1 = e_1^t X = \frac{1}{2}X_1 - \frac{\sqrt{2}}{2}X_2 + \frac{1}{2}X_3$$

$$Z_2 = e_2^t X = \frac{\sqrt{2}}{2}X_1 - \frac{\sqrt{2}}{2}X_3$$

We have that the contribution of the variables to $Z_1$ and $Z_2$ is the same as for $\rho > 0$, since the absolute values of the loadings are the same.

However we have a different interpretation of $Z_1$, since we have to take the signs of the loadings into account. In this case there is a contrast between the variable $X_2$ and the variables $X_1$ and $X_3$.

## 2.3

Consider

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} X_1 - X_2 \\ X_2 - X_3 \end{pmatrix}$$

We have that

$$Z = AX = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

So $Z$ is a Gaussian and in order to know its distribution we have to compute $\mu_Z$, the mean of $Z$, and its covariance matrix $\mathbf{\Sigma}_Z$.

$$\mu_Z = A\mu = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix} = \begin{pmatrix} 2 \\ -3 \end{pmatrix}$$

$$\mathbf{\Sigma}_Z = A\ \mathbf{\Sigma} A^t = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 2-2\rho & 2\rho-1 \\ 2\rho-1 & 2-2\rho \end{pmatrix}$$

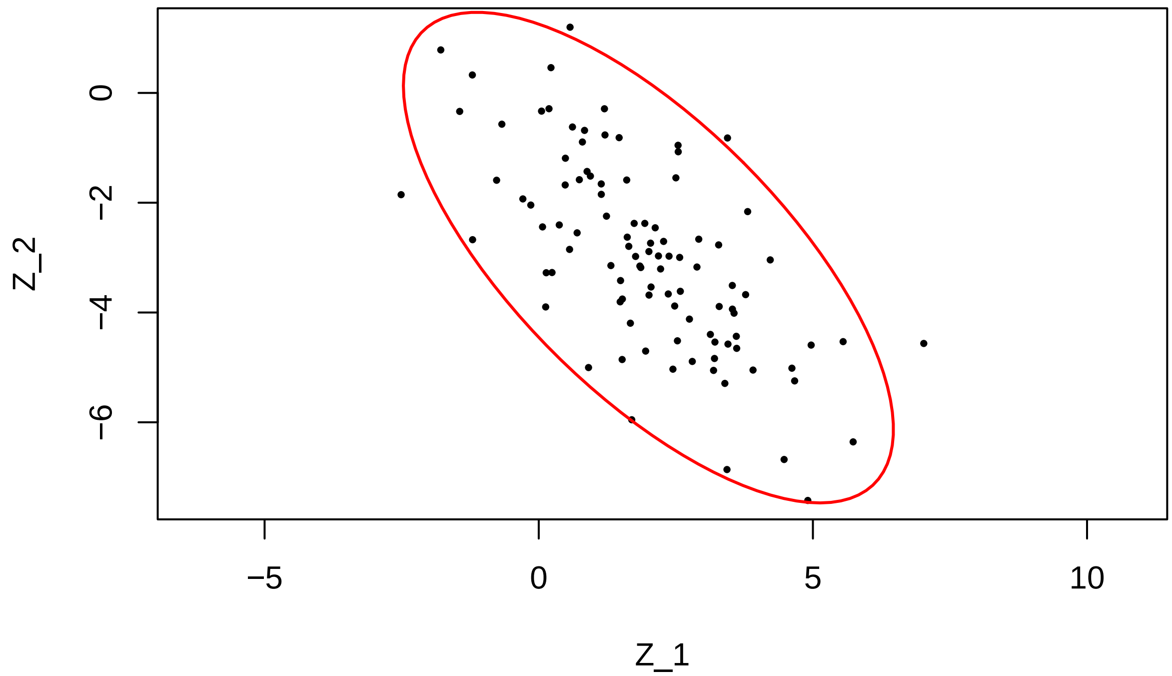
## 2.4

Let $\rho = -2/3$, then

$$\mu_Z = \begin{pmatrix} 2 \\ -3 \end{pmatrix}, \ \mathbf{\Sigma}_Z = \begin{pmatrix} 10/3 & -7/3 \\ -7/3 & 10/3 \end{pmatrix}$$

In order to sketch an ellipse, we simulate a sample of size 100 from the distribution of $Z$.

```
library(MASS)
library(ellipse)

rho<-(-2/3)
x <- c(2-2*rho, 2*rho-1,2*rho-1, 2-2*rho)
s<-round(matrix(x, nrow = 2, ncol = 2),3)
mu<-c(2,-3)

sample_size <- 100

sample_distribution <- mvrnorm(n = sample_size, mu, Sigma = s)

z<-round(matrix(c(sample_distribution[,1],sample_distribution[,2]), ncol = 2, nrow = 100),3)

plot(z, pch=16,asp=1, cex=1/2, xlab = "Z_1", ylab = "Z_2")
lines(ellipse(x=s[,1:2], centre=mu[1:2], level = 0.95), lwd=1.5, col="red")
```




As we can see, the ellipse contains around the 95% of the points generated.

If we consider the equation of the ellipse $(z-\mu_Z)^t\mathbf{\Sigma}_Z{}^{-1}(z-\mu_Z)=c^2$ we have that the constant $c^2$ we need in this case is Chi-squared quantile of order 0.95 ($\alpha = 0.05$) with 2 degrees of freedom, which is 5.991.

If we compute eigenvalues and eigenvectors of $\mathbf{\Sigma}_Z{}^{-1}$ (which is positive definite), we get that the axes of the ellipse are in the direction of the eigenvectors and are proportional to the inverse of the square root of the eigenvalues.

In fact

$\mathbf{\Sigma}_Z{}^{-1} = \begin{pmatrix} 10/17 & 7/17 \\ 7/17 & 10/17 \end{pmatrix}$

Compute its eigenvalues:

$det(\mathbf{\Sigma}_Z{}^{-1} - \lambda I) = 0$ has solutions $\lambda_1 = 1$ and $\lambda_2 = \dfrac{3}{17}$.

Compute the eigenvectors solving the system $\mathbf{\Sigma}_Z{}^{-1}e_i = \lambda_i e_i$ for $i = 1, 2$:

- $e_1 = (1, 1)$
- $e_2 = (1, -1)$

So the ellipse centered in the point $(2, -3)$ has one semiaxis in the direction of the vector $(1, 1)$ with length equal to $c = 2.448$, and the other semiaxis in the direction of the vector $(1, -1)$ with length equal to $c\sqrt{\dfrac{17}{3}} = 5.828$.

### 2.5

If we consider $\rho = 2/3$ we have

$\mathbf{\Sigma}_Z = \begin{pmatrix} 2/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix}$ and so $\mathbf{\Sigma}_Z{}^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$

Compute the eigenvalues of $\mathbf{\Sigma}_Z{}^{-1}$:

$det(\mathbf{\Sigma}_Z{}^{-1} - \lambda I) = 0$ has solutions $\lambda_1 = 1$ and $\lambda_2 = 3$.

Compute the eigenvectors solving the system ${\boldsymbol{\Sigma}_Z}^{-1} e_i = \lambda_i e_i$ for $i = 1, 2$:

- $e_1 = (1, 1)$
- $e_2 = (1, -1)$

So the ellipse centered in the point $(2, -3)$ would have one semiaxis in the direction of the vector $(1, 1)$ with length equal to $c = 2.448$, and the other semiaxis in the direction of the vector $(1, -1)$ with length equal to $c\sqrt{\frac{1}{3}} = 1.413$.

### Exercise 3

```
pendigits<-read.table("data/pendigits.txt", sep=",",header =TRUE)
names(pendigits)<-c(paste0(rep(c("x","y"),8),rep(1:8,each=2)),"digit")
pander(summary(pendigits))
```

Table 13: Table continues below

| x1 | y1 | x2 | y2 |
|---|---|---|---|
| Min. : 0.00 | Min. : 0.00 | Min. : 0.00 | Min. : 0.00 |
| 1st Qu.: 6.00 | 1st Qu.: 76.00 | 1st Qu.: 20.00 | 1st Qu.: 72.00 |
| Median : 32.00 | Median : 89.00 | Median : 40.00 | Median : 91.00 |
| Mean : 38.81 | Mean : 85.12 | Mean : 40.61 | Mean : 83.77 |
| 3rd Qu.: 65.00 | 3rd Qu.:100.00 | 3rd Qu.: 58.00 | 3rd Qu.:100.00 |
| Max. :100.00 | Max. :100.00 | Max. :100.00 | Max. :100.00 |

Table 14: Table continues below

| x3 | y3 | x4 | y4 |
|---|---|---|---|
| Min. : 0.00 | Min. : 0.00 | Min. : 0.00 | Min. : 0.0 |
| 1st Qu.: 18.00 | 1st Qu.: 49.00 | 1st Qu.: 28.00 | 1st Qu.: 23.0 |
| Median : 53.00 | Median : 71.00 | Median : 54.00 | Median : 43.0 |
| Mean : 49.77 | Mean : 65.58 | Mean : 51.22 | Mean : 44.5 |
| 3rd Qu.: 78.00 | 3rd Qu.: 86.00 | 3rd Qu.: 74.00 | 3rd Qu.: 64.0 |
| Max. :100.00 | Max. :100.00 | Max. :100.00 | Max. :100.0 |

Table 15: Table continues below

| x5 | y5 | x6 | y6 |
|---|---|---|---|
| Min. : 0.00 | Min. : 0.0 | Min. : 0.00 | Min. : 0.00 |
| 1st Qu.: 29.00 | 1st Qu.: 7.0 | 1st Qu.: 23.00 | 1st Qu.: 11.00 |
| Median : 60.00 | Median : 33.0 | Median : 73.00 | Median : 30.00 |
| Mean : 56.87 | Mean : 33.7 | Mean : 60.52 | Mean : 34.82 |
| 3rd Qu.: 89.00 | 3rd Qu.: 54.0 | 3rd Qu.: 97.00 | 3rd Qu.: 55.00 |
| Max. :100.00 | Max. :100.0 | Max. :100.00 | Max. :100.00 |

| x7 | y7 | x8 | y8 | digit |
|---|---|---|---|---|
| Min. : 0.00 | Min. : 0.00 | Min. : 0.00 | Min. : 0.00 | Min. :0.000 |
| 1st Qu.: 42.00 | 1st Qu.: 5.00 | 1st Qu.: 0.00 | 1st Qu.: 0.00 | 1st Qu.:2.000 |
| Median : 53.00 | Median : 27.00 | Median : 40.00 | Median : 9.00 | Median :4.000 |
| Mean : 55.02 | Mean : 34.93 | Mean : 47.29 | Mean : 28.84 | Mean :4.431 |
| 3rd Qu.: 68.00 | 3rd Qu.: 47.00 | 3rd Qu.:100.00 | 3rd Qu.: 51.00 | 3rd Qu.:7.000 |
| Max. :100.00 | Max. :100.00 | Max. :100.00 | Max. :100.00 | Max. :9.000 |

```
head(pendigits)
```

```
##    x1  y1 x2  y2  x3  y3  x4  y4  x5 y5  x6  y6 x7 y7  x8 y8 digit
## 1   0  89 27 100  42  75  29  45  15 15  37   0 69  2 100  6     2
```

```
## 2    0  57 31  68  72  90 100 100  76 75  50 51 28 25  16  0     1
## 3    0 100  7  92   5  68  19  45  86 34 100 45 74 23  67  0     4
## 4    0  67 49  83 100 100  81  80  60 60  40 40 33 20  47  0     1
## 5  100 100 88  99  49  74  17  47   0 16  37  0 73 16  20 20     6
## 6    0 100  3  72  26  35  85  35 100 71  73 97 65 49  66  0     4
```

```r
dim(pendigits)
```

```
## [1] 10991    17
```

We consider only the first 16 variables, excluding the column digit.

### 3.1

Let's perform a principal component analysis on the standardized variables.

```r
pendigits.pca<- prcomp(X, scale = TRUE)
pander(summary(pendigits.pca))
```

Table 17: Principal Components Analysis (continued below)

| | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 |
|---|---|---|---|---|---|---|
| **x1** | 0.05066 | 0.03322 | 0.4512 | -0.2524 | 0.02836 | -0.5236 |
| **y1** | 0.1338 | 0.1253 | 0.3397 | -0.09392 | -0.4521 | 0.2077 |
| **x2** | -0.2625 | 0.1168 | 0.201 | -0.3576 | 0.4167 | -0.02727 |
| **y2** | -0.302 | 0.2421 | 0.2056 | 0.04589 | -0.3051 | 0.161 |
| **x3** | -0.2514 | 0.09057 | -0.3056 | -0.2982 | 0.2855 | 0.4391 |
| **y3** | -0.4238 | 0.05089 | 0.005033 | 0.1175 | 0.007355 | 0.01998 |
| **x4** | -0.1291 | -0.1999 | -0.4003 | -0.1922 | -0.3611 | 0.1023 |
| **y4** | -0.3755 | -0.2423 | -0.04429 | 0.1162 | 0.08174 | -0.1447 |
| **x5** | -0.02832 | -0.4446 | 0.02916 | -0.09824 | -0.3919 | 0.07551 |
| **y5** | -0.112 | -0.4748 | -0.07755 | 0.07532 | 0.1542 | -0.2132 |
| **x6** | 0.05601 | -0.3873 | 0.3422 | 0.01084 | 0.06138 | 0.3952 |
| **y6** | 0.2983 | -0.316 | -0.1501 | 0.02988 | 0.1179 | -0.147 |
| **x7** | 0.134 | -0.06212 | 0.2565 | 0.5623 | 0.2955 | 0.3639 |
| **y7** | 0.4194 | 0.0177 | -0.1555 | -0.148 | 0.08739 | 0.04749 |
| **x8** | 0.001823 | 0.3012 | -0.296 | 0.4874 | -0.1164 | -0.2173 |
| **y8** | 0.3459 | 0.1834 | -0.1368 | -0.2286 | 0.07733 | 0.1435 |

Table 18: Table continues below

| | PC7 | PC8 | PC9 | PC10 | PC11 | PC12 |
|---|---|---|---|---|---|---|
| **x1** | -0.1114 | 0.4764 | 0.04265 | -0.2084 | -0.07004 | 0.2638 |
| **y1** | 0.6707 | 0.1002 | -0.07929 | 0.008563 | -0.3246 | -0.08009 |
| **x2** | 0.00306 | -0.05811 | -0.6431 | -0.01757 | -0.03426 | -0.1396 |
| **y2** | -0.02921 | 0.1691 | -0.1742 | 0.3235 | 0.6275 | -0.02799 |
| **x3** | 0.2473 | 0.1272 | -0.002112 | -0.1911 | -0.1153 | 0.1931 |
| **y3** | -0.04127 | 0.1762 | 0.1201 | 0.4779 | -0.1629 | 0.334 |
| **x4** | -0.09887 | 0.4756 | -0.06027 | -0.359 | 0.1373 | 0.08331 |
| **y4** | 0.08868 | 0.2177 | 0.1676 | 0.2069 | -0.3314 | -0.03598 |
| **x5** | -0.355 | -0.08655 | -0.4049 | 0.05683 | -0.2217 | -0.136 |
| **y5** | 0.3426 | 0.1343 | -0.03611 | 0.1503 | 0.1071 | -0.4414 |
| **x6** | -0.1606 | -0.1331 | -0.01893 | 0.09854 | -0.1321 | 0.3971 |
| **y6** | 0.3591 | 0.04095 | -0.2658 | 0.1892 | 0.3283 | 0.3835 |

| | PC7 | PC8 | PC9 | PC10 | PC11 | PC12 |
|---|---|---|---|---|---|---|
| **x7** | -0.07293 | 0.4545 | -0.104 | -0.2421 | 0.05536 | -0.1493 |
| **y7** | -0.01828 | 0.1687 | -0.08703 | 0.2944 | -0.004369 | 0.2189 |
| **x8** | 0.005871 | 0.03605 | -0.4965 | -0.02084 | -0.3157 | 0.1813 |
| **y8** | -0.2291 | 0.3591 | 0.02515 | 0.4448 | -0.1939 | -0.3542 |

| | PC13 | PC14 | PC15 | PC16 |
|---|---|---|---|---|
| **x1** | -0.2906 | 0.08907 | -0.04465 | -0.02379 |
| **y1** | 0.115 | 0.001655 | 0.05864 | -0.003507 |
| **x2** | 0.3211 | -0.1447 | 0.1129 | 0.007611 |
| **y2** | -0.2009 | -0.1092 | -0.274 | -0.02009 |
| **x3** | -0.4481 | 0.2513 | -0.2082 | -0.01846 |
| **y3** | 0.1583 | 0.3282 | 0.4944 | -0.1125 |
| **x4** | 0.2634 | -0.2887 | 0.2416 | 0.002166 |
| **y4** | 0.2553 | -0.229 | -0.5949 | 0.22 |
| **x5** | -0.1035 | 0.4722 | -0.172 | -0.01075 |
| **y5** | -0.3532 | -0.1166 | 0.23 | -0.3524 |
| **x6** | -0.2347 | -0.5214 | 0.1185 | 0.003394 |
| **y6** | 0.07197 | 0.1756 | 0.001117 | 0.4735 |
| **x7** | 0.1513 | 0.2159 | -0.01709 | -0.00683 |
| **y7** | 0.24 | -0.01838 | -0.3004 | -0.6699 |
| **x8** | -0.2969 | -0.2296 | 0.008691 | -0.03891 |
| **y8** | -0.1848 | -0.1169 | 0.1402 | 0.3725 |

Table 20: Table continues below

| | PC1 | PC2 | PC3 | PC4 | PC5 |
|---|---|---|---|---|---|
| **Standard deviation** | 2.172 | 1.797 | 1.605 | 1.109 | 1.031 |
| **Proportion of Variance** | 0.2948 | 0.2018 | 0.1611 | 0.07687 | 0.06644 |
| **Cumulative Proportion** | 0.2948 | 0.4966 | 0.6576 | 0.7345 | 0.801 |

Table 21: Table continues below

| | PC6 | PC7 | PC8 | PC9 | PC10 |
|---|---|---|---|---|---|
| **Standard deviation** | 0.8929 | 0.7789 | 0.7404 | 0.641 | 0.5461 |
| **Proportion of Variance** | 0.04983 | 0.03792 | 0.03426 | 0.02568 | 0.01864 |
| **Cumulative Proportion** | 0.8508 | 0.8887 | 0.923 | 0.9486 | 0.9673 |

Table 22: Table continues below

| | PC11 | PC12 | PC13 | PC14 | PC15 |
|---|---|---|---|---|---|
| **Standard deviation** | 0.4588 | 0.3351 | 0.2837 | 0.2408 | 0.1851 |
| **Proportion of Variance** | 0.01316 | 0.00702 | 0.00503 | 0.00363 | 0.00214 |
| **Cumulative Proportion** | 0.9805 | 0.9875 | 0.9925 | 0.9961 | 0.9983 |

| | PC16 |
|---|---|
| **Standard deviation** | 0.1667 |
| **Proportion of Variance** | 0.00174 |
| **Cumulative Proportion** | 1 |

The first 5 components explain 80% of the whole data variability while the first 6 components explain almost 85%. But we could set k = 5 < 16 = p since 80% is good enough to our purpose.

The standard deviations of PCs are:

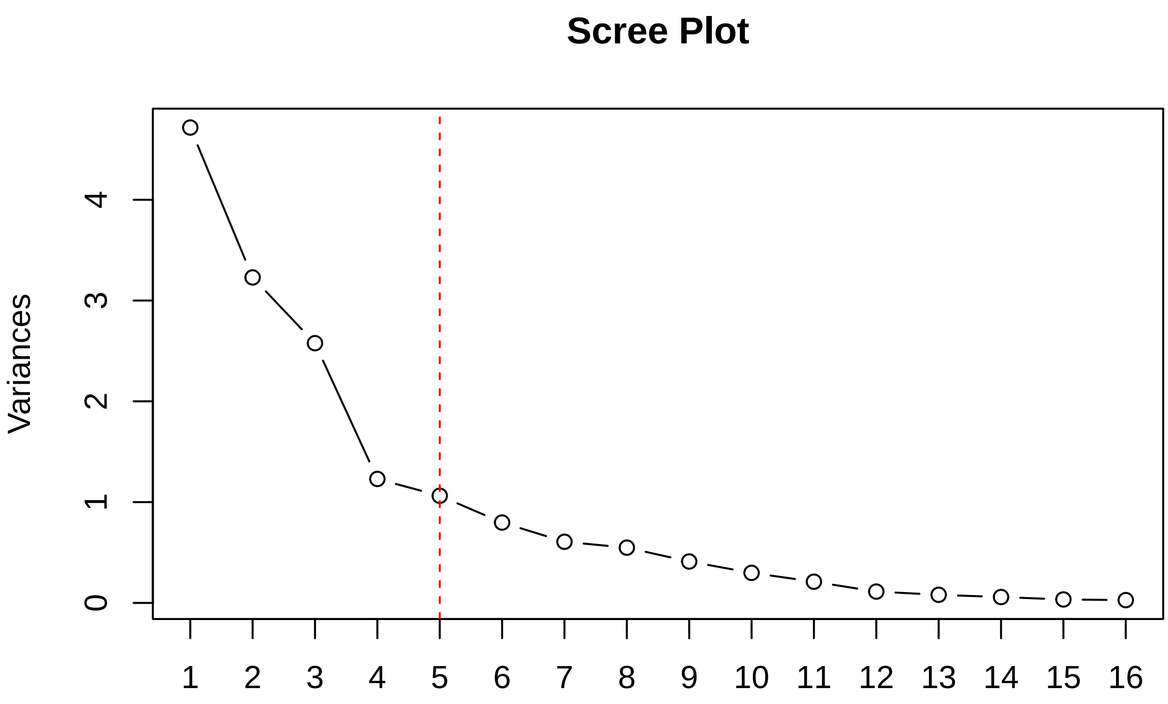
```
##  [1] 2.1717023 1.7969951 1.6052845 1.1090179 1.0310554 0.8928969 0.7788843
##  [8] 0.7403863 0.6409729 0.5461442 0.4588082 0.3351097 0.2836961 0.2408487
## [15] 0.1851186 0.1667406
```

In order to check if k=5 PCs are enough, we analyze whether the vector of the eigenvalues is larger than their sample mean (that is equal to 1):

```
##  [1]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE
```

We get that the first 5 principal components are enough.

Another method to determine the number of PCs to be retained is a graphical representation known as scree plot, which shows the eigenvalues for each individual PC.



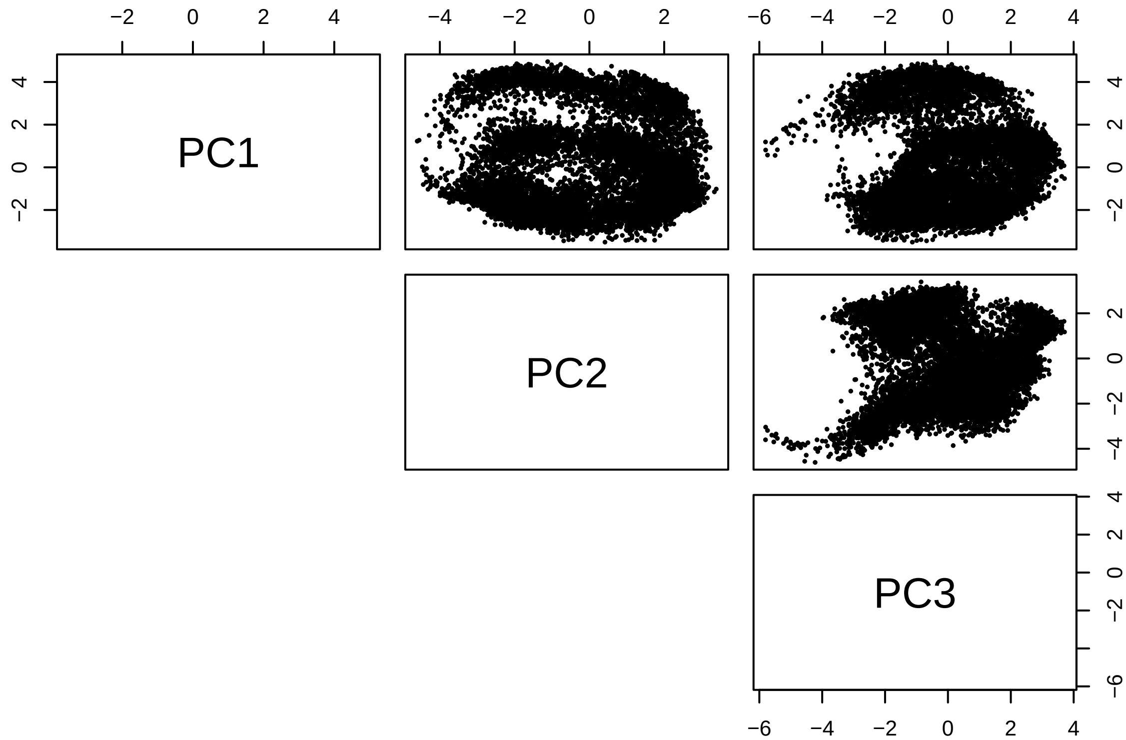

As we can see it forms a steep curve followed by a bend and then a straight-line trend. The elbow is shown at k+1 = 5 and so the scree plot tells us to consider only the first 4 components.

However considering the fact that both the previous two methods gave us 5 PCs, we retain k=5 PCs.

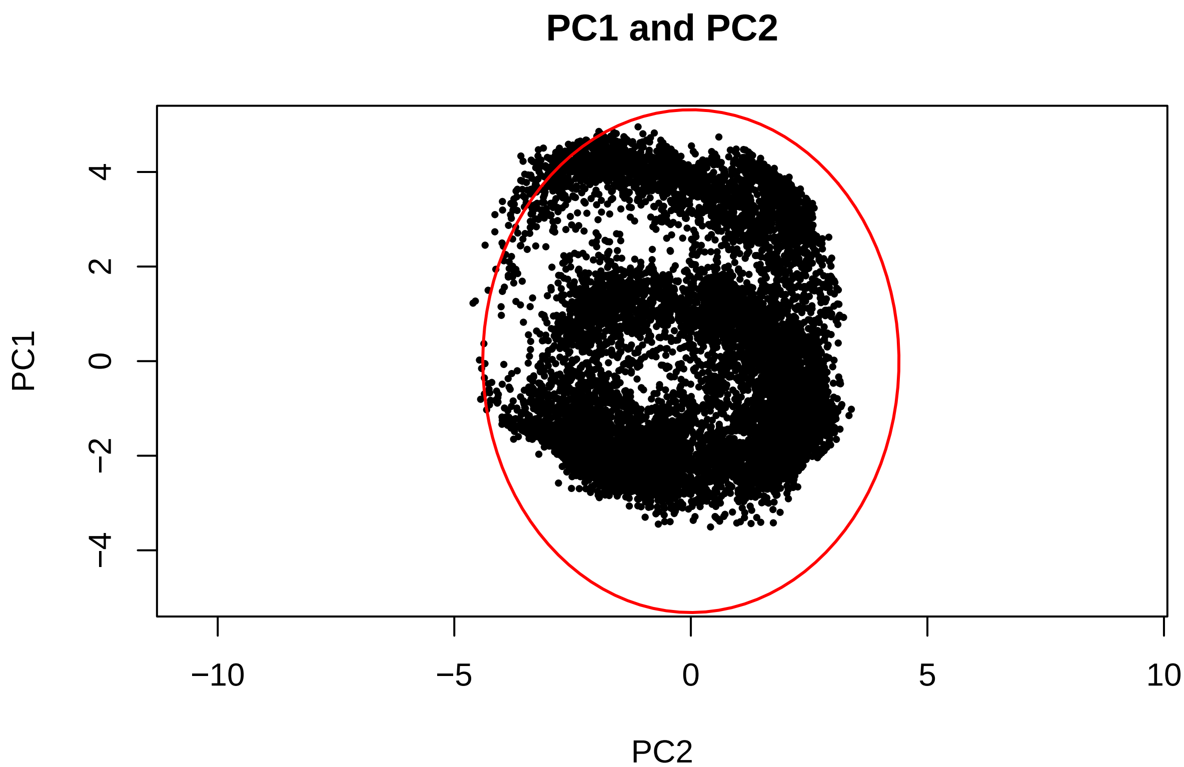
### 3.2

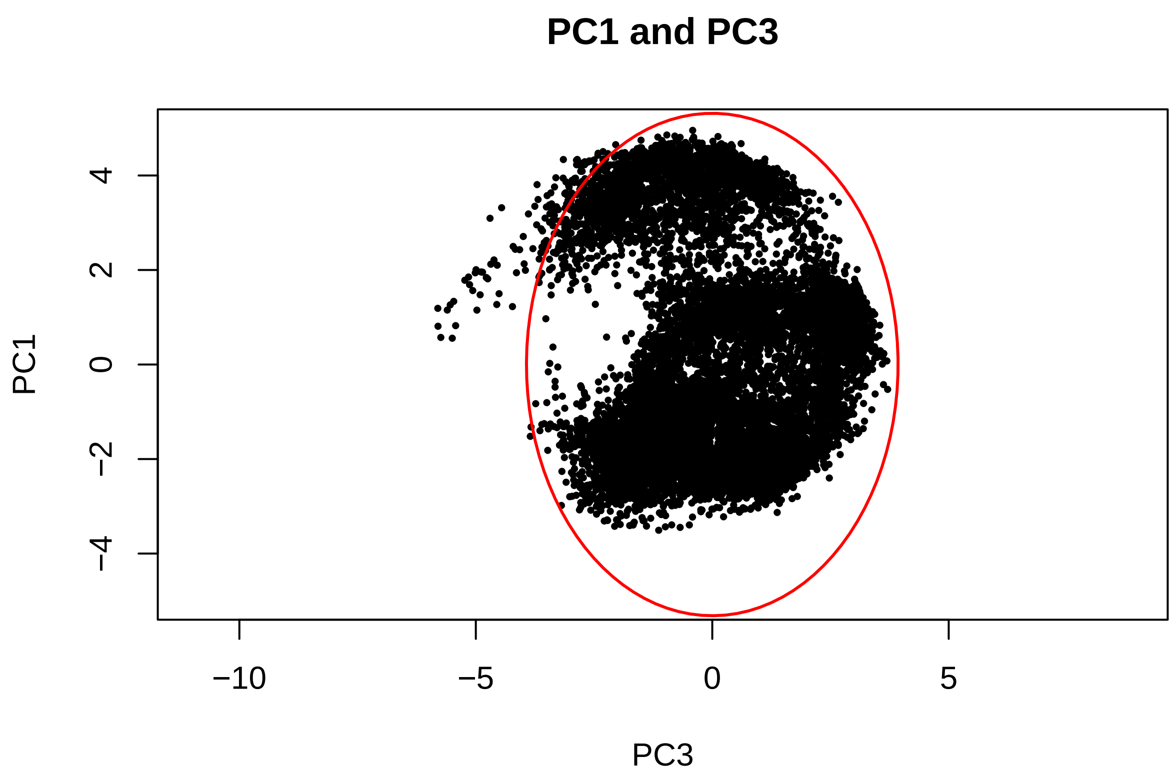
We want to check multivariate normality through the first three PCs.

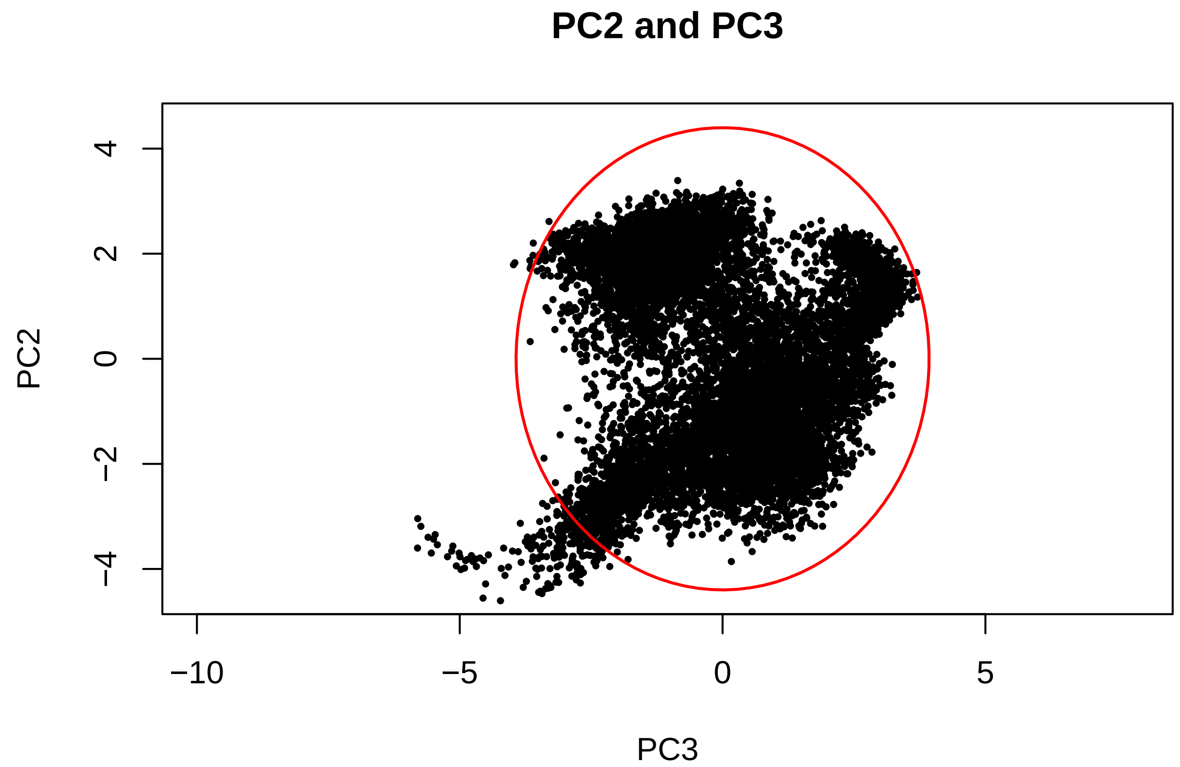
We consider the scatter plots for the first three PCs: the observation (black points) represent the scores for PC1, PC2 and PC3.



We evaluate the bivariate normality:

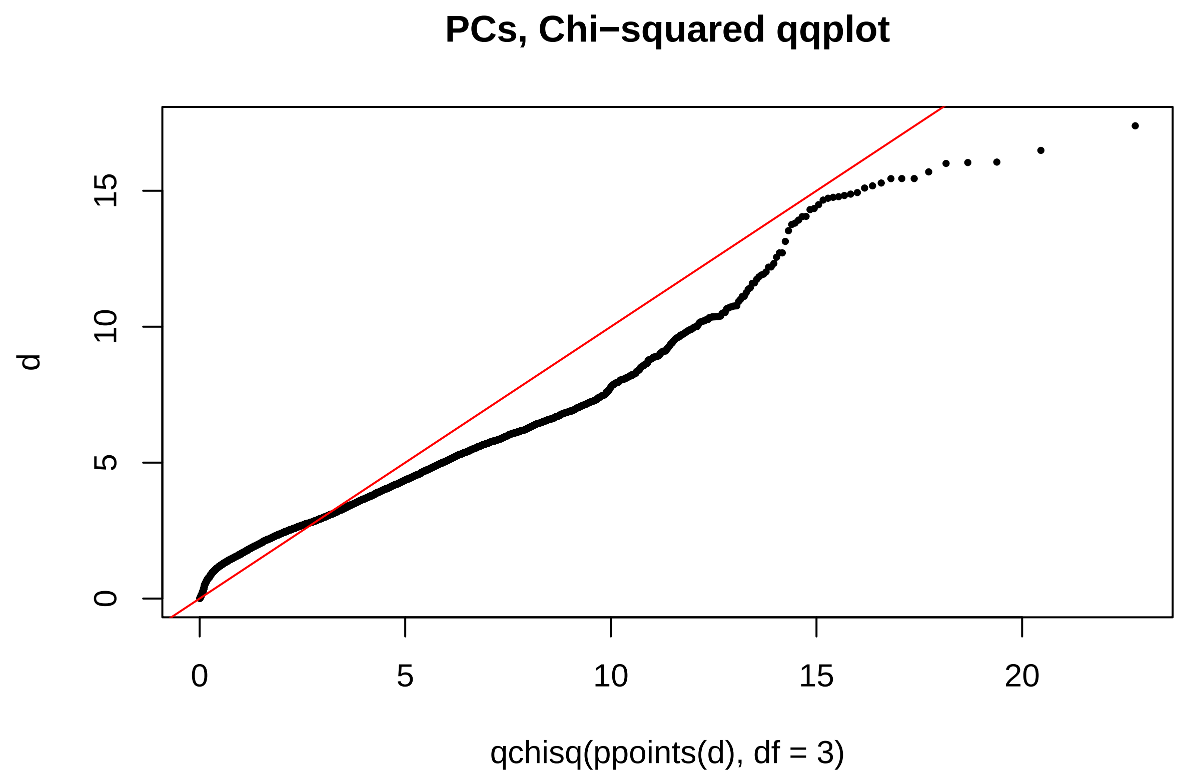





If bivariate normality was respected, we would expect to have around 550 (5% of 10992) points out of the ellipse. From the first two plots though, it seems that they have not a bivariate normal distribution since there are less points than what we expected. So we can conclude that the observed data are not bivariate normally distributed.

To determine the multivariate normality, let's check the Chi-square QQ-plot of squared Mahalanobis distances.

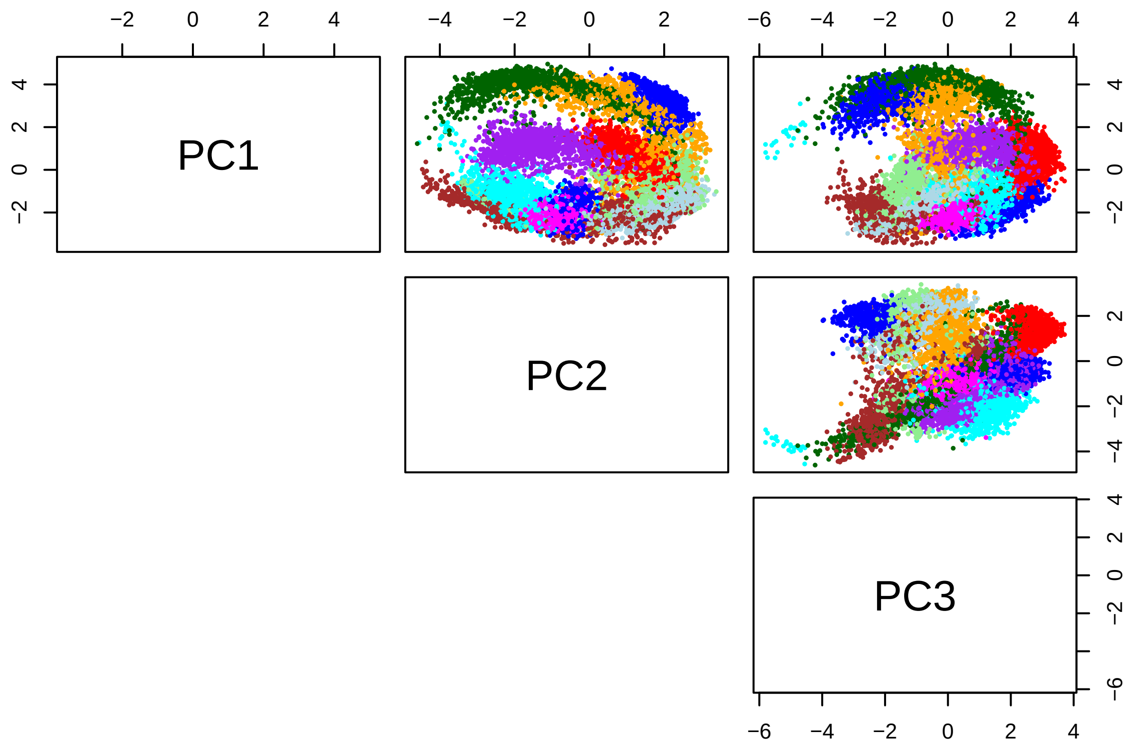


It does not seem to have a multivariate normal distribution since it doesn't follow a linear trend (red line). In fact from the previous plots we have already said that it was not bivariate normal distributed, and so it couldn't be multivariate.

### 3.3

We make scatter plots with the first three principal components, color coding the observations according to the digit class.

```
pairs(pca3,pch=16, cex=1/2,lower.panel=NULL, col= digit.col)
```



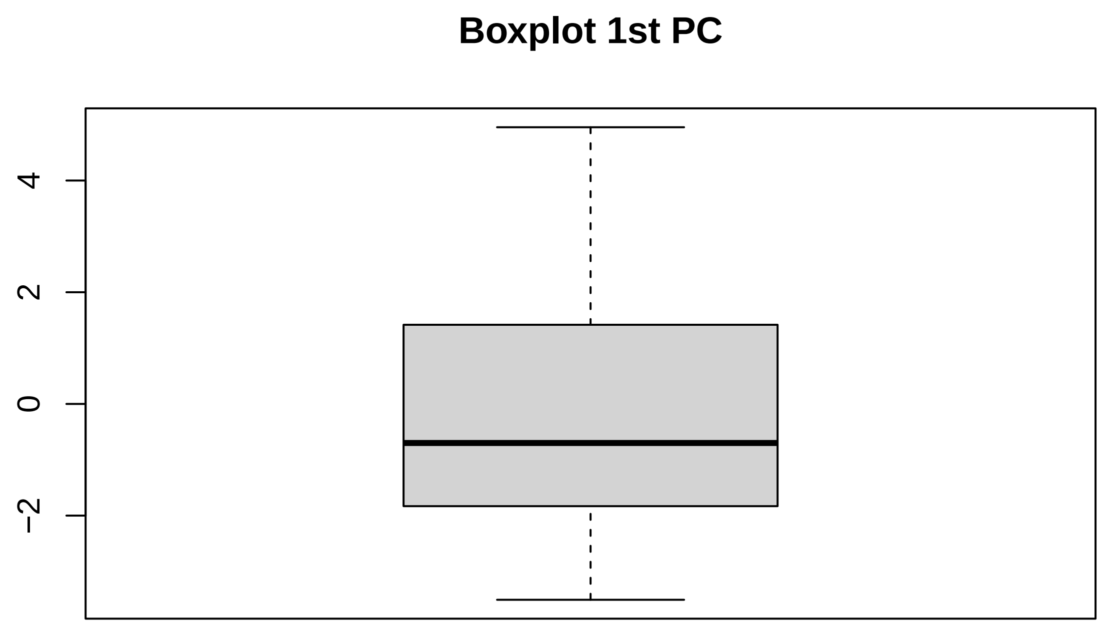

We can see that the points form some clusters of colors.

Notice that it seems that all points are gathered together, but there are some points forming a tail out of the cloud with respect to the component of PC3.

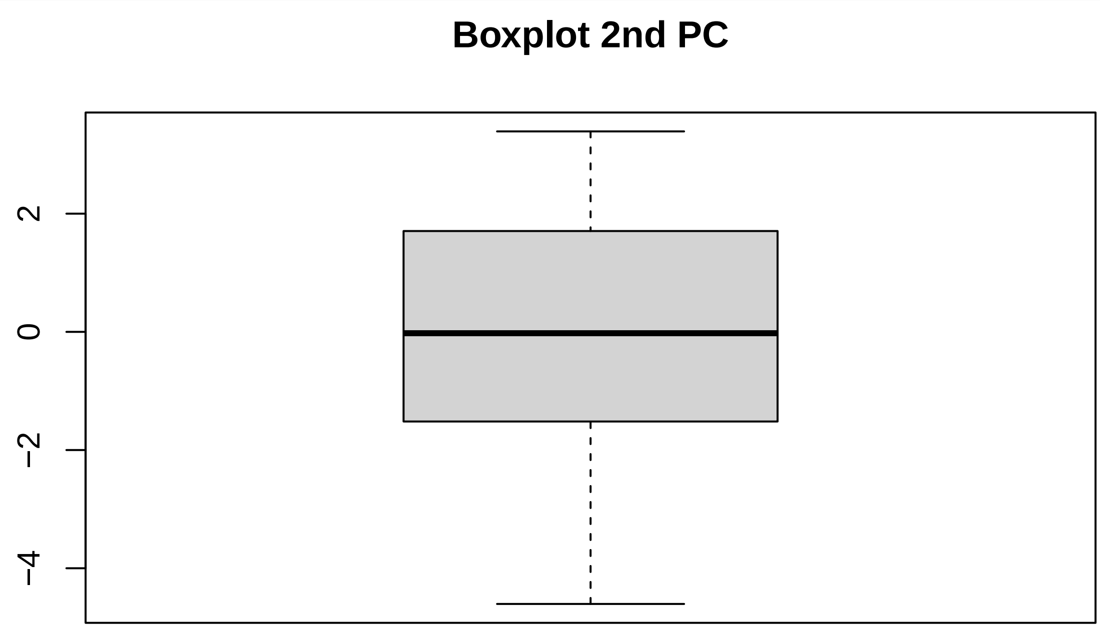
### 3.4

From the boxplot of each of the first 3 PCs we can identify some possible univariate outliers.

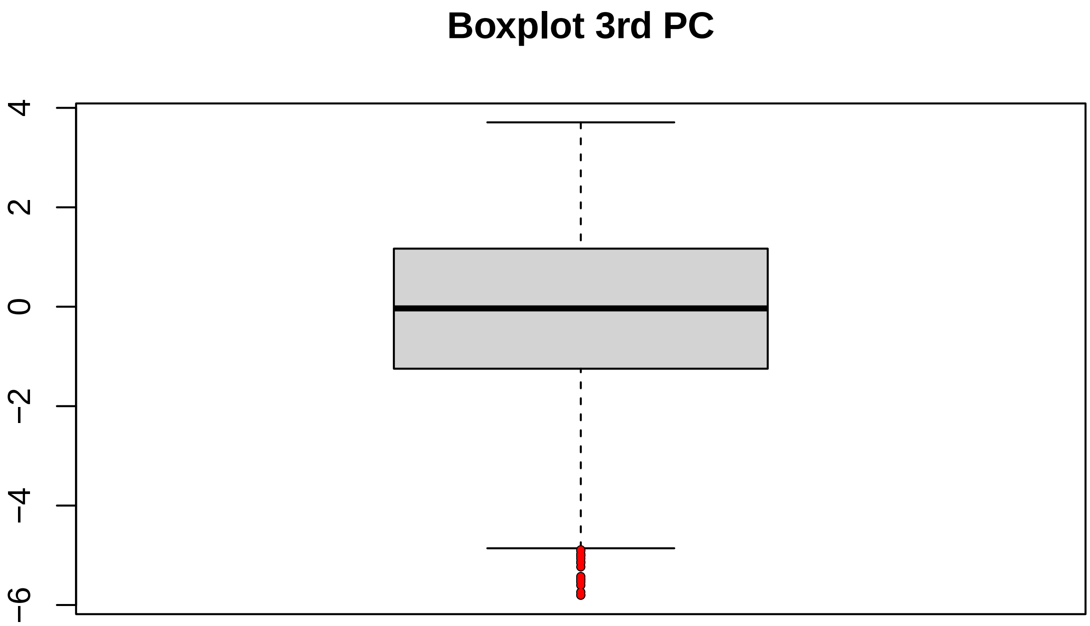
```
boxplot(pca3[,1], main = "Boxplot 1st PC", cex=1/2)
```



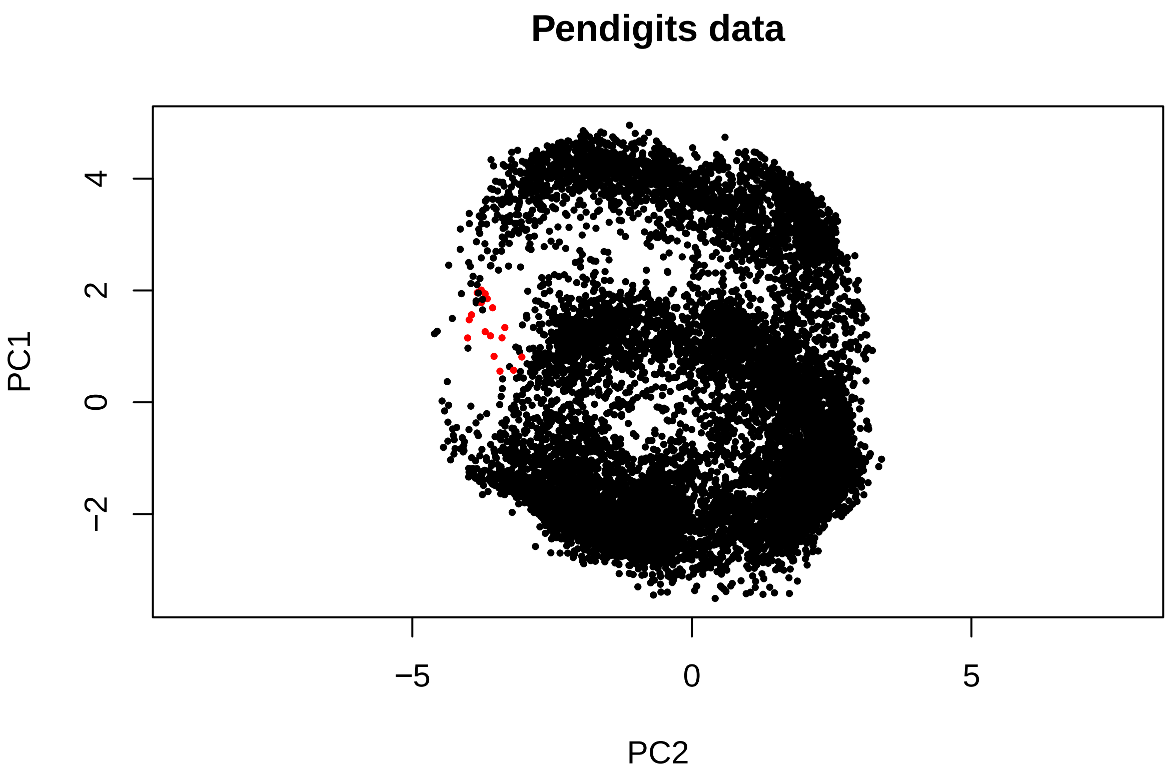

```
boxplot(pca3[,2], main = "Boxplot 2nd PC", cex=1/2)
```

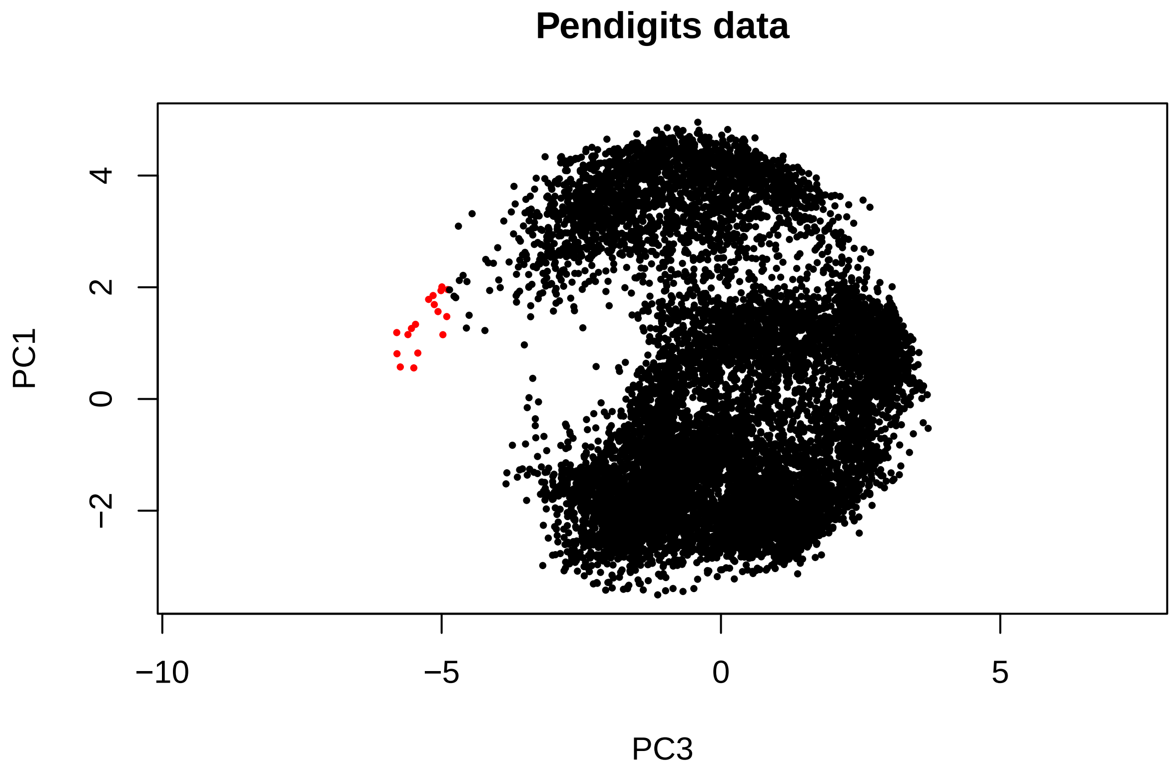


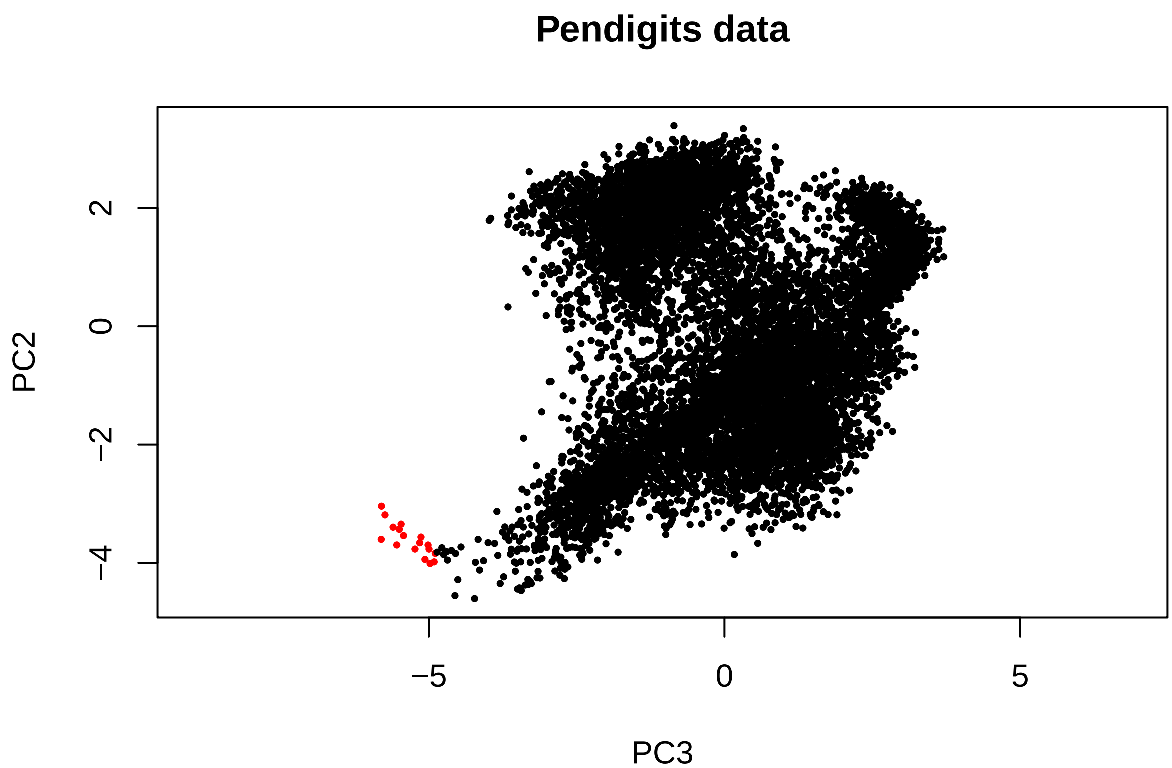

```
index <- which(pendigits.pca$x[,3] < (quantile(pendigits.pca$x[,3], .25)-
1.5*(quantile(pendigits.pca$x[,3], .75)-
quantile(pendigits.pca$x[,3], .25))))


boxplot(pendigits.pca$x[,3], main = "Boxplot 3rd PC", cex=1/2)
points(rep(1,17),pendigits.pca$x[index,3], pch=16, col="red", cex=1/2)

col.index.out<-rep("black",10992)
col.index.out[index]<-"red"
```




There are some univariate outliers for the 3rd PC: we colored the 17 observations in red in the box plot in order to visualize them in the scatter plots.



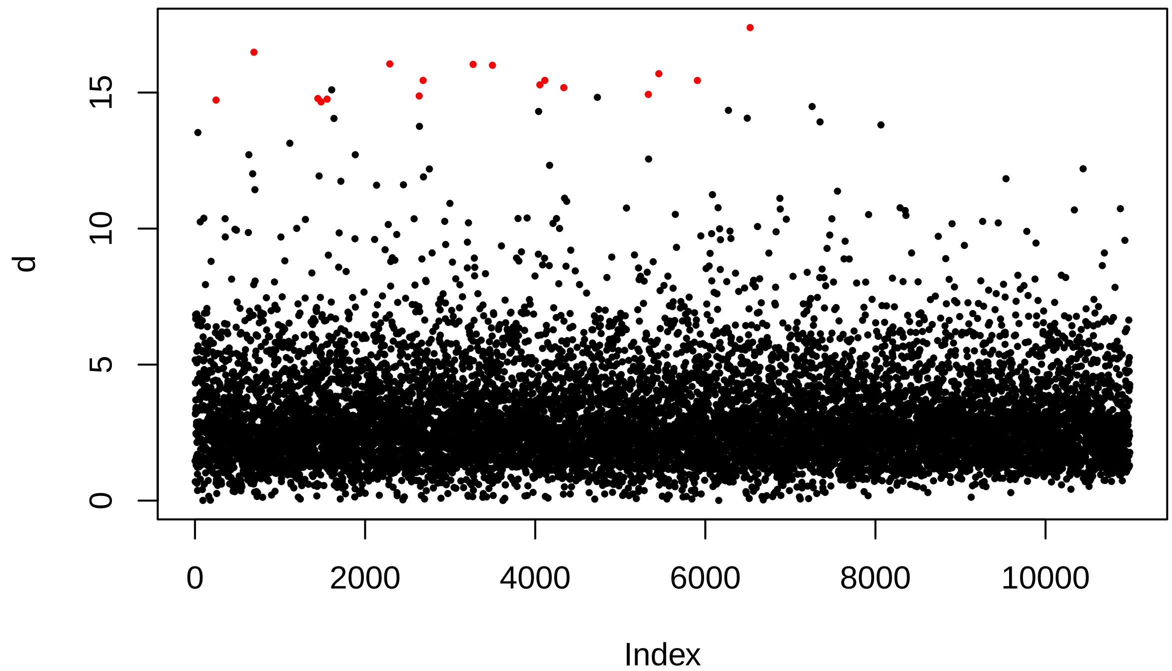



Those observation can be considered a suspect observation since they have a low score with respect to the 3rd PC.

Since we don't have the assumption of multivariate normality, we can not determine the multivariate outliers using the method with squared Mahalanobis distance. However we can plot the squared Mahalanobis distance and observe those points.

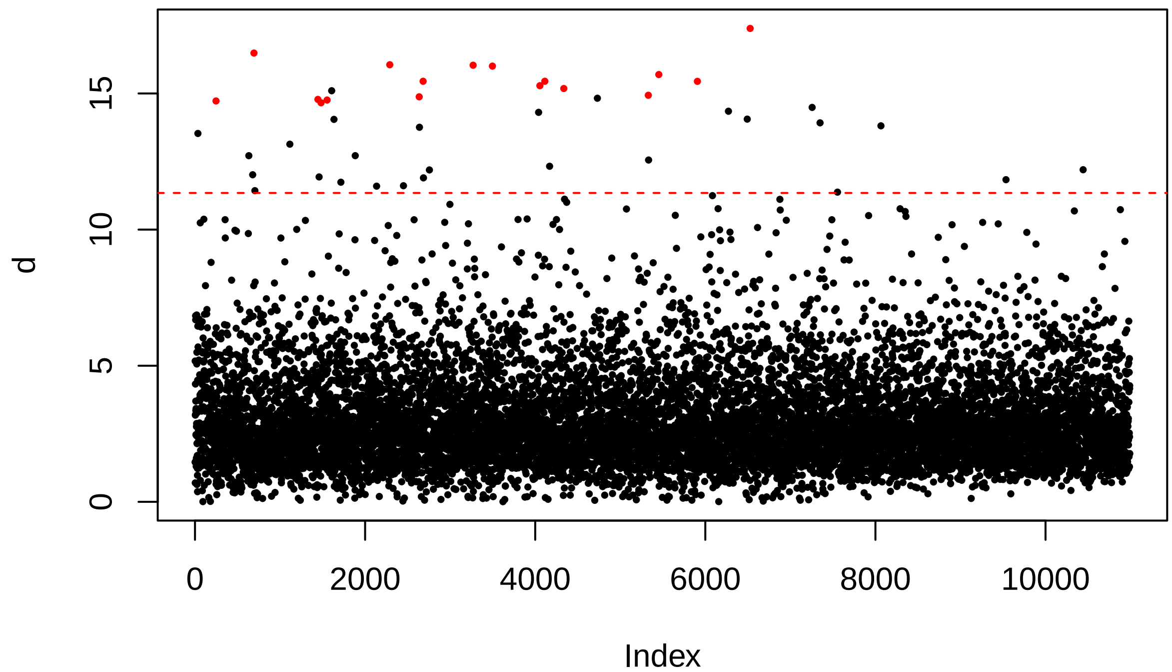
```
d<-mahalanobis(pca3,center=colmean,cov=s)
plot(d,pch=16, cex=1/2,  col= col.index.out)
```




From the plot with squared Mahalnobis distance we have noticed that those values, having a larger Mahalanobis distance, are located in the upper side of the plot.

Even though it is not multivariate Gaussian, we decide to plot the 99% quantile in order to visualize the points above the line.

```
d<-mahalanobis(pca3,center=colmean,cov=s)
plot(d,pch=16, cex=1/2,  col= col.index.out)
abline(h=qchisq(0.99,df=3), lty=2, col="red")
```



```
sum(d>qchisq(0.99,df=3))/dim(pca3)[1]
```

```
## [1] 0.004003275
```

From the plot, we obtain that there are not only the 17 flagged points above the line, corresponding to the 0.4% of points, instead of 1%.

In conclusion, we have 17 univariate outliers identified by the boxplot for the 3rd PC, but we can not say anything for sure about multivariate outliers.