Analysis of House Prices , USA

Sai Mounica Pothuru (RUID: 198008261)

MSDS596: Regression and Time Series

Dec 08 2020

1 Introduction

A home is a place where you feel safe and secure; a place where you experience emotional warmth and feel surrounded by love and affection; a place where there are no constraints on your development and where you don't have to fight for your rights continually. Everyone needs to have a stable home, but we all know how difficult it could be to secure a perfect fairytale house. Price plays a significant role in acquiring a place. It varies from several thousand to millions. Neighborhood comps, Location, Home size and usable space, Age, and condition are a few of the critical factors that affect the houses' pricing. In this project, we will analyze the importance of these factors and how each element affects the outcome. Multiple linear regression used will guide several people in finding the perfect space in the location

2 Dataset

The data set **House Sales in King County, USA** is attained from Kaggle and is an open data set. The data was a collection of 21613 houses, collected in King County, including Seattle, the USA, from May 2014 and May 2015. It contains 19 attributes namely *id, date, price, number of bedrooms, bathrooms, sqft_living, sqft_loft, floors, waterfront, view, sqft_above, sqft_basement, yr_built, yr_renovated, zip code, lat, long.* Price is the response variable making the rest 19 variables predictors.

All the attributes are assumed to be continuous for analysis purposes. The data has no missing entries and is clean. The predictors are all numeric except date and location coordinates. All the observations are unique. All the analyses and visualizations presented are performed in R v 4.0.3. Upon looking at the data, few glitches observed, such as the yr_renovated, are 0 for a few observations, which does not make sense—assigning NA values to such records.

3 Data Analytic strategy

Uni-variate, Bi-variate analysis, and visualizations are performed to understand data distribution. Correlation (Figure 2) between different attributes is plotted to detect highly correlated predictors. Predictors date and id had the least impact on the response and hence removed during the model generation. The VIF and skewness of the predictor variables are calculated. None of the variables had VIF factor¿10. So, the threshold is changed to 5, giving two predictors with high multicollinearity. Hence, one of the variables *sqft_above* will not be used in the model.

A linear model is generated initially with the other predictors. The QQ plot is mapped and observed that transformation is required. Boxcox is applied to get a suitable lambda for transformation. The lambda resulted in 0 effecting to devise a log transformation of the generated model. Even though the transformation leads to a more stable model, having more than 15 variables seemed unnecessary. The forward selection method is implemented to get the variables that broadly explained the variance in the outcome. AIC and BIC criteria are applied to get these variables.

A model is generated with the filtered variables, and the eight predictors contributed to 75% of the variance in the response

Finally, model sensitivity is tested for the final model. The dataset is split to train and test data at a 70:30 ratio. The model is generated with the train set and is predicted using the test set. The RMSE for the obtained actual and predicted values resulted in 0.264, showing that the model is stable.

4 Results

The descriptive statistics are provided in Table 1(zipcode and location coordinates are removed). The data shows few irregularities in *yr_renovated* and hence modified the 0 values to NA. There are few outliers observed but did not delete them in-case if they are reasonable observations.

From the correlation plot and the VIF Table (Table 2), we can observe that there is a high correlation between predictors *sqft_living* and *sqft_above*. Hence, we cannot use both the predictors, removing *sqft_above* from further model generation. Various models are generated using forward selection. AIC and BIC criteria are applied to the complete data-set for selecting the predictors that explain the maximum variance. The BIC plot shows the number of predictors that can be used to describe the maximum variance.

Model validation is performed using training and testing data sets, and the results are as shown in Table 3. The values are logarithmic results as log transformation is applied to the final model as suggested by Boxcox. The results are not far from expected, following the RMSE to be 0.264.

5 Conclusion

The analysis shows that the two predictors, location and size have the utmost importance in deciding the price of a house. The model generated is a reasonable one that explains 75% of variance in the response. However, the pricing of homes does not depend only on the given attributes. The research should be conducted in such a way that they should identify the reason behind the highest price difference of houses in the same area. For example, what is the crime rate? How is schooling in the neighborhood? etc., as these play an essential role in deciding the worth of the house for the estimated price.

References

Power transform - https://en.wikipedia.org/wiki/Power_transform

Variable Selection - https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Dataset - https://www.kaggle.com/

id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long
1000102	20140916T000000	280000	6	3	2400	9373	2	0	0	3	7	2400	0	1991	0	98002	47.3262	-122.214
1000102	20150422T000000	300000	6	3	2400	9373	2	0	0	3	7	2400	0	1991	0	98002	47.3262	-122.214
1200019	20140508T000000	647500	4	1.75	2060	26036	1	0	0	4	8	1160	900	1947	0	98166	47.4444	-122.351
1200021	20140811T000000	400000	3	1	1460	43000	1	0	0	3	7	1460	0	1952	0	98166	47.4434	-122.347
2800031	20150401T000000	235000	3	1	1430	7599	1.5	0	0	4	6	1010	420	1930	0	98168	47.4783	-122.265
3600057	20150319T000000	402500	4	2	1650	3504	1	0	0	3	7	760	890	1951	2013	98144	47.5803	-122.294
3600072	20150330T000000	680000	4	2.75	2220	5310	1	0	0	5	7	1170	1050	1951	0	98144	47.5801	-122.294
3800008	20150224T000000	178000	5	1.5	1990	18200	1	0	0	3	7	1990	0	1960	0	98178	47.4938	-122.262
5200087	20140709T000000	487000	4	2.5	2540	5001	2	0	0	3	9	2540	0	2005	0	98108	47.5423	-122.302
6200017	20141112T000000	281000	3	1	1340	21336	1.5	0	0	4	5	1340	0	1945	0	98032	47.4023	-122.273
7200080	20141104T000000	239000	4	2	1980	10585	1.5	0	0	2	6	1980	0	1924	0	98055	47.4836	-122.214
7200179	20141016T000000	150000	2	1	840	12750	1	0	0	3	6	840	0	1925	0	98055	47.484	-122.211
7200179	20150424T000000	175000	2	1	840	12750	1	0	0	3	6	840	0	1925	0	98055	47.484	-122.211
7400062	20140521T000000	299800	2	1	790	5240	1	0	0	4	6	790	0	1925	0	98118	47.5303	-122.288
7600057	20140805T000000	520000	3	2	1410	2700	2	0	0	4	7	1410	0	1902	0	98122	47.6029	-122.302
7600065	20140605T000000	465000	3	2.25	1530	1245	2	0	0	3	9	1050	480	2014	0	98122	47.6018	-122.297
7600125	20141218T000000	630000	5	1	3020	4800	2	0	0	3	7	3020	0	1901	0	98122	47.6025	-122.313
7600136	20140718T000000	411000	2	2	1130	1148	2	0	0	3	9	800	330	2007	0	98122	47.6023	-122.314
9000025	20141203T000000	496000	2	1	1420	4635	2	0	0	4	7	1420	0	1941	1973	98115	47.68	-122.304

Figure 1: Data Snippet

	Mean	SD	Median	Min	Max	Variance	N
price	540182.2	367362.2	450000	75000	7700000	1.35E+11	21613
bedrooms	3.370842	0.930062	3	0	33	0.865015	21613
bathrooms	2.114757	0.770163	2.25	0	8	0.593151	21613
sqft_living	2079.9	918.4409	1910	290	13540	843533.7	21613
sqft_lot	15106.97	41420.51	7618	520	1651359	1.72E+09	21613
floors	1.494309	0.539989	1.5	1	3.5	0.291588	21613
waterfront	0.007542	0.086517	0	0	1	0.007485	21613
view	0.234303	0.766318	0	0	4	0.587243	21613
condition	3.40943	0.650743	3	1	5	0.423467	21613
grade	7.656873	1.175459	7	1	13	1.381703	21613
sqft_above	1788.391	828.091	1560	290	9410	685734.7	21613
sqft_basement	291.509	442.575	0	0	4820	195872.7	21613
yr_built	1971.005	29.37341	1975	1900	2015	862.7973	21613
yr_renovated	84.40226	401.6792	0	0	2015	161346.2	21613

Table 1: Descriptive Analysis of the predictors

VIF TABLE	
bedrooms	1.648341
bathrooms	3.345278
sqft_living	8.323209
sqft_lot	1.103305
floors	1.983642
waterfront	1.202648
view	1.399366
condition	1.247249
grade	3.137628
sqft_above	6.809475
yr_built	2.428187
yr_renovated	1.147099
zipcode	1.647325
lat	1.177113
long	1.768878

Table 2: VIF values of the complete model

Model Validation	
predicted	**actual**
12.82077592	13.31132948
14.39950181	14.02252473
13.21868339	12.64432758
13.22379889	13.18063229
12.56874675	12.14950229
12.38098088	12.43995829
12.89098217	12.70381303
13.03284321	12.98997419
13.30546035	13.48561664
12.19755227	12.25719343
12.99203596	13.30468493
12.91052581	12.97154049
14.39867708	13.85473127
12.77273318	12.79385931
12.98250659	12.84792653
12.75114329	12.66032792
12.50353163	12.40287222
12.46337706	11.9381932
13.26079236	13.45883561

Table 3: Sample of expected Vs Actual Prices(Log)

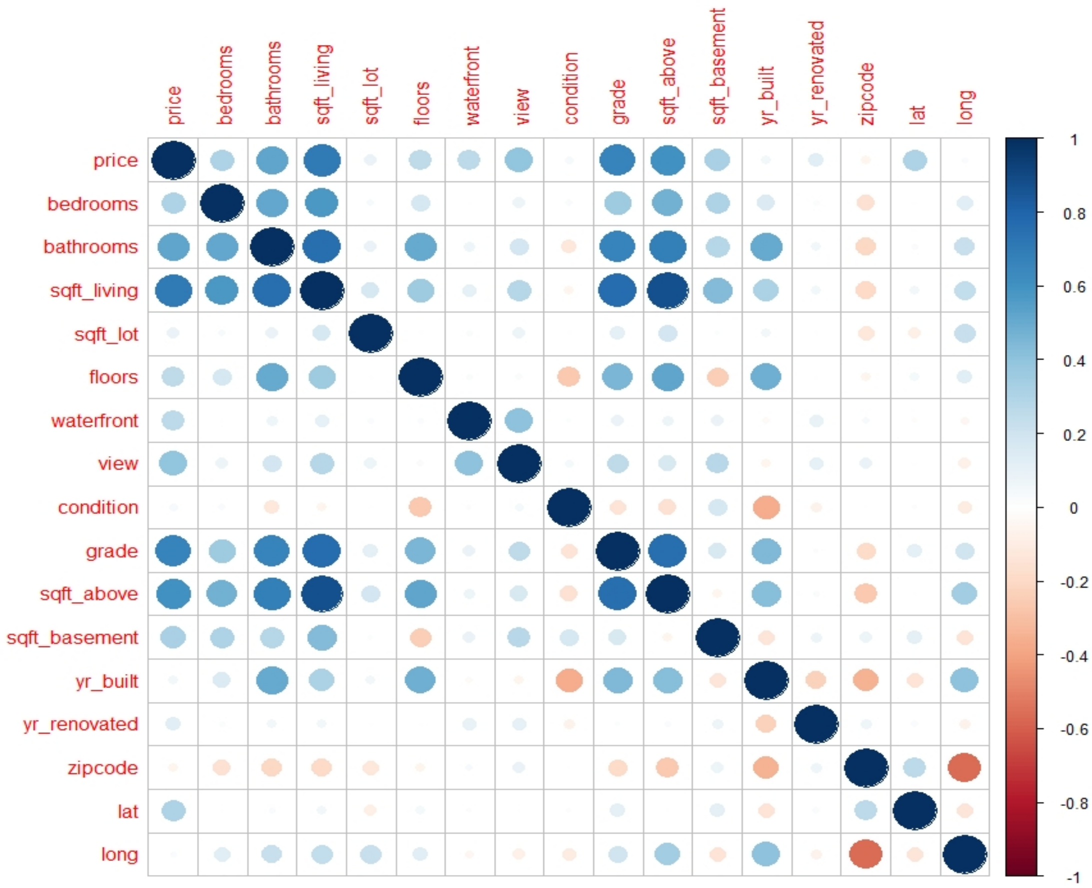

Figure 2: Correlation plot

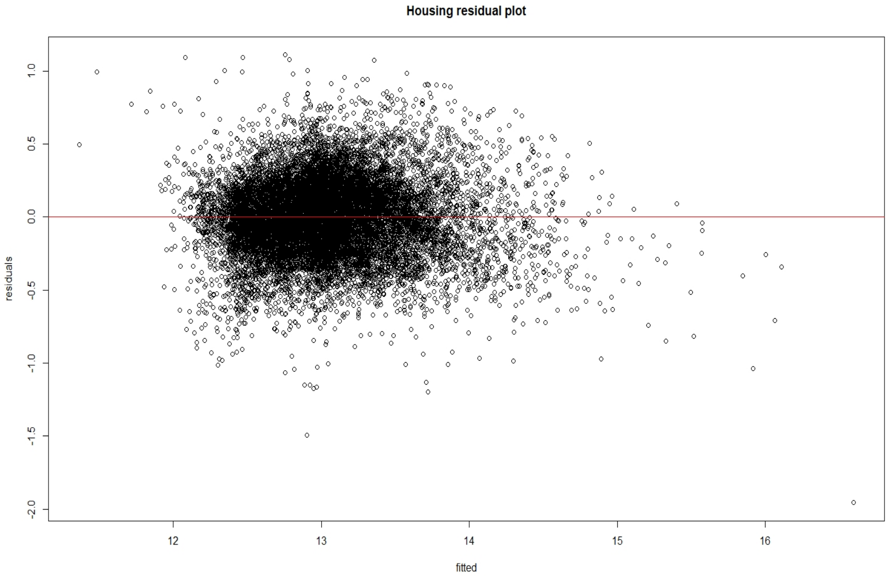

Figure 3: Residual Plot

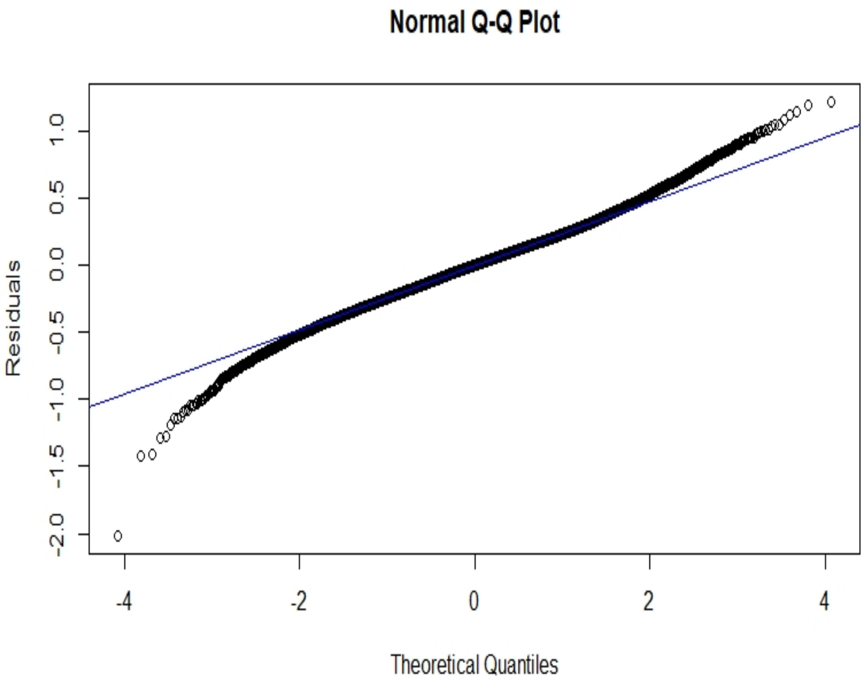

Figure 4: QQ Plot

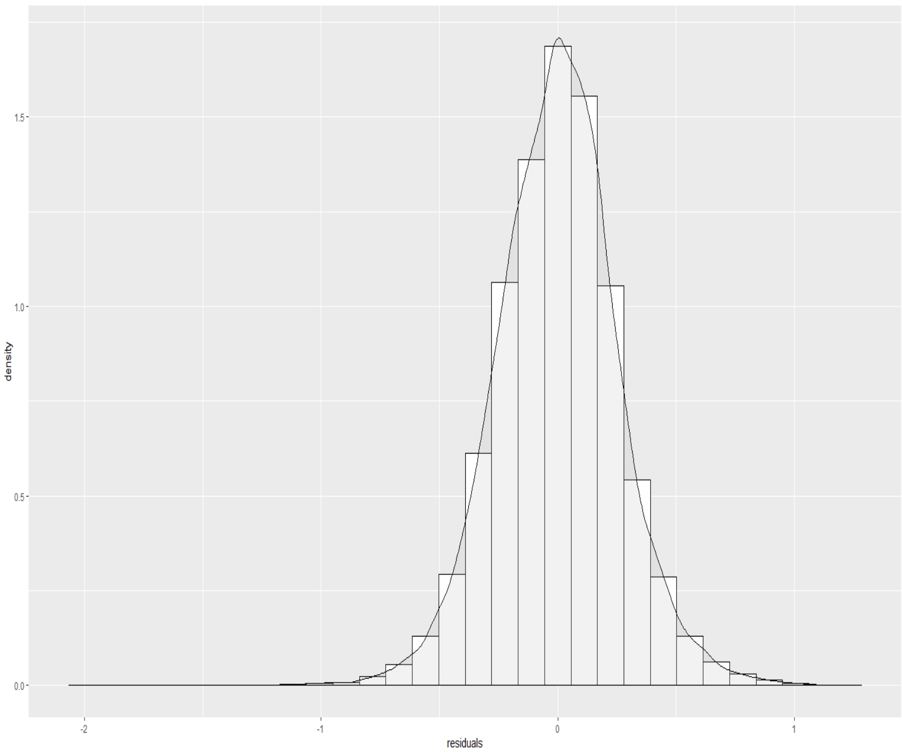

Figure 5: Density Plot

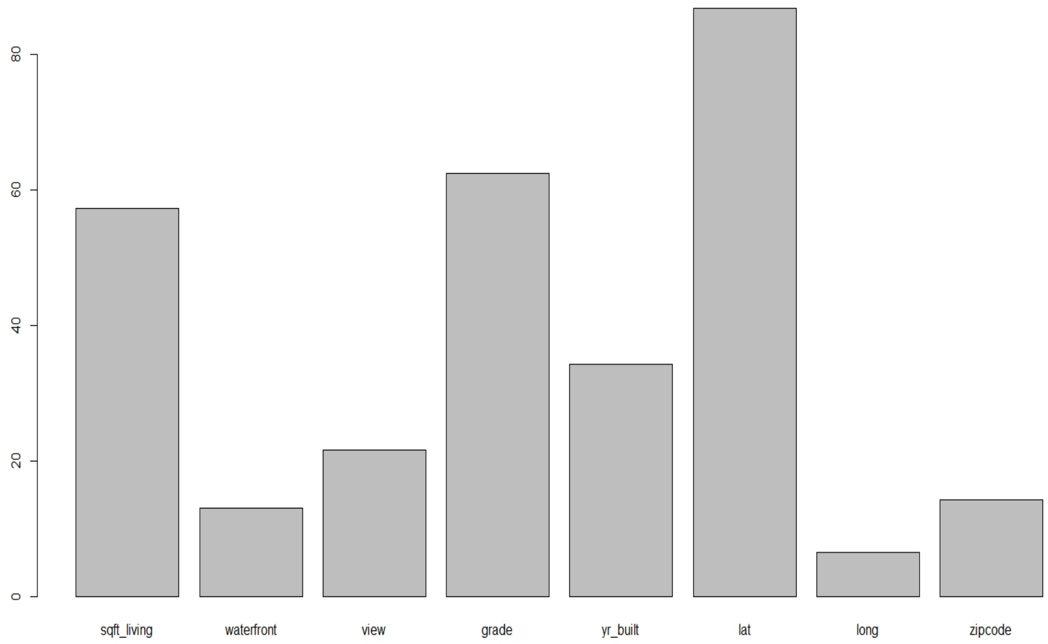

Figure 6: Variable Ranking

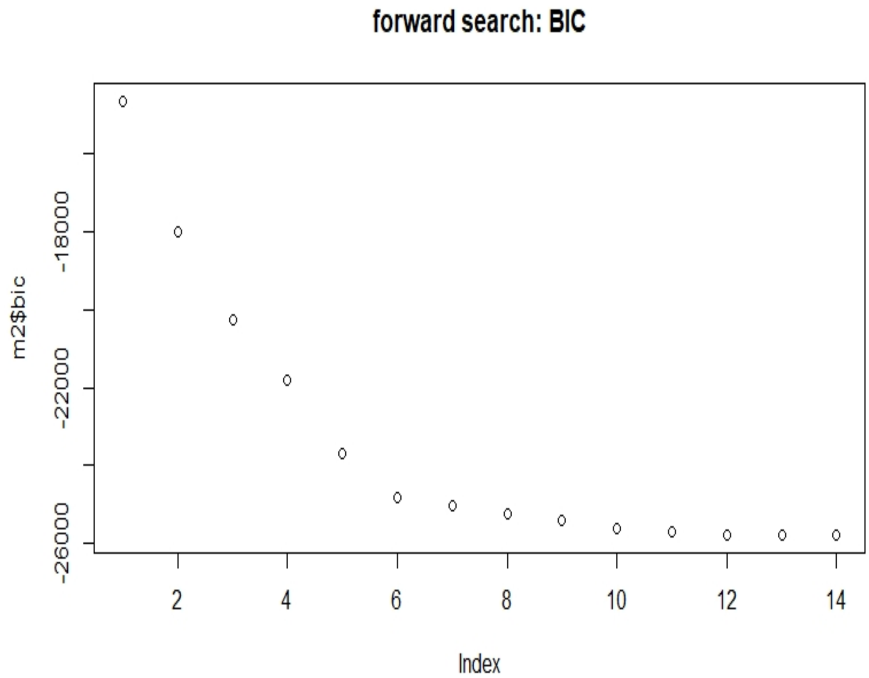

Figure 7: BIC Criterion