**Machine Learning Capstone Project Report on House Price Prediction using Advanced Regression Techniques**

Table of Contents

**1.Introduction**3

**2.Problem Statement4**

2.1Data Set (Feature descriptions) ………………………………………………………………………… 4

**3.Project Plan ………………………………………………………………………………………………………. 6**

3.1 Flow Chart……………………………………………………………………………………………… 7

**4.Data Insights - Exploratory Data Analysis……………………………………………………………………… 8**

4.1. Five Point Summary - Statistical Description……………………………………………...…………... 8

4. 2. EDA and Visualizations………………………………………………………………….………. ….. 9

4.2.1: Univariate and Bivariate Analysis …………………………………………………………. 9

4.2.2: Geographical data & Observations………………………………………………………… 9

4.2.3: Correlation plots (Heat Map) ...……………………………………………………………. 9

4.3 outliers Summary……………………………………………………………………………………… 12

**5: Model Selection and Building……………………………….………………………………………………… 13**

5.1 Selection of Model………………………………...………………………………………………… 13

5.2 Model Baselining………………………………….………………………………………………… 13

5.2.1 Feature Scaling………………………….………………………………………………… 13

5.2.2 Data Segregation………………………...………………………………………………... 13

5.2.3 Model Training and Testing…………….…………………………………………………. 13

**Chapter 1 : INTRODUCTION**

In machine learning (ML), statistical methods are used to empower machines to learn without being programmed explicitly. The field focuses on letting algorithms learn from the provided data, collect insights, and make predictions on unanalyzed data based on the gathered information.

In general, ML is based on three key models of learning algorithms:

● Supervised learning - a dataset is present with inputs and known outputs

● Unsupervised learning - the machine learns from a dataset that comes with input variables only

● Reinforcement learning - algorithms are used to select an action

This project is implemented using supervised machine learning algorithms since the given dataset has input variables with known output price. The outcome of our project is to make predictions on the sales prices of the houses of Seattle with the dataset provided.

We (below members) are working on this project as an extension of our PGML Course.

* Indra Chatterjee
* Manju Kiruthiga
* Ranganathan Sriraman
* Sunil Kumar
* Shyamla

**Chapter 2: Problem Statement**

**Creation of Intelligent Regression based data model to predict house/home prices on basis of sales data in Seattle region from 2014 to 2015. Data models take into account various features like area, location, amenities and condition**

**2.1 Data Set (Feature descriptions)**

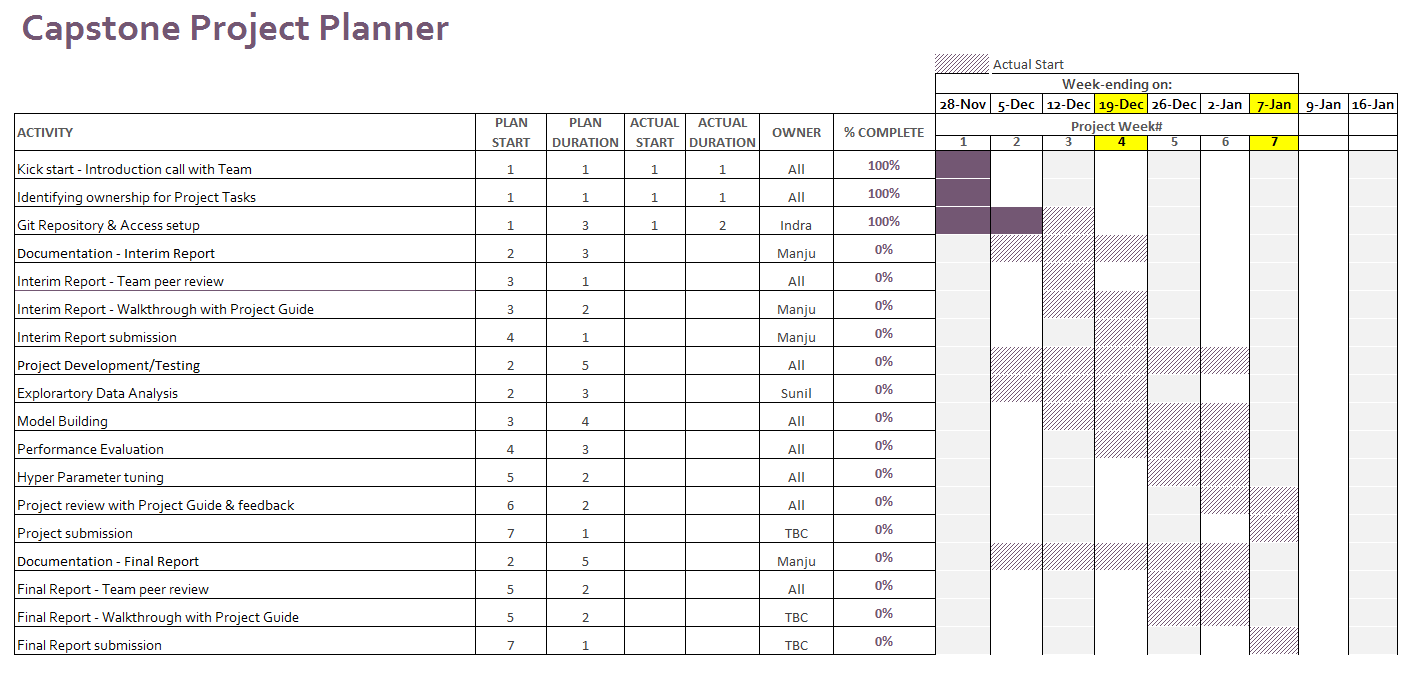
The dataset contains House Sale data of Seattle, WA area from 2014-May through 2015-May. The shape of the dataset is 21613\*23 (23 attributes and 21613 observations). Below information would describe the features given.

**Price is the target attribute and the remaining acts as predictors.**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Feature Name** | **Feature Description** |
| 1 | cid | a unique 7 to 10 digit ID representing the house property |
| 2 | dayhours | represents when the date/time when the sale happened |
| 3 | price | selling price of the house - This will be our target variable |
| 4 | room\_bed | represents number of bedrooms in the house |
| 5 | room\_bath | represents number of bathrooms in the house |
| 6 | living\_measure | represents square footage of living area in the home (carpet area) |
| 7 | lot\_measure | represents square footage of the lot |
| 8 | ceil | represents number of floors/levels in the house |
| 9 | coast | is a boolean variable representing whether the house has a water front or not |
| 10 | sight | is a boolean variable representing whether the house has been viewed by potential clients or not |
| 11 | condition | represents overall condition of the house on a scale of 5 |
| 12 | quality | grade given to housing unit based on grading system ranges from 1 to 13 |
| 13 | ceil\_measure | represents square footage of the house except basement |
| 14 | basement | represents square footage of the basement |
| 15 | yr\_built | year when this house was built originally, it basically represents the age of the house |
| 16 | yr\_renovated | year when the house was renovated. We assume that the last renovation year is captured in the data set |
| 17 | zipcode | represents zipcode of the property |
| 18 | lat | represents latitude of the property |
| 19 | long | represents longitude of the property |
| 20 | living\_measure15 | represents square footage of living area in the home (carpet area) after renovation 2015 |
| 21 | lot\_measure15 | represents square footage of lot area in the home after renovation 2015 |
| 22 | furnished | is a boolean variable representing whether the property is furnished (personal property) |
| 23 | total\_area | represents total sum of both living and lot measure |

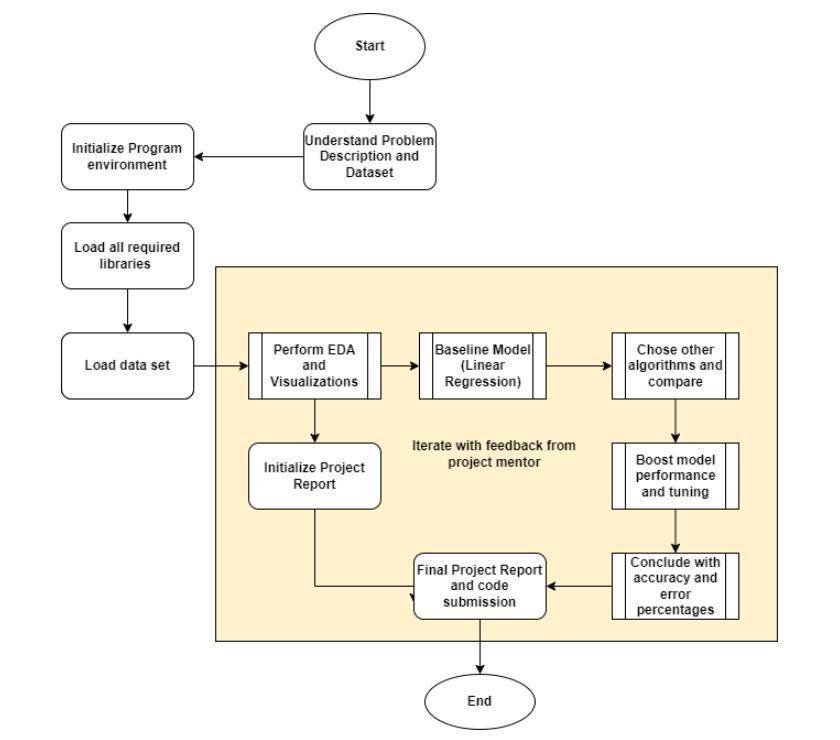
**Chapter 3: Project Plan**

* Here is the snapshot of the project plan
* The project plan will be reviewed and updated periodically
* It will be maintained in GitHub



****

**3.1: Flow Chart**

Flow chart below depicts the project life cycle.

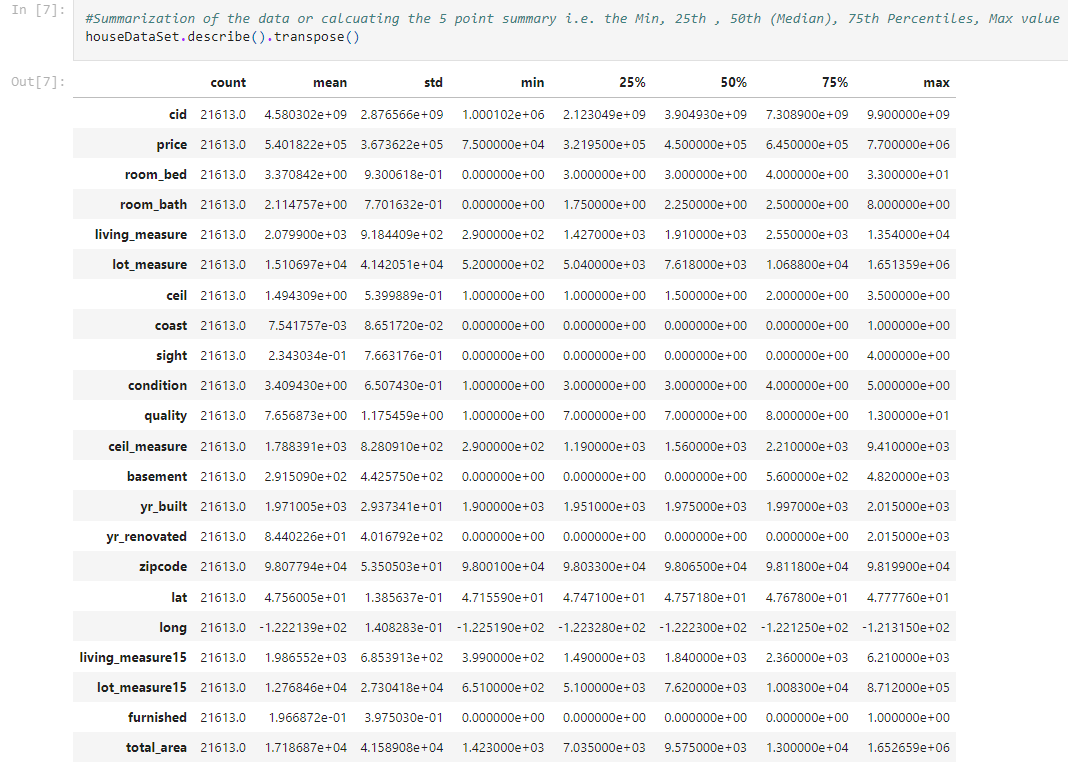
**Chapter 4: Data Insights - Exploratory Data Analysis**

### **4.1. Five Point Summary - Statistical Description**

The five-point summary involves the calculation of 5 summary statistical quantities, namely:

* **Median**: The middle value in the sample, called as the 50th percentile or the 2nd quartile.
* **1st Quartile**: The 25th percentile.
* **3rd Quartile**: The 75th percentile.
* **Minimum**: The smallest observation in the sample.
* **Maximum**: The largest observation in the sample.

Below code and output depicts the 5-point summary statistical quantities of the given House price prediction dataset.



### **4.2. EDA and Visualizations**

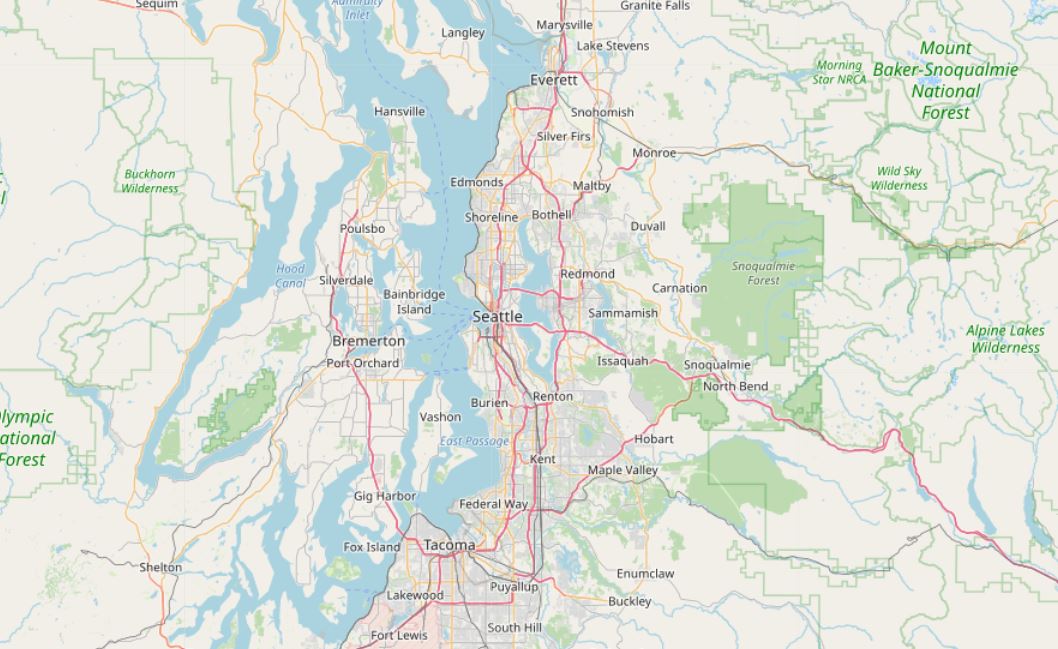
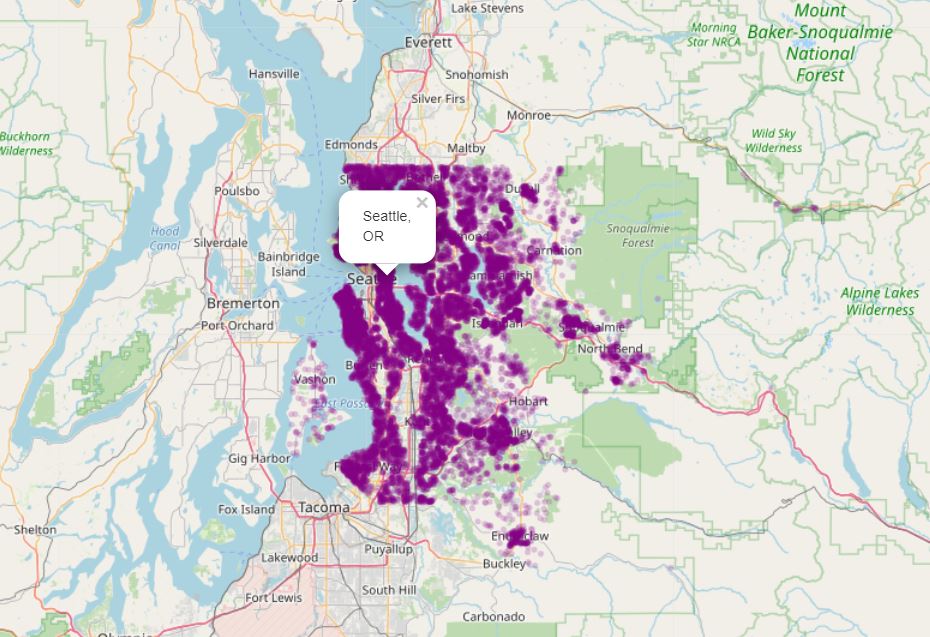
### **4.2.1: Univariate and Bivariate Analysis (Selective features)**

TBC

**4.2.2: Geographical data & Observations**

In the Dataset given, we have Latitude, Longitude & Zipcode of each property. We examined further to see which Geographical area this data represents. The Geo information might not be directly relevant to the model building, it was interesting to find out that the data is from properties in Seattle, WA, USA area. (zipcodes 98001 to 98199). Plotting this data using Folium library over the Seattle map to see the distribution of the properties.

Plain Map of Seattle on the left and properties in the dataset plotted on the right

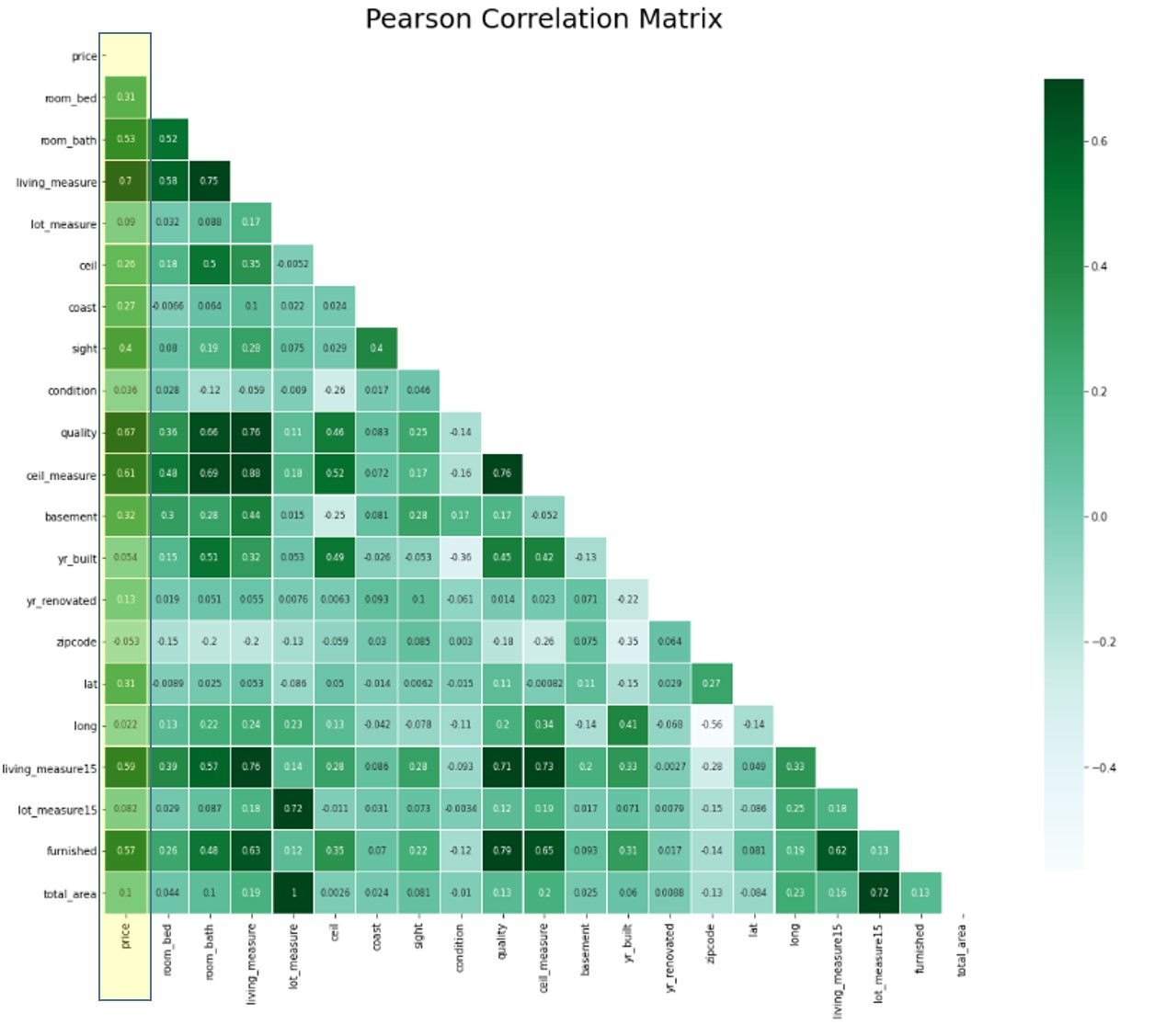
**4.2.3: Correlation plots (Heat Map)**

Correlation is a statistical term describing the degree to which two variables move in coordination with one another. If the two variables move in the same direction, then those variables are said to have a positive correlation. If they move in opposite directions, then they have a negative correlation.

Here we are going with Pearson Correlation. There are other types of correlation also available like

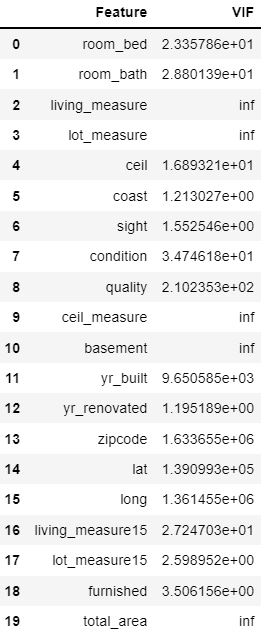
* Positive, Negative or Zero Correlation
* Linear or Curvilinear Correlation
* Scatter Diagram Method
* Pearson's Product Moment Co-efficient of Correlation
* Spearman's Rank Correlation Coefficient

|  |  |
| --- | --- |
| **correlation coefficient** | **Means** |
| 1 | Features are highly Positively correlated |
| 0 | Features are not correlated |
| -1 | Features are highly Negatively correlated |



Observation from the Correlation heat map chart (Target Variable Vs Features):

* Columns like living\_measure (0.7), quality (0.67), ceil\_measure(0.61), living\_measure15(0.59), furnished (0.57), rooms\_bath (0.53) are Highly Positively correlated
* Columns like lot\_measure, condition, zipcode, longitude, yr\_built, lot\_measure15, total\_area are Not correlated with Price
* Interesting observation here is on the yr\_built feature. We usually assume that age of a property would be a key factor in determining the price. however, the yr\_built has a correlation closer to 0. We will examine further during the model building to see if any new column ("building\_age") can be derived based on yr\_built and yr\_renovation
* Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model. This means that an independent variable can be predicted from another independent variable in a regression model. From the above heat map, it is clearly visible that there is a possibility for Multicollinearity hence we checked for VIF (Variable Inflation Factor)



Based on above correlation heat map/VIF and summary listed, we plan to drop all parameters with a correlation less than 0.25 with target variable 'price'.

##### cid: Since this is a unique representation on house and it will not have any implication on the model. Correlation is negative

##### dayhours: It represents the date house was sold, which really does not help in house price prediction also, data is in string form cannot be used in computation

##### zipcode: Looking into the data and searching the zip-code we have seen that it specifically belongs to a particular area and hence we can drop this parameter too

##### "lot\_measure" and 'lot\_measure15': Weak correlation when it comes to predict the price of the house. We can say that customers do not care on the lot\_measure compare to living measure

##### "lat" and "long": Since they represent a 3D view of the house and buyers do not really care about the lattitude and longitude degree of the house. This would reduce the compute power needs

##### 'condition': Very weak correlation with price

##### 'yr\_built' and 'yr\_renovated': Both have very weak correlation when it comes to predicting price of the house

##### 'total\_area': Since it sums up both living\_measure and lot\_measure and since lot\_measure is the bigger value and does not have a strong co-relation therefore, total\_area itself has a week correlation

## **4.3 Outliers Summary (Identification and handling)**

**TBC**

**Chapter 5: Model Selection and Building**

**5.1 Selection of Model**

The problem at hand is prediction of house property prices. Here the Target variable is given & the Price feature is a continuous variable. Hence this would be a **Supervised - Regression problem**.

For the Base model, we wanted to start with a **Simple Linear Regression**.

**5.2 Model Baselining**

**5.2.1 Feature Scaling**

Feature scaling is a **method used to normalize the range of independent variables or features of data.** In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

We are using **zscore function from scipy** python library for feature scaling.

**5.2.2 Data Segregation**

We should start with separating features for our model from the target variable. Notice that in our case all columns except *‘Price’* are features that we want to use for the model.

* Our target variable is ‘*Price’* - y\_data
* As discussed in the 4.2.3 section, after dropping the mentioned variables, we have total 11 predictor variables for our model building – x\_data

**5.2.3 Model Training and Testing**

We are using **train\_test\_split** from the function **scikit-learn** to divide features data (x\_data) and target data (y\_data) even further into train and test.

**Controlling test-train split fraction:** We can control the train\_test split fraction by using the ***test\_size*** parameter. Note that we had it set to 0.3 in our example. It can be any number between 0.0 and 1.0. Do not need to specify the fraction for the train set as by the default it will use all the remaining data that is not taken for the test set.