# MULTIVARIATE STATISTICAL ANALYSIS

REPORT FOR MULTIVARIATE STATISTICAL ANALYSIS - PROBLEM SET 1

**Bruno Casella**
Dipartimento di Informatica
Università di Torino
casella@di.unito.it

April 4, 2022

**ABSTRACT**

This report describes the results and the implementation of Problem Set 1 of Multivariate Statistical Analysis. All the R code is available at the following GitHub link: https://github.com/CasellaJr/MultivariateStatAnalysis. The exercises has been solved in collaboration with the PhD student Lorenzo Paletto.

## Exercise 1

The **air pollution** data consists of 7 measurements recorded at n = 41 cities in the United States. Variables are

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

We ignore the SO2 variable and concentrate on the remaining 6, two of which relate to human ecology (Manuf. and Pop) and four to climate (Neg. Temp, Wind, Precip, Days).

The **sample mean vector** is:

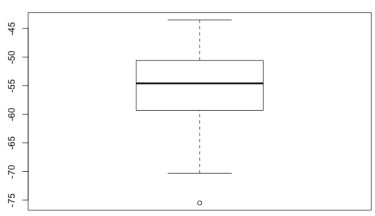
| Neg. Temp | Manuf. | Pop. | Wind | Precip. | Days |
|---|---|---|---|---|---|
| -55.763415 | 463.097561 | 608.609756 | 9.443902 | 36.769024 | 113.902439 |

The **sample correlation matrix R** is:

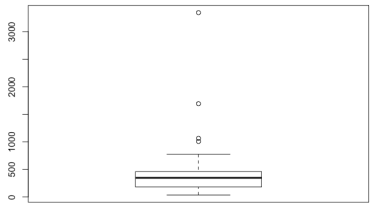
$$\begin{pmatrix} 1.00000000 & 0.19004216 & 0.06267813 & 0.34973963 & -0.38625342 & 0.43024212 \\ 0.19004216 & 1.00000000 & 0.95526935 & 0.23794683 & -0.03241688 & 0.13182930 \\ 0.06267813 & 0.95526935 & 1.00000000 & 0.21264375 & -0.02611873 & 0.04208319 \\ 0.34973963 & 0.23794683 & 0.21264375 & 1.00000000 & -0.01299438 & 0.16410559 \\ -0.38625342 & -0.03241688 & -0.02611873 & -0.01299438 & 1.00000000 & 0.49609671 \\ 0.43024212 & 0.13182930 & 0.04208319 & 0.16410559 & 0.49609671 & 1.00000000 \end{pmatrix}$$

Manuf. and Pop. are strongly positive correlated. Days and Neg. Temp have a moderate uphill positive relationship. Days and Precip. are moderately positive correlated. Neg. Temp and Precip. are negatively correlated.

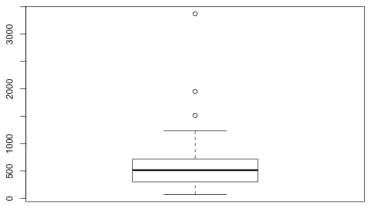
These are the boxplots of each variable.



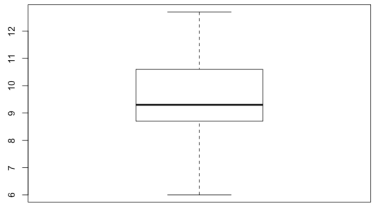
(a) Boxplot Neg Temp



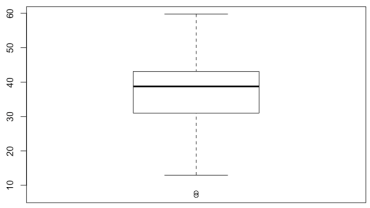
(b) Boxplot Manuf



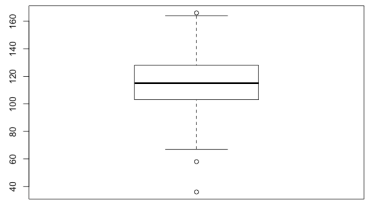
(c) Boxplot Pop



(d) Boxplot Wind



(e) Boxplot Precip

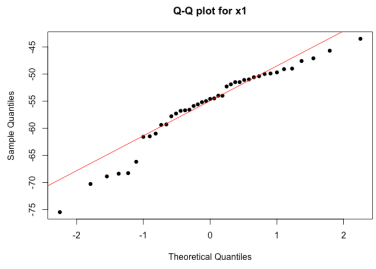


(f) Boxplot Days

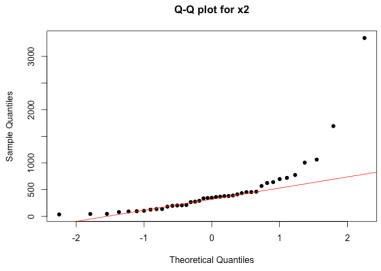
Figure 1: Boxplots of all the 6 variables

Boxplot of Neg Temp shows only 1 outlier that corresponds to the 9-th observation. It is about the city of Miami.
Boxplot of Manuf shows 4 outliers. However, the text of the exercise asks for only 2 outliers. So, the most external outliers are the observations 11 (Chicago) and 29 (Philadelphia).
Boxplot of Pop shows 3 outliers. The most external are the same of the previous boxplot.
In boxplot of Wind there are no outliers.
Boxplot of Precip shows 2 outliers that correspond to observation 1 (Phoenix) and 23 (Alburquerque).
Boxplot of Days shows 3 outliers. The most external are the same of the previous boxplot.

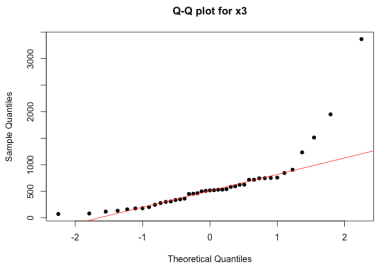
Construct a normal Q-Q plot for each variable and comment about normality.



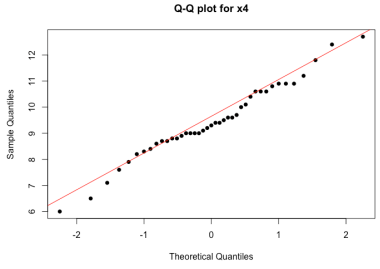

(a) Q-Q plot Neg Temp



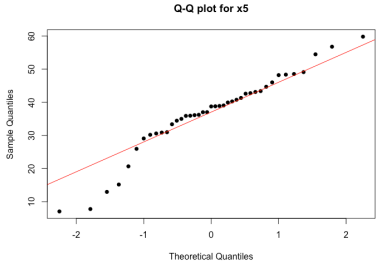

(b) Q-Q plot Manuf



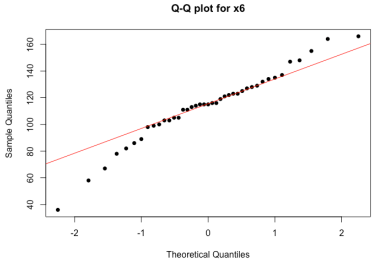

(c) Q-Q plot Pop




(d) Q-Q plot Wind




(e) Q-Q plot Precip

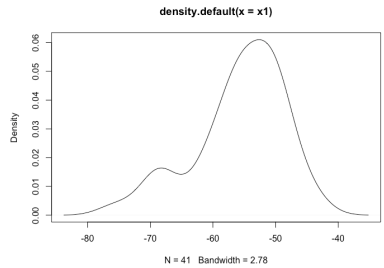



(f) Q-Q plot Days

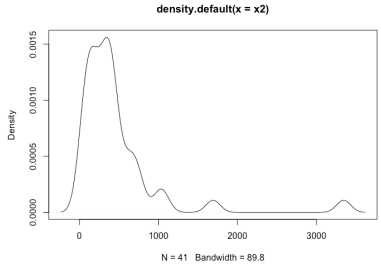
Figure 2: Q-Q plots of all the 6 variables

Q-Q plot of Neg Temp shows a quite normal distribution.
Q-Q plot of Manuf shows a right skewed distribution.
Q-Q plot of Pop shows a right skewed distribution.
Q-Q plot of Wind shows a normal distribution.
Q-Q plot of Precip shows a little left skewed distribution.
Q-Q plot of Days shows a normal distribution.

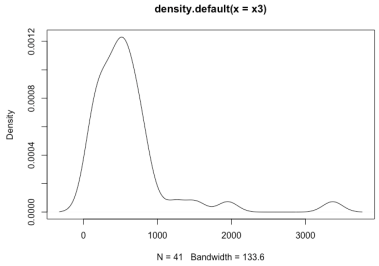
These interpretations can be confirmed plotting the density or computing the skewness of each variable.



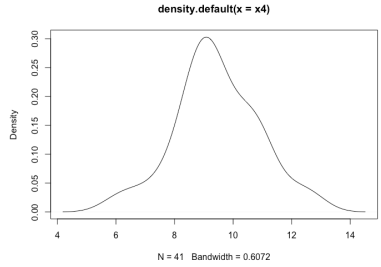

(a) Density of Neg Temp



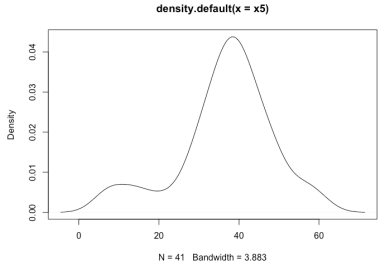

(b) Density of Manuf



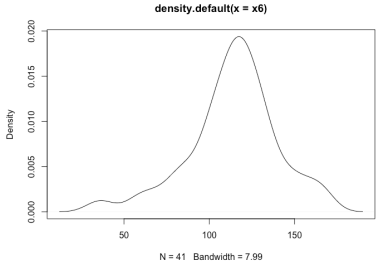

(c) Density of Pop




(d) Density of Wind




(e) Density of Precip




(f) Density of Days

Figure 3: Density of all the 6 variables

Values of skewness:

| Neg. Temp | Manuf. | Pop. | Wind | Precip. | Days |
|---|---|---|---|---|---|
| -0.8540294 | 3.616089 | 3.052242 | 0.002776073 | -0.7186492 | -0.5708491 |

Scatterplots:

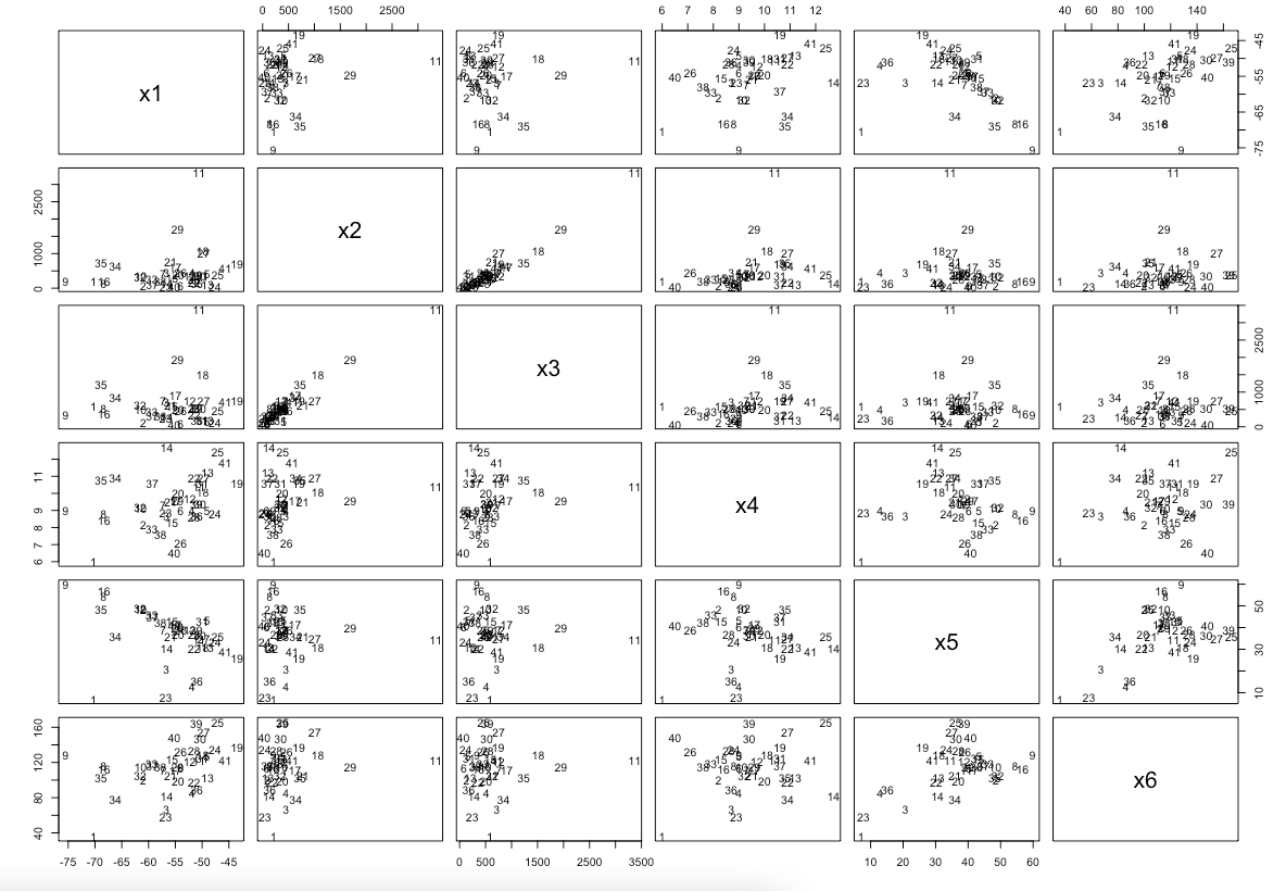



Figure 4: Scatterplots

We can see that the outliers are the observations 11 and 29, that are the same detected with the boxplots of Manuf and Pop.

Now I am going to plot the chi-squared Q-Q plot of the squared Mahalanobis distances.

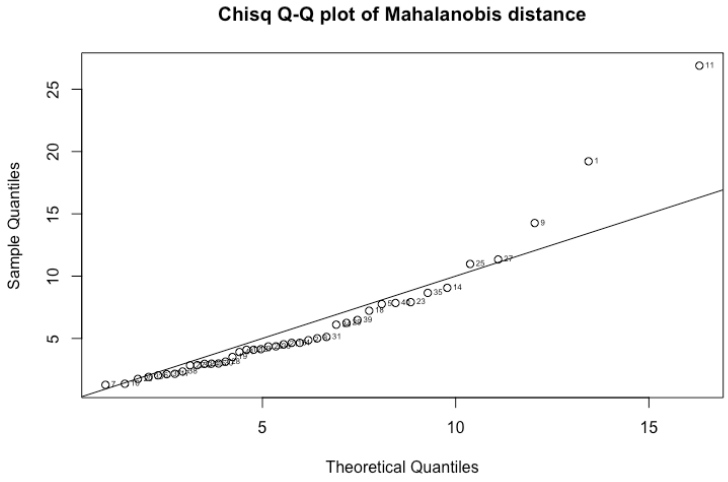



Figure 5: Chi-squared Q-Q plot of the squared Mahalanobis distances

We can see that observations 1, 9 and 11 are outliers. All these 3 observations have been previously detected using the boxplots.
This is the same plot after removing the outliers.

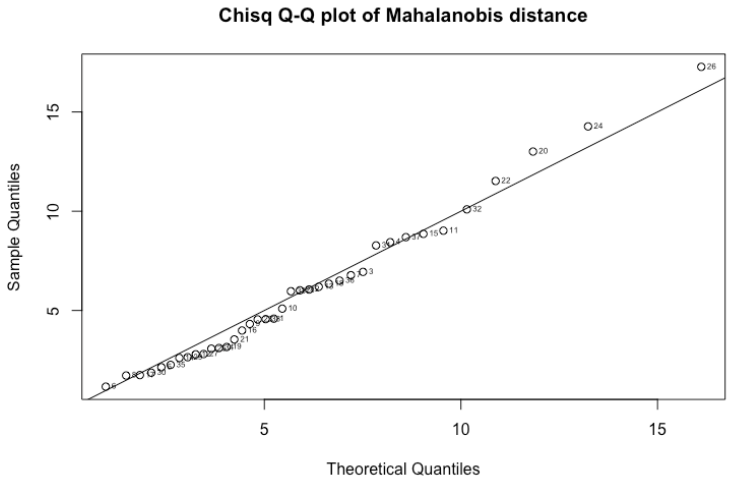



Figure 6: Chi-squared Q-Q plot of the squared Mahalanobis distances without outliers

## Exercise 2

## Exercise 3