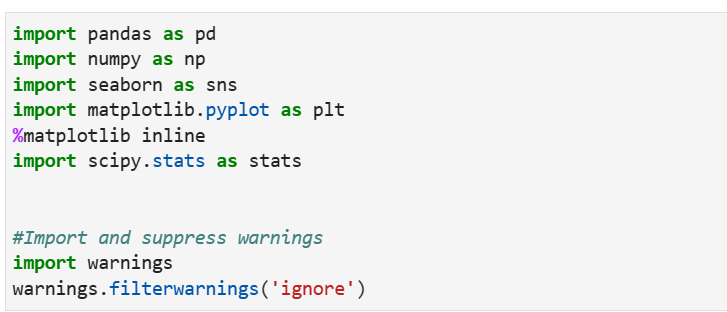
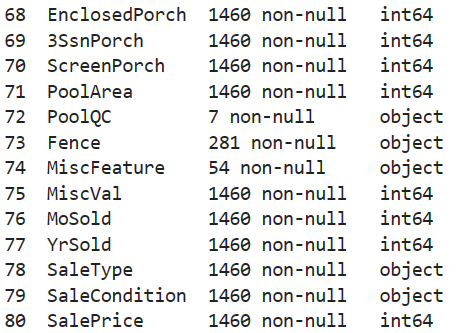
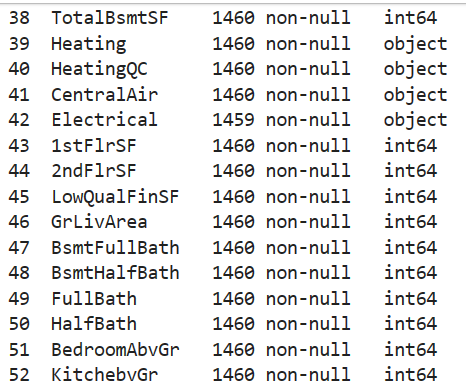
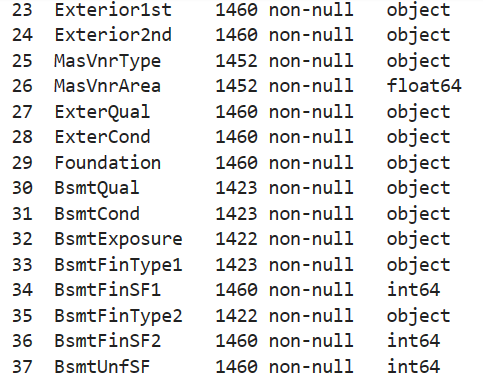
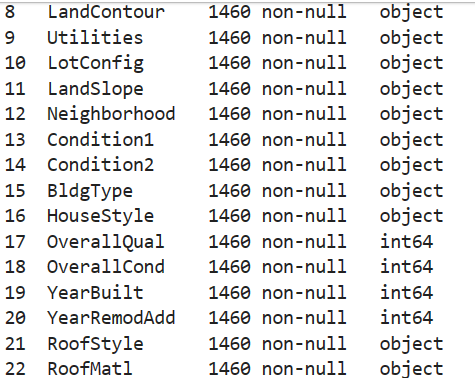
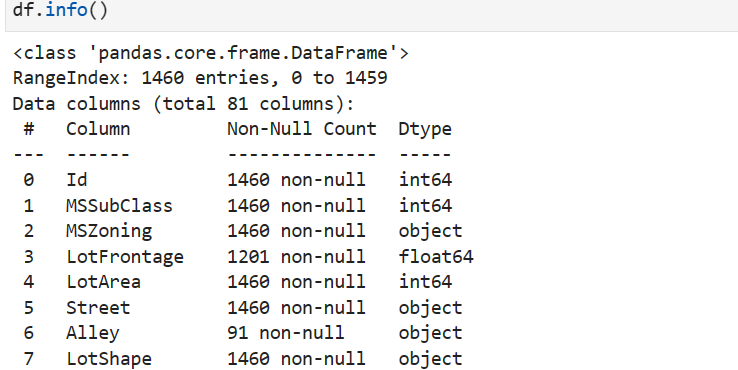
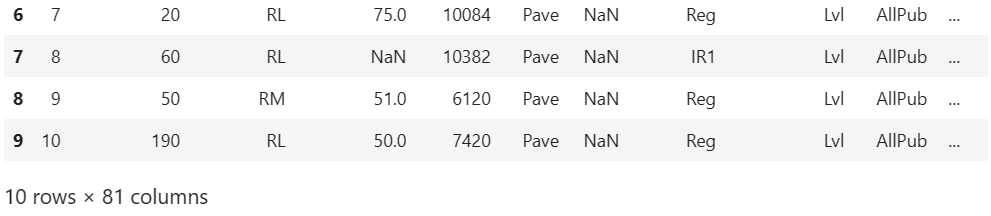
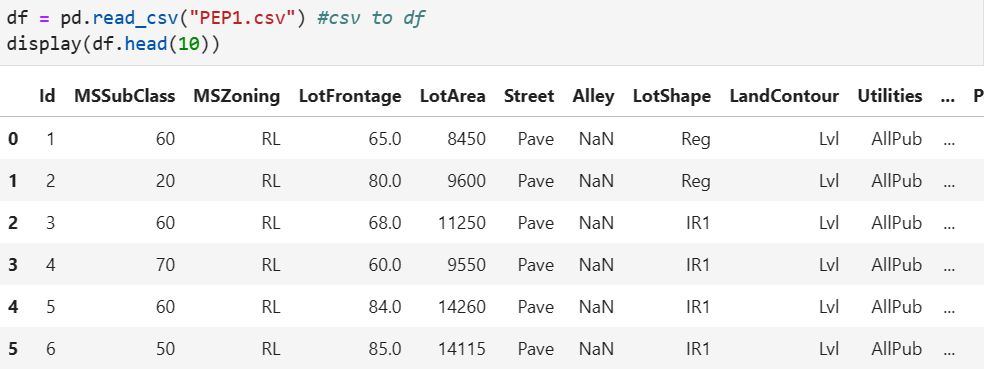
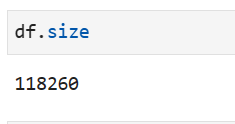
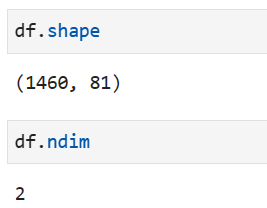
**Project 1: Feature Engineering**

**Note:**

1. Download the “PEP1.csv” using the link given in the Feature Engineering project problem statement
2. For a detailed description of the dataset, you can download and refer to data\_description.txt using the link given in the Feature Engineering project problem statement

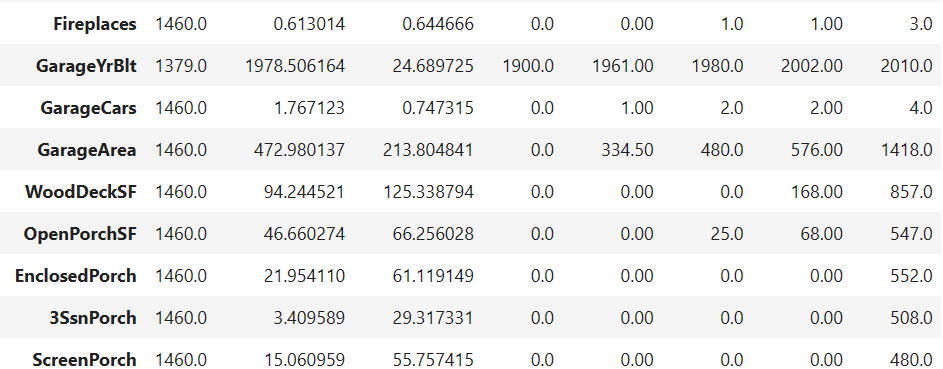
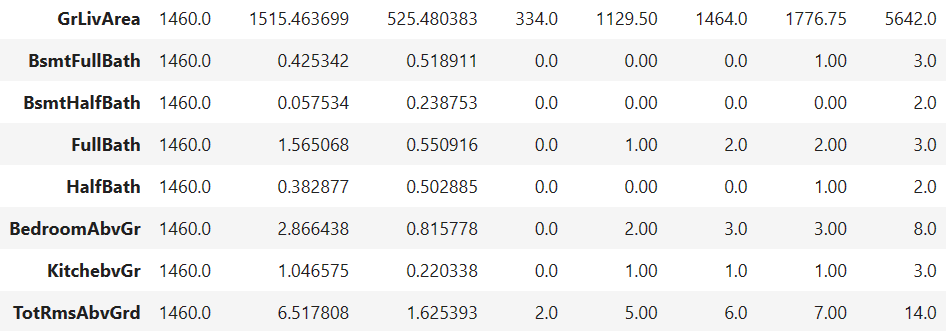
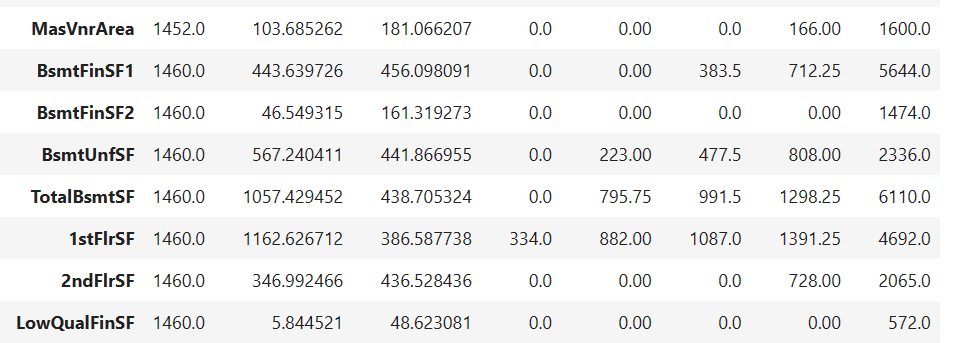
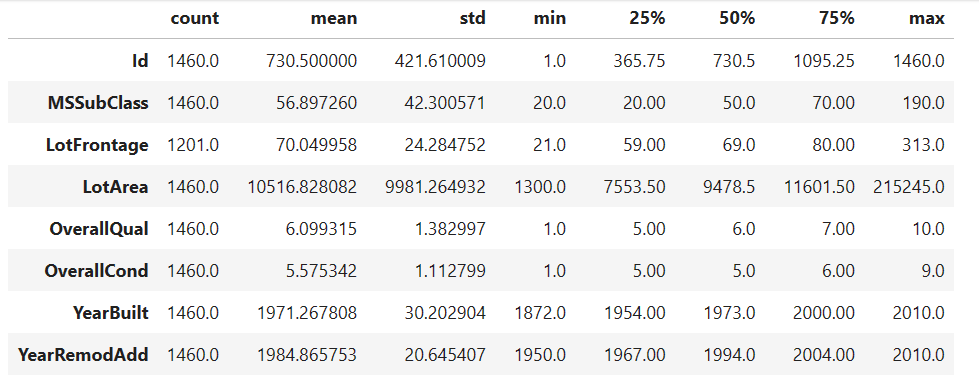
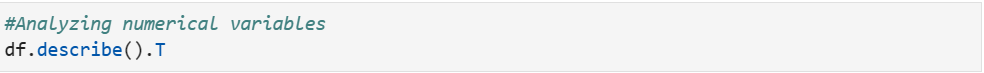
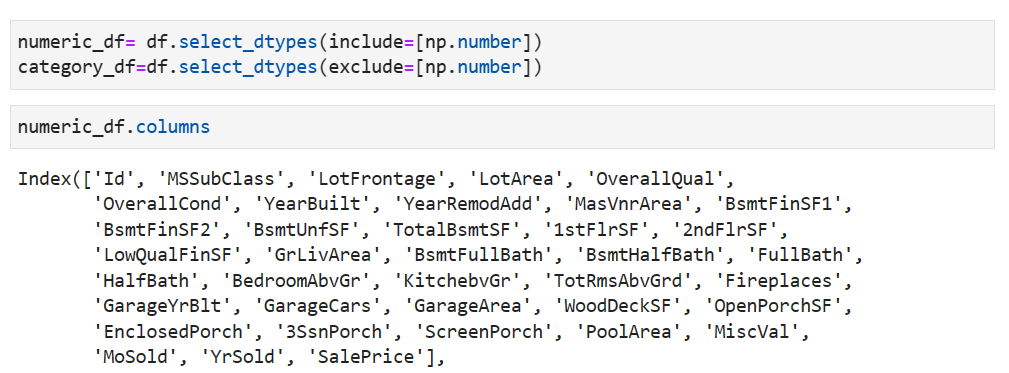
**Perform the following steps:**

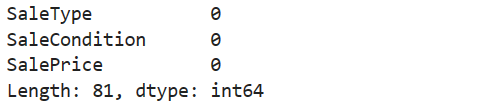
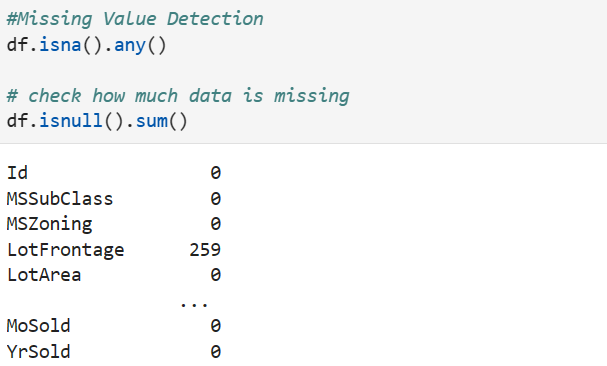
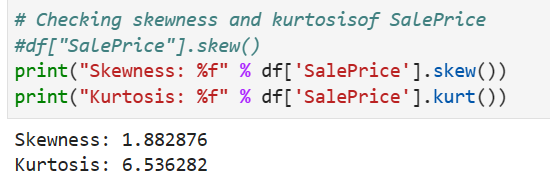
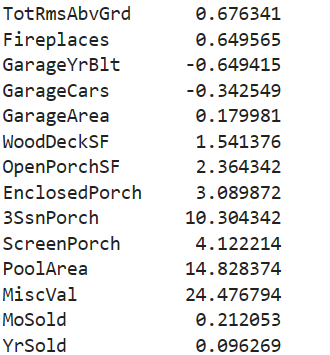
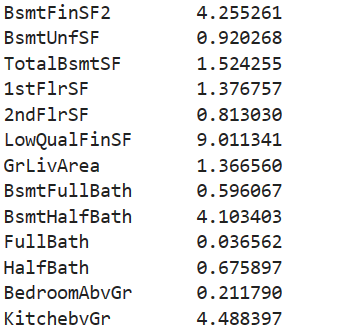
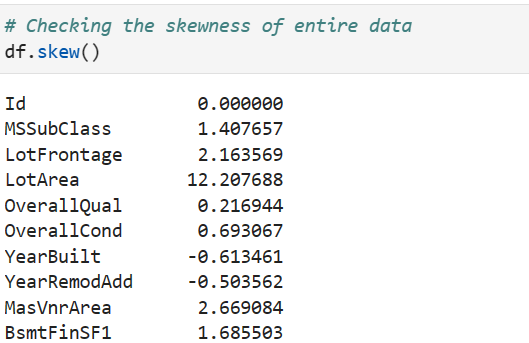
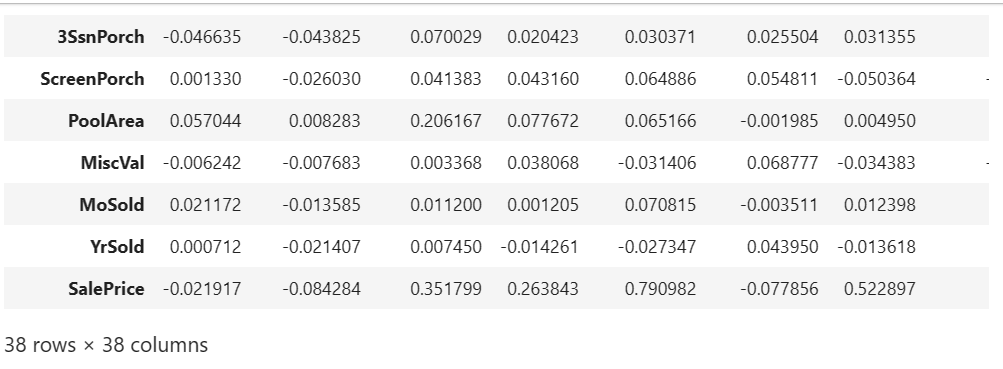
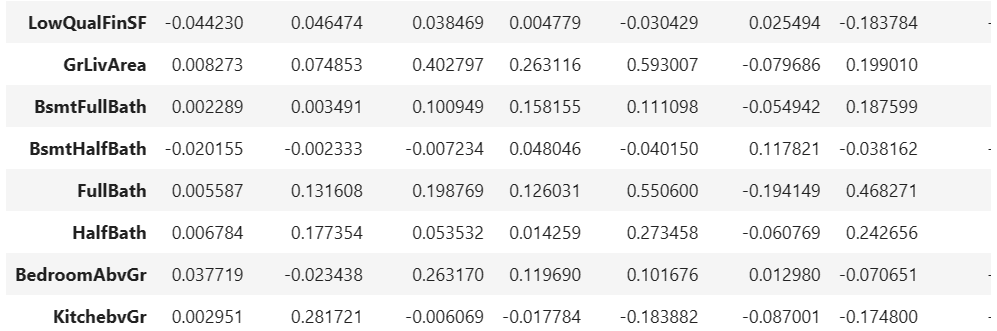
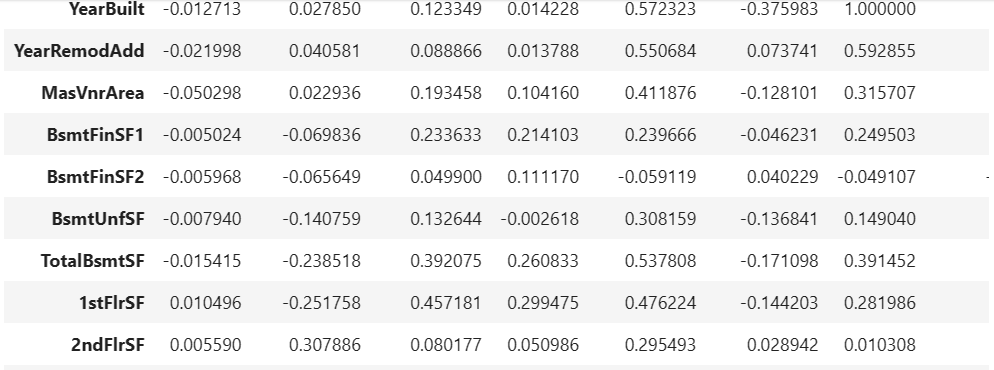
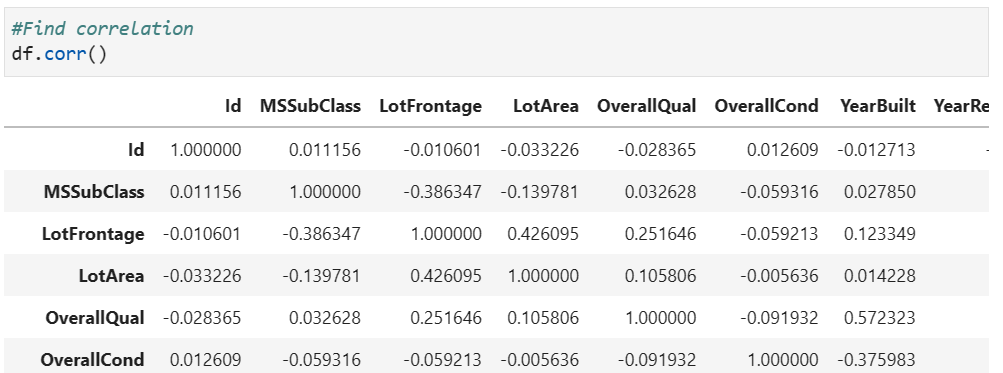
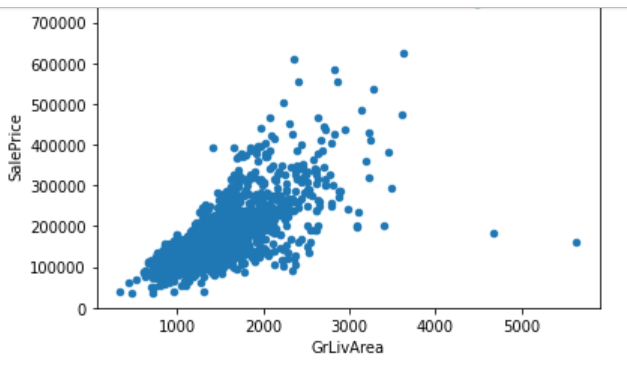
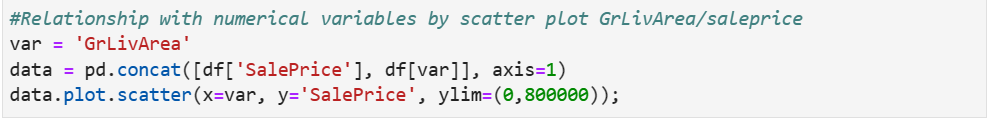
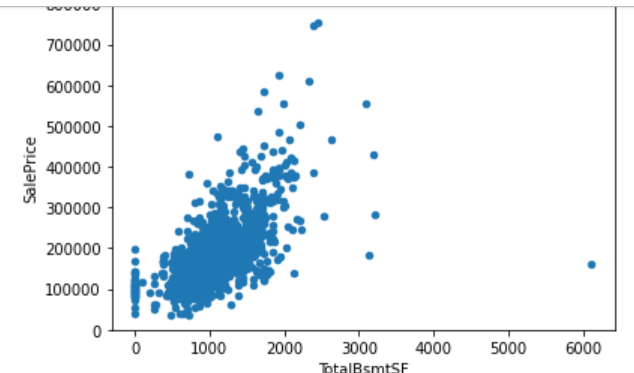
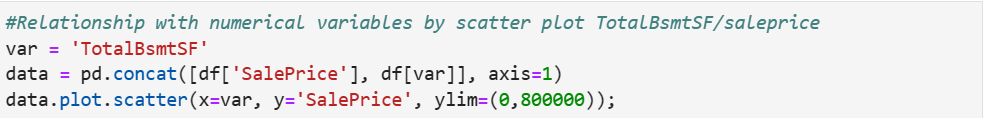
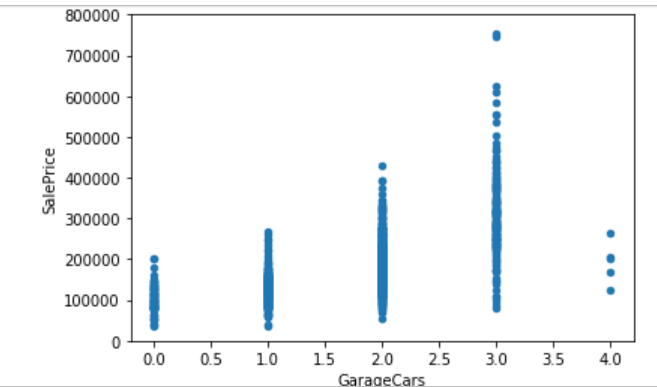
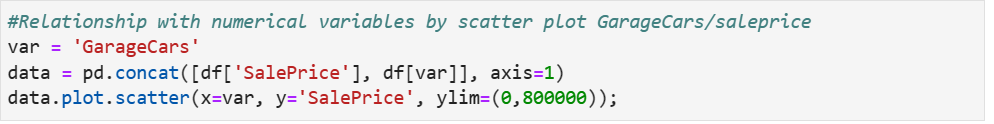
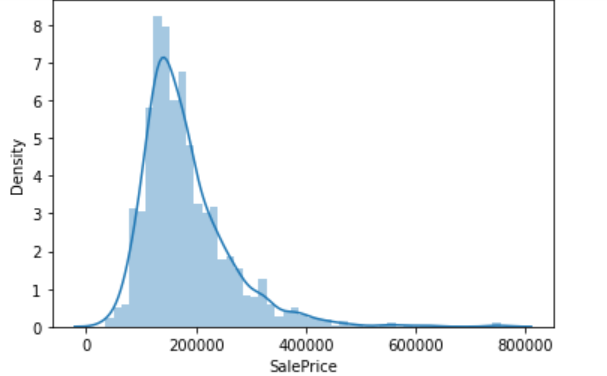
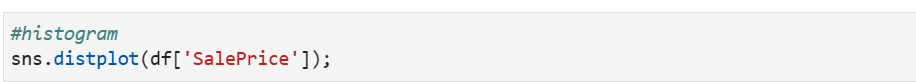
1. Understand the dataset:
2. 
3. 
   1. Identify the shape of the dataset  
      

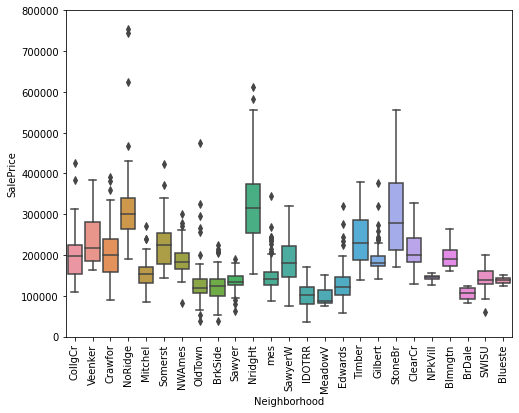
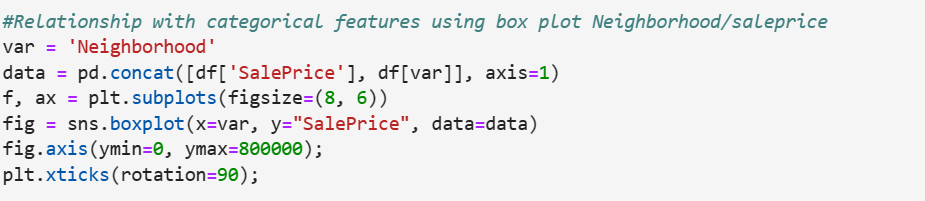
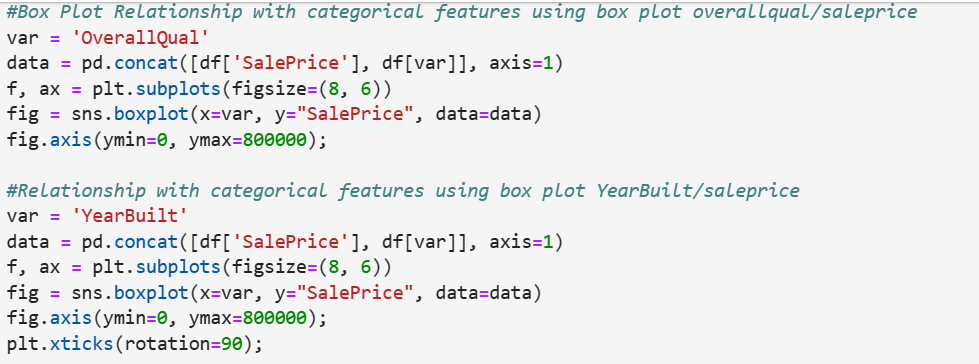


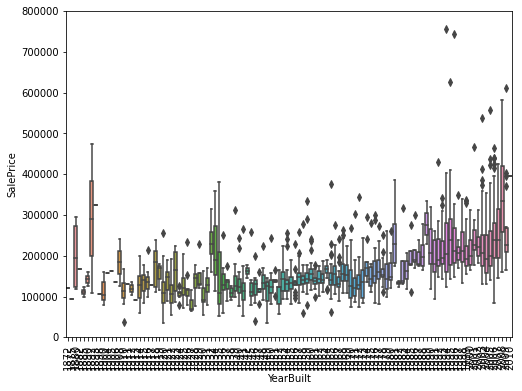
* 1. Identify variables with null values
  2. Identify variables with unique values

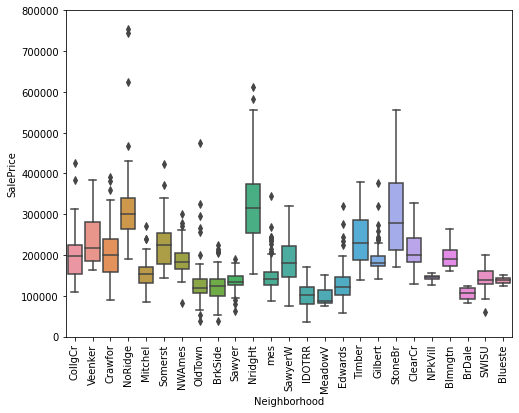
1. Generate a separate dataset for numerical and categorical variables



1. EDA of numerical variables:
   1. Missing value treatment  
      
   2. Identify the skewness and distribution
   3. Identify significant variables using a correlation matrix 
   4. Pair plot for distribution and density
2. EDA of categorical variables
   1. Missing value treatment
   2. Count plot and box plot for bivariate analysis
   3. Identify significant variables using p-values and Chi-Square values

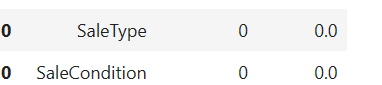
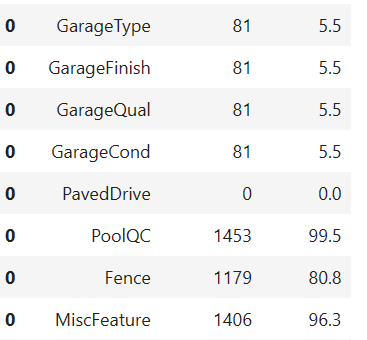
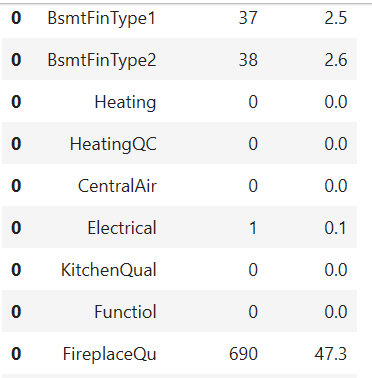
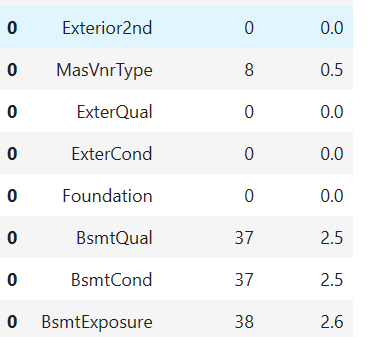
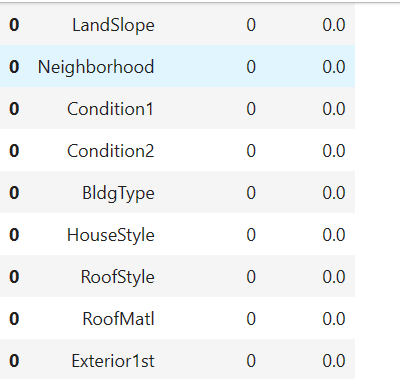
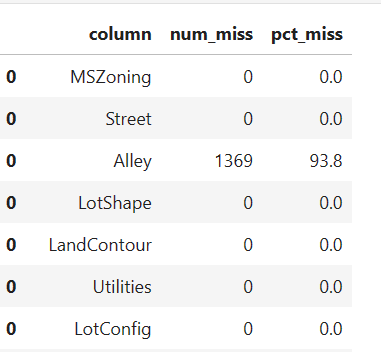
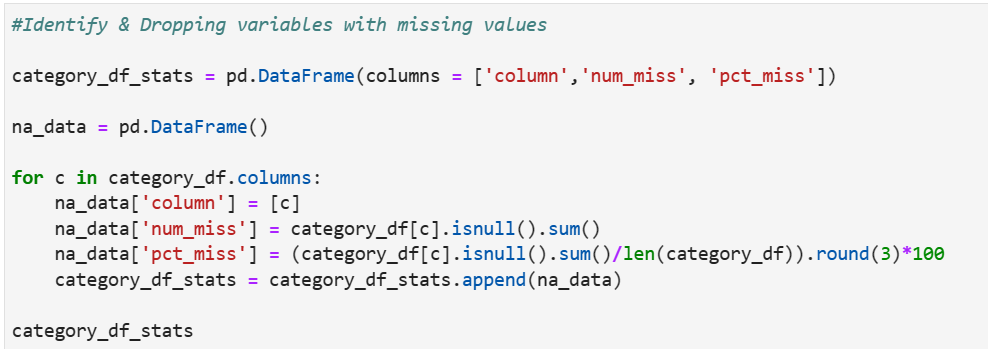
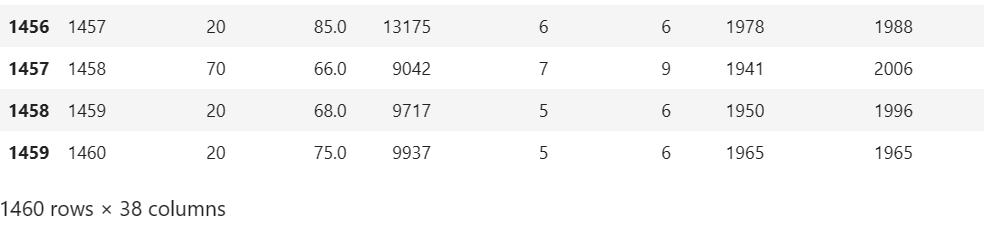
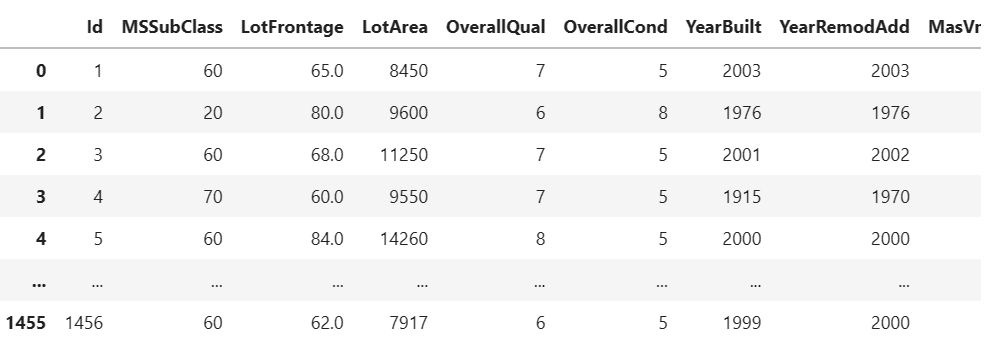




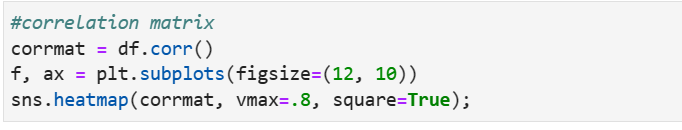
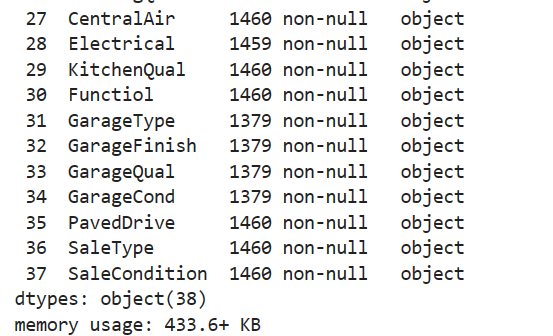
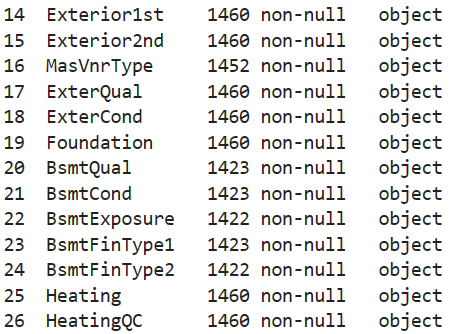
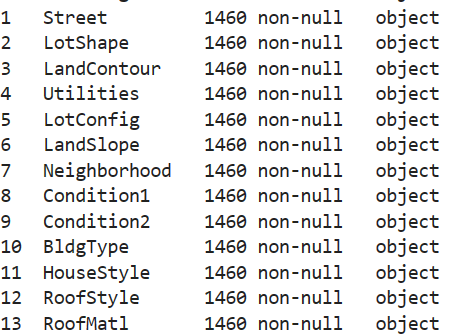
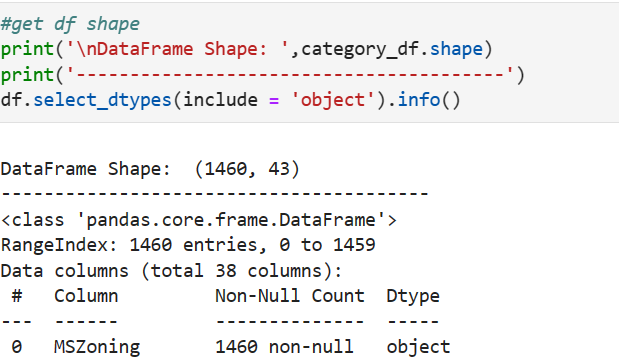
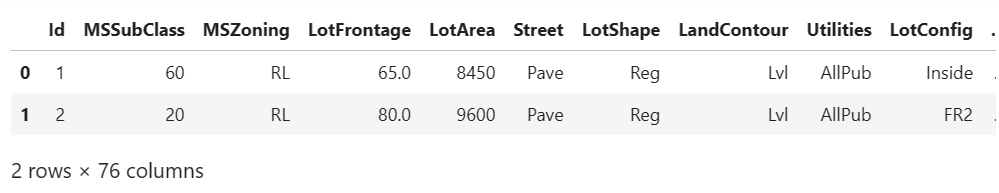
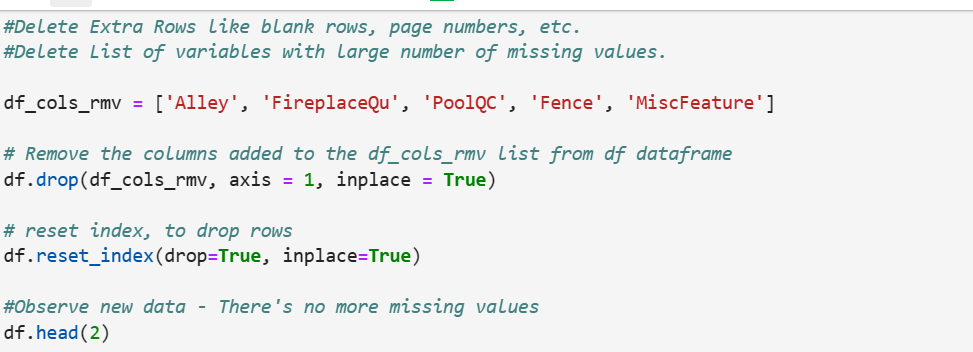


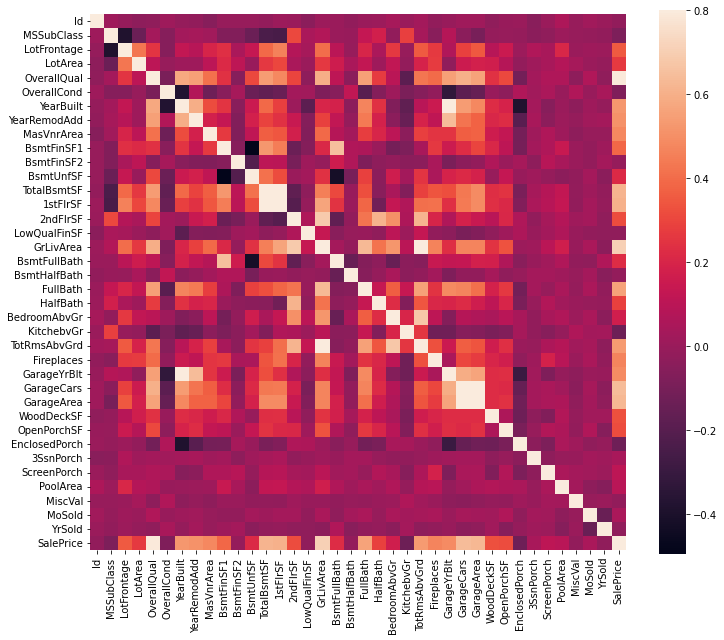
GrLivArea' and 'TotalBsmtSF' seem to be linearly related with 'SalePrice' The box plot shows Sales prices increase with the Overall quality and throughout the Years Built. Sales prices is the highest in Neighbourhood NoRidge.

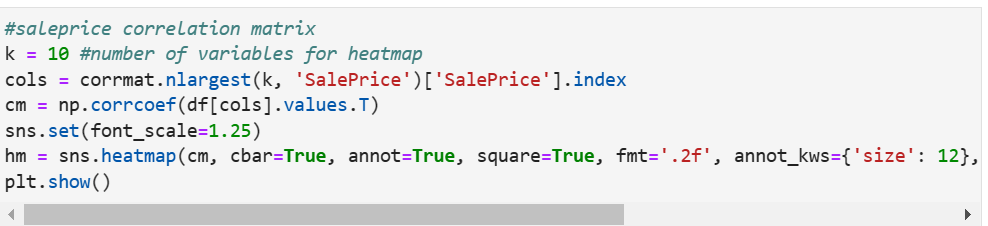
1. Combine all the significant categorical and numerical variables

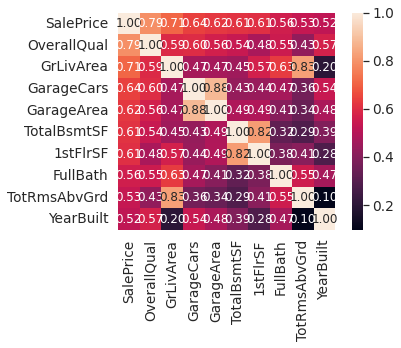


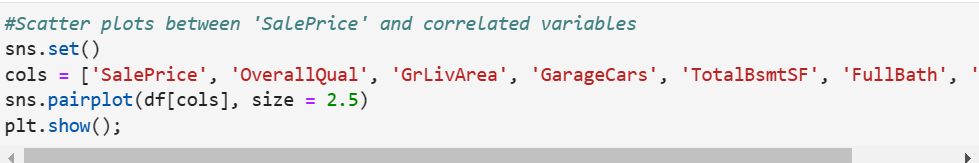
1. Plot box plot for the new dataset to find the variables with outliers

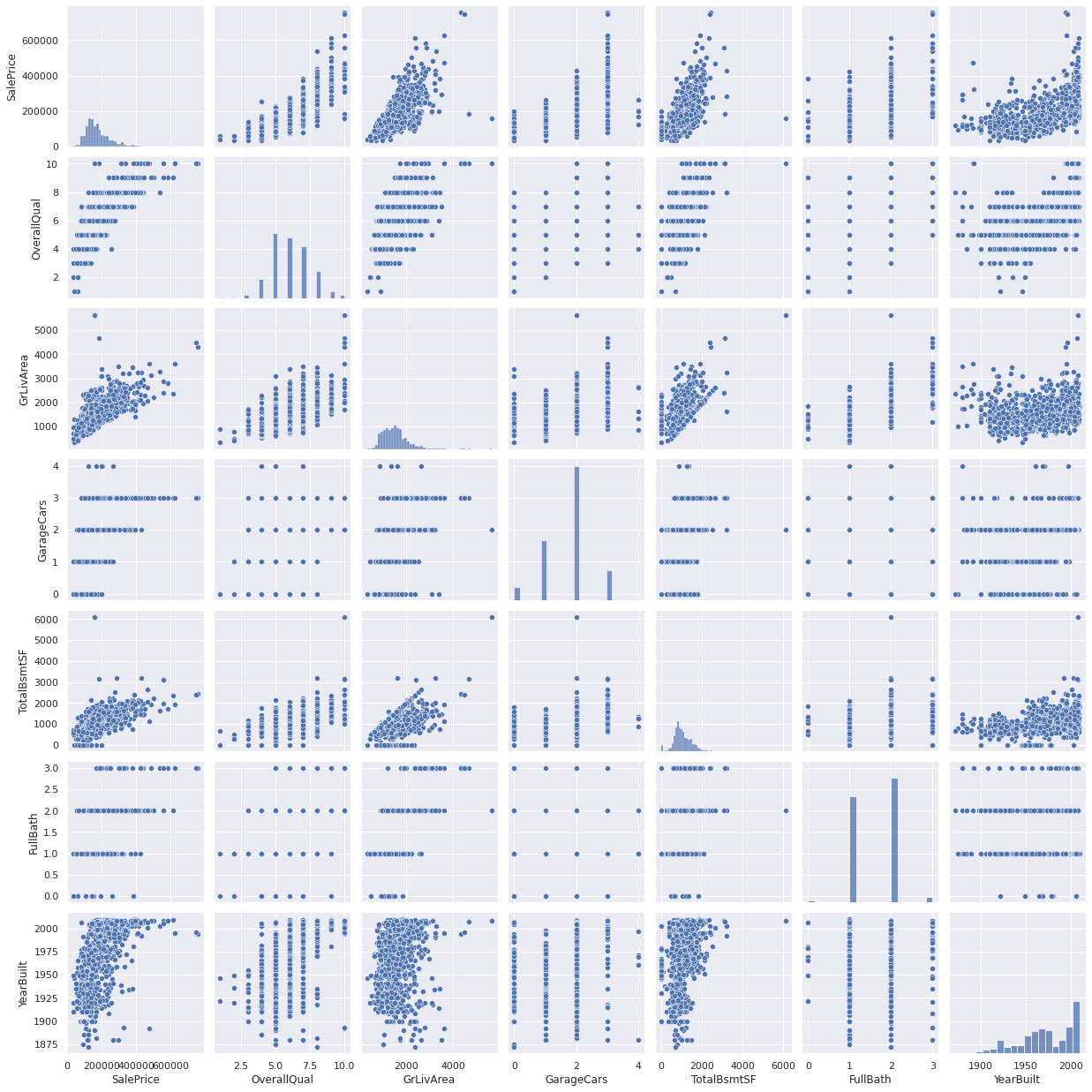


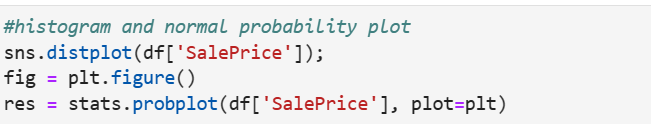


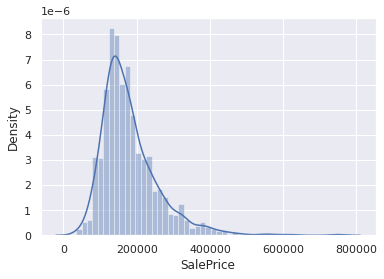


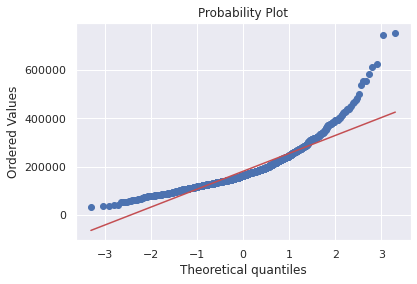


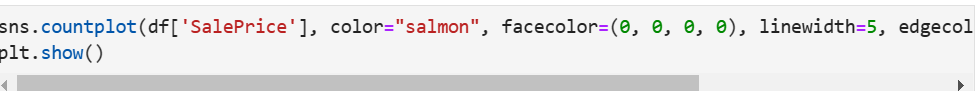


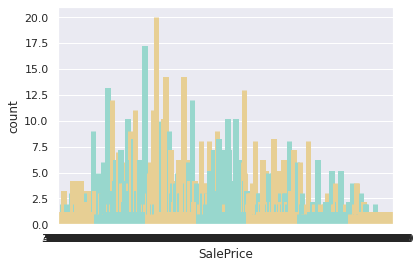


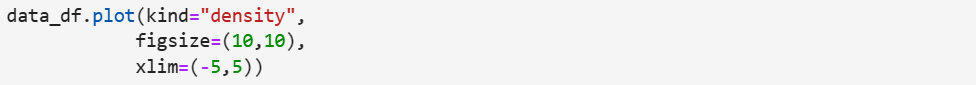
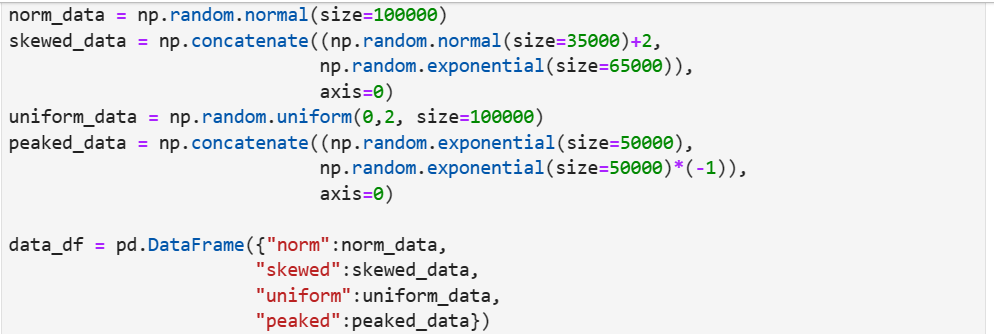
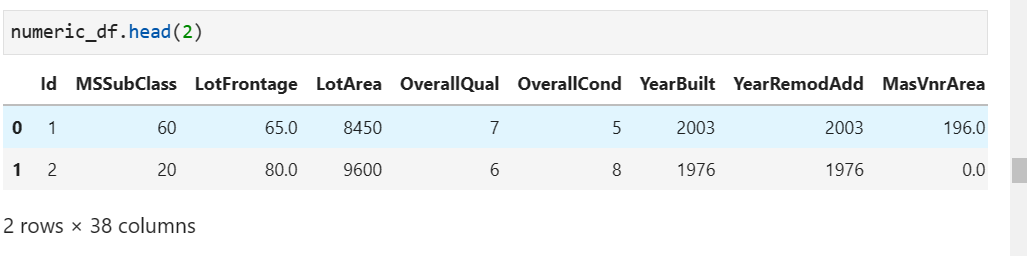
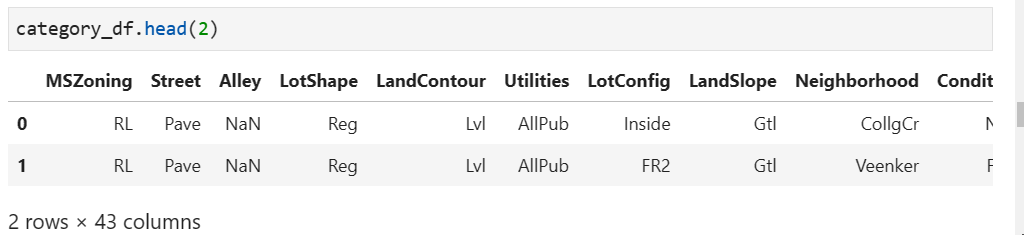






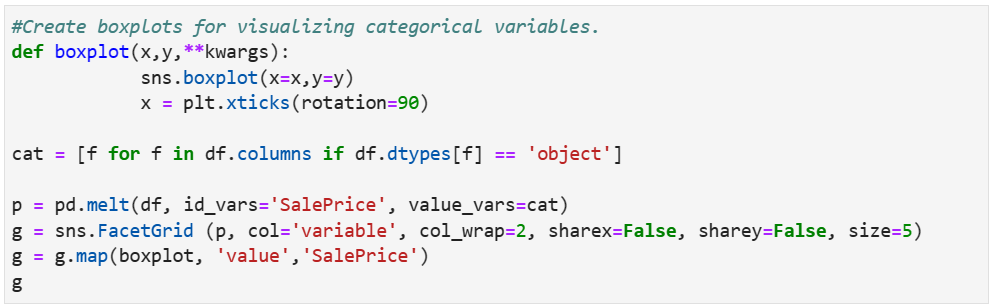
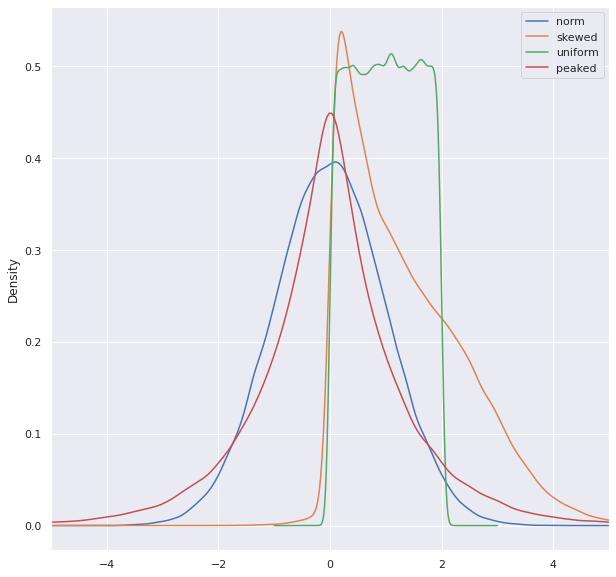




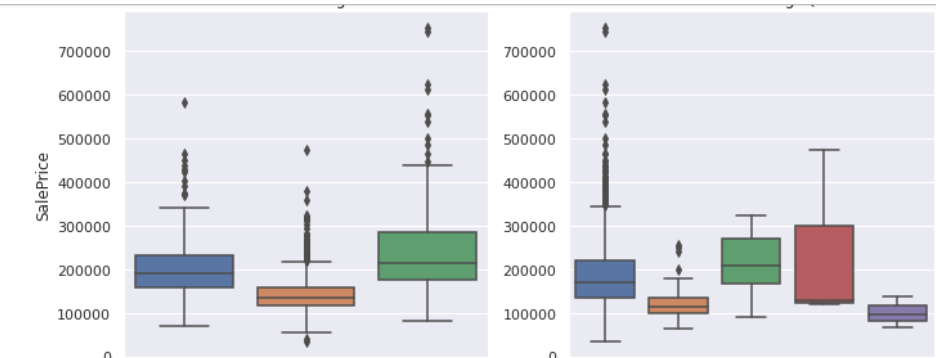
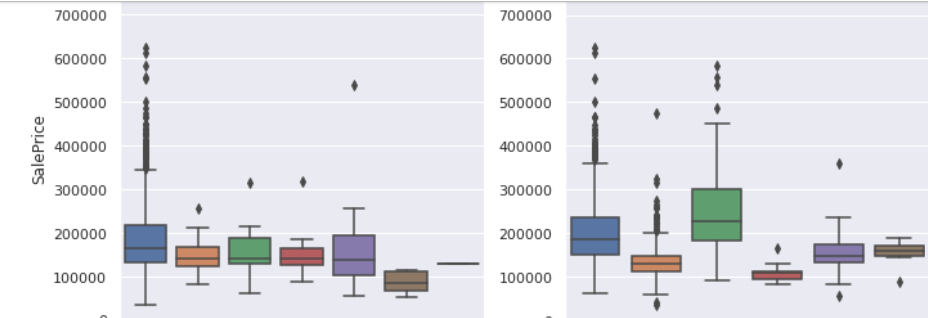
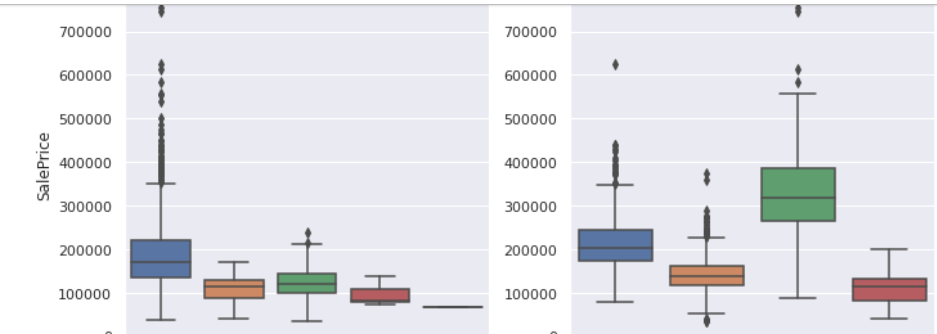
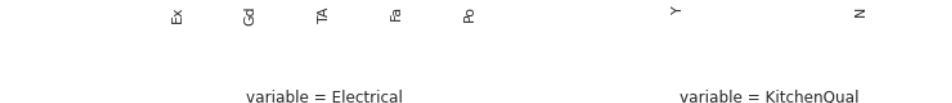
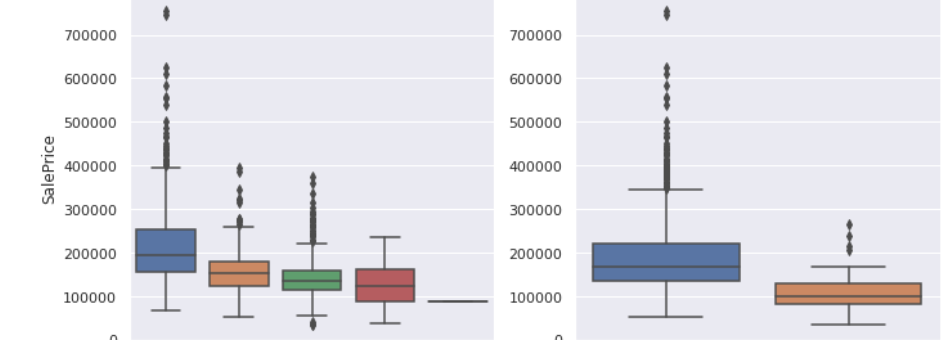
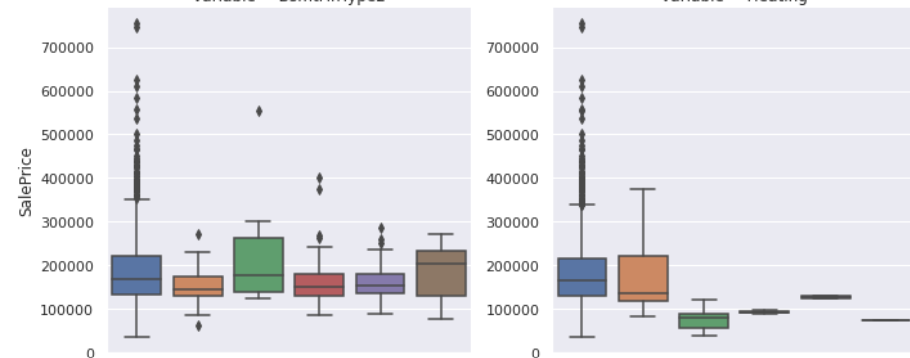
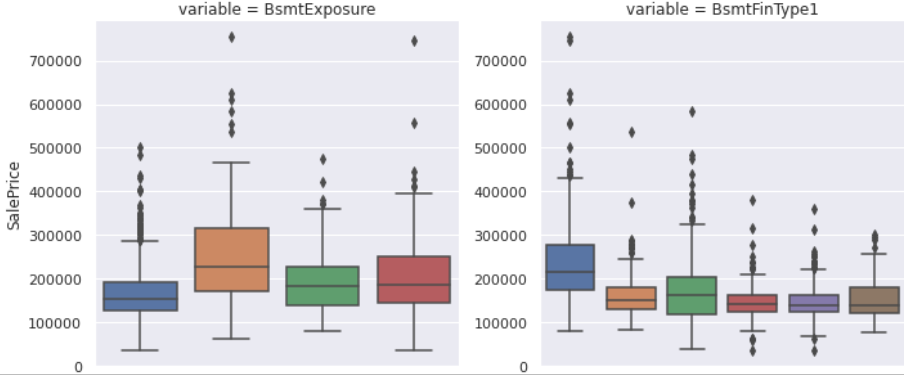
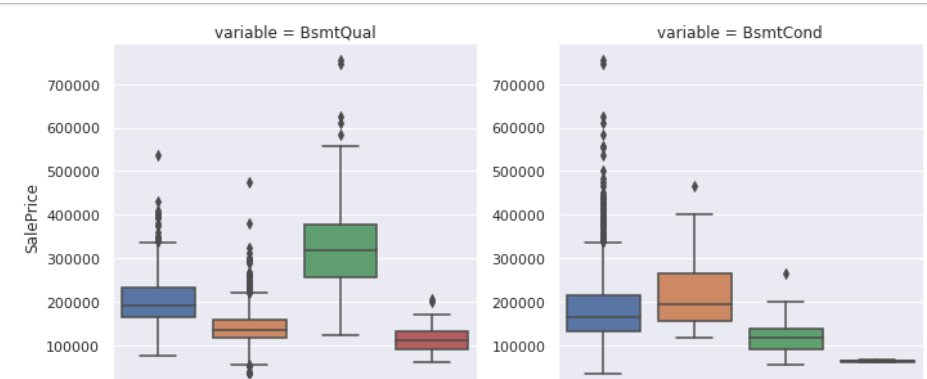
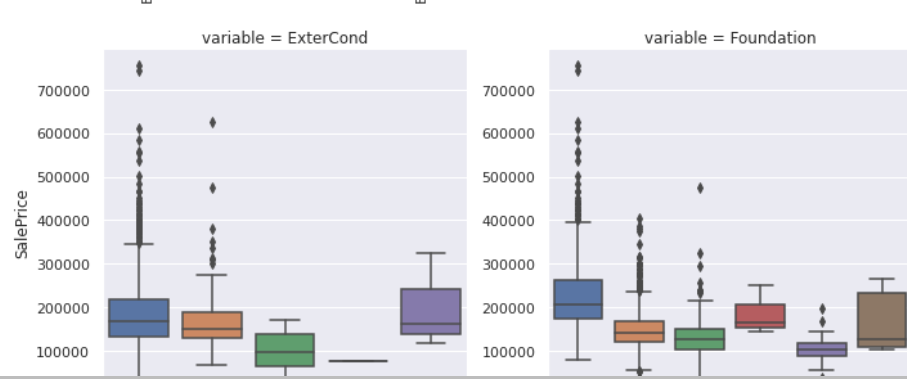
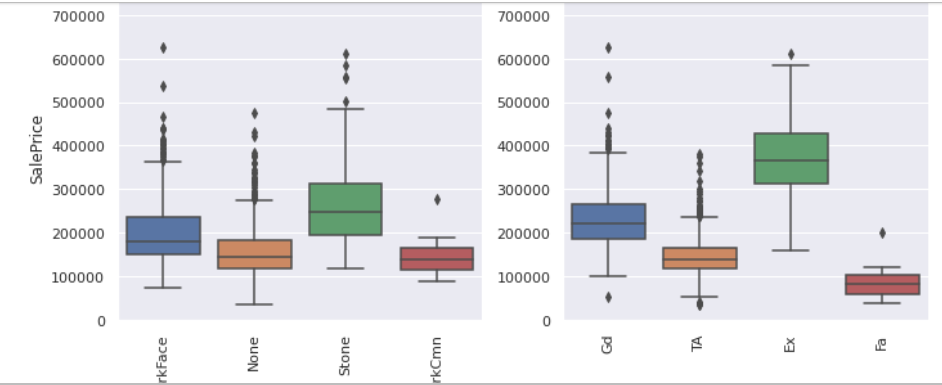
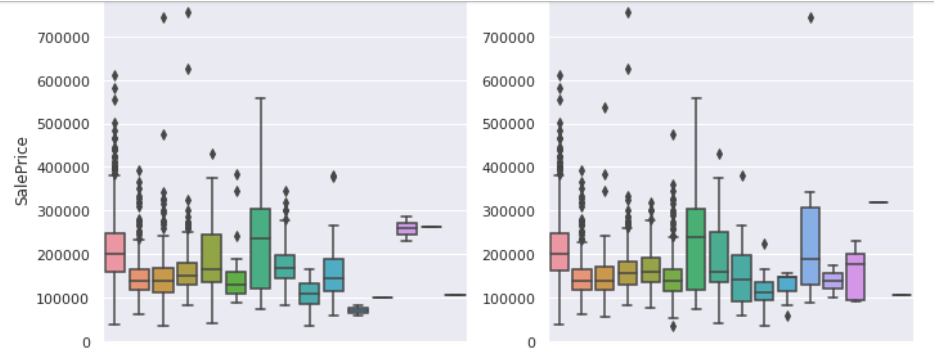
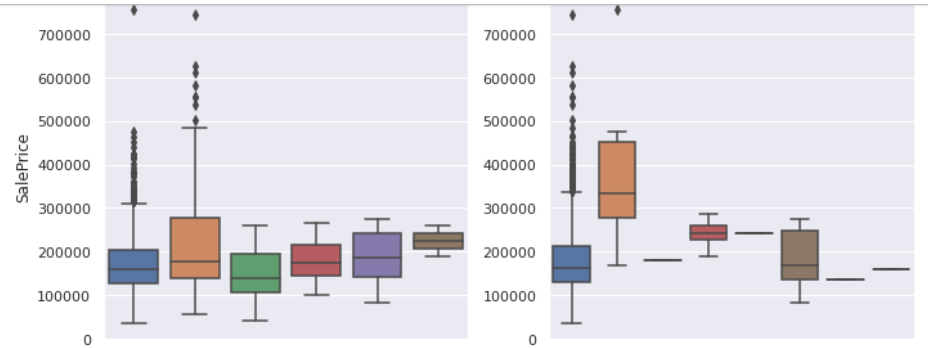
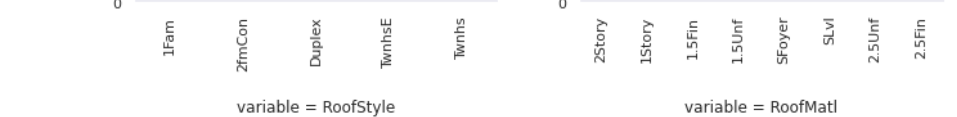
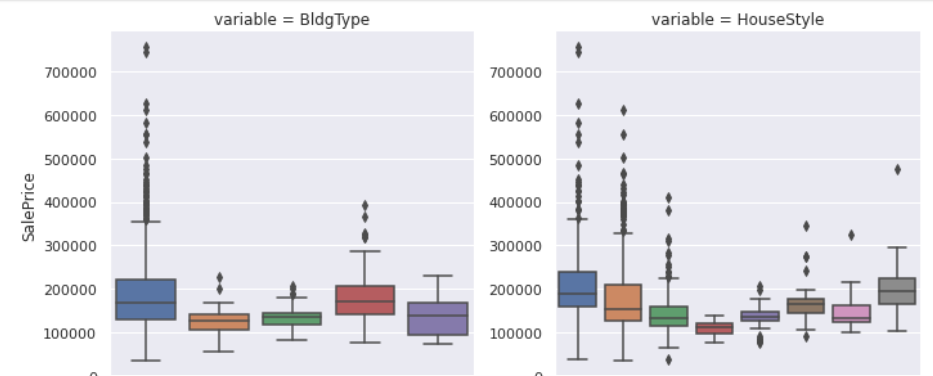
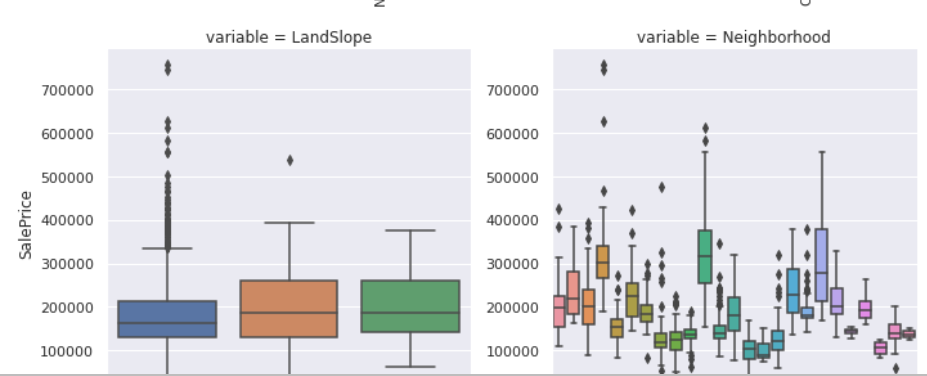
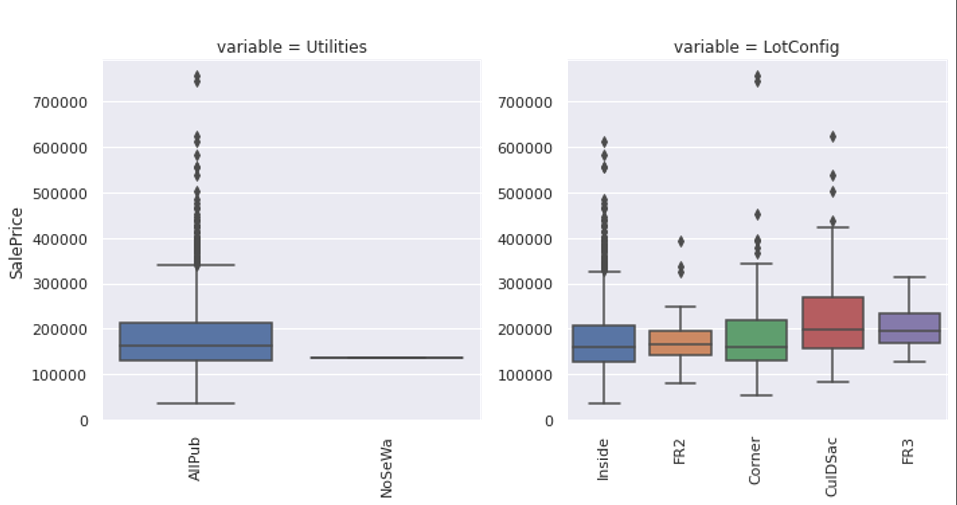
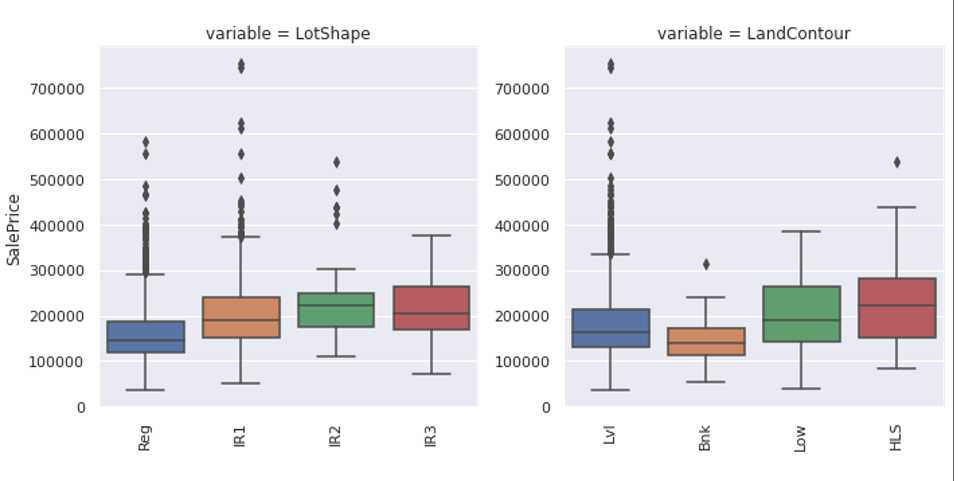
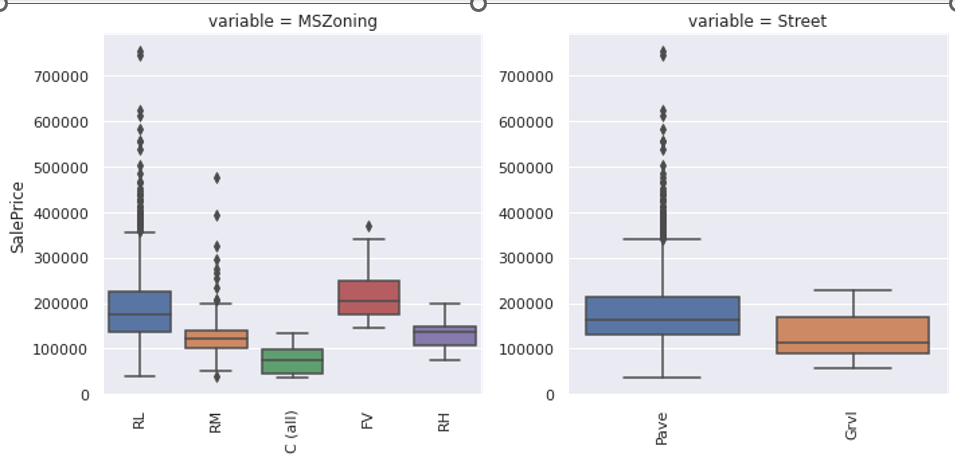


[207]:

<AxesSubplot:ylabel='Density'>



<seaborn.axisgrid.FacetGrid at 0x7f3ba6cc8c90>



**Note:** The last two points are performed to make the new dataset ready for training and prediction.