REPORT ON PROPERTY - PRICE DATA

(Submitted by: Navneeth Sreenivasan)

This report aims to present the summarizations and findings on the property-price data. The project was undertaken from 1st July to 15th July and this report gives a detailed summary of the findings and the patterns observed on data cleaning and data analysis tasks performed. The first task was to perform data cleaning, where we preprocess and clean the data to ensure accuracy and consistency. This involved finding the types of the data, handling missing values with the help of heatmaps, encoding the categorical values to numerical ones using LabelEncoder, checking for duplicates and handling outliers by checking their skewness. The second task given was to perform exploratory data analysis. This involved analyzing the variable 'Sale_Price' since we believe this is the target variable to be predicted. We checked its skewness and visualized it to find its distribution using a bar graph ggplot. After this, we plotted a heatmap to identify the correlation between the target and the other features and checked the correlation between these attributes with the target variable. We also plot the six most correlated variables with the target variable and see how they are related. Python language was used to perform the data preprocessing and cleaning as well as the exploratory data analysis. Jupyter Notebook was used to write the code and visualize the results.

Task 1 – Data Preprocessing

The first task we had was to perform data preprocessing and cleaning, where we preprocess and clean the data to ensure accuracy and consistency. This involved finding the types of the data, handling missing values with the help of heatmaps, encoding the categorical values to numerical ones using LabelEncoder, checking for duplicates and handling outliers by checking their skewness.

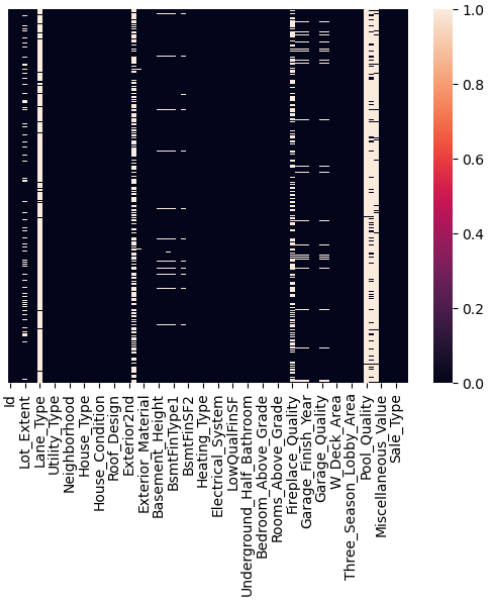
1) Types of data:

The data consists of 1459 rows along 81 columns. More than half of the data consists of object type variables, and the rest consists of integer and floating-point variables. We can see that 38 of the columns consist of integer and float variables, whereas the rest consist of object variables.

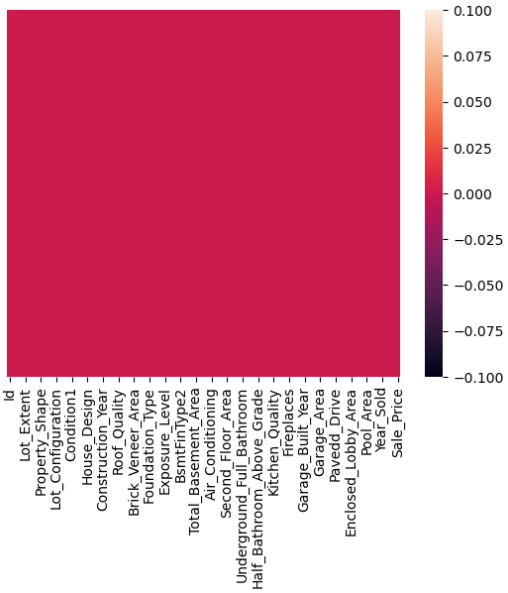
2) Checking for missing values:

Using isnull() function in python, we try and see how many null values are present in the data. We see that a few columns have missing values, for example, Lot_Extent variable has 259 missing values. We plot a heatmap to see the null values, and the results are interesting. We can see that *Brick_Veneer_Type*, *Fireplace_Quality* and *Fence_Quality* variables consists of a lot of missing values. However, *Miscellaneous_Feature*, *Pool_Quality* and *Lane_Type* variables

consists of mostly only missing values. These two variables have only a few actual values present in the dataset. The heatmap which was plotted is given below.



Not satisfied with this, we use info() function in python, to check for missing values. The output reinforces our heatmap. Therefore, we drop the columns with more than 50% of its value missing, which includes most of the above-mentioned columns. After this, we replace the missing values of the object variables with the mode of those columns and the missing values of the integer and float variables with the mean of those columns. Plotting a heatmap and running info() again, reinforces what we have seen, that is, there are no more missing values. We see there are 76 columns now. The new heatmap is shown below.



3) Encoding the data:

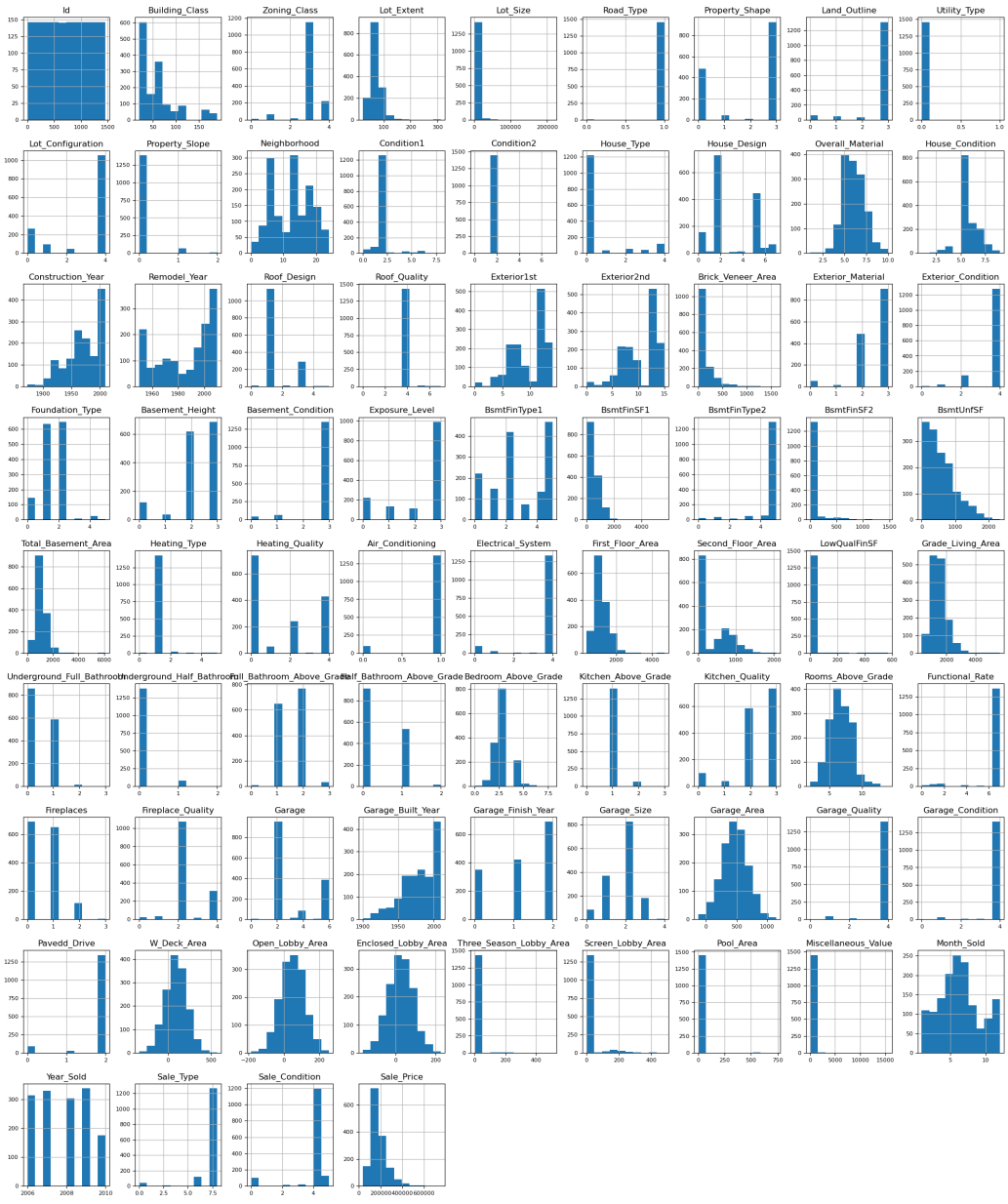
We use LabelEncoder module from sklearn library to convert the object types to numerical data types. We use fit_transform() from LabelEncoder to convert the object type data to appropriate numeric type data. We can now see that most of the object types have been replaced with integer data type, with a few float datatypes as required.

4) Checking for duplicates:

We see that there are no duplicates in the data, using the duplicated.sum() function

5) Handling the outliers:

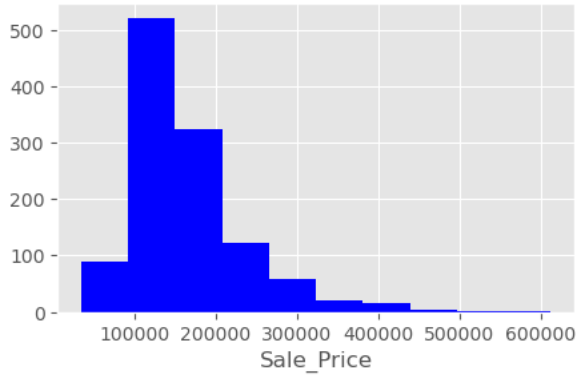
We first plot a histogram of each of the columns and see the results. The histogram is shown below.



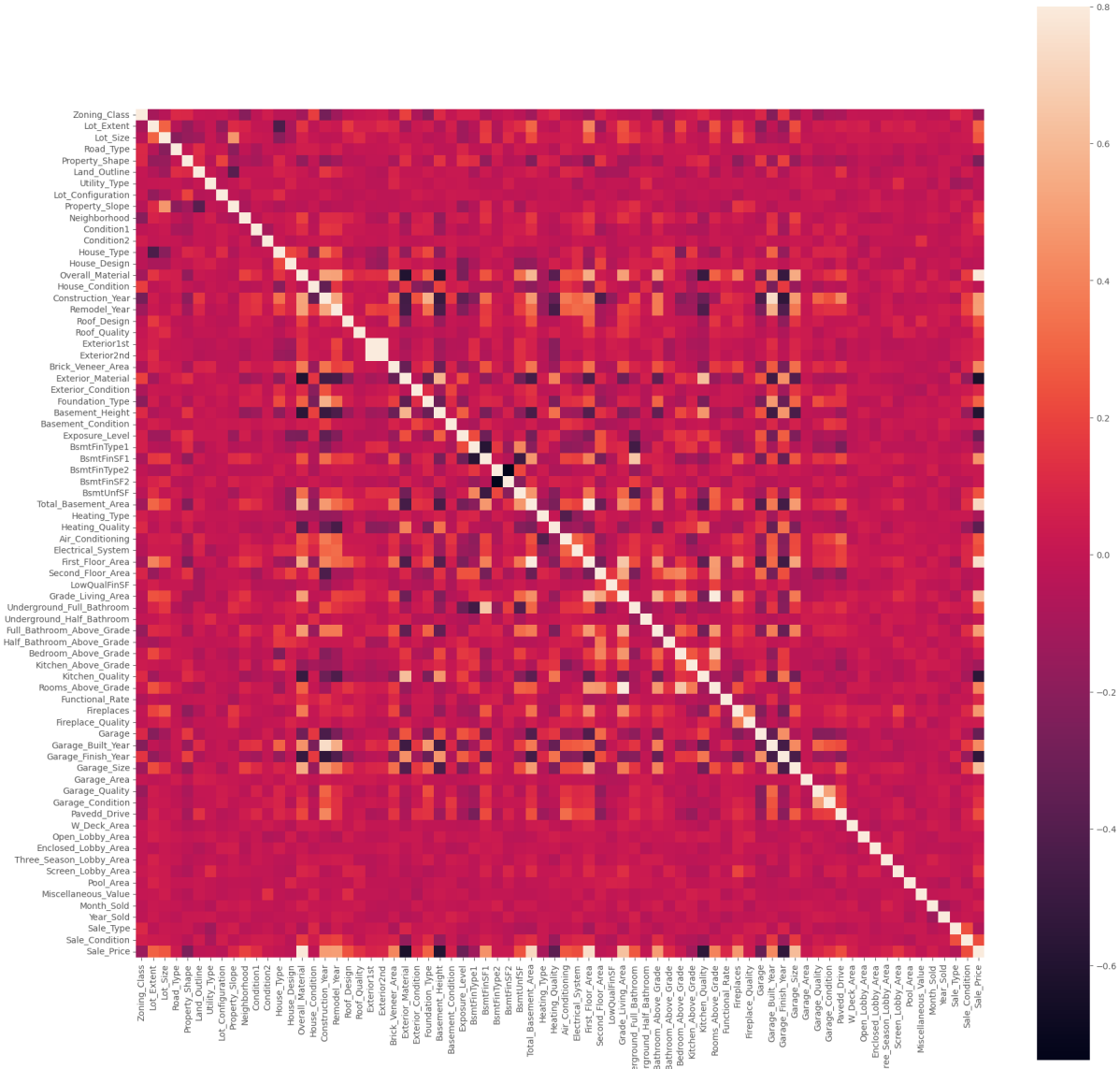
We then use the skew() function to check the skewness of each of the variables. This lets us remove the column *Id* since its skewness is 0, and it's not required for us anyway. The rest of the variables are skewed either to the positive or to the negative side. This makes us realize that outliers exist for all the variables, and we need to remove it. We make a function called, *outlier_detection*() to handle the outliers. If the skewness is 0, then it is normally distributed. We apply Z-score formula on it, where we find the difference between the mean of that variable to its values and then divide it by the standard deviation of that particular variable. If it is not normally distributed, then we apply InterQuartileRange to it. Finally, when we check the data we have, we see that it has 1160 rows and 75 columns to it.

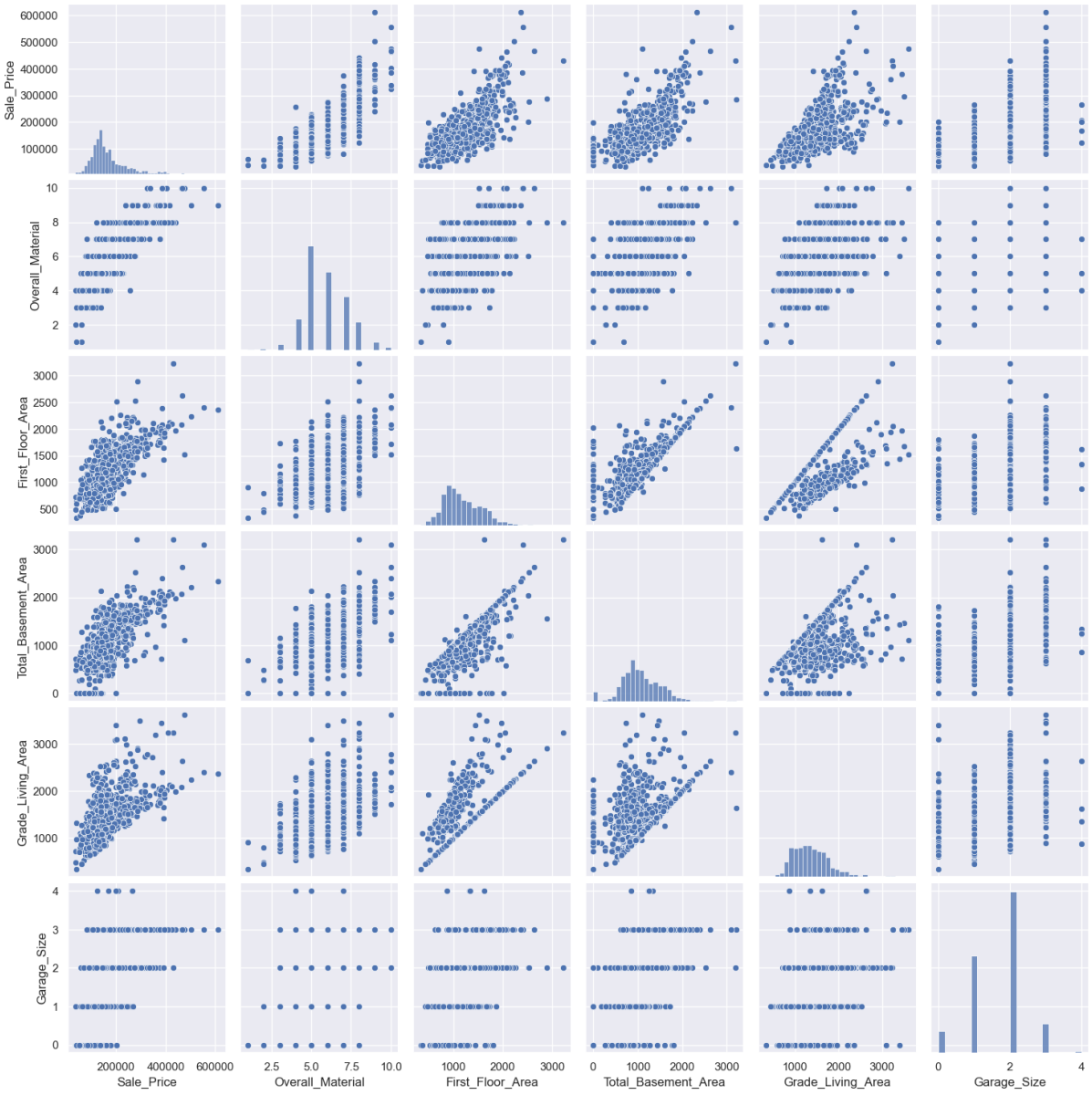
Task 2 – Exploratory Data Analysis

We believe Sale_Price is the value that will be predicted, given the dataset, so we analyze it. First, we describe it and then check its skewness. After this, we visualize the target Sale_Price variable using a bar graph using ggplot. The visualization is as given below.



We also plot a heatmap matrix to identify the correlation between the target variable and the other feature variables. We can see form the heatmap that some correlation exists between all the variables, but there does indeed seem to be one with the target variable. We try to see the correlation between each of the variables to the Sale_Price variable using corr() function, which computes the pairwise correlation between each of the variables to the Sale_Price variable. This makes us see that almost all the variables have some correlation with the target Sale_Price variable. We also use seaborn library to plot the six variables most correlated with Sale_Price using Seaborn's pairplot() function and this also shows that some correlation exists between Sale_Price and the variables. The heatmap and the pairplot are given below.


Zoning_Class
Lot_Extent
Lot_Size
Road_Type
Property_Shape
Land_Outline
Utility_Type
Lot_Configuration
Property_Slope
Neighborhood
Condition1
Condition2
House_Type
House_Design
Overall_Material
House_Condition
Construction_Year
Remodel_Year
Roof_Design
Roof_Quality
Exterior1st
Exterior2nd
Brick_Veneer_Area
Exterior_Material
Exterior_Condition
Foundation_Type
Basement_Height
Basement_Condition
Exposure_Level
BsmtFinType1
BsmtFinSF1
BsmtFinType2
BsmtFinSF2
BsmtUnfSF
Total_Basement_Area
Heating_Type
Heating_Quality
Air_Conditioning
Electrical_System
First_Floor_Area
Second_Floor_Area
LowQualFinSF
Grade_Living_Area
Underground_Full_Bathroom
Underground_Half_Bathroom
Full_Bathroom_Above_Grade
Half_Bathroom_Above_Grade
Bedroom_Above_Grade
Kitchen_Above_Grade
Kitchen_Quality
Rooms_Above_Grade
Functional_Rate
Fireplaces
Fireplace_Quality
Garage
Garage_Built_Year
Garage_Finish_Year
Garage_Size
Garage_Area
Garage_Quality
Garage_Condition
Pavedd_Drive
W_Deck_Area
Open_Lobby_Area
Enclosed_Lobby_Area
Three_Season_Lobby_Area
Screen_Lobby_Area
Pool_Area
Miscellaneous_Value
Month_Sold
Year_Sold
Sale_Type
Sale_Condition
Sale_Price
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Sale_Price
Overall_Material
First_Floor_Area
Total_Basement_Area
Grade_Living_Area
Garage_Size