**KING COUNTY HOUSING**

**URVI NILESHKUMAR PATEL (0770850)**

DAB 402, M018, 2020-2021

A capstone project submitted

in partial fulfilment of the requirements for the degree of

**DATA ANALYTICS FOR BUSINESS**

in

**Data Analysis**

SAINT CLAIR COLLEGE FOR APPLIED ARTS AND TECHNOLOGY

Mississauga, ONTARIO, CANADA

SUBMITTED TO

**Professor Savita Sherawat**

**Abstract**

To foresee King County's home costs, I picked the lodging cost dataset that was sourced from Kaggle. This dataset contains house deal costs for King County, which incorporates Seattle. It incorporates homes sold between May 2014 and May 2015

**Introduction:**

In one's life, the most costly and biggest buy that the individual makes are generally a home. People should know the sensible worth of their resources and housing prices. The problem of house pricing is typically explained on a basis of hedonic demand theory. It states that each characteristic of a house (such as size or location) has some contribution in its price. Since our data consists of different features of houses, it fits the hedonic theory. Over the years, the problem of property evaluation was solved in many ways. Statistical tools used by analysts to explain house prices range from simple linear regression to more complex techniques such as artificial neural networks. In the literature we can distinguish two trends, these are publications describing linear models compared to advanced machine learning algorithms (Din, Hoesli, and Bender [2001](https://pbiecek.github.io/xai_stories/story-house-sale-prices.html#ref-house_prices_1)) and (Selim [2009](https://pbiecek.github.io/xai_stories/story-house-sale-prices.html#ref-house_prices)). In the (Selim [2009](https://pbiecek.github.io/xai_stories/story-house-sale-prices.html#ref-house_prices)) research, can see artificial neural network prediction errors as compared to linear regression. We can conclude that we reduce the interpretability to an increase in the quality of model fitting. The second trend in articles is the contribution of machine learning to price prediction (Conway [2018](https://pbiecek.github.io/xai_stories/story-house-sale-prices.html#ref-house_prices_3)) and (Park and Bae [2015](https://pbiecek.github.io/xai_stories/story-house-sale-prices.html#ref-house_prices_2)). The authors often consider variables that describe the location of a property as an important element of pricing modeling. These include relationships such as postal codes, distances from public transport (bus stops, subways), parks, or cultural sites (Law [2017](https://pbiecek.github.io/xai_stories/story-house-sale-prices.html#ref-house_prices_4)) and (Heyman and Sommervoll [2019](https://pbiecek.github.io/xai_stories/story-house-sale-prices.html#ref-house_prices_5)).

Here, we will apply a few relapses and prescient techniques to concentrate on house deal cost in King County, Washington, USA and investigate the best model for expectation.

For instance, highlight positioning with Random Forest, RFE, and direct models was examined, and straight models were assessed in certain works. Numerous relapses, rope relapse and k-Nearest Neighbors Regression were additionally examined.

**Software:** RStudio, Jupyter(python) Excel, Word, PowerPoint.

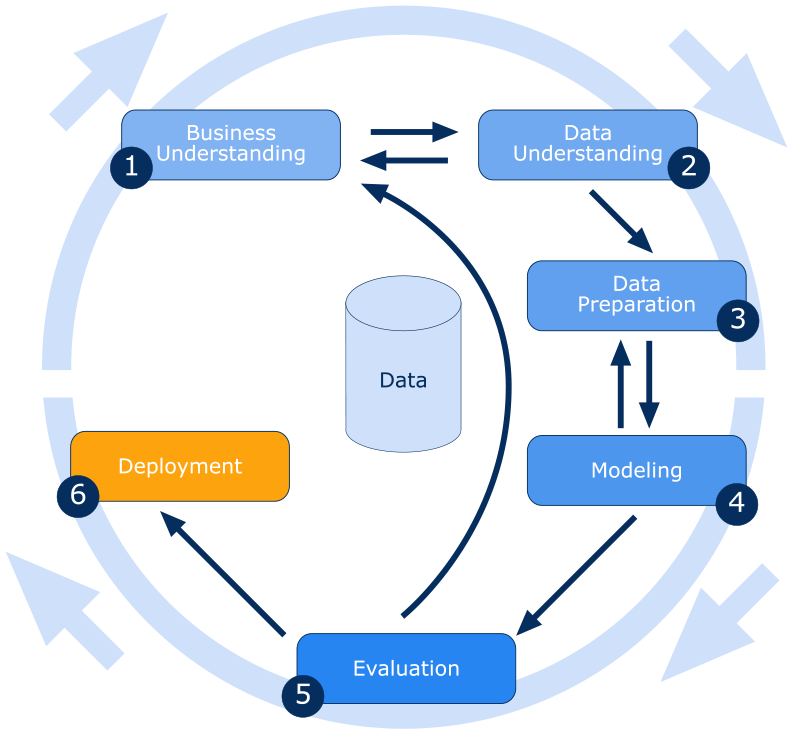
**Depiction**

In this dataset the business cost of houses in King County, Seattle is available. It incorporates homes sold between May 2014 and May 2015. Prior to doing anything we should initially think about the dataset that it contains, what are its provisions and what is the design of information.

By noticing the information, we can realize that the cost is reliant upon different elements like bedrooms (which is most reliant element), washrooms, sqft\_living (second most significant component), sqft\_lot, floors and so forth the cost is likewise reliant upon the area of the house where it is available. Different provisions like waterfront, see are less reliant upon the cost.

**Overall Methodology**

Our overall methodology contains a total of 6 steps for data mining.



**Define the Goal**

Understanding the company or activity in which your data project is embedded is critical to its success and the first step in any successful data analytics project. It's not enough to just download a cool open dataset. You must determine a clear objective of what you want to do with data to be motivated, direction, and purpose: a specific question to answer, a product to produce, and so on. The purpose of an analytics project is not to show off how well you know a new tool, but to find relevant trends in the data you've provided. As a result, it would be more productive to focus on asking questions of the data than worrying about which tool to employ. So, our goal is to predict the price of the house. To figure out the best model, we used a supervised learning approach on dataset variables.

The general thought of relapse is to look at three things:

a. Which locations within the King County area have the highest average house

prices?

b. Which house attributes increase sale price?

c. Does time of the year have an impact on house sales?

**Get the data**

Once we have gotten our goal figured out, we started looking for our data, the second phase of a data analytics project. Mixing and merging data from as many data sources as possible is what makes a data project great, so we looked as far as possible. We looked for various open-source datasets on the internet. There is a plethora of datasets available on the Internet to supplement what you already have. For example, census data may be used to calculate the average revenue in the district where your user lives, and OpenStreetMap can show you how many coffee shops are on a certain street. There are open data platforms in several nations (like data.gov in the U.S.). After checking various datasets, we choose the King County Housing dataset from kaggle.

Dataset link: [House Sales in King County, USA | Kaggle](https://www.kaggle.com/harlfoxem/housesalesprediction)

**Data Preparation**

After we obtained our data, we began working on it in the third data analytics project phase. We began investigating to discover what we had and how we might connect it all to meet our initial goal. This dataset comprises house selling prices for King County, which includes Seattle, which we discovered. Homes sold between May 2014 and May 2015 are included. It has 21597 rows and 21 columns (features).

In data exploration we tried to find out these answers first:

* Which features are categorical?

These values classify the samples into sets of similar samples. Within categorical features are the values nominal, ordinal, ratio, or interval based. Among other things this helps us select the appropriate plots for visualization.

Categorical: id, waterfront, zip code.

* Which features are numerical?

These values change from sample to sample. Within numerical features are the values discrete, continuous, or time series based? Among other things this helps us select the appropriate plots for visualization.

* + Continuous: price, bathrooms, floors, lat, long.
  + Discrete: date, bedrooms, sqft\_living, sqft\_lot, view, condition, grade, sqft\_above, sqft\_basement, yr\_built, yr\_renovated, sqft\_living15, sqft\_lot15.
* Which features contain blank, null, or empty values?

We can check for missing values with pandas isnull() and is.na in R. This indicates whether values are missing or not. Then we can sum all the values to check every column.

**Detailed data dictionary and structure of the data**

| **Attribute** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| id | 21613 | 4580301 | 28765 | 1000 | 212304 | 3904 | 7308 | 990 |
| price | 21613 | 540182.1 | 367362 | 75000 | 321950 | 450000 | 645000 | 7700000 |
| bedrooms | 21613 | 3.37 | 0.93006 | 0 | 3 | 3 | 4 | 33 |
| bathrooms | 21613 | 2.114 | 0.7701 | 0 | 1.75 | 2.25 | 2.5 | 8 |
| sqft\_living | 21613 | 2079.8 | 918.44 | 290 | 1427 | 1910 | 2550 | 13540 |
| sqft\_lot | 21613 | 15106.9 | 41420.5 | 520 | 5040 | 7618 | 10688 | 1651359 |
| floors | 21613 | 1.494 | 0.53998 | 1 | 1 | 1.5 | 2 | 3.5 |
| waterfront | 21613 | 0.0075 | 0.08651 | 0 | 0 | 0 | 0 | 1 |
| view | 21613 | 0.23 | 0.76 | 0 | 0 | 0 | 0 | 4 |
| condition | 21613 | 3.4 | 0.65 | 1 | 3 | 3 | 4 | 5 |
| grade | 21594 | 7.65 | 1.17 | 1 | 7 | 7 | 8 | 13 |
| sqft\_above | 21613 | 1788.39 | 828.09 | 290 | 1190 | 1560 | 2210 | 9410 |
| sqft\_basement | 21613 | 291.5 | 442.57 | 0 | 0 | 0 | 560 | 4820 |
| yr\_built | 21613 | 1971 | 29.37 | 1900 | 1951 | 1975 | 1997 | 2015 |
| Attribute | count | mean | std | min | 25% | 50% | 75% | max |
| yr\_renovated | 21613 | 84.4 | 401.67 | 0 | 0 | 0 | 0 | 2015 |
| zipcode | 21613 | 98077.93 | 53.5 | 98001 | 98033 | 98065 | 98118 | 98199 |
| lat | 21613 | 47.56 | 0.13 | 47.15 | 47.471 | 47.5718 | 47.678 | 47.7776 |
| Long | 21613 | -122.21 | 0.14 | -122.5 | -122.328 | -122.23 | -122.125 | -121.315 |
| sqft\_living15 | 21613 | 1986.55 | 685.3 | 399 | 1490 | 1840 | 2360 | 6210 |
| sqft\_lot15 | 21613 | 12768.45 | 27304.1 | 651 | 5100 | 7620 | 10083 | 871200 |

**Data Cleaning:**

Cleaning your data is the next (and, by far, the most hated) stage. You've most likely observed that, despite having a country feature, you have varied spellings or even missing data. It's time to examine each of your columns to ensure that your data is uniform and free of errors.

* **Changing the datatype of the date column:** Date attribute data type is Character so, I just change it in Date Data type
* **Missing values:**

|  |  |
| --- | --- |
| Attribute | Missing Value |
| Grade | 19 |

**Grade:** we find out the mean of the grade attribute and replace the missing values with

the mean value of the grade.

* **Removing unwanted columns:** We removed unwanted columns which do not have any relation with price. For Example: ID
* **Converting the price:**  from Dollar to units of 1000 Dollar to improve readability.
* **Here, the house with 33 bedrooms is interesting. Let's check it.**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **id** | **date** | **price** | **bedrooms** | **bathrooms** | **sqft\_living** | **sqft\_lot** | **floors** | **waterfront** | **view** | **condition** | **grade** |
| 7.13E+09 | 2014-10-13 | 221900 | 33 | 1 | 1180 | 5650 | 1 | 0 | 0 | 3 | 7 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **sqft\_above** | **sqft\_basement** | **yr\_built** | **yr\_renovated** | **zipcode** | **lat** | **long** | **sqft\_living15** | **sqft\_lot15** |
| 1180 | 0 | 1955 | 0 | 98178 | 47.5112 | -122.257 | 1340 | 5650 |

**With 1 bathroom, a sale price of $221900 it is likely that this house has 3 bedrooms and the 33 was a data entry error. That's why we replaced 33 with 3.**

* **After that I created 2 new column using date, year built, and year renovated.**

For that I just split the date column in to 3-part Date, Month and Year

**Age:** Converting the year-built column into an age column as it made the model more interpretable. Since the latest year in the dataset was 2015, I created this column by subtracting year built from 2015.

**Renovated:** I changed the year renovated column into a binary column. 1 for homes renovated in the past 10 years or built within the past 5 (most likely not needing renovations), and 0 for homes that were not renovated within 10 years.

* **Drop the unusable column.**

After Creating the above 2 column my dataset does not need column id, Date, and year renovated so I just drop this columns for cleaning purposes.

* **Remove Outliers.**

Outlier: An outlier is a data point in a data set that is distant from all other observations. A data point that lies outside the overall distribution of the dataset.

First, I check the boxplot to see whether houses with waterfront or without waterfront have outliers or not.

Chart, box and whisker chart

Description automatically generated

Above images depict my dataset have some Outliers. So, I need to remove this outlier.

After that I removed outliers using 3 sigma

Formula: (*μ + mean)/ σ < = 3*

After that, a total 2826 outliers removed.

Now new size of data is: (18787, 21)

Then I store the new data set in new data and give the name KC\_clean1.

After removing outlier, the boxplot of waterfront looks like this

Graphical user interface

Description automatically generated

So, from above boxplot I can say all houses with waterfront have outliers so after removing outliers all houses with waterfront removed so, I can say that dataset do not need the attribute waterfront for predicting price. So, I Drop the waterfront column for the accurate result.

And then I plot the boxplot for **grade**, **bathrooms**, and **bedrooms** with outliers or without outliers to understand the other attribute respectively.

Chart, box and whisker chart

Description automatically generated

Boxplot for grade with or without outliers.

**Chart, scatter chart

Description automatically generated**

Here, the graph depicts the boxplot of bathrooms with outliers and without outliers

Chart, timeline, box and whisker chart

Description automatically generated

Boxplot for bedrooms with or without outliers

After removing Outliers, the distplot is plotted for sqft living to see if the data is skewed or not.

**A density plot** is a representation of the distribution of a numeric variable. It uses a kernel density estimate to show the probability density function of the variable

A picture containing icon

Description automatically generated

Distplot of sqft\_living with outliers it depicts the data is left skewed.

Chart, histogram

Description automatically generated

But After removing outliers it is center skewed so, I can use this attribute for modeling

Chart

Description automatically generated

Distplot of sqft\_living with outliers it depicts the data is left skewed.

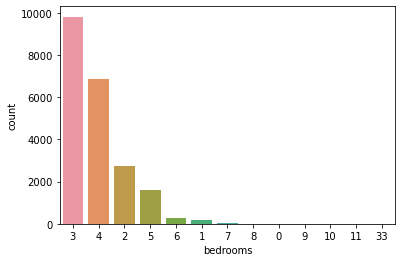
Chart, histogram

Description automatically generated

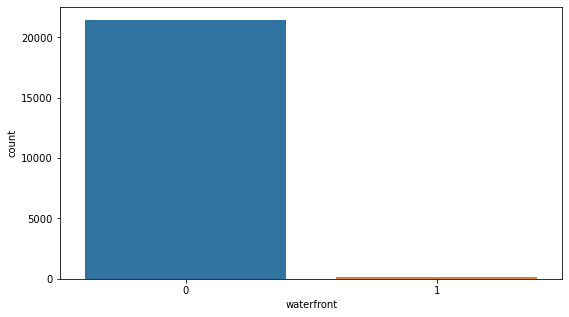
But after removing outliers it is center skewed so, I can use this attribute for modeling

Now that we have clean data, it's time to alter it to maximize its worth. To start the data enrichment phase of the project, you should combine all your diverse sources and group logs to limit down your data to the most important attributes. Here, we added one more column to our data that is the age of the house. We calculated it by subtracting the year built from the year sold. After that we divided the dataset into a training and testing set by randomly selecting 80% of the data.

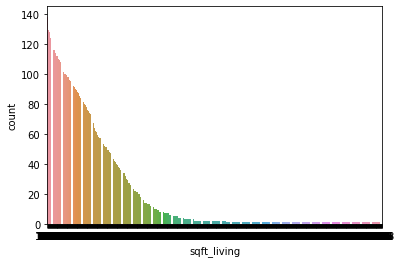
1. **Data Visualization:** Data visualization gives us a clear idea of what the information means by giving it visual context through maps or graphs. This makes the data more natural for the human mind to comprehend and therefore makes it easier to identify trends, patterns, and outliers within large data sets.

****

Number of houses per bedrooms

****

Number of houses with waterfront

****

Number of houses per sqft\_living

****

Latitude and Longitude with Price

**let's see how the price changed along the years**

**Chart, scatter chart

Description automatically generated**

**The price changed along the grade**

**Shape

Description automatically generated**

For the two categorical variables(view and grade) we draw boxplots to understand the relationship.

Chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

**linear relationship between features and price**

To get an idea of the linear relationship between features and price, we will use a regplot for each feature.

Chart, scatter chart

Description automatically generated**Chart, icon

Description automatically generated**

linear relationship between bedrooms and price, bathrooms and price

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

linear relationship between sqft\_living and price, sqft\_lot and price

Chart

Description automatically generated with medium confidenceChart

Description automatically generated

linear relationship between floors and price, view and price

Icon

Description automatically generatedChart, icon

Description automatically generated

linear relationship between grade and price, condition and price

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

linear relationship between sqft\_above and price, sqft\_basement and price

Chart, scatter chart

Description automatically generatedShape, rectangle

Description automatically generated

linear relationship between renovated and price, age, and price

There are several features where there is a clear linear relationship with price, but there are some features where the relationship is not so clear. We can further look at the correlation between all of our features and price to get a better idea of which features have a linear relationship with the dependent variable. Prior to doing so, let's look at correlation.

**Correlation**: Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate). It's a common tool for describing simple relationships without making a statement about cause and effect.

**Correlation between price and other attribute:**

| **Name: Price** | **Dtype: float64** |
| --- | --- |
| Price | 1.000000 |
| sqft\_living | 0.702044 |
| grade | 0.656089 |
| sqft\_above | 0.605566 |
| sqft\_living | 0.585374 |
| bathrooms | 0.525134 |
| view | 0.397346 |
| sqft\_basement | 0.323837 |
| bedrooms | 0.315427 |
| lat | 0.306919 |
| waterfront | 0.266331 |
| floors | 0.256786 |
| yr\_renovated | 0.126442 |
| sqft\_lot | 0.089655 |
| sqft\_lot15 | 0.082456 |
| yr\_built | 0.053982 |
| condition | 0.036392 |
| long | 0.021571 |
| zipcode | -0.053168 |

After checking **correlation**, I create plotting in **heatmaps.**

**Heatmaps** visualize the data in 2-D colored maps making use of color variations like hue,

saturation, or luminance. Heatmaps describe relationships between variables in the form of colors instead of numbers. These variables are plotted on both axes. The color changes describe the relationship between two values according to the intensity of the color in a particular block.

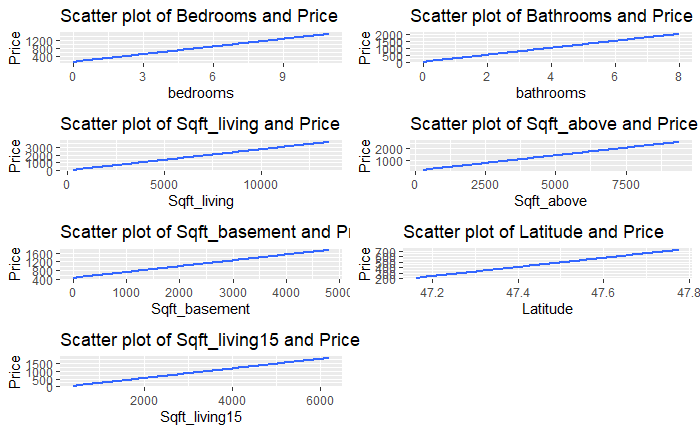
**Treemap chart

Description automatically generated with medium confidence**

According to our corrplot price is positively correlated with bedroom, bathroom, Sqft\_living, view, grade, sqft\_above, sqft\_basement, lat, sqft\_living 15.

Next, we will draw some scatter plots to determine the relationship between these variables.

**Association of different variables with respect to Price.**



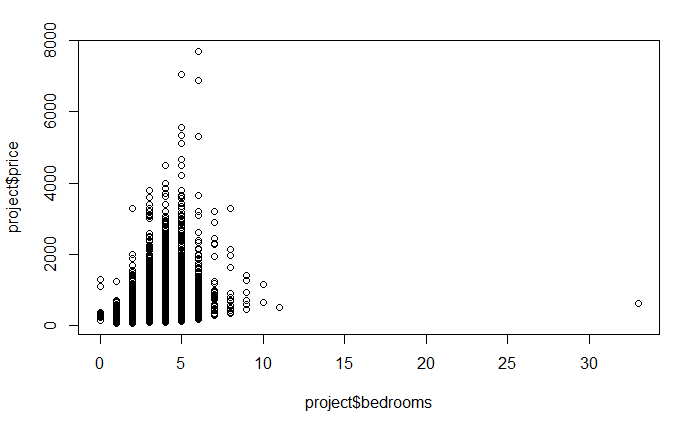
From these scatter plots, we conclude that the relationship between price and bedroom, bathroom, Sqft\_living,sqft\_above, sqft\_basement, lat, sqft\_living15 is linear.

**Data Modeling**

Now that we have a decent dataset (or more), it's time to start studying it by creating graphs. Visualization is the finest approach to examine and convey your discoveries when working with massive amounts of data, and it is the next phase of any data analytics project.

**Multiple linear Regression:** Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. Multiple regression is an extension of linear (OLS) regression that uses just one explanatory variable.

**Plot of Multiple Regression**

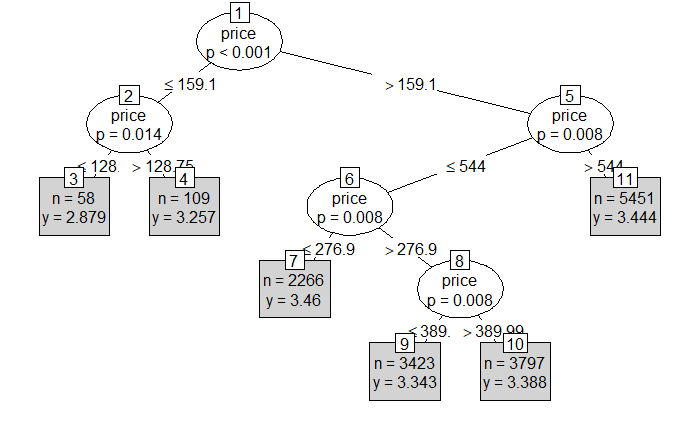
****

**Checking unique values of Condition attribute for Decision Tree**

[1] 3 5 4 1 2

**Decision Tree:** A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

**Decision tree for condition:**

****

**Table for Decision tree:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| project\_ctree\_prediction1 | 1 | 2 | 3 | 4 | 5 |
| 2.879310345 | 1 | 2 | 23 | 7 | 0 |
| 3.256880734 | 1 | 4 | 32 | 15 | 3 |
| 3.343266141 | 2 | 6 | 927 | 345 | 78 |
| 3.388201211 | 2 | 6 | 1139 | 428 | 98 |
| 3.443771785 | 1 | 6 | 1521 | 618 | 272 |
| 3.45984113 | 1 | 25 | 549 | 326 | 71 |

**Evaluation Matrix Decision Tree:**

MAE: 78461.55996097215

MSE:14001782968.965452

RMSE: 118329.12984115725

R2: 0.7276427439009999

**Applying KNN (KN Table):** The K-nearest neighbors (KNN) algorithm is a type of supervised machine learning algorithm. [KNN](https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm) is extremely easy to implement in its most basic form, and yet performs quite complex classification tasks. It is a lazy learning algorithm since it doesn't have a specialized training phase. Rather, it uses all the data for training while classifying a new data point or instance. KNN is a non-parametric learning algorithm, which means that it doesn't assume anything about the underlying data. This is an extremely useful feature since most of the real-world data doesn't really follow any theoretical assumption e.g. linear-separability, uniform distribution, etc.

K Nearest Neighbor is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. It is mostly used to classifies a data point based on how its neighbors are classified

**Evaluation Matrix K Nearest Neighbors:**

MAE: 129808.49547631718

MSE:30672429334.523396

RMSE: 175135.4599574952

R2: 0.38806313605199916

**Clustering:** Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.

|  |  |  |  |
| --- | --- | --- | --- |
| K-means clustering with 3 clusters of sizes | 20172 | 1153 | 288 |

**Cluster means:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | price | bedrooms | bathrooms | sqft\_living | sqft\_lot | floors | waterfront | view |
| 1 | 531.1024 | 3.362731 | 2.092145 | 2029.652 | 8347.167 | 1.490631 | 0.006742019 | 0.227444 |
| 2 | 665.629 | 3.509107 | 2.41327 | 2762.224 | 64905.751 | 1.535993 | 0.02254987 | 0.2766696 |
| 3 | 673.9173 | 3.385417 | 2.503472 | 2867.646 | 289206.587 | 1.585069 | 0.003472222 | 0.5451389 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | condition | grade | sqft\_above | sqft\_basement | yr\_built | yr\_renovated | zipcode | lat |
| 1 | 3.411858 | 7.607401 | 1738.533 | 291.1195 | 1970.324 | 84.01517 | 98080.03 | 47.56214 |
| 2 | 3.402428 | 8.37294 | 2460.735 | 301.4892 | 1979.716 | 88.03816 | 98050.27 | 47.54267 |
| 3 | 3.267361 | 8.274306 | 2588.809 | 278.8368 | 1983.865 | 96.95833 | 98042.43 | 47.48375 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | long | sqft\_living15 | sqft\_lot15 |
| 1 | -122.2248 | 1946.82 | 7931.056 |
| 2 | -122.072 | 2573.16 | 55542.745 |
| 3 | -122.0202 | 2420.993 | 180342.316 |

# **Random forests:**

# Random forests are a supervised learning algorithm. Random forests create decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting.

**Evaluation Matrix Random Forest:**

MAE: 53488.89229312283

MSE:6516063286.617214

RMSE: 80722.13628625803

R2: 0.8699998852581527

# **Conclusion:**

The variables that will heavily affect the price change of each house in King County heavily are the bedrooms, bathrooms, sqft\_living, floors, waterfront, condition, grade, sqft\_above, sqft\_living15, sqft\_lot15, yr\_gap, renov variables.

Accuracy Table.

|  | **Linear regression** | **KNN** | **Decision Tree** | **Random\_Forest** |
| --- | --- | --- | --- | --- |
| **R2** | 6.759024e-01 | 3.880631e-01 | 7.276427e-01 | 8.699999e-01 |
| **MAE** | 9.233212e+04 | 1.298085e+05 | 7.846156e+04 | 5.348889e+04 |
| **MSE** | 1.624491e+10 | 3.067243e+10 | 1.400178e+10 | 6.516063e+09 |
| **RMSE** | 1.274555e+05 | 1.751355e+05 | 1.183291e+05 | 8.072214e+04 |

After checking the accuracy of every model, I can say that **Random Forest** model is best fit for predicting the Price of KC Housing.

**GitHub link:**

<https://github.com/urvinil333/-kc_house_price_10>

**Reference:**

<https://www.kaggle.com/harlfoxem/housesalesprediction>

<https://udspace.udel.edu/bitstream/handle/19716/21667/RR17-10.pdf?sequence=1&isAllowed=y>

<https://geodacenter.github.io/data-and-lab/KingCounty-HouseSales2015/>

[6 essential steps to the data mining process - BarnRaisers, LLC (barnraisersllc.com)](https://barnraisersllc.com/2018/10/01/data-mining-process-essential-steps/)