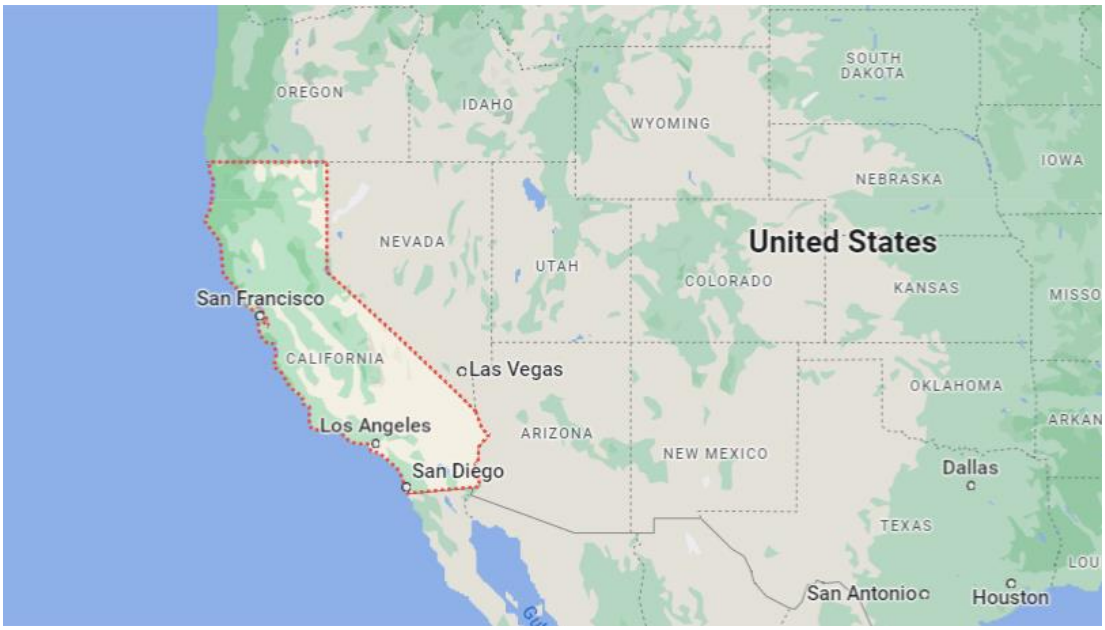


California Housing Price Prediction Using Multiple Linear Regression

DATA 603 project report
by

Farhana Kausar Shaik
Ashique Jaman
Ali Haghighat

Contents

1. Introduction

1.1 Context

Housing price prediction is a significant problem in the real estate industry. The ability to predict the price of a house accurately is crucial for buyers, sellers, and real estate agents for several reasons.

1. **Real estate investment**: Many people invest in real estate as a means of generating wealth. By predicting housing prices, investors can make informed decisions about which properties to buy and when to sell them to maximize their returns.
2. **Home buying and selling**: For individuals looking to buy or sell a home, knowing the likely selling price can help them make better decisions about the timing of their purchase or sale, and also give them a better idea of what they can afford.
3. **Economic forecasting**: Housing prices are closely linked to the overall health of the economy, so predicting housing prices can provide valuable insights into broader economic trends and help policymakers make informed decisions.
4. **Risk management**: Lenders and other financial institutions need to manage their risk exposure when lending money for real estate purchases. Predicting housing prices can help them assess the likelihood of loan defaults and make more informed lending decisions.

Overall, accurate housing price predictions can help individuals and organizations make better-informed decisions and manage risk, which can have important economic and social implications.

1.2 Objective

Under the scope of this project, we are going to apply different multiple linear regression techniques and also its assumptions to analyze historical selling price and property features like location, number of rooms, age, ocean proximity etc. of Californian Housing sells data to create predictive models that can forecast the price of a specific house in a given location. These valuable insights can be used for decision-making in the real estate industry. We are also going to evaluate the model's effectiveness in accurately predicting the housing price with our test data.

2. Methodology

2.1 Data

There are 10 columns and 20640 records in the tabular data out of which 207 rows contain missing values. This data set is licensed under the Apache2.0 open-source license. It is retrieved from the Kaggle website, and the column names and the definitions are summarized below:

1. **longitude**: A measure of how far west a house is; a higher value is farther west
2. **latitude**: A measure of how far north a house is; a higher value is farther north
3. **housing_median_age**: Median age of a house within a block; a lower number is a newer building
4. **total_rooms**: Total number of rooms within a block
5. **total_bedrooms**: Total number of bedrooms within a block
6. **population**: Total number of people residing within a block
7. **households**: Total number of households, a group of people residing within a home unit, for a block
8. **median_income**: Median income for households within a block of houses (measured in tens of thousands of US Dollars)
9. **median_house_value**: Median house value for households within a block (measured in US Dollars)
10. **ocean_proximity**: Location of the house w.r.t ocean/sea. This column includes four different categories each states the the proximity to the ocean. These four categories are: <1H ocean, INLAND, NEAR BAY, NEAR OCCEAN.
    a. **<1H ocean:** Less than one hour distance from ocean
    b. **INLAND:** Far from ocean
    c. **NEAR BAY:** Near to the bay
    d. **NEAR OCCEAN:** Near to the ocean

In census data collection, a block is the smallest geographic unit for which population and demographic information is collected. The U.S. Census Bureau defines a block as "the smallest geographic unit for which the Census Bureau tabulates decennial census data" and notes that blocks are typically bounded by visible features such as streets, streams, and railroad tracks.

So, the total number of people residing within a block refers to the population count of a specific geographic area bounded by visible features such as streets, roads, streams, and railroad tracks.

2.2 Approach

Our dataset has enormous data and the major problem for this would be overfitting. This is the reason why our dataset has been splitted into train and test datasets and assessed. Then, after processing our data, we have found out that our model needs to be improved. For this, we have applied log transformation and box cox transformation to our model. We checked the normality and homoscedasticity for each of these transformations, but the result was not what we expected, both

normality and homoscedasticity were not met. The last step remaining was to compare the RMSE for the scaled data and transformed data between training data and testing data. The result was, the RMSE for the testing data is slightly higher than the training data in both the cases.

2.3 Workflow

1. The dataset has been splitted into train and test data.
2. Performed hypothesis test with F-statistic (anova).
3. Performed Individual t-test.
4. Applied backward - elimination procedure.
5. Checked for AIC, BIC, CP, r-squared and adjusted r-squared values and found model 9 has the least AIC, BIC, and CP values and the highest r-squared ad adjusted r-squared values.
6. VIF test for multicollinearity. Found four predictors (population, households, total rooms and total bedrooms) with multicollinearity = 1 and removed households and total bedrooms as they give similar information to population and total rooms respectively. Found two significant predictors (latitude and longitude) even though multicollinearity is present and retained them.
7. Combined them using PCA. Performed VIF test again and found no multicollinearity.
8. Considered higher order model.
9. Considered interaction terms.
10. Combined higher order model and model with interaction terms.
11. Checked for homoscedasticity and normality. Both did not hold true.
12. Checked for outliers using cook's distance and removed them.
13. Applied log transformation and box-cox transformation.
14. Checked for linear assumption, homoscedasticity and normality assumptions, but they were not met.
15. Calculated RMSE for scaled training data and scaled testing data and compared them.
16. Calculated RMSE for final training data and final testing data and compared them.

It was particularly hard when VIF for latitude and longitude came out to be the same. We overcame this by combining both the variables using PCA. The model was improved in terms of adj R squared and RMSE, although the linear model assumptions especially normality and homoscedasticity were not met.

2.4 Workload Distribution

The group members shared the workload equally, with each member contributing an equitable amount of effort to complete the tasks.

3. Main Results of the Analysis

3.1 Finding the best first order model

The first step is to find the best model in doing the single t-test and the ANOVA table for the model. The results are summarized in the tables below:

```
Analysis of Variance Table

Model 1: median_house_value ~ 1
Model 2: median_house_value ~ longitude + latitude + housing_median_age +
    total_rooms + total_bedrooms + population + households +
    median_income + factor(ocean_proximity)
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1  16283 2.1671e+14
2  16272 7.7017e+13 11 1.3969e+14 2683.1 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 1: The ANOVA table of the full linear model

The above output shows that Fcal=2683.1 with df= 11 (p-value< 2.2e-16 < = 0.05), indicating that we should clearly reject the null hypothesis. It provides compelling evidence against the null hypothesis H0. In other words, the large F-test suggests that at least one of the independent variables in the data set must be related to the median price of houses. The next table will show independent t-tests to see which predictor is significant and which one should be dropped from the model.

```
Call:
lm(formula = median_house_value ~ longitude + latitude + housing_median_age +
    total_rooms + total_bedrooms + population + households +
    median_income + factor(ocean_proximity), data = traindf)

Residuals:
    Min      1Q  Median      3Q     Max
-554991  -42812  -10585   28695  743753

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -2.265e+06  9.838e+04 -23.027  < 2e-16 ***
longitude                       -2.678e+04  1.140e+03 -23.495  < 2e-16 ***
latitude                        -2.551e+04  1.124e+03 -22.689  < 2e-16 ***
housing_median_age               1.089e+03  4.946e+01  22.007  < 2e-16 ***
total_rooms                     -5.649e+00  8.913e-01  -6.338 2.39e-10 ***
total_bedrooms                   9.978e+01  7.713e+00  12.937  < 2e-16 ***
population                      -3.655e+01  1.196e+00 -30.551  < 2e-16 ***
households                       4.431e+01  8.299e+00   5.340 9.44e-08 ***
median_income                    3.907e+04  3.802e+02 102.764  < 2e-16 ***
factor(ocean_proximity)INLAND   -3.941e+04  1.967e+03 -20.030  < 2e-16 ***
factor(ocean_proximity)NEAR BAY -4.035e+03  2.146e+03  -1.880   0.0601 .
factor(ocean_proximity)NEAR OCEAN  3.370e+03  1.752e+03   1.924   0.0544 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68800 on 16272 degrees of freedom
Multiple R-squared:  0.6446,    Adjusted R-squared:  0.6444
F-statistic:  2683 on 11 and 16272 DF,  p-value: < 2.2e-16
```

Figure 2: Individual t-test to identify significant predictors

From the output of the t-test, all the variables of the Full model are significant enough (with $\alpha = 0.05$) to be in the model.

The values of the adjusted R square and the RMSE are 0.644365 and 68797.339813 dollars respectively. So, the next step will be a stepwise regression procedure. And for this data, we used Backward Elimination Method since the other two methods had some errors. So, as per the output, there is no predictor to be eliminated from the full model. Next, we are going to evaluate the first-order model based on all the possible regression selection procedures.

```
      rsquare AdjustedR         cp      AIC
[1,] 0.4745273 0.4744950 7779.15990 415346.5
[2,] 0.5879406 0.5878394 2588.44546 411393.4
[3,] 0.5969257 0.5968019 2179.05593 411036.4
[4,] 0.6173188 0.6171778 1247.34233 410192.9
[5,] 0.6299904 0.6298313  669.16519 409646.6
[6,] 0.6331405 0.6330053  526.93452 409505.4
[7,] 0.6430159 0.6428185   76.78172 409067.0
[8,] 0.6439829 0.6437641   34.51043 409024.9
[9,] 0.6446055 0.6443653    8.00000 408998.3
```

Figure 3: AIC and CP values based on ols_step_best_subset

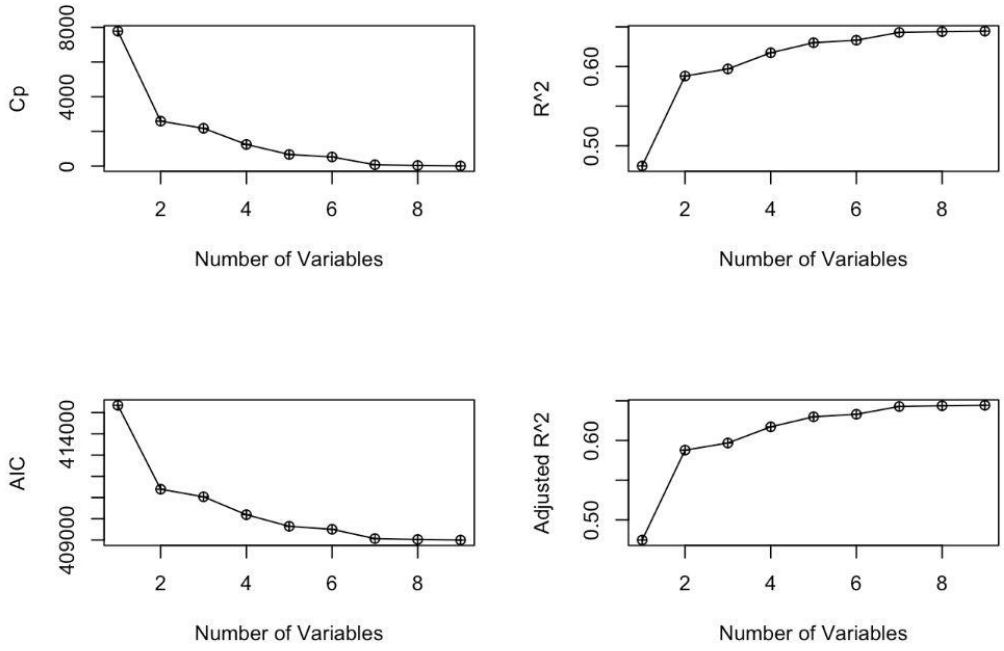


Figure 4: AIC AND CP VS number of Predictors

From Figure 3: AIC and CP values based on ols_step_best_subset and Figure 4, the best CP and AIC belong to the model that includes all 9 predictors, which is the same model as Figure 2. Now the other controlling criteria are BIC, RMSE and Adjusted R^2 which are addressed next.

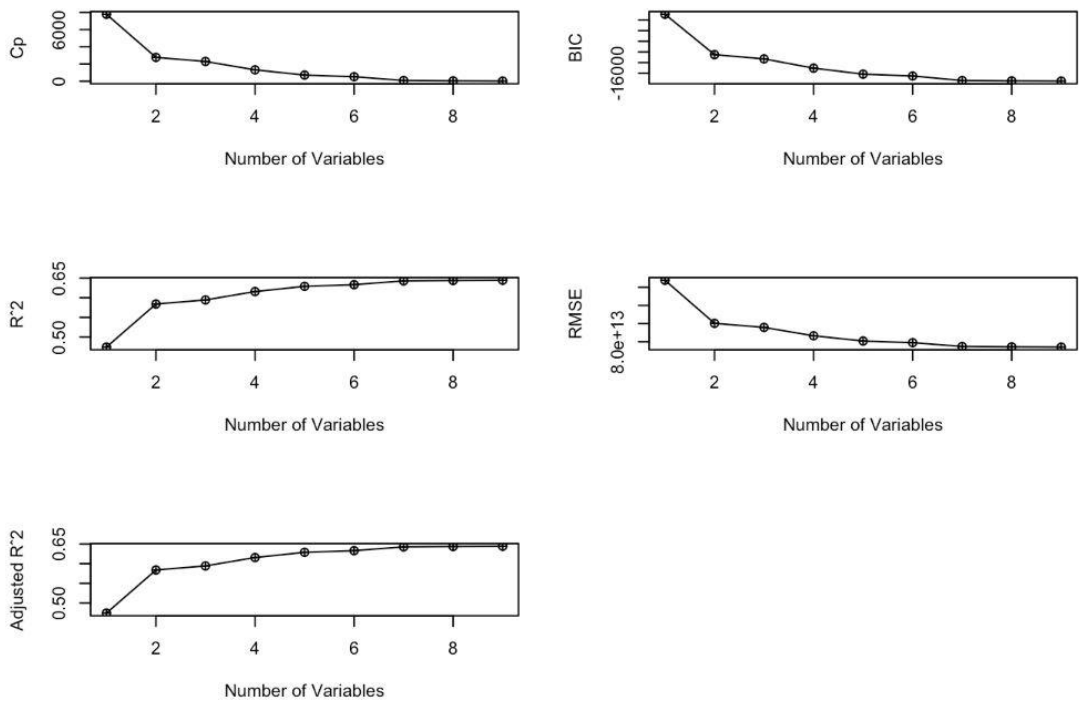


Figure 5: BIC, RMSE and Adjusted R vs Number of Predictors

	rsquare	cp	BIC	RMSE	AdjustedR
[1,]	0.4745273	7779.15990	-10458.66	1.138738e+14	0.4744950
[2,]	0.5839940	2769.14367	-14252.88	9.015154e+13	0.5839429
[3,]	0.5943812	2295.55953	-14654.94	8.790056e+13	0.5943064
[4,]	0.6156472	1323.87902	-15522.18	8.329206e+13	0.6155528
[5,]	0.6289357	717.45404	-16085.44	8.041234e+13	0.6288218
[6,]	0.6331405	526.93452	-16261.33	7.950113e+13	0.6330053
[7,]	0.6428307	85.26263	-16687.53	7.740120e+13	0.6426771
[8,]	0.6437789	43.84795	-16721.12	7.719572e+13	0.6436038
[9,]	0.6443998	17.41985	-16739.83	7.706116e+13	0.6442032

Figure 6: BIC, RMSE, and Adjusted R square for the first order linear model

Based on Figure 6 and Figure 5, we also see that model 9 is the best choice if we consider BIC and Adjusted R-squared. Up to this step, we recognized the significant predictors as well as the number of predictors based on the criteria. The values of

RMSE for (Figure 6:) are not in accordance with the RMSE from the summary table so, we do not trust the results of RMSE from this figure, and it will not have an effect on our model selection.

Before finalizing the first-order model, we are going to check the multicollinearity issue. In this part we use ggpairs plot and the VIF test.

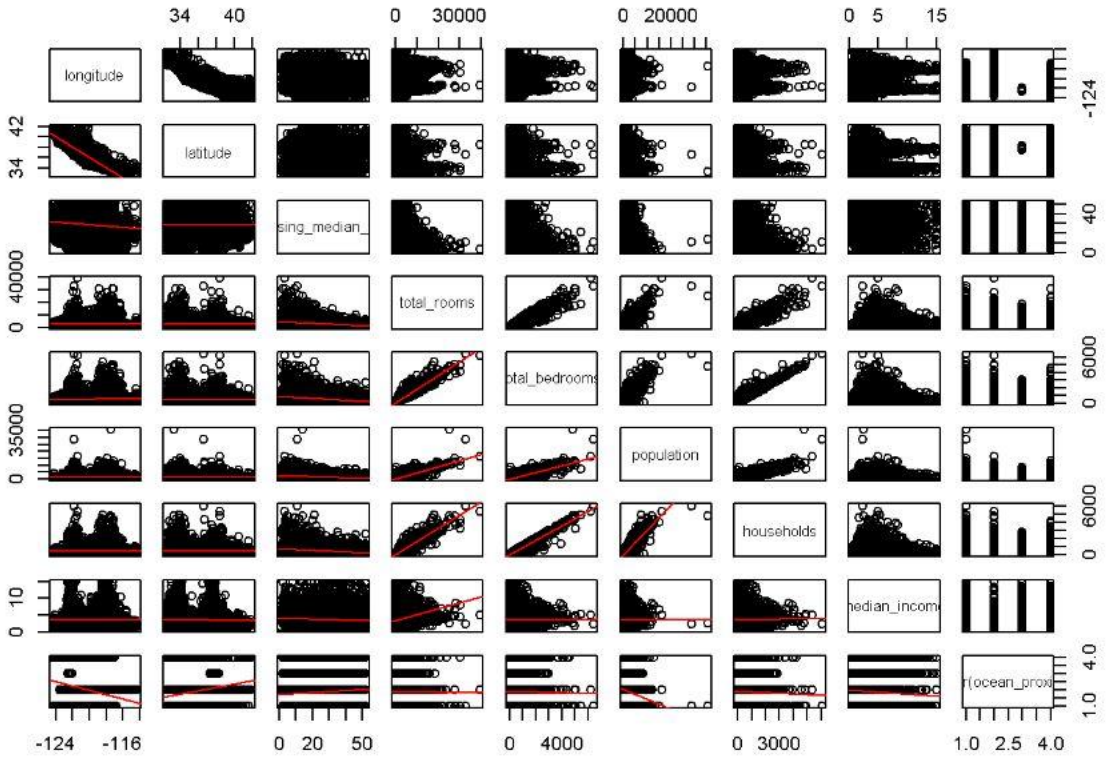


Figure 7: Pair-wise scatter plot to identify correaltion between independent variables

```
Call:
imcdiag(mod = full_model, method = "VIF")


 VIF Multicollinearity Diagnostics

                                    VIF detection
longitude                         17.9905         1
latitude                          19.8747         1
housing_median_age                 1.3321         0
total_rooms                       12.9399         1
total_bedrooms                    35.9855         1
population                         6.3387         0
households                        34.3829         1
median_income                      1.7826         0
factor(ocean_proximity)INLAND      2.8869         0
factor(ocean_proximity)NEAR BAY    1.5634         0
factor(ocean_proximity)NEAR OCEAN  1.1987         0

Multicollinearity may be due to longitude latitude total_rooms total_bedrooms households regressors

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```

Figure 8: The VIF test result for the first order model

From the scatter plot (Figure 7), we can observe that the longitude, and latitude are highly correlated. On the other hand, population and households, total rooms, and total bedrooms are highly correlated. So, we use the VIF test to diagnose the multicollinearity to identify the correlation between independent variables and the strength of that correlation.

Also, the above table (Figure 8) validates our findings from the plot. So, we should find a remedy for the highly correlated predictors. Since the total bedrooms variable is considered in the total rooms we can remove it from the model. There are three predictors left that are relevant: households, population, and total rooms. The population is a better estimator of how crowded a block can be than households. Also, the total rooms can give approximately similar information compared to households. Therefore, the household variable can also be removed from the model. For latitude and longitude, since each of them cannot express the location of a block solely, we cannot drop one and keep the other. In this case, we use principal component analysis (PCA) to combine these two values. The PCA analysis is summarized below:

```
Importance of components:
                          PC1     PC2
Standard deviation     1.3873 0.27454
Proportion of Variance 0.9623 0.03769
Cumulative Proportion  0.9623 1.00000
Standard deviations (1, .., p=2):
[1] 1.3873096 0.2745398

Rotation (n x k) = (2 x 2):
                                PC1       PC2
Housing_Data.latitude  -0.7071068 0.7071068
Housing_Data.longitude  0.7071068 0.7071068
```

Figure 9: Summary of PCA analysis

The first principal component (PC1) is the linear combination of the original variables that accounts for the NW-SE direction, and the second principal component (PC2) captures the location along the NE-SW direction. Instead of using latitude and longitude in the model, we use PC1 and PC2, and name them NW_SE and NE_SW.

Since we add new columns to our data set, it is important to split the test and train data with the same seed to have the same train and test data. After all these changes the result of the VIF will be like Figure 10. We saw the multicollinearity occurred between latitude and longitude. However, after PCA analysis, it is observed that the two components of location are not highly correlated. Also, the result of the individual t-test shows that both predictors related to location should be kept in the model. So, based on all the modifications here, we can build our first-order model with significant predictors and eliminate the effect of multicollinearity. At the end of this part, the final first-order linear model is:

$$\begin{aligned} \widehat{median\_house\_value} &= 58010 + 1100NW\_SE - 80330NE\_SW + 1017housing\_median\_age + 15.11total\_rooms \\ &\quad - 25.13population + 33670median\_income - 46740INLAND - 1620NEAR\ BAY \\ &\quad + 3007\ NEAR\ OCEAN \end{aligned}$$

From the above model, when we are heading towards NW, the value of NW_SE will increase, but the House_price will decrease. On the other hand, when we are moving towards the SW direction (getting closer to the beach), the NE_SW value in the equation is decreasing (the negative number increases), so the price of houses will be higher. The other information that can be retrieved from the model is the difference between the blocks near the beach and away from it. The ocean_proximity variable has four factors: <1H ocean, INLAND, NEAR BAY, NEAR OCEAN. The interception represents the price of blocks less than one hour from the beach. The difference between the blocks NEAR OCEAN and INLAND is, by average, 49747 dollars when the other predictors are held constant and equal (Although it is not possible because the location changes). The reason why the INLAND and NEAR BAY factors have negative signs is the price of the blocks less than 1 hour from the ocean is higher since the intercept is positive.

```
Call:
imcdiag(mod = modified_first_order_model, method = "VIF")


 VIF Multicollinearity Diagnostics

                                     VIF detection
NW_SE                             1.4582         0
NE_SW                             2.7813         0
housing_median_age                1.3264         0
total_rooms                       4.7629         0
population                        4.5004         0
median_income                     1.3401         0
factor(ocean_proximity)INLAND     2.8406         0
factor(ocean_proximity)NEAR BAY   1.5604         0
factor(ocean_proximity)NEAR OCEAN 1.1982         0

NOTE:  VIF Method Failed to detect multicollinearity


0 --> COLLINEARITY is not detected by the test
```

Figure 10: VIF test after tackling multicollinearity

```
Call:
lm(formula = median_house_value ~ NW_SE + NE_SW + housing_median_age +
    total_rooms + population + median_income + factor(ocean_proximity),
    data = traindf)

Residuals:
    Min      1Q  Median      3Q     Max
-501105  -45344  -11810   30377  491081

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       5.801e+04  2.611e+03  22.222   <2e-16 ***
NW_SE                             1.106e+03  4.819e+02   2.294   0.0218 *
NE_SW                            -8.033e+04  3.342e+03 -24.037   <2e-16 ***
housing_median_age                1.017e+03  5.072e+01  20.051   <2e-16 ***
total_rooms                       1.511e+01  5.557e-01  27.195   <2e-16 ***
population                       -2.513e+01  1.036e+00 -24.263   <2e-16 ***
median_income                     3.367e+04  3.388e+02  99.396   <2e-16 ***
factor(ocean_proximity)INLAND    -4.674e+04  2.006e+03 -23.305   <2e-16 ***
factor(ocean_proximity)NEAR BAY  -1.602e+03  2.203e+03  -0.727   0.4673
factor(ocean_proximity)NEAR OCEAN 3.007e+03  1.800e+03   1.671   0.0948 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 70700 on 16274 degrees of freedom
Multiple R-squared:  0.6246, Adjusted R-squared:  0.6244
F-statistic:  3009 on 9 and 16274 DF,  p-value: < 2.2e-16
```

Figure 11: The summary of the first order model after removing multicollinearity

3.2 Interaction and higher-order terms

Interaction and higher-order terms can be useful in capturing non-linear relationships between the predictor variables and the response variable. When two or more predictor variables interact, the relationship between each predictor and the response variable may change depending on the value of the other predictor(s). In this case, adding interaction terms to the model can improve its fit. Now, similar to the previous part we use ggpairs plot to see if there is any nonlinear relationship between prediction variables and the response. Based on the Figure 14, we used 50 sample data (since train data points are 16248 the plots will not show the pattern clearly, and it takes a while to draw all the plots, so we used sample data.) to gain an initial intuition on how to implement higher powers for each predictor, and then we will modify it based on p_value. We started from 11 for each predictor, and after considering many steps, the higher-order model is determined. We also did the same for the interaction terms. Some of them are significant and others are not significant when combined with higher-order

terms. Figure 12 shows the final model and all the interactions and higher-order terms, and Figure 13 shows the summary of the model with each t-test.

```
final_house_price_model_1= lm(
   median_house_value~ (
      poly(NW_SE,4)+
      poly(NE_SW,2)+
      poly(housing_median_age,4)+
      poly(total_rooms,5)+
      poly(population,3)+
      poly(median_income,3)+
      factor(ocean_proximity)+

      NW_SE:NE_SW + NW_SE:housing_median_age
      +NW_SE:median_income+NW_SE:factor(ocean_proximity)+

      NE_SW:housing_median_age+NE_SW:median_income+NE_SW:factor(ocean_proximity)+

      housing_median_age:total_rooms+housing_median_age:population+
      housing_median_age:median_income+housing_median_age:factor(ocean_proximity)+

      total_rooms:median_income+total_rooms:factor(ocean_proximity)+

      population:median_income+

      median_income:factor(ocean_proximity)



), data=traindf
```

Figure 12: Model with interaction and higher order terms

(Intercept)	43.466	< 2e-16	***
poly(NW_SE, 4)1	-3.610	0.000308	***
poly(NW_SE, 4)2	-13.517	< 2e-16	***
poly(NW_SE, 4)3	5.625	1.88e-08	***
poly(NW_SE, 4)4	-19.370	< 2e-16	***
poly(NE_SW, 2)1	4.411	1.04e-05	***
poly(NE_SW, 2)2	4.418	1.00e-05	***
poly(housing_median_age, 4)1	-1.763	0.077911	.
poly(housing_median_age, 4)2	6.869	6.69e-12	***
poly(housing_median_age, 4)3	3.350	0.000809	***
poly(housing_median_age, 4)4	9.795	< 2e-16	***
poly(total_rooms, 5)1	-7.958	1.87e-15	***
poly(total_rooms, 5)2	-10.484	< 2e-16	***
poly(total_rooms, 5)3	5.289	1.25e-07	***
poly(total_rooms, 5)4	-2.460	0.013910	*
poly(total_rooms, 5)5	3.418	0.000633	***
poly(population, 3)1	9.958	< 2e-16	***
poly(population, 3)2	9.665	< 2e-16	***
poly(population, 3)3	-13.479	< 2e-16	***
poly(median_income, 3)1	22.818	< 2e-16	***
poly(median_income, 3)2	-11.693	< 2e-16	***
poly(median_income, 3)3	-19.753	< 2e-16	***
factor(ocean_proximity)INLAND	-9.317	< 2e-16	***
factor(ocean_proximity)NEAR BAY	-11.444	< 2e-16	***
factor(ocean_proximity)NEAR OCEAN	3.119	0.001817	**
NW_SE:NE_SW	-7.241	4.64e-13	***
NW_SE:housing_median_age	8.493	< 2e-16	***
NW_SE:median_income	5.360	8.44e-08	***
factor(ocean_proximity)INLAND:NW_SE	2.988	0.002810	**
factor(ocean_proximity)NEAR BAY:NW_SE	-7.543	4.84e-14	***
factor(ocean_proximity)NEAR OCEAN:NW_SE	-12.370	< 2e-16	***
NE_SW:housing_median_age	-11.244	< 2e-16	***
NE_SW:median_income	-11.601	< 2e-16	***
factor(ocean_proximity)INLAND:NE_SW	-0.078	0.937539	
factor(ocean_proximity)NEAR BAY:NE_SW	-23.426	< 2e-16	***
factor(ocean_proximity)NEAR OCEAN:NE_SW	6.207	5.53e-10	***
housing_median_age:total_rooms	13.884	< 2e-16	***
housing_median_age:population	-10.218	< 2e-16	***
housing_median_age:median_income	3.223	0.001272	**
factor(ocean_proximity)INLAND:housing_median_age	2.616	0.008917	**
factor(ocean_proximity)NEAR BAY:housing_median_age	-2.352	0.018707	*
factor(ocean_proximity)NEAR OCEAN:housing_median_age	-0.279	0.780447	
median_income:total_rooms	16.492	< 2e-16	***
factor(ocean_proximity)INLAND:total_rooms	-5.633	1.80e-08	***
factor(ocean_proximity)NEAR BAY:total_rooms	0.355	0.722587	
factor(ocean_proximity)NEAR OCEAN:total_rooms	0.257	0.797164	
median_income:population	-18.005	< 2e-16	***
factor(ocean_proximity)INLAND:median_income	6.569	5.22e-11	***
factor(ocean_proximity)NEAR BAY:median_income	1.156	0.247669	

Figure 13: Summary of model with interaction and higher order terms

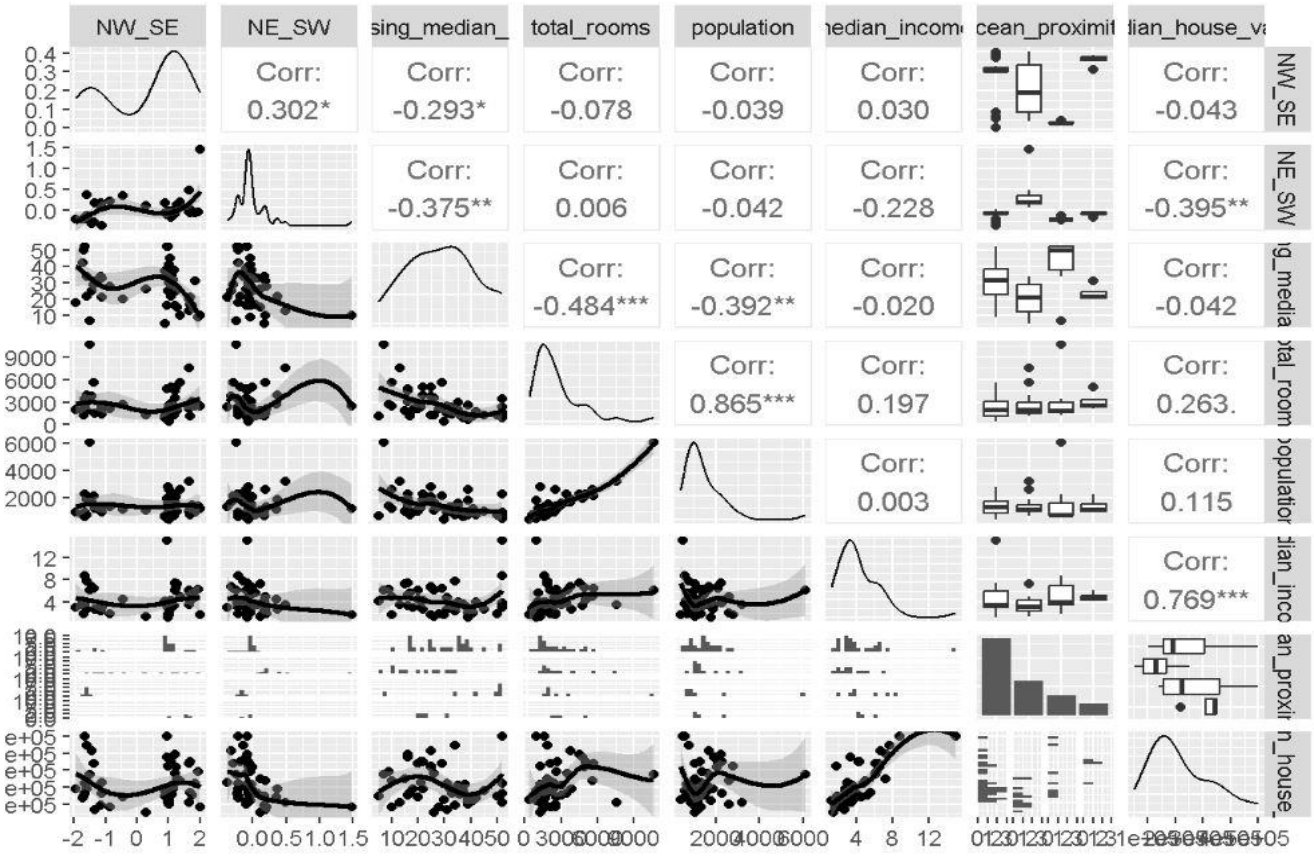


Figure 14: ggpairs plot to find any nonlinear pattern between predictors and the response

3.3 Checking the Regression Assumptions and modifying the initial model

In this section, we check the assumption of a linear regression model. Some assumptions such as Residual Independence assumption is not straightforward to calculate. However, other important conditions such as normality, equal variance and cook's distance for outliers can be checked. We also considered the multicollinearity assumption and addressed it in the previous part. The model that we consider in this part is called final_house_price_model_1 and it includes higher order terms and interaction terms.

```
studentized Breusch-Pagan test

data:  final_house_price_model_1
BP = 1434.4, df = 49, p-value < 2.2e-16
```

Figure 15: bp test for equal Variance assumption

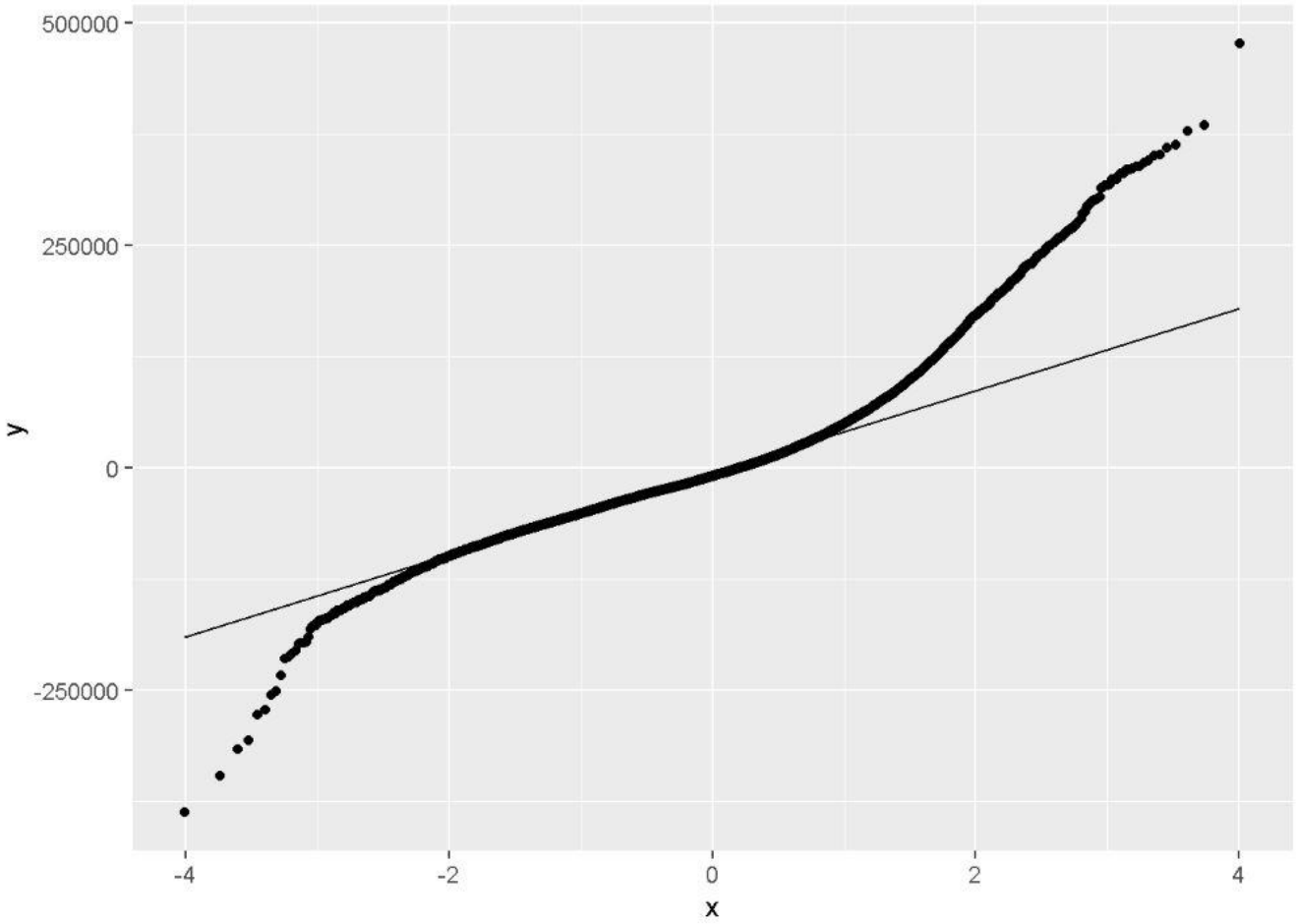


Figure 16: QQplot for the residuals of the model define in part

3.2 Interaction and higher-order terms

```
traindf[cooks.distance(final_house_price_model_1)>0.5,]
```

	longitude <dbl>	latitude <dbl>	housing_median_age <int>	total_rooms <int>	total_bedrooms <int>	population <int>	households <int>
9876	-121.79	36.64	11	32627	6445	28566	6082
13135	-121.44	38.43	3	39320	6210	16305	5358
15356	-117.42	33.35	14	25135	4819	35682	4769

3 rows | 1-8 of 13 columns

Figure 17: Outliers with cook's distance greater than 0.5

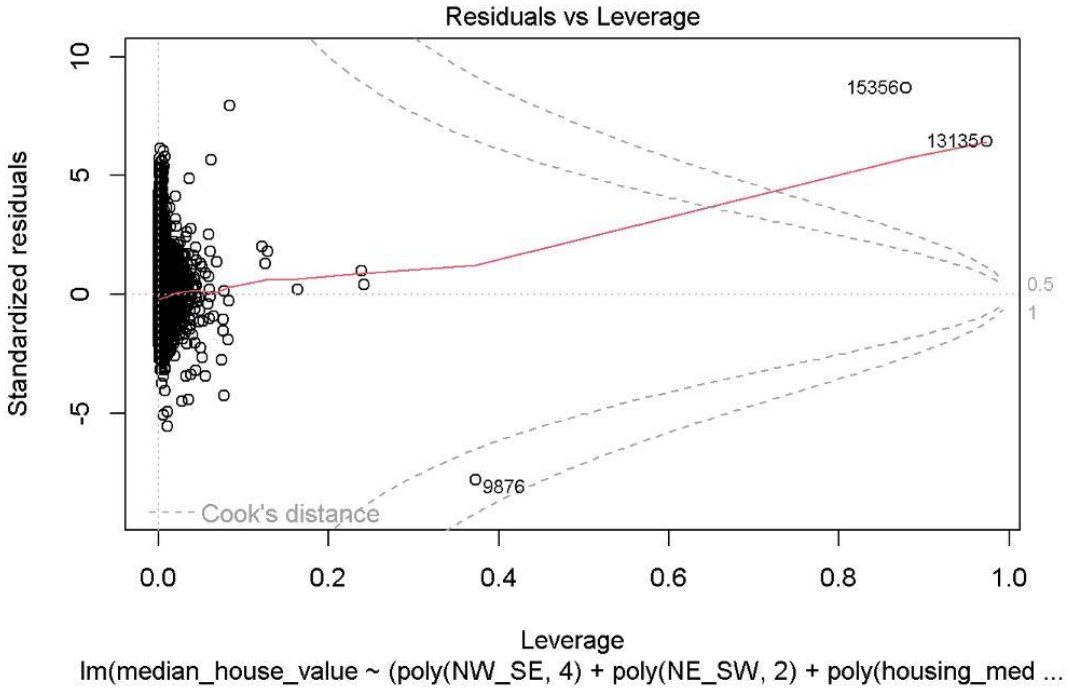


Figure 18: Residuals vs Leverage plot for detecting outliers or influential points

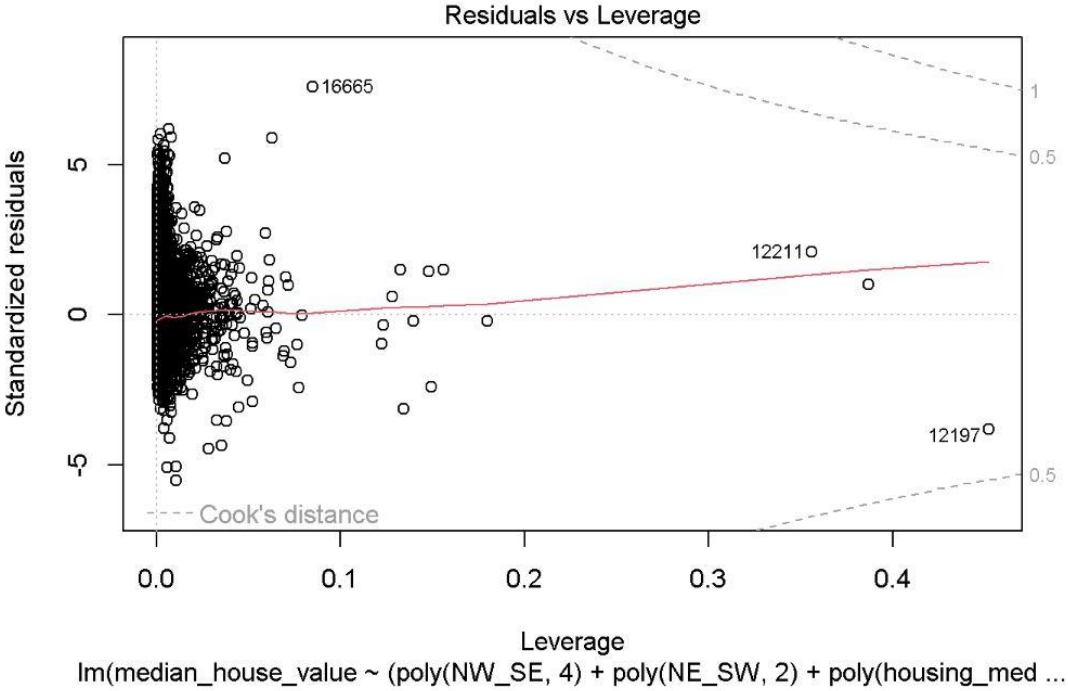


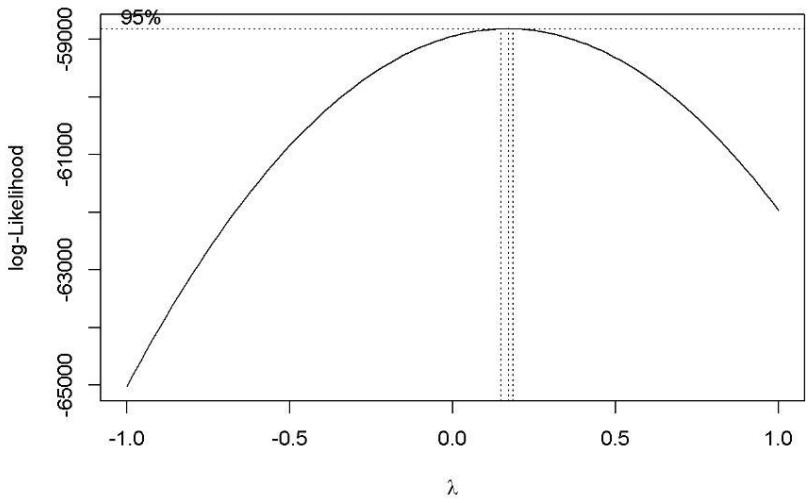
Figure 19: Residual Vs Leverage plot after removing outliers

Firstly, we used the studentized Breusch-Pagan test (Figure 15) to check the homoscedasticity. Also, the Shapiro-Wilk test showed an error because the data points are more than 5000. However, from the QQ plot (Figure 16) we can understand that the normality assumption for residuals is not met. Considering 0.5 as our cut-off for the cook's distance, three outliers were identified (Figure 17, Figure 18) and removed from the model. There are around 16000 data points in our train data. So, removing 3 outliers does not affect the model so much. The Residual Vs Leverage plot after removing outliers are shown in Figure 19.

The main model assumptions are not met here. So, we should other methods such as log transformation and box-cox transformation to see if the model will improve or not.

3.4 Modifying model by log transformation and Box-Cox transformation

In this part, to mitigate the effect of non-normality and heteroscedasticity, we use log transformation and Box_Cox transformation and perform the model diagnostic again to see how the model will improve. Since the model from the log transformation is not improved significantly (All the calculations are available in the appendix), in this section we summarize the final best model obtained from the Box_Cox transformation. By individual t-tests, some insignificant predictors were dropped from the model (Just higher order terms and the extra interaction terms).




```
# best lamda value
best_lambda = box_cox_model $ x[which(box_cox_model $ y == max(box_cox_model $ y))]
best_lambda
```

```
## [1] 0.1717172
```

Figure 20: Finding Best Lambda for Box_Cox transformation

```
##                                                t value Pr(>|t|)
## (Intercept)                                    312.991  < 2e-16 ***
## poly(NW_SE, 4)1                                -4.023 5.77e-05 ***
## poly(NW_SE, 4)2                               -15.315  < 2e-16 ***
## poly(NW_SE, 4)3                                11.959  < 2e-16 ***
## poly(NW_SE, 4)4                               -27.863  < 2e-16 ***
## poly(NE_SW, 2)1                                 3.773 0.000162 ***
## poly(NE_SW, 2)2                                 6.264 3.84e-10 ***
## poly(housing_median_age, 4)1                   -1.449 0.147494
## poly(housing_median_age, 4)2                    5.646 1.67e-08 ***
## poly(housing_median_age, 4)3                    5.529 3.26e-08 ***
## poly(housing_median_age, 4)4                    9.262  < 2e-16 ***
## poly(total_rooms, 4)1                          -3.652 0.000261 ***
## poly(total_rooms, 4)2                         -12.465  < 2e-16 ***
## poly(total_rooms, 4)3                          12.635  < 2e-16 ***
## poly(total_rooms, 4)4                          -6.314 2.80e-10 ***
## poly(population, 3)1                            4.684 2.83e-06 ***
## poly(population, 3)2                           15.631  < 2e-16 ***
## poly(population, 3)3                          -13.069  < 2e-16 ***
## poly(median_income, 3)1                        35.016  < 2e-16 ***
## poly(median_income, 3)2                       -22.106  < 2e-16 ***
## poly(median_income, 3)3                        -7.331 2.39e-13 ***
## factor(ocean_proximity)INLAND                 -15.353  < 2e-16 ***
## factor(ocean_proximity)NEAR BAY                -9.663  < 2e-16 ***
## factor(ocean_proximity)NEAR OCEAN               3.141 0.001685 **
## NW_SE:NE_SW                                   -11.010  < 2e-16 ***
## NW_SE:housing_median_age                        8.801  < 2e-16 ***
## NW_SE:median_income                             3.949 7.88e-05 ***
## factor(ocean_proximity)INLAND:NW_SE             6.041 1.57e-09 ***
## factor(ocean_proximity)NEAR BAY:NW_SE          -5.390 7.15e-08 ***
## factor(ocean_proximity)NEAR OCEAN:NW_SE       -14.452  < 2e-16 ***
## NE_SW:housing_median_age                      -12.055  < 2e-16 ***
## NE_SW:median_income                            -7.536 5.10e-14 ***
## factor(ocean_proximity)INLAND:NE_SW            -3.725 0.000196 ***
## factor(ocean_proximity)NEAR BAY:NE_SW         -21.155  < 2e-16 ***
## factor(ocean_proximity)NEAR OCEAN:NE_SW         6.023 1.75e-09 ***
## housing_median_age:total_rooms                 10.820  < 2e-16 ***
## housing_median_age:population                  -6.900 5.39e-12 ***
## factor(ocean_proximity)INLAND:housing_median_age   1.782 0.074701 .
## factor(ocean_proximity)NEAR BAY:housing_median_age -2.007 0.044803 *
## factor(ocean_proximity)NEAR OCEAN:housing_median_age -0.838 0.402318
## median_income:total_rooms                      13.025  < 2e-16 ***
## factor(ocean_proximity)INLAND:total_rooms      -3.096 0.001967 **
## factor(ocean_proximity)NEAR BAY:total_rooms      1.244 0.213696
## factor(ocean_proximity)NEAR OCEAN:total_rooms    0.957 0.338812
## median_income:population                      -14.107  < 2e-16 ***
## factor(ocean_proximity)INLAND:median_income     12.410  < 2e-16 ***
## factor(ocean_proximity)NEAR BAY:median_income    2.712 0.006692 **
## factor(ocean_proximity)NEAR OCEAN:median_income  0.254 0.799525
```

Figure 21: Individual t-test for the final model

```
Asymptotic one-sample Kolmogorov-Smirnov test

data:  box_cox_model_4$residuals
D = 0.050475, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Figure 22: Kolmogrov_Smirov test for the final model

(Intercept)	4.163e+01
poly(NW_SE, 4)1	-5.330e+01
poly(NW_SE, 4)2	-5.040e+01
poly(NW_SE, 4)3	4.660e+01
poly(NW_SE, 4)4	-7.717e+01
poly(NE_SW, 2)1	7.712e+01
poly(NE_SW, 2)2	3.143e+01
poly(housing_median_age, 4)1	-9.733e+00
poly(housing_median_age, 4)2	1.600e+01
poly(housing_median_age, 4)3	1.396e+01
poly(housing_median_age, 4)4	2.213e+01
poly(total_rooms, 4)1	-6.656e+01
poly(total_rooms, 4)2	-6.748e+01
poly(total_rooms, 4)3	4.877e+01
poly(total_rooms, 4)4	-1.580e+01
poly(population, 3)1	8.665e+01
poly(population, 3)2	7.658e+01
poly(population, 3)3	-4.845e+01
poly(median_income, 3)1	2.138e+02
poly(median_income, 3)2	-6.707e+01
poly(median_income, 3)3	-1.879e+01
factor(ocean_proximity)INLAND	-4.360e+00
factor(ocean_proximity)NEAR BAY	-1.114e+01
factor(ocean_proximity)NEAR OCEAN	8.586e-01
NW_SE:NE_SW	-1.178e+00
NW_SE:housing_median_age	1.291e-02
NW_SE:median_income	4.388e-02
factor(ocean_proximity)INLAND:NW_SE	4.157e-01
factor(ocean_proximity)NEAR BAY:NW_SE	-3.500e+00
factor(ocean_proximity)NEAR OCEAN:NW_SE	-8.315e-01
NE_SW:housing_median_age	-1.233e-01
NE_SW:median_income	-6.366e-01
factor(ocean_proximity)INLAND:NE_SW	-1.995e+00
factor(ocean_proximity)NEAR BAY:NE_SW	-2.107e+01
factor(ocean_proximity)NEAR OCEAN:NE_SW	2.950e+00
housing_median_age:total_rooms	2.023e-05
housing_median_age:population	-2.324e-05
factor(ocean_proximity)INLAND:housing_median_age	1.076e-02
factor(ocean_proximity)NEAR BAY:housing_median_age	-1.355e-02
factor(ocean_proximity)NEAR OCEAN:housing_median_age	-4.632e-03
median_income:total_rooms	1.549e-04
factor(ocean_proximity)INLAND:total_rooms	-7.172e-05
factor(ocean_proximity)NEAR BAY:total_rooms	4.583e-05
factor(ocean_proximity)NEAR OCEAN:total_rooms	2.941e-05
median_income:population	-3.823e-04
factor(ocean_proximity)INLAND:median_income	5.558e-01
factor(ocean_proximity)NEAR BAY:median_income	1.069e-01
factor(ocean_proximity)NEAR OCEAN:median_income	8.125e-03

Figure 23: Final Model Coefficients

```
bptest(box_cox_model_4)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  box_cox_model_4
## BP = 1314.7, df = 47, p-value < 2.2e-16
```

Figure 24: Equal Variance test for the final model

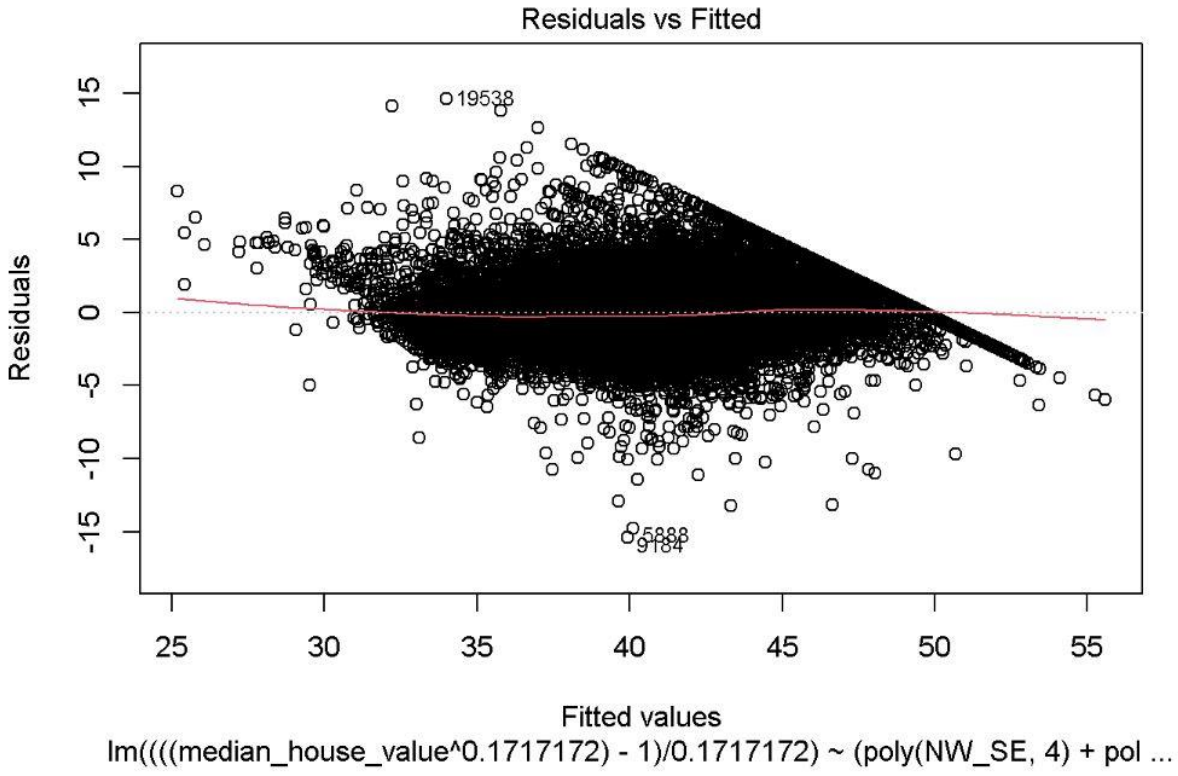


Figure 25: Residuals Vs Fitted plot for final model

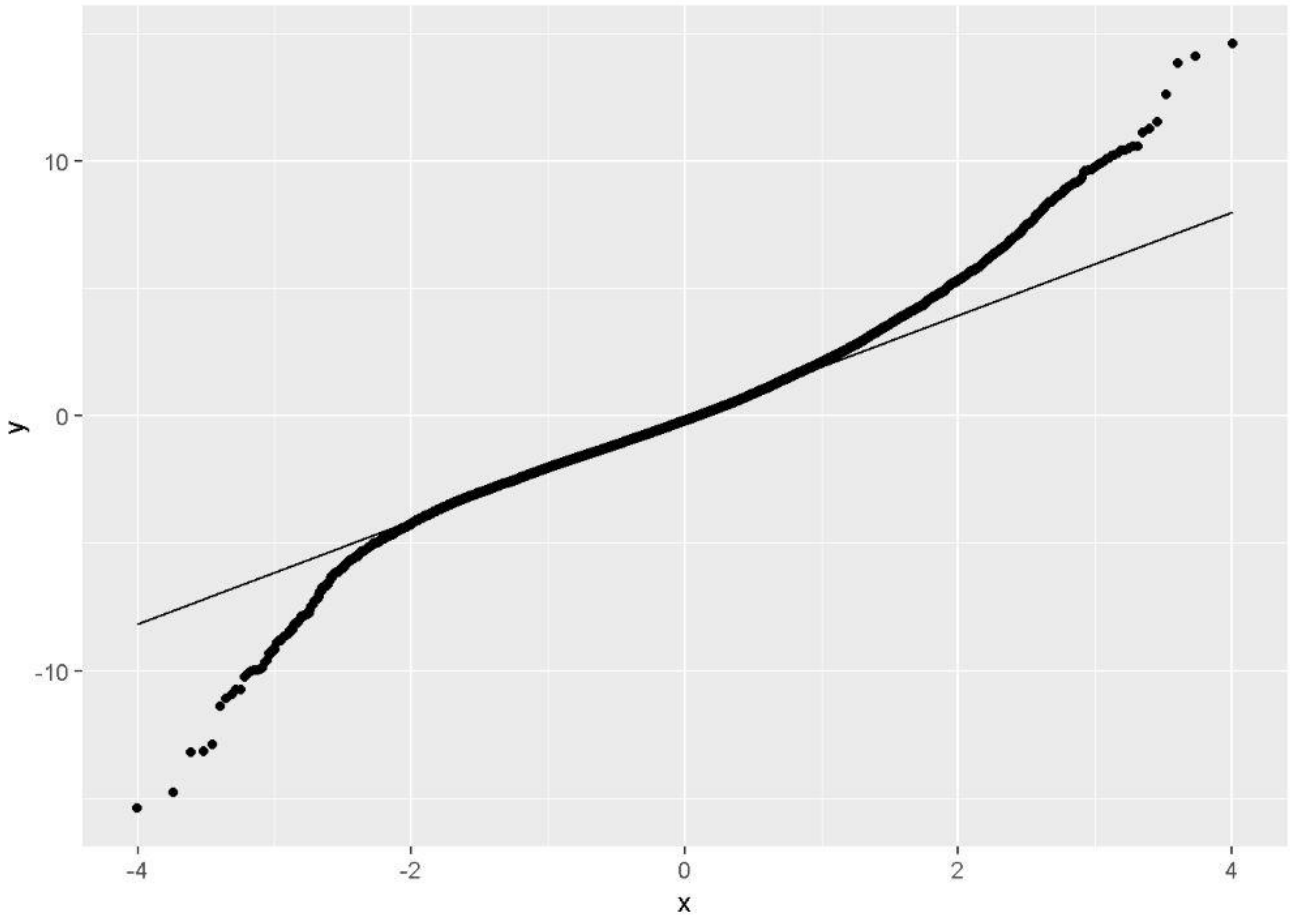


Figure 26: QQplot of the residuals of the final model

It is worth mentioning that the H0 for the normality test is the residuals are normally distributed and the H0 for the equal variance test is the Variance of all error terms are the same. The alternative hypothesis for both is against the normal condition. So, instead of repeating this many times, we just mention it once.

First, we found the best lambda for Box-Cox transformation (Figure 20), which is equal to 0.1717172. Then, after two attempts and removing the insignificant predictors from the final model, the final model and the coefficients are summarized in (Figure 23) and (Figure 21). The adjusted R-squared for this final model is 0.734, which is better than the final model before the transformation which was 0.7059. We will talk about the RMSE in the next part. Since the transformation is applied to the model and the effect of higher values in the response has been reduced, the linearity assumption seems to be met (Figure 25). Also, homoscedasticity cannot be distinguished only by the residuals vs fitted values plot. However, the result of the BP test still shows that the model is far away from the equal variance condition (Figure 24). Finally, from the QQplot and the Kolmogrov_Smirov test (Figure 22), we do not have enough statistical evidence to accept the normality assumption although the QQplot shows some improvement (Figure 26).

3.5 RMSE for training and test data

We divided the data set into train (80%) and test data (20%). Although our approach to find the best linear model did not lead to meeting all the conditions, it is worth analyzing and comparing the RMSE for the train and test data. Our approach here is using the final model to predict the test data. Then the RMSE is calculated by R based on the formula $\sqrt{\frac{\Sigma(y_i-\widehat{y_i})^2}{n-p-1}}$. All the codes are available in the appendix. Since the response variable is transformed by the Box-Cox model, we can transform the true value with the same model, or we can normalize data (subtracting each data point from mean and dividing by SD) and then compare the RMSE. The important thing is the response has no unit or changed unit here.

RMSE	Train Data	Test Data	Difference (%)
Transformed Y	2.334	2.501	+6.693
Normalized Y	0.5986	0.63072	+5.380

Table 1: RMSE Comparison between the Train and Test data

Figure 27: Predicted Price Vs Median Income for the Test Data

Finally, we plotted Predicted Price for the test data Vs Median Income to see how predicted price fits the true values. Median Income is one of the predictors selected that enable us to draw a 2D plot. Also, there is no data leakage happened since the

model is created and fitted based on the train data, and the test data is just used to show how the model works on the test data in terms of RMSE.

4 Conclusion and Future Scope

4.1 Conclusion

To summarize our findings from the analysis, the response variable, house price in the model is significantly ($\alpha = 0.05$) dependent on location (combination of latitude and longitude), house age, total rooms, population, median income, and proximity from the ocean. Then their interaction and higher-order terms have been added to the final linear model to capture the nonlinear relationship between the predictors and the response. Next, the model diagnostic was applied to our model to check the conditions such as Linearity, Normality, and Homoscedasticity. All three conditions were not met, so we used the Box-Cox transformation to improve the model. Although the final model (after transformation) conditions were not held based on statistic tests, the QQ plot and the residual plot showed a little improvement. Finally, RMSE was calculated and compared for both the training and the test data to see how well the model will perform on new data. The results indicated that the RMSE for the final model is 6.69% more than the RMSE of the train data.

4.2 Future Scope

There is a significant future scope for this housing price prediction project. Here are a few potential areas of development:

1. **Advanced ML algorithms:** Instead of multiple linear regression we can use sophisticated Machine learning algorithms which is capable of handling large amounts of data and not dependent on the normality of the residuals. There is a lot of potential for developing more advanced algorithms that can make more accurate predictions and better identify trends in the market.s
2. **Integration of additional features**: Housing price prediction models can be enhanced by incorporating more features about the house like the built-up area, carpet area, garden area, architecture of the building, materials used, number of floors, number of washrooms per floor etc. Some economic indicators like inflation, Consumer Price Index (purchasing power of people) and Housing loan rate make a great impact on the housing price of any society. Apart from these, to understand and predict the real trend we need to incorporate time series data of at least past 5 years. This way more accurate predictions can be made.
3. **Enhanced visualization tools**: Simple but effective data visualization tools can help users understand the trends and patterns in housing price data more effectively. Developing more advanced and user-friendly visualization tools can improve the user experience and make it easier to analyze large amounts of data.
4. **Personalized recommendations**: With the help of machine learning algorithms, personalized recommendations can be made to users about which properties they should consider buying or selling based on their unique preferences and requirements.

In conclusion, as the real estate industry continues to evolve and generate more data, the future scope for housing price prediction project is vast, and there are many opportunities for innovation and growth in this field.

5. References

[1] California Housing Prices | Kaggle. (n.d.). Retrieved March 8,2023, from https://www.kaggle.com/datasets/camnugent/california-housing-prices

[2] Pace, R. K., & Barry, R. (1997). Sparse Spatial Autoregressions. Statistics and Probability Letters, 33(3), 291-297. Retrieved March 8, 2023, from https://doi.org/10.1016/S0167-7152(96)00128-X

[3] Torgo, L. (n.d.). California Housing dataset. University of Porto. https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html

6. Appendix

The source code for this project has been added to the below .rmd file.

DATA603_Final_Project_FAA_20230405.Rmd