Preliminary analysis and summary statistics

1. How many observations do we have?

The dataset we collect has 395 observations from two Portugal high school student ; Gabriel Pereira, Mousinho da Silveira. 349 observations were collected from Gabriel Pereira high school and rest of them are from Mousinho da Silveira high school. However, while we analyze this dataset we will not focus on which school students are in but other variables.

2. What information/features/characteristics do you have for each observation?

For each 395 observations, we have 33 informations ; and we categorize 30 information that can affect one's study performance and below columns are 30 informations.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
1	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
2	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no	yes	yes	no	no	4	3	4	1	1	3	6	5	6	6
3	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no	no	yes	yes	no	5	3	3	1	1	3	4	5	5	6
4	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3	yes	no	yes	no	yes	yes	yes	no	4	3	2	2	3	3	10	7	8	10
5	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes	yes	yes	yes	yes	yes	yes	3	2	2	1	1	5	2	15	14	15
6	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes	yes	no	yes	yes	no	no	4	3	2	1	2	5	4	6	10	10
7	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2	0	no	yes	yes	yes	yes	yes	yes	no	5	4	2	1	2	5	10	15	15	15
8	GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	no	no	no	no	yes	yes	yes	no	4	4	4	1	1	3	0	12	12	11
9	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0	yes	yes	no	no	yes	yes	no	no	4	1	4	1	1	1	6	6	5	6
10	GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	0	no	yes	yes	no	yes	yes	yes	no	4	2	2	1	1	1	0	16	18	19
11	GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2	0	no	yes	yes	yes	yes	yes	yes	no	5	5	1	1	1	5	0	14	15	15
12	GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother	1	2	0	no	yes	yes	no	yes	yes	yes	no	3	3	3	1	2	2	0	10	8	9
13	GP	F	15	U	GT3	T	2	1	services	other	reputation	father	3	3	0	no	yes	no	yes	yes	yes	yes	no	5	2	2	1	1	4	4	10	12	12
14	GP	M	15	U	LE3	T	4	4	health	services	course	father	1	1	0	no	yes	yes	yes	yes	yes	yes	no	4	3	3	1	3	5	2	14	14	14
15	GP	M	15	U	GT3	T	4	3	teacher	other	course	mother	2	2	0	no	yes	yes	no	yes	yes	yes	no	5	4	3	1	2	3	2	10	10	11
16	GP	M	15	U	GT3	A	2	2	other	other	home	other	1	3	0	no	yes	no	no	yes	yes	yes	yes	4	5	2	1	1	3	0	14	16	16
17	GP	F	16	U	GT3	T	4	4	health	other	home	mother	1	1	0	no	yes	no	no	yes	yes	yes	no	4	4	4	1	2	2	4	14	14	14
18	GP	F	16	U	GT3	T	4	4	services	services	reputation	mother	1	3	0	no	yes	yes	yes	yes	yes	yes	no	3	2	3	1	2	2	6	13	14	14
19	GP	F	16	U	GT3	T	3	3	other	other	reputation	mother	3	2	0	yes	yes	no	yes	yes	yes	no	no	5	3	2	1	1	4	4	8	10	10
20	GP	M	17	U	GT3	T	3	2	services	services	course	mother	1	1	3	no	yes	no	yes	yes	yes	yes	no	5	5	5	2	4	5	16	6	5	5
21	GP	M	16	U	LE3	T	4	3	health	other	home	father	1	1	0	no	no	yes	yes	yes	yes	yes	no	3	1	3	1	3	5	4	8	10	10
22	GP	M	15	U	GT3	T	4	3	teacher	other	reputation	mother	1	2	0	no	no	no	no	yes	yes	yes	no	4	4	1	1	1	1	0	13	14	15
23	GP	M	15	U	GT3	T	4	4	health	health	other	father	1	1	0	no	yes	yes	no	yes	yes	yes	no	5	4	2	1	1	5	0	12	15	15
24	GP	M	16	U	LE3	T	4	2	teacher	other	course	mother	1	2	0	no	no	no	yes	yes	yes	yes	no	4	5	1	1	3	5	2	15	15	16
25	GP	M	16	U	LE3	T	2	2	other	other	reputation	mother	2	2	0	no	yes	no	yes	yes	yes	yes	no	5	4	4	2	4	5	0	13	13	12
26	GP	F	15	R	GT3	T	2	4	services	health	course	mother	1	3	0	yes	yes	yes	yes	yes	yes	yes	no	4	3	2	1	1	5	2	10	9	8
27	GP	F	16	U	GT3	T	2	2	services	services	home	mother	1	1	2	no	yes	yes	no	no	yes	yes	no	1	2	2	1	3	5	14	6	9	8
28	GP	M	15	U	GT3	T	2	2	other	other	home	mother	1	1	0	no	yes	yes	no	yes	yes	yes	no	4	2	2	1	2	5	2	12	12	11
29	GP	M	15	U	GT3	T	4	2	health	services	other	mother	1	1	0	no	no	yes	no	yes	yes	yes	no	2	2	4	2	4	1	4	15	16	15
30	GP	M	16	U	LE3	A	3	4	services	other	home	mother	1	2	0	yes	yes	no	yes	yes	yes	yes	no	5	3	3	1	1	5	4	11	11	11
31	GP	M	16	U	GT3	T	4	4	teacher	teacher	home	mother	1	2	0	no	yes	yes	yes	yes	yes	yes	yes	4	4	5	5	5	5	16	10	12	11
32	GP	M	15	U	GT3	T	4	4	health	services	home	mother	1	2	0	no	yes	yes	no	no	yes	yes	no	5	4	2	3	4	5	0	9	11	12
33	GP	M	15	U	GT3	T	4	4	services	services	reputation	mother	2	2	0	no	yes	no	yes	yes	yes	yes	no	4	3	1	1	1	5	0	17	16	17
34	GP	M	15	R	GT3	T	4	3	teacher	at_home	course	mother	1	2	0	no	yes	no	yes	yes	yes	yes	yes	4	5	2	1	1	5	0	17	16	16
35	GP	M	15	U	LE3	T	3	3	other	other	course	mother	1	2	0	no	no	no	yes	no	yes	yes	no	5	3	2	1	1	2	0	8	10	12
36	GP	M	16	U	GT3	T	3	2	other	other	home	mother	1	1	0	no	yes	yes	no	no	yes	yes	no	5	4	3	1	1	5	0	12	14	15
37	GP	F	15	U	GT3	T	2	3	other	other	other	father	2	1	0	no	yes	no	yes	yes	yes	no	no	3	5	1	1	1	5	0	8	7	6
38	GP	M	15	U	LE3	T	4	3	teacher	services	home	mother	1	3	0	no	yes	no	yes	yes	yes	yes	no	5	4	3	1	1	4	2	15	16	18

Although all these information can influence one's grade, we select 6 variables that seem to affect most on student's performance;

1) sex

(M,F)

2) parents education level

(mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - econdary education or 4 - higher education), father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 â€“ higher education) and to measure this we average two values; mother education level and that of fathers.

3) travel time to school

home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

4) paid for additional education

extra paid classes within the course subject (Math or Portuguese) (binary: yes or no

5) the number of going out with their friends.

going out with friends (numeric: from 1 - very low to 5 - very high)

6) the number of absences

number of school absences (numeric: from 0 to 93)

This survey measured student's performance with scale 0 to 20 with checking math and Portuguese grade.

1. first period grade

(numeric: from 0 to 20)

2. second period grade

(numeric: from 0 to 20)

3. final grade

(numeric: from 0 to 20)

3. What are the min/max/mean/median/sd values for each of these features?

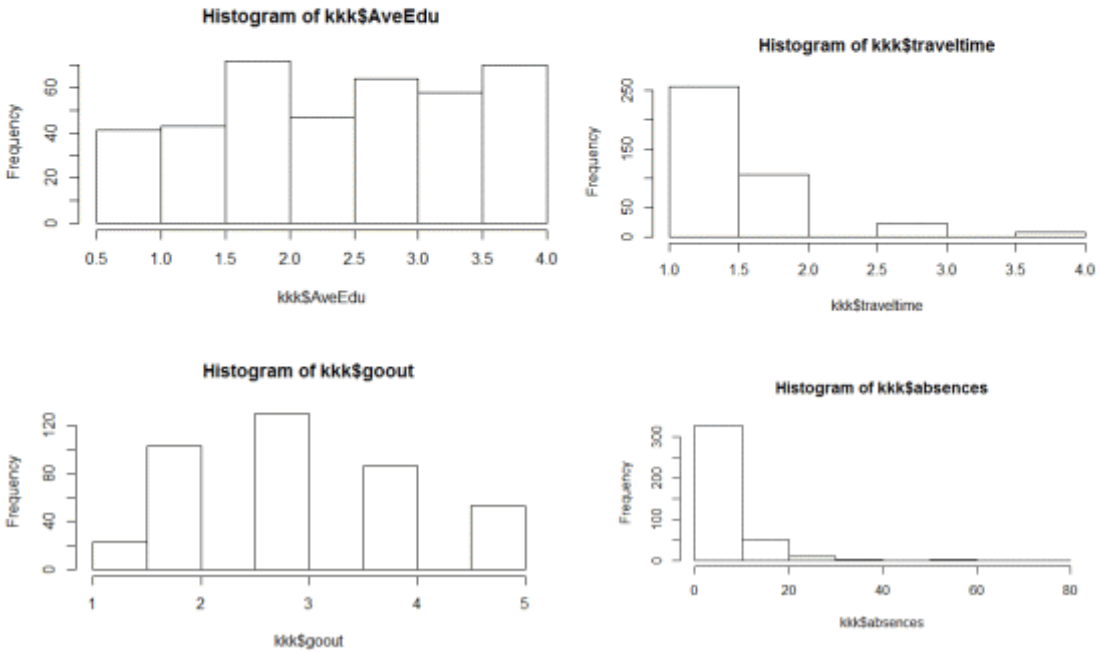
To calculate these, we omitted all information that we do not focus on and left only 9 variables; sex, parents education level, travel time to school, whether they paid for additional education, the number of going out with friends, the number of absences for semester, first period grade, second period grade and final grade. To measure the binary information, we changed women to 0 men to 1 and no paid to 0, paid 1. Also to simplify we average mother's and father's education level. For example if mother graduate only primary school and father has bachelor degree it would be

(1+4)/2=2.5 and like this, we average all grades of three periods, and if one get 14,15 and 10 at each period the value would be 13.

Since it does not have any meaning to calculate min/max/median/sd for sex and paid or not paid that were changed to 0 and 1 by us, we only calculated other variables statistical value.

4. What are the min/max/mean/median/sd values for each of these features? What is the distribution of the core features(show a histogram)?

	AveEdu	Traveltime	Paid	goout	absences	Avegrade
min	0.5	1	0	1	0	1.333333
max	4	4	1	5	75	19.33333
mean	2.635443	1.448101	0.458228	3.108861	5.708861	10.67932
median	2.5	1	0	3	4	10.66667
sd	0.983369	0.697505	0.498884	1.113278	8.003096	3.696786



5. Are the obvious trends in data and are the differences statistically significant?

We analyzed relationship between sex and grade.

Q : Does the difference between sex exist?

A : P-value is less than 0.05, so we could infer that there exists the differences of grade between sex.

<Two Sample t-test
data: kkk2$Avegrade and kkk2$Avegrade.1
t = -2.015, df = 393, p-value = 0.04459
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.4773516 -0.0181749
sample estimates:
mean of x mean of y
10.32532 11.07308>

6. Provide a bullet list of the next 5-10 tasks you will perform in analyzing your dataset.

1) relationship between sex and grade.
2) relationship between average level of education of parents and grade.
3) relationship between travel time to school and grade.
4) relationship between playing time of the subject and grade.
5) relationship between the number of absence in one semester and study performance.
6)relationship between whether they paid for additional education between grade.
7) relationship between independent variables.