Basilico Eleonora, Caporali Francesco, Malgieri Luigi Maria
31 March 2023

UNIVERSITÀ DI TORINO

Multivariate Statistical Analysis
Problem Set 1

Exercise 1

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

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

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

Before starting our analysis, we change a couple of variable names in order to avoid spaces, and add the variable *Density* representing the population density.

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

1.1

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

```
R = round(cor(st), 2)
library(corrplot)
```

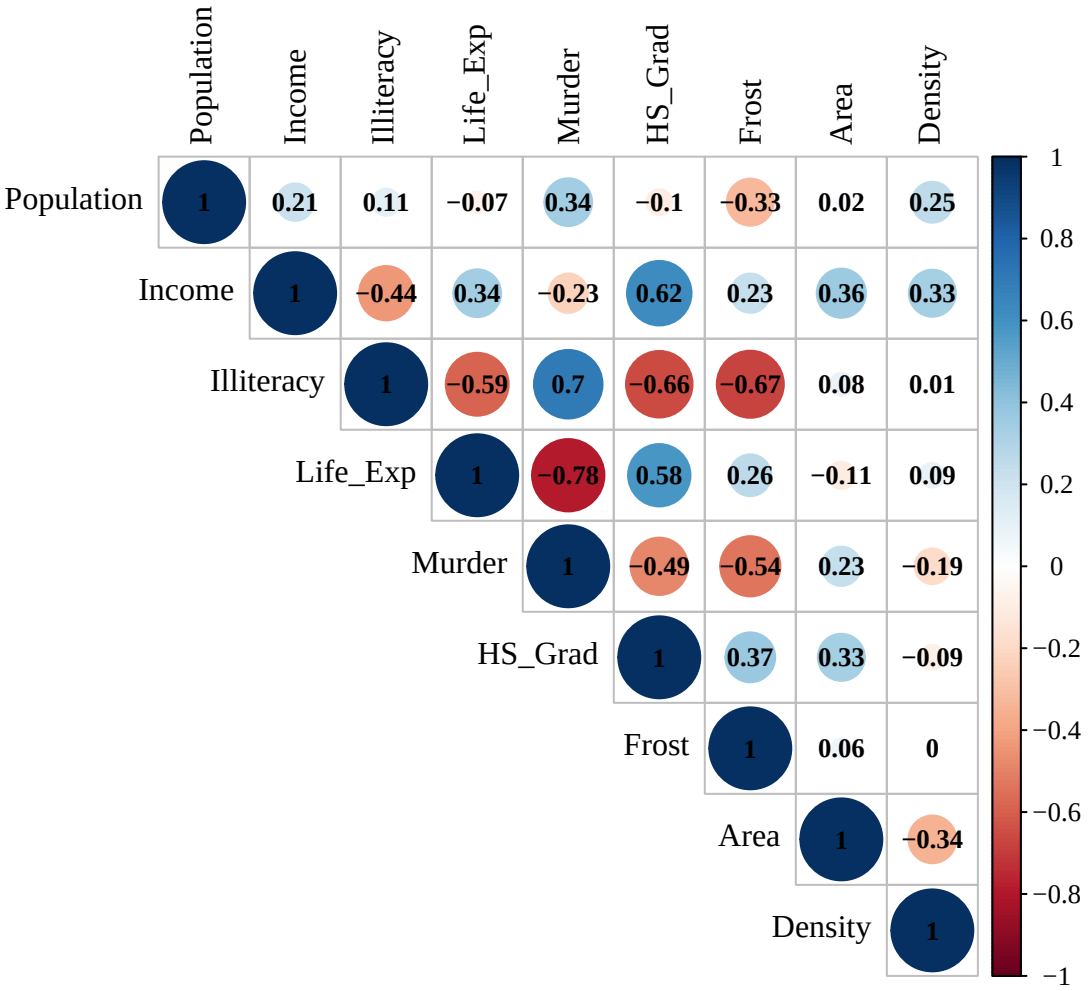
```
## Warning: package 'corrplot' was built under R version 4.2.3
```

```
corrplot(R, type = "upper", method = "circle", tl.col = "black", addCoef.col = "black",
         number.cex = 0.8)
```

By analyzing the correlation matrix, we can observe that:

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

- The variables with the higher positive correlation (0.7) are *Murder* and *Illiteracy*, which is credible since the two variables are intuitively bind;



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

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

1.2

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

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

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

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

In the following dataframes the first column refers to the index of the corresponding observation, while the second one refers to the variable with respect to this observation is a potential univariate outliar.

```
mat_1
```

	row	col
California	5	1
New_York	32	1
Alaska_2	2	2
Louisiana	18	3
Alaska_8	2	8
Massachusetts	21	9
New_Jersey	30	9
Rhode_Island	39	9

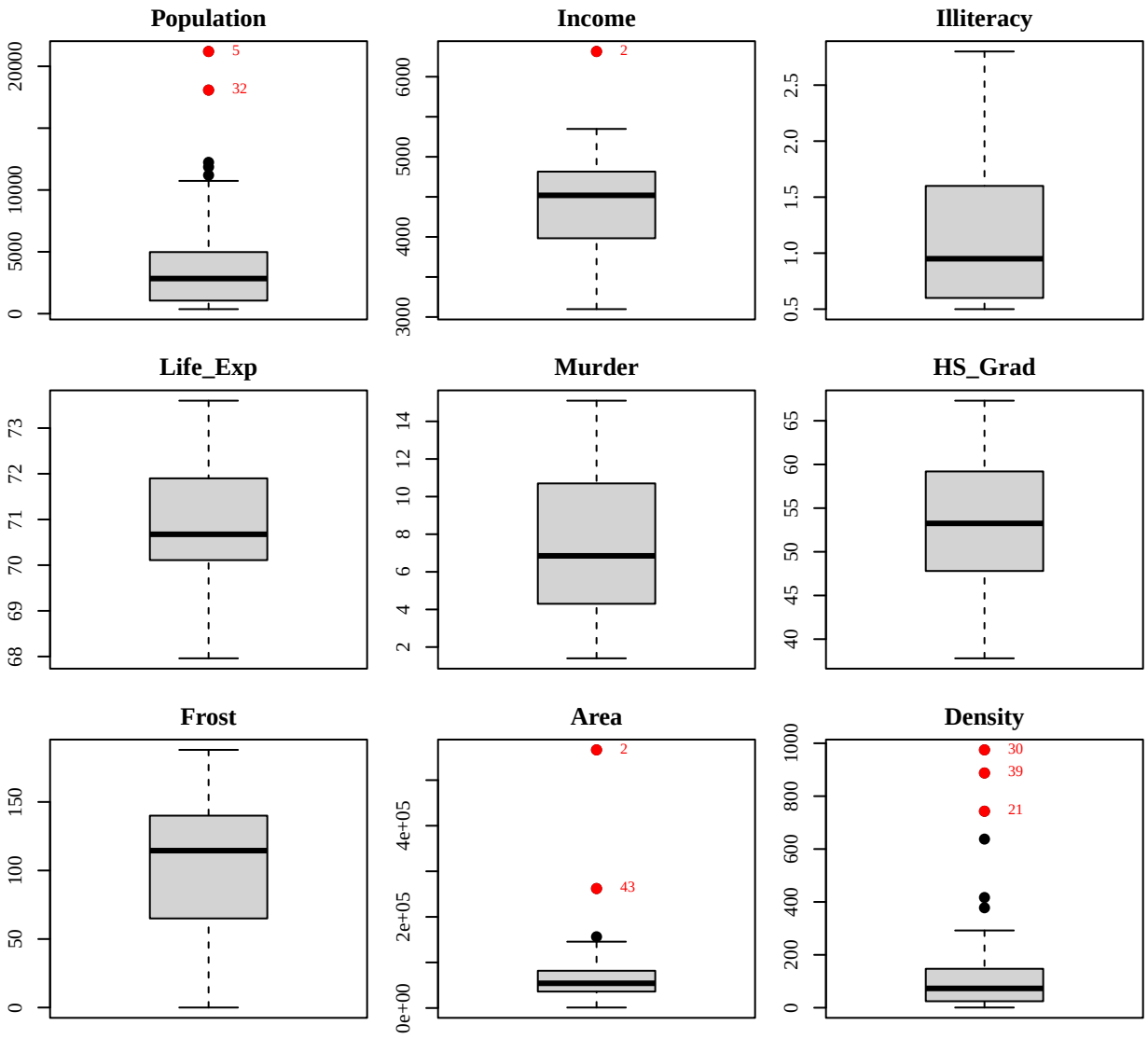
```
mat_2
```

	row	col
California	5	1
New_York	32	1
Alaska_2	2	2
Louisiana	18	3
Alaska_8	2	8
Texas	43	8
Massachusetts	21	9
New_Jersey	30	9
Rhode_Island	39	9

Note: In the tables above the names Alaska_2 and Alaska_8 both refer to the observation 2. We just had to change the rownames since in a dataframe we cannot have two rows with the same name. We choose the numbers 2 and 8 to highlight the index of the variable with respect to they are potential univariate outliers.

1.3

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



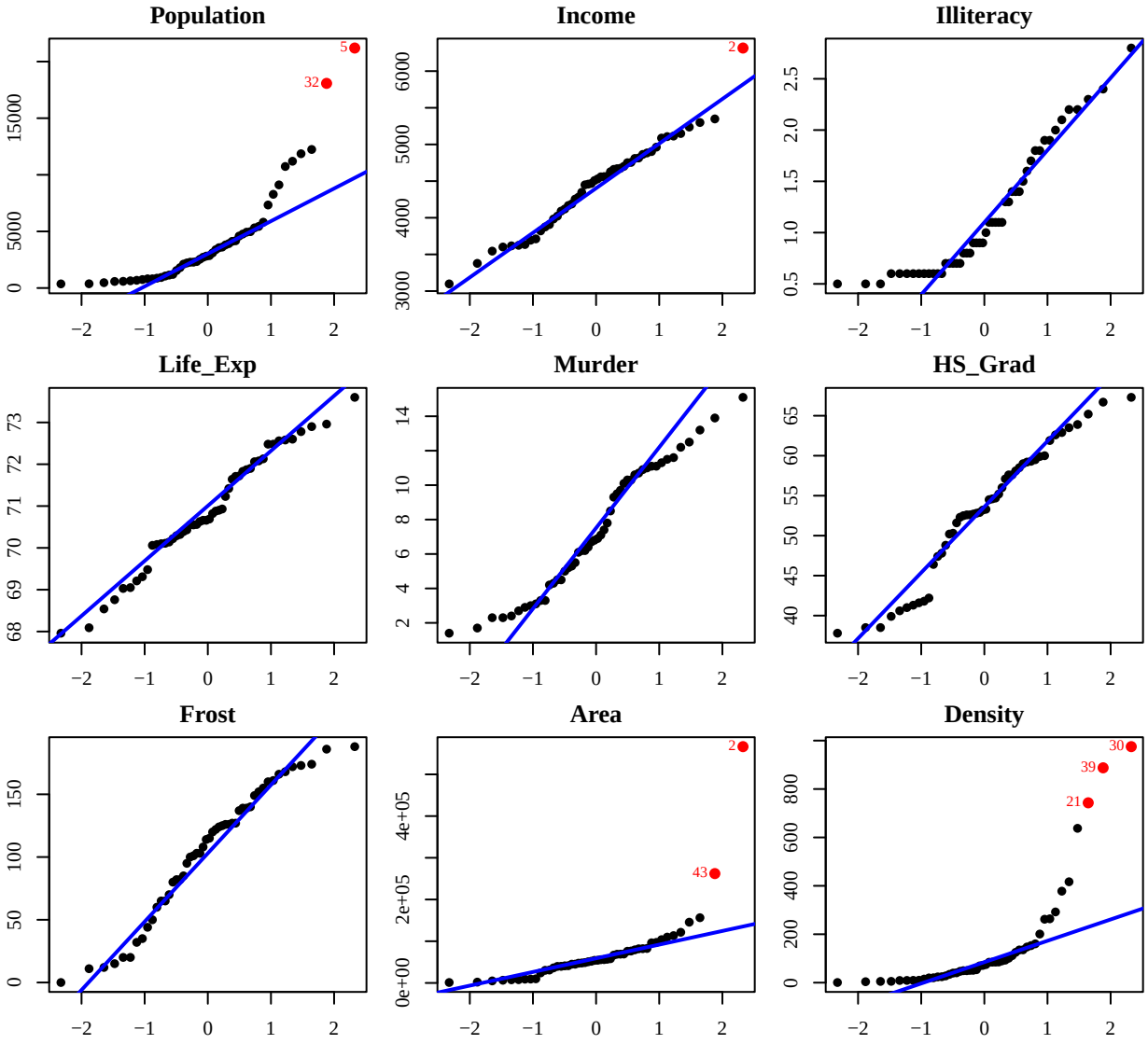
We can make the following considerations:

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

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

1.4

In order to check whether each variable is normally distributed or not, we first examine the relationship between the theoretical and the sample quantiles through the corresponding Q-Q plots.



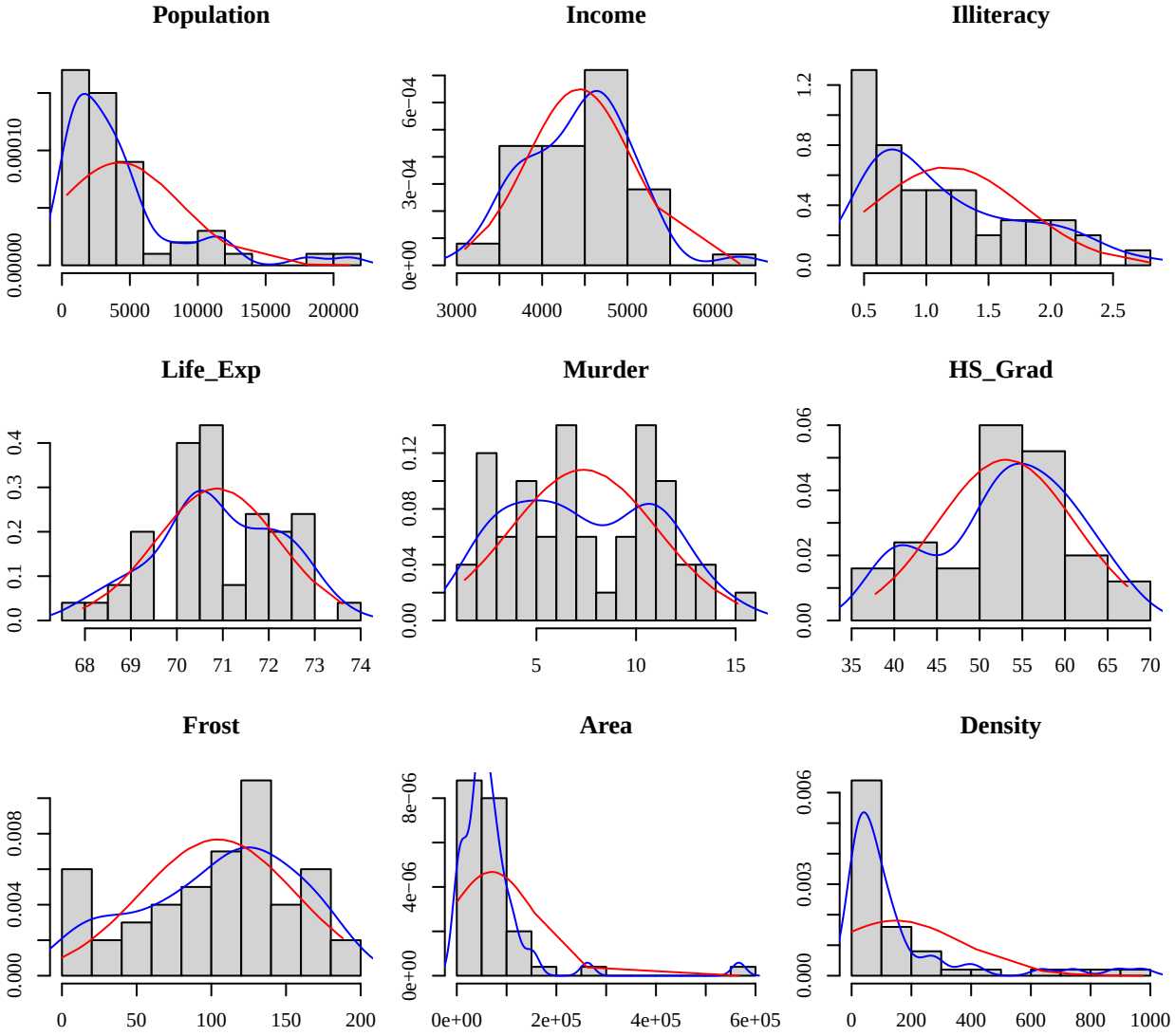
From the Q-Q plots we can observe that:

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

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

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

We can draw the same conclusions by observing the histograms of the single variables. In the following plots the blue line represents the empirical density, while the red line the theoretical density.



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

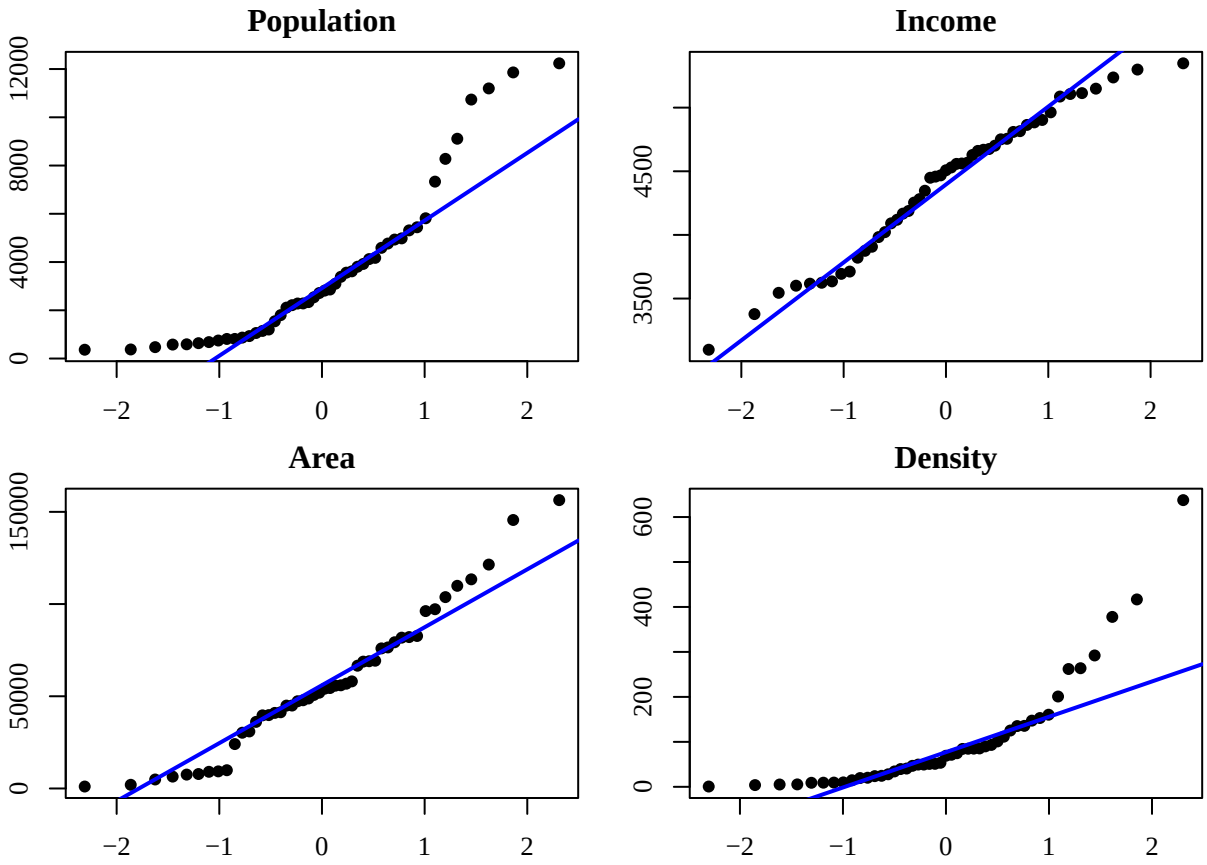
	Population	Income	Illiteracy	Life_Exp	Murder	HS_Grad	Frost	Area	Density
p.value	2e-07	0.4300105	0.0001396	0.4423285	0.0474463	0.0458156	0.0526747	0	0

As we can see, the variables *Income* and *Life_Exp* can be considered normally distributed, since their p-value is grater than 0.05. Moreover, the p-values of the variables *Murder*, *HS_Grad* and *Frost* are really close to

0.05, but the only one for which the null hypothesis is not rejected is *Frost*. However, its p-value is ≈ 0.0527, thus it is reasonable to doubt its normality, taking also in account the corresponding Q-Q plot and histogram presented above.

Now, it may be interesting to see how the previous tests change if we remove from each variable the observations we identified as their outliers. Let's take a look at the Q-Q plots and at the results of the Shapiro-Wilk test.

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

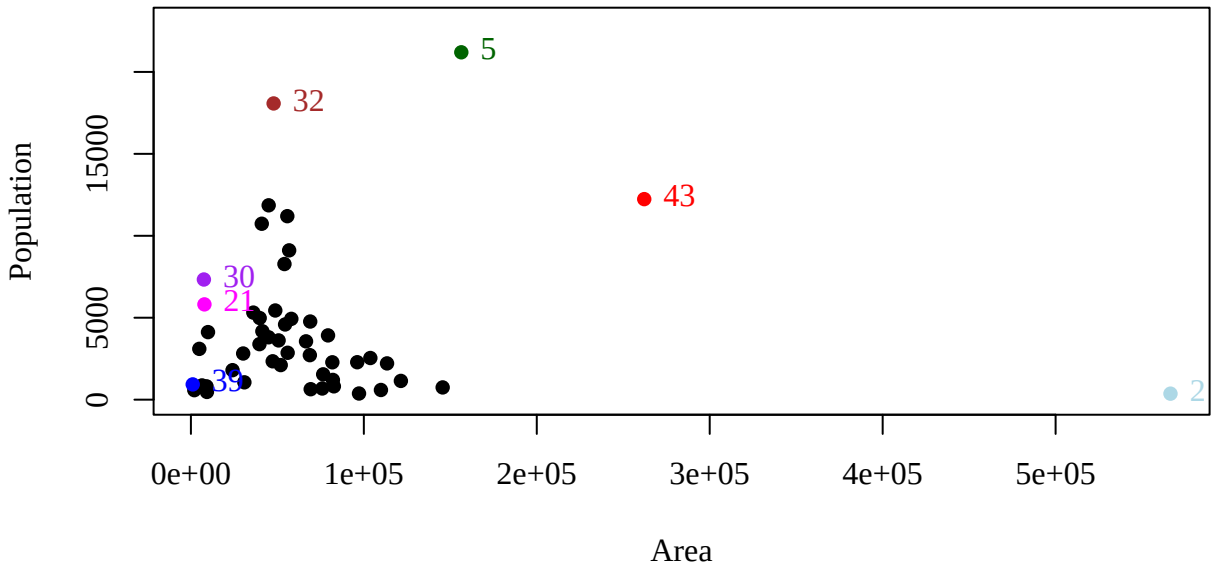


We can note that:

- The variables *Population* and *Density* follows the same non-gaussian beheaviour that we observed before; in fact, their p-values are almost negligible;
- As for the variable *Income*, we can still say that its distribution is gaussian. However, from the Q-Q plot we can observe the presence of more pronounced tails. Its p-value is still grater than 0.05 but a bit lower than the previous one;
- Finally, by removing its outliers, the variable *Area* seems to become gaussian: its p-value is very close to 0.05 but there is a huge difference between this value and the previous p-value.

1.5

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



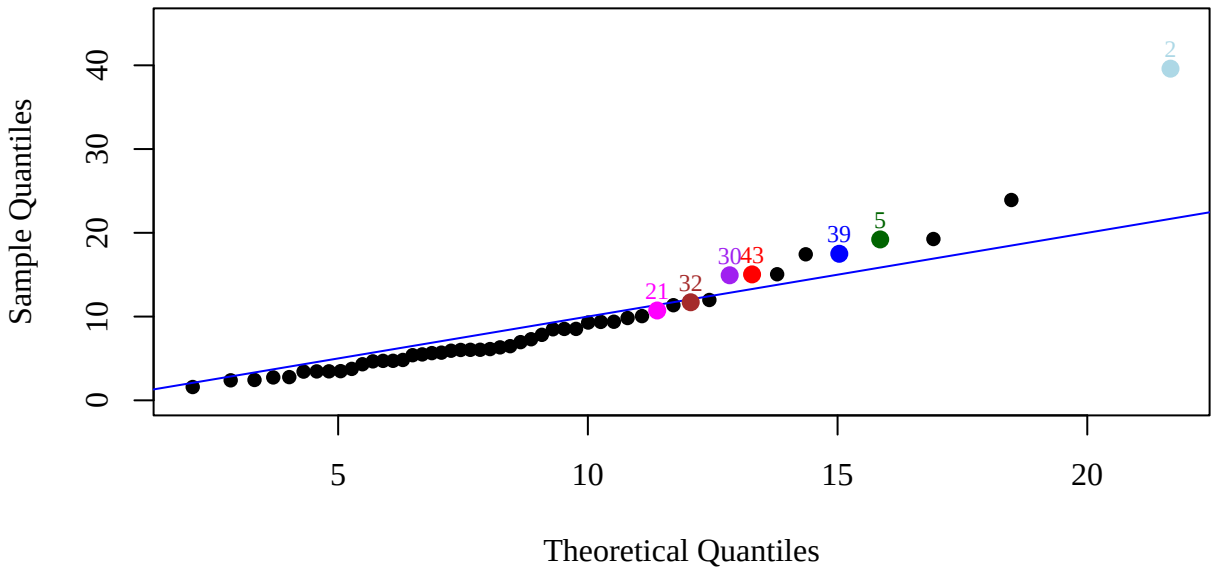
From the plot we can observe that all the mass is roughly concentrated in the rectangle $[0, 15\times10^4]\times[0, 15\times10^3]$. Also, unlike the other univariate outliers, the points corresponding to the observations $2, 5, 32$ and 43 seems to be really far from the other points. Hence, we can identify them as bivariate outliers.

1.6

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

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

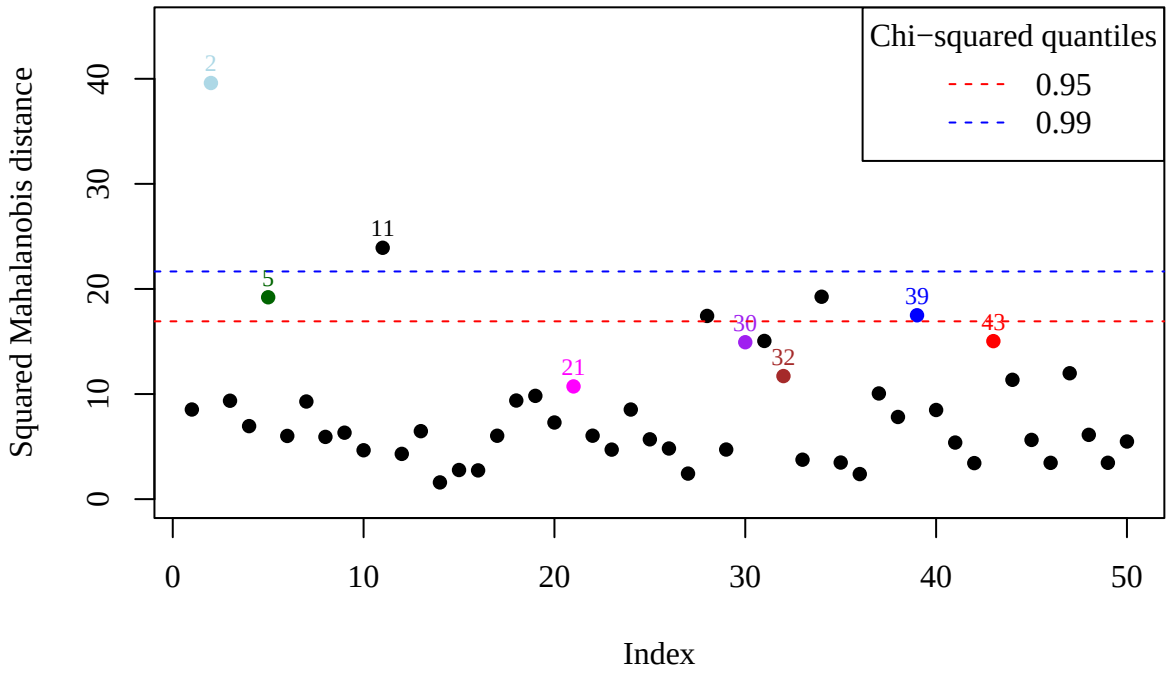
Hence, we can check the multivariate normality by looking at the Q-Q plot of the squared Mahalanobis distances *vs* a χ^2_9.



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

1.7

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



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