**PCA**

To perform PCA we used the label encoder to convert non-numeric features to numeric values. We then applied PCA without limiting the number of components to get a sense of the data. The quick tailing off of variance in the dataset explained by each subsequent component suggests that not many features in the housing training dataset are strongly correlated. We then limited the number of components for the analysis. It is worth noting that a reduction in the dimensionality of the data by half -- from 80 columns in the dataset to 40 components -- still encompasses 80% of the total variation in the dataset. If one wishes to drastically reduce dimensionality we would suggest using 15 components, which is sufficient to incorporate ~50% of the total variance.

**K Means Model**

To create a k means model, we first plotted all the numerical columns against each other to find some sort of clustering pattern. Sale Price vs. Garage Area gave a somewhat prominent cluster, although there was overlapping. This led to a problem when using the K means classifier. The model chose random variables and approximated the closest points as a part of the cluster of those variables. We noticed that the model classified several variables correctly. The problem was when there was overlapping. The MS Zoning column was used as the hyper parameter because it created quite prominent clusters indicating that sale price and garage area, proportional to the actual area of the house, were dependent on the zone the house was located in. When trying to fit data into this model, we realized that the k means classifier works well when the data has clear-cut clusters. Overlapping boundaries or points prevent the classifier from being as accurate as possible since this model classifies all the points near the chosen centroid as one cluster.