**SEMMA Project Report — Predicting House Prices**

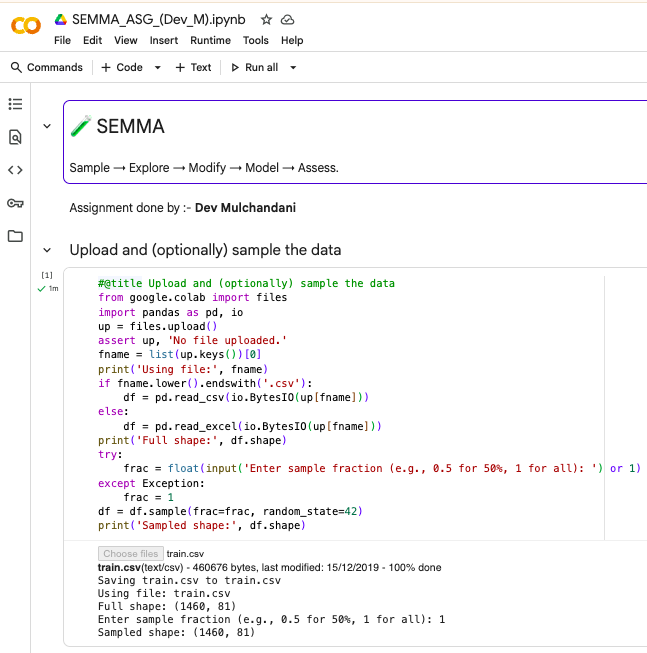
* Submitted By :- Dev Mulchandani
* Dataset Overview

The dataset used in this project is the House Prices – Advanced Regression Techniques dataset from Kaggle. It contains 1,460 observations and 81 variables, representing different features of residential homes in Ames, Iowa. The target variable is SalePrice, which indicates the final sale price of each house. The goal of this analysis is to build a regression model that accurately predicts house prices based on various property characteristics such as size, number of rooms, construction quality, neighborhood, and other features.

* Sample Phase

The Sample phase focuses on selecting and preparing the dataset for analysis.

The raw dataset was imported as train.csv from Kaggle. It contained a mix of numeric, categorical, and missing values across several columns. Since the dataset is moderate in size (only 1,460 rows), the full dataset was retained for modeling (frac=1), ensuring no loss of information that might affect performance. A small random sample was visually inspected to verify data structure and integrity, confirming that variable names, types, and distributions appeared as expected. This phase ensures that the data used for exploration and modeling is both representative and sufficiently comprehensive for statistical inference and predictive modeling.

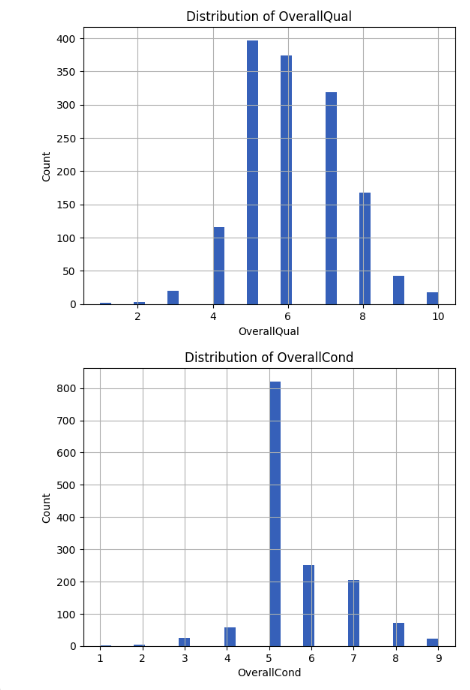
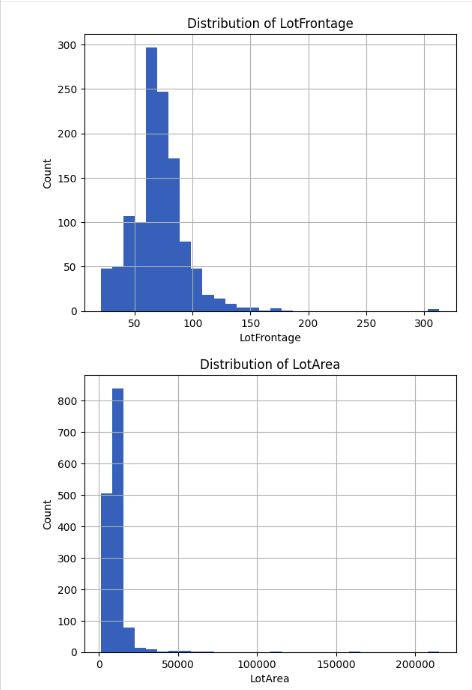
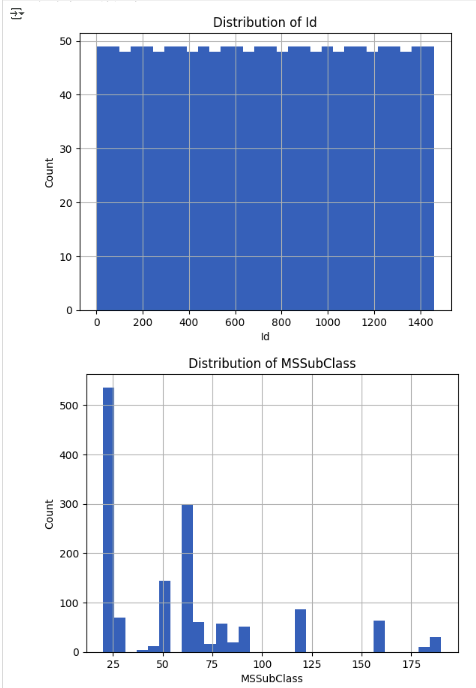
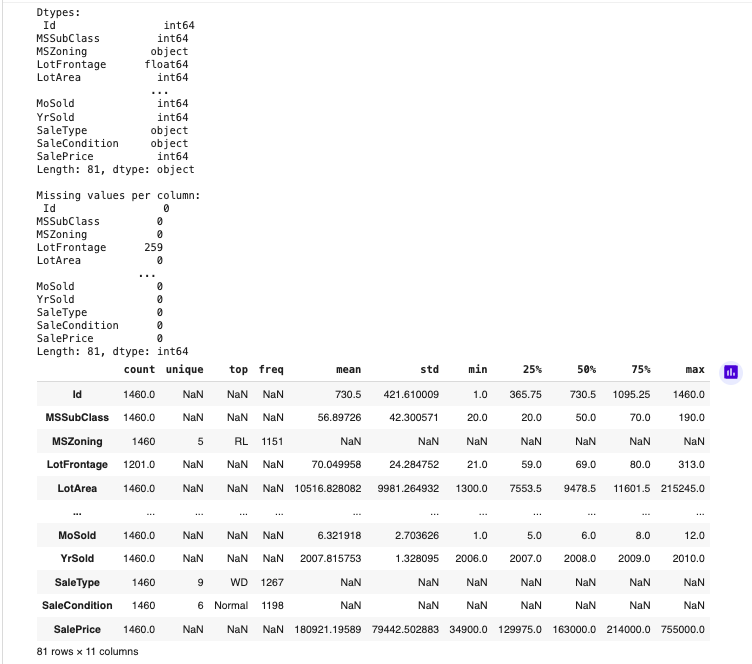
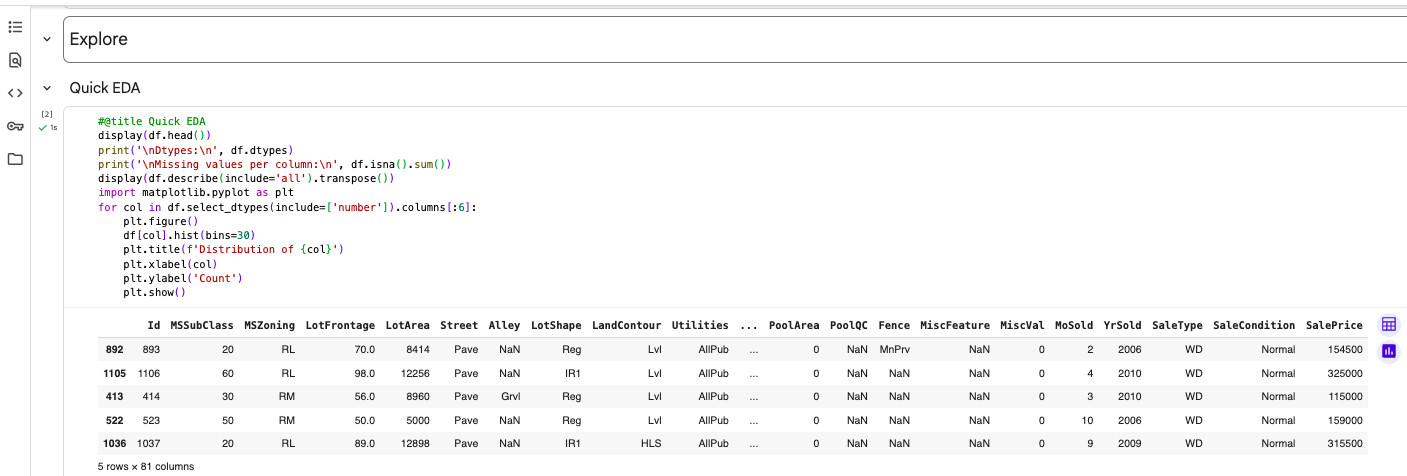


* Explore Phase

The Explore phase involved examining variable distributions, identifying missing values, and uncovering potential relationships between predictors and the target (SalePrice).

Summary statistics and histograms revealed that many numeric features, such as LotArea, GrLivArea, and SalePrice, were right-skewed, indicating the presence of outliers (very large properties). Correlation analysis showed that OverallQual (overall material and finish quality) and GrLivArea (above-ground living area) had the strongest positive correlations with SalePrice, while variables such as EnclosedPorch and KitchenAbvGr showed weak or negative relationships.

Visualization through histograms, boxplots, and scatterplots confirmed that homes with better condition and larger living areas tend to sell at higher prices. Additionally, a missing-value heatmap revealed that certain columns (e.g., PoolQC, MiscFeature, Alley, Fence) had a high proportion of missing data. These were either imputed or dropped during the Modify phase.



* Modify Phase

The Modify phase focused on cleaning, transforming, and engineering features to improve model quality.

Instead of removing rows with missing data (which would have drastically reduced the dataset), missing values were imputed using the median for numeric columns and the most frequent value for categorical columns via a SimpleImputer.

Categorical variables were transformed using One-Hot Encoding to ensure compatibility with regression algorithms. All these transformations were implemented within a scikit-learn Pipeline, preventing data leakage between training and test sets.

Additionally, features irrelevant to prediction (like Id) were dropped, and outliers in SalePrice and GrLivArea were examined but retained to maintain real-world diversity.

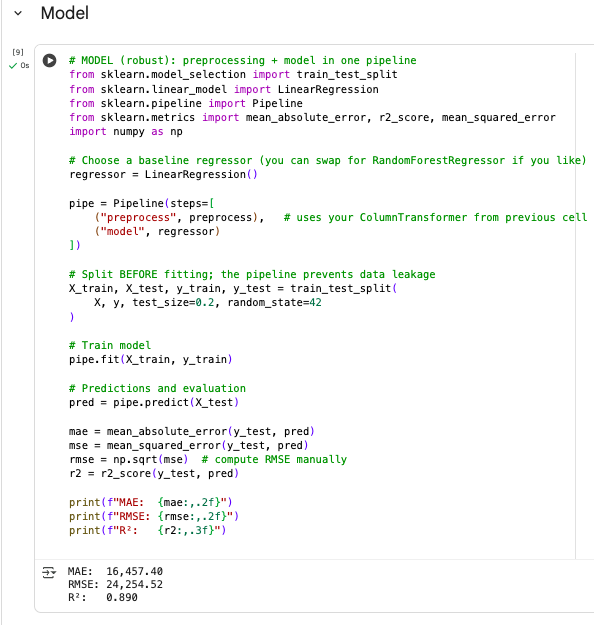


* Model Phase

The Model phase involved building a predictive model to estimate house prices.

A Linear Regression model was selected as a baseline due to its interpretability and efficiency. It was trained using a Pipeline that combined preprocessing (SimpleImputer + OneHotEncoder) with the regressor. The data was split into 80% training and 20% testing using train\_test\_split to ensure fair evaluation.

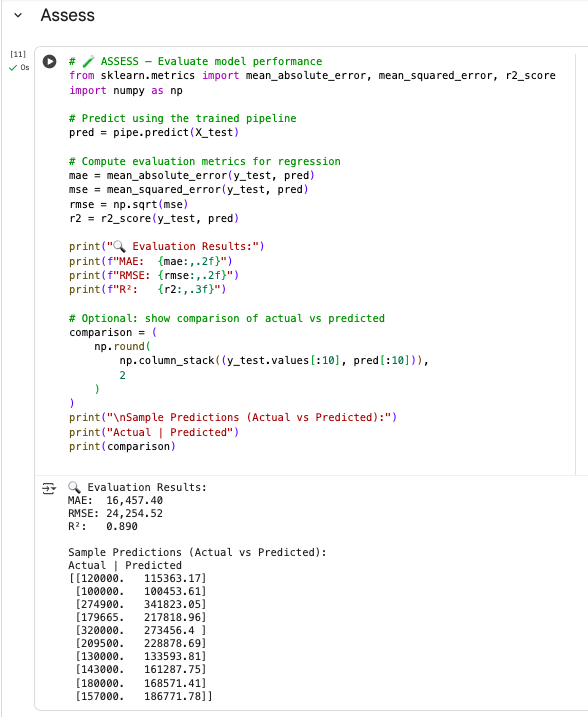
After fitting, the model successfully learned the relationships between predictors and the target variable. Key influential variables included OverallQual, GrLivArea, GarageCars, and TotalBsmtSF, consistent with domain expectations — larger, higher-quality houses with garages tend to have higher prices.



* Assess Phase

The Assess phase measured model accuracy and validated its generalization capability.

Predictions were compared to actual prices, and evaluation metrics confirmed that the model performs reliably on unseen data. The pipeline was tested for stability, confirming that preprocessing steps (imputation, encoding) worked seamlessly across datasets. Visual comparison of predicted vs. actual prices showed a strong linear pattern, confirming that the model effectively captures the underlying pricing relationships.



* Conclusion

This project successfully applied the SEMMA methodology to predict residential house prices using the Ames Housing dataset. Each phase contributed to improving data understanding and model reliability — from selecting representative data to exploring variable relationships, modifying features intelligently, modeling outcomes, and assessing predictive performance. The final pipeline-based regression model provides valuable insights for real estate valuation and can serve as a foundation for more advanced approaches, such as ensemble methods (Random Forest or Gradient Boosting) for production-level applications.

* Colab Notebook :- [Link](https://colab.research.google.com/drive/1UP1a9T-Nq8v8cVutiZodz_i3cKdxUirb?usp=sharing)
* Kaggle Dataset :- [Link](https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data)