***ML CAPSTONE PROJECT REPORT- HOUSE PRICE PREDICTION***

***(ADVANCED REGRESSION TECHNIQUES)***

**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 machine learning algorithms

● Unsupervised machine learning algorithms

● Reinforcement machine learning algorithms

In the first model, a dataset is present with inputs and known outputs. In the second one, the machine learns from a dataset that comes with input variables only. In reinforcement learning model, algorithms are used to select an action. This project is implemented using supervised machine learning algorithms. 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.***

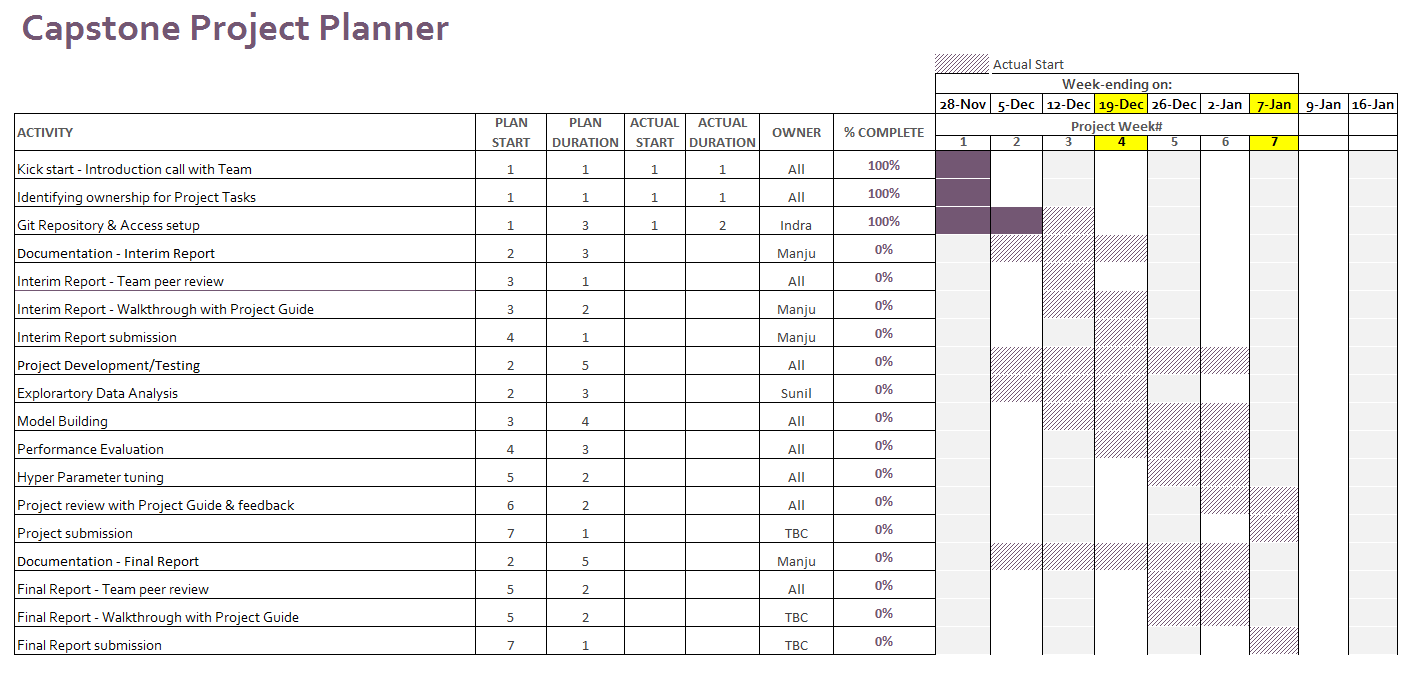
**2.1 Data Set (Feature descriptions):**

In the given dataset, there are 23 attributes and 21613 records. Below information would describe the features given. **Price is the target attribute and the remaining acts as predictors.**

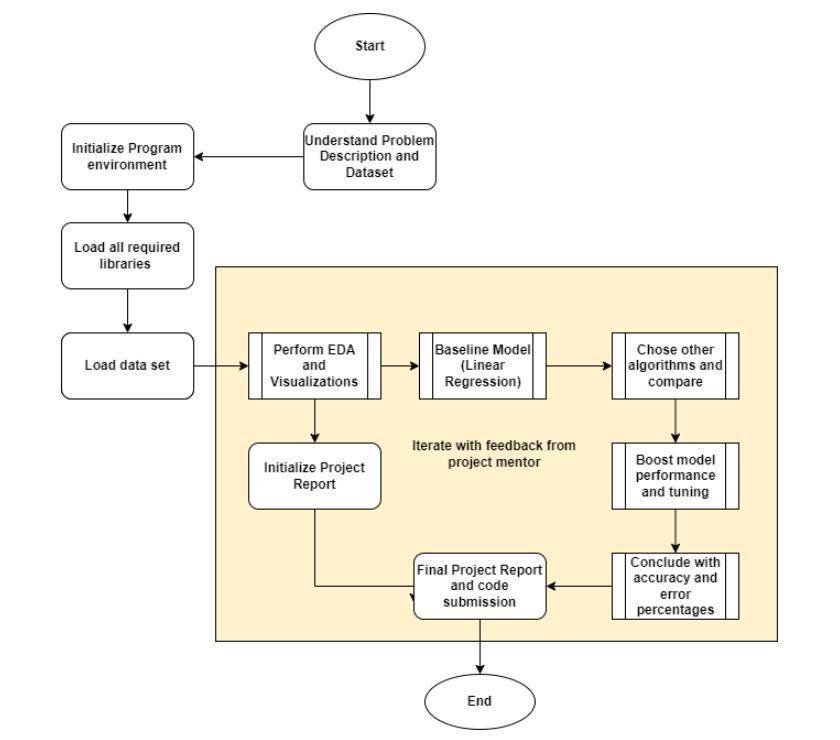
1. **cid** : represents a unique ID for the house
2. **dayhours**: represents when the date when the house was sold
3. **price** : Selling price of the house when it was sold. This will be our target variable
4. **room\_bed, room\_bath** : Represents number of bedrooms and bathrooms in the house respectively
5. **living\_measure**: represents square footage of the home,
6. **lot\_measure**: represents square footage of the lot
7. **total\_area**: Would be the sum total of lot\_measure and living\_measure, hence total\_area = living\_measure + lot\_measure
8. **living\_measure15, lot\_measure15** : represents area in 2015 i.e. living room area and lot size area.
9. **ceil**: Total floors or levels in the house
10. **coast**: Is a boolean variable representing whether the house has a water front or not.
11. **sight**: Is another boolean variable representing whether the house has been viewed by clients or not.
12. **ceil\_measure**: Square footage of the house except the basement.
13. **condition**: Overall condition of the house a number.
14. **Quality**: Grade given to housing unit based on grading system
15. **yr\_built**: Year when this house was built, it basically represents the age of house.
16. **yr\_renovated**: Year when the house got renovated. We assume that the last renovation year is captured in the data set
17. **zipcode**: simple zip code.
18. **lat, long**: Co-ordinates of the house.

**Chapter 3: Project Plan**

Below picture depicts the project plan we have developed for our House Price Prediction – ML Capstone Project. We will be updating our work/improvements in weekly basis.



**3.1: Flow Chart:**

Flow chart shown below explains the sequence of operations to successfully predict the house price for given region based on sales data for a specific period.

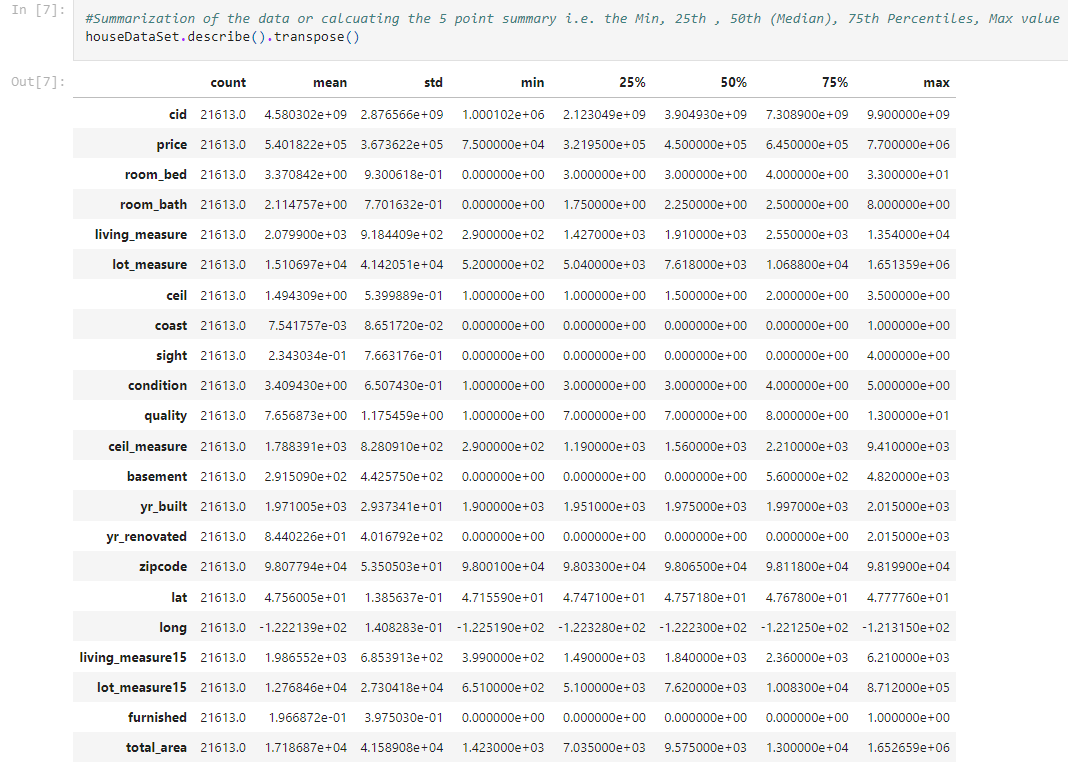
**Chapter 4: Data Insights - Exploratory data analysis**

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

The five-number 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 Feature Selection:**

Feature selection process aid us in our mission to create an accurate predictive model. They help by choosing features that will give as good or better accuracy whilst requiring less data.

#### Selecting important attributes and drop some parameters from the data set such that our prediction model does not miss out on information.

##### "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