Exercise 1 - MSA

Matteo Bandiera - Samuele Fonio - Luca Macis

Exercise 1

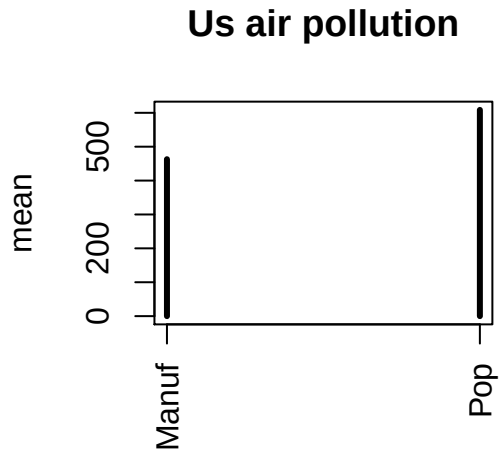
Point 1.1

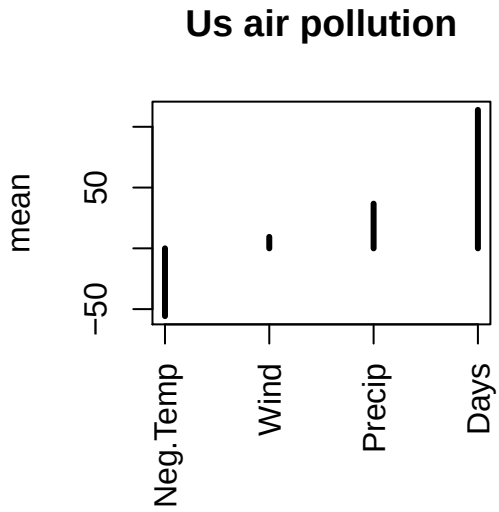
Compute the sample mean and the correlation matrix and comment on the correlations.

Let's start by commenting on the sample mean of the variables. Here we find a summary of the data

Neg.Temp	Manuf	Pop	Wind	Precip	Days
Min. :-75.50	Min. : 35.0	Min. : 71.0	Min. : 6.000	Min. : 7.05	Min. : 36.0
1st Qu.:-59.30	1st Qu.: 181.0	1st Qu.: 299.0	1st Qu.: 8.700	1st Qu.:30.96	1st Qu.:103.0
Median :-54.60	Median : 347.0	Median : 515.0	Median : 9.300	Median :38.74	Median :115.0
Mean :-55.76	Mean : 463.1	Mean : 608.6	Mean : 9.444	Mean :36.77	Mean :113.9
3rd Qu.:-50.60	3rd Qu.: 462.0	3rd Qu.: 717.0	3rd Qu.:10.600	3rd Qu.:43.11	3rd Qu.:128.0
Max. :-43.50	Max. :3344.0	Max. :3369.0	Max. :12.700	Max. :59.80	Max. :166.0

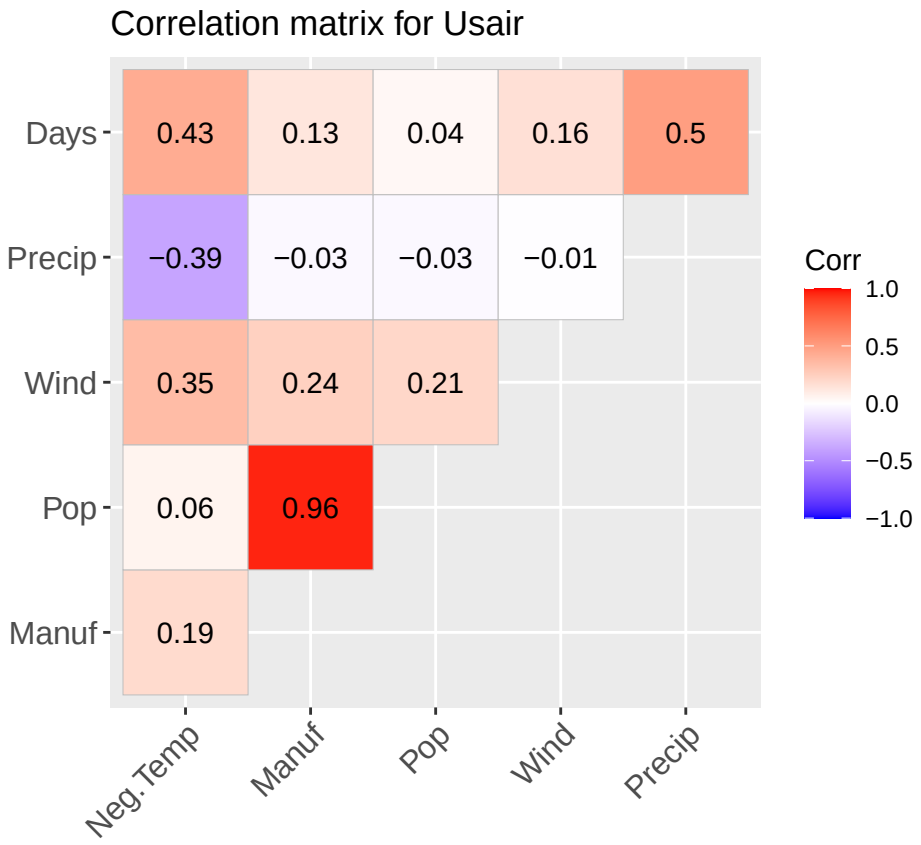
And now let's see the plots of the sample mean:





For a better view we decided to distinguish between the two kind of data that we are treating. In fact as we can see Manuf and Pop are something related to the humans, and have bigger scales. We will refer to them as "human ecology". While the other variables deal with the nature. We will refer to them as "climate".
As we can see the scales are so different, and we have to remember that are all different kind of measures. The only note to take into account is the negative mean of the Neg. Temp, that was predictable since the definition of this variable.
In order to present the correlation matrix in a suitable way we used a specific package, that presented us this

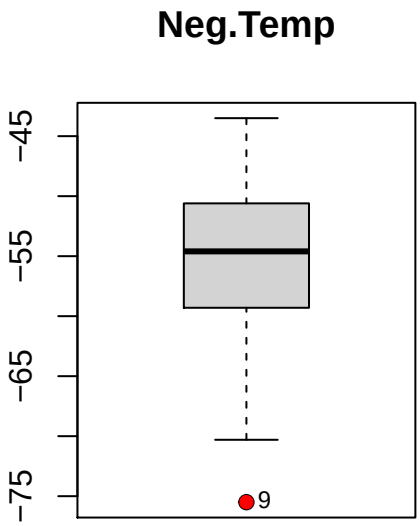
output.

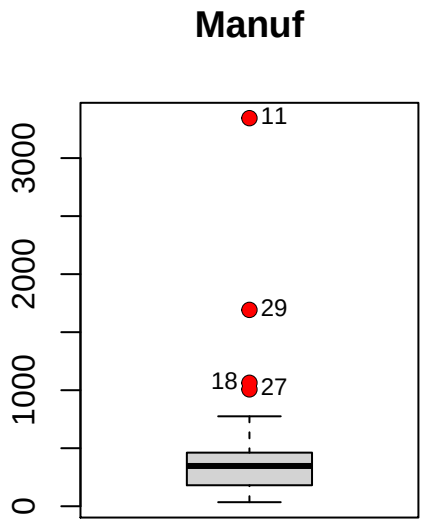


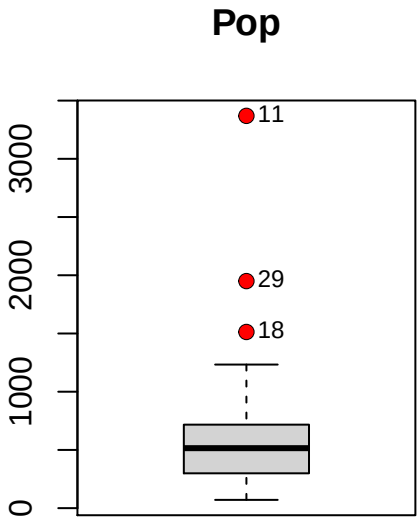
The only strong correlation between the variables is present between Pop and Manuf. This was predictable for what they represent. The second stronger correlation in between Days and precipitations, which was also predictable always by their definitions. It was expected a weak correlation between the human ecology and climate, because the presence of population does not effect directly the climate behavior. It's also interesting noting the negative correlation of the precipitation with respect to the negative temperature that is in any case justified by the definition of the variables.

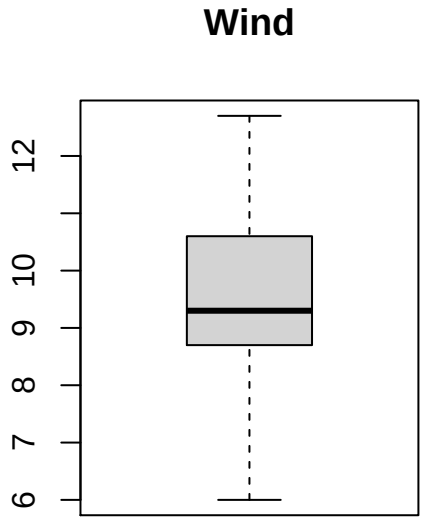
Point 1.2

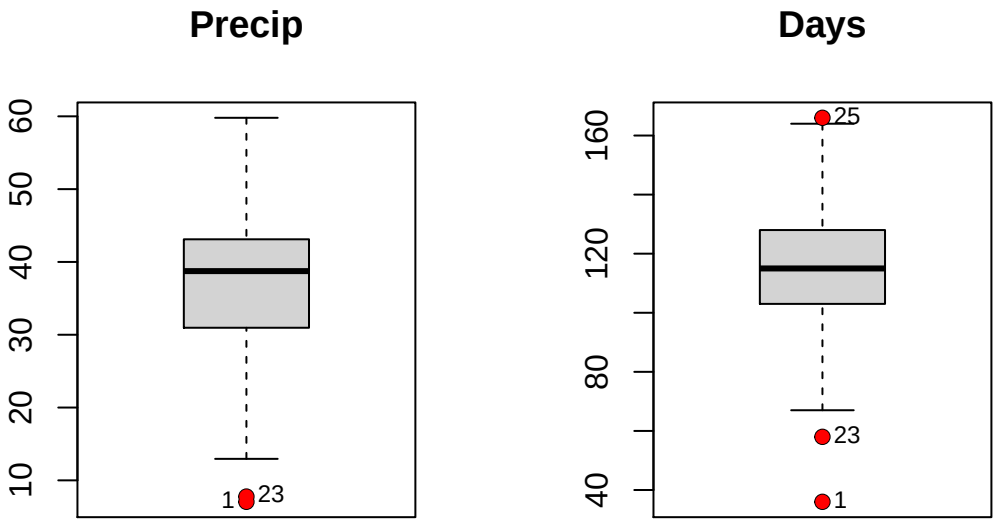
In point 1.2 we are going to study the outliers starting from the boxplot.







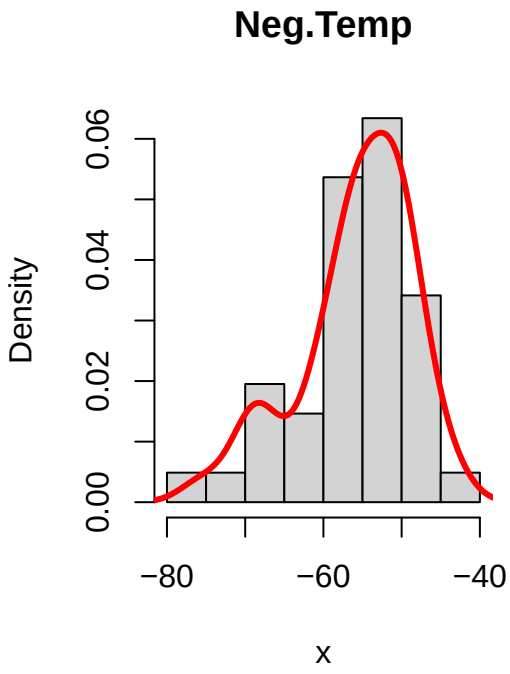


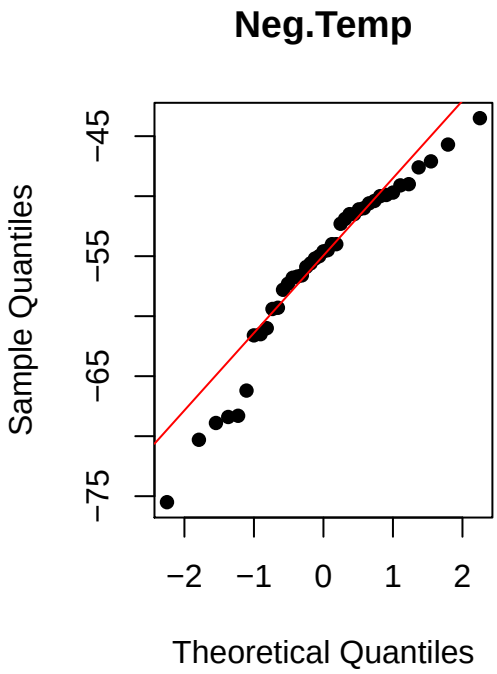


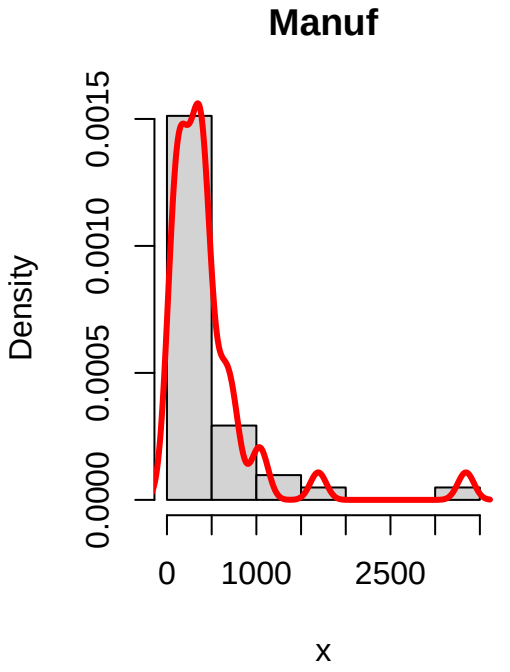
These are the outliers detected from the boxplots. We stored them in a list.
In this way we will be able to compare them in future. For the moment let's just comment a little bit on this output.
The manuf and pop variables have almost the same outliers, which is not surprising since they are strongly correlated. The wind variable has no outliers, while we can comment about the observation 23 as an outlier for days and precip. We can see from the correlation matrix that the correlation between these two variables is the second strongest among all, so we could expect a common outlier.
So for the moment we have to recall that we have the indexes of the observations detected in a list and that these will be useful in the future analysis. But for doing further analysis of the outliers, we have to check the normality of our variables.

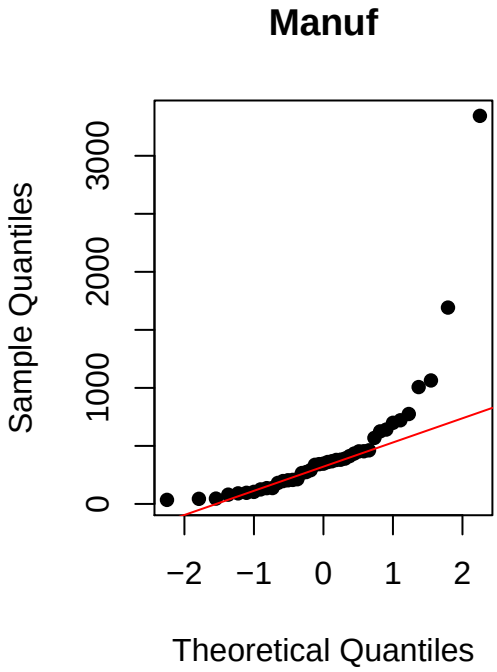
Point 1.3

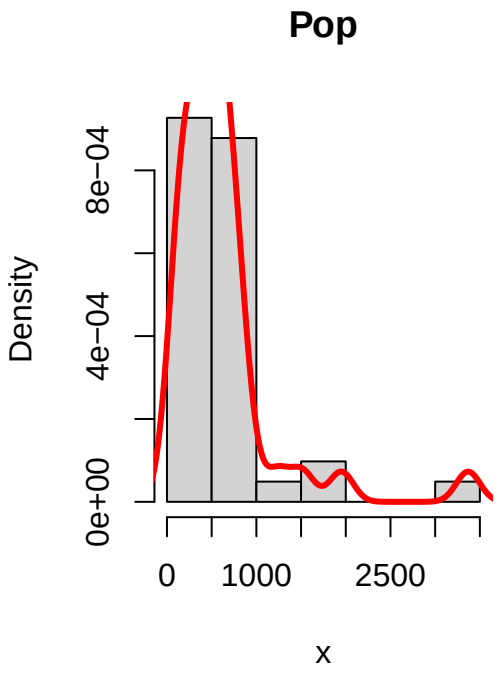
In this point we have to discuss the normality. We will use two important tools: histograms and qqplots.

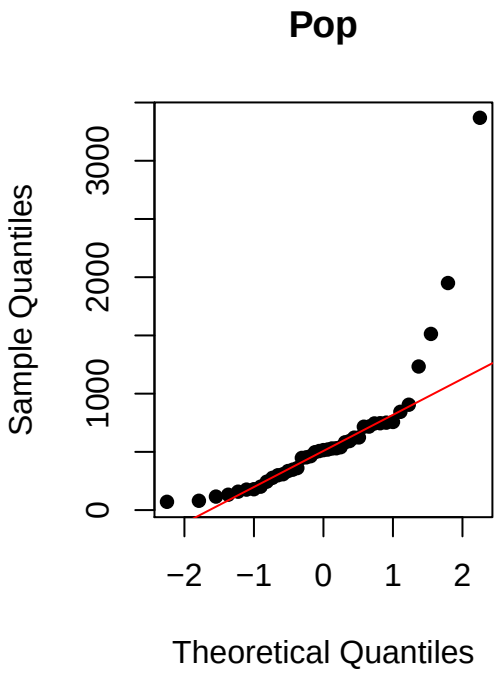


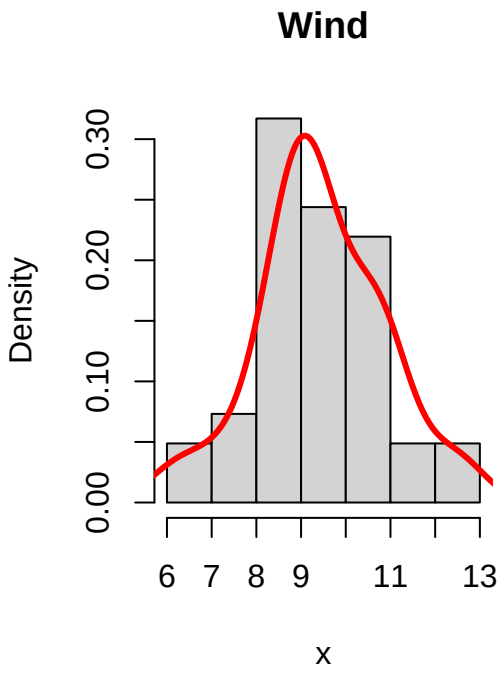


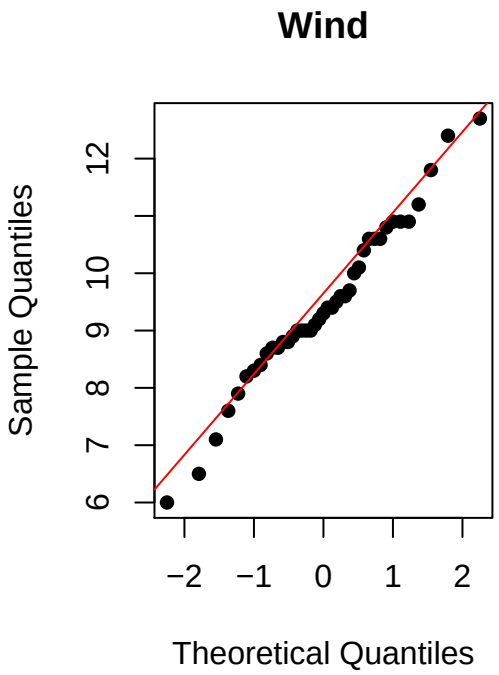


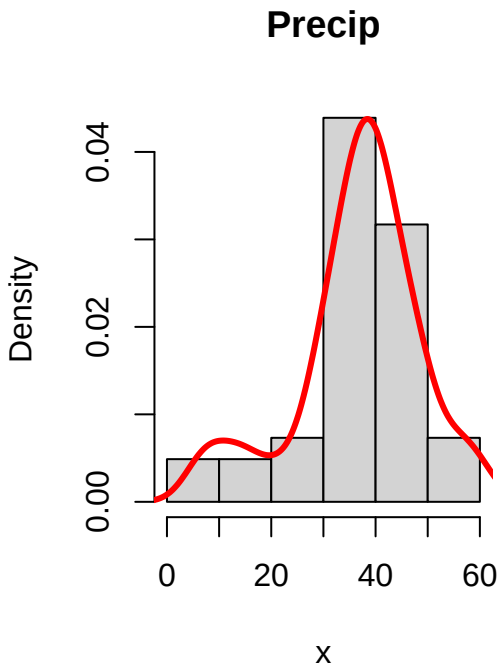


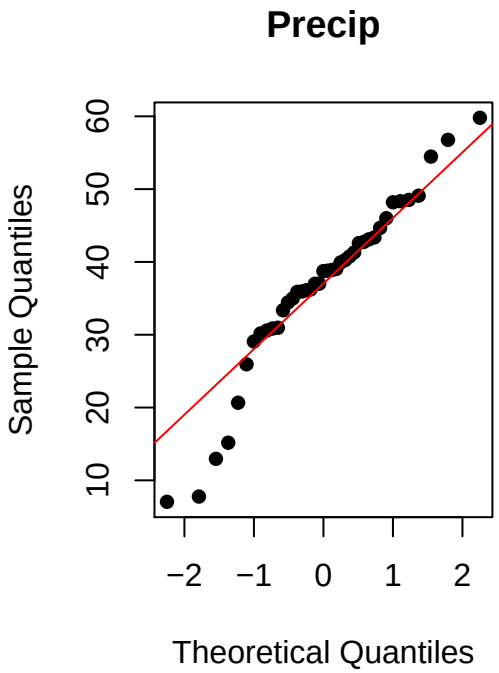

Pop
Density
0e+00
4e−04
8e−04
0
1000
2500
x

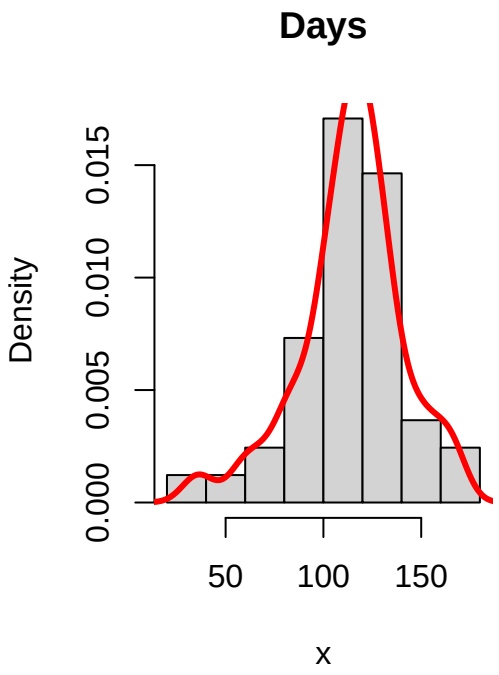

Pop
Sample Quantiles
0
1000
2000
3000
−2
−1
0
1
2
Theoretical Quantiles

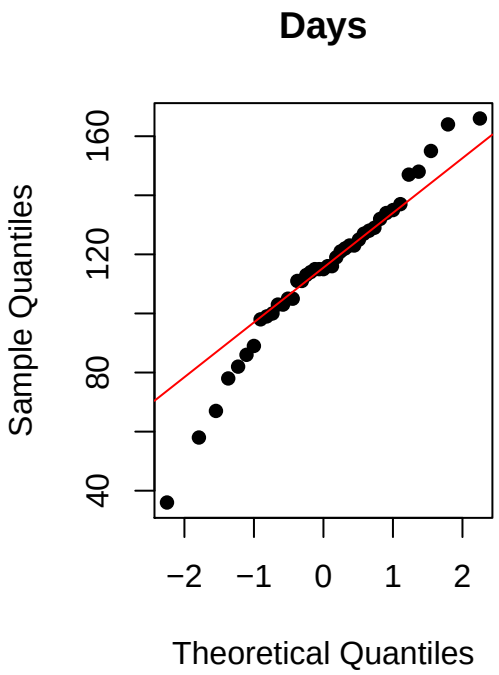

Wind
Density
0.00
0.10
0.20
0.30
6
7
8
9
11
13
x


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

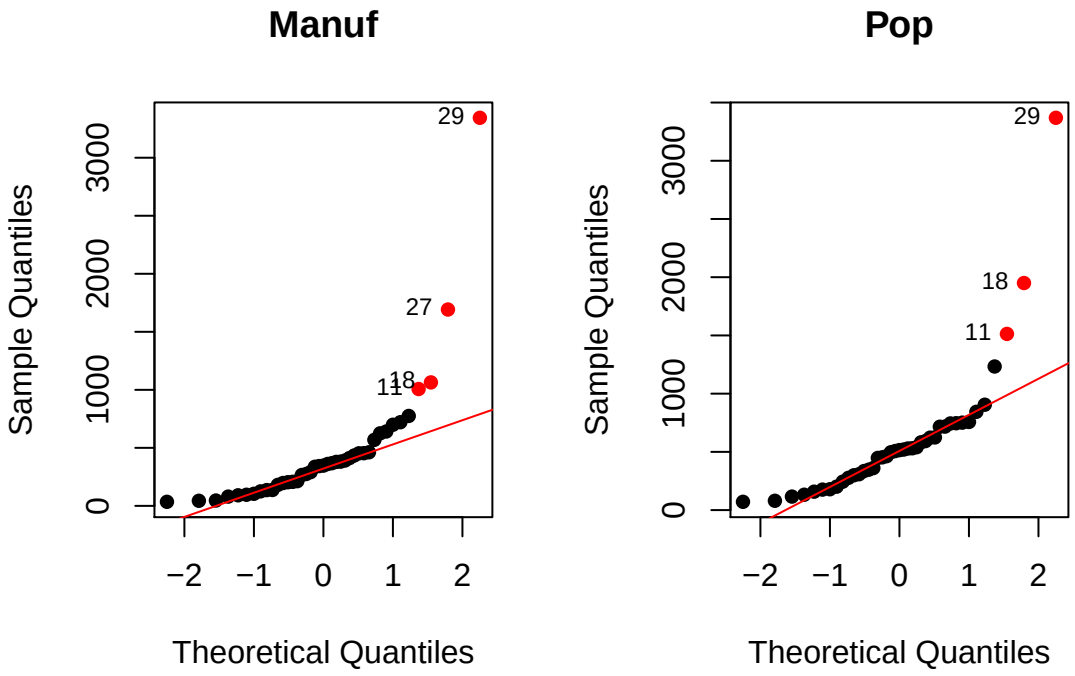




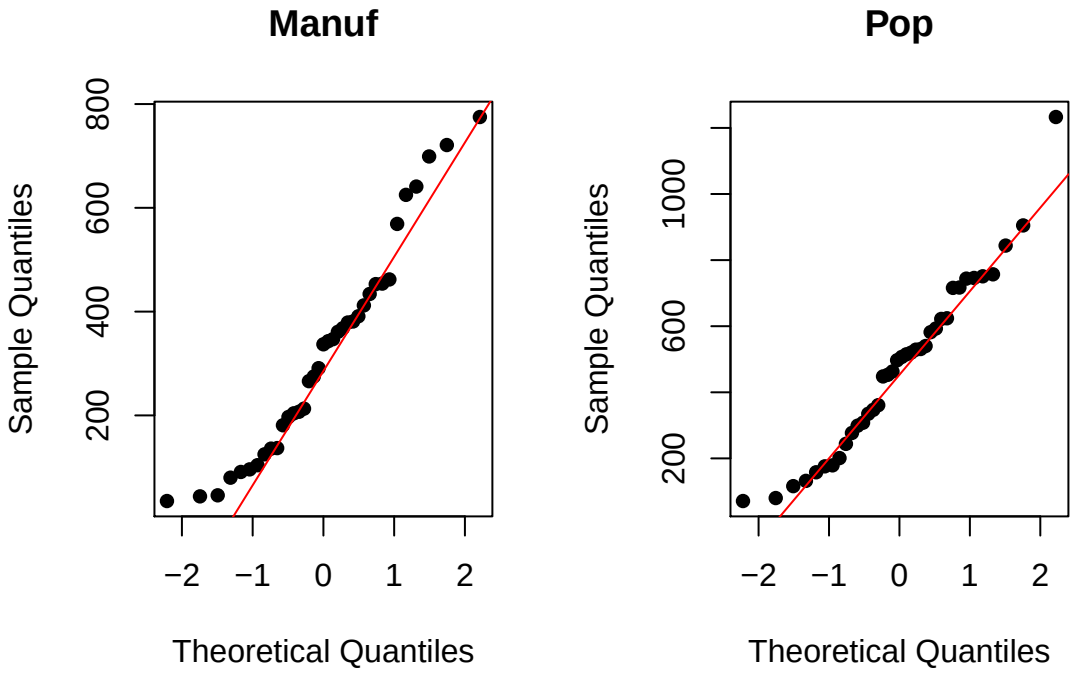




As we can see there seems to be no problem of normality for the "natural" variables: the histogram shows a normality density pattern and the quantiles seem to not behave in a strange way. This is not true for the human ecology. Let's study in particular the qqplots. We can color the outliers found in point 1.2 and add a text to recognize them.

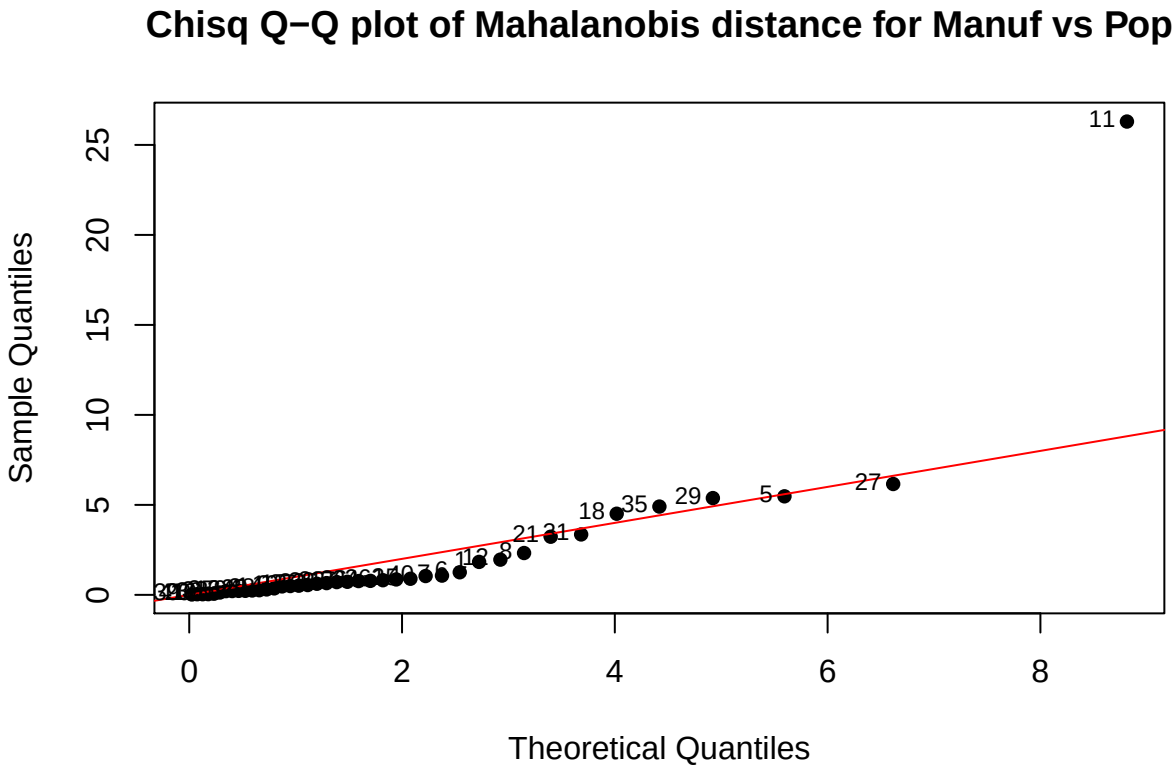


As we can see the highlighted observations are the outliers detected in point 1.2. So the question is: if they don't behave as the sample, may them affect also the normality? So in order to answer we can try to cut them off and comment:

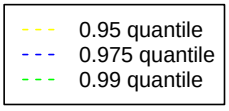


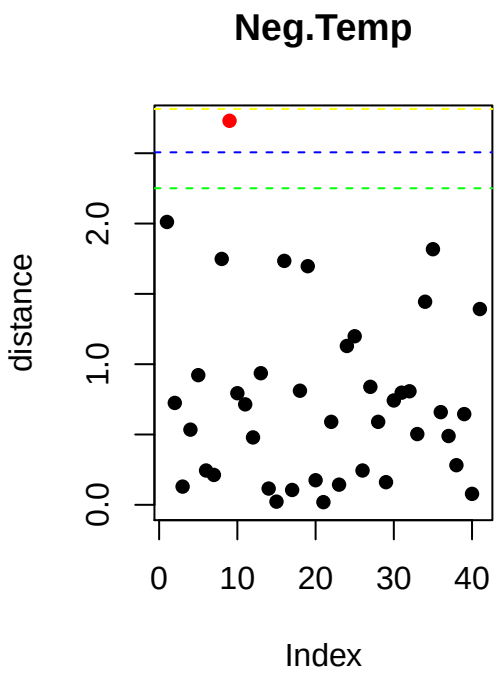
Without those observation the sample shows a normal behavior. This means that the outliers found in

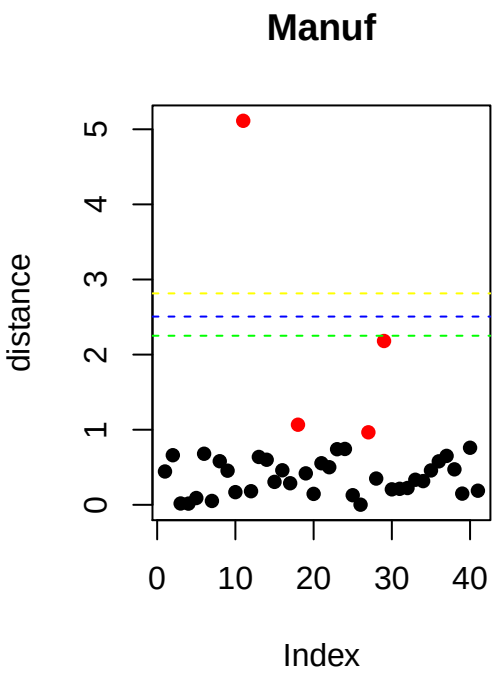
point 1.2 affect also the normality, hiding it. Anyway this problem could be solved doing a bivariate normal analysis.

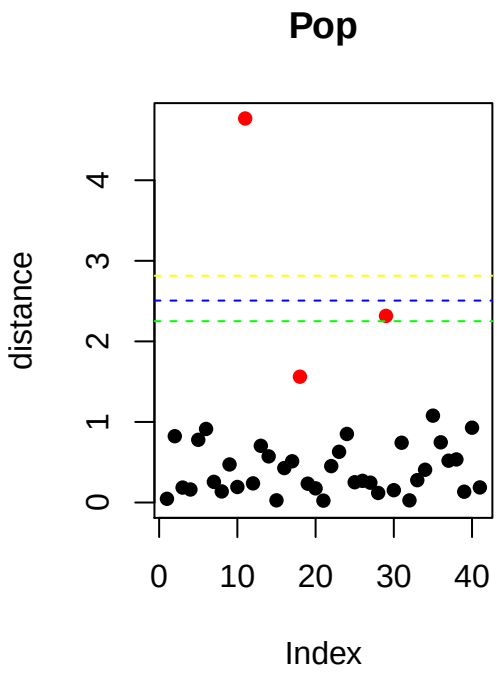


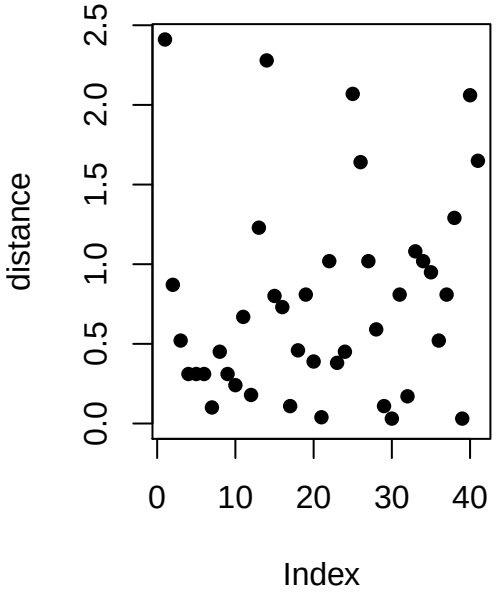
This shows us that the bivariate behavior of (Manuf,Pop) agrees with a normal behavior of the single variables. So we can conclude that they cannot be considered normal in the univariate case, even if the observations that make them non-normal are outliers, but this does not affect the multivariate analysis.
Furthermore, observation 11 will be interesting to study.
Anyway, we can do another step in order to detect the outliers. In fact we can see which observations alter the normality with the following plots.

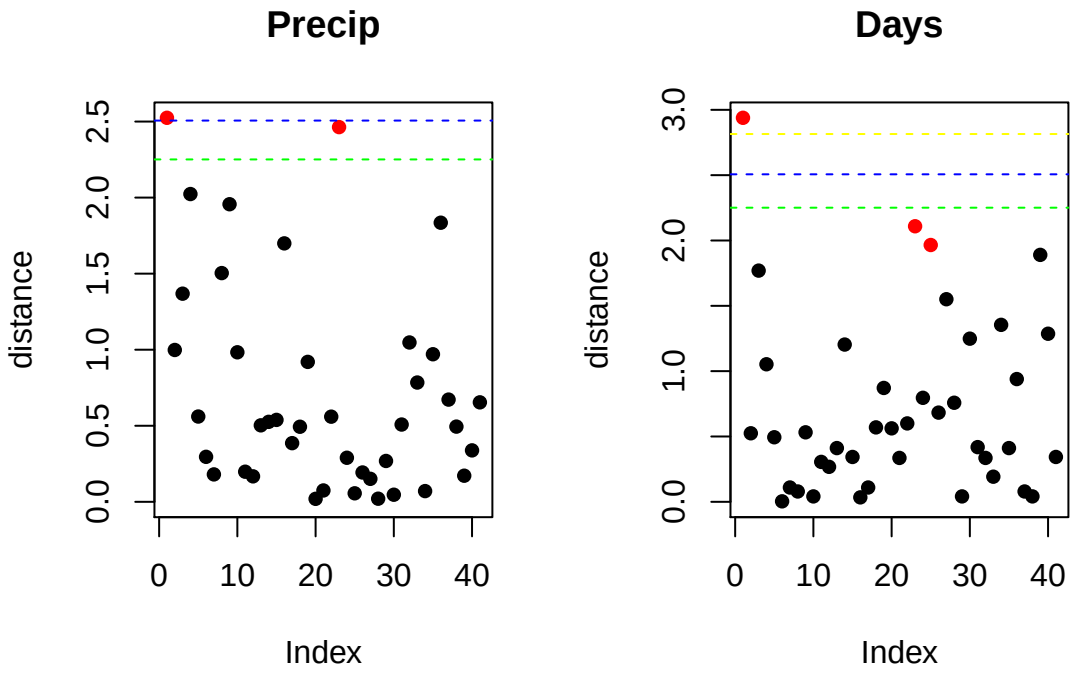



Neg.Temp
distance
0.0
1.0
2.0
0
10
20
30
40
Index


Manuf
distance
0
1
2
3
4
5
0
10
20
30
40
Index


Pop
distance
0
1
2
3
4
0
10
20
30
40
Index

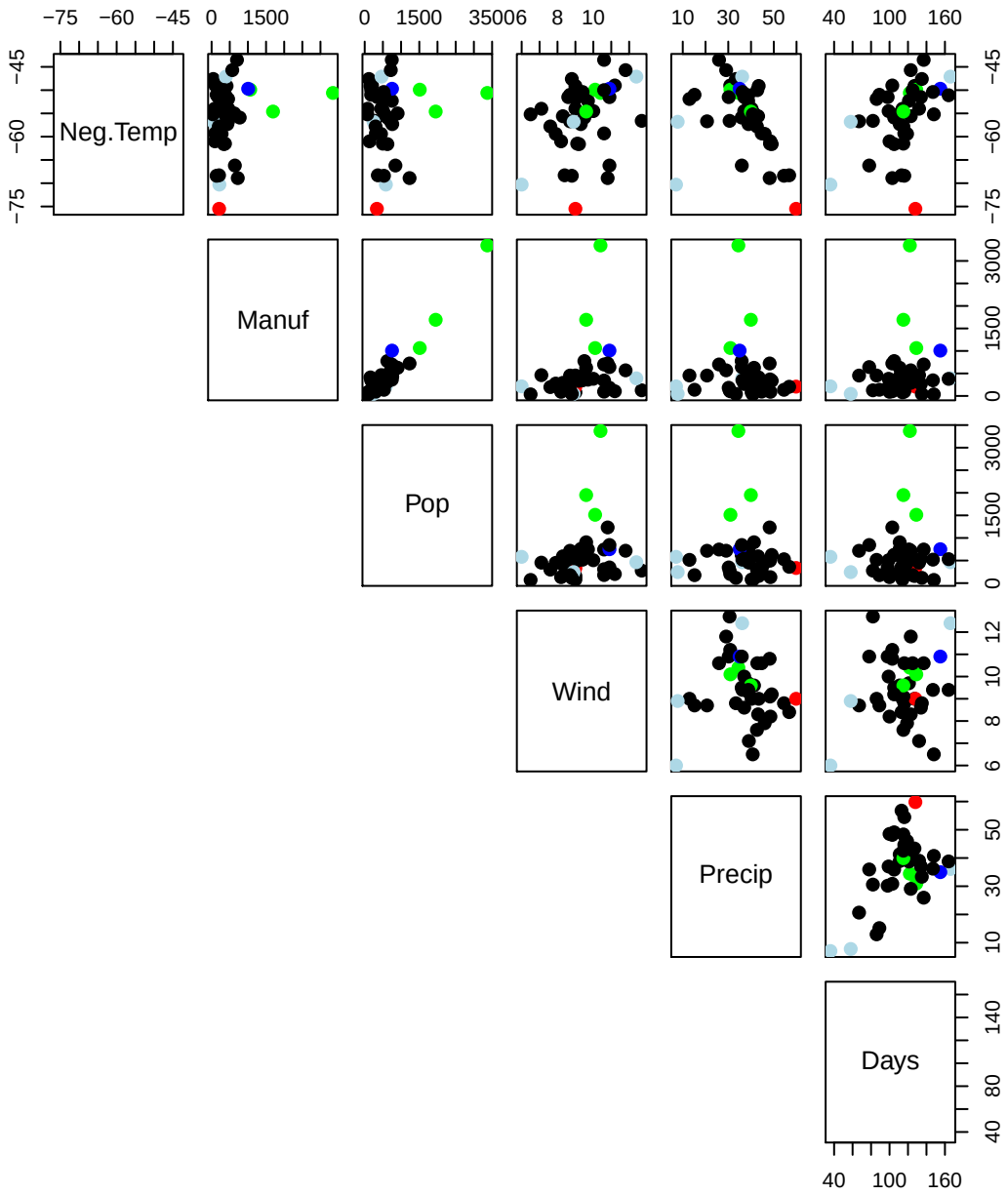

Wind
distance
0.0
0.5
1.0
1.5
2.0
2.5
0
10
20
30
40
Index



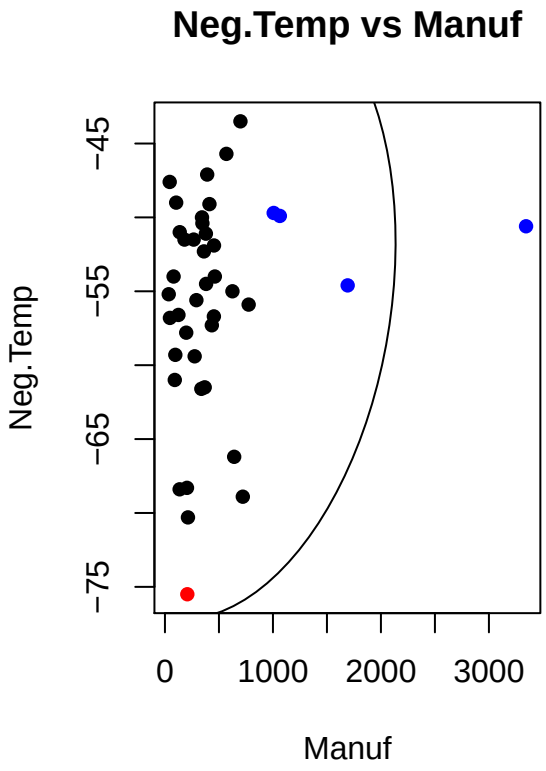
We made these plot standardizing the sample, and then comparing the distance from the mean with the normal quantiles of order 0.95 (green), 0.975 (blue) and 0.99 (yellow) weighted with the sample dimension. We highlighted in red the outliers detected in point 1.2.
First of all we can notice that some red points are over the 0.99 percentile, which means that they are far from the mean in a sample and theoretical way. In general we can say that the behavior is not strange, since only the outliers already detected are, sometimes, over the 0.99 (or less) quantiles. This shows us that our variables do not disagree with a normal pattern. For the Human ecology we can notice that our outliers have a great weight also in this analysis.
The wind behavior seems to be the most standard in this sense: it did not have any outlier and the quantiles are not strange.

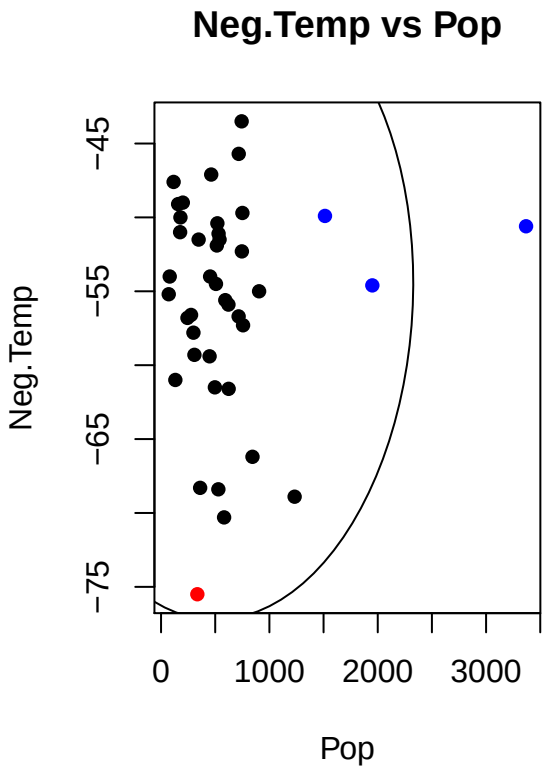
Point 1.4

We can use scatterplots to detect the outlier found at point 1.2. In particular we can study case by case the following plots, in which we colored with a scale of colors the outliers of point 1.2:

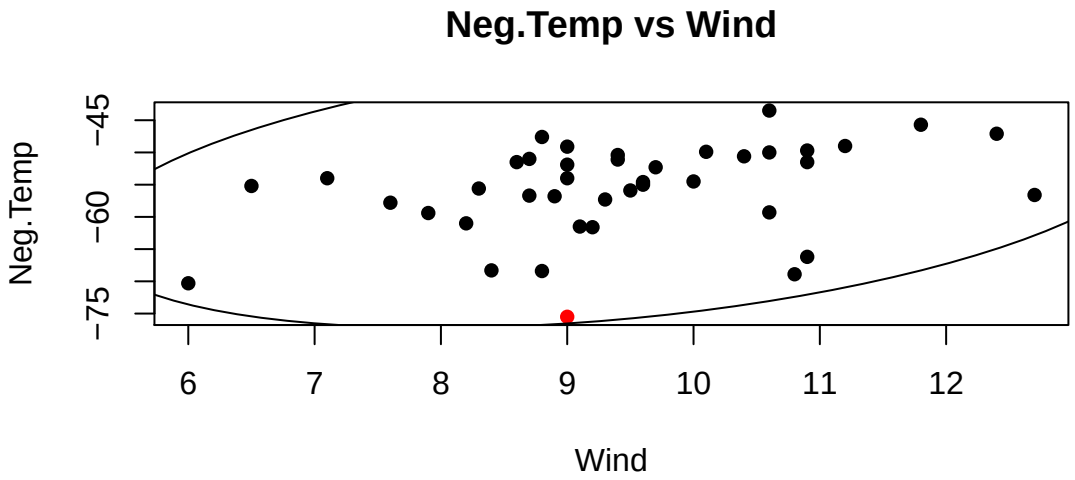


This image is not so clear, but can give us an idea of what is happening: we are matching the variables two by two and studying their relation. We'll see only the most important and emblematic cases. In the following we are coloring the outliers of point 1.2, in particular red is referred to the second variable and blue to the first. Furthermore we add the ellipse of level 0.95 (weighted with the sample size) just to see what the points do.

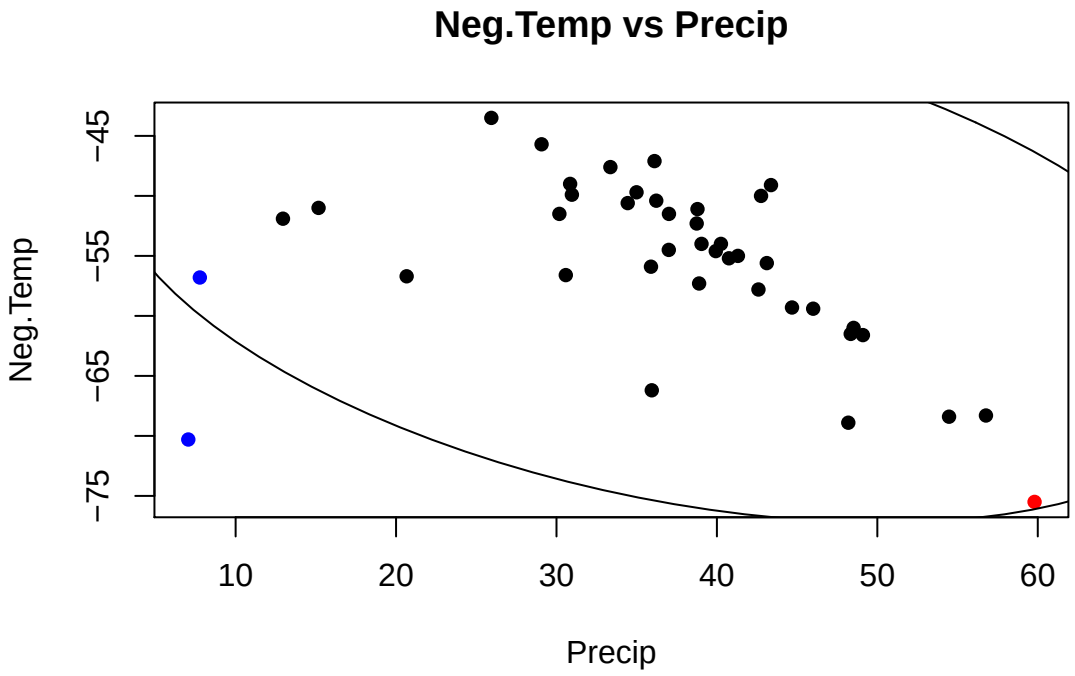




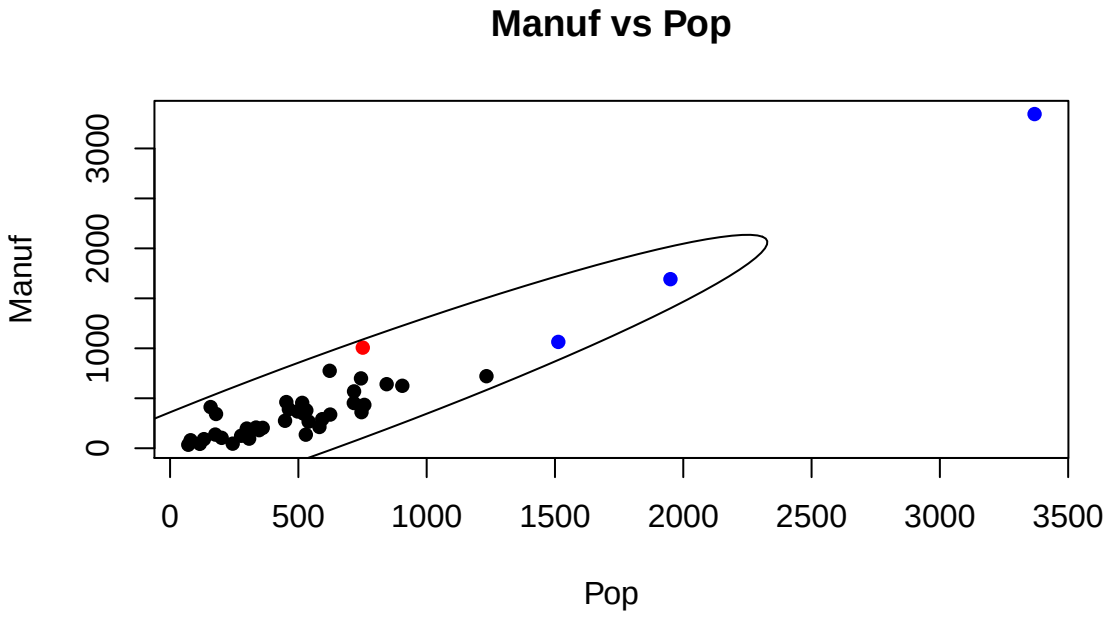
We decided to put this graphs because they are emblematic: the outliers are clearly out of the range if we consider just the single axes,and as we can see they are all inside the ellipse of level 0.95. We'll check upper levels later on. For the moment it's enough to note that the univariate outliers are detectable from these plots. Furthermore recall that Manuf and Pop were strongly correlated, so we expect almost the same behavior when we compare them to any variable, as in this case. In fact these two graphs seem to be very similar. However from this plot is also possible to see the weak correlation between Neg. Temp and bot Manuf and Pop: the points does not concentrate around any line.



For this plot we had just one observation of point 1.2. Since also from previous analysis the behavior seem regular, we could expect that all the points fall inside the ellipse. As we can see there isn't a strong line shape, as there is not a strong correlation, but the points seem to be more ordered than the previous two graphs.



This plot is interesting since we can recall that the correlation between these two variables was negative. As a consequence the points are quite concentrated around a line with negative slope. The outliers found in point 1.2 are always present.



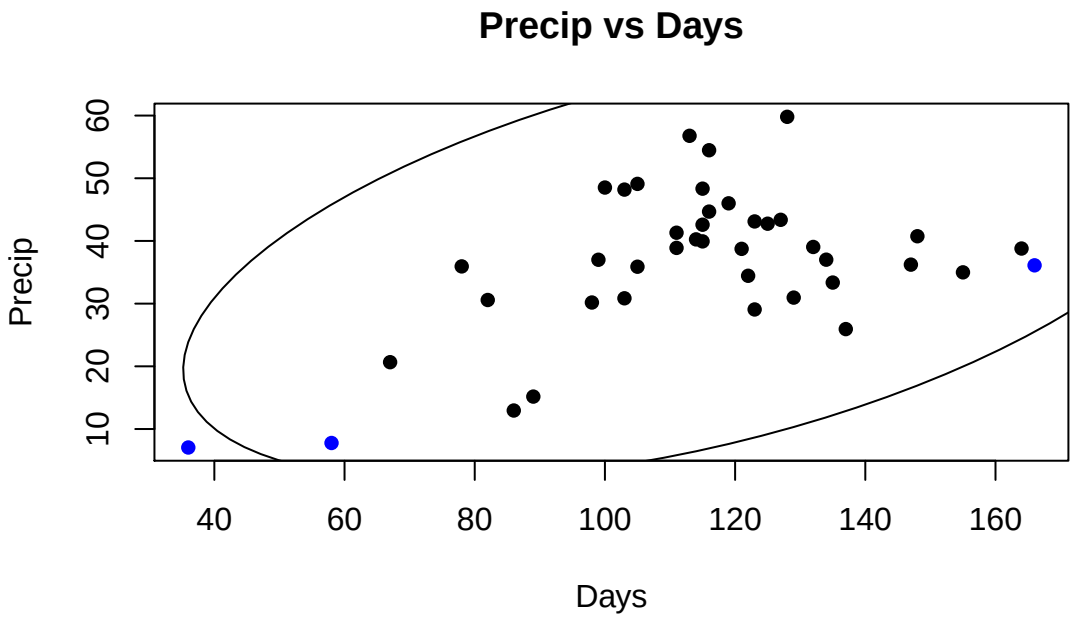
This plot is interesting since it's clear that the correlation between these two variables is strong. As we can see the outliers (those in common are in blue) are far from the other observations, as they are univariate outliers. Probably that point so far from the others is observation 11 that we saw in the QQplot of the Mahalanobis distance for these two variables. We can check it just by watching the data:

```
usair1[11,]
```

```
##         Neg.Temp Manuf  Pop Wind Precip Days
## Chicago    -50.6  3344 3369 10.4  34.44  122
```

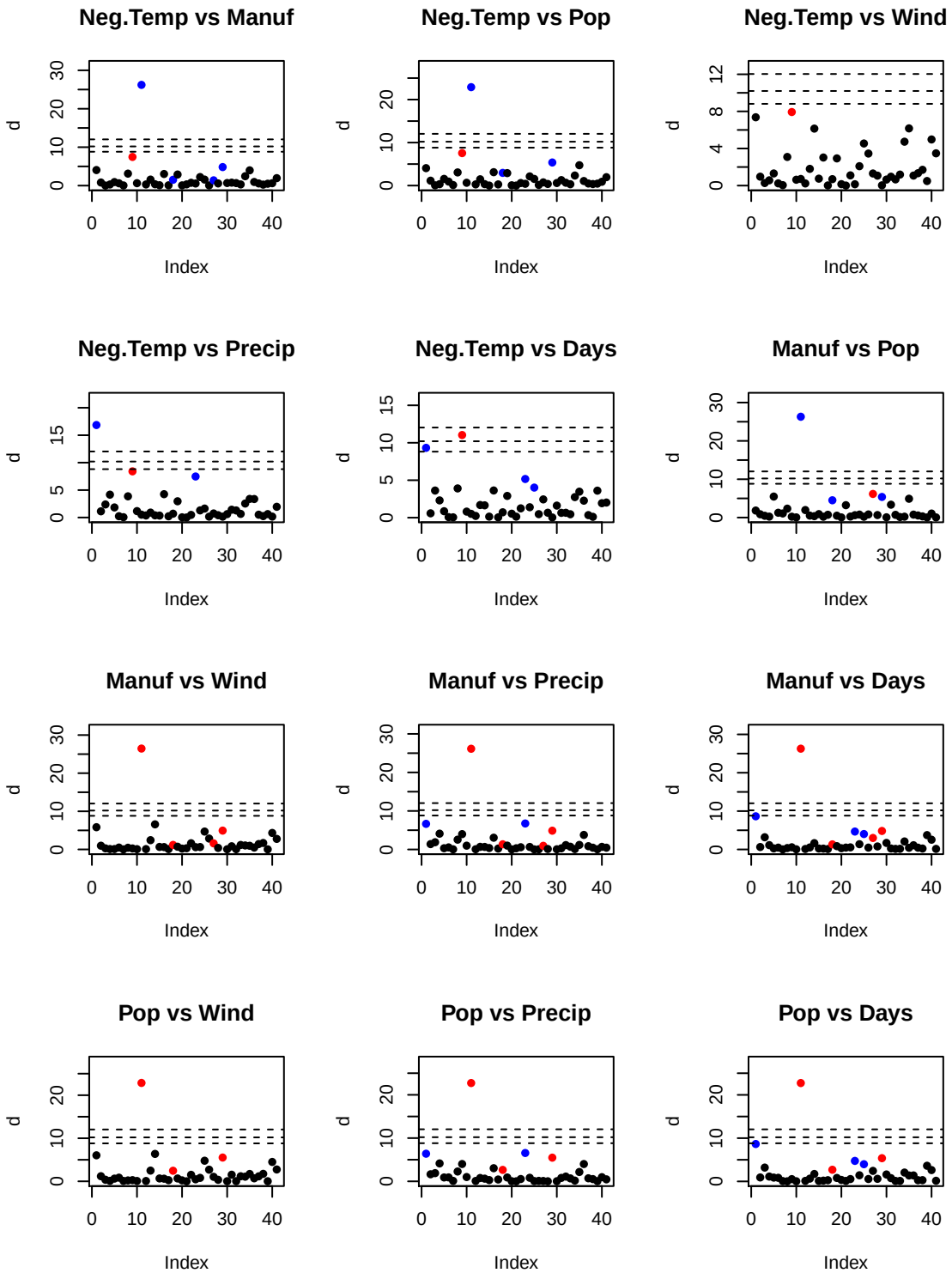
As we expected it's observation 11, as we can see from the values of Manuf and Pop.

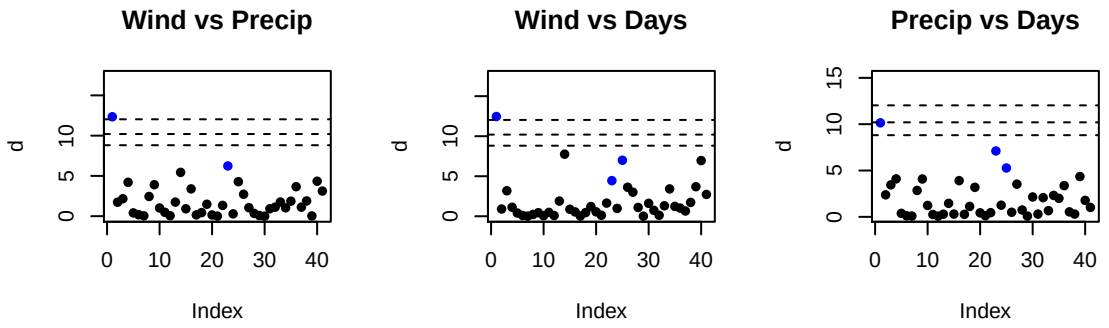
Note: All the comparison between a variable and Manuf/Pop have the same pattern that we have already seen, so we will neglect further comparison with these two variables.



We put this last plot just to see a better line shape since the correlation between Precip and Days is the second biggest.
So in general it's easy to detect the previous outliers from these plots, it's enough to restrict to one axis and see which points are far from the other observations. Sometimes this is not clear because there can be points colored that are very near other observations but we found them anyway in our point 1.2. This means that in a bivariate sense their behavior is not strange.

Now we can also study the Mahalanobis distance (for bivariate case) to detect possible hidden outliers. We already saw something drawing the ellipses in the previous plots, but we can do more looking at the Mahalanobis distance. However we have to remember that the Mahalanobis distance is standardized with respect to the standard deviation.


Neg.Temp vs Manuf
Neg.Temp vs Pop
Neg.Temp vs Wind
Neg.Temp vs Precip
Neg.Temp vs Days
Manuf vs Pop
Manuf vs Wind
Manuf vs Precip
Manuf vs Days
Pop vs Wind
Pop vs Precip
Pop vs Days
d
Index

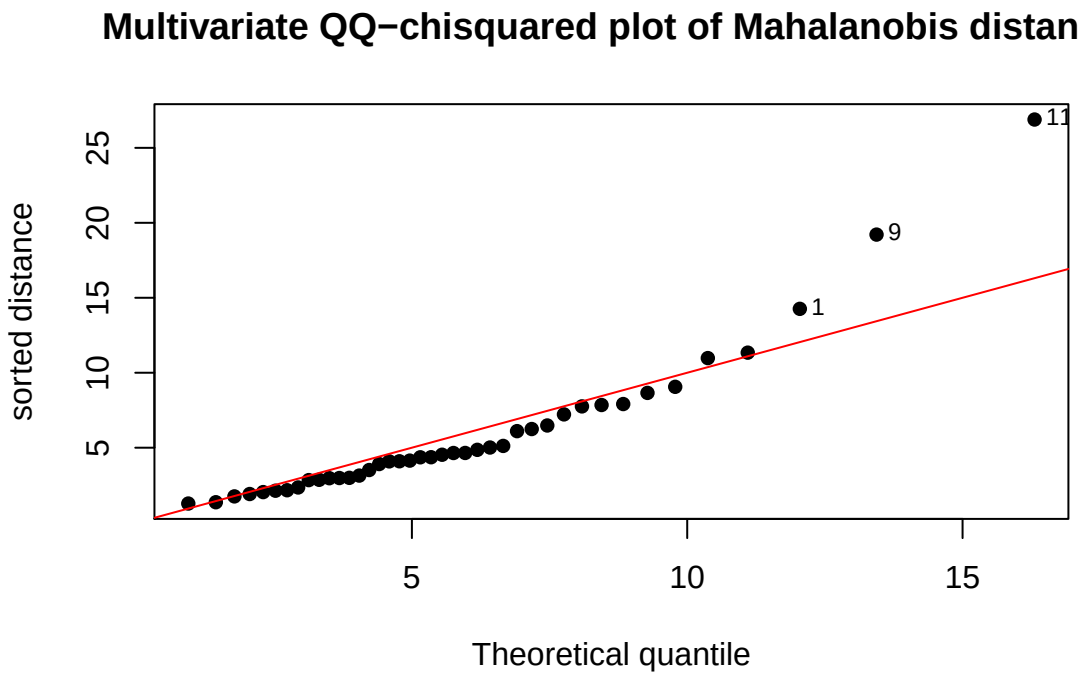


As we can see there are no new outliers to detect. This means that standardizing the data they don't represent any strange behavior. Indeed after standardization they are all below level 0.95.

To conclude: points detected in point 1.2 are detectable from these scatterplots if we focus on each axis singularly. If we want to study the bivariate case, we can do it but, as we saw, there is no significant result to highlight.

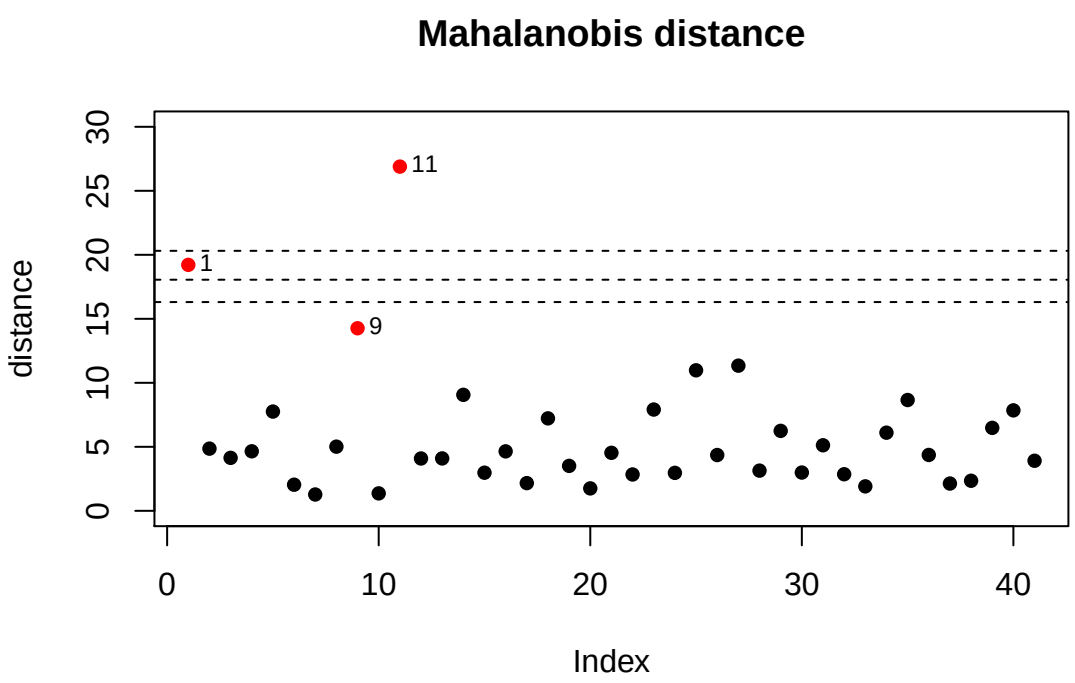
Point 1.5 and 1.6

Observe that we already did a sort of multivariate normality analysis, but in a bivariate sense. And everything seem to work quite well.We now move on the multivariate (p-dimensional) case:



As we can see the behavior is quite linear and so we can assume the singular normality behavior. In fact all the points are near the line bisecting the first quadrant, except for those three observations that are our new possible multivariate outliers. This plot confirm also the good multivariate behavior of the Mahalanobis distance that, theoretically should behave as a chisquared distribution.
It's quite interesting noting that there is also observation 9 among the new possible outliers. Observation 9 was an outlier just for Neg. Temp, but in a multivariate sense its importance is big.

We can also compare the Mahalanobis distance for detecting them:



As we can see observation 11 is a real outlier, while with respect to the 0.95 quantile observation 9 is not.
So from these plots we can conclude also our analysis about multivariate outliers. In fact as we know a scatter plot of six variables is not possible to perform, but from these last two plots we can see that there are observations number 1,9 and 11 that don't behave as wanted.
In particular, watching at the Mahalanobis distance plot, we can conclude that observation number 9 is not a very multivariate outlier. This was predictable, since as we stated before observation 9 is an outlier only for the first variable.
Observation 1 is under level 99%. However its behavior is on the edge. In fact also its bivariate analysis shows that in any analysis that involved precip and/or days (the variables for which it was a univariate outlier) it was very close to the quantiles, and in some cases it also went over the 99% level. If we don't consider it officially an outlier we can anyway consider that it's a point of particular interest.
Observation 11 is well known since it's been detected in all previous analysis. The over 99% level with respect to the chisquared quantile confirms that it is clearly a multivariate outlier. Furthermore we can recall the fact that it was a univariate outlier for Manuf and pop, but also the only bivariate outlier for these ones and for all bivariate comparison involving manuf and/or Pop.

Exercise 2

In this exercise we had to perform PCA on a simulated sample $X = (X_1, X_2, X_3)$ distributed as $N_3(\mu, \Sigma)$, with :

$$\mu = [1 \quad -1 \quad 2]^t$$
$$\Sigma = \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix}$$

Let's first of all detect the eigenvalues and eigenvector for studying the PCs: Since the problem is symmetric and we have to order the eigenvalues, for now we can assume $\rho > 0$. In fact the case $\rho < 0$ present the same eigenvalues:

$$\lambda_1 = 1 + \sqrt{2}\rho$$
$$\lambda_2 = 1$$
$$\lambda_3 = 1 - \sqrt{2}\rho$$

While the eigenvectors are:

$$a_1 = (1, \sqrt{2}, 1)$$
$$a_2 = (1, 0, -1)$$
$$a_3 = (1, -\sqrt{2}, 1)$$

To see for which ρ the PC1 and PC2 are enough to describe the data we can check that

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} \geq 0.80$$

and we get $|\rho| \geq 0.2828$.
The meaning of the eigenvectors is the contribution of the original variables to each PC. It can be easily seen that a_1, a_2 and a_3 are perpendicular, But they are not unitary normed. To have this we should divide by their euclidean norm.
If we standardize them and analyse each contribution to the first two PCs we get that the contribution of the first variable to the first PC1 is less than the one for PC2. The second variable contributes only to the PC1, since $a_2^2 = 0$. The third variable contributes in a way similar to the first variable, but for the PC2 the contribution is negative.

We can easily see that Z, as defined, is a linear transformation of X with a particular matrix $A \in \Re^{(2,3)}$:

$$Z = AX$$
$$A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$$

So we get that:$Z \sim N_2(A\mu, A\Sigma A^t) = N_2(\mu_z, \Sigma_z)$, with:

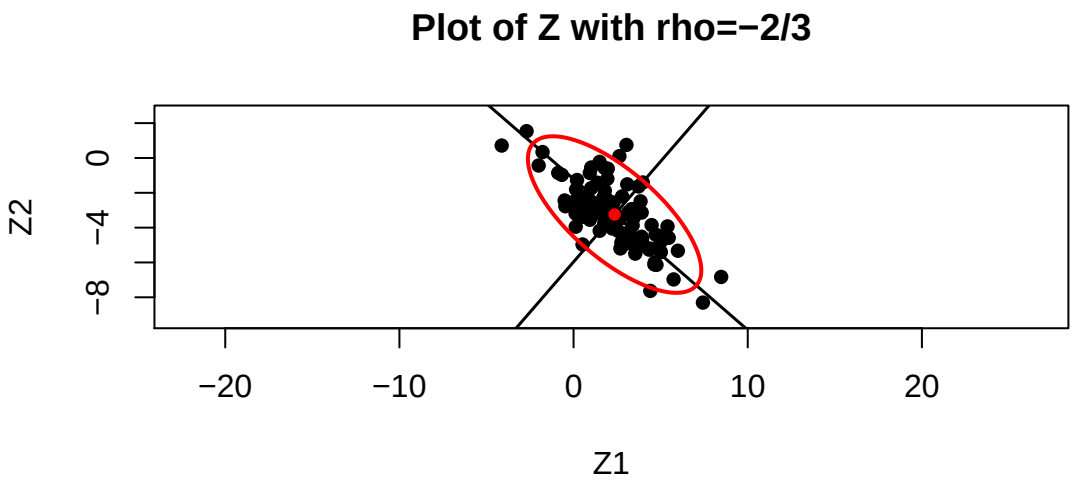
$$\mu_z = (2, -3)$$

$$\Sigma_z = \begin{pmatrix} -2\rho + 2 & 2\rho - 1 \\ 2\rho - 1 & -2\rho + 2 \end{pmatrix}$$

If $\rho = -2/3$:

$$\Sigma_z = \begin{pmatrix} \frac{10}{3} & -\frac{7}{3} \\ -\frac{7}{3} & \frac{10}{3} \end{pmatrix}$$

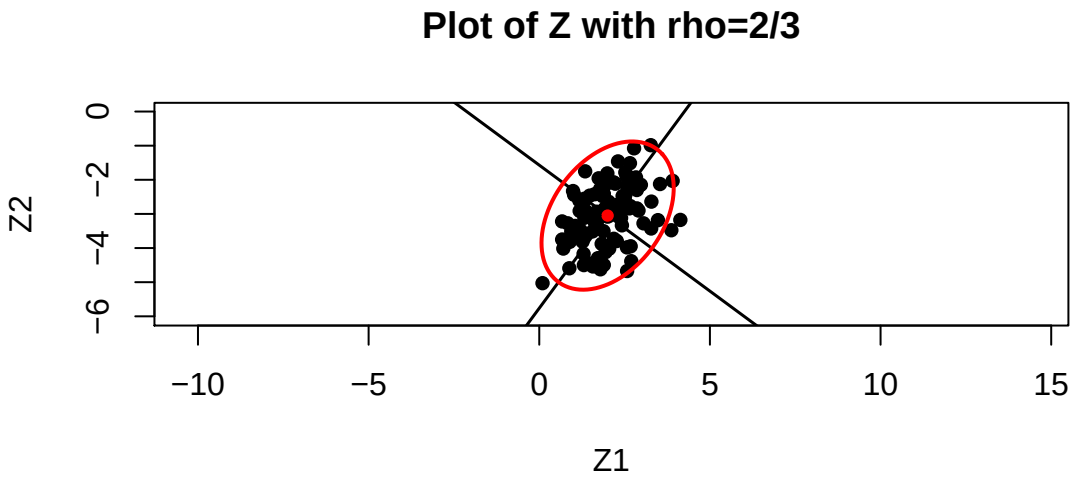
So now we can perform our simulation and show the ellipse. This Ellipse will be quite large since the determinant is $|\Sigma_z| = 2$. Furthermore from the sign of the covariance we can see that the form will be:



If $\rho = 2/3$:

$$\Sigma_z = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

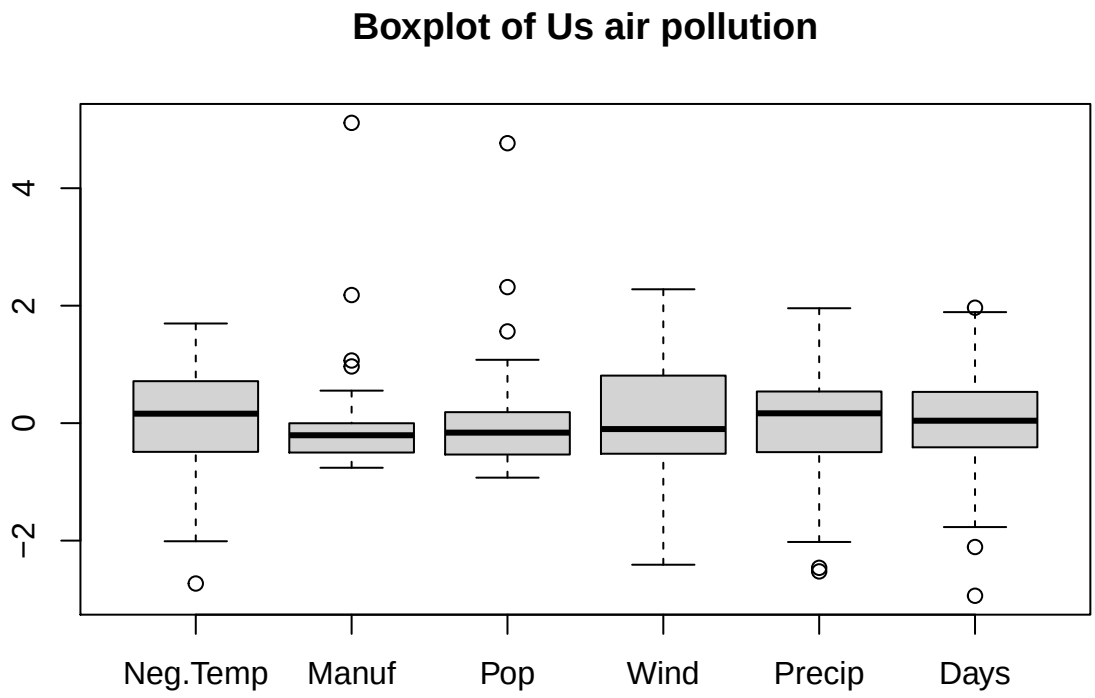
Then it would change the orientation of the ellipse because of the sign of the covariance and will be really more concentrated around the sample mean, since the determinant becomes $|\Sigma_z| = 0.7$:



Exercise 3

Point 3.1

For this point we have to perform a Principal Component Analysis. We have standardized data of Usair pollution and made a boxplot to spot any possible univariate outlier. Note, all these plots non-standardized are already present at point 1.2, but they were not scaled. In any case the outliers are the same. This is just a reminder of those plots.




```
## Importance of components:
##                           PC1   PC2    PC3    PC4    PC5     PC6
## Standard deviation     1.482 1.225 1.1810 0.8719 0.3385 0.18560
## Proportion of Variance 0.366 0.250 0.2324 0.1267 0.0191 0.00574
## Cumulative Proportion  0.366 0.616 0.8485 0.9752 0.9943 1.00000
```

We have performed principal component analysis on standardized data ,which will give us 6 new variables, which are linear combinations of the original ones, having as a aim a data dimensionality reduction preserving all or most of information provided by all original data. Furthermore we presented a summary of the basic feature of these new six variables.

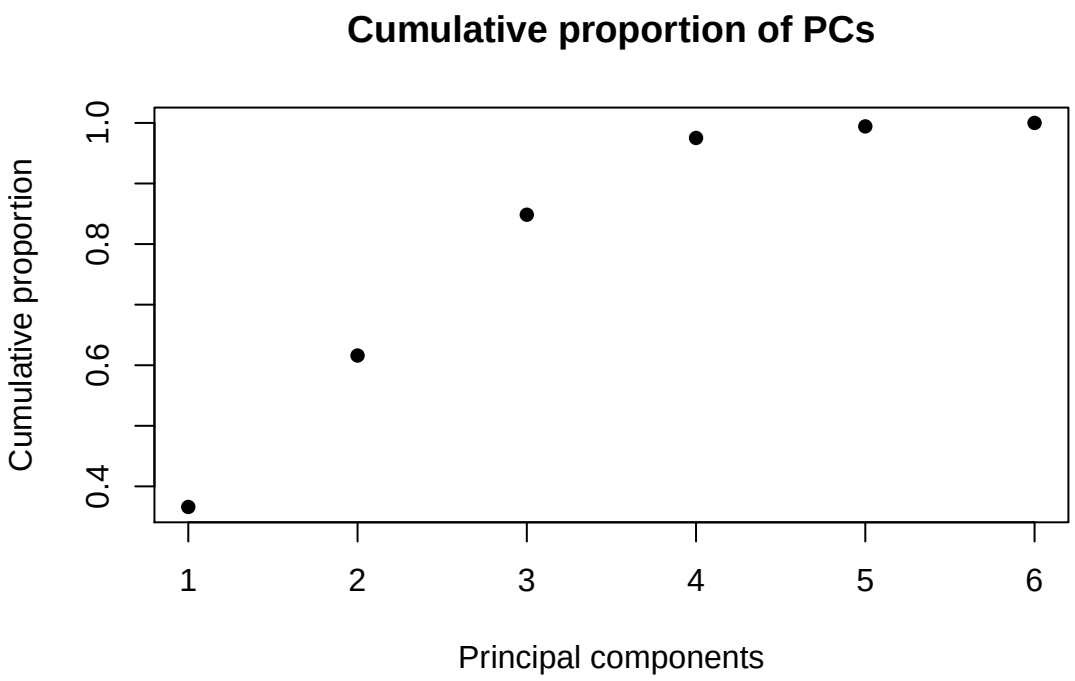
Now we present the standard deviation of each principal component and the coefficients of rotation of the axis given by the linear transformation performed in principal components which gives us back a sense of how original variables "load" high or low on each principal component.

```
## Standard deviations (1, .., p=6):
## [1] 1.4819456 1.2247218 1.1809526 0.8719099 0.3384829 0.1855998
##
## Rotation (n x k) = (6 x 6):
```

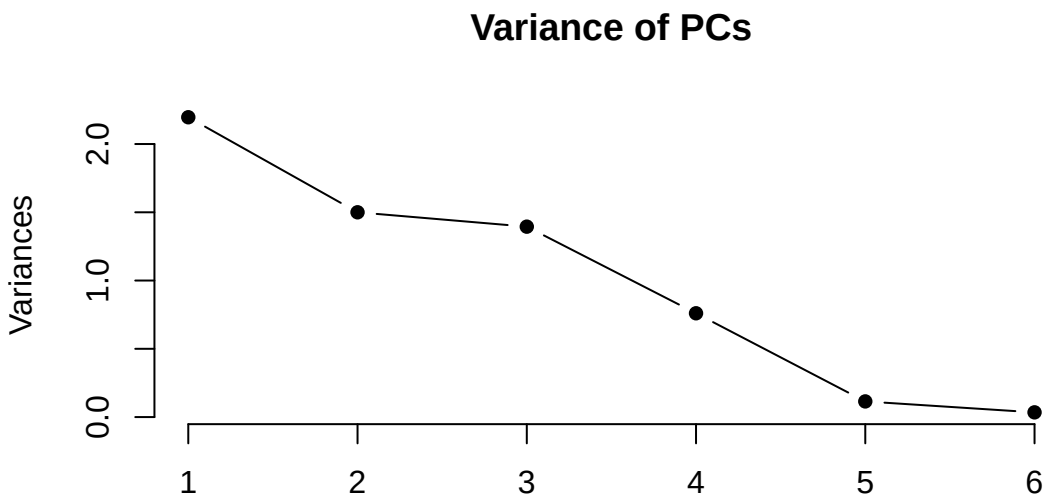
```
##                 PC1        PC2         PC3         PC4         PC5         PC6
## Neg.Temp  0.32964613 -0.1275974  0.67168611  0.30645728 -0.55805638  0.13618780
## Manuf     0.61154243  0.1680577 -0.27288633  0.13684076 -0.10204211 -0.70297051
## Pop       0.57782195  0.2224533 -0.35037413  0.07248126  0.07806551  0.69464131
## Wind      0.35383877 -0.1307915  0.29725334 -0.86942583  0.11326688 -0.02452501
## Precip   -0.04080701 -0.6228578 -0.50456294 -0.17114826 -0.56818342  0.06062222
## Days      0.23791593 -0.7077653  0.09308852  0.31130693  0.58000387 -0.02196062
```

Point 3.2

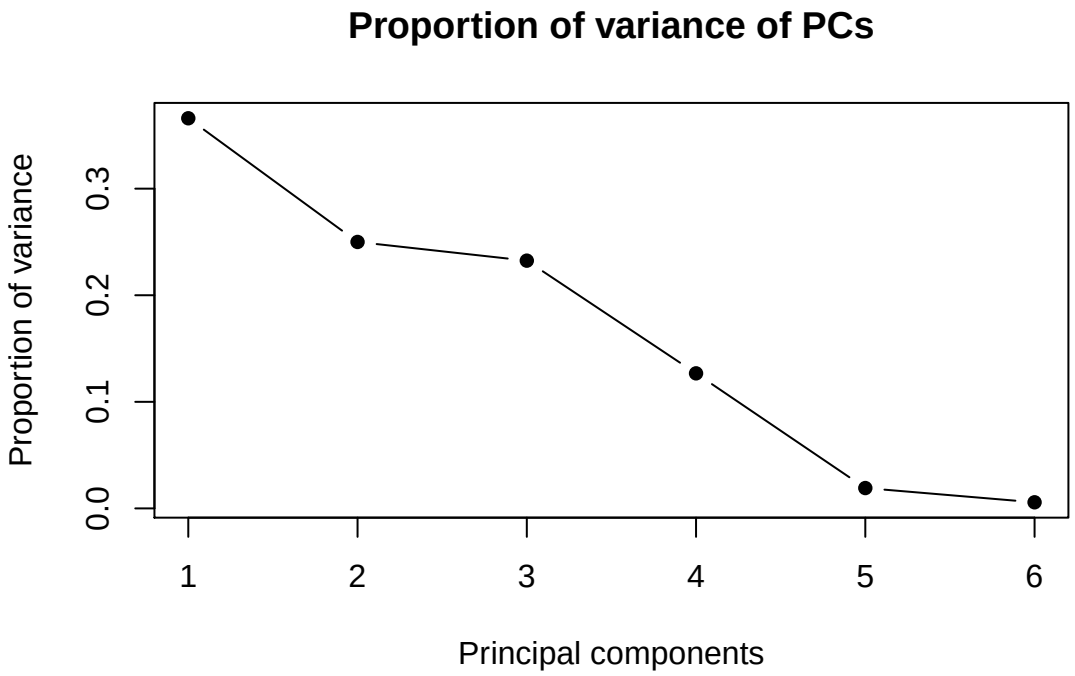
Now we plot the cumulative proportion of variance explained by the principal components, to show how many principal components account for the majority of total variance and taking the decision to retain some accordingly, to achieve the data dimensionality we were aiming.



Here we produced this plot having in mind basically the same thing as before and take the decision to retain just three of the six PCs we have found. In fact, as we can see, the variance is decreasing, so the first PCs explain the majority of the variance.



Now we will produce a graph representing the proportion of variance (not the variance itself) of each PC in a decreasing order , in order to see the proportion explained by each PC.



Our suggestion, based on three preceding plots, is to retain just the first three PCs in order to get a satisfactory data reduction, while preserving the most of the information. This holds true because the first three PCs account for almost 85% of the total variance as we can see in the first table.

Note: From the screeplot of variances we can see that the "elbow" is in the PC5. This should suggest us

to take the first 4 PCs. However, as stated before, the first three PCs explain more than the 85% of the information, that is a satisfactory proportion of information and the dimensionality reduction would be important. So in front of an important dimensionality reduction we retain that to perform a satisfactory analysis we can consider the first three PCs.

Point 3.3

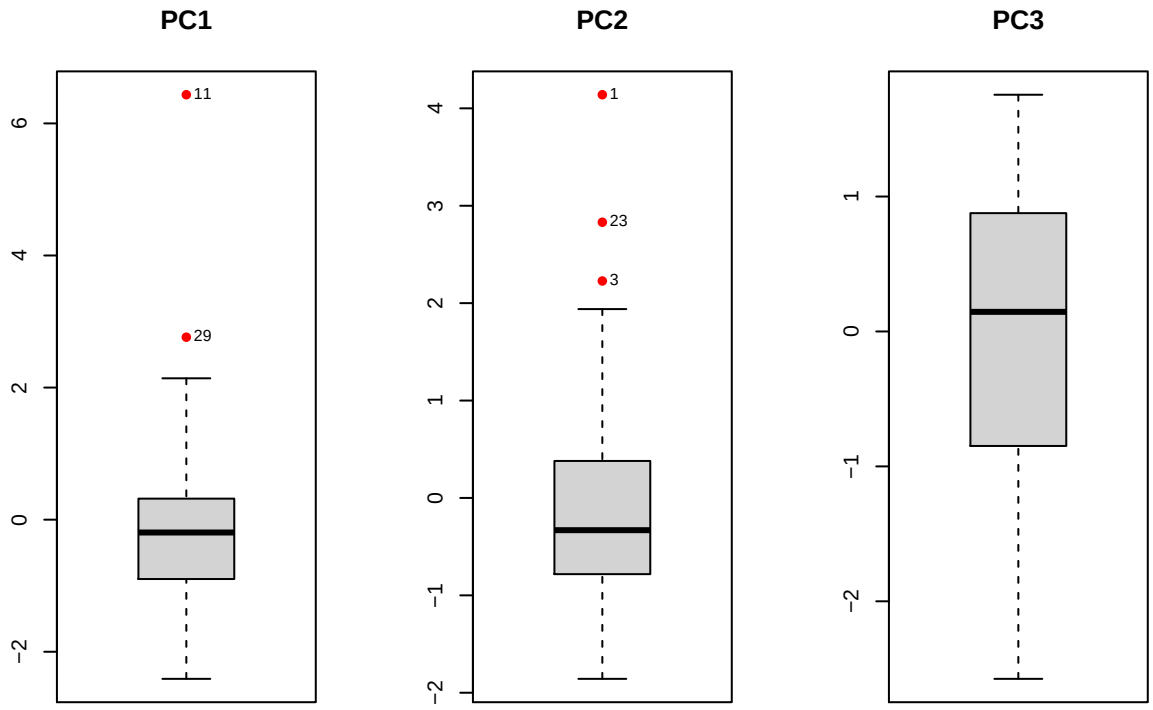
The first three principal components are a linear combination of the original variables and in the interpretation is related to the loadings. In fact the rotation coefficient of each variable on the principal component gives us a hint on how the variables influence the values of the principal component. Given that, we propose the interpretation of the first three principal components: the first one, having high loadings on human ecology, is a component which accounts for a lot of information given by the two original variables Man. and Pop.; the last two instead have high loadings on Climate factors, so they represent the information given by the original variables Neg.Temp.,Days and Precipitation. The loading of the variable Wind was really high only on PC4, but it's clear that the information present in the first three PCs is enough to continue considering it. In any case its behavior was quite regular with respect to every analysis performed.

Point 3.4

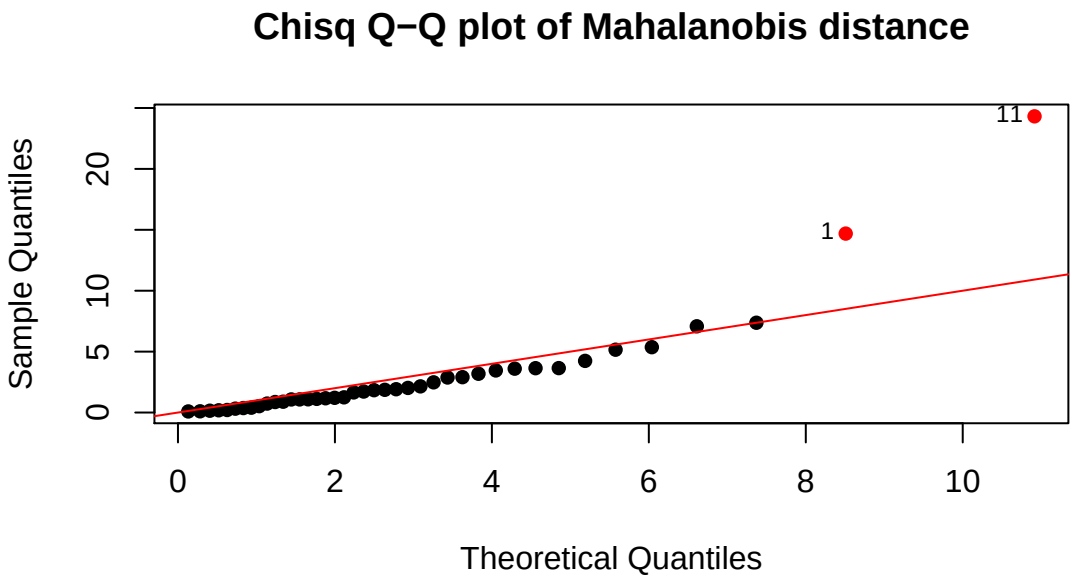
Now we can study the scores. The scores are the data transformed by applying the linear combination proposed by the PCs.

PC1	PC2	PC3
Min. :-2.4102	Min. :-1.8572	Min. :-2.5746
1st Qu.:-0.8989	1st Qu.:-0.7823	1st Qu.:-0.8487
Median :-0.1944	Median :-0.3308	Median : 0.1452
Mean : 0.0000	Mean : 0.0000	Mean : 0.0000
3rd Qu.: 0.3183	3rd Qu.: 0.3797	3rd Qu.: 0.8768
Max. : 6.4340	Max. : 4.1397	Max. : 1.7544

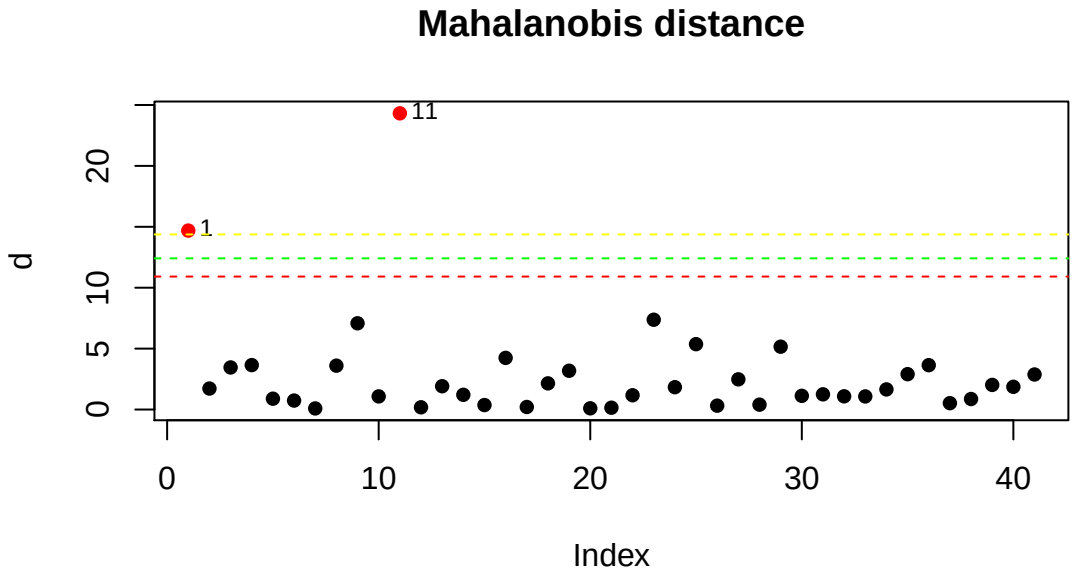
Now we can start studying the normality, but before doing it let's see the three boxplots.



One interesting thing to note is that observation 11 is an outlier for PC1, while we recognized it also as an outlier in point 1.2 for Manuf and Pop. This confirm that their influence in this PC is important.
Now we'll check the multivariate normality for these three PCs.



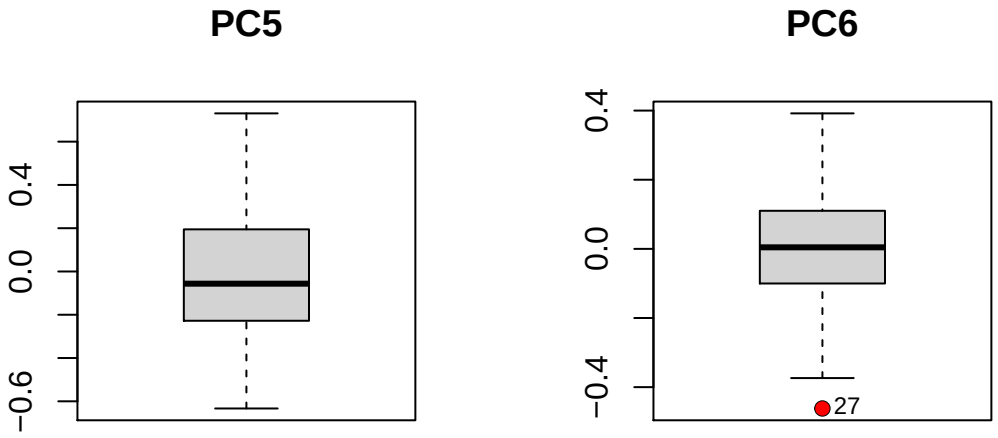
We have computed the Mahalanobis distances of the values of our three principal components and produced a graph which compares them with quantiles of a chi squared distribution in order to check for multivariate normality and to check for multivariate outliers which we have highlighted in red and labeled with the observation they correspond to. From this we can deduce that observation 1 and 11 are also multivariate outliers. We detected them also in exercise 1, but this time we were not able to detect the observation 9, which does not appear from this analysis.



We produced a scatterplot of Mahalanobis distances together with quantile lines of chi squared distribution which correspond to 95%,97.5% and 99% to highlight even more the observations which are multivariate outliers.

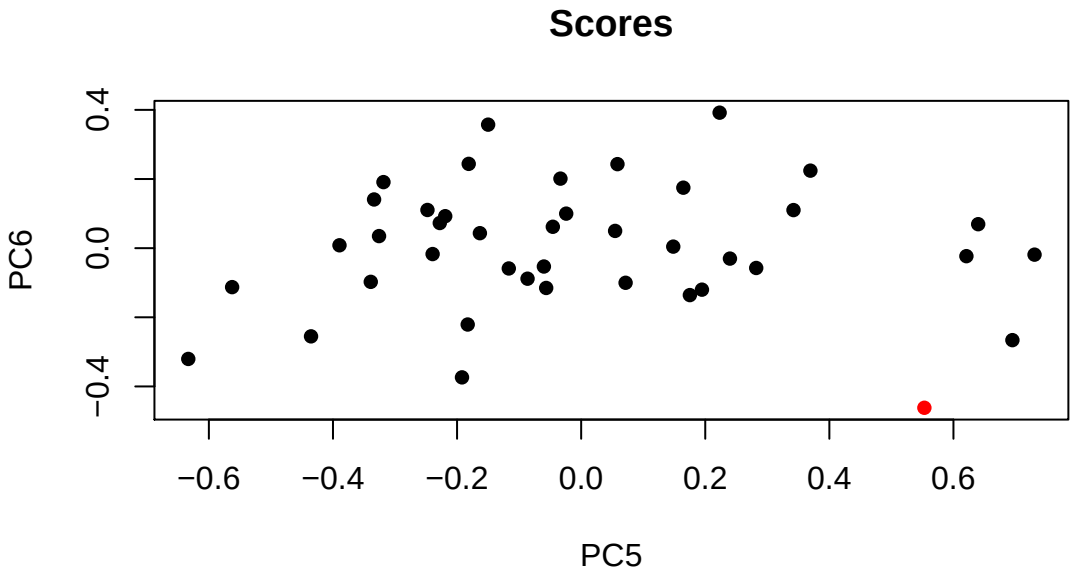
Point 3.5

Now by looking at the last two PCs we can start looking for suspect observations. Recall that a suspect observation is a univariate outlier for the last PCs.
We will first draw the two boxplot for each PCs interested.



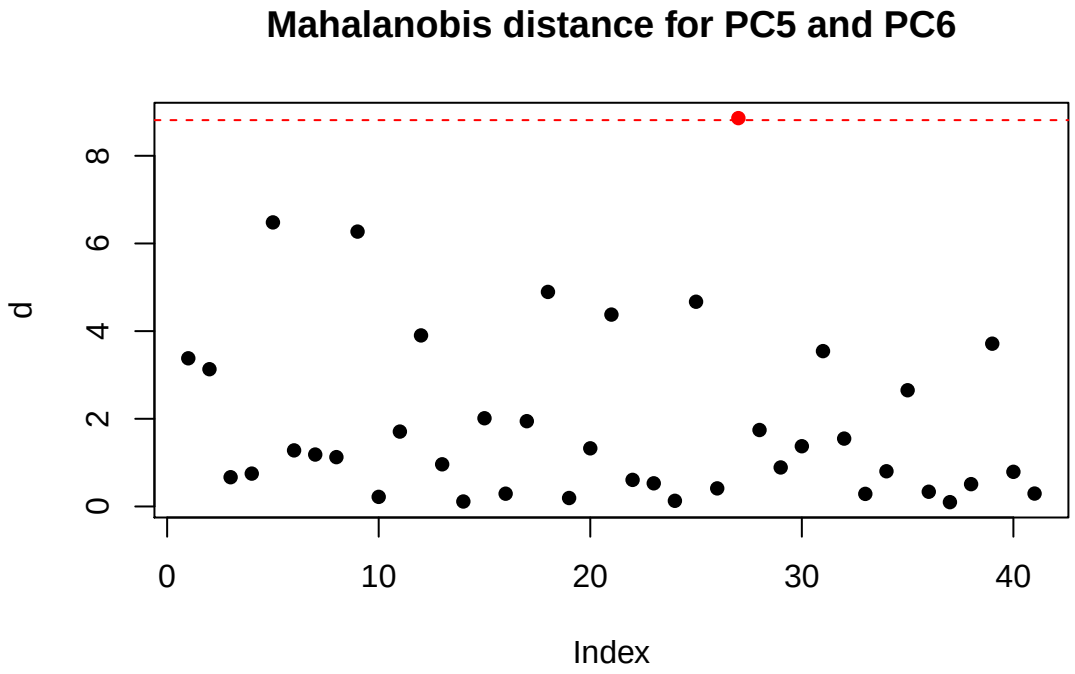
In the PC5 we couldn't find any outlier, while in PC6 there is one: score 27.

As a further analysis we can check the bivariate plot to see if it is a multivariate outlier. We can see that it is not, since it doesn't seem to be strange for PC5.



In fact if we restrict to the axis of PC5 we can see that it is not an outlier for PC5.

When we were looking for multivariate outliers with the original variables we checked if from the bivariate plot there could arise some interesting comment and then we compared the Mahalanobis distance. In this case we had just one possible suspect observation, that didn't behave strangely with respect to PC5, so we conclude that it's not a multivariate outlier. We can confirm this also looking at the comparison of the Mahalanobis distance with the chi-squared quantile (95%).



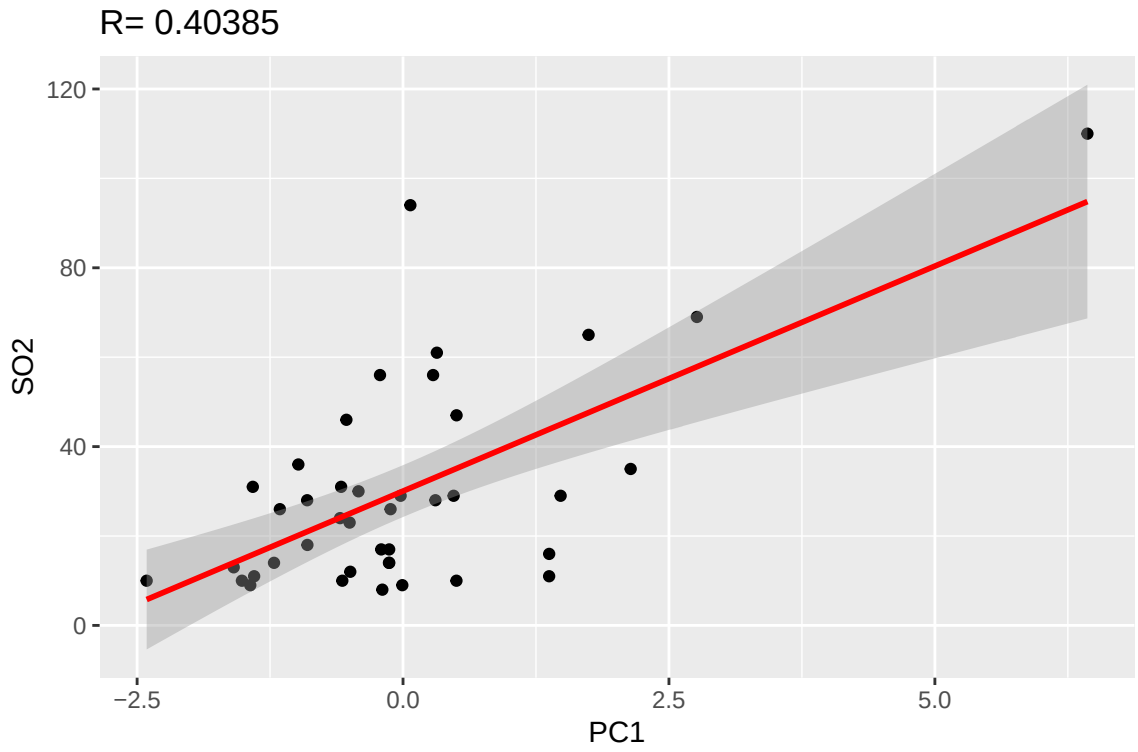
As we can see it's just a little bit over the 95% quantile.
To conclude observation 27 is a suspect observation since it's an outlier for PC6, but it's not a multivariate outlier since from the last plot we can see it's just over the 95% quantile.
Furthermore we can see that it was an outlier for Manuf, and as expected Manuf has an important impact on PC6.

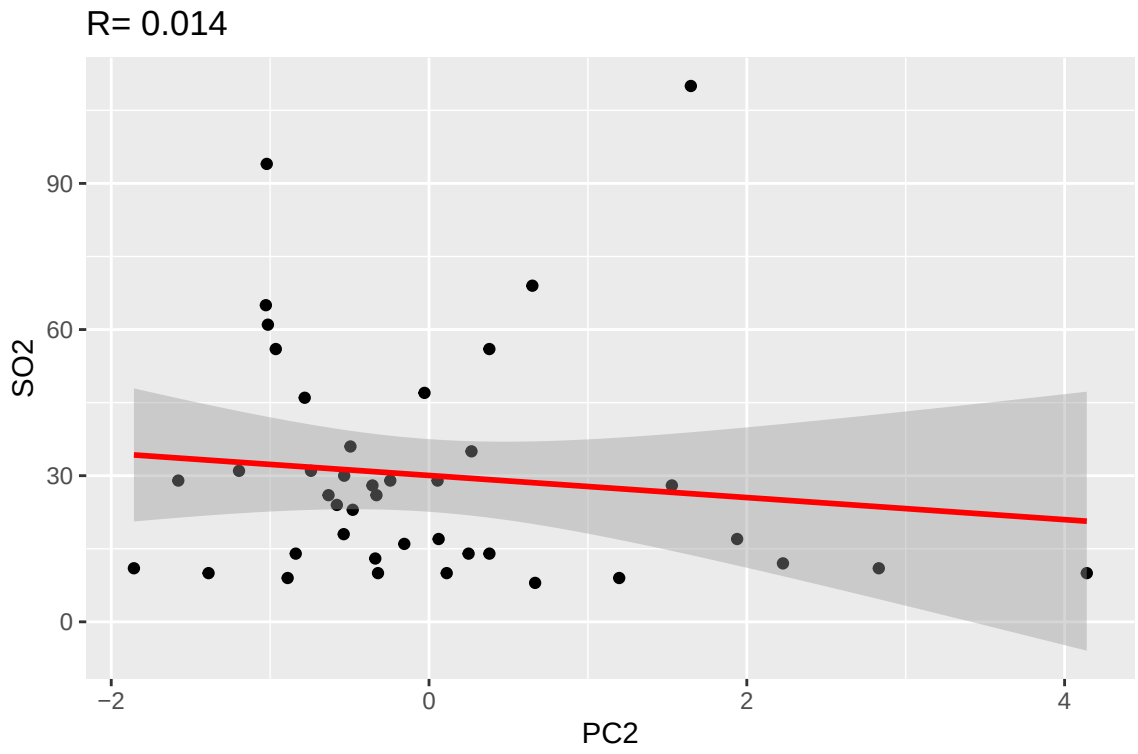
Point 3.6

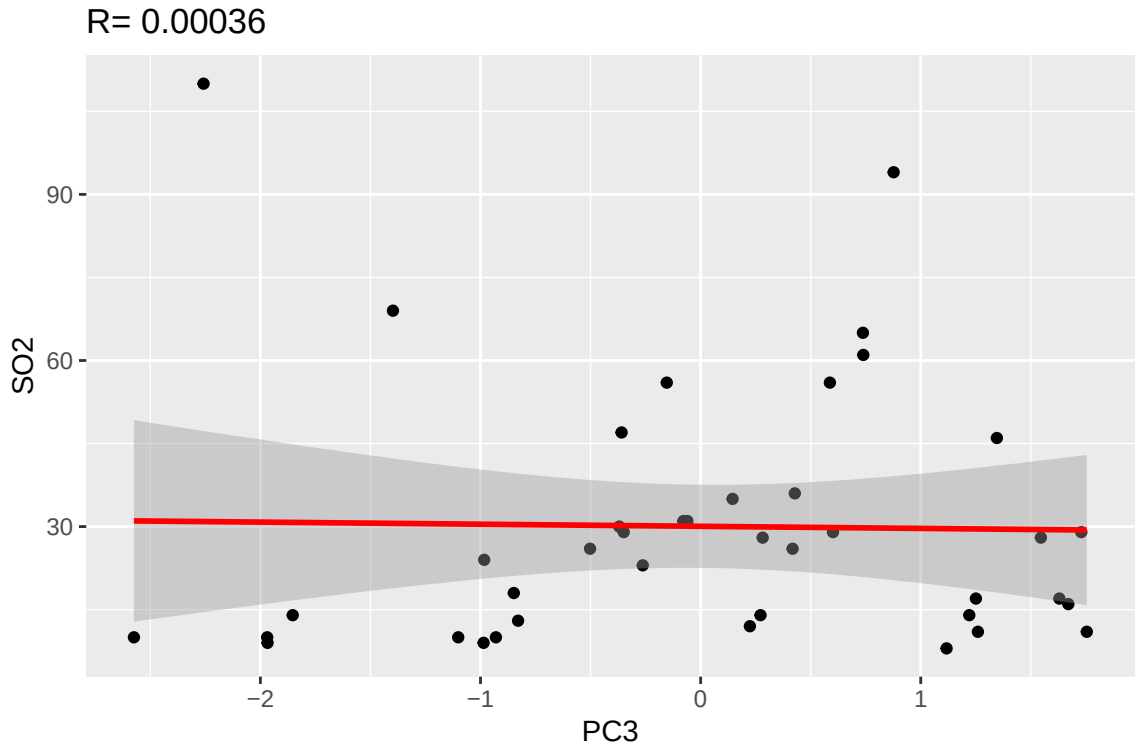
During this analysis we will study what happens when we take into account the PCA with respect to the prediction of the SO2 values, that we have excluded from the beginning.
We will associate the first 3 PCs with every city and then use them to explore their relationship with SO2.

SO2	PC1	PC2	PC3
Min. : 8.00	Min. :-2.4102	Min. :-1.8572	Min. :-2.5746
1st Qu.: 13.00	1st Qu.:-0.8989	1st Qu.:-0.7823	1st Qu.:-0.8487
Median : 26.00	Median :-0.1944	Median :-0.3308	Median : 0.1452
Mean : 30.05	Mean : 0.0000	Mean : 0.0000	Mean : 0.0000
3rd Qu.: 35.00	3rd Qu.: 0.3183	3rd Qu.: 0.3797	3rd Qu.: 0.8768
Max. :110.00	Max. : 6.4340	Max. : 4.1397	Max. : 1.7544

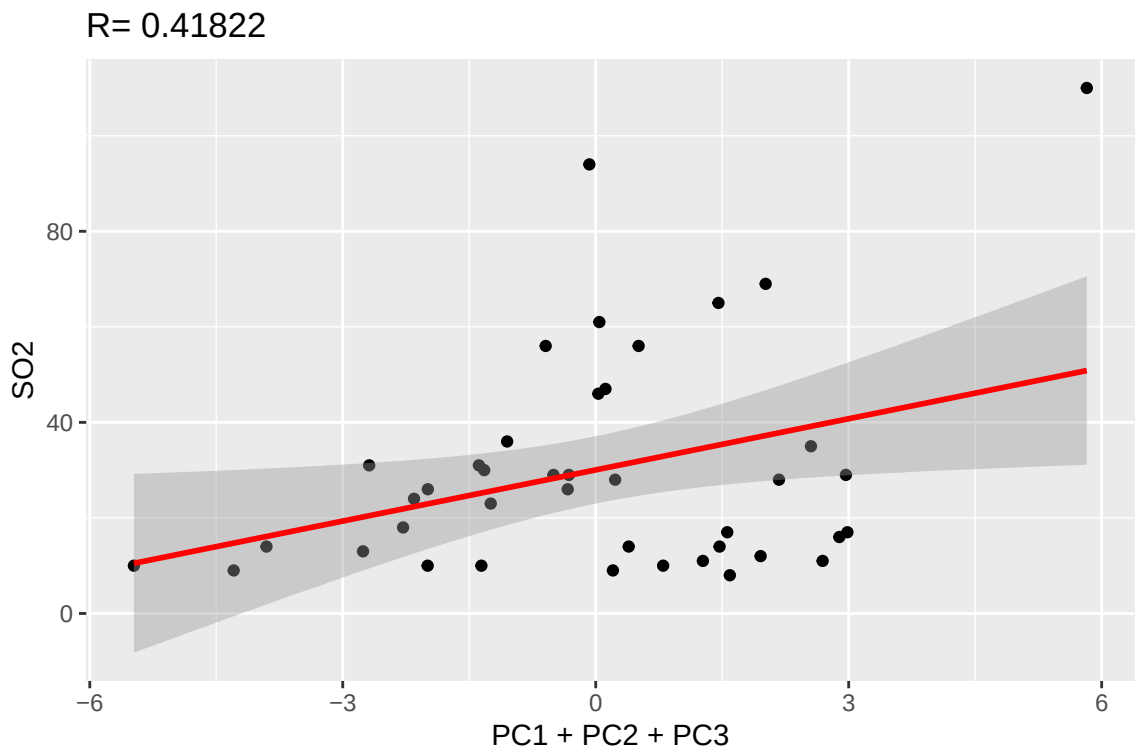
Now we can start our analysis. First of all, we will do 3 scatterplots to visualize the correlation between SO2 and each PC and we will build a linear model to see the relationship between SO2 and the first PC. Then we will calculate the R-squared, an important statistical measure which is a regression model that represents the proportion of the difference or variance in statistical terms for a dependent variable which can be explained by an independent variable or variables. In short, it determines how well data will fit the regression model. (In the following plots the title is the multiple R-squared).





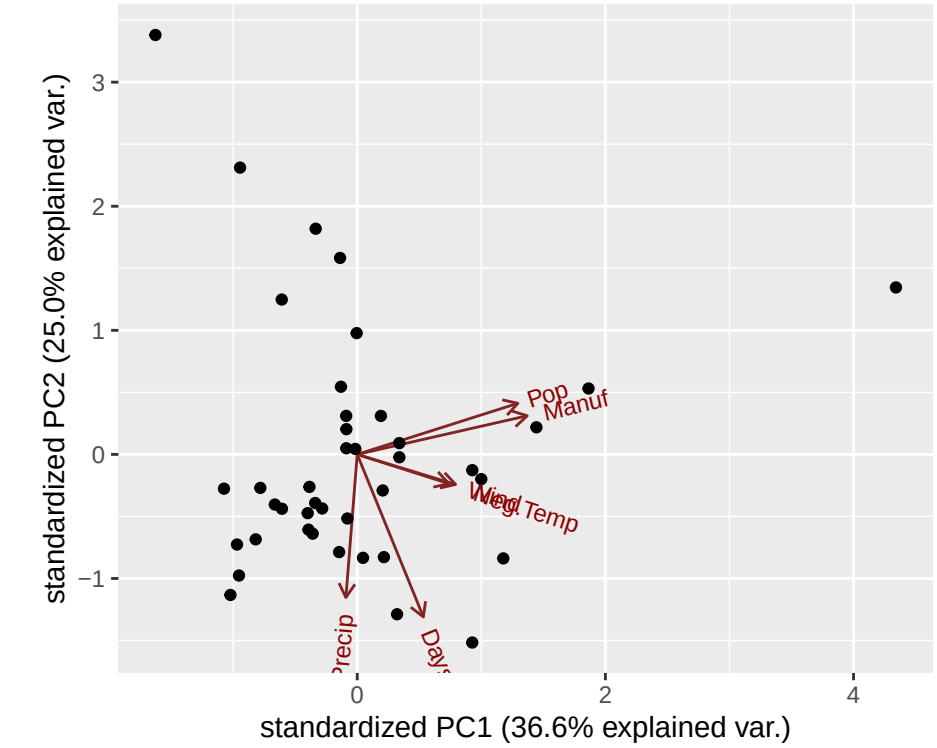


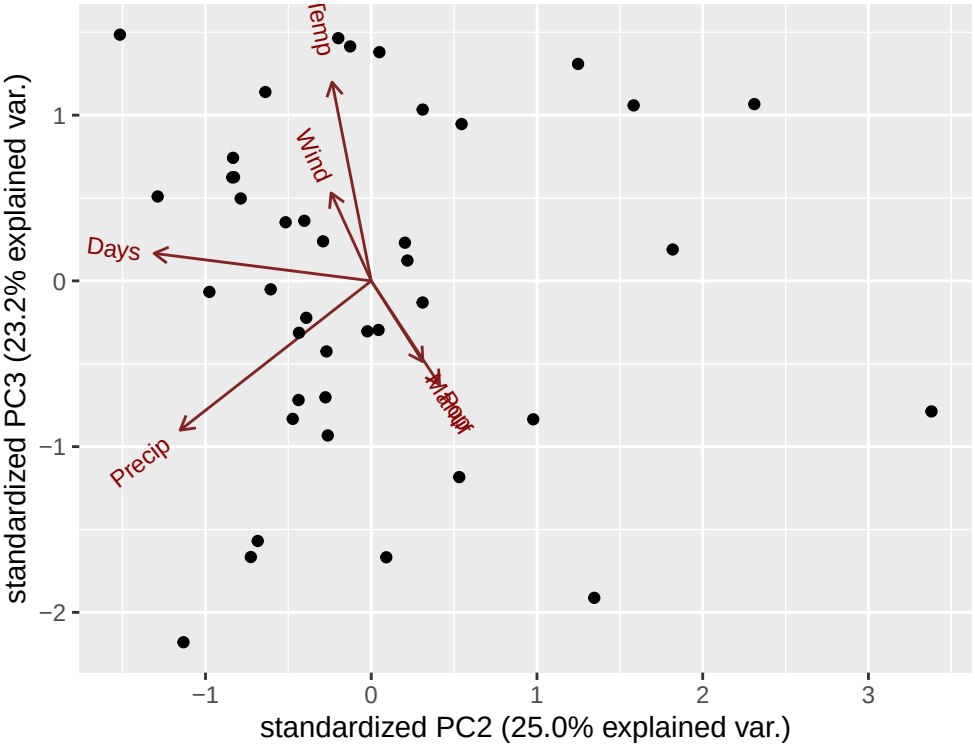
Now we know that the first PC is the one which explain the most correlation with SO2.
We can then build a linear model considering all the 3 PC together, knowing that the relationship with the first PC is much stronger than the others. (the linear models with PC2 and PC3 have an R2 of 0.014 and 0.0003649 respectively, so there is almost no correlation between those PCs and SO2).

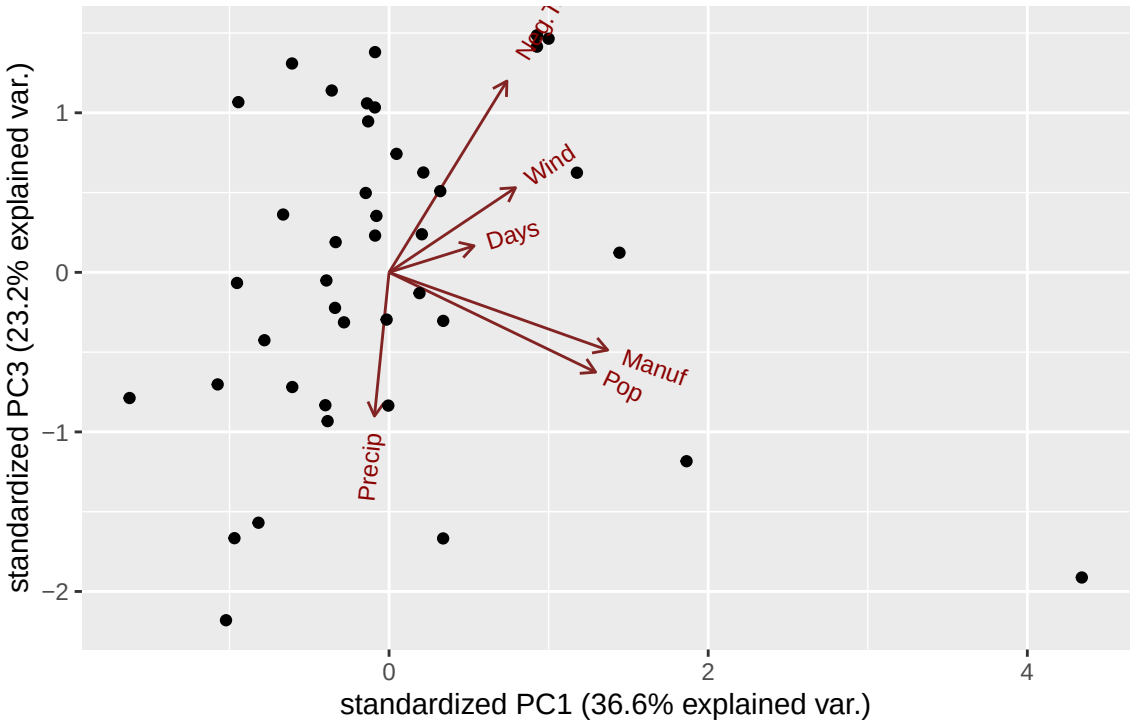


As expected, R is a bit higher (thanks to the more information gained) than the one with only the first PC, but not in a relevant way.

Now it's time to plot our PCA. We will make 3 biplots, which include both the position of each sample in terms of PC1, PC2 and PC3 (every plot will show a different pair each times and it also will show how the initial variables map onto this. We will use the ggbiplot package, which offers a user-friendly and pretty function to plot biplots. A biplot is a type of plot that will allow you to visualize how the samples relate to one another in our PCA, so which samples are similar and which are different) and will simultaneously reveal how each variable contributes to each principal component.


standardized PC2 (25.0% explained var.)
standardized PC1 (36.6% explained var.)
Pop
Manuf
Wind
Temp
Precip
Days
3
2
1
0
−1
0
2
4

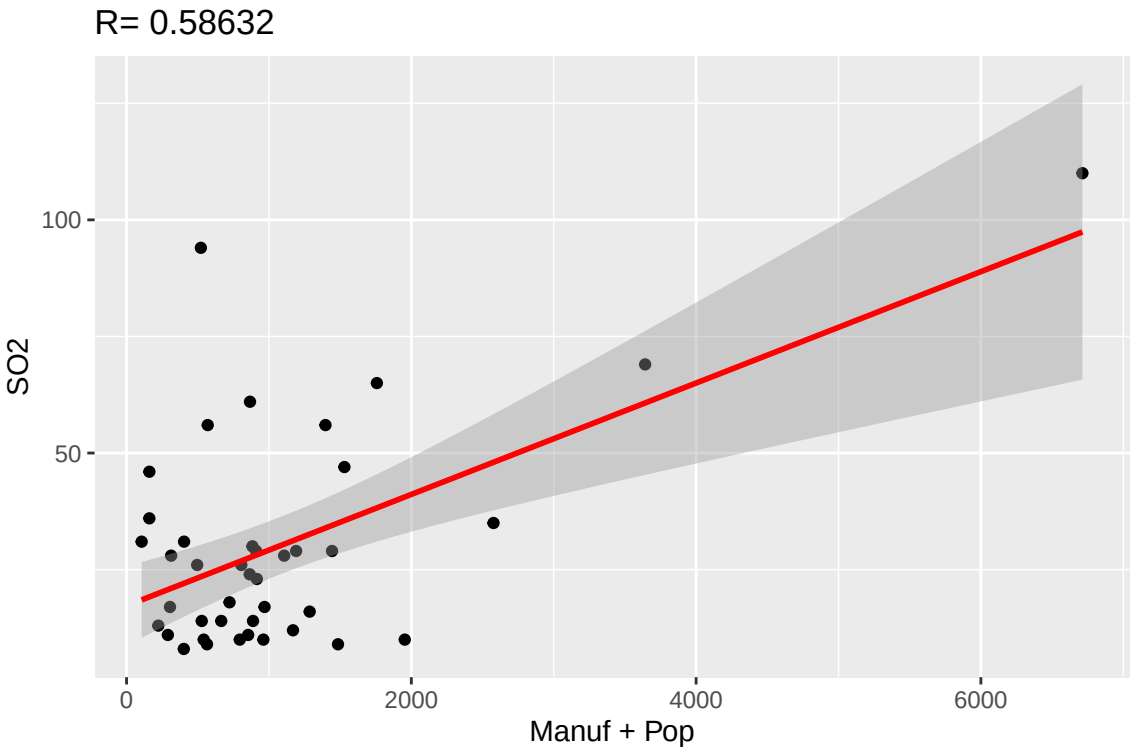

standardized PC3 (23.2% explained var.)
standardized PC2 (25.0% explained var.)
Temp
Wind
Days
Precip
1
0
−1
−2
−1
0
1
2
3

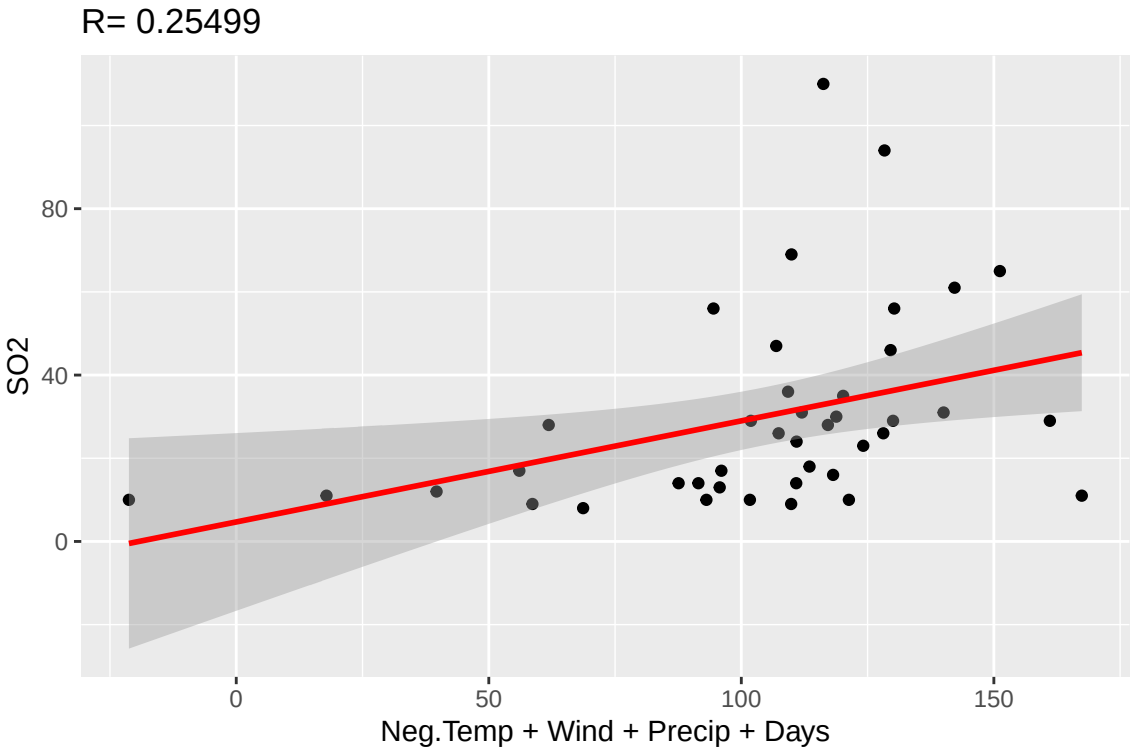


We can see how in the first and in the third biplot 'Manuf' and 'Pop' (the two Human ecology variables) are the ones which contribute the most to PC1, the PC that in the previous linear models has the highest R-squared. We can ignore the second biplot because PC2 and PC3 have almost no correlation with SO2. In fact they seem to be more affected from the climate variables. Furthermore, in the next exercise we can expect that the Human ecology variables have a higher R-squared than the Climate variables.

Point 3.7

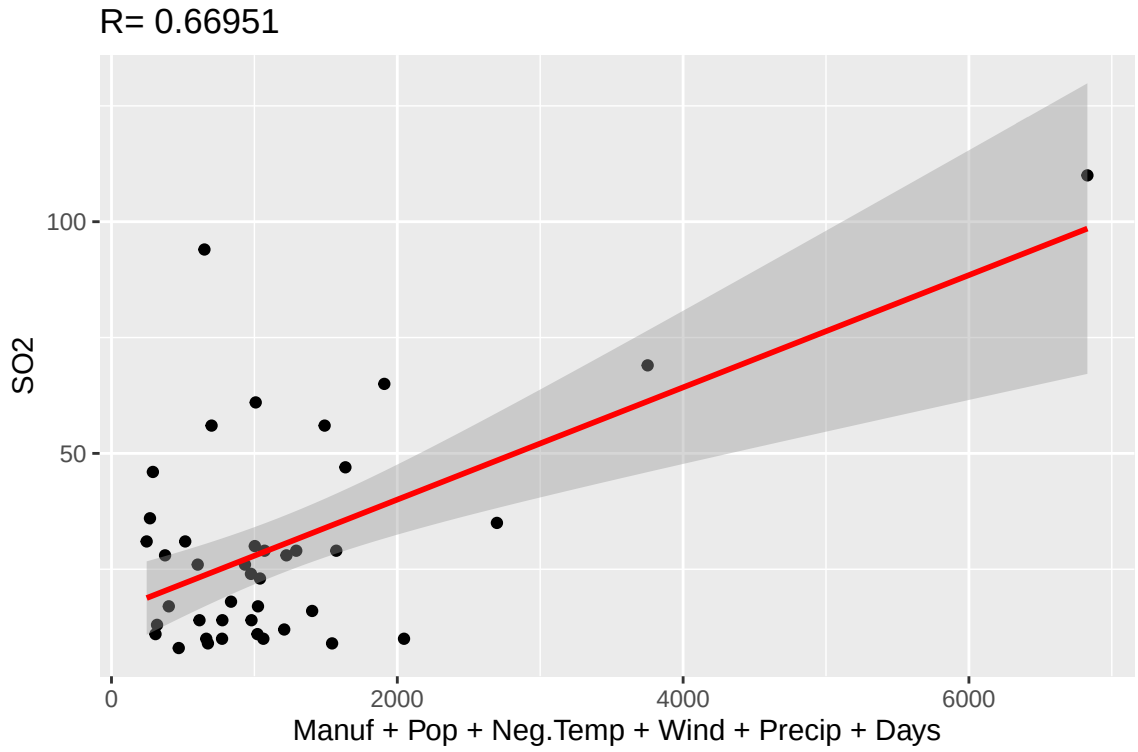
In this last point we have to investigate about the use of multiple linear regression on the air pollution data using the human ecology and climate variables.





As expected, the Human ecology variables have a stronger correlation with SO2 than the Climate variables.

We can confirm this stronger relationship with a further analysis.




```
## $coefficients
##                 Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 111.72848064 47.31810073  2.361221 0.0240867374
## Neg.Temp      1.26794109  0.62117952  2.041183 0.0490557189
## Manuf         0.06491817  0.01574825  4.122245 0.0002277862
## Pop          -0.03927674  0.01513274 -2.595482 0.0138461970
## Wind         -3.18136579  1.81501910 -1.752800 0.0886503978
## Precip        0.51235896  0.36275507  1.412410 0.1669175999
## Days         -0.05205019  0.16201386 -0.321270 0.7499724652
##
## $r.squared
## [1] 0.6695118
##
## $adj.r.squared
## [1] 0.6111904
```

If the p-value for a variable is less than our significance level (let's presume 0.025), our sample data provide enough evidence to reject the null hypothesis for the entire population. In our case the data favor the hypothesis that there is a non-zero correlation. According to the p-values of the predictors, 'Manuf' and 'Pop', are the most relevant (<0.025) to predict SO2. It confirms our previous analysis in the exercise 3.6.