Semester Project

Assignment

DATA ANALYSIS REPORT TO PREDICT

STUDENT'S FINAL GRADE

BY: SHLOKA SARDAL

DATE: 24/09/2023

TABLE OF CONTENTS

- **Introduction**

  - Education is of utmost importance for long term progress of any country. The given dataset contains data of students from two Portuguese Secondary schools namely, GP (Gabriel Pereira) and MS (Mousinho da Silveira). The data set consists of categorical variables like school, sex, address, famsize, Pstatus, Mjob, Fjob, reason, guardian, schoolsup, famsup, paid, activities, nursery, higher, internet and romantic. Numeric variables like age, Medu, Fedu, travel time, study time, failures, famrel, free time, go out, Dalc (Weekday alcohol consumption), Walc (Weekend alcohol consumption), health, absences, G1(first period grade),G2(second period grade) and explained variable like G3(Final grade). In order to predict the final grade results of the students, detailed analysis of the different variables mentioned in the data set is carried out.
  - To predict the accurate final grades following data tools and processes are performed: -
  - **Performing EDA** (Exploratory Data Analysis) enables to find data patterns, spot anomalies, testing hypothesis and assumptions. As a part EDA (Exploratory Data Analysis), **descriptive statistics** of all numeric variables is performed to get a basic information about the **central tendency measures** and the **degree of dispersion of the variables**. **Correlation table** of all the numeric variables is prepared to showcase the relationship between the explanatory variables and the explained variable G3(Final grade).
  - **Pivot tables** and **Pivot charts** are made so that one gets a detailed account of the various variables affecting students final grades and **dashboards** in excel are prepared to present the data visually in a better way.
  - **Multiple Regression model** is run to check the **null hypothesis**, **P-value** (critical level of probability) is calculated to check if the parameter estimates are **statistically significant**.

- **Initial Assumptions and hypotheses**
- Initial Assumptions

  It is assumed that -

- The Data values have **normal distribution** (bell curve).
- The Data values have **homogeneity of variance**.
- Data has a **linear relationship**.

- Hypotheses

- Students with **more study time** are more likely to secure **better grades**.
- Students with **high score in quality of family relationship** are more likely to get **better grades**.
- Students who **go out with friends** very often are more likely to score **lower grades**.
- Students with **high Dalc** (workday alcohol consumption) score are more likely to get **lower grades**.
- Students with **good health score** are more likely to get **better grade**.
- Students with **more absences** are more likely to secure **lower grades**.
- Students with **good score in G1**(First period grade) are more likely to secure **better G3**(Final grade).
- Students with **good score in G2**(Second period grade) are more likely to secure **better G3**(Final grade).

- **Exploratory Data Analysis**

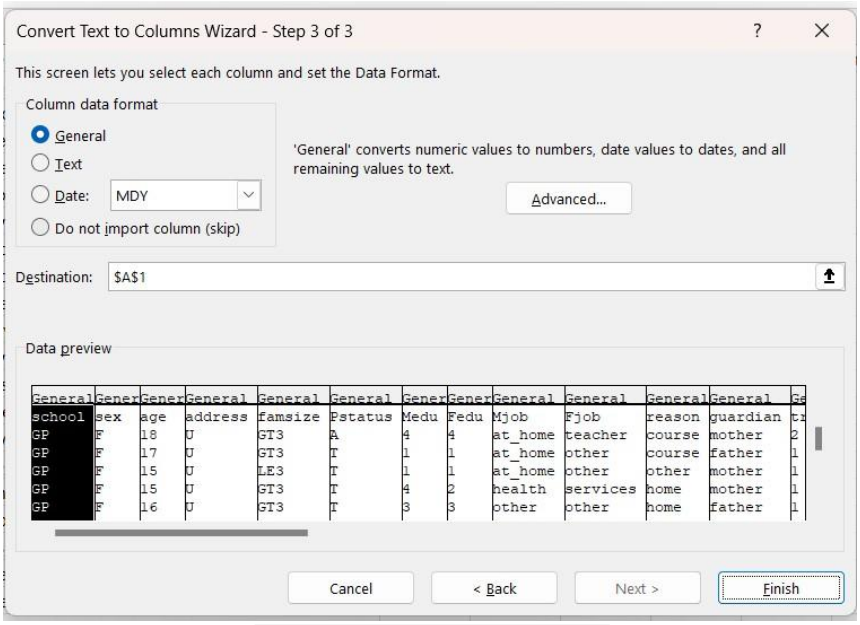


Figure 1

➢ As seen in Figure 1, first the data set is converted from text to columns in order to further work on it.

❖ **<u>Descriptive Statistics</u>**

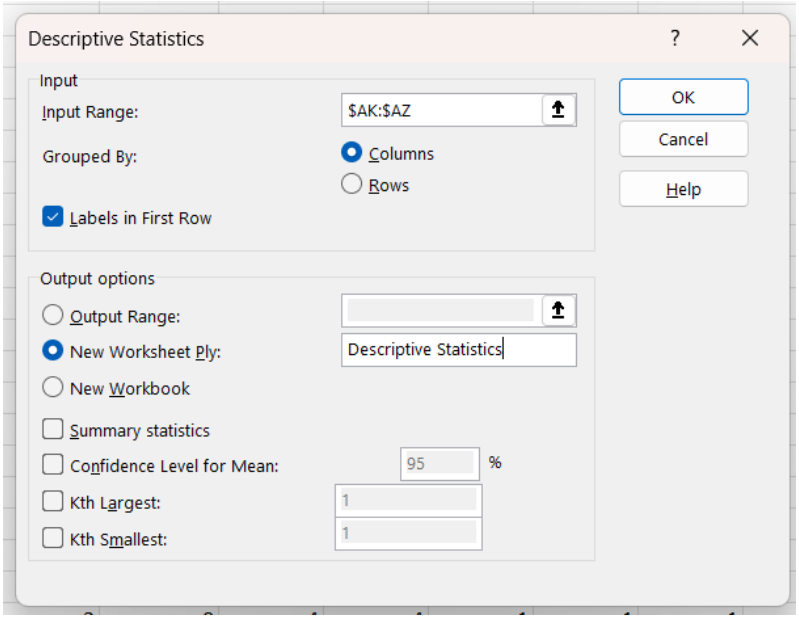


Figure 2

➢ Configuration of Descriptive Statistics of numeric variables is seen in Figure 2.

1) Age

age	
Mean	16.6962
Standard Error	0.064205
Median	17
Mode	16
Standard Deviation	1.276043
Sample Variance	1.628285
Kurtosis	-0.00122
Skewness	0.46627
Range	7
Minimum	15
Maximum	22
Sum	6595
Count	395

Figure 3

The table in Figure 3, states the descriptive statistics for the variable Age.

- Average age of the group of students is 16 years old.
- Standard Error is 0.064, the closer the standard error value is to zero the more accurate the mean value is.
- Median is 17
- Mode is 16, which means most of the students are 16 years old.
- Standard Deviation is 1.276
- Standard Variance is 1.628
- Kurtosis is -0.001 indicating the distribution is not normal and has a slightly flatter curve.
- Skewness is 0.466 indicates the distribution is relatively symmetric and positively skewed.
- Range is 7 (22-15)
- Minimum age is 15
- Maximum age is 22
- Total count of Students is 395.

2) Medu (Mother's Education) and Fedu (Father's education)

Medu		*Fedu*	
Mean	2.749367	Mean	2.521519
Standard Error	0.055082	Standard Error	0.054753
Median	3	Median	2
Mode	4	Mode	2
Standard Deviatic	1.094735	Standard Deviat	1.088201
Sample Variance	1.198445	Sample Varianc	1.18418
Kurtosis	-1.09001	Kurtosis	-1.19854
Skewness	-0.31838	Skewness	-0.03167
Range	4	Range	4
Minimum	0	Minimum	0
Maximum	4	Maximum	4

Figure 4

The table in Figure 4, states the descriptive statistics for the variables Medu (Mother's Education) and Fedu (Father's education)

- The average of Medu is 2.749 and average of Fedu is 2.521 which means on an average students mother's and father's education is from 5th grade to 9th grade.
- Standard Error of Medu is 0.055 whereas Standard Error of Fedu is 0.054
- Median of Medu is 3 whereas median of Fedu is 2
- Mode of Medu is 4 which means most of the mothers of the students have gained higher education, whereas Mode of Fedu is 2, which means most of the students fathers education is 5th grade to 9th grade.
- Standard Deviation of Medu is 1.094 whereas Standard Deviation of Fedu 1.088, in case of Medu data points are more dispersed from the mean as compared to Standard Deviation of Fedu.
- Sample Variance of Medu is 1.198, whereas Sample Variance of Fedu is 1.184.

- Kurtosis of Medu -1.090, indicating a flatter curve than normal distribution, similarly Kurtosis of Fedu is -1.198 indicating a flatter curve than the normal distribution.
- The Skewness of Medu is -0.31, while skewness of Fedu is -0.031 both indicating the distribution is relatively symmetric but negatively skewed.
- Range of both Medu and Fedu is 4 (4-0)
- Minimum education of both Medu and Fedu is 0 , indicating minimum education of the mother and father of a student is none.
- The maximum education of both Medu and Fedu is 4 which means, maximum education of the mother and father of a student is higher education.

3) Travel time (home to school) and Study time (weekly study time)

traveltime		*studytime*	
Mean	1.448101	Mean	2.035443
Standard Error	0.035095	Standard Error	0.042227
Median	1	Median	2
Mode	1	Mode	2
Standard Deviation	0.697505	Standard Deviation	0.83924
Sample Variance	0.486513	Sample Variance	0.704324
Kurtosis	2.34419	Kurtosis	-0.01443
Skewness	1.607029	Skewness	0.632142
Range	3	Range	3
Minimum	1	Minimum	1
Maximum	4	Maximum	4

Figure 5

The table in Figure 5, states the descriptive statistics for the variables travel time and study time

- Mean of travel time is 1.448, which means the average travel time of the students from home to school is less than 15 minutes, while mean of study time is 2.035, which means on an average each student studies for 2 to 5 hours per week.
- Standard Error of travel time is 0.035, while Standard Error of study time is 0.042.
- Median of travel time is 1, while median of study time is 2.
- Mode of travel time is 1 which means most of the students travel to school in less than 15 minutes, while mode of study time is 2, most of the students study for 2 to 5 hours per week.
- Standard Deviation of travel time is 0.697, indicating the data points are clustered around the mean, while Standard Deviation of study time is 0.839, indicates that the data points are clustered around the mean.
- Sample Variance of travel time is 0.486, while Sample Variance of study time is 0.704.
- Kurtosis is 2.344, positive kurtosis means the distribution is more peaked than normal distribution, while kurtosis of study time is -0.014, indicates that the curve is flatter than the normal distribution curve.

- Skewness of travel time is 1.607, which means the distribution is heavily skewed, while the skewness of study time is 0.632, which indicates the distribution is moderately skewed. Both the distributions are skewed on the right side.
- Range of both travel time and study time is 3 (4-1)
- Minimum of both travel time and study time is 1, which means minimum travel time by a student from home to school is less than 15 minutes, while minimum weekly study time for each student is less than 2 hours.
- Maximum of both travel time and study time is 4, which means maximum travel time by a student home to school is more than 1 hour, while maximum weekly study time, each student is more than 10 hours.

4) Failures

Failures	
Mean	2.703797
Standard Error	0.034229
Median	3
Mode	3
Standard Deviation	0.680296
Sample Variance	0.462803
Kurtosis	2.127109
Skewness	-1.98268
Range	2
Minimum	1
Maximum	3

Figure 6

The table in Figure 6, states the descriptive statistics for the variable Failures.

- Mean is 2.703, the average number of past class failures is 2.
- Standard Error is 0.034.
- Median is 3.
- Mode is 3, most of the students have 3 past class failures.
- Standard Deviation is 0.680, the data points are clustered around the mean.
- Sample Variance is 0.462
- Kurtosis is 2.127, indicates a peaked curve of distribution than the normal distribution.
- Skewness is -1.982, indicates the distribution is negative and heavily skewed.
- Range 2 (3 -1)
- Minimum 1, indicates minimum number of past class failures is 1.
- Maximum 3, indicates maximum number of past class failures is 3.

5) Famrel (Family relationships)

famrel	
Mean	3.944304
Standard Error	0.045116
Median	4
Mode	4
Standard Deviation	0.896659
Sample Variance	0.803997
Kurtosis	1.139772
Skewness	-0.95188
Range	4
Minimum	1
Maximum	5

Figure 7

The table in Figure 7, states the descriptive statistics for the variable Famrel (Family relationships)

- Mean is 3.944, indicates that the average quality of family relationships of the students is good.
- Standard Error is 0.045, the closer the standard error value is to zero the more accurate the mean value is.
- Median is 4.
- Mode is 4, most of the students have good quality family relationships.
- Standard Deviation is 0.896.
- Sample Variance is 0.803.
- Kurtosis is 1.139 indicates a peaked distribution curve as compared to normal distribution.
- Skewness is -0.951, indicates the distribution is negative and moderately skewed.
- Range is 4 (5 -1)
- Minimum is 1, indicates minimum number of quality of family relationships is 1 which is very low.
- Maximum is 5, indicates maximum number of quality of family relationships is 5 which is very high.

6) Free time

freetime	
Mean	3.235443
Standard Error	0.050258
Median	3
Mode	3
Standard Deviation	0.998862
Sample Variance	0.997725
Kurtosis	-0.30181
Skewness	-0.16335
Range	4
Minimum	1
Maximum	5

Figure 8

The table in Figure 8, states the descriptive statistics for the variable free time

- Mean is 3.235, indicates on average the students have quite good free time.
- Standard Error is 0.050.
- Median is 3.
- Mode is 3, which means most of the students have a good amount of free time.
- Standard Deviation is 0.998.
- Sample Variance is 0.997.
- Kurtosis is -0.301, indicating the distribution curve flatter than the normal distribution.
- Skewness is -0.163, indicates that distribution is symmetric and negatively skewed.
- Range is 4 (5-1)
- Minimum is 1, indicates minimum free time available to students is 1(very low)
- Maximum is 5, indicates maximum free time available to students is 5 (very high)

7) Go out

goout	
Mean	3.108861
Standard Error	0.056015
Median	3
Mode	3
Standard Deviation	1.113278
Sample Variance	1.239388
Kurtosis	-0.77025
Skewness	0.116502
Range	4
Minimum	1
Maximum	5

Figure 9

The table in Figure 9, states the descriptive statistics for the variable go out

- Mean is 3.108, which means on average students spend good time out with their friends.
- Standard Error is 0.056.
- Median is 3.
- Mode is 3, most of the students go out quite often with their friends.
- Standard Deviation is 1.113, indicates the distribution is dispersed.
- Sample Variance is 1.239.
- Kurtosis is -0.770, indicates the distribution curve is flatter than the normal distribution curve.
- Skewness is 0.116, indicates relatively symmetric distribution.
- Range is 4 (5-1)
- Minimum is 1, the minimum score of go out with friends is 1(very low).
- Maximum is 5, the maximum score of going out with friends is 5 (very high)

8) **Dalc (work day alcohol consumption) and Walc (weekend alcohol consumption of students**)

S	T	U	V
Dalc		*Walc*	
Mean	1.481013	Mean	2.291139
Standard Error	0.044818	Standard Error	0.064801
Median	1	Median	2
Mode	1	Mode	1
Standard Deviation	0.890741	Standard Deviation	1.287897
Sample Variance	0.79342	Sample Variance	1.658678
Kurtosis	4.759492	Kurtosis	-0.79085
Skewness	2.190762	Skewness	0.61196
Range	4	Range	4
Minimum	1	Minimum	1
Maximum	5	Maximum	5

Figure 10

The table in Figure 10, states the descriptive statistics for the variables Dalc and Walc

- Mean of Dalc is 1.481, on an average the alcohol consumption of students on a work day is very low. Whereas Mean of Walc is 2.291, indicates average weekend alcohol consumption of students is not very low.
- Standard Error of Dalc is 0.044, while Standard Error of Walc is 0.064
- Median of Dalc is 1, Median of Walc is 2
- Mode of both Dalc and Walc is 1, most of the students consume very low alcohol on a work day.
- Standard Deviation of Dalc is 0.890, the data points are clustered, while Standard Deviation of Walc is 1.287, indicates data points are dispersed.
- Sample Variance of Dalc is 0.793, whereas Sample Variance of Walc is 1.658
- Kurtosis of Dalc is 4.759, indicates a peaked curve of distribution than the normal distribution whereas Kurtosis of Walc is -0.790, indicates the distribution curve is flatter than the normal distribution curve.

- Skewness of Dalc is 2.190, indicates the distribution is heavily skewed, whereas Skewness of Walc is 0.611 , indicates the distribution is moderately skewed.
- Range of both Dalc and Walc is 4 (5 -1)
- Minimum of both Dalc and Walc is 1, indicates the minimum alcohol consumption of students on a workday is 1(very low).
- Maximum of both Dalc and Walc is 5, indicates the maximum alcohol consumption of students on a workday is 5(very high).

9) Health (current health status of students)

health	
Mean	3.55443
Standard Error	0.069954
Median	4
Mode	5
Standard Deviation	1.390303
Sample Variance	1.932944
Kurtosis	-1.01408
Skewness	-0.4946
Range	4
Minimum	1
Maximum	5

Figure 11

The table in Figure 11, states the descriptive statistics for the variable health

- Mean is 3.554, indicates the current average health status of students is good.
- Standard Error is 0.069
- Median is 4
- Mode is 5, indicates the health status of most of the students is very good.
- Standard Deviation is 1.390, the distribution is dispersed.
- Sample Variance is 1.932
- Kurtosis is -1.014, indicates the flatter distribution curve as compared to normal distribution curve.
- Skewness is -0.494, indicates distribution is symmetric and negatively skewed.
- Range is 4 (5 -1)
- Minimum is 1, indicates minimum health status score of students is 1 (very low)
- Maximum is 5, indicates maximum health status score of students is 5 (very high)

10) Absences (number of school absences)

absences	
Mean	5.708861
Standard Error	0.402679
Median	4
Mode	0
Standard Deviatior	8.003096
Sample Variance	64.04954
Kurtosis	21.71915
Skewness	3.671579
Range	75
Minimum	0
Maximum	75

Figure 12

The table in Figure 12, states the descriptive statistics for the variable absences

- Mean is 5.708, on average the number of school absences by students is 5 days.
- Standard Error is 0.402
- Median is 4
- Mode is 0, indicates most of the students zero number of absences.
- Standard Deviation is 8.003, the datapoints are highly dispersed.
- Sample Variance is 64.049
- Kurtosis is 21.719 the distribution curve is highly peaked than the normal distribution curve.
- Skewness is 3.671, indicates the distribution is heavily skewed.
- Range is 75 (75 - 0)
- Minimum is 0, minimum number of absences by students is 0.
- Maximum is 75, maximum number of absences by students is 75.

11) G1 (First period grade), G2(second grade period) and G3(Final grade)

G1		G2		G3	
Mean	10.90886	Mean	10.71392	Mean	10.41519
Standard Error	0.167007	Standard Error	0.189262	Standard Error	0.230517
Median	11	Median	11	Median	11
Mode	10	Mode	9	Mode	10
Standard Deviation	3.319195	Standard Deviatior	3.761505	Standard Deviatio	4.581443
Sample Variance	11.01705	Sample Variance	14.14892	Sample Variance	20.98962
Kurtosis	-0.69383	Kurtosis	0.627706	Kurtosis	0.403421
Skewness	0.240613	Skewness	-0.43165	Skewness	-0.73267
Range	16	Range	19	Range	20
Minimum	3	Minimum	0	Minimum	0
Maximum	19	Maximum	19	Maximum	20

Figure 13

The table in Figure 13, states the descriptive statistics for the variable G1, G2 and G3

- Mean for G1 is 10.908, for G2 is 10.713, for G3 is 10.415 which means average score obtained by students for G1, G2 and G3 is 10.
- Standard Error for G1 is 0.167, for G2 is 0.189, for G3 is 0.230
- Median for G1, G2 and G3 is 11
- Mode for G1 and G3 is 10, most of students have obtained 10 score while Mode for G2 is 9, most of the students obtained 9 as their score.
- Standard Deviation for G1 is 3.319, for G2 is 3.761, for G3 is 4.581 in case of G3 the data points are more dispersed as compared to G1 and G2.
- Sample Variance for G1 is 11.017, for G2 14.148, for G3 20.989.
- Kurtosis is for G1 is -0.693, indicates slightly flatter curve than the normal distribution curve, Kurtosis for G2 is 0.627 the distribution curve is peaked than the normal distribution curve. Kurtosis for G3 is 0.403, the distribution curve is peaked as compared to normal distribution curve.

- Skewness for G1 is 0.240, indicates positive and relatively symmetric distribution, for G2 the skewness is -0.431, indicates negative and relatively symmetric distribution, for G3 the skewness is -0.732, indicates negative distribution and moderately skewed.
- Range for G1 is 16 (19 -3), for G2 is 19 (19 -0), for G3 is 20 (20 -0).
- Minimum for G1 is 3, the minimum score obtained by the students is 3, Minimum for G2 and G3 is 0 minimum score obtained by students is zero.
- The maximum for G1 and G2 is 19, the maximum score obtained by the students is 19 while maximum for G3 is 20, maximum score obtained by students is 20.

❖ **Pivot Tables and Pivot Charts**

1) **Total number of students in the school based on school and age.**

School and age	Count of sex
GP	
15	82
16	104
17	86
18	57
19	18
20	1
22	1
MS	
17	12
18	25
19	6
20	2
21	1
Grand Total	395

Figure 14

School and age	Count of sex
GP	349
MS	46
Grand Total	395

Figure 15

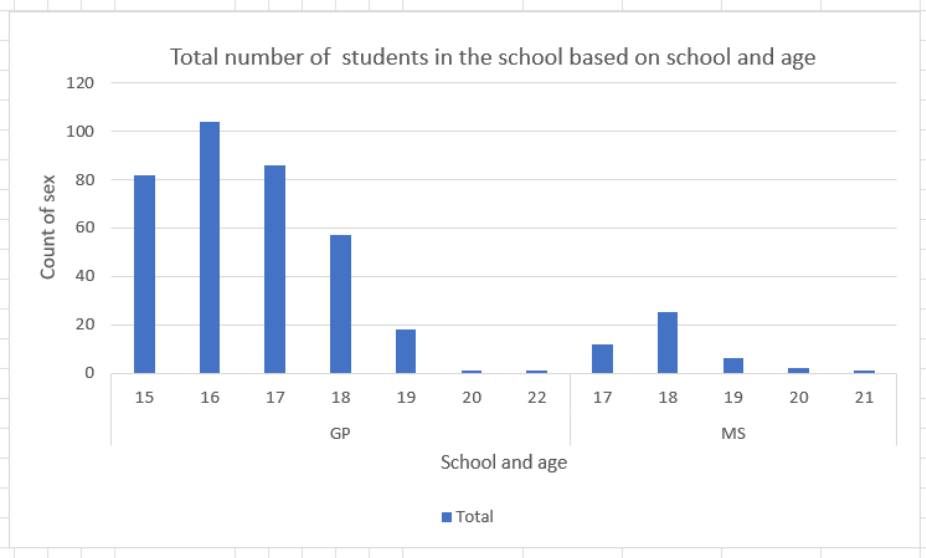


Figure 16

- Pivot table in Figure 14, gives a detailed description of students from both the schools (GP and MS) based on their school and age.
- In Figure 15, the total number of students going to the GP school is 349 and total number of students going to MS school i.e 46 is shown.
- In Figure 16, it is clear from the pivot chart that majority of students go to GP school, also highest number of students from GP school are in the age group of 16 years old. Highest number of students from MS school belong to the age group of 18 years old.

2) Classification of students by address and family size

Address and fam size	Count of sex
R	
GT3	68
LE3	20
U	
GT3	213
LE3	94
Grand Total	**395**

Figure 17

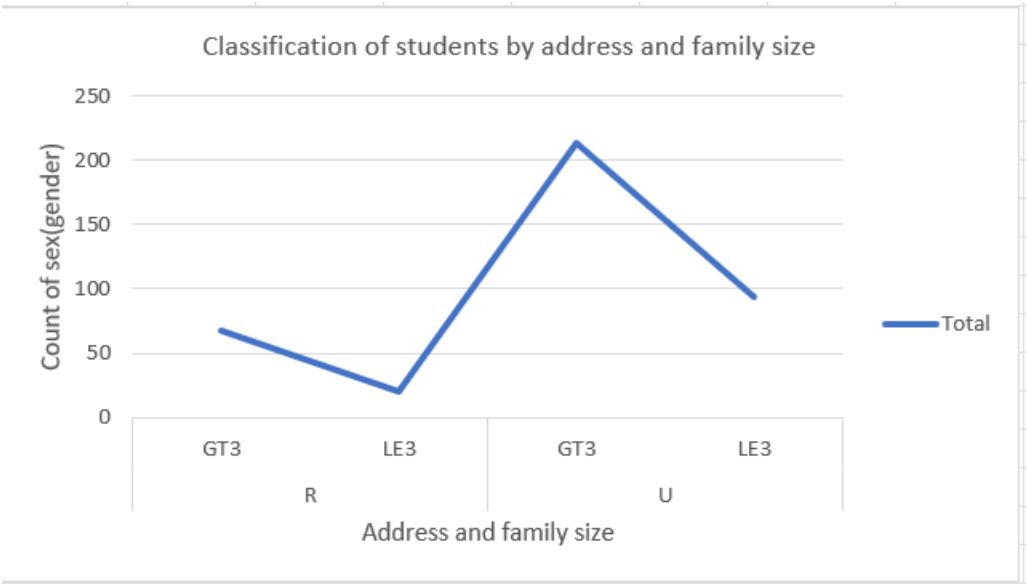


Figure 18

- In Figure 17 it seen that a detailed description of students coming from urban and rural areas based on their family size is given.
- Figure 18 shows a visual presentation in the form a Pivot chart, where it can be seen that highest number of the students come from the urban area and belong to a family size greater than 3.

3) Classification of number of students based on mother's job

M job	Count of sex
at_home	59
health	34
other	141
services	103
teacher	58
Grand Total	**395**

Figure 19

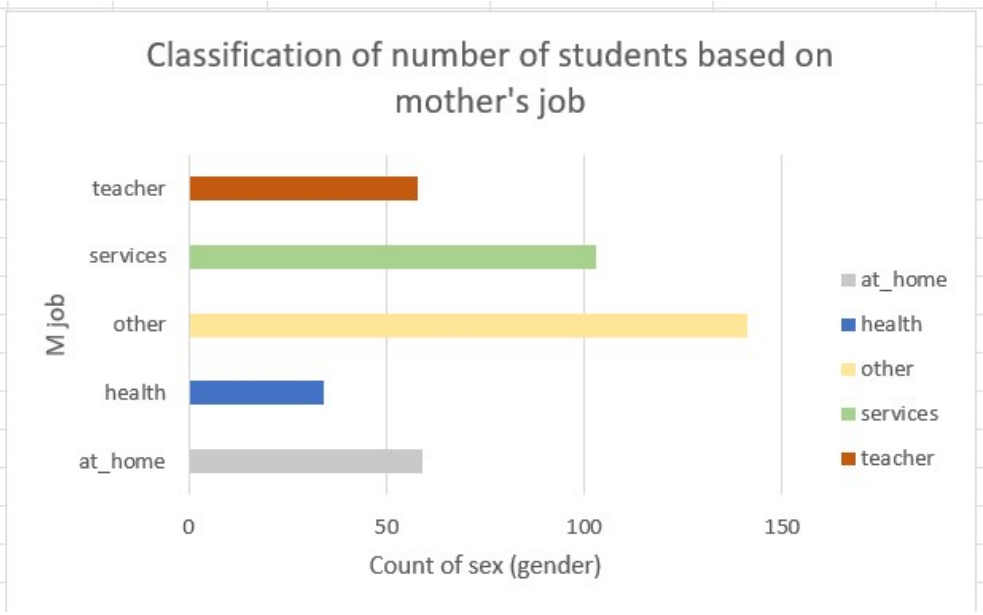


Figure 20

- It is clear from Figure 19 that the number of students whose mother's job is from other category is the highest.
- In figure 20, Pivot Chart shows the visual presentation of classification of students based on mother's job. The number of students whose mother's job is from health category is the lowest.

4) Total number of students based on father's job

Fjob	Count of sex
at_home	20
health	18
other	217
services	111
teacher	29
Grand Total	**395**

Figure 21

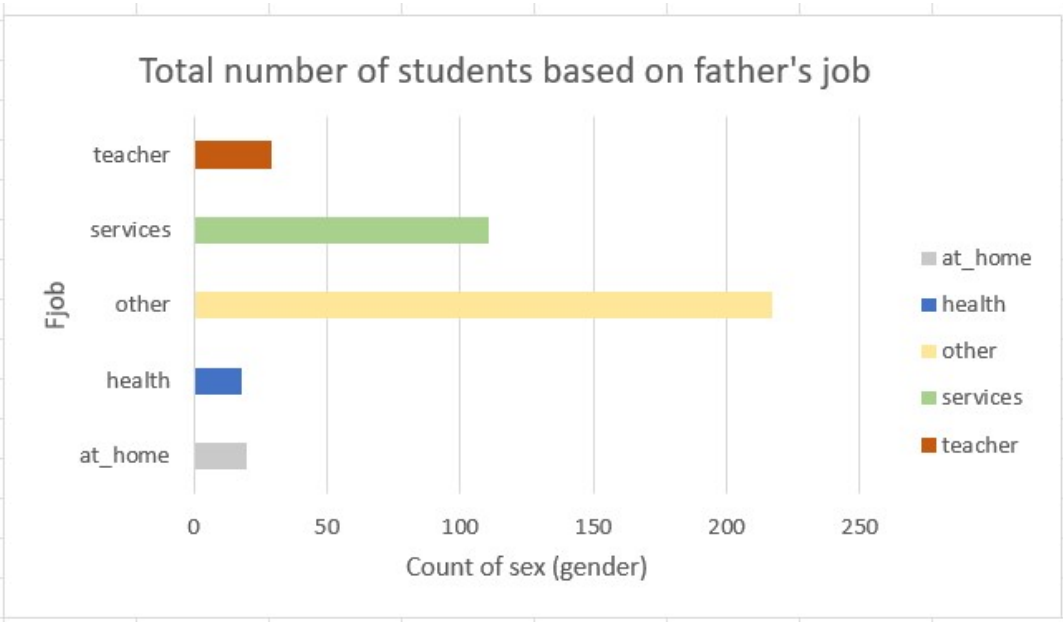


Figure 22

- It is clear from Figure 21 that the number of students whose father's job is from other category is the highest.
- In figure 22, Pivot Chart shows the visual presentation of classification of students based on father's job. The number of students whose father's job is from health category is the lowest.

❖ Dashboard 1

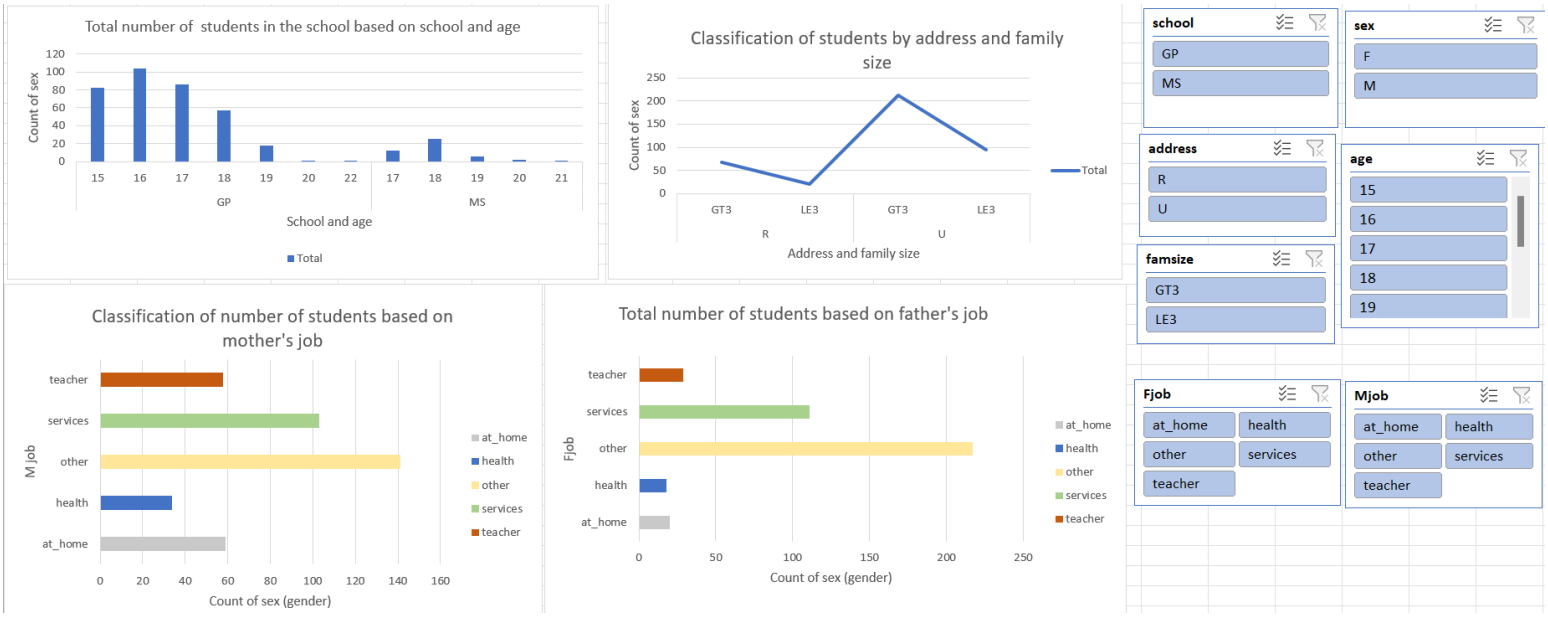


Figure 23

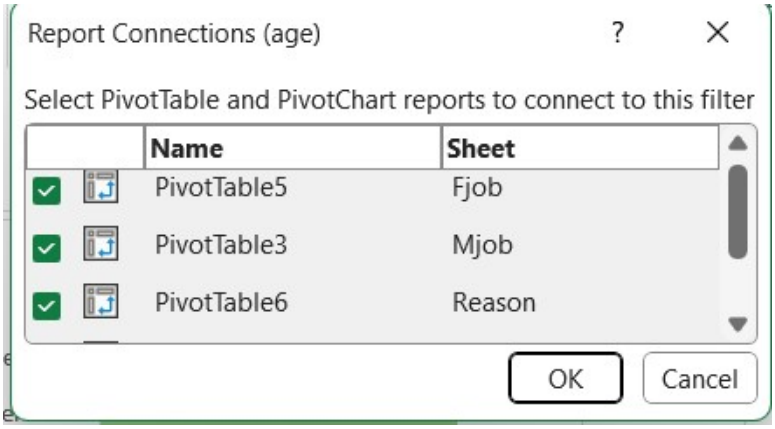


Figure 24

- As it can be seen in Figure 23, a dashboard is created in order to get a overview of the various Pivot Charts.
- First chart in the dashboard presents the total number of students in the school based on age and school.
- Second chart presents the classification of students by address and family size.
- Third chart presents the classification of number of students based on mother's job.
- The fourth chart presents the total number of students based on father's job.
- Figure 24, shows connections established between different pivot charts.

5) Classification of students based on reason.

Reason	Count of sex
course	145
home	109
other	36
reputation	105
Grand Total	**395**

Figure 25

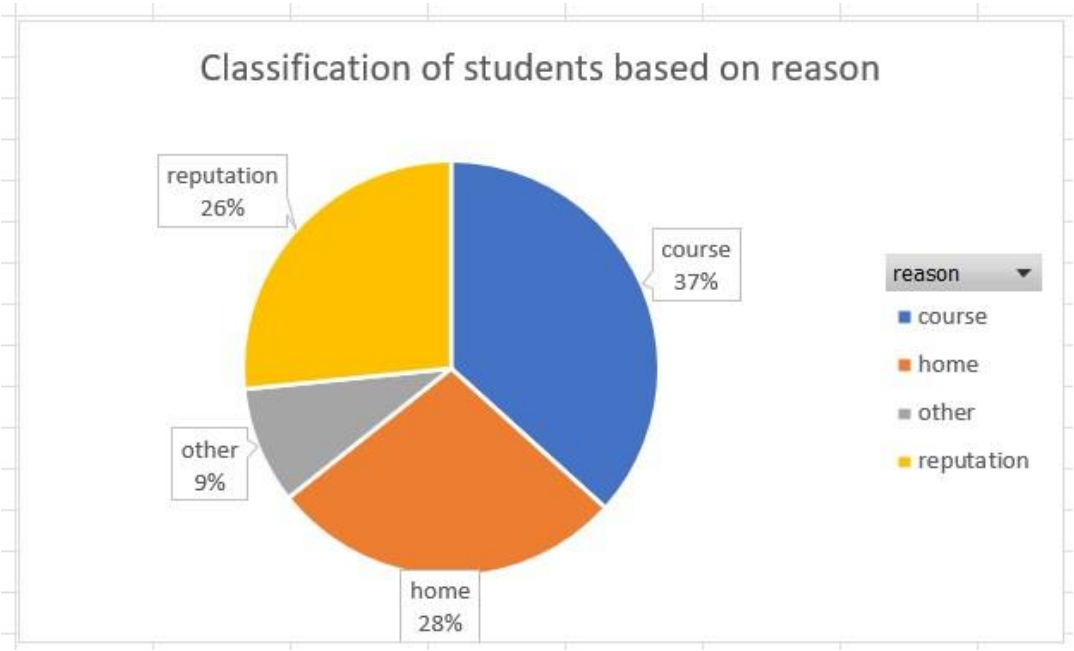


Figure 26

- As seen in Figure 25, detailed classification of students is given based on reason as to why they chose the respective school,145 students have opted for their respective school based on course preference.
- In Figure 26, Pie chart shows the classification of students based on reason. 37% of the students chose the respective school based on course preference followed by home (28%), reputation (26%) and other reasons(9%).

6) Classification of number of students by guardian and G3

Count of sex G3																			
Guardian	**0**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**Grand Total**
father	8		1	4	1	5	6	16	8	9	8	8	6	4	2	4			90
mother	25	1	6	11	7	21	18	37	35	20	19	19	26	11	4	7	5	1	273
other	5				1	6	4	3	4	2	4		1	1		1			32
Grand Total	**38**	**1**	**7**	**15**	**9**	**32**	**28**	**56**	**47**	**31**	**31**	**27**	**33**	**16**	**6**	**12**	**5**	**1**	**395**

Figure 27

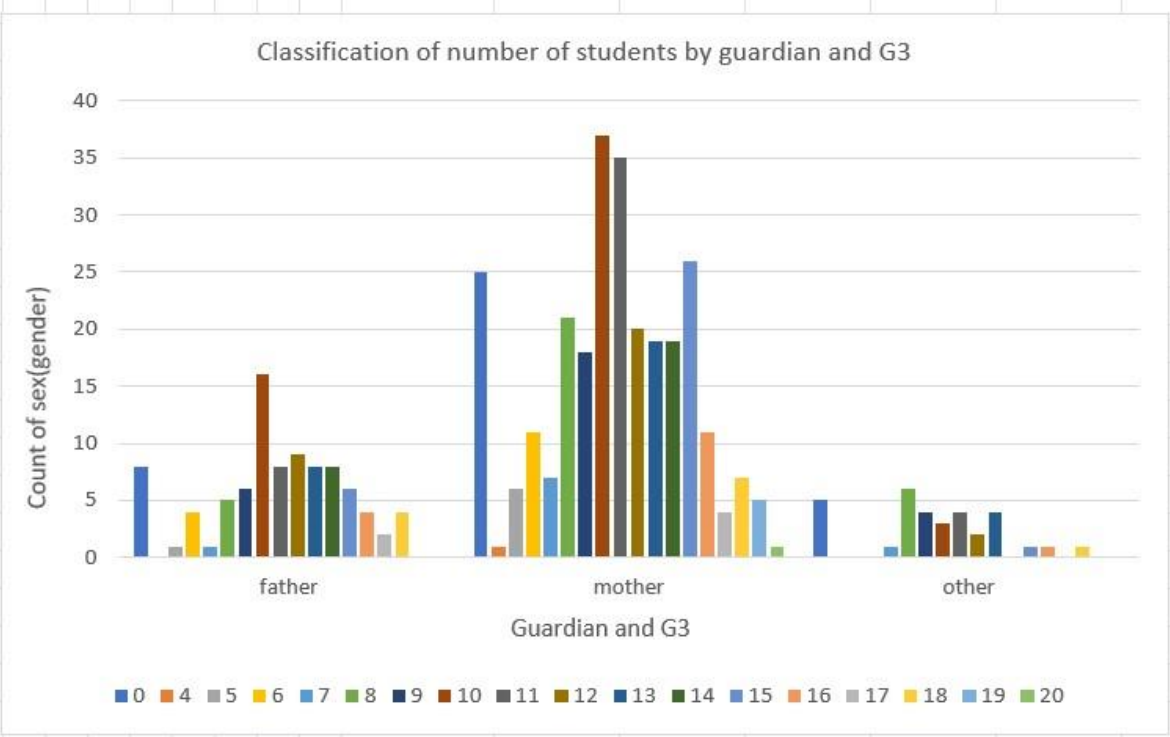


Figure 28

- Figure 27, gives detailed description of the students G3(Final grade) and as to how many students are taken care of by the father, mother or other guardian.
- Figure 28, clearly shows that maximum number of students are taken care of by the mother out of them the highest number of students (37) belong to the category who have scored 10 as G3. Out of the students being taken care of by the father, the highest number of students (16) have got 10 score as well for the G3, while the highest number of students (6) in the other category have scored 8 as G3.

7) Count of students based on family relation and age

Count of sex Age									
Famrel	15	16	17	18	19	20	21	22	Grand Total
1	2	3	1	2					8
2	2	7	5	3	1				18
3	14	20	20	10	4				68
4	40	48	48	43	16				195
5	24	26	24	24	3	3	1	1	106
Grand Total	**82**	**104**	**98**	**82**	**24**	**3**	**1**	**1**	**395**

Figure 29

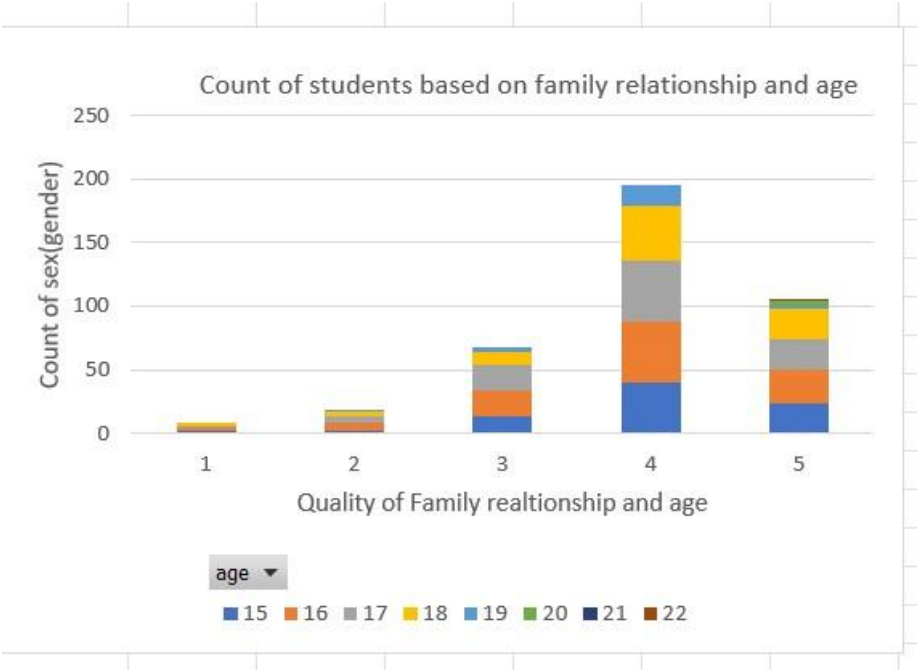


Figure 30

- In Figure 29, it is seen that highest number of students across all age groups belongs to the category of family relationship with score of 4, followed by family relationship with score of five.
- Figure 30, shows visual presentation of family relationship based on age. It can be seen that least number of students belong to category of family relationship with score 1.

8) Count of students based on travel time and age

Count of sex	Age								
Travel time	15	16	17	18	19	20	21	22	Grand Total
1	60	72	61	44	16	2	1	1	257
2	17	23	31	30	5	1			107
3	2	8	4	6	3				23
4	3	1	2	2					8
Grand Total	**82**	**104**	**98**	**82**	**24**	**3**	**1**	**1**	**395**

Figure 31

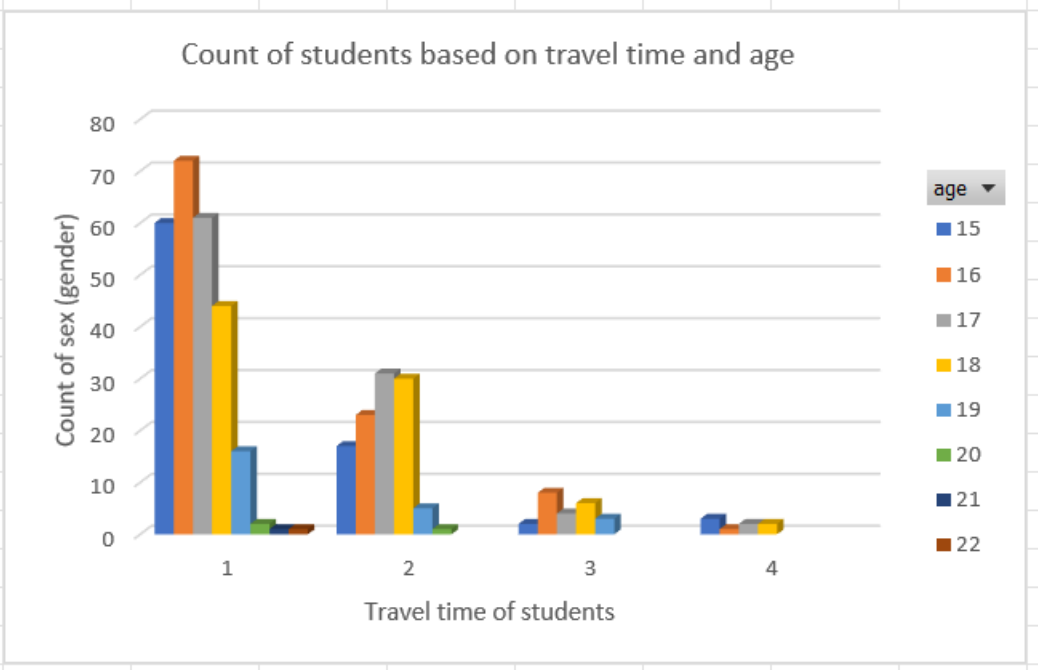


Figure 32

- The table in Figure 31, gives a detailed description of the number of students about their travel time based on age.
- It is clear from the Figure 32, highest number of students from all age groups belong to the travel time category 1, followed by travel time category 2.

❖ Dashboard 2

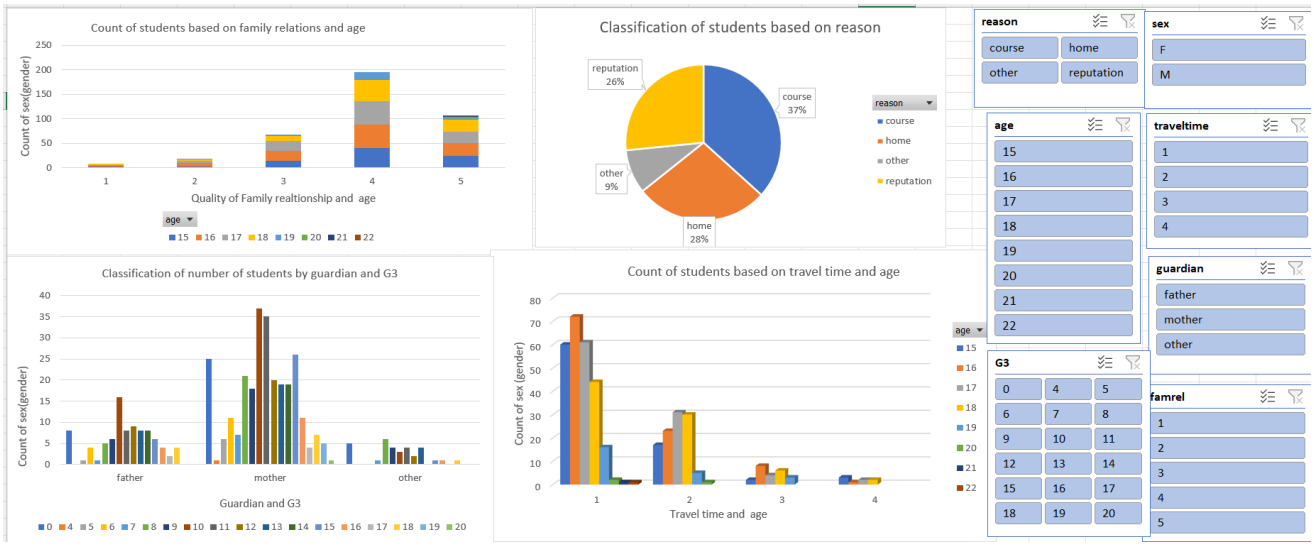


Figure 33

- Figure 33, shows the dashboard wherein various factors affecting the students data are shown in the charts.
- First chart shows Count of students based on family relations and age.
- Second chart shows Classification of students based on reason.
- Third chart shows Classification of number of students by guardian and G3.
- Fourth chart shows Count of students based on travel time and age.

9) Count of students based on schoolsup and G3

Count of sex	G3																		
Schoolsup	0	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Grand Total
no	37	1	5	9	7	26	22	46	38	27	29	26	32	16	5	12	5	1	344
yes	1		2	6	2	6	6	10	9	4	2	1	1		1				51
Grand Total	**38**	**1**	**7**	**15**	**9**	**32**	**28**	**56**	**47**	**31**	**31**	**27**	**33**	**16**	**6**	**12**	**5**	**1**	**395**

Figure 34

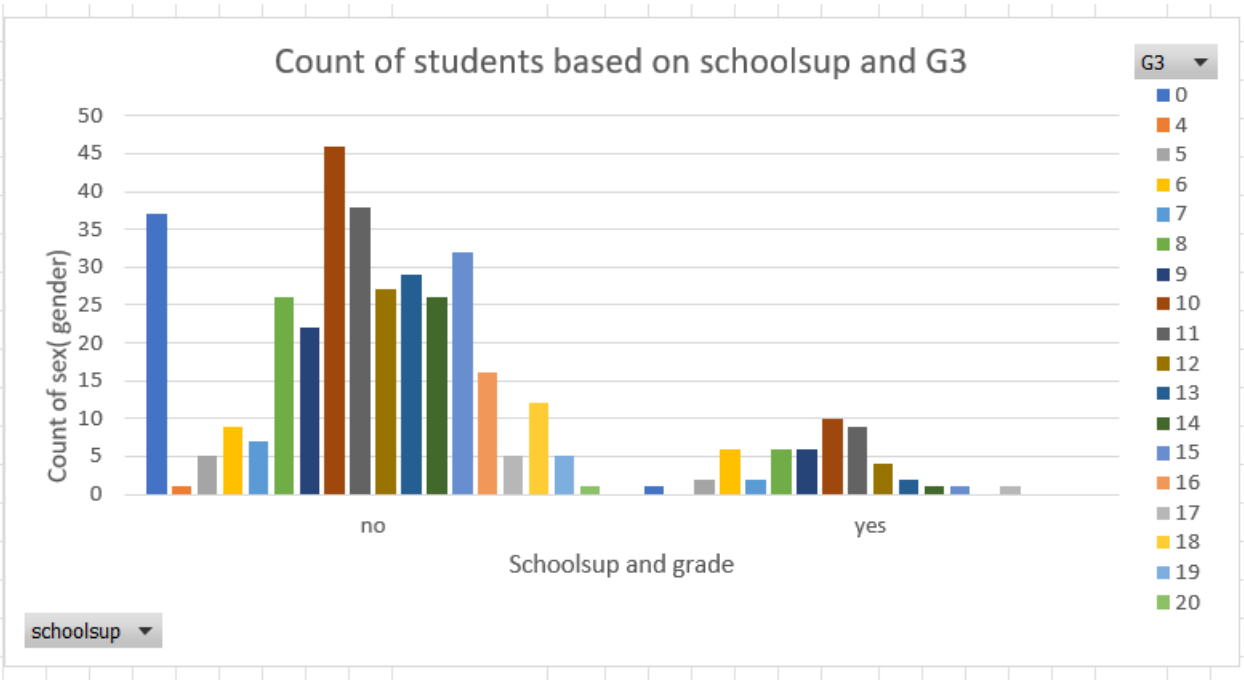


Figure 35

- It is seen from Figure 34 that, majority of the students do not have school support.Out of 395 students only 51 students have school support.
- It is seen from Figure 35 that highest number of students have got 10 score in both the cases that is with or without school support.

10) Count of students based on Famsup and G3

Count of sex	G3																	
Famsup	**0**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**Grand Total**
no	15	1	2	2	4	11	10	24	16	13	15	9	15	7	1	4	3	1	153
yes	23		5	13	5	21	18	32	31	18	16	18	18	9	5	8	2		242
Grand Total	**38**	**1**	**7**	**15**	**9**	**32**	**28**	**56**	**47**	**31**	**31**	**27**	**33**	**16**	**6**	**12**	**5**	**1**	**395**

Figure 36

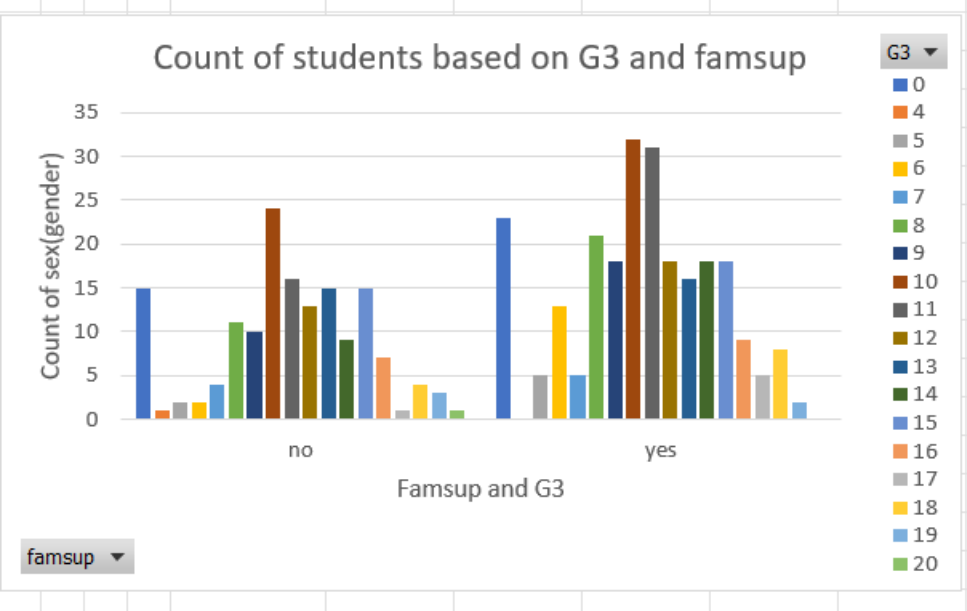


Figure 37

- It is seen in Figure 36 that, 242 students out 395 students have family support.
- It is seen in Figure 37 that, highest number of students with or without family support have scored 10 as G3.

11) Count of students based on G3 and paid classes

Count of sex	G3																		
Paid Classes	0	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Grand Total
no	30		5	9	7	12	16	26	21	19	14	14	14	12	5	6	3	1	214
yes	8	1	2	6	2	20	12	30	26	12	17	13	19	4	1	6	2		181
Grand Total	**38**	**1**	**7**	**15**	**9**	**32**	**28**	**56**	**47**	**31**	**31**	**27**	**33**	**16**	**6**	**12**	**5**	**1**	**395**

Figure 38

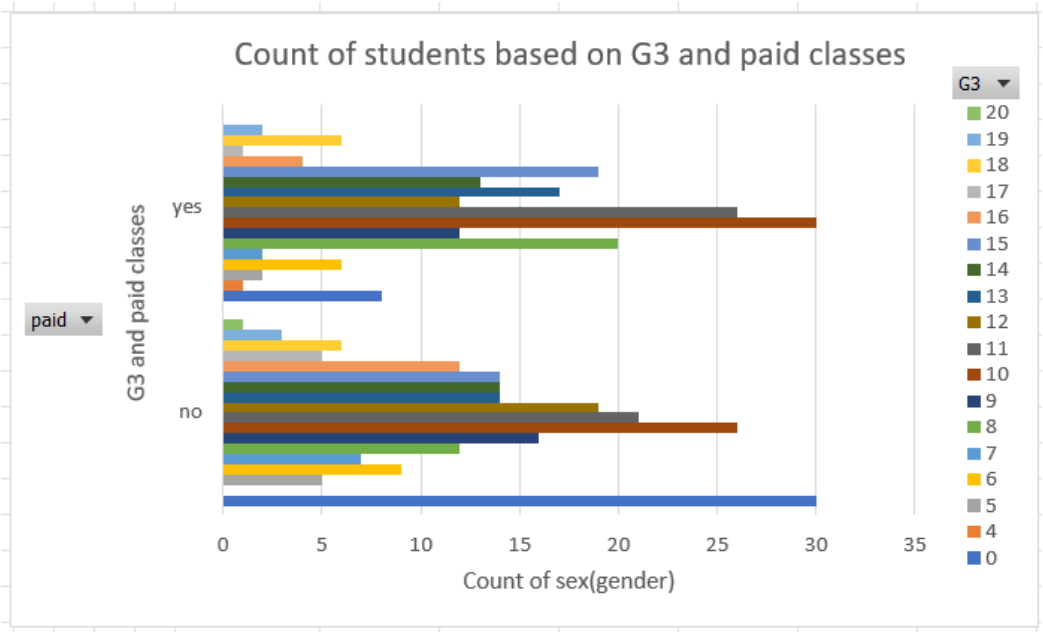


Figure 39

- It is seen in Figure 38, that out of 395 students 214 students have not taken extra paid classes, which means almost 50% of the students have not taken extra paid classes.
- It is further observed in Figure 39, that highest number of students who have taken extra paid classes have got 10 score as their grade(G3), whereas highest number of students who did not take extra paid classes have got 0 score as their grade(G3). Not taking up extra paid classes has clearly impacted the grade of these students.

12) Count of students based on extra -curricular activities and G3

Count of sex	G3																	
extra -curricular activities	0	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Grand Total
no	17	1	4	10	4	19	10	27	24	15	19	10	15	9	2	4	4		194
yes	21		3	5	5	13	18	29	23	16	12	17	18	7	4	8	1	1	201
Grand Total	**38**	**1**	**7**	**15**	**9**	**32**	**28**	**56**	**47**	**31**	**31**	**27**	**33**	**16**	**6**	**12**	**5**	**1**	**395**

Figure 40

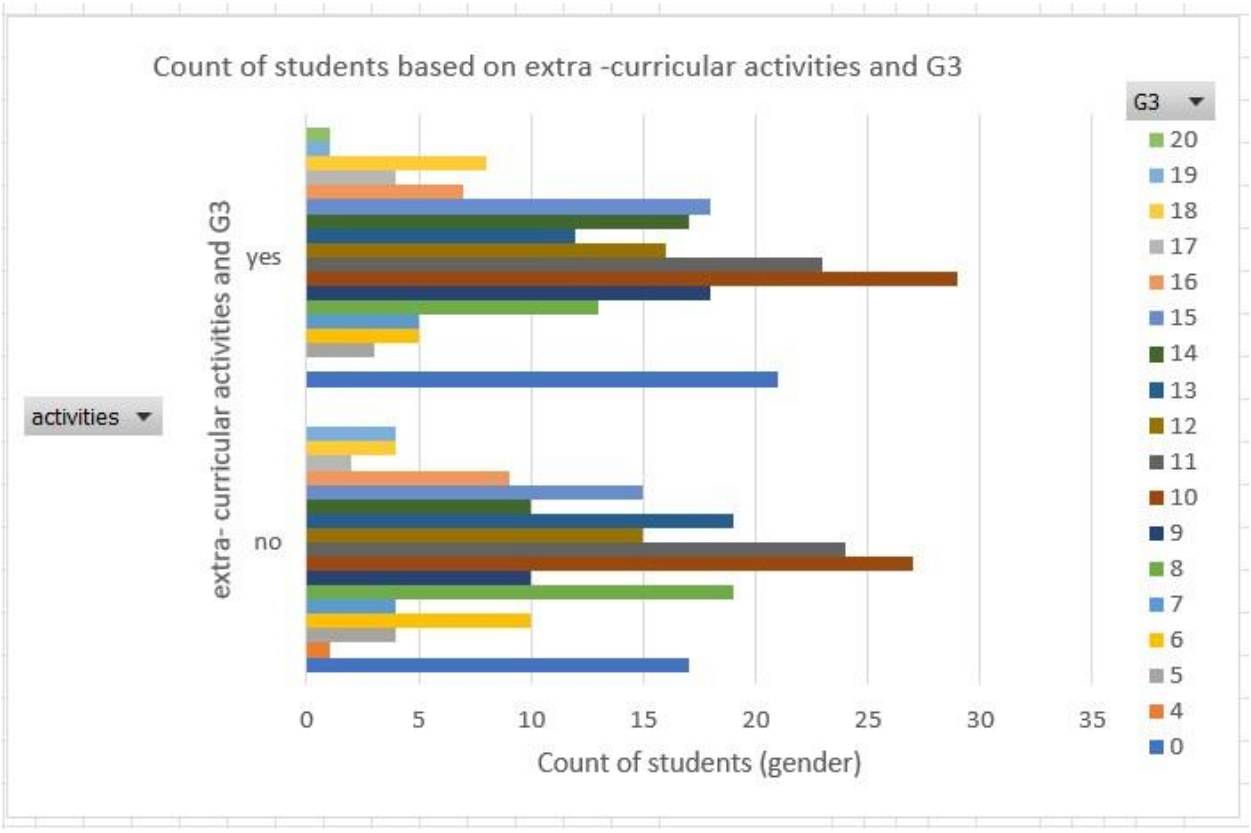


Figure 41

- As seen in Figure 40, out of 395 students 201 students have taken up extra-curricular activities.
- As seen in Figure 41, the highest number of students who have taken up extra-curricular activities have scored 10 as grade(G3), likewise highest number of students who have not taken up extra-curricular activities have scored 10 as grade(G3) as well.

❖ Dashboard 3

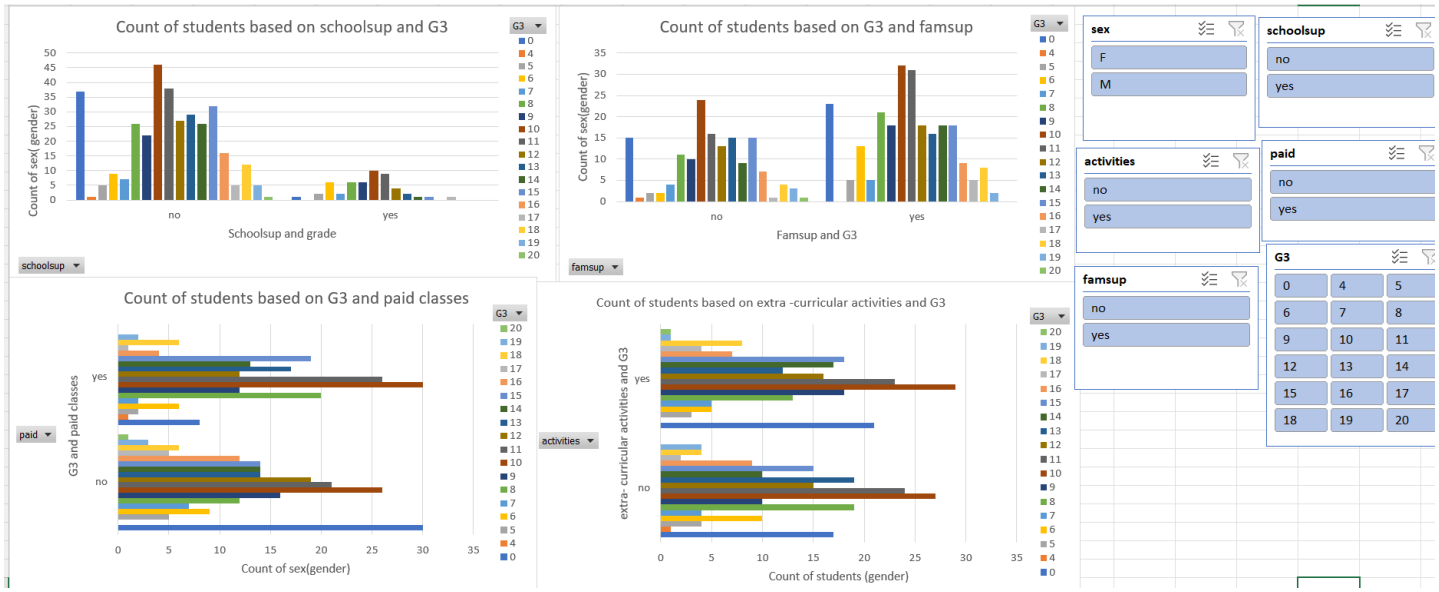


Figure 42

➢ Figure 42, shows the dashboard wherein various factors affecting the students data are shown in the charts.

➢ First chart shows Count of students based on schoolsup and G3.

➢ Second chart shows Count of students based on famsup and G3.

➢ Third chart shows Count of students based on G3 and paid classes.

➢ Fourth chart shows Count of students based on extra-curricular activities and G3.

13) Count of students based on higher education and G3.

Count of sex	G3																		
Higher educ	0	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Grand Total
no	6				1	4	2	3		3	1								20
yes	32	1	7	15	8	28	26	53	47	28	30	27	33	16	6	12	5	1	375
Grand Total	**38**	**1**	**7**	**15**	**9**	**32**	**28**	**56**	**47**	**31**	**31**	**27**	**33**	**16**	**6**	**12**	**5**	**1**	**395**

Figure 43

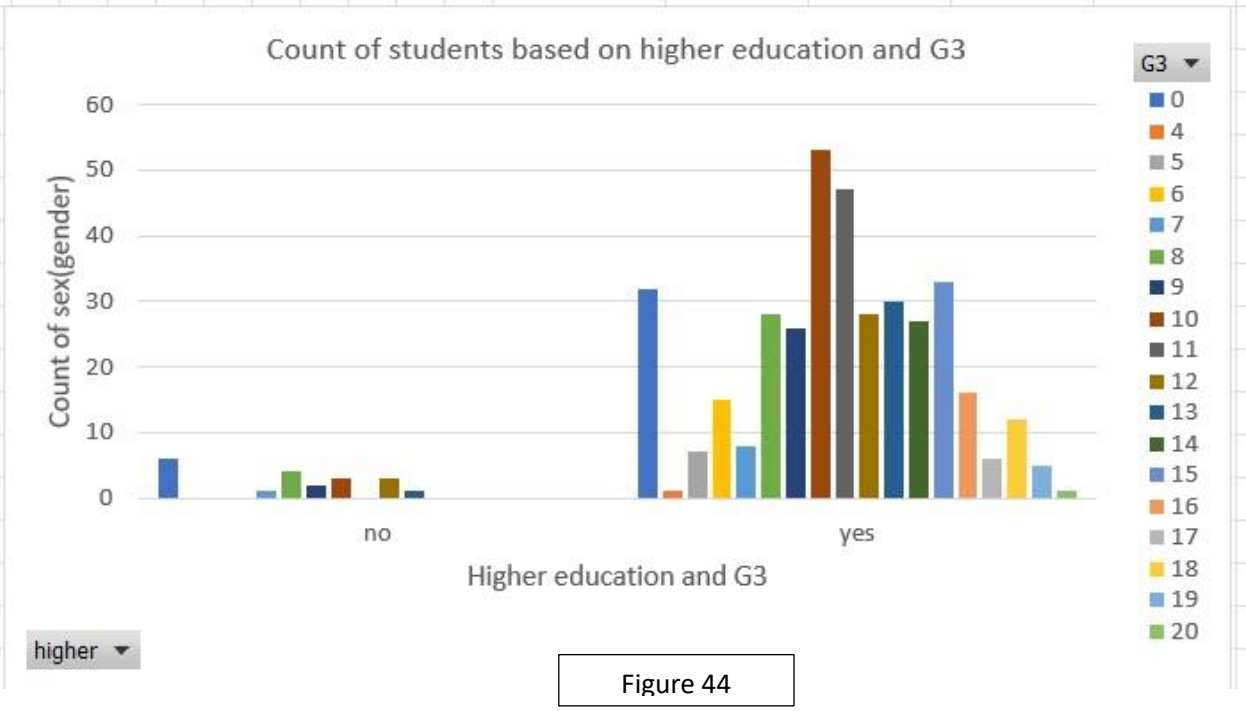


Figure 44

- It is seen in Figure 43, that out of 395 students 375 students have taken up higher education.
- It is observed in Figure 44, that highest number of students (53) who have taken up higher education have scored 10 as their grade(G3), whereas highest number of students (6) who have not taken up higher education have scored 0 as their grade (G3).

14) Count of students based on Internet and G3

Count of sex	G3																		
Internet	**0**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**Grand Total**
no	8		1	5	3	3	6	12	6	8	4	5	2		1	2			66
yes	30	1	6	10	6	29	22	44	41	23	27	22	31	16	5	10	5	1	329
Grand Total	**38**	**1**	**7**	**15**	**9**	**32**	**28**	**56**	**47**	**31**	**31**	**27**	**33**	**16**	**6**	**12**	**5**	**1**	**395**

Figure 45

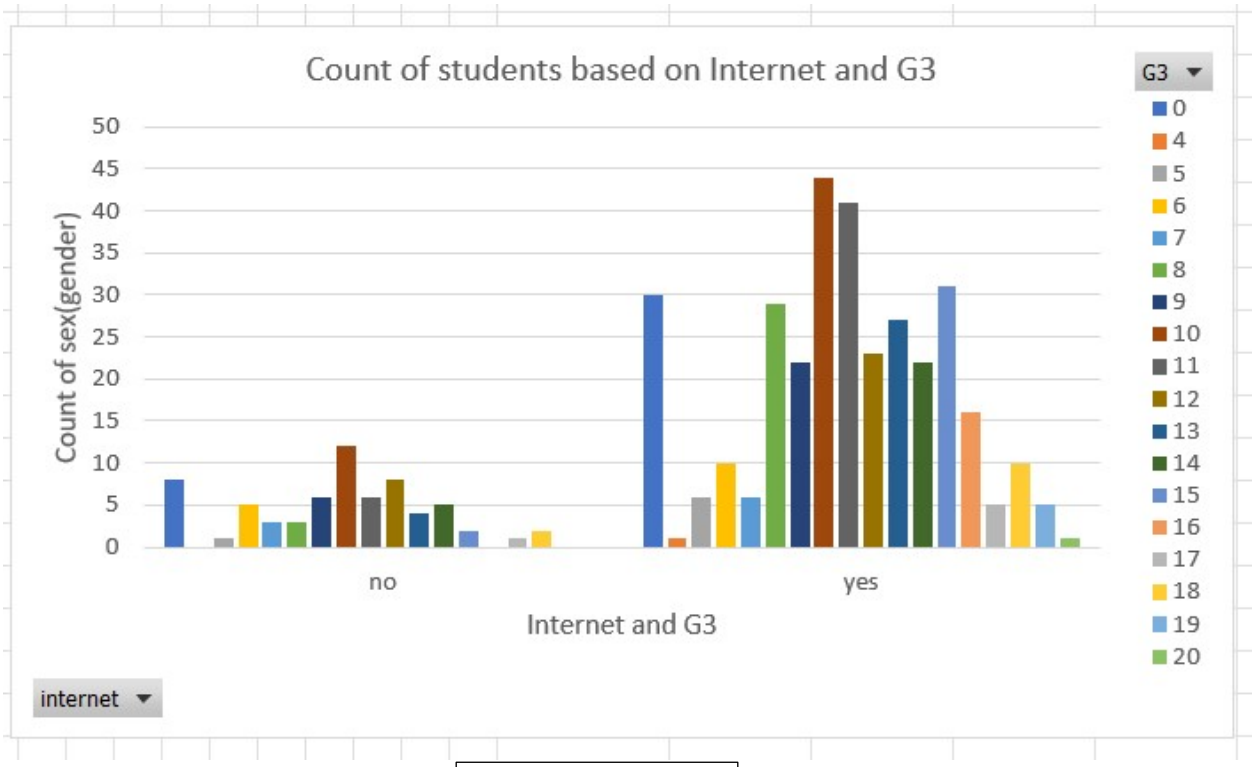


Figure 46

- It is seen in Figure 45,out of 395 students 329 students have access to internet.
- It is seen in Figure 46, that the highest number of students(44) who have access to internet have scored 10,followed by score 11 as their G3 whereas, highest number of students (12) who do not have access to internet have scored 10 as well followed by score 0 as the G3.

15) Count of students based on romantic relationships and G3

Count of sex	G3																		
Romantic	0	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Grand Total
no	18		5	13	4	22	16	45	30	17	19	19	24	9	5	11	5	1	263
yes	20	1	2	2	5	10	12	11	17	14	12	8	9	7	1	1			132
Grand Total	**38**	**1**	**7**	**15**	**9**	**32**	**28**	**56**	**47**	**31**	**31**	**27**	**33**	**16**	**6**	**12**	**5**	**1**	**395**

Figure 47

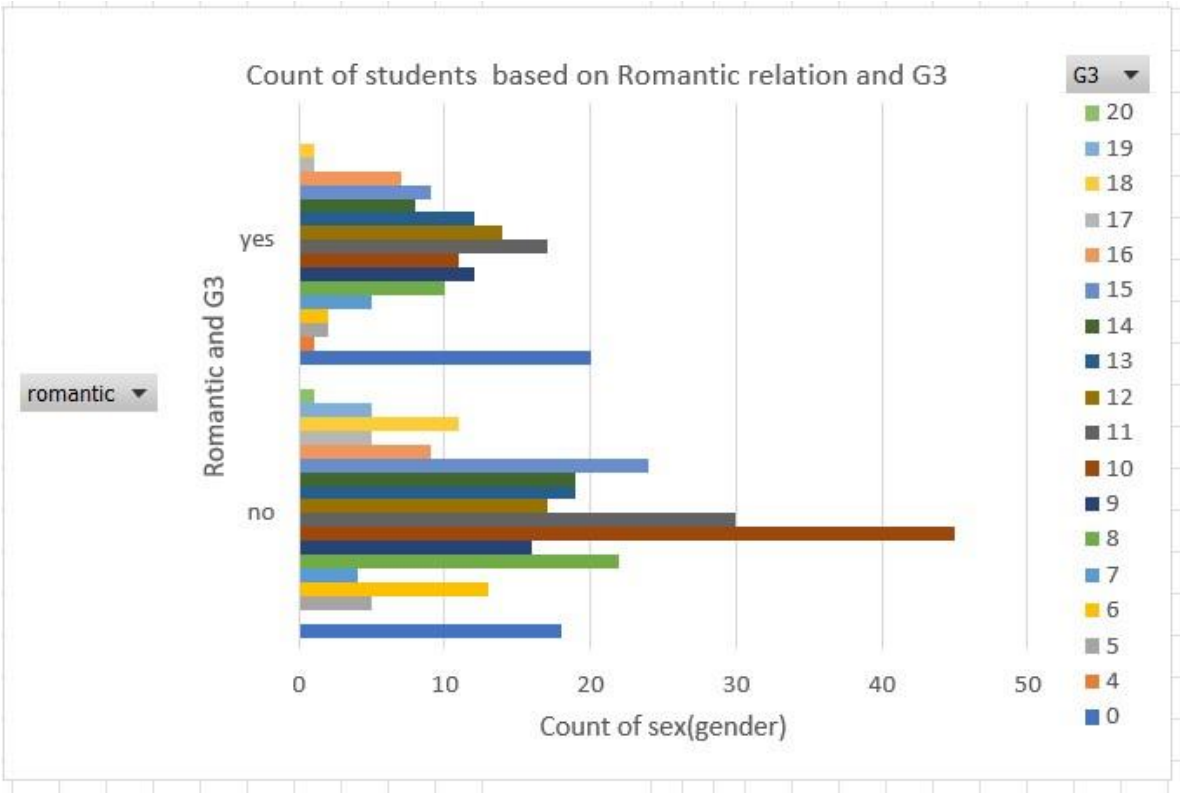


Figure 48

- It is seen in Figure 47, that out of 395 students 263 students are not in a romantic relationship.
- In Figure 48, it is observed that, the highest number of students (45) who are not in a romantic relationship have got 10 score as G3, whereas the highest number of students (20) who are in a romantic relationship have got 0 score as G3.

16) Count of students based on P Status and G3

Count of sex	G3																		
P Status	**0**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**Grand Total**
A	2			3	2	1	3	5	8	3	4	1	3	1	1	2	2		41
T	36	1	7	12	7	31	25	51	39	28	27	26	30	15	5	10	3	1	354
Grand Total	**38**	**1**	**7**	**15**	**9**	**32**	**28**	**56**	**47**	**31**	**31**	**27**	**33**	**16**	**6**	**12**	**5**	**1**	**395**

Figure 49

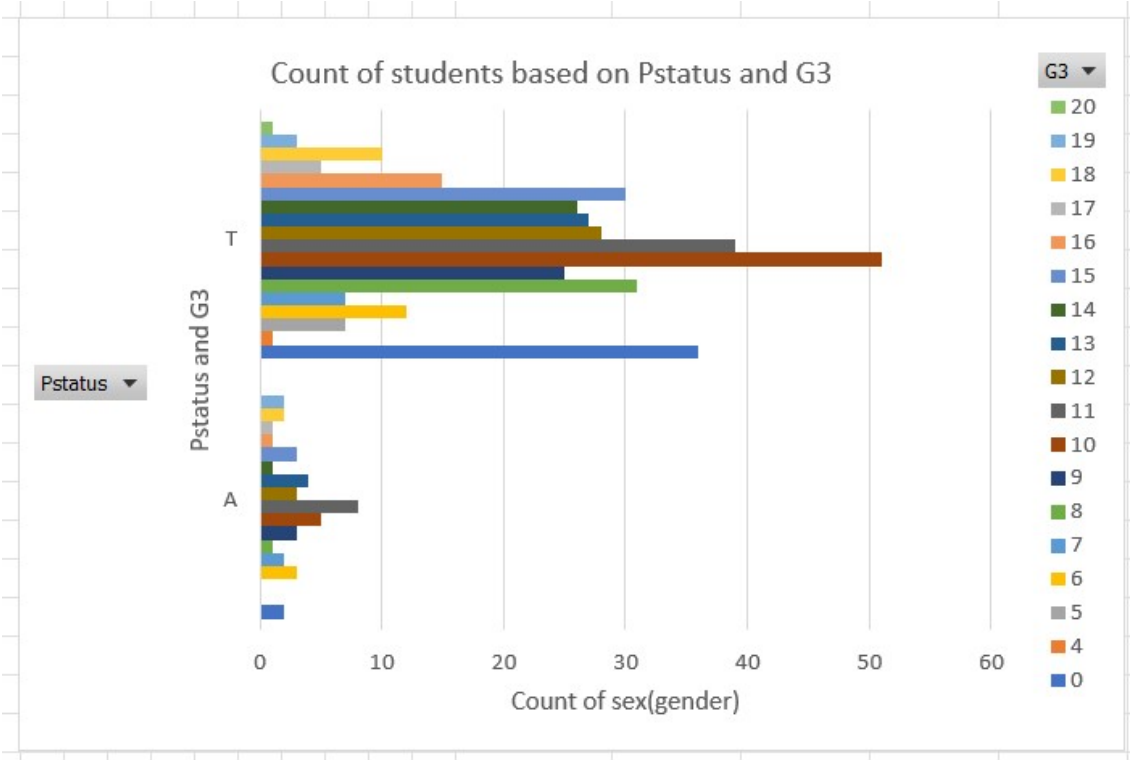


Figure 50

- It is seen in Figure 49, out of 395 students 354 students have their parent together.
- It is observed in Figure 50, that highest number of students (51) whose parents are together have got 10 score as G3, whereas the highest number of students (8) whose parents are not together have got 11 score as G3.

❖ **Dashboard 4**

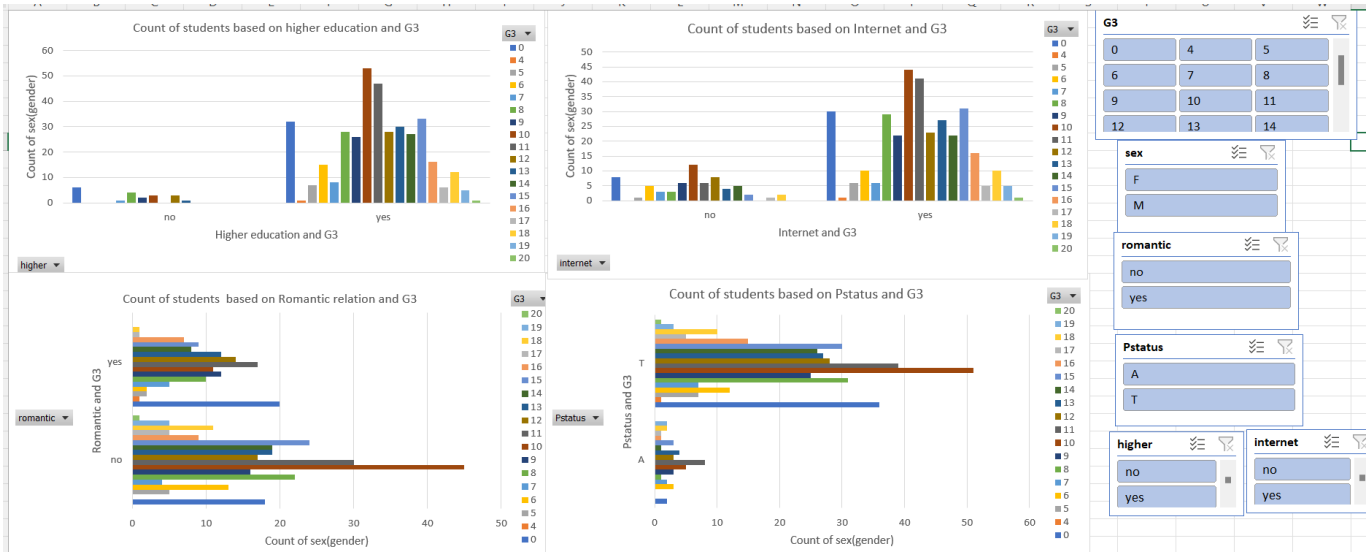


Figure 51

➢ Figure 51, shows the dashboard wherein various factors affecting the students data are shown in the charts.

➢ First chart shows Count of students based on higher education and G3.

➢ Second chart shows Count of students based on internet and G3.

➢ Third chart shows Count of students based on romantic relation G3.

➢ Fourth chart shows Count of students based on Pstatus and G3.

❖ **Regression Analysis**

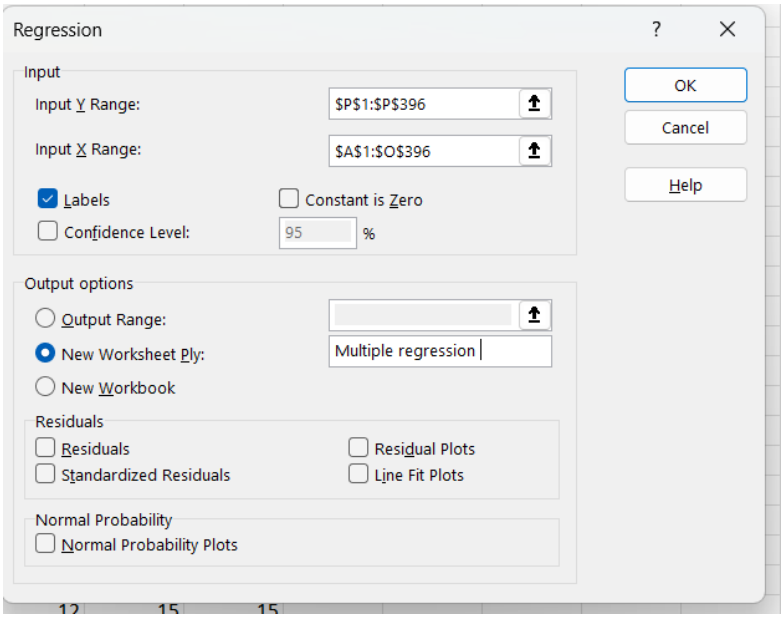


Figure 52

➢ As it is seen in Figure 52, configuration for multiple regression is displayed.

Regression Statistics	
Multiple R	0.914728398
R Square	0.836728042
Adjusted R Square	0.830266091
Standard Error	1.887498248
Observations	395

Figure 53

➢ The Regression statistics block of regression output is seen in the Figure 53.

➢ The R square is 83%, which means that variability of the explanatory variable explains about 83% of variability of the explained variable.

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	15	6919.664648	461.311	129.4854	8.5893E-139
Residual	379	1350.244212	3.56265		
Total	394	8269.908861			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.61024121	1.733017244	-0.35213	0.724939	-4.01777419	2.797292	-4.01777419	2.797291763
age	-0.20501415	0.079795471	-2.56925	0.010574	-0.361911435	-0.04812	-0.36191144	-0.048116869
Medu	0.117190295	0.115784364	1.012143	0.312116	-0.110469896	0.34485	-0.1104699	0.344850486
Fedu	-0.11930911	0.113980968	-1.04675	0.295884	-0.343423384	0.104805	-0.34342338	0.104805171
traveltime	0.127924884	0.14275444	0.896118	0.370758	-0.15276503	0.408615	-0.15276503	0.408614797
studytime	-0.11543284	0.120581192	-0.9573	0.339024	-0.352524763	0.121659	-0.35252476	0.121659081
Failures	0.006719121	0.37656195	0.017843	0.985773	-0.733693168	0.747131	-0.73369317	0.747131409
famrel	0.361687259	0.147074811	2.459206	0.01437	0.072502446	0.650872	0.072502446	0.650872073
freetime	0.042554047	0.103857656	0.409734	0.682232	-0.161655339	0.246763	-0.16165534	0.246763434
goout	-0.00522061	0.100704383	-0.05184	0.958683	-0.203229893	0.192789	-0.20322989	0.192788675
Dalc	-0.12834267	0.144075191	-0.8908	0.3736	-0.411629501	0.154944	-0.4116295	0.154944162
Walc	0.155299112	0.107414092	1.445798	0.14906	-0.055903091	0.366501	-0.05590309	0.366501315
health	0.050203893	0.070277158	0.71437	0.475438	-0.087978076	0.188386	-0.08797808	0.188385861
absences	0.040901422	0.012370481	3.306373	0.001035	0.01657805	0.065225	0.01657805	0.065224794
G1	0.167910345	0.056588576	2.967213	0.003196	0.056643454	0.279177	0.056643454	0.279177235
G2	0.980006212	0.050333111	19.47041	5.08E-59	0.881039085	1.078973	0.881039085	1.078973339

Figure 54

- As seen in Figure 54, a multiple regression is run, where the target variable is variable G3 and input variables are age, Medu, Fedu, travel time, study time, Failures, famrel, free time, gout, Dalc, Walc, health, absences, G1 and G2.

- G3 = α + β1*age + β2* Medu + β3* Fedu + β4* traveltime + β5*studytime + β6 *Failures + β7* Famrel + β8* freetime + β9* go out + β10* Dalc + β11* Walc + β12* health + β13* absences +β14* G1+ β15* G2 + ξ

  Where

  - -0.610 for Intercept α
  - -0.205 for slope β1
  - 0.117 for slope β2
  - -0.119 for slope β3
  - 0.127 for slope β4
  - -0.115 for slope β5
  - 0.006 for slope β6
  - 0.361 for slope β7
  - 0.042 for slope β8
  - -0.005 for slope β9

- -0.128 for slope $\beta 10$
- 0.155 for slope $\beta 11$
- 0.050 for slope $\beta 12$
- 0.040 for slope $\beta 13$
- 0.167 for slope $\beta 14$
- 0.980 for slope $\beta 15$

- In multiple regression model, where the explanatory variable is age, the estimate of slope co -efficient is - 0.205, which means:
  - With increase in every year of age, there will be a decrease in G3 (Final grade) on average by -0.205 units.
- In multiple regression model, where the explanatory variable is Medu, the estimate of slope co-efficient is 0.117, which means:
  - The higher the level of mother's education, it positively influences the student's grade G3(Final grade) on an average by 0.117 units.
- In multiple regression model, where the explanatory variable is Fedu, the estimate of slope co-efficient is -0.119, which means:
  - The higher the level of father's education, it negatively influences the student's grade G3(Final grade) on an average by -0.119 units.
- In multiple regression model, where the explanatory variable is traveltime, the estimate of slope co-efficient is 0.127, which means:
  - The more the travel time, it positively influences the student's grade G3(Final grade) on an average by 0.127 units.
- In multiple regression model, where the explanatory variable is study time, the estimate of slope co-efficient is -0.115 which means:
  - The more the study time, it negatively influences the student's grade G3(Final grade) on an average by -0.115 units.
- In multiple regression model, where the explanatory variable is Failures, the estimate of slope co-efficient is 0.006 which means:
  - With every failure, the student's grade G3(Final grade) will increase on an average by 0.006 units.
- In multiple regression model, where the explanatory variable is famrel, the estimate of slope co-efficient is 0.361, which means:
  - The better the quality of family relationships, it positively influences the student's grade G3(Final grade) on an average by 0.361 units.

- In multiple regression model, where the explanatory variable is free time, the estimate , of slope co-efficient is 0.042, which means:
  - The more freetime the student has, it positively influences the student's grade G3(Final grade) on an average by 0.042 units.
- In multiple regression model, where the explanatory variable is go out, the estimate of slope co-efficient is - 0.005, which means:
  - The higher the intensity of going out, it results in decrease in the student's grade G3(Final grade) on an average by 0.005 units.
- In multiple regression model, where the explanatory variable is Dalc, the estimate of slope co-efficient is - 0.128, which means:
  - The higher the level of consumption of alcohol on a work day, it negatively influences the student's grade G3(Final grade) on an average by -0.128 units.
- In multiple regression model, where the explanatory variable is Walc, the estimate of slope co-efficient is 0.155, which means:
  - On the contrary, higher the level of consumption of alcohol on a weekend positively influences the student's grade G3(Final grade) on an average by 0.155 units.
- In multiple regression model, where the explanatory variable is health, the estimate of slope co-efficient is 0.050, which means:
  - The higher the level of health of the student, it positively influences the student's grade G3(Final grade) on an average by 0.050 units.
- In multiple regression model, where the explanatory variable is absences, the estimate of slope co-efficient is 0.040, which means:
  - With every increase in the number of absences, it results in increase in the student's grade G3(Final grade) on an average by 0.040 units.
- In a multiple regression model, where the explanatory variable is G1, the estimate of slope co-efficient is 0.167, which means:
  - With increase in every unit of G1, it results in increase in the student's grade G3(Final grade) on an average by 0.167 units.

- In a multiple regression model, where the explanatory variable is G2, the estimate of slope co-efficient is 0.980, which means:
  - With increase in every unit of G2, it results in increase in the student's grade G3(Final grade) on an average by 0.980 units.

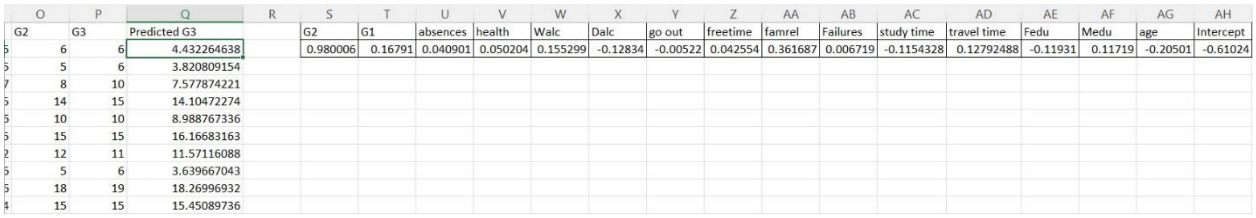
❖ **Significance of parameter estimate and Null hypothesis**

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.61024121	1.733017244	-0.35213	0.724939	-4.01777419	2.797292	-4.01777419	2.797291763
age	-0.20501415	0.079795471	-2.56925	0.010574	-0.361911435	-0.04812	-0.36191144	-0.048116869
Medu	0.117190295	0.115784364	1.012143	0.312116	-0.110469896	0.34485	-0.1104699	0.344850486
Fedu	-0.11930911	0.113980968	-1.04675	0.295884	-0.343423384	0.104805	-0.34342338	0.104805171
traveltime	0.127924884	0.14275444	0.896118	0.370758	-0.15276503	0.408615	-0.15276503	0.408614797
studytime	-0.11543284	0.120581192	-0.9573	0.339024	-0.352524763	0.121659	-0.35252476	0.121659081
Failures	0.006719121	0.37656195	0.017843	0.985773	-0.733693168	0.747131	-0.73369317	0.747131409
famrel	0.361687259	0.147074811	2.459206	0.01437	0.072502446	0.650872	0.072502446	0.650872073
freetime	0.042554047	0.103857656	0.409734	0.682232	-0.161655339	0.246763	-0.16165534	0.246763434
goout	-0.00522061	0.100704383	-0.05184	0.958683	-0.203229893	0.192789	-0.20322989	0.192788675
Dalc	-0.12834267	0.144075191	-0.8908	0.3736	-0.411629501	0.154944	-0.4116295	0.154944162
Walc	0.155299112	0.107414092	1.445798	0.14906	-0.055903091	0.366501	-0.05590309	0.366501315
health	0.050203893	0.070277158	0.71437	0.475438	-0.087978076	0.188386	-0.08797808	0.188385861
absences	0.040901422	0.012370481	3.306373	0.001035	0.01657805	0.065225	0.01657805	0.065224794
G1	0.167910345	0.056588576	2.967213	0.003196	0.056643454	0.279177	0.056643454	0.279177235
G2	0.980006212	0.050333111	19.47041	5.08E-59	0.881039085	1.078973	0.881039085	1.078973339

Figure 55

- In Figure 55 , we can see the significance of the parameter estimates by observing the p-value, the parameter estimate associated with the variables age, famrel, absences, G1 and G2 are statistically significant since the p -value associated with the respective variables is less than the threshold value that is 5% (0.05).This means for these variables we reject null hypothesis which states that the parameter estimate associated with variable equals zero in favour of alternative one which states that the parameter estimate associated with variable does not equal zero.
- The P-value of parameter estimates associated with variables Medu, Fedu, studytime, travel time, Failures, free time, go out,Dalc, Walc, and health is not statistically significant. This means for these variables we accept null hypothesis.

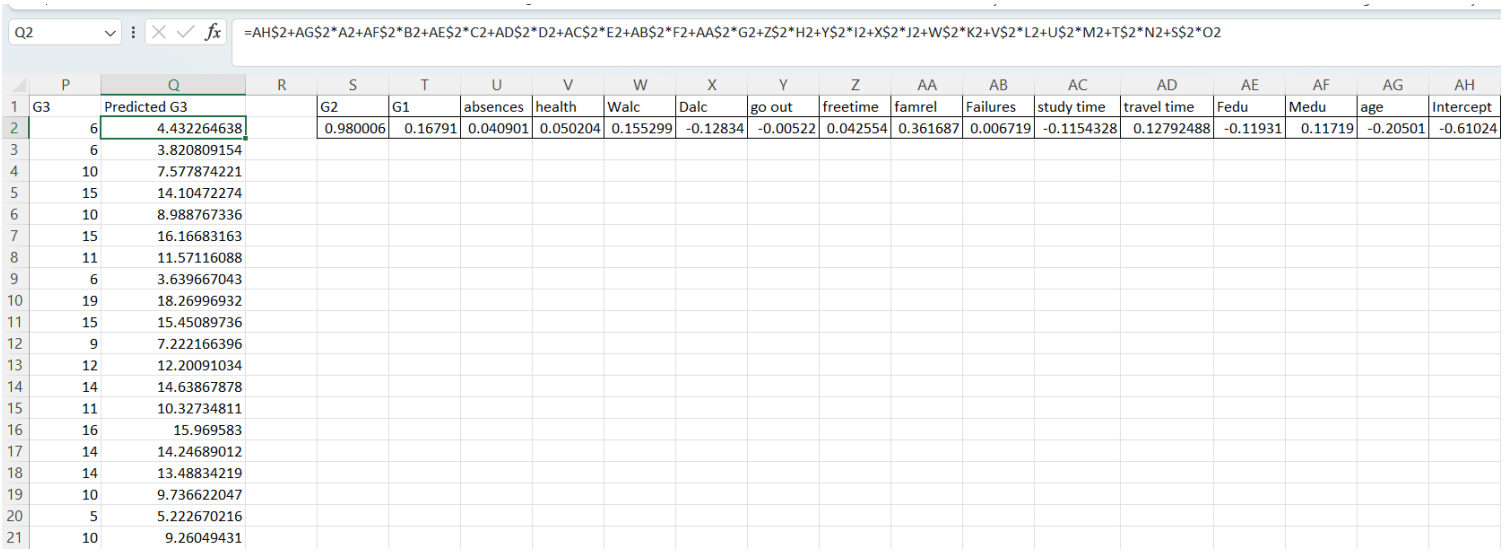
❖ **Prediction of G3(Final Grade)**



	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
	G2	G3	Predicted G3		G2	G1	absences	health	Walc	Dalc	go out	freetime	famrel	Failures	study time	travel time	Fedu	Medu	age	Intercept
5	6	6	4.432264638		0.980006	0.16791	0.040901	0.050204	0.155299	-0.12834	-0.00522	0.042554	0.361687	0.006719	-0.1154328	0.12792488	-0.11931	0.11719	-0.20501	-0.61024
5	5	6	3.820809154																	
7	8	10	7.577874221																	
5	14	15	14.10472274																	
5	10	10	8.988767336																	
5	15	15	16.16683163																	
2	12	11	11.57116088																	
5	5	6	3.639667043																	
5	18	19	18.26996932																	
4	15	15	15.45089736																	

Figure 56

➢ As seen in Figure 56, intercept and parameter estimates of slope co-efficient from the multiple regression model is used to predict G3(Final grade).



Q2 =AH$2+AG$2*A2+AF$2*B2+AE$2*C2+AD$2*D2+AC$2*E2+AB$2*F2+AA$2*G2+Z$2*H2+Y$2*I2+X$2*J2+W$2*K2+V$2*L2+U$2*M2+T$2*N2+S$2*O2

	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
1	G3	Predicted G3		G2	G1	absences	health	Walc	Dalc	go out	freetime	famrel	Failures	study time	travel time	Fedu	Medu	age	Intercept
2	6	4.432264638		0.980006	0.16791	0.040901	0.050204	0.155299	-0.12834	-0.00522	0.042554	0.361687	0.006719	-0.1154328	0.12792488	-0.11931	0.11719	-0.20501	-0.61024
3	6	3.820809154																	
4	10	7.577874221																	
5	15	14.10472274																	
6	10	8.988767336																	
7	15	16.16683163																	
8	11	11.57116088																	
9	6	3.639667043																	
10	19	18.26996932																	
11	15	15.45089736																	
12	9	7.222166396																	
13	12	12.20091034																	
14	14	14.63867878																	
15	11	10.32734811																	
16	16	15.969583																	
17	14	14.24689012																	
18	14	13.48834219																	
19	10	9.736622047																	
20	5	5.222670216																	
21	10	9.26049431																	

Figure 57

➢ Predicted G3 = α + β1*age + β2* Medu + β3* Fedu + β4* traveltime + β5*studytime + β6 *Failures + β7* Famrel + β8* freetime + β9* go out + β10* Dalc + β11* Walc + β12* health + β13* absences +β14* G1+ β15* G2 + ξ

➢ The formula mentioned above is used in Figure 57 , to get the predicted G3(Final grade)

❖ **Descriptive Statistics of G3 and Predicted G3**

G3	
Mean	10.41519
Standard Error	0.230517
Median	11
Mode	10
Standard Deviatic	4.581443
Sample Variance	20.98962
Kurtosis	0.403421
Skewness	-0.73267
Range	20
Minimum	0
Maximum	20
Sum	4114
Count	395

Figure 58

Predicted G3	
Mean	10.41518987
Standard Error	0.210860807
Median	10.28679041
Mode	#N/A
Standard Deviation	4.19077566
Sample Variance	17.56260063
Kurtosis	0.260032356
Skewness	-0.279634222
Range	22.1138976
Minimum	-1.653947456
Maximum	20.45995015
Sum	4114
Count	395

Figure59

- The mean (average) of both Actual G3 and Predicted G3 is the same.
- Median of Actual G3 is 11 and median of Predicted G3 is 10.286.
- Mode of Actual G3 is 10 whereas Mode of Predicted G3 states #N/A which means no value is available.
- Standard deviation of Actual G3 is 4.581 and standard deviation of Predicted G3 is 4.190, the data points of Actual G3 are more dispersed as compared to Predicted G3.
- In case of kurtosis the distribution curve of Actual G3 is more peaked as compared to Predicted G3.
- The Skewness of Actual G3 is -0.732 which means the distribution is moderately skewed on the left side, whereas the Skewness of Predicted G3 is -0.279 which indicates a relatively symmetric distribution.
- Range of Actual G3 is 20 and Range of Predicted G3 is 22.113.
- Minimum of Actual G3 is 0 whereas minimum of Predicted G3 is -1.653.
- Maximum of Actual G3 is 20 whereas maximum of Predicted G3 is 20.45.

❖ **Correlation of actual value to predicted value**

	G3	*Predicted G3*
G3	1	
Predicted G3	0.914728398	1

Figure 60

➢ It is seen in Figure 60, that the correlation between the variables G3 and Predicted is 91% which is good.

- **Trends, Patterns and Anomalies**

  ❖ **CORRELATION**

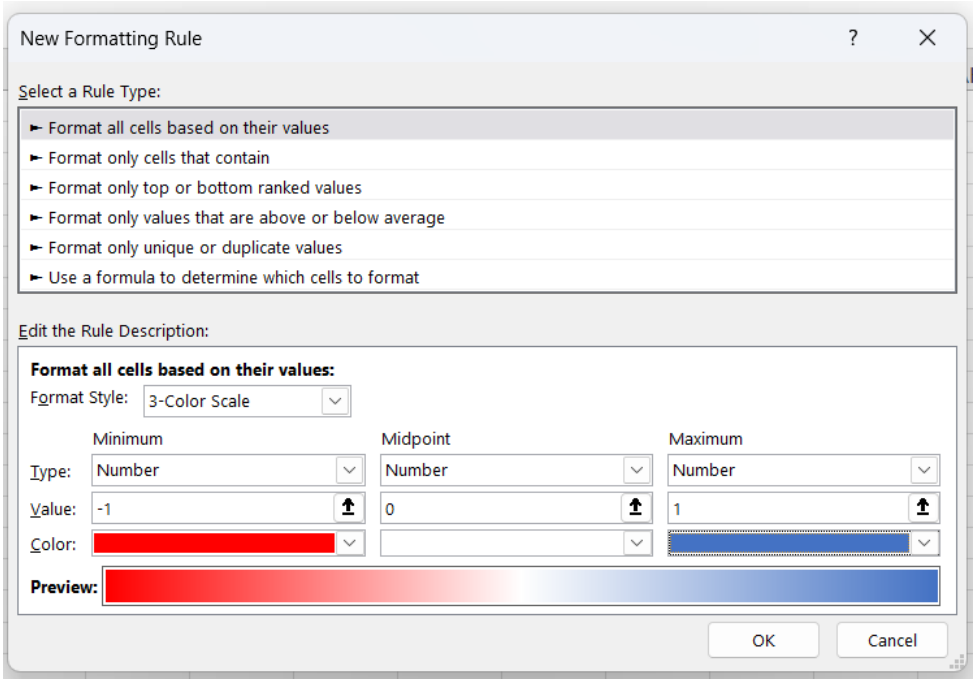
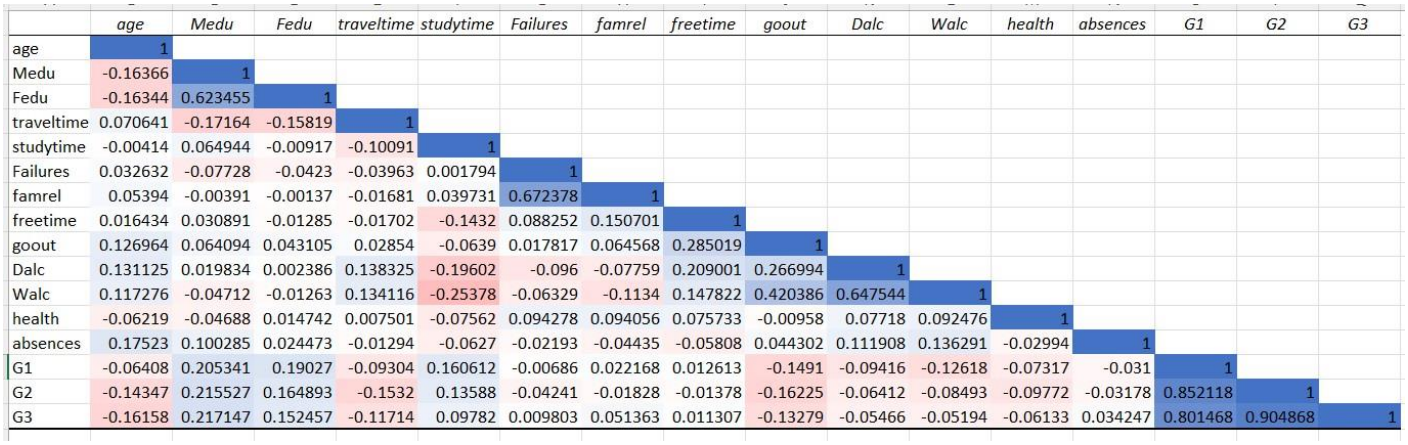


Figure 61

➢ As it is seen in Figure 61, the correlation table displays a pattern using conditional formatting (color scale), where the variables with a strong and positive correlation are displayed in shades of blue, whereas the variables with negative correlation between them are displayed in shades of the red. The darker the shade of the color (red/blue) the stronger the positive /negative correlation is displayed between the variables. The variables with very weak or no correlation are displayed in white colour.



	age	Medu	Fedu	traveltime	studytime	Failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
age	1															
Medu	-0.16366	1														
Fedu	-0.16344	0.623455	1													
traveltime	0.070641	-0.17164	-0.15819	1												
studytime	-0.00414	0.064944	-0.00917	-0.10091	1											
Failures	0.032632	-0.07728	-0.0423	-0.03963	0.001794	1										
famrel	0.05394	-0.00391	-0.00137	-0.01681	0.039731	0.672378	1									
freetime	0.016434	0.030891	-0.01285	-0.01702	-0.1432	0.088252	0.150701	1								
goout	0.126964	0.064094	0.043105	0.02854	-0.0639	0.017817	0.064568	0.285019	1							
Dalc	0.131125	0.019834	0.002386	0.138325	-0.19602	-0.096	-0.07759	0.209001	0.266994	1						
Walc	0.117276	-0.04712	-0.01263	0.134116	-0.25378	-0.06329	-0.1134	0.147822	0.420386	0.647544	1					
health	-0.06219	-0.04688	0.014742	0.007501	-0.07562	0.094278	0.094056	0.075733	-0.00958	0.07718	0.092476	1				
absences	0.17523	0.100285	0.024473	-0.01294	-0.0627	-0.02193	-0.04435	-0.05808	0.044302	0.111908	0.136291	-0.02994	1			
G1	-0.06408	0.205341	0.19027	-0.09304	0.160612	-0.00686	0.022168	0.012613	-0.1491	-0.09416	-0.12618	-0.07317	-0.031	1		
G2	-0.14347	0.215527	0.164893	-0.1532	0.13588	-0.04241	-0.01828	-0.01378	-0.16225	-0.06412	-0.08493	-0.09772	-0.03178	0.852118	1	
G3	-0.16158	0.217147	0.152457	-0.11714	0.09782	0.009803	0.051363	0.011307	-0.13279	-0.05466	-0.05194	-0.06133	0.034247	0.801468	0.904868	1

Figure 62

- The correlation between explained variable G3 and the variable age is about -16%.
- The correlation between explained variable G3 and the variable Medu is about 21%.
- The correlation between explained variable G3 and the variable Fedu is about 15%.
- The correlation between explained variable G3 and the variable travel time is about - 11%.
- The correlation between explained variable G3 and the variable study time is about 9%.
- The correlation between explained variable G3 and the variable Failures is about 0.9%
- The correlation between explained variable G3 and the variable famrel is about 5%.
- The correlation between explained variable G3 and the variable free time is about 1.1%.
- The correlation between explained variable G3 and the variable go out is about -13%.
- The correlation between explained variable G3 and the variable Dalc is about - 5%.
- The correlation between explained variable G3 and the variable Walc is about - 5%.
- The correlation between explained variable G3 and the variable health is about - 6%.
- The correlation between explained variable G3 and the variable absences is about 3%.
- The correlation between explained variable G3 and the variable G1 is about 80%.
- The correlation between explained variable G3 and the variable G2 is about 90%.
- The variable G3 has a weak positive correlation with variables Study time ,failures famrel,free time and absences.
- The variable G3 has a weak negative correlation with variables traveltime,Dalc,Walc and health .
- The variable G3 has strong positive correlation with both variables G1 and G2, but the correlation between G2 and G3 is more stronger as compared to G3 and G1.

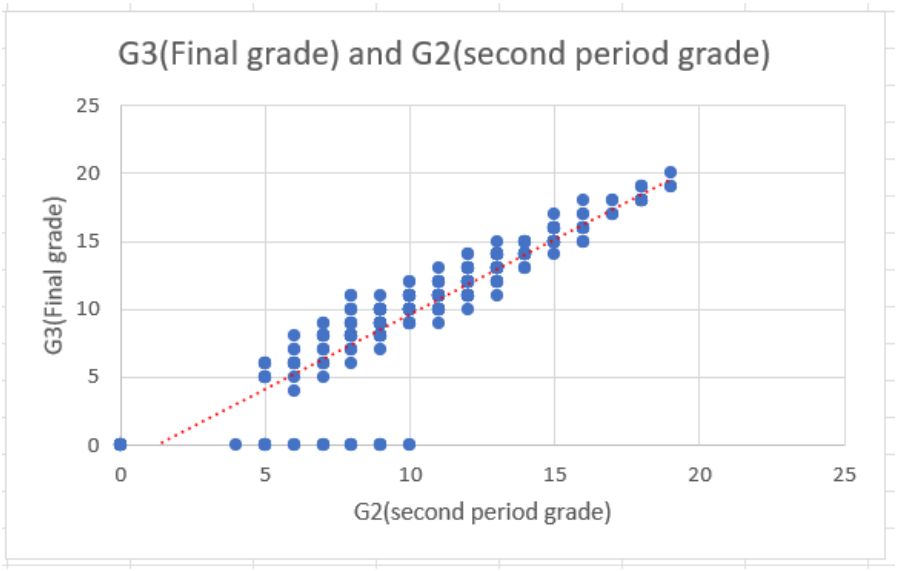
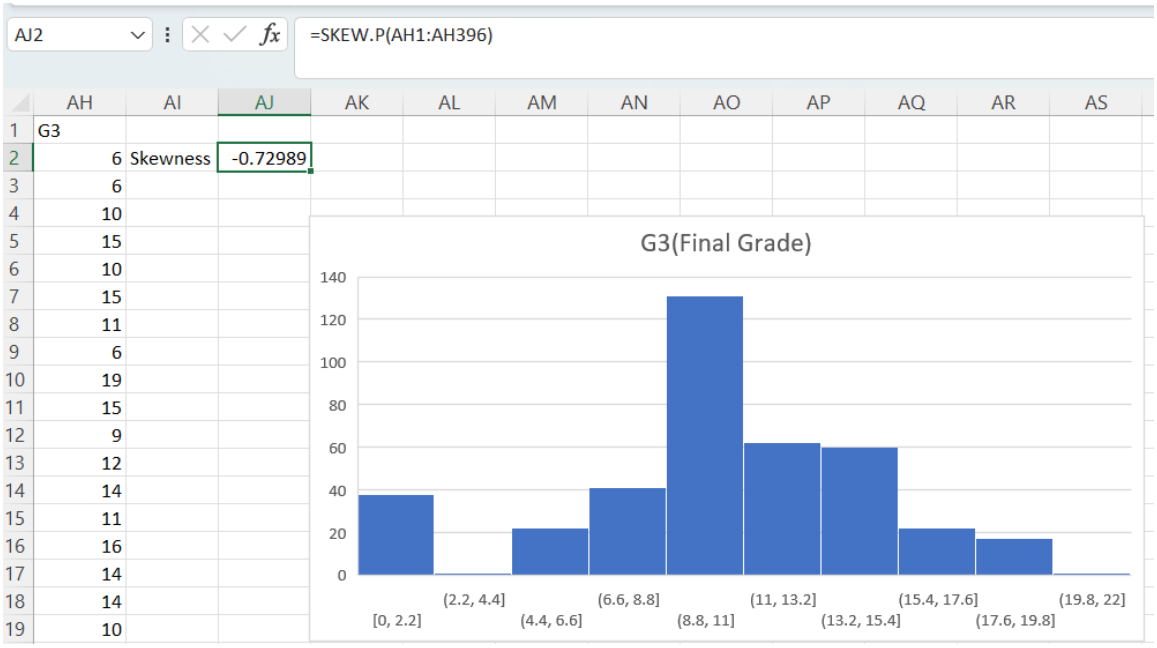


Figure 63

➢As seen in Figure 63 , the variabe G3 has maximum positive correlation with the variable G2.

❖ **SKEWNESS**

1) Histogram of variable G3(Final grade)



	AH	AI	AJ
1	G3		
2	6	Skewness	-0.72989
3	6		
4	10		
5	15		
6	10		
7	15		
8	11		
9	6		
10	19		
11	15		
12	9		
13	12		
14	14		
15	11		
16	16		
17	14		
18	14		
19	10		

Figure 64

➢ As seen in Figure 64, the skewness of the variable G3 is -0.729, which means the distribution is moderately skewed on the left side.

❖ **TRENDLINE**

1) Scatter plot for the variable G3 and age

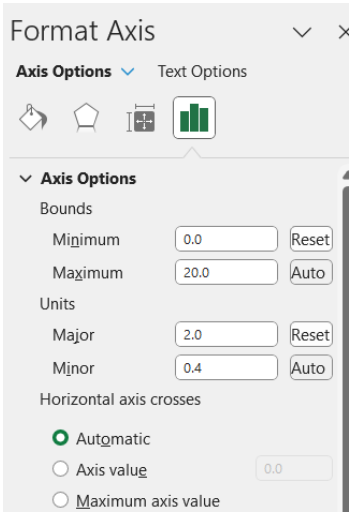


Figure 65

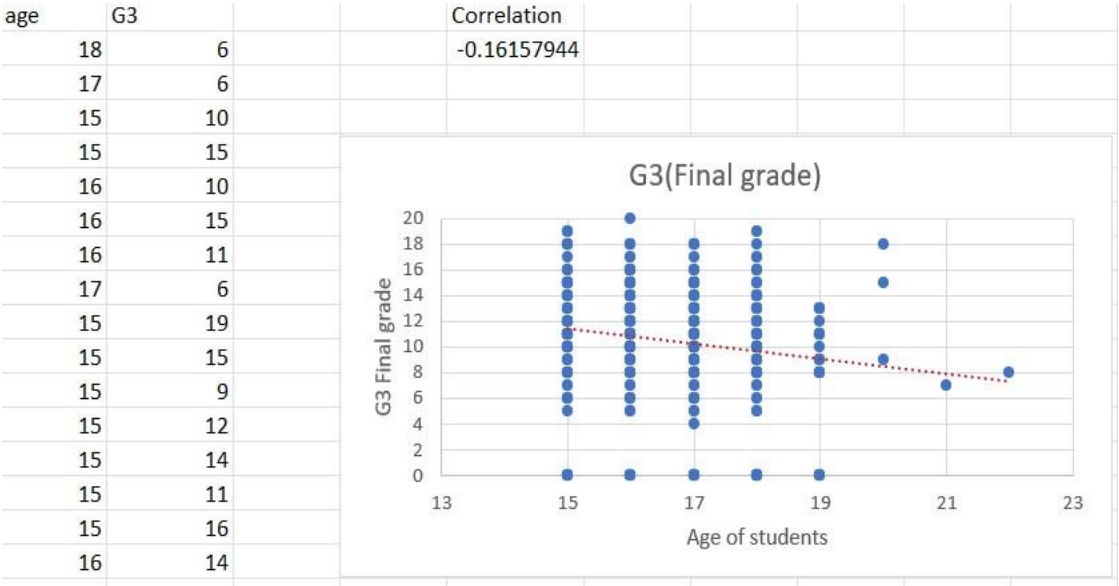


Figure 66

- It is seen in Figure 65, that the minimum and maximum bound limits of the variable G3 are set, similarly minimum and maximum bound limits of the variable Age are set.
- A scatter plot between variables G3 and Age is seen in Figure 66, it is observed that the grade score of students in the age group 15 years to 18 years ranges between 0 to 20 units.
- There is a decline in the grade score to 13 units (highest score) for the age group 19 years old.
- The grade scores increase for the age group 20 years old as compared to the age group 19 years old.

- The grade score decreases again for the age group 21 years old and increases by a few units for the age group 22 years old.
- A declining trendline in the scatter plot and a negative correlation as seen in Figure 66, confirms the same.

2) Scatter plot for Medu and G3

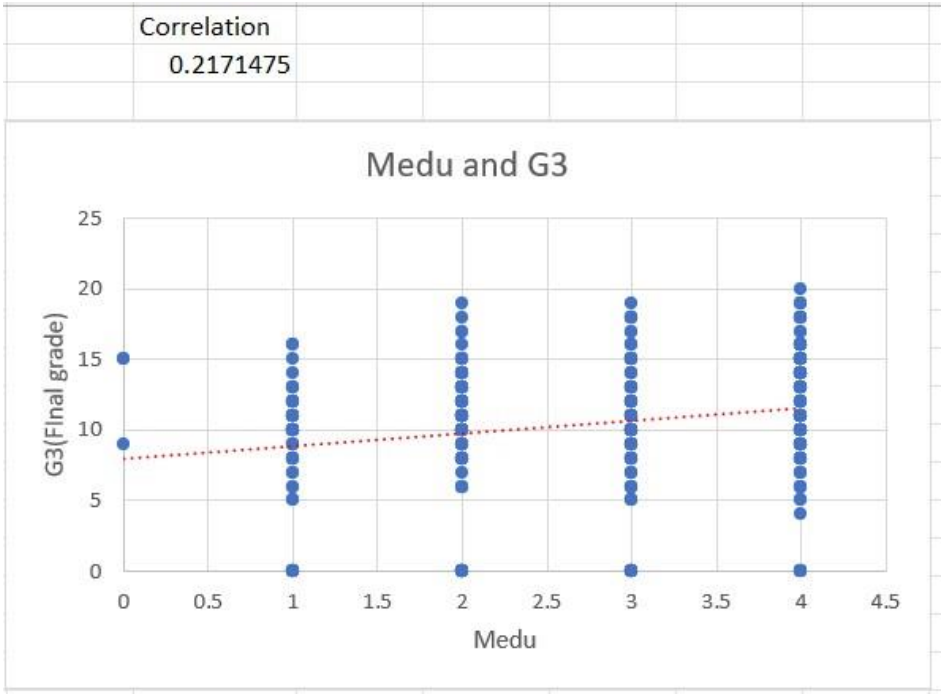


Figure 67

- It is seen in the Figure 67, that the variables Medu and G3 have a positive correlation.
- It is also observed that the increase in the level of mother's education leads to an overall increase in the level of G3 scores of the students.
- An upward trendline in the scatter plot and a positive correlation as seen in Figure 67, confirms the same.

3) Scatter plot for Fedu and G3

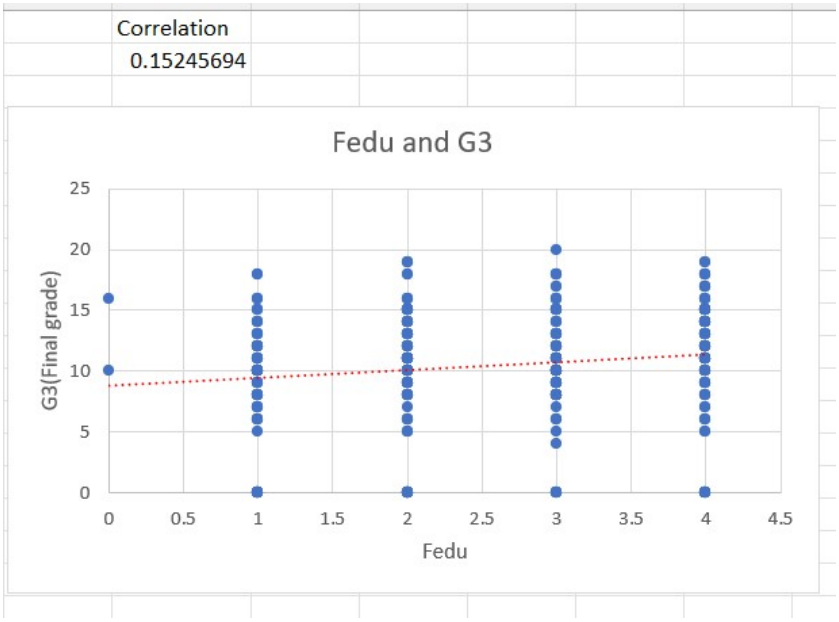


Figure 68

- It is seen in Figure 68, that the variables Fedu and G3 have a positive correlation.
- It is also observed that the increase in the level of father's education leads to an overall increase in the level of G3 scores of the students for level 1, 2 and 3 of Fedu(Father's education). Incase of Level 4 of Fedu it is observed that there is a slight decline in the highest score as compared to Level 3 of Fedu.
- An upward trendline in the scatter plot and a positive correlation as seen in Figure 68, confirms the same.

4) Scatter plot of study time and G3

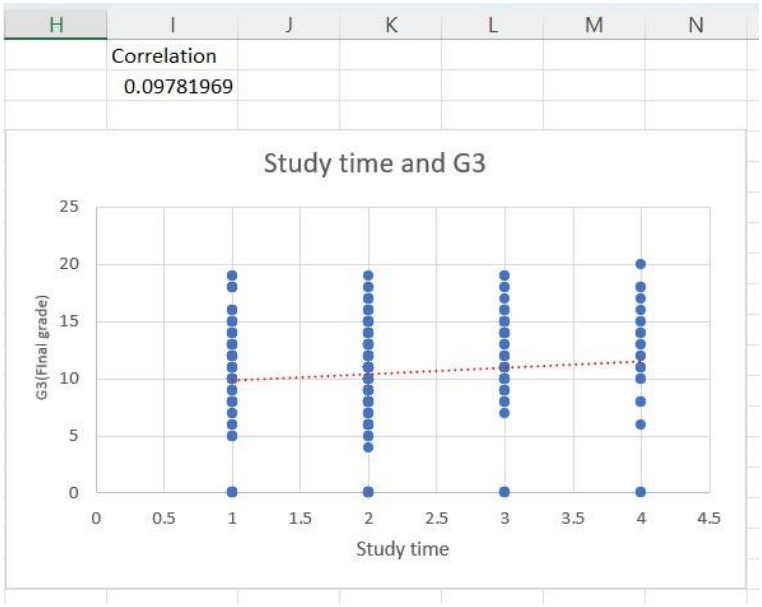


Figure 69

- It is seen in Figure 69, that the correlation between G3 and study time is positive.
- It is further observed that increase in study time leads to an overall increase in G3 scores of the students.
- An upward trendline in the scatter plot and a positive correlation as seen in Figure 69, confirms the same.

5) Scatter plot of G3 and Absences

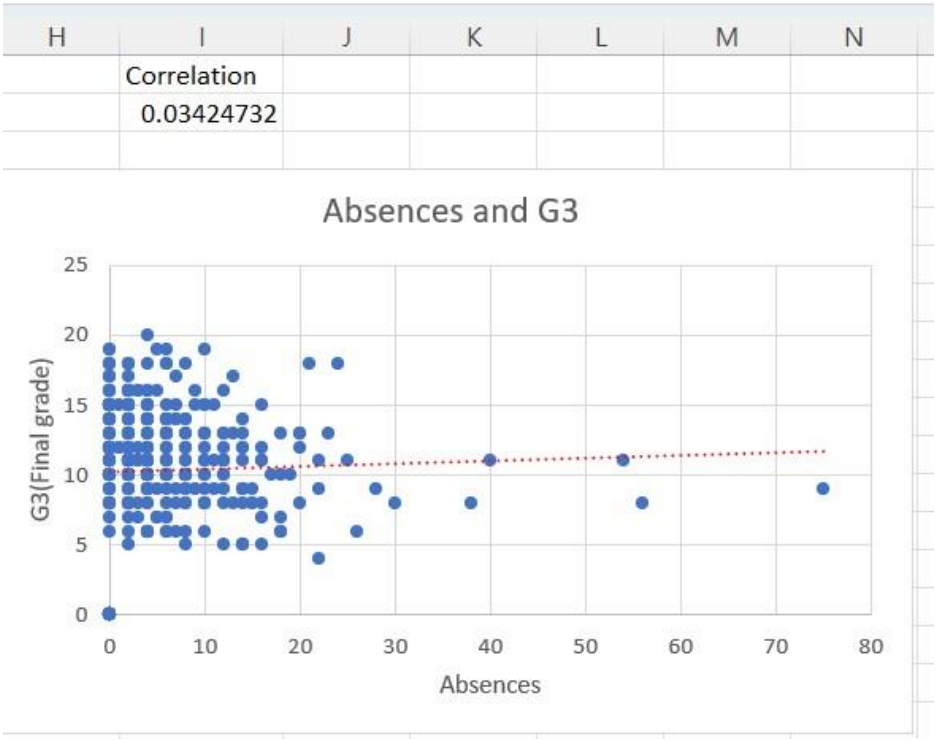


Figure 70

- In Figure 70, it is observed that most of the students have absences recorded between 0 to 30 days.
- It is further observed that as the number of absent days of the students increases further after 30 days, the G3 score decreases.
- An upward trendline in the scatter plot and a relatively weak positive correlation as seen in Figure 70, confirms the same.

❖ **ANOMALIES**

failures	number of past class failures (numeric: n if 1<=n<3, else 3)

Figure 71

P
failures
0
0
3
0
0
0
0
0
0
0
0
0
0

Figure 72

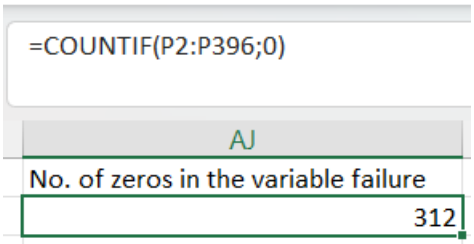


Figure 73

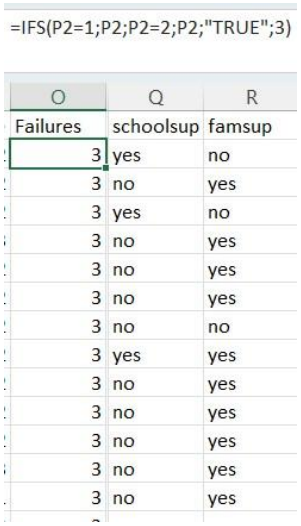


Figure 74

- The description of variable failure as seen in Figure 71, is given in the dataset description.
- However, it is observed in Figure 72 and 73, that out of 395 values there are 312 values in the variable failures which do not follow this description. Therefore, the necessary data cleaning process needs to be implemented.
- As seen in Figure 74, Ifs function is used to fill in values as per the condition mentioned in the dataset description.

- **Discussion**

❖ **Initial Assumptions**

1) The Data values have normal distribution (bell curve)

- After doing Exploratory data analysis it is observed that not all data values in the data set have a normal distribution.
- The variables Age, Medu, Fedu, study time, free time, go out, health, G1, G2 have normal (symmetric) distribution since the skewness of these variables ranges between -0.5 to 0.5.
- The variables famrel, Walc, G3 have asymmetric distribution since the values are moderately skewed i.e the values range between -1 to -0.5 and 0.5 to 1.
- The variables travel time, failures, Dalc, absences have asymmetric distribution since the values are heavily skewed i.e the values range between below -1 to above 1.

2) The Data values have homogeneity of variance

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
G1	395	4309	10.90886	11.01705
G2	395	4232	10.71392	14.14892
G3	395	4114	10.41519	20.98962

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	48.84219	2	24.4211	1.587311	0.20491	3.003338
Within Groups	18185.3	1182	15.3852			
Total	18234.14	1184				

Figure 75

- As it is seen in Figure 75, the average of G1(First period grade) is the highest followed by G2(Second period grade) and G3(Final grade).
- The variance of the variables G1, G2 and G3 is not the same.
- The p- value is more than 5% (0.05) which means that the values are not statistically significant.
- F critical value is greater than F statistic, we can conclude that the test is not significant.

3) Data has a linear relationship.

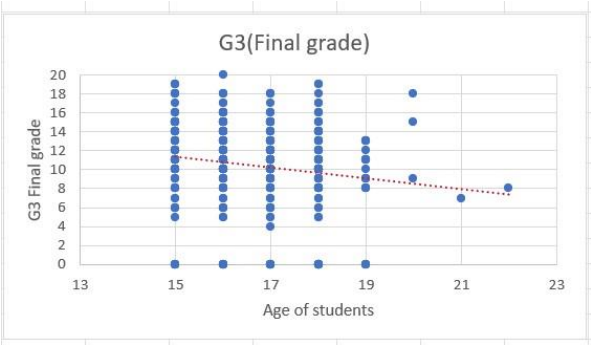


Figure 76

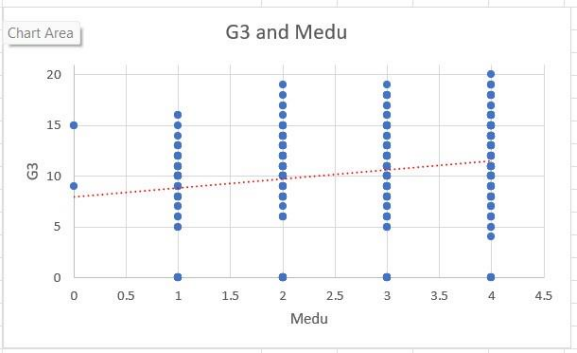


Figure 77

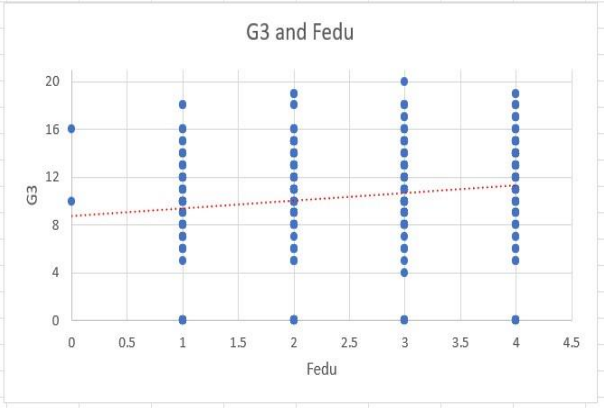


Figure 78

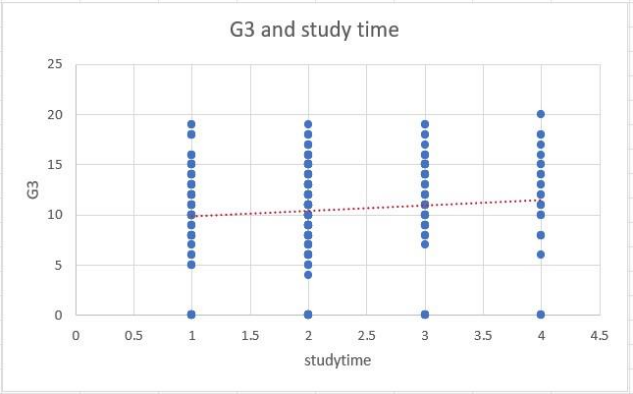


Figure 79

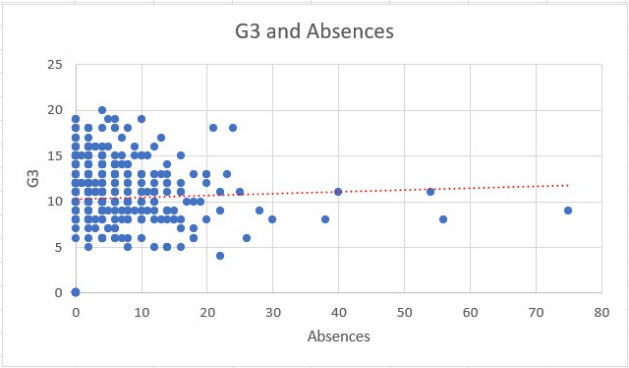


Figure 80

- The scatter plot of variable G3 and Age presented in Figure 76, displays a negative linear relationship with a declining trendline.
- The scatter plot of variable G3 and Medu presented in Figure 77, displays a positive linear relationship.
- The scatter plot of variable G3 and Fedu presented in Figure 78, displays a positive linear relationship.
- The scatter plot of variable G3 and study time presented in Figure 79, displays a weak positive linear relationship.
- The scatter plot of variable G3 and absences presented in Figure 80, displays a weak positive linear relationship.

❖ **Hypotheses**

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.61024121	1.733017244	-0.35213	0.724939	-4.01777419	2.797292	-4.01777419	2.797291763
age	-0.20501415	0.079795471	-2.56925	0.010574	-0.361911435	-0.04812	-0.36191144	-0.048116869
Medu	0.117190295	0.115784364	1.012143	0.312116	-0.110469896	0.34485	-0.1104699	0.344850486
Fedu	-0.11930911	0.113980968	-1.04675	0.295884	-0.343423384	0.104805	-0.34342338	0.104805171
traveltime	0.127924884	0.14275444	0.896118	0.370758	-0.15276503	0.408615	-0.15276503	0.408614797
studytime	-0.11543284	0.120581192	-0.9573	0.339024	-0.352524763	0.121659	-0.35252476	0.121659081
Failures	0.006719121	0.37656195	0.017843	0.985773	-0.733693168	0.747131	-0.73369317	0.747131409
famrel	0.361687259	0.147074811	2.459206	0.01437	0.072502446	0.650872	0.072502446	0.650872073
freetime	0.042554047	0.103857656	0.409734	0.682232	-0.161655339	0.246763	-0.16165534	0.246763434
goout	-0.00522061	0.100704383	-0.05184	0.958683	-0.203229893	0.192789	-0.20322989	0.192788675
Dalc	-0.12834267	0.144075191	-0.8908	0.3736	-0.411629501	0.154944	-0.4116295	0.154944162
Walc	0.155299112	0.107414092	1.445798	0.14906	-0.055903091	0.366501	-0.05590309	0.366501315
health	0.050203893	0.070277158	0.71437	0.475438	-0.087978076	0.188386	-0.08797808	0.188385861
absences	0.040901422	0.012370481	3.306373	0.001035	0.01657805	0.065225	0.01657805	0.065224794
G1	0.167910345	0.056588576	2.967213	0.003196	0.056643454	0.279177	0.056643454	0.279177235
G2	0.980006212	0.050333111	19.47041	5.08E-59	0.881039085	1.078973	0.881039085	1.078973339

Figure 81

- Students with more study time are more likely to secure better grades.

  The above-mentioned hypotheses does not hold true. It can be seen from Figure 81, in the multiple regression model, that the variable study time negatively influences the variable G3 (Final grade).

  The parameter estimate associated with the variable study time is not statistically significant.

- Students with high score in quality of famrel are more likely to get better grades.

  The above-mentioned hypotheses hold true. It can be seen from Figure 81, in the multiple regression model that the variable famrel positively influences the variable G3 (Final grade).

  The parameter estimate associated with the variable famrel is statistically significant.

- Students who go out with friends are more likely to score lower grades.

  The above-mentioned hypotheses hold true. It can be seen from Figure 81, in the multiple regression model that the variable go out negatively influences the variable G3 (Final grade).

  The parameter estimate associated with the variable go out is not statistically significant.

- Students with very high Dalc (workday alcohol consumption) score are more likely to get lower grades.

  The above-mentioned hypotheses hold true. It can be seen from Figure 81, in the multiple regression model that the variable Dalc negatively influences the variable G3 (Final grade). The parameter estimate associated with the variable Dalc is not statistically significant.

- Students with good health score are more likely to get better grade.

  The above-mentioned hypotheses hold true. It can be seen from Figure 81, in the multiple regression model that the variable health positively influences the variable G3 (Final grade). The parameter estimate associated with variable health is not statistically significant.

- Students with more absences are more likely to secure lower grades.

  The above-mentioned hypotheses does not hold true. On the contrary it can be seen from Figure 81 , in the multiple regression model that the variable absences positively influences the variable G3 (Final grade). The parameter estimate associated with the variable absences is statistically significant.

- Students with good score in G1(First period grade) are more likely to secure better G3(Final grade)

  The above-mentioned hypotheses holds true. It can be seen from Figure 81, in the multiple regression model that the variable G1 positively influences the variable G3 (Final grade). The parameter estimate associated with the variable G1 is statistically significant.

- Students with good score in G2(Second period grade) are more likely to secure better G3(Final grade)

  The above-mentioned hypotheses holds true. It can be seen from Figure 81, in the multiple regression model that the variable G2 positively influences the variable G3 (Final grade). The parameter estimate associated with the variable G2 is statistically significant.

- **Conclusion and reflection**

- To predict the final grade that is G3, in the given dataset, EDA (Exploratory Data Analysis) is performed, wherein Descriptive statistics of all the numeric variables are analyzed, pivot tables and pivot charts are prepared to get a detailed visual presentation of how various explanatory variables affect the explained variable G3(Final grade).Correlation table is prepared to get an insight into how each explanatory variable affects the explained variable G3(Final grade).ANOVA analysis is performed to analyze the variance of the variables G1, G2 and G3. The multiple regression model is used to test the null hypothesis and p-value of the parameter estimates and to predict the G3 (final grade).
- By performing the above-mentioned analysis, it is observed that, most of the students go to the "GP" secondary school, more than 50% of the students come from Urban area.37% of the students have taken admission in the respective schools because of the course preference. The majority of the students take less than 15 minutes to travel from home to school. Most of the students have good family relations. Almost 87% of the students do not have school support (extra educational support). More than 50% of the students have not paid for extra paid classes. Almost 50% of the students have participated in extra-curricular activities. Almost 95% of the students have taken up higher education. Almost 83% of the students have access to the internet and 66% of the students are not involved in a romantic relationship. Almost 89% of the student's parents are together.
- It is further observed in the regression model that, as the age of the student increases there is a reduction in the G3(final grade), as the intensity of going out with friends increases there is a reduction in the G3(final grade). Increase in Dalc (workday alcohol consumption) leads to reduction in G3 (final grade). After that, the final grade (G3) of the students is predicted using multiple regression model. Correlation between actual G3 (final grade) and predicted G3 (final grade) is calculated which is almost 91%.
- In the beginning of the analysis, it was clear to me what approach to follow in terms of EDA(Exploratory Data Analysis) which includes Descriptive statistics of numeric variables, Pivot tables and Pivot charts wherein the relations between various variables(numeric and categorical) can be analyzed and how they affect the explained variable G3(final grade). Initially, I had decided to prepare a scatter plot of variables G1 and G3, and a scatter plot of G2 and G3, but while doing the analysis I realized that variable G3 has the strongest positive correlation with variable G2 as compared to G1, so I decided to prepare a scatter plot of G3 and G2 only.

What surprised me was the initial assumptions and hypothesis, as to how some of my initial assumptions and hypothesis hold true and some don't. While working with this dataset I realized, especially in the multiple regression model how relationship between explanatory variables and explained variables can be analyzed and interpreted for real data and the intercept and the slope co-efficient of parameter estimates can be further used for prediction. Also, I got to know how the application of central tendency measures, measures of shape, Pivot Charts, Pivot Tables, Dashboard works as I proceeded with the project.