House Price Investigation for Cities in the State of Washington

Graduate Course:
Quantitative Analysis for Business

Team Members:
Sagar Bansal,
Yiqing Bu,
Parth Padia,
Ziting Yuan

Abstract

On an individual level, the real estate markets form as one of the most reliable ways for investors to grow their personal wealth. They qualify as a safe and potential investment option to secure income stability. On a much more global scale, property prices are becoming increasingly important and beneficial, the ability to identify price trends in properties serve as markers for indicating the overall market condition and the economic health of a country. This makes it a somewhat interesting investment category for analysis as there exists no standard source of real estate valuation. Our analysis emphasizes on investigating the different factors that are correlated with the price of a house in ten cities in the state of Washington. The cities that we studied were Seattle, Renton, Bellevue, Redmond, Kirkland, Issaquah, Kent, Auburn, Sammamish and Federal Way. We built a multiple linear regression model with variables such as number of bedrooms, number of bathrooms, square footage of living area, square footage of lot, number of floors, city, and age of the house out of 18 different variables in the beginning. After our analysis, we found that only square footage of lot does not add much to the average natural log of house price values when other features of the house remain same. The detailed explanation is provided below in the paper.

Section 1 - Introduction

Real estate is one of the largest asset classes worldwide**.** It is a successful long-term investment strategy, which has established itself as a safe and efficient method of investment that secures income stability for investors of all types. Apart from its value-appreciative nature, real estate owners stand to benefit from rental income that can arise as a part of owning and investing in properties. Thus, we would like to analyse some real-estate properties in one of the top real-estate markets in the United states (Washington) in order to gain some insight into factors that may or may not affect price trends in real-estate markets.

The purpose of writing this paper is to estimate and examine how the factors including number of bedrooms, number of bathrooms, sqft_living, sqft_lot, number of floors, age of house, and cities affect the house prices in Washington state during the summer season. The method of analysis we used is multiple linear regression which can help us address the following specific questions related to our topic:

- Is our model able to explain most of the variability in the response?
- Is our model statistically significant at 5% significance level?
- How different variables are related with the house price?
- Are there any independent variables that turned out to be uncorrelated with the response?
- How can the houses in the State of Washington become more valuable after we fully study this dataset?
- What are the potential problems or issues with our model?

In Section 2, we will have more detailed explanations of the data, and patterns in the data themselves. Section 3 contains our hypothesized model, evaluations of the model and interpretations from the model. We will also illustrate our modified model. Section 4 includes a summary of the key findings and concluding remarks.

Section 2 - Data Characteristics

Our chosen dataset comprises house-property related data in terms of price and numerous variables that may be associated with it, this data is pertaining to the period of May, 2014 - July 2014. The dataset has been sourced from Kaggle - the online web-based data repository with a large number of public datasets that are best-suited for carrying out data analysis. (The link to the original dataset has been referenced in the appendix). The data provided is cross-sectional in nature as the data records are a group of houses from May, 2014 to July 2014 and not for a particular house over a certain period of time. Also, this is an observational study since our data was collected first and then, we hypothesized the model.

Specifically, the dataset provides us with the price of houses spanning across multiple cities in the state of Washington, with each house having 18 characteristics such as the number of bedrooms, number of bathrooms, area of the house, location, year it was built etc. Hence, we first started with selecting variables that we want to study in our analysis based on our initial observations and previous knowledge about the data.

I) Data Used

We eliminated some unnecessary variables such as date, street name, state with zip, country, waterfront, view, yr_renovated, condition, sqft_above and sqft_basement. In our dataset, Date entries range from May, 2014 to July, 2014. As there is no clear explanation of this field on Kaggle, we assumed that the data has been collected/posted on these dates. Furthermore, these dates belong to the summer period, therefore based on our domain knowledge we believe that house prices tend to have the same variation for months in the same season. Hence, we decided to remove the date column from our study.

Subsequently, we were of the opinion that the State and Country will not add much to our analysis as all the data observations belong to the Washington, United States. Also, we believe that exploring relationships in terms of cities instead of street names and zip code will be more efficient as it is easier to take actions focused on a city than individual zip code areas or street names. The Waterfront, View, Yr_renovated variables have a majority of values as zero so we believe that these variables will not add much to our analysis. For the Condition variable, the dataset does not have any explanation about the meaning of different condition values. For instance, we are not sure that a condition value of 1 represents best or worst condition.

In addition, we got rid of the columns “sqft-above” and “sqft-basement” as their aggregate has already been recorded in the column “sqft_living”. Finally, due to the presence of 42 different cities in the dataset, we will focus our study on ten cities in the state of Washington. We selected these cities by looking at the top ten cities with the maximum number of data observations in our dataset. The resulting ten cities are Seattle, Renton, Bellevue, Redmond, Kirkland, Issaquah, Kent, Auburn, Sammamish and Federal Way.

After removing unneeded variables, we had one response and seven independent variables in total. **Price** is a quantitative variable that shows the price of the houses listed. This will be used as our response variable in the analysis. **Bedrooms** is a quantitative variable depicting the number of bedrooms for each row entry of house data. **Bathrooms** is a quantitative variable depicting the number of bathrooms for each row entry

of house data. **Sqft_Living** is a quantitative variable that describes the total area (in ft^2) of the house and the basement (if applicable).

Sqft_Lot is a quantitative variable that describes the total area (in ft^2) of the lot that the house has been built upon or the area occupied by the house's property. **Floors** is a quantitative variable that describes the number of floors a house might have. There are no "0" values in this data field. In the case of a house not having an upper level, the living room itself is considered as a floor. Thus, the minimum value is 1. **City** is a qualitative variable that tells us the name of the city in which the specified house is located.

Ageofthehouse (in year) is a new quantitative variable estimated from the **Yr_built** variable. Observations in this column were calculated by taking the difference of the built year of the house and '2014' since this dataset pertains to that year. In other words, the age of the house will tell us about the number of years it has been since its construction. We also cleaned some data records with zero value of bedrooms, bathrooms or house price variables and unusually high house price value. Practically, these values are not reasonable and can be a result of data entry problems. After cleaning the data, we have 3,412 data observations in the dataset.

II) Primary Relationships

The Table 1 and Table 2 below give us a summary of our response (price) and all of our variables. Specifically, Table 1 shows us the measures of central tendencies for all our concerned quantitative variables, such as the mean value, maximum and minimum values, standard deviation etc. for each variable. Table 2 gives us the count and percentage of the number of observations for each city in the dataset.

Table 1: Numerical Summary of Quantitative Variables

	N	Minimum	Maximum	Mean	Std. Deviation
price	3412	87500.0000	7062500.000	559877.4125	353808.5124
bathrooms	3412	.75	8.00	2.1491	.77291
sqft_living	3412	370	13540	2100.01	936.518
sqft_lot	3412	638	1074218	11753.26	28512.876
floors	3412	1.0	3.5	1.525	.5531
bedrooms	3412	1	9	3.38	.916
Age	3267	0	114	45.47	31.910
Valid N (listwise)	3267				

We further conducted some basic explorations using scatter and box plots by plotting each of our independent variables with our response. We noticed some primary relationships between our response and some of our independent variables. We can observe that there is a positive correlation between price and bathrooms (Fig.1), indicating that houses with a greater number of bathrooms tend to have higher prices. We can also observe that there exists a positive correlation between price and sqft_living indicating that houses with more square footage of the living area tend to have higher prices. (Fig 2.)

The boxplot of price by city (Fig. 3) shows different ranges of house price values (minimum to maximum) for different cities. This observation gives us a strong indication that the city name can be correlated with house prices. From the rest of our statistical plots which are referenced in the appendix, we cannot see any initial observations that could indicate a correlation of Bedrooms, Sqft_lot, Floors and Age variables with our response variable. Notwithstanding, we would like to confirm these observations using our model.

Table 2: Numerical Summary of Qualitative Variable (City)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Auburn	175	5.1	5.1	5.1
	Bellevue	281	8.2	8.2	13.4
	Federal	145	4.2	4.2	17.6
	Issaquah	186	5.5	5.5	23.1
	Kent	183	5.4	5.4	28.4
	Kirkland	187	5.5	5.5	33.9
	Redmond	234	6.9	6.9	40.8
	Renton	291	8.5	8.5	49.3
	Sammamish	171	5.0	5.0	54.3
	Seattle	1559	45.7	45.7	100.0
	Total	3412	100.0	100.0	

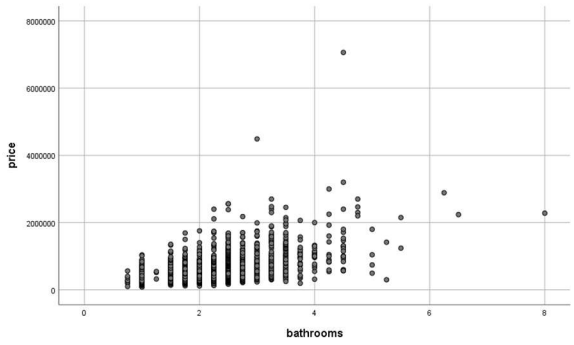

Fig. 1: Scatter plot of Price vs Bathrooms

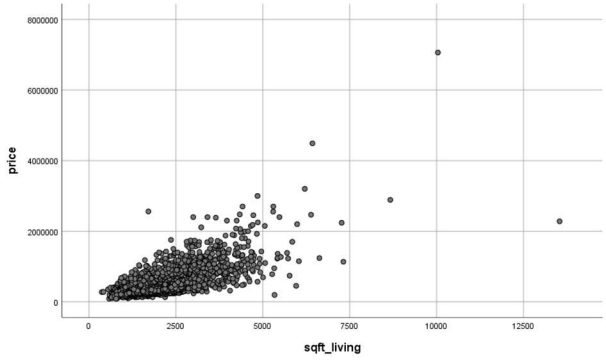

Fig. 2: Scatter plot of Price vs Sqft_living

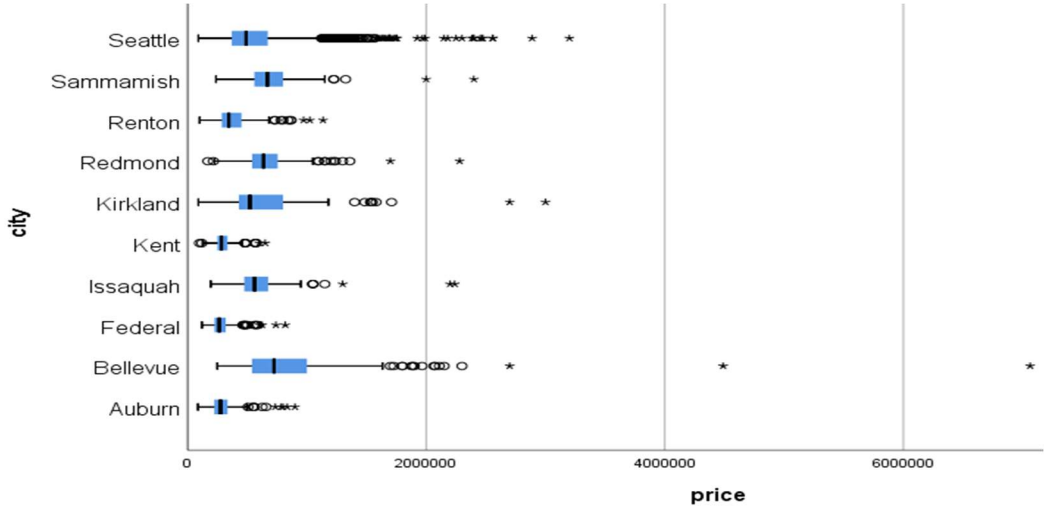

Fig. 3: Box plot of Price by City

Section 3 - Model Selection and Interpretation

As per the nature of the data defined in the preceding section, our model should contain House Price as the target variable and others as independent variables. Given that our response is continuous and we want to understand the correlation between other variables and our response, the chosen model was formulated using multiple linear regression technique. This section has been divided into two subsections: Original Model and Transformed Model. Based on our exploration, we initial came up with our original model. However, we found that this model violates certain assumptions during the model validity stage. We then updated our initial model by transforming our response to natural log of house price and used the transformed model to make interpretations.

I) Original model

For the thought process, we had about five steps. First, we hypothesized the model using the multiple linear regression technique. Next, we needed to fit the model by using SPSS, from which we got the model summary and could analyze the coefficients of the model. During this process, we performed dummy coding, which is the process of transforming a qualitative variable into binary variables (0 or 1). To achieve this, we have 10 different cities, so we identified 10 qualitative variables. Then, we created 9 dummy variables by choosing one variable as the reference level. We chose to assign Federal Way as the reference level, and coded the other cities as CitySeattle, CitySammamish, CityKirkland, CityRenton, CityBellevue, CityRedmond, and CityIssaquah. We then checked the model utility and assumptions, after which we modified our model with a new response.

We assumed the model follows a linear regression method, which means that the relationship between our response ~ House price and our independent variables is linear. The model needs to meet the following requirements: the error terms are normally distributed, are independent, has a mean value of zero and a constant standard deviation. The complete hypothesized model is:

$$\text{Price}_i = \beta_0 + \beta_1\text{Bedroom}_i + \beta_2\text{Bathroom}_i + \beta_3\text{Sqft\_living}_i + \beta_4\text{Floors}_i + \beta_5\text{Sqft\_lot}_i + \beta_6\text{AgeoftheHouse}_i + \beta_{c1}\text{CitySeattle}_i + \beta_{c2}\text{CityRenton}_i + \beta_{c3}\text{CityBellevue}_i + \beta_{c4}\text{CityRedmond}_i + \beta_{c5}\text{CityKirkland}_i + \beta_{c6}\text{CityIssaquah}_i + \beta_{c7}\text{CityKent}_i + \beta_{c8}\text{CityAuburn}_i + \beta_{c9}\text{CitySammamish}_i + \varepsilon_i$$

$$\epsilon_i \overset{Indp}{\sim} Normal(0, \sigma)$$

where the Price_i refers to the price of the ith house. The variable Sqft_living_i is the square footage of living area of the ith house, Sqft_lot_i refers to the square footage of lot of the ith house, Floors_i refers to the number of floors of the ith house, Bedroom_i and Bathroom_i refers to the number of each amenity of the ith house and Ageofthehouse_i refers to simply as the age of the ith house. CitySeattle_i, CityRenton_i, CityBellevue_i, CityRedmond_i, CityKirkland_i, CityIssaquah_i, CityKent_i, CityAuburn_i and CitySammamish_i are nine dummy variables where 1 represents that the ith house is located in Seattle, Renton, Bellevue, Redmond, Kirkland, Issaquah, Kent, Auburn and Sammamish city, respectively and Federal Way is retained as the reference level. β_0 represents the intercept, β_1, β_2, β_3, β_4, β_5 and β_6 represent slope of the

corresponding quantitative variables and β_{c1}, β_{c2}, β_{c3}, β_{c4}, β_{c5}, β_{c6}, β_{c7}, β_{c8} and β_{c9} represent coefficients of the corresponding dummy variables.

Table 3 & 4: Model Summary and ANOVA

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.796[a]	.633	.631	214783.0340

a. Predictors: (Constant), CItySammamish, sqft_lot, cityKirkland, CityKent, floors, CityRenton, CityAuburn, cityBellevue, CityRedmond, bedrooms, CityIssaquah, Age of the house, sqft_living, bathrooms, CitySeattle

b. Dependent Variable: price

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.703E+14	15	1.802E+13	390.660	.000[b]
	Residual	1.567E+14	3396	4.613E+10		
	Total	4.270E+14	3411			

a. Dependent Variable: price

b. Predictors: (Constant), CItySammamish, sqft_lot, cityKirkland, CityKent, floors, CityRenton, CityAuburn, cityBellevue, CityRedmond, bedrooms, CityIssaquah, Age of the house, sqft_living, bathrooms, CitySeattle

From the summary of the model fitting shown in Table.3, the multiple coefficient of determination (R^2) value is 0.633, indicating that approximately 63.3% of variability in the house price can be explained by the independent variables. It means the linear relationship between the house price and the independent variables is relatively strong. In our model summary, the root mean square error is $214,783, which means that after taking our independent variables into account, our error has a standard deviation of about $214,783. It indicates that the addition of more variables can further improve the model. As seen in Table 4, the p-value of F-test is 0.000, which is less than 0.05, so the model is statistically significant overall. The test result further supports the statement that the model is useful.

For model validity, we generated residual analysis to check the model assumptions. First, we check for lack of fit and homogeneity by using the Residual vs. Fitted Value Plot as shown in Fig. 4. The reason we check the plot is to see that the standard error is constant. The residual plot shows that the residuals are not evenly spreading out and display a multiplicative pattern. In other words, our model shows lack of fit and the standard error is not constant. It violates our initial assumption and therefore, the model is invalid. We will use the variance stabilizing transformation on the response to improve the model. Because the identified pattern is Multiplicative, we want to create a transformed response with the natural log function.

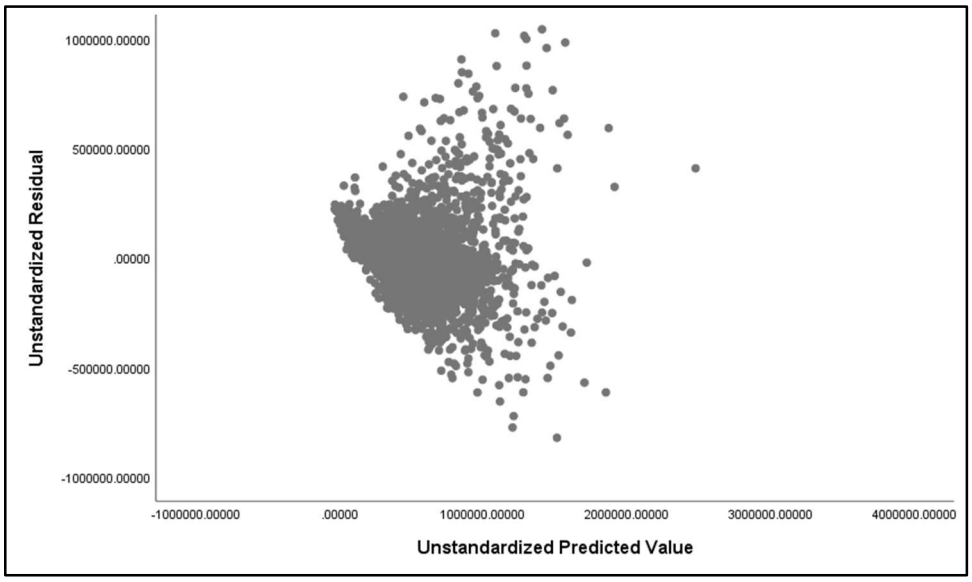

Fig. 4: Scatter plot of Unstandardized Residual vs Unstandardized Predicted Value

II) Transformed Model

In the transformed model, the relationship is described between our independent variables and transformed response. The transformed model still needs to meet the following requirements: the error terms are normally distributed, are independent, has a mean value of zero and a constant standard deviation. The complete transformed model is:

$$\ln(\text{Price}_i) = \beta_0 + \beta_1\text{Bedroom}_i + \beta_2\text{Bathroom}_i + \beta_3\text{Sqft\_living}_i + \beta_4\text{Floors}_i + \beta_5\text{Sqft\_lot}_i + \beta_6\text{AgeoftheHouse}_i + \beta_{c1}\text{CitySeattle}_i + \beta_{c2}\text{CityRenton}_i + \beta_{c3}\text{CityBellevue}_i + \beta_{c4}\text{CityRedmond}_i + \beta_{c5}\text{CityKirkland}_i + \beta_{c6}\text{CityIssaquah}_i + \beta_{c7}\text{CityKent}_i + \beta_{c8}\text{CityAuburn}_i + \beta_{c9}\text{CitySammamish}_i + \varepsilon_i$$

$$\epsilon_i \overset{Indp}{\sim} Normal(0, \sigma)$$

where the $\ln(\text{Price}_i)$ refers to the natural log of price of the ith house. The variable Sqft_living_i is the square footage of living area of the ith house, Sqft_lot_i refers to the square footage of lot of the ith house, Floors_i refers to the number of floors of the ith house, Bedroom_i and Bathroom_i refers to the number of each amenity of the ith house and Ageofthehouse_i refers to simply as the age of the ith house. CitySeattle_i, CityRenton_i, CityBellevue_i, CityRedmond_i, CityKirkland_i, CityIssaquah_i, CityKent_i, CityAuburn_i and CitySammamish_i are nine dummy variables where 1 represents that the ith house is located in Seattle, Renton, Bellevue, Redmond, Kirkland, Issaquah, Kent, Auburn and Sammamish city, respectively and Federal Way is retained as the reference level. β_0 represents the intercept, β_1, β_2, β_3, β_4, β_5 and β_6 represent slope of the corresponding quantitative variables and β_{c1}, β_{c2}, β_{c3}, β_{c4}, β_{c5}, β_{c6}, β_{c7}, β_{c8} and β_{c9} represent coefficients of the corresponding dummy variables.

Table 5 & 6: Model Summary and ANOVA of the transformed model

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.833[a]	.694	.693	.28757

a. Predictors: (Constant), citySammamish, sqft_lot, cityKirkland, cityKent, floors, cityRenton, cityAuburn, cityBellevue, cityRedmond, bedrooms, cityIssaquah, Age of the house, sqft_living, bathrooms, citySeattle

b. Dependent Variable: logprice

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	638.013	15	42.534	514.325	.000[b]
	Residual	280.846	3396	.083		
	Total	918.859	3411			

a. Dependent Variable: logprice

b. Predictors: (Constant), citySammamish, sqft_lot, cityKirkland, cityKent, floors, cityRenton, cityAuburn, cityBellevue, cityRedmond, bedrooms, cityIssaquah, Age of the house, sqft_living, bathrooms, citySeattle

With reference to Table 5, we can see that multiple coefficient of determination (R^2) for the transformed model is 0.694. This indicates that 69.4% of the variability in the model can be explained by the independent variables. We can also observe that our adjusted coefficient of determination is 0.693. This means that 69.3% of the variability is explained after accounting for the model size. Since there isn't much difference between the coefficient of determination and the adjusted coefficient of determination, we can conclude that the model is good after accounting for the size of the model. The standard error of the model is 0.28757, given a moderately high coefficient of determination is good. However, addition of more independent variables can further improve the model. In addition, we can observe that the p-value of F-test (0.000) is less than 0.05, thus the model is statistically significant and can be considered useful at 5% significance level. Overall, the transformed model is useful.

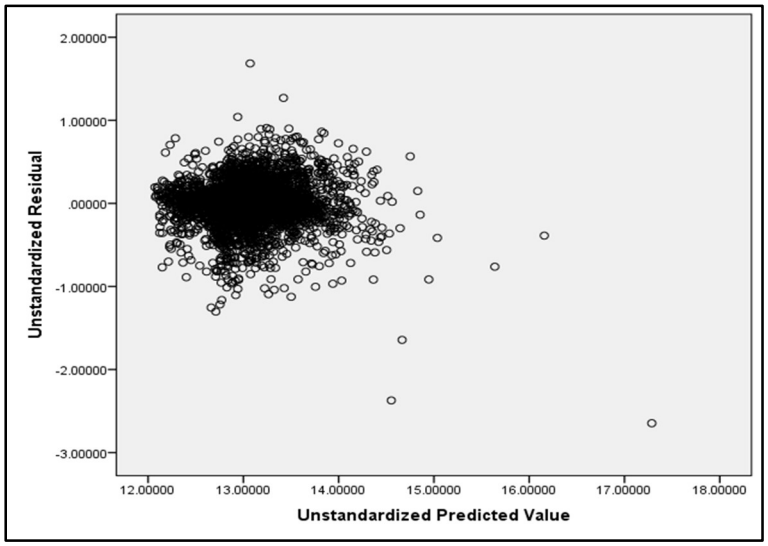

Fig. 5: Scatter plot of Unstandardized Residual vs Unstandardized Predicted Value for the transformed model

To check for lack of fit and homogeneity, we produced the Residual vs. Fitted Value Plot as shown in Fig. 5. Upon observing the pattern of the plot in the transformed model, we can see that the residuals are spreading out evenly and that there is no other significant pattern that would warrant further transformation of the model. The transformed model confirms a linear relationship between the new response and independent variables. This pattern also gives no indication of heterogeneity issues.

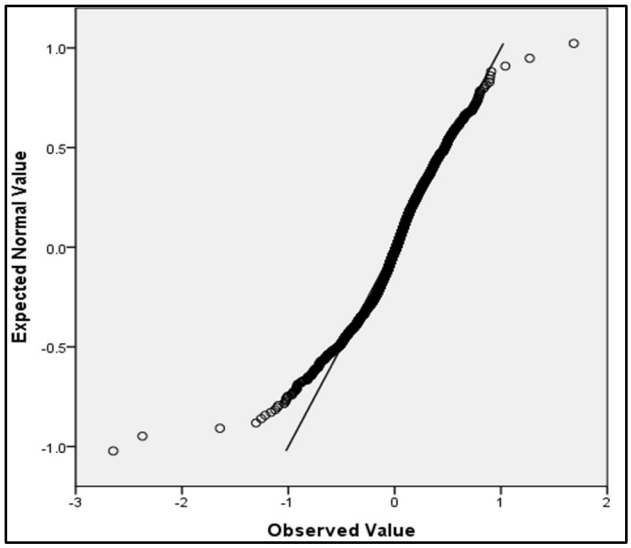

Fig. 6: Normal Q-Q of Unstandardized Residual for the transformed model

We then checked the normality of the model to determine that the errors are normally distributed. We constructed a Quantile-Quantile (Q-Q) plot of the residuals to evaluate this assumption. The Q-Q plot (Fig. 6) displays that there are some deviations from normal towards both ends of line, but they are still close enough to the diagonal line which represents normal. Additionally, we also constructed a histogram showing the distribution of the unstandardized residuals (Fig. 7). This distribution is almost normal in this

plot. Hence, we conclude that there is no strong violation of normality and that the error terms are approximately normal. Notwithstanding, we do suggest some future measures to investigate this issue further for more clarity. We also checked the assumption that error terms are independent of each other. For this, there is no evidence to suggest that the data has a time series structure and there is no need to perform the Durbin-Watson test. Thus, we can conclude that the assumption is valid and residuals are independent of each other. Overall, the transformed model does not violate any assumptions.

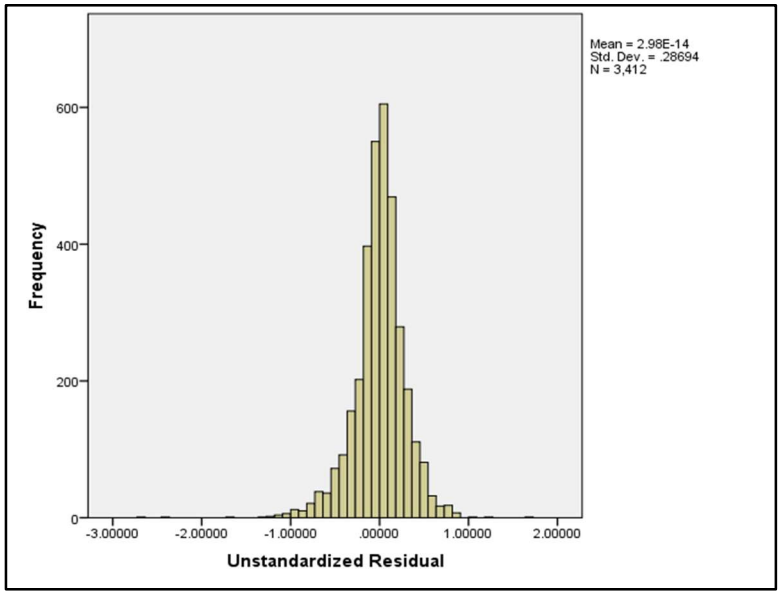

Fig. 7: Histogram of Unstandardized Residuals for the transformed model

As observed in Table 7, we will interpret the co-efficient values of each term to understand the relationship between the independent variables and natural log of house price. In the following discussions, we will not use the unit "$" for our response as the unit of natural log of house price is dissimilar to the unit of house price. We will interpret our response as unitless. The model estimates that the mean of the natural log of house price is zero when number of bedrooms, number of bathrooms, square footage of living area (ft^2), square footage of lot (ft^2), number of floors, and age of the house (years) are set to zero and city is Federal Way. This however, is not possible as the value of our quantitative independent variables cannot be equal to zero practically. Given this, only the variable "Age of the house" can assume that value.

In terms of bedrooms, one unit increase in the number of bedrooms is associated with the mean of the natural log of the house price decreasing by 0.039, when all other variables are held constant. This is a crucial finding as it indicates that while larger houses are high-priced, larger houses with too many bedrooms are unappealing. It is in contrast to the general believe that houses with a greater number of bedrooms tends to have higher price. In terms of bathrooms, one unit increase in the number of bathrooms is associated with the mean of the natural log of the house price increasing by 0.101, when all other variables are held constant. In terms of square footage of living area, one square foot increase in the square footage of living area is associated with the mean of the natural log of the house price changing by

approximately zero, when all other variables are held constant. This relationship will need further investigation in future research.

In terms of square footage of lot, one square foot decrease in the square footage of lot is associated with the mean of the natural log of the house price decreasing by 3.012e-7, when all other variables are held constant. We found that this co-efficient value is not statistically significant at 5% significance level. This indicates that the square footage of lot does not provide values relative to the variability in our response i.e., this variable will not add much to the mean value of natural log of house price. In terms of floors, one unit increase in the number of floors is associated with the mean of the natural log of the house price increasing by 0.103, when all other variables are held constant. In terms of age, one-year increase in the age of the house is associated with the mean of the natural log of the house price increasing by 0.003, when all other variables are held constant.

Table 7: Coefficients for the transformed model

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	11.585	.036		324.792	.000
	bedrooms	-.039	.007	-.069	-5.577	.000
	bathrooms	.101	.011	.151	8.942	.000
	sqft_living	.000	.000	.566	34.836	.000
	sqft_lot	-3.012E-7	.000	-.017	-1.641	.101
	floors	.103	.011	.109	8.971	.000
	Age of the house	.003	.000	.164	12.047	.000
	citySeattle	.579	.026	.556	22.219	.000
	cityRenton	.215	.029	.116	7.343	.000
	cityBellevue	.771	.030	.408	26.002	.000
	cityRedmond	.664	.031	.323	21.657	.000
	cityKirkland	.642	.032	.282	20.148	.000
	cityIssaquah	.532	.032	.233	16.498	.000
	cityKent	.047	.032	.020	1.454	.146
	cityAuburn	.016	.032	.007	.504	.614
	citySammamish	.599	.033	.252	18.214	.000

a. Dependent Variable: logprice

For houses in Seattle, the model estimates that the change in average of the natural log of the house price is 0.579 higher in Seattle than in Federal Way, when all other variables are held constant. For houses in Renton, the model estimates that the change in average of the natural log of the house price is 0.215 higher in Renton than in Federal Way, when all other variables are held constant. For houses in Bellevue, the model estimates that the change in average of the natural log of the house price is 0.771 higher in Bellevue than in Federal Way, when all other variables are held constant. For houses in Redmond, the model estimates that the change in average of the natural log of the house price is 0.664 higher in Redmond than in Federal Way, when all other variables are held constant.

For houses in Kirkland, the model estimates that the change in average of the natural log of the house price is 0.642 higher in Kirkland than in Federal Way, when all other variables are held constant. For houses in Issaquah, the model estimates that the change in average of the natural log of the house price is 0.532

higher in Issaquah than in Federal Way, when all other variables are held constant. For houses in Sammamish, the model estimates that the change in average of the natural log of the house price is 0.599 higher in Sammamish than in Federal Way, when all other variables are held constant.

For houses in Kent, the model estimates that the change in average of the natural log of the house price is 0.047 higher in Kent than in Federal Way, when all other variables are held constant. For houses in Auburn, the model estimates that the change in average of the natural log of the house price is 0.016 higher in Auburn than in Federal Way, when all other variables are held constant. For Kent and Auburn cities, we discovered that the co-efficient values are not statistically significant at 5% significance level. This indicates that these variables will not add much to the mean value of natural log of house price. In other words, the average houses prices in Kent and Auburn are not much different than the average house prices in Federal Way for the same number of bedrooms, number of bathrooms, square footage of living area (ft^2), square footage of lot (ft^2), number of floors, and age of the house (years).

Section 4 - Summary and Concluding Remarks

In conclusion, the variables we chose to analyze in the report are playing an important role in association with the natural log of house price in the ten chosen cities in the state of Washington. Some of the variables which include locations(cities), sqft_living, age of house, number of floors, number of bathrooms and number of bedrooms have a strong impact on the determination of house prices. The model indicates that Federal Way, Kent and Auburn are the least expensive cities on average among the ten cities which we have studied for houses with same number of bathrooms, number of bedrooms, age, square footage of living area, square footage of lot and number of floors.

In addition, we observe that change in square footage of living area does not affect our response value even though it is found to be statistically significant in our model. We can also note that number of bedrooms share a negative correlation with the natural log of house price when other independent variables are taken into account. Thus, if any type of analysis needs to be conducted in the future using this model, further investigation would be required to determine why the square footage of living area has a co-efficient value of zero and why the number of bedrooms has a negative correlation with the response. The drawback of the model is that the data we collected refers to houses that have been sold at a specific time, therefore, we cannot use the model to predict or infer the future prices of a house. We can only use the model for understanding the relationship between house prices and its correlated factors. Additionally, the dataset we analyzed is imbalanced in terms of cities, because the number of observations varies greatly for different cities. Therefore, this model would be limited because of the available data.

Appendix

1) Original Dataset: https://www.kaggle.com/shree1992/housedata

2) Graphs for reference:

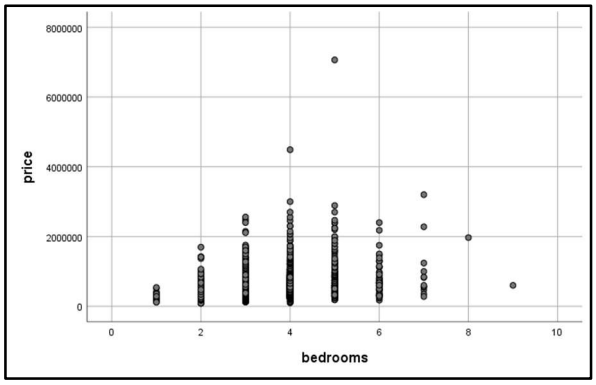

Fig. a: Scatter plot of Price vs Bedrooms

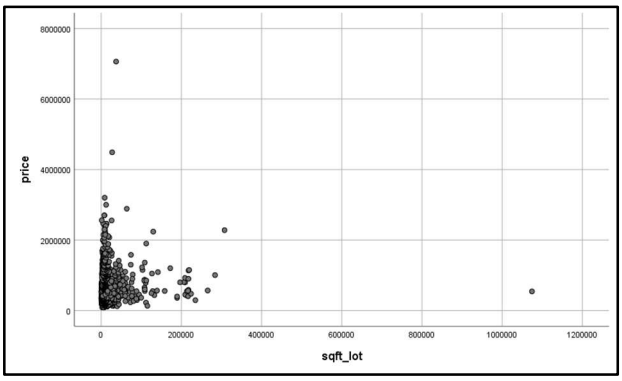

Fig. b: Scatter plot of Price vs Sqft_lot

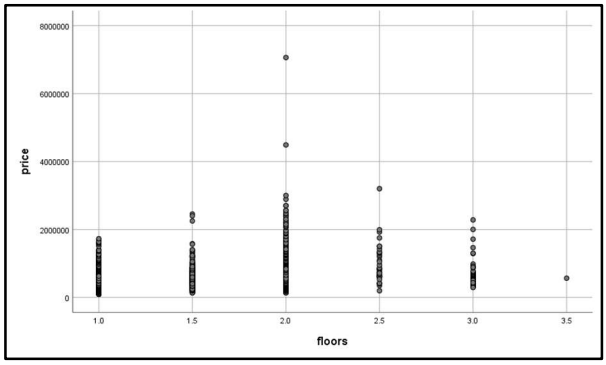

Fig. c: Scatter plot of Price vs Floors

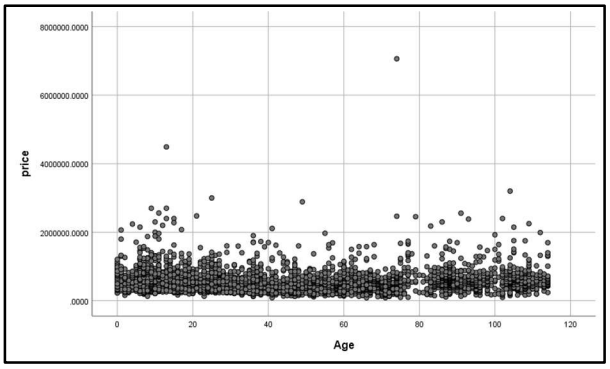

Fig. d: Scatter plot of Price vs Age