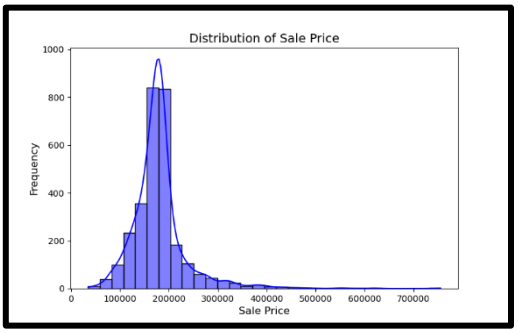
Exercise 3

The Ames Housing dataset was compiled by Dean De Cock for use in data science education. There are 2919 observations and 81 variables, and it is divided by 12 floats, 26 integers, and 43 objects.

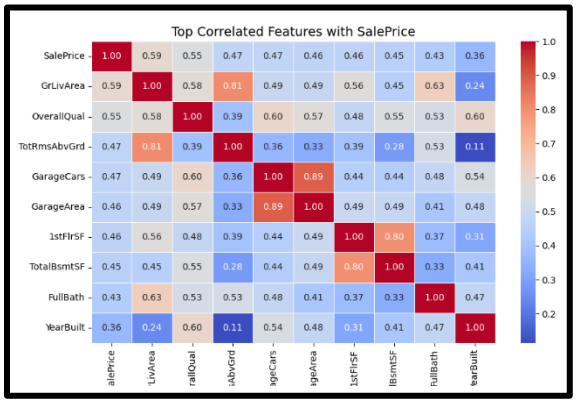
To summarize, there are too many missing values in this data. Therefore, I separate into two types which is numerical variables and categorical variables. Firstly, I drop columns with more than 40 percent of missing values. Then, I impute missing numerical values with the median and impute missing categorical values with the mode. After computing the missing values, there are 75 variables which are 38 numerical variables and 37 categorical variables.



From housing dataset, sale price is the dependent variable. Thus, I decide to plot the histogram of sale price to see the distribution of this variable.

The histogram of sale price shows the right-skewed distribution because most data is on the left, with a long tail on the right.

To see the correlation between the variables, I compute all numeric variables in correlation matrix and get the top 10 variables which are correlated with sale price.



The heat map above provides information about top 10 correlated features with sale price. The size of garage in car capacity has a positively strong relationship with the size of garage in square feet which is 0.89. However, the variable which is the highest correlation with the sale price is the above grade (ground) living area square feet (0.588).

Patcharapa Pinnarat 100434786