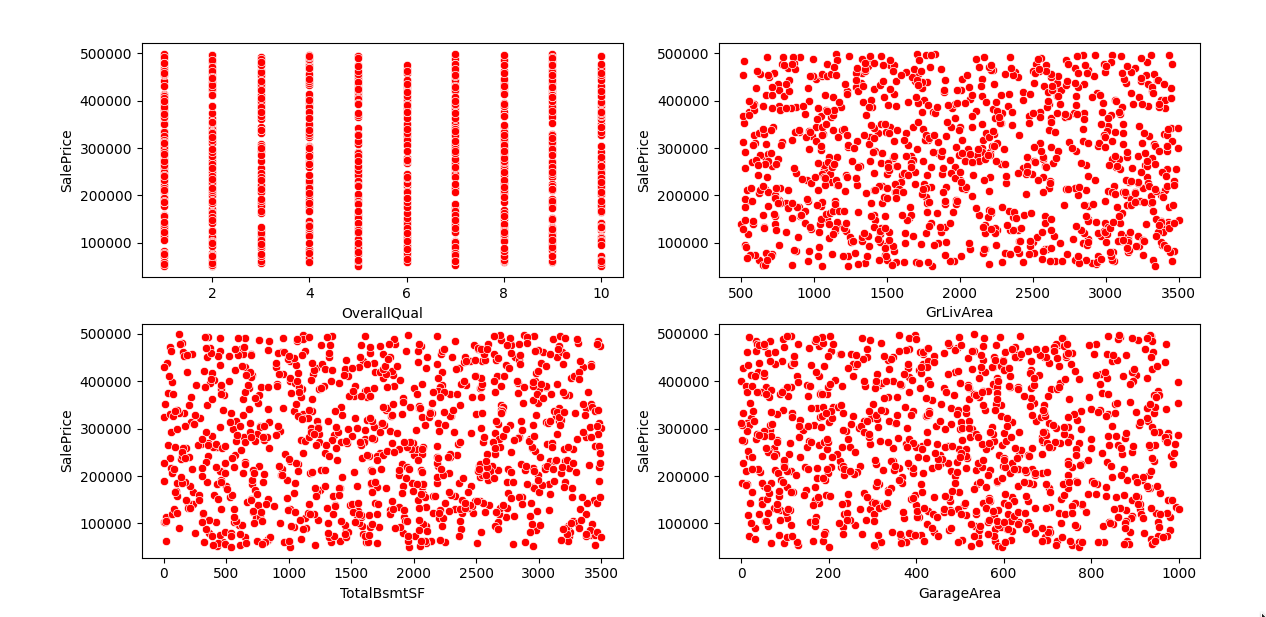
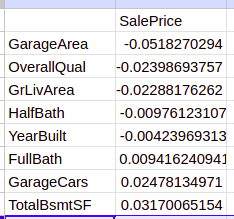
**Data Analysis:**

1. There are no null values or missing values in the given dataset. There is no requirement for data imputation

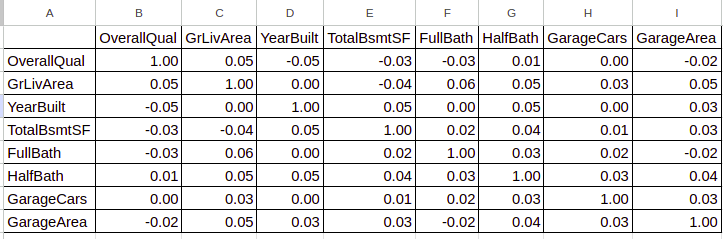
**SCATTERPLOT BETWEEN SALES PRICE AND INDEPENDENT VARIABLES:**  Scatter Plot showed that there is no correlation between any of the independent variables and sales price (dependent variable). These seem random data and model training might not help to identify the patterns for training and prediction, but continuing with the exercise.



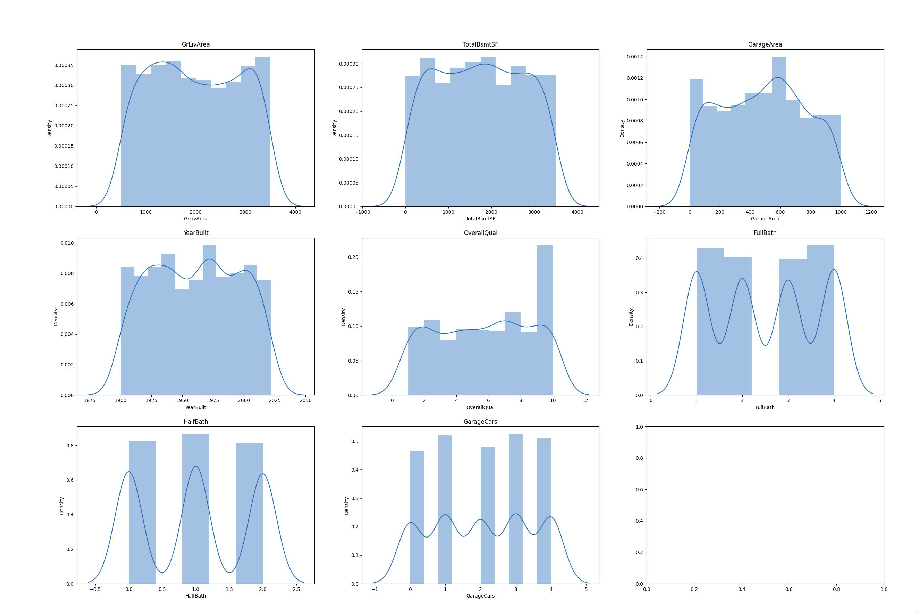
All the correlations between the target variables and independent variables are minimal:



Correlation within independent variables are minimal:



**DISTRIBUTION PLOT** shows that most of them are of either multi modal or uniform distributions.



**Data preparation:**

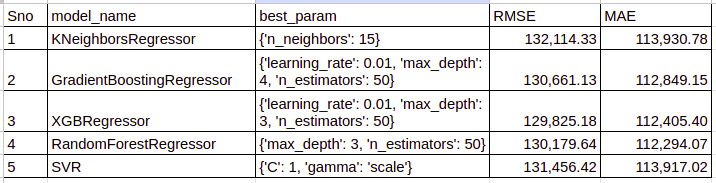
OverallQual, FullBath, HalfBath, GarageCars are considered as categorical variables and dummy encoded in data\_transform.py.

Data normalization is done by minmaxscaler (range standardization) method to have standardized weight updates across all the features. Outliers samples are removed from the Isolation forest method. Code folder structure and their purpose is given in the README.MD file.

**Model Experimentation based on RMSE and MAE Metric:**

Tried Regression algorithms like gradient boosting regressor, Random forest Regressor, Knn Neighbor regressor etc. models were experimented, PFB result:

XGB regressor performed well when compared to the regression models and the best parameter is based on the GridsearchCV algorithm.



XGB regressor is the best model considered for the regression model.

**Future Scope:**

1. Improve data quality for the given use case.
2. Explore with more data transformation & model techniques for better model performance.
3. Better feature engineering techniques to create relevant features for model performance improvement.