What makes the real estate properties expensive or affordable?*

King County, Washington, USA

Reem Abughazaleh

20 April 2022

Abstract

The real estate properties are one of the most significant proportion of a household's portfolio. This paper aims at understanding the relationship between its price and the variables that describe the real estate property for instance the square footage of the apartments interior living space, square footage of the land space, whether the property has waterfront view or not and the square footage of the interior housing space that is above ground level. The simple and the multiple regression has been used to understand the relationship and for prediction. Having a waterfront view has a positive and significant impact on the property prices.

1 Introduction

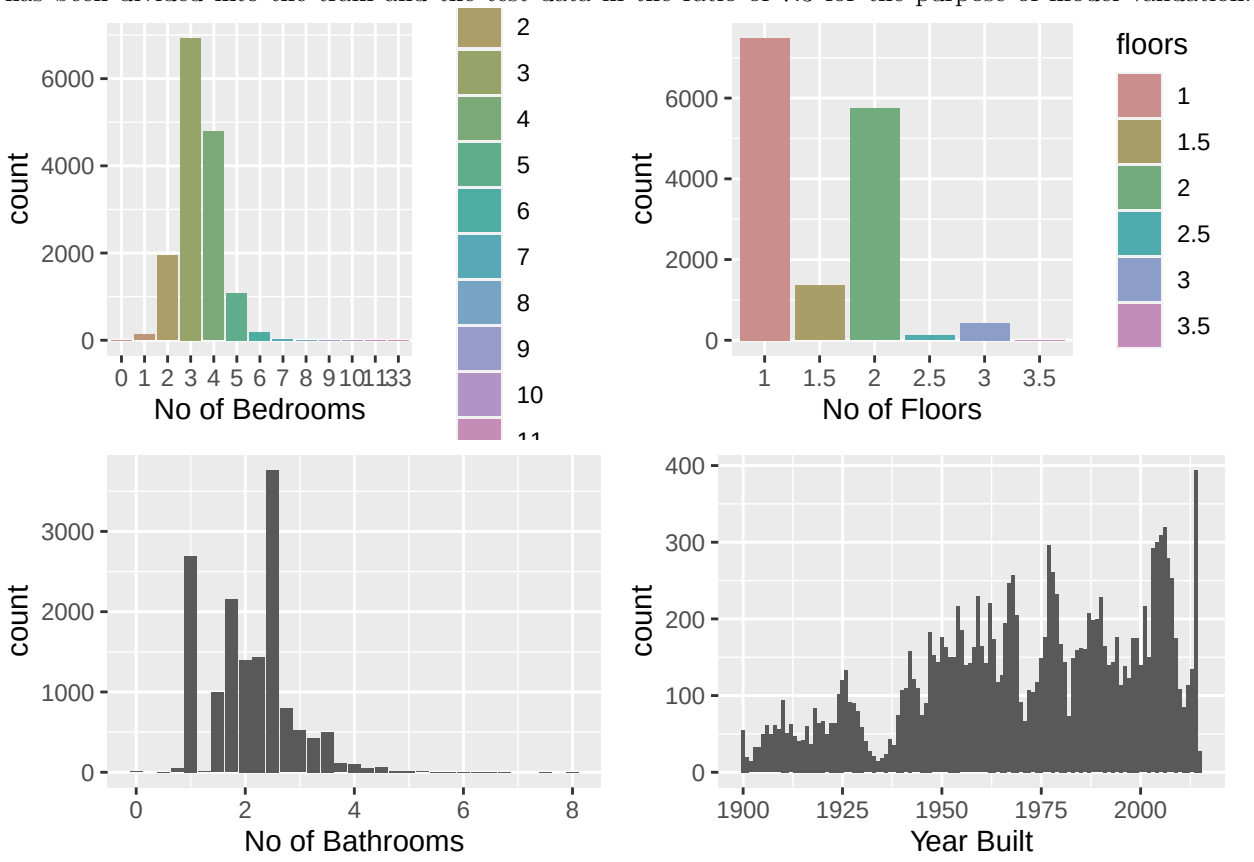
The housing market is one of the most pivotal parts of any public economy. The real estate is one of the most significant investments in a household's portfolio, accounting for the majority of private households' wealth in highly developed countries. Subsequently, perceptions of the housing market and exact forecasts of land costs are supportive for land purchasers and merchants as well as financial subject matter experts. Notwithstanding, land determining is a muddled and troublesome assignment attributable to many immediate and aberrant factors that unavoidably impact the exactness of expectations (De Nadai 2018).As a general rule, factors impacting the real estate costs could be quantitative or subjective. The quantitative factors conceivably incorporate square footage of the apartments interior living space, square footage of the land space, and the square footage of the interior housing space that is above ground level (Chou 2022). The categorical variables include whether the property faces the waterfront view or not, how many bedrooms and bathrooms it has, how many floors the property has, what's the grade type of the property, the type of the view, the condition of the apartment, the year in which it was built, the year in which it was last renovated etc. This paper consists of the multiple linear regression models. For the sake of ease of interpretation, the continuous variables have been used in the models. One dummy variable whether the property has waterfront view or not has also been used in the third model. The three models have approximately same R-squared and adjusted r-squared. There is no detection of multicollinearity and the residuals are normal and with constant variance as shown in the residuals vs fitted plot. None of the models include the location of the property which is a very important variable in determining the cost since it's beyond the scope of this paper.

2 Data

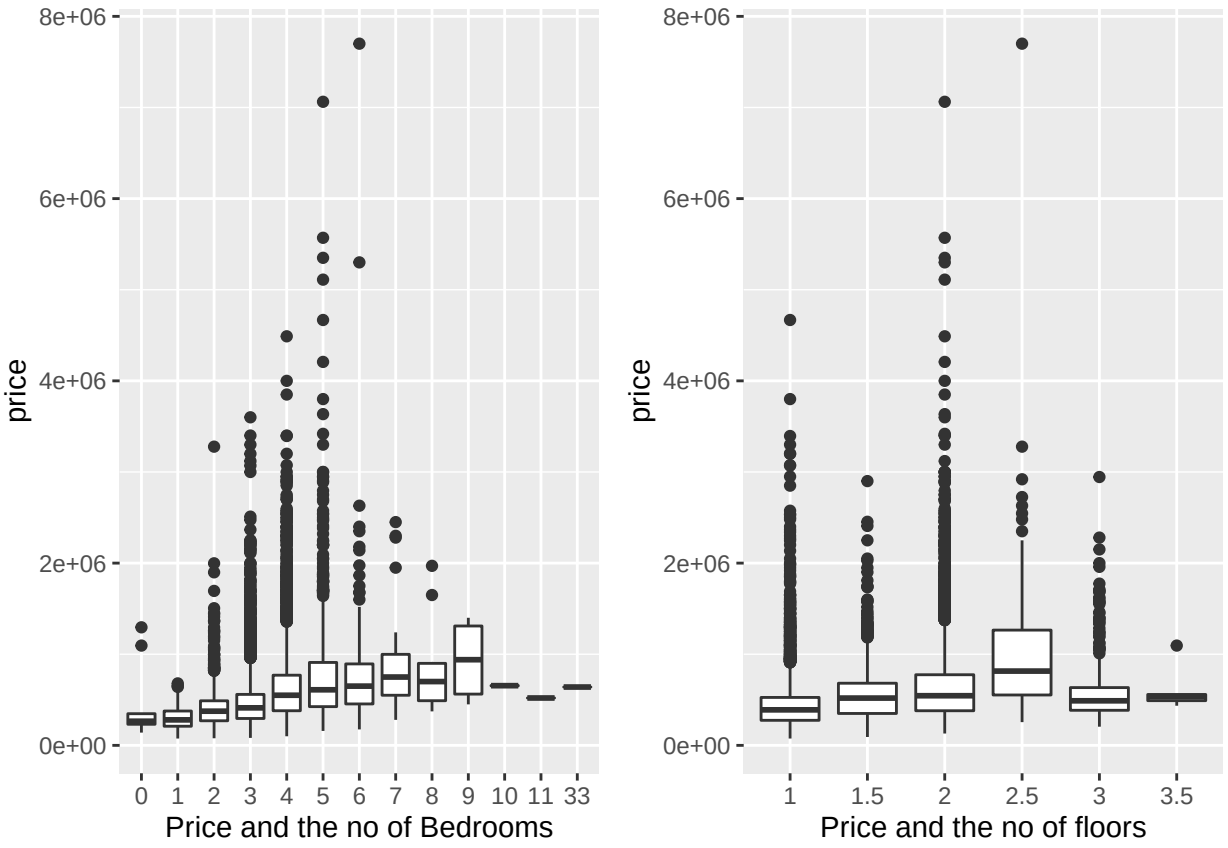
The data set contains the real estate prices of the houses belonging to King County located in the state of United States of Washington. The properties that were sold or bought during May 2014 and May 2015 in

*Code and data are available at: LINK.

the before mentioned location comprises of this data.
The data also contains several other features like the number of bedrooms, number of floors, number of bathrooms where .5 means a toilet without the shower, square footage of the apartments interior living space, square footage of the land space, a dummy variable for whether the apartment was overlooking the waterfront or not, view, condition and the grade of the apartment, square footage of the interior housing space that is above ground level, square footage of the basement, year of the construction and the last renovation, zip code, latitude and longitude, the square footage of interior housing living space for the nearest 15 neighbors and the square footage of the land lots of the nearest 15 neighbors.The data set has been divided into the train and the test data in the ratio of 7:3 for the purpose of model validation.

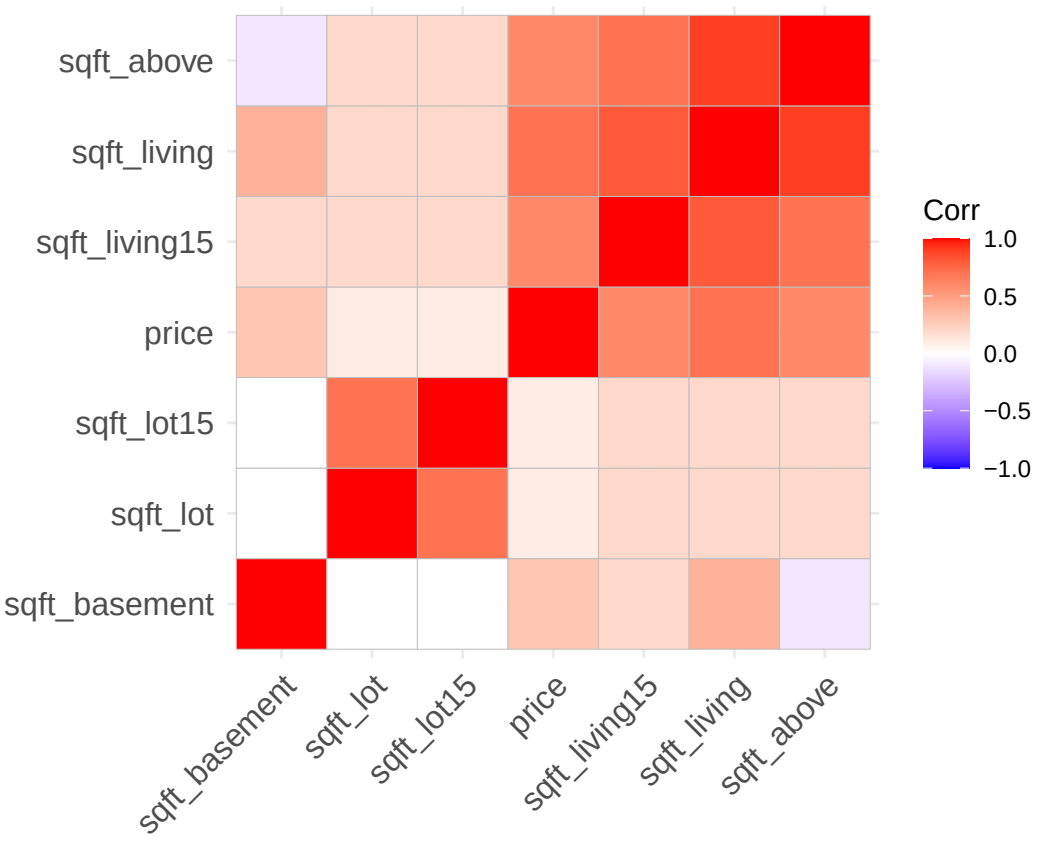


The above bargraphs show the count of houses based on four features: no of bedrooms in the house, no. of floors, no. of bathrooms and the year built. The maximum houses have three bedrooms and the least number of houses have only one bedroom or more than 6 bedrooms. 7486 houses have only one floor whereas only six houses have more than three floors. 24.87% of the houses have 2.5 bathrooms whereas only 6.5% of the houses have 1.5 bathroom. Most of the houses have been built recently whereas only some of the houses have been built more than 100 years ago.The oldest house was constructed in 1900 and the recent most construction happened in 2015.



The houses with eight bedrooms have the highest inter-quartile price range. The most affordable house having eight bedrooms costs USD 373,000 and the most expensive house having eight bedrooms costs USD 1,970,000. However, the most expensive house consists of six bedrooms worth USD 7,700,000.Similarly, the house with 2.5 floors have the highest inter-quartile range. Also, the most expensive house has 2.5 floors.Also, the median price of the house having 2.5 floors is the highest, even greater than the houses having 3 and 3.5 floors.The most expensive house has 2.5 floors and has eight bedrooms in it and costs USD 7,700,000.

The outliers in the box plots suggest the prices of the luxurious houses with the same features. For instance, a property that has five bedrooms costs USD 158,550 and also costs 7,062,500. One of the most important factors that cause a difference in the prices is the location of the property: latitude and longitude.These outliers hold important information about the dataset, hence they are not to be removed from the data set.



This is a correlation plot which suggests that very few variables have high correlation among themselves. Square footage of the apartments interior living space is highly correlated with the square footage of the interior housing space that is above ground level and the square footage of interior housing living space for the nearest 15 neighbors. Remember that high correlation doesn't necessarily indicate the presence of multicollinearity.

3 Model

The multiple linear regression model has been used where the log transformation has been done on the price to reduce it's absolute values otherwise we would get inflated estimates. The log prices have been regressed on square footage of the apartments interior living space. The regression equation is:

$$\log(\text{price}) = \alpha + \beta_1(\text{sqft\_living}) + \epsilon \tag{1}$$

The second model includes two more predictors which are square footage of the land space and the square footage of the interior housing space that is above ground level. The regression equation for the second model is:

$$\log(\text{price}) = \alpha + \beta_1(\text{sqft\_living}) + \beta_2(\text{sqft\_lot}) + \beta_3(\text{sqft\_above}) + \epsilon \tag{2}$$

The third regression model includes a dummy variable whether the property holds a waterfront view or not. The regression equation becomes:

$$\log(\text{price}) = \alpha + \beta_1(\text{sqft\_living}) + \beta_2(\text{sqft\_lot}) + \beta_3(\text{sqft\_above}) + \beta_4(\text{waterfront}_1) + \epsilon \tag{3}$$

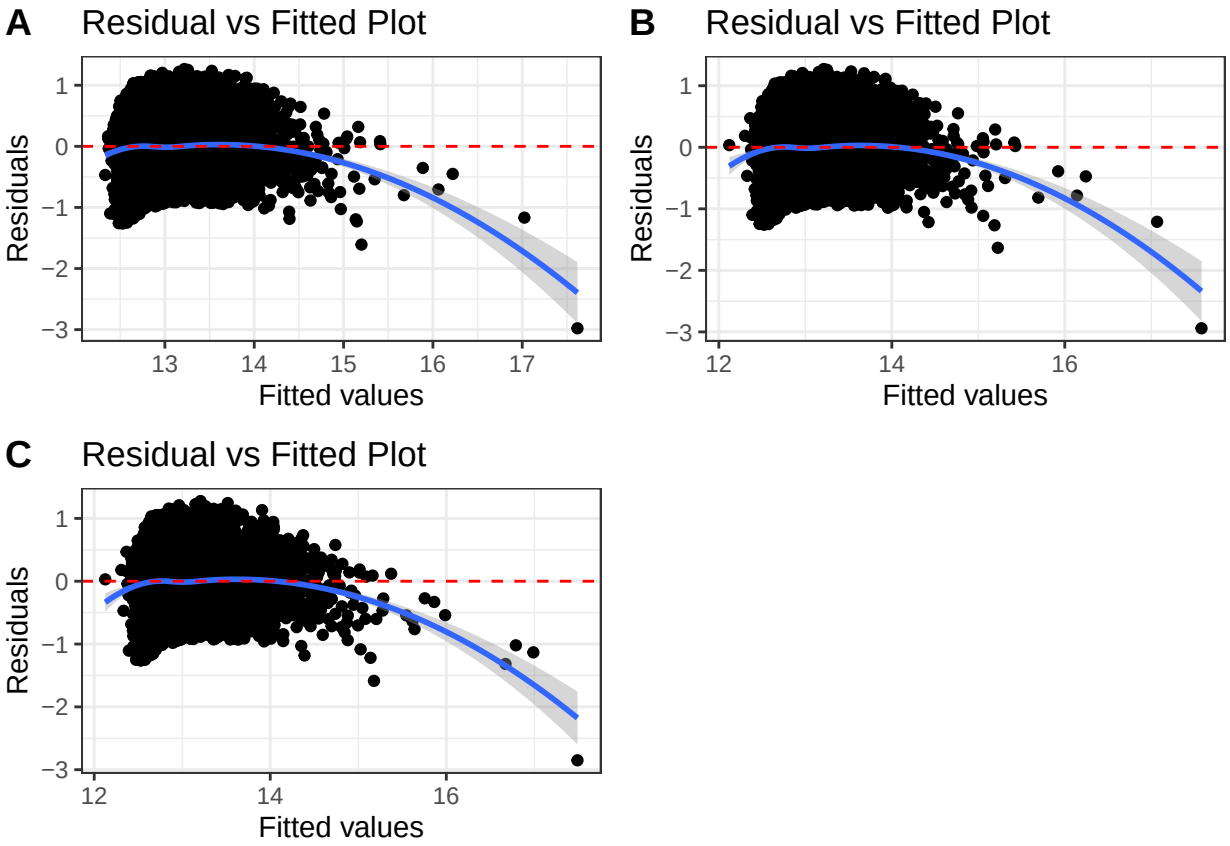
	Model 1	Model 2	Model 3
(Intercept)	12.217	12.219	12.227
	0.008 (0.000)	0.008 (0.000)	0.008 (0.000)
sqft_living	0.000	0.000	0.000
	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
sqft_lot		0.000	0.000
		0.000 (0.000)	0.000 (0.000)
sqft_above		0.000	0.000
		0.000 (0.010)	0.000 (0.089)
waterfront1			0.607
			0.035 (0.000)
Num.Obs.	15 129	15 129	15 129
R2	0.485	0.485	0.495
R2 Adj.	0.485	0.485	0.495
AIC	13 410.6	13 390.1	13 097.1
BIC	13 433.5	13 428.2	13 142.8
Log.Lik.	−6702.314	−6690.046	−6542.533
F	14 222.043	4755.930	3711.404
RMSE	0.38	0.38	0.37

4 Results

The interpretation of the coefficients would be : average change in the log price when there is one unit change in the predictor variable keeping the other predictor variables that are there in the model constant. The intercept means the average log price when all the predictor variables in the model are equal to zero.The table consists of the regression estimates(coefficients), standard errors of the coefficients and their p-values. The p-values for all the coefficients are less than 0.05, hence significant at 5% level of significance for the all the predictor variables for the three regression models. The waterfront variable seems to impact the real estate price the most since it has the largest estimated value in comparison to other variables in the models.

The models have the similar values for the parameters on model selection. There is no significant improvement in the scores as we increase the number of variables from model 1 to model 3. The r-square of Model 1 and 2 is 48.5% where model 3 has a r-square of 49.5%. However, the model 1 has the highest AIC of 13410.6 and model 3 has the least AIC of 13097.1 probably because of two additional variables in the third model which reduces the AIC score. Similarly, the first model has the highest F-score (14222.043) whereas the third model has the smallest F-score (3711.404). However, all the F-scores are significant at 5% level of significance.The root mean squared error (RMSE) is more or less the same for all the three models.

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```



The residual plot for all the three models look the same. There doesn't seem to be any difference in any of the three models despite having the different number of independent variables. The three residuals plots against the fitted values suggest the normality of the residuals and the constant variance of the residuals which abide by the assumptions of simple and multiple linear regression. Hence, we can say that the regression results are reliable.

VIF SCORES FOR MODEL 2

```
## sqft_living    sqft_lot  sqft_above
##        4.36        1.04        4.38
```

VIF SCORES FOR MODEL 3

```
## sqft_living    sqft_lot  sqft_above  waterfront
##        4.40        1.04        4.39        1.02
```

As can be seen from the above table, the VIF scores are less than 10 for both the models which doesn't indicate the presence of multicollinearity in the model.

5 Discussion

The property prices are significantly depended upon the location of the property: whether it is in a rural area or a posh urban area; latitudes and longitudes to be very precise. However, including latitudes and longitudes or the considering the location of the property is beyond the scope of this paper.

The data set also includes the dates when a property was sold. Some of the properties have been sold multiple times and hence included in the data set repetitively. Since, the time factor also fluctuates the prices, the time series regression can be executed on the model which might improve the prediction accuracy.

The variable selection can be done through the feature selection algorithms like principal component analysis, or lasso or ridge regression can also be executed to get the significant variables in the model.

Appendix

A References

Chou, JS., Fleshman, DB. & Truong, DN. Comparison of machine learning models to provide preliminary forecasts of real estate prices. J Hous and the Built Environ (2022).

M. De Nadai and B. Lepri, "The Economic Value of Neighborhoods: Predicting Real Estate Prices from the Urban Environment," 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 2018, pp. 323-330, doi: 10.1109/DSAA.2018.00043.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. Ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." Journal of Open Source Software 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. Dplyr: A Grammar of Data Manipulation. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2015. Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963