PROJECT 2

PROBABILITY AND STATISTICAL DATA ANALYSIS

SECI / SCSI 2143

Title:

USA House Sales Price & How other factors affect its Variation

Name: Lee Sze Yuan

Subject: Probability & Statistical Data Analysis

Section: 02

Lecturer: Dr. Chan Weng Howe

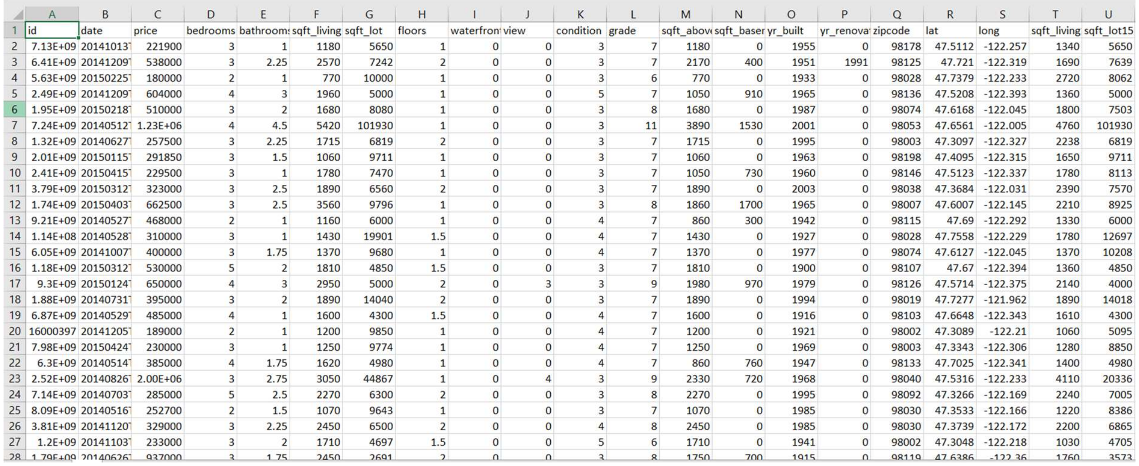
Date: 28/06/2020 (Sunday)

Table of Contents

1.0 Introduction

1.1 Dataset

For this project 2, we have visited a few websites and go through a few datasets which is suitable for this project. Since project 2 require us to conduct lots of hypothesis tests like linear regression, correlation etc. So, we look for dataset which have many numerical variables. These are the 2 datasets I found on Kaggle (https://www.kaggle.com/) and plan to use.

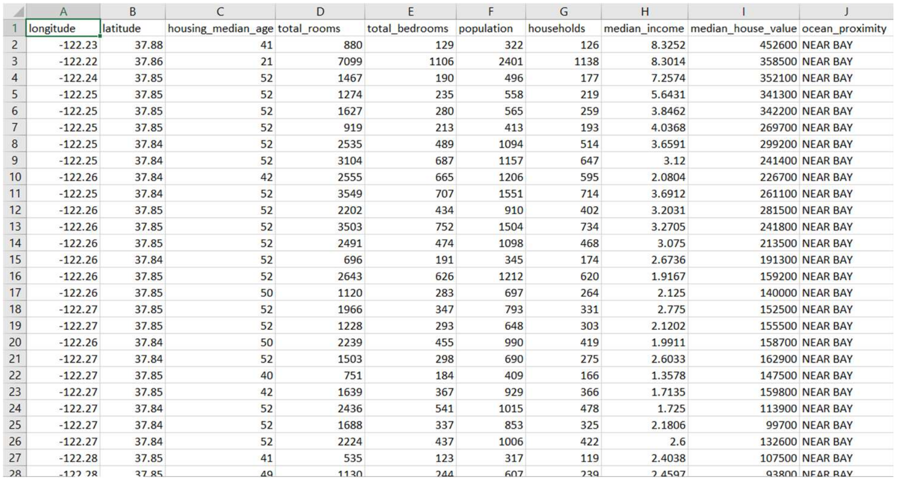


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	id	date	price	bedrooms	bathroom	sqft_living	sqft_lot	floors	waterfron	view	condition	grade	sqft_abov	sqft_baser	yr_built	yr_renova	zipcode	lat	long	sqft_living	sqft_lot15
2	7.13E+09	20141013T	221900	3	1	1180	5650	1	0	0	3	7	1180	0	1955	0	98178	47.5112	-122.257	1340	5650
3	6.41E+09	20141209T	538000	3	2.25	2570	7242	2	0	0	3	7	2170	400	1951	1991	98125	47.721	-122.319	1690	7639
4	5.63E+09	20150225T	180000	2	1	770	10000	1	0	0	3	6	770	0	1933	0	98028	47.7379	-122.233	2720	8062
5	2.49E+09	20141209T	604000	4	3	1960	5000	1	0	0	5	7	1050	910	1965	0	98136	47.5208	-122.393	1360	5000
6	1.95E+09	20150218T	510000	3	2	1680	8080	1	0	0	3	8	1680	0	1987	0	98074	47.6168	-122.045	1800	7503
7	7.24E+09	20140512T	1.23E+06	4	4.5	5420	101930	1	0	0	3	11	3890	1530	2001	0	98053	47.6561	-122.005	4760	101930
8	1.32E+09	20140627T	257500	3	2.25	1715	6819	2	0	0	3	7	1715	0	1995	0	98003	47.3097	-122.327	2238	6819
9	2.01E+09	20150115T	291850	3	1.5	1060	9711	1	0	0	3	7	1060	0	1963	0	98198	47.4095	-122.315	1650	9711
10	2.41E+09	20150415T	229500	3	1	1780	7470	1	0	0	3	7	1050	730	1960	0	98146	47.5123	-122.337	1780	8113
11	3.79E+09	20150312T	323000	3	2.5	1890	6560	2	0	0	3	7	1890	0	2003	0	98038	47.3684	-122.031	2390	7570
12	1.74E+09	20150403T	662500	3	2.5	3560	9796	1	0	0	3	8	1860	1700	1965	0	98007	47.6007	-122.145	2210	8925
13	9.21E+09	20140527T	468000	2	1	1160	6000	1	0	0	4	7	860	300	1942	0	98115	47.69	-122.292	1330	6000
14	1.14E+08	20140528T	310000	3	1	1430	19901	1.5	0	0	4	7	1430	0	1927	0	98028	47.7558	-122.229	1780	12697
15	6.05E+09	20141007T	400000	3	1.75	1370	9680	1	0	0	4	7	1370	0	1977	0	98074	47.6127	-122.045	1370	10208
16	1.18E+09	20150312T	530000	5	2	1810	4850	1.5	0	0	3	7	1810	0	1900	0	98107	47.67	-122.394	1360	4850
17	9.3E+09	20150124T	650000	4	3	2950	5000	2	0	3	3	9	1980	970	1979	0	98126	47.5714	-122.375	2140	4000
18	1.88E+09	20140731T	395000	3	2	1890	14040	2	0	0	3	7	1890	0	1994	0	98019	47.7277	-121.962	1890	14018
19	6.87E+09	20140529T	485000	4	1	1600	4300	1.5	0	0	4	7	1600	0	1916	0	98103	47.6648	-122.343	1610	4300
20	16000397	20141205T	189000	2	1	1200	9850	1	0	0	4	7	1200	0	1921	0	98002	47.3089	-122.21	1060	5095
21	7.98E+09	20150424T	230000	3	1	1250	9774	1	0	0	4	7	1250	0	1969	0	98003	47.3343	-122.306	1280	8850
22	6.3E+09	20140514T	385000	4	1.75	1620	4980	1	0	0	4	7	860	760	1947	0	98133	47.7025	-122.341	1400	4980
23	2.52E+09	20140826T	2.00E+06	3	2.75	3050	44867	1	0	4	3	9	2330	720	1968	0	98040	47.5316	-122.233	4110	20336
24	7.14E+09	20140703T	285000	5	2.5	2270	6300	2	0	0	3	8	2270	0	1995	0	98092	47.3266	-122.169	2240	7005
25	8.09E+09	20140516T	252700	2	1.5	1070	9643	1	0	0	3	7	1070	0	1985	0	98030	47.3533	-122.166	1220	8386
26	3.81E+09	20141120T	329000	3	2.25	2450	6500	2	0	0	4	8	2450	0	1985	0	98030	47.3739	-122.172	2200	6865
27	1.2E+09	20141103T	233000	3	2	1710	4697	1.5	0	0	5	6	1710	0	1941	0	98002	47.3048	-122.218	1030	4705
28	1.79E+09	20140626T	937000	3	1.75	2450	2691	2	0	0	3	8	1750	700	1915	0	98119	47.6386	-122.36	1760	3573

Figure 1.0: USA House Sales Price

The first dataset here is USA house Sales Price. This dataset contains the house sale prices for the County, USA. It includes homes sold in USA during the time between May 2014 and May 2015. This dataset is definitely great to conduct linear regression and correlation test. We will use this dataset as my main dataset and perform most of the data analysis and hypothesis testing on the data of this dataset.

Link: https://www.kaggle.com/harlfoxem/housesalesprediction



	A	B	C	D	E	F	G	H	I	J
1	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
2	-122.23	37.88	41	880	129	322	126	8.3252	452600	NEAR BAY
3	-122.22	37.86	21	7099	1106	2401	1138	8.3014	358500	NEAR BAY
4	-122.24	37.85	52	1467	190	496	177	7.2574	352100	NEAR BAY
5	-122.25	37.85	52	1274	235	558	219	5.6431	341300	NEAR BAY
6	-122.25	37.85	52	1627	280	565	259	3.8462	342200	NEAR BAY
7	-122.25	37.85	52	919	213	413	193	4.0368	269700	NEAR BAY
8	-122.25	37.84	52	2535	489	1094	514	3.6591	299200	NEAR BAY
9	-122.25	37.84	52	3104	687	1157	647	3.12	241400	NEAR BAY
10	-122.26	37.84	42	2555	665	1206	595	2.0804	226700	NEAR BAY
11	-122.25	37.84	52	3549	707	1551	714	3.6912	261100	NEAR BAY
12	-122.26	37.85	52	2202	434	910	402	3.2031	281500	NEAR BAY
13	-122.26	37.85	52	3503	752	1504	734	3.2705	241800	NEAR BAY
14	-122.26	37.85	52	2491	474	1098	468	3.075	213500	NEAR BAY
15	-122.26	37.84	52	696	191	345	174	2.6736	191300	NEAR BAY
16	-122.26	37.85	52	2643	626	1212	620	1.9167	159200	NEAR BAY
17	-122.26	37.85	50	1120	283	697	264	2.125	140000	NEAR BAY
18	-122.27	37.85	52	1966	347	793	331	2.775	152500	NEAR BAY
19	-122.27	37.85	52	1228	293	648	303	2.1202	155500	NEAR BAY
20	-122.26	37.84	50	2239	455	990	419	1.9911	158700	NEAR BAY
21	-122.27	37.84	52	1503	298	690	275	2.6033	162900	NEAR BAY
22	-122.27	37.85	40	751	184	409	166	1.3578	147500	NEAR BAY
23	-122.27	37.85	42	1639	367	929	366	1.7135	159800	NEAR BAY
24	-122.27	37.84	52	2436	541	1015	478	1.725	113900	NEAR BAY
25	-122.27	37.84	52	1688	337	853	325	2.1806	99700	NEAR BAY
26	-122.27	37.84	52	2224	437	1006	422	2.6	132600	NEAR BAY
27	-122.28	37.85	41	535	123	317	119	2.4038	107500	NEAR BAY
28	-122.28	37.85	49	1130	244	607	239	2.4597	93800	NEAR BAY

Figure 2.0: California Housing Price

Then, the second dataset here, we have this dataset “California Housing Price”. Based on the description given on Kaggle. This is a dataset used in the second chapter of Aurélien Géron's recent book 'Hands-On Machine learning with Scikit-Learn and TensorFlow'. The author of this book used this dataset to show how to perform various data analysis and also implementing machine learning algorithm. This dataset contains information from the 1990 California census. We will use this dataset as my second dataset. So that we can perform data analysis and hypothesis tests which require 2 samples together with the first dataset.

Link: https://www.kaggle.com/camnugent/california-housing-prices

1.2 Aim of Study

The purpose of this study is to study the house sales price in USA and California. This study wishes to find out the trend of the house sale price in USA and to see how house sales price are affected by other factors like total square feet, number of bedrooms etc. Then, if possible, to see how the house sales price has changed over time when compared to the house sales price in USA nowadays. Below is a list of purposes we have summarised for this study

- First, this study hopes to find out whether the claim of actual population parameter of USA houses sales prices can be rejected or should not be rejected. Or to
- Tests whether there is a change in the population parameters or not. These questions or hypothesis will be tested out by using the sample statistic of USA houses sales prices in between May 2014 and May 2015.
- Besides, this study also aims to observe and study the relationships between house sales price value and other variables. Example of variables include number of floors, number of bedrooms, number of bedrooms etc.
- This study aims to find out how each variable affect each other. Lastly, this study will also focus in studying whether there is difference in house sales price when the houses have different characteristics.

The target population we will use in this study is the Prices of house sales in USA and California.

2.0 Hypothesis Testing

Commonly Used Value

No	Dataset	$\bar{x}$	S	μ	N
1	USA	54008.1	367127.2	385000	21613
2	California	206855.1	115395.8	-	20604

2.1 1 Sample Hypothesis testing

2.1.1

To test whether the sample mean has any different from the claim of population mean

$H_0 : u = 385000$

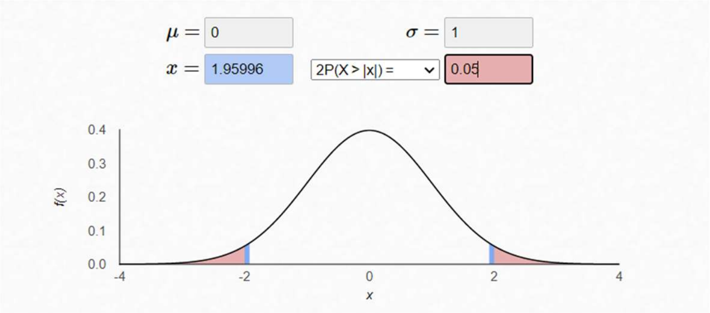
$H_1 : u \neq 385000$

$$\alpha = 0.05$$

$$z = \frac{\bar{x} - u}{s/\sqrt{n}} = 62.104$$

$$c.v. z_{0.05} = 1.960$$

$$p(z > 62.104) = 0$$



$\mu =$ 0

$\sigma =$ 1

$x =$ 1.95996

2P(X > |x|) = 0.05

Figure 3.0 Critical Region

Analysis: Since statistic value = 62.104 > 1.960 and P-value is lee than 0.05. We reject H_0 at a significance level of 0.05

Conclusion: There is sufficient evidence that population mean of house sales price is not equal to 385000.

2.1.2

To test whether the sample variance has any different from the claim of population variance

H_0: σ = 400000

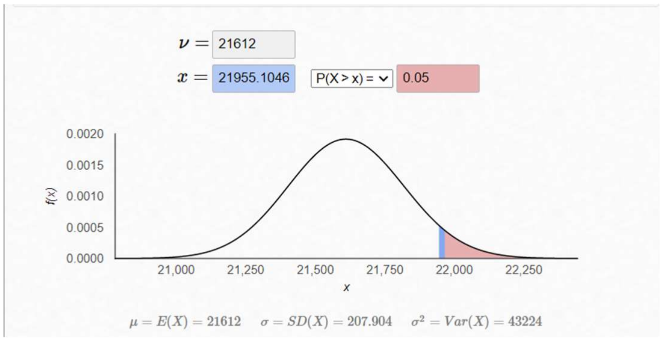
H_1 : σ < 400000

$$\alpha = 0.05$$

df = n-1 = 21612

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = 18205.73$$

$$c.v.\chi^2{}_{0.05,n-1} = 21955.1$$



$\nu =$ 21612

$x =$ 21955.1046 P(X > x) = 0.05

$\mu = E(X) = 21612$ $\sigma = SD(X) = 207.904$ $\sigma^2 = Var(X) = 43224$

Figure 4.0 Critical Region

Analysis: Since statistic value = 18205.73 < 21955.1. We do not reject H_0 at a significance level of 0.05

Conclusion: There is insufficient evidence that population variance of house sales price is less than 400000.

2.2 Correlation

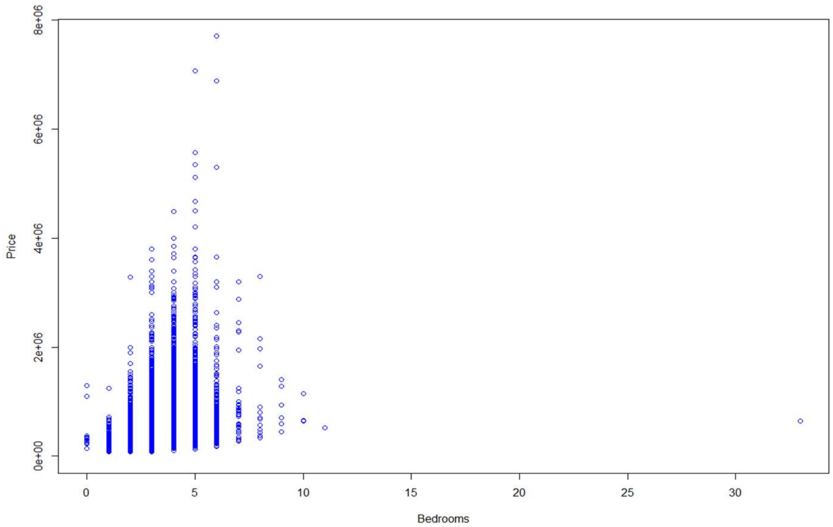
2.2.1

To measure strength of the linear relationship between price and number of bedrooms

X: No of Bedrooms Y: House Sales Price

H_0: p = 0

H_1 : p ≠ 0



r = 0.3083496

$\alpha = 0.05$

df = n-2 = 21611

Figure 5.0: Correlation of No of Bedrooms VS House Sales Price

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = 47.65139$$

P-value = 0

$c.v. t_{df,-\alpha/2}$ = -1.960 $c.v. t_{df,\alpha/2}$ = 1.960

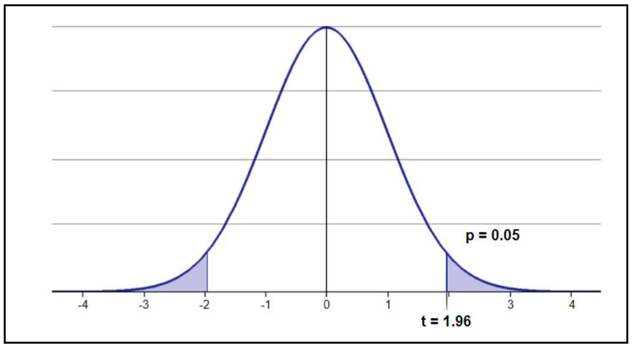


Figure 6.0: Critical Region

Analysis: Since statistic value = 47.65139 > 1.960 and P-value < 0.05. We reject H_0 at a significance level of 0.05

Conclusion: There is sufficient evidence of a linear relationship between No of Bedrooms and House Sales Price

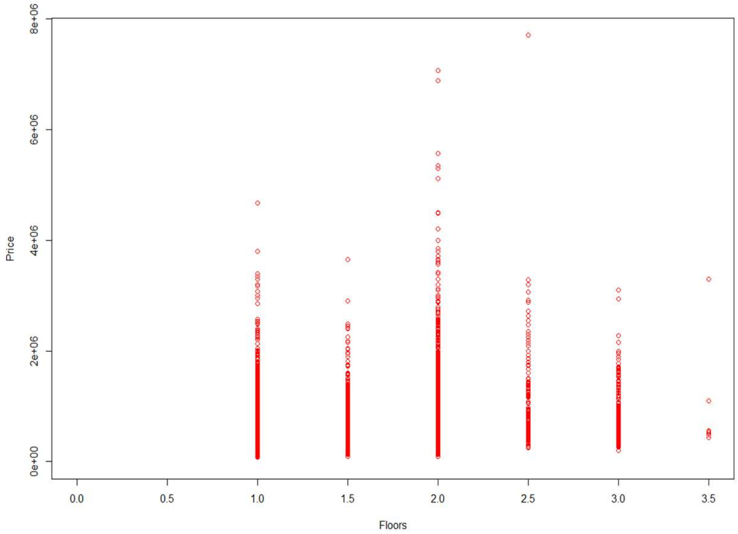
2.2.2

To measure strength of the linear relationship between price and number of floors

X: No of Floors Y: House Sales Price

H_0: p = 0

H_1 : p ≠ 0



r = 0.2567939

$\alpha = 0.05$

df = n-2 = 21611

Figure 7.0: Correlation of No of Floors VS House Sales Price

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = 39.06029$$

P-value = 1.58101e-322

$c.v.\, t_{df,-\alpha/2}$ = -1.960 $c.v.\, t_{df,\alpha/2}$ = 1.960

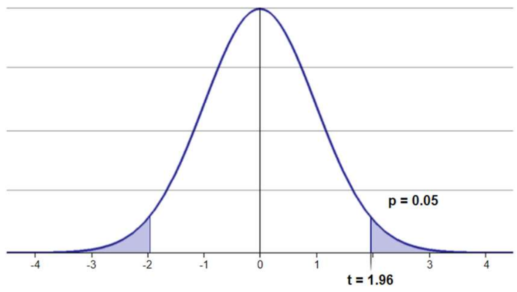


Figure 8.0: Critical Region

Analysis: Since statistic value = 39.06029 > 1.960 and P-value < 0.05. We reject H_0 at a significance level of 0.05

Conclusion: There is sufficient evidence of a linear relationship between No of Floors and House Sales Price

2.3 Regression

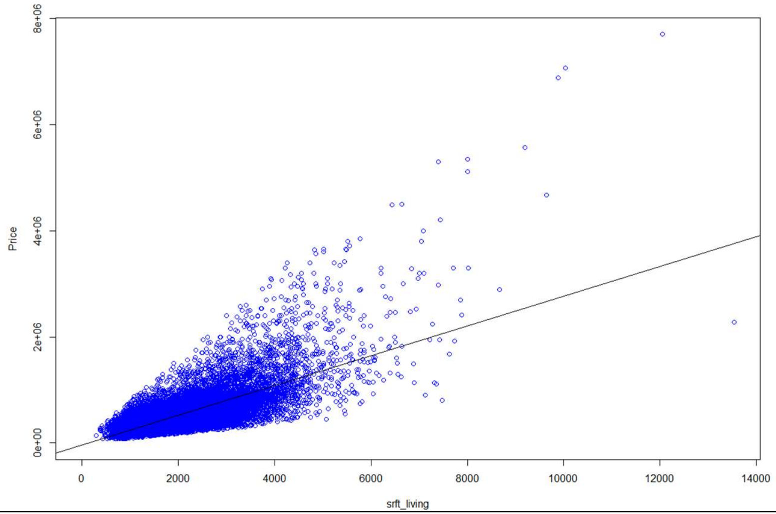
2.3.1

To predict the value of house price based on the value of sqft_living

Dependent: House Sales Price Independent: sqft_living

H_0: β1 = 0

H_1 : β1 ≠ 0



ŷ = -43580.7 + 280.6x

b_0 = -43580.7
$b_1 = 280.6$

$R^2 = 0.4928532$

$\alpha = 0.05$

df = n-2 = 21611

Figure 9.0: Regression of sqft_living VS House Sales Price

S_{b_1} =1.936

$$t = \frac{b_1 - \beta_1}{S_{b_1}} = 144.920$$

P-value = 0

$c.v.\,t_{df,-\alpha/2}$ = -1.960 $c.v.\,t_{df,\alpha/2}$ = 1.960

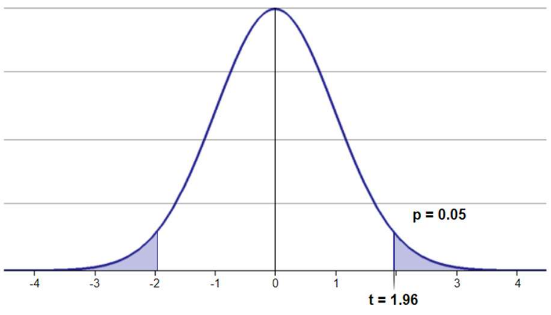


Figure 10.0: Critical Region

Analysis: Since statistic value = 144.920 > 1.960 and P-value < 0.05. We reject H_0 at a significance level of 0.05

Conclusion: There is sufficient evidence that the variable, sqft_living affects House Sales Price.

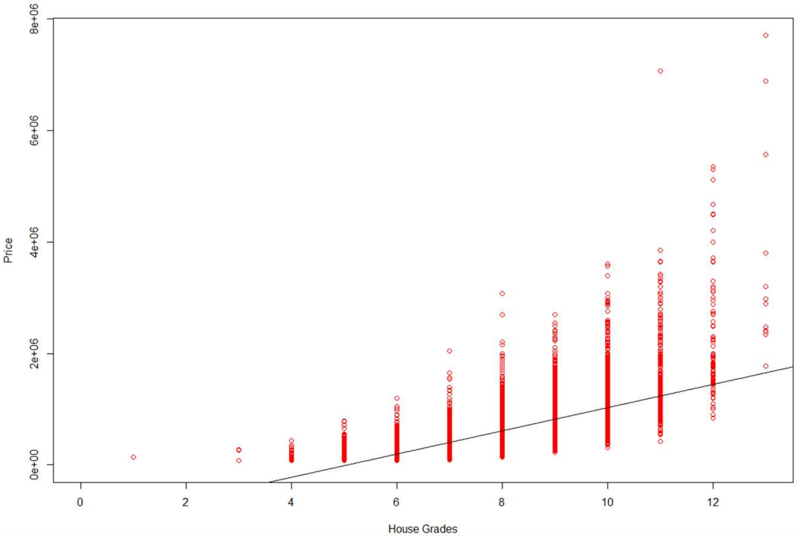
2.3.2

To predict the value of house price based on the value of house grade

Dependent: House Sales Price   Independent: House Grade

H_0: β1 = 0

H_1 : β1 ≠ 0



ŷ = -1056045+ 208458x

$b_0 =$ -1056045

b_1 =208458

$R^2 =$ 0.4454685

$\alpha = 0.05$

df = n-2 = 21611

Figure 11.0: Regression of House Grade VS House Sales Price

S_{b_1} =1582

$$t = \frac{b_1 - \beta_1}{S_{b_1}} = 131.76$$

P-value = 0

$c.v. t_{df,-\alpha/2}$ = -1.960 $c.v. t_{df,\alpha/2}$ = 1.960

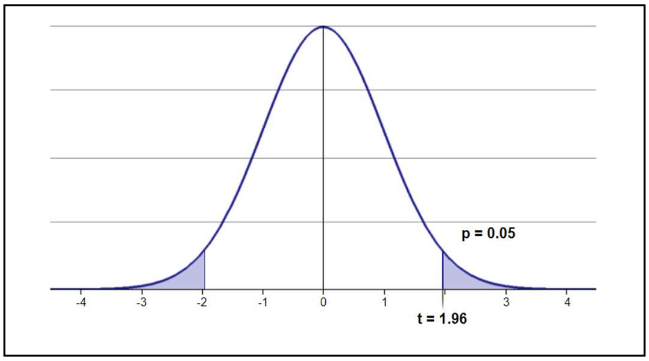


Figure 12.0: Critical Region

Analysis: Since statistic value = 131.76 > 1.960 and P-value < 0.05. We reject H_0 at a significance level of 0.05

Conclusion: There is sufficient evidence that the variable, House Grade affects House Sales Price.

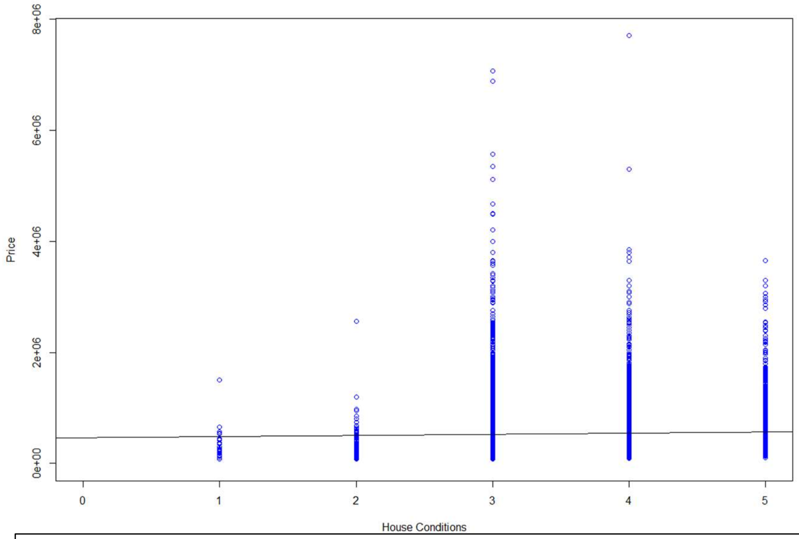
2.3.3

To predict the value of house price based on the house condition ratings

Dependent: House Sales Price Independent: House Condition

H_0: β1 = 0

H_1 : β1 ≠ 0



$\hat{y}$ = 470147 + 20514x

b_0 = 470147
b_1 = 20514

$R^2 = 0.00132218$

$\alpha = 0.05$

df = n-2 = 21611

Figure 13.0: Regression of House Condition VS House Sales

S_{b_1} =3835

$$t = \frac{b_1 - \beta_1}{S_{b_1}} = 5.349$$

P-value = 0

$c.v.\, t_{df,-\alpha/2}$ = -1.960 $c.v.\, t_{df,\alpha/2}$ = 1.960

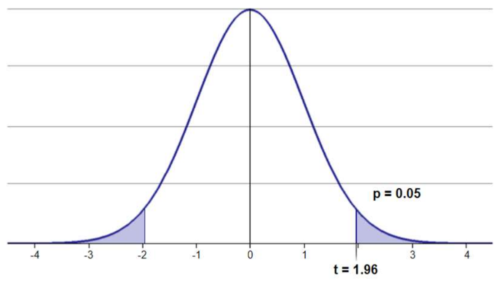


Figure 14.0: Critical Region

Analysis: Since statistic value = 5.349 > 1.960 and P-value < 0.05. We reject H_0 at a significance level of 0.05

Conclusion: There is sufficient evidence that the variable, House Condition affects House Sales Price.

2.4 Chi Square test of independence

2.4.1

To test whether grade and condition is independent or not

Ho: Grade is independent of the condition

H1: Grade is dependent of the condition

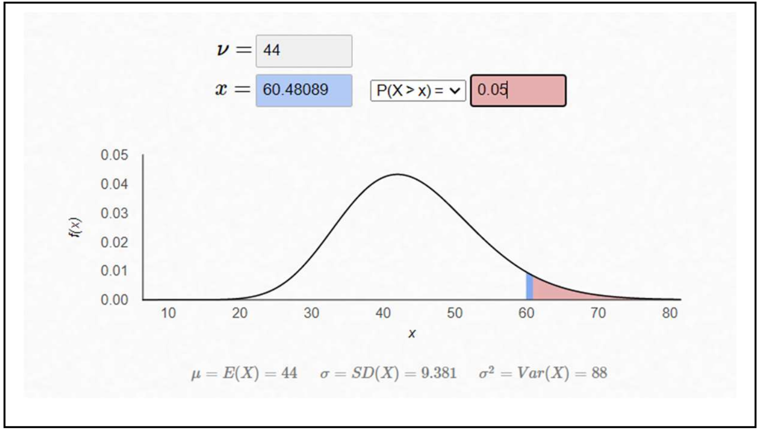
$\alpha = 0.05$

df = (R-1)(C-1) = 44

$$\chi^2 = \Sigma \frac{(0-E)^2}{E} = 2225.6$$

P-value = 0

$$\chi^2{}_{df,\alpha} = 60.481$$



$\nu =$ 44

$x =$ 60.48089 P(X > x) = 0.05

$\mu = E(X) = 44$ $\sigma = SD(X) = 9.381$ $\sigma^2 = Var(X) = 88$

Figure 15.0: Critical Region

Analysis: Since statistic value = 2225.6 > 60.481and P-value < 0.05. We reject H_0 at a significance level of 0.05

Conclusion: There is sufficient evidence that the variable, House Grade is independent of the variable, House Condition

2.5 2 Two Means, independent samples, unknown σ2

2.5.1

To test whether USA has same population of House Sales Price as California

Ho: u_1 of House Sales Price in USA = u_2 of House Sales Price in California

H1: u_1 of House Sales Price in USA ≠ u_2 of House Sales Price in California

$$\alpha = 0.05$$

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = 127.0313$$

$$v = \frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{\frac{(\frac{S_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{S_2^2}{n_2})^2}{n_2 - 1}} = 26023.39$$

P-value = 0

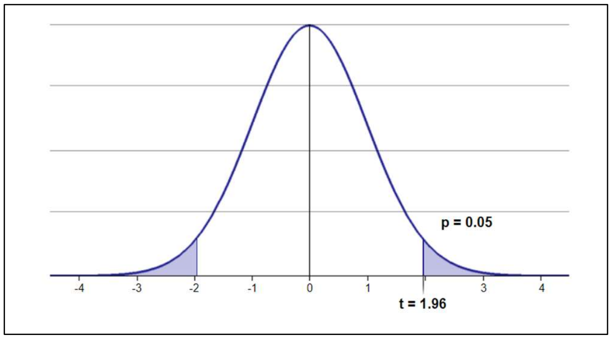


Figure 16.0: Critical Region

$c.v.\, t_{df,-^{\alpha}/_{2}}$ = -1.960 $c.v.\, t_{df,^{\alpha}/_{2}}$ = 1.960

Analysis: Since statistic value = 127.0313 > 1.960 and P-value < 0.05. We reject H_0 at a significance level of 0.05

Conclusion: There is sufficient evidence that the mean of House Sales Price in USA is not equal to mean of House Sales Price in California

3.0 Discussion

The focus of my study and analysis is to study the house sales price in USA. My study aims to study the house sales price data found on Kaggle and to see in what way, different factors affect the house sales price. After conducting several hypothesis tests, there are several findings to be discussed.

(A) Is House Sale Price in USA really as low as 385000?

To validate this claim from the **statista** website (link), I decide to run a 1-Sample Hypothesis testing on the mean of the house sales price in the dataset. For this test, the null hypothesis here will be 385000 and we want to find out whether this hypothesis should be accepted or should be rejected in a significance level of 0.05.

Then, throughout the process of Hypothesis testing, firstly, I found that the sample house sales price mean of the dataset is 540088.1 which deviates a lot from the claim by the website. Then, after further calculation, I obtain a result that showing the House Sales Price in USA has changed significantly over these few years because I obtain a P-value of roughly 0 (can't be calculated by RStudio). This result led to a conclusion that we should reject null hypothesis. There is insufficient evidence that the House Sales Price is equal to 38500. The house sales price in USA is much higher than the claim which mean the house sales price was more expensive back in the time.

(B) Is the deviation of house sale price in USA very big?

First, I claim that the house sale price in USA has a population standard deviation of 400000 based on the sample standard deviation of 367127.2 that I calculated in 1st test. Again, I decide to run a 1-sample Chi-square on the Hypothesis of the standard deviation at a significance level of 0.05. The critical value for this test is 21955.1

After calculation, I get a test statistic of 18205.73. So, as we can see, the test statistic is smaller than 21955.1 by around 10000. So, I can conclude that the null hypothesis on the population standard deviation is 400000 should not be rejected. The house sales price has this deviation.

(C) Is number of floors or bedrooms related to the house sales price?

As we commonly knew, the total square feet of a house is related to the house price in a linear proportional relationship. Because higher square feet mean more usable space and there will be more rooms for some purposes like study room, entertainment room, garage etc. (Gomez, 2019) But what wonders me next is will the number of rooms or floors related to the house sales price in the same way as total square feet do. The reason I have this question is because with more rooms, people can create different rooms of different purposes. Then, with more floors, the usable space also increases.

So, I decide a do a correlation analysis on No of Bedrooms VS Price and No of Floors VS Price. This is to check whether there is relationship between these 2 variables and the House Sales Price.

The independent variables will be No of Bedrooms and No of Floors and dependent variables will be Price.

As what was shown on the last section 2.0, I managed to get the Sample Correlation Coefficient, r for both (x, y) pairs. For the first (x, y) pair: No of Bedrooms VS Price. The r value is only 0.3083496. And for the second (x, y) pair: No of Floors VS Price. The r value is 0.2567939. The r value for each (x, y) pair indicates that the linear relationship between independent variable, x and dependent variable, y is POSTIVE but WEAK. But before I make a conclusion that there is a correlation exists between each (x, y) pairs.

I have done significance test for correlation on each (x, y) pair at a significance level of 0.05 and degree of freedom of 21611. Then, the t-value test statistic for both test (1st Pair: 47.65139, 2nd Pair: 39.06029) has exceeded the critical value 1.960 a lot. This indicates that there is correlation exists in both pair, but the relationship is just a weak positive relationship.

The possible reason behind this result is there could be some house that have more rooms or floors, but the total square feet is still small. So, people less prefer this kind of house and they still prefer house with bigger total square feet even if there is less rooms or floors

(D) Can we predict the house sales price based on the sqft_living, grade and house conditions to predict the house sales price?

After I finished in studying the relationship between No of Bedrooms and Price and the relationship between No of Floors and Price. I want to see whether can we predict the house sales price based on the sqft_living, grade and house. Can sqft_living, grade and condition helps people to predict the house price easily.

So, I will do a fit a linear regression model to predict the house sales price based on these 3 variables. These 3 variables are sqft_living, grades and conditions. My expectation for these tests is that great portion of the total variation in the house sales price is explained by variation in the sqft_living, grade and house conditions.

After I execute the analysis on R Studio to try to find the coefficient of determination value, R^2, I get quite a different result from my expectation. Only 49.29% and 44.54% of the variation of house sales price can be explained by sqft_living and grade respectively. However, for house condition, this independent variable can only explain less than 1% of the variation of house sales price which is 00.13%.

For further confirmation, I have conducted t test on each linear regression models. The result of t test shows that all 3 variables have a linear relationship with the house sales price. Just 2 of the variables have a stronger linear relationship than another one.

First, I expect that sqft_living could explain like 80% of the variation in the house sales price. However, it turns out that sqft_livig could only explain roughly 50% of that. This shows that sqft_living is not the biggest factor that are causing the variation in the house sales price. This also shows that there is other factor affecting the variations of house sales price. This is something interesting for further investigation.

Then, I expect that grade and house conditions could explain roughly 60% of the variation in the house sales price. But it turns out that Grade could only explain 44.54%. So, just like sqft_living, grade can only a portion of the variations.

For house conditions, it is the weirdest one. According to Opendoor.com, homes that are newer appraise at a higher value (in better condition). But the house condition could only explain 00.13%. This shows that house condition is not an important factor in affecting the house sales price. House sales price will be affected more by the grade and sqft_living than the factor, house conditions

(E) Is grade related to conditions in any ways?

After I found out that Garde and house conditions affect the variation of house sales price differently. I want to further investigate these 2 variables which seem connected in some ways. My hypothesis is that a house with good house conditions will have a high grade too. So, I decided to run a Chi Square test of Independence on the grade variables and house conditions variable on a significance level of 0.05 and a degree of freedom of 44.

To do this Chi Square test of Independence, I have calculated the X2 value which is 2225.6 and the critical value is 60.481. So, the X2 value is more than the critical value a lot, this shows that Grade is dependent of the house conditions. House Conditions have some kind of relationship with the Grade.

(F) Mean house sales price in USA and Mean house sales price in California

So, I also decide to do a test between USA and California. The purpose I do this test is to study is the mean house sales price in USA same as that in California. So, I decided to do a Two Means, independent samples, unknown σ2 test at a significance level of 0.05. Then the result I get is that the t value is much higher than that of critical value and the p-value is less than 0.05. So, the mean of house sales price in USA is different from that of California. The reason behind this result could be that the location, economic, market and other factors causing the difference between mean of house sales price in USA and California. (Unknown, 2019)

(G) Possible Factor in affecting the house sales price

First, according to Opendoor.com, Economic and local market (if there is a lot of buyer) is a factor that might affect house sales price. For example, if employment or wage growth slows, then fewer people might be able to afford a home. Then, there will be less demand for house. SO, this will affect the house sales price. (Gomez, 2019)Then, the location of a property is also an important factor. For instance, if it is close to places or things like public transport, shops, schools and restaurants. Then the price will be higher. (Unknown, 2019) Lastly, if the house is at an areas with good reputations, low crime ratesm then it will have a higher price too. (Whitten, 2017)

4.0 Conclusion

In this project, I learned a lot of things and can make a few conclusions:

- First, the USA house price has varied a lot from the past
- And there are many factors affecting & also not affecting the house sales price. The price is not necessarily affected by the factors like total square feet, grade, No of floors only. There are other factors that are not obvious like location, economic and market which are affecting house sales price
- And the general price in USA is different from that of California

In a nutshell, although we might be able to predict house sales price based on these factors, but there is a lot of unforeseen variables we need to take in consider also before predicting the house sales price.

5.0 References

Bognar, M. (2019). *Normal Distribution*. Retrieved from Department of Statistics and Actuarial Science: https://homepage.divms.uiowa.edu/~mbognar/applets/normal.html

Bognar, M. (2019). *Student's t-Distribution*. Retrieved from Department of Statistics and Actuarial Science: https://homepage.divms.uiowa.edu/~mbognar/applets/t.html

Bognar, M. (2020). *Chi-Square Distribution*. Retrieved from Department of Statistics and Actuarial Science: https://homepage.divms.uiowa.edu/~mbognar/applets/chisq.html

Gomez, J. (27 3, 2019). *8 critical factors that influence a home's value*. Retrieved from opendoor.com: https://www.opendoor.com/w/blog/factors-that-influence-home-value

Unknown. (03 May, 2019). *7 Factors Impacting Price in Malaysia's Property Market*. Retrieved from propertyguru.com.my: https://www.propertyguru.com.my/property-guides/7-factors-impacting-price-in-malaysia-s-property-market-13759

Whitten, R. (2017). *What influences a property's value?* Retrieved from finder.com: https://www.finder.com.au/what-influences-a-propertys-value