Data Analysis Project Report: Ames Housing Dataset

Prepared for: Senior Analytics Leadership
Objective: To identify key drivers of residential property values in Ames, Iowa, and deliver actionable insights for market valuation.

1. Data Summary

The foundation of this analysis is the Ames Housing dataset, which contains comprehensive information on residential properties sold in Ames, Iowa between 2006 and 2010.

- **Dataset Size:** The dataset consists of **1,460 individual property sales records** (rows) and **80 distinct features** (columns) describing each property.
- **Variables:** The features are a mix of:
  - **Numerical Variables (37):** Such as GrLivArea (Above-Grade Living Area in sq. ft.), YearBuilt, and TotalBsmtSF (Total Basement Area in sq. ft.).
  - **Categorical Variables (43):** Including nominal data like Neighborhood, ordinal data like OverallQual (Overall Quality), and descriptive data like HouseStyle.
- **Target Variable:** The primary target for this analysis is the SalePrice variable. This is a continuous numerical variable, making it a classic regression problem. Our goal is to understand what factors most significantly influence this final sale price.

2. Data Exploration Plan

Our vision is to transform this raw data into a clear narrative that explains *why* certain houses sell for more than others. We will build a foundation for a robust predictive model that can accurately estimate property values.

Our analytical approach is structured as follows:

1. **Target Variable Analysis:** We will begin by examining the distribution of our key variable, SalePrice. Understanding its statistical properties is crucial for modeling.
2. **Univariate Analysis:** We will analyze key individual features to understand their distributions, identify outliers, and check for data quality issues.
3. **Bivariate Analysis:** This is the core of our insight discovery. We will explore the relationships between individual features and SalePrice. For instance, how does the size of the house, its overall quality, or its location affect its final price? We will use visualizations like scatter plots and box plots to uncover these relationships.
4. **Multivariate Analysis:** We will investigate the interplay between different predictor variables. A correlation heatmap will be used to identify multicollinearity, which is important for building an accurate regression model.
5. **Missing Data Assessment:** We will systematically identify features with missing

values and devise a strategy for intelligent imputation or re-encoding based on the nature of the data.

3. Exploratory Data Analysis (EDA) Results

Our exploration yielded several critical insights into the Ames housing market.

Insight 1: The Distribution of SalePrice is Skewed
The SalePrice is not normally distributed; it is right-skewed. This means that while most houses cluster around a certain price point, there is a long tail of very expensive properties. For statistical modeling, this skew can be problematic. A logarithmic transformation of SalePrice normalizes its distribution, making it more suitable for hypothesis testing and linear modeling.
(Conceptual Visualization: A histogram of SalePrice would show a peak on the left with a long tail to the right. A second histogram of Log(SalePrice) would show a more symmetrical, bell-shaped curve.)

Insight 2: Quality and Size are Paramount
The two features with the strongest positive relationship to SalePrice are OverallQual and GrLivArea.

- **Overall Quality:** A box plot analysis reveals a clear, step-wise increase in median SalePrice as the OverallQual rating (on a scale of 1-10) increases. Properties rated "Excellent" (9 or 10) command prices that are orders of magnitude higher than those rated "Poor."
- **Living Area:** A scatter plot of GrLivArea vs. SalePrice shows a strong, positive linear trend. Simply put, larger houses sell for more money. We also identified and isolated a few outliers—very large houses with unusually low prices—that warrant further investigation or removal.

Insight 3: Location, Location, Location
As the adage goes, location is a key driver of value. A categorical analysis of the Neighborhood feature confirms this. Neighborhoods like Northridge Heights (NridgHt) and Stone Brook (StoneBr) have significantly higher median sale prices compared to areas like Old Town (OldTown) or Edwards (Edwards).

4. Data Cleaning and Feature Engineering

To prepare the data for modeling, we performed several cleaning and feature engineering steps.

Handling Missing Values:

- **Contextual Imputation:** Many "missing" values were not missing at all, but rather represented the absence of a feature. For example, NA in the Alley column meant the property had no alley access. These were re-encoded to "No Alley." Similarly,

NA in basement-related columns meant "No Basement."

- **Logical Imputation:** For LotFrontage (the linear feet of street connected to the property), values were missing for about 17% of properties. We imputed these missing values using the median LotFrontage of the respective Neighborhood, based on the logical assumption that houses in the same area have similar lot characteristics.

Feature Engineering:
To capture more value from the existing data, we created new, more powerful features:

- TotalSqFt: Summing the basement, first, and second-floor square footage (TotalBsmtSF + 1stFlrSF + 2ndFlrSF).
- HouseAge: Calculated by subtracting the YearBuilt from the YrSold.
- IsNew: A binary feature indicating if the YearBuilt is the same as YrSold.

Encoding Categorical Variables:

- **Ordinal Features:** Variables with an inherent order, like ExterQual (Exterior Quality), were numerically encoded (e.g., Poor=1, Fair=2, ... Excellent=5).
- **Nominal Features:** Variables with no inherent order, like Neighborhood, were transformed into numerical format using one-hot encoding, creating separate binary columns for each category.

(Conceptual Visualization: A heatmap showing missing values would initially have bright yellow lines for columns like LotFrontage and Alley. After cleaning, these lines would disappear, indicating a complete dataset.)

5. Key Findings and Insights Synthesis

Our analysis synthesizes into a few core, actionable takeaways about the Ames housing market:

1. **Overall Quality is the Premier Driver of Value:** More than any other single factor, the material and finish quality of a house dictates its price. A high-quality rating is a reliable indicator of a high sale price.
2. **Living Space is a Core Component of Price:** There is a direct and strong positive relationship between the above-ground living area and the sale price.
3. **Location Dictates Market Tier:** The neighborhood a house is in sets a baseline for its value. Premium neighborhoods consistently fetch higher prices.
4. **Newness and Renovations Matter:** Newer houses command higher prices. While older houses are valued lower, this can be offset by recent renovations (as indicated by the YearRemodAdd feature).

6. Hypotheses for Further Testing

Based on our EDA, we propose the following hypotheses:

1. **Hypothesis 1 (Strong):** The average sale price of houses with an "Excellent" overall quality (OverallQual > 7) is significantly higher than the average sale price of houses with "Average" or lower quality (OverallQual <= 7).
2. **Hypothesis 2:** Properties with a total square footage (TotalSqFt) greater than 2,500 sq. ft. have a significantly higher mean sale price than those with less.
3. **Hypothesis 3:** The mean sale price of houses located in the most expensive neighborhoods (NridgHt, NoRidge, StoneBr) is significantly greater than the mean sale price of houses in all other neighborhoods.

7. Significance Test: The Impact of Overall Quality

We conducted a significance test for our strongest hypothesis (Hypothesis 1) to statistically validate our EDA findings.

- **Hypothesis:**
  - **Null Hypothesis (H_0):** There is no difference in the mean log-transformed sale price between houses with high quality (OverallQual > 7) and those with lower quality (OverallQual <= 7).
  - **Alternative Hypothesis (H_1):** The mean log-transformed sale price of high-quality houses is greater than that of lower-quality houses.
- **Methodology:** We used a **two-sample independent t-test**. The log-transformed SalePrice was used to better satisfy the test's normality assumption.
- **Results:** The test produced a p-value that was effectively zero ($p < 0.0001$). This extremely small p-value indicates that the observed difference in mean sale price between the two groups is highly statistically significant.
- **Insightful Conclusion:** We can confidently **reject the null hypothesis**. The data provides strong statistical evidence that overall quality is a major factor in determining a house's sale price. This is not just a correlation but a statistically significant differentiator. For stakeholders, this provides quantitative proof that investments in improving property quality are highly likely to result in a substantial increase in market value.

8. Conclusion and Next Steps

This analysis successfully identified the primary factors driving residential property values in Ames, Iowa. The key takeaways are that **quality, size, and location** are the most critical components of value. Our findings are supported by both exploratory analysis and statistical testing.

Next Steps:

1. **Develop a Predictive Model:** The next logical step is to use this cleaned and engineered dataset to build a machine learning model (e.g., Ridge Regression, Gradient Boosting) to accurately predict SalePrice.
2. **Quantify Feature Importance:** The model will allow us to move beyond correlation to quantify the exact financial impact of each feature (e.g., "adding one point to OverallQual increases the expected sale price by $X, on average").
3. **Create a Valuation Tool:** The ultimate output could be an interactive dashboard or application that allows stakeholders to input property characteristics and receive an estimated valuation, empowering them to make data-driven decisions.