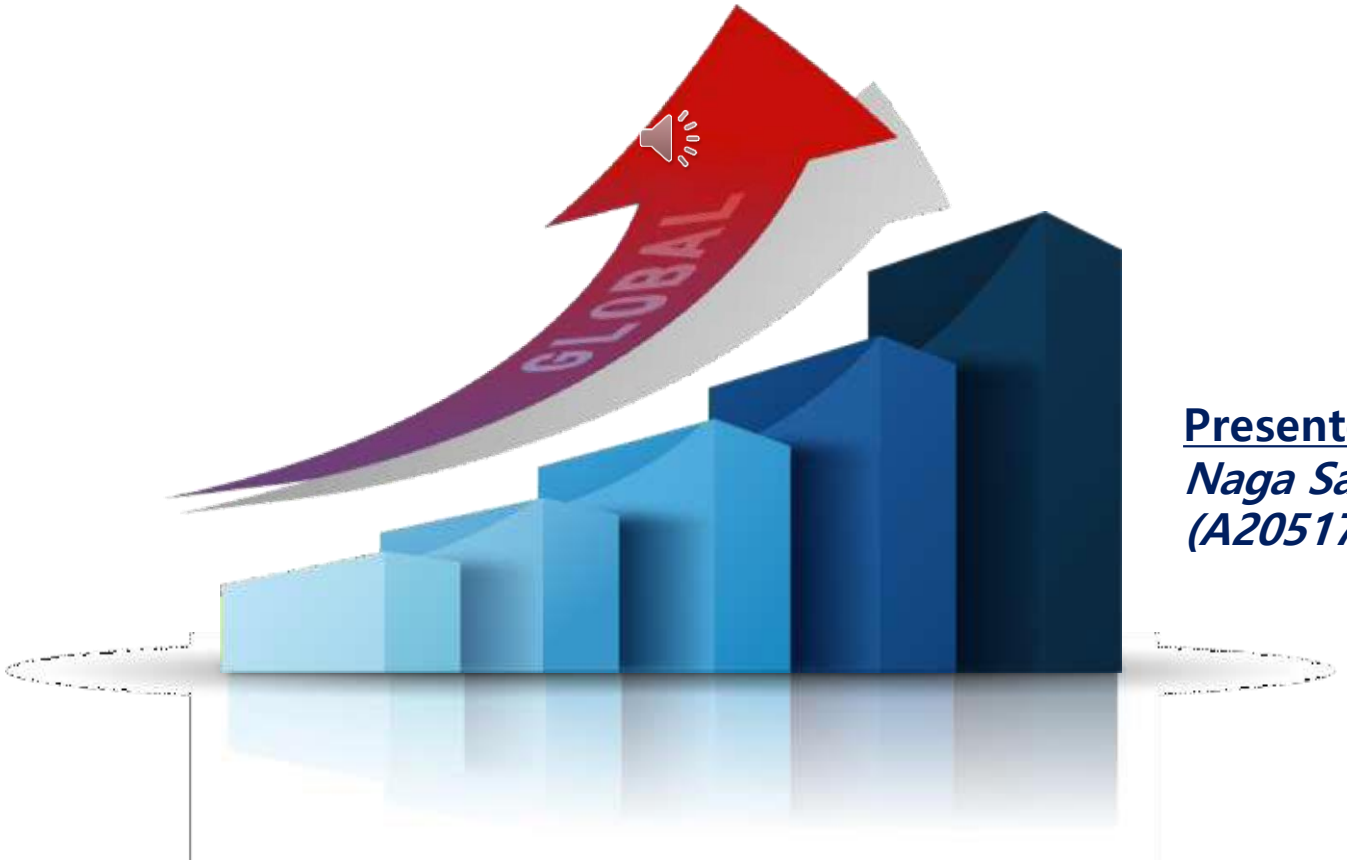
USA

HOUSE PRICE PREDICTION



Submitted to :
Prof. Yong Zheng

Presented By :
Naga Satya Silpa Annadevara
(A20517818)

Objective

This project is prepared to predict the price of house in ‘USA’ by building Predictive models using the concept of **Multiple Linear Regression**

Introduction & Motivation:

- A complicated and dynamic system, the housing market. The location, size, amenities, and condition of the property are just a few of the variables that might impact a home's price.
- The main driving force/the biggest motivation behind this project of house price forecasting and the development of a linear regression model is COVID-19.
- It is very difficult to search house in USA during and after Covid. Even if you find a house it is very difficult to get a perfect price for the same.
- To overcome such problem Regression Techniques can be used.
- Using Regression Techniques it will be easy to know the price of house based on the area available, number of bedrooms, facilities available.

Dataset Description

Dataset Size : 5000 x 7

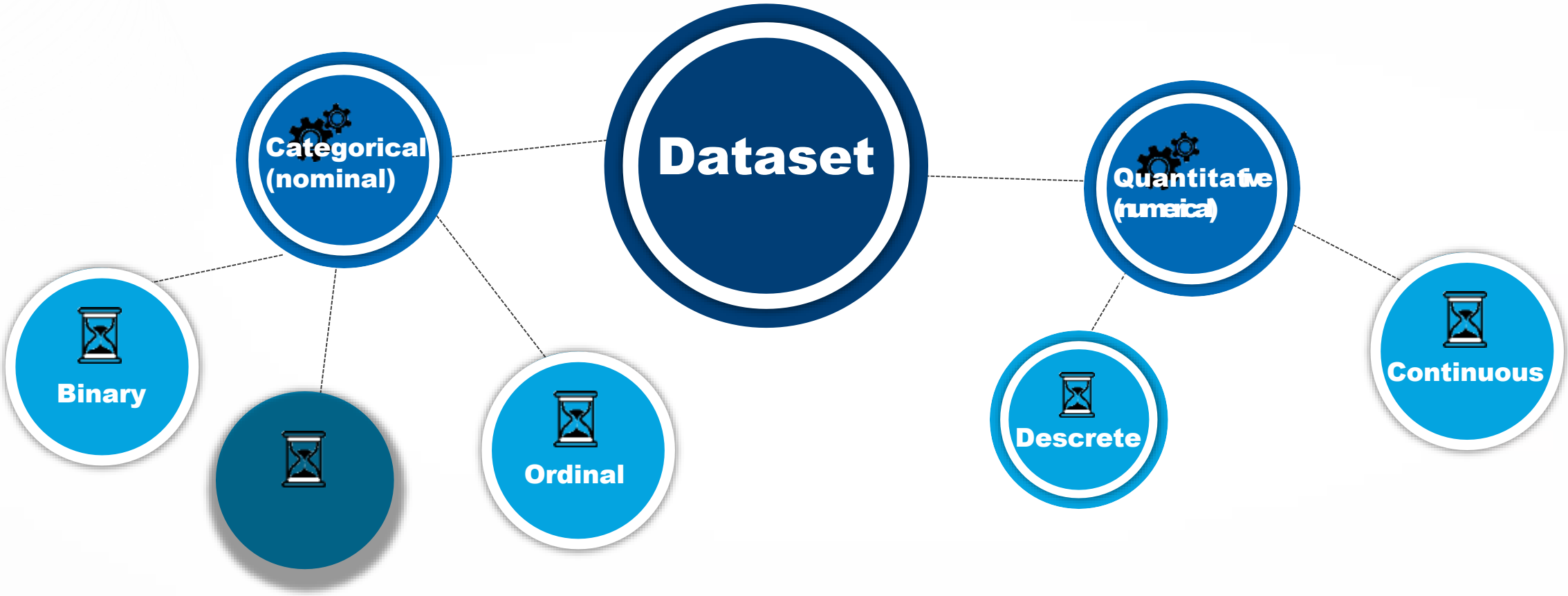
Following are the parameters/variables in dataset for house price prediction of 'USA' :

- **Price**
- **Avg Area Income**
- **Avg Area Housing age**
- **Avg Area Number of rooms**
- **Avg Area Number of bedrooms**
- **Area population**
- **Address**

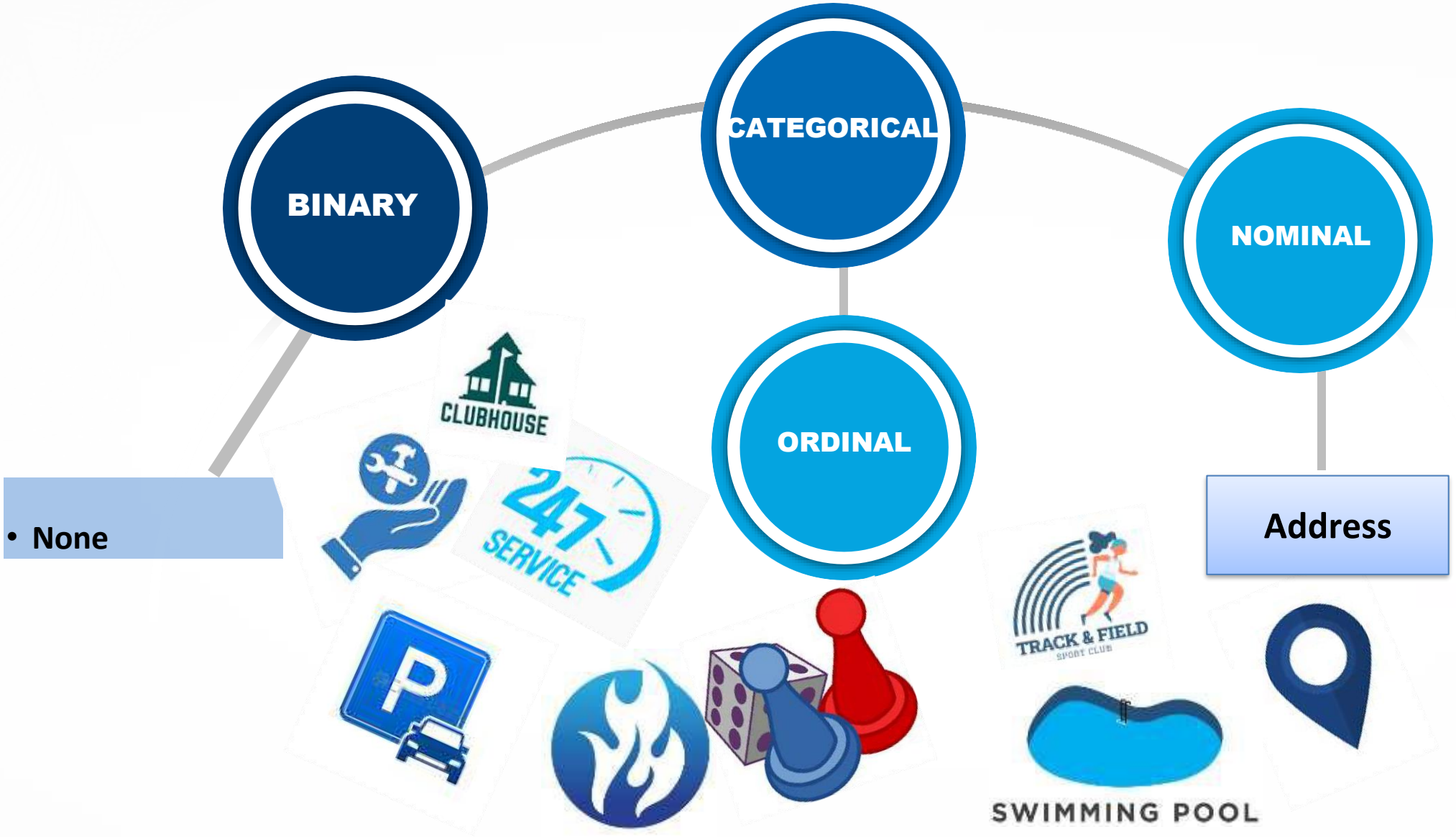
- **Source: Kaggle.com**
- **URL : https://www.kaggle.com/datasets/vikramamin/housing-linear-regression**

Dataset Type

Dataset is a collection of different variables of different data types.



Categorical Attributes



Quantitative Attributes

1.Price
2. Avg. Area Income
3.Avg Area Housing Age
4.Avg Area Number of rooms
5. Avg Area Number of bedrooms
6.Avg Population

Snapshot of dataset

The following snapshot shows head part of data taken from Dataset used in House price prediction.

Untitled1 | data

Filter

	Avg..Area.Income	Avg..Area.House.Age	Avg..Area.Number.of.Rooms	Avg..Area.Number.of.Bedrooms	Area.Population	Price	Address
1	79545.46	5.682861	7.009188	4.09	23086.801	1059033.6	208 Michael Ferry Apt. 674 Laurabury, NE 37010-5101
2	79248.64	6.002900	6.730821	3.09	40173.072	1505890.9	188 Johnson Views Suite 079 Lake Kathleen, CA 48958
3	61287.07	5.865890	8.512727	5.13	36882.159	1058988.0	9127 Elizabeth Stravenue Danieltown, WI 06482-3489
4	63345.24	7.188236	5.586729	3.26	34310.243	1260616.8	USS Barnett FPO AP 44820
5	59982.20	5.040555	7.839388	4.23	26354.109	630943.5	USNS Raymond FPO AE 09386
6	80175.75	4.988408	6.104512	4.04	26748.428	1068138.1	06039 Jennifer Islands Apt. 443 Tracyport, KS 16077
7	64698.46	6.025336	8.147760	3.41	60828.249	1502055.8	4759 Daniel Shoals Suite 442 Nguyenburgh, CO 20247
8	78394.34	6.989780	6.620478	2.42	36516.359	1573936.6	972 Joyce Viaduct Lake William, TN 17778-6483
9	59927.66	5.362126	6.393121	2.30	29387.396	798869.5	USS Gilbert FPO AA 20957
10	81885.93	4.423672	8.167688	6.10	40149.966	1545154.8	Unit 9446 Box 0958 DPO AE 97025
11	80527.47	8.093513	5.042747	4.10	47224.360	1707045.7	6368 John Motorway Suite 700 Janetbury, NM 26854
12	50593.70	4.496513	7.467627	4.49	34343.992	663732.4	911 Castillo Park Apt. 717 Davisborough, PW 78603
13	39033.81	7.671755	7.250029	3.10	39220.361	1042814.1	209 Natasha Stream Suite 961 Huffmanland, NE 52457
14	73163.66	6.919535	5.993188	2.27	32326.123	1291331.5	829 Welch Track Apt. 992 North John, AR 26532-5136
15	69391.38	5.344776	8.406418	4.37	35521.294	1402818.2	PSC 5330, Box 4420 APO AP 08302

Exploratory Data Analysis (EDA)

Checked for the missing values and found there are no missing values

R 4.3.0 · ~/

```
> na_count
                                 na_count
Avg..Area.Income                        0
Avg..Area.House.Age                     0
Avg..Area.Number.of.Rooms               0
Avg..Area.Number.of.Bedrooms            0
Area.Population                         0
Price                                   0
Address                                 0
```

```
> data = data[, !(names(data) %in% c("Address"))]
Error in exists(cacheKey, where = .rs.WorkingDataEnv, inherits = FALSE) :
  invalid first argument
Error in assign(cacheKey, frame, .rs.CachedDataEnv) :
  attempt to use zero-length variable name
> str(data)
'data.frame':   5000 obs. of  6 variables:
 $ Avg..Area.Income            : num  79545 79249 61287 63345 59982 ...
 $ Avg..Area.House.Age         : num  5.68 6 5.87 7.19 5.04 ...
 $ Avg..Area.Number.of.Rooms   : num  7.01 6.73 8.51 5.59 7.84 ...
 $ Avg..Area.Number.of.Bedrooms: num  4.09 3.09 5.13 3.26 4.23 4.04 3.41 2.42 2.3 6.1 ...
 $ Area.Population             : num  23087 40173 36882 34310 26354 ...
 $ Price                       : num  1059034 1505891 1058988 1260617 630943 ...
>
```

#column 'Address' has been removed, Now we have only 6 variables

Files | Plots | Packages | Help | Viewer | Presentation

Removed 1 non-numeric variable (address) which is not useful as It has got 5000 unique values in it.

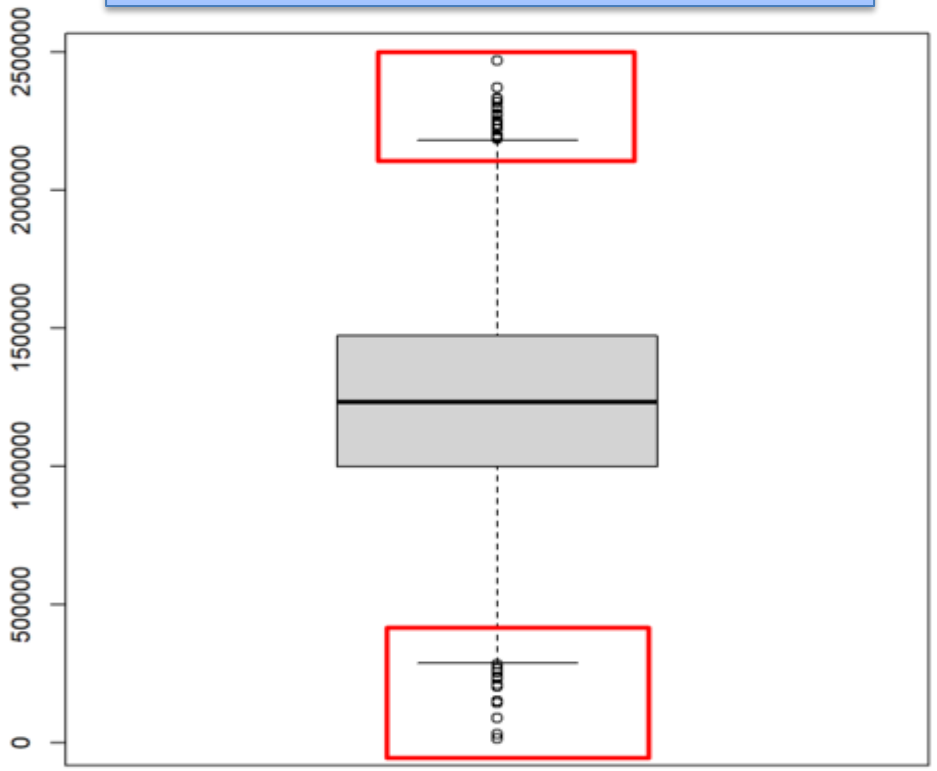
Data Analysis (Descriptive statistics)

Performed Descriptive statistics on the variables Price & Avg. Area Income

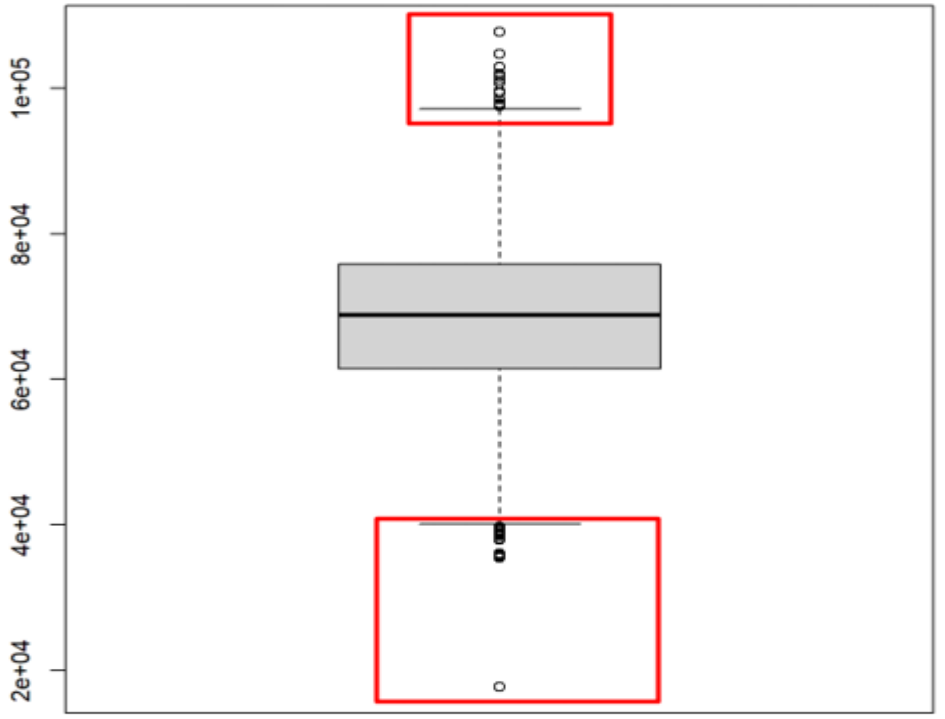
```
> describe(Price)
   vars    n    mean       sd  median trimmed      mad      min     max   range skew kurtosis
X1    1 5000 1232073 353117.6 1232669 1232160 350330.4 15938.66 2469066 2453127    0    -0.06
         se
X1 4993.84
> summary(Price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15939  997577 1232669 1232073 1471210 2469066
> var(Price)
[1] 124692058202
>
>
> describe(Avg..Area.Income)
   vars    n     mean       sd   median  trimmed      mad      min      max    range  skew
X1    1 5000 68583.11 10657.99 68804.29 68611.84 10598.27 17796.63 107701.8 89905.12 -0.03
   kurtosis     se
X1     0.04 150.73
> summary(Avg..Area.Income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17797   61481   68804   68583   75783  107702
> var(Avg..Area.Income)
[1] 113592777
>
```

Data Visualisation (Box-Whisker Plot)

Boxplot for the variable ‘Price’. Outliers Found

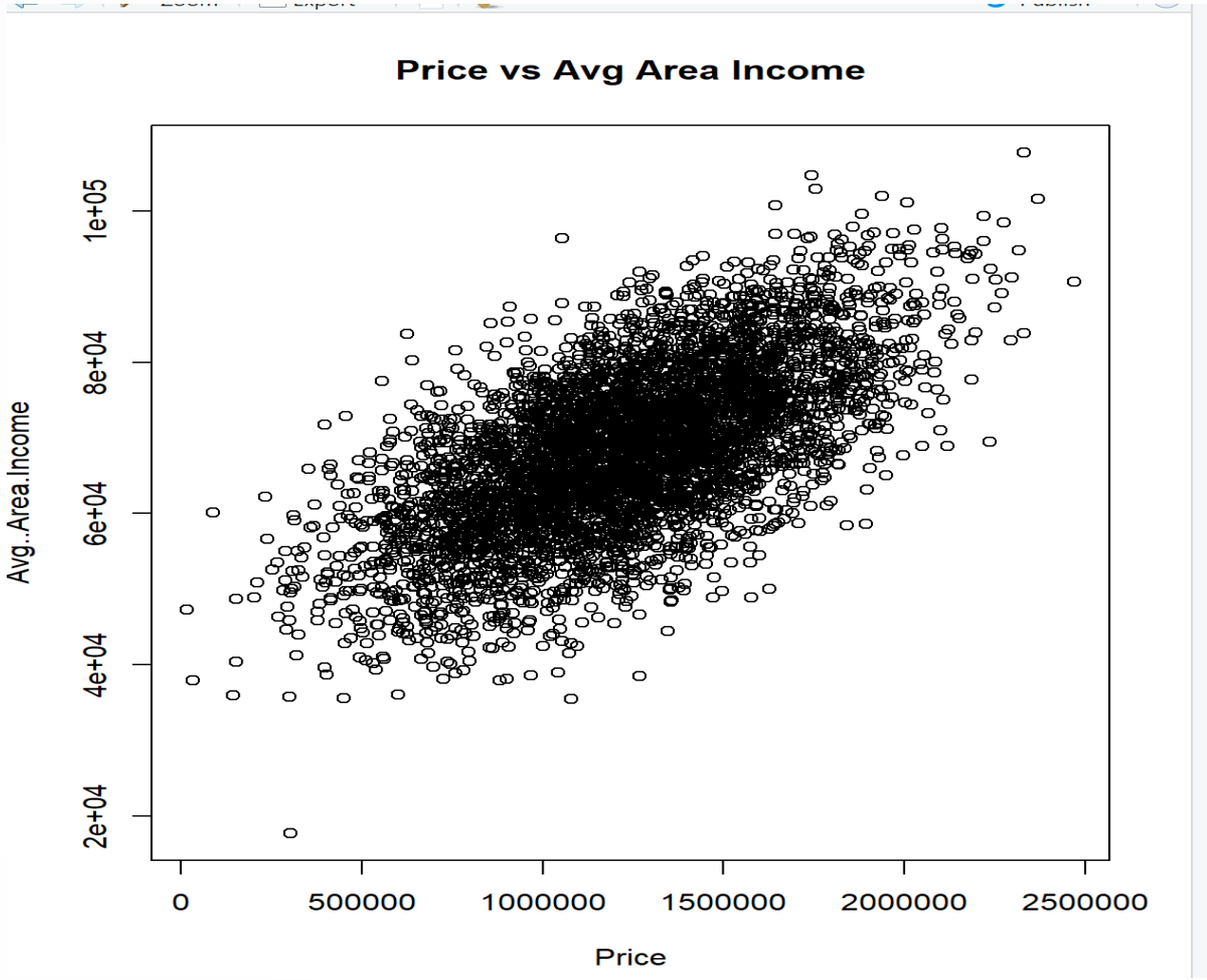


Boxplot for the variable ‘Avg Area Income’. Outliers Found



Note :But we focus more on Influential points in the regression analysis after building the model. So, outliers were not removed.

Data Visualisation (Scatterplot)



Observation : In the given Scatter plot it shows that as the ‘Price’ of house increases with increase in ‘Avg area Income

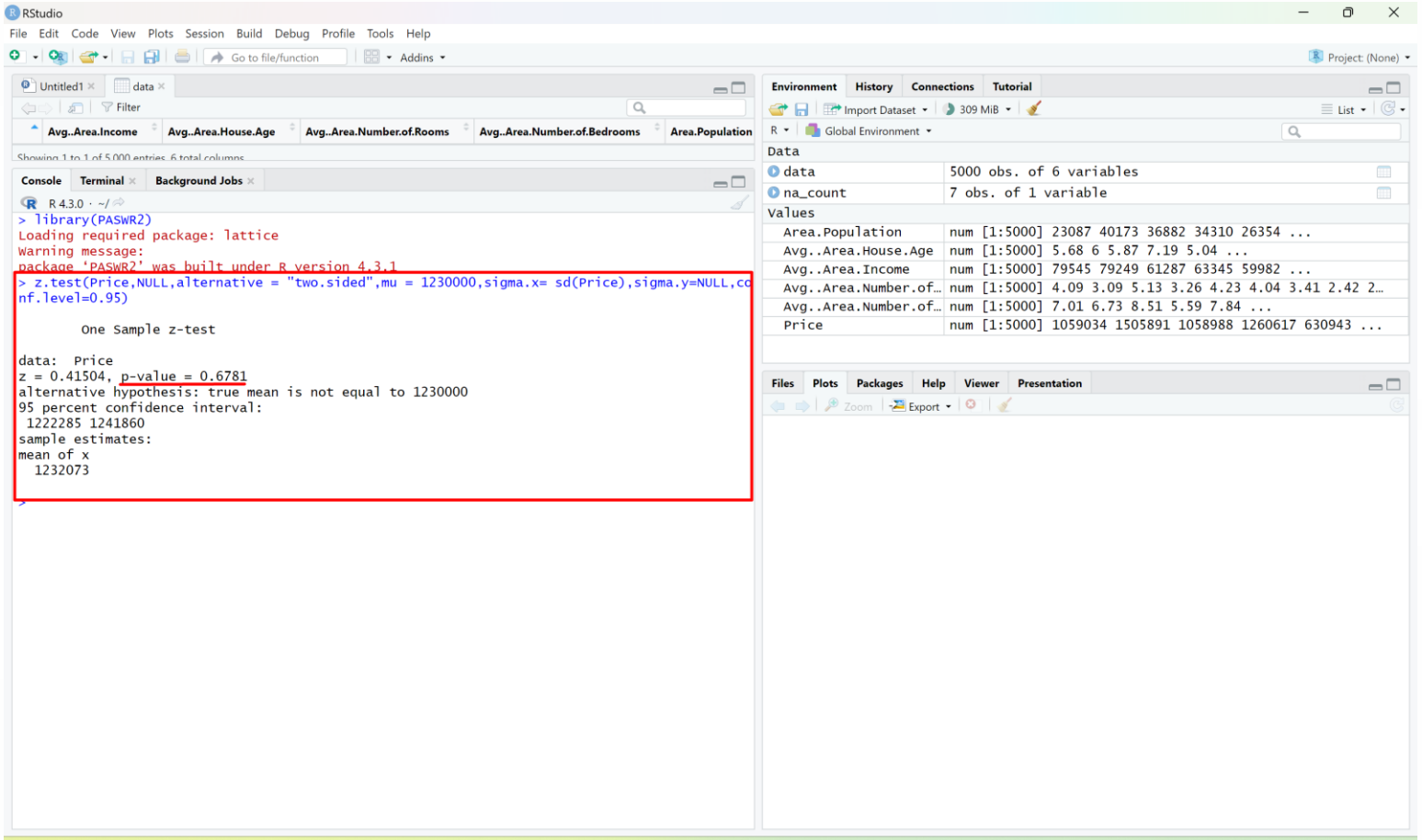
Hypothesis testing (Validating a claim~self-developed)

We were told that the average housing price will be greater than 1.23 million, but we were thinking it is not greater than 1.23 million (In other words, it could be less than or equal to 1.23 million). By using 95% as confidence level, validate the claim/Assumption.??)

Null hypothesis:
The average housing price is greater than 1.23 million.
$H0$: μ > 1.23 million

Alternative hypothesis: the average housing price is less than or equal to 1.23 million
Ha: μ <= 1.23 million

Conclusion: Based on 95% confidence level we fail to reject null hypothesis as p-value is larger than alpha(0.05)



Examined linearity b/w Price & X-variables(Correlation analysis)

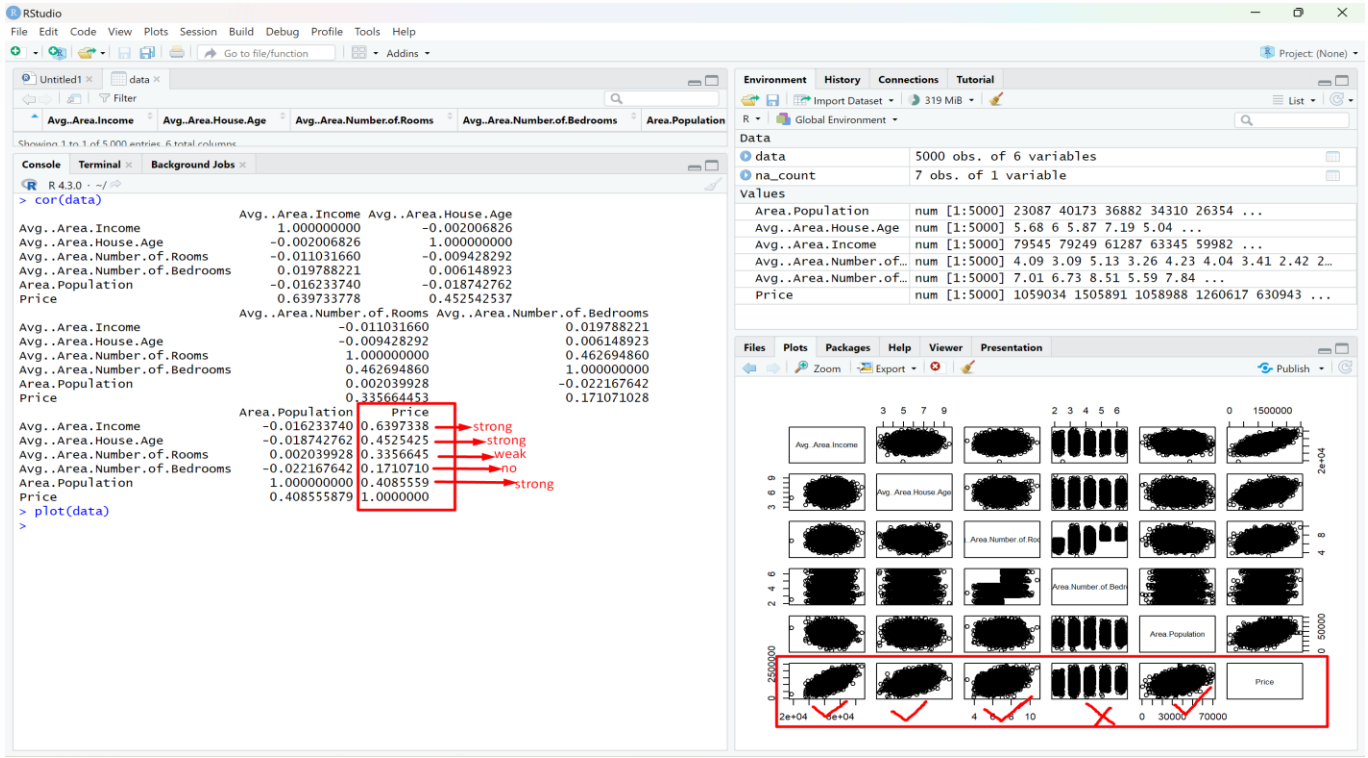
1.Price & Avg..Area.Income = 0.639 - strong Correlation - +ve correlation

2.Price & Avg..Area.House.Age = 0.452 - strong Correlation - +ve correlation

3.Price & Avg..Area.Number.of.Rooms = 0.335 - weak Correlation – +ve correlation

4.Price & Avg..Area.Number.of.Bedrooms = 0.171 - No Correlation – -ve correlation

5.Price & Area.Population = 0.408 – strong Correlation – +ve correlation



Applied Transformation on Avg Area no.of bedrooms

Square Transformation
Log Transformation
Square root Transformation
Power Transformation

```
 Avg..Area.Number.of.Bedrooms2 = Avg..Area.Number.of.Bedrooms * Avg..Area.Number.of.Bedrooms
 cor(Price, Avg..Area.Number.of.Bedrooms2)
1] 0.1678041
 logAvg..Area.Number.of.Bedrooms = log(Avg..Area.Number.of.Bedrooms)
 cor(Price, logAvg..Area.Number.of.Bedrooms)
1] 0.1704294
 Avg..Area.Number.of.BedroomsR = 1/Avg..Area.Number.of.Bedrooms
 cor(Price, Avg..Area.Number.of.Bedrooms)
1] 0.171071
 Avg..Area.Number.of.BedroomsSqrt = sqrt(Avg..Area.Number.of.Bedrooms)
 cor(Price, Avg..Area.Number.of.BedroomsSqrt)
1] 0.1712992
```

As the correlation didn't improve after transformation on Avg..Area.Number.of.Bedrooms variable, dropped that variable.

Now we have only 4 x-variables and 1 y-variable (price)

Data Split & Building predictive models by using linear regression

1. As our data is less than 1 million rows it is small data, N-fold cross evaluation is used for the data split. Taken N=5. So that we decompose the data into 5 folds and conduct 5 rounds of evaluation and calculate the average
2. we set the independent variables as X = Avg Area Income, Avg Area House Age , Avg Area No.of rooms, Avg area population. and target variables as Y = price.
3. Fitting the train set to multiple linear regression and getting the score of the model.
4. Fitting the train set to full linear regression model, Lasso regression model, feature selection models and getting the score of the model.
5. Calculate the model score to understand how our model performed along with the explained variance score.

Building MLR & Findings:

MODEL	RMSE
M1 : Full model using all x-variables By using Train function & lm method	101251.2
M2: Base model using only 1 x-variable By using Train function & lm method	271358.3
M3: Based on backward method. Used 4 x-variables & train function & LeapBAckward method(3 x- var got used)	101243.6
M4: Based on Forward method. Used 4 x-variables train function & LeapForward method (3 x- var got used)	**101165.9** **Lesser than other models**
M5: Based on stepwise method. Used 4 x-variables train function & LeapSeq method (3 x- var got used)	101201.0
Lasso regression model with all 4 x-variables by using train function & lasso method	106454.4

Avg..Area.Income
Avg..Area.House.Age
Area.Population

Selected Model: M4 as it has less RMSE, to improve the model by seeking different methods like (VIF & removing influential points) & to perform model diagnosis for the model M4.

Model diagnosis & improvements for M4

F-test

Null hypothesis : (H0): The coefficients of all x-variables are zero and there is no linear relationship with Price

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

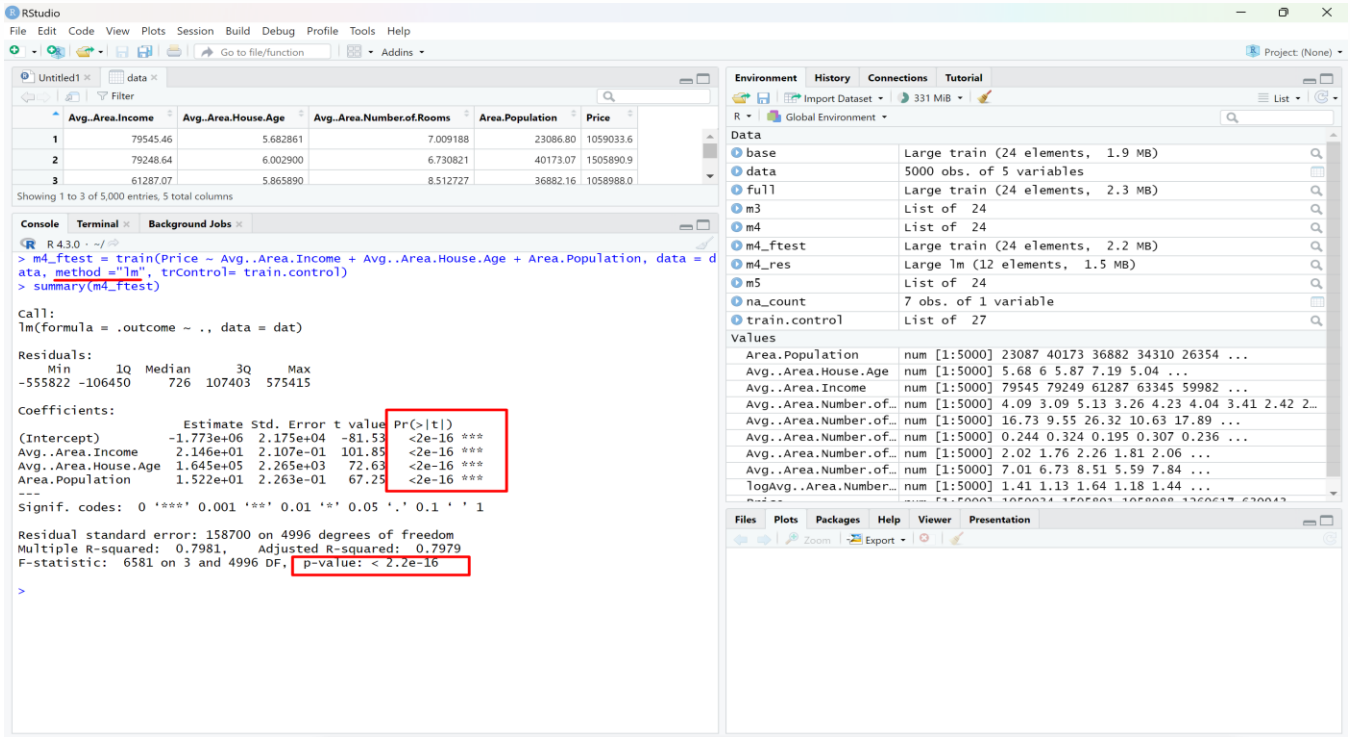
Alternative Hypothesis: (Ha): At least one of the coefficients of the x-variables (Avg..Area.Income, Avg..Area.House.Age, Area.Population) is not zero and can affect Price.

$$H_a : \beta_j \neq 0$$

β1 = coefficient of " Avg..Area.Income,"

β2 = coefficient of " Avg..Area.House.Age,"

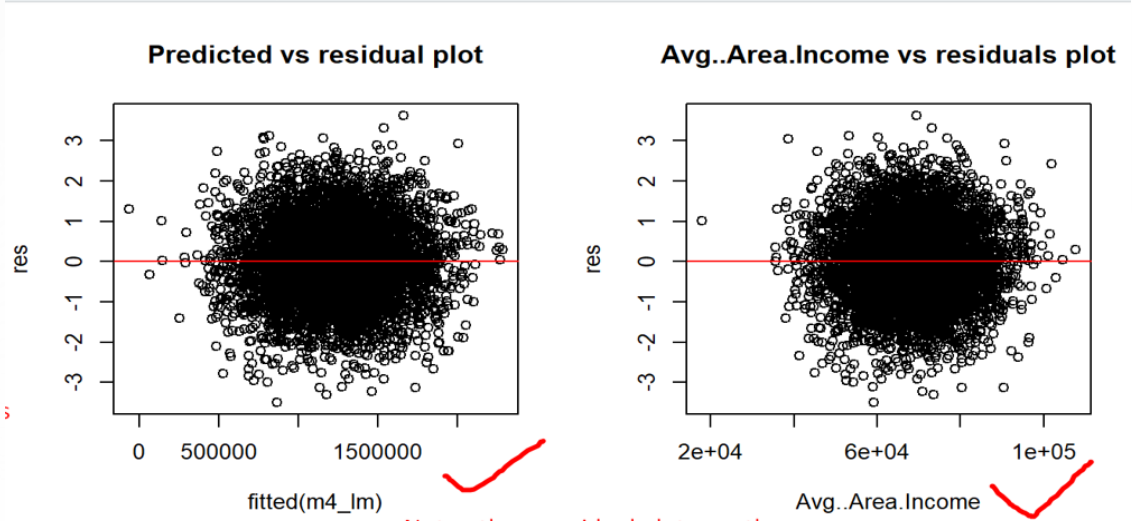
β3 = coefficient of " Area.Population."



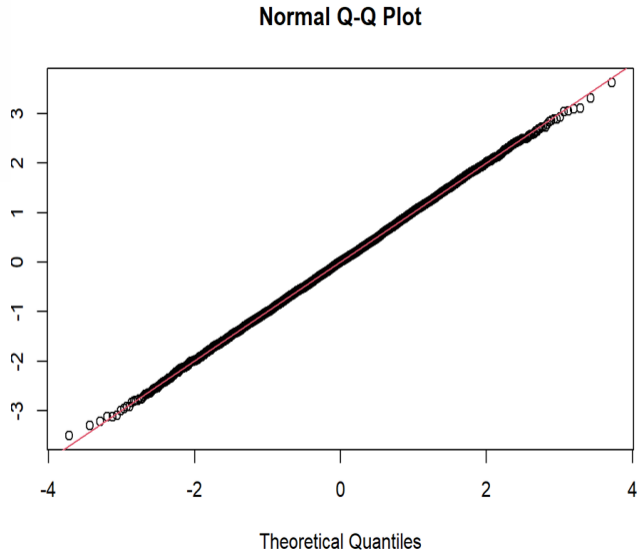
Conclusion on F-test: p-value < 0.05 for model m4. At 95% confidence level, we can say that at least 1 x variables among (Avg..Area.Income , Avg..Area.House.Age , Area.Population) has a significant linear relationship with Price and can affect the value of y-variable ~ Price

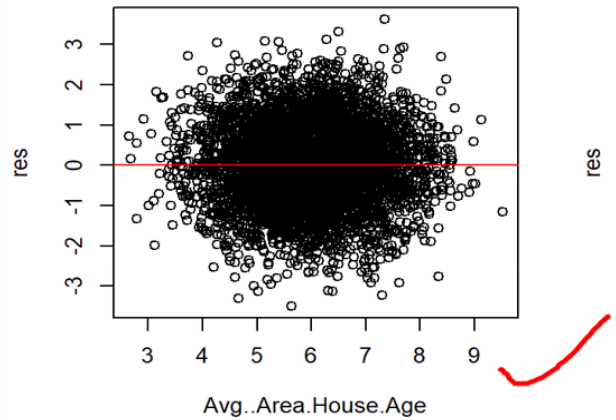
Model diagnosis & improvements for M4, cond..,

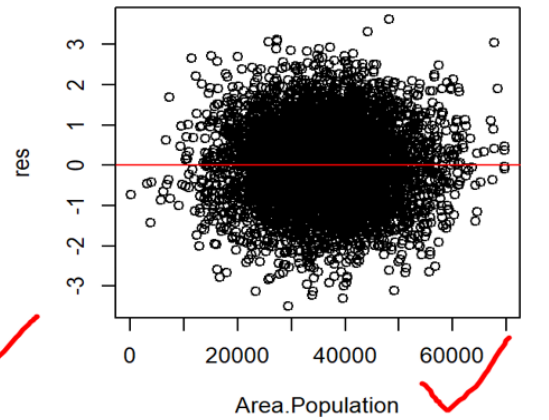
Residual Analysis



Note : these residual plots are the same plots we got initially.







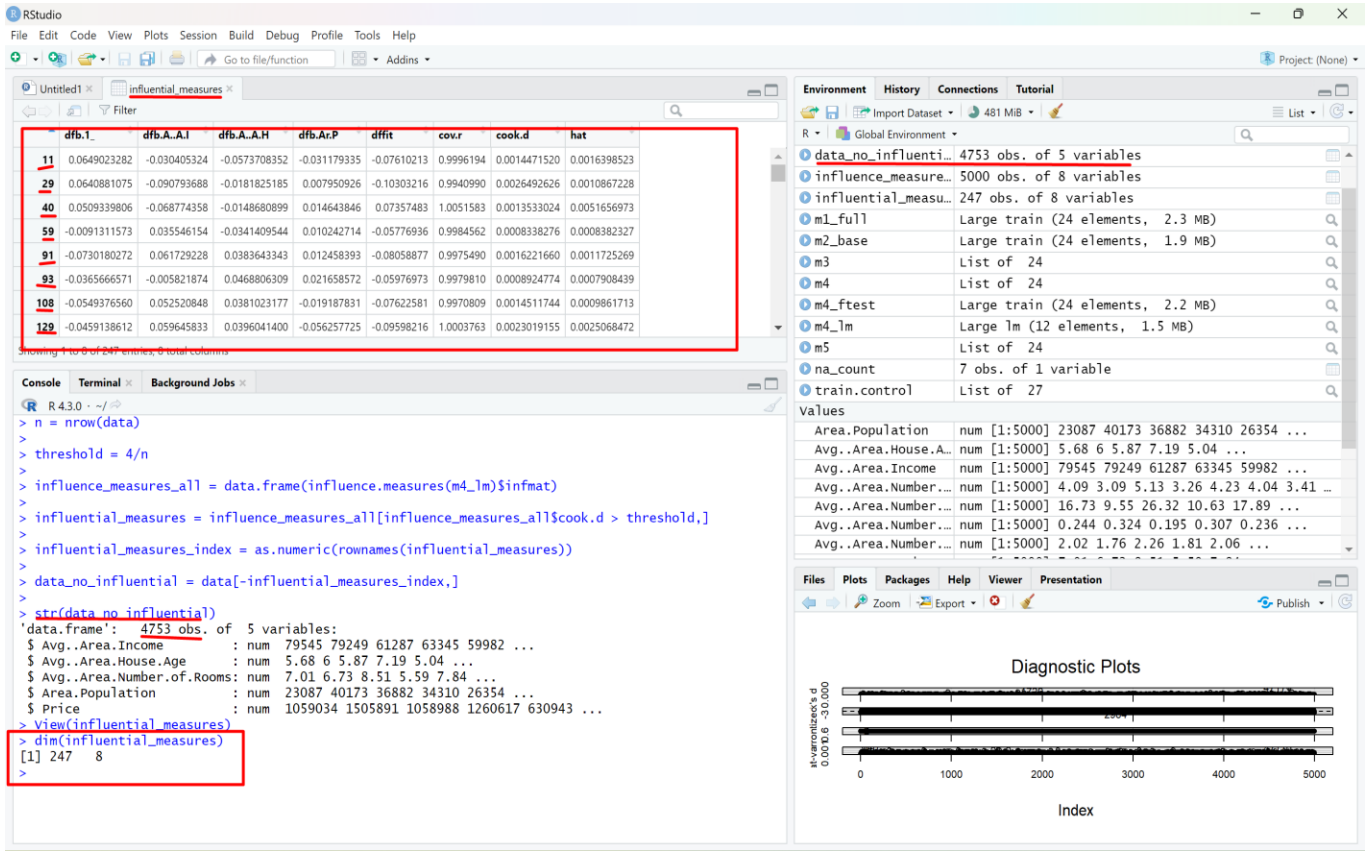
Conclusion on residual analysis: I felt residual constant variance is ok ok. Tried transformation on y-var(Price) but no improvement. Linearity exists b/w y and each x-variable. Residual follows normality. Hence, m4 passed model diagnosis.

Improvements for M4, cond.,

Checked multicollinearity by VIF & Influential points

```
> vif(m4_res)
  Avg..Area.Income Avg..Area.House.Age      Area.Population
          1.000269             1.000357             1.000616
```

No high collinearity between x-variables



Found 247 influential points & removed them.
So we need to build a new model with no influential points with the same x-variables.

Models We got:

MODEL	Function used	Method used	Variables USED	RMSE	Model dignosis	Improvements
M1 (full model)	Train function	lm method	All 4 X-varibles	101251.2		
M2 (base model)	Train function	lm method	1 x-variable	271358.3		
M3	Train function	LeapBackward	3 X-variables	101243.6		
M4	**Train function**	**LeapFoward**	**3 X-variables**	**101165.9 Lesser than other models**	**So selected this model. & moved performed model dignosisi**	
M4	**Lm function**	**lm method**	**3 X-variables**		**Performed f-test for m4**	
M4	**Lm function**	**lm method**	**3 X-variables**		**Performed res analysis for m4. So used lm function rather than train function. Because res cant be calculated for the model built with train function**	**~Calculated Vif for m4(no vif) ~identified influential points & removed by cooks.d &built final model m6 with new data(data_no_influential points**
M5	Train function	LeapSeq	All 4 X-variables	101201.0		
Lasso regression model	Train function	Lasso metho	All 4 X-variables	106454.4		
*********	*******	*********	Removed infl.points	*******	**********	*******
Final_model_M6 (built from improving M4)	Lm function	lm method	3 X-variables	142848.8	M4 Model built after removing influential points.	Built on new data with no influential points

Findings, Explanations & Conclusion:

1. After removing influential points, I re-built the same model m4 with the new name “final model m6” with the new data “data_no_influential” by using the same x-variables that got used in the model m4 but not with train method. I built the final model m6 with lm method.

2. But I found that RMSE has increased after removing influential points in the final model m6 compared to previous best model m4

3. So, I felt that those influential points were not truly influential in a negative way, or that their presence was actually helping to improve the predictive performance of my model

4. I would like to pick model M4 which has least RMSE (with influential points) and interpret the model:

```
coef(m4$finalModel, 3)
      (Intercept)    Avg..Area.Income Avg..Area.House.Age     Area.Population
    -1.772988e+06        2.145624e+01        1.644960e+05        1.521688e+01
```

Interpreting the best model: M4:

$$Y = \beta 0 + \beta 1 * X1 + \beta 2 * X2 + \beta 3 * X3 + e$$

$$Price = (-1.773e{+}06) + 2.146e{+}01 * X1 + 1.645e{+}05 * X2 + 1.522e{+}01 * X3 + e$$

X1 = Avg..Area.Income
X2 = Avg..Area.House.Age
X3 = Area.Population

The intercept term is -1.773e+06. It represents the value of the target variable (Price) when all predictor variables (X1, X2, and X3) are zero. However, it's important to note that in this context, having all predictor variables equal to zero might not be meaningful, especially for features like income, age, and population. The intercept is essentially the y-intercept of the regression line.

The coefficient of X1 (Avg..Area.Income) is 2.146e+01. This means that, for each one-unit increase in Avg..Area.Income, the Price is expected to increase by 2.146e+01 (21.46 units), holding all other predictors constant.

The coefficient of X2 (Avg..Area.House.Age) is 1.645e+05. This indicates that, for each one-unit increase in Avg..Area.House.Age, the Price is expected to increase by 1.645e+05 (164,500 units), assuming all other predictors remain constant

The coefficient of X3 (Area.Population) is 1.522e+01. This means that, for each one-unit increase in Area.Population, the Price is expected to increase by 1.522e+01 (15.22 units), holding all other predictors constant.

Limitations & Future Work:

Limitations:
One finding indicates that the model M4 constructed with feature selection and the leap forward approach yielded superior results (least RMSE value), however there were influential points. To determine if the pattern is present similarly in other data or not, this must be checked with other data set. It was impossible for us to double-check this with another dataset because we could only work with one at a time.

Future Work:
1. Collect new variables like “year built” , whether related information like Temperature, rainfall, snow,floods ,tornado(for ex: no.of tornado’s per year, no.of floods per year etc., & school district rating (from 1-10)
2. Improve correlation of x-variable with y-variable by applying transformations.
5. Handling Influential points.
6. Trying different advance regression techniques like polynomial regression, ridge regression, or knn regression, special neural network model, Regression Tree.

Thank you