Basilico Eleonora, Caporali Francesco, Malgieri Luigi Maria
02 April 2023

UNIVERSITÀ DI TORINO

Multivariate Statistical Analysis
Problem Set 1

Exercise 1

Consider the dataset `state.x77`, which contains 8 variables recorded to the 50 states of the United States of America in 1977.

```
st = as.data.frame(state.x77)
head(st)
```

	Population	Income	Illiteracy	Life Exp	Murder	HS Grad	Frost	Area
Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708
Alaska	365	6315	1.5	69.31	11.3	66.7	152	566432
Arizona	2212	4530	1.8	70.55	7.8	58.1	15	113417
Arkansas	2110	3378	1.9	70.66	10.1	39.9	65	51945
California	21198	5114	1.1	71.71	10.3	62.6	20	156361
Colorado	2541	4884	0.7	72.06	6.8	63.9	166	103766

Preliminary to our analysis, we rename some variables to avoid spaces, and add the variable *Density* representing the population density.

```
st[, 9] = st$Population * 1000 / st$Area
names(st)[c(4, 6, 9)] = c("Life_Exp", "HS_Grad", "Density")
```

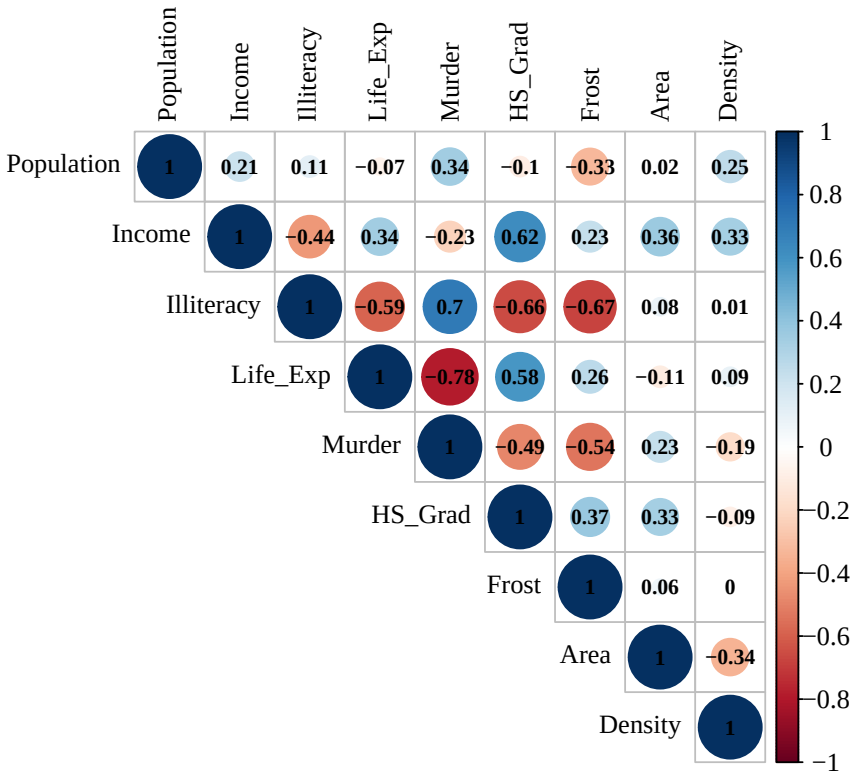
1.1

In order to compute and visualize the correlation matrix we use the function `corrplot`. The plot is displayed at the biginning of the next page.

```
cor_mat = round(cor(st), 2)
corrplot(cor_mat, type = "upper", method = "circle", tl.col = "black",
    addCoef.col = "black", number.cex = 0.65, tl.cex = 0.8, cl.cex = 0.8)
```

By analyzing the correlation matrix, we can observe that:

- the variables with the higher negative correlation (-0.78) are *Murder* and *Life_Exp*, which is reasonable since more murders imply an overall reduction of life expectancy;
- there are high negative correlations also between the variable *Illiteracy* and the variables *Frost*, *HS_Grad* and *Life_Exp* (-0.67, -0.66 and -0.59 respectively). The first one is unexpected: there are no natural considerations to justify this value. The second one is reasonable, since more graduates imply less illiterate citizens; however, it is quite odd that this correlation is lower, though only slightly, than the previous one. As for the third one, it makes sense too, since if a major percentage of citizens is educated, then the overall life expectancy should increase;
- the variables with the higher positive correlation (0.7) are *Murder* and *Illiteracy*, which is indeed a plausible relation;



- there are two other high positive correlations between the variable *HS_Grad* and the variables *Income* and *Life_Exp* (0.62 and 0.58 respectively). As we expect, the graduates percentage is highly correlated with both the variables *Income* and *Life_Exp*, which are two important indicators of well-being.

Moreover, we can note that the variables which are correlated the most with the others (both in a negative and in a positive sense) are *HS_Grad*, *Illiteracy* and *Murder*. On the contrary, the ones which are less correlated with the others are *Population*, *Area* and *Density*. Note also that the variable *Density* was derived from *Population* and *Area*, hence it is reasonable that it follows their behaviour in terms of correlations. However, it is quite surprising that the correlation between *Density* and the variables *Population* and *Area* is not so high, since, as we just said, it was derived from them.

1.2

In order to detect potential univariate outliers we first scale our dataset and then identify them as the values x such that

$$|x| > \Phi^{-1}(0.99),$$

where Φ is the cumulative distribution function (CDF) of a $\mathcal{N}(0,1)$.
We used the 99th percentile since by taking lower values the potential outliers would have been too many. However, we noticed the presence of another potential outlier for the variable *Area* by taking the 98.75th percentile.

```
scale_st = round(scale(st), 3)
quant = c(qnorm(0.99), qnorm(0.9875))
mat_1 = which(abs(scale_st[, ]) > quant[1], arr.ind = T)
mat_1 = as.data.frame(mat_1)
mat_2 = which(abs(scale_st[, ]) > quant[2], arr.ind = T)
mat_2 = as.data.frame(mat_2)
```

In the following dataframes (respectively `mat_1` and `mat_2`) the first column refers to the index of the corresponding observation, while the second one refers to the variable with respect to this observation is a potential univariate outlier.

```
mat_1
```

	row	col
California	5	1
NewYork	32	1
Alaska_Income	2	2
Louisiana	18	3
Alaska_Area	2	8
Massachusetts	21	9
NewJersey	30	9
RhodeIsland	39	9

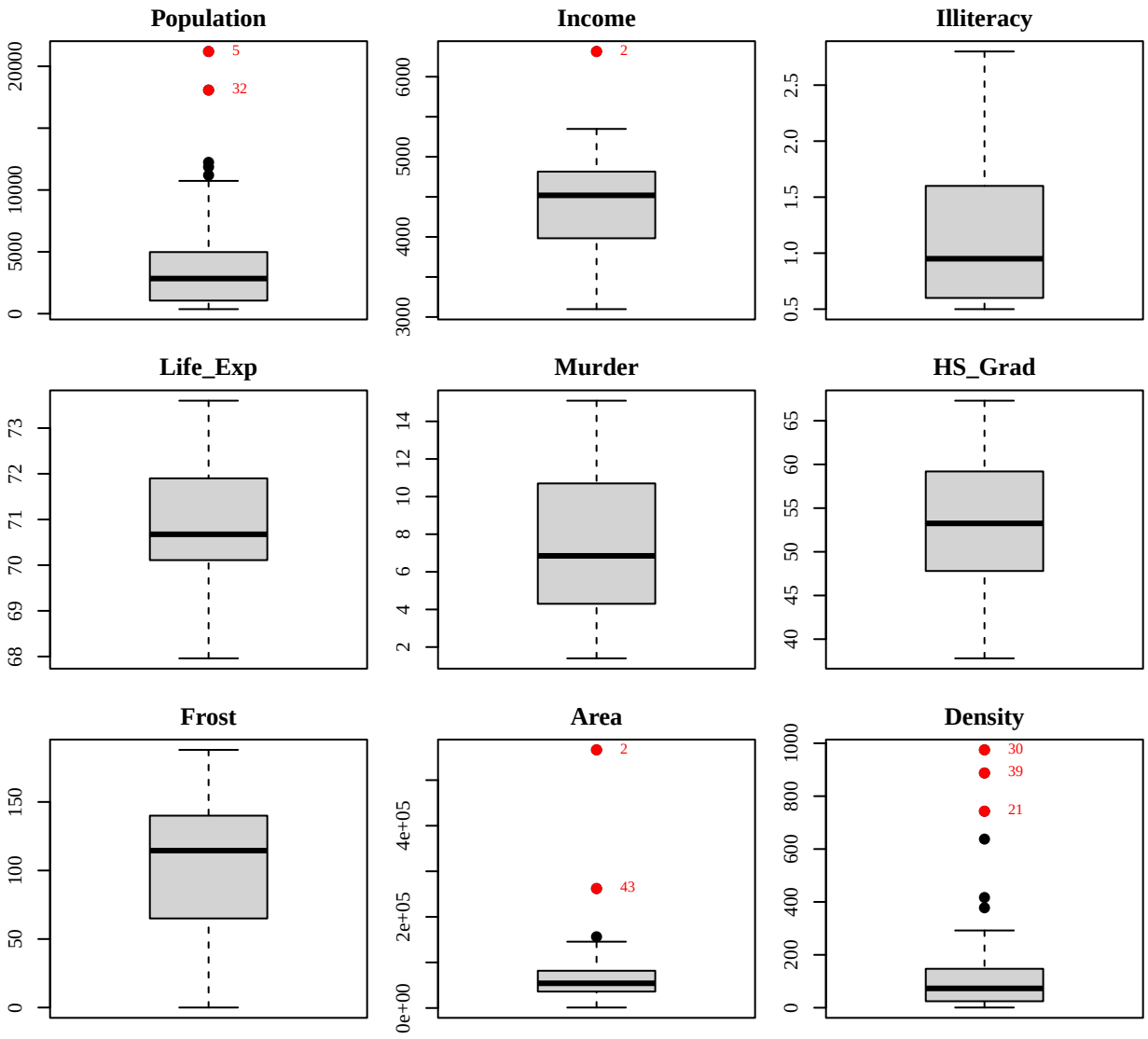
```
mat_2
```

	row	col
California	5	1
NewYork	32	1
Alaska_Income	2	2
Louisiana	18	3
Alaska_Area	2	8
Texas	43	8
Massachusetts	21	9
NewJersey	30	9
RhodeIsland	39	9

Note: In the tables above the names Alaska_Income and Alaska_Area both refer to observation 2 (i.e. Alaska). As it constitutes a potential univariate outlier for both variables, we opted for a relabelling to display them separately.

1.3

We report below the boxplots corresponding to each variable. We highlighted in red the potential univariate outliers found in point 1.2.



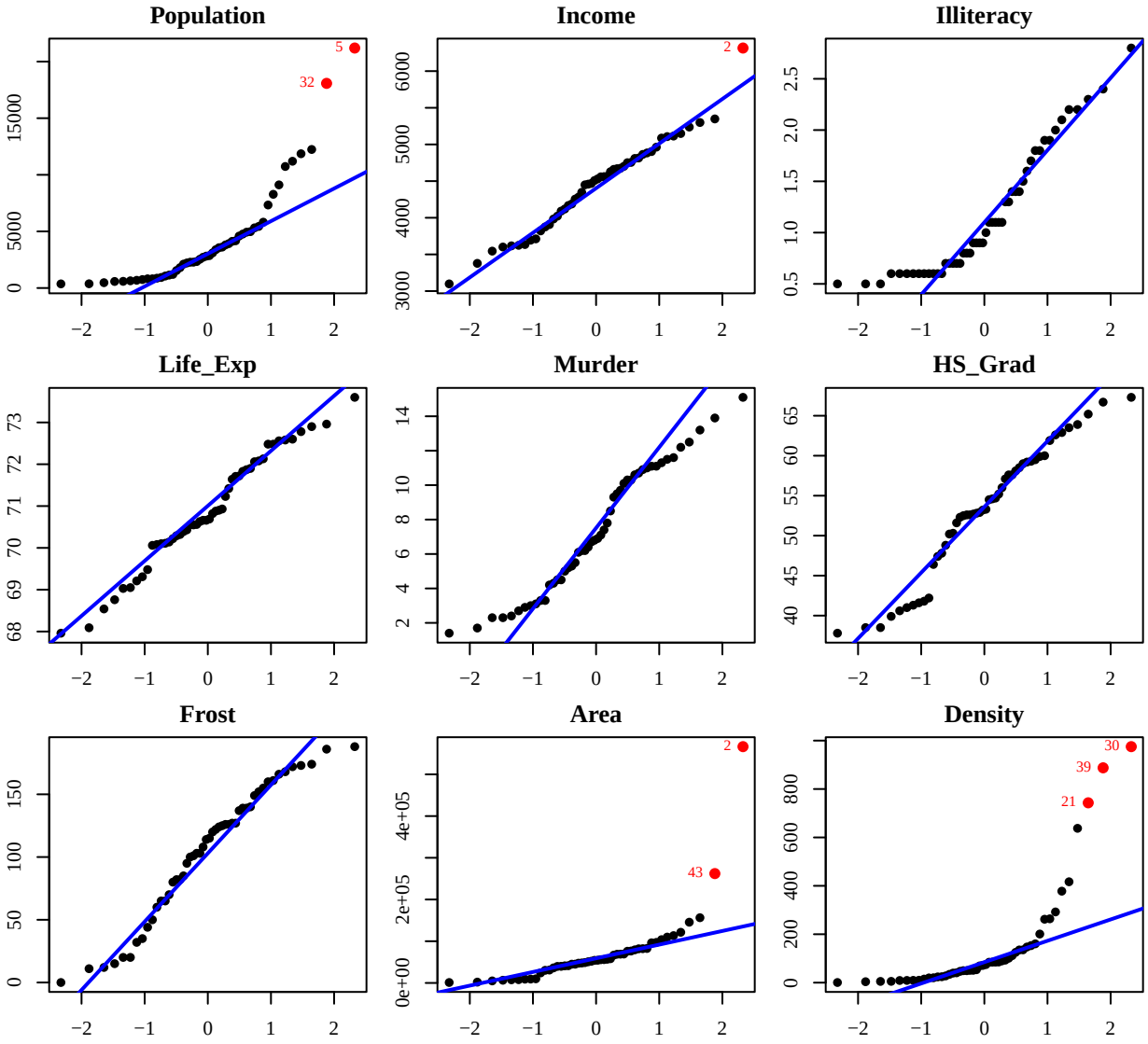
We can make the following considerations:

- according to what we found in point 1.2, the variables *Life_Exp*, *Murder*, *HS_Grad* and *Frost* seems not to have any potential univariate outlier;
- the potential outlier we identified for the variable *Illiteracy* (observation 18) does not show up in the corresponding boxplot. This is plausible, since the variable's distribution seems to have very fat tails. For this reason we choose to do not consider this observation as an outlier;
- the variable *Income* seems to have only the observation 2 as potential outlier, which is consistent with what we obtained in the previous point. Note also that the observation 2 is a potential outlier both for the variable *Income* and the variable *Area*;
- as for the remaning variables, the boxplots generated many other potential univariate outliers, but the ones that we did not detect in the previous point are not in the outer tails (at least 98.75th percentile) of the distribution. Hence, we will not think of them as outliers.

In conclusion, by looking at the boxplots we infer that observations $2, 5, 21, 30, 32, 39$ and 43 are potential univariate outliers.

1.4

In order to investigate the normality of the data, we first compare the theoretical and the sample quantiles with the aid of the corresponding Q-Q plots.



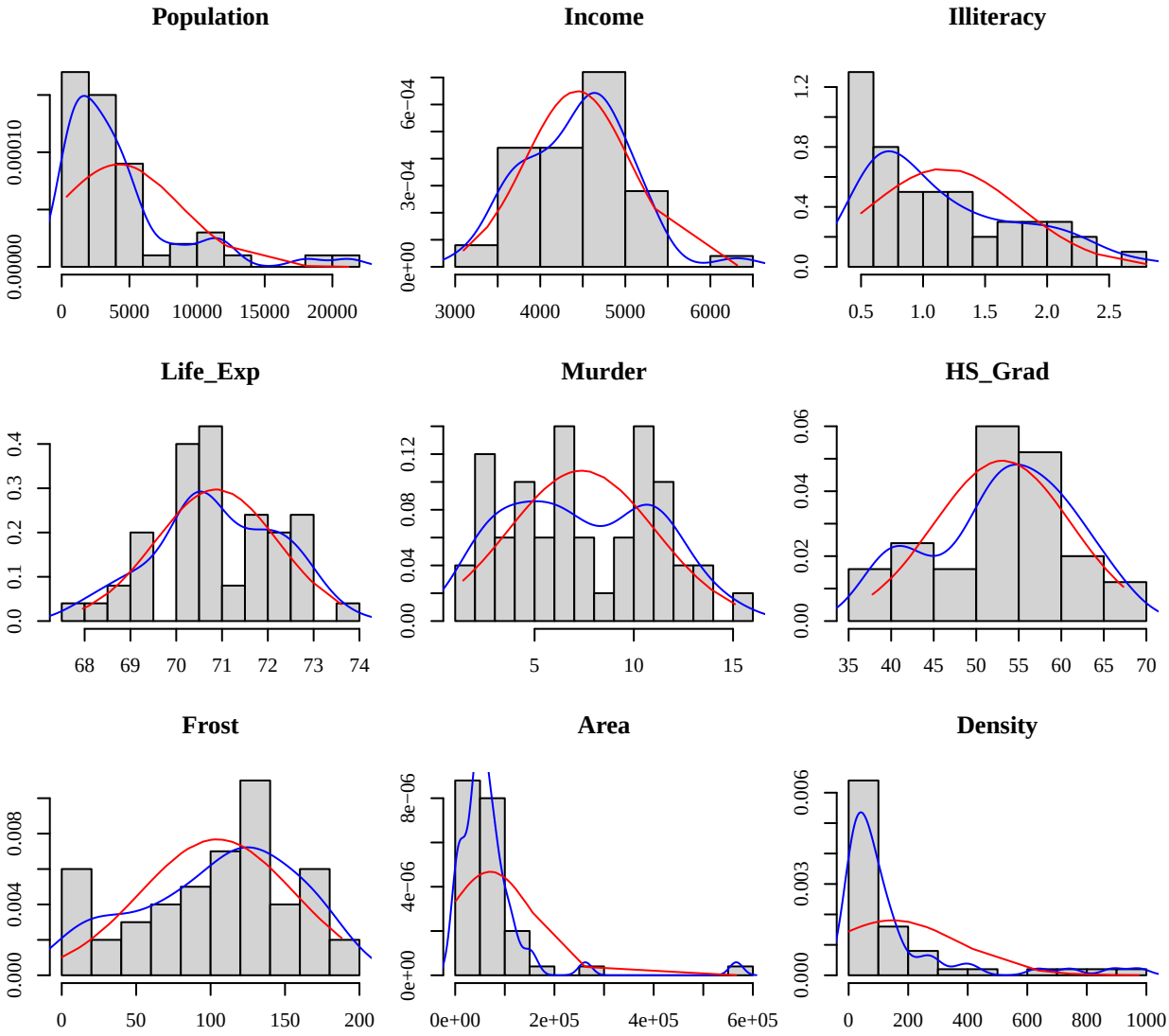
From the Q-Q plots we can observe that:

- all the values corresponding to the variable *Income* lie very close to the Q-Q line, except for the observation 2, which was previously identified as an univariate outlier;
- also the variable *Life_Exp* seems to have a distribution close to normal: the values are more spread out with respect to the ones corresponding to the variable *Income*, but they still are very close to the blue line (which represents the linear relationship between the sample and the theoretical quantiles);
- the variables *Murder*, *HS_Grad* and *Frost* have a very similiar behaviour: most of the points lie near the Q-Q line, but they have a thinner right tail and a heavier left tail. Note the absence of univariate outliers;
- also the variables *Population*, *Area* and *Density* have very similar shapes, which are pretty far from being linear. All of them have heavy tails, which is also caused by the presence of more than one outlier;

- the trajectory of the variable *Illiteracy* is very atypical, indeed on the left side of the plot we can observe that a consistent percentage of the points share the same values.

In conclusion, we can infer a gaussian behaviour only for the variables *Income* and *Life_Exp*.

We can draw the same conclusions by observing the histograms of the single variables. In the following plots the blue line represents the empirical density, while the red line the theoretical gaussian density (under sample mean and variance).



Another possible way to assess normality is by taking the Shapiro-Wilk test: if the returned p-value is less than the chosen significance level, i.e 0.05, we can reject the null hypothesis that the data are normally distributed. If the p-value is grater than the chosen significance level, we fail to reject the null hypothesis. By performing the Shapiro-Wilk test, we obtain the following results:

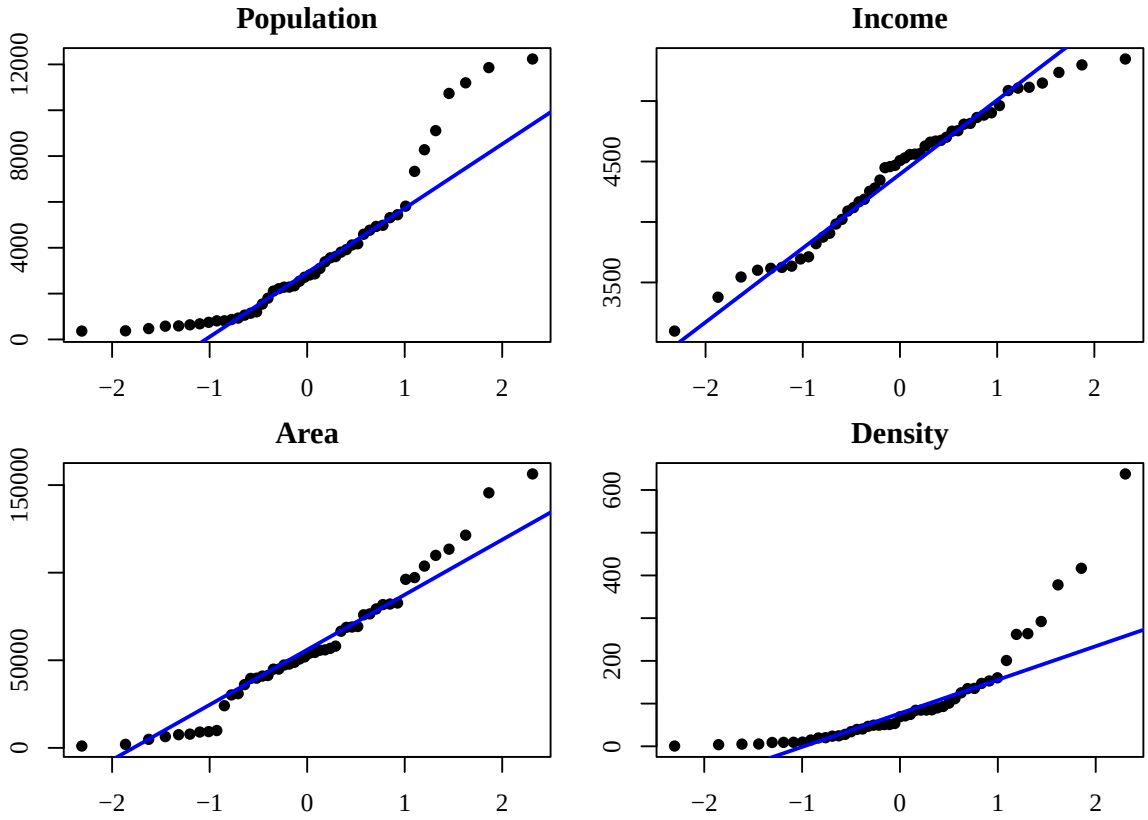
	Population	Income	Illiteracy	Life_Exp	Murder	HS_Grad	Frost	Area	Density
p.value	2e-07	0.4300105	0.0001396	0.4423285	0.0474463	0.0458156	0.0526747	0	0

As we can see, the variables *Income* and *Life_Exp* can be considered normally distributed, since their p-value

is grater than 0.05. Moreover, the p-values of the variables *Murder*, *HS_Grad* and *Frost* are really close to 0.05, but the only one for which the null hypothesis is not rejected is *Frost*. However, its p-value is ≈ 0.0527, thus it is reasonable to doubt its normality, taking also in account the corresponding Q-Q plot and histogram presented above.

Now, it may be of interest to see how the previous tests change if for each variable we remove the observations we identified as their respective outliers (where any). We report the adjusted Q-Q plots and the results of the Shapiro-Wilk test.

	Population	Income	Area	Density
p.value	1.39e-05	0.2493674	0.0600198	1e-07

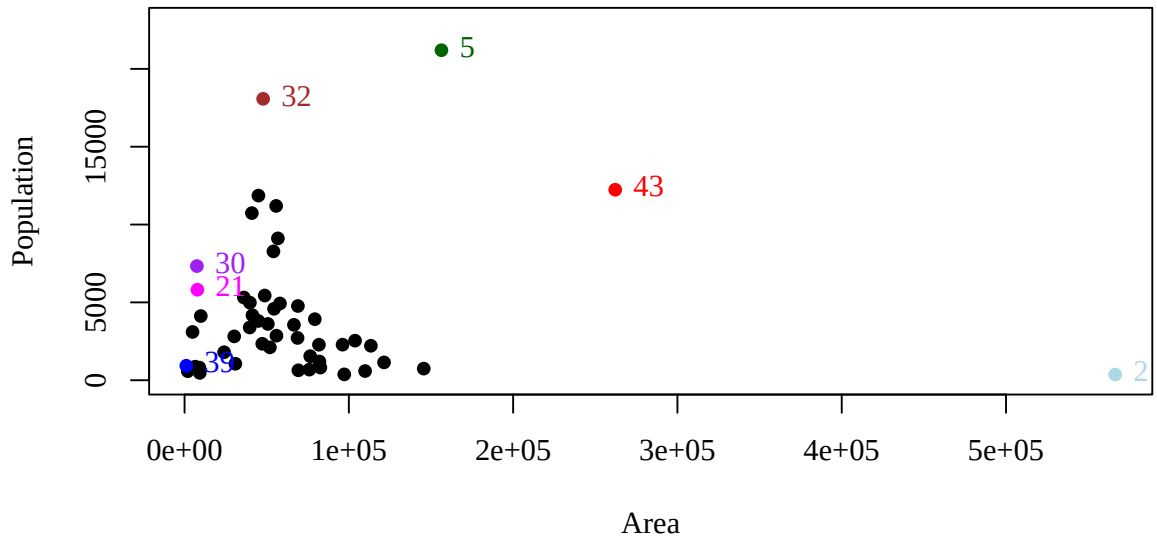


We can observe that:

- the variables *Population* and *Density* follow the same non-gaussian behaviour that we observed before; as a matter of fact, their p-values are almost negligible;
- as for the variable *Income*, we can still argue that its distribution is gaussian. However, from the Q-Q plot we can observe the presence of more pronounced tails. Its p-value is still greater than 0.05 yet a slightly lower than the previous one;
- finally, by removing its outliers, the variable *Area* seems to become gaussian: its p-value is very close to 0.05 but there is a huge difference between the current and the previous p-value.

1.5

We report the scatterplot of the variables *Area vs Population*, where have colored all the potential univariate outliers.



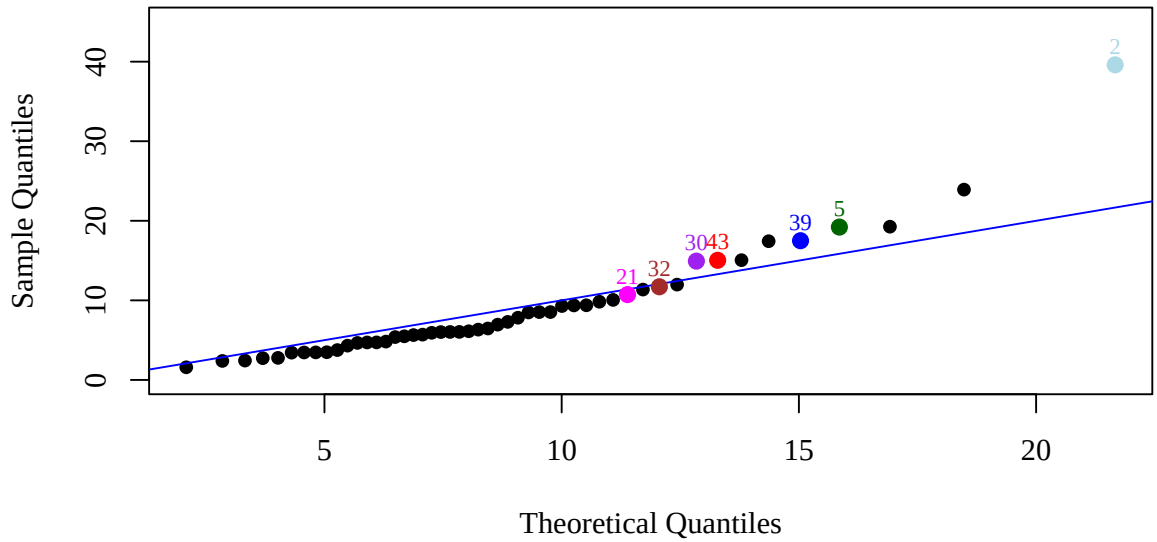
From the plot we can observe that all the mass is roughly concentrated in the rectangle $[0, 15\times10^6]\times[0, 15\times10^3]$. Furthermore, unlike the other univariate outliers, only the points corresponding to the observations $2, 5, 32$ and 43 seems to be really far from the others. Hence, we can identify them as bivariate outliers.

1.6

The squared Mahalanobis distance is a distance function that quantifies the gap between an observation and the sample mean, weighted by the inverse of the covariance matrix. If our variables are distributed as a 9-dimensional multivariate normal, then we will have

$$d \sim \sum_{i=1}^{9} \mathcal{N}(0,1)^2 \sim \chi^2_9.$$

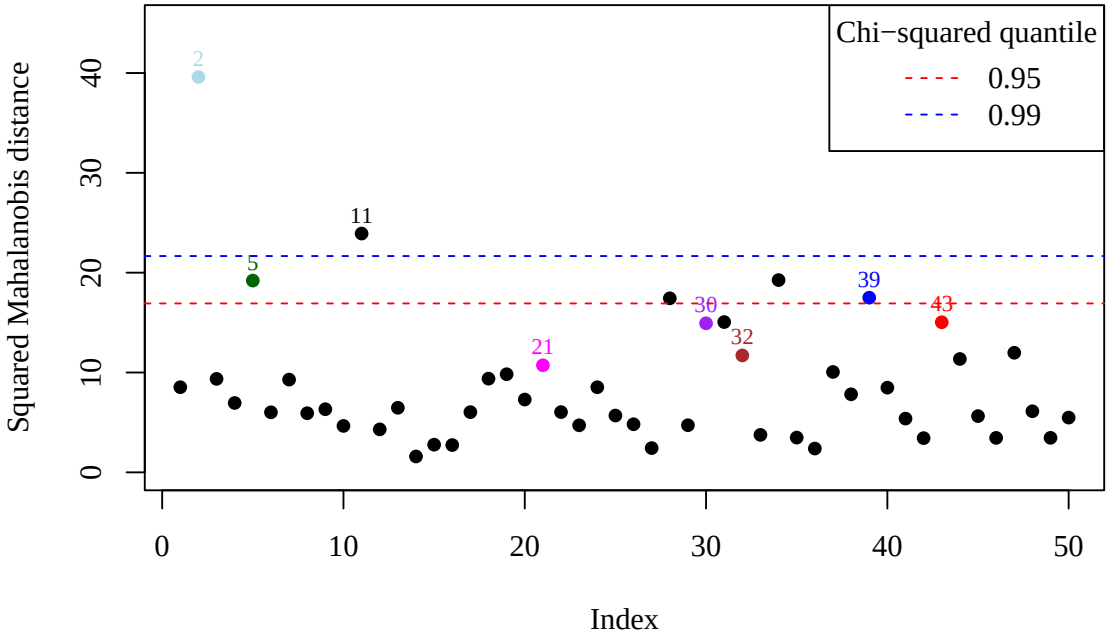
Hence, we can check the multivariate normality by looking at the Q-Q plot of the squared Mahalanobis distances *vs* a χ^2_9.



The Chi-sqared Q-Q plot of Mahalanobis distance shows that the majority of the points are close to the Q-Q line. The most evident exceptions are the point corresponding to the observation 2, which was previously detected as an univariate outlier, and the second to last point, which corresponds to the observation 11. Nevertheless, we can consider d χ^2_9-distributed, and, hence, say that our variables are jointly distributed as a multivariate gaussian.

1.7

In order to identify the multivariate outliers, we can plot the vector of the squared Mahalanobis distances and then add some threshold lines corresponding to different levels of the theoretical quantiles of a χ^2_9 (in particular, we used $\alpha_1 = 0.95$ and $\alpha_2 = \frac{(n-0.5)}{n} = 0.99$ with $n =$ `nrow(st)`).



The observations 2 and 11 are above the blue line (which corresponds to $\chi^2_9(\alpha_2)$), hence they can be confidently considered as multivariate outliers. This confirms what we previously noticed in the Chi-squared Q-Q plot. Note also that some observations lie in the strip delimited by the two lines. Nonetheless, we choose not to consider them as multivariate outliers, taking also in account what we observed in the previous point.
Finally, we can notice that the majority of the observations identified as univariate outliers cannot be considered as multivariate outliers with the only exception of the observation 2.

Exercise 2

2.1

First observe that Σ is invertible since $\det(\Sigma) = -2\rho^3 - 3\rho^2 + 1$ which is greater than 0, $\forall \rho \in \left(-1, \frac{1}{2}\right)$. We compute the inverse of Σ by exploiting the identity

$$\Sigma = (1+\rho)I - \rho aa^T \text{ with } a = (1,1,-1)$$

and applying the following theorem, which is known as the Neumann Series Theorem:

Theorem (Neumann Series)**.** *Let T be a linear mapping $T : \mathbb{R}^n \to \mathbb{R}^n$. If the series $\sum_{i=0}^{\infty} T^i$ converges, then $I - T$ is invertible and it holds*

$$(I-T)^{-1} = \sum_{i=0}^{\infty} T^i.$$

We first rewrite

$$\Sigma = (1+\rho)(I - caa^T) \text{ with } a = (1,1,-1) \text{ and } c = \frac{\rho}{1+\rho}.$$

Let $A \coloneqq I - caa^T$; it holds

$$\begin{aligned}
A^{-1} = (I - caa^T)^{-1} &= \sum_{i=0}^{\infty} (caa^T)^i = \\
&= \sum_{i=0}^{\infty} c^i (aa^T)^i = I + \sum_{i=1}^{\infty} c^i (\|a\|^2)^{i-1} aa^T = \\
&= I + caa^T \sum_{i=1}^{\infty} (c\|a\|^2)^{i-1} = I + caa^T \sum_{i=0}^{\infty} (c\|a\|^2)^i = \\
&= I + caa^T \frac{1}{1 - c\|a\|^2} = I + \frac{\rho}{1+\rho} \frac{1}{1 - 3\frac{\rho}{1+\rho}} aa^T = \\
&= I + \frac{\rho}{1-2\rho} aa^T.
\end{aligned}$$

Note that to pass from the third to the fourth line it is necessary to have $c\|a\|^2 < 1$ but since we are considering $\rho \in \left(-1, \frac{1}{2}\right)$ it is always true indeed

$$c\|a\|^2 < 1 \iff \frac{\rho}{1+\rho} 3 < 1 \iff 3\rho < 1 + \rho \iff \rho < \frac{1}{2}.$$

Thus

$$\Sigma^{-1} = (1+\rho)^{-1} A^{-1} = \frac{1}{1+\rho} \left(I + \frac{\rho}{1-2\rho} aa^T \right).$$

We can compute Σ^{-1} in many other ways; for example we can suppose that Σ^{-1} is of the same form Σ, i.e.

$$\Sigma^{-1} = yI + kaa^T.$$

and than find the values for y and k. We obtain:

$$\begin{aligned}
\Sigma\Sigma^{-1} &= ((1+\rho)I - \rho aa^T)(yI + kaa^T) = \\
&= (1+\rho)yI + (1+\rho)kaa^T - \rho y aa^T - k\rho aa^T aa^T = \\
&= (1+\rho)yI + ((1+\rho)k - \rho y - 3k\rho) aa^T.
\end{aligned}$$

This leads to the following system:

$$\begin{cases} (1+\rho)y = 1 \\ (1+\rho)k - \rho y - 3k\rho = 0 \end{cases} \iff \begin{cases} y = \frac{1}{(1+\rho)} \\ k + \rho k - \rho y - 3k\rho = 0 \end{cases}.$$

By solving the second equation we obtain

$$k - 2\rho k - \frac{\rho}{1+\rho} = 0 \iff k(1-2\rho) = \frac{\rho}{1+\rho} \iff k = \frac{\rho}{(1+\rho)(1-2\rho)}.$$

Hence

$$\Sigma^{-1} = \frac{1}{1+\rho} I + \frac{\rho}{(1+\rho)(1-2\rho)} aa^T = \frac{1}{1+\rho}\left(I + \frac{\rho}{1-2\rho} aa^T\right).$$

2.2

We find the eigenvalues of Σ by computing the roots of the characteristic polynomial $p(\lambda) = \det(\Sigma - \lambda I)$.

$$\begin{aligned}
\det(\Sigma - \lambda I) &= \det \begin{pmatrix} 1-\lambda & -\rho & \rho \\ -\rho & 1-\lambda & \rho \\ \rho & \rho & 1-\lambda \end{pmatrix} = \\
&= (1-\lambda)\det \begin{pmatrix} 1-\lambda & \rho \\ \rho & 1-\lambda \end{pmatrix} + \rho \det \begin{pmatrix} -\rho & \rho \\ \rho & 1-\lambda \end{pmatrix} + \rho \det \begin{pmatrix} -\rho & \rho \\ 1-\lambda & \rho \end{pmatrix} = \\
&= (1-\lambda)\left((1-\lambda)^2 - \rho^2\right) + \rho\left(-\rho(1-\lambda) - \rho^2\right) + \rho\left(-\rho^2 - \rho(1-\lambda)\right) = \\
&= (1-\lambda)(1-\lambda+\rho)(1-\lambda-\rho) - 2\rho^2(1-\lambda+\rho) = \\
&= (1-\lambda+\rho)\left((1-\lambda)(1-\lambda-\rho) - 2\rho^2\right) = \\
&= (1-\lambda+\rho)(1-\lambda-\rho-\lambda+\lambda^2+\lambda\rho-2\rho^2) = \\
&= (1-\lambda+\rho)\left(\lambda^2 + \lambda(\rho-2) - 2\rho^2 - \rho + 1\right)
\end{aligned}$$

Hence

$$\begin{aligned}
p(\lambda) = 0 &\iff 1-\lambda+\rho = 0 \text{ or } \lambda^2 + \lambda(\rho-2) - 2\rho^2 - \rho + 1 = 0 \\
&\iff \lambda = 1+\rho \text{ or } \lambda = \lambda_{1,2}
\end{aligned}$$

with $\lambda_{1,2}$ roots of $p(\lambda) = \lambda^2 + \lambda(\rho-2) - 2\rho^2 - \rho + 1$.

$$\begin{aligned}
\lambda_{1,2} &= \frac{-\rho + 2 \pm \sqrt{(\rho-2)^2 - 4(-2\rho^2 - \rho + 1)}}{2} = \\
&= \frac{-\rho + 2 \pm \sqrt{9\rho^2}}{2} = \\
&= \frac{-\rho + 2 \pm 3|\rho|}{2} = \\
&= (1+\rho, 1-2\rho).
\end{aligned}$$

Hence the eigenvalues (with multiplicity) are $\{1+\rho, 1+\rho, 1-2\rho\}$.
A faster way to find the spectrum (set of eigenvalues, meant with multiplicity) is reported below. We exploit some basic properties of the spectrum.
We denote with $\mathrm{Sp}(\Sigma)$ the spectrum of the matrix Σ (as a linear operator).

$$\begin{aligned}
\mathrm{Sp}(\Sigma) &= \mathrm{Sp}\left((1+\rho)\left(I - \frac{\rho}{1+\rho} aa^t\right)\right) = \\
&= (1+\rho)\,\mathrm{Sp}\left(I - \frac{\rho}{1+\rho} aa^t\right) = \\
&= (1+\rho)\left(1 - \frac{\rho}{1+\rho}\,\mathrm{Sp}\left(aa^t\right)\right).
\end{aligned}$$

Since $(aa^T)a = \|a\|^2 a$ and $\mathrm{rank}\left(aa^T\right) = 1$, it holds that

$$\mathrm{Sp}\left(aa^T\right) = \left\{0, 0, \|a\|^2\right\}.$$

Hence, recalling that the spectrum of a polynomial of an operator (here aa^T) is the polynomial of the spectrum

$$\begin{aligned}\mathrm{Sp}\,(\Sigma) &= (1+\rho)\left(1-\frac{\rho}{1+\rho}\{0,0,\|a\|^2\}\right) = \\ &= (1+\rho)\left\{1,1,1-3\frac{\rho}{1+\rho}\right\} = \\ &= \{1+\rho,1+\rho,1+\rho-3\rho\} = \\ &= \{1+\rho,1+\rho,1-2\rho\}\,,\end{aligned}$$

where the multiplications and translations of sets are meant componentwise.

2.3

First we need to write the eigenvalues of Σ in ascending order. We can distinguish the following two cases:

1. if $\rho \in \left[0, \frac{1}{2}\right)$, then $1+\rho \geq 1-2\rho$. This leads to

$$\begin{cases}\lambda_1 = 1+\rho \\ \lambda_2 = 1+\rho \\ \lambda_3 = 1-2\rho\end{cases}, \text{ with } \lambda_1 \geq \lambda_2 \geq \lambda_3.$$

   Now we find ρ such that the first two principal components (PCs) account for more than 80% of the total variation of Z.
   Since λ_i corresponds to the variance of the ith PC $\forall i \in \{1,2,3\}$ and the variation up to the kth PC corresponds to the sum of the first k eigenvalues, we just need to find ρ such that

$$\lambda_1 + \lambda_2 > 0.8(\lambda_1+\lambda_2+\lambda_3).$$

   By solving the inequality we get

$$2(1+\rho) > \frac{4}{5}3 \iff 1+\rho > \frac{6}{5} \iff \rho > \frac{1}{5}.$$

2. if $\rho \in (-1,0)$ then $1+\rho \leq 1-2\rho$. This leads to

$$\begin{cases}\lambda_1 = 1-2\rho \\ \lambda_2 = 1+\rho \\ \lambda_3 = 1+\rho\end{cases}, \text{ with } \lambda_1 \geq \lambda_2 \geq \lambda_3.$$

   By using the same argument we used in the previous point we obtain that ρ has to satisfy the following condition:

$$(1-2\rho)+(1+\rho) > \frac{4}{5}3 \iff 2-\rho > \frac{12}{5} \iff \rho < -\frac{2}{5}.$$

Hence, if $\rho \in \left[0, \frac{1}{2}\right)$, it must be $\rho > \frac{1}{5}$, while, if $\rho \in (-1,0)$, it must be $\rho < -\frac{2}{5}$.
Finally, $\forall \rho \in \left(-1, -\frac{2}{5}\right) \cup \left(\frac{1}{5}, \frac{1}{2}\right)$ PC1 and PC2 account for more than 80% of the total variation of Z.

2.4

In order to find the conditional distribution of $Y = (Y_1, Y_2)$ given $X = x$ we use the following result we have seen in class.

Proposition. *Let $X = (X_1, X_2) \sim \mathcal{N}_p(\mu, \Sigma)$ with*

$$\mu = \begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix}, \Sigma = \left(\begin{array}{c|c}\Sigma_{1,1} & \Sigma_{1,2} \\ \hline \Sigma_{1,2}^T & \Sigma_{2,2}\end{array}\right)$$

where the dimension of X_1 is $q < p$. Then, the conditional distribution of $X_2|X_1 = x$ is $\mathcal{N}_{p-q}\left(\tilde{\mu}, \tilde{\Sigma}\right)$ with

$$\begin{cases} \tilde{\mu} = \mu_2 + \Sigma_{1,2}^T \Sigma_{1,1}^{-1}(X_1 - \mu_1), \\ \tilde{\Sigma} = \Sigma_{2,2} - \Sigma_{1,2}^T \Sigma_{1,1}^{-1} \Sigma_{1,2} \end{cases}.$$

By applying the proposition to

$$Z = (X, Y_1, Y_2) \sim \mathcal{N}_3(\mu, \Sigma) \text{ with}$$

$$\mu = \mathbb{E}[Z] = \begin{pmatrix} \mathbb{E}[X] \\ \mathbb{E}\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \end{pmatrix} \coloneqq \begin{pmatrix} \mu_1 \\ \hline \mu_{2,1} \\ \mu_{2,2} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ and}$$

$$\Sigma = \left(\begin{array}{c|cc} \operatorname{Var}(X) & \operatorname{Cov}(X, Y_1) & \operatorname{Cov}(X, Y_2) \\ \hline \operatorname{Cov}(X, Y_1) & \multicolumn{2}{c}{\multirow{2}{*}{$\operatorname{Var}(Y)$}} \\ \operatorname{Cov}(X, Y_2) & & \end{array}\right) \coloneqq \left(\begin{array}{c|c} \Sigma_{1,1} & \Sigma_{1,2} \\ \hline \Sigma_{1,2}^T & \Sigma_{2,2} \end{array}\right),$$

we obtain $(Y_1, Y_2)|X = x \sim \mathcal{N}_2\left(\tilde{\mu}, \tilde{\Sigma}\right)$ with

$$\begin{cases} \tilde{\mu} = \mu_2 + \Sigma_{1,2}^T \Sigma_{1,1}^{-1}(x - \mu_1), \\ \tilde{\Sigma} = \Sigma_{2,2} - \Sigma_{1,2}^T \Sigma_{1,1}^{-1} \Sigma_{1,2} \end{cases}.$$

By substituting what we found we get

$$\begin{aligned} \tilde{\mu} &= \mathbb{E}\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} + \begin{pmatrix} \operatorname{Cov}(X, Y_1) \\ \operatorname{Cov}(X, Y_2) \end{pmatrix} \frac{x - \mathbb{E}[X]}{\operatorname{Var}(X)} = \\ &= \begin{pmatrix} 0 \\ 2 \end{pmatrix} + \begin{pmatrix} -\rho \\ \rho \end{pmatrix} \frac{x - 1}{1} = \begin{pmatrix} 0 - \rho(x-1) \\ 2 + \rho(x-1) \end{pmatrix} = \\ &= \begin{pmatrix} -\rho(x-1) \\ 2 + \rho(x-1) \end{pmatrix} \end{aligned}$$

and

$$\begin{aligned} \tilde{\Sigma} &= \operatorname{Var}(Y) - \frac{1}{\operatorname{Var}(X)} \begin{pmatrix} \operatorname{Cov}(X, Y_1) \\ \operatorname{Cov}(X, Y_2) \end{pmatrix} \begin{pmatrix} \operatorname{Cov}(X, Y_1) & \operatorname{Cov}(X, Y_2) \end{pmatrix} = \\ &= \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} - \begin{pmatrix} -\rho \\ \rho \end{pmatrix} \begin{pmatrix} -\rho & \rho \end{pmatrix} = \\ &= \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} - \rho^2 \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} = \\ &= \begin{pmatrix} 1 - \rho^2 & \rho + \rho^2 \\ \rho + \rho^2 & 1 - \rho^2 \end{pmatrix}. \end{aligned}$$

Hence, it holds

$$(Y_1, Y_2)|X = x \sim \mathcal{N}_2\left(\begin{pmatrix} -\rho(x-1) \\ 2 + \rho(x-1) \end{pmatrix}, \begin{pmatrix} 1 - \rho^2 & \rho + \rho^2 \\ \rho + \rho^2 & 1 - \rho^2 \end{pmatrix}\right).$$

2.5

Let $\rho = \frac{1}{5}$. Then according to what we found in the previous point we obtain

$$\begin{cases} \mu_Y \coloneqq \begin{pmatrix} -\rho(x-1) \\ 2 + \rho(x-1) \end{pmatrix}\Big|_{x=0, \rho=1/5} \\ \Sigma_Y \coloneqq \begin{pmatrix} 1 - \rho^2 & \rho + \rho^2 \\ \rho + \rho^2 & 1 - \rho^2 \end{pmatrix}\Big|_{\rho=1/5} \end{cases} \iff \begin{cases} \mu_Y = \frac{1}{5}\begin{pmatrix} 1 \\ 9 \end{pmatrix} \\ \Sigma_Y = \frac{6}{25}\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix} \end{cases}.$$

Since Σ_Y is invertible ($\det(\Sigma_Y) = \frac{6}{25}15 > 0$) the random variable $(Y-\mu_Y)^T\Sigma_Y^{-1}(Y-\mu_Y)$ is well defined. Moreover, since $Y \sim \mathcal{N}_2(\mu_Y, \Sigma_Y)$, then

$$(Y-\mu_Y)^T\Sigma_Y^{-1}(Y-\mu_Y) \sim \chi^2_2.$$

Considering the 2-dimensional space $y = (y_1, y_2)$ and letting $c \in \mathbb{R}^+$ we have that

$$(y-\mu_Y)^T\Sigma_Y^{-1}(y-\mu_Y) = c^2$$

defines a contour line of the density function of Y which is an ellipse that contains the following percentage of the mass:

$$\mathbb{P}\left((Y-\mu_Y)^T\Sigma_Y^{-1}(Y-\mu_Y) \leq c^2\right).$$

By imposing this probability to 0.95, since $(Y-\mu_Y)^T\Sigma_Y^{-1}(Y-\mu_Y) \sim \chi^2_2$, we get that

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

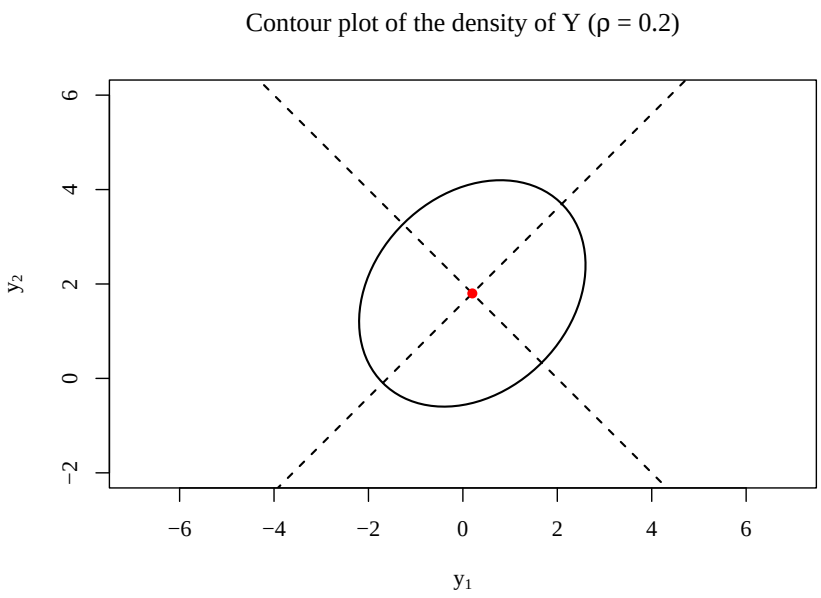
```
c = sqrt(qchisq(0.95, df = 2))
```

An explicit computation reveals

$$c = 2.4477468.$$

This ellipse is centered in μ_Y, its axes have length $c\sqrt{\lambda_1}$, $c\sqrt{\lambda_2}$ and directions e_1, e_2, where (λ_1, e_1) and (λ_2, e_2) are the eigenpairs of the matrix Σ_y.

```
rho = 0.2
mu_y = 1 / 5 * c(1, 9)
sigma_y = 6 / 25 * matrix(c(4, 1, 1, 4), nrow = 2)
eig = eigen(sigma_y, symmetric = T)
```



Exercise 3

3.1

First of all we divide each variable by `weight` in order to equalize out the different types of servings of each food.

```
nutritional = read.table("data/nutritional.txt")
nutritional = nutritional[, -6] / nutritional[, 6]
head(round(nutritional, 3))
```

fat	food.energy	carbohydrates	protein	cholesterol	saturated.fat
0.133	1.667	0.133	0.000	0.133	0.013
0.375	3.750	0.125	0.000	0.250	0.062
0.035	3.175	0.776	0.141	0.000	0.004
0.000	3.175	0.776	0.106	0.000	0.004
0.000	0.303	0.030	0.030	0.000	0.000
0.035	2.469	0.741	0.141	0.000	0.004

After the standardization of the dataset, carried out with the command `scale`, we perform the Principal Components Analysis.

```
nt = scale(nutritional)
nt_pca = prcomp(nt)
as.data.frame(nt_pca$rotation)
```

	PC1	PC2	PC3	PC4	PC5	PC6
fat	-0.5572394	0.0987008	-0.2750890	0.1304014	-0.4546980	0.6166958
food.energy	-0.5361507	0.3567665	0.1370762	0.0745468	-0.2729547	-0.6974301
carbohydrates	0.0245536	0.6716316	0.5684779	-0.2861681	0.1568663	0.3444441
protein	-0.2352271	-0.3738430	0.6388770	0.5991035	0.1538186	0.1189985
cholesterol	-0.2525045	-0.5213044	0.3256120	-0.7170962	-0.2102965	-0.0029044
saturated.fat	-0.5313507	-0.0192336	-0.2611169	-0.1496468	0.7913619	0.0216043

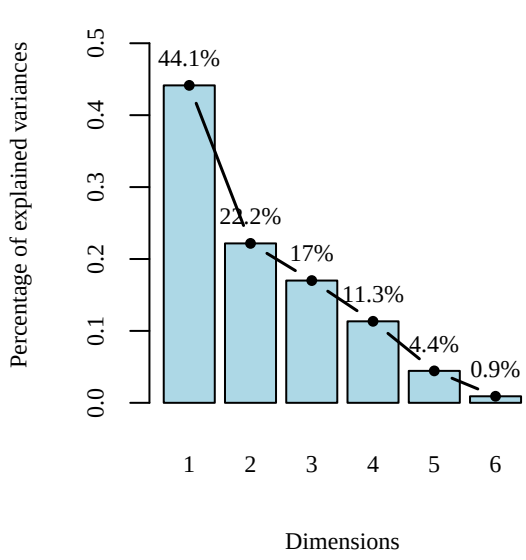
3.2

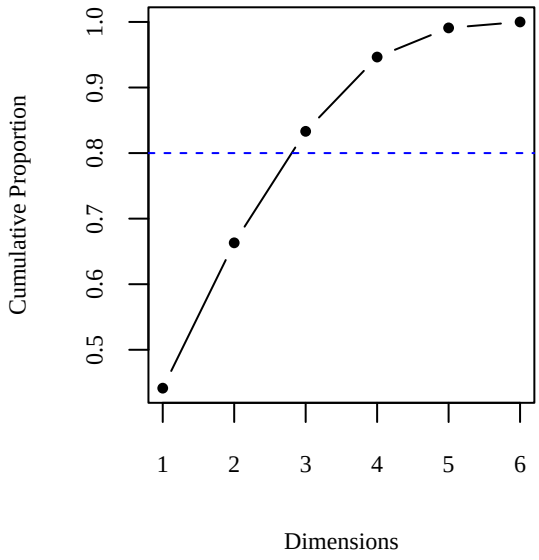
In order to decide how many components to retain we first observe the proportions and the cumulative proportions of explained variances.

```
nt_sum = as.data.frame(summary(nt_pca)$importance)[-1, ]
nt_sum
```

	PC1	PC2	PC3	PC4	PC5	PC6
Proportion of Variance	0.44143	0.22169	0.17002	0.11334	0.04442	0.00909
Cumulative Proportion	0.44143	0.66312	0.83314	0.94649	0.99091	1.00000

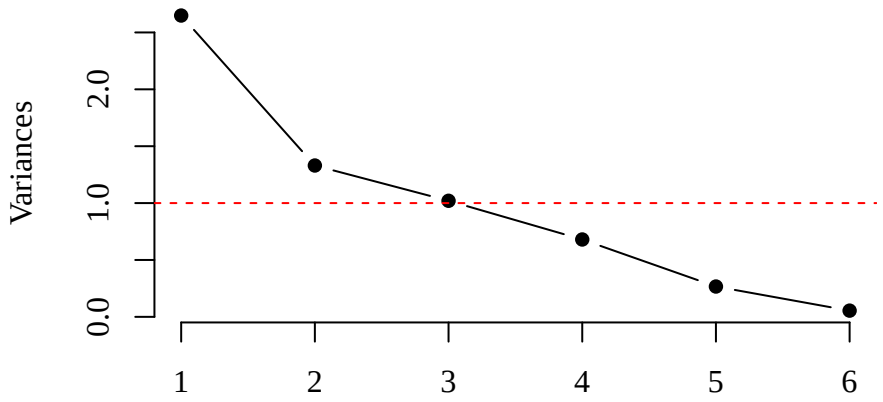
For a more immediate visualization we can plot the values reported above.





The first three principal components explain the 83.3% of the variability of the data, and if we add a fourth component we arrive near to the 95%, which exceeds our requirements. Hence, it seems reasonable to retain only the first three PCs.

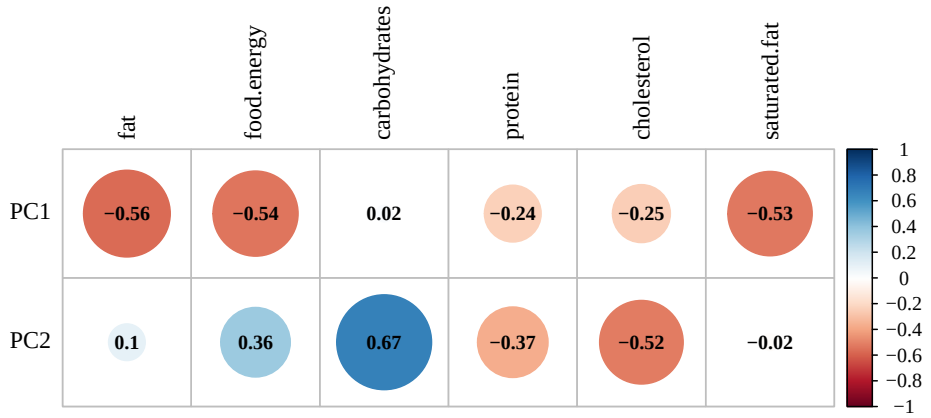
A robustness check for this proposed number of PCs can be performed by looking at the eigenvalues associated to each component. In order to identify the ones which capture most of the variance of the original variables, we want to identify those that are sized above average. As the original data was previously standardized, this means to look for eigenvalues greater than 1.



The screeplot is not so easy to interpret. Due to the absence of an evident *elbow* in the curve, it would not be a useful selection criterium. In conclusion, as only the eigenvalues corresponding to the first three PCs are greater than 1, we can confirm the choice of retaining only the first three PCs.

3.3

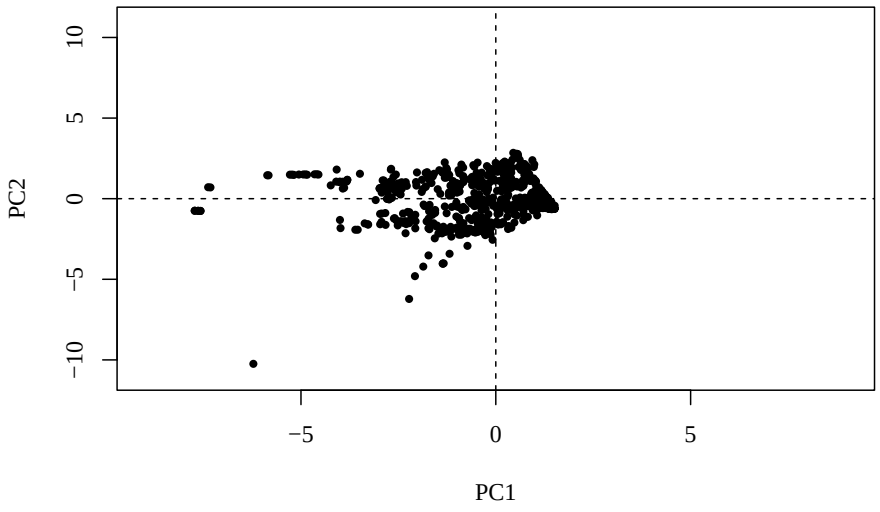
In order to give an interpretation to the first two PCs we look at their loadings (we report them in the following representation).



PC1. Except for the loading associated to the variable *carbohydrates*, which is almost negligible, the other loadings are all negative and lie in the interval $[-0.24, -1]$. The most influencial variables are *fat*, *satured.fat* and *food.energy*, which are related to how much a food is dietetic (or not). Since these loadings are negative, we choose to interpret the first PC as a measure of the dietary power of a certain food item.

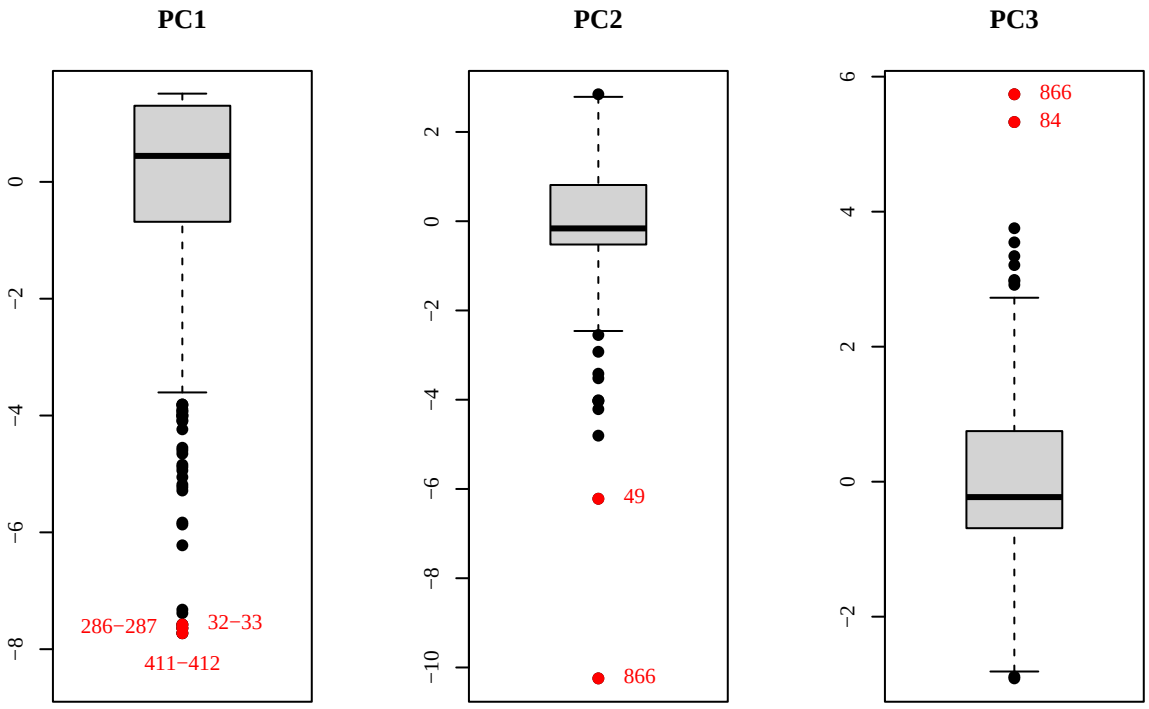
PC2. The loadings of the second PC are more complicated to analyse. The most influencing in a positive sense are the ones corresponding to the variables *carbohydrates* and *food.energy*, while the negative ones are *protein* and *colhesterol*. Hence, PC2 identifies how intense in carbohydrates a serving is, and at the same time low in cholesterol. A possible interpretation could be that the component is an indicator of how intensly a food is made of cereals.

Here we report the scatterplot of our dataset projected on the plane PC1 *vs* PC2.



3.4

In order to identify univariate outliers, we display the boxplots with respect to the first three principal components.



In the figure above we can spot a consistent amount of outliers for each principal component. We identified the most extreme (which are displayed in red) by scaling our PCs and then comparing them with some quantiles (with levels over 0.9999) of a $\mathcal{N}(0,1)$. It is remarkable that some outliers have multiplicity equal to 2: note that in the first boxplot the outliers are actually six.

We now want to measure how these outliers score in the original variables. We build a dataframe with *min*, *mean* and *max* of each variable.

	fat	food.energy	carbohydrates	protein	cholesterol	saturated.fat
min	-0.585	-1.164	-0.954	-0.778	-0.379	-0.562
mean	0.000	0.000	0.000	0.000	0.000	0.000
max	4.583	3.498	3.053	8.753	18.170	7.107

We now print the original dataset restricted to the outliers.

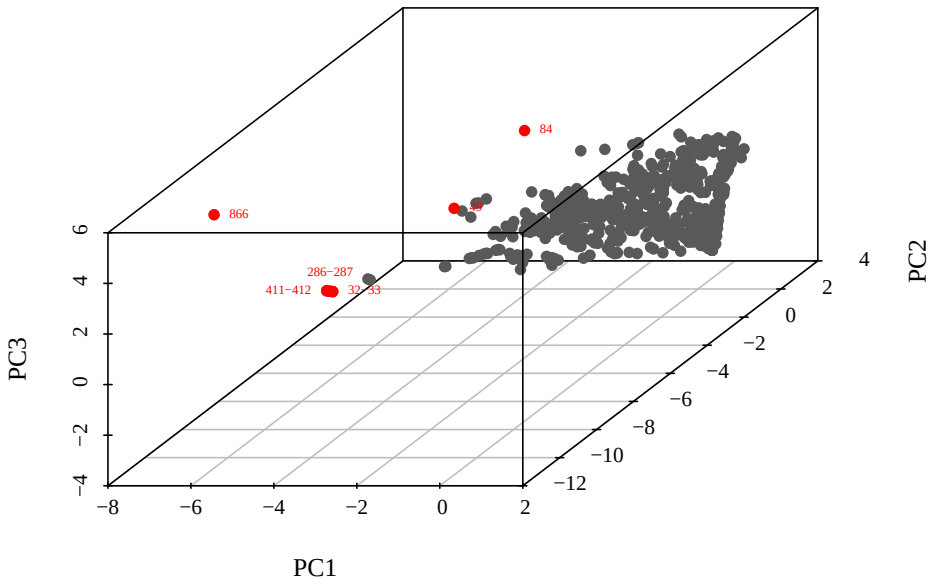
	fat	food.energy	carbohydrates	protein	cholesterol	saturated.fat
32	3.550	2.452	-0.954	-0.778	2.878	6.999
33	3.550	2.452	-0.954	-0.778	2.878	6.999
286	3.476	2.526	-0.954	-0.778	2.899	7.107
287	3.476	2.526	-0.954	-0.778	2.899	7.107
411	3.623	2.539	-0.954	-0.679	2.857	7.079
412	3.623	2.539	-0.954	-0.679	2.857	7.079
49	-0.327	-0.389	-0.954	2.002	8.948	-0.260
866	0.935	0.659	-0.954	1.184	18.170	0.861
84	-0.585	0.681	-0.954	8.753	-0.379	-0.562

We can observe that:

- as for the observations $32-33$, $286-287$, $411-412$, they score very low on *carbohydrates* and *protein*. However, given how low these two variables load on the first PC, this would not be enough to explain their outlier behaviour. The latter is instead due to their high values in *food.energy*, *fat* and *saturated.fat*. If we were to pick only two, they would be these last two;
- as for the observations 49 and 866, they score extremely high in *cholesterol* and attain the minimum in *carbohydrates*;
- finally, the observation 84 reaches the extremes in almost all variables but *food.energy*. If we were to choose only two of them, any random choice would be good.

3.5

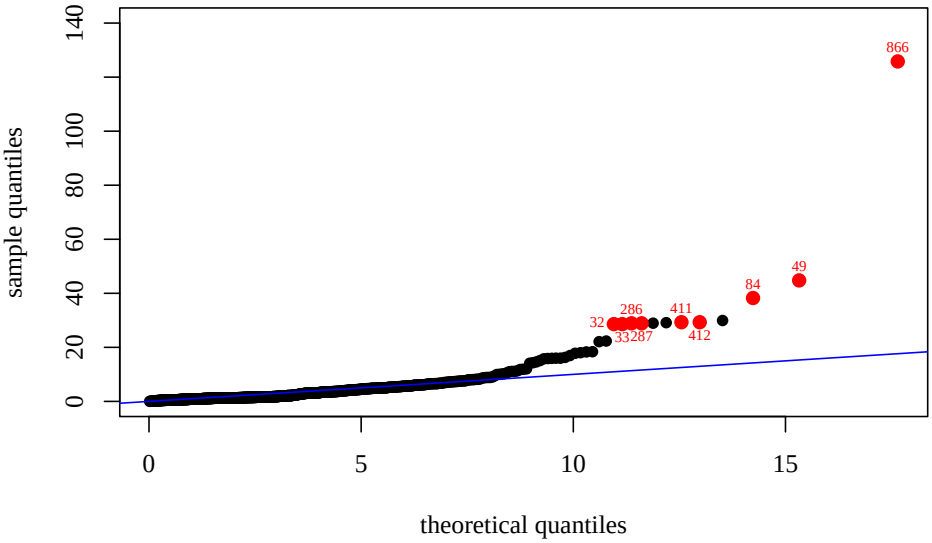
In the figure below we report a 3D scatterplot of the first three principal components.



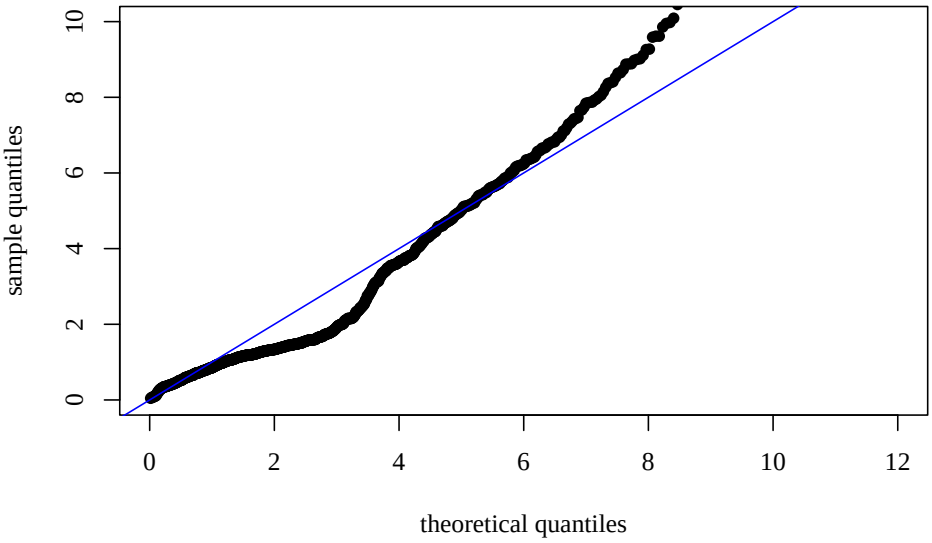
The outliers found in the previous point are labeled and highlighted in red. The plot shows that the joint distribution of these principal components has high variance. Nevertheless, the red points seem to be significantly far from the cloud. We hypotize that they could be multivariate outliers, as it will be discussed in more detail in the following.

3.6

We investigate multivariate normality through the first three principal components by computing the Mahalanobis distances, and then comparing them with the quantiles of a χ^2_3 with a Q-Q plot.



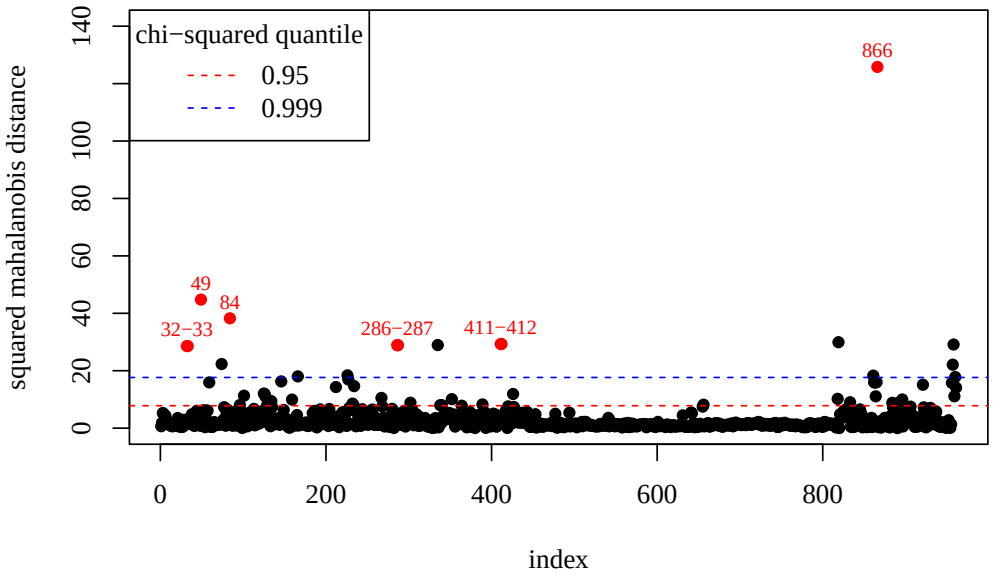
The approximation is rather good for the observations that have a low score in the Mahalanobis sense. These points also account for a large part of the overall mass. However, the empirical distribution appears to have a heavy right tail. Hence, the normality assumption is not plausible.
To corroborate this thesis, we display the same plot after removing the outliers and zooming in on the part that seems to fit best.



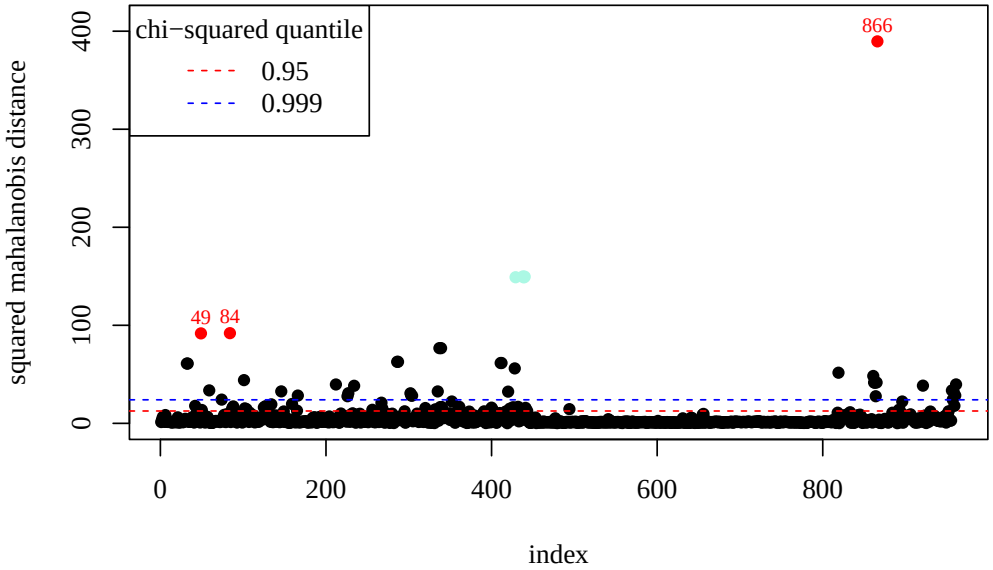
This further advocates against normality. Indeed, what at first sight seems an alignment to the Q-Q line, from a closer perspective reveals a mismatch also for smaller values of the Mahalanobis distance.

3.7

The Mahalanobis distance is not only of great use to asses the normality of the sample. It can also be employed to detect multivariate outliers. The following plot shows the squared Mahalanobis distances. We also displayed the quantiles of a χ^2_3 for the levels $\alpha = 0.95$ and $\alpha = 0.99$.



We observe that all the extreme univariate outliers we previously identified turned out to be also multivariate outliers indeed (despite not being the only ones). If we were to select only a handful of them, we would choose the most extreme ones, namely being observations 866, 49 and 84.
It can be of interest to asses whether these outliers were also multivariate outliers with respect to the original variables. Therefore, we also plot the corresponding squared Mahalanobis distances, with the quantiles of a χ^2_6 of levels $\alpha = 0.95$ and $\alpha = 0.99$.



Not surprisingly, these three outliers are indeed multivariate outliers also for the joint distribution of the original variables. It is worth to remark that also observations 429, 438, 439, 440 (marked with •) turn out to be multivariate outliers for the original variables. However, they are not so with respect to the three principal components.