Ames Housing Data Analysis Report

[1] Dataset Summary

The Ames Housing dataset contains detailed information on 2,930 residential property sales in Ames, Iowa, with 81 different features describing each home. These features cover a wide range of aspects, including lot size, neighborhood, year built, number of rooms, quality ratings, and various amenities. The main variable of interest is SalePrice, which records the final sale price of each house. This rich dataset provides an excellent opportunity to explore what factors influence home values and to practice a variety of data analysis and machine learning techniques.

[2] Data Exploration Plan

To get the most out of this dataset, I started by outlining a plan for exploration. My goal was to understand which features have the biggest impact on house prices and to uncover any interesting trends or patterns. I focused on questions like: How are home prices distributed? Which neighborhoods and house styles are most valuable? Are there any missing values or outliers that need attention? My approach included looking at summary statistics, visualizing distributions and relationships, checking for missing data, and brainstorming ideas for new features that could help with prediction. I used Python, pandas, matplotlib, and seaborn for the analysis and visualizations.

[3] Exploratory Data Analysis (EDA)

Diving into the Ames Housing dataset, the first thing that stands out is the wide range of home sale prices. Most homes sell for between $130,000 and $210,000, but there are a few that go for much more—up to $755,000! The average sale price is about $180,800, but the distribution is skewed to the right, meaning there are some high-priced outliers pulling up the mean. Lot sizes also vary a lot, from cozy 1,300 square foot lots to sprawling properties over 215,000 square feet. Most homes were built in the latter half of the 20th century, with a median year built of 1973, but there are some historic homes dating back to the 1800s. When looking at above-ground living area (Gr Liv Area), the average is around 1,500 square feet, but again, there are some very large homes in the mix. Not surprisingly, bigger homes and newer homes tend to fetch higher prices. The correlation between Gr Liv Area and SalePrice is especially strong (0.71), and newer homes (Year Built) also show a solid positive relationship (0.56). Certain neighborhoods—like NoRidge, StoneBr, and NridgHt—stand out for their higher average sale prices, and the 2Story house style is particularly popular and valuable. There are also some clear outliers, both in terms of price and size, which will need to be considered in any modeling.

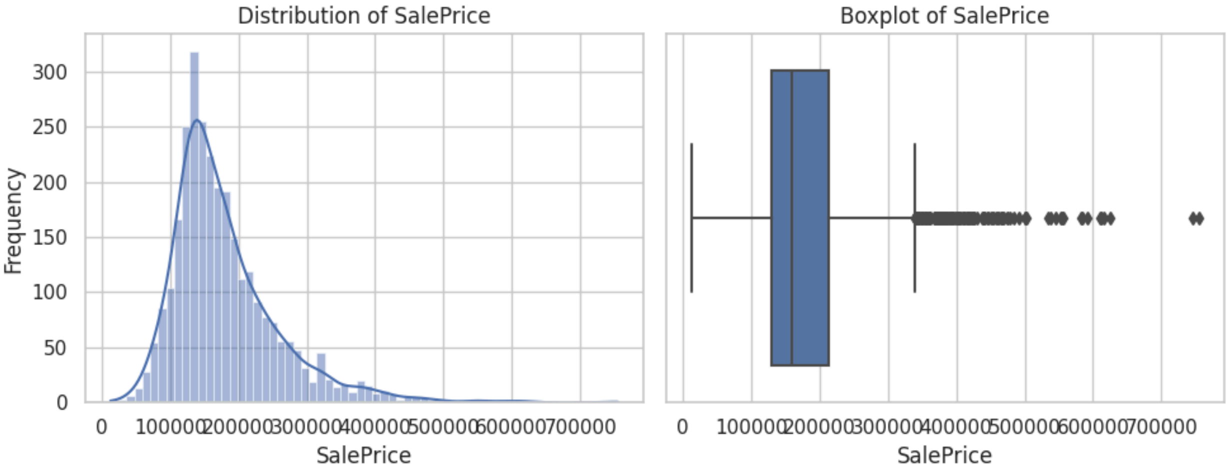


Figure 1: Distribution and Boxplot of SalePrice

The distribution plot shows that most homes are clustered between $100,000 and $300,000, but there are a number of high-value outliers. The boxplot confirms the right-skewed nature of SalePrice and highlights several extreme values above $500,000.

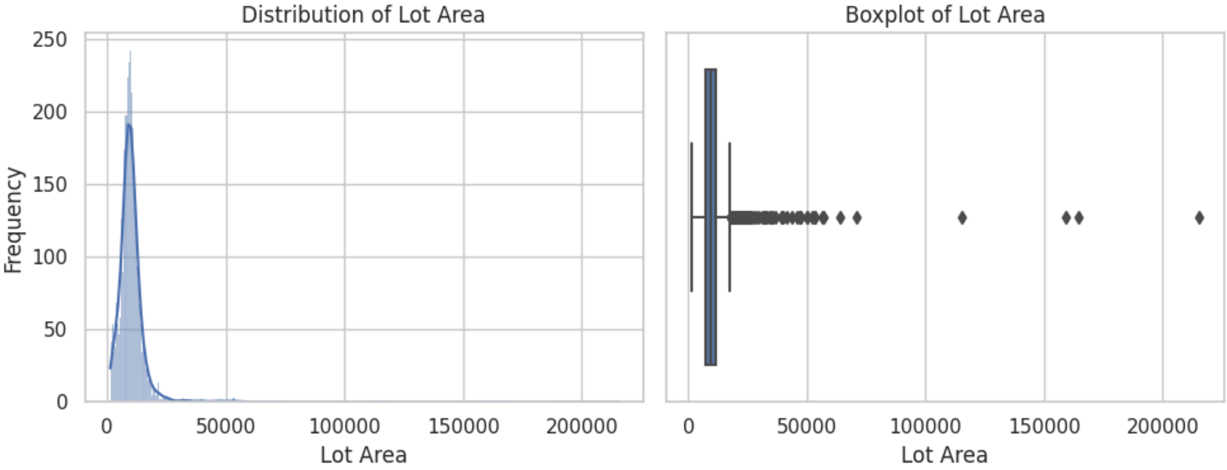


Figure 2: Distribution and Boxplot of Lot Area

Most properties have a lot area under 20,000 square feet, but there are a few very large lots that stand out as outliers. The boxplot makes it clear that the majority of homes are on much smaller lots, with a long tail of larger properties.

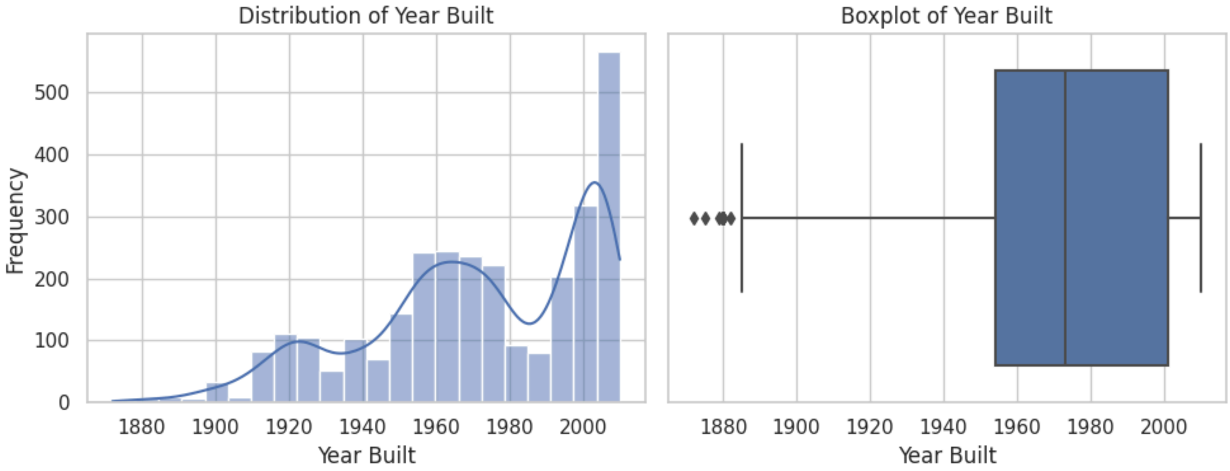


Figure 3: Distribution and Boxplot of Year Built

The distribution of Year Built shows several peaks, with a large number of homes built in the 1960s and 2000s. The boxplot reveals a few very old homes, but most properties were built after 1950.

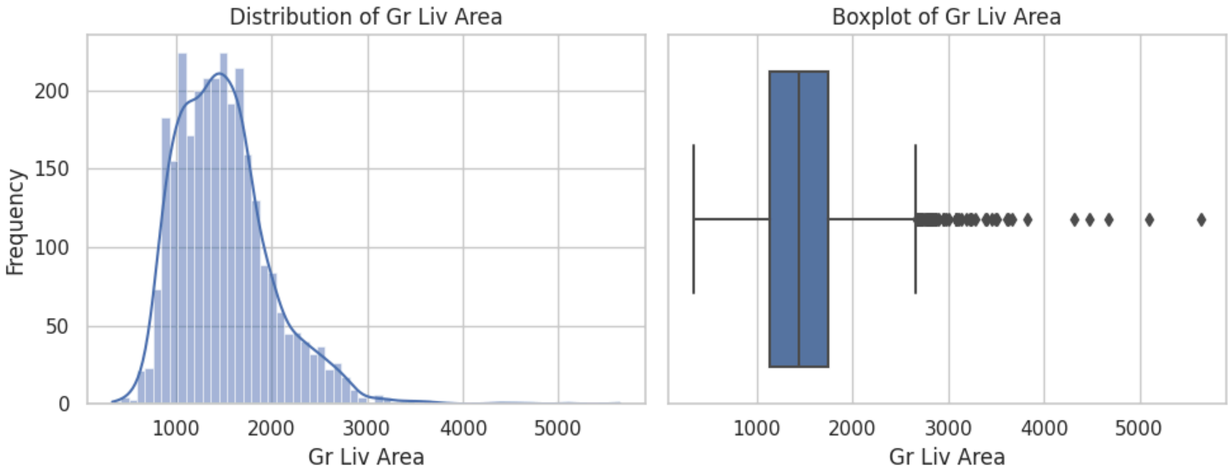


Figure 4: Distribution and Boxplot of Gr Liv Area

Above-ground living area is mostly between 1,000 and 2,500 square feet, with a few very large homes. The boxplot shows that while most homes are of moderate size, there are some significant outliers.

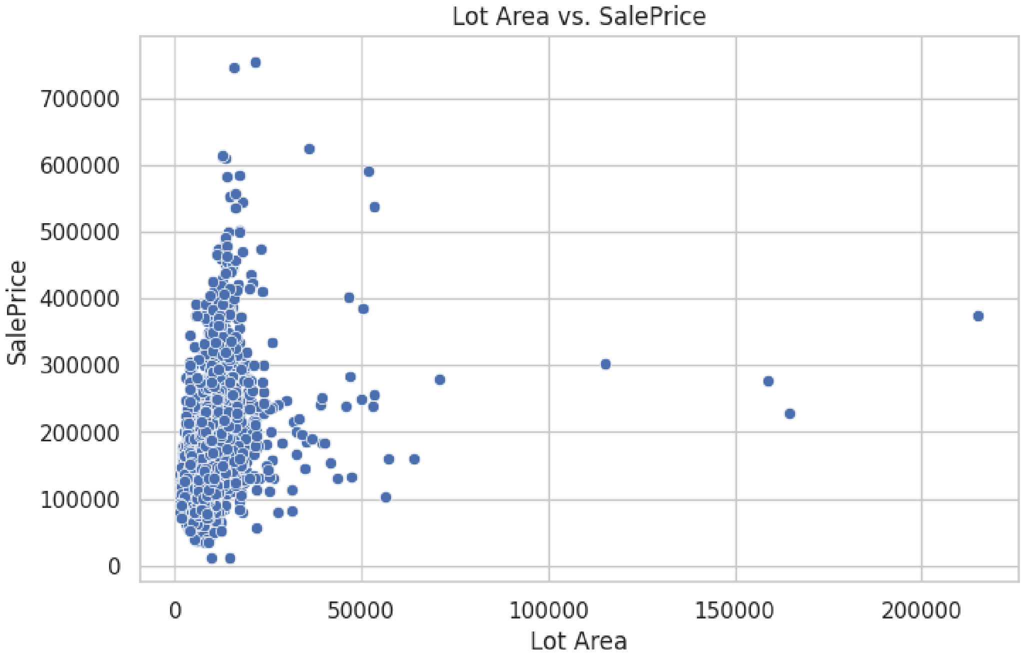


Figure 5: Scatter Plot of Lot Area vs. SalePrice

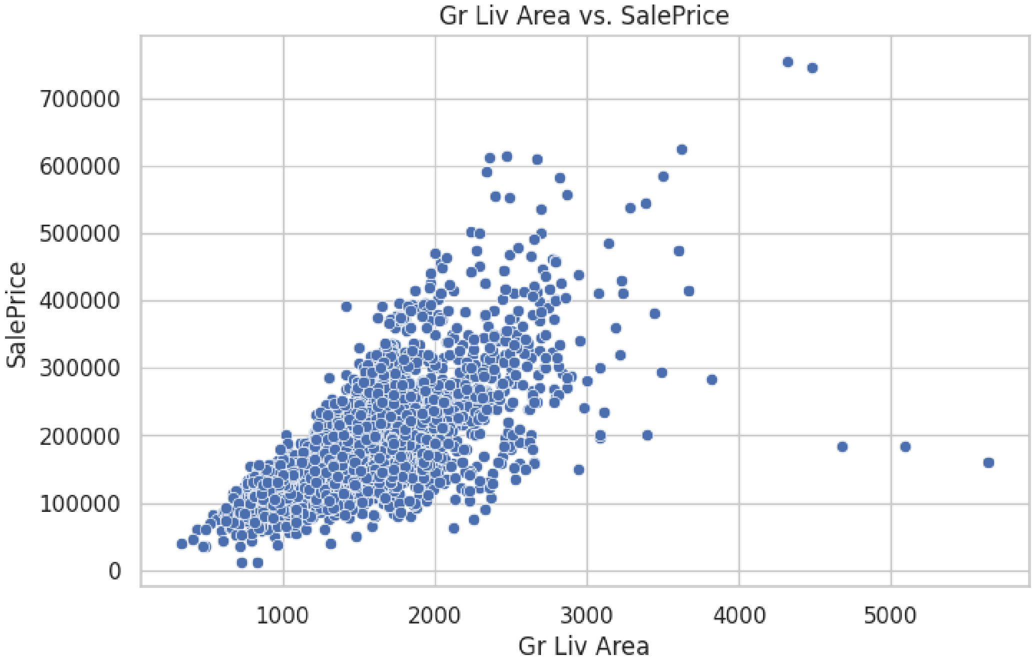


Figure 6: Scatter Plot of Gr Liv Area vs. SalePrice

This scatter plot shows that while larger lots can command higher prices, the relationship is weak and there is a lot of variability. Most high-priced homes do not necessarily have the largest lots.

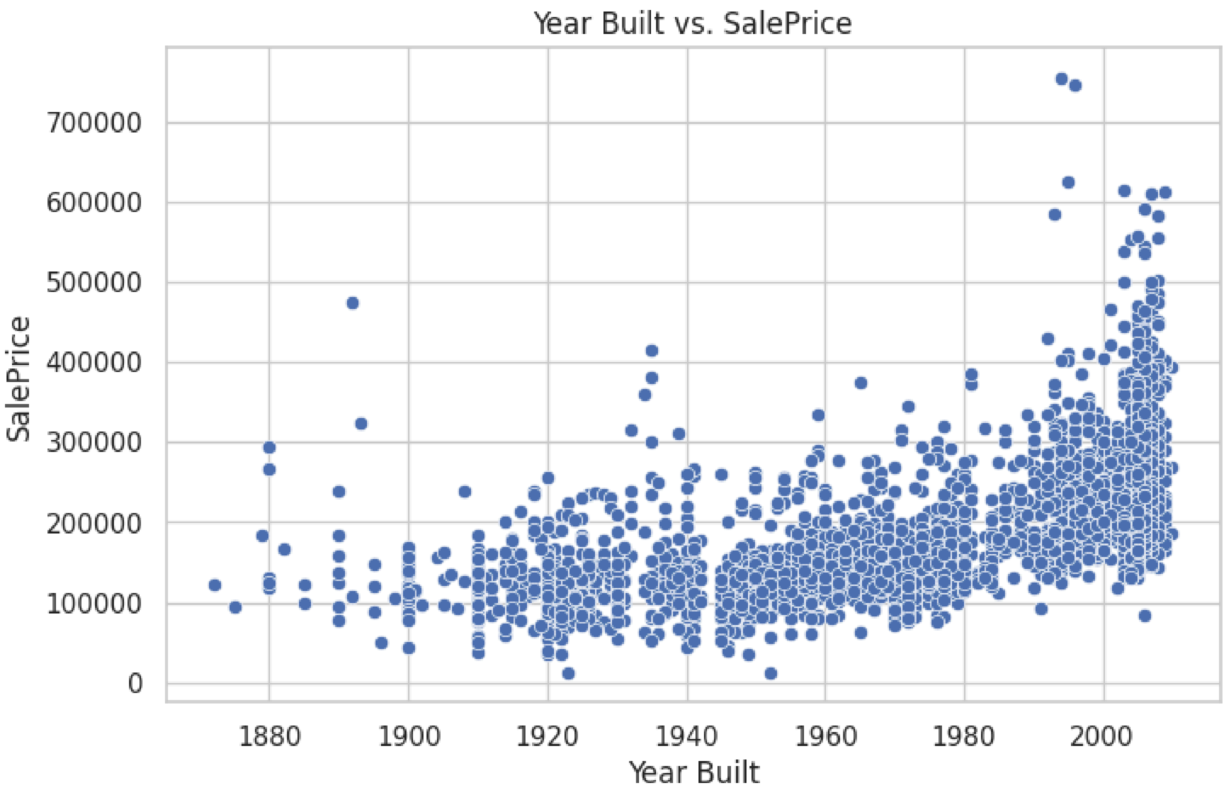


Figure 7: Scatter Plot of Year Built vs. SalePrice

Newer homes generally have higher sale prices, especially those built after 2000. The plot shows a positive trend, with some older homes still achieving high prices, likely due to renovations or location.

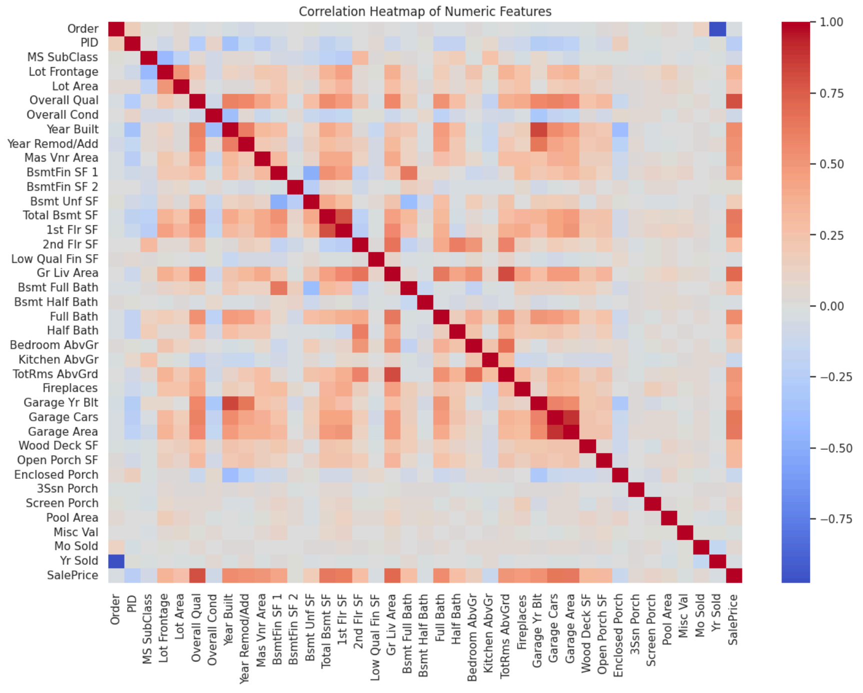


Figure 8: Correlation Heatmap of Numeric Features

The heatmap highlights strong positive correlations between SalePrice and features like Gr Liv Area, Overall Qual, and Garage Area. It also shows which variables are less related to price, helping to guide feature selection for modeling.

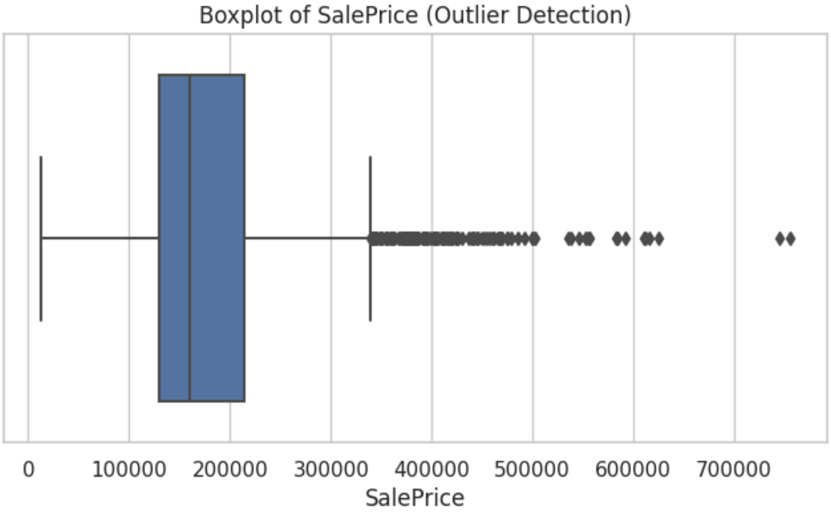


Figure 9: Boxplot of SalePrice (Outlier Detection)

This boxplot makes it easy to spot outliers in SalePrice, with several homes selling for much more than the typical range. These outliers may need special handling in modeling.

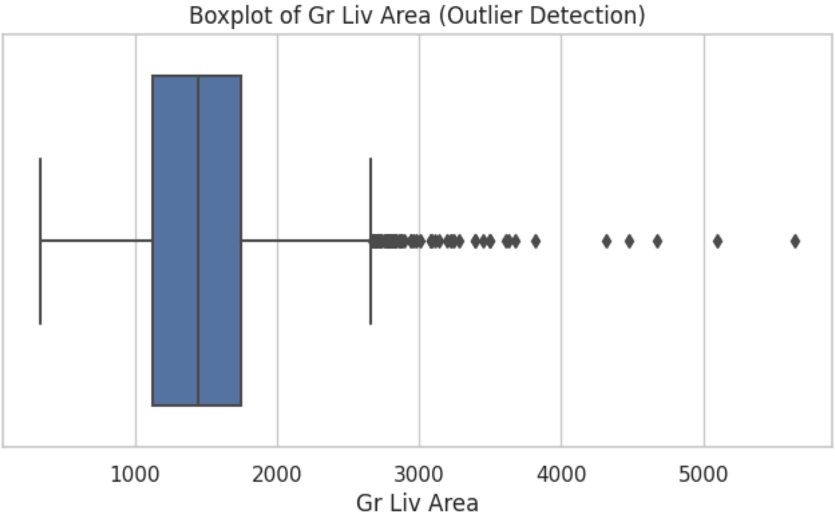
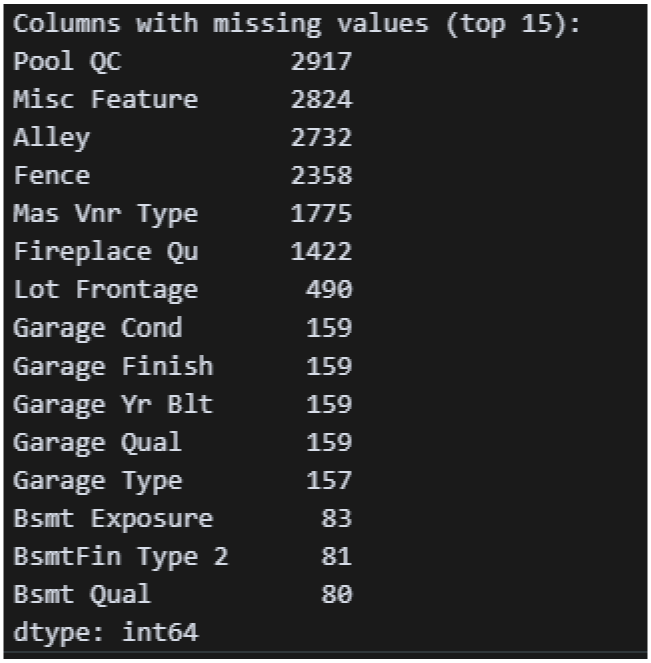


Figure 10: Boxplot of Gr Liv Area (Outlier Detection)

The boxplot shows that most homes have a Gr Liv Area below 2,500 square feet, but there are a few very large homes that could influence model results if not addressed.

[4] Data Cleaning & Feature Engineering

As with any real-world dataset, there are some messy parts to clean up. Several columns have missing values, especially those related to amenities that not every home has, like pools, alleys, or fences. For example, almost all homes are missing a value for Pool QC, and many are missing data for Misc Feature, Alley, and Fence. To get the data ready for analysis, I created a couple of new features: TotalBathrooms, which combines all full and half baths (both above and below ground), and Age, which measures how old the home was at the time of sale. I also converted the Neighborhood column into a numeric format to make it easier to use in models. Infinite values were replaced with NaN, and I summarized the missing values to decide how best to handle them—either by imputing or excluding them, depending on the situation.



```
Columns with missing values (top 15):
Pool QC           2917
Misc Feature      2824
Alley             2732
Fence             2358
Mas Vnr Type      1775
Fireplace Qu      1422
Lot Frontage       490
Garage Cond        159
Garage Finish      159
Garage Yr Blt      159
Garage Qual        159
Garage Type        157
Bsmt Exposure       83
BsmtFin Type 2      81
Bsmt Qual           80
dtype: int64
```

Figure 11: Summary Table of Top 15 Columns with Missing Values in the Ames Housing Dataset

This table shows which columns have the most missing data, with Pool QC, Misc Feature, and Alley missing in the vast majority of homes. This helps prioritize which variables may need to be dropped or imputed.

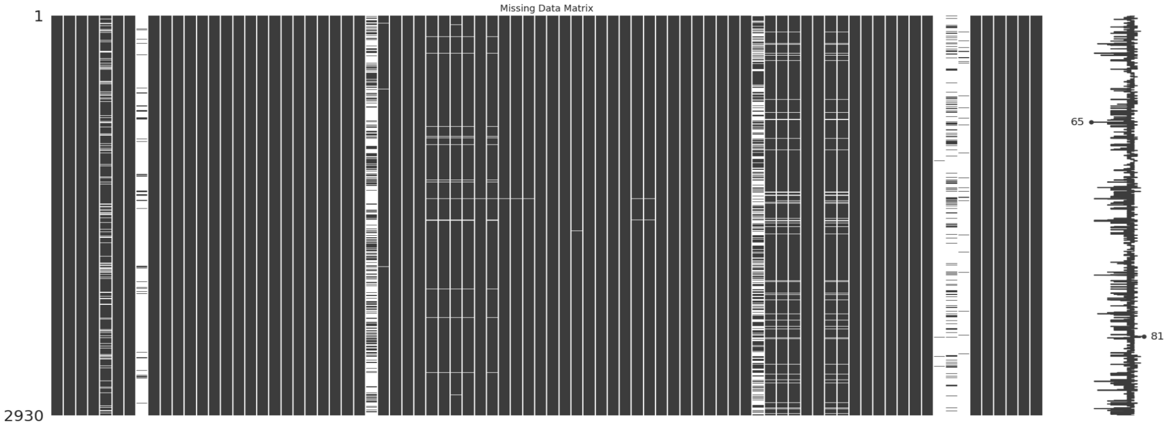


Figure 12: Missing Data Matrix Visualization for the Ames Housing Dataset

The missing data matrix provides a visual overview of where missing values occur in the dataset. Columns with lots of white lines have more missing data, making it easy to spot problem areas.

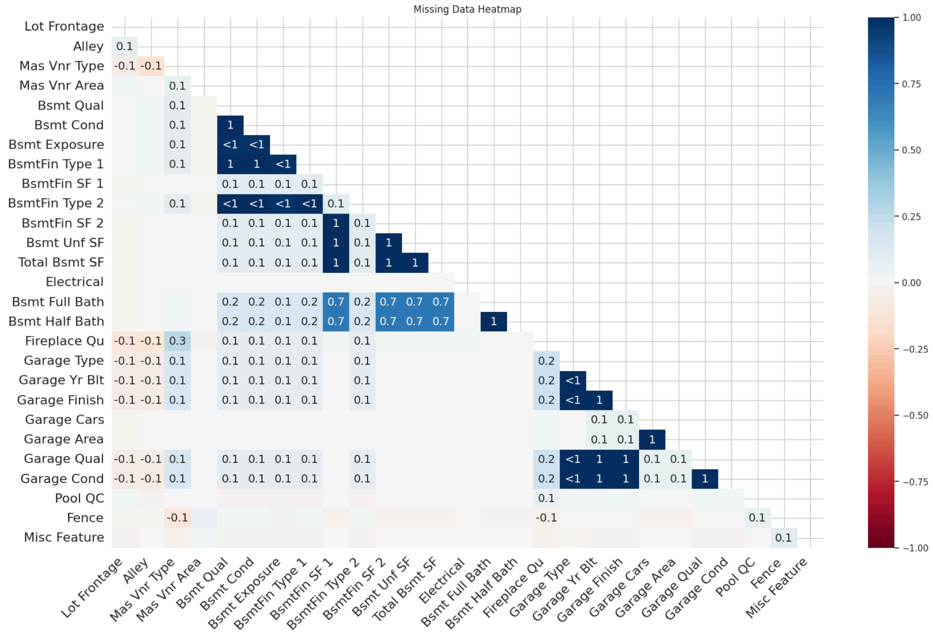
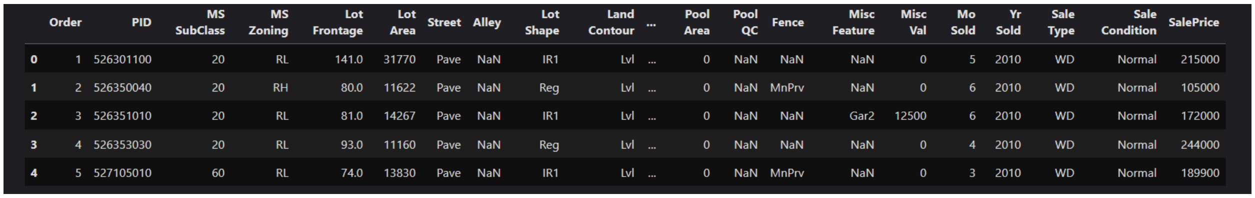


Figure 13: Missing Data Heatmap Showing Correlations in Missingness Between Columns

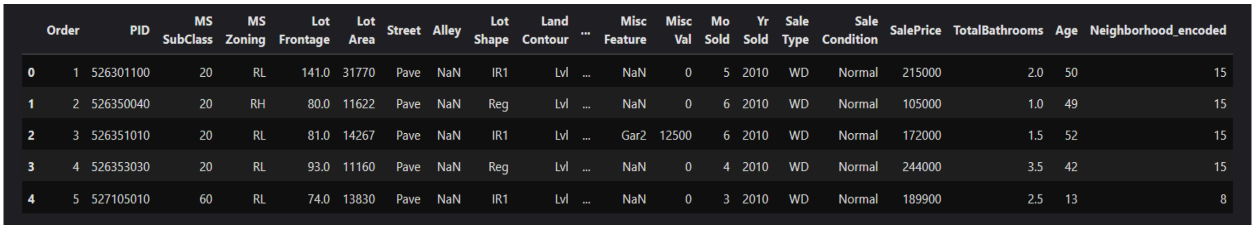
The heatmap shows which columns tend to have missing values together. For example, several basement-related features are often missing in the same homes, which may indicate a structural reason (like no basement).

Before:



	Order	PID	MS SubClass	MS Zoning	Lot Frontage	Lot Area	Street	Alley	Lot Shape	Land Contour	...	Pool Area	Pool QC	Fence	Misc Feature	Misc Val	Mo Sold	Yr Sold	Sale Type	Sale Condition	SalePrice
0	1	526301100	20	RL	141.0	31770	Pave	NaN	IR1	Lvl	...	0	NaN	NaN	NaN	0	5	2010	WD	Normal	215000
1	2	526350040	20	RH	80.0	11622	Pave	NaN	Reg	Lvl	...	0	NaN	MnPrv	NaN	0	6	2010	WD	Normal	105000
2	3	526351010	20	RL	81.0	14267	Pave	NaN	IR1	Lvl	...	0	NaN	NaN	Gar2	12500	6	2010	WD	Normal	172000
3	4	526353030	20	RL	93.0	11160	Pave	NaN	Reg	Lvl	...	0	NaN	NaN	NaN	0	4	2010	WD	Normal	244000
4	5	527105010	60	RL	74.0	13830	Pave	NaN	IR1	Lvl	...	0	NaN	MnPrv	NaN	0	3	2010	WD	Normal	189900

After:



	Order	PID	MS SubClass	MS Zoning	Lot Frontage	Lot Area	Street	Alley	Lot Shape	Land Contour	...	Misc Feature	Misc Val	Mo Sold	Yr Sold	Sale Type	Sale Condition	SalePrice	TotalBathrooms	Age	Neighborhood_encoded
0	1	526301100	20	RL	141.0	31770	Pave	NaN	IR1	Lvl	...	NaN	0	5	2010	WD	Normal	215000	2.0	50	15
1	2	526350040	20	RH	80.0	11622	Pave	NaN	Reg	Lvl	...	NaN	0	6	2010	WD	Normal	105000	1.0	49	15
2	3	526351010	20	RL	81.0	14267	Pave	NaN	IR1	Lvl	...	Gar2	12500	6	2010	WD	Normal	172000	1.5	52	15
3	4	526353030	20	RL	93.0	11160	Pave	NaN	Reg	Lvl	...	NaN	0	4	2010	WD	Normal	244000	3.5	42	15
4	5	527105010	60	RL	74.0	13830	Pave	NaN	IR1	Lvl	...	NaN	0	3	2010	WD	Normal	189900	2.5	13	8

Figure 16: before/after screenshot of feature engineering

This before-and-after comparison shows the addition of new features like TotalBathrooms and Age, which help capture more information about each home and improve model performance.

[5] Key Findings & Insights

A few things became clear during the analysis. First, the size and quality of the home (Gr Liv Area and Overall Qual) are the biggest drivers of price, along with the neighborhood. Newer homes and those in the most desirable neighborhoods (NoRidge, StoneBr, NridgHt) consistently sell for more. The data also has some outliers and missing values, so it’s important to use robust methods when building models. Overall, the dataset is rich and offers a lot of potential for predictive modeling.

[6] Hypothesis Formulation

1. Houses in the "NridgHt" neighborhood have higher average SalePrice than those in "NAmes".
2. Houses with more than 2 bathrooms have higher SalePrice than those with 2 or fewer.
3. Newer houses (Age < 10 years) have higher SalePrice than older houses (Age ≥ 10 years).

[7] Hypothesis Testing & Significance Analysis

To see if the difference in sale prices between "NridgHt" and "NAmes" is real or just due to chance, I ran a t-test. The results were very clear: the t-statistic was 23.28 and the p-value was less than 0.0001. This means there is a statistically significant difference in average sale price between these two neighborhoods—homes in "NridgHt" really do sell for more.

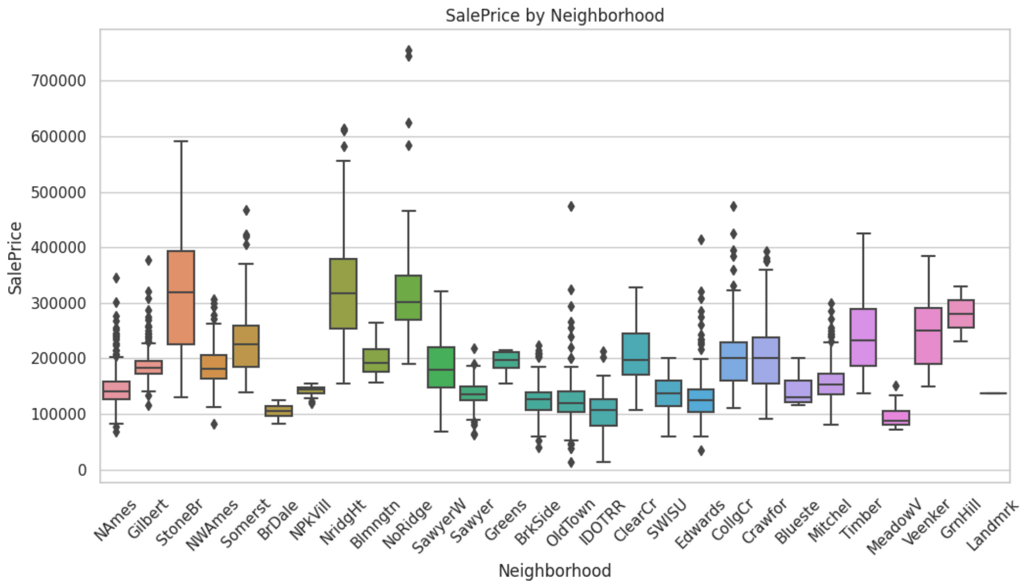


Figure 8: Boxplot of SalePrice by Neighborhood

This boxplot visually confirms that homes in "NridgHt" and a few other neighborhoods have much higher median sale prices than those in "NAmes" and most other areas, supporting the statistical test results.

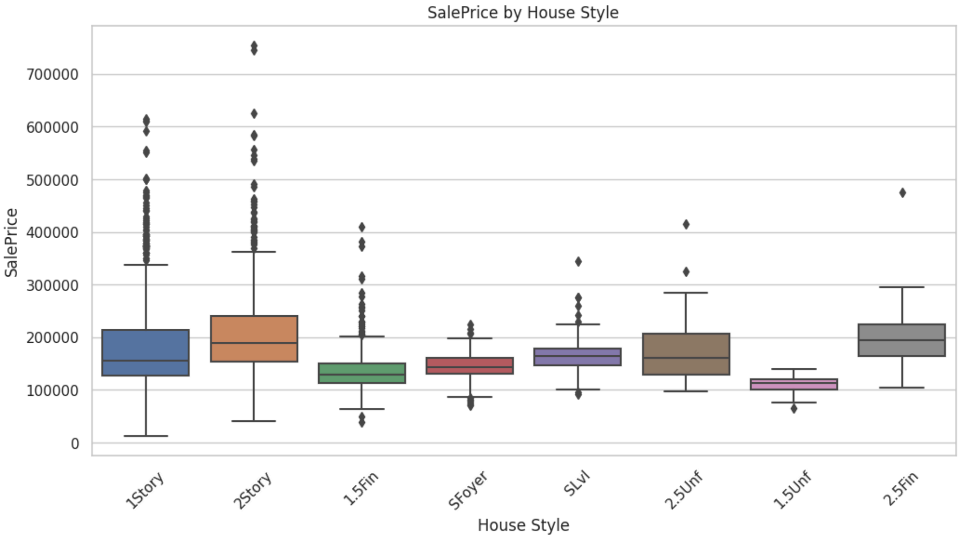


Figure 9: Boxplot of SalePrice by House Style

The boxplot shows that 2Story homes tend to have the highest sale prices, while 1.5Unf and 1.5Fin homes are generally less expensive. This highlights the importance of house style in price prediction.

[8] Conclusion & Next Steps

Working with the Ames Housing dataset has shown just how much information you can uncover with careful analysis. The most important factors for predicting price are home size, quality, and location, but there’s still more to explore. Next, I’d recommend building some predictive models—like linear regression or tree-based models—to see how well we can estimate sale prices. It would also be useful to try more advanced feature engineering, validate the models with cross-validation, and look for non-linear relationships that might improve predictions even further.