**Objectives**

To assess the data and prepare a new dataset for training and prediction.

To create a box plot to identify the variables with outliers.

Using the dataset, find the factors influencing price negotiations while buying a house.

Text

Description automatically generated with medium confidence

The table shows a dataset's chi-square test results for independence for several categorical variables. The chi-square test is used to determine if there is a significant association between two categorical variables. The p-value indicates the probability of observing such an extreme result if there were no associations between the variables. A p-value less than 0.05 is commonly used as a threshold for statistical significance, meaning there is strong evidence of an association between the variables.

The table shows that all the variables have p-values less than 0.05, indicating they are all significantly associated with the outcome variable.

The variables with the highest chi-square values and smallest p-values are Neighborhood, SaleType, and KitchenQual. This indicates that these variables have the strongest association with the outcome variable.

The neighborhood has the highest chi-square value and smallest p-value, which means it is the most significant variable. This suggests that the location of the property has a strong influence on the outcome variable.

Text

Description automatically generated with medium confidence

This list shows the correlation coefficients between the dataset's target variable SalePrice and other numerical variables.

OverallQual has the highest correlation coefficient of 0.79, indicating a strong positive correlation between the overall quality of the house and the sale price.

GrLivArea, "above grade (ground) living area square feet," has the second highest correlation coefficient of 0.71, indicating a strong positive correlation between the living area and sale price.

GarageCars and GarageArea are strongly correlated with SalePrice, with correlation coefficients of 0.64 and 0.62, respectively. These two variables represent different aspects of the garage space, with GarageCars indicating the number of cars that can fit in the garage and GarageArea indicating the size of the garage in square feet.

Therefore, the top three most strongly correlated variables with SalePrice are OverallQual, GrLivArea, and GarageCars or GarageArea. Since GarageCars and GarageArea are highly correlated, we can group them and consider them as a single variable in our analysis.