Problem Set 1

Matteo Allegrini (423094)
Leonardo Federico De Blasio (40293)
Simone Maria Gervasoni (115376)
Alice Leto (232032)
Marinella Nigro (312312)

Contents

Exercise 1

For $i = 1, \dots, n$, let Y_i be i.i.d. random variables taking values in $\{1, 2, \dots, p, p+1\}$ with probabilities $\pi_1, \dots, \pi_p, \pi_{p+1} > 0$, $\sum_{j=1}^{p+1} \pi_j = 1$. If we code Y_i via the one-hot vector $Y_i = (Y_{i1}, \dots, Y_{i,p+1})$ where $Y_{ij} = 1$ when $Y_i = j$, then $\sum_{i=1}^{n} Y_i$ has multinomial distribution, Multinomial$(n, \pi_1, \dots, \pi_{p+1})$, and the multivariate Central Limit Theorem (CLT) implies

$$\sqrt{n}(\hat{\pi} - \pi) \xrightarrow{d} \mathcal{N}_{p+1}(0, \mathrm{diag}(\pi) - \pi\pi^T)$$

for $\hat{\pi} = n^{-1} \sum_{i=1}^{n} Y_i$ and $\pi = (\pi_1, \dots, \pi_{p+1})$. Assume n is sufficiently large so that $\sqrt{n}(\hat{\pi} - \pi)$ is normally distributed according to the CLT above and let X be the vector of the first p coordinates.

Point 1

What is the distribution of X? Justify your answer.

Let $Y_i \in \{1, 2, \dots, p\}$ $i = 1, \dots, n$ with probability $\pi_1, \dots, \pi_p, \pi_{p+1} > 0$ and $\sum_{j=1}^{p+1} \pi_j = 1$; i.e.

$$Y_i = \begin{cases} 1 & \text{with probability } \pi_1 \\ 2 & \text{with probability } \pi_2 \\ & \vdots \\ p+1 & \text{with probability } \pi_{p+1} \end{cases}$$

The one-hot vector $Y_i = (Y_{i1}, \dots, Y_{ip+1})$ is such that $Y_i = 1$ when $Y_i = j$. So Y_i codified as one hot vector is distributed as $Y_i \sim \text{Multinomial}(1, \pi_1, \dots, \pi_{p+1})$. Further $\sum_{i=1}^{n} Y_i$ has multinational distribution $\sum_{i=1}^{n} Y_i \sim \text{Multinomial}(n, \pi_1, \dots, \pi_{p+1})$.

By the CLT we can say

$$(X_1, \dots, X_{p+1}) \sim \mathcal{N}_{p+1}(0, \mathrm{diag}(\pi) - \pi\pi^T).$$

Where the covariance matrix is:

$$\begin{aligned} \Omega &= \begin{pmatrix} \pi_1 & & & \\ & \pi_2 & & \\ & & \ddots & \\ & & & \pi_{p+1} \end{pmatrix} - \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_{p+1} \end{pmatrix} \begin{pmatrix} \pi_1 & \dots & \pi_{p+1} \end{pmatrix} \\ &= \begin{pmatrix} \pi_1 & & & \\ & \pi_2 & & \\ & & \ddots & \\ & & & \pi_{p+1} \end{pmatrix} - \begin{pmatrix} \pi_1^2 & \pi_1\pi_2 & \dots & \pi_1\pi_{p+1} \\ \pi_2\pi_1 & \pi_2^2 & \dots & \pi_2\pi_{p+1} \\ \vdots & & \ddots & \\ \pi_{p+1}\pi_1 & & & \pi_{p+1}^2 \end{pmatrix} \\ \Omega &= \left(\begin{array}{ccc|c} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \dots & -\pi_1\pi_{p+1} \\ -\pi_2\pi_1 & \pi_2(1-\pi_2) & \dots & -\pi_2\pi_{p+1} \\ \vdots & \vdots & & \vdots \\ \hline -\pi_{p+1}\pi_1 & -\pi_{p+1}\pi_2 & \dots & \pi_{p+1}(1-\pi_{p+1}) \end{array}\right) \end{aligned}$$

If we take the first p components of a Gaussian vector of dimension p+1 , it is still a Gaussian vectorr of dimension p in fact:

$$(X_1, \dots, X_{p+1}) \sim \mathcal{N}_{p+1}(0, \Omega)$$

- The new mean vector $\underline{0}_p$ is the vector of dimension $p < p+1$.
- The element $\Omega_{p+1,p+1} = \underbrace{\pi_{p+1}}_{>0} \underbrace{(1 - \pi_{p+1})}_{>^{(1)}\ 0} > 0$

(1) $1 - \pi_{p+1} > 0 \iff \pi_{p+1} < 1$ and this is always true because π_{p+1} is a probability that is defined $(0, 1)$.

So we can say thatthe marginal distribution of the Gaussian vector of dimension p+1 is still a Gaussian:

$$(X_1, \ldots, X_p) \sim \mathcal{N}_p(0, \Sigma)$$

where

$$\Sigma := \Omega_{p\times p} = \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & \ldots & -\pi_1\pi_p \\ -\pi_2\pi_1 & \pi_2(1-\pi_2) & \ldots & -\pi_2\pi_p \\ \vdots & & \ddots & \\ -\pi_p\pi_1 & -\pi_p\pi_2 & \ldots & \pi_p(1-\pi_p) \end{pmatrix}$$

Point 2

Let Σ be the $p \times p$ covariance matrix of X. Find the inverse of Σ.

We want to find the inverse of the covariance matrix:

$$\Sigma = \begin{pmatrix} \pi_1^2 & -\pi_1\pi_2 & \ldots & -\pi_1\pi_p \\ -\pi_2\pi_1 & \pi_2^2 & \ldots & -\pi_2\pi_p \\ \vdots & \vdots & \ddots & \vdots \\ -\pi_p\pi_1 & -\pi_p\pi_2 & \ldots & \pi_p^2 \end{pmatrix}$$

First of all we need to check if the matrix is invertible, so we need to prove that $det(\Sigma) \neq 0$.
Since evaluating the determinant is not quite easy, we find a **Lemma** that give us an equivalent condition for the invertibility.

Sherman Morrison Formula:

Suppose $A \in \mathbb{R}^{n\times n}$ invertible and $u, v \in \mathbb{R}$ are column vectors. $A + uv^T$ is invertible $\iff 1 + v^T A u \neq 0$. In this case:

$$(A + uv^T) = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}. \tag{1}$$

In our case we have:

$$\Sigma = diag(\pi_1, \ldots, \pi_p) - \pi\pi^T = A + uv^T$$

where:

- $A = \text{diag}(\pi_1, \ldots, \pi_p) \in \mathbb{R}^{p\times p}$
- $u = \begin{pmatrix} -\pi_1 \\ -\pi_2 \\ \vdots \\ -\pi_p \end{pmatrix}$
- $v^T = \begin{pmatrix} \pi_1 & \ldots & \pi_p \end{pmatrix}$

Remark: in our case there is a minus so to be coherent with the notation of the formula (1) we take the minus inside the vector u.

Is A invertible?

$$|A| = \left| \begin{pmatrix} \pi_1 & & \\ & \ddots & \\ & & \pi_p \end{pmatrix} \right| = \pi_1\pi_2 \ldots \pi_p > 0$$

so it is invertible beacuse the determinant is bigger than 0

To apply the **Lemma** we check that $1 + v^T A^{-1} u \neq 0$.

Since A is a diagonal matrix we know that the inverse is as such:

$$A^{-1} = \begin{pmatrix} \frac{1}{\pi_1} & & \\ & \ddots & \\ & & \frac{1}{\pi_p} \end{pmatrix}$$

now we calculate :

$$1 + v^T A^{-1} u = 1 + \begin{pmatrix} \pi_1 & \dots & \pi_p \end{pmatrix} \begin{pmatrix} \frac{1}{\pi_1} & & \\ & \ddots & \\ & & \frac{1}{\pi_p} \end{pmatrix} \begin{pmatrix} -\pi_1 \\ \vdots \\ -\pi_p \end{pmatrix} = 1 + \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} -\pi_1 \\ \vdots \\ -\pi_p \end{pmatrix} =$$

$$= 1 - (\pi_1 + \dots + \pi_p) \neq 0 \Leftrightarrow \pi_1 + \dots + \pi_p \neq^{(2)} 1$$

(2) this is always true because by hypothesis $\sum_{i=1}^{p+1} \pi_i = 1$ and $\pi_i > 0$ so $\sum_{i=1}^{p} \pi_i < 1$ because π_{p+1} is missing.

The hypothesis of the **Lemma** are verified now we need to find the inverse:

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{\underbrace{1 + v^t A^{-1} u}_{1-(\pi_1+\dots+\pi_p)=\pi_{p+1}}}.$$

The numerator of the fraction is the following:

$$A^{-1} u v^T A^{-1} = \underbrace{\begin{pmatrix} \frac{1}{\pi_1} & & \\ & \ddots & \\ & & \frac{1}{\pi_p} \end{pmatrix}}_{p \times p} \underbrace{\begin{pmatrix} -\pi_1 \\ \vdots \\ -\pi_p \end{pmatrix}}_{p \times 1} \begin{pmatrix} \pi_1 & \dots & \pi_p \end{pmatrix} \begin{pmatrix} \frac{1}{\pi_1} & & \\ & \ddots & \\ & & \frac{1}{\pi_p} \end{pmatrix}$$

$$= \begin{pmatrix} -1 \\ -1 \\ \vdots \\ -1 \end{pmatrix} \begin{pmatrix} \pi_1 & \dots & \pi_p \end{pmatrix} \begin{pmatrix} \frac{1}{\pi_1} & & \\ & \ddots & \\ & & \frac{1}{\pi_p} \end{pmatrix}$$

$$= \begin{pmatrix} -\pi_1 & \dots & -\pi_p \\ -\pi_1 & \dots & -\pi_p \\ \vdots & & \vdots \\ -\pi_1 & \dots & -\pi_p \end{pmatrix} \begin{pmatrix} \frac{1}{\pi_1} & & \\ & \ddots & \\ & & \frac{1}{\pi_p} \end{pmatrix} = \underbrace{\begin{pmatrix} -1 & \dots & -1 \\ -1 & \dots & -1 \\ \vdots & & \vdots \\ -1 & \dots & -1 \end{pmatrix}}_{p \times p}$$

Now we can substitue the quantity above to find the inverse of the matrix Σ:

$$\begin{aligned}
\Sigma^{-1} &= A^{-1} - \begin{pmatrix} -\frac{1}{\pi_{p+1}} & \cdots & -\frac{1}{\pi_{p+1}} \\ \vdots & \ddots & \vdots \\ -\frac{1}{\pi_{p+1}} & \cdots & -\frac{1}{\pi_{p+1}} \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{\pi_1} & & \\ & \ddots & \\ & & \frac{1}{\pi_p} \end{pmatrix} + \begin{pmatrix} \frac{1}{\pi_{p+1}} & \cdots & \frac{1}{\pi_{p+1}} \\ \vdots & \ddots & \vdots \\ \frac{1}{\pi_{p+1}} & \cdots & \frac{1}{\pi_{p+1}} \end{pmatrix} \\
&= \begin{pmatrix} \frac{1}{\pi_1} + \frac{1}{\pi_{p+1}} & \frac{1}{\pi_{p+1}} & \cdots & \frac{1}{\pi_{p+1}} \\ \frac{1}{\pi_{p+1}} & \frac{1}{\pi_2} + \frac{1}{\pi_{p+1}} & \cdots & \frac{1}{\pi_{p+1}} \\ \vdots & \vdots & & \vdots \\ \frac{1}{\pi_{p+1}} & \frac{1}{\pi_{p+1}} & \cdots & \frac{1}{\pi_p} + \frac{1}{\pi_{p+1}} \end{pmatrix}
\end{aligned}$$

Point 3

Let $\pi = (\pi_0, \ldots, \pi_0, 1 - p\pi_0)$ for some $0 < \pi_0 < 1/p$. Find the eigenvalues of Σ. How large should p be such that the proportion of variance explained by the last (population) principal component account for less than 20% of total variation of X?

3.1 Find the eigenvalues Σ

$$\begin{aligned}
\Sigma &= diag(\lambda_1, \ldots, \lambda_n) - \pi_0 \pi^T \\
&= \pi_0 \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix} + \pi_0^2 \underbrace{\begin{pmatrix} -1 & -1 & \ldots & -1 \\ -1 & -1 & \ldots & -1 \\ \vdots & \vdots & \vdots & \vdots \\ -1 & -1 & \ldots & -1 \end{pmatrix}}_{J_p} \\
&= \pi_0 I_p + \pi_0^2 J_p \\
&= \pi_0 I_p - \pi_0^2 A
\end{aligned}$$

where $A = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \ldots & 1 \end{pmatrix}$.

At first we want to find the eigenvalues of the matrix A. We observe that:

$$Av = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \ldots & 1 \end{pmatrix} \underbrace{\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}}_{v} = pv.$$

so by definition of eigenvalues and eigenvectors we can say that the eigenvalue of our matrix A is $\lambda_1 = p$. The rank$(A) = 1$ so λ_1 is the only eigenvalue different from 0.

Spectral theorem (for real symmetric matrices)

Let $A \in \mathbb{R}^{n \times n}$ be real and symmetric. Then

1. The eigenvalues of A are real .
2. A is diagonalizable.
3. There is an orthonormal basis of $\mathbb{R}^n$ consiting on the eigenvectors of A.

In short, A may be orthonormaly diagonalized: $A = VDV^T$ where $V \in \mathbb{R}^{n\times n}$ is an orthonormal matrix of eigenvectors of A and D is a real diagonal matrix of eigenvalues of A.

So by the **Spectral Theorem** we can say that $\exists Q \in \mathbb{R}^{p\times p}$ orthonormal matrix such that $A = QDQ^T$ where $D = \begin{pmatrix} p & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}$.

Now we find the eigenvalue of $\Sigma = \pi_0(I_p - \pi_0 A)$; we recover A with QDQ^T:

$$\begin{aligned} \Sigma &= \pi_0(I_p - \pi_0 QDQ^T) \\ &= \pi_0(QI_pQ^T - \pi_0 QDQ^T) \\ &= \pi_0 Q(I_p - \pi_0 D)Q^T \end{aligned}$$

Definition Similar Matrix

A matrix A is said similar to $B \iff \exists P$ invertible such that $A = P^{-1}BP$.

From this definition and by how we wrote Σ we can say that the matrices Σ and $T := \pi_0(I_p - \pi_0 D)$ are similar so we can proceed by finding the eigenvalues of T (since two similar matrices share the same eigenvalues).

$$\begin{aligned} |T - \lambda I_p| &= \left| \begin{pmatrix} \pi_0(1-p\pi_0) - \lambda & & & \\ & \pi_0 - \lambda & & \\ & & \ddots & \\ & & & \pi_0 - \lambda \end{pmatrix} \right| = 0 \\ &= [\pi_0(1-p\pi_0) - \lambda] \cdot (\pi_0 - \lambda)^{p-1} = 0 \end{aligned}$$

So the eigenvalues of Σ are $\tilde{\lambda}_1 = \pi_0(1-p\pi_0)$ and $\tilde{\lambda}_2 = \dots = \tilde{\lambda}_p = \pi_0$.

3.2 In order to find p such that the proportion of the variance explained by the last principal component is less than 20% we need to find the smallest eigenvalue.

$$\tilde{\lambda} = (\pi_0(1-\pi_0 p), \underbrace{\pi_0, \dots, \pi_0}_{p-1})$$

We check that $\pi_0 > \pi_0(1-\pi_0 p)$:

$$\pi_0 > \pi_0(1-\pi_0 p) \iff 1 > (1-\pi_0 p) \iff \pi_0 p > 0 \quad \forall p > 0$$

So we have that the smallest eigenvalue $\tilde{\lambda}_p = \pi_0(1-\pi_0 p)$.

Now we find the total variance as the sum of the eigenvalues of Σ:

$$\begin{aligned} Var(X) &= (1-\pi_0 p)\pi_0 + \pi_0(p-1) = \\ &= \pi_0 - \pi_0^2 p + \pi_0 p - \pi_0 \\ &= \pi_0 p(1-\pi_0) \end{aligned}$$

And using the formula $\frac{\tilde{\lambda}_p}{Var(X)}$ we set the proportion less then 20%

$$
\begin{aligned}
&\frac{\tilde{\lambda}_p}{Var(X)} < \frac{1}{5} \\
&\frac{\pi_0(1-\pi_0 p)}{\pi_0 p(1-\pi_0)} < \frac{1}{5} \quad 0 < \pi_0 < \frac{1}{p} \\
&\frac{(1-\pi_0 p)}{p(1-\pi_0)} - \frac{1}{5} < 0 \\
&\frac{5\pi(1-\pi_0 p) - \pi_0 p(1-\pi_0)}{5\pi_0 p(1-\pi)} < 0 \\
&\frac{5\pi_0 - 5\pi_0^2 p - \pi_0 p + \pi_0^2 p}{5\pi_0 p(1-\pi_0)} < 0 \\
&\frac{-\pi_0^2 - \pi_0 p + 5\pi_0}{5\pi_0 p(1-\pi_0)} <^3 0 \\
&\frac{\pi_0 p(-4\pi_0 - 1) + 5\pi_0}{5\pi_0 p(1-\pi_0)} < 0 \\
&p(-4\pi_0 - 1) + 5 < 0 \\
&p > \frac{5}{4\pi_0 + 1}
\end{aligned}
$$

(3) The denominator is always greater than 0.

We want to eliminate the dependence on π_0 in the previous inequality, so as to find a lower bound for p. To do so, we consider the worst case scenario in which the first $p-1$ principal components contribute as little as possible to the total explained variance, thus forcing the last principal component to assume the greatest value possible (i.e. when $\pi_0 \to 0$). Mathematically we can prove it by showing the the function $f(\pi_0) = \frac{\pi_0(1-p\pi_0)}{\pi_0 p(1-\pi_0)}$ is decreasing in its domain $\left(0, \frac{1}{p}\right)$

$$
\begin{aligned}
f'(\pi_0) &= \frac{-p(p(1-\pi_0)) - (1-\pi_0\pi)(-p)}{(p(1-\pi_0))^2} && (2) \\
&= \dots && (3) \\
&= \frac{(1-p)}{p(1-\pi_0)^2} < 0 \iff p > 1 && (4)
\end{aligned}
$$

In that limiting case, we obtain a simple lower bound that no longer depends on π_0.

$$
p > \frac{5}{4\pi_0 + 1} \underset{\pi_0 \to_0}{\to} 5
$$

Point 4

Perform a simulation study with $p = 3$, $\pi_0 = 1/4$ and $N = 1000$ Monte Carlo samples of $n = 100$ multinomially distributed Y_i. For $X = (X_1, X_2, X_3)$, make a scatterplot of the N values of X_2 vs X_1 and sketch the ellipse corresponding to the contour of the (theoretical limiting) bivariate density of (X_1, X_2) which contains 95% probability.

We firstly initialize the variables such that they are coherent with the text, then we sample from a uniform discrete random variable (since each outcome is equally likely with probability $\pi_0 = \frac{1}{4}$). After we codify Y_i with one-hot random vectors.($Y_i \sim$ Multinomial $\left(1, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$). Once we sample $n = 100$ Y_i, we sum

the columns to get the random variable which is distributed as a $\sum_{i=1}^{100} Y_i \sim \text{Multinomial}\left(100, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$, we repeat this process $N = 1000$ and we apply the CLT on this 1000 sample to get the random vector $\sqrt{n}(\hat{\pi} - \pi) \sim \mathcal{N}_4(\underline{0}, \text{diag}(\pi) - \pi\pi^T)$.

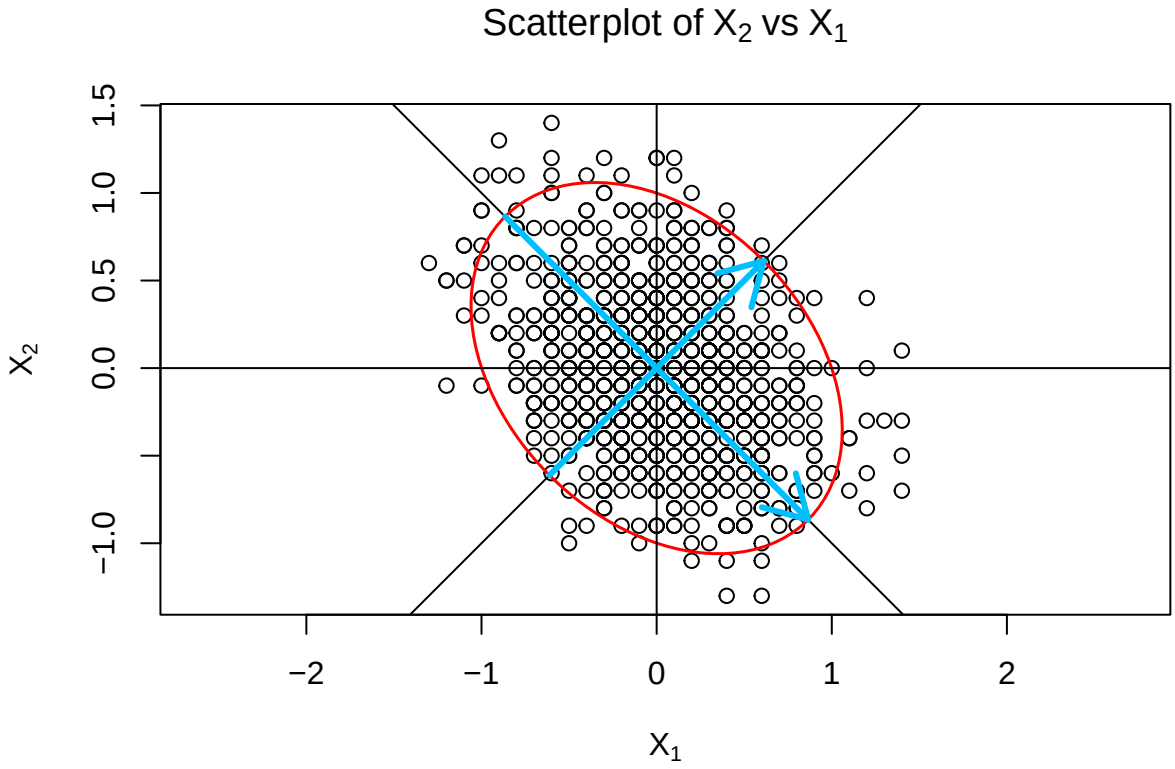
```
#We set seed to standardize result
set.seed(4567)
p = 3
pi_0 = 1/4
N = 1000
n= 100

norm_sample<- matrix(NA, nrow=N, ncol=4)
for (i in (1:N)){
  # Matrix nx4 for codified vector inizialed as full of NAs
  Y_hot_i = matrix(NA, nrow=n, ncol=4)
  # Samples from a uniform discrete (not yet codified)
  Y_i = sample(4,n,replace =T)
  for (j in (1:n)){
      # This transforms the Yi (not yet codified) into the codified vector
      Y_hot_i[j,] = as.numeric(1:4 %in% Y_i[j])
    }
  M = Y_hot_i %>%
  colSums()
  norm_sample[i,]<-(M/n - pi_0)*sqrt(n)
}
X <- norm_sample[,1:3]
```

We then pick the first 3 columns to get X, we plot the first two, and draw the contour line for the theoretical bivariate distribution parameterized as such $\mathcal{N}_2\left(\underline{0}, \begin{bmatrix} \frac{3}{16} & -\frac{1}{16} \\ -\frac{1}{16} & \frac{3}{16} \end{bmatrix}\right)$.

Furthermore we compute the eigenvectors and we use them to find the axes (using this formula $c\sqrt{\tilde{\lambda}_j e_j}$), of the theoretical ellipse which follows the expression $(x-\mu)^T\Sigma^{-1}(x-\mu) = c^2$ where $c = F_{\chi^2_2}(0.95)$.

```
Sigma = matrix( c(3/16 , -1/16, -1/16,3/16) ,nrow=2 , ncol=2)
sigma.e <- eigen(Sigma)
#eigenvectors
P <- sigma.e$vectors
#eigenvalues
lambda <- sigma.e$values
mu <- c(0,0)
e1 <- P[,1]
e2 <- P[,2]
# Direction of the first eigenvector
b <- e1[2]/e1[1]
a <- b*mu[1]+ mu[2]
# Direction of the second eigenvector
b2 <- e2[2]/e2[1]
a2 <- b2*mu[1]+ mu[2]
cc<-sqrt(qchisq(0.95, 2))
# Points touched by eigenvectors
x<-cbind(cc*sqrt(lambda[1])*e1+mu,
         -cc*sqrt(lambda[1])*e1+mu,
         cc*sqrt(lambda[2])*e2+mu,
         -cc*sqrt(lambda[2])*e2+mu)
```

```r
plot(X[,1],X[,2], xlab=expression(X[1]), ylab = expression(X[2]),
     main=expression("Scatterplot of"~ X[2] ~ "vs"~ X[1]), asp= 1 )
lines(ellipse(x=Sigma,centre=mu,level=0.95),col="red",lwd=1.5)
# Direction of the first and second axes
abline(a = a , b=b)
abline(a= a2, b=b2)
abline(h=0,v=0)
arrows(x[1,1],x[2,1],x[1,2],x[2,2],code=2,col="deepskyblue",lwd=3)
arrows(x[1,3],x[2,3],x[1,4],x[2,4],code=2,col="deepskyblue",lwd=3)
```

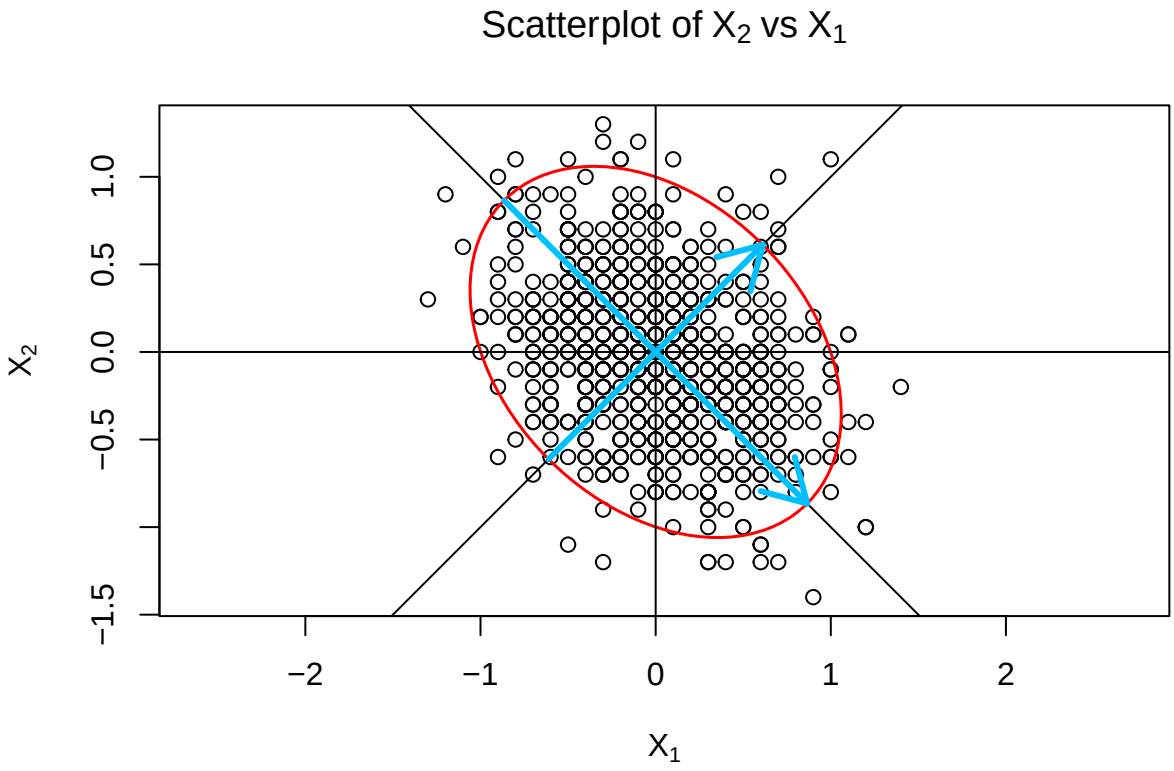


As we can see from the plot above our simulated point fit perfectly in the theoretical ellipse, which is a strong indication that our simulation process is correct. The ellipse is also tilted to the left because the correlation between the two variable is negative, further the observations have equal spread for both the X_1 X_2 variables (since the variance is equal).

We also decide to sample directly from a multinational distribution and we obtain very similiar results.

```r
set.seed(4567)

norm_samplev2<- matrix(NA, nrow=N, ncol=4)
for (i in (1:N)){
  M<-rmultinom(1,size=n,prob=c(1/4,1/4,1/4,1/4))
  norm_samplev2[i,]<-(M/n - pi_0)*sqrt(n)
}
Xv2 <- norm_samplev2[,1:3]
plot(Xv2[,1],Xv2[,2], xlab=expression(X[1]), ylab = expression(X[2]),
     main=expression("Scatterplot of"~ X[2] ~ "vs"~ X[1]), asp= 1 )
```

```r
lines(ellipse(x=Sigma,centre=mu,level=0.95),col="red",lwd=1.5,asp=1)
abline(a = a , b=b)
abline(a= a2, b=b2)
abline(h=0,v=0)
arrows(x[1,1],x[2,1],x[1,2],x[2,2],code=2,col="deepskyblue",lwd=3)
arrows(x[1,3],x[2,3],x[1,4],x[2,4],code=2,col="deepskyblue",lwd=3)
```



Point 5

Find the conditional distributions of $(X_1, X_2) \mid X_3 = x_3$ and of $X_3 \mid (X_1 = x_1, X_2 = x_2)$.

Given a Gaussian random vector we know that conditioning on it the conditioned vector is still Gaussian with parameters :

$$\overline{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$$
$$\overline{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

We know that $(X_1, X_2, X_3) \sim \mathcal{N}_3(\underline{0}, \Sigma)$ where

$$\Sigma = \begin{pmatrix} \frac{3}{16} & -\frac{1}{16} & -\frac{1}{16} \\ -\frac{1}{16} & \frac{3}{16} & -\frac{1}{16} \\ -\frac{1}{16} & -\frac{1}{16} & \frac{3}{16} \end{pmatrix}$$

$$\left(\begin{array}{cc|c} \frac{3}{16} & -\frac{1}{16} & -\frac{1}{16} \\ -\frac{1}{16} & \frac{3}{16} & -\frac{1}{16} \\ \hline -\frac{1}{16} & -\frac{1}{16} & \frac{3}{16} \end{array}\right) = \left(\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right)$$

5.1 We need to find the distribution of $(X_1, X_2)|X_3 = x_3 \sim \mathcal{N}_2(\mu_{12|3}, \Sigma_{12|3})$.

So now we have to apply the formula above adapting the notation to our conditioned vector.

The mean vector is:

$$\begin{aligned}\mu_{12|3} &= \mu_{12} + \Sigma_{12}\Sigma_{22}^{-1}(x_3 - \mu_3) \\ &= \begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix} + \begin{pmatrix}-\frac{1}{16} \\ -\frac{1}{16}\end{pmatrix}\frac{16}{3}(x_3 - \mu_3) \\ &= \begin{pmatrix}0 \\ 0\end{pmatrix} + \begin{pmatrix}-\frac{1}{3} \\ -\frac{1}{3}\end{pmatrix} x_3 = \begin{pmatrix}-\frac{1}{3}x_3 \\ -\frac{1}{3}x_3\end{pmatrix}\end{aligned}$$

$$\begin{aligned}\Sigma_{12|3} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\ &= \begin{pmatrix}\frac{3}{16} & -\frac{1}{16} \\ -\frac{1}{16} & \frac{3}{16}\end{pmatrix} - \begin{pmatrix}-\frac{1}{16} \\ -\frac{1}{16}\end{pmatrix}\frac{16}{3}\begin{pmatrix}-\frac{1}{16} & -\frac{1}{16}\end{pmatrix} \\ &= \begin{pmatrix}\frac{3}{16} & -\frac{1}{16} \\ -\frac{1}{16} & \frac{3}{16}\end{pmatrix} - \begin{pmatrix}\frac{1}{48} & \frac{1}{48} \\ \frac{1}{48} & \frac{1}{48}\end{pmatrix} \\ &= \begin{pmatrix}\frac{1}{6} & -\frac{1}{12} \\ -\frac{1}{12} & \frac{1}{6}\end{pmatrix}.\end{aligned}$$

So a the end

$$(X_1, X_2)|X_3 = x_3 \sim \mathcal{N}_2\left(\begin{bmatrix}-\frac{1}{3}x_3 \\ -\frac{1}{3}x_3\end{bmatrix}, \begin{bmatrix}\frac{1}{6} & -\frac{1}{12} \\ -\frac{1}{12} & \frac{1}{6}\end{bmatrix}\right).$$

5.2 We now have to find the Gaussian random variable:

$$X_3|(X_1 = x_1, X_2 = x_2) \sim \mathcal{N}_1(\mu_{3|12}, \Sigma_{3|12}).$$

We get the mean:

$$\begin{aligned}\mu_{3|1,2} &= \mu_3 + \Sigma_{12}\Sigma_{11}^{-1}\begin{pmatrix}x_1 - \mu_1 \\ x_2 - \mu_2\end{pmatrix} \\ &= 0 + \begin{pmatrix}-\frac{1}{16} & -\frac{1}{16}\end{pmatrix}32\begin{pmatrix}\frac{3}{16} & \frac{1}{16} \\ \frac{1}{16} & \frac{3}{16}\end{pmatrix}\begin{pmatrix}x_1 \\ x_2\end{pmatrix} \\ &= \begin{pmatrix}-\frac{1}{16} & -\frac{1}{16}\end{pmatrix}\underbrace{\begin{pmatrix}6 & 2 \\ 2 & 6\end{pmatrix}}_{\Sigma_{11}^{-1}}\begin{pmatrix}x_1 \\ x_2\end{pmatrix} \\ &= \begin{pmatrix}-\frac{1}{2} & -\frac{1}{2}\end{pmatrix}\begin{pmatrix}x_1 \\ x_2\end{pmatrix} \\ &= -\frac{1}{2}x_1 - \frac{1}{2}x_2\end{aligned}$$

And the variance:

$$
\begin{aligned}
\Sigma_{3|12} &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \\
&= \frac{3}{16} - \begin{pmatrix} -\frac{1}{16} & -\frac{1}{16} \end{pmatrix} \begin{pmatrix} 6 & 2 \\ 2 & 6 \end{pmatrix} \begin{pmatrix} -\frac{1}{16} \\ -\frac{1}{16} \end{pmatrix} \\
&= \frac{3}{16} - \begin{pmatrix} -\frac{1}{2} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} -\frac{1}{16} \\ -\frac{1}{16} \end{pmatrix} \\
&= \frac{3}{16} - \frac{1}{16} = \frac{1}{8}
\end{aligned}
$$

So at the end:

$$
X_3|(X_1 = x_1, X_2 = x_2) \sim \mathcal{N}\left(-\frac{1}{2}x_1 - \frac{1}{2}x_2, \frac{1}{8}\right)
$$

Exercise 2

The Boston data contains housing values in 506 suburbs of Boston. We will work with all variables ad except for zn, chas, rad and medv.

```
X<-Boston[,-c(2,4,9,14)]
head(X)
```

```
##      crim indus   nox    rm  age    dis tax ptratio  black lstat
## 1 0.00632  2.31 0.538 6.575 65.2 4.0900 296    15.3 396.90  4.98
## 2 0.02731  7.07 0.469 6.421 78.9 4.9671 242    17.8 396.90  9.14
## 3 0.02729  7.07 0.469 7.185 61.1 4.9671 242    17.8 392.83  4.03
## 4 0.03237  2.18 0.458 6.998 45.8 6.0622 222    18.7 394.63  2.94
## 5 0.06905  2.18 0.458 7.147 54.2 6.0622 222    18.7 396.90  5.33
## 6 0.02985  2.18 0.458 6.430 58.7 6.0622 222    18.7 394.12  5.21
```

Let's start by analyzing the variables presented here:

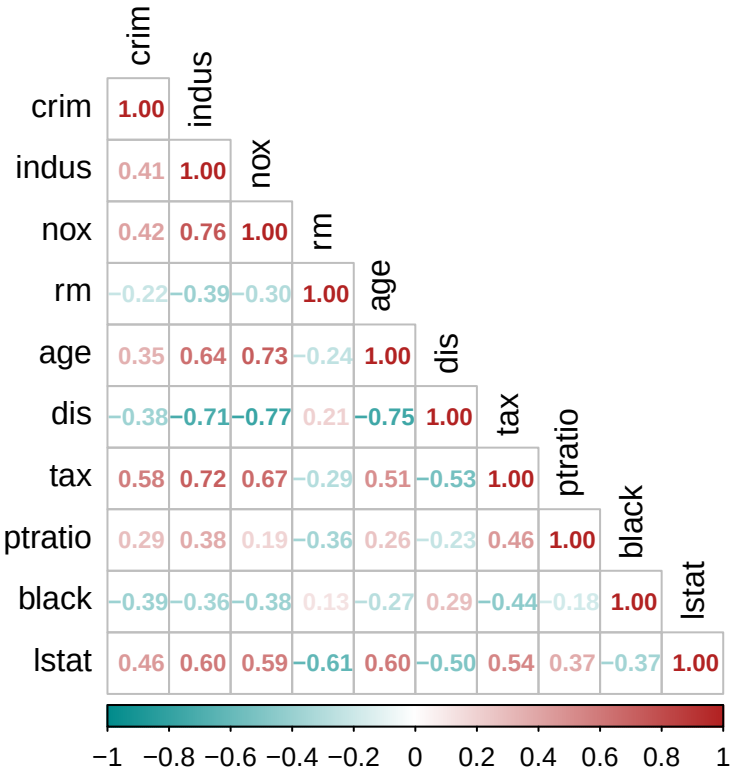
- **crime**: represents the per capita crime rate by town
- **indus**: represents the proportion of non-retail business acres per town
- **nox**: represents the concentration of nitrogen oxides (parts per 10 million)
- **rm**: represents the average number of rooms per dwelling
- **age**: represents the proportion of owner-occupied units built prior to 1940
- **dis**: represents the weighted distances to five Boston employment centers
- **tax**: represents the property tax rate per $10,000
- **ptratio**: represents the pupil-teacher ratio by town
- **black**: represents $1000(B - 0.63)^2$, where B is the proportion of Black residents by town

Point 1

Compute the correlation matrix R and comment on the largest 4 correlations.

With the following code we plot the correlation matrix of our dataset.

```
R = cor(X)

col_inv = colorRampPalette(c("darkcyan", "white", "firebrick"))(300)

corrplot(R, col = col_inv, method = "number",type="lower",tl.col    = "black",
         title = "Correlation Table",number.cex =0.75,  mar = c(0,0,1,0)

)
```



We immediately see that the data exhibit a noticeable dependence structure, as evidenced by the above correlation matrix, which shows moderately high correlation (both positive and negative) among the variables in the dataset.

Let's now find, with the following code, the 4 largest correlations.

```
p= ncol(X)

b =  sort(abs(R[lower.tri(R)]), decreasing = T)[1:4]
indices = (matrix( abs(R) %in% b , ncol=p ) & lower.tri(R)) %>% which(arr.ind = TRUE)


result_table = tibble(
  Variable1 = rownames(R)[indices[, 1]],
  Variable2 = rownames(R)[indices[, 2]],
  Correlation = R[indices]
)

print(result_table)
```

```
## # A tibble: 4 x 3
##   Variable1 Variable2 Correlation
##   <chr>     <chr>           <dbl>
## 1 nox       indus           0.764
## 2 age       nox             0.731
## 3 dis       nox            -0.769
## 4 dis       age            -0.748
```

The variable *nox* quantifies the concentration of nitrogen oxides (measured in parts per 10 million), while the variable *indus* represents the proportion of non-retail business acres per town. Correlation is an index ranging from -1 to 1, where -1 indicates an inverse correlation, meaning that high values of one variable correspond to low values of the other, and vice versa while values close to 1 indicate a perfect linear correlation. As expected, there is a positive correlation between *nox* and *indus*, since highly industrialized areas tend to emit more nitrogen oxides that are a common byproduct of industrial activities (e.g. all combustion processes).

The variable *age* describes the proportion of owner-occupied units built before 1940. A strong correlation with *nox* is therefore unsurprising. Focusing specifically on residential buildings, older houses are expected to have higher emissions, leading to an increased concentration of nitrogen oxides in those areas.

The variable *dis* represents the weighted mean distance to five major employment centers in Boston. It is reasonable to expect that is negatively correlated with the variable *nox* beacause areas farther from Boston are more rural and generally experience lower pollution levels.

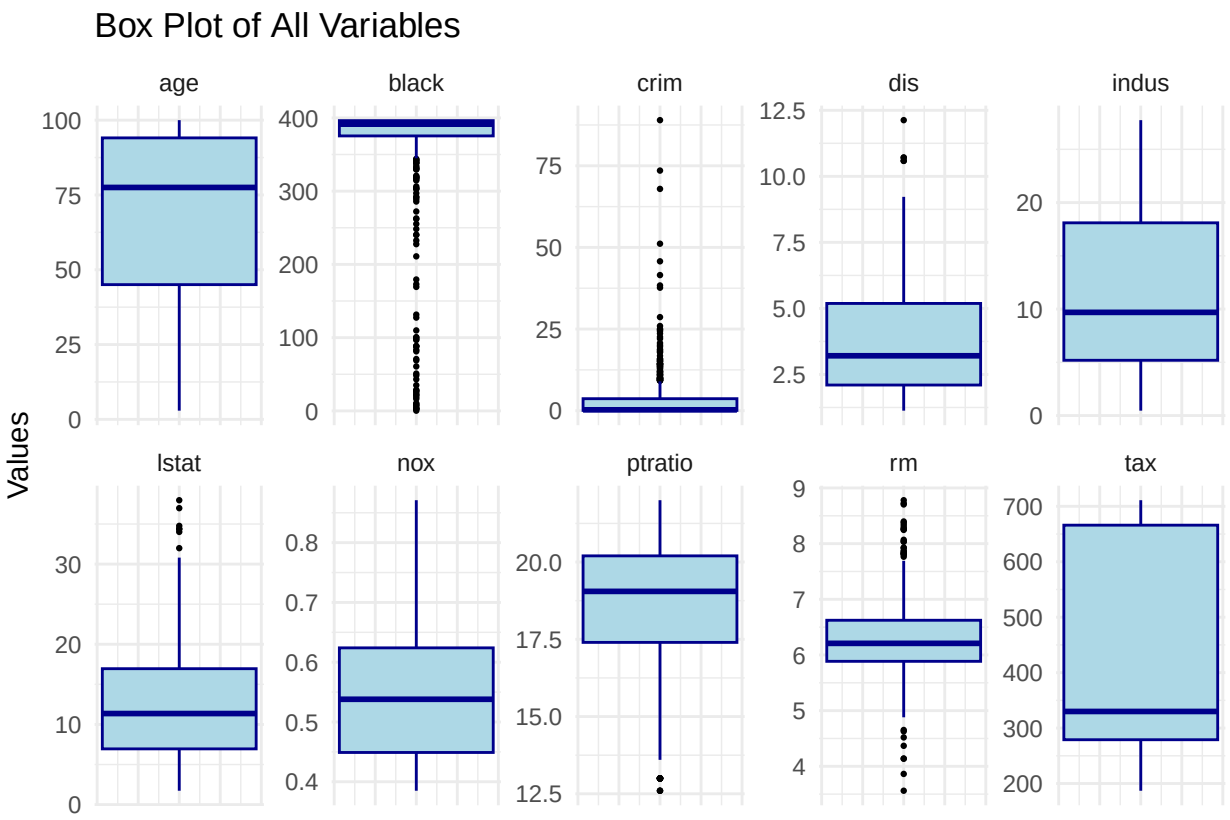
Finally, a correlation is observed between the *age* of buildings and their distance from the city center (*dis*). The closer a neighborhood is to Boston, the more likely it is to contain older buildings. This phenomenon can be explained by the city's historical expansion: Boston likely grew outward from an initial urban core, where older houses were originally built, with newer constructions developing over time as the city expanded.

Point 2

Identify the 3 most extreme univariate outliers.

First, we present a boxplot to visualize the distribution of each variable and identify potential outliers. The variables are plotted on their original scales, without standardization, to allow for first look of their individual behavior, since they naturally operate on different scales.

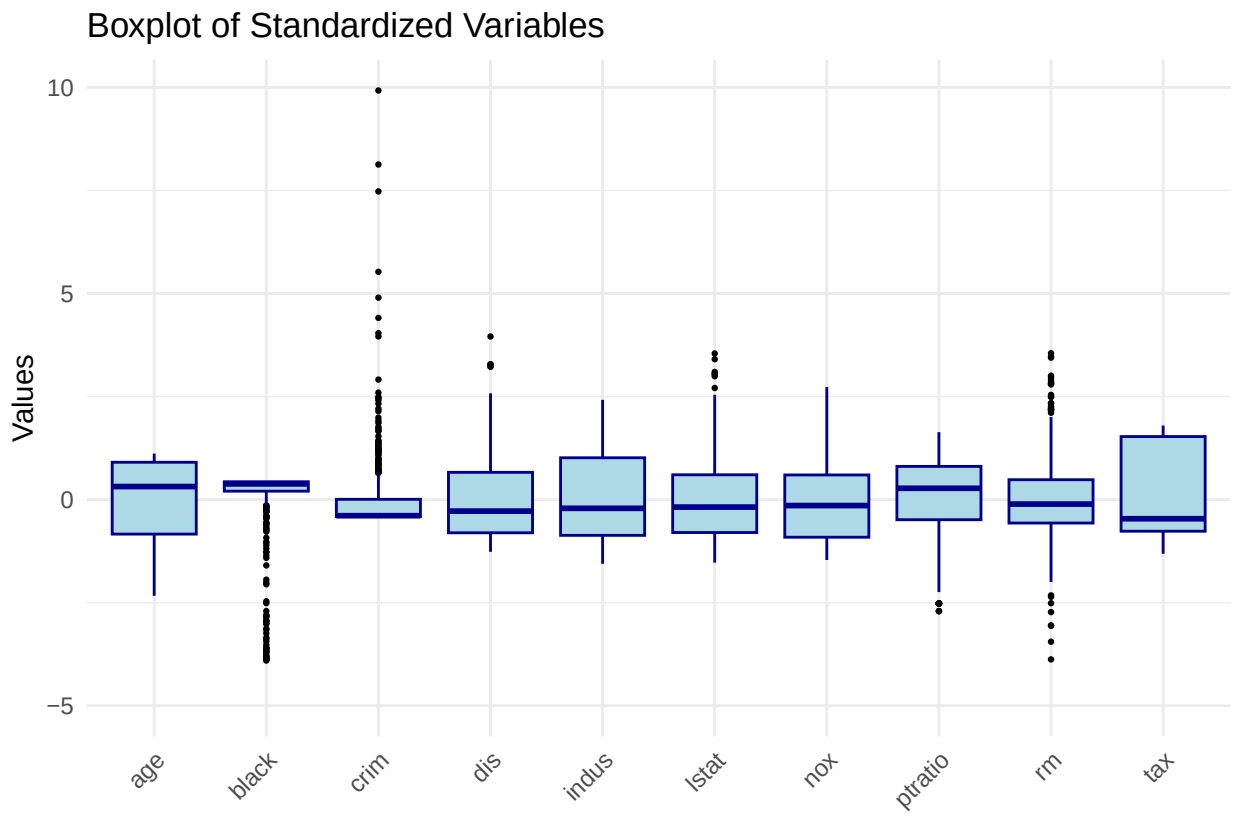
```
df_long <- X %>%
  pivot_longer(cols = everything(),
               names_to = "Variabile",
               values_to = "Valore")

ggplot(df_long, aes(x = 1, y = Valore)) +
  geom_boxplot(fill = "lightblue", color = "darkblue",
               outlier.colour = "black", outlier.shape = 16, outlier.size = 0.8) +
  facet_wrap(~ Variabile, scales = "free_y", ncol = 5, nrow = 2) +
  labs(title= "Box Plot of All Variables", y = "Values", x="") +
  theme_minimal() +
  theme(
    axis.text.x = element_blank(),
    axis.ticks.x = element_blank()
  )
```



From the boxplot above, we observe that some variables are more prone to outliers than others. This is particularly evident in variables such as *black*, *crime*, and *rm*, compared to variables like *age*, *indus*, *nox*, and *tax*, which seem to exhibit fewer atypical behaviors.

We now consider the standardized data, so that all variables share the same scale, allowing for a single comparable visualization.

```
X_scaled = as.data.frame(scale(X))
df_long <- X_scaled %>%
  pivot_longer(cols = everything(), names_to = "Variabile", values_to = "Valore")

ggplot(df_long, aes(x = Variabile, y = Valore)) +
  geom_boxplot(fill = "lightblue", color = "darkblue", outlier.colour = "black",
               outlier.shape = 16, outlier.size = 0.8) +
  coord_cartesian(ylim = c(-5, max(df_long$Valore))) +
  labs(title = "Boxplot of Standardized Variables", x="",  y = "Values") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



This graph confirms that the variables *black*, *crime*, and *rm* have more outliers.

Additionally, some outliers deviate significantly from the mean. We now proceed to identify the three most extreme univariate outliers.

The following code compute for each variable the univariate distance (using the absolute value) of each observation from the corresponding column mean, and then stores in a separate dataframe the variable name and the index line of the three most extreme univariate outliers.

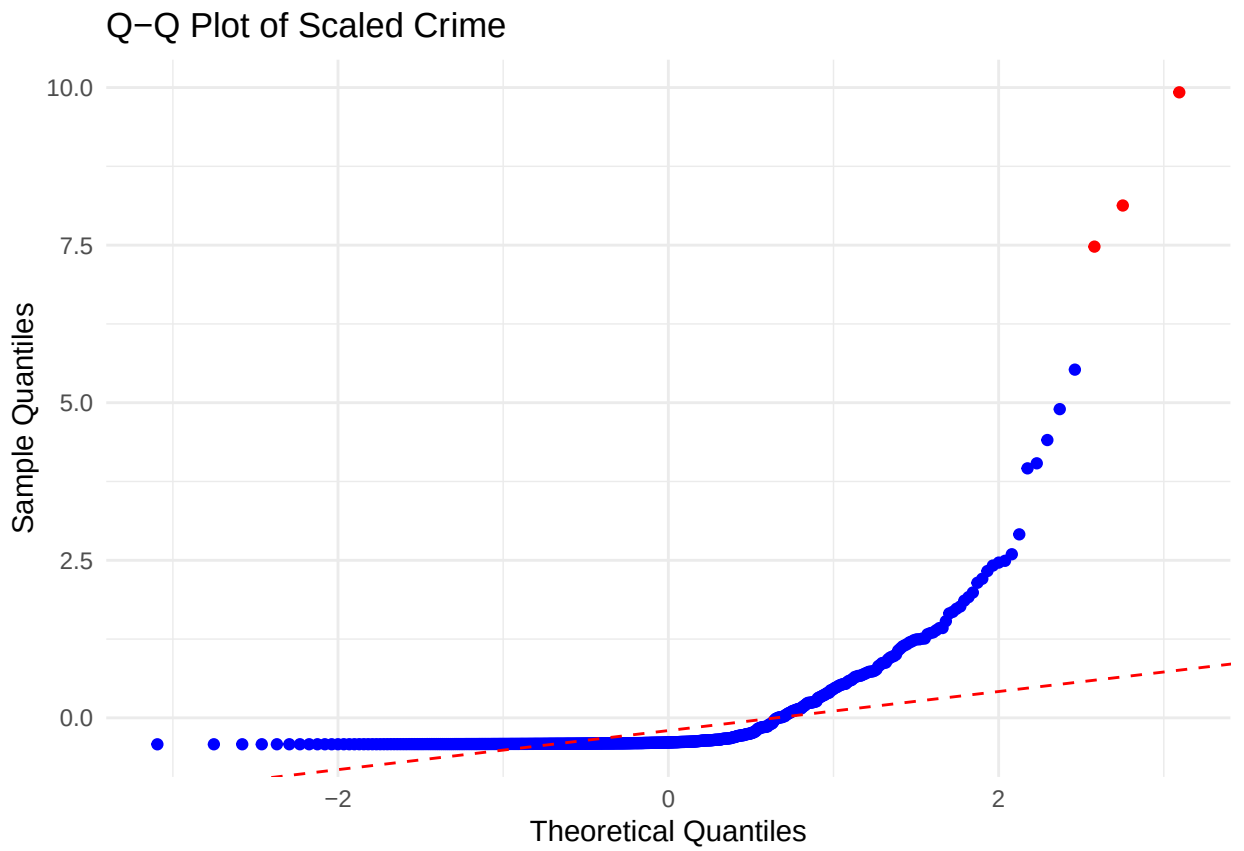
```
k = 3
Outliers =  data.frame(outliers = c("max_outlier1", "max_outlier2", "max_outlier3"),
                       col_name = c(0,0,0),
                       row_index = c(0,0,0),
                       row.names = 1)

maxoutliers = rep(0, k)
X_dist = abs(X_scaled)

for (i in (1:p)) {
  sorted = sort(X_dist[,i], decreasing = T, index.return = T)
  for (j in (1:k)) {
    if (sorted$x[j] > maxoutliers[j]) {
      maxoutliers[j] = sorted$x[j]
      Outliers[j,1] =  names(X)[i]
      Outliers[j,2] =  sorted$ix[j]
    }
  }
}
```

```
Outliers
```

```
##              col_name row_index
## max_outlier1     crim       381
## max_outlier2     crim       419
## max_outlier3     crim       406
```

As we see from the above output, the three most extreme univariate outliers are all in the *crim* variable. This naturally raises the question of the behavior of crime. To investigate further, we create a QQ plot to assess whether the crime variable can be approximated by a normal distribution. We compute the empirical quantiles and compare them with the theoretical quantiles of the normal distribution. For the data to be normally distributed, the points should closely follow the red line, which would indicate a good approximation to the normal distribution.

```
scaled_crim = X_scaled$crim
scaled_crim_df = as.data.frame(x = scaled_crim)

qq_data <- as.data.frame(qqnorm(scaled_crim, plot.it = FALSE))
names(qq_data) <- c("theoretical", "sample")

qq_data$index <- 1:nrow(qq_data)

qq_data$color <- ifelse(qq_data$index %in% Outliers[,2], "red", "blue")

y_quantiles = quantile(scaled_crim, probs = c(0.25, 0.75))
x_quantiles = qnorm(c(0.25, 0.75))

slope = diff(y_quantiles) / diff(x_quantiles)
intercept = y_quantiles[1] - slope * x_quantiles[1]

ggplot(qq_data, aes(x = theoretical, y = sample)) +
  geom_point(aes(color = color)) +
  geom_abline(slope = slope,
              intercept = intercept, color = "red", linetype ="dashed") +
  scale_color_identity() +
  theme_minimal()+
  labs(title = "Q-Q Plot of Scaled Crime",
       x = "Theoretical Quantiles", y = "Sample Quantiles")
```



We can observe from the plot that the distribution of *crim* does not follow a perfectly normal pattern. In particular, the previously identified outliers stand out significantly. So we can conclude saying that the variable *crim* has the three most univariate outliers indicating that the crime rate distribution has extremly high values for some neighborhoods.

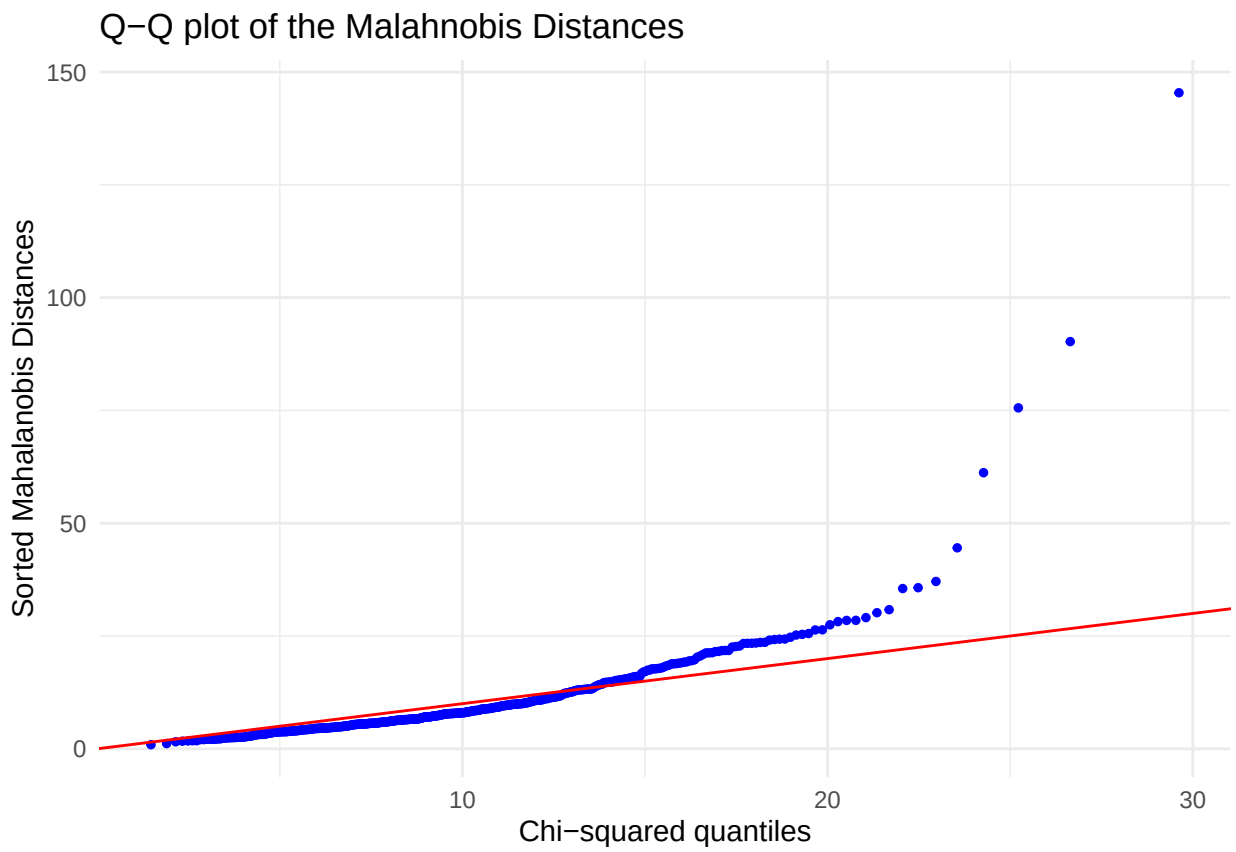
Point 3

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

Mahalanobis distance measures how "unusual" a point is compared to a group of data, considering not only its distance from the mean but also how the variables are distributed and correlated with each other. With the following code, we construct the chi-square Q-Q plot of the squared Mahalanobis distances of our data.

```
n <- nrow(X_scaled)
bar.x<-colMeans(X_scaled)
S = cov(X_scaled)

d <- mahalanobis(X_scaled, center = bar.x, cov = S)

chi2_quantiles <- qchisq(ppoints(d), df = p)

sorted_d <- sort(d)

ggplot(cbind(chi2_quantiles, sorted_d), aes(x = chi2_quantiles, y = sorted_d)) +
  geom_point( color = "blue", size = 1) +
  geom_abline(slope = 1, intercept = 0, color = "red") +
  theme_minimal() +
```

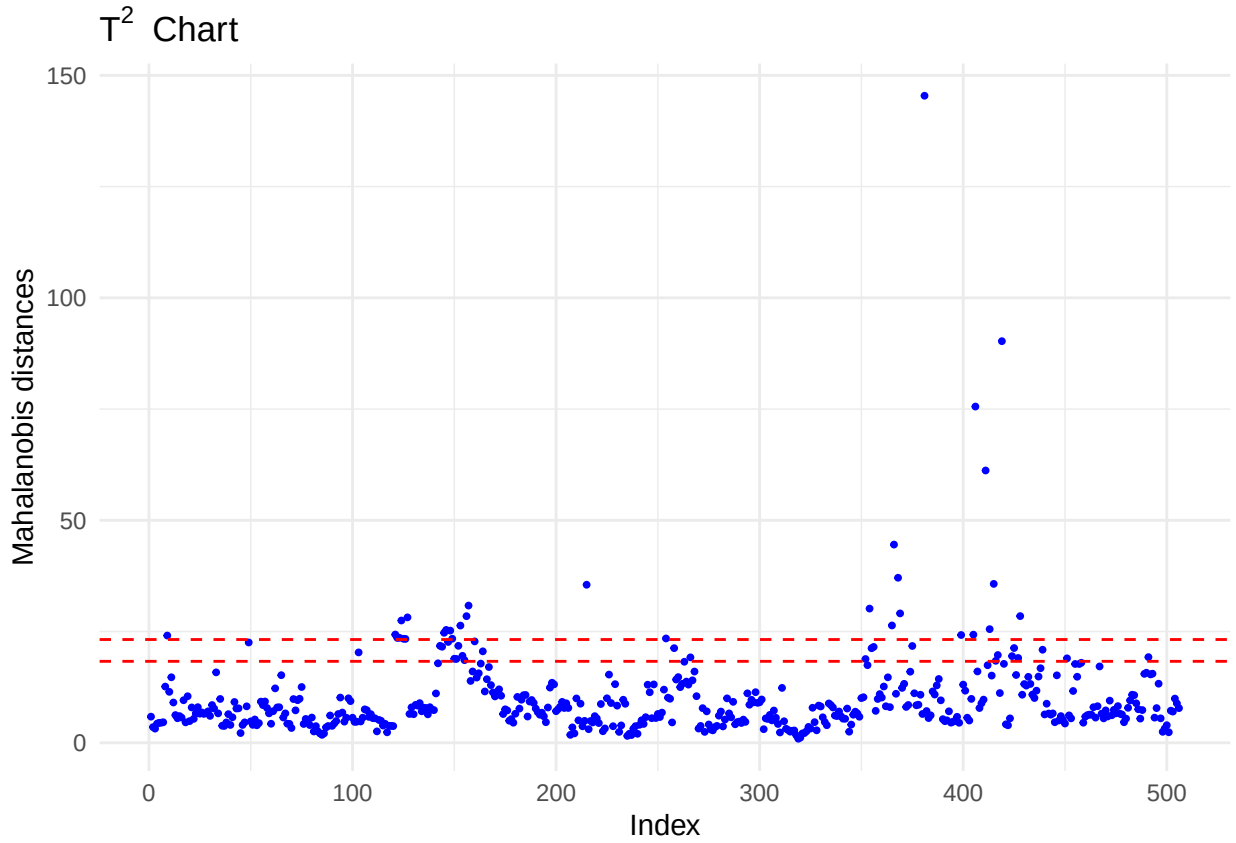
```
  labs(title = "Q-Q plot of the Malahnobis Distances", x = "Chi-squared quantiles",
       y = "Sorted Mahalanobis Distances")
```



As we can see from the above figure, there seem to be several points deviating above the theoretical quantile line on the right-hand side. This pattern suggests the presence of heavier tails in the empirical distribution compared to the theoretical chi-square distribution, indicating a departure from multivariate normality. In particular, some observations may be considered outliers with respect to the multivariate normal model.

Let's verify it better with the T^2 chart with the two levels 0.95 and 0.99.

```
ggplot(data = cbind(index = 1:n, d), aes(x = index, y = d)) +
  geom_point(color = "blue", size = 0.7) +
  geom_hline(yintercept = qchisq(0.95,df=p), lty = 2, color = "red") +
  geom_hline(yintercept = qchisq(0.99,df=p), lty = 2, color = "red") +
  theme_minimal() +
  labs(title = expression(T^2~" Chart"), x = "Index", y = "Mahalanobis distances")
```



The T^2 chart above reveals that a substantial number of observations (over 30 out of approximately 500) lie above the 0.99 threshold. Under the assumption of multivariate normality, one would expect around 1% of the observations (roughly 5 out of 500) to exceed this limit. The presence of a disproportionately large number of such points suggests the existence of potential outliers or a deviation from the multivariate normal model.

The following script counts the number of observations which exceed the 0.99 level in the T^2 chart.

```
alpha = 0.99
quant = qchisq(alpha,df=p)
extreme_obs = 0

for (i in 1:n ) {
  if(d[i] > quant){
    extreme_obs = extreme_obs +1
  }
}
extreme_obs
```

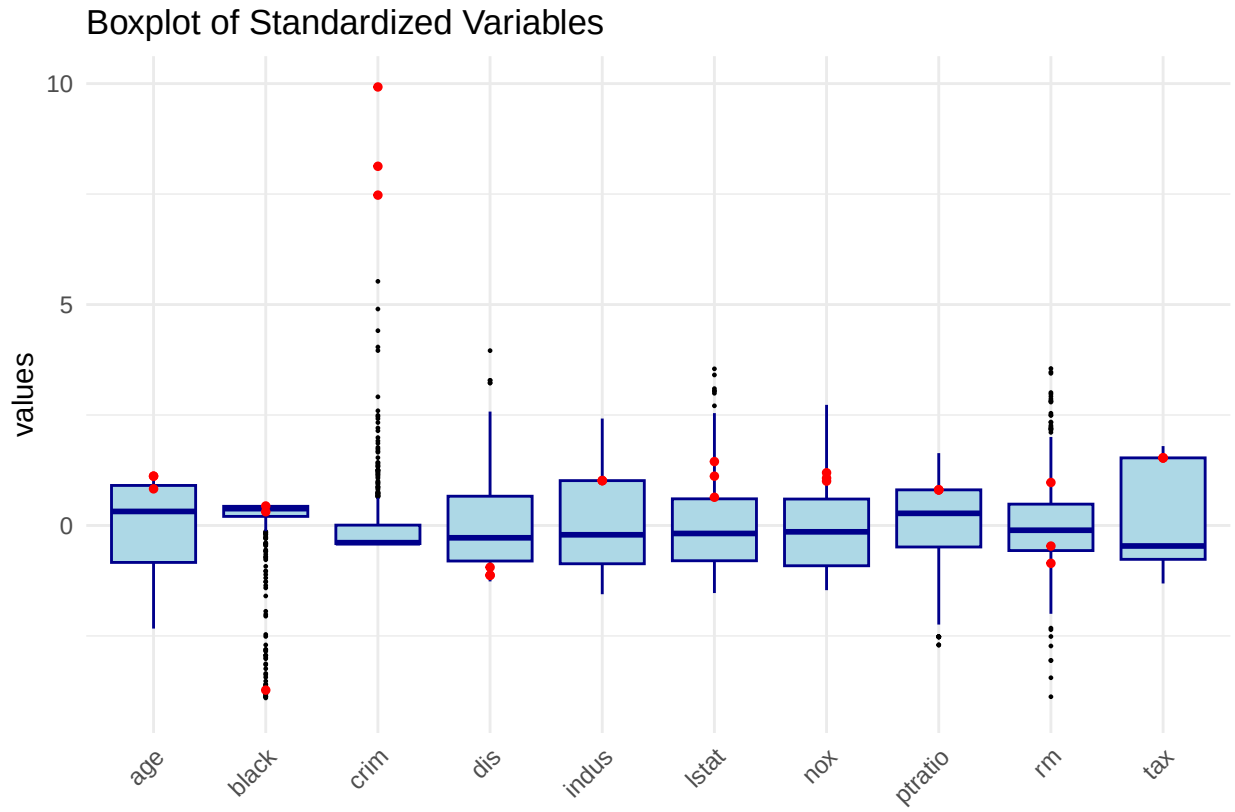
```
## [1] 31
```

Point 4

Are the univariate outliers identified in point 2. also multivariate outliers? Justify your answer.

In order to verify if the observations corresponding to the univariate outliers found in point 2 are also multivariate outliers, let's first visualize them in the boxplot figure presented before.

```r
df_long = X_scaled %>%
  mutate(Riga = row_number()) %>%
  pivot_longer(cols = -Riga, names_to = "Variabile", values_to = "Valore")

d = df_long %>%
  filter(Riga == Outliers[1,2] | Riga == Outliers[2,2] | Riga == Outliers[3,2] )

ggplot(df_long, aes(x = Variabile, y = Valore)) +
  geom_boxplot(fill = "lightblue", color = "darkblue", outlier.colour = "black",
               outlier.shape = 16, outlier.size = 0.5) +
  geom_point(data = d, aes(x = Variabile, y = Valore),
             color = "red", size = 1) +
  coord_cartesian(ylim = c(-4, max(df_long$Valore))) +
  labs(title = "Boxplot of Standardized Variables", x = "", y = "values") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "none")
```



In the above figure we colored in red the values (for each variables) of the three observations which correspond to the univariate outliers found before.

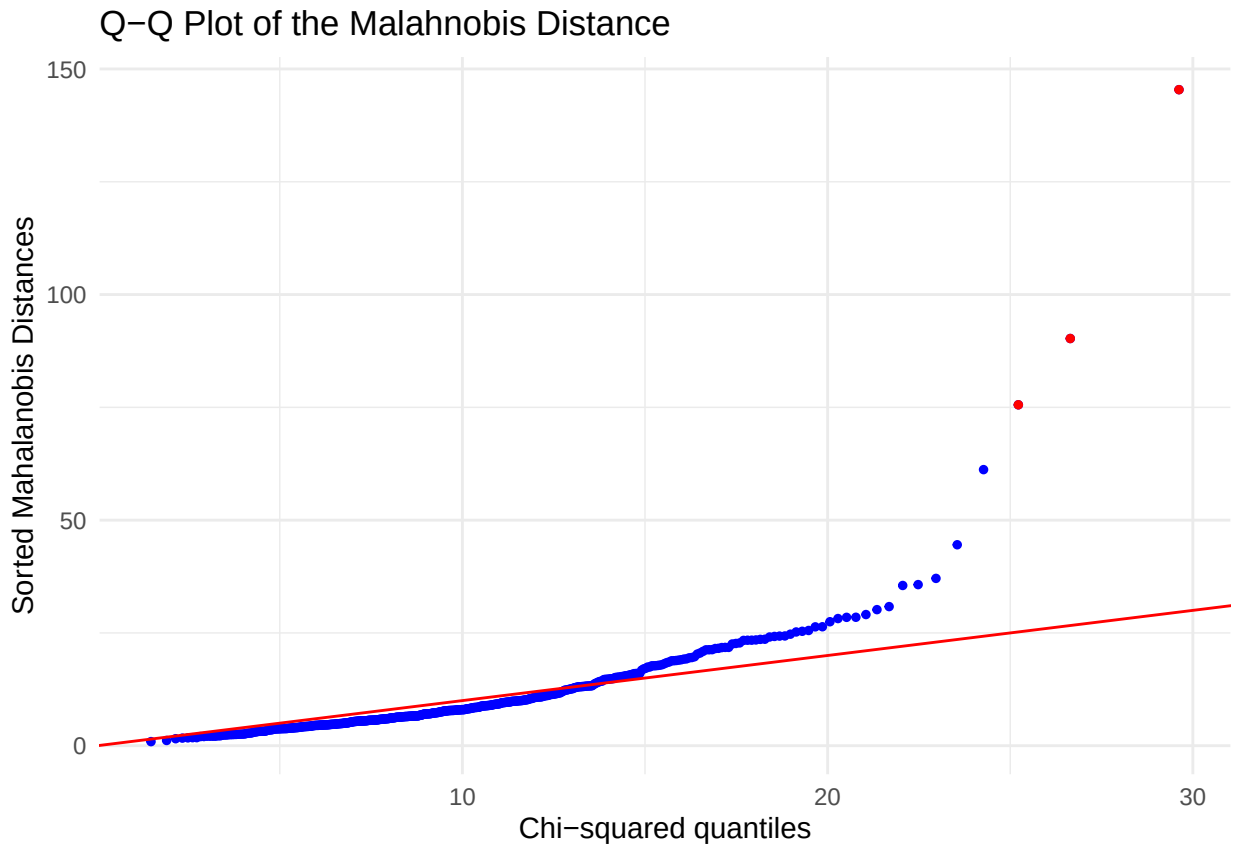
As we can see, the values of those observations in the variables which are not *crim* are almost always near the corresponding boxes (which represents the range 1-3 quartile) and therefore not so extreme.

On the other hand, we also need to take in consideration the structure of dependencies of our data in order to say (with confidence) that our univariate outliers are also multivariate outliers.

To do that, let's see the positions of those outlier in the chi-square Q-Q plot of the squared Mahalanobis

distances.

```
r_names_out = as.character(Outliers[,2])

ggplot(cbind(chi2_quantiles, sorted_d), aes(x = chi2_quantiles, y = sorted_d)) +
  geom_point( color = "blue", size = 1) +
  geom_point(data = cbind(chi2_quantiles, sorted_d)[r_names_out,],
             aes(x = chi2_quantiles, y=sorted_d),color = "red", size = 1) +
  geom_abline(slope = 1, intercept = 0, color = "red") +
  theme_minimal()+
  labs(title = "Q-Q Plot of the Malahnobis Distance",
       x = "Chi-squared quantiles", y = "Sorted Mahalanobis Distances")
```



As we can see, the outliers that we found before are also clearly multivariate outliers, since they corresponds to the three observations with the highest Mahalanobis distances.

Point 5

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

With the following code we perform the PCA on our standardized dataset. Furthermore we also print, for each principal component, it's own standard deviation, proportion and cumulative proportion of explained variability.

```
pca = prcomp(X, center = T, scale. = T)
summary(pca)
```

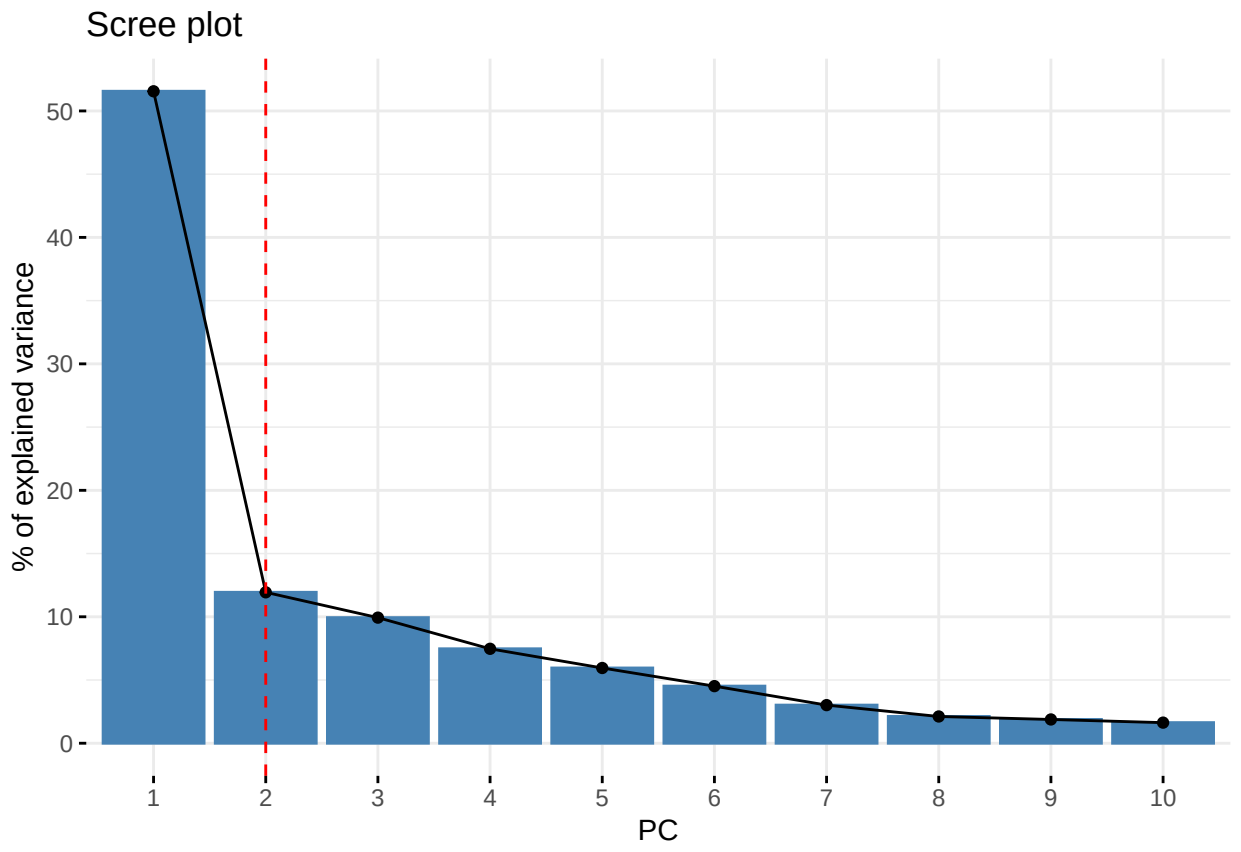
```
## Importance of components:
##                            PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.2706 1.0925 0.99663 0.86418 0.77144 0.67202 0.54922
## Proportion of Variance 0.5155 0.1194 0.09933 0.07468 0.05951 0.04516 0.03016
## Cumulative Proportion  0.5155 0.6349 0.73422 0.80890 0.86841 0.91358 0.94374
##                            PC8     PC9    PC10
## Standard deviation     0.46001 0.43367 0.40364
## Proportion of Variance 0.02116 0.01881 0.01629
## Cumulative Proportion  0.96490 0.98371 1.00000
```

To decide how many components to retain there is not a precise criteria, but there exist some "rules of thumb" which may help in the decision process.

One method consist in taking the components which (together) explains the $70 - 80\%$ of the total variability of the data. As we can see from the third row of the summary of our PCA, the first 3 components explains (more or less) the 73% of the total variability of the original data and therefore they seem enough to achieve a satisfactory lower-dimensional representation. We could also take the first 4 components, which in total would explain the 80% of the total variability, in order to have an even more accurate representation, but in this way we would loose the possibility to visualize graphically the results of the PCA.

The second method consists in the construction of the so-called "screeplot", which would indicate the number of components to retain in correspondence of the "elbow".

```
fviz_eig(pca, xlab = "PC", ylab = "% of explained variance")+
  geom_vline(xintercept = 2, linetype = "dashed", color = "red")
```



As we can see from the above figure, this second method would advise us to retain just the first two components, since the "elbow" appears in corrispondence of the second eigenvalue. On the other hand, since the first

two explain only the 63% of the total variability, we still think it would be better to retain the first three components.

The third method consists in taking all the components with an eigenvalue greater or equal than one. Let's then print the eigenvalues of our principal components.

```
pca$sdev^2
```

```
##  [1] 5.1553996 1.1935312 0.9932782 0.7468114 0.5951263 0.4516087 0.3016373
##  [8] 0.2116125 0.1880704 0.1629244
```

As we can see, the eigenvalues of the first three PC are all ≥ 1 (the third one is almost equal to one) so this method tells us to retain the first three PC for a good lower-dimensional representation.

In conclusion, considering all the three method, it would be advisable to chose the first three components, that strike a balance between a good reduction of dimensional while still explaining a great amount of variability.

Point 6

Interpret the first 3 principal components by selecting for each principal component the variables with correlation greater (in absolute value) than 0.4 with that principal component.

In order to interpret properly the first 3 PC we have to look at their correlations with our original variables.

```
corr_CP = cor(X_scaled, pca$x[,1:3])
corr_CP = cbind(corr_CP, Explained_var = apply(corr_CP^2, 1, sum) )


round(corr_CP, 3)
```

```
##             PC1    PC2    PC3 Explained_var
## crim      0.620  0.114 -0.485         0.632
## indus     0.867 -0.086  0.114         0.773
## nox       0.857 -0.322  0.114         0.851
## rm       -0.493 -0.634 -0.382         0.790
## age       0.788 -0.311  0.250         0.780
## dis      -0.792  0.397 -0.183         0.819
## tax       0.819  0.052 -0.262         0.743
## ptratio   0.482  0.581 -0.058         0.573
## black    -0.511  0.020  0.624         0.651
## lstat     0.793  0.270  0.169         0.731
```

The first 3 columns of the above output shows the correlation of each original variable with the first 3 PC.

The fourth column shows instead the proportion of variability of each variable explained by the first 3 PC. As we can see, there is only one variable (ptratio) whose explained variability is under the 0.6 (probably because it has low correlation with the other variables), and this confirm the fact that the first 3 PCs achieve a satisfactory lower-dimensional representation of our data.

As we can see from the table, the first PC has a high correlation (in module greater than 0.4) with all the 10 original variables (it is not so strange since it explains alone more than the 50% of the total variability of our data).

A pratical interpretation of the first principal component is that the PC1 is a general quality indicator (the lower the better) of the considered housing. In particular, an observation with a high PC1 score generally indicates an industrial area, leading to significant pollution levels. Additionally, such areas are subject to high property tax burdens, consist predominantly of old buildings, and are characterized by elevated crime rates. The population primarily comprises low-income residents, and the neighborhood often suffers from a

shortage of qualified teachers. These areas are also typically far from major employment districts, have low presence of black communities and are characterized by smaller housing units.

The second component has high positive correlation with the variable *ptratio* and a high negative correlation with the variable *rm* (the lower the better). Thus, we can interpret the second principal component as quality of housing and schooling indicator for an area since richer area will have more room per dwelling and more teacher for each pupil. For instance, an observation with a high score of PC2 corresponds in general to an area characterized by relatively small houses and with shortage of teachers.

The third principal component has a high positive correlation with the variable *black* and a significant negative correlation with *crim* (the higher the better). Therefore, we can think of PC3 as an indicator of areas with a high black population and with a low crime rate. So, an observation with high score in the third component is generally represented by a multicultural area with low rate of criminality.

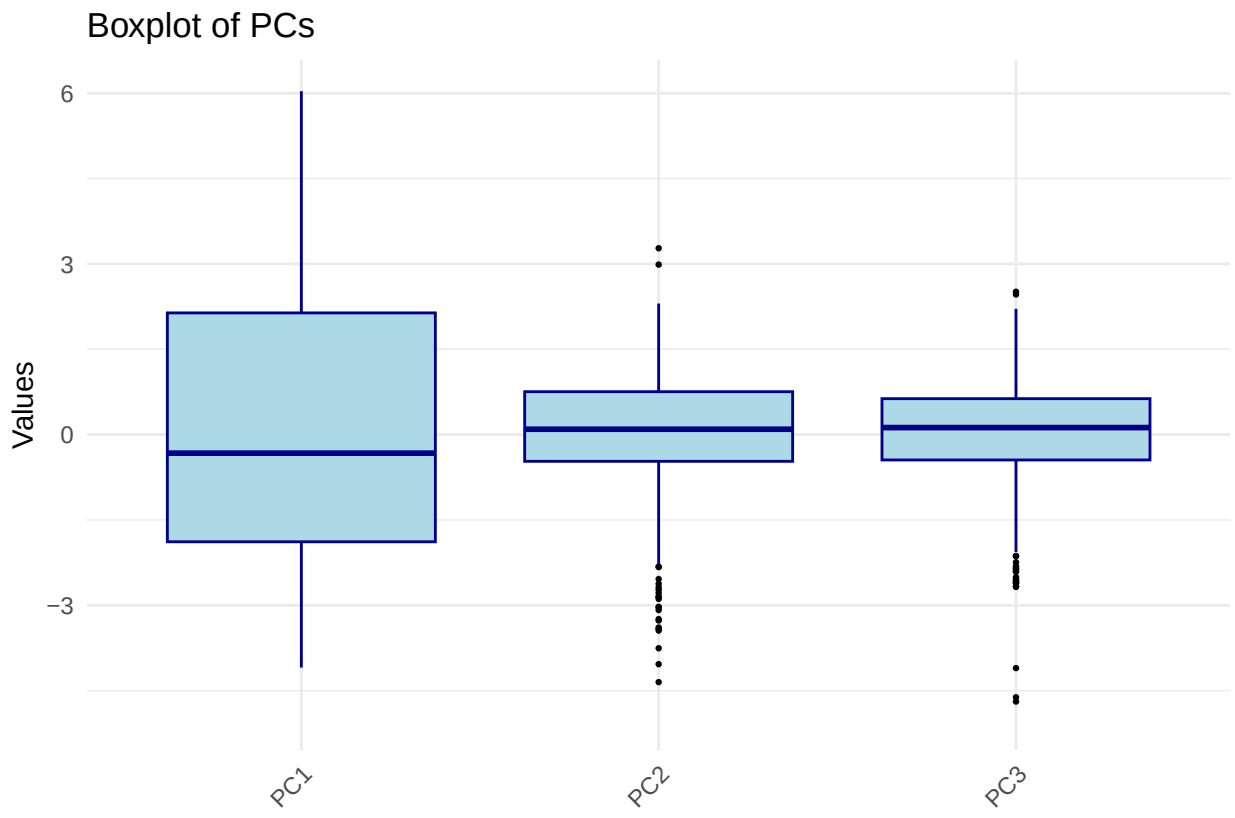
We can also extract those variable with a high correlation with each PC with the following code.

```
corr_CP = cor(X_scaled, pca$x[,1:3])
result = apply(corr_CP, 2, function(col) rownames(corr_CP)[abs(col) > 0.4])

# Print the results
print(result)
```

```
## $PC1
##  [1] "crim"    "indus"   "nox"     "rm"      "age"     "dis"     "tax"
##  [8] "ptratio" "black"   "lstat"
##
## $PC2
## [1] "rm"      "ptratio"
##
## $PC3
## [1] "crim"  "black"
```

Furthermore we show the boxplot of the three principal components, and as we can predict from the theory the variance of each components goes down (since they explain less variance). We also notice that the first principal component has no univariate outlier, but we can see the second has a few, to be more specific we can interpret the negative outliers as the rich part of the city while the positive ones represent the poorer parts. The third also has a few outliers in particular we can interpret the positive outlier as the more diverse and/or safer part of town, while the negative ones the more segregated and/or more dangerous part of town.

```
X <- tibble(as.data.frame(pca$x)[,1:3])

df_long <- X %>%
  pivot_longer(cols = everything(),
               names_to = "Variabile",
               values_to = "Valore")

ggplot(df_long, aes(x = Variabile, y = Valore)) +
  geom_boxplot(fill = "lightblue", color = "darkblue", outlier.colour = "black",
               outlier.shape = 16, outlier.size = 0.8) +
  coord_cartesian(ylim = c(-5, max(df_long$Valore))) +
  labs(title = "Boxplot of PCs", x="",  y = "Values") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Point 7

Describe the 3 outliers identified in point 2. in terms of the first 3 principal components.

To describe the three outliers identified in point (2) in terms of the first 3 PC we first look at the standardized values of those outliers for our original variables.

```
X_scaled[Outliers[,2],]
```

```
##         crim    indus      nox         rm       age        dis      tax
## 381 9.924110 1.014995 1.003687  0.9726003 0.8286338 -1.1295680 1.529413
## 419 8.128839 1.014995 1.072726 -0.4663057 1.1163897 -0.9462094 1.529413
## 406 7.476247 1.014995 1.193543 -0.8562763 1.1163897 -1.1253414 1.529413
##       ptratio      black     lstat
## 381 0.8057784  0.4406159 0.6381316
## 419 0.8057784 -3.7266503 1.1156516
## 406 0.8057784  0.3099404 1.4461347
```

We immediately notice that all the three outliers exhibit a particularly high value in the variable *crim*, while the values for all the other variables are not particularly extreme.

In order to give an interpretation of these outliers using the 3 PCs we print their score:

```
pca_scores = as.data.frame(pca$x)
pca_scores[Outliers[,2], 1:3]
```

```
##          PC1       PC2       PC3
## 381 4.792497 0.1171367 -4.616393
```

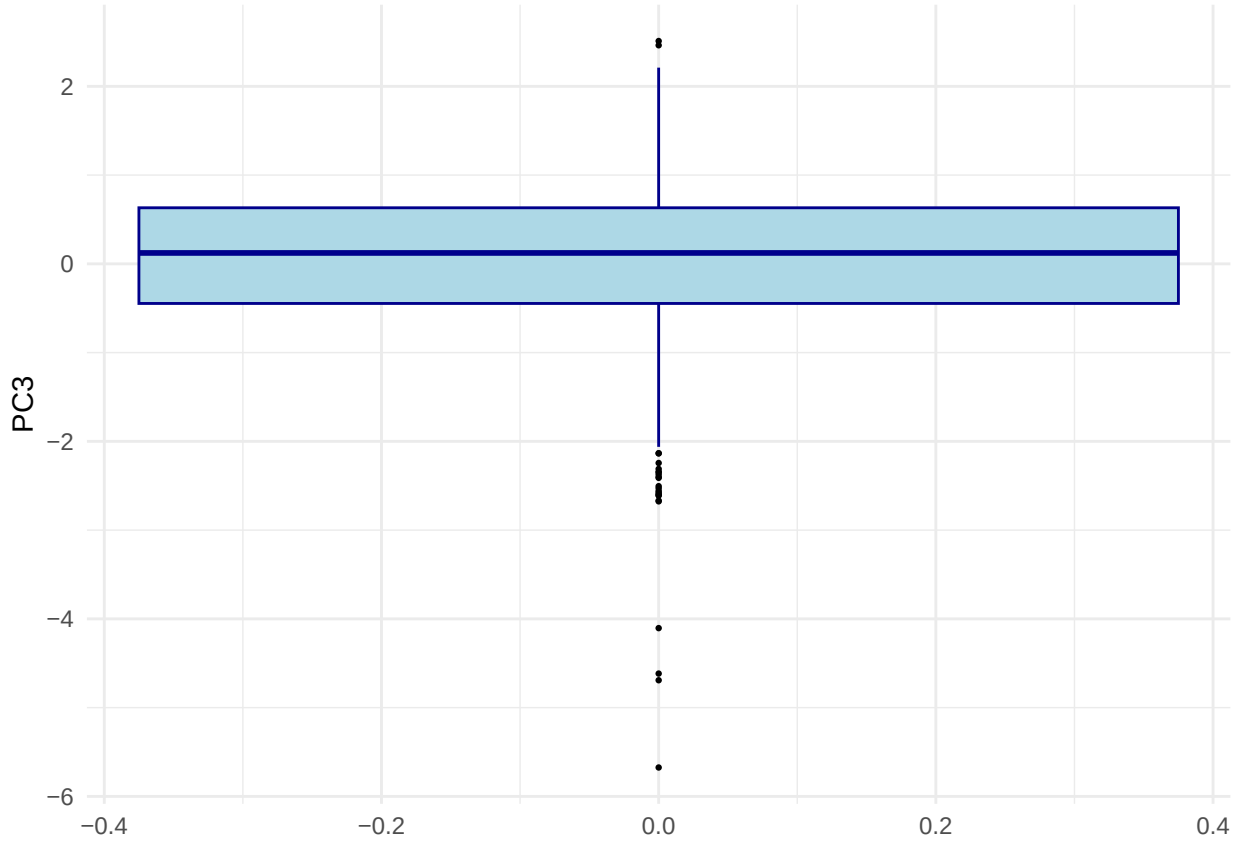
```
## 419 5.780708 0.7698542 -5.674614
## 406 5.003244 0.9833960 -2.576814
```

So, as we can see in the output above, all the three outliers have simultaneously high PC1 score and low PC3 score. This suggests that the areas represented by the three outliers found before are probably the kind of areas with heavily industrialized areas with significant pollution levels with higher rate of criminality and generally lower black population.

PC2 is fairly average and from this we can deduce that such areas have average quality schooling and average quality home.

Regarding PC3 we know that all of the three observations are also univariate outliers for PC3 .

```
X[,3] %>%
  ggplot(aes(y = PC3)) +
  geom_boxplot(fill = "lightblue", color = "darkblue",
               outlier.colour = "black", outlier.shape = 16, outlier.size = 0.8) +
  theme_minimal()
```



This is due to the fact that they are also an univariate outlier in the variable crime and the variable black doesn't compensate for it, so we know for certain that these areas aren't extremely diverse.

With the following code we also created a 3D view of the graph defined by the first 3 PC, in which we highlighted in red the three outliers found before.

```
outlier_pcs <- pca_scores[Outliers[,2], 1:3]
outlier_pcs$Index <- Outliers[,2]

plot_ly(data = pca_scores, x = ~PC1, y = ~PC2, z = ~PC3,
```

```
      type = "scatter3d", mode = "markers",
      marker = list(size = 3, color = 'gray', opacity = 0.5)) %>%
add_trace(data = outlier_pcs, x = ~PC1, y = ~PC2, z = ~PC3,
          type = "scatter3d", mode = "markers+text",
          marker = list(size = 6, color = 'red'),
          text = ~Index,
          textposition = "top center") %>%
layout(title = "3D PCA Plot with Outliers Highlighted",
       scene = list(xaxis = list(title = "PC1"),
                    yaxis = list(title = "PC2"),
                    zaxis = list(title = "PC3")))
```

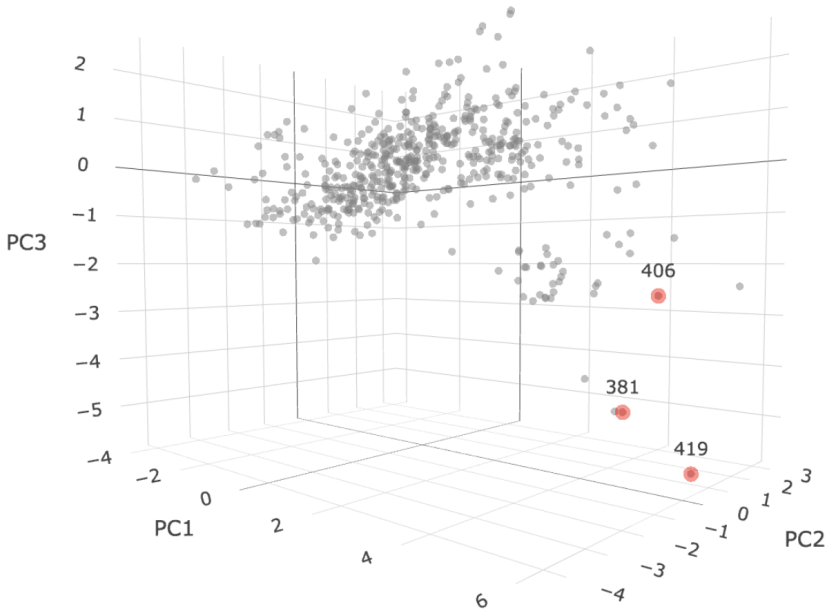


Figure 1: *
3D PCA Plot with Outliers Highlighted