1st HOME ASSIGNMENT

By Stefano Agresti

FIRST EXERCISE

1. INTRODUCTION

The dataset chosen for the first exercise measures the interest rates on the 10-Year Treasury bonds of the United States Government (link to data https://fred.stlouisfed.org/series/DGS10, source "FRED"). It's divided into two columns, one for the date and one for the interest rate value on that day, and it contains 15056 rows.

The measurement is performed daily from 1962-01-02 to 2019-09-17.

There are 643 rows with missing values. It's interesting to notice how most of the missing values relates to holidays (e.g. there's no data for the 25th of December of any year), which explains its absence.

2. QUESTION a)

Construct a stem-and-leaf and histogram. Impose the empirical density estimate on the histogram. Discuss the results focusing on the shape of the plots and number of modes.

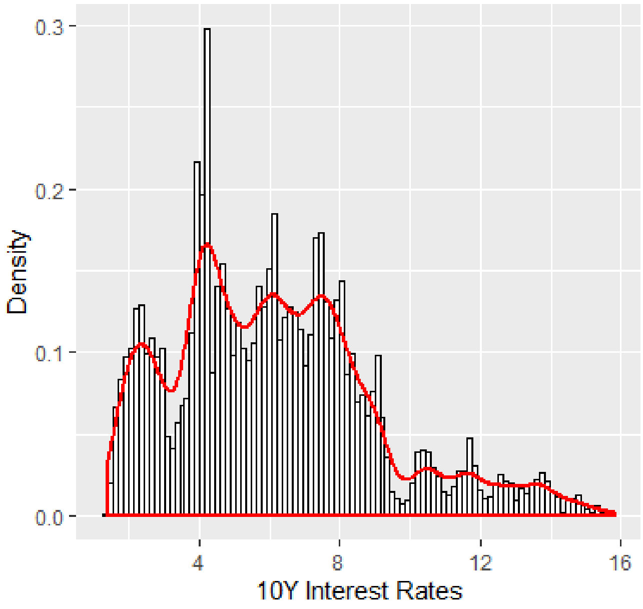
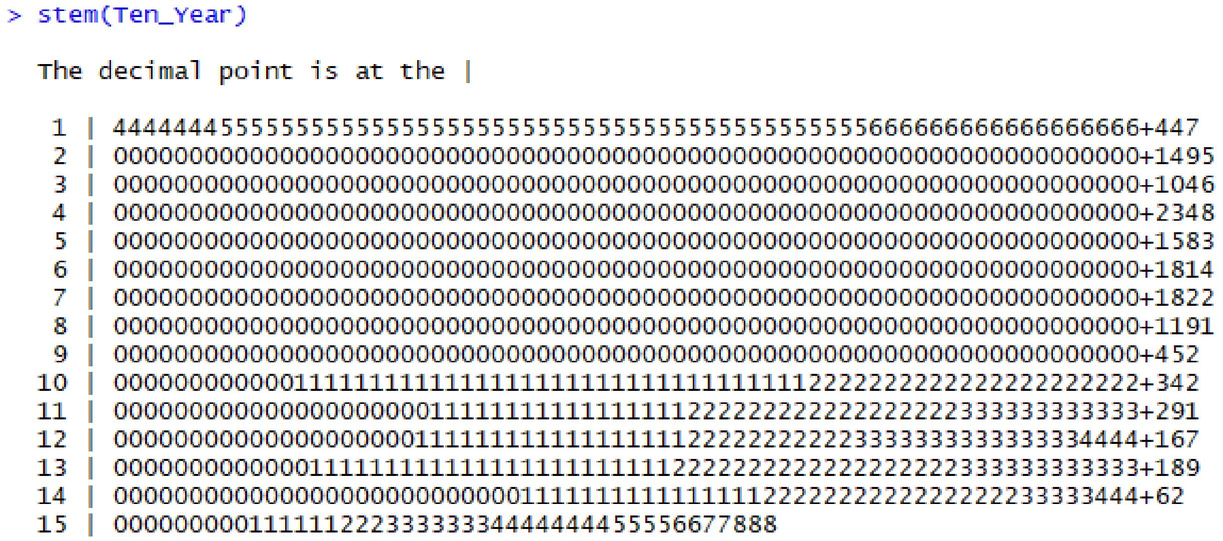


Figure 1: Histogram

Both stem-and-leaf and histogram clearly show that most of the data is found between 2% and 7-8%, with a particularly strong concentration at 4%. This is in line with the usual interest rate of a 10-Year Government bond.

We can also observe that there are peaks up to 15%, but in general (as easily observed on the stem-and-leaf) the number of times it got beyond 10% is small in comparison to the time period analyzed. We can speculate that these events occurred in moments of crisis or political tensions and are therefore concentrated in specific moments.

Values don’t seem to follow a normal gaussian distribution, but this can be explained by the fact that the analyzed time period covers over 50 years of US history, with all consequent fluctuations.



```
> stem(Ten_Year)

  The decimal point is at the |

   1 | 44444445555555555555555555555555555555555555555666666666666666666+447
   2 | 00000000000000000000000000000000000000000000000000000000000000000000+1495
   3 | 00000000000000000000000000000000000000000000000000000000000000000000+1046
   4 | 00000000000000000000000000000000000000000000000000000000000000000000+2348
   5 | 00000000000000000000000000000000000000000000000000000000000000000000+1583
   6 | 00000000000000000000000000000000000000000000000000000000000000000000+1814
   7 | 00000000000000000000000000000000000000000000000000000000000000000000+1822
   8 | 00000000000000000000000000000000000000000000000000000000000000000000+1191
   9 | 00000000000000000000000000000000000000000000000000000000000000000000+452
  10 | 00000000000011111111111111111111111111111111112222222222222222222222+342
  11 | 00000000000000000000011111111111111111222222222222222222333333333333+291
  12 | 00000000000000000000111111111111111111222222222223333333333333334444+167
  13 | 00000000000001111111111111111111111112222222222222222222333333333333+189
  14 | 00000000000000000000000000011111111111111112222222222222222233333444+62
  15 | 00000000011111122233333334444444455556677888
```

Figure 2: Stem-and-leaf

3. QUESTION b)

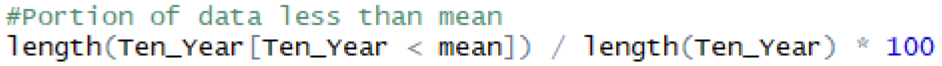
Compute the mean and median. Based solely on that, conclude whether the distribution is skewed. Find the proportion of the data which are less than the mean value

Value of mean: 6.138451

Value of median: 5.89

Portion of data below mean: 53.94436%

Based solely on this, we can say that the distribution is slightly right-skewed (median is less than mean).



```
#Portion of data less than mean
length(Ten_Year[Ten_Year < mean]) / length(Ten_Year) * 100
```

Figure 3: Code to compute portion of data below mean

4. QUESTION c)

Compute the 1st and 3rd quartiles, the 90th quantile and the mode. Explain the meaning of the obtained quantities. Find the value that cuts off the top 25% of the data.

1st quartile: 4.07

3rd quartile: 7.79

90th quantile: 10.16

Mode: 4.19

The 1st quartile is the value for which 25% of the values in the dataset is smaller and 75% is bigger.

The 3rd quartile is the value for which 75% of the values in the dataset is smaller and 25% larger (therefore, it's also the value that cuts off the top 25% of the values).

The 90th quantile is the value for which 90% of the values in the dataset is smaller and 10% larger.

The mode is the most frequent value found in the dataset.

5. QUESTION d)

Compute the range, the sample standard deviation and the IQR. Construct the boxplot of the data. Comment on the boxplot including skewness, outliers etc.

Range: 14.47

Sample Standard Deviation: 2.889889

IQR: 3.72

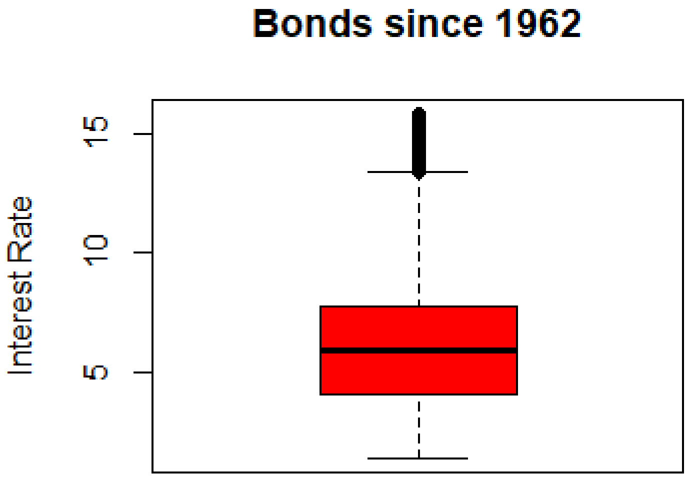


Figure 4: Boxplot

Observing the boxplot and knowing 1st and 3rd quartile, we can confirm the speculation we made observing the graphs in question a). Most values are found between 4.07% and 7.79% (approximately 50% of them) and only a quarter of them goes beyond 7%.

We can also confirm that the highest values reached by the interest rates are indeed outliers (as confirmed by the *boxplot* function, lowest outlier is 13.38%) and therefore exceptional cases.

The distribution appears right skewed – as guessed in question b) -, with 75% of the values concentrated below 7.79% and the remaining 25% going up until 15.8%.

Data is clearly not symmetrical, since the top 25% of values is extended on a much longer space than the lower 75%, which again tells us that interest rates tend to be compacted on moderately low figures.

6. QUESTION e)

Check whether the empirical distribution is normal by examining the QQ-plot.

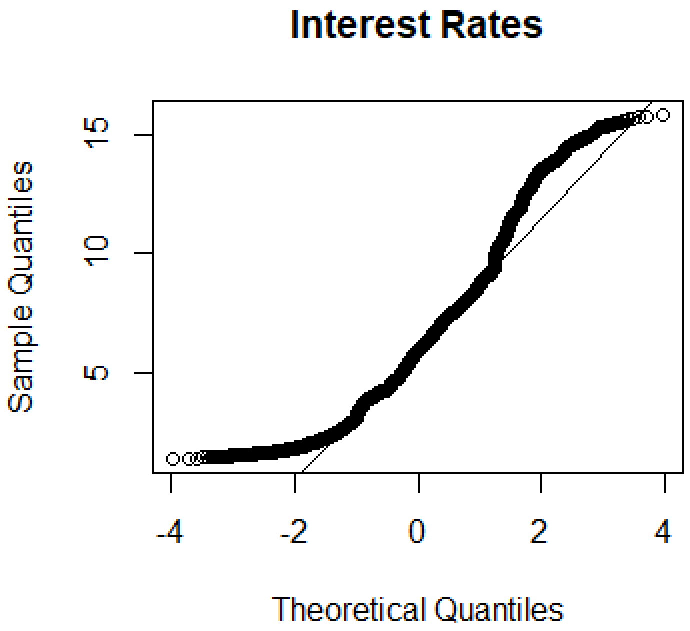


Figure 5: QQ-Plot

From the QQPlot we can notice how the data follows a normal distribution approximately from 2% to 10%, while outside of this range assumes a quite different shape.

Therefore, overall distribution is not normal.

SECOND EXERCISE

1. INTRODUCTION

For this exercise we're going to analyze the distribution of both 10-Year and 2-Year interest rates on US government bonds (link to 2-Year dataset https://fred.stlouisfed.org/series/DGS2). On top of that, we're going to analyze if a correlation between the two values exists.

Length of the second dataset is 11297, therefore we're going to cut approximately 4000 rows in the 10-Year dataset.

Number of missing values is equal for both datasets.

2. QUESTION a)

Create side-by-side boxplots. Compare the centers and spreads.

We can see from the side by side boxplot how the center is similar in both distribution (around 5%), while the 2-Year dataset has a much longer space between 1st quartile and the center. Also, its values spread on a wider range of values.

We can therefore assume that the 2-Year bonds have more and larger fluctuations around the center than the 10-Year ones. This probably reflects the greater volatility on short-term trust in the government.

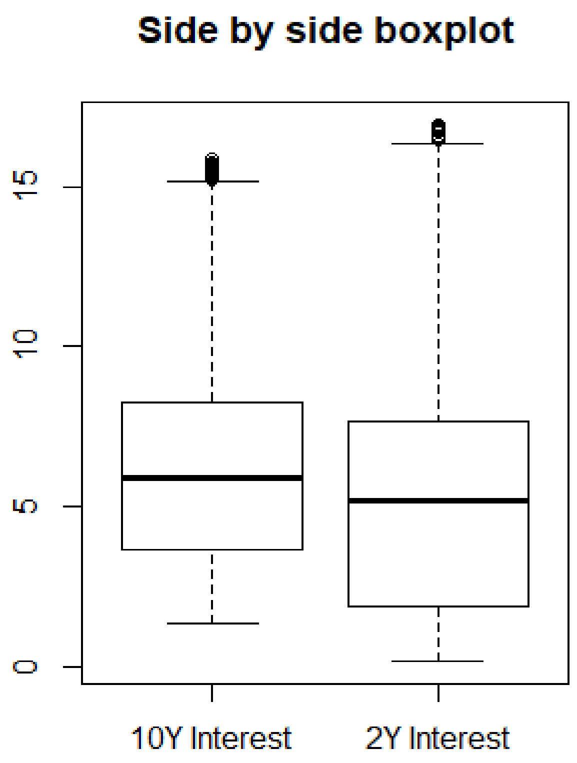


Figure 6: Side by side boxplot

3. QUESTION b)

Draw the scatter plot. Comment on the possible dependence and presence of outliers.

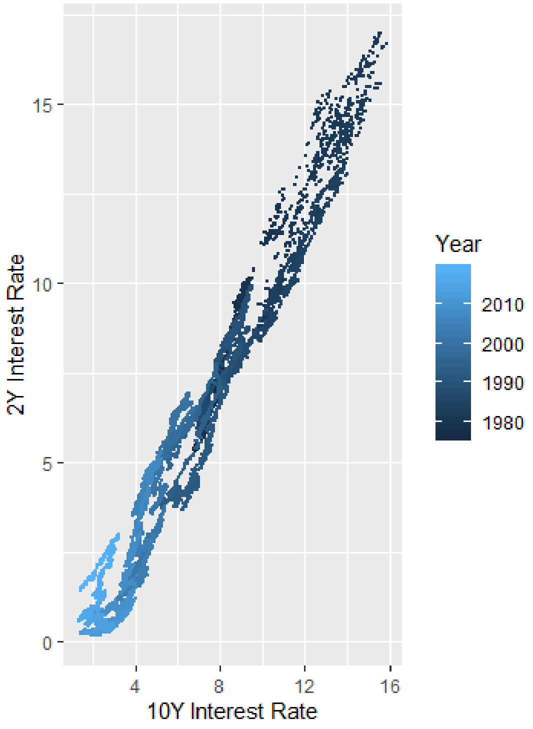


Figure 7: Scatter plot

From this plot it's evident how the two datasets are deeply related. There's basically no outlier and data seem to follow a linear distribution.

This could be foreseen as both types of bonds tend to be related to the same type of events (political and economic crisis mainly).

It's also interesting to see how the value of both interest rates is gone down during the years, with top values in the '80s and bottom ones in the 2010s.

4. QUESTION c)

Compute Pearson's and Spearman's coefficient of correlation. Interpret and compare their values. Are their values consistent with the scatter plot?

Pearson: 0.9766628

Spearman: 0.9720328

The values are almost identical and close to 1, which is consistent with the linear distribution we observed in point b).

5. QUESTION d)

Add the marginal distributions to the scatter plot. For that purpose, use histogram and box plot.

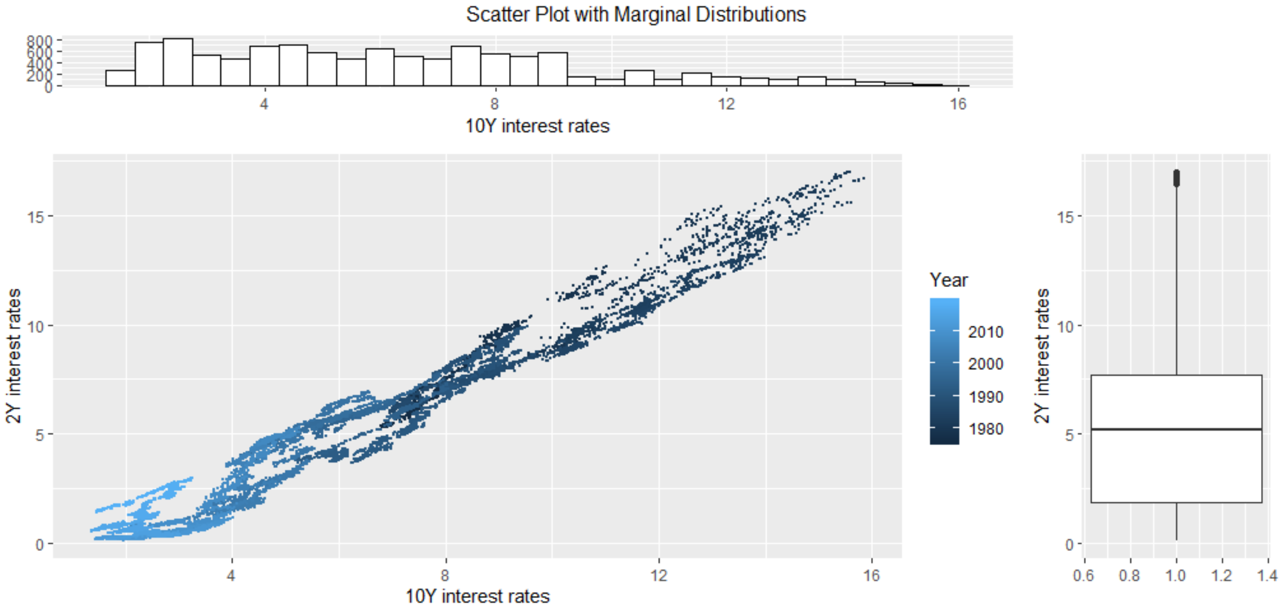
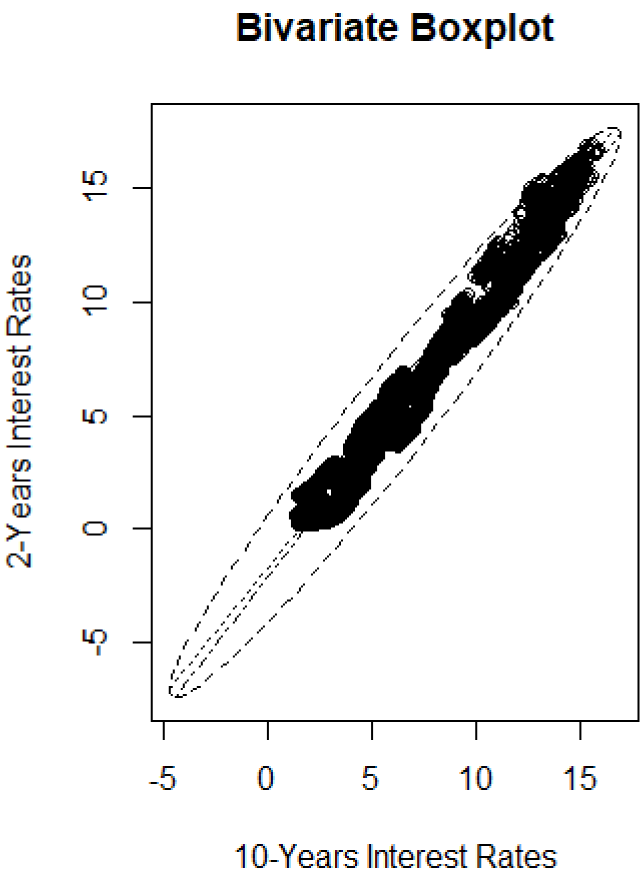


Figure 8: Scatter plot with marginal distributions

This plot summarizes all the information observed in the previous questions.

6. QUESTION e)

Depict the bivariate box plot. Comment on the outliers. Remove the outliers, if any, and re-compute the Pearson correlation coefficient.



As we could've guessed in the former questions, data is so deeply connected that it doesn't present any outlier in the bivariate boxplot.

Figure 9: Bivariate boxplot

7. QUESTION f)

Create the convex hull. Remove the observations lying on the hull and re-compute the correlation coefficient.

New correlation coefficient: 0.9766589

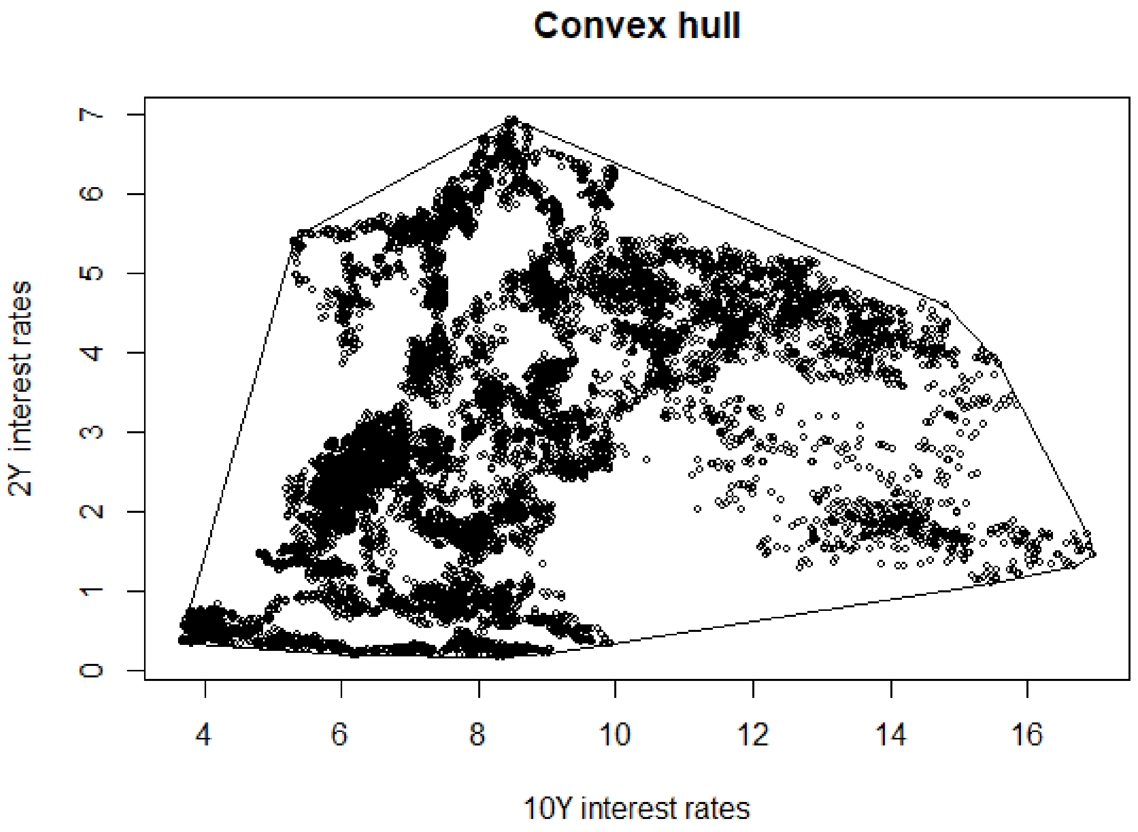


Figure 10: Convex hull

THIRD EXERCISE

1. INTRODUCTION

The dataset chosen for the third exercise is a collection of some SAT results in the state of New York (link to data https://catalog.data.gov/dataset/sat-results-e88d7, source "FRED"). It's divided into six columns, though we will only focus on the last three ("ReadingScore", "MathScore" and "WritingScore"). It contains 478 rows, but for the goals of this exercise I decided to limit them 50 (to avoid having unreadable graphs).

There's no missing value.

2. QUESTION a)

Pick up a dataset which has three variables (from source 2 or 3) and create the bubble plot. Interpret the result.

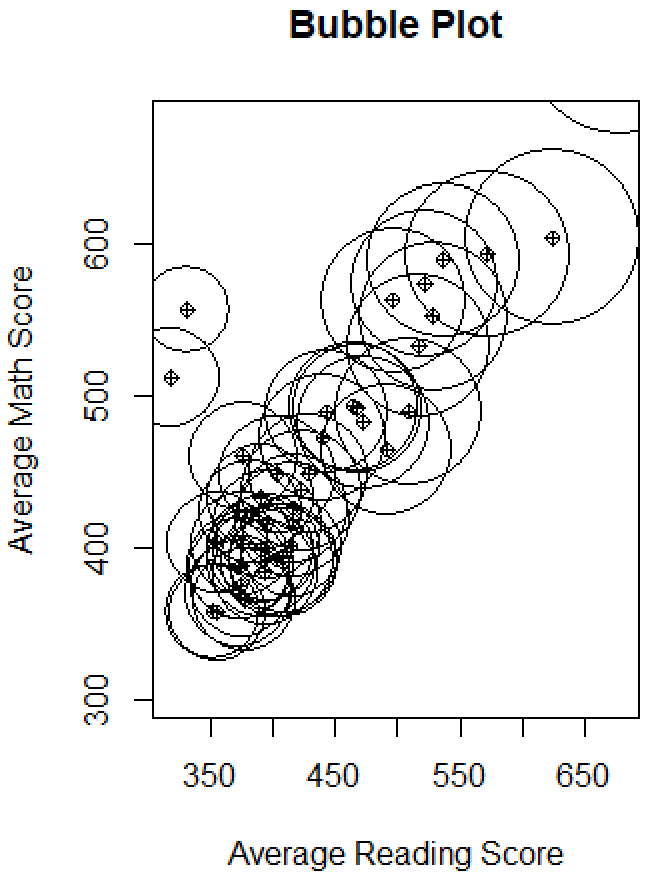


Figure 11: Bubble plot

It’s clear from the graph how results are quite connected to each other. Students with higher Reading Score tend to have higher scores in the two other topics as well.

It’s interesting to see how the only two evident outliers got a high score in math and a low score in Reading and Writing. We can speculate that they’re two students with a strong talent for numbers who didn’t bothered to study too much for the other parts of the exam.

3. QUESTION b)

Use data source 2 or 3. Create the glyph plot of all observations, Section 2.3. Do any stars look alike?

Observing this graph, we can confirm our initial impression. Most stars look symmetrical and they’re all quite similar to each other. We can also see how the 49th student appears to have gotten the best results in all fields.

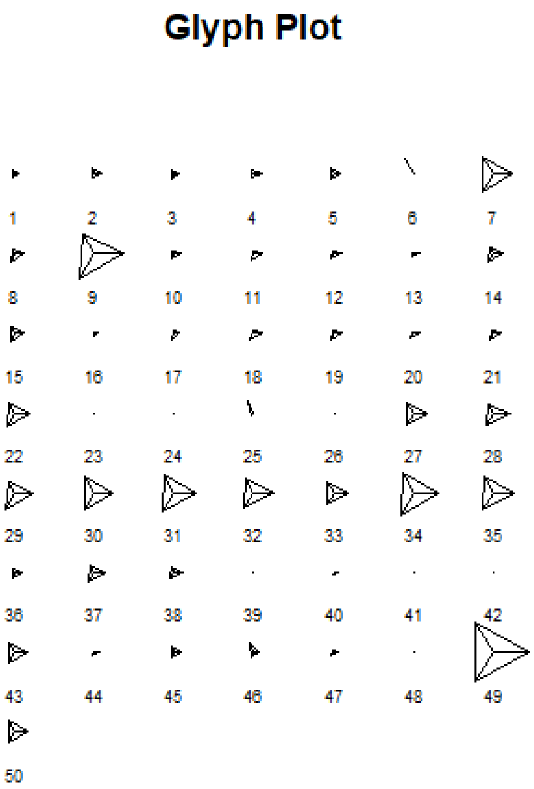


Figure 12: Glyph plot

4. QUESTION c)

Use data source 2 or 3. Create the scatter plot matrix and analyze it.

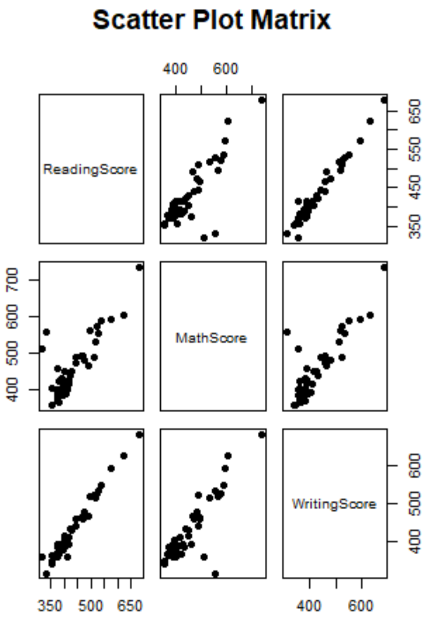


Figure 13: Scatter plot matrix

As already seen in the preceding points, we can again assume from this matrix how a linear relation exists between the scores in the three parts of the exam.

However, it's also interesting to observe how this correlation is much stronger between Reading and Writing score, while it contains some noise when Math scores are taken into consideration.

As in question a), we can speculate that this comes from the fact that Math scores are more associated to talent rather than studying.