A Project Report on

Data Storytelling and Communication

Under supervision of Prof. Dr. Swati Chandana,
Professor, SRH university

\

Submitted by
Medisetti Narendra, Matrikel Number: 11012094
M.Sc in Big data & Business Analytics, SRH University, Heidelberg

Abstract:

Our aim of this project is to analyse House Prices data and predict the better explainability methods and visualization.

In the academic talk there is a far-reaching accord that house estimation changes considerably influence cash related quality and genuine financial action. Therefore, it is important to have timely information on House prices increase every year, so there is a need for a system to predict house prices in the future. The aim of this project is to measure the house price movements in real estate markets in United States of America and forecast near-term price developments. Our framework gives a conclusive dwelling place value forecast model to profit a purchaser and dealer or a real estate to settle on a superior educated choice framework on different highlights. This project utilizes both the relating to pricing model (Continuous data). This project used various machine learning algorithms, such as linear Regression, Random Forest (RF), boostting techniques like XGboost, Catboost to predict house prices.

For the House Price dataset, I use Python Programming techniques to predict the house prices in The United State of America

Introduction:

Today world the further era of big data, more and more people begin to engage in data analysis and mining. Machine learning is a common mean for data analysis, has been more and more attention.

The real estate market is rapidly evolving. A recent report published. Estimates the size of the professionally managed real estate investment market in $9 trillion in 2018, increasing a total of $1.1 trillion since the previous year. Of course, the real market size is expected to be much larger when counting assets which are not professionally managed or that are not object of investment. However, looking at the market evolution from a global perspective turns out to be too simplistic. Although the market at a global scale is very tightly correlated, there are many aspects influencing the behaviour of markets at a local scale, such as political instability or the emergence of highly demanded “hot spots” that can shift rapidly. In addition, different market segments evolve at different places, such as high-end luxury houses.

The housing market in the United of America has always been a topic of national attention, with news sites dedicating sections that report on key news that can potentially affect housing prices and the trends in recent months. Not only are the trends in housing market of concern

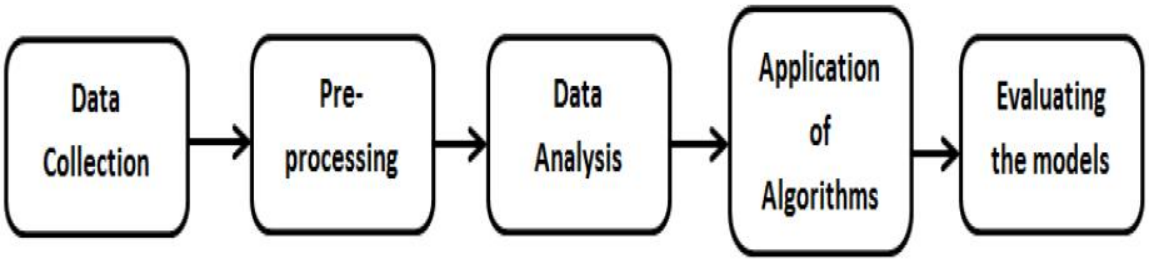
to buyers and owners, they reflect the current economic situation and social sentiments in the country. For many people, buying a property is one of the most important decision and purchase in life. Besides the affordability of a house, other factors such as the desirability of the location and the long-term investment prospects also affect the decision-making process. This project justifies the value of houses to the customer who want buy and sell their houses in the United States of America.

Software requirement:

- Python Programming language.
- MS Excel: To Import, Export and Analyse.
- Tableau: It's a visualization tool, to present the insights pictorially.
- Packages: Pandas, Numpy, Seaborn, Matplotlib, Sklearn.

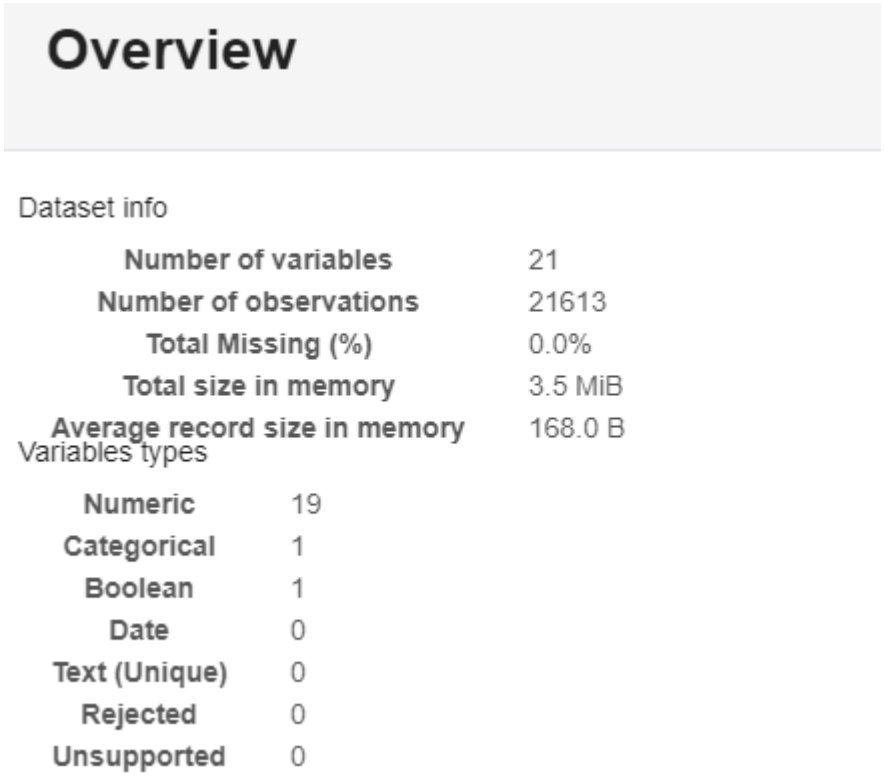
Methodology:

Methodology represents a description about the framework that is undertaken in project. It consists of various milestones that need to be achieved in order to fulfil the objective. We have undertaken different data mining and machine learning concepts. The following diagram (figure 1) represents step-wise tasks that need to be completed.



Data Collection:

The dataset used in this project was an open source dataset from Kaggle website. It consists of 21613 rows and 21 features that have the possibility of affecting the property prices. It contains 19 numeric and categorical variables.



Overview

Dataset info

Number of variables	21
Number of observations	21613
Total Missing (%)	0.0%
Total size in memory	3.5 MiB
Average record size in memory	168.0 B

Variables types

Numeric	19
Categorical	1
Boolean	1
Date	0
Text (Unique)	0
Rejected	0
Unsupported	0

Following table represent a brief description about most important parameters that affect the selling price of the house.

Features	Description	Datatype
ID	It is the unique numeric number assigned to each house being sold	Integer
Date	It is the date on which the house was sold out	Object
Price	It is the price of house.	Float
Bedrooms	It determines number of bedrooms in a house.	Integer
Bathrooms	It determines number of bathrooms in a bedroom of a house	Float

Sqft_living	It is the measurement variable which determines the measurement of house in square foot.	Integer
Sqft_lot	It is also the measurement variable which determines square foot of the lot.	Integer
Floors	It determines total floors means levels of house.	Float
Waterfront	This feature determines whether a house has a view to waterfront 0.	Integer
View	This feature determines whether a house has been viewed or not.	Integer
Condition	It determines the overall condition of a house on a scale of 1 to 5.	Integer
Grade	It determines the overall grade given to the housing unit, based on King County grading system on a scale of 1 to 11.	Integer
Sqft_above	It determines square footage of house apart from basement.	Integer
Sqft_basement	It determines square footage of the basement of the house.	Integer
Yr_built	It determines the date of building of the house.	Integer
Yr_renovated	It determines year of renovation of house.	Integer
Zip Code	It determines the zip code of the location of the house.	Integer
Latitude	It determines the latitude of the location of the house.	Float
Longitude	It determines the longitude of the location of the house.	Float
Sqft_living15	Living room area in 2015(implies-some renovations)	Integer
Sqft_lot15	lot Size area in 2015(implies-- some renovations)	integer

The over quality to rate the house overall condition and price of the house, Location, which Year house was built and year of renovated, the house contains a number of Bedrooms and bathrooms, Living area and basement. The price at which house is sold in particular location. There is condition feature which will rate the building. Price feature is a dependent variable on several other independent variables like bedroom, living area, bathroom... etc. Some parameters had numerical values and other object which is converted into categorical variables. Following integer and categorical show a most important features that affect the price of the house.

Data Pre-processing:

It is a process of transforming the raw, complex data into systematic understandable knowledge. It involves the process of finding out missing and redundant data in the dataset. Entire dataset checked for NaN and whichever observation consists of NaN will be deleted. Thus, this brings uniformity in the dataset. However, in our dataset, there was no missing values found meaning that every record was constituted its corresponding feature values.

I have compared my dataset with the actual ground truth (http/info.kingcounty.govs:/) in order to come to a conclusion that I am working on the correct data set.

However, I have invested brief time on Trifecta to find the incorrect data, fortunately the data set was almost achieved the goal of uniqueness, but It has helped to observe the patterns of each column. Which eventually played a vital role.

bedrooms
Numeric

Distinct count	13
Unique (%)	0.1%
Missing (%)	0.0%
Missing (n)	0
Infinite (%)	0.0%
Infinite (n)	0
Mean	3.3708
Minimum	0
Maximum	33
Zeros (%)	0.1%

Data Analysis:

Before applying a model to dataset, we have to analyse the characteristics of our dataset. Thus, we need to analyse our data, study the different features and relationship among the columns.

We can also find out the outliers present in our dataset. Outliers occur due to some kind of experimental errors and they need to be excluded from the dataset. Data profiling help to understand the dataset better by using statistics, histogram, finding a relation between the depended and Independ variable and Normalize data.

Statistics view:

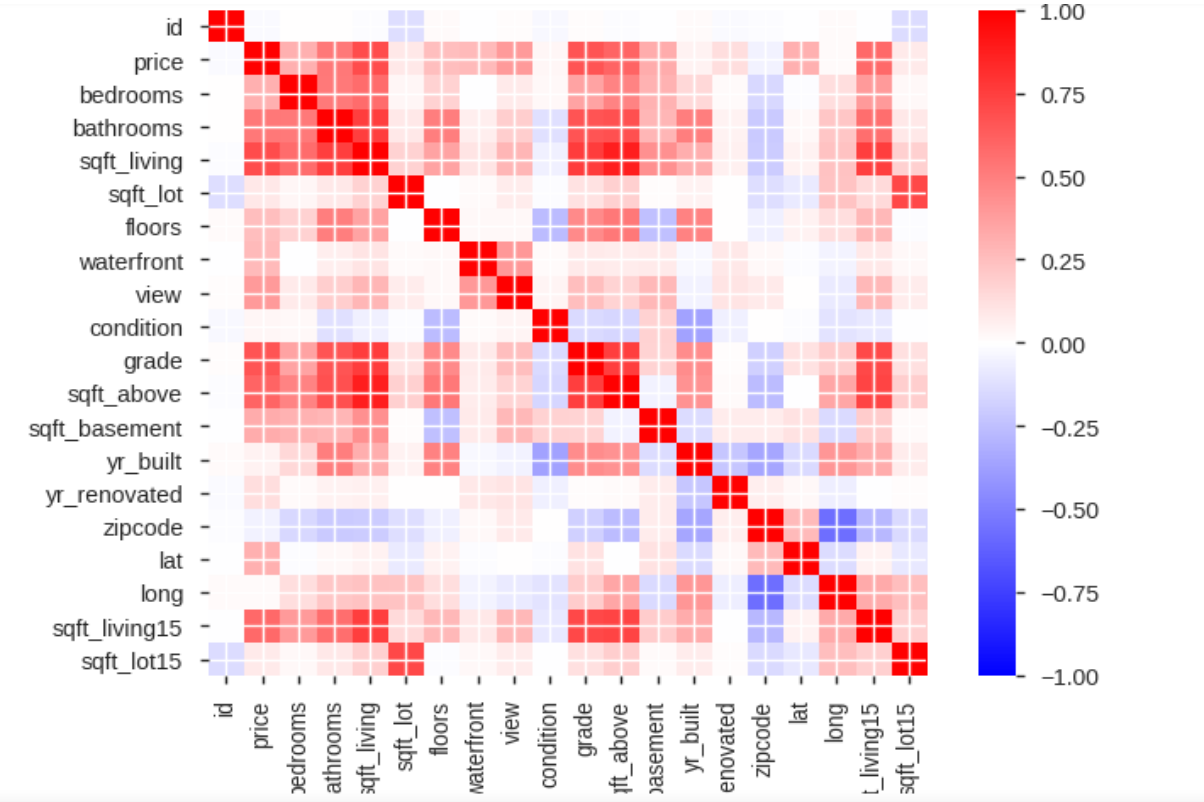
Minimum	1
5-th percentile	3
Q1	3
Median	3
Q3	4
95-th percentile	5
Maximum	5
Range	4
Interquartile range	1

Descriptive statistics

Standard deviation	0.65074
Coef of variation	0.19087
Kurtosis	0.52576
Mean	3.4094
MAD	0.56072
Skewness	1.0328
Sum	73688
Variance	0.42347
Memory size	168.9 KiB

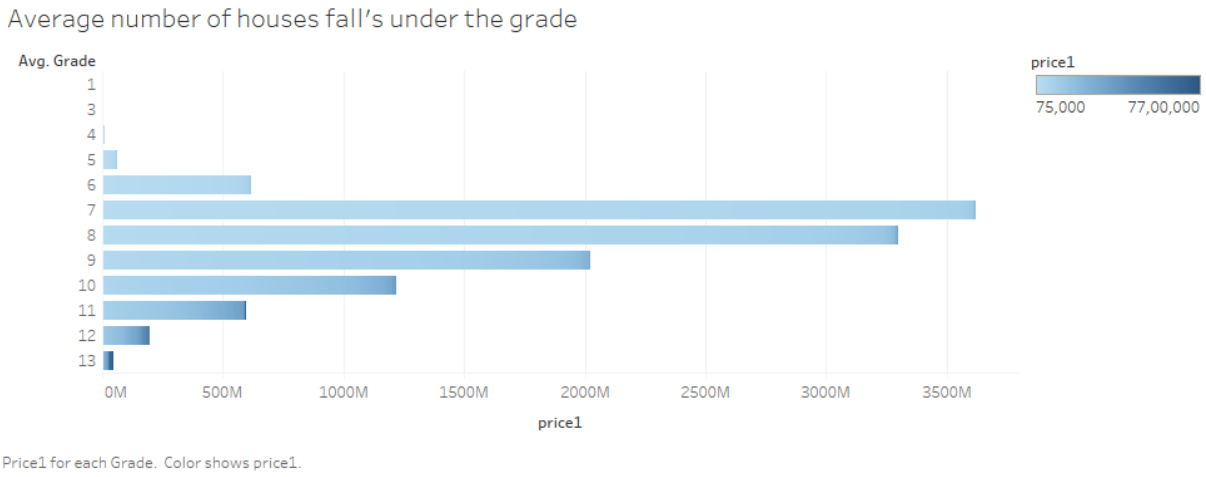
Correlation:

The correlation matrix helps to identify the relation between two features which has a strong feature dependence one from another (darkest colours - the positive correlation, lightest colours - the negative). A correlation number gives the degree of association
between two variables. The correlation number exists between +1 to -1. It means serious difficulties in improving the predictions based on this data. However, for such data sets, we have the ability to reduce the dimensionality. In this project the target variable is price which has a lower correlation in independent variables like condition, date, zip code, longitude, sqft_living15 and remaining independent variable has a higher correlation. This is useful to know, because some machine learning algorithms like linear regression can have poor performance if there are highly correlated input variables in your data.

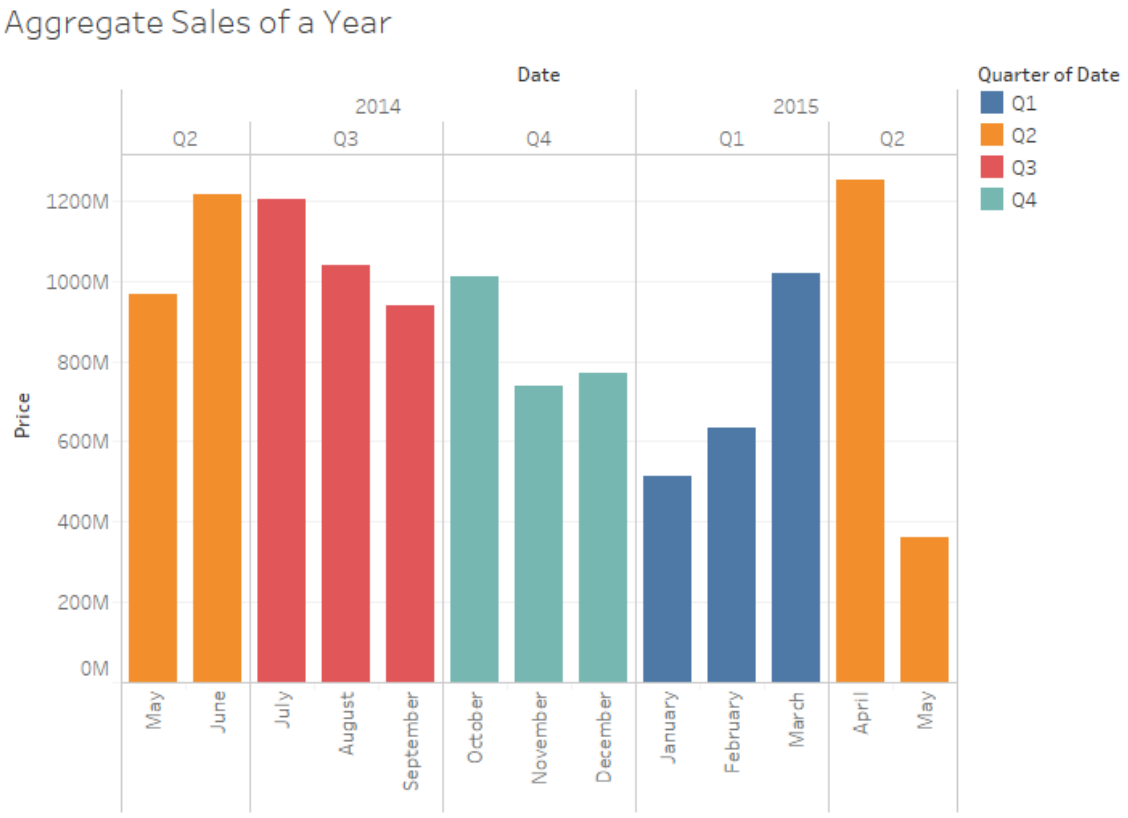


Average number of housing units under the grade.

It determines the overall grade given to the housing unit, based on King County grading system on a scale of 1 to 11.



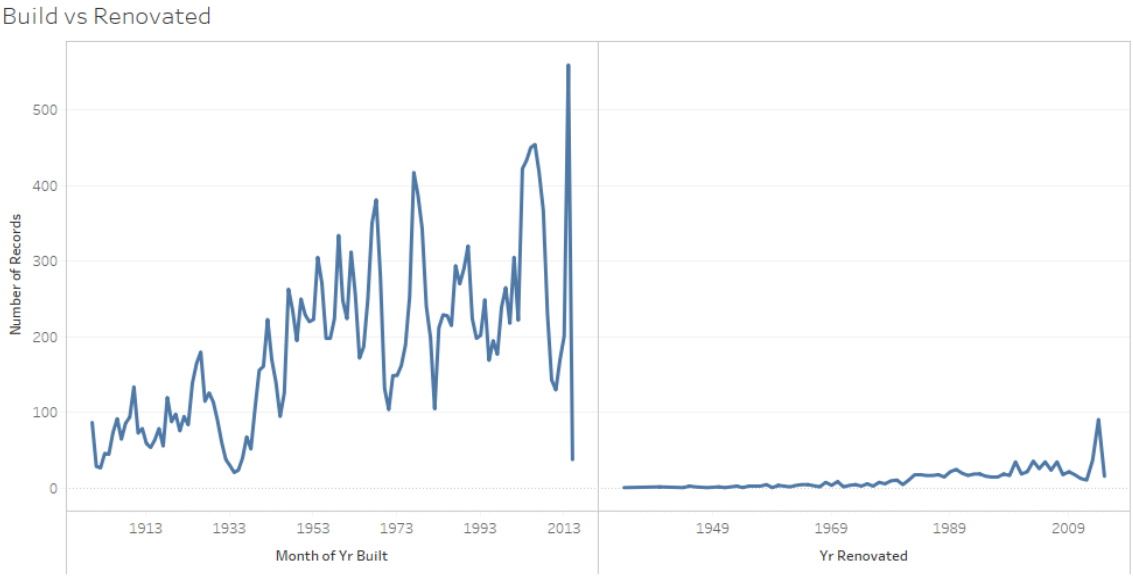
Annual sale of houses in The United Stated of America



Sum of Price for each Date Month broken down by Date Year and Date Quarter. Color shows details about Date Quarter.

The above graph shows the aggregate sales of a year. However, we cannot come to a conclusion based on the yearly data, but the actual figure states Q2 and Q3 has the greatest number of sales compared to Q1 and Q4.

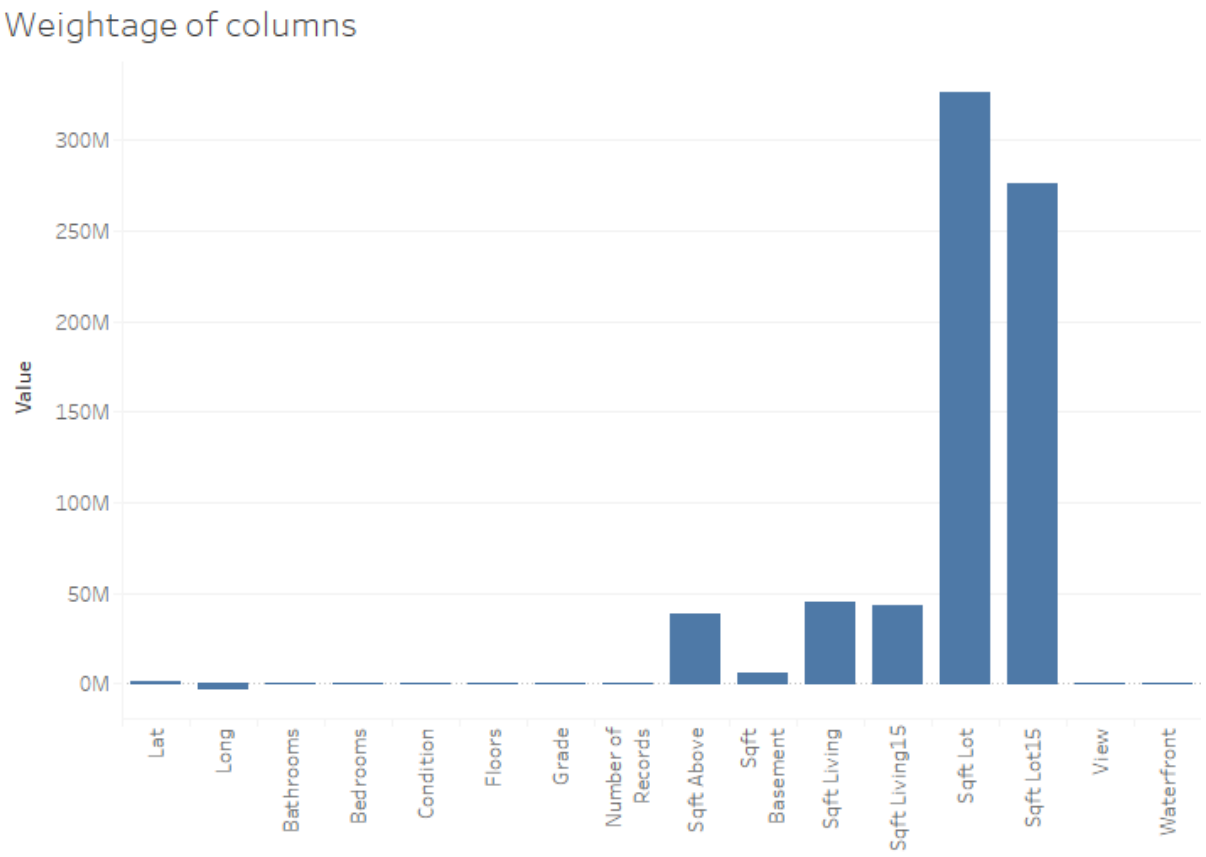
Build vs Renovation



The trends of sum of Number of Records for Yr Built Month and Yr Renovated. The data is filtered on Yr Renovated and Yr Built Year. The Yr Renovated filter excludes #Error. The Yr Built Year filter keeps 116 of 116 members.

The figure clearly illustrates renovation of the houses till 2009 was very minimal and it took a steep after it followed by the declined fashion. The pattern of construction of houses has irregular fashion, which has no specific demonstration to identify special insights. Overall, Renovation of houses has less priority in comparison with the construction.

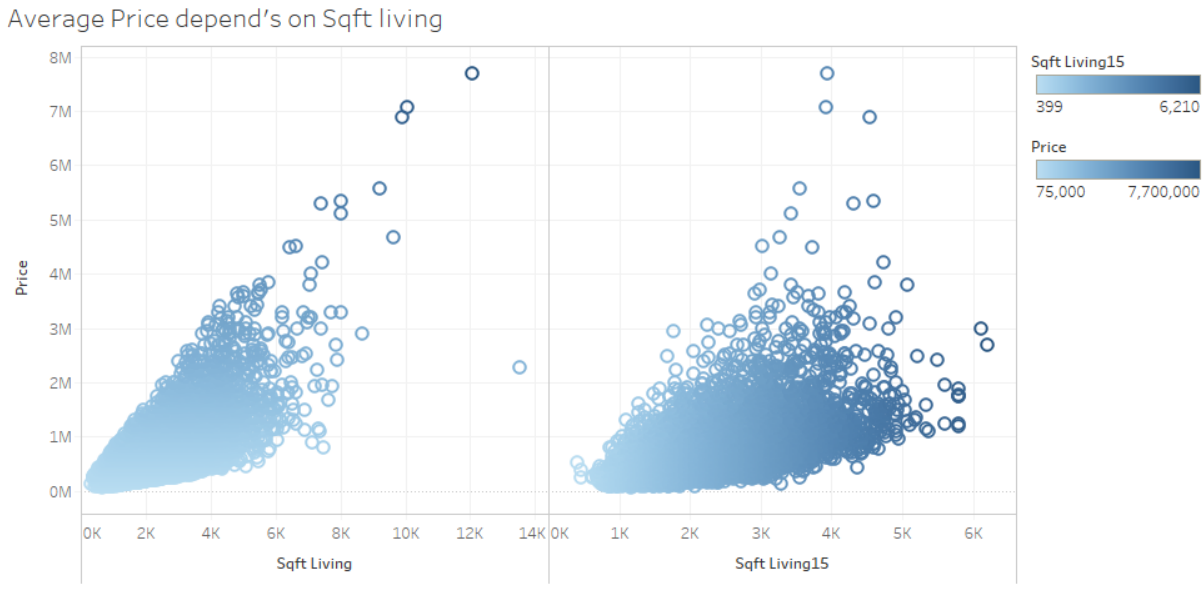
Weightage of Features



Avg. Lat, Avg. Long, Bathrooms, Bedrooms, Condition, Floors, Grade, Number of Records, Sqft Above, Sqft Basement, Sqft Living, Sqft Living15, Sqft Lot, Sqft Lot15, View and Waterfront.

The above figure shows which has more weightage in order to predict the prices of the house. We can clearly see that Square foot plays vital role and dominates the rest of the files. Overall, here we can make a statement that price is directly proportion to the sqft living.

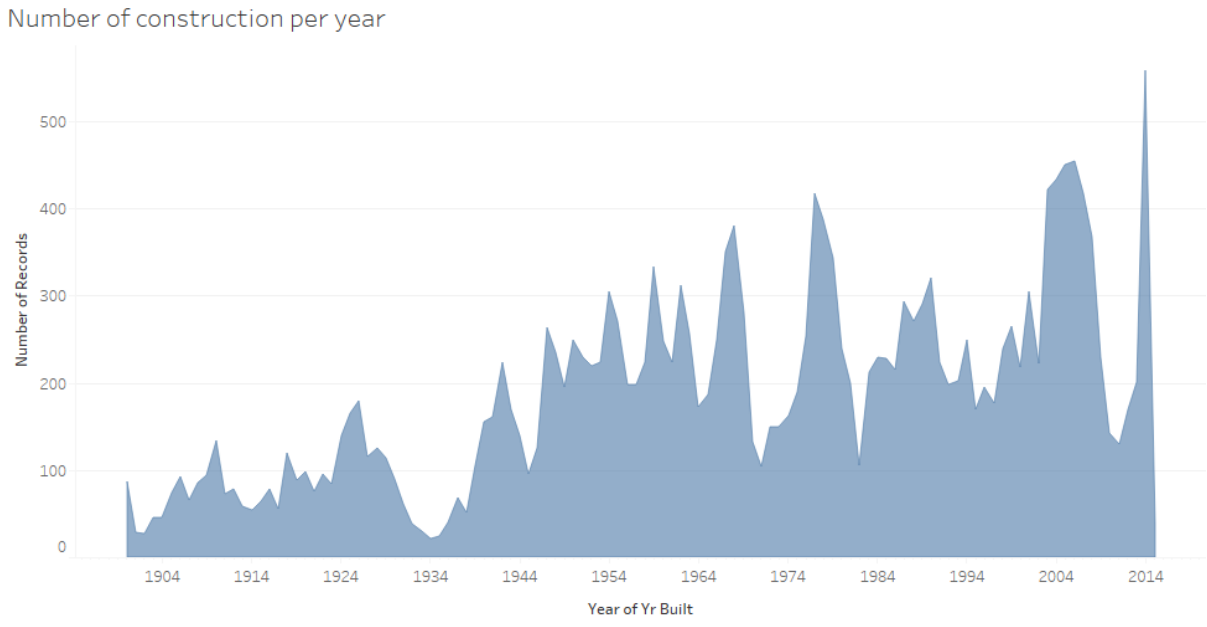
Average price depends on sqft living



Sqft Living and Sqft Living15 vs. Price. For pane Sqft Living: Color shows Price. For pane Sqft Living15: Color shows Sqft Living15.

The above figure shows the average price depending on Sqft living. As I mention earlier price is directly proportion to the square foot. The two pictures show the major difference of price and the sight increase in the size of the rooms (Sqft) before and after the renovation.

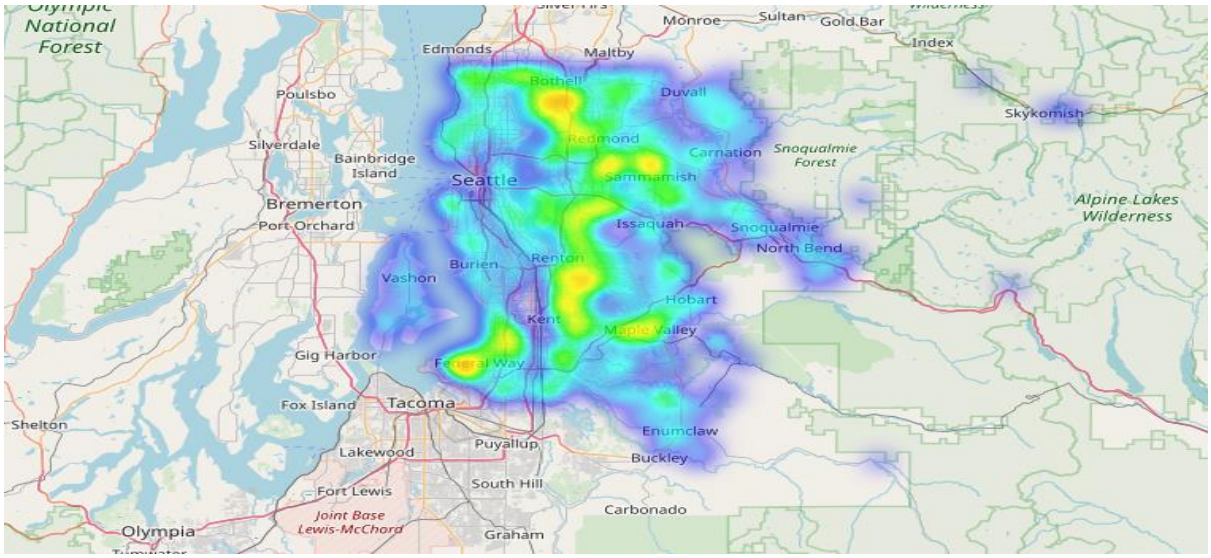
Number of constructions per Year:



The plot of sum of Number of Records for Yr Built Year.

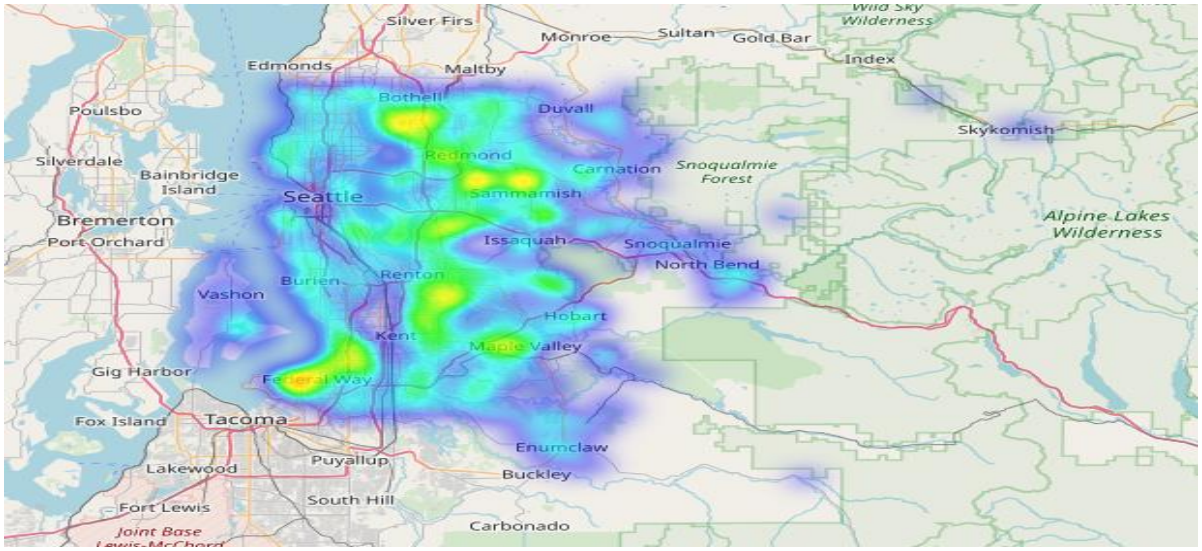
The above graph shows the evolution of the construction over decayed. It's quite obvious to have an inclined fashion as the population grows the number of construction of buildings spikes. Ironically, during the beginning of 1930's, end of 1960's and beginning of 2010's has steep declined fashion.

1940-1960



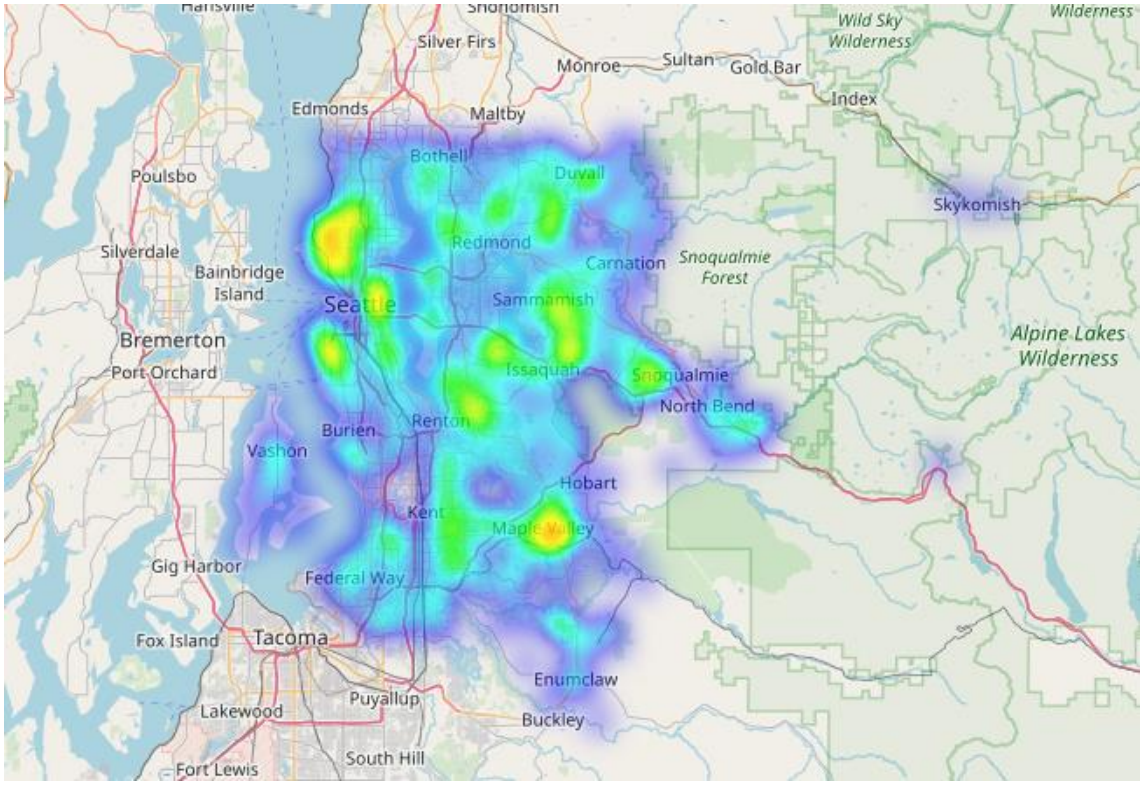
The above are heatmaps shown between the interval of 30 years. In comparision with one another , there was no much difference between them , but the boundaries have picked up the density which implies the growth of the construstion has incresed gradually. Here we can see Skykomish has contain soe datapoints , which we can considered has outlayer,but as my model is only predicting the prices the location of the building has less priority.I have corss verifed the data which doesn't has any abnormlities ,tho I have proceded with the further analysis but applying meachin learing algorithams.

1960 – 1990



The above are heatmaps shown between the interval of 30 years. In comparision with one another , there was no much difference between them , but the boundaries have picked up the density which implies the growth of the construstion has incresed gradually. Here we can see Skykomish has contain soe datapoints , which we can considered has outlayer,but as my model is only predicting the prices the location of the building has less priority.I have corss verifed the data which doesn't has any abnormlities ,tho I have proceded with the further analysis but applying meachin learing algorithams.

1990 – 2015



The above are heatmaps shown between the interval of 30 years. In comparision with one another , there was no much difference between them , but the boundaries have picked up the density which implies the growth of the construstion has incresed gradually. Here we can see Skykomish has contain soe datapoints , which we can considered has outlayer,but as my model is only predicting the prices the location of the building has less priority.I have corss verifed the data which doesn't has any abnormlities ,tho I have proceded with the further analysis but applying meachin learing algorithams.

Application of Algoritham:

After completing the data pre-processing and data analysis, Next step is to apply an appropriate machine learning model that fits our dataset. The data contains continuous value. so, I choose a regression algorithm. I have selected four algorithms to predict the dependent variable in our dataset. Next training the model to predict the continuous values. The four

algorithms are Linear Regression, Random forest, XGboost, Catboost. These algorithms were implemented with the help of python's SciKit-learn Library. The predicted outputs obtained from these algorithms were saved in comma separated value file. This file
was generated by the code at run time.

1. Linear Regression:

linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship.

$$y_i = \alpha + \beta x_i + \varepsilon_i .$$

It determines how much of the total variation in Y (dependent variable) is explained by the variation in X (independent variable). Mathematically, it can be written as:

$$R - Square = 1 - \frac{\sum(Y_{actual} - Y_{predicted})\text{^}2}{\sum(Y_{actual} - Y_{mean})\text{^}2}$$

The value of R-square is always between 0 and 1, where 0 means that the model does not model explain any variability in the target variable (Y) and 1 meaning it explains full variability in the target variable.

OLS Regression Results

Dep. Variable:	price	**R-squared (uncentered):**	0.881
Model:	OLS	**Adj. R-squared (uncentered):**	0.881
Method:	Least Squares	**F-statistic:**	9177.
Date:	Wed, 11 Sep 2019	**Prob (F-statistic):**	0.00
Time:	23:49:57	**Log-Likelihood:**	-2.3746e+05
No. Observations:	17290	**AIC:**	4.750e+05
Df Residuals:	17276	**BIC:**	4.751e+05
Df Model:	14		
Covariance Type:	nonrobust		

RMSE: 212789.37425322627

Initially, I have applied the standard liner regression algorithm which is most widely used to understand the insights of the data. The following are the observations of the outcomes. The outcome of the predicted line has less accuracy, the algorithm basically helps me to observe the fashion of the data points, which are lying far from the actual predicted line. Further, I have found the RMS value to identify the error between the actual line and the predicted line, which are not significant, to overcome this I have picked a new algorithm (Random forest) to reduce the RMS value.

2. Random Forest

A random forest is a data construct applied to machine learning that develops large numbers of random decision trees analysing sets of variables. This type of algorithm helps to enhance the ways that analyse complex data. Random forest run on concept of bagging technic. With increase in computational power, I choose random forest algorithms which perform very intensive calculations. Leverage the strengths of different kind of models and Reduce overfitting.

```
print(models_cross)
```

```
           Model     Score
0  Random Forest  0.862353
```

```
np.sqrt(mean_squared_error(y_test, predictions))
```

```
144253.20061582504
```

Secondly, after observing the outcome of the liner regression, I have picked up random forest to get the better outcomes. The final come out come has a better RMs value, but which has not giving me the satisfactory results. Now coming to the outcome, the RMS value has reduced by half, which implies my direction of choosing the algorithm has betterment. My intention is not limited to reduce the RMS value, but to corelated the prices with the global standards which gives the meaningful insights. The main disadvantage in random forest algorithm is Speed and Explicability. So, I have chosen XGboost algorithm to improve the accuracy and speed.

3. XGBoost

The Xgboost also known as Extreme Gradient Boosting. It is an open source machine learning library with the so-called tree algorithm with gradient boosting. Using XGBoost, target variables can be more accurately by combining several simpler and weaker models and making estimates. The software addresses machine learning problems in the areas of regression, ranking. It works fast and delivers accurate output.

RMSE:

```
from sklearn.metrics import mean_squared_error, r2_score
np.sqrt(mean_squared_error(y_test, y_pred))
```

```
182348.6389199035
```

By now have gained better understanding of the data and the outcome which I am looking forward. XGboost has larger community support which help me to achive the outcome .

4. Catboost Algorithm:

Catboost is an open source machine learning algorithm. It can work with any data types to help solve a complex of problems that businesses face today. It provides best-in-class accuracy. Advantage of cataboost are it has high Performance, Handling Categorical features automatically, Robust, Easy-to-use. In catboost can handle categorical variables without performing a one-hot encoding. It clearly signifies that CatBoost mostly performs better for both tuned and default models.

```
print(model.get_best_score())
```

```
{'learn': {'RMSE': 142001.1841480544}, 'validation': {'RMSE': 149481.99244222036}}
```

Evaluvating the model:

Algorithm	RMSE
Linear Algorithm	212789.37425322627
Random Forest Algorithm	144253.20061582504
XGboost Algorithm	182348.638919035
Catboost	149481.99244222036

By seeing above table we can conclude that random forest has less RMSE value and accurate and catboost algorithm works fast and delivers accurate output.

Proposed Solution:

The actual observation from the data set which I found has major impact are Square foot , location and the year of construction respectively. However the overall accuracy which I have tried to improve is limited to the data which I have took to analyse.

My Idea of innovation is to predict the price automatically considering the factors influencing the price which are mentioned below. It has a broad scope and difficult to integrate the dimesions together as many factors cannot be fit under one platform. To bulid a prediction model we have to focus on integrating the social and economical factors .

Factors Influencing the outcome:

The scope of the project has sevaral dimessions to be consider inorder predict accuratly, Which includes priorty of the people which cannot be measured generically. The other major factors are the following.

- Depriceation Rate of the Building
- Expendicatre rate of the particular geografical location
- Increase in the population which creates the nessasity of accomodaction.
- Rasie/ Fall in the Realestate marks which has sevaral influnesing factors.
- Requried trusted soures to compare the ground truth.
- Slightly a marginal effect on the emotinal factors of the people.

To conclude, above are major factors to be considered to predict the accurate price of the building which are not mentioned in the data set.

Conclusion:

Therefore, to conclude it has been observed that the major factors that influenced on predicting the house prices were mainly the Square Feet of the property, the bedrooms and the bathrooms. By performing various algorithms for the analysis of house prices, it was observed that Random Forest algorithm produced results with better RMS value. The main drawback of using other algorithms was lack of sufficient data to train the model and produce desired results.