Introduction:

Real estate / Housing prices topic has gotten my interest. Various factors contribute to the sale prices of houses / residential properties like size of property, the localities / neighborhood, age of property, number of rooms, amenities, building quality and other factors. I would like to build a possible model to be able to predict the prices based on various factors and identify which ones are really effective.

Research questions:

1) Explore how the Sale price depend on the Area of lot / built up area
2) How does the number of rooms / kitchen / bathrooms affect the sales price
3) Explore if the number of stories in the building affect the sale price
4) How does the location / zip code affect the sale price
5) Does having the garage / pool etc. other features affect sale price
6) Explore if any other factors contribute to the sale price (e.g. quality, condition, year built etc.), if available

Addressing the problem statements and propose approach suitability:

I will still need to explore the datasets further to identify more data trimming needs and identify the variables which may have higher effect on the housing sale prices and keep only the relevant variables. This is because one of the dataset has 81 columns in it and hence keeping the relevant information will help for better results. I will be exploring the datasets to identify the outliers, missing values / null values and clean the data. Also exploration through variety of graphs and methods could help in determining more relevant variables and in cleaning relatively small amount of data. After these steps, preparing the model and evaluating it will help in predicting the house prices and confirming the factors affecting the prices.

Data Sources:

1) https://www.kaggle.com/camnugent/california-housing-prices
   - Following variables are part of this dataset, as provided on the source link

1. longitude: A measure of how far west a house is; a higher value is farther west

2. latitude: A measure of how far north a house is; a higher value is farther north

3. housingMedianAge: Median age of a house within a block; a lower number is a newer building

4. totalRooms: Total number of rooms within a block

5. totalBedrooms: Total number of bedrooms within a block

6. population: Total number of people residing within a block

7. households: Total number of households, a group of people residing within a home unit, for a block

8. medianIncome: Median income for households within a block of houses (measured in tens of thousands of US Dollars)

9. medianHouseValue: Median house value for households within a block (measured in US Dollars)

10. oceanProximity: Location of the house w.r.t ocean/sea

2) https://www.kaggle.com/arnavkulkarni/housing-prices-in-london
- Following variables are part of this dataset, as provided on the source link

1. Property Name

2. Price

3. House Type - Contains one of the following types of houses (House, Flat/Apartment, New Development, Duplex, Penthouse, Studio, Bungalow, Mews)

4. Area in sq ft

5. No. of Bedrooms

6. No. of Bathrooms

7. No. of Receptions

8 Location

9. City/County - Includes London, Essex, Middlesex, Hertfordshire, Kent, and Surrey.

10. Postal Code

3) https://www.kaggle.com/gpandi007/usa-housing-dataset?select=housing_train.csv
- There are about 81 variables in the dataset. So, I am not mentioning all of them here. Some of the common ones are Zoning, Lot Area, Lot Shape, Neighborhood, Housestyle (number of stories), Utilities, Overall Quality, condition, Year Built and renovated, Total Basement area, Pool area, Sale Price etc.

R Packages:

ggplot2 for visualization

For models building and evaluations of the possible models:

QuantPsyc, car, caTools

Types of Plots

As of now, I am planning for below plots for analysis using various variables and determination of factors.

Histograms

Scatter Plots

Future learning :

I will be exploring to see if there are any additional techniques / packages which we have not used so far and learning more about them, if those can be used effectively for this analysis. Also, if I come across more factors affecting the sale prices during exploration, I will be using them as well.