Introduction

Research Problem & Description

"How do the number of bedrooms, number of bathrooms, square footage of living space, lot size, number of floors, waterfront status, and view quality (interior of the house) influence house prices in King County?"

As of June 2024, house prices in King County are up 7.7% from the previous year. The median price of all houses sold was $892,000 (Redfin, 2024). We think that understanding the factors that drive house prices is crucial for buyers, sellers, and policy makers. According to economist Taylor Marr, the main factor that affects the price of a property is the house itself （Williams, 2021). There is a report that shows that buyers of higher-priced homes will place a different value on features in the home, such as the number of bedrooms or bathrooms, than consumers buying lower-priced homes (Zietz, J., Zietz, E.N. & Sirmans, G.S., 2008). Therefore, this study focuses on identifying the impact of various housing characteristics, such as the number of bedrooms, bathrooms, living space square footage, lot size, number of floors, waterfront status, and view quality on house prices in King County, Washington. By examining how these features affect prices, we will provide valuable insights that can help potential homebuyers make informed decisions and contribute to a broader understanding of the market. The dataset we used is derived from house sales in King County, USA (Kiyakoglu, 2019). We have cleaned the dataset to contain 8 variables for 21,613 observations to demonstrate a better statistical relationship.

Data Characteristics

Table 1. Introduction of Variables in Dataset

Variable Name	Type	Description
`price`	numerical	Price of the house
`bedrooms`	numerical	Number of bedrooms
`bathrooms`	numerical	Number of bathrooms
`sqft_living`	numerical	Square footage of the living space
`sqft_lot`	numerical	Square footage of the lot
`floors`	numerical	Number of floors
`waterfront`	categorical with a level of 5	Whether the house is a waterfront property
`view`	categorical with a level of 2	Quality of the view

Challenges on This Study

The primary challenge in studying house prices is the complex interplay between various factors influencing them. The causation of multicollinearity among variables, such as living space square footage and the number of bedrooms or bathrooms may be highly correlated, can make it difficult to determine the individual effect of each variable on house prices. Additionally, variability in house prices may increase with the size or value of a home, leading to heteroscedasticity, where the variance of errors is not constant. This

can violate assumptions of linear regression models.

Methodologies

To address the research problem, we first used the IQR method to identify outliers. Values that are below Q1 – 1.5IQR or above Q3 + 1.5IQR were removed from our dataset. We used the `sum(is.na())` function in R to detect missing values, and the result is that this dataset does not contain any missing values. We decided to employ the Multiple Linear Regression (MLR) as our primary analytical tool for the study. MLR is particularly useful for this study since it allows us to model the relationship between house prices (dependent variable) and multiple housing features (independent variables). The linear regression model will provide coefficients for each predictor. Additionally, a comprehensive Exploratory Data Analysis (EDA) will also be conducted to visualize and understand the distributions and relationships between variables. Moreover, we will also access the presence of multicollinearity among predictor variables by calculating the Variance Inflation Factor (VIF) and examining the correlation matrix.

Results

Firstly, we discovered that the distribution of logged house prices shows a rightward skew by illustrating the histogram of price, indicating that while most houses are priced within a certain range, there are some significantly higher-priced homes. We also observed that there are positive relationships between house prices and numerical variables such as the number of bedrooms, bathrooms, and living space. Notably, the boxplots demonstrate that waterfront homes tend to have higher prices, and house prices increase steadily with better view quality. These observations can draw to our conclusion, which is all the variables considered in this study could potentially influence house prices in King County.

The correlation matrix further supports the findings. There are strong positive correlations between house prices and variables like square footage of living space (0.702) and the number of bathrooms (0.525). Additionally, the Variance Inflation Factor (VIF) values for these variables are all below the threshold of 5, which means multicollinearity is not a significant concern in this dataset. This analysis highlights the importance of considering multiple features when analyzing house prices in the region.

Relevance to STA302

The research project is highly related to STA302, as it applies fundamental concepts of statistical modelling and data analysis learned in the course to a realistic problem. The project reinforces the theoretical knowledge of Multiple Linear Regression and practical aspects such as data cleaning and interpretation of results. By working on a real dataset, we manage to follow along the process of conducting research, from statistical analysis to model evaluation. In our dataset, we will analysis how multiple factors influence the house price by using Multiple Linear Regression. It helps us find the relationship between multiple independent variables and a dependent variable (house prices). Data cleaning is also very crucial, since we are dealing with real-

world data, and it must be realistic.

Methodologies

Model Formulation:

The house price, Y , is the dependent variable, and the following independent variables, X , are considered:

- X_1 : : Number of bedrooms
- X_2 : : Number of bathrooms
- X_3 : : Square footage of living space
- X_4 : : Lot size (square footage of the lot)
- X_5 : : Number of floors
- X_6 : :Waterfront status (a categorical binary variable: 1 if the house is on the waterfront, 0 otherwise)
- X_7 : : View quality (a categorical variable representing the quality of the view, has 5 levels)

The model can be represented as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \epsilon$$

where β_0 is the intercept, $\beta_1, \ldots, \beta_7$ are the coefficients for the corresponding independent variables, and ϵ is the error term.

Objective:

This research's main objective is to investigate the impact of various housing characteristics on house prices in King County, Washington. Specifically, we aim to understand how the number of bedrooms, number of bathrooms, square footage of living space, lot size, number of floors, waterfront status, and view quality influence house prices. The goal is to develop a regression model that captures the relationships between these variables and house prices, providing insights that can guide buyers and real estate investors.

Estimation process & statistical inference procedures

In our chosen dataset, the first column is the ID of the house and is not an appropriate variable, so we do not reference this column. The second column is the date, which is also not an appropriate variable. The third to the tenth column （from PRICE to VIEW） is the main part of our analysis. The third column is the dependent variable, y, while the column after that are the independent variables, x. This is because we think that these data give a good indication of the effect of home interiors on the price of a home. We found no missing values in these plots by running the dataset in R studio. So, we get Cleaning Data. After the implementation of the IQR method, we know the rows which contain outliers, and we may avoid these rows in the following method

to ensure all the data could not heavily skew analysis results. Then, we used Exploratory Data Analysis (EDA).

Since our chosen data set is about the "house price", the price of which is a huge value, such as "$221,900" in the second row of the data set. Histogram might not look like normal distribution. To ensure our dependent variable Y follows Normal distribution, we will check if the histogram is normally distributed. If not, we will take $\log$ to lower the value to make sure the histogram looks like normal distribution.

To detect the possible multi-collinearity, we will check the GVIF (Generalized Variance Inflation Factor). It is used when predictors are not binary or continuous but categorical with multiple levels. For categorical variables with more than one level, GVIF is used instead of VIF. Since our chosen dependent variable X_7 is categorical more than one level, we decided to use GVIF. It measures how much larger the variance is because of stronger muti-collinearity. Generally, when GVIF > 1 means some muti-collinearity presents, which is acceptable. When GVIF > 5, there is several muti-collinear, which is unacceptable. We do not want multi-collinearity, thus the smaller VIF, the better. If any GVIF > 5, we will the corresponding dependent variable X, until all GVIF do not larger than 5, we will consider this model as a reduced model. If no GVIF > 5, we will keep the full model.

Now, we will get the appropriate model. We will check the assumptions: Linearity; Constant Variance; Uncorrelated Errors; and Normality. By drawing the residual plots, we will see if the model violates linearity, constant variance or uncorrelated error. If the residual plot has a curve, then it violates linearity. If the residual plot spreads out, then it violates constant variance. If the residual plot has a cluster, then it violated uncorrelated errors. By drawing the QQ-plot, we will see if it shows a straight line, then normality holds. By applying the transformations to the variables if any of the model violates the 4 assumptions and re-fit the model, and check whether the violations are resolved.

Since our n (sample size) is larger than the condition of AIC_c, we will use Akaike information criterion (AIC) to check if the model is good. If AIC is small, then we can conclude that our model is good. Otherwise, we will create and compare different models to identify the one with the lowest AIC. We may also simplify the model.

To check the outlier and leverage point. We will calculate the value for the points, if the outlier meets the cut-off interval which is between [-2, 2], then it is an outlier. If $h_{ii} > 2\frac{P}{n}$, then it is a leverage point. Since we have a huge dataset, we may face lots of outliers and leverage points. We will state if all the points whether they are one of them, if we do have a lot, we will state in the limitation.

Additionally, we will consider the adjusted R^2_{adj}, which adjusts for the number of predictors in the model, to ensure that the inclusion of additional variables genuinely improved the model's explanatory power.

Novelties, advantages or importance

One of the novelties of our approach was the application of the Generalized Variance Inflation Factor (GVIF) to assess multicollinearity in categorical predictors with multiple levels. This method will allow us to

effectively identify and address multicollinearity, which is not typically handled well by standard VIF estimate of categorical variables.

The AIC model will ensure that our final model be both precise and highly predictive. This method will provide a clear advantage in identifying the most significant predictors, leading to more reliable and interpretable results.

The methodologies will be crucial for accurately modeling the complex relationships between housing characteristics and prices. We will be able to provide insights for homebuyers and real estate professionals, helping them understand which factors most significantly influence house prices in King County. Our approach ensures that these findings are both statistically stable and practically relevant.

Results

Step 1: We performed EDA first.

Table 2. summary table of numerical variables in the whole dataset

Variables	Price	Bedrooms	Bathrooms	Living	Lot	Floors
Min	75,000.0	0	0	290.0	520.00	1
Q1	321,950.0	3	1.75	1,427.0	5,040.00	1
Median	450,000.0	3	2.25	1,910.0	7,618.00	1.50
Mean	540,088.1	3.37	2.11	2.079.9	15,106.97	1.49
Q3	645,000.0	4	2.50	2,550.0	10,688.00	2
Max	7,700,000.0	33	8	13,540.0	1,651,359.00	3.5

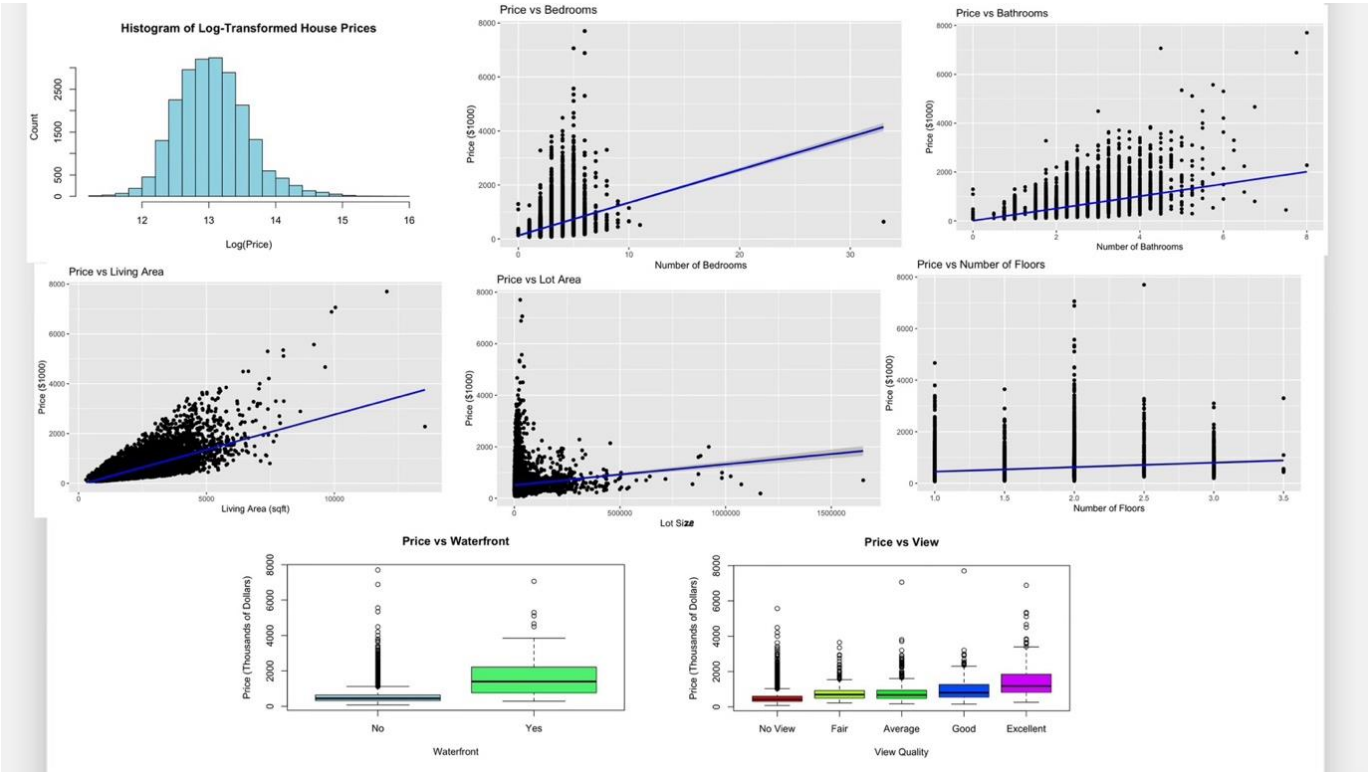


Figure 1. histogram of log transformed house price & scatterplots of house price vs. other numerical variables & boxplots of categorical variables

According to the above, we found that the normal distribution of logged house price is right-skewed. Besides, there are also positive linear relationships between house price and other numerical variables. According to the boxplot of price and waterfront, it is obviously reflected that houses that are waterfront tend to have higher prices. The boxplot of price and view also demonstrates a stable increasing trend in house price when the quality of view is improving. Therefore, all the variables might potentially influence the house price in King County, Washington.

Step 2: Check the Correlation matrix

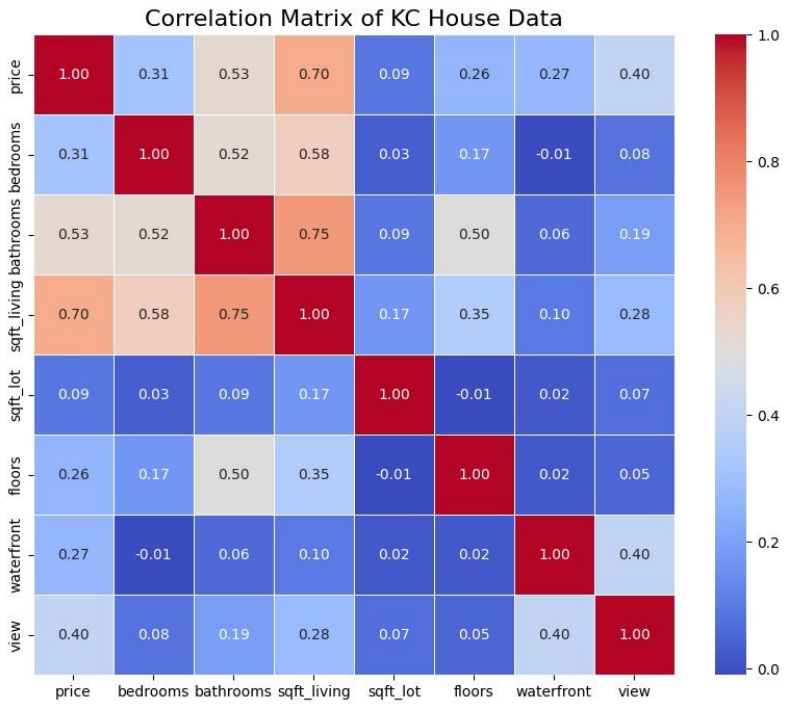


Figure 2. Correlation matrix of the King County house data

Some of the values are large. For example, in Figure 2, sqft_living & bathrooms has the value of 0.75. This might be potential multi-collinearity. We will re-check it uses VIF in the following step.

Step 3: Summarize the initial model

Table 2. Estimated Coefficient and Standard Error for variables in model

Variable	**Estimated Coefficient**	**Standard Error**
intercept	67640	7277
X_1	-46450	2237
X_2	6604	3600
X_3	281.8	3.047
X_4	-0.389	0.04083
X_5	8905	3580
X_6	552900	20910
X_7	73740	2464

The actions taken above summarized to our initial model:

$$\hat{y} = 67640 + (-46450)\ X_1 + 6604\ X_2 + 281.8\ X_3 + (-0.389)\ X_4 + 8905\ X_5 + 552900\ X_6 + 552900\ X_7 + \epsilon$$

Step 4: Check the QQ-plot

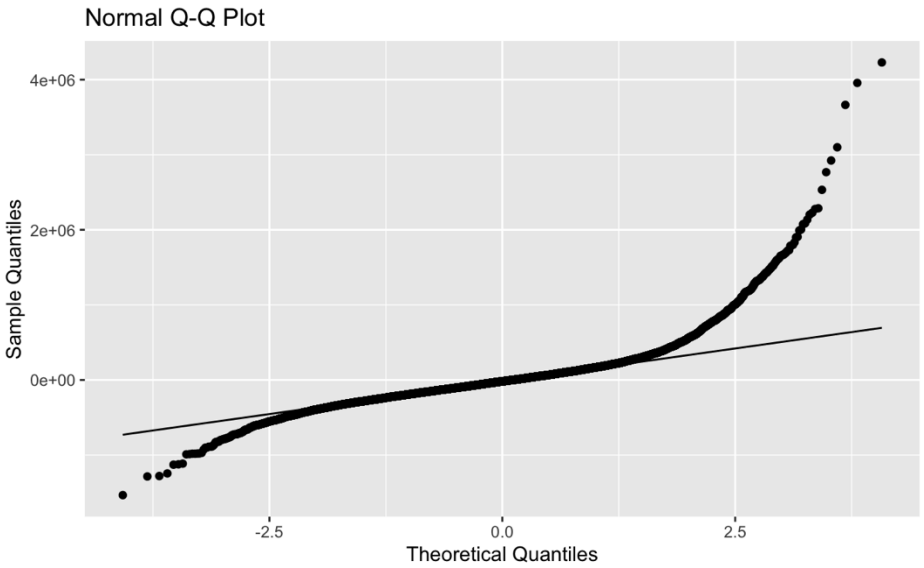


Figure 3. Normal Q-Q plot of initial model

After drawing the initial model's QQ-plot, we can clearly see it is not Normal. Then we decided to log on to this model.

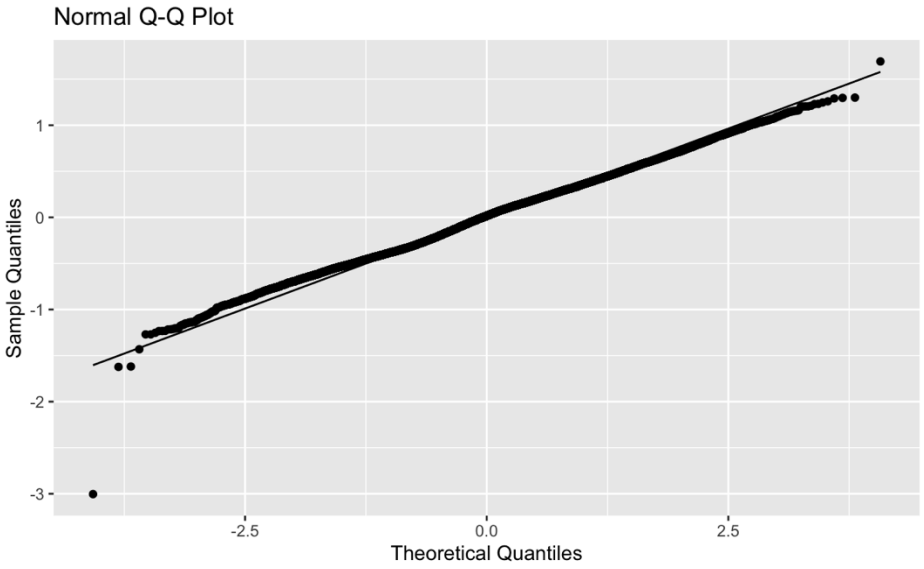


Figure 4. Normal Q-Q plot of Log transformed model

We can see the Normal Q-Q Plot of the log model is perfectly linear. We denote this model as second model.

Step 5: Check the Residual Plots & Apply EDA for log transformed model & Model assumptions

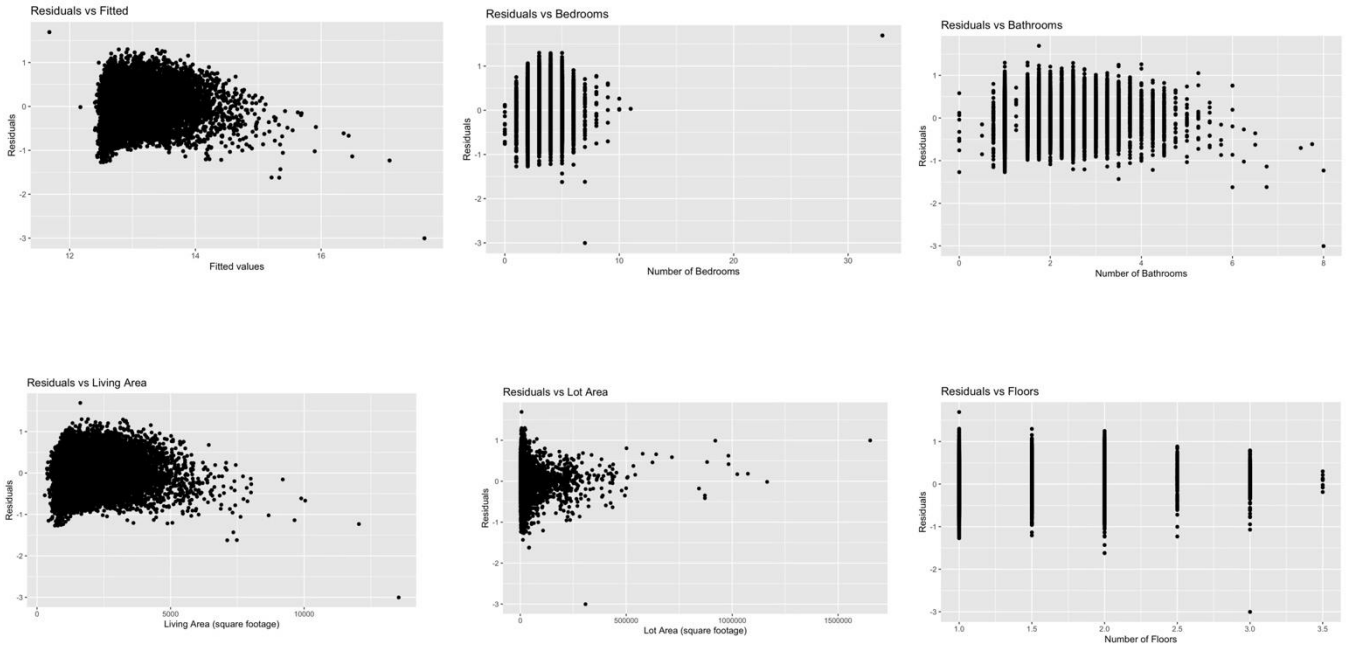


Figure 5. Residual plots for log transformation model

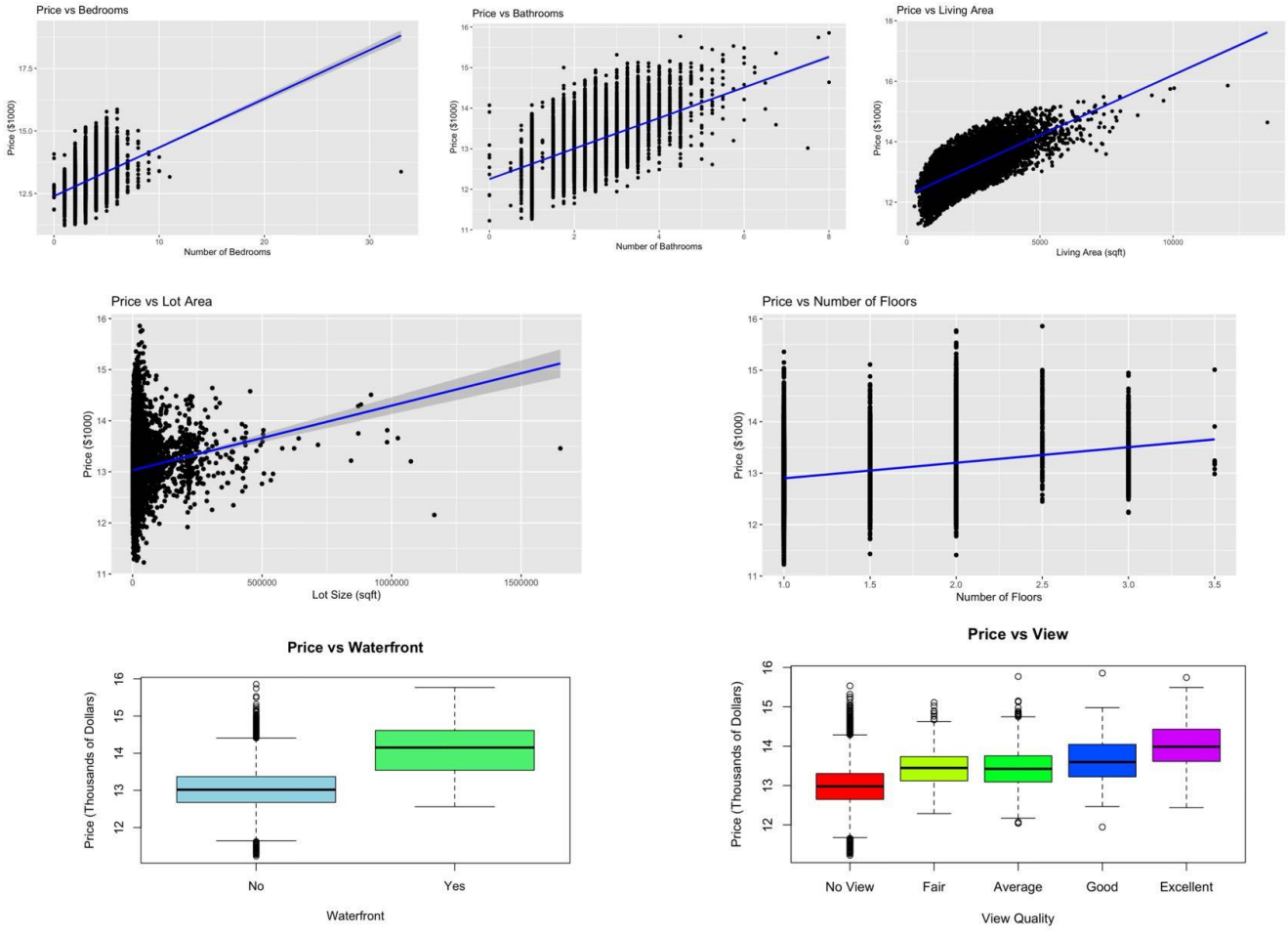


Figure 6. scatterplots of logged price vs. numerical variables & boxplots of price vs. categorical variables

After drawing both boxplot and residuals plot, we realized our log transformation model (second model) is

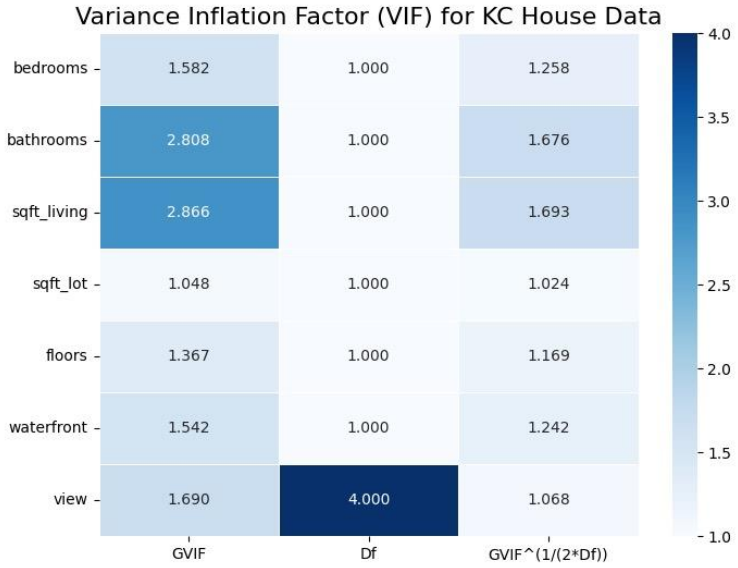
much better than the initial model, since they show a strong linear relationship. Therefore, we will use the second model in the following steps.

For the 4 model assumptions, we can say that the second model is better than the initial model. Since the QQ-plot in step 4 shows a straight line, so normality holds. By comparing the two models, we can say the second model is much better than the initial model. There is no curve on the residual plots, so linearity holds. There is no spreading out pattern, so constant variance holds. There is no cluster for the points on the plots, so uncorrelated error holds.

Step 6: Summarize the Log Transformed Model

$\hat{y}$ = 12.2300 + (-0.0383) X_1 + 0.0296 X_2 + 0.0003 X_3 + (-2.945e-07) X_4 + 0.0688 X_5 + 0.2824 X_6 + 0.0990 X_7 + ϵ

Step 7: Check VIF



After checking each predictor in the model, we would say that the model contains few multi-collinearities since GVIF for all the dependent variables are less than 5, which is great. This ensures that predictors in our model should not be removed.

Step 8: Comparison Between Initial and Transformed Models

The AIC value of our initial model is evaluated to be 597323.4, while the AIC value of the log transformed model decreases to 17787.68. This effectively proves that the transformed model demonstrates the linear relationship between price and the independent variables more accurately than the initial model.

	Initial Model	**Log Transformed Model**
N of Predictors	7	7
R^2_{adj}	0.5611	0.5195
AIC	597323.4	17787.68
(G)VIF Violation (>5)	0	0

Therefore, our final model remains the same:

$\hat{y}$ = 12.2300 + (-0.0383) X_1 + 0.0296 X_2 + 0.0003 X_3 + (-2.945e-07) X_4 + 0.0688 X_5 + 0.2824 X_6 + 0.0990 X_7 + ϵ

This model demonstrates a better understanding of how the number of bedrooms, number of bathrooms, living space, lot space, number of floors, waterfront, and view quality impact house prices in King County. According to our model formula, we see variables such as X_1 and X_4 (number of bedrooms and lot area) have negative coefficients, which is interesting as it indicates that an increase in number of bedrooms and a large lot may cause a decrease in house prices. Other variables except X_1 and X_4 demonstrate positive relationships with house prices. Moreover, variables such as X_3, X_5, X_6, and X_7 (living space, number of floors, waterfront status, and view quality) significantly impact house prices. Overall, the model explains about 52% of the variance in house prices.

Conclusion

This study aimed to explore the impact of various housing characteristics on house prices in King County, Washington. By analyzing factors such as the number of bedrooms, number of bathrooms, square footage of living space, lot size, number of floors, waterfront status, and view quality, we developed a regression model that provides valuable insights into how the interior factors effect house prices. And we find that: Each additional bedroom or the lot size associated with a slightly decrease in the price, suggesting that larger numbers of bedrooms do not necessarily lead to higher prices and maybe after a certain point the lot size does not add much value. The other five factors are different from them, these factors have positive effects on house prices, which means the larger they are, the higher the price.

These findings confirm that housing characteristics such as living space, bathrooms, and location-related features like waterfront status and view quality are the most influential factors in determining house prices in King County. The study concludes that although traditional factors like the number of bedrooms and lot size play a role, but modern buyers in King County prefer living space, bathrooms, and location-related factors such as waterfront and views. Our log-transformed model effectively captures the relationships between these variables and provides a clear understanding of how each contributes to the overall price, making it a useful tool for predicting house prices in this region.

Limitation & Improvement:

Throughout the report, the data we collect from the website is the data of houses being sold all over the King Country, so the data size is extremely big which lead to many outliers and leverage points during the analysis of the data. By checking either the residual plot or the QQ plot, the outliers are especially obvious. To solve this problem in the future, we can use an advanced data set. For instance, we will categorize and organize the data by separating them into different groups according to their location. Creating the model with the housing

data in the neighborhood, and then comparing the results between various neighborhoods will provide more accurate conclusions. Secondly, by comparing the R^2_{adj} in the initial model and log transformed model, the R^2_{adj} is smaller in the log transformed model. Since the term cannot improve our log transformed model fit by an enough amount, the value of adjusted R^2_{adj} decreases by 0.04. However, we still choose to use the log transformed model. Due to the large number in the data set, we need to reduce the skewness of our measurement factors. Thus, we continue using the log transformed model to make our data follow a normal distribution or approximately normal distribution. Same solution to the previous shortcoming, if we can organize the data, or find some new updated data set, we can use the initial model to avoid the decreasing of R^2_{adj}.

Acknowledgement

Ziqi Cheng: Ziqi provided valuable references for our report, contributed to the Introduction part by adding insightful sentences and citations, and helped draft the sections on limitations and improvements.

Xinpeng Qi: Xinpeng was instrumental in formulating the model, defining the objectives, and detailing the estimation process and statistical inference procedures in Methodologies. Xinpeng also highlighted the novelties, advantages, and importance of our approach. Worked the result part with Yiyun Zhang.

Yihang Xu: Yihang contributed to the conclusion by writing two key paragraphs, providing a concise summary of our findings. Yihang also did some research about our project, chatted them with teammates.

Yiyun Zhang: Yiyun planned the project outline and concluded detailed steps, managed all aspects related to coding, and contributed research question description, data characteristics, challenges, methodologies, results summary, and relevance to STA302 in the Introduction part and mainly contributed to the Results part. Yiyun also provided all the output values in R (e.g. coefficients, AIC etc.) and all the graphs as well.

Reference

Kiyakoglu, B. Y. (2019). Predicting House prices. Retrieved from https://www.kaggle.com/code/burhanykiyakoglu/predicting-house-prices/data

Redfin. (2024). King County, WA housing market: House Prices & Trends. Retrieved from https://www.redfin.com/county/118/WA/King-County/housing-market

Williams, D. (2021). 6 features that determine a home's final sale price. Retrieved from https://money.com/how-to-price-a-home/

Zietz, J., Zietz, E.N. & Sirmans, G.S. Determinants of House Prices: A Quantile Regression Approach. J Real Estate Finance Econ 37, 317–333 (2008)