Analysis of House Sales in King County, WA

Nuoer Chen, Keer Liang, Zihan Ni, Shangyu Wang

Introduction

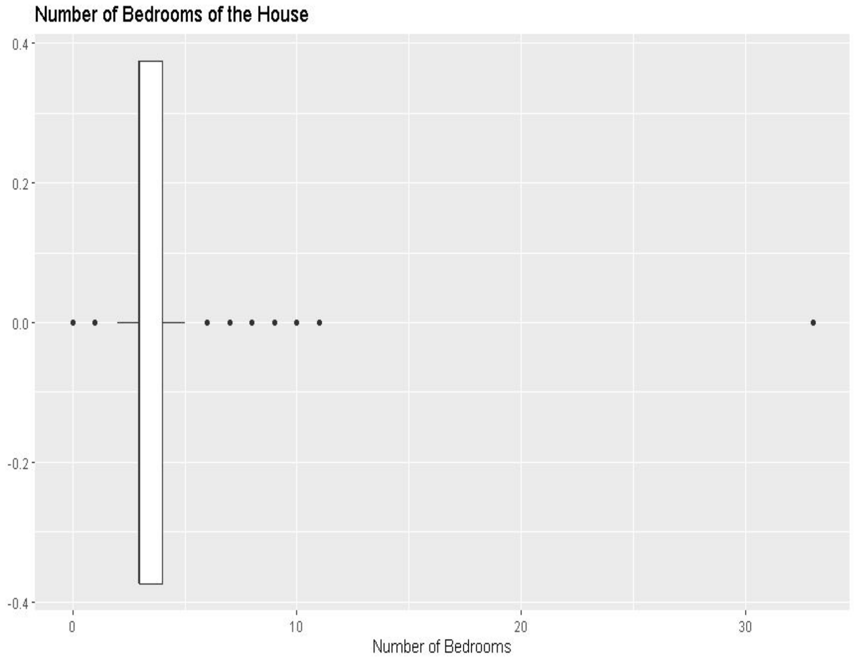
This data provides the information of sale records of the houses in King County, WA from 2014 to 2016. After the data cleaning, there are a total of 21586 sales records of houses in the dataset. We want to explore the relationships between the sale price of the houses, in USD, and other variables, such as the living area in square feet, the number of bathrooms and bedrooms. We came up with a couple business questions in order to help us find the best predictor of the price by the approach of running different tests.

Our business questions are:

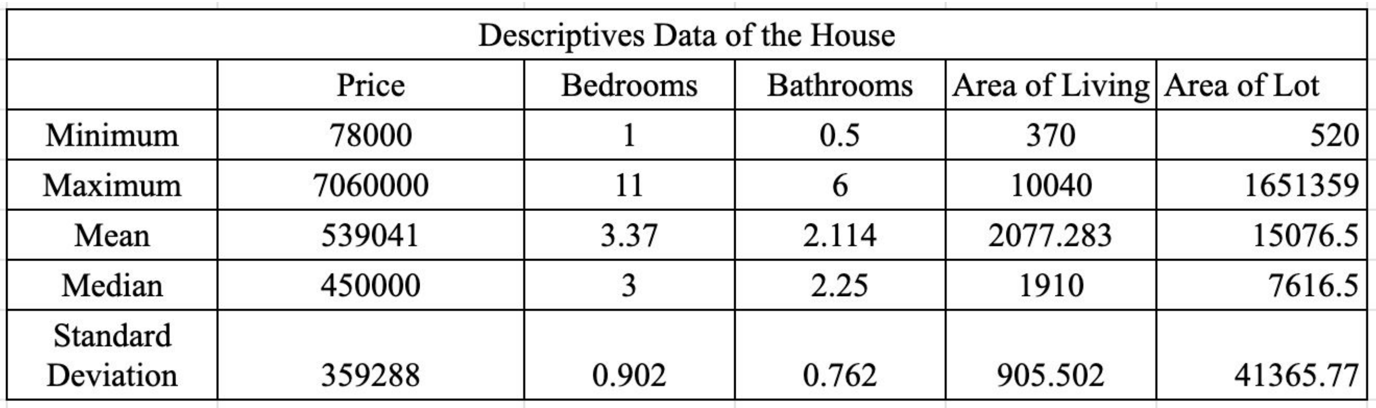
- “Does the living area affect the price of the house. If so, what is relationship between them?”(SLR model)
- “Is the area of lot a good predictor of the price of the houses?”(SLR model)
- “Does the renovated house sell at a higher price than the house that are not renovated before?” (Two sample t-test)
- “How can we predict a house price based on the information of the house?”(MLR model)

Data Preparation

In order to investigate the business questions above, the raw data needs to be filtered. Since all the tests are investigating the house price, we filter out the data that has either 0 bedrooms or 0 bathrooms because these sales record can be considered as selling lands or property. Moreover, the extreme outliers, which includes the house with 33 bedrooms (see graph on the right)and the houses with over six bathrooms, are filtered out according to the boxplot that we plot for the data cleaning.



DATA SUMMARY



Descriptives Data of the House

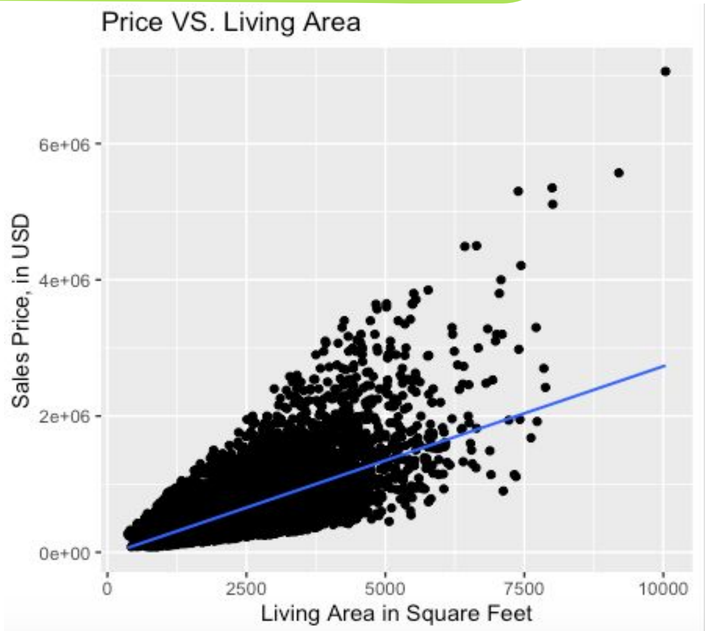
	Price	Bedrooms	Bathrooms	Area of Living	Area of Lot
Minimum	78000	1	0.5	370	520
Maximum	7060000	11	6	10040	1651359
Mean	539041	3.37	2.114	2077.283	15076.5
Median	450000	3	2.25	1910	7616.5
Standard Deviation	359288	0.902	0.762	905.502	41365.77

Price: the price of the house, in USD
Bedrooms: the number of bedrooms in the house
Bathrooms: the number of bathrooms in the house
Area of Living: the living area of the house, in square feet
Area of Lot: the lot area of the house, in square feet
Waterfront: Whether the house is waterfront(1) or not(0)

Living area vs. price (linear model)

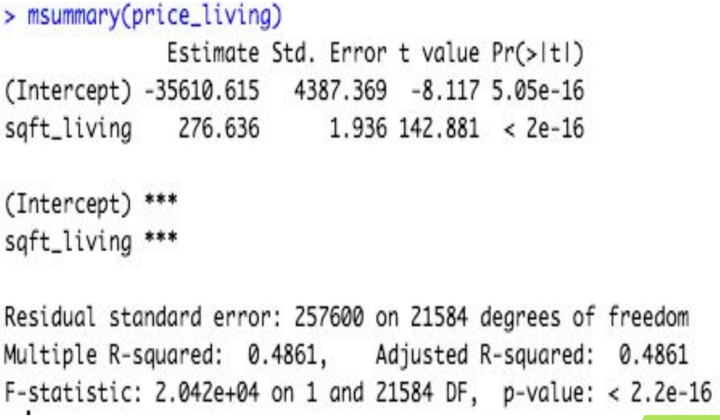
H0: There is no relationship between living area in square feet and the price of the house in dollar.

Ha: There is relationship between living area in square feet and the price of the house in dollar.



LIVING AREA VS. PRICE (LINEAR MODEL)

According to the result of linear regression on the left, the P-value of the variable living area, which is about 2.2e-16, is smaller than the alpha of 0.05, so we could reject the null hypothesis. Thus, there is statistically significant evidence that the house price, in USD, is associated with the living area, in square feet. As each square feet increase in area, the house price will increase by 276.636 dollars. And based on the R-squared, about 48.61% of the variation of the house price could be explained by the living area of the house. Since the P-value of the model is also about 0, we could also conclude that the model itself is also statistically significant.

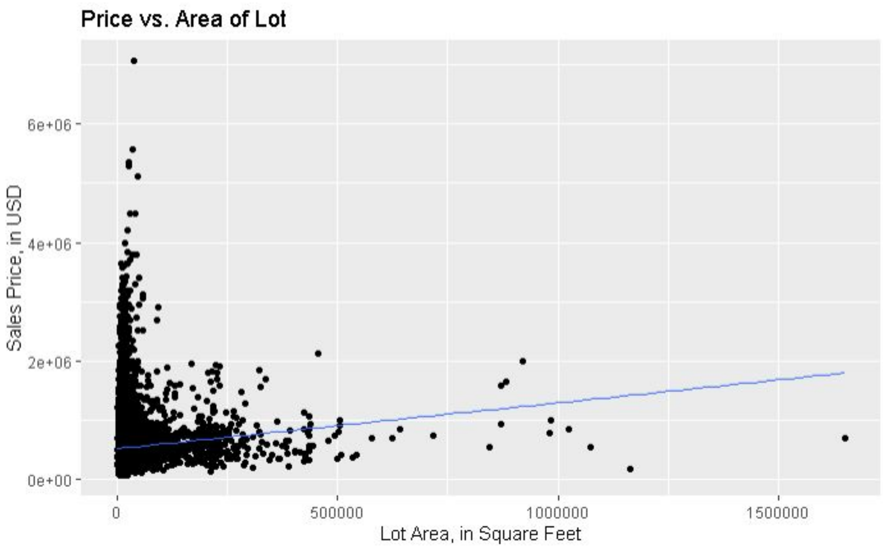


```
> msummary(price_living)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -35610.615   4387.369  -8.117 5.05e-16
sqft_living    276.636      1.936 142.881  < 2e-16

(Intercept) ***
sqft_living ***

Residual standard error: 257600 on 21584 degrees of freedom
Multiple R-squared:  0.4861,	Adjusted R-squared:  0.4861
F-statistic: 2.042e+04 on 1 and 21584 DF,  p-value: < 2.2e-16
```

SLR Price vs. Lot Area

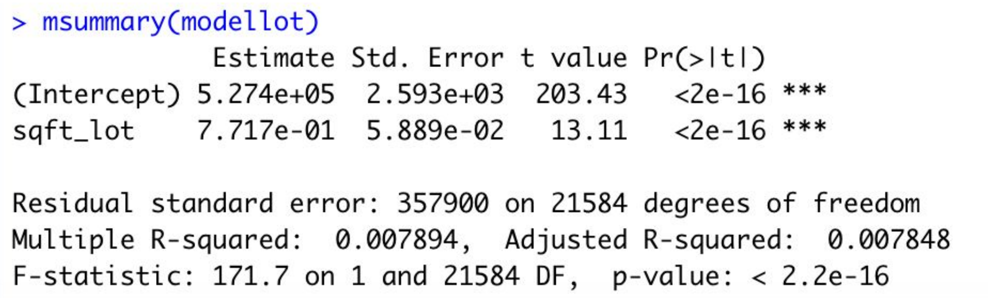
H0: There is no relationship between price and lot area.

Ha: There exists relationship between price and lot area.



SLR Price vs. Lot Area

We could see that the p-value is really small; therefore, we would reject the null hypothesis. Then there is statistically significant enough evidence that the lot area is a predictor for price. From the slope, we could see that there would be an increase of 0.7717 dollars per one square feet increase of lot area. However, the R-squared is relatively small, which indicates that the lot area can explain only 0.789% of variation on the price of the house. This indicates that although it is statistically significant, it is not a practically significant predictor for the house price. It indeed surprises us because we would think that normally the area of a house should be the most important factor for house prices, so the area of the living room and lot should both be good predictors for house prices. But it turns out that lot area has a small R-squared and is probably not a good predictor for price.

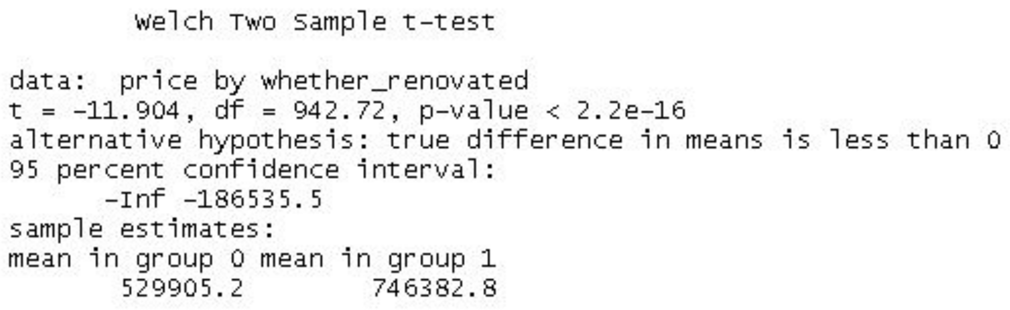


```
> msummary(modellot)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.274e+05  2.593e+03  203.43   <2e-16 ***
sqft_lot    7.717e-01  5.889e-02   13.11   <2e-16 ***

Residual standard error: 357900 on 21584 degrees of freedom
Multiple R-squared:  0.007894,  Adjusted R-squared:  0.007848
F-statistic: 171.7 on 1 and 21584 DF,  p-value: < 2.2e-16
```

Two Sample T-Test

H0: The mean price of the unrenovated house is the same as the renovated

HA: The mean price of the unrenovated house is less than the renovated



```
        Welch Two Sample t-test

data:  price by whether_renovated
t = -11.904, df = 942.72, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -186535.5
sample estimates:
mean in group 0 mean in group 1 
       529905.2        746382.8 
```

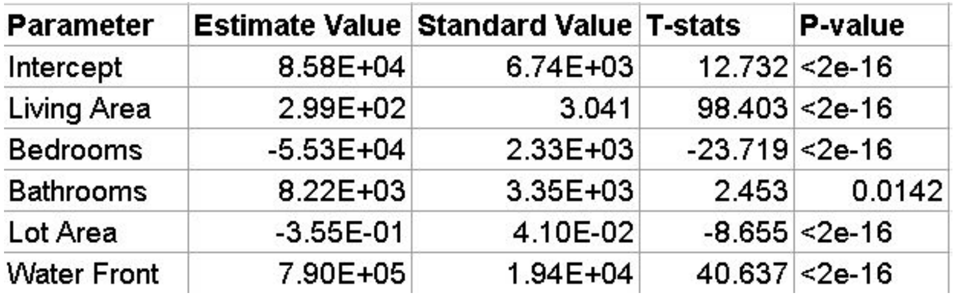
From the raw data, the new variable of whether_renovated is created to categorize whether the house is renovated. The mean price of the renovated is 746382.8 USD, which is indicated by group 1, and the mean price of the unrenovated house is 529905.2 USD, which is indicated by group 0 . The mean of the renovated house price is higher than the unrenovated. Therefore, we wanted to check if there is statistically significant evidence or not for the null hypothesis. We conduct a Two Sample T-Test. As it shows, the p-value, 2.2e-16, is extremely small; therefore, we would reject the null hypothesis. Then there is statistically significant evidence that the average price of the renovated house is higher than the house that was not renovated before.

Multiple Linear Regression

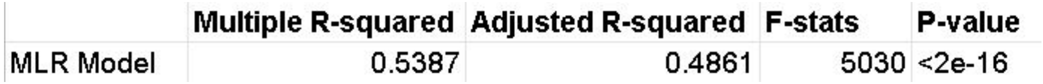
MLR Results:

Dependent Variable: Price of House

Independent Variables: Living Area, Bedrooms, Bathrooms, Lot Area, Waterfront



Parameter	Estimate Value	Standard Value	T-stats	P-value
Intercept	8.58E+04	6.74E+03	12.732	<2e-16
Living Area	2.99E+02	3.041	98.403	<2e-16
Bedrooms	-5.53E+04	2.33E+03	-23.719	<2e-16
Bathrooms	8.22E+03	3.35E+03	2.453	0.0142
Lot Area	-3.55E-01	4.10E-02	-8.655	<2e-16
Water Front	7.90E+05	1.94E+04	40.637	<2e-16



	Multiple R-squared	Adjusted R-squared	F-stats	P-value
MLR Model	0.5387	0.4861	5030	<2e-16

Summary:

Based on the multiple linear regression model, it indicates how different independent variables would affect the sale price of the house, in USD. The independent variables are: the living and lot area of the house in square feet, the number of bedrooms and bathrooms, and whether the house is waterfront(1)/or not(0). All of the variables and the model itself are statistically significant since the p-value are all smaller than the alpha 0.05. According to the R-squared, about 48.61% of variation on the price of the house can be explained by this model.

Conclusion

Our projects of investigating what and how factors would affect the sales price of the house, in King County, WA demonstrates that most of the price of the house can not be simply predicted by the variables that we used. From the two simple linear regression models, both the living area and lot area of the house are statistically significant predictors of the price of the house. However, people seem to value the living area more than lot area of the house since the linear relationship between lot area and the price of the house is strongly weak. Moreover, we have the confidence to conclude that the price of the house would be higher if the house is renovated based on the two-sample t-test we conduct. We select the variables that we think would be a strong predictor for the house price. It is surprised to see that the increase in the number of bedrooms and lot area would cause a drop in the price of the house. The limitations of the model, though it is statistically significant, would be that it would only predict less than half of the price of the house by using these variables.

Thanks!

Data Source :
https://www.kaggle.com/harlfoxem/housesalesprediction