



Project Write up

Visualization of Vietnam High School Education Dataset

2021 to 2022

Ivy Do

Abstract

This is the Project Write-up file which gets along with the Project Code file and Readme file. The document plays a primary role in explaining how I chose the specific preprocessing techniques and dealt with the autocorrelation phenomenon in the data set for optimizing the visualization part. The document is divided into three primary parts which are: preprocessing part, autocorrelation addressing part, and visualization part.

Table of Contents

Abstract	1
I. Introduction.....	3
II. Preprocessing stage.....	4
2.1. Finding and fixing problem – missing values	4
2.2. Categorizing variables	7
2.3. Checking duplicate observations	8
III. Autocorrelation addressing.....	9
IV. Visualization stage.....	12
4.1. Linear relationship between two variables.....	12
4.2. Multivariate plots.....	13
4.3. Gender visualization	14
4.4. Visualization for two discrete variables	14
4.5. Boxplot for spotting outliers.....	16
4.6. The distribution of unchosen cases in the optional subjects.	20
V. Conclusion	21
VI. References	22

I. Introduction

The data set is taken from Portal of the Ministry of Education and Training of Vietnam for the final examination for high school pupils in Nam Dinh province in 2022. The data set has 4,003 cases as observations and 15 variables in each 15 columns. The purpose of this document is to figure out problems in the data set as the raw data and fixing them for better usage of visualization. The visualization stage contains the scatter plots, box plots, pie charts, bar charts for showing the insights in the data set.

Hence, the document contains the first stage which is preprocessing. After that, the visualization stage shows all the necessary illustrations for discovering the insight from the data set. Please note that even though there are 15 variables with 15 columns respectively, there are some columns that will be not used for playing a vital role in discovering information, for example, the column of the year that the examination has been taken part in for the student (the “year” column) and the column of the date that the student enrolled for the optional subjects (the “enroll_date” column).

II. Preprocessing stage

2.1. Finding and fixing problem – missing values

Missing values, also known as missing, not filled or not updated values in the data set, can be the result of an error process in the data entry process. Cleaning input data is always accompanied by processing missing values.

It is necessary to first understand the nature of missing values, and then come up with a suitable solution to handle missing values. Usually, there are 2 ways to handle missing values:

- Method 1: Eliminate missing values (in case those missing values are not important to our data or the number of missing values is too small - only about less than 3% of the total number of observations in a given variable).
- Method 2: Replace missing values with another value. Which value to replace with will depend on the nature of the missing values in those cases.

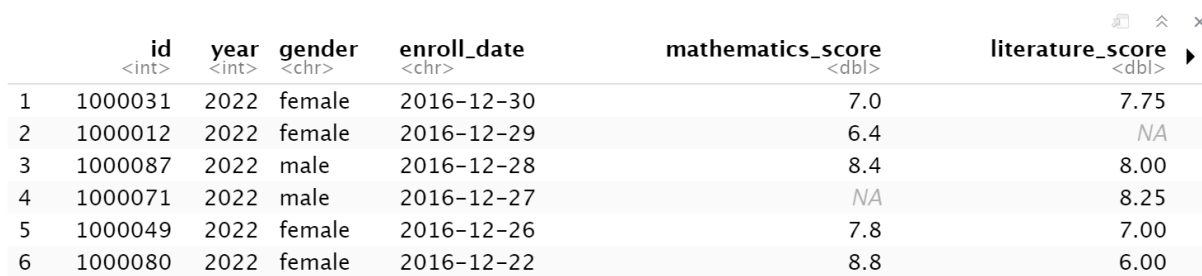
The question now is: So, if it is necessary to replace missing values with another value, which value should be replaced?

In case the variable with missing values is a variable - numeric: It is possible to replace missing values with values such as: 0, median, mean, etc. depending on the particular case.

The case of a variable with missing values is a categorical variable: It is possible to group the cases of missing values into a group, name it missing.

At first, the data set is imported from the csv file which names vietnam.csv by the function `df <- read.csv("vietnam.csv", header = TRUE, stringsAsFactors = FALSE)`. The data set is now names as `df`.

The function `head(df)` has been utilized for presenting the data set 6 first rows of the data set. Now we spot some missing value which names NA, as follows:



	id <int>	year <int>	gender <chr>	enroll_date <chr>	mathematics_score <dbl>	literature_score <dbl>
1	1000031	2022	female	2016-12-30	7.0	7.75
2	1000012	2022	female	2016-12-29	6.4	NA
3	1000087	2022	male	2016-12-28	8.4	8.00
4	1000071	2022	male	2016-12-27	NA	8.25
5	1000049	2022	female	2016-12-26	7.8	7.00
6	1000080	2022	female	2016-12-22	8.8	6.00

Figure 1: The first 7 rows of the data set

Therefore, the first preprocessing step that I would deal with is the missing value issue. The missing values appear in the mathematic score and literature score columns, this might have influence on the statistics measurements in each column if there are some missing values on it. In order to solve this problem, the missing values will suppose to be replaced by average value of the respective column to keep the statistical value in control and reflect the best of the data.

However, please note that there are solely the missing values from the compulsory subject columns will be fixed, which are columns `mathematic_score`, `literature_score`, and `foreign_language_score` column. The missing values in the compulsory subjects might appear because the students might be absent from the exam and get no score. The missing values in the optional subjects will be kept as original, as the missing values in the optional subjects appear because of some unchosen cases.

The data set after identifying and dealing with the missing values in the columns `mathematic_score`, `literature_score`, and `foreign_language_score` is as follows, as all the missing values are replaced by average values of respective variables.

	id	year	gender	enroll_date	mathematics_score	literature_score
1	1000031	2022	female	2016-12-30	7.000000	7.750000
2	1000012	2022	female	2016-12-29	6.400000	6.999112
3	1000087	2022	male	2016-12-28	8.400000	8.000000
4	1000071	2022	male	2016-12-27	7.108598	8.250000
5	1000049	2022	female	2016-12-26	7.800000	7.000000
6	1000080	2022	female	2016-12-22	8.800000	6.000000

Figure 2: The data set after replacing the missing value by the average values

2.2. Categorizing variables

In the vietnam.csv data set, there are some non-numeric data, which are the column gender and the column foreign_language_type. Hence, it is necessary to categorizing the variables and turn them into factor variables with function *factor*, as follows:

The gender column is the nominal scale variable. In this scale the numbers are only used to classify objects, they have no other meaning. In essence, the nominal scale is the classification and naming of expressions and assigning them a corresponding number. The nominal scale helps to convert the individuals answering this question into expressions of the variable “gender”. We can convention put Male = 1, Female = 0. These numbers are nominal because we cannot add or calculate the average value of “gender”. Statistical operations that can be used for nominal scales include counting, calculating the frequency of an expression, determining the mode value, and performing a number of tests (these tests will be in the visualization section below)

```
df$gender <- factor(df$gender, levels = c("female", "male"), labels = c(0, 1))
```

The foreign_language_type is ordinal scale, is a scale where the numbers on the nominal scale are arranged according to some convention of order or superiority, but we do not know the distance between them. This means that any hierarchical scale is a nominal scale but cannot be inverted. In the foreign_language_type variable, the

answer is conventionally N1 to N7. However, N7 is not 7 times higher than N1, but only shows people with higher N7 language proficiency than people with N1 language proficiency.

```
df$foreign_language_type <- factor(df$foreign_language_type,  
  
    levels = c("N1", "N2", "N3", "N4", "N5", "N6", "N7"),  
  
    labels = c(1, 2, 3, 4, 5, 6, 7))
```

2.3. Checking duplicate observations

The first column which names *id* is the student's number which is unique for each student over the time. Hence, it is necessary to check if there is any similar id number in the total observations, if any, there would be solved immediately, such as remove the unnecessary observations.

In order to quickly review rows with duplicates, I would like to use the function *duplicated()* from the *tidyverse* package. By default, all columns are considered when evaluating duplicates - the rows returned by the function are 100% duplicates considering the values in all columns.

In order to spot the position of duplicated observation in df data frame, the function *duplicated(df)* has been utilized. Then using the function *df[duplicated(df)]* to extract the duplicated elements (if any). The results show there is no duplicated element in the dataframe df.

III. Autocorrelation addressing

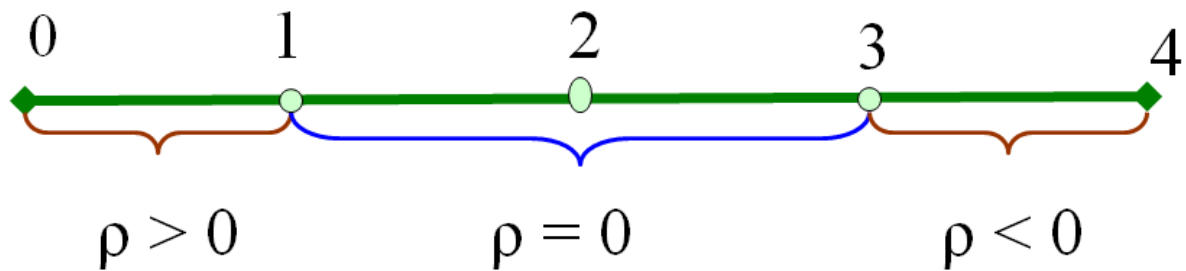
Autocorrelation is the phenomenon where the noise term at time t (also known as error) usually denoted u_t is correlated with the noise term at time $(t-1)$ or any other plural term in the past. In time-series data the autocorrelation is called Autocorrelation and in panel data the autocorrelation is called Serial Correlation.

The general formula is as follows:

$$U_{it} = \beta * U_{it-1} + c_{it}$$

U is the noise term at t and $t-1$, Coefficient $\beta \neq 0$ has autocorrelation and vice versa ($i = 0$ when it's time-series). This phenomenon violates the hypothesis in the classical linear regression model that assumes that autocorrelation does not exist in the disturbances u_i .

There are numerous methods to examine autocorrelation phenomenon. In the document, the autocorrelation will be tested by Durbin-Waston test. According to Mukhtar M. Ali, the autocorrelation phenomenon could be detected by using the test Durbin-Watson test with three ranges as follows:



Whereas p is the autocorrelation level. When $p = 0$, it means there is no autocorrelation. When $p > 0$ or $p < 0$, it means that the model has autocorrelation phenomenon. The Durbin Watson point will vary from 0 to 4. When the Durbin Watson point varies from 0 to 1, it means $p > 0$, so that the model has autocorrelation. When the Durbin Watson point varies from 1 to 3, it means that $p = 0$, so that there is no autocorrelation. When the Durbin Watson point varies from 3 to 4, it means $p < 0$, so that the model has autocorrelation.

At first, we come up with the model that the variable y is mathematical score (the dependent variable), and the variable x is gender (the independent variable). Then the expected model is $y = a * x$ while a is a constant number to indicate the relationship between the independent and dependent variables. Here we have the hypotheses that state as follows:

- Null hypothesis: H_0 : The model does not occur autocorrelation
- Alternative hypothesis: H_1 : The model occurs autocorrelation

By using the function `dwtest(y~x)` from the `lmtest` package, we have the result of Durbin-Watson test as below:

```
Durbin-Watson test

data:  y ~ x
DW = 1.7006, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Figure 3: Durbin-Watson test

The Durbin-Watson test can be linearly mapped to the Pearson correlation between the values and their lags. Durbin-Watson always produces a range of experimental numbers from 0 to 4. Values close to 0 indicate a greater degree of positive correlation, values close to 4 indicate a greater degree of negative autocorrelation, while values close to 4 indicate a greater degree of negative autocorrelation. Values closer to the middle indicate less autocorrelation. In this case, the Durbin-Watson score is 1.7006 which is closed to 2 (the middle point). Therefore, it is less autocorrelation for the model between the independent variable (gender) and the dependent variable (mathematical score). Hence, accept the null hypothesis: H_0 : The model does not occur autocorrelation.

IV. Visualization stage

4.1. Linear relationship between two variables

Continuing exploring the relationship between pupils' gender and their mathematical score. There is one rumor that male does better mathematical job and tasks than female, the part will give us the evidence that shows this rumor is right or wrong. The independent variable (x variable) is gender, and the dependent variable (y variable) is mathematical score. We would like to use the function *smoothScatter()* and *scatter.smooth()* to explore whether the gender and mathematical score variables are linearly or nonlinearly related. At first, the *smoothScatter()* function shows there might have similar points. Then the function *scatter.smooth()* with a straight regression line (the red line) will answer the question: "Are that the two variable linearly?" and the answer is no, they are nonlinearly with no red regression line.

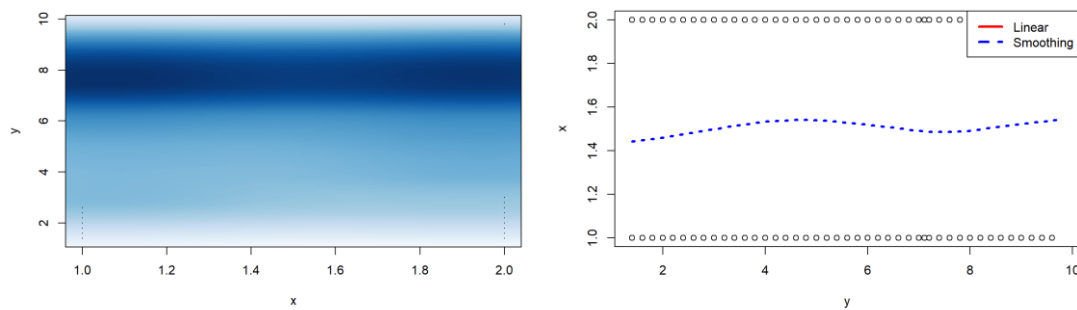


Figure 4: Scatter plot for linear relationship between gender and mathematical score of students

4.2. Multivariate plots

The functions `plot()` and `pairs()` have been utilized in this part to visualize the variables `mathematical_score`, `literature_score`, and `foreign_language_score`. The following picture indicates the relationship and similarity among the three variables. In general, the scores of mathematical, literature and foreign language subjects are somewhat high.

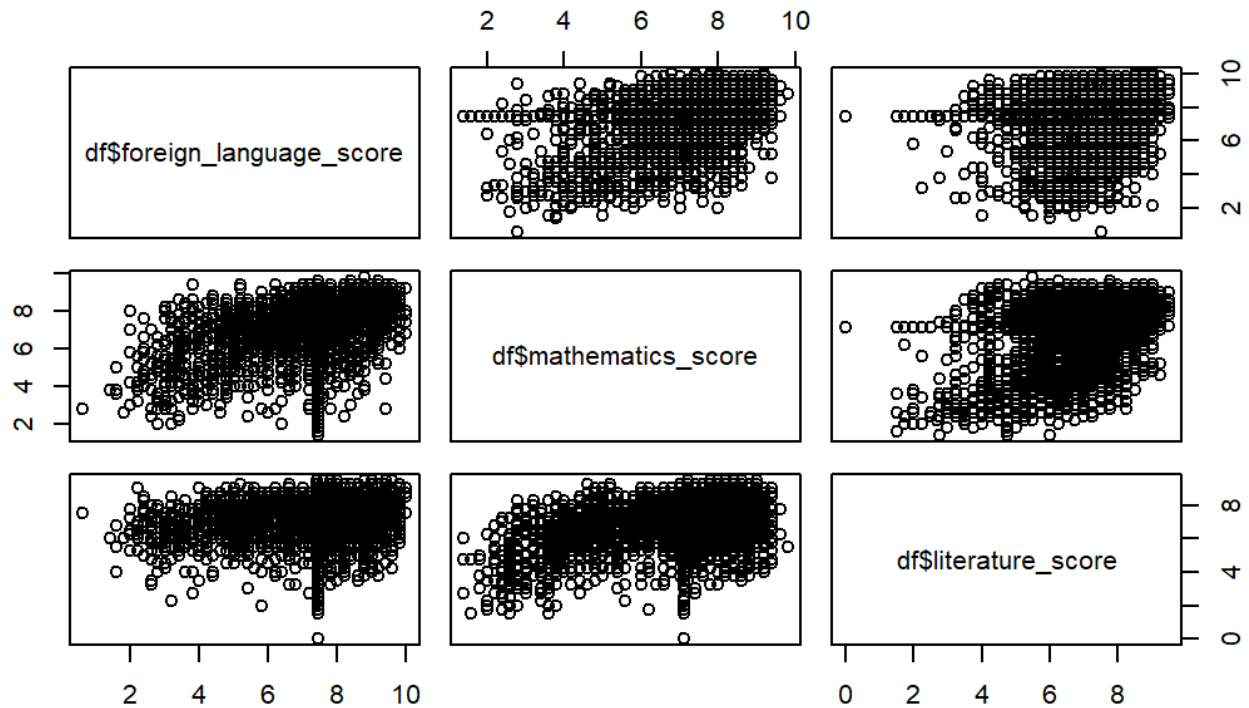


Figure 5: The multivariate plots for mathematical, literature and foreign language scores

4.3. Gender visualization

The bar plot has been used for representing male and female frequencies in horizontal lines. The result shows that there is the evenly distribution for gender in this examination while they are equal in the number of female and male pupils. The function `barplot()` has been used in this case, we might choose the vertical or horizontal bar style (as I choose the horizontal style) as below:

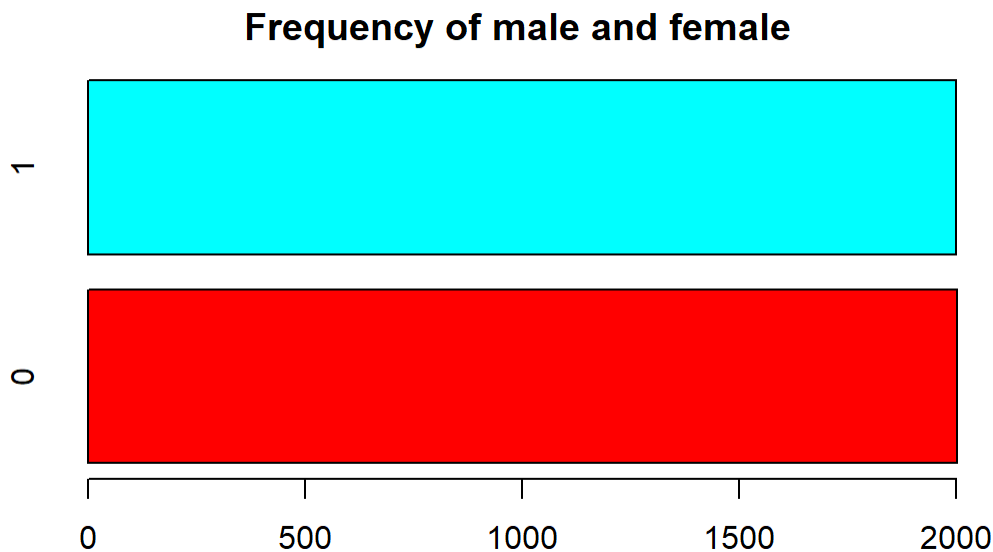


Figure 6: The bar chart of gender frequency

4.4. Visualization for two discrete variables

The `literature_score` column is a continuous variable. We can divide it into groups based on the number of points. The `cut` function has the function of "cutting" a continuous variable into many discrete groups. The frequency of the three groups is as below. The picture below is the result of combining `literature_score` and gender

variables into three groups. It shows there are 28 female and 24 male pupil in the first group (the lowest score group), the second group is much more pupil with 549 female and 531 male pupils. Then the highest score group which varies from 6.33 to 9.51 grade, has 1425 female and 1446 male students.

In this phrase, the `cut()` function has been utilized for dividing the variable `literature_score` into three groups, specifically the function looks like: `cut(df$literature_score,3)`. Then grouping the variable `gender` and `literature_score` into a table for presenting the number of male and female pupils in each literature score group

	lit (-0.0095,3.17]	(3.17,6.33]	(6.33,9.51]
	52	1080	2871
lit			
	(-0.0095,3.17]	(3.17,6.33]	(6.33,9.51]
0	28	549	1425
1	24	531	1446

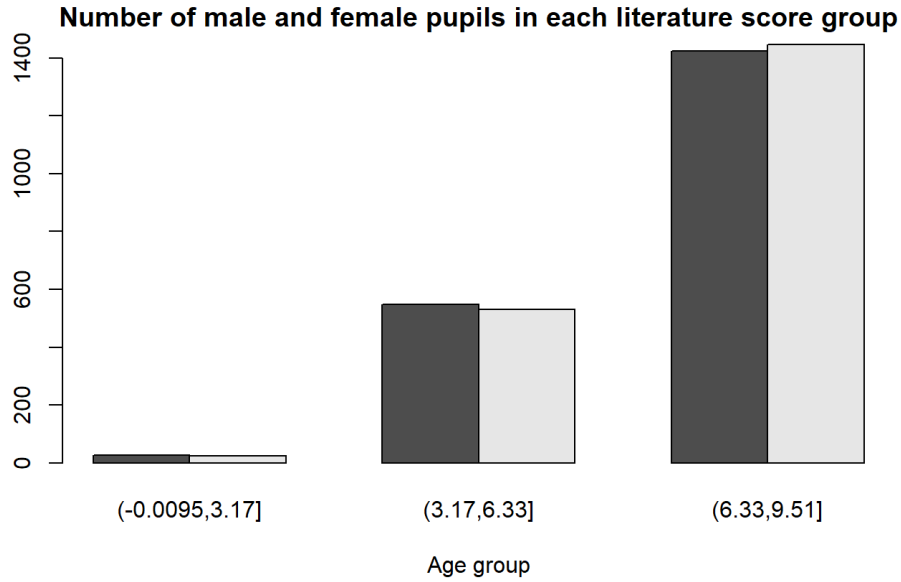


Figure 7: The bar chart of number of male and female pupils in three literature score group

It seems that even the literature score has been divided into three separately groups, there is no difference between the female and male students in the distribution in each group.

4.5. Boxplot for spotting outliers

The following boxplots are used for indicating outliers and spotting quartiles. Besides detecting outliers with boxplot, we can use the normalized residual Scatter graph to improve the linear regression results. However, this document only covers the method of finding outliers via boxplot. In the boxplot plots, outliers are denoted by circles (o).

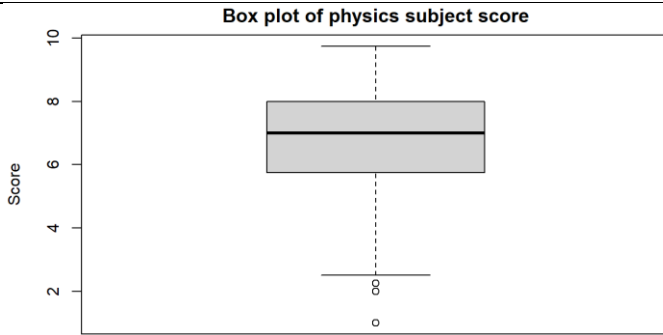


Figure 8: Box plot of physics subject score

There are three outliers in the physics subject score which are around grade 2, they could be seen as *extreme outliers*

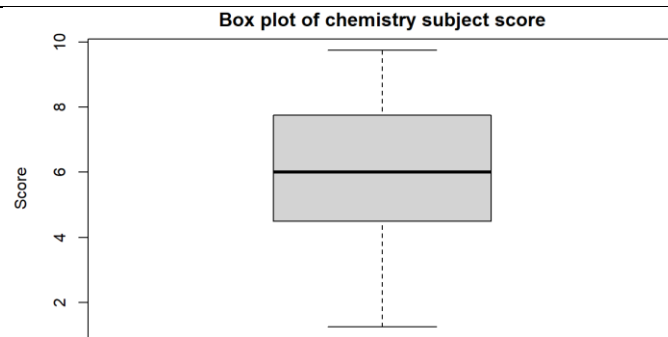


Figure 9: Box plot of chemistry subject score

In the chemistry subject score, there is no outlier, and the median is lower than the physics subject. However, the interquartile range is wider.

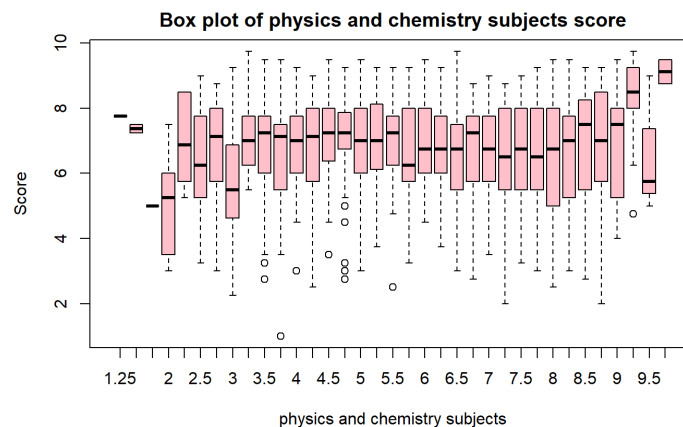


Figure 10: Box plot of physics and chemistry subjects score

There are some outliers which belongs to the cases that have lower score than average. The median score is above 5, which is the acceptable score, the minimum score to pass each subject in Vietnam.

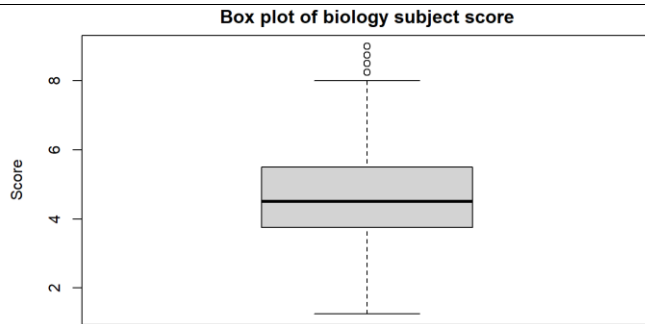


Figure 11: Box plot of biology subject score

In general, the median is around 4 and 5 grade; however, there are four extreme outliers they are higher than grade 8 in the biology score

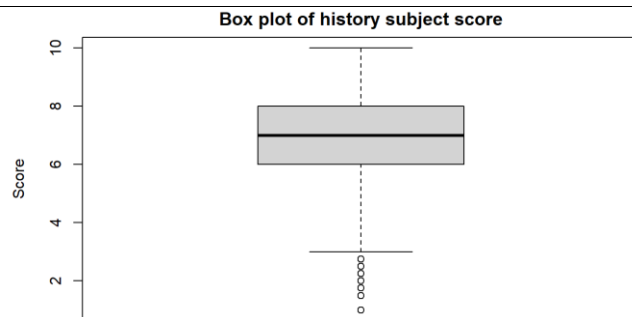


Figure 12: Box plot of history subject score

The history subject has the median is around 7, which is quite high. Nevertheless, there are 7 outliers which are under the grade 3.

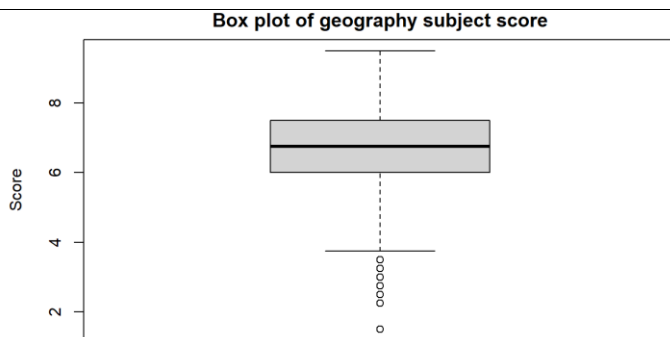


Figure 13: Box plot of geography subject score

The geography score has the same figure likes the history score: median and outlier numbers. However, the 7 outliers are under grade 4.

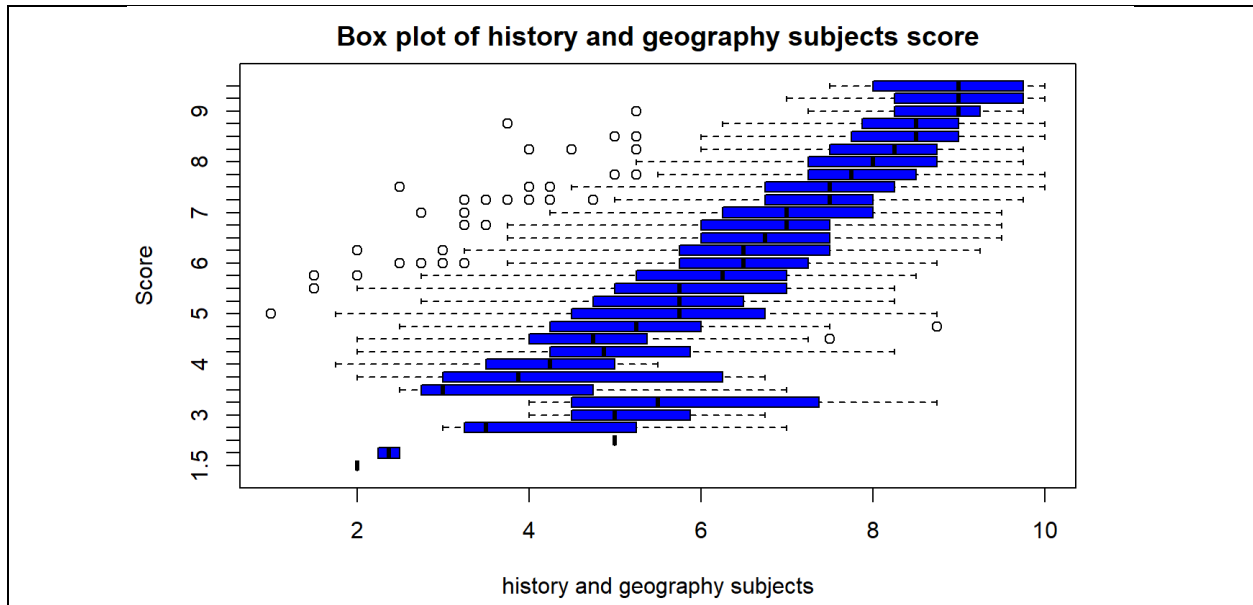


Figure 14: Box plot of history and geography subjects score

There are a lot of outliers in the case, and that have a bad influence on the interquartile range and the median score.

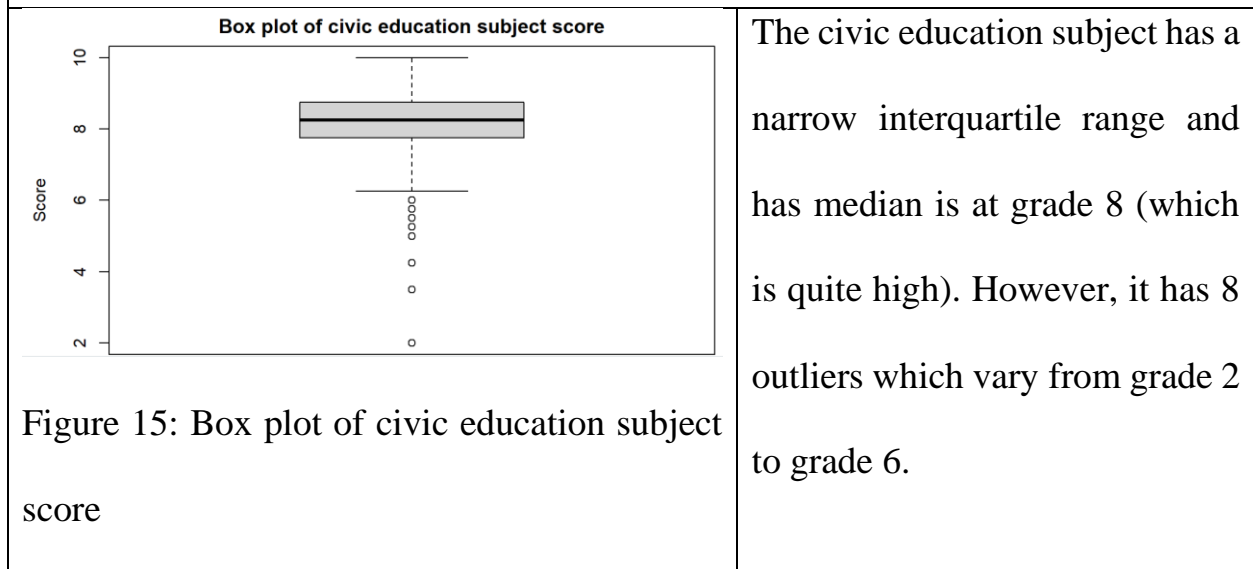


Figure 15: Box plot of civic education subject score

In general, all the outliers in the box plots above come from scores that are too low or too high for the interquartile range. It is quite necessary to detect extreme outliers

or outliers in the total number of students taking the test. The reason is because each year, each province will sum up the average score of each subject, the highest score, the lowest score to evaluate for next year's exam questions. The appearance of outliers will cause the statistics on those subjects to be changed and not displayed correctly. In addition, the appearance of many outliers and processing them will make the linear regression model more suitable.

However, in this document, we only stop at detecting outliers, sometimes detecting outliers and paying attention to each observation in outliers will help teachers pay more attention to each weak student. The removal of outliers from a study should be done with caution and consideration, as doing so may detract from the practicality of the study as well as significantly reduce the sample size.

4.6. The distribution of unchosen cases in the optional subjects.

The pie chart below illustrates the distribution of unchosen in the optional subjects. Besides three compulsory subjects, students could choose one optional natural sciences course which is physics, chemistry, or biology; and one optional social science which could be history, geography, or civic education either, they all are optional. However, the number of unchosen optional subjects are equally distributed which is 15 cases. Hence, we have the following pie chart which is divided equally into 6 proportions.

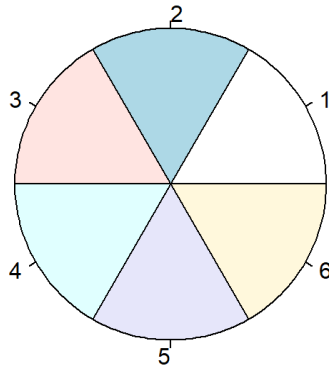
The distribution of optional subject

Figure 16: Pie chart of the unchosen cases in optional subject distribution

V. Conclusion

By using the preprocessing techniques which include finding and fixing missing values, categorizing the non-numeric variables, and examining the duplicate observations, addressing autocorrelation; the visualization part has been optimized. The scatter plot has been used for exploring the linear relationship between two variables: pupils' gender and their mathematical score; however, they are nonlinearly. In the part of multivariable plot, it shows that the variables `mathematical_score`, `literature_score`, and `foreign_language_score` has similar trend. Furthermore, exploring the gender distribution in the literature score, there is no difference between the female and male students in the distribution in three divided groups. The outliers in each subject score are vary, except the chemistry subject score, the other subject scores have their own outliers. The last insight belongs to the distribution of unchosen in the six optional subjects (physics, chemistry, biology,

history, geography, or civic education), although they are nonmandatory, there is the evenly distribution in the six subjects that each subjects has 15 unchosen cases.

VI. References

John Maindonald (2003). *Data Analysis and Graphics Using R – An Example Approach*. Cambridge University Press.

Julian Faraway (2004). *Linear Models with R*. Chapman & Hall/CRC

Peter Dalgaard (2004). *Introductory Statistics with R*. Springer Publisher.

Stephen C. Few (2012). *Show me the numbers – designing tables and graphs to enlighten*. Analytics Press.