Effect of Socio-Economic Factors on Student's Performance

Jishant Acharya

25/03/2022

References

1. S. Kotsiantis , C. Pierrakeas & P. Pintelas ; Predicting Students' Performance In Distance Learning Using Machine Learning Techniques
2. Paulo Cortez and Alice Silva ; Using Data Mining to Predict Secondary School Student Performance
3. Juan-José Navarro1, Javier García-Rubio, Pedro R. Olivares ; The Relative Age Effect and Its Influence on Academic Performance
4. Kawtar Tani, Elizabeth Dalzell, Nathan Ehambaranathan, Sheela Murugasu and Anne Steele ; Evaluation of factors affecting students' performance in tertiary education
5. Thomas R. Ford ; Social Factors Affecting Academic Performance: Further Evidence
6. Yaoran Li , Jeff Allen , Alex Casillas ; Relating psychological and social factors to academic performance: A longitudinal investigation of high-poverty middle school students

Introduction :

School has become a place to compete against the world to prove yourself and set yourself apart from the crowd to excel. This is done to measure everyone's grasping power with the same yardstick marks. This is in no way the right manner to judge a person's intelligence. Each individual is different, everyone comes from a varied set of backgrounds. There can never be a single yardstick that can measure each and everyone. This analysis is trying to prove that.
There are various effects that have been in place in the background and have different effects on the ability of the student to perform in exams.
.
In the following exploratory analysis document we use the data set[2] and try to back the findings that were made by many researches.
We all know that there is a dependence of socio-economic parameters on marks. We will try to help back the conclusion made with respect to age that with increase in the age marks or the ability to perform goes down[3].
We also know that poverty has been an obstacle in obtaining quality education. As a domino effect , it affects the students ability to perform due to all the other circumstances that come with poverty[6].
The are various studies that back the fact that the gender of the student also plays a pivotal role in being an over-achiver, females are more likely to be over-achievers when compared to male[5].
Conclusively, we have seen in real world that there are many indirectly influencing factors that affect the marks and the overall ability to learn. There are studies that have been successful in showcasing the said effect of indirect factors on marks of the student[1].

Central Idea

The main idea is to find the factors that directly or indirectly affect the student performance.Here we will try to peg everything against the G3 value to measure performance on a single standard scale and also because

G1 and G2 have a very strong correlation with G3 which shows an influence.This means if we are able to understand the influence on G3 we will easily be able to comment on the overall performance of the student.

.

We will be running a analysis that works to yield relation with with final marks for the above mentioned reasons but we also will be working in finding indirect influences by working on 2 columns that are dependent on each other which in-turn has an effect on the final marks.

Dataset Description

The following is an extensive data set that has many social factors also included in the records. There are 3 types of marks. G1, G2 and G3 are three marks that correspond to the first period, second period and the final marks. There are other factors present in the data set that range from being related socially to being able to show economical situation of the student.

```
##    ï..Variable        Type                                     Description
## 1
## 2       school categorical                          One of the two schools
## 3          sex categorical                                  Male or Female
## 4          age  continuous                                 Age from 18 -22
## 5      address categorical                                  Urban or Rural
## 6      famsize categorical         LT3 - Less Than 3 ; GT3 - Greater than 3
## 7      Pstatus categorical             Living 'T' (Together) or 'A' (Apart)
## 8         Medu categorical     Level of Education ( 5 levels - From 0 to 4)
## 9         Fedu categorical     Level of Education ( 5 levels - From 0 to 4)
## 10        Mjob categorical                                   Types of Jobs
## 11        Fjob categorical                                   Types of Jobs
## 12  traveltime categorical           Level of Travel (4 levels - From 1 - 4)
## 13   studytime categorical        Level of Study Time (4 levels - From 1-4)
## 14     failure categorical               Past failures ( n if 1<=n<3 or 4
## 15   schoolsup categorical               Yes / No for support from school
## 16      famsup categorical               Yes / No for support from family
## 17        paid categorical                     Yes / No for paid classes 
## 18     nursery categorical                Yes / No for nursery attendance
## 19    internet categorical                     Yes / No for availability
## 20       goout categorical                                  Level from 1-5
## 21    romantic categorical Yes / No from involvement in romantic activities
## 22    freetime categorical                     Level of Free time from 1-5
## 23      health categorical                   Quality of health from 1-5
## 24          G1  continuous                          Marks for first period
## 25          G2  continuous                         Marks for second period
## 26          G3  continuous                          Marks for third period
```

```
##   school sex age address famsize Pstatus Medu Fedu    Mjob    Fjob traveltime
## 1     GP   F  18       U     GT3       A    4    4 at_home teacher          2
## 2     GP   F  17       U     GT3       T    1    1 at_home   other          1
## 3     GP   F  15       U     LE3       T    1    1 at_home   other          1
##   studytime failures schoolsup famsup paid nursery internet romantic freetime
## 1         2        0       yes     no   no     yes       no       no        3
## 2         2        0        no    yes   no      no      yes       no        3
## 3         2        3       yes     no  yes     yes      yes       no        3
##   goout health absences G1 G2 G3
## 1     4      3        6  5  6  6
## 2     3      3        4  5  5  6
## 3     2      3       10  7  8 10
```

Analysis

Gender Distribution

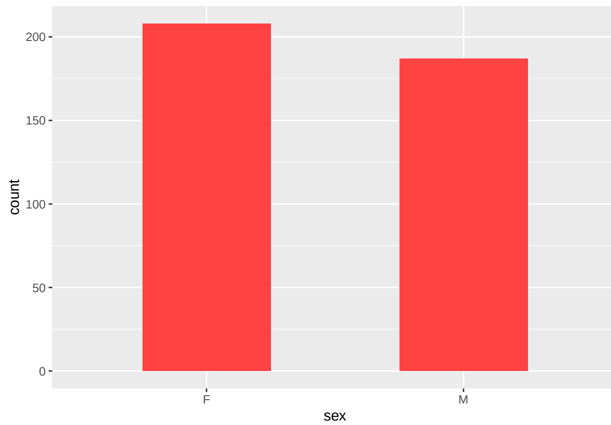

Figure 1: Population distribution between genders

```r
paste("Total number of students in the sample", nrow(student.mat))
```

```
## [1] "Total number of students in the sample 395"
```

```r
paste("Number of Male Students: ",nrow(student.mat[student.mat$sex == 'M',]))
```

```
## [1] "Number of Male Students:  187"
```

```r
paste("Number of Female Students: ",nrow(student.mat[student.mat$sex == 'F',]))
```

```
## [1] "Number of Female Students:  208"
```

> This graph shows that there is no class bias in the dataset, the number of male is almost similar to the number of females.

Frequency Distribution of Male Population

```r
#Plots the Male Age Frequency distribution, along side plotting the mean and the median lines

ggplot(student.mat[student.mat$sex == "M",], aes(age)) + geom_histogram(fill=primary,binwidth = 1) +
geom_vline(xintercept=mean(student.mat[student.mat$sex == "M",]$age),size=2, color=secondary) +
geom_vline(color=third, size=2,xintercept=median(student.mat[student.mat$sex == "M",]$age))
```

> This shows that the age of males is skewed to the left and has a tail towards the right.The Green line is the median of the distribution and the blue line is the mean of the distribution.

Frequency Distribution of Female Population

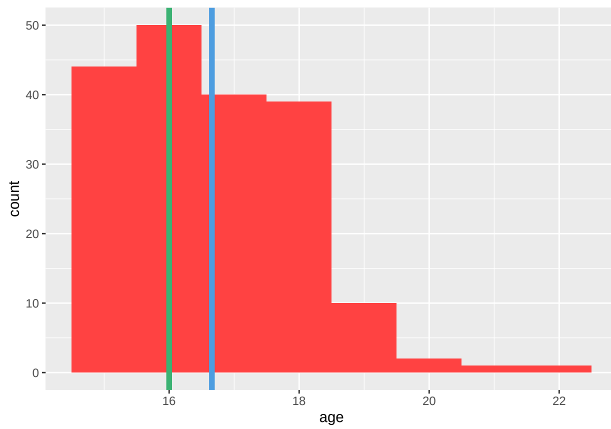

Figure 2: Male Population frequency distribution

```
#Plots the Female Age Frequency distribution, along side plotting median and mean line

ggplot(student.mat[student.mat$sex == "F",], aes(age)) + geom_histogram(fill=primary,binwidth = 1) +
geom_vline(xintercept=mean(student.mat[student.mat$sex == "F",]$age),size=2, color=secondary) +
geom_vline(color=third, size=2,xintercept=median(student.mat[student.mat$sex == "F",]$age))
```

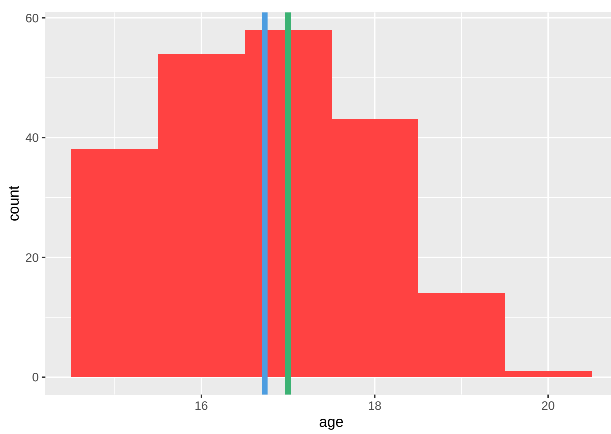

Figure 3: Female population frequency distribution

This shows that the distribution is slightly skewed to the right but not by a lot. Here also the green line shows the median and the blue line shows the mean of the distribution.

Dependance of Age on Marks (Liner Regression)

```
#Gives the Liner Model Coefficient and Intercept
lm(G3 ~ age, data=student.mat)
```

```
##
## Call:
```

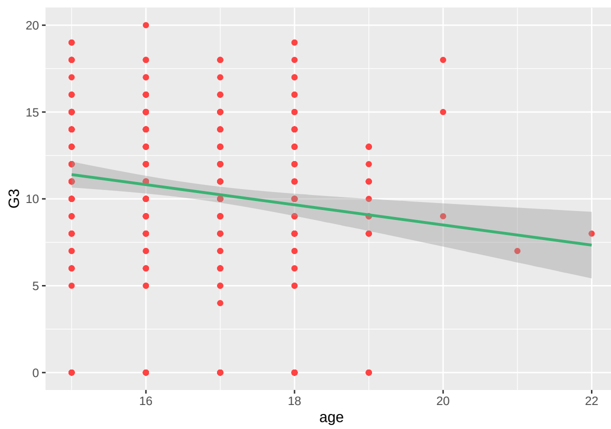

Figure 4: Age v/s G3 Regression

```
## lm(formula = G3 ~ age, data = student.mat)
##
## Coefficients:
## (Intercept)          age
##     20.1011      -0.5801
```

Result : *The purpose of the test was to yield the relationship between age and marks of the student. The sample space was of 395 student and their final semester marks we taken into consideration. We performed linear regression on the variables and found a negative correlation. This shows that not only there is a dependence of age on marks but it's negative in nature which means as age progresses, marks of the student will come down.The estimation formula comes out to be* $\hat{y} = 20.1011 - (0.5801)age$.

Dependance of Sex on the Living Arrangement

To get the dependence as they both are categorical in nature, we prefer using the Chi Square Test of Independence to test the significance of one on the other.

$\mathbf{H_0}$: *There is no significant dependence between sex and living arrangement*
$\mathbf{H_1}$: *There is significant dependence between sex and living arrangement*

```
#Does a Chi Square Test Of Independence

chisq.test(student.mat$sex,student.mat$Pstatus)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  student.mat$sex and student.mat$Pstatus
## X-squared = 0.090428, df = 1, p-value = 0.7636
```

Here we see the P-value of the test comes out to be 0.76 which is greater than 0.05 which is the threshold value to reject the Null Hypothesis and shift to the alternate hypothesis.

Result : *We seek to determine if there is a significant dependence of sex on the living arrangement'. 395 Students from both the schools combined were used as a sample. The sample population constituted 208* *female and 187 male candidates. A Chi Square Test revealed the X^2 value of 0.09, degree of freedom to be 1 and the p-value to be 0.76. This means that we reject the null hypothesis and select the alternate hypothesis that says that there is a significant dependence between the 2 metrics.*

Difference in Male & Female popoulaion with respect to marks

> We will work this out using a Independant Sample T-Test as the sample space is the same but we have divided them into groups to find if gender influences marks.
> Here we will use all the marks namely G1, G2 & G3 to understand if different period marks have different effect.

We will use the same hypothesis for all the three tests between the separated population.

$\mathbf{H_0}$: *There is no difference between the marks of the corresponding periods for male and female*
$\mathbf{H_1}$: *There is difference between the marks of the corresponding periods for male and female*

```
pulled_df <- as.data.frame(as.data.frame(student.mat)[,c("sex","G1","G2","G3")]) #Making a selective
gender.groupings <- group_by(pulled_df, sex)
get_summary_stats(gender.groupings) #getting summary
```

```
## # A tibble: 6 x 14
##   sex   variable     n   min   max median    q1    q3   iqr   mad  mean    sd
##   <chr> <chr>    <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 F     G1         208     4    19     10     8    13     5  2.96 10.6   3.23
## 2 F     G2         208     0    18     10     8    13     5  2.96 10.4   3.64
## 3 F     G3         208     0    19     10     8    13     5  4.45  9.97  4.62
## 4 M     G1         187     3    19     11     9    14     5  4.45 11.2   3.39
## 5 M     G2         187     0    19     11     9    14     5  2.96 11.1   3.87
## 6 M     G3         187     0    20     11     9    14     5  4.45 10.9   4.50
## # ... with 2 more variables: se <dbl>, ci <dbl>
```

```
#Helps find out outliers in the corresponding columns
identify_outliers(gender.groupings,G1)
```

```
## [1] sex        G1         G2         G3         is.outlier is.extreme
## <0 rows> (or 0-length row.names)
```

```
identify_outliers(gender.groupings,G2)
```

```
## # A tibble: 13 x 6
##    sex      G1    G2    G3 is.outlier is.extreme
##    <chr> <int> <int> <int> <lgl>      <lgl>
##  1 F        12     0     0 TRUE       FALSE
##  2 F         8     0     0 TRUE       FALSE
##  3 F        11     0     0 TRUE       FALSE
##  4 F         4     0     0 TRUE       FALSE
##  5 F         7     0     0 TRUE       FALSE
##  6 F         6     0     0 TRUE       FALSE
##  7 F         7     0     0 TRUE       FALSE
##  8 M         9     0     0 TRUE       FALSE
##  9 M        10     0     0 TRUE       FALSE
```

```
## 10 M          5     0     0 TRUE       FALSE
## 11 M          5     0     0 TRUE       FALSE
## 12 M          7     0     0 TRUE       FALSE
## 13 M          6     0     0 TRUE       FALSE
```

```
identify_outliers(gender.groupings,G3)
```

```
## # A tibble: 38 x 6
##    sex      G1    G2    G3 is.outlier is.extreme
##    <chr> <int> <int> <int> <lgl>      <lgl>
##  1 F        12     0     0 TRUE       FALSE
##  2 F         8     0     0 TRUE       FALSE
##  3 F        11     0     0 TRUE       FALSE
##  4 F         4     0     0 TRUE       FALSE
##  5 F         6     7     0 TRUE       FALSE
##  6 F         6     7     0 TRUE       FALSE
##  7 F         8     7     0 TRUE       FALSE
##  8 F         6     5     0 TRUE       FALSE
##  9 F         7     0     0 TRUE       FALSE
## 10 F        10     9     0 TRUE       FALSE
## # ... with 28 more rows
```

```
#Does the T-Tests
t.test(male_students$G1, female_students$G1)
```

```
##
##  Welch Two Sample t-test
##
## data:  male_students$G1 and female_students$G1
## t = 1.8237, df = 383.79, p-value = 0.06898
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.04764732  1.26715575
## sample estimates:
## mean of x mean of y
##  11.22995  10.62019
```

```
t.test(male_students$G2, female_students$G2)
```

```
##
##  Welch Two Sample t-test
##
## data:  male_students$G2 and female_students$G2
## t = 1.8077, df = 382.38, p-value = 0.07144
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.060092  1.430978
## sample estimates:
## mean of x mean of y
##  11.07487  10.38942
```

```
t.test(male_students$G3, female_students$G3)
```

```
##
##  Welch Two Sample t-test
##
## data:  male_students$G3 and female_students$G3
## t = 2.0651, df = 390.57, p-value = 0.03958
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.04545244 1.85073226
## sample estimates:
## mean of x mean of y
## 10.914439  9.966346
```

Result: *We did this test to determine if gender influences any test scores. The tests were conducted for marks from all the periods that are namely G1, G2 and G3 respectively. The sample space was populated with a total of 395 students of which 208 are female and 187 are male candidates. We performed t-tests for finding if there is a difference in means of both the groups. In all the three tests pegged to the test period we found that there is a difference in the sample means for all the test periods. This test reveals that gender does influence marks irrespective of the test period in question.*

Difference between the 2 schools

```
paste("Number of Students from GP school", nrow(gp_school))
```

```
## [1] "Number of Students from GP school 349"
```

```
paste("Number of Students from MS school", nrow(ms_school))
```

```
## [1] "Number of Students from MS school 46"
```

> We will use the independant t-test across schools for all the evaluation periods and try to understand if school also makes a difference when it comes to marks.

$\mathbf{H_0}$: *There is -no difference between the marks of the students from different schools of the corresponding periods*
$\mathbf{H_1}$: *There is difference between the marks of the students from different schools of the corresponding periods*

```
get_summary_stats(school.grouping)
```

```
## # A tibble: 6 x 14
##   school variable     n   min   max median    q1    q3   iqr   mad  mean    sd
##   <chr>  <chr>    <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 GP     G1         349     3    19   11       8  13    5     4.45 10.9   3.32
## 2 GP     G2         349     0    19   11       9  13    4     2.96 10.8   3.81
## 3 GP     G3         349     0    20   11       8  14    6     4.45 10.5   4.62
## 4 MS     G1          46     6    19   10.5     8  13    5     3.71 10.7   3.35
## 5 MS     G2          46     5    18   10       8  12.8  4.75  3.71 10.2   3.38
## 6 MS     G3          46     0    19   10       8  12.8  4.75  2.96  9.85  4.24
## # ... with 2 more variables: se <dbl>, ci <dbl>
```

```r
identify_outliers(school.grouping, G1)
```

```
## [1] school     G1         G2         G3         is.outlier is.extreme
## <0 rows> (or 0-length row.names)
```

```r
identify_outliers(school.grouping, G2)
```

```
## # A tibble: 13 x 6
##    school    G1    G2    G3 is.outlier is.extreme
##    <chr>  <int> <int> <int> <lgl>      <lgl>
##  1 GP        12     0     0 TRUE       FALSE
##  2 GP         8     0     0 TRUE       FALSE
##  3 GP         9     0     0 TRUE       FALSE
##  4 GP        11     0     0 TRUE       FALSE
##  5 GP        10     0     0 TRUE       FALSE
##  6 GP         4     0     0 TRUE       FALSE
##  7 GP         5     0     0 TRUE       FALSE
##  8 GP         5     0     0 TRUE       FALSE
##  9 GP         7     0     0 TRUE       FALSE
## 10 GP         6     0     0 TRUE       FALSE
## 11 GP         7     0     0 TRUE       FALSE
## 12 GP         6     0     0 TRUE       FALSE
## 13 GP         7     0     0 TRUE       FALSE
```

```r
identify_outliers(school.grouping, G3)
```

```
## # A tibble: 4 x 6
##   school    G1    G2    G3 is.outlier is.extreme
##   <chr>  <int> <int> <int> <lgl>      <lgl>
## 1 MS         7     6     0 TRUE       FALSE
## 2 MS         6     5     0 TRUE       FALSE
## 3 MS         7     5     0 TRUE       FALSE
## 4 MS         6     5     0 TRUE       FALSE
```

```r
t.test(ms_school$G1, gp_school$G1)
```

```
##
##  Welch Two Sample t-test
##
## data:  ms_school$G1 and gp_school$G1
## t = -0.50699, df = 57.297, p-value = 0.6141
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.3160832  0.7842532
## sample estimates:
## mean of x mean of y
##  10.67391  10.93983
```

```r
t.test(ms_school$G2, gp_school$G2)
```

```
##
##  Welch Two Sample t-test
##
## data:  ms_school$G2 and gp_school$G2
## t = -1.0902, df = 61.128, p-value = 0.2799
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.6624404  0.4892748
## sample estimates:
## mean of x mean of y
##  10.19565  10.78223
```

```
t.test(ms_school$G3, gp_school$G3)
```

```
##
##  Welch Two Sample t-test
##
## data:  ms_school$G3 and gp_school$G3
## t = -0.95555, df = 60.054, p-value = 0.3431
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.9863568  0.7020663
## sample estimates:
## mean of x mean of y
##  9.847826 10.489971
```

Result: *We conducted the test to understand if they type of school has an influence on the marks that for all the periods. The sample space has 349 student from the GP school and 46 from the MS school. We ran T-Tests of independence and we found that the means are not equal. Which means that we reject the null hypothesis of equal means and understand that school has significant influence on the marks a student gets.*

Multiple Regression for Correlation

Between G1,G2 and G3 (Multiple Regression)

```
ggplot(student.mat, aes(x=G1, y=G3)) + geom_point(color=primary) +
geom_smooth(method='lm',color=third)
```

```
ggplot(student.mat, aes(x=G2, y=G3)) + geom_point(color=primary) +
geom_smooth(method='lm',color=third)
```

```
# Multiple Regression
lm(G3 ~ G1+G2, data = student.mat)
```

```
##
## Call:
## lm(formula = G3 ~ G1 + G2, data = student.mat)
##
## Coefficients:
## (Intercept)           G1           G2
##     -1.8300       0.1533       0.9869
```

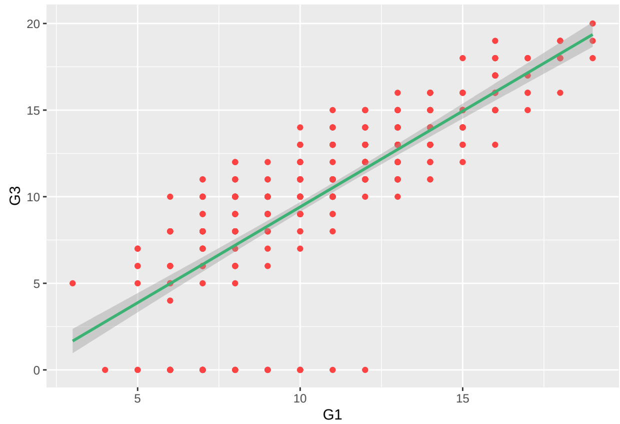

Figure 5: G1 marks vs G3 Marks with Regression Line

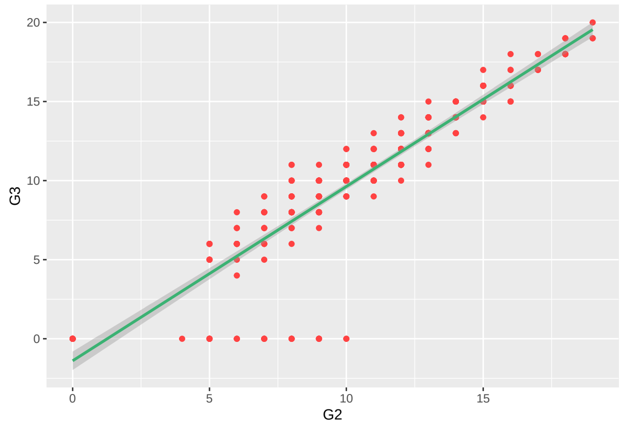

Figure 6: G2 marks vs G3 Marks with Regression Line

Result: *We conducted test to see if there is a correlation between the G1, G2 v/s G3 marks. This would help us understand if there is a dependence on the final marks by the past marks. We did a multiple regression between G1, G2 and G3 to get the regression formula that will help us get the approximate relation between the said variables.* $\hat{y} = -1.83 + (0.1533)G1 + (0.9869)G2$ *came out to be the final multiple regression relationship.*

ANOVA between Family Support, School Support and G3 marks.

$\mathbf{H_0}$: *There is no dependence of Marks on the Family and School Support*
$\mathbf{H_1}$: *There is dependence of Marks on the Family and School Support*

```
identify_outliers(student.mat, G3)
```

```
##  [1] school     sex        age        address    famsize    Pstatus
##  [7] Medu       Fedu       Mjob       Fjob       traveltime studytime
## [13] failures   schoolsup  famsup     paid       nursery    internet
## [19] romantic   freetime   goout      health     absences   G1
## [25] G2         G3         is.outlier is.extreme
## <0 rows> (or 0-length row.names)
```

```
shapiro_test(student.mat, G3)
```

```
## # A tibble: 1 x 3
##   variable statistic        p
##   <chr>        <dbl>    <dbl>
## 1 G3           0.929 8.84e-13
```

```
levene_test(student.mat,G3 ~ famsup*schoolsup)
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic      p
##   <int> <int>     <dbl>  <dbl>
## 1     3   391      3.32 0.0200
```

```
anova_test(student.mat, G3 ~ famsup*schoolsup)
```

```
## ANOVA Table (type II tests)
##
##             Effect DFn DFd     F     p p<.05      ges
## 1           famsup   1 391 0.371 0.543       0.000948
## 2        schoolsup   1 391 2.472 0.117       0.006000
## 3 famsup:schoolsup   1 391 0.739 0.391       0.002000
```

```
model <- lm(G3 ~ famsup+schoolsup,data = student.mat)
anova_school_grouping <- group_by(student.mat, schoolsup)
anova_test(anova_school_grouping, G3 ~ famsup, error = model)
```

```
## # A tibble: 2 x 8
##   schoolsup Effect   DFn   DFd     F     p `p<.05`      ges
## * <chr>     <chr>  <dbl> <dbl> <dbl> <dbl> <chr>      <dbl>
## 1 no        famsup     1   392 0.089 0.765 ""      0.000228
## 2 yes       famsup     1   392 1.02  0.313 ""      0.003
```

Result: *We did the test in order to check if there is any influence of family and school's support on the marks of the student. We performed an Two-Way ANOVA in order to test the hypotheses. The p-values for all the relations came back positive that shows that school support and family support individually also have an influence on the student's marks. Combined family and school support yields a p-value of 0.391 which satisfies* $p > 0.05$ *and hence we reject the null hypothesis and understand that family and school support together have a significant influence on the final period marks. Furthermore the posthoc reveals that there is an impact of school support that influences the student's marks. It shows that there is a significance effect on marks of the student when family supports but that effect increases when the school also supports at the same time.*

Conclusion

We were able to understand that there are a lot of factors to be considered that have a direct or an indirect influence on the marks of the student.