**Data analytics using Excel**

**House Price Prediction**

**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:**

Descriptions and names of the columns (features) are given below.

**Id:** unique id

**Date:** Date house was sold

**Price:** Price of the sold house (Target Variable)

**Bedrooms:** Number of Bedrooms

**Bathrooms:** Number of bathrooms

**Living area:** square footage of the living space

**Lot area:** square footage of the lot

**Number of floors:** Total floors in the house

**Waterfront:** Whether the house is on a waterfront (1: yes, 0: no)

**Number of views:** number of special views

**Condition:** Condition of the house on a scale of 1-5 (1 being the lowest, 5 being the highest)

**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)

**Area of the house:** square footage of house apart from basement

**Area of Basement:** Square footage of the basement

**Built year:** Built year

**Renovation year:** Year when the house was renovated

**Postal code:** Postal code of the house

**Living\_area\_renov:** Living room area currently (after renovations)

**Lot\_area\_renov:** Lot area currently (after renovations)

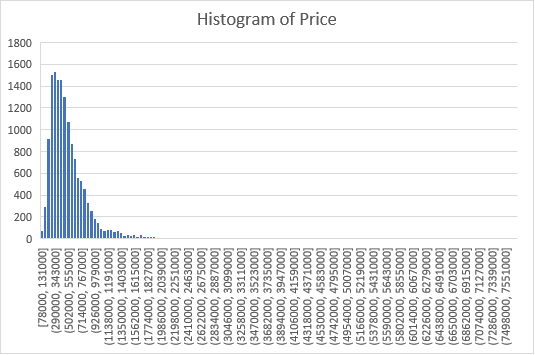
**Number of Schools nearby:** Number of schools in the vicinity of the house

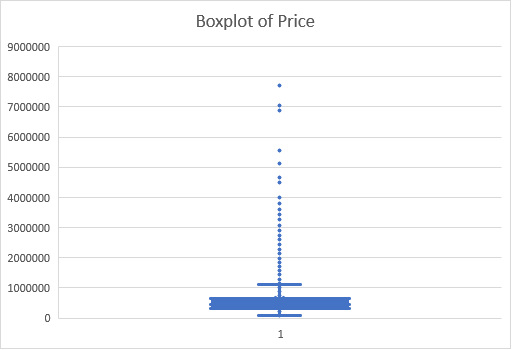
**Distance from the airport:** Distance in KMs from nearest Airport

**Descriptive Statistics:**

1. **Price Target**

|  |  |
| --- | --- |
| ***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 |





**Inference:**

There are some extreme values (not considering as outliers), so it is right skewed. The range is higher than 3 standard deviations because of extreme values.

1. **Number of bedrooms**

**Inference:**

* Mostly people prefer 3-bedroom house. Maximum price for 6-bedroom house.
* The average price is high for 8-bedroom house.
* Above 10-bedroom house, min, max and average price is same for house.

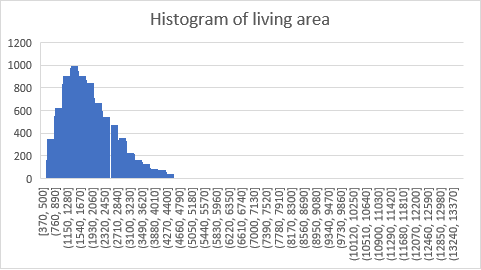
1. **number of bathrooms**

**Inference:**

* Highest number of houses has 2.5 bathrooms followed by 1,1.5.
* The data is not normal
* Number of bathroom increases, min and average price also increases. But there is a drop in 7.5 and then price increases.

1. **living area**

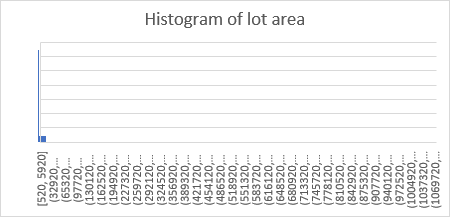
|  |  |
| --- | --- |
| ***living area*** |  |
| Mean | 2098.263 |
| Standard Error | 7.677208 |
| Median | 1930 |
| Mode | 1400 |
| Standard Deviation | 928.2757 |
| Sample Variance | 861695.8 |
| Kurtosis | 6.073617 |
| Skewness | 1.538337 |
| Range | 13170 |
| Minimum | 370 |
| Maximum | 13540 |
| Sum | 30676605 |
| Count | 14620 |



**Inference:**

* Due to some extreme values/outliers, the distribution is right skewed.
* The range is higher than 3 standard deviations
* The size of living area increases, price also increases

1. **lot area**



**Inference:**

* Highly right skewed.
* There is no change in price as lot area increases and no relationship exists between price and lot area

1. **number of floors**

**Inference:**

* Most of the house has 1 floor followed by 2 floor
* Number of floors increases, price also increases. But there is a drop in 3 floor and then it increases

1. **waterfront present**

**Inference:**

* Most houses have no waterfront.
* Price is high for house which has waterfront

1. **number of views**

**Inference:**

* Maximum house has 0 views
* As number of views increases, price also increases

1. **condition of the house**

**Inference:**

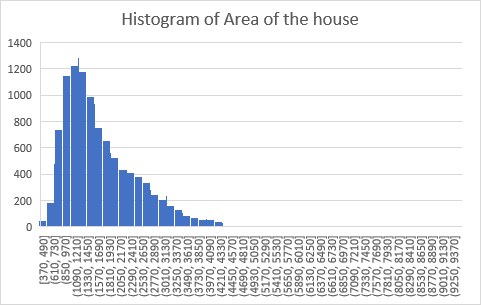
* condition of most house is 3 (medium)
* As condition of house increases, price is also increasing.

1. **grade of the house**

**Inference:**

* Most of the house has grade 7 ,followed by 8
* As the grade increases, price also increases

1. **Area of the house (excluding basement)**

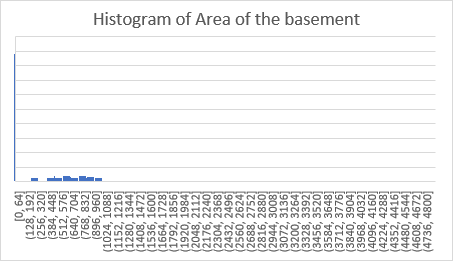


**Inference:**

Due to some outliers, the data is right skewed

As area of the house increases , price also increases and has linear relationship

1. **Area of the basement**



**Inference:**

* Data is normal. Most houses have no basement
* We can see a lot of datapoints at 0 because most of the houses have no basement.
* In houses having no basement, there is lot of variation in price.
* 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.

1. **Built Year**

**Inference:**

* A lot of ups and downs are there in terms of House Price w.r.t the year of built.
* On an average the house prices are stagnant through the years.
* As a general linear trend, the prices are somewhat decreasing as the built year is increasing.

1. **Renovation Year**

**Inference:**

Maximum houses are not renovated

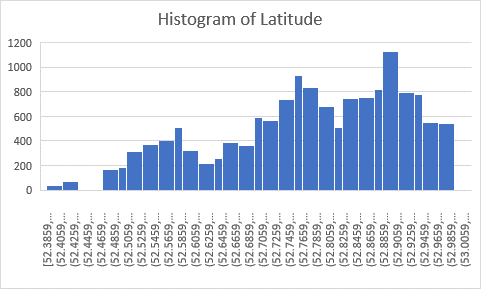
In 1988 and 2002, price is high for renovation

1. **Postal Code**

**Inference:**

The postal code 122070-71 has the high price

1. **Latitude**



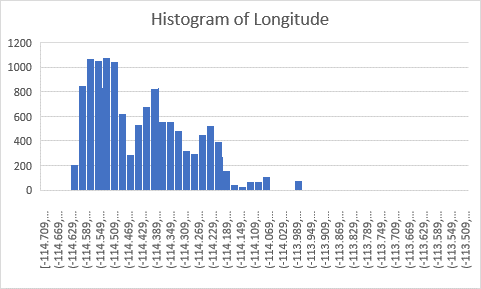
**Inference:**

Data is not normal

The house price is high between the latitude 52.8 -52.9

And most houses also built in that latitude

1. **Longitude**



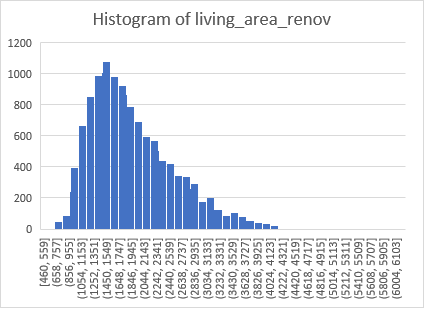
**Inference:**

Data is not normal

The house price is high between the longitude -114.6 to -114.2

And most houses also built in that longitude

1. **living\_area\_renov**



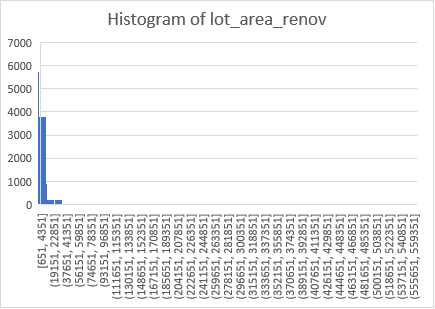
**Inference:**

It is slightly right skewed due to some outliers

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

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

1. **lot\_area\_renov**



**Inference:**

It is highly right skewed due to outliers

We can observe there is no pattern

As the lot area increases, the house price also increases.

1. **Number of schools nearby**

**Inference:**

Most of the house has 3 schools in nearby

As number of school increases, price also increases

It has linear trend but there is a slight drop in 2

1. **Distance from the airport**

**Inference:**

* All the houses are lies between 50- 80 km distance from the airport
* There is no Linear or Exponential Trend
* Between Avg price and distance of the house from the airport.
* 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 the important variables**

1. **Hypothesis 1\_No of bedrooms**

To check whether the no. of bedrooms has any effect on the house price

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

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

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ANOVA** |  |  |  |  |  |  |  |  |
| ***Source of Variation*** | ***SS*** | ***df*** | ***MS*** | ***F*** | ***P-value*** | ***F crit*** |  |  |
| Between Groups | 2.14E+14 | 10 | 2.14E+13 | 177.7427 | 0 | 1.83135 |  |  |
| Within Groups | 1.76E+15 | 14608 | 1.21E+11 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Total | 1.97E+15 | 14618 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| **Conclusion** |  |  |  |  |  |  |  |  |
| **Since the p-value is quite less than 0.05 (alpha-level), we reject the null hypothesis.** | | | | | | | |  |
| **Meaning that the avg 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.** | | | | | | | | |

1. **Hypothesis 2\_ No of floors**

To check whether the no. of floors has any effect on the house price

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

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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ANOVA** |  |  |  |  |  |  |  |
| ***Source of Variation*** | ***SS*** | ***df*** | ***MS*** | ***F*** | ***P-value*** | ***F crit*** |  |
| Between Groups | 1.83E+14 | 5 | 3.65E+13 | 297.7222 | 0 | 2.214711 |  |
| Within Groups | 1.79E+15 | 14614 | 1.23E+11 |  |  |  |  |
|  |  |  |  |  |  |  |  |
| Total | 1.97E+15 | 14619 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| **Conclusion:** |  |  |  |  |  |  |  |
| **Since the p-value is quite less than 0.05 (alpha-level), we reject the null hypothesis.** | | | | | | | |
| **The avg 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** | | | | | | | |

1. **Hypothesis 3\_waterfront present**

To check whether the waterfront has any effect on the house price

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

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **t-Test: Two-Sample Assuming Unequal Variances** | | |  |  |  |
|  |  |  |  |  |  |
|  | *0* | *1* |  |  |  |
| Mean | 530417.4 | 1641902 |  |  |  |
| Variance | 1.16E+11 | 1.33E+12 |  |  |  |
| Observations | 14508 | 112 |  |  |  |
| Hypothesized Mean Difference | 0 |  |  |  |  |
| df | 111 |  |  |  |  |
| t Stat | -10.1815 |  |  |  |  |
| P(T<=t) one-tail | 6.83E-18 |  |  |  |  |
| t Critical one-tail | 1.658697 |  |  |  |  |
| P(T<=t) two-tail | 1.37E-17 |  |  |  |  |
| t Critical two-tail | 1.981567 |  |  |  |  |
|  |  |  |  |  |  |
| **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.** | | | | | |

1. **Hypothesis 4\_Number of schools**

To check whether the no. of schools nearby has any effect on the house price

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

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

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **ANOVA** |  |  |  |  |  |  |  |  |  |
| **Source of Variation** | **SS** | **df** | **MS** | **F** | **P-value** | **F crit** |  |  |  |
| Between Groups | 3.56E+11 | 2 | 1.78E+11 | 1.318099 | 0.267675 | 2.996346 |  |  |  |
| Within Groups | 1.97E+15 | 14617 | 1.35E+11 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| Total | 1.97E+15 | 14619 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
| **Conclusion:** |  |  |  |  |  |  |  |  |  |
| **Since the p-value is greater than 0.05 (alpha-level), we accept the null hypothesis.** | | | | | | | |  |  |
|  |  |  |  |  |  |  |  |  |  |
| **We can conclude that number of schools nearby is not an important variable in predicting house price.** | | | | | | | | | |

1. **Hypothesis 5\_number of views**

To check whether the no. of views has any effect on the house price

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

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

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ANOVA** |  |  |  |  |  |  |  |
| ***Source of Variation*** | ***SS*** | ***df*** | ***MS*** | ***F*** | ***P-value*** | ***F crit*** |  |
| Between Groups | 3.26E+14 | 4 | 8.16E+13 | 723.5651 | 0 | 2.37254 |  |
| Within Groups | 1.65E+15 | 14615 | 1.13E+11 |  |  |  |  |
|  |  |  |  |  |  |  |  |
| Total | 1.97E+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.** | | | | | | | |

1. **Hypothesis 6\_Lattitude**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVA** |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 1 | 1.74765E+14 | 1.75E+14 | 1419.3062 | 1.6078E-296 |  |
| Residual | 14618 | 1.79997E+15 | 1.23E+11 |  |  |  |
| Total | 14619 | 1.97474E+15 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | -41434140 | 1114125.66 | -37.1898 | 2.45E-289 | -43617966.85 | -39250313 |
| Lattitude | 795052.247 | 21103.65193 | 37.67368 | 1.61E-296 | 753686.424 | 836418.07 |

To check whether the latitude has any effect on the house price

|  |  |
| --- | --- |
| **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 between variables:**

***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) | 0.61 |
| *4* | Living room area after renovations (Sq ft) | 0.58 |
| *5* | No. of bathrooms | 0.53 |
| *6* | number of views | 0.39 |
| *7* | Area of the basement (Sq ft) | 0.33 |
| *8* | number of bedrooms | 0.3 |
| *9* | Lattitude | 0.29 |
| *10* | waterfront present (0-Yes / 1-No) | 0.26 |

**linear regression model on the data of year 2016:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ***Regression Statistics*** |  |  |  |  |  |  |  |  |
| Multiple R | 0.836902073 |  |  | | |  |  |  |
| R Square | 0.70040508 |  |  |  |  |  |  |  |
| Adjusted R Square | 0.700138427 |  |  |  |  |  |  |  |
| Standard Error | 201259.3265 |  |  |  |  |  |  |  |
| Observations | 14620 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| **ANOVA** |  |  |  |  |  |  |  |  |
|  | ***df*** | ***SS*** | ***MS*** | ***F*** | ***Significance F*** |  |  |  |
| Regression | 13 | 1.38311E+15 | 1.06393E+14 | 2626.653509 | 0 |  |  |  |
| Residual | 14606 | 5.91621E+14 | 40505316493 |  |  |  |  |  |
| Total | 14619 | 1.97474E+15 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | ***Coefficients*** | ***Standard Error*** | ***t Stat*** | ***P-value*** | ***Lower 95%*** | ***Upper 95%*** | ***Lower 95.0%*** | ***Upper 95.0%*** |
| **Intercept** | -68141440.41 | 11599774.98 | -5.874376055 | 4.3367E-09 | -90878465.77 | -45404415.05 | -90878465.77 | -45404415.05 |
| **number of bedrooms** | -33417.30274 | 2249.615089 | -14.85467577 | 1.4918E-49 | -37826.8327 | -29007.77278 | -37826.8327 | -29007.77278 |
| **number of bathrooms** | 40302.17084 | 3860.204773 | 10.44042304 | 1.99177E-25 | 32735.6815 | 47868.66018 | 32735.6815 | 47868.66018 |
| **living area** | 172.9820269 | 3.971256683 | 43.55851074 | 0 | 165.1978617 | 180.766192 | 165.1978617 | 180.766192 |
| **number of floors** | 17421.26101 | 3903.872062 | 4.462559409 | 8.15921E-06 | 9769.178254 | 25073.34376 | 9769.178254 | 25073.34376 |
| **waterfront present** | 592278.0433 | 20884.48459 | 28.35971559 | 2.8408E-172 | 551341.8134 | 633214.2732 | 551341.8134 | 633214.2732 |
| **number of views** | 45576.5247 | 2556.767603 | 17.82583784 | 2.47313E-70 | 40564.93698 | 50588.11241 | 40564.93698 | 50588.11241 |
| **condition of the house** | 28004.72708 | 2758.168357 | 10.15337842 | 3.84964E-24 | 22598.36842 | 33411.08573 | 22598.36842 | 33411.08573 |
| **grade of the house** | 101641.0998 | 2603.536593 | 39.0396279 | 0 | 96537.83897 | 106744.3607 | 96537.83897 | 106744.3607 |
| **Built Year** | -2606.627928 | 82.38447948 | -31.63979361 | 1.4467E-212 | -2768.111922 | -2445.143933 | -2768.111922 | -2445.143933 |
| **Postal Code** | 275.1385007 | 93.68087219 | 2.936976292 | 0.003319468 | 91.51214849 | 458.7648529 | 91.51214849 | 458.7648529 |
| **Lattitude** | 542659.2482 | 13284.15542 | 40.85011287 | 0 | 516620.6242 | 568697.8721 | 516620.6242 | 568697.8721 |
| **Longitude** | -90033.32658 | 13904.40083 | -6.475167658 | 9.77327E-11 | -117287.7099 | -62778.94322 | -117287.7099 | -62778.94322 |
| **living\_area\_renov** | 19.55011401 | 4.122001589 | 4.74286911 | 2.12704E-06 | 11.47046981 | 27.6297582 | 11.47046981 | 27.6297582 |

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.

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

**Inference:**

* 70% variation is explained by the model
* We can observe there are some negative values that means some of the predicted values are higher than the actual price.
* We need further examination to understand more about data. Regularization can be done.