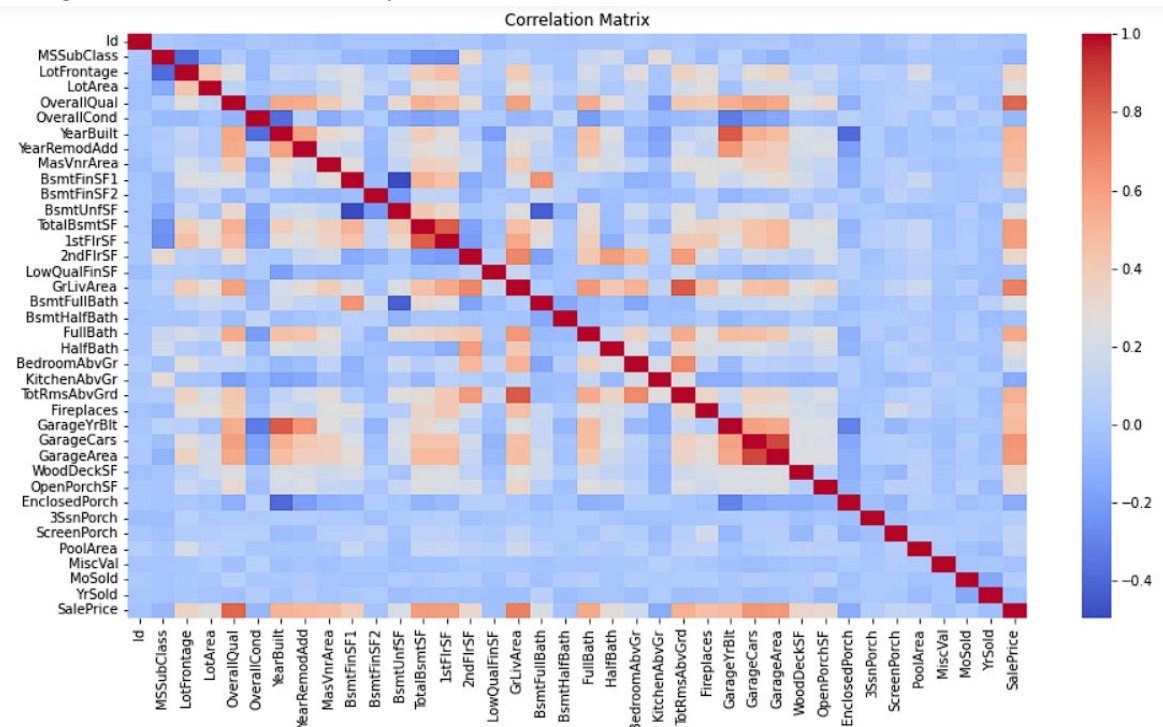


The train.csv dataset has 1460 rows and 81 columns.

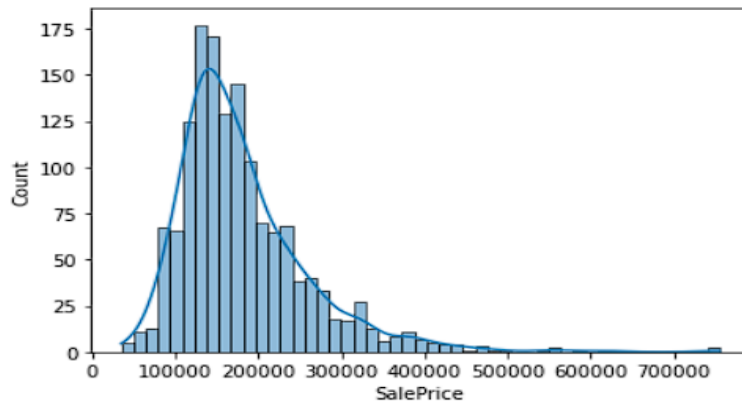
There are 259 null rows in 'LotFrontage' column. The missing values in 'LotFrontage' column may be due to the fact that the frontage measurements were not recorded or were missing for some houses in the dataset. It is also possible that the missing values may be due to data entry errors or other issues with data collection or processing. We may choose to impute the missing values using an appropriate method or drop the column altogether.

1. Looking at the correlation heatmap for numerical variables:



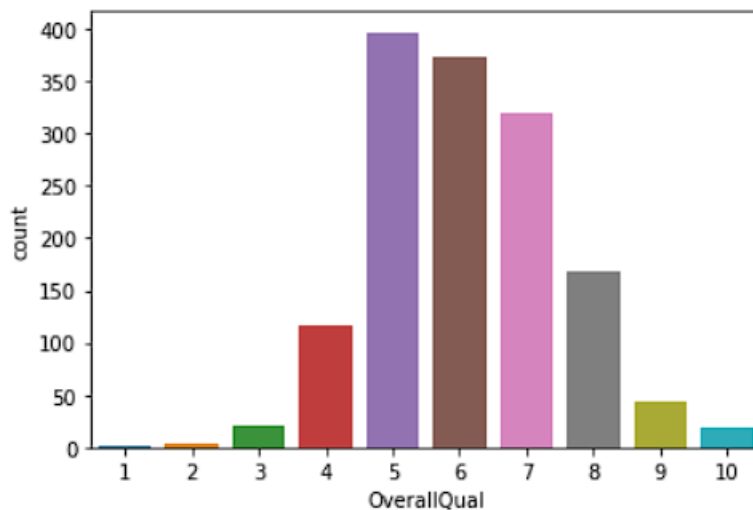
By looking at the heatmap, we can identify which variables are highly correlated with each other. We can see that the variables 'GarageArea' and 'GarageCars' are highly correlated with each other, with a correlation coefficient of more than 0.8. This makes sense because larger garages tend to be able to accommodate more cars. We can also see that the variable 'OverallQual' is highly correlated with the target variable 'SalePrice', with a correlation coefficient of almost 0.8. This suggests that the overall quality of a house is a good predictor of its sale price.

2. Distribution of the target variable:



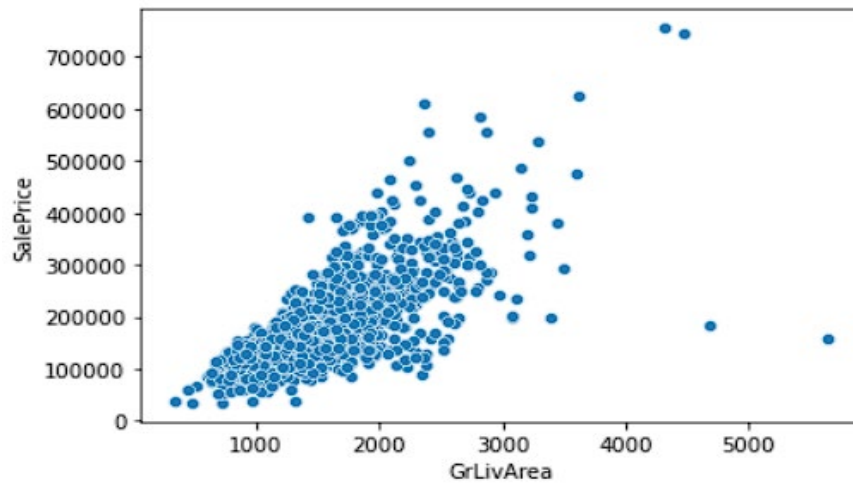
The histogram is skewed to the right, with a long tail on the right side, this suggests that there are some houses with very high sale prices, which may be due to factors such as location, size, or amenities. On the other hand, if the histogram is roughly symmetric, this suggests that the sale prices of houses are more evenly distributed, with fewer extreme values.

3. Distribution of categorical variables:



This variable represents the overall material and finish quality of the house, rated on a scale of 1 to 10. The countplot shows the frequency of each rating value in the dataset. By visualizing the distribution of the OverallQual variable, we can gain insights into the quality of the houses in the dataset. In this case, most of the houses have medium (neither high nor low) quality ratings, this suggests that the houses in the dataset are generally of average quality, which may be due to factors such as location, construction materials, or design.

4. Scatterplot to see relationship between 2 variables:



We can see that there is a positive correlation between the 'GrLivArea' and 'SalePrice' variables. The points on the scatter plot generally form an upward-sloping trend, indicating that houses with larger living area above ground level tend to have higher sale prices.