Naga Satya Silpa: Building Housing Price Predictive Models Using Regression Techniques

For example: Group 5: Build a recommender system using collaborative filtering

Your submissions:

- Group number_Report.pdf
- Group number_ R Codes.txt (with necessary comments in the codings)
- Group number_ R Outputs.pdf (the snapshots of key outputs)
- Group number_Slide.pdf (your presentation slide)
- Group number_video.mp4 (your recorded presentation video)

First Name	Last Name	Email (hawk.iit.edu)	Student ID
Naga Satya Silpa	Annadevara	nannadevara@hawk.iit.edu	A20517818

Table of Contents

1. Introduction

Introduce the background and motivations.

Introduction:

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, as well as the state of the economy. Regression techniques has been utilized to anticipate home price trends more and more recently. The linear regression method is one of the most widely used machine learning techniques for this.

Background:

-A simple but successful statistical model called linear regression may be used to forecast a continuous variable, like the cost of a house, from a number of independent elements. The model suggests that the relationship between the dependent variable and the independent variables is linear, which indicates that a change in one variable is associated with a proportionate change in the other variable.

-In order to create a linear regression model, we must first gather data on home prices and other property characteristics. A statistical software program will be used to fit the model to the data after we obtain them. The model's coefficients will be output by the program and used to predict the dependent variable from the independent variables.

- After fitting the model, we can use it to forecast the price of a property based on its attributes. For instance, we may use the model to forecast a house's price if we know its location, size, and condition.

Motivations:

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. The epidemic has had a profound effect on many economic factors, and the real estate market is no exception. Before the COVID-19 outbreak, we were residents of Chicago's downtown. Situations altered after the outbreak began. Our ability to pay so much rent decreased because of a rise in apartment leasing rates. The demand for additional room in the house for office purposes ultimately developed because of firms implementing work-from-home programs, and apartments in cities often do not have this kind of huge area. After taking all of this into consideration, we made the decision to relocate from Chicago, Illinois to Aurora, Illinois. It is a suburb in the state of Illinois where we may purchase or rent larger homes for less money. When we made the decision to leave the city for the suburbs and purchase a home during the epidemic, we discovered that the property market is quite unstable and unpredictable, which caused us a lot of anxiety. To estimate the price of homes, we utilized applications like Redfin and Zillow. So I was really inspired by all the experiences I went through during epidemic that led me to embark on this project this semester.

The requirement for housing price forecasts and the application of linear regression models are motivated by COVID-19 in the following ways:

1.Market Uncertainty: Due to shifts in consumer tastes, the rise of remote labor, and general economic uncertainty, the housing market fluctuated during the epidemic. For prospective buyers, sellers, and investors seeking consistency in their judgments, predicting house prices becomes essential.

2.Demand Shift: Demand Patterns Changed Due to COVID-19. People looked for homes with more space, home offices, and access to outdoor facilities. Robust modeling strategies like linear regression are needed to predict how these changes would impact house prices.

3.Economic Recovery: Predicting home prices can be useful in assessing the situation of the real estate market and its contribution to the overall economic recovery as nations recover from the financial effects of the epidemic.

4.Government Policies: To boost the housing market during and after the epidemic, governments may employ several stimuli plans or policies. Predictive models can be used to evaluate the efficacy of such treatments.

5.Risk evaluation: For financial institutions, forecasting house prices aids in evaluating the risk involved in mortgage lending in an uncertain economic environment.

6. Disruptions to Supply chain networks: The pandemic had an impact on supply networks for construction, which might have an influence on real estate pricing. The effect of these interruptions on housing expenses can be examined using a regression model.

7. Migration Patterns: COVID-19 had an impact on migratory patterns as individuals migrated from urban regions to suburban or rural locations. Understanding how these changes will impact housing costs can help with infrastructure planning and urban planning.

8. Real estate industry adaptation: To respond to shifting market conditions, real estate professionals and businesses must modify their strategy. Predictive models can help you spot opportunities and danger zones.

Project Scope and Objectives:

The primary purpose/objective/aim of the project is to:

1. Develop an efficient pricing prediction model.
2. Verify the model's accuracy.
3. Identify the crucial house price factors that contribute to the model's accuracy.
4. Predict the unseen data (such as house price given the features like location, size, and number of rooms).

2. Data

Introduce your data, such as where did you get it (provide the URL if possible), how large it is, what are the variables/features, what are the variable types, etc.

About the data:

Title: "Housing Price Prediction Dataset"

Source: Our dataset for this project is taken from **Kaggle.com**.

The URL: https://www.kaggle.com/datasets/vikramamin/housing-linear-regression

Data description & size: This dataset shows housing information for 5000 homes in all over the USA. The dataset contains 5000 rows, where each row represents a different property in USA. There are 7 columns, including one target variable ("Price") and six features describing the characteristics of each property.

Data Usage: This dataset may be used for a variety of tasks, such as running a linear regression to forecast house prices based on the given attributes. It may also be used for data exploration, visualization, and other kinds of linear regression models to learn more about the US housing market. It may be used to create different types of linear regression models and analyze the results.

The variables/features are as follows:

- Avg. Area Income: Average income of the Area
- Avg. Area House Age: Average Age of the House in the Area
- Avg. Area Number of Rooms: Average Number of Rooms in a House in the Area
- Avg. Area Number of Bedrooms: Average Number of Bedrooms in the House in the Area
- Area Population: Average Population in the Area
- Price: Average Price of the House in the Area
- Address: Address of the Area

Variable data types:

Both numerical and category variables are present in the dataset. The continuous numerical variables that reflect the homes and the neighborhood's varied characteristics include population, income, age, and the number of rooms in each house. Text information about each property's location in the USA is contained in the category variable "Address."

Variable/columns/features	Data type
Avg. Area Income	Numerical / continuous data~ decimal
Avg. Area House Age	Numerical / continuous data~ decimal

Avg. Area Number of Rooms	Numerical / continuous data~ decimal
Avg. Area Number of Bedrooms	Numerical / continuous data~ decimal
Area Population	Numerical / continuous data~ decimal
Price	Numerical / continuous data~ decimal
Address	Qualitative / nominal ~ string

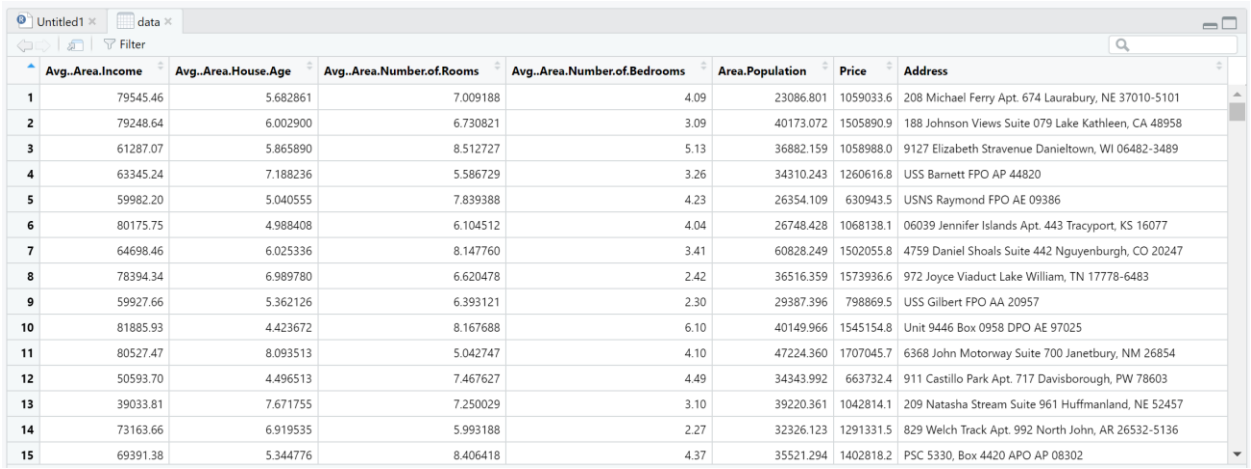
Independent variables/predictors/x-variables:

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

Dependent variable/response variable/target variable/y-variable:

- Price

Here is the sneak peek of the dataset:



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

3. Problems to be Solved

List the problems you want to solve. Note: they are problems, not solutions. For example, hypothesis testing, ANOVA, classification, cluttering, etc., are all solutions.

While performing linear regression analysis, we may meet several issues that must be addressed and resolved. Some of these issues can be resolved during the data pre-processing step, while others can be solved during the post-processing phase, and yet others may be solved while or after the linear regression model has been built. When doing linear regression analysis on the provided dataset with predictor variables (Avg. Area House Age, Avg. Area Income, Avg. Area Number of Rooms, Avg. Area Number of Bedrooms, Area Population, Address) and the response variable (House Price), The challenges/ problems that may arise and that I would like to overcome are:

Data pre-processing problems: Problems need to be resolved Before building the model:

1. Missing values:

If the dataset contains missing values, they must be treated carefully since linear regression does not operate with incomplete data. Missing values in rows or columns must be addressed using one of the options offered before generating the model.

2. Non-Numeric Data:

Not all feature/variables may have a substantial impact on predicting the response variable. If the dataset contains categorical or non-numeric variables, they must be encoded or converted into numeric formats before being employed in the linear regression model. if the columns that don't offer any useful data for creating the model may be deleted. Our data has one categorical variable~ 'Address' and we need to handle this variable either by converting into numerical variable or by deleting the variable.

3. Data Analysis (Descriptive statistics):

To obtain insights, identify trends, and come to meaningful conclusions, data analysis for a dataset entails a thorough investigation of the data. Although calculating mean, median, mode, and standard deviation as well as drawing box plots and histograms are important parts of descriptive statistics and data visualization, the process of data analysis involves much more.

4. Validate claim/Assumption:

A claim/Assumption should be developed by our own and we should validate that assumption is true or not based on different techniques.

The claim/Assumption: 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.??

4. non-linearity:

Nonlinearity is a statistical word that refers to a scenario in which there is no straight-line or pattern observed between an independent variable and a dependent variable.
The relationship between variables might, however, be more complicated and non-linear in real-world situations. It is possible for the data points to exhibit a curved pattern, exponential growth, logarithmic behavior, or some other non-linear trend. It is difficult to effectively capture the underlying patterns when fitting a linear regression model to such data, which results in subpar prediction performance.

5. data split:

Before running a linear regression analysis, the data must be divided. It is based on the size of the data. Data must be separated using a single technology if it is tiny and has less than 1 million rows. A different kind of technology must be employed if the data is huge, defined as having more than 1 million rows. Therefore, it's crucial to consider the data's size while choosing the appropriate solution.

6. Building MLR:

Once data is split according to the size of the data, we need to build different multiple linear regression models to compare them based on the selected metrics. Building models is a crucial step in regression analysis. Hence, this step needs to be performed with extra care.

Data post-processing problems: Problems need to be resolved After building the model:

8. Model diagnosis:

After building the model, it needs to be tested to see if there are any issues with the model, we built that could impact the model's functionality and the accuracy of its predictions. To make sure that the linear regression model is appropriate for the provided data and that the results can be relied upon for generating predictions and inferences, model diagnosis is an essential step. There are different solutions to solve this problem which we will discuss in the solutions section.

6. Multicollinearity:

Multi collinearity refers to the features or x-variables that you will use to design your linear regression model having a high degree of correlation. Because of this, separating out the unique influences of each predictor on the response variable can be difficult. Therefore, it's crucial to spot the multi-collinearity issue in the x-variable prior to building the model during the pre-processing phase and resolve it.

9. Influential points: Influential points are the outliers that heavily influence the estimated regression coefficients. They have a high leverage on the model parameters. If removed, the parameter estimates change. You may have many outliers but not all of them are influential points. Excessive values
of the predictor variables are frequently linked to influential points. So, it is important to identify the influential points and remove them. Influential points cannot be identified in advance. You need to build a model first, then identify influential points with respect to the model you built. There are different metrics on how to solve this problem which we will discuss in the next section.

4. Solutions

List the solutions for the problems in part 3. Make sure your solutions can solve the problems in part 3 one by one. If you are going to build predictive models, clearly indicate the dependent and independent variables. In this section, you just provide your plan as solutions. You do not need to give technical details. You will give the details of the solutions and outputs in Section 5.

For the linear regression analysis on our dataset to be reliable and valid, I believe that all the above-mentioned data pre-processing and post-processing issues must be resolved. When these problems are handled properly, decision-making and inference can benefit from more precise and significant insights. The solutions for the above problems are as follows:

-We are predicting the target variable 'Price' which is numerical variable and building predictive model on numerical variable. So, this is called Regression model.

Independent variables/predictors/x-variables:

- X1 = Avg. Area Income
- X2 = Avg. Area House Age
- X3 = Avg. Area Number of Rooms
- X4 = Avg. Area Number of Bedrooms
- X5 = Area Population
- X6 = Address

Dependent variable/response variable/target variable/y-variable:

- Y = Price

Data pre-processing problems: Problems need to be resolved Before building the model:

1. Handling Missing values:

Solution: fill in by mean value

Missing values in rows and columns can be removed if more than 80% of values are missing. Otherwise, we should try to fill in missing values by the following ways:

For numerical variables: we need to fill in them by using mean or median value.

For nominal variable: by using the value with largest relative frequency.

2. Handling Non-numeric data (removing / converting irrelevant variables):

Our data collection contains non-numerical data. A categorical/nominal variable called "Address" has no numerical value. It should be ignored and not used in the construction of our model because it has 5000 unique values. It is necessary to eliminate the variable "address" from the data set as a result.

3.Performing data analysis (descriptive statistics)

- Next I am interested in knowing more about the column's 'Price & Income, so I would like to apply descriptive statistics initially by calculating mean, median, variance, Q1, Q2, Q3 and visualize these columns by histogram or boxplot on these 2 variables.

4. Validate the claim/Assumption:

Solution: Hypothesis testing

I would like to use hypothesis testing to validate the claim/assumption I developed above:
(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.??)

Initially I would like to state Null and alternative hypothesis:

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

Based on Ha we can observe that it is a two-tail test. As the sample size is larger than 30, it is z-test. We draw the conclusion based on the p-value in the z-test.

- If p-value is less than alpha (95%confidence level) = reject null hypothesis
- If p-value is larger than alpha = do not reject null hypothesis (accept null hypoth)

5. Handling non-linearity: Examine linear relation b/w X & y variables
Preferred Solution: check correlation (not always clear in scatterplot)
Nonlinearity b/w 2 variables can be observed by
a. scatterplot
b. Correlation analysis by Pearson correlation
If there is no straight-line pattern observed in the scatterplot between y and x variables it indicates that there is non-linear relationship b/w the variables.
If the correlation values b/w y and z variables are in b/w -0.1 & +0.1 indicates no linearity
If the correlation value is in b/w 0.1 and 0.4- it indicates week correlation/ weak linearity
If weak or no linearity/correlation, don't drop the variables, don't drop the variables.
Solution:
a. Applying transformation on x-variables:

- log transformation.
- Square root transformation
- Reverse transformation

b. Adding higher order terms in the regression model to straighten up the line. In other words, it is basically power transformation on x-variables. It is basically we increase the power of x-variables to 2 or 3. If we build regression models by using higher order terms, then it is called polynomial regression model.

6. Handling data split:

As our data is less than 1 million rows it is small data. So, we need to **use N-fold cross evaluation.**

I would like to take N=5.

So that we decompose the data into 5 folds and conduct 5 rounds of evaluation and calculate the average by using R. The screenshots and coding will be shown in the next step.

7. Building 5 different models:

Initially I would like to build:

1.full model by using train function & lm method & with all x-variables **~M1**

2.base model by using train function & lm method & with single x-variable **~M2**

Then I would like to use feature selection process in n-fold cross evaluation and build models by using:

3.backward elimination by using train function & LeapBackward method **~M3**

4.Forward method by using train function & LeapForward method **~M4**

5.stepwise method using using train function & LeapSeq method **~M5**

Lasso model by using train function & lasso method

After the best model selected among the 5 models above, I would like to perform model diagnosis and calculate VIF and identify and remove influential points and build the final model

6. Final model using lm function **~M6**

Data post-processing problems: Problems need to be resolved After building the model:

8. Model diagnosis:

Solution:

a. Goodness of fit test (F-test):

F-test is a statistical test for the purpose of hypothesis testing. While performing F-test on every model, we need to state null and alternative hypothesis according to that model. In this F-test, we do want to reject null hypothesis. For that we need to look at the p-value in the output we produce.

If the p-value is less than Alpha (assuming using 95% confidence), We reject null hypothesis and one of our x-variables can have significant linear relation with our y-variable (price) and can affect the value of target variable. To know which variables are effective, we need to look at the individual t-test for the p-value. The variables whose p-value is lesser than alpha re affective and the remaining x-variables need to be eliminated.

If the p-value is larger than Alpha (assuming using 95% confidence), we fail to reject null hypothesis and no x-variables can have significant linear relation with our y-variable (price) and can affect the value of target variable.

If our model failed F-test:

Solutions:

- Try transformations on X-variable or y-variable.

b. Residual analysis:

In the residual analysis I would like to check the residual for the following assumptions:
a. Residual must have constant variance. (if not constant, try transformation on y-variable)
b. Residual must have linear relationship with each of our x-variable. (If no linearity, try transformation on x-variables)
c. Residual should follow normal distribution. (No proper solution as our data size is small.)

7. Handling multicollinearity among x-variables:

Solution: Calculation VIF (variation Inflation factor)

Multicollinearity can be identified by the scatterplot or by the correlation analysis. If the correlation is strong b/w x-variables no need to use all of them. The suggested way to solve this problem is to calculate VIF (variation inflation factor). VIF can be calculated only in the post-processing phase.

So, I would like to go with VIF and check which variables have VIF value larger than 4. Those pairs of variables need to be removed.

9.Handling influential points:

Solution: Cooks distance ()

Influential points are different from outliers. Though we remove outliers by IQR method, there might still be chances of having influential points and need to check and removed after the model was built. I would like to use Cook distance as a metric to eliminate.

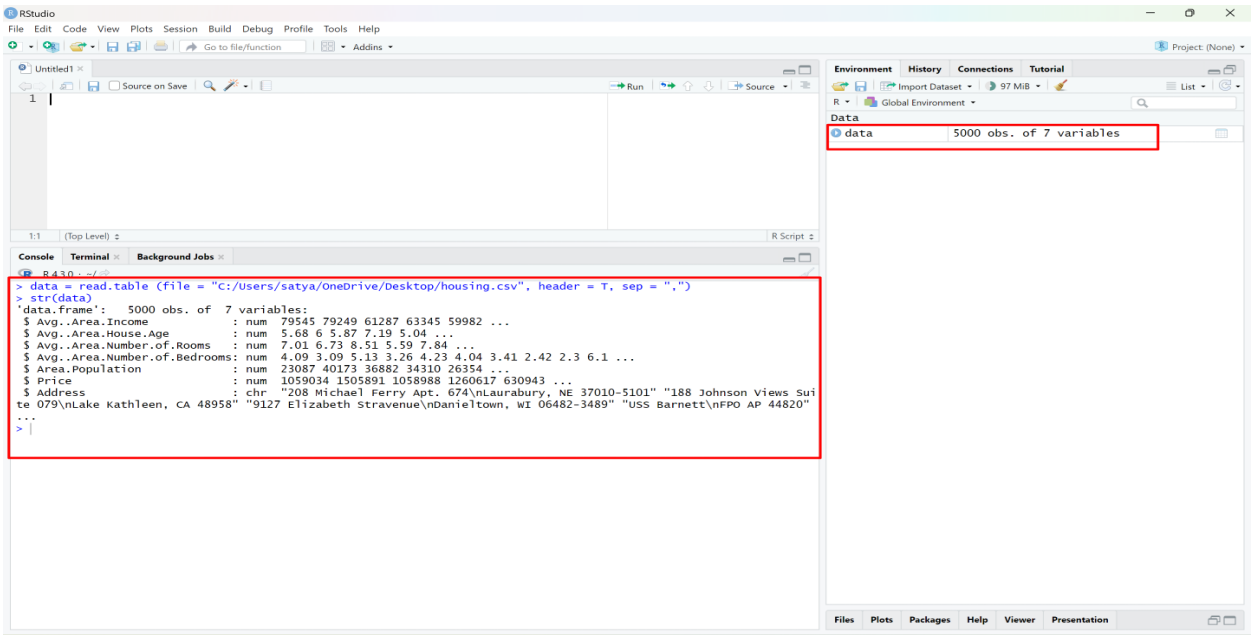
5. Experiments and Results

5.1. Methods and Process

Solve the problems your proposed one by one. Give the necessary codes, snapshots, and explanations.

Initially I loaded my dataset into R-studio using the code:

```
#Load the housing.csv data into R:
data = read.table (file = "C:/Users/satya/OneDrive/Desktop/housing.csv", header = T, sep = ",")
#To get a brief idea on the loaded data:
str(data)
```



1. Then I checked for missing values using the code:

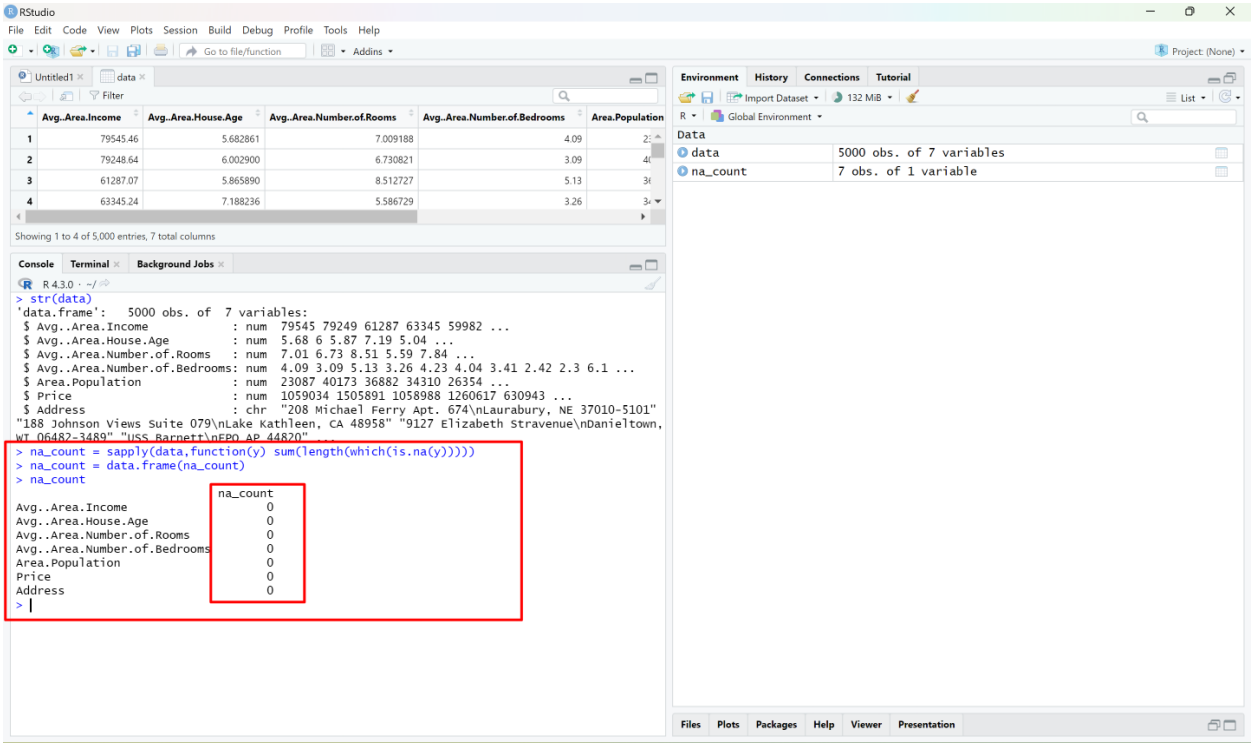
#count the no.of missing values for each column

na_count = sapply(data,function(y) sum(length(which(is.na(y)))))

na_count = data.frame(na_count)

na_count

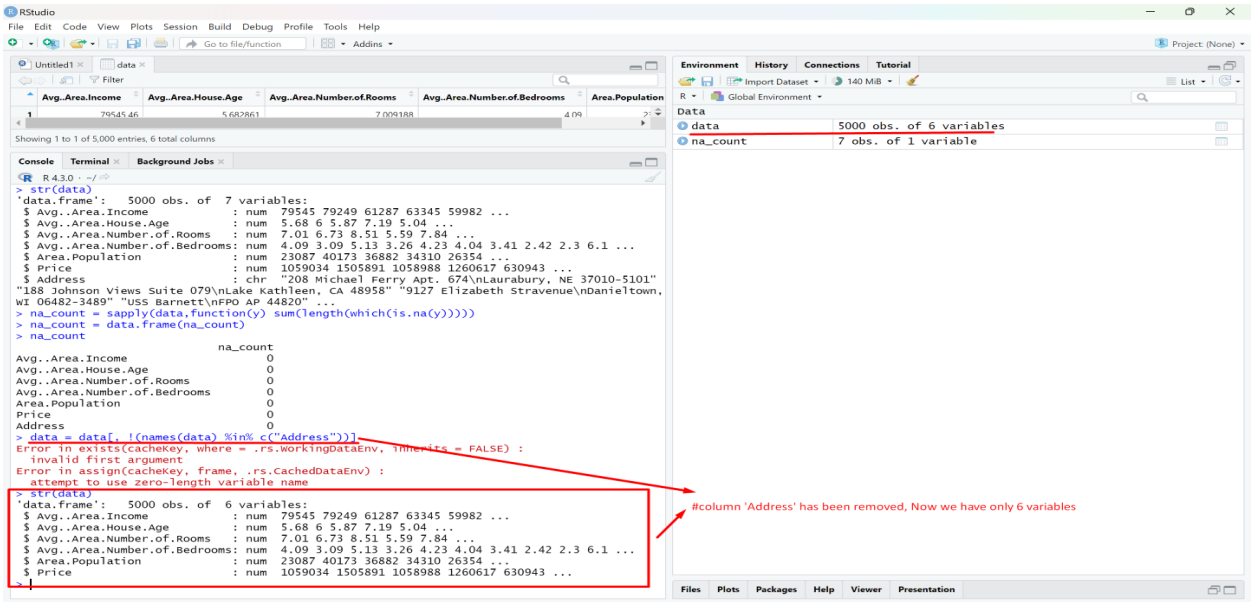
No missing values found in our data set.



2. Then I handled non-numeric data. Removed column 'Address' by using the code:

data = data[, !(names(data) %in% c("Address"))]

str(data)



Now we have the final variables:

- X1 = Avg. Area Income
- X2 = Avg. Area House Age
- X3 = Avg. Area Number of Rooms
- X4 = Avg. Area Number of Bedrooms
- X5 = Area Population
- Y = Price

3. Then I did Descriptive statistics on columns 'Price & Income':

To calculate the descriptive statistics, library 'psych' is installed and loaded.

#library 'psych' installed

install.packages("psych")

library(psych)

Q1, Q2, Q3, average value and variance value can be calculated by using 2 functions.

1. describe ()
2. summary ()
3. var ()

And I Would like to visualize the columns by boxplot by using boxplot () function.

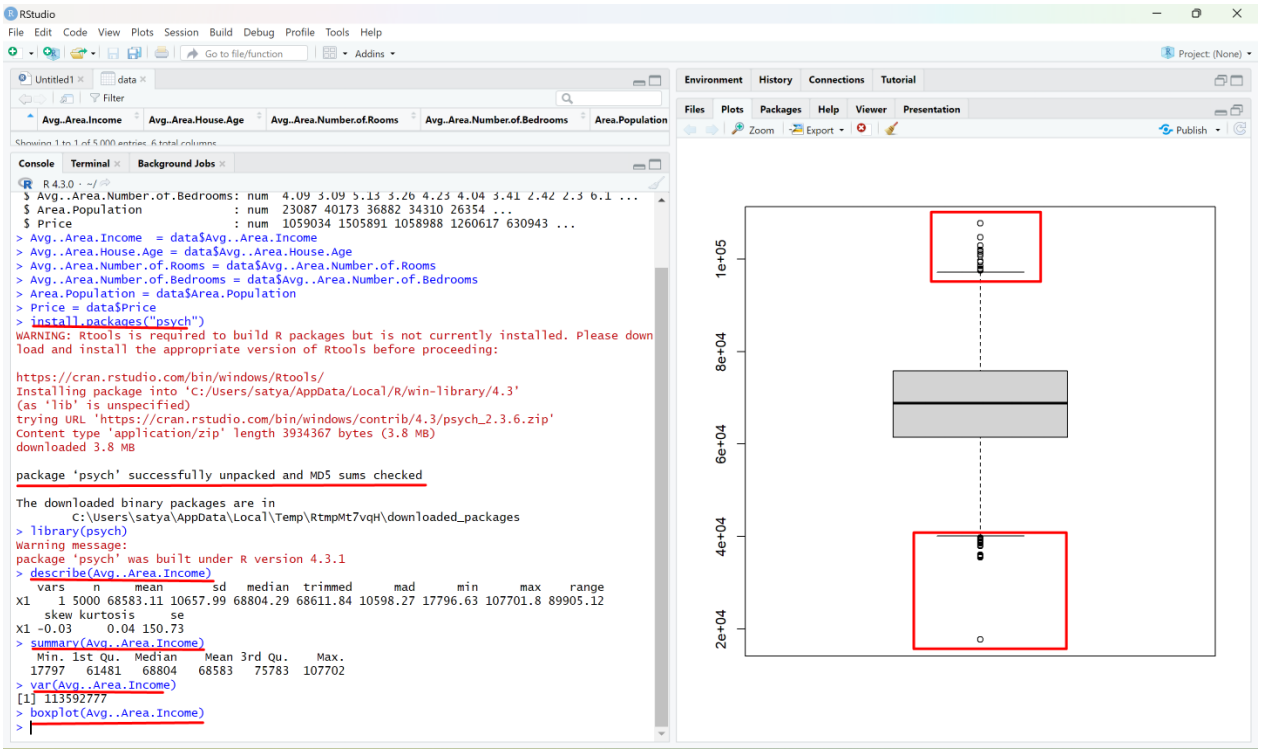
#descriptive statistics on the column 'Income':

describe(Avg..Area.Income)

summary(Avg..Area.Income)

var(Avg..Area.Income)

boxplot(Avg..Area.Income)



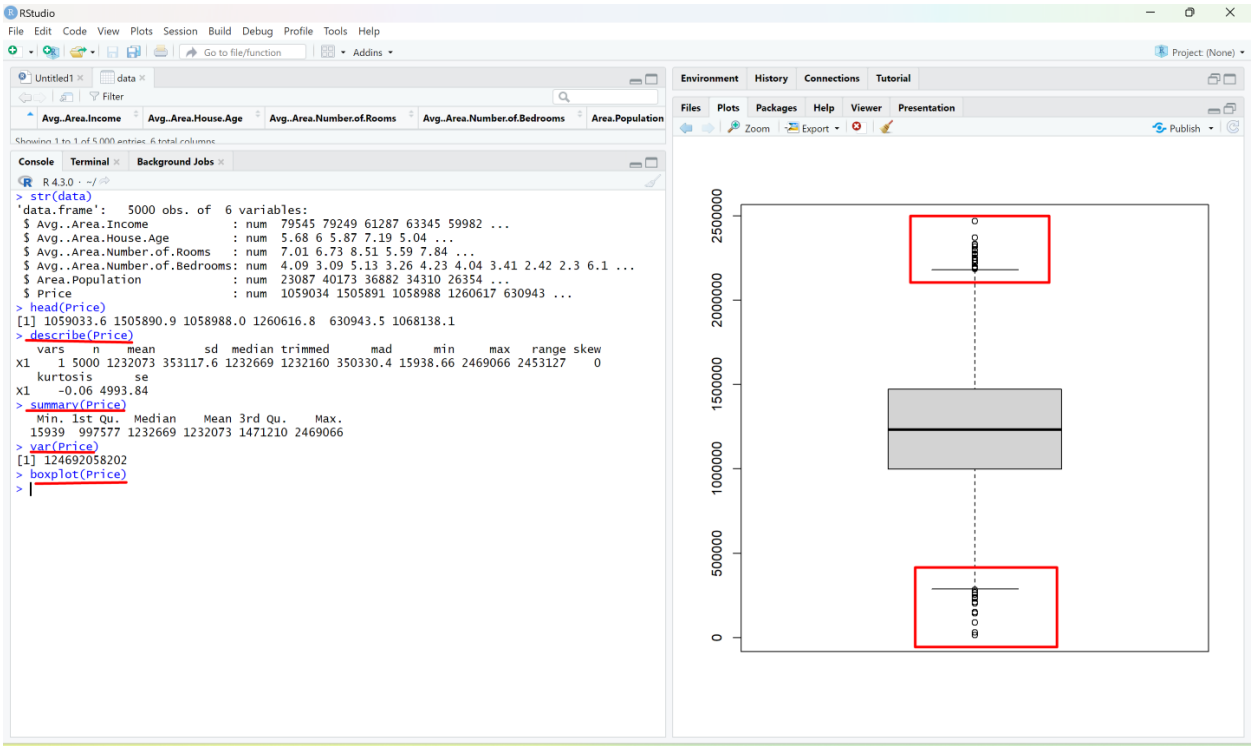
#descriptive statistics on the column 'Price':

describe(Price)

summary(Price)

var(Price)

boxplot(Price)



Interpretation/observations of the boxplots:

	For the column 'Income'	For the column 'Price'
The distribution:	Not normal distribution	Normal distribution
Skewness	-0.03 Slightly negatively skewed	0 No skewness
Potential outliers	There are outliers towards the maximum value and min value	There are outliers towards the maximum value and min value
Variance	113592777	124692058202
Standard deviation	10657.99	353117.6
Min value	17796.63	15938.66
Max value	107701.8	2469066
Q1	61481	997577
Q2/Median	68804.29	1232669
Q3	75783	1471210
Range	89905.12	2453127
Mean	68583.11	1232073
Median/Q2	68804.29	1232669

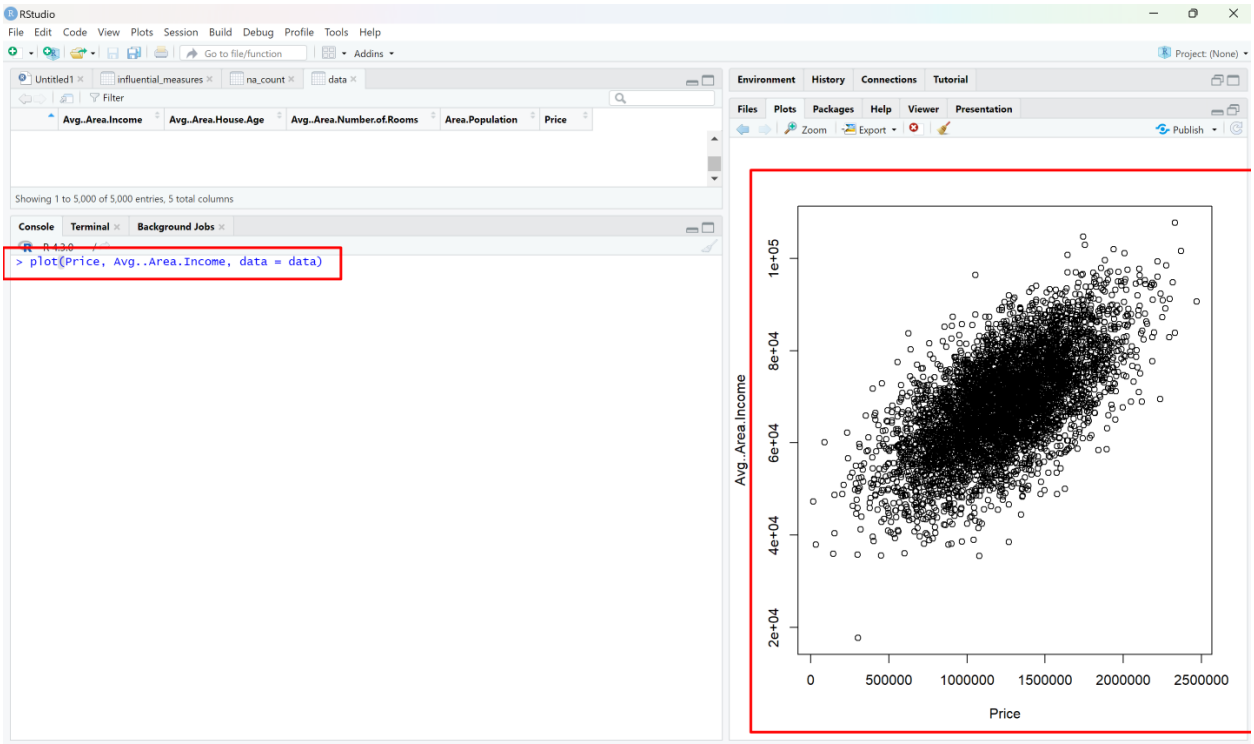
Note: data point = vector with value in multiple variables. In this case we have 6 variables. Just because we identified some data points have outliers in 2 variables, we cannot conclude these data points are outliers in other variables too. Why? Because there are other variables too in our data set.

So I would like to look at the residual plot (res vs pred values / res vs each x-var plots) for the observations which goes beyond the threshold (-3 , +3) after I build the models. Then I would like to consider them as outliers. So, the outliers are not removed in this phase.

But in linear regression we only care about influential points which are also outliers which greatly affect the fitted model. So, I would like to focus more on the influential points after building the models.

Also, I would like to draw scatterplot for these 2 variables:

plot(Price, Avg..Area.Income, data = data)



In the given Scatter plot it shows that as the 'Price' of house increases with increase in 'Avg area Income

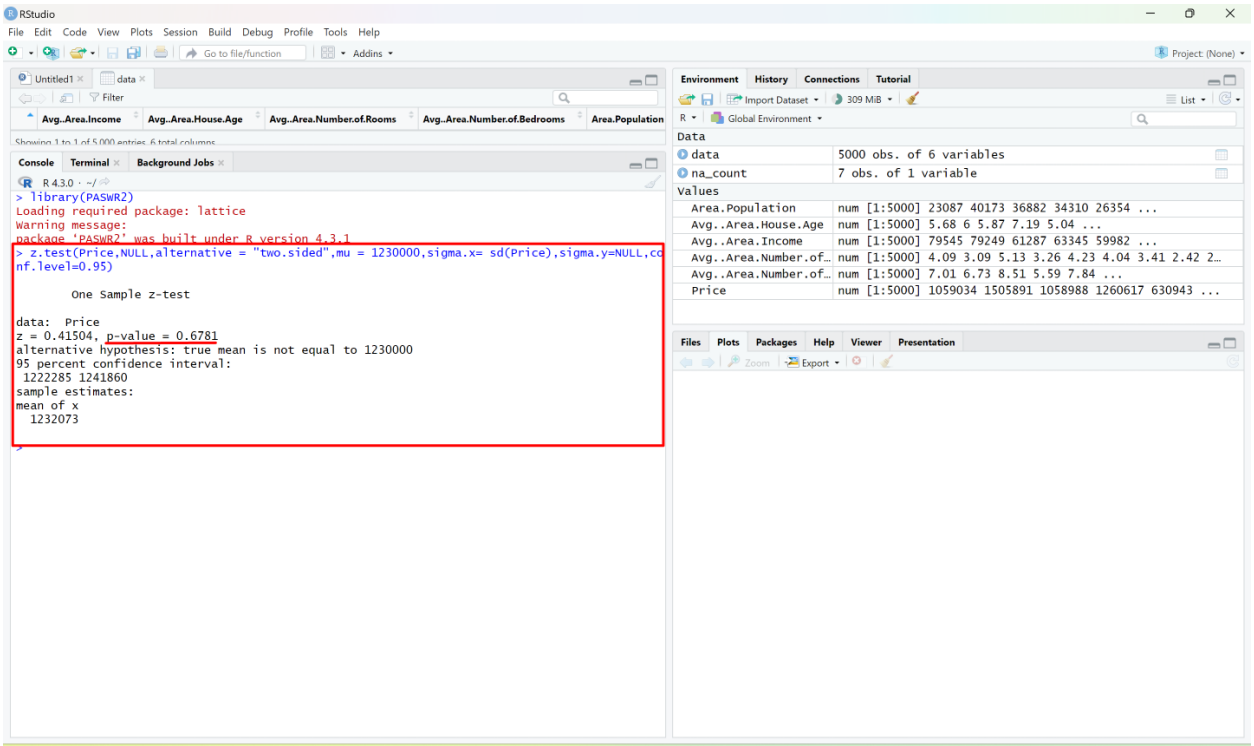
4. Validate the claim/Assumption by using **Hypothesis testing:**

To perform hypothesis testing library "PASWR2" is installed.

install.packages("PASWR2")

library(PASWR2)

z.test(Price,NULL,alternative = "two.sided",mu = 1230000,sigma.x= sd(Price),sigma.y=NULL,conf.level=0.95)



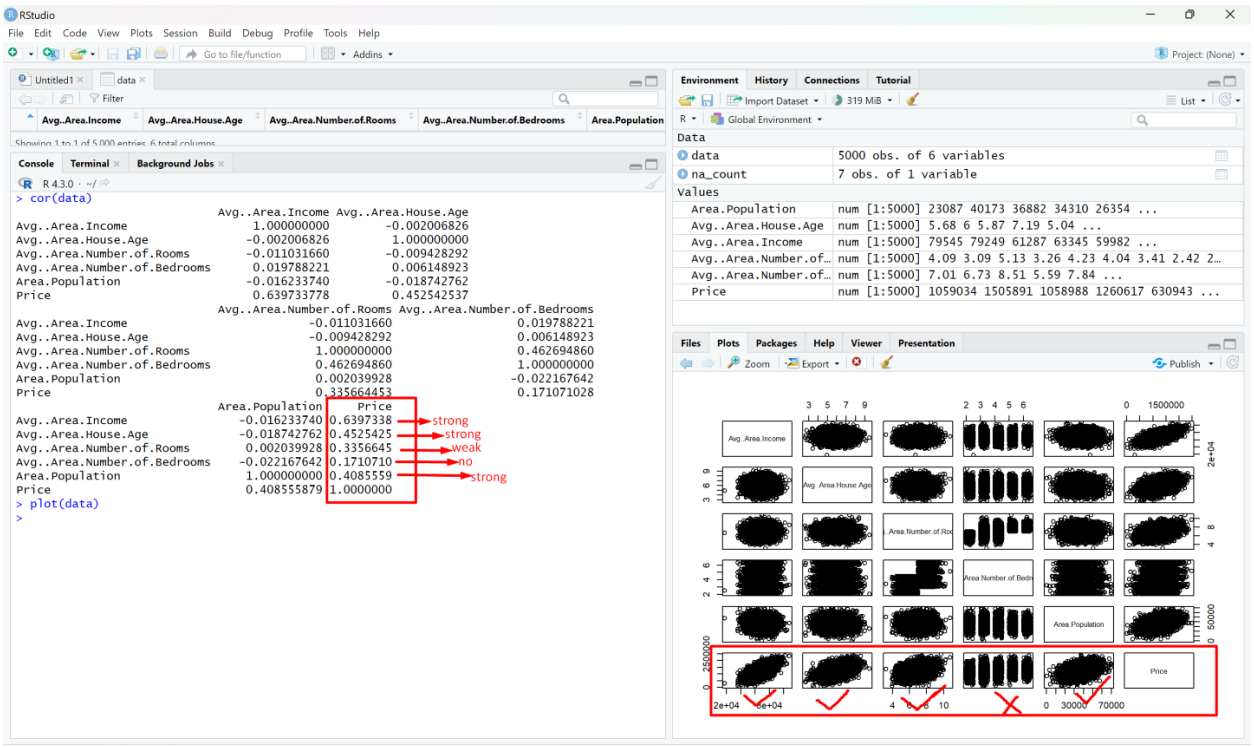
Conclusion:

Based on the z-test performed at a 95% confidence level, we fail to reject the null hypothesis. The p-value is 0.6781, which is greater than the significant level of 0.05. Therefore, we fail to reject null hypothesis (In other words, we accept null hypothesis & reject alt hypothesis). Hence, we cannot conclude that the average housing price is less than or equal to 1.23 million. Hence, I have 95% confidence to reject alternative hypothesis and accept null hypothesis.

5. Handling non-linearity: Examine linear relation b/w X & y variables (check correlations)

cor(data)

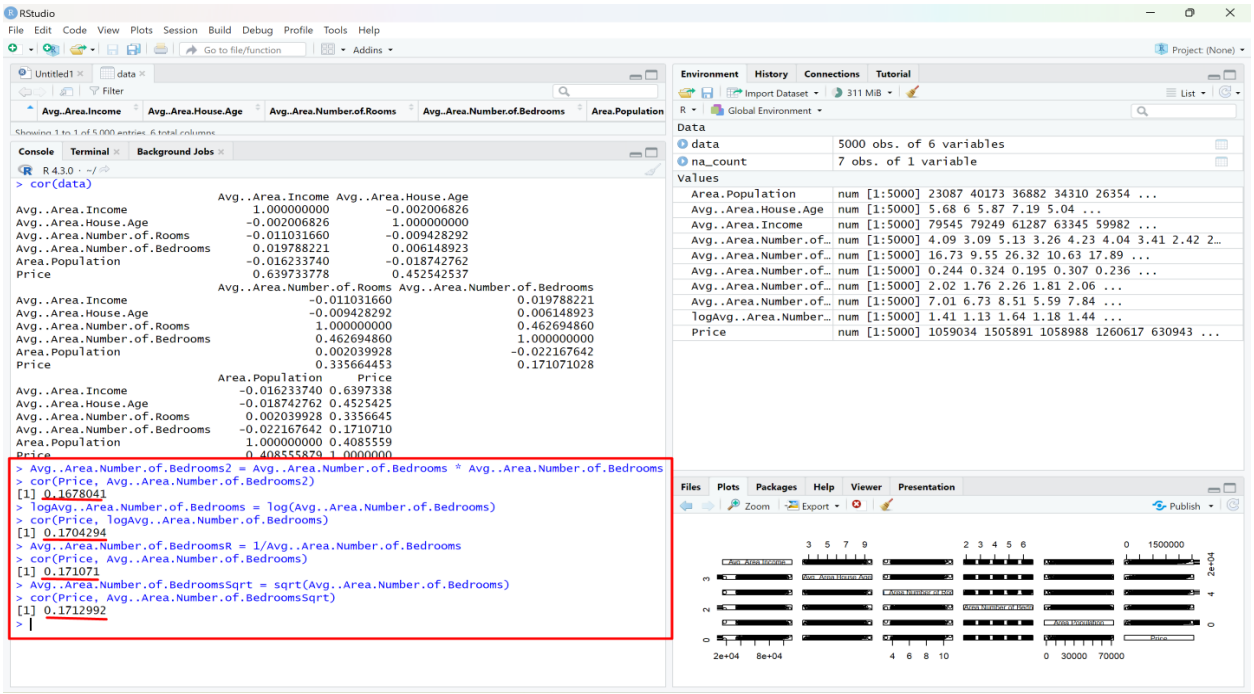
- We noticed that the correlation values are as follows:
  - Price & Avg..Area.Income = 0.639 - strong Correlation - +ve correlation
  - Price & Avg..Area.House.Age = 0.452 - strong Correlation - +ve correlation
  - Price & Avg..Area.Number.of.Rooms = 0.335 - weak Correlation – +ve correlation
  - Price & Avg..Area.Number.of.Bedrooms = 0.171 - No Correlation – -ve correlation
  - Price & Area.Population = 0.408 – strong Correlation – +ve correlation



- As Price(Y-variable) & Avg..Area.Number.of.Bedrooms (X-var) has no correlation , we can try transformation on this x- variable & re-calculate the correlation with Price (y-variable) again.
  - Square transformation: X' = X * X
  - Log transformation: X' = logX
  - Inversion transformation: X' = 1/X
  - Square root transformation: X' = sqrt(X)

Transformation on Avg..Area.Number.of.Bedrooms (X-var): (code)

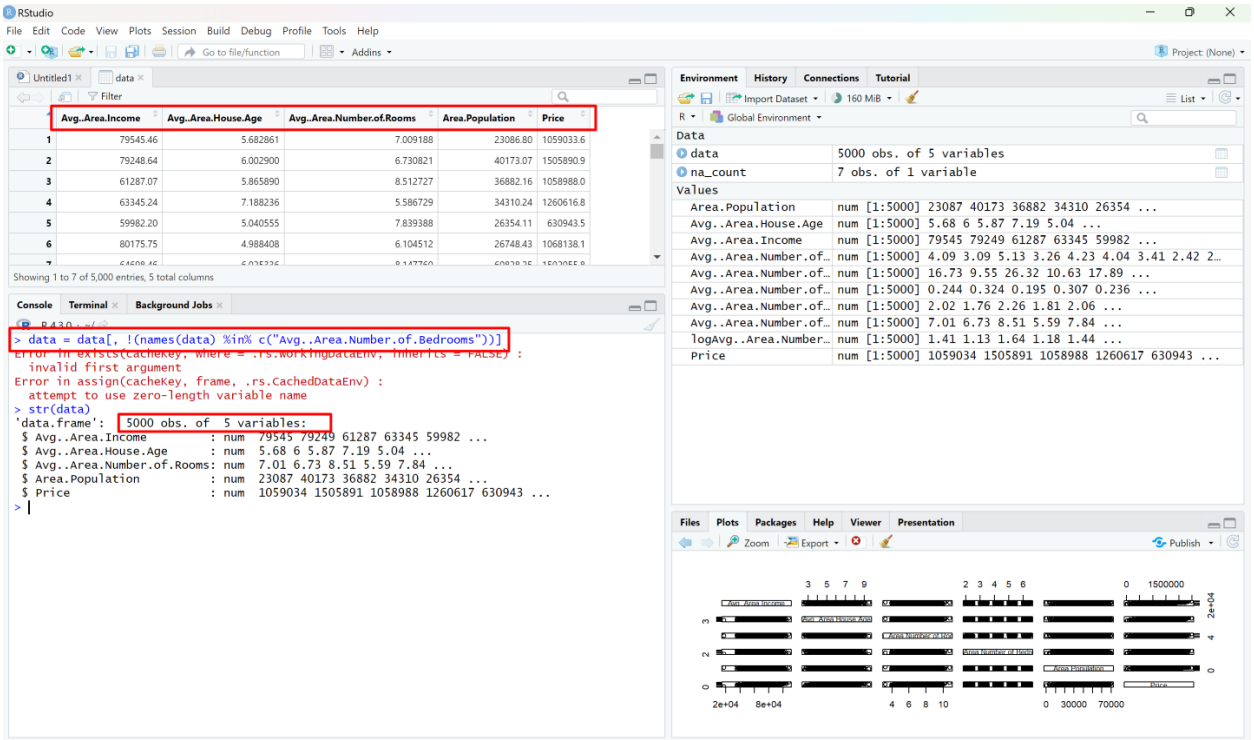
```
Avg..Area.Number.of.Bedrooms2 = Avg..Area.Number.of.Bedrooms *
Avg..Area.Number.of.Bedrooms
cor(Price, Avg..Area.Number.of.Bedrooms2)
logAvg..Area.Number.of.Bedrooms = log(Avg..Area.Number.of.Bedrooms)
cor(Price, logAvg..Area.Number.of.Bedrooms)
Avg..Area.Number.of.BedroomsR = 1/Avg..Area.Number.of.Bedrooms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



As the correlation didn't improve after transformation on Avg..Area.Number.of.Bedrooms variable, we decided to drop that variable.

#droping the x-variable:

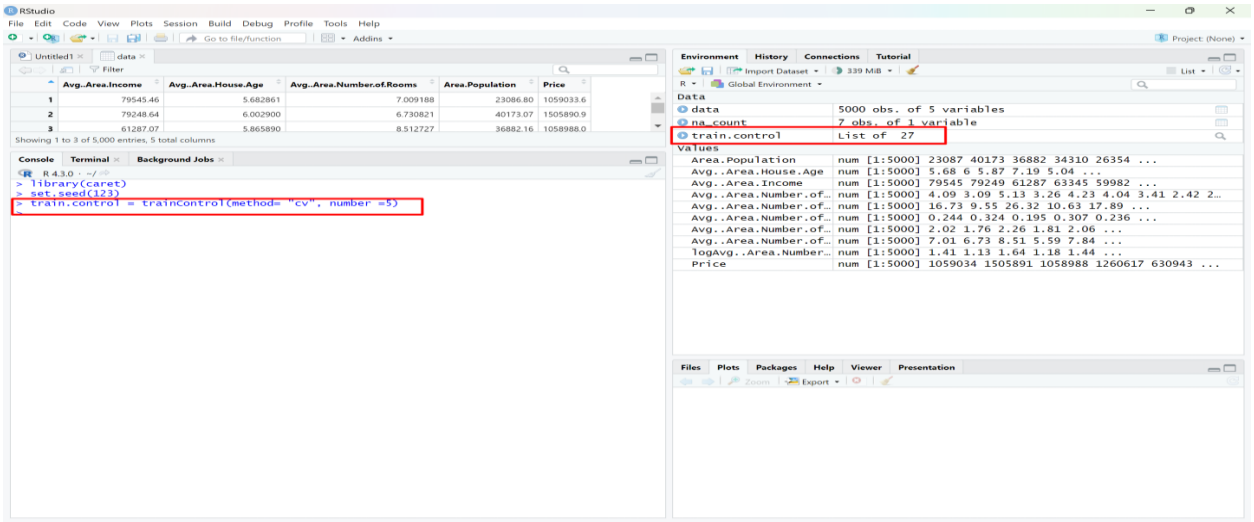
data = data[, !(names(data) %in% c("Avg..Area.Number.of.Bedrooms"))]



Now we have 4 x-variables (avg area income, Avg area housing age, Avg area no.of rooms, Avg area population) and 1 -variable (Price)

6. data split by N-fold cross evaluation:

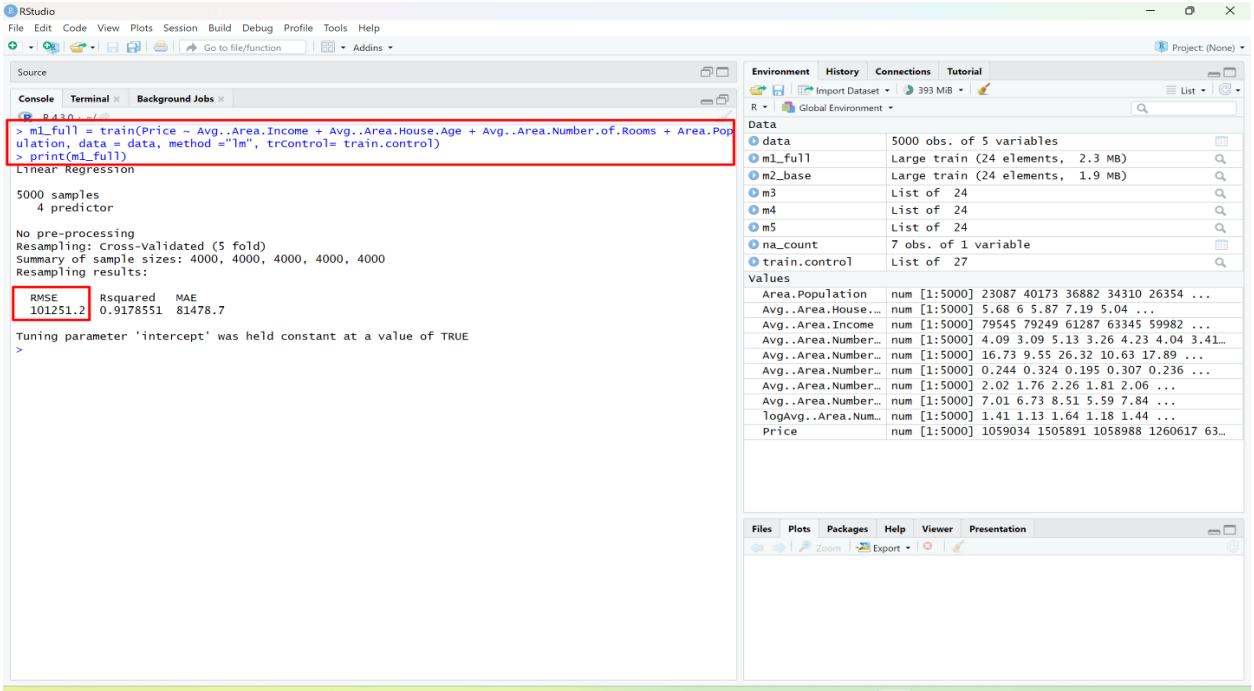
```
#library caret installed
install.packages("caret")
library(caret)
set.seed(123)
train.control = trainControl(method= "cv", number =5)
```



7.Build 5 different multiple linear regression models to compare them:

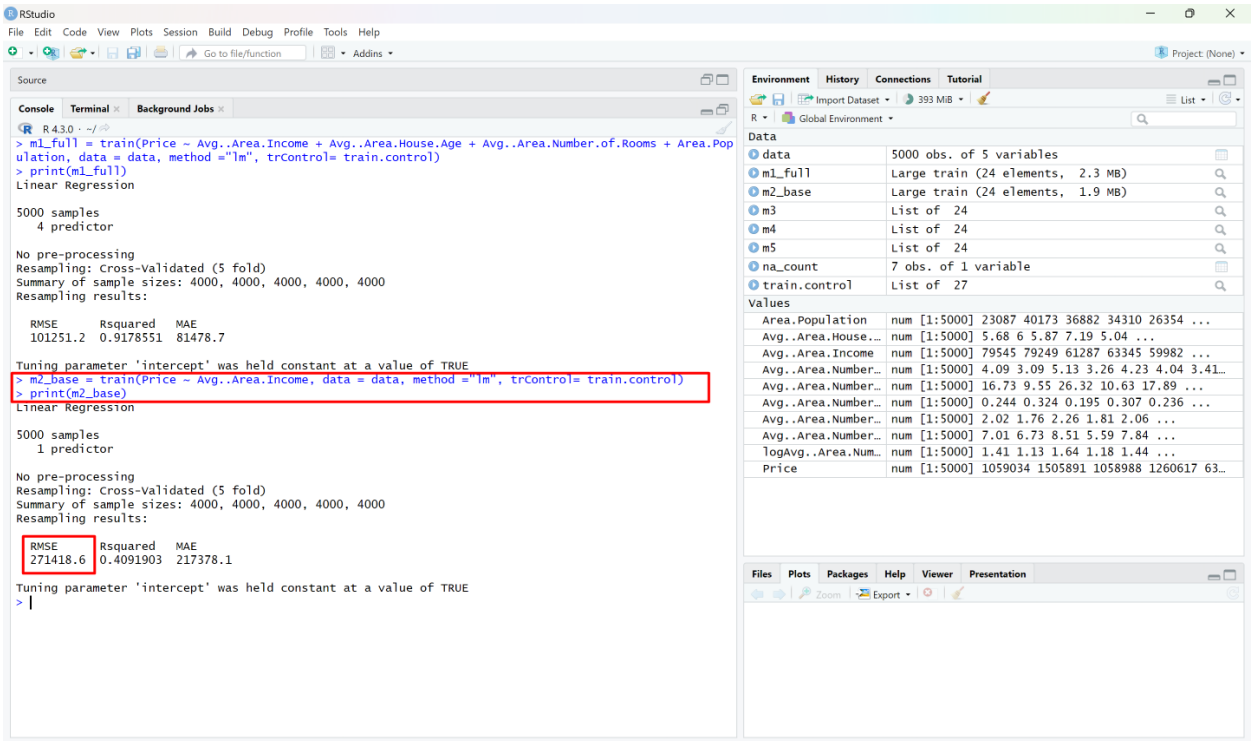
Building **Full model** by using all x-variables & target variable: **M1**

```
m1_full = train(Price ~ Avg..Area.Income + Avg..Area.House.Age + Avg..Area.Number.of.Rooms
+ Area.Population, data = data, method ="lm", trControl= train.control)
print(m1_full)
```



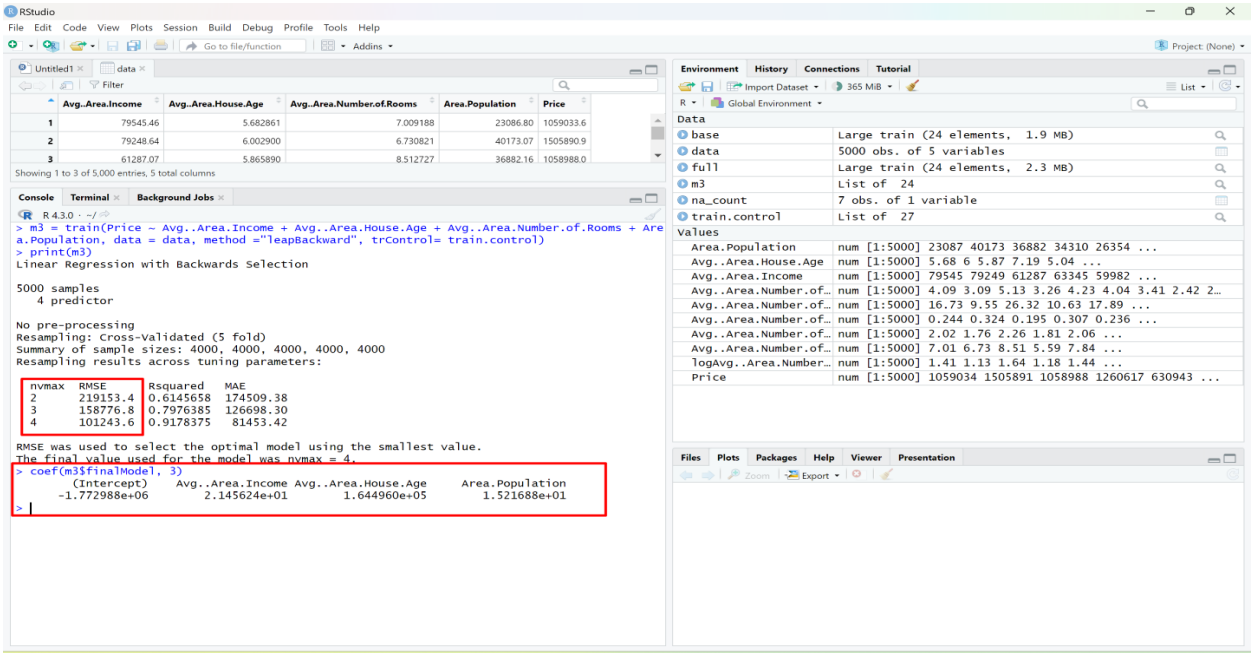
Building **base model** with just one x-variable & target variable:**M2**

```
m2_base = train(Price ~ Avg..Area.Income, data = data, method ="lm", trControl= train.control)
print(m2_base)
```



Building m3 by backward method: **M3**

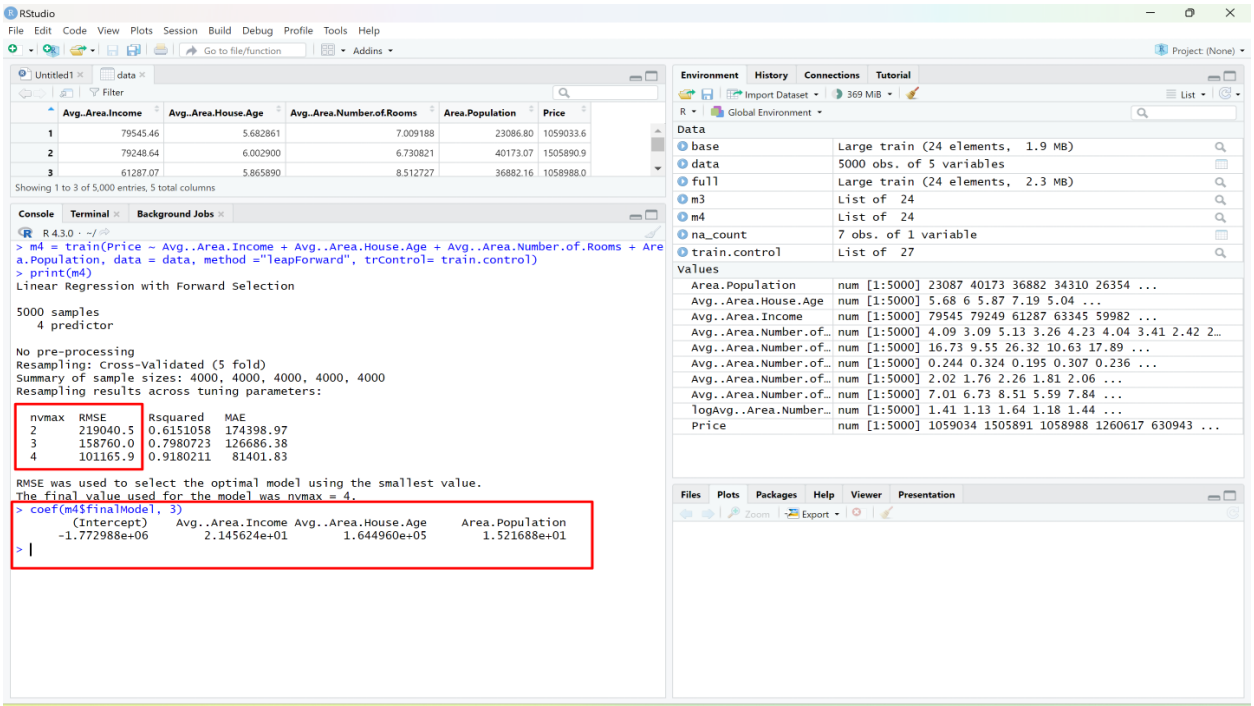
```
m3 = train(Price ~ Avg..Area.Income + Avg..Area.House.Age + Avg..Area.Number.of.Rooms + Area.Population, data = data, method ="leapBackward", trControl= train.control)
print(m3)
coef(m3$finalModel, 3)
```



Observation: M3 used 3 x-var's (Avg..Area.Income Avg..Area.House.Age Area.Population) and 1 intercept

Building m4 by forward method: **M4**

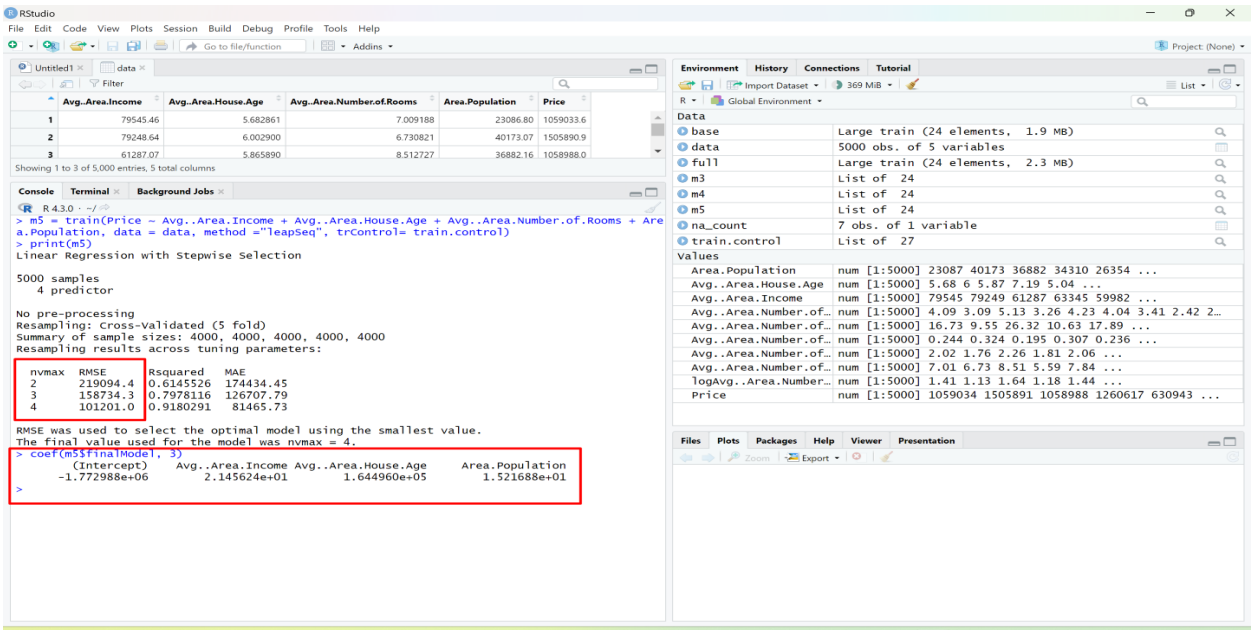
```
m4 = train(Price ~ Avg..Area.Income + Avg..Area.House.Age + Avg..Area.Number.of.Rooms + Area.Population, data = data, method ="leapForward", trControl= train.control)
print(m4)
coef(m4$finalModel, 3)
```



Observation: M4 used 3 x-var's (Avg..Area.Income Avg..Area.House.Age Area.Population) and 1 intercept

building m5 by forward method: **M5**

```
m5 = train(Price ~ Avg..Area.Income + Avg..Area.House.Age + Avg..Area.Number.of.Rooms + Area.Population, data = data, method ="leapSeq", trControl= train.control)
print(m5)
coef(m5$finalModel, 3)
```

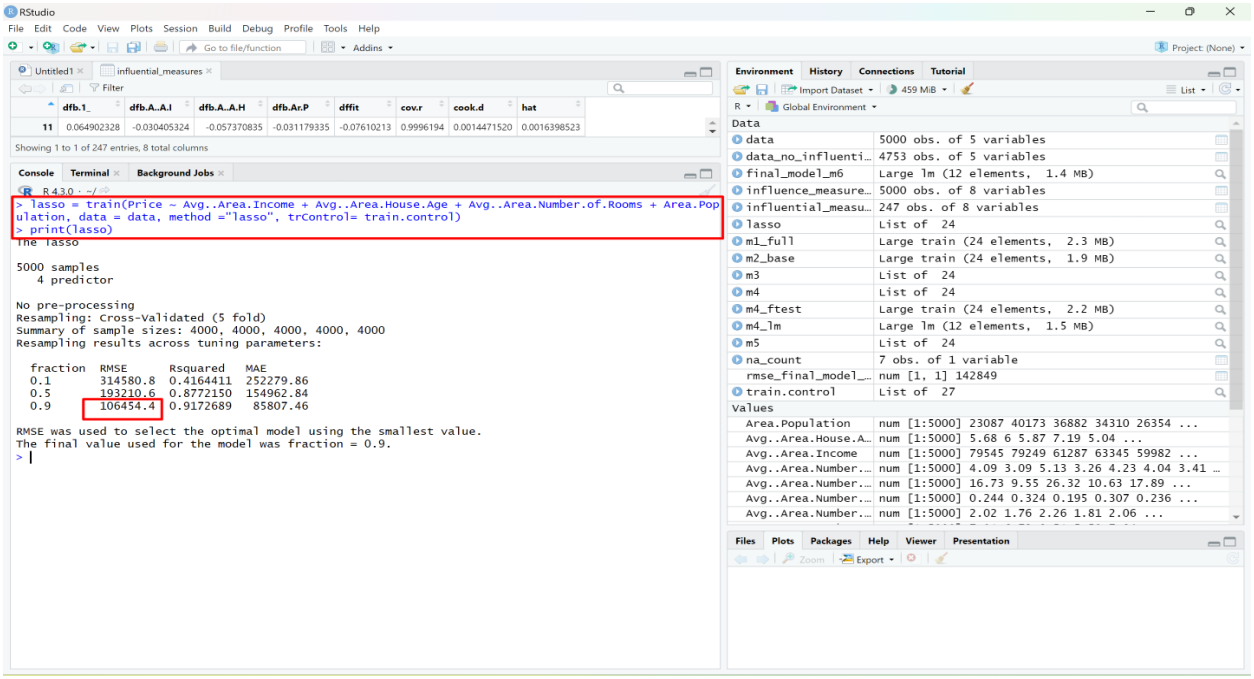


Observation: M5 used 3 x-var's (Avg..Area.Income Avg..Area.House.Age Area.Population) and 1 intercept

building lasso regression model: **Lasso**

code:

```
Install.packages("lars")
install.packages("elasticnet")
library(lars)
library(elasticnet)
lasso = train(Price ~ Avg..Area.Income + Avg..Area.House.Age + Avg..Area.Number.of.Rooms + Area.Population, data = data, method ="lasso", trControl= train.control)
```



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

Selected Model: M4 as it has less RMSE, so I would like to pick m4 to improve the model by seeking different methods like (VIF & removing influential points) & to perform model diagnosis for the model M4.

7. Model diagnosis:

F-test on M4:

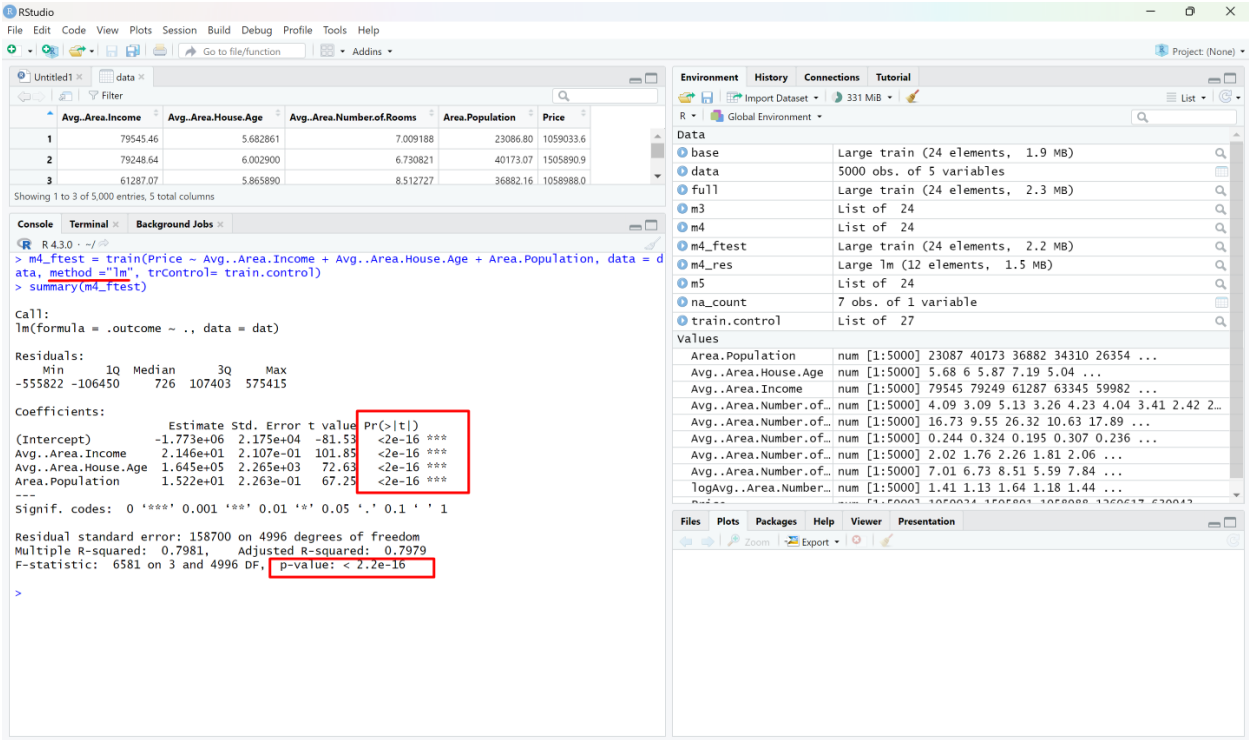
- Null hypothesis : (H0): The coefficients of all x-variables are zero and there is no linear relationship with Price.
- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

- 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.
- $H_a : \beta_j \neq 0$

Note : In these hypotheses,
β1 = coefficient of " Avg..Area.Income,"
β2 = coefficient of " Avg..Area.House.Age,"
β3 = coefficient of " Area.Population."

#build m4 again by using lm method with the x-variable that got used in m4:

m4_ftest = train(Price ~ Avg..Area.Income + Avg..Area.House.Age + Area.Population, data = data, method ="lm", trControl= train.control)
summary(m4_ftest)



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.

******By conducting an F-test, we can observe that there is sufficient evidence to reject the null hypothesis and conclude that the independent variables have a significant impact on the dependent variable. *********

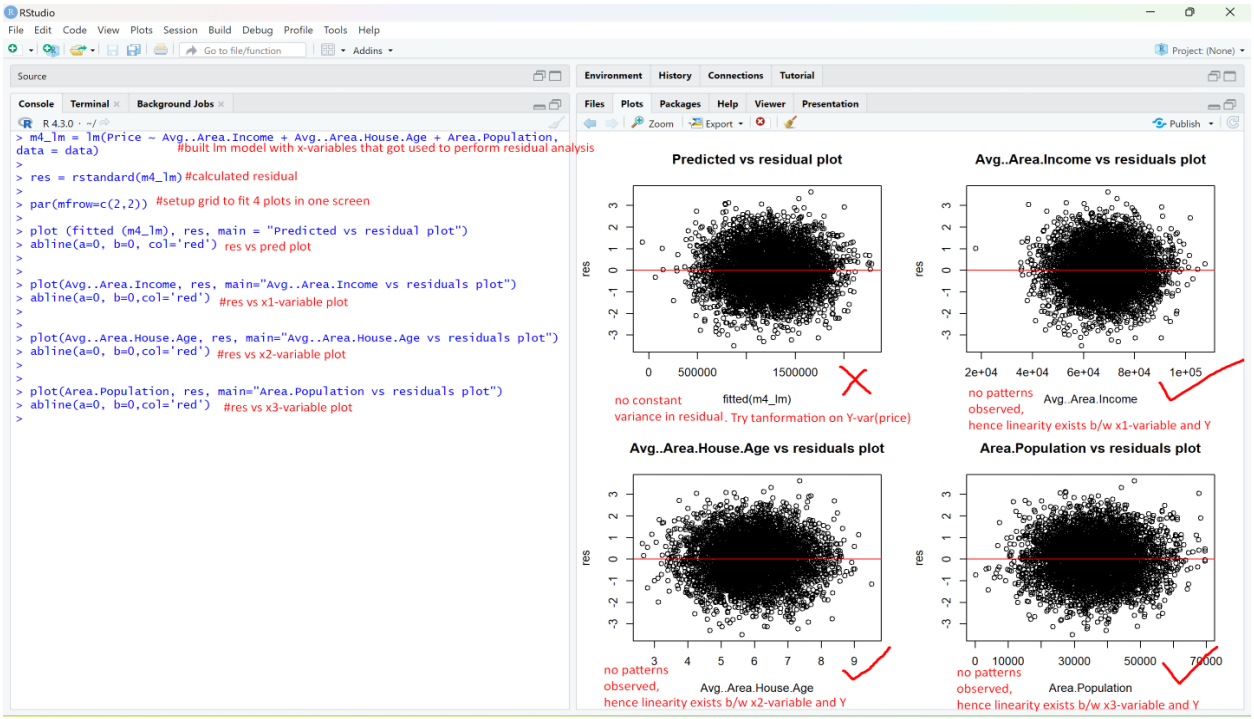
Residual analysis on M4:

- Goals in residual analysis:

1. Validate the constant variance: Plot residuals vs predicted values: To check constant variance for the residuals.
2. Validate the linearity relationship: Plot residuals vs each x-variable: To check linearity assumptions for Y and the x-variable.

m4_lm = lm(Price ~ Avg..Area.Income + Avg..Area.House.Age + Area.Population, data = data)
#calculate residual for the model m4:
res = rstandard(m4_lm)
#res vs pred plot:
par(mfrow=c(2,2))

```
plot (fitted (m4_lm), res, main = "Predicted vs residual plot")
abline(a=0, b=0, col='red')
#res vs x1 variable plot:
plot(Avg..Area.Income, res, main="Avg..Area.Income vs residuals plot")
abline(a=0, b=0,col='red')
#res vs x2 variable plot:
plot(Avg..Area.House.Age, res, main="Avg..Area.House.Age vs residuals plot")
abline(a=0, b=0,col='red')
#res vs x3 variable plot:
plot(Area.Population, res, main="Area.Population vs residuals plot")
abline(a=0, b=0,col='red')
```



Observation: Constant variance: I felt Spread is not too bad Still I'd like to try transformation on Y-variable (Price)

Observation: No patterns observed; hence linearity exists b/w X1- variable and Y-variable.

Observation: No patterns observed; hence linearity exists b/w X2- variable and Y-variable

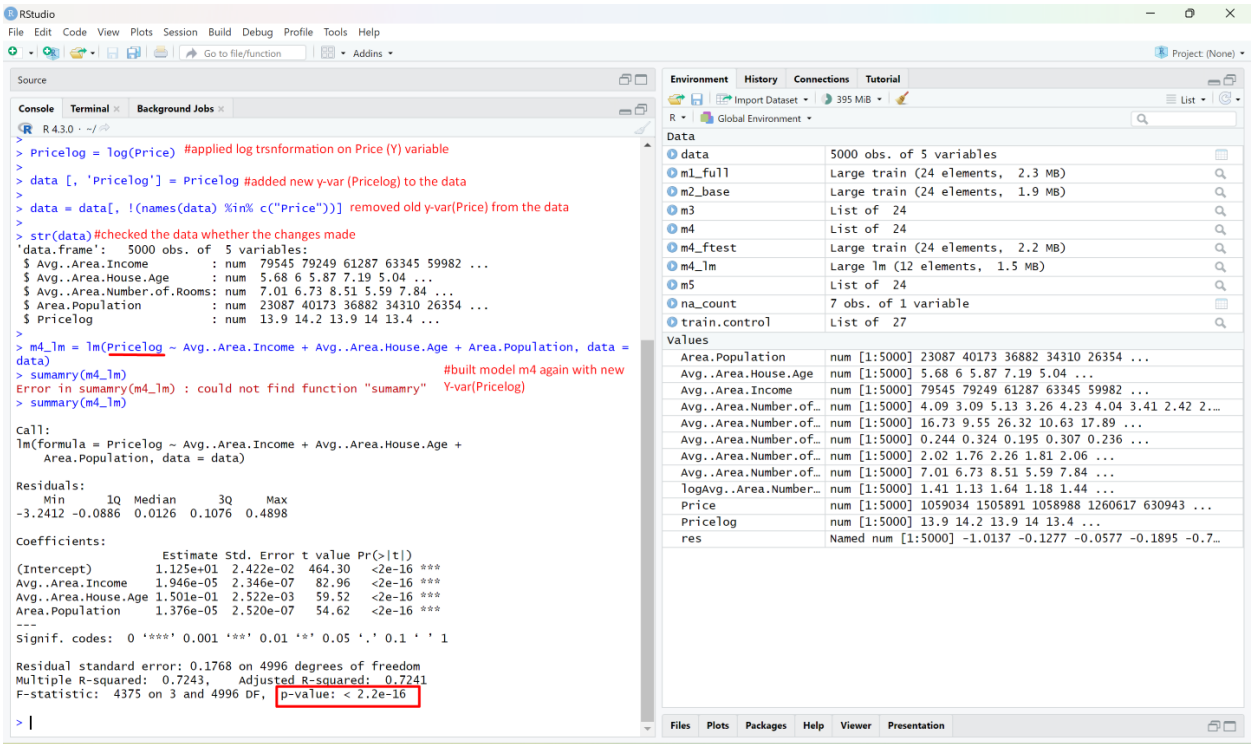
Observation: No patterns observed; hence linearity exists b/w X3- variable and Y-variable

Tranformation on Y-variable & rebuilding the model again:

```
Pricelog = log(Price)
#add new y_variable (Pricelog) to the data:
data [, 'Pricelog'] = Pricelog
#drop original y-variable (Price) from the data:
data = data[, !(names(data) %in% c("Price"))]
Re_check the data whether the changes made:
str(data)
```

#re-build m4_lm again with new y-variable:

m4_lm = lm(Pricelog ~ Avg..Area.Income + Avg..Area.House.Age + Area.Population, data = data)



Re-perform residual analysis on m4_lm again:

#re-calculate residual for the model m4_lm:

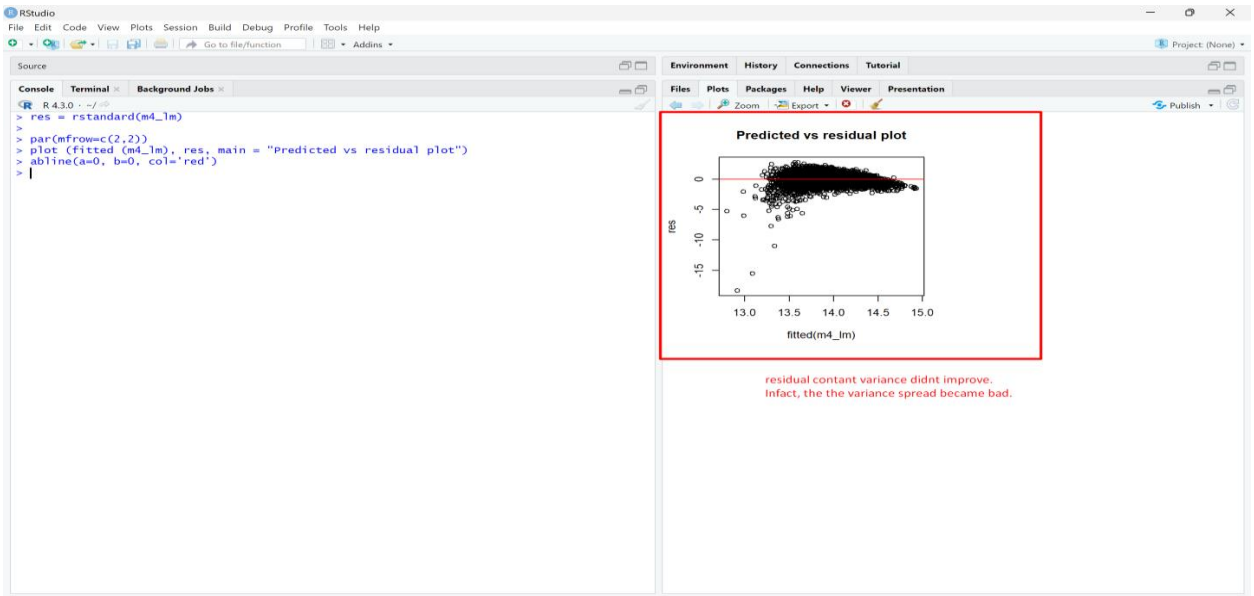
res = rstandard(m4_lm)

#res vs pred plot:

par(mfrow=c(2,2))

plot (fitted (m4_lm), res, main = "Predicted vs residual plot")

abline(a=0, b=0, col='red')



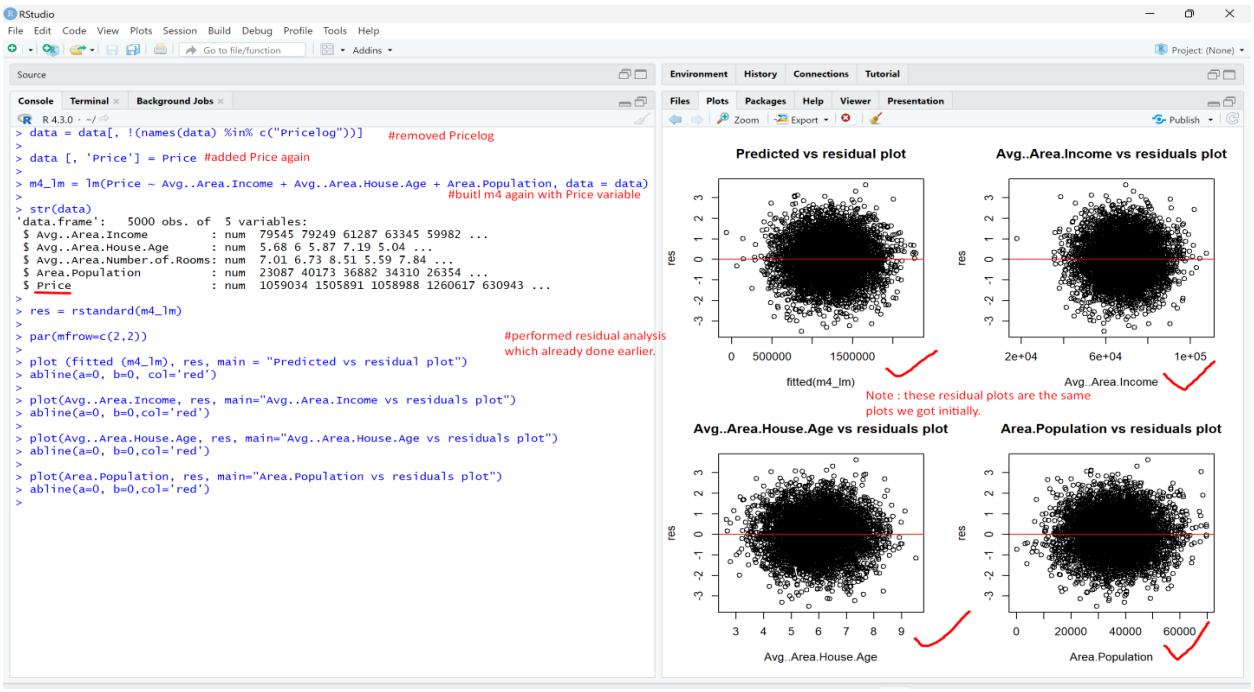
Note:

Observation: As the log transformation on Y-variable(price) didn't improve the variance. In fact, the variance spread got much worse. So, I decided to go with the previous model where I used non-transformed original y-variable (Price)

As residual constant variance got much more worse Deleting the Pricelog variable and re-adding the original Price variable to the data and rebuilding the model again with original Price variable:

```
#No improvement, residual spread got much more worse
#so Adding original Price variable & removing Pricelog variable & rebuilding the m4 again:
#drop Pricelog to data:
data = data[, !(names(data) %in% c("Pricelog"))]
#add Price to data:
data [, 'Price'] = Price
#rebuild m4 again with Price variable
m4_lm = lm(Price ~ Avg..Area.Income + Avg..Area.House.Age + Area.Population, data = data)
#check whether the changes made:
str(data)
#calculate residual for the model m4:
res = rstandard(m4_lm)
#res vs pred plot:
par(mfrow=c(2,2))
plot (fitted (m4_lm), res, main = "Predicted vs residual plot")
abline(a=0, b=0, col='red')

#res vs x1 variable plot:
plot(Avg..Area.Income, res, main="Avg..Area.Income vs residuals plot")
abline(a=0, b=0,col='red')
#res vs x2 variable plot:
plot(Avg..Area.House.Age, res, main="Avg..Area.House.Age vs residuals plot")
abline(a=0, b=0,col='red')
#res vs x3 variable plot:
plot(Area.Population, res, main="Area.Population vs residuals plot")
abline(a=0, b=0,col='red')
```

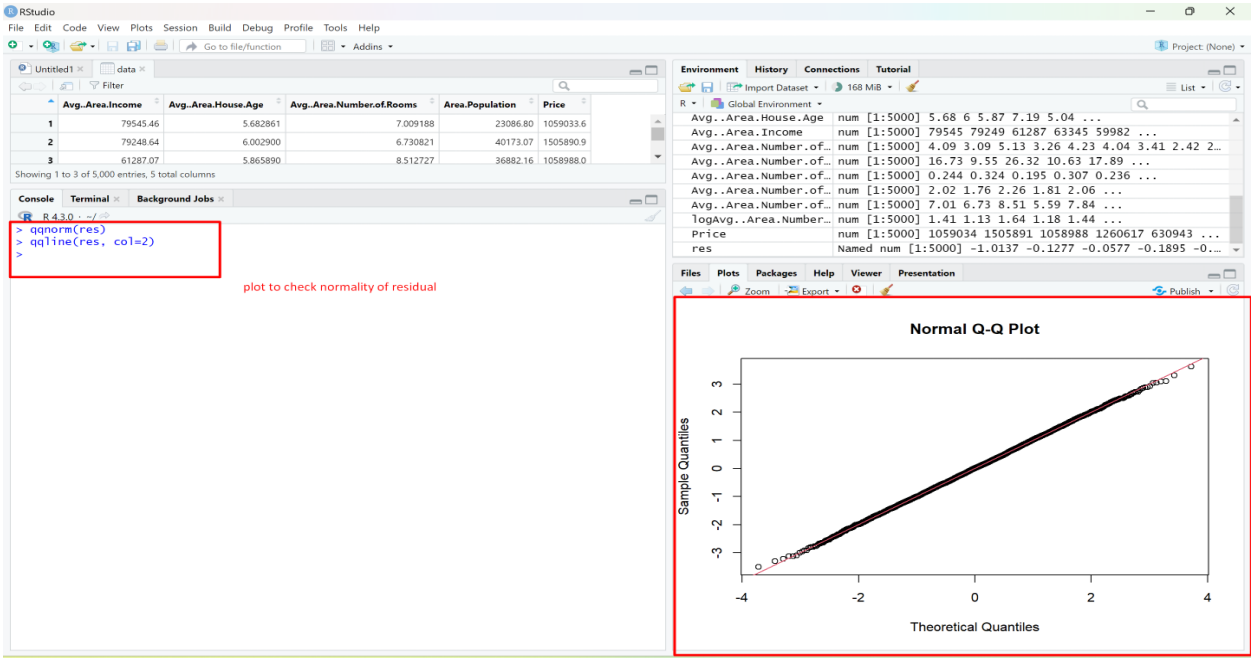


3. Validate normal distribution of residuals: Draw normal probability plot of residuals: To check normality assumption for the error terms; if points lie close to a line, the errors can be assumed to be approximately normal. Otherwise, the assumption of normality is not satisfied.

#qqplot to check normality of residual:

qqnorm(res)

qqline(res, col=2)



Observation: The points lie close to the line, hence the residual follows normal distribution

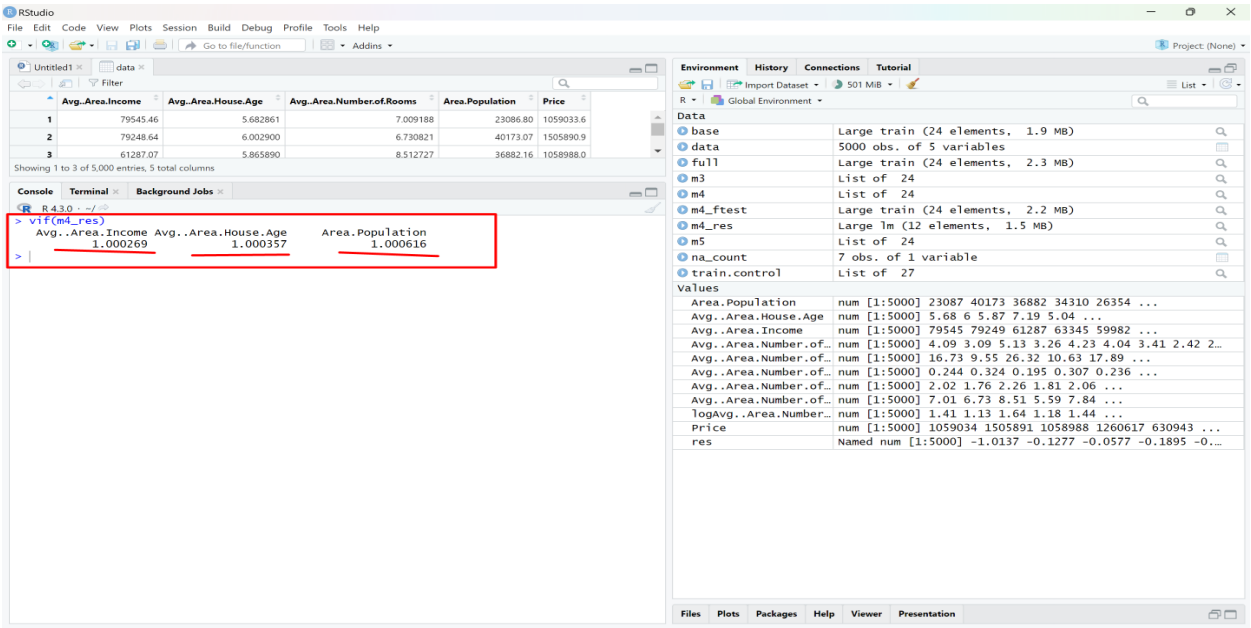
Conclusion: M4 qualified in the f- test and residual analysis. Hence M4 passed model diagnosis.

8. Checking multicollinearity for the selected model above (M4) by calculating VIF:

VIF for M4:

Library “car” installed.

```
install.packages("car")
library(car)
vif(m4_res)
```



Observation: Checked for multicollinearity for the selected model: M4. All the absolute values are less than 4. There is no high correlation between the independent variables.

9. Checking for Influential points by cooks.distance:

#library “stats” is installed

If we use cook distance, any points with cook distance > 4/n are influential points. We need to remove them. And re-build the model and check the RMSE. Here, n = data size = 5000; 4/5000 = 0.0008

So, we need to remove any data points with values larger than 0.0008:

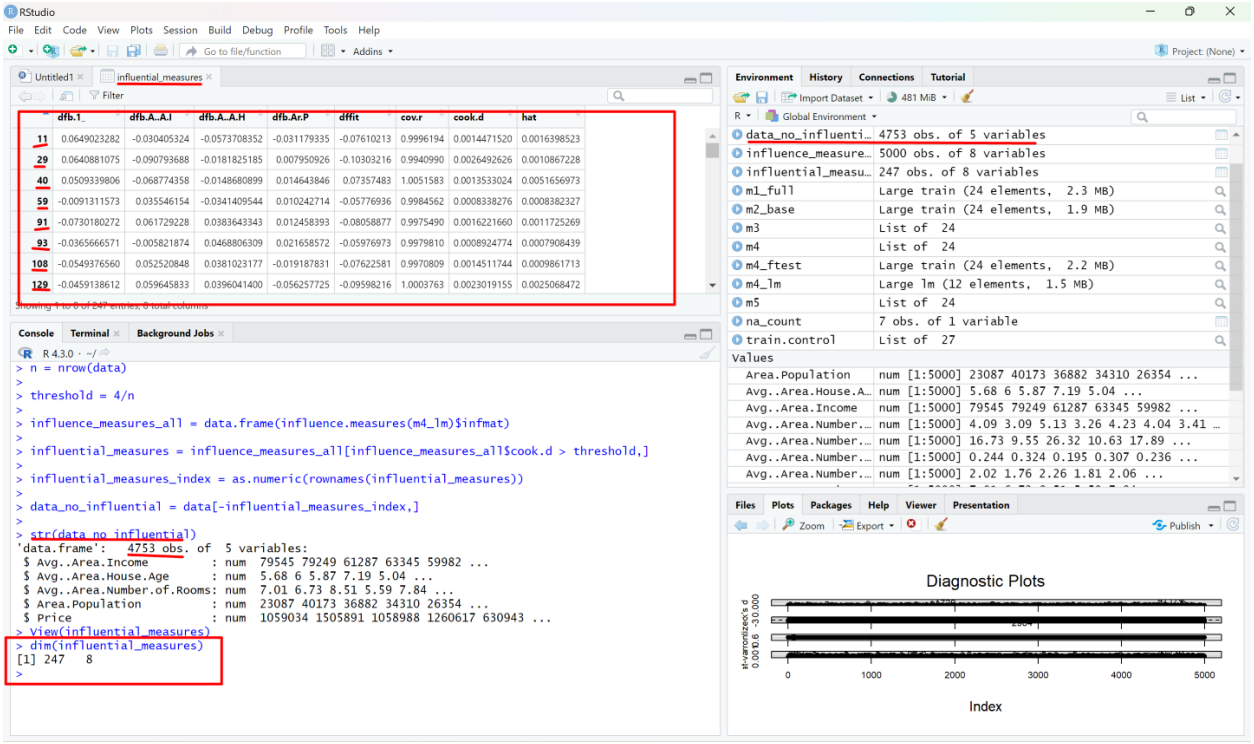
247 influential points were observed and removed from the data and created new data called data_no_influential

Code:

```
install.packages("stats")
library(stats)

n = nrow(data)
threshold = 4/n
influence_measures_all = data.frame(influence.measures(m4_lm)$infmat)
influential_measures = influence_measures_all[influence_measures_all$cook.d > threshold,]
```

influential_measures_index = as.numeric(rownames(influential_measures))
data_no_influential = data[-influential_measures_index,]



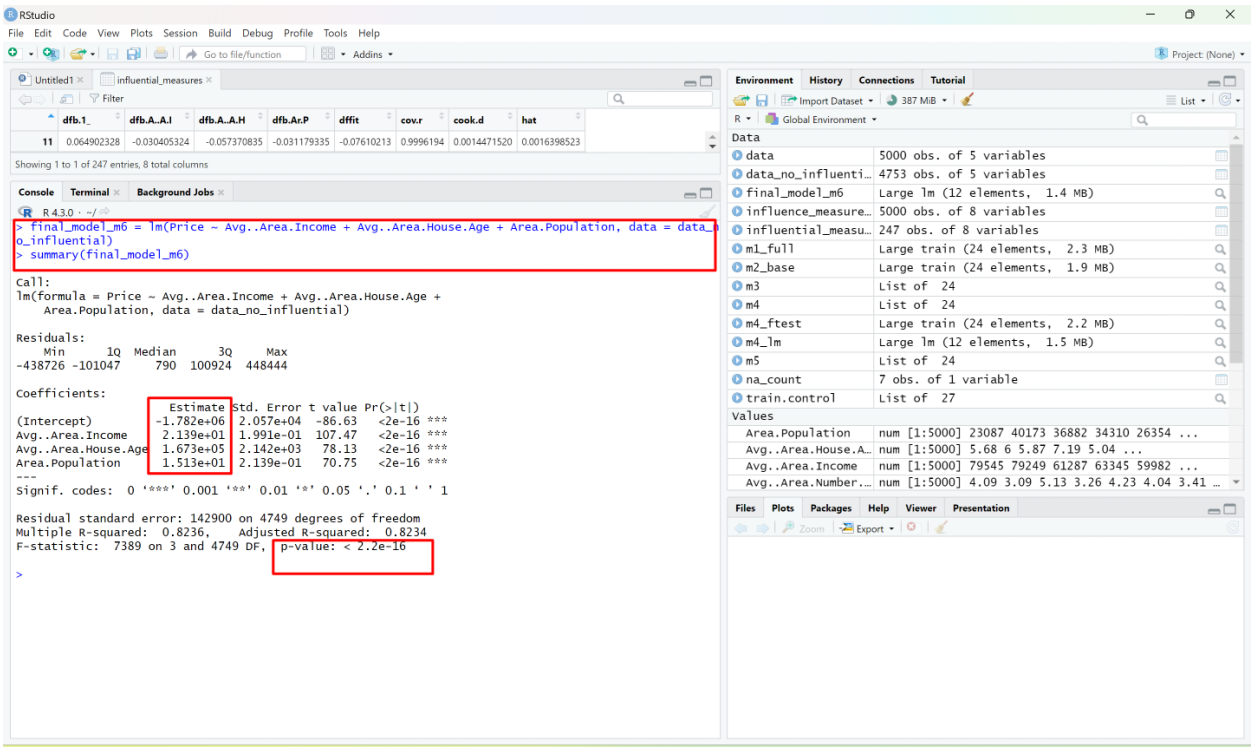
After removing Influential points We got new data called “data_no_influential”

Rebuild the m4 again with the name “final model m6” with the new data (data_no_influential_points):

Code:

final_model_m6 = lm(Price ~ Avg..Area.Income + Avg..Area.House.Age + Area.Population, data = data_no_influential)

summary(final_model_m6)



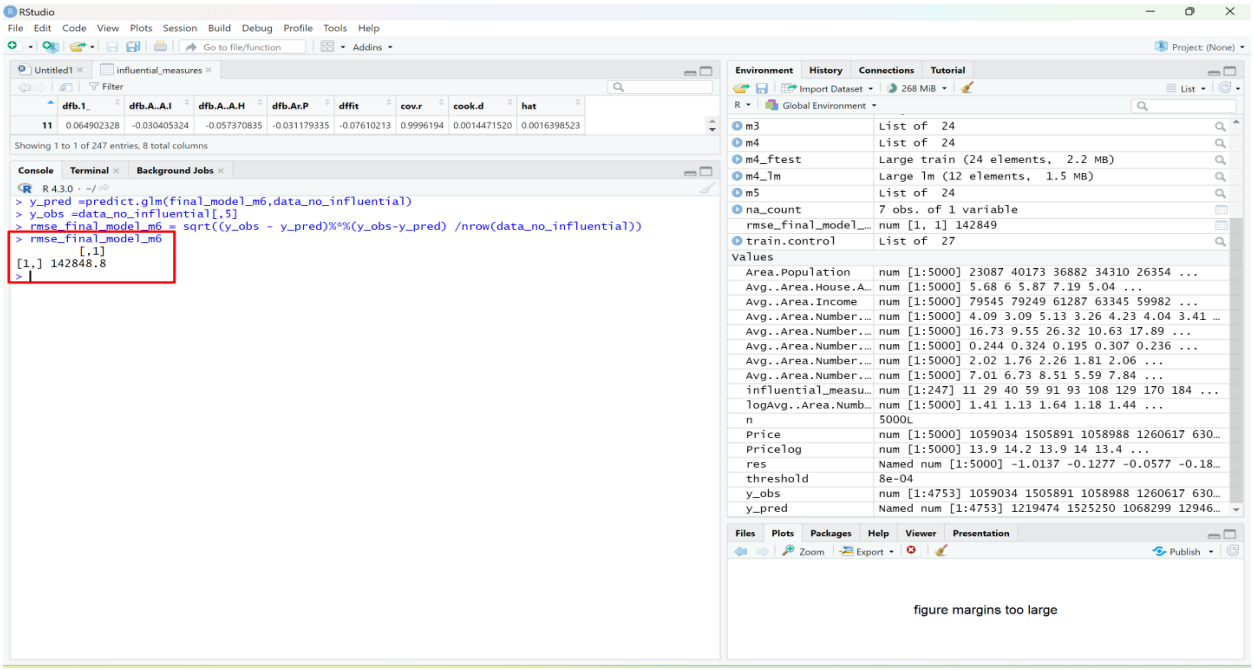
RMSE for the final model M6:

Code: y_pred =predict.glm(final_model_m6,data_no_influential)

y_obs =data_no_influential[,5]

rmse_final_model_m6 = sqrt((y_obs – y_pred)%*%(y_obs-y_pred) /nrow(data_no_influential))

rmse_final_model_m6



5.2. Evaluations and Results

Given the same problem, you may have several solutions or build several models. Evaluate your solutions based on selected metrics and compare them.

Given the problem "Predicting housing price using linear regression model" I have built. 6 models:

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

5.3. Findings

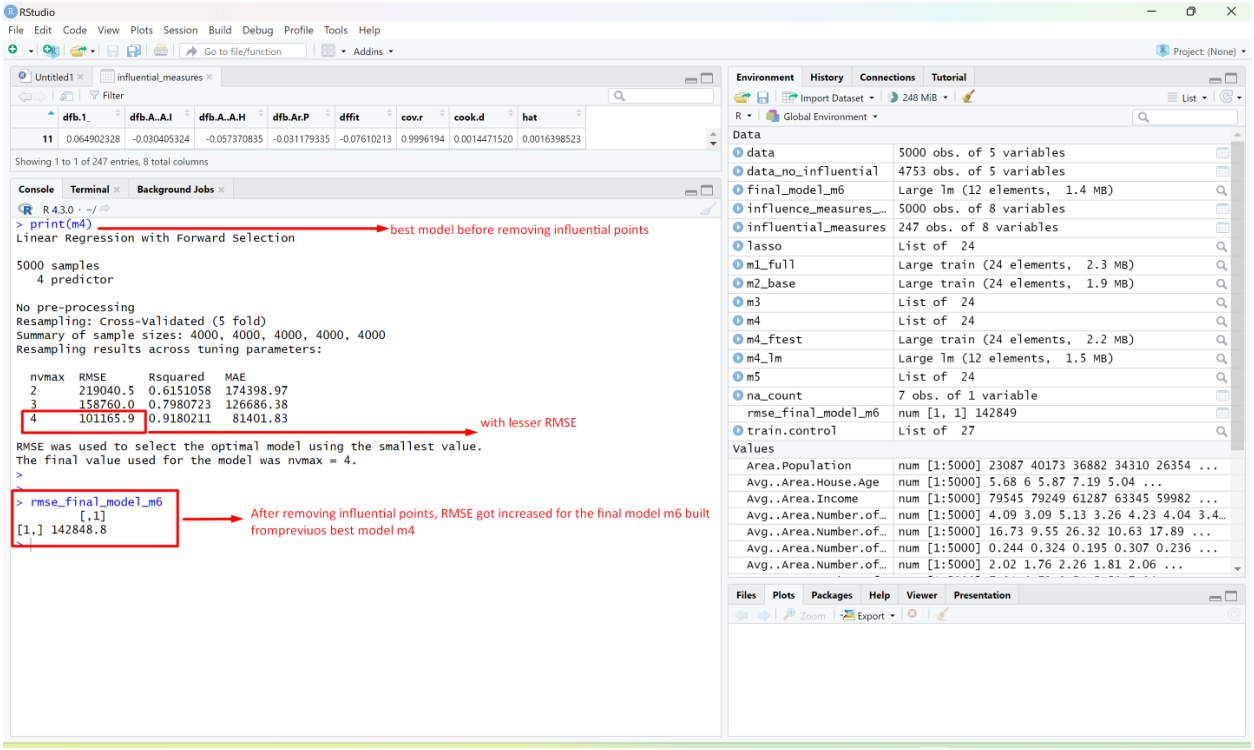
Provide the summary of your findings, explanations, conclusions.

Interpretation of the best model above: Write the equation for the final model:Explain the co-efficients and intercept:

Findings & explanations:

1. Found that there is 1 non-numeric variable which has 5000 unique values and removed.
2. Found there are no missing values and duplicates in the data
3.Outliers were found but didn't remove as we focus more on Influential points
4.Addressed non-linearity b/w one x-variable and y-variable and removed that x-variable (Avg..Area.Number.of.Bedrooms)
5. We built multiple linear regression models using different combinations of predictor variables to explore their impact on the response variable, "house price."
6. I found that model m4 is the best model with lesser RMSE **101165.9** compared to the other 5 models I built. So, I decided to pick M4 and move forward by performing model diagnosis (m4 passed model diagnosis) and seeking further improvements like Checked for VIF of the model m4 (passed VIF test). Checked for Influential points, found 247 influential points and removed them from the data.
7. 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.

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



Model	RMSE	ADJR2
M4 – with influential points	101165.9	79.79
Final model_m6 (no influential points) (same model built from m4)	142848.8	82.34

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. Here are some potential causes for this to happen:

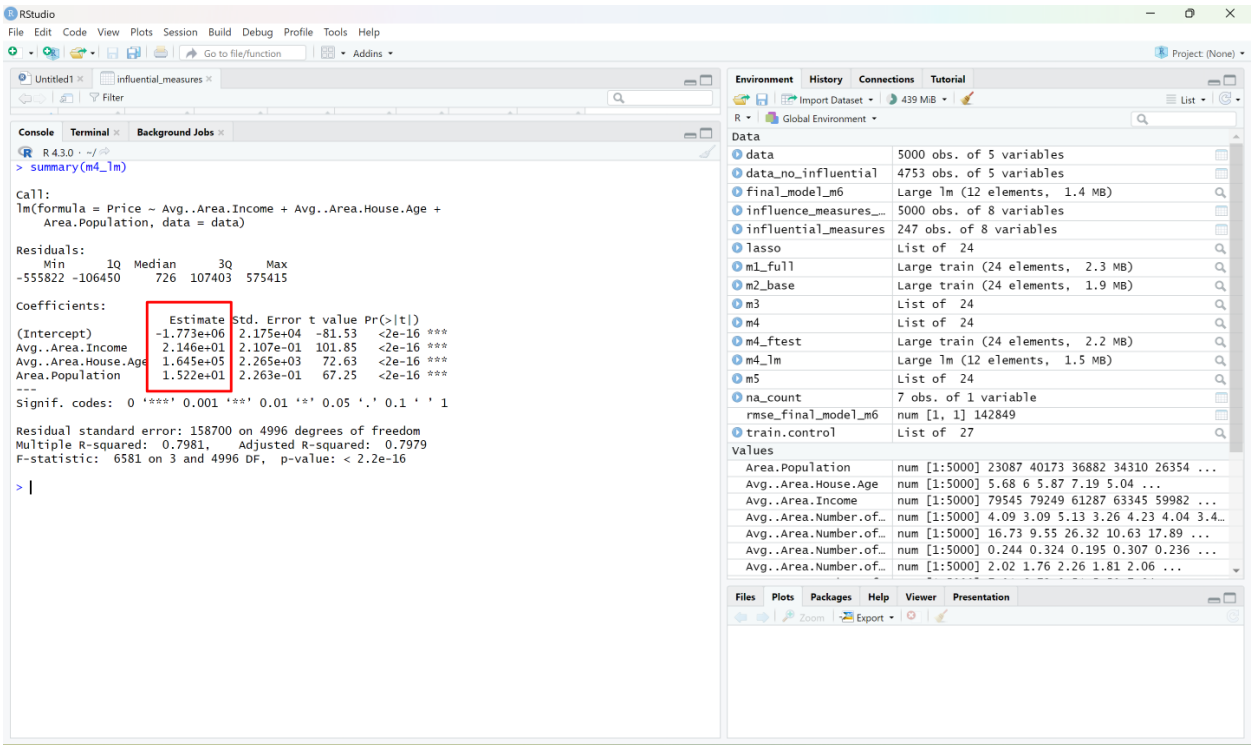
Data quality: The influential points may have indicated significant trends in the data or held essential information. Eliminating them might have resulted in the loss of crucial data, lowering the model's accuracy.

Overfitting: The influential points may have been outliers, but their existence helped to regularize the model and avoid overfitting. The model performed poorly on new, unforeseen data after you eliminated these points because it was more prone to overfitting to the remaining data.

Sample size: Removing the influential points may have resulted in a smaller sample size overall if they made up a sizable fraction of the dataset. Smaller datasets may result in less reliable models and higher RMSE values.

Conclusion and Interpretation of the best model from the finding above:

From the above findings and observations, I would like to pick model M4 which has least RMSE (with influential points) and interpret the model:



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

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

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

Explaining the intercept and slope:

$\beta 0$ ~ explaining the intercept:

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.

Explaining the slopes:

$\beta 1$ ~

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.

$\beta 2$ ~

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.

$\beta 3$ ~

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.

The coefficients (slopes) essentially quantify the change in the target variable (Price) associated with a one-unit change in each predictor variable while keeping all other predictors constant. For example, if the income (X1) increases by one unit and the age (X2) and population (X3) remain unchanged, the expected change in the Price is 2.146e+01 units. Similarly, if the age (X2) increases by one unit and the income (X1) and population (X3) are unchanged, the expected change in the Price is 1.645e+05 units.

6. Conclusions and Future Work

6.1. Conclusions

A short summary of your whole project and conclusions, such as what you want to, why you want to do so, which solutions you use, and which findings or final results you get finally.

Short summary of my whole Project:

What do I want to do?

I wanted to predict the housing price different states in the USA by building a n accurate predictive model using linear regression techniques.

Why do I want to do?

Because of the personal experience of not affording higher lease during covid and also to provide better understanding about the linear regression techniques, building predictive models and classification models and different ways to improve this models to get better model with accuracy and model performance.

Which solutions did I use?

- Collected housing price dataset from Kaggle.com (The URL is attached above) which consists of 5000 instances and 7 variables among which 1 variable is categorical

- checked for missing values and found none

- Handled non-numeric data (variable – Address) by removing it as it has 5000 unique values

-Performed descriptive statistics on the Price variable and income variable by knowing the mean, median, mode , variance and distribution of the variables

- Validated the claim/assumption (self -developed) on the Price variable by using hypothesis testing

-Handled non-linearity b/w Y and X—variables by Checked correlation values. I Used Pearson correlation in correlation analysis

-Found 1 X-variable Avg..Area.Number.of.Bedrooms ha no correlation with Y-variable Price and tried log, sqrt, power transformations. As I didn't see the improvement, Dropped that X-variable.

- Handled data split by N-fold crosse evaluation.

- Built 6 different multiple linear regression models by using different techniques like feature selection and lasso method to compare them

-performed model diagnosis for the best model m4 by f-test & residual analysis

-Chekced vif for the model m4

-identified influential points in model m4 & removed them

-Built a new model "final model m6" with no influential points.

Final result:

Rmse got increased after removing influential points. So decided to keep them and pick model m4 as best model (model with influential points)

Conclusion:

In the current study, on the USA housing dataset, it has been tried to fit different types of regression models which include Simple Linear Regression Model, multiple linear regression model, backward elimination model, forward elimination model, Stepwise model & lasso regression model. We are finding the M4 model(with influential points) as the most appropriate model giving the highest values of the R-square and the minimum Root Mean Square Error Value. Hence, it can be concluded that the M4 model is the most appropriate model for this dataset and should be used for predicting housing prices.

In conclusion, this linear regression model aims to predict the Price based on the values of the Avg..Area.Income, Avg..Area.House.Age, and Area.Population variables. The coefficients of the

predictor variables indicate the strength and direction of their relationships with the Price, while the intercept represents the Price value when all predictor variables are zero.

I hope that my research has contributed not only to the
advancement of property evaluation methodology and empirical findings, but also to the presentation of an alternative approach to house value valuation.

6.2. Limitations

Introduce the limitations of your work.

However, we must remember that these linear regression tools and procedures have their own drawbacks. Careful feature selection is necessary since there are typically several feasible characteristics from which researchers might choose and incorporate in models. In conclusion, MLR is typically the poorest model since y and x typically have non-linear relationships. MLR often performs substantially worse than predictive models that can better capture non-linear relationships.

Here, we've employed regression models like lasso regression, feature selection, and simple linear multiple linear regression. The outcomes, however, may have been improved if we had focused on the K closest neighbor regression or the Regression Tree. 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 also relevant factors. To determine if the pattern is present similarly in other data or not, this must be checked again. It was impossible for us to double-check this with another dataset because we could only work with one at a time.

6.3. Potential Improvements or Future Work

Introduce and discuss possible methods to improve or extend your work in the future.

Improvements/Future work:

Several new areas of interest have been ignited by the results of our study and analysis, which we will now describe. We will offer some recommendations in this part for how the results reported in this thesis may be elaborated upon in subsequent research as well as for how the body of literature on price prediction might be widened.

1. collect new variables:
future research may involve collecting new variables to improve the model prediction such as “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.:
For future work may involve improving the , the correlation b/w the y variable and x-variable by applying different transformations.In our data, Avg area no.of bedrooms has no no linear relationship with Price. So future work may focus on this point to get better predictive power to the model.

3. Normalizing/scaling the features:

Scaling or normalizing is used to normalize the range of independent variables or features of data. By scaling/normalizing the independent variables, we ensure that each feature contributes approximately proportionately to the final distance. Among other reasons like being a pre-requisite step for PCA (principal component analysis), it also makes gradient descent converge much faster compared to un-scaled features.

4. handling outliers:

Future work might also focus on the outliers in the data. If the outliers has negative effect on the model, They can be removed.

5. Influential points:

In this thesis the influential points are not removed because the rmse of the model decreased with no influential points. So we need to cross check this similar pattern with other dataset and see if the other data set also had the same issue.

6. Trying different advance regression techniques: Based on the project's findings, we can offer recommendations for further improvements in the model, such as trying different regression techniques (e.g., polynomial regression, ridge regression, or knn regression, special neural network model, Regression Tree) or incorporating additional relevant features to enhance the prediction accuracy.