# *FINAL ASSIGNMENT*

*Exploring what influences the value of houses.*

*Jessica Krook*

*393985: Exploratory Data Analysis and Visualization COM SCI X 450.2 (Fall 2023)*

Introduction

The housing market is always evolving and is continually becoming more complex. As a small, private housing organization, understanding the market and future trends is important for sustained growth and strategic edge.

In this report, I will analyze data provided by the organization and provide insights found in the data. I will first take you through a summary of the data, then I will explain my cleaning process. After that, I will highlight interesting findings about the data using both single and two-variable visualizations. After that I will go through some modeling I did on the data using single regression and k-means clustering. Finally, I will use a different cleaning method than I originally used to show that how you clean data affects the insights you gain from analysis.

## 1. Data summary, oddities, and outliers

In the housing dataset:

There are NAs present in the numerical variables of the dataset. NA's affect analyses such as a correlation matrix, so they should not be in the dataset. The two variables that contain them are:

- "sqft"
- "lotsize"

There are extreme outliers present in the data. They are data points far away from the next nearest point compared to the rest of the data. These are the result of data capturing errors and need to be addressed. The extreme outliers are found in the numerical variables:

- Beds: the max is 999, where the 3rd Qu is 4
- Baths: the max is 25, and then the 3rd Qu is 2.5
- Year: 2111 has not happened yet, and 1495 is too long ago for a regular house to be built and still standing
- Soldprice: min is 664, and the 1st Qu is 974,500. 664 is also too low of a value for a house to be sold in 2019.

I plan to delete the NAs and extreme outliers as they are a small part of the dataset and are data-capturing mistakes. Deleting the NAs will not change the story of the data, and the story will be error-free and free from influencing numbers.

In the school dataset

There are outliers in the "size" variable. I will investigate and possibly remove them during the data-cleaning process.

## 2. Data cleaning

In the housing dataset:

I changed the character variables neighborhood, type, levels, cooling, heating, fireplace, elementary, middle, and high to factor variables.

The “type” variable had inconsistencies. It had 4 values that were duplicates with incorrect spelling. I updated “town house” to townhouse and “condominium” to “condo”. The resulted in the “types” values: “townhouse”, “condo”, “single-family home” and “multi-family home”, making analysis easier.

After changing the characters to factors and re-running the summary I found the variable levels has “?” values and, cooling, heating, and fireplace have empty values, so I removed the lines containing “?” and the lines with empty values.

For the numerical variables, I deleted the lines with NA in “sqft” and “lotsize”.

I also deleted all the lines containing extreme outliers for “beds,” “baths”, “year” and “soldprice”.

In the school dataset:

I changed the character variable “school” to a factor.

I searched for outliers in the “size” variable and decided not to remove any as the outliers were not extreme outliers and I wanted to ensure the merge was possible to match all housing lines with the corresponding high school.

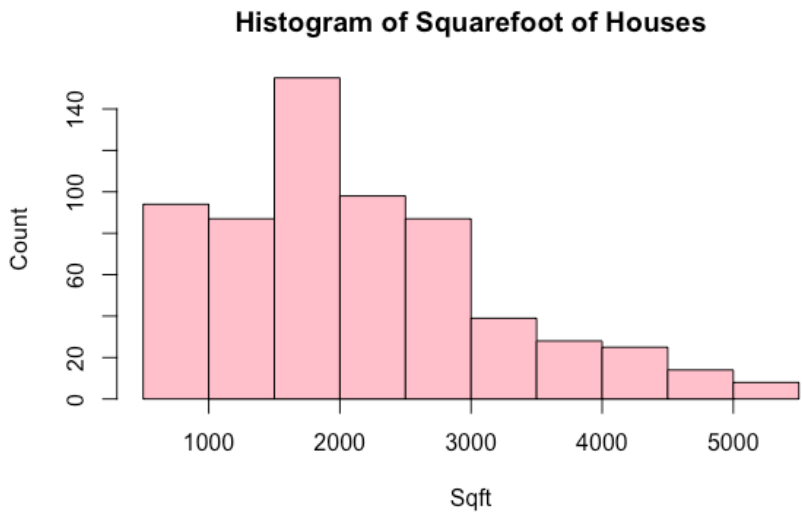
Merging the housing and schools datasets:

I performed a left merge on the housing, high school, and school, school variables.

## 3. One-variable visuals

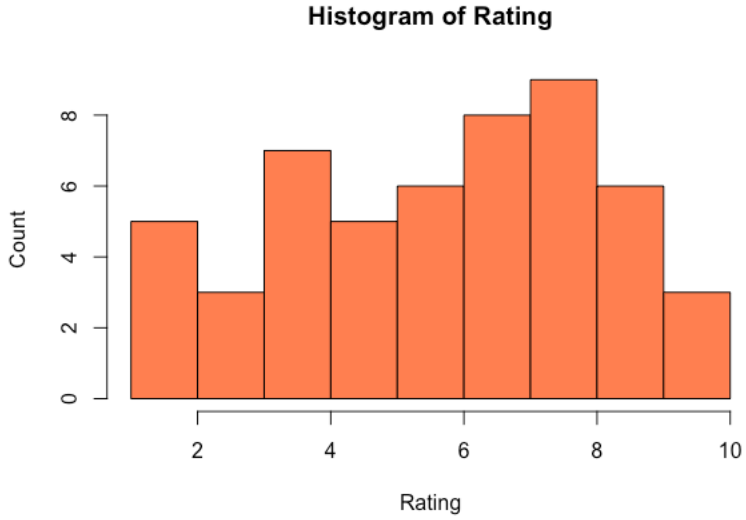
Histograms

*Visual 1: Histogram of Square Foot of Houses*




Most houses are between 1500 and 3000 sqft. With the most falling between 1500 and 2500 sqft

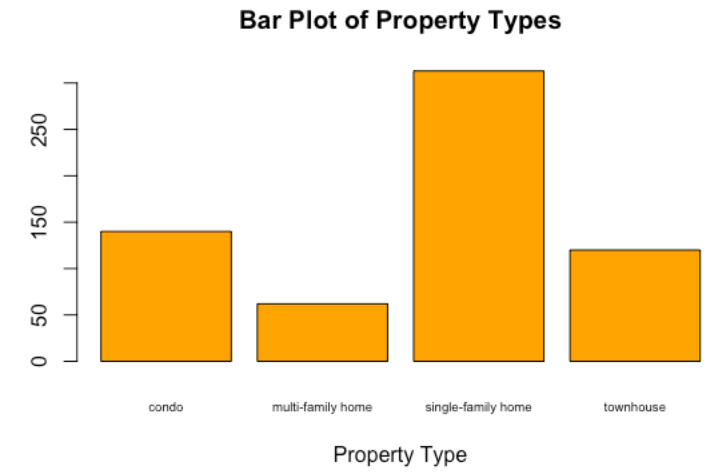
*Visual 2: Histogram of distribution of School Ratings*




Most schools have a rating between 6 and 87 There is a peak at 1 and a peak at 4. These could be further investigated as to why there are large groups of low-rated schools.

## Bar plot

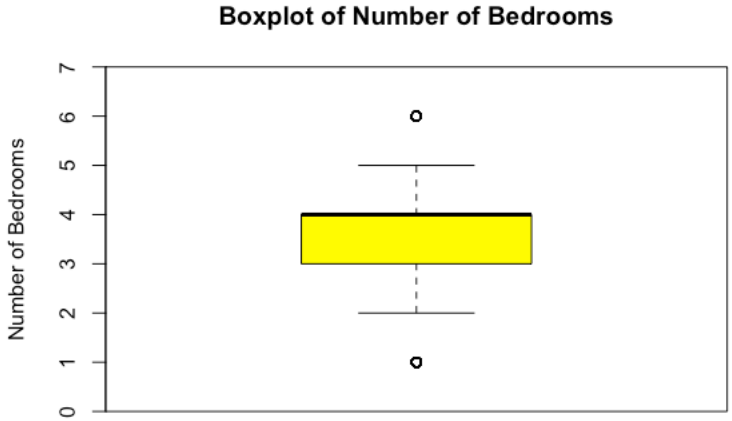
*Visual 3: Bar Plot of the Count of Property Types*




Most property types are single-family homes, and the least are multi-family homes. We can further look at the characteristics that make up each home type in further analysis for example we can look at how many bedrooms are in each type of home.

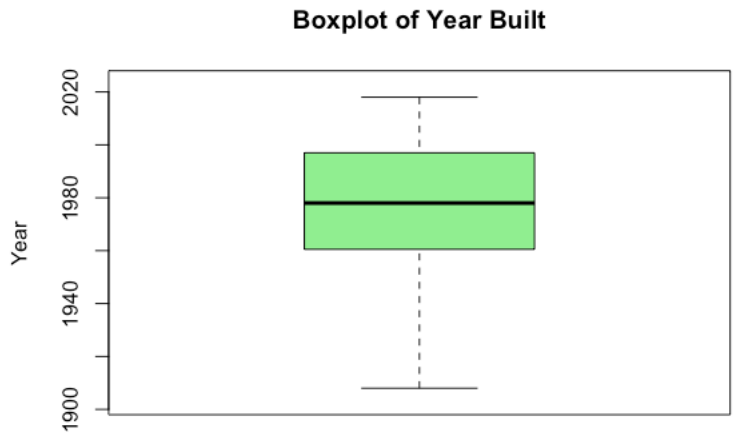
## Box plots

*Visual 4: Box Plot showing the number of Bedrooms in Houses*



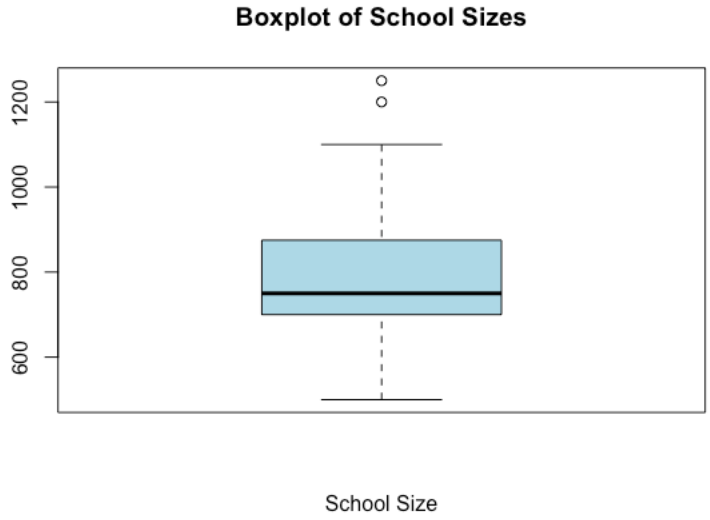

There are two outliers in the bedrooms variable. But since there are multiple points at these outliers. We will accept they are part of the dataset and will not skew the data. The median is skewed at the very top of the meaning there are more houses with bedrooms of 4 and more than of less than 4 in the dataset.

*Visual 5: Box Plot showing the Years houses were built*




There are more data points for houses built in later years. The houses are concentrated to be built between 1970 and 1990. The oldest houses were built around 1900 and the newest were built in 2018.

*Visual 6: Box Plot showing the Size of Schools*

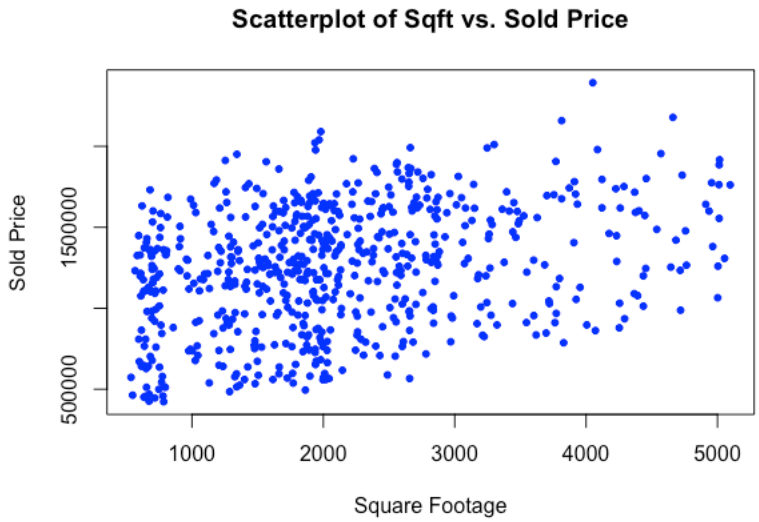



There is a large range of school sizes. The smallest school is around 400 people while the largest school is around 1200. Even though the largest school is around 1200, most school sizes are between 1000 pupils. It would be interesting to look at the relationship between school size and school rating and whether bigger schools have higher ratings than smaller schools.

## 4. Two-variable visuals

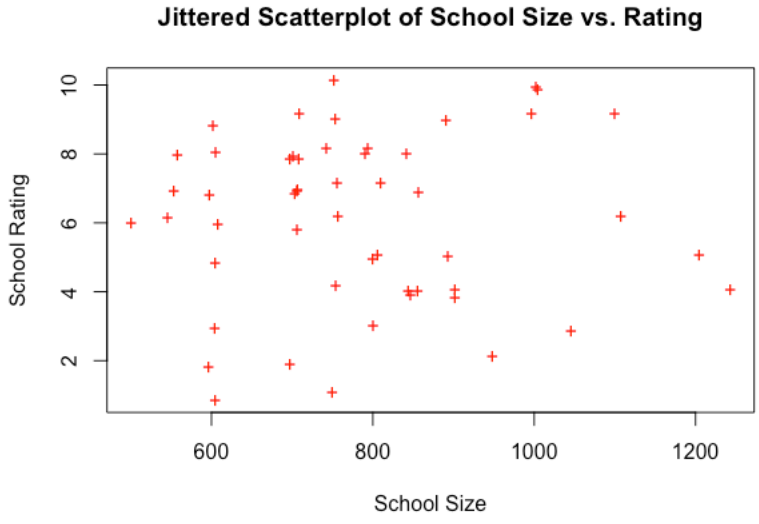
Scatter plot

*Visual 7: Scatterplot showing the Square foot of a house vs its Selling Price*




Most houses are under 3000 feet and under $1500000. There is a relationship between the size of a house and its sold price – the bigger the house, the higher the selling price. It is a weak relationship, but it is visible in the graph. Since the relationship is weak, other factors may influence the selling price of a house.

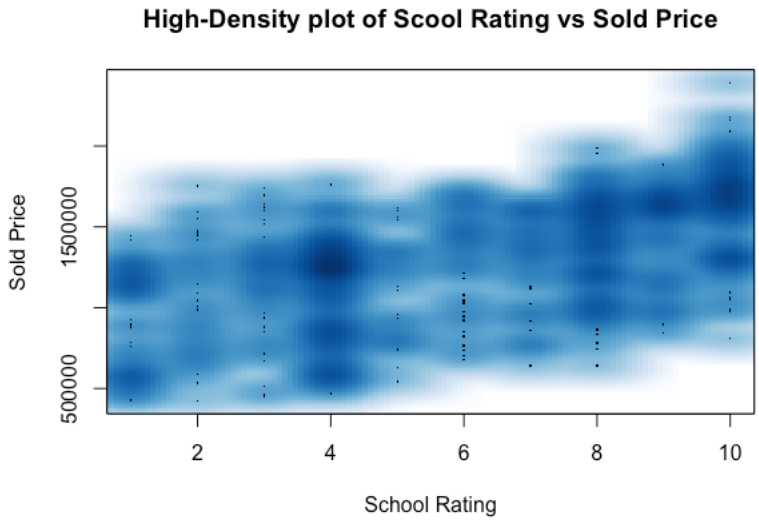
*Visual 8: Scatterplot showing the Size of a school vs its Rating*




The scatterplot shows that schools with pupils between 700 and 1100 have the highest rating. However, the highest concentration of 8+ ratings falls between 700 and 850. We can also see that schools between 850 and 1000 have a dip in school ratings. There is no true relationship between school size and its rating.

High-density plot

*Visual 9: High-Density Plot showing a School's Rating vs the Selling Price of a House*




We can see that as the school rating increases, so does the sold price of a house. This shows buyers want to live in districts with schools with high ratings. There is a high concentration of sold price over 1500000 with schools rated 10.

This discovery will serve as the foundation of our analysis.

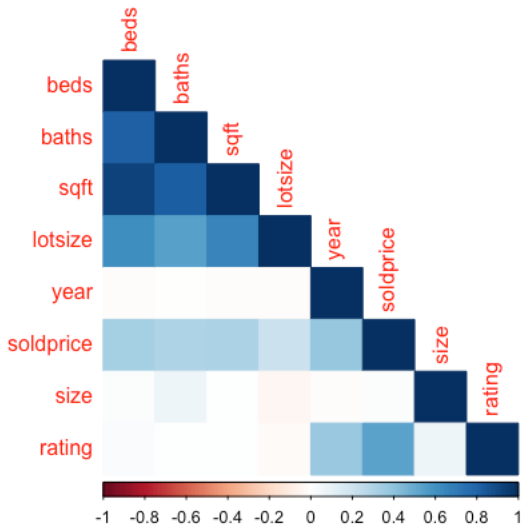
There is also a high concentration of schools rated 4 with a selling price just below 1500000. A possible explanation is that these are smaller houses with one bedroom where people are not interested in schools.

This can be something explored in the future.

## 5. Analysis

Exploring relationships between the data

*Visual 10: Correlation Matrix to help visualize relationships in the data*



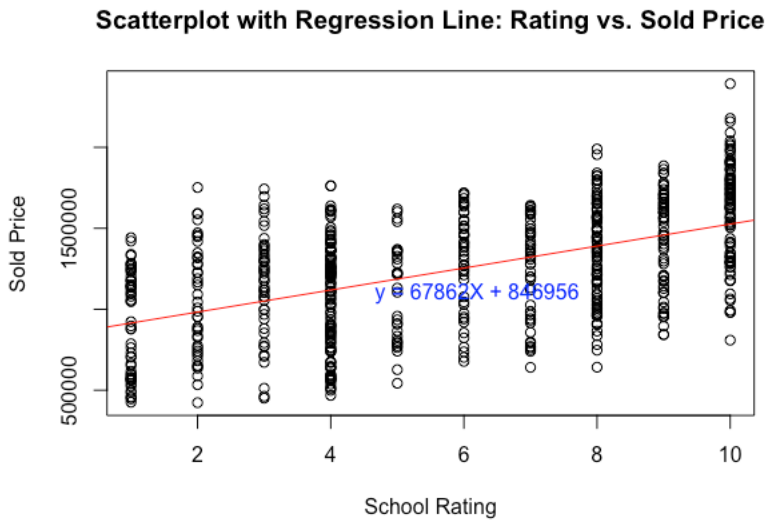

In order to decide what regression, I would investigate, I created a correlation matrix between all variables in the “merged_data” dataset. From this, I found that there is a high correlation between:

- beds and baths,
- sqft and bed,
- sqft and beds,
- lotsize and beds,
- lotsize and sqft
- soldprice and rating

The correlation between “soldprice” and “rating” is the most interesting to me.

## Regression result

*Visual 10: Regression Visualisation of School Rating vs Selling Price*




We can see from the regression result that the rating is highly significant to the sold price

Summary of Sold Price vs. School Rating:

Model Fit:
The model explores the relationship between the sold price of properties and their school ratings.

Coefficients:

- Intercept: The sold price is estimated to be $846,956 when the school rating is zero.
- Rating: for each unit school rating increase, the sold price is expected to increase by $67,862.

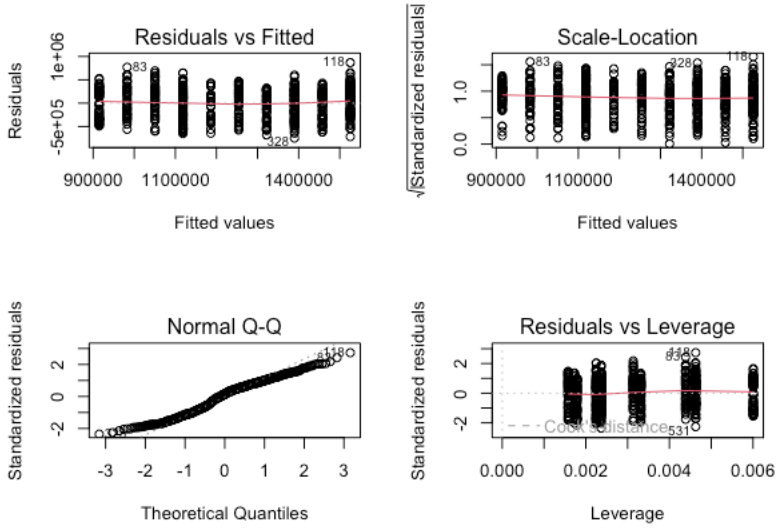
Model Performance:

- The model's residual standard error is approximately $318,000.
- The model explains 28.2% of the variability in sales prices (Multiple R-squared).

Overall Fit:
The F-statistic is 248.6 with a very low p-value (< 2.2e-16). This means it is a highly significant model.

Statistical Significance:
The intercept and coefficients are highly statistically significant. This suggests a strong relationship between school rating and sold price.

*Visual 11: Diagnostic Plots of the Regression model*




The diagnostic plots look normal.

## Clustering result

*Visual 12: K-Means Clustering visual of Square Foot vs Selling Price*


While the clustering is not perfect, we can see three clusters where square footage and sold price are similar. From this, we can make predictions of the price of a house based on its square footage.

## 6. Sensitivity Analysis

### Recleaning the data

Empty values

I used my judgment to replace the empty factors values:

> For the “levels” category:
> The category with the most values is “1 level”, so we will update the “?” to “1”.
>
> For the “cooling”, “heating”, and “fireplaces” categories:
> The most frequent value is “No”, so we will update the empty values to “No”.

For the numerical variables sqft and lotsize, I used single imputation to fill in the missing values because both variables had extreme outliers. by using the median, the imputation will not be influenced by the extreme outliers.

Outliers

I used my judgment and applied different methods for each variable’s outliers.

Baths:

The upper bound for “baths” is 4.75, therefore I rounded up to 5 and kept all outliers that equal 5. I also chose to keep the outliers because the 5 bathrooms correspond with 6 bedrooms, and it makes sense to have 5 bathrooms in a 6-bedroom house. I was left with one other outlier “25”. I updated it to the mean of the “baths” variable, 2.001 so I rounded to 2 and replaced the 25 with 2.

Beds:

5 and 6 bedrooms are possible, so I only updated the extreme outlier of 999. I replaced the value with the median of the dataset, which was 4

Year:

Year was a bit trickier as the range is so big it is difficult to just use the mean or median from the whole dataset. I decided to subset the data by each neighborhood The outliers fell in as most neighborhoods develop around the same time. I then used the median for each neighborhood. This is by no means the best method, but it was one that I could make the best guess.
For the outlier 2111, I used the median of the orange neighborhood. This was 1980.
For the outlier 1498 - I used the median of the silver neighborhood. This was 1959.

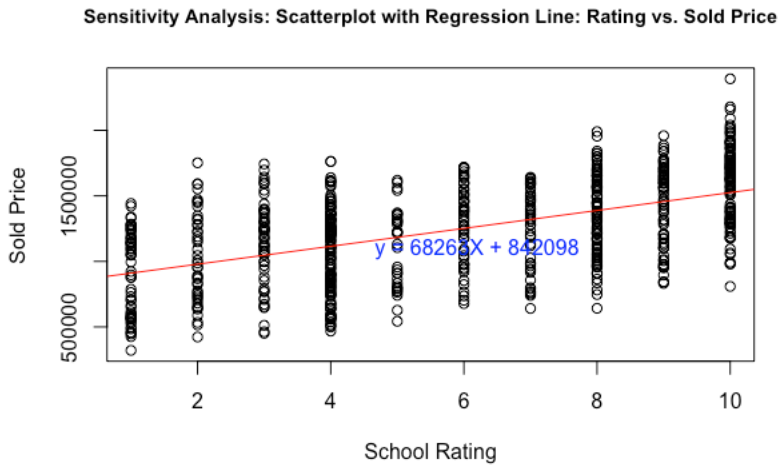
Soldprice:

I chose to change the soldprice outlier “664” as it was an extreme outlier to the median soldprice of “1244858 ” so the new value was not influenced by the outlier.

**Comparing new results to the original run of the analysis**

Regression Model

*Visual 13: Re-run of Regression Visualisation of School rating vs Selling Price*




The regression analysis returned similar results after re-running the model with the new dataset. This makes sense as very little was changed in the “soldprice” variable. The second data had one extra point added at the median point of the data. This would not influence the data enough to change the regression model.

Summary of Sold Price vs. School Rating for the second data set:

Coefficients:

- Intercept: The sold price is estimated to be $842,098 when the school rating is zero.
- Rating: for each unit school rating increase, the sold price is expected to increase by $68,263.

While the sold price is estimated to be lower than the first model, the expected increase in sold price is larger than the first model.

Statistical Significance:
The intercept and coefficients are highly statistically significant. This suggests a strong relationship between school rating and sold price.

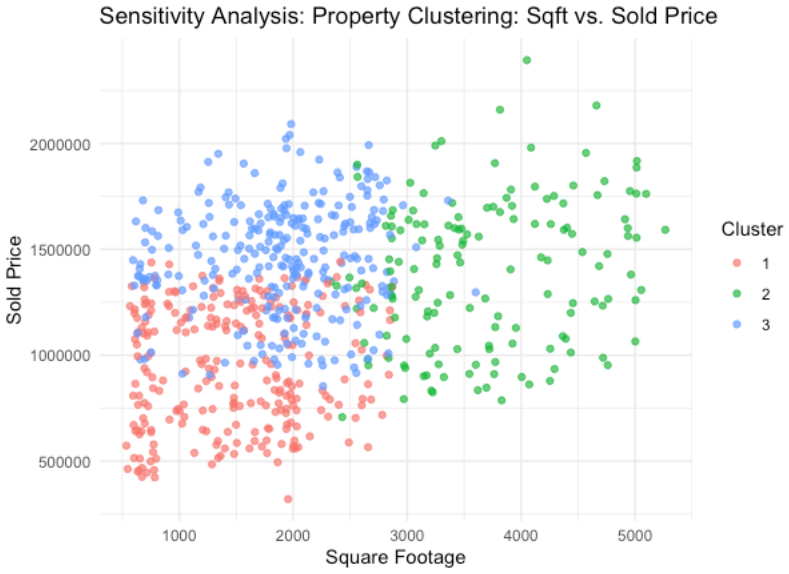
Model Performance:

- The model's residual standard error is approximately $315,700.
- The model explains 28.55% of the variability in sales prices (Multiple R-squared).

Overall Fit:
The F-statistic is 272.1 with a very low p-value (< 2.2e-16). This means it is a highly significant model.

## Clustering

*Visual 14: Re-run of K-Means Clustering visual of Square Foot vs Selling Price*




The clusters look largely the same between the two graphs we can see three clusters with minimal overlapping where square foot and sold prices are similar. From this, we can make predictions of the price of a house based on its square footage that will be similar to the first k-means clustering.

The model did not change too much as there were very few changes in the two variables used in the analysis. There was one value change in “sold price” that was an outlier. Sqft’s missing values were imputed with the median, meaning 6 values were added in the middle of the model. These would have little effect on a K-means clustering model.

## Conclusion

This analysis has shown that school rating has a significant influence on the sale price of houses. House values are higher when schools have higher ratings. In the future, a multiple regression could be run on this data to see if other variables have a significant impact on house values and if certain variables in combination result in higher house values.