2nd HOME ASSIGNMENT

By Stefano Agresti

1ST EXERCISE

1. INTRODUCTION

The dataset used comes from FRED and it's a monthly report about unemployment rate in the US from 1948 to 2019 (https://fred.stlouisfed.org/series/UNRATE).

It contains 860 rows and no missing values.

2. QUESTION 1A)

Obtain a 97% confidence interval for the population mean.

***Interval**: (5.622449, 5.865923)*

We are 97% sure that the population mean is found in this interval.

```
> t.test(Unemployment_Rate, conf.level = 0.97)

        One Sample t-test

data:  Unemployment_Rate
t = 102.57, df = 859, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
97 percent confidence interval:
 5.622449 5.865923
sample estimates:
mean of x
 5.744186
```

Figure 1: T-test results

3. QUESTION 1b)

Perform a t-test on whether the population mean is equal to the sample median. Clearly state the null and alternative hypotheses, provide the p-value.

***Null hypothesis**: Mean is equal to median (5.6)*

***Alternative hypothesis**: Mean is not equal to median (5.6)*

***P-value**: 0.0102*

```
> t.test(Unemployment_Rate, mu = median)

        One Sample t-test

data:  Unemployment_Rate
t = 2.5745, df = 859, p-value = 0.0102
alternative hypothesis: true mean is not equal to 5.6
95 percent confidence interval:
 5.634264 5.854108
sample estimates:
mean of x
 5.744186
```

Figure 2: T-test results, checking mean = median

Since the p-value is quite small, we can reject the null hypothesis and assume that the mean is not equal to the median.

4. QUESTION 1C)

Obtain a 95% confidence interval for the population standard deviation.

***Interval**: (1.568255, 1.723902)*

We are 95% sure that the population standard deviation is found in this interval.

5. QUESTION 1D)

Find some dataset with a categorical variable. For that variable, compute the proportion of some level. Obtain a 99% confidence interval for that proportion.

Still using the same dataset, we select the rows where UNRATE > 6%, using this condition as categorical variable.

As a result, we obtain that 296 rows out of 860 respect the condition.

Therefore, we compute the confidence interval on this proportion, obtaining the following interval:

***Interval**: (0.3032309, 0.3875601)*

This means that we are 99% sure that the true proportion of unemployment rates greater than 6% is in this interval.

6. QUESTION 1E)

Perform a hypothesis test on whether the population proportion is equal to 1/2. Clearly state the null and alternative hypotheses, provide the p-value.

***Null hypothesis**: True proportion = 0.5*

***Alternative hypothesis**: True proportion <> 0.5*

***P-value** < 2.2e-16*

Since we were using the same condition, it's evident that the probability of it representing 50% of the entire population would be low. To confirm that, the p-value is almost insignificant, being in the order of e^-16.

7. QUESTION 1F)

Come up with some data for calculating the confidence intervals between proportions of two populations (in fact, you need just four numbers). Obtain a 99% confidence interval for the difference between proportions.

We split the dataset in two parts: rows referring to data until 1980 (included) and rows referring to data after 1980 (excluded). Then, we apply again the 6% condition, obtaining the following results:

***Until 1980**: 94 out of 396*

***After 1980**: 202 out 464*

The confidence interval for the difference on the two proportions is:

***Interval**: (0.1147103, 0.2812319)*

Since zero is not in the interval we can safely assume that the true proportions are different.

8. QUESTION 1G)

Perform an appropriate hypothesis test for the difference between proportions (perhaps, using imaginary data). Draw a conclusion.

As default hypothesis, we assume the two proportions to be equal. As a result, we obtain a p-value in the order of e^-9, which tells us that the hypothesis should be rejected.

This confirms the result of the previous point, where the confidence interval for the difference between the two proportions was shown not to include the zero value.

As a conclusion, we can affirm again that the two proportions are likely to be different from each other.

9. INTRODUCTION TO QUESTION 2

For this question we're using three datasets from FRED referring to the global prices of coffee, sugar and cotton. All of them are monthly reports and contain 355 rows. Initially the datasets included opening and closing prices, plus other similar information. To comply with exercise requests, those values have been replaced with log returns.

```r
coffee <- getSymbols("PCOFFOTMUSDM", src = "FRED", auto.assign = FALSE)
coffee <- na.omit(coffee)
coffee$log_returns <- diff(log(coffee), lag = 1)
coffee <- coffee[,2]
coffee <- na.omit(coffee)
```

Figure 3: R code used to compute log returns

10. QUESTION 2A)

Perform the Jarque-Bera for normality. State clearly the null and alternative hypothesis.

***Null hypothesis**: population is normally distributed*

***Alternative hypothesis**: population is not normally distributed*

The test has been executed using the *jb.norm.test* function in package "*normtest*", obtaining the following results:

***Coffee**: JB = 256.68, p-value < 2.2e-16*

***Sugar**: JB = 4.4477, p-value = 0.0965*

***Cotton**: JB = 202.4, p-value < 2.2e-16*

From these numbers, we can conclude that neither the coffee dataset nor the cotton one are normally distributed. The sugar dataset instead is characterized by a JB value small enough, preventing us from rejecting the null hypothesis.

11. QUESTION 2B)

Check whether the (univariate) empirical distribution of log returns for each stock is normal by examining the QQ-plot. Use the command qqPlot() from car package instead of the built-in function. Discuss whether the observations are within the confidence interval.

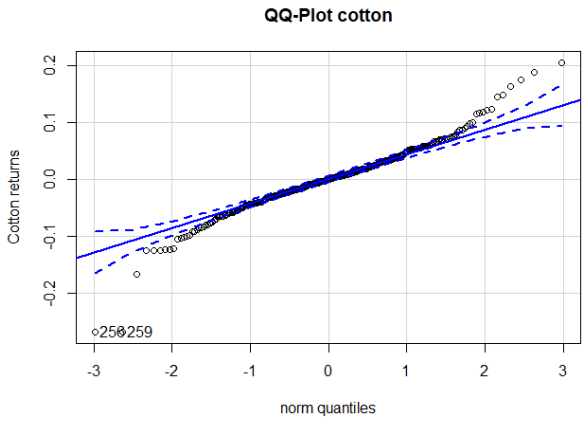

Figure 4: QQ-Plot for cotton log retuns

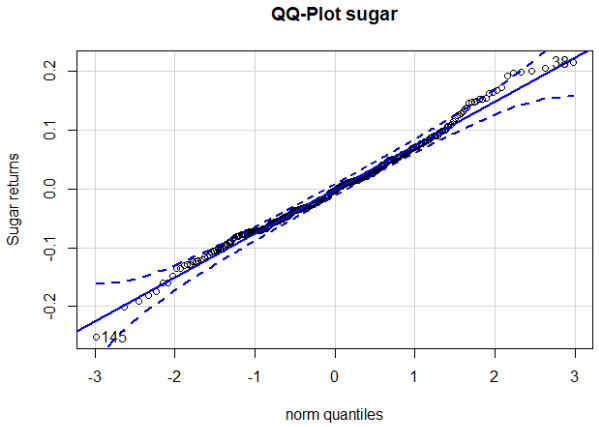

Figure 5: QQ-PLot for sugar log returns

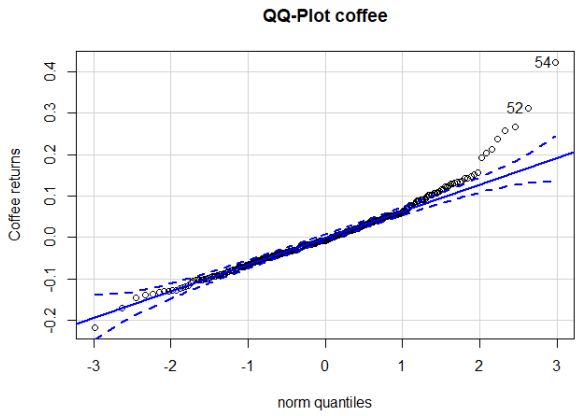

Figure 6: QQ-Plot for coffee log returns

From the graphs, we can observe how the sugar dataset is the one that seems to follow the most a normal distribution. Instead, coffee and cotton show a certain number of outliers towards the beginning and the end of the graph. This can be interpreted by assuming that their distribution, though following a bell shape, is probably skewed.

In general, however, observations tend to be inside the confidence interval, therefore, as we said, the distributions still maintain a bell shape (i.e. we're not handling an exponential or other kinds of distribution).

12. INTRODUCTION TO QUESTION 3

For this and the following question I used the "Crabs" dataset from the MASS package. The dataset contains various columns, but we will only focus on the Front Lobe dimensions for question 3. Only for this question we will also limit the length of the dataset to 50 rows (instead of the initial 200).

13. QUESTION 3)

Use a built-in set from 2 to perform the X^2-test for homogeneity (uniform distribution). Describe the data and discuss the result.

Using the Chi-Squared test, we get:

***X-squared**: 33.859*

***p-value**: 0.9509*

We also know that, for alpha = .05 and df = 49:

$$X_{alpha}\text{-squared}: 67.5$$

Observing the *p-value*, so close to 1, and the *X-squared*, much lower than the corresponding X_{alpha}, we can assume that the dimensions of crabs' front lobes are uniformly distributed.

14. INTRODUCTION TO QUESTION 4

To answer this question, we have to build a two-variable matrix. In order to do that, we divide the dataset between male/female and between crabs with front lobe greater/smaller than 14.

15. QUESTION 4)

Get a two-way contingency table from sources 3. Conduct a X^2-test for association (independence) between the variables.

***X-squared**: 0.35911*

***X_{alpha}-squared**: 3.841*

***P-value**: 0.549*

From these results, we can conclude that there's no relation between gender and having front-lobe greater than 14 ($X < X_{alpha}$).

2nd EXERCISE

1. INTRODUCTION

We use again the CRABS dataset from MASS, which contains 200 rows and 8 columns (each corresponding to a different characteristic of the crab).

2. QUESTION 1A)

Build a simple regression model (command lm). Provide the estimates of the model's parameters. Draw the scatter plot and the regression line.

To answer the question, we analyze the body length of crabs against their widths. The results of the linear model are:

***Intercept**: 1.089919*

***Slope**: 1.100266*

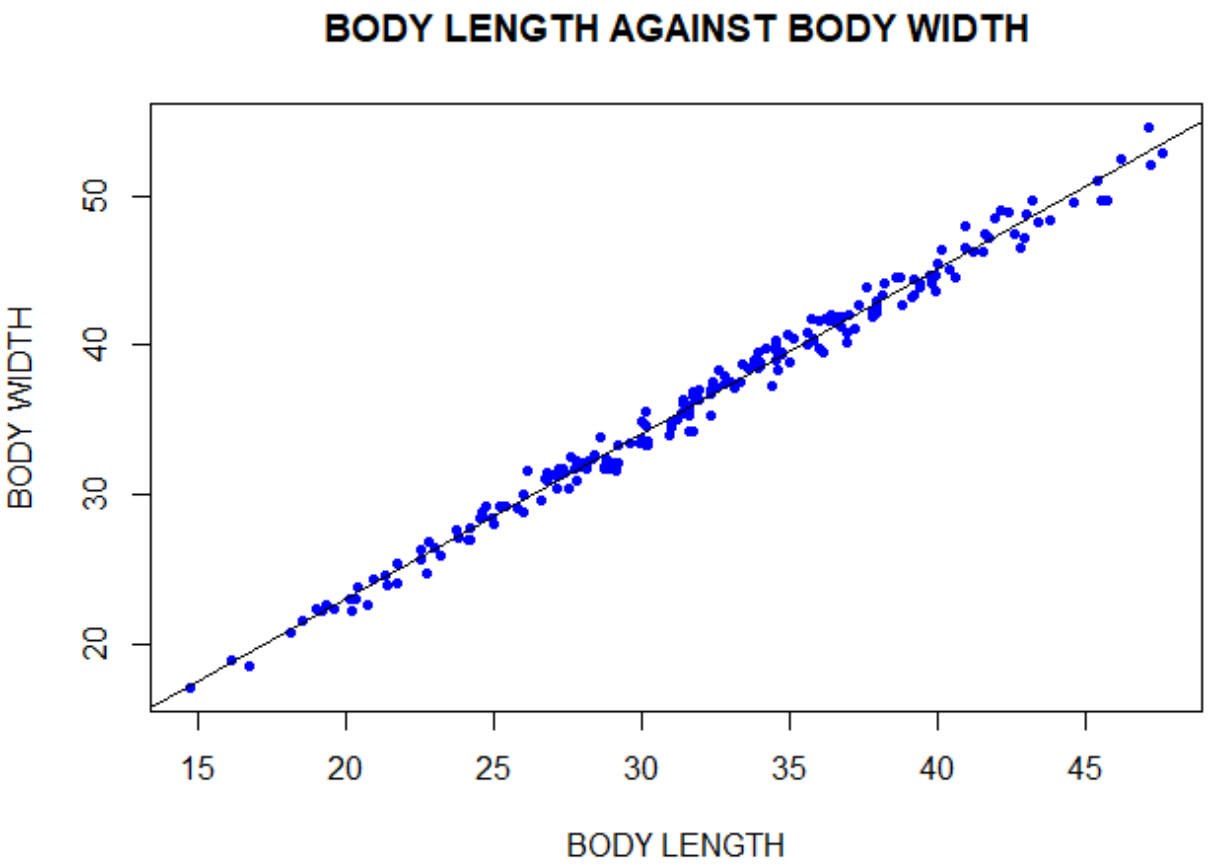

Figure 7: Graphical representation of the model

It's evident that a linear relation exists between the two characteristics of the crabs.

3. QUESTION 1B)

Analyze the summary statistics (command summary()) focusing on:

i. The t-test for the slope. Explain.

ii. The F-test. Explain.

iii. R^2 coefficient. Explain.

```
> summary(stats)

Call:
lm(formula = CW ~ CL, data = crabs)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7683 -0.6088  0.1075  0.5394  1.8092

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.089919   0.257490   4.233 3.53e-05 ***
CL          1.100266   0.007831 140.504  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7864 on 198 degrees of freedom
Multiple R-squared:  0.9901,    Adjusted R-squared:   0.99
F-statistic: 1.974e+04 on 1 and 198 DF,  p-value: < 2.2e-16
```

Figure 8: Result of the command "summary"

i. ***T-value for the slope**: 140.504.*

Such a big number for the t-value, combined with an almost irrelevant p-value (<2e-16), implies that we can reject the null hypothesis (i.e. no relationship exists between the two variables).
The t-value, together with the p-value, is used to check whether we are forcing to fit a model on a collection of data where no relationship exists or one of a bigger order is needed.

ii. ***F-test**: 1.974e+04.*

The F-test measures whether we're including variables in our model that are not significant. A big figure for the F-test, like the one we got, means that we can reject the null hypothesis (i.e. we're including a non-significant variable). This makes sense since our model only includes one variable and we already showed in the previous point how the model is not forced on the data.

iii. ***Multiple R-squared**: 0.9901*
***Adjusted R-squared**: 0.99*

Multiple R-squared is used to check how well the model fits the data. It's computed by subtracting the residuals from the variance of the data, which is then normalized. Another way of seeing this, R-squared measures the amount of variability explained by the variables used in our model. Hence, the bigger it is, the more our model explains the variations in the data.
Our model is therefore a good one, since it explains more than 99% of the data variability.
Adjusted R-squared is used to take into consideration the number of variables we're using. Multiple R-squared will always go up if we add variables, even if they're not really correlated. Adjusted R-squared will instead get worse if we add too many non-significant variables.
Again, our model is good from this point of view since we got 0.99 as a result. We could've guessed this as we're only using one variable, which we already proved to be significant.

4. QUESTION 1C)

Plot the residuals against fitted values and comment on the model's adequacy. Examine the qq-plot for the residuals.

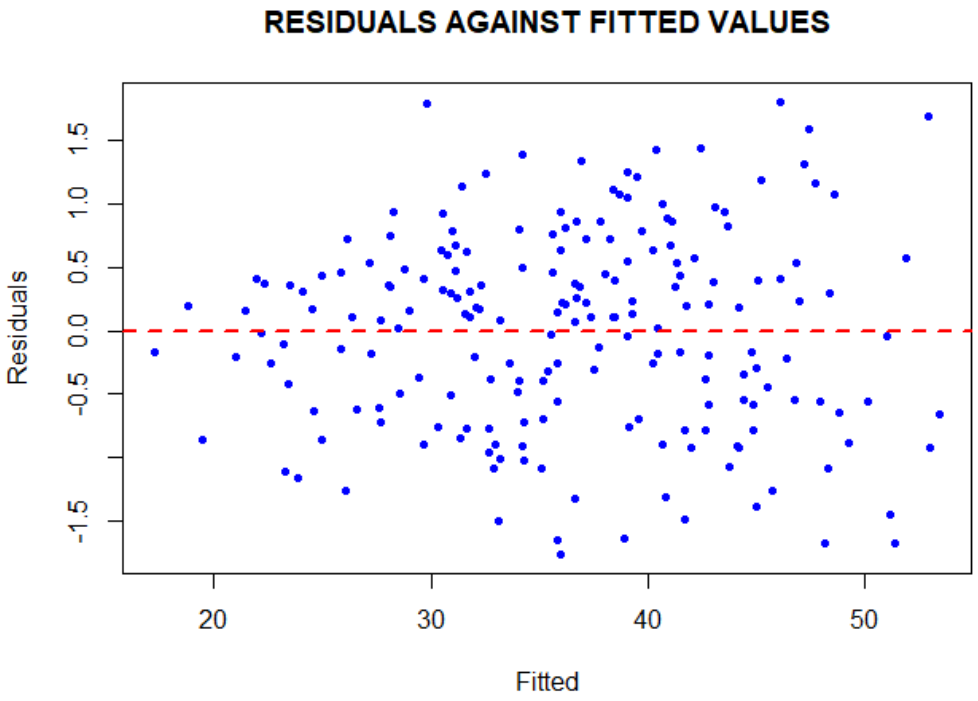

Figure 9: Scatterplot showing residuals' distribution

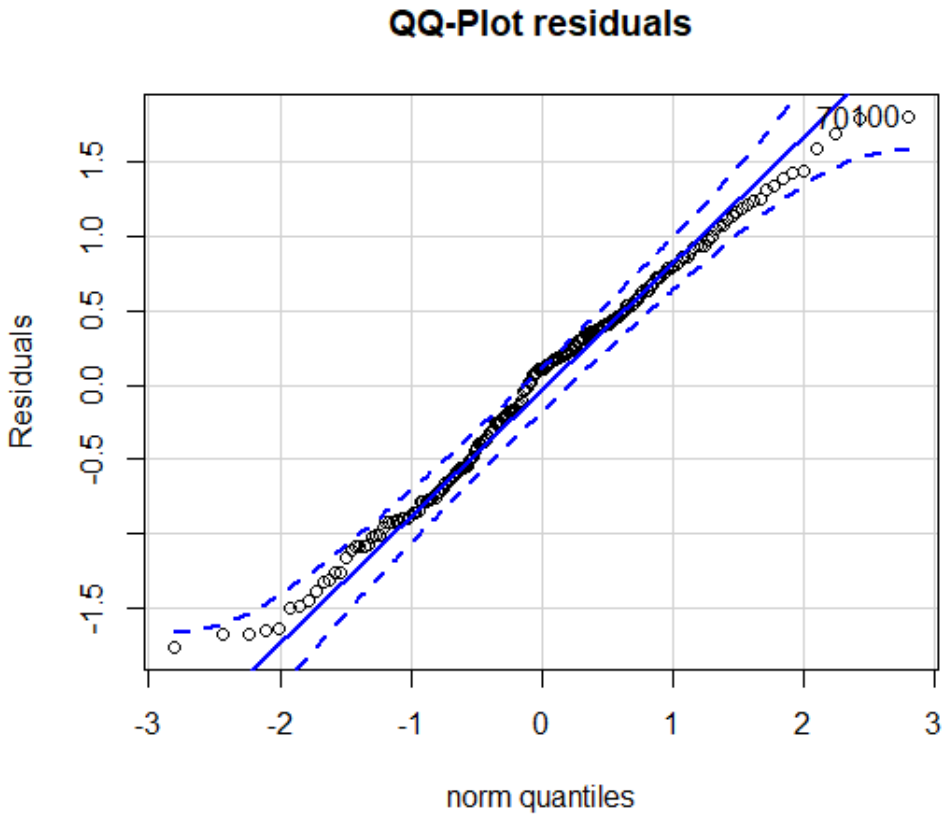

Figure 10: QQ-Plot for the residuals

From the graph we can observe how the points seem to be scattered randomly around 0. There's no outlier, nor any specific pattern.

This is a good sign for our model, because it means that the errors are equally distributed through the data, without any bias. Also, the fact that they bounce around 0 is an indicator that our linear model is a good estimator of the data.

The qq-plot as well seems to confirm that the residuals follow a bell-shaped distribution, though not perfectly formed

5. QUESTION 1D)

Make predictions for several new values of the independent variable. For each predicted value, compute and plot the confidence intervals for the mean and single value.

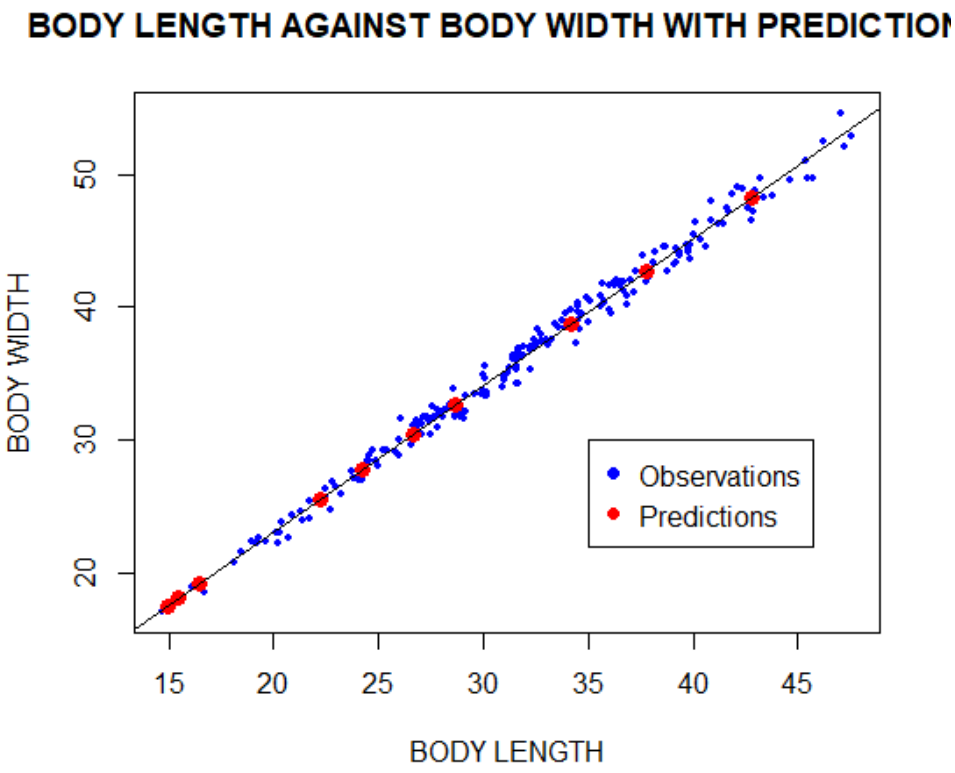

Figure 11: Plot with relation CL-CW with predicted values in red

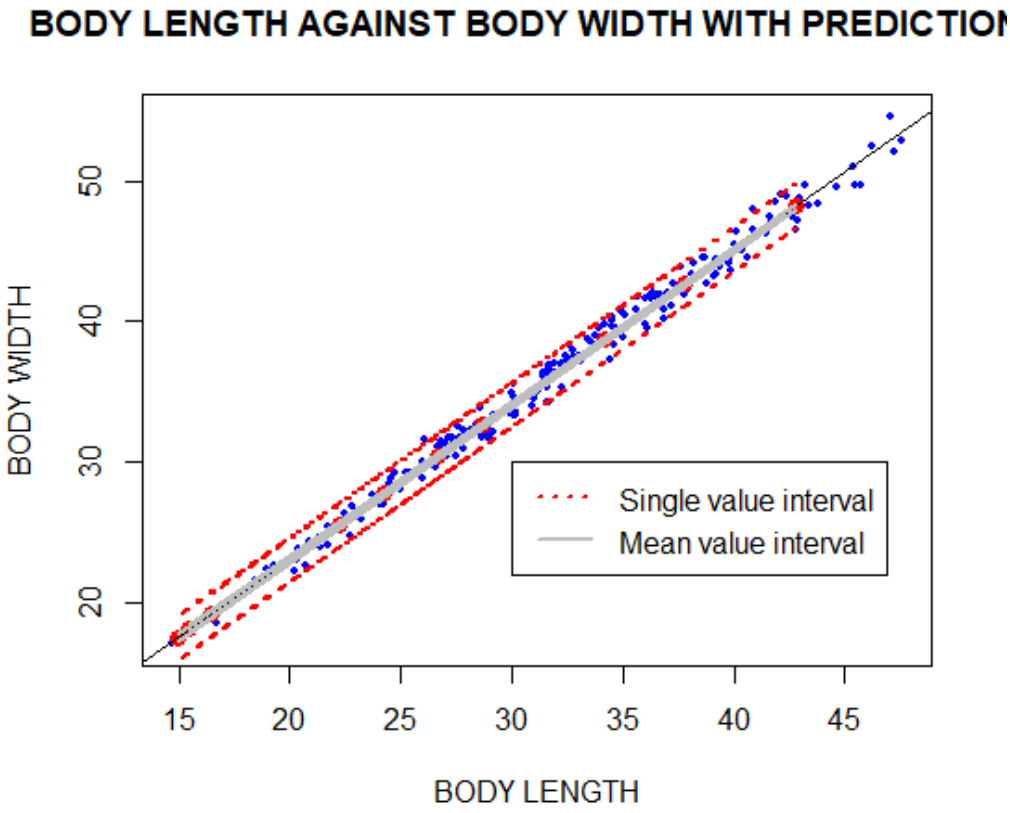

Figure 12: Same plot as in figure 10, but with confidence interval highlighted

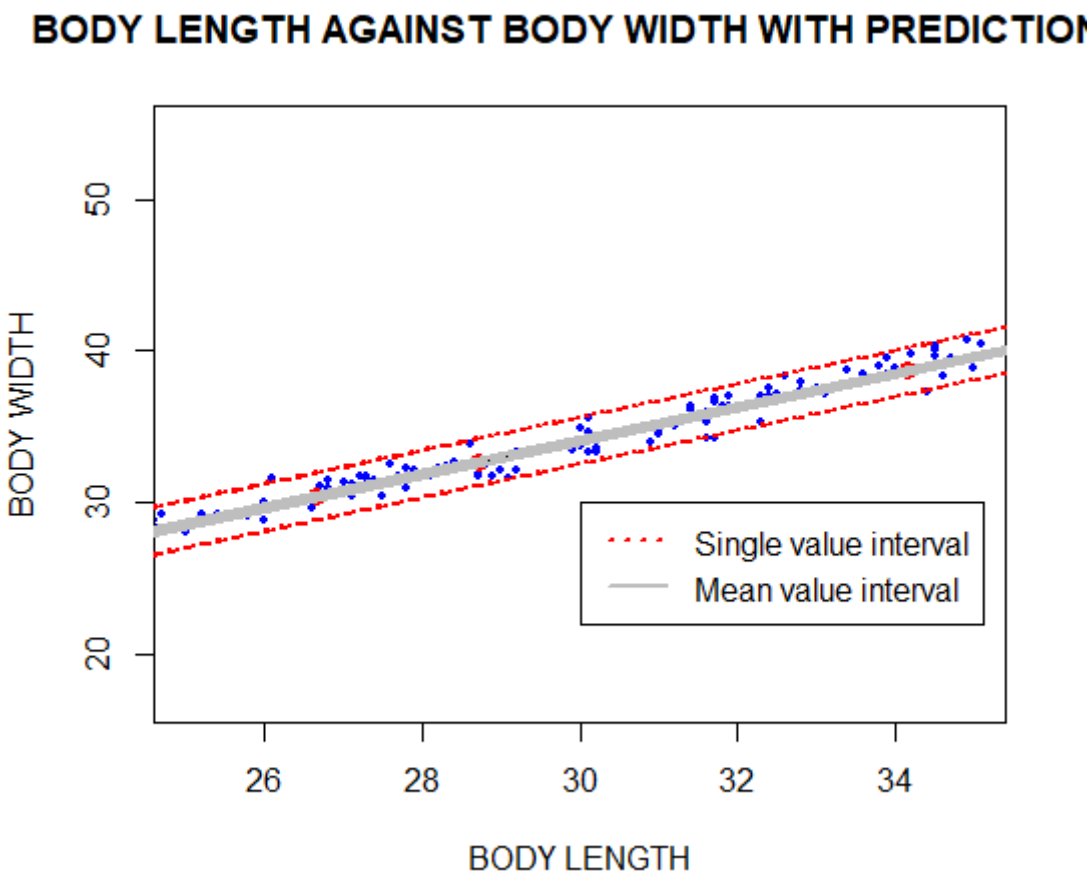

Figure 13: Close-up of the plot in figure 11

These three graphs show the distribution of CL with respect to CW, with some predictions added. In the table below there are the actual numerical values for these predictions.

CL values	CW predictions	Confidence interval (single)	Confidence interval (mean)
28.70688	32.67511	31.11952, 34.23070	32.55354, 32.79669
26.64143	30.40257	28.84558, 31.96956	30.26420, 30.54094
22.23369	25.52288	23.99072, 27.11504	25.36509, 25.74067
14.93827	17.52598	15.94884, 19.10313	17.23909, 17.81287
42.82880	48.21298	46.64948, 49.77648	48.01437, 48.41160
34.15984	38.67482	37.11979, 40.22985	38.56067, 38.78898
16.43327	19.17088	17.59745, 20.74431	18.90518, 19.43659
37.81047	42.69148	41.13428, 44.24868	42.55082, 42.83215
24.21851	27.73671	26.17724, 29.29618	27.57282, 27.90060
15.48727	18.13004	16.55429, 19.70578	17.85096, 18.40911

6. INTRODUCTION TO QUESTION 2

Once again, we're going to use the "CRABS" dataset, employing also values from the other columns.

7. QUESTION 2A)

Choose the response and explanatory variables.

***Response variable**: Crabs' body width (CW)*

***Explanatory variables**: Crabs' body length (CL), front lobe dimension (FL), rear width (RW)*

8. QUESTION 2B)

Build a multivariate linear model (command lm). Provide the estimates of the model's parameters.

***Intercept**: 0.40241*

***Crabs' body length**: 1.31908*

***Crabs' rear width**: 0.33285*

***Crabs' front lobe**: -0.67878*

9. QUESTION 2C)

Analyze the summary statistics (command summary()) with the emphasis on:

i. t-test for slopes. Explain.

ii. Overall F-test. Explain.

iii. R2 and adjusted R2 coefficients. Explain.

```
> summary(stats)

Call:
lm(formula = CW ~ CL + RW + FL, data = crabs)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6201 -0.4497  0.0230  0.3820  1.3015

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.40241    0.21954   1.833   0.0683 .
CL           1.31908    0.02969  44.429  < 2e-16 ***
RW           0.33285    0.03991   8.341 1.29e-14 ***
FL          -0.67878    0.06469 -10.492  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6091 on 196 degrees of freedom
Multiple R-squared:  0.9941,    Adjusted R-squared:  0.994
F-statistic: 1.101e+04 on 3 and 196 DF,  p-value: < 2.2e-16
```

Figure 14: Result of the "summary" command

i. ***T-value for CL**: 44.429*
***T-value for RW**: 8.341*
***T-value for FL**: -10.92*

All t-values are significant, meaning that we can reject the null hypothesis (i.e. coefficient equal to 0). However, it's interesting to notice how the t-value for the body length is much greater than the other, meaning that this coefficient is probably the most important one.

ii. ***F-test**: 1.101e+04*

The value for the F-test is again quite big, so we can safely assume that we're not including non-significant variables.

iii. ***Multiple R-squared**: 0.9941*
***Adjusted R-squared**: 0.994*

They're both close to 1 and close to each other. This means that the variables we chose explain almost all variation in the response variable and that all of them are significant to it (otherwise adjusted R-squared would be smaller).

10. QUESTION 2D)

Plot the residuals against fitted values and comment on the model's adequacy.

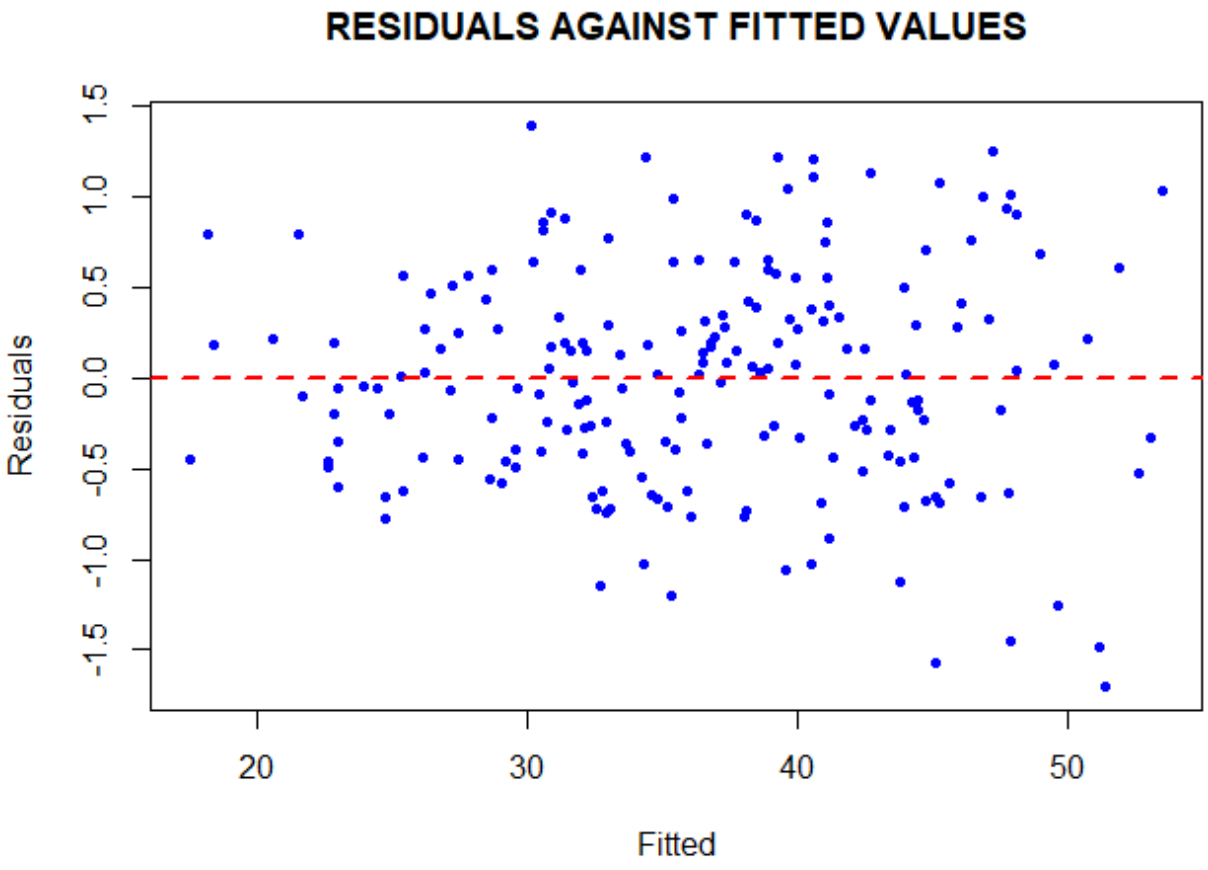

Figure 15: Plot with residuals' distribution

This plot is similar to the one we had in the previous question. Values are uniformly distributed around 0, without evident outliers or specific pattern. As before, this implies that we can assume that our linear model fits well the data.

11. QUESTION 2E)

Play with your model by adding or removing the explanatory variables. Alternatively, add a non-linear term(s) to your model:

i. Choose the best one by the AIC criterion (command stepAIC), see p. 295 of [1].

ii. For each model, watch the value of the adjusted R2. Explain

Single variable models				
Statistics	CL (Crabs' Length)	FL (Front Lobe)	RW (Rear Width)	BD (Body Depth)
Coefficient (est)	1.100266	2.1732	2.75437	2.22455
Intercept (est)	1.089919	2.5493	1.32798	5.20298
T-value (slope)	140.504	51.743	29.122	54.111
F-test	1.974e+04	2677	848.1	2928
Multiple Rsquared	0.9901	0.9311	0.8107	0.9367

Between the single variable models, it's clear how the best one is the one using the crabs' length as explanatory variable. The model largely outperforms the others on all statistics.

However, the other models don't perform poorly, as their statistics are still fairly good. Among them, the weakest one is the one using the rear width. This is the only model where multiple R-squared goes under 90%, F-test is under 1000 and t-test is below 50.

2- variable models				
Statistics	CL + sex (dummy variable)	CL + FL	CL + BD	FL + BD
1st coefficient (est)	1.103790	1.33329	1.44291	0.8351
2nd coefficient (est)	-0.476751	-0.48486	-0.72442	1.3828
Intercept (est)	1.215151	1.16418	0.25302	3.9999
1st t-value (slope)	146.716	38.741	41.079	3.333
2nd t-value (slope)	-4.462	-6.917	-9.922	5.408

F-test	1.082e+04	1.223e+04	1.478e+04	1544
Multiple R-squared	0.991	0.992	0.9934	0.94
Adjusted R-squared	0.9909	0.9919	0.9933	0.9394

From the 2-variable models, we can observe how there's no improvement from the single variable model with CL we analyzed in the previous points. That model remains the one with best F-test and the t-value for CL is always the largest in every model where the variable is present.

We can assume that CL is strictly related to CW and therefore can provide a good estimator by itself.

3-variable models		
Statistics	CL + RW + BD	CL + FL + BD
1st coefficient (est)	1.39451	1.45501
2nd coefficient (est)	0.24147	-0.10520
3rd coefficient (est)	-0.78683	-0.64311
Intercept (est)	-0.39326	0.36307
1st t-value (slope)	42.839	39.903
2nd t-value (slope)	6.722	-1.217
3rd t-value (slope)	-11.807	-6.503
F-test	1.208e+04	9876
Multiple R-squared	0.9946	0.9934
Adjusted R-squared	0.9945	0.9933

Again, no real improvement. It's actually interesting to notice how the coefficient for CL is always the largest, confirming that this variable is the most significant one.

5-variable model	
Statistics	CL + RW + B
1st coefficient (est)	1.45393
2nd coefficient (est)	-0.34628
3rd coefficient (est)	-0.52365
4th coefficient (est)	0.19379
5th coefficient (est)	-0.28314
Intercept (est)	0.15156

1st t-value (slope)	40.219
2nd t-value (slope)	-4.323
3rd t-value (slope)	-6.041
4th t-value (slope)	2.843
5th t-value (slope)	-1.823
F-test	7970
Multiple R-squared	0.9952
Adjusted R-squared	0.995

Out of curiosity, we can build a model using all of the variables, though the obtained results only confirm what we already observed: the CL variable is enough to create an estimation of CW, while other variables are less precise and mostly superfluous.

PS: the command stepAIC will be used in point 14, question 3b), on the same dataset. Answer to this question can be found there.

12. INTRODUCTION TO QUESTION 3

We're using once more the CRABS dataset, using crabs' sex as binary variable.

13. QUESTION 3A)

Build a logistic regression model (command glm). Comment on the significance of the coefficients.

```
Call:
glm(formula = sex_num ~ CL, family = binomial, data = crabs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3604  -1.1723   0.0111   1.1728   1.3858

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.95791    0.66358  -1.444    0.149
CL           0.02983    0.02019   1.478    0.139

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 277.26  on 199  degrees of freedom
Residual deviance: 275.04  on 198  degrees of freedom
AIC: 279.04

Number of Fisher Scoring iterations: 4
```

Figure 16: Result of the command glm

$$p(x) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 x}}$$

Figure 17: Formula used in glm models

Intercept and CL estimates are used in the model in place of B_0 and B_1. The value of 0.02983 in CL means that an increase of 1 in CL implies an increase of 0.02983 in the probability of the crab being a male.

However, we can notice how the p-value is pretty big, as is the AIC. Both indicates that this is probably not a good estimator for predicting the crabs' sex.

14. QUESTION 3B)

Use stepAIC command to select the best model.

***Best AIC**: 27*

Using the stepAIC command, we find that the best model to predict the crabs' sex is the one built with the combination of CL + FL + RW.

15. QUESTION 3C)

Make a prediction based on the entire dataset. State the threshold of acceptance. Compare the forecast with the actual observations. Comment on the result.

First, we generate new variables for CL, FL, RW inside the range of observation. Here's what we get:

***CL**: 18.45791*

***FL**: 17.32709*

***RW**: 18.08571*

Using the model computed in the previous point, we make a prediction using the *predict()* function and imposing 0.5 as threshold. The result we get is the following:

***Response**: 2.220446e-16*

Being the value so low, we can say that the model predicted a female specimen.

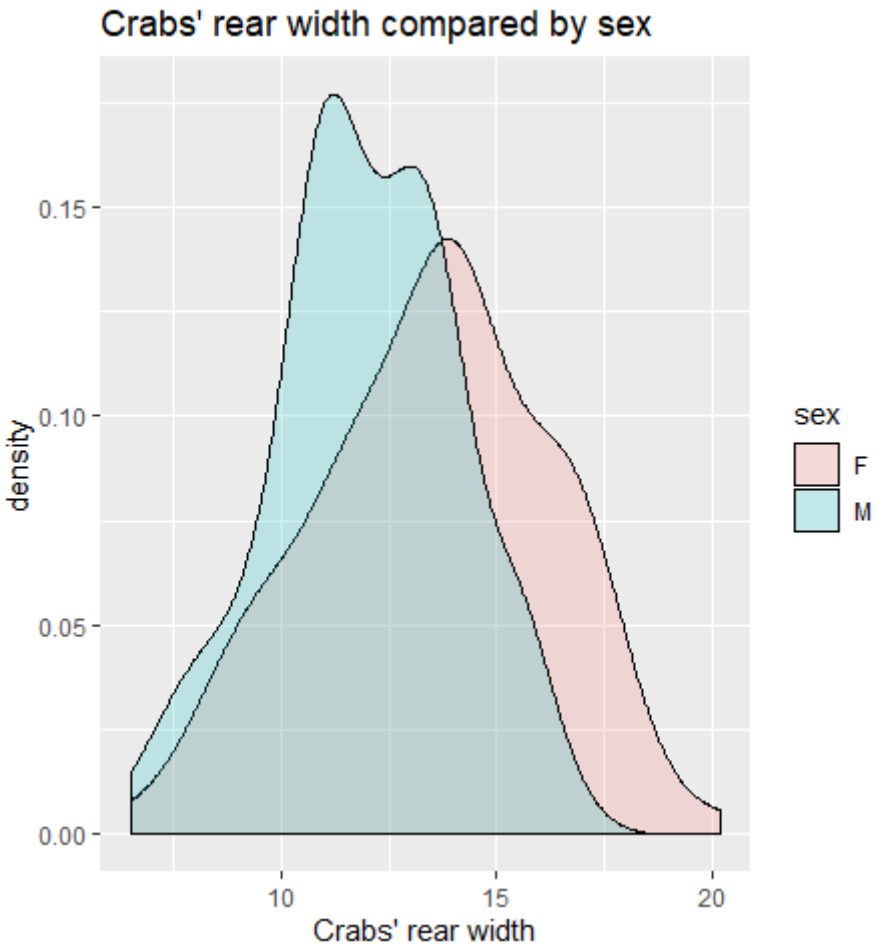

Figure 18: Distribution of CW divided by sex

Observing the actual data, it's interesting to notice how a rear width greater than ~17.5 is associated with female crabs. Since in our case the value was more than 18, we can say that our forecast is supported by real data.

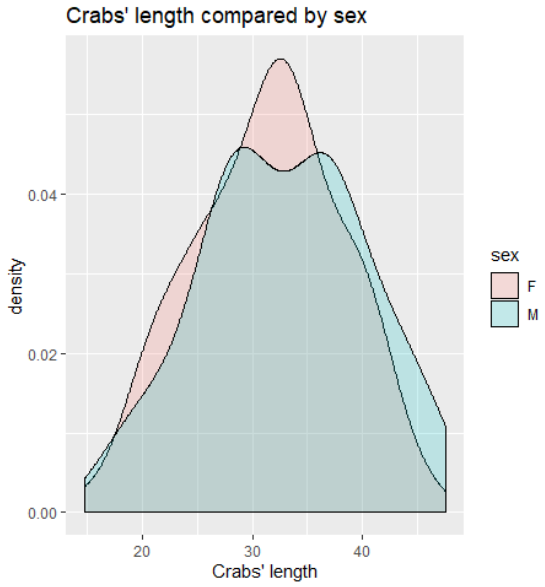

Figure 19: CL distribution divided by sex

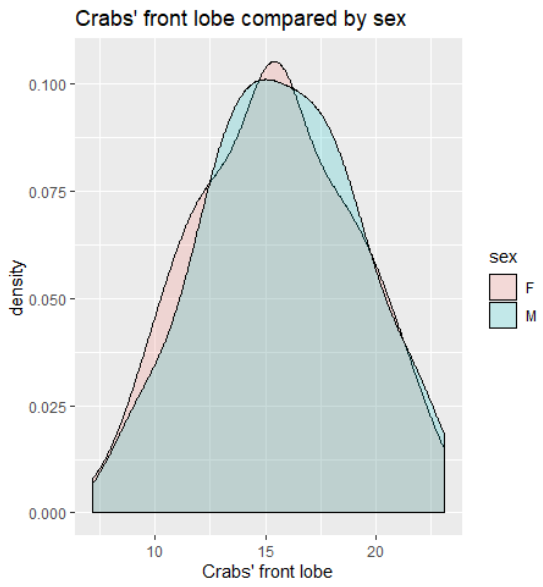

Figure 19: FL distribution divided by sex

Unfortunately, less information comes from the two other parameters, as we can see how the distributions of male and female basically overlap.

In conclusion, just by observing these distributions we can say that our prediction makes sense based on the rear width value. However, it's probably necessary to look for more data on crabs to have more precise forecasts.

16. QUESTION 3D)

Divide the entire set into training and test subsets. Rebuild the model using only the training subset. Make predictions for the test subset. Comment.

We divide the set into two subsets of length 150 and 50 rows (respectively train and test sets). Rebuilding the model using only the train and applying the *predict()* function on the test data, we obtain 49 correct predictions out of 50.

It's certainly good to see such a good percentage of correct guesses, though we should be aware of a too high success rate, because it might depend on some bias in the data or on similar issues. However, these results should encourage to try again the model on a bigger dataset to check if it's a good predictor in a real environment as well.

17. QUESTION 4A)

Conduct the linear discriminant analysis (command lda, package MASS) using training and test subsets. Compare the forecast with the actual observations. Comment on the results.

Prior probabilities of the two groups:

- ***Male**: 52%*
- ***Female**: 48%*

Coefficients of linear discriminants:

- ***CL**: 0.7602307*
- ***CW**: -0.1841087*
- ***RW**: -1.6595610*
- ***FL**: -0.1897579*
- ***BD**: 0.2144373*

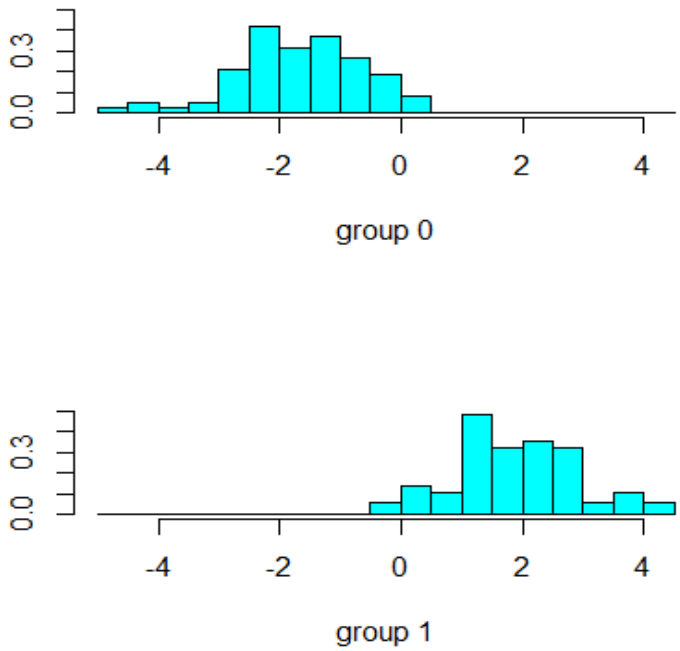

Figure 20: Graphical result of lda command

After prediction, we obtain 47 correct guesses out of 50 values in the test set, with a success rate of 94%. As in the previous question, such a high rate may lead us to question whether the data is biased or has some kind of issues, but, again, the goodness of our predictions should bring us to try the model on a larger dataset.

18. QUESTION 4B)

Conduct the quadratic discriminant analysis (command qda). Comment.

```
Call:
qda(sex_num ~ CL + CW + RW + FL + BD, data = crabs)

Prior probabilities of groups:
  0   1
0.5 0.5

Group means:
      CL     CW     RW     FL     BD
0 31.360 35.830 13.487 15.432 13.724
1 32.851 36.999 11.990 15.734 14.337
```

Figure 21: Result of the command qda

This time the probabilities for a crab of being male/female are split evenly.

Below this information, *qda* shows the mean of the various characteristics divided by the gender. We can notice how the characteristics that differ the most are CL, CW and RW, while FL and BD are similar. This agrees with the model we previously found, which was using CL and RW. Curiously though, that model also included FL, parameter that's instead the most similar between males and females.

19. QUESTION 5A)

Conduct the KNN classification (command knn(), package class) using training and test subsets. Compare the forecast with the actual observations. Comment on the results.

Using the KNN classification, we obtain 45 correct predictions out of 50 values in the test set, which is a 90% success rate. Although it's still very high, it is worse compared to the previous models.

However, we need to point out once again, that these tests should be repeated on a larger dataset in order to confirm their actual goodness.

20. Question 5b)

Play with the number of nearest neighbors K.

Number of K neighbors	Correct predictions (out of 50 values)
1	45
2	43
5	47

8	47
10	42
20	44
30	39
50	31
75	24
100	26
150	24

It's interesting to notice how the model doesn't improve its performances increasing the number of K-neighbors. On the contrary, as can be seen on the table, the correct predictions are only around 50% for 75, 100 and 150 neighbors.

21. QUESTION 6)

Compare the quality of classification obtained by algorithms 3-5 for the test subset.

Model	Percentage of correct predictions
Logistic Regression Model	98%
Discriminant Analysis	94%
KNN Classification	94% (best) – 48% (worst)

Considering the small dimensions of the dataset, we don't have enough information to clearly declare one of the models better than the other.

However, the one that gave the best result was the Logistic Regression Model, with 49 correct predictions out of the 50 rows in the test set.

The worst one was probably the KNN classification. Although it got the same results as the Discriminant Analysis when using 5 and 8 neighbors, it was worse in the other cases. It even went below 40 out of 50 correct predictions when using more than 20 of them.

In conclusion, as previously stated, the models should be tested on larger datasets to be sure of their quality, but for the moment we can be satisfied with the obtained results, which look encouraging enough.

3RD EXERCISE

1. INTRODUCTION

For this exercise we will use a dataset taken from Kaggle.com with information about a series of red wines (link https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009/download). The dataset includes 12 columns and 1599 rows, but for the purposes of this exercise I decided to limit them to 100.

I've also modified the variable *quality* to make it binary (1 if quality is greater than 5, 0 otherwise).

2. QUESTION 2A)

Plot the individuals in the plane corresponding to the first two principal components (PCs), see [4], p.31. Comment on the resulting cloud.

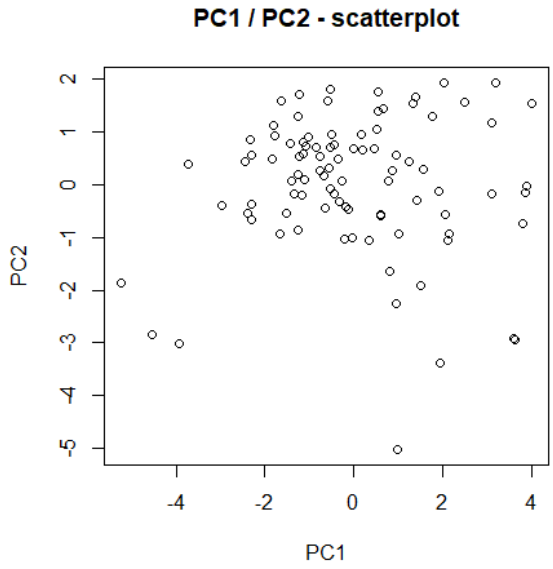

Figure 22: Scatterplot relative to PC1/PC2

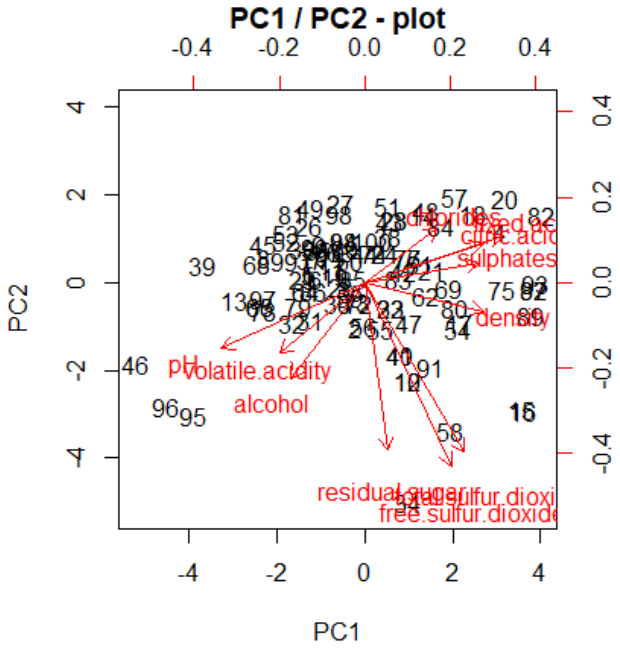

Figure 23: Visualization of variables and rows wrt PCs

The scatterplot doesn't really give many information by itself. Data seems to be clustered around (0,1) with several outliers.

However, when we visualize the variables and their importance for the PCs, we can notice how some of them are much more important than others. In particular, residual sugar, free sulfur dioxide and total sulfur dioxide have the longest arrows, meaning that they explain a great part of data variability.

3. QUESTION 2B)

Justify the choice of the PCs by plotting the eigenvalues, [4],p.32. Calculate how much of the total variability is explained by the first two PCs.

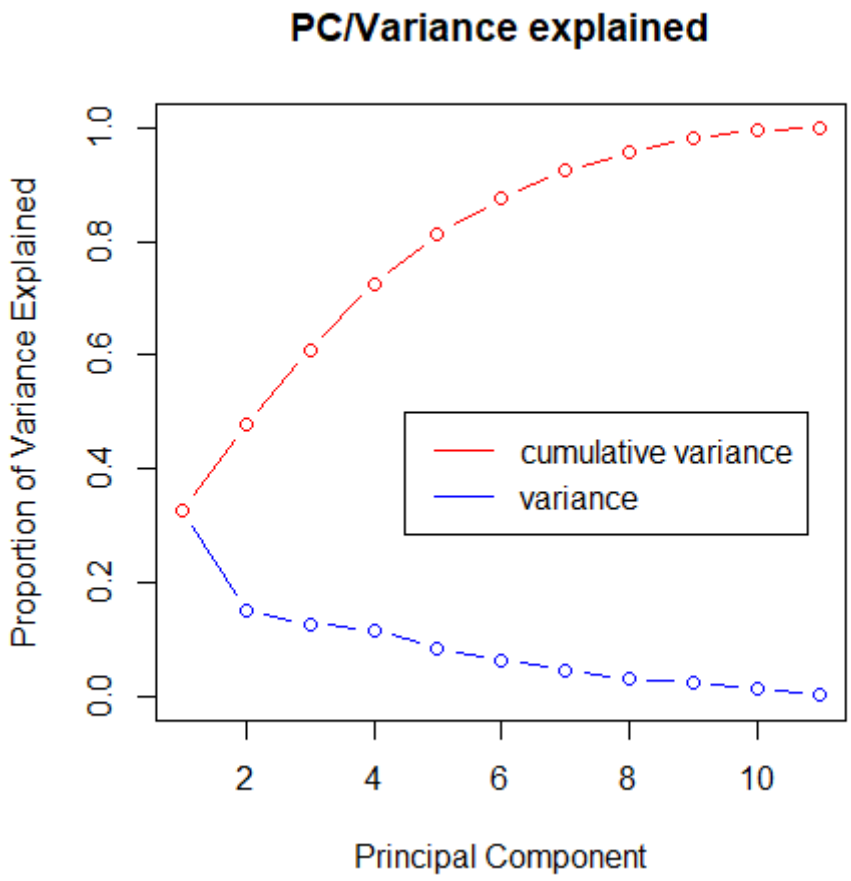

Figure 24: Variance explained by the PCs

As we can see, the first two PCs explain together roughly 50% of the variance, while combining the first five of them would instead explain more than 80% of it.

This shows the usefulness of PCA: instead of using eleven variables to observe the data, we can reduce its dimensions to a much smaller number, certainly easier to handle, but still maintaining most of the information.

4. QUESTION 2C)

Discuss the quality of the PCA representation: provide cos2 and the contributions for each individual, [4], p.34.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
1	2.804842e-01	6.470333e-03	2.198550e-01	1.839408e-01	0.0073423088
2	6.069289e-03	1.543136e-01	1.397677e-03	6.283236e-01	0.0098169315
3	6.899791e-02	9.917838e-03	2.371935e-02	6.274256e-01	0.0002827248
4	4.577380e-01	6.405671e-02	1.089301e-01	1.685427e-01	0.0160915607
5	2.804842e-01	6.470333e-03	2.198550e-01	1.839408e-01	0.0073423088
6	2.441844e-01	1.773028e-03	2.075154e-01	1.507706e-01	0.0274212219
7	1.198217e-01	3.763374e-02	4.051796e-02	1.491742e-01	0.4571874266
8	6.137181e-01	3.716806e-02	4.930595e-02	1.467257e-02	0.0886893683
9	3.026489e-01	2.320344e-01	2.199209e-01	1.070393e-01	0.0135006309
10	7.901496e-02	4.348665e-01	6.916729e-02	1.086672e-01	0.1915351084
11	1.751048e-01	2.225271e-02	5.234064e-02	1.770220e-01	0.2892550568
12	7.901496e-02	4.348665e-01	6.916729e-02	1.086672e-01	0.1915351084
13	6.697652e-01	1.245243e-02	1.747869e-01	2.849392e-04	0.0545186986
14	1.567639e-01	2.145005e-01	3.043641e-02	7.849297e-02	0.0432754933
15	4.935683e-01	3.242103e-01	1.463166e-04	1.010028e-01	0.0212978301
16	5.073417e-01	3.348520e-01	2.165601e-04	9.467244e-02	0.0183775465
17	3.392259e-01	6.387347e-02	5.708259e-02	4.166316e-01	0.0485231558
18	2.826815e-01	1.128493e-01	2.203872e-01	1.381878e-01	0.1471147424
19	7.367115e-02	4.121309e-03	5.683340e-01	1.000595e-01	0.0392958476
20	4.422965e-01	1.612255e-01	1.431644e-01	1.333345e-03	0.0729288024
21	2.248177e-01	7.997837e-03	2.226715e-02	3.776412e-01	0.1933721434
22	6.913263e-02	6.362207e-02	1.001669e-01	2.392419e-01	0.0327246422
23	1.337682e-01	6.173909e-01	1.600381e-02	4.590133e-04	0.0016239303
24	6.073062e-02	1.314365e-01	1.316500e-01	3.538273e-02	0.0909053806
25	1.115541e-01	3.167065e-03	1.843663e-02	1.056666e-01	0.0516716435
26	2.486092e-01	2.638166e-01	2.156965e-02	2.608348e-02	0.1153548803
27	6.220460e-02	7.043188e-01	4.294893e-02	7.217060e-02	0.0052370043
28	1.337682e-01	6.173909e-01	1.600381e-02	4.590133e-04	0.0016239303
29	4.292126e-01	1.293423e-03	1.009006e-01	2.709072e-01	0.0325531218
30	4.984055e-01	1.522590e-01	6.868145e-02	1.190008e-01	0.0012851150
31	4.214531e-01	2.002615e-01	1.466084e-01	5.398000e-02	0.0375665081
32	5.122264e-01	1.643005e-01	8.115479e-03	3.137763e-02	0.0356031049
33	8.215397e-02	7.502499e-02	1.728768e-02	1.171972e-01	0.0913241944
34	1.975387e-02	5.074287e-01	2.526599e-01	9.325659e-03	0.0689306560
35	1.202103e-01	5.582846e-02	6.475325e-02	7.172444e-02	0.0687995543
36	3.266326e-02	1.712830e-02	6.137460e-01	7.457121e-02	0.2253107142

Figure 205: Cos2 for rows 1-36

36	3.266326e-02	1.712830e-02	6.137460e-01	7.457121e-02	0.2253107142
37	2.507384e-01	4.563480e-02	1.034243e-01	3.731752e-02	0.2781571934
38	9.424972e-02	3.754506e-01	7.515165e-02	2.002974e-01	0.0189548649
39	5.153159e-01	5.478117e-03	1.194958e-01	1.753845e-01	0.0054580753
40	6.161163e-02	2.578590e-01	7.967711e-02	1.520580e-01	0.2789720682
41	6.161163e-02	2.578590e-01	7.967711e-02	1.520580e-01	0.2789720682
42	2.107657e-01	1.987120e-02	4.995341e-01	2.569003e-02	0.0270077809
43	1.992857e-02	1.273343e-01	1.100326e-01	1.695207e-02	0.5234595023
44	6.201126e-06	8.621427e-02	1.254095e-02	4.573733e-03	0.1835976773
45	7.538404e-01	1.036913e-01	4.741732e-02	4.763683e-02	0.0138436710
46	5.042154e-01	6.369775e-02	1.619392e-01	2.200466e-01	0.0355301364
47	9.841217e-02	8.496204e-02	3.426572e-02	1.722751e-01	0.0470465770
48	2.333016e-01	3.360169e-01	2.298806e-02	3.165470e-01	0.0171022546
49	2.171246e-01	4.140329e-01	1.552076e-03	4.657204e-02	0.0525936847
50	1.116672e-02	2.178471e-02	1.435535e-01	1.531753e-01	0.1991332634
51	4.479411e-02	4.844884e-01	1.232065e-01	9.522067e-02	0.0036729574
52	5.674712e-01	1.555650e-01	1.178824e-03	1.060345e-02	0.0173513309
53	5.253590e-01	2.040291e-01	1.698921e-04	4.365461e-03	0.0147187238
54	4.997473e-01	1.281089e-01	5.653947e-03	6.088359e-02	0.1976162987
55	1.930057e-02	1.895551e-01	1.651271e-01	4.453552e-05	0.0557259514
56	4.523039e-04	2.316695e-01	4.335876e-01	1.957978e-01	0.0002835895
57	2.341286e-01	2.121729e-01	1.791453e-01	7.636168e-02	0.0169192806
58	1.940180e-01	5.984735e-01	1.829675e-02	8.469385e-02	0.0074932250
59	8.091838e-02	8.494950e-02	2.890837e-01	1.610874e-02	0.0035505689
60	6.184689e-02	1.839416e-01	4.137767e-02	3.459514e-01	0.0691282525
61	2.572338e-01	3.019385e-02	2.688843e-01	1.528094e-01	0.0645666788
62	2.963918e-01	1.364749e-02	8.756934e-02	7.353611e-03	0.1118850169
63	2.387669e-01	4.024328e-01	1.438982e-01	2.235927e-04	0.0851767099
64	6.094088e-01	1.199190e-02	5.830485e-05	2.691182e-01	0.0214076613
65	4.866563e-01	2.515253e-02	7.741187e-02	1.049873e-04	0.2798103065
66	4.866563e-01	2.515253e-02	7.741187e-02	1.049873e-04	0.2798103065
67	3.546657e-01	2.409376e-01	6.619549e-02	1.043758e-04	0.1137841241
68	7.834895e-01	2.354881e-02	1.163313e-02	1.274337e-04	0.0517513253
69	3.023229e-01	1.306881e-03	4.672427e-03	4.915599e-01	0.0028018455
70	2.244028e-01	6.154244e-02	1.655560e-02	4.381580e-02	0.1012963446
71	2.646289e-01	1.264420e-01	1.514681e-01	2.290721e-01	0.1120007756
72	5.855841e-03	7.835483e-02	2.228096e-02	4.521026e-02	0.3587387471

Figure 26: Cos2 for rows 36-72

73	1.250982e-02	6.199249e-02	3.670994e-02	6.809722e-02	0.4091872366
74	1.346423e-02	1.407908e-01	4.108168e-01	1.164389e-01	0.0418972899
75	7.047290e-01	2.320841e-03	5.233543e-02	1.574237e-01	0.0171436644
76	6.646105e-02	2.172506e-02	8.936574e-02	4.590736e-01	0.0637931758
77	6.646105e-02	2.172506e-02	8.936574e-02	4.590736e-01	0.0637931758
78	6.775935e-01	5.521863e-02	1.452421e-03	4.095685e-02	0.0847104853
79	5.196275e-01	6.553163e-02	1.281096e-01	5.140701e-02	0.0233248182
80	4.516780e-01	3.448708e-02	6.862313e-02	1.364889e-01	0.1836396368
81	3.637841e-01	3.384871e-01	6.343635e-04	2.319324e-02	0.0344853298
82	3.820748e-01	5.684783e-02	1.848787e-01	3.366686e-03	0.1460249028
83	1.456138e-01	1.362786e-03	5.618960e-03	1.083813e-01	0.2747098447
84	1.312866e-01	7.052906e-02	2.312295e-01	1.994108e-01	0.1958709881
85	8.281024e-03	6.564039e-04	1.559533e-01	6.773740e-01	0.0106641085
86	6.866713e-01	2.265171e-02	3.259018e-02	5.065729e-02	0.0311494635
87	5.078827e-01	7.436054e-04	1.777423e-01	1.472935e-02	0.0048865513
88	1.273815e-01	4.668741e-01	1.206279e-01	1.801082e-01	0.0492237228
89	8.504557e-01	3.249697e-02	2.698173e-02	1.219445e-02	0.0390885545
90	2.096778e-01	1.077756e-01	2.637068e-01	9.918458e-02	0.0234899324
91	1.793247e-01	2.952694e-01	6.350867e-02	4.846399e-05	0.4000548871
92	5.078827e-01	7.436054e-04	1.777423e-01	1.472935e-02	0.0048865513
93	5.123196e-01	5.468563e-05	1.660471e-01	1.637521e-02	0.0062751745
94	1.273815e-01	4.668741e-01	1.206279e-01	1.801082e-01	0.0492237228
95	4.081720e-01	2.428442e-01	1.675273e-01	1.410618e-02	0.0748082366
96	4.057259e-01	1.574534e-01	2.368787e-01	1.739100e-01	0.0128617873
97	6.753719e-01	1.797296e-02	2.159879e-03	4.606490e-02	0.1941092742
98	8.006854e-02	5.799194e-01	3.323467e-02	6.826982e-04	0.0169100389
99	4.143608e-01	2.881009e-02	8.177529e-02	2.586875e-01	0.0375755474
100	1.017797e-02	3.483734e-01	3.501004e-01	9.177999e-02	0.1869108207

Figure 27: Cos2 for rows 73-100

In these images we reported the cos^2 values for each individual. To study the PCA a bit more in depth, we can plot the cos^2 values and see their distribution.

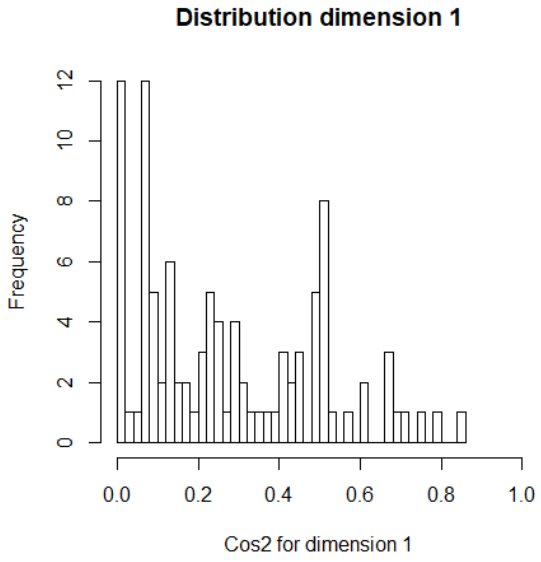

Figure 28: Cos2 distribution for dimension 1

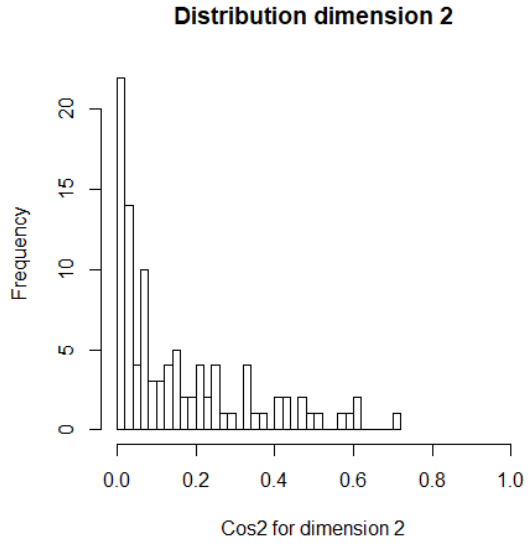

Figure 29: Cos2 distribution for dimension 2

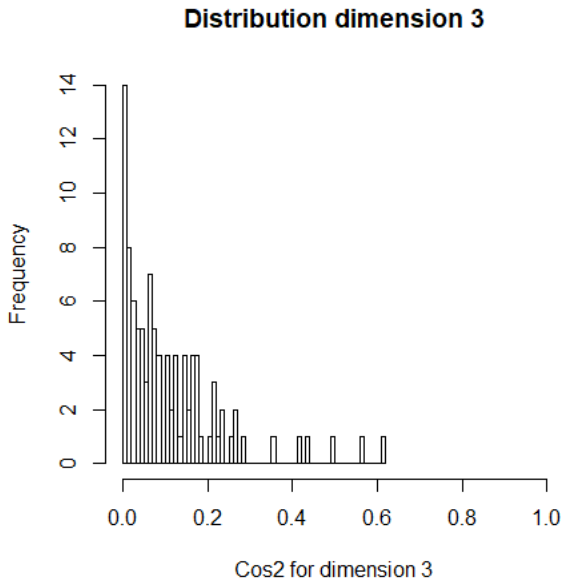

Figure 30: Cos2 distribution for dimension 3

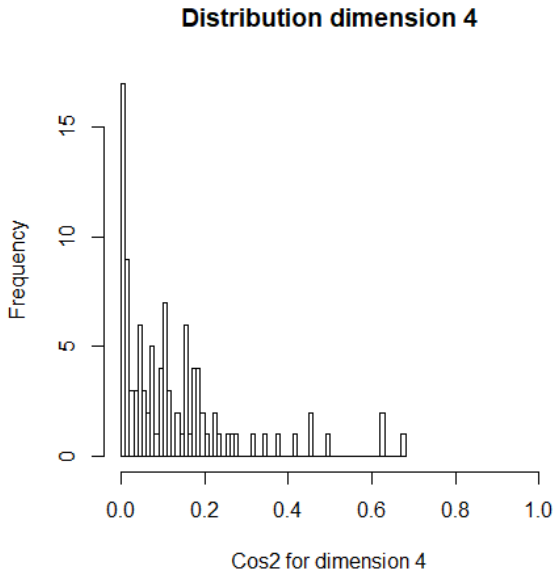

Figure 31: Cos2 distribution for dimension 4

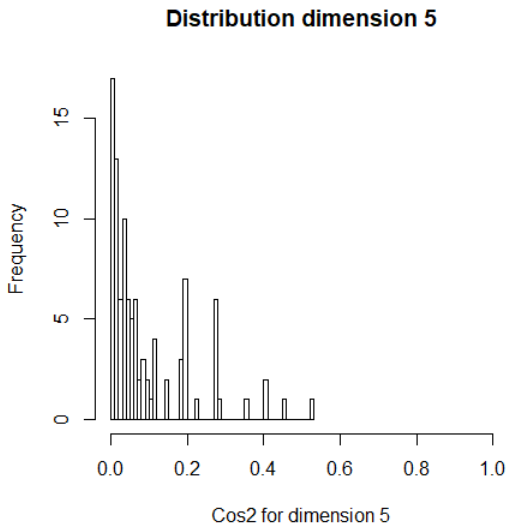

Figure 32: Cos2 distribution for dimension 5

It's clear from the plots how the most important dimension is the first one, with values mostly of e^{-1} order. The others tend to be closer to 0, but still present numbers of similar magnitude.

From this rapid look, we can be satisfied by the results, since it implies that the dimensions (or at least the first one) explain a significant part of the observations.

5. QUESTION 2D)

If there are categorical variables, paint the individuals with different colors according to the categories. Draw the confidence ellipses and interpret them, [4], p. 36.

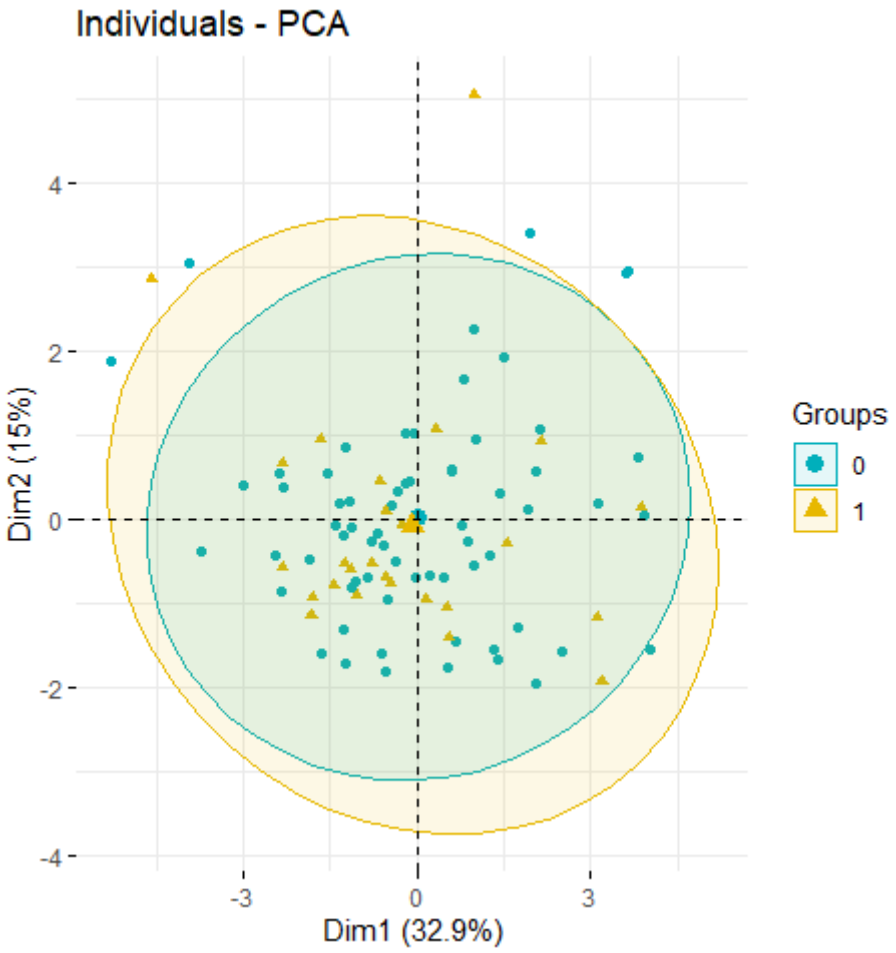

Figure 33: Scatterplot of the rows wrt the dimensions, divided into two groups

From this final graph we can infer that the information provided to us are not enough to gain a significant insight into the quality of the wine. As we can see, the two confidence ellipses basically overlap on each other, meaning that the quality of the wine doesn't really depend on the variables we have in the dataset.

6. QUESTION 3A)

Using the graphical output of pca command, discuss correlation between the variables including presence of groups of variables that are closely related.

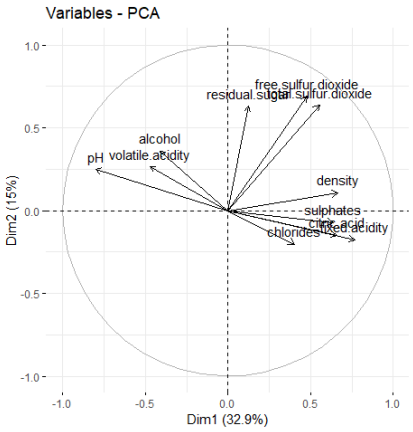

Figure 34: Correlation between variables and dimensions

We can see how the variables seem to be divided into three main groups pointing in different directions. Alcohol, pH and volatile acidity have an opposite direction with respect to density, sulphates, citric acid, fixed acidity and chlorides, which indicates a negative correlation.

We can also observe how some characteristics almost overlap with each other, like free sulfur dioxide and total sulfur dioxide or citric acid and fixed acidity. The meanings of these strict correlations are easy to interpret (obviously a wine with greater quantity of free sulfur dioxide will have a greater total sulfur dioxide).

Other correlations can be probably explained by a greater knowledge of wine chemistry (for example, more alcohol seems to produce higher pH).

7. QUESTION 3B)

Discuss the quality of the PCA representation: provide cos2 and the contributions for variables.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
fixed.acidity	0.58974749	0.030821196	0.125846126	6.052154e-04	0.0023298848
volatile.acidity	0.22316153	0.070465957	0.002800058	4.909713e-01	0.0068620729
citric.acid	0.43368789	0.021974011	0.008849270	3.538145e-01	0.0016424109
residual.sugar	0.01537742	0.398672155	0.236057558	5.055850e-07	0.2216097201
chlorides	0.16101416	0.041357612	0.192575576	9.093487e-02	0.2190551865
free.sulfur.dioxide	0.22991679	0.476361797	0.028464522	6.029246e-03	0.1104240529
total.sulfur.dioxide	0.30602799	0.406026667	0.129551386	1.458888e-03	0.0656451787
density	0.44083924	0.011598590	0.372694798	1.190571e-02	0.0375217291
pH	0.64097034	0.061915552	0.008604513	9.483438e-02	0.0004975305
sulphates	0.40667438	0.005226612	0.217941019	1.882511e-02	0.0774473324
alcohol	0.17079539	0.129085468	0.094453556	2.293871e-01	0.1978406331

Figure 35: Variability explained by dimensions

From these results, the PCA dimensions seem to explain a good portion of the variables' observations.

We can also notice how the PCs contribute to some of the variables much more than to others (for example dimension 1 has a 0.6 value for fixed acidity, but only 0.02 for residual sugar).

8. QUESTION 3C)

Plot the correlations between variables using pairs function. Compare the result with that of 3a.

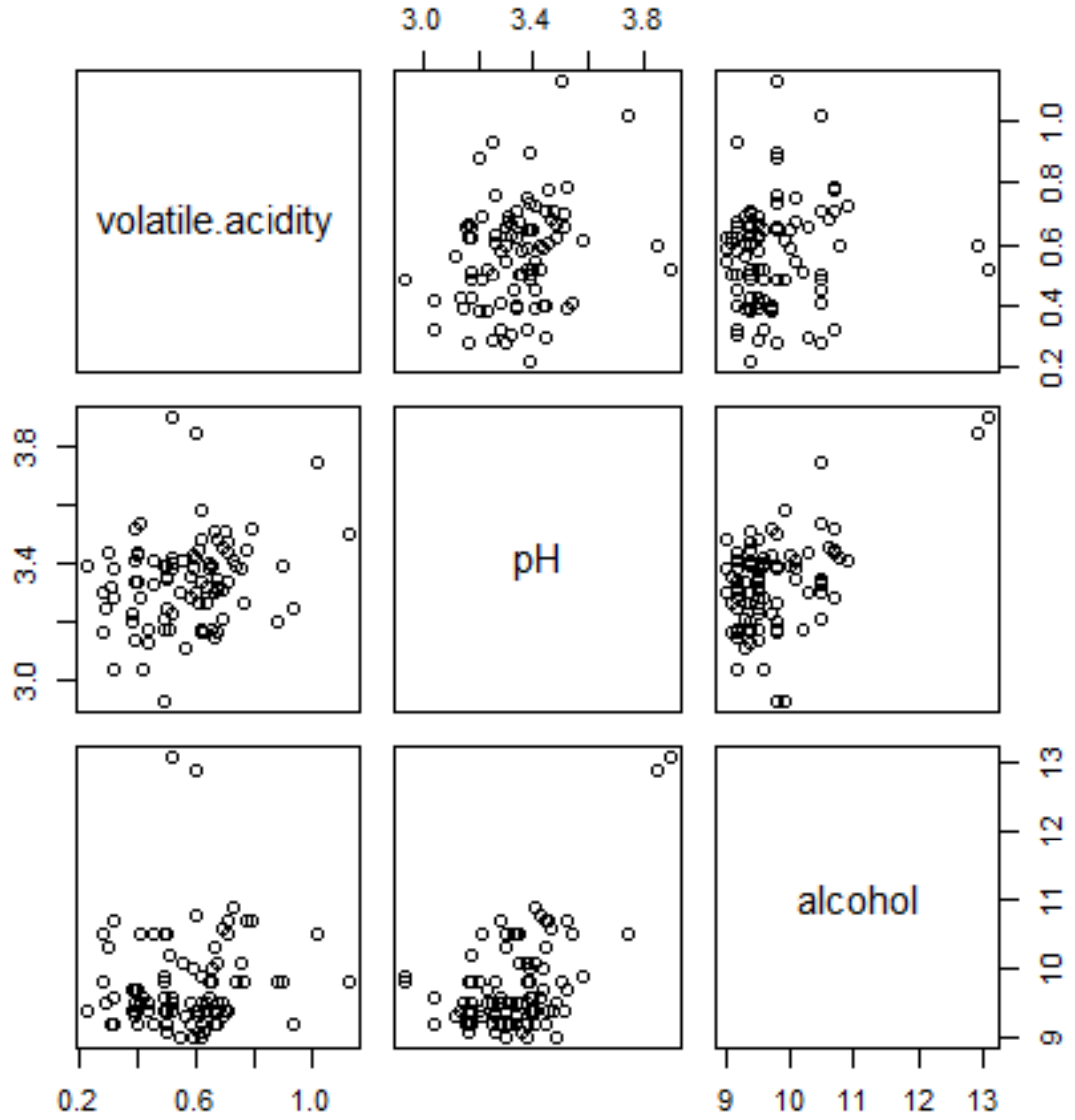

Figure 36: Correlation between some of the variables

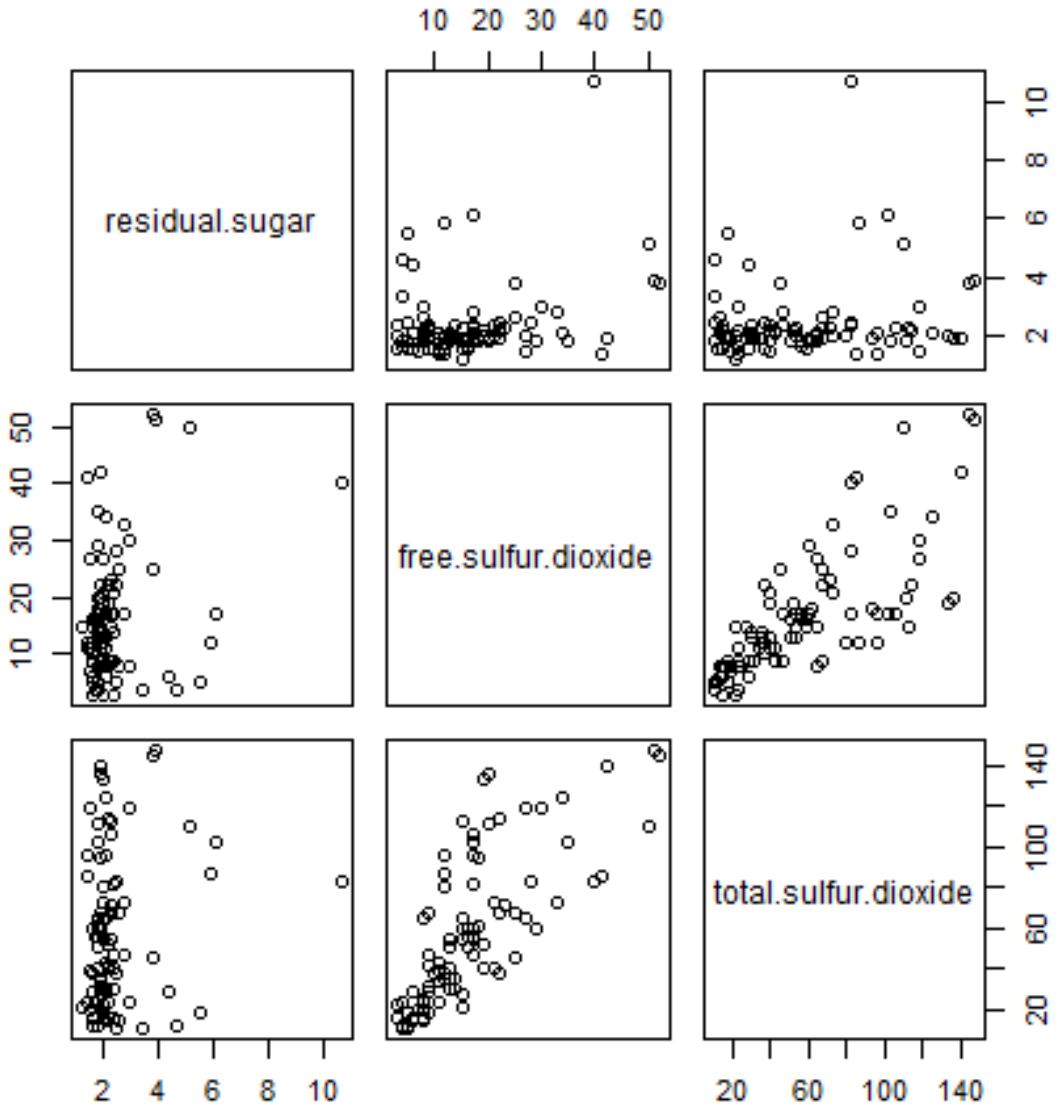

Figure 37: Correlation between other variables

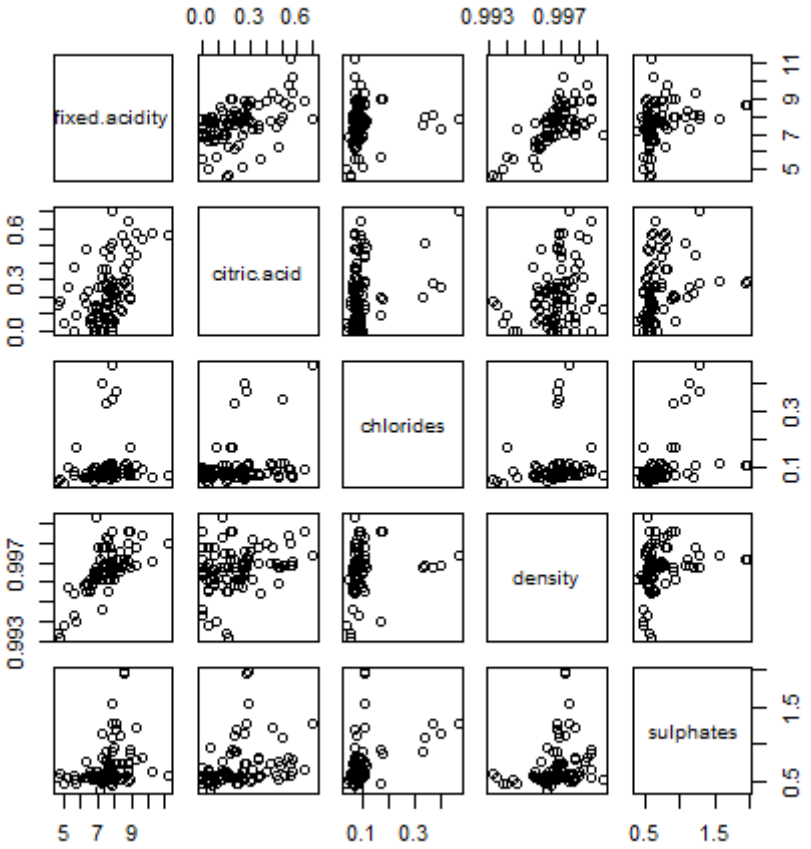

Figure 38: Correlation between more variables

To make the matrix plot readable, I decided to split the variables into the three main groups we observed in question 3a).

From the resulting plots it's sometimes easier to notice the correlations and sometimes more complicated. For example, correlation between free sulfure dioxide and total sulfure dioxide is quite evident, as it is the one between density and fixed acidity. Less clear and noisier appear the relationships between alcohol and pH or between residual sugar and the two dioxides, which instead looked rather strong in the previous graph.

4TH EXERCISE

For this exercise I will be using a dataset containing information about Peron's election results divided by cities and countryside (http://users.stat.ufl.edu/~winner/data/peron.txt). Dataset contains 6 rows and 4 columns: Percentage obtained (1=>70, 2=60-70, 3=50-60, 4=40-50, 5=30-40, 6=<30) and the three different regions (Big city, Township and countryside). The value in the cell is the number of counties in the region that obtained the percentage relative to that row.

1. QUESTION 2A)

Do the X2 test for independence and interpret it, see Section 2.2.2 of [4].

Applying the chi-squared test to the first two columns of the dataset, we obtain the following results:

***X-squared**: 45.107*

***P-value**: 7.366e-05*

Given the small p-value, we can assume that a statistically significant relation exists between rows and columns variables.

2. QUESTION 2B)

Perform the CA, get the 2D representation of row and column profiles

- *Separately*
- *in the same graph*

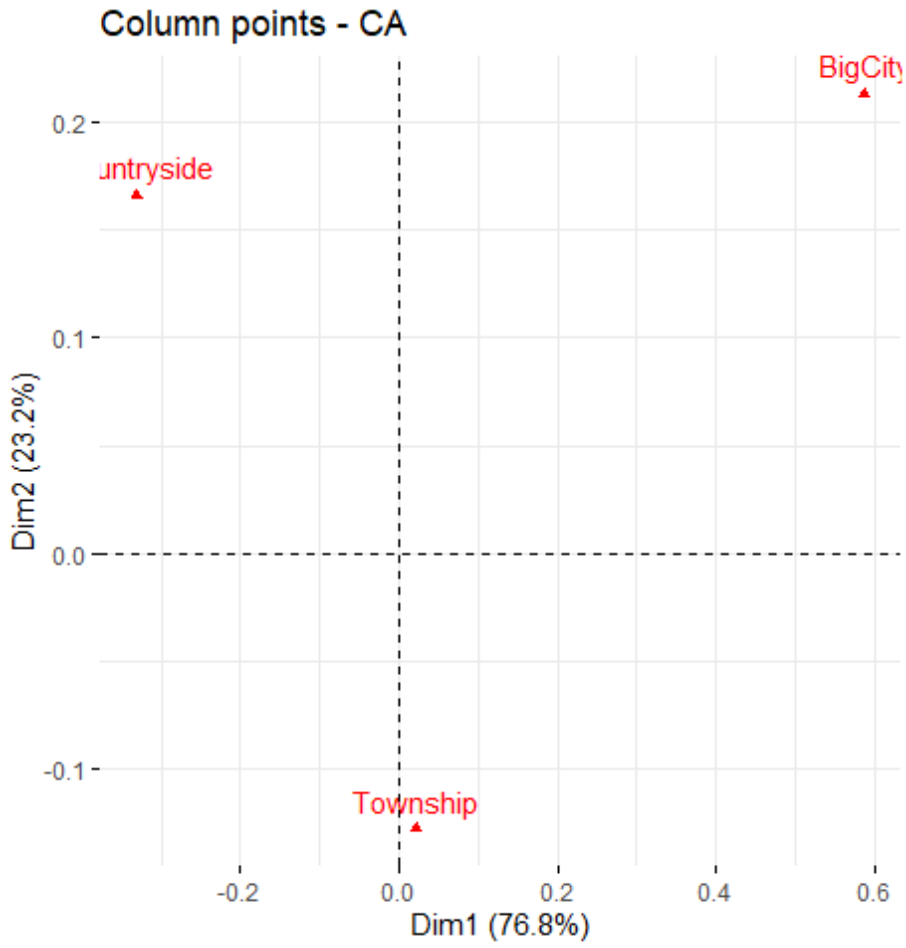

Figure 39: Only column profile

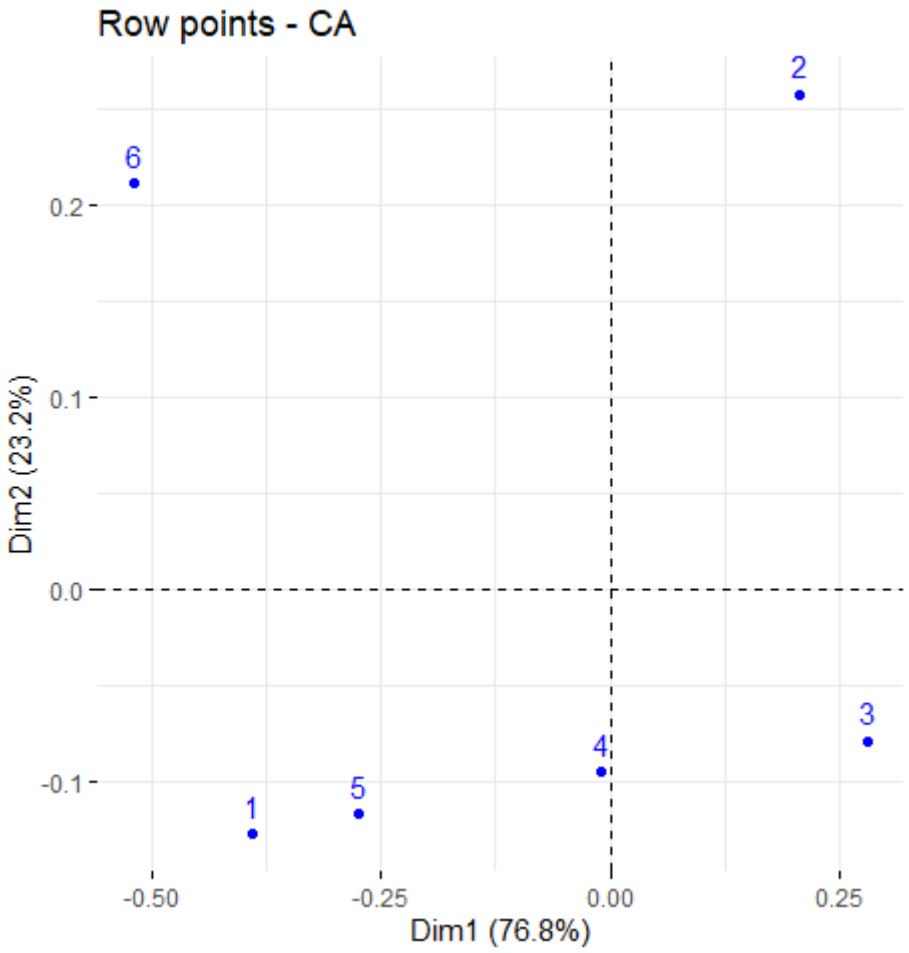

Figure 40: Only row profile

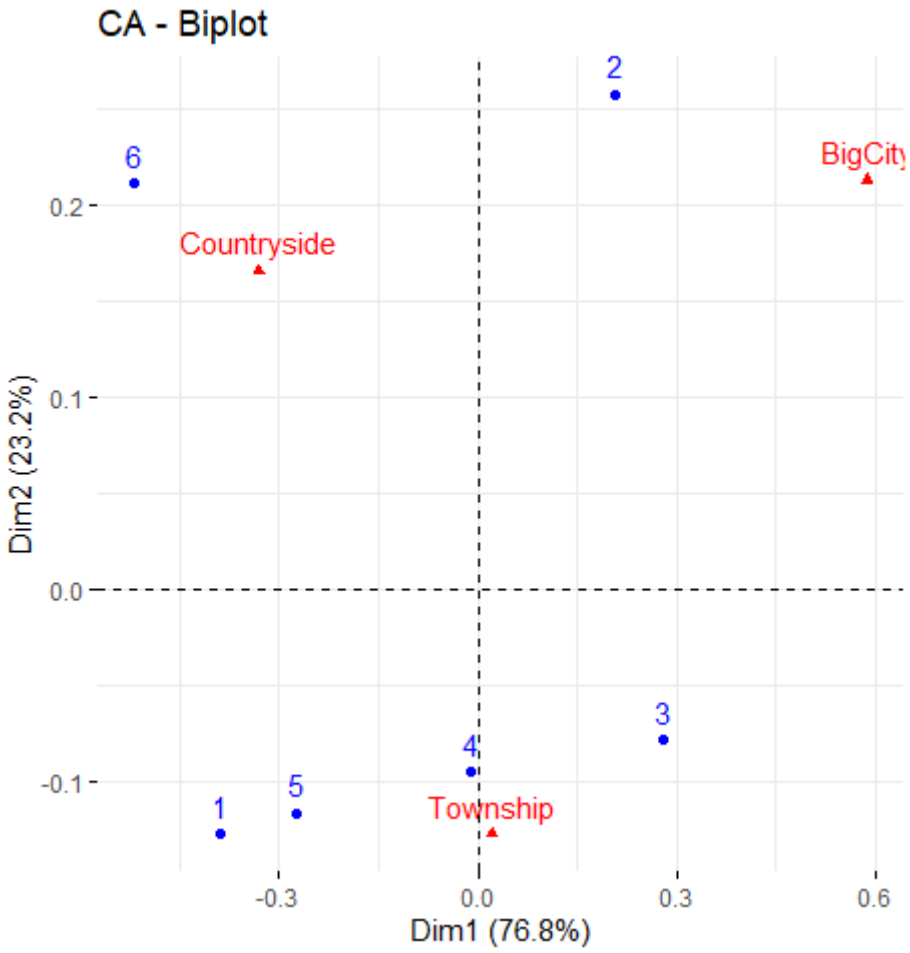

Figure 41: Joint plot

3. QUESTION 2C)

Analyze the patterns obtained in item 2b. Focus on the total variability, similarities/dissimilarities and the conclusions that can be made from the simultaneous representation of rows and columns.

We can observe how places with 1,3,4,5 percentages tend to behave similar, while 2 and 6 percentages are outliers.

It's also interesting to notice that the first group of rows is associated with township, while countryside and big cities are associated respectively to 6 and 2.

We can infer that Peron got less votes in the countryside and more in cities, while got mixed results in townships.

4. QUESTION 2D)

Provide the table and graph of eigenvalues, justify the choice of principal components.

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.07739142	76.77266	76.77266
Dim.2	0.02341455	23.22734	100.00000

Figure 42: EIgenvalue table

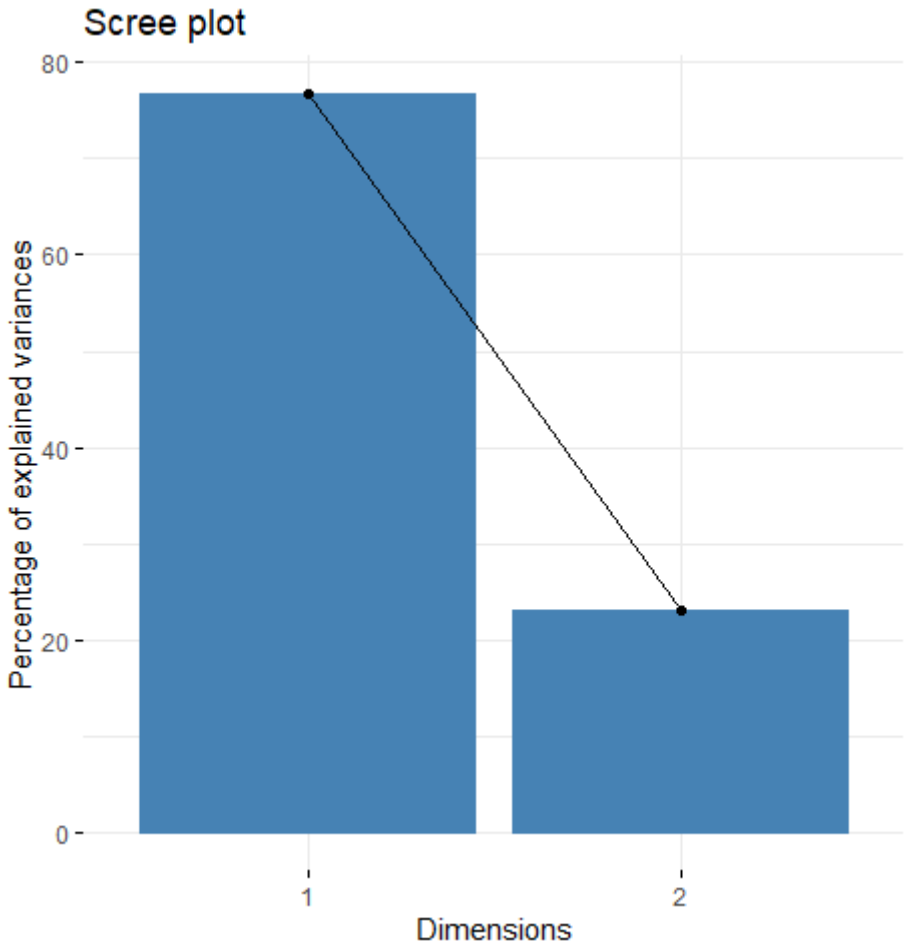

Figure 43: Graphical representation of eigenvalues

From the eigenvalues we can see how the first dimension is sufficient to explain most of the variability, almost 80%.

This exceptional result can be explained by the small dimensions of the dataset, both in terms of rows and columns.

5. QUESTION 2E)

Discuss the quality of the CA representation based on cos^2 for rows and columns.

	Dim 1	Dim 2
BigCity	0.88374558	0.1162544
Township	0.02784692	0.9721531
Countryside	0.79853555	0.2014645

Figure 44: Cos2 for columns

	Dim 1	Dim 2
1	0.90288177	0.09711823
2	0.39039401	0.60960599
3	0.92678671	0.07321329
4	0.01344103	0.98655897
5	0.84541416	0.15458584
6	0.85709478	0.14290522

Figure 45: Cos2 for rows

As we can guess from these images, the CA representation is very good at explaining the variability for this dataset.

The cos^2 values for the first dimension in the columns are close to 1 for two of the three columns and most of the values are around 0.9 regarding the rows.

This means, as stated before, that the first dimension alone can explain the great majority of the variability in the dataset.

Again, we need to remind that such good results depend mostly on the small scale of the dataset.

5TH EXERCISE

1. QUESTION 1A)

Plot a two-dimensional MDS configuration representing the cities. Compare the result with the actual geographical location of the cities across the country.

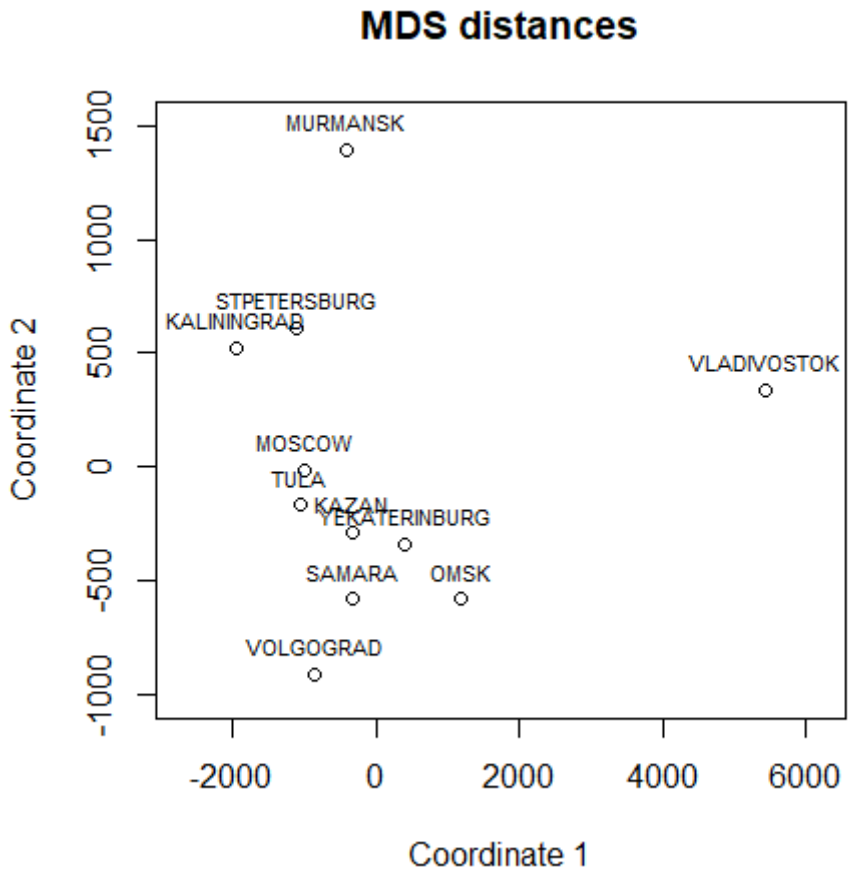

Figure 46: MDS scaled map

The map looks much more compact than the real one (for example, Kazan is more to the east with respect to Moscow, as is Yekaterinburg), but in general cities positions are respected.

2. QUESTION 1B)

Based on the computed eigenvalues, discuss the quality of representation in the 2D space.

```
> cumsum(abs(eig)) / sum(abs(eig))
 [1] 0.8952647 0.9960954 0.9970208 0.9970437 0.9970551 0.9970551 0.9970610 0.9970685 0.9970902
[10] 0.9974427 1.0000000
> cumsum(eig^2) / sum(eig^2)
 [1] 0.9874650 0.9999907 0.9999918 0.9999918 0.9999918 0.9999918 0.9999918 0.9999918 0.9999918
[10] 0.9999919 1.0000000
```

Figure 47: Cumulative sums of the eigenvalues

These values are very high and confirm what we said in the previous question, that is, the 2D representation is a good representation of the real-world distances.

3. QUESTION 1C)

Plot the Shepard diagram and discuss it.

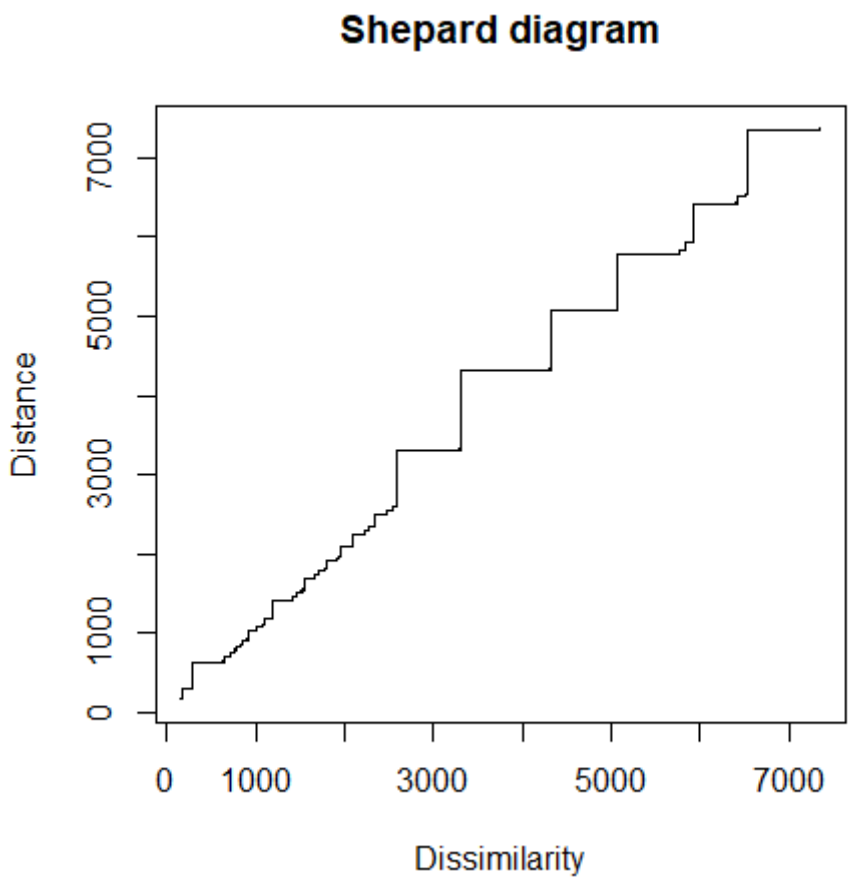

Figure 48: Shepard diagram

The Shepard diagram shows some discrepancies in the upper right part of the line. This probably depends on the huge distances between the more eastern cities and the others. Clearly an MDS scaling cannot show them properly on a small scale 2D graph, therefore causing some dissimilarities.

4. QUESTION 1D)

Check whether the MDS configuration you obtained does restore the original distances in a sufficiently high dimensional space.

***5-dimension max difference**: 190*

***8-dimension max difference**: 99*

***10-dimension max difference**: 3e-11*

We can see that using all 10 dimensions, distances are correctly restored. The same can't be said for 8- or 5-dimension space, for which the maximum difference with the real distance are respectively 99 and 190 (we have to consider however that these values are still low in comparison with the distances we used).

5. QUESTION 2)

Get the airline distances between 10-12 cities/towns worldwide or, alternatively, a rank-order dissimilarity matrix such as Tables 4.4 or 4.5 in [2], e.g., http://prtools.org/disdatasets/.

- *Do the same analysis as for task 1.*
- *Pay special attention to the negative eigenvalues, if any. In that regard, discuss which dimensions you choose to represent the MDS solution and why.*

I got the distances between 10 cities worldwide: New York, Ottawa, Seattle, Casablanca, London, Rome, Instanbul, Moscow, Tokyo and Queenstown (in South Africa)

6. QUESTION 2A)

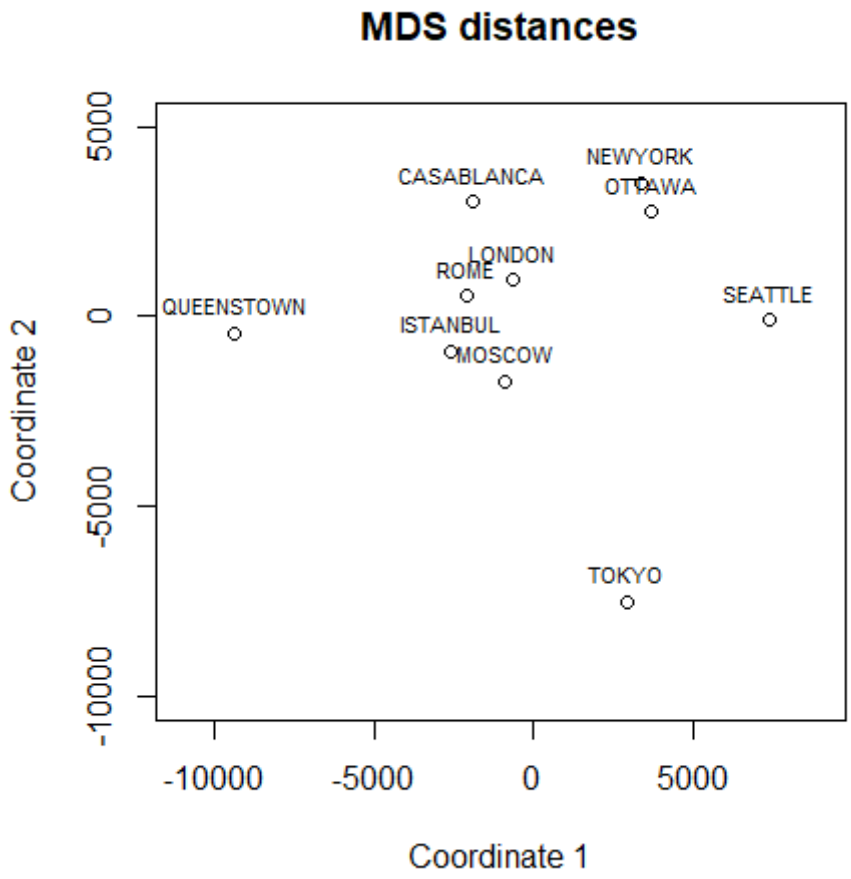

Figure 49: MDS scaled map

This time the scaled distances are much more different from the real ones, but, provided that the graph is rotated, the proportions are still more or less maintained, with the exception of New York, Ottawa and Seattle. The reason these three cities are wrongly positioned depends on the fact that we are translating a 3D world into a 2D graph. When distances become too large, the map will start wrapping on itself.

7. QUESTION 2B)

```
> world_eig
 [1] 191050118.8  90475050.3  26589661.4   2917399.4    447146.6   -106484.1   -163711.8   -729206.4
 [9]  -7140677.0 -14774997.1
> #Let's evaluate them
> cumsum(abs(world_eig)) / sum(abs(world_eig))
 [1] 0.5713316 0.8418955 0.9214113 0.9301357 0.9314729 0.9317914 0.9322809 0.9344616 0.9558157
[10] 1.0000000
> cumsum(world_eig^2) / sum(world_eig^2)
 [1] 0.7991894 0.9784202 0.9939005 0.9940869 0.9940913 0.9940915 0.9940921 0.9941038 0.9952202
[10] 1.0000000
```

Figure 50: Eigenvalues and cumulative sums

First of all, it's worth noticing how the eigenvalues highest in absolute value are the ones related to the three American cities, which are in fact the ones worst positioned.

Moreover, it's important to see that this time the cumulative sum is far worse than before, with a mere 0.57 as first value. Again, this reflects the problems of translating a 3D world in a 2D map.

Observing these values, we can conclude that a 3- or 4- dimensional space would fit well for this dataset, while the 2-dimensional one that we used will not be good enough.

8. QUESTION 2C)

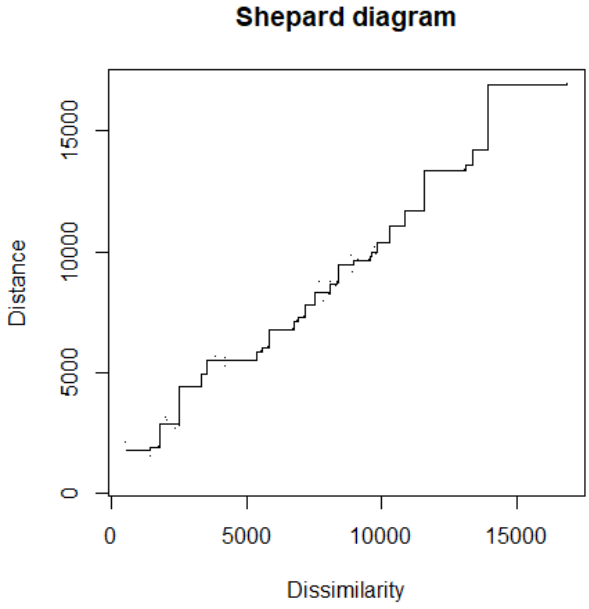

Figure 51: Shepard diagram

What we've just seen is confirmed by the Shepard diagram. This time the dissimilarities are much more emphasized, reflecting the problems related to the MDS scaling.

9. QUESTION 2D)

***5-dimension max difference**: 610*

***8-dimension max difference**: 190*

***9-dimension max difference**: 3e-11*

With respect to the first question, we get this time much worse results. In a 9-dimension space (maximum possible), distances are correctly restored, but using lower dimensions, errors become quite large.

6TH EXERCISE

1. INTRODUCTION TO EXERCISE 6

The dataset used in this question contains various measures regarding climate in Australia, particularly focusing on the rain, recovered from Kaggle (https://www.kaggle.com/jsphyg/weather-dataset-rattle-package). The dataset was built for machine learning purposes, so a couple of empty columns had to be removed. The number of rows had to be decreased as well, to avoid an excessive amount of computation. Final dataset has 18 columns and 2000 rows.

2. QUESTION 1)

Do the hierarchical clustering (preceded by the PCA) using command HCPC from FactoMineR package:

(a) Clearly name the recommended (by HCPC) clusters.

(b) Explain the meaning of the barplot in the upper-right corner of the output.

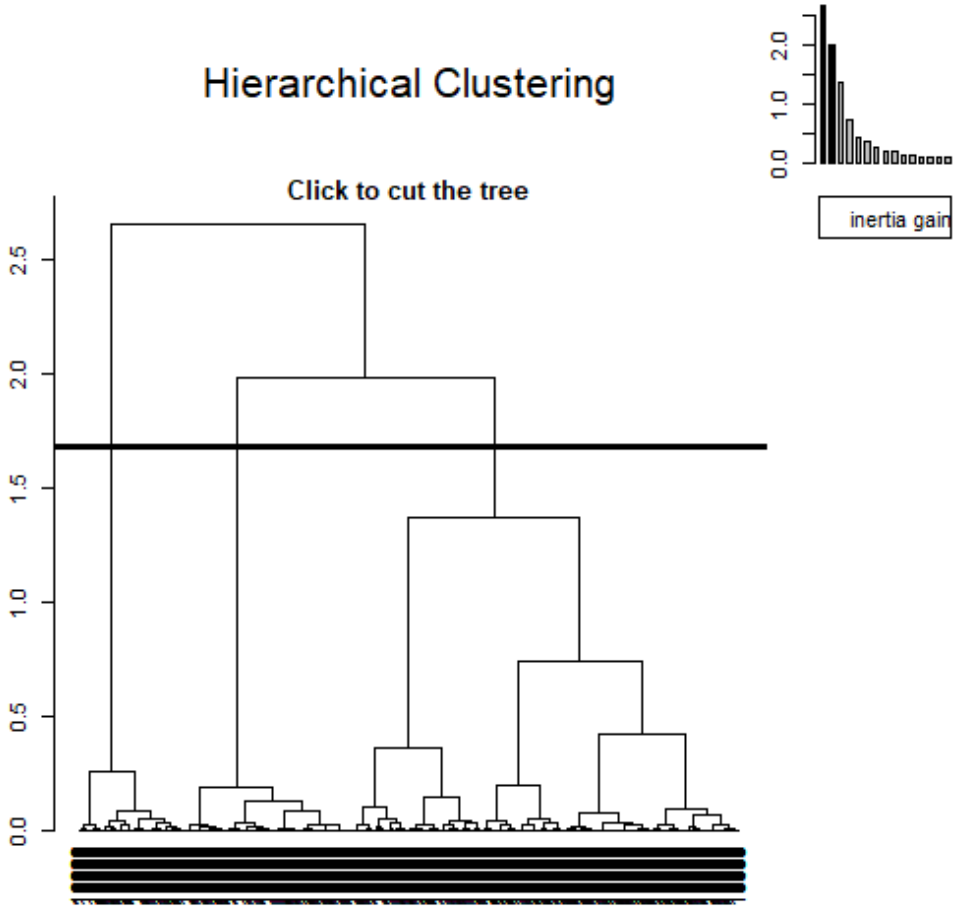

Figure 52: HCPC hierarchical clustering

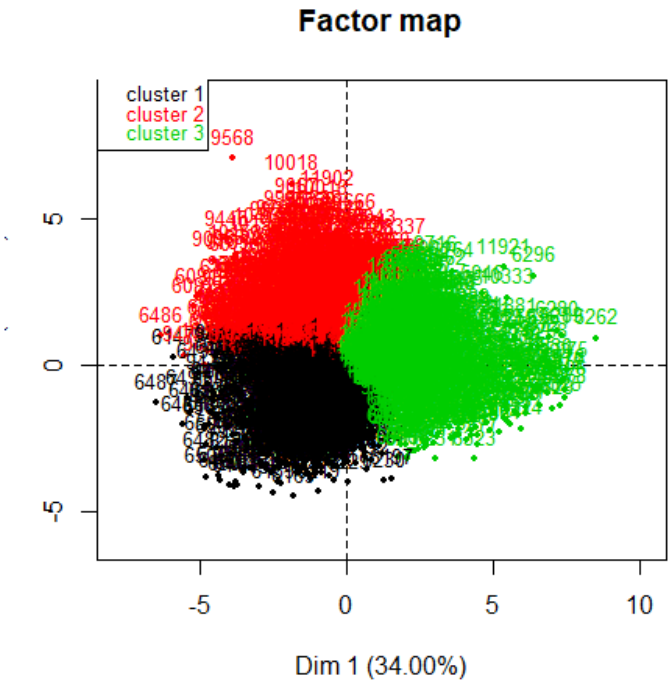

Figure 53: 3-clusters division of the data

The function hcpc suggests cutting on the third level. Observing the scatterplot at this level, it's visually clear how the clustering was performed.

This function allows us to observe the information characterizing the three clusters. I reported the most significative here:

Parameter (mean)	1st cluster	2nd cluster	3rd cluster
MaxTemp	20.78	23.5	30.9
Rainfall	0.74	7.84	0.48
WindGustSpeed	32.84	40.04	43.36
Humidity3pm	46.9	71.26	40.53
Sunshine	9.16	4.21	10.65

From these simple data, the three clusters seem to be referring to three different climate areas. Trying to make guesses, the last cluster refers to a desertic area, the second one to a seashore and the first to a colder inland area.

The barplot in the upper right corner represents the loss in *between-inertia* when going from *n* clusters to *n-1*. By *between-inertia* we indicate how "close" clusters are, (i.e. the higher the *between-inertia*, the more the clusters are separated). Clearly the first bar is the highest, as one cluster can't be separated from itself, while it goes down as soon as n increases.

This plot is useful to see if it's worth dividing the data into a greater number of clusters. If the loss in *between-inertia* is high, then the new clusters we create won't overlap, otherwise they might end up being too close

to each other. In general, it's useless to have more clusters if they're not really separated from each other, while it makes sense to split a cluster that's sparsely populated.

3. QUESTION 2)

Perform the K-means clustering, choosing K according to the results of hierarchical clustering.

(a) Plot the results

(b) Compare distribution of points over clusters with that of hierarchical approach

In the following graphs we're going to take the mean of some of the parameters, grouping them through the K-mean clustering (K = 3).

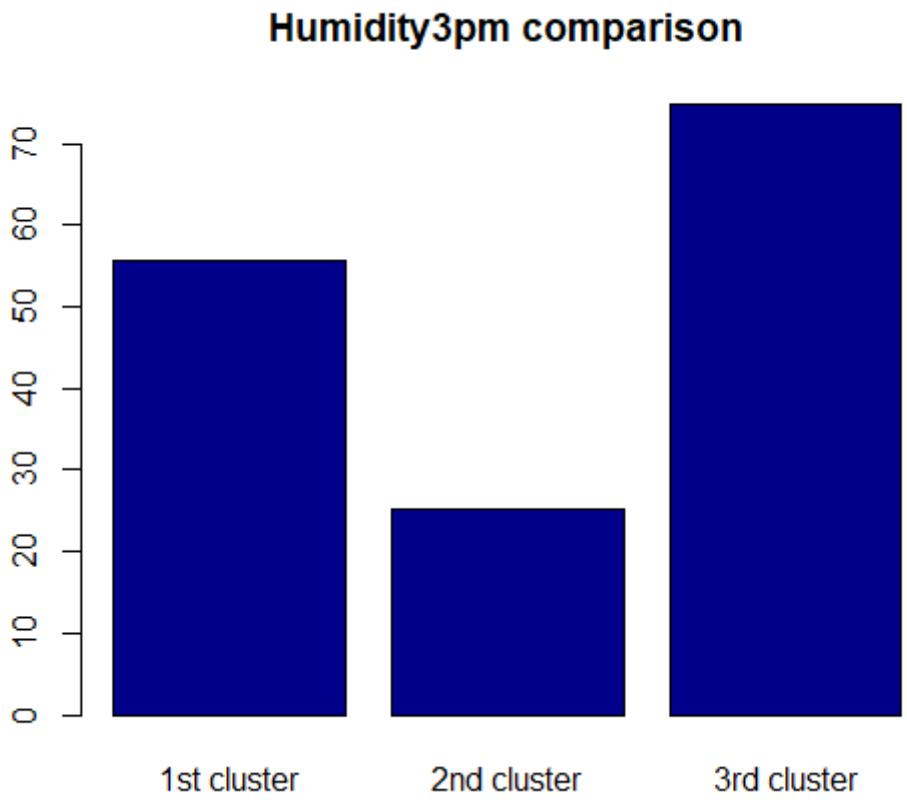

Figure 54: Mean of humidity at 3pm

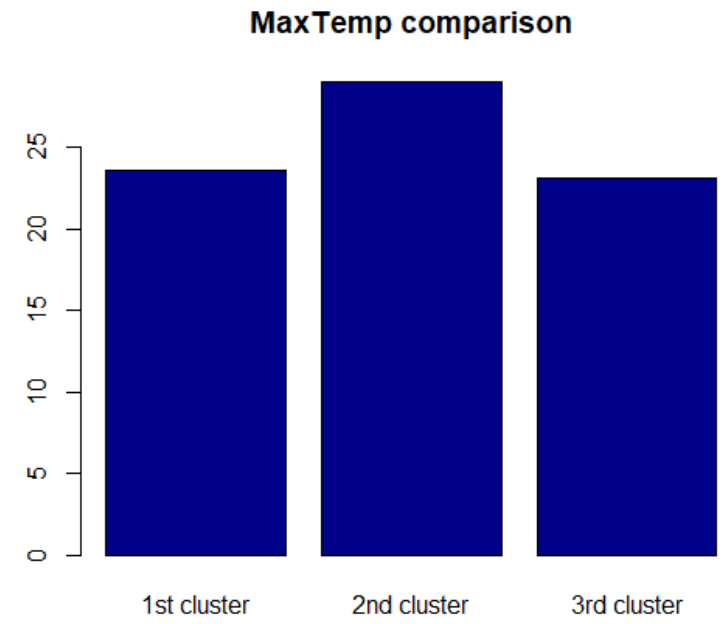

Figure 55: Mean of the max temperature

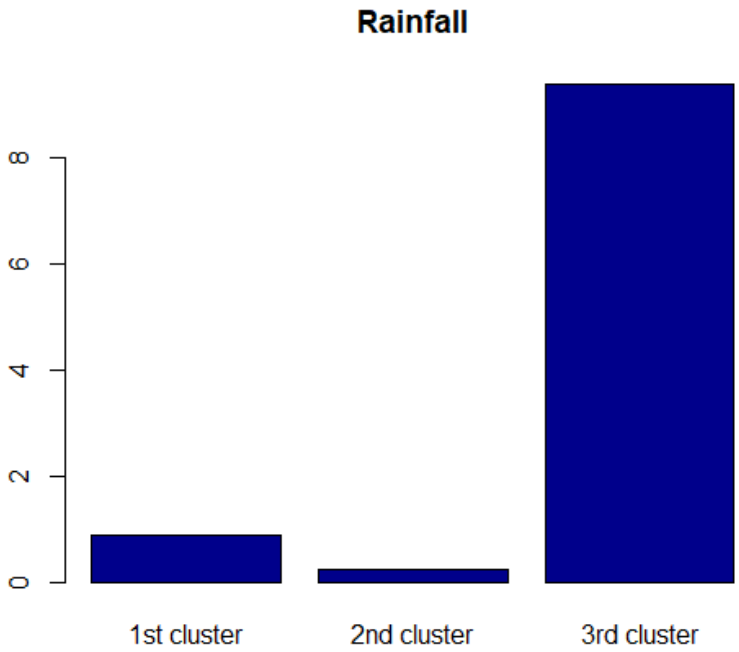

Figure 56: Mean of rainfall

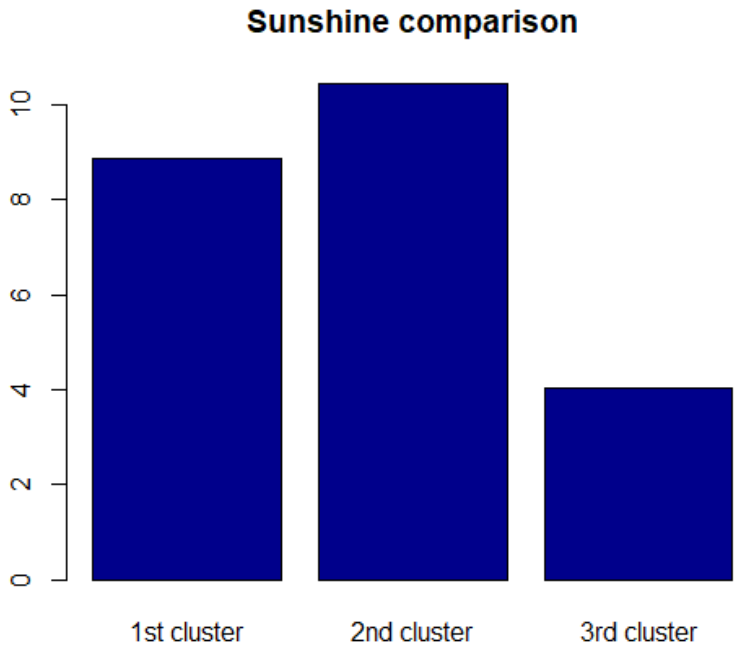

Figure 57: Mean of sunshine

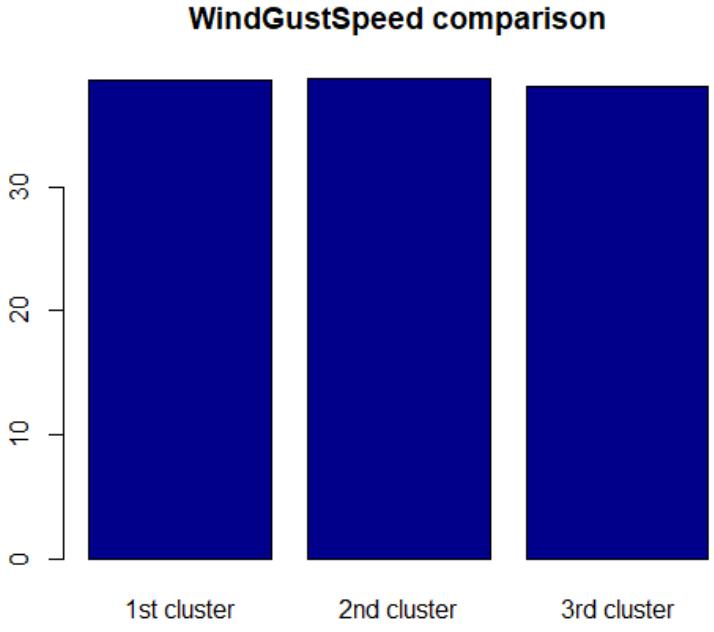

Figure 58: Mean of wind gust

These results reassure us about what we said earlier. We obtain a rainy and humid group, a hotter and dry one and a humid climate with little rain.

In conclusion, results are similar using both techniques.