Data Analytics Using Excel

Project – House Price Prediction

Report by Saurabh Tayal

Problem Statement:

In this project, we are going to look at a number of houses sold in the year 2016 and 2017 in a fictional state by a well-known real estate agency. The agency has trained auditors who measure and map all the relevant features for the properties along with information related to the geography around it. The agency wants to understand the relevance of the parameters that they collect in relation to the price of the house. They have hired you to create a model which makes use of the available information to predict the monetary value of a house.

You are expected to use the data of the year 2016 to create a regression model where the price is the dependent variable. Identify the factors that are the driving factors for house prices. Using the model, you are expected to predict the selling prices of the houses sold in 2017.

Deliverables:

• Create an excel report that contains all the meaningful information such as relevant charts, pivot tables etc.

• Create a few hypotheses around the important variables and validate them using the data

• Mention all the variable which are highly correlated

• Build a linear regression model on the data of year 2016. Predict the price for year 2016 using this regression model, plot the regressed values against the actual values to understand the difference

• Using the above linear regression model, predict the prices of the houses sold in the year 2017. Interpret your findings from the model.

Data Dictionary:

1. **Id**: unique id
2. **Date**: Date house was sold
3. **Price**: Price of the sold house (Target Variable)
4. **Bedrooms**: Number of Bedrooms
5. **Bathrooms**: Number of bathrooms

6. **Living area**: Square footage of the living space
7. **Lot area**: Square footage of the lot
8. **Number of floors**: Total floors in the house
9. **Waterfront**: Whether the house is on a waterfront (1: yes, 0: no)
10. **Number of views**: number of special views
11. **Condition**: Condition of the house on a scale of 1-5 (1 being the lowest, 5 being the highest)
12. **Grade of the house**: Grade of the house based on Foundation, Drainage and Fire Prevention on a scale of 1-13 (1 being the lowest and 13 being the highest)
13. **Area of the house**: Square footage of house apart from basement
14. **Area of Basement**: Square footage of the basement
15. **Built year**: Built year
16. **Renovation year**: Year when the house was renovated
17. **Postal code**: Postal code of the house
18. **Living_area_renov**: Living room area currently (after renovations)
19. **Lot_area_renov**: Lot area currently (after renovations)
20. **Number of Schools nearby**: Number of schools in the vicinity of the house
21. **Distance from the airport**: Distance in KMs from nearest Airport

Descriptive information of the data provided using appropriate plots for the variables along with the insights:

1. Price

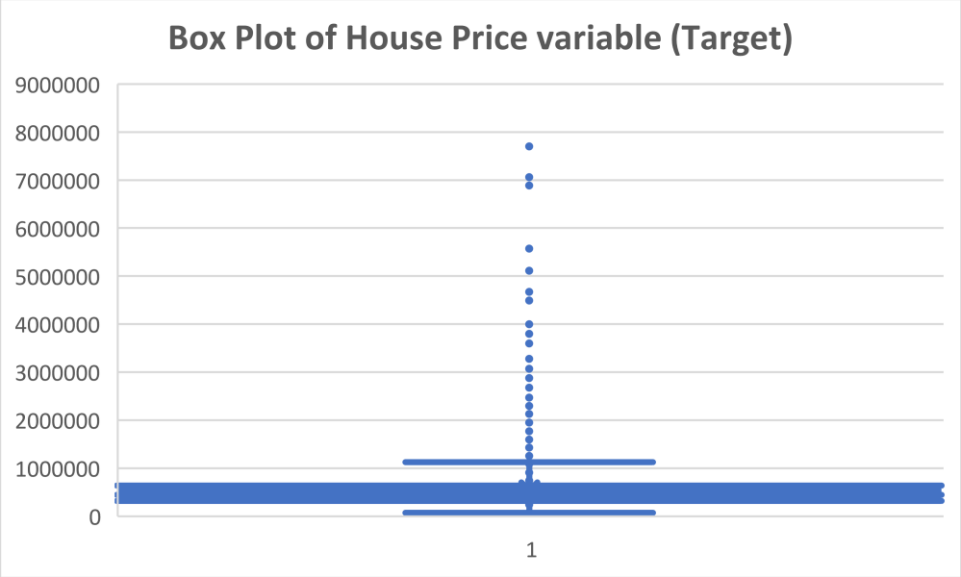
Price	
Mean	538932.2183
Standard Error	3039.638394
Median	450000
Mode	450000
Standard Deviation	367532.3808
Sample Variance	1.3508E+11
Kurtosis	40.32191815
Skewness	4.269297721
Range	7622000
Minimum	78000
Maximum	7700000
Sum	7879189032
Count	14620

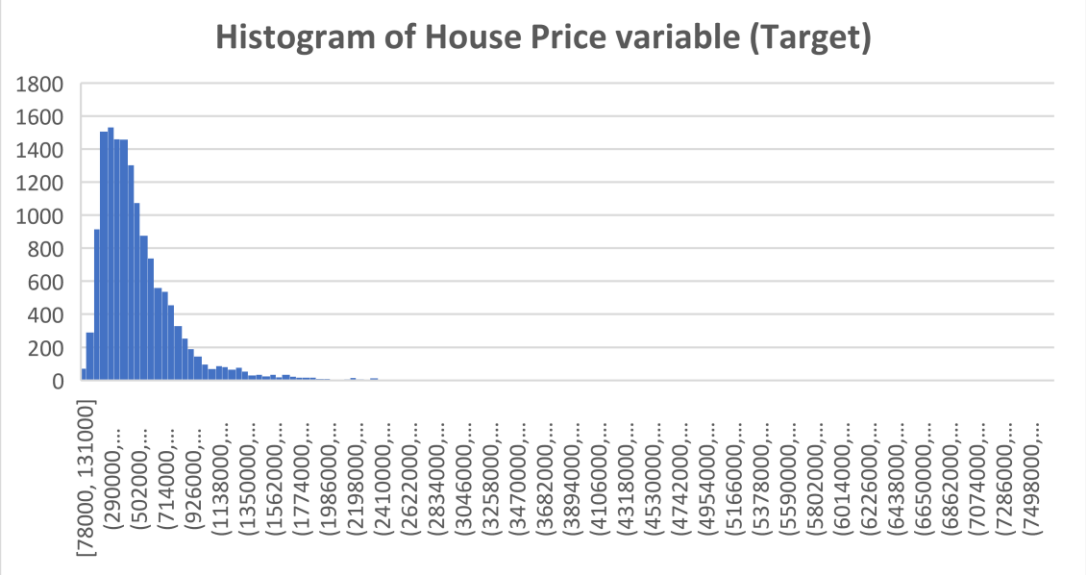
Observations:

1. We can see that there are a lot of outliers on the higher value side of the data.

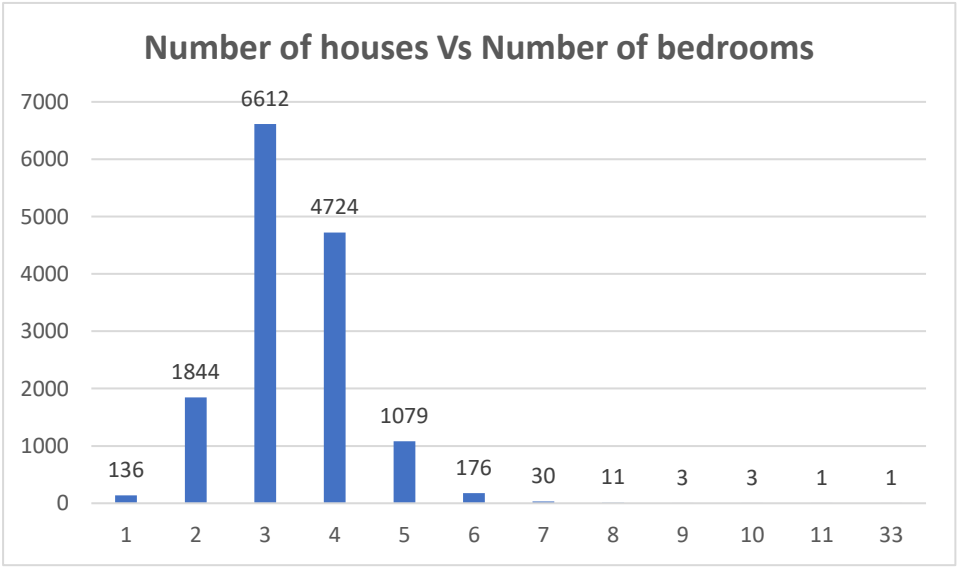
2. Due to these outliers, the data is highly skewed towards the right.

3. There is very high standard deviation.





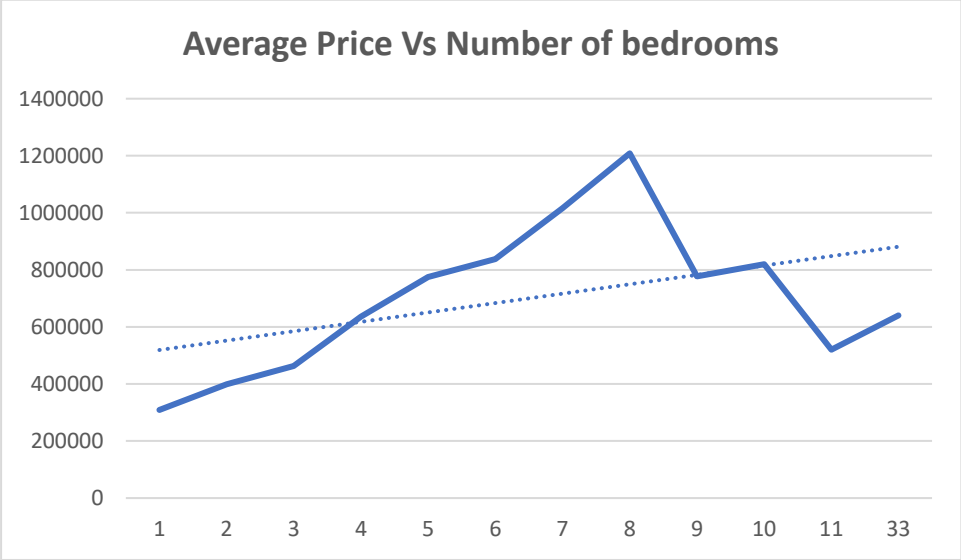
2. Number of Bedrooms



Observations:

1. As we can see maximum number of houses are with 3 bedrooms.

2. The plot is right-skewed.

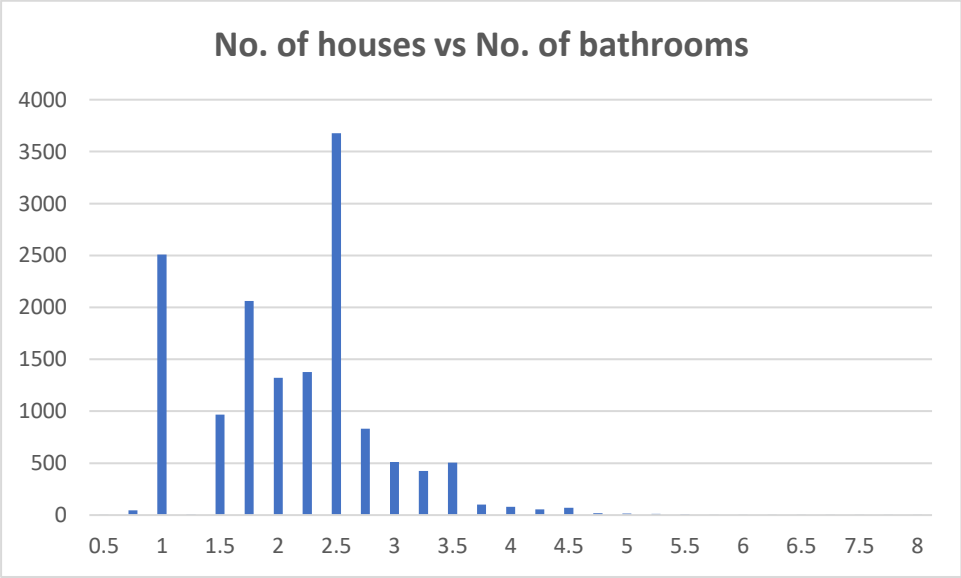


Observations:

1. The maximum average price is for 8 bedrooms house.

2. As the no. of bedrooms increase, price of house also increases, but starts to decrease after 8 bedrooms are reached.

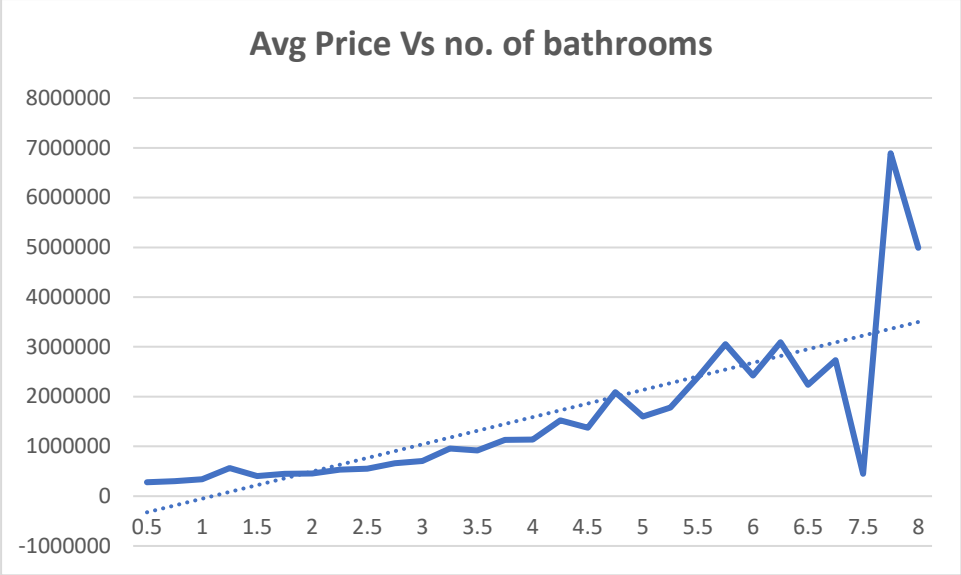
3. Number of Bedrooms



Observations:

1. Max number of houses have 2.5 no. of bathrooms, followed by 1 and 1.5.

2. Data is not normal.



Observations:

1. As the no. of bathrooms increase, the house price also increases.

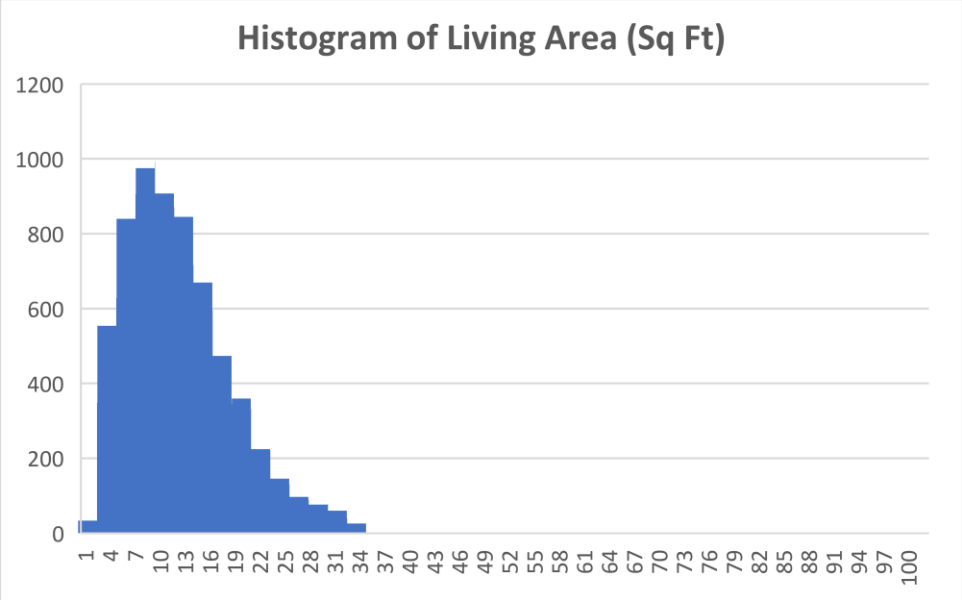
2. A somewhat linear trend can be seen, but there's a sharp drop at 7.5 bathrooms which picks up quite strongly at 7.75 indicating that there's an anomaly in the observed data.

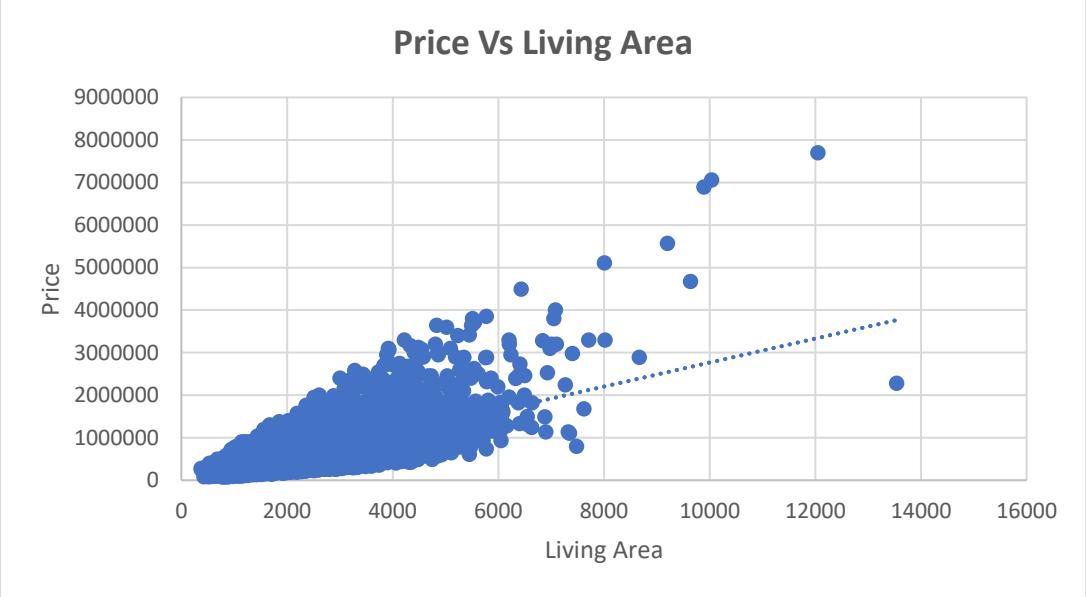
4. Living Area

living area	
Mean	2098.262996
Standard Error	7.677207969
Median	1930
Mode	1400
Standard Deviation	928.2757212
Sample Variance	861695.8146
Kurtosis	6.073617146
Skewness	1.538336624
Range	13170
Minimum	370
Maximum	13540
Sum	30676605
Count	14620

Observations:

1. The Living area distribution is highly skewed towards right, showing that there are a lot of outliers with quite a high value.





Observations:

1. We can see a general linear trend that as the living area increases, the price of the house increases.

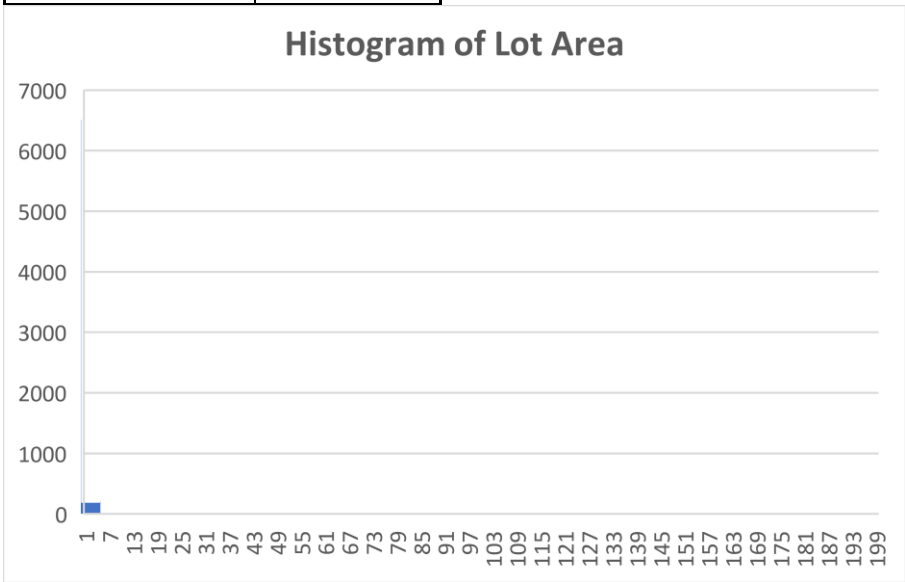
2. Maximum number of houses have an area between 2000-6000 square ft.

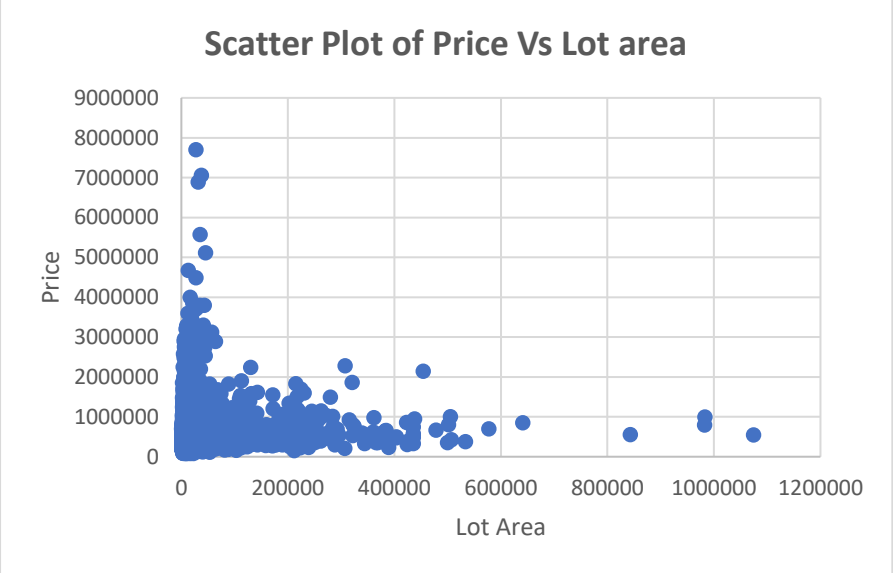
3. There are some outliers as well due to which the data is skewed.

5. Lot Area

lot area	
Mean	15093.28112
Standard Error	313.6102908

Median	7620
Mode	5000
Standard Deviation	37919.6213
Sample Variance	1437897680
Kurtosis	164.7572734
Skewness	10.15520609
Range	1073698
Minimum	520
Maximum	1074218
Sum	220663770
Count	14620





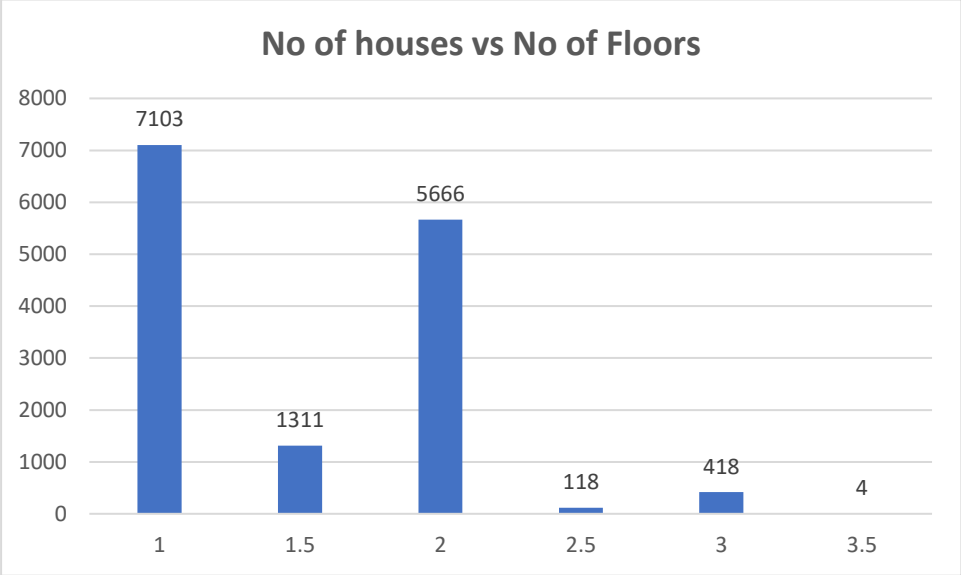
Observations:

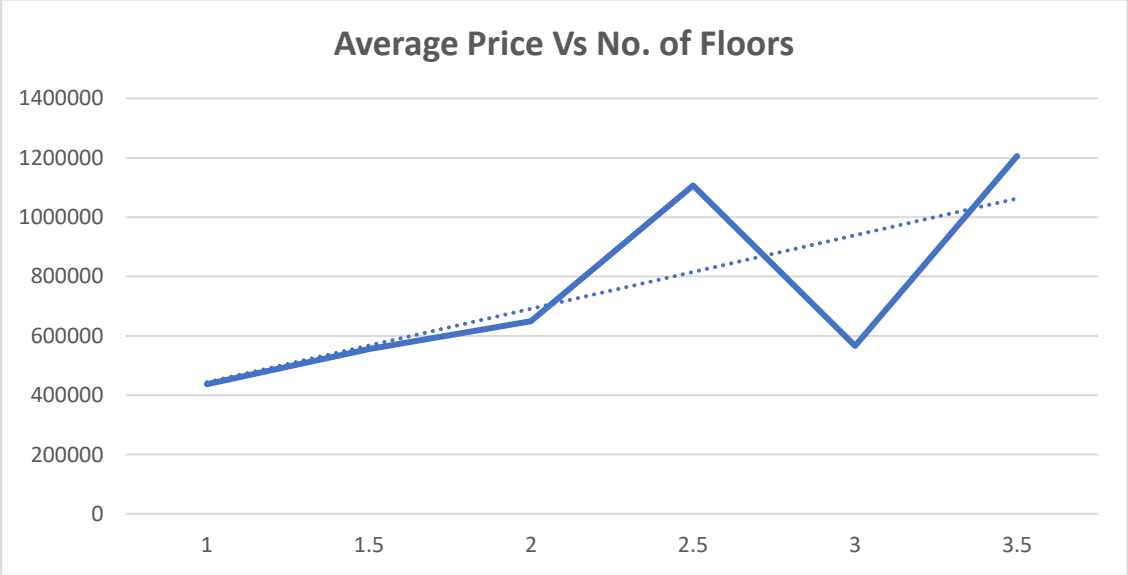
1. The data is heavily skewed towards the right.

2. The standard deviation is quite high, greater than the mean.

3. Most number of houses have a lot area of 5000 square ft.
4. We can see no relationship between Lot are and Price of the houses.

6. Number of Floors

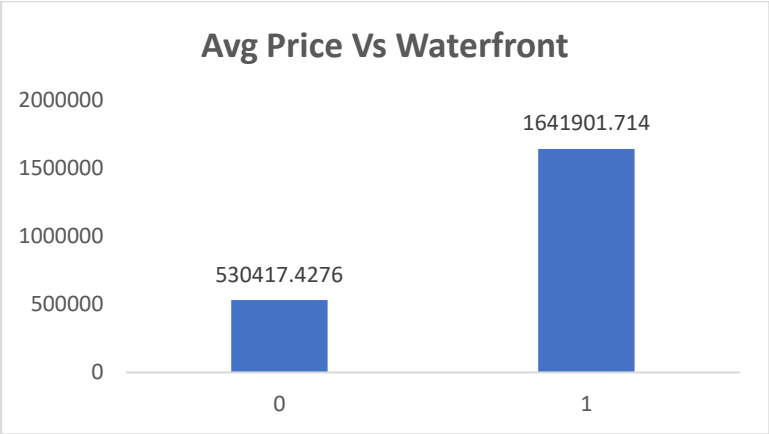




<u>Observations:</u>
1. Max no. of houses have 1 floor followed by 2 floors.
2. As a general trend we can see that as no. of floors increase, the price of house also increases.
3. But there is an exception with 3 floors whose average price drops as compared to houses with 2.5 floors indicating that observed data can be faulty.

7. Waterfront

Waterfront	No of houses
0	14508
1	112
Grand Total	**14620**



Observations:

1. As we can see, almost every house has no Waterfront, except 112 houses out of 14620 houses.

2. And the average price of the house with waterfront is much higher than those without a waterfront.

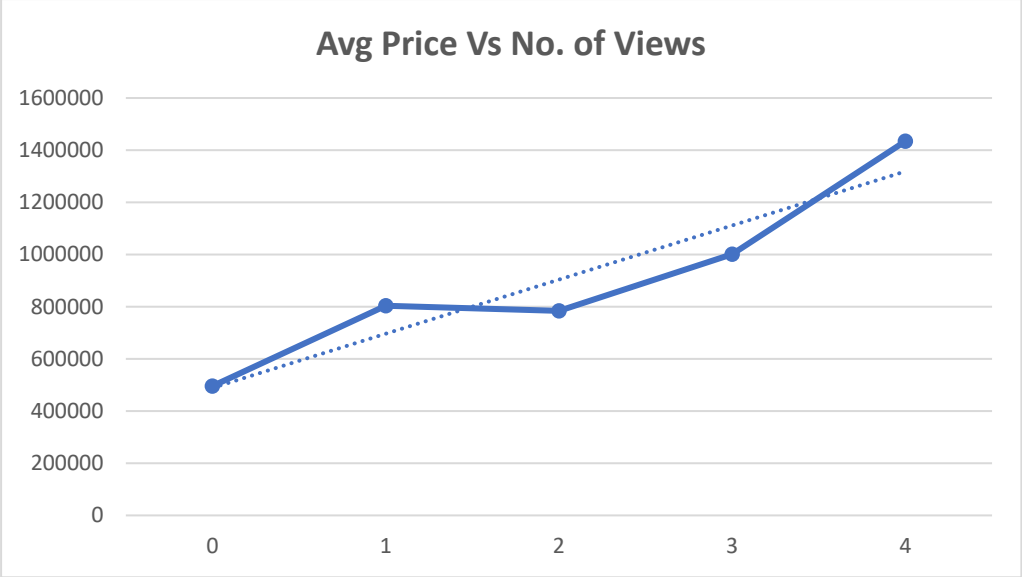
8. Number of views

No. of views	No of houses
0	13198
1	219
2	636
3	351
4	216
Grand Total	**14620**

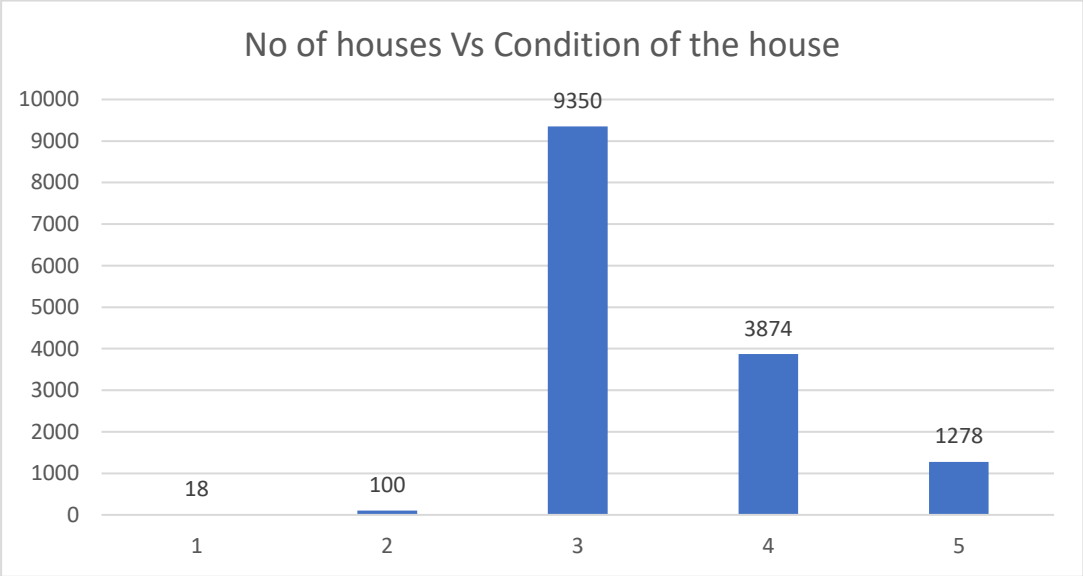
Observations:

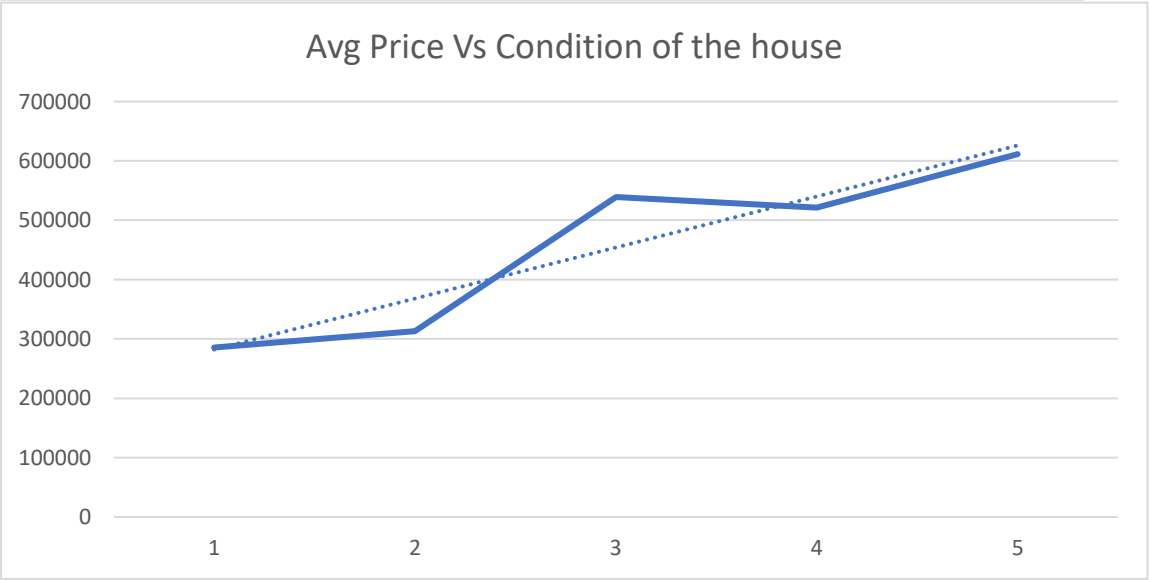
1. Maximum number of houses have 0 views.

2. We can see that as the no. of views increase, the price of the house also increase.



9. Condition of the House





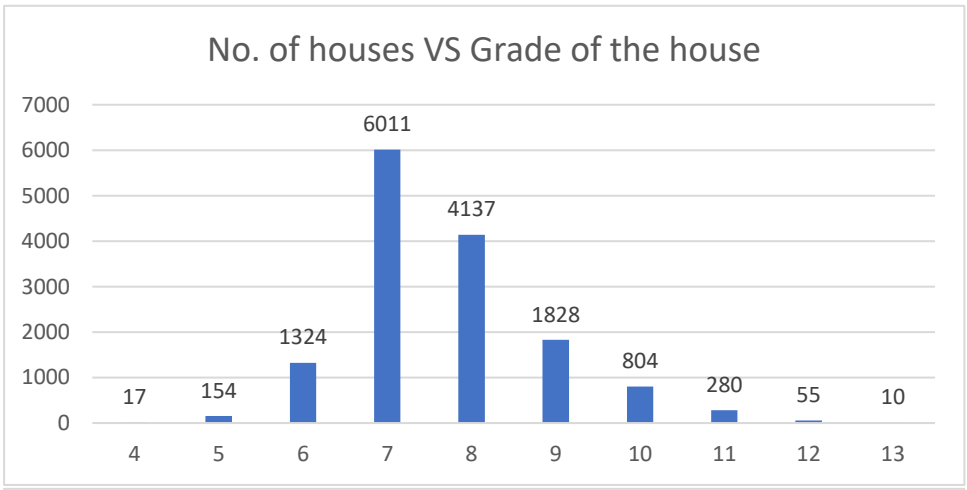
Observations:

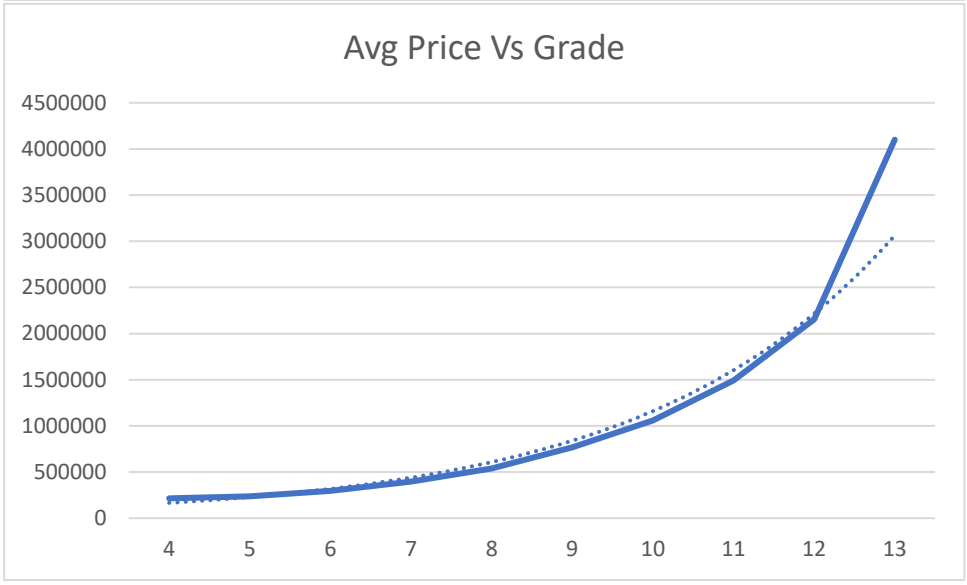
1. Max no. of houses have a condition rating of 3 followed by 4 and then by 5.

2. There is a general linear trend b/w Avg Price and Condition of the house.

3. As the condition of the house is improved, its price is also increased.

10. Grade of the House





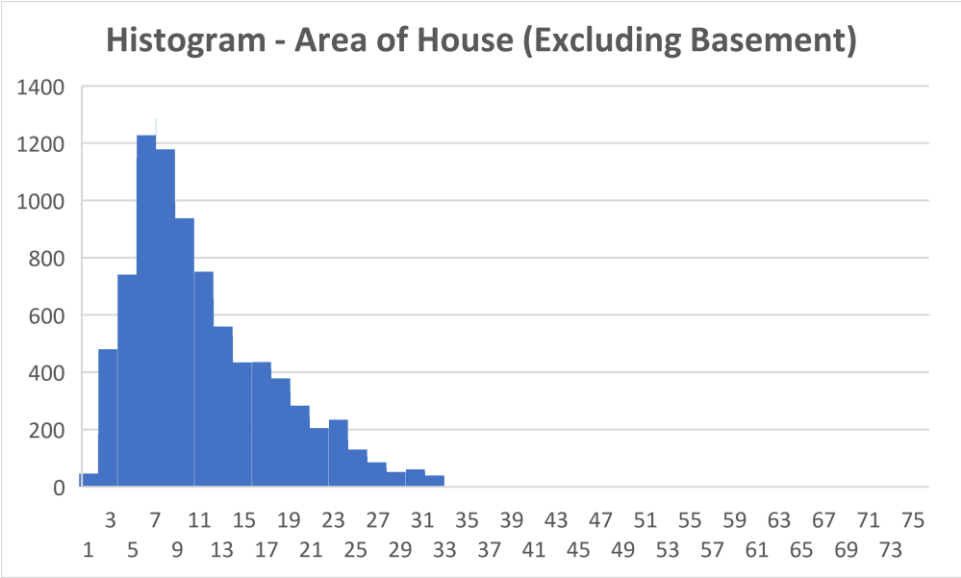
Observations:

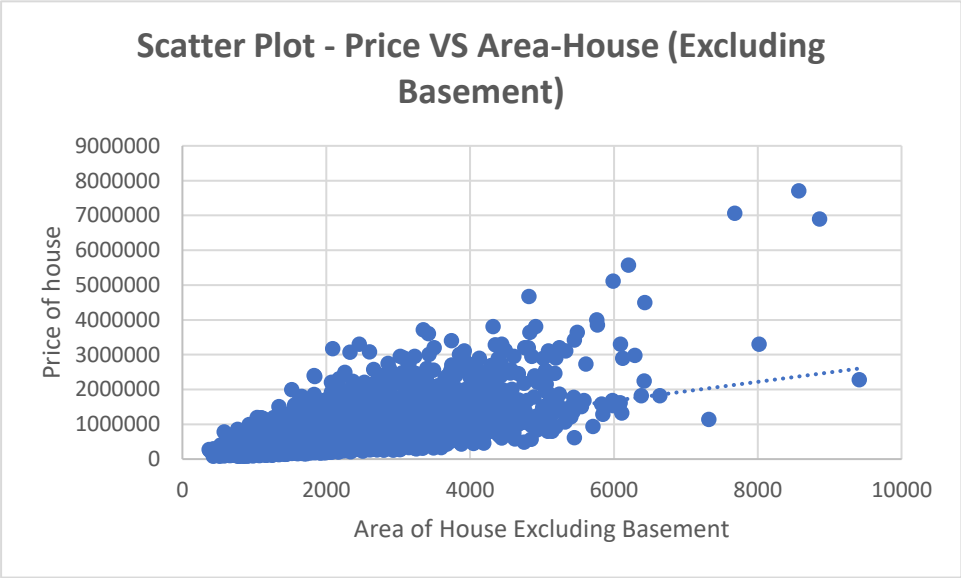
1. Maximum number of houses are with a grade of 7 followed by grade 8, 9 and 3.

2. The Price of the house increases exponentially with respect to Grade of house.

3. It is worth noticing that the price increases sharply after grade 12.

11. Area of the House (Excluding Basement)



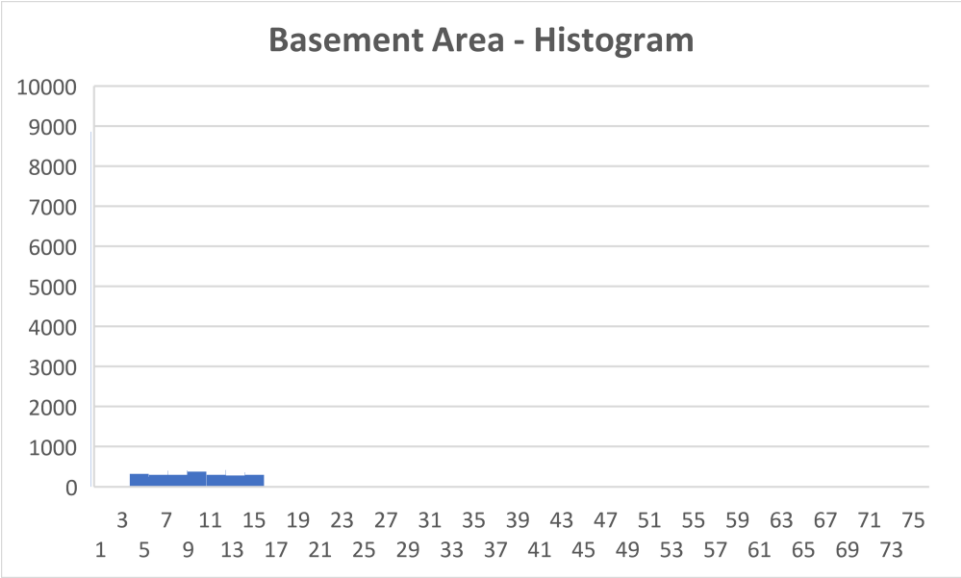


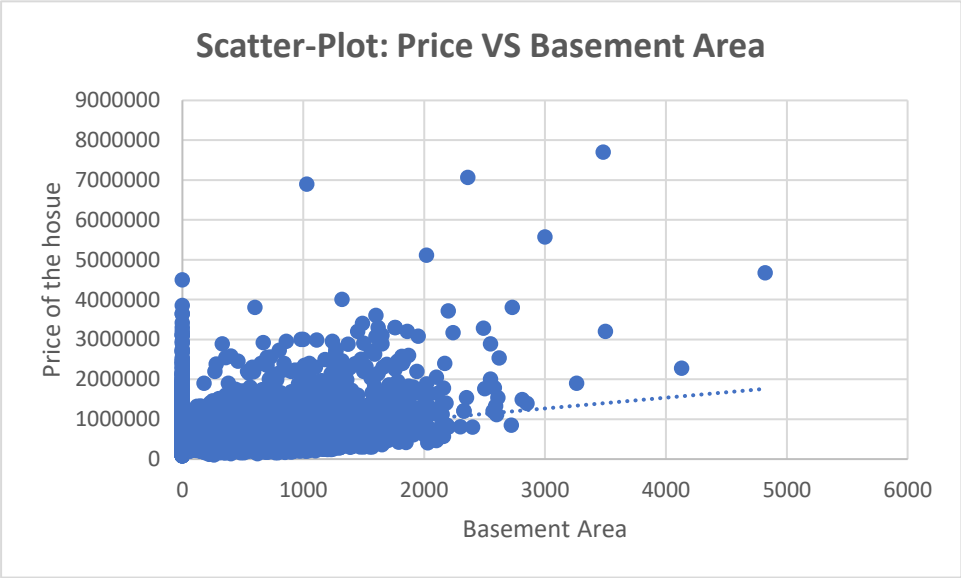
Observations:

1. The data is skewed towards the right showing the presence of outliers having higher values.

2. We can see a general linear trend that as the area of house increases, the price of the house increases.

12. Basement Area





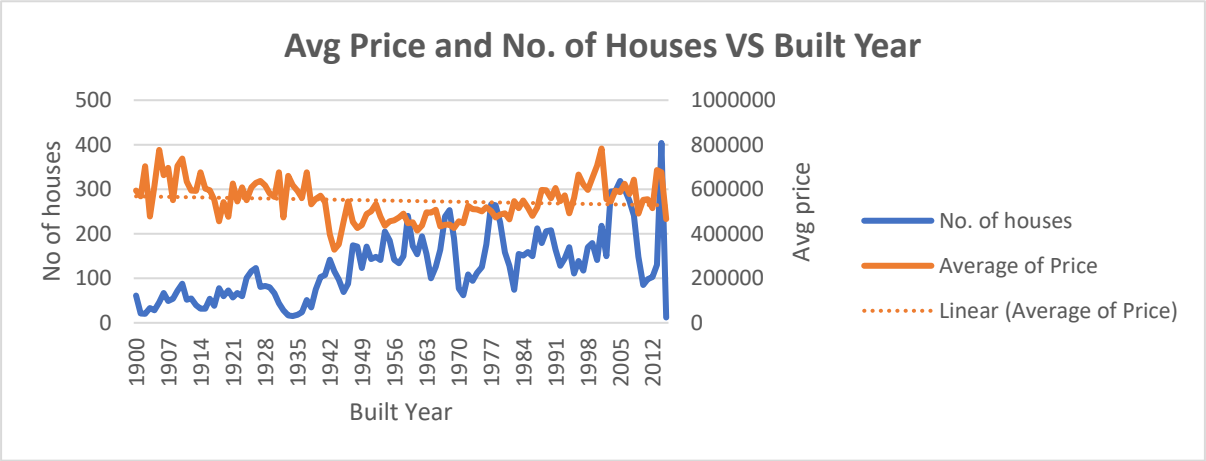
<u>Observations:</u>

1. We can see a lot of datapoints at 0 indicating that most of the houses have no basement.

2. In houses having no basement, there is lot of variation in price on account of other variables we can say.

3. Excluding them we can see that there is general linear trend between the basement area and the house price, as basement area increases house price also increases.

13. Built Year



Observations:

1. A lot of ups and downs are there in terms of House Price w.r.t the built year.

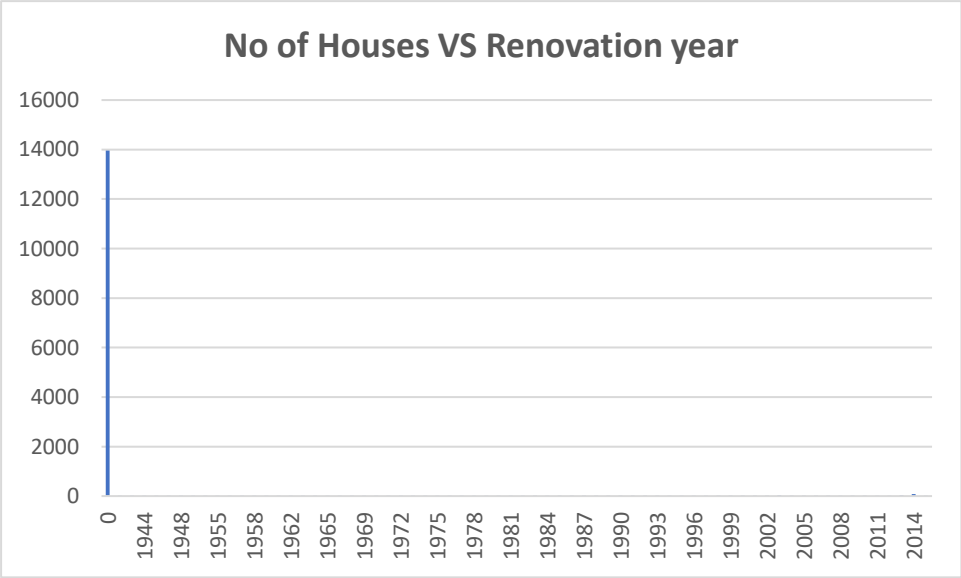
2. On an average the house prices are stagnant through the years.

3. As a general linear trend, the prices are somewhat decreasing as the built year is increasing.

4. There are a few years in which the no. of houses built has dropped drastically.

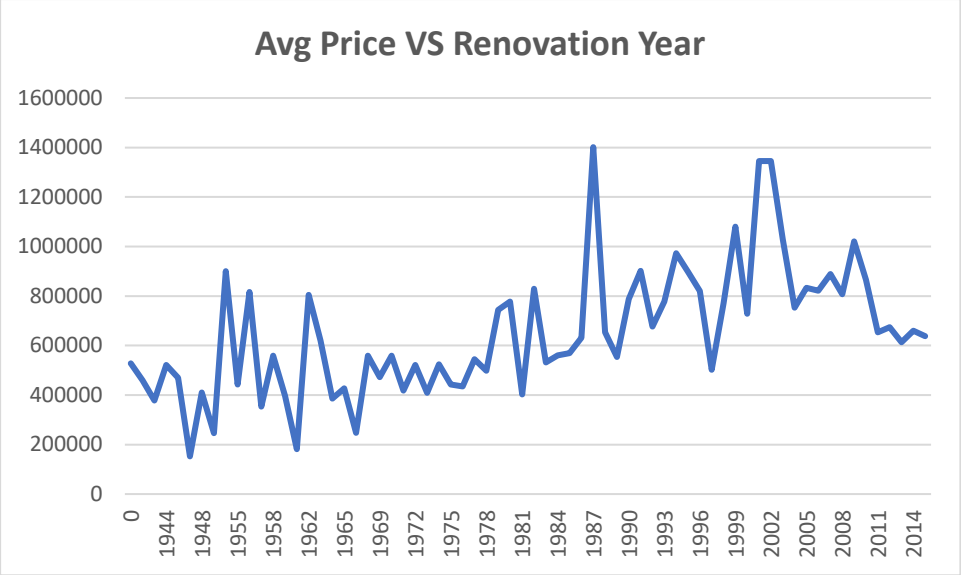
5. For example, in the time of World War 1 & 2, Indo-Pak/China war, 2008 Recession.

14. Renovation Year



Observations:

1. Most of the houses are not renovated.

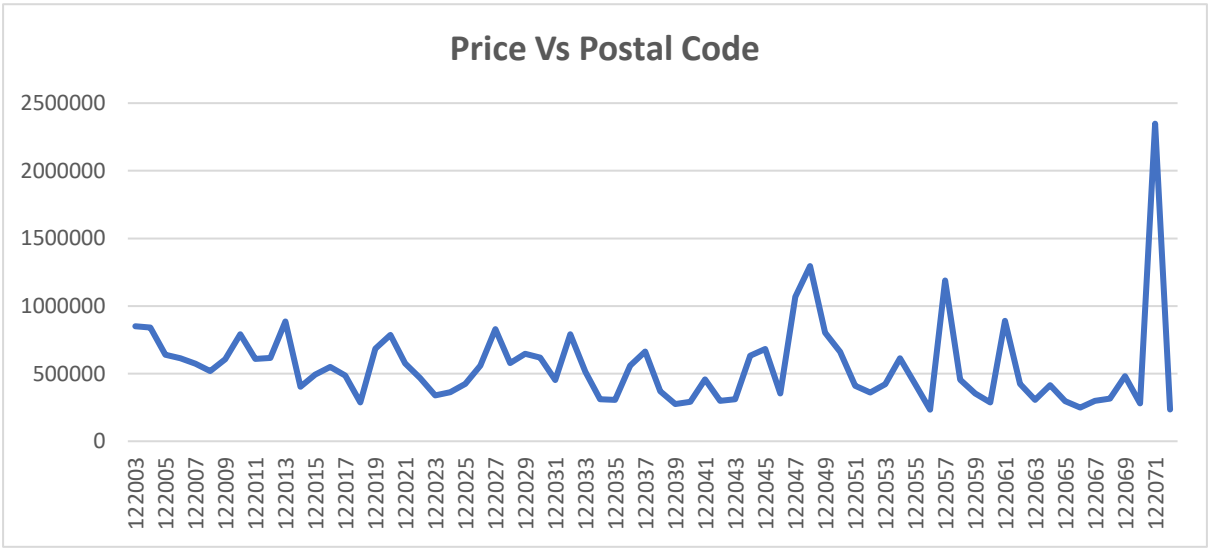


Observations:

1. The houses renovated between 1987-88 and 2002-03 have the highest average price.

2. Apart from that a general linear trend can be seen where the average price of house increases as the renovation year becomes more recent.

15. Postal Code



Observations:

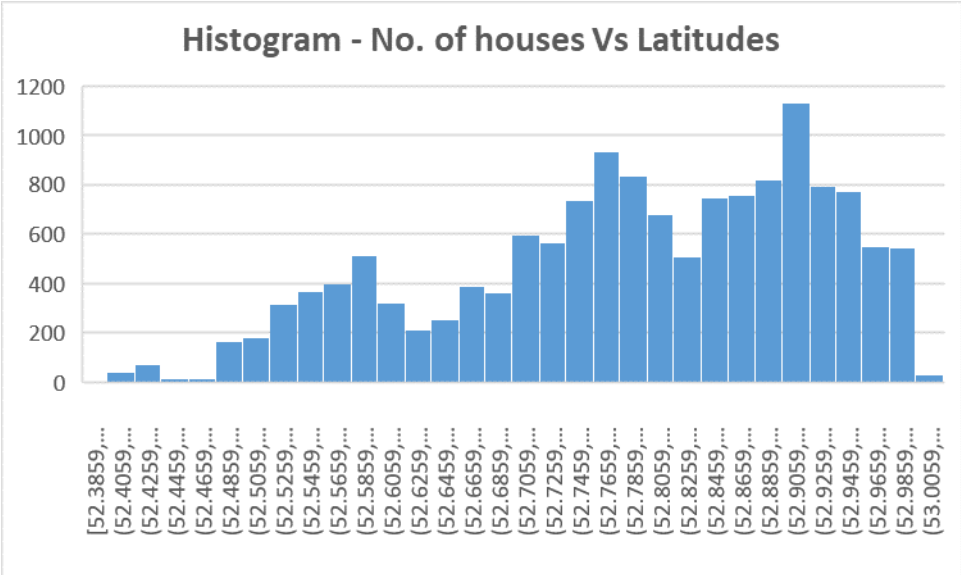
1. Average Price of houses in area with postal codes given below are higher as compared to other areas:

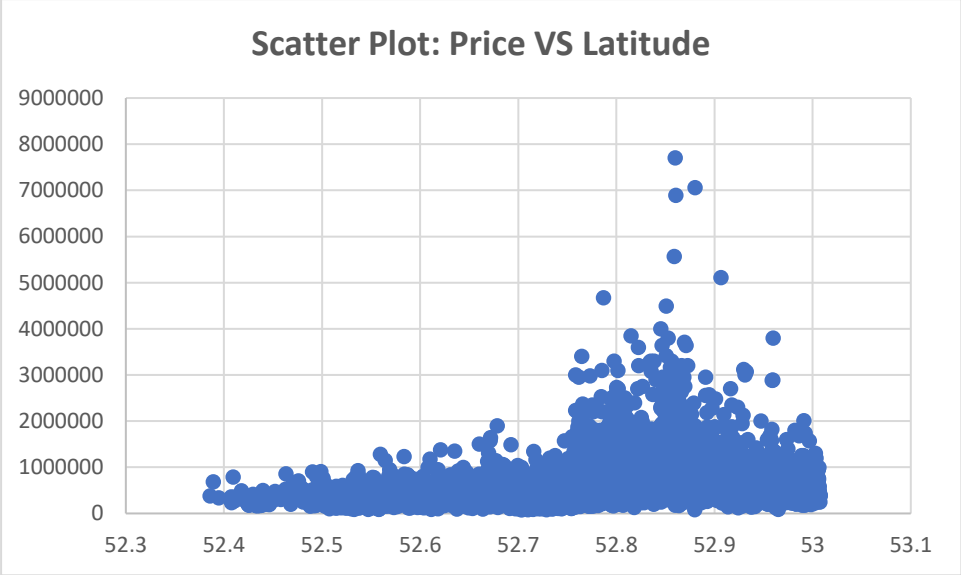
a) 122047 – 122049

b) 122057 - 122058

c) 122071 – 122072

16. Latitude





<u>Observations:</u>

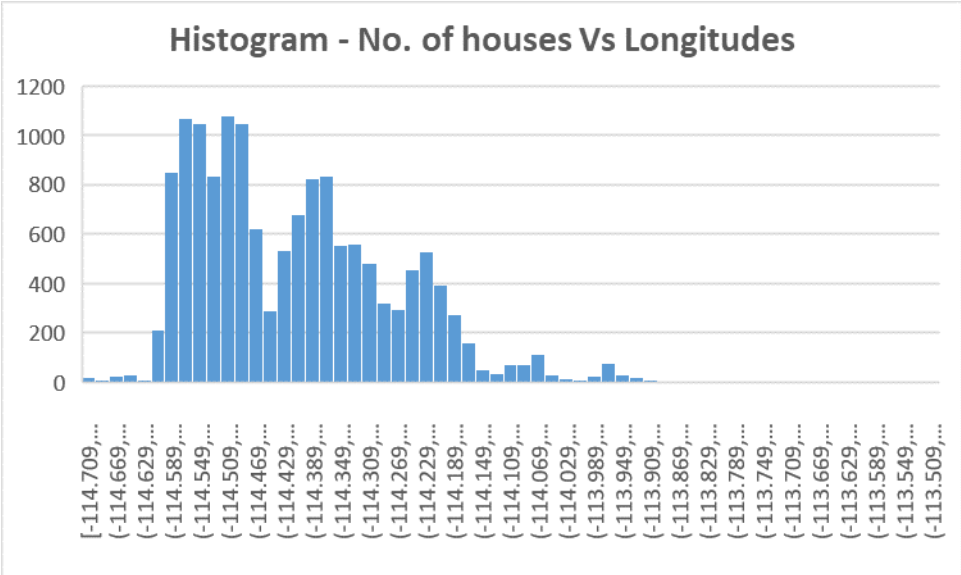
1. We can see that maximum no. of houses are between latitudes 52.7 - 52.9.

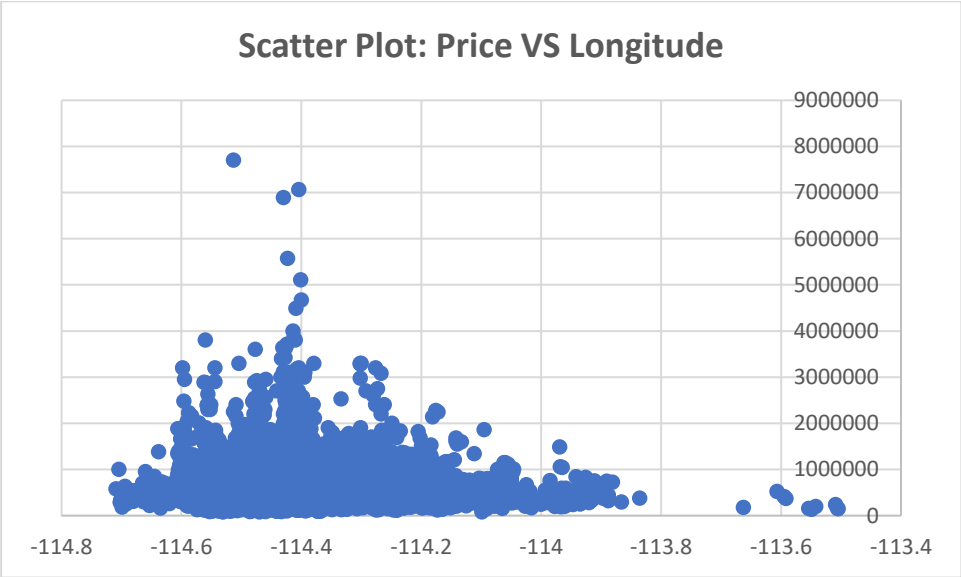
2. Also, we can see different clusters of houses around the following latitudes:

- **a) 52.9**
- **b) 52.7**
- **c) 52.5**

3. The price of the many houses located in between the latitudes of 52.7 to 52.9 are higher than other houses.

17.Longitude





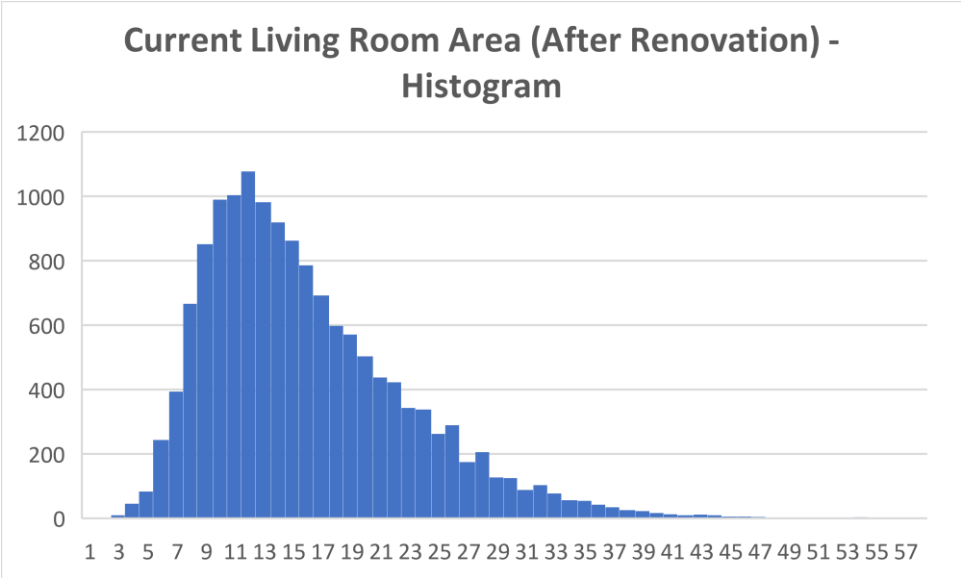
<u>Observations:</u>

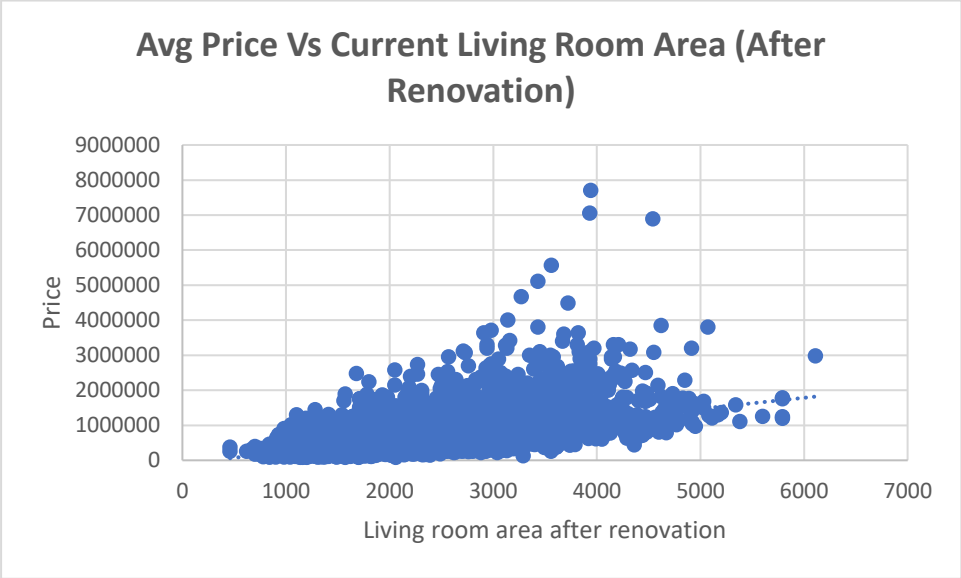
1. We can see 3 different distributions of houses around the following longitudes:

- **a) -114.5**
- **b) -114.3**
- **c) -114.2**

2. Some of the houses located between longitudes -114.6 to -114.2 are costlier than the rest.

18.Living Room Area After Renovation





<u>Observations:</u>

1. The Living Room area is skewed towards the right side due to lot of outliers having high values.

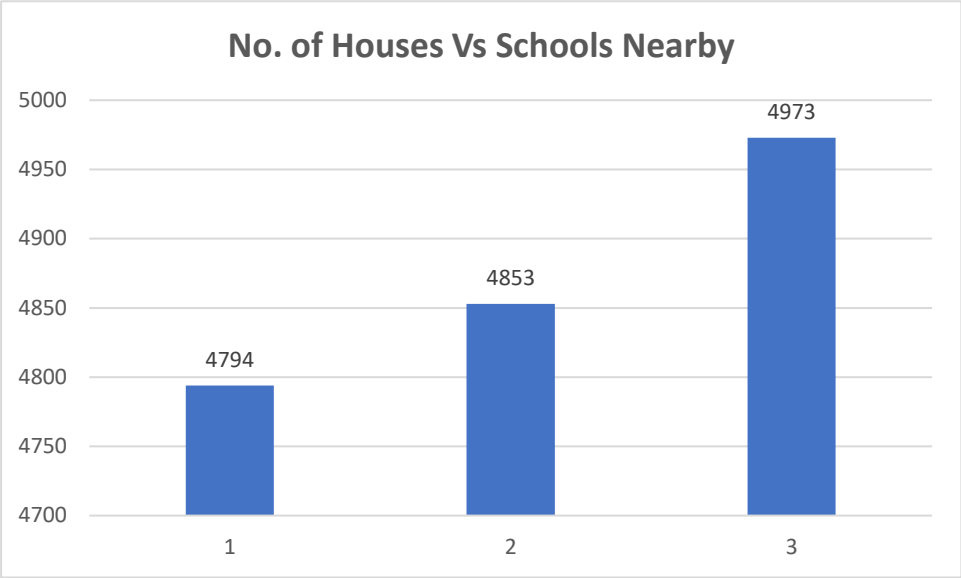
2. The standard deviation is also quite high.

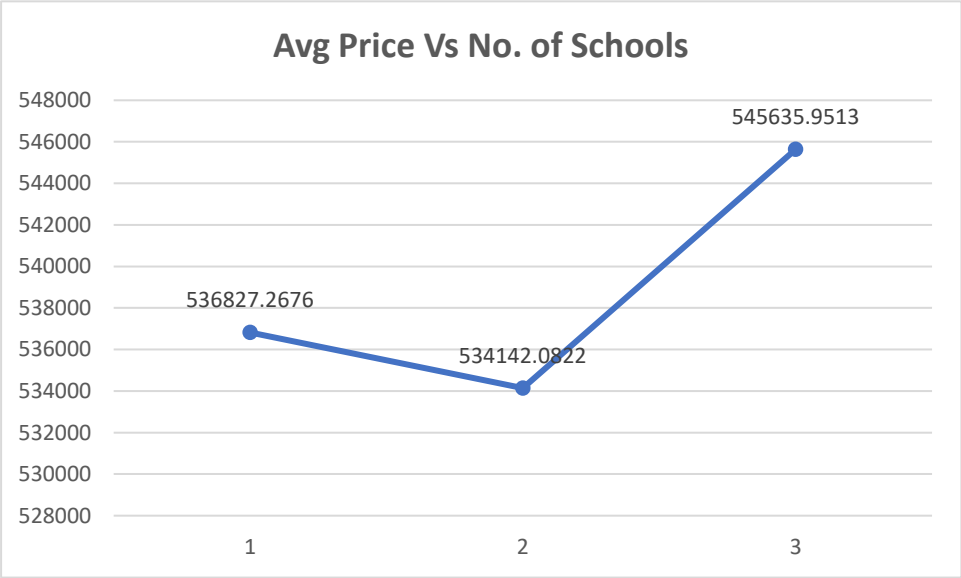
3. We can observe a general linear trend between the living room area and price.

4. As the living room area increases, the house price also increases.

5. Although there are a few outliers having nearly same area but very high price.

19. Number of Schools Nearby





Observations:

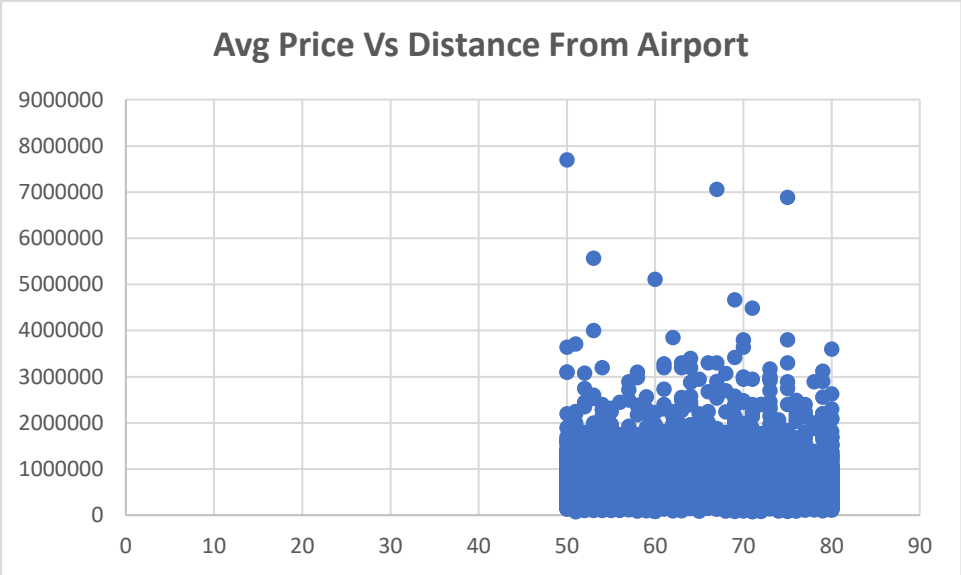
1. Most of the houses have 3 schools nearby.

2. After analysing the graph, we can say that there is no such relation between no. of schools nearby and the average price of the houses.

20. Distance from the Airport

Distance from the airport	
Mean	64.95095759
Standard Error	0.073904325
Median	65
Mode	54
Standard Deviation	8.936007828

Sample Variance	79.8522359
Kurtosis	-1.203048214
Skewness	0.00611433
Range	30
Minimum	50
Maximum	80
Sum	949583
Count	14620



Observations:

1. There is no Linear or Exponential Trend Between Avg price and distance of the house from the airport.

2. Although it is worth noticing that there are a few outliers having high prices with respect to the distance from airport when compared with other houses having the same distance.

A Few Hypotheses Around Some Variables along with data validation

1) **Hypothesis Test 1: To check whether the no. of bedrooms has any effect on the house price**

H0 (Null Hypothesis): The average price of the houses for each category is same

H1 (Alt Hypothesis): The average price of the houses is different across categories

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	2.14221E+14	11	1.94746E+13	**161.591972**	**0**	**1.78930249**
Within Groups	1.76051E+15	14608	1.20517E+11			
Total	1.97474E+15	14619				

Conclusion:

Since the p-value is 0 for an alpha level of 0.05,
We reject the Null-Hypothesis.
Meaning that the average price is different for different no. of bedrooms at 95% confidence.
So, we can conclude that no. of bedrooms is an important variable for predicting house-price.

2) Hypothesis Test 2: To check whether the no. of floors has any effect on the house price

H0 (Null Hypothesis): The average price of the houses for each category is same
H1 (Alt Hypothesis): The average price of the houses is different across categories.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	1.82555E+14	5	3.6511E+13	**297.7221698**	**0**	**2.214710987**
Within Groups	1.79218E+15	14614	1.22634E+11			
Total	1.97474E+15	14619				

Conclusion:

Since the p-value is 0 at an alpha-level of 0.05, we reject the null hypothesis meaning that,
The average price of the houses is different with different no. of views.
So, we can conclude that no. of views is an important variable for predicting house price

3) Hypothesis Test 3: To check whether the waterfront has any effect on the house price

H0 (Null Hypothesis): The average price of the houses for each category is same
H1 (Alt Hypothesis): The average price of the houses is different across categories.

t-Test: Two-Sample Assuming Unequal Variances

	0	*1*
Mean	530417.4276	1641901.714
Variance	1.16452E+11	1.33385E+12

Observations	14508	112
Hypothesized Mean Difference	0	
df	111	
t Stat	**-10.18151258**	
P(T<=t) one-tail	**6.8264E-18**	
t Critical one-tail	**1.658697265**	
P(T<=t) two-tail	**1.36528E-17**	
t Critical two-tail	**1.981566757**	

Conclusion:

Since the p-value is quite less than 0.05 (alpha-level), we reject the null hypothesis.
Meaning that the price of house is different for the houses with and without waterfront
We can conclude that waterfront is an important variable in predicting house price.

4) Hypothesis Test 4: To check whether the no. of views has any effect on the house price.

H0 (Null Hypothesis): The average price of the houses for each category is same
H1 (Alt Hypothesis): The average price of the houses is different across categories.

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	3.26421E+14	4	8.16054E+13	**723.565112**	**0**	**2.372539954**
Within Groups	1.64831E+15	14615	1.12782E+11			
Total	1.97474E+15	14619				

Conclusion:

Since the p-value is quite less than 0.05 (alpha-level), we reject the null hypothesis.
Meaning that the price of house is different for the houses different no. of views
We can conclude that no. of views is an important variable in predicting house price.

5) Hypothesis Test 5: To check whether the latitude has any effect on the house price

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.29748998
R Square	0.088500288
Adjusted R Square	0.088437934
Standard Error	350904.3322
Observations	14620

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-41434139.9	1114125.66	-37.1898264	**2.4494E-289**
Latitude	795052.2468	21103.65193	37.67368081	**1.6078E-296**

Conclusion:

The p-value is less than 0.05, leading to the conclusion that Latitude has an effect on the house price and is a significant predictor of the same.

Correlation Table Between Independent and Target Variable

Correlation	*Price*
living area	0.712169477
grade of the house	0.67181438
Area of the house (excluding basement)	0.61522042
living_area_renov	0.584924464
number of bathrooms	0.531734563
number of views	0.395973098
Area of the basement	0.330202327
number of bedrooms	0.308460143
Latitude	0.29748998
waterfront present	0.263686554
number of floors	0.262731825
Renovation Year	0.133172648
lot area	0.081991997
lot_area_renov	0.075535167
Built Year	0.050307111
condition of the house	0.041376376
Longitude	0.024414019
Number of schools nearby	0.009889891
Distance from the airport	0.003803696
Postal Code	-0.11590817

Top 10 Variables which are highly correlated with the Price Variable (Target) are:

1	Living Area (Sq Ft)	0.71
2	Grade of the house (1 - 13)	0.67
3	Area of the house (excluding basement) (Sq ft)	0.61
4	Living room area after renovations (Sq ft)	0.58
5	No. of bathrooms	0.53

6	**No. of views**	**0.39**
7	**Area of the basement (Sq ft)**	**0.33**
8	**No. of bedrooms**	**0.3**
9	**Latitude**	**0.29**
10	**waterfront present (0-Yes / 1-No)**	**0.26**

A linear regression model on the data of year 2016

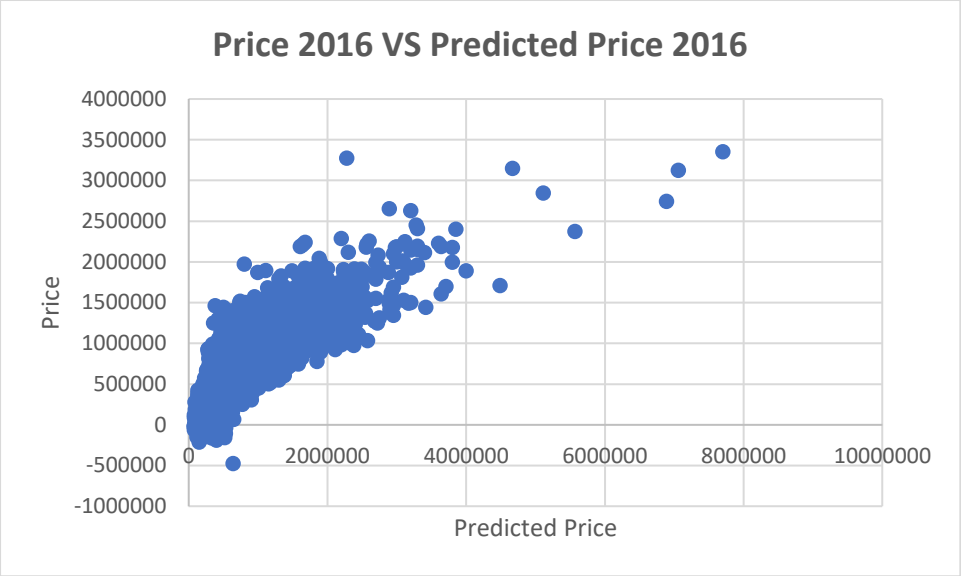
SUMMARY OUTPUT

Regression Statistics	
Multiple R	**0.836796361**
R Square	**0.700228149**
Adjusted R Square	**0.69998188**
Standard Error	201311.855
Observations	14620

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**-34458081.1**	1739199.566	-19.81260906	**3.13328E-86**
number of bedrooms	**-33550.03407**	2249.748123	-14.91279567	**6.33811E-50**
number of bathrooms	**39977.94092**	3859.632974	10.35796439	**4.70179E-25**
living area	**173.719517**	3.964344375	43.82049101	**0**
number of floors	**16747.20756**	3898.137122	4.296207915	**1.74868E-05**
waterfront present	**593095.6568**	20888.07945	28.39397745	**1.1284E-172**
number of views	**45829.04888**	2555.988413	17.93006911	**3.96991E-71**
condition of the house	**28017.22701**	2758.884952	10.15527196	**3.7761E-24**
grade of the house	**101158.9435**	2599.033764	38.92175043	**0**
Built Year	**-2601.885589**	82.39015239	-31.58005554	**8.477E-212**
Latitude	**530824.7607**	12661.58471	41.92403818	**0**
Longitude	**-94516.0489**	13823.98173	-6.837107478	**8.40138E-12**
living_area_renov	**19.26109753**	4.12190231	4.67286609	**2.99677E-06**

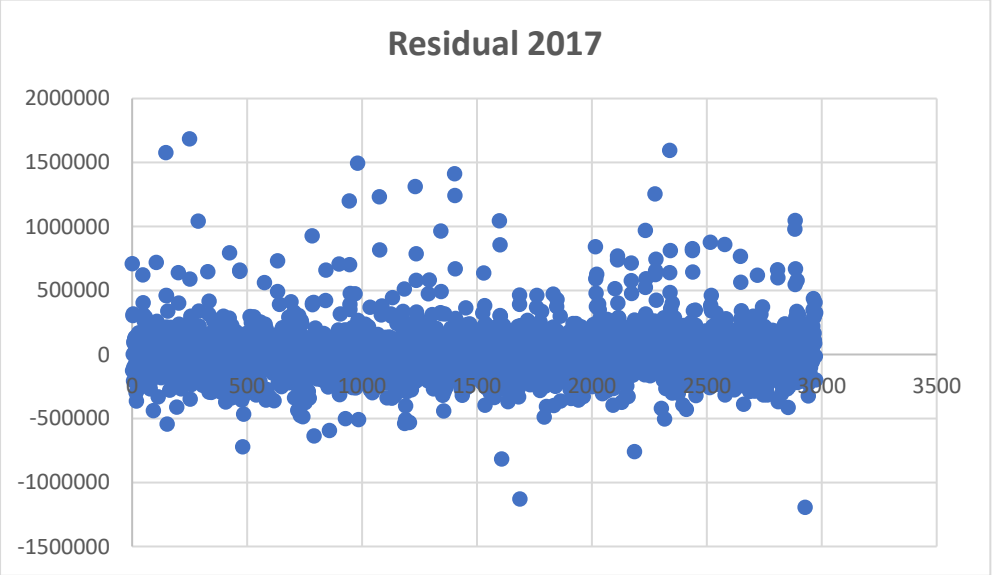
NOTE: Using this regression model, the price for year 2016 has been predicted which can be found in the excel file.

Plotting the Regressed Values against the Actual Values to understand the difference.



<u>NOTE:</u> Using the above linear regression model, the prices of the houses sold in the year 2017 has been predicted which can be found in the excel file.

Plotting the Residuals (Actual Value – Predicted Value) to understand the difference.



<u>Findings are given below</u>:

1. There are 26 cases in which the price predicted is coming out to be negative.

2. The data is approximately 70% explained by the variables.

3. We can observe that a lot of residuals are near 0 while some are far above and some are far below the 0 mark.

4. Further examination needs to be done as to determine why some predictions are coming out to be negative.

5. Regularization needs to be done because there is one variable (living_area) which has the highest influence on the model, so its effect needs to be piped down to make the model more accurate and less biased.

THANK YOU