Nick Tran

Write up:
The analysis of the dataset reveals several interesting findings and correlations among the variables. Let's explore each observation in detail.

'OverallQual', 'GrLivArea', and 'TotalBsmtSF' exhibit a strong correlation with 'SalePrice'. This implies that the overall quality of a house, the above-ground living area, and the total basement area have a significant impact on the sale price of a property.

'GarageCars' and 'GarageArea' are correlated variables. However, the number of cars that can fit into a garage is dependent on the garage area. Therefore, one of these variables can be removed to avoid redundancy. To determine the correlation between them, further analysis is required.

'TotalBsmtSF' and '1stFlrSF' show a high correlation. Since these variables provide similar information about the size of the house, it is possible to drop one of them to avoid multicollinearity issues.

'YearBuilt' and 'TotRmsAbvGrd' are highly correlated. In this case, it is suggested to discard 'TotRmsAbvGrd' since it provides redundant information compared to 'YearBuilt'.

Based on the Kaiser-Meyer-Olkin (KMO) model value, which is greater than 0.5, we can conclude that the dataset is suitable for factor analysis. This indicates that the variables in the dataset are interrelated and can be represented by unobserved factors.

By examining the eigenvalues, it is found that only two values are above 1. Therefore, it is determined that only one factor or unobserved variable should be chosen for further analysis.

Additionally, the Scree plot supports the decision to select only one factor, as there are only two values after the elbow point. This ensures a concise representation of the data while capturing the most significant information.

Further exploration of the data reveals interesting patterns. The scatter plot of 'TotalBsmtSF' and 'GrLivArea' shows a distinct line below which most 'TotalBsmtSF' values fall. This aligns with the expectation that the basement area is usually smaller than the living area. The exponential increase in 'SalePrice' with respect to 'YearBuilt' in recent years suggests that newer houses tend to have higher prices.

It can be concluded that the distribution of 'SalePrice' varies across different neighborhoods, making it a potentially valuable predictor of neighborhood characteristics.

To gain a better understanding of the data, swarmplots (similar to boxplots) are utilized. These plots not only display the distribution of numerical variables but also show the number of observations at each value. A denser swarmplot indicates a higher concentration of data points at a particular value.

In summary, the analysis provides insights into the relationships and correlations among the variables in the dataset. By considering the suggestions for removing redundant variables and utilizing factor analysis, we can achieve a more concise representation of the data while capturing the essential information. The identified patterns in the scatter plot and the importance of 'SalePrice' in predicting neighborhood characteristics further contribute to the overall understanding of the dataset.