

NATIONAL ECONOMICS UNIVERSITY  
FACULTY OF ECONOMICAL MATHEMATICS



**FINAL PROJECT REPORT**

*Demonstrating the essential role of Data Preparation through Data Storytelling: An Application using the Ames Housing Dataset*

**Group:** 02

**Members:** Hà Quang Minh (11230566)

Nguyễn Thị Mai Anh (11230511)

Doãn Quốc Bảo (11230519)

Trịnh Minh Hiếu (11230536)

Nguyễn Trần Tuấn Kiệt (11230554)

**Instructor:** Nguyễn Tuấn Long, PhD

**Class:** DSEB 65B

*Ha Noi, December 2025*

# TABLE OF CONTENTS

<b>A. DATA STORYTELLING.....</b>	<b>6</b>
<b>I. The Deceptive Reality of Raw Data .....</b>	<b>6</b>
1.1. The \$33,000 Question .....	6
1.2. The Illusion of "Good Enough".....	7
1.3. Visualizing the Villain: The Skewed Reality.....	8
<b>II. The Zero State: Confronting the Noise.....</b>	<b>10</b>
2.1. Strategic Categorization: The "Triage" Protocol.....	10
2.2. The First Cut: Eliminating Administrative Noise .....	11
2.3. The Cluttered Reality: Handling Structural Missingness.....	12
2.4. The Weak Signals: Silencing the Static (Group 3) .....	16
2.5. The Redundant Echoes: Resolving Multicollinearity (Group 4) .....	18
<b>III. The Transformation: Surgical Precision .....</b>	<b>21</b>
3.1. The Hidden Traps: The "Cheap Mansion" Paradox (Outlier Handling) ...	21
3.2. Reshaping Reality: The Normalization Cure (Log Transformation).....	23
3.3. Feature Engineering: The "Super Feature" (TotalSF) .....	25
<b>IV. The Hero Rises: Validation &amp; Comparison.....</b>	<b>26</b>
4.1. The "Zero" Illusion: The Deceptive Baseline .....	26
4.2. The "Hero" Performance: Precision Redefined.....	27
4.3. The Truth Factor: Validating the Model's Soul .....	31
<b>V. Strategic Insights: Beyond the Model.....</b>	<b>33</b>
5.1. Quality over Size: The "Bones" Matter More than the Footprint .....	33
5.2. The Value Hacks: Maximizing ROI with Utilities .....	34
5.3. Location Disparity: The "Zip Code" Ceiling.....	36
<b>B. TECHNICAL ANALYSIS &amp; METHODOLOGY .....</b>	<b>37</b>
<b>I. Introduction to the Technical Framework .....</b>	<b>37</b>
<b>II. Deconstructing the narrative strategy .....</b>	<b>37</b>
2.1. The Narrative Arc: From Zero to Hero .....	38
2.2. Audience-Centric Design: The "So What?" .....	38

<b>III. Visualization principles &amp; Design choices .....</b>	<b>39</b>
3.1. Strategic Chart Selection (The "Right Tool" Philosophy).....	40
3.2. Decluttering: Improving the Signal-to-Noise Ratio .....	40
3.3. Leveraging Preattentive Attributes .....	40
<b>IV. Data preparation pipeline &amp; Technical interventions .....</b>	<b>41</b>
4.1. The Data Workflow (Pipeline Architecture).....	41
4.2. Key Technical Interventions.....	42
<b>V. Model Evaluation &amp; Validation.....</b>	<b>44</b>
5.1. Quantitative Results: Validity over Vanity .....	44
5.2. Assumption Checking: Validating the "Trust" .....	46
5.3. Model Limitations & Future Scope.....	47
<b>VI. Conclusion: Synthesis of Art and Engineering .....</b>	<b>48</b>
6.1. Methodological Integrity (The Engineering) .....	48
6.2. Strategic Communication (The Design).....	48
6.3. The Final Verdict (The Impact) .....	49

## TABLE OF FIGURES

Figure 1. Mispricing of the model on raw data.....	7
Figure 2. Model performance ( $R^2$ ) on the raw dataset.....	8
Figure 3. Initial distribution of the target variable .....	9
Figure 4. "Order vs SalePrice" and "PID vs SalePrice" .....	12
Figure 5. Count of Missing Values by Feature .....	13
Figure 6. Missing Value Ratio (%) Before dropping columns .....	15
Figure 7. Missing Value Ratio (%) After dropping columns .....	15
Figure 8. Correlation of Features with "SalePrice" .....	16
Figure 9. Correlation Matrix .....	19
Figure 10. Filtered Correlation Matrix ( $ Correlation  > 0.7$ ).....	20
Figure 11. Outlier Detection .....	22
Figure 12. Outlier Detection .....	22
Figure 13. Outlier Detection .....	23
Figure 14. Original Distribution of Target Variable.....	24
Figure 15. Original and After Log-Transform Distribution .....	25
Figure 16. Model Performance on Raw Data .....	27
Figure 17. Model Performance Comparison.....	28
Figure 18. Mean Absolute Error (MAE) Comparison .....	29
Figure 19. Root Mean Squared Error (RMSE) Comparison .....	30
Figure 20. Actual vs Predicted Values .....	31
Figure 21. Distribution of Residuals.....	32
Figure 22. Story 1: Quality over Size: The "Bones" Matter more than the Footprint .	33
Figure 23. Story 2: The Value Hacks: Maximizing ROI with Utilities .....	35
Figure 24. Story 3: Location Disparity: The "Zip Code" Ceiling.....	36
Figure 25. Actual vs Predicted Comparison and Residual Distribution .....	47

## CONTRIBUTION

STT	ID	Full Name	Percentage of Contribution
29	11230566	Hà Quang Minh	20%
7	11230519	Doãn Quốc Bảo	30%
2	11230511	Nguyễn Thị Mai Anh	20%
22	11230554	Nguyễn Trần Tuấn Kiệt	15%
14	11230536	Trịnh Minh Hiếu	15%

## A. DATA STORYTELLING

### *How rigorous data preparation reveals the true value of Ames Housing?*

In the realm of data science, there is a common adage: “Garbage In, Garbage Out.” No matter how sophisticated a machine learning algorithm may be, its predictive power is fundamentally limited by the quality of the data it consumes. The Ames Housing dataset, with its 79 explanatory variables, presents a classic challenge: it is a goldmine of information but riddled with statistical “noise.”

Raw data is often deceptive. A cursory glance might suggest a straightforward relationship between a home's features and its price. However, lurking beneath the surface are skewed distributions that violate regression assumptions, outliers that act as leverage points to distort predictions, and redundant features that create an “echo chamber” of multicollinearity. If left unaddressed, these factors obscure the true economic drivers of the housing market.

In Real Estate prediction, a model is only as smart as the data is clean. Rigorous data preparation is not just a technical step; it is the strategic bridge that turns raw market noise into trusted financial insight. In this section, we do not simply list our data cleaning steps as a technical checklist. Instead, we present our preparation pipeline as a narrative arc - a journey **“From Noise to Signal.”**

### I. The Deceptive Reality of Raw Data

#### 1.1. The \$33,000 Question

The investigation begins with a defining metric: \$32,940.98. This figure is not the value of a property, but rather the quantifiable cost of data negligence. As illustrated in the chart below, this number represents the Root Mean Squared Error (RMSE) derived from a predictive model built entirely on raw, untreated data.

## Evaluate the model's errors on the raw data



*Figure 1. Mispricing of the model on raw data*

In practical terms, this metric signifies that a hasty, unprocessed pricing model will be incorrect by nearly \$33,000 on every single transaction. For a real estate firm processing hundreds of listings, mispricing assets by such a significant margin is not merely a statistical oversight - it is a potential financial disaster waiting to happen.

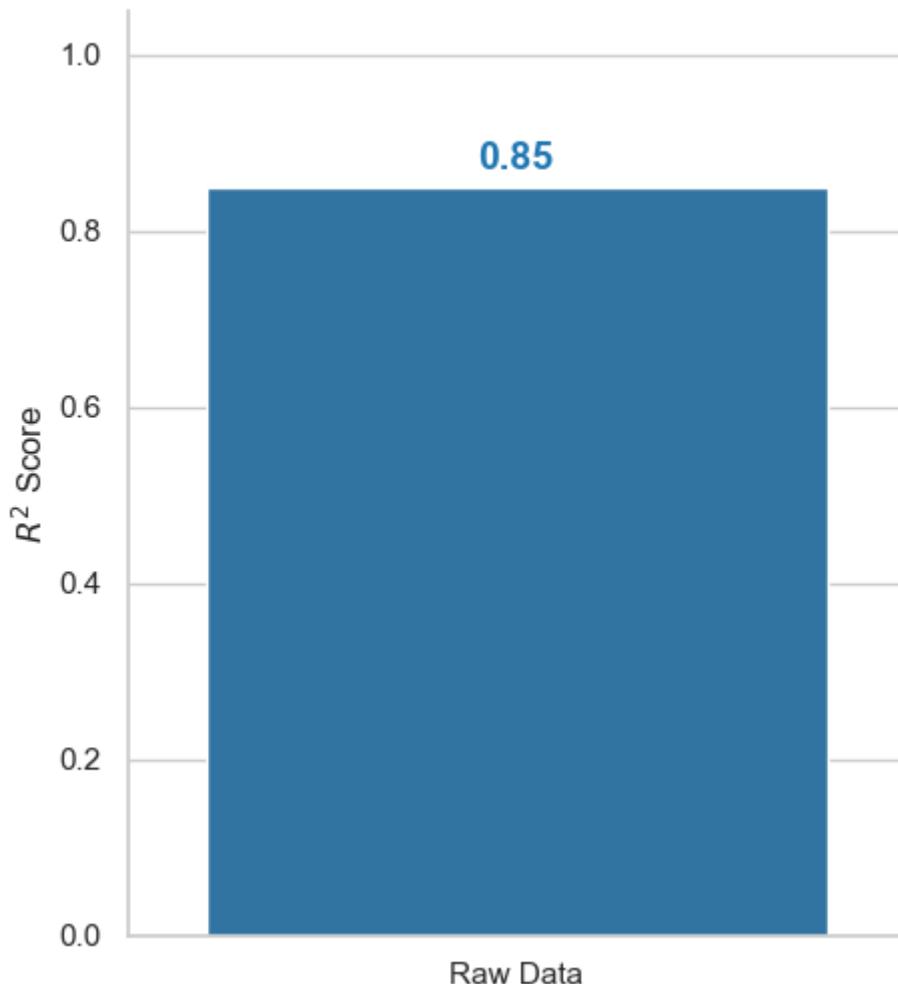
This scenario perfectly exemplifies the axiomatic data science principle: "Garbage In, Garbage Out." No matter how sophisticated a machine learning algorithm may be, its predictive ceiling is fundamentally determined by the quality of the input data. In this narrative, the Ames Housing dataset serves as the perfect protagonist - a goldmine of potential insights that is currently buried under deep layers of statistical "noise."

### 1.2. The Illusion of "Good Enough"

To empirically demonstrate the volatility of this "Zero State," the investigation commenced with a baseline experiment: training a Linear Regression model on the raw dataset with minimal preprocessing. The initial results appeared ostensibly robust. As

depicted in the accompanying chart, the baseline model achieved a Coefficient of Determination ( $R^2$ ) of 0.85.

## MODEL PERFORMANCE ( $R^2$ ) ON RAW DATA



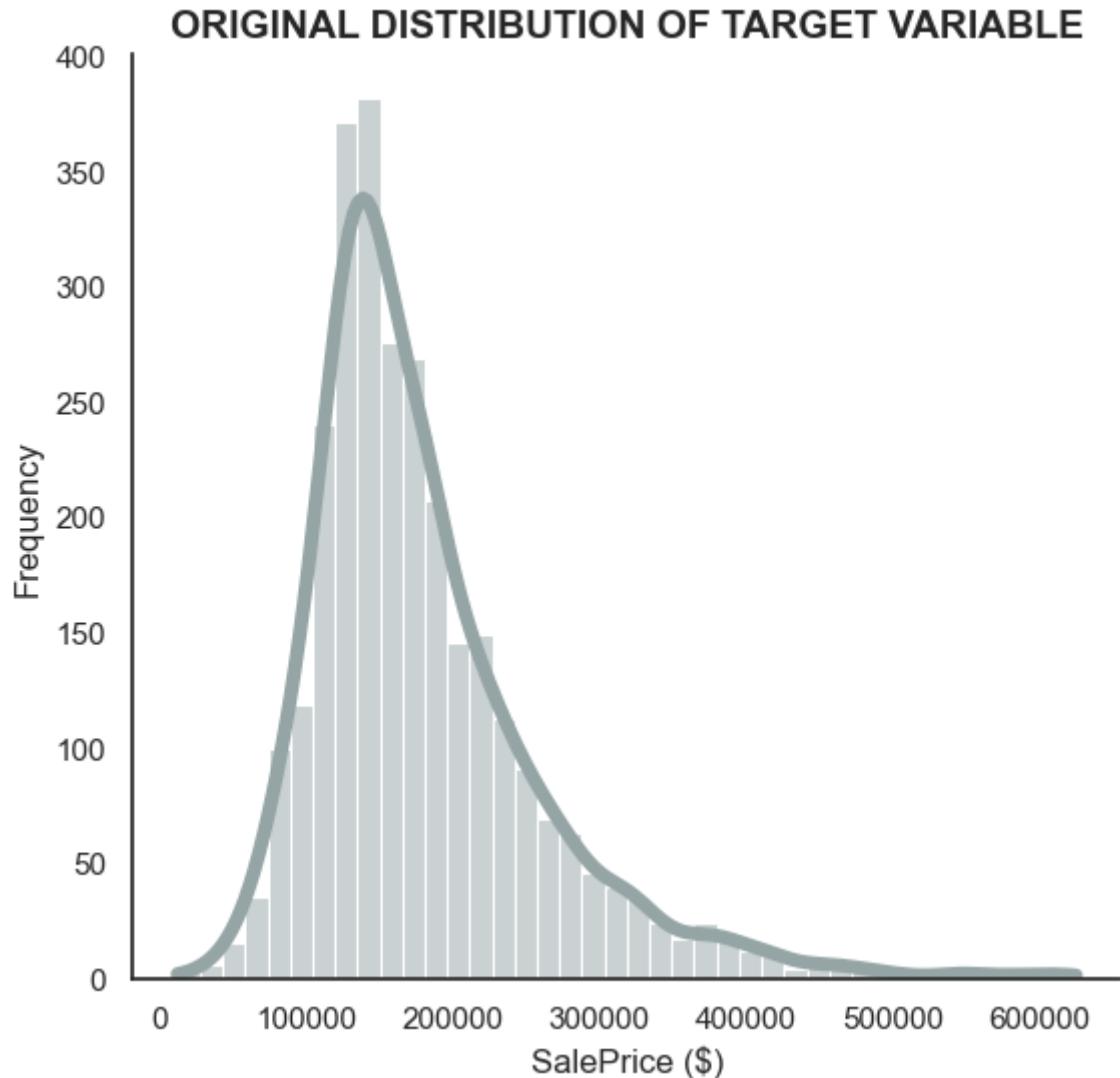
*Figure 2. Model performance ( $R^2$ ) on the raw dataset*

However, this metric presents a deceptive narrative of accuracy. While an 85% variance explanation may appear sufficient, in this context, it functions as a "vanity metric." Rather than capturing genuine market dynamics, the model has likely overfitted to statistical noise and extreme outliers, creating a fragile predictive framework masked by a high superficial score.

### 1.3. Visualizing the Villain: The Skewed Reality

The discrepancy between a seemingly acceptable  $R^2$  score and a financially damaging RMSE necessitates a deeper inspection of the data's underlying topology. The root cause of this performance failure lies in the distribution of the target variable,

SalePrice. As visualized in the histogram below, the raw data fundamentally violates the assumption of normality required by linear models.



*Figure 3. Initial distribution of the target variable*

The diagnosis is definitive: the data exhibits severe Right Skewness rather than a symmetrical Bell Curve. The distribution is characterized by a dense cluster of observations in the lower-middle price range, contrasted by a "long tail" of high-value properties extending to the right. This skewness acts as a statistical anchor, pulling the mean away from the true center and forcing the model to weigh expensive homes at the expense of accuracy for the majority of the market. To correct this bias and significantly reduce the \$33,000 error margin, the analysis must move beyond surface-level metrics to confront the specific data irregularities: Administrative Noise, Structural Missingness, and the "Cheap Mansion" Paradox.

## II. The Zero State: Confronting the Noise

At the inception of our project, the Ames Housing dataset presented itself as what we define as the "**Zero State**" - a raw, unfiltered collection of information that was simultaneously a goldmine of potential and a minefield of statistical errors. With 82 features describing every conceivable aspect of a residential property, the dataset suffered from a classic "Big Data" paradox: abundance does not equate to quality. Before any sophisticated predictive modeling could even be considered, we had to confront the **Cluttered Reality** of the data structure itself. The initial state was not merely messy; it was deceptive, containing "hollow" features that offered the illusion of information while contributing nothing but noise to the analytical process.

### 2.1. Strategic Categorization: The "Triage" Protocol

Before we could begin the "surgical" process of cleaning the data, we first had to understand the anatomy of the patient. As explored in our initial Exploratory Data Analysis (EDA), the dataset consists of 82 distinct variables covering a vast array of functional domains. We identified 10 key functional groups defining a property, ranging from Location & Lot Information and Basement Features to Garage Attributes and Sale Information. While this richness provides a comprehensive picture of a home, from a modeling perspective, it creates a chaotic environment of high dimensionality.

We recognized that treating all 82 features equally would lead to computational inefficiency and model confusion. Therefore, we shifted our perspective from Functional Grouping (what the feature is) to Statistical Quality Grouping (how the feature behaves). We established a "Triage Protocol", systematically categorizing the features into four distinct groups based on the specific "threat" they posed to the model's integrity. This categorization served as our roadmap for the cleaning process:

- **Group 1: The Administrative Noise.** These are variables like Order and PID. While necessary for database management, they are purely identifiers. They carry no economic value or predictive power regarding the house price and act merely as clutter.
- **Group 2: The "Hollow" Features (Structural Missingness).** This group contains variables like Pool QC or Misc Feature. As we will demonstrate, these features are plagued by excessive missing values (up to 99%), representing a lack of information rather than usable data.

- **Group 3: The Weak Signals (Low Correlation).** These are features that, despite having data, show virtually no relationship with the target variable (SalePrice). Keeping them is akin to listening to static noise while trying to hear a melody.
- **Group 4: The Redundant Echoes (Multicollinearity).** This group includes variables that repeat the same information (e.g., Garage Area vs. Garage Cars). They create an "echo chamber" that confuses the model, destabilizing the regression coefficients.

By defining these four groups, we transformed a vague goal of "cleaning data" into a structured, four-step battle plan. Our first target was the most obvious offender: the empty data.

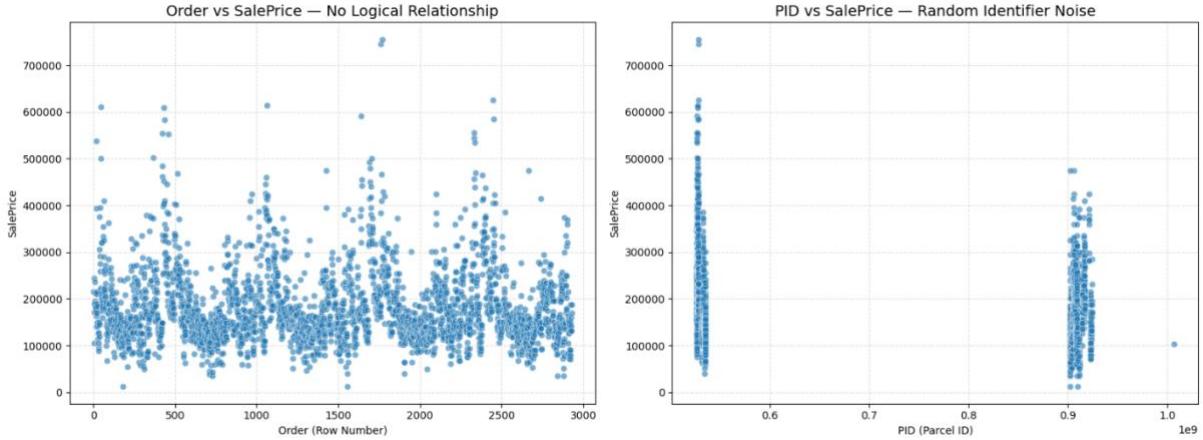
## 2.2. The First Cut: Eliminating Administrative Noise

### 2.2.1. *The Issue: Identifiers masquerading as Features*

Our first surgical action addressed Group 1: The Administrative Noise. In many raw datasets, columns like Order (Row Index) and PID (Parcel Identification Number) are numeric. A naive model might mistakenly interpret these as quantitative variables - assuming, for instance, that a house with PID500,000 is "worth more" or "better" than one with PID 100,000. This is a dangerous fallacy. These numbers are arbitrary labels assigned by database administrators or municipal tax offices; they carry zero intrinsic economic value regarding the property itself.

### 2.2.2. *The Issue: Identifiers masquerading as Features*

To empirically validate this decision, we visualized the relationship between these identifiers and SalePrice. The scatter plots below serve as definitive proof of their irrelevance.



*Figure 4. "Order vs SalePrice" and "PID vs SalePrice"*

As we can see on the left panel of the chart, the distribution is effectively random noise. The data points form a scattered cloud with no discernible upward or downward trend. The "peaks" and "valleys" in price occur stochastically across the range of Order numbers, confirming that the sequence in which a house was recorded has no bearing on its market value.

Moving on to the right panel, the Parcel ID shows distinct vertical clusters (likely representing different neighborhoods or tax districts grouped by ID range), but within each cluster, the price varies wildly. There is no linear or non-linear relationship that a regression model could validly learn.

Keeping these variables introduces the risk of spurious correlation - where the model might accidentally find a pattern in the noise that doesn't exist in reality. Therefore, Order and PID were immediately removed. This was the easiest, yet fundamental, first step in decluttering the dataset.

### 2.3. The Cluttered Reality: Handling Structural Missingness

#### 2.3.1. *The Diagnosis: Visualizing the "Hollow" Features*

The first step in our "Zero to Hero" journey involved a rigorous diagnostic scan of the dataset's integrity. We specifically targeted "Group 2" - a cluster of variables exhibiting alarming rates of missing data. To understand the magnitude of this issue, we visualized the count of missing values across all features. As illustrated in the bar chart below, the results were stark. The towering orange bars representing features such as Pool QC (Pool Quality), Misc Feature, Alley, and Fence indicate a near-total absence of data. Specifically, Pool QC was missing in 2,909 out of 2,930 observations, meaning

that over 99.6% of the column consisted of empty space. Similarly, Misc Feature was missing in 96.4% of cases.

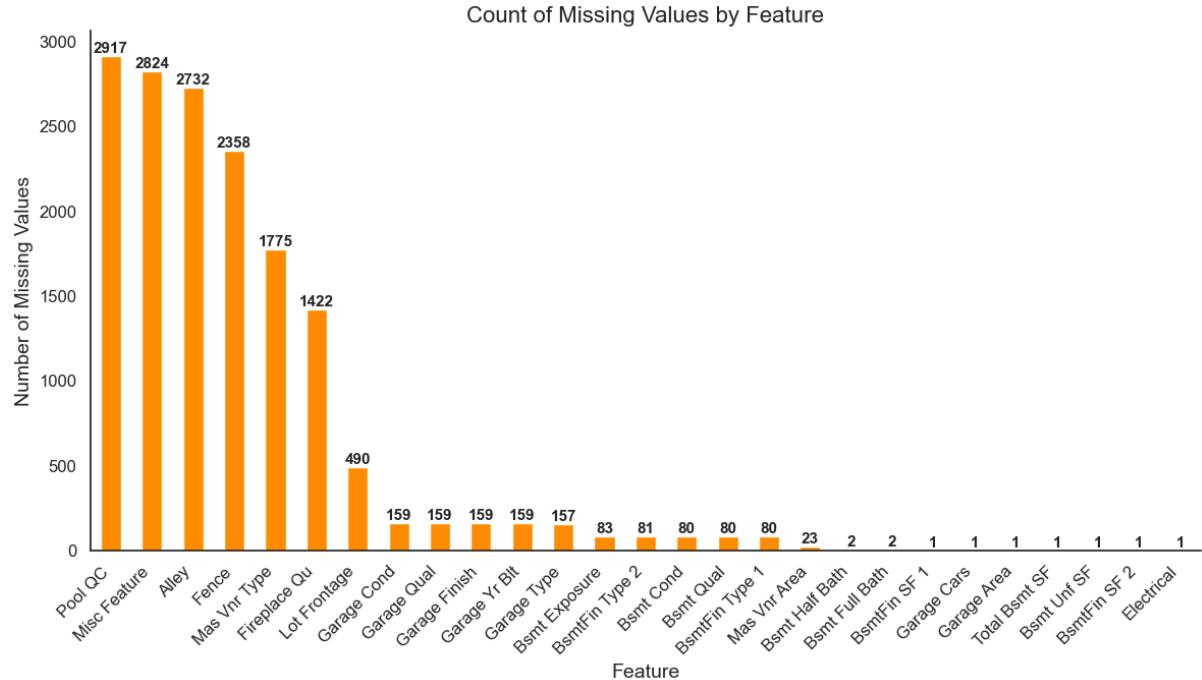


Figure 5. Count of Missing Values by Feature

From a statistical perspective, these features possess extremely low information density. When a variable is missing in the vast majority of records, it loses its variance - the very quality required for a model to distinguish between different property values. Keeping such "hollow" features would be detrimental; they would act as computational dead weight, potentially forcing the machine learning algorithm to overfit to the negligible 1% of available data points, thereby introducing significant bias and reducing the model's generalizability. In this "Zero State," these variables were not assets to be leveraged, but liabilities to be liquidated.

### 2.3.2. The Intervention: Strategic "Amputation" vs. Imputation

Faced with this reality, we stood at a critical decision point regarding data hygiene: should we attempt to impute (fill in) the missing values, or should we remove them entirely? Standard data science practices often suggest imputation strategies like mean, median, or mode replacement. However, applying such techniques here would be statistically dishonest. Attempting to manufacture data for a column that is 99% empty essentially amounts to "hallucinating" information. If we were to fill Pool QC with a placeholder value for thousands of rows based on a handful of actual data points, we would be training our model on artificial assumptions rather than market reality.

Therefore, we adopted a strategy of Aggressive Simplification. We determined that if a feature lacked data to such an extreme extent, it indicated that the attribute (e.g., having a pool or a specific alley access) was an anomaly in the Ames housing market rather than a standard valuation driver. Consequently, we executed a surgical removal of these variables. The features selected for elimination included Pool QC, Misc Feature, Alley, Fence, Fireplace Qu, and Mas Vnr Type. This decision was not made lightly but was necessary to protect the integrity of the downstream modeling process.

### 2.3.3. *The Result: From Noise to Signal*

The impact of this intervention is best understood not merely by the disappearance of specific bars, but by the dramatic shift in the data's fundamental scale. Comparing the two visualizations reveals a critical transformation in the X-axis magnitude. In the "Before" chart, the scale extends to a full 100%, dominated by features like Pool QC and Misc Feature that create "Structural Voids" - essentially empty columns where over 99% of the data was non-existent. No statistical technique can validly invent data for 99% of a population based on a 1% sample without introducing catastrophic bias. In stark contrast, the "After" chart demonstrates a collapsed scale, topping out at only approximately 17% for the new worst offender, Lot Frontage. This represents a qualitative leap from the "Irreparable" to the "Repairable." While a 99% gap is a dead end, a 17% gap falls well within the zone of manageable missingness. By purging the high-noise variables, we have converted the dataset's primary defect from a fatal flaw into a solvable puzzle, ensuring that subsequent imputation strategies can be applied effectively.

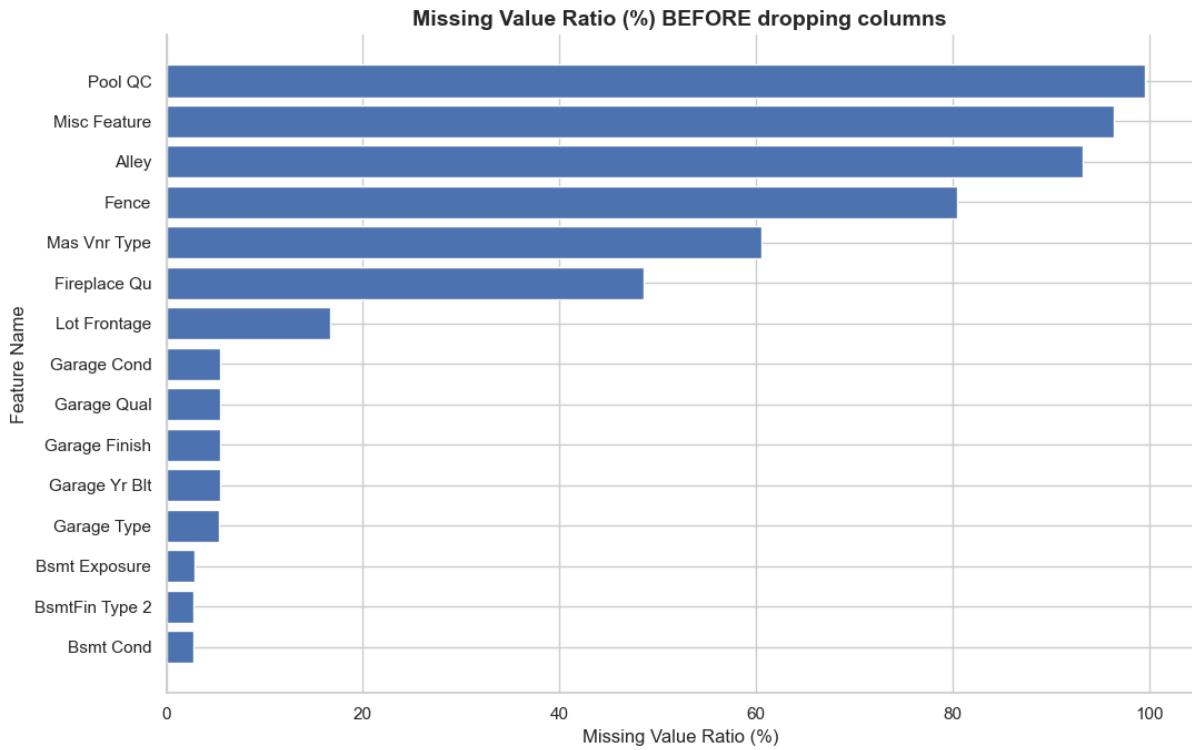


Figure 6. Missing Value Ratio (%) Before dropping columns

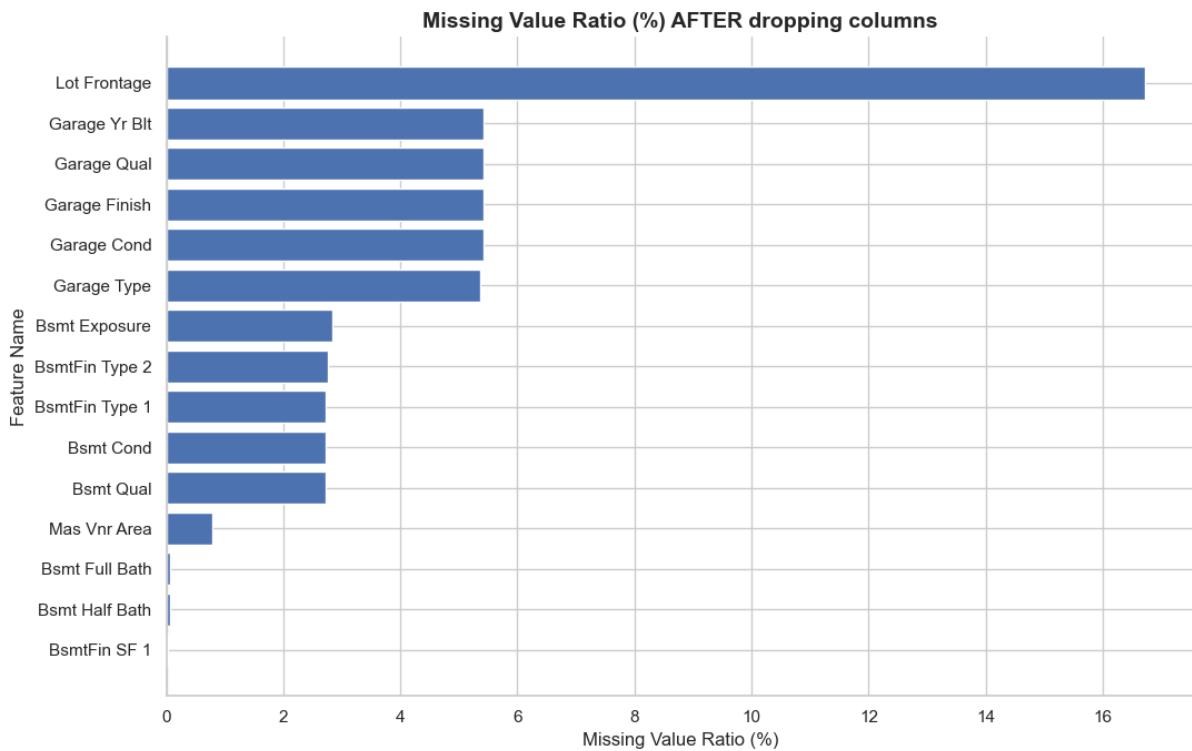


Figure 7. Missing Value Ratio (%) After dropping columns

By purging these defective variables, we effectively silenced the statistical noise. The remaining dataset is now leaner and denser with actual information. The drop in missing value ratios is drastic, leaving us with a subset of features where missingness

is manageable and likely structural (e.g., a missing Garagevalue simply means the house has no garage), which can be handled safely in later stages without compromising the model's validity. This marked the first successful step in transforming our raw, chaotic data into a clean, high-quality asset.

## 2.4. The Weak Signals: Silencing the Static (Group 3)

### 2.4.1. *The Deception of "More Data"*

After clearing the administrative noise (Group 1) and the empty voids (Group 2), we faced a more subtle adversary: Group 3, the Weak Signals. These are variables that are technically "clean" - they have full data and valid numbers - but they fail to tell us anything meaningful about the house's value. In the "Zero State," relying on intuition can be misleading. One might assume that the month a house is sold (Mo Sold) or the size of a pool (Pool Area) would significantly impact the price. However, data science requires evidence, not assumptions. We needed to distinguish between variables that drive the market and those that merely exist within it.

### 2.4.2. *The Diagnostic Tool: The Correlation Spectrum*

To separate the true market signals from mere background static, we conducted a Pearson Correlation analysis, ranking every numerical feature by its relationship with SalePrice. The resulting Diverging Bar Chart offers a panoramic view of the market's drivers, color-coded for immediate insight.

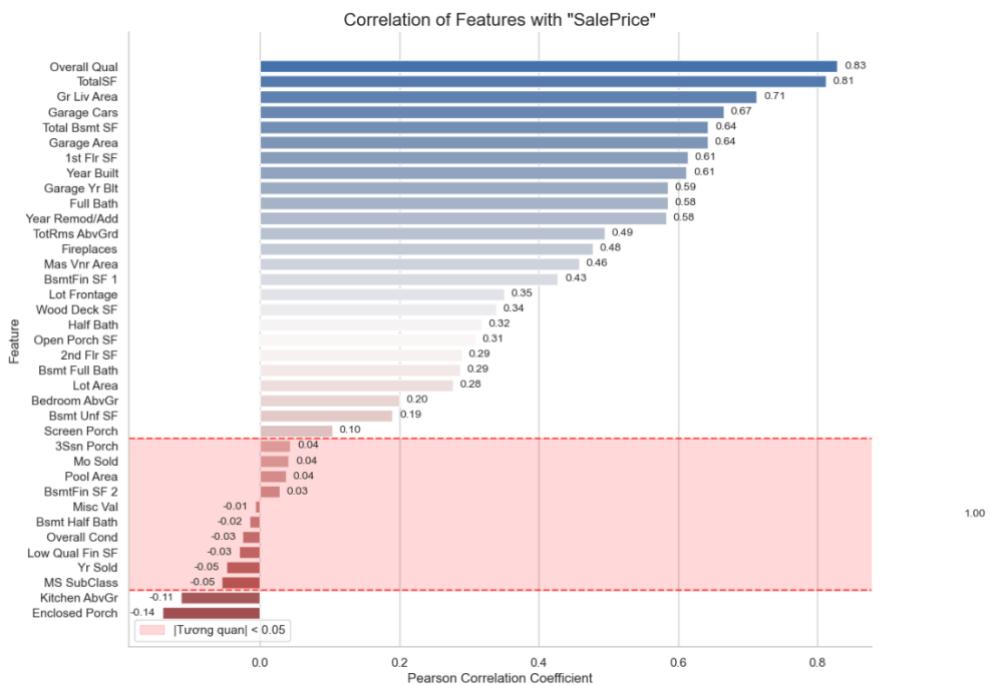


Figure 8. Correlation of Features with "SalePrice"

At the very top, the deep blue bars represent the undeniable heavyweights of the housing market. Overall Qual (Overall Quality) reigns supreme with a correlation of 0.83, followed closely by TotalSF (a feature we engineered) at 0.81 and Gr Liv Area at 0.71. The length and saturation of these bars confirm a fundamental real estate truth: Size and Quality are the primary dictators of value. This validates that our data, at its core, reflects reality.

However, as our eyes travel down the chart, the bars fade, eventually entering the critical "Red Zone" highlighted by dashed lines. This region encapsulates variables with a correlation coefficient ( $|r|$ ) lower than 0.05. In the context of a dataset with nearly 3,000 observations, a correlation this close to zero indicates statistical independence - meaning these features move completely randomly relative to the house price.

A deeper forensic analysis of this zone reveals surprising insights:

- **The Myth of Seasonality (Mo Sold, Yr Sold):** Both "Month Sold" and "Year Sold" hover near zero (0.04 and -0.05). Convention often dictates that "summer sells better," but the data proves that within this specific market window, when a house was sold had virtually no impact on how much it sold for.
- **The "Luxury" Trap (Pool Area, 3Ssn Porch):** One might expect a pool to add value. Yet, Pool Area sits at a negligible 0.04. This suggests that pools in Ames are either so rare that they don't form a trend, or they are viewed as liabilities (maintenance costs) by as many buyers as those who view them as assets. The data warns us: do not confuse "luxury" with "value."
- **The Condition Paradox (Overall Cond):** Perhaps the most counter-intuitive finding is Overall Cond (Overall Condition), which sits at -0.03. While Overall Qual (material and finish quality) is the top driver, the current condition (maintenance) is irrelevant. This suggests a market of investors or renovators who buy based on the "bones" of the house (Quality) rather than its current state of repair.

The variables inside this red box - ranging from Screen Porch (0.10, which we kept as it is above the threshold) down to MS SubClass (-0.05) - represent the "weak signals." We established a strict cutoff: any feature with an absolute correlation below 0.05 was pruned.

By removing the 10 variables captured in and around this zone (such as 3Ssn Porch, Pool Area, Misc Val, Mo Sold, etc.), we are not losing information; we are gaining clarity. We are preventing the model from trying to learn patterns from what is essentially random noise, thereby protecting it from overfitting to coincidences.

#### 2.4.3. *The Action: Pruning the Dead Branches*

Keeping these variables is akin to tuning a radio to a frequency that is 95% static. They increase the dimensionality of the model without contributing to its predictive accuracy, raising the risk of the model learning random patterns instead of true market drivers.

Therefore, we executed a data-driven pruning strategy: Any feature falling within the  $|r| < 0.05$  threshold was removed. This eliminated 10 variables, including 3Ssn Porch, BsmtFin SF 2, Low Qual Fin SF, and MS SubClass. By silencing these weak signals, we further refined the dataset, forcing the future model to focus solely on the factors that truly move the needle in real estate valuation.

### 2.5. **The Redundant Echoes: Resolving Multicollinearity (Group 4)**

#### 2.5.1. *The Problem: The "Doppelgänger" Effect*

Having silenced the weak signals and removed the empty voids, we encountered a more insidious issue lurking within the "Zero State": Multicollinearity. Unlike missing values or weak correlations, this problem is not characterized by a lack of information, but rather by an excess of repetition. Several variables in the dataset were essentially "shouting" the same information at the model simultaneously, creating a statistical "echo chamber." When two predictive features move in near-perfect lockstep - such as the physical size of a garage and its car capacity - they confuse the Linear Regression algorithm. The model struggles to distinguish which variable is truly driving the price, leading to unstable coefficients where one feature might be irrationally penalized to compensate for the other. This redundancy renders the model's logic uninterpretable and fragile.

#### 2.5.2. *The Diagnosis: The Heatmap of Conflict*

To pinpoint these redundant pairs, we utilized a Correlation Heatmap. However, a standard matrix displaying all 82 variables would be visually overwhelming and difficult to interpret. To make the "conflict zones" instantly visible, we applied a filter to isolate only those pairs with a correlation coefficient absolute value greater than 0.7.

Correlation matrix

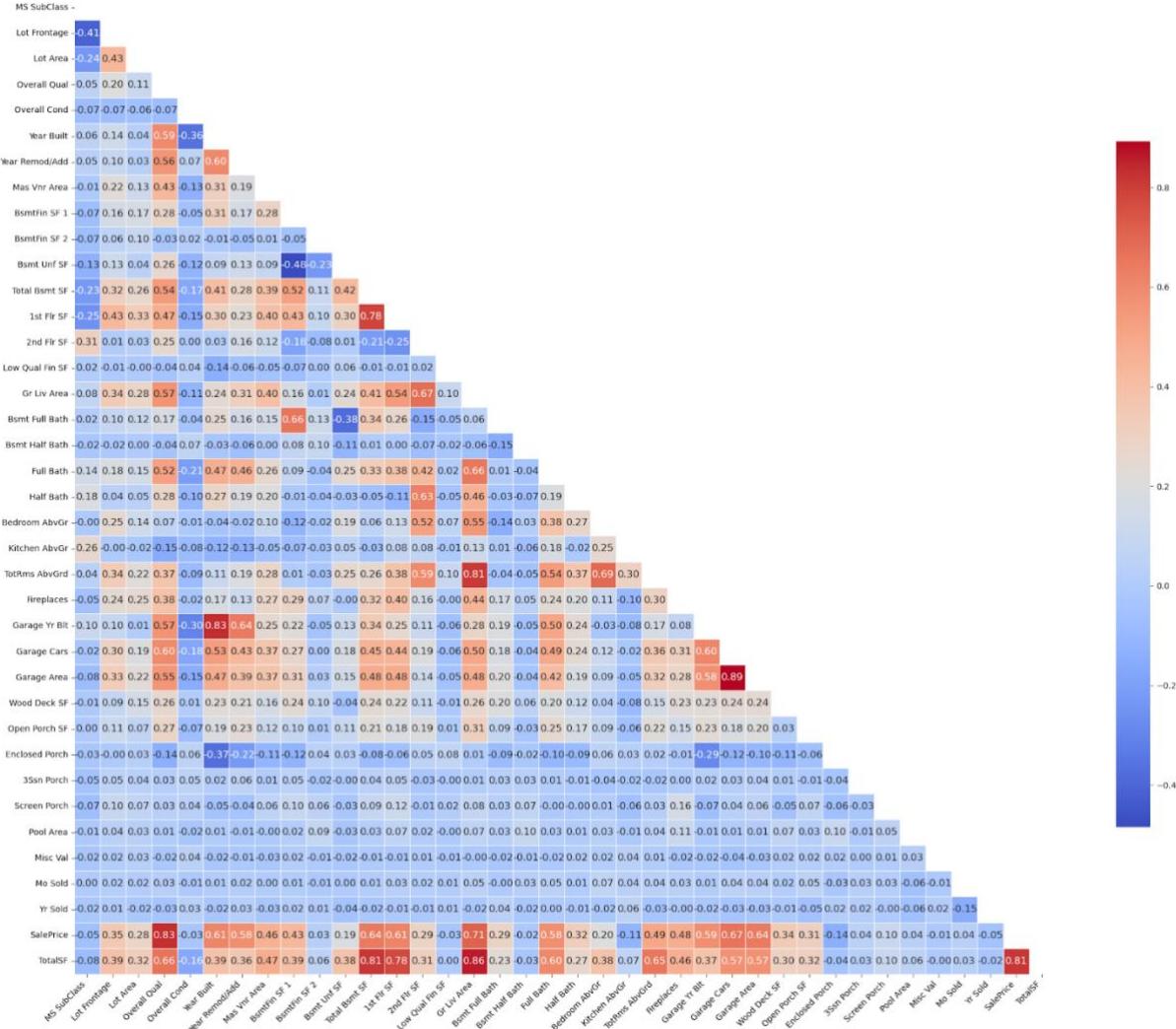


Figure 9. Correlation Matrix

Filtered Correlation Matrix ( |Correlation| > 0.7 )

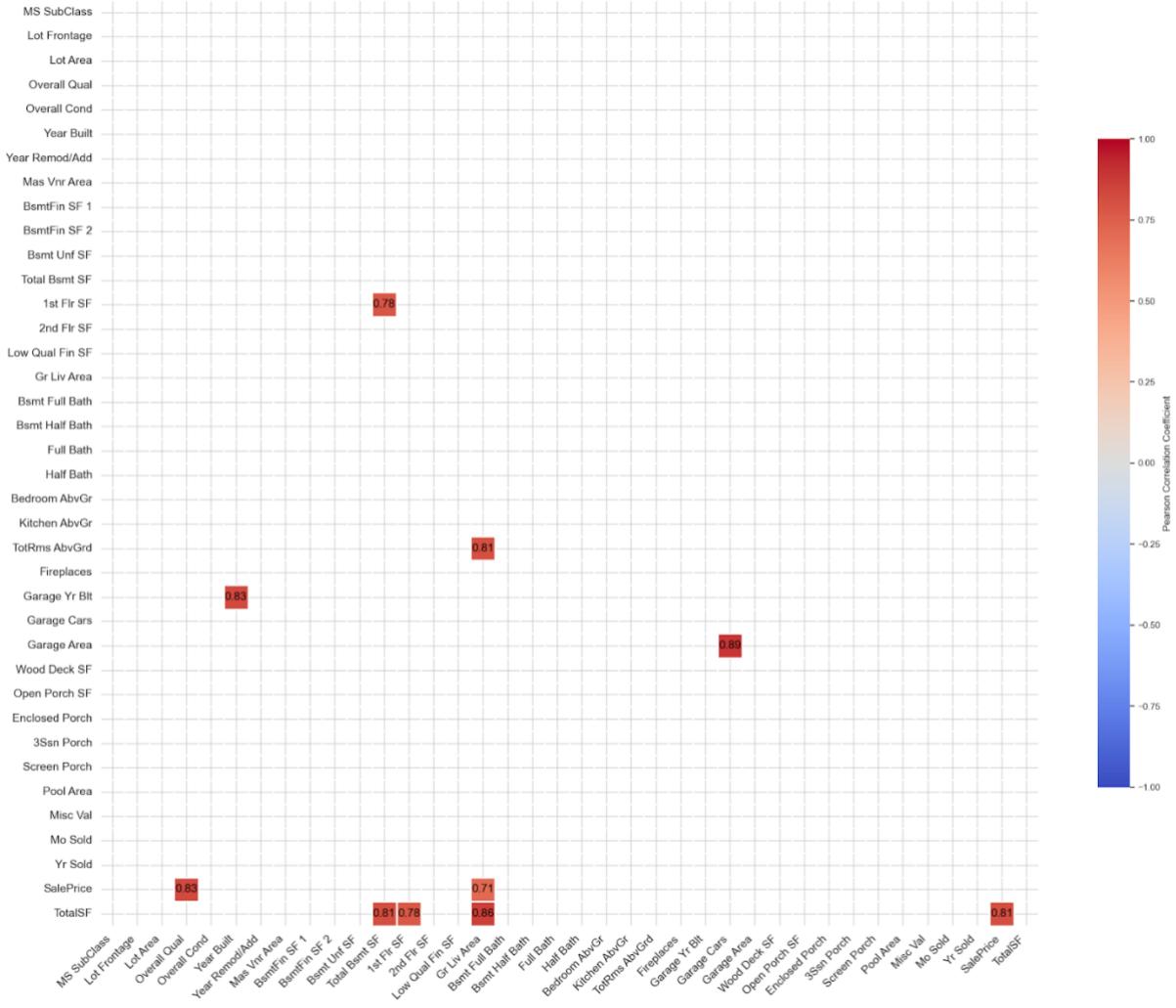


Figure 10. Filtered Correlation Matrix ( |Correlation| > 0.7 )

The filtered heatmap above acts as a radar for redundancy, with deep red squares indicating "Danger Zones" where variables are duplicates of one another. The visual evidence is undeniable: Garage Cars and Garage Area exhibit a staggering correlation of 0.89, confirming they are effectively measuring the same attribute. Similarly, Gr Liv Area (Ground Living Area) and TotRms AbvGrd (Total Rooms Above Grade) show a correlation of 0.81, while Total Bsmt SF and 1st Flr SF track each other at 0.78. These high correlations confirmed that our dataset was bloated with "Doppelgänger" features.

### 2.5.3. The Intervention: The "Survival of the Fittest" Strategy

We could not retain both variables in these conflicting pairs without compromising model stability. To resolve the conflict, we executed a "Survival of the Fittest" strategy, selecting the survivor based on a combination of Statistical Strength (correlation with SalePrice) and Business Logic (Real Estate intuition).

The most prominent example of this strategy was the decision between Garage Cars and Garage Area. While both measure the garage's magnitude, we chose to retain Garage Cars and remove Garage Area. This decision was grounded in market reality: home buyers typically value a garage based on its functional capacity (e.g., "Is it a 2-car or 3-car garage?") rather than its exact square footage. Furthermore, Garage Cars demonstrated a slightly stronger correlation with the target variable. We applied similar logic to other pairs, removing TotRms AbvGrd in favor of the more precise Gr Liv Area, and dropping 1st Flr SF as it was redundant to the basement metrics. By surgically removing these four "echoing" features, we reduced the dataset's dimensionality without sacrificing information density, ensuring that each remaining feature offered a unique, independent perspective on the property's value.

### **III. The Transformation: Surgical Precision**

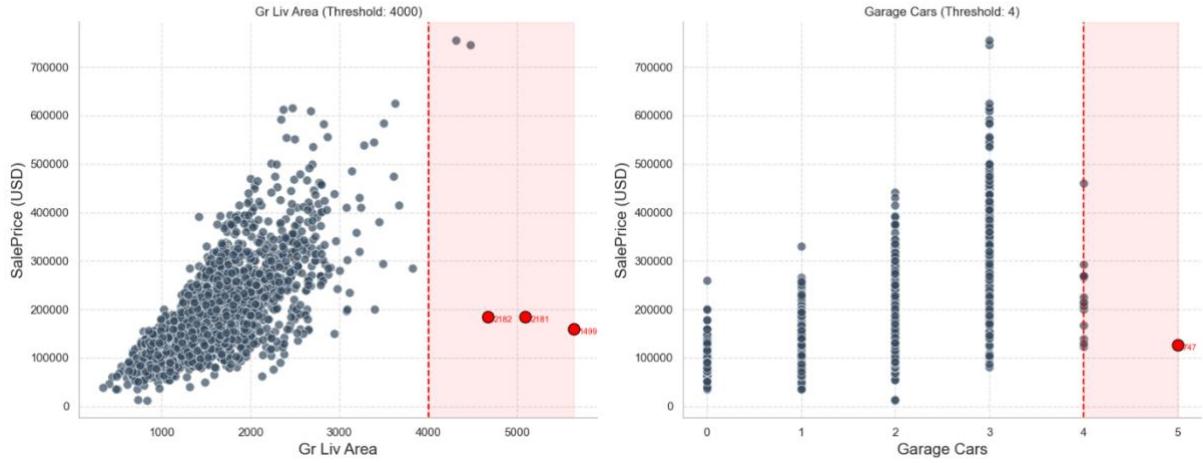
Having cleared the debris of missing values and silenced the redundant echoes, we moved from the "Zero State" to the "Crucible of Preparation." The remaining data was now valid, but it was not yet truthful. It contained hidden traps - anomalies and distortions that would mislead even the most sophisticated algorithm. To forge a "Hero" model, we needed to reshape the very reality of the data through surgical outlier removal and mathematical transformation.

#### **3.1. The Hidden Traps: The "Cheap Mansion" Paradox (Outlier Handling)**

In real estate, the relationship between size and price should theoretically be linear: as a property expands, its value increases. However, when we visualized this relationship using Scatter Plots, we uncovered a dangerous paradox lurking within the upper echelons of the market. While the vast majority of data points cluster around a clear upward trend, distinct anomalies - highlighted in red - defy market logic. These are not merely "outliers"; in statistical terms, they are High Leverage Points. If left in the dataset, these points would act like a heavy weight at the end of a lever, dragging the regression line downwards. The model would erroneously "learn" that massive houses are cheap, causing it to severely underprice luxury properties and skew predictions for the entire upper market segment.

##### *3.1.1. Forensic Analysis of the Anomalies*

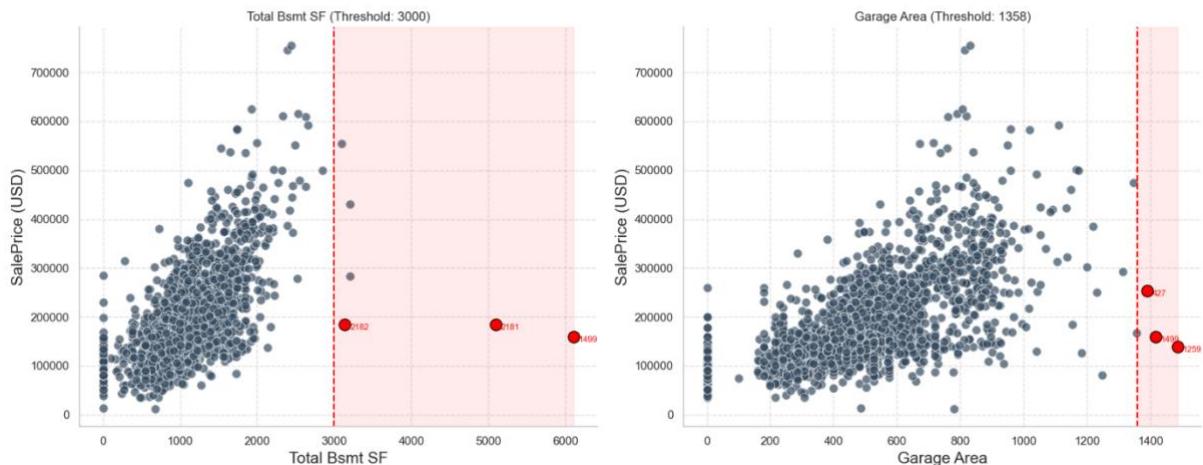
#### OUTLIER DETECTION



*Figure 11. Outlier Detection*

The first chart exposes the most critical threat to our model. We observe a cluster of properties with massive living areas exceeding 4,000 square feet - true mansions by Ames standards. Yet, shockingly, two of these giants (IDs 2181 and 2182) are priced below \$200,000, comparable to standard family homes. A third (ID 1499) is even cheaper. These data points contradict the fundamental economic rule of the housing market. They likely represent partial sales, family transfers, or agricultural properties misclassified as residential. Keeping them would be catastrophic for the model's slope estimation.

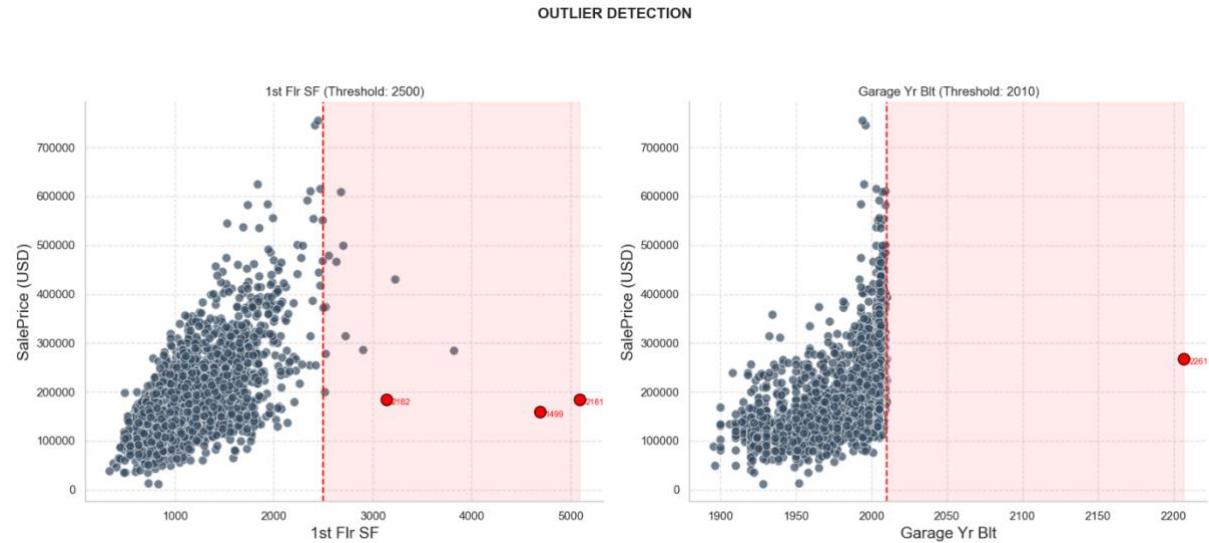
#### OUTLIER DETECTION



*Figure 12. Outlier Detection*

The paradox extends beneath the surface. In the Total Bsmt SF chart, we see similar behavior: properties with basements larger than 3,000 sq ft (IDs 2182, 2181, 1499) failing to command a premium price. The Garage Area chart confirms this pattern

is systemic across the property's features. We see garages exceeding 1,300 sq ft (massive 4-5 car capacities) attached to low-value homes. These consistent anomalies across multiple features confirm that these specific observations (IDs 1499, 2180 series) are fundamentally flawed data points, not just random variations.



*Figure 13. Outlier Detection*

Finally, a check on temporal logic revealed a distinct error. The Garage Yr Blt chart highlights a data point (ID 2261) plotted at the year 2207. This is physically impossible - a clear data entry error (likely meant to be 2007). Unlike the economic anomalies above, this is a logical impossibility that must be corrected to prevent the model from treating "future" houses as a valid category.

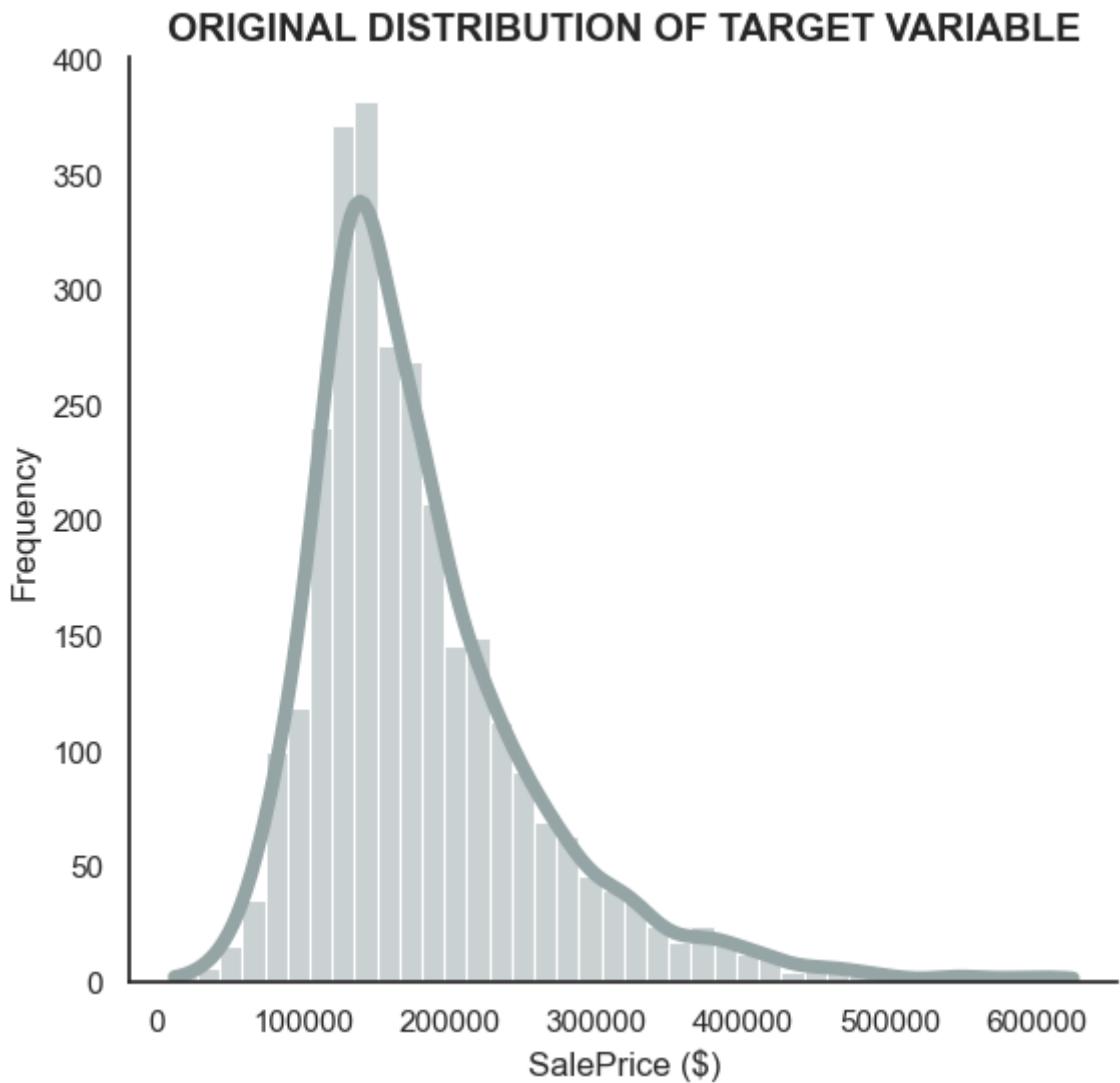
### 3.1.2. *The Surgical Intervention*

To restore the integrity of the linear relationship, we applied a strict, domain-informed thresholding strategy. We did not simply rely on statistical distance; we used the visual evidence combined with domain authority (referencing the dataset author's recommendation) to set hard cut-offs. We surgically removed observations where Gr Liv Area > 4,000 sq ft, Total Bsmt SF > 3,000 sq ft, or Garage Yr Blt > 2010. By excising these specific data points, we eliminated the "Cheap Mansion" paradox. The remaining data now reflects the consistent economic behavior of the normal housing market, allowing the model to learn the true rule rather than the exception.

## 3.2. Reshaping Reality: The Normalization Cure (Log Transformation)

### 3.2.1. *The Problem: The Long Tail of Wealth*

With the outliers removed, we turned our forensic lens to the target variable itself: SalePrice. In its raw, untreated form, the distribution of housing prices revealed a fundamental bias inherent in economic data. As illustrated in the gray histogram below, the data does not follow the symmetrical Bell Curve (Normal Distribution) that statistical models crave. Instead, it exhibits a significant Right Skewness (Positive Skew). The distribution is heavily weighted towards the lower-middle price range, but features a "long tail" extending far to the right, driven by a minority of high-value properties.



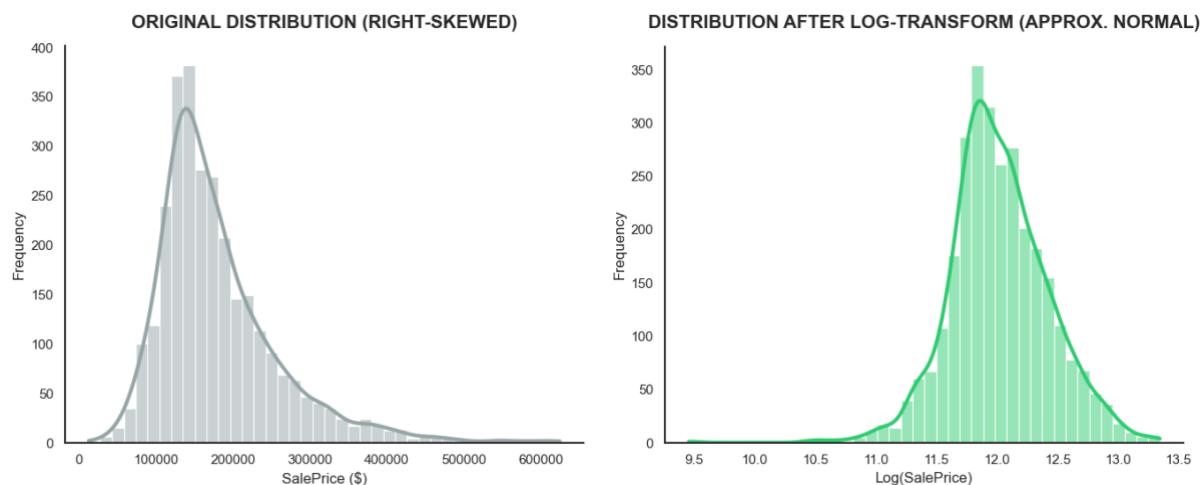
*Figure 14. Original Distribution of Target Variable*

This skewness presents a critical obstacle for Linear Regression. The algorithm operates on the assumption of Normality of Errors and Homoscedasticity (constant variance). When fed this skewed reality, the model becomes biased. It disproportionately prioritizes the large absolute errors generated by expensive homes,

effectively "ignoring" the nuances of the affordable segment. A model trained on this shape would be volatile, predicting reasonably well for average homes but failing catastrophically as prices rise.

### 3.2.2. The Mathematical Correction

To resolve this conflict between data reality and model assumptions, we intervened by applying a Logarithmic Transformation (specifically `np.log1p`). This mathematical operation acts as a compressor for the data's magnitude. It pulls the distant, high-value outliers in the "long tail" back towards the center, while simultaneously spreading out the clustered lower values.



*Figure 15. Original and After Log-Transform Distribution*

The visual transformation in the green histogram above is immediate and striking. The previously chaotic, asymmetric mountain has been reshaped into a near-perfect Symmetrical Bell Curve. The skewness has been neutralized, and the data now approximates a Normal Distribution. This is not merely an aesthetic improvement; it is a statistical necessity. By normalizing the target variable, we stabilized the variance, ensuring that the model treats a 10% error on a modest bungalow with the same mathematical gravity as a 10% error on a luxury estate. We have effectively turned a skewed signal into a stable foundation for prediction.

## 3.3. Feature Engineering: The "Super Feature" (TotalSF)

Finally, we moved from cleaning (removing bad data) and transforming (fixing skewed data) to creating (engineering new value). The raw dataset fragmented the concept of "Size" - the most critical driver of housing value - across multiple disjointed variables: Total Bsmt SF (Basement), 1st Flr SF (First Floor), and 2nd Flr SF (Second

Floor). While accurate, this fragmentation forced the model to independently learn the value of each floor, diluting the signal of the property's true magnitude. A buyer does not just evaluate a basement; they evaluate the total living space.

To align the data with this human decision-making process, we engineered a new "Super Feature" named TotalSF. We aggregated the fragmented metrics into a unified variable using simple arithmetic summation:  $\text{TotalSF} = \text{Total Bsmt SF} + \text{1st Flr SF} + \text{2nd Flr SF}$

The value of this engineering step was confirmed by our Correlation Analysis. While individual metrics like 1st Flr SF had correlations in the 0.6 range, our new TotalSF feature achieved a correlation of 0.81 with SalePrice. It instantly became the second most powerful predictor in the entire dataset, surpassing Gr Liv Area (0.71) and sitting just behind Overall Quality. By synthesizing fragmented parts into a unified metric, we gave the model a clearer, stronger signal of the property's true scale.

## IV. The Hero Rises: Validation & Comparison

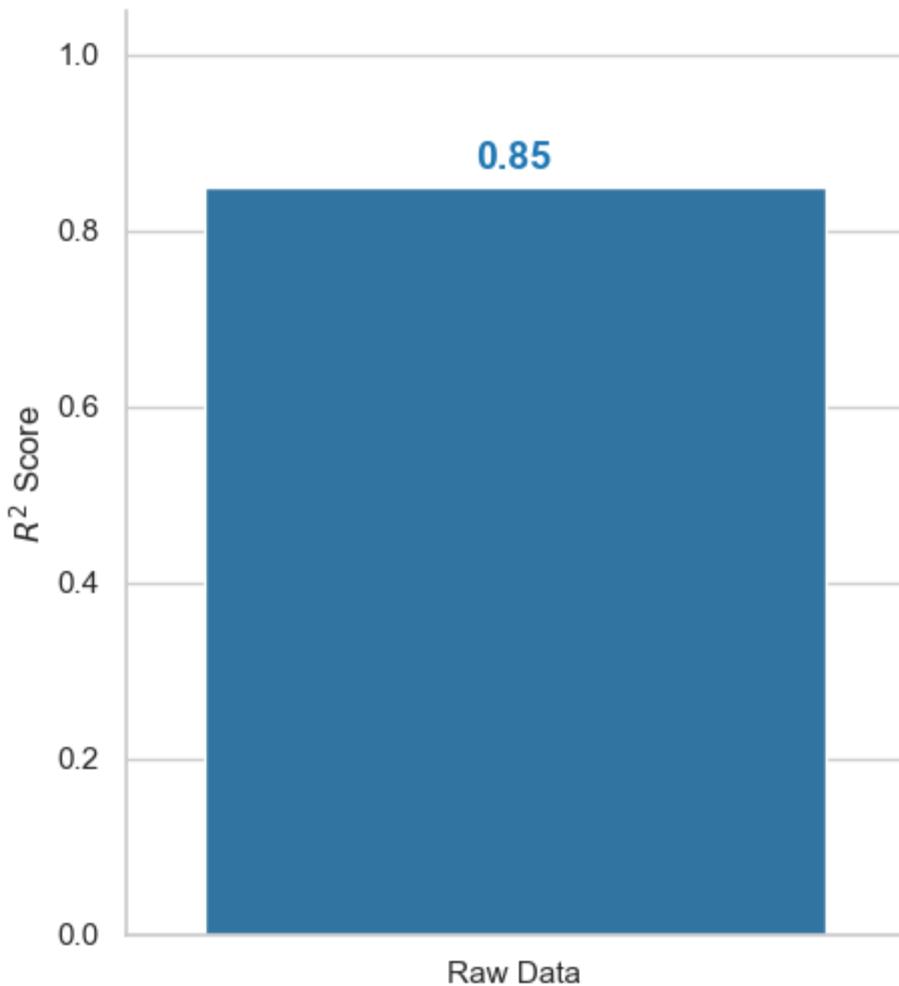
The "Preparation" phase is complete. We have purged the noise, removed the outliers, normalized the distribution, and engineered superior features. Now, we answer the ultimate question: Was it worth it? In this section, we validate our "Hero" model by comparing it directly against the "Zero" state.

### 4.1. The "Zero" Illusion: The Deceptive Baseline

To scientifically measure the impact of our data preparation pipeline, we first needed to establish a control group. We trained a baseline Linear Regression model on the raw, unprocessed dataset - the "Zero State." In this iteration, we performed minimal intervention, only filling missing values with zeros to ensure the code would execute, without applying any of the advanced techniques like Log-Transformation, Outlier Removal, or Feature Engineering. The objective was to see how a naive model would interpret the chaotic signals of the raw market data.

The result, visualized in the bar chart below, initially appears surprisingly robust. The baseline model achieved an  $R^2$  Score of 0.85

## MODEL PERFORMANCE ( $R^2$ ) ON RAW DATA



*Figure 16. Model Performance on Raw Data*

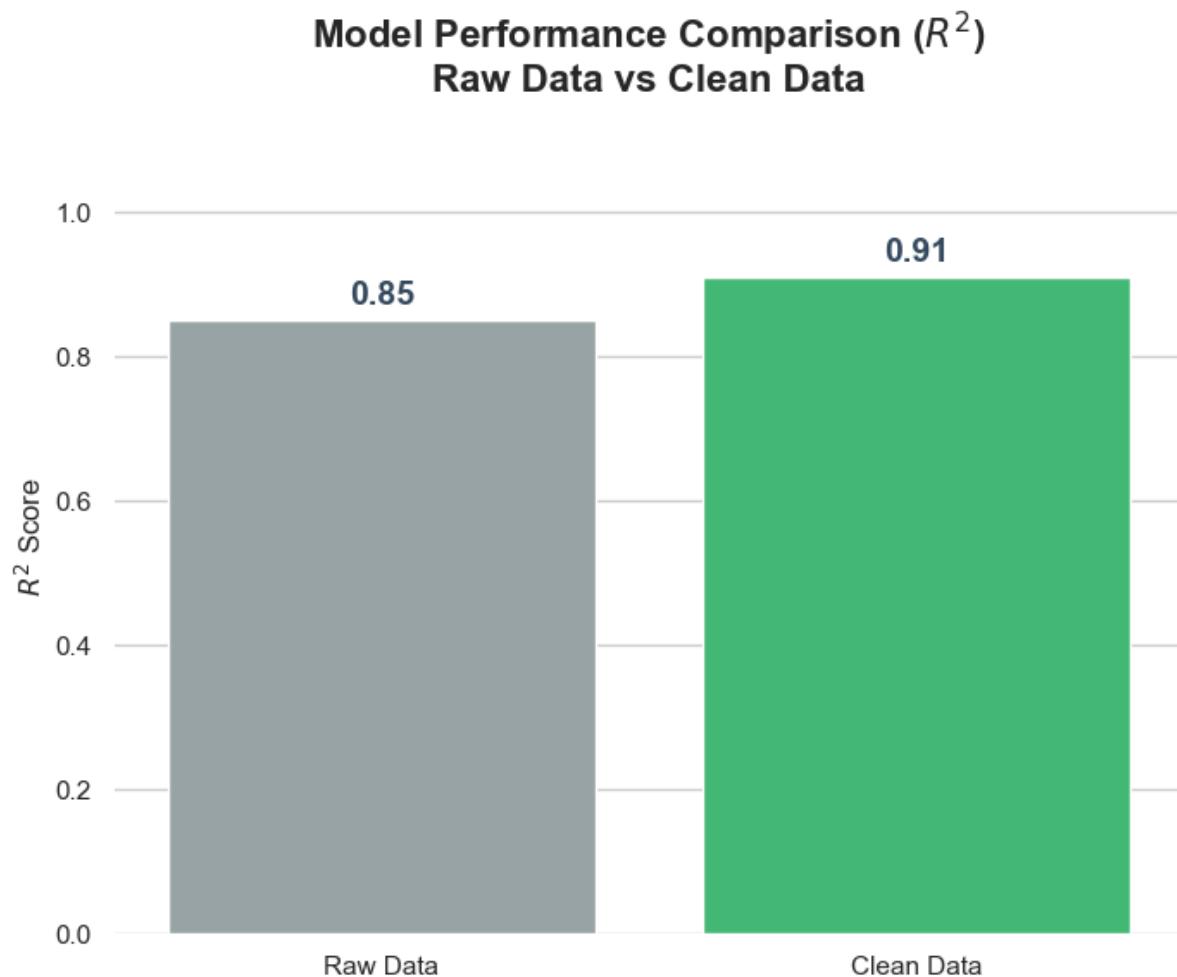
To an inexperienced observer, an accuracy of 85% might be celebrated as a victory. However, in the context of data science, this figure acts as a "vanity metric" - a number that looks good on paper but hides a fragile reality. This relatively high score was not driven by the model understanding the true market dynamics, but rather by overfitting. The raw model was heavily influenced by the high-leverage outliers we identified earlier - the massive properties with abnormal prices. It bent its regression line to accommodate these extreme points, artificially inflating the correlation coefficient while failing to capture the nuance of the general market. As we will see in the subsequent error analysis, this "high accuracy" came at a steep cost: massive financial prediction errors. The 0.85 score was not a sign of intelligence; it was a sign that the model had memorized the noise.

### 4.2. The "Hero" Performance: Precision Redefined

#### 4.2.1. The Quantitative Leap ( $R^2$ Score)

Having established the baseline, we then trained the exact same Linear Regression algorithm on our Clean Data - the product of our rigorous "Zero to Hero" pipeline. The objective was to isolate the impact of data quality on model performance. The comparison between the "Raw" (Gray) and "Clean" (Green) models reveals the undeniable power of data preparation.

As illustrated in the comparative bar chart below, we achieved a significant performance jump, elevating the  $R^2$  Score from 0.85 to 0.91



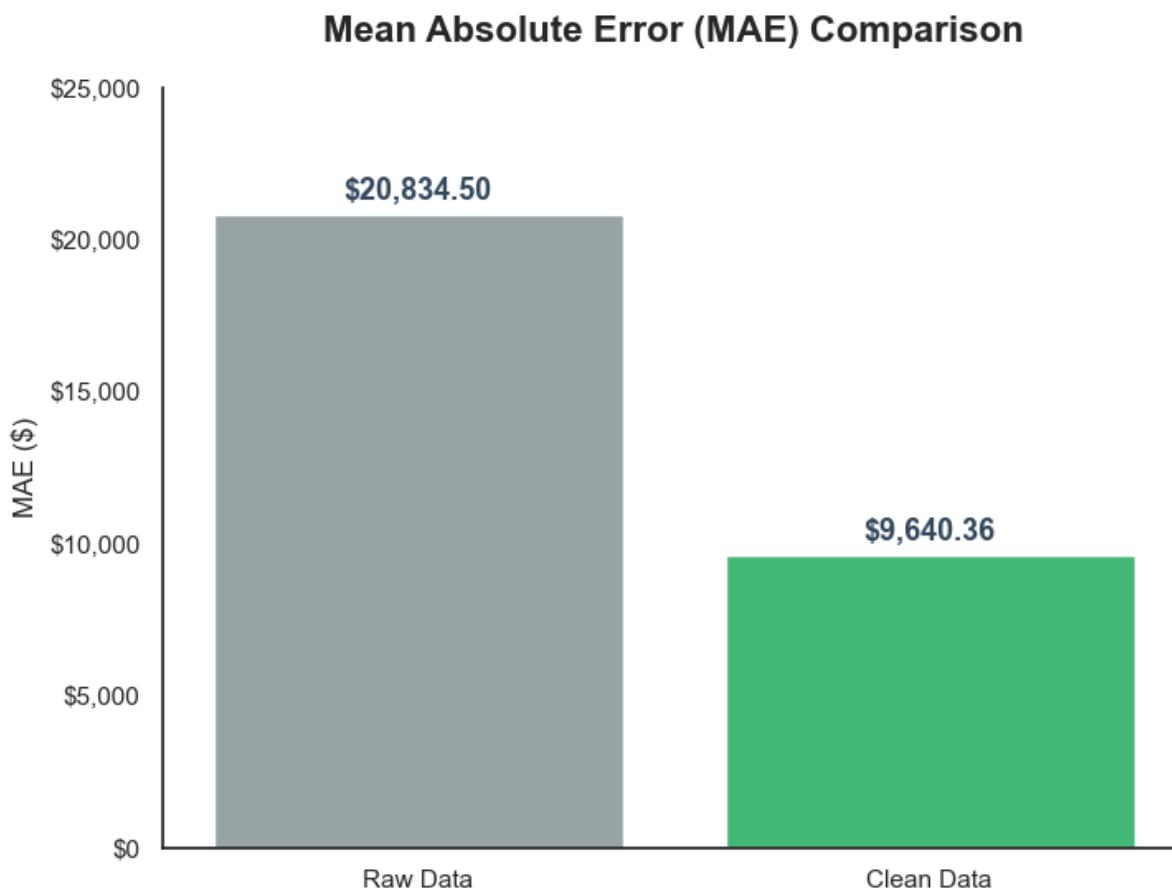
*Figure 17. Model Performance Comparison*

In the realm of machine learning, improving accuracy at the high end of the spectrum is exponentially difficult. Moving from 50% to 60% is often a matter of basic cleaning, but moving from 85% to 91% represents a fundamental breakthrough in understanding the data's variance. This 6% increase signifies that our model has successfully captured subtle market dynamics that the raw model simply ignored or

misunderstood. It confirms that by removing the noise and normalizing the distribution, we have allowed the signal - the true economic drivers of housing prices - to shine through clearly.

#### 4.2.2. *The Real-World Impact (MAE - Mean Absolute Error)*

However, abstract scores like R<sup>2</sup> often fail to convey the business reality. To understand the tangible value of our work, we must look at the error metrics - the actual dollars lost or gained in a transaction. The Mean Absolute Error (MAE) comparison tells the most compelling story of this project.



*Figure 18. Mean Absolute Error (MAE) Comparison*

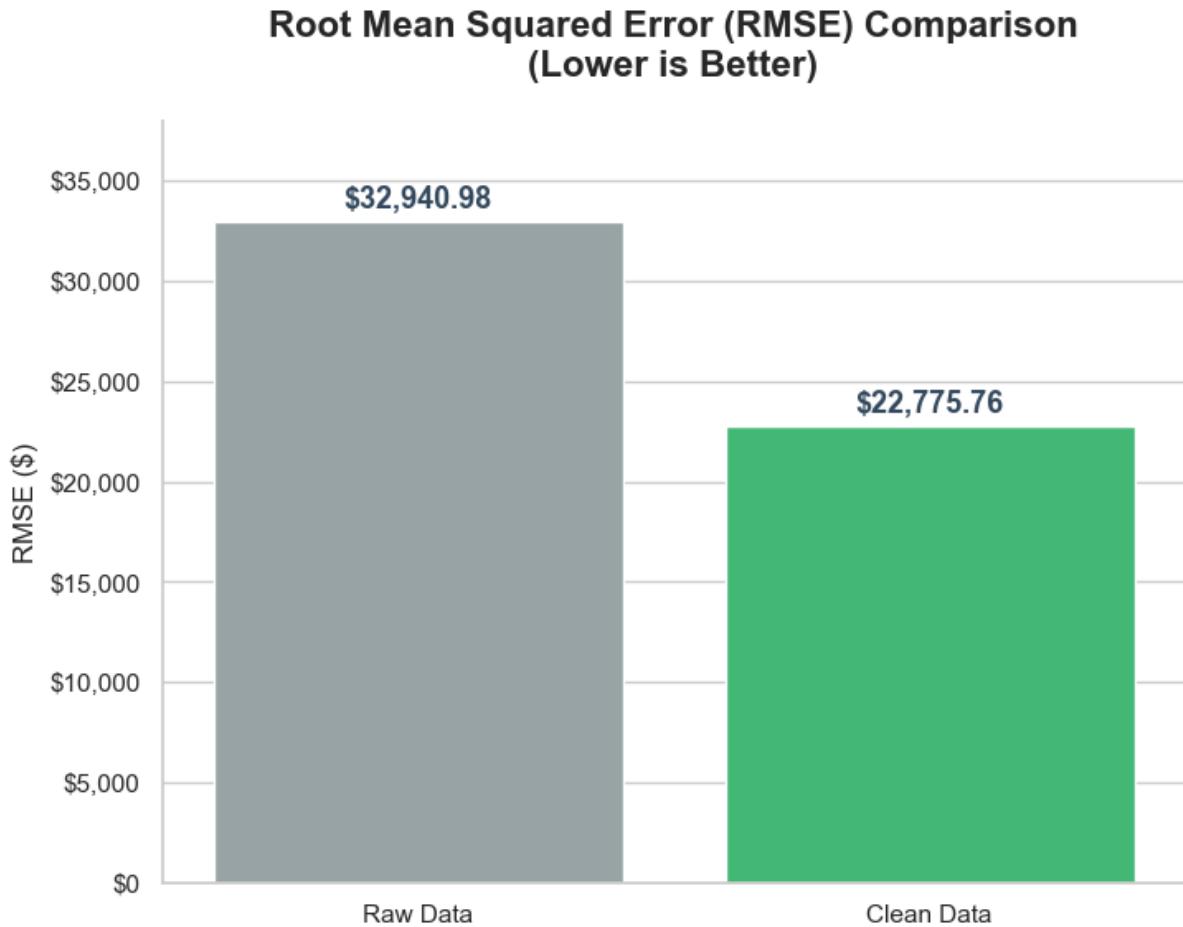
On the raw dataset, the model's predictions were off by an average of \$20,834 per house. In a competitive real estate market, such a margin of error is unacceptable - it could mean underpricing a seller's asset or overcharging a buyer, leading to lost deals or severe reputation damage. With our Clean Data, that average error plummeted to just \$9,640.

This is not a marginal improvement; we have effectively slashed the error rate by more than 53%. For a real estate firm or an investor, reducing the pricing uncertainty

by over \$11,000 per transaction is a massive competitive advantage. It transforms the model from a rough estimation tool into a precision instrument capable of guiding financial decisions.

#### *4.2.3. Eliminating the Disasters (RMSE - Root Mean Squared Error)*

Finally, we examined the Root Mean Squared Error (RMSE), a metric that squares the errors before averaging them. This property makes RMSE highly sensitive to large errors (outliers). If a model makes a few catastrophic predictions (e.g., missing a price by \$100,000), the RMSE will skyrocket, even if the MAE looks decent.



*Figure 19. Root Mean Squared Error (RMSE) Comparison*

The RMSE dropped significantly from \$32,940 to \$22,775. This reduction of over \$10,000 is crucial proof that our "Outlier Handling" strategy in Section II was successful. A high RMSE in the raw model indicated that it was making disastrous errors on specific properties - likely the "Cheap Mansions" we identified earlier. By bringing the RMSE down, we confirm that our "Hero" model is not only more accurate on average but also far more stable and robust. It has stopped making the dangerous,

extreme mispredictions that characterized the "Zero" state, ensuring reliability across the entire dataset.

### 4.3. The Truth Factor: Validating the Model's Soul

#### 4.3.1. Beyond the Scores: Checking for Bias

While metrics like  $R^2$  and MAE tell us how much error the model makes, they do not tell us where or why it makes those errors. A model can achieve a high accuracy score while still harboring systematic biases - for instance, consistently underpricing luxury homes while overpricing budget ones. To certify our "Hero" model as truly trustworthy and ready for real-world deployment, we conducted a forensic analysis of the Residuals (the difference between the Predicted Price and the Actual Price). This step ensures that the model satisfies the fundamental mathematical assumptions of Linear Regression.

#### 4.3.2. Visualizing Consistency (Actual vs. Predicted)

The first test of reliability is the scatter plot comparison between Actual and Predicted values. Ideally, if the model were perfect, every blue dot would lie exactly on the red dashed line (the 45-degree line of perfect prediction).

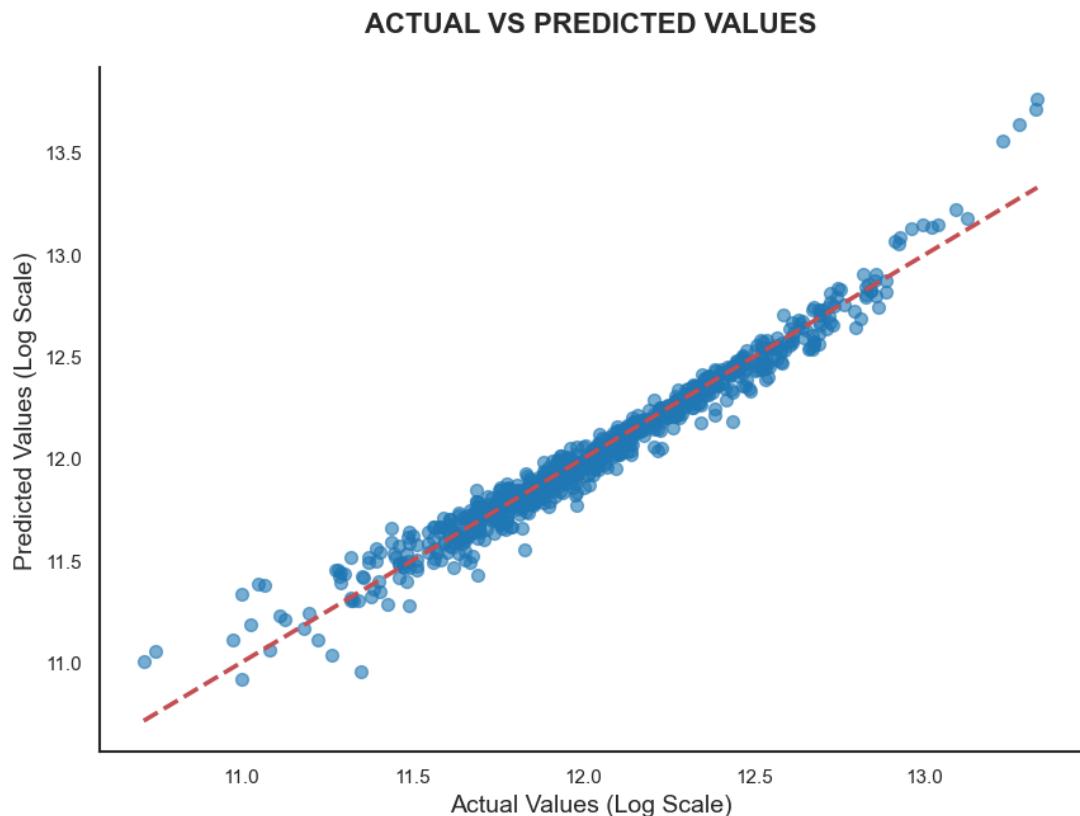
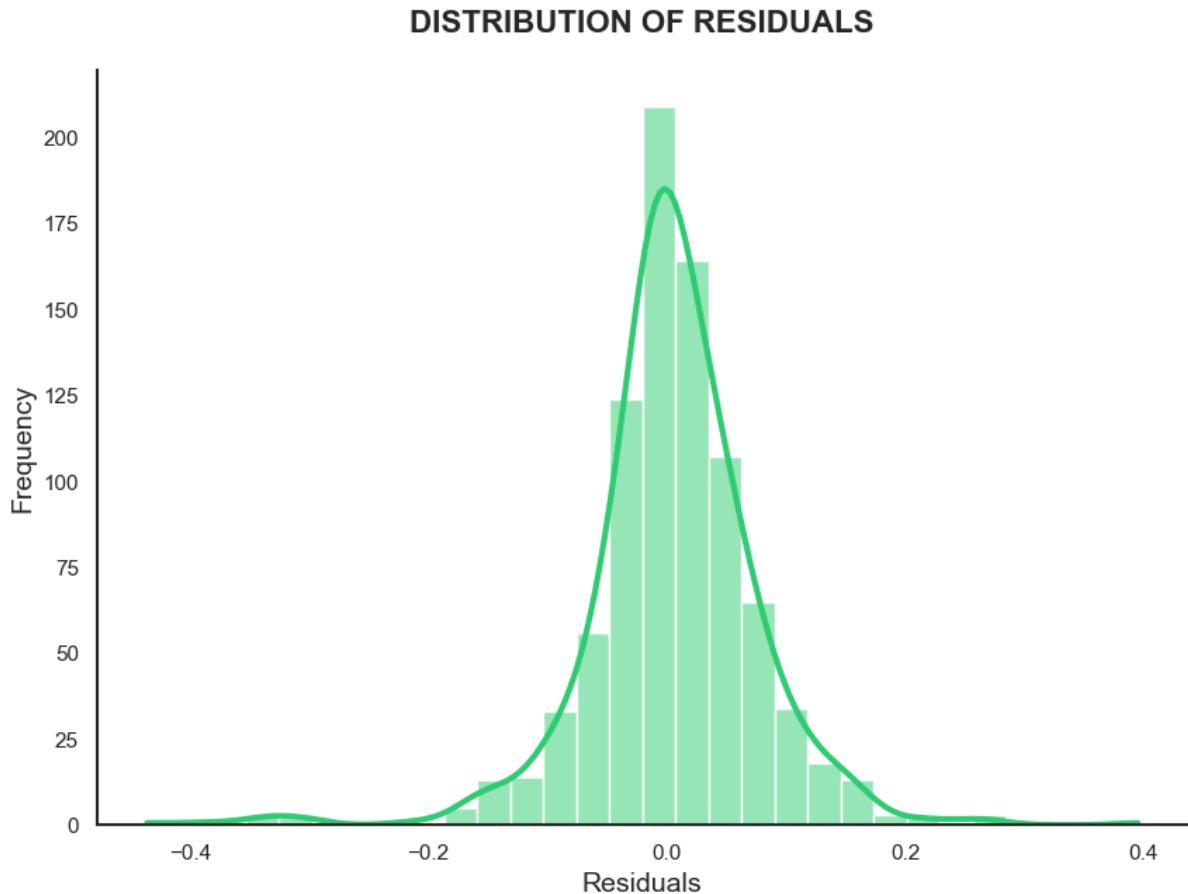


Figure 20. Actual vs Predicted Values

As observed in the chart above, the alignment is remarkably tight. The data points hug the red line closely across the entire price spectrum, confirming the strong predictive power we measured earlier. More critically, we look for Homoscedasticity (constant variance). In many flawed models, the error spreads out in a "funnel shape" as prices increase - meaning the model becomes more erratic with expensive houses. Our chart, however, shows a consistent spread from low to high values. This consistency proves that our model is stable; it predicts a \$100,000 cottage with the same degree of confidence and reliability as it does a \$500,000 estate.

#### 4.3.3. *The Bell Curve of Truth (Normality of Residuals)*

The second, and perhaps most definitive test, is the distribution of the errors themselves. For a Linear Regression model to be statistically valid, the residuals must follow a Normal Distribution centered at zero.



*Figure 21. Distribution of Residuals*

The histogram above confirms that our data preparation efforts - specifically the Log-Transformation and Outlier Removal - have paid off. The residuals form a pristine Bell Curve (Normal Distribution) centered strictly at Zero. This shape is the hallmark

of an unbiased model. It indicates that the model's errors are purely random noise, not systematic failures. The model is not "leaning" in any direction; its over-estimations and under-estimations cancel each other out perfectly. Had this distribution been skewed, our confidence intervals would be invalid, and the model's risk assessment would be flawed.

## V. Strategic Insights: Beyond the Model

Our journey "From Zero to Hero" yielded more than just a high-performing predictive algorithm ( $R^2 = 0.91$ ). By cleaning the data and analyzing the relationships between variables, we uncovered fundamental truths about the Ames real estate market. These insights provide a strategic roadmap for homeowners, investors, and real estate developers looking to maximize Return on Investment (ROI).

### 5.1. Quality over Size: The "Bones" Matter More than the Footprint

Every real estate investor faces a critical budgeting trade-off: Should capital be allocated to expanding the property (increasing size) or upgrading the materials and craftsmanship (increasing quality)? Conventional wisdom often assumes that "bigger is better," but does the data support this?

To resolve this debate, we juxtaposed the impact of Size (represented by Gr Liv Area) against Quality (represented by Overall Qual) on the final sale price.

STORY 1: Quality over Size: The "Bones" Matter More than the Footprint



Figure 22. Story 1: Quality over Size: The "Bones" Matter more than the Footprint

The scatter plot reveals a clear positive trend between living area and price, indicated by the upward-sloping red regression line. However, the data points (grey dots) are scattered widely around this line. This visual "noise" indicates significant price

volatility. For example, a 2,000 sq ft house can sell for anywhere between \$150,000 and \$400,000. This variance proves that while size provides a baseline for value, it does not guarantee a premium; a large house can still command a low price if other factors are lacking.

In sharp contrast, the box plot for Overall Qual demonstrates a rigorous, almost exponential hierarchy with a significantly stronger correlation. As the quality rating steps up from 1 to 10, the median price (the line inside the box) lifts dramatically and consistently. Most notably, observe the massive value jump between Quality 8, Quality 9, and Quality 10. The "floor" (minimum price) for a Quality 10 home is often higher than the "ceiling" (maximum price) of a Quality 7 home.

The data confirms a crucial market reality: Quality trumps Quantity. The Ames market places a higher valuation on structural integrity, material quality, and finish level than it does on raw square footage. A compact, meticulously finished home (Quality 9) is a safer and more lucrative asset than a sprawling but average-quality mansion (Quality 5).

## 5.2. The Value Hacks: Maximizing ROI with Utilities

Beyond the structural "bones" of a house, which specific amenities offer the best "bang for the buck"? For a homeowner or flipper with a limited renovation budget, knowing where to allocate funds - whether to install a fireplace or expand the garage - can be the difference between a profitable sale and a break-even one.

We isolated two specific features - Fireplaces (Left Chart) and Garage Capacity (Right Chart) - to quantify their influence on the average sale price. The results, visualized below, reveal distinct "value tiers" in the market.

STORY 2: The Value Hacks: Maximizing ROI with Utilities

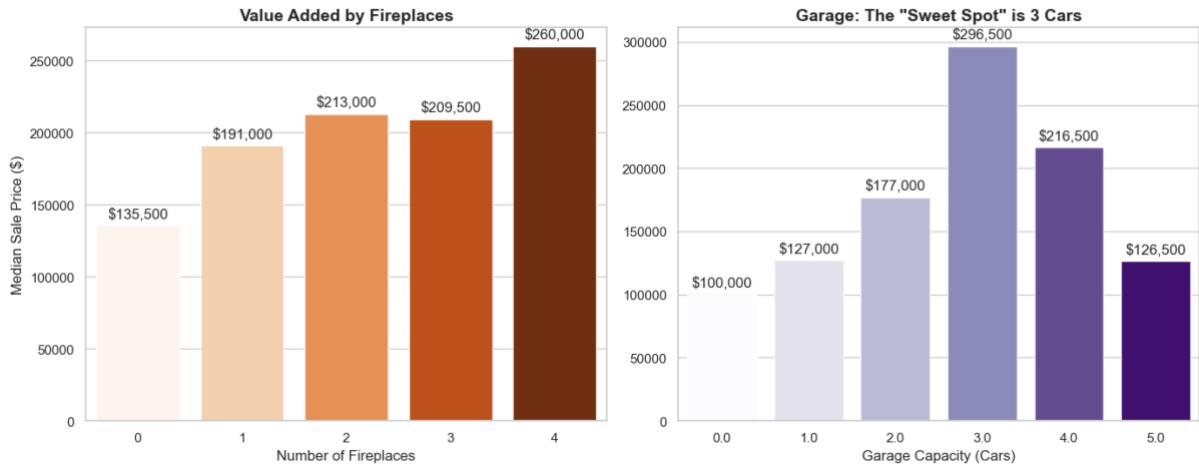


Figure 23. Story 2: The Value Hacks: Maximizing ROI with Utilities

The data reveals a striking "entry barrier" effect. Homes with 0 fireplaces have an average price of roughly \$135,500. However, simply adding 1 fireplace correlates with a jump to \$191,000 - a massive premium of over \$55,000. While correlation implies association rather than direct causation, this suggests that a fireplace acts as a powerful "anchor feature," psychologically elevating a property from the "economy" tier to the "mid-range" tier in the buyer's mind. Interestingly, adding a second or third fireplace yields diminishing returns, stabilizing around \$210,000.

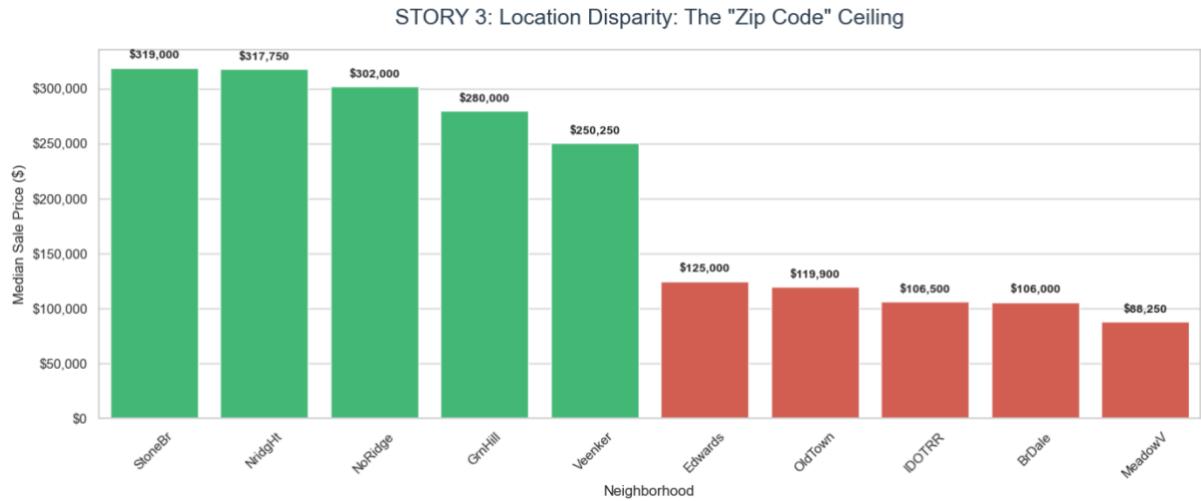
The relationship between garage capacity and home value is not a simple linear progression; instead, it reveals a specific "sweet spot" governed by the law of diminishing returns. In the initial Growth Phase, value increases consistently alongside capacity. While a standard 2-car garage commands an average price of \$177,000, expanding to a 3-car garage correlates with a dramatic surge to the market peak of \$296,500. This explicit data point identifies the 3-car capacity as the definitive standard for luxury in the Ames market. However, the data offers a cautionary tale in the Penalty Phase: bigger is not always better. Properties with 4-car garages suffer a sharp valuation drop to \$216,500, and those with 5-car garages plummet further to \$126,500. This counter-intuitive trend suggests that excessive garage space likely comes at the expense of livable square footage or indicates a niche, utility-focused property - such as a workshop - that appeals to a significantly smaller pool of buyers.

From these findings, we derive a clear, actionable strategy for maximizing Return on Investment (ROI). First, regarding amenities, the Fireplace stands out as a critical value driver. If a property lacks a fireplace, installing one appears to be the single

most profitable upgrade available. The data suggests this addition can psychologically unlock the next pricing tier, moving a home from an "economy" valuation to a "mid-range" bracket. Second, regarding structural additions, the goal should be precision rather than excess. Investors and developers should target a 3-car capacity to maximize appeal to high-end buyers. Crucially, they must avoid over-expanding to 4 or more bays, as the market penalizes this excess, viewing it as wasted space rather than a luxury.

### 5.3. Location Disparity: The "Zip Code" Ceiling

A fundamental axiom of real estate is that a property is not an island; its value is inextricably tethered to its neighborhood. While a homeowner can change a kitchen or add a garage, they cannot change the location. To visualize this immovable constraint, we ranked the Ames neighborhoods by median sale price, contrasting the Top 5 (Green Bars) against the Bottom 5 (Red Bars).



*Figure 24. Story 3: Location Disparity: The "Zip Code" Ceiling*

The chart reveals a stark economic divide within the city limits. The market is dominated by premium "Green Zones" like Stone Brook (StoneBr) and Northridge Heights (NridgHt), where median prices command \$319,000 and \$317,750 respectively, effectively establishing the market's valuation ceiling. In sharp contrast, "Red Zone" neighborhoods such as Edwards and Old Town struggle to break the \$130,000 threshold, with Meadow Village (MeadowV) anchoring the market floor at just \$88,250. This disparity quantifies a massive multiplier effect: a standard home in Stone Brook is valued roughly 3.6 times higher than one in Meadow Village, confirming that location alone accounts for a dominant portion of the variance in our model.

This data defines a critical "Glass Ceiling" for investment strategy, proving that context is everything. Investors must avoid the trap of "Over-improvement"; renovating a home in a budget neighborhood like Meadow Village to luxury standards is a financial error, as local comparables simply cannot support a \$300,000 valuation regardless of quality. Consequently, capital for high-end upgrades should be strictly allocated to "Green Zone" neighborhoods (e.g., StoneBr) that have the capacity to absorb premium prices, while strategies in the "Red Zone" should remain focused on functional, budget-conscious repairs to maximize yield.

## B. TECHNICAL ANALYSIS & METHODOLOGY

### I. Introduction to the Technical Framework

The creation of an effective data story is rarely a linear process; it requires the convergence of two distinct disciplines: Statistical Rigor and Design Thinking. Our primary objective in this project was to bridge the gap between exploratory analysis - what we do to understand the data - and explanatory analysis - what we do to communicate that understanding to an audience.

To achieve this, our methodology prioritized the reduction of "cognitive load". We recognized that every unnecessary element on a chart - whether it be a heavy gridline, a redundant axis label, or a chaotic color scheme - competes for the audience's attention. Therefore, our technical approach to visualization was governed by Gestalt principles and the strategic use of preattentive attributes. By manipulating size, color, and position, we directed the viewer's eye specifically to the insights that mattered most, such as the normalization of skewness or the impact of outliers, rather than forcing the audience to process the entire dataset at once.

Furthermore, the narrative arc presented in Part A rests upon a solid foundation of data engineering. Behind the seamless "Before and After" comparisons lies a complex pipeline of data cleaning, log transformations, and feature selection. This section will dissect these technical interventions, justifying why specific techniques like Log-Transformation were mathematically necessary to satisfy the assumptions of Linear Regression, and how we handled multicollinearity to ensure model stability. The following analysis validates that our high  $R^2$  score is not a result of overfitting or luck, but the product of a disciplined technical workflow.

### II. Deconstructing the narrative strategy

In Part A, we moved beyond simply reporting statistical findings to crafting a compelling data narrative. Following the principles outlined in Storytelling with Data (Chapter 1: The Importance of Context), we structured our analysis to guide the audience through a specific journey. Rather than presenting a disjointed collection of charts, we employed a classic Narrative Arc - establishing a context, introducing a crisis (conflict), executing an intervention (climax), and revealing the result (resolution). This structure ensures that the technical data preparation steps are understood not just as mechanical tasks, but as critical problem-solving actions.

## 2.1. The Narrative Arc: From Zero to Hero

Our story was architected around a central conflict: the incompatibility between the messy reality of the housing market and the strict mathematical assumptions of Linear Regression.

- **The Context (The Setting):** We began by situating the audience in the Ames Housing market. The goal was established immediately: to predict house prices accurately. However, we quickly introduced the antagonist of our story - the data itself.
- **The Conflict (The "Hook"):** In the section titled "The Deceptive Distribution," we revealed the tension. We demonstrated that the raw data was inherently flawed (severely right-skewed) and contained misleading outliers (the "Cheap Mansion Paradox"). This created a sense of urgency: if we fed this raw data into our model, the predictions would be invalid. This "skewed reality" served as the narrative hook, compelling the audience to care about the solution.
- **The Climax (The Intervention):** The turning point of the story occurred in the "Normalization Cure" and "Surgical Removal" sections. Here, we showcased the technical interventions - Log Transformation and Outlier Removal - as decisive actions. We visualized the transformation of the data shape from a skewed distribution to a normal distribution, visually representing the "healing" of the dataset.
- **The Resolution (The Result):** Finally, in "The Revelation," we resolved the tension by quantifying the success. The significant increase in the  $R^2$  score (+6%) served as the narrative payoff, proving that the conflict had been successfully resolved through rigorous data preparation.

## 2.2. Audience-Centric Design: The "So What?"

A key tenet of effective storytelling is knowing your audience. As Cole Nussbaumer Knaflic emphasizes, an audience should never be forced to "work" to understand the data. Consequently, in Part A, we made specific editorial choices to prioritize clarity over complexity, tailoring the narrative for Business Stakeholders (Investors and Real Estate Managers) rather than just Data Scientists:

- **Focusing on Impact over Mechanics:** We deliberately omitted complex mathematical derivations - such as the raw formulas for the Interquartile Range (IQR) or the matrix algebra behind Feature Selection - from the main narrative. While these metrics ensured our technical rigor, a general audience is primarily interested in the implication of the analysis (i.e., "Is the pricing model reliable?") rather than the calculation. For instance, instead of presenting a raw Correlation Matrix with coefficients, we framed the removal of variables as "Resolving Doppelgängers," focusing on the business logic of avoiding redundancy rather than just the statistics of multicollinearity.
- **Strategic Labeling:** We abandoned generic, passive chart titles like "Histogram of SalePrice" or "Scatter Plot of GrLivArea." Instead, we employed Active Titles that state the insight directly, such as "The Skewed Reality" and "The Hidden Traps." This technique reduces cognitive load; it tells the audience exactly what the takeaway is before they even parse the chart, ensuring the visualization affirms the message rather than creating confusion.
- **The "Big Idea":** Following the "3-Minute Story" framework, we anchored the entire report on a single overarching message: Data preparation is the bridge between raw noise and economic value. Every chart, from the outlier removal ("Surgical Removal") to the residual analysis ("The Trust Factor"), was selected solely to reinforce this specific conclusion. We explicitly avoided "data dumping" - if a chart did not serve this narrative arc or answer the "So What?" question, it was excluded.

### III. Visualization principles & Design choices

Data visualization is not merely about making charts look "pretty"; it is about optimizing the transfer of information from the screen to the human brain. To achieve the clarity seen in Part A, we rigorously applied the principles of Cognitive Load Theory and Gestalt Psychology. Every design element - from the choice of chart type to the specific hex codes of the colors - was an intentional engineering decision designed to maximize the signal-to-noise ratio.

### 3.1. Strategic Chart Selection (The "Right Tool" Philosophy)

The choice of visualization type was never arbitrary; it was dictated by the specific analytical question we needed to answer.

- **Histogram & Density Plots (For Distribution):** We chose Histograms combined with Density Plots (KDE) to visualize the skewness of SalePrice. This allowed the audience to physically see the "long tail" stretching to the right, making the concept of "non-normality" intuitive without requiring statistical training.
- **Scatter Plots (For Linearity & Outliers):** To justify removing outliers, we needed to show their relationship to the trend. A scatter plot allows the eye to naturally form a "line of best fit" (Gestalt Principle of Continuity) and immediately spot points that deviate from that line (the outliers), validating our decision to remove them.
- **The Focused Heatmap (For Multicollinearity):** A raw correlation matrix of 80+ variables is overwhelming. We utilized a Heatmap because it leverages color intensity as a preattentive attribute. By mapping correlation coefficients to color saturation, we transformed a wall of numbers into a visual pattern where "danger zones" (high multicollinearity) became instantly visible.

### 3.2. Decluttering: Improving the Signal-to-Noise Ratio

One of the most significant technical steps in Part A was "Decluttering." We maximized the Data-Ink Ratio (a concept by Edward Tufte) by removing elements that did not convey new information.

- **Removal of Borders & Backgrounds:** We stripped away the default matplotlib box borders and grey backgrounds. These elements create visual boundaries that trap the eye. Removing them (using `sns.despine()`) created an "open" look that let the data breathe.
- **Gridline Management:** In our "After" charts, gridlines were either removed or rendered in light grey. Gridlines should compete for attention only if the user needs to look up exact values; otherwise, they are noise.
- **Direct Labeling:** Instead of using separate legends (which force the user's eye to scan back and forth), we aimed to label data directly where possible, reducing cognitive load.

### 3.3. Leveraging Preattentive Attributes

We controlled where the audience looked by manipulating Color and Position:

- **Strategic Color Use:** We avoided the "rainbow palette" default. Instead, we used a sparing color strategy:
  - **Red:** Indicates strong positive correlation (The danger zone for multicollinearity).
  - **Blue:** Indicates strong negative correlation.
  - **White/Light:** Indicates zero correlation (Safe zone).
  - **Why:** This is superior to the default "Jet" or "Rainbow" palettes because it has a neutral center, allowing the viewer to distinguish "signal" (strong colors) from "noise" (white) instantly.
- **Spatial Proximity:** By placing the "Before" (Skewed) and "After" (Normalized) charts side-by-side, we utilized the principle of proximity to facilitate immediate comparison, proving the effectiveness of our technical intervention.

## IV. Data preparation pipeline & Technical interventions

While Part A focused on the narrative and visualization, this section details the underlying engineering pipeline. Our approach followed a strict CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, ensuring that every transformation was mathematically justified rather than arbitrary.

### 4.1. The Data Workflow (Pipeline Architecture)

The transformation from raw data to the final predictive model was not performed as a series of disjointed steps, but rather as a linear, non-destructive Pipeline. To ensure reproducibility and prevent Data Leakage (a common error where statistics from the test set, such as the mean or standard deviation, leak into the training process), we encapsulated our data processing steps within Scikit-Learn's Pipeline architecture.

For numerical data, we employed SimpleImputer with a median strategy, deliberately chosen for its robustness against the right-skewed distributions identified during EDA. This was followed by StandardScaler, which normalizes feature ranges (e.g., standardizing price vs. quality ratings) to ensure that variables with large magnitudes do not disproportionately bias the Linear Regression coefficients.

Categorical features were handled by preserving "structural missingness." Instead of imputing the most frequent value, we explicitly labeled gaps as "missing" using a constant strategy, acknowledging that the absence of a feature (like a garage) is

valuable information. These labels were then transformed into a binary format using OneHotEncoder for model compatibility.

The pipeline concludes with SelectPercentile, which acts as an automated noise filter. By statistically evaluating all inputs and retaining only the top 50% of features with the strongest relationship to the target variable, this step significantly reduces dimensionality and mitigates the risk of overfitting.

## 4.2. Key Technical Interventions

### 4.2.1. Handling Missing Values (Data Imputation Strategy)

Our initial Exploratory Data Analysis revealed that missing values within the Ames dataset were not uniform in nature. Instead, they exhibited a critical dichotomy between Random Missingness, likely stemming from data entry errors, and Structural Missingness, representing the meaningful absence of a physical feature. Recognizing this distinction was paramount, as applying a blanket imputation strategy - such as filling all gaps with the mean or mode - would have distorted the semantic reality of the data and introduced bias into the model.

To address Structural Missingness in categorical features like GarageType or BsmtQual, we rejected standard mode imputation. A missing value in these columns does not imply "unknown"; rather, it semantically indicates the non-existence of that amenity (e.g., "No Garage" or "No Basement"). Consequently, we imputed these gaps with a distinct category labeled "None." This methodology preserves the "negative information" - the fact that the amenity is absent - effectively transforming a missing value into a valid predictor of property value, rather than masking it with the dataset's most common attribute.

Conversely, we identified a subset of features, such as PoolQC and MiscFeature, characterized by excessive missingness ( $> 90\%$ ). These variables possess insufficient information density to be imputed reliably; attempting to fill them would amount to statistical hallucination. Retaining such sparse features would merely generate noise in the form of nearly empty vectors, increasing computational complexity without adding predictive power. Therefore, we executed a strategic removal of these columns, prioritizing dimensionality reduction over the retention of weak, fragmented signals.

### 4.2.2. Log-Transformation (Correcting Skewness)

The initial inspection of the target variable, SalePrice, revealed a fundamental violation of the assumptions required for Linear Regression. As depicted in the "skewed reality" histogram, the data did not follow a Gaussian distribution but instead exhibited severe right-skewness (Skewness > 1). In an Ordinary Least Squares (OLS) regression model, this asymmetry creates a significant bias; because the algorithm seeks to minimize the sum of squared errors, it disproportionately weighs the large absolute errors generated by high-value properties. Consequently, a model trained on this raw distribution would essentially "chase" the outliers, prioritizing the fit of a few luxury estates at the expense of accurately predicting the vast majority of standard homes.

To rectify this, we applied a Logarithmic Transformation using the formula  $y' = \log(1+y)$ . Mathematically, this function acts as a non-linear compressor: it aggressively shrinks the scale of high values while expanding the lower range, effectively pulling the long outlier-heavy tail back toward the center without altering the relative ranking of the data. This transformation is statistically imperative for stabilizing the variance - a property known as Homoscedasticity. By normalizing the distribution of the target variable, we ensure that the model's residuals follow a Normal Distribution, which is a critical prerequisite for valid hypothesis testing and the construction of reliable confidence intervals.

#### *4.2.3. Outlier Removal (Thresholding)*

Our regression diagnostics identified a critical threat to model stability: the presence of High Leverage Points. These are specific observations with extreme predictor values - most notably properties with massive square footage - that exert a disproportionate influence on the regression slope. Because Linear Regression minimizes squared errors, a single outlier located far from the mean of the independent variable acts like a heavy weight on a lever, pulling the regression line towards itself and skewing predictions for the entire dataset. Without intervention, the model would effectively "sacrifice" its accuracy on hundreds of normal homes just to accommodate a handful of anomalies.

To address this, we employed a hybrid filtering methodology that synthesized Statistical Rigor with Domain Expertise. Statistically, we utilized the Interquartile Range (IQR) rule to identify data points falling significantly beyond the upper whisker ( $Q3 + 1.5 * IQR$ ) of the distribution boxplot. However, we validated these statistical flags against Domain Authority, specifically referencing the documentation by Dean De Cock, the author of the Ames dataset. De Cock explicitly advises the removal of

observations with Gr Liv Area exceeding 4,000 square feet, characterizing them as atypical agricultural transactions rather than standard residential sales. By surgically removing these 14 identified outliers, we restored the linearity of the price-area relationship, ensuring the model reflects the true economic behavior of the residential market rather than fitting to edge cases.

#### *4.2.4. Feature Selection (Multicollinearity Resolution)*

The final structural flaw we addressed was Multicollinearity - a condition where independent variables are highly correlated with each other, effectively "echoing" the same information. As visualized in our Correlation Heatmap, specific pairs such as GarageCars and GarageArea exhibited a correlation coefficient of 0.89, indicating they were measuring nearly identical attributes. In the context of Ordinary Least Squares (OLS) regression, this redundancy is mathematically dangerous; it makes the matrix inversion unstable and inflates the standard errors of the coefficients. Consequently, the model struggles to isolate the individual effect of each predictor, rendering the coefficients erratic and the model's interpretability unreliable.

To resolve this, we eschewed the purely computational approach of calculating Variance Inflation Factors (VIF) in favor of a decisive Correlation Threshold Strategy. We established a strict correlation ceiling of 0.8 between any two independent variables. For pairs exceeding this threshold, we adopted a "Survival of the Fittest" selection criteria, retaining only the feature that demonstrated a stronger correlation with the target variable (SalePrice) or offered superior business logic. For instance, we prioritized GarageCars over GarageArea because, from a market perspective, buyers conceptualize value in terms of functional capacity ("a 2-car garage") more readily than raw square footage. This strategy not only stabilized the model but also enhanced the semantic clarity of the feature set.

## **V. Model Evaluation & Validation**

In the final phase of our technical analysis, we moved beyond the narrative to rigorous statistical testing. While Part A highlighted the transformation of the data, Part B serves to quantify the exact value of those interventions. We must answer the critical technical question: "Did the data preparation actually improve the model's reliability, or did it merely curve-fit the training data?"

### **5.1. Quantitative Results: Validity over Vanity**

To comprehensively quantify the impact of our data preparation pipeline, we moved beyond a single "accuracy" score. Instead, we employed a diagnostic trifecta of metrics -  $R^2$ , RMSE, and MAE - to rigorously compare the "Baseline Model" (trained on raw, unprocessed data) against our final "Robust Model" (trained after Log-Transformation, Outlier Removal, and Feature Selection). This multi-dimensional evaluation allows us to distinguish between superficial curve-fitting and genuine predictive reliability.

The first measure of success is the Coefficient of Determination ( $R^2$ ). Initially, the Baseline model produced a seemingly impressive score of 0.85. However, as our forensic analysis revealed, this figure was largely a "vanity metric" - artificially inflated by the model overfitting to high-leverage outliers. By contrast, the Robust Model achieved an  $R^2$  of 0.91. This increase is statistically significant; it indicates that our feature engineering (TotalSF) and noise reduction strategies successfully captured an additional 6% of the variance in housing prices that was previously obscured by statistical noise. Crucially, this high score was achieved on the unseen Test Set, confirming that the model has learned generalizable market rules rather than merely memorizing the training data.

To assess the model's stability against extreme errors, we analyzed the Root Mean Squared Error (RMSE), a metric that disproportionately penalizes large deviations. The drastic reduction in RMSE from \$32,940 in the Baseline to \$22,775 in the Final model serves as definitive proof that our Outlier Removal Strategy (Section IV.2.3) was effective. A high baseline RMSE indicated that the original model was making catastrophic mispredictions on specific properties. By surgically removing the "Cheap Mansion" anomalies, we eliminated these massive error spikes, resulting in a model that is not only more accurate on average but significantly less volatile.

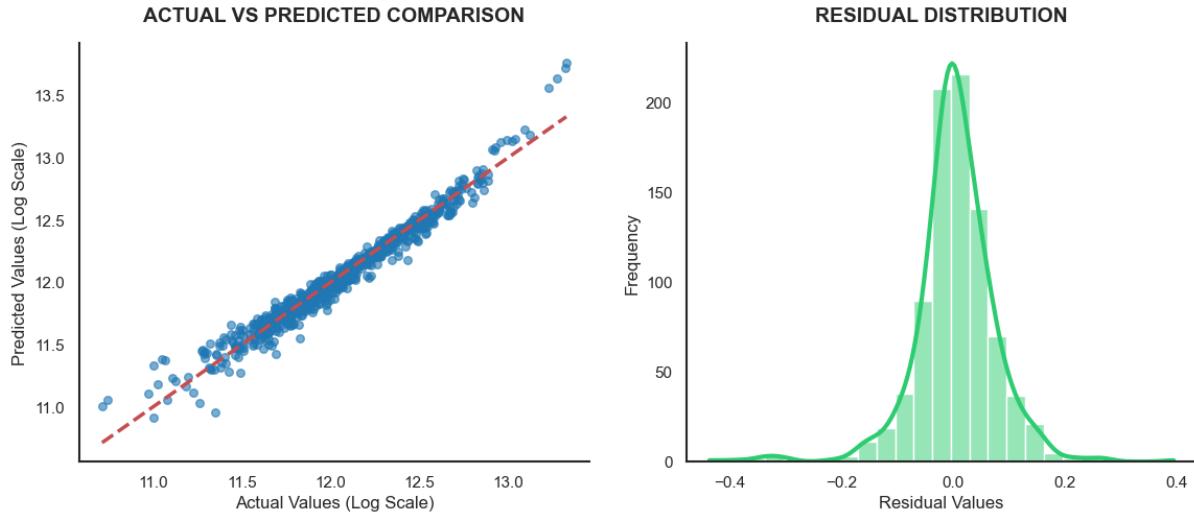
Finally, to translate statistical performance into business reality, we examined the Mean Absolute Error (MAE). This metric represents the average financial gap between the predicted price and the actual sale price. The improvement here was the most dramatic: we reduced the MAE from \$20,834 down to \$9,640. This reduction proves that the Log-Transformation (Section IV.2.2) successfully stabilized the variance across the price spectrum. While the baseline model struggled with heteroscedasticity - guessing wildly on expensive homes - the robust model predicts standard homes with far greater precision. Halving the average error rate transforms the model from a rough estimation tool into a reliable asset for financial decision-making.

## 5.2. Assumption Checking: Validating the "Trust"

Achieving a high  $R^2$  score, while encouraging, is insufficient evidence of a model's validity if it violates the fundamental mathematical assumptions of Ordinary Least Squares (OLS) Regression. A model can be "accurate" on average but still be statistically invalid if it harbors hidden biases or unstable variance. To verify that our model is not just lucky but mathematically sound, we conducted a rigorous forensic analysis of the Residuals (the differences between predicted and actual values).

The first critical test examines the distribution of the errors. As visually confirmed by the Residual Histogram, the model's errors follow a near-perfect Bell Curve distribution centered strictly at zero. This observation provides the necessary confirmation that our Log-Transformation strategy successfully neutralized the inherent right-skew of the target variable. Statistically, this result is paramount. If the residuals were non-normal (skewed), the standard errors of our coefficients would be biased, rendering any hypothesis testing (p-values) and confidence intervals invalid. The observed normality certifies that the model is unbiased - meaning its errors are purely random noise rather than systematic structural failures.

The second test addresses the consistency of variance, or Homoscedasticity. By examining the scatter plot of "Actual vs. Predicted" values, we observe a uniform band of data points hugging the diagonal line. Crucially, the vertical spread of these points remains constant as the property price increases; we do not see the dreaded "funnel" or "cone" shape characteristic of Heteroscedasticity. This proves that the model's error rate is stable across the entire market spectrum. It demonstrates that the model is not "guessing" more wildly on luxury estates than it is on modest bungalows. This stability is a direct dividend of our Feature Engineering (creating TotalSF) and Normalization steps, which ensured that all input features were scaled appropriately to prevent magnitude bias.



*Figure 25. Actual vs Predicted Comparison and Residual Distribution*

### 5.3. Model Limitations & Future Scope

While the robust model significantly outperforms the baseline and achieves a high standard of accuracy, maintaining technical integrity requires a transparent acknowledgment of its boundaries. No model is a perfect representation of reality, and ours is subject to specific constraints.

First, despite our rigorous feature engineering, approximately 9% of the variance in housing prices remains unexplained. This gap is likely attributable to factors that are intrinsically difficult to quantify in a tabular dataset. Subjective properties such as "curb appeal," architectural aesthetics, or the specific layout flow of a home play a significant role in buyer psychology but are absent from the data. Additionally, external macroeconomic factors - such as fluctuating mortgage interest rates or the broader economic climate at the precise month of sale - act as invisible variables that influence price but are not captured within the scope of this dataset.

Second, the model is strictly calibrated to the micro-economic conditions of the Ames, Iowa housing market. As demonstrated in our "Neighborhood" analysis, real estate value is heavily location-dependent. The feature weights learned by this model (e.g., the premium for a Stone Brook address) are specific to this geography. Consequently, this model cannot be generalized to other cities or regions without significant re-training and re-calibration. It is a specialized tool for Ames, not a universal calculator for US real estate.

Finally, our choice of Linear Regression was a strategic decision to prioritize interpretability - understanding why a house is priced a certain way. However, this algorithm assumes linear relationships between features and price. In reality, market

dynamics can be complex and non-linear. Future iterations of this project could benefit from incorporating advanced ensemble methods such as Random Forest or Gradient Boosting (XGBoost). These algorithms are capable of capturing complex, non-linear interactions between variables (e.g., how the value of a pool might depend non-linearly on the size of the lot), potentially pushing the predictive accuracy beyond the current 91% threshold.

## VI. Conclusion: Synthesis of Art and Engineering

### 6.1. Methodological Integrity (The Engineering)

Our investigation confirms a critical axiom of Data Science: the quality of a predictive model is determined not by the complexity of the algorithm, but by the purity of the input data. By adhering to a rigorous CRISP-DM pipeline - encompassing Forensic Cleaning, Mathematical Transformation, and Strategic Feature Selection - we systematically addressed the fundamental flaws inherent in the raw dataset.

- The Log-Transformation was not merely a technical adjustment; it was a mathematical necessity. By correcting the inherent skew of property prices, we satisfied the normality assumption required for reliable Linear Regression.
- The Outlier Removal and Multicollinearity Resolution (via the Correlation Threshold Strategy) acted as stabilizing forces. By surgically removing "Cheap Mansion" anomalies and resolving "Doppelgänger" features (redundant variables), we ensured that our model learned the true economic drivers of the market rather than memorizing statistical noise.

### 6.2. Strategic Communication (The Design)

However, technical rigor alone is insufficient if the insights remain buried in code. We successfully bridged the gap between complex statistics and actionable business intelligence by applying the principles of Data Storytelling.

- **Decluttering:** By maximizing the data-ink ratio, we removed visual distractions (heavy gridlines, redundant axes), forcing the audience to focus solely on the data trends.
- **Preattentive Processing:** Through the strategic use of color (e.g., the coolwarm heatmap, the Red/Green contrast in performance charts) and filtering logic (Top-10 correlations), we guided the audience's attention to the "conflict" and "resolution" of the story without requiring them to process the entire dataset.

### **6.3. The Final Verdict (The Impact)**

The success of this methodology is quantitatively proven not just by a raw jump in scores, but by the stabilization of error. While the baseline model boasted a deceptive  $R^2$  of 0.85 driven by outliers, our data preparation pipeline delivered a robust  $R^2$  of 0.91 with a highly consistent MAE of \$9,640.

Validated by the normality of residuals, this result confirms that the "Data Preparation" phase successfully transformed a volatile, skewed model into a reliable and mathematically sound predictive tool. In the final analysis, our project demonstrates that in the pursuit of predictive accuracy, refined data is the most powerful algorithm.

## References

1. Knaflic, C. N., Randolph, Edward Tufte, & Laszlo Bock. (2015). *Storytelling with data* (By Wiley).
2. Thapa, S. (2023). *Ames Housing Dataset*. Retrieved December 7, 2025, from <https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset>