



Statistical Analysis on House Price

Author: Duy Le Duc

Student Id: bs21don032

BDS21 Batch

TABLE OF CONTENTS

1.INTRODUCTION	3
Purpose of House Price Prediction	3
Purpose of the study	3
Data set	3
2. DATA UNDERSTANDING	4
Sample data	4
Key statistics	4
Histogram	5
Boxplots	6
Correlation scatterplot and correlation matrix	7
3. TRAIN AND SPLIT DATA	9
4. REGRESSSION ANALYSIS	10
Regression model 1	10
Regression model 2	12
Model Validation	13
Hypothesis testing	16
Prediction for test data	16
5. CONCLUSION	17
6. REFERENCE	18

1. INTRODUCTION

a. Purpose of house price prediction:

House prices vary greatly depend on many factors: location, number of rooms, ... House prices prediction can help the owner to determine the selling price of a house and assist customer in finding the right house to purchase.

b. Purpose of this study:

This study focuses on predicting house prices based on the attributes of houses and people in the neighborhood such as average house age, area population, ... This study also focuses on how each of those attributes affect the price of the predicted house.

c. Dataset:

We have a data set containing of house prices based on test results. The data consists of the following features:

1. Average Area Income
2. Average Area House Age
3. Average Area Number of Rooms
4. Average Area Number of Bedrooms
5. Area Population
6. Price
7. Address

Price is the dependent variable in the data set, whereas others are independent variables.

The dataset contains 499 unique rows, each representing a house with its attributes.

2. DATA UNDERSTANDING

a. Sample Data:

The sample data (first 6 rows) is shown below:

```
> head(mydata)
# A tibble: 6 × 7
  `Avg. Area Income` `Avg. Area House Age` `Avg. Area Number of Rooms` `Avg. Area Number of Bedrooms` `Area Population` Price Address
      <dbl>          <dbl>          <dbl>          <dbl>          <dbl>    <dbl> <dbl> <chr>
1      79545.         5.68            7.01            4.09        23087. 1059034. "208 Michael Ferry Apt. 674\nLa...
2      79249.         6.00            6.73            3.09        40173. 1505891. "188 Johnson Views Suite 079\nL...
3      61287.         5.87            8.51            5.13        36882. 1058988. "9127 Elizabeth Stravenue\nDani...
4      63345.         7.19            5.59            3.26        34310. 1260617. "USS Barnett\nFPO AP 44820"
5      59982.         5.04            7.84            4.23        26354. 630943. "USNS Raymond\nFPO AE 09386"
6      80176.         4.99            6.10            4.04        26748. 1068138. "06039 Jennifer Islands Apt. 44...
```

With respect to the NOIR classification (Categorical type [Nominal, Ordinal], Continuous type [Ratio, Interval]), the data in data set can be classified into interval data of Continuous type, except for 'Address' column, which falls into Nominal data of Categorical type. We have conducted test to search for Null values. No NULL value was found in the dataset.

b. Key statistics of the data:

Next is some of the key statistics and observations from the attributes of the dataset:

```
> describe(mydata)                                     #Basic info about the data
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Avg. Area Income	1	499	68133.07	10349.44	68494.98	68209.80	10507.89	17796.63	97112.36	79315.73	-0.25	0.78	463.30
Avg. Area House Age	2	499	6.05	0.99	6.03	6.05	1.05	3.69	8.56	4.87	-0.01	-0.53	0.04
Avg. Area Number of Rooms	3	499	6.97	1.05	7.04	6.98	1.05	3.24	9.71	6.47	-0.11	-0.08	0.05
Avg. Area Number of Bedrooms	4	499	3.99	1.23	4.07	3.94	1.33	2.00	6.50	4.50	0.35	-0.71	0.06
Area Population	5	499	35718.23	9901.83	35799.64	35704.72	9502.61	172.61	69575.45	69402.84	0.04	0.42	443.27
Price	6	499	1226420.02	346732.00	1228810.75	1228910.60	325820.77	152071.87	2469065.59	2316993.72	-0.02	0.41	15521.86
Address*	7	499	250.00	144.19	250.00	250.00	185.32	1.00	499.00	498.00	0.00	-1.21	6.45

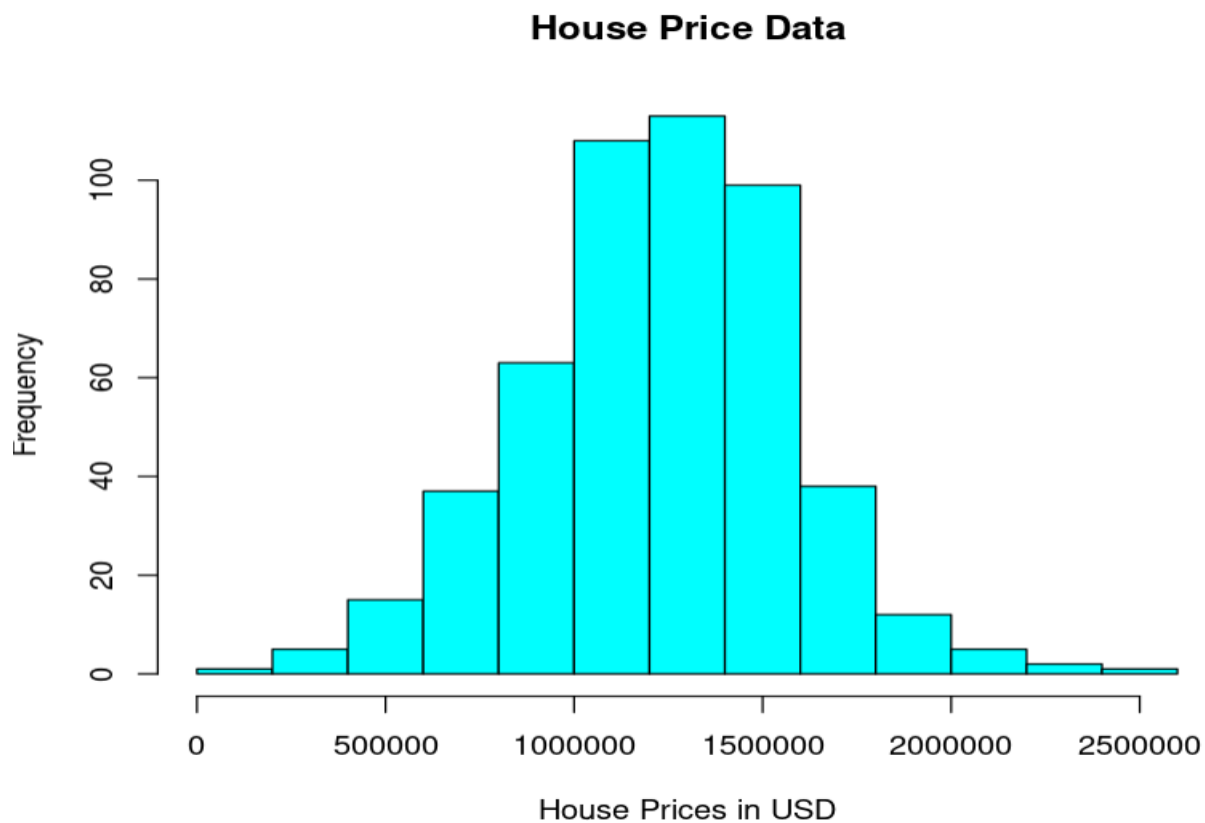
2. DATA UNDERSTANDING

From the table above, the following statement can be inferred:

The range of different features varies greatly. For instance, with Average Area Number of Rooms, Average Area House Age and Average Area Number of Bedrooms, ranges are only from 4.5 to 6.47. Whereas for Price, the range is 2316993.72. Mean of Price is 1226420.02, and it is less than median, which is 1228810.75, indicating a negative skewness of -0.02. We can also drop column Address since it has 499 unique categorical values

c. Histogram:

Next is the histogram of our dependent variable - Price - to understand its distribution.

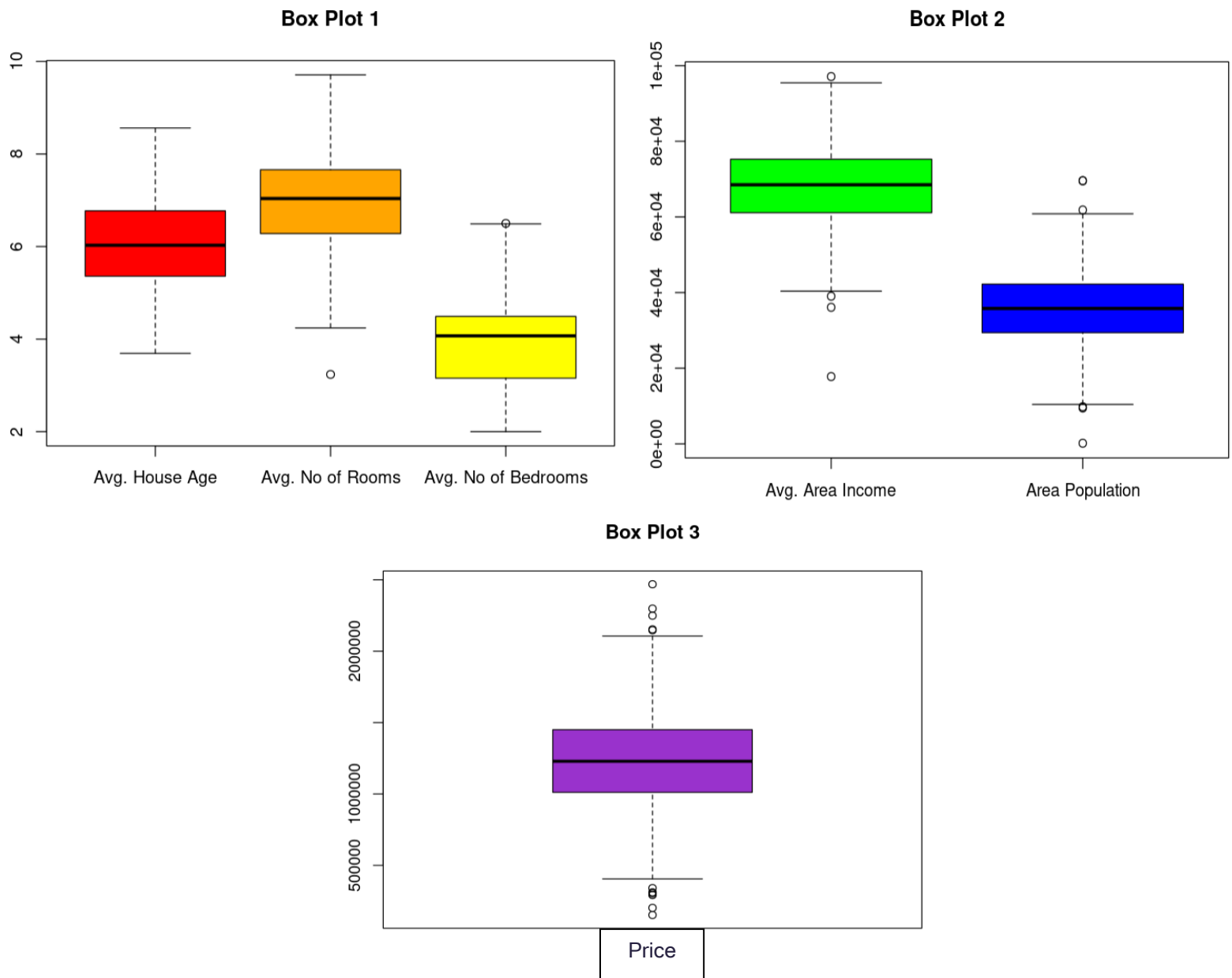


As we can see from the histogram above, it follows normal distribution with slightly negative skewness

2. DATA UNDERSTANDING

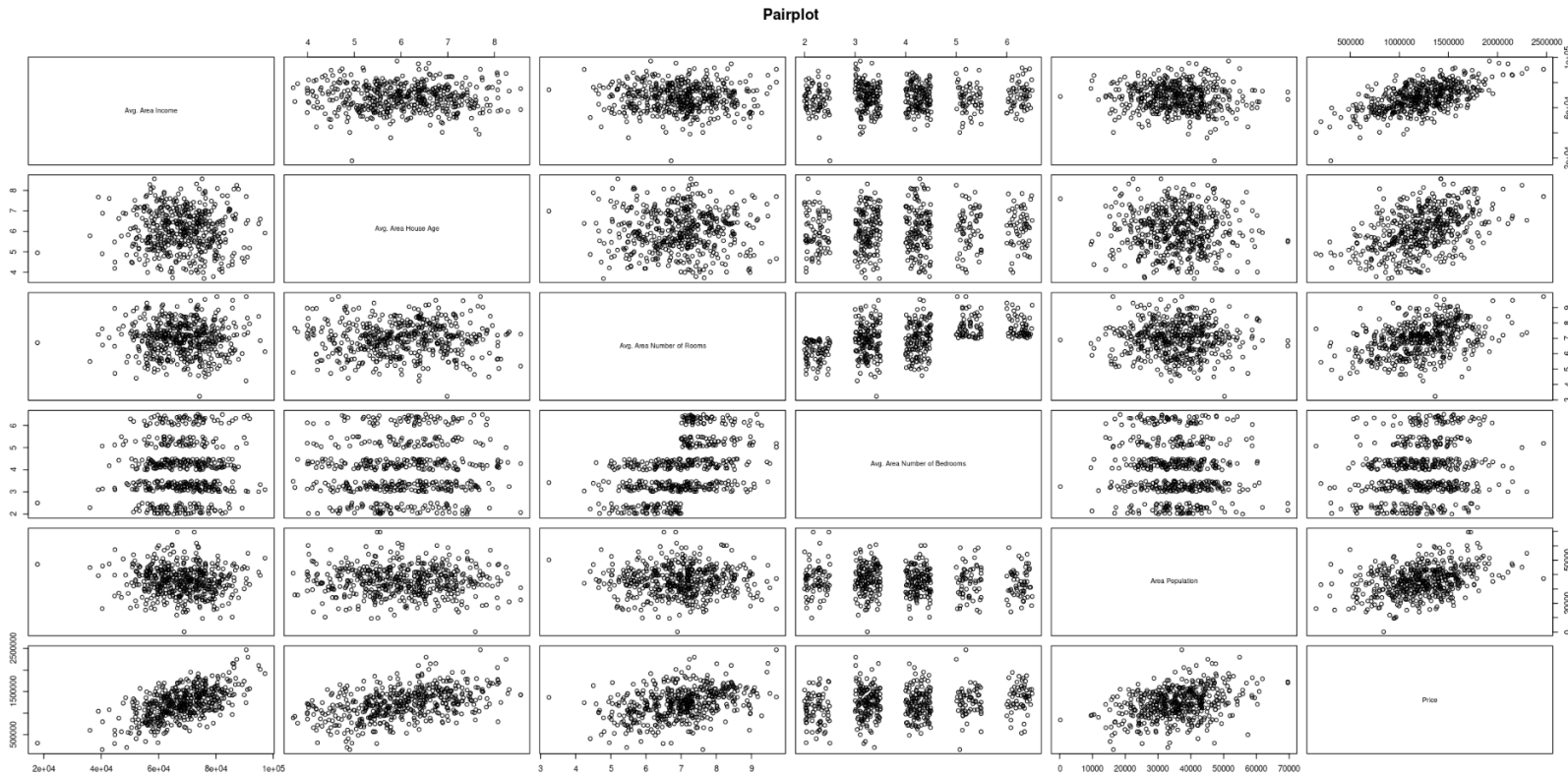
d. Boxplots:

The boxplots below show distributions of each attribute and their outliers. We divided into 3 different boxplots to make it easier for analyzing since the range of each attribute varies immensely.



As we can observed from the above figures, there are few outliers present in the dataset. After examining the dataset, it was found that there's some correlation between the output and the outlier values. Moreover, as we have observed in earlier section, the skewness of Price is only -0.02. Therefore, we didn't remove any outliers and proceed with the old dataset.

e. Correlation Scatterplots and Correlation Matrix:



	Avg. Area Income	Avg. Area House Age	Avg. Area Number of Rooms	Avg. Area Number of Bedrooms	Area Population	Price
Avg. Area Income	1	-0.016885602	-0.000202535	0.040777824	-0.082362821	0.604051053
Avg. Area House Age	-0.016885602	1	0.02550368	0.05901759	-0.022644997	0.46552784
Avg. Area Number of Rooms	-0.000202535	0.02550368	1	0.466878941	-0.003109368	0.384991043
Avg. Area Number of Bedrooms	0.040777824	0.05901759	0.466878941	1	-0.08251301	0.17981334
Area Population	-0.082362821	-0.022644997	-0.003109368	-0.08251301	1	0.364618715
Price	0.604051053	0.46552784	0.384991043	0.17981334	0.364618715	1

2. DATA UNDERSTANDING

As we can observe from the correlation matrix and correlation plot above, there is a strong linear relationship between the first independent variable and the dependent variable (Price): 0.60405. Whereas, others independent variables are somewhat less than 0.5 (the mark for linear correlation): 0.4655, 0.3849, 0.3646 for 'Avg. Area House Age', 'Avg. Area Number of Rooms', 'Area Population' respectively. It's important to note that it's only 0.1798 for 'Avg. Area Number of Bedrooms'. This shows that there's a weak linear relationship between the independent variable and dependent variable.







Next is the information about relationship between independent variables. As we can see from the plot and the correlation matrix, there is no relationship between almost all of variables since the correlation score are extremely low. All of them are <0.1 except for correlation score between 'Avg. Area Number of Rooms' and 'Avg. Area Number of Bedrooms', which is 0.4668. Since this value is exceptionally high, we might speculate a linear relationship between these two variables.

3. TRAIN AND TEST SPLIT

In this step, data is split into training and testing data with the following code:

```
#Train and Test Split
install.packages('caTools')
library('caTools')
set.seed(123)
split <- sample.split(mydata_new$Price, SplitRatio = 0.70)
train_data <- subset(mydata, split == T)
test_data <- subset(mydata, split == F)
```

The dataset is split into 70-30 ratio, where 70% of the data is for training and 30% of the data is for testing. As we can see from the figure above, 499 objects of mydata_new is split into 349 objects for train_data and 150 objects for test_data.

 mydata_new	499 obs. of 6 variables	
 train_data	349 obs. of 7 variables	
 test_data	150 obs. of 7 variables	

4. REGRESSION ANALYSIS

Now that the model is split into training and testing data, we will use training data to create model 1. We will create a multiple linear regression model that takes the form:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{k-1} x_{k-1} + \varepsilon \dots$$

Where Y is the dependent variable, x_1, x_2, \dots are independent variables, ε is the error, β_0 is the intercept, β_1, β_2, \dots are the measures of total effect of the predictor variables x_1, x_2, \dots respectively.

For hypothesis testing and the settings of confidence limits, we also assume that ε is normally distributed.

X variables x_1, x_2, \dots for the model are as follows:

- Average Area Income
- Average Area House Age
- Average Area Number of Rooms
- Average Area Number of Bedrooms
- Area Population

Y variable for the model is:

- Price

a. Regression model 1:

First, we create the model by the following code:

```
> #Create Model1
> model1 <- lm(Price~Avg. Area Income'+ `Avg. Area House Age` + `Avg. Area Number of Rooms` + `Avg. Area Number of Bedrooms` + `Area Population`, data = train_data)
> summary(model1)
```

Call:

```
lm(formula = Price ~ `Avg. Area Income` + `Avg. Area House Age` +
  `Avg. Area Number of Rooms` + `Avg. Area Number of Bedrooms` +
  `Area Population`, data = train_data)
```

4. REGRESSION ANALYSIS

Output of model 1:

Statistic	Value	Criteria
Residual standard error	104700	
Multiple R - squared	0.9123	> 0.6
Adjusted R - squared	0.911	> 0.6

Model	df	F	P - value
Regression	5	713.2	< 2.2e-16
Residual	343		
Total	348		

Criteria:

P-value < 0.05 of the above F test indicates that Model-1 is good for predicting the result - Price of the house.

Regression Output Coefficients and P-value:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-2.671e+06	6.648e+04	-40.179	<2e-16	***
`Avg. Area Income`	2.142e+01	5.429e-01	39.454	<2e-16	***
`Avg. Area House Age`	1.659e+05	5.518e+03	30.072	<2e-16	***
`Avg. Area Number of Rooms`	1.333e+05	6.157e+03	21.651	<2e-16	***
`Avg. Area Number of Bedrooms`	-7.091e+03	5.130e+03	-1.382	0.168	
`Area Population`	1.504e+01	5.823e-01	25.833	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The regression output above illustrates that for attribute 'Avg. Area Number of Bedrooms', the P-value is higher than 0.05 (0.168). Therefore, this parameter has less or no impact on the result. Therefore, we will create a new model after removing the attribute 'Avg. Area Number of Bedrooms'.

4. REGRESSION ANALYSIS

b. Regression model 2:

First, we again create model 2 by the following code:

```
> #Create Model2
> model2 <- lm(Price~`Avg. Area Income`+ `Avg. Area House Age`+ `Avg. Area Number of Rooms`+ `Area Population`, data = train_data)
> summary(model2)
```

Call:
lm(formula = Price ~ `Avg. Area Income` + `Avg. Area House Age` +
`Avg. Area Number of Rooms` + `Area Population`, data = train_data)

Second is the output of model 2:

Statistic	Value	Criteria
Residual standard error	104400	
Multiple R - squared	0.918	> 0.6
Adjusted R – squared	0.9127	> 0.6

Model	df	F	P - value
Regression	4	888.7	< 2.2e-16
Residual	344		
Total	348		

Adjusted R-squared and Multiple R-squared both improve by a decent amount (1% and 0.6% respectively)

Criteria:

P-value < 0.05 of the above F test indicates that Model-1 is good for predicting the result - Price of the house.

4. REGRESSION ANALYSIS

Regression Output Coefficients and P-value:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-2.669e+06	6.655e+04	-40.10	<2e-16	***
`Avg. Area Income`	2.137e+01	5.422e-01	39.41	<2e-16	***
`Avg. Area House Age`	1.658e+05	5.524e+03	30.01	<2e-16	***
`Avg. Area Number of Rooms`	1.292e+05	5.398e+03	23.93	<2e-16	***
`Area Population`	1.512e+01	5.801e-01	26.07	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Model equation is:

Y (price) = - 2.669e+06 + 2.137e+01 * 'Avg. Area Income' + 1.658e+05 * 'Avg. Area House Age' + 1.292e+05 * 'Avg. Area Number of Rooms' + 1.512e+01 * 'Area Population'

c. Model validation:

In this step, we will perform VIF (Variance Inflation Factor) test and Step AIC to analyze whether the model is optimum.

1. Variance Inflation Factor:

Variance Inflation Factor is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

Formula:
$$VIF_i = \frac{1}{1 - R_i^2}$$

Where R_i^2 is the coefficient of determination of the regression equation for regressing X_i on other X 's. Criteria: $VIF > 5$ is an indication of multicollinearity. Solution for multicollinearity: Remove one or more highly correlated independent variables.

Result:

```
> vif(model1)
      `Avg. Area Income`      `Avg. Area House Age`
           1.012144                1.007388
`Avg. Area Number of Rooms` `Avg. Area Number of Bedrooms`
           1.309988                1.328635
      `Area Population`
           1.017276

>
> vif(model2)
      `Avg. Area Income`      `Avg. Area House Age`      `Avg. Area Number of Rooms`
           1.006745                1.006811                1.004276
      `Area Population`
           1.006881
```

As we can observe from the output of VIF test, there's no value that have $VIF > 5$. This clearly shows that there's no multicollinearity between independent variables. However, we can notice that in model 1, VIF score for 'Avg. Area Number of Rooms' and 'Avg. Area Number of Bedrooms' are quite high 1.30 and 1.32 respectively, compared to other values (1.02, 1.007, 1.01)

Therefore, let's validate it further by step AIC:

2. Step AIC:

Step AIC is one of the most frequently used search method for feature selection.

Formula: $AIC = 2k - 2\ln(\hat{L})$

Where k is the number of model parameter and L is the maximum value of the likelihood function of the model.

Step AIC is performed on model 1 (using all input parameters):

```
Step: AIC=8073.83
Price ~ `Avg. Area Income` + `Avg. Area House Age` + `Avg. Area Number of Rooms` +
`Area Population`
```

	Df	Sum of Sq	RSS	AIC
<none>			3.7795e+12	8073.8
+ `Avg. Area Number of Bedrooms`	1	2.0932e+10	3.7585e+12	8073.9
- `Avg. Area Number of Rooms`	1	6.2932e+12	1.0073e+13	8413.9
- `Area Population`	1	7.4685e+12	1.1248e+13	8452.5
- `Avg. Area House Age`	1	9.8936e+12	1.3673e+13	8520.6
- `Avg. Area Income`	1	1.7061e+13	2.0840e+13	8667.7

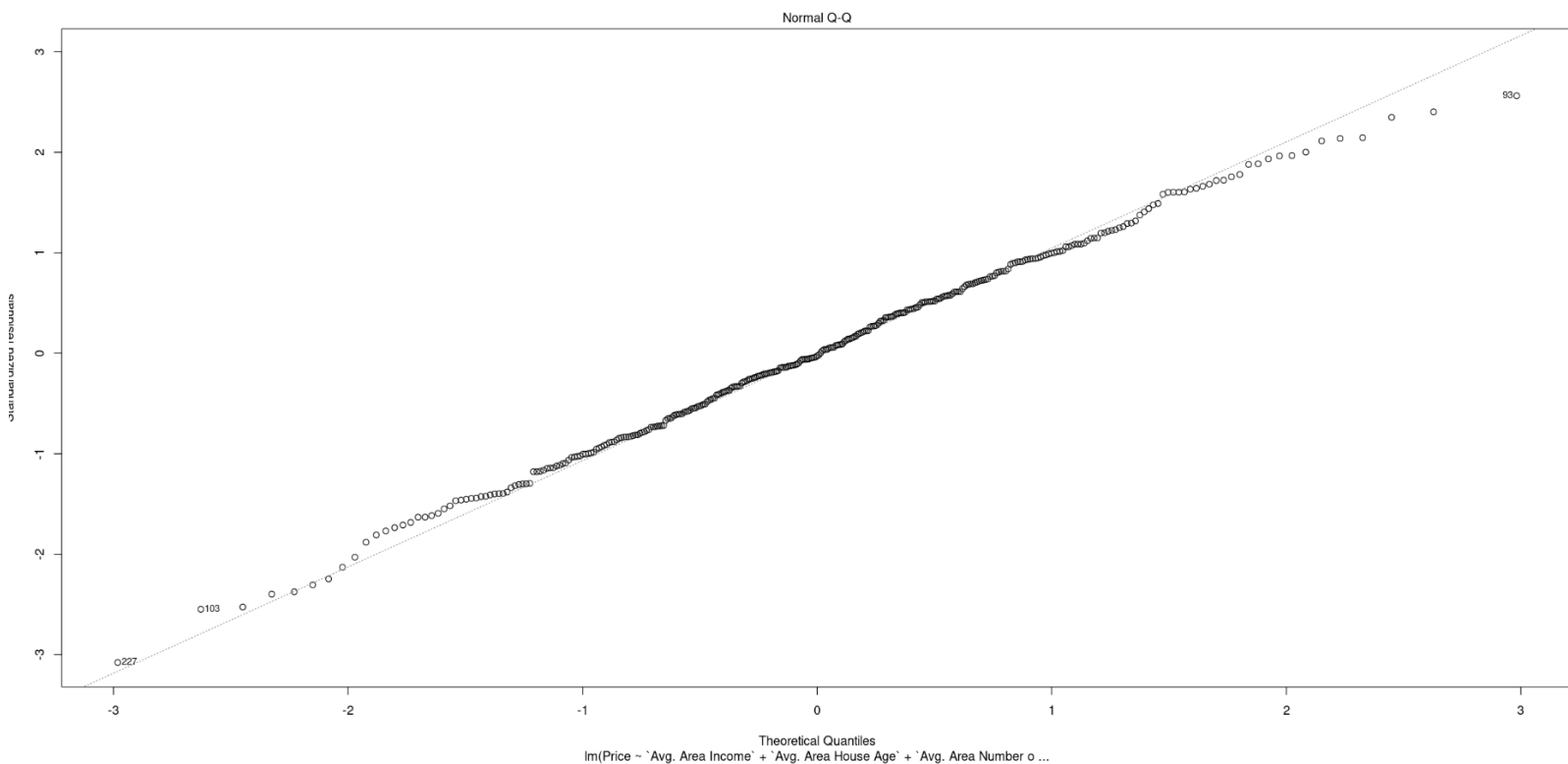
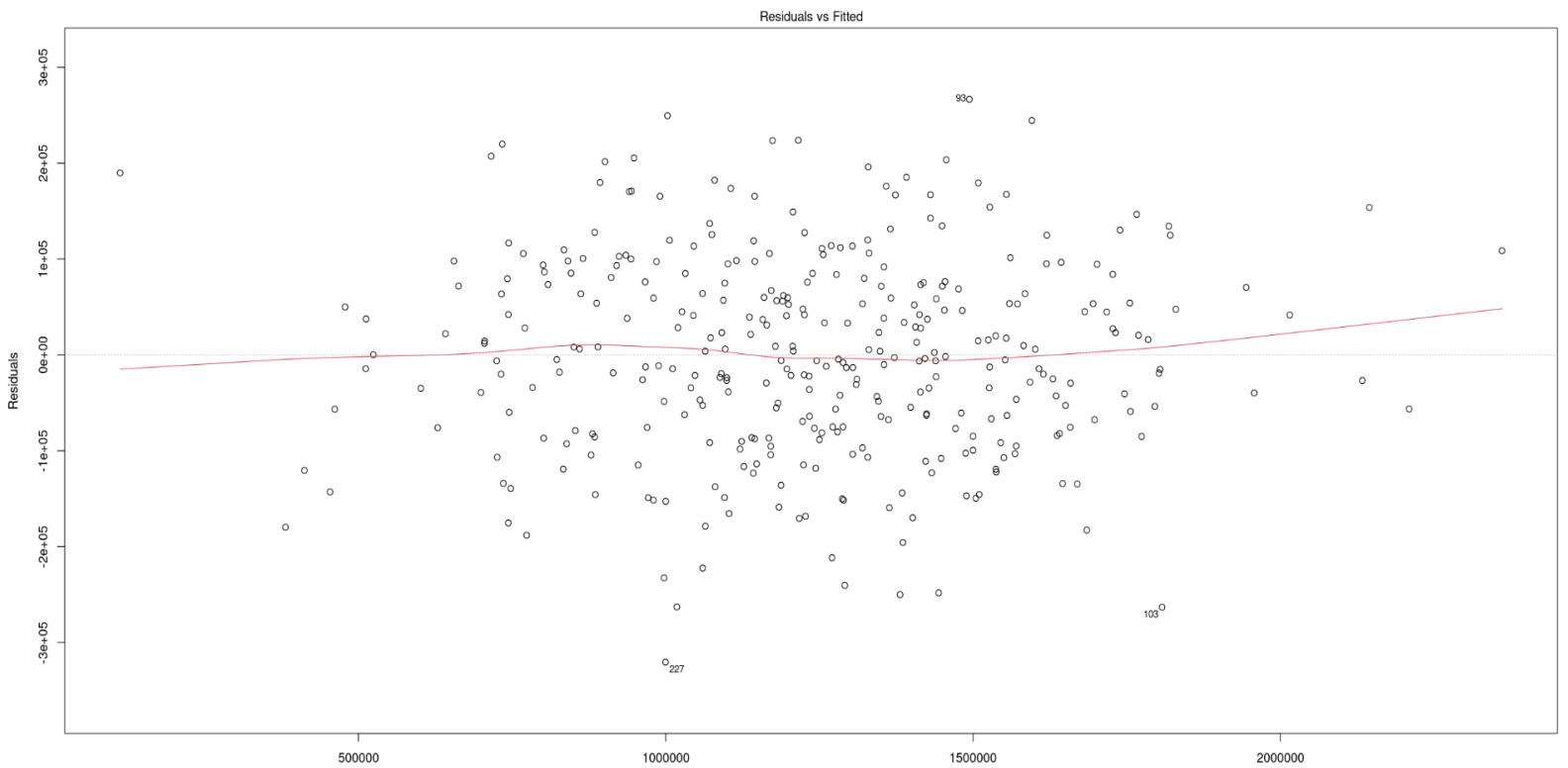
```
Call:
lm(formula = Price ~ `Avg. Area Income` + `Avg. Area House Age` +
`Avg. Area Number of Rooms` + `Area Population`, data = train_data)
```

Coefficients:

(Intercept)	`Avg. Area Income`	`Avg. Area House Age`
-2.669e+06	2.137e+01	1.658e+05
`Avg. Area Number of Rooms`	`Area Population`	
1.292e+05	1.512e+01	

From the result above, it can be seen that model 2 is the result of step AIC. We have already deleted attribute 'Avg. Area Number of Bedrooms'.

c. Model validation:



The residual vs fitted plot shows that the data points 'bounce randomly' around the residual line. No data points really stand out from the random pattern. Along with the residual vs fitted plot, the QQ plot also suggest that the data is linearly distributed.

d. Hypothesis Testing on model 2:

$H_0: B_1 = B_2 = \dots = B_{k-1} = 0$

$H_1: B_i \neq 0$, for at least one i .

ANOVA output:

```
> summary(res.aov)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
`Avg. Area Income`	1	1.538e+13	1.538e+13	1400.3	<2e-16	***
`Avg. Area House Age`	1	1.005e+13	1.005e+13	914.6	<2e-16	***
`Avg. Area Number of Rooms`	1	6.153e+12	6.153e+12	560.0	<2e-16	***
`Area Population`	1	7.469e+12	7.469e+12	679.8	<2e-16	***
Residuals	344	3.779e+12	1.099e+10			

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can observed from the result above, all the p-values are significant to the output.

Therefore, we can reject the Null Hypothesis. Therefore, the model can be used for prediction.

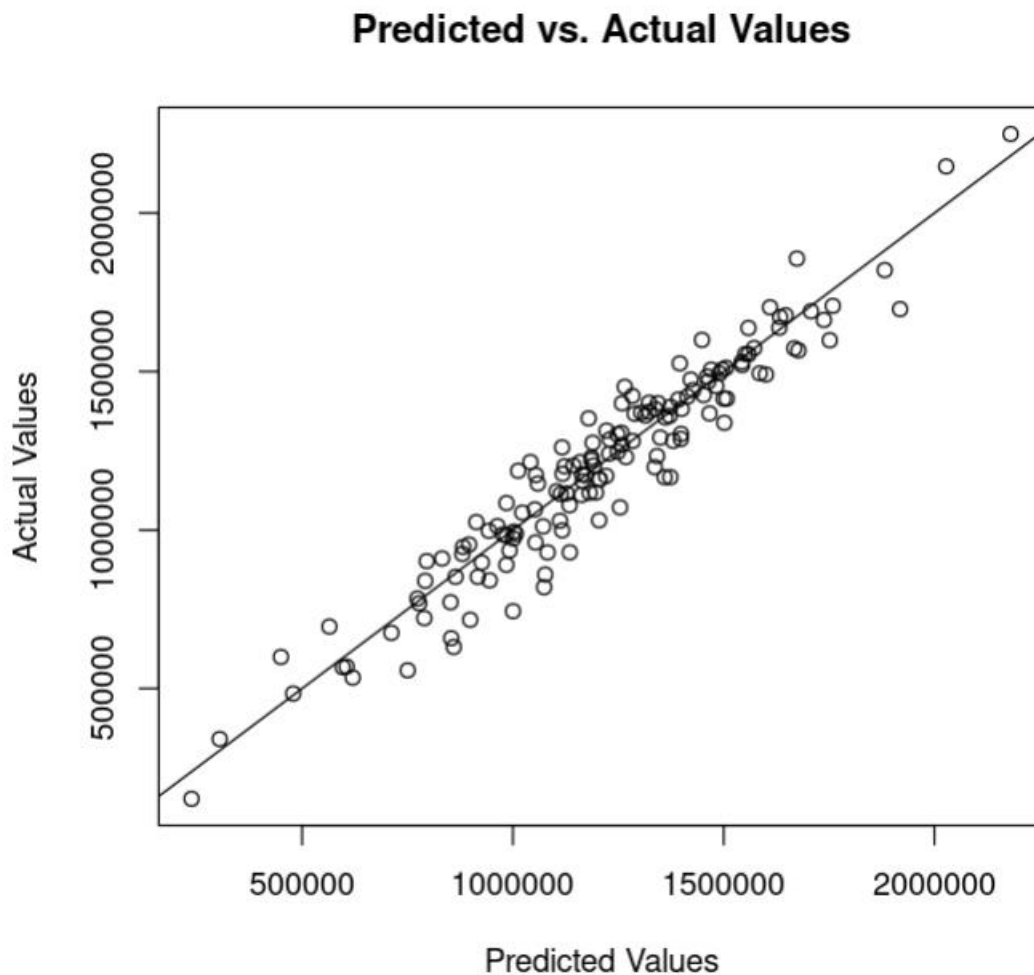
e. Prediction for test data:

The dependent variable (Price) was predicted for test data. The results of the prediction (some error metrics) are as follow:

```
> accuracy(p,test_data$Price)
```

	ME	RMSE	MAE	MPE	MAPE
Test set	-12040	96699.61	74330.63	-2.101168	7.336536

Predicted vs Actual Values Plot:



5. Conclusion:

Model 2 explain price by 4 independent variables: 'Avg. Area Income', 'Avg. Area House Age', 'Avg. Area Number of Rooms', 'Area Population' with an R-squared of 0.9127, better than model 1 - 0.911 - where there's an additional variable 'Avg. Number of Bed Rooms'. This clearly suggest that the speculation that there's some linear relationship between 'Avg. Area Number of Rooms' and 'Avg. Area Number of Bed Rooms' are correct. This has been proved by step AIC. Looking at the plots above, we can conclude that 4 variables in model 2 has a visible linear relationship.

6. REFERENCE

https://en.wikipedia.org/wiki/Variance_inflation_factor

[https://thesai.org/Downloads/Volume8No10/Paper_42-](https://thesai.org/Downloads/Volume8No10/Paper_42-Modeling_House_Price_Prediction_using_Linear_Regression.pdf)

[Modeling_House_Price_Prediction_using_Linear_Regression.pdf](https://thesai.org/Downloads/Volume8No10/Paper_42-Modeling_House_Price_Prediction_using_Linear_Regression.pdf)

<https://ashutoshr.medium.com/what-is-stepaic-in-r-a65b71c9eeba>

Dataset:

<https://www.kaggle.com/code/kaiyungtan/usa-housing-linear-models/data>