COVER PAGE

• The Project Title:

House Sales in King County, USA (Homes sold between may 2014 and may 2015)

• Author's Name:

Harlfoxem

• Course Project Advisor/Professor's Name:

Professor: Duoduo Liao, Ph.D.

GTA: Akanksh R. Telkala

• University Name, Course Number and Name:

George Mason University

AIT-580

Lakshmi Prasanna Gundabolu

• Date Of Submission:

12/17/2019

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

PROJECT TITLE

HOUSE SALES IN KING COUNTY, USA

(HOMES SOLD BETWEEN MAY 2014 AND MAY 2015)

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

ABSTRACT

Occupants in each network in the area are confronting an exceptional challenge in finding and keeping a home they can bear. A typical U.S. family devotes about a quarter of its annual income and half or more of its net worth to housing. Moderate lodging is a basic segment of our locale's foundation, and we should act together, over all degrees of government and all divisions, to address this emergency and guarantee the wellbeing and decency of our networks and the monetary imperativeness of our locale. Recent research has emphasized the importance of housing supply in determining house prices in different U.S. markets. This dataset contains house deal costs for King County, which incorporates Seattle. It incorporates homes sold between May 2014 and May 2015. In this project, I have Preprocessed the dataset(checking for missing values, duplicate values and deleting some attributes which are not useful), Performed some visualizations and Implemented Correlation Analysis to show which factors influencing the price of the house and what would be the important factors we need to consider before buying the house in that Area. Moreover, I also predicted the price of the houses using Random Forest Regression and Ordinary Least Square Regression techniques. Additionally, Conducted Hypothesis test for predicted value to determine the probability that a given condition is true and computing the p-value to know whether the null hypothesis is true or alternative is true.

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

INTRODUCTION

The project is about how were the sales of house prices in king county Area between the period May 2014 and May 2015. The dataset contains house sales prices for king county which includes Seattle. The data of these sales come from the official public records of the home sales in the king county area, Washington State. The dataset contains 21613 rows and 21 columns. Each represents a home sold from May 2014 through May 2015. The Main Aim for the project is predicting house prices based on some given attributes. The Motivation of this project is preprocessing the data, analyzing it, visualizing and performing some tests based on the predicted values. The purpose of report is about finding accuracies for different regression techniques and explaining the best regression techniques in such a way that how it works for predicting the price of houses.

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

OBJECTIVES

- Performing data preprocessing and cleaning to check for missing values and duplicate values in the dataset.
- Visualizing the data in such a way that everyone can understand how the price of the house will be determined.
- Using Correlation matrix to know the relationship between the given attributes.
- Plotting the graphs like Scatter plot and Box plot to eliminate the outliers for predicting the price of the houses.
- Performing the Hypothesis tests to check the probability of particular condition.
- Comparing accuracies and efficiencies for different Regression techniques and implementing the best Regression Model to predict the price of this dataset.
- Predict the house sales prices by using linear regression model techniques.

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

THE DATASET

• DATASET SELECTION

Hearing from low income families in all parts of the county who are struggling to find and keep a home they can afford today. The dataset which I have selected is reasonably large and complex dataset from Regional affordable housing task force data.

Data Source: Harlfoxem. "House Sales in King County, USA." *Kaggle*, 25 Aug. 2016, https://www.kaggle.com/harlfoxem/housesalesprediction

• DATASET DESCRIPTION

Describing the dataset:

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015.The dataset contains 21613 rows and 21 columns. Each represents a home sold from May 2014 through May 2015.

Who (company, agency, organization) collected the data?

Regional affordable housing task force data is an Organization which I have selected data from. The Regional Affordable Housing Task Force was created in 2017 to bring together representatives from King County, the City of Seattle and other cities with the goal of developing a regional plan to address the affordable housing crisis in King County.

Who they are and what do they do?

The members of the Regional Affordable Housing Task Force have immersed themselves in affordable housing data and policy to fully understand the economic drivers of the affordable housing crisis, how it is affecting individuals and families, and what solutions are available.

What is their role or purpose?

The main purpose of them is according to their estimates, they need 156,000 more affordable homes today and another 88,000 affordable homes by 2040 to ensure that no low-income or working households are cost burdened.

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

NEED

Why did they collect this data?

A coordinated, countywide effort to build affordable housing is not just about housing. It is also about building healthy and welcoming communities where all families and people, regardless of income, race, family size or need, are able to live near good schools, transit, jobs, and green spaces. King County is booming and finding ways to safely and affordably house our residents is a key component of ensuring prosperity continues and is shared into the future. So, they collected this data to tell people that these are the ranges of prices in that area which they can afford.

What potential questions and specific questions could be answered by studying this data?

- Smaller the house lesser the price how do we justify?
- In general, what is the most preferred type of house?
- Which is the costliest place to live in and why is it costly?
- If we are buying a house mostly which grade houses are available?
- Which Variables will have Strong Correlation with Attribute ‘Price’ in the dataset?
- How to know whether our Predicted Value is Correct or not?
- Finding the Accuracy of Different Regression Techniques?
- Identifying the best statistical model that fits dataset?

Is there any privacy, quality, ethical, or other issues with this dataset?

- There are no such issues in this dataset. The quality of dataset is very good and it does not have any missing values or duplicate data.

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

DATASET SCHEMA.

The description of proposed dataset is as follows:

id – Identification number of home sale

date- Date of the home sale

price - Price of each home sold

bathrooms - Number of bathrooms, where 0.5 accounts for a room with a toilet but no shower

floors - Number of floors
bedrooms - Number of bedrooms

sqft_living - Square footage of the apartment's interior living space

sqft_lot - Square footage of the land space

waterfront - A dummy variable for whether the apartment was overlooking the waterfront or not

view - An index from 0 to 4 of how good the view of the property was

condition - An index from 1 to 5 on the condition of the apartment,

BUILDING CONDITION

Relative to age and grade. Coded 1-5.

1 = Poor- Worn out. Repair and overhaul needed on painted surfaces, roofing, plumbing, heating and numerous functional inadequacies.

2 = Fair- Badly worn. Much repair needed. Many items need refinishing or overhauling.

3 = Average- Some evidence of deferred maintenance and normal obsolescence with age in that a few minor repairs are needed, along with some refinishing.

4 = Good- No obvious maintenance required but neither is everything new. Appearance and utility are above the standard.

5= Very Good- All items well maintained, many having been overhauled and repaired as they have shown signs of wear, increasing the life expectancy and lowering the effective age with little deterioration.

yr_built - The year the house was initially built.

yr_renovated - The year of the house's last renovation.

zipcode - What zipcode area the house is in.

grade - An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.

sqft_above - The square footage of the interior housing space that is above ground level.

sqft_basement - The square footage of the interior housing space that is below ground level.

lat - Latitude

long - Longitude

sqft_living15 - The square footage of interior housing living space for the nearest 15 neighbors.

sqft_lot15 - The square footage of the land lots of the nearest 15 neighbors.

Data Types of the Dataset

```
# printing datatypes
lp.dtypes
```

```
id                int64
date             object
price           float64
bedrooms          int64
bathrooms       float64
sqft_living       int64
sqft_lot          int64
floors          float64
waterfront        int64
view              int64
condition         int64
grade             int64
sqft_above        int64
sqft_basement     int64
yr_built          int64
yr_renovated      int64
zipcode           int64
lat             float64
long            float64
sqft_living15     int64
sqft_lot15        int64
dtype: object
```

Sample values of all the attributes in the dataset:

```
lp.head(3)
```

	id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view
0	7129300520	20141013T000000	221900.0	3	1.00	1180	5650	1.0	0	0
1	6414100192	20141209T000000	538000.0	3	2.25	2570	7242	2.0	0	0
2	5631500400	20150225T000000	180000.0	2	1.00	770	10000	1.0	0	0

Continue….

grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
7	1180	0	1955	0	98178	47.5112	-122.257	1340	5650
7	2170	400	1951	1991	98125	47.7210	-122.319	1690	7639
6	770	0	1933	0	98028	47.7379	-122.233	2720	8062

• DATASET PRE-PROCESSING

Data cleaning or data preparation is an essential part of statistical analysis. To ensure high quality data, we need to perform data preprocessing and data cleaning with the dataset.

Following are some of the steps:

- Removed the inconsistencies in the dataset. Checked for missing values, this dataset has no missing values.
- Later, identifying duplicate values and dropping them from the dataset.
- Checked for Null values in the dataset. As there are no null values in the data, directly performed feature engineering.
- By identifying some incompatible data types for analysis, Performed Feature engineering by transforming the data types.
- Creating a new attribute named house age from two existing variables in dataset by doing subtraction from one variable from another.
- Modulating the values of yr_renovated attribute to two values 0 and 1 and renaming it as renovated.
- Dropping some of the variables like date, year renovated and year built.
- Later on, while doing Regression for predicting the house prices some more attributes which are not necessary like zipcode, id, lat and long also will be dropped.
- Plotted correlation matrix (heat map) to know the relationship between all variables and also for knowing which attributes are affecting the price of the house.

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

THE SYSTEM

• System Architecture

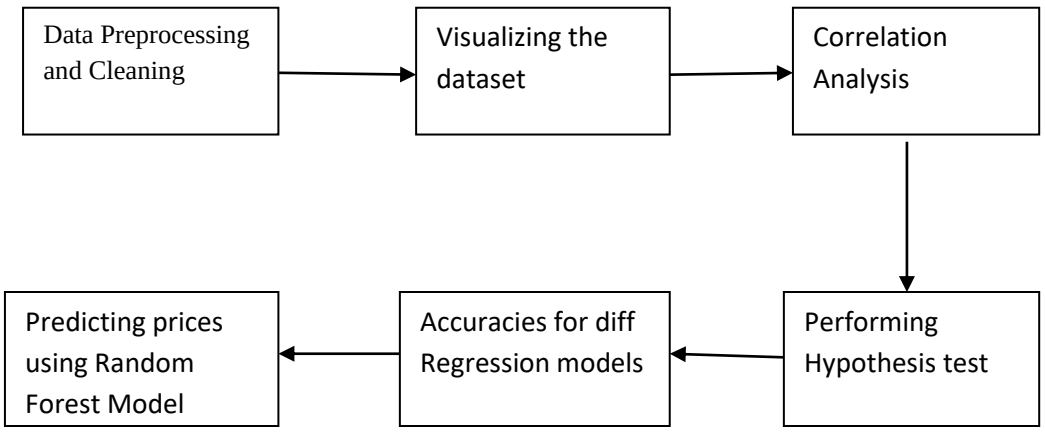


Fig: System Flow Architecture for predicting house sales prices

Data Processing:

The first step is to understand the dataset and analyzing it in such a way that we need to make dataset free from missing values, outliers, and duplicate values which is called descriptive analysis.

The second step is exploring the dataset by visualizing it with different plots and finding relationships between all the attributes which is called exploratory analysis.

Then Performing the Hypothesis test, which is called Inferential Analytics. After that finding accuracies for different regression techniques and finding the best fit model which fits our data.

Finally predicting the house prices by using Random forest regression model.

Descriptive Analytics: Descriptive Analytics takes raw data and parses that data to draw conclusions that are useful and understandable for people. It is very important as we can make informed strategic decisions based on historic data. From this, we can draw so many insights from dataset like which attributes are affecting the price of the house and what will be the important features in the dataset.

Diagnostic Analysis: It is type of data analytics in which data mining and correlation techniques are categorized.

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

Data Visualization: Here, some of the visualizations are performed on dataset to get clear meaning of what data is about. These visualization can interpret the best relationships between the data.

Linear regression and Radom Regression models: By using this, we can predict what will be the housing price in future.

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

Data Analytics Algorithms:

Here I have used two algorithms mainly to predict the house sales price:

Random Forest Regression Algorithm:

Random Forest is based on ensemble learning which is a type of supervised machine learning algorithm. Random Forest Algorithm combines multiple algorithm of same type which is multiple decision trees resulting a forest of trees. This algorithm can be used for regression as well as classification.

In this algorithm we have multiple trees and every tree is trained on a training set. Even if a data point is introduced it can't effect all trees as it would only impact one tree. So, it is considered as a stable algorithm. This can be used on both numeric and categorical features of datasets. It performs well even when there is missing data or when data is not scaled properly.

Building the model:

First, splitted the data into train and test set, so we can test MSE. I performed bagging by taking mtry=13(attribute split).Then call randomForest() function and then compute the test MSE for 500 bagged trees(default value). Then plotted a graph to observe the deviation between test values and the predicted values.

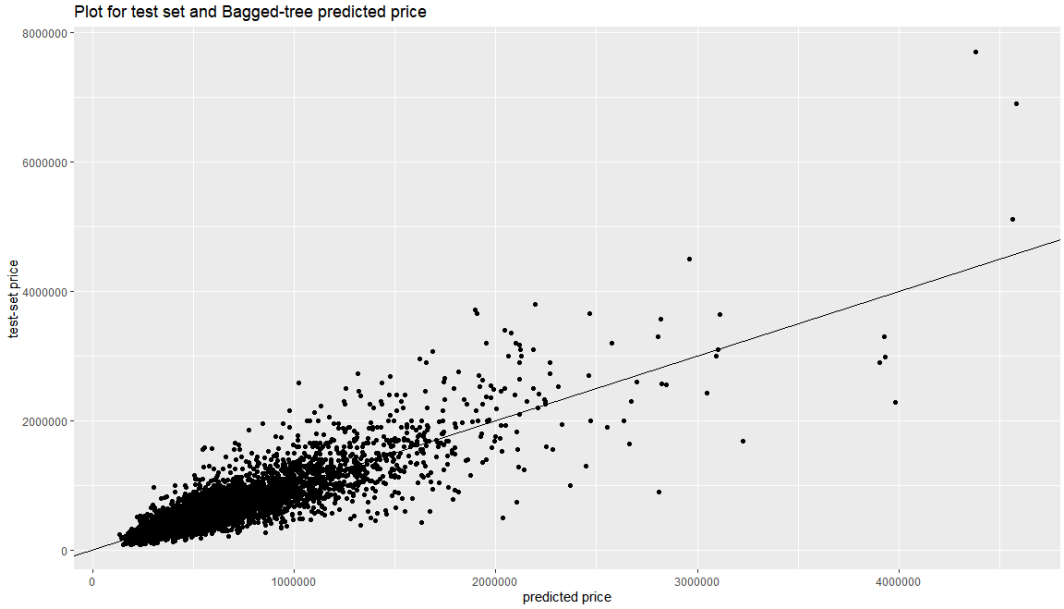


Fig1:Scatterplot for test set vs bagged tree predicted (price)

Initially there is not much deviation between the predicted and test set values in the scatterplot, but after a certain point there was a lot of deviation. This deviation cause due to some exceptional cases in dataset.

Variable Importance plot:

Plotted the Variable importance for the dataset and then did random forest by taking mtry=8 and then using four most important variables in the dataset.

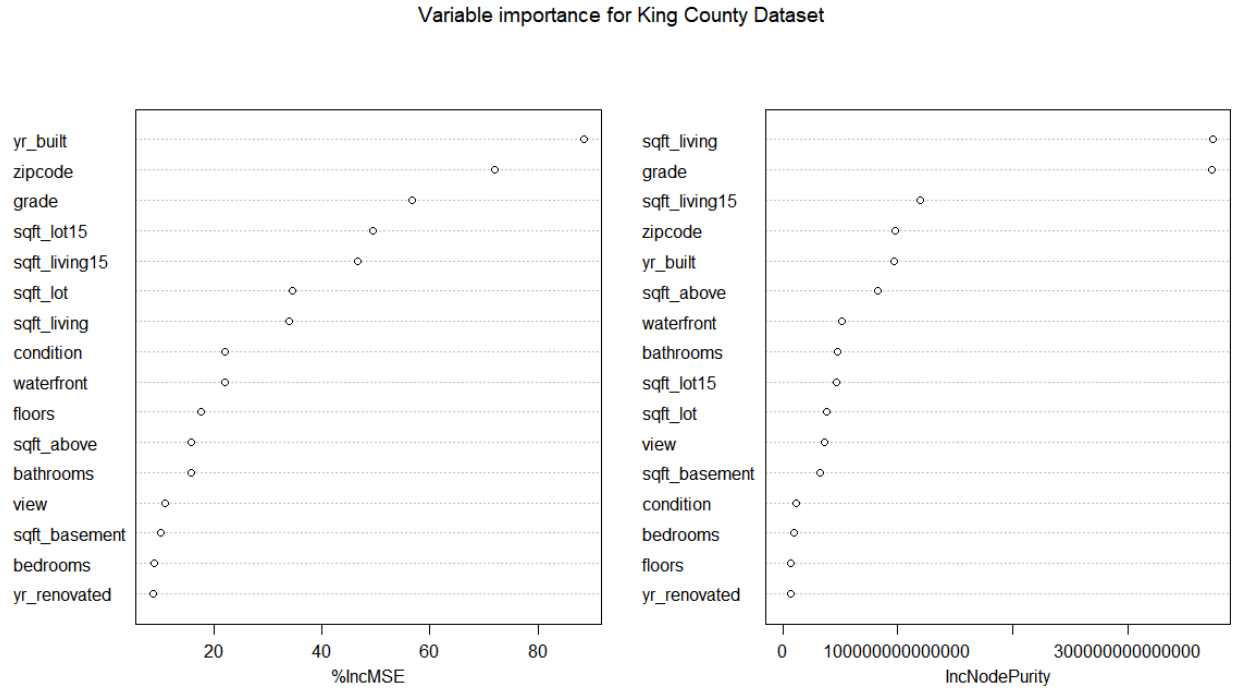


Fig2: Variable importance for King County Dataset

From the above graph we can say that sqft_living is the first important attribute in the dataset. Grade is also almost equally important as sqft_living. Yr_renovated is the least important variable used to predict the price in this dataset.

Yr_built and zipcode attributes have highest percentage increase in MSE.

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

Ordinary Least Square Regression Algorithm:

Ordinary Least Square Regression Algorithm is a statistical method of analysis that estimates the relationship between one or more independent variables and a dependent variable; the method estimates the relationship by minimizing the sum of the squares in the difference between the observed and predicted values of the dependent variable configured as a straight line.

In this project, I have taken dependent variable as price and all other strongly correlated variables as independent variables.

Here, firstly grouping all the attributes which have good correlation which is price into a single attribute. Creating a data frame and splitting attributes into two variables for prediction. Performing the algorithm code to predict the prices.

```
x = sm.add_constant(x) #adding a constant
```

```
C:\Users\GMU\Anaconda3\lib\site-packages\numpy\core\from
be removed in a future version. Use numpy.ptp instead.
  return ptp(axis=axis, out=out, **kwargs)
```

```
model = sm.OLS(y,x).fit()
predicted = model.predict(x)
model.summary()
```

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.662
Model:	OLS	**Adj. R-squared:**	0.662
Method:	Least Squares	**F-statistic:**	3845.
Date:	Tue, 17 Dec 2019	**Prob (F-statistic):**	0.00
Time:	14:15:42	**Log-Likelihood:**	-2.9590e+05
No. Observations:	21613	**AIC:**	5.918e+05
Df Residuals:	21601	**BIC:**	5.919e+05
Df Model:	11		

```
print(predicted)
```

```
0         282194.93
1         648230.86
2         310478.71
3         453567.72
4         563929.25
             ...
21608     464909.78
21609     518035.32
21610     242950.41
21611     429681.92
21612     242793.72
Length: 21613, dtype: float64
```

Fig 3: Performing Ordinary least square algorithm and predicted price

PROPOSED DEVELOPMENT SYSTEM:

Software Requirements:

Operating System – Windows/Mac OS (latest version)

The Software which I have used for this project are namely:

- Framework: Anaconda Jupyter IDE
- Python notebook in Jupyter lab
    1. For data preprocessing and performing correlation analysis between all the attributes.
    2. Comparing accuracies for various types of Regression models and finding which has the best accuracy for the dataset.
    3. Predicting house prices using ordinary least square regression technique.
- Used R Notebook in Jupyter lab for inferential analytics like hypothesis testing.
    1. Hypothesis test using predicted variable price to check whether it is correct or not.
    2. Hypothesis test to check whether the condition we have assumed is correct or not by comparing two grade conditions.(>7 & <7)
    3. Hypothesis test to check whether the condition we have given is correct or not by comparing condition variable conditions in the dataset.(>3 & <3)
- Used R studio for Inferential analysis of the dataset.
    1. Plotted some of the important plots like scatterplot, boxplot and histograms for visualization analysis of the dataset.
    2. Checking accuracy for linear regression technique.
    3. Predicting house sales prices using Random forest regression technique (ML) which is used for statistical analysis of data.

- Tableau For visualizing the graphs plots and charts to get better understanding of the dataset variables and their relationships.
    1. Plotted Geomap for better understanding at which places the average house sales prices are minimum and maximum.
    2. Plotted bar graphs for better understanding of relationship between two variables in the dataset.
    3. Comparing boxplots for identifying the outliers.
    4. By visualizing the data, it has been found that most preferred type of houses consist of which attributes mainly.

Hardware requirements

- Hardware - Intel Pentium (dual core) compatible processor or any multiprocessor based computer with a minimum 2Ghz or greater processor , a 64-bit system and network interface card.
- Disk Space – The space should be a minimum of 1GB.
- Memory – The recommended memory is 4GB or more.

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

Experimental Results and Analysis:

Research Question:

Smaller the house lesser the price how do we justify?

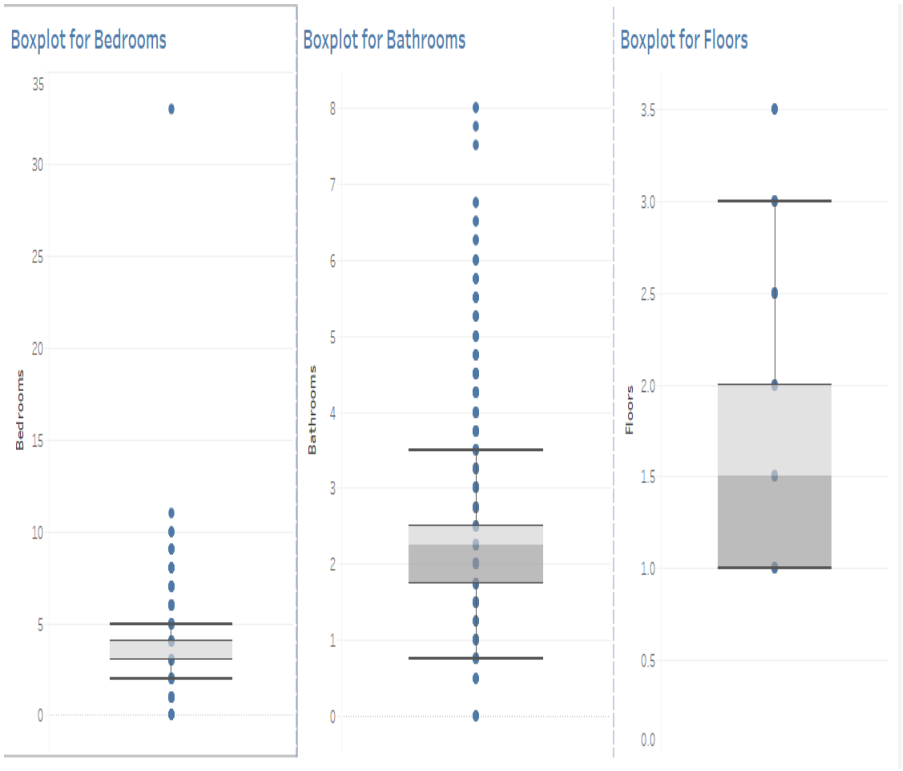
- From this scatter plot we can conclude that the house which has less sqft_living has lowest price compare to houses which has high sqft_living value.

- Predicted value compare to the test value has only 50% of accuracy as we got R squared value of 0.493

In general, what is the most preferred type of house?

From this Boxplots we have this information.

- Many houses have 3 bedrooms, 2.5 bathrooms and 1.5 floors which are available in reasonable prices.
- So many of people are interested in buying this type of houses.



AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

Which is the costliest place to live in and why is it costly?

From the below Geomap we identified that Medina is the costliest place in King County with an average price of $2,161,300. The cheapest place in King County is Auburn with an Average price of $234,284.

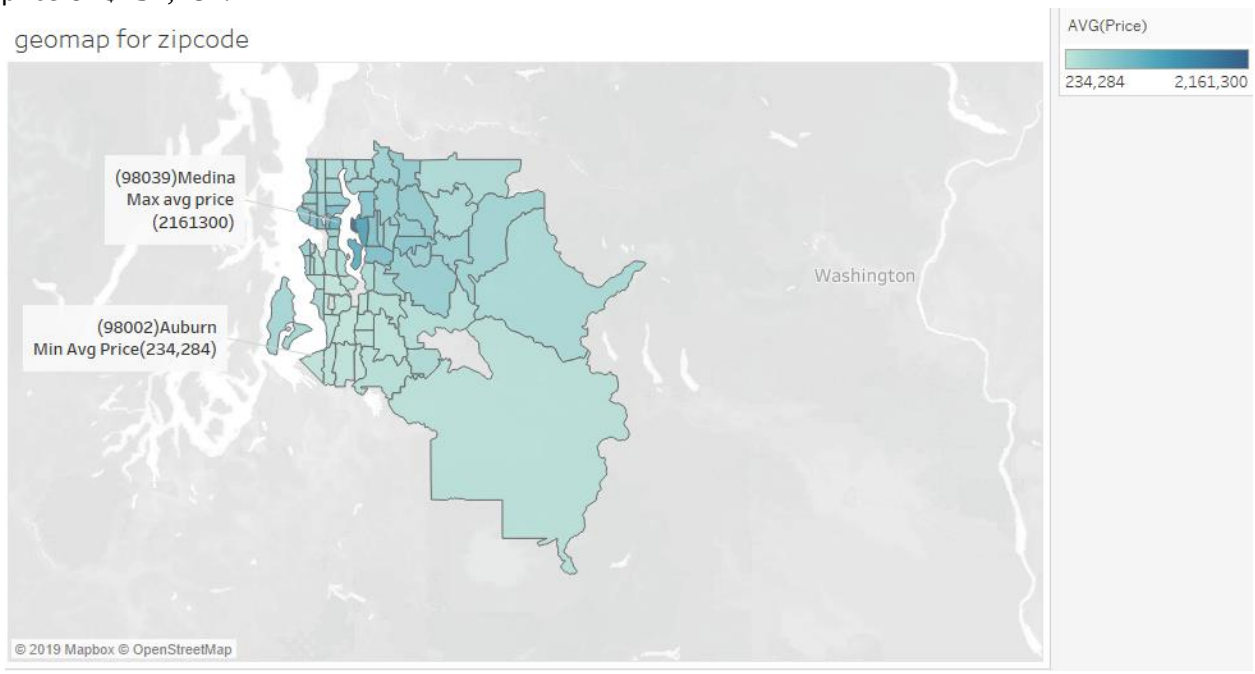
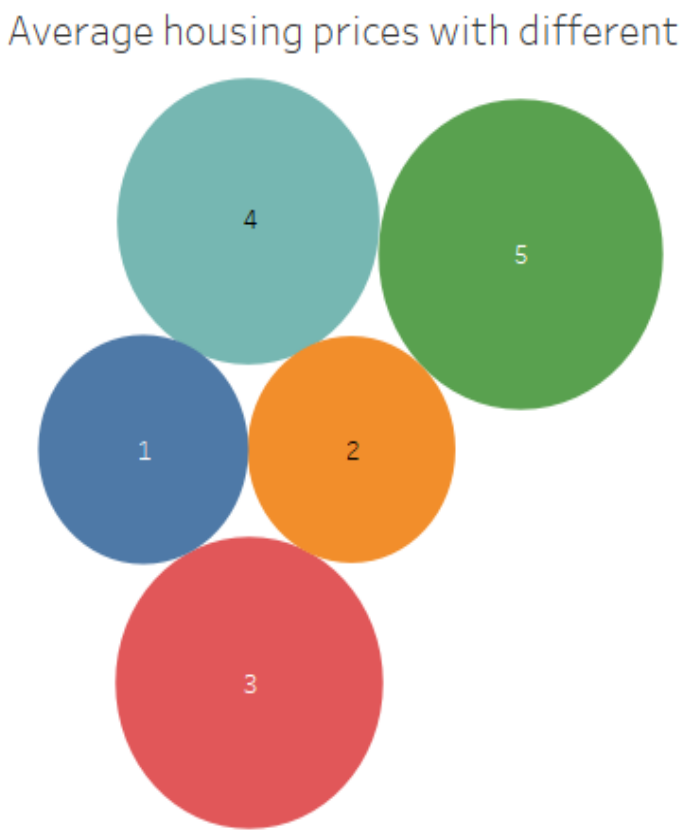


Fig1: Geomap representing the Average Prices for houses in King County

- The place Medina is a costliest place in King County because all the houses have a grade which is greater than or equal to 7 and it also has a waterfront which is shown in the below graph. The second costliest place which is Mercer island also has good grade and waterfront.

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

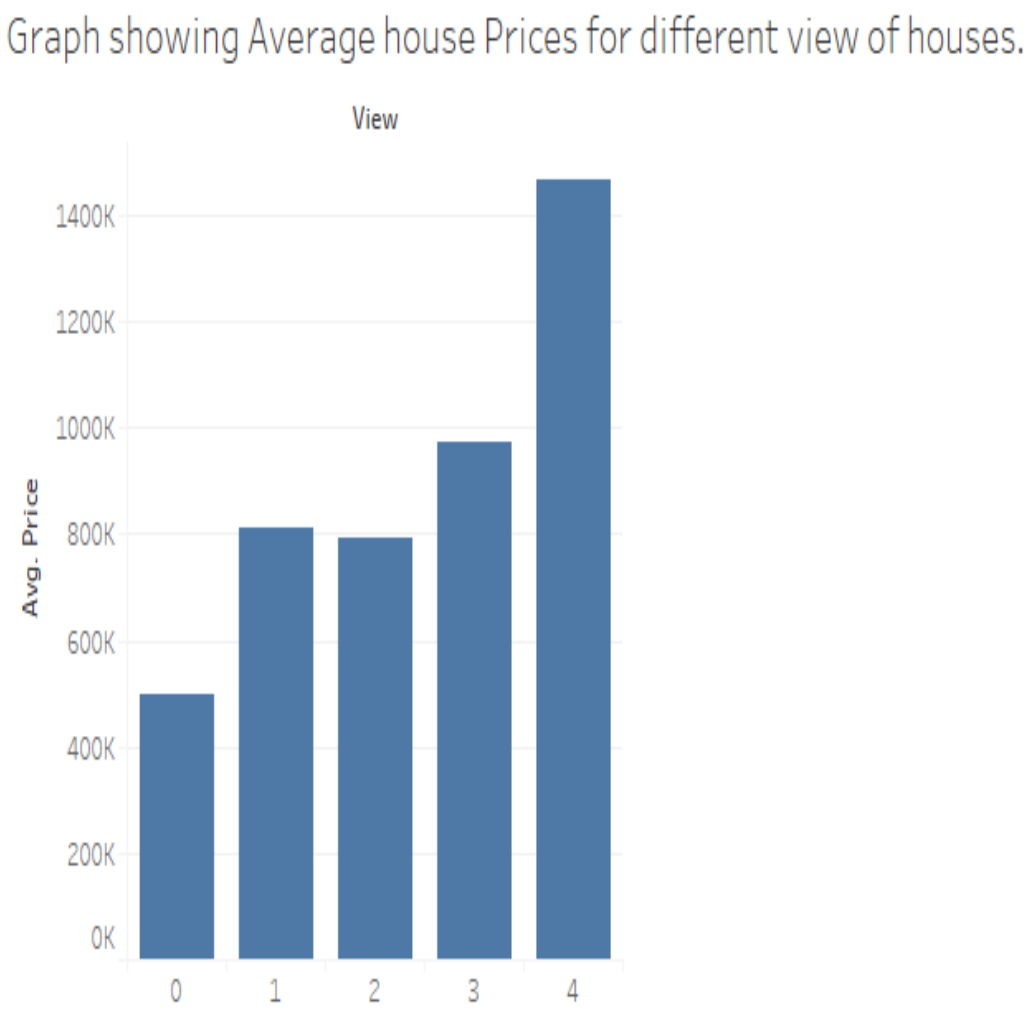
What will be the Average Price of the House with Different Conditions?



- Color shows details about Condition.
- Size shows average of Price.
- From this it is concluded that the houses with condition 1 has minimum average price and the houses with **condition 5 has maximum** average housing price.

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

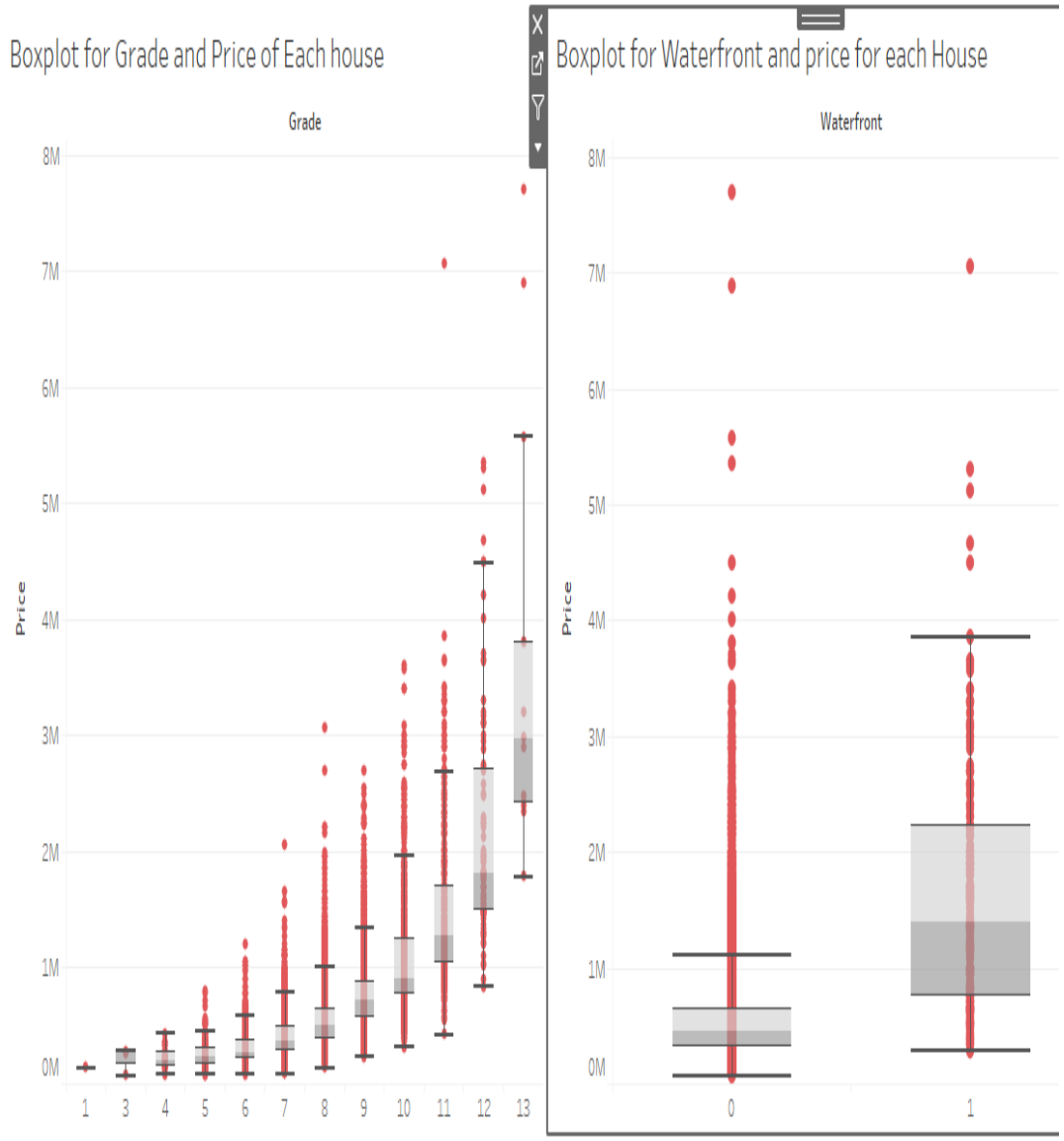
Which has Highest Average Price with Different Views of Houses?



- The house which has highest view has Maximum Average Housing Price of 1,464,363.
- The house with lowest view has Minimum Average Housing Price of 496,624.

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

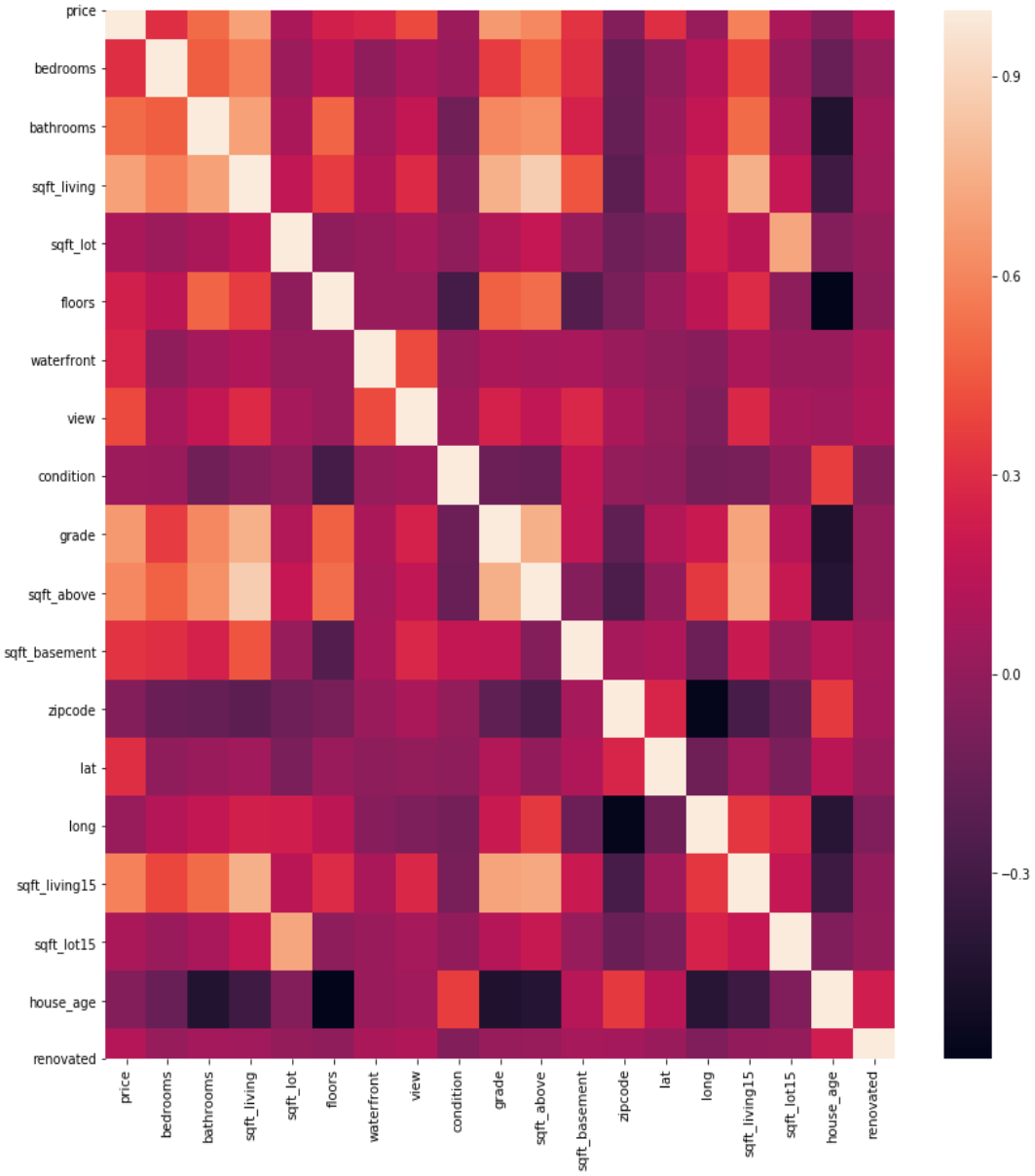
Comparing the Boxplots for Grade with Price and Waterfront with Price



- The Houses with Highest Grade has Maximum Average Prices and the houses with Lowest grades has Minimum Average Housing prices.
- The Houses with Waterfront Option have Maximum Average Housing Prices and Houses Without Waterfront Option has Minimum Average Housing Prices. Here '0' represents House without Waterfront And '1' represents Houses with Waterfront .

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

CORRELATION ANALYSIS



- From this we can say that the attributes which are effecting variable price are grade, sqft_living, sqft_living15, bedrooms, bathrooms and floors .
- These are variables which are strongly correlated to the variable price.

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

HYPOTHESIS TEST

One sample t-test for checking whether the predicted price is correct or not

```
t.test(lp$price, mu=540182.1587933121, alternative='greater')
```

```
        One Sample t-test

data:  lp$price
t = 2.6555e-12, df = 21612, p-value = 0.5
alternative hypothesis: true mean is greater than 540182.2
95 percent confidence interval:
 536071.8      Inf
sample estimates:
mean of x
 540182.2
```

Two sample t-test for comparing two values means and predicting whether the predicted value for both of them having same variable is correct or not .

```
t.test(lp$price[lp$grade>7],lp$price[lp$grade<7], alternative ='less')
```

```
        Welch Two Sample t-test

data:  lp$price[lp$grade > 7] and lp$price[lp$grade < 7]
t = 82.735, df = 12193, p-value = 1
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 428188.1
sample estimates:
mean of x mean of y
 714879.5  295038.8
```

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

REGRESSION ANALYSIS:

Identifying the best statistical model that fits dataset? (inferential analytics)

Accuracy for different statistical models:

```
Multiple Linear Regression Model Score is  69.0
Decision tree  Regression Model Score is  76.0
Random Forest Regression Model Score is  88.0
```

	Model	Score	Explained Variance Score
2	Random forest Regression	0.88	0.85
1	Decision Tree	0.76	0.73
0	Multiple Linear Regression	0.69	0.53

Here we can say that random forest regression has best accuracy compared to other two models**.**

For identifying the best statistical model:

- Firstly, we need to find correlation matrix for variables in the dataset.
- Implementing Hypothesis test to determine whether conditions which are drawn by using correlation are satisfying the data or not.
- Finding the accuracy of the best model.
- Plotting the variable importance in the dataset.
- Performing best regression model that suits for the dataset.

LINEAR REGRESSION MODEL:

By this we can tell that residual standard error for linear regression model is 65% which means we only have 65% chance of predicting our data would be correct.

```
Residual standard error: 216900 on 21601 degrees of freedom
Multiple R-squared:  0.6515,    Adjusted R-squared:  0.6514
F-statistic:  3672 on 11 and 21601 DF,  p-value: < 0.00000000000000022
```

Fig: Scatterplot for Linear regression model for sqft_living and price

Interpreting the results:

- I have predicted the price of the house which is 88% accurate and its mean is about 540182.1587934085
- I also calculated accuracies for various regression techniques, for random forest I got 88% , Decision tree is 76% and 69% for linear regression model.
- So I have chosen Random forest regression model for predicting the house prices.

AIT-580 Final Project Report
Lakshmi Prasanna Gundabolu
Professor: Duoduo Liao, Ph.D.

CONCLUSION:

Insights drawn from visualization:

1. I found that places which are closer to the coast are costlier than non-coastal areas.
2. I found that grade and price are directly proportional.
3. I found that as a square feet of living is increasing price is also increasing; price has highest correlation with sqft_living.
4. I also found that average house prices of built houses are cheaper than renovated houses.
5. The most preferred type of house consists of 3 bedrooms, 2.5 bathrooms and 1.5 floors.

Insights drawn from statistical models:

1. I found that multiple linear regression is not best fit model for our data and we don't have any outliers.
2. I calculated accuracies for different models and we got random forest is best model as it has accuracy of 88.03%
3. I plotted predicted vs test value and we found that the model fits our data.

REFERENCES:

[1] Harlfoxem. “House Sales in King County, USA.” *Kaggle*, 25 Aug. 2016, https://www.kaggle.com/harlfoxem/housesalesprediction.

[2] Girgin, Samet. “Random Forest Regression in 5 Steps with Python.” *Medium*, Medium, 7 Apr. 2019, https://medium.com/@sametgirgin/random-forest-regression-in-5-steps-with-python-ee4259eca0de.

[3] “Linear Regression R.” *DataCamp Community*, https://www.datacamp.com/community/tutorials/linear-regression-R.