House Prices Prediction Report

ALY 6010 Probability Theory and

Introductory Statistics Project

Adrian An

12-14-2022

Introduction

There has been significant growth in real estate over the years. Haggins et al., 2019 elaborated that real estate was one of the most profitable ventures for investment. This study aims to predict the house price using multiple linear regression in King County, Washington State, US. Multiple linear regression is used to establish whether there is a significant relationship between one dependent variable and several independent variables. I evaluated whether there is a statistically significant relationship between the dependent and independent variables at a 5% significance ($\alpha = .05$). The data was obtained online from the Kaggle website. I obtained the first six rows of each column of the data to have a general understanding of the data. I used graphs such as boxplots to check for outliers to clean the data. The outliers in this dataset were acceptable even if they are extreme since there is the possibility of such observations. The outliers were not deleted but kept as part of the data. The data contained 21 columns and 21613 rows.

The null hypothesis states that there is no statistically significant relationship between the price of the house and the independent variables.

H0: $\beta i = 0$

The alternative hypothesis states that there is a statistically significant relationship between the price of the house and the independent variables.

H1: $\beta i \neq 0$

```
##           id            date   price bedrooms bathrooms sqft_living sqft_l
ot
## 1 7129300520 20141013T000000  221900        3      1.00        1180     56
50
## 2 6414100192 20141209T000000  538000        3      2.25        2570     72
42
## 3 5631500400 20150225T000000  180000        2      1.00         770    100
00
```

```
## 4 2487200875 20141209T000000  604000        4      3.00         1960     50
00
## 5 1954400510 20150218T000000  510000        3      2.00         1680     80
80
## 6 7237550310 20140512T000000 1225000        4      4.50         5420   1019
30
##   floors waterfront  view condition  grade sqft_above sqft_basement  yr_built
## 1      1          0    0         3     7       1180             0     1955
## 2      2          0    0         3     7       2170           400     1951
## 3      1          0    0         3     6        770             0     1933
## 4      1          0    0         5     7       1050           910     1965
## 5      1          0    0         3     8       1680             0     1987
## 6      1          0    0         3    11       3890          1530     2001
##   yr_renovated zipcode     lat     long sqft_living15  sqft_lot15
## 1            0   98178 47.5112 -122.257          1340        5650
## 2         1991   98125 47.7210 -122.319          1690        7639
## 3            0   98028 47.7379 -122.233          2720        8062
## 4            0   98136 47.5208 -122.393          1360        5000
## 5            0   98074 47.6168 -122.045          1800        7503
## 6            0   98053 47.6561 -122.005          4760      101930
```

Variable description

Price refers to the price of the house in dollars. Bedrooms refer to the number of bedrooms contained in a house. Bathrooms refer to the number of bathrooms included in a house.Yr_built refers to the year the house was built. sqft_living refers to the living area of the house in square feet. Grade refers to a score given according to the quality of the house.

This technique is employed to investigate whether there is a relationship between one dependent variable and several independent variables.

Data Analysis

- **Measures of Central Tendency**

I obtained the mean, first quartile, median, mean, third quartile, and a maximum of all the columns in the data. Mean is mainly used when the data is approximately normally distributed.

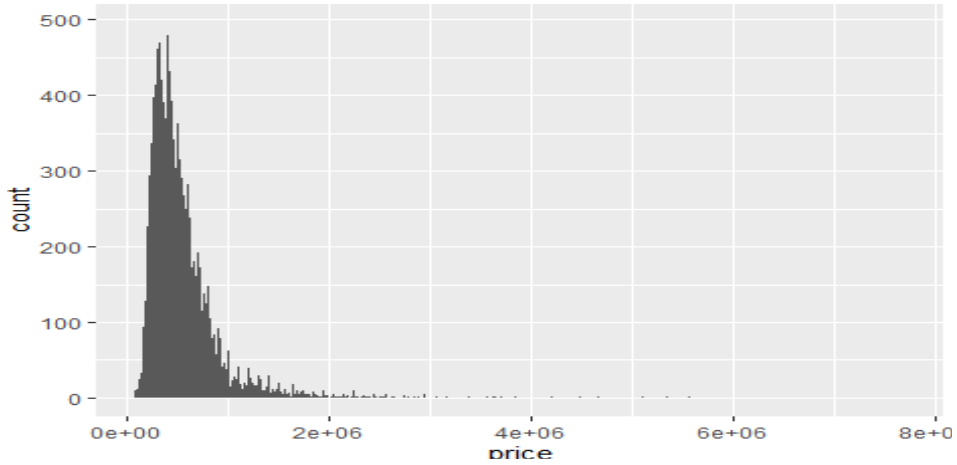
```
##      price            bedrooms        bathrooms      sqft_living
##  Min.   :  75000   Min.   : 0.000   Min.   :0.000   Min.   :  290
##  1st Qu.: 321950   1st Qu.: 3.000   1st Qu.:1.750   1st Qu.: 1427
##  Median : 450000   Median : 3.000   Median :2.250   Median : 1910
##  Mean   : 540088   Mean   : 3.371   Mean   :2.115   Mean   : 2080
##  3rd Qu.: 645000   3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.: 2550
##  Max.   :7700000   Max.   :33.000   Max.   :8.000   Max.    13540
##     sqft_lot           floors        waterfront            view
##  Min.   :    520   Min.   :1.000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.:   5040   1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median :   7618   Median :1.500   Median :0.000000   Median :0.0000
##  Mean   :  15107   Mean   :1.494   Mean   :0.007542   Mean   :0.2343
##  3rd Qu.:  10688   3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.0000
##  Max.    1651359   Max.   :3.500   Max.   :1.000000   Max.   :4.0000
##      grade
##  Min.   : 1.000
##  1st Qu.: 7.000
##  Median : 7.000
##  Mean   : 7.657
##  3rd Qu.: 8.000
##  Max.   :13.000
```

- **Price**

The minimum price of the house in King County was $75,000, while the maximum price was $7,700,000. The mean price of the houses was $540,088. The range of the price of the house was considerably huge. Therefore, the mean didn't give a representative measure of central tendency.

Figure 1.

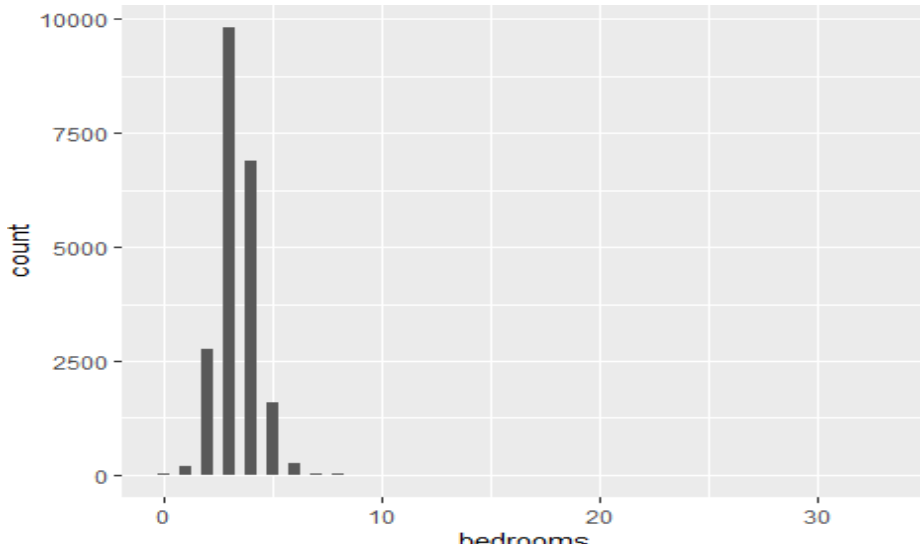
Price histogram



The histogram above was positively skewed. Therefore only a few houses were highly priced compared to other homes.

Figure 2.

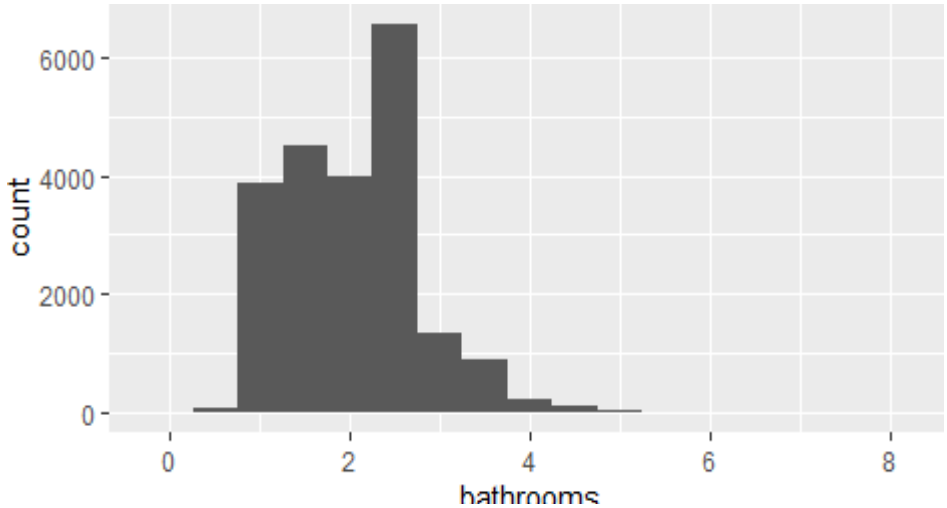
The number of bedrooms histogram.



The number of bedrooms in a house was approximately normally distributed. We expect that the more the number of bedrooms a home has, the more the price of the house should be.

Figure 3.

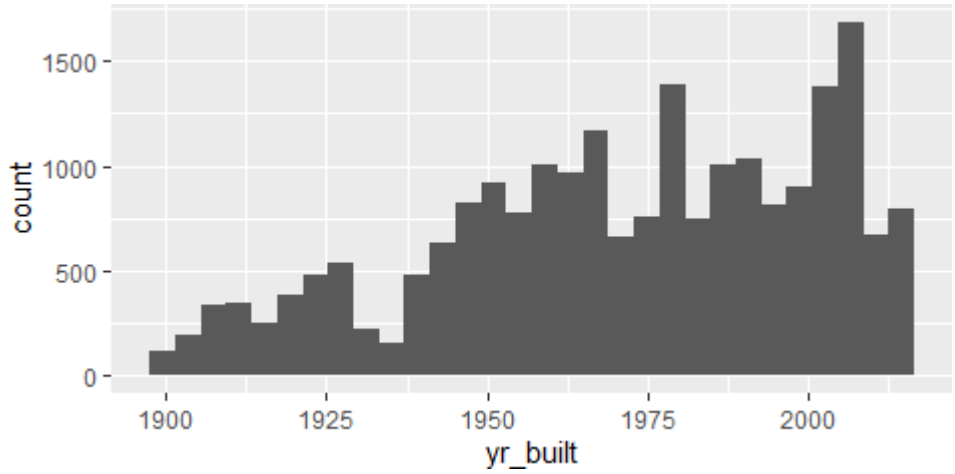
The number of bathrooms histogram.



The number of bathrooms was positively skewed, meaning only a few houses had many bathrooms. We expect that the more the number of bathrooms, the more a house should cost.

Figure 4.

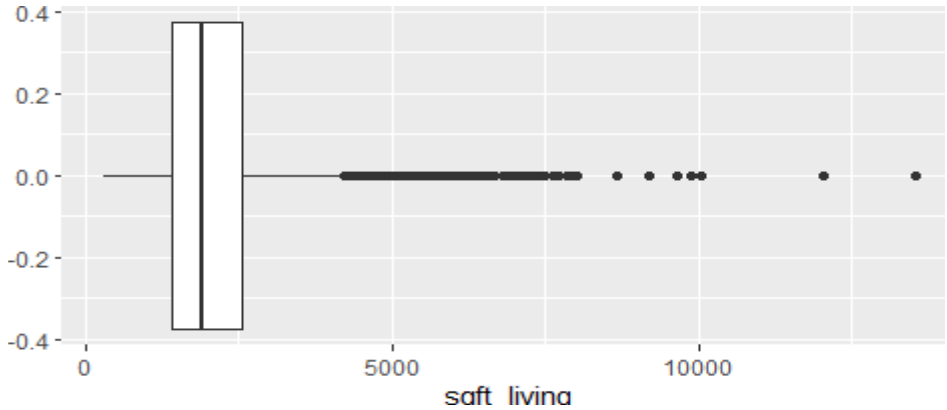
A histogram of the years the houses were built.



The year the house was built was negatively skewed. Therefore, there were few houses built earlier compared to many built recently. We expect the houses built recently to be more expensive than those built earlier, which are considered old, except for unique homes such as traditional castles with historical significance.

Figure 5.

House living area in square feet boxplot.

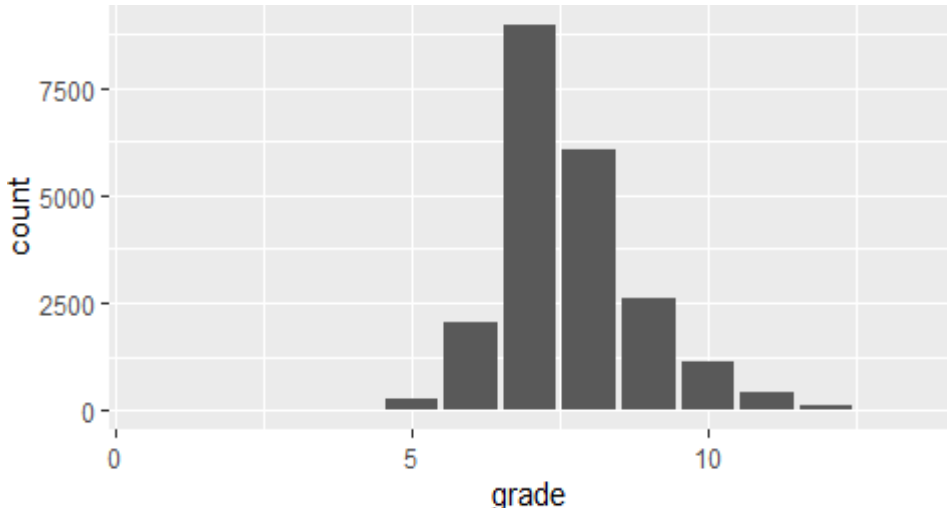


The living area of the houses was not approximately normal, and it was positively skewed. The boxplot above also indicates that there were some extreme values. We leave the outliers since it's possible to have a house with more than 5000 square feet in the living area. We expect a place

with a large living area to be expensive compared to a home with a smaller living area (Lee, 2016).

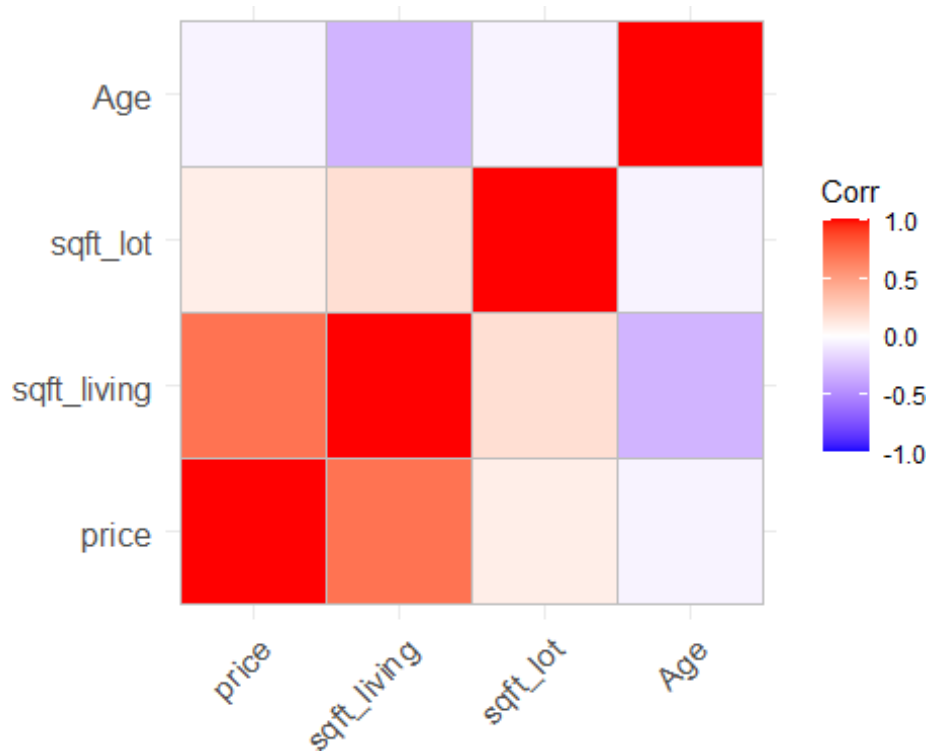
Figure 6.

Grade of the house bar plot



Each house was graded according to the King County grading system, where one implies a poor grade and 13 denotes excellent quality. We expect the higher the grade; the more expensive a house should cost.

Correlation Analysis



The correlation analysis was conducted to establish the relationship between variables. The variables obtained from the data entailed price, square foot, square lot, and age of the property. A correlation chart with four variables was generated to show the strength of the relationship. The reason for using a few variables is to avoid overlapping and unclear displays of the correlation coefficient. The analysis results show that the price of the house and square foot had a strong positive correlation (r=0.7). Relationship between other variables exhibited a very weak relationship. Their correlation coefficient was lower than 0.3.

Multiple linear regression.

This technique is employed to investigate whether there is a relationship between one dependent variable and several independent variables.

$$\hat{\mathbf{y}} = a + b_i x_i$$

$$\hat{\mathbf{y}} = a + b_1 x_1 + b_2 x_2 + \ldots + \ldots b_n x_n$$

Where;

$\hat{y}$ is the dependent variable

a is the intercept of the model

bi is the coefficient of the independent variable

xi is the independent variable.

```
##
## Call:
## lm(formula = price ~ bathrooms + sqft_living + bedrooms  + grade +
##     yr_built)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1346944  -117295   -12142    90730  4432536
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.204e+06  1.219e+05   59.12   <2e-16 ***
## bathrooms    5.964e+04  3.473e+03   17.17   <2e-16 ***
## sqft_living  1.830e+02  3.356e+00   54.54   <2e-16 ***
## bedrooms    -4.789e+04  2.108e+03  -22.72   <2e-16 ***
## grade        1.324e+05  2.201e+03   60.14   <2e-16 ***
## yr_built    -4.071e+03  6.416e+01  -63.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 227800 on 21607 degrees of freedom
## Multiple R-squared:  0.6151, Adjusted R-squared:  0.615
## F-statistic:  6905 on 5 and 21607 DF,  p-value: < 2.2e-16
```

The model is statistically significant in predicting the house price, $F(5,21607) = 6905$, $p < 0.05$.

R2 is used to indicate the amount of variation that is explained by the independent variable (Madhuri et al., 2019). The model accounted for 61.51% of the variation in house prices.

House price = 7.204e+06 + 5.964e+04 (bathrooms) + 1.830e+02 (sqft_living) -4.789e+04 (bedrooms) + 1.324e+05 (grade) -4.071e+03 (yr_built)

The intercept means when all the other independent variables are zero, the house price is 7. 204e+06. An increase in the number of bathrooms by one unit resulted in an increase in the house price by 5.964e+04 dollars. The relationship between the piece of the house and the number of bathrooms was as expected. The variable number of bathrooms in the house statistically significantly predicted the price of the home, $p < .05$.

An increase in the living area by one square foot resulted in an increase in the price of the house by 1.830e+02 dollars. The relationship between the size of the house's living room in square feet and the cost of the house was as expected. The variable living area in square feet of the house statistically significantly predicted the price of the home, $p < .05$.

An increase in the number of bedrooms by one unit resulted in a reduction in the price of the house by 4.789e+04 dollars. The results of the relationship between the number of bedrooms and the cost of the house were contrary to what was expected. We expected the relationship between the number of bedrooms and the house price to be positive. Therefore, we expected an increase in the number of bedrooms would increase the cost of the house. The variable number of bedrooms of the house statistically significantly predicted the price of the home, $p < .05$.

An increase in the grade of the house by one unit resulted in an increase in the price of the house by 1.324e+05. The relationship between the house's grade and the house's price was as anticipated. As the grade of the house increases, we expect the cost of the house to increase

also. The variable grade of the house statistically significantly predicted the price of the home, $p < .05$.

An increase in the year in which the house was built by one unit resulted in a decrease in the price of the house by 4.071e+03 units. This was contrary to our expectations since we expected houses built recently to be more expensive compared to homes built earlier. The variable year built statistically significantly predicted the price of the house, $p < .05$.

Summary

The model accounted for 62% of the variation in prices in King County. There were two variables whose results were contrary to what was expected, i.e., the year the house was built and the number of bedrooms in the house. Each of the five independent variables (size, grade, bedrooms, bathroom, and year the house was built) statistically significantly predicted the price of the home. Further investigation needs to be done to help us understand why an increase in the number of bedrooms resulted in a reduction in the cost of the house and also why ancient houses were more costly compared to recently built.

References

Shanu Sushmita. (2022, Fall). ALY6010: Probability Theory and Introductory Statistics, Fall 2022 CPS Quarter, Canvas Modules, from https://northeastern.instructure.com/courses/126362/modules.

Higgins, D. M., Rezaei, A., & Wood, P. (2019). The value of a tram station on local house prices: an hedonic modelling approach. Pacific Rim Property Research Journal, 25(3), 217-227.https://doi.org/10.1080/14445921.2019.1693323

Lee, J. (2016). Measuring the value of apartment density?. *International Journal Of Housing Markets And Analysis*, *9*(4), 483-501. https://doi.org/10.1108/ijhma-08-2015-0047

Madhuri, C. R., Anuradha, G., & Pujitha, M. V. (2019, March). House price prediction using regression techniques: a comparative study. *In 2019 International conference on smart structures and systems (ICSSS) (pp. 1-5). IEEE.* https://doi.org/10.1109/ICSSS.2019.8882834

Dataset retrieved from: https://www.kaggle.com/datasets/swathiachath/kc-housesales-data