Individual Assignment: SA3

Yash Srivastava
Yash_srivastava_ampba2021s@isb.edu
PG ID: **12010060**

Objective: To predict house prices in Seattle based on the data provided.

DATASET: The dataset contains house sale data for Seattle, Washington. Following are the features/ variables available.

Snapshot of the Dataset:

```
'data.frame':	999 obs. of  25 variables:
 $ id             : num  7129300520 6414100192 5631500400 2487200875 1954400510 ...
 $ date           : chr  "20141013T000000" "20141209T000000" "20150225T000000" "20141209T000000" ...
 $ price          : num  221900 538000 180000 604000 510000 ...
 $ bedrooms       : Factor w/ 8 levels "0","1","2","3",..: 4 4 3 5 4 5 4 4 4 4 ...
 $ bathrooms      : Factor w/ 6 levels "0","1","2","3",..: 2 3 2 4 3 5 3 2 2 3 ...
 $ sqft_living    : int  1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...
 $ sqft_lot       : int  5650 7242 10000 5000 8080 101930 6819 9711 7470 6560 ...
 $ floors         : Factor w/ 3 levels "1","2","3": 1 2 1 1 1 1 2 1 1 2 ...
 $ waterfront     : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ view           : Factor w/ 5 levels "0","1","2","3",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ condition      : Factor w/ 5 levels "1","2","3","4",..: 3 3 3 5 3 3 3 3 3 3 ...
 $ grade          : int  7 7 6 7 8 11 7 7 7 7 ...
 $ sqft_above     : int  1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...
 $ sqft_basement  : int  0 400 0 910 0 1530 0 0 730 0 ...
 $ yr_built       : int  1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...
 $ yr_renovated   : int  0 1991 0 0 0 0 0 0 0 0 ...
 $ zipcode        : int  98178 98125 98028 98136 98074 98053 98003 98198 98146 98038 ...
 $ lat            : num  47.5 47.7 47.7 47.5 47.6 ...
 $ long           : num  -122 -122 -122 -122 -122 ...
 $ sqft_living15  : int  1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ...
 $ sqft_lot15     : int  5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 ...
 $ is_renovated   : Factor w/ 2 levels "0","1": 1 2 1 1 1 1 1 1 1 1 ...
 $ house_age      : num  65 29 87 55 33 19 25 57 60 17 ...
 $ type_livingsqft: Factor w/ 2 levels "Only groundfloor",..: 1 2 1 2 1 2 1 1 2 1 ...
 $ Grades         : F
actor w/ 5 levels "Very Low","Moderate",..: 3 3 2 3 4 5 3 3 3 3 ...
```

Questions?

1. Identify the variables that can predict house prices in Seattle. Explain why you think those variables predict house prices.
2. Create a correlation matrix of the identified variables and comment on the correlations.
3. Use regression to predict house prices by looking at the impact of the identified variables on the house prices. Explain how the identified variable impact house prices.
4. Based on your regression analysis, what factors are more important in predicting house prices? Explain.
5. Are there any potential interaction effects? Check for interaction effects and comment.

How do we aim to answer these?

We have used R studio for the analysis and have done a comprehensive cross comparison of models using excel. The following steps where done to answer some particular questions.

1. Reading the dataset to understand from a domain perspective and analyze the variables that may or may not affect house prices.
2. Selecting all continuous variables and performing an Exploratory Data Analysis:
    a. Checking for null values.
    b. Finding out linear relationship using scatterplots.
    c. Checking for correlation between variables which includes
        i. Dependent vs Independent
        ii. Within Independent
    d. Checking variable distribution
    e. Analyzing the categorical variables

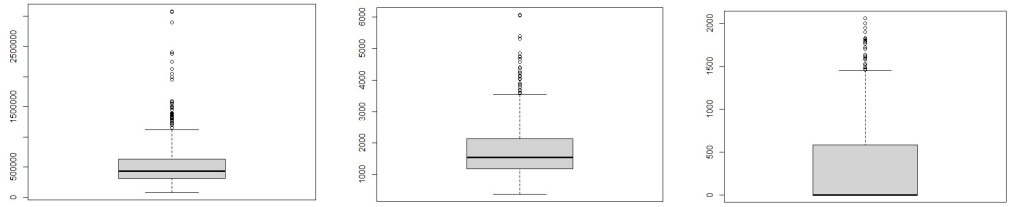
The goal of exploratory data analysis is to obtain confidence in this data to a point where we're ready to engage a machine learning algorithm. Another side benefit of EDA is to refine our selection of feature variables that will be used later for machine learning.

3. Outlier detection and treatment: It is important because the model may not be correctly predicting the value because of the variability caused by outliers.
4. Checking for the concentration of Zeros for Independent variables.
5. Finalizing the data set for multiple regression[Continuous variables].
6. Running multiple iterations and cross-comparing different models based on coefficients, standard error, adjusted R square, VIFs and other properties.
7. Finalizing the model then adding all categorical variables to the final dataset then reiterating to find suitable model.
8. Converting categorical vars into dummy variables.
9. Creating subsets from all variables using forward or backward elimination methods.
10. Checking for log linear model to see the model becomes better.
11. Testing against certain assumptions such as Homoskedasticity and Normality. If the results are not suitable then again eliminating variables or transforming them then running model again.
12. Checking for interaction effects.

OUTLIER TREATMENT

We tried to find out if there are any outliers among the continuous variables. We have first plotted different boxplots to check these outliers.

Some of the boxplots:



As we can see in the plot above there are lot of outliers which need to be treated. All continuous variable had a lot of outliers.

We have used the method of flooring and capping to treat these.
To do achieve this we have distributed the outliers into 100 percentiles and capped top 1% of data points to their 99th percentile value.

Q1. Identify the variables that can predict house prices in Seattle. Explain why you think those variables predict house prices.

From the data we can infer that there might be few variables that have an impact on sales.
Square feet living and Square feet living for the nearest 15 neighbors
sqft_living and sqft_living15: Well it is fairly obvious that a house with more area will cost more compared to a house with less area. People may look for other factors as well such as amenities provides or the availability of a friendly neighborhood so they cut down bit but while looking for a house it is usually the first choice criteria.

Also, the type of houses that are in the neighborhood also drives the house prices. For example, a posh locality will have expensive houses.

Basement area could be important it provides a utility space for people to store necessary items.

Renovation: A newly renovated house has good look and feel, therefore it will drive the price of the house

Ground and multistory: A house with multiple floors will definity be expensive compared to a house which has only ground floor.

Age: Extremely old house will have higher monetary values

Condition: Generally speaking, a house which has a good condition will drive more customers therefore impacting the price.

Bedrooms and Bathrooms: A house more no.of bedrooms or bathrooms will generally be more expensive.

Q2. Create a correlation matrix of the identified variables and comment on the correlations

	price	sqft_living	sqft_lot	sqft_above	sqft_basement	sqft_living15	sqft_lot15
price	1.00	0.70	0.15	0.58	0.37	0.65	0.16
sqft_living	0.70	1.00	0.23	0.86	0.46	0.81	0.26
sqft_lot	0.15	0.23	1.00	0.25	0.01	0.27	0.83
sqft_above	0.58	0.86	0.25	1.00	-0.05	0.76	0.29
sqft_basement	0.37	0.46	0.01	-0.05	1.00	0.25	0.00
sqft_living15	0.65	0.81	0.27	0.76	0.25	1.00	0.28
sqft_lot15	0.16	0.26	0.83	0.29	0.00	0.28	1.00
house_age	-0.07	-0.32	-0.11	-0.41	0.08	-0.34	-0.13

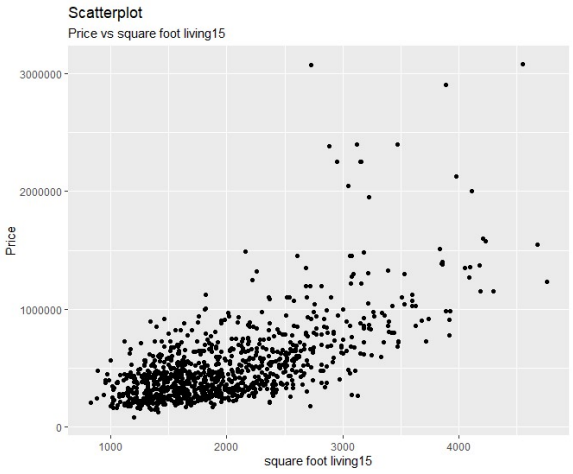
Inference: We can see a very strong correlation between **Price-square feet living** and **Price-square feet living15.** This is in line with our explanation regarding the predictor variables.

Among all Independent variable we also see that some are highly correlated:

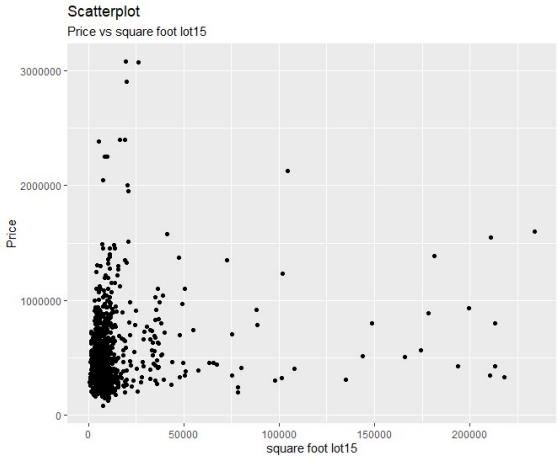
1. **square feet living** and **square feet living of the neighborhood -** We think that if might infer that if the house is spacious then generally speaking, mot houses in the neighborhood will also be spacious.
2. **square feet above** and **square feet living** and **square feet living of the neighborhood**

These correlation between independent variables tell us that there might be a multicollinearity issue when we will try to run regression and fit our model. This inference suggests that almost the same variability is explained by 2-3 variables which will make our model redundant.

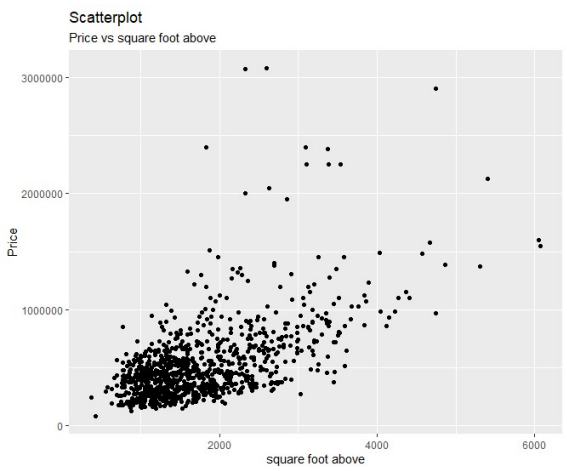
Adding some graphs to showcase the relationships between Price and other variables:



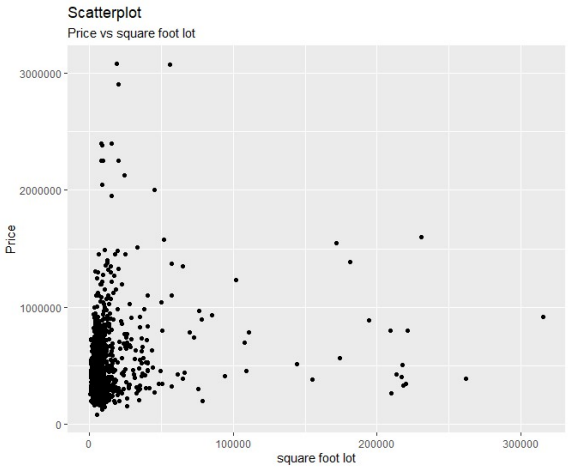
Price vs Square foot living15



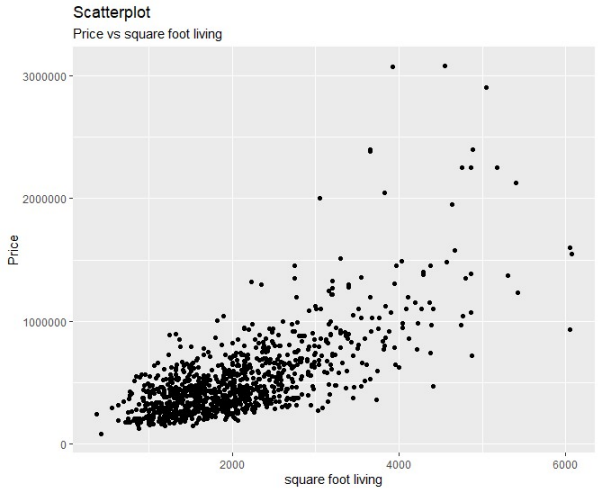
Price vs Square Foot lot



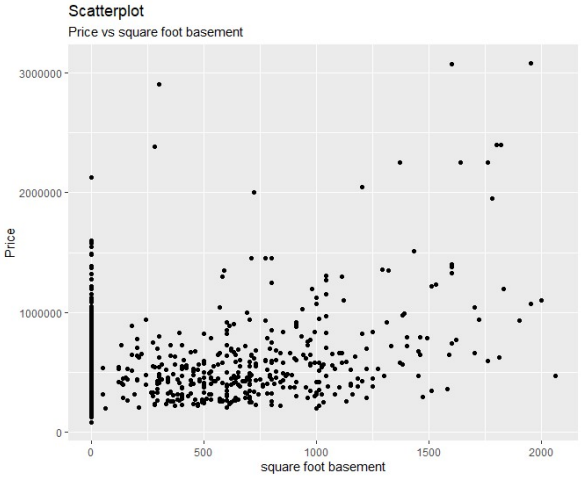
Price vs square foot above



Price vs square foot lot



Price vs square foot living



Price vs square foot basement

Q3 & Q4:

Use regression to predict house prices by looking at the impact of the identified variables on the house prices. Explain how the identified variable impact house prices.

Based on your regression analysis, what factors are more important in predicting house prices? Explain.

ITERATION 1

First, we have done 2 iterations:

1. Model with outliers
2. Model without outlier(treated outliers)

Then we have compared them to see which one is better and used that as benchmark moving forward.

Outliers
Model 1

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-117242	23469.1439	-4.996	6.93E-07	***
sqft_living15	130.8876	19.5251	6.704	3.41E-11	***
sqft_lot15	-0.1729	0.5399	-0.32	0.749	
sqft_basement	244.6436	18.6239	13.136	<	2E-16
sqft_above	177.1696	16.0255	11.055	<	2E-16
sqft_lot	-0.2396	0.4648	-0.516	0.606	
sqft_living	NA	NA	NA	NA	
Multiple R-squared: 0.5212		Adjusted R-squared: 0.5188			

Model 2
Outliers treated

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-89276.3548	21279.8539	-4.195	2.96859E-05	***
sqft_living15_cap	124.9371	17.7477	7.04	3.6E-12	***
sqft_lot15_cap	0.418	0.5781	0.723	0.469858	
sqft_basement_cap	-369.8116	118.8325	-3.112	0.001911	**
sqft_above_cap	-432.9676	120.2657	-3.6	0.000334	***
sqft_lot_cap	-0.5829	0.486	-1.199	0.230692	
sqft_living_cap	599.3017	119.5826	5.012	6.38915E-07	***
Multiple R-squared: 0.5517,		Adjusted R-squared: 0.549			

After both model we saw that model 1 considered sqft_living non-significant but we have assumed that it has to used as from an intuitive sense this variable should impact prices.

In terms of standard error of intercepts and adjusted R square Model 2 seems to be performing better.
Also, Model 1 had very high VIFs

sqft_living15_cap	sqft_lot15_cap	sqft_basement_cap	sqft_above_cap	sqft_lot_cap	sqft_living_cap
3.118829	3.907482	63.559311	189.072415	3.917217	242.714079

We have therefore selected Outlier treatment as we are getting better results.

After this we have multiple no of iteration for different set of combinations variables. We performed multiple iterations and cross comparing models on the basis of adjusted R squares, estimates and standard errors and VIFs as shown in below images. We carried out exhaustive search for creating subsets of variables with gave the best model in terms of all attributes mentioned above.

We also ran a Log-linear model to check if there were some improvements. Model diagnostics were carried out to check for Normality and Homoskedasity.

Model 3 - Running against 4 variables

Variables:
sqft_living15_cap
sqft_basement_cap
sqft_above_cap
sqft_living_cap

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-87301.46	21196.05	-4.119	4.12582E-05	***
sqft_living15_cap	121.52	17.51	6.939	7.09E-12	***
sqft_basement_cap	-352.25	118.07	-2.983	0.002921	**
sqft_above_cap	-415.3	119.48	-3.476	0.000531	***
sqft_living_cap	582.6	118.89	4.9	1.1175E-06	***

Multiple R-squared: 0.5509 Adjusted R-squared: 0.5491
F-statistic: 304.8 p-value: < 0.00000000000000022

VIF

sqft_living15_cap	sqft_basement_cap	sqft_above_cap	sqft_living_cap
3.03689	62.7632	186.66369	239.98316

Model 4 - Running against 5 variables

Variables:
sqft_living15_cap
sqft_lot15_cap
sqft_lot_cap
sqft_basement_cap
sqft_above_cap

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-101864.3415	21386.1547	-4.763	2.19019E-06	***
sqft_living15_cap	134.0563	17.8673	7.503	1.38E-13	***
sqft_lot15_cap	0.2266	0.5838	0.388	0.698	
sqft_lot_cap	-0.3394	0.4894	-0.694	0.488	
sqft_basement_cap	219.7938	16.9399	12.975	<	2E-16
sqft_above_cap	165.2085	14.9266	11.068	<	2E-16

Multiple R-squared: 0.5403 Adjusted R-squared: 0.538
F-statistic: 244.2 p-value: < 0.00000000000000022

VIF

sqft_living15_cap	sqft_lot15_cap	sqft_lot_cap	sqft_basement_cap	sqft_above_cap
3.086044	3.890428	3.878087	1.260989	2.843451

Model 5 - Running against 5 variables keeping log of price

Variables:
sqft_living15_cap
sqft_lot15_cap
sqft_lot_cap
sqft_basement_cap
sqft_above_cap

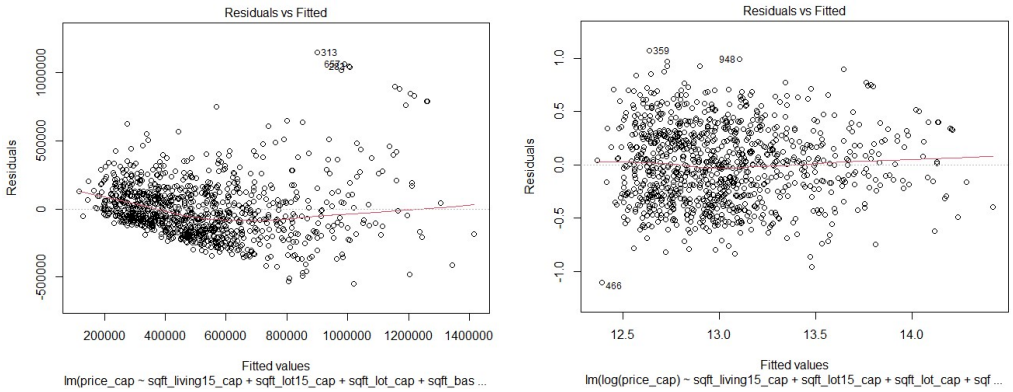
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	12.01559808	0.035869938	334.977	<	2E-16
sqft_living15_cap	0.000221712	2.99679E-05	7.398	2.93E-13	***
sqft_lot15_cap	-3.558E-07	9.792E-07	-0.363	0.716	
sqft_lot_cap	-2.655E-07	8.209E-07	-0.323	0.746	
sqft_basement_cap	0.000340223	2.84125E-05	11.974	<	2E-16
sqft_above_cap	0.000266379	2.50357E-05	10.64	<	2E-16

Multiple R-squared: 0.5182 Adjusted R-squared: 0.5158
F-statistic: 213.6 p-value: < 0.00000000000000022

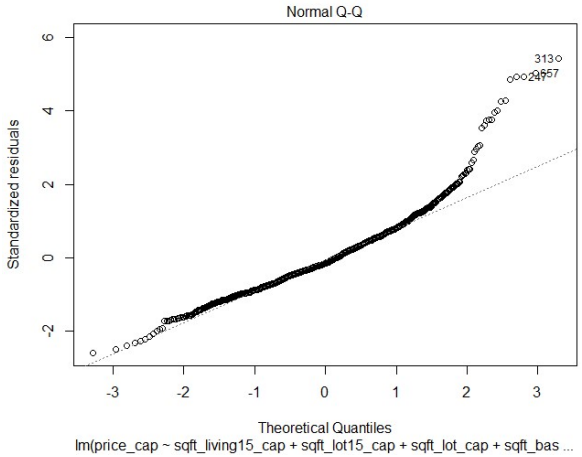
VIF

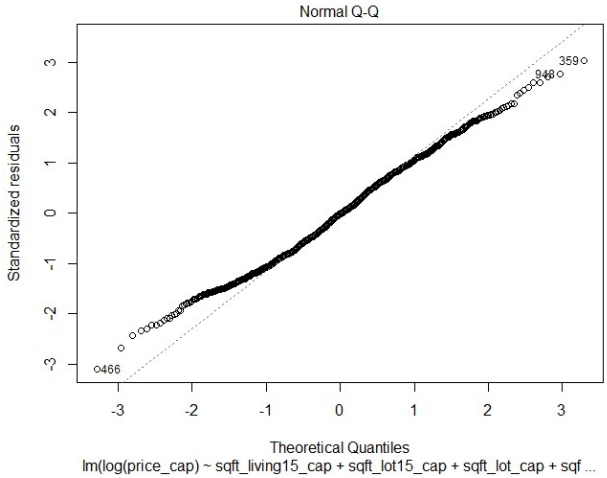
sqft_living15_cap	sqft_lot15_cap	sqft_lot_cap	sqft_basement_cap	sqft_above_cap
3.086044	3.890428	3.878087	1.260989	2.843451

Model Diagnostics:



It is clear from the image that we don't have homoskedasity for non-log model.
After doing Log-linear transformation it is not completely homoscedastic but very much close to it.
In terms of Normality we have observed that both models do not meet the conditions. This was verified by plotting graphs.



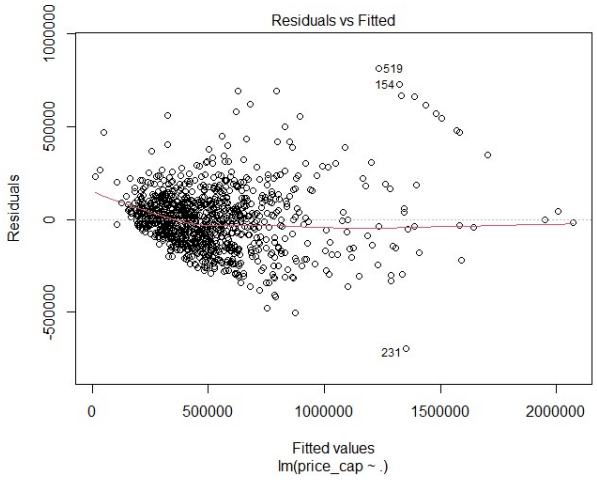


At this point it is clear that although we have resolved hemoskedasity by taking a log-linear model, we couldn't reduce the normality issue, therefore we decided to introduce all categorical variables. By doing this we also reduced the chance of committing an omitted variable bias.

Categorical + Continuous

Firstly, we converted categorical variables into factors. We can't iterate through factor variables therefore we have to convert them into dummy variables then we have added all continuous variables then ran our regression.

Inference: Adjusted R square: 0.7088
Although we are getting a high adj R square, but we cant select this model there are lot of variables now. Since there are lot of variables now, we will a lot of difficulty in controlling them.



Also, this model is violating the homoskedacity assumption.

We have used a package called Boruta to get important variables. Then again we did sub setting these variables. After doing multiple iterations we selected our model and did a Log-linear transformation to check for other improvements.

We tried to fix the homoskedacity of using weighted-least square regression meathod. It resulted in a higher adjusted R square also.

Comparing our Models and our diagnostics:

Variables	Adj R squa	BIC
1	0.521741	-724.053
2	0.570773	-826.208
3	0.59475	-877.729
4	0.614661	-922.156
5	0.637109	-976.217
6	0.649619	-1005.36
7	0.656556	-1019.44
8	0.661279	-1027.38
9	0.665607	-1034.32
10	0.670353	-1042.71
11	0.673586	-1046.66
12	0.676605	-1050.05
13	0.678266	-1049.3
14	0.679754	-1048.04
15	0.681553	-1047.78
16	0.682905	-1046.14
17	0.684374	-1044.88
18	0.685642	-1043.02
19	0.686797	-1040.81

The above image shows a subset of 19 variables sorted by their Adj R square values and BIC values.
Based on that we have run multiple regression models.

We have taken variable subsets till the point when Adj R values increases through a small margin only.

So we have taken 12 to 15 variables and ran iterations on it.

After multiple iterations we have selected a model with 12 variables.
We have fixed the normality and homoskedacity using weighted least square method.

Weighted least square regression against 12 Variables with log of price	Estimate	Std. Error	t value	Pr(>|t|)				VIF
is_renovated1	(Intercept)	11.91781	0.053229	223.896	<	2E-16	***	is_renovated	1.051811
sqft_living_cap	is_renovated1	0.355474	0.066109	5.377	9.45203E-08	***		sqft_living_ca	4.832855
sqft_living15_cap	sqft_living_cap	0.000238	2.11E-05	11.306	<	2E-16	***	sqft_living15	3.761121
house_age	sqft_living15_c	0.000265	2.53E-05	10.475	<	2E-16	***	house_age	1.605453
bedrooms_2	house_age	0.005338	0.00044	12.137	<	2E-16	***	bathrooms_4	1.643994
bathrooms_4	bathrooms_4	-0.00503	0.048228	-0.104	0.917			floors_1	1.765183
floors_1	floors_1	-0.20256	0.02636	-7.685	3.7E-14	***		bedrooms_2	1.11791
view_1	bedrooms_2	0.067557	0.040193	1.681	0.0931	.		view_1	1.280091
view_2	view_1	0.220677	0.053308	4.14	3.77536E-05	***		view_2	1.054081
condition_2	view_2	0.019478	0.052154	0.373	0.7089			condition_2	1.004901
`type_livingsqft_Only groundfloor`	condition_2	-0.45648	0.202841	-2.25	0.0246	*		`type_livingsq	1.382046
waterfront_1	`type_livingsqf	-0.12757	0.022966	-5.555	3.57415E-08		***	waterfront_1	1.020306
	waterfront_1	0.524363	0.117751	4.453	9.42884E-06	***			
	Adjusted R-squ	0.7045							

Variables selected:

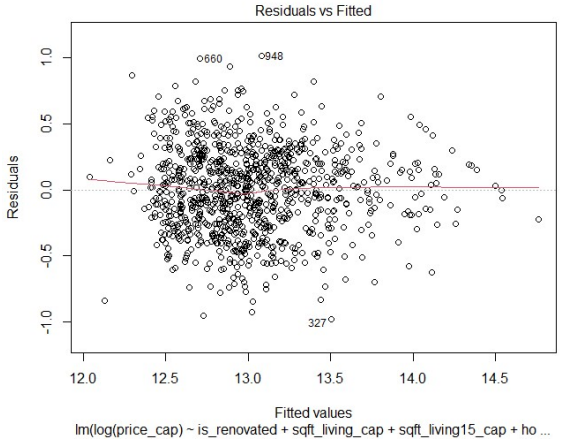
is_renovated1

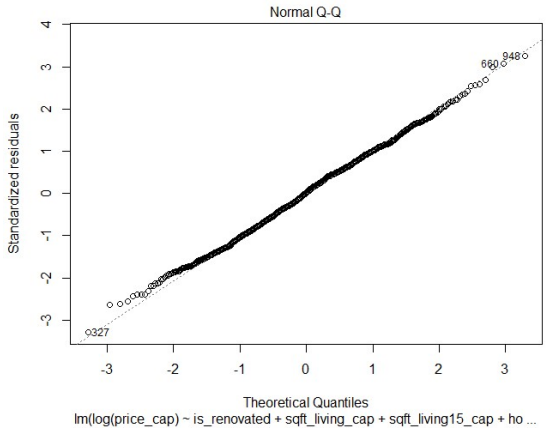
sqft_living_cap

sqft_living15_cap
house_age
bedrooms_2
bathrooms_4
floors_1
view_1
view_2
condition_2
`type_livingsqft_Only groundfloor`
waterfront_1

As you can see in the image, the VIF looks good.
Adj R square = 0.7045 which is good as well.

Diagnostics:



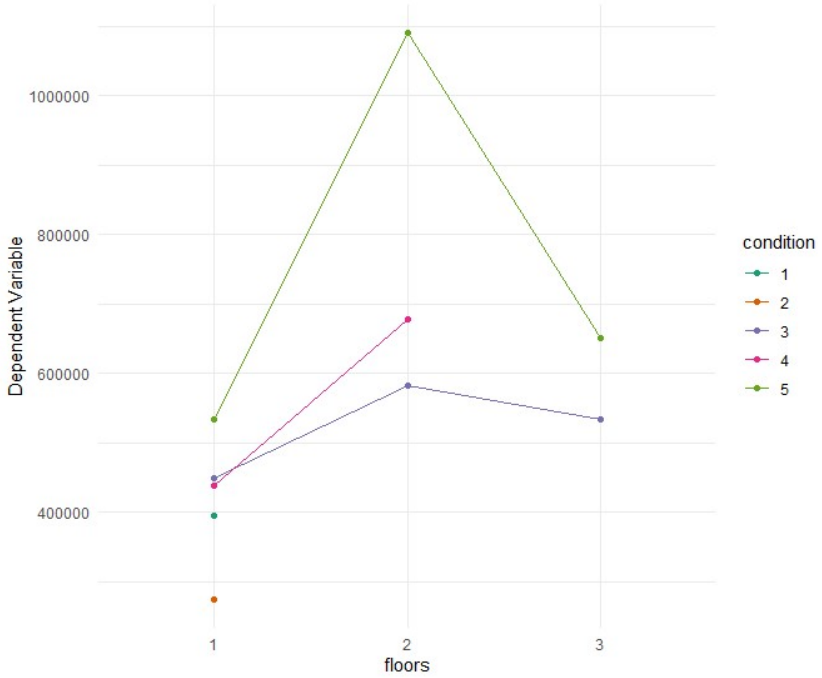


Thus, we have resolved our assumptions also.

This will be our final model for predicting house prices in Seattle.

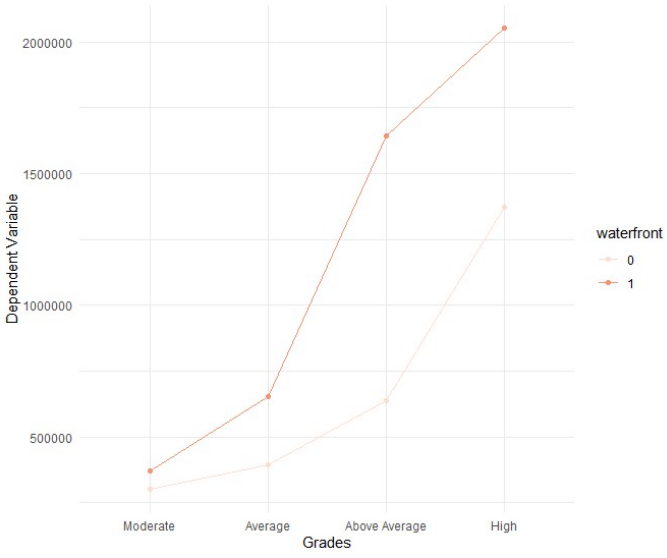
Q5. Are there any potential interaction effects? Check for interaction effects and comment.

Let us check for interaction between Floors and Condition:



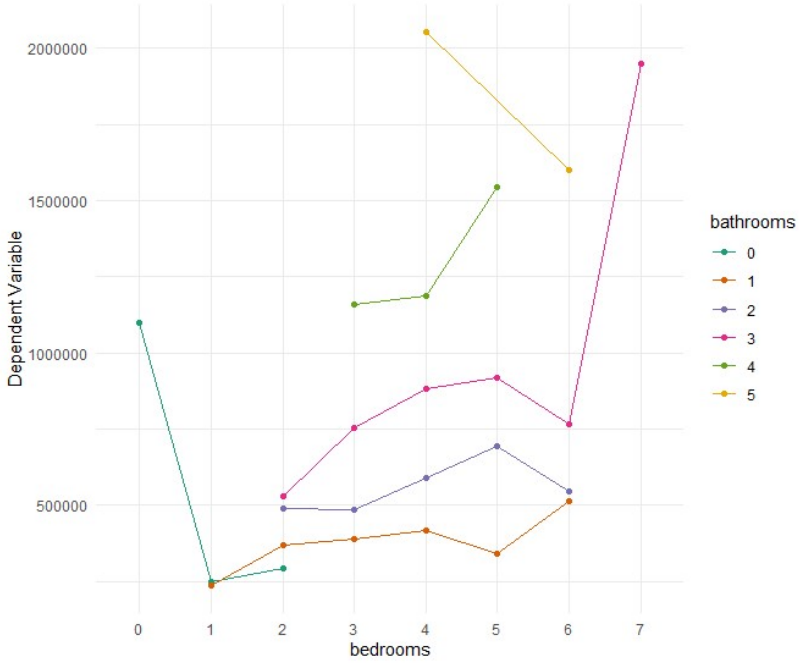
As seen from the fig. there is no interaction between Floors and Condition.

Let us check for interaction between Grades and waterfront:



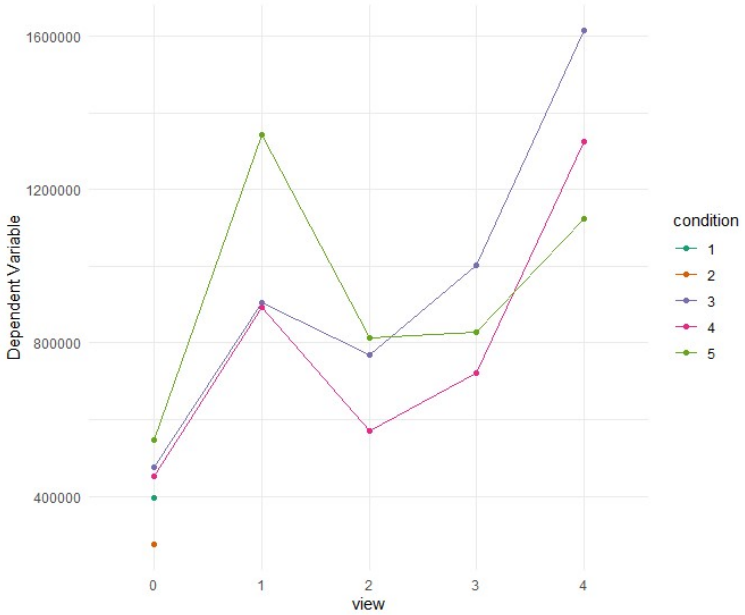
As seen from the fig. there is no interaction between Grad level and waterfront.

Let us check for interaction between bedroom and bathroom:



No interaction between bedroom and bathroom.

Let us check for interaction between view and condition:



There is interaction between View and condition.