Variables correlation

Here are some of the most stronger correlation between variable for the dataset Train.csv, please grab the file from Kaggle (https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/code). For each one of the correlations mentioned in this document, we created a plots that briefly describe the correlation between the variables and also analyzed the distribution of the dataset.

It's worth noting that there are also many variables that are highly correlated with each other, which could lead to issues of multicollinearity if all of them were included in a linear regression model. Therefore, it may be necessary to remove some of these variables or use a method like principal component analysis to reduce the dimensionality of the dataset.

We can see that there is correlation between the variables in the dataset, there many variables that are highly correlated with each other, for example we see a correlation between sale price and the above grade living area and first floor square feet.
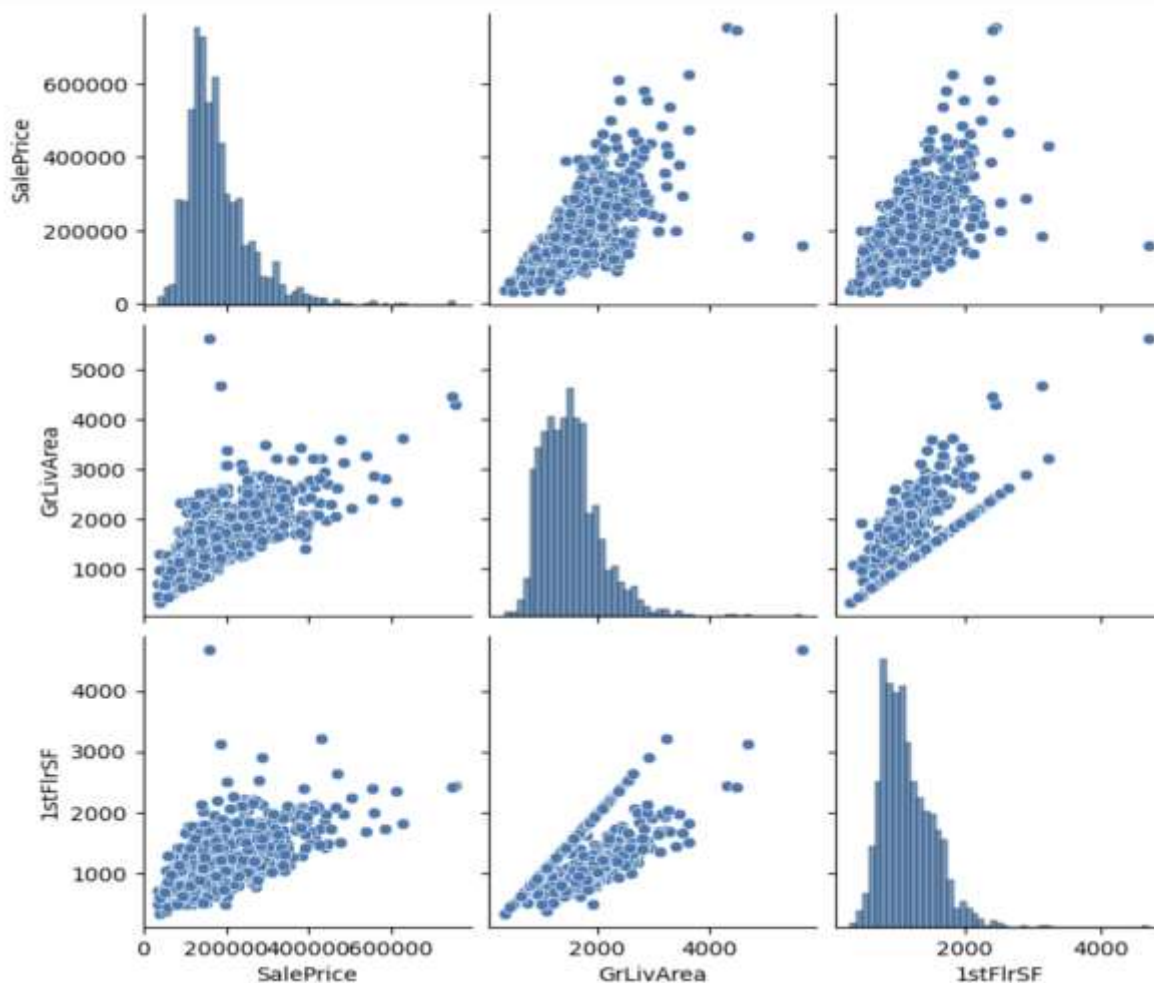
Here it is the distribution of each variable in the dataset, we can see that the result of plotting them show a negative, positive, and distribution with zero skewness. We can look for variables that are normally distributed (bell-shaped curve), variables that are skewed (have a tail on one side), and variables that have a bimodal distribution (two peaks).
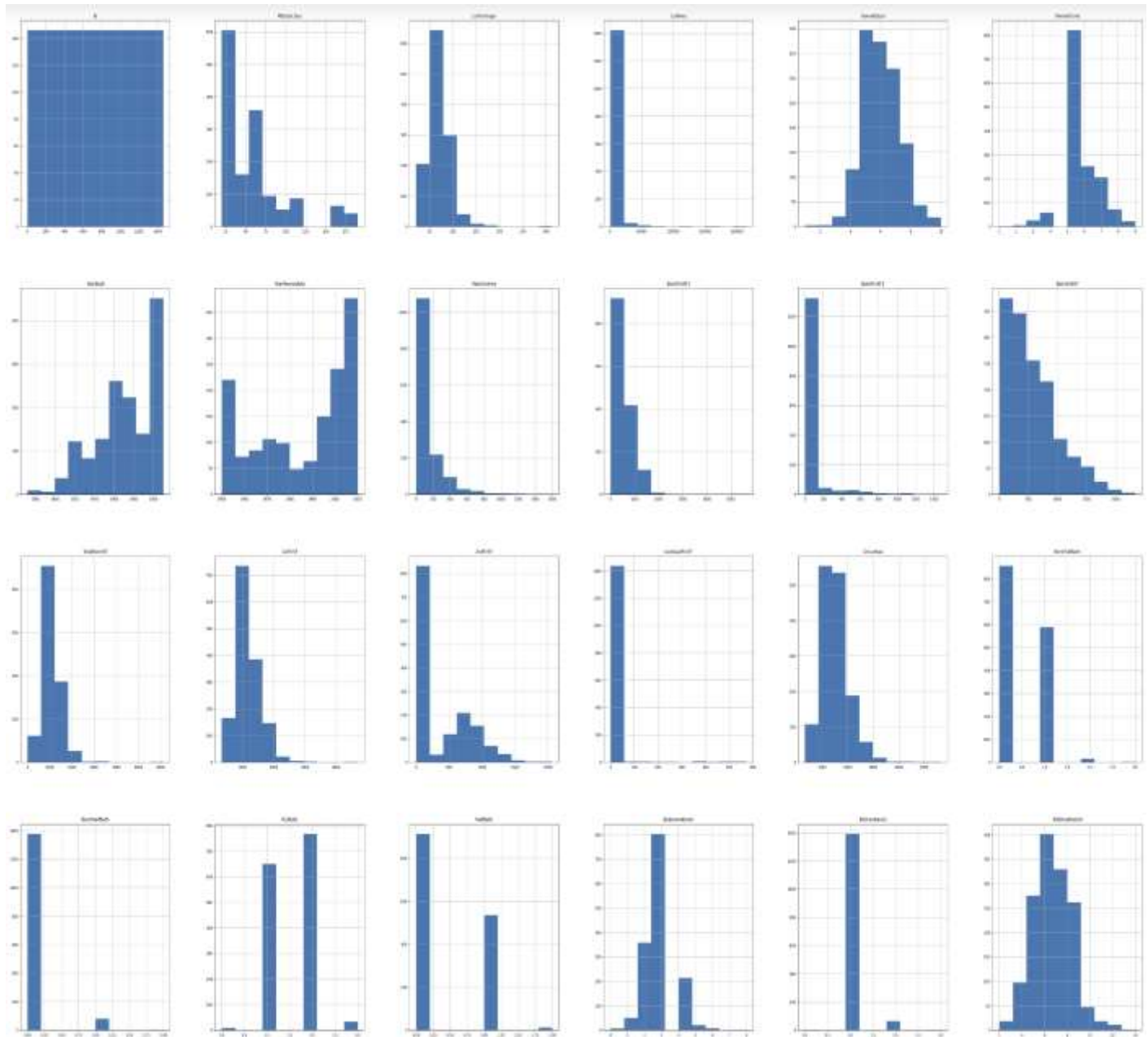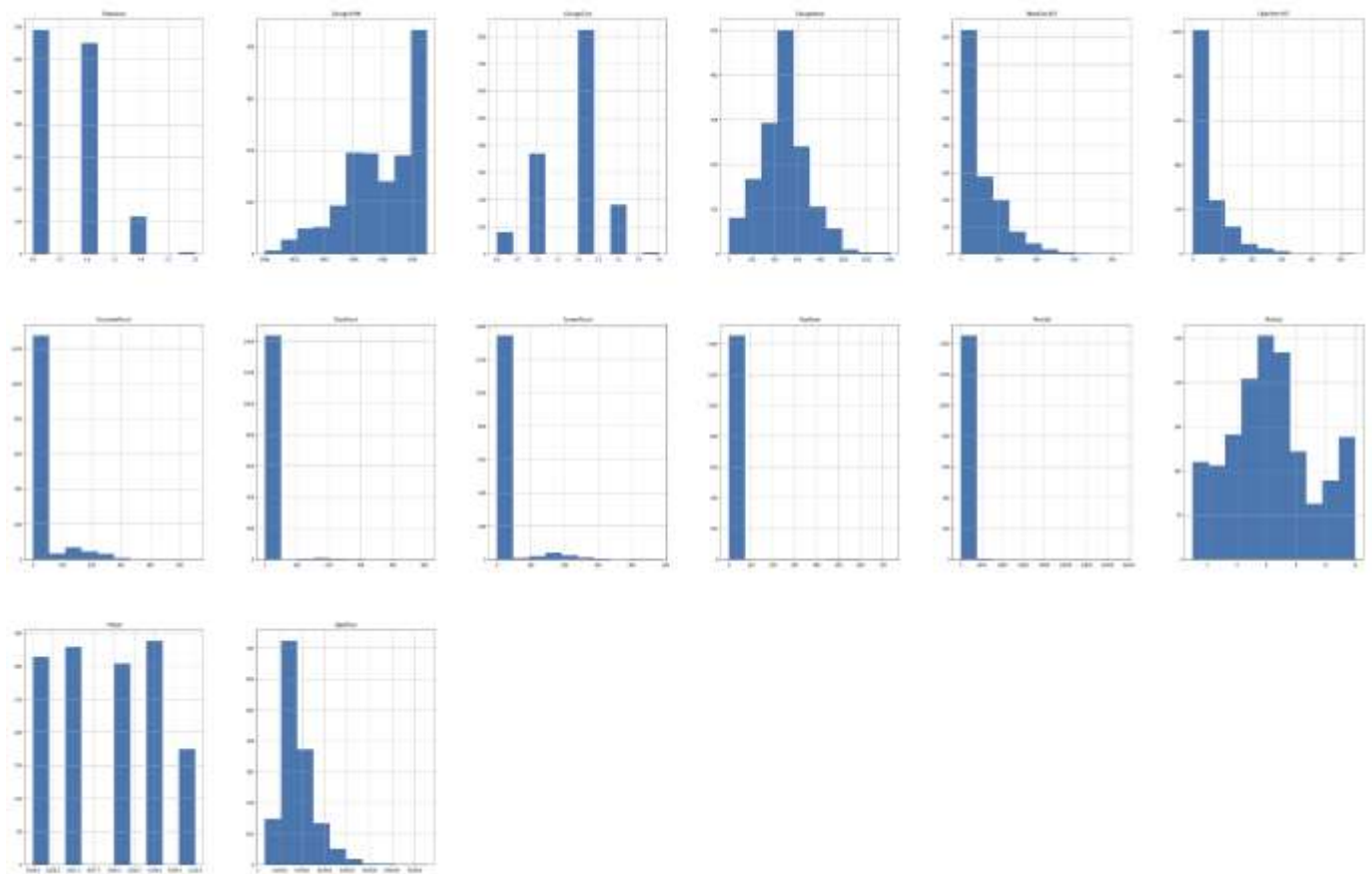
** Keep in mind that some variables may have missing values or outliers that can affect the distribution.