AMES HOUSING EDA REPORT

1. Introduction

The Ames housing dataset contains information from the Ames Assessor's Office to compute the assessed values of individual residential properties sold in Ames, Iowa between 2006 and 2010. The dataset was originally developed by the Ames Assessor's Office for tax assessment purposes, but because of the depth and spread of observations, it can be used to develop models to predict the selling price of a home. The dataset itself contains 82 variables, thoroughly covering all aspects of a property, inside-and-out, that could predict the home selling price.

This level of granularity, however, may be unnecessary for the purpose of developing a linear regression model that predicts the price of home; therefore, variable selection will be implemented prior to fitting the model to the data. In the dataset, overly specific variables—such as "miscellaneous feature" (features not covered in other categories), "screen porch" (square footage of screened porch), and "masonry veneer area" (square footage of masonry veneer), for example—may not add meaningfully to the conclusions drawn from the model and could be removed from the dataset. Additionally, extreme outliers of the response variable (i.e., SalePrice) will need to be manipulated as these values can cause downstream inaccuracies. One way to do that is to apply a logarithmic transformation to rescale the data and reduce extreme values.

The Ames housing dataset will allow us to examine relationships between property characteristics and sale price by constructing regression models aimed at predicting selling price. This is because the dataset consists of very detailed information regarding the structure, location, and qualitative attributes of the homes sold in Ames, indicating caution when attempting to generalize results to other markets. Additionally, the model would capture associations rather than casual effects as many variables are subjective assessments or context-dependent measurements (e.g., ordinal variables like "Overall Quality"). Therefore, care must be taken when interpreting results and avoiding casual claims as the model would have a limited scope of interference outside Ames.

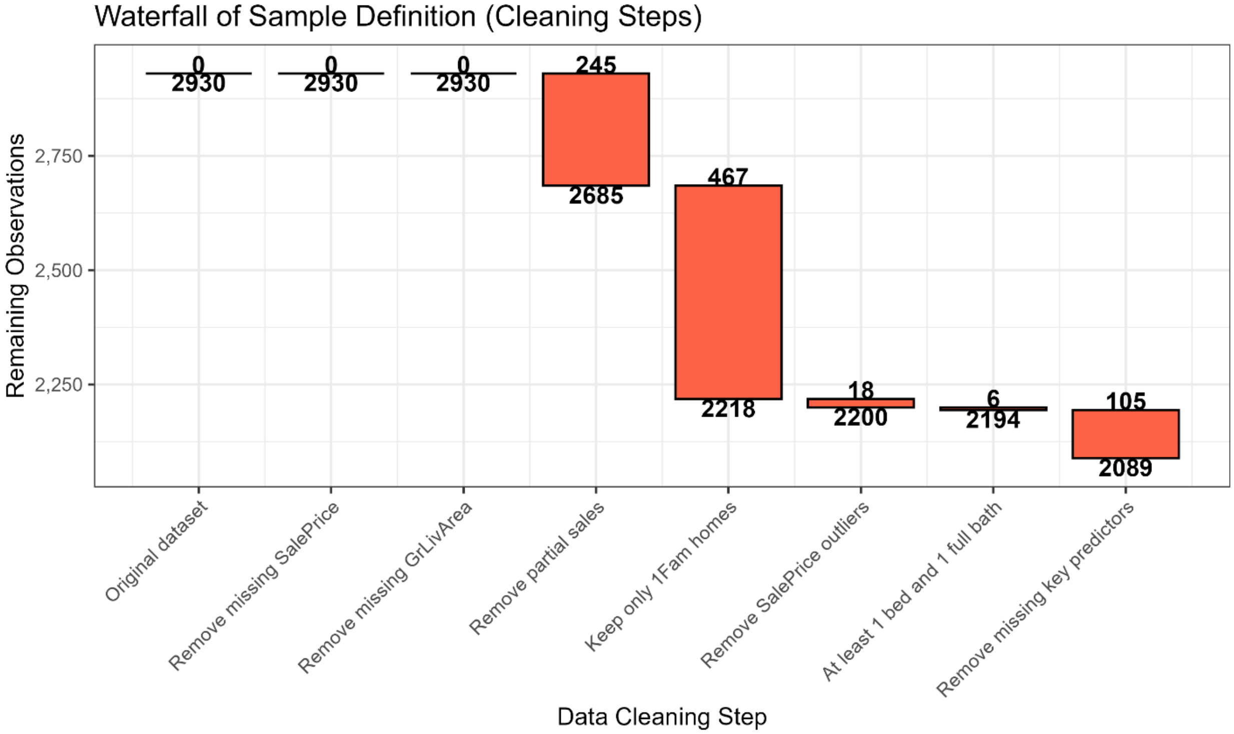
2. Sample Definition

To ensure the data used in the model represents a typical residential housing purchase, a series of drop conditions were implemented to define the final sample. Observations were removed if they did not meet quality, completeness, or population-of-interest criteria. We excluded any properties with missing values for the response variable (*SalePrice*), observations with implausible or zero finished living area, properties with partial or incomplete sales, properties not classified as a residential single-family home, properties that had at least 1 bedroom above ground and 1 full bathroom, and transactions with extreme outliers that were inconsistent with the main housing data. We also removed records with extensive missingness in key predictor variables used in the model. Key predictor variables were determined based on absolute correlation value to response variable with a threshold of 0.5 or higher. Table 1 and Figure 1 present a visual representation of the cleaning process defined above, indicating both the number of remaining observations after each step and the number of observations removed until our final sample dataset.

Table 1. Waterfall Summary of Sample Definition.

Condition	Remaining Observations	Number of Observations Removed
Original dataset	2930	0
Remove missing SalePrice	2930	0
Remove missing GrLivArea	2930	0
Remove partial sales	2685	245
Keep only single-family homes	2218	467
Remove SalePrice extreme outliers (<$465,500)	2200	18
At least 1 bed and 1 bath	2194	6
Remove missing key predictors	2089	105
Final Analytical Sample		**2091**

Figure 1. Waterfall Chart of Sample Definition.



After applying these drop conditions, our final analytical sample consists of 2,089 residential home sale transactions in Ames. These methods were chosen to highlight what we would consider a typical transaction in Ames.

3. Data Quality Check

The twenty variables selected for the data quality review were chosen because they are commonly used in regression modeling of housing prices and represent a broad scope of the dataset. They include structural features (e.g., square footage, number of rooms, bathrooms, and garage size), temporal characteristics (e.g., year built and year remodeled), and marketing descriptors (e.g., neighborhood, zoning, and sale condition). The selected variables contain both numeric and categorical/ordinal types to ensure the data quality assessment covered different data types and their potential structuring errors. Moreover, these variables were relevant to the objective of predicting a typical residential home transaction in Ames. Therefore, assessing the quality and logical consistency of these variables is essential to building a valid and interpretable predictive model.

The data quality check found that the selected variables were generally clean and consistent with expectations based on the data dictionary. No observations were deemed as impossible values such as negative or zero sale price, negative square footage, or negative room counts. In addition, no extreme housing size or sale price outliers were detected by quantile analysis. One illogical value was detected: a case where the year a home was remodeled (*YearRemodel*) was before the year the home was built (*YearBuilt*). Since a property cannot be remodeled before it was built, this was classified as a data entry error and removed before subsequent analysis, leaving the final sample of 2,088 transactions. Moreover, *LotFrontage* was the only variable with missing values, with 401 missing observations (roughly ~14% of the data), whereas all other variables were complete. Tables A1 and A2 in the appendix illustrate the results of the data quality data quality assessment described above.

4. Initial Exploratory Analysis

Following the data quality assessment, an initial exploratory data analysis (EDA) was conducted to better understand the relationships between the response variable (*SalePrice*) and a selected group of predictors. Ten variables were chosen from the broader set of data quality assessment variables because they contain important characteristics of a home such as home size, quality, age, and location. Similar to the data quality assessment, these variables cover different data types, ensuring robust interpretation of the results from the model and are a representative sample of what a typical transaction in Ames.

The variables selected were: *SalePrice*, *GrLivArea*, *LotArea, TotalBsmtSF*, *GarageArea*, *OverallQual*, *FullBath*, *BedroomAbvGr*, *YearBuilt*, and *Neighborhood*. Five of these variables are continuous, while the other variables are discrete or categorical. The EDA will be divided into two sections to clearly present the findings of the EDA.

4.1 EDA of Continuous Variables

The continuous variables examined were *GrLivArea*, *LotArea*, *TotalBsmtSF*, and *GarageArea*. For each of these variables, histograms were produced to assess distributional characteristics, and scatterplots with LOESS smoothers were generated to evaluate their relationship with *SalePrice*.

The histograms indicated the distribution of the data had a degree right skewness, with most of the homes sales in the lower to middle range of the distribution, while a smaller number of high-priced transactions in the right tail. *LotArea* demonstrated the strongest skewness from all the predictors, driven by a small number of very large lots; suggesting that transformations like logarithms may be appropriate for modeling. *GrLivArea*, *TotalBsmtSF*, and *GarageArea* exhibited were approximately unimodal with slight skewness and several high-value homes that pull the distribution to the right.

Scatterplots of each predictor against *SalePrice* indicate a clear positive relationship. The strongest relationship was observed between *SalePrice* and *GrLivArea*, where *SalePrice* increases steadily with living area and the LOESS smoother shows a mostly consistent trend. *TotalBsmtSF* and *GarageArea* show a similar positive association with *SalePrice*, although with slightly more dispersion. *LotArea* displayed a much weaker but still positive relationship, likely due to the influence of large lots that do not proportionately increase market value. Figures A1-A4 in the appendix illustrate the relationships described above.

These results suggest that measures of home size are strongly related to sale price, with above-ground living area being the clear strongest continuous predictor. These variables will be key candidates in subsequent regression model development.

4.2 EDA of Discrete & Categorical Variables

The discrete and categorical variables examined were *OverallQual*, *FullBath*, *BedroomAbvGr*, *YearBuilt*, and *Neighborhood*. Boxplots of *SalePrice* by each category were generated to visualize distributional differences.

OverallQual showed the clearest patterns when compared to *SalePrice* of the discrete and categorical variables. Median *SalePrice* consistently increased with quality rating, and both the spread and upper tail of *SalePrice* were higher for homes with quality ratings of 7 or higher. This suggests that perceived construction and finish quality is a crucial determinant of home value. Similar, but weaker patterns were observed for *FullBath* and *BedroomAbvGr*, where homes with a greater number of full bathrooms and bedrooms generally tended to sell for high prices than homes with fewer bathrooms or bedrooms; however, diminishing results appear beyond three bedrooms, indicating that bedroom count alone may not fully capture home desirability. Figures A5-A7 in the appendix illustrate the relationships described above.

Neighborhood demonstrated substantial variation in SalePrice distribution, with some neighborhoods clustering around significantly higher price levels, while others showed lower and tighter distributions. This confirms that location has a strong effect in the housing market and suggests that including neighborhood indicators will be important for capturing spatial price variation. *YearBuilt* behaved more like an ordinal variable, with newer homes tending to have higher prices, though this is not strictly linear due to remodeling, variation in neighborhood desirability, and interest rates levels. Figures A8 and A9 in the appendix illustrate the relationships described above.

4.3 Summary of Initial EDA findings

The initial EDA confirms that several key variables exhibit strong relationships with *SalePrice*. Larger homes, higher-quality construction, more bathrooms, and favorable neighborhoods are generally associated with higher sale prices. The EDA also revealed skewed distributions and the presence of high-value outliers in multiple variables, which lead to data transformations. Most importantly, the EDA provided insight into which predictors are likely going to be the most influential in the modeling stage. *GrLivArea*, *OverallQual*, and *Neighborhood* appear the most promising, while *LotArea* may require careful treatment due to skewness.

5. Initial Exploratory Analysis for Modeling

The response variable in this problem is *SalePrice*, which is defined as the transaction price of each property in the dataset. An inspection on the distribution of *SalePrice* revealed moderate right skewness, with a relatively small number of high-prices properties concentrated on the upper tail of the distribution. Figure A10 in the appendix illustrates the distribution of the response variable in a histogram. Since many statistical modeling techniques, such as linear regressions, assume approximate normality of residuals and homoscedasticity, it is common to consider a transformation to the response variable. One technique is the logarithmic transformation, log(*SalePrice*). When examining the distribution of log(*SalePrice*), the distribution becomes more symmetrical and closer to normal. Figure A11 in the appendix visualizes the distribution of log(*SalePrice*) in a histogram.

To explore the functional relationship between sale price and home attributes, the variables *GrLivArea*, *TotalBsmtSF*, and *LotArea* were examined in both *SalePrice* and log(*SalePrice*) with LOESS smoothers. *GrLivArea* shows the strongest positive relationship when plotted against SalePrice as seen in Figure A1. When plotted against log(*SalePrice*), the relationship looked almost identical, indicating *GrLivArea* as a robust predictor for sale price. Figure A12 illustrates the relationship between *GrLivArea* and log(*SalePrice*). *TotalBsmtSF* also exhibited a positive association with *SalePrice*, although with greater dispersion than *GrLivArea*. This is as evident in Figure A4, as a fan-shaped pattern was visible at the higher end, indicating increasing variance at higher prices. When plotted against log(*SalePrice*), the fan-shaped pattern, heteroscedasticity, was slightly reduced, supporting the use of the transformation as shown in Figure A13. Lastly, *LotArea* showed the weakest visual relationship with *SalePrice* due to extreme skew and a small set of large lot properties as seen in Figure A2. When plotted against log(*SalePrice*), the heteroscedasticity becomes reduced, however *LotArea* itself remains highly skewed and dominated by a very few large lot properties, suggesting that a transformation of *LotArea* may be necessary, illustrated in Figure A14.

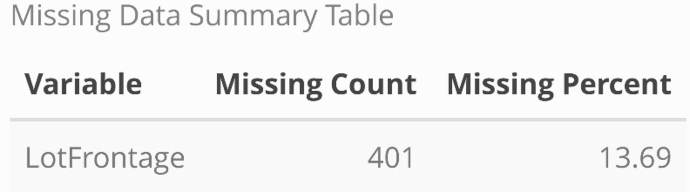
Based on the initial EDA modeling results, several potential difficulties and modeling concerns emerged. First, *SalePrice* exhibits right skew and heavy tails, violating approximate normality assumptions and showing signs of heteroscedasticity. Applying a log transformation mitigates these issues and is likely going to improve model performance. Second, the data contained influential and extreme observations, particularly very large homes and lots that occur infrequently but may disproportionately affect estimated coefficients. Third, the EDA indicates that transformations of certain predictor variables may also be warranted, such as *LotArea* that displays extremely high skewness and is a strong candidate for log transformation, while *GrLivArea* and *TotalBsmtSF* show moderate skewness and may benefit form transformation depending on residuals analysis. Finally, centering and scaling the predictors may further enhance model interpretability and performance during estimation.

6. Conclusions

Overall, the EDA indicates that the dataset is generally suitable for modeling, but several features warrant caution and intentional preprocessing. The distribution of the response variable, *SalePrice*, is right skewed with heavy trails, suggesting potential violations of normality; however, the log transformation of SalePrice improves symmetry and stabilizes variance, making it a strong candidate for subsequent modeling. In addition, a small number of extreme but valid observations (e.g., very large homes and lots) may disproportionately influence the estimates of the model, so care should be taken to mitigate their influence in later stages. Finally, the EDA clearly suggests that some predictors, particularly *LotArea*, may benefit from transformation due to skewness and outliers. Collectively, these findings indicate the need for careful model checking, potential variable transformation, and robust methods as the model process begins.

A. Appendix

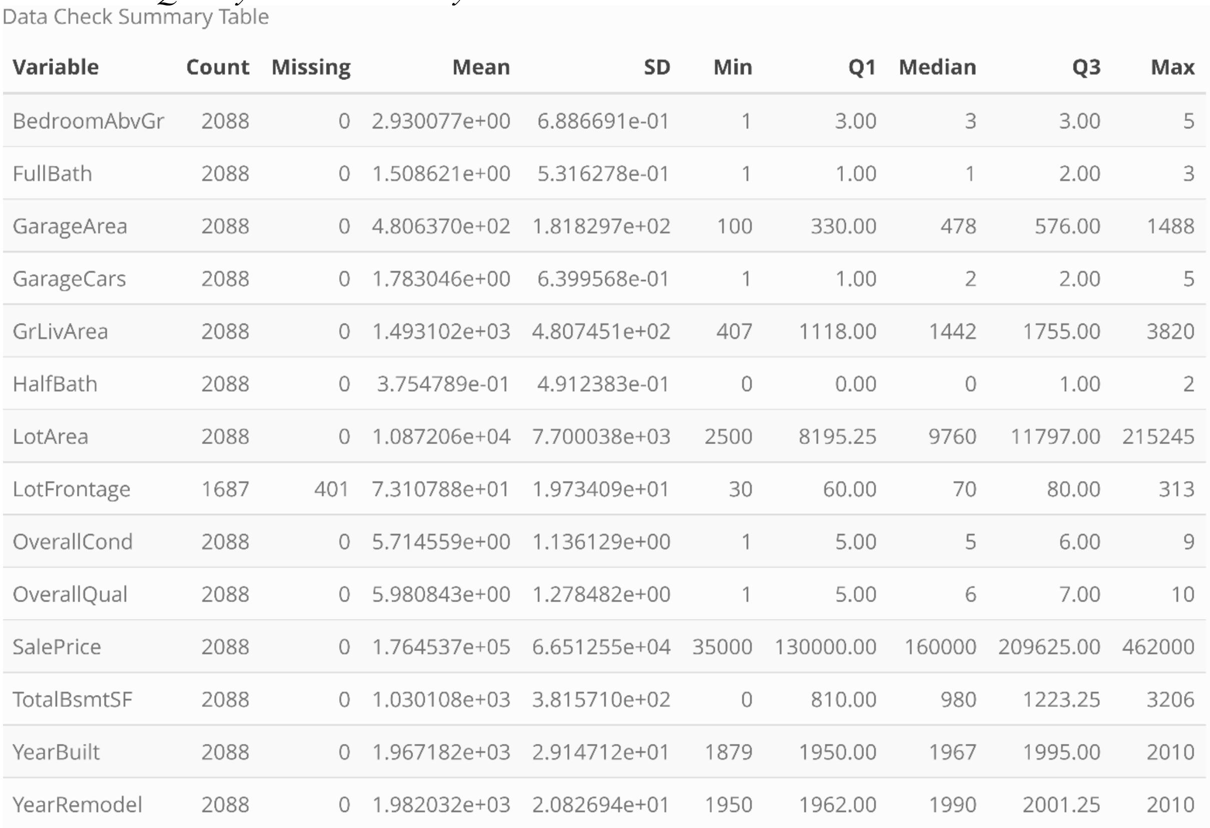
Table A1. Missing Data Summary Table.



Missing Data Summary Table

Variable	Missing Count	Missing Percent
LotFrontage	401	13.69

Table A2. Data Quality Check Summary Table.



Data Check Summary Table

Variable	Count	Missing	Mean	SD	Min	Q1	Median	Q3	Max
BedroomAbvGr	2088	0	2.930077e+00	6.886691e-01	1	3.00	3	3.00	5
FullBath	2088	0	1.508621e+00	5.316278e-01	1	1.00	1	2.00	3
GarageArea	2088	0	4.806370e+02	1.818297e+02	100	330.00	478	576.00	1488
GarageCars	2088	0	1.783046e+00	6.399568e-01	1	1.00	2	2.00	5
GrLivArea	2088	0	1.493102e+03	4.807451e+02	407	1118.00	1442	1755.00	3820
HalfBath	2088	0	3.754789e-01	4.912383e-01	0	0.00	0	1.00	2
LotArea	2088	0	1.087206e+04	7.700038e+03	2500	8195.25	9760	11797.00	215245
LotFrontage	1687	401	7.310788e+01	1.973409e+01	30	60.00	70	80.00	313
OverallCond	2088	0	5.714559e+00	1.136129e+00	1	5.00	5	6.00	9
OverallQual	2088	0	5.980843e+00	1.278482e+00	1	5.00	6	7.00	10
SalePrice	2088	0	1.764537e+05	6.651255e+04	35000	130000.00	160000	209625.00	462000
TotalBsmtSF	2088	0	1.030108e+03	3.815710e+02	0	810.00	980	1223.25	3206
YearBuilt	2088	0	1.967182e+03	2.914712e+01	1879	1950.00	1967	1995.00	2010
YearRemodel	2088	0	1.982032e+03	2.082694e+01	1950	1962.00	1990	2001.25	2010

Figure A1. Scatterplot of SalePrice & GrLivArea.

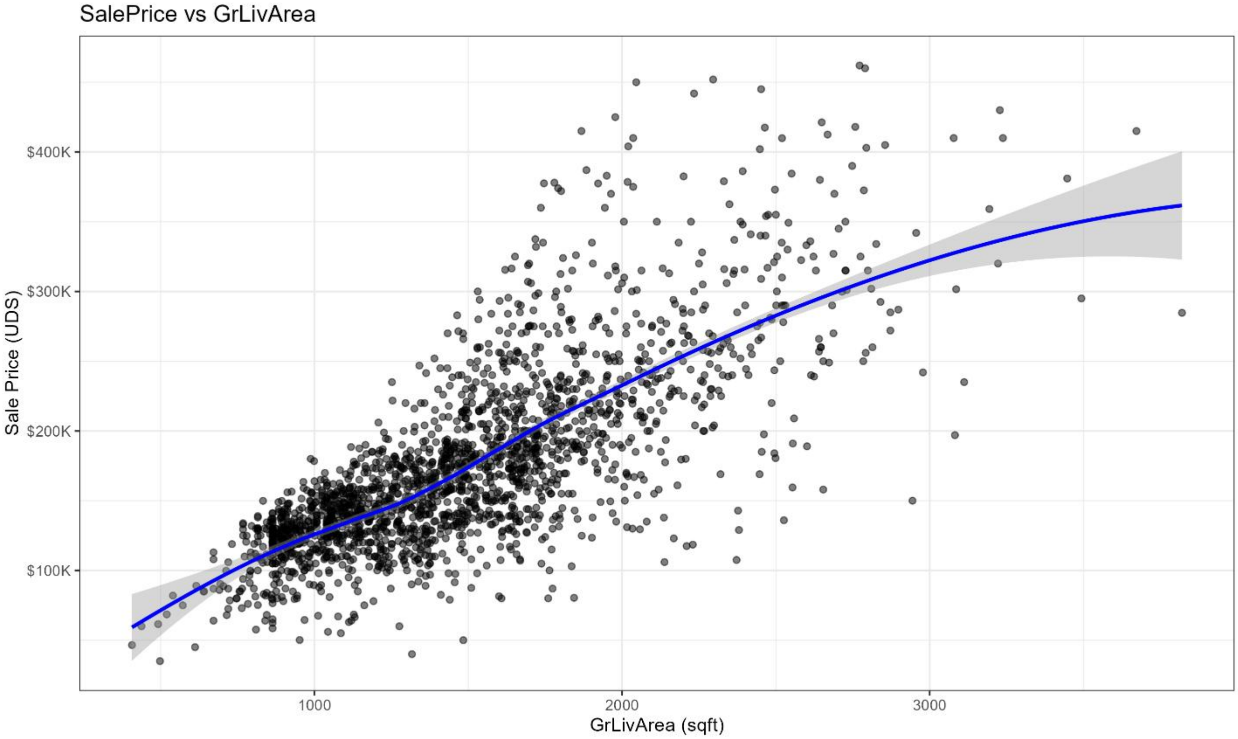


Figure A2. Scatterplot of SalePrice & LotArea.

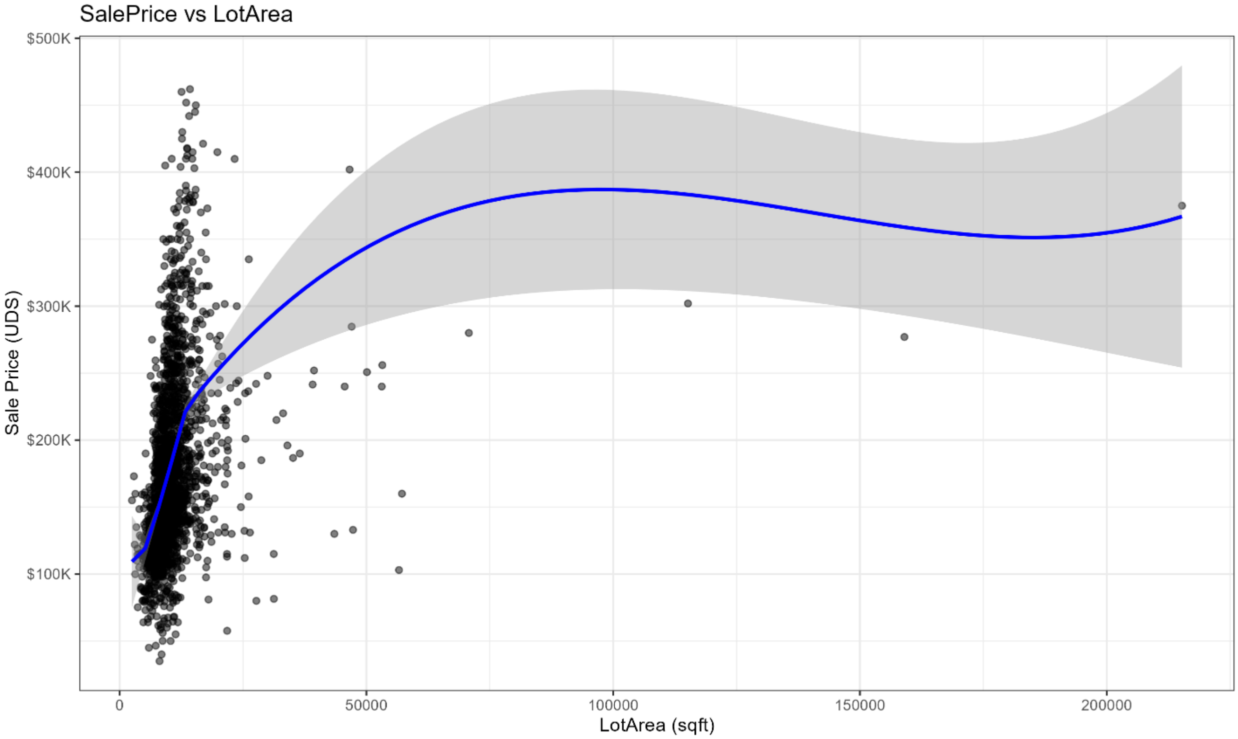


Figure A3. Scatterplot of SalePrice & GarageArea.

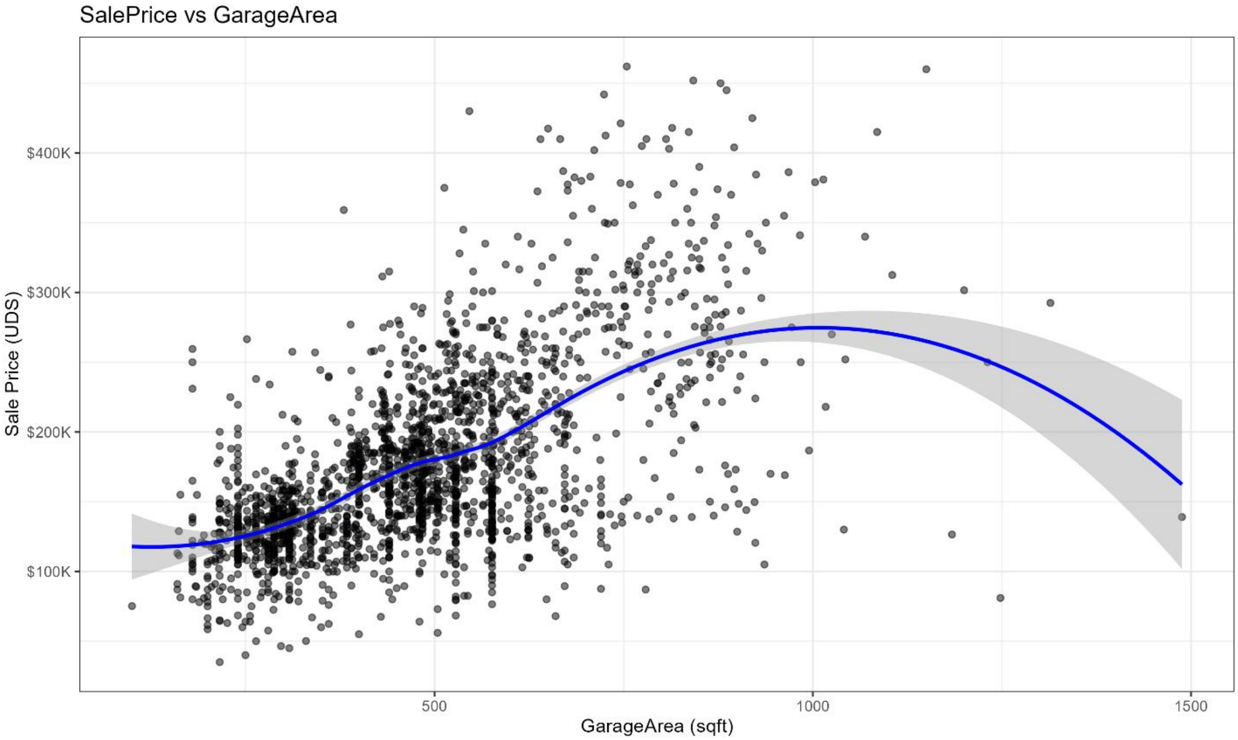


Figure A4. Scatterplot of SalePrice & TotalBsmtSF.

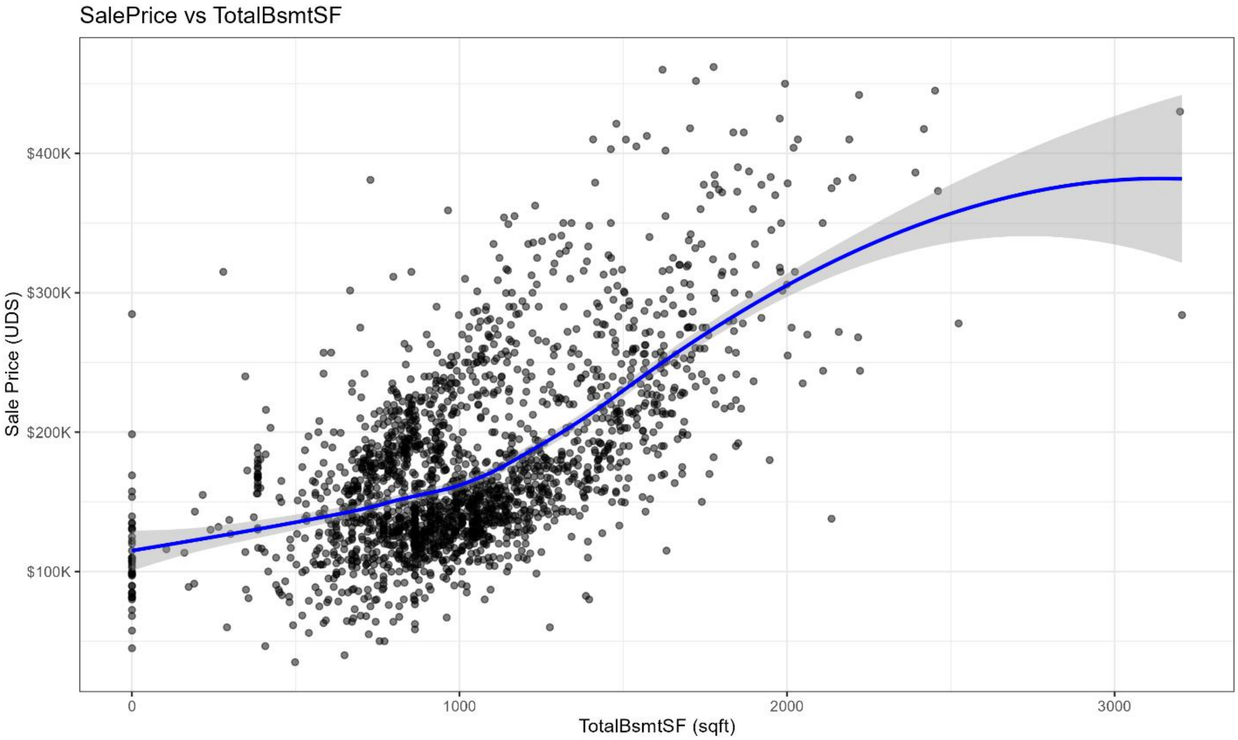


Figure A5. Boxplot of SalePrice by OverallQual.

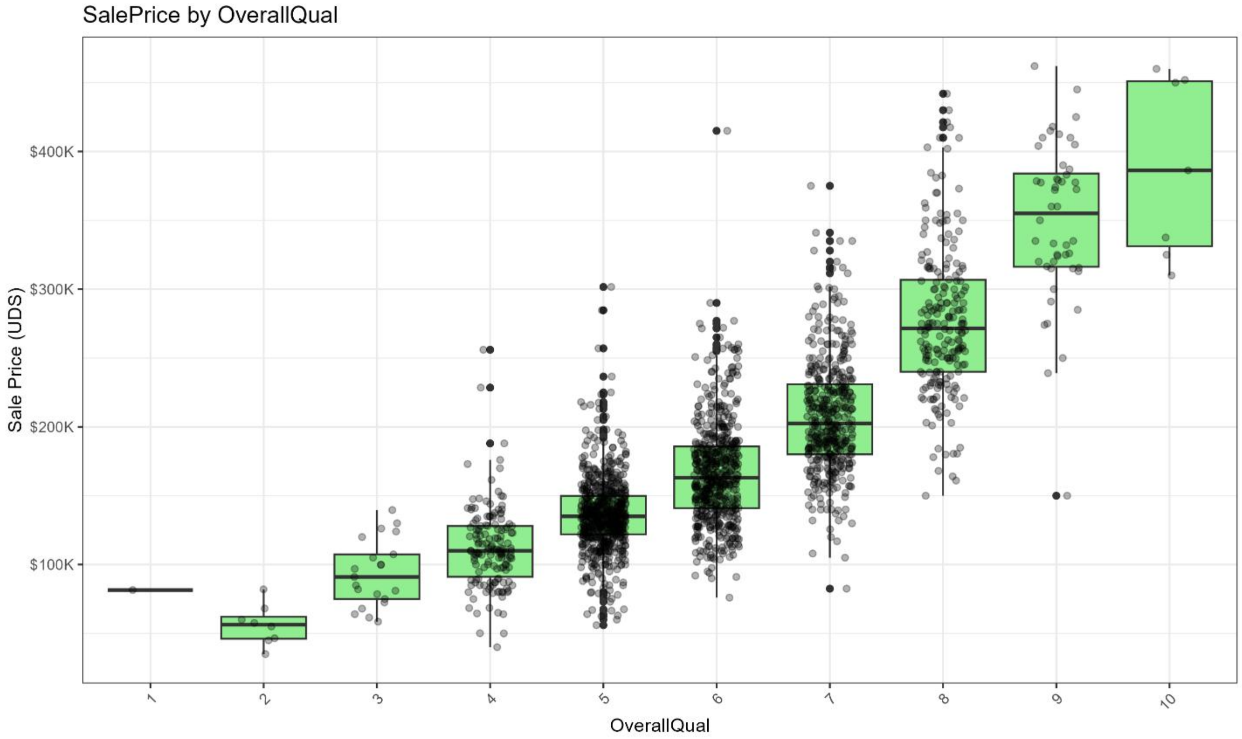


Figure A6. Boxplot of SalePrice by BedroomAbvGr.

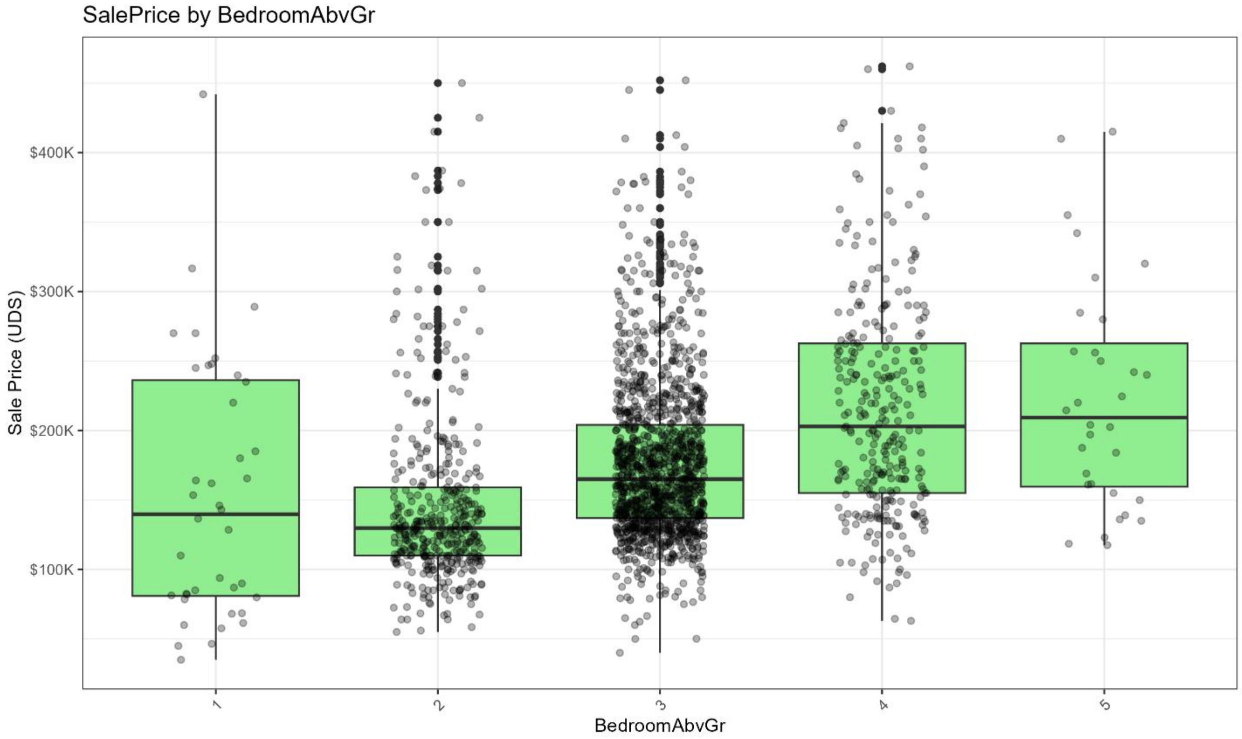


Figure A7. Boxplot of SalePrice by FullBath.

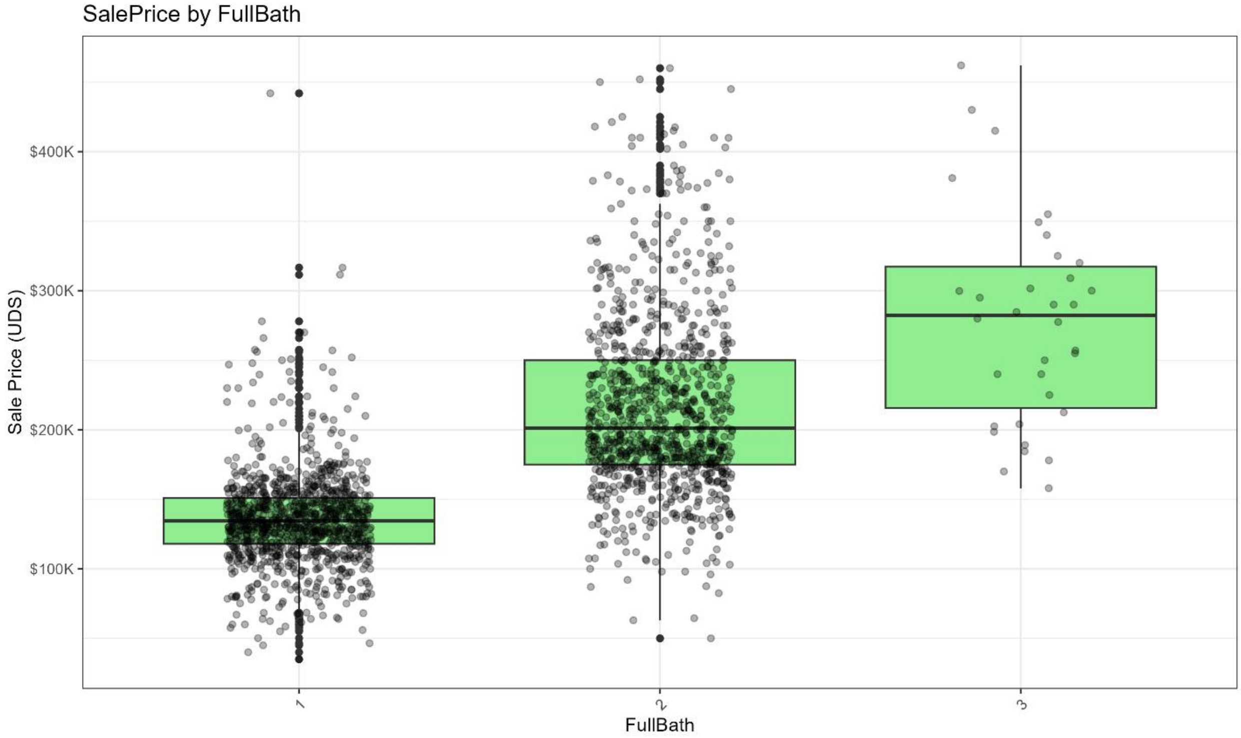


Figure A8. Boxplot of SalePrice by Neighborhood.

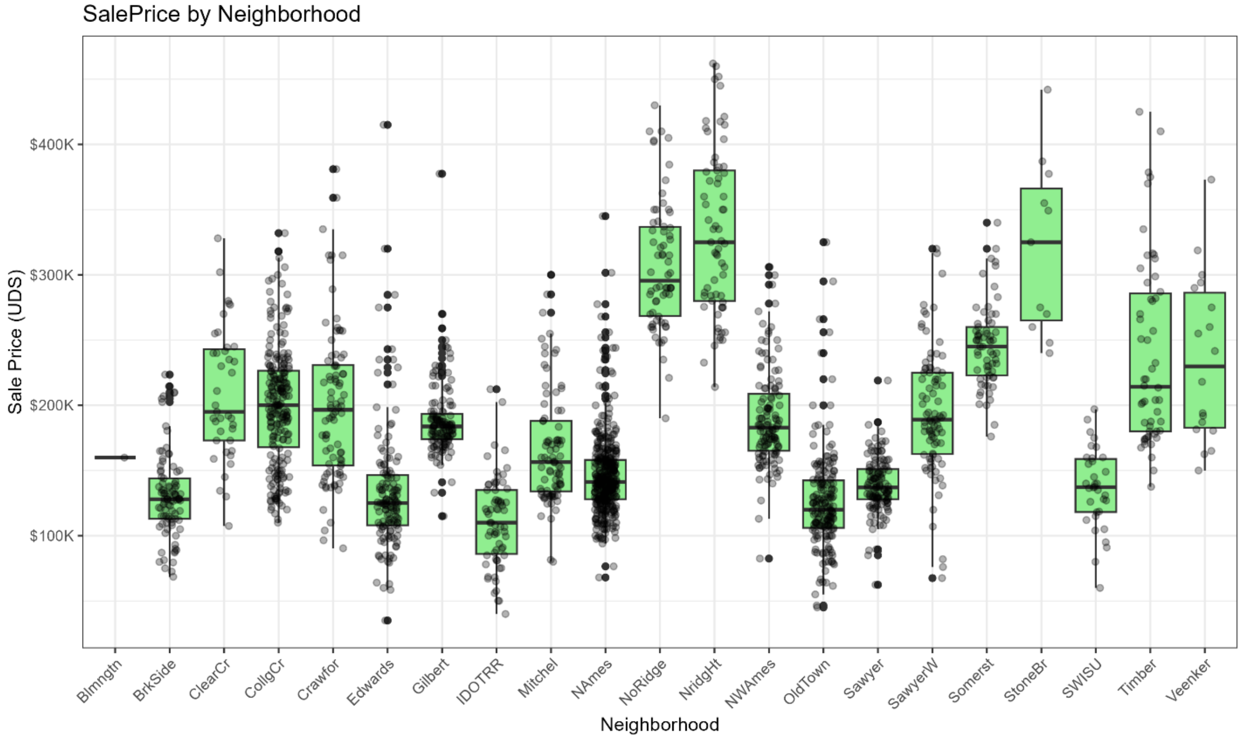


Figure A9. Boxplot of SalePrice by YearBuilt.

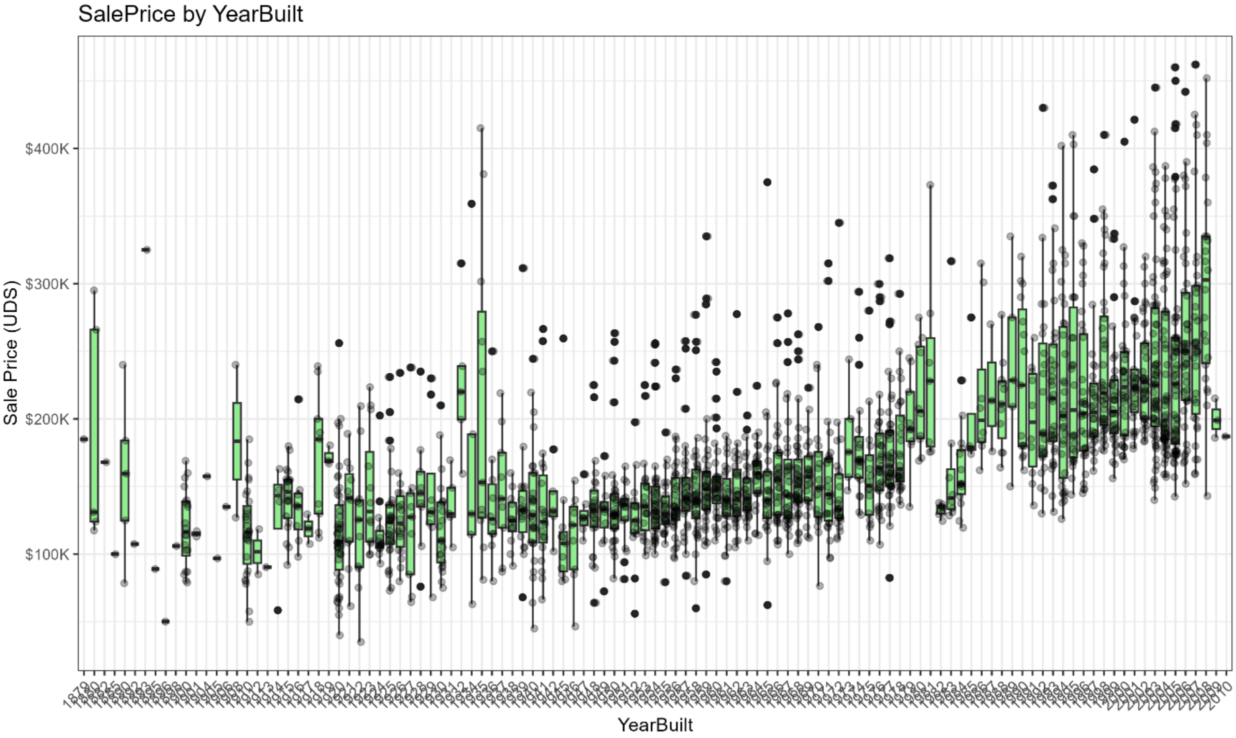


Figure A10. Histogram of SalePrice.

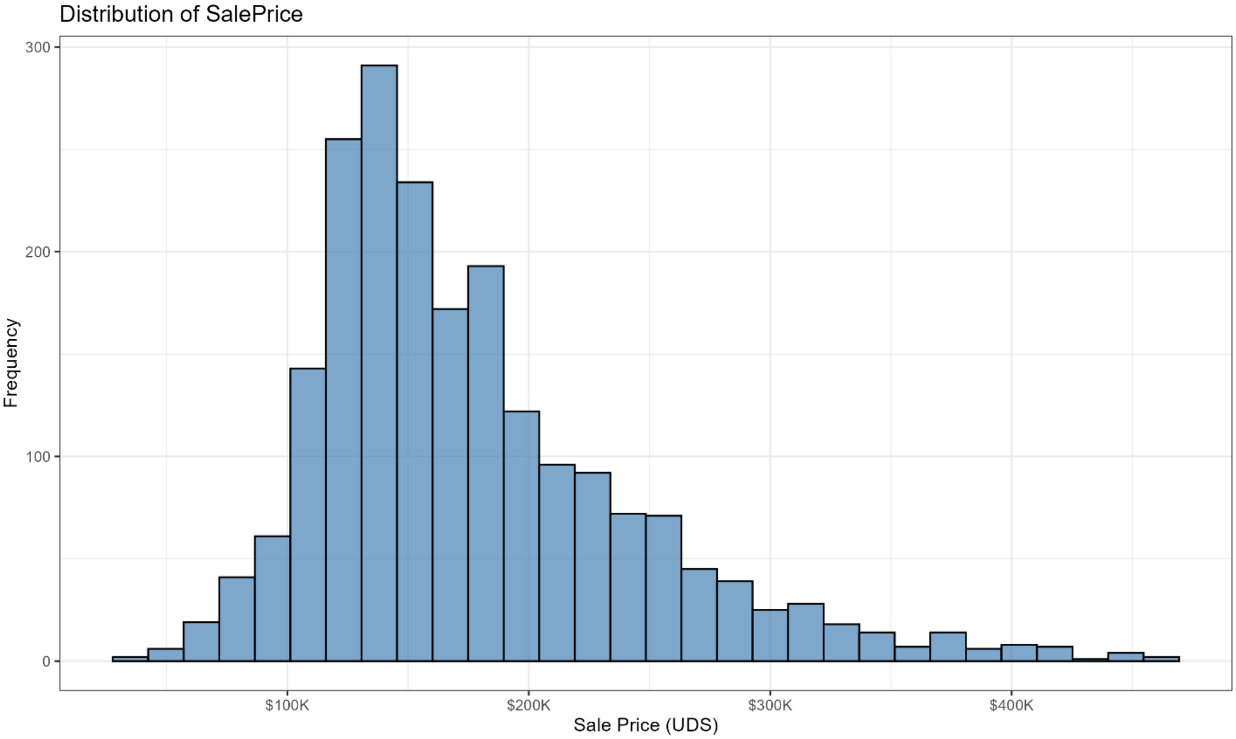


Figure A11. Histogram of logSalePrice.

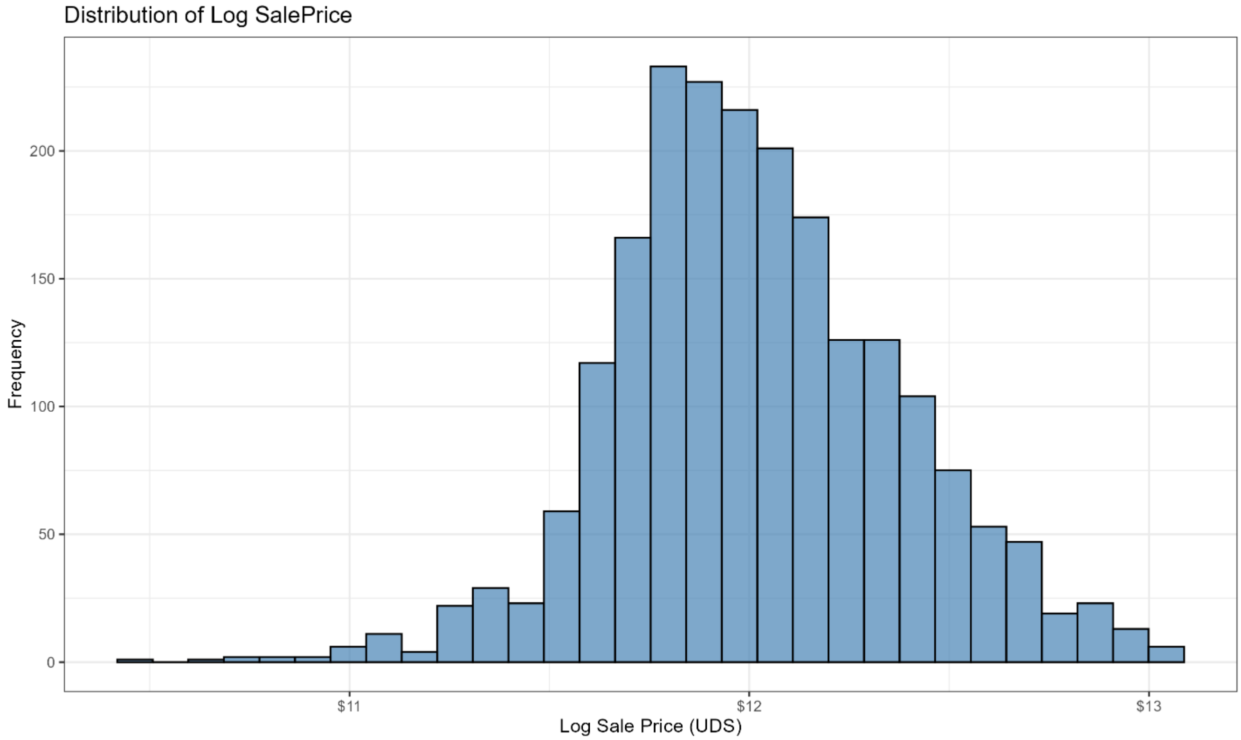


Figure A12. Scatterplot of log(SalePrice) vs GrLivArea.

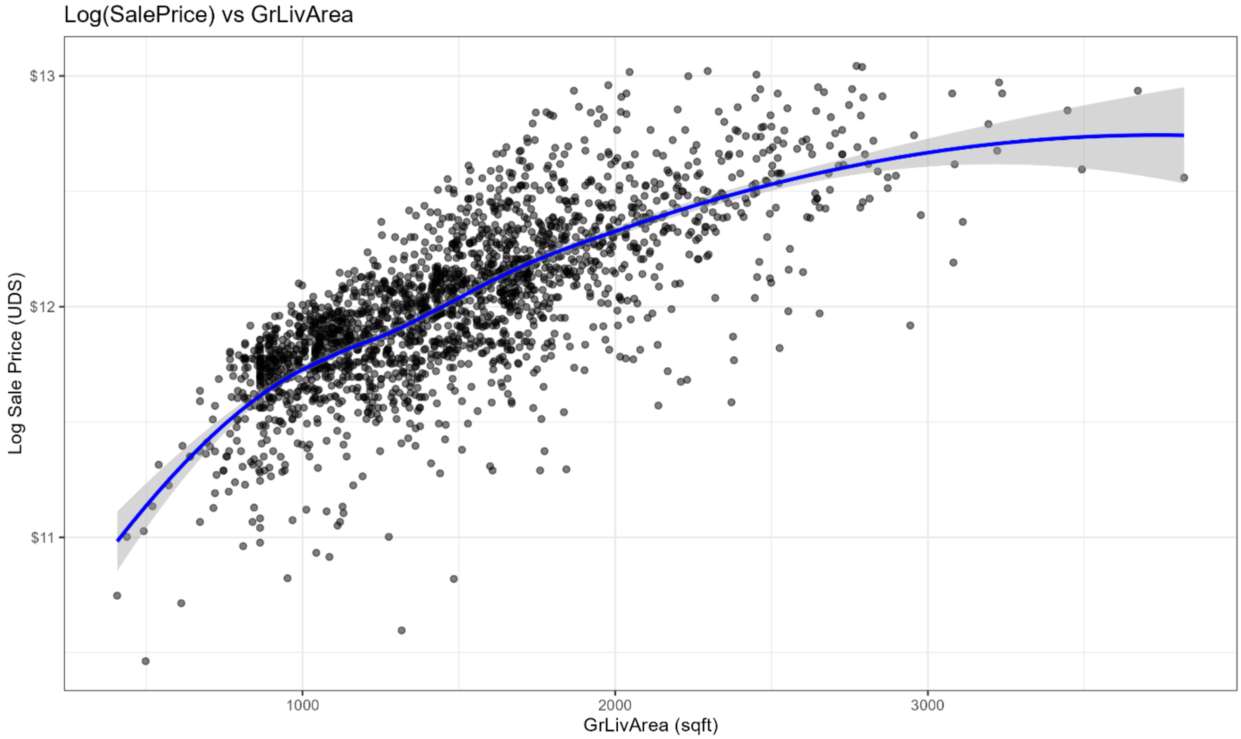


Figure A13. Scatterplot of log(SalePrice) vs TotalBsmtSF.

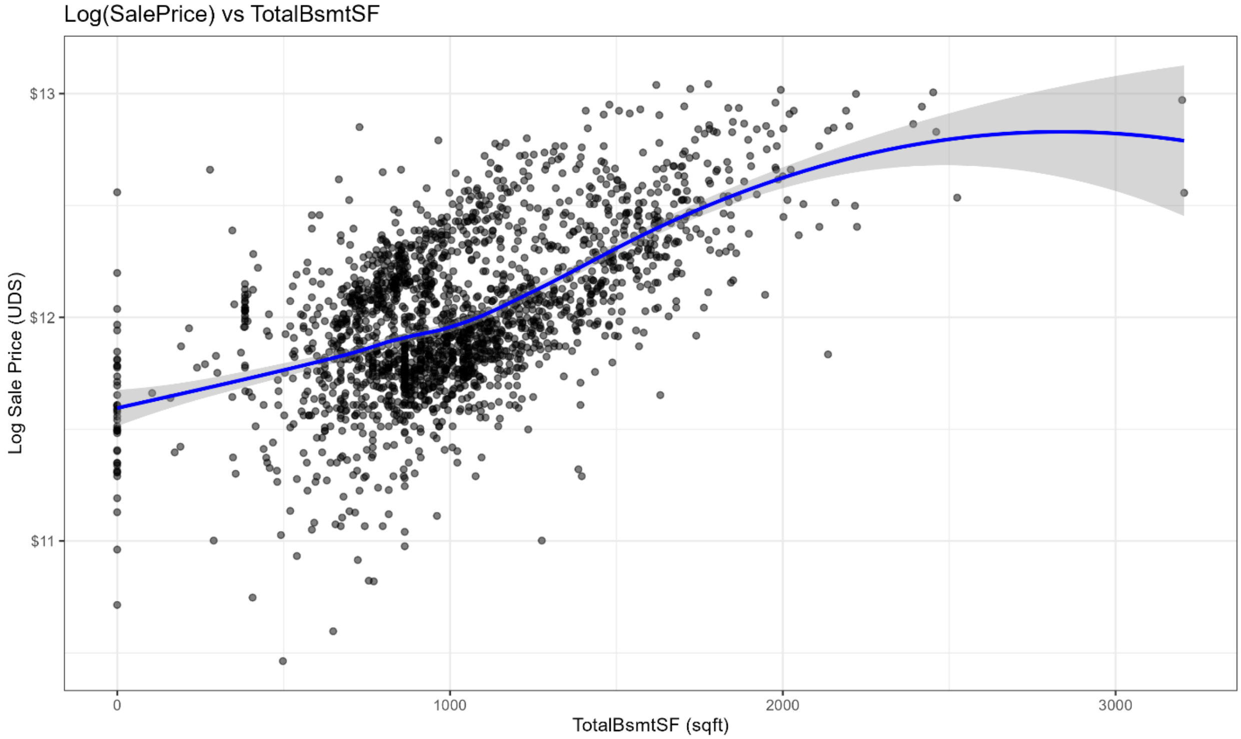
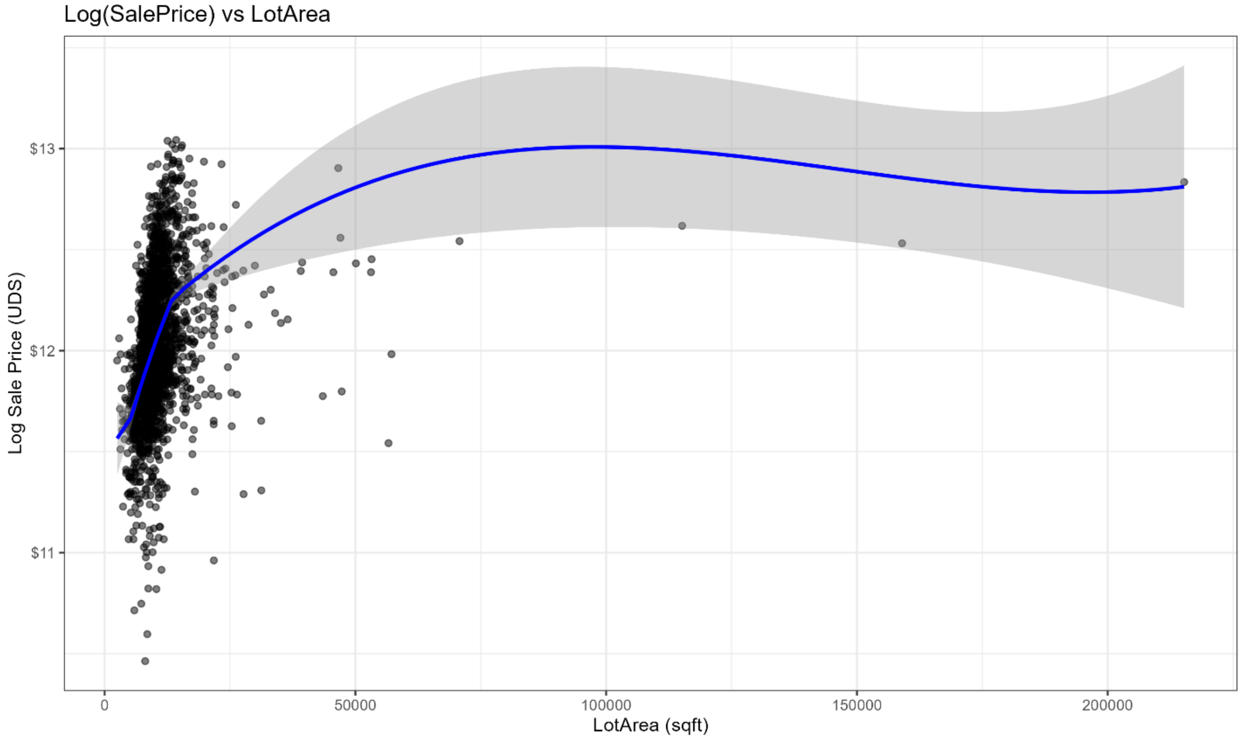


Figure A14. Scatterplot of log(SalePrice) vs LotArea.


Log(SalePrice) vs LotArea
Log Sale Price (UDS)
$13
$12
$11
LotArea (sqft)