HOUSING PRICE IN THE STATE OF WASHINGTON

Valerie Liang, Miaoboyang Xu

Professor Disha Shende

Economics 271: Business Analysis

Spring 2021

I. Introduction

This research paper aims to examine what factor affects housing prices the most and specifically analyze 21,613 properties from Washington. Our dataset is about houses built in 1995 and were revalued in 2014 and 2015 for resales.

To accomplish this task, we had to do some data cleaning then apply our data to python. The purpose of the statistic was to help create specific models for further analysis so that we can answer our research questions based on the data and finally decide if we reject or fail to reject our null hypotheses. With python, we can generate a summary table that shows all relevant quantitative data and their respective mean, standard deviation, minimum, and maximum. In the methodology, we employ corresponding boxplots, and a statistical summary table of all relevant variables that we focus on analyzing: housing price (price, in U.S. dollars), which are our dependent variables, the number of floors (floors), the number of bathrooms (Bathrooms), size of the living room (sqft_living) and the number of great schools (grade). Since our data focuses on birth weights in Washington, we came up with four hypotheses for testing and then answer our research question:

Research Question: Does "GRADE" have the greatest impact on housing prices among the number of bathrooms, floors, and size of living rooms?

For this research question, our main explanatory variable is 'grade' and our outcome variable is housing price. At the same time, we will hold the independent variables of "bathrooms", "floors", "size of living rooms" as controls.

Our assumptions on the relationships between the independent variables and dependent variables are:

1. *House prices increase with the increase in grade.*
2. *House prices increase with the increase of living room size.*
3. *House prices increase with the increase in the number of floors.*
4. *House prices increase with the increase in the number of bathrooms.*

II. Literature Review

To better understand housing prices in Washington, we collected the median sales price of houses sold in the past two decades from the Washington government and FRED. We searched for the median instead of the mean value because the median home price is a common measurement used to compare real estate prices in different markets and periods. It is less biased than the average price since it is not as heavily influenced by a small number of very highly-priced homes. Back to left graph, housing prices in Washington are more expensive than the overall situation in the U.S. except for the recovery period after the great recession from 2011 to 2016. Specifically, home prices in Washington accelerated quickly from 2002 through 2007, increasing by over $121,000, a gain of 64%. During the recession period, from 2007 to 2019, the price went down by approximately 24%. After the recession, house prices recovered to the highest price before the great recession in 2016, Median prices in 2019 exceeded 2009 values by 58.9%.

Figure 1. Housing Prices in the State of Washington and U.S.

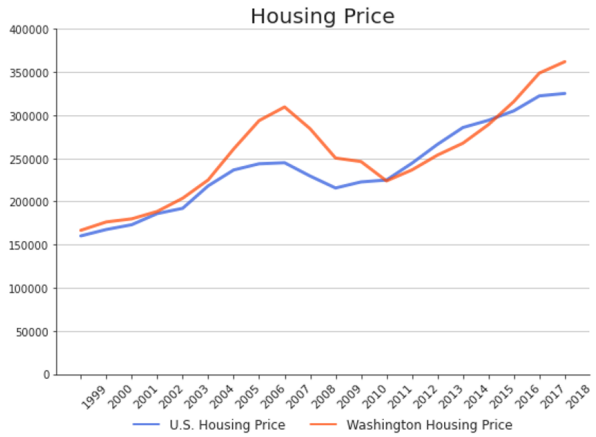
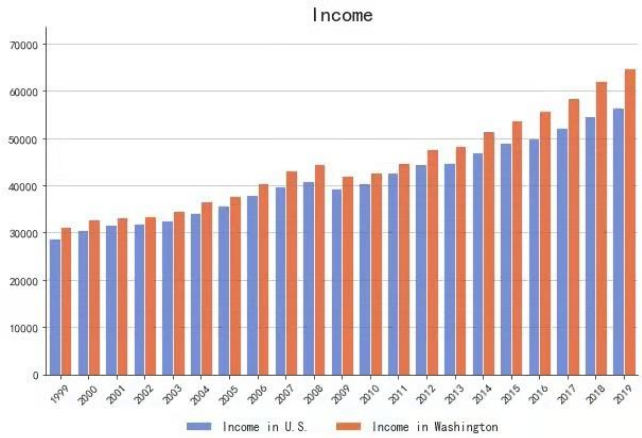


Figure 2. Personal Income in Washington and U.S.



Personal income also plays a significant role in terms of purchasing a real estate property. Personal income includes all income earned by Washington households, including wages, self-employment income, interest, dividends, rent, social security, and other transfer payments. In the graph, Washington's per capita personal income has been higher than that of the United States overall in almost every year since 1980. Comparing personal income graphs with housing price graphs, we realized that during the recession, personal income was growing, but the price of the house dropped, which seems unusual, so this is the reason why we were interested in researching deeply in this time period.

After a brief look at housing prices and personal income in the State of Washington, we turned into researching what external factors determine the price of a home. In terms of housing price determinants, there are two main categories. The first would be the value of the house itself, such as the number of bedrooms, number of bathrooms, and housing facility. The second category is about neighborhoods. Different definitions are used in the previous literature since it is an awfully broad term, which not only means the neighbor, the residents, and their characteristics. Meanwhile, it could be the living conditions such as grade location, the quality of local public goods, and crime rates. Our results and suggestions shed light on what a 'neighborhood' is in terms of what external factor matters to households when they are making decisions on purchasing the property.

Based on the data sets, our research is limited by the monotony of variables. We lack variables from category two such as crime rates, transportation, or conveniences index. Therefore, this paper mainly focuses on category one data. Some variables are simple. For instance, square feet of living describes the size of the living room. In a contrasting manner, it is necessary to provide definitions for other variables before moving on. First, although we describe floors as the number of floors in each property, we are still searching for a more detailed definition since it could mean properties on a particular level, which is an opposite way to define this variable. Hence, we may need future research on this particular variable. Second, "grade" describes how many great schools in your district, the bigger the number, the better. In the following analysis, we change the name of this variable into "school" when we are explaining the data, which is for the purpose of better understanding, but we will keep to use grade inorder to guarantee the integrity and primitiveness of the dataset. The third variable is bathrooms. Since

most of the values are decimal instead of integers, we make some clarifications. One bathroom is called a full bathroom which contains four key items: a bathtub, a shower, a toilet, and a sink with running water. The half bathroom contains a toilet and sink. The quarter bathroom has only a shower in it. Three quarter bathroom contains four key items except for a bathtub.

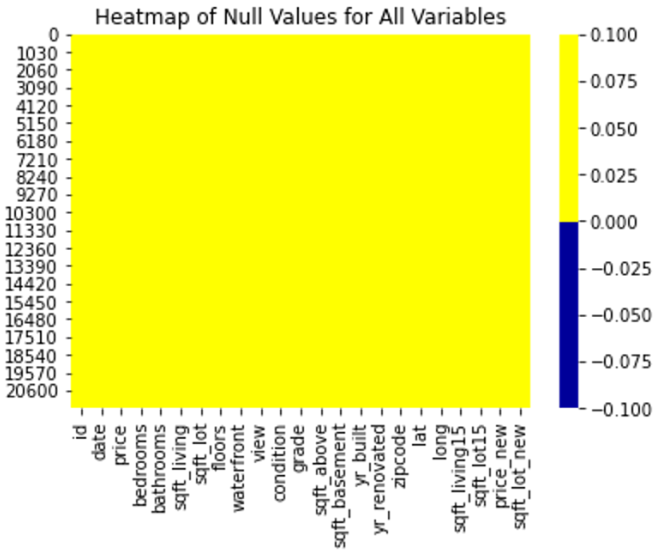
III. Methodology and Statistical Results

1. Rules and Assumptions

Before stepping into our research methods and data, we made some assumptions. The purpose of making those assumptions is to try to block unobserved heterogeneity as much as possible and make our results precise. First, we are only looking at residential properties, since the price of commercial properties such as industrial property, shops, and offices are sensitive to interest rate instead of the variables we are researching for such as the number of bathrooms. Second, we unify all the units of area into square feet. Third, the housing price is the price before property tax.

2. Data Cleaning

Firstly, we checked the data type of all variables. Our results show that four variables are floating (bathrooms, floors, lat, and long), which means that the data of those variables are in decimals. The only variable (date) is the object which includes information of the year, month, and day. The rest of the variables are integers. More importantly, all of our variables are numerical numbers with no string values.

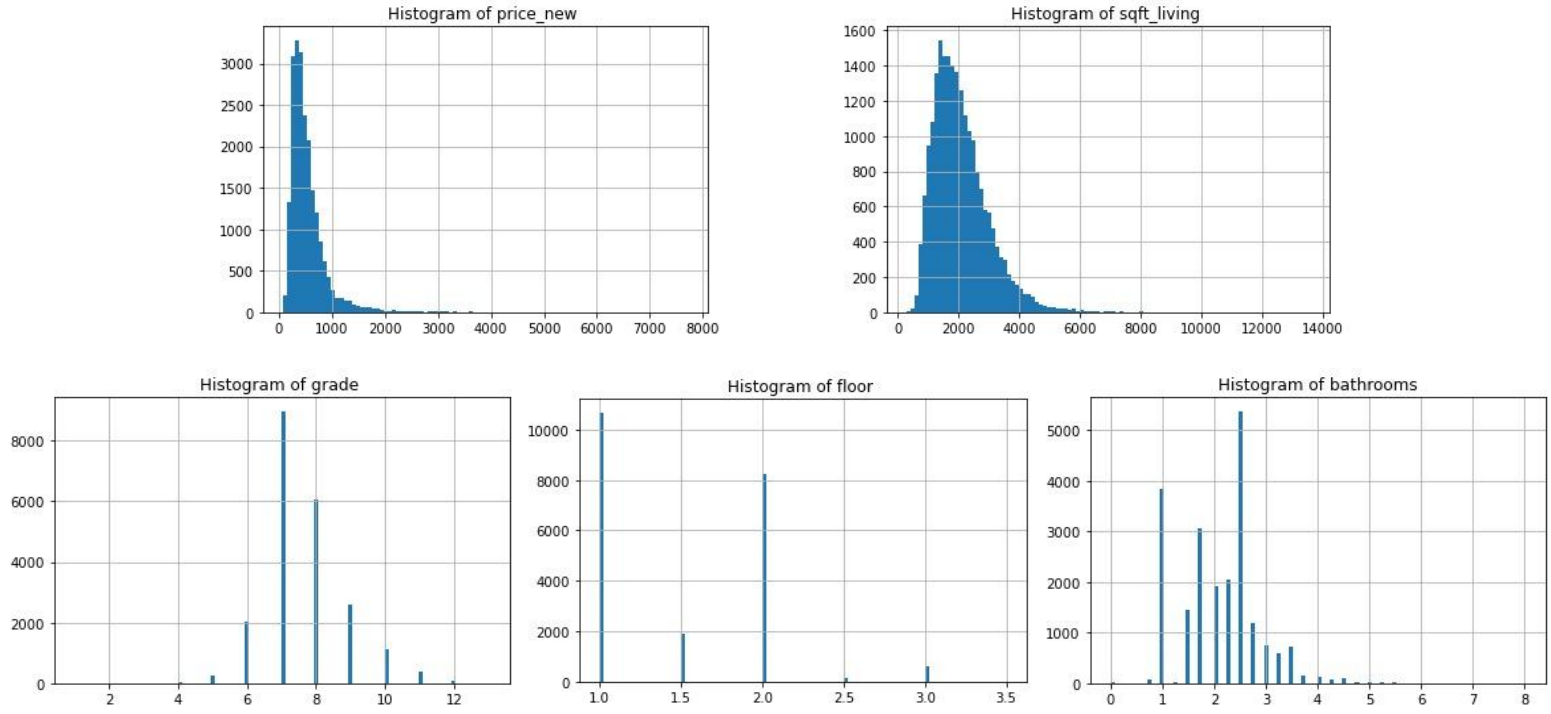


We also checked if there is any missing value in our dataset. In this heatmap, the whole graph shows as yellow which tells us that there is no null value for all variables in our dataset. Next, we checked if there is any string value for each variable. The data shows that except for the variable of “date”, all of the other variables are integers. Moreover, we

noticed that if we use the original index of the price, it will cause some errors in the regression model. To avoid the error, we remake another column by using the formula of *df['price_new'] = df['price']/1000*.

3. Analysis of Data distribution

The figures below are the histograms of all the variables. The histograms of price_new, sqft_living, and grades show that the data are normally distributed. The histograms of floors and bathrooms show a decreasing trend from left to right. We will say that the data for these two variables skew to the right.



From table 1 below, it shows the minimum, maximum, and mean values for all of our variables we used for analysis.

Table 1. Summary Statistics on Quantitative data

Variable	Minimum	Maximum	Mean
The Total Numbers of Schools Around the Housing Area (grade)	1.0	13.0	7
The Area of the Living Room in Square Feet(sqft_living)	290	13540	2079

Total Numbers of Floors (floors)	1	3.5	1
Total Numbers of Bathrooms (bathrooms)	0	8	2
Housing Price (price_new)	75	7700	540

4. Preparation for Regression Model

In our analysis, the mathematical equation of our regression model is:

$$Price\_new = \beta_0 + \boldsymbol{\beta}_1 grade + \boldsymbol{\beta}_2 sqft\_living + \boldsymbol{\beta}_3 floors + \boldsymbol{\beta}_4 bathrooms + \epsilon$$

We used only one regression model, where $\boldsymbol{\beta}_1$ is the impact of total numbers of school on housing price; $\boldsymbol{\beta}_2$ is the impact of the size of the living room on housing price; $\boldsymbol{\beta}_3$ is the impact of total numbers of floors on housing price; $\boldsymbol{\beta}_4$ is the impact of total numbers of bathrooms on housing price; and ϵ is an error term. Subsequently, we tested whether there is statistical significance between the two variables. We setted up the null hypothesis and alternative hypothesis:

$H_0: \beta_i = 0$ (there is no relationship between the two variables)

$H_0: \beta_i \neq 0$ (there is a relationship between the two variables)

Table 2. Statistically Significance of all the independent variables

Independent Variables		***P*-value**	**Significance**
X_1	grade	0.000	Yes
X_2	sqft_living	0.000	Yes
X_3	floors	0.000	Yes
X_4	bathrooms	0.000	Yes

Based on Table 2, ***p***-value of X_1 (grade), X_2 (sqft_living), X_3 (floors), and X_4(bathrooms) are all 0.000 which is also lesser than the default α value 5%. As a result, all of the dependent

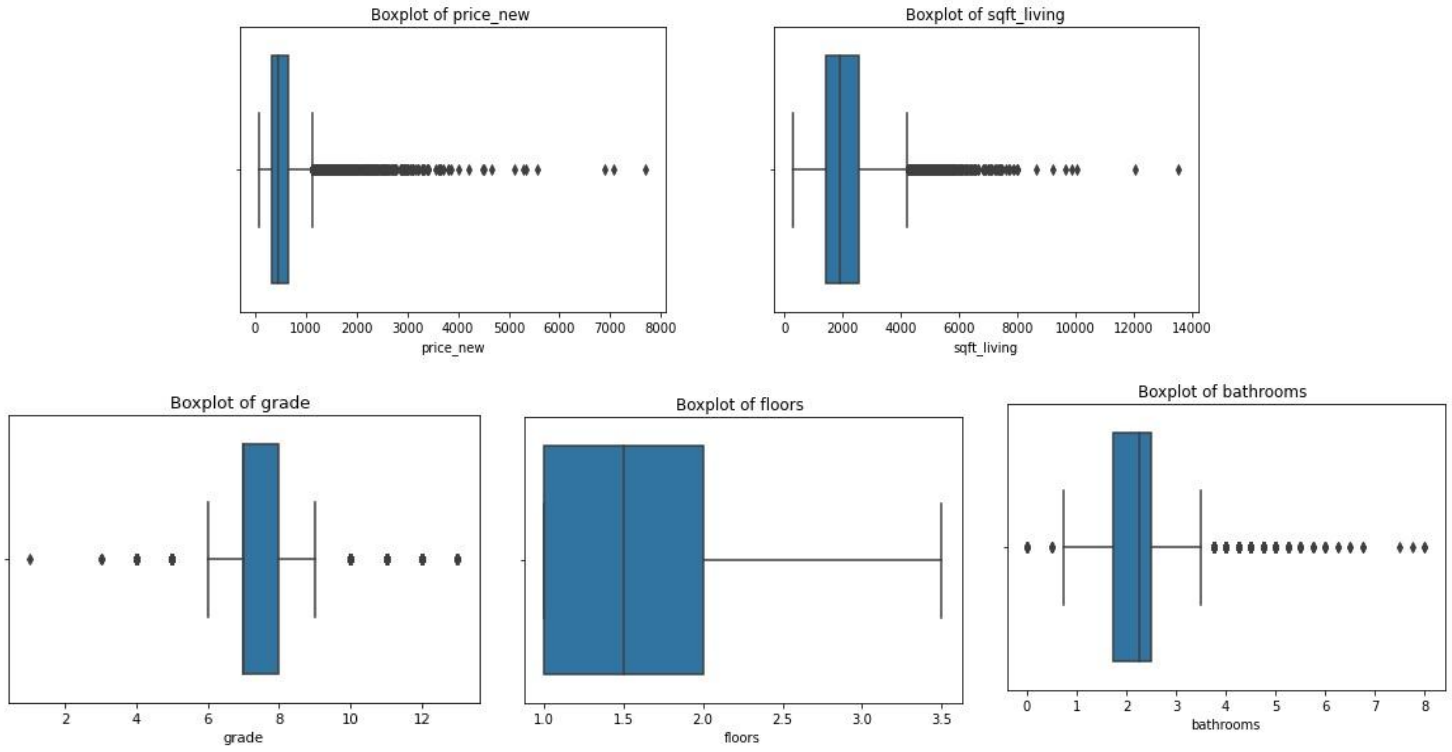
variables are statistically significant. For all the variables that satisfied ***p***-value < α value 5%, we can reject the null hypothesis and continue our interpretation of the regression model.

The purpose of the multiple regression model is to analyze whether the independent variables x_i (more than one) have an impact on the dependent variable on Y. In our scenario, we have five independent variables x_i: the total numbers of grades (schools around the house area), the total area of the living room in square feet, the total number of floors, and the total numbers of bathrooms. Our dependent variable is the housing price as Y. Since we divided the price by 1000 previously, we will need to multiply the coefficient by 1000 when we interpret the housing price for all variables.

5. Analysis of Boxplot

The primary purpose of these boxplots is to find if there are any outlier on the variables that we will use for research. The outliers are the values that excess from the range between lower bound and upper bound.

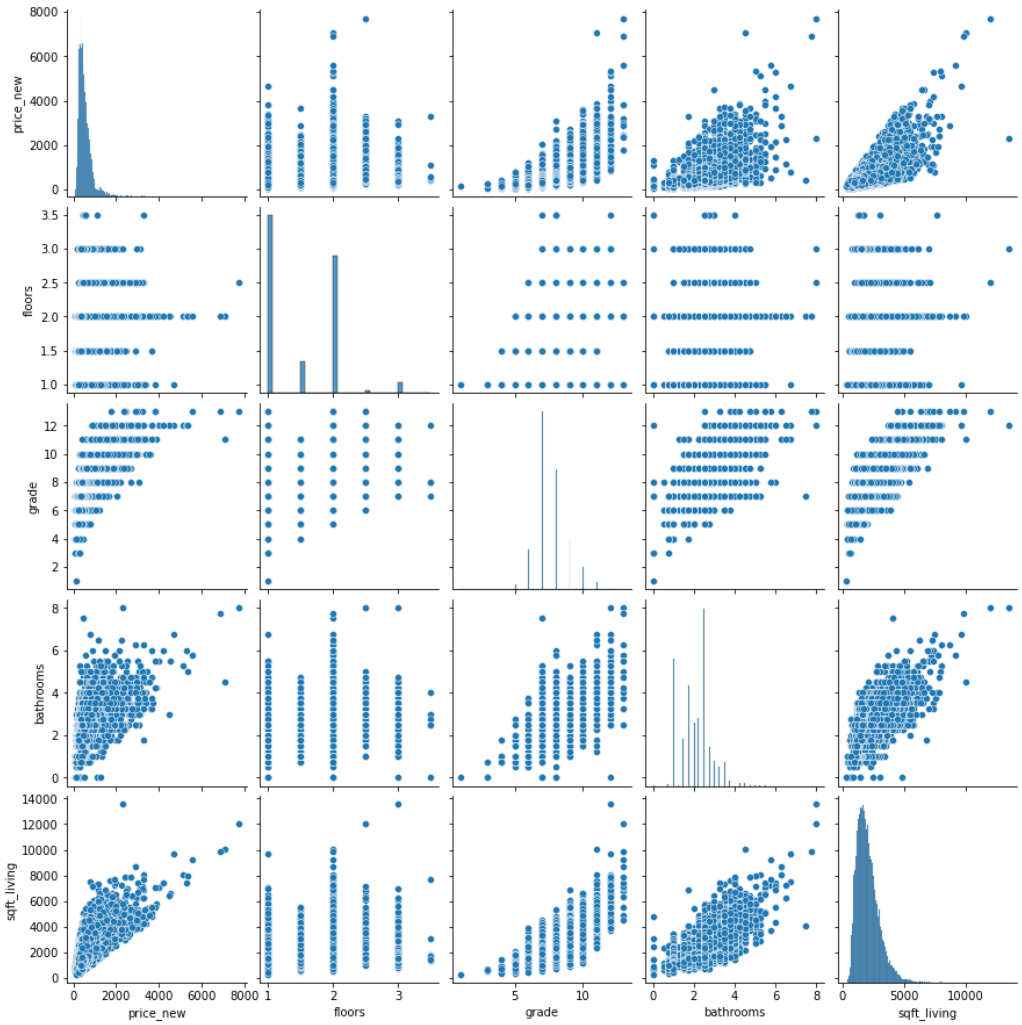
The figure below shows that the boxplot of price_new, sqft_living, grade, bathrooms, and floors. From the boxplot of price_new sqft_living, grade, and bathrooms, we can see that there are lots of outliers clustered above the upper bound, especially for the variable of price_new and sqft_living. Since the numbers of outliers are too many, it may lead to a big consistency for our analysis if we delete them. Therefore, we decided to keep all the outliers.



6. Analysis of Pairplot

From the pairplot below, it shows the correlation between different variables. Firstly, The graphs of “sqft_living”, “bathrooms”, and “grade” versus “price_new” in the first column. The dots show an obvious upward trend visually. As the “price_new” increases, the numbers of “sqft-living”, numbers of 'bathrooms’, and the numbers of “grade” also increase. Therefore, we can conclude that the dependent variable “price_new” has a positive correlation with the independent variables “sqft_living”, “bathroom” and “grade”.

Differently, the second graph of “floors” versus “price_new” does not have a recognizable trend. The dots are located more like normally distributed ones. Therefore, we concluded that there is no strong correlation between the independent variable “floor” and the dependent variable “price_new”.



7. Results of Regression Model

For understanding the relationship between different variables better, we ran the multilinear regression model. In total, we generated the regression model by four times separately. At each time, we will add a new variable to the previous model. In model 1, we generate the regression model between "grade" and "price_new". Because "grade" is our main explanatory variable, we will focus mostly on analyzing how the coefficient of "grade" changes in different models. In model 2, we added one more independent variable "sqft_living" to model 1. In model 3, we added the variable "floors" to model 2. In model 4, we added the variable "bathrooms" to model 3. As we mentioned early, we will need to multiply the coefficient value by 1000 when we do the interpretation for the regression model.

Table 2: Result of the multilinear regression model on all variables

Does higher grade have a higher housing price?

	Model 1	Model 2	Model 3	Model 4
Intercept	-1056.04***	-598.11***	-602.58***	-601.35***
	(12.26)	(13.30)	(13.26)	(13.24)
grade	208.46***	98.55***	107.64***	109.45***
	(1.58)	(2.24)	(2.35)	(2.36)
sqft_living		0.18***	0.18***	0.20***
		(0.00)	(0.00)	(0.00)
floors			-43.96***	-34.65***
			(3.54)	(3.76)
bathrooms				-26.67***
				(3.65)
R-squared	0.45	0.53	0.54	0.54
R-squared Adj.	0.45	0.53	0.54	0.54

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

When we interpret the result of the regression model, we will follow some rules. If the coefficient value is positive, we will interpret the result as: ceteris paribus, on average, with an increase or decrease in x, the housing price y will increase or decrease by 1000* (the coefficient value) US dollars. Therefore, we interpreted the result in model 4 as: ceteris paribus, on average, with an increase in one number of grade, the housing price will increase by $109,450; ceteris paribus, on average, with an increase in one ft^2 of living room, the housing price will increase by $200; ceteris paribus, on average, with an increase in one floor, the housing price will drop by $34,650; ceteris paribus, on average, with an increase in one bathroom, the housing price will decrease by $26,670.

The R-squared provides the information of how much x% of variations in Y is explained by variations in x. The result, it shows that the R-squared value in the four models is close to 50% similarly. It means that approximately fifty percent of variations in the housing price are explained by variations in the numbers of "grade", the area of the living room, the total number of floors and bathrooms.

When we compare all coefficient values of "grade" with all of the other variables' coefficient values, we noticed that the coefficient values of "grade" were the largest. Therefore, it proved that the main explanatory variable "grade" does have the greatest impact on our dependent variable "price_new". Next, when we compare the coefficient values of "sqft_living" horizontally, we noticed that the coefficient values did not have a significant change. As a result, we conclude that the independent variable of "sqft-living" has a very stable impact on the dependent variable "price_new".

On the other hand, the result of the variables "floors" and "bathrooms" violated our assumptions. The coefficient values of these two variables in both model 3 and model 4 were negative, which means that with an increase in the numbers of floors or an increase in the numbers of bathrooms, the housing price will decrease. Based on our studies so far, we could not give a proper explanation for this result. We will need to do more research on the consumer's behavior and gather more data to analyze this further.

IV. Conclusion & Recommendation

1. Conclusion:

In conclusion, our main explanatory variable "grade" has a great impact on the housing price based on the analysis of multilinear regression models. Furthermore, the data analysis of the paris plot and regression model also proves that our assumptions on the "grade", "sqft_living", and "price" are correct. However, our assumptions on the "floors" and "bathrooms" are wrong.

2. Recommendation & Limitation

As for recommendations, we suggest consumers who have a limited budget to buy houses that are far away from school because it will be a lot cheaper. However, if the consumer thinks of

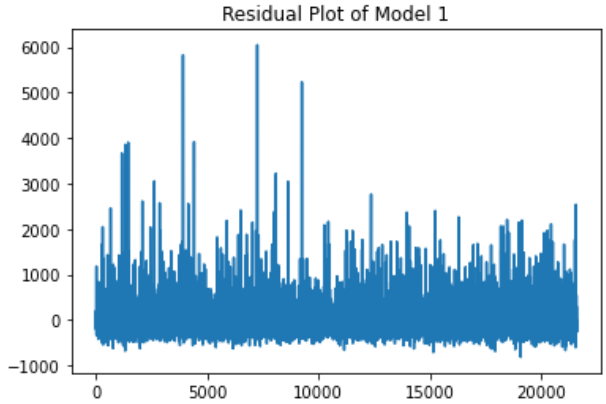
buying the house with an investing purpose, we suggest the investors buy the houses that are near the school area. The more total number of schools the better. In our literature review, we also show that the housing price in Washington keeps increasing in a rapid upward trend in recent decades. As the State of Washington develops, the reputation and popularity of it also increases, and that will have a positive effect on the housing price. Based on this trend, we predict that the housing price of Washington will keep increasing even more in the future.

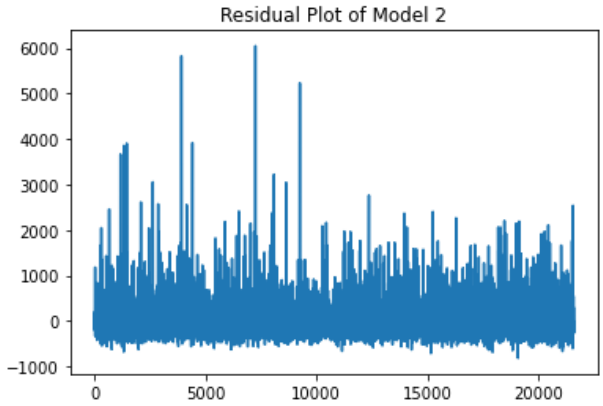
Besides, there are some limitations on our analysis. Since our dataset only provides the numbers of floors and bathrooms, we could not develop a better analysis to explain its negative impact on the housing price. We will need to look for more data such as the size of each floor and the survey of consumers' desire on having more or less of the total number of floors and bathrooms.

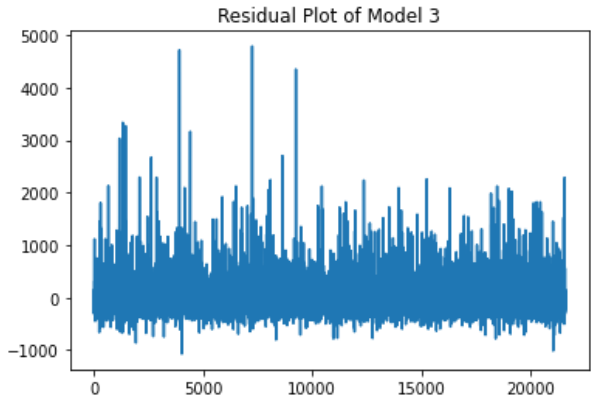
Moreover, we have countless outliers in most of our variables. Even though we decided to keep the outliers for inconsistency. In some way, the outliers may cause a disproportionate effect on theon regression model, which can result in misleading interpretations. Therefore, the regression model may not be fully reliable.

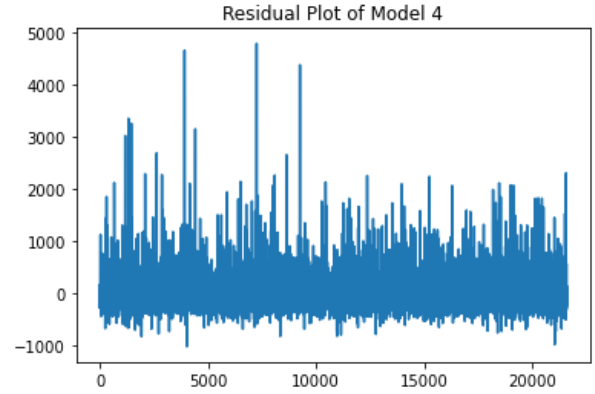
Last but not least, our dataset did not provide enough variables that are one of the determinants under the neighborhood category. Consequently, we could not analyze if “grade” has the greatest impact on the housing price among all of the other important determinants under the neighborhood category.

Appendix


Residual Plot of Model 1
6000
5000
4000
3000
2000
1000
0
−1000
0
5000
10000
15000
20000


Residual Plot of Model 2
6000
5000
4000
3000
2000
1000
0
−1000
0
5000
10000
15000
20000


Residual Plot of Model 3
5000
4000
3000
2000
1000
0
−1000
0
5000
10000
15000
20000


Residual Plot of Model 4
5000
4000
3000
2000
1000
0
−1000
0
5000
10000
15000
20000

Reference

Engle, Robert F., David M. Lilien, and Mark Watson. "A Dynamic Model of Housing Price Determination." 07 Mar. 2002. Web. 02.Apr. 2021.

Kiel, Katherine A., and Jeffrey E. Zabel. "Location, Location, Location: The 3l Approach to House Price Determination." 10 Jan. 2008. Web. 02 Apr. 2021.

"Median Home Price." 15 July 2020. Web. 02 Apr. 2021.

"Median Sales Price of Houses Sold for the United States." 28 Jan. 2021. Web. 02 Apr. 2021.

Oloke, O. C., Y. A. Olawale, and A.S Oni. "Price Determination for Residential Properties in Lagos State, Nigeria: The Principal-Agent Dilemma." 01 Jan. 1970. Web. 02 Apr. 2021.

"Trends in Home Buyer Preferences." Web. 02 Apr. 2021.

"Washington and U.S. per Capita Personal Income." 14 July 2020. Web. 02 Apr. 2021.