Problem Set 1

Palagi Valerio, Pezzetti Lucia, Testa Federico

Exercise 1

The data set *state.x77* in the package *datasets* contains 8 variables:

- *Population*: population estimate as of July 1, 1975
- *Income*: per capita income (1974)
- *Illiteracy*: illiteracy (1970, percent of population)
- *Life Exp*: life expectancy in years (1969–71)
- *Murder*: murder and non-negligent manslaughter rate per 100,000 population (1976)
- *HS Grad*: percent high-school graduates (1970)
- *Frost*: mean number of days with minimum temperature below freezing (1931–1960) in capital or large city
- *Area*: land area in square miles

```
head(state.x77)
```

```
           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
```

```
colnames(state.x77)
```

```
[1] "Population" "Income"     "Illiteracy" "Life Exp"   "Murder"
[6] "HS Grad"    "Frost"      "Area"
```

We can immediately notice that state.x77 is not in the form of a data frame, so we first of all need to coerce it into a data frame:

```
st <- as.data.frame(state.x77);
```

We perform also some other useful transformations of the variables. In particular, we remove the spaces inside the variables names by renaming *Life Exp* with *Life.Exp* and *HS Grad* with *HS.grad*. Moreover we add an additional variable, called *Density*, that accounts for the population density.

```
names(st)[4] = "Life.Exp";
names(st)[6] = "HS.Grad";
st[,9] = st$Population * 1000 / st$Area;
colnames(st)[9] = "Density";

names(st)
```

```
[1] "Population" "Income"     "Illiteracy" "Life.Exp"   "Murder"
[6] "HS.Grad"    "Frost"      "Area"       "Density"
```

```
n <- nrow(st)
p <- ncol(st)
```

1.1 Compute the correlation matrix and comment on the most relevant relationships among variables (up to 10).

The simplest way to compute the correlation matrix, let's call it $\mathbf{R}$, is by means of the build-in function *cor()*. We round the matrix at the third decimal.

```
           Population Income Illiteracy Life.Exp Murder HS.Grad  Frost   Area
Population      1.000  0.208      0.108   -0.068  0.344  -0.098 -0.332  0.023
Income          0.208  1.000     -0.437    0.340 -0.230   0.620  0.226  0.363
Illiteracy      0.108 -0.437      1.000   -0.588  0.703  -0.657 -0.672  0.077
Life.Exp       -0.068  0.340     -0.588    1.000 -0.781   0.582  0.262 -0.107
Murder          0.344 -0.230      0.703   -0.781  1.000  -0.488 -0.539  0.228
HS.Grad        -0.098  0.620     -0.657    0.582 -0.488   1.000  0.367  0.334
```

```
Frost          -0.332  0.226     -0.672    0.262 -0.539   0.367  1.000  0.059
Area            0.023  0.363      0.077   -0.107  0.228   0.334  0.059  1.000
Density         0.246  0.330      0.009    0.091 -0.185  -0.088  0.002 -0.341
           Density
Population   0.246
Income       0.330
Illiteracy   0.009
Life.Exp     0.091
Murder      -0.185
HS.Grad     -0.088
Frost        0.002
Area        -0.341
Density      1.000
```

Now, since the correlation matrix is symmetrical and we are interested in studying how the variables are correlated among each other, we can reduce our attention only on the upper triangular part of the correlation matrix and then look at the highest correlations in terms of absolute value. We will then analyze the most relevant relationships. As a first step we consider the first ten highest (in term of absolute value) correlations.

```
    First_var Second_var Correlation
1    Life.Exp     Murder      -0.781
2  Illiteracy     Murder       0.703
3  Illiteracy      Frost      -0.672
4  Illiteracy    HS.Grad      -0.657
5      Income    HS.Grad       0.620
6  Illiteracy   Life.Exp      -0.588
7    Life.Exp    HS.Grad       0.582
8      Murder      Frost      -0.539
9      Murder    HS.Grad      -0.488
10     Income Illiteracy      -0.437
```

By looking at these correlations we decide to focus on the first 8 ones. The reasons behind this choice are related to the fact that correlations below 0.5 begins to be too weak and, more importantly, a relevant part of the interpretation of the two latter correlations follows naturally from the previous eight, so it is not particularly worthy of attention.

The first clear-cut correlation is negative and between *Life.Exp* and *Murder* (-0.781). Although at first glance it may seem natural to interpret this relationship (a higher number of murders clearly reduce life expectancy), asserting that the number of murders have, directly, such a huge impact on the life span seems too extreme. To better clarify what we mean let's consider the regression of *Life.Exp* on *Murders*:

```r
lm(data = st, Life.Exp~Murder)
```

```
Call:
lm(formula = Life.Exp ~ Murder, data = st)

Coefficients:
(Intercept)       Murder
    72.9736      -0.2839
```

What we get is that if we increase the rate of murders in 100 000 inhabitants of ten unit, we have a decrease of 2.8 years in the average life expectancy of the state. Such a huge reduction seems unjustified if we only associate it to the previous "natural" interpretation. Instead, a possible explanation of this could be that the states in which there is a higher murder rate are also the states with a worse situation of violence and widespread criminality, as a consequence inhabitants of these states live in a more dangerous and stressful environment and this is reflected in a lower life expectancy.

The greatest positive correlation among the variables is between Illiteracy and Murder (0.703). This tells us that states with a higher percentage of illiterate individuals tend to have a higher murder rate, and viceversa. This relationship seems particularly interesting, as it's the second highest in absolute value but doesn't have an immediate "natural" interpretation. A possible one could be that states with a higher illiteracy rate tend to have a lower overall level of education, which in turn could be related to more underdeveloped areas in a broad sense (economically, culturally, ...).

We try to contextualize this interpretation and these correlations in the particular geo-political situation of the US and the socio-economical divide between Northern States and Southern States at the time, in order to validate our interpretation and further elaborate on it. In general, Northern States were historically more industrialized, urbanized and educated whereas Southern States had a more rural economy, a lower education level and a widespread possession of firearms. In this framework, illiteracy seems to be an effective indicator of overall (under)development and is "geographically correlated" to murder rates (through the North-South divide). Indeed, let's look at the states with values of *Murder* and *Illiteracy* above the third quartile:

```r
summary(st)
```

```
   Population        Income       Illiteracy       Life.Exp
 Min.   :  365   Min.   :3098   Min.   :0.500   Min.   :67.96
 1st Qu.: 1080   1st Qu.:3993   1st Qu.:0.625   1st Qu.:70.12
 Median : 2838   Median :4519   Median :0.950   Median :70.67
```

```
 Mean   : 4246   Mean   :4436   Mean   :1.170   Mean   :70.88
 3rd Qu.: 4968   3rd Qu.:4814   3rd Qu.:1.575   3rd Qu.:71.89
 Max.   :21198   Max.   :6315   Max.   :2.800   Max.   :73.60
     Murder          HS.Grad          Frost             Area
 Min.   : 1.400   Min.   :37.80   Min.   :  0.00   Min.   :  1049
 1st Qu.: 4.350   1st Qu.:48.05   1st Qu.: 66.25   1st Qu.: 36985
 Median : 6.850   Median :53.25   Median :114.50   Median : 54277
 Mean   : 7.378   Mean   :53.11   Mean   :104.46   Mean   : 70736
 3rd Qu.:10.675   3rd Qu.:59.15   3rd Qu.:139.75   3rd Qu.: 81163
 Max.   :15.100   Max.   :67.30   Max.   :188.00   Max.   :566432
    Density
 Min.   :  0.6444
 1st Qu.: 25.3352
 Median : 73.0154
 Mean   :149.2245
 3rd Qu.:144.2828
 Max.   :975.0033
```

```
filter(st, Illiteracy > 1.575, Murder > 10.675)
```

	Population	Income	Illiteracy	Life.Exp	Murder	HS.Grad	Frost	Area	Density
Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708	71.29053
Georgia	4931	4091	2.0	68.54	13.9	40.6	60	58073	84.91037
Louisiana	3806	3545	2.8	68.76	13.2	42.2	12	44930	84.70955
Mississippi	2341	3098	2.4	68.09	12.5	41.0	50	47296	49.49679
North Carolina	5441	3875	1.8	69.21	11.1	38.5	80	48798	111.50047
South Carolina	2816	3635	2.3	67.96	11.6	37.8	65	30225	93.16791
Tennessee	4173	3821	1.7	70.11	11.0	41.8	70	41328	100.97271
Texas	12237	4188	2.2	70.90	12.2	47.4	35	262134	46.68223

As expected, they all are Southern rural states.

Similarly, if we consider all the states whose *Murder* rate is above the third quartile and whose life expectancy is below the first quartile, we obtain:

```
filter(st, Murder > 10.675, Life.Exp < 70.12)
```

```
               Population Income Illiteracy Life.Exp Murder HS.Grad Frost
Alabama              3615   3624        2.1    69.05   15.1    41.3    20
Alaska                365   6315        1.5    69.31   11.3    66.7   152
Georgia              4931   4091        2.0    68.54   13.9    40.6    60
Louisiana            3806   3545        2.8    68.76   13.2    42.2    12
Mississippi          2341   3098        2.4    68.09   12.5    41.0    50
Nevada                590   5149        0.5    69.03   11.5    65.2   188
North Carolina       5441   3875        1.8    69.21   11.1    38.5    80
South Carolina       2816   3635        2.3    67.96   11.6    37.8    65
Tennessee            4173   3821        1.7    70.11   11.0    41.8    70
                 Area     Density
Alabama         50708  71.2905261
Alaska         566432   0.6443845
Georgia         58073  84.9103714
Louisiana       44930  84.7095482
Mississippi     47296  49.4967862
Nevada         109889   5.3690542
North Carolina  48798 111.5004713
South Carolina  30225  93.1679074
Tennessee       41328 100.9727062
```

As we can see, with the only exception of Alaska (whose situation, as we will see and as expected, is exceptional and not comparable to the one of the other US states), this results support our speculations.

To proceed we notice that, quite surprisingly, *Frost* is significantly negatively correlated with *Illiteracy* (-0.672). This relationship may seem quite counterintuitive as in states with a very harsh climate (let's for example think at the Alaska) we expect lower civilization, where the term civilization is used in its broadest sense. Nevertheless, if we contextualize it, then this relationship appears nearly natural: the states with a warmer climate are the Southern ones and therefore, as already addressed, the ones with, on average, a lower level of education. To a further confirm, we can filter our data frame by retaining only the states with values for *Frost* below the first quartile and for *Illiteracy* above the third one.

```
filter(st, Illiteracy > 1.575, Frost < 66.25)
```

```
               Population Income Illiteracy Life.Exp Murder HS.Grad Frost
```

Alabama	3615	3624	2.1	69.05	15.1	41.3	20
Arizona	2212	4530	1.8	70.55	7.8	58.1	15
Arkansas	2110	3378	1.9	70.66	10.1	39.9	65
Georgia	4931	4091	2.0	68.54	13.9	40.6	60
Hawaii	868	4963	1.9	73.60	6.2	61.9	0
Louisiana	3806	3545	2.8	68.76	13.2	42.2	12
Mississippi	2341	3098	2.4	68.09	12.5	41.0	50
South Carolina	2816	3635	2.3	67.96	11.6	37.8	65
Texas	12237	4188	2.2	70.90	12.2	47.4	35

	Area	Density
Alabama	50708	71.29053
Arizona	113417	19.50325
Arkansas	51945	40.61989
Georgia	58073	84.91037
Hawaii	6425	135.09728
Louisiana	44930	84.70955
Mississippi	47296	49.49679
South Carolina	30225	93.16791
Texas	262134	46.68223

The following correlation we analyze in some sense reflect what we have already implicitly addressed in the above interpretations. Indeed, the fact that *Illiteracy* and *HS.Grad* are negatively correlated (-0.657) can be viewed as a consequence of the fact that both a high percentage of illiterates and non-graduates reflect an overall cultural framework that gives less "value" to education.

Another straightforward relationship is the positive correlation between *HS.Grad* and *Income* (0.620). This can be explained by enhancing both a direct and an indirect linkage. Specifically, it is reasonable that high-school graduates are usually more highly-qualified and therefore can apply to jobs with higher wages and, at the same time, states with a large share of graduates are more likely to invest more on the educational system, being more urbanized and attractive for large firms that can afford to pay more their employees.

Taking into account everything we have already discussed above, the two next correlations, *Illiteracy-Life.Exp* (-0.588) and *HS.Grad-Life.Exp* (0.582), seem to have a well-grounded explanation. Explicitly, we have already argued that both *Illiteracy* and *HS.Grad* may be considered as global indicators of the level of education of a states, which is more inadequate in the Southern states. Furthermore, these very states are also the ones with a higher murder rate and so, as justified above, a more widespread criminality. If we also recall our diagnosis of the correlation between *Murder* and *Life.Exp*, our impression is that the historical background support the fact that states with a significant level of illiteracy or, in general, a low percentage of high-school graduates are also the ones with a shorter life expectancy. In addition to this, we may also guess how these aspects interact with each other: more educated individuals might

be more conscious about their lifestyles, pay more attention to healthy habits, have jobs that are less physically demanding and are more likely to live in urban areas where it is more easy to have access to healthcare facilities.

Finally, the data reveal that *Murder* is negatively correlated with *Frost*(-0.539). Once again this linkage may be contextualized in the South-North divide of the US. States that have warmer climates also have a higher murder rate whereas states with a more harsh climate tend to have a lower number of murders.

```
st %>% filter((Murder > 10.675 & Frost < 66.25) |
                (Murder < 4.350 & Frost > 139.75)) %>% arrange(Frost)
```

	Population	Income	Illiteracy	Life.Exp	Murder	HS.Grad	Frost
Florida	8277	4815	1.3	70.66	10.7	52.6	11
Louisiana	3806	3545	2.8	68.76	13.2	42.2	12
Alabama	3615	3624	2.1	69.05	15.1	41.3	20
Texas	12237	4188	2.2	70.90	12.2	47.4	35
Mississippi	2341	3098	2.4	68.09	12.5	41.0	50
Georgia	4931	4091	2.0	68.54	13.9	40.6	60
South Carolina	2816	3635	2.3	67.96	11.6	37.8	65
Iowa	2861	4628	0.5	72.56	2.3	59.0	140
Wisconsin	4589	4468	0.7	72.48	3.0	54.5	149
Minnesota	3921	4675	0.6	72.96	2.3	57.6	160
Maine	1058	3694	0.7	70.39	2.7	54.7	161
South Dakota	681	4167	0.5	72.08	1.7	53.3	172
New Hampshire	812	4281	0.7	71.23	3.3	57.6	174
North Dakota	637	5087	0.8	72.78	1.4	50.3	186

	Area	Density
Florida	54090	153.022740
Louisiana	44930	84.709548
Alabama	50708	71.290526
Texas	262134	46.682231
Mississippi	47296	49.496786
Georgia	58073	84.910371
South Carolina	30225	93.167907
Iowa	55941	51.143169
Wisconsin	54464	84.257491
Minnesota	79289	49.452005
Maine	30920	34.217335

```
South Dakota    75955   8.965835
New Hampshire    9027  89.952365
North Dakota    69273   9.195502
```

```
st %>% filter((Murder > 10.675 & Frost > 139.75) |
                (Murder < 4.350 & Frost < 66.25)) %>% arrange(Frost)
```

```
           Population Income Illiteracy Life.Exp Murder HS.Grad Frost   Area
Washington       3559   4864        0.6    71.72    4.3    63.5    32  66570
Oregon           2284   4660        0.6    72.13    4.2    60.0    44  96184
Alaska            365   6315        1.5    69.31   11.3    66.7   152 566432
Nevada            590   5149        0.5    69.03   11.5    65.2   188 109889
              Density
Washington 53.4625207
Oregon     23.7461532
Alaska      0.6443845
Nevada      5.3690542
```

The two above data frames might help in giving a better visual representation of the correlation.

To conclude the analysis of the correlations, we can notice the fact that *Population*, *Area* and *Density* are little correlated to any other variable. This is interesting and deserve some remarks. To begin with, the geographical extension of the states is strongly related to the history of the country: the smaller eastern states correspond to the first thirteen British colonies or, more generally, to the territories colonized in the first phases of the European expansion. Conversely, the larger, "ruler-drawn" central and western states are the one colonized later during the so-called westward expansion. This suggests that *Area* is more associated to a east-west subdivision, rather than a south-north one and separates *Area* from the other variables. Similarly, both *Population* and *Density* may be regarded as "cross-sectional" variables, meaning that they are too influenced by historical, cultural and geographical features to display a clearly distinguishable pattern and this place them in a unique situation and do not allow for meaningful correlation with the other variables.

1.2 Find univariate outliers, up to 3 per variable, up to 10 in total.

To discuss the univariate outliers we start by considering the standardized values $z_{ij} = \frac{x_{ij} - \bar{x}}{s_{jj}}$ and examine them for large or small values compared to the quantiles of standard normal distribution. In particular we choose to use as threshold the $\frac{n-0.5}{n}$ normal quantile.

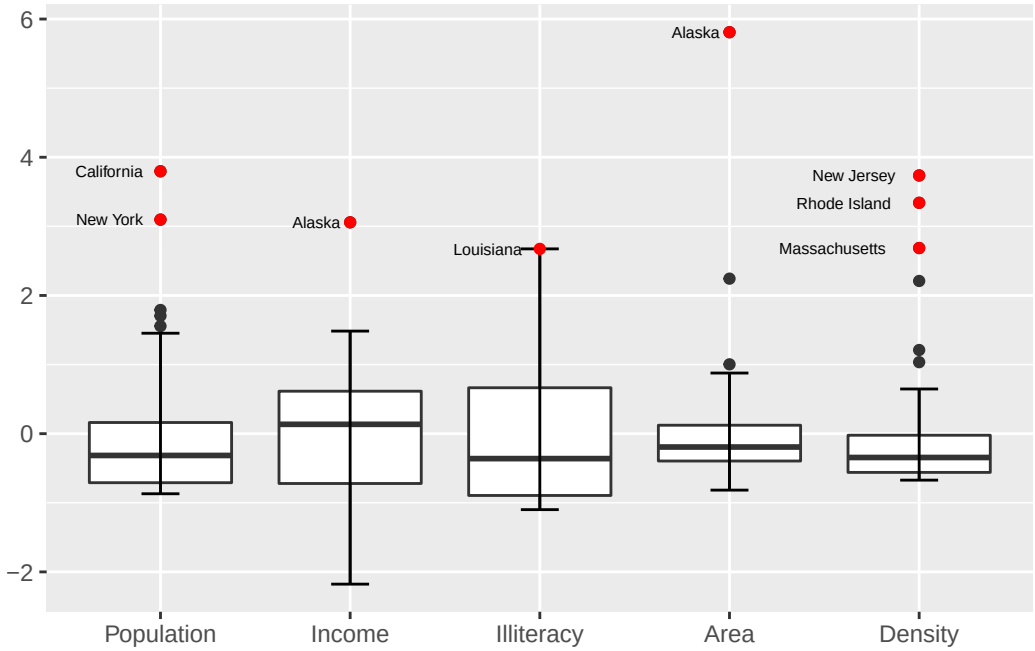
```
           Population Income Illiteracy Life.Exp Murder HS.Grad  Frost   Area
Alabama        -0.141 -1.321      1.526   -1.362  2.092  -1.462 -1.625 -0.235
Alaska         -0.869  3.058      0.541   -1.169  1.062   1.683  0.915  5.809
Arizona        -0.456  0.153      1.034   -0.245  0.114   0.618 -1.721  0.500
Arkansas       -0.479 -1.721      1.198   -0.163  0.737  -1.635 -0.759 -0.220
California      3.797  1.104     -0.115    0.619  0.792   1.175 -1.625  1.003
Colorado       -0.382  0.729     -0.771    0.880 -0.157   1.336  1.184  0.387
           Density
Alabama     -0.353
Alaska      -0.672
Arizona     -0.587
Arkansas    -0.491
California  -0.062
Colorado    -0.564

[1] 2.326348
```

We summarize the results in a data frame where The obtained potential univariate outliers with the corresponding variable are:

```
    variable       outlier
1 Population    California
2 Population      New York
3     Income        Alaska
4 Illiteracy     Louisiana
5       Area        Alaska
6    Density Massachusetts
7    Density    New Jersey
8    Density  Rhode Island
```

1.3 Make a boxplot of any variable plotting the corresponding outliers, if any, found in point 2 in red

As the above procedure shows we have eight potential univariate outliers. We can add more evidence to whether they are actually outliers by considering the boxplots of the interested variables and marking them in red. We moreover notice that, since the variables have very different scales, we consider the scaled data.



From the boxplots we can make some considerations:

- For all the potential outliers that we have identified, except for the one of *Illiteracy*, we have further evidence to support claim that they are univariate outliers.
- We have failed in identifying some other "outlier-candidates" for the variables *Population*, *Area* and *Density*. Indeed, even if the detected ones exhibits a more extreme behavior, the boxplots still suggest that these observations deviates strongly with respect to the other.
- As already said, the boxplot of the variable *Illiteracy* shows no outliers. The potential one that we have detected seems to correspond exactly to the upper acceptance limit of the boxplot, thus we may conclude that, even if the observation is certainly peculiar, there are not enough indications to consider it an outlier.

We may make an hypothesis to explain why the boxplots present so many discrepancies with respect to our predictions: the reason why we have detected a "false" outlier for illiteracy and just some of the outliers of *Population*, *Area* and *Density* may be related to an unjustified assumption of normality in the criterion used (i.e the comparison to the quantiles of order $\frac{n-0.5}{n}$ of the gaussian distribution). We'll argue more about this in the following section.

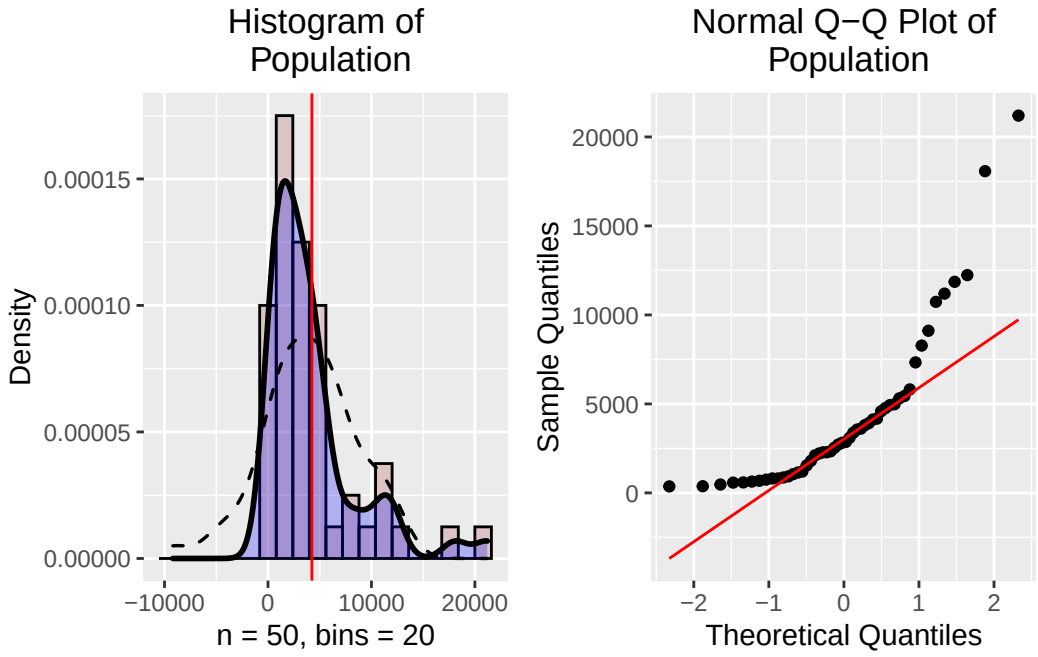
1.4 Comment about normality of each variable

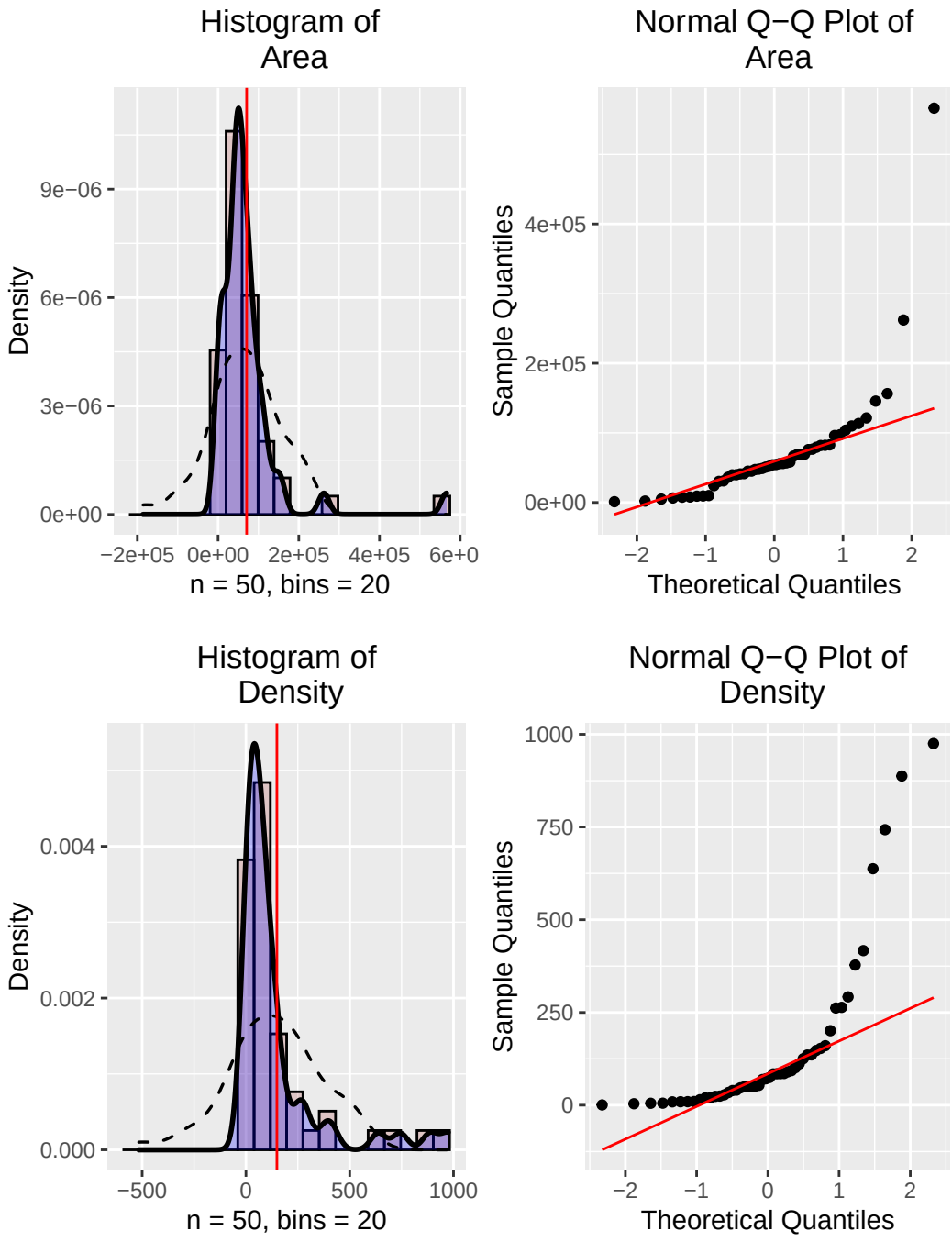
To study the normality of each variable we rely on two main tools: the histograms of the data and the QQ normal plots.

Before proceeding we recall that the distribution of a standard gaussian distribution is symmetrical around the mean and has a bell-like shape. Therefore, if the distribution of a variable is significantly skewed (to the left or to the right) it is likely not to be normally distributed. To better interpret the histograms we draw not only the histograms themselves, but also the sample density of the considered variable, the (dotted) density of a gaussian (generated by considering a normal sample of size $n = 50$ equal to the number of observations of our variables) and finally a vertical red line at position given by the sample mean.

For what concern the Q-Q plots, instead, we emphasize the fact that they play a vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other. In our case the theoretical quantile that we are going to consider are the one corresponding to the normal distribution and we plot them against the sample quantiles. If the two distributions which we are comparing are sufficiently similar, meaning that our variable is reasonably gaussian, then the points in the Q-Q plot will approximately lie on a straight line.

We start by analyzing the three variables (*Population*, *Area*, *Density*) for which we have failed to detect some upper outlier candidates.





As expected, there is little evidence to support the assumption of normality for *Population* and *Density*. They are both skewed to the right and present lower and (even more dramatically) upper quantiles that deviates from the theoretical gaussian ones. The plots also allow for a discussion on the kurtosis of the distribution: both the variables exhibit a heavy right tail

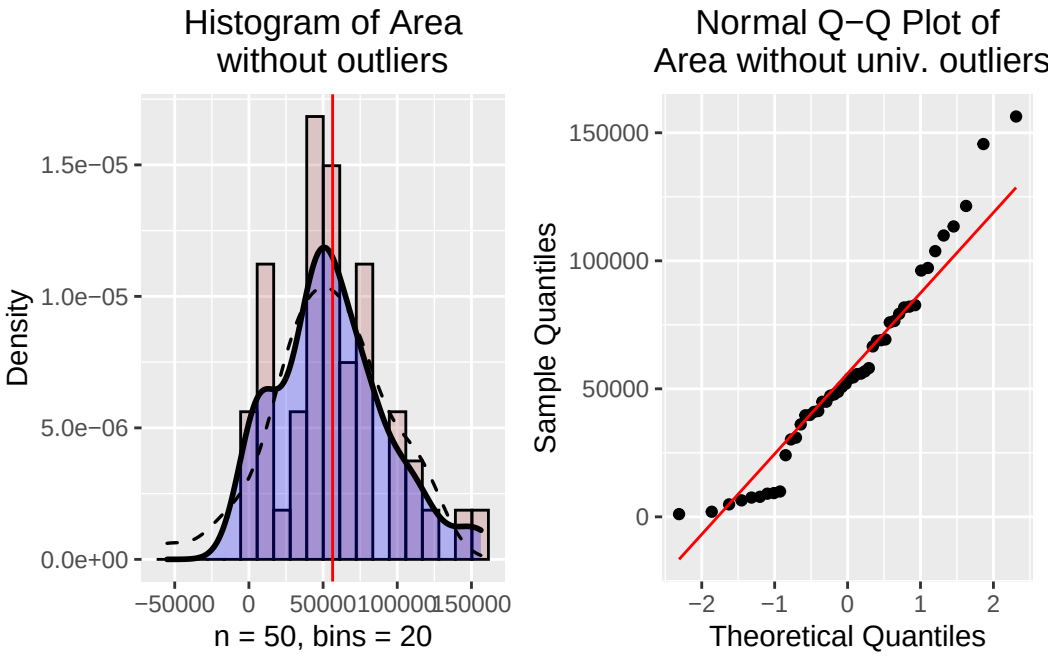
and a light left one. However, for what concern the left tail of the distribution, we think that this behavior is introduced by the lower bound these variables have (they need to be positive). Indeed, the presence of a lower bound that is "close" to the mean with respect to the total variability of the data, artificially creates the effect of a lighter-than-normal left tail.

Area presents a more delicate situation. It is still a bit skewed to the right, however (except from the last two points) the deviation from the straight line is much more suppressed and makes it difficult to safely reject the hypothesis of normality of the variable. Nevertheless, both ends of the Q-Q plots display that the tails of the sample distribution are quite different to the gaussian ones. Therefore we believe that there is no enough evidence to support the gaussianity of the variable *Area*.

Despite this, the presence of two observations that deviates significantly from the others suggests the possibility to study the normality of the variable without considering them. First of all, it is meaningful to identify those outliers, indeed we only know that the most extreme one is Alaska. We rapidly obtain that the two outliers are:

```
[1] "Alaska" "Texas"
```

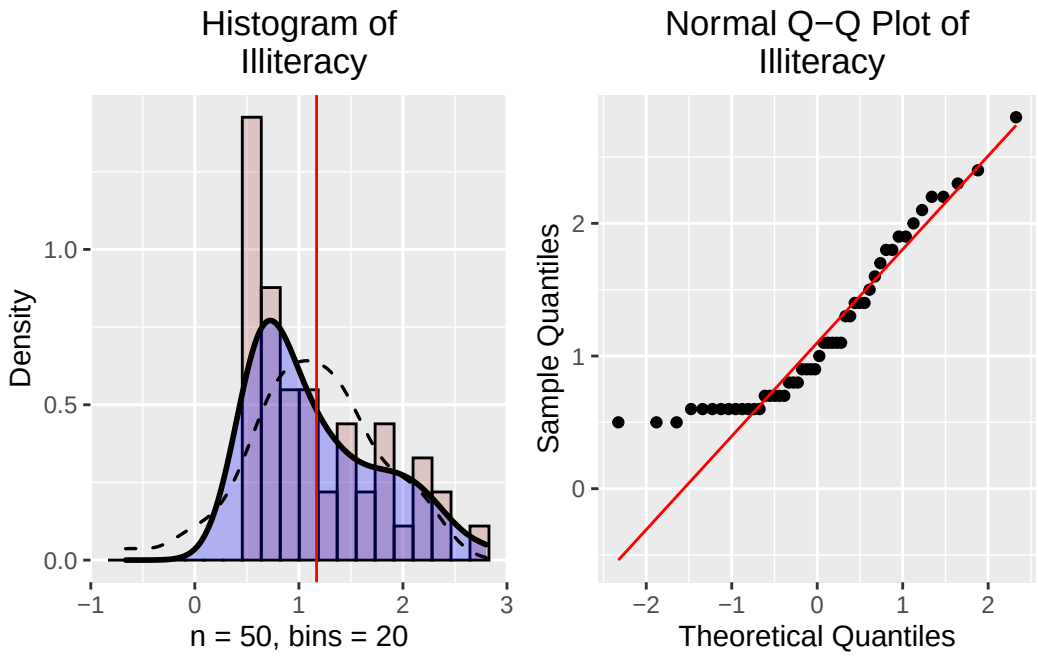
Now, we remove these values and plot the histogram and the Q-Q plot of *Area*:



Even if the Q-Q plot is closer to an acceptable one, even now there are some features that make us hesitant to not reject the hypothesis of normality. In particular, the trend of the lower quantiles suggests the possibility to have a bimodal distribution and this may make the sample

distribution meaningfully different to the Gaussian one. As a consequence, even in the absence of the outliers, we do not feel safe enough in not rejecting the gaussianity of the variable, but we suspend our judgment as we will need to make further tests.

Let's now take into account the variable *Illiteracy*. We recall that, by comparison between the standardized data and the normal quantile of order $\frac{n-0.5}{n}$, we have wrongly identified a potential outlier.

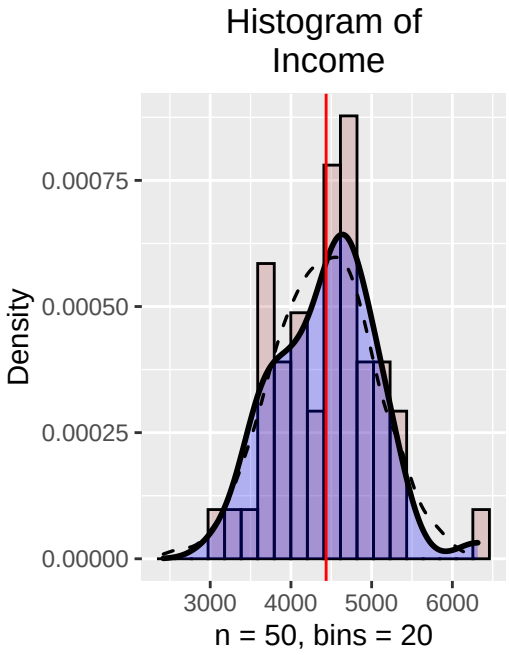


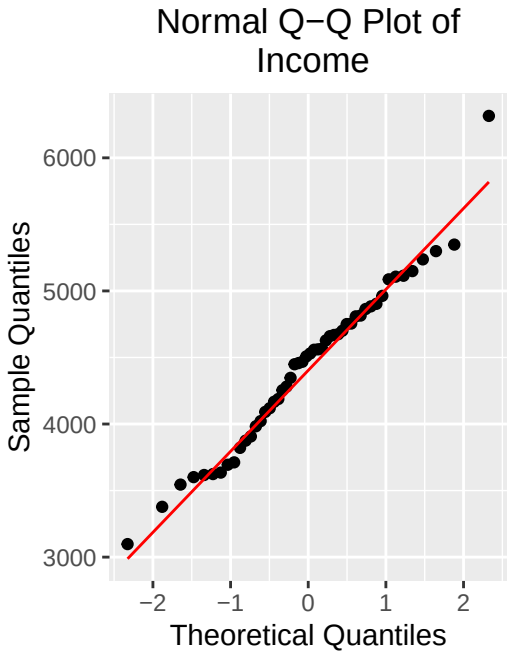
The *Illiteracy* Q-Q plot exhibits some very interesting features:

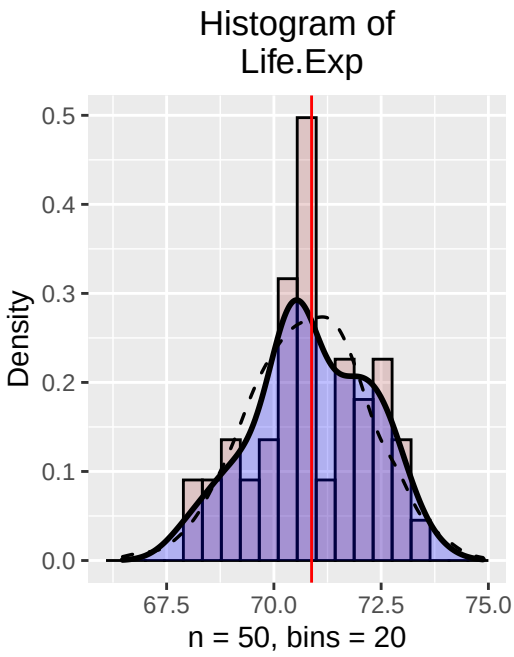
- The stairstep pattern, in which only specific, separated heights ("sample quantiles") are attained, shows the data values are discrete.
- There is a large number of values at the value of 0.6, far more than any other value. This concentration of values tends to skew the data to the right.
- Apart from this "spike" at 0.6, a closer look shows that the remaining points are initially slightly lower than the reference line (for values between 0.7 and 1.5) and then slightly greater (for values between 1.6 and 2.3) before roughly returning to the line at the end (values 2.4 and 2.8). This "curvature" indicates a certain form of non-normality.

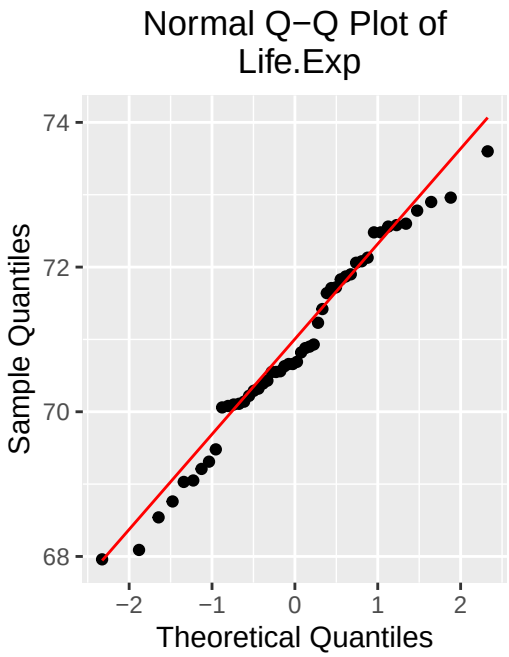
Putting everything together, even for this variable there is no enough evidence to support the normality assumption.

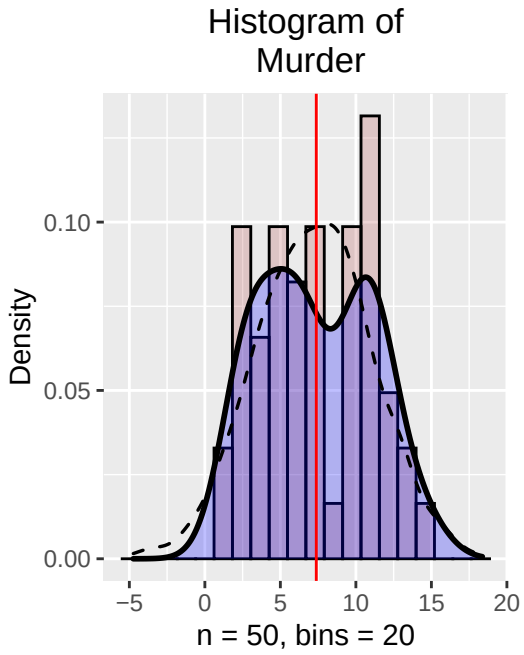
Finally we consider the remaining variables.

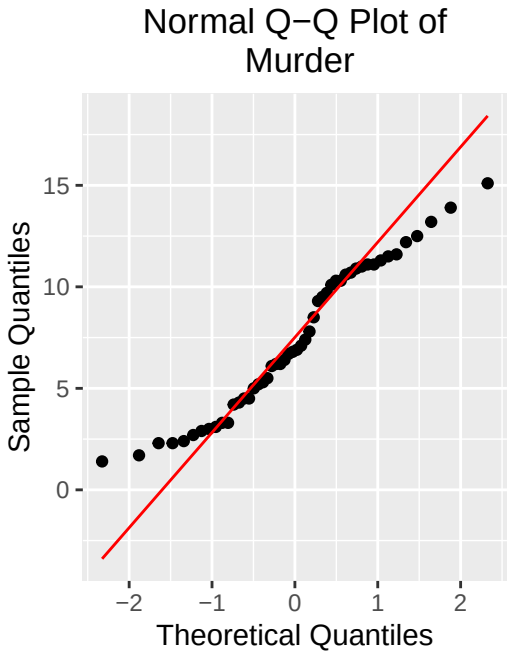

Histogram of
Income
Density
0.00075
0.00050
0.00025
0.00000
3000
4000
5000
6000
n = 50, bins = 20

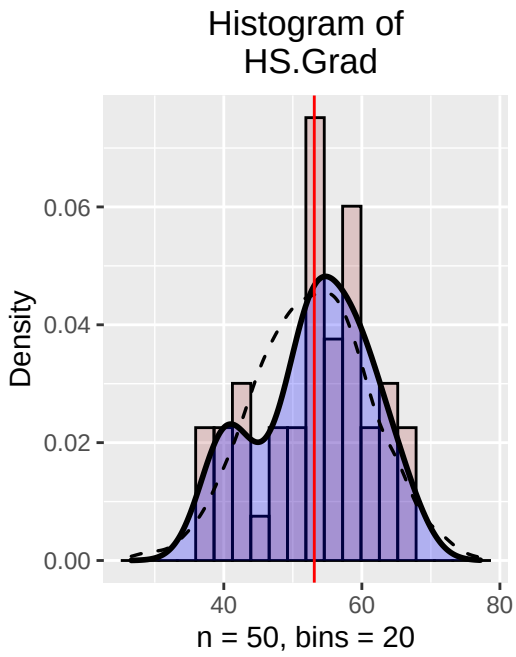

Normal Q−Q Plot of
Income
Sample Quantiles
6000
5000
4000
3000
−2
−1
0
1
2
Theoretical Quantiles

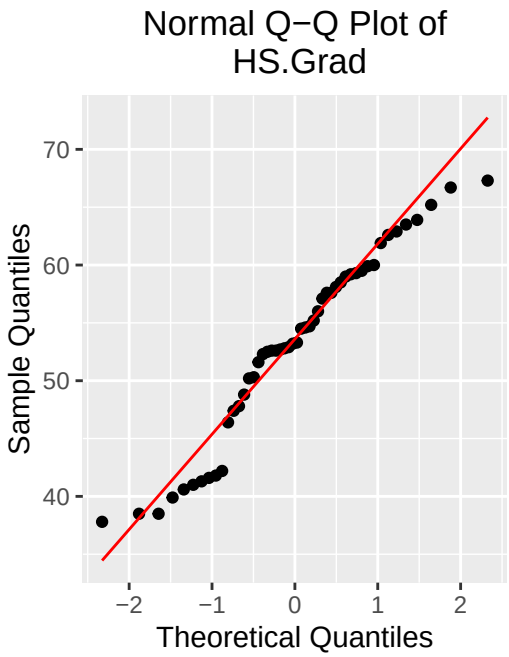

Histogram of
Life.Exp
Density
0.5
0.4
0.3
0.2
0.1
0.0
67.5
70.0
72.5
75.0
n = 50, bins = 20

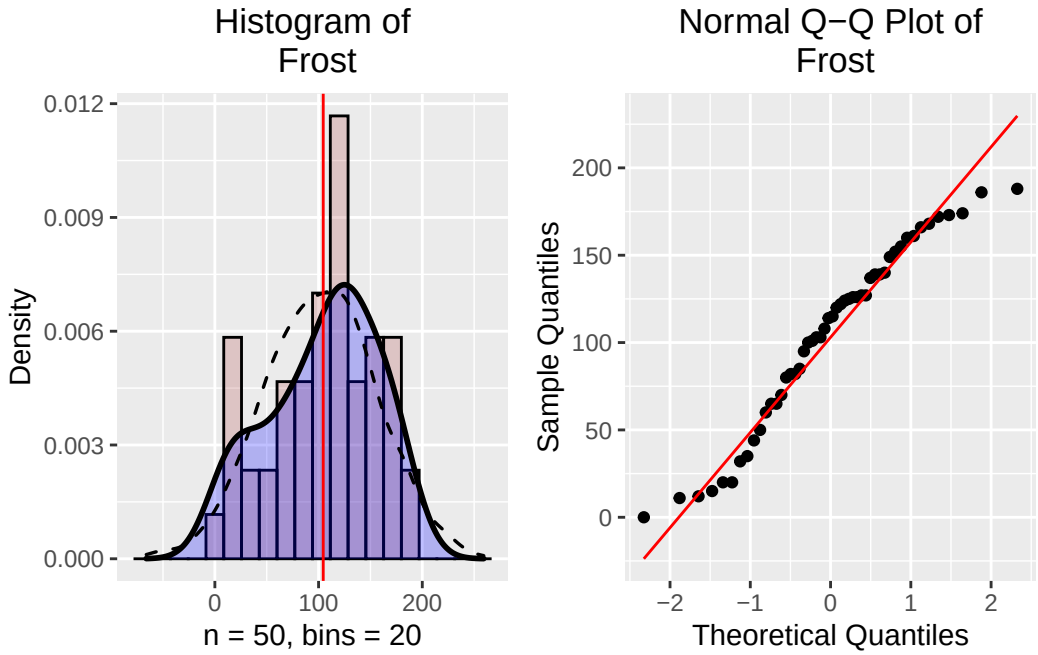

Normal Q−Q Plot of
Life.Exp
Sample Quantiles
74
72
70
68
−2
−1
0
1
2
Theoretical Quantiles


Histogram of
Murder
Density
0.10
0.05
0.00
−5
0
5
10
15
20
n = 50, bins = 20


Normal Q−Q Plot of
Murder
Sample Quantiles
15
10
5
0
−2
−1
0
1
2
Theoretical Quantiles


Histogram of
HS.Grad
Density
0.06
0.04
0.02
0.00
40
60
80
n = 50, bins = 20


Normal Q−Q Plot of
HS.Grad
Sample Quantiles
70
60
50
40
−2
−1
0
1
2
Theoretical Quantiles



For all these variables but *Murder* and *HS.Grad* there seems to be enough evidence to not reject the assumption of univariate gaussianity. We consequently focus on the two remaining variables.

The plots of *HS.Grad* are quite controversial: the jump in the Q-Q plot together with the "wave-like" trend of the sample quantiles suggests that the variable has a bimodal probability density. In addition the highest and lowest quantiles present spikes of quite similar entities followed by very thin tails. For all these reasons we are more inclined to reject the normality of *HS.Grad*.

For what concern *Murder*, similarly, there are multiple "red flags" that encourage us to reject the hypothesis of normality. Not only the sample density of *Murder* is bimodal, but it also have light tails on both sides (since the sample quantiles are higher than the theoretical ones in the left-bottom of the Q-Q plot and lower than the theoretical ones in its the right-top).

We can validate our speculations by performing a Shapiro test. The Shapiro test is used to test whether a variable is normally distributed or not. The null hypothesis states that the considered variable is normally distributed, therefore if the p-value is greater than 0.05 (standard threshold), then the normality of the data is not rejected. The results of the tests will be saved in a data frame.

```
    Variable      p.value Statistics
1 Population 1.906393e-07  0.7699920
2     Income 4.300105e-01  0.9769037
```

```
3 Illiteracy 1.396258e-04  0.8831491
4   Life.Exp 4.423285e-01  0.9772400
5     Murder 4.744626e-02  0.9534691
6    HS.Grad 4.581562e-02  0.9531029
7      Frost 5.267472e-02  0.9545618
8       Area 7.591835e-11  0.5717872
9    Density 7.262254e-10  0.6372746
```

The variables that "pass" the Shapiro test, meaning that they have a sufficiently high p-value, are:

```
shap %>% filter(p.value > 0.05)
```

```
  Variable    p.value Statistics
1   Income 0.43001048  0.9769037
2 Life.Exp 0.44232853  0.9772400
3    Frost 0.05267472  0.9545618
```

This suits and validates our previous discussion.

To conclude, we use the Shapiro test also to check if the non-gaussianity of *Area* is greatly influenced by the presence of the two outliers Alaska and Texas.

```
    Shapiro-Wilk normality test

data:  area[, 1]
W = 0.95441, p-value = 0.06002
```

The p-value returned by the test is higher than 0.05, therefore we can conclude that after removing the observations relative to Alaska and Texas, that are by far the states with the largest territorial extension, we do not reject anymore the assumption that *Area* is gaussian.

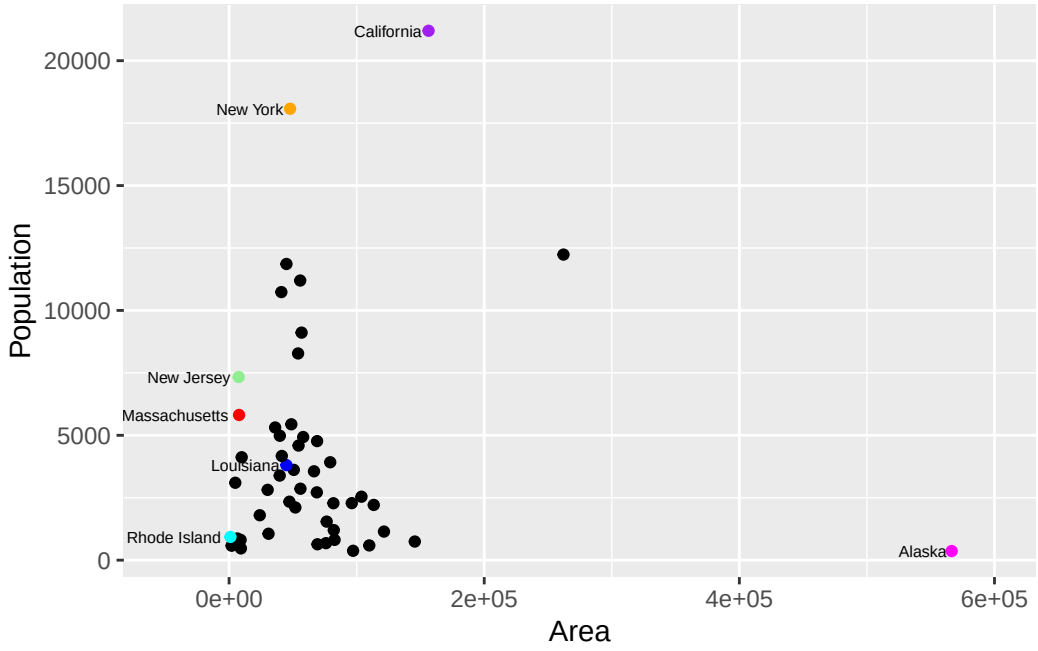
1.5 Make a scatter plot of Area vs Population, colour-coding the outliers found in point 2 with a different colours. Choose among the following colour names. Can they be considered bivariate outliers?

In this section we are clearly interested in analyzing whether the univariate outliers of *Population* and *Area* can also be considered as bivariate outliers (for this pair of variable). Nevertheless, we think that it would be also interesting to verify if any of the univariate outliers for the other

variables shows a peculiar behavior for this bivariate distribution. To accomplish this we firstly recall which are the univariate outliers (restricted to those found in section 1.2) of *Area* and *Population*:

```
    variable    outlier
1 Population California
2 Population   New York
3       Area     Alaska
```

We can observe that the two variables have no univariate outliers in common (among the ones identified in section 1.2).

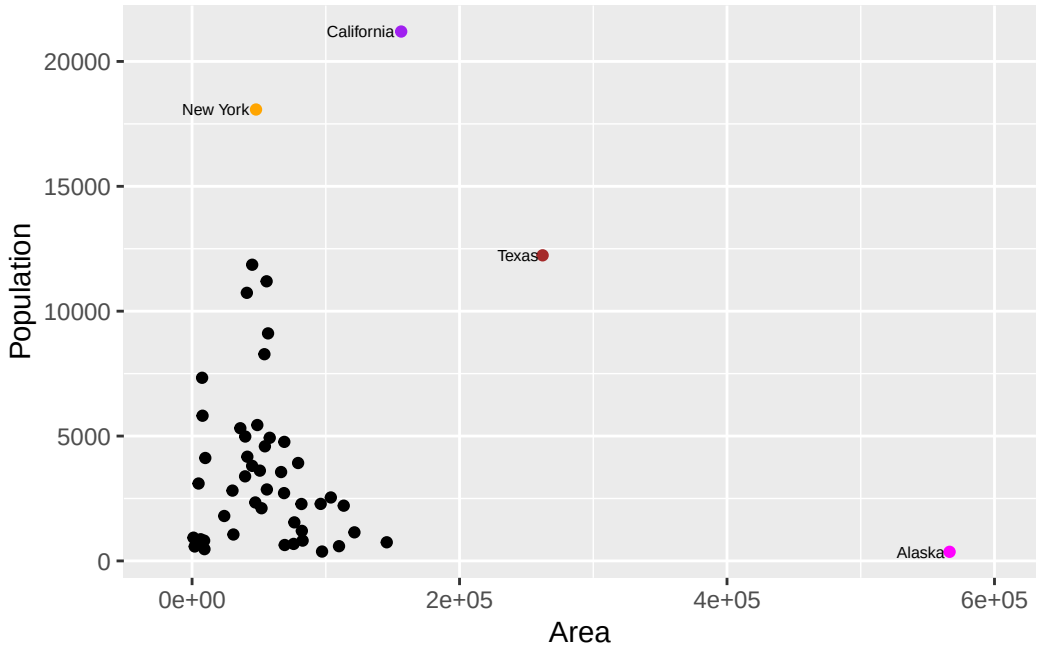


The scatter plot reveals that the three univariate outliers for *Area* (Alaska) and *Population* (California and New York), can also be regarded as bivariate outliers. Indeed, even if their values are extreme only for one of the two variables, these values are so "out-of-proportion" with respect to the other data that they relevantly isolate the univariate outliers also in the bi-dimensional scatter plot. In addition to these three states, we also notice the presence of a fourth potential bivariate outlier. We can easily identify it as Texas since it corresponds to the states with the second largest geographical extension (and therefore is the second outlier of *Area* that we have discussed about in the previous section). To be more complete, we can also notice that Texas is the third most populated state so, by combining this information to the boxplot of *Population* plotted in section 1.3, we can conclude that Texas was not only a univariate outlier for *Area* but also for *Population*.

Another aspect worth-noticing of the plot is that all the univariate outliers of the other variables do not exhibit any special behavior, but they are inside the "cloud" of the points.

Finally, we have decided not to perform the analysis of the bivariate outliers with the aid of the ellipse containing the $\frac{n-0.5}{n}\% = 0.99\%$ of the joint population distribution because this approach relies on the assumption of bivariate Gaussian distribution, that in this context does not hold. Indeed, the data lead to reject the Gaussianity of both *Area* and *Population* and therefore it is highly improbable that their joint distribution is Normal.

To outline the drawn conclusions, we produce again the scatter plot of *Area* vs *Population* , this time only enhancing the identified bivariate outliers. To do so it is enough to retain only the two states with the largest territorial extensions and the two with the largest population.

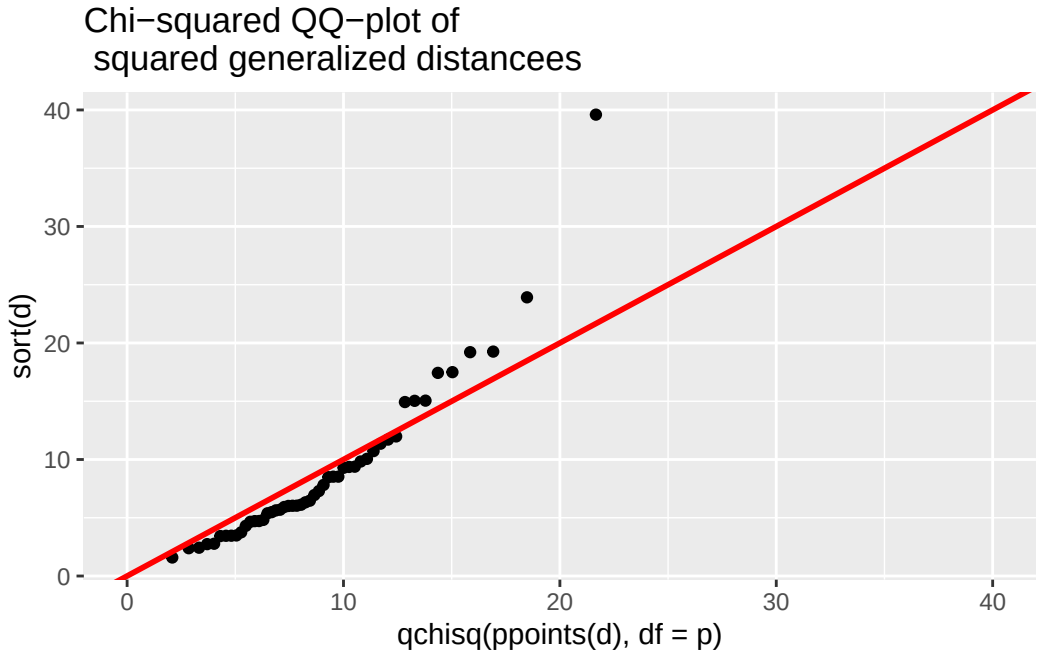


1.6 Construct a chi-square Q-Q plot of the squared Mahalanobis distances and comment about multivariate normality.

From the previous sections we know that there is not enough evidence for univariate normality of six out of the variables hence we expect that the multivariate distribution is not Gaussian (otherwise all the marginal distributions would have been Gaussian).

To confirm this, we compute the squared generalized distances $d_i^2 = (x_i - \bar{x})^T \mathbf{S}^{-1}(x_i - \bar{x}), \quad i = 1, \ldots, n$. Under the jointly multivariate assumption, each of the squared distances should behave like χ_p^2 random variables (with p corresponding to the number of variables of the considered data set). We can, therefore, address the problem of multivariate normality by plotting the increasingly arranged values of the squared Mahalanobis distances against the

theoretical quantiles of the chi-square distribution with p degrees of freedom. The multivariate normality will not be rejected if the plot resembles a straight line through the origin.



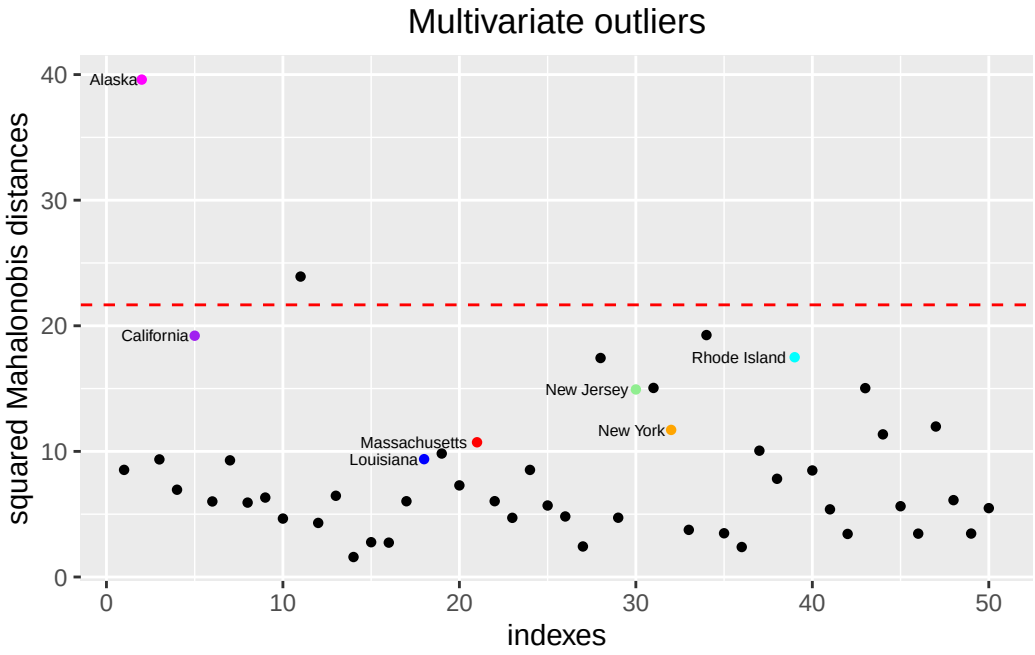
As expected the plot indicates a lack of evidence for multivariate normality:

- the great majority of the squared distances (approximately all the ones between 0 and 10) are slightly below the straight line. This correspond to the fact that the points at the left end of the plot are more densely concentrated than for a χ^2_p distribution.
- the largest values, conversely, display an increasing trend that indicates a heavier right tail in the squared distances distribution.

The two remarks together give a clear indication that we lack statistical evidence to assess the Chi-squared distribution, and so underlying multivariate normal distribution. Therefore, combining this plot with the already discussed results of univariate gaussianity, we conclude that there are enough elements to reject the assumption of multivariate gaussianity.

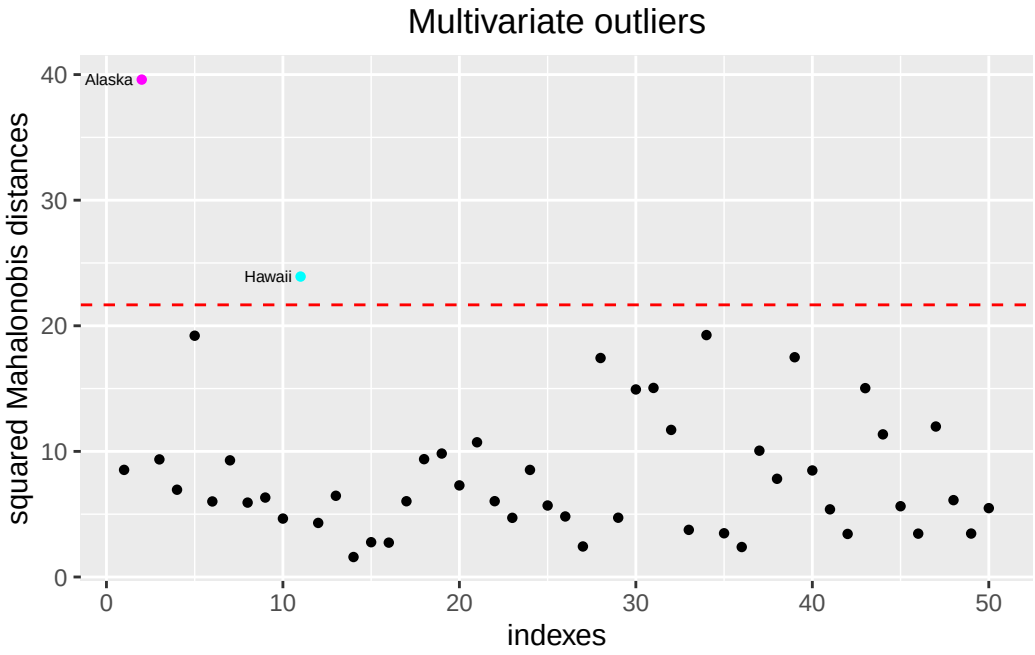
1.7 Identify multivariate outliers, if any, and compare with the univariate outliers previously found

Despite the lack of evidence to support multivariate normality (but never forgetting this potential source of complications and mistakes), we try to detect multivariate outliers using again the squared Mahalanobis distance. This is useful since it may reveals unexpected multivariate outliers, that do not exhibit particularly extreme behaviors for any of the single variables.



The comparison of the squared Mahalanobis distances with the $\frac{n-0.5}{n}\% = 0.99\%$ quantile of the χ^2_p distribution highlights the presence of two multivariate outliers, which correspond to the states of:

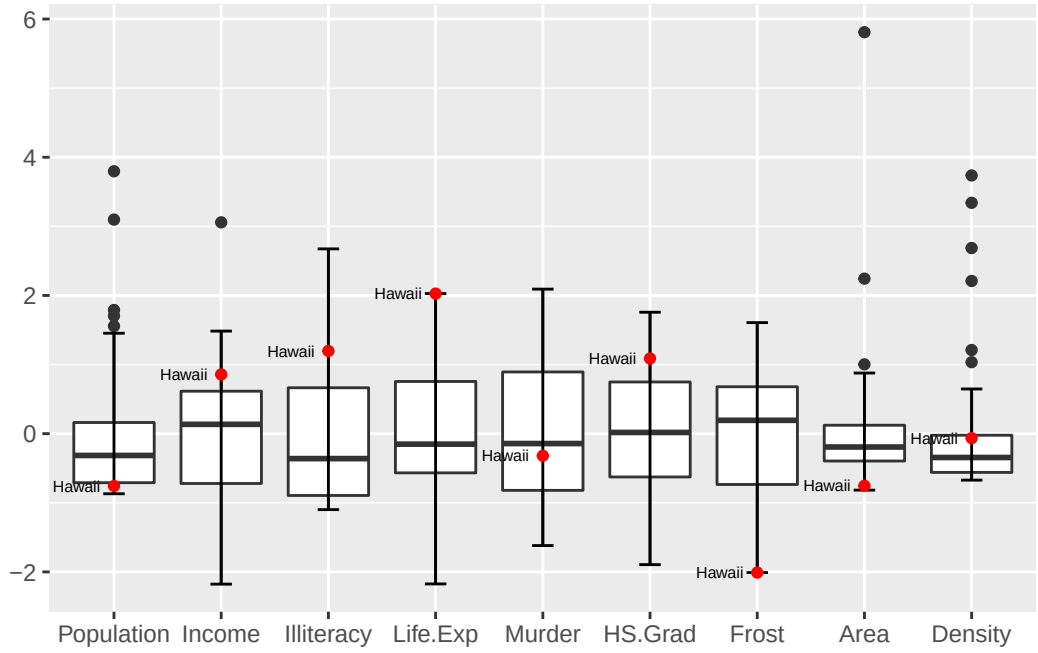
```
[1] "Alaska" "Hawaii"
```



The comparison shows that Alaska and Hawaii are the only two potential multivariate outliers. Alaska was also a univariate outlier for the two distinct variables *Income* and *Area*, whereas Hawaii is not one of the univariate outliers (which we have found in section 1.2) for any of the variables. Therefore we may argue that, although both of them can be considered as multivariate outliers, the reason for that has different explanations:

- the squared Mahalanobis distance of Alaska is, probably, so large because of the fact that it is both a univariate and a bivariate outlier for some of the variables.
- conversely, it is more likely that the Mahalanobis distance of Hawaii is large due to the global interactions in the joint distribution.

To have a confirmation of these speculations we mark the observation related to Hawaii in all the boxplots of the variables:



Interestingly, these two multivariate outliers correspond to the two latest states to have joined USA and this may have played a relevant role in isolating them from a cultural and historical point of view. In addition, they are also the only two non-contiguous states of the US and are characterized by two quite unique geographical contexts.

All these peculiarities may play an important role in explaining the exceptionality of these states.

Exercise 2

Let us consider $Z = (X, Y_1, Y_2)$ a Gaussian random vector $N_3(\mu, \mathbf{\Sigma})$ where

$$\mu = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \quad \mathbf{\Sigma} = \begin{bmatrix} 1 & -\rho & \rho \\ -\rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

with $-1 < \rho < 0.5$.

2.1 Find the inverse of $\mathbf{\Sigma}$

First of all, the matrix $\mathbf{\Sigma}$ is invertible since the determinant is different from 0 for every value of ρ considered:

$$det(\mathbf{\Sigma}) = 1 - 3\rho^2 - 2\rho^3 = (1+\rho)(1-\rho-2\rho^2) = (1+\rho)^2(1-2\rho) \neq 0$$

We note that the covariance matrix of Z can be written as $\mathbf{\Sigma} = (1+\rho)I - \rho aa^T$ where $a = (1, 1, -1)$ tri-dimensional vector. If we define $\mathbf{\Sigma}^* \coloneqq \frac{1}{1+\rho}\mathbf{\Sigma}$, we note that we have a matrix of the form: $\mathbf{\Sigma}^* = I - \frac{\rho}{1+\rho}aa^T$. If we find the inverse of $\mathbf{\Sigma}^*$, then simply $\mathbf{\Sigma}^{-1} = \frac{1}{1+\rho}(\mathbf{\Sigma}^*)^{-1}$. The form of $\mathbf{\Sigma}^*$ suggests that we may look for an inverse matrix using the Neumann series technique, which can be stated as:

> **Theorem:** If T is a bounded linear operator on a normed vector space X, and if the Neumann series $\sum_{k=0}^{\infty} T^k$ converges in the operator norm, then $\mathrm{Id} - T$ is invertible and its inverse is given by the Neumann series:
>
> $$(I-T)^{-1} = \sum_{k=0}^{\infty} T^k$$

In our particular instance the normed space is the space of real valued 3x3 matrices and the operator norm of a matrix is given by the spectral radius, which is trivially bounded for any matrix in the space. The advantage of this method is that by the associative property of matrix product we manage to simplify greatly the powers of aa^T, using the inner product $(\cdot, \cdot)$:

$$(aa^T)^2 = (aa^T)(aa^T) = a(a^Ta)a^T = (a,a)aa^T = ||a||^2aa^T = 3aa^T \tag{1}$$

Thus, more generally (by simple induction):

$$(aa^T)^n = ||a||^{n-1}aa^T = 3^{n-1}aa^T = \frac{3^n}{3}aa^T, \forall n \geq 1 \tag{2}$$

So the inverse of $\mathbf{\Sigma}^*$ is $\sum_{n\geq 0}\left(\frac{\rho}{1+\rho}aa^T\right)^n = I + \frac{1}{3}\left[\sum_{n\geq 1}\left(\frac{3\rho}{1+\rho}\right)^n\right]aa^T$, whenever the geometric series above converges. The series does not converge for all values of ρ considered, but since the theorem above does not say anything whenever the series is not-convergent, we can use it to guess effectively the form of the inverse

$$(\mathbf{\Sigma}^*)^{-1} = I + \beta aa^T \text{ for some } \beta$$

Combining our guess with the condition of invertibility

$$I = \mathbf{\Sigma}^*(\mathbf{\Sigma}^*)^{-1}$$

(it is sufficient to check for the left/right inverse condition, since we know the inverse exists and thus it's the unique left/right inverse), we get:

$$\begin{aligned} I &= (I - \frac{\rho}{1+\rho} aa^T)(1 + \beta aa^T) = \\ &= I + \left(\beta - \frac{3\beta\rho}{1+\rho} - \frac{\rho}{1+\rho}\right) aa^T = \\ &= I + \left(\frac{1-2\rho}{1+\rho}\beta - \frac{\rho}{1+\rho}\right) aa^T \\ &\implies \beta = \frac{\rho}{1-2\rho} \end{aligned}$$

and the solution is well defined since ρ is smaller than $\frac{1}{2}$.
To conclude, the inverse of $\mathbf{\Sigma}$ is

$$\mathbf{\Sigma}^{-1} = \frac{1}{1+\rho}(\mathbf{\Sigma}^*)^{-1} = \frac{1}{1+\rho} I + \frac{\rho}{(1-2\rho)(1+\rho)} aa^T$$

which can be written explicitly as

$$aa^T = \begin{bmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix} \implies \mathbf{\Sigma}^{-1} = \frac{1}{(1-2\rho)(1+\rho)} \begin{bmatrix} 1-\rho & \rho & -\rho \\ \rho & 1-\rho & -\rho \\ -\rho & -\rho & 1-\rho \end{bmatrix}$$

2.2 Find the eigenvalues of $\mathbf{\Sigma}$

A possible solution would be to find the solution of the characteristic polynomial, which is the standard procedure. However, we note that:

- the sum of the eigenvalues is the trace of the matrix $\mathbf{\Sigma}$: $\lambda_1 + \lambda_2 + \lambda_3 = Tr(\mathbf{\Sigma}) = 3$
- the eigenvalues of the inverse $\mathbf{\Sigma}^{-1}$ are the reciprocal of the eigenvalues of $\mathbf{\Sigma}$, and using the same property of the trace we mentioned above: $\frac{1}{\lambda_1} + \frac{1}{\lambda_2} + \frac{1}{\lambda_3} = Tr(\mathbf{\Sigma}^{-1}) = \frac{3(1-\rho)}{(1-2\rho)(1+\rho)}$
- since $\mathbf{\Sigma} = (1+\rho)I - \frac{\rho}{1+\rho} aa^T$ it's easy to see that a is an eigenvector of $\mathbf{\Sigma}$ (not-normalized), because: $\mathbf{\Sigma}a = (1+\rho)a - \rho a(a^T a) = (1 + \rho - 3\rho)\, a = (1-2\rho)a$.

Let us set $\lambda_1 = 1 - 2\rho$, eigenvalue corresponding to the eigenvector a. Then by combining the first two conditions we get the following system:

$$\begin{cases} 1 - 2\rho + \lambda_2 + \lambda_3 = 3 \\ \dfrac{1}{1-2\rho} + \dfrac{1}{\lambda_2} + \dfrac{1}{\lambda_3} = \dfrac{3(1-\rho)}{(1-2\rho)(1+\rho)} \end{cases} \quad \begin{cases} \lambda_2 + \lambda_3 = 2(1+\rho) \\ \dfrac{\lambda_2 + \lambda_3}{\lambda_2\lambda_3} = \dfrac{3(1-\rho) - (1+\rho)}{(1-2\rho)(1+\rho)} \end{cases}$$

Substituting sum in the first into the second yields:

$$\begin{cases} \lambda_2 + \lambda_3 = 2(1+\rho) \\ \dfrac{1(1+\rho)}{\lambda_2\lambda_3} = \dfrac{2(1-2\rho)}{(1-2\rho)(1+\rho)} \end{cases} \quad \begin{cases} \lambda_2 + \lambda_3 = 2(1+\rho) \\ \dfrac{1(1+\rho)}{\lambda_2\lambda_3} = \dfrac{2}{1+\rho} \end{cases}$$

$$\begin{cases} \lambda_2 + \lambda_3 = 2(1+\rho) \\ \lambda_2 \lambda_3 = (1+\rho)^2 \end{cases} \implies \lambda_2 = \lambda_3 = 1 + \rho$$

So the eigenvalues are $\lambda_1 = 1 - 2\rho, \lambda_2 = \lambda_3 = 1 + \rho$.

2.3 Let PC1 and PC2 be the first two (population) principal components of Z. Find ρ such that they account for more than 80% of the total variation of X.

We recall that the total variation of Z is equal to the total variance of the principal components and that the (population) covariance matrix of the principal components is, by construction, a diagonal matrix with the eigenvalues of $\mathbf{\Sigma}$ on the diagonal in decreasing order. hence the proportion of variance of Z explained by the first two principal components is given by the ratio between the sum of the two largest eigenvalues of $\mathbf{\Sigma}$ (i.e.: the variance of the first two PCs) and the sum of all eigenvalues of $\mathbf{\Sigma}$, which is equal to its trace. In other words, we want to find the values of $\rho \in (-1, \frac{1}{2})$ such that

$$\frac{\lambda_{(1)} + \lambda_{(2)}}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{\lambda_{(1)} + \lambda_{(2)}}{3} > 0.8 = \frac{4}{5}$$

where $\lambda_{(j)}$ is the j-th eigenvalue of $\mathbf{\Sigma}$ in decreasing order.
The order of the eigenvalues depends on the values of the parameter, in particular:

$$1 + \rho \geq 1 - 2\rho \iff \rho \geq 0$$

.
Either $\rho \in (-1, 0]$, which implies

$$\lambda_{(1)} + \lambda_{(2)} = \lambda_1 + \lambda_2 = 1 - 2\rho + 1 + \rho = 2 - \rho$$

and so we ask that

$$2 - \rho > \frac{12}{5} \iff \rho < -\frac{2}{5}$$

Or instead $\rho \in [0, \frac{1}{2})$: which implies

$$\lambda_{(1)} + \lambda_{(2)} = 2(1 + \rho) = 2 + 2\rho$$

and thus the inequality

$$2 + 2\rho > \frac{12}{5} \iff \rho > \frac{1}{5}$$

To sum up, the values of the parameter that satisfy the condition are $\rho \in (-1, -\frac{2}{5}) \cup (\frac{1}{5}, \frac{1}{2})$.

2.4 Find the conditional distribution of $Y = (Y_1, Y_2)$ given $X = x$.

To compute the conditional distribution, we make use of the formula

> **Proposition:** Let $Z = (X, Y) \sim N_p(\mu, \mathbf{\Sigma})$ with
>
> $$\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_x & \mathbf{\Sigma}_{xy} \\ \mathbf{\Sigma}_{yx} & \mathbf{\Sigma}_y \end{bmatrix}$$
>
> where $\mu_x \in \mathbb{R}^q$, $q < p$, and $\mathbf{\Sigma}_y$ positive definite. Then the conditional distribution of $Y \mid X = x \sim N_{p-q}(\mu_{y|x}, \mathbf{\Sigma}_{y|x})$ with
>
> $$\mu_{y|x} = \mu_y + \mathbf{\Sigma}_{yx}\mathbf{\Sigma}_x^{-1}(x - \mu_x) \quad \text{and} \quad \mathbf{\Sigma}_{y|x} = \mathbf{\Sigma}_y - \mathbf{\Sigma}_{yx}\mathbf{\Sigma}_x^{-1}\mathbf{\Sigma}_{xy}$$

In our case, we have:

- $Y = (Y_1, Y_2)$ bi-dimensional (X is one-dimensional and consistent in notation)
- $\mu_x = 1$ and $\mu_y = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$
- $\mathbf{\Sigma}_x = \begin{bmatrix} 1 \end{bmatrix}$, $\mathbf{\Sigma}_{xy} = \begin{bmatrix} -\rho & \rho \end{bmatrix} = \mathbf{\Sigma}_{yx}^T$, and $\mathbf{\Sigma}_y = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$

Since $\mathbf{\Sigma}_{22}$ is positive definite (principal sub-matrix of the positive definite $\mathbf{\Sigma}$), the proposition can be applied and we find that $(Y_1, Y_2)|X = x \sim N_2(\mu_{y|x}, \mathbf{\Sigma}_{y|x})$ with

$$\mu_{y|x} = \begin{bmatrix} 0 \\ 2 \end{bmatrix} + \begin{bmatrix} -\rho \\ \rho \end{bmatrix} (1)^{-1}(x - 1) = \begin{bmatrix} -\rho(x-1) \\ 2 + \rho(x-1) \end{bmatrix}$$

$$\mathbf{\Sigma}_{y|x} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} - \begin{bmatrix} -\rho \\ \rho \end{bmatrix} \begin{bmatrix} -\rho & \rho \end{bmatrix} = \begin{bmatrix} 1 - \rho^2 & \rho(1+\rho) \\ \rho(1+\rho) & 1 - \rho^2 \end{bmatrix}$$

$$= (1 + \rho) \begin{bmatrix} 1 - \rho & \rho \\ \rho & 1 - \rho \end{bmatrix}$$

Note that the covariance matrix of the conditional distribution is, as expected, independent of the value x, and positive definite given $\rho \in \left(-1, \frac{1}{2}\right)$ (positive trace $tr(\mathbf{\Sigma}_{y|x}) = 2(1 - \rho)$ and positive determinant $det(\mathbf{\Sigma}_{y|x}) = (1 + \rho^2)(1 - 2\rho)$).

2.5 Let $\rho = 0.2$, and Σ_y and μ_y be the corresponding covariance matrix and the mean vector of the distribution (Y_1, Y_2) given X=0. Sketch the ellipse

$$(y - \mu_y)^T \mathbf{\Sigma}_y^{-1} (y - \mu_y) = c^2$$

in the 2-dimensional space $y = (y_1, y_2)$ by setting the constant "c" such that the ellipse contains 0.95 probability with respect to the conditional distribution of Y.

First of all, the given expression is well defined since the matrix $\mathbf{\Sigma}_y$ - which we called $\mathbf{\Sigma}_{y|x}$ above - is invertible, since we already commented its positive definiteness.
The key result we exploit here is that the random variable associated to the squared Mahalanobis distance $Q = (Y-\mu_y)^T \mathbf{\Sigma}_y^{-1} (Y-\mu_y)$, where we use as a notation that $Y = (Y_1, Y_2)$, is distributed as a $\mathcal{X}_2^2$, since $Y \mid X = 0 \sim N_2(\mu_y, \mathbf{\Sigma}_y)$ with:

- $\mu_y^T = \begin{bmatrix} 0.2 & 1.8 \end{bmatrix}$
- $\mathbf{\Sigma}_y = \begin{bmatrix} 0.96 & 0.24 \\ 0.24 & 0.96 \end{bmatrix}$

(by the computation of the previous point with the given value for ρ).
Therefore we can set $c^2 = \chi_2^2(0.05)$, upper 5-percentile of a chi-square distribution with two degrees of freedom, in order to ensure that $1 - \alpha = 0.95$ of the "probability mass" is contained within the ellipse. More formally, if E is the 2-dimensional area contained within the ellipse:

$$E = \{y \in \mathbb{R}^2 : (y - \mu_y)^T \mathbf{\Sigma}_y^{-1} (y - \mu_y) \leq \chi_2^2(0.05)\}$$

then

$$\mathbb{P}(Y \in E \mid X = 0) = \mathbb{P}((Y - \mu_y)^T \mathbf{\Sigma}_y^{-1} (Y - \mu_y) \leq \chi_2^2(0.05) \mid X = 0) = 0.95$$

Let us implement the sketch in R.
First of all we compute:

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

```
[1] 2.447747
```

The ellipse is centered around the mean μ_y of the distribution $Y \mid X = 0$. Its axes oriented as the eigenvectors e_1 and e_2 of the variance-covariance matrix $\mathbf{\Sigma}_y$ and their respective length is given by $c\sqrt{\lambda_1}$ and $c\sqrt{\lambda_2}$, where λ_1 and λ_2 are the eigenvalues of the matrix $\mathbf{\Sigma}_y$.

```
rho=0.2
mu_y=c(rho,2-rho)
```

```r
Sigma_y=matrix(c(1-rho^2,rho+rho^2,rho+rho^2,1-rho^2), nrow=2)
eig=eigen(Sigma_y, symmetric=T)
eig$values
```

```
[1] 1.20 0.72
```

```r
eig$vectors[,1] #first eigenvector
```

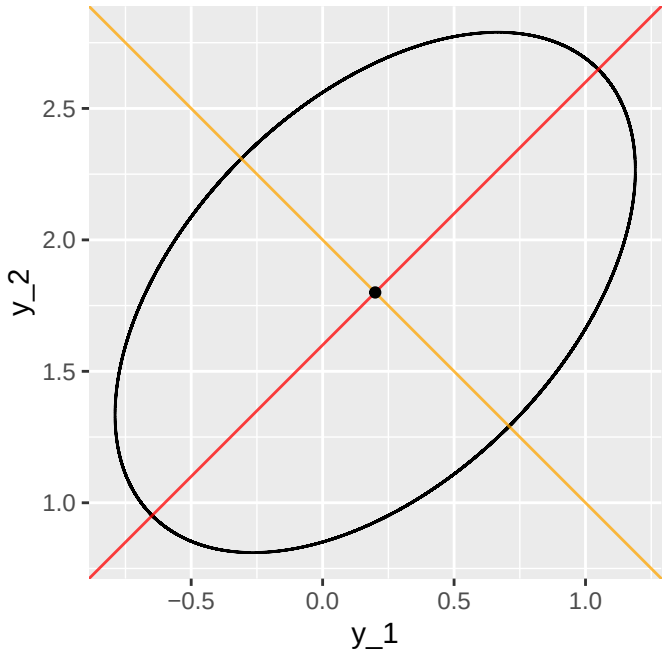
```
[1] 0.7071068 0.7071068
```

```r
eig$vector[,2]  #second eigenvector
```

```
[1] -0.7071068  0.7071068
```

We note that the eigenvectors, and so the directions of the axes, correspond to those of the two bisectors of the plane (i.e.: $\{(y_1, y_2) \in \mathbb{R}^2 : y_1 = y_2\}$ and $\{(y_1, y_2) \in \mathbb{R}^2 : y_1 = -y_2\}$, the first being the direction of the main axis of the ellipse since it's relative to the largest eigenvalue).

Finally, the plot:



Exercise 3

3.1 To equalize out the different types of servings of each food, first divide each variable by weight of the food item. Next, because of the wide variations in the different variables, standardize each variable. Perform Principal Component Analysis on the transformed data.

First of all we import the data and we perform the requested transformations.

```
nutritional <- read.table("data/nutritional.txt")
head(nutritional)
```

```
  fat food.energy carbohydrates protein cholesterol weight saturated.fat
1   2          25             2       0           2  15.00           0.2
2   6          60             2       0           4  16.00           1.0
3   1          90            22       4           0  28.35           0.1
4   0          90            22       3           0  28.35           0.1
5   0          10             1       1           0  33.00           0.0
6   1          70            21       4           0  28.35           0.1
```

```
nutritional=nutritional/nutritional$weight
nutritional=nutritional[,-6]

summary(nutritional)
```

```
      fat            food.energy      carbohydrates        protein
 Min.   :0.00000   Min.   :0.0000   Min.   :0.00000   Min.   :0.000000
 1st Qu.:0.00000   1st Qu.:0.6024   1st Qu.:0.04721   1st Qu.:0.008333
 Median :0.03125   Median :1.8421   Median :0.13636   Median :0.033333
 Mean   :0.11324   Mean   :2.2533   Mean   :0.23806   Mean   :0.069961
 3rd Qu.:0.13295   3rd Qu.:3.4091   3rd Qu.:0.38201   3rd Qu.:0.094714
 Max.   :1.00000   Max.   :9.0244   Max.   :1.00000   Max.   :0.857143
  cholesterol       saturated.fat
 Min.   : 0.0000   Min.   :0.00000
 1st Qu.: 0.0000   1st Qu.:0.00000
 Median : 0.0000   Median :0.00800
 Mean   : 0.2560   Mean   :0.03717
 3rd Qu.: 0.2898   3rd Qu.:0.04735
 Max.   :12.5294   Max.   :0.50714
```

```r
nutritional=scale(nutritional)
nt=as.data.frame(nutritional)

n <- nrow(nt)
```

Now we are ready to proceed with the principal component analysis.

```r
nutritional.pca<-prcomp(nt)
nutritional.pca
```

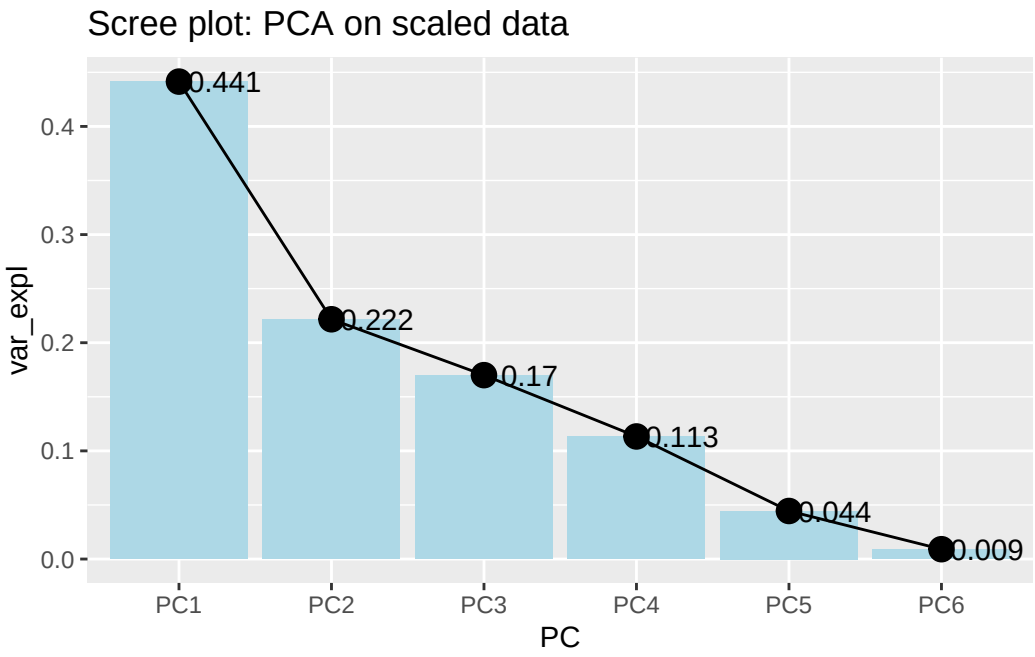
```
Standard deviations (1, .., p=6):
[1] 1.6274498 1.1533146 1.0100127 0.8246633 0.5162603 0.2335648

Rotation (n x k) = (6 x 6):
                      PC1         PC2        PC3         PC4        PC5
fat           -0.55723936  0.09870077 -0.2750890  0.13040139 -0.4546980
food.energy   -0.53615066  0.35676646  0.1370762  0.07454684 -0.2729547
carbohydrates  0.02455362  0.67163163  0.5684779 -0.28616806  0.1568663
protein       -0.23522713 -0.37384298  0.6388770  0.59910351  0.1538186
cholesterol   -0.25250455 -0.52130441  0.3256120 -0.71709615 -0.2102965
saturated.fat -0.53135067 -0.01923360 -0.2611169 -0.14964683  0.7913619
                       PC6
fat            0.616695791
food.energy   -0.697430105
carbohydrates  0.344444078
protein        0.118998503
cholesterol   -0.002904374
saturated.fat  0.021604346
```

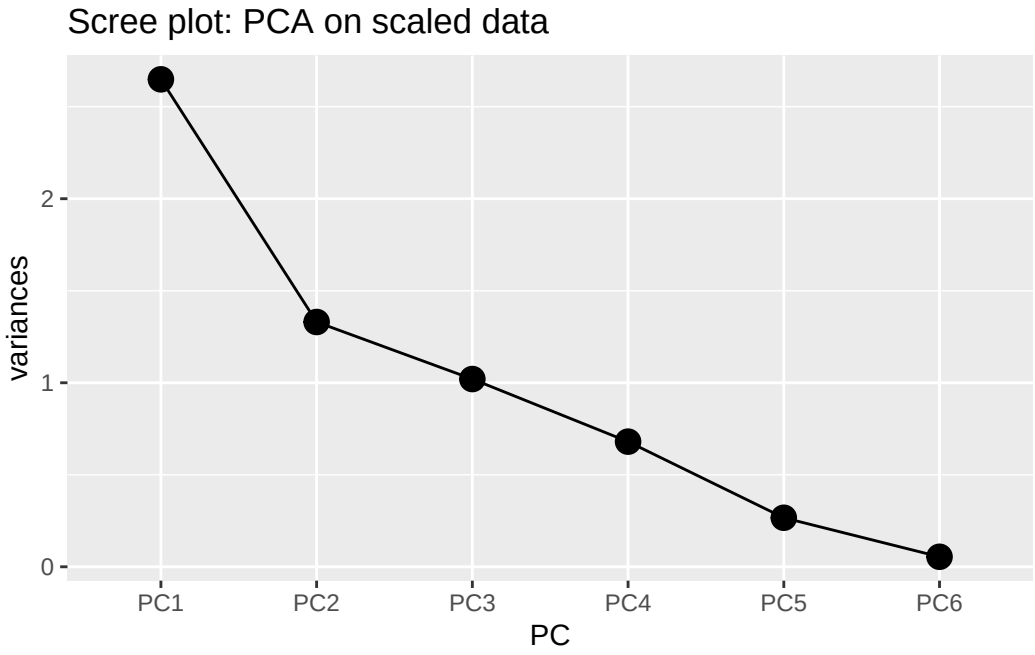
3.2 Decide how many components to retain in order to achieve a

satisfactory lower-dimensional representation of the data. Justify your answer.

In order to decide how many components to retain, we need to understand how much proportion of the variance each component explains.

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6
Standard deviation     1.6274 1.1533 1.0100 0.8247 0.51626 0.23356
Proportion of Variance 0.4414 0.2217 0.1700 0.1133 0.04442 0.00909
Cumulative Proportion  0.4414 0.6631 0.8331 0.9465 0.99091 1.00000
```





As we can see, retaining the first three principal components gives us a cumulative proportion of the variance more or less equal to 83%, which is good but not ideal. Anyway, even though the gain of adding the fourth component is not negligible (11%), we choose not to retain it as this goes in favor of dimensionality reduction. Looking at the screeplot, we can observe again

that there is not a fully-satisfactory criterion for the number of PC's to retain (i.e. the "elbow rule" does not apply). Indeed, the only "bends" in the screeplot seem to suggest that we either keep one or four PC's and the first solution is not acceptable due to the low variance explained whereas the second solution does not provide a good dimensionality reduction. Hence, as a compromise between variance explained and dimensions retained, we opt to keep the first three PC's.

3.3 Give an interpretation to the first two principal components

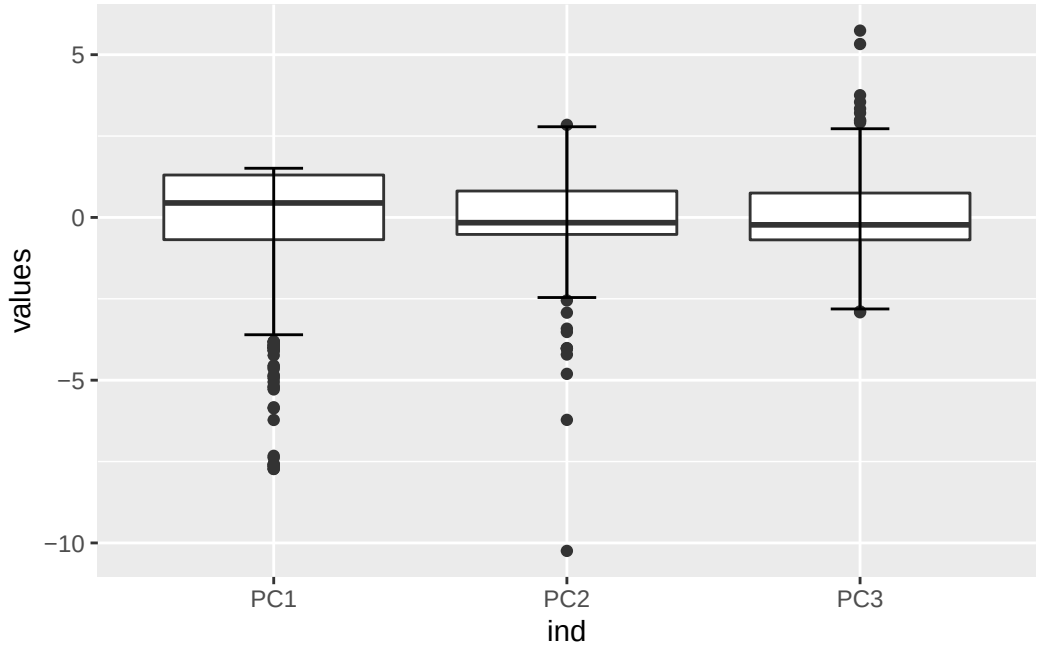
```
nutritional.pca$rotation[,1:2]
```

```
                     PC1         PC2
fat           -0.55723936  0.09870077
food.energy   -0.53615066  0.35676646
carbohydrates  0.02455362  0.67163163
protein       -0.23522713 -0.37384298
cholesterol   -0.25250455 -0.52130441
saturated.fat -0.53135067 -0.01923360
```

In order to give an interpretation of the first two principal components we can look at the above table. We start from the first PC: the weights of *fat* and *saturated.fat* are negative and quite high (in absolute value) and almost equal to the value of *food.energy*. This suggests that this component is a measure of how much fat a food contains, which could also explain why *food.energy* has a relevant coefficient: fats account for the most part of calories in a food. In particular, a food with a high percentage of fat (especially if saturated) will score low and viceversa.
On the other hand, in the second PC the weights of *protein* and *carbohydrates* are significant and opposite in sign and the weight of *cholesterol* (which is a substance only presents in food of animal origin) has the same sign (negative) of *protein*. Now, since in general a food with animal origin has more protein than carbohydrates and a food with vegetable origin has more carbohydrates than protein, we can think of this PC as an indicator of the origin of the food. In particular, a food of animal origin will score low and a plant based food will score high.

3.4 Identify univariate outliers with respect to the first three principal components, up to 3 per component. These points correspond to foods that are very high or very low in what variable (up to 2 variables per observation)?

To answer this question, a good strategy is to plot the boxplots relative to the first three principal components:



As we can see, we have many outliers for each principal component. We focus on the most extreme observation, starting from the first PC:

```
         fat food.energy carbohydrates    protein cholesterol saturated.fat
411 3.622901    2.539105    -0.9537757 -0.6794926    2.857026      7.079055
412 3.622901    2.539105    -0.9537757 -0.6794926    2.857026      7.079055
286 3.475875    2.526043    -0.9537757 -0.7778908    2.899137      7.106775
```

We can observe that, according to the boxplot, these three outliers score very low on the first PC. Furthermore, they have a very high percentage of fats and even more of saturated fats. These two facts combined match our previous interpretation of the first PC, because we expect that a food with an extreme value of fat is an outlier for the first PC. We note that the first two have identical values, but since we do not know whether the frequency of equal observations is relevant in the general framework of the analysis or not, we do not discard the copies. The third is very similar to the previous two. In general, the possibility of finding almost equal yet not identical observations in the dataset could be related to the fact that the original variables are integers (except for *saturated.fat*) and since they are measured on different weights, they have different approximations propagated by the rescaling (e.g. a 20g food with 0g proteins could actually have 2g of proteins if measured in 100g, nevertheless the rescaling of our data is not able to keep it into account). For the sake of completeness, we also underline the fact that the fourth outlier is actually identical to the third one.

```
         fat food.energy carbohydrates    protein cholesterol saturated.fat
```

```
286 3.475875    2.526043    -0.9537757 -0.7778908    2.899137      7.106775
287 3.475875    2.526043    -0.9537757 -0.7778908    2.899137      7.106775
```

No we move to the second PC:

```
           fat food.energy carbohydrates  protein cholesterol saturated.fat
866  0.9349277  0.65926945    -0.9537757 1.184284   18.169910     0.8611680
49  -0.3268590 -0.38916254    -0.9537757 2.001856    8.947732    -0.2596103
819 -0.1596343 -0.03968521    -0.6238323 2.230777    6.761927    -0.1172893
```

As we can see, these outliers score very low on the second PC. Moreover, they have a high percentage of proteins and a very high percentage of cholesterol. These facts suggest the animal origin and again match our interpretation of the second PC.
Now we conclude with the third PC:

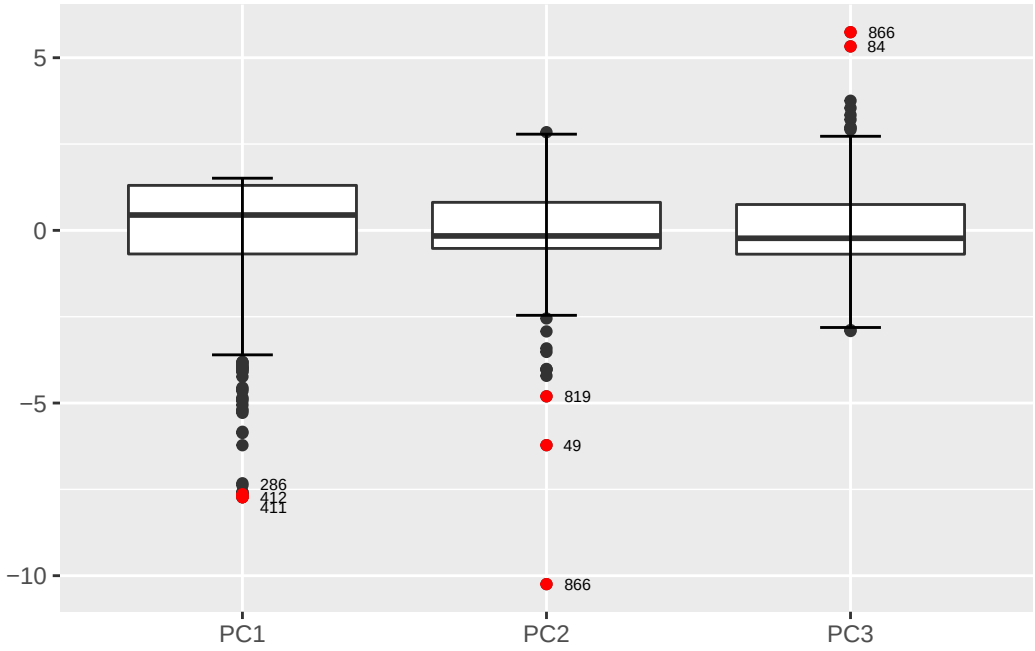
```
           fat food.energy carbohydrates  protein cholesterol saturated.fat
866  0.9349277   0.6592695    -0.9537757 1.184284  18.1699105     0.8611680
84  -0.5852973   0.6809761    -0.9537757 8.752671  -0.3789479    -0.5620426
```

Finally, we can observe that the both observations have all values close to the average, except for *cholesterol* for the first one and *protein* for the second, which are remarkably high. This could explain the fact that they are outliers since both *protein* and *cholesterol* have a relevant weight in the PC3.

We also notice that, for the same reason, observation *866* was also an univariate outlier for the second PC.

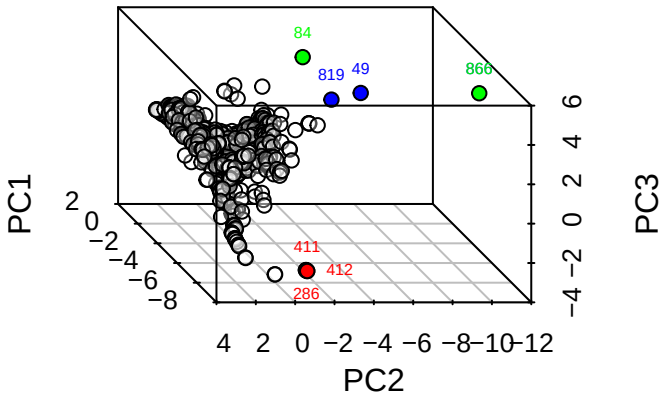
We decided to discuss only these two observations, because they are by far the most extreme observations.

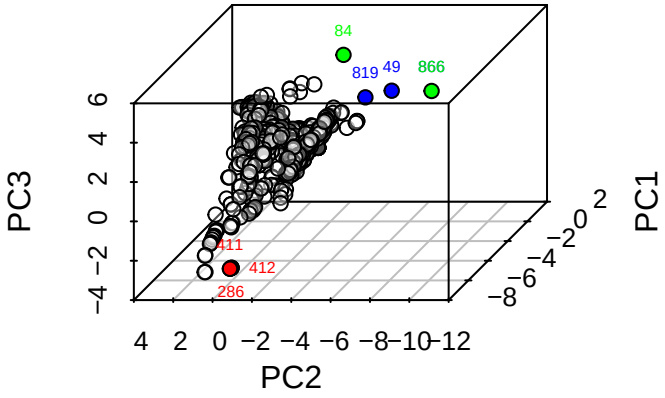
To give a visual representation, we mark them in red inside the boxplots:

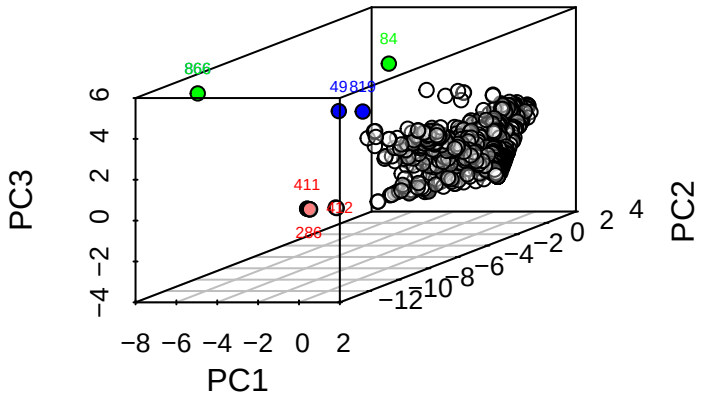


3.5 Make a 3-d scatter plot with the first three principal components, while color coding these outliers.

As requested we make a 3-d scatterplot with the first three principal components. We plot in red the ones relative to the first, in blue the ones relative to the second and in green the ones relative to the third. We can appreciate that these outliers are potential "3-d outliers", because they are very far from the cloud containing most of the point.

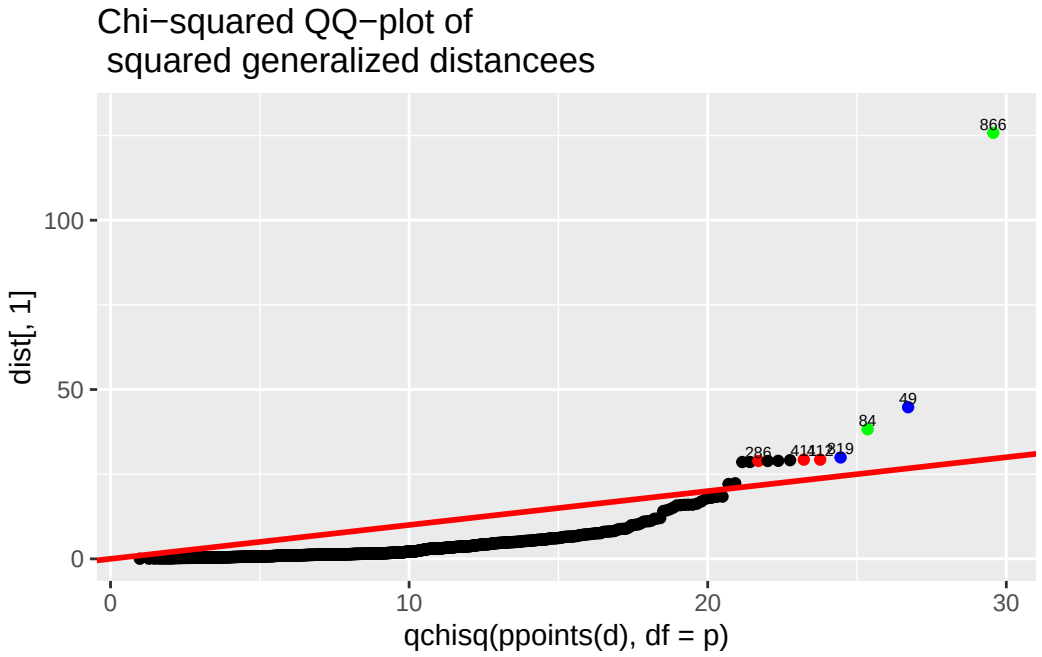

84
819
49
866
411
412
286
PC1
PC2
PC3
2
0
−2
−4
−6
−8
4
2
0
−2
−4
−6
−8
−10
−12
−4
−2
0
2
4
6


84
819
49
866
411
412
286
PC3
PC2
PC1
−4
−2
0
2
4
6
4
2
0
−2
−4
−6
−8
−10
−12
2
0
−2
−4
−6
−8



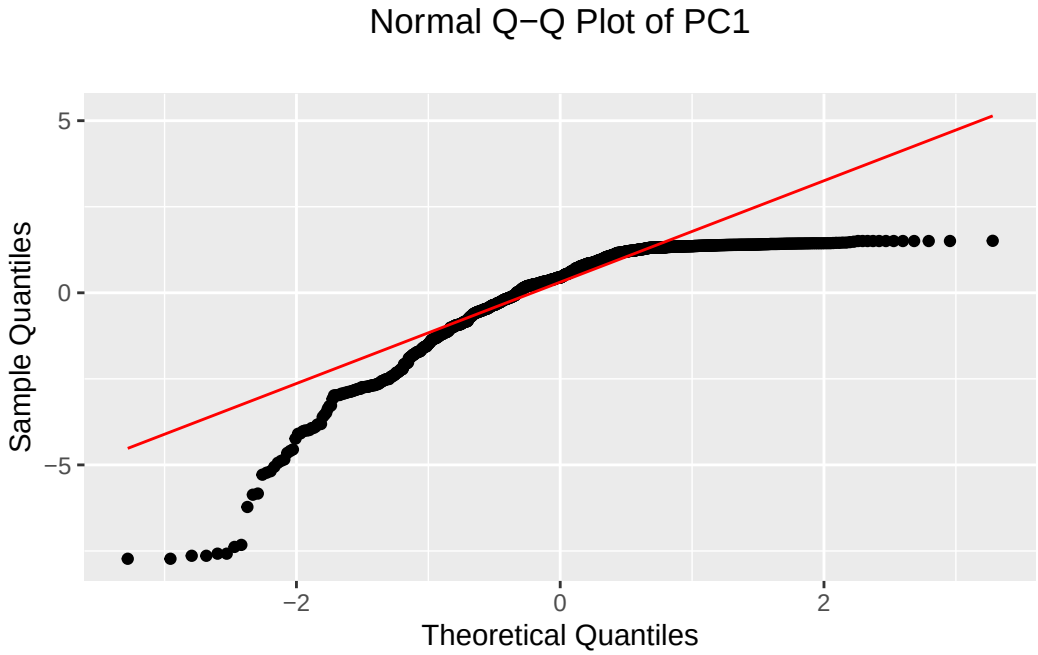
3.6 Investigate multivariate normality through the first three principal components.

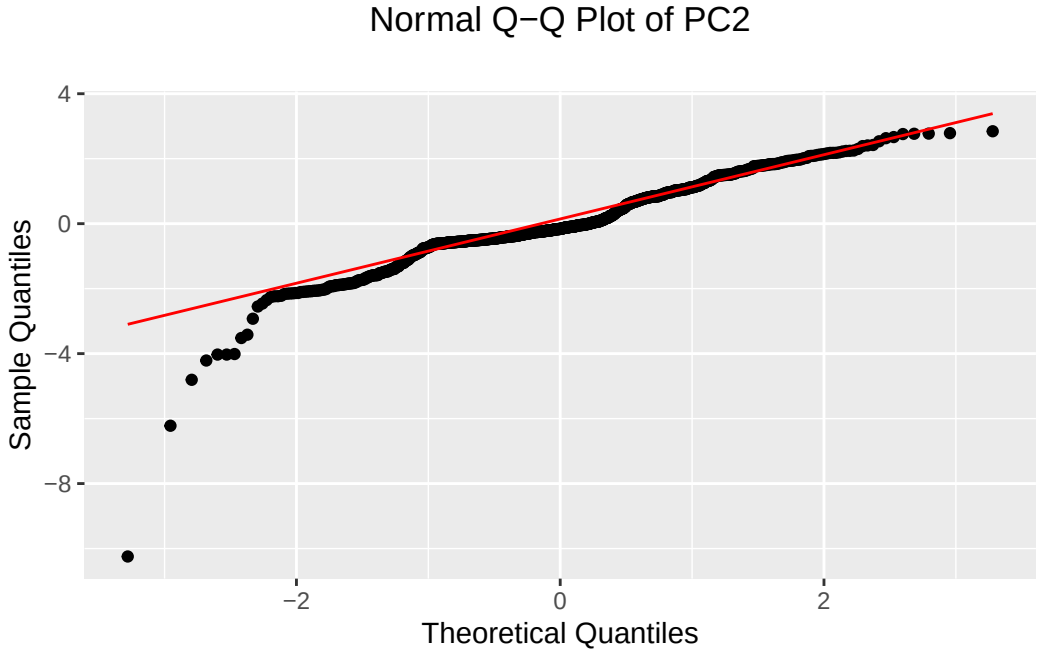
To answer this question, we plot in three different graphs the theoretical quantile of a normal distribution in correspondence quantile followed by the first three principal components.

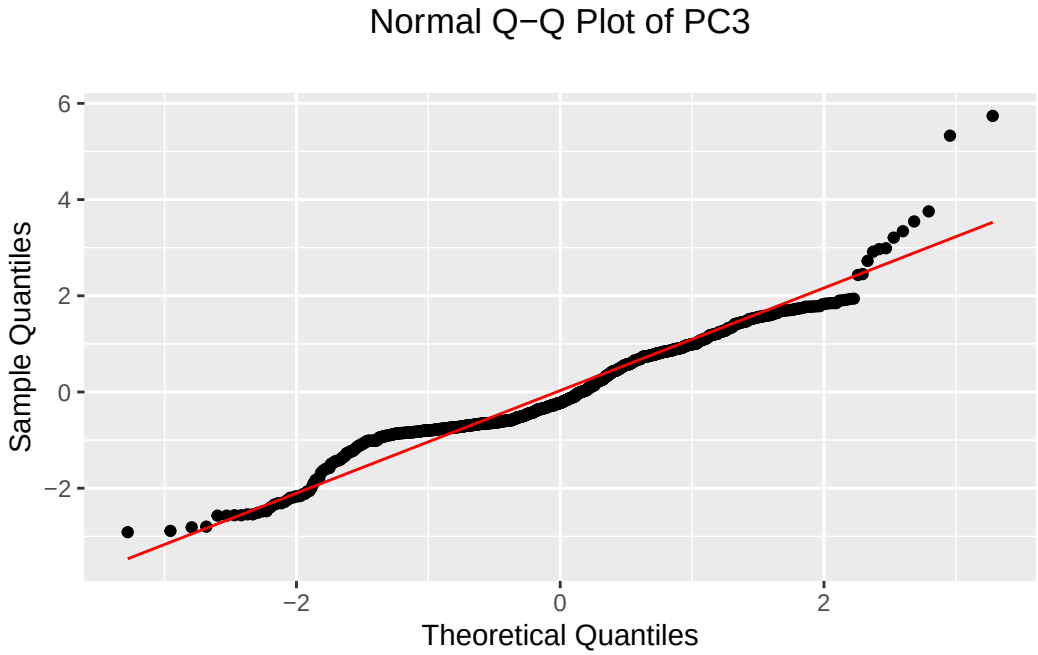


Looking at the above plot, we do not have sufficient statistical evidence to assess multivariate normality of the data. Indeed, the pattern at the top right seems to suggest that the largest values of the squared Mahalanobis distances are more spread out than for a χ^2_3-distribution, i.e. the tail is heavier. This indicates that the multivariate distribution of our data is more right skewed than you would expect to see with a multivariate normal.

We can find a further confirmation by studying the univariate normality of the first three PC's, since a multivariate Gaussian requires each marginal to be Gaussian too. We perform it by means of the QQ-plot for each PC:


Normal Q−Q Plot of PC1
Sample Quantiles
Theoretical Quantiles
5
0
−5
−2
0
2


Normal Q−Q Plot of PC2
Sample Quantiles
Theoretical Quantiles
4
0
−4
−8
−2
0
2

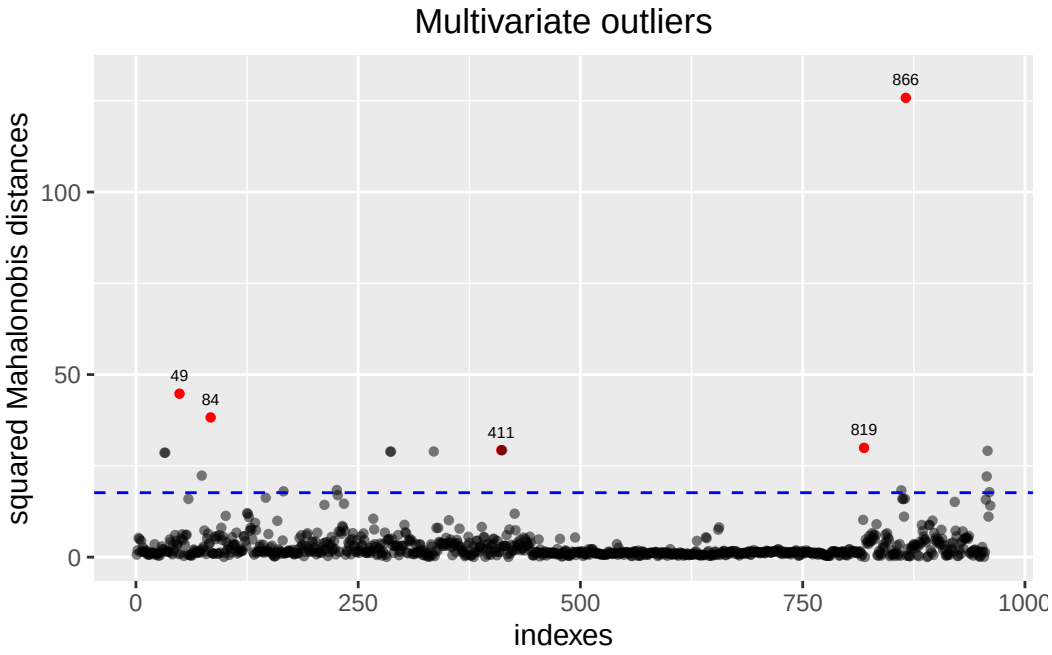


In particular in the first two plots, there is a high evidence of non normality, since the trend of the extreme quantiles moves away from the theoretical normal one. The third appears less pathological, but it still displays a few controversial features that would require further investigation.

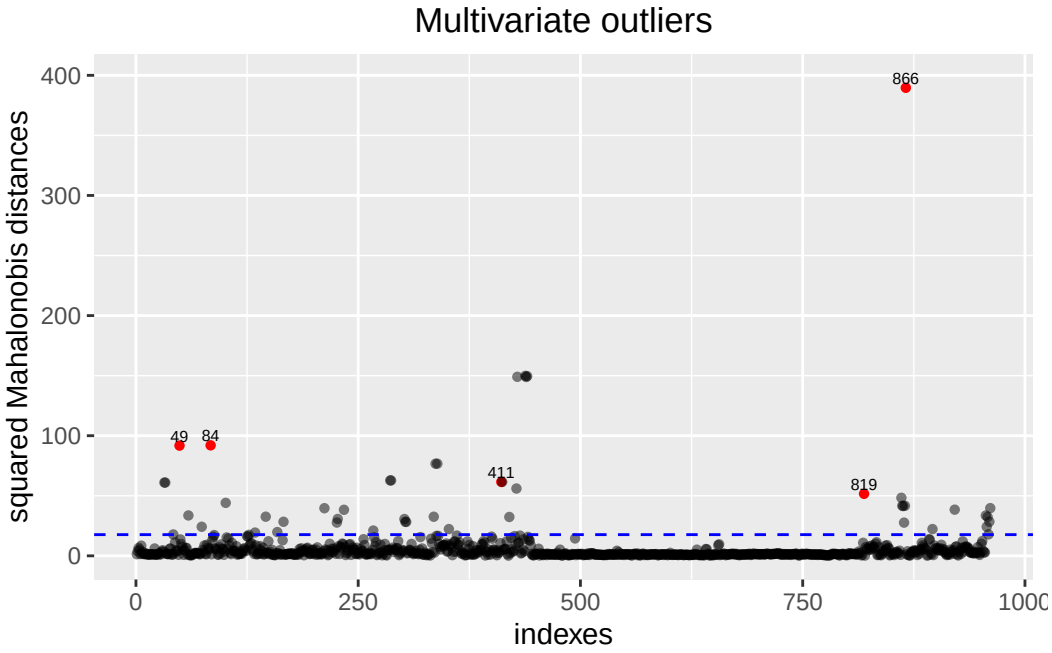
3.7 Find multivariate outliers through the first three principal components, up to 5 in total. Are they the most extreme observations with respect to the 6 original variables?

We start by looking for the multivariate outliers with respect to the first three principal components, by looking for their indexes.

```
[1] 866  49  84 819 411
```



Then, we can compare them graphically with the outliers of the original variables.



We can appreciate that all the multivariate outliers detected for the first three principal component are also detected when we refer to the original variables. Meanwhile, it can be

noticed that the second and the third most extreme observations according to the original variables and their relative squared Mahalanobis distance are not in the five most extreme observations according to the first three PC's and their relative distance. This can be due to the fact that for these observations the remaining PC's play a central role in explaining their variability (keeping in mind that the first 3 only account for the 83% of the total variation).

We can now detect the multivariate outliers for the original variables that are not among the five most extreme outliers for the first three PCs. Since:

```
[1] 84
```

we have four of those outliers, namely:

```
[1] 438 440 429 439
```

To conclude, we represent also them in the above plot.