Inference: Seattle Housing Price

Pushya Jain, Karn Shoker, Parthkumar Patel, Arav Dewan

1. Introduction

1.1 Data

The dataset being used for this study is the Housing Price Prediction - Seattle dataset from Kaggle, originally sourced from the following link:
https://www.kaggle.com/datasets/samuelcortinhas/house-price-prediction-seattle.
This dataset contains detailed information on houses sold in Seattle, Washington, USA, between August and December 2022. Although it represents one of the latest datasets available, it's important to acknowledge potential discrepancies relative to current market conditions.

1.2 Variables

Variable	**Description**	**Datatype**
beds	Number of bedrooms	Numerical
baths	Number of bathrooms	Numerical
size	Total area of property (sqft)	Numerical
size_units	Units for size (sqft)	Categorical
lot_size	Total area of land where property is located (sqft/acre)	Numerical
lot_size_units	Units for lot size (sqft/acre)	Categorical
zip_code	Postal code of the property	Numerical
price	Price of the property (USD)	Numerical

1.3 Research Question

How do various covariates, such as physical and geographical features, correlate with the market price of houses in Seattle?

1.4 Motivation

- Seattle is one of North America's fastest growing and developing cities and it has a lot of promising job opportunities, particularly in the STEM fields which many STATs students see as an option to settle down in. Especially for BC residents, this is a great option since it's very close to the province.
- Understanding the housing market is crucial for prospective residents, buyers, and investors. This study aims to uncover relationships between housing features and market prices to provide valuable insights into the Seattle housing market.

2. Analysis

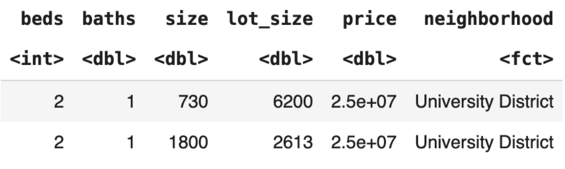
2.1 Data Preprocessing

- **Unit Conversion**: `lot_size`, includes properties measured originally in both `sqft` and `acre`, were converted entirely to `sqft` for uniformity. Additionally the columns `lot_size_unit` are `size_unit` which specify units were removed after conversion of `lot_size` to `sqft` since these columns no longer provide useful information.
- **Handling Missing Data**: Rows with `NA` values were removed. Linear regression requires complete data to produce unbiased coefficients and valid inferences. Missing values could lead to inaccurate predictions and incorrect conclusions.
- **Mapping ZIP code to neighborhood**: `zip_code` was mapped to corresponding neighborhood based on online resources. This improves interpretability for readers unfamiliar with Seattle ZIP codes and simplifies the analysis. Some ZIP codes which are within the same neighborhood were grouped to address imbalances in the dataset, as certain ZIP codes had very few properties in the dataset. .
- **Combine Train and Test sets**: The original dataset includes a train and test set. Since the objective of this study is to derive insights rather than evaluate predictive performance, combining the train and test datasets for final inference is reasonable. This approach will provide more robust results by leveraging the full dataset.

2.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is performed to visually examine the relationships between covariates and the response variable. This process provides insights that help guide our decisions when constructing linear regression models, ensuring they accurately capture the underlying data patterns.

To begin, we examine the distribution of price to identify potential extreme values. Outliers in the price variable can significantly impact model interpretability and may need to be addressed to prevent distortion of results. As shown in Figure 1, there appear to be two extreme outliers, notably a house with 2 bedrooms and 1 bathroom priced at USD $25 million. This value significantly deviates far beyond the third quartile and is excluded from further visual analysis and linear modeling since extreme values such as this can disproportionately influence the regression line, distorting the results from its true trend and leading to biased parameter estimates, thus reducing model accuracy and reliability. The outlier (price of $ 25 million for a house with 2 beds and 1 bathroom) significantly skews the average, making it unrepresentative of the dataset. Visual identification indicates it deviates notably from other data points [*Frost, J. (2023, May 18).]*. Further rigorous testing, such as residual analysis, will be performed to identify any additional outliers after fitting the linear regression model.



beds	baths	size	lot_size	price	neighborhood
<int>	<dbl>	<dbl>	<dbl>	<dbl>	<fct>
2	1	730	6200	2.5e+07	University District
2	1	1800	2613	2.5e+07	University District

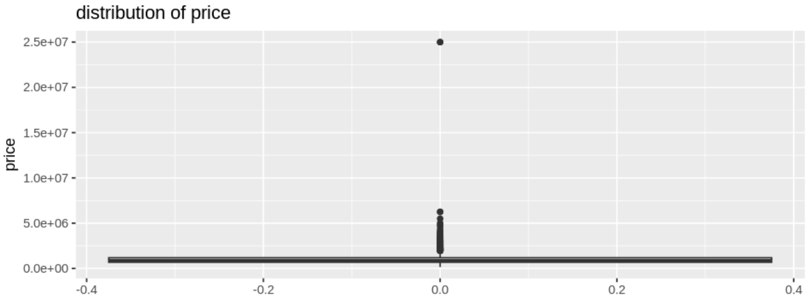


Figure 1: Detecting outliers in the data

Analysis of continuous covariates

Now let's analyze all the continuous covariates and examine their relationship with the price.

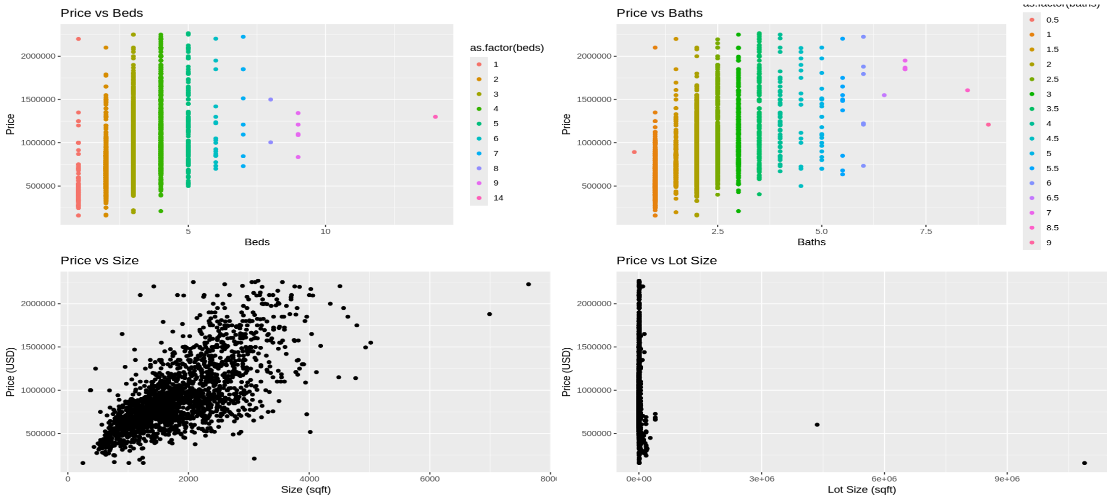


Figure 2: Plotting price vs continuous variables

Price vs. Beds and **Price vs. Baths**: Examining the plots, a moderate positive linear relationship between `price` and `beds` as well as `price` and `price` can be observed which suggests that houses with more amenities (bedrooms and bathrooms) generally command higher prices.

Price vs. Size: A strong positive linear relationship can be observed which indicates that larger houses are associated with higher prices, potentially making `size` a key predictor of `price`.

Price vs. Lot Size: The relationship is obscured by extreme values in `lot_size`, making it difficult to interpret. Applying a log transformation to `lot_size` could mitigate the effect of these extreme values and reveal a clearer relationship between `lot_size` and price.

Log Transformation of lot size

The log-transformed lot size exhibits a weak positive linear relationship with house price based on plot. While this transformation improves interpretability, the relationship remains weak, indicating that `log_lot_size` alone is not a strong predictor of `price`. Aside from size, all other covariates show only weak to moderate positive linear relationships with `price`. The variability suggests that the influence of other factors, such as neighborhood desirability and broader marker conditions play a role in the price of properties.

To further investigate the role of geographical factors, we next examine how house prices vary across neighborhoods. This analysis provides insight into how location impacts housing market trends.

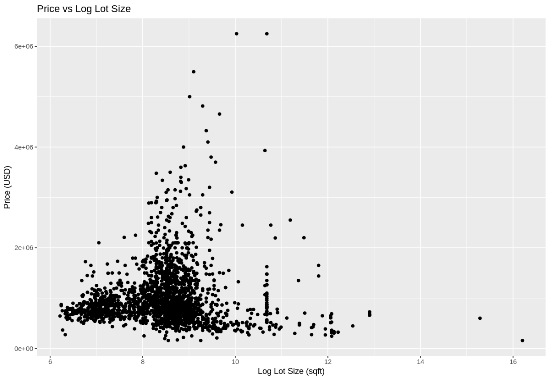


Figure 3: Price vs Log Lot Size plot

Price by Neighbourhood

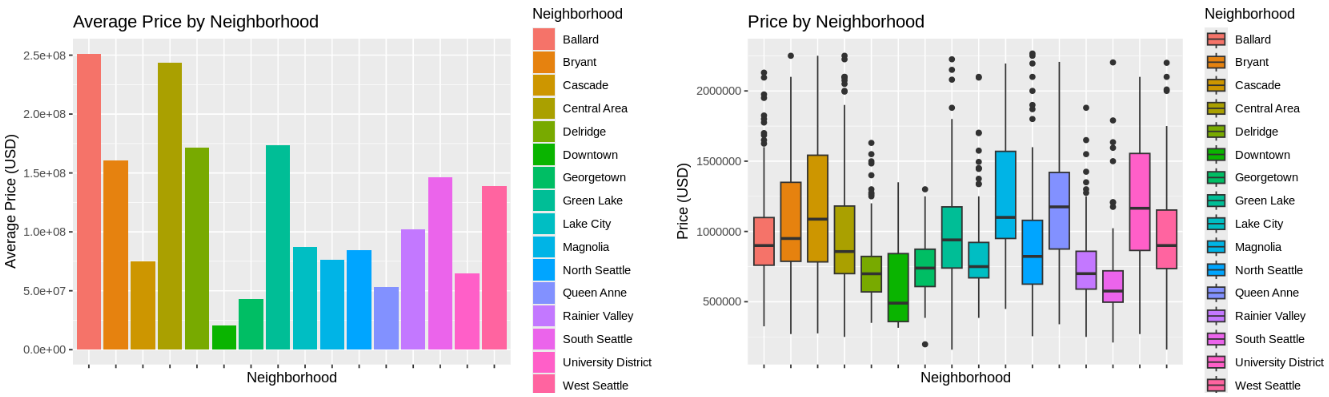


Figure 4: Price vs Neighbourhood plots

Average Price by Neighborhood: Neighborhoods such as Ballard and Central Area exhibit the highest average house prices, while Downtown and Georgetown have the lowest. However, averages can be misleading due to the influence of extreme values.

Price Distribution by Neighborhood: Extreme values in Central Area and Ballard inflate their average house prices. By examining the distributions: Cascade, University District, and Lake City show the highest median house prices. Downtown and South Seattle have the lowest median prices. This distribution highlights the variability in house prices across neighborhoods and provides a clearer understanding beyond averages.

Role of Neighborhood: The wide price ranges across neighborhoods indicate that location is a critical determinant of house price, offering valuable predictive information for modeling and analysis.

Conclusion of EDA

Based on the visualisations above, we have decided to apply a log transformation on `lot_size` to ensure the transformed lot size has a linear relation with the response variable `price` and will be included as a separate covariate in the model called `log_lot_price`.

2.3 Relationships Between Covariates

To examine the relationships between the continuous covariates and to look for potential multicollinearity, we will examine the correlation matrix of the covariates. This matrix provides insight into how covariates are related to each other. High correlation values may indicate multicollinearity, which can affect the stability and interpretability of regression coefficients in linear models.

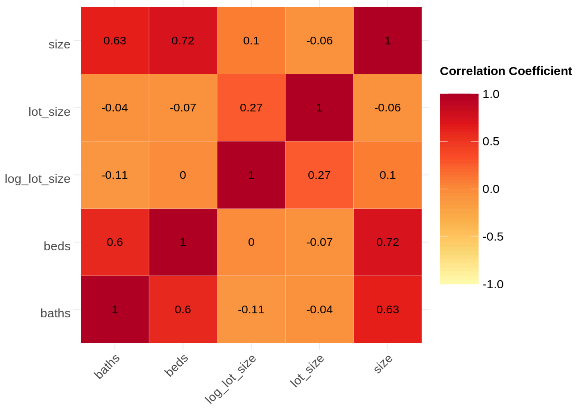


Figure 5: Correlation Matrix Plot

High Correlations: The correlation matrix reveals notable relationships among some of the continuous covariates

- `beds` and `size`: Correlation of 0.72 suggests a strong linear relationship.
- `baths` and `size`: Correlation of 0.63 indicates a moderately strong relationship.
- `beds` and `baths`: Correlation of 0.60 suggests these variables are also moderately correlated.

Multicollinearity Concern: The high correlations indicate potential multicollinearity when including beds, baths, and size together as covariates in the model. Multicollinearity can destabilize coefficient estimates and reduce interpretability. To address this concern, we will proceed cautiously and evaluate multicollinearity using Variance Inflation Factor (VIF) after fitting the model. This ensures the model remains robust and interpretable.

2.4. Model Selection

Backward Selection

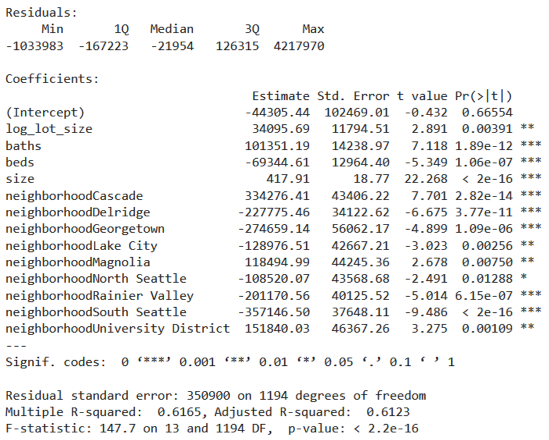
To refine our model, we employ a backward selection approach, which iteratively removes the least significant covariate until all remaining predictors are statistically significant at the 5% level. The least significant covariate is determined as the one with the highest p-value exceeding this threshold. This approach enhances interpretability, ensures only meaningful predictors are retained, and eliminates neighborhoods that do not add predictive value. It also improves model parsimony by reducing complexity while maintaining accuracy.

Initial Model: Covariates included in the initial model are `baths`, `beds`, `size`, `lot_size`, `log_lot_size`, and `neighborhood` (with neighborhoods represented as dummy variables). We are using an additive model instead of one with interactions because there are numerous potential interactions, and it's unclear which interaction would be meaningful to include in our model.

Simplifying Neighborhoods: We will apply a carefully chosen rule to neighborhood categorical variable. If a specific neighborhood's dummy variable is the least significant predictor in an iteration, we attempt to group it into a new neighborhood category labeled `Other`, which combines neighborhoods that do not provide useful information to the model. The `Other` category is created to improve the model's parsimony and interpretability. However, if the `Other` neighborhood remains insignificant, we exclude these neighborhoods from the model as well. We will track these dropped neighborhoods and interpret them as being similar to the baseline neighborhood *Ballard*. This approach allows us to compare the results of the baseline with the dropped neighborhoods post-analysis and ensure our model only includes coefficients which provide useful information to the model.

Model from Backward Selection: After several iterations, the final model kept the following predictors:

- **Continuous predictors**: `beds`, `baths`, `size`, and `log_lot_size`.
- **Categorical predictors**: Only neighborhoods that had a significant impact on `price`.



```
Residuals:
     Min       1Q   Median       3Q      Max
-1033983  -167223   -21954   126315  4217970

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -44305.44  102469.01  -0.432  0.66554
log_lot_size                     34095.69   11794.51   2.891  0.00391 **
baths                           101351.19   14238.97   7.118 1.89e-12 ***
beds                            -69344.61   12964.40  -5.349 1.06e-07 ***
size                               417.91      18.77  22.268  < 2e-16 ***
neighborhoodCascade             334276.41   43406.22   7.701 2.82e-14 ***
neighborhoodDelridge           -227775.46   34122.62  -6.675 3.77e-11 ***
neighborhoodGeorgetown         -274659.14   56062.17  -4.899 1.09e-06 ***
neighborhoodLake City          -128976.51   42667.21  -3.023  0.00256 **
neighborhoodMagnolia            118494.99   44245.36   2.678  0.00750 **
neighborhoodNorth Seattle      -108520.07   43568.68  -2.491  0.01288 *
neighborhoodRainier Valley     -201170.56   40125.52  -5.014 6.15e-07 ***
neighborhoodSouth Seattle      -357146.50   37648.11  -9.486  < 2e-16 ***
neighborhoodUniversity District 151840.03   46367.26   3.275  0.00109 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 350900 on 1194 degrees of freedom
Multiple R-squared:  0.6165, Adjusted R-squared:  0.6123
F-statistic: 147.7 on 13 and 1194 DF,  p-value: < 2.2e-16
```

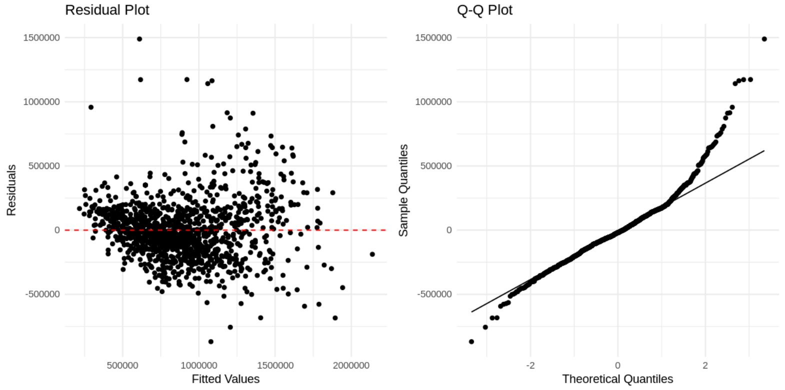


Figure 6: Summary of the final model using backward selection

Figure 7: Residual and Q-Q plots

The diagnostic plots in Figure 7 reveal violations of key assumptions in the initial linear model:

Homoscedasticity: The residual plot indicates non-constant variance among residuals. A fanning pattern is visible, suggesting that the variance of residuals increases with higher fitted values. A moderate negative linear relationship can also be observed among some residuals further confirming heteroscedasticity.

Normality: The Q-Q plot shows significant deviations from normality specifically near the tails, indicating that the residuals are not normally distributed.

Multicollinearity: Despite concerns raised by the correlation matrix, the Variance Inflation Factor (VIF) for the covariates is below the threshold of 10 with the closest being 9 for the Dummy variable `neighbourhoodMagnoli`., suggesting no significant multicollinearity issues in the model.

Addressing Violations: To address the assumption violations, we propose a transformation of the response variable price, such as a log or square root transformation. Transformations can help stabilize variance (addressing heteroscedasticity) and improve the normality of residuals.

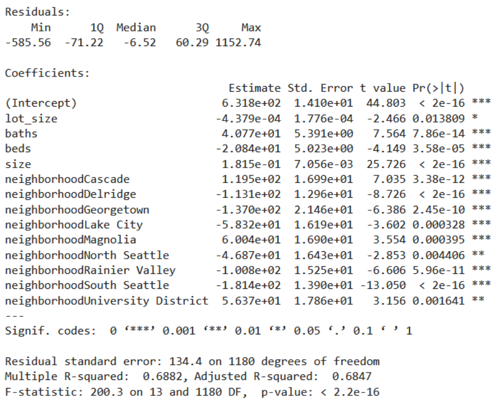
2.5 Transforming Response Variable and Outlier Rejection

Next Steps:

- Apply the log and square root transformations to the response variable `price`.
- Reapply the backward selection process to refine the model with the transformed response.
- Reassess model diagnostics (residuals, Q-Q plot, VIF) to ensure the revised model satisfies the assumptions.

Furthermore, we are also removing the data points that result in standard residuals being higher than some threshold, $|e_i^*| > 3$, to account for the outliers present in the data. The threshold is chosen based on the generally accepted threshold among the scientific community. This will help our model to focus on actual trends and relations present in the data.

Square root price model

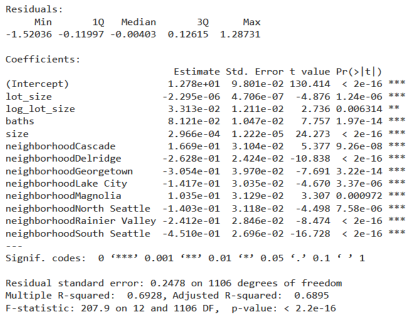


```
Residuals:
    Min      1Q  Median      3Q     Max 
-585.56  -71.22   -6.52   60.29 1152.74 

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     6.318e+02  1.410e+01  44.803  < 2e-16 ***
lot_size                       -4.379e-04  1.776e-04  -2.466 0.013809 *  
baths                           4.077e+01  5.391e+00   7.564 7.86e-14 ***
beds                           -2.084e+01  5.023e+00  -4.149 3.58e-05 ***
size                            1.815e-01  7.056e-03  25.726  < 2e-16 ***
neighborhoodCascade             1.195e+02  1.699e+01   7.035 3.38e-12 ***
neighborhoodDelridge           -1.131e+02  1.296e+01  -8.726  < 2e-16 ***
neighborhoodGeorgetown         -1.370e+02  2.146e+01  -6.386 2.45e-10 ***
neighborhoodLake City          -5.832e+01  1.619e+01  -3.602 0.000328 ***
neighborhoodMagnolia            6.004e+01  1.690e+01   3.554 0.000395 ***
neighborhoodNorth Seattle      -4.687e+01  1.643e+01  -2.853 0.004406 ** 
neighborhoodRainier Valley     -1.008e+02  1.525e+01  -6.606 5.96e-11 ***
neighborhoodSouth Seattle      -1.814e+02  1.390e+01 -13.050  < 2e-16 ***
neighborhoodUniversity District  5.637e+01  1.786e+01   3.156 0.001641 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 134.4 on 1180 degrees of freedom
Multiple R-squared:  0.6882,	Adjusted R-squared:  0.6847 
F-statistic: 200.3 on 13 and 1180 DF,  p-value: < 2.2e-16
```

Log price model



```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.52036 -0.11997 -0.00403  0.12615  1.28731 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 1.278e+01  9.801e-02 130.414  < 2e-16 ***
lot_size                   -2.295e-06  4.706e-07  -4.876 1.24e-06 ***
log_lot_size                3.313e-02  1.211e-02   2.736 0.006314 ** 
baths                       8.121e-02  1.047e-02   7.757 1.97e-14 ***
size                        2.966e-04  1.222e-05  24.273  < 2e-16 ***
neighborhoodCascade         1.669e-01  3.104e-02   5.377 9.26e-08 ***
neighborhoodDelridge       -2.628e-01  2.424e-02 -10.838  < 2e-16 ***
neighborhoodGeorgetown     -3.054e-01  3.970e-02  -7.691 3.22e-14 ***
neighborhoodLake City      -1.417e-01  3.035e-02  -4.670 3.37e-06 ***
neighborhoodMagnolia        1.035e-01  3.129e-02   3.307 0.000972 ***
neighborhoodNorth Seattle  -1.403e-01  3.118e-02  -4.498 7.58e-06 ***
neighborhoodRainier Valley -2.412e-01  2.846e-02  -8.474  < 2e-16 ***
neighborhoodSouth Seattle  -4.510e-01  2.696e-02 -16.728  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2478 on 1106 degrees of freedom
Multiple R-squared:  0.6928,	Adjusted R-squared:  0.6895 
F-statistic: 207.9 on 12 and 1106 DF,  p-value: < 2.2e-16
```

Figure 8: Summary of Square root and Log price model post backward selection

Analysis of Transformations: The model with `sqrt_price` selected the `lot_size` instead of `log_lot_size` from backward selection but the model with `log_price` selected `lot_size` as well and it didn't select beds. Both plots have similar R^2_{adj} so we are going to check for any linear regression assumption violations again using residual vs. fitted values plot, QQ plot and the VIF of the model.

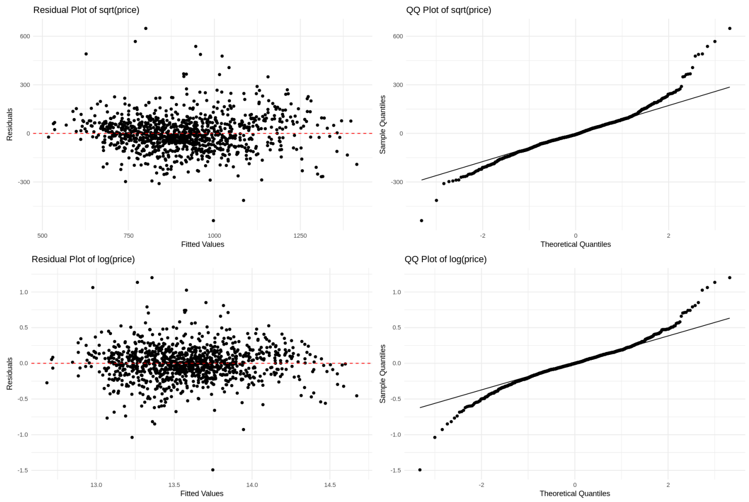


Figure 9: Residual and QQ plot of square root and log price models'

2.6 Final model

Both models (response `sqrt_price` and `log_price`) successfully address the initial assumption violations and exhibit improvements.

- Both models show an increased R^2_{adj} , indicating better explanatory power and improved fit.
- Both models demonstrate minimal heteroscedasticity, with relatively constant variance across fitted values.
- Residuals for both models follow a normal distribution for most values, with slight deviations at the tails. The model with `sqrt_price` has slightly less deviation from the normal model, making it a better candidate.
- No Variation Inflation Factor (VIF) exceeds threshold of 10, confirming that multicollinearity is not a concern in either model.

Conclusion of Model Selection: Given the model with `sqrt_price`'s follows the normal distribution more closely than the model with `log_price`, we will select the model with `sqrt_price` as our final model.This model is a good candidate and meets all assumptions of linear regression model.

- Residuals are approximately normal.
- Variance is relatively constant.
- No significant multicollinearity is present.
- This model is robust and provides reliable insights into the relationships between covariates and house prices.

2.7 Findings

Base-line Neighborhoods
Any inference we make for the baseline neighborhood *Ballard* likely applies to *Green Lake*, *Downtown*, *Central Area*, *Queen Anne*, *Bryant*, and *West Seattle* as well since these are neighborhoods whose effects were statistically insignificant in our model so they likely exhibit similar characteristics as the baseline neighborhood *Ballard*. Moving forward, these neighborhoods will be grouped together and referred to collectively as the baseline neighborhoods. Any inferences made about Ballard can be reasonably applied to this group, as the model suggests they share comparable influences on house prices.

Beds: 1-unit increase in beds reduces `sqrt_price` by −20.8.

Baths: 1-unit increase in baths increases `sqrt_price` by 40.8.

Size: 1-unit increase in size reduces `sqrt_price` by 0.18.

Lot Size: 1-unit increase in lot_size reduces `sqrt_price` by −0.00044.

Neighborhood	Effects on sqrt_price relative to baseline neighborhood
Cascade	Strong positive effect of +119.5
Delridge	Strong negative effect of −113.1
Georgetown	Strong negative effect of −137.0
Lake City	Negative effect of −58.3
Magnolia	Positive effect of +60.0
North Seattle	Negative effect of −46.9
Ranier Valley	Strong negative effect of −100.8
South Seattle	Strong negative effect of −181.4
University District	Positive effect of +56.4

Expensive Neighborhoods: University District, Cascade, and Magnolia are significantly more expensive than baseline neighborhoods, likely due to prestigious factors such as high-quality amenities, luxury properties, and low crime rates. These neighborhoods are ideal for high-budget buyers seeking premium locations.

Cheaper Neighborhoods: Delridge, Georgetown, Rainier Valley, and South Seattle have lower property prices, potentially due to factors like lower-end housing stock, limited infrastructure, isolation from services, or higher crime rates. Further investigation is needed to understand these influences.

Baseline Neighborhoods: Baseline neighborhoods (e.g. Ballard, Green Lake, Downtown) represent more affordable neighborhoods that appeal to middle-class buyers.

Bathrooms: Bathrooms have the strongest positive impact on price, with more bathrooms significantly increasing home value. Buyers in Seattle prioritize this feature when evaluating properties.

Size: Larger homes (size) significantly increase prices, reflecting a consistent preference for spacious properties across all neighborhoods.

Lot Size: Lot size has a slight negative effect on price, suggesting that buyers in Seattle prioritize house size and amenities over lot size, likely due to urban preferences.

Bedrooms: More bedrooms slightly decrease house prices, contrary to expectations. This could indicate the influence of confounding variables or an interaction with other features, suggesting a preference for smaller, functional, and well-designed homes. Further exploration is warranted.

3. Conclusion

This study aimed to identify the physical and geographical factors influencing house prices in Seattle. Using a linear regression model refined through backward selection, we ensured compliance with regression assumptions. The final model, with covariates beds, baths, size, lot_size, and neighborhoods, achieved an adjusted R^2 of 0.685, demonstrating robust explanatory power. Our analysis suggests that bathrooms and house size significantly increase property prices, highlighting buyer preferences for functional and spacious homes. Lot size seems to have minimal negative effect, and additional bedrooms seem to slightly decrease prices, a counter-intuitive result warranting further investigation into potential confounding factors. Neighborhood desirability is pivotal. Areas like University District, Cascade, and Magnolia command higher prices due to potential factors such as prestige, safety, and amenities. Conversely, neighborhoods like Delridge, Georgetown, and South Seattle are associated with lower prices, potentially due to limited infrastructure or higher crime rates.

This analysis offers actionable insights for buyers, investors, and policymakers. Buyers can identify neighborhoods and features that align with their preferences, while policymakers can address disparities across neighborhoods. Future research could examine external factors, such as market trends and socio-economic influences, to deepen understanding and guide decision-making in Seattle’s housing market.

4. References

1. United States Postal Service. (2024). *United States Zip Code*. UnitedStatesZipCodes.org . https://www.unitedstateszipcodes.org/98104/

2. Frost, J. (2023, May 18). *5 ways to find outliers in your data*. Statistics By Jim. https://statisticsbyjim.com/basics/outliers/

5. Appendix

A. Data Preprocessing

1. **Transformations**: Converted `lot size` from acres to square feet. Missing values in `lot_size_units` were omitted.
2. **Categorical Variables**: Mapped `zip_code` to neighbourhoods and grouped insignificant neighbourhoods into "Other" during backward selection.
3. **Removed Variables**: `size_units` and `lot_size_units` were excluded as they added no value to the analysis.

B. EDA Highlights

1. **Outlier Detection**: Identified extreme price outliers using the IQR method and standardized residuals.
2. **Visualizations**: Explored relationships between price and predictors (e.g., beds, baths, size, lot size) using scatter plots and boxplots. Neighbourhoods like Cascade and Magnolia showed higher prices, while South Seattle and Rainier Valley had lower prices.
3. **Correlation**: Identified multicollinearity concerns between `beds`, `baths`, and `size`.

C. Model Selection

1. **Initial Model**: Included `baths`, `beds`, `size`, `lot_size`, `log_lot_size`, and `neighbourhood`.
2. **Backward Selection**: Iteratively removed insignificant predictors, grouping neighbourhoods with high p-values into "Other."
3. **Transformations**: Tested `sqrt_price` and `log_price` as response variables. Final model based on `sqrt_price`.

D. Final Model

1. **Predictors**: `lot_size`, `baths`, `beds`, `size`, and significant neighbourhoods. Baseline neighborhoods grouped as "Ballard," "Green Lake," "Downtown," "Central Area," "Queen Anne," "Bryant," and "West Seattle."

E. Diagnostics

1. **Residual Plot**: Minimal heteroscedasticity observed.
2. **Q-Q Plot**: `sqrt_price` transformation showed residuals closer to normality than other transformations.

F. Future Work - Include interaction terms (e.g., `baths * lot_size`) to improve accuracy, and explore external factors like crime rates and proximity to amenities.