**Report**

**Title: *Big Data Solutions to the Affordable Housing Crisis in Urban Areas***

**Topic:** Smart City - Housing Affordability

**Subject code:** DS-670-HYB2-23WNTR

**Subject:** Capstone: Big Data & Bus Analy

**Professor:** Reda Mastouri

19/12/2023

**Team Members**

1. Saikiran Reddy Durgareddygari
2. Guna Venkata Sai Bharath Bodepudi;
3. Sahil Vericherla.

**Regression Model Evaluation Report**

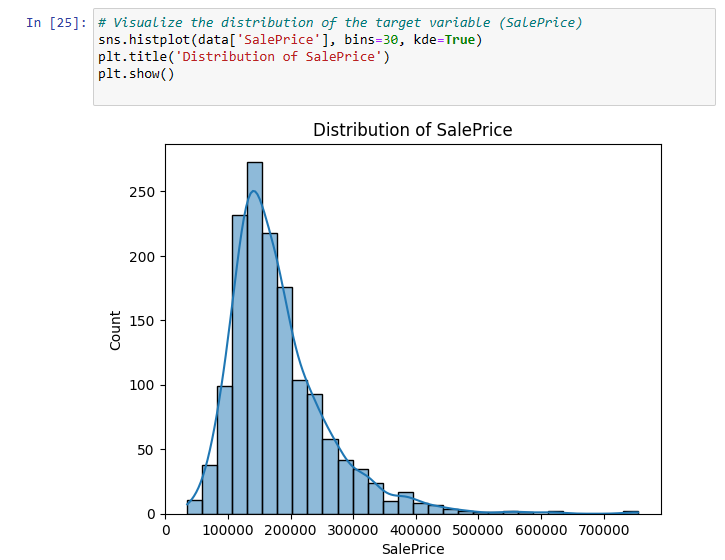
**Dataset Overview:**

* **Dataset Source:** Provided CSV file ("train.csv")
* **Number of Entries:** 1460
* **Number of Features:** 72 (after preprocessing)

**Exploratory Data Analysis:**

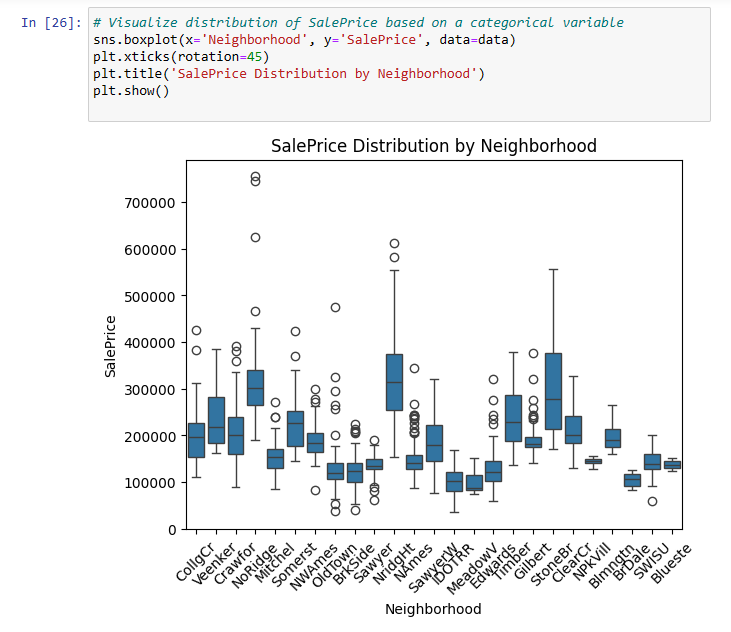
Target Variable Distribution:

* The distribution of the target variable "SalePrice" is right-skewed, indicating a concentration of lower-priced houses with a few higher-priced outliers.



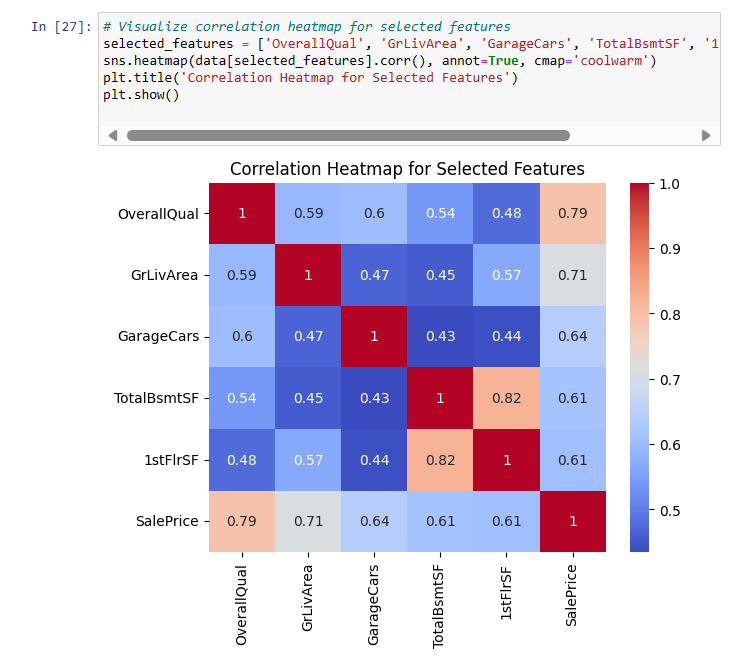
Neighborhood-wise SalePrice Distribution:

* SalePrice varies significantly across different neighborhoods, suggesting that location is a key factor in determining house prices.



Correlation Heatmap:

* Strong positive correlations observed between "OverallQual," "GrLivArea," "GarageCars," "TotalBsmtSF," and the target variable "SalePrice."



**Data Preprocessing:**

Columns Removed:

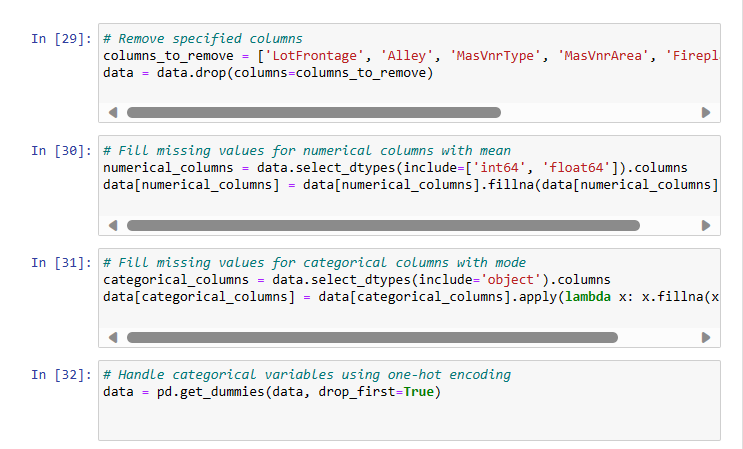
* The following columns were removed due to high missing values or deemed irrelevant: 'LotFrontage', 'Alley', 'MasVnrType', 'MasVnrArea', 'FireplaceQu', 'PoolQC', 'Fence', 'MiscFeature'.

Handling Missing Values:

* Numerical columns filled with mean values.
* Categorical columns filled with mode values.

One-Hot Encoding:

* Categorical variables were encoded using one-hot encoding to facilitate model training.



**Regression Model Evaluation:**

Linear Regression:

* Mean Absolute Error (MAE): 20,424.4996
* Mean Squared Error (MSE): 2,664,008,986.8477
* Root Mean Squared Error (RMSE): 51,614.0387
* Mean Absolute Percentage Error (MAPE): 12.1069

Random Forest Regressor:

* Mean Absolute Error (MAE): 17,798.3270
* Mean Squared Error (MSE): 842,636,815.2384
* Root Mean Squared Error (RMSE): 29,028.2072
* Mean Absolute Percentage Error (MAPE): 10.8326

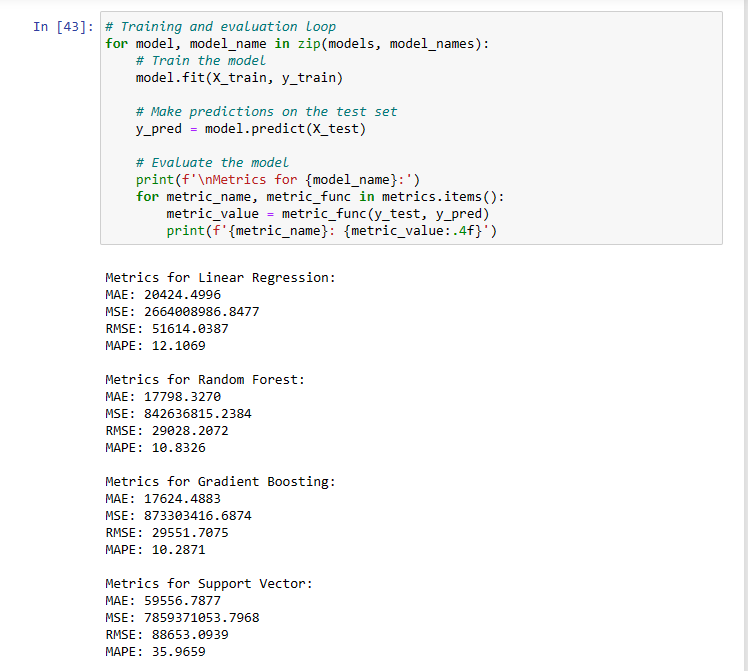
Gradient Boosting Regressor:

* Mean Absolute Error (MAE): 17,624.4883
* Mean Squared Error (MSE): 873,303,416.6874
* Root Mean Squared Error (RMSE): 29,551.7075
* Mean Absolute Percentage Error (MAPE): 10.2871

Support Vector Regressor:

* Mean Absolute Error (MAE): 59,556.7877
* Mean Squared Error (MSE): 7,859,371,053.7968
* Root Mean Squared Error (RMSE): 88,653.0939
* Mean Absolute Percentage Error (MAPE): 35.9659





**Summary and Recommendations:**

1. **Best Performing Model:** The Random Forest Regressor exhibited the best performance among the models evaluated, showing lower MAE, MSE, RMSE, and MAPE.
2. **Feature Importance:** Further analysis of feature importance in the Random Forest model may provide insights into the key factors influencing house prices.
3. **Outlier Detection:** Explore the impact of outliers on the models and consider outlier removal or transformation for improved performance.
4. **Hyperparameter Tuning:** Fine-tune hyperparameters for Random Forest and Gradient Boosting models to potentially enhance performance further.
5. **Additional Feature Engineering:** Explore additional feature engineering techniques to create new informative features that could improve model accuracy.
6. **Ensemble Methods:** Consider ensemble methods, such as stacking or blending, to combine predictions from multiple models for improved robustness.