Factors affect the house price in Pierce County in 2020

Xinjing Guo
1005086620

Introduction

House becomes more important for people, not only a place that can shelter from wind and rain, but it also can increase the index of people's happiness with family numbers. Cattaeo (2009) points out that a lot of countries put money on housing especially in the United States. It means that housing takes a quite significant role in people's minds. Since housing prices connect to people's income and wealth level, it is also necessary to study what factors affect house price as a prerequisite. Such as the material of the roof and the number of bedrooms.

The dataset is about the real estate sales for Pierce County, WA in 2020 (https://www.piercecountywa.gov/736/Data-Downloads). The data has 16814 samples and 19 variables which contain categorical and numerical variables. Each column shows the basic information about the house such as the sale date, house sum of the square feet.
As people pursue more comfortable and safe houses, the research provides people a guideline on choosing a house, knowing what factors are important for most people so that they can also notice. Moreover, real estate can also gain more information about people's concerns on choosing a house. They would pay more effort and money into these factors to make sure people are satisfied.

Next, through the method, result, and discussion sections to study **what factors affect the house price in Pierce County in 2020**. The method section shows how to create a full model, reduce the model, and automate the model. Then, get the best model by comparing. Finally, using the testing model to check whether it is a final model is appropriate. The result section shows the tables and graphs I got in code and check the goodness of the final model. The discussion model shows how the final model interprets the content and the goal.

Methods

- Create full training model:

In this project, in order to study which variables, correlate to the house price, firstly, separate data into two independent datasets which are a training dataset and a test dataset. The training dataset contains 80 percent of the original data and others in the test dataset. Use box plot and bar plot to have a general understanding of training data and put variables into a multiple linear regression model expect some has no correlation with house price such as “sale date”. Based on the Johansson (2017), he illiterates that the age, size and conditions of house can affect the price of house. Therefore, these variables should in the model.

- Full training model checking:

After fatting the model, check whether variables are independent of each other by checking the scatter plot. But before using the residual plot, make sure conditions 1 and 2 satisfy. Condition 1 proves response variable is a single function of a linear combination of the predictors. And condition 2 shows each predictor is a linear function with each other. Then, using residual plots make sure it has linearity of the relationship, uncorrelated errors, and constant variance. Using Q-Q plot checks whether points lie on the straight line smoothly. If the distribution of response variable is quite skewed, use the transform let become normal. After that, the full model is created. Lastly, use VIF to check whether there are some variables have multicollinearity. If does, drop one and check again until all values of variables are less than 5.

- Compare full, reduce and automated model

Based on the full model, check which variables are significant to the price of house (p-value < 0.05). Put these variables to a new model which is the reduce mode. Check VIF to make sure there is no multicollinearity.
Moreover, using the automated selection get the automated model and check VIF. Next, use F-test to check whether these reduce and automated models are better than full model. In F-test, the null hypothesis is the different coefficients between the full model and reduce model/automated model is 0. Alternative hypothesis is at least one of them is not zero. If the p-value is less than 0.05, we have evidence to reject alternative hypothesis test which there is no difference between two models. So, the value greater than 0.05 be better. Moreover, calculate the value of BIC, AIC and adjusted R-square. We hope value of BIC and AIC are smaller, adjusted R-square be larger. Finally, get the final model. But still need to check whether it satisfies the assumptions.

- Model validation

Use the same final training model but the test data to check whether variables are valid.
Firstly, compare EDA and determine whether there has a huge difference between testing model and final training model. Moreover, check whether scatter plots satisfy the assumptions. Observe whether the significant variables in final training model are still in testing model and whether the change of adjusted R-square is large. Moreover, check whether leverage points, outliers, and influential points are reasonable. Lastly, using VIF make sure each predictor is independent . If these values are reasonable, it means the regression model is acceptable. If not, find reason and discuss.

Result

- Description of Data

At the beginning to build the model, the variable about the materials used on the interior walls for all sample are dry wall. Also, the variable which is type of waterfront does not include any information. The "sale data" does not provide useful information on housing price. I delete these variables. Moreover, two variables, total square footage of the attached garage and total detached garage square footage are quite similar. So, I only choose one. Then, put others into the regression model.

Table1: Summary for training data

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
sale_price	2000	348000	418100	462033	525000	6130000
House_squared_feet	200	1320	1774	1880	2352	9510
attached_garage_square_feet	0.0	0.0	421.0	364.2	528.0	2816.0
bathrooms	0.000	2.000	2.000	2.317	3.000	8.000
stories	0.000	1.000	2.000	1.557	2.000	3.000

Based on the table 1, there are three numerical variables. The sale price is the predictor which shows the minimum price of the house is $2000 and the maximum price is $6130000. These values are quite reasonable. These numbers show the basic information about house, and they are all quite reasonable.

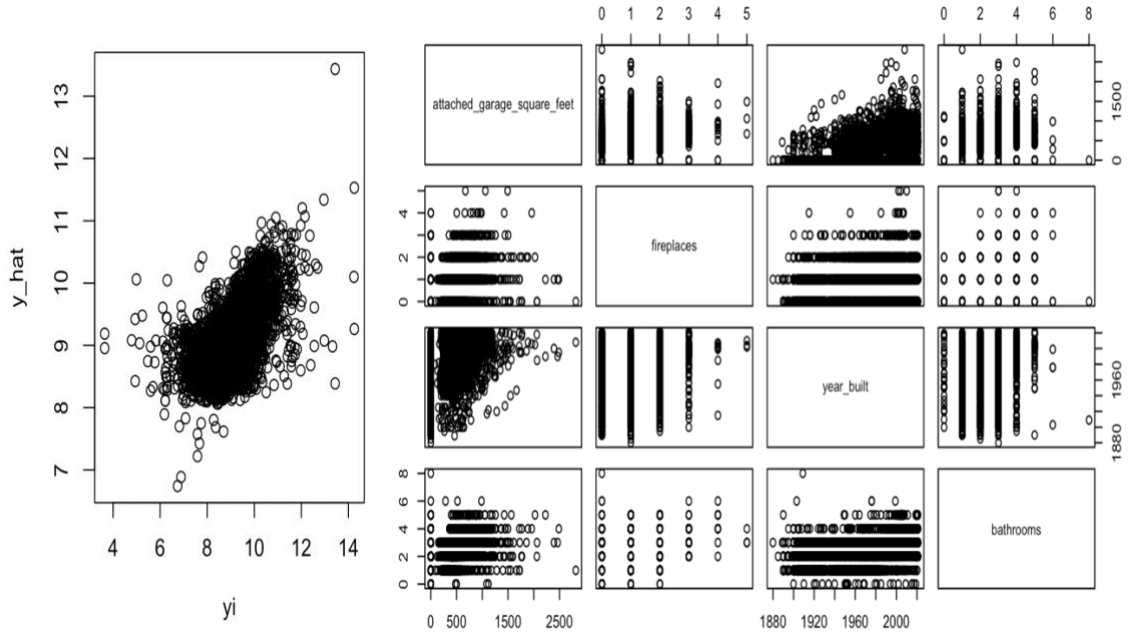
- Process of Obtaining Final Model

Table 2: Checking model

Model	AIC	BIC	Adj R^2	ANOVA
Transformed full model:	14875.27	15205.57	0.3841541	
	$sale\_price^{0.17} = \beta_0 + \beta_1 X_{attic\_finished\_square\_feet} + \beta_2 X_{basement\_square\_feet} + \beta_3 X_{attached\_garage\_square\_feet} + \beta_4 X_{fireplaces} + \beta_5 X_{hvac\_description} + \beta_6 X_{exterior} + \beta_7 X_{roof\_cover} + \beta_8 X_{year\_built} + \beta_9 X_{bathrooms} + \beta_{10} X_{bedrooms} + \beta_{11} X_{stories} + \beta_{12} X_{utility\_sewer} + \epsilon$			
Reduce model:	15364.48	15567.16	0.3605384	$2.2*10^{-16}$ (reduce model vs. trans_full_model)
	$sale\_price^{0.17} = \beta_0 + \beta_1 X_{attic\_finished\_square\_feet} + \beta_2 X_{basement\_square\_feet} + \beta_3 X_{attached\_garage\_square\_feet} + \beta_4 X_{fireplaces} + \beta_6 X_{exterior} + \beta_8 X_{year\_built} + \beta_9 X_{bathrooms} + \beta_{12} X_{utility\_sewer} + \epsilon$			

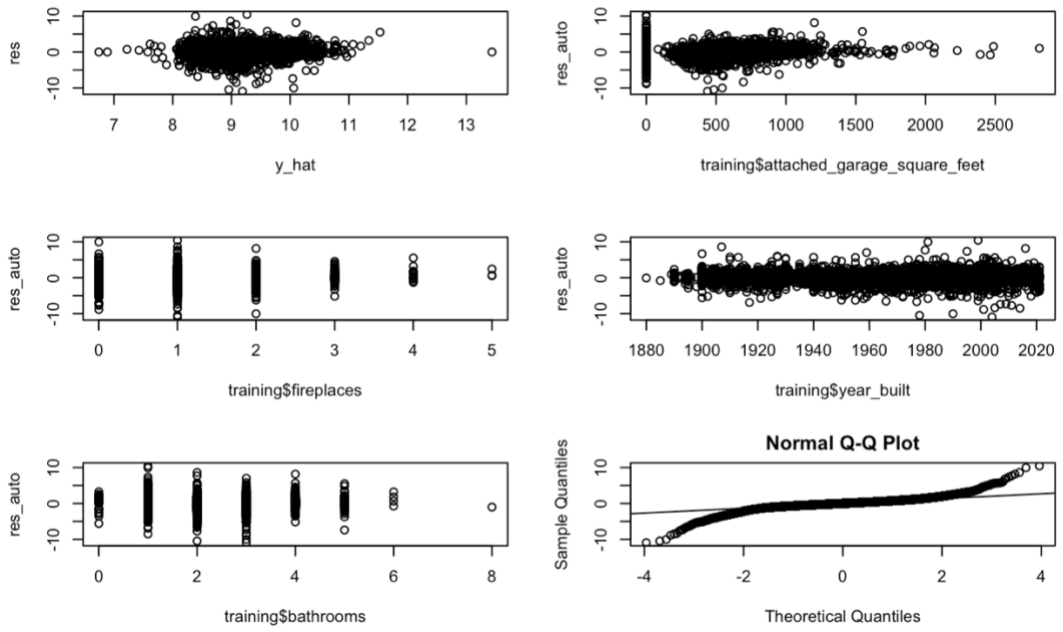
Automated model:	14873.26	15567.16	0.3841549	0.3712 (auto_model vs. trans_full_model)
	$sale\_price^{0.17} = \beta_0 + \beta_1 X_{attic\_finished\_square\_feet} + \beta_2 X_{basement\_square\_feet} + \beta_3 X_{attached\_garage\_square\_feet} + \beta_4 X_{fireplaces} + \beta_5 X_{hvac\_description} + \beta_6 X_{exterior} + \beta_7 X_{roof\_cover} + \beta_8 X_{year\_built} + \beta_9 X_{utility\_sewer} + \epsilon$			

This table shows models and values of anova, AIC, BIC and adjusted R-square of them. Using F-test check which model is better. Then, calculate the value of AIC, BIC, and adjusted R-square. By compare these values, the automated model is the best.



Graph 1: Condition 1, Condition 2 for training data

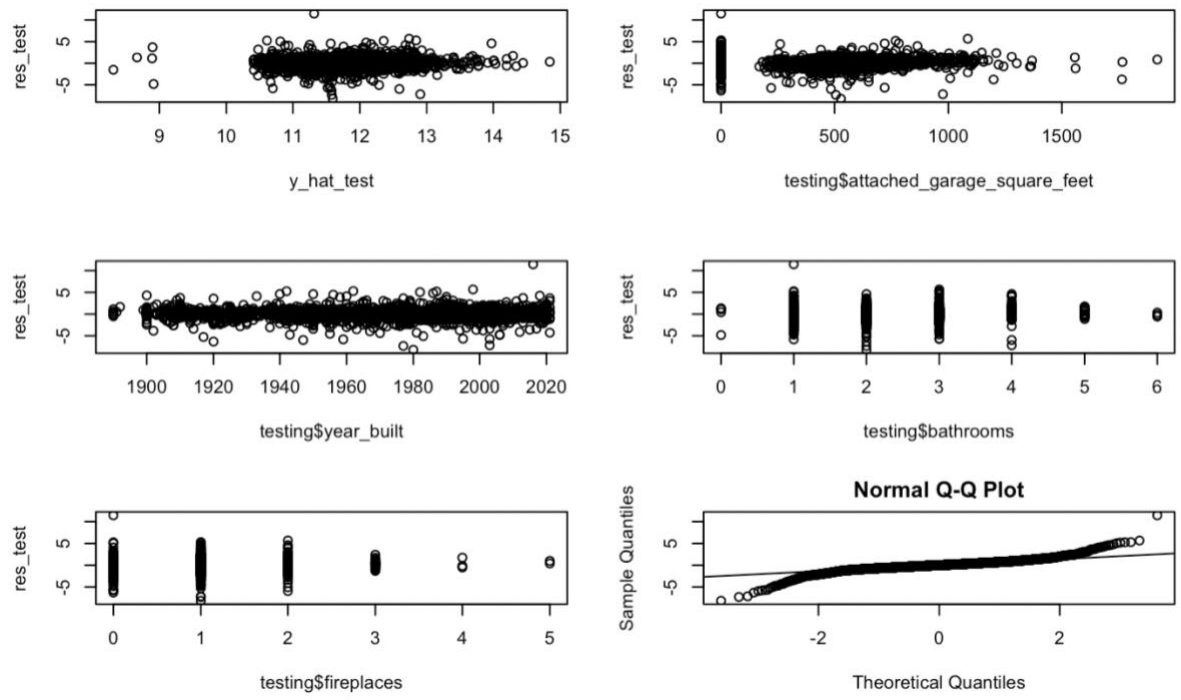
Graph 1 shows two conditions for residual plots. First plot shows points have no pattern. The second plots show the correlation between numerical variables to check whether every variable is a linear function to another. Based on the plots, it shows response variables can be represents by the linear regression model and there is no correlation between each variable.



Graph 2: Residual plots for testing model for automated model

Based on the graph 2, there are six residual plots for automated model. The first plot shows the residuals and predictor plots. Next four scatter plots are the residuals and the fitted values. The last is a Q-Q plot. Since there is no pattern in these plots, it satisfies three assumptions which are linearly of the relationship, uncorrelated errors, and the constant variance.

- Goodness of Final Model



Graph 3: Residual plots for testing model

Graph 3 shows six residual plots about the testing model. The first model is about the residuals versus predictor plots, and the next four plot is about the residuals versus fitted value plots. Since these plots have no pattern, it means they satisfy the assumption, linearity of the relationship, uncorrelated errors, and constant variance. Moreover, the last plot is about the QQ plot. Since these most points lie on the straight line, the distribution is normal. Therefore, the final model is quite appropriate.

Discussion

In this project, study what factors affect the house price and based on the plot, it shows the price of housing has correlation with the living area of attic, total square footage of the basement, total square footage of the garage, number of fireplaces, predominant heating source, material for the roof, year of building and sewage disposal system.

Final model:

$$\text{sale\_price}^{0.17} = \beta_0 + \beta_1 X_{\text{attic\_finished\_square\_feet}} + \beta_2 X_{\text{basement\_square\_feet}} + \beta_3 X_{\text{attached\_garage\_square\_feet}} + \beta_4 X_{\text{fireplaces}} + \beta_5 X_{\text{hvac\_descriptionForced Air}} + \beta_6 X_{\text{exterior}} + \beta_7 X_{\text{roof\_cover}} + \beta_8 X_{\text{year\_built}} + \beta_9 X_{\text{utility\_sewer}} + \epsilon$$

· β_0: The intercept of the model which is 0.1406

· β_1: Control other variables constant, increase one unit finished living area in the attic, the sale price increases 0.08505 dollar.

· β_5: Control other variables constant, increase one unit of the sale price decrease 0.3862 dollars.

· β_8: Control other variables constant, increase one unit of year built, the sale price increases 0.0007355 dollars.

· ϵ : The error term.

All these aspects affect the price of house. And they are quite acceptable but some variables that I think are important have not been included in the model, such as number of floors of the house. It is because Gomez (2019) points out the house area is also important to the house price.

Moreover, the outliers and leverage points are quite reasonable. Only an influential point is not reasonable since there is no bathroom and bedroom in a house. Also, because there are some extreme points in data, the β would have some change in different models.

Based on the research result, it means people would like to focus on these areas. For example, if the house is very old, and has fewer bathrooms, the sale price would be lower. This research give public a general understanding on house marking in Pierce County in 2020. And the range of the house price. Moreover, it provides more people references on buying the house. Besides, since these factors are significant to consumers, it also promotes real estate to pay more attention to these points. For example, they would provide a better sewage disposal system, better roof material and better heating source.

- Limitations of Analysis

There are some limitations in this study. Firstly, after transform response variable to $(sale\_price)^{0.17}$, there are still some outliers not fit perfectly on the line. Overall, the distribution approaches to the normal. Second problem is there are some different significant variables in both automated model and the testing model. It is because for the testing and training data, there are a lot of extreme points which the house price is quite high, but they are all reasonable. It is because the area of the entire property is quite large. And these extreme points may change the significance of the variables. Moreover, the adjusted R^2 of the final model is not large, so we need to explore other methods such as non-linear regression model to fit.

Reference:

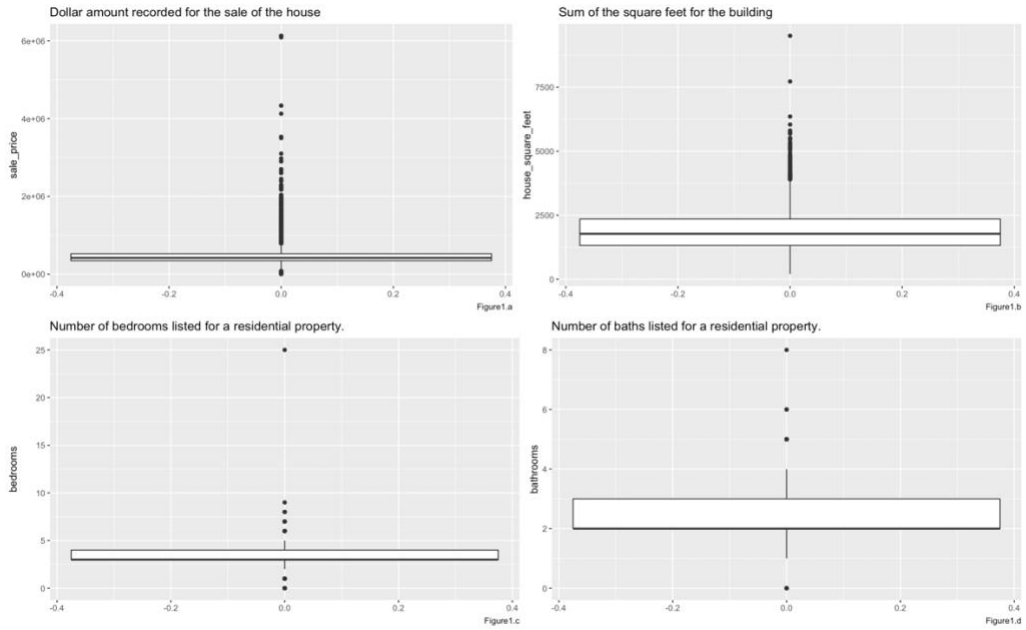
Cattaneo, M. D., Galiani, S., Gertler, P. J., Martinez, S., & Titiunik, R. (2009). Housing, health, and happiness. *American Economic Journal: Economic Policy*, *1*(1), 75–105. https://doi.org/10.1257/pol.1.1.75

Data downloads: Pierce County, WA - official website. Data Downloads | Pierce County, WA - Official Website. (n.d.). Retrieved December 17, 2021, from https://www.piercecountywa.gov/736/Data-Downloads

Johansson, A. (2017, August 4). *6 factors that influence a home's value*. Inman. Retrieved December 17, 2021, from https://www.inman.com/2017/08/07/6-factors-that-influence-a-homes-value/

8 critical factors that influence a home's value. Opendoor. (2019, September 19). Retrieved December 17, 2021, from https://www.opendoor.com/w/blog/factors-that-influence-home-value

Appendix:



Graph: box plot for training model

Table: VIF for final model:

	attic_finished_square_feet	basement_square_feet	attached_garage_square_feet	fireplaces	bathrooms
VIF	1.099567	1.296472	1.944213	1.279812	1.994587
	hvac_description	exterior	roof_cover	year_built	utility_sewer
VIF	2.898276	2.971638	1.587020	2.334327	1.687303

Table: common significant variables

Variable	auto_model	test_auto_model
basement_square_feetyes	***	***
attached_garage_square_feet	***	***
fireplaces	***	***
roof_coverComposition Shingle	*	*
year_built	***	**
bathrooms	***	***
utility_sewerSEWER/SEPTIC AVAIL	***	**
utility_sewerSEWER/SEPTIC INSTALLED	***	***
utility_sewerSEWER/SEPTIC NO	***	**

Based on the table, it shows there are quite more same significant variables in both automated model and testing model with same variables. Therefore, it means that the regression model is quite good since testing data can fit in this model.