**Final Project**

**Project Title**: - House Price Prediction

For the House price prediction, I would like to use the below datasets

* Kaggle Datasource available from https://www.kaggle.com/aariyan101/usa-housingcsv

**Data Set Details**

The dataset has the below fields

* Avg. Area Income:
* Avg. Area House Age
* Avg. Area Number of Rooms
* Avg. Area Number of Bedrooms
* Area Population
* Address

Business Problem/Data: With the growing price of houses every year it is very appropriate to understand what different parameters impact the price of houses. By understanding different variables which directly impact House Price it will help customers to understand what parameters must be considered for purchasing a house thereby planning budget efficiently and getting a good house as per the price.

**Graphical Analysis:**

##### seaborn Library is used to study and analyze different columns of the Dataset.

Bar plot from seaborn is used to study the relationship between “Area population” and “price” variables on the graph

Line plot from seaborn is used to study the relationship between “Area population” and “price” variables on the graph

Bar plot from seaborn is used to study the relationship between “Area population” and “price” variables on the graph

Dist plot from seaborn is used to understand Housing price distribution which is a bell-shaped curve which indicates the data is normally distributed

Heat plot from seaborn is used to understand the co-relation between variables of a Data set

##### 

**Dimensionality & Feature Reduction and Feature Engineering:**

After studying the initial data, it is understood that “field” is mostly textual which is not required for Analysis so removing the field from the Dataset using “del” method of Pandas Data Frame

It is also identified that there are NAN values from the rows of the columns, so the data set is clear for Model Training

**Model Selection & Evaluation:**

I have selected the Linear Regression for training the Data set and testing the data set. USA\_Housing.csv Data Frame is divided into two sets for analysis one being ‘X’ as Data frame which contains all other column data except Price column

‘y’ as a Data Frame which contains ‘Price’ column only

Using sklearn the library the ‘X’ and ‘y’ Data Frames are divided into Train and Test Data sets respectively where Train Data set percentage is 70% and Test Data set percentage is 30%.

The Train Datasets are fit in to the model and prediction is performed and co-relation is calculated to understand the Price increase for a single unit increase of other column variables.

Coefficient Analysis of variables which impact the price of house in $

# 1 unit increase of Avg income in a particular area will increase the house price by $21.617635

# 1 unit increase of House Age in a particular area will increase the house price by $165221.119872

# 1 unit increase of Rooms in a particular area will increase the house price by $121405.376596

# 1 unit increase of Number of Bedrooms in a particular area will increase the house price by $1318.718783

# 1 unit increase of Population in a particular area will increase the house price by $15.225196

From the above Co-efficient analysis results it is understood that the major parameters to be taken in to consideration while purchasing a house are Age of the House in a given Locality and No of Rooms

It is understood that 1 unit increase of House Age in a given areas will increase the price by $165221.119872

1 unit increase in number of rooms in a particular area will increase the price of the house by $121405.376596

**Conclusion:**

The Model performed well when tested with Test Data and when Scatter plots are performed on the prediction of Test Data. The Scatter plot above is in linear shape which is good indication the model has performed well for Test Data based on the Trained Data

Also, Histogram plot is generated to check the on the Test price variable with respect to predictions and the shape of the curve is bell shaped. The histogram plot is bell shaped which means the data is normally distributed and the Model has performed good predictions

Mean Absolute Error, Mean squared Error and Sqaure root are generated to analyze the performance of the Model and the Error to calculate the predictions of Error Data on Data set.

MAE(Mean Absolute Error) : 81257.55795856068

MSE(Mean Squared Error): 10169125565.897734

RMSE(Square Root of the Mean Squared Error) : 100842.08231635111

From the above metrics information, the error percentage is very minimal so it can be stated that Linear Regression model performed very well.

And from the Analysis of the Co-efficient Data, it can be stated that for a medium family of 4 or 5 the options to consider while purchasing a house is No of Rooms an increase in Number of rooms will increase the price by above 100K. Age of the House are likely to increase the price of the house by above 100k possibly to rapid development of areas as many towns are moving towards the status of becoming cities with rapid urbanization preferably in a new construction with decent number of rooms might bring down the price of the house rather than purchasing an old house in cities.

**Potential Challenges which may be overlooked:**

The potential challenges which can missed are: The Data Set with latest prices is not available on the Kaggle as the current Data set considered from Kaggle can be few years old also there could be other parameters which might also impact the price of the houses like school district ratings, Parks, Maintenace of the city or Town, Property Tax which could differ from community to community these parameters are not looked in to as they are not available in current Dataset, so the above mentioned parameters might also need to be taken in to account while purchasing a house.