HO CHI MINH UNIVERSITY OF TECHNOLOGY

OFFICE FOR INTERNATIONAL STUDY PROGRAM

PROBABILITY AND STATISTIC

PROJECT REPORT

Instructor: Prf. Nguyen Tien Dung

Class: DTQ1

Group: 07

Member:

Nguyễn Văn Quốc Chương	1950004
Lương Thị Minh Oanh	1950031
Lê Tử Quân	2053372
Nguyễn Minh Khiêm	2052531
Nguyễn Huy Trường	1752583

Table of Contents

1.PROBLEM

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

Attribute Information:

- sex - student's sex (binary: 'F' - female or 'M' - male)
- age - student's age (numeric: from 15 to 22)
- studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4- >10 hours)
- failures - number of past class failures (numeric: n if 1<=n<3, else 4)
- higher - wants to take higher education (binary: yes or no)
- absences - number of school absences (numeric: from 0 to 93)

these grades are related with the course subject, Math or Portuguese:

- G1 - first period grade (numeric: from 0 to 20)
- G2 - second period grade (numeric: from 0 to 20)
- G3 - final grade (numeric: from 0 to 20, output target)

Steps:

1. Import data: grade.csv
2. Data cleaning: NA (Not available)
3. Data visualization
    a. Transformation
    b. Descriptive statistics for each of the variables
    c. Graphs: hist, boxplot, pairs.
4. Fitting linear regression models: We want to explore what factors may affect the final grade.
5. Predictions

2.SOLUTION:

2.1. Import data :

At first, installing the libraries for commands and functions is needed to solve the problem in a clear way.

```
install.packages("dplyr")
install.packages("GGally")
install.packages("corrplot")
install.packages("ggpubr")
install.packages("broom")
```

```
[ ] library(ggplot2)
    library(devtools)
    library(GGally)
    library(corrplot)
    library(dplyr)
    library(broom)
    library(ggpubr)
```

After building a group of libraries, inputting the dataset and organizing the variables or factors from the dataset in columns are the following steps.

```
[ ] #https://drive.google.com/file/d/1xBHBU-hB6K4xQv4UTFEzcvjyQKqWWjpZ/view?usp=sharing
```

```
[ ] system("gdown --id 1xBHBU-hB6K4xQv4UTFEzcvjyQKqWWjpZ")
```

```
[ ] gradeData <- read.table("grade.csv", header = TRUE, sep = ",")
    View(gradeData)
```

2.2. Data cleaning: NA

Locating the null value in any factors and replacing them is the significant stage in data cleaning. In order to complete this step, by using the "summary" command.

```
summary(gradeData)
```

```
     sex                 age          studytime        failures     
 Length:395         Min.   :15.0   Min.   :1.000   Min.   :0.0000  
 Class :character   1st Qu.:16.0   1st Qu.:1.000   1st Qu.:0.0000  
 Mode  :character   Median :17.0   Median :2.000   Median :0.0000  
                    Mean   :16.7   Mean   :2.035   Mean   :0.3342  
                    3rd Qu.:18.0   3rd Qu.:2.000   3rd Qu.:0.0000  
                    Max.   :22.0   Max.   :4.000   Max.   :3.0000  

    higher             absences            G1              G2       
 Length:395         Min.   : 0.000   Min.   : 3.00   Min.   : 0.00  
 Class :character   1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00  
 Mode  :character   Median : 4.000   Median :11.00   Median :11.00  
                    Mean   : 5.709   Mean   :10.91   Mean   :10.72  
                    3rd Qu.: 8.000   3rd Qu.:13.00   3rd Qu.:13.00  
                    Max.   :75.000   Max.   :19.00   Max.   :19.00  
                                                     NA's   :5      
       G3       
 Min.   : 0.00  
 1st Qu.: 8.00  
 Median :11.00  
 Mean   :10.42  
 3rd Qu.:14.00  
 Max.   :20.00  
```

From the result, there are 5 NA values in G2 column, so the next step is the change in those values into the median calculated by rest values in this column.

```
[ ] gradeData[is.na(gradeData)] = median(gradeData$G2, na.rm = TRUE)
    head(gradeData)
```

2.3. Data visualization

2.3.1. Transformation

To utilize R program to calculate, all factors or values from the dataset must be transferred to numeric type. Before the transformation process is coded, several implies are established for thorough understanding.

- School: GP = 0
- School: MS = 1

- Sex: Male = 0
- Sex: Female = 1

- Address: U = 0
- Address: R = 1

- Famsize: GT3 = (
- Famsize: LE3 = 1

- Pstatus: A = 0
- Pstatus: T = 1

- Jobs: at_home = 0
- Jobs: services = 1
- Jobs: teacher = 2
- Jobs: health = 3
- Jobs: other = 4

- Reason: course = 0
- Reason: home = 1
- Reason: reputation = 2
- Reason: other = 3

- Guardian: father = 0
- Guardian: mother = 1
- Guardian: other = 3

- Everything else: no = 0
- Everything else: yes = 1

And then,

```
[ ] gradeData[gradeData == "GP"] <- 0
    gradeData[gradeData == "MS"] <- 1

    gradeData[gradeData == "M"] <- 0
    gradeData[gradeData == "F"] <- 1

    gradeData[gradeData == "U"] <- 0
    gradeData[gradeData == "R"] <- 1

    gradeData[gradeData == "GT3"] <- 0
    gradeData[gradeData == "LE3"] <- 1

    gradeData[gradeData == "A"] <- 0
    gradeData[gradeData == "T"] <- 1

    gradeData[gradeData == "at_home"] <- 0
    gradeData[gradeData == "services"] <- 1
    gradeData[gradeData == "teacher"] <- 2
    gradeData[gradeData == "health"] <- 3
    gradeData$Mjob[gradeData$Mjob == "other"] <- 4
    gradeData$Fjob[gradeData$Fjob == "other"] <- 4

    gradeData[gradeData == "course"] <- 0
    gradeData[gradeData == "home"] <- 1
    gradeData[gradeData == "reputation"] <- 2
    gradeData$reason[gradeData$reason == "other"] <- 3
```

```r
gradeData[gradeData == "father"] <- 0
gradeData[gradeData == "mother"] <- 1
gradeData$guardian[gradeData$guardian == "other"] <- 3

gradeData[gradeData == "yes"] <- 0
gradeData[gradeData == "no"] <- 1

head(gradeData)
```

A data.frame: 6 × 34

	X	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	⋯	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
	<int>	<chr>	<chr>	<int>	<chr>	<chr>	<chr>	<int>	<int>	<chr>	⋯	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<dbl>	<int>
1	1	0	1	18	0	0	0	4	4	0	⋯	4	3	4	1	1	3	6	5	6	6
2	2	0	1	17	0	0	1	1	1	0	⋯	5	3	3	1	1	3	4	5	11	6
3	3	0	1	15	0	1	1	1	1	0	⋯	4	3	2	2	3	3	10	7	8	10
4	4	0	1	15	0	0	1	4	2	3	⋯	3	2	2	1	1	5	2	15	14	15
5	5	0	1	16	0	0	1	3	3	4	⋯	4	3	2	1	2	5	4	6	10	10
6	6	0	0	16	0	1	1	4	3	1	⋯	5	4	2	1	2	5	10	15	11	15

2.3.2. statistics for each of the variables

After the data cleaning and transformation have been done, *class(gradedata* and *summary* command is used to form all the variables into the separate table containing calculating information such as min, 1st Qu., median, mean, 3rd Qu., and max.

```
[ ]  class(gradeData$school) <- "numeric"
     class(gradeData$sex) <- "numeric"
     class(gradeData$address) <- "numeric"
     class(gradeData$famsize) <- "numeric"
     class(gradeData$Pstatus) <- "numeric"
     class(gradeData$Mjob) <- "numeric"
     class(gradeData$Fjob) <- "numeric"
     class(gradeData$reason) <- "numeric"
     class(gradeData$guardian) <- "numeric"
     class(gradeData$schoolsup) <- "numeric"
     class(gradeData$famsup) <- "numeric"
     class(gradeData$paid) <- "numeric"
     class(gradeData$activities) <- "numeric"
     class(gradeData$nursery) <- "numeric"
     class(gradeData$higher) <- "numeric"
     class(gradeData$internet) <- "numeric"
     class(gradeData$romantic) <- "numeric"

     summary(gradeData)
```

Here is the result:

```
       X             school            sex              age
 Min.   :  1.0   Min.   :0.0000   Min.   :0.0000   Min.   :15.0
 1st Qu.: 99.5   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:16.0
 Median :198.0   Median :0.0000   Median :1.0000   Median :17.0
 Mean   :198.0   Mean   :0.1165   Mean   :0.5266   Mean   :16.7
 3rd Qu.:296.5   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:18.0
 Max.   :395.0   Max.   :1.0000   Max.   :1.0000   Max.   :22.0
    address          famsize          Pstatus            Medu
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:2.000
 Median :0.0000   Median :0.0000   Median :1.0000   Median :3.000
 Mean   :0.2228   Mean   :0.2886   Mean   :0.8962   Mean   :2.749
 3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:4.000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :4.000
      Fedu            Mjob            Fjob           reason
 Min.   :0.000   Min.   :0.000   Min.   :0.000   Min.   :0.000
 1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.000
 Median :2.000   Median :2.000   Median :4.000   Median :1.000
 Mean   :2.522   Mean   :2.241   Mean   :2.762   Mean   :1.081
 3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000
 Max.   :4.000   Max.   :4.000   Max.   :4.000   Max.   :3.000
    guardian        traveltime      studytime        failures
 Min.   :0.0000   Min.   :1.000   Min.   :1.000   Min.   :0.0000
 1st Qu.:1.0000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000
 Median :1.0000   Median :1.000   Median :2.000   Median :0.0000
 Mean   :0.9342   Mean   :1.448   Mean   :2.035   Mean   :0.3342
 3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
 Max.   :3.0000   Max.   :4.000   Max.   :4.000   Max.   :3.0000
```

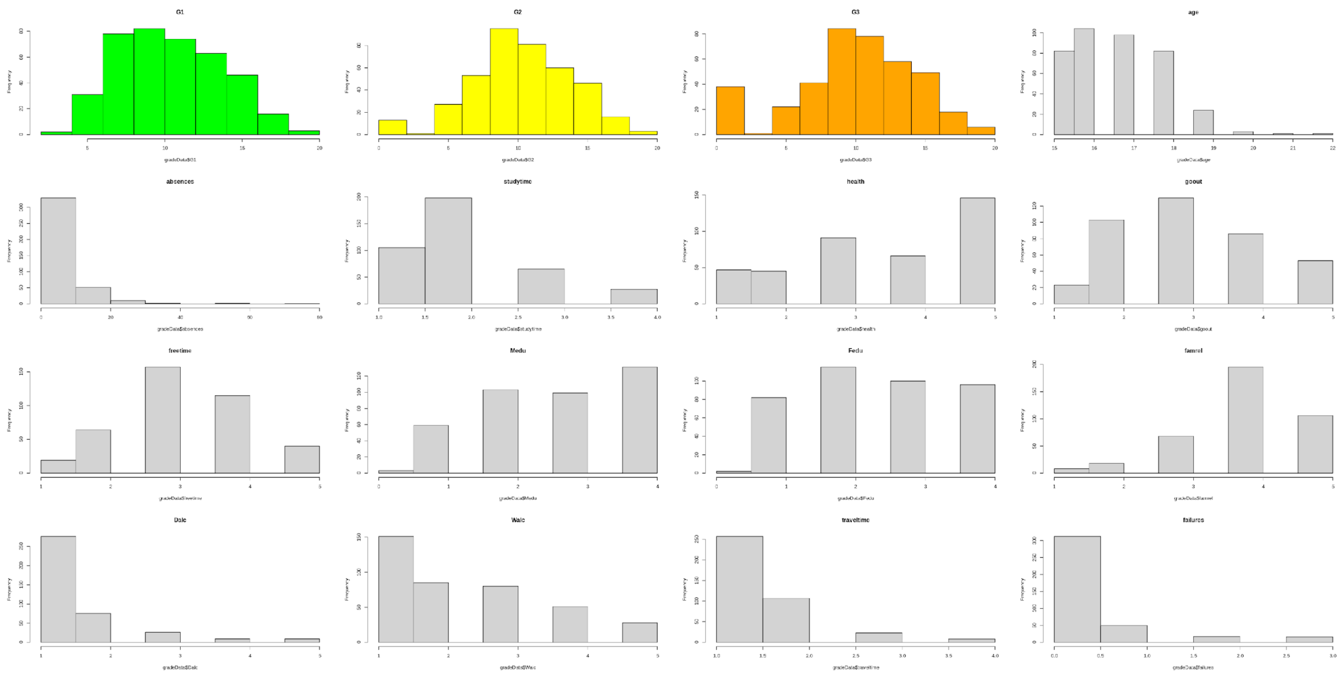
```
   schoolsup        famsup           paid           activities    
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :1.0000   Median :0.0000   Median :1.0000   Median :0.0000  
 Mean   :0.8709   Mean   :0.3873   Mean   :0.5418   Mean   :0.4911  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
    nursery           higher           internet         romantic     
 Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :0.0000   Median :0.00000   Median :0.0000   Median :1.0000  
 Mean   :0.2051   Mean   :0.05063   Mean   :0.1671   Mean   :0.6658  
 3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000  
     famrel         freetime         goout            Dalc      
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000  
 1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000  
 Median :4.000   Median :3.000   Median :3.000   Median :1.000  
 Mean   :3.944   Mean   :3.235   Mean   :3.109   Mean   :1.481  
 3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:2.000  
 Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000  
      Walc          health         absences            G1       
 Min.   :1.000   Min.   :1.000   Min.   : 0.000   Min.   : 3.00  
 1st Qu.:1.000   1st Qu.:3.000   1st Qu.: 0.000   1st Qu.: 8.00  
 Median :2.000   Median :4.000   Median : 4.000   Median :11.00  
 Mean   :2.291   Mean   :3.554   Mean   : 5.709   Mean   :10.91  
 3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.: 8.000   3rd Qu.:13.00  
 Max.   :5.000   Max.   :5.000   Max.   :75.000   Max.   :19.00  
```

```
       G2              G3       
 Min.   : 0.00   Min.   : 0.00  
 1st Qu.: 9.00   1st Qu.: 8.00  
 Median :11.00   Median :11.00  
 Mean   :10.72   Mean   :10.42  
 3rd Qu.:13.00   3rd Qu.:14.00  
 Max.   :19.00   Max.   :20.00  
```

2.3.3. Graphs: hist, boxplot, pairs

Hist

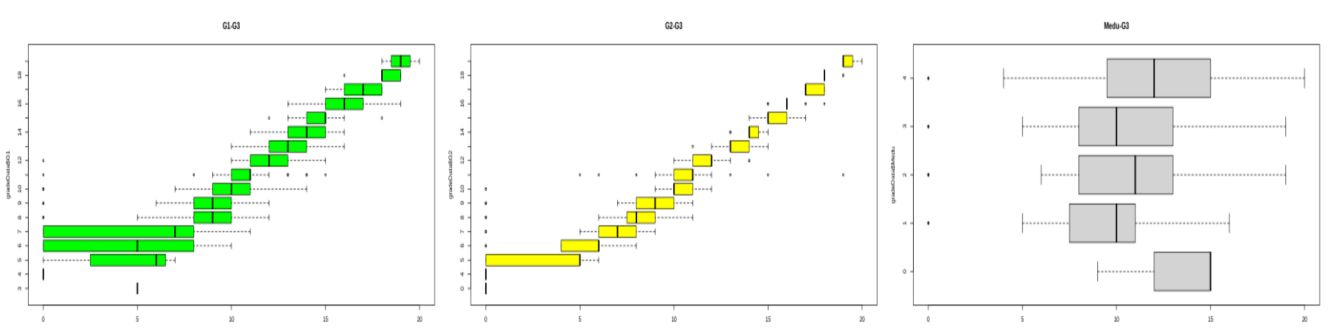
```
[ ] options(repr.plot.width=30, repr.plot.height=15)
    par(mfrow=c(4,4))
    hist(gradeData$G1, main = "G1", col = "green")
    hist(gradeData$G2, main = "G2", col = "yellow")
    hist(gradeData$G3, main = "G3", col = "orange")
    hist(gradeData$age, main = "age")
    hist(gradeData$absences, main = "absences")
    hist(gradeData$studytime, main = "studytime")
    hist(gradeData$health, main = "health")
    hist(gradeData$goout, main = "goout")
    hist(gradeData$freetime, main = "freetime")
    hist(gradeData$Medu, main = "Medu")
    hist(gradeData$Fedu, main = "Fedu")
    hist(gradeData$famrel,  main = "famrel")
    hist(gradeData$Dalc, main = "Dalc")
    hist(gradeData$Walc, main = "Walc")
    hist(gradeData$traveltime, main = "traveltime")
    hist(gradeData$failures, main = "failures")
```

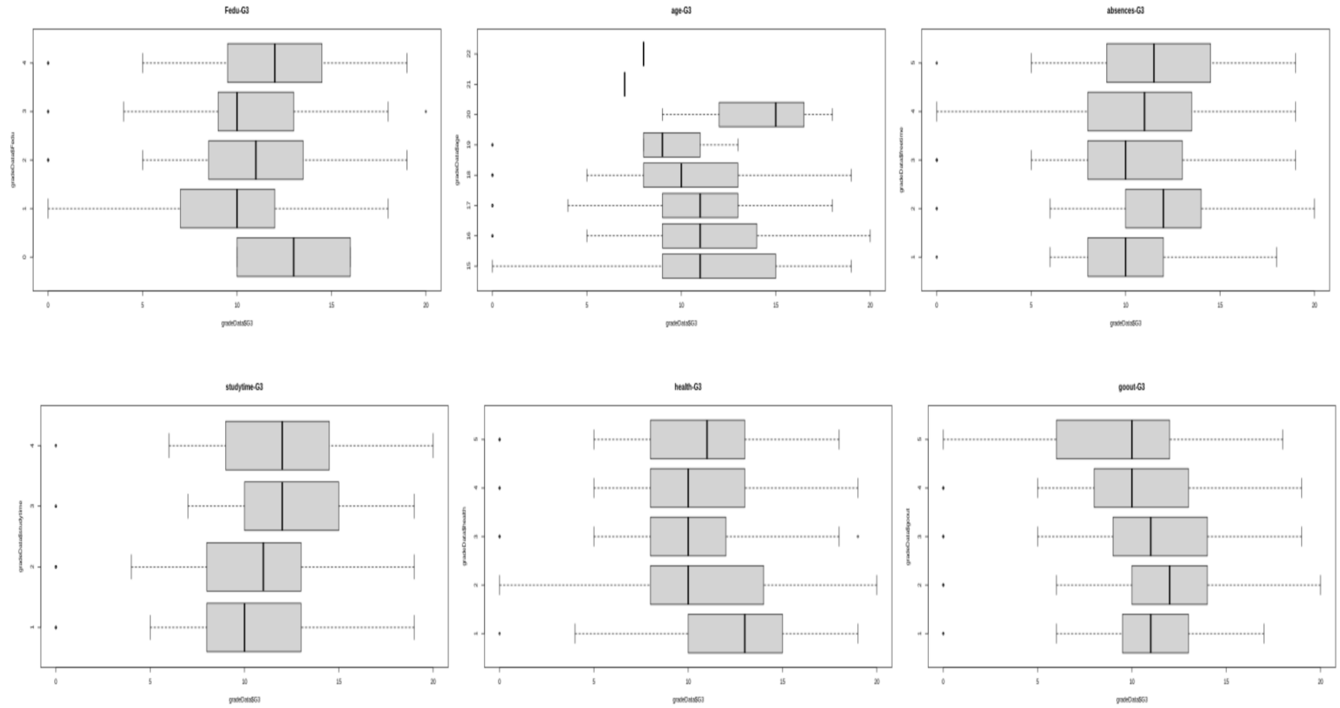


Boxplot

Using *boxplot* command to compare final grade G3 with G1, G2, Medu, Fedu, age, absences, studytime, health and goout.

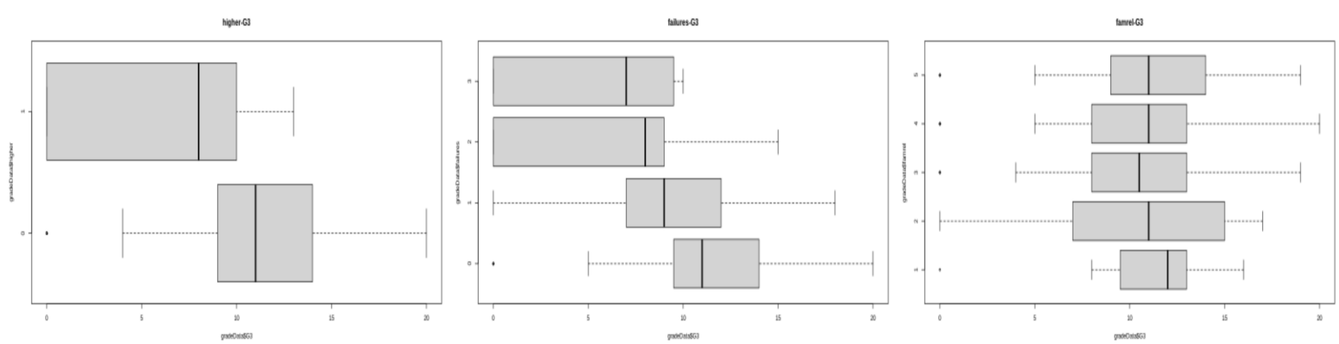
```
options(repr.plot.width=30, repr.plot.height=15)
par(mfrow=c(3,3))
boxplot(gradeData$G3 ~ gradeData$G1, horizontal = TRUE, main = "G1-G3", col = "green")
boxplot(gradeData$G3 ~ gradeData$G2, horizontal = TRUE, main = "G2-G3", col = "yellow")
boxplot(gradeData$G3 ~ gradeData$Medu, horizontal = TRUE, main = "Medu-G3")
boxplot(gradeData$G3 ~ gradeData$Fedu, horizontal = TRUE, main = "Fedu-G3")
boxplot(gradeData$G3 ~ gradeData$age, horizontal = TRUE, main = "age-G3")
boxplot(gradeData$G3 ~ gradeData$freetime, horizontal = TRUE, main = "absences-G3")
boxplot(gradeData$G3 ~ gradeData$studytime, horizontal = TRUE, main = "studytime-G3")
boxplot(gradeData$G3 ~ gradeData$health, horizontal = TRUE, main = "health-G3")
boxplot(gradeData$G3 ~ gradeData$goout, horizontal = TRUE, main = "goout-G3")
```

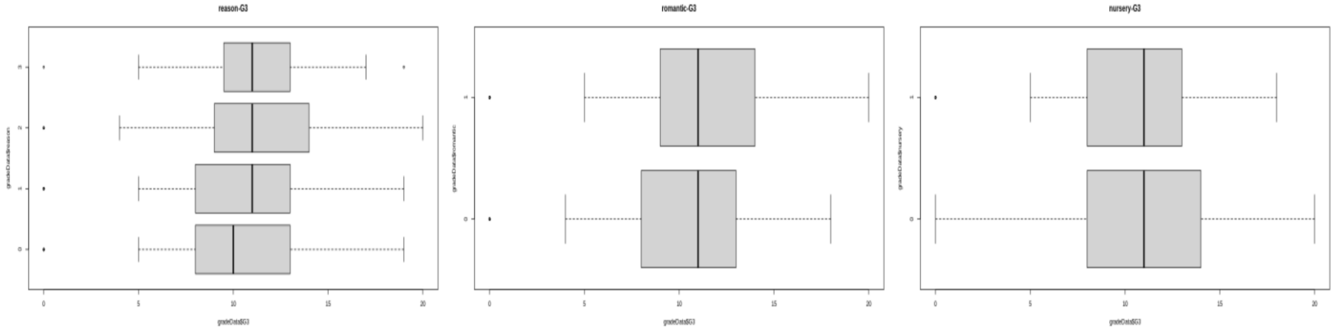


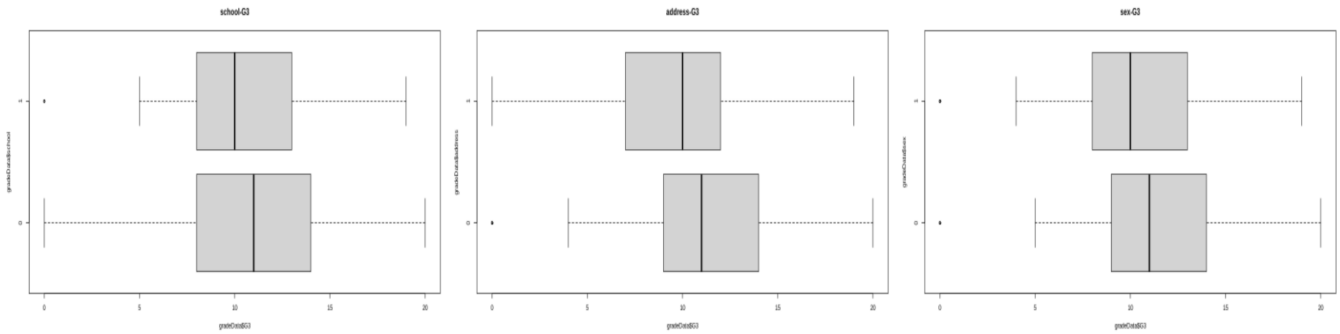


Continuously, using *boxplot* command to compare final grade G3 with school, address, sex, higher, failures, famrel, reason, romantic and nursery.

```
[ ]  options(repr.plot.width=30, repr.plot.height=15)
     par(mfrow=c(3,3))
     boxplot(gradeData$G3 ~ gradeData$school, horizontal = TRUE, main = "school-G3")
     boxplot(gradeData$G3 ~ gradeData$address, horizontal = TRUE, main = "address-G3")
     boxplot(gradeData$G3 ~ gradeData$sex, horizontal = TRUE, main = "sex-G3")
     boxplot(gradeData$G3 ~ gradeData$higher, horizontal = TRUE, main = "higher-G3")
     boxplot(gradeData$G3 ~ gradeData$failures, horizontal = TRUE, main = "failures-G3")
     boxplot(gradeData$G3 ~ gradeData$famrel, horizontal = TRUE, main = "famrel-G3")
     boxplot(gradeData$G3 ~ gradeData$reason, horizontal = TRUE, main = "reason-G3")
     boxplot(gradeData$G3 ~ gradeData$romantic, horizontal = TRUE, main = "romantic-G3")
     boxplot(gradeData$G3 ~ gradeData$nursery, horizontal = TRUE, main = "nursery-G3")
```



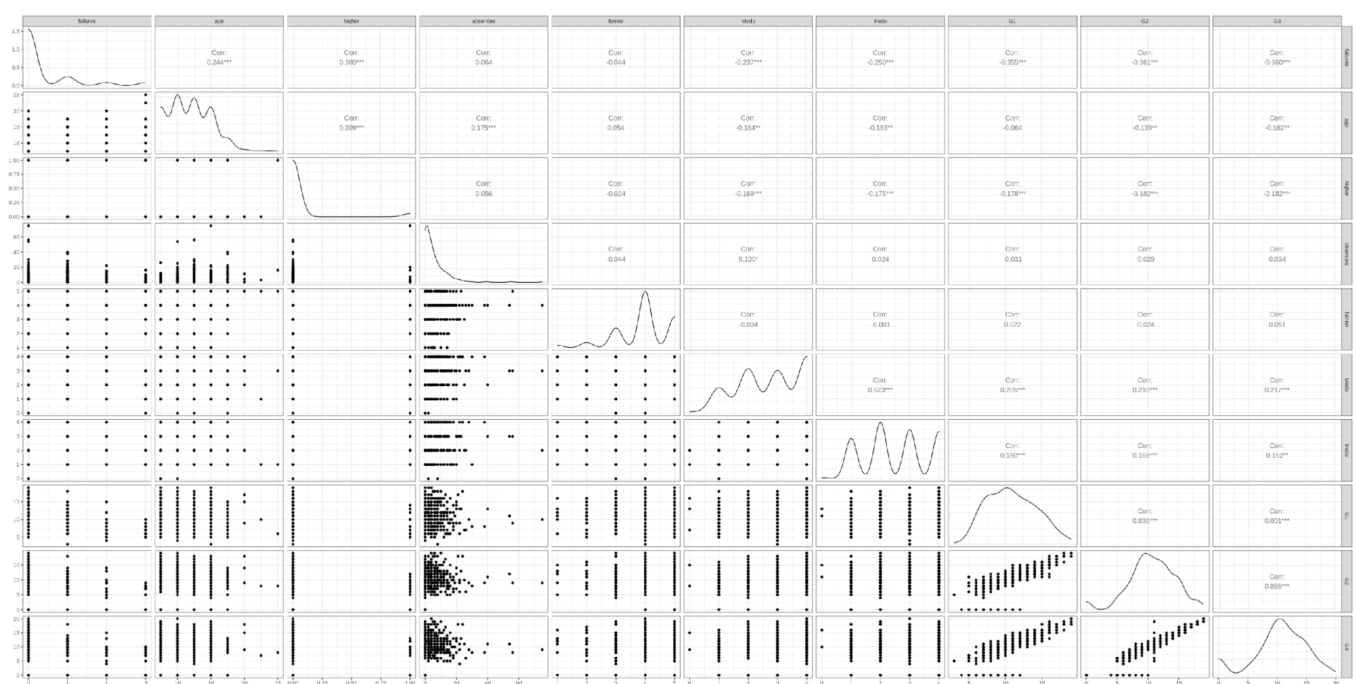




Pairs

Using *pairs* command to show the statistical relationship between variables (failures, age, higher, absences, famrel, Medu, Fedu, G1, G2 and G3).

```
[ ]  options(repr.plot.width=30, repr.plot.height=15)
     ggpairs(subData) + theme_bw()
```



2.3.4. Fitting linear regression models

First, using below command to confirm that G3 is a function of the other values and $data = grade$ to confirm that R has to compute on dataset called grade.

```
[ ]  LinearModel <- lm(G3 ~ .,data=gradeData)
     summary(LinearModel)
```

```
[ ] Call:
    lm(formula = G3 ~ ., data = gradeData)

    Residuals:
        Min      1Q  Median      3Q     Max
    -7.5690 -0.6073  0.2500  1.0744  5.7061

    Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
    (Intercept) -3.827865   2.388814  -1.602  0.10996
    X           -0.004157   0.001663  -2.499  0.01291 *
    school1      0.939343   0.402766   2.332  0.02025 *
    sex1        -0.224026   0.238265  -0.940  0.34774
    age          0.043876   0.143198   0.306  0.75948
    address1    -0.027359   0.276990  -0.099  0.92138
    famsize1     0.092757   0.231310   0.401  0.68866
    Pstatus1    -0.298755   0.341233  -0.876  0.38189
    Medu         0.108200   0.153445   0.705  0.48119
    Fedu        -0.158381   0.131043  -1.209  0.22762
    Mjob1        0.316354   0.376445   0.840  0.40127
    Mjob2        0.175186   0.491379   0.357  0.72166
    Mjob3        0.006884   0.527828   0.013  0.98960
    Mjob4        0.266285   0.337450   0.789  0.43058
    Fjob1       -0.183476   0.501502  -0.366  0.71469
    Fjob2       -0.115774   0.612794  -0.189  0.85026
    Fjob3        0.373380   0.678883   0.550  0.58267
    Fjob4        0.023485   0.484967   0.048  0.96140
    reason1     -0.110051   0.261144  -0.421  0.67371
    reason2      0.180781   0.272287   0.664  0.50717
    reason3      0.342835   0.387481   0.885  0.37688
    guardian1    0.241905   0.257476   0.940  0.34811
```

Based on p-value, constructing 6 models more by eliminating one by one variable from the low p-value to the lowest.

```
[ ]  LinearModel_1 <- lm(G3 ~ X +school+ famrel + absences + G1 + G2 , data = gradeData)
     LinearModel_2 <- lm(G3 ~ school + famrel + absences + G1 + G2, data= gradeData)
     LinearModel_3 <- lm(G3 ~ famrel + absences + G1 + G2, data = gradeData)
     LinearModel_4 <- lm(G3 ~ absences + G1 + G2, data = gradeData)
     LinearModel_5 <- lm(G3 ~ G1 + G2, data = gradeData)
     LinearModel_6 <- lm(G3 ~ G2, data = gradeData)
```

```
[ ] anova(LinearModel_6,LinearModel_5,LinearModel_4,LinearModel_3,LinearModel_2,LinearModel_1,LinearModel)
```

A anova: 7 × 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	393	1642.932	NA	NA	NA	NA
2	392	1565.603	1	77.328443	20.6497771	7.590171e-06
3	391	1534.502	1	31.101716	8.3053980	4.195276e-03
4	390	1495.395	1	39.106886	10.4430976	1.347272e-03
5	389	1494.942	1	0.452328	0.1207896	7.283874e-01
6	388	1425.370	1	69.572059	18.5785134	2.118998e-05
7	352	1318.155	36	107.215052	0.7952970	7.961416e-01

Then, by *anova* command, the comparison between regression models are built.

Observing the Anova data table from the model 1 to 7, the result has illustrated that the model 2 seems to be the finest model to be built a fitting linear regression model compared to other models because of the p-values (the model 2 has smallest value, p2 ~ 0.019).

model 2: `G3 ~ school + famrel + absences + G1 + G2`

Then, having the fitting model below:

```
[ ] guardian3    -0.052696   0.474736  -0.111  0.91168
    traveltime    0.101313   0.160785   0.630  0.52903
    studytime    -0.099203   0.137499  -0.721  0.47109
    failures     -0.193218   0.167203  -1.156  0.24863
    schoolsup1   -0.449382   0.326983  -1.374  0.17021
    famsup1      -0.125593   0.230105  -0.546  0.58554
    paid1        -0.256815   0.226228  -1.135  0.25706
    activities1   0.323779   0.210157   1.541  0.12430
    nursery1      0.221102   0.259206   0.853  0.39424
    higher1      -0.247778   0.513630  -0.482  0.62982
    internet1     0.096594   0.294700   0.328  0.74328
    romantic1     0.209608   0.225673   0.929  0.35362
    famrel        0.347329   0.116769   2.975  0.00314 **
    freetime      0.025411   0.112307   0.226  0.82113
    goout        -0.015578   0.107024  -0.146  0.88436
    Dalc         -0.212028   0.156003  -1.359  0.17497
    Walc          0.210583   0.117030   1.799  0.07281 .
    health        0.044192   0.076199   0.580  0.56232
    absences      0.041264   0.013654   3.022  0.00269 **
    G1            0.305410   0.060386   5.058 6.85e-07 ***
    G2            0.873046   0.052079  16.764  < 2e-16 ***
    ---
    Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Residual standard error: 1.935 on 352 degrees of freedom
    Multiple R-squared:  0.8406,    Adjusted R-squared:  0.8216
    F-statistic:  44.2 on 42 and 352 DF,  p-value: < 2.2e-16
```

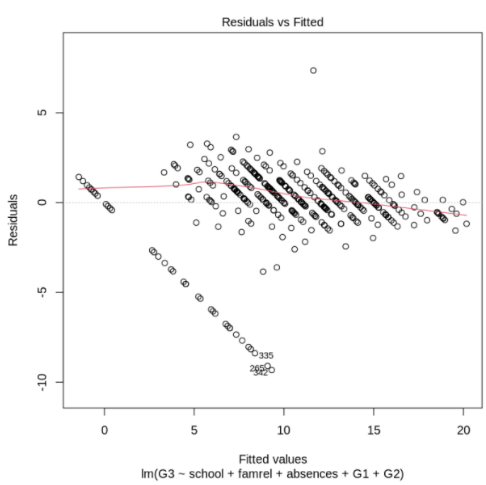
$$\boldsymbol{G3} = -3.77114 + 0.93638 \times G2 + 0.23115 \times G1 + 0.35501 \times famrel + 0.03726 \times absences + 0.10628 \times \text{school1}$$

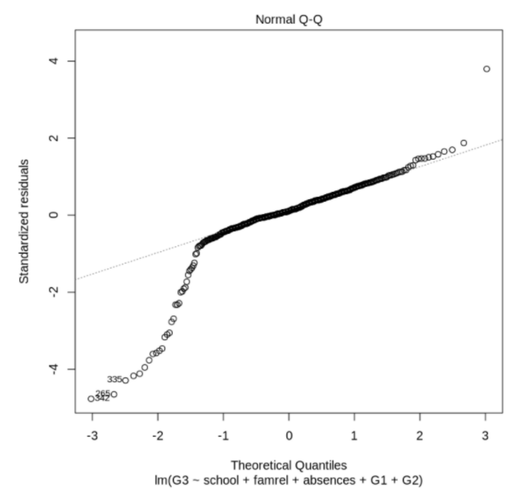
```
[ ]  summary(LinearModel_2)
```

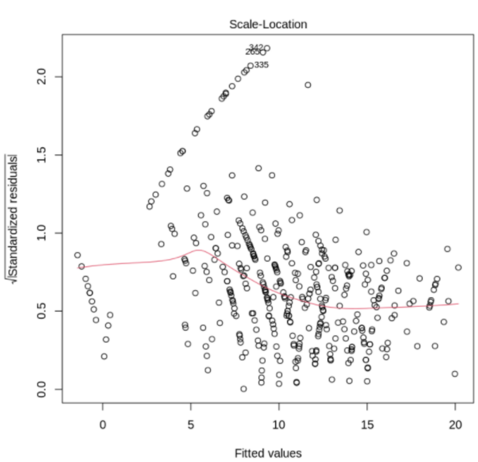
```
Call:
lm(formula = G3 ~ school + famrel + absences + G1 + G2, data = gradeData)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.3242 -0.4523  0.2072  1.0080  7.3526 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.77114    0.56316  -6.696 7.49e-11 ***
school1      0.10628    0.30980   0.343  0.73173    
famrel       0.35501    0.11080   3.204  0.00147 ** 
absences     0.03726    0.01241   3.002  0.00285 ** 
G1           0.23115    0.05443   4.247 2.72e-05 ***
G2           0.93638    0.04870  19.226  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.96 on 389 degrees of freedom
Multiple R-squared:  0.8192,	Adjusted R-squared:  0.8169 
F-statistic: 352.6 on 5 and 389 DF,  p-value: < 2.2e-16
```

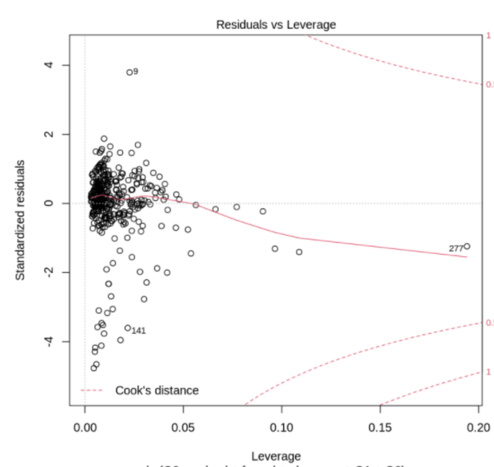
Following that, plotting that model:

```
[ ]  plot(LinearModel_2)
```









2.4. Predictions

2.4.1. Evaluation

First, in order to evaluate whether those students passed or failed based on final grade, the condition order: *if their final grade is not less than 10, they are passed* , is used to *evaluate*. After that step,the prediction data also is built as the same function above but predict_G3.

```
[ ] evaluate = gradeData$G3
    evaluate = ifelse(evaluate >=10,"pass","fail")
    observe = table(evaluate)
    View (observe)
```

```
evaluate
fail pass
 130  265
```

```
[ ] Predict_G3 = predict(LinearModel_2,gradeData)
    Predict_G3 = ifelse(Predict_G3>=10, "pass", "fail")
    observe = table(Predict_G3)
    View (observe)
```

```
Predict_G3
fail pass
 185  210
```

The percent error for students who failed is $\frac{185-130}{130}\times 100\% = 42.31\%$

The percent error for students who passed is $\frac{265-210}{265}\times 100\% = 20.75\%$

2.4.2. Prediction a new data

First, creating a data frame to predict the final grade. As below, the new data frame is given as an example

```
newd = data.frame(school = 1,famrel =5,absences =20, G1 =10, G2 =11)
```

Then, using *predict* command to compute G3 (final grade) from the others factor in the data frame.

```
G3_predict = predict(LinearModel_2,newd)
```

And using *round* command to round the result

```
round(G3_predict, digits = 4)
```

1: 11.4671

Finally, the final result computed by R is 11.4671.

REFERENCES

Our souce code: https://colab.research.google.com/drive/1zOCpF4MARuGzPXpdbI-peQ7Amtbc-aNv?usp=sharing (we run directly on the google collab and then converting to the R file)

1. R-tutor.com. 2021. *Estimated Multiple Regression Equation | R Tutorial*. [online] Available at: <http://www.r-tutor.com/elementarystatistics/multiple-linear-regression/estimated-multiple-regressionequation> [Accessed 23 May 2021].
2. Advstats.psychstat.org. 2021. *Relative Importance of Predictors -- Advanced Statistics using R*. [online] Available at: <https://advstats.psychstat.org/book/mregression/importance.php> [Accessed 23 May 2021].
3. Youtube.com. 2021. *R Stats: Multiple Regression - Variable Selection*. [online] Available at: <https://www.youtube.com/watch?v=HP3RhjLhRjY&t=408s> [Accessed 23 May 2021].
4. Phillips, N., 2021. *YaRrr! The Pirate's Guide to R*. [online] Bookdown.org. Available at: <https://bookdown.org/ndphillips/YaRrr/comparingregression-models-with-anova.html> [Accessed 28 May 2021].
5. Nguyễn Văn, T., 2006. *PHÂN TÍCH SỐ LIỆU VÀ TẠO BIỂU ĐỒ BẰNG R*. Ho Chi Minh City: Nhà xuất bản Đại học Bách Khoa TP. Hồ Chí Minh.
6. Archive.ics.uci.edu. 2021. *Wine quality dataset*. [online] Available at: <https://archive.ics.uci.edu/ml/datasets/Wine+Quality?fbclid=IwAR22sb8xlcpIyexBIFWHbA7DQtuk2F_WGsMffU-CWPIzdGhCv5_karnGWiw> [Accessed 01 June 2021].