UTM

UNIVERSITI TEKNOLOGI MALAYSIA

Probability & Statistical Data Analysis (SECI2143)

Section 02

Group 11

Project 2

Group Members

Liew Yvonne A21EC0045

Lau Yee Chi A21EC0042

Sam Chia Yun A21EC0127

Table of Content

Introduction or Background

This is an educational data set found in the website Kaggle which is collected from a learning management system (LMS) called Kalboard 360. Kalboard 360 is a multi-agent LMS, which has been designed to facilitate learning through the use of leading-edge technology. Such a system provides users with synchronous access to educational resources from any device with Internet connection.

The data is collected using a learner activity tracker tool, which is called experience API (xAPI). The xAPI is a component of the training and learning architecture (TLA) that enables monitoring learning progress and learner's actions like reading an article or watching a training video. The experience API helps the learning activity providers to determine the learner, activity and objects that describe a learning experience.

The dataset consists of 480 student records and 16 features. The features are classified into three major categories: (1) Demographic features such as gender and nationality. (2) Academic background features such as educational stage, grade Level and section. (3) Behavioral features such as raised hand on class, opening resources, answering surveys by parents, and school satisfaction.

We are interested in this dataset since we ourselves as students are curious what are the criteria/variables that affect the students' performance and what are the impact of this criteria/variable on the students' performance in class and their grades/marks.

We expect to see whether the mean of the medium class has higher occurrence of students raising hands in class or the mean of the high class has higher occurrence of students raising hands in class. Besides, we hope to see what is the relationship between the number of discussions participated and the number of raised hands in class.

Dataset

For hypothesis testing, correlation testing and Chi-Square test, the chosen dataset is students' academic performance in class dataset. "Raisedhand" stands for the number of times a student raise hand in class. "Discussion" stands for the number of times a student participates in a discussion group. "Class" stands for the class a student is in. "ParentschoolSatisfaction" stands for the satisfaction of parents towards the school. The "class" is classified into three levels, which are high (H), medium (M) and low (L) according to their total grade or mark. Students in high level have their total mark in interval including values from 90-100, students in medium level have their total mark in interval including values from 70-89, students in low level have their total mark in interval including values from 0-69. As the "class" is in the form of the alphabet (L,M,H), we changed the alphabet to numbers according to their levels. Low (L) changes to 0, medium (M) changes to 1 and high (H) changes to 2.

We choose "raisedhand" and "class" to do hypothesis testing (2 samples test). We will focus on two classes, which are high class (2) and medium class (1). We wish to determine if the mean of raised hand in the medium class is smaller than the high class. The possible outcome of the test will be that high class has a higher mean of raising hand than medium class. We choose "discussion" and "raisehands" to do correlation. The reason why we choose these two variables is because these two variables seem to be linked together and they are both ratio-type data. The claim is that "students who participate more in a discussion group are more likely to have a higher occurrence of raising hands in class". The possible outcome of the test will be a linear relationship and will produce a positive correlation. For the Chi-Square test, we use "ParentschoolSatisfaction" and "class" as the variables. We wish to determine whether the relationship between the class the student is in and the satisfaction of the parent towards school is dependent or independent. The possible outcome will be the relationship between the class the student is in and the satisfaction of the parent towards school is dependent.

The chosen dataset for regression analysis is a real estate data set. It shows the price of a house per unit area depending on different aspects such as "distance to the nearest MRT station" and "house age". The variables that we choose are "house age" and "house price per unit area". The reason why we choose these variables is that "house age" is an independent variable while "house price per unit area" is a dependent variable. We wish to determine whether the house age will affect the price of a house per unit area. The possible outcome of the test will be a negative linear relationship between house age and house price. The claim is that "the older the house age, the cheaper is house price per unit area".

Data analysis

Test 1: Hypothesis test

Hypothesis test on mean raised hand in class 1 and mean raised hand in class 2 to determine the relationship between two of them. We would like to know which class has the higher mean raised hand. Assume the confidence level to be 95%, significant level, α = 0.05. The parameters of interest are $\mu 1$ and $\mu 2$.

H_o: $\mu 1 = \mu 2$ (mean raised hand of both class is same)

H_1: $\mu 1 < \mu 2$ (mean raised hand of class 2 is greater than class 1)

α = 0.05

Class 1 who raised hand in class

n1=211

$\bar{x}1$=48.938

s1=26.894

Class 2 who raised hand in class

n2=142

$\bar{x}2$=70.433

s2=22.558

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

Test statistic, t_0= -8.117

$$v = \frac{\left(\frac{S_1^2}{n1}+\frac{S_2^2}{n2}\right)^2}{\frac{\left(S_1^2/n1\right)^2}{n1-1}+\frac{\left(S_2^2/n2\right)^2}{n2-1}}.$$

v=334.476

v≈334

Critical value, $t_{\alpha=0.05,\,334}$= -1.645

Conclusion:

Using α = 0.05, we reject H_o if $t_o < t_{0.05,\,334}$=-1.645. Since t_o= -8.117 < $t_{0.05,\,334}$= -1.645, we reject the null hypothesis. There is sufficient evidence to support mean raised hand in class 2 is more than mean raised hand in class 1.

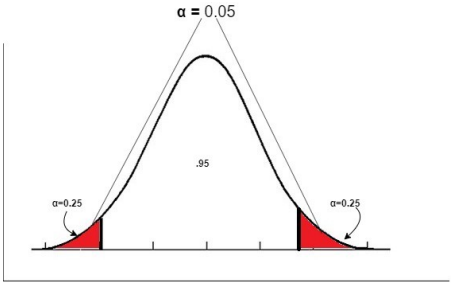
Test 2: Correlation Test

Correlation analysis to investigate the relationship between the number of discussions participated and the number of raised hands in class.

This test is to measure the strength of the relationship between the number of discussions participated and the number of raised hands in class. Assume the confidence level to be 95%, significant level, α = 0.05. The independent variable is the number of discussions participated by students while the dependent variable is the number of raised hands in class. Since both the variables are ratio scale data, hence Pearson's product-moment correlation using cor.test() function in R to obtain the correlation efficient (r).

H_o: ρ = 0 (no linear equation exists)

H_1: ρ ≠ 0 (A linear equation does exists)



The shaded region is the rejection region.

$\alpha = 0.05$,

$\alpha/2 = 0.025$ as it is 2 tailed test

Correlation coefficient, r = 0.339386,

Sample, n = 480,

Degrees of Freedom, $d.f = 480 - 2 = 478$

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Test statistic, t =7.888258

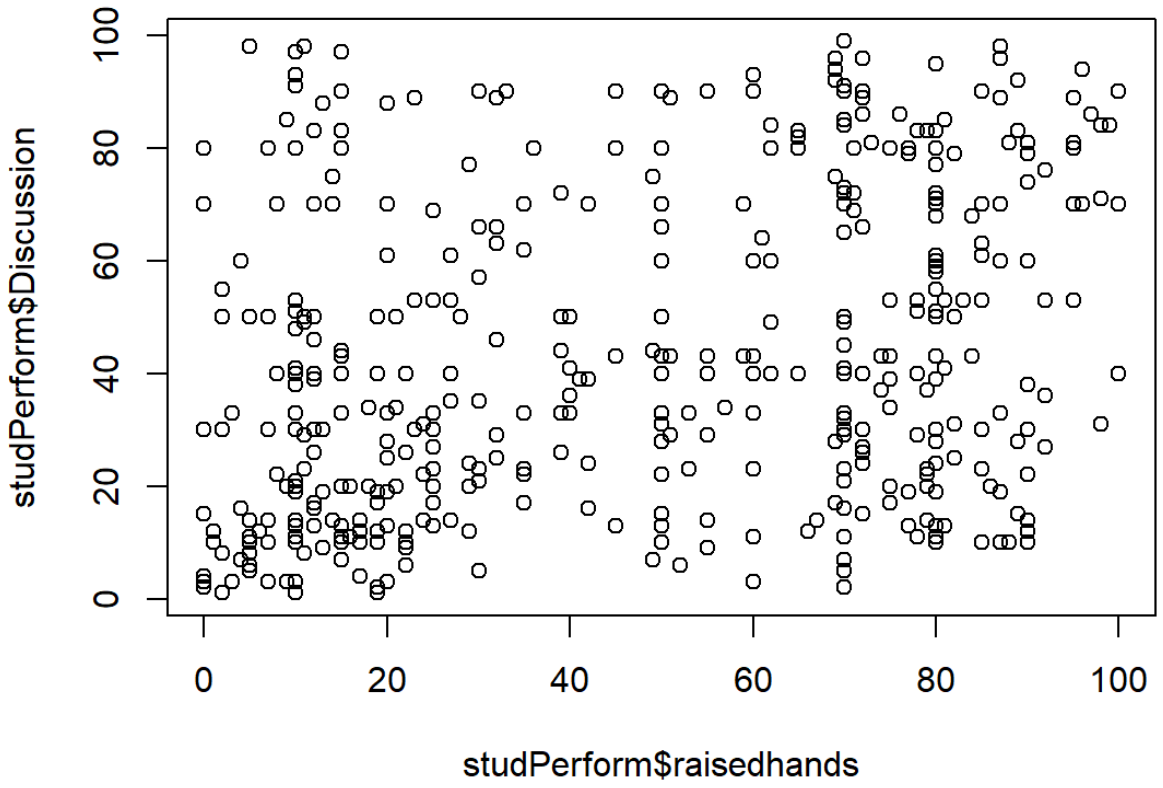
Critical value, $-t_{\alpha/2=0.025,\,478} = -1.968066$

$t_{\alpha/2=0.025,\,478} = 1.968066$

Conclusion:

Since the test statistic, t =7.888258 is not in between the $-t_{\alpha/2=0.025,\,478} = -1.968066$ and $t_{\alpha/2=0.025,\,478} = 1.968066$. It falls within the rejection region. Hence we reject the null hypothesis, H_0.

Since the correlation coefficient, r = 0.339386 which is positive and falls within 0 and 0.5, hence it has weak linear relationship between the number of discussions participated and the number of raised hands in class. There is sufficient evidence to prove that there is a linear correlation between the number of discussions participated and the number of raised hands in class.



Test 3: Regression test

This test is to predict the house price per unit area based on the house age. We wish to determine if there is a linear relationship between house age and house price per unit area. Assume the confidence level to be 95%, significant level, $\alpha = 0.05$. Independent variable (x) that is used in this test is "house age" while the dependent variable (y) is "house price per unit area".

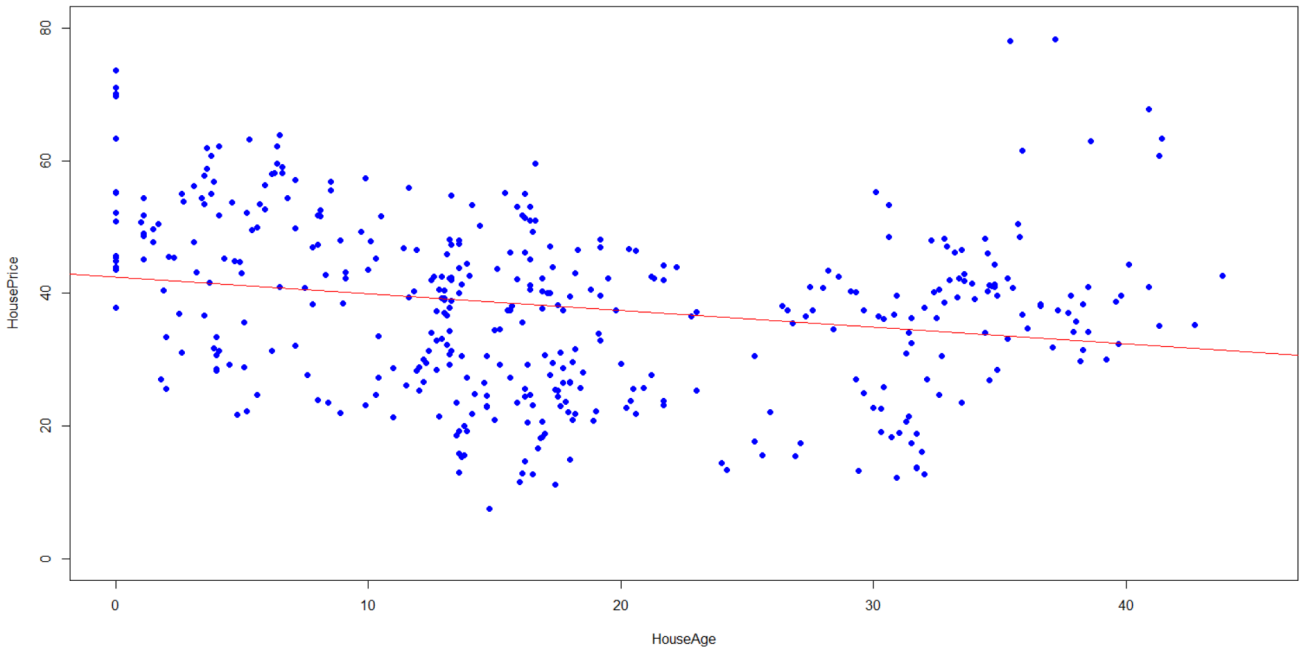
Estimated Regression Model, $\hat{y} = b_0 + b_1 x$
From the R script,
Estimated regression intercept, $b_0 = 42.4347$,
Estimated regression slope, $b_1 = -0.2515$
$\hat{y} = 42.4347 + (-0.2515)x$

H_o: $\beta_1 = 0$ (no linear relationship exists)
H_1: $\beta_1 \neq 0$ (linear relationship does exists)

$\alpha = 0.05$,
$\alpha/2 = 0.025$ as it is a 2 tailed2 test.
Sample, $n = 414$,
Degrees of Freedom, $d.f = 414 - 2 = 412$
Test statistic, $t = \frac{b_1 - \beta}{S_{b_1}}$
From the R script, $b_1 = -0.25149$, $S_{b_1} = 0.05752$, $t = -4.372$
Critical value, $-t_{\alpha/2=0.025,\,414} = -1.960$
$t_{\alpha/2=0.025,\,414} = 1.960$

Conclusion: Since $t = -4.372 < -t_{\alpha/2,\,412} = -1.960$, we reject the null hypothesis. There is sufficient evidence to prove that there is a linear relationship between the house age and house price per unit area.

From the graph, we can see it is a negative linear relationship as it has a negative slope which has a value of -0.2515. We can say that the house price per unit area depends on the house age. When the house age is older, the house price per unit area is lower. From the R script, only 4.434% of the variation in house price per unit area is explained by variation in house age. From the graph, we can see that the values are widely scattered. It is a weak linear relationship. There are some houses with a very old house age, however the price of the house is still quite high. This might be due to some other factors, such as the house is very near to the MRT station or it has many convenience stores nearby.


HousePrice
0
20
40
60
80
0
10
20
30
40
HouseAge

Optional test (Chi-Square Two Contingency Table):

This test is to determine whether the relationship between the class the student is in and the satisfaction of the parent towards school is dependent or independent. Assume the confidence level to be 95%, significant level, α = 0.05.

Ho : Variables are independent

H_1: Variables are dependent

α = 0.05

Class	Good	Bad	Total
0	43	84	127
1	131	80	211
2	118	24	142
Total	292	188	480

Class	Good, G	EG	Bad, B	EB	Total
0	42	77.26	84	49.74	254
1	131	128.36	80	82.64	422
2	118	86.38	24	55.62	284
Total	292	292	188	188	960

$$e_{ij} = \frac{(i\,th\ \ row\ total)\ x(\ j\,th\ \ row\ total)}{Total\ sample\ size}$$

e_{11} = 77.26

e_{12} = 49.74

e_{21} = 128.36

e_{22} = 82.64

e_{31} = 86.38

e_{32} = 55.62

Degrees of Freedom, $d.f$ = (3-1) (2-1)
= 2

Critical value, $\chi 2_{k=2,\ \alpha = 0.05}$ = 5.991

Cell, ij	Observed Count, Oij	Expected Count, eij	$\frac{O_{ij} - e_{ij}}{e_{ij}}$
1,1	43	77.26	15.192
1,2	84	49.74	23.598
2,1	131	128.36	0.054
2,2	80	82.64	0.084
3,1	118	86.38	11.575
3,2	24	55.62	17.976

$\chi 2$= 15.192 + 23.598 + 0.054 + 0.084 + 11.575 + 17.946
= 68.479

Test statistic : $\chi 2$ = 68.479

Critical value : $\chi 2_{k=2,\ \alpha = 0.05}$ = 5.991

Since test statistic value = 68.479 > critical value = 5.991, thus we reject Ho at α = 0.05. There is evidence that the variables are dependent.

Conclusion

From all the activities done in Project 2, we have learned to process the data by filtering the dataset in order to find out the useful variable for analysis purposes. Besides, we learn to analyze the data by finding the relationship between variables using statistical analytic methods like hypothesis test, correlation test and regression test. From the results, we find the interesting finding that the number of discussion groups participated by students does not have much relationship with the number of students who raise hands in class. In short, the progress of finding the result through analysis is going smoothly and we are coming to the end of the project thus, successfully finding out the result from the analysis.

Appendix

Students' Academic Performance Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	gender	NationalIT	PlaceofBir	StageID	GradeID	SectionID	Topic	Semester	Relation	raisedhanc	VisITedRes	Announcer	Discussion	ParentAns	Parentsch	StudentAb	Class
2	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	15	16	2	20	Yes	Good	Under-7	M
3	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	20	20	3	25	Yes	Good	Under-7	M
4	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	10	7	0	30	No	Bad	Above-7	L
5	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	30	25	5	35	No	Bad	Above-7	L
6	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	40	50	12	50	No	Bad	Above-7	M
7	F	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	42	30	13	70	Yes	Bad	Above-7	M
8	M	KW	KuwaIT	MiddleSch	G-07	A	Math	F	Father	35	12	0	17	No	Bad	Above-7	L
9	M	KW	KuwaIT	MiddleSch	G-07	A	Math	F	Father	50	10	15	22	Yes	Good	Under-7	M
10	F	KW	KuwaIT	MiddleSch	G-07	A	Math	F	Father	12	21	16	50	Yes	Good	Under-7	M
11	F	KW	KuwaIT	MiddleSch	G-07	B	IT	F	Father	70	80	25	70	Yes	Good	Under-7	M
12	M	KW	KuwaIT	MiddleSch	G-07	A	Math	F	Father	50	88	30	80	Yes	Good	Under-7	H
13	M	KW	KuwaIT	MiddleSch	G-07	B	Math	F	Father	19	6	19	12	Yes	Good	Under-7	M
14	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	5	1	0	11	No	Bad	Above-7	L
15	M	lebanon	lebanon	MiddleSch	G-08	A	Math	F	Father	20	14	12	19	No	Bad	Above-7	L
16	F	KW	KuwaIT	MiddleSch	G-08	A	Math	F	Mum	62	70	44	60	No	Bad	Above-7	H
17	F	KW	KuwaIT	MiddleSch	G-06	A	IT	F	Father	30	40	22	66	Yes	Good	Under-7	M
18	M	KW	KuwaIT	MiddleSch	G-07	B	IT	F	Father	36	30	20	80	No	Bad	Above-7	M
19	M	KW	KuwaIT	MiddleSch	G-07	A	Math	F	Father	55	13	35	90	No	Bad	Above-7	M
20	F	KW	KuwaIT	MiddleSch	G-07	A	IT	F	Mum	69	15	36	96	Yes	Good	Under-7	M
21	M	KW	KuwaIT	MiddleSch	G-07	B	IT	F	Mum	70	50	40	99	Yes	Good	Under-7	H
22	F	KW	KuwaIT	MiddleSch	G-07	A	IT	F	Father	60	60	33	90	No	Bad	Above-7	M
23	F	KW	KuwaIT	MiddleSch	G-07	B	IT	F	Father	10	12	4	80	No	Bad	Under-7	M
24	M	KW	KuwaIT	MiddleSch	G-07	A	IT	F	Father	15	21	2	90	No	Bad	Under-7	M
25	M	KW	KuwaIT	MiddleSch	G-07	A	IT	F	Father	2	0	2	50	No	Bad	Above-7	L
26	M	KW	KuwaIT	MiddleSch	G-07	B	IT	F	Father	0	2	3	70	Yes	Good	Above-7	L
27	M	KW	KuwaIT	MiddleSch	G-07	A	IT	F	Father	8	7	30	40	Yes	Good	Above-7	L
28	M	KW	KuwaIT	MiddleSch	G-07	B	IT	F	Father	19	19	25	40	Yes	Bad	Under-7	M
29	M	KW	KuwaIT	MiddleSch	G-08	A	Arabic	F	Father	25	15	12	33	No	Bad	Above-7	L
30	M	KW	KuwaIT	MiddleSch	G-08	A	Science	F	Father	75	85	52	43	Yes	Good	Under-7	M

Processed Students' Academic Performance Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	gender	NationalIT	PlaceofBir	StageID	GradeID	SectionID	Topic	Semester	Relation	raisedhanc	VisITedRes	Announcer	Discussion	ParentA	ParentschoolSatisfaction	StudentAb	Class
2	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	10	7	0	30	No	Bad	Above-7	0
3	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	30	25	5	35	No	Bad	Above-7	0
4	M	KW	KuwaIT	MiddleSch	G-07	A	Math	F	Father	35	12	0	17	No	Bad	Above-7	0
5	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Father	5	1	0	11	No	Bad	Above-7	0
6	M	lebanon	lebanon	MiddleSch	G-08	A	Math	F	Father	20	14	12	19	No	Bad	Above-7	0
7	M	KW	KuwaIT	MiddleSch	G-07	A	IT	F	Father	2	0	2	50	No	Bad	Above-7	0
8	M	KW	KuwaIT	MiddleSch	G-07	B	IT	F	Father	0	2	3	70	Yes	Good	Above-7	0
9	M	KW	KuwaIT	MiddleSch	G-07	A	IT	F	Father	8	7	30	40	Yes	Good	Above-7	0
10	M	KW	KuwaIT	MiddleSch	G-08	A	Arabic	F	Father	25	15	12	33	No	Bad	Above-7	0
11	M	KW	KuwaIT	MiddleSch	G-07	A	IT	F	Father	4	5	40	16	Yes	Good	Above-7	0
12	F	KW	KuwaIT	lowerlevel	G-07	A	IT	F	Father	2	19	10	50	Yes	Good	Above-7	0
13	M	KW	KuwaIT	lowerlevel	G-05	A	English	F	Father	8	22	9	40	No	Bad	Above-7	0
14	M	KW	KuwaIT	MiddleSch	G-07	B	Science	F	Father	12	11	8	40	No	Bad	Above-7	0
15	M	KW	KuwaIT	MiddleSch	G-07	A	English	F	Father	10	12	17	30	No	Bad	Above-7	0
16	M	KW	KuwaIT	MiddleSch	G-07	B	Science	F	Mum	8	6	4	22	Yes	Good	Above-7	0
17	M	KW	KuwaIT	MiddleSch	G-06	A	IT	F	Father	0	0	0	4	No	Bad	Above-7	0
18	F	KW	KuwaIT	MiddleSch	G-07	A	IT	F	Father	14	13	3	70	No	Bad	Above-7	0
19	M	KW	KuwaIT	HighSchoo	G-09	A	IT	F	Father	10	12	7	33	Yes	Good	Above-7	0
20	M	KW	KuwaIT	MiddleSch	G-07	A	Quran	F	Father	20	12	15	70	No	Good	Above-7	0
21	M	KW	KuwaIT	lowerlevel	G-05	A	English	F	Father	7	10	1	30	No	Bad	Above-7	0
22	F	KW	KuwaIT	HighSchoo	G-12	A	English	F	Father	20	14	12	70	No	Bad	Above-7	0
23	M	KW	KuwaIT	HighSchoo	G-12	A	English	F	Father	39	15	16	50	No	Good	Above-7	0
24	M	KW	KuwaIT	HighSchoo	G-12	A	English	F	Mum	12	50	8	30	No	Bad	Above-7	0
25	M	KW	KuwaIT	MiddleSch	G-07	A	Math	F	Father	16	14	6	20	Yes	Good	Above-7	0
26	M	USA	USA	MiddleSch	G-08	B	Math	F	Father	19	5	4	1	Yes	Good	Above-7	0
27	M	KW	KuwaIT	MiddleSch	G-07	A	Math	F	Father	5	2	6	5	Yes	Good	Above-7	0
28	F	Jordan	Jordan	lowerlevel	G-05	A	English	F	Mum	21	10	28	20	No	Good	Above-7	0
29	M	KW	KuwaIT	HighSchoo	G-10	A	IT	F	Father	0	5	7	2	No	Bad	Above-7	0
30	M	KW	KuwaIT	HighSchoo	G-12	A	English	F	Father	13	5	18	19	No	Bad	Above-7	0

Real Estate Dataset

	A	B	C	D	E	F	G	H
1	No	transaction date	house age	distance to the nearest MRT station	number of convenience stores	latitude	longitude	house price of unit area
2	1	2012.917	32	84.87882	10	24.98298	121.54024	37.9
3	2	2012.917	19.5	306.5947	9	24.98034	121.53951	42.2
4	3	2013.583	13.3	561.9845	5	24.98746	121.54391	47.3
5	4	2013.5	13.3	561.9845	5	24.98746	121.54391	54.8
6	5	2012.833	5	390.5684	5	24.97937	121.54245	43.1
7	6	2012.667	7.1	2175.03	3	24.96305	121.51254	32.1
8	7	2012.667	34.5	623.4731	7	24.97933	121.53642	40.3
9	8	2013.417	20.3	287.6025	6	24.98042	121.54228	46.7
10	9	2013.5	31.7	5512.038	1	24.95095	121.48458	18.8
11	10	2013.417	17.9	1783.18	3	24.96731	121.51486	22.1
12	11	2013.083	34.8	405.2134	1	24.97349	121.53372	41.4
13	12	2013.333	6.3	90.45606	9	24.97433	121.5431	58.1
14	13	2012.917	13	492.2313	5	24.96515	121.53737	39.3
15	14	2012.667	20.4	2469.645	4	24.96108	121.51046	23.8
16	15	2013.5	13.2	1164.838	4	24.99156	121.53406	34.3
17	16	2013.583	35.7	579.2083	2	24.9824	121.54619	50.5
18	17	2013.25	0	292.9978	6	24.97744	121.54458	70.1
19	18	2012.75	17.7	350.8515	1	24.97544	121.53119	37.4
20	19	2013.417	16.9	368.1363	8	24.9675	121.54451	42.3
21	20	2012.667	1.5	23.38284	7	24.96772	121.54102	47.7
22	21	2013.417	4.5	2275.877	3	24.96314	121.51151	29.3
23	22	2013.417	10.5	279.1726	7	24.97528	121.54541	51.6
24	23	2012.917	14.7	1360.139	1	24.95204	121.54842	24.6
25	24	2013.083	10.1	279.1726	7	24.97528	121.54541	47.9
26	25	2013	39.6	480.6977	4	24.97353	121.53885	38.8
27	26	2013.083	29.3	1487.868	2	24.97542	121.51726	27
28	27	2012.667	3.1	383.8624	5	24.98085	121.54391	56.2
29	28	2013.25	10.4	276.449	5	24.95593	121.53913	33.6
30	29	2013.5	19.2	557.478	4	24.97419	121.53797	47
31	30	2013.083	7.1	451.2438	5	24.97563	121.54694	57.1