

House Price Prediction with Big Data

Ziyu Cui^{1, a, *}

¹ Santa Clara University, Santa Clara, CA 95053, United States

a. zcui2@scu.edu

*corresponding author

Abstract: Many of the current models for house price prediction focus on the house price index itself or include too few factors. In this research, however, different aspects of housing properties (bedrooms, bathrooms, grade, view, etc.) will be incorporated to generate a more accurate prediction model. Doing so not only remedies for a lack of variables in current models but also benefits the consumers by providing an accurate estimation of housing values. This paper examines the estimation methodology of a multiple linear regression model. To get the optimal prediction power and keep the model as simple as possible, different variable combinations will be tested. Through comparing different regression models and analyzing the regression results, the predictive model introduced in this paper has a high house price prediction power. This paper provides a potential solution for the prediction of the house price in the King County, Washington.

Keywords: House Price Prediction, Data Analysis, Linear Regression, Predictive Model, Pricing

1. Introduction

House price prediction has always been an interesting area to do research on. There has been a lot of research done on this topic and new research keeps emerging these days. The decision on this topic was inspired by the recent shifts in house prices in California due to the pandemic. House prices had been dramatically surging and plunging in the last few months and it must be essential for consumers and investors to accurately estimate the actual value of a property. Thus, in the paper, the author decided to construct a model that can help people to estimate the real value of a house based on factors such as property characteristics and locations. The approach was inspired by Zestimate, which uses various variables and a neural network approach to provide an estimation of house value to consumers. According to the paper *Zillow's Estimates of Single-Family Housing Values*, the prediction model could be able to produce an estimation within 10% of the actual transaction value [1].

2. Datasets and Variables

To discover the main factor that determines the house price, the author selected the dataset including 21,613 King County house prices with house characteristics from May 2014 to May 2015. Investigating the correlation between house values and house characteristics, the paper will be looking into each housing feature's correlation with house market values [2].

Bedrooms: the number of bedrooms.

Bathrooms: the number of bathrooms, where 0.5 refers to bathrooms with no showers.

Sqftliving: the square footage of the house.

Sqftlot: the square footage of the land space.

Floors: the number of building floors.

Waterfront: 1 if the house has a waterfront view and 0 otherwise.

View: an index from 0 to 4. 0 indicates the worst view and 4 indicates the best view.

Condition: an index from 1 to 5 on the condition of the apartment.

Grade: an index from 1 to 13, where a score of 1-3 indicates the building is short of construction and design, 7 has an average level of construction and design, and 11-13 has a high-quality level of construction and design.

Sqftabove: the square footage of the house without counting the basement.

Sqftbasement: the square footage of the basement.

Yr-built: the year the house was built.

Yr-renovated: the year of the house's last renovation.

Zipcode: the zip code of the house.

Sqftliving15: the living space size in 2015 (after innovation).

Sqftlot15: the land space area in 2015 (after innovation).

2.1. Descriptive Table

To perform further analysis of the data, since there are both continuous and quantitative data in the data set, it is essential to present the data in the form of a descriptive table to visualize the range of each variable. A descriptive table is also a great tool to examine if there exists any extreme outlier or abnormal variable value [3].

Table 1 presents a total of 21,612 pieces of housing information, indicating a large enough dataset to draw further conclusions on the relationship between house prices and other housing features. Removing one house with 33 bedrooms reduces potential extreme outlier bias since the median of the bedrooms is 3 with a standard deviation of 0.9. From the information, it is important to notice that the land space (sqftlot) has a large standard deviation of 41,421, with a minimum size of 520 and a maximum size of 1,651,359. With the most bathroom counts being 8, the standard deviation is 0.8 with a mean of 2.1. Finally, the mean value is 7.7 for grade and 3.4 for condition. Two numbers are close to the median value of the two features, which means that the data collected have included a fair amount of good and bad design, as well as houses in good and poor conditions. On the one hand, such information shows that the dataset includes a variety of house types. On the other hand, it also alerts people to the possibility that the effectiveness of the regression model would be downplayed by the presence of extreme outliers.

Table 1. King county house data (continue).

Statistic	N	Mean	St. Dev.	Min	Median	Max
Price	21,612	540,177.5	367,370.1	75,000	450,000	7,700,000
Bedrooms	21,612	3.4	0.9	0	3	11
Bathrooms	21,612	2.1	0.8	0.0	2.2	8.0
Sqftlot	21,612	15,107.4	41,421.4	520	7,619	1,651,359
Floors	21,612	1.5	0.5	1.0	1.5	3.5
Waterfront	21,612	0.01	0.1	0	0	1
View	21,612	0.2	0.8	0	0	4
Condition	21,612	3.4	0.7	1	3	5

Table 1. (continued).

Grade	21,612	7.7	1.2	1	7	13
Yr-built	21,612	1971.0	29.4	1,900	1,975	2,015
Yr-renovated	21,612	84.4	401.7	0	0	2,015
Sqftliving15	21,612	1986.6	685.4	399	1,840	6,210
Sqftlot15	21,612	12,768.8	27,304.8	651	7,620	871,200

2.2. Housing Distribution

To better understand the distribution of the dataset, the paper plotted the houses by longitude and latitude. From the map shown in Figure 1, it can be clearly observed that the homes in the dataset are spread out across the King County.

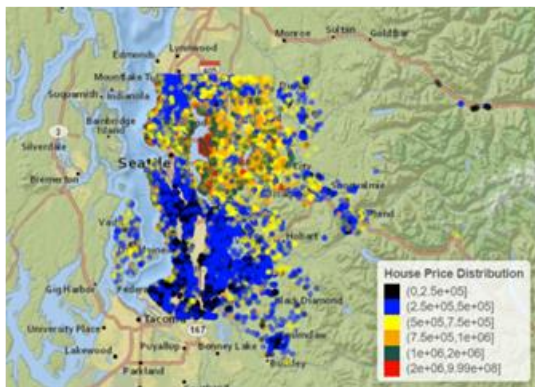


Figure 1. King county house distribution [4].

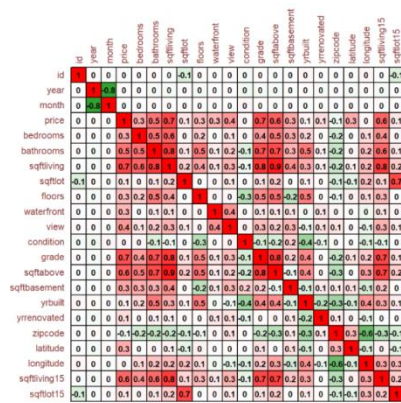


Figure 2. Variable correlation table [5].

2.3. Correlation Table

From Figure 2, it is important to notice that bathrooms, sqftliving, grade, sqftabove, and sqftliving15 have a relatively higher correlation with the price factor, therefore, it would be reasonable to focus on these factors in further analysis.

2.4. Plot Graph for Variables in the Base Model

From Figure 3, there exists a strong and positive correlation between the house sales price and the square footage of the living space, as the line of best fit trends upwards and the data points follow the pattern. In addition, it is also important to notice that most data points are located on the left side of the graph and there are a few extreme outliers spread around the graph. For instance, there are three points located around 10,000 square footage of living space with prices around and above 7,000,000 dollars, which could cause potential bias in the regression results. As seen in Figure 4, there also exists a clear pattern of increase in the house price as the index of building construction design increases. Therefore, it is reasonable to conclude that there is a positive correlation between the grade and the house selling price.

According to Figure 5, there exists a positive correlation between the house market value and the square footage above the basement. Like the correlation graph between price and square footage of living space shown in Figure 3, it is important to pay attention to the three outliers located around

8500 square feet with house price over 7,000,000 dollars. Plotting bathroom variables as shown in Figure 6 results in a positive correlation between the house price and the number of bathrooms in the building.

As seen in Figure 7, there exists a positive correlation between the view index and the house price. However, this correlation is more gradual compared to the previous one. According to Figure 8, there is a positive correlation between the number of bedrooms and the house sales price.

Finally, as shown in Figure 9, for the binary waterfront variable, houses with a waterfront view have a relatively higher price than those without a waterfront view.

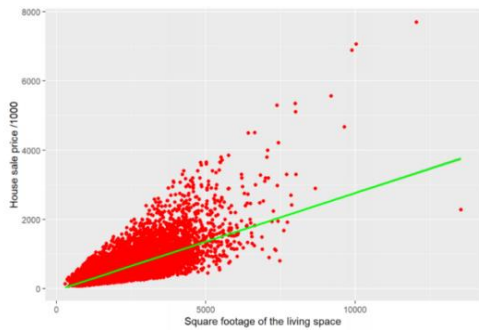


Figure 3. Correlation between house price and living space area.

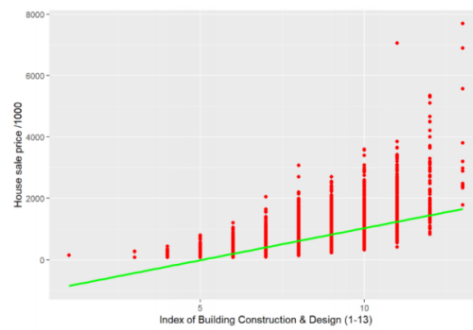


Figure 4. Correlation between house price and grade index.

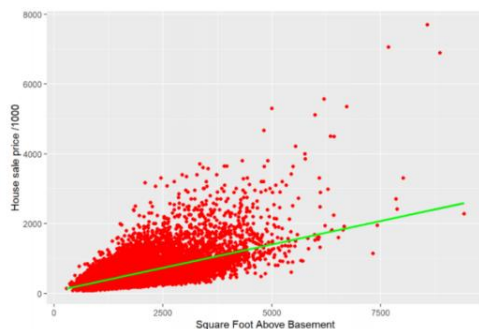


Figure 5. Correlation between house price and area above basement.

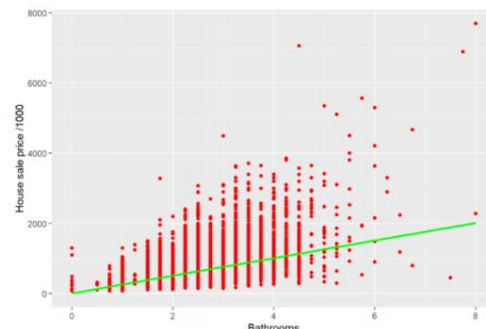


Figure 6. Correlation between house price and number of bathrooms.

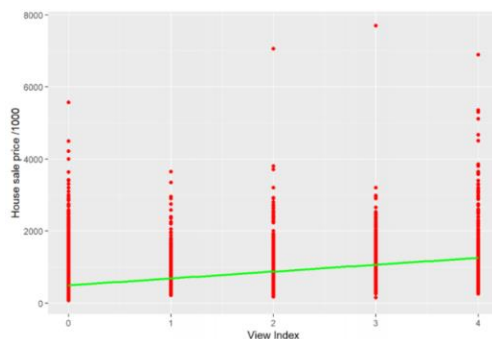


Figure 7. Correlation between house price and view index.

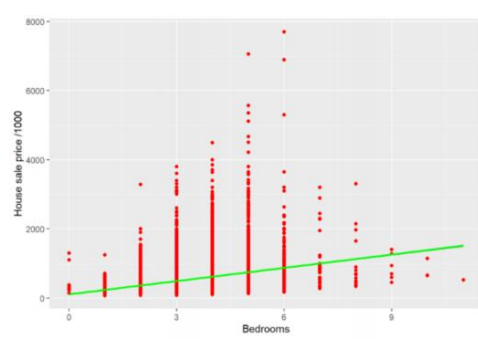


Figure 8. Correlation between house price and number of bedrooms.

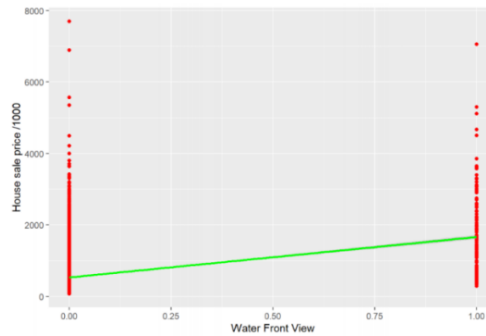


Figure 9. Correlation between house price and waterfront view.

3. Model Selection

After observing all the variables of interest that have a positive correlation with the house price, it would be reasonable to conduct further analysis. The model used to analyze the correlation between the independent variable (house price) and the dependent variables will be linear regression.

$$Y_i = \beta_0 + \beta_1 X_i \quad (1)$$

The final model of this research includes multiple independent variables to predict the dependent price variable. Thus, a multiple regression model will be introduced to explain a more accurate estimation [6].

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_i X_i \quad (2)$$

4. Result and Analysis

4.1. First Regression Analysis

The first regression table (Table 2) shows the correlation between the independent variables and the house price.

Table 2. First regression table.

	Dependent Variables: Price					
Sqftliving	184.528	234.822	252.69	209.106	233.108	233.774
Grade	98,638.39	110,076.6	115,149.2	104,006.9	97,802	110,119.9
Sqftabove		-77,762	-76.931	-41.49	-44.962	-49.574
Bathrooms			-37,098.19	-29,907.69	-19,956.71	-20,574.48
View				92,223.54	88,606.23	61,595.69
Bedrooms					-35,297.97	-32,218.98
Waterfront						585,038
Constant	-598,890.8	-652,008.5	-651,045.8	-575,276	-472,740	-490,775.9
Observations	21,612	21,612	21,612	21,612	21,612	21,612
R2	0.535	0.542	0.544	0.576	0.581	0.596
Adjusted R2	0.535	0.541	0.544	0.576	0.588	0.596
Residual Std. Error	250,646.3	248,880.9	248,221.2	239,233	237,944.2	233,409.7
F Statistic	12,408.42	8,492.936	6,433.566	5,871.054	4,984.943	4,561.693

4.1.1. Regression 1

Holding other variables the same, one additional square foot increase in the living space would increase the house value by 184.528 dollars. One additional grade index would increase the house value by 98,638.39 dollars.

4.1.2. Regression 2

Holding other variables the same, one additional square foot increase in the living space would increase the house value by 234.822 dollars. One additional grade index would increase the house value by 110,076.6 dollars. One additional square foot of the building apart from the basement would decrease the house value by 77.762 dollars.

4.1.3. Regression 3

Holding other variables the same, one additional square foot increase in the living space would increase the house value by 252.69 dollars. One additional grade index would increase the house value by 115,149.2 dollars. One additional square foot of the building apart from the basement would decrease the house value by 76.931 dollars. One additional bathroom count would decrease the house value by 37,098.19 dollars.

4.1.4. Regression 4

Holding other variables the same, one additional square foot increase in the living space would increase the house value by 209.106 dollars. One additional grade index would increase the house value by 104,006.9 dollars. One additional square foot of the building apart from the basement would decrease the house value by 41.49 dollars. One additional bathroom count would decrease the house value by 29,907.69 dollars. One additional view index would increase the house value by 92,223.54 dollars.

4.1.5. Regression 5

Holding other variables the same, one additional square foot increase in the living space would increase the house value by 233.108 dollars. One additional grade index would increase the house value by 97,802 dollars. One additional square foot of the building apart from the basement would decrease the house value by 44.962 dollars. One additional bathroom count would decrease the house value by 19,956.71 dollars. One additional view index would increase the house value by 88,606.23 dollars. One more bathroom in the house would decrease the house price by 35,297.97 dollars.

4.1.6. Regression 6

Holding other variables the same, one additional square foot increase in the living space would increase the house value by 233.774 dollars. One additional grade index would increase the house value by 100,119.9 dollars. One additional square foot of the building apart from the basement would decrease the house value by 49.574 dollars. One additional bathroom count would decrease the house value by 20,574.48 dollars. One additional view index would increase the house value by 61,595.69 dollars. One more bathroom in the house would decrease the house price by 32,218.98 dollars. A house with a waterfront view would raise the house value by 585,638 dollars. It is important to note that all the variables included in the first regression table are statistically significant, which means that it is reasonable to reject the null hypothesis and conclude that there is a strongly positive or negative correlation between the dependent variables and the price. Of all the regression models,

however, it is important to note that the optimal R2 value is only 0.596, meaning that only 59.6 percent of the variation in the dependent variable can be explained by the independent variables. As a result, the author decided to further develop the 6th regression model since it explains the most variance in price.

4.2. Second Regression Analysis

For the second regression table (Table 3), the author has replaced the variables of sqftabove, bedrooms, and bathrooms since they have a limited impact on the R2. The paper added sqftliving15 and yr-built. In addition, it can be clearly observed from Figure 1 that houses located in the northwest generally have a higher price. According to the paper House Prices and Relative Location, the walking distance to key locations such as parks, schools, and hospitals has a high explanatory power on house prices, thus it would be reasonable to consider latitude and longitude to the regression analysis [7,8].

Table 3. Second regression table.

Dependent Variables: Price				
Sqftliving	164.751	159.189	166.488	0.0002
Grade	93,850.1	133,157.4	108,952.9	0.174
Sqftliving15	3.447	14.039	24.641	0.0001
View	70,447.32	47,477.68	50,664.2	0.061
Waterfront	590,653.6	597,910.4	611,701.9	0.39
Yr-built		-3,284	-2,395.714	-0.003
Latitude			557,820.6	1.354
Longitude			-108,761.3	-0.058
Constant	-548,902.3	5,618,776	-35,805,938	-54.527
Observations	21,612	21,612	21,612	21,612
R2	0.589	0.642	0.686	0.753
Adjusted R2	0.589	0.642	0.685	0.753
Residual Std. Error	235,598.9	219,898.3	206,054	0.262
F Statistic	6,187.91	6,451.975	5,886.408	8,237.456

4.2.1. Regression 1

Holding other variables the same, one additional square footage of living space would increase the house price by 164.751 dollars. One more grade index would increase the house market value by 93,850.1 dollars. One additional square footage of living in 2015 would raise the house price by 3.447 dollars. One more view index would increase the house price by 70,447.32 dollars. A house that has a waterfront view would be worth 590,653.6 dollars more than those that do not.

4.2.2. Regression 2

Holding other variables the same, one additional square footage of living space would increase the house price by 159.189 dollars. One more grade index would increase the house market value by 133,157.4 dollars. One additional square footage of living in 2015 would raise the house price by 14.039 dollars. One more view index would increase the house price by 47,477.68 dollars. A house that has a waterfront view would be worth 597,910.4 dollars more than those that do not. One more later the house was built would decrease the house value by 3,284 dollars.

4.2.3. Regression 3

Holding other variables the same, one additional square footage of living space would increase the house price by 166,488 dollars. One more grade index would increase the house market value by 108,952.9 dollars. One additional square footage of living in 2015 would raise the house price by 24,641 dollars. One more view index would increase the house price by 50,664.2 dollars. A house that has a waterfront view would be worth 611,701.9 dollars more than those that do not. One more later the house was built would decrease the house value by 2395.714 dollars. One more degree increase in latitude would increase the house value by 557,820.6 dollars. One more degree increase in longitude would decrease the house price by 108,761.3 dollars.

The impact of coordinates can be explained in Figure 1. It can be clearly observed that houses to the northwest are generally more expensive than the surrendering area.

4.2.4. Regression 4

For this last regression, the author has taken the log of price. Doing so would reduce the lack of generalization of the dependent variable, normalizing it so that it would be easier to draw a correlation. The adjusted R² for this regression is 0.753, which means that 75.3% of the variance in price has been explained by this regression model. With all the independent variables being statistically significant, this paper decided to accept it as the final regression model.

5. Conclusion

The Final Model of King County House Price Prediction:

$$\log(\text{price}) = -54.527 + (0.0002)\text{sqftliving} + (0.174)\text{grade} + (0.0001)\text{sqftliving15} + (0.061)\text{view} + (0.39)\text{waterfront} - (0.003)\text{yr - built} + (1.354)\text{latitude} - (0.058)\text{longitude}$$

The model means that while holding other variables constant, one additional square footage of living space would increase the house price by 0.02%. One additional grade index would increase the house price by 17.4%. One more square footage of living space in 2015 would increase the house value by 0.01%. One additional view index would increase the house price by 6.1%. A house with a waterfront view would raise the property value by 39%. One year increase in the year built would decrease the house price by 0.3%. One degree increase in latitude would raise the house price by 135.4% (1 degree latitude ~ 69 miles). One degree increase in longitude would decrease the house value by 5.8% (1 degree longitude ~ 54.6 miles). According to the regression result, it is surprising to see how the property location can have such a huge impact on the house market value. One degree latitude increase would raise the property value by 135.4%, which means that one additional mile to the north would raise the house price by 1.96%. The waterfront variable is one of the most significant variables economically. Having a waterfront view makes a 39% difference in the house value. Grade is the second economically significant factor; each level makes a 17.4% difference. Since there are 13 levels of grade, the difference between a low grade and a high grade can easily cause a larger effect on the price than the waterfront does.

However, it is important to realize that the adjusted R² for the final model is 0.753, meaning that 75.3% of the variation of the house price is explained by the model and 24.7% of the price variations are undermined by the variables. This is caused by the omitted variable bias, meaning that some confounding variables are not included in the regression model. As a result, the prediction made by this model may not be exactly accurate and should be used only for basic reference but not for detailed analysis. Most importantly, the external validity of the house price prediction model is low since the dataset only included the housing data in King County, WA. If the prediction of this model is applied to other states or cities, the result may not be accurate due to economic, cultural, and geographic

factors. For instance, if the model is to predict the house price for an inland city in Texas, the results will not be as accurate since there would be very few houses with a waterfront view.

References

- [1] Hollas, D., et al.: *Zillow's Estimates of Single-Family Housing Values: Semantic Scholar*. *Zillow's Estimates of Single-Family Housing Values. Semantic Scholar*. (1970). <https://www.semanticscholar.org/paper/Zillow-%E2%80%99-s-Estimates-of-Single-Family-Housing-Hollas-Rutherford/ca1a7e03f08380dce18ff5f766f1bd4301a42201#paper-header>.
- [2] Harlfoxem. *House Sales in King County, USA*. Kaggle. (2016). <https://www.kaggle.com/datasets/harlfoxem/housesalesprediction>.
- [3] Spriestersbach, A., et al.: *Descriptive Statistics: The Specification of Statistical Measures and Their Presentation in Tables and Graphs. Part 7 of a Series on Evaluation of Scientific Publications. Deutsches Arzteblatt International, Deutscher Arzte Verlag*. (2009). <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2770212/>.
- [4] Ambarish, B.: *Tutorial Housesales Kingcounty EDA&Modelling*. Kaggle. (2018). <https://www.kaggle.com/code/ambarish/tutorial-housesales-kingcounty-eda-modelling/report>.
- [5] Dhakre, A.: *EDA: Linear Regression: K-Fold CV: ADJ R2=0.87*. Kaggle. (2017). <https://www.kaggle.com/code/amitdhakre13/eda-linear-regression-k-fold-cv-adj-r2-0-87>.
- [6] Marill, K. A.: *Advanced Statistics: Linear Regression, PART II ... - Wiley Online Library*. Wiley Online Library. <https://onlinelibrary.wiley.com/doi/pdf/10.1197/j.aem.2003.09.006>.
- [7] Heyman, A. V., Sommervoll, D. E.: *House Prices and Relative Location*. ScienceDirect. (2019). <https://www.sciencedirect.com/science/article/pii/S0264275118312241>.
- [8] Rivas, R., Patil, D., Hristidis, V. et al.: *The impact of colleges and hospitals to local real estate markets*. *J Big Data* 6, 7 (2019). <https://doi.org/10.1186/s40537-019-0174-7>.