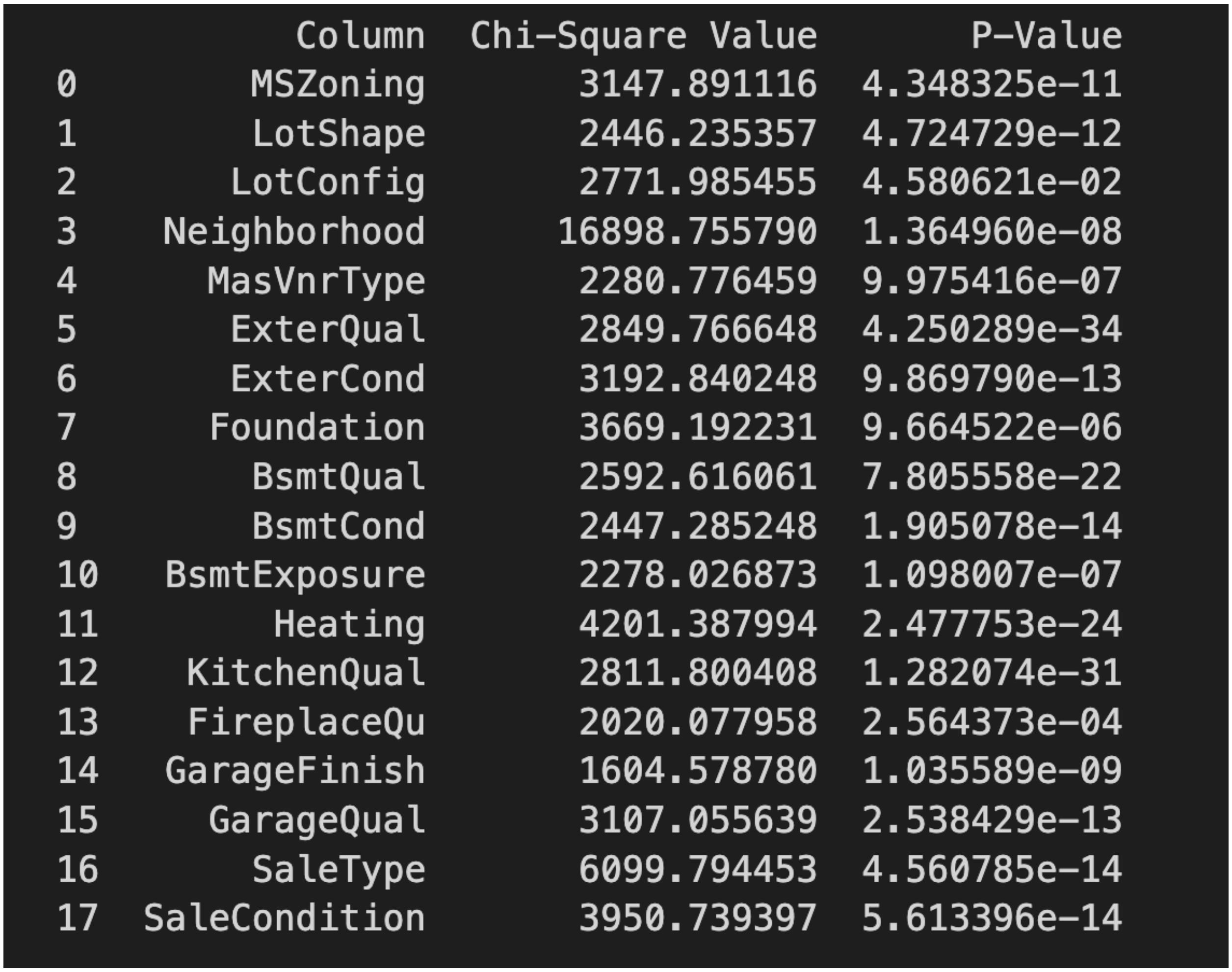
Objectives

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.



```
              Column  Chi-Square Value       P-Value
0           MSZoning       3147.891116  4.348325e-11
1           LotShape       2446.235357  4.724729e-12
2          LotConfig       2771.985455  4.580621e-02
3       Neighborhood      16898.755790  1.364960e-08
4         MasVnrType       2280.776459  9.975416e-07
5          ExterQual       2849.766648  4.250289e-34
6          ExterCond       3192.840248  9.869790e-13
7         Foundation       3669.192231  9.664522e-06
8           BsmtQual       2592.616061  7.805558e-22
9           BsmtCond       2447.285248  1.905078e-14
10      BsmtExposure       2278.026873  1.098007e-07
11           Heating       4201.387994  2.477753e-24
12       KitchenQual       2811.800408  1.282074e-31
13       FireplaceQu       2020.077958  2.564373e-04
14      GarageFinish       1604.578780  1.035589e-09
15        GarageQual       3107.055639  2.538429e-13
16          SaleType       6099.794453  4.560785e-14
17     SaleCondition       3950.739397  5.613396e-14
```

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.

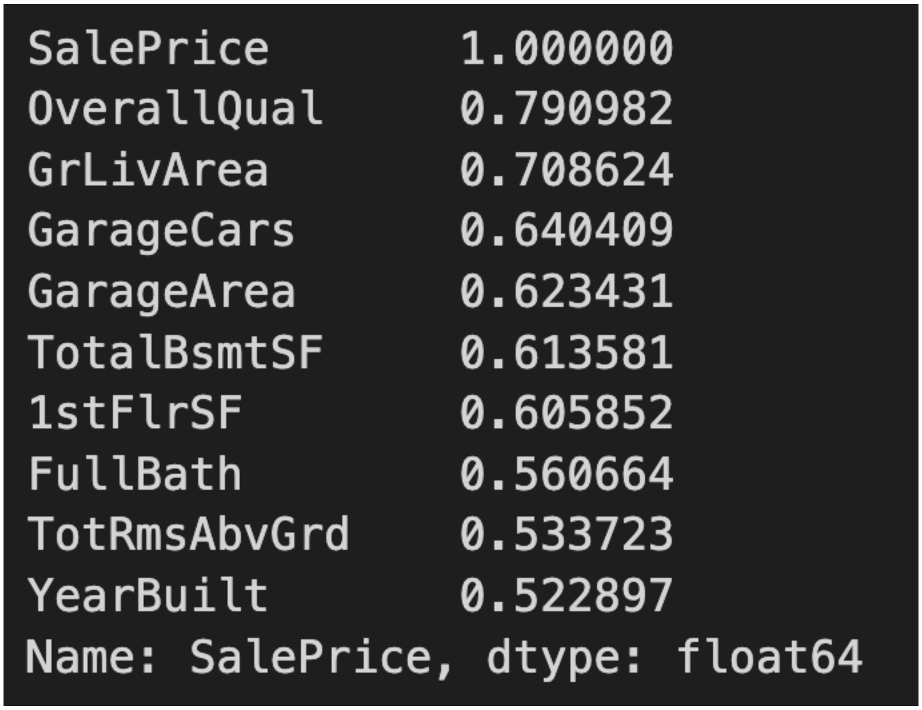
The variables with the highest chi-square values and smallest p-values are Neighborhood, SaleType, and KitchenQual. This indicates that these variables have the strongest association with the outcome variable.

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.

An interesting observation is that KitchenQual had a lower p-value than SaleType, which suggests a more substantial statistical significance. However, the ranking in this list is based on the Chi-Square Value, not the p-value.

The Chi-Square Value measures how much the observed frequencies differ from the expected frequencies under a specific statistical model. In this case, it measures how much the observed frequencies of the target variable SalePrice differ across the categories of each categorical variable relative to what would be expected if the target variable and categorical variable were independent. A higher Chi-Square Value indicates a stronger association between the categorical and target variables.

In this list, SaleType has a higher Chi-Square Value than KitchenQual, indicating a stronger association between SaleType and SalePrice. This does not necessarily mean that SaleType is a better predictor of SalePrice than KitchenQual, as there could be other factors not captured by this analysis. It only suggests that SaleType is more strongly associated with SalePrice in this dataset.



```
SalePrice       1.000000
OverallQual     0.790982
GrLivArea       0.708624
GarageCars      0.640409
GarageArea      0.623431
TotalBsmtSF     0.613581
1stFlrSF        0.605852
FullBath        0.560664
TotRmsAbvGrd    0.533723
YearBuilt       0.522897
Name: SalePrice, dtype: float64
```

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.

Looking at the violin plots for the categorical variables against the numerical variables, we can make the following observations and conclusions:

Neighborhood: The SalePrice of houses in some neighborhoods (e.g., NridgHt and NoRidge) tends to be higher than in others. The OverallQual and GrLivArea tend to be higher for houses in these neighborhoods.

SaleType: Houses sold under new construction have higher SalePrice, OverallQual, and GrLivArea than other sale types.

KitchenQual: Houses with excellent kitchen quality tend to have higher SalePrice, OverallQual, and GrLivArea than those with fair or poor kitchen quality.

OverallQual: Higher quality is associated with higher SalePrice, GrLivArea, and GarageArea.

GrLivArea: Houses with larger living areas tend to have higher SalePrice, OverallQual, and GarageArea.

GarageArea: Houses with larger garage areas tend to have higher SalePrice, OverallQual, and GrLivArea.

Overall, these violin plots help us understand the relationship between the categorical and the numerical variables and how each contributes to the SalePrice of a house.