# Multivariate Statistical Analysis Problem set 1

Lorenzo Sala 943481 - Stefano Sperti 947676 - Francesco Virgili 1101739

Università di Torino

```
knitr::opts_chunk$set(echo = TRUE,fig.align="center")
getwd()
library(moments)
library(ellipse)
library(corrplot)
library(scatterplot3d)
library(lemon)
library(MASS)
library(knitr)
library(formatR)
```

### Esercizio 1

Consider the *air pollution* data from which the variable *SO2* has been removed. As a preliminary step we select the variables of interest and change the variable "Neg.Temp" into "Temp" which has inverted sign and it is easier to interpret. It consists of 6 variables and 41 observations, each corresponding to a different city.

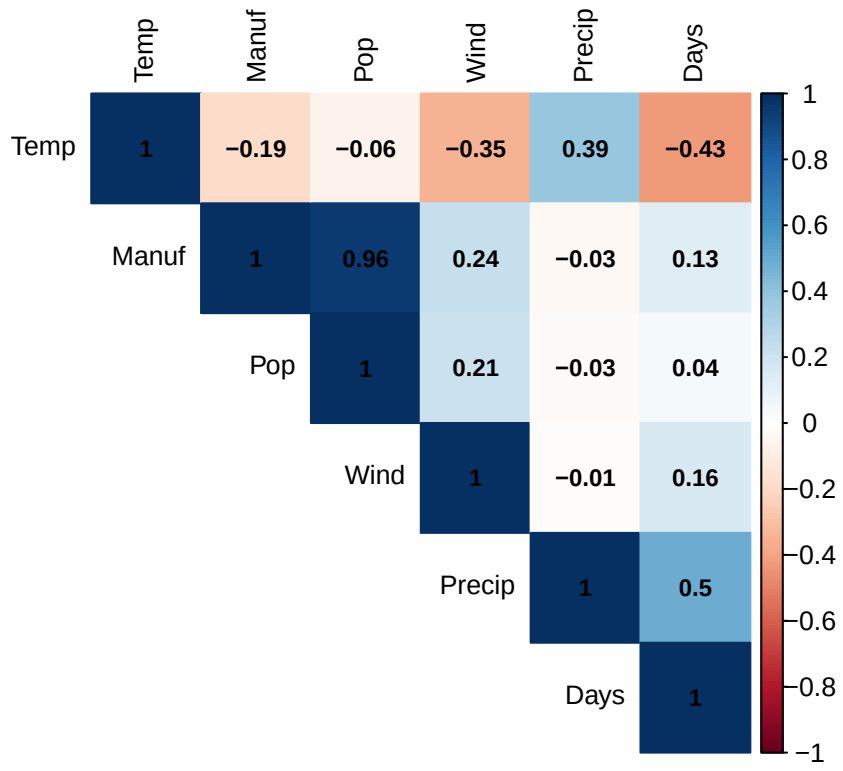
```
#setwd("C:/Users/fravi/Desktop/Universita/Laurea Magistrale/Anno 1/2° semestre/Multivariate statistical
usair<-read.table("data/usair.txt",header=TRUE)
usair <- usair[, !(names(usair) == "SO2")]
usair$Neg.Temp <- -usair$Neg.Temp
names(usair)[names(usair)=="Neg.Temp"] <- "Temp"
head(usair)
```

| | Temp | Manuf | Pop | Wind | Precip | Days |
|---|---|---|---|---|---|---|
| Phoenix | 70.3 | 213 | 582 | 6.0 | 7.05 | 36 |
| Little Rock | 61.0 | 91 | 132 | 8.2 | 48.52 | 100 |
| San Francisco | 56.7 | 453 | 716 | 8.7 | 20.66 | 67 |
| Denver | 51.9 | 454 | 515 | 9.0 | 12.95 | 86 |
| Hartford | 49.1 | 412 | 158 | 9.0 | 43.37 | 127 |
| Wilmington | 54.0 | 80 | 80 | 9.0 | 40.25 | 114 |

## 1.1

The sample correlation matrix can be obtained using the *cov()* function, which calculates the covariance matrix. However, for a clearer visualization, we've utilized the *corrplot* package.

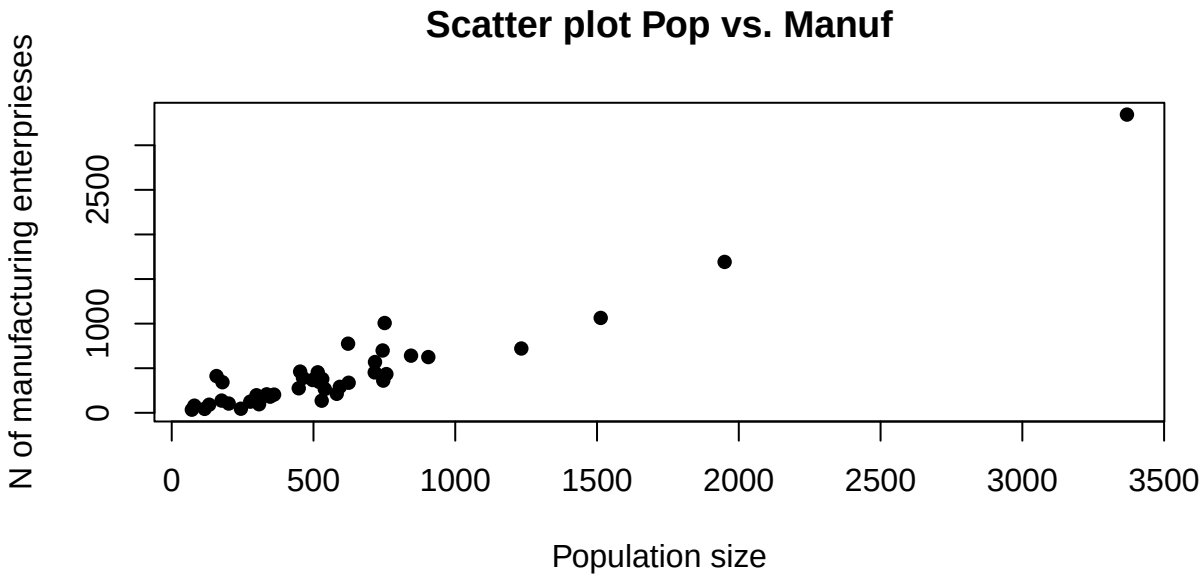
```
n <- dim(usair)[1]#obs
p <- dim(usair)[2]#var
R = round(cor(usair), 3)
corrplot(R, type = "upper", method = "color", tl.col = "black",
addCoef.col = "black", number.cex = 0.75, tl.cex = 0.8, cl.cex = 0.8)
```




For improving clarity we sort the absolute values of the correlation matrix in descending order.

```
##  [1] 0.955 0.496 0.430 0.386 0.350 0.238 0.213 0.190 0.164 0.132 0.063 0.042
## [13] 0.032 0.026 0.013
```

The correlation matrix shows that most of the couples of variables have low correlations; the most evident exception is the correlation between the size of the population (Pop) of the city and the number of manufacturing enterprises with more than 20 employees (Manuf), which is very close to 1:

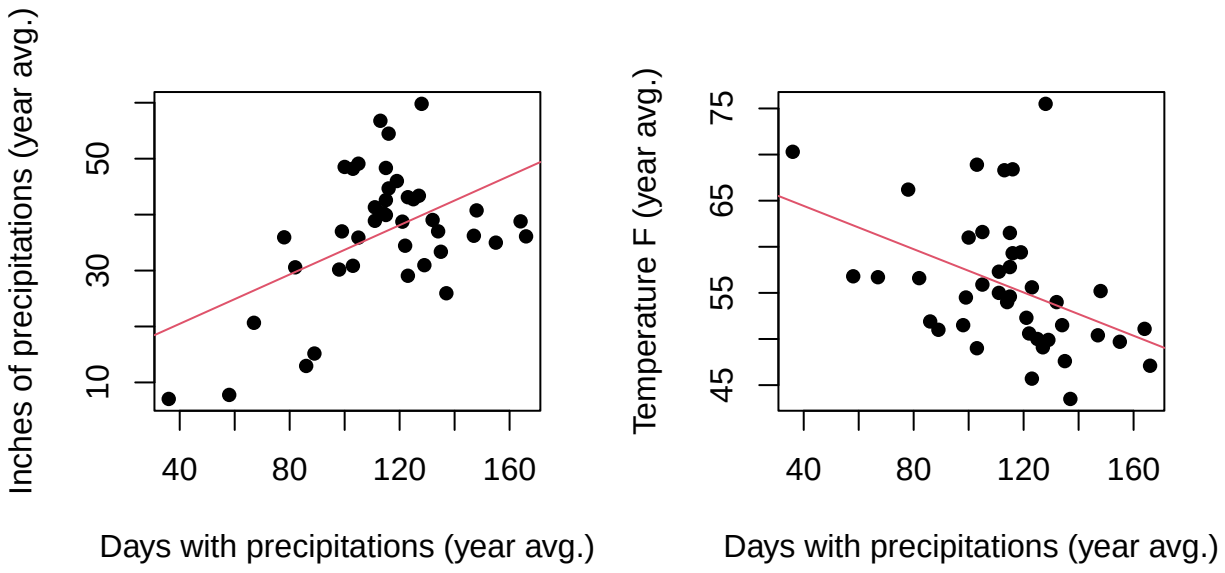


For the variables Population and Manufactory, the high correlation value ($\rho = 0.96$) is intuitive, as it suggests that cities with larger populations tend to have more operating manufactory industries. This correlation aligns with the expectation that higher population densities correspond to greater manufactory activity.

We also explore scatter plots for other pairs of variables with significant correlations ($\rho > 0.4$). These include the relationship between the average annual number of days with precipitation and both the annual average inches of precipitation, and the average annual temperatures.

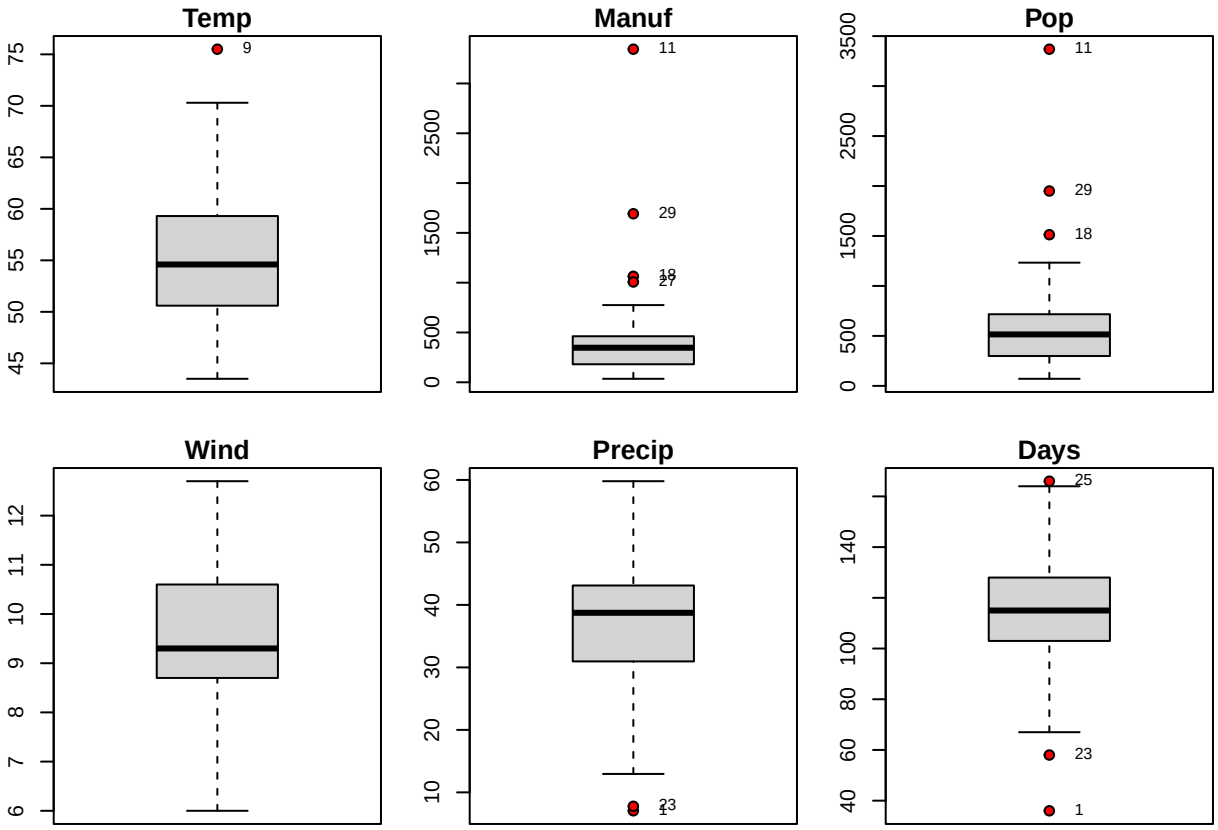
It is logical that as the number of days with precipitation increases, so does the amount of precipitation recorded annually, while temperatures tend to be lower. Indeed the variables Temperature and Days ($\rho = -0.430$) and Precipitation and Days ($\rho = 0.496$) exhibit moderate correlations, indicating that an increase in the average number of days with precipitation per year is associated with lower temperatures and higher precipitation levels. This relationship is consistent with expectations in meteorology, as cities experiencing colder climates tend to have more frequent and intense precipitation events.

The remaining correlation coefficients have relatively low absolute values, suggesting weak or negligible relationships between the corresponding variables.




## 1.2

We begin our analysis by examining outliers. Initially, we inspect the boxplots of each variable to identify potential univariate outliers. It's important to note that we cannot display the boxplots side by side due to the considerable differences in the scales of the variables.




For each variable, we take into account only the most extreme outliers checking which are the data that differ from the median of more than the interquantile distance multiplied by 2.5 (in R, the outliers pointed out by the boxplots are points that differ from the median more than 1.5 times the interquantile distance, here the multiplying constant is modified, in order to consider at most the two or three most significant outliers).

```
## [1] "Indices of outliers identified using the enlarged whiskers:"
```

```
## [1] 11 18 29  1 23
```

```
## [1] "Corresponding row names of outliers:"
```

```
## [1] "Chicago"      "Detroit"      "Philadelphia" "Phoenix"      "Alburquerque"
```

Examining the boxplots reveals that Wind is the only variable without potential outliers. Observation 9 appears to be an outlier for Temperature, and therefore, even without enlarging the whiskers, we decide to mark it because it could be useful for further analysis. Observations 11 and 29 could be outliers for both Manufacturing and Population, while observations 18 and 27 are positioned near the edges (whiskers). However Observation 27 disappears if we enlarge the whisker slightly, and therefore we do not consider it as an outlier. In particular, 18 results as an outlier in the enlarged whiskers for the variable Population and therefore we have decided to mark it as an outlier. For Precipitation and Days, observations 1 and 23 might be outliers for both variables and they result also as extreme outlier for the variable Precip. Similarly to observation 27 for Manuf, observation 25 for Days is likely not an outlier due to its proximity to the upper whisker.

In summary, observations 1, 9, 11, 18, 23, and 29 are potential outliers based on their individual positions within the data distribution.

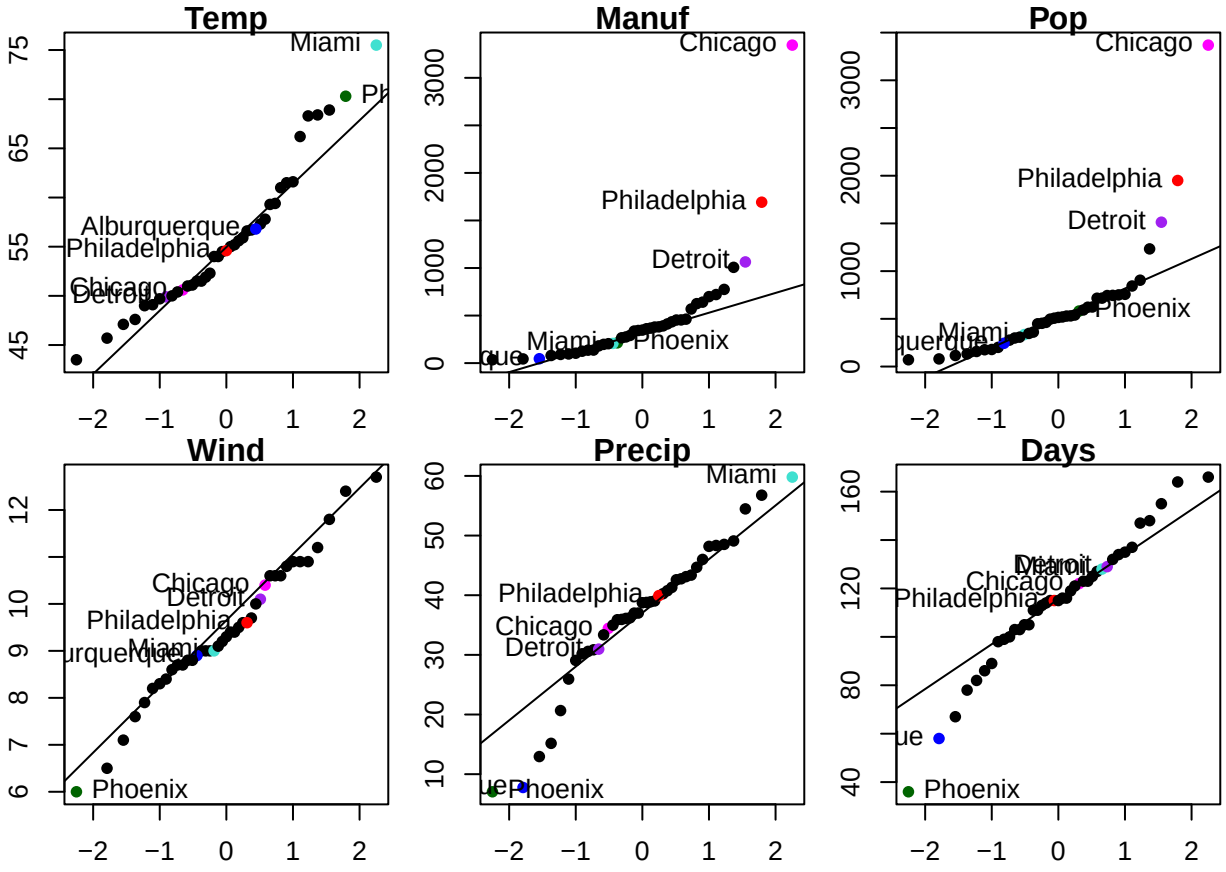
```
outliers <- c(1, 9, 11, 18,23, 29)
col.index<-rep("black",41)
```

```
col.index[outliers] <- lookup<-c("darkgreen", "turquoise", "magenta", "purple",
          "blue", "red")
```

## 1.3

Normal Q-Q plots are useful tools to assess normality in data. Each plot displays the theoretical quantiles of a normal distribution on the x-axis and the sample quantiles of the observed data on the y-axis. Ideally, data points should fall close to a straight diagonal line plotted in blue. Deviations from this line suggest potential departures from normality.

```
#3. Construct a normal Q-Q plot for each variable and comment about normality.
par(mfrow = c(2, 3))
par(mar = c(2, 2, 1, 1), cex = 0.8)
for (i in 1:p) {
  qqnorm(usair[,i],main=names(usair)[i],pch=16,col=col.index)
  qqline(usair[,i])
  qqx <- qqnorm(usair[,i],main=names(usair)[i],plot=F)$x
  qqy <- qqnorm(usair[,i],main=names(usair)[i],plot=F)$y
  text(qqx[outliers],qqy[outliers],row.names(usair[outliers,]),pos=c(4,2,2,2,2,2))
}
```




We see that the only variables that cluster closely around the Q-Q line and the sample mean and have no problems in the tails are Temp and Wind, which seem to be distributed pretty similarly to a normal. The variables which have the most significantly different observations from the theoretical expectations are Manuf and Pop which display unusual tail behavior; their distribution is similar, because of the almost perfect correlation between the two: an outlier for one will almost certainly be an outlier for the other. In both cases,

the outliers are Philadelphia, Chicago, and Detroit which we already identified in the previous point using boxplots. Also Phoenix and Alburquerque have values that considerably differ from the theoretical quantiles;

To continue the investigation of normality in the data, we can calculate both skewness and kurtosis for each variable.

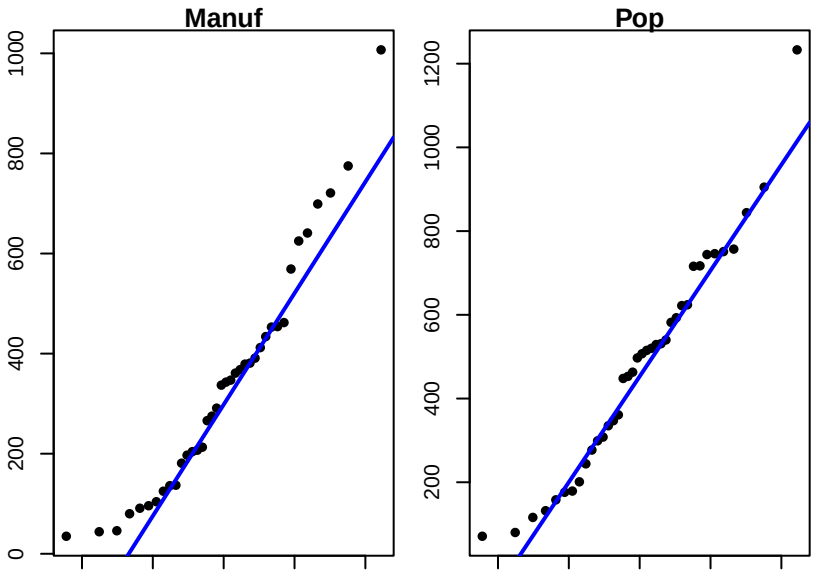
```
print(apply(usair,2,skewness))
```

```
##         Temp        Manuf          Pop         Wind       Precip         Days
##  0.854029361  3.616089258  3.052241663  0.002776073 -0.718649219 -0.570849126
```

```
print(apply(usair,2,kurtosis))
```

```
##      Temp     Manuf       Pop      Wind    Precip      Days
##  3.247125 18.209433 14.263346  3.215074  3.672754  3.908682
```

Manufacturing and Population clearly demonstrate non-normality. Their rightward skew and heavy tails are strong indicators of distributions that differ from the normal pattern. We now examine how their distributions change after removing the potential outliers.



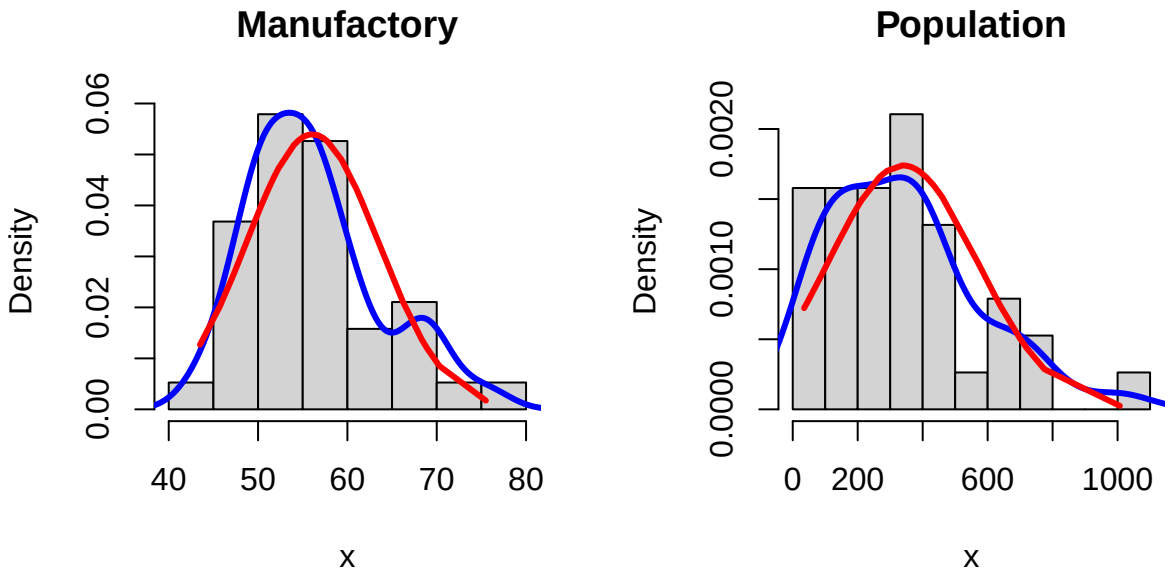

```
print(apply(new_usair[,2:3],2,skewness))
```

```
##     Manuf       Pop
## 0.8463644 0.4881390
```

```
print(apply(new_usair[,2:3],2,kurtosis))
```

```
##    Manuf      Pop
## 3.388033 3.182224
```

```
## Warning in 1:selected_columns: numerical expression has 2 elements: only the
## first used
```

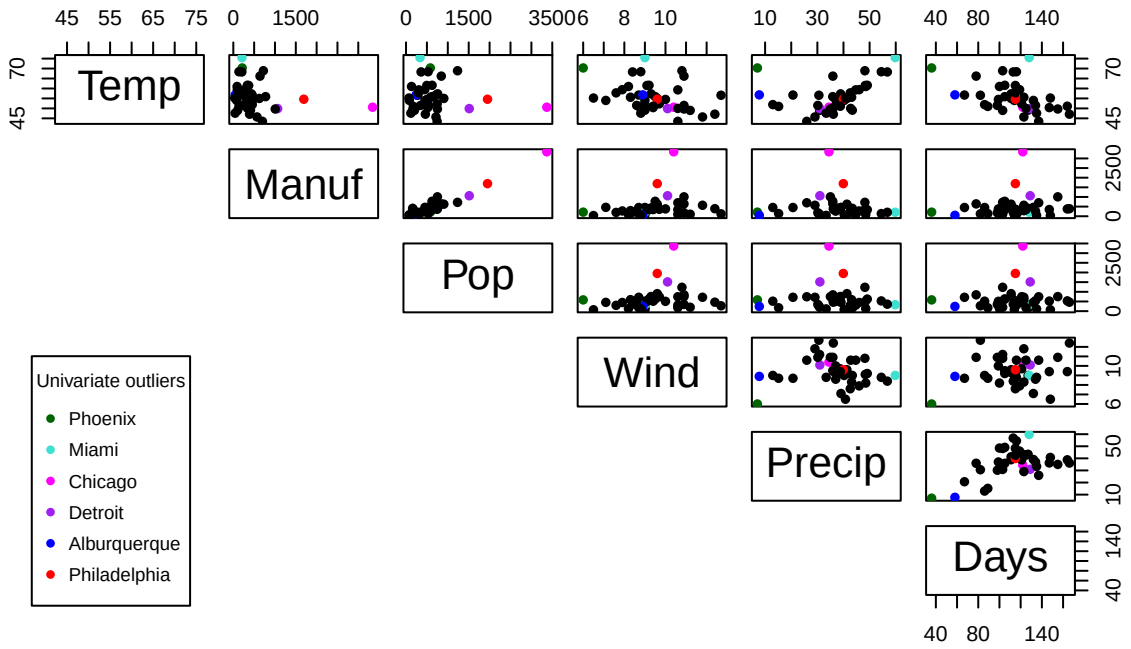



Removing observations 11 and 29 influences the normality of Manufacturing and Population. Manufacturing exhibits a slight rightward skew, but its right tail becomes lighter. Without outliers, the distribution of the variable appears gaussian, however, it's important to note that the presence of 3 outliers in 44 observations is highly unlikely under the assumption of a truly Gaussian variable. This discrepancy highlights a challenge in assuming normality in datasets with limited observations. However, for the remainder of our analysis, we assume that all variables are normally distributed.

## 1.4

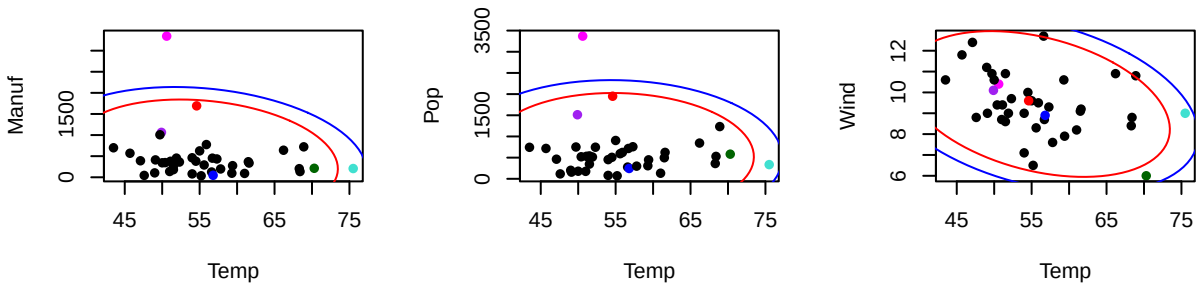
Looking at all the possible pairs of scatter plots we confirm our impressions about Chicago, Philadelphia and Phoenix: all these points seem to "break" the normality of the data. Detroit and Alburquerque do not seem to differ that much from the other points in the various bivariate representations.

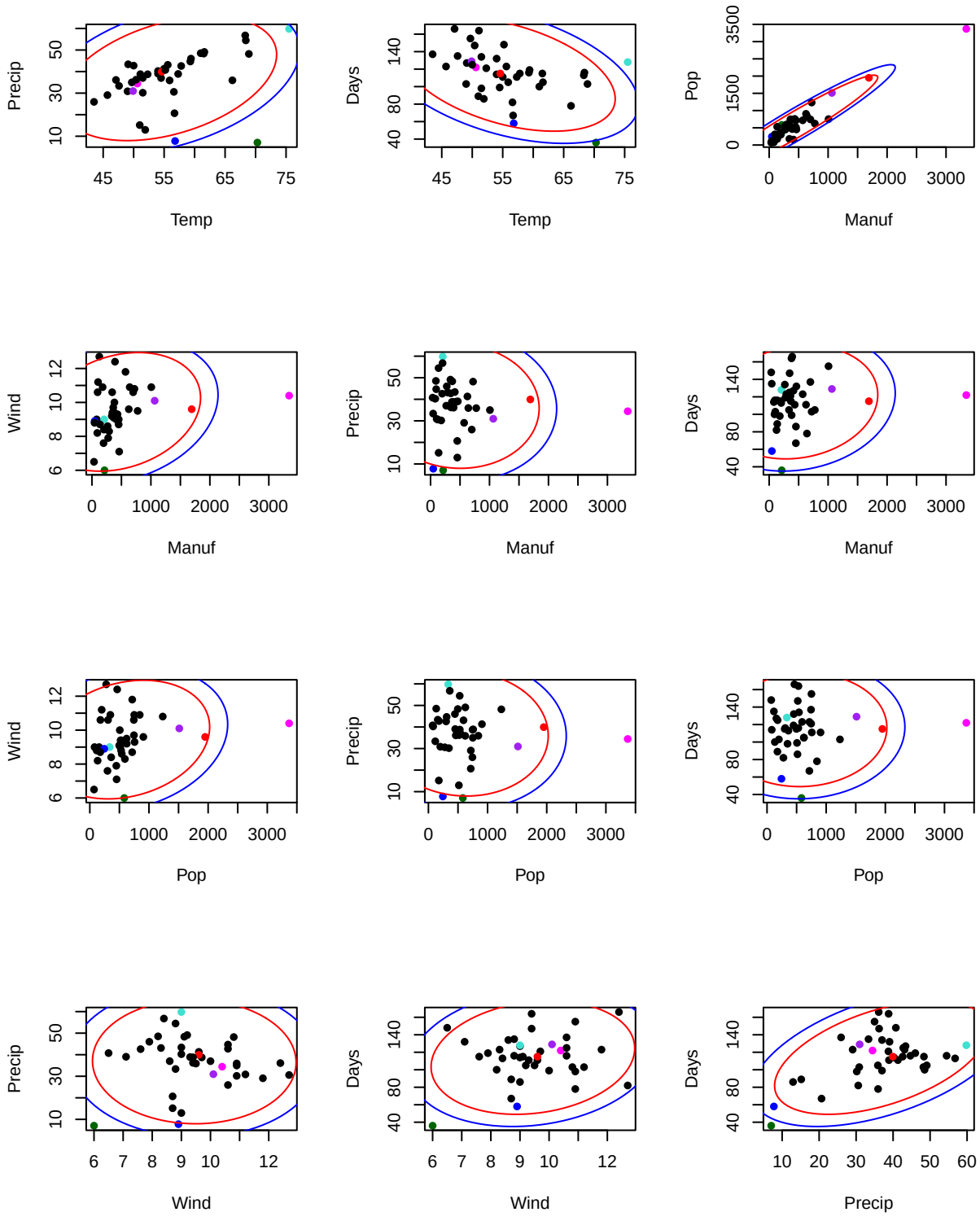
```
# Create a scatterplot matrix with colored points
pairs(usair, pch = 16, lower.panel = NULL, col=col.index)
# Enable clipping for legend outside plot area
par(xpd = TRUE)

# Add legend for univariate outliers
legend(x = "bottomleft",
       legend = rownames(usair[outliers,]),
       col = lookup,
       pch = 16,
       title = "Univariate outliers",
       cex = 0.6)
```



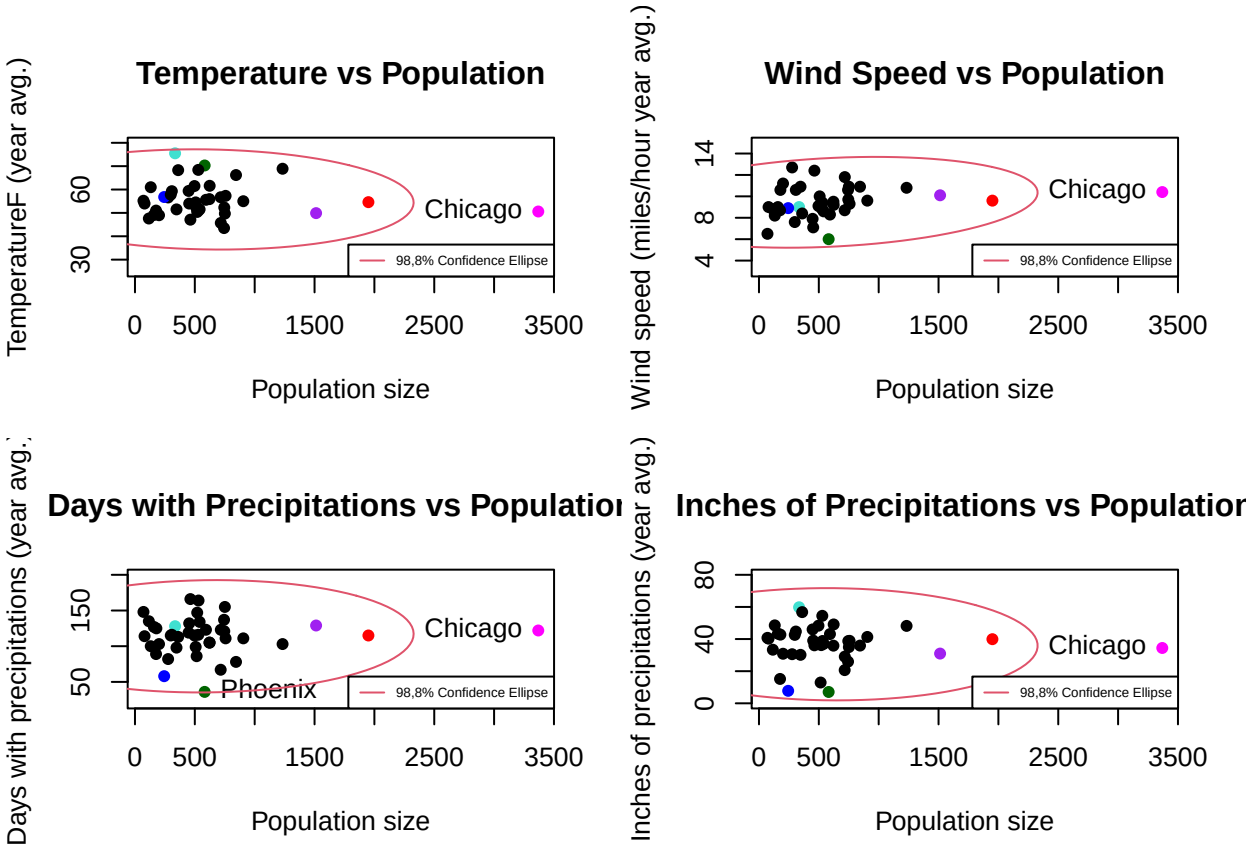

In order to have a more detailed analysis, we plot some of the scatter plots with the contours plot of a an appropriate Gaussian, and check which are the points that fall outside the ellipses, which are at levels 0.95 and 0.988.

```
S = cov(usair)
par(mfrow=c(1,3))

for(i in 1:5){
  for(j in (i+1):6){
    plot(usair[,j] ~ usair[,i], xlab=names(usair)[i], ylab=names(usair)[j], col=col.index,pch=16)
    lines(ellipse(x=S[c(i,j),c(i,j)], centre=colMeans(usair)[c(i,j)], level=0.95), col="red",lwd=1)
    lines(ellipse(x=S[c(i,j),c(i,j)], centre=colMeans(usair)[c(i,j)], level=0.988), col="blue",lwd=1)
}
}
```

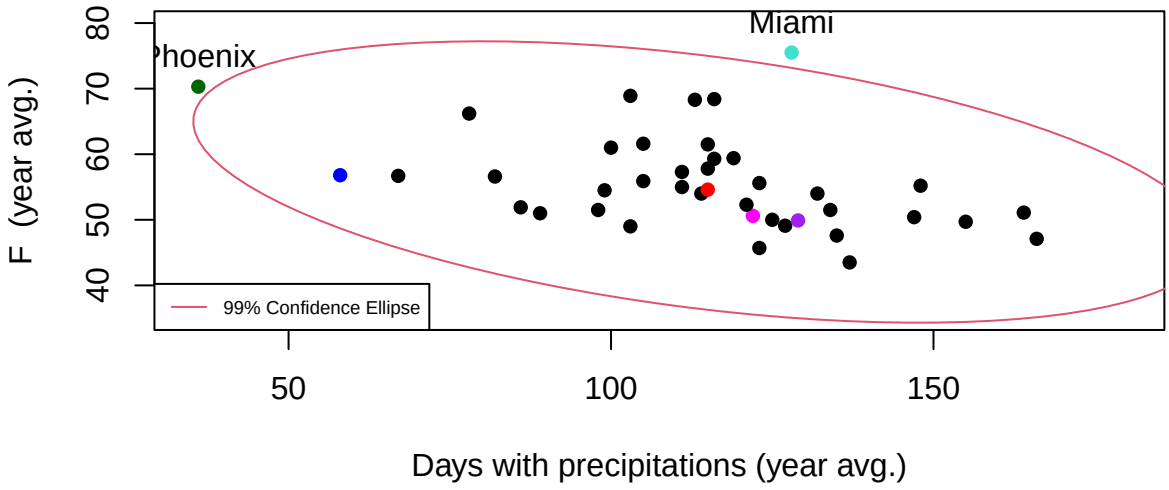





In particular we delve our analysis in some scatter plot for further investigation:

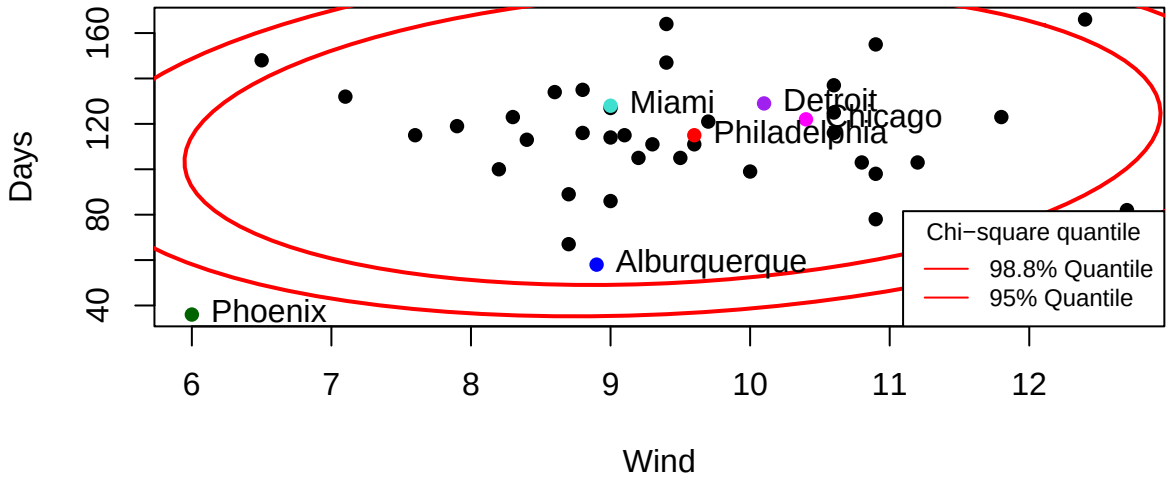


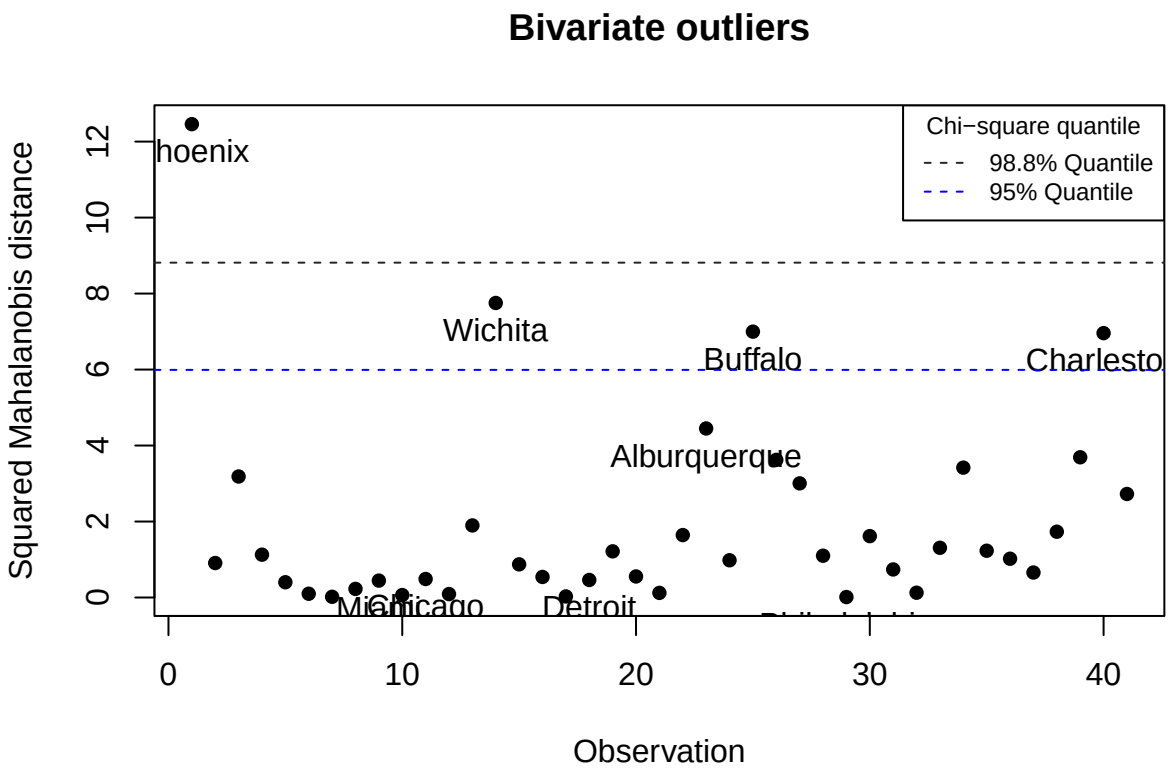

In the above scatterplot it appears that at level 0.988 Chicago is the only outlier, while the others fall inside the ellipses every time. Actually from this bivariate perspective, one could spot also the outlier Miami, which appears to be an outlier (at level 0.99) in the scatter-plot of the variables Temp and Days.




In the next scatter plots we investigate the relationship between Wind and Days and Days, we have identified

three bivariate outliers that were not apparent in the univariate analysis. To investigate their significance, we employ the Mahalanobis distance for calculating the indices of this outliers.

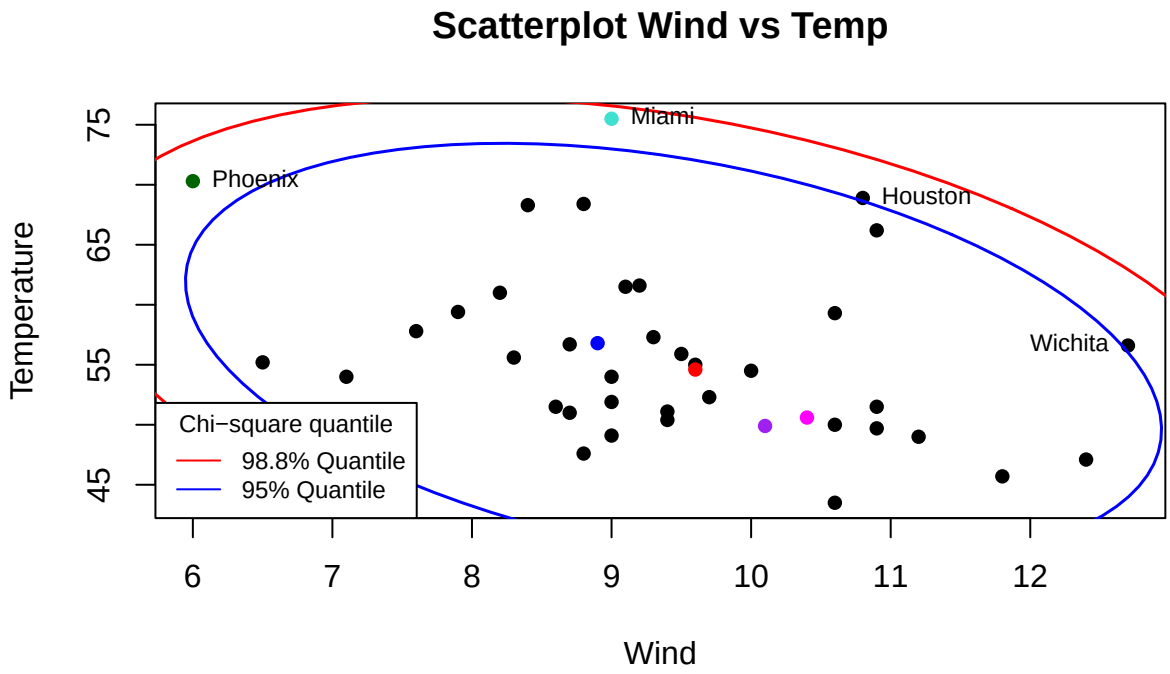



```
## Indices of outliers: 1 14 25 40 and names: Phoenix Wichita Buffalo Charleston
```

We also observed that in the scatterplot of Wind and Temp, there is another potential outlier, Houston, at the 95% confidence level.

```
plot(usair$Temp~usair$Wind, main="Scatterplot Wind vs Temp", xlab="Wind", ylab="Temperature",pch=16,col=
lines(ellipse(x=(S[c(4,1),c(4,1)]), centre=x_bar[c(4,1)],level=0.95),
      col="blue",lwd=1.5)
lines(ellipse(x=(S[c(4,1),c(4,1)]), centre=x_bar[c(4,1)],level=0.988),
      col="red",lwd=1.5)
text(x=usair$Wind[1],y=usair$Temp[1], pos=4, labels="Phoenix",cex=0.75)
text(x=usair$Wind[9],y=usair$Temp[9], pos=4, labels="Miami",cex=0.75)
text(x=usair$Wind[14],y=usair$Temp[14], pos=2, labels="Wichita",cex=0.75)
text(x=usair$Wind[35],y=usair$Temp[35], pos=4, labels="Houston",cex=0.75)
# Add legend for chi-square quantiles
legend(x = "bottomleft", legend = c("98.8% Quantile", "95% Quantile"),
       col = c("red","blue"), lty = 1, title = "Chi-square quantile", cex = 0.75,bg="white")
```




At the end of the analysis, it was observed that the outliers previously identified as univariate outliers in cities such as Phoenix, Chicago, and Miami remained outliers in the bivariate analysis at a level of 99% (using the continuity correction). Additionally, new outliers emerged (at 95% level) in the bivariate analysis, including Wichita, Buffalo, Houston, and Charleston. On the other hand, some cities that exhibited univariate outlier behaviour (observation 29, Philadelphia) did not display bivariate outlier behaviour.

## 1.5

Now we check the multivariate normality using the squared Mahalanobis distance: such quantity should be distributed as a chi-squared with 6 degrees of freedom, if we assume that our data come from a multivariate normal of dimension 6.

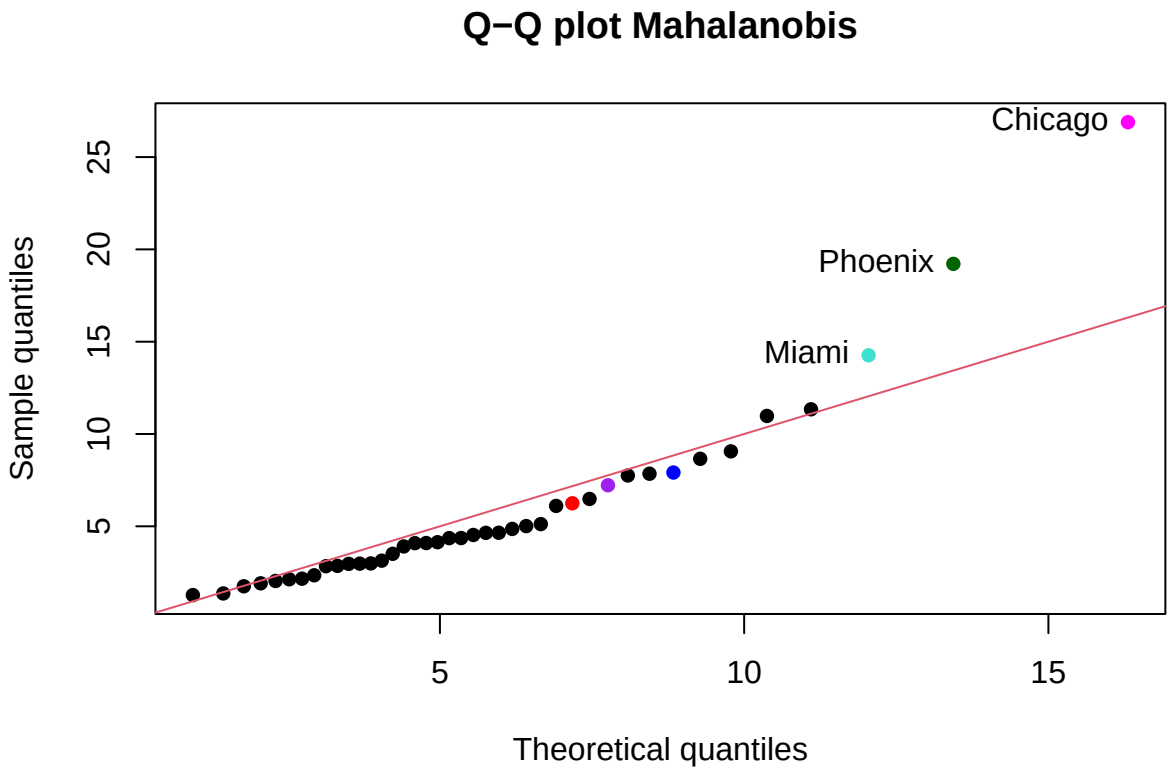
```
d <- mahalanobis(usair, center = x_bar, cov = cov(usair))
```

Therefore, to assess multivariate normality, we can construct a Q-Q plot. This plot compares the quantiles of the observed squared Mahalanobis distances against the expected quantiles of a chi-square distribution with 6 degrees of freedom.

```
d2 <- mahalanobis(usair,colMeans(usair),cov(usair))
qqplot(qchisq(ppoints(d2),df=p),d2,xlab="Theoretical quantiles",ylab="Sample quantiles",pch=16,main="Q-
which(d2>14)
```

```
## Phoenix   Miami Chicago
##       1       9      11
```

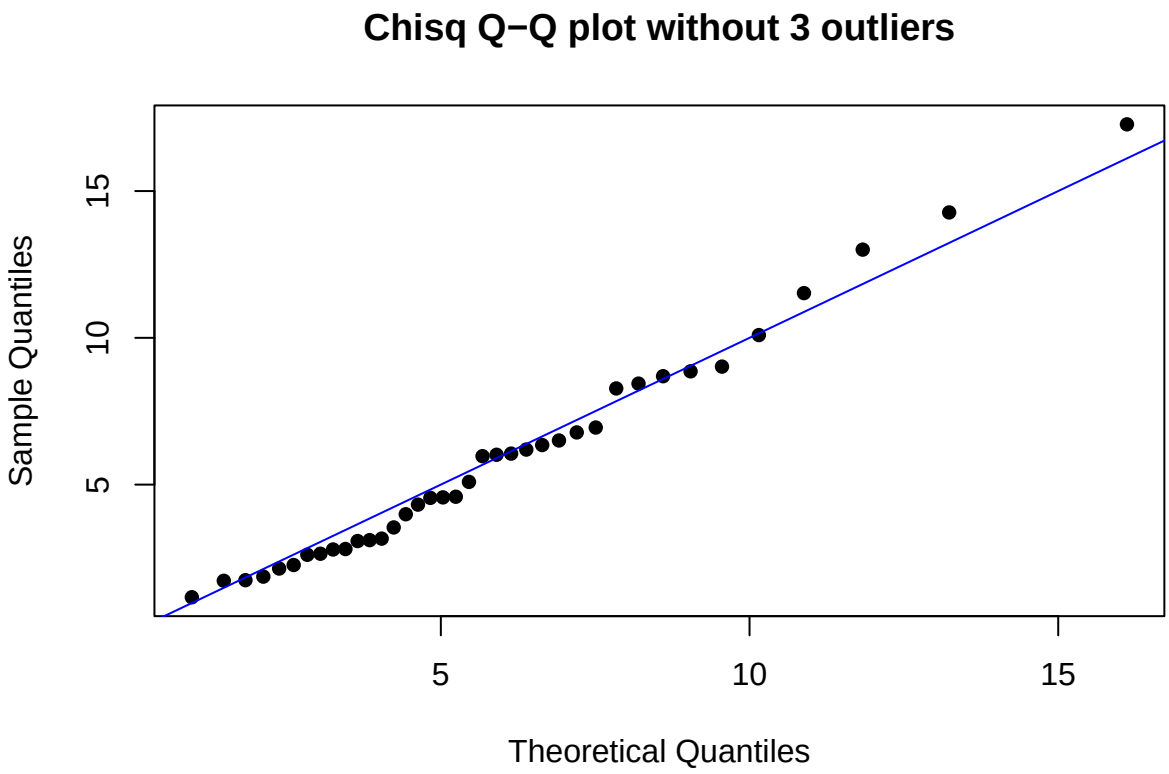
```
text((qchisq(ppoints(d2),df=p))[39:41],sort(d2)[39:41],names(sort(d2)[39:41]),pos=2)
abline(0,1,col=2)
```




In general the observations seem to be distributed as a $\chi^2$, except for the observations in the right tail, that are more distant from the theoretical quantiles and are Chicago, Phoenix and Miami.

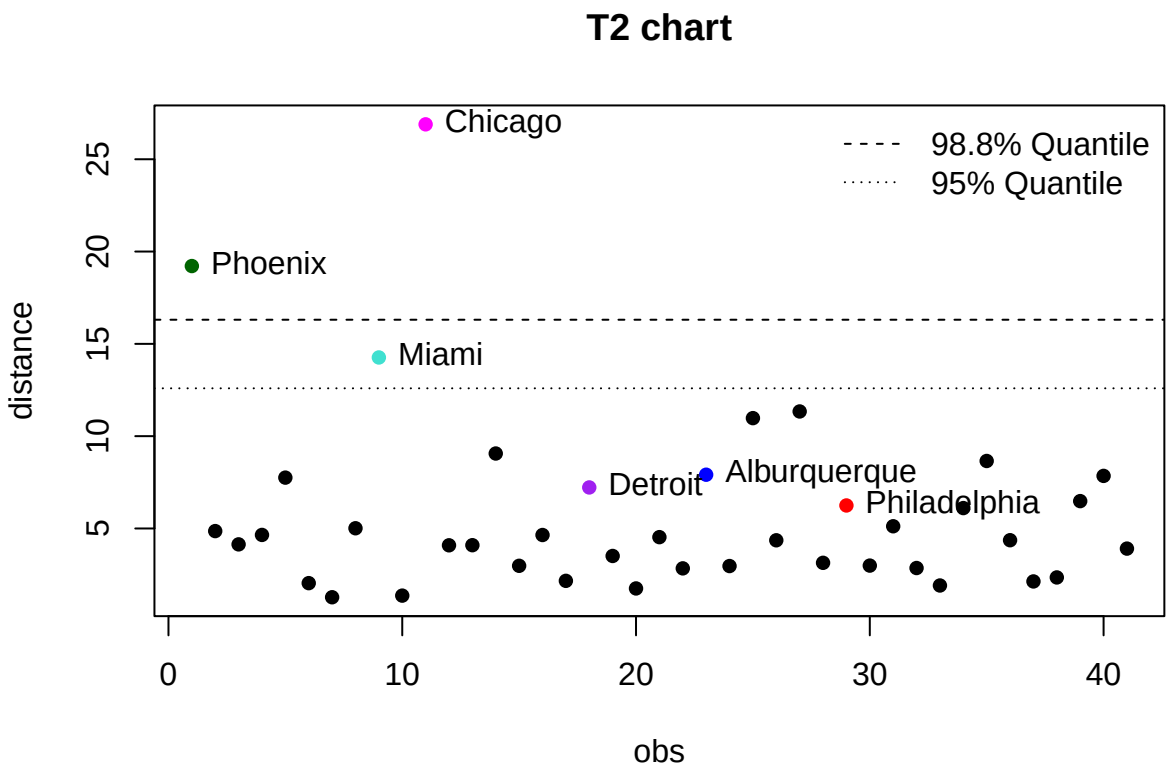
By removing the outliers, we find a multivariate normality in the Q-Q plot.

```
new_usair <- usair[-c(1,9, 11), ]
d_new <- mahalanobis(new_usair, center = colMeans(new_usair), cov = cov(new_usair))
# Generate a chi-square Q-Q plot of Mahalanobis distances
plot(qchisq(ppoints(d_new),df=p),sort(d_new),main="Chisq Q-Q plot without 3 outliers",
xlab="Theoretical Quantiles",ylab="Sample Quantiles",pch=16)

# Add diagonal reference line
abline(0, 1, col = "blue",lwd=1)
```




## 1.6

To further investigate the potential multivariate outliers identified in the Q-Q plot (observations 1, 9, and 11), we plot the squared Mahalanobis distances for each observation with critical values from the chi-square distribution with 6 degrees of freedom $\chi^2_6$. We can determine whether the observations are statistically unlikely to belong to the assumed multivariate normal distribution.




We observe that the only two values falling outside the 99th percentile (using the continuity correction) of the $\chi^2$ distribution are Phoenix and Chicago, while the other univariate outliers are considerably lower. Another noteworthy observation is Miami, which exceeds the 95th percentile and could potentially be considered an outlier. Thus, it is notable that the cities identified as outliers in the univariate analysis are not necessarily outliers in the multivariate context. Conversely, some cities that were not identified as outliers for any individual variable were only flagged after considering multiple variables simultaneously, as observed in the bivariate analysis. Specifically, Miami was not deemed an "extreme" outlier in the univariate analysis (in comparison of other that are marked in the figure) but persists as an outlier in the multivariate analysis.

In conclusion, our analysis reveals a mixed pattern in outlier consistency. While some outliers identified in bivariate analysis do not persist as outliers in multivariate analysis, others consistently maintain outlier status across all analyses conducted.

## Exercise 2

First we compute the covariance matrix of $X$:

$$S_x = \text{Var}(X) = V\text{Var}(Z)V^T + \text{Var}(\varepsilon) = \sigma_z^2 V I_q V^T + \sigma^2 I_p$$

Where we used the independence of $Z$ and $\varepsilon$, then, the total variance is given by

$$\text{tr}(S_x) = \sigma_z^2 \text{tr}(VV^T) + \sigma^2 \text{tr}(I_p) = \sigma_z^2 \text{tr}(V^T V) + \sigma^2 p = \sigma_z^2 q + \sigma^2 p$$

Where we used the properties of the trace and the fact that $V^T V = I_q$ since the columns of $V$ are orthogonal and with unit norm.

Now notice that $VV^T$ is a symmetric $p \times p$ matrix with rank $q < p$, thus we can perform the spectral decomposition and we know that only $q$ eigenvalues will be not null since the rank of the matrix is $q$:

$$VV^T = PDP^T$$

In particular notice that since

$$(VV^T)V = V$$

then by definition, each column of $V$ is an eigenvector for $VV^T$ and 1 is the relative eigenvalue, which therefore has multiplicity $q$. Thus the matrix $D$ of the spectral decomposition is defined as

$$D = \begin{pmatrix} I_q & O_{q,q-p} \\ O_{q-p,q} & O_{q-p,q-p} \end{pmatrix}$$

Therefore, since the matrix $P$ in the spectral decomposition can be chosen to be orthogonal we get that

$$P^T S_x P = \sigma_z^2 P^T (VV^T) P + \sigma^2 P^T P = \sigma_z^2 D + \sigma^2 I_p = \begin{pmatrix} (\sigma_z^2 + \sigma^2) I_q & O_{q,q-p} \\ O_{q-p,q} & \sigma^2 I_{q-p} \end{pmatrix}$$

Thus it is clear that the biggest $q$ eigenvalues are $\sigma_z^2 + \sigma^2$, therefore we get that

$$\frac{\sum_{i=1}^q \lambda_i}{\text{tr}(S_x)} = \frac{(\sigma_z^2 + \sigma^2)q}{\sigma_z^2 q + \sigma^2 p} \geq 0.8 \implies 5q\sigma_z^2 + 5q\sigma^2 \geq 4q\sigma_z^2 + 4p\sigma^2 \implies q\sigma_z^2 \geq (4p - 5q)\sigma^2$$

$$(1+\delta)\sigma^2 \geq \frac{4p - 5q}{q}\sigma^2 \implies \delta \geq \frac{4p - 6q}{q}$$

Notice that, since $\delta > -1$ we get that if $q > \frac{4}{5}p$, then the condition is satisfied for all $\delta$.

### 2.2

In our problem we have that:

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \quad \boldsymbol{\Sigma} = \frac{1}{9}\begin{bmatrix} 5 & -4 & 2 \\ -4 & 5 & 2 \\ 2 & 2 & 8 \end{bmatrix} + \frac{1}{3}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \frac{1}{9}\begin{bmatrix} 8 & -4 & 2 \\ -4 & 8 & 2 \\ 2 & 2 & 11 \end{bmatrix}$$

Let define

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

In the representation of the second matrix, the initial matrix can be denoted as follows:

$$\boldsymbol{\Sigma} = \frac{1}{9}\left[\begin{array}{cc|c} 8 & -4 & 2 \\ -4 & 8 & 2 \\ \hline 2 & 2 & 11 \end{array}\right] = \left[\begin{array}{cc|c} \frac{8}{9} & -\frac{4}{9} & \frac{2}{9} \\ -\frac{4}{9} & \frac{8}{9} & \frac{2}{9} \\ \hline \frac{2}{9} & \frac{2}{9} & \frac{11}{9} \end{array}\right]$$

With

$$\Sigma_{11} = \begin{bmatrix} \frac{8}{9} & -\frac{4}{9} \\ -\frac{4}{9} & \frac{8}{9} \end{bmatrix}, \quad \Sigma_{12} = \begin{bmatrix} \frac{2}{9} \\ \frac{2}{9} \end{bmatrix}, \quad \Sigma_{21} = \begin{bmatrix} \frac{2}{9} & \frac{2}{9} \end{bmatrix}, \quad \Sigma_{22} = \frac{11}{9}$$

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \boldsymbol{\mu}_2 = [1].$$

Recall that if $X = (X_1, X_2) \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

Then the conditional distribution of $X_1$ given $X_2 = x_2$ is a multivariate normal of dimension $p - q$:

$$X_1|X_2 = x_2 \sim \mathcal{N}_{p-q}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(x_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$$

Therefore, we can compute the parameters of the distributions $X_1, X_2|X_3 = -1$:

Now let calculate the parameter of the conditional distirbution:

$$\boldsymbol{\mu}_c = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12} \left(\boldsymbol{\Sigma}_{22}\right)^{-1} (-1 - \boldsymbol{\mu}_2)$$

$$\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \left(\boldsymbol{\Sigma}_{22}\right)^{-1} \boldsymbol{\Sigma}_{21}$$

Substituting the matrix above we get:

$$\boldsymbol{\mu}_c = \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} \frac{2}{9} \\ \frac{2}{9} \end{bmatrix} \left(\frac{11}{9}\right)^{-1} (-1 - 1) = \begin{bmatrix} \frac{7}{11} \\ \frac{7}{11} \end{bmatrix}$$

$$\boldsymbol{\Sigma}_c = \begin{bmatrix} \frac{8}{9} & -\frac{4}{9} \\ -\frac{4}{9} & \frac{8}{9} \end{bmatrix} - \begin{bmatrix} \frac{2}{9} \\ \frac{2}{9} \end{bmatrix} \left(\frac{11}{9}\right)^{-1} \begin{bmatrix} \frac{2}{9} & \frac{2}{9} \end{bmatrix} = \begin{bmatrix} \frac{28}{33} & -\frac{16}{33} \\ -\frac{16}{33} & \frac{28}{33} \end{bmatrix}$$

In the 2-dimensional space of $x = (x_1, x_2)$ by setting the constant $c$ such that the ellipse contains 0.95 probability with respect to the joint distribution of $X_1, X_2|X_3 = -1$, the equation of the ellipse is:

$$(x - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1} (x - \boldsymbol{\mu}_c) = c^2$$

with $c^2 = \chi^2_{0.95}$.

After some calculation we get:

$$\frac{7}{4}\left(x_1 - \frac{7}{11}\right)^2 + 2\left(x_1 - \frac{7}{11}\right)\left(x_2 - \frac{7}{11}\right) + \frac{7}{4}\left(x_2 - \frac{7}{11}\right)^2 = \chi^2_{0.95} = 5.991 \sim 6 \tag{1}$$

That is:

$$\frac{7}{4}\left(x_1^2 + x_2^2\right) + 2x_1x_2 - \frac{7}{2}\left(x_1 + x_2\right) - \frac{83}{22} = 0 \tag{2}$$

A Geogebra visualization of the ellipse in the figure 1.

Recall that if $X = (X_1, X_2) \sim \mathcal{N}_p(\mu, \Sigma)$ with

$$\mu = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

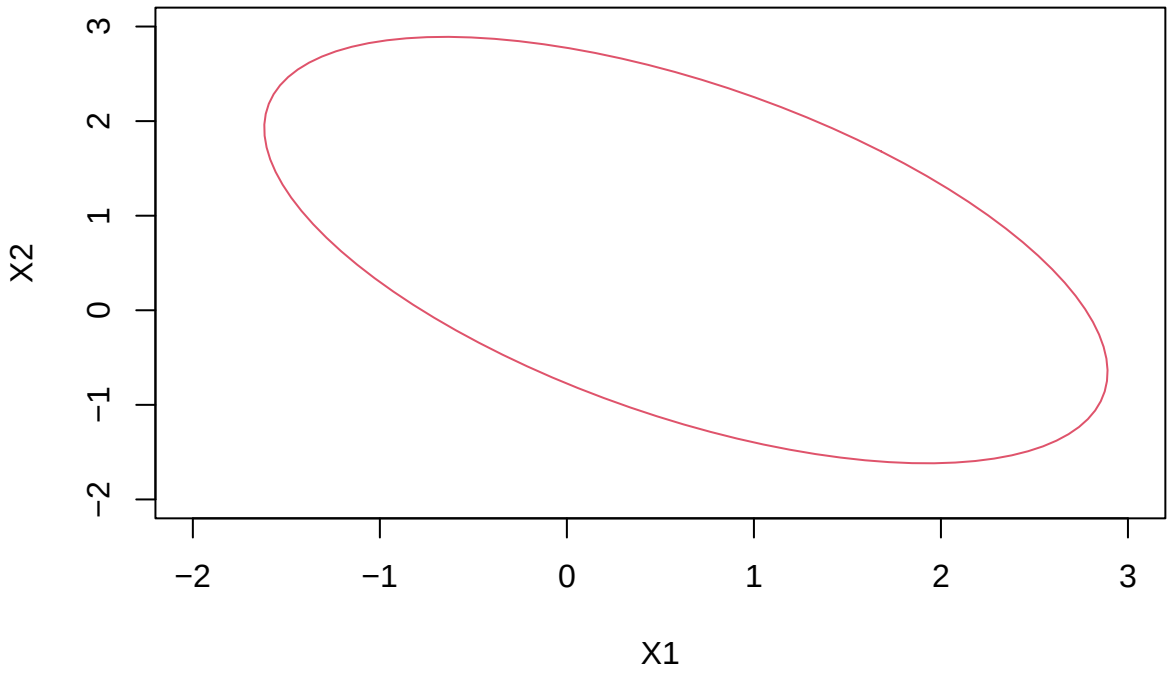
Then the conditional distribution of $X_1$ given $X_2 = x_2$ is a multivariate normal of dimension $p - q$:

$$X_1|X_2 = x_2 \sim \mathcal{N}_{p-q}(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(x_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$$

The same calculation are done through the software:

```
V=matrix(c(-1/3,2/3,2/3,2/3,-1/3,2/3),nrow=3,ncol=2)
I3=matrix(c(1,0,0,0,1,0,0,0,1),nrow=3,ncol=3)
Sig=V%*%t(V)+I3/3
mu=c(1,1,1)
mu1=mu[1:2]
mu2=mu[3]
Sig11=Sig[1:2,1:2]
Sig12=Sig[1:2,3]
Sig21=Sig[3,1:2]
Sig22=Sig[3,3]
muc=mu1+Sig12%*%solve(Sig22)%*%(-1-mu2)
Sigc=Sig11-Sig12%*%solve(Sig22)%*%Sig21
```

Then we can represent the ellipse:



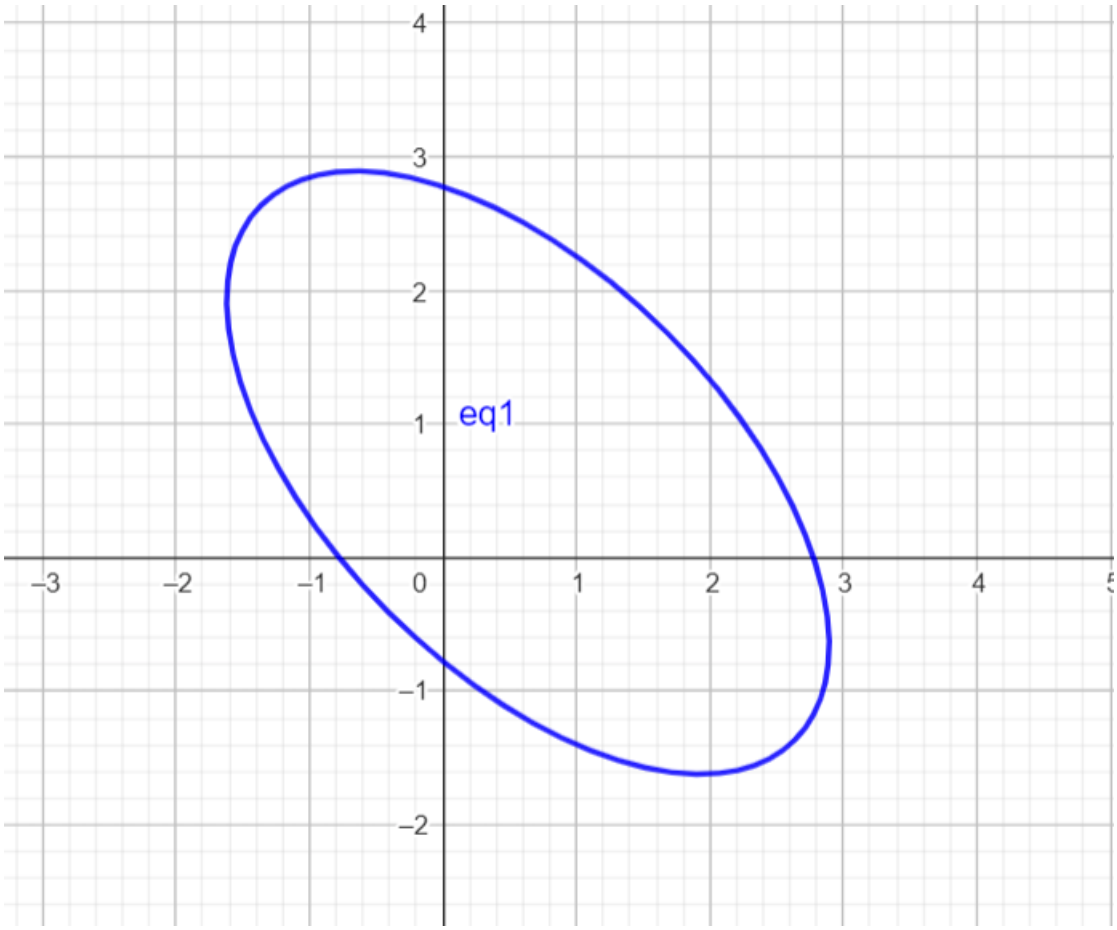



Figure 1: Representation using geogebra

## 3: PCA

# 3.1

We import the data into the object `pendigits` and we remove the last column (the one with the `digit` attribute), since it won't be of use for now.

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

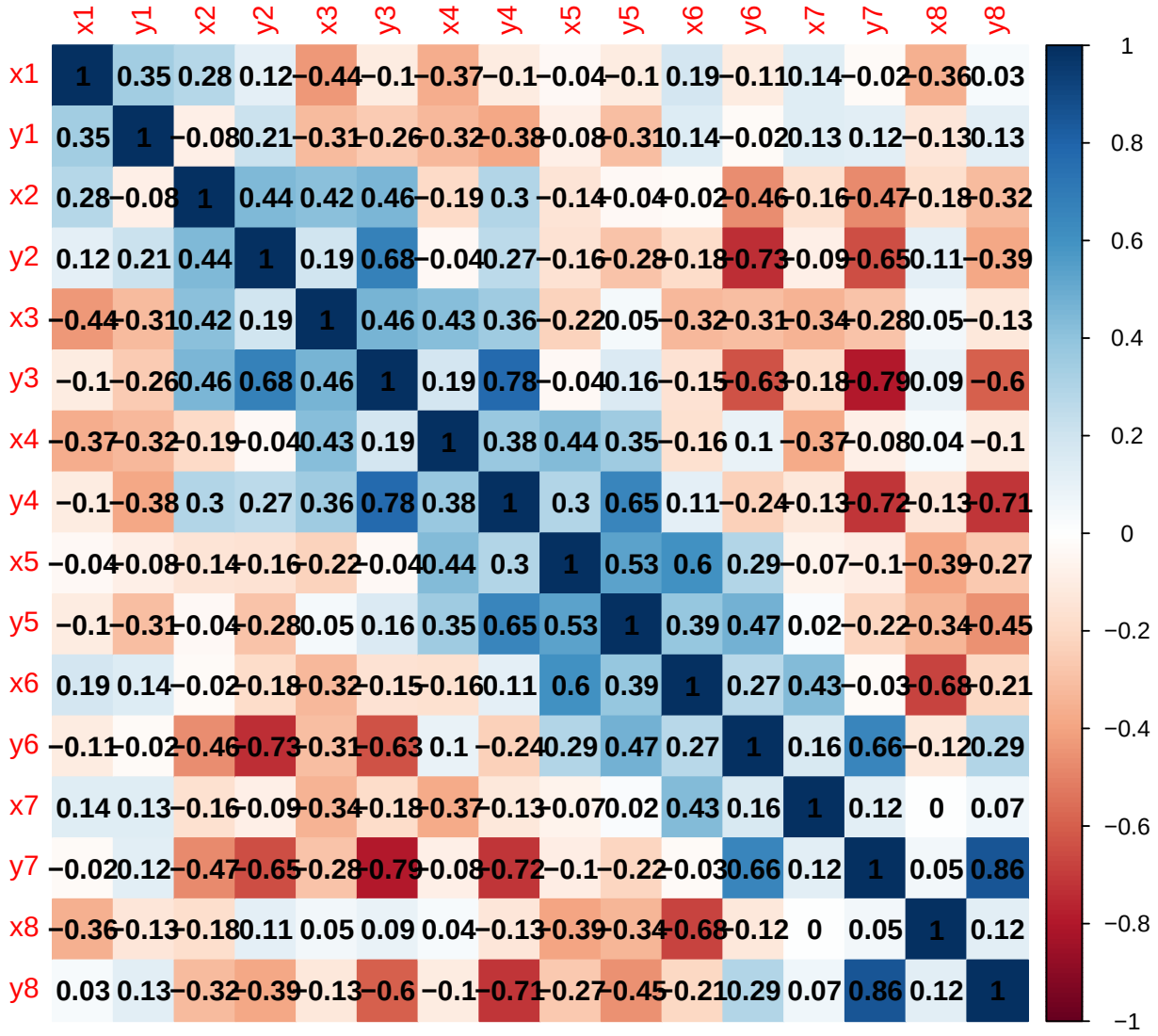
| x1 | y1 | x2 | y2 | x3 | y3 | x4 | y4 | x5 | y5 | x6 | y6 | x7 | y7 | x8 | y8 | digit |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 47 | 100 | 27 | 81 | 57 | 37 | 26 | 0 | 0 | 23 | 56 | 53 | 100 | 90 | 40 | 98 | 8 |
| 0 | 89 | 27 | 100 | 42 | 75 | 29 | 45 | 15 | 15 | 37 | 0 | 69 | 2 | 100 | 6 | 2 |
| 0 | 57 | 31 | 68 | 72 | 90 | 100 | 100 | 76 | 75 | 50 | 51 | 28 | 25 | 16 | 0 | 1 |
| 0 | 100 | 7 | 92 | 5 | 68 | 19 | 45 | 86 | 34 | 100 | 45 | 74 | 23 | 67 | 0 | 4 |
| 0 | 67 | 49 | 83 | 100 | 100 | 81 | 80 | 60 | 60 | 40 | 40 | 33 | 20 | 47 | 0 | 1 |
| 100 | 100 | 88 | 99 | 49 | 74 | 17 | 47 | 0 | 16 | 37 | 0 | 73 | 16 | 20 | 20 | 6 |

```
pend=pendigits[,-ncol(pendigits)]
```

We calculate multivariate mean, variance matrix and correlation matrix. Due to the number of variables it is difficult to extrapolate information from the correlation matrix. What we see is that the height values tend to have a strong positive (or negative) correlation with other y values (for example, the `y3` variable has a correlation of 0.78 with the variable `y4` and `y7` and `y8` have an even higher correlation of 0.857).

```
bar.x=colMeans(pend) #finding multivariate mean
n<-nrow(pend)
S<-var(pend) #finding variance matrix
R<-cor(pend) #finding correlation matrix
```

```
corrplot(R,method="color",
         type="full", order="original",
         title="Visualization of correlation matrix",
         addCoef.col = "black",
         insig = "pch",mar=c(0,0,1,0))
```




This visualization of the data helps in showing how the correlation matrix has a sort of bi-diagonal nature, with consecutive coordinates tending to be more correlated (especially in the $y$ dimension) and coordinates captured in further away moments tending to be less or negatively correlated.

We now perform a principal components analysis on the data. To improve readability and interpretation we set the option `scale` to true so that the variables are scaled to have unit variance before the analysis:

```
pendigits.pca=prcomp(pend, scale. =T)
as.data.frame(summary(pendigits.pca)$importance)
```

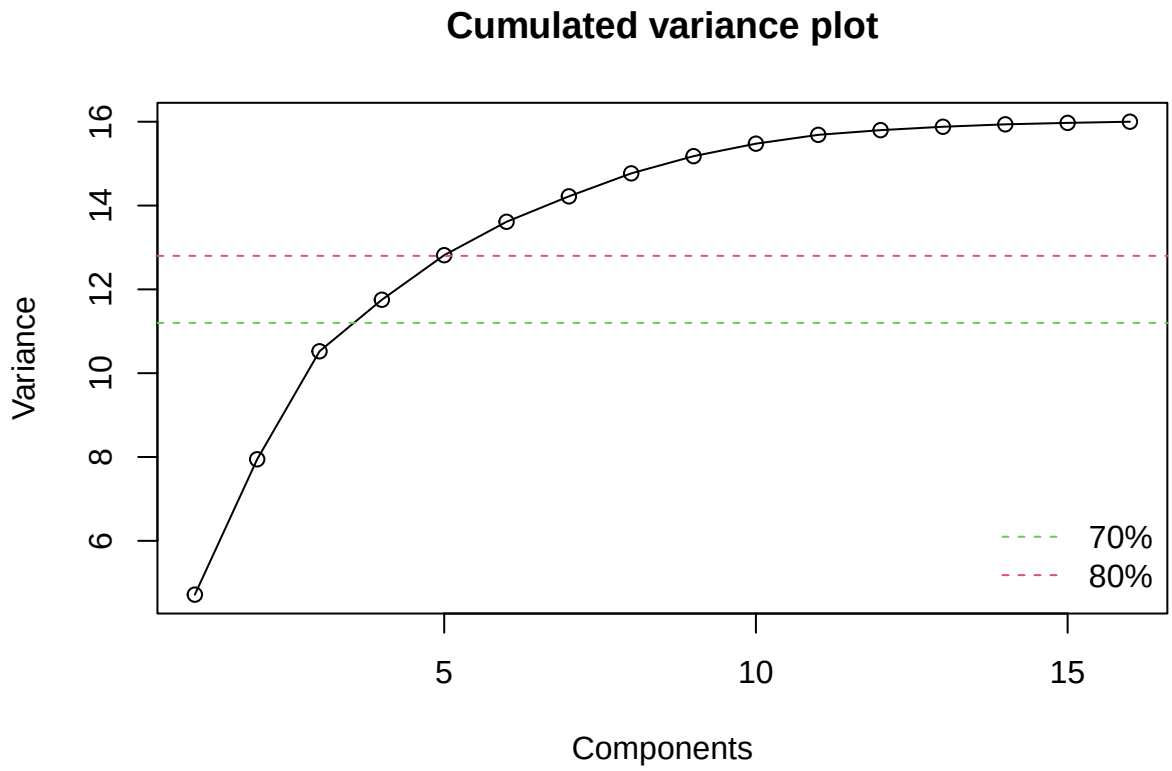
```
##                             PC1       PC2       PC3       PC4       PC5        PC6
## Standard deviation     2.17177 1.796973 1.605243 1.108937 1.031067 0.8929386
## Proportion of Variance 0.29479 0.201820 0.161050 0.076860 0.066440 0.0498300
## Cumulative Proportion  0.29479 0.496610 0.657660 0.734520 0.800960 0.8507900
##                               PC7        PC8        PC9       PC10       PC11
```

```
## Standard deviation     0.7788784 0.7404401 0.6409556 0.5461052 0.4588166
## Proportion of Variance 0.0379200 0.0342700 0.0256800 0.0186400 0.0131600
## Cumulative Proportion  0.8887100 0.9229700 0.9486500 0.9672900 0.9804500
##                            PC12      PC13      PC14      PC15      PC16
## Standard deviation     0.3351073 0.2836918 0.2408363 0.1851067 0.1667316
## Proportion of Variance 0.0070200 0.0050300 0.0036300 0.0021400 0.0017400
## Cumulative Proportion  0.9874700 0.9925000 0.9961200 0.9982600 1.0000000
```

```
plot(x=1:16,y=cumsum(pendigits.pca$sdev^2),type='ol',ylab="Variance",xlab = "Components",main="Cumulate
```

```
## Warning in plot.xy(xy, type, ...): plot type 'ol' will be truncated to first
## character
```

```
abline(h=0.7*16,col=3,lty=2)
abline(h=0.8*16,col=2,lty=2)
legend("bottomright", legend = c("70%","80%"), lty = 2, col = c(3,2), bty = "n")
```

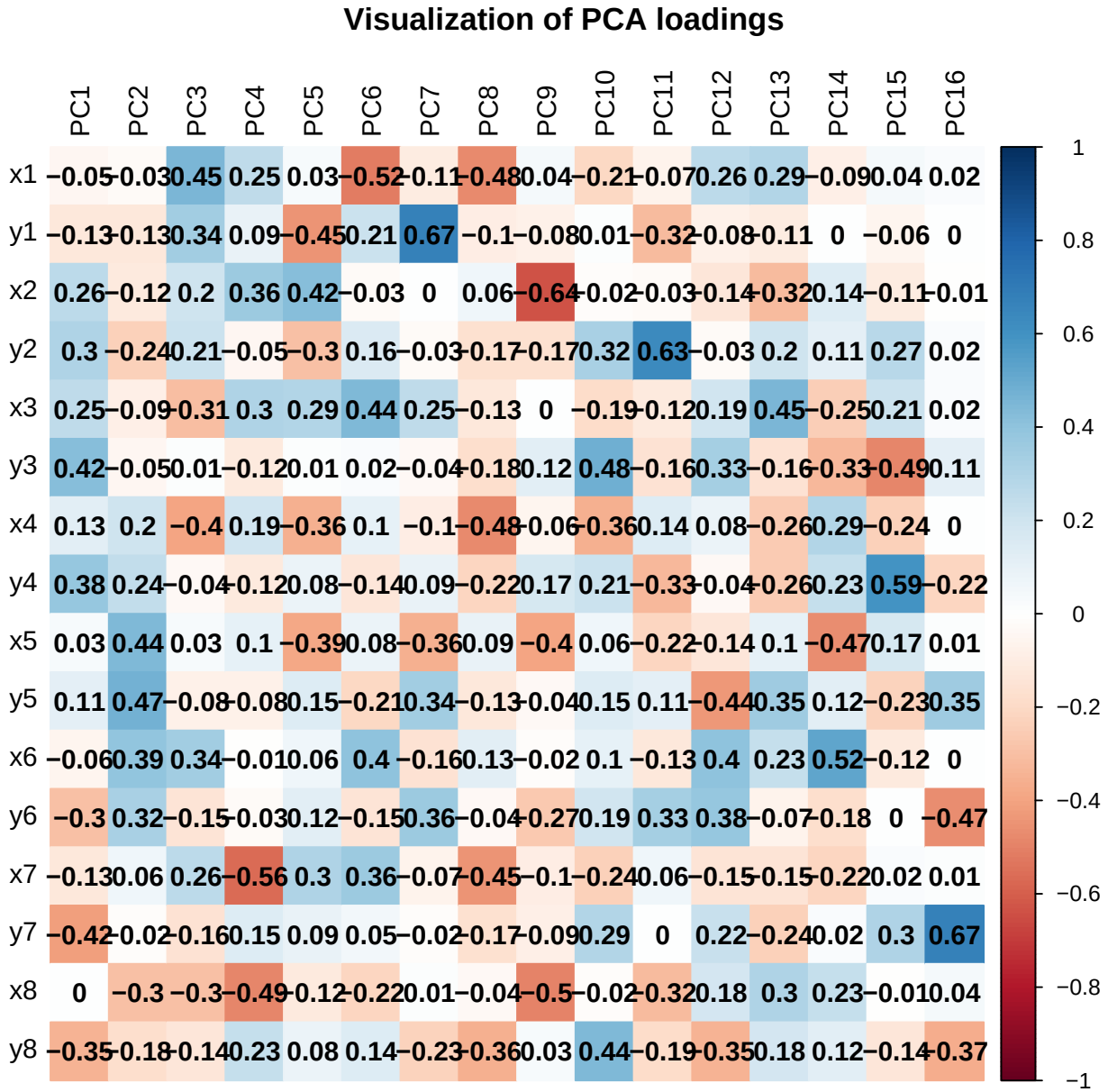



The summary reports the standard deviation of each principal component as well as the individual and cumulative proportion of the explained variance. From this last information we can already perform a choice regarding dimension reduction by retaining the principal components that cumulatively explain at least 80% of the total variance. On this criterion's basis we can then choose to retain the first 5 principal components, with the sixth one explaining $\approx 85\%$ of the total variance. If we choose to accept that the cumulative explanation by the principal components is at least 70% of the total variance, then we may decide to retain four principal components.

We then bring our attention to the `rotation` element of `pendigits.pca`. We can visualize how much each variable influences each principal component (that is, the measure of its loadings). Finding an interpretation of the relevance of each pen coordinate for each principal component is not a trivial task: it is worth noting,

though, that for the first principal component the biggest (positive) loadings seem to belong to the coordinates of early pen strokes, while for the second principal component the same can be said for the central pen strokes.

```
corrplot(round(pendigits.pca$rotation, 2), method = "color"
    , tl.col = "black", addCoef.col = "black", number.cex = 1,title = "Visualization of PCA loadings",ma
```




We investigate further the variable reduction. Another criterion to choose how many components we should retain is by looking at the associated eigenvalues. Since the aim of variable reduction is to have enough components to explain as much as possible of the variance of the original variables we can select the eigenvalues that score above the average value of the eigenvalues associated to the principal component. Since we normalized the data before analyzing, it is enough to check for eigenvalues greater than 1.

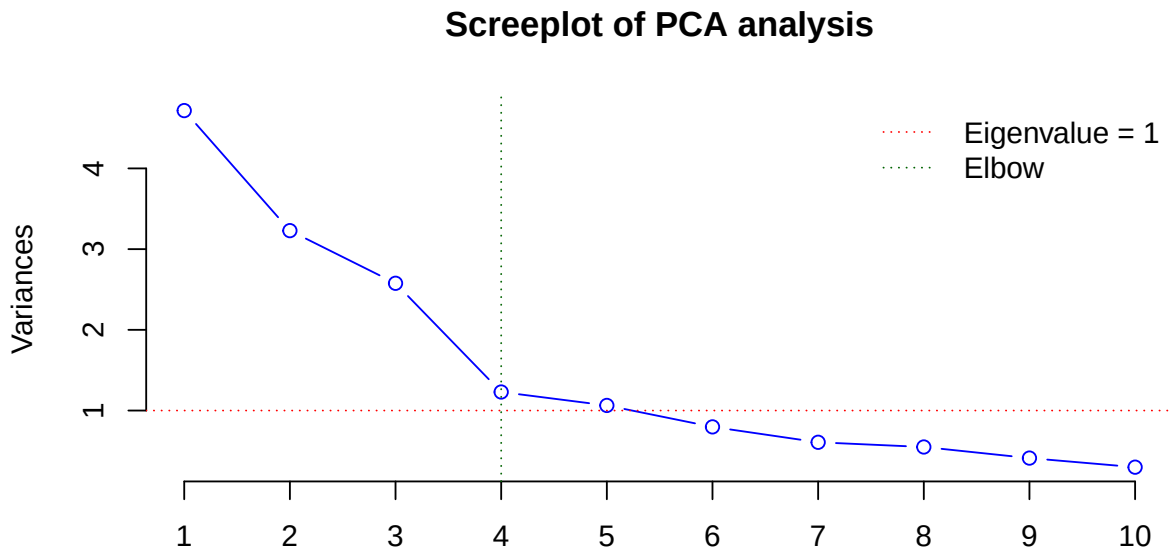
```
pendigits.pca$sdev[pendigits.pca$sdev^2>=1]
```

```
## [1] 2.171770 1.796973 1.605243 1.108937 1.031067
```

This criterion agrees with the previous one in retaining 5 principal components.

Another criterion is to look for "elbow" or sudden value drops between eigenvalues. The screeplot of the data allows us to find a (not so pronounced) elbow at the fourth eigenvalue. The screeplot criterion suggests choosing $k$ eigenvalues, where the elbows happens at eigenvalue $k+1$. The screeplot also visualizes the a red line at variance 1, to help visualize the threshold below which, according to the previous criterion, we should stop retaining variables.

```
screeplot(pendigits.pca,type = "l",col="blue",main = "Screeplot of PCA analysis")
abline(h = 1, col = "red", lty = 3)
abline(v = 4, col = "darkgreen", lty = 9)
legend("topright", legend = c("Eigenvalue = 1", "Elbow"),
       col = c("red", "darkgreen"), lty = c(3, 9), bty = "n")
```




In this case, the elbow method suggests choosing 3 principal components. This criterion is not in agreement with the previous two, but it is important noting that the "elbow" that we can find in this screeplot is not particularly obvious, so in absence of large evident drops in value the criterion may partially fail.

So, in the end, the three criteria do not agree. Since two out of three propose to retain 5 variables, we find reasonable accepting this as an answer (which is still a great improvement from the initial 16 variables).

## 3.2

We plot the first three principal component against each other. For readability's sake we chose to color code the observation in the following way:

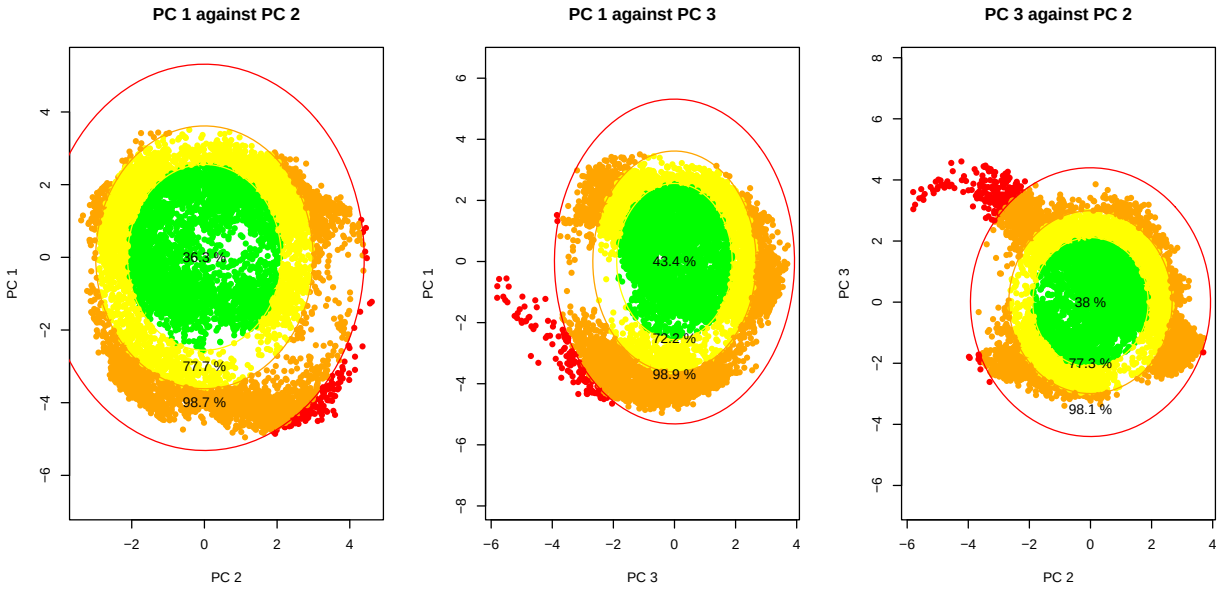
- as green the observaions that fall inside the 50% confidence level ellipse;

- as yellow the observations that fall between the 50% and the 75% confidence level ellipse;
- as orange the observations that fall between the 75% the 95% confidence level ellipse;
- as red the observations that fall outside these ellipses;

We then pasted the actual cumulative percentage of the data over the plot to compare actual and theoretical proportions.

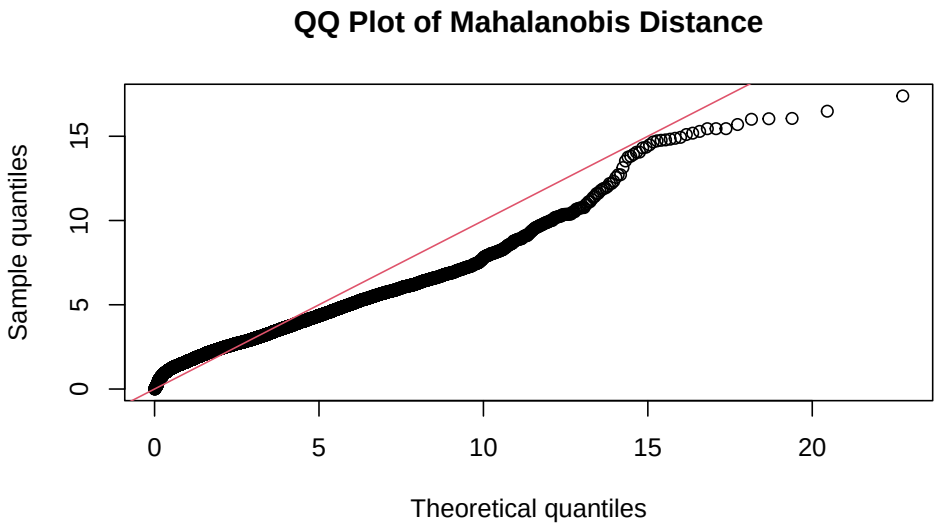
```
pendigits.pca3=pendigits.pca$x[,1:3]
par(mfrow=c(1,3))
d1=mahalanobis(pendigits.pca3[,c(2,1)],center = c(0,0),cov=var(pendigits.pca3[,c(2,1)]))
index1 <- which(d1>qchisq(0.95,2))
index1.2 <- which(d1>qchisq(0.75,2)&d1<=qchisq(0.95,2))
index1.3 <- which(d1>qchisq(0.5,2)&d1<=qchisq(0.75,2))
col.index1 <- rep("green",nrow(pendigits.pca3))
col.index1[index1] <- "red"
col.index1[index1.2] <- "orange"
col.index1[index1.3] <- "yellow"
p1 <- round(length(index1)/nrow(pendigits.pca3),3)
p1.2 <- round(length(index1.2)/nrow(pendigits.pca3),3)
p1.3 <- round(length(index1.3)/nrow(pendigits.pca3),3)

plot(pendigits.pca3[,1]~pendigits.pca3[,2],asp=1,col=col.index1,pch=16,xlab = "PC 2",ylab = "PC 1",main=
lines(ellipse(x=var(pendigits.pca3[,c(2,1)]),center=c(0,0),level=0.95),col="red")
lines(ellipse(x=var(pendigits.pca3[,c(2,1)]),center=c(0,0),level=0.75),col="orange")
lines(ellipse(x=var(pendigits.pca3[,c(2,1)]),center=c(0,0),level=0.5),col="yellow")
text(0,0,paste((1-p1-p1.2-p1.3)*100,"%"))#should be 50%
text(0,-3,paste((1-p1.2-p1)*100,"%"))#should be 75%
text(0,-4,paste((1-p1)*100,"%"))#should be 95%

d2=mahalanobis(pendigits.pca3[,c(3,1)],center = c(0,0),cov=var(pendigits.pca3[,c(3,1)]))
index2 <- which(d2>qchisq(0.95,2))
index2.2 <- which(d2>qchisq(0.75,2)&d2<=qchisq(0.95,2))
index2.3 <- which(d2>qchisq(0.5,2)&d2<=qchisq(0.75,2))
col.index2 <- rep("green",nrow(pendigits.pca3))
col.index2[index2] <- "red"
col.index2[index2.2] <- "orange"
col.index2[index2.3] <- "yellow"
p2 <- round(length(index2)/nrow(pendigits.pca3),3)
p2.2 <- round(length(index2.2)/nrow(pendigits.pca3),3)
p2.3 <- round(length(index2.3)/nrow(pendigits.pca3),3)

plot(pendigits.pca3[,1]~pendigits.pca3[,3],asp=1,col=col.index2,pch=16,xlab = "PC 3",ylab = "PC 1",main=
lines(ellipse(x=var(pendigits.pca3[,c(3,1)]),center=c(0,0),level=0.95),col="red")
lines(ellipse(x=var(pendigits.pca3[,c(3,1)]),center=c(0,0),level=0.75),col="orange")
lines(ellipse(x=var(pendigits.pca3[,c(3,1)]),center=c(0,0),level=0.5),col="yellow")
text(0,-0,paste((1-p2-p2.2-p2.3)*100,"%"))#should be 50%
text(0,-2.5,paste((1-p2.2-p2)*100,"%"))#should be 75%
text(0,-3.7,paste((1-p2)*100,"%"))#should be 95%

d3=mahalanobis(pendigits.pca3[,c(3,2)],center = c(0,0),cov=var(pendigits.pca3[,c(3,2)]))
index3 <- which(d3>qchisq(0.95,2))
index3.2 <- which(d3>qchisq(0.75,2)&d3<=qchisq(0.95,2))
index3.3 <- which(d3>qchisq(0.5,2)&d3<=qchisq(0.75,2))
col.index3 <- rep("green",nrow(pendigits.pca3))
```

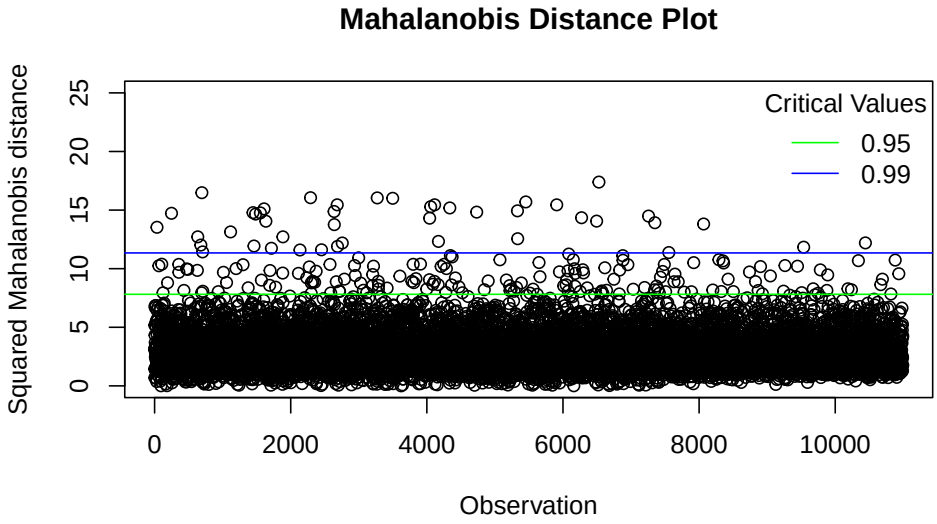
```r
col.index3[index3] <- "red"
col.index3[index3.2] <- "orange"
col.index3[index3.3] <- "yellow"
p3 <- round(length(index3)/nrow(pendigits.pca3),3)
p3.2 <- round(length(index3.2)/nrow(pendigits.pca3),3)
p3.3 <- round(length(index3.3)/nrow(pendigits.pca3),3)

plot(pendigits.pca3[,2]~pendigits.pca3[,3],asp=1,col=col.index3,pch=16,xlab = "PC 2",ylab = "PC 3",main=
lines(ellipse(x=var(pendigits.pca3[,c(3,2)]),center=c(0,0),level=0.95),col="red")
lines(ellipse(x=var(pendigits.pca3[,c(3,2)]),center=c(0,0),level=0.75),col="orange")
lines(ellipse(x=var(pendigits.pca3[,c(3,2)]),center=c(0,0),level=0.5),col="yellow")
text(0,0,paste((1-p3-p3.2-p3.3)*100,"%"))#should be 50%
text(0,-2,paste((1-p3.2-p3)*100,"%"))#should be 75%
text(0,-3.5,paste((1-p3)*100,"%"))#should be 95%
```




If the original data were normally distributed, also the principal components should be since they are linear combinations of the original variables; in particular each couple of variables form a multinormal vector should be again normally distributed, however it is difficult to be convinced of the normality of the PCs from these scatter plots. Indeed, we notice that in each case we have less probability mass than expected in the center of the distribution (50% is never reached) and that we have lighter tails than expected; Moreover, apart from the first plot, we observe a peculiar pattern in the tail observations, with some clusters of points that could be considered outliers under the assumption of normality indeed it seems that there is a mass point far from the distirbution.

To investigate further, we consider the Mahalanobis distance and examine whether observations fall outside the critical values of $\chi^2$ at the 0.95 and 0.99 confidence levels. In this case, we observe a significant portion of the data exceeding each of these two thresholds. Additionally, the Q-Q plot indicates that the Mahalanobis distance does not follow a $\chi^2$ distribution with 3 degrees of freedom since it deviates too much from the straight line. In the Mahalanobis distance Q-Q plot, the sample quantiles are lower than the $\chi^2_3$ values for quantiles, suggesting that the Mahalanobis distances in your data are generally more concentrated than what would be expected under the $\chi^2_3$ distribution.



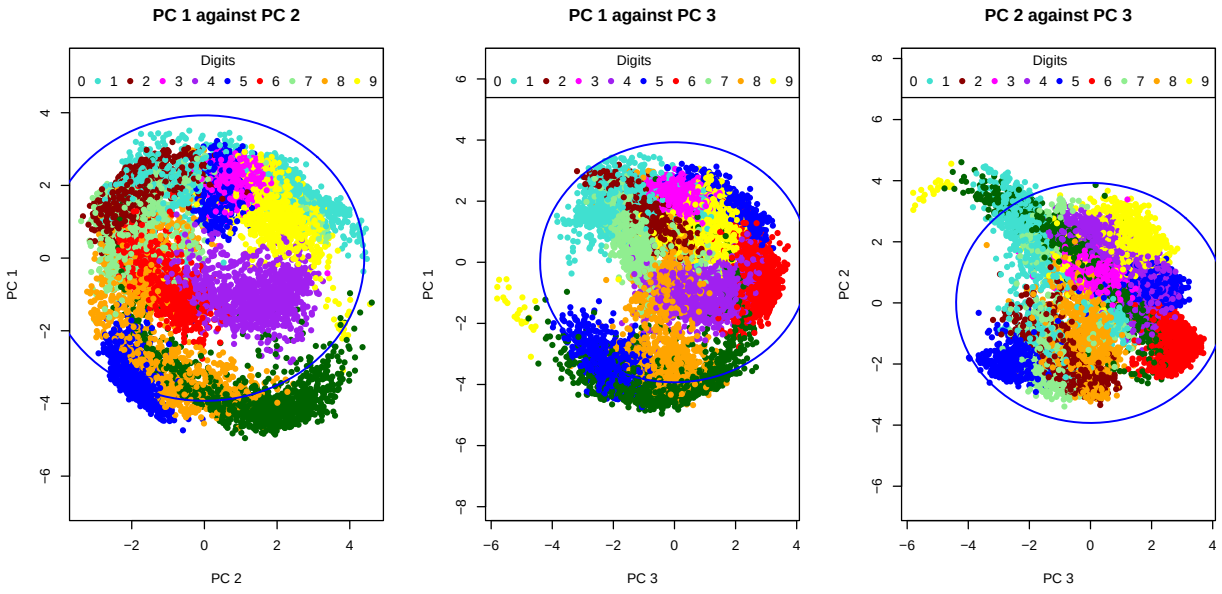

## 3.3

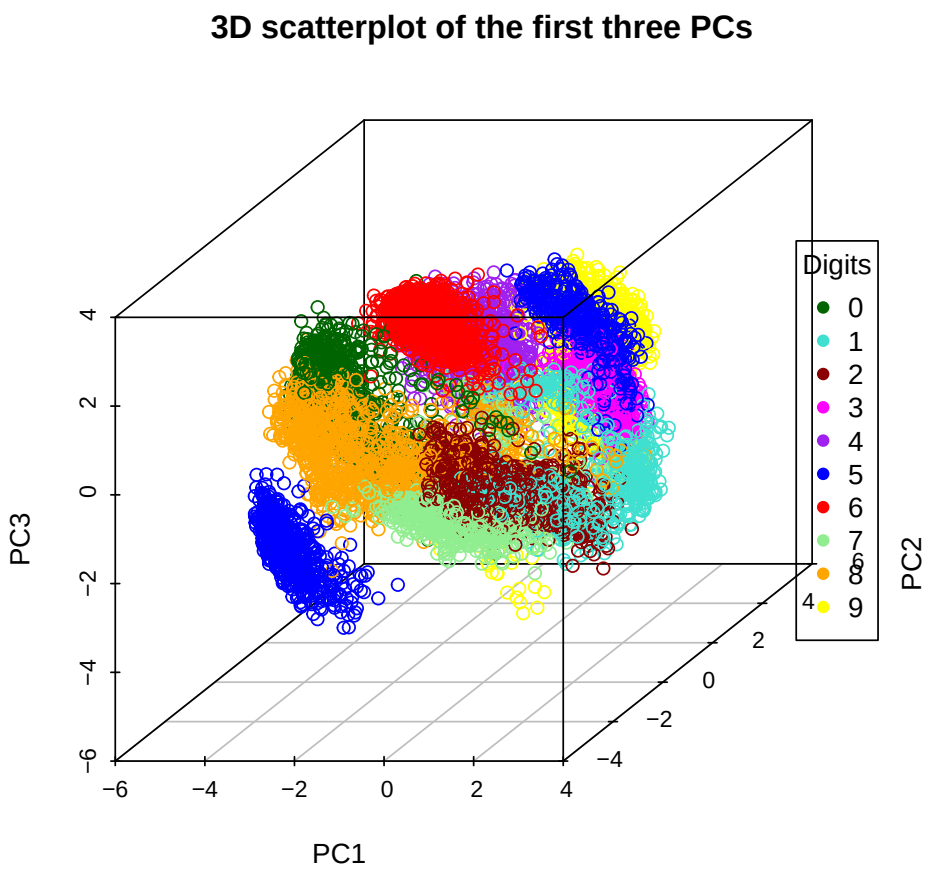
We can take a look to the scatter plots of the PCs, color-coded by which digit each observation represents:

```
lookup=c("darkgreen", "turquoise", "darkred", "magenta", "purple",
         "blue", "red", "lightgreen","orange","yellow") #vector of colors
col.index=pendigits$digit
for (i in 0:9){
  col.index[col.index==i]=lookup[i+1]
} #assigning colors to digits
```

```
par(mfrow=c(1,3))
plot(pendigits.pca3[,1]~pendigits.pca3[,2],asp=1,col=col.index,pch=16,xlab = "PC 2",ylab = "PC 1",main=
lines(ellipse(x=diag(pendigits.pca$sdev[2:3]^2),
              centre=c(0,0),level=0.95), col="blue",lwd=1.5)
```

```
legend(x = "topright",
       legend = c(0,1, 2,3,4,5,6,7,8,9),
       col = lookup,
       horiz = T,
       pch = 16,
       title = "Digits",
       cex = 1)
plot(pendigits.pca3[,1]~pendigits.pca3[,3],asp=1,col=col.index,pch=16,xlab = "PC 3",ylab = "PC 1",main=
lines(ellipse(x=diag(pendigits.pca$sdev[2:3]^2),
              centre=c(0,0),level=0.95), col="blue",lwd=1.5)
legend(x = "topright",
       legend = c(0,1, 2,3,4,5,6,7,8,9),
       col = lookup,
       horiz = T,
       pch = 16,
       title = "Digits",
       cex = 1)
plot(pendigits.pca3[,2]~pendigits.pca3[,3],asp=1,col=col.index,pch=16,xlab = "PC 3",ylab = "PC 2",main=
lines(ellipse(x=diag(pendigits.pca$sdev[2:3]^2),
              centre=c(0,0),level=0.95), col="blue",lwd=1.5)
legend(x = "toprigh",
       legend = c(0,1, 2,3,4,5,6,7,8,9),
       col = lookup,
       horiz = T,
       pch = 16,
       title = "Digits",
       cex = 1)
```

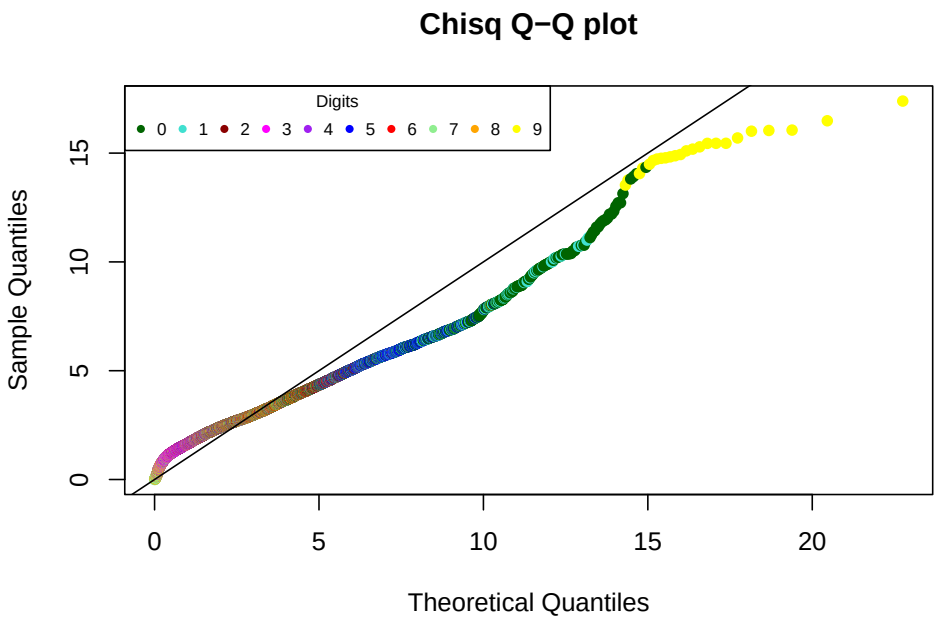





In the 2D and 3D scatterplots, after applying color coding, we observe that digits tend to cluster together. Interestingly, we notice that the previously mentioned mass points all correspond to the same digit. Moreover, different digits tend to show different behaviours. For example:

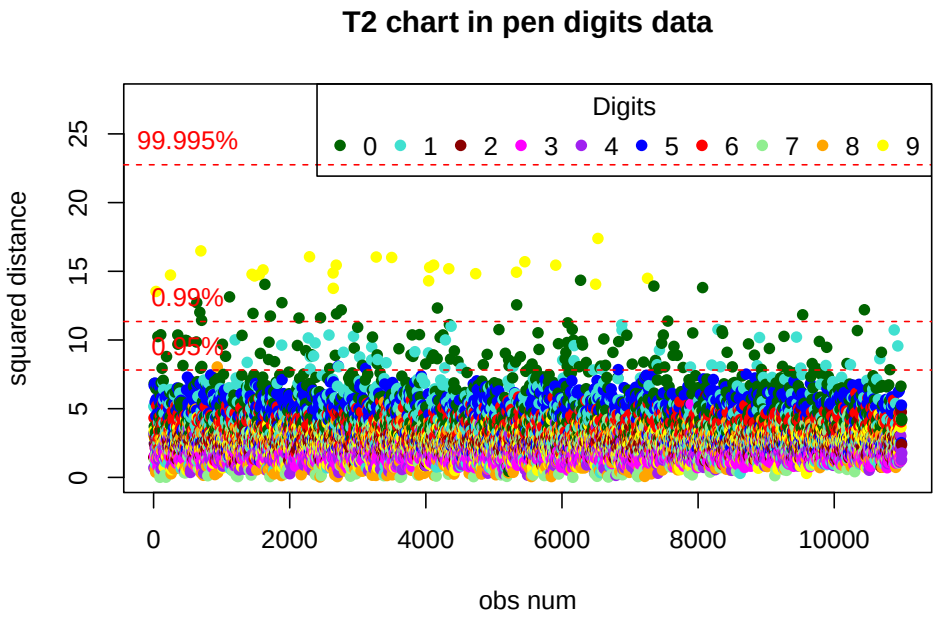
- from the 2d plots we can see how the 0 digit is clustered in a group that largely falls outside the ellipsis, signaling what could be a particularly non-normal or outlier-ish behaviour;
- from the 2d and 3d plots we can see how the 5 digit tends to cluster in two different zones away from each other;
- there are digits that looks more normally distributed on the scatterplot: for example, the 4 digit seems to show less dispersion.

We can further investigate this aspect with the chi square $Q - Q$ plot of the multivariate Mahalanobis distance.




Here the color coding reveals an interesting fact about the distribution: there are digits that are especially off the normal expectation, for example the 9 digit which is much more concentrated at the "tail" of the distribution than it should be under normality assumption. More information about this can be desumed from the color-coded Mahalanobis distance plot:

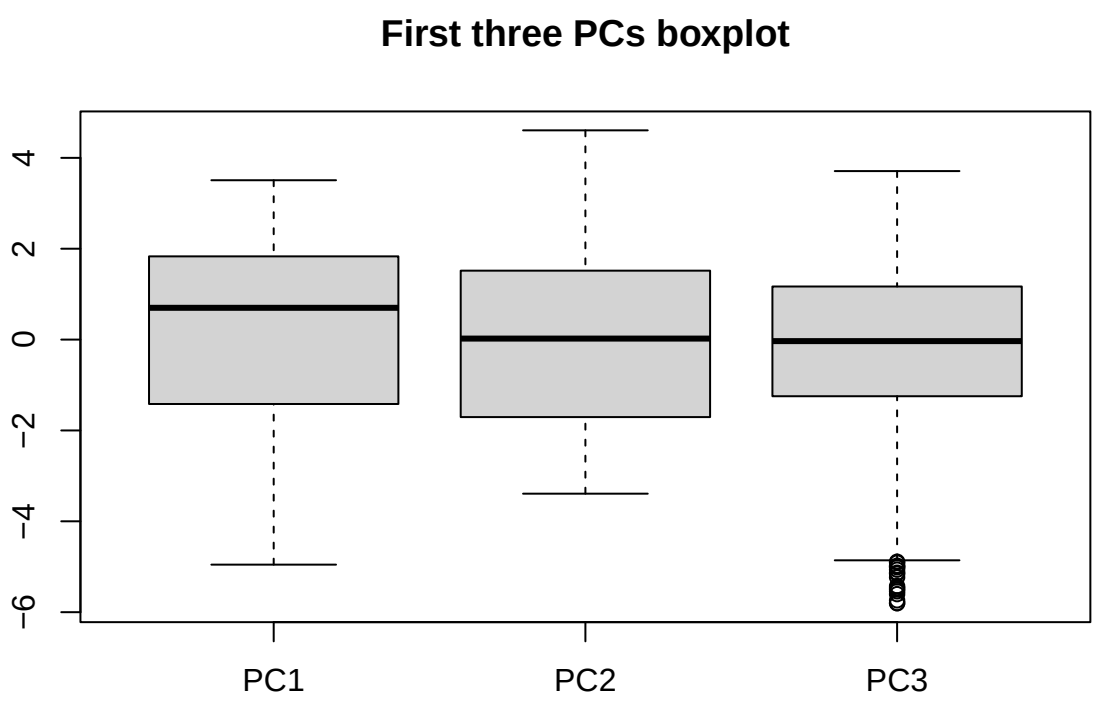
```
d<-mahalanobis(pendigits.pca$x[,1:3],center=c(0,0,0),cov=cov(pendigits.pca$x[,1:3]))
plot(d,ylim=c(0,qchisq(1-0.05/n,df=3)),
     pch=16,
     col=col.index,
     main="T2 chart in pen digits data", xlab="obs num",
     ylab="squared distance")
abline(h=qchisq(0.95,df=3), lty=2, col="red")
abline(h=qchisq(0.99,df=3),lty=2, col="red")
abline(h = qchisq(1 - 0.5/n, df = 3), lty = 2, col="red")

text(x = 500, y = qchisq(0.95, df = 3), labels = paste0(0.95, "%"), pos = 3, col="red")
text(x = 500, y = qchisq(0.99, df = 3), labels = paste0(0.99, "%"), pos = 3, col="red")
text(x = 500, y = qchisq(1 - 0.5/n, df = 3), labels = paste0(round(1 - 0.5/n,5)*100, "%"), pos = 3, col=
#abline(h=qchisq(1-1/3,df=3),lty=2)
legend(x = "topright",
       legend = c(0,1, 2,3,4,5,6,7,8,9),
       col = lookup,
       horiz = T,
       pch = 16,
       title = "Digits",
       cex = 1)
```




Here, we can observe more clearly that the digits deviating most from normality are 9, 0, and 1. Particularly, some points could be considered outliers at the 99% confidence level. However, the number of outliers is relatively low compared to the sample size. By employing continuity correction, we find that no observations could be deemed outliers (since the threshold is approximately 1) due to the large sample size. This means that we are able to approximate the "real" distribution accurately, and there are no outliers present.

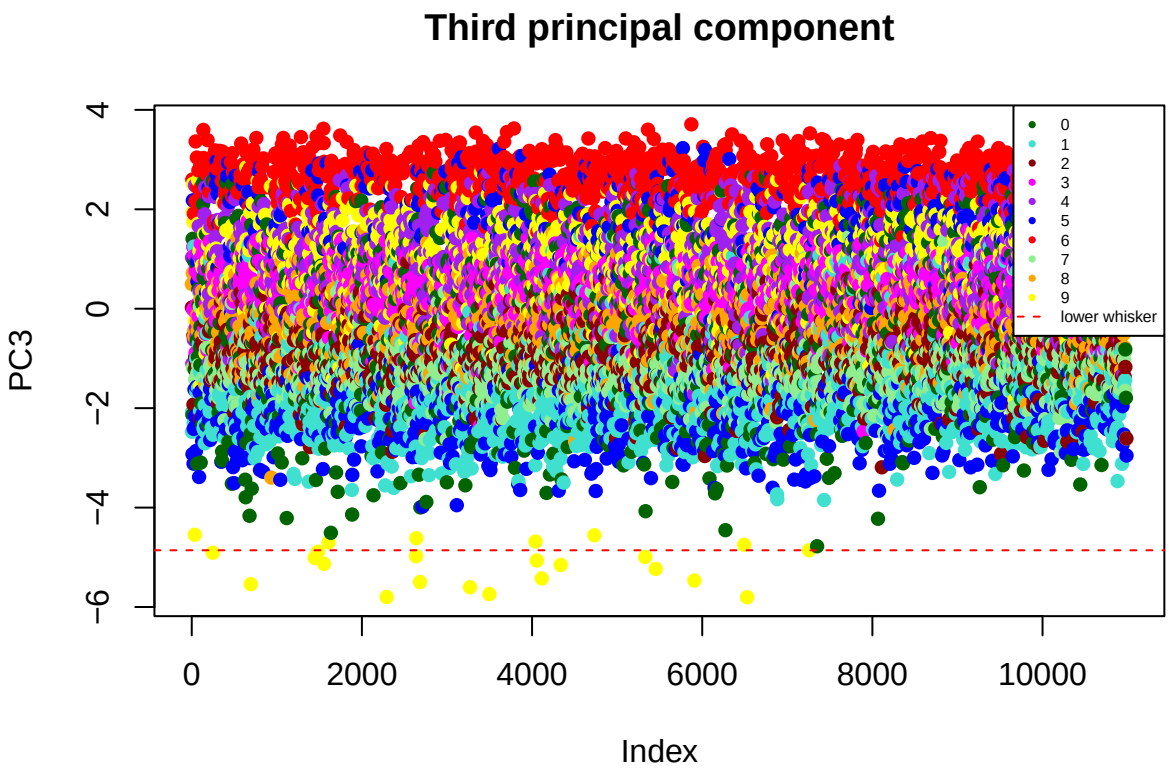
## 3.4

We start by analyzing a boxplot of the first three principal components.



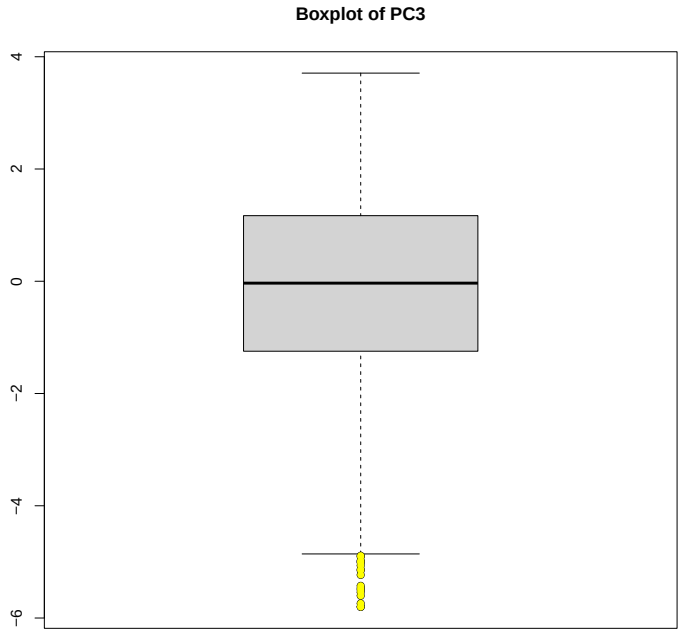

The boxplot function reports as outlier any data point which is further than 1.5 × the interquartile range from the box. From the boxplot we can see that on this criterion's basis only the third principal component has outliers in the low-end of the data. To further investigate this matter, we can plot the third principal component alongside the line that represents the limit of the lower whisker of the boxplot,

```
plot(pendigits.pca$x[,3],col=col.index,main = "Third principal component",ylab="PC3",pch = 16)
legend(x = "topright",
       legend = c(0,1, 2,3,4,5,6,7,8,9,"lower whisker"),
       col = append(lookup,"red"),
       bg="white",
       horiz = F,
       pch = c(16,16,16,16,16,16,16,16,16,16,NA),
       lty = c(NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,2),
       cex = 0.5)
abline(h=boxplot.stats(pendigits.pca$x[,3])$stats[1],lty=2,col="red")
```




From this plot we can already see how all outliers are 9 digits. To confirm that:

```
boxplot(pendigits.pca$x[,3], main="Boxplot of PC3")
index=which(pendigits.pca$x[,3] %in% boxplot.stats(pendigits.pca$x[,3])$out)
points(rep(1,length(index)),
pendigits.pca$x[index,3],pch=16,col=col.index[index])
```

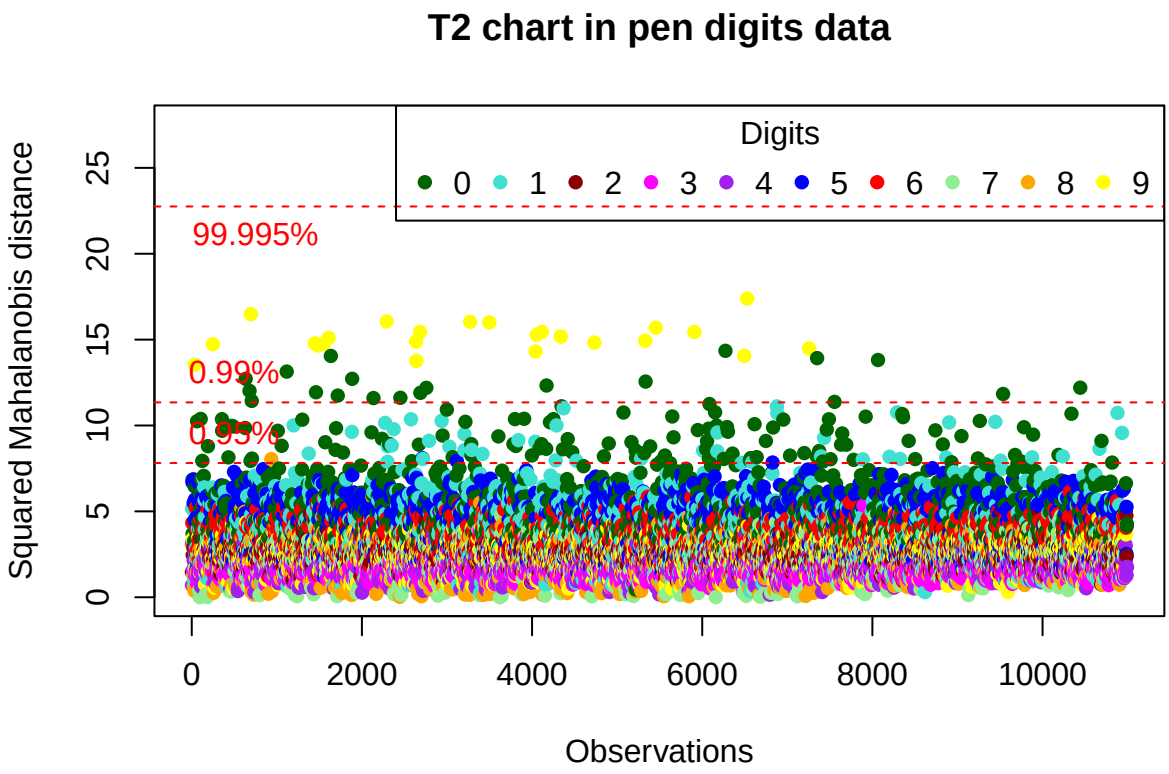



```
pendigits$digit[index]
```

```
##  [1] 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9
```

Analyzing the Mahalanobis distance from the theoretical quantiles, we can once more see the difference between the different digits.

```
d <- mahalanobis(pendigits.pca$x[,1:3], center = c(0,0,0), cov = cov(pendigits.pca$x[,1:3]))
plot(d, ylim = c(0, qchisq(1-0.05/n, df = 3)),
     pch = 16,
     col = col.index,
     main = "T2 chart in pen digits data", xlab = "Observations",
     ylab = "Squared Mahalanobis distance")
abline(h = qchisq(0.95, df = 3), lty = 2, col="red")
abline(h = qchisq(0.99, df = 3), lty = 2, col="red")
abline(h = qchisq(1 - 0.5/n, df = 3), lty = 2, col="red")
text(x = 750, y = qchisq(1 - 0.5/n, df = 3), labels = paste0(round(1 - 0.5/n,5)*100, "%"), pos = 1, col=
text(x = 500, y = qchisq(0.95, df = 3), labels = paste0(0.95, "%"), pos = 3, col="red")
text(x = 500, y = qchisq(0.99, df = 3), labels = paste0(0.99, "%"), pos = 3, col="red")

legend(x = "topright",
       legend = c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9),
       col = lookup,
       horiz = TRUE,
       pch = 16,
       title = "Digits",
       cex = 1)
```




Some digit 9 emerges as some of the most problematic points, potentially being a multivariate outlier across all levels of $\chi^2_3$ critical values. Additionally, numerous instances of digit 9 appear as univariate outliers with respect to the third component. Interestingly, in the multivariate context, many of these points could be considered outliers at the 99% confidence level, but not when applying continuity correction, likely due to the large sample size.