Data Quality and Preprocessing Report
House Price Prediction Project

1 Introduction

The raw dataset, while comprehensive, contained several issues that would negatively impact the performance of regression models. These issues included a significant number of missing values, a skewed target variable (`SalePrice`), the presence of outliers, and a mix of numerical and categorical data types. This report provides a systematic overview of the pipeline developed to address these challenges.

2 Verification of Data Cleaning

The following table provides a high-level summary of the dataset's state before and after the preprocessing pipeline was executed on the training data.

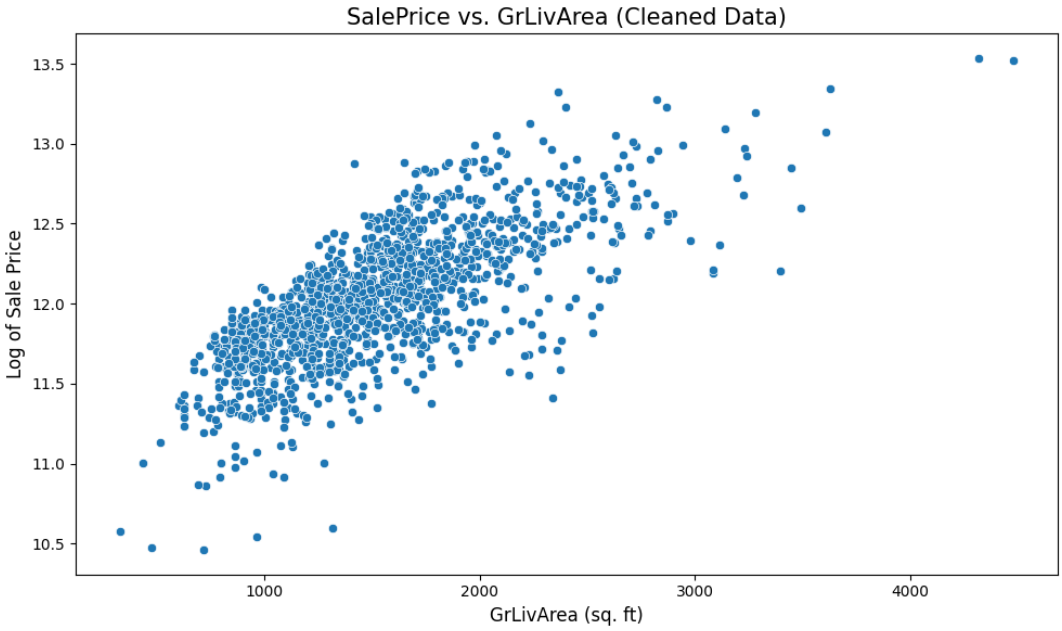
Metric	Before (Original Data)	After (Cleaned Data)
Shape	(1460, 81)	(1458, 286)
Missing Values	Present in 19 features	None
Data Types	Mixed (numerical, categorical)	All numerical
Target Variable Skew	Right-skewed	Normalized (Log-transformed)
Outliers (`GrLivArea`)	Present	Removed

3 Preprocessing Strategy

The following steps were systematically applied to both the training and test datasets to ensure consistency.

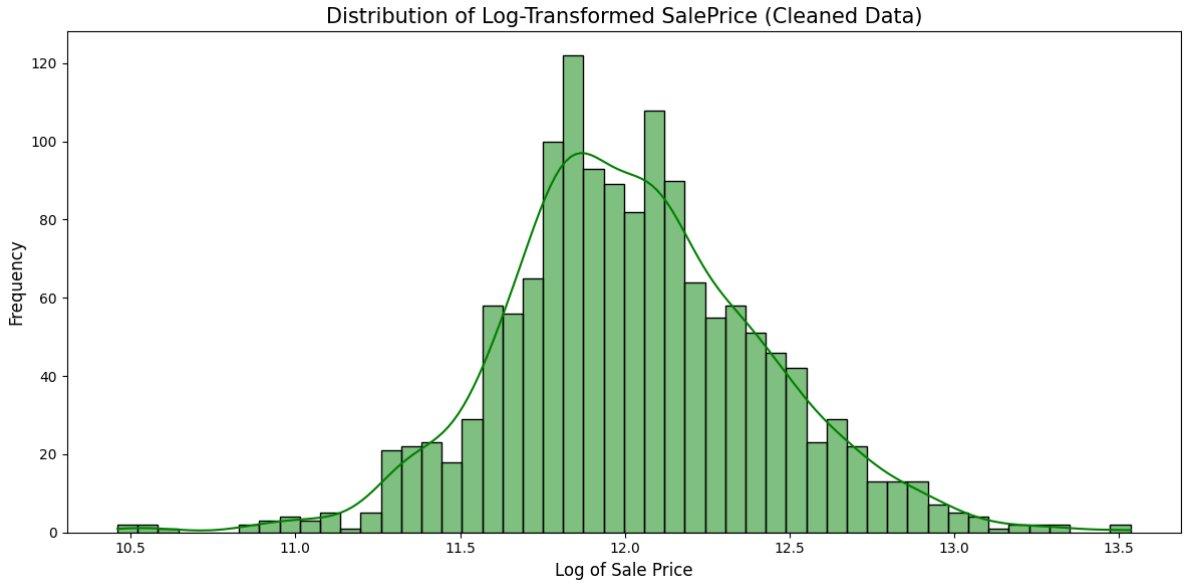
3.1 Handling Outliers

Based on the initial EDA, two outliers were identified in the `GrLivArea` feature where houses had a very large living area (> 4000 sq. ft.) but a disproportionately low `SalePrice`. These data points were removed from the training set to prevent them from negatively influencing the model.



3.2 Target Variable Transformation

The target variable, `SalePrice`, was found to be strongly right-skewed. To normalize its distribution, a logarithmic transformation (`np.log1p`) was applied. This is a standard practice that helps linear models satisfy their assumption of normality.



3.3 Missing Value Imputation

A multi-faceted strategy was employed to handle missing values, treating different features based on the meaning of their missing data as described in the data dictionary.

- **NA means "None":** For features like `PoolQC`, `Alley`, `Fence`, and `FireplaceQu`, a missing value indicates the absence of that feature. These were filled with the string literal `'None'`.
- **NA means Zero:** For numerical features related to basements and garages (e.g., `GarageArea`, `TotalBsmtSF`), a missing value implies a value of 0. These were filled with `0`.
- **Grouped Median Imputation:** For `LotFrontage`, missing values were imputed using the median `LotFrontage` of the respective `Neighborhood`. This is a more accurate approach than using the global median.
- **Mode Imputation:** For a few remaining categorical features with a small number of missing values (e.g., `Electrical`), the most frequent value (mode) was used.

3.4 Feature Engineering and Encoding

- **Type Conversion:** Some numerical features like `MSSubClass` and `YrSold` were correctly converted to strings as they represent categories, not continuous values.
- **One-Hot Encoding:** All categorical features were converted into a numerical format using one-hot encoding (`pd.get_dummies`). This process expanded the feature set from 81 to 286 columns and allows the model to interpret these variables.

4 Conclusion

The preprocessing pipeline successfully addressed all identified data quality issues. The final dataset is now clean, entirely numerical, and free of missing values and significant outliers. This processed data provides a solid and reliable foundation for the model building and training phase of the project.