Exploratory Data Analysis Report
House Prices Dataset

1 Distribution of the Target Variable (SalePrice)

The target variable, `SalePrice`, exhibited a pronounced right-skewed distribution. Since this skewness violates the linear model assumptions of normality, a logarithmic transformation (`np.log1p`) was applied to normalize the data. The transformation resulted in a distribution closer to Gaussian, improving model suitability.

1.1 Statistical Summary

The following tables summarize the descriptive statistics of `SalePrice` before and after log transformation. Post-transformation, the mean and median values are much closer, confirming the normalization effect.

(a) Original SalePrice

Statistic	Value
count	1,460.00
mean	180,921.20
std	79,442.50
min	34,900.00
25%	129,975.00
50%	163,000.00
75%	214,000.00
max	755,000.00

(b) Log-Transformed SalePrice

Statistic	Value
count	1460.0000
mean	12.0241
std	0.3994
min	10.4603
25%	11.7751
50%	12.0015
75%	12.2737
max	13.5345

Figure 1: Statistical Summary of `SalePrice` Before and After Log Transformation

1.2 Visualizations

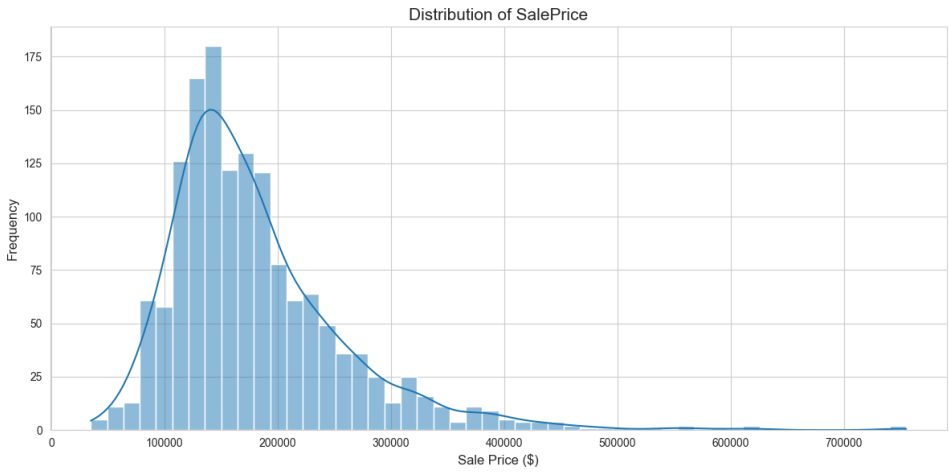

Figure 2: Distribution of the original SalePrice (right-skewed).

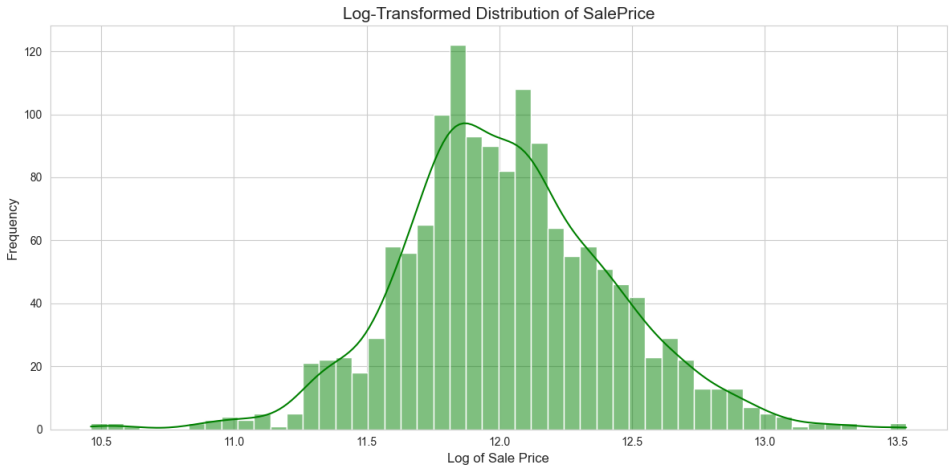

Figure 3: Distribution after log-transformation (approximately normal).

2 Missing Values and Feature Types

2.1 Missing Values

The dataset contains several features with significant missingness. For example, features like `PoolQC`, `Alley`, and `Fence` have more than 80% missing values, typically indicating the absence of those features rather than data collection errors. These will be treated as categorical features with "None" or similar placeholders during preprocessing.

Table 1: Top Features with Missing Values

Feature	**Total Missing**	**Percent Missing**
PoolQC	1453	99.52%
MiscFeature	1406	96.30%
Alley	1369	93.77%
Fence	1179	80.75%
MasVnrType	872	59.73%
FireplaceQu	690	47.26%
LotFrontage	259	17.74%
Garage-related Features	81	5.55%
Basement-related Features	37–38	2.5–2.6%
Electrical	1	0.07%

2.2 Feature Types

The dataset consists of both numerical and categorical variables, requiring distinct pre-processing pipelines.

Table 2: Summary of Feature Types

Feature Type	**Count**
Total Features	81
Numerical	38
Categorical	43

3 Correlation Analysis

A correlation heatmap identified the features most strongly correlated with `SalePrice`. `OverallQual`, `GrLivArea`, and `GarageCars` were the top predictors, indicating that overall material quality and living area are strong determinants of housing prices.

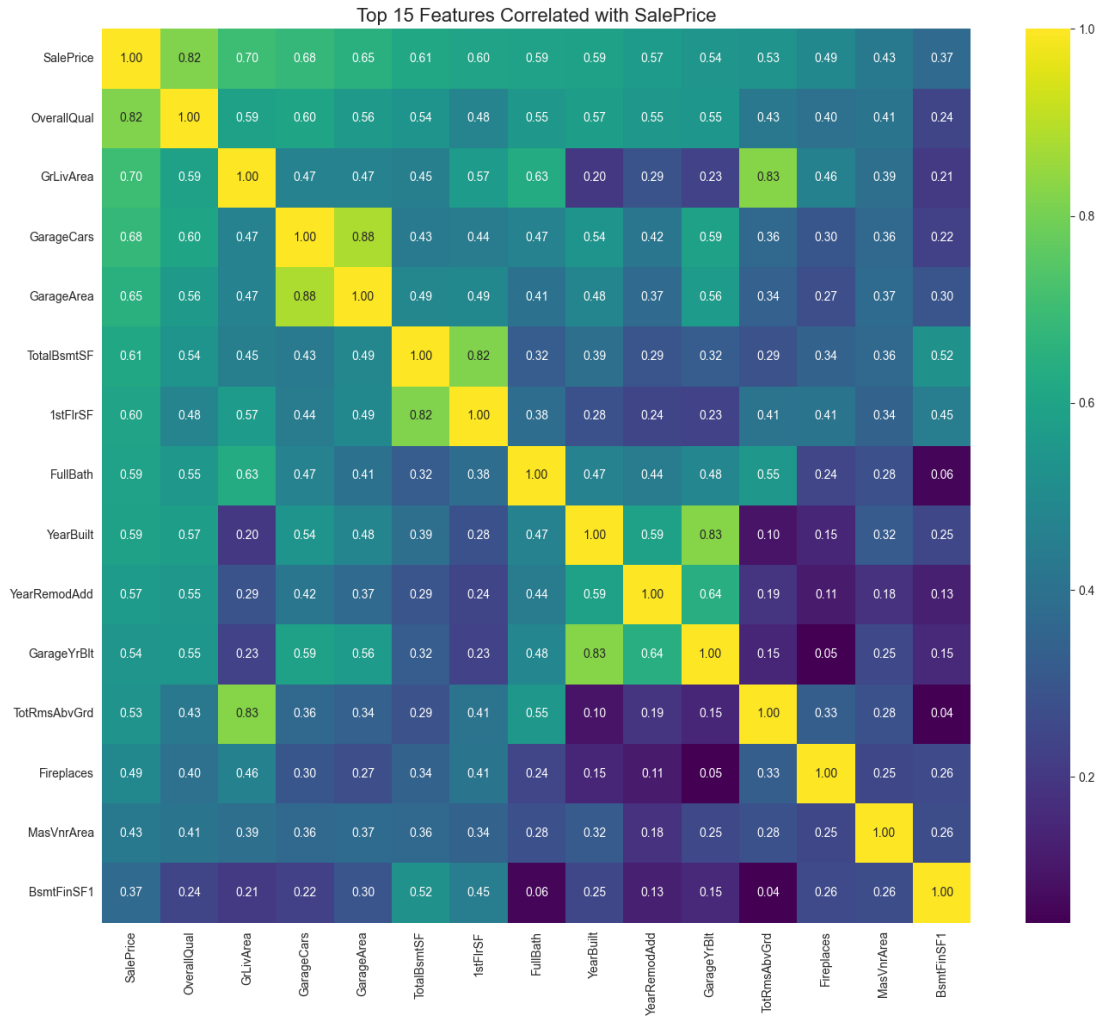

Figure 4: Correlation matrix of the top 15 numerical features.

4 Outlier Analysis

Outliers were detected through boxplots and scatterplots. The relationship between `OverallQual` and `SalePrice` was positive, as expected. However, a few points in the `GrLivArea` vs. `SalePrice` plot indicated potential outliers — large living areas with disproportionately low sale prices. These will be addressed in the preprocessing phase.

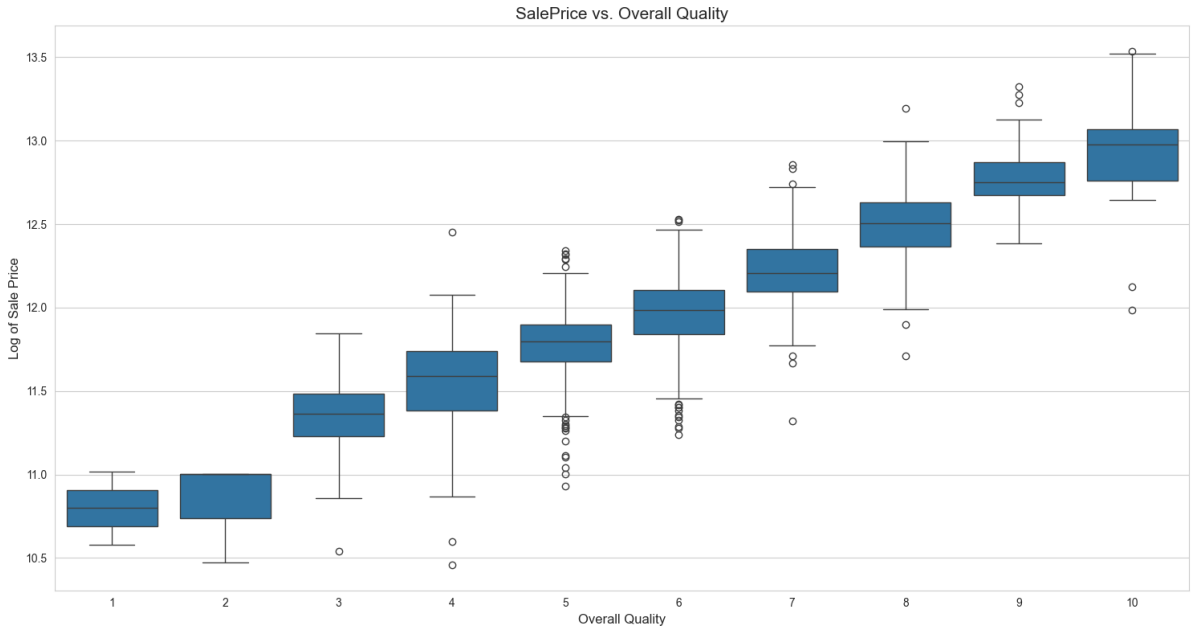

Figure 5: SalePrice vs. Overall Quality.

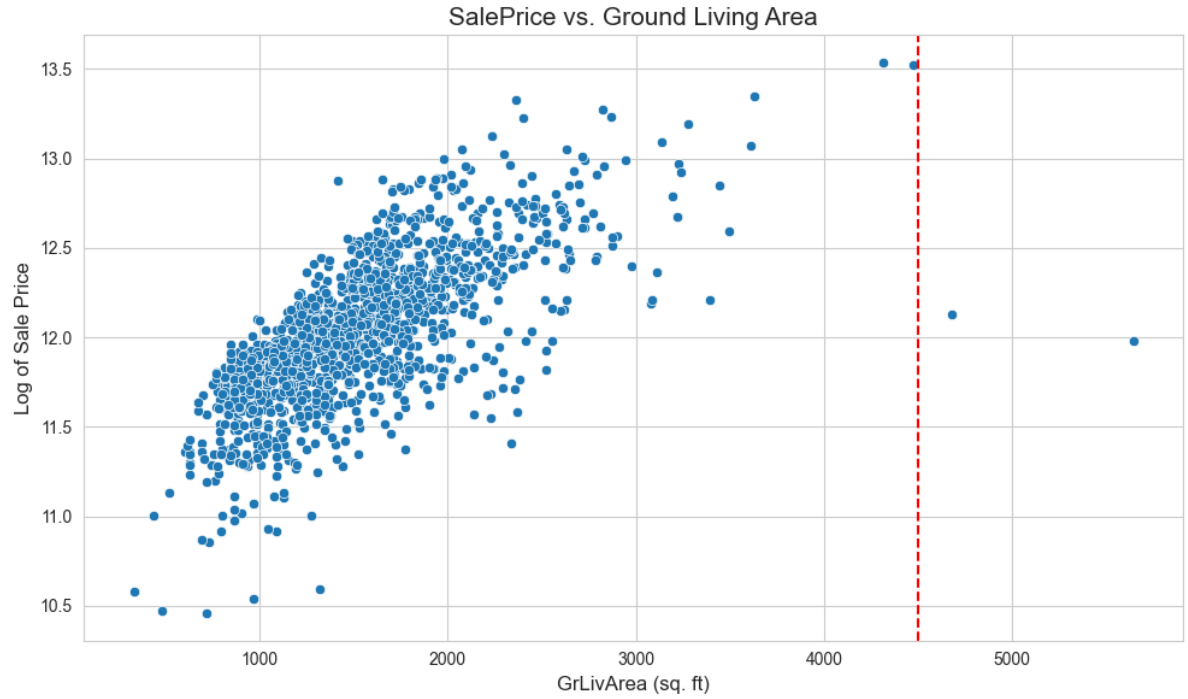

Figure 6: SalePrice vs. Ground Living Area (Outlier Detection).