EECS 4412 A Data Mining

Project Phase 2

Group Member:

Trevor Ramtahal (216771875)

Ka Hei Chan (216328387)

Yongjie Ba (215122856)

Date: June 14, 2022

Professor: Habib-ur Rehman

Appendix A

I **Yongjie Ba** student ID # **215122856** acknowledge that I have contributed at least 30% time and effort to the preparation of this report and work discussed herein.

Student Signature:_____

I *Trevor Ramtahal* student ID # *216771875* acknowledge that I have contributed at least 30% time and effort to the preparation of this report and work discussed herein.

Student Signature: _____

I *Ka Hei Chan* student ID # *216328387* acknowledge that I have contributed at least 30% time and effort to the preparation of this report and work discussed herein.

Student_Signature: _____

Task 1: Describing the Data

For the Family status data set:

sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	guardian	failures	famsup	famrel	absences	G1	G2	G3	
F	18	U	GT3	A	4	4	at_home	teacher	mother	0	no	4	6	5	6	6	
F	17	U	GT3	T	1	1	at_home	other	father	0	yes	5	4	5	5	6	
F	15	U	LE3	T	1	1	at_home	other	mother	3	no	4	10	7	8	10	
F	15	U	GT3	T	4	2	health	services	mother	0	yes	3	2	15	14	15	
F	16	U	GT3	T	3	3	other	other	father	0	yes	4	4	6	10	10	
M	16	U	LE3	T	4	3	services	other	mother	0	yes	5	10	15	15	15	
M	16	U	LE3	T	2	2	other	other	mother	0	no	4	0	12	12	11	
F	17	U	GT3	A	4	4	other	teacher	mother	0	yes	4	6	6	5	6	
M	15	U	LE3	A	3	2	services	other	mother	0	yes	4	0	16	18	19	
M	15	U	GT3	T	3	4	other	other	mother	0	yes	5	0	14	15	15	
F	15	U	GT3	T	4	4	teacher	health	mother	0	yes	3	0	10	8	9	
F	15	U	GT3	T	2	1	services	other	father	0	yes	5	4	10	12	12	
M	15	U	LE3	T	4	4	health	services	father	0	yes	4	2	14	14	14	
M	15	U	GT3	T	4	3	teacher	other	mother	0	yes	5	2	10	10	11	
M	15	U	GT3	A	2	2	other	other	other	0	yes	4	0	14	16	16	
F	16	U	GT3	T	4	4	health	other	mother	0	yes	4	4	14	14	14	
F	16	U	GT3	T	4	4	services	services	mother	0	yes	3	6	13	14	14	
F	16	U	GT3	T	3	3	other	other	mother	0	yes	5	4	8	10	10	
M	17	U	GT3	T	3	2	services	services	mother	3	yes	5	16	6	5	5	
M	16	U	LE3	T	4	3	health	other	father	0	no	3	4	8	10	10	
M	15	U	GT3	T	4	3	teacher	other	mother	0	no	4	0	13	14	15	
M	15	U	GT3	T	4	4	health	health	father	0	yes	5	0	12	15	15	
M	16	U	LE3	T	4	2	teacher	other	mother	0	no	4	2	15	15	16	
M	16	U	LE3	T	2	2	other	other	mother	0	yes	5	0	13	13	12	
F	15	R	GT3	T	2	4	services	health	mother	0	yes	4	2	10	9	8	
F	16	U	GT3	T	2	2	services	services	mother	2	yes	1	14	6	9	8	
M	15	U	GT3	T	2	2	other	other	mother	0	yes	4	2	12	12	11	
M	15	U	GT3	T	4	2	health	services	mother	0	no	2	4	15	16	15	
M	16	U	LE3	A	3	4	services	other	mother	0	yes	5	4	11	11	11	
M	16	U	GT3	T	4	4	teacher	teacher	mother	0	yes	4	16	10	12	11	
M	15	U	GT3	T	4	4	health	services	mother	0	yes	5	0	9	11	12	
M	15	U	GT3	T	4	4	services	services	mother	0	yes	4	0	17	16	17	
M	15	R	GT3	T	4	3	teacher	at_home	mother	0	yes	4	0	17	16	16	
M	15	U	LE3	T	3	3	other	other	mother	0	no	5	0	8	10	12	
M	16	U	GT3	T	3	2	other	other	mother	0	yes	5	0	12	14	15	
F	15	U	GT3	T	2	3	other	other	father	0	yes	3	0	8	7	6	
M	15	U	LE3	T	4	3	teacher	services	mother	0	yes	5	2	15	16	18	
M	16	R	GT3	A	4	4	other	teacher	mother	0	yes	2	7	15	16	15	
F	15	R	GT3	T	3	4	services	health	mother	0	yes	4	2	12	12	11	
F	15	R	GT3	T	2	2	at_home	other	mother	0	yes	4	8	14	13	13	
F	16	U	LE3	T	2	2	other	other	mother	1	yes	3	25	7	10	11	
M	15	U	LE3	T	4	4	teacher	other	other	0	yes	5	8	12	12	12	
M	15	U	GT3	T	4	4	services	teacher	father	0	yes	4	2	19	18	18	
M	15	U	GT3	T	2	2	services	services	father	0	yes	5	0	8	8	11	
F	16	U	LE3	T	2	2	other	at_home	father	1	no	4	14	10	10	9	
F	15	U	LE3	A	4	3	other	other	mother	0	yes	5	8	8	8	6	
F	16	U	LE3	A	3	3	other	services	mother	0	yes	2	12	11	12	11	
M	16	U	GT3	T	4	3	health	services	mother	0	no	4	4	19	19	20	
M	15	U	GT3	T	4	2	teacher	other	mother	0	yes	4	2	15	15	14	
F	15	U	GT3	T	4	4	services	teacher	father	1	yes	4	2	7	7	7	

Adding a new feature: Ave G - the average grade of G1, G2 and G3

Then classify the data:

Set yellow background colour to the student whose guardian is the mother;

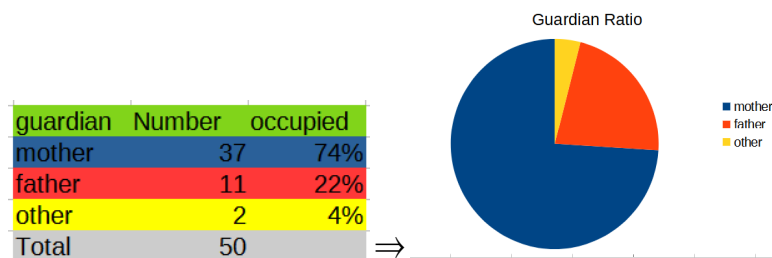Set blue background colour to the student whose guardian is the father;

Set white background colour to the student whose guardian is the other;

For detail please see the picture below;

sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	guardian	failures	famsup	famrel	absences	G1	G2	G3	Ave G
F	18	U	GT3	A		4	at_home	teacher	mother	0	no	4	6	5	6	6	6
F	16	U	LE3	T		2	other	at_home	father	1	no	4	14	10	10	9	10
M	16	U	LE3	T		3	health	other	father	0	no	3	4	8	10	10	9
F	15	U	LE3	T		1	at_home	other	mother	3	no	4	10	7	8	10	8
M	16	U	LE3	T		2	other	other	mother	0	no	4	0	12	12	11	12
M	15	U	LE3	T		3	other	other	mother	0	no	5	0	8	10	12	10
M	15	U	GT3	T		3	teacher	other	mother	0	no	4	0	13	14	15	14
M	16	U	LE3	T		2	teacher	other	mother	0	no	4	2	15	15	16	15
M	15	U	GT3	T		2	health	services	mother	0	no	2	4	15	16	15	15
M	16	U	GT3	T		3	health	services	mother	0	no	4	4	19	19	20	19
F	15	U	LE3	A		3	other	other	mother	0	yes	5	8	8	8	6	7
M	15	U	LE3	A		2	services	other	mother	0	yes	4	0	16	18	19	18
M	16	U	LE3	A		4	services	other	mother	0	yes	5	4	11	11	11	11
F	16	U	LE3	A		3	other	services	mother	0	yes	2	12	11	12	11	11
F	17	U	GT3	A		4	other	teacher	mother	0	yes	4	6	6	5	6	6
M	16	R	GT3	A		4	other	teacher	mother	0	yes	2	7	15	16	15	15
M	15	U	GT3	A		2	other	other	other	0	yes	4	0	14	16	16	15
M	15	U	GT3	T		4	health	health	father	0	yes	5	0	12	15	15	14
F	17	U	GT3	T		1	at_home	other	father	0	yes	5	4	5	5	6	5
F	16	U	GT3	T		3	other	other	father	0	yes	4	4	6	10	10	9
F	15	U	GT3	T		3	other	other	father	0	yes	3	0	8	7	6	7
F	15	U	GT3	T		1	services	other	father	0	yes	5	4	10	12	12	11
M	15	U	LE3	T		4	health	services	father	0	yes	4	2	14	14	14	14
M	15	U	GT3	T		2	services	services	father	0	yes	5	0	8	8	11	9
M	15	U	GT3	T		4	services	teacher	father	0	yes	4	2	19	18	18	18
F	15	U	GT3	T		4	services	teacher	father	1	yes	4	2	7	7	7	7
M	15	R	GT3	T		3	teacher	at_home	mother	0	yes	4	0	17	16	16	16
F	15	R	GT3	T		4	services	health	mother	0	yes	4	2	10	9	8	9
F	15	R	GT3	T		4	services	health	mother	0	yes	4	2	12	12	11	12
F	15	U	GT3	T		4	teacher	health	mother	0	yes	3	0	10	8	9	9
F	15	R	GT3	T		2	at_home	other	mother	0	yes	4	8	14	13	13	13
F	16	U	GT3	T		4	health	other	mother	0	yes	4	4	14	14	14	14
M	15	U	GT3	T		4	other	other	mother	0	yes	5	0	14	15	15	15
F	16	U	GT3	T		3	other	other	mother	0	yes	5	4	8	10	10	9
M	16	U	LE3	T		2	other	other	mother	0	yes	5	0	13	13	12	13
M	15	U	GT3	T		2	other	other	mother	0	yes	4	2	12	12	11	12
M	16	U	GT3	T		2	other	other	mother	0	yes	5	0	12	14	15	14
F	16	U	LE3	T		2	other	other	mother	1	yes	3	25	7	10	11	9
M	16	U	LE3	T		3	services	other	mother	0	yes	5	10	15	15	15	15
M	15	U	GT3	T		3	teacher	other	mother	0	yes	5	2	10	10	11	10
M	15	U	GT3	T		2	teacher	other	mother	0	yes	4	2	15	15	14	15
F	15	U	GT3	T		2	health	services	mother	0	yes	3	2	15	14	15	15
M	15	U	GT3	T		4	health	services	mother	0	yes	5	0	9	11	12	11
F	16	U	GT3	T		4	services	services	mother	0	yes	3	6	13	14	14	14
M	17	U	GT3	T		2	services	services	mother	3	yes	5	16	6	5	5	5
F	16	U	GT3	T		2	services	services	mother	2	yes	1	14	6	9	8	8
M	15	U	GT3	T		4	services	services	mother	0	yes	4	0	17	16	17	17
M	15	U	LE3	T		3	teacher	services	mother	0	yes	5	2	15	16	18	16
M	16	U	GT3	T		4	teacher	teacher	mother	0	yes	4	16	10	12	11	11
M	15	U	LE3	T		4	teacher	other	other	0	yes	5	8	12	12	12	12

Based on the dataset, we can get:

1. The guardian ratio between father, mother and others:

guardian	Number	occupied
mother	37	74%
father	11	22%
other	2	4%
Total	50	

⟹

Guardian Ratio

- mother
- father
- other

2. The average grade of students whose guardian is a mother or father and then get the average:

guardian	Mjob	Ave G
mother	at_home	9
	health	15
	other	11
	services	12
	teacher	13

guardian	Fjob	Ave G
father	at_home	10
	health	14
	other	8
	services	12
	teacher	13

⟹

P_Job	Ave G
at_home	10
health	15
other	10
services	12
teacher	13

Then we can transform the tables into graphs:

Relationship between Mother's Job and Children's Grade (Mother as Guardian)

	at_home	health	other	services	teacher
Ave G					

Relationship between Father's Job and Children's Grade (Father as Guardian)

	at_home	health	other	services	teacher
Ave G					

Relationship between Parent's Job and Children's Grade

	at_home	health	other	services	teacher
Ave G					

$\Rightarrow$
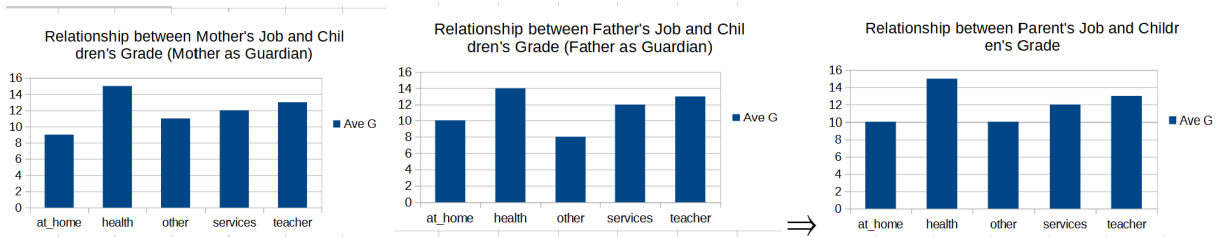
We can find that the guardian did less influence on students' grades.

3. We can also know whether a student has or has not their family support and also support impacts their grades:

famsup	Number	occupied	Ave G
No	10	20%	12
Yes	40	80%	12
Total	50		

$\Rightarrow$

Family Support Ratio

- No
- Yes

Student Grade with Family Support

- ■ No
- —— Mean (No)
- ◆ Yes
- —— Mean (Yes)
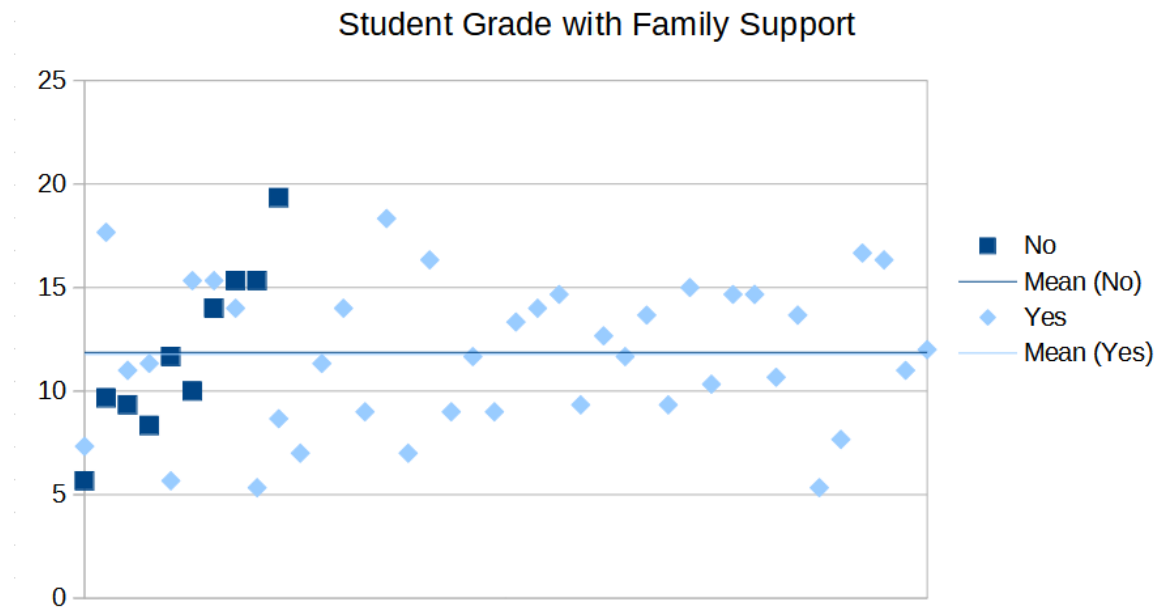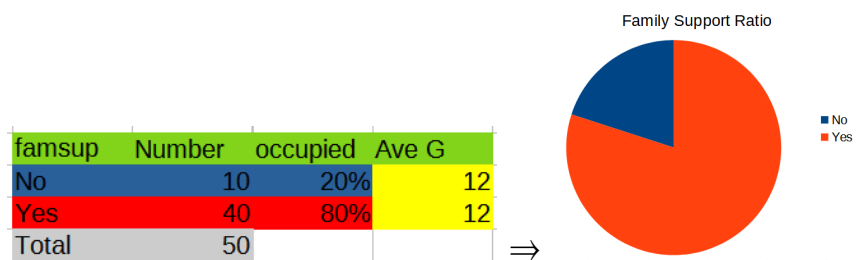
And we can find that the mean of yes and no are the same.

=> Family support did less influence on students' grades.

4. We can get the relationship between students' absences and their grades:

Relationship between absences and Grade
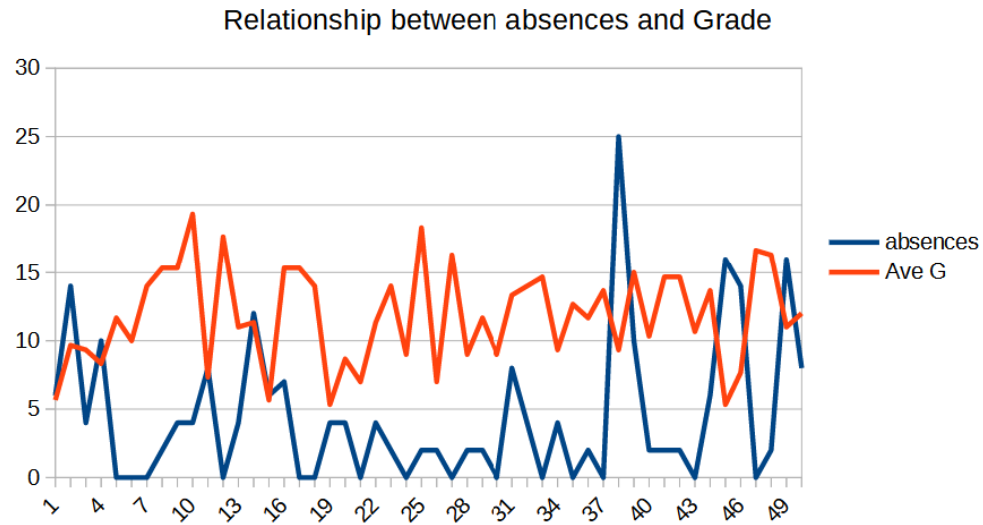
Usually high average grade means fewer absences.

5. The closely Relationship between G1, G2 and G3

Relationship between G1,G2 and G3
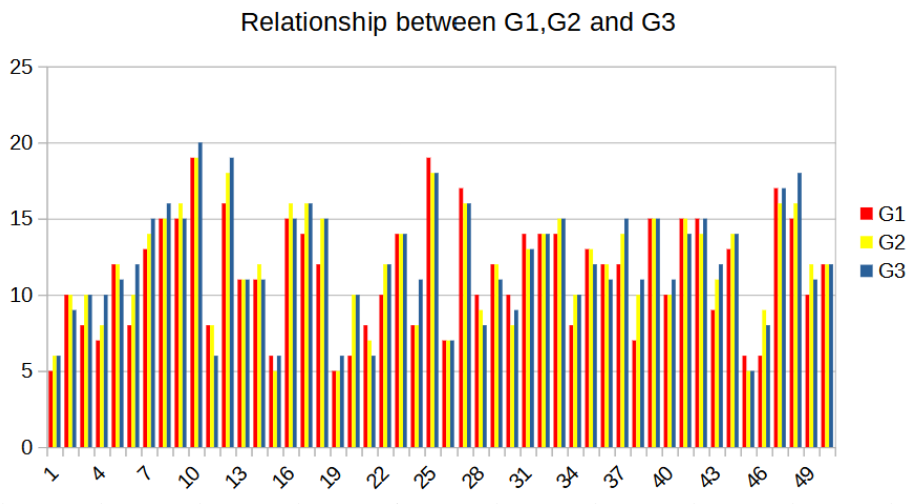
The G3 is high related to G1 and G2.

6. The relationship between the relationship of students' parents and students' Grade

Pstatus	Ave G
A	11.17
T	11.91

For the data set of students' spend time on:

| sex	age	address	traveltime	studytime	failures	romantic	freetime	goout	Dalc	Walc	absences	G1	G2	G3
F	18	U	2	2	0	no	3	4	1	1	6	5	6	6
F	17	U	1	2	0	no	3	3	1	1	4	5	5	6
F	15	U	1	2	3	no	3	2	2	3	10	7	8	10
F	15	U	1	3	0	yes	2	2	1	1	2	15	14	15
F	16	U	1	2	0	no	3	2	1	2	4	6	10	10
M	16	U	1	2	0	no	4	2	1	2	10	15	15	15
M	16	U	1	2	0	no	4	4	1	1	0	12	12	11
F	17	U	2	2	0	no	1	4	1	1	6	6	5	6
M	15	U	1	2	0	no	2	2	1	1	0	16	18	19
M	15	U	1	2	0	no	5	1	1	1	0	14	15	15
F	15	U	1	2	0	no	3	3	1	2	0	10	8	9
F	15	U	3	3	0	no	2	2	1	1	4	10	12	12
M	15	U	1	1	0	no	3	3	1	3	2	14	14	14
M	15	U	2	2	0	no	4	3	1	2	2	10	10	11
M	15	U	1	3	0	yes	5	2	1	1	0	14	16	16
F	16	U	1	1	0	no	4	4	1	2	4	14	14	14
F	16	U	1	3	0	no	2	3	1	2	6	13	14	14
F	16	U	3	2	0	no	3	2	1	1	4	8	10	10
M	17	U	1	1	3	no	5	5	2	4	16	6	5	5
M	16	U	1	1	0	no	1	3	1	3	4	8	10	10
M	15	U	1	2	0	no	4	1	1	1	0	13	14	15
M	15	U	1	1	0	no	4	2	1	1	0	12	15	15
M	16	U	1	2	0	no	5	1	1	3	2	15	15	16
M	16	U	2	2	0	no	4	4	2	4	0	13	13	12
F	15	R	1	3	0	no	3	2	1	1	2	10	9	8
F	16	U	1	1	2	no	2	2	1	3	14	6	9	8
M	15	U	1	1	0	no	2	2	1	2	2	12	12	11
M	15	U	1	1	0	no	2	4	2	4	4	15	16	15
M	16	U	1	2	0	no	3	3	1	1	4	11	11	11
M	16	U	1	2	0	yes	4	5	5	5	16	10	12	11
M	15	U	1	2	0	no	4	2	3	4	0	9	11	12
M	15	U	2	2	0	no	3	1	1	1	0	17	16	17
M	15	R	1	2	0	yes	5	2	1	1	0	17	16	16
M	15	U	1	2	0	no	3	2	1	1	0	8	10	12
M	16	U	1	1	0	no	4	3	1	1	0	12	14	15
F	15	U	2	1	0	no	5	1	1	1	0	8	7	6
M	15	U	1	3	0	no	4	3	1	1	2	15	16	18
M	16	R	2	3	0	yes	4	3	1	1	7	15	16	15
F	15	R	1	3	0	no	3	2	1	1	2	12	12	11
F	15	R	1	1	0	no	3	1	1	1	8	14	13	13
F	16	U	2	2	1	yes	3	3	1	2	25	7	10	11
M	15	U	1	1	0	yes	4	3	2	4	8	12	12	12
M	15	U	1	2	0	no	3	3	1	2	19	18	18	18
M	15	U	1	1	0	no	4	1	1	1	0	8	8	11
F	16	U	2	2	1	no	3	3	2	2	14	10	10	9
F	15	U	1	2	0	yes	2	2	1	1	8	8	8	6
F	16	U	1	2	0	no	3	5	1	4	12	11	12	11
M	16	U	1	4	0	no	2	2	1	1	4	19	19	20
M	15	U	1	2	0	no	3	3	2	2	2	15	15	14
F	15	U	1	2	1	no	4	4	1	1	2	7	7	7

Based on the dataset, we can get:

1. Relationship between students who spent more time on studying and their grades, and relationship between students who spend more time on travel and their grades:

Relationship between studying time and Grade

Relationship between travel time and Grade
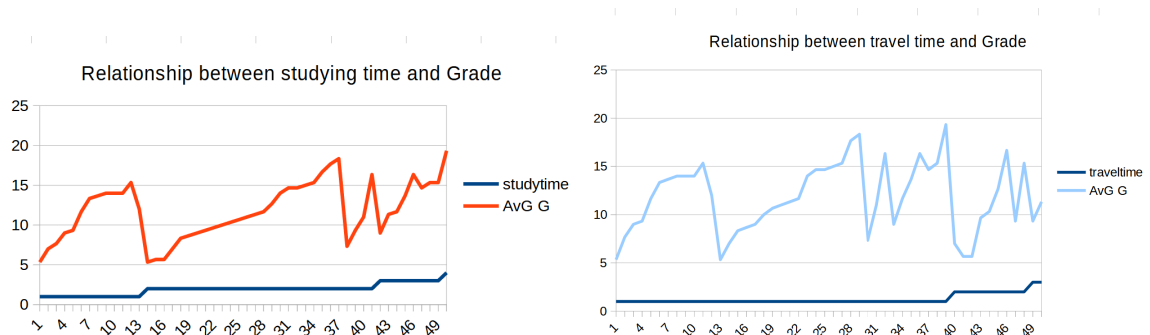
As the time spent on studying increases, the student's grade is also increasing, also decreasing as travelling time increases overall.

2. The ratio of students who has a romantic relationship or not:

   romantic: (nominal, binary) student in a romantic relationship, 'yes' or 'no'

romantic	number	occupied	Ave G
yes	8	16%	13
no	42	84%	12
Total	50		

3. The ratio of free time that students have:

   freetime: (interval) student's free time after school, ranked from 1 - very low up to 5 - very high

freetime	number	occupied	Ave G
1	2	4%	8
2	9	18%	13
3	18	36%	11
4	15	30%	12
5	6	12%	12
Total	50	100%	

=>

Ratio of free time that students have
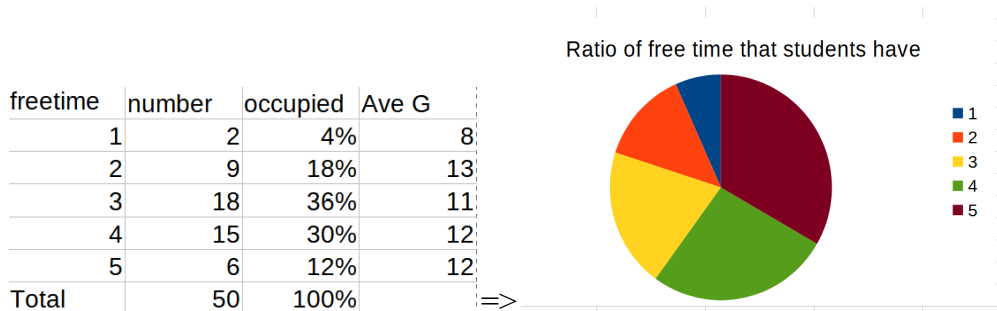
Based on this table, we can easily get the ratio of free time that students have and also know that less free time does not mean high grade.

4. The relationship between Dalc, Walc and students' Grade:

   Dalc - (interval) same ranking system as freetime, but for how much alcohol the student consumes on workdays

   Walc - (interval) same ranking system as freetime, but for how much alcohol the student consumes on weekends

   DW - Dalc and Walc: if both Dalc and Walc is 1, then DW is 11.

DW	number	Ave G
11	28	12
12	8	11
13	4	12
14	1	11
22	2	12
23	1	8
24	4	11
34	1	11
55	1	11

Then we can translate the data into graphics:

Relationship between Dalc, Walc and students' Grade
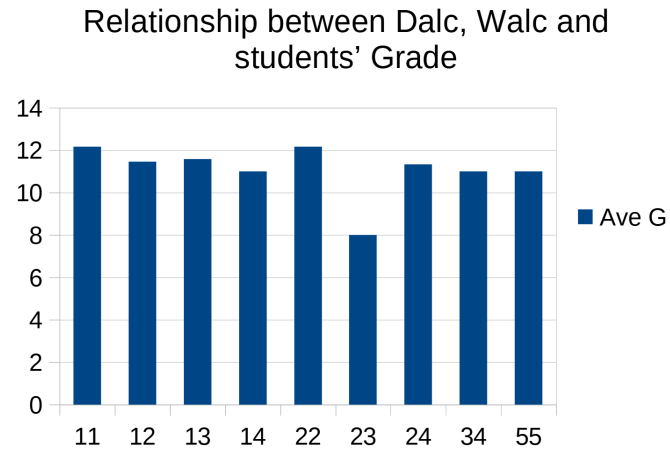
Based on the graph, the students who consume alcohol on weekends will get more grades.

5. The distribution of both males and females spent studying

Distribution of Male and Female spent time on studying
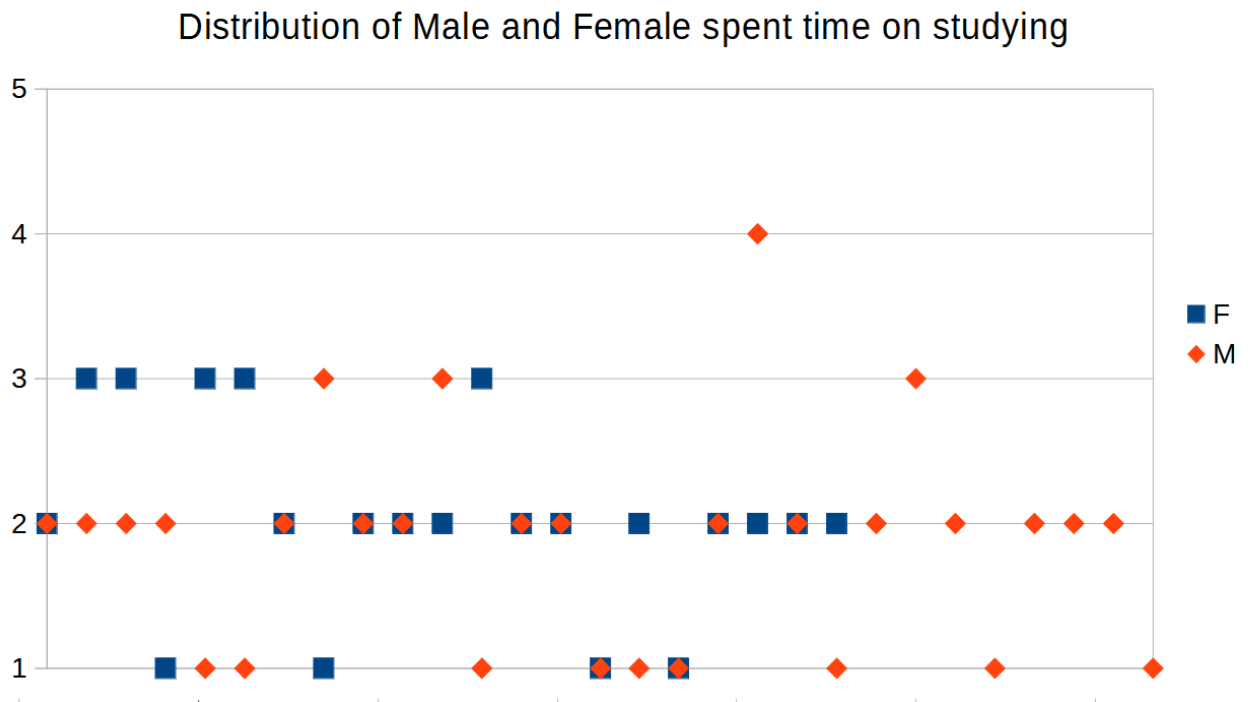
For the data set of support that students received:

| sex	age	address	failures	schoolsup	famsup	paid	activities	internet	absences	G1	G2	G3
F	18	U	0	yes	no	no	no	no	6	5	6	6
F	17	U	0	no	yes	no	no	yes	4	5	5	6
F	15	U	3	yes	no	yes	no	yes	10	7	8	10
F	15	U	0	no	yes	yes	yes	yes	2	15	14	15
F	16	U	0	no	yes	yes	no	no	4	6	10	10
M	16	U	0	no	yes	yes	yes	yes	10	15	15	15
M	16	U	0	no	no	no	no	yes	0	12	12	11
F	17	U	0	yes	yes	no	no	no	6	6	5	6
M	15	U	0	no	yes	yes	no	yes	0	16	18	19
M	15	U	0	no	yes	yes	yes	yes	0	14	15	15
F	15	U	0	no	yes	yes	no	yes	0	10	8	9
F	15	U	0	no	yes	no	yes	yes	4	10	12	12
M	15	U	0	no	yes	yes	yes	yes	2	14	14	14
M	15	U	0	no	yes	yes	no	yes	2	10	10	11
M	15	U	0	no	yes	no	no	yes	0	14	16	16
F	16	U	0	no	yes	no	no	yes	4	14	14	14
F	16	U	0	no	yes	yes	yes	yes	6	13	14	14
F	16	U	0	yes	yes	no	yes	no	4	8	10	10
M	17	U	3	no	yes	no	yes	yes	16	6	5	5
M	16	U	0	no	no	yes	yes	yes	4	8	10	10
M	15	U	0	no	no	no	no	yes	0	13	14	15
M	15	U	0	no	yes	yes	no	yes	0	12	15	15
M	16	U	0	no	no	no	yes	yes	2	15	15	16
M	16	U	0	no	yes	no	yes	yes	0	13	13	12
F	15	R	0	yes	yes	yes	yes	yes	2	10	9	8
F	16	U	2	no	yes	yes	no	yes	14	6	9	8
M	15	U	0	no	yes	yes	no	yes	2	12	12	11
M	15	U	0	no	no	yes	no	yes	4	15	16	15
M	16	U	0	yes	yes	no	yes	yes	4	11	11	11
M	16	U	0	no	yes	yes	yes	yes	16	10	12	11
M	15	U	0	no	yes	yes	no	yes	0	9	11	12
M	15	U	0	no	yes	no	yes	yes	0	17	16	17
M	15	R	0	no	yes	no	yes	yes	0	17	16	16
M	15	U	0	no	no	no	yes	yes	0	8	10	12
M	16	U	0	no	yes	yes	no	yes	0	12	14	15
F	15	U	0	no	yes	no	yes	no	0	8	7	6
M	15	U	0	no	yes	no	yes	yes	2	15	16	18
M	16	R	0	no	yes	no	yes	yes	7	15	16	15
F	15	R	0	yes	yes	yes	yes	yes	2	12	12	11
F	15	R	0	yes	yes	yes	yes	no	8	14	13	13
F	16	U	1	no	yes	no	yes	yes	25	7	10	11
M	15	U	0	no	yes	no	no	yes	8	12	12	12
M	15	U	0	no	yes	no	yes	yes	2	19	18	18
M	15	U	0	yes	yes	no	no	yes	0	8	8	11
F	16	U	1	yes	no	no	yes	yes	14	10	10	9
F	15	U	0	yes	yes	yes	yes	yes	8	8	8	6
F	16	U	0	no	yes	no	no	yes	12	11	12	11
M	16	U	0	no	no	no	yes	yes	4	19	19	20
M	15	U	0	no	yes	yes	no	no	2	15	15	14
F	15	U	1	yes	yes	no	yes	yes	2	7	7	7

Based on the data set, we can get:

1. The relationship between Internet support and Grades:

Relationship between Internet support and Grade s
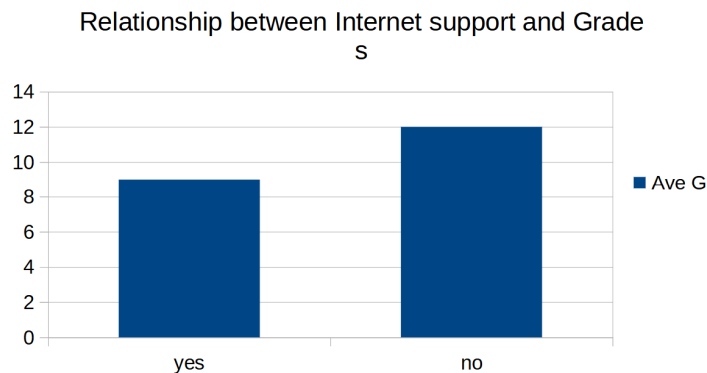
By using this graph, we can quickly get how the internet influences students' grades.

2. The relationship between family support or school support or both or none and students' grades:

Relationship between supports and Grades
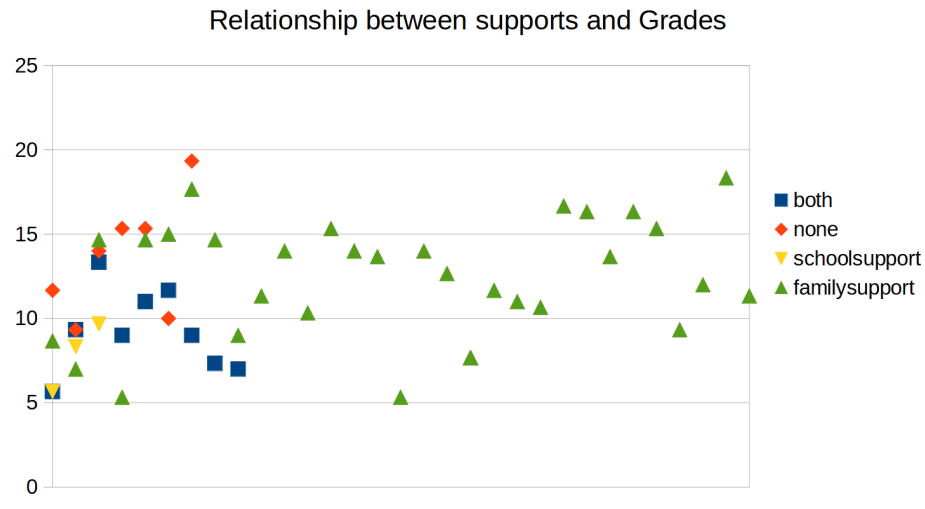
Using this technology, we can quickly figure out how different support influences students' grades.

3. The ratio of paid courses and unpaid courses and their influences on students' grades:

paid	number	occupied	Ave G
yes	23	46%	12.71
no	27	54%	11.01
		difference	8.49%

The ratio of paid and unpaid
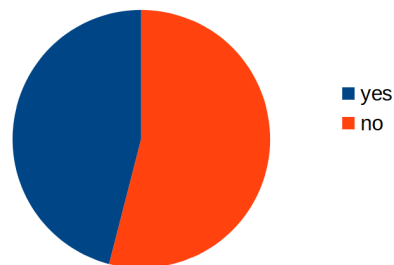
We can find that the students who paid for extra courses will have 2.83% grades higher than the students who are not paid for extra courses.

Also by using this chart and table, we can quickly know whats the occupied of paid courses on total students courses. And we also know the impact that having paid courses on students' grades.

4. The relationship between having or not having activities and their grades:

The relationship between activities and grades
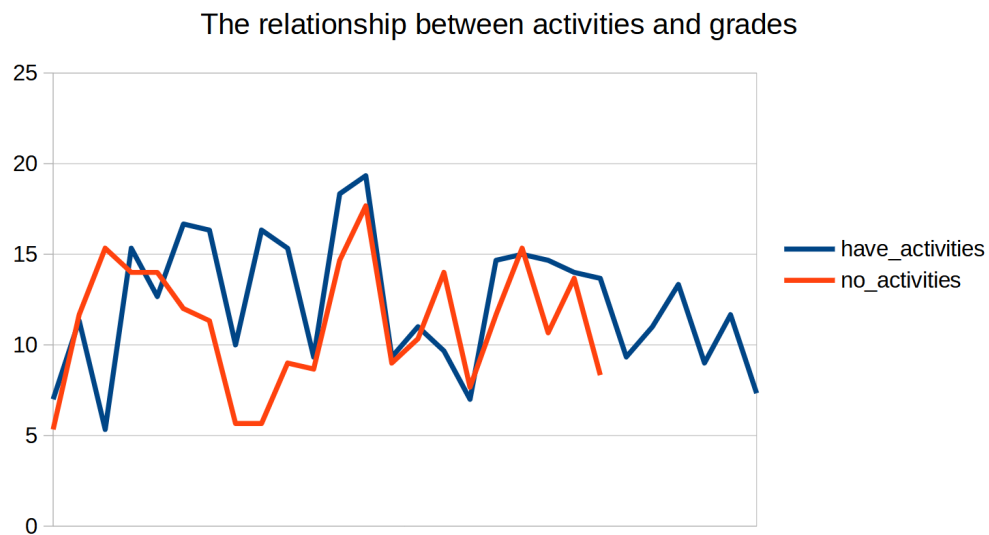
By using this graph, we can quickly get what's the difference between having activities and not having activities that impact students' grades.

Task 2: Basic Statistical Analysis

In 216771875-216328387-215122856-T2.csv

The selected dimensions are bold below.

Nominal: sex, address, Pstatus, **Mjob**, Fjob, guardian, famsup

Ordinal: **Medu**, Fedu, famrel

Interval:**famsize**

Ratio: age, failures, absences, G1, G2, **G3**

For preprocessing, by using simple random sampling (get 50 objects from the original dataset) will help us save more time.

Mjob	Number	Occupied
at_home	4	8%
health	8	16%
other	16	32%
services	13	26%
teacher	9	18%
Total	50	
Mean	can not determine	
Mode	other	

For the mother's job, other have more numbers, so the mode of Mjob is other.

But the mean can not be determined, because the attribute of a mother's job is nominal.

Medu	Number
0	0
1	2
2	12
3	10
4	26
Total	50
Mean	3.2
Mode	4

Based on the table we can quickly get the mode is 4, which is post-secondary education

also the mean of the mother's education level is 3.2 (≈ 3), which is secondary education.

famsize	Number
GT3	34
LE3	16
Total	50
Mean	can not determine
Mode	GT3

By counting the dataset, we can get there are 34 students whose family members are greater than 3, and 16 students whose family members are less than or equal to 3, so the mode is GT3, but we can not get the mean since this interval can not be calculated in numbers.

G3	
Total Mark	604
Number	50
Mean	12.08
Mode	11

For G3, we can get the mean of 12.08 (≈12), and after sorting the G3 of all students, we can quickly get the mode of G3 is 11.

Task 3: Standardization and Normalization

In 216771875-216328387-215122856-T3.csv

The selected dimensions are:

Interval:**freetime**

Ratio: **G3**

Z-score standardization formula: z-score = x - mean / stdev; (x: the element to be compared; mean: mathematical mean of the attribute; stdev: standard deviation of attribute)

Min-Max normalization formula: xscale = x-xmin/(xmax-xmin); (x: element to be compared; xmin: minimum value in attribute; xmax: maximum value of an attribute)

For the attribute freetime, after performing Z-score standardization and Min-Max normalization as seen in the graph, although from the original data we can analyze the attribute based on the occurrence of the value, after the processes we can analyze how many elements are close/far from the mathematical mean.

freetime(original)

freetime(Z-Score standardization)

freetime(Min-Max normalization)
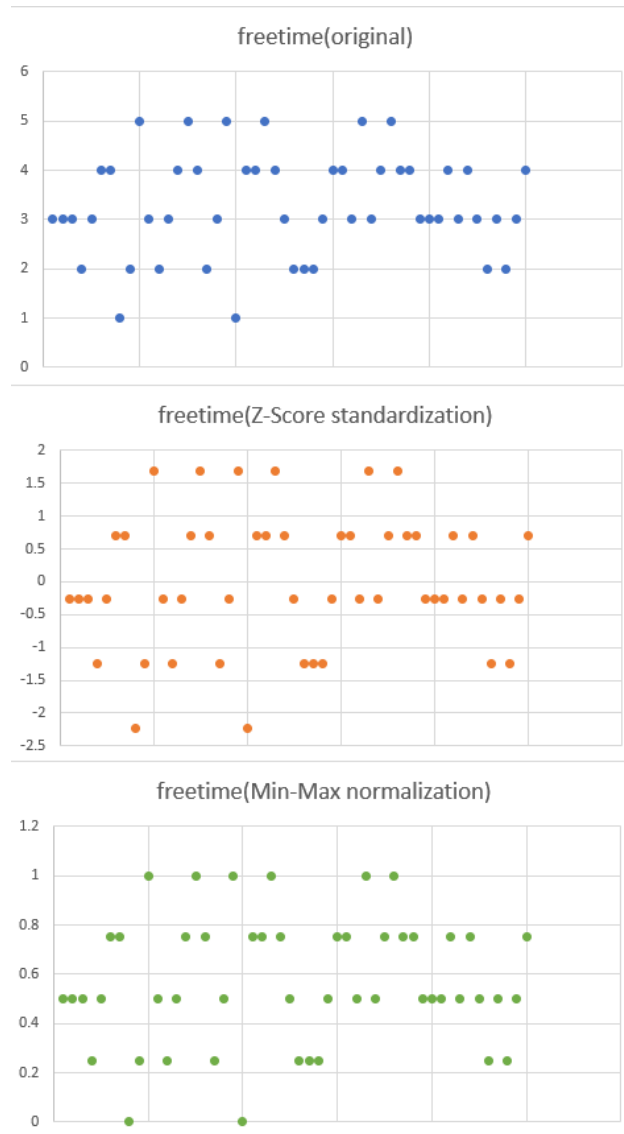
Similar to freetime, the G3 attribute also benefits on the two processes. By standardization and normalization, every datapoint has the same scale thus each feature is equally important.

G3(Original)

G3(Z-Score standrdization)

G3(Min-Max normalization)
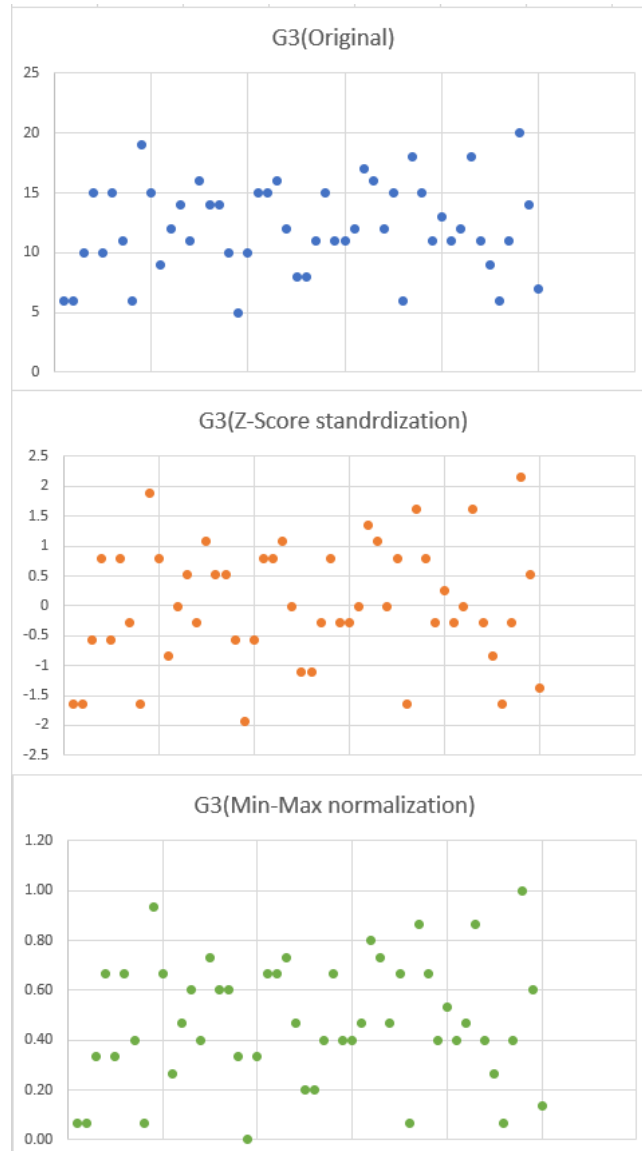
Despite having a similar scatter pattern, the two processes should result in better scatter (more distinct than the original) if the range of values of the attributes were larger. On another hand, as both attributes do not have outlier due to the small range, performing Min-Max normalization and Z-score standardization results in a clear chart

Task 4: Principal Component Analysis

For task 4, the dataset regarding family-grade relations has the most, i.e. 14 attributes, said dataset is chosen. In order to ease the calculation process, we have dropped several insignificant attributes and resulted in the below spreadsheet, focusing on mother/father's education level, family's relationship, failures, and the three grades.

Medu	Fedu	failures	famrel	G1	G2	G3
4	4	0	4	5	6	6
1	1	0	5	5	5	6
1	1	3	4	7	8	10
4	2	0	3	15	14	15
3	3	0	4	6	10	10
4	3	0	5	15	15	15
2	2	0	4	12	12	11
4	4	0	4	6	5	6
3	2	0	4	16	18	19
3	4	0	5	14	15	15
4	4	0	3	10	8	9
2	1	0	5	10	12	12
4	4	0	4	14	14	14
4	3	0	5	10	10	11
2	2	0	4	14	16	16
4	4	0	4	14	14	14
4	4	0	3	13	14	14
3	3	0	5	8	10	10
3	2	3	5	6	5	5
4	3	0	3	8	10	10
4	3	0	4	13	14	15
4	4	0	5	12	15	15
4	2	0	4	15	15	16
2	2	0	5	13	13	12
2	4	0	4	10	9	8
2	2	2	1	6	9	8
2	2	0	4	12	12	11
4	2	0	2	15	16	15
3	4	0	5	11	11	11
4	4	0	4	10	12	11
4	4	0	5	9	11	12
4	4	0	4	17	16	17
4	3	0	4	17	16	16
3	3	0	5	8	10	12
3	2	0	5	12	14	15
2	3	0	3	8	7	6
4	3	0	5	15	16	18
4	4	0	2	15	16	15
3	4	0	4	12	12	11
2	2	0	4	14	13	13
2	2	1	3	7	10	11
4	4	0	5	12	12	12
4	4	0	4	19	18	18
2	2	0	5	8	8	11
2	2	1	4	10	10	9
4	3	0	5	8	8	6
3	3	0	2	11	12	11
4	3	0	4	19	19	20
4	2	0	4	15	15	14
4	4	1	4	7	7	7

To perform PCA, standardization is needed to compare every component in the same scale.

Using the Z-score standardization result in the dataset below:

After Standarization:

Medu	Fedu	failures	famrel	G1	G2	G3
0.85	1.13	-0.33	-0.04	-1.73	-1.67	-1.65
-2.35	-2.01	-0.33	1.02	-1.73	-1.95	-1.65393
-2.35	-2.01	4.14	-0.04	-1.18	-1.11	-0.56582
0.85	-0.96	-0.33	-1.11	0.99	0.58	0.794323
-0.21	0.08	-0.33	-0.04	-1.46	-0.55	-0.56582
0.85	0.08	-0.33	1.02	0.99	0.86	0.794323
-1.28	-0.96	-0.33	-0.04	0.17	0.02	-0.29379
0.85	1.13	-0.33	-0.04	-1.46	-1.95	-1.65393
-0.21	-0.96	-0.33	-0.04	1.26	1.70	1.882437
-0.21	1.13	-0.33	1.02	0.72	0.86	0.794323
0.85	1.13	-0.33	-1.11	-0.37	-1.11	-0.83785
-1.28	-2.01	-0.33	1.02	-0.37	0.02	-0.02176
0.85	1.13	-0.33	-0.04	0.72	0.58	0.522295
0.85	0.08	-0.33	1.02	-0.37	-0.55	-0.29379
-1.28	-0.96	-0.33	-0.04	0.72	1.14	1.066352
0.85	1.13	-0.33	-0.04	0.72	0.58	0.522295
0.85	1.13	-0.33	-1.11	0.45	0.58	0.522295
-0.21	0.08	-0.33	1.02	-0.91	-0.55	-0.56582
-0.21	-0.96	4.14	1.02	-1.46	-1.95	-1.92596
0.85	0.08	-0.33	-1.11	-0.91	-0.55	-0.56582
0.85	0.08	-0.33	-0.04	0.45	0.58	0.794323
0.85	1.13	-0.33	1.02	0.17	0.86	0.794323
0.85	-0.96	-0.33	-0.04	0.99	0.86	1.066352
-1.28	-0.96	-0.33	1.02	0.45	0.30	-0.02176
-1.28	1.13	-0.33	-0.04	-0.37	-0.83	-1.10988
-1.28	-0.96	2.65	-3.24	-1.46	-0.83	-1.10988
-1.28	-0.96	-0.33	-0.04	0.17	0.02	-0.29379
0.85	-0.96	-0.33	-2.18	0.99	1.14	0.794323
-0.21	1.13	-0.33	1.02	-0.10	-0.26	-0.29379
0.85	1.13	-0.33	-0.04	-0.37	0.02	-0.29379
0.85	1.13	-0.33	1.02	-0.64	-0.26	-0.02176
0.85	1.13	-0.33	-0.04	1.53	1.14	1.33838
0.85	0.08	-0.33	-0.04	1.53	1.14	1.066352
-0.21	0.08	-0.33	1.02	-0.91	-0.55	-0.02176
-0.21	-0.96	-0.33	1.02	0.17	0.58	0.794323
-1.28	0.08	-0.33	-1.11	-0.91	-1.39	-1.65393
0.85	0.08	-0.33	1.02	0.99	1.14	1.610409
0.85	1.13	-0.33	-2.18	0.99	1.14	0.794323
-0.21	1.13	-0.33	-0.04	0.17	0.02	-0.29379
-1.28	-0.96	-0.33	-0.04	0.72	0.30	0.250266
-1.28	-0.96	1.16	-1.11	-1.18	-0.55	-0.29379
0.85	1.13	-0.33	1.02	0.17	0.02	-0.02176
0.85	1.13	-0.33	-0.04	2.08	1.70	1.610409
-1.28	-0.96	-0.33	1.02	-0.91	-1.11	-0.29379
-1.28	-0.96	1.16	-0.04	-0.37	-0.55	-0.83785
0.85	0.08	-0.33	1.02	-0.91	-1.11	-1.65393
-0.21	0.08	-0.33	-2.18	-0.10	0.02	-0.29379
0.85	0.08	-0.33	-0.04	2.08	1.98	2.154466
0.85	-0.96	-0.33	-0.04	0.99	0.86	0.522295
0.85	1.13	1.16	-0.04	-1.18	-1.39	-1.3819

After standardization, we can compute the covariance matrix for the data set by using

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Resulting the covariance matrix below:

Pearson's r	Medu	Fedu	failures	famrel	G1	G3	G2
Medu	-	0.642	-0.355	-0.032	0.385	0.349	0.357
Fedu	0.642	-	-0.346	0.004	0.122	0.036	0.075
failures	-0.355	-0.346	-	-0.141	-0.404	-0.363	-0.396
famrel	-0.032	0.004	-0.141	-	-0.045	0.016	-0.071
G1	0.385	0.122	-0.404	-0.045	-	0.896	0.936
G3	0.349	0.036	-0.363	0.016	0.896	-	0.962
G2	0.357	0.075	-0.396	-0.071	0.936	0.962	-

After computing covariance, we then can calculate the Eigenvalue

$$\det (A - \lambda I) = 0$$

by solving lamda, then choose the largest two lambda as principal components.

Using Microsoft Excel add-ins, analyse-it, the result table shows that the first component has a strong relation toward the three grades and average relation toward Medu,Fedu, and failures. The second component shows strong relation toward Medu and Fedu. Hence these two components are chosen and highlighted.

				Component			
	1	2	3	4	5	6	7
Medu	-0.328	0.503	-0.172	-0.350	0.697	-0.036	-0.033
Fedu	-0.185	0.690	-0.121	-0.125	-0.672	0.084	0.024
failures	0.321	-0.267	-0.302	-0.842	-0.155	-0.022	-0.025
famrel	0.006	0.092	0.928	-0.354	-0.017	-0.039	-0.061
G1	-0.503	-0.214	-0.024	-0.073	-0.154	-0.786	0.231
G2	-0.507	-0.262	-0.038	-0.057	-0.115	0.194	-0.787
G3	-0.495	-0.277	0.039	-0.139	-0.037	0.578	0.567

The below biplot indicates the relations between different attributes. In terms of the length of the vector, we can discover the following G1, G2, and G3. Medu and Fedu have the second strongest representation, and famrel has the least representation. Apart from that, it also indicates that the G1, G2 and G3 have a positive correlation, indicating that if a student's grade is already high in the first exam, he/she will tend to score high in the following two exams. On the other hand, it shows that Medu has a stronger positive correlation with the grades than Fedu, showing that a mother's educational level tends to contribute more to the student's academic achievement in this dataset. Furthermore, it shows that the failures attribute has a negative correlation with the grades, which is common knowledge, indicating that with more failures the grades of the students will be lower.
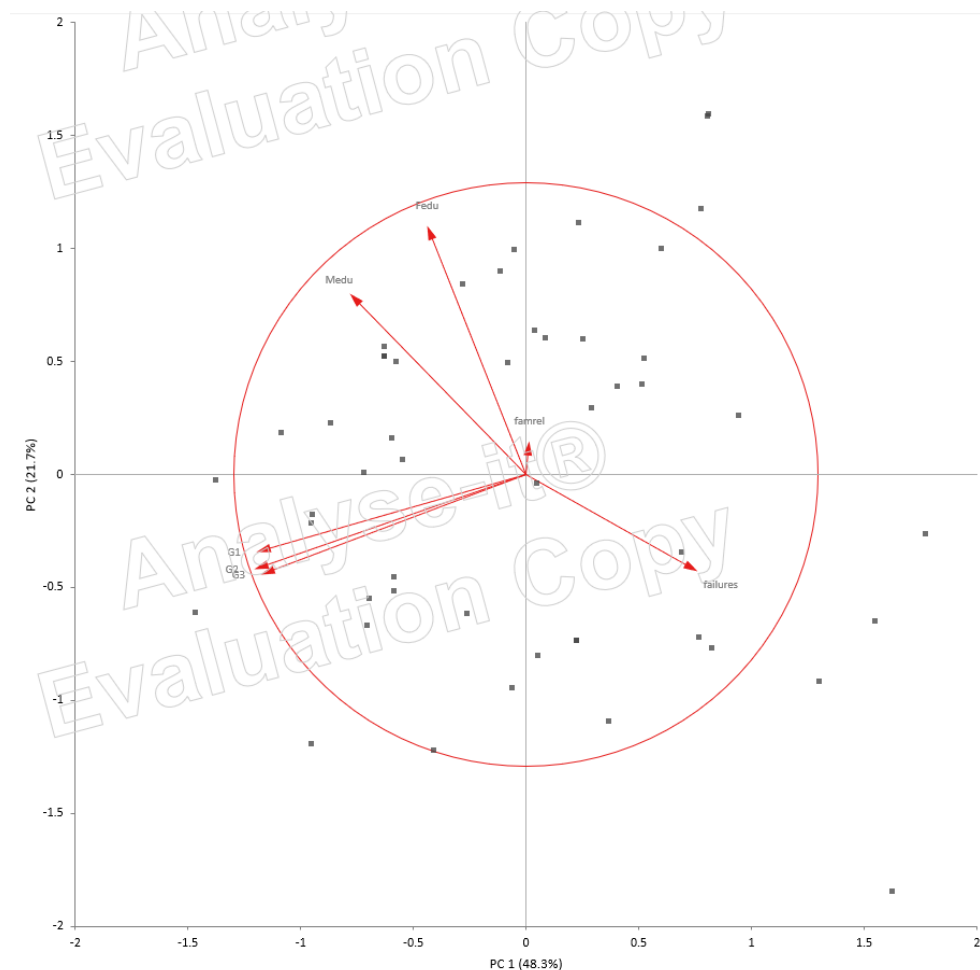
Task 5: Similarity Measurement

- The object type to be used for this task will be from timer.csv, or the table from task one which was focused on what students did in their free time. To fit the requirements of the task, the 5 numeric dimensions to be picked are the following

    - Age (ratio), which describes the student's age

    - Studytime (ordinal), which describes numerically the amount of study time a student has

    - Freetime (interval), describes numerically the amount of free time a student has

    - Absences (ratio), describes the amount of time a student has been absent

    - G3 (ratio), the numeric value of the student's final grade

- As one of the 5 questions within task one was how the ratio of a student's study time and free time affect their grades (G3), the focus will be on these three attributes

- The filtered data can be found in 216771875-216328387-215122856-T5Data.csv

The following CSV files for euclidean and cosine distance matrices were formed using the file EECS4412_Task5.ipynb (using Jupyter Notebook, the sklearn imported functions).

- 216771875-216328387-215122856-T5EU.csv

- 216771875-216328387-215122856-T5CO.csv

- 216771875-216328387-215122856-T5MA.csv

Notes:

- when checking the jupyter notebook file the 216771875-216328387-215122856-T5Data.csv file must be in the same folder/directory as the notebook file, keep all files from TASK_5 folder within the TASK_5 folder

- Images of the code (if necessary) and the code itself can be found within the TASK_5 folder, all imports are stated at the beginning of the code and consists of only imports available when installing Jupyter Notebook and Anaconda

Task 6: Classification

- The original classifier to be used with the data table from task 5 would have been G3, and through using regression and a decision tree classifier for testing, we would have possibly been able to predict how study time and free time affect a student's grades. However, as we are only working with 50 objects, this becomes infeasible as it would be basically impossible to create an accurate classifier with 20 different outcomes. Instead, as shown in EECS4412_Task6.ipynb we have done the following preprocessing steps:
    - As we want to answer the question of how study time and free time affect one's grades, we will be removing age and absences from the table
    - To give our classifier an easier time, we need to reduce the total number of possible classifications. Based on the value of G3, a new value is added to each object, called rating. The rating could have one of 3 values, 'Failing' if G3 is between 0 and 10, 'Passing' if G3 is between 10 and 15, and 'Strongly_Passing' if G3 is between 15 and 20. With the reduction of classes, with our total of 50 objects, the classifier should have an easier time making classification versus having the original 20 from G3
    - G3 is removed as it is basically redundant
- Thus, our final table is located in 216771875-216328387-215122856-T6Data.csv and contains the attributes studytime, freetime and rating.

- **Decision Tree Classification Analysis**

  - The split we decided to go with for the data set was a 70 - 30 split, 70 for training data and 30 for testing, the code for all the decision tree classification can be found within EECS4412_Task6.ipynb

  - The Accuracy of this decision tree in our case happens to be 0.666. This indicates to us that there isn't a correlation between how a student spends their time and their grades, thus these attributes are 1, unusable (at least by themselves) to predict a student's grade, and 2, have no to little effect on a student's grades

  - The Accuracy is affected by the data sample used for training. Take for instance the snippet of code below from EECS4412_Task6.ipynb

```
In [44]:   #Setup for Decision Tree Classifer
           #Marking data as feature or target
           feature_cols = ['studytime', 'freetime'] # feature
           x = table[feature_cols]
           y = table.rating # target

           #Spliting data into training (70%) and test (30%)
           x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=3)

In [45]:   #Create Decision Tree classifer object
           clf = DecisionTreeClassifier()
```

    - If we change the random_state to 1, we will receive an accuracy of 0.333, while a random_state of 2 would be 0.533. While increasing the amount of training data may assist in having better accuracy, the best possible classifier developed was is random_state of 3, with an accuracy of 0.666

      - On top of this, analysis from task 1 (page 7-8) has indicated that study time and free time have no guaranteed effect on one's grades as well. While there is a slight upward trend in terms of study time versus grades, there may not be enough data (mainly in the case of the number of students who spend the most time studying) to accurately train the

classifier to determine one's grades. Free time also appears to have little

correlation with grades, and other than students with the least free time,

they all had about the same average for grades. With more data on

students who spend more time, we may be able to accurately see that

upward trend in our analysis, and train our classifier better.

- The Decision Tree generated can be viewed below, or will be within the task 6

folder (decision_tree_graph.pdf), or will be generated once

EECS4412_Task6.ipynb is run:
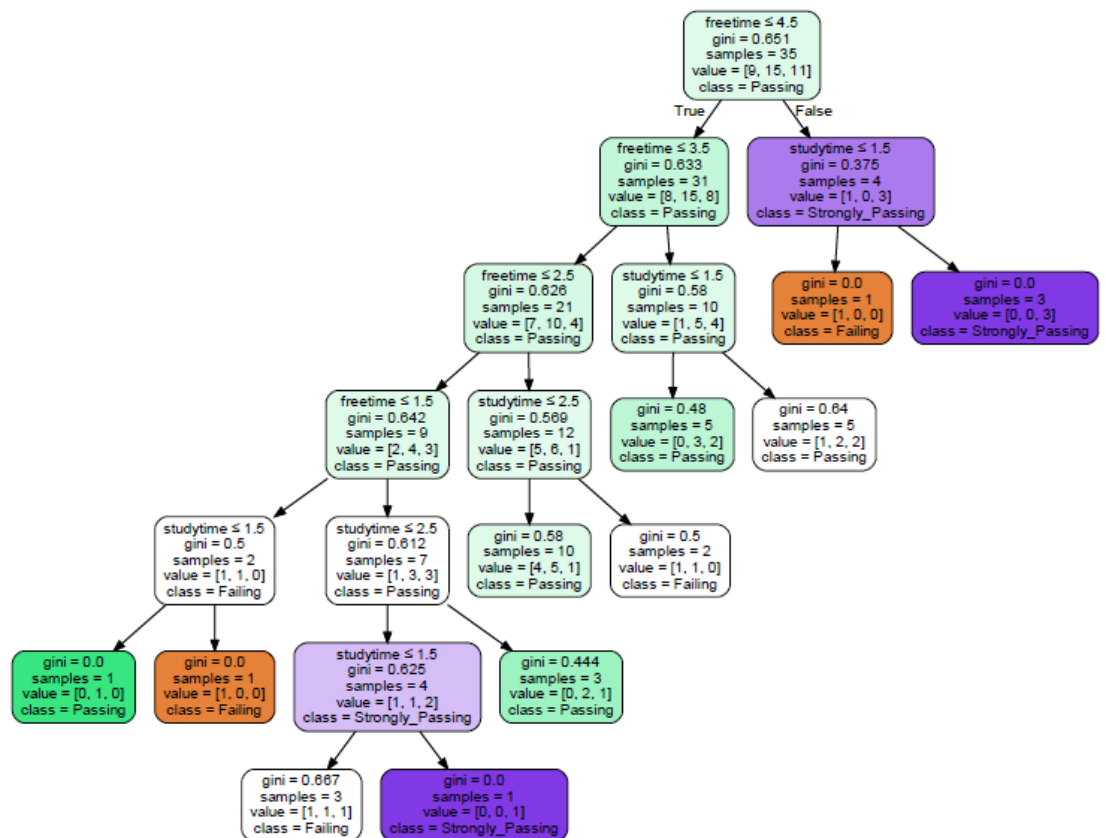
Decision Tree Table Generated by Task 6

Notes

- Images of the code and the code itself (for jupyter notebook) can be found within the TASK_6 folder all imports are from the jupyter and anaconda installation except for the ones notated with #graphviz

- Much like task 5, keep all files from the TASK_6 folder within the TASK_6 folder for the Jupyter Notebook code to work properly

   - To install graphviz, in the Anaconda Navigator type the following

      - *conda install graphviz*

      - *conda install python-graphviz*

   - More Documentation at https://graphviz.readthedocs.io/en/stable/manual.html

   - More Troubleshooting can be found from

     https://stackoverflow.com/questions/33433274/anaconda-graphviz-cant-import-after-installation

   - ALTERNATIVELY, the code snippet (from Task6_IMG04.png) can be commented out and the two import lines annotated with #graphviz can be removed, code will function but no visual graph will be created

Task 7: Progress on your Objectives

1. From table 1, based on the quality of family relationships and the parents' cohabitation status, we had raised the question of "***How does the quality of family relationships affect the students' performance and would students be taken care of by single parents who struggle on their performances?***"

From our analysis in task 1 (page 6), we can see that a good relationship between students' parents will lead to higher grades for students.

Pstatus	Ave G
A	11.17
T	11.91

Found:

A: students' parents apart from each other;

T: students' parents living together;

2. In table 1, based on the attributes of family information, we have asked the question "***Is there a positive correlation between parents' high level of education or whose job is teachers and students' performance?***" From our analysis in task 1 (page 4-5), we can see that parents with jobs we typically associate with requiring education such as health and teaching had a higher average grade in comparison to other jobs and parents who stayed at home. While this does answer the question we can possibly use this data for further analysis.

3. From table 2, we originally raised the question of "***how does the ratio between time spent on studying and time spent on not studying affect the students' performances?***" Through the analysis, we have done on task 1 (page 7-8) and using the attributes of study time and free time in the decision tree classifier in task 6, we can come to the conclusion that study time and free time are not major contributing factors to a student's grades with our current data. We can come to the conclusion due to two main reasons, one is the fact that while there is an upward trend between study time and grades, we may not have enough evidence within 50 students to say this is without a doubt accurate and there is little correlation between free time and grade. Secondly, our decision tree classifier had a max accuracy of 66.66%, pointing to the idea that these attributes are not the best to use for predicting grades. From our analysis on page 7, we have also found that lower travelling time resulted in higher

grades, so we can instead ask "***how does the ratio between time spent on studying and time spent travelling affect the students' performances?***"

4. In table 3, we can find whether extra paid classes have a positive, neutral or negative effect on the students' performances by asking the question "***Is there a correlation between extra paid classes and the students' performances?***" From our analysis in task 1, we can see there is a positive correlation between having paid for classes and grades, as paying resulted in a higher average grade compared to not doing so. This answers the question we raised and we can use this data further to aid in future attempts at classification.

5. In table 3, we have a question: "***Whether it is more beneficial for their performance if students accept support not only from school but also from family?***" Through analysis done in task one (page 11), we can make several observations

    a. Having only school support typically resulted in lower grades

    b. Having no support typically had higher grades, family support and no support are comparable in their averages

    c. Having support from both school and family typically resulted in lower grades

    This somewhat answers our question with our limited results from 50 students, which we can use to assist in a classifier.