IBM Data Analyst Professional Certificate

Course: Data Analysis with Python

Final Project – Data Assessment Report

Exploratory Data Analysis and Hypothesis Testing on Ames Housing Dataset

Submitted by: Laura Puerto

Date: March 30, 2025

1. Dataset Description

The dataset used in this project is the Ames Housing dataset, which contains detailed information on residential property sales in Ames, Iowa.
It includes 2,930 observations and 80 variables describing various aspects of each property, such as lot size, year built, number of bedrooms, neighborhood, and overall quality.

The dataset is rich in both **numerical** and **categorical** features, making it ideal for exploratory data analysis and hypothesis testing. The target variable for this project is SalePrice, which represents the selling price of each house. This will allow us to investigate which features influence housing prices the most and to validate assumptions using statistical tests.

2. Initial Plan for Data Exploration

Before beginning data cleaning or feature engineering, an initial plan for data exploration is defined to guide the analytical process.

The main objectives of this exploration are:

- To examine the distribution of key numerical variables such as SalePrice, Gr Liv Area, and Lot Area.
- To explore the relationships between selected features and the target variable SalePrice.
- To investigate the impact of categorical variables like Neighborhood, HouseStyle, and OverallQual.
- To identify outliers and detect potential anomalies in the data.
- To assess missing values and determine the appropriate treatment.
- To detect variables that may require transformation (e.g., log-scaling) or encoding (e.g., one-hot).

This exploration strategy will ensure that further steps such as cleaning and modeling are based on a solid understanding of the dataset's structure and content.

3. Data Cleaning and Feature Engineering

We started the cleaning process by identifying missing values across the dataset. Several variables were found to contain null values, including features related to garage, basement, and pool information.

Variables with missing values will be analyzed individually to determine the most appropriate strategy:

- If the missing values represent the absence of a feature (e.g., no garage or no pool), we may replace them with a string such as "None" or use a dedicated category.
- For numerical variables, we may use imputation strategies such as replacing with 0 or the median, depending on the context.
- Variables with excessive missing values and low importance may be dropped entirely.

This analysis helps us ensure the dataset is clean and ready for further exploration and modeling.

Detailed analysis of missing values

Specifically, we observed that the following variables had a very high number of missing values:

- Pool QC (2917 missing out of 2930)
- Misc Feature (2824 missing)
- Alley (2732 missing)
- Fence (2358 missing)
- Fireplace Qu (1422 missing)

In these cases, the missing values likely indicate the absence of the feature (e.g., no pool, no alley access, no fireplace). These will be handled by replacing missing values with a distinct category such as "None".

Garage and basement-related variables (e.g., Garage Type, Garage Finish, Garage Qual, Bsmt Qual, Bsmt Cond) had 80–159 missing values. These will also be treated as "None" when appropriate.

Finally, variables with only 1 or 2 missing entries (e.g., Electrical, Garage Area) will be imputed using the **mode** or **median**, depending on the type of feature.

This more detailed analysis allows us to define a clear cleaning strategy adapted to each case.

Outlier Analysis

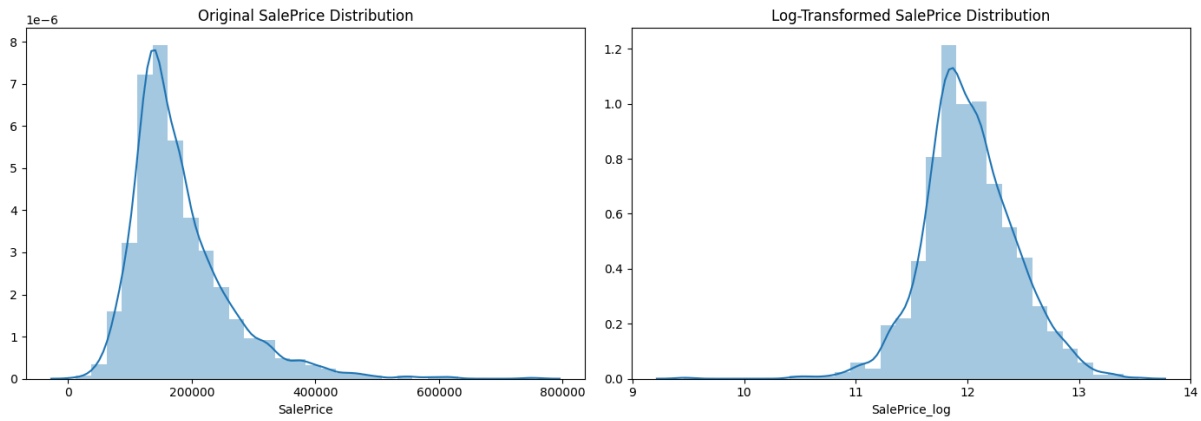
We visualized the distribution of the target variable SalePrice and found that it is **right-skewed**, with several extreme values on the higher end of the scale. These outliers could impact statistical testing and modeling.

A similar pattern is observed in the predictor variable Gr Liv Area, which also displays some unusually high values. Boxplots further confirm the presence of potential outliers.

In the next steps, we may consider applying a **log-transformation** to reduce skewness and improve the distribution for modeling.

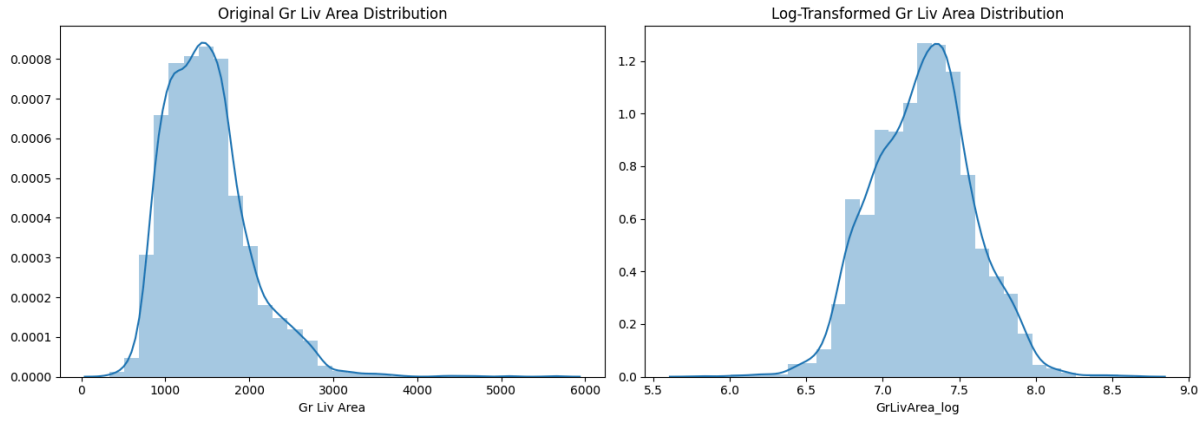
Log-Transformation of SalePrice

To reduce skewness and minimize the impact of extreme outliers, we applied a log-transformation to the target variable SalePrice.
The transformed variable SalePrice_log shows a much more symmetrical distribution, which is beneficial for statistical modeling and hypothesis testing.
This approach is commonly used in housing price prediction tasks and ensures that the model assumptions of linearity and normality are better satisfied.



Log-Transformation of Gr Liv Area

Following a similar reasoning, we also applied a log-transformation to the Gr Liv Area variable, which exhibited a right-skewed distribution with several outliers. The transformed version, GrLivArea_log, shows a more symmetric and compact distribution, which is expected to enhance the model's performance and its ability to detect linear patterns.



4. Key Findings and Insights

Based on the exploratory visualizations, we identified several strong relationships between housing characteristics and sale price:

- GrLivArea_log shows a clear **positive linear relationship** with SalePrice_log. Larger homes tend to have higher sale prices.
- Overall Qual exhibits a **stepwise increase in price** as the quality score rises, confirming it as one of the most influential predictors.
- The Neighborhood variable reveals **significant differences in median prices** across areas. For instance, StoneBr and NridgHt are associated with higher values, while BrDale and MeadowV tend to have lower ones.
- Different HouseStyle categories also correspond to **notable price differences**, suggesting that architectural design plays a role in perceived property value.

These findings will guide the formulation of hypotheses and selection of features for statistical modeling.

5. Hypotheses Formulation

Based on the exploratory analysis, we propose the following hypotheses:

1. **Overall Quality and Price**
   H_0: There is no significant difference in SalePrice_log between houses with low and high OverallQual.
   H_1: Higher OverallQual is associated with a higher SalePrice_log.

2. **Neighborhood and Price**
   H_0: The mean SalePrice_log is the same across all neighborhoods.
   H_1: At least one neighborhood shows a significantly different mean SalePrice_log.

3. **Living Area and Price**
   H_0: There is no correlation between GrLivArea_log and SalePrice_log.
   H_1: There is a significant positive correlation between GrLivArea_log and SalePrice_log.

6. Hypothesis Testing

To test whether overall quality has a significant impact on housing prices, we grouped the dataset into two categories:

- **Low quality**: Overall Qual ≤ 5
- **High quality**: Overall Qual > 5

We then conducted an **independent two-sample t-test** comparing the average SalePrice_log between both groups.

- **T-statistic**: 42.683
- **P-value**: 1.58×10^{-302}

The p-value is far below the conventional threshold of 0.05, indicating a **strong statistically significant difference** between the two groups.
This supports the hypothesis that **higher overall quality is associated with higher sale prices**.

The boxplot also visually confirms this finding, showing a clear upward shift in the distribution for higher quality homes.

7. Suggestions for Next Steps

To extend this analysis and gain deeper insights, the following next steps are recommended:

- **Include additional predictors** such as basement area, garage size, or number of bathrooms to improve the explanatory power of the model.
- **Incorporate interaction effects**, for example between neighborhood and quality, to explore whether the impact of quality depends on location.
- **Build a multivariate regression model** using both categorical and numerical features to predict SalePrice_log.
- **Apply cross-validation** to evaluate model robustness and prevent overfitting.
- **Conduct feature selection** techniques to identify the most relevant variables.
- **Explore non-linear models** such as decision trees or random forests to capture complex patterns in the data.

These steps would help improve the accuracy and interpretability of the analysis, and move towards a production-ready prediction model.

8. Data Quality Assessment and Additional Data Request

Overall, the dataset provides rich and detailed information about housing characteristics, with a good balance of numerical and categorical variables.

The number of observations (2,930) is sufficient for exploratory analysis and hypothesis testing.

However, several variables presented **a high proportion of missing values**, especially those related to pools, alleys, and fences. In most cases, these missing values were likely **not random**, but indicated the absence of the feature. This required careful treatment to preserve the underlying information without introducing bias.

Some categorical features contained **many levels** (e.g., Neighborhood), which adds complexity to the analysis and may require dimensionality reduction techniques.

If additional data were available, the following would improve the analysis:

- **Renovation year or remodeling information**, to understand how updates impact value.
- **Energy efficiency or heating system data**, as they can influence modern buyer preferences.
- **School quality or proximity to amenities**, which are known to affect housing prices.
- **Market trends or macroeconomic variables**, to account for time-related price shifts.

Despite these limitations, the dataset is of **high quality and well-suited for educational and modeling purposes**.