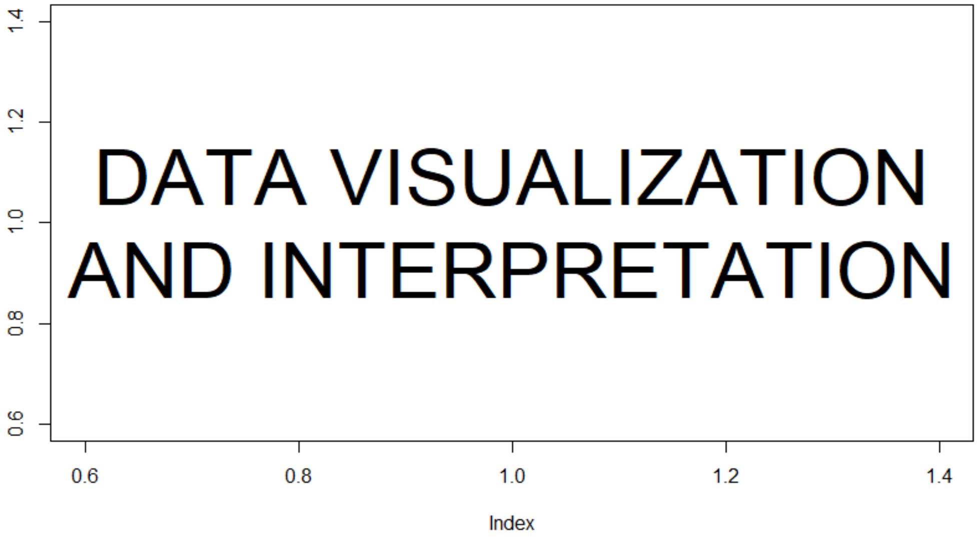
~~~
```{r}
plot (1, type = 'n'); text (1, 1, 'DATA VISUALIZATION
AND INTERPRETATION', cex = 3)
```
~~~



Members:

Lahiru Lowe
COADDSFT191F-002

Amir Sohil
COADDSFT191F-006

Sanjula Meneripitiya
COADDSFT191F-012

Pasan Maneth
COADDSFT191F-023

Batch:

COADDSFT191F

Abstract

The objective of conducting this research was to derive a relationship between the price of a house and the other factors such as number of rooms, area of the house, number of bathrooms, number of floors, etc. This would be useful for real-estate agencies to properly value their houses as well as for the population to evaluate the price of a house with its available features before buying a house.

Different variables would affect the price of the house both positively and negatively. The variables analyzed include number of bedrooms, number of bathrooms, area of the house, number of floors, availability of a waterfront, condition, grade (given by the local authority) and the year built. Each and every variable among these were analyzed individually to derive the relationship between these variables and the price of a house. Some of these variables were categorical and some were numerical. Therefore, appropriate data visualization techniques were used to analyze different data sets accordingly.

Programming languages like Python and data visualization software like R were used during the research to make the results more accurate. Initially, the data sets were analyzed individually. Then relationships among the variables were derived. Finally, all the insights derived through the previous analysis were blended together and used to formulate a multiple regression model to derive the price of a house.

During our analysis, a strong negative correlation between price and year built was apparent. All the other variables showed a strong or a moderate positive correlation with price and the other variables.

By considering all these factors, the final output of the research was derived as a multiple regression model which can be easily used by the average layman to approximate a value of a house before buying or selling a house. So, in conclusion, with the help of this research, a buyer no longer needs to pay an excessive value on a property.

Acknowledgement

This research was supported by the National Institute for Business Management (NIBM). We thank all of our colleagues from NIBM who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper.

We express our heartfelt thanks to Mr. Thurairasa Balakumar [Lecturer - Statistics] for assistance with data visualization techniques and Mrs. Thisuri Liyanage [Lecturer - English] for comments that greatly improved the manuscript.

We would also like to show our gratitude to the founders and developers of all the open-source software which were used in making this research a success, the most important of which were the Python program and the RStudio visualization software.

Contents

Introduction

Investment is a business activity in which most people are interested in this globalization era. There are several entities that are often used for investment, for example, gold, stocks and property. In particular, investment in property has increased significantly over the years. Housing price trends are not only the concern of buyers and sellers, but they also indicate the current economic situation. There are many factors which have an impact on house prices, such as numbers of bedrooms and bathrooms. Even the number of floors in a house, its condition, grade, the year it was built in and the living space affect the house's price. Manual house predication becomes difficult with the increase of these factors, hence there are many machine learning tools for house price prediction that use linear regression. The aim is to make a model which can give us a good house pricing prediction based on these variables. Linear regression has been used for this dataset and consequently it can be said to give a fair amount of accuracy.

In this report, the insights of the given data have been gained by descriptive analysis methods using the R programming language, data visualization and by seeing the relationships between each variable. Subsequently, a model has been created to predict the house's price using Python programming language. A model like this would be very valuable for a real estate agent who could make use of the information provided on a daily basis.

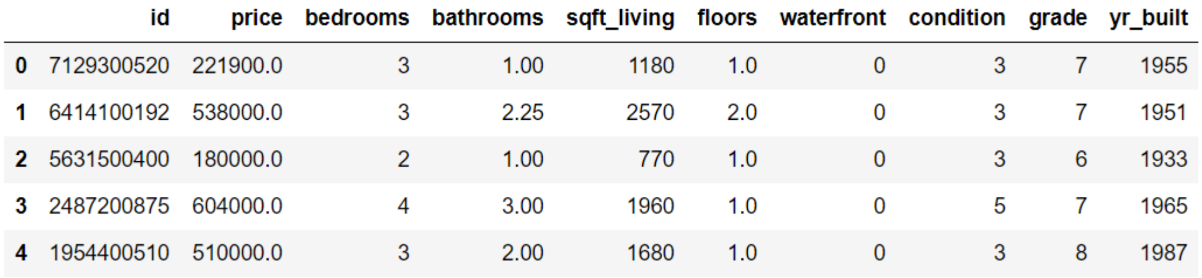
The following link leads to a GitHub repository containing all codes, markdown files, notebook that might be considered necessary to reference the conjoint English-Statistics assignment: https://github.com/StatisticsEnglishAssignment.

Methodology

The dataset used in this project is a part of a dataset consisting of many more features which can be accessed from https://www.kaggle.com/harlfoxem/housesalesprediction. This dataset contains house sale prices for King County in the USA, which includes Seattle. It includes homes sold between May 2014 and May 2015.

The subset used by us consists of 21613 entries and represents aggregate information of about 10 features of homes from various suburbs.

This is an overview of the subset used, with its original features:



	id	price	bedrooms	bathrooms	sqft_living	floors	waterfront	condition	grade	yr_built
0	7129300520	221900.0	3	1.00	1180	1.0	0	3	7	1955
1	6414100192	538000.0	3	2.25	2570	2.0	0	3	7	1951
2	5631500400	180000.0	2	1.00	770	1.0	0	3	6	1933
3	2487200875	604000.0	4	3.00	1960	1.0	0	5	7	1965
4	1954400510	510000.0	3	2.00	1680	1.0	0	3	8	1987

Figure 1 - Data subset overview

The features can be summarized as follows:

- **id:** Identity of each house.
- **price:** Price of respective house.
- **bedrooms:** Number of bedrooms in the house.
- **bathrooms:** Number of bathrooms in the house.
- **sqft_living:** Living space of the house.
- **floors:** Number of floors in the house.
- **waterfront:** Presence or absence of a waterfront.
- **condition:** Score of condition.
- **grade:** Score of grading.
- **yr_built:** Year the house was built in.

For the purpose of the project the dataset has been reprocessed as follows:
The values of the 'waterfront' attribute, i.e. 1's and 0's have been replaced with 'Yes' and 'No'.

This data has been analyzed descriptively, visualized and a model has been found as per the aim of this project. Various methods were used for this purpose, inclusive of linear regression methods, machine learning algorithms, Ridge regression methods, various visualization techniques, various statistical procedures, etc.

Using these methods and various other unspecified ones, results were obtained which could be used by agents, buyers, financial institutions like banks and any other party interested in pricing a house.

Descriptive Analysis

Price

Minimum: 75,000

Maximum: 7,700,000

Mean: 540,182

First Quartile (Q1): 321,950

Median (Q2): 450,000

Third Quartile (Q3): 645,000

The five number summary and the mean of the variable 'price' has been provided above. As could be deduced from observation, mean has most probably been affected by the extremity of the outliers. This could be deemed evident by the existence of large differences between the mean and both the minimum and the maximum. Therefore, in order to have a fairer analysis, the mean with the suspected outliers eliminated will be provided.

Mean with outliers eliminated: 476,591.2

Mode: 350,000 and 450,000

As seen, the variable 'price' seems to have two modes, which essentially means that both of them occur at equal frequencies in the provided dataset.

Interquartile Range (IQR): 323,050

Range: 7,625,000

As can be observed, the IQR looks almost Lilliputian in comparison to the variable's range. This is because the presence of outliers does not affect the IQR of any given variable. Such measures like Q1, Q2, Q3 and IQR, which remain unaffected by extreme values, are known as resistant measures.

Standard deviation: 367362.2

Variance: 1.34955e+11

A large standard deviation, such as the one obtained here, means that the values in the data set are farther away from the mean, on average. This further proves the earlier statement, where it was mentioned that the mean has most probably been affected by the extremity of the variable's outliers.

Coefficient of variation: 68.0071

Above is a measure of relative dispersion representing the degree of variability relative to the mean. It is usually given as a percentage (like the one given here) and its high nature here suggests the presence of extreme values in the dataset.

Bedrooms

Counts of various values of the variable 'bedrooms' are given below:

No bedrooms: 13

1 bedroom: 199

2 bedrooms: 2760

3 bedrooms: 9824

4 bedrooms: 6882

5 bedrooms: 1601

6 bedrooms: 272

7 bedrooms: 38

8 bedrooms: 13

9 bedrooms: 6

10 bedrooms: 3

11 bedrooms: 1

33 bedrooms: 1

Using the given data, it can be observed that the number of bedrooms range from none to 33. Here, most of the data can be seen concentrated at a single region.

Mean: 3.370842

The mean number of bedrooms in each house is as mentioned above. This could be rounded off to 3 (the value of the median), hence showing that the data is indeed concentrated around that area (as that value has the most frequency

Square feet of living

Minimum: 290 **First Quartile (Q1): 1,427**

Maximum: 13,540 **Median (Q2): 1,910**

Mean: 2,080 **Third Quartile (Q3): 2,550**

The five number summary and the mean of the variable 'sqft_living' has been provided above. Here too, mean has most probably been affected by the extremity of the outliers. Therefore, here too, the mean with the suspected outliers eliminated will be provided.

Mean with outliers eliminated: 1,998.913 ***Mode: 1,300***

As can be observed, the mean has reduced in value after the elimination of suspected outliers. This may seem smaller in comparison to the 'price' variable but, nonetheless, is still very significant.

Interquartile Range (IQR): 1,123 **Range: 13,250**

The IQR is considerably smaller than the variable's range. This too can be concluded as due to the reasons mentioned above. Hence, the IQR can be said to give a fairer analysis than the range.

Standard deviation: 918.4409 **Variance: 843,533.7**

The standard deviation obtained has a highly considerable value with respect to many of the values of the variable. Therefore, the statement that the mean has been affected by the extremity of the variable's outliers is plausible.

Coefficient of variation: 44.15794

The coefficient of variation of the variable 'sqft_living' is lower than that of the 'price' variable but it is still very high in comparison to typically acceptable coefficients of variation.

Bathrooms

Counts of various values (values have been rounded off to the nearest whole number as number of bathrooms in a house cannot be a floating point):

No bathrooms: 14 **1 bathroom: 3,933** **2 bathrooms: 13,851**

3 bathrooms: 2,527

4 bathrooms: 1,201

5 bathrooms: 57

6 bathrooms: 24

7 bathrooms: 2

8 bathrooms: 4

Through observation of the given data, it can be concluded that the number of bathrooms range from none to 8. Here too, most of the data can be seen concentrated at a single region.

Mean: 2.058715

The mean number of bathrooms in each house is as given above. This could be rounded off to 2, hence showing that the mean is almost the same as the mode and the median.

Floors

Counts of various values (values have been rounded off to the nearest whole number as number of floors in a house cannot be a floating point):

1 floor: 10,680

2 floors: 10,312

3 floors: 613

4 floors: 8

It is noticeable that most of the houses have 1 or 2 floors, with the majority of them having 1. Houses with 3 or 4 floors are present but their proportion in the total percentage is almost trivial.

Mean: 1.534956

The mean number of floors in each house is as provided above. When rounded off, it takes the value of 2 floors per house (which is also the variable's median).

Waterfronts

Houses with a presence of a waterfront: 163

Houses with an absence of a waterfront: 21,450

As observable, houses with an absence of a waterfront make up for a gargantuan proportion of the total houses. In fact, houses with no waterfront make up around 99.246% of the total houses, thereby making it crystal clear how significant of a part is made up by such houses.

Condition

Counts of various values of the variable 'condition' are given below:

1: 30

2: 172

3: 14,031

4: 5,679

5: 1,701

The value of condition of most of the houses can be seen to be 3. The value making up the least proportion of houses can be observed to be 1. The proportion making up the conditional value of 2 is can be deemed almost picayune too. The proportion of houses making up the values 4 and 5 are comparatively significant.

Mean: 3.40943

The mean value of condition of each house is as given above. It is slightly larger than the mode, indicating that the values larger than 3 (the median) make for a significantly larger portion of the data than those smaller.

Grade

Counts of various values of the variable 'grade' are given below:

1: 1

3: 3

4: 29

5: 242

6: 2,038

7: 8,981

8: 6,068

9: 2,615

10: 1,134

11: 399

12: 90

13: 13

The value of the grade of most of the houses can be observed to be 7. The number of houses making up most of the values is almost insignificant, indicating concentration of data at a certain point.

Mean: 7.656873

The mean value of grade of each house is as provided. It is observed to be slightly larger than the mode and the median, both of which are 7.

Year built

Minimum: 1900

First Quartile: 1951

Mean: 1971

Median: 1975

Mode: 2014

Third Quartile: 1997

Maximum: 2015

IQR: 46

Range: 115

The differences between the mean, median and the mode can be attributed to the presence of outliers. The large difference between the range and the IQR can be attributed to the same fact too.

Visualizing and interpreting the relationship between variables

Price visualization

These graphs below show the visualization of variables with respect to house pricing. Let's check them out one by one.

1. **First graph shows the visualization of bedrooms and house price.**

The insight gained from this graph clearly shows that if a house has around five bedrooms, the price has increased but if it increases to more than five the price hasn't increased. That means people are not willing to buy houses with more bedrooms. The same scenario is seen below five bedrooms as the price has decreased. It shows that people don't usually go for houses with a low number of bedrooms and due to this, the price of houses are comparatively low. So we can come up with a conclusion like this: keeping the number of bedrooms around five is better to generate more sales.

2. **2nd, 6th, 7th and 8th graphs show the visualization of price with respect to the number of bathrooms, the house's grade, the year the house was built in and the area of living.**

According to the respective graphs, when the number of bathrooms increased the price has also increased. The price vs. grade graph shows that when the grade is high, the price is high, so maintaining a good grade is a good marketing strategy as well. When we look at the graph showing the year the house was built in, there is no big deviation but the price has comparatively increased in houses built around 2000. The last graph shows that with increased living area, the price has also increased. As per these graphs, the conclusion drawn is that people could be seen to be willing to buy houses with more facilities. So, from a seller's perspective, facilities could be increased to potentially sell the respective house for a higher price.

3. **Price visualization with respect to number of floors and the condition.**

The visualization with floors shows that until around two floors the price has increased. Here too, the scenario is the same as with bedrooms as when number of floors increase, the price has reduced because people are not willing to buy a house with more floors. The graph with condition shows that generally houses with a condition of around 3 are more expensive.

4. **Price visualization with respect to the presence or absence of a waterfront.**

So according to the graph below which considers the presence of a waterfront, when a waterfront is present, then the price is generally high and with its absence, the price is comparatively low.

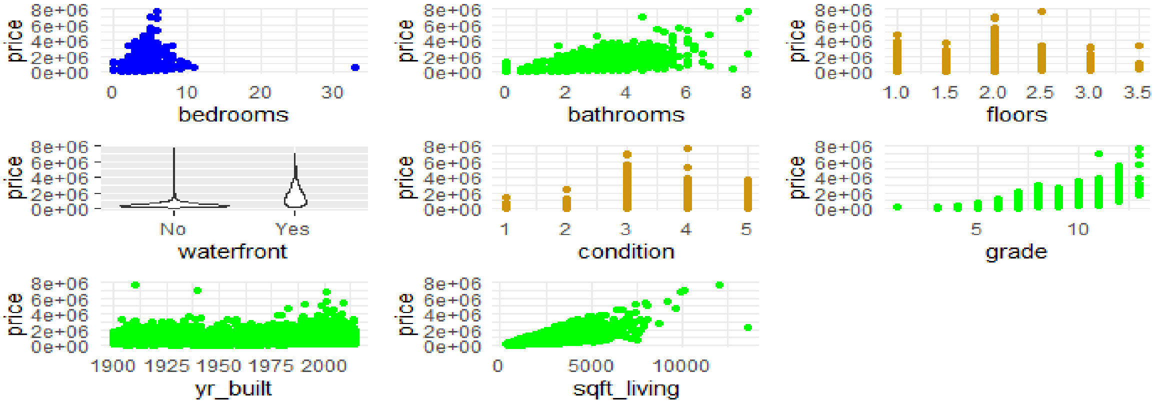


Figure 2 - Visualizations of price with other variables

Visualizations of the number of washrooms respective to other variables.

The graphs below show how the constructions of the number of washrooms respective to other variables look like.

When an analysis is done for the first two graphs below, it could be seen how constructions maintain their number of washrooms respective to number of bedrooms and area of living.

The first graph shows that when the number of bedrooms is high, the number of washrooms have increased. But there is an outlier in the graph. When the analysis is done, it shows that that particular house was built in 1947 so it is possible.

When an analysis is done for the number of washrooms with respect to area of living, grade and the year built, it could be seen that the number of washrooms have increased with their increase. But when the condition graph is analyzed, houses with a condition of 3 could be seen to have more washrooms and when the condition increased the number of washrooms have reduced. The same could be seen when the condition is below 3 as this too reduced the number of washrooms.

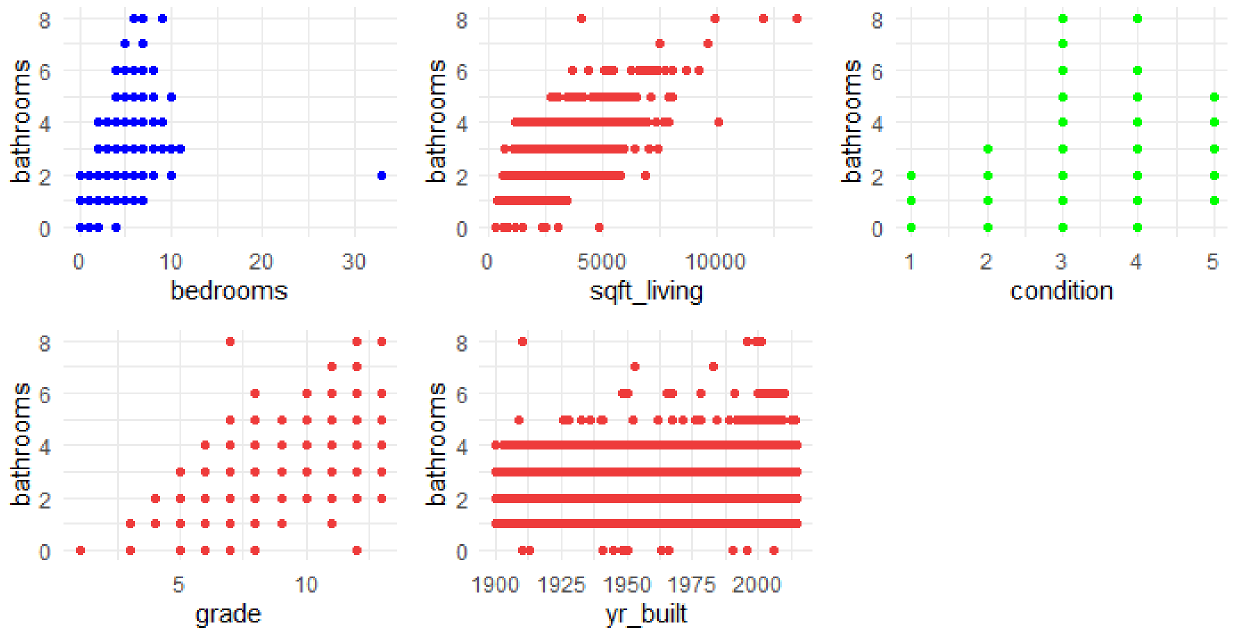


Figure 3 - Visualizations of bathrooms with other variables

Visualizations of living area respective to other variables.

The graphs below show the visualizations of living area respective to other variables.

With the scrutiny of the above mentioned graphs of living area respective to the number of bedrooms, bathrooms and grade it shows that with the increase of number of bedrooms, bathrooms and grade the living area has also increased.

The scrutiny of the number of floors and condition graphs shows that when the number of floors increased, the living area has decreased and the largest living area can be seen in houses having around two floors. Houses with a condition of 3 could also be seen to have more living area.

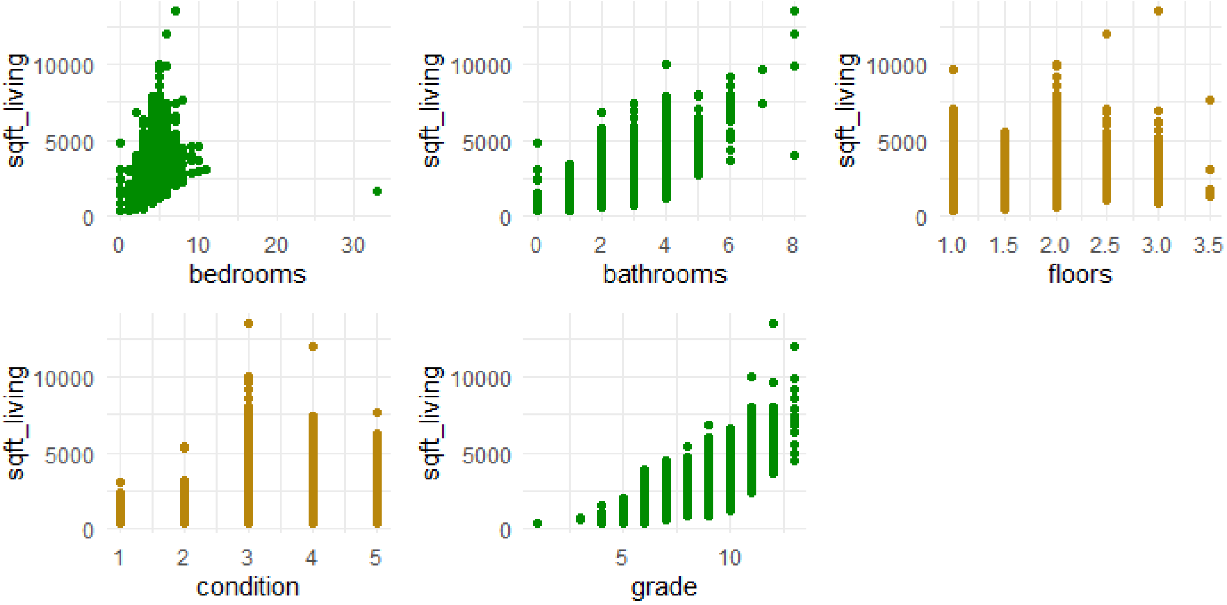


Figure 4 - Visualization of living area with other variables

Visualizations of the number of floors respective to other variables.

The graphs given below show the visualizations of the number of floors respective to other variables.

When the condition, grade and the year built graphs are looked at, they show that when the condition and grade are high, the floors numbers have increased, so the number of floors is the variable that has probably been used to measure the condition. Also it could be noted that when the year built is close to 2000, the number of floors have increased, an acceptable reason for which maybe the use of new technology.

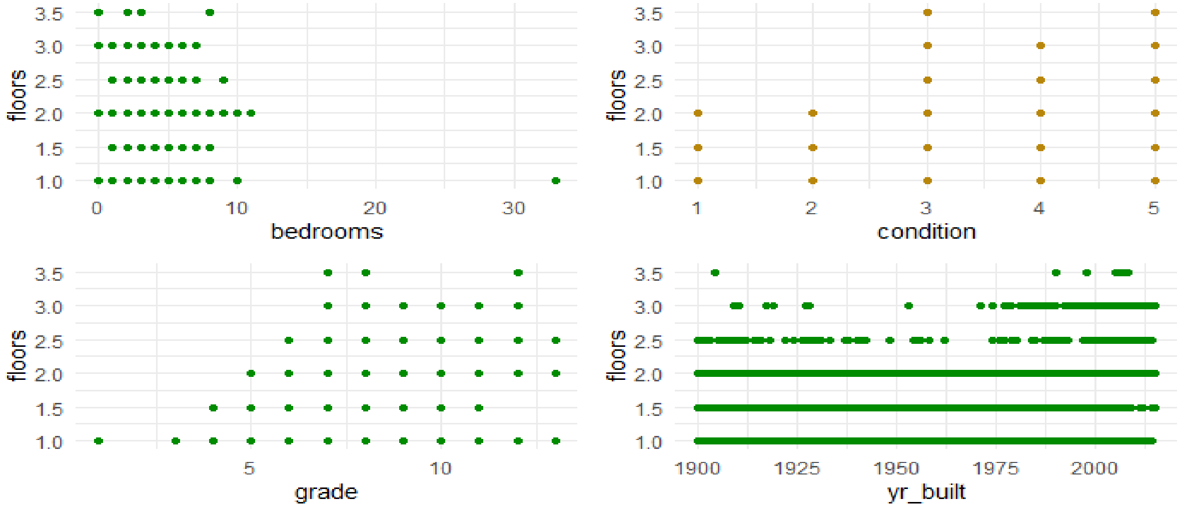


Figure 5 - Visualization of floors with other variables

Visualizations of waterfront respective to other variables.

According to the violin plots below, it could be seen that if a waterfront is present, the living area has increased. A look at the condition graph shows that when a waterfront is present, the house's condition has increased which can essentially be taken to mean that the presence of a waterfront adds value to a house.

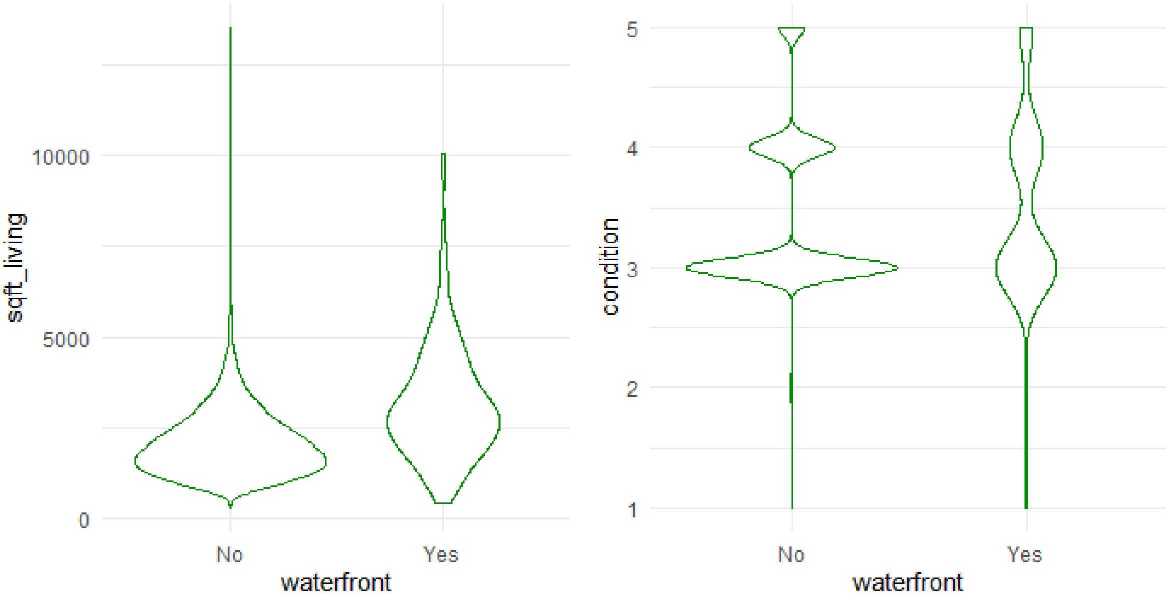


Figure 6 - Visualization of waterfront with other variables

Visualizations of waterfront respective to other variables.

With the scrutiny of the first graph below, when the grade is high, the condition is also high. Also when the year built is around 2000, the condition is comparatively high.

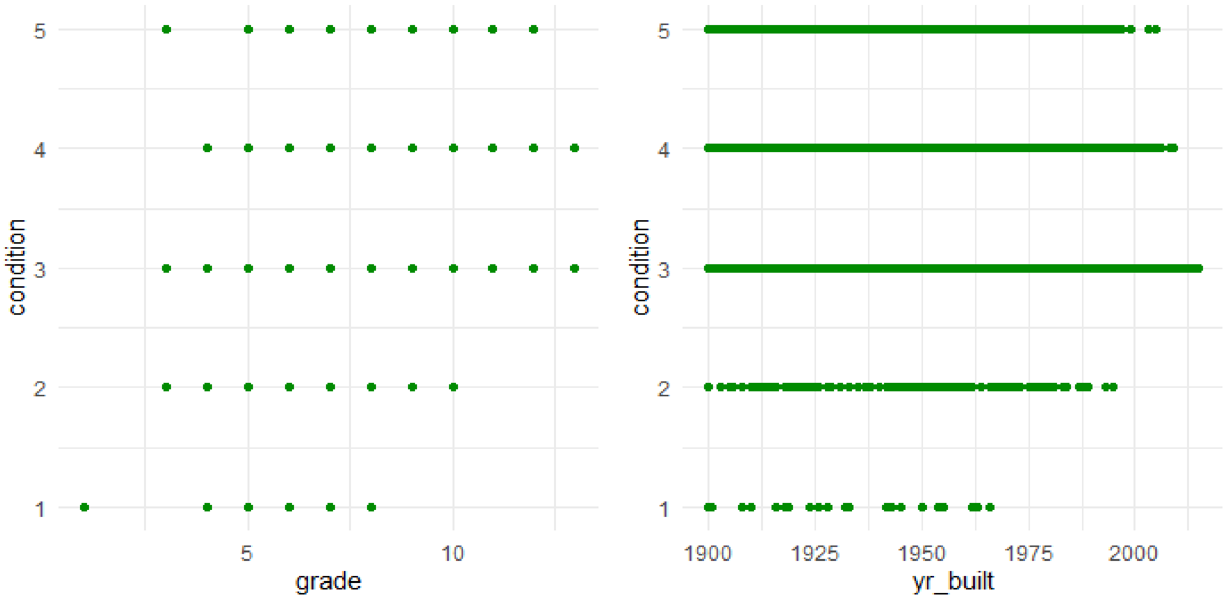


Figure 7 - Visualization of condition with other variables

Further visualizations

Since there were unrealistic values for floors, number of rooms and number of bedrooms, those values have been filtered before arriving at these illustrations. Also, the outliers in prices of houses have also been removed in order to obtain a better insight on the data.

Variable 1- Price

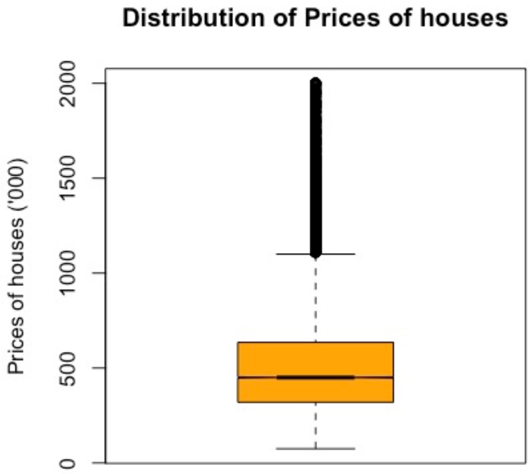


Figure 8 - Boxplot of price range

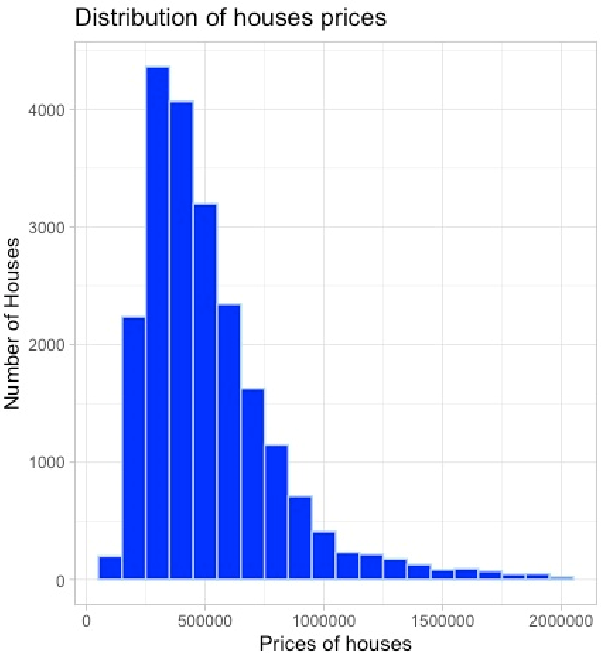


Figure 9 - Frequency of houses based on price

With the above illustrations (boxplot and histogram), the price range of the houses available can be clearly seen. There were houses beyond the price of 2 Million in the data set, but those were removed to have a better understanding on the data set, because they were outliers. Majority of the houses lie on the price range of 300,000 to 500,000. Median price of a houses is approximately 350,000.

Variable 2- Bedrooms

This illustration depicts the number of houses based on the availability of number of bedrooms in each house.

Majority of the houses are having 3 bedrooms. i.e. 9500 houses. There are few houses with very large number of bedrooms too. As an example, there is a house with 33 bedrooms in the data set. However, the most common number of bedrooms per house across the data set are 1 bedroom, 2 bedrooms, 3 bedrooms, 4 bedrooms and 5 bedrooms.

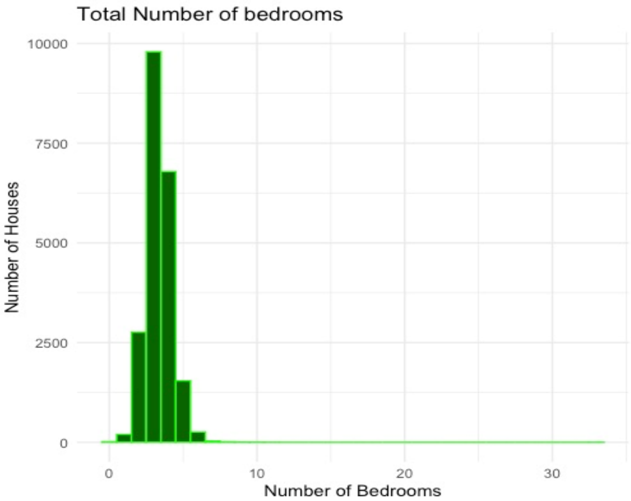


Figure 10 - Frequency of houses based on number of bedrooms

Variable 3- Bathrooms

This histogram depicts the number of houses based on the number of bathrooms per house. It is clear that most of the houses are having 2 bathrooms. But there are also outliers in the data set. For example, there is one house with 8 bathrooms. That was found in the house with 33 bedrooms. There are approximately 4000 houses with only one bathroom and approximately 2500 houses with three bathrooms and 1700 houses with 4 bathrooms and negligible number of houses with more than 4 bathrooms. In the given data set, there are negligible number of houses without any bathroom too. It is quite common that the number of bathrooms in a house would increase with the number of bedrooms as the number of people living in them would increase.

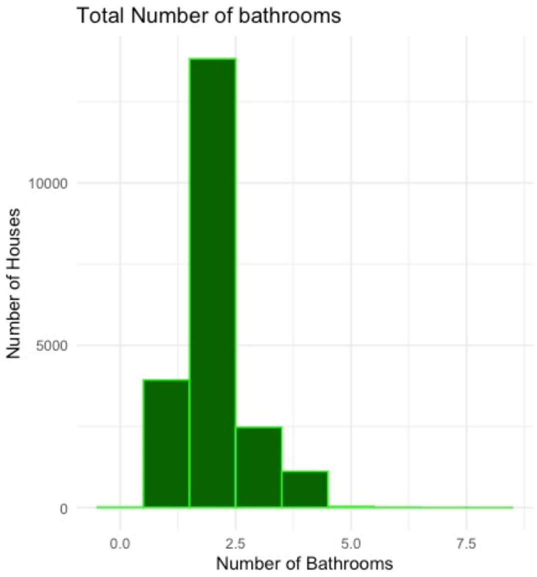
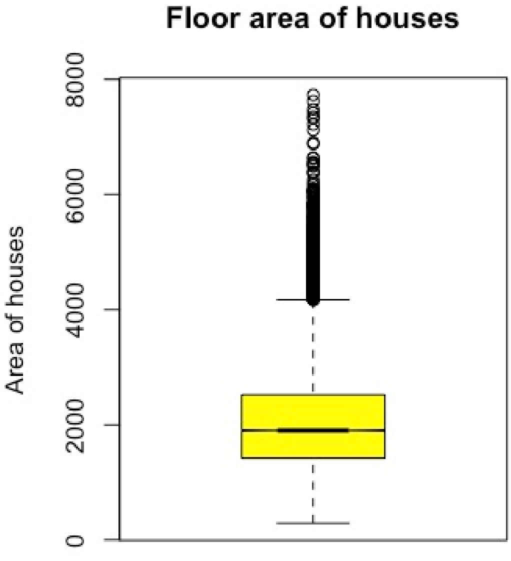


Figure 11 - Frequency of houses based on number of bathrooms

Variable 4- House Size



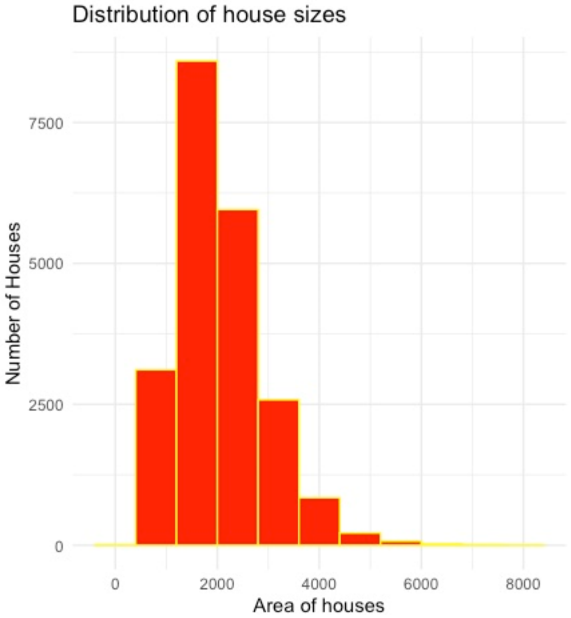


Figure 12 - Boxplot of floor area

Figure 13 - Frequency of houses based on floor area

The diagram on the left side depicts the distribution of the size of the room in square feet (sqft). Median size of a house is approximately 2000 sqft. There are more than 8000 houses with the size rounding to 2000 sqft. Also, there are negligible number of houses with huge size, more than 5000 sqft. Diagram on the right depicts the number of houses in each house size bracket. This illustration also confirms the median size of a house. The second most common house size bracket is between 2400 and 3200 sqft, which accounts to approximately 5500 houses.

Variable 5- Floors

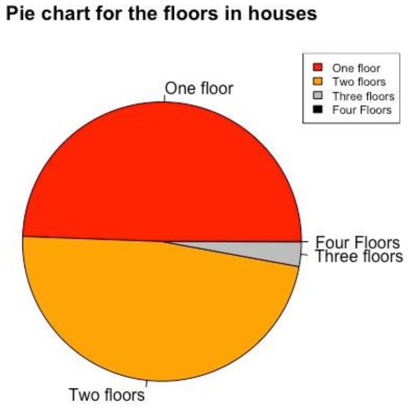


Figure 14 - Pie chart on number of floors

This pie chart illustrates the houses based on the number of floors. Majority of the houses are single floored, which accounts to roughly 50% of the total number of houses. And more than 45% of the houses are of two floors. The maximum number of floors in a house is 4. But, the ere are only a negligible number of houses with 4 floors. The balance among all the houses contain 3 floors., which accounts to approximately 4.5%.

Variable 6- Condition

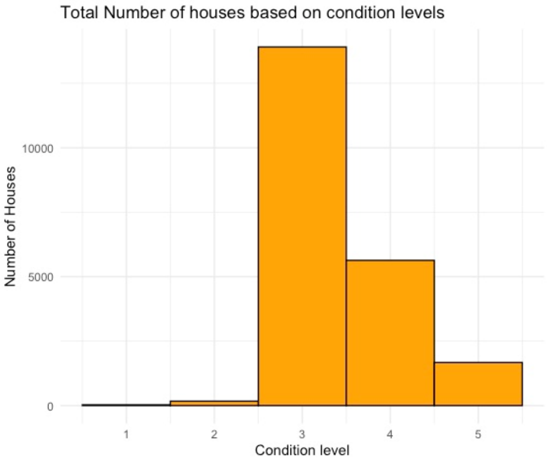


Figure 15 - Frequency of houses based on condition level

Condition level can be considered as a rating given for all the houses by a local authority. Necessary criterion might need to be satisfied to have the condition levels to achieve. Most of the houses are of condition level 3. It is more than 13000 houses. There are only a negligible number of houses with lower condition levels such as 1 or 2. All the other houses are having condition levels above 3. The maximum condition level permitted by the local authority might be 5(Houses with premium facilities). That may be the reason for us to not have any outliers above the value of 5.

Variable 7- Grade

Grade can be considered as a rating given for every house. Most of the houses are under grade 7 which approximates to more than 8000 houses. The second most frequent grade among the houses is 8. It is noticeable that almost all the houses are having grades between grade 4 and grade 11. Only a negligible number of houses are having grades beyond the range of grade 4 and grade 11.

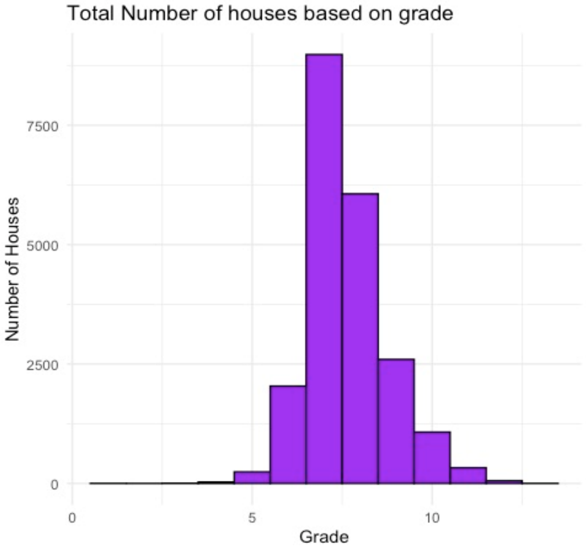
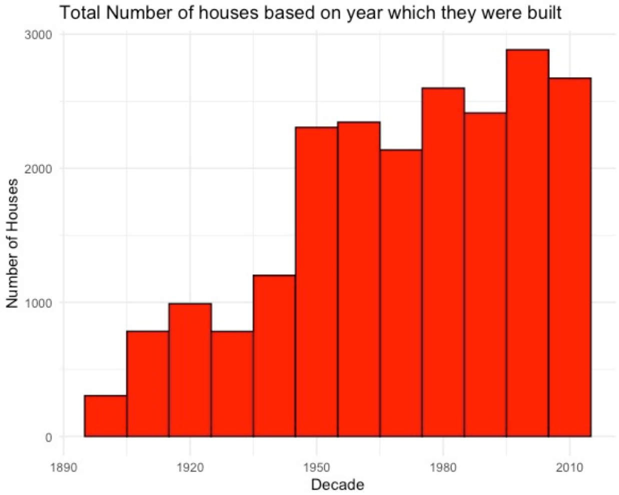


Figure 16 - Frequency of houses based on grade

Variable 8- Year Built



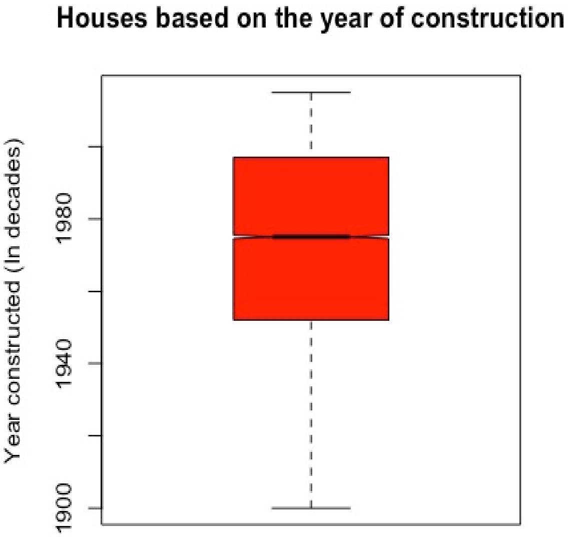


Figure 17 - Frequency of houses based on year they were built in

Figure 18 - Boxplot of houses based on year they were built in

These two illustrations depict the years in which the houses were built. Oldest house among the data set was constructed towards the end of the 19th century. The latest houses were built in 2010. Most of the houses were built in the last few decades of 20th century. i.e. Between 1950 and 2000. The median year which a house constructed is around year 1970.

Fitting the model for house price

The final goal of this entire report or most analytics project for that matter, is to create a model based on the data for the purposes of inference or prediction. The goal of this report—to be precise: is to create a model for the house price. Say, given respective values for an attribute *(**"features"** in machine learning terms)* of a certain hypothetical house, to predict its price.

Even though the R language was used to analyze the former part of this report, the rest will be modelled using the Python language. The usual stack of libraries (Numpy, pandas, Matplotlib and seaborn) used for data analysis is used here. And additionally, the popular machine learning library ***scikit-learn*** will be used to choose the algorithm from.

Modelling data primarily falls under two categories, i.e. classification and regression. A model for classification aims to predict a class label to which a data point belongs, and the classes have no continuity between them. A model for regression aims to predict a continuous value given a set of inputs based on a regression algorithm. Where this report is concerned, the goal is to predict the house price, given respective values for an attribute of a certain hypothetical house. Hence, this is a task for a regression algorithm.

All the code workings done in this chapter are available in the GitHub page created for this report. Please refer the code working in companion to the rest of the modelling process.

Taking a first look at the data suggests, that the features with an observable variation with price are bathrooms, square feet living and grade. However, with intuition and some further in-depth analysis, it was evident that classifying the data with respective to the waterfront feature (a house having or not having a water front), yields better results, which will be further elaborated in the latter part of the report.

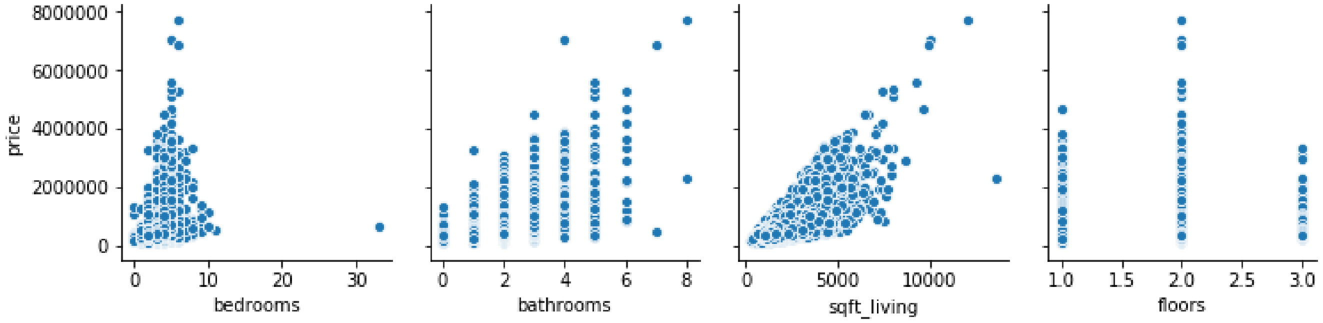


Figure 19 - Price vs. other variables (1)

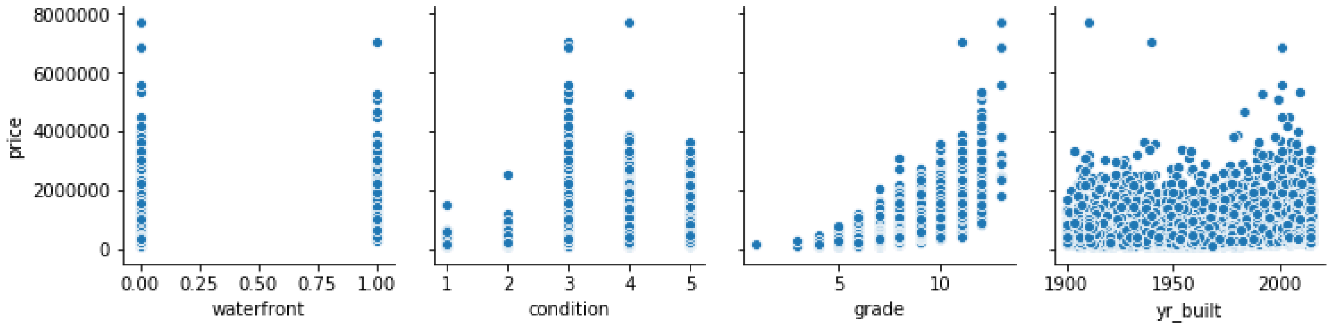


Figure 20 - Price vs. other variables (2)

The data was also split into the training and testing sets for the purpose of analyzing the ***generalized performance*** of the model as well as to identify whether the algorithm is ***"overfitting"*** or ***"underfitting".*** Any production ready model must generalize well, meaning – its predictions (or classifications) must have an acceptable accuracy on never before seen data *(new data on which the model was not trained on*). The only way to achieve this is to split the data we have into two sets: the training set and the test set, and train the model on the training data and test the accuracy of the model with the test data. One of the reasons *scikit-learn* was used for this report was that it provides functions and methods for all of these tasks. An in-depth explanation of overfitting and underfitting is out of the scope of this report. But a brief one would be: "*an overfitted model focuses too much on the training data"* and an "underfitted model focuses too little on the training data".

Mathematical representation of a multiple linear regression algorithm is as follows:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j .$$

The accuracy of the model is calculated by the residual sum of squares:

$$\begin{aligned} \mathrm{RSS}(\beta) &= \sum_{i=1}^{N} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2 . \end{aligned}$$

Fitting the model on a linear regression yielded the following results:

Weights or gradients of each feature (β_i):

```
[-3.96041604e+04  5.92646630e+04  1.72795058e+02  2.15157377e+04
  6.98616313e+05  1.65258301e+04  1.28021984e+05 -3.84904717e+03]
```

Intercept of the model (β_0):

```
6724065.73122059
```

Accuracy on the training data: `64%`
Accuracy on the testing data: `65%`

Aiming to further model accuracy

After further EDA (Exploratory Data Analysis) on the data the following pattern was uncovered.

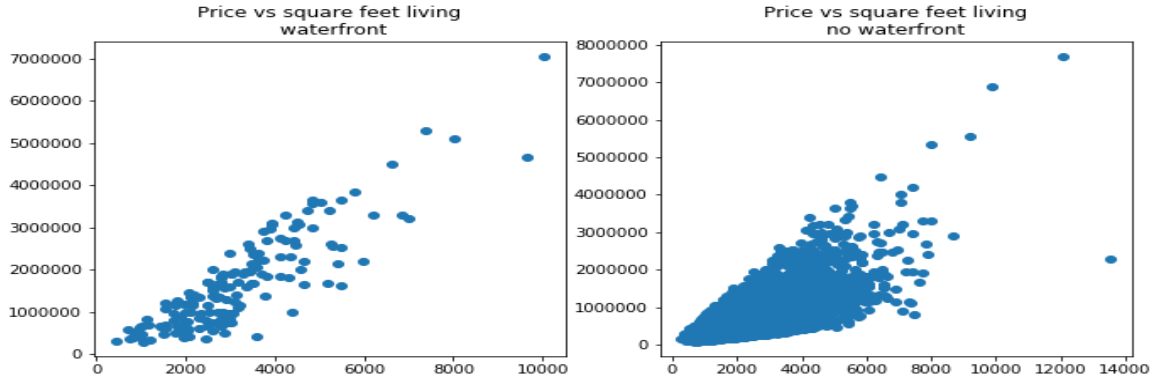


Figure 19 - Price vs. living area (differentiated with waterfront availability)

Houses with a waterfront (usually a beach or a lake) showed a close positive relation with available living space and the price of the property. This also makes intuitive sense as waterfront properties are usually considered luxury properties. And feeding the model based on this discreet classification yielded instantly better results.

Weights or gradients of each feature (β_i):

```
[ 2.11727725e+04  1.38609849e+05  4.82828272e+02 -3.33051381e+04
 -8.73114914e-11  1.38230014e+05  1.01218068e+05 -1.63501281e+03]
```

Intercept of the model (β_0):

```
1619031.15257297
```

Accuracy on the training data: `79%`
Accuracy on the testing data: `73%`

Observing the training and the testing accuracies, it is evident that the model is clearly *overfitting* as the accuracy of the testing data is far greater but the training data suffers (*the model is not generalizing well*). The remedy for this predicament is to **regularize** the model. Regularization refers to constricting the model such that it chooses weights as close to zero as possible. Such a linear model is Ridge regression, where the strength of the regularization is given by the *alpha* parameter.

Ridge regression ('subject to' provides the regularization parameter):

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2,$$

$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 \leq t,$$

Measure of accuracy for Ridge regression:

$$\text{RSS}(\lambda) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta,$$

Applying Ridge regression with different regularization strengths (alpha values) yielded the following results:

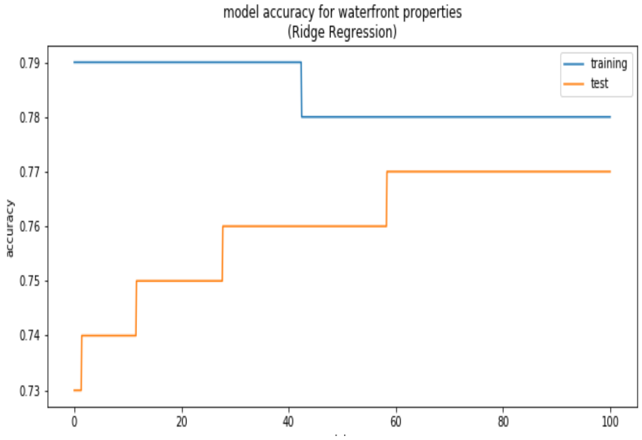


Figure 20 - Ridge regression for properties with waterfronts

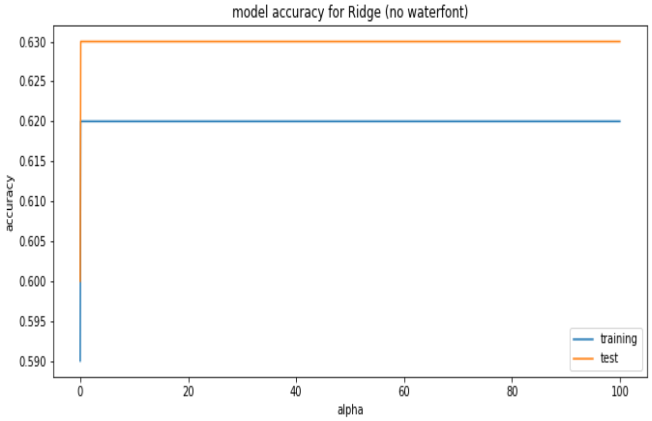


Figure 23 – Ridge regression for properties with no waterfronts

Finally, it is evident that *waterfront model* takes a district improvement from regularizing whereas the *no waterfront model* didn't see an improvement whatsoever.

Therefore, as the model with the best accuracy is Ridge regression, we will choose this as our algorithm for the final model (with an alpha=100).

Final model for water front properties:

```
Weights: [12845.7329075  57306.39134966   547.25284937  1919.21312129
          0.          57748.86588098 58568.80316605  -828.37504575]

Intercept: 652781.782596465

Training accuracy: 78%
Testing accuracy: 77%
```

For example, the model could predict the price of a house with 3 bedrooms, 3 bathrooms, 1000 square feet living space, 2 floors, with water front, condition and grade of 9 and 3 respectively built in 2008 to cost 446398.54 dollars with 77% accuracy.

Final model for no water font properties:

```
Weights: [-35128.6495305   51075.14152215    163.14405638  25527.46097554
          0.           19417.40794916 130900.6035829   -3797.65319682]

Intercept: 6603712.982146138

Training accuracy: 62%
Testing accuracy: 63%
```

For example, the model could predict the price of a house with 3 bedrooms, 3 bathrooms, 4000 square feet living space, 2 floors, no water front, condition and grade of 9 and 3 respectively built in 2008 to cost 296954.46 dollars with 63% accuracy.

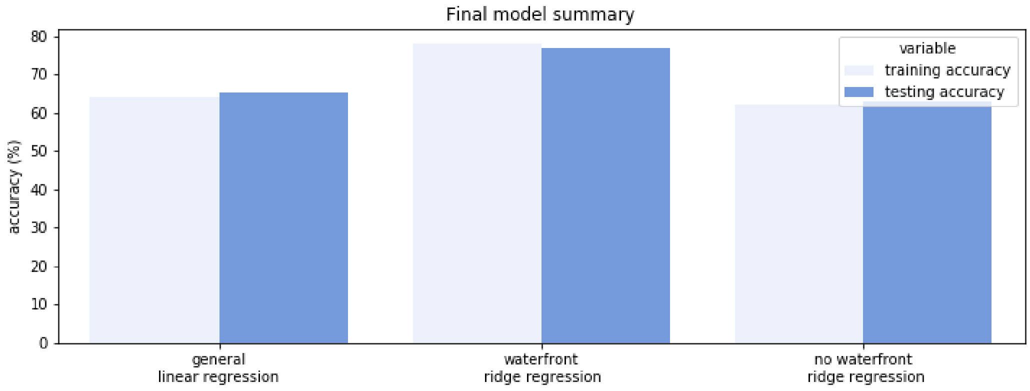


Figure 21 - Final model summary

Outcome

Computation of descriptive statistics were done with the aid of the statistical programming language R. Descriptive statistics provided the foundation to grasp the statistical parameters of the data, thereby providing the baseline for understanding the variations and correlations among the features. Exploratory data analysis in terms of single variables and inter-variable relationships were also performed using the R language and the outcome of this was to gain a deeper visual understanding of the data. The final "house price model" outcome is a machine learning Ridge regression model which predicts the price of a hypothetical house given its respective features. However, it is important to note that two models were produced in the process with the aim of maximizing the model accuracy. It is crucial that the correct model is implemented according to the situation. The regression model was computed with the aid of the Python programming language and its usual stack of libraries used for data analysis and machine learning, i.e. 'NumPy', 'pandas', 'Matplotlib' and 'scikit-learn'. Note that additionally the Python statistical visualization package 'seaborn' was also used for some visualizations.

In a classical setting of real estate pricing, the agent would use a real estate appraiser to appraise and calculate the market value of a house and the interested buyer usually negotiates a price with the real estate agent. This approach to pricing is both time and cost ineffective to both the buyer and the seller. In the aspect of the seller, real estate appraisers cost a lot and the process could take weeks and at best days. In terms of the buyer, they would either have to conduct their own research or hire a trained professional to do it for them. Scaling this process in terms of buyer and sellers for multiple houses with multiple buyers and sellers, the time and capital spent is enormous.

The model's outcome is to eliminate this completely, as the results are instant and cost is considerably low and as the model's approach scales unlike an appraiser, who needs to be paid for every appraisal or be paid on a regular basis. The model is a one-time investment thereby both the agent and the buyer have the ability to take advantage of economies of scale. The production models could be kept up to date with new data and with having different price options and price ranges like; one-time use, weekly, monthly, yearly payments and full purchase as well would provide the customer with the necessary flexibility and the motivation to purchase causing maximum industry disruption. The model could also be used by financial institutions like banks and any other party interested in pricing a house.

Conclusion

The model created with intuition gained by descriptive statistics as well as visualizations serves as an excellent substitute for the classical methods of pricing a house, however the model is not without its shortcomings. The maximum model accuracy obtained was 77% which might not be adequate for certain institutions as well as certain scenarios. This could be attributed to the fact that data available was of poor quality, such as the number of bathrooms and floors having point values and features that are known to have a big impact on house price such as location information (crime rates, resident ethnicities, proximity to cities and other public infrastructure and pollution indexes) not being available in the data. The model would perform fine in most situations with satisfactory results but it would benefit greatly with better and appropriate data.

References

Hastie, T., Tibshirani, R. and Friedman, J., 2016. *The Elements Of Statistical Learning*. 2nd ed. Springer.

Kaggle.com. 2020. *House Sales In King County, USA*. [online] Available at: <https://www.kaggle.com/harlfoxem/housesalesprediction> [Accessed 11 July 2020].

Python.org. 2020. *Welcome To Python.Org*. [online] Available at: <https://www.python.org/> [Accessed 15 July 2020].

Cran.r-project.org. 2020. *The Comprehensive R Archive Network*. [online] Available at: <https://cran.r-project.org/> [Accessed 16 July 2020].

Rstudio.com. 2020. [online] Available at: <https://rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf> [Accessed 21 July 2020].

Datanovia. 2020. *How To Combine Multiple Ggplots Into A Figure - Datanovia*. [online] Available at: <https://www.datanovia.com/en/lessons/combine-multiple-ggplots-into-a-figure/> [Accessed 22 July 2020].

Numpy.org. 2020. *Numpy*. [online] Available at: <https://numpy.org/> [Accessed 23 July 2020].

Sthda.com. 2020. *Ggplot2 - Easy Way To Mix Multiple Graphs On The Same Page - Articles - STHDA*. [online] Available at: <http://www.sthda.com/english/articles/24-ggpubr-publication-ready-plots/81-ggplot2-easy-way-to-mix-multiple-graphs-on-the-same-page/> [Accessed 24 July 2020].

Pandas.pydata.org. 2020. *Pandas Documentation — Pandas 1.1.0 Documentation*. [online] Available at: <https://pandas.pydata.org/docs/> [Accessed 23 July 2020].

Scikit-learn.org. 2020. *Scikit-Learn: Machine Learning In Python — Scikit-Learn 0.23.2 Documentation*. [online] Available at: <https://scikit-learn.org/stable/> [Accessed 26 July 2020].

Lowe, L., Sohil, A., Meneripitiya, S. and Maneth, P., 2020. *Statisticsenglishassignment - Overview*. [online] GitHub. Available at: <https://github.com/StatisticsEnglishAssignment> [Accessed 6 August 2020].