REGRESSION STUDY ON HOUSE PRICES

INTRODUCTION

House ownership has always been part of American Dream. Because of that, many people accept owning a home as the right, without considering the benefits and risks. Owning a house provides stable and safe environment. Home ownership is a reasonable safe long-term investment and provides a rent-free retirement.

This study seeks to determine if a relationship exists between house price and the features, conditions of house using the Multiple Regression on SAS Enterprise Miner. This project is helpful for the people who wants to buy a house, also to the real estate people, one can get a clear vision about the price of a house.

LITERATURE REVIEW

In our point of view, we believe that the increment of a housing price is due to features and conditions of the house, I have found out several more important variables that leads to the factors in determination of housing price. These factors include number of bedrooms, area of the living and lot, floor, area of basement, condition, grade, year it built, zip code, year it renovated, longitude and latitude.

1) One who has studied using an empirical analysis has shown that income (demography trends) and nominal interest rates are the key explanatory factors in housing price. On the other hand, the equity returns may also have been an influential factor in the determination of housing price (Pages & Maza, 2003).
2) The study of Yazgi & Dokmeci (2007), in which, applied the multiple regression model and it was found that the most important factor affecting the housing prices is the size of the floor area. The second and third most effective factors are the road surface ratio and the floor area, and fourth factor is the distance of housing to the seashore. This means, the bigger the size of floor the more money a purchaser will have to pay. This study is obviously different from the previous study. It looked upon the geographical area.
3) The studies of Hiebert & Roma (2010) by using the empirical and regression analysis have shown that even though a particular location is associated with

facilities, population differences across cities influenced more. Therefore, even if a particular location is fully facilitated, people will still prefer living among a population.

4) The price expectations that is predictable contribute to the increased of housing price. Thus, it was found that the price of housing is influenced by the regional economic activity, the regional housing stock, and the user cost of capital (Grimes et al, 2004).

DATA AND METHODOLOGY

Empirical data for this study is obtained from a house sales prices for king county, Seattle. It includes houses sold between May 2014 and May 2015. The target variable is the price of house and is measured by the dependent variables that gives the information of a housing community. The dataset has following columns:

- price: The price of the house
- bedrooms: Number of bedrooms
- bathrooms: Number of bathrooms
- sqft_living: Area of living area
- sqft_lot: Area of lot
- sqft_base: Area of basement
- floors: Number of floors
- yr_built: Year the house is built.
- yr_renovated: Year the house is renovated.
- zipcode: The zipcode of the house
- lat: The latitude of the house
- long: The longtitude of the house

CODE

```
proc contents data = work.price;
run;
```

```
proc freq data = work.price;
tables floors waterfront view;
run;
proc sgplot data = work.price;
histogram bedrooms;
run;
proc sgplot data = work.price;
histogram bathrooms;
run;
proc sgplot data = work.price;
histogram sqft_living;
run;
proc sgplot data = work.price;
histogram floors;
run;
proc sgplot data = work.price;
histogram yr_built;
run;

proc reg data = work.price;
model price = bedrooms bathrooms sqft_living sqft_lot floors waterfront view
condition grade sqft_above sqft_basement yr_built yr_renovated sqft_living15
sqft_lot15;
run;
/* removed variable sqft_basement as it is endogeneous variable */
proc reg data = work.price plots(maxpoints=none);
model price = bedrooms bathrooms sqft_living sqft_lot floors waterfront view
condition grade sqft_above yr_built yr_renovated sqft_living15 sqft_lot15;
plot  residual.*price.;
run;
```

ANLYSIS AND RESULTS

Contents of dataset:

The SAS System

The CONTENTS Procedure

Data Set Name	WORK.PRICE	**Observations**	21613
Member Type	DATA	**Variables**	21
Engine	V9	**Indexes**	0
Created	04/27/2020 19:14:10	**Observation Length**	176
Last Modified	04/27/2020 19:14:10	**Deleted Observations**	0
Protection		**Compressed**	NO
Data Set Type		**Sorted**	NO
Label			
Data Representation	WINDOWS_64		
Encoding	wlatin1 Western (Windows)		

The SAS System

The FREQ Procedure

floors	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	10680	49.41	10680	49.41
2	10151	46.97	20831	96.38
3	774	3.58	21605	99.96
4	8	0.04	21613	100.00

waterfront	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	21450	99.25	21450	99.25
1	163	0.75	21613	100.00

view	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	19489	90.17	19489	90.17
1	332	1.54	19821	91.71
2	963	4.46	20784	96.16
3	510	2.36	21294	98.52
4	319	1.48	21613	100.00

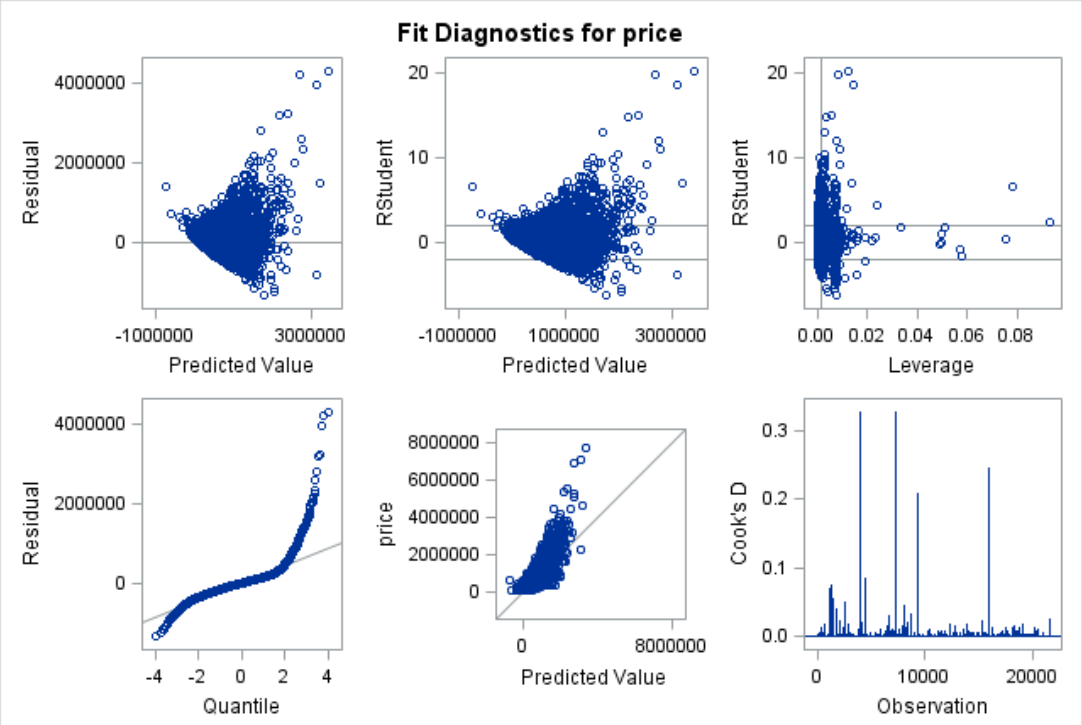
Frequency of each attribute is shown above, it clearly defines that people majorly choose first floor. Major number of houses does not have waterfront and there is no view for many houses.

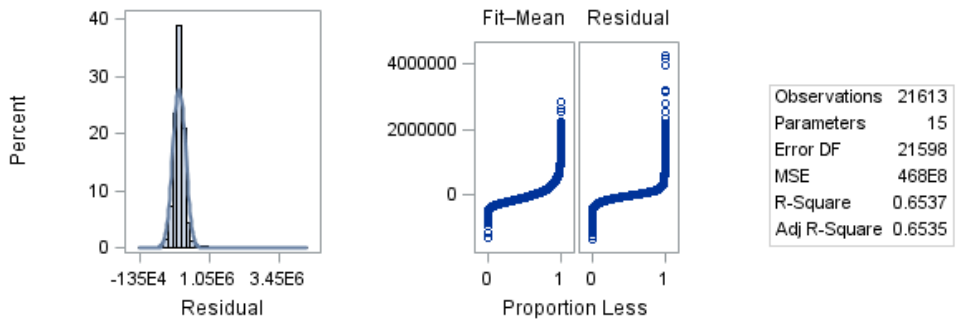
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6069890	137284	44.21	<.0001
bedrooms	1	-39782	2026.37560	-19.63	<.0001
bathrooms	1	46478	3474.49151	13.38	<.0001
sqft_living	1	166.52773	4.65572	35.77	<.0001
sqft_lot	1	-0.00398	0.05131	-0.08	0.9382
floors	1	23100	3337.10718	6.92	<.0001
waterfront	1	578886	18648	31.04	<.0001
view	1	43396	2274.18647	19.08	<.0001
condition	1	19374	2496.88307	7.76	<.0001
grade	1	119994	2248.46173	53.37	<.0001
sqft_above	1	-5.69664	4.53522	-1.26	0.2091
yr_built	1	-3504.94890	70.20995	-49.92	<.0001
yr_renovated	1	10.84089	3.91380	2.77	0.0056
sqft_living15	1	24.90152	3.60592	6.91	<.0001
sqft_lot15	1	-0.55733	0.07837	-7.11	<.0001

Parameter estimates are shown in the above figure.

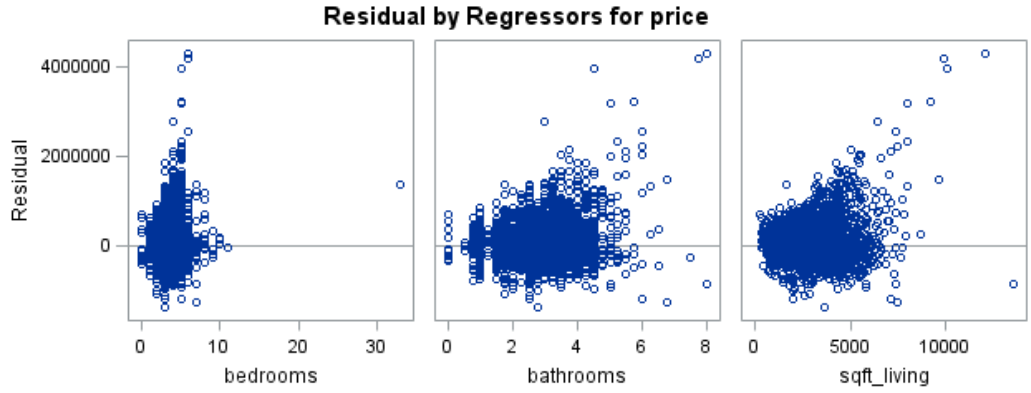
The SAS System

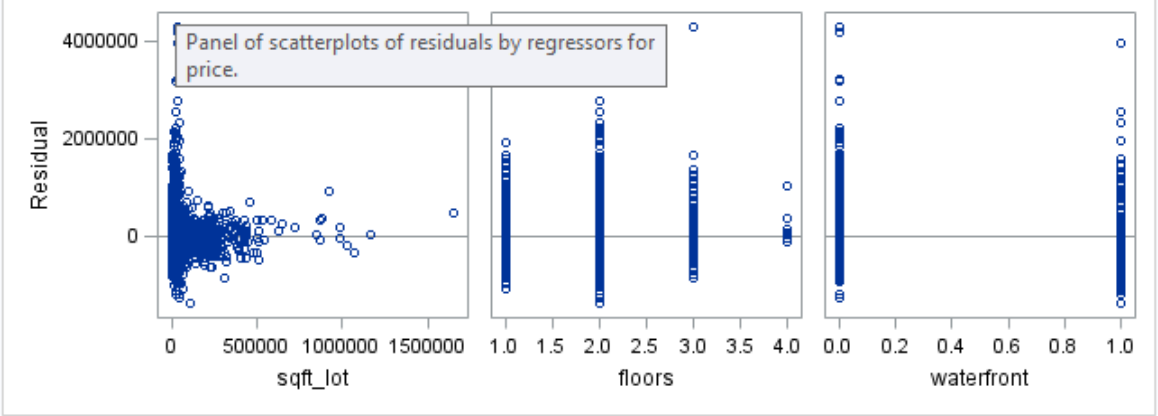
The REG Procedure
Model: MODEL1
Dependent Variable: price

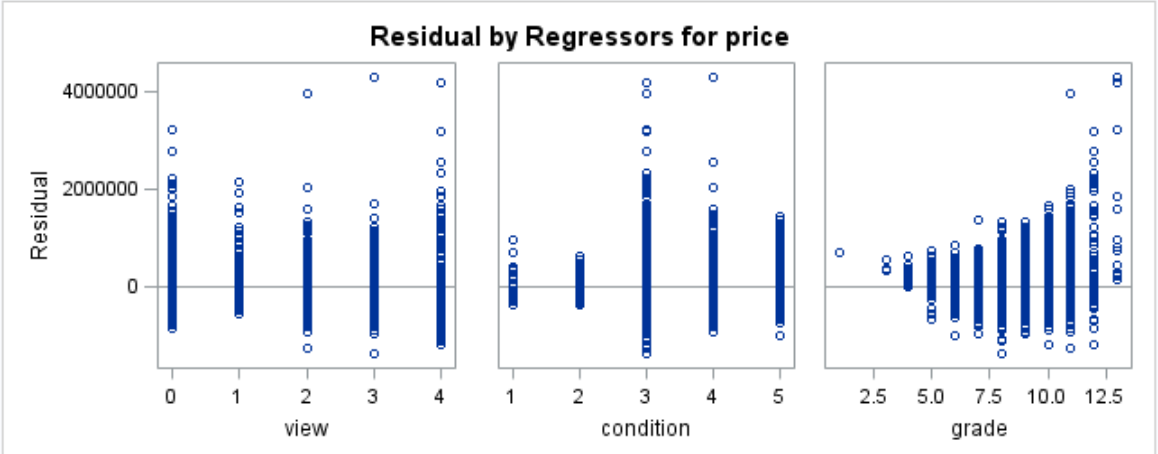


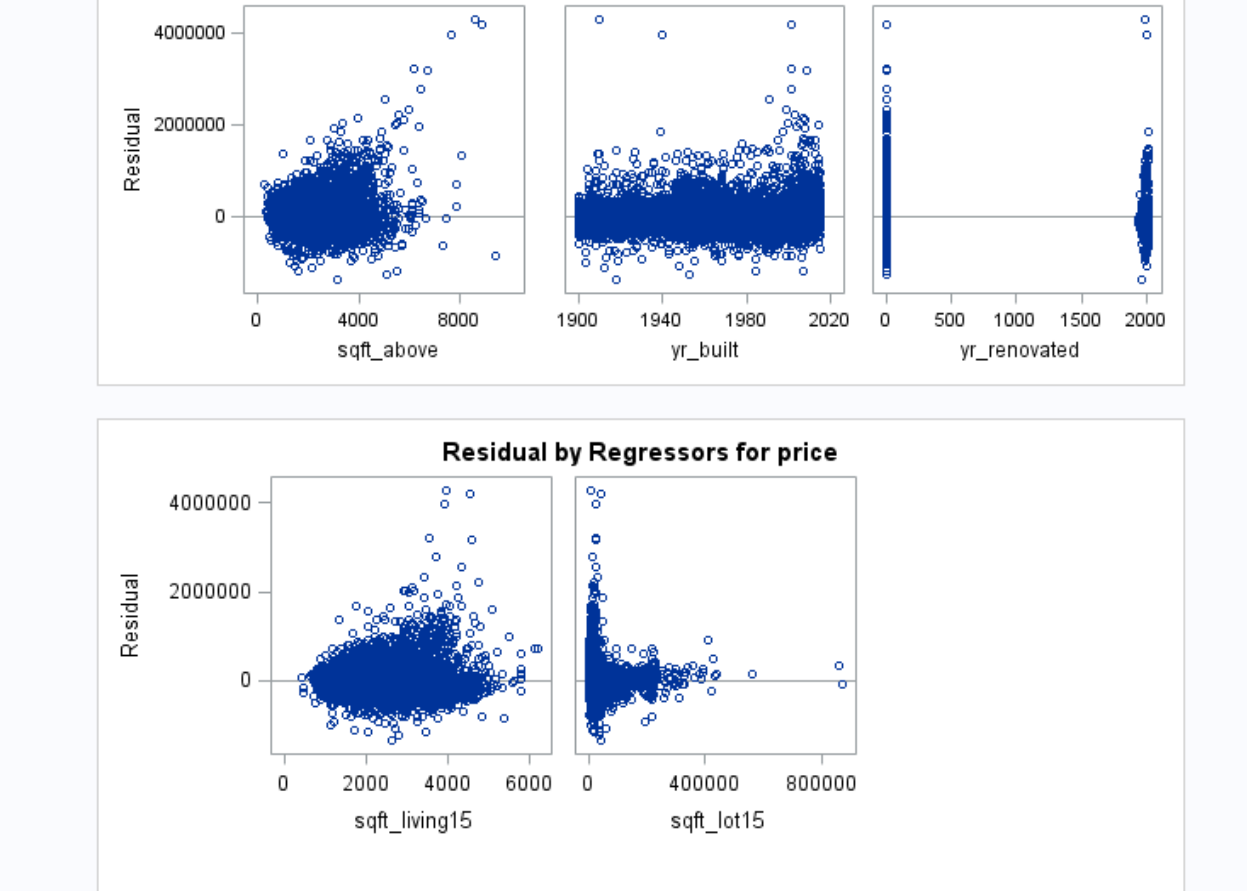


Observations	21613
Parameters	15
Error DF	21598
MSE	468E8
R-Square	0.6537
Adj R-Square	0.6535




Panel of scatterplots of residuals by regressors for price.
Residual
4000000
2000000
0
0
500000
1000000
1500000
sqft_lot
1.0
1.5
2.0
2.5
3.0
3.5
4.0
floors
0.0
0.2
0.4
0.6
0.8
1.0
waterfront


Residual by Regressors for price
Residual
4000000
2000000
0
0
1
2
3
4
view
1
2
3
4
5
condition
2.5
5.0
7.5
10.0
12.5
grade



The above residual plots show classifies according the number of bedrooms, bathrooms and the area of sqft_liv, sqft_lot and number of views, condition, grades, waterfront, year built and yr_renovated.

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: price

Number of Observations Read	21613
Number of Observations Used	21613

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	1.906724E15	1.361946E14	2912.63	<.0001
Error	21598	1.009924E15	46760050110		
Corrected Total	21612	2.916648E15			

Root MSE	216241	R-Square	0.6537
Dependent Mean	540182	Adj R-Sq	0.6535
Coeff Var	40.03107		

The F-value is 2912.63 and P value is < 0.05. So, the regression model is significant.

The P-value for the t-statistic of the selected variables is all <=0.05, so all the variables are significant in the model.

Above observation is the result of the regression procedure, it shows the value of

$r^2 = 0.6537$ which means 65.37% is explained by the taken independent variables.

CONCLUSION

In future we can also include the latitude, longitude, elevation of the house in the model to predict the house price with more accuracy. Future work can also include demographic variable like income, number of children in the model to explain the variability in the house pricing and to predict the house pricing more effectively.