# Project Report-5502

**Project Name:** Housing PriceEstimation

**Team Members:**

**Keerthini Veluru** (Keerthiniveluru@my.unt.edu) – Data Wrangling, pre-processing

**Pravallika Bollu** (Pravallikabollu@my.unt.edu) – EDA and visualizations

**Tanuja Madala** (NagaLakshmiTanujaMadala@my.unt.edu) – Data Preparation

**Sravya Jampana** (Sravyajampana@my.unt.edu) – Model development and Evaluation

Project proposal, Design and Final Report will involve collective effort of all the team members.

**Communication:**

* Discord
* E-Mail
* Messages
* Zoom meetings on weekend to discuss developments in the project

**Workflow:**

**Includes -** Data pre-processing, Data preparation, Model training and refinement, Prediction, Model Evaluation and Final Project presentation

Diagram

Description automatically generated

**Project Abstract:**

The real estate sector is an important industry with many stakeholders ranging from regulatory bodies to private companies and investors. Among these stakeholders, there is a high demand for a better understanding of the industry operational mechanism and driving factors. Today there is a large amount of data available on relevant statistics as well as on additional contextual factors, and it is natural to try to make use of these to improve our understanding of the industry.

In some cases, non-traditional variables have proved to be useful predictors of real estate trends. For example, it is observed that Seattle apartments close to specialty food stores such as Whole Foods experienced a higher increase in value than average. This project can be considered as a further step towards more evidence-based decision making for the benefit of these stakeholders. The project focused on assessment value for residential properties in King County, the 12th-most populous in the United States, between years 2014-15. King county is predominantly urban and suburban county in Washington state.

**Data Specification:**

This Project uses **SUPERVISED** Learning which helps to solve the problem a lot of the potential users face ‘Estimation of housing price’. With the housing market, being so diverse and with so many factors influencing the prices coming into play, one can easily be overwhelmed. By the means of this project, we want to highlight the important components of such a market in a purely unbiased, data-driven way. The main objective is to obtain a model that can help the customers to

**About the Dataset:**

Source: https://www.kaggle.com/harlfoxem/housesalesprediction

The dataset has total 21 columns and 21,613 records

|  |  |
| --- | --- |
| <class 'pandas.core.frame.DataFrame'>  RangeIndex:21613 entries, to 21612  Data columns (total 21 columns): |  |
| id | 21613 non-null int64 |
| date | 21613 non-null object |
| price | 21613 non-null float64 |
| bedrooms | 21614 non-null int64 |
| bathrooms | 21615 non-null float64 |
| sqft\_living | 21616 non-null int64 |
| sqft\_lot | 21616 non-null int64 |
| floors | 21618 non-null float64 |
| waterfront | 21616 non-null int64 |
| view | 21617 non-null int64 |
| condition | 21618 non-null int64 |
| grade | 21619 non-null int64 |
| sqft\_above | 21620 non-null int64 |
| sqft\_basement | 21621 non-null int64 |
| yr\_built | 21622 non-null int64 |
| yr\_renovated | 21623 non-null int64 |
| zipcode | 21624 non-null int64 |
| lot | 21628 non-null float64 |
| long | 21629 non-null float64 |
| sqft\_living15 | 21623 non-null int64 |
| sqft\_lot15 | 21624 non-null int64 |
| dtypes: float64(5), int64(15), object(1) |  |
| memory usage:3.5+MB |  |

**Project Design:**

**1.Tools, Technology, and frameworks:**

We followed the CRISP DM methodology to develop the machine learning model.

Technology — Jupyter notebook for code submission

Language — Python

Modules/Libraries — pandas, SciKit-Learn, Matplotlib, Seaborn, Numpy

API — statsmodels.api

Google Doc’s - Store the code

Google Collab - Version Control

Memory — 16GB RAM and above is recommended, but 8GB will do the job.

Graphical User Interface

Hardware — minimum i5 processor, but i3 will do the job

Internet

**2.** **Data Cleaning:**

There are no null values, but we found some inconsistencies in the dataset. We observed that the datatype of date variable is object in the dataset, we changed it into datetime format

Table

Description automatically generated with low confidence Graphical user interface, text, application

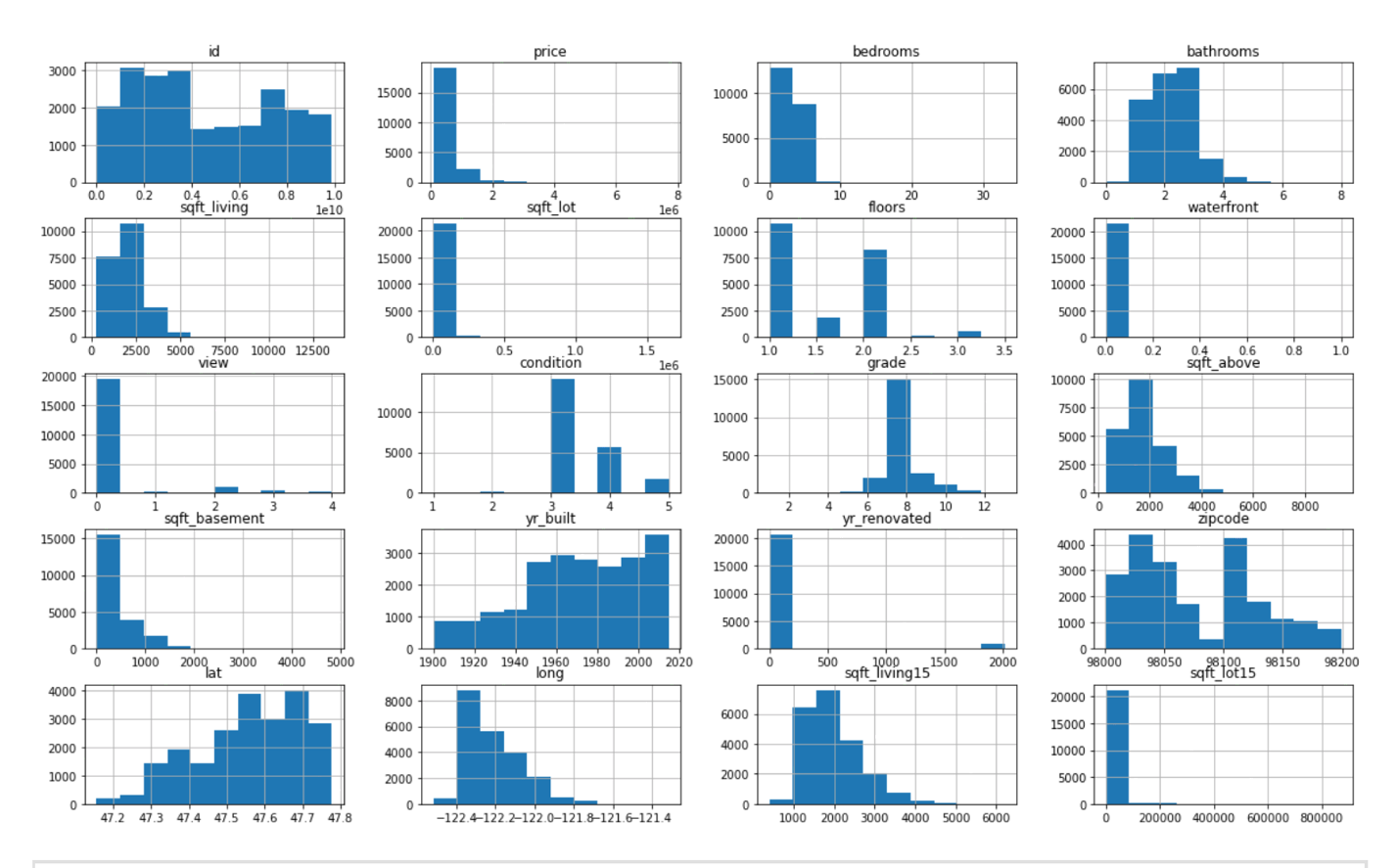
Description automatically generated

**3.Exploratory Data Analysis and Visualization:**

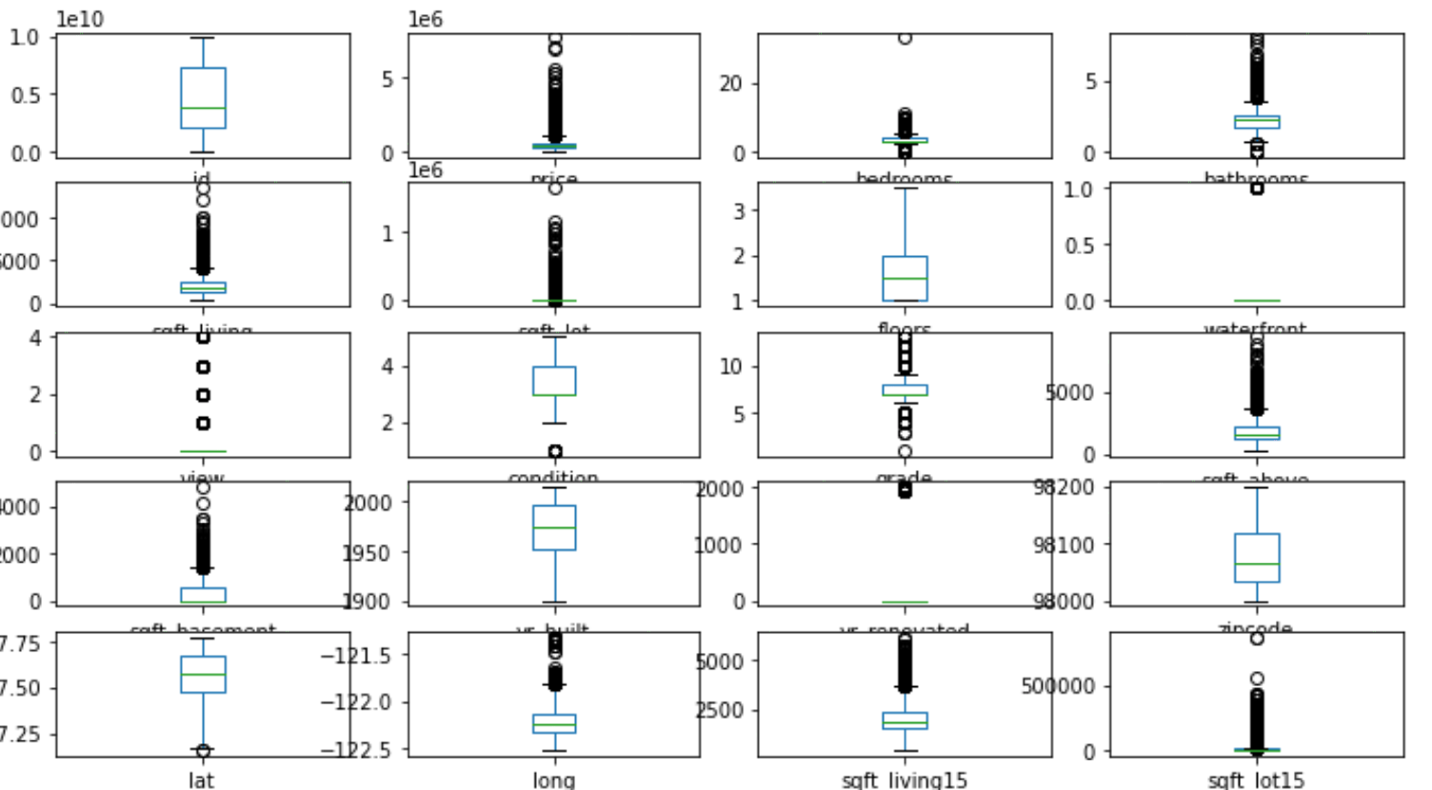
In total, we have 21 columns and 21,613 records in the dataset. The summary statistics of the numerical variables in the dataset is analyzed carefully to draw conclusions and make necessary assumptions required for the successful accomplishment of the project objective – estimation of residential housing price.

**Univariate Analysis:**

**Histogram –** To analyze the distribution



**Boxplot**- To analyze the outliers



We wanted to include only the Residential properties for this project. And to identify whether if a house is Residential property or not, we made the following assumptions:

* A Residential house would have a minimum of 1 bedroom and 1 bathroom – So records that have the number of bedrooms and bathrooms as 0 are eliminated from the dataset.
* We observed a record, where the number of bedrooms is 33 and the living area (sq\_ft living) is just 1620. This record could either be an error or that property could be for commercial use, like a motel. So, we removed this record from the dataset to protect the integrity of our study.

We removed all such inconsistent records in this project. We are now left with 21,596 records that are used to develop the linear regression model.

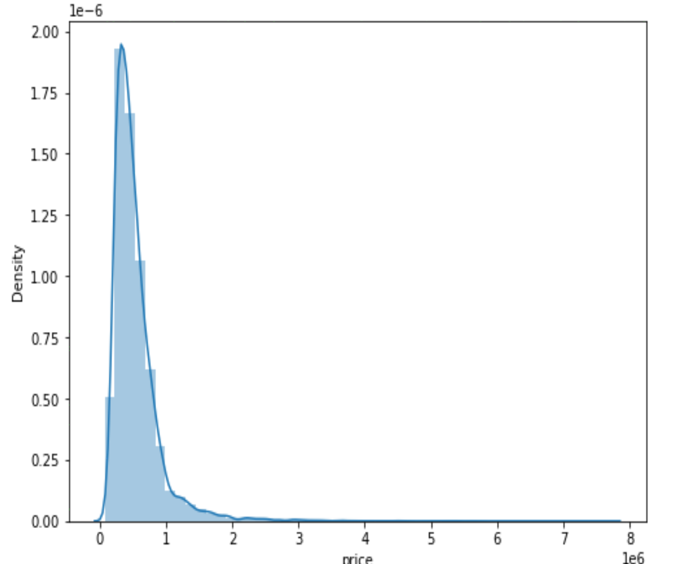
Table

Description automatically generated

Table

Description automatically generated

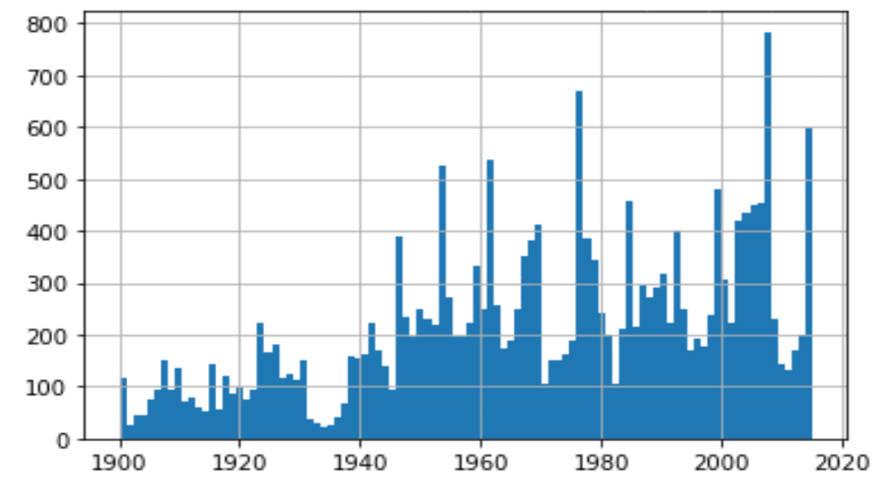
**Target Variable:** Housing Price distribution:



It is evident from the graph that most of the houses are within a million dollars, and there are very small number of houses that are more than 2 million. With this distribution, we can understand the general price range for houses in the area.

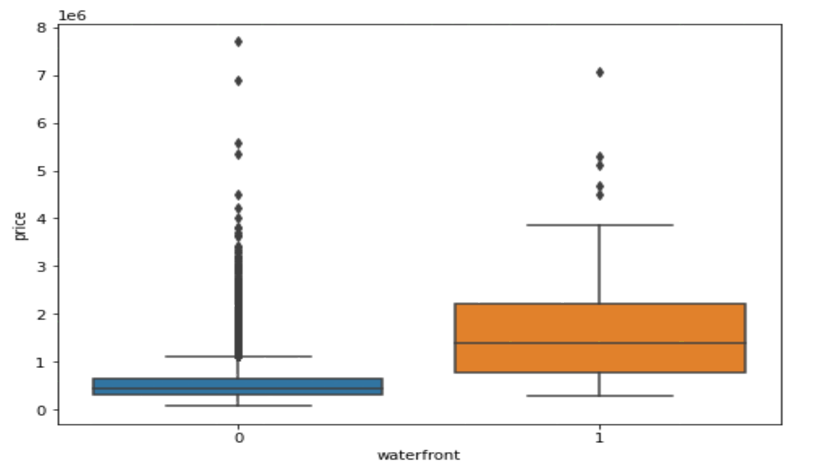
**Relation between Price and other Categorical variables:**

* **Price Vs year\_built**



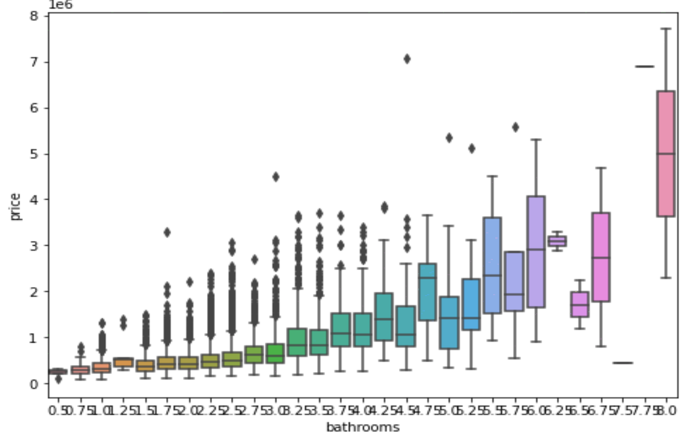
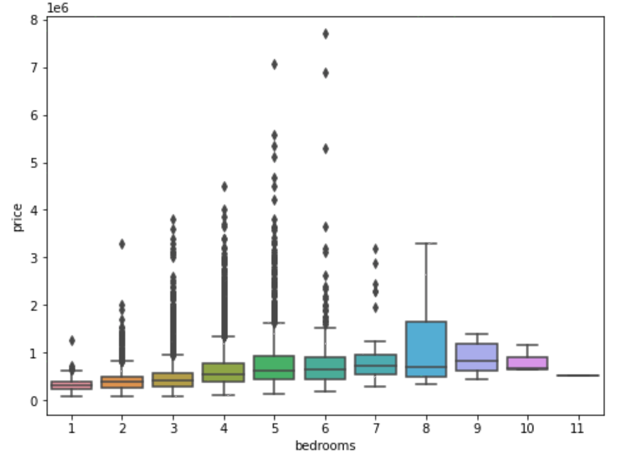
Observation: The houses built in the recent years are more expensive than old properties. So, we can understand that age of the property is important when it comes to pricing

* **Price Vs Waterfront:**

****

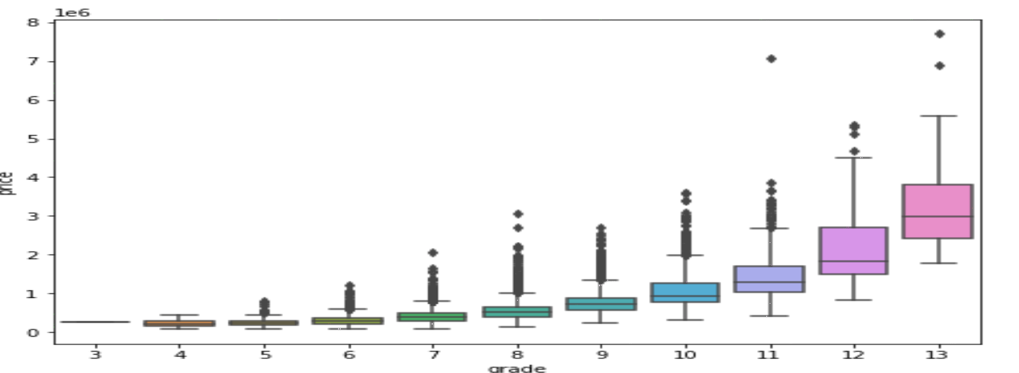
Observation: Waterfront properties come at a premium

* **Price Vs Bedrooms and Bathrooms:**



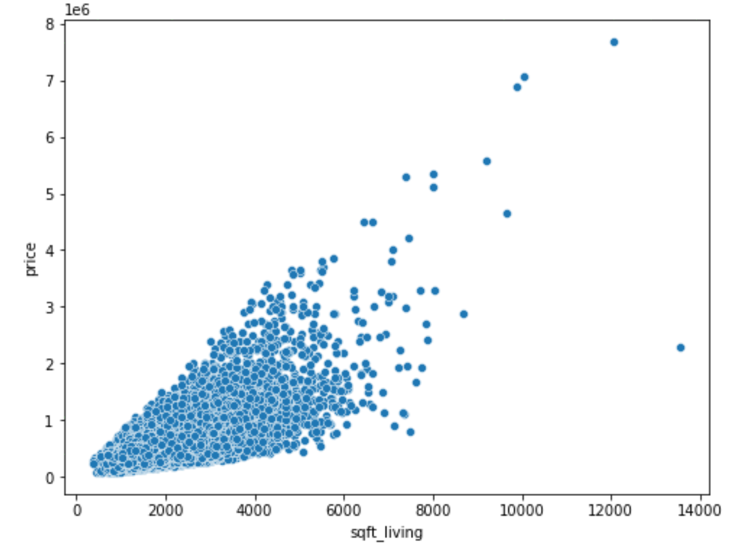
Observation: There is a positive linear relationship between price and these features. There are some exception cases, but in general we can say there is a positive correlation between the variables.

* **Price Vs Grade:**



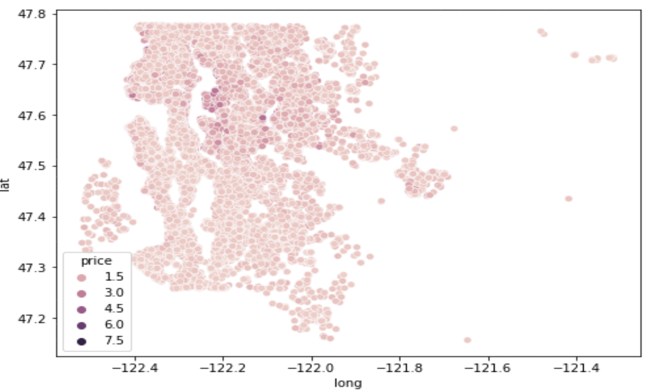
Observation: It is evident that better construction quality/grade, higher the price

* **Price Vs sqft\_living:**

****

Observation: It is evident that more the living area, higher the price

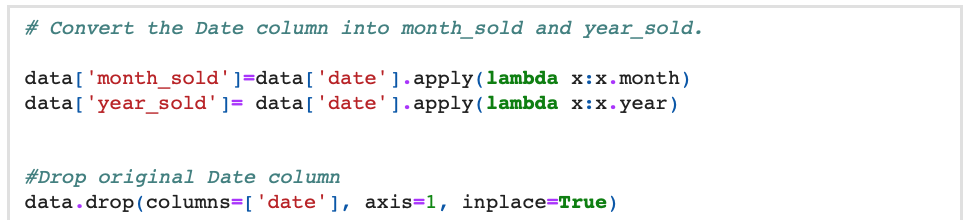
* **Price Vs Location:**

****

Observation: Most houses with higher price are distributed near Seattle city. This may be because, Seattle is a big city with a lot of corporates like Amazon, Microsoft, Costco.

**4. Feature Engineering and Selection:**

* The dataset has captured the housing price values for one year. We wanted to split the date into month and year columns to see if there are price increases for any part of year. So, we added two new features month\_sold and year\_sold. Then dropped the date column.



* We added a new feature ‘Age’ (how old the property is) for better analysis. Age is calculated from year\_built variable: The latest built year is 2015, so age = (2015 – year\_built). Then dropped the year\_built column

Text

Description automatically generated with medium confidence

* We converted the year\_renovated into a categorical feature ‘renovated’, which takes values either 0 or 1.

’1’ – For homes either renovated within past 10 years or built within the past 5 years.

‘0’ – For homes that don’t fall under category 1 (Assumption is generally houses should be renovated every 10 years)

Graphical user interface, text

Description automatically generated

* Seattle is important city in King County, where most houses are listed. From the Price Vs latitude and longitude comparison, we can find that houses near to center of the Seattle city (downtown area) have higher prices. So, for a better analysis of house prices, we calculated distance variable from coordinates of house and coordinates of Seattle downtown area. This is done using Geopy module in python. This feature is going to be key in understanding the housing market.

Graphical user interface, text, application

Description automatically generated

Finally, to select all the important features that affect house prices, we developed a correlation heat map.

Graphical user interface

Description automatically generated with low confidence

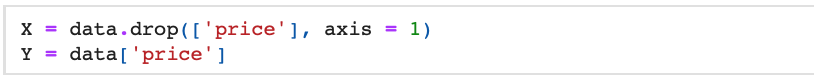
Basing on correlation of all the features with our target variable price, we identified 16 variables that influence the house prices.

**Text, letter

Description automatically generated**

**5. Model Development and Evaluation:**

We developed a linear model for our price prediction project. We chose the Linear regression model, since it has produced better results. Even before developing a model, we should separate the dataset into input, output arrays. The input array has all the chosen features from the above step except the price. And Price will be the output array

****

Next step is, to split the dataset into train and test sets. We’ll train the model using the train set and then evaluate the model.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

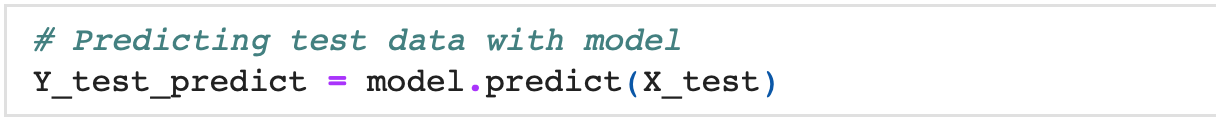
Description automatically generated

Even, we can check the normality of errors using a distribution plot – difference between actual target values from train set and the predicted values of the target variable. i.e, Y\_train – Y\_predict

Chart, histogram

Description automatically generated

Once, we are satisfied with the performance of our model we could predict the values for the test set, using the trained model. And then Evaluate the model and compare the R-squared values



Graphical user interface, text, application

Description automatically generated

**6. Prediction:**

Finally, we have a good working model which can help us predict the house price with certain features. Let’s see how this works,

Mr. X wants to buy a house in king county. He has certain requirements: 4 bedrooms, 3 bathrooms, A living area of 3400 sqft, a lot size of 7000 sqft, 2 Story home with a water-front, and a decent view of category 4, and construction grade 3. The area above basement should be 3000sqft and the basement area of 400sqft. The neighbors living in the vicinity of 15 houses should be around living-2800sqft and lot-size 6000sqft. The house should be 5years old (built year:2010) and should be in 5.5 miles distance range from Seattle downtown for easy commute.

Graphical user interface, text, application, chat or text message

Description automatically generated

**Model predicts that the price of such house is estimated at 1.19 million dollars**

**Project Milestones:**

Coming up with a project idea. We always wanted to pursue a prediction model, but with lots of data and options available deciding on the best one was the challenge. Once we had finalized the idea, we had a few notable milestones.

* Obtaining the cleaned data set
* Finding outliers in number of bedrooms and bathrooms. We made an assumption that a residential house will have at least 1 bedroom and 1 bathroom. And also, we have seen there are 33 bedrooms in one house, so may be that is a Motel/commercial property. So, we deleted all such records from the dataset.
* Coming up with new feature creation ideas: Inferring that the houses built after 2000 are more expensive. We wanted to check if age of the property affects price so calculated age of the property for better analysis.
* Developing other new features like month\_sold, year\_sold, renovated, and distance. Assessing which features are important and influential to house price by looking into the correlation matrix.
* Deciding on the method to split the dataset into test and train sets. Splitting the dataset into training and testing data sets. Fitting linear regression model to trained data set followed by prediction of the model.
* Evaluating the models in terms of accuracy and deciding which model works best for this dataset. (R- squared, F-statistic, p-value,) with an R-square score of 71.
* Calculating residuals and checking the normality of errors.
* Final prediction using arbitrary values for all selected features.

**Repository/Archive:**

<https://drive.google.com/drive/folders/1URpNHsc6g8NghRf-LbWidJgyvD4cya_X?usp=sharing>

**Resources and Related Projects:**

All the resources used are open source and accessible online. We are taking inspiration from some of the inferences and visualization ideas from these resources

Google slides project presentation: <https://docs.google.com/presentation/d/1A-tn-B245rW2ySFGEqE4rmoyC0AEYBKK2HoJzZkDFSo/edit?usp=sharing>

Similar machine Learning project from data Analytics class (but we are using a different dataset, just taking some inspiration from the project)

<https://drive.google.com/file/d/1PqwszIwCI5gwvi9UmFkb8hAjEsl4FAgw/view?usp=sharing>

Source Code:

<https://www.kaggle.com/skotnis/housing-sales-in-king-county-washington?scriptVersionId=44269181&cellId=7>

<https://studygyaan.com/data-science-ml/linear-regression-machine-learning-project-for-house-price-prediction>

<https://machinelearningmastery.com/linear-regression-for-machine-learning/> - Understanding Linear regression model and making predictions with linear regression

<https://www.youtube.com/watch?v=NUXdtN1W1FE> – Understanding Linear regression, its application, identifying dependent and independent variables, use cases

Mckinsey report, https://www.mckinsey.com/industries/capital-projects-and-infrastructure/ our-insights/getting-ahead-of-the-market-how-big-data-is-transforming-real-estate. - How to draw data driven insights from the real estate analytics project.

B. de Ville, Decision trees, Wiley Interdisciplinary Reviews: Computational Statistics 5 (2013), no. 6, 448–455. – Understanding the decision tree model and how it works for our project

T. G. Dietterich, an experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization, Machine learning 40 (2000), no. 2, 139– 157. - Understanding the decision tree model and how to construct a decision tree model

**Appendix**

****