#### Data Analysis



DATA ANALYSIS WITH STUDENT INFORMATION DATABASE

Table of Contents

1. [Dataset Description 2](#_bookmark0)
2. [MATLAB Code 3](#_bookmark1)
3. [Execution Outcome 8](#_bookmark2)

[Centered Data Matrix 12](#_bookmark3)

[Boxplot Graphs 13](#_bookmark4)

[Age 13](#_bookmark5)

[GPA 13](#_bookmark6)

[Absences 14](#_bookmark7)

[Fedu 14](#_bookmark9)

[G1 15](#_bookmark10)

[G2 15](#_bookmark8)

[G3 16](#_bookmark11)

[Health 16](#_bookmark12)

[Traveltime 17](#_bookmark13)

[Studytime 17](#_bookmark14)

[Medu 18](#_bookmark15)

1. [Interpretation of the skewness and kurtosis values 18](#_bookmark16)
2. [References 20](#_bookmark17)

# 1. Dataset Description

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student ages, grades, demographic, social and school related features and it was collected by using school reports and questionnaires

**Dataset size:** 396 x 19

This dataset has enough data to reach reliable information because the size of dataset is very large, and it is very detailed.

The reason to choose this dataset is it is clear to understand, and it is interesting to learn information about Portuguese students. Because the education system is different from Turkey.

The student information dataset gives personal and educational information about Portuguese students. For example, their GPA or their mother’s and father’s education level. The Education level of mother of Portuguese students is 3 (rounding 2.97) and father of Portuguese students is 2 (rounding 2.5). It is interesting that the education level of mothers is higher then fathers. Also the study time of the students is double of travel time

**Important note:** the target attribute G3 has a strong correlation with attributes G2 and G1.

This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful (see paper source for more details).

**Attribute Information:**

1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. sex - student's sex (binary: 'F' - female or 'M' - male)
3. age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: 'U' - urban or 'R' - rural)
5. gpa – student’s GPA grades (numeric: from 0.00 to 4.00)
6. Medu – mother’s education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
7. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade,
8. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')
9. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other')
10. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
11. guardian - student's guardian (nominal: 'mother', 'father' or 'other')
12. traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
13. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
14. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
15. health - current health status (numeric: from 1 - very bad to 5 - very good)
16. absences - number of school absences (numeric: from 0 to 93)
17. G1 - first period grade (numeric: from 0 to 20)
18. G2 - second period grade (numeric: from 0 to 20)
19. G3 - final grade (numeric: from 0 to 20, output target)

# MATLAB Code

%%Reading data

A = [age gpa Medu Fedu traveltime studytime health absences G1 G2 G3];

%%%% calculate mean %%% mAge = mean(age);

mGPA = mean(gpa);

mMedu = mean(Medu); mFedu = mean(Fedu);

mTraveltime = mean(traveltime); mStudytime = mean(studytime); mHealth = mean(health); mAbsences = mean(absences);

mG1 = mean(G1); mG2 = mean(G2); mG3 = mean(G3);

%%% calculate median %%%

medAge = median(age); medGPA = median(gpa); medMedu = median(Medu); medFedu = median(Fedu);

medTraveltime = median(traveltime); medStudytime = median(studytime); medHealth = median(health); medAbsences = median(absences); medG1 = median(G1);

medG2 = median(G2); medG3 = median(G3);

%%% calculate sum %%% sAge = sum(age);

sGPA = sum(gpa); sMedu = sum(Medu); sFedu = sum(Fedu);

sTraveltime = sum(traveltime); sStudytime = sum(studytime); sHealth = sum(health); sAbsences = sum(absences);

sG1 = sum(G1); sG2 = sum(G2); sG3 = sum(G3);

%%% calculate max %%%

maxAge = max(age); maxGPA = max(gpa); maxMedu = max(Medu); maxFedu = max(Fedu);

maxTraveltime = max(traveltime); maxStudytime = max(studytime); maxHealth = max(health); maxAbsences = max(absences); maxG1 = max(G1);

maxG2 = max(G2); maxG3 = max(G3);

%%% calculate range %%% rAge = range(age);

rGPA = range(gpa); rMedu = range(Medu); rFedu = range(Fedu);

rTraveltime = range(traveltime); rStudytime = range(studytime); rHealth = range(health); rAbsences = range(absences);

rG1 = range(G1); rG2 = range(G2); rG3 = range(G3);

%%% calculate skewness %%%

skAge = skewness(age); skGPA = skewness(gpa); skMedu = skewness(Medu); skFedu = skewness(Fedu);

skTraveltime = skewness(traveltime); skStudytime = skewness(studytime); skHealth = skewness(health); skAbsences = skewness(absences); skG1 = skewness(G1);

skG2 = skewness(G2); skG3 = skewness(G3);

%%% calculate kurtosis %%%

kuAge = kurtosis(age); kuGPA = kurtosis(gpa); kuMedu = kurtosis(Medu); kuFedu = kurtosis(Fedu);

kuTraveltime = kurtosis(traveltime); kuStudytime = kurtosis(studytime);

kuHealth = kurtosis(health); kuAbsences = kurtosis(absences); kuG1 = kurtosis(G1);

kuG2 = kurtosis(G2); kuG3 = kurtosis(G3);

%%% calculate number of outliers %%% noAge = isoutlier(age);

noGPA = isoutlier(gpa);

noMedu = isoutlier(Medu); noFedu = isoutlier(Fedu);

noTraveltime = isoutlier(traveltime); noStudytime = isoutlier(studytime); noHealth = isoutlier(health); noAbsences = isoutlier(absences); noG1 = isoutlier(G1);

noG2 = isoutlier(G2); noG3 = isoutlier(G3);

%% inputs %%

answ = input('For mean press 1\nFor median press 2\nFor sum press 3\nFor max press 4\nFor range press 5\nFor skewness press 6\nFor kurtosis press 7\nFor boxplot press 8\nFor number of outliers press 9\n');

if answ == 1

display(mean(A));

fprintf('Mean of ages: %f', mAge); fprintf('\nMean of GPAs: %f', mGPA); fprintf('\nMean of Medu: %f', mMedu); fprintf('\nMean of Fedu: %f', mFedu);

fprintf('\nMean of travel time: %f', mTraveltime); fprintf('\nMean of study time: %f', mStudytime); fprintf('\nMean of health: %f', mHealth); fprintf('\nMean of Absences: %f', mAbsences); fprintf('\nMean of G1: %f', mG1);

fprintf('\nMean of G2: %f', mG2); fprintf('\nMean of G3: %f', mG3);

elseif answ == 2 display(median(A));

fprintf('Median of ages: %f', medAge); fprintf('\nMedian of GPAs: %f', medGPA); fprintf('\nMedian of Medu: %f', medMedu); fprintf('\nMedian of Fedu: %f', medFedu); fprintf('\nMedian of travel time: %f', medTraveltime); fprintf('\nMedian of study time: %f', medStudytime); fprintf('\nMedian of health: %f', medHealth); fprintf('\nMedian of Absences: %f', medAbsences); fprintf('\nMedian of G1: %f', medG1); fprintf('\nMedian of G2: %f', medG2); fprintf('\nMedianof G3: %f', medG3);

elseif answ == 3 sumA = sum(A); display(sumA);

fprintf('Sum of ages: %f', sAge); fprintf('\nSum of GPAs: %f', sGPA); fprintf('\nSum of Medu: %f', sMedu);

fprintf('\nSum of Fedu: %f', sFedu); fprintf('\nSum of travel time: %f', sTraveltime); fprintf('\nSum of study time: %f', sStudytime); fprintf('\nSum of health: %f', sHealth); fprintf('\nSum of Absences: %f', sAbsences); fprintf('\nSum of G1: %f', sG1);

fprintf('\nSum of G2: %f', sG2); fprintf('\nSum of G3: %f', sG3);

elseif answ == 4 display(max(A));

fprintf('Max of ages: %f', maxAge); fprintf('\nMax of GPAs: %f', maxGPA); fprintf('\nMax of Medu: %f', maxMedu); fprintf('\nMax of Fedu: %f', maxFedu);

fprintf('\nMax of travel time: %f', maxTraveltime); fprintf('\nMax of study time: %f', maxStudytime); fprintf('\nMax of health: %f', maxHealth); fprintf('\nMax of Absences: %f', maxAbsences); fprintf('\nMax of G1: %f', maxG1);

fprintf('\nMax of G2: %f', maxG2); fprintf('\nMax of G3: %f', maxG3);

elseif answ == 5 display(range(A));

fprintf('Range of ages: %f', rAge); fprintf('\nRange of GPAs: %f',rGPA); fprintf('\nRange of Medu: %f', rMedu); fprintf('\nRange of Fedu: %f', rFedu); fprintf('\nRange of travel time: %f', rTraveltime); fprintf('\nRange of study time: %f', rStudytime); fprintf('\nRange of health: %f', rHealth); fprintf('\nRange of Absences: %f', rAbsences); fprintf('\nRange of G1: %f', rG1);

fprintf('\nRange of G2: %f', rG2); fprintf('\nRange of G3: %f', rG3);

elseif answ == 6 display(skewness(A));

fprintf('Skewness of ages: %f', skAge ); fprintf('\nSkewness of GPAs: %f',skGPA); fprintf('\nSkewness of Medu: %f', skMedu); fprintf('\nSkewness of Fedu: %f', skFedu); fprintf('\nSkewness of travel time: %f', skTraveltime); fprintf('\nSkewness of study time: %f', skStudytime); fprintf('\nSkewness of health: %f', skHealth); fprintf('\nSkewness of Absences: %f', skAbsences); fprintf('\nSkewness of G1: %f', skG1); fprintf('\nSkewness of G2: %f', skG2); fprintf('\nSkewness G3: %f', skG3);

elseif answ == 7 display(kurtosis(A));

fprintf('Kurtosis of ages: %f', kuAge ); fprintf('\nKurtosis of GPAs: %f',kuGPA); fprintf('\nKurtosis of Medu: %f', kuMedu); fprintf('\nKurtosis of Fedu: %f', kuFedu); fprintf('\nKurtosis of travel time: %f', kuTraveltime); fprintf('\nKurtosis of study time: %f', kuStudytime); fprintf('\nKurtosis of health: %f', kuHealth);

fprintf('\nKurtosis of Absences: %f', kuAbsences); fprintf('\nKurtosis of G1: %f', kuG1); fprintf('\nKurtosis of G2: %f', kuG2); fprintf('\nKurtosis G3: %f', kuG3);

%%%% calculate boxplots seperatly %%% elseif answ == 8

display(boxplot(A));

answ2 = input('\nFor boxplot of age press 1\nFor boxplot of gpa press 2\nFor boxplot of medu press 3\nFor boxplot of Fedu press 4\nFor boxplot of traveltime press 5\nFor boxplot of studytime press 6\nFor boxplot of health press 7\nFor boxplot of abcenses press 8\nFor boxplot of G1 press 9\nFor boxplot of G2 press 10\nFor boxplot of G3 press 11\n');

if answ2 == 1

bpAge = boxplot(age); display(bpAge);

elseif answ2 == 2

bpGPA = boxplot(gpa); display(bpGPA);

elseif answ2 == 3

bpMedu = boxplot(Medu); display(bpMedu);

elseif answ2 == 4

bpFedu = boxplot(Fedu); display(bpFedu);

elseif answ2 == 5

bpTraveltime = boxplot(traveltime); display(bpTraveltime);

elseif answ2 == 6

bpStudytime = boxplot(studytime); display(bpStudytime);

elseif answ2 == 7

bpHealth = boxplot(health); display(bpHealth);

elseif answ2 == 8

bpAbsences = boxplot(absences); display(bpAbsences);

elseif answ2 == 9

bpG1 = boxplot(G1); display(bpG1);

elseif answ2 == 10 bpG2 = boxplot(G2); display(bpG2);

elseif answ2 == 11 bpG3 = boxplot(G3); display(bpG3);

else

fprintf("Invalid number");

end

elseif answ == 9

fprintf('\nNumber of outliers of ages: %f', sum(noAge)); fprintf('\nNumber of outliers of GPAs: %f', sum(noGPA)); fprintf('\nNumber of outliers of Medu: %f', sum(noMedu)); fprintf('\nNumber of outliers of Fedu: %f', sum(noFedu)); fprintf('\nNumber of outliers of traveltime: %f', sum(noTraveltime)); fprintf('\nNumber of outliers of studytime: %f', sum(noStudytime)); fprintf('\nNumber of outliers of health: %f', sum(noHealth)); fprintf('\nNumber of outliers of absences: %f', sum(noAbsences)); fprintf('\nNumber of outliers of G1: %f', sum(noG1));

fprintf('\nNumber of outliers of G2: %f', sum(noG2)); fprintf('\nNumber of outliers of G3: %f', sum(noG3));

else

fprintf("Unvalid number.");

end

%%% CENTERED DATA MATRIX %%% tA = A' ;

meanA = mean(A);

fprintf("\n");

cmat = A' - meanA' ; fprintf("\n"); disp(cmat);

# Execution Outcome

For mean press 1 For median press 2 For sum press 3 For max press 4

For For range press 5 For skewness press 6 For kurtosis press 7 For boxplot press 8

For number of outliers press 9

1

16.6962

2.0289

2.7494

2.5215

1.4481

2.0354

3.5544

5.7089 10.9089 10.7139 10.4152

Mean of ages: 16.696203 Mean of GPAs: 2.028911 Mean of Medu: 2.749367 Mean of Fedu: 2.521519

Mean of travel time: 1.448101 Mean of study time: 2.035443 Mean of health: 3.554430 Mean of Absences: 5.708861 Mean of G1: 10.908861

Mean of G2: 10.713924 Mean of G3: 10.415190

2

17.0000

1.9600

3.0000

2.0000

1.0000

2.0000

4.0000

4.0000 11.0000 11.0000 11.0000

Median of ages: 17.000000 Median of GPAs: 1.960000 Median of Medu: 3.000000 Median of Fedu: 2.000000 Median of travel time: 1.000000 Median of study time: 2.000000 Median of health: 4.000000 Median of Absences: 4.000000 Median of G1: 11.000000

Median of G2: 11.000000 Median of G3: 11.000000

3

sumA =

6595.0

801.420

1086.0

996.0

572.00

804.00

1404.00

2255.0 4309.0 4.2320 4114.00

Sum of ages: 6595.000000 Sum of GPAs: 801.420000

Sum of Medu: 1086.000000 Sum of Fedu: 996.000000

Sum of travel time: 572.000000 Sum of study time: 804.000000 Sum of health: 1404.000000 Sum of Absences: 2255.000000 Sum of G1: 4309.000000

Sum of G2: 4232.000000 Sum of G3: 4114.000000

4

22.0000

3.9900

4.0000 4.0000

4.0000

4.0000

5.0000

75.0000 19.0000 19.0000 20.0000

Max of ages: 22.000000 Max of GPAs: 3.990000

Max of Medu: 4.000000 Max of Fedu: 4.000000

Max of travel time: 4.000000 Max of study time: 4.000000 Max of health: 5.000000

Max of Absences: 75.000000 Max of G1: 19.000000

Max of G2: 19.000000

Max of G3: 20.000000

5

7.0000

3.9800

4.0000

4.0000

3.0000

3.0000

4.0000

75.0000 16.0000 19.0000 20.0000

Range of ages: 7.000000 Range of GPAs: 3.980000 Range of Medu: 4.000000 Range of Fedu: 4.000000

Range of travel time: 3.000000 Range of study time: 3.000000 Range of health: 4.000000 Range of Absences: 75.000000 Range of G1: 16.000000

Range of G2: 19.000000 Range of G3: 20.000000

6

0.4645

0.0092 -0.3172 -0.0316

1.6009

0.6297 -0.4927

3.6576 0.2397 -0.4300 -0.7299

Skewness of ages: 0.464498 Skewness of GPAs: 0.009155 Skewness of Medu: -0.317170 Skewness of Fedu: -0.031552 Skewness of travel time: 1.600920 Skewness of study time: 0.629739 Skewness of health: -0.492723 Skewness of Absences: 3.657622 Skewness of G1: 0.239699 Skewness of G2: -0.430004 Skewness G3: -0.729887

7

2.9836 1.8163 1.9086

1.8014

5.2995

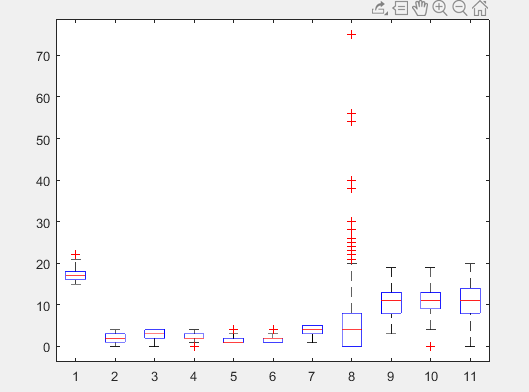
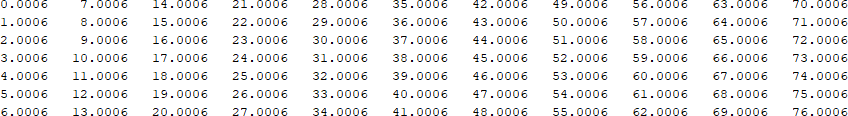
2.9706

1.9836

24.4300 2.2998 3.6046 3.3832

Kurtosis of ages: 2.983642 Kurtosis of GPAs: 1.816286 Kurtosis of Medu: 1.908583 Kurtosis of Fedu: 1.801427 Kurtosis of travel time: 5.299470 Kurtosis of study time: 2.970599 Kurtosis of health: 1.983561 Kurtosis of Absences: 24.430045

Kurtosis of G1: 2.299771 Kurtosis of G2: 3.604637 Kurtosis G3: 3.383181



For boxplot of age press 1 For boxplot of gpa press 2 For boxplot of medu press 3 For boxplot of Fedu press 4

For boxplot of traveltime press 5 For boxplot of studytime press 6 For boxplot of health press 7 For boxplot of abcenses press 8 For boxplot of G1 press 9

For boxplot of G2 press 10 For boxplot of G3 press 11

The graphs of the above values can be found in the "**Boxplot Graphs**" section.

8

9

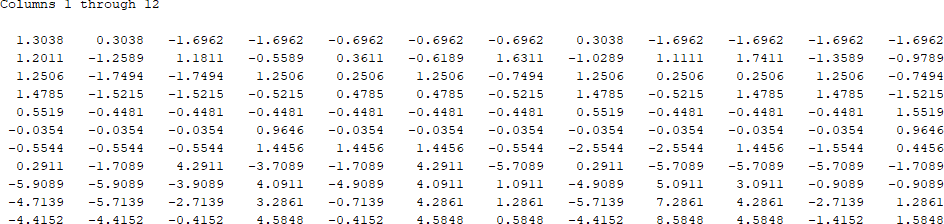
Number of outliers of ages: 1.000000 Number of outliers of GPAs: 0.000000 Number of outliers of Medu: 0.000000 Number of outliers of Fedu: 0.000000

Number of outliers of traveltime: 138.000000

Number of outliers of studytime: 197.000000 Number of outliers of health: 0.000000 Number of outliers of absences: 14.000000 Number of outliers of G1: 0.000000

Number of outliers of G2: 13.000000 Number of outliers of G3: 0.000000

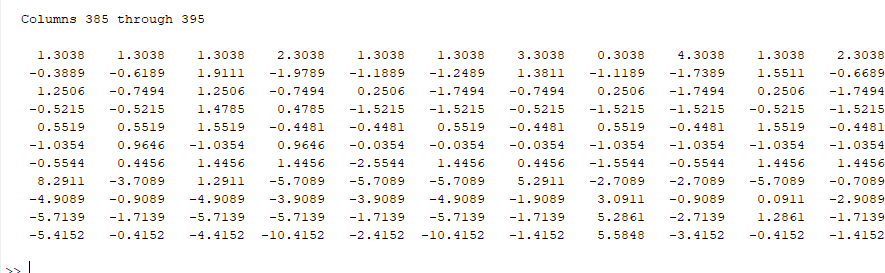
## Centered Data Matrix



**.**

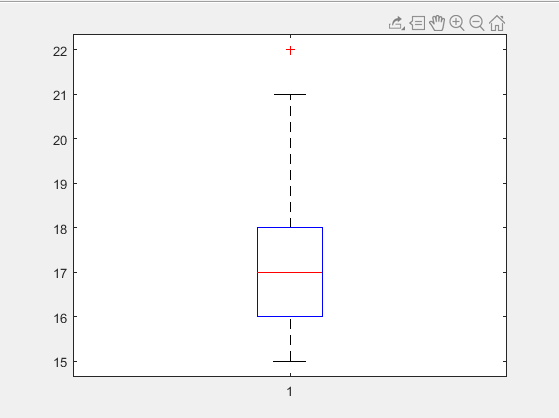
**.**

**.**

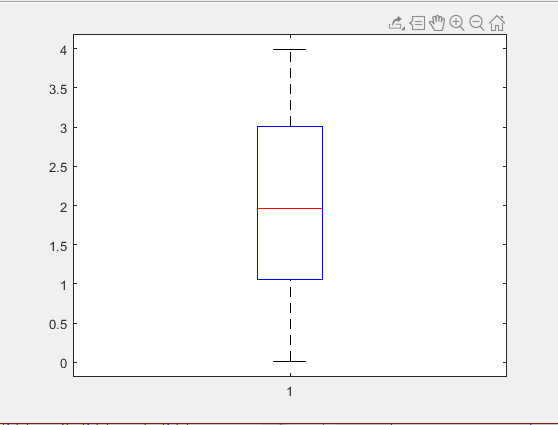


## Boxplot Graphs

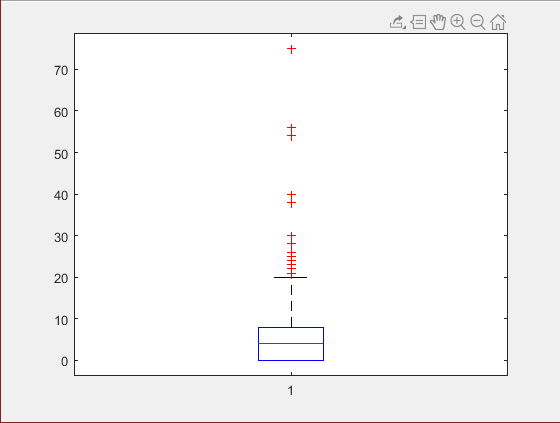
### Age



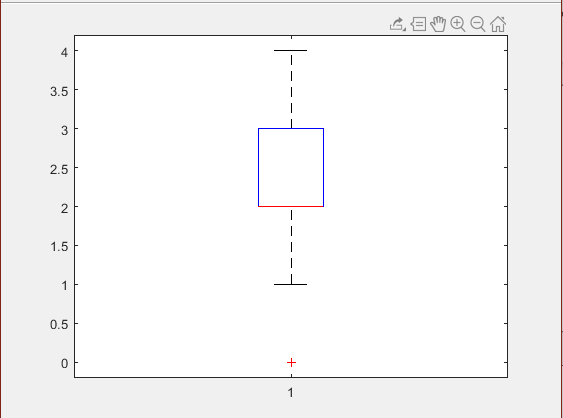
### GPA



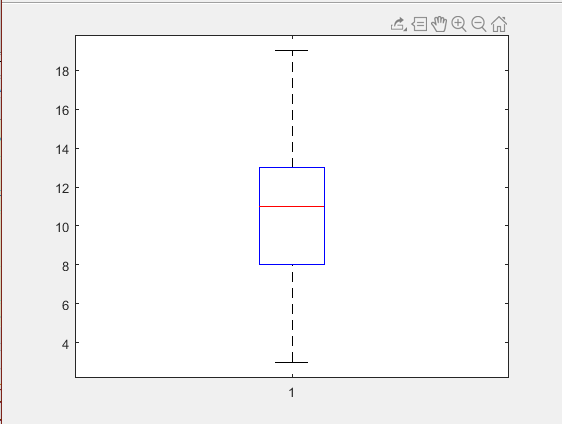
### Absences



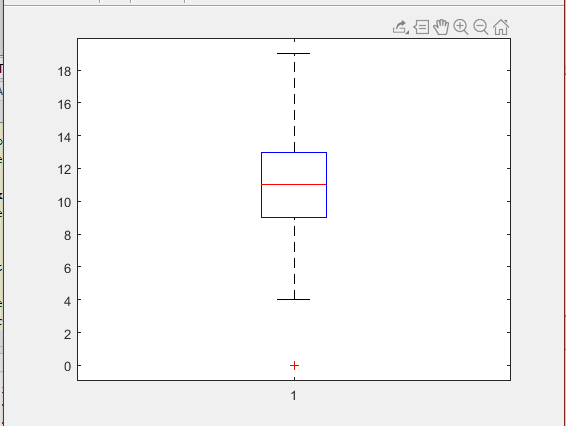
### Fedu



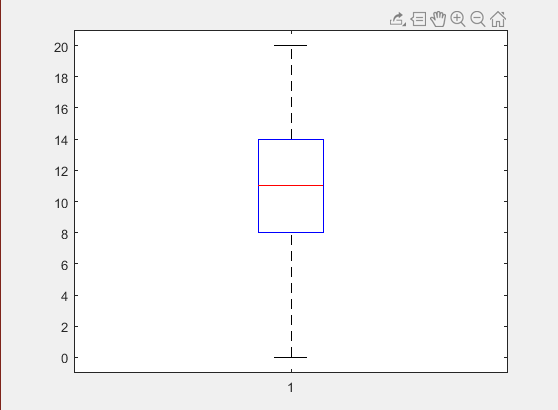
### G1



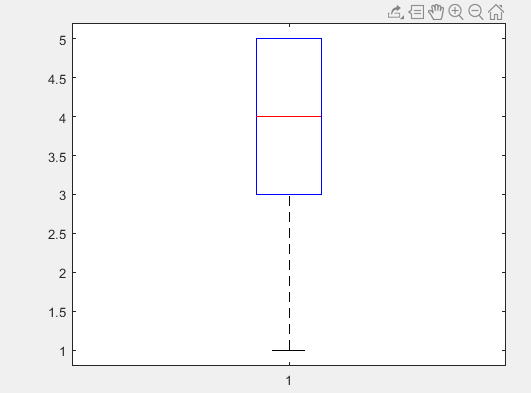
### G2



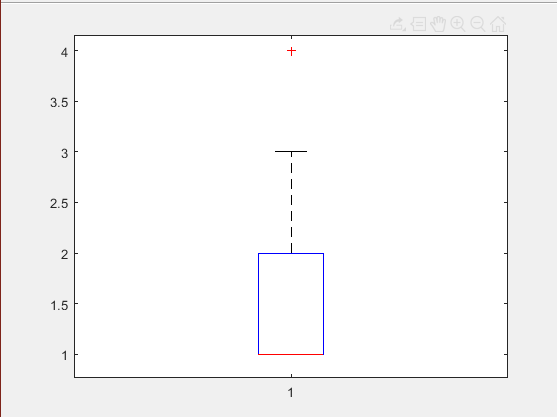
### G3



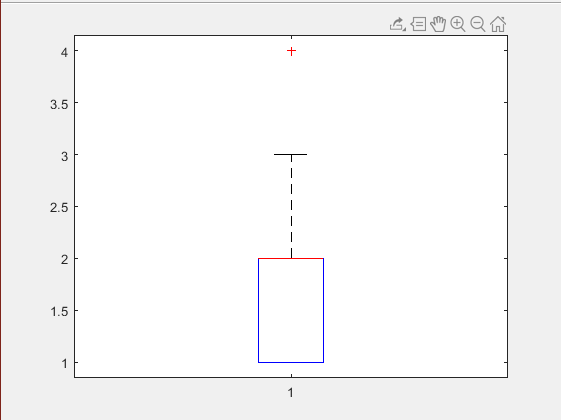
### Health



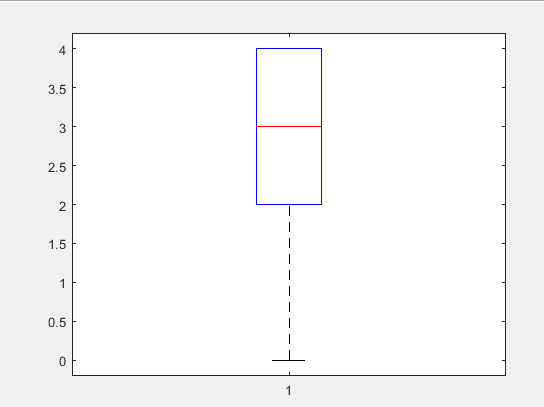
### Traveltime



### Studytime



### Medu

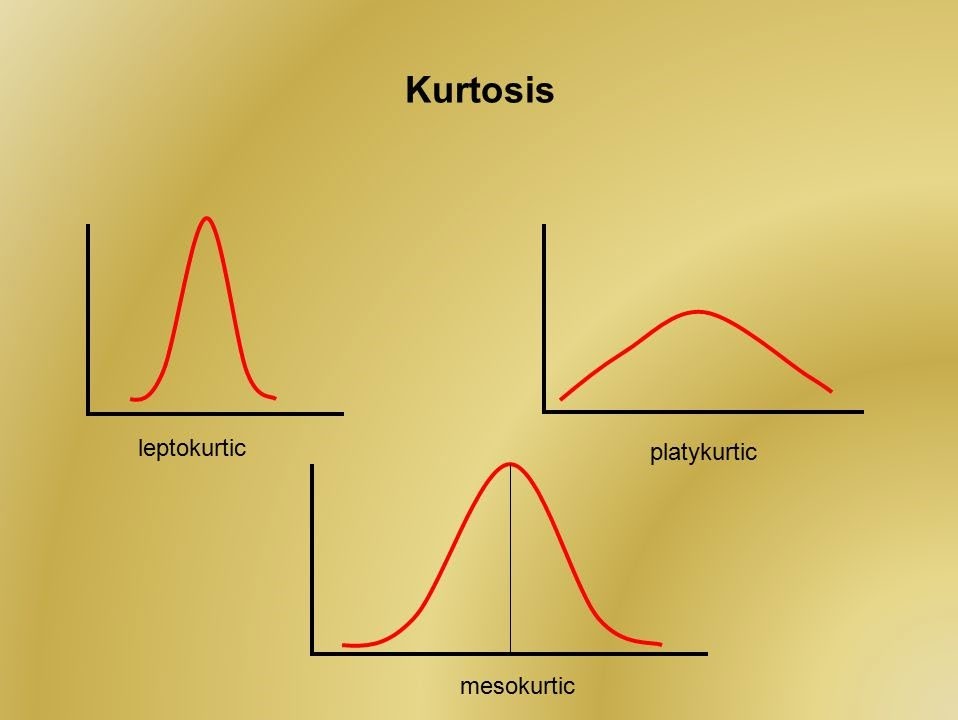


As an evaluating the result, let’s take kurtosis values as an example; in this data set kurtosis values mostly “Platykurtic” so outcomes have not extreme deviations from the mean. Having Platykurtic outcomes is not bad for the dataset. In boxplot outcomes, some values and also graphs are outliers but most of them close to mean. Most of the outcomes are in 3rd quartile or 1st quartile. Lastly Skewness outcomes are mostly skewed to the left. If test results are skewed left it means the average of the group was lower than the median of the group which means that either people got worse scores than expected or some people got lower scores than expected.

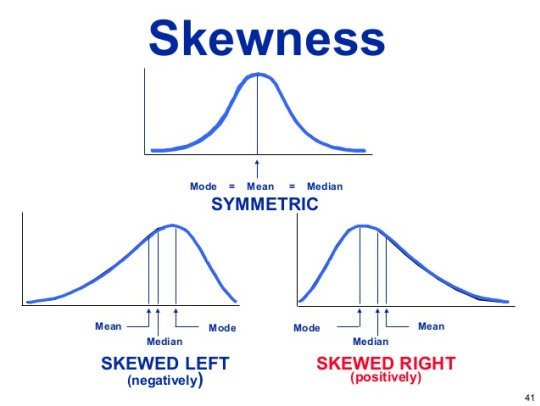
# Interpretation of the skewness and kurtosis values

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Mean** | **Median** | **Skewness** | **Kurtosis** | **Interpreting**  **Kurtosis** | **Interpreting**  **Skewness** |
| *Age* | 16.696203 | 17.000000 | 0.464498 | 2.983642 | Platykurtic | Skewed to the left |
| *GPA* | 2.028911 | 1.960000 | 0.009155 | 1.816286 | Platykurtic | Skewed to the right |
| *Absence* | 5.708861 | 4.000000 | 3.657622 | 24.430045 | Leptokurtic | Skewed to the right |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *FEDU* | 2.521519 | 2.000000 | -0.031552 | 1.801427 | Platykurtic | Skewed to the right |
| *G1* | 10.908861 | 11.000000 | 0.239699 | 2.299771 | Platykurtic | Skewed to the left |
| *G2* | 10.713924 | 11.000000 | -0.430004 | 3.604637 | Leptokurtic | Skewed to the left |
| *G3* | 10.415190 | 11.000000 | -0.729887 | 3.383181 | Leptokurtic | Skewed to the left |
| *Health* | 3.554430 | 4.000000 | -0.492723 | 1.983561 | Platykurtic | Skewed to the left |
| *MEDU* | 2.749367 | 3.000000 | -0.317170 | 1.908583 | Platykurtic | Skewed to the left |
| *Studytime* | 2.035443 | 2.000000 | 0.629739 | 2.970599 | Platykurtic | Skewed to the right |
| *Traveltime* | 1.448101 | 1.000000 | 1.600920 | 5.299470 | Leptokurtic | Skewed to the right |



**KURTOSİS**



**SKEWNESS**

# References

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

Documentation. (n.d.). Retrieved from https://[www.mathworks.com/help/.](http://www.mathworks.com/help/)