T.C.
FATİH SULTAN MEHMET VAKIF ÜNİVERSİTESİ

DATA SCIENCE
MIDTERM PROJECT

House Prices: Advanced Regression Techniques

Merve YAVUZ
1621221003

HOUSE PRICES: Advanced Regression Techniques

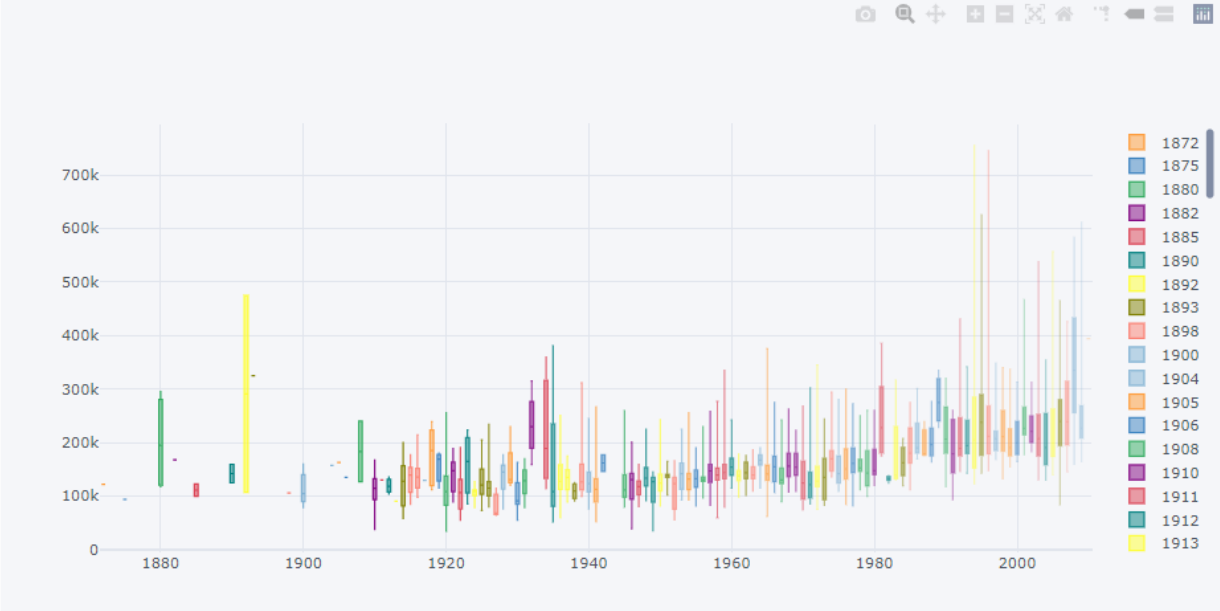
1.Explanatory Data Analysis

The data consists of 1460 rows and 81 columns. It contains 43 objects, 35 int, 3 float type data.

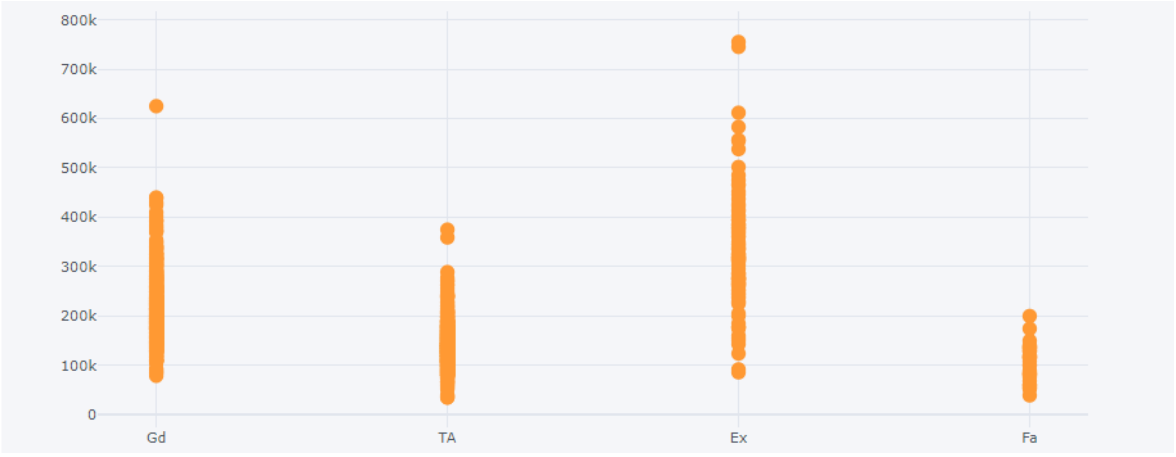
The dataset contains null values in 19 columns.

Below are the graphs containing the effect of five columns on the sales price.

In the boxplot chart, it is observed that the price average increases as the building year of the building increases.



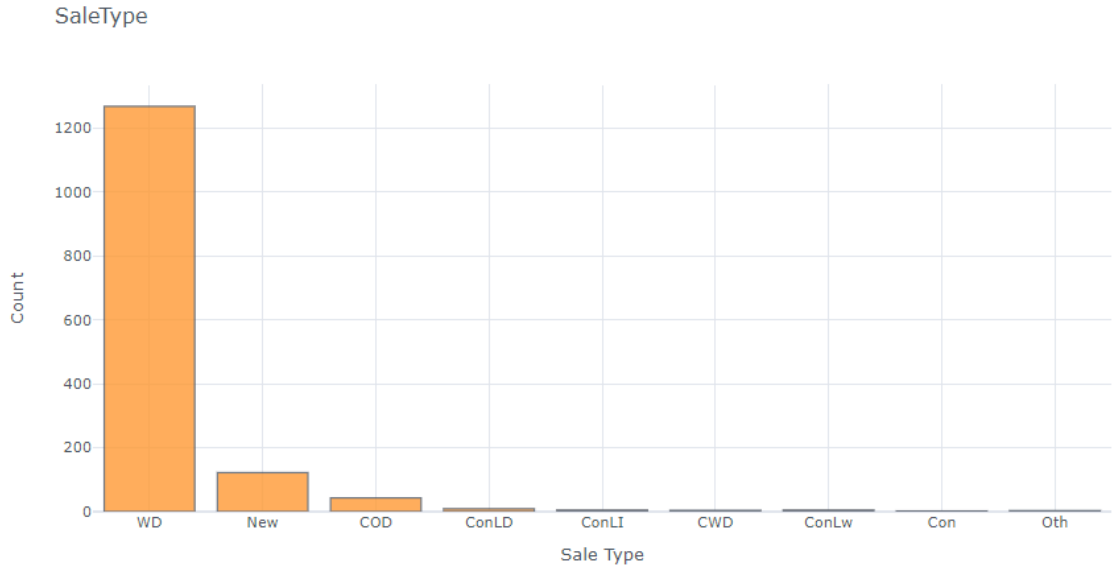
In the graphic below, the distribution of kitchen quality on sales prices can be seen. Although it can be observed that high quality kitchen houses and low quality houses are sold at the same price, in general, it can be said that the quality increases the sale price of the house.



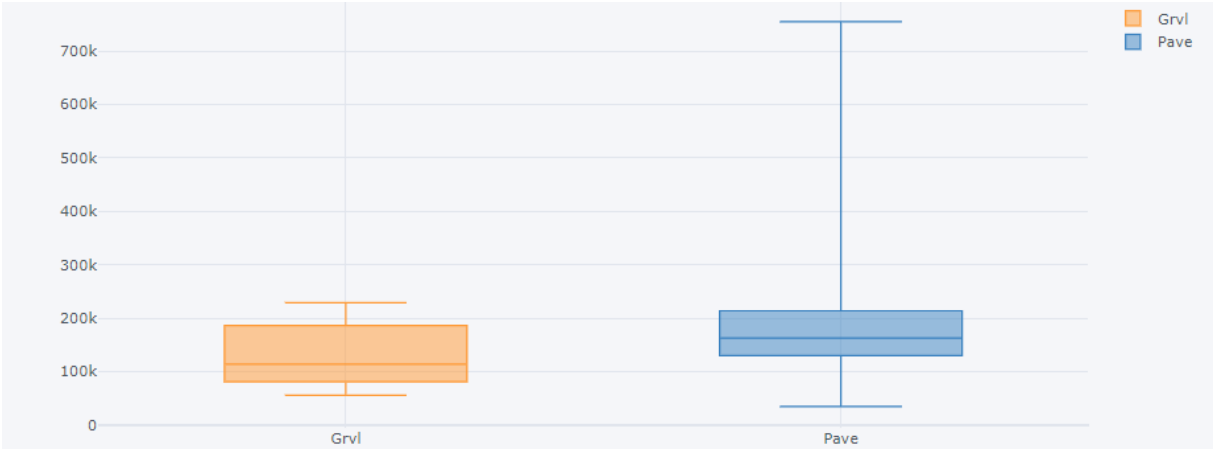
First floor square meter values are analyzed with describe method

```
count    1460.000000
mean     1162.626712
std       386.587738
min       334.000000
25%       882.000000
50%      1087.000000
75%      1391.250000
max      4692.000000
Name: 1stFlrSF, dtype: float64
```

Below is the distribution of sales types in the data. The most observed houses are the houses of WD sales type.



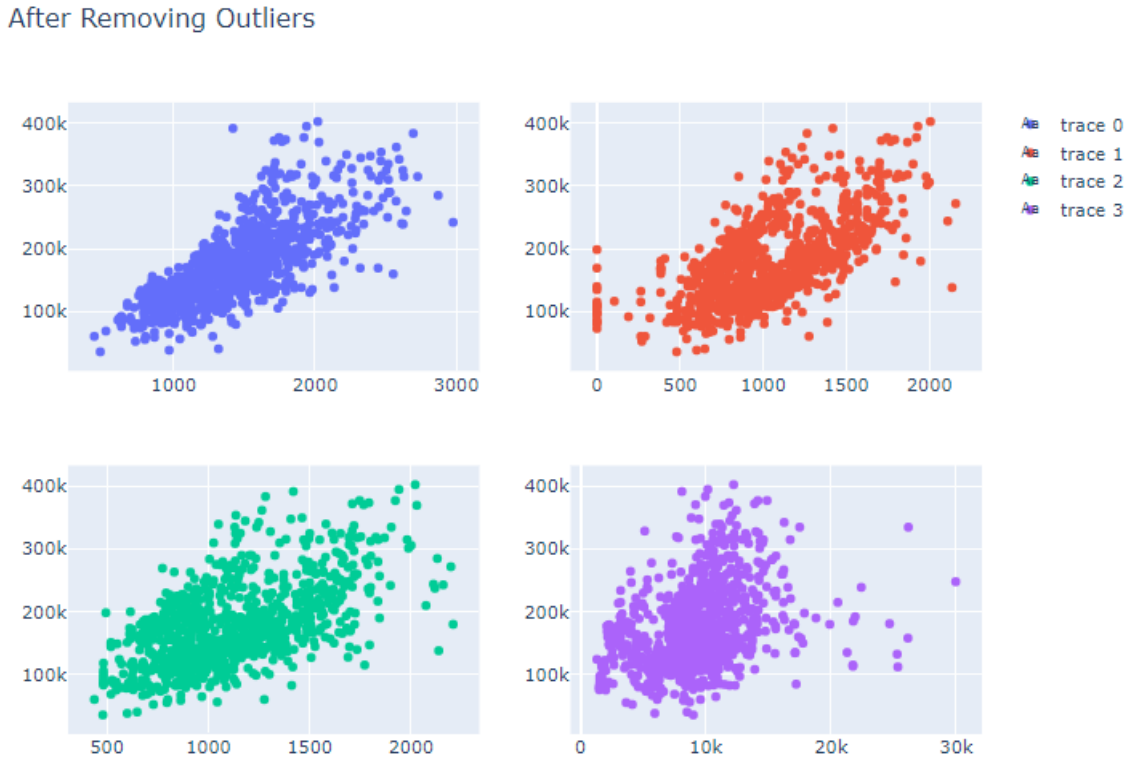
In the boxplot chart below, the distribution of the street type in the sales price is observed. It is seen that the streets of Pave type are sold at higher prices on average.



2.Preprocessing

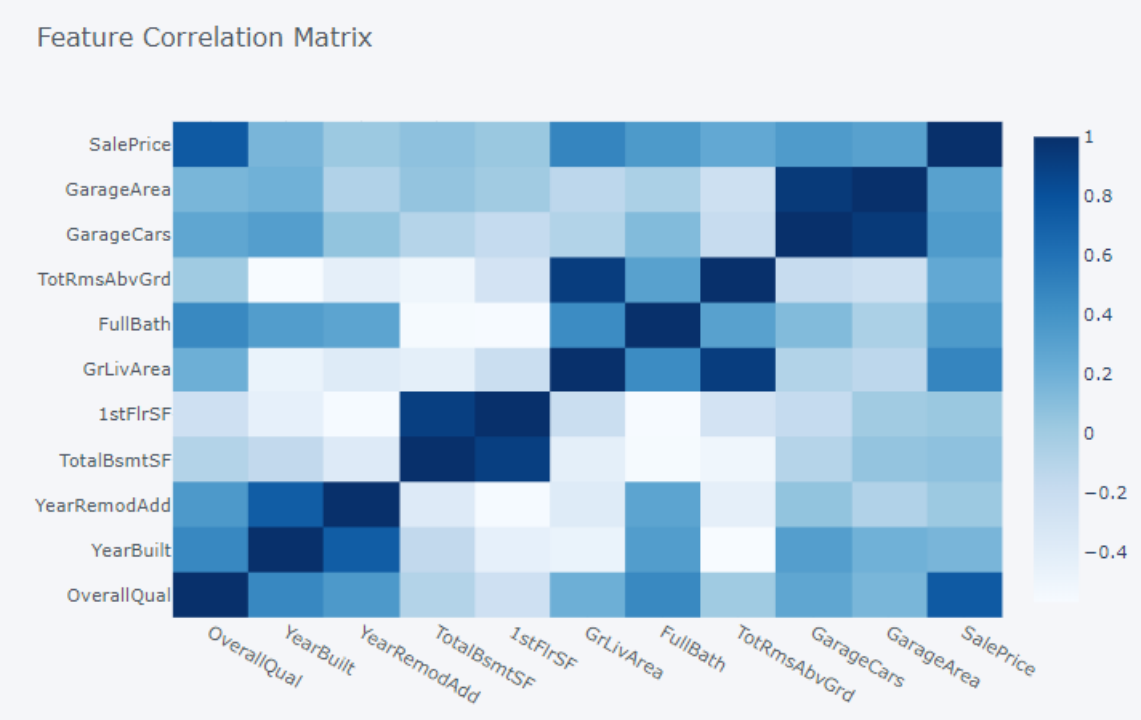
Categorical values in the data are filled with the most frequent value, and numerical values are filled with average.

Outlier values are removed from the data according to Z score. Below is the GrLivArea, TotalBsmtSF, 1stFlrSF, LotArea values before and after the outlier is removed.



LabelEncoder was used to process categorical columns. In this way, categorical data were converted into numbers.

Correlation matrix is used for feature selection. Those whose absolute value is greater than 0.5 are taken.



3. Model

Gradient Boosting Regression was used as a model and an accuracy value of 87.71292419045736 was obtained.

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

Boosting is a method of converting weak learners into strong learners. Therefore, it has been determined that it gives better results than linear regression.

KAGGLE :

https://www.kaggle.com/merveyavuz/house-price-prediction