Report on Ames Housing Data Set

Esther Waweru - 171444

2024-10-27

Introduction

The Ames Housing dataset provides a comprehensive look at housing data in Ames, Iowa. It serves as a rich alternative to the well-known Boston Housing dataset, containing detailed information on the properties sold in the area. This analysis aims to explore the relationships between various property features and house sale prices using advanced multiple regression techniques.

The dataset consists of more than 80 variables across nearly 3,000 observations. Approximately half of these observations belong to a test set where the sale price is not disclosed. Each observation provides details on the features listed in Table 1.

Lot…Location	Building	Meta.Data
Neighborhood	Basements & Garages	Functionality
Streets & Alley Quality	Square Footage	Miscellaneous Features & Prices
Lot Area & Shape	Quality & Condition	Sale Type & Condition
Zoning	Roof & Siding Materials	Sale Date (month and year)

Objectives

The primary objective of this analysis is to build a predictive model for house prices using multiple regression analysis. The project involves:

1. Preprocessing the data to handle missing values and categorical variables.
2. Conducting exploratory data analysis to identify significant predictors.
3. Building and evaluating a regression model to predict house sale prices.
4. Improving the model and providing insights for stakeholders.

Metric of Success

Performance of the model is based on R2, adjusted R2.

Data Cleaning and Preprocessing

Handling Missing data

Missing data can occur due to various reasons, such as data entry errors, lack of information, or irrelevant attributes for specific instances. To ensure the dataset is suitable for analysis, we first investigated the extent of missing data in each variable. Below are the steps taken to identify, handle, and impute or exclude missing values, along with a discussion of the potential impacts of these choices on the analysis.

1. **Identification of Missing Values**

The initial inspection of the dataset revealed several columns with missing values. Missing values were identified through the function is.na() to count the NA values in each column.

Variables	No..of.missing
LotFrontage	259
Alley	1369
BsmtQual	37
BsmtCond	37
BsmtExposure	38
BsmtFinType1	37
BsmtFinType2	38
Electrical	1
FireplaceQu	690
GarageType	81
GarageYrBlt	81
GarageFinish	81
GarageQual	81
GarageCond	81
PoolQC	1453
Fence	1179
MiscFeature	1406

2. **Handling Missing Data by Column**

- Basement Columns (BsmtExposure and BsmtFinType2)

The basement-related columns had missing values that corresponded to buildings without basements. However, some variables, such as BsmtExposure, had inconsistencies that suggested data entry errors rather than the absence of a basement. We assumed that missing values in BsmtExposure represented "No Exposure" rather than the absence of a basement. For BsmtFinType2, the missing values were assumed to indicate that a building only had one basement area, so we replaced missing values with "None."

- Fireplace (FireplaceQu) and Number of Fireplaces (Fireplaces)

Missing values in FireplaceQu indicated the absence of a fireplace, based on the data description. Therefore, we replaced missing values with "No Fireplace" and ensured that the Fireplaces variable had a value of 0 where FireplaceQu was "No Fireplace."

- Garage-Related Columns (GarageType, GarageFinish, GarageQual, GarageCond, and GarageYrBlt)

Missing values for most garage-related columns signified a lack of a garage on the property. We replaced missing values with "No Garage" in relevant columns, except for GarageYrBlt, where we kept the year column intact.

- Lot Frontage (LotFrontage)

Missing values in LotFrontage could potentially affect analyses related to property size or neighborhood characteristics. To retain the numeric consistency of this variable, we replaced missing values with the mean LotFrontage, minimizing potential skew from large or small lot sizes.

- Masonry Veneer (MasVnrType and MasVnrArea)

For MasVnrType, missing values were replaced with "None," assuming no masonry veneer was present. For MasVnrArea, missing values were imputed with the mean value.

- Other Categorical Features with Missing Values

Alley: Missing values were assumed to indicate that the property had no alley access, so we replaced missing values with "No Alley."

PoolQC: Missing values were assumed to represent properties without a pool, so we filled in missing values with "No Pool."

Fence: Missing values were assumed to mean no fence on the property, so we filled them with "No Fence."

MiscFeature: Missing values here were filled with "None," assuming no miscellaneous features.

- Electrical

Since it had only one missing values, we dropped it.

Checking Data Type

Data types play a crucial role in data analysis, as they determine how operations are performed on the data. In the dataset, I initially identified several variables with the character data type that were more appropriately treated as factors. Additionally, some numerical variables represented categorical data, which warranted conversion to factor type for proper analysis.

To transform to factors, we used the following R code:

```r
# Extract columns with character datatype
col1 <- train[, sapply(train, is.character)]

col <- colnames(col1)

factor_columns <- c('MSSubClass','OverallQual','OverallCond',col)

# Convert specified columns from factor
train[factor_columns] <- lapply(train[factor_columns], as.factor)

str(train)
```

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical step in the data analysis process, allowing us to understand the dataset, identify patterns, and reveal insights. In this section, we will create visualizations for categorical and numerical variables, as well as examine the relationship between the target variable, `SalePrice`, and other key variables.

- Categorical Variables

Bar plots are useful for visualizing the frequency or proportion of categories within categorical variables. Here is how we can build bar plots for selected categorical variables in our dataset.

```
# Extract columns with character datatype
cat_c <- names(train)[sapply(train, is.factor)]

# Create barplot for each categorical variable in the dataset
for(col in cat_c) {
 bars <- ggplot(train, aes(x = train[[col]])) +
    geom_bar(fill = "blue") +
    labs(title = paste("Bar Plot of", col), x = col, y = "Frequency") +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

 print(bars)
}
```

Some of the insights from the bar plots are :

1. Most type of dwelling is 1-STORY 1946 & NEWER ALL STYLES, followed by 2-STORY 1946 & NEWER and 1-1/2 STORY FINISHED ALL AGES.
2. Most houses are in the Residential Low Density zone.
3. Majority of the houses have paved street and no alley access.
4. Most houses have a regular shape followed by slightly irregular
5. For LandContour, majority of the houses are near flat;For utilities majority have all public utilities - Electricity, Water, Gas and Sewer.
6. Majority of the houses lot configuration is inside and have gentle slope.
7. Majority of the houses are located in North Ames, are in a normal proximity to various conditions.
8. Most of the buildings are single family detached, one story followed by two story.
9. Rates the overall material and finish of the house are mainly in the range average(5),above average(6) and good(7) and overall condition of the house is mainly average(5).
10. Type of roof is mainly Gable with most of the roof material as Standard (Composite) Shingle
11. Most exterior covering of houses is Vinyl Siding and have no Masonry veneer

- Numerical Variables

Density plots are a great way to visualize the distribution of continuous numerical variables. They provide insights into the data's underlying distribution, which can be particularly useful for identifying skewness and the presence of outliers.

```
# Extract columns with character datatype
numer_c <- names(train)[sapply(train, is.numeric)]

# Create barplot for each categorical variable in the dataset
for(num in numer_c) {
 density <- ggplot(train, aes(x = train[[num]])) +
    geom_density(fill = "lightblue") +
    labs(title = paste("Distribution Plot of", num), x = num, y = "Density") +
    theme_minimal()
 print(density)
}
```

From these, we can see that most of the variables are skewed to the right.

- Comparing Sale Price with Other Variables

1. Sale Price vs Year of Sale

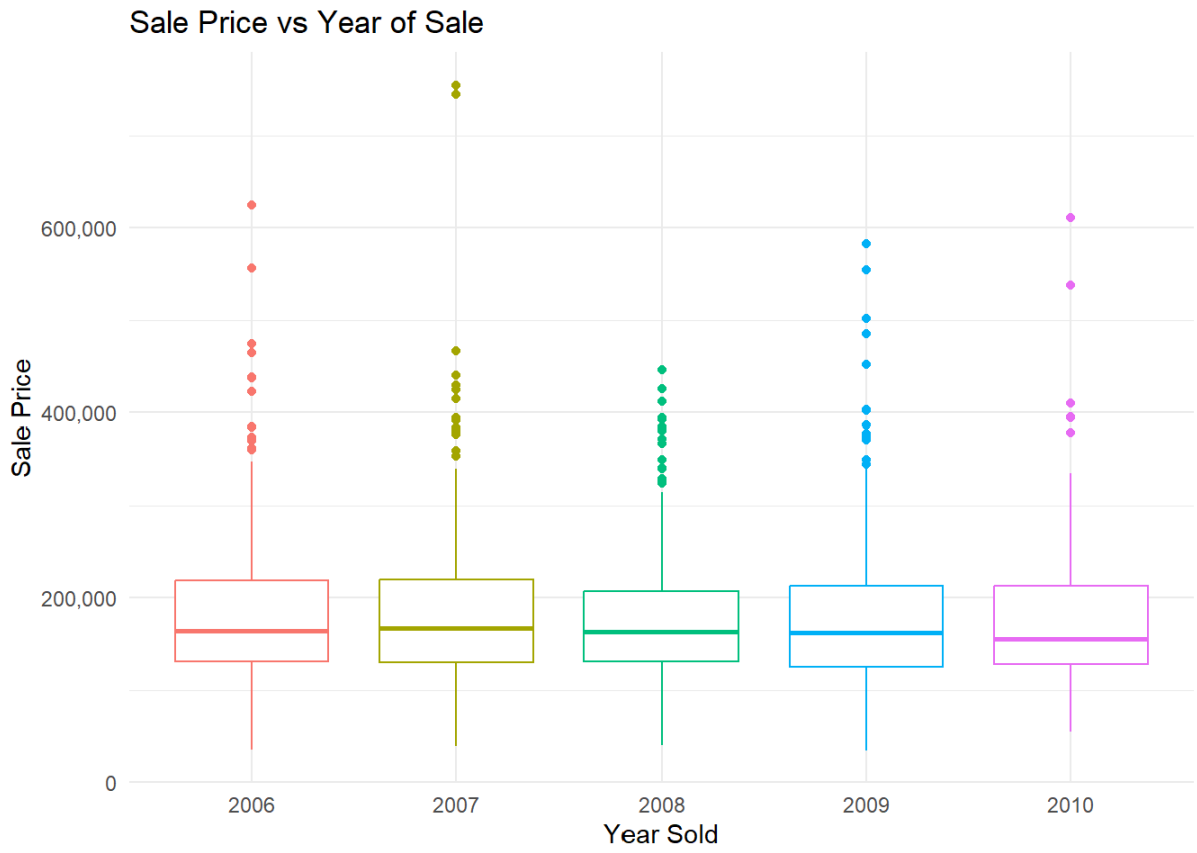

Figure 1: Price vs Year Sold

There is an almost similar pattern across the years when it comes to Sale Prices.

2. Sale Price VS Year Built VS Renovated

From the plot, we can say that newer houses are sold at higher prices and house build after 1940 were mostly renovated later.

3. Sale Price vs Overall Quality

Houses with Very Good, Excellent and Very Excellent Quality are more expensive.

4. Sale Price vs Overall Condition

Houses with Very Good, Excellent and Very Excellent Overall Condition are more expensive.

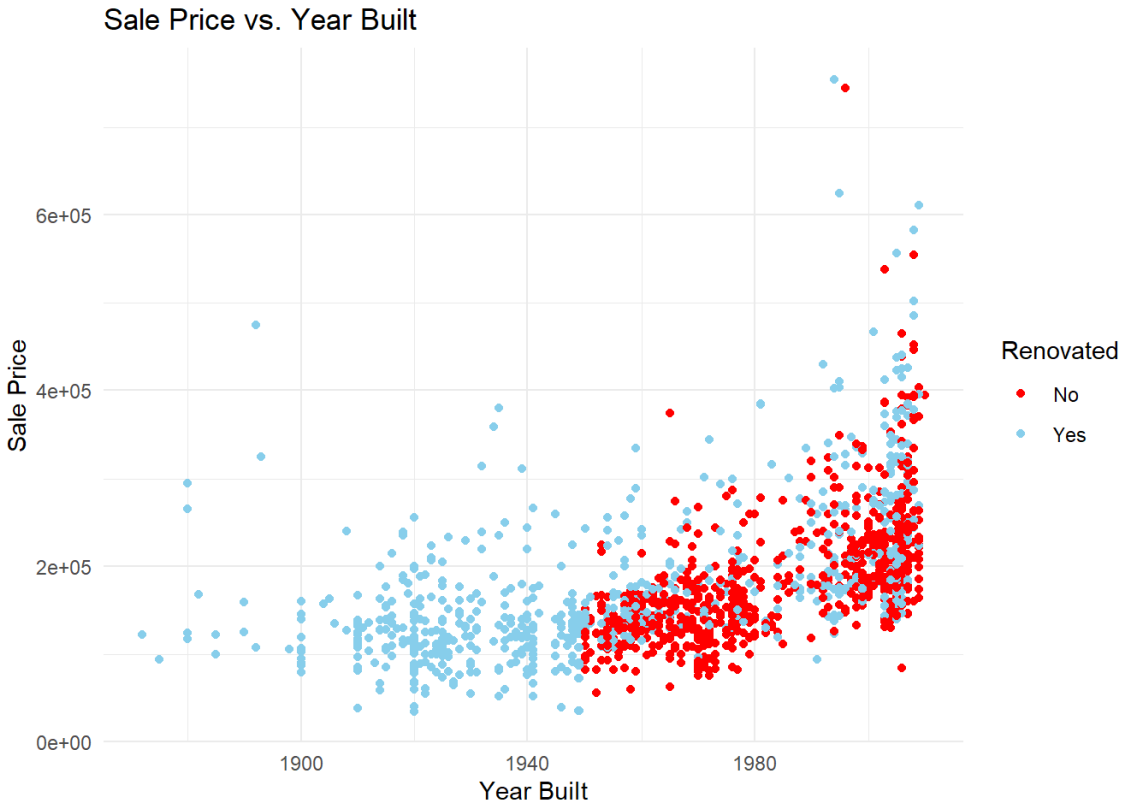

Figure 2: Sale Price VS Year Built VS Renovated

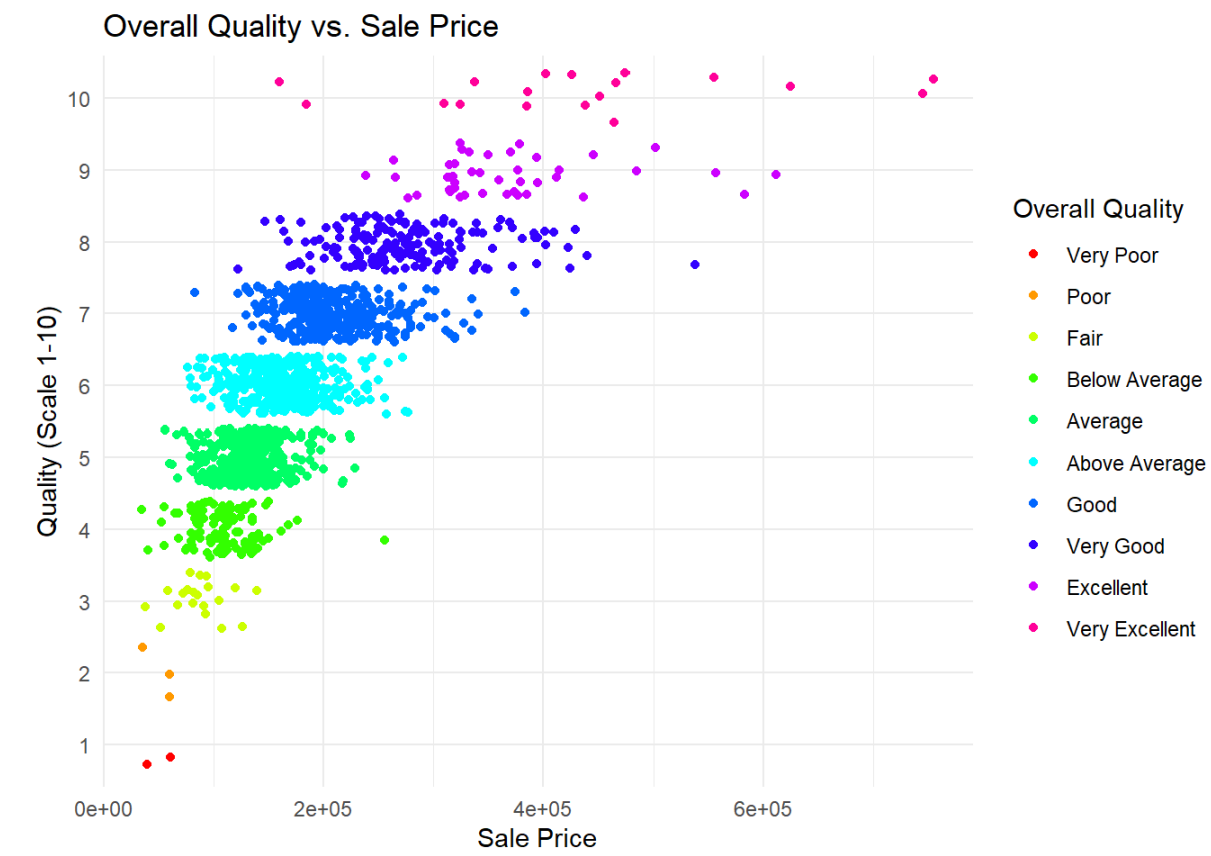

Figure 3: Sale Price vs Overall Quality

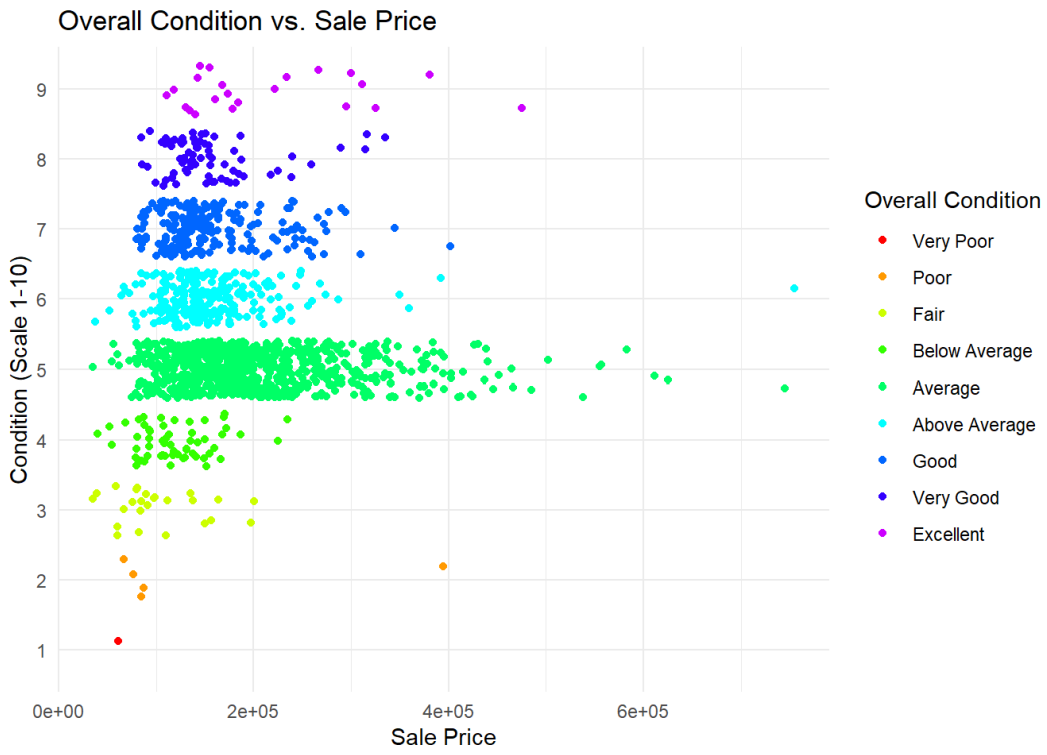

Figure 4: Sale Price vs Overall Condition

5. Sale Price VS Neighbourhood

Northridge Heights and Stony Brook are the most expensive neighbourhoods as per the Sale Prices.

- Correlation

- GrLivArea (Above Ground Living Area): Has the highest positive correlation with sale price, meaning larger living areas are strongly associated with higher prices.
- GarageCars and GarageArea: Both variables related to garage capacity and area show high positive correlations, indicating that homes with larger garages tend to sell for higher prices.
- Attributes like TotalBsmtSF (Total Basement Area), X1stFlrSF (First Floor Square Footage), FullBath (Number of Full Bathrooms), and TotRmsAbvGrd (Total Rooms Above Ground) show moderate positive correlations. This suggests that these features are still important but not as strongly associated with price as the top attributes.
- Features like HalfBath, LotArea, ScreenPorch, and BedroomAbvGr (Bedrooms Above Ground) have weaker correlations with sale price. While they may add some value, their influence on price is smaller compared to the more strongly correlated attributes.
- Variables such as LowQualFinSF (Low Quality Finished Square Footage), EnclosedPorch, and KitchenAbvGr (Kitchens Above Ground) show very low or even slightly negative correlations with sale price, suggesting they either do not significantly influence price or may slightly decrease it in some cases.

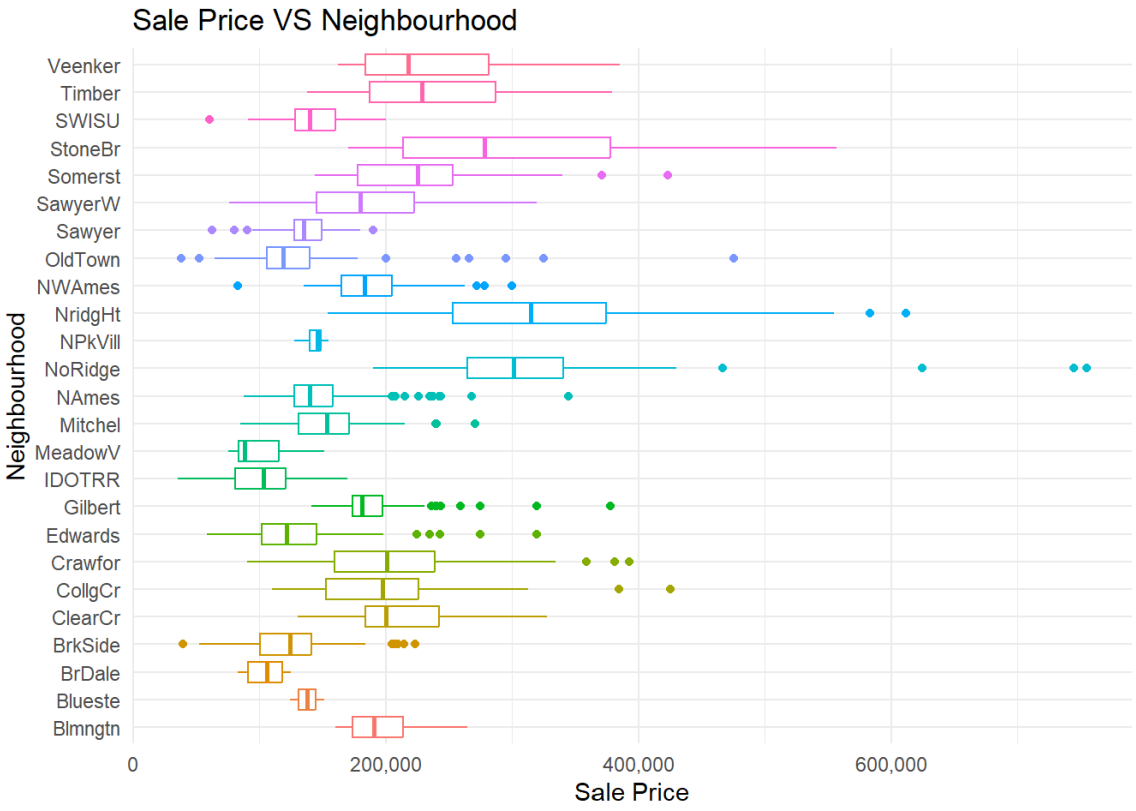

Figure 5: Sale Price VS Neighbourhood

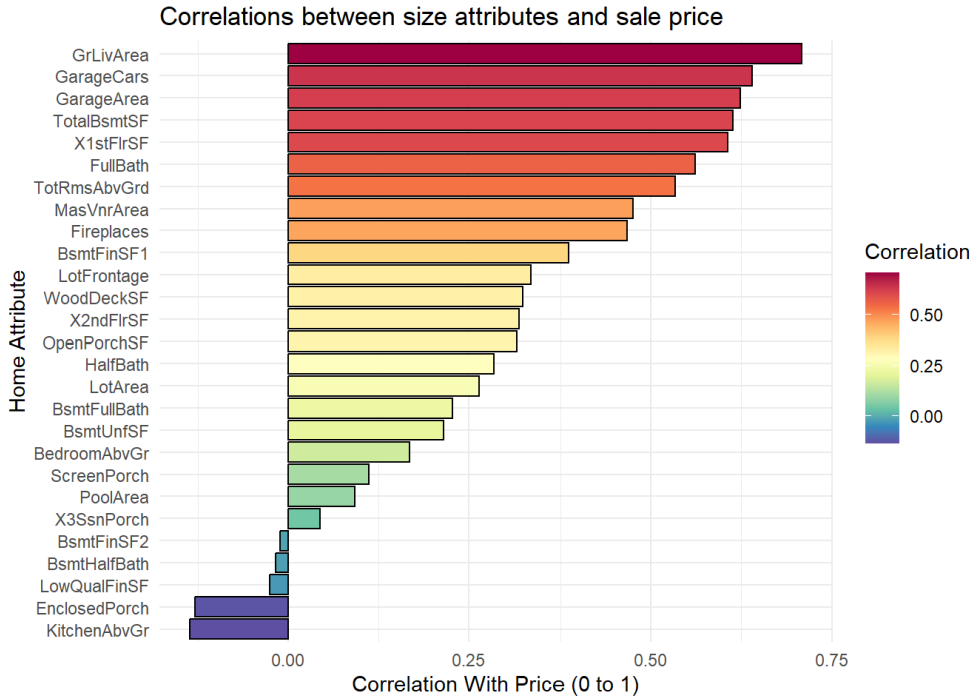

Figure 6: Size Attributes Correlation with Sale Price

Feature Engineering

Feature engineering is an essential step in the data preprocessing pipeline, as it helps create new informative features from existing ones while improving model performance. In our analysis, we combined several related variables into a single feature to capture more meaningful information. For instance, we merged `BsmtFinSF1`, `BsmtFinSF2`, `X1stFlrSF` and `X2ndFlrSF` into a new variable called `TotalFinSF`, which describe the finished square feet area of the house.

Additionally, for categorical variables, we removed those where the mode accounted for 90% or more of total observations. This approach helps eliminate features with low variance that may not contribute significantly to the model's predictive power. By focusing on more informative variables, we aim to enhance the robustness and generalizability of our analysis.

From 80 variables, we reduced them to 56.

Model Building

We first One hot encoded the categorical variables using code:

```
#One hot encode categorical variables
cat_dat <- t_data[,sapply(t_data, is.factor)]
num_dat <- t_data[,sapply(t_data, is.numeric)]

encode <- onehot(cat_dat)
cat_encoded <- as.data.frame(predict(encode, cat_dat))

# Combine encoded categorical data with numeric data
train_data <- cbind(num_dat, cat_encoded)
```

The model fitting process began by initially fitting a linear regression model using all available predictor variables. This comprehensive approach provided a baseline understanding of how the various features interact with the target variable, `SalePrice`. However, to refine the model and identify the most significant predictors, we performed stepwise regression using the Akaike Information Criterion (AIC) as the selection criterion. This method allowed us to systematically add or remove predictors based on their contribution to the model, resulting in a more parsimonious and interpretable model.

After identifying the optimal set of predictors through stepwise regression, we checked for multicollinearity among the variables using Variance Inflation Factor (VIF) scores. Multicollinearity can distort the estimated coefficients and inflate the standard errors, leading to unreliable statistical inferences. We dropped all predictors with a VIF greater than 5, as these indicated significant multicollinearity, and refitted the model using the remaining predictors identified from the stepwise regression.

To further enhance model performance, we addressed the skewness of the target variable, `SalePrice`, which often violates the assumptions of linear regression. We applied a log transformation to `SalePrice`, which stabilized the variance and made the distribution more symmetric. Finally, we refitted the linear regression model using the refined set of predictors after stepwise regression and the removal of variables with high VIF. This final model provides a robust framework for understanding the relationships within the data and generating reliable predictions.

Results

Table 3: Model Comparison Metrics

Metrics	Initial.Model	Stepwise.Model	Remove.VIF.5	log_transform
Rsquared	0.8788	0.8758	0.8394	0.8809
Adjusted Rsquared	0.8645	0.8695	0.8315	0.8751
F-statistic	61.8200	137.8000	106.8000	151.2000
p-value	0.0000	0.0000	0.0000	0.0000

p-value for all models show they are statistically significant.

1. **Initial Model**

- R-squared: 0.8788, indicating that 87.88% of the variability in the sale price is explained by the model.
- Adjusted R-squared: 0.8645, suggesting a good model fit after adjusting for the number of predictors.
- F-statistic: 61.82, indicating the overall model is significant.

2. **Stepwise Model**

- R-squared: 0.8758, slightly lower than the Initial Model.
- Adjusted R-squared: 0.8695, slightly higher than the Initial Model, indicating a better-adjusted model with potentially fewer predictors.
- F-statistic: 137.8, significantly higher than the Initial Model, suggesting a stronger overall model fit.

3. **Remove.VIF.5 Model**

- R-squared: 0.8394, lower than both the Initial and Stepwise Models, indicating a decrease in the explained variance.
- Adjusted R-squared: 0.8315, also lower, indicating that the model is less effective.
- F-statistic: 106.8, higher than the Initial Model but lower than the Stepwise Model.

4. **Log_Transform Model**

- R-squared: 0.8809, the highest among all models, indicating the best model fit.
- Adjusted R-squared: 0.8751, also the highest, showing that this model performs best even after accounting for the number of predictors.
- F-statistic: 151.2, the highest value, indicating a very strong overall model fit.

Discussion

- The Log_Transform Model performs the best overall, as indicated by the highest R-squared, Adjusted R-squared, and F-statistic values. This suggests that transforming the target variable improved the model's performance.
- The Stepwise Model is also strong, with a high Adjusted R-squared and F-statistic, indicating that automatic variable selection improved the model over the Initial Model.
- The Remove.VIF.5 Model has the lowest R-squared and Adjusted R-squared values, indicating that removing predictors with high multicollinearity (VIF > 5) resulted in a less predictive model. However, it still has a statistically significant F-statistic and p-value, showing that it's a valid but potentially less effective model.

Conclusion

The analysis revealed several significant predictors of sale price, highlighting that overall quality and condition (OverallQual) are critical, as homes with higher ratings tend to command significantly higher prices. Additionally, the number of garage spaces (GarageCars) positively correlates with property value, emphasizing the importance of garage capacity for resale potential. Neighborhoods also play a vital role, with certain areas exhibiting much higher average sale prices, suggesting that buyers should focus on neighborhoods with a history of appreciation. Features related to basements, such as finishing quality and conditions, were found to enhance home value. Finally, addressing the skewness of the target variable, SalePrice, through log transformation improved model reliability, indicating that while prices are generally high, opportunities exist for buyers willing to invest in properties needing upgrades or renovations.

When it comes to fitting the model, The Log_Transform Model emerges as the best performing model, with the highest R-squared and Adjusted R-squared values, indicating superior predictive accuracy after data transformation. This suggests that addressing issues like skewness or non-linearity significantly improved model performance. The Stepwise Model also performed well, showing that automatic variable selection enhanced predictive power without sacrificing model simplicity. In contrast, the Remove.VIF.5 Model, which reduced multicollinearity by eliminating predictors with VIF > 5, resulted in a less accurate model, highlighting the trade-off between minimizing multicollinearity and maintaining predictive accuracy. Overall, data transformation appears to offer the most substantial improvement in this analysis.

Potential buyers should conduct thorough inspections focusing on overall quality, basement conditions, and garage capacity, as properties excelling in these areas are likely to yield better investment returns. It is crucial to consider long-term value by evaluating properties in high-quality neighborhoods with strong appreciation potential. Seeking professional valuations can help validate pricing and provide leverage during negotiations. Buyers should also remain open to properties needing cosmetic upgrades, as these can offer substantial returns on investment after renovations. Finally, staying informed about market conditions, including interest rates and housing supply, will empower buyers to make timely and strategic purchasing decisions.