Problem Set 1

Luca Frattegiani (1013326)

06/04/2022

Exercise 1: Air Pollution Data

Data Description

The Air Pollution Data consist of 7 measurements (variables) recorded in 41 cities (observations) of the United States. The first variable "*S02*" (which rapresents the sulphur dioxide content in micrograms per cubic meter) has been deleted, leading to a total of 6 Variables we can group in two main topics:

Human Ecology:

Manuf: Number of manufacturing enterprises employing 20 or more workers.

Pop: Population size (1970 census) in thousands.

Climate:

SO2: Sulphur dioxide content in micrograms per cubic meter.

Neg.Temp: Average annual temperature in Fo (negative values).

Wind: Average annual wind speed in miles per hour.

Precip: Average annual precipitation in inches.

Days: Average number of days with precipitation per year.

Data importation

We display the first rows of the dataset:

```
##               SO2 Neg.Temp Manuf Pop Wind Precip Days
## Phoenix        10    -70.3   213 582  6.0   7.05   36
## Little Rock    13    -61.0    91 132  8.2  48.52  100
## San Francisco  12    -56.7   453 716  8.7  20.66   67
## Denver         17    -51.9   454 515  9.0  12.95   86
## Hartford       56    -49.1   412 158  9.0  43.37  127
## Wilmington     36    -54.0    80  80  9.0  40.25  114
```

1.1) Compute the Sample Mean Vector, Correlation Matrix and comment on correlations:

We compute the Sample Mean Vector $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \quad j = 1, ..., p$ for all of the $p = 6$ variables considered:

```
##   Neg.Temp      Manuf        Pop       Wind     Precip       Days
## -55.763415 463.097561 608.609756   9.443902  36.769024 113.902439
```

Now we look for the Correlation Matrix $\mathbf{R}$ which entries are defined as $r_{jk} = \frac{s_{jk}}{s_j s_k}$. Looking at the measurements included in the dataset we expect to see an high level of positive correlation between the number of people registered (*"Pop"*) and the amount of manufacturing enterprises (*"Manuf"*) in the city and also between the variables *"Precip"* and *"Days"*:

```
##          Neg.Temp  Manuf    Pop   Wind Precip  Days
## Neg.Temp    1.000  0.190  0.063  0.350 -0.386 0.430
## Manuf       0.190  1.000  0.955  0.238 -0.032 0.132
## Pop         0.063  0.955  1.000  0.213 -0.026 0.042
## Wind        0.350  0.238  0.213  1.000 -0.013 0.164
## Precip     -0.386 -0.032 -0.026 -0.013  1.000 0.496
## Days        0.430  0.132  0.042  0.164  0.496 1.000
```

In order to have a clearer view of the correlations we can put them in a matrix ordering them from the highest to the lowest ones:

```
##    Pairs.of.variables Correlations
## 1         Pop : Manuf        0.955
## 2       Days : Precip        0.496
## 3     Days : Neg.Temp        0.430
## 4     Wind : Neg.Temp        0.350
## 5        Wind : Manuf        0.238
## 6          Wind : Pop        0.213
## 7    Manuf : Neg.Temp        0.190
## 8         Days : Wind        0.164
## 9        Days : Manuf        0.132
## 10     Pop : Neg.Temp        0.063
## 11         Days : Pop        0.042
## 12      Precip : Wind       -0.013
## 13       Precip : Pop       -0.026
## 14     Precip : Manuf       -0.032
## 15  Precip : Neg.Temp       -0.386
```

From the previous Matrix of ordered correlations we can conclude that:

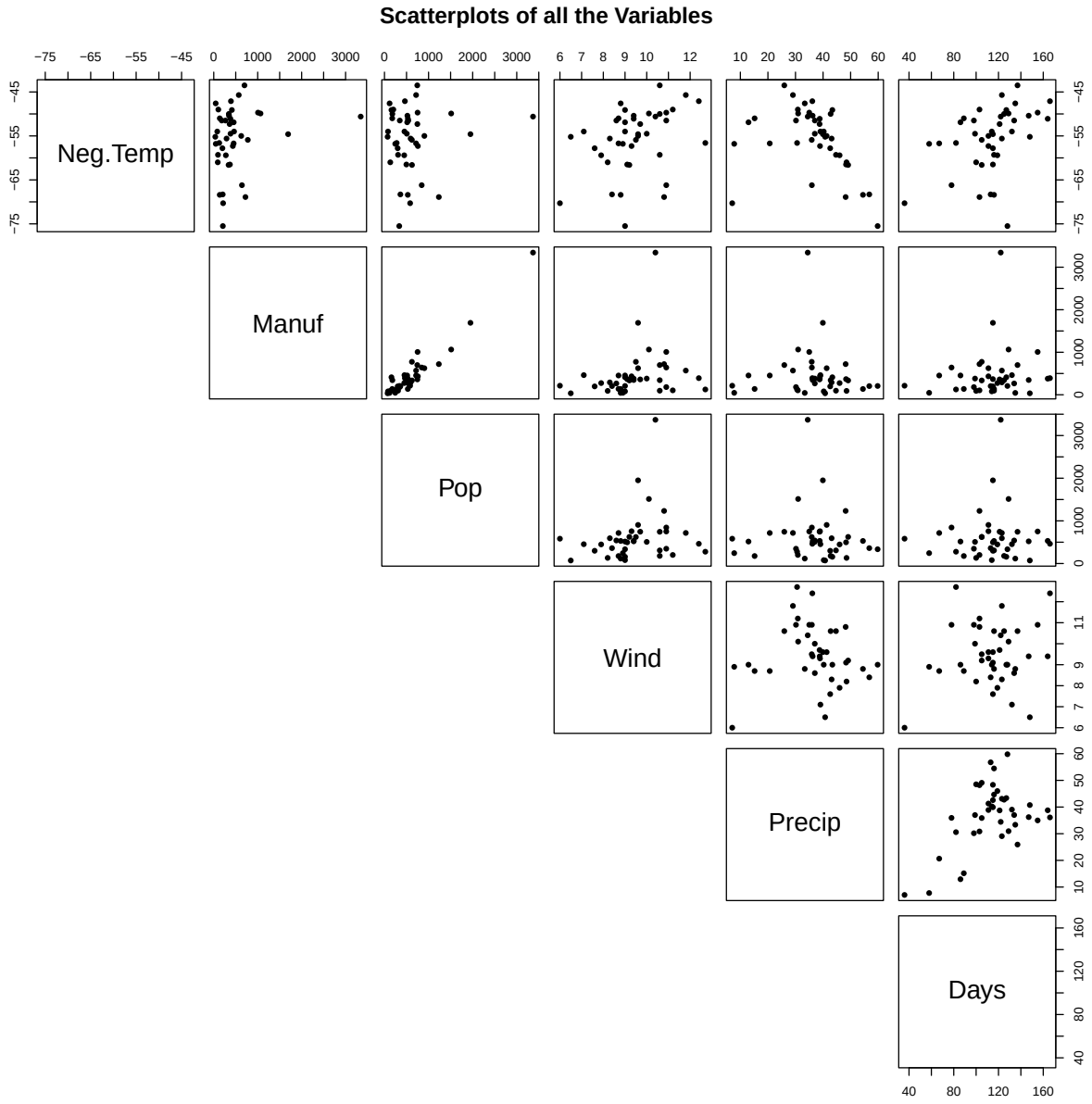
- There's a really high positive correlation between the social variables "Pop" and "$Manuf$" ($r = 0.955$) which suggest a possible linear dependence between them.
- Other relevant positive correlations are the ones between the climate measurements "$Days$" and "$Precip$" ($r = 0.496$), which is an obvious consequence since an higher number of days with

precipitations may lead to an higher value of the annual average precipitation level;

"$Days$" has another quite high correlation with the temperature ($r = 0.430$), which suggest that the more days with precipitations we have the less the recorded temperature is (since the "$Neg.Temp$" has negative values);

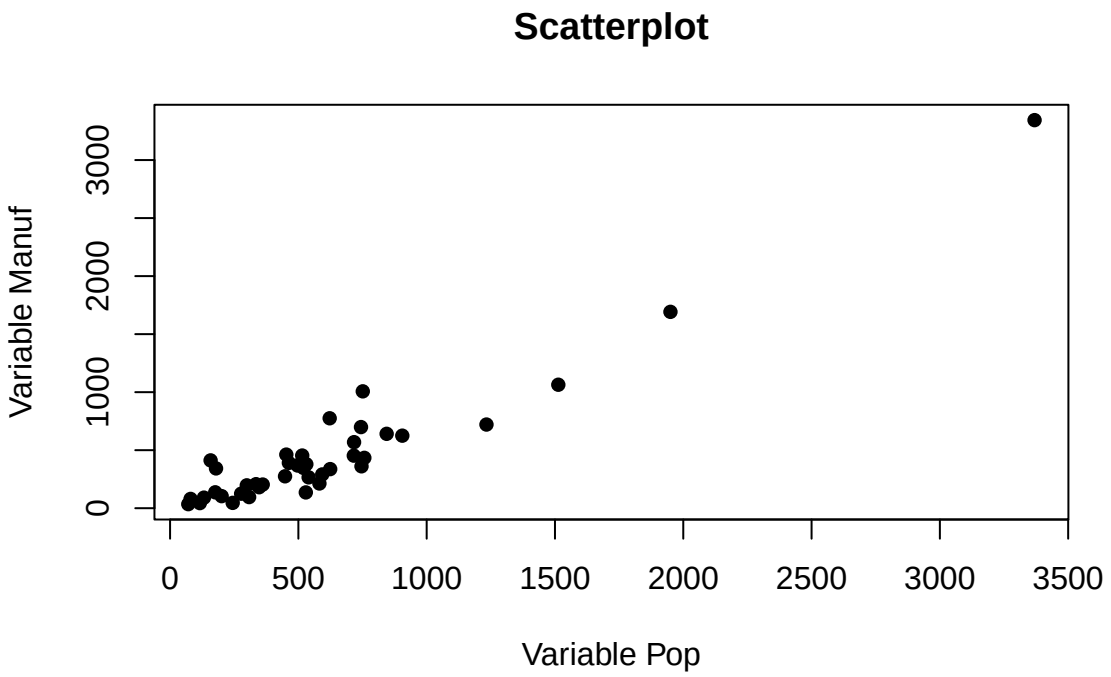
- We can also see that in general, each variable has at least one quite high correlation (around 0.5) with the others except for the variable "$Wind$".
- The other correlations are less significant.

We can see if the above considerations hold when we look at the scatterplots of the measurements:



So the plot confirms that the strong correlation between *"Manuf"* and *"Pop"* is maybe linked to a

linear relationship that we underline below through their scatterplot:



1.2) Make a boxplot of each variable and comment about the presence of outliers (no more than two per variable). Identify these observations:

We draw the boxplots for each variable in the Dataset and we comment about the presence of univariate outliers:

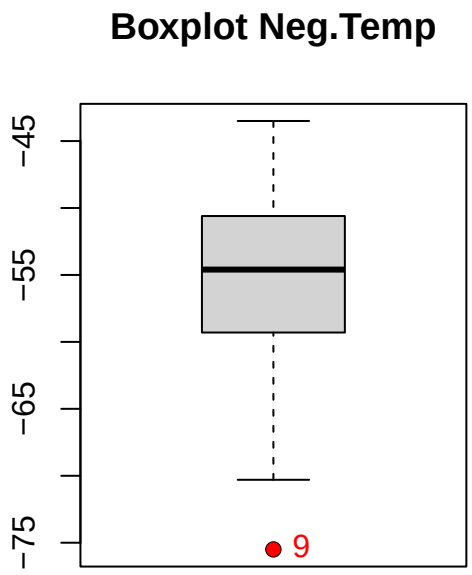
- **Neg.Temp**: The observation 9 is considered as an outlier for the first variable since it hires a very low level of temperature.
- **Manuf**: Looking at this measurement, the boxplot detects 4 outliers in total, but since two of them are really close to the upper whisker we can consider just observations 11 and 29 to be "*strong*" outliers.
- **Pop**: Here the boxplot confirms that observations 11 and 29 are outliers also for this variable (and this is coherent with the strong relationship we noticed before between these two measurements).
- **Wind**: No outlier has been detected for this variable.
- **Precip**: Observation 23 and 1 can be considered as outliers since they're under the lower whisker.
- **Days**: Also here the graphs underline that observations 23 and 1 are outlier cause they hire a low value (especially for 1). Also in that case, we noticed before that "*Precip*" and "*Days*" had a particular relationship.

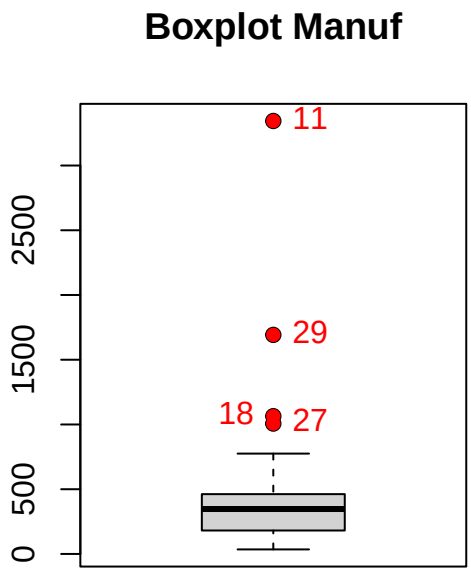
Here we summarize the detected observations:

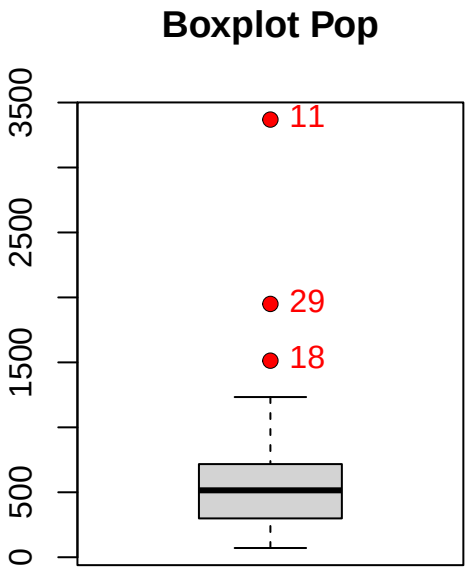
```
## $Neg.Temp
## [1] 9

## $Manuf
## [1] 11 18 27 29

## $Pop
## [1] 11 18 29

## $Wind
## integer(0)

## $Precip
## [1]  1 23

## $Days
## [1]  1 23 25
```

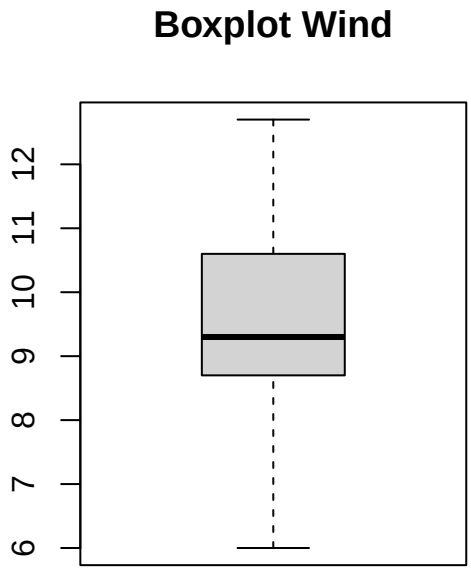
To conclude, we found some outliers for each of the variable (except from "$Wind$") and in particular there were some recurrent observations:

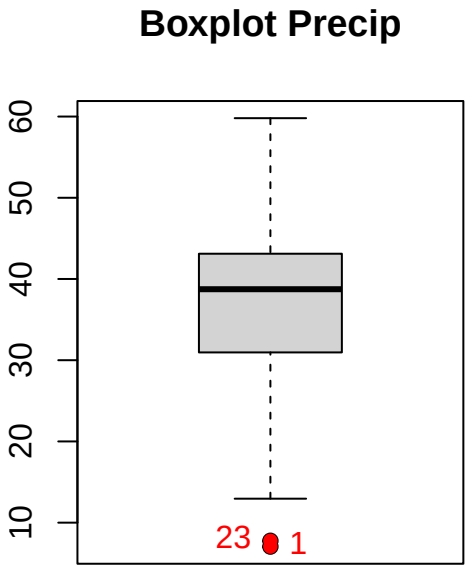
- "11"and"29" : The cities of Chicago and Philadelphia appears as outliers for the Human variables "Pop" and "$Manuf$".
- "1"and"23 : " Phoenix and Alburquerque appears as outliers for the Climate variables "$Precip$" and "$Days$".

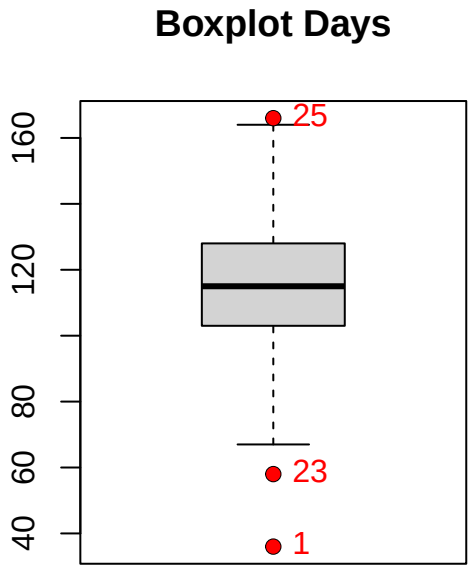





Boxplot Pop
3500
2500
1500
500
0
11
29
18


Boxplot Wind
12
11
10
9
8
7
6


Boxplot Precip
60
50
40
30
20
10
23
1


Boxplot Days
160
120
80
60
40
25
23
1

1.3) Construct a normal Q-Q plot for each variable and comment about normality:

We plot the two Q-Q plots for each variable, considering all the observation collected in the first case while in the second plot we'll remove the outliers detected at point *"1.2)"* in order to see how the normality changes.

The Q-Q plots will show the relationship between the *"Theoretical Quantiles"* (the quantiles of the $\mathcal{N}(0,1)$) and the *"Sample Quantiles"* observed for each measurement ($x_{(1)} \leq ... \leq x_{(i)} \leq x_{(n)} \quad i = 1, ..., n$ where $x_{(i)}$ is the $\frac{i-0.05}{n}$ quantile). The two lines plotted respectively show:

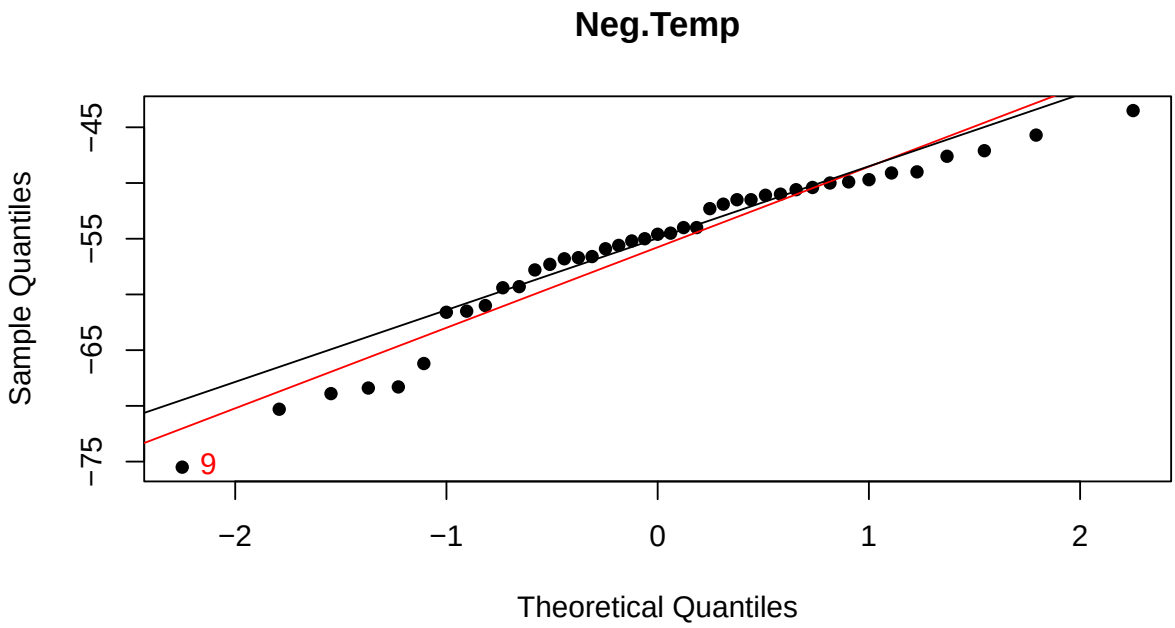
- **Red line:** It's made of the theoretical values of the observed quantiles we should obtain under the hypothesis of normality (defined as $\hat{x_{(i)}} = \sigma \cdot x_{(i)} + \mu$ where σ and μ are respectively the standard deviation and the mean of the Variable).
- **Black line:** It's the so called Q-Q line which passes through the 1^{th} and the 3^{rd} theoretical normal quantiles.

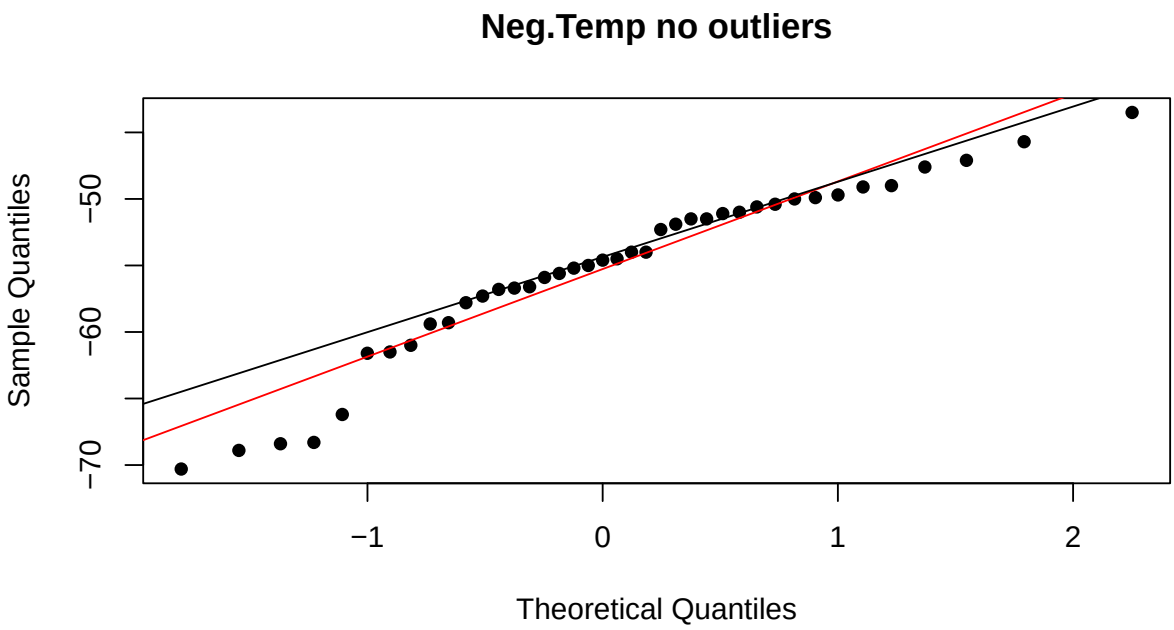
Then we'll plot the histogram with the estimated density function (the red curve) compared with the density curve it should have under assumption of normality (the black curve) to obtain another element for the evaluation, and finally we'll show the results of the Shapiro-Wilks normality test (which for low level of the *p-value* rejects the hypothesis of normality).

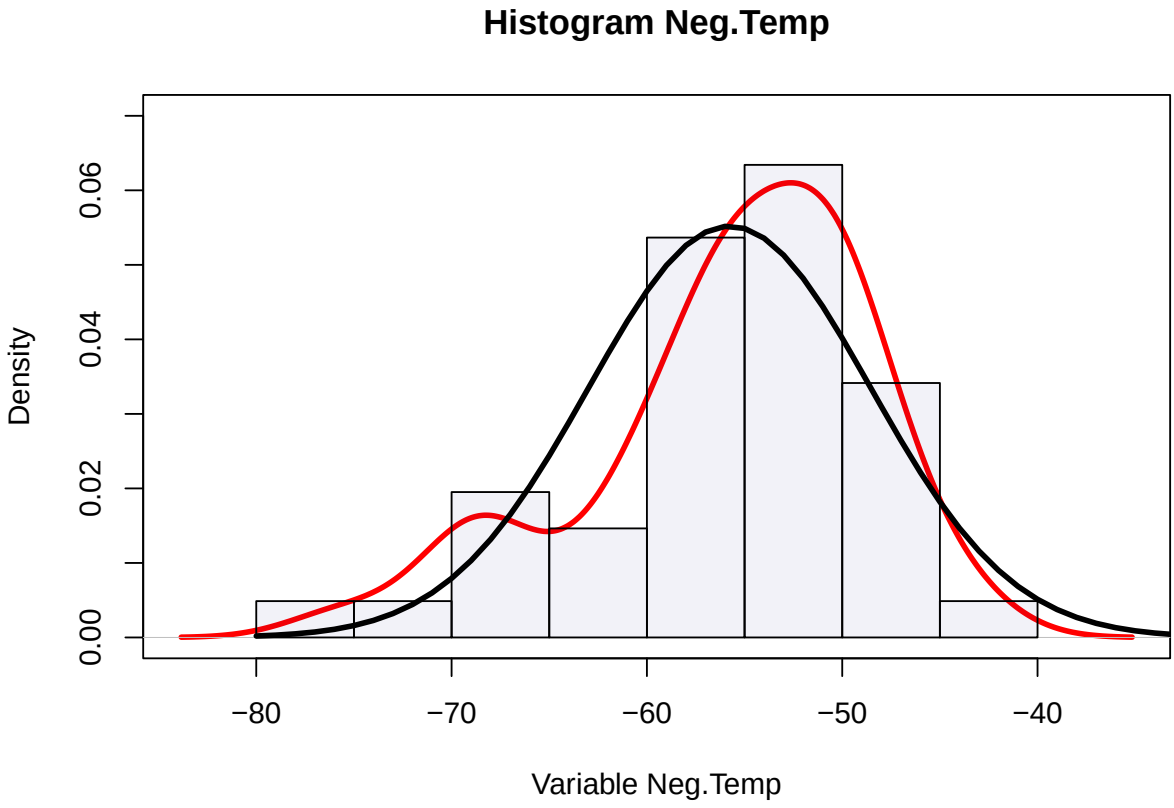
We start with the results of the Shapiro-Wilks tests:

```
##   Variable Statistic's value p-value observed
## 1 Neg.Temp         0.9355419     2.214972e-02
## 2    Manuf         0.6054834     2.781101e-09
## 3      Pop         0.6804922     3.622798e-08
## 4     Wind         0.9805736     6.972580e-01
## 5   Precip         0.9421444     3.725311e-02
## 6     Days         0.9653990     2.419457e-01
```

Then we show the plots and we comment:

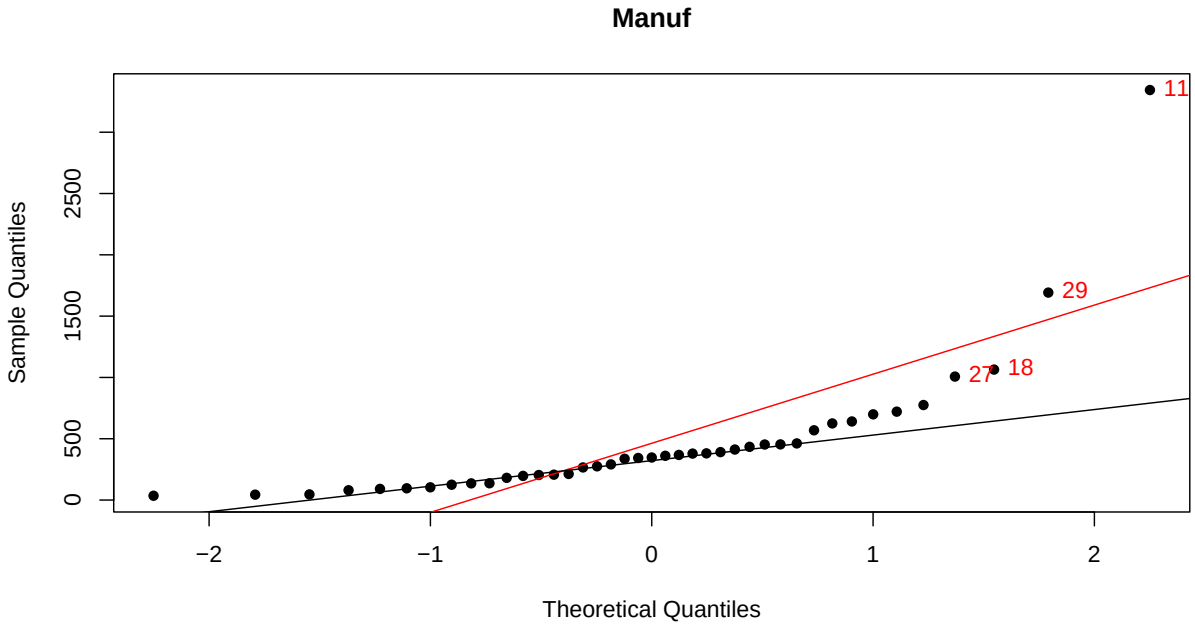


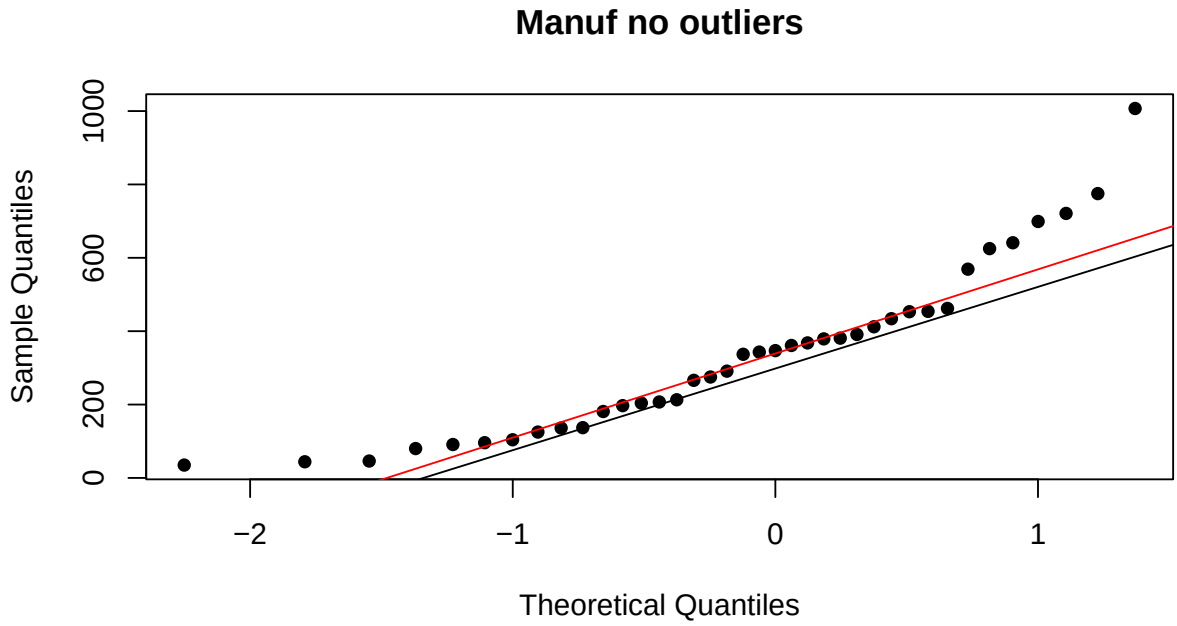


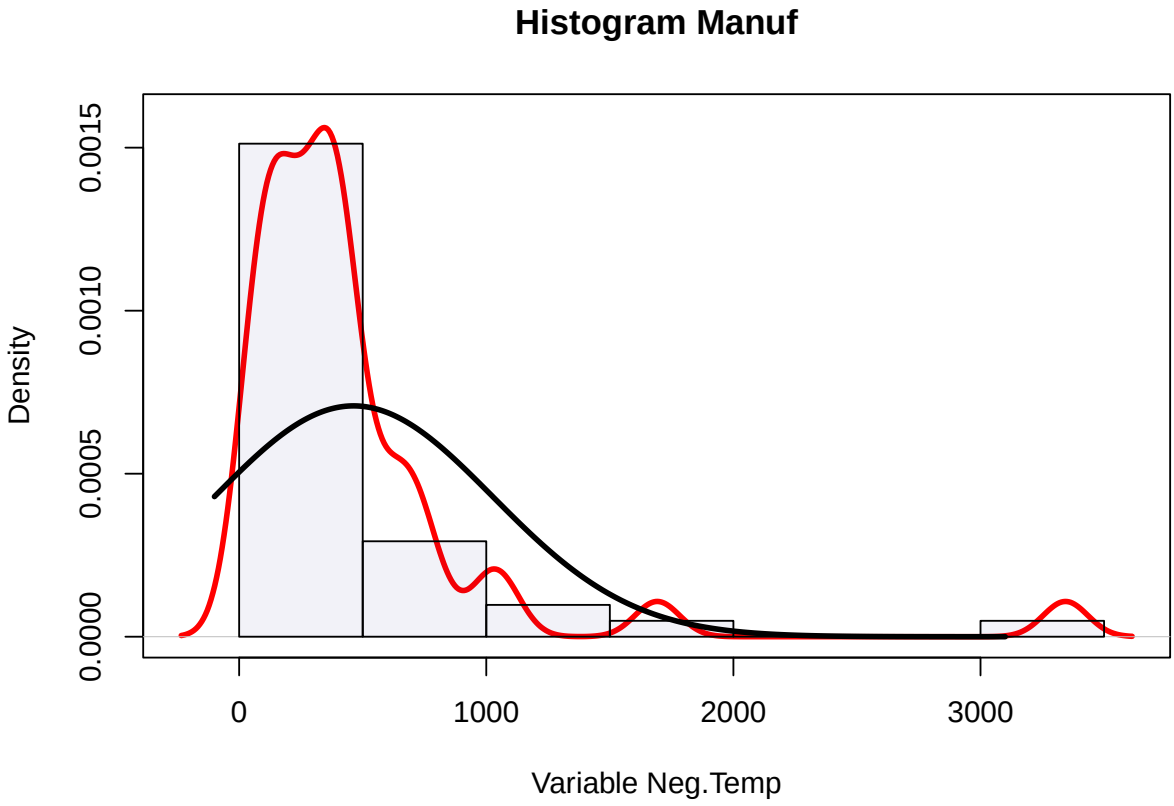


This variable seems to be quite normally distributed except for some low observations which deviate from the lines. If we remove the 9^{th} observation, the middle quantiles seems to be closer to the red line.

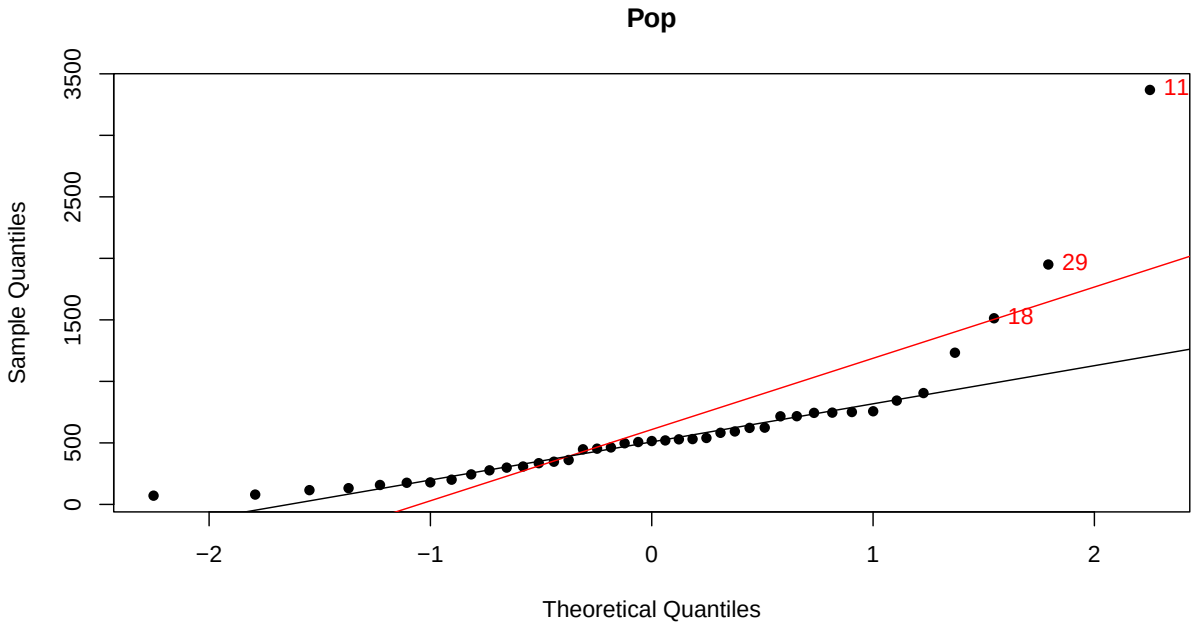
The histogram confirms that *"Neg.Temp"* has a non normal behavior in particular with respect to the first sample quantiles, but the result of the Shapiro-wilks test returns a value of the statistics which is really close to "1", confirming our doubts on the normality of this variable.

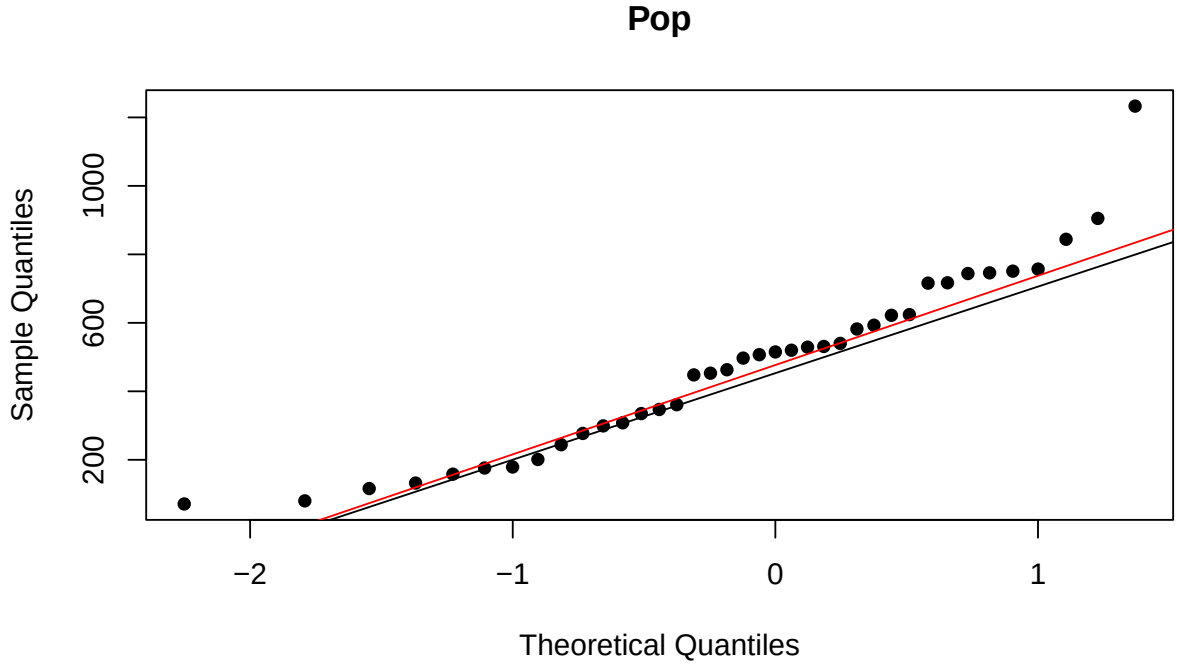


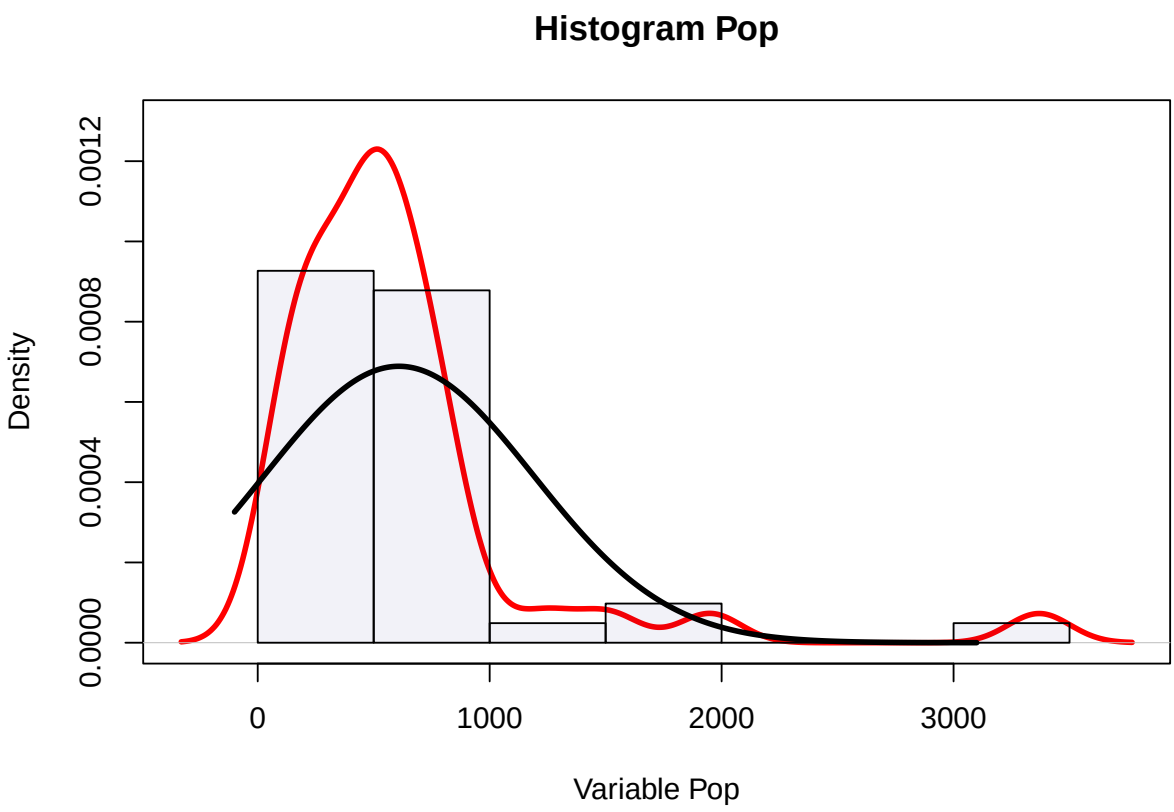




The 2^{nd} variable doesn't seem to be normally distributed especially because the highest observations which deviates significantly from the two lines. Without the 3 outliers considered before (observations *"11"*, *"18"* and *"29"*), the fit of the sample quantiles with the theoretical ones improve remarkably but we have excluded around the 7% of the observations. Finally, the histogram shows that *"Manuf"* is pretty much different from a Gaussian distribution, and this is confirmed also by the Shapiro-Wilks test.


Pop
Sample Quantiles
Theoretical Quantiles
0
500
1500
2500
3500
−2
−1
0
1
2
11
29
18

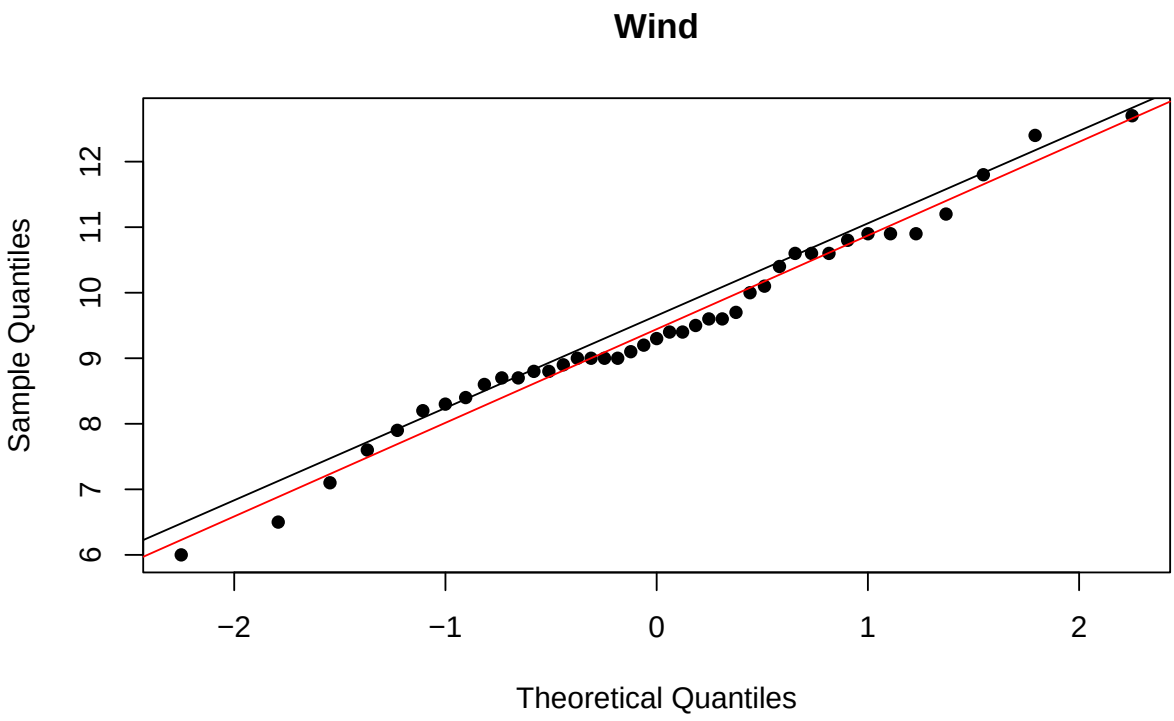

Pop
Sample Quantiles
Theoretical Quantiles
200
600
1000
−2
−1
0
1

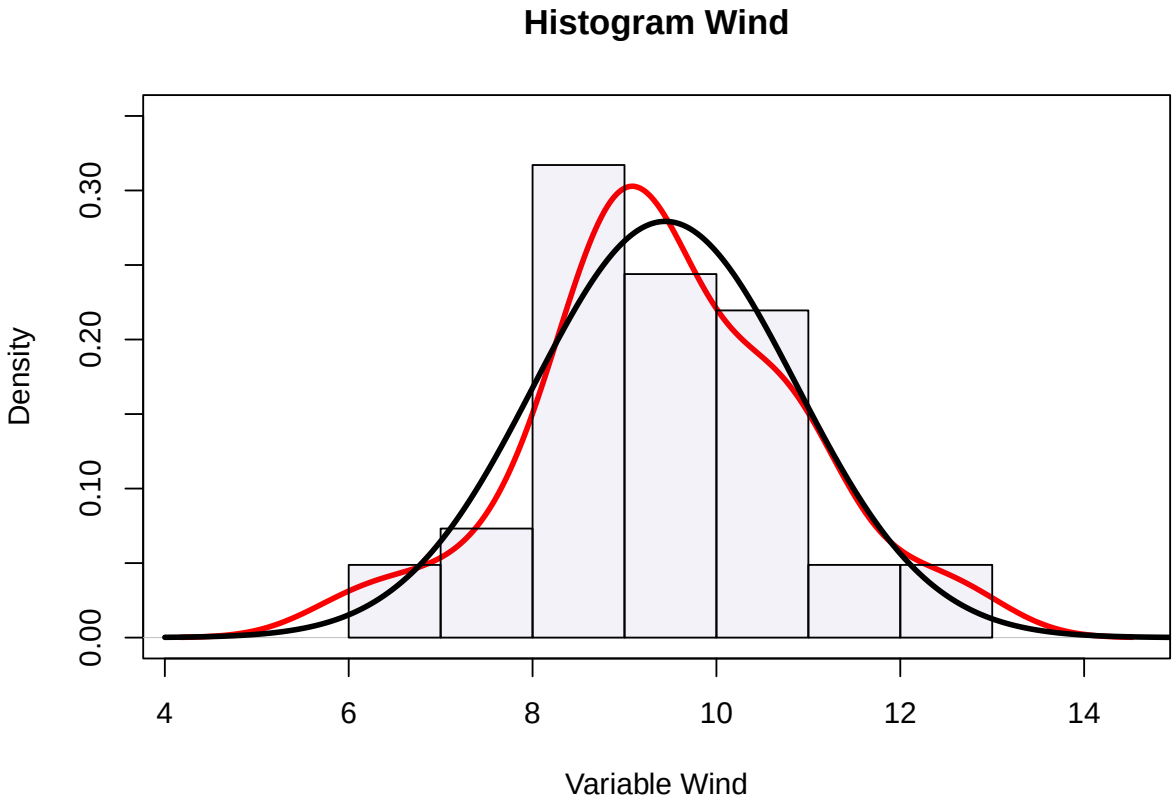


Also the variable *"Pop"* doesn't seem to be normally distributed as we can detect from the previous Q-Q plot, which shows some high and low observations distant from the red and the black line depicted and also the Shapiro-Wilks test rejects the hypothesis of normality.

Even by removing the outliers (again, observations *"11"*, *"18"* and *"29"*) there're some points which deviate considerably, though the fit with 2 lines improves.

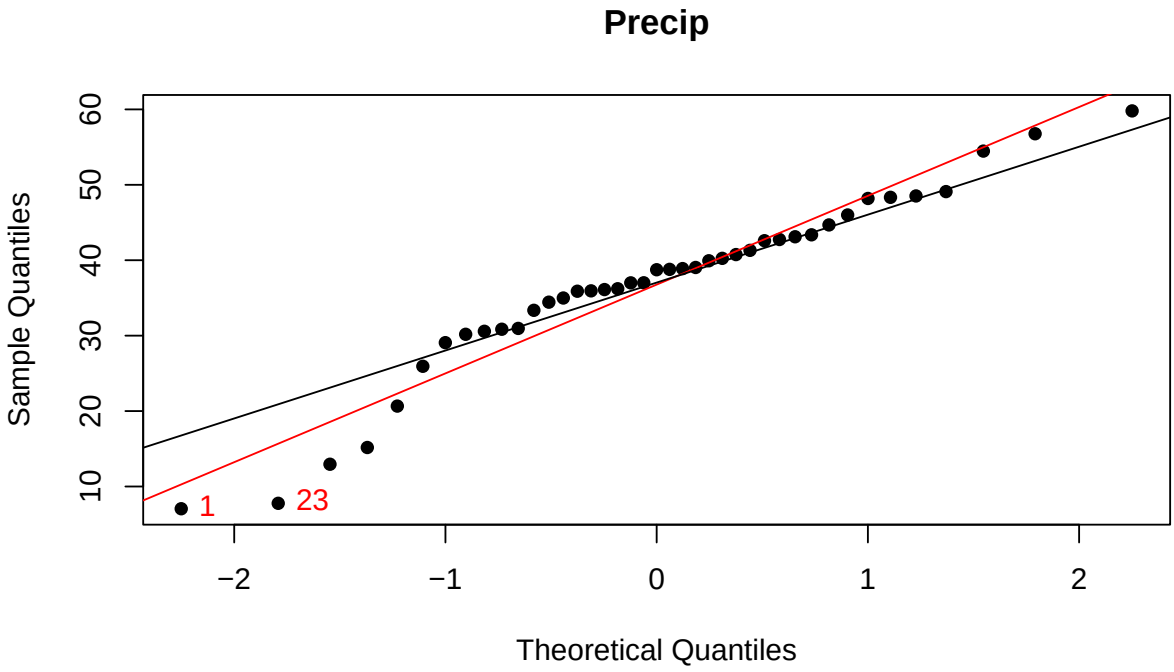
These results are confirmed by the histogram of the variable which shows an high frequency for the first values hired by the variable and a low frequency for the others. The considerations made for the normality of the variables *"Manuf"* and *"Pop"* are coherent with the high correlation we detected before for these two measurements, indeed, they seems to have a similar distribution which is non Gaussian.

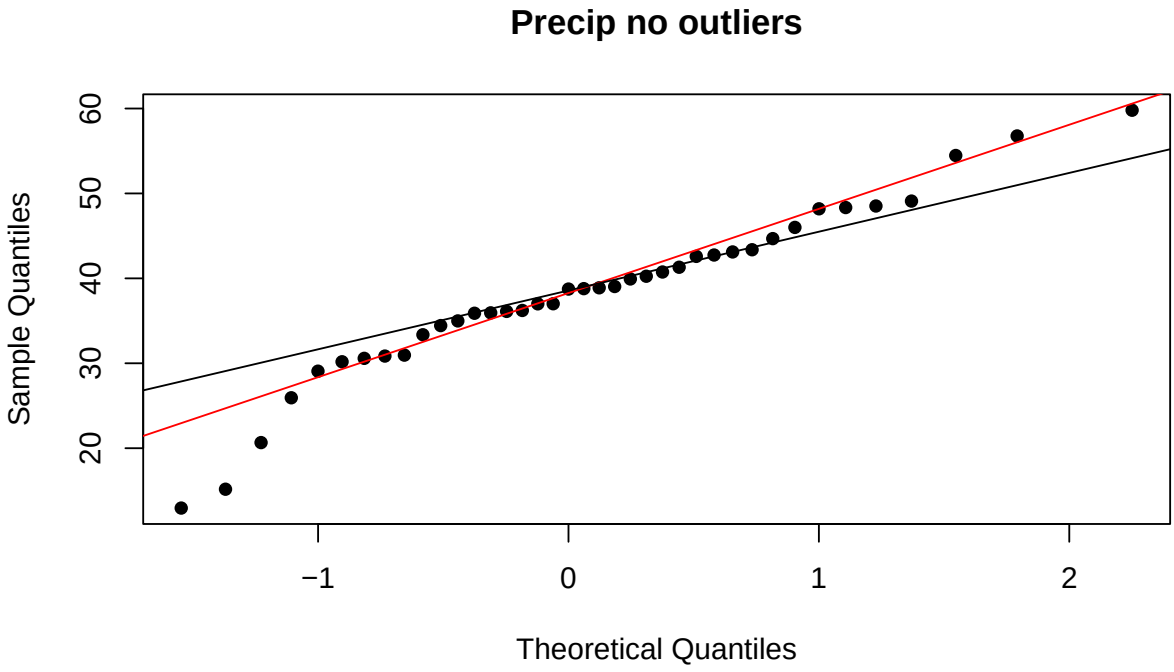

Wind
Sample Quantiles
Theoretical Quantiles
6
7
8
9
10
11
12
−2
−1
0
1
2

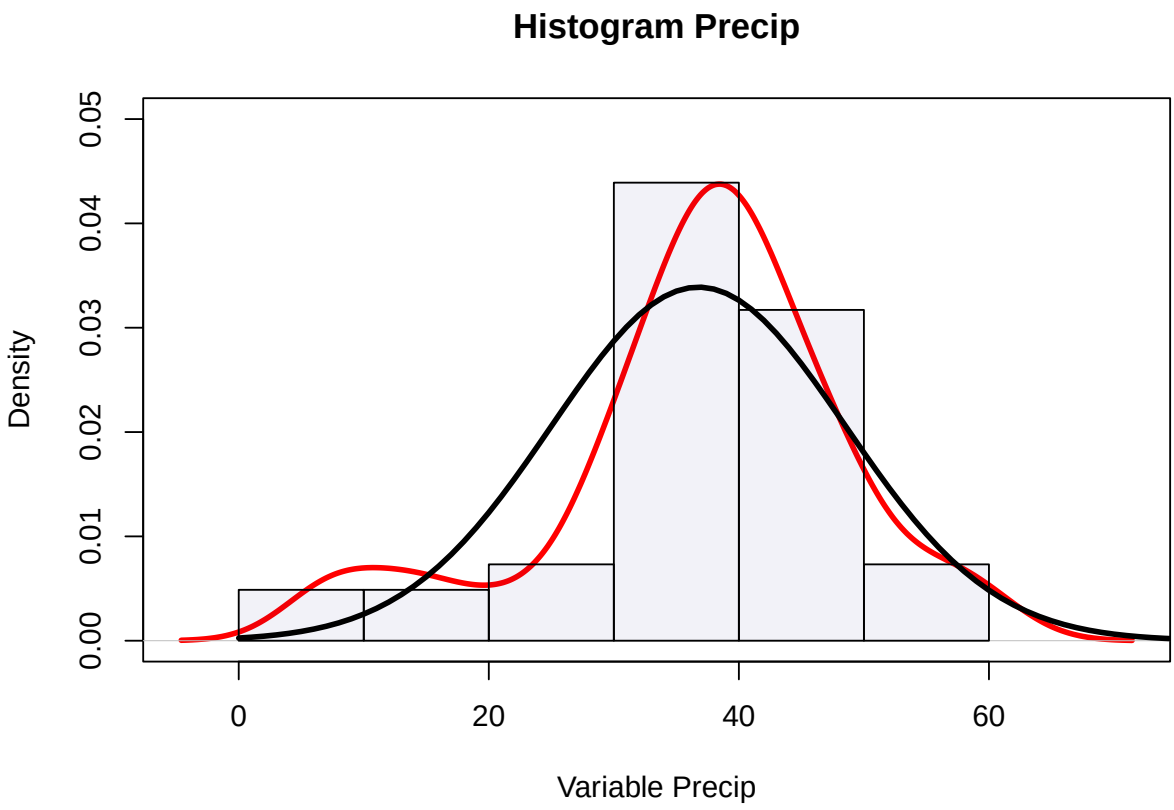

Histogram Wind
Density
Variable Wind
0.00
0.10
0.20
0.30
4
6
8
10
12
14

The data collected variable *"Wind"* seems to show Normal distribution which is evident both from the good approximation they give of the two lines of the Q-Q plot and from the shape of the histogram. Also the Shapiro-Wilks test confirms this conclusion.

We didn't detected any outlier for this variable so there's no need to make a second Q-Q plot.



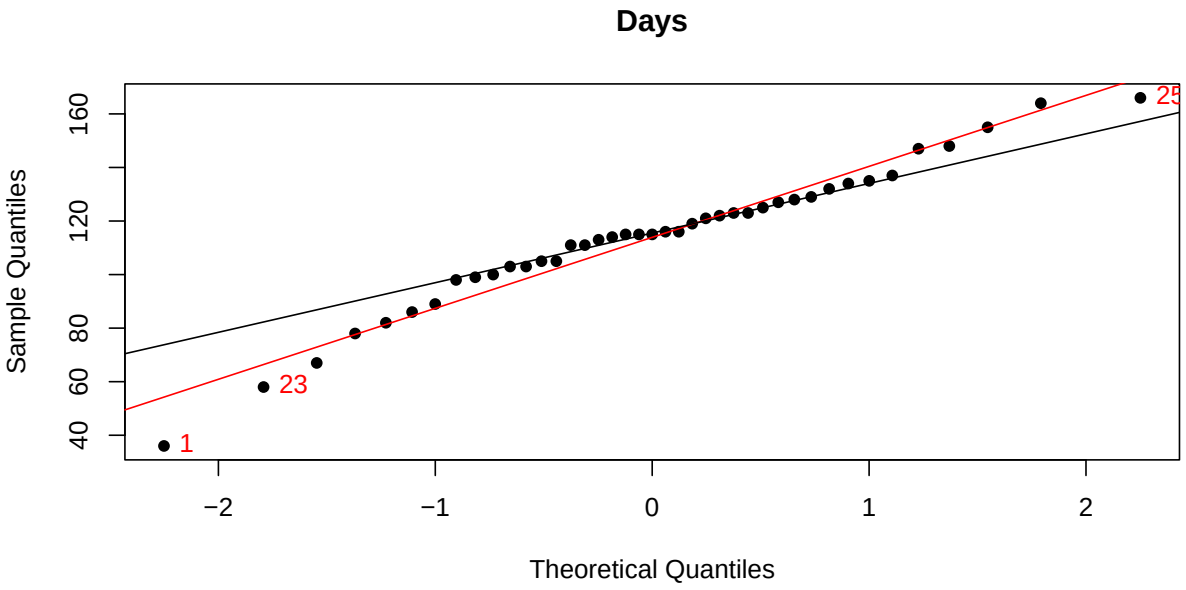


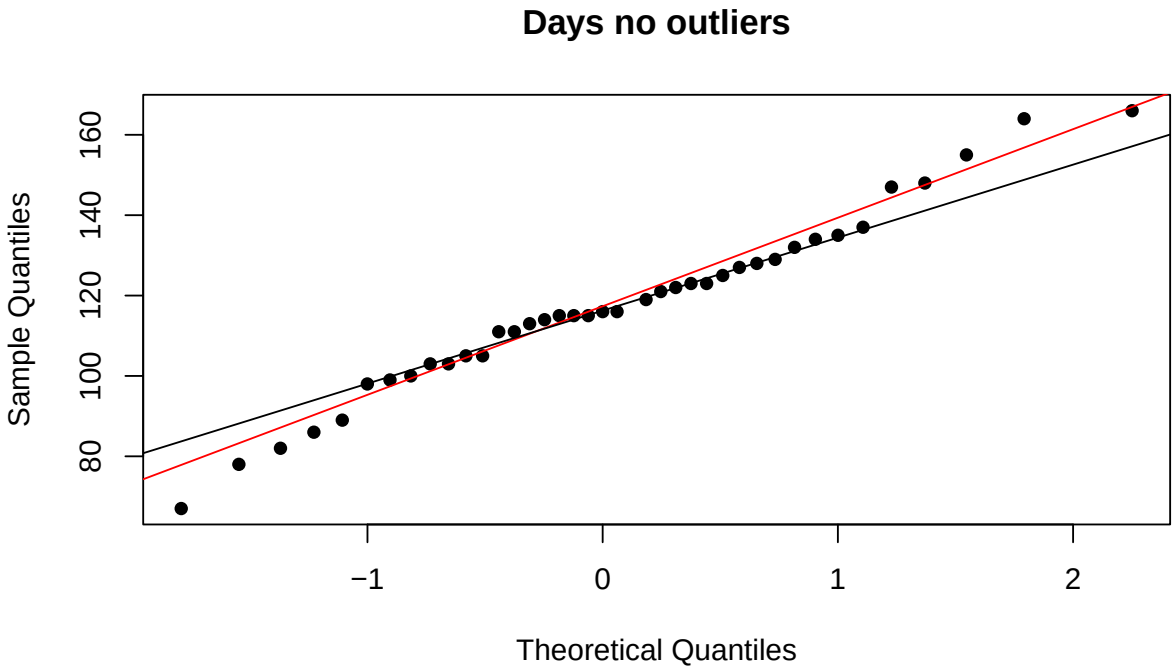


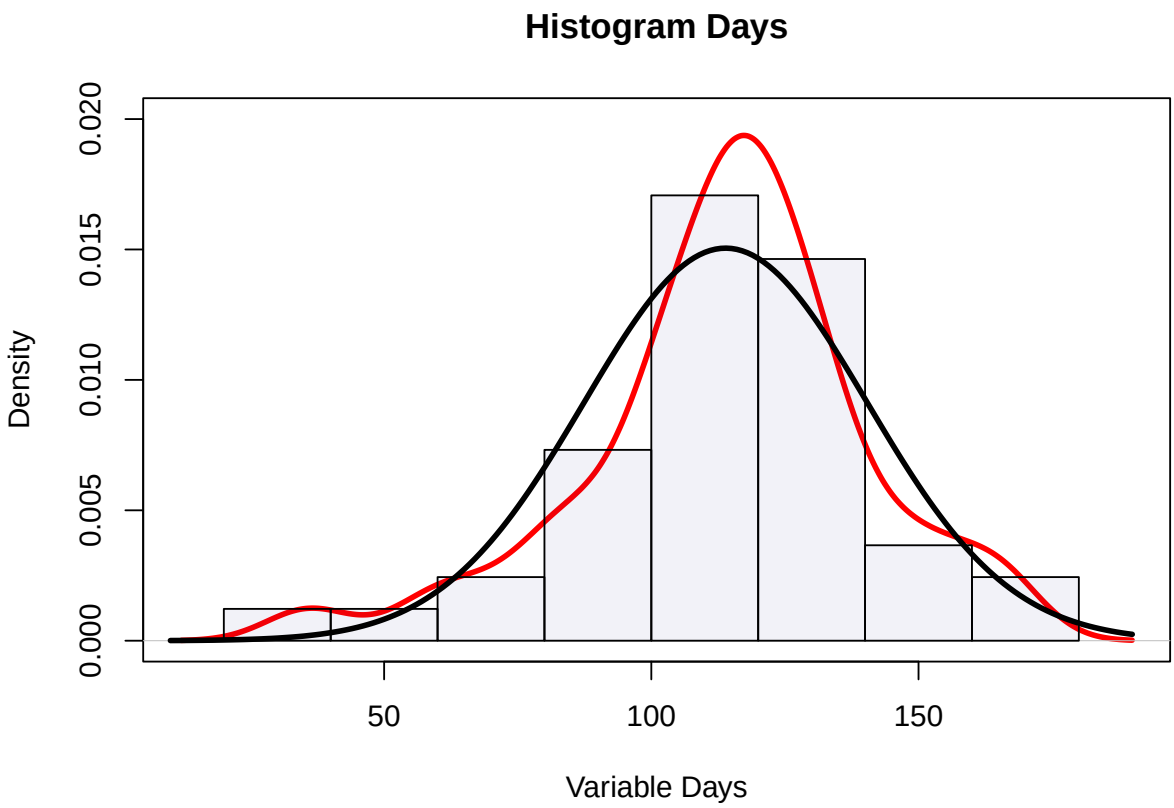
Variable *"Precip"* deviates from the Normal distribution especially for the lowest quantiles which are very far away from the two lines.

If we remove the two lowest observations (number *"1"* and *"23"*), we still observe a non Gaussian behaviour in particular for the smallest and largest measurements.

Looking at the histogram, the plot confirms that the variable has an unusual high frequency for the first values recorded and so we can conclude that probably this variable is not normally distributed. But like we saw for the variable *"Neg.Temp"*, Shapiro-Wilks test hires an high value, so the evaluation of univariate normality is not totally clear.


Days
Sample Quantiles
Theoretical Quantiles
40
60
80
120
160
−2
−1
0
1
2
1
23
25


Days no outliers
Sample Quantiles
Theoretical Quantiles
80
100
120
140
160
−1
0
1
2



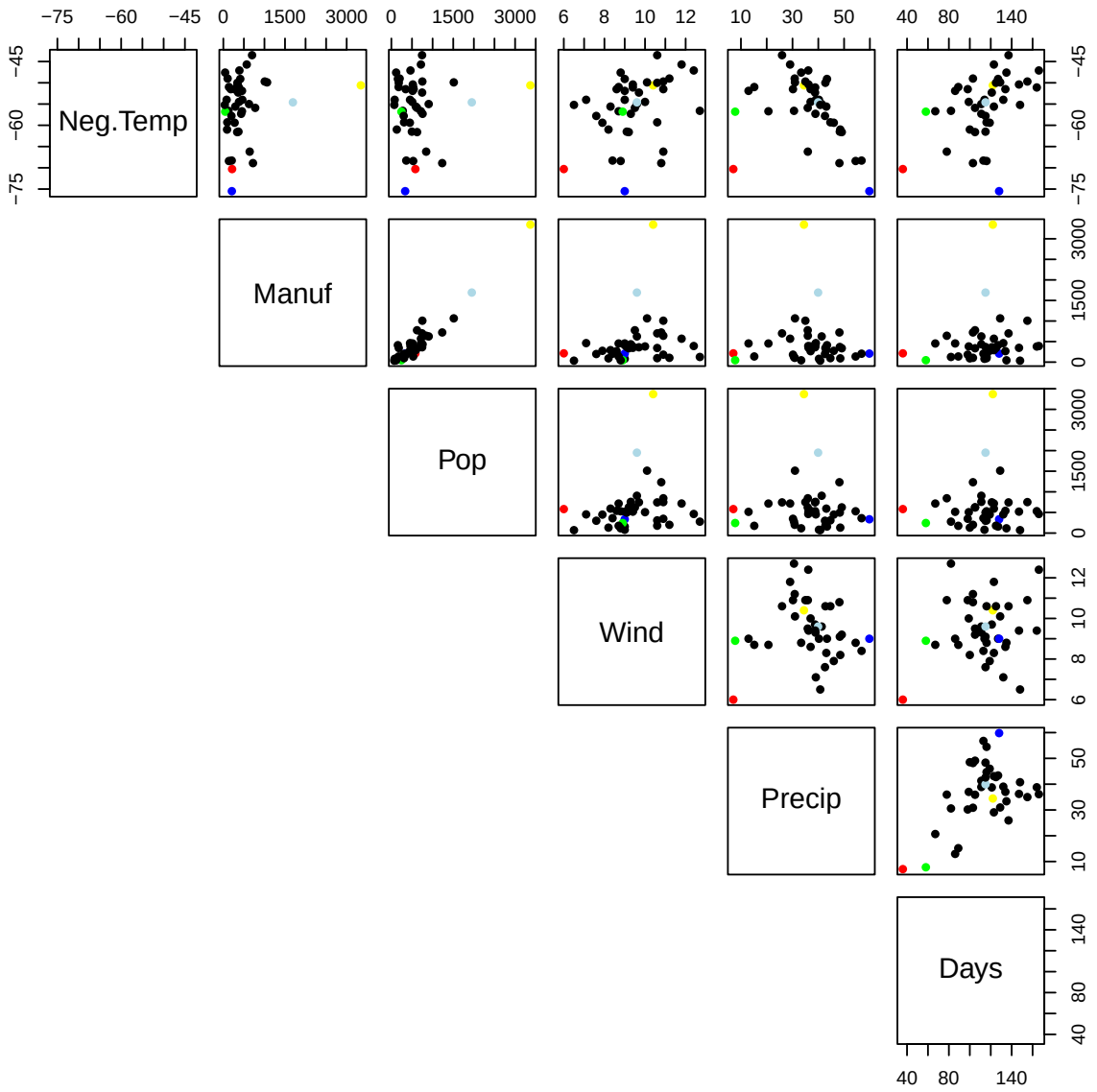
The Q-Q plot of the last variable seems to suggest that it isn't normally distributed in particular because it shows some lower quantiles which are distant from the red and the black line.

Removing the observations we detected as outliers (number *"11"* and *"23"*), the fit improve in particular with the red line, so the variables performs a good approximation of the Gaussian distribution. Indeed, here Shapiro-Wilks test doesn't reject the hypothesis of normality.

The histogram of *"Days"* is similar to the one observed for the variable *"Pop"* and this is consistent with the significant correlation we noticed above between the two variables.

1.4) By using scatterplots, comment on whether the outliers at point 1.2) can be detected from them:

We plot all the possible scatterplots we can obtain from the variables coloring the outliers detected at point *"1.2)"*:



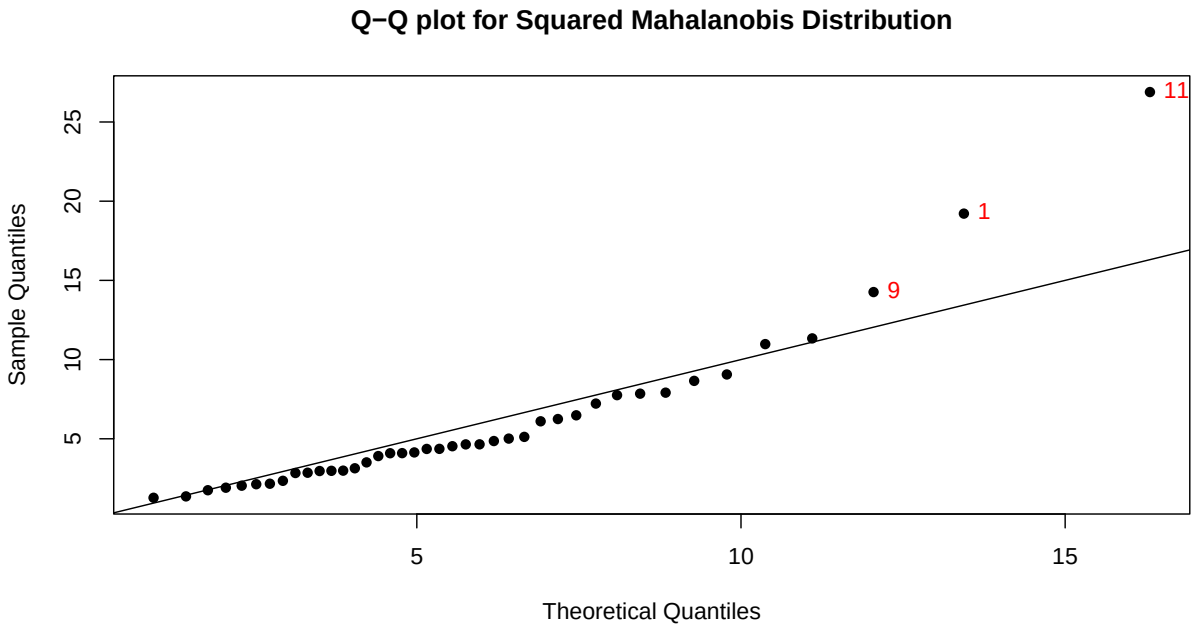
We observe that:

- **Outlier 1 red:** This observations may could be detected looking at the scatterpltos of the variable "$Wind$" vs "$Precip$", "$Neg.Temp$" and "$Days$" cause it appears at the extreme left lower corner of the plot but also from the scatterplots which involves the variable "$Days$" (indeed this observation was the strangest one we detected for this measurement).
- **Outlier 9 blue:** This outlier appears as an extreme observation just in some of the scatterplots in which we consider the variable "$Neg.Temp$", which is the only one that shows this point as an outlier.

- **Outlier 11 yellow:** This observation appears as a clear outlier in all the scatterplots that involves the two variables "Pop" and "$Manuf$".
- **Outlier 23 green:** This observation is not detectable as an outlier in any of the scatterplots cause it's always to much closer to the cloud of points.
- **Outlier 29 lightblue:** The same we said before for outlier 11.

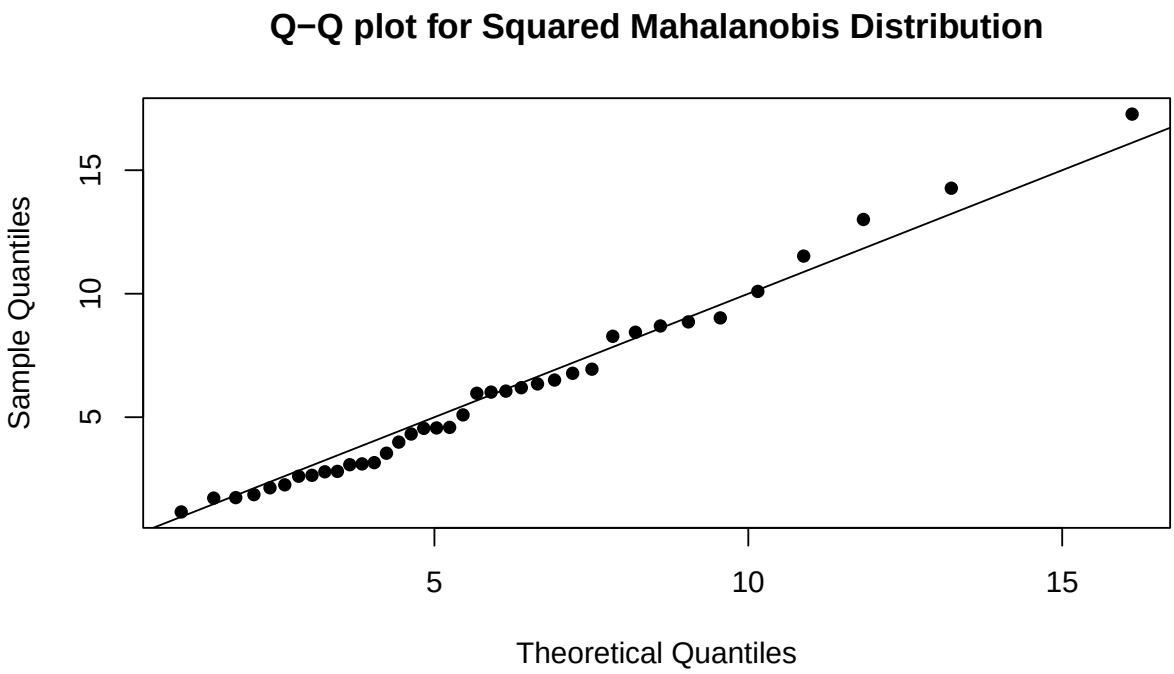
1.5) Construct a chi-square Q-Q plot of the squared Mahalanobis distances and comment about normality:

We show the Q-Q plot of the observed squared Mahalanobis distances, defined as: $(X-\mu)^T\Sigma(X-\mu)$ If data X are multivariate normally distributed, we expect the squared Mahalanobis distance random variable to be distributed as a χ^2_p, and so we'll see the observed distances to be close to the lines of the Q-Q plot.



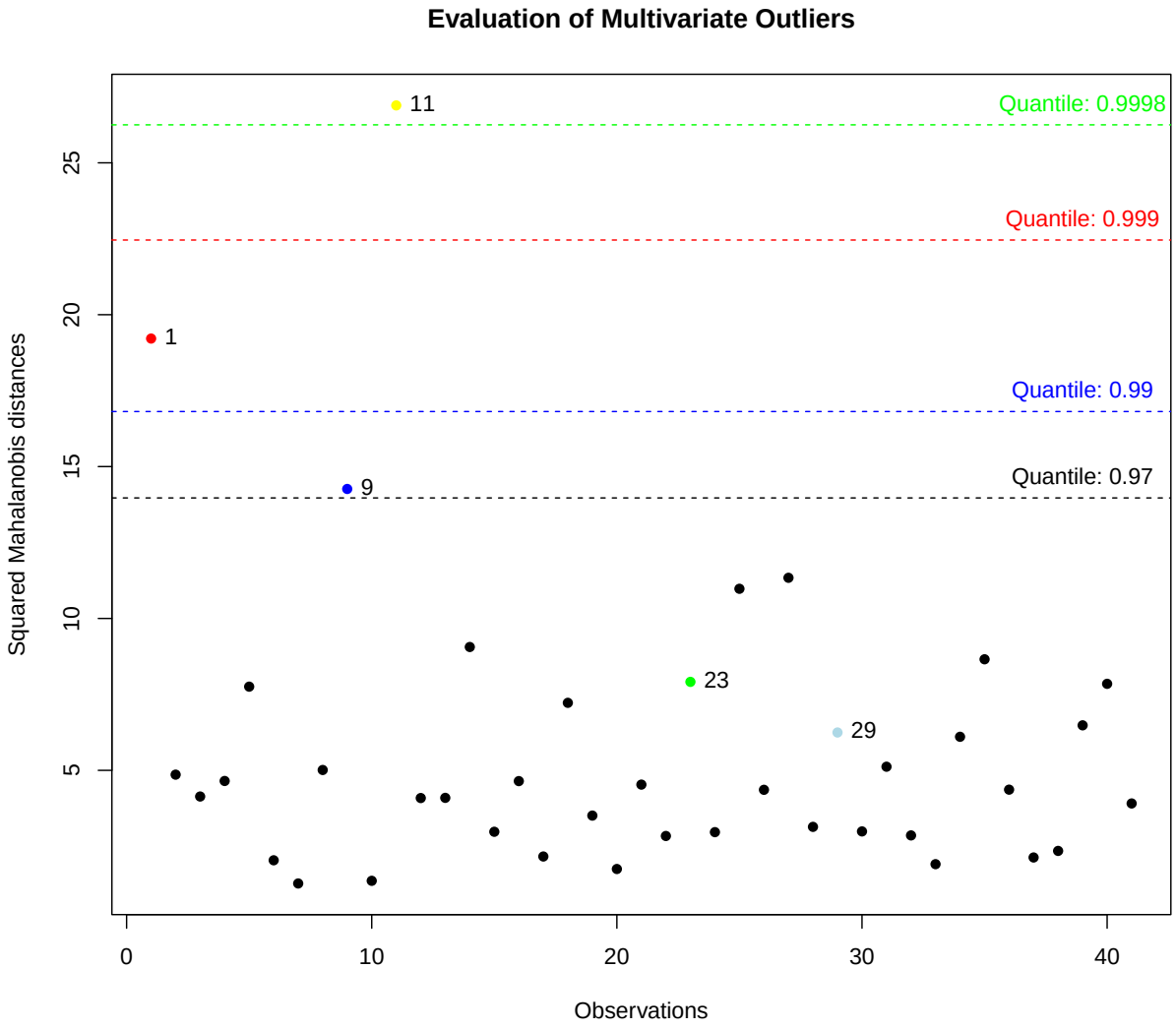
We see that the distances observed on the sample are really close to the line that passes through the 1^{st} and the $3rd$ quantile of the χ^2_6 except from the highest values which are observations *"9"*, *"1"* and *"11"* that we detected before as univariate outliers (we saw that *"11"* was an extreme outlier for the variables *"Pop"* and *"Manuf"* while *"1"* was considered as an outlier both for *"Precp"* and *"Days"* and *"9"* is the observation related to the minimum value for the *"Neg.Temp"* measurement). Producing the same plot without the three outliers we see that the points give a very good approximation of the χ^2_6 distribution, so we can conclude that data are not perfectly multivariate normally distributed (in particular variables *"Manuf"* and *"Pop"* are distant from a Gaussian distribution). But if we consider once again the Shapiro-Wilks tests without considering observations "1", "9" and "11" we obtain better results also for univariate normality:

```
##   Variable Statistic's value p-value observed
## 1 Neg.Temp         0.9535691     1.167743e-01
## 2    Manuf         0.8379978     6.866797e-05
## 3      Pop         0.8566081     1.876200e-04
## 4     Wind         0.9791330     6.867439e-01
## 5   Precip         0.9437118     5.533868e-02
## 6     Days         0.9834781     8.355821e-01
```


Q−Q plot for Squared Mahalanobis Distribution
Sample Quantiles
5
10
15
5
10
15
Theoretical Quantiles

1.6) Identify multivariate outliers, if any, and compare with the answer to point 1.2):

To detect potential multivariate outliers, we plot again the squared Mahalanobis distances to see if there're values which are over the highest quantiles of the χ^2_6 distribution (since we saw above that can be considered as multivariate Gaussian).



We can immediately see that there're some observations which hire an high value: indeed, *"1"* is over the 99%-quantile and *"11"* is over the 99,98%-quantile. Since they were detected before as univariate and bivariate outlier and since they appear now also as multivariate outlier, we can conclude that these two observations may can be excluded from the dataset. In addiction, we saw at the previous point that removing these observation the observed squared Mahalanobis distances fitted then Q-Q line of the theoretical quantiles of the χ^2_6 pretty much better than before.
Another particular observation which can be considered as a potential outlier is observation *"9"* that is over the 97%-quantile and was seen to be the only univariate outlier for the variable *"Neg.Temp"*.

Exercise 2:

Description of the problem:

Let $X = (X_1, X_2, X_3) \sim \mathcal{N}_3(\mu, \Sigma)$, where:

$$\mu = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ o & \rho & 1 \end{bmatrix}, \text{ with } |\rho| < \sqrt{2}/2.$$

2.1) Let PC_1 and PC_2 be the first two population principal components of X. Find ρ such that they account for more than 80 percent of total vairation of X:

From the theory, we know that the first two Principal Components are two linear combinations of the original variables $X = (X_1, X_2, X_3)$ defined as:

$PC_1 = a_1^T X$, $PC_2 = a_2^T X$

Where the vector of coefficients are respectively equal to the eigenvectors of the Population Covariance Matrix Σ, related to the first two eigenvalues of Σ put in decreasing order, i.e.:

$a_i = e_i$ where $\Sigma e_i = \lambda_i e_i$, $i = 1, 2, 3$ and we set $\lambda_1 \geq \lambda_2 \geq \lambda_3$

In addiction we consider the eigenvectors for the different eigenvalues such that they're orthogonal between each other and with unitary norm, i.e.

$$e_i e_j^T = \begin{cases} 1 & if \quad j = i \\ 0 & if \quad j \neq i \end{cases}$$

In addiction, we know that is required to determine ρ in order to give at least the 80% of the total variance to be explained by PC_1 and PC_2. The proportion of variance is computed through the ratio $\frac{\lambda_1 + ... + \lambda_k}{\lambda_1 + ... + \lambda_p}$ since we know that the variance explained by each $j - th$ component is equal to the $j - th$ eigenvalue λ_j and the total variance of the data is $trace(\Sigma) = {s_1}^2 + ... + {s_p}^2 = \lambda_1 + ... + \lambda_p$.
Since in that case $p = 3$, $k = 2$ and $trace(\Sigma) = 3$, we need to impose the following condition: $\frac{\lambda_1 + \lambda_2}{3} = 0.8$

So, to find the vectors of coefficients for PC_1 and PC_2 we first need to determine λ_1 and λ_2 finding the solutions of the characteristic polynomial of the Covariance Matrix $p_\Sigma(x) = 0$:

$$p_\Sigma(x) = \det(\Sigma - \lambda I_3) = \begin{vmatrix} (1-\rho) & \rho & 0 \\ \rho & (1-\rho) & \rho \\ 0 & \rho & (1-\rho) \end{vmatrix} = 0$$

The computation leads to the following solutions: $\lambda = 1 \pm \rho\sqrt{2}$ and $\lambda = 1$. So we need to distinguish two cases we can face in order to extract correctly the first two eigenvalues:

- $\rho > 0$**:**
  In that case we choose the eigenvalues $\lambda = 1 + \rho\sqrt{2}$ and $\lambda = 1$, so considering the two constraints we put on the values acceptable for ρ and on the minimum percentage of variance we want to explain with the first 2 Principal Components, we find the following range of solutions: $\frac{\sqrt{2}}{5} < \rho < \frac{\sqrt{2}}{2}$
- $\rho < 0$**:**
  In that case we choose the eigenvalues $\lambda = 1 - \rho\sqrt{2}$ and $\lambda = 1$, so we find the following range of solutions: $-\frac{\sqrt{2}}{2} < \rho < -\frac{\sqrt{2}}{5}$.

2.2) Give an interpretation to PC_1 and PC_2 in terms of the original variables:

To give an interpretation of the Principal Components in terms of the original variables we need to determine the coefficients of each components: a_{jk}. These coefficients are called *"loadings"* and they express how much the k-th variables loads on the j-th component. So we need to compute the vector of coefficients a_1 and a_2, i. e. the eigenvectors e_1 and e_2 solving for each of the the system defined by:

$$(\Sigma - \lambda_j I_3)x_j = 0_{R^3} \quad j = 1, 2$$

In that way we'll find the structure of the eigenvectors and finally we'll need to search for the ones which have norm equal to "1", i. e.: $e_j = \frac{x_j}{||x_j||} \quad j = 1, 2$.
We still need to distringuish the results depending on the sign of ρ and finally we obtain:

- $\rho > 0$:
  In that case the eigenvectors are $e_1 = \begin{bmatrix} \frac{1}{2} \\ \frac{\sqrt{2}}{2} \\ \frac{1}{2} \end{bmatrix}$, $e_2 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ 0 \\ -\frac{\sqrt{2}}{2} \end{bmatrix}$.
  Looking at the loadings, we can state that PC_1 is a weighted average of all the 3 original variables and we can underline an higher importance of X_2 with respect to the other two measurements. Instead, PC_2 is a weighted difference between X_1 and X_3 without considering at all the 2^{nd} variable.
- $\rho < 0$:
  In that case the eigenvectors are $e_1 = \begin{bmatrix} \frac{1}{2} \\ -\frac{\sqrt{2}}{2} \\ \frac{1}{2} \end{bmatrix}$, $e_2 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ 0 \\ -\frac{\sqrt{2}}{2} \end{bmatrix}$.
  Here, PC_1 is a weighted average of X_1, X_2 and X_3 with alternate signs, indeed there's a positive relationship with X_1 and X_3 while the 2^{nd} variables expresses a negative contribution which is the highest one in absolute terms (like we saw in the previous case). PC_2 remains the same as before.

2.3) Find the distribution of Z:

We define Z to be:

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

Since Z is defined as a vector of linear combinations of normally distributed Random Variable, we know that: $Z_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ where we can compute the mean and the standard deviation of the distribution using the known vector of coefficients of Z_1, which is $a_1 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$:

$\mu_1 = a_1^T \mu = 2$,
$\sigma_1 = a_1^T \Sigma a_1 = 2(1 - \rho)$
And we say that $Z_1 \sim \mathcal{N}(2, 2(1 - \rho))$

In the same way we find the distribution of Z_2, with vector of coefficients $a_2 = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$:

$\mu_2 = a_2^T \mu = -3$,
$\sigma_2 = a_2^T \Sigma a_2 = 2(1 - \rho)$
And we say that $Z_2 \sim \mathcal{N}(-3, 2(1 - \rho))$

Then we need to compute the covariance between the two Random Variables in order to obtain the Covariance Matrix. We apply the following formula:

$s_{12} = a_1^T \Sigma a_2 = 2\rho - 1$

Finally, we can state correctly the distribution of Z:

$Z \sim \mathcal{N}_2(\mu_z, \Sigma_z)$, where $\mu_z = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$ and $\Sigma_z = \begin{bmatrix} 2(1-\rho) & 2\rho - 1 \\ 2\rho - 1 & 2(1-\rho) \end{bmatrix}$

2.4) Setting $\rho = -2/3$, sketch the ellipse defined by the Squared Mahalanobis distance Random Variable such tthat it contains 0.95 probability with respect to the joint distribution of Z:

When $\rho = -2/3$, the Population Mean Vector and the Population Covariance Matrix are:

$\mu_z = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$ and $\Sigma_z = \begin{bmatrix} \frac{10}{3} & -\frac{7}{3} \\ -\frac{7}{3} & \frac{10}{3} \end{bmatrix}$

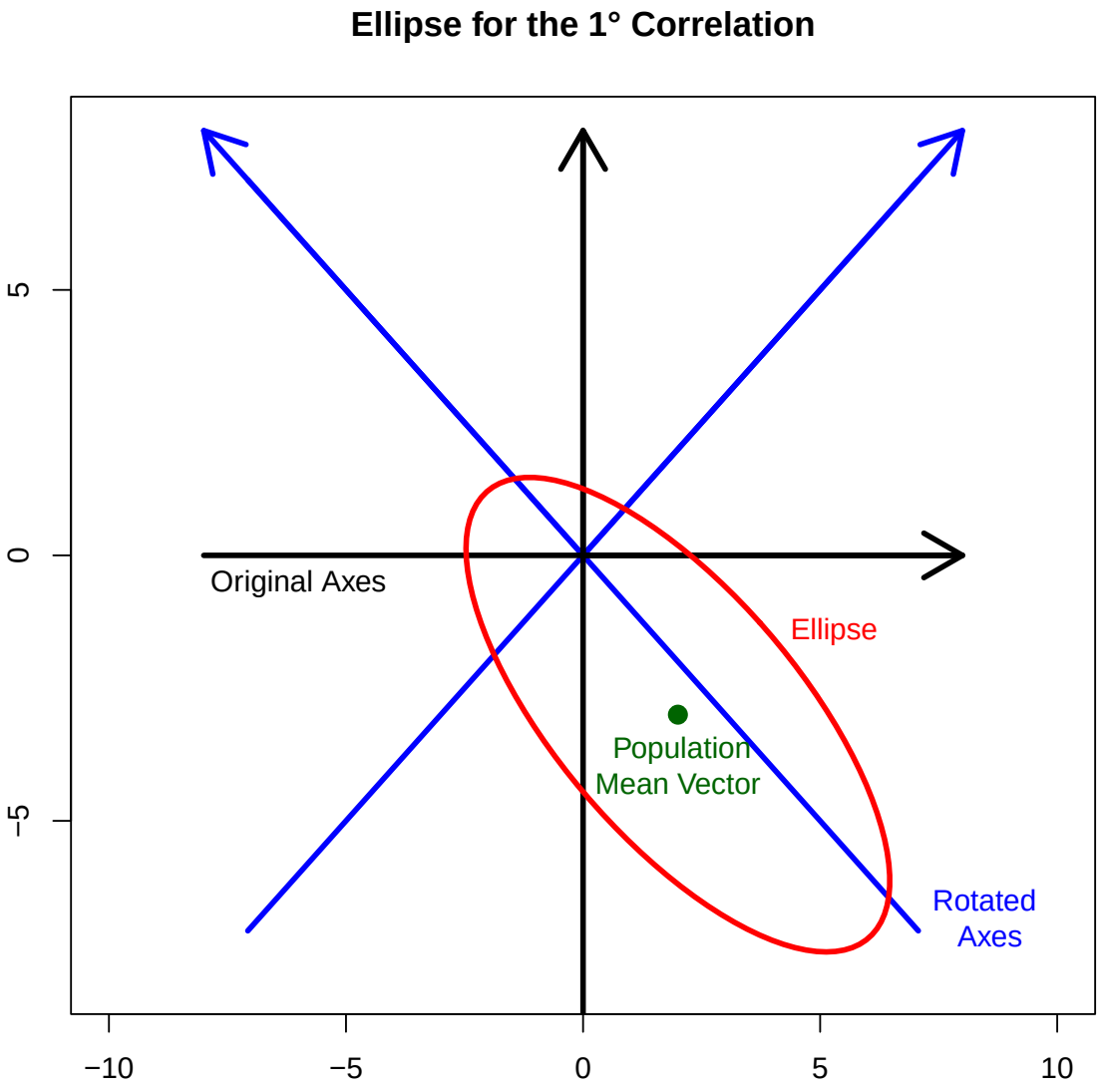
The ellipse we need to plot given by:$(z - \mu_z)^T \Sigma (z - \mu_z) = c^2$. From the theory, we know that since $Z \sim \mathcal{N}_2(\mu_z, \Sigma_z)$, then the Squared Mahalanobis distance Random Variable is distributed as a X_2^2 and we know that the set of points given by the above equation forms an ellipse on the 2-dimensional spaze $z = (z_1, z_2)$, centered around the Population Mean Vector μ_z, with axes oriented in the direction determined by the eigenvectors of the Population Covariance Matrix Σ_z and half-lengths of the axes proportional to the square roots of the eigenvalues of Σ_z.
To draw the ellipse, first of all we need to find the correct value of c using the quantiles of the X_2^2. In particular, since it's required that the ellipse contains 0.95 probability with respect to the joint distribution of Z, we set $\alpha = 0.05$ and we consider the quantile $X^2_{2,0.05}$. So, we find c in such a way:

$c = \sqrt{X^2_{2,\alpha}} = \sqrt{5.99} \simeq \pm 2.448$

Then we need to determine the eigenvalues and eigenvectors of Σ_Z (like we did at point 2.1)) when $\rho = -2/3$, so we can sketch the ellipse. Proceeding by computing first the solutions of the characteristic polynomial $p_{\Sigma_z}(x) = 0$ (so we find the eigenvalues λ_1 and λ_2) and solving the system given by $((\Sigma_z - \lambda_j I_2)x_j = 0_{R^2} \quad j = 1, 2)$ we obtain the following results:

- $\lambda_1 = \frac{17}{3}$ and $e_1 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix}$:
  So the longest ax corresponds to the bisector of the 2^{nd} and 4^{th} quadrants and it's the one in which the ellipse shows the longest direction.
- $\lambda_2 = 1$ and $e_2 = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$:
  So the shortest ax corresponds to the bisector of the 1^{st} and 3^{rd} quadrants and it's the one in which the ellipse shows the shortest direction.

Here we plot the ellipse:



2.5) Comment on how the ellipse would change with $\rho = 2/3$ (no need to draw it):

Changing the sign of the correlation ρ we see a change in the Covariance matrix Σ_z which becomes: $\Sigma_z = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{bmatrix}$ and so we will expect a change in the eigenvectors and eigenvalues of Σ_z which will produce some effects in the shape of the ellipse.

Computing the new eigenvalues and eigenvectors of Σ_z we find that:

- $\lambda_1^{'} = 1$ and $e_1^{'} = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}$: So the longest ax corresponds to the bisector of the 1^{st} and 3^{rd} quadrants and it's the one in which the ellipse shows the longest direction.
- $\lambda_2^{'} = \frac{1}{3}$ and $e_2^{'} = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix}$: So the shortest ax corresponds to the bisector of the 2^{nd} and 4^{th} quadrants and it's the one in which the ellipse shows the shortest direction.

About the effects on the ellipse shifting the value of ρ from $-\frac{2}{3}$ to $\frac{2}{3}$ we can state that:

- **Centre:** It remains the same since the Population Mean Vector μ_z is not affected by the value of ρ.
- **Direction of the axes:** The 2 ellipses show the same pair of eigenvectors (e_1, e_2) and so in both cases the axes correspond to the bisectors of the quadrants.
- **Orientation:** The 2^{nd} ellipse has inverted axes with respect to the previous one, since $e_1^{'} = e_2$ and $e_2^{'} = e_1$. This leads to a different orientation since the 1^{st} ellipses stretches more along the bisector of the 2^{nd} and 4^{th} quadrant (resulting "right oriented") while the other ellipses stretches more along the bisector of the 1^{st} and 3^{rd} quadrant (resulting "left oriented").
- **Dimensions:** We recall the formulas to determine the vertices of the ellipses:

  $\pm c\lambda_1^{1/2} e_1$ for the horizontal ones
  $\pm c\lambda_2^{1/2} e_2$ for the vertical ones

  Then, since when $\rho = -\frac{2}{3}$ the eigenvalues are $\lambda_1 = \frac{17}{3}$, $\lambda_2 = 1$ and when $\rho = \frac{2}{3}$ the eigenvalues $\lambda_1^{'} = 1$, $\lambda_2{'} = \frac{1}{3}$; it follows that:

  $c\lambda_1^{1/2} e_1 > c\lambda_1^{'1/2} e_1^{'}$ and $c\lambda_2^{1/2} e_2 > c\lambda_2^{'1/2} e_2^{'}$

  i. e., the 2^{nd} ellipse is smaller than the 1^{st} one.

Exercise 3: Pen digit Data

Data Description

The **Pen Digit** data set was created by collecting 250 samples from 44 writers.

These writers were asked to write 250 digits in random order inside boxes of 500 by 500 tablet pixel resolution.

The raw data on each of $n = 10992$ handwritten digits consisted of a sequence, (x_t, y_t), $t = 1, 2, ..., T$, of tablet coordinates of the pen at fixed time intervals of 100 milliseconds, where x_t and y_t were integers in the range 0-500.

These data were then normalized to make the representations invariant to translation and scale distortions. The new coordinates were such that the coordinate that had the maximum range varied between 0 and 100. Usually x_t stays in this range, because most integers are taller than they are wide.

Finally, from the normalized trajectory of each handwritten digit, 8 regularly spaced measurements, (x_t, y_t), were chosen by spatial resampling, which gave a total of $p = 16$ variables. The data includes a *"class attribute"*, column *"digit"*, coded $0, 1, ..., 9$, about the actual digit.

Data importation

We display the first rows of the dataset:

```
##    x1  y1 x2  y2  x3  y3  x4  y4 x5 y5  x6 y6  x7 y7  x8 y8 digit
## 1  47 100 27  81  57  37  26   0  0 23  56 53 100 90  40 98     8
## 2   0  89 27 100  42  75  29  45 15 15  37  0  69  2 100  6     2
## 3   0  57 31  68  72  90 100 100 76 75  50 51  28 25  16  0     1
## 4   0 100  7  92   5  68  19  45 86 34 100 45  74 23  67  0     4
## 5   0  67 49  83 100 100  81  80 60 60  40 40  33 20  47  0     1
## 6 100 100 88  99  49  74  17  47  0 16  37  0  73 16  20 20     6
```

3.1) Perform a principal component analysis on the standardized variables. Report standard deviations. Decide how many components to retain in order to achieve a satisfactory lower-dimensional representation of the data. Justify your answer:

First of all, we standardize the data and we remove from the *"class attribute"* in order to perform the Principal Component analysis.

```
##             x1         y1         x2          y2         x3          y3
## [1,]  0.2389437  0.9174502 -0.5164799 -0.14476365  0.2120092 -1.05839444
## [2,] -1.1330073  0.2392153 -0.5164799  0.84669693 -0.2278669  0.34918565
## [3,] -1.1330073 -1.7338315 -0.3646368 -0.82313142  0.6518852  0.90480937
## [4,] -1.1330073  0.9174502 -1.2756953  0.42923984 -1.3128945  0.08989458
## [5,] -1.1330073 -1.1172544  0.3186571 -0.04039938  1.4729872  1.27522518
## [6,]  1.7860374  0.9174502  1.7991271  0.79451480 -0.0225914  0.31214407
##             x4          y4          x5          y5         x6         y6
## [1,] -0.8248144 -1.48795710 -1.66596334 -0.39249002 -0.1211211  0.6701144
## [2,] -0.7267010  0.01675246 -1.22653849 -0.68605132 -0.6306673 -1.2841642
## [3,]  1.5953147  1.85584192  0.56045591  1.51565845 -0.2820304  0.5963680
## [4,] -1.0537455  0.01675246  0.85340581  0.01115677  1.0588804  0.3751289
## [5,]  0.9739302  1.18708212  0.09173607  0.96523100 -0.5502126  0.1907630
## [6,] -1.1191544  0.08362844 -1.66596334 -0.64935616 -0.6306673 -1.2841642
##             x7         y7           x8         y8
## [1,]  2.0137285  1.6607506 -0.174506993  1.9328771
```

```
## [2,]  0.6258058 -0.9934123  1.262260860 -0.6385288
## [3,] -1.2098337 -0.2997106 -0.749214134 -0.8062292
## [4,]  0.8496643 -0.3600325  0.472038541 -0.8062292
## [5,] -0.9859753 -0.4505154 -0.006884077 -0.8062292
## [6,]  0.8048926 -0.5711591 -0.653429611 -0.2472279
```

Secondly, we perform PCA on the standardized variables through the built-in command we find in R and we display the first rows of the Principal Components obtained:

```
##            PC1         PC2          PC3         PC4          PC5         PC6
## x1 -0.05068856 -0.03325051  0.451265448 -0.25236086  0.028138114 -0.52350244
## y1 -0.13382327 -0.12534138  0.339642019 -0.09379930 -0.451900048  0.20825345
## x2  0.26251698 -0.11691748  0.201017586 -0.35772400  0.416567660 -0.02755209
## y2  0.30197656 -0.24219672  0.205563498  0.04599805 -0.304927844  0.16126446
## x3  0.25131576 -0.09062510 -0.305548548 -0.29830804  0.285678513  0.43909340
## y3  0.42375760 -0.05098217  0.005069603  0.11755889  0.007455499  0.02006344
##             PC7         PC8          PC9         PC10        PC11        PC12
## x1  0.110257579 -0.47680835  0.042582174 -0.208348582  0.07006321 -0.26374500
## y1 -0.670818476 -0.09883220 -0.079360557  0.008417937  0.32468237  0.08011023
## x2 -0.002906442  0.05836556 -0.643071163 -0.017575754  0.03422481  0.13960585
## y2  0.028845132 -0.16928359 -0.174198314  0.323636372 -0.62744518  0.02787127
## x3 -0.247305368 -0.12657555 -0.002175622 -0.191163159  0.11538009 -0.19308481
## y3  0.040922393 -0.17633014  0.120033020  0.477884290  0.16312697 -0.33394801
##           PC13         PC14        PC15         PC16
## x1  0.2906300 -0.089062393  0.04463247  0.023800306
## y1 -0.1149746 -0.001654317 -0.05863589  0.003488715
## x2 -0.3210719  0.144694827 -0.11289579 -0.007622158
## y2  0.2008865  0.109226618  0.27397394  0.020128625
## x3  0.4480635 -0.251236546  0.20813113  0.018484989
## y3 -0.1583327 -0.328153558 -0.49441960  0.112435258
```

Then we report the standard deviations of each principal components, that we know to be equal to the square roots of the decreasingly ordered eigenvalues $\lambda_1 \geq ... \geq \lambda_p$ of the Sample covariance Matrix S:

```
##  [1] 2.1717703 1.7969730 1.6052434 1.1089372 1.0310671 0.8929386 0.7788784
##  [8] 0.7404401 0.6409556 0.5461052 0.4588166 0.3351073 0.2836918 0.2408363
## [15] 0.1851067 0.1667316
```

The, to decide how many Principal Components we should retain in order to achieve a satisfactory dimension reduction of data, we can exploit 3 different method to make this evaluation:

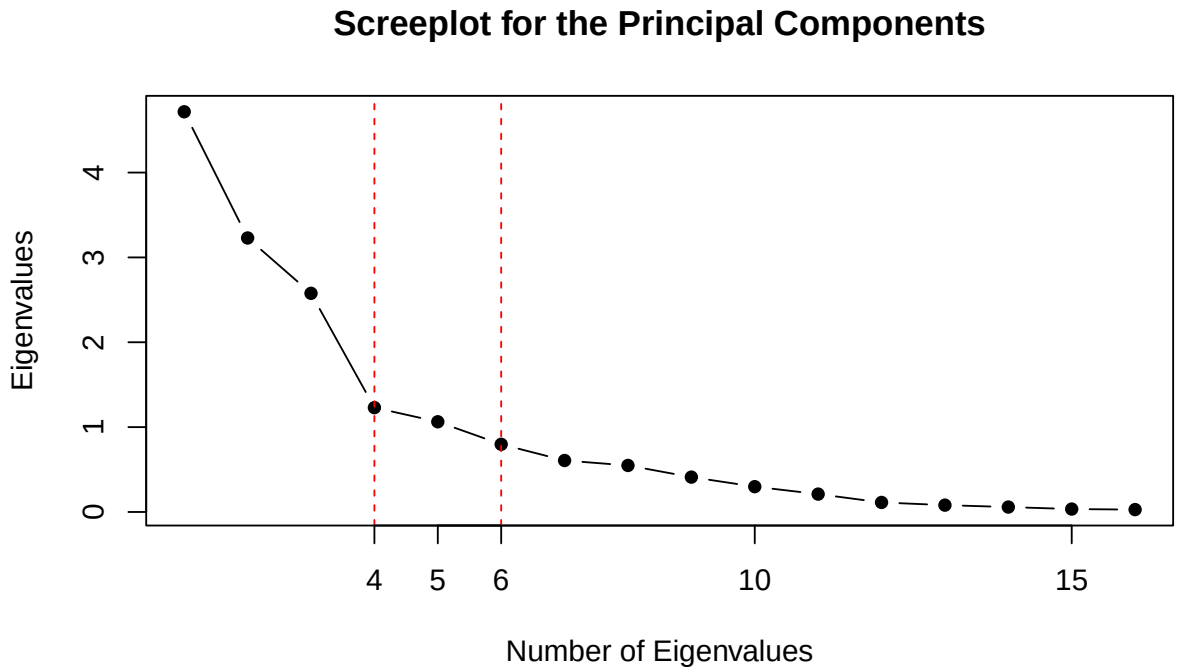
Method 1: Proportion of Variance explained

We choose the first k Principal components that together explain a sufficient proportion of the total variance we have in the data. We say that the proportion has an acceptable level when we reach around the 80%. We show the result:

```
## Importance of components:
##                            PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.1718 1.7970 1.6052 1.10894 1.03107 0.89294 0.77888
## Proportion of Variance 0.2948 0.2018 0.1610 0.07686 0.06644 0.04983 0.03792
## Cumulative Proportion  0.2948 0.4966 0.6577 0.73452 0.80096 0.85079 0.88871
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.74044 0.64096 0.54611 0.45882 0.33511 0.28369 0.24084
```

```
## Proportion of Variance 0.03427 0.02568 0.01864 0.01316 0.00702 0.00503 0.00363
## Cumulative Proportion  0.92297 0.94865 0.96729 0.98045 0.98747 0.99250 0.99612
##                           PC15    PC16
## Standard deviation     0.18511 0.16673
## Proportion of Variance 0.00214 0.00174
## Cumulative Proportion  0.99826 1.00000
```

Method 2: Screeplot
We plot the relationship between the number of eigenvalues (so we'll have $j = 1, ..., p$ on the x-ax) and the value assumed by the corresponding eigenvalue (so we'll have the corresponding $\lambda_1, ..., \lambda_p$ on the y-ax).This graphical method helps to determine a correct number of Principal Components to retain cause we can see the point in which the variance explained by the $k + 1$-th component is considerably low compared to the previous k-th components. This point is called "*elbow*" and usually, if it appears at the point $x = k + 1$ we choose to keep just the first k components.



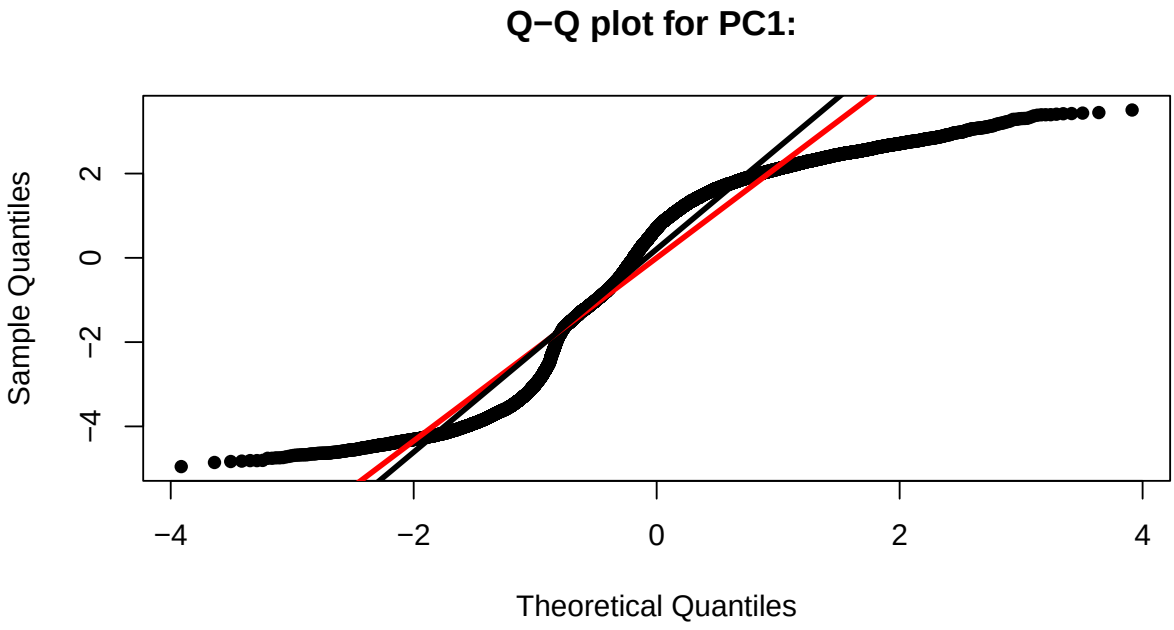
Method 3: Eigenvalues higher than the their mean value
This last method consists in evaluating which eigenvalues are higher than their average value and in retaining just the components associated to those λ_j's. We show the results:

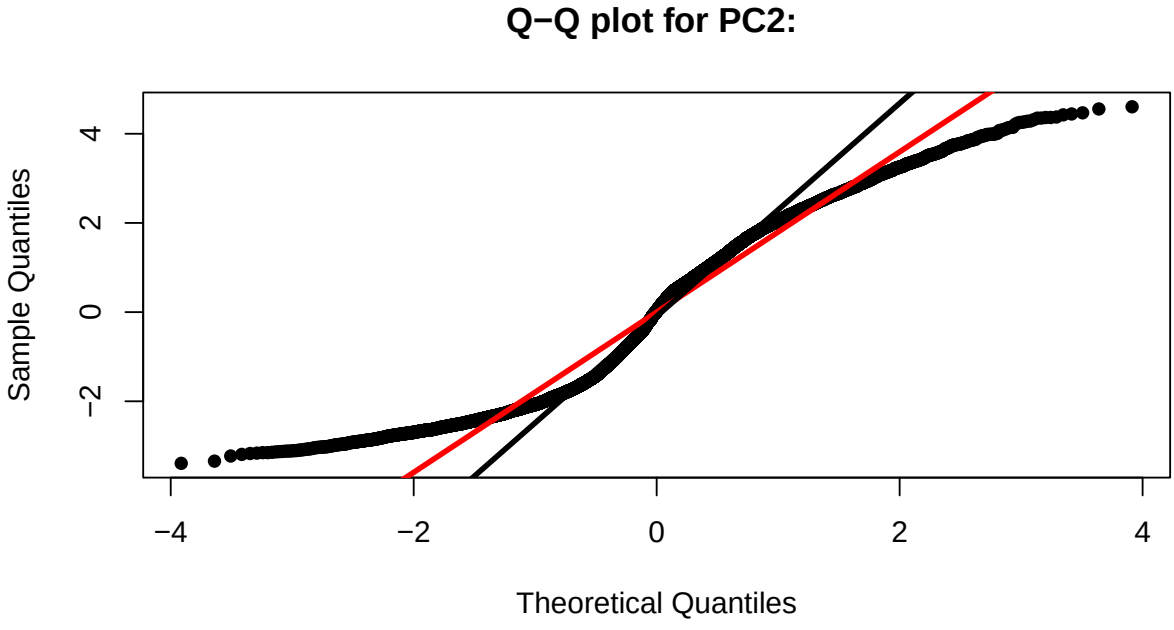
```
##  [1]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [13] FALSE FALSE FALSE FALSE
```

Comparing the 3 methods we can deduce that maybe a correct number of components to retain is $k = 5$ since together they explain around the 80% of the total variance, they're the middle number of components to retain between the two possible *"elbows"* detected in the Screeplot and they're related to the eigenvalues which are higher than the average value, so they satisfies both the 3 methods used in the evaluation.

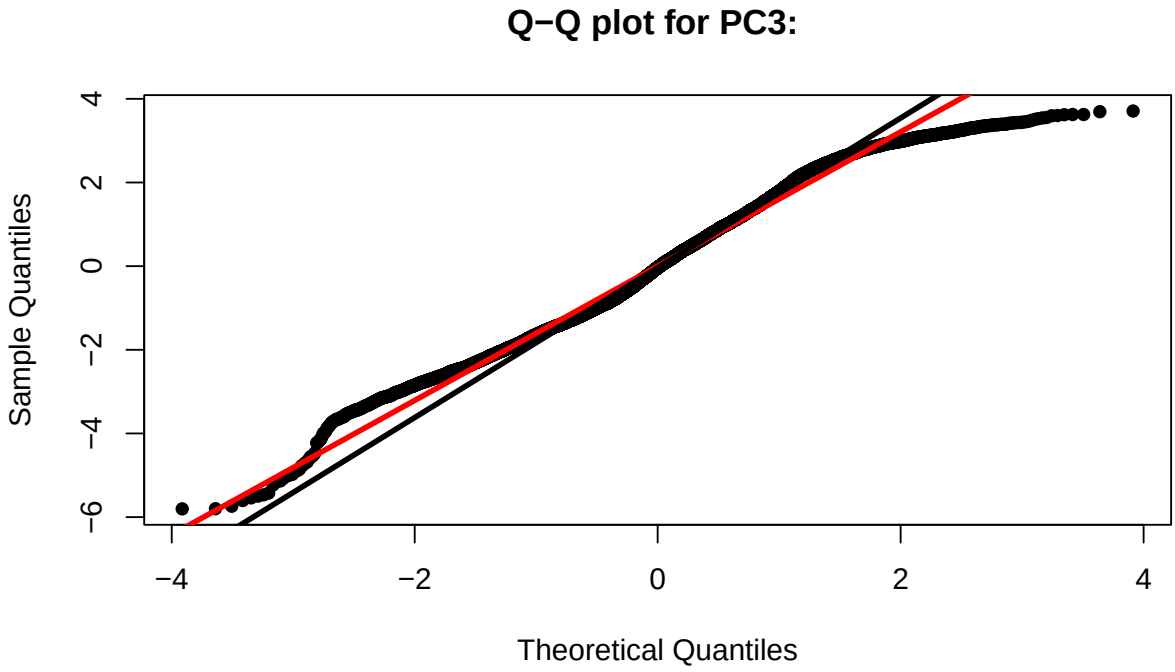
3.2) Investigate multivariate normality through the first three principal components:

Univariate Normality of the Principal Components: To investigate about multivariate normality of data, we can plot the Q-Q plot of the first 3 Principal Components, in order to see if they can be considered as univariate Gaussians and try to conclude, as a consequence, about the normality of the original variables (since PC_1, PC_2 and PC_3 are linear combinations of the original measurements, so if the original variables were normally distributed, we should obtain the same conclusion for these linear combinations).



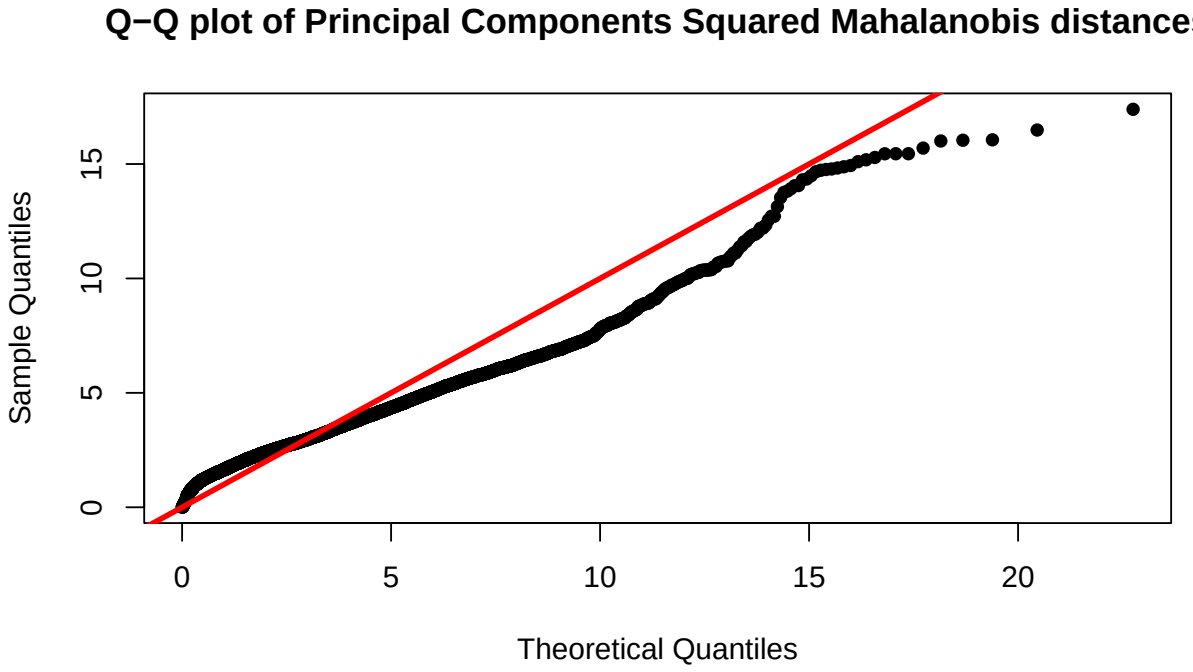


We can see clearly that the Q-Q plots of PC_1 and PC_2 shows that these components are not normally distributed since the observations never follows the two lines they should match with under the assumption of Gaussian data.



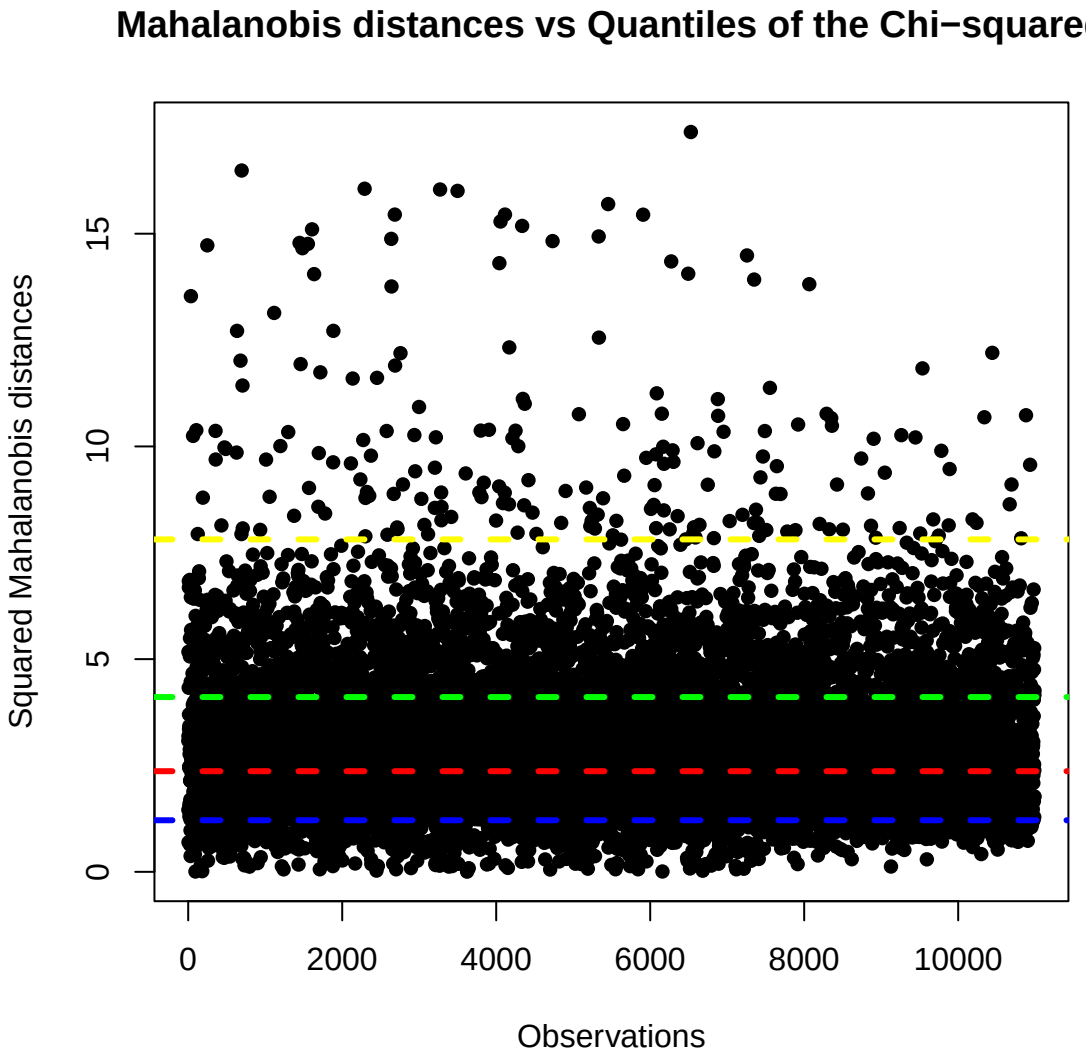
The situation is a little bit different if we look at the Q-Q plot of PC_3 because in that case the points fit well with the two lines a part for the highest values observed which deviate from the theoretical quantiles. To sum up, the Q-Q plots doesn't show a clear Normal Distribution for the first 3 Principal Components.

Multivariate Normality of th Principal Components: To study the multivariate normality of the Principal Components we can look first of all at the Q-Q plot the Squared Mahalanobis distances computed for PC_1, PC_2 and PC_3 against the theoretical quantiles of the χ^2_3 to evaluate if these 3 Principal Components are multivariate Gaussian or not (like we did n Exercise 1, point 1.6).



We can clearly see that the points deviate considerably from the red line which represents the value they should hire if data were multivariate Gaussian. So from this first plot we can't assume a normality assumption for our Principal Components.
Another tool we can use to evaluate multivariate normality, is t plot again the squared Mahalanobis distances against the values of the quantiles of χ^2 and use a numerical criterion to evaluate if those distances can be considered distributed as a χ^2_3:

Numerical Criterion:
If the Squared Mahalanobis distances are distributed as a χ^2_3 we should expect to find around the $100 \cdot (1-\alpha)\%$ of observations above the quantile χ^2_α. So, we choose the levels $\alpha = 0.25, 0.5, 0.75, 0.95$ to make our evaluation and if data are multivariate Gaussian we should expect to see respectively the 75%, 50%, 25% and 5% of the observations above the chosen quantiles.
We first plot the distances and the chosen quantiles (0.25 in blue, 0.5 in red, 0.75 in green and 0.95 in yellow):



And then we compute the percentage of elements above those quantiles to see if they match or not with the expected ones under assumption of multivariate normality:

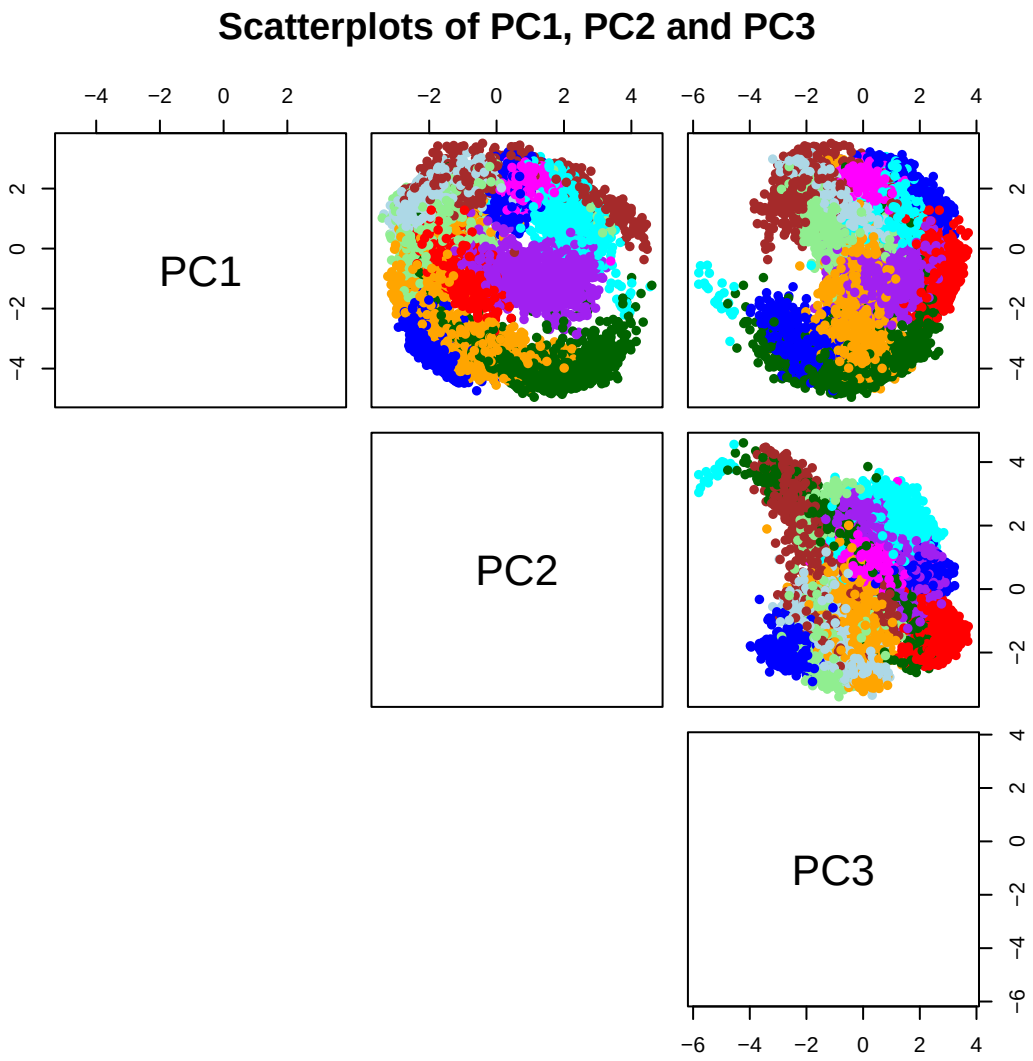
```
##                 Number of elements Observed Percentage Expected Percentage
## Quantile 0.25:               10093              91.82%                 75%
## Quantile 0.5:                 6441               58.6%                 50%
## Quantile 0.75:                2195              19.97%                 25%
## Quantile 0.95:                 203               1.85%                  5%
```

This second tool used to evaluate multivariate normality seems to confirm the consideration we deduced before looking at the Q-Q plot since the observed percentages of data above the quantiles are really different from the ones we would expect to see under assumption of multivariate normality.
So, according to the previous considerations, we can conclude that PC_1, PC_2 and PC_3 are not normally distributed.

Conclusions: Summing up all the evaluation of normality performed over PC_1, PC_2 and PC_3 we can conclude that data doesn't show a multivariate Normal Distribution.

3.3) Make scatterplots with the first three principal components, while color coding the observations according to the digit class:

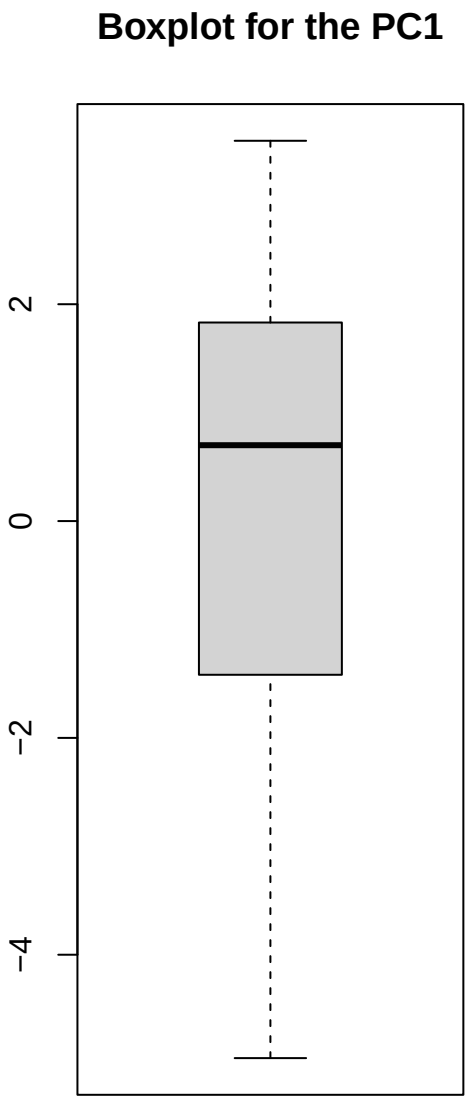
We print the scatterplot matrix of the first 3 Principal Components coloring the observations with respect to the variable *"digit"*:

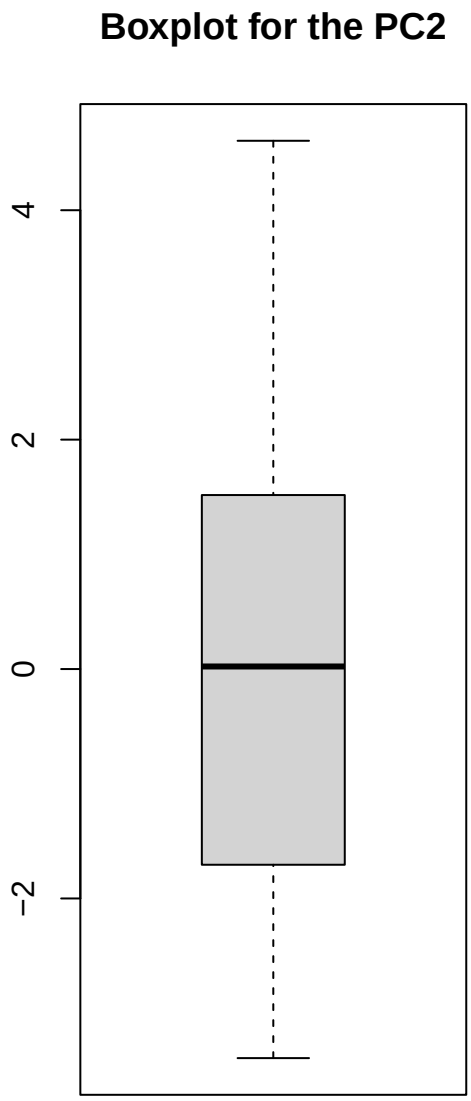


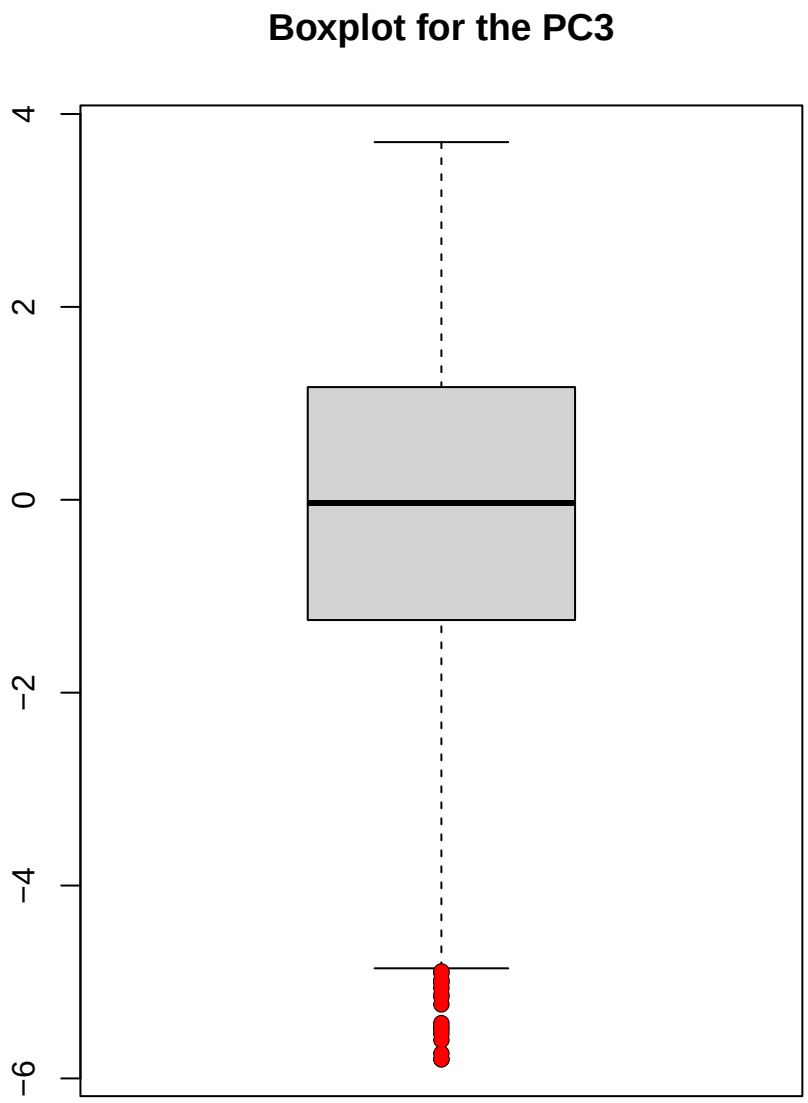
The scatterplots above shows that data tends to form a sort of clusters with respect to the classifying *"digit"* variable, except for the digits *"5"*, *"2"* and *"8"* which seems to be more scattered along the plane and not very concentrated in a cluster.
We can also notice the scatterplots (PC_3, PC_1) and (PC_3, PC_2) have such a *"tail"* of observations which deviates from the cloud of points in which most of the observations are concentrated and we'll see later that these points have a particular meaning.
Finally, one could have supposed that maybe (PC_2, PC_1) corresponds to a bivariate Gaussian couple of variables since their scatterplot has a quite-elliptical shape but we concluded before (looking at their Q-Q plots) that these two components can't be considered as normally distributed. Also the cloud of points of the scatterplot (PC_3, PC_1) shows a similar behaviour and this confirms the doubts about the possible normality of the 3^{rd} Principal Component we detected at point 3.2).

3.4) Comment about outliers with respect to the first three principal components:

In that case we'll investigate only univariate outliers of PC_1, PC_2 and PC_3 through the boxplot representation since we can't use neither the ellipses in the scatterplots of the components to search for possible bivariate outliers neither the high quantiles of the χ^2_3 to look for possible multivariate outliers since in point 3.2) we concluded that PC_1, PC_2 and PC_3 are not multivariate Gaussian and so these techniques are not effective. However, we can still plot both the scatterplots and the Squared Mahalanobis distances in order to show the position of the observations that will be detected as univariate outliers. We start plotting the boxplots of the Principal Components:

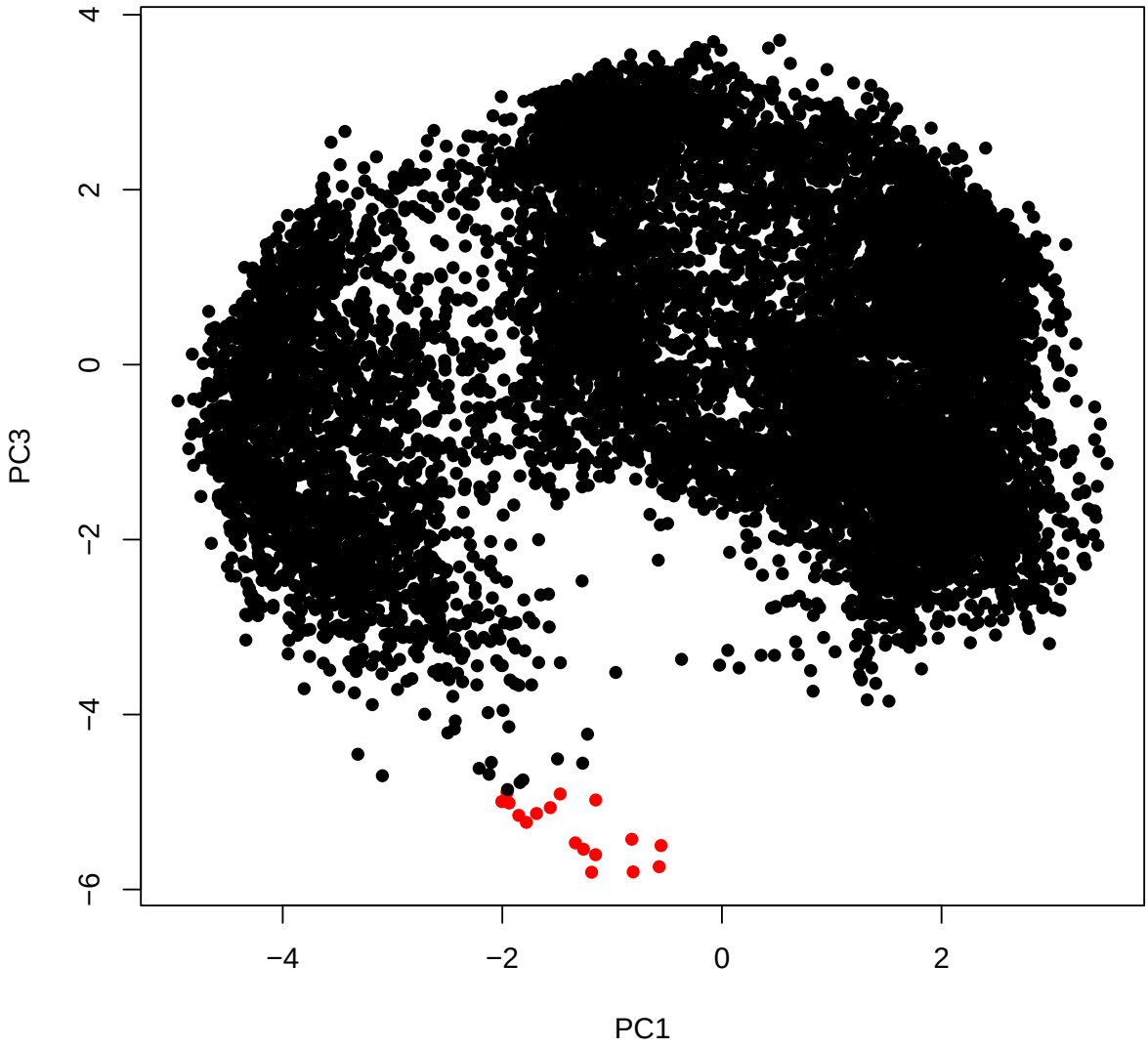


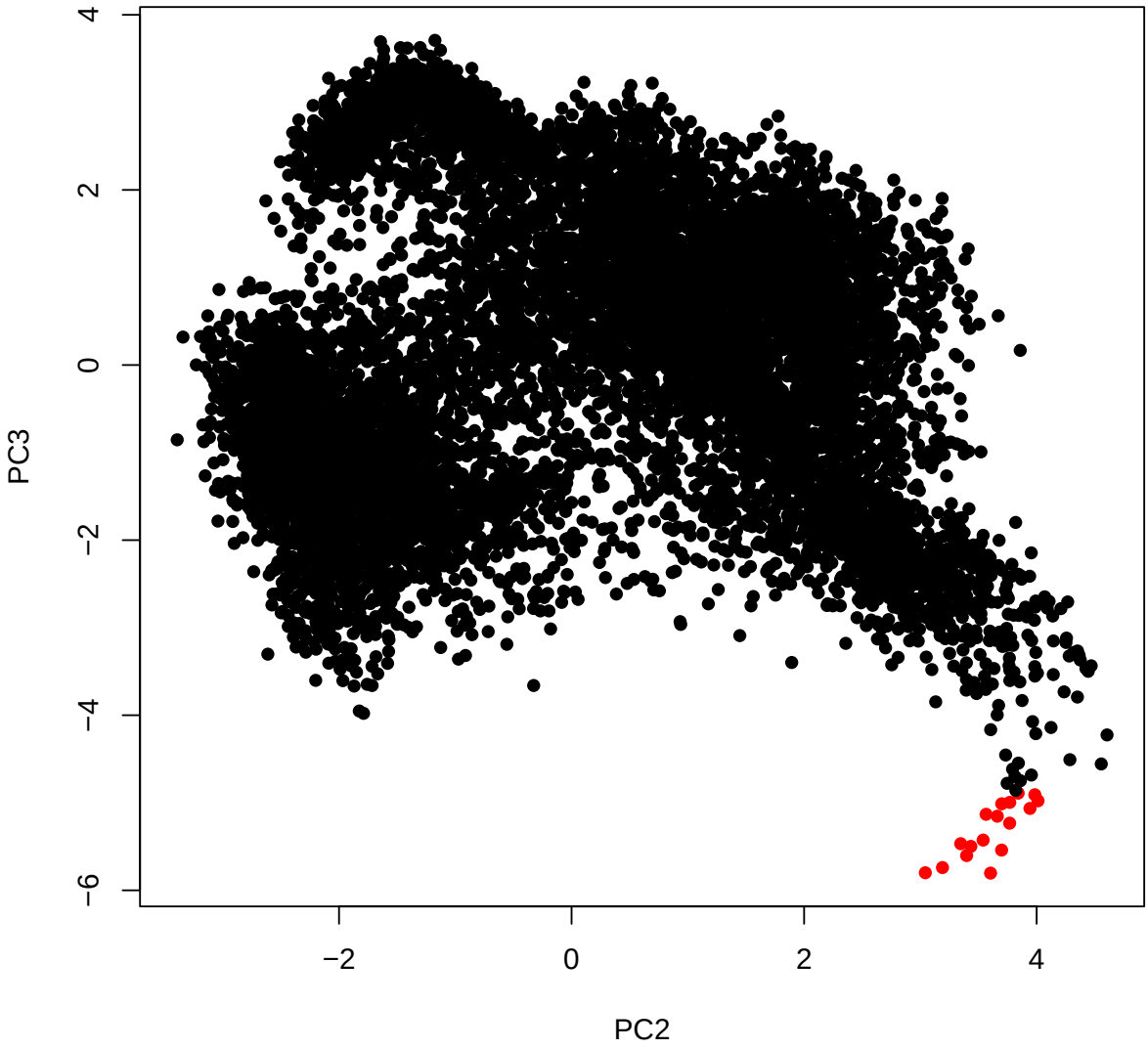




The results shows that there isn't any univariate outlier for PC_1 and PC_2 while PC_3 shows a total of 17 observations down the 1^{st} whisker which could be considered as potential outliers due to their low value.

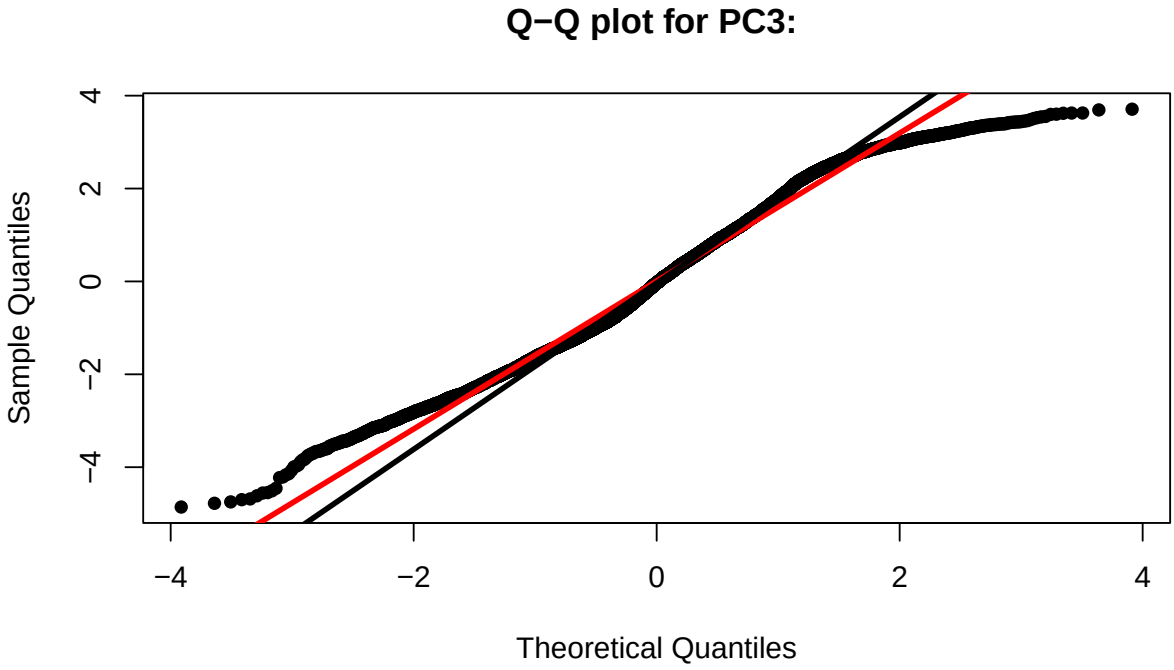
Now, we color in red these observations in the two scatterplots which concerns the 3^{rd} Principal Components to see if they hire a strange position:



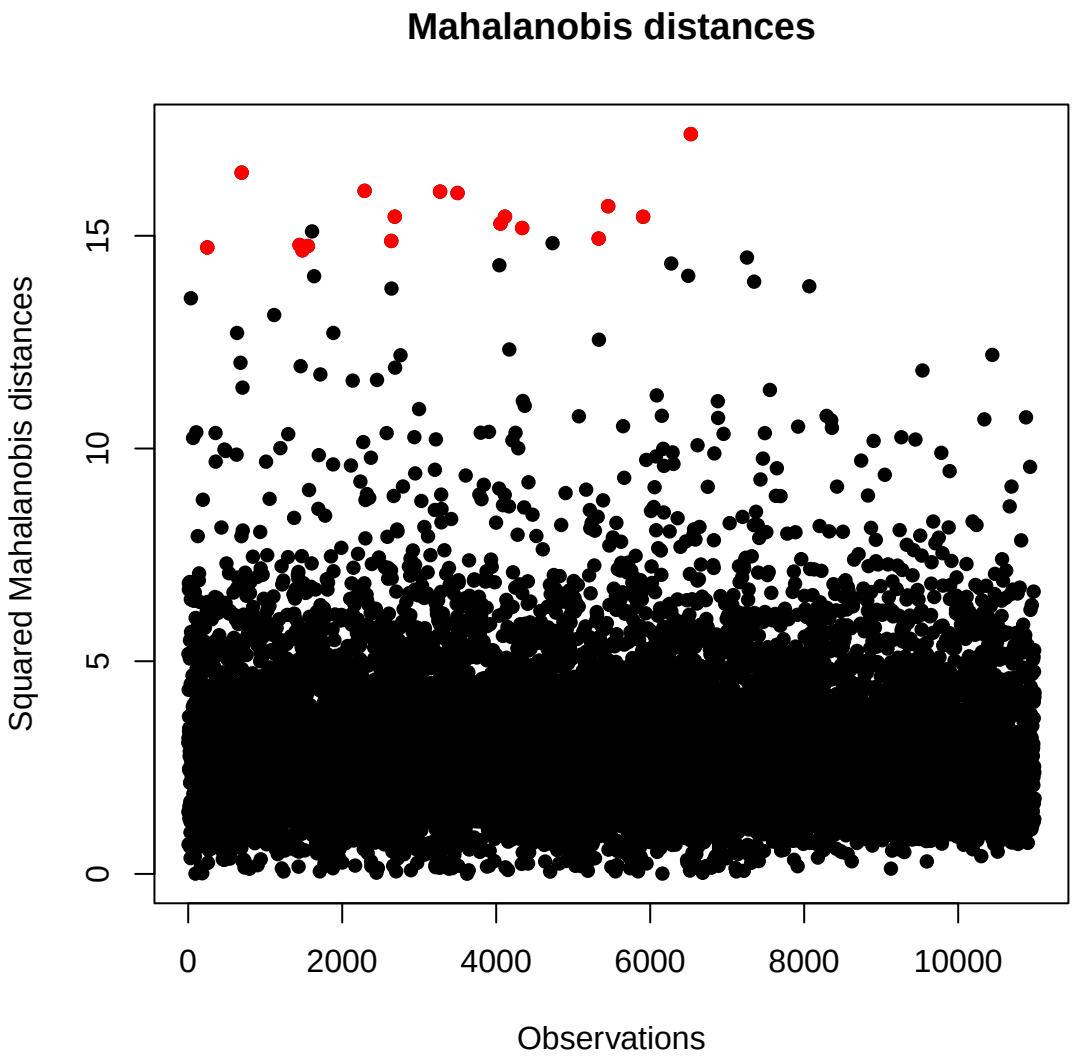


We see that the detected outliers are found in the *"tails"* of the two scatterplots (as we anticipated before) and so they were graphically detectable since they deviates from the main cloud of points. We could also observe that removing these observations, these scatterplots hire a more elliptical shape, so this suggest that those outliers could be removed in order to make the 3^{rd} Principal Component closer to a Gaussian.

But if we plot again the Q-Q plot of PC_3 removing the 17 outliers we see that the situation doesn't change significantly for the univariate normality of this components:



Finally we plot the observed Squared Mahalanobis distances of the 3 Principal Components underlining in red the detected outliers:



Even though we can't use the quantiles of the χ^2_3 to determine if the 17 observations are also multivariate outliers, cause (PC_1, PC_2, PC_3) are not multivariate Gaussian, we observe that these observations are related to really high values of the Mahalanobis distance.

The last consideration we can make about the detected univariate outliers is that all of them are related to observations that match with the digit *"9"*. This is probably related to some particular features that this digit has.