Project Report

On

“Alcohol and other factors effect on students’ performance”

By

Vineeth Matrupally M538M626

Under the guidance of

Dr. Murtaza Nasir

ASSISTANT PROFESSOR

W. Frank Barton School of Business,

Wichita State University,1845 Fairmount St, Wichita, KS 67260

(August 2022 – December 2022)

CONTENT:

1. INTRODUCTION

Schools, colleges, and universities are nugatory without the students. Students plays a vital role in the educational institutions and are the maximum crucial asset for any instructional organization. Academic achievement has an important role to make them successful people so that they can make their place in the outer world. Schools were built for the learners to enhance all their skills like learning, thinking, communication, technical, and whatnot. Measuring student academic performance is a challenging aspect of academic literature.

They were built to educate students (learners) and to overcome their failures. Students' academic performance is affected by several factors which include students' learning skills, parental background, classmates' influence, teachers' quality, and infrastructure. Academic performance includes your grades, accolades, extra curriculum activities, experiences that demonstrate accomplishment, and students' passion. In recent research where researchers have found binge drinking is lowering academic performance including lowered GPA.

Peer pressure or we can say that students do try to impose things, especially bad habits. They start drinking to look cool ignoring the fact that it may ruin their school life. Improved student grades and Faculty resource allocation. According to a recent study released by America's Promised Alliance; the U.S. has a 30% rate of students who are failing to graduate high school. A national study indicates that college students with an "A" average consume an average of only 3.3 drinks per week, while "D" students consume an average of 9 drinks per week.

Factors like an unhealthy environment, Procrastination on assignments, difficulty planning and organizing to complete tasks, missing classes, etc. are the issues that affect students' grades. But Alcohol is the cover of the book. The results we found with our selected data set did surprise us.

2. LITERATURE REVIEW

In educational institutions, success is measured by academic performance or how well a student meets standard eligibility by that institution. The student’s performance can be recognized by parents, teachers, and employers. Education is not the only road to be successful in the real working world, much an attempt is made to identify, evaluate, track, and inspire the development of college students in schools. Countless studies have been done, to define the factors which are affecting students' universal achievement. Any student's achievement and performance in the school or a university is depended on different types of factors, such as personal, teacher, and institutional factors. School plays a crucial role in producing better quality students who will be great leaders of their countries.

Researchers have seen that family, home, demographic, school, and environment are factors that contribute to a student's academic performance. The role of a family is to provide for, educate, and protect their children. As much as we say teachers play an important part in a student’s life, so do the family, and the parents, which form an integral part of education.

Socioeconomic factors like attendance in the class, annual income of the family, their parents and guardian’s education, the ratio of the teacher-student in the school, number of trained teachers in school, sex of student, and travel time from home to school, all effects on the student’s performances. Teachers and Professors may forget the students who are lagging or are not that good at cramming books. This may lead to one of the reasons which may affect his/her performance. Due to this, students don’t feel like having an open mind or opening with their teachers about their feelings.

The same situation is with the parents of the students, they may have a well-balanced relationship with their kids or worst. The high school era decides whether the student is going to be a great person in the real world or not. Having a good relationship with teachers and parents makes a huge difference.

3. METHODOLOGY

In our project, we are using Multi Linear Regression Model. Regression analysis is the study of relationships between variables. If there are several explanatory variables, it is called multiple regression.

General Multiple Regression Equation: Predicted: Y = a + b1X1 + b2X2 + … + bkXk Collectively a and bs in the equation is called the regression coefficients. By applying Multi Linear Regression Model, we get this coefficient.

Procedure:

1. Data Acquisition
2. Exploratory Data Analysis
3. Data Preprocessing
4. Model Building
5. Model Evaluation

4. DATA ACIQUISTION

4.1 DATA DESCRIPTION

The data attributes include student grades, demographic, social and school related features and it was collected by using school reports and questionnaires. This data approach student achievement in secondary education (High school) of two Portuguese schools namely Gabriel Pereira school, the rest belong to Mousinho da Silveira school.

The datasets are provided regarding the student performance in Portuguese language subject. There are 33 variables in the data; and each variable has its own weightage. There are 649 Observations in total. The target variable is G3 (final grade), and some other variables which are affecting this variable. The 28 variables are categorical and remaining 5 are real numerical features (non – categorical). The variables are bifurcated further in Measures; 17 are Nominal and 11 are Ordinal and the rest of the 5 variables are Scale. The scale variables are Age, Absence, G1, G2, G3. Our target variable is scale, so we are using a regression model for prediction in this project.

`

4.2 VARAIABLE DESCRIPTION

Table 4.2.1: *Variable Description Table:*

#	Variable	Description
1	School	student's school
2	Sex	student's sex
3	Age	student's age
4	Address	student's home address type
5	Famsize	family size
6	Pstatus	parent's cohabitation status
7	Medu	mother's education
8	Fedu	father's education
9	Mjob	mother's job
10	Fjob	father's job
11	Reason	reason to choose this school
12	guardian	student's guardian
13	Traveltime	home to school travel time
14	Study time	weekly study time
15	Failures	number of past class failures
16	Schoolsup	extra educational support
17	Famsup	family educational support
18	Paid	extra paid classes within course subject
19	Activities	extracurricular activities
20	Nursery	attended nursery school
21	Higher	wants to take higher education
22	Internet	Internet access at home
23	Romantic	with a romantic relationship
24	Farrell	quality of family relationships
25	Free time	free time after school
26	Gout	going out with friends
27	Dalc	workday alcohol consumption
28	Walc	weekend alcohol consumption
29	Health	current health status
30	Absences	number of school absences
31	G1	first period grade
32	G2	second period grade
33	G3	final grade

- The data set downloaded from https://www.kaggle.com/datasets/student-performance

5. EXPLORATORY DATA ANALYSIS

We have created distribution charts and tables using tableau tool and SPSS. Given below are the explanations of each figure.

Figure 5.1: The school graph shows the distribution based on number of students studying in GP and MS school. In the graph, we can observe that GP (Gabriel Pereira school) has got the more numbers of school than the MS (Mousinho da Silveira school).

Figure 5.2: In the age graph, we can observe that the students of age 17 are the maximum to join the school and least is the age 22.

Figure 5.3: The Graph representing distribution based on gender and non-presence in class (absentees):

In the gender graph, F and M represents as female and male respectively. The number of females is more than the number of males in the schools.

Figure 5.4: In the absences graph, we can see that the observation is not linear.

Figure 5.5: The graph shows the relation between students consuming alcohol Weekly vs their Grades:

This graph follows a decreasing trend of students consuming less alcohol weekly performing the best and the worse who drink the most.

Figure 5.6: The graph shows the relation between students consuming alcohol daily vs their Grades:

The graph describes that the students who drink less gets better grades. We have decreasing trend except for the highest level of alcohol consumption

Figure 5.7: The graph shows the relation between G3 versus failure.

Most of the students have graded 11, and 15 people got 0 grading in their academics.

Figure 5.8: The graph shows the relation between G3 versus G3.

We can see as the G2 increases the G3 is increasing and there is a linear relationship between G2 and G3

Figure 5.9, 5.10: The graphs show the alcohol consumption depending on the living area in both the workdays and the weekend, and in both graphs, we can observe that the people are consuming very low alcohol on the scale 1-5

Figure 5.11. The graph shows the relation between students consuming alcohol daily vs their Grades. The graph describes that the students who drink less gets better grades. We have decreasing trend except for the highest level of alcohol consumption.

Figure 5.12: Mean value of G3 over weekend alcohol consumption. The graph illustrates the relation between students consuming alcohol Weekly vs their Grades. This graph follows a decreasing trend, students consuming less alcohol weekly perform the best and worse who drink the most.

Figure 5.13: Relationship between Study Time vs their Grades:

The students who study for => 10 hours and 2 to 5 hours (3,4) has the highest grade of 19 and are scoring above 6 marks. The graphs illustrate that 2, 4 students have topped with 19 whereas 1, 2 has most chances of getting failed and the highest scored marks are 18 more times gets lesser grading and vice versa.

Figure 5.14: Relation between students Failure vs their Grades:

This graph shows a decreasing trend, students' failure is getting decreased, person who fails more times gets lesser grading and vice versa.

Figure 5.15: Relation between student's Famrel (Family relationship) vs their G3 (Final Grades) in this graph: We can see that students who have excellent relations with their family are getting the zero grades (in the first row). In contrast, the students who have a neutral relationship with their family are scoring the highest score i.e., 19.

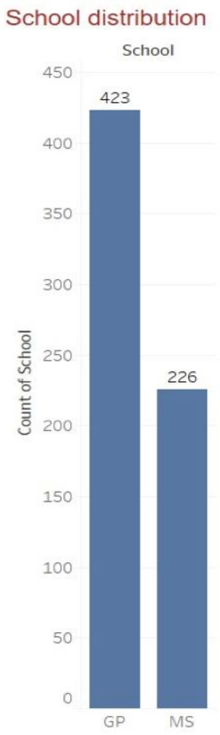


Figure 5.1 : School Distribution

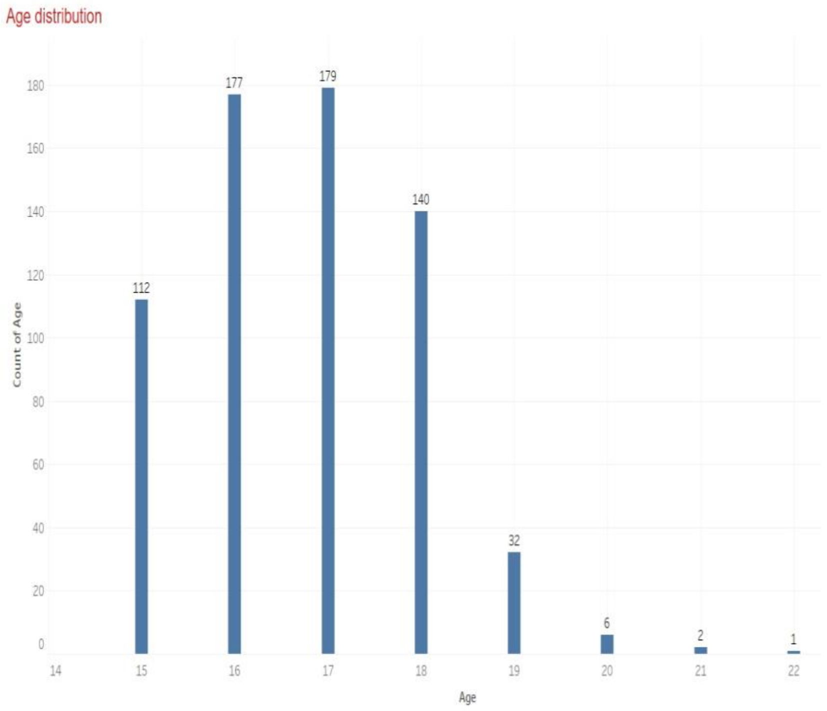


Figure 5.2 : Age Distribution

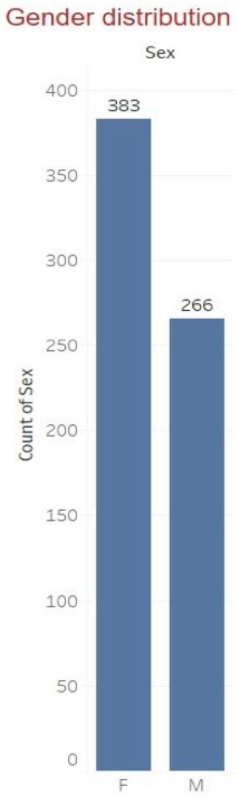


Figure 5.3 : Gender Distribution

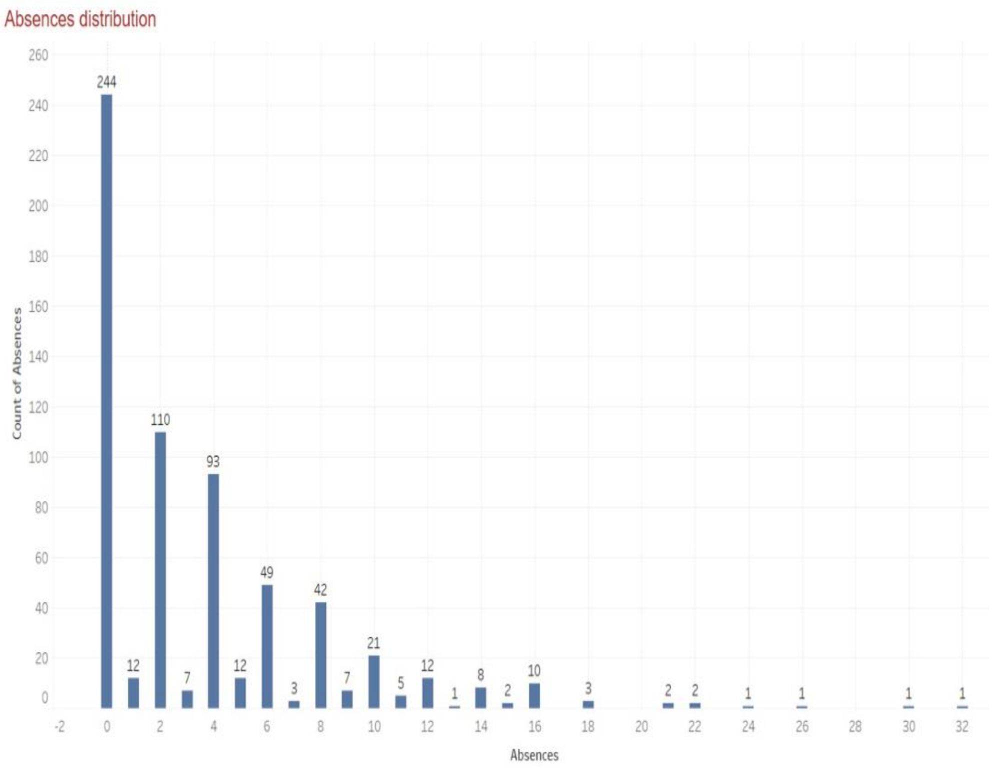


Figure 5.4 : Absences Distribution

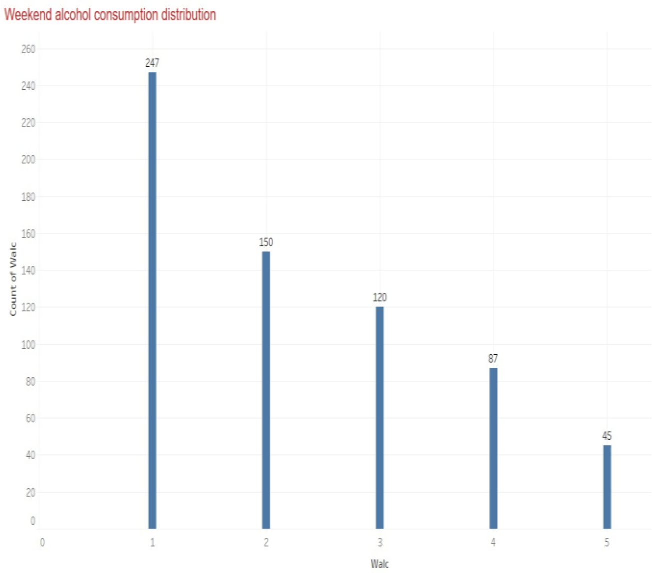


Figure 5.5: Weekend alcohol consumption distribution

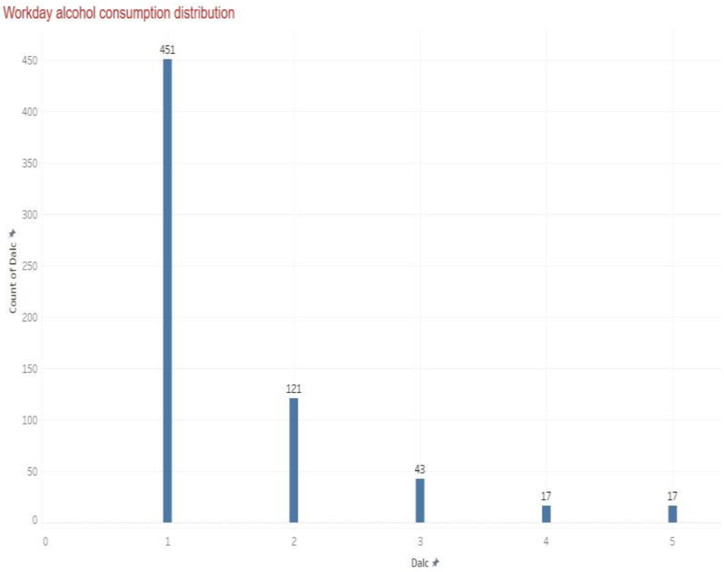


Figure 5.6 : Workday alcohol consumption distribution

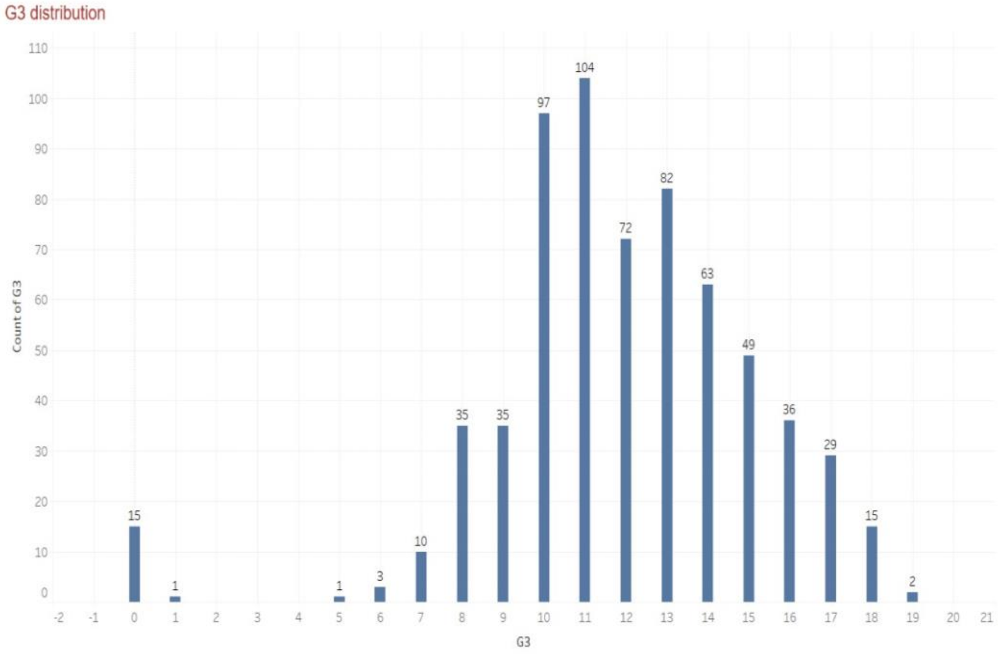


Figure 7: G3 Distribution

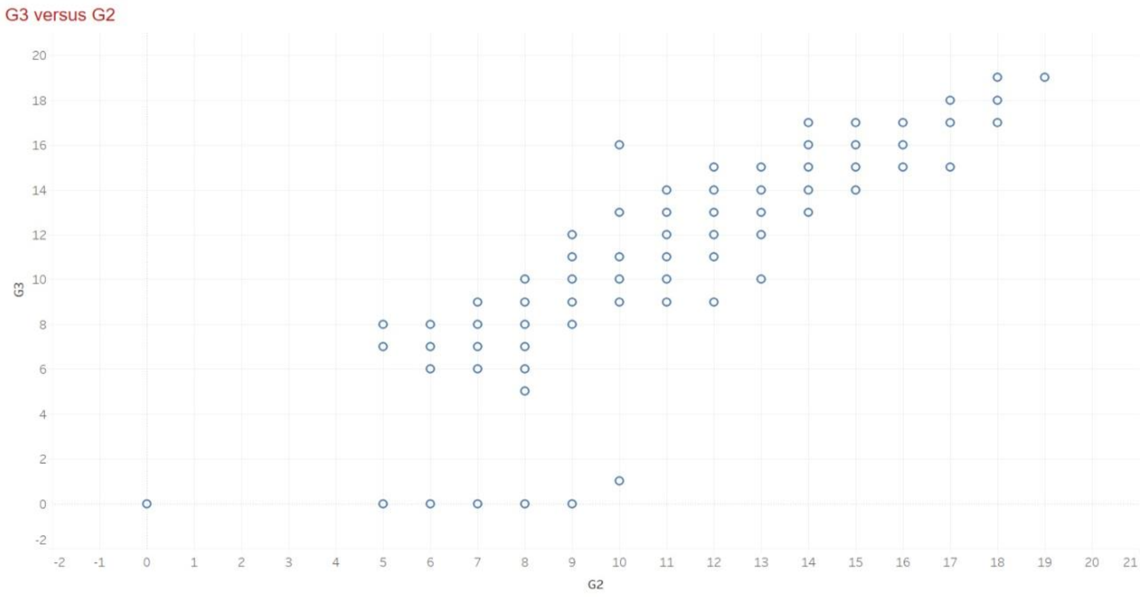


Figure 5.8 : G3 versus G2

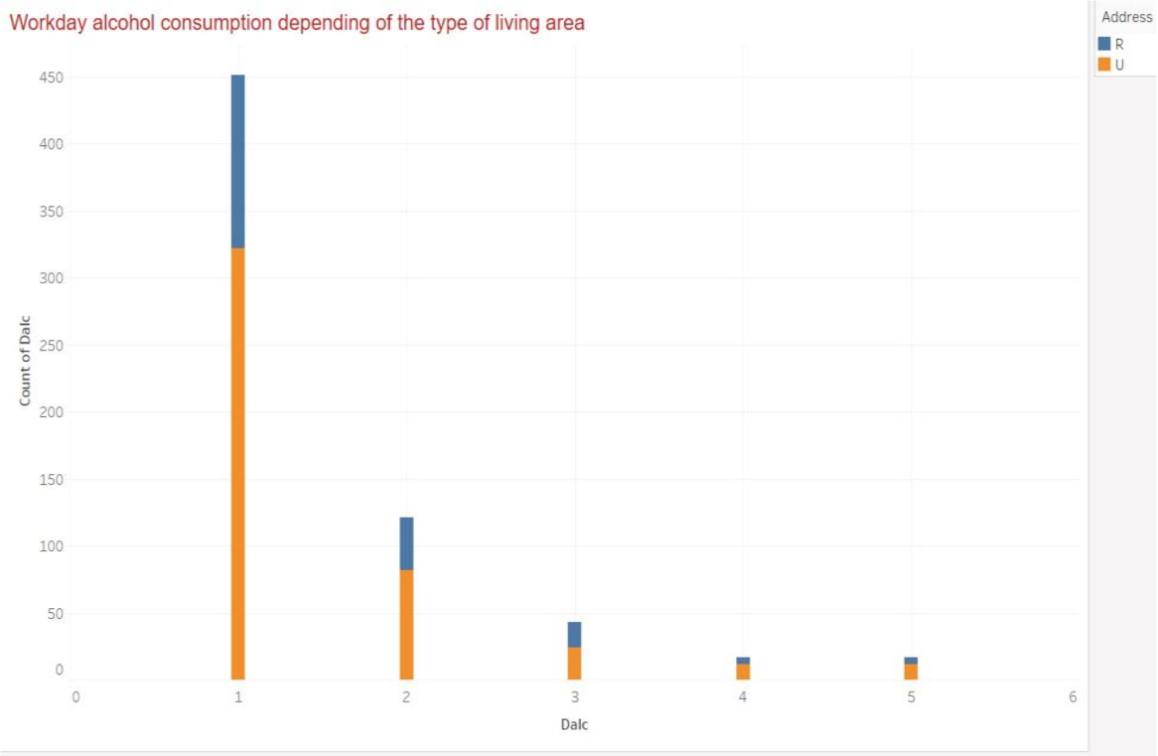


Figure 5.9 : Workday Alcohol consumption depending on the type of living area

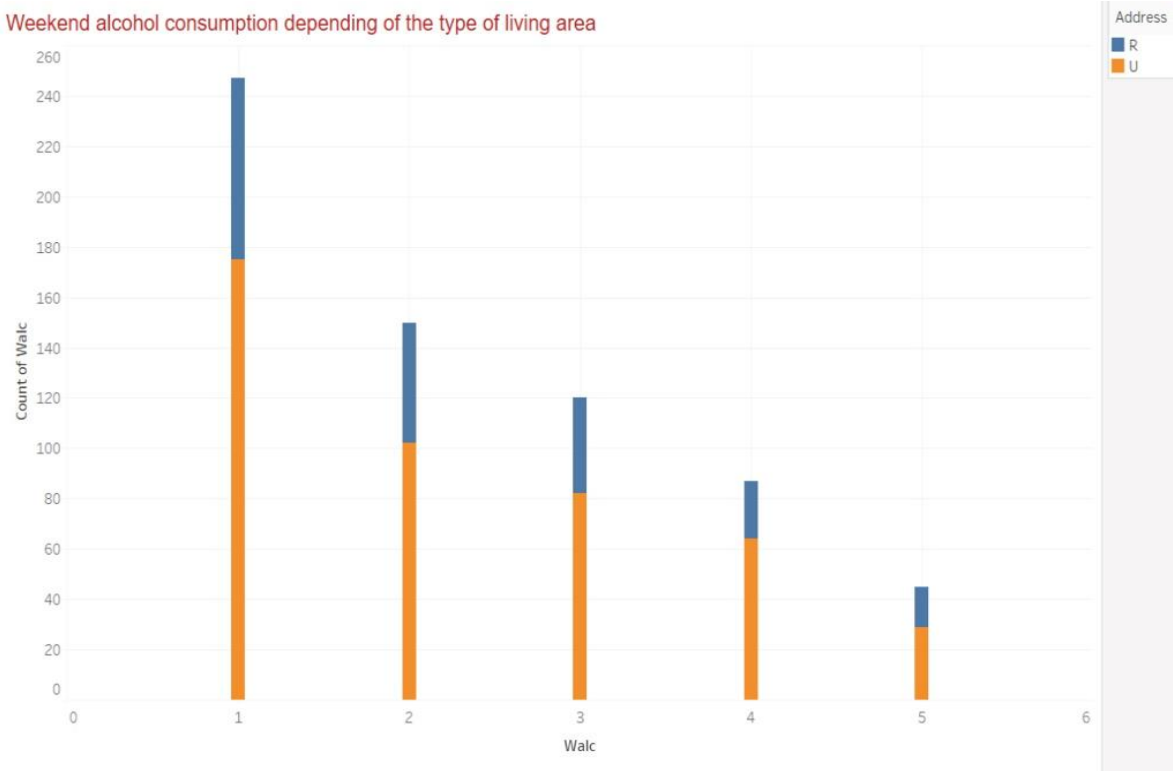


Figure 5.10: Weekend Alcohol consumption depending on the type of living area

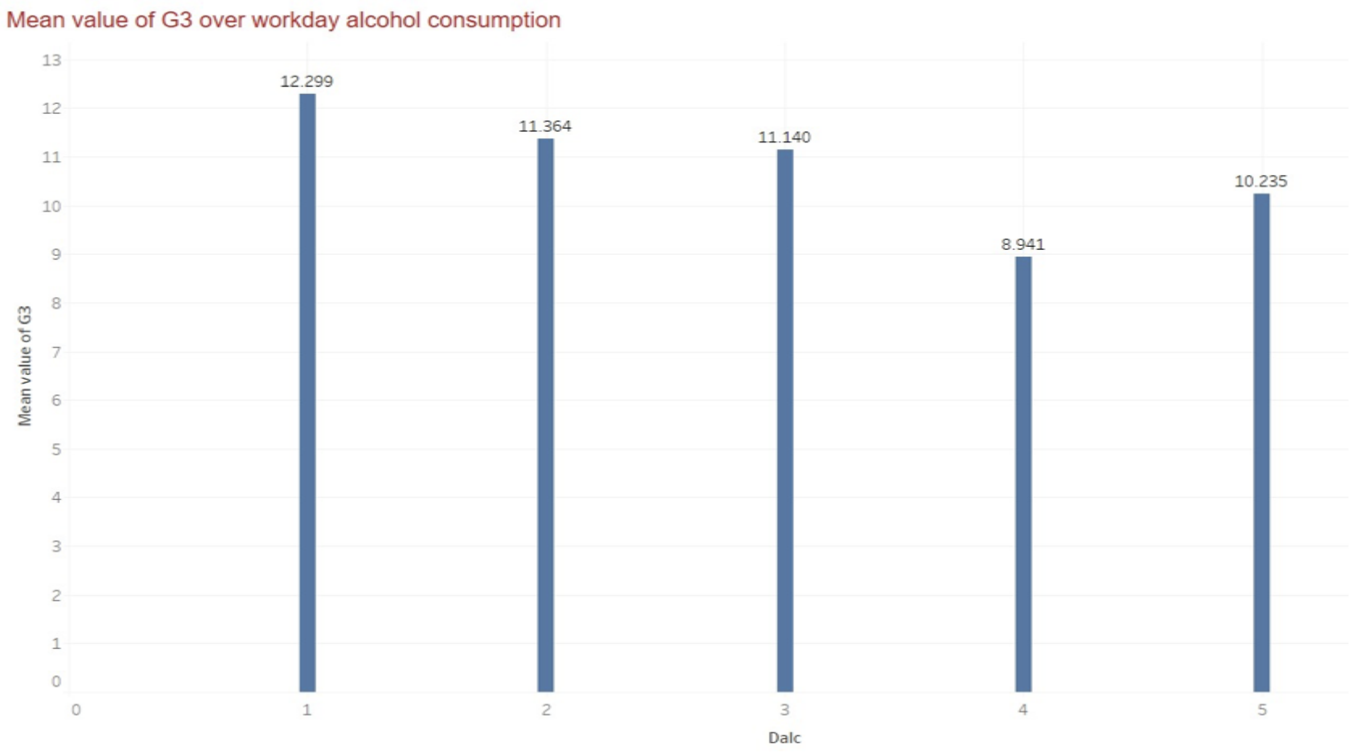


Figure 5.11: Mean value of G3 over workday alcohol consumption

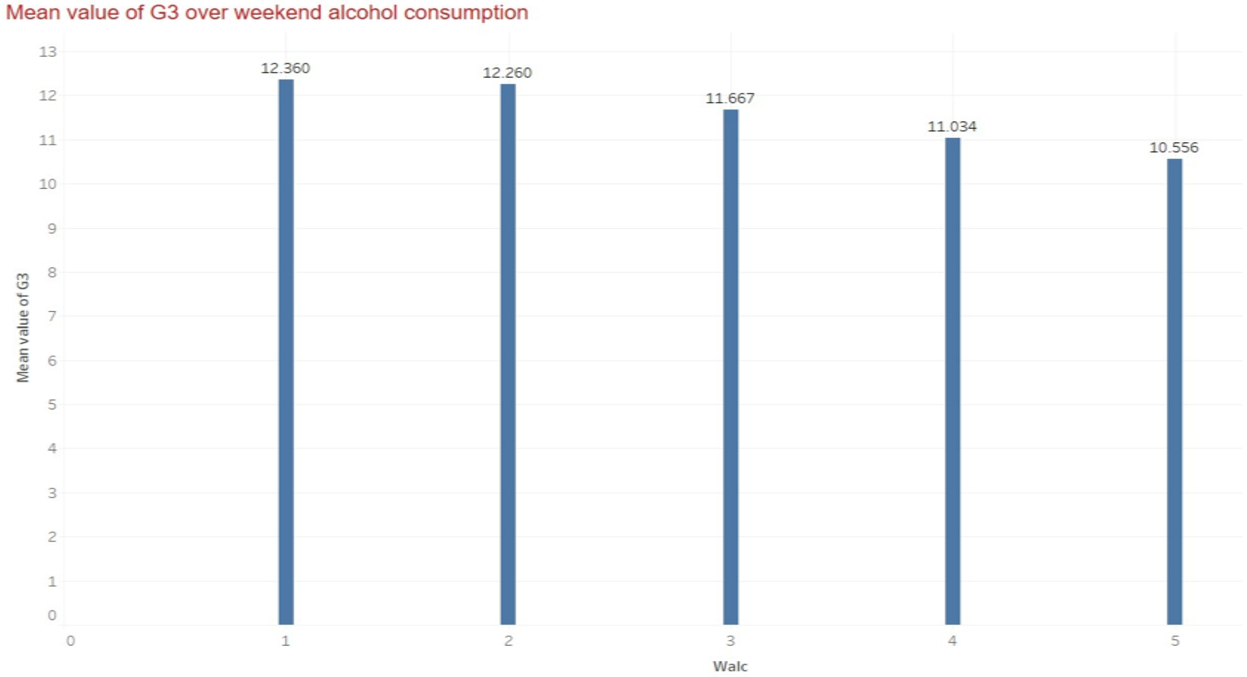


Figure 5.12: Mean value of G3 over weekend alcohol consumption

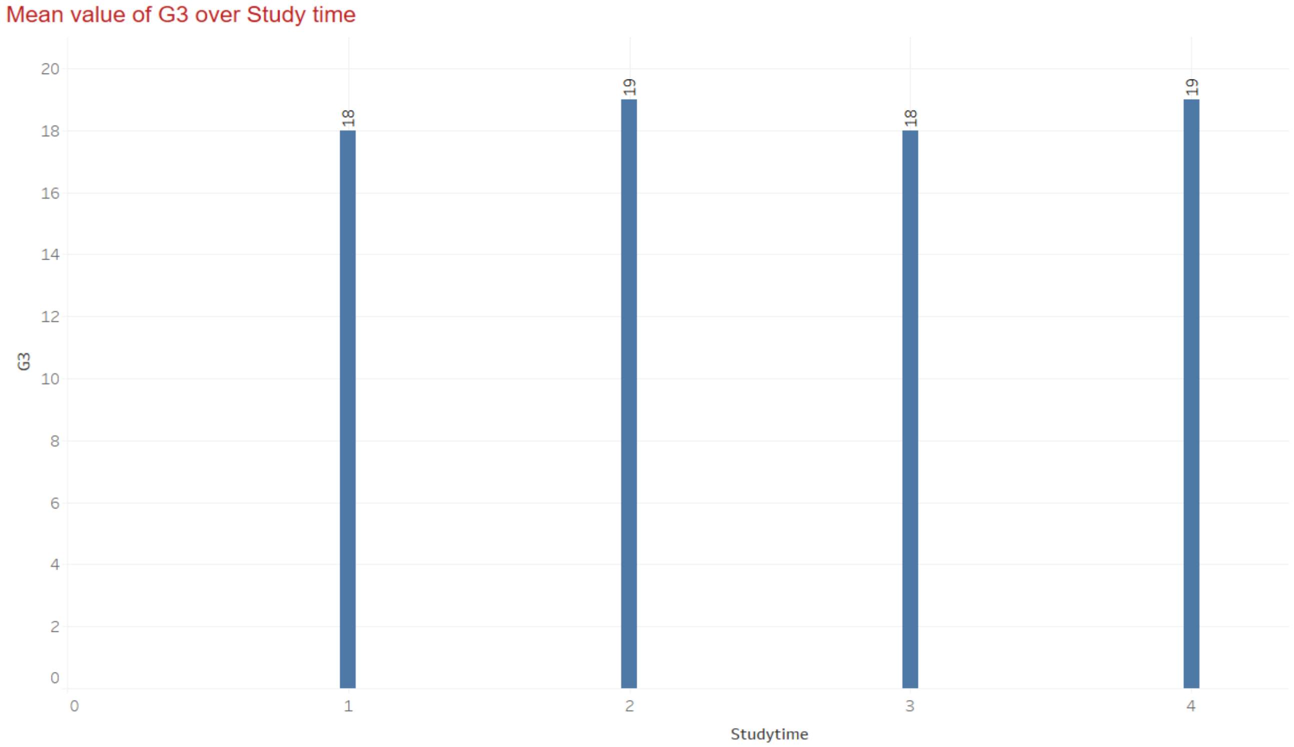


Figure 5.13: Mean value of G3 over study time

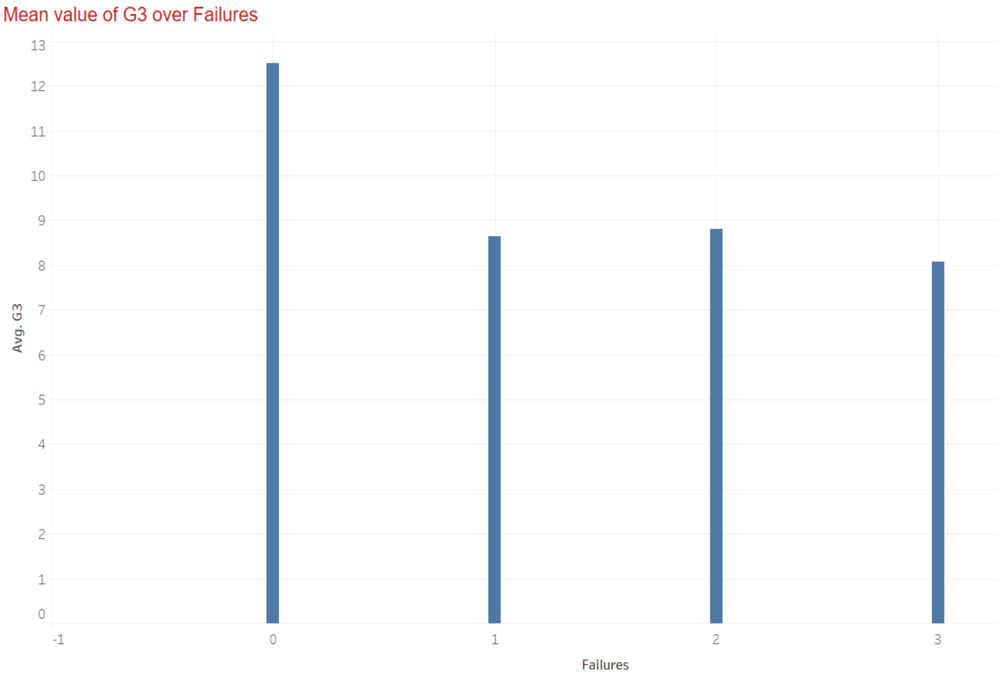
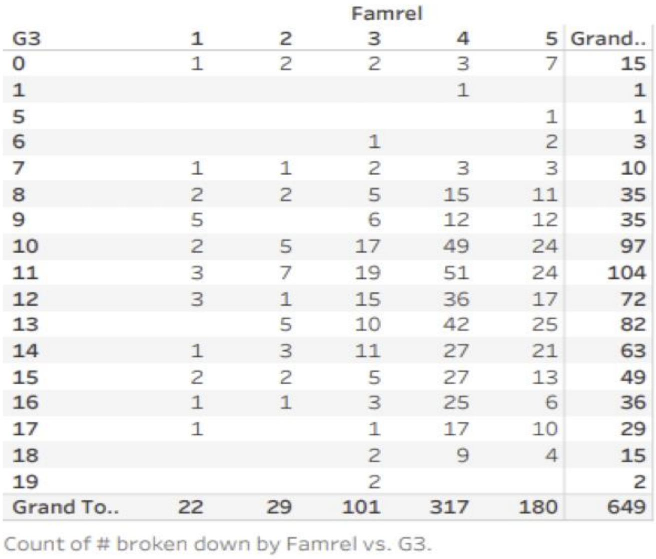


Figure 5.14: Mean value of G3 over failures



			Famrel			
G3	1	2	3	4	5	Grand..
0	1	2	2	3	7	15
1				1		1
5					1	1
6			1		2	3
7	1	1	2	3	3	10
8	2	2	5	15	11	35
9	5		6	12	12	35
10	2	5	17	49	24	97
11	3	7	19	51	24	104
12	3	1	15	36	17	72
13		5	10	42	25	82
14	1	3	11	27	21	63
15	2	2	5	27	13	49
16	1	1	3	25	6	36
17	1		1	17	10	29
18			2	9	4	15
19			2			2
Grand To..	22	29	101	317	180	649

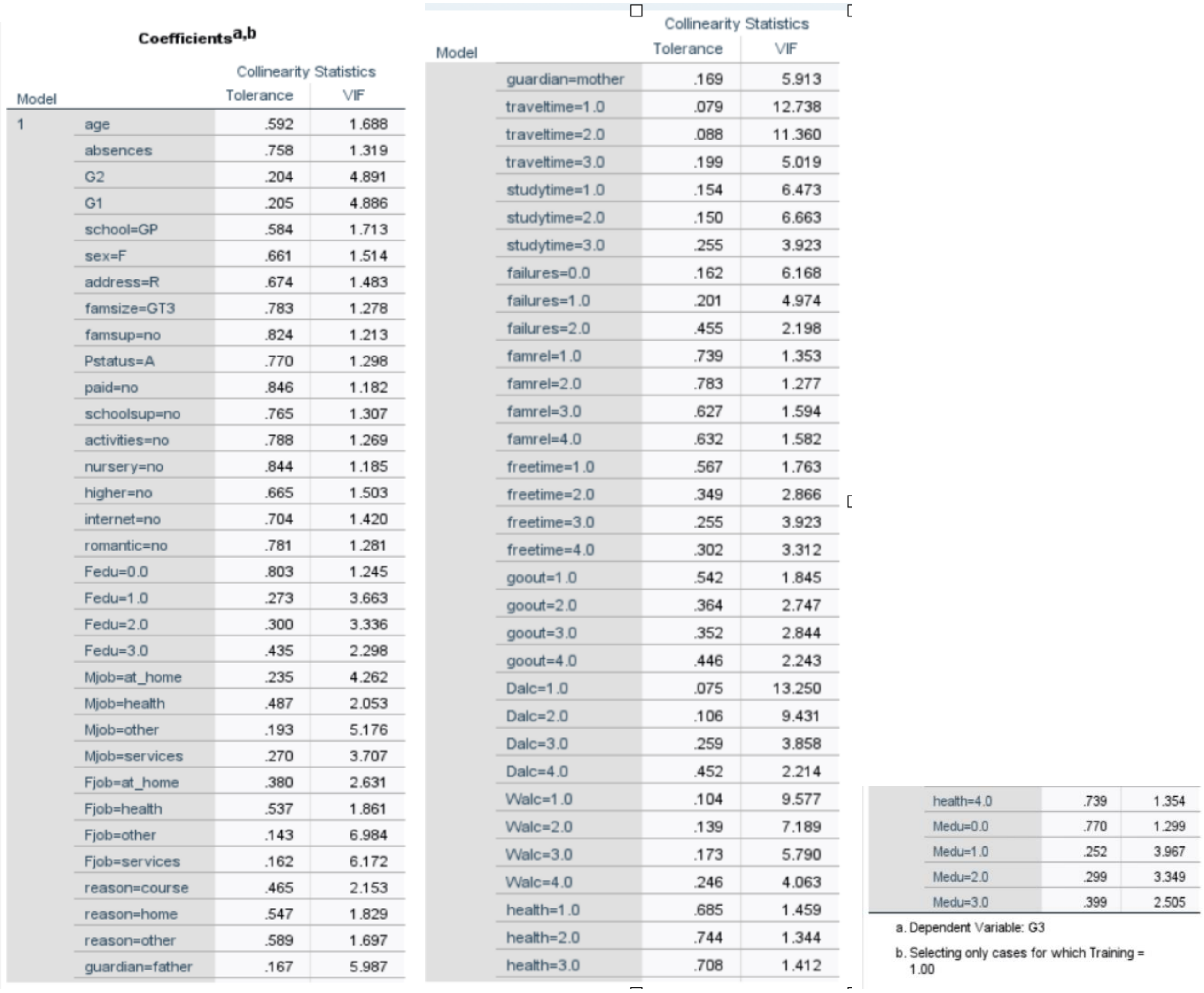
Count of # broken down by Famrel vs. G3.

-

Figure 5.15: Influence of family relation on their grades

6. BUILD MODEL

We have created dummy variables for all the 28 categorical predictors. We have split the data into Training and Test in the ratio 3:1 i.e. (train 488, Test 161) to prevent overfitting. Found multicollinearity in our data set so removed the variables which have the VIF value greater than 5. Now ran a multi linear regression with the selected variables on the training dataset



Coefficients[a,b]

Model		Collinearity Statistics	
		Tolerance	VIF
1	age	.592	1.688
	absences	.758	1.319
	G2	.204	4.891
	G1	.205	4.886
	school=GP	.584	1.713
	sex=F	.661	1.514
	address=R	.674	1.483
	famsize=GT3	.783	1.278
	famsup=no	.824	1.213
	Pstatus=A	.770	1.298
	paid=no	.846	1.182
	schoolsup=no	.765	1.307
	activities=no	.788	1.269
	nursery=no	.844	1.185
	higher=no	.665	1.503
	internet=no	.704	1.420
	romantic=no	.781	1.281
	Fedu=0.0	.803	1.245
	Fedu=1.0	.273	3.663
	Fedu=2.0	.300	3.336
	Fedu=3.0	.435	2.298
	Mjob=at_home	.235	4.262
	Mjob=health	.487	2.053
	Mjob=other	.193	5.176
	Mjob=services	.270	3.707
	Fjob=at_home	.380	2.631
	Fjob=health	.537	1.861
	Fjob=other	.143	6.984
	Fjob=services	.162	6.172
	reason=course	.465	2.153
	reason=home	.547	1.829
	reason=other	.589	1.697
	guardian=father	.167	5.987
	guardian=mother	.169	5.913
	traveltime=1.0	.079	12.738
	traveltime=2.0	.088	11.360
	traveltime=3.0	.199	5.019
	studytime=1.0	.154	6.473
	studytime=2.0	.150	6.663
	studytime=3.0	.255	3.923
	failures=0.0	.162	6.168
	failures=1.0	.201	4.974
	failures=2.0	.455	2.198
	famrel=1.0	.739	1.353
	famrel=2.0	.783	1.277
	famrel=3.0	.627	1.594
	famrel=4.0	.632	1.582
	freetime=1.0	.567	1.763
	freetime=2.0	.349	2.866
	freetime=3.0	.255	3.923
	freetime=4.0	.302	3.312
	goout=1.0	.542	1.845
	goout=2.0	.364	2.747
	goout=3.0	.352	2.844
	goout=4.0	.446	2.243
	Dalc=1.0	.075	13.250
	Dalc=2.0	.106	9.431
	Dalc=3.0	.259	3.858
	Dalc=4.0	.452	2.214
	Walc=1.0	.104	9.577
	Walc=2.0	.139	7.189
	Walc=3.0	.173	5.790
	Walc=4.0	.246	4.063
	health=1.0	.685	1.459
	health=2.0	.744	1.344
	health=3.0	.708	1.412
	health=4.0	.739	1.354
	Medu=0.0	.770	1.299
	Medu=1.0	.252	3.967
	Medu=2.0	.299	3.349
	Medu=3.0	.399	2.505

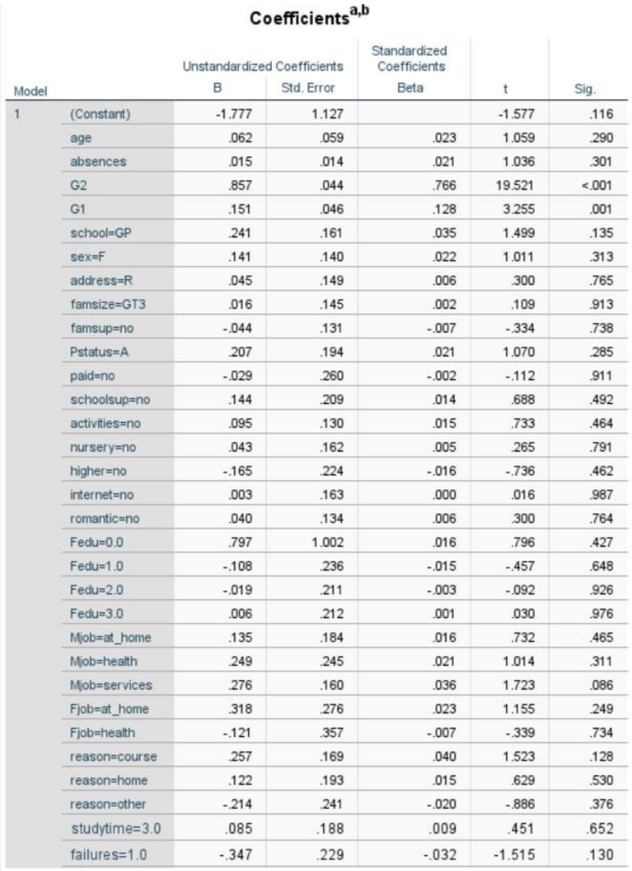
a. Dependent Variable: G3

b. Selecting only cases for which Training = 1.00

Figure 10: VIF

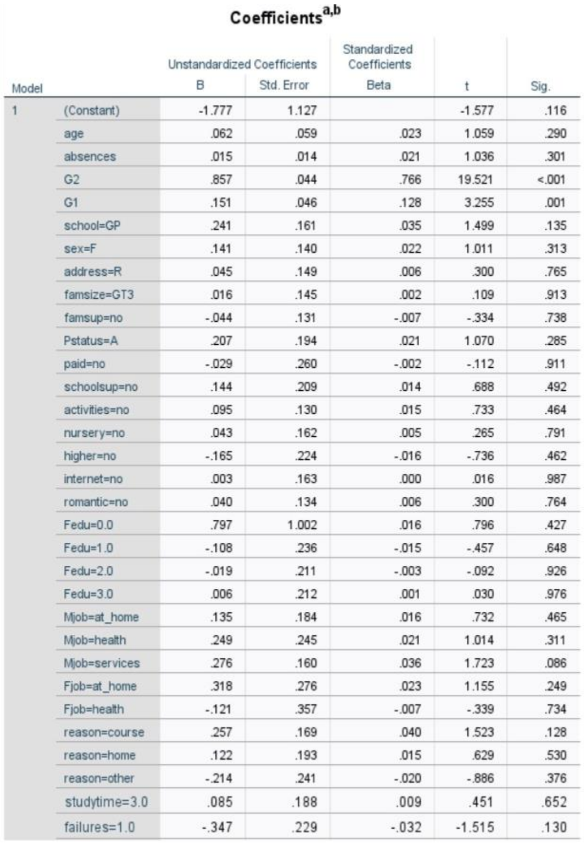
7. PREDECTIVE MODEL RESULTS

The regression coefficients from the multi linear regression model as shown in the figure below



Coefficients^{a,b}

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.777	1.127		-1.577	.116
	age	.062	.059	.023	1.059	.290
	absences	.015	.014	.021	1.036	.301
	G2	.857	.044	.766	19.521	<.001
	G1	.151	.046	.128	3.255	.001
	school=GP	.241	.161	.035	1.499	.135
	sex=F	.141	.140	.022	1.011	.313
	address=R	.045	.149	.006	.300	.765
	famsize=GT3	.016	.145	.002	.109	.913
	famsup=no	-.044	.131	-.007	-.334	.738
	Pstatus=A	.207	.194	.021	1.070	.285
	paid=no	-.029	.260	-.002	-.112	.911
	schoolsup=no	.144	.209	.014	.688	.492
	activities=no	.095	.130	.015	.733	.464
	nursery=no	.043	.162	.005	.265	.791
	higher=no	-.165	.224	-.016	-.736	.462
	internet=no	.003	.163	.000	.016	.987
	romantic=no	.040	.134	.006	.300	.764
	Fedu=0.0	.797	1.002	.016	.796	.427
	Fedu=1.0	-.108	.236	-.015	-.457	.648
	Fedu=2.0	-.019	.211	-.003	-.092	.926
	Fedu=3.0	.006	.212	.001	.030	.976
	Mjob=at_home	.135	.184	.016	.732	.465
	Mjob=health	.249	.245	.021	1.014	.311
	Mjob=services	.276	.160	.036	1.723	.086
	Fjob=at_home	.318	.276	.023	1.155	.249
	Fjob=health	-.121	.357	-.007	-.339	.734
	reason=course	.257	.169	.040	1.523	.128
	reason=home	.122	.193	.015	.629	.530
	reason=other	-.214	.241	-.020	-.886	.376
	studytime=3.0	.085	.188	.009	.451	.652
	failures=1.0	-.347	.229	-.032	-1.515	.130



Coefficients^{a,b}

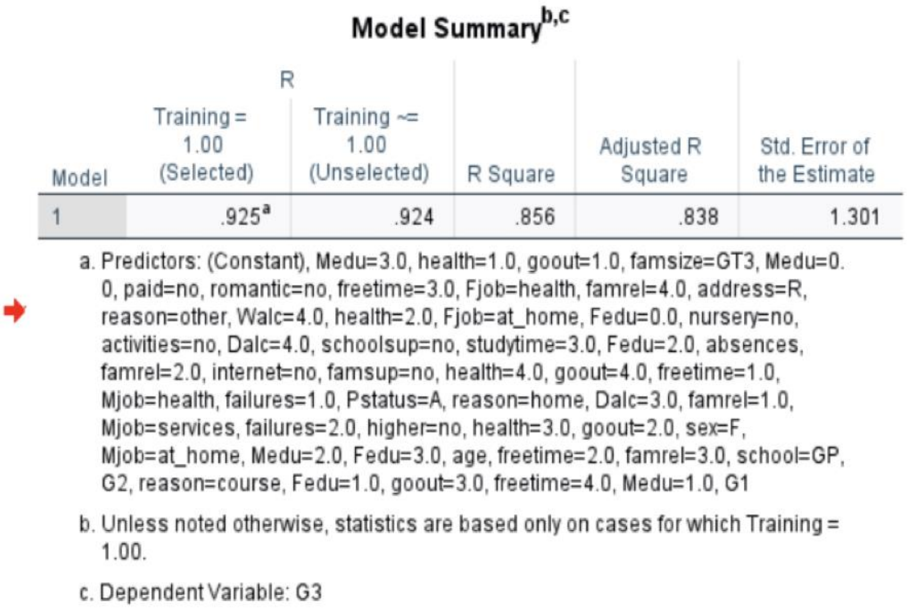
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.777	1.127		-1.577	.116
	age	.062	.059	.023	1.059	.290
	absences	.015	.014	.021	1.036	.301
	G2	.857	.044	.766	19.521	<.001
	G1	.151	.046	.128	3.255	.001
	school=GP	.241	.161	.035	1.499	.135
	sex=F	.141	.140	.022	1.011	.313
	address=R	.045	.149	.006	.300	.765
	famsize=GT3	.016	.145	.002	.109	.913
	famsup=no	-.044	.131	-.007	-.334	.738
	Pstatus=A	.207	.194	.021	1.070	.285
	paid=no	-.029	.260	-.002	-.112	.911
	schoolsup=no	.144	.209	.014	.688	.492
	activities=no	.095	.130	.015	.733	.464
	nursery=no	.043	.162	.005	.265	.791
	higher=no	-.165	.224	-.016	-.736	.462
	internet=no	.003	.163	.000	.016	.987
	romantic=no	.040	.134	.006	.300	.764
	Fedu=0.0	.797	1.002	.016	.796	.427
	Fedu=1.0	-.108	.236	-.015	-.457	.648
	Fedu=2.0	-.019	.211	-.003	-.092	.926
	Fedu=3.0	.006	.212	.001	.030	.976
	Mjob=at_home	.135	.184	.016	.732	.465
	Mjob=health	.249	.245	.021	1.014	.311
	Mjob=services	.276	.160	.036	1.723	.086
	Fjob=at_home	.318	.276	.023	1.155	.249
	Fjob=health	-.121	.357	-.007	-.339	.734
	reason=course	.257	.169	.040	1.523	.128
	reason=home	.122	.193	.015	.629	.530
	reason=other	-.214	.241	-.020	-.886	.376
	studytime=3.0	.085	.188	.009	.451	.652
	failures=1.0	-.347	.229	-.032	-1.515	.130

Figure 7.1: Regression Coefficients

The performance metrics of this model is shown in the below table

Table 7. 1: *Performance metrics of the model*

R-Squared (Train)	0.856	R-Squared (Test)	0.833967
Adjusted R-Squared (Train)	0.838	Adjusted R-Squared (Test)	0.746997
Standard error of estimate (Train)	1.301	Standard error of estimate (Test)	1.327029
MAE (Train)	0.719262	MAE (Test)	0.84472
RMSE (Train)	1.266692	RMSE (Test)	1.318761



Model Summary[b,c]

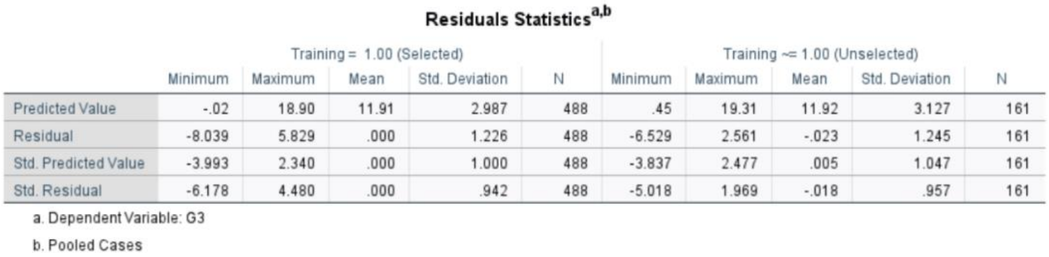
	R				
Model	Training = 1.00 (Selected)	Training ~= 1.00 (Unselected)	R Square	Adjusted R Square	Std. Error of the Estimate
1	.925[a]	.924	.856	.838	1.301

a. Predictors: (Constant), Medu=3.0, health=1.0, goout=1.0, famsize=GT3, Medu=0.0, paid=no, romantic=no, freetime=3.0, Fjob=health, famrel=4.0, address=R, reason=other, Walc=4.0, health=2.0, Fjob=at_home, Fedu=0.0, nursery=no, activities=no, Dalc=4.0, schoolsup=no, studytime=3.0, Fedu=2.0, absences, famrel=2.0, internet=no, famsup=no, health=4.0, goout=4.0, freetime=1.0, Mjob=health, failures=1.0, Pstatus=A, reason=home, Dalc=3.0, famrel=1.0, Mjob=services, failures=2.0, higher=no, health=3.0, goout=2.0, sex=F, Mjob=at_home, Medu=2.0, Fedu=3.0, age, freetime=2.0, famrel=3.0, school=GP, G2, reason=course, Fedu=1.0, goout=3.0, freetime=4.0, Medu=1.0, G1

b. Unless noted otherwise, statistics are based only on cases for which Training = 1.00.

c. Dependent Variable: G3

Figure 11: Model Summary



Residuals Statistics[a,b]

	Training = 1.00 (Selected)					Training ~= 1.00 (Unselected)				
	Minimum	Maximum	Mean	Std. Deviation	N	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-.02	18.90	11.91	2.987	488	.45	19.31	11.92	3.127	161
Residual	-8.039	5.829	.000	1.226	488	-6.529	2.561	-.023	1.245	161
Std. Predicted Value	-3.993	2.340	.000	1.000	488	-3.837	2.477	.005	1.047	161
Std. Residual	-6.178	4.480	.000	.942	488	-5.018	1.969	-.018	.957	161

a. Dependent Variable: G3

b. Pooled Cases

Figure 7.3: Residual Statistics

The Multi linear regression gave the R-squared value of 0.85 which indicates a good accuracy. Hence, we can use this model to predict the student performance when all the predictors are specified.

CONCLUSION

Surprisingly, we have concluded that Alcohol consumption has less effect on student's grade. The students who get good marks in G1 and G2 will also get good grade in G3. The most important feature which effected on students' performance is 'Failures' (number of past class failure). Students who have failed more no of classes are categorized as they will not be able to get good grades in future too.

Another factor we concluded is 'Study time', students are not using their Study time wisely to study for their classes. Parents education also matters, whether they can help their child in studying. G1, G2, Failures, higher education, school, studytime, Dalc, are some of the factors which are affecting student performance.

According to the data and the observations made, how is having a good relationship with the family affects a student's performance. Well, here you should always have a different eye on things and that's what the Analysts do. So, having a very good relationship with family means that are lenient with them and not much focusing on their academics; students should have a balanced relationship with their family to achieve good academic grades and enjoy their personal life.

REFERENCES

1. https://econweb.ucsd.edu/~jbetts/Pub/A39%20Research%20Brief%20RB_803JBRB.pdf

   Public policy institute of California 500 Washington Street, Suite 800 • San Francisco, California 94111

2. https://globaljournals.org/GJMBR_Volume12/3-Factors-Affecting-Students-Academic.pdf by Irfan Mustaq & Shabana Nawaz Khan, Mohammad Ali Jinnah University, Islamabad, Pakistan.

3. Data Set: https://www.kaggle.com/datasets/student-performance