EAS506 - Statistical Data Mining I

Homework 1 – Question 1

Paul M Girdler

09/14/18

Abstract

This report summarizes the data preparation, cleaning and exploratory data analysis (EDA) process undertaken to prepare the data for linear modelling to predict *First Period Grades* (variables *G1.x*, and *G1.y*) from a dataset from **UCI machine learning repository**.

The final cleaned data set was saved as:
data.set named "g"
"cleaned_student_data.rdata"
27 variables x 380 students

Content

1 Introduction

The *Student Performance Dataset* is based upon two datasets of the academic performance of Portuguese students in two different classes: Math and Portuguese. The dataset is available on the **UCI machine learning repository**.

This report summarizes the data preparation, cleaning and exploratory data analysis (EDA) process undertaken to prepare the data for linear modelling to predict *First Period Grades* (variables *G1.x*, and *G1.y*)

2 Method

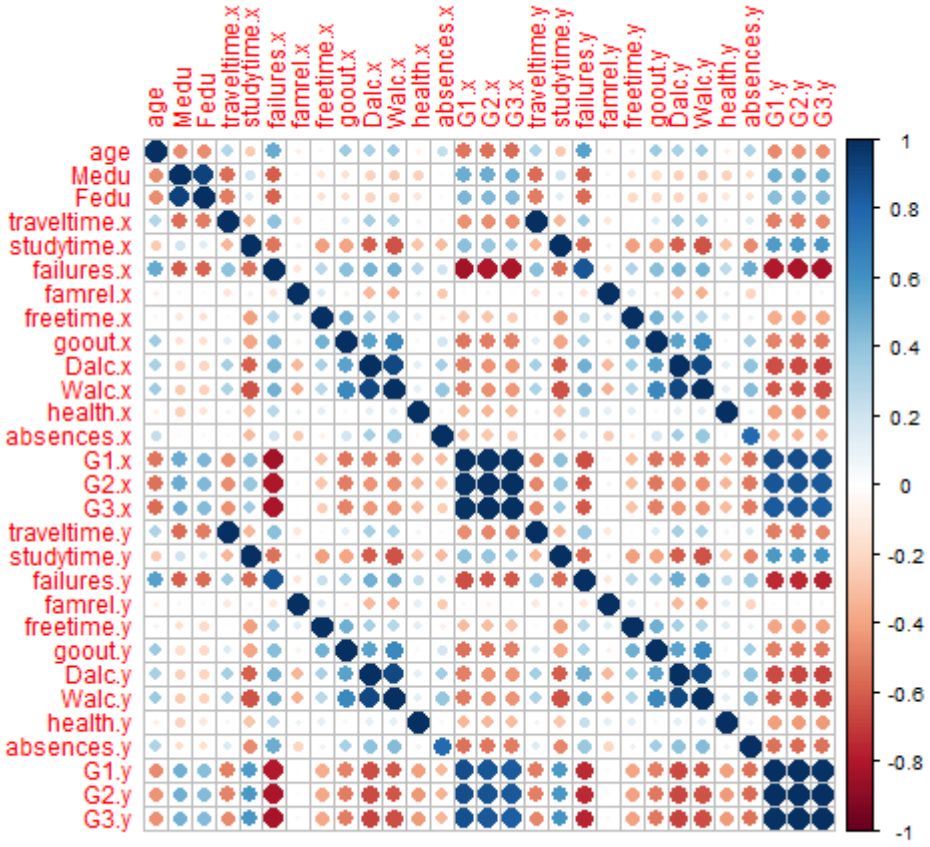
2.1 Initialization Steps

- Clear the memory
- Install and load all required libraries.
- Import and merge data.
- Briefly examine the data.

2.2 Examining for Redundant Data

- Create a correlation plot
- Examine for redundant data produced during the merge.
- Drop redundant data

2.2.1 Figure – Corrplot of Student Data



2.2.2 Duplicated columns(numeric):

From 2.2.1 Figure – Corrplot of Student Data it is evident the following were imported and duplicated in the merge:

traveltime, studytime, famrel, freetime, goout, Dalc, Walc, health

2.2.3 Duplicated columns(categorical):

Also by inspection the following cat columns(categorical) are duplicated:

guardian, schoolsup, famsup, paid, higher, romantic, activities

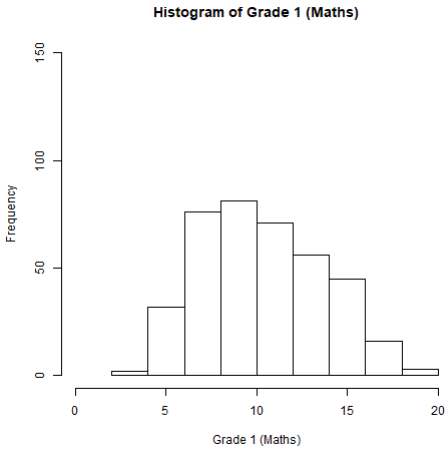
2.3 Descriptions for the Predictors

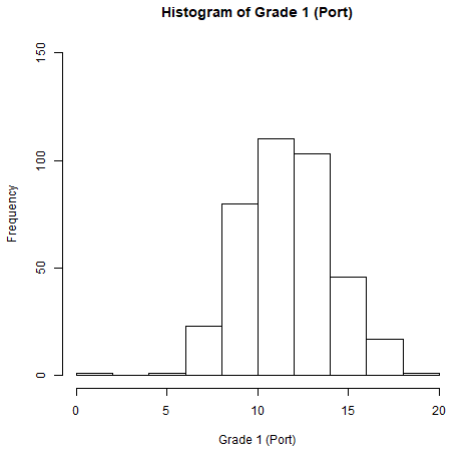
- For readability a reference vector was created with all the descriptions of the predictors.
- These vectors will be referenced in titles and labels in all plots. This will enhance readability and should make creating plots more streamlined.

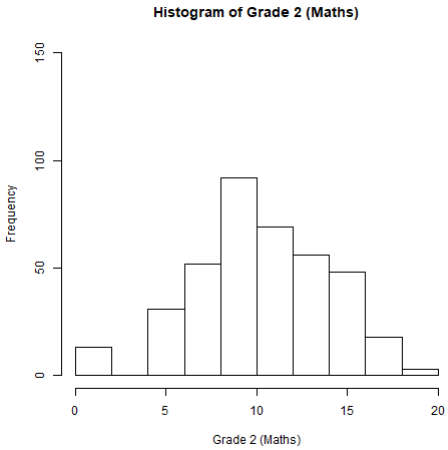
2.4 Examine all *Grade* variables

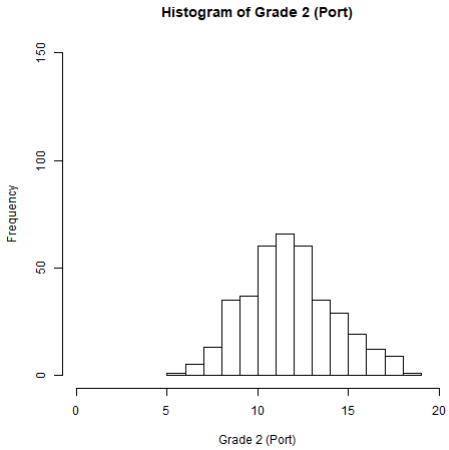
- Histograms wer plotted for all Grade variables
- Boxplots were plotted for *First Period Grades* (variables *G1.x*, and *G1.y*)
- Possible outliers were noted as were general trends in both cases.

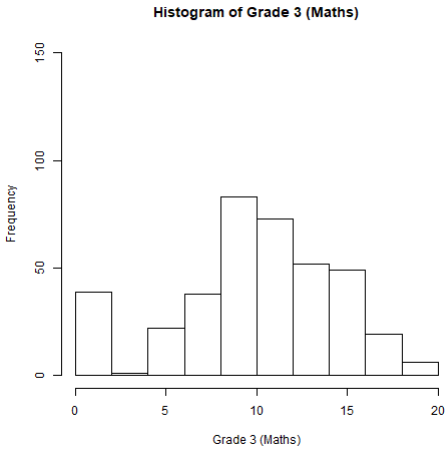
2.4.1 Histogram Plots of *Grade* (G1, G2, G3) Variables

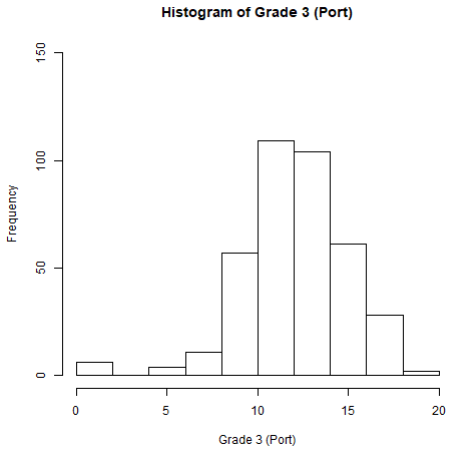




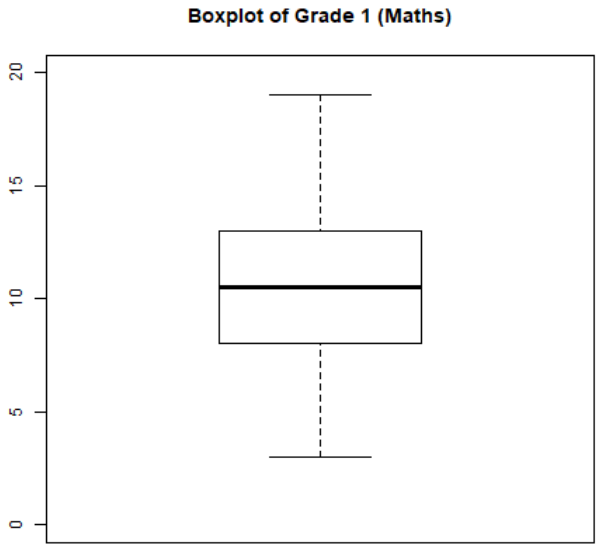


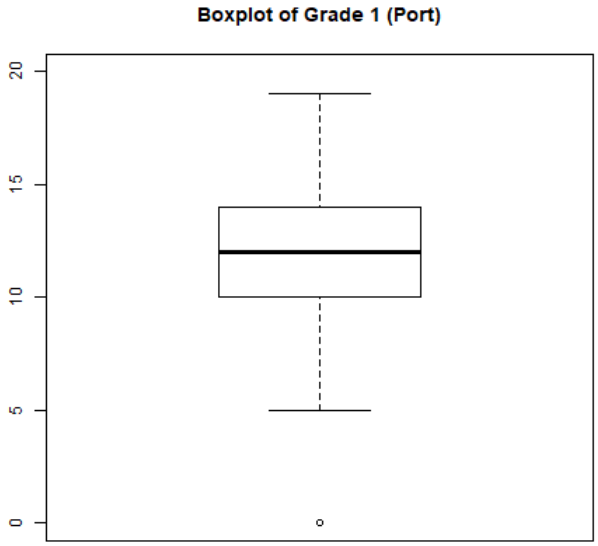






2.4.2 Boxplots of First Period Grade (*G1.x* and *G1.y*) Variables





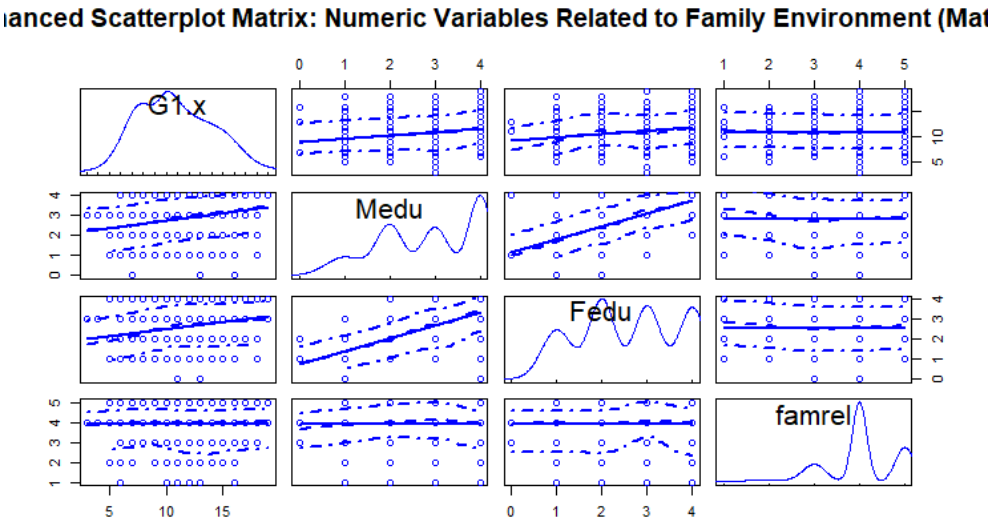
- G1 generally appears to be normally distributed.
- Outliers appear to be for *Grades* (*G1.x* and *G1.y* less than or equal to 5).
- A subset was created with these 2 row entries removed.

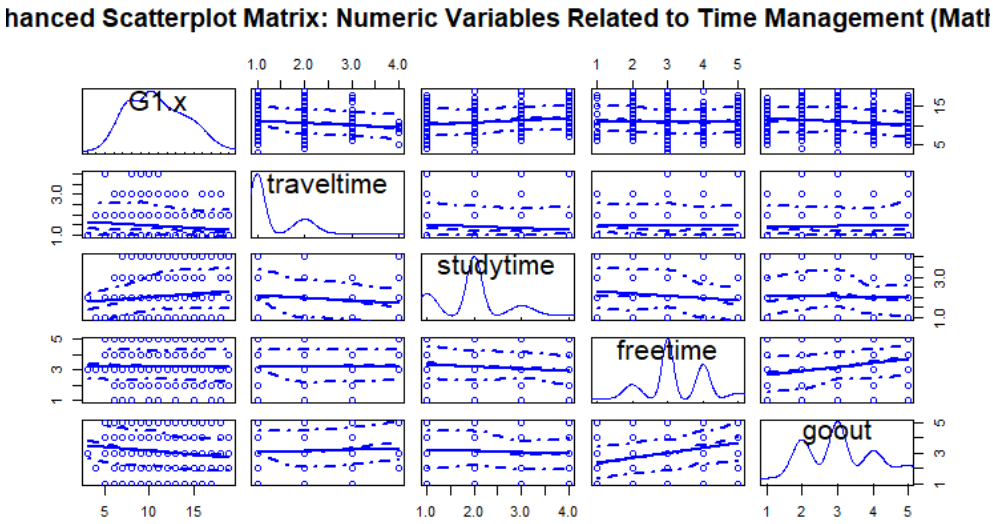
3 Results

3.1 Enhanced Scatterplots Matrix of Numerical Variables

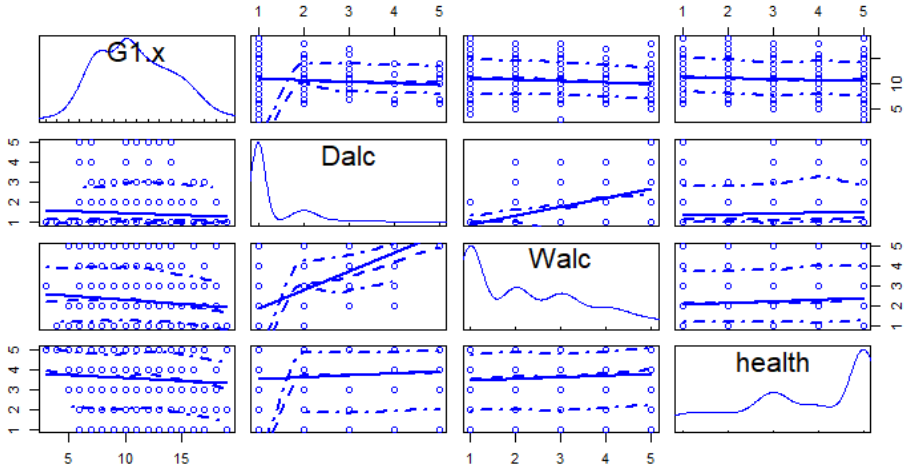
A data subset was then created of all the numeric variables. These were grouped broudly into the following feature families: *Family Environment, Time Management, Health,* and *Misc*. This commonsense approach was undertaken because there were too many variables to plot neatly on a single plot.

3.1.1 Scatterplots

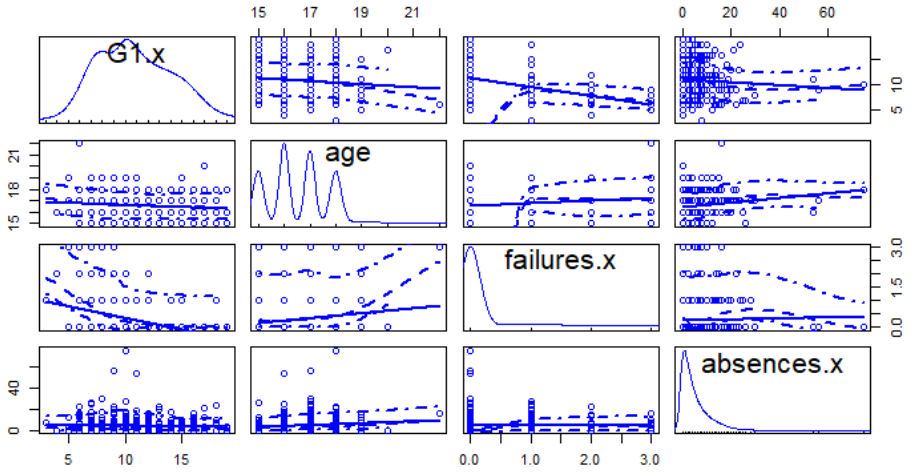




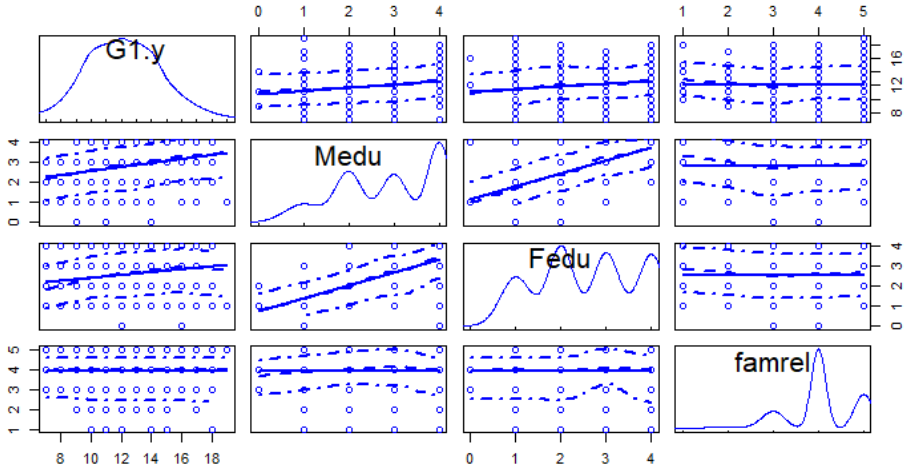
Enhanced Scatterplot Matrix: Numeric Variables Related to Health (Maths)



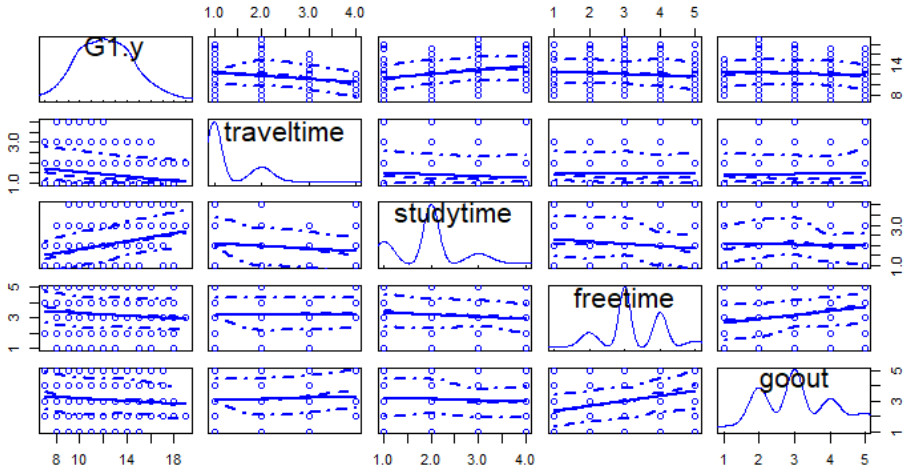
Enhanced Scatterplot Matrix: Misc Numeric Variables (Maths)

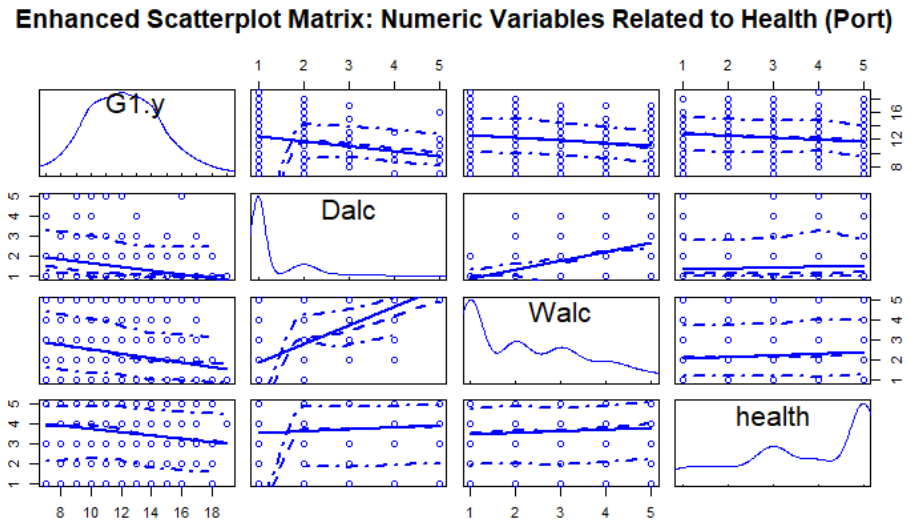


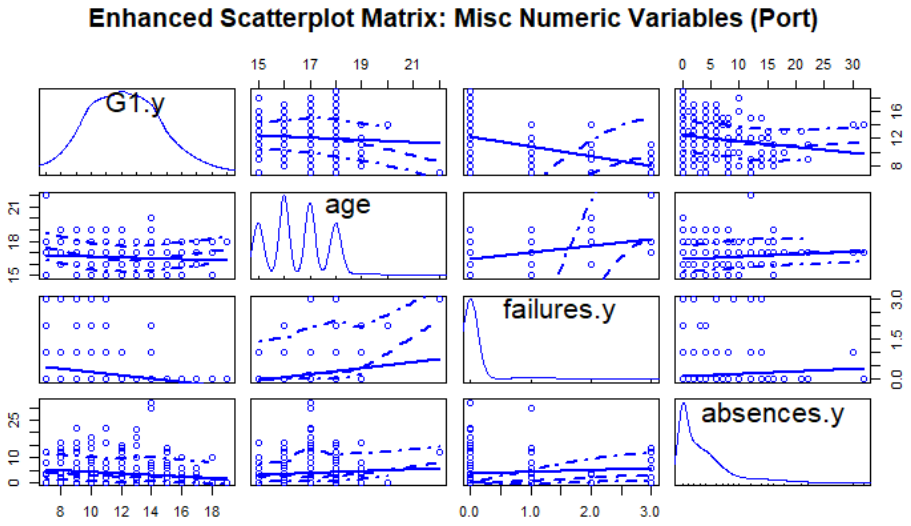
hanced Scatterplot Matrix: Numeric Variables Related to Family Environment (Po



ıhanced Scatterplot Matrix: Numeric Variables Related to Time Management (Poı





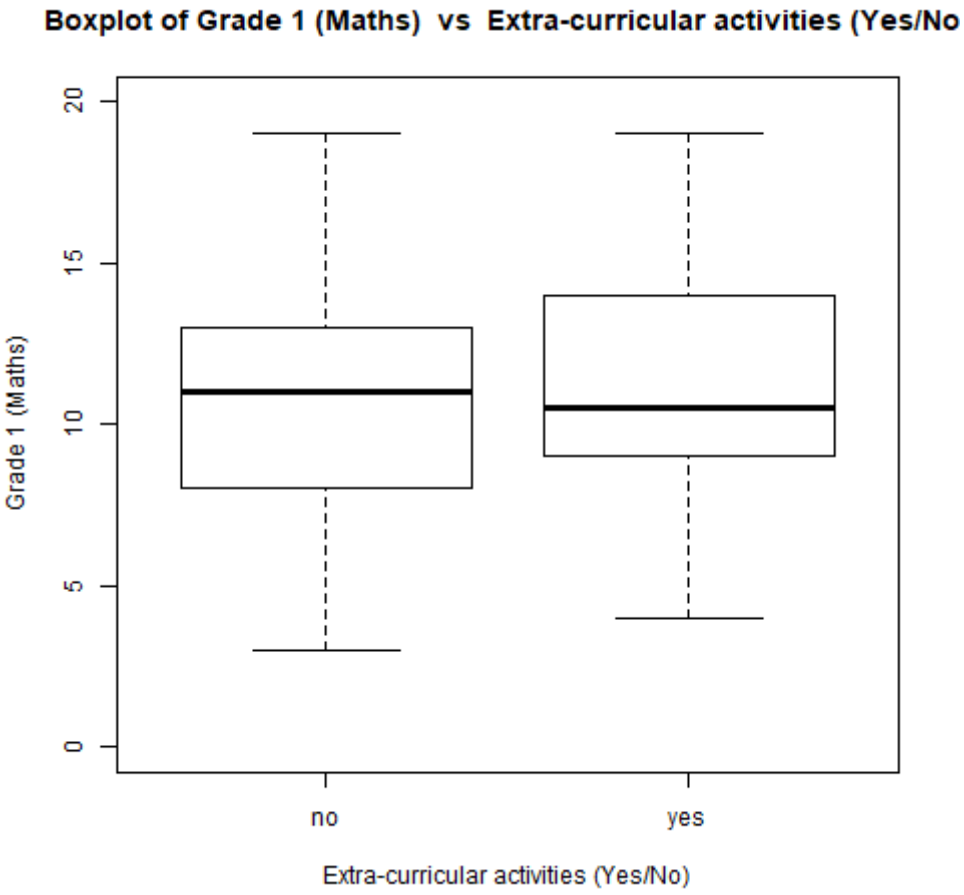


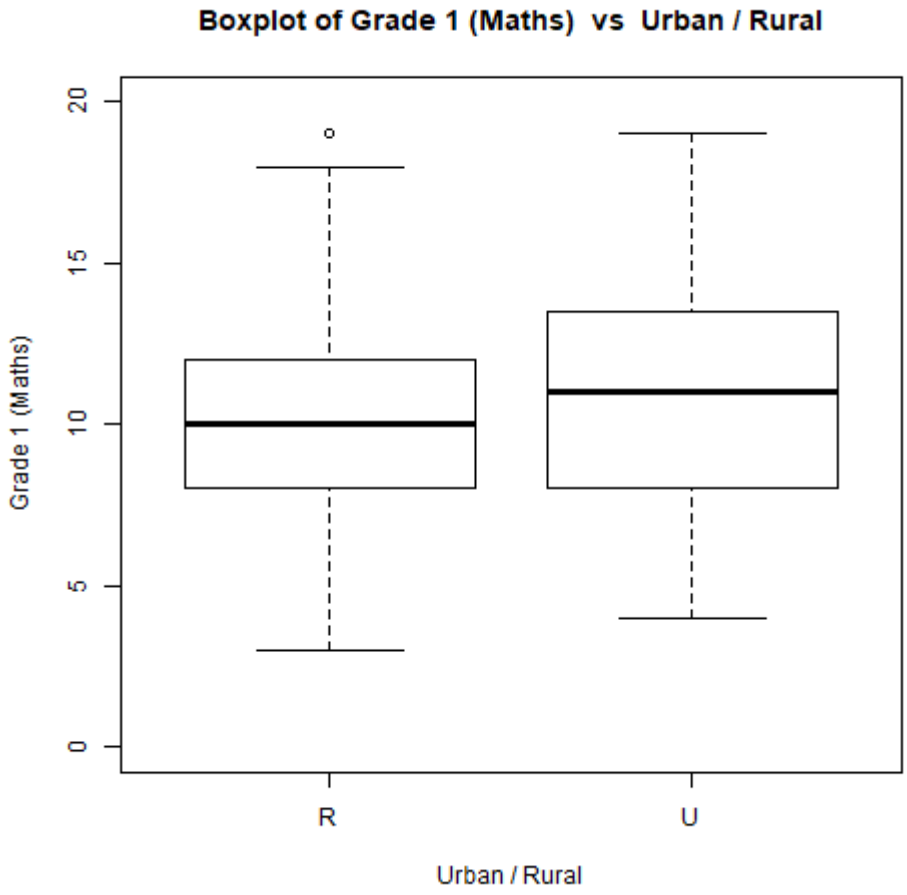
3.1.2 Dropped Numerical Variables

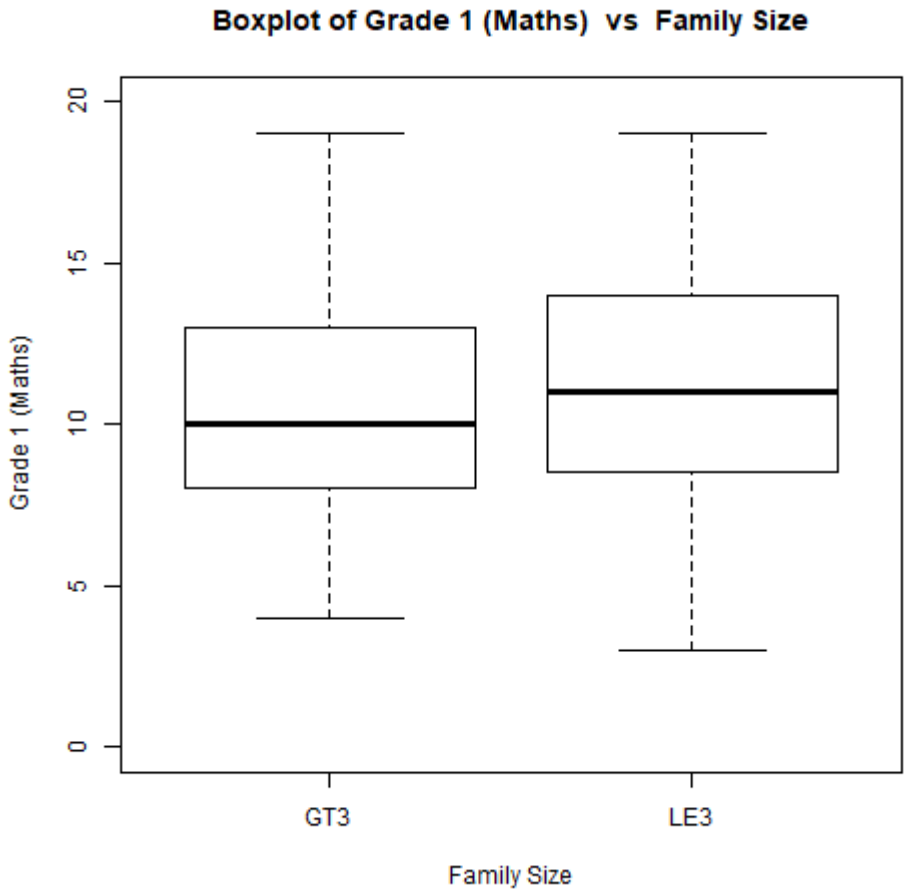
After examining the plots above I made the decision to drop the following variables due to insignificance or redundancy: *Fedu, famrel, freetime, goout, and Walc*. See Discussion for more details. Variables have now been reduced from the original 53 down to 33.

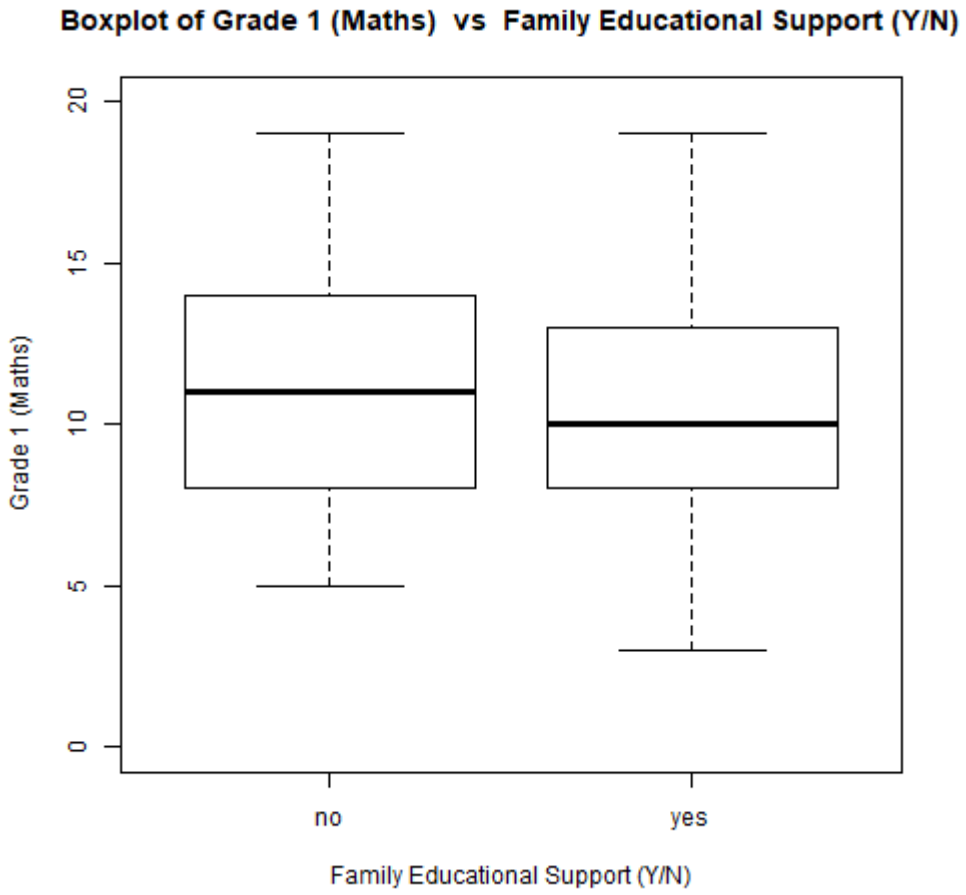
3.2 Boxplots of Categorical Features

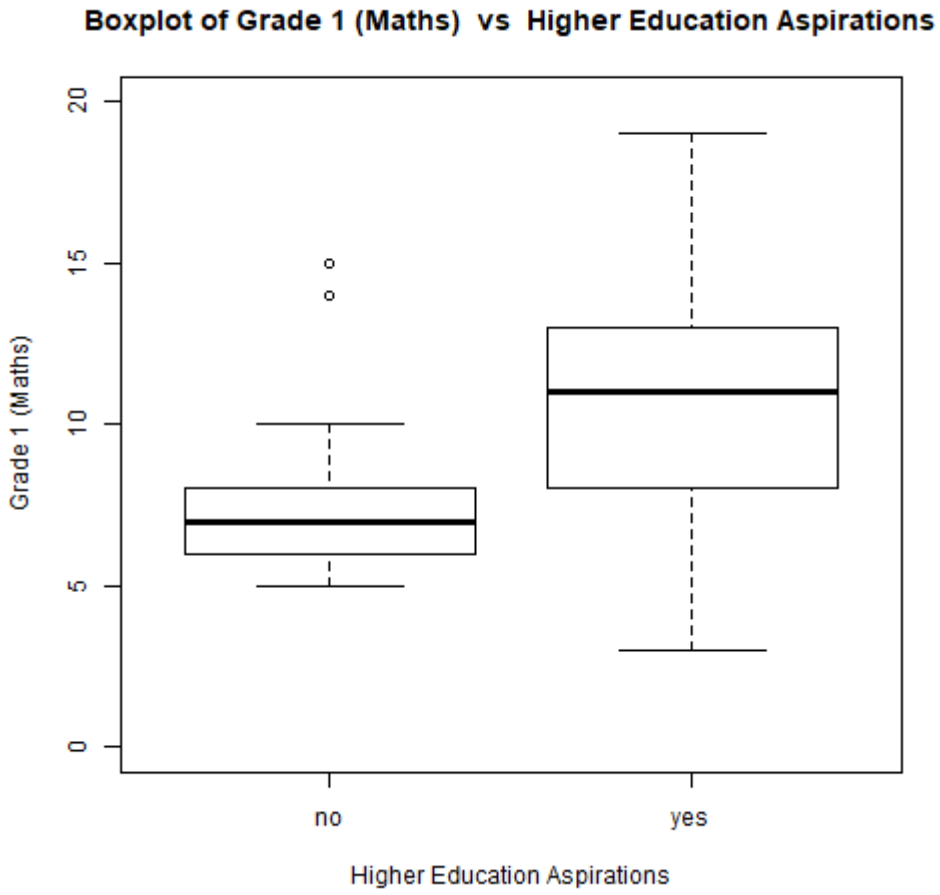
Boxplots were made comparing the predictor feature with response variable First Period Grade (G1).

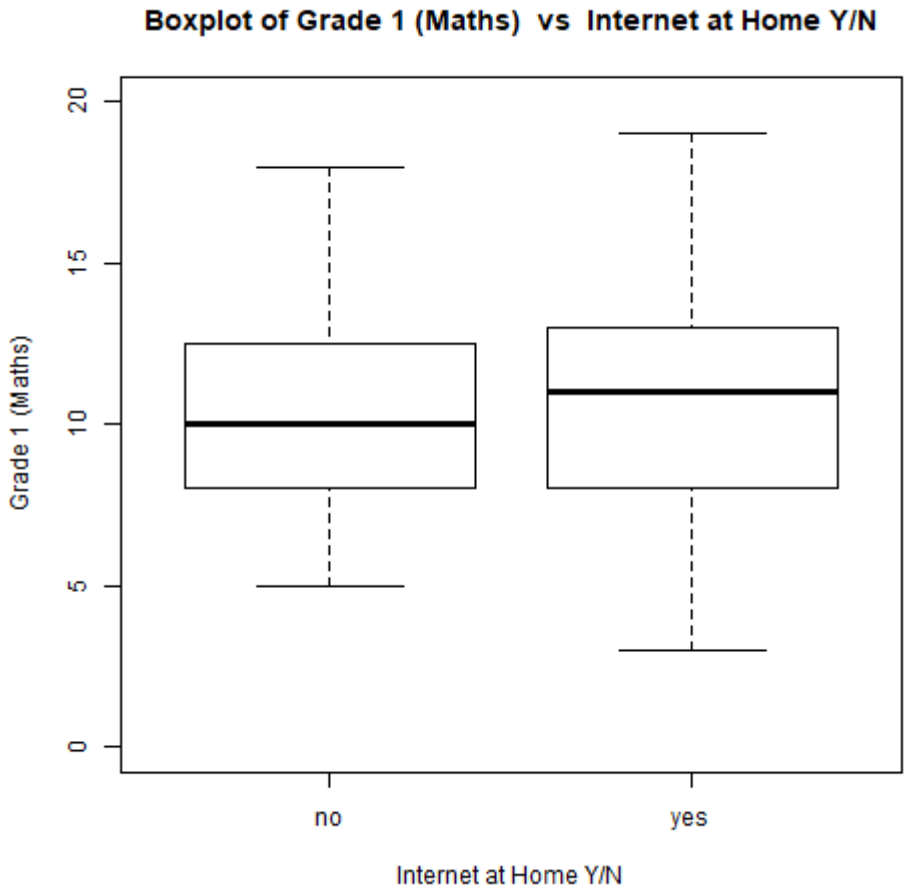

Boxplot of Grade 1 (Maths) vs Extra-curricular activities (Yes/No
Grade 1 (Maths)
0
5
10
15
20
no
yes
Extra-curricular activities (Yes/No)

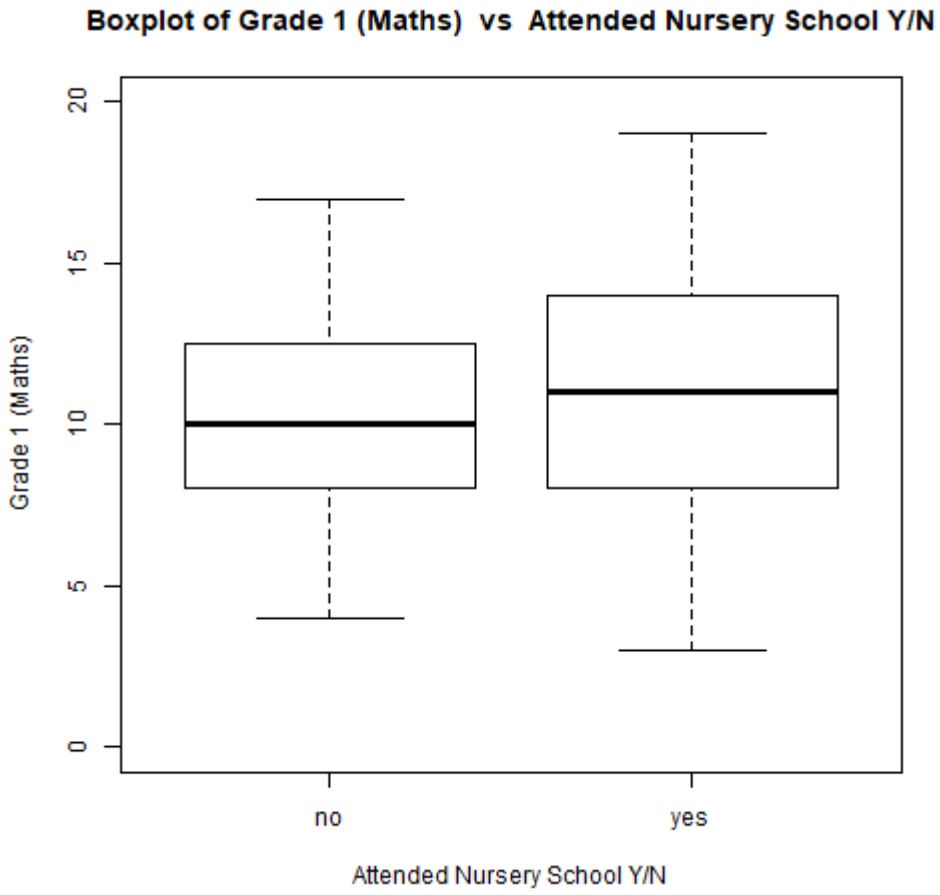

Boxplot of Grade 1 (Maths) vs Urban / Rural
Grade 1 (Maths)
0
5
10
15
20
R
U
Urban / Rural

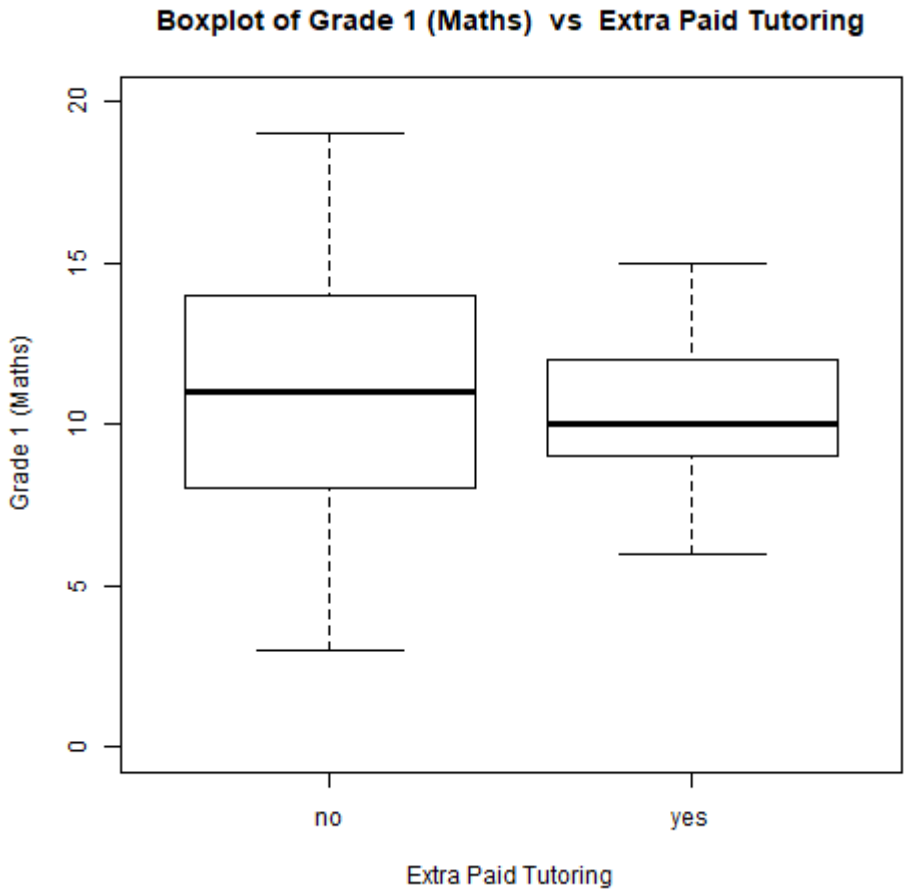

Boxplot of Grade 1 (Maths) vs Family Size
Grade 1 (Maths)
0
5
10
15
20
GT3
LE3
Family Size

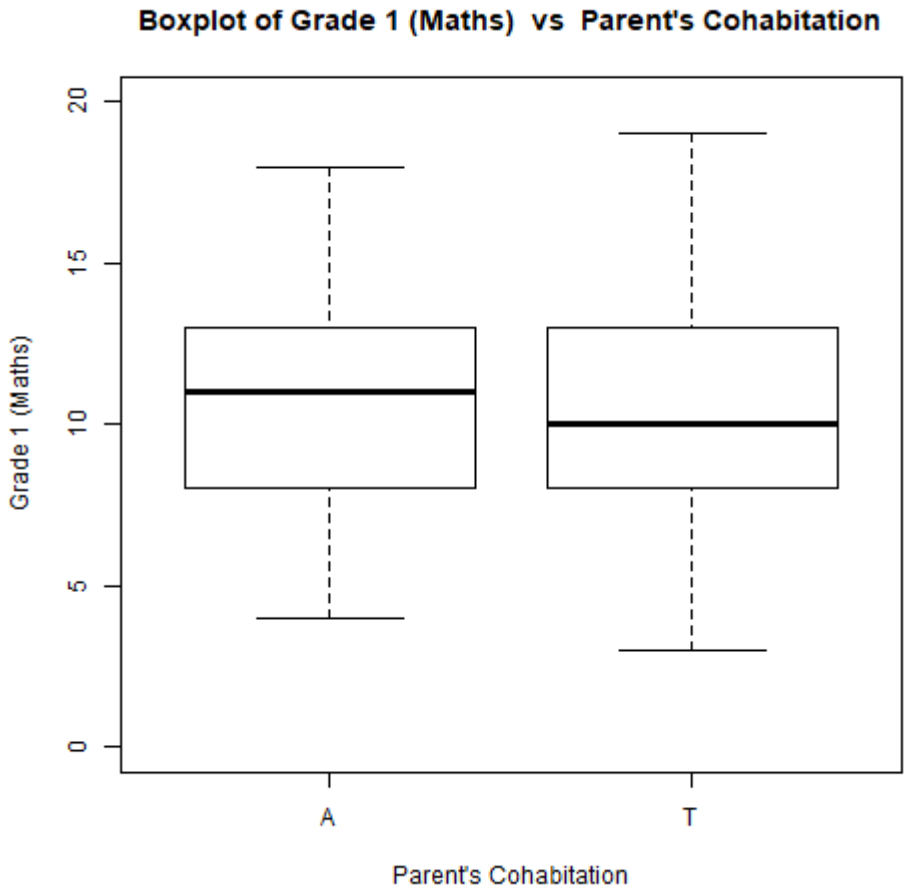

Boxplot of Grade 1 (Maths) vs Family Educational Support (Y/N)
Grade 1 (Maths)
0
5
10
15
20
no
yes
Family Educational Support (Y/N)

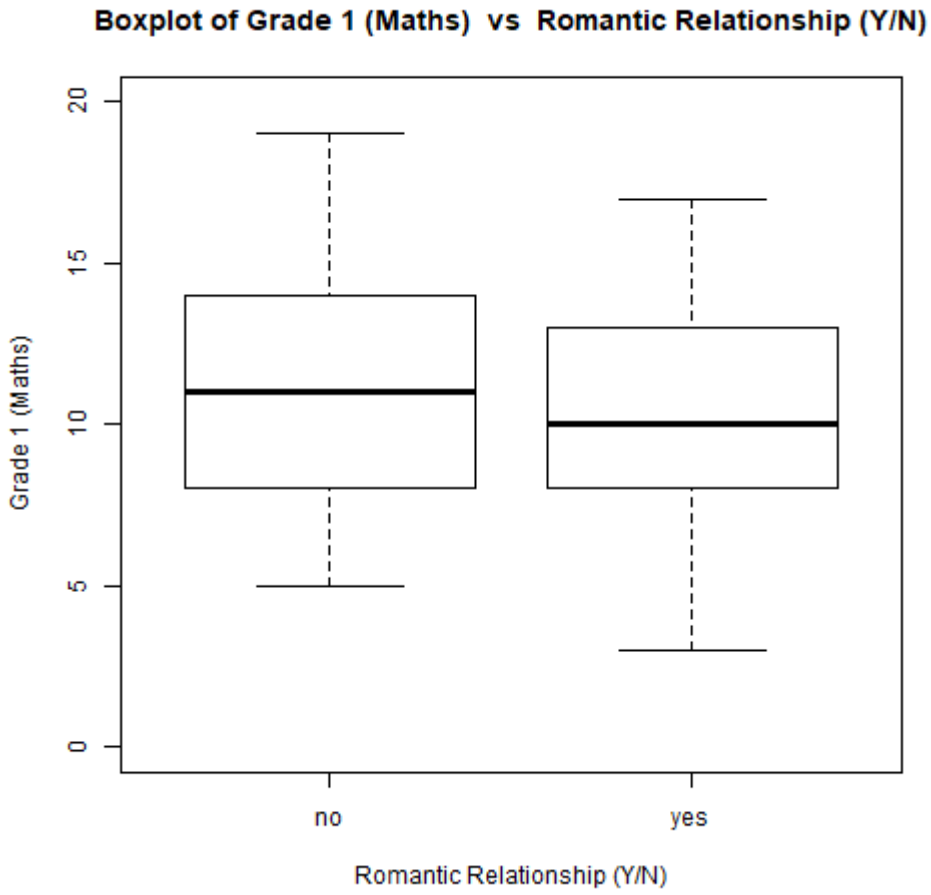

Boxplot of Grade 1 (Maths) vs Higher Education Aspirations
Grade 1 (Maths)
0
5
10
15
20
no
yes
Higher Education Aspirations

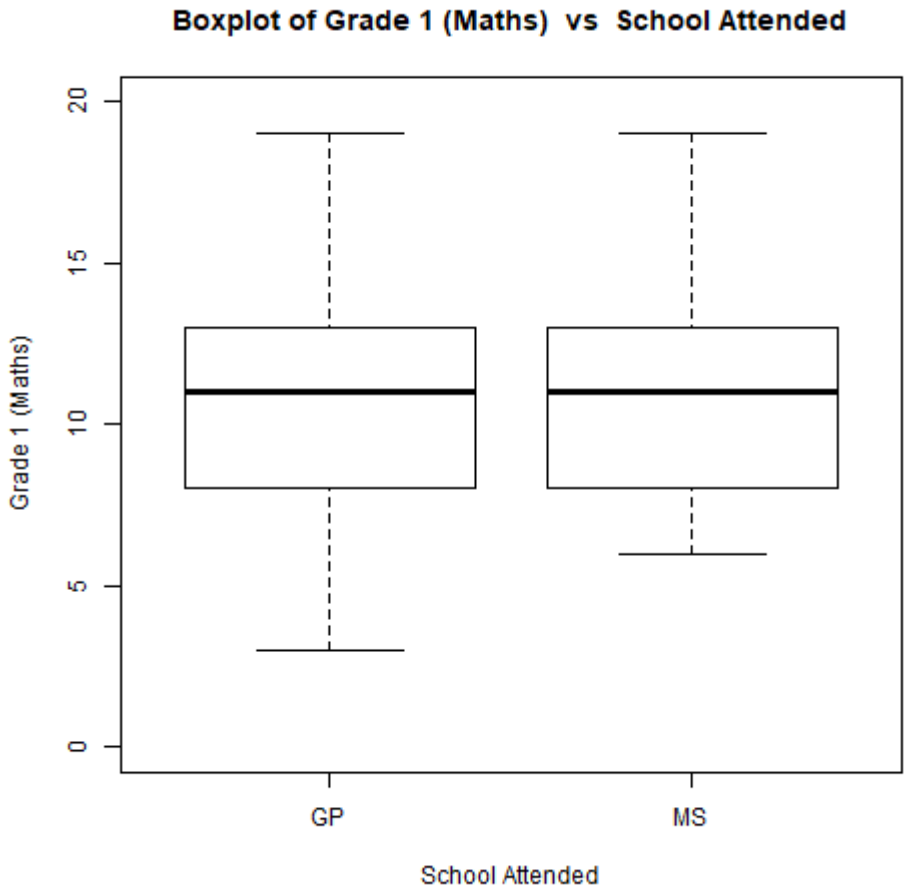

Boxplot of Grade 1 (Maths) vs Internet at Home Y/N
Grade 1 (Maths)
0
5
10
15
20
no
yes
Internet at Home Y/N

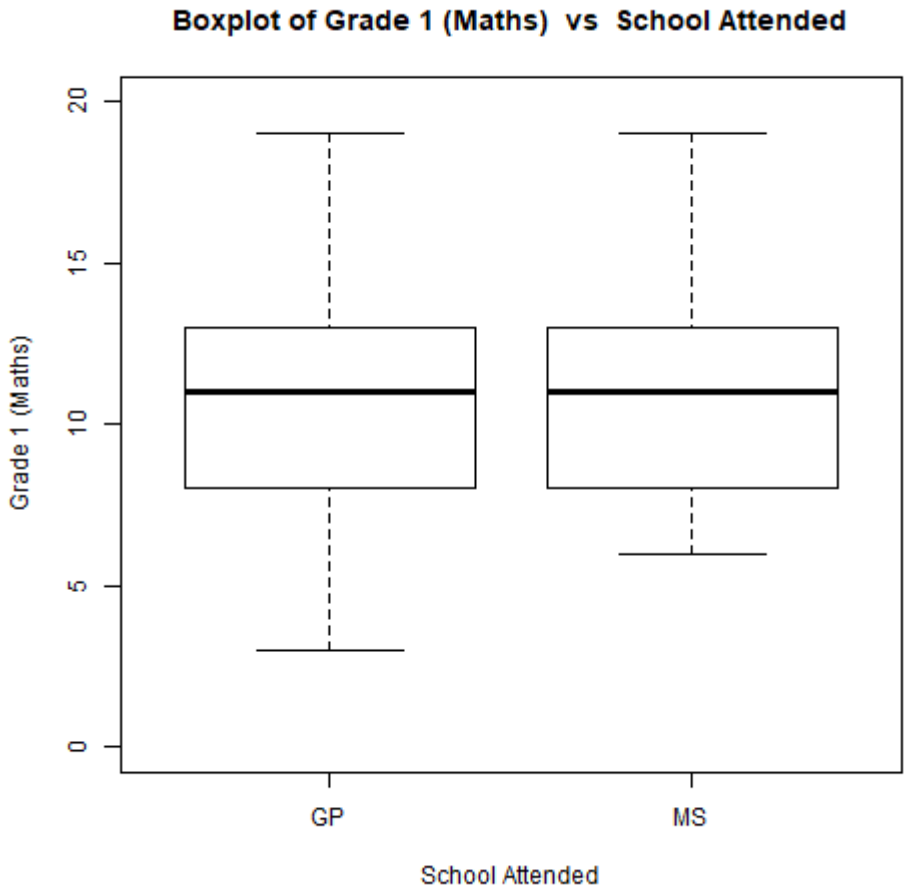

Boxplot of Grade 1 (Maths) vs Attended Nursery School Y/N
Grade 1 (Maths)
0
5
10
15
20
no
yes
Attended Nursery School Y/N

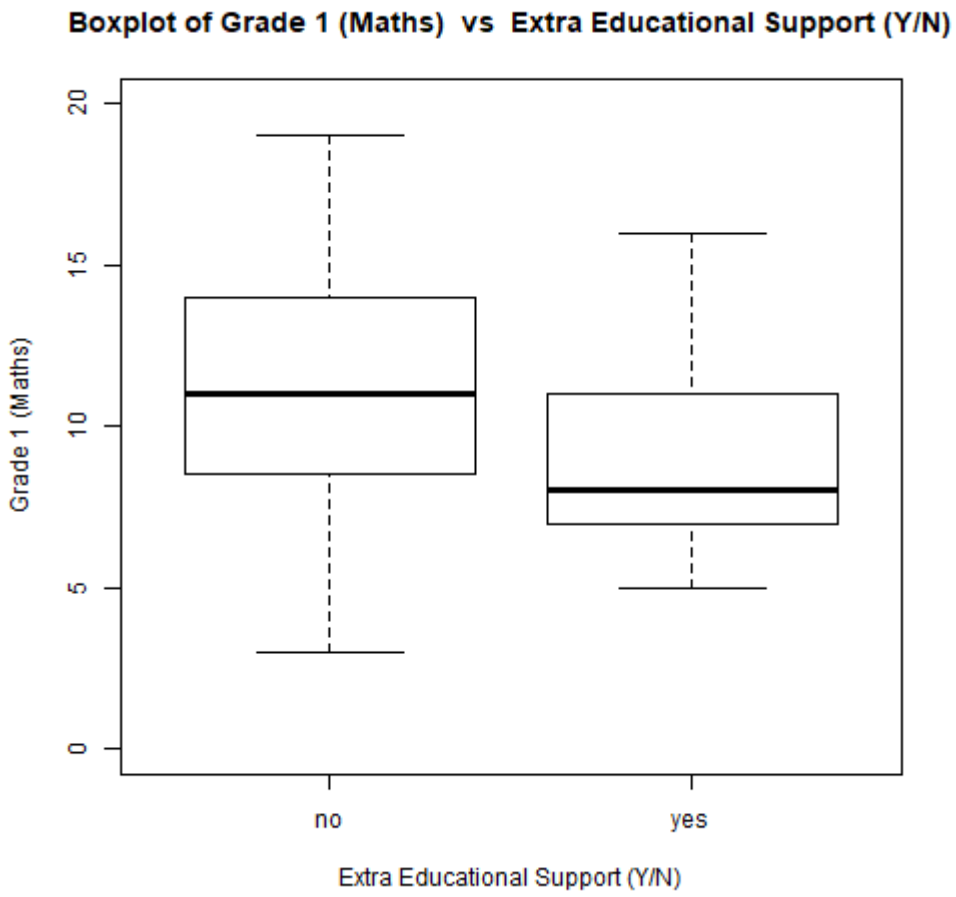

Boxplot of Grade 1 (Maths) vs Extra Paid Tutoring
Grade 1 (Maths)
0
5
10
15
20
no
yes
Extra Paid Tutoring

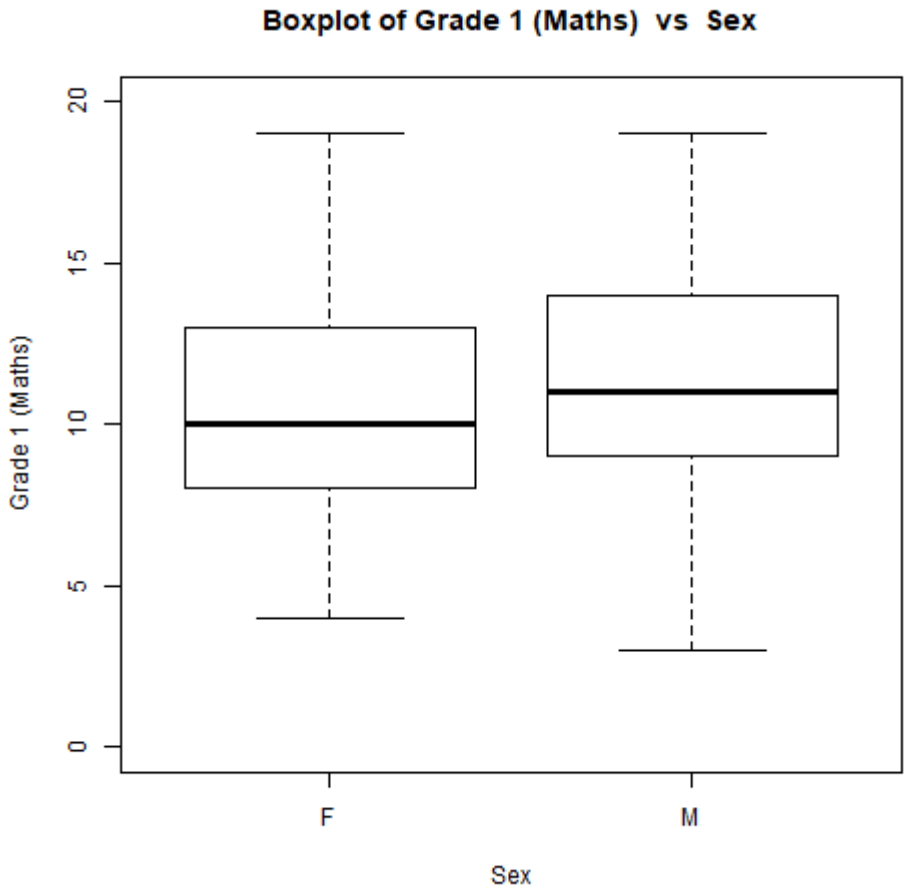

Boxplot of Grade 1 (Maths) vs Parent's Cohabitation
Grade 1 (Maths)
0
5
10
15
20
A
T
Parent's Cohabitation

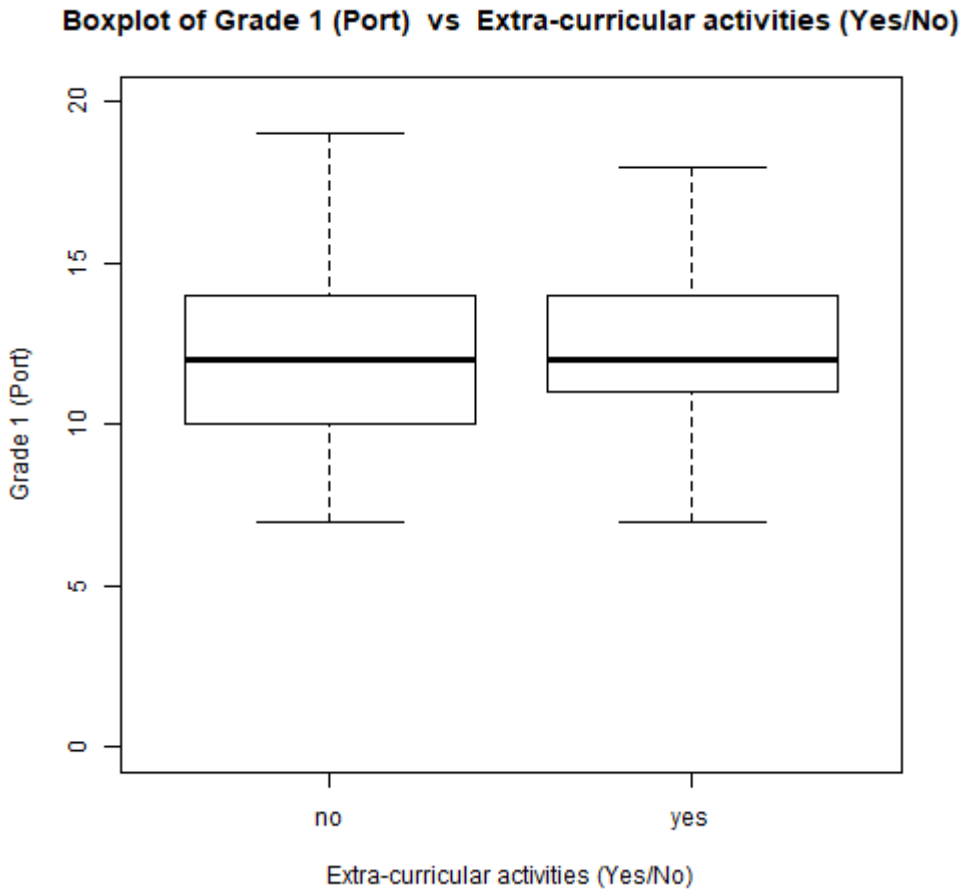

Boxplot of Grade 1 (Maths) vs Romantic Relationship (Y/N)
Grade 1 (Maths)
0
5
10
15
20
no
yes
Romantic Relationship (Y/N)

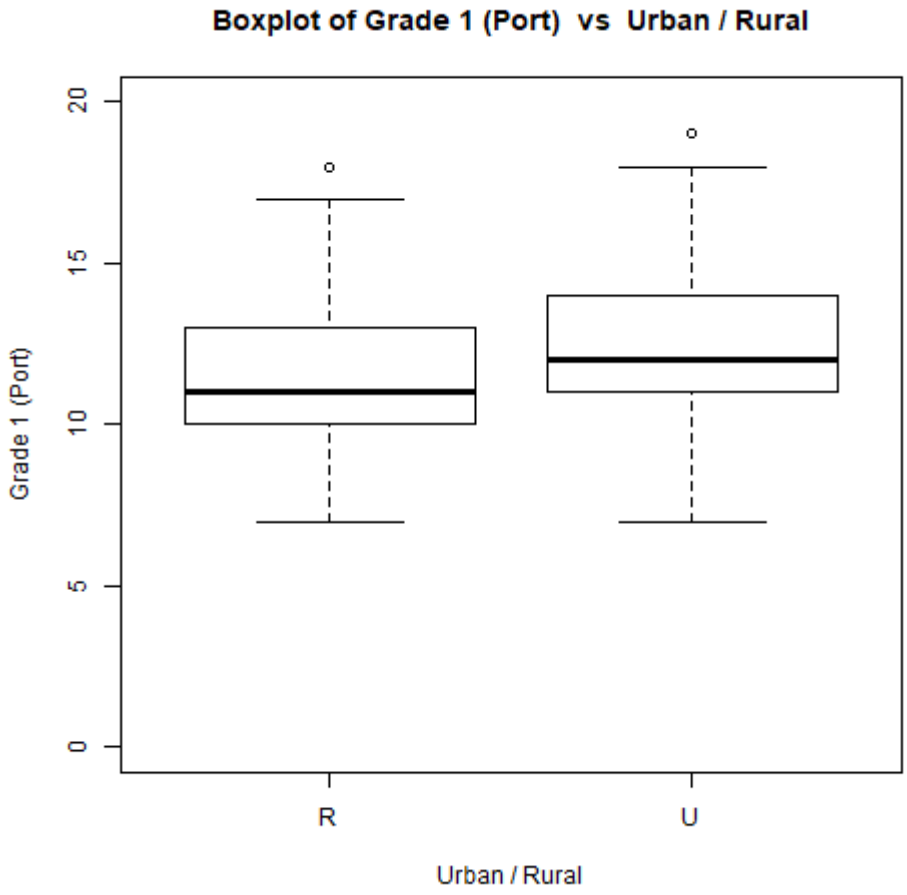

Boxplot of Grade 1 (Maths) vs School Attended
Grade 1 (Maths)
0
5
10
15
20
GP
MS
School Attended

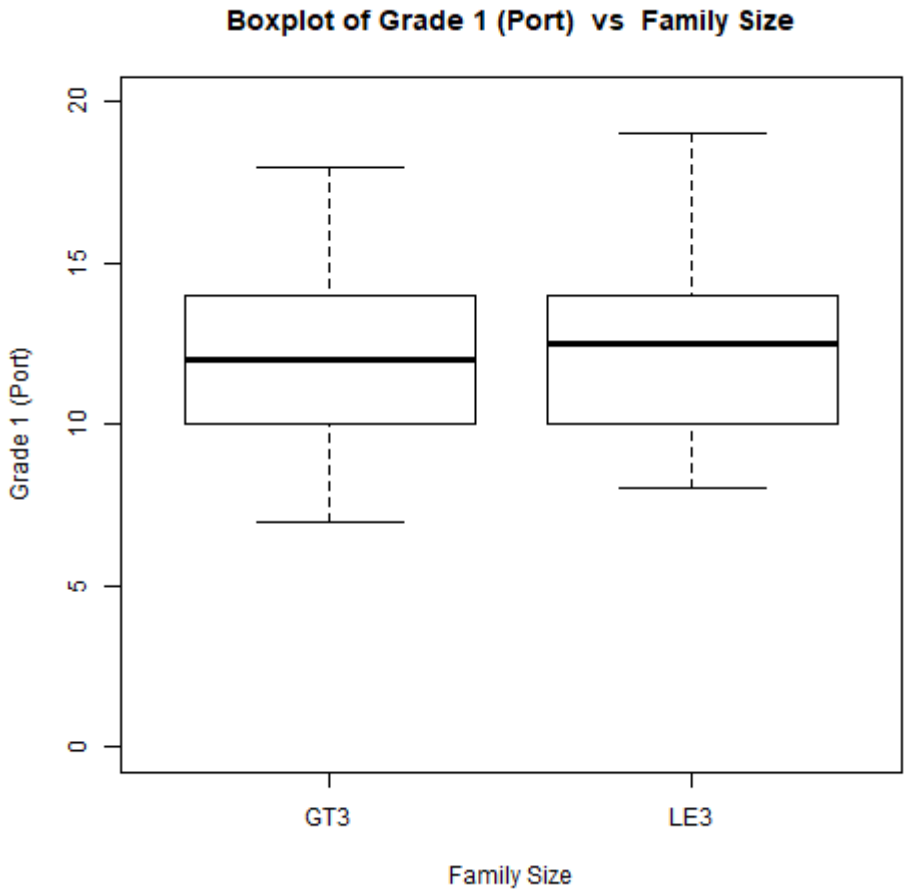

Boxplot of Grade 1 (Maths) vs School Attended
Grade 1 (Maths)
0
5
10
15
20
GP
MS
School Attended

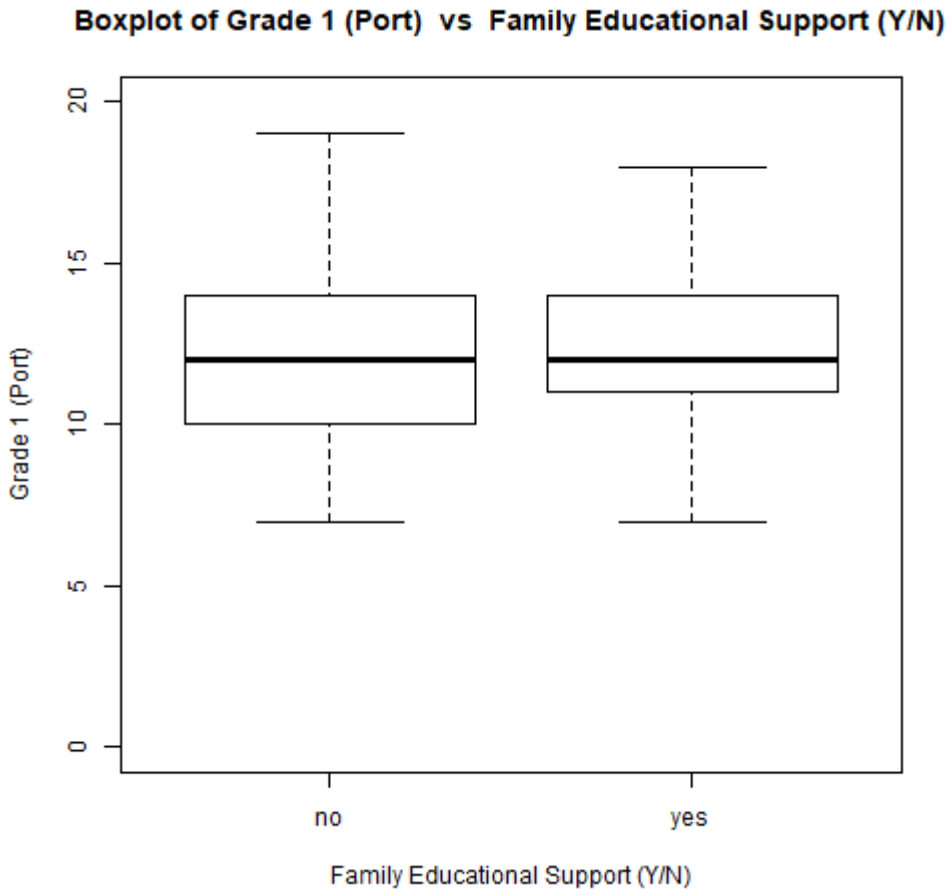

Boxplot of Grade 1 (Maths) vs Extra Educational Support (Y/N)
Grade 1 (Maths)
0
5
10
15
20
no
yes
Extra Educational Support (Y/N)

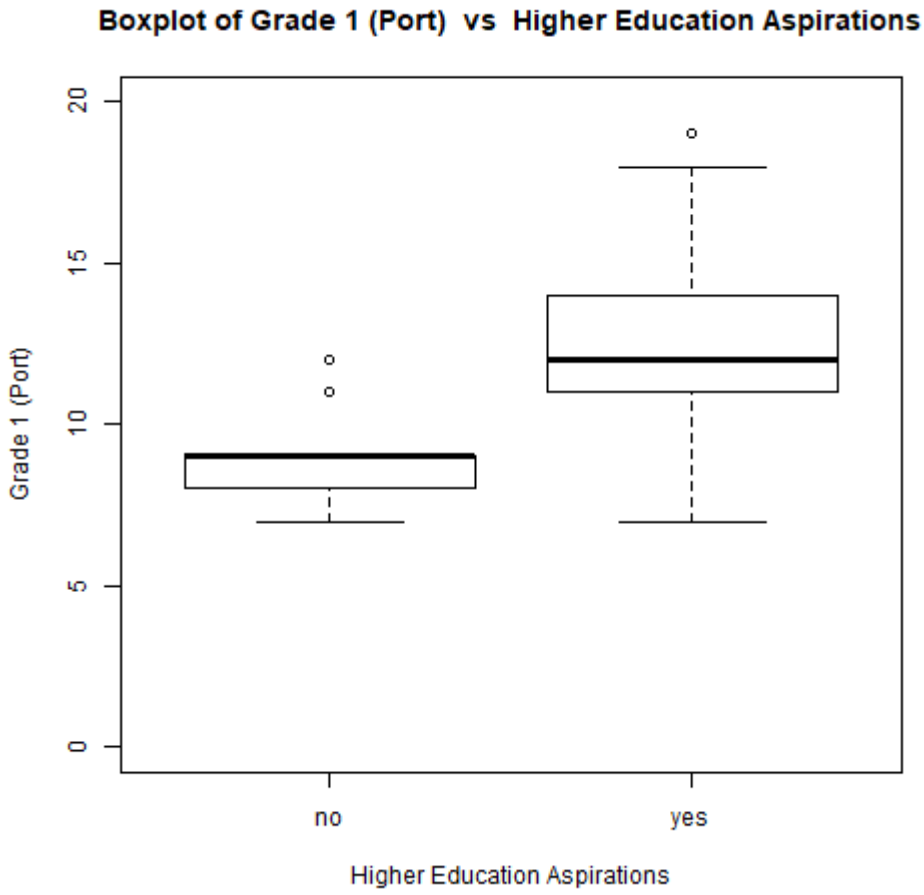

Boxplot of Grade 1 (Maths) vs Sex
Grade 1 (Maths)
0
5
10
15
20
F
M
Sex

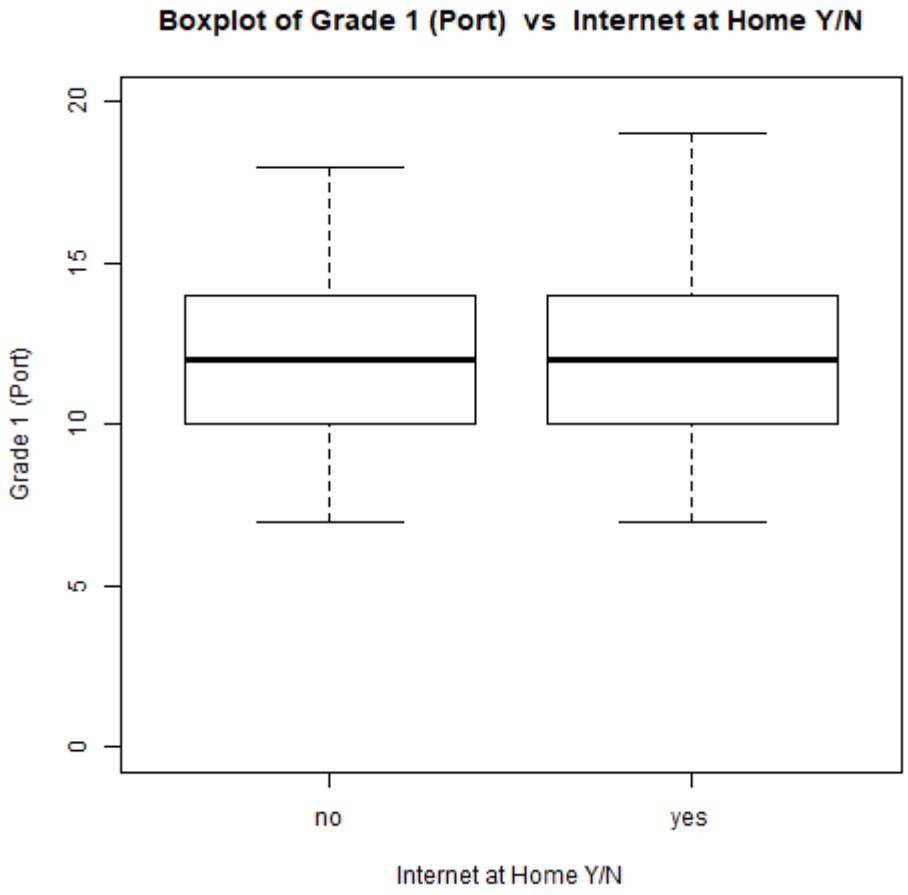

Boxplot of Grade 1 (Port) vs Extra-curricular activities (Yes/No)
Grade 1 (Port)
0
5
10
15
20
no
yes
Extra-curricular activities (Yes/No)

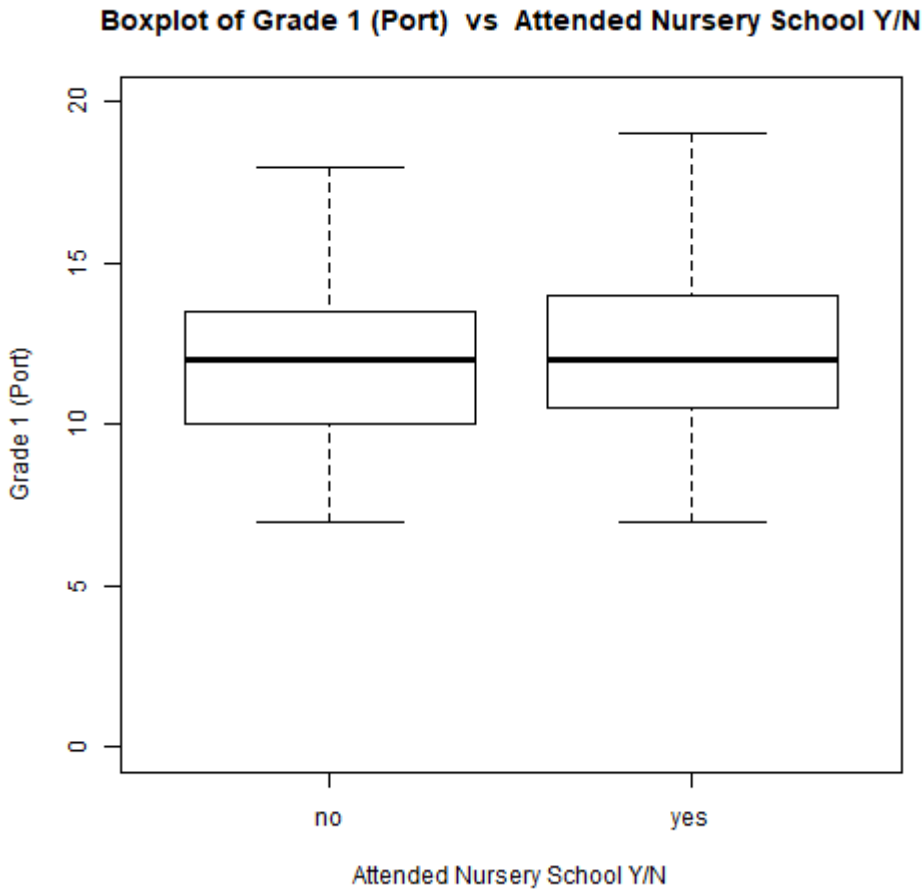

Boxplot of Grade 1 (Port) vs Urban / Rural
Grade 1 (Port)
0
5
10
15
20
R
U
Urban / Rural

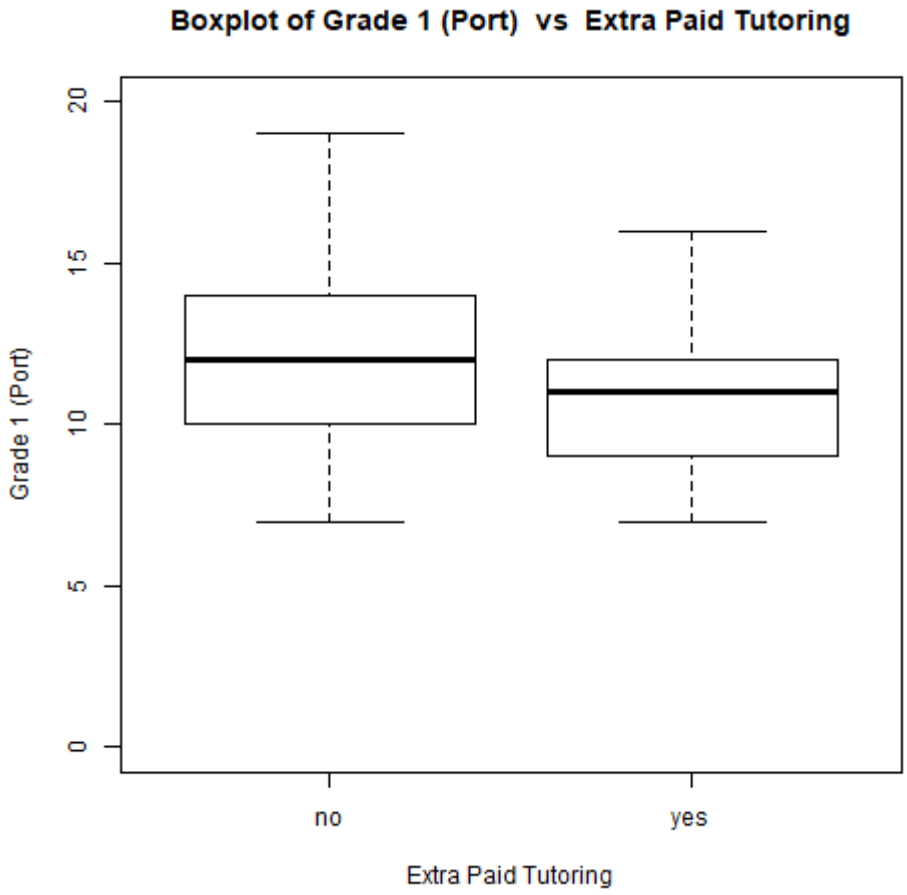

Boxplot of Grade 1 (Port) vs Family Size
Grade 1 (Port)
0
5
10
15
20
GT3
LE3
Family Size

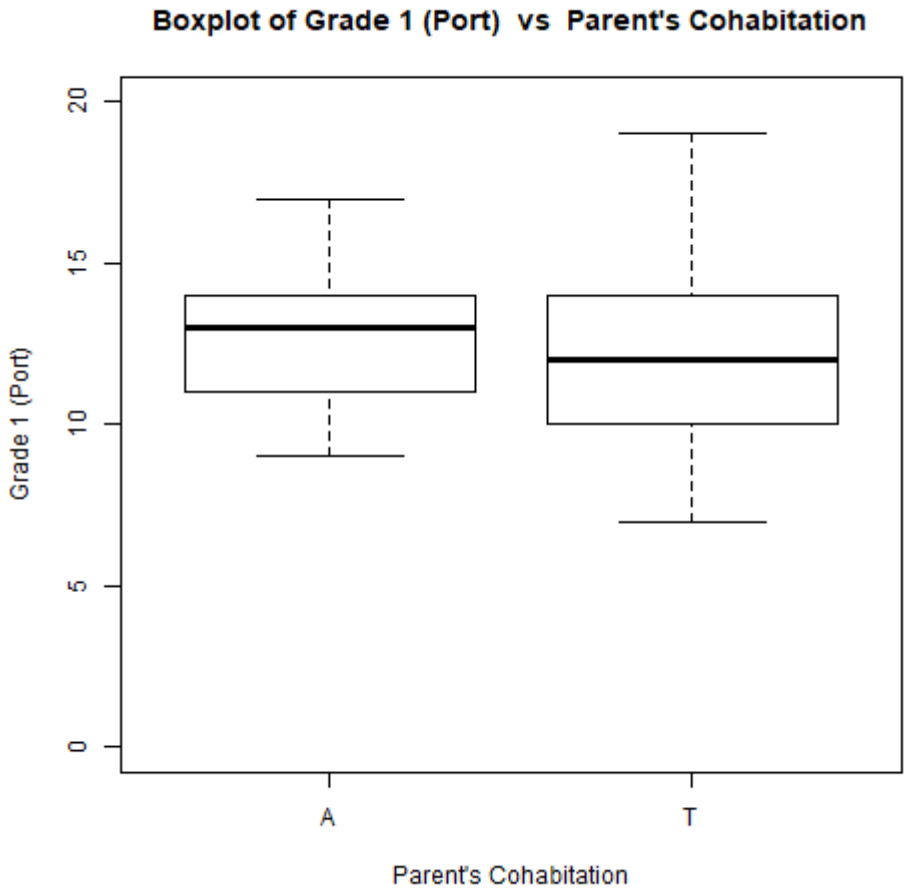

Boxplot of Grade 1 (Port) vs Family Educational Support (Y/N)
Grade 1 (Port)
0
5
10
15
20
no
yes
Family Educational Support (Y/N)

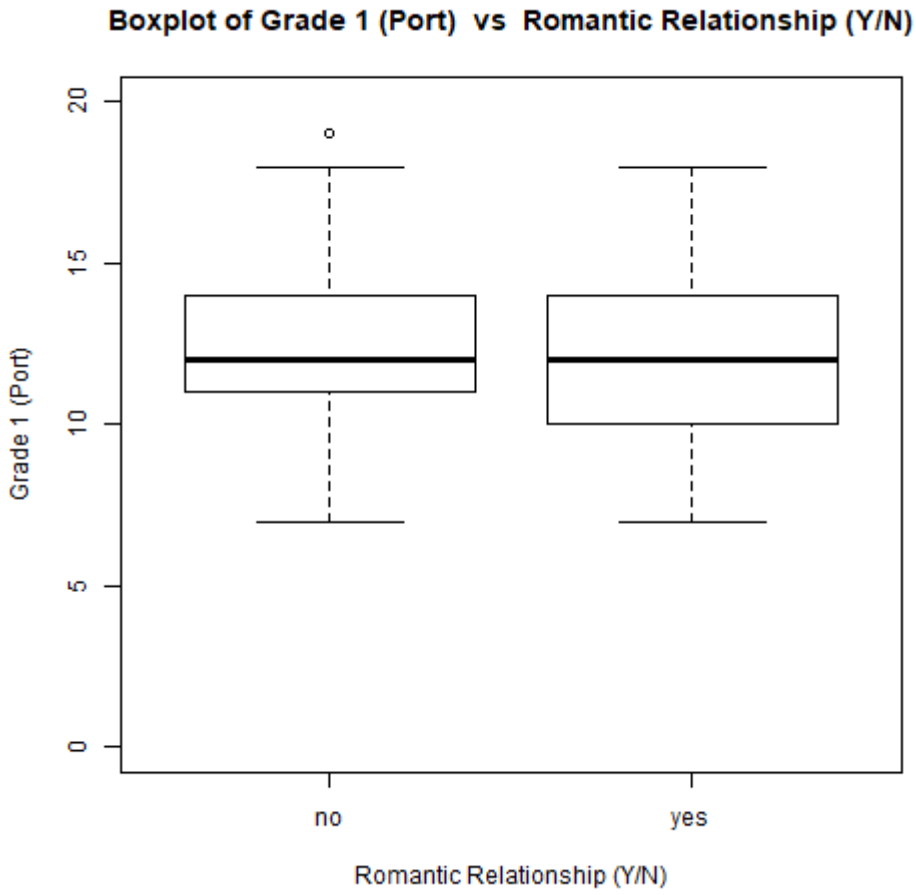

Boxplot of Grade 1 (Port) vs Higher Education Aspirations
Grade 1 (Port)
0
5
10
15
20
no
yes
Higher Education Aspirations

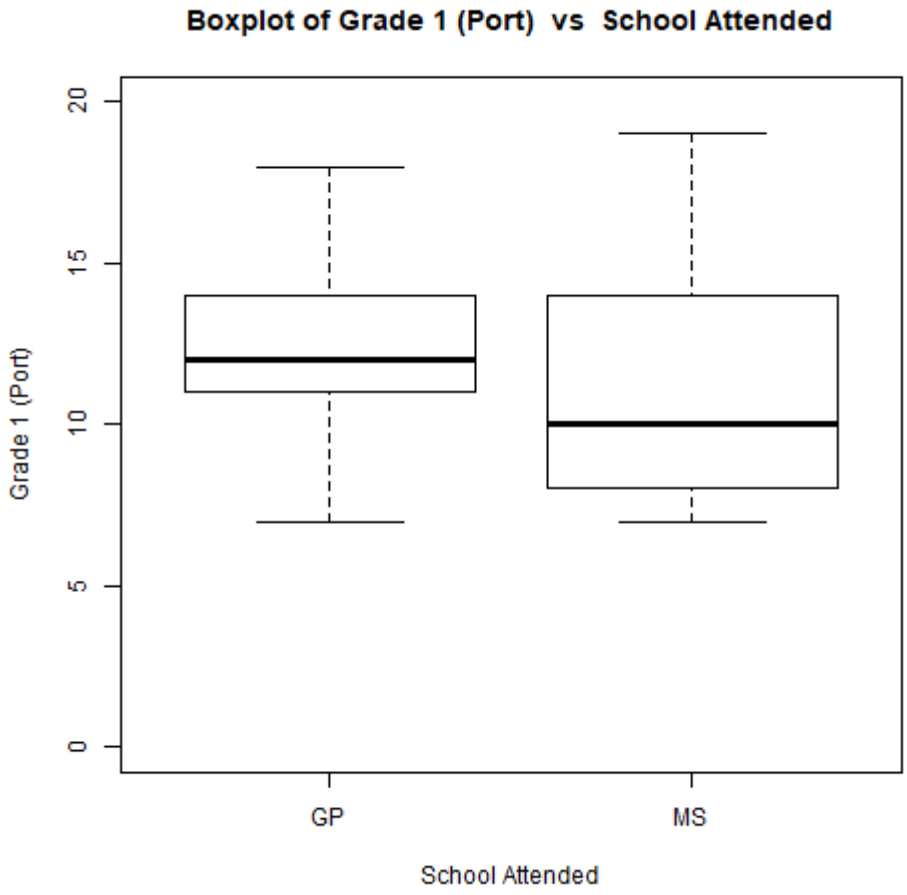

Boxplot of Grade 1 (Port) vs Internet at Home Y/N
Grade 1 (Port)
0
5
10
15
20
no
yes
Internet at Home Y/N

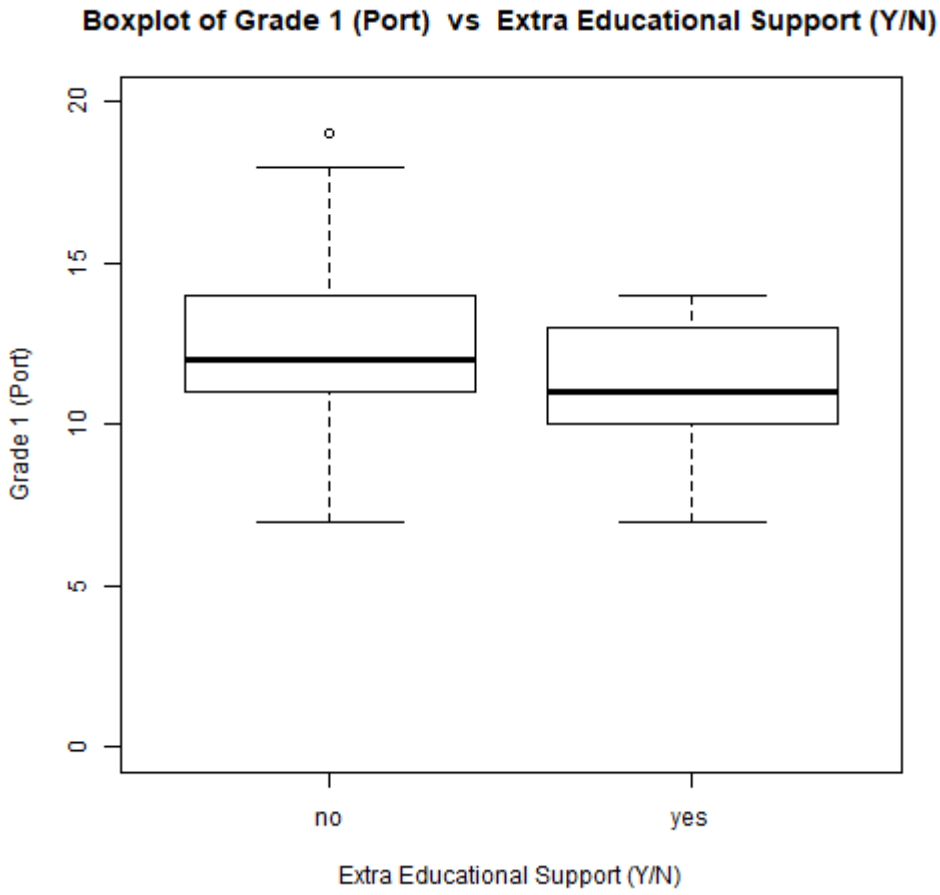

Boxplot of Grade 1 (Port) vs Attended Nursery School Y/N
Grade 1 (Port)
0
5
10
15
20
no
yes
Attended Nursery School Y/N

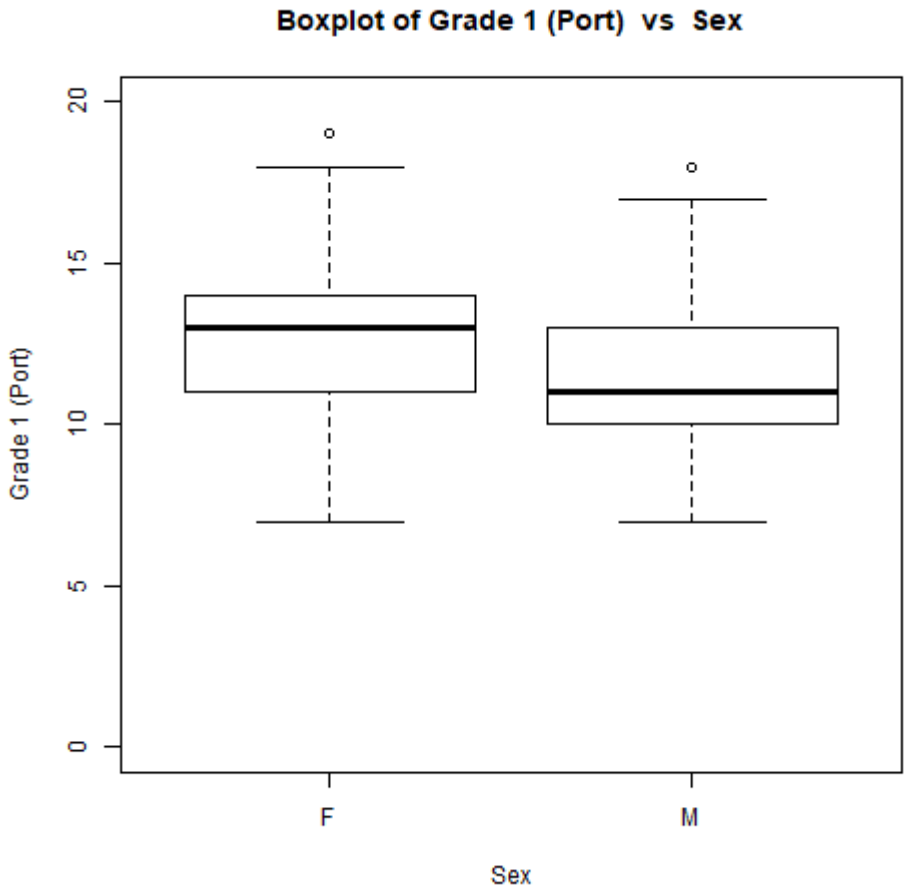

Boxplot of Grade 1 (Port) vs Extra Paid Tutoring
Grade 1 (Port)
0
5
10
15
20
no
yes
Extra Paid Tutoring


Boxplot of Grade 1 (Port) vs Parent's Cohabitation
Grade 1 (Port)
0
5
10
15
20
A
T
Parent's Cohabitation


Boxplot of Grade 1 (Port) vs Romantic Relationship (Y/N)
Grade 1 (Port)
0
5
10
15
20
no
yes
Romantic Relationship (Y/N)


Boxplot of Grade 1 (Port) vs School Attended
Grade 1 (Port)
0
5
10
15
20
GP
MS
School Attended


Boxplot of Grade 1 (Port) vs Extra Educational Support (Y/N)
Grade 1 (Port)
0
5
10
15
20
no
yes
Extra Educational Support (Y/N)



3.2.1 Dropped Categorical Features

After examining the plots above I made the decision to drop the following features due to insignificance or redundancy: *famsize, Pstatus, famsup, nursery, internet,* and *romantic*. Variables have now been reduced from the original 53 down to 27.

4 Discussion

4.1 Enhanced Scatterplots Matrix Discussion

4.1.1 Obervations and Comments from Family Environment

- Fedu and Medu seem to be related.
- Both Fedu and Medu seem to have a slight positive effect on G1
- Medu is a better predictor of G1.

- **Fedu can be dropped as a predictor.**

- Tested a combined variable ave(Fedu, Medu) but it was no better a predictor than Medu.
- famrel seems to have no effect on G1.
- **famrel can be dropped as a predictor.**

4.1.2 Observations and Comments from Time Management

- Study time has a slight positive effect on G1.
- Travel time has a slight negative effect on G1.
- Travel time has a very slight negative effect on study time. i.e. slightly related.
- freetime has a positive effect on goout
- freetime and gout don't have seem to have an effect on G1
- **freetime and goout can be dropped as predictors.**

4.1.3 Observations from Health

- Dalc, Walc and Health have a slight negative effect on G1
- Dalc and Walc are related.
- Tested a variable weighted average of Dalc and Walc to reduce features. It wasn't any more effective.
- Surprisingly Dalc and Walc have no effect on health.
- **Walc can be dropped as a predictor.**

4.1.4 Observations Misc

- age has a negative effect on G1.
- failures has a negative effect on G1
- absences has a negative effect on G1

4.2 Boxplots observations and comments

- The variable higher appeared to have some predictive power.
- G1 scores have a different spread relative to the school the data was drawn from.
- G1 scores from each subject differed by sex.
- A number of features were dropped due to insignificance.

4.3 Cleaned Data

The final cleaned data set was saved as:
data.set named "g"
"cleaned_student_data.rdata"
27 variables x 380 students

5 Appendices

5.1 Appendix 1: Title of appendix

5.3 Appendix 2: Another title