29.12.2023

EDA Report for Housing Dataset

By Darshak Vasoya

Executive Summary

Exploring, Visualizing and analysis for housing database, these tasks are being done in the report. Pandas for data transformation, NumPy for data calculation, matplotlib and seaborn for visualizing data, skillsnetwork for download dataset, SciPy for implementing statistics methods on dataset, these python libraries are going to be used. First, exploring different variables and their types and seeing how different variables correlate with each other. Second, removing duplicate value, handling missing value, finding Outliers and transforming categorical variables into dummies for further computation will be done. At the end, we will check hypotheses by chi-square test and t-test.

INDEX

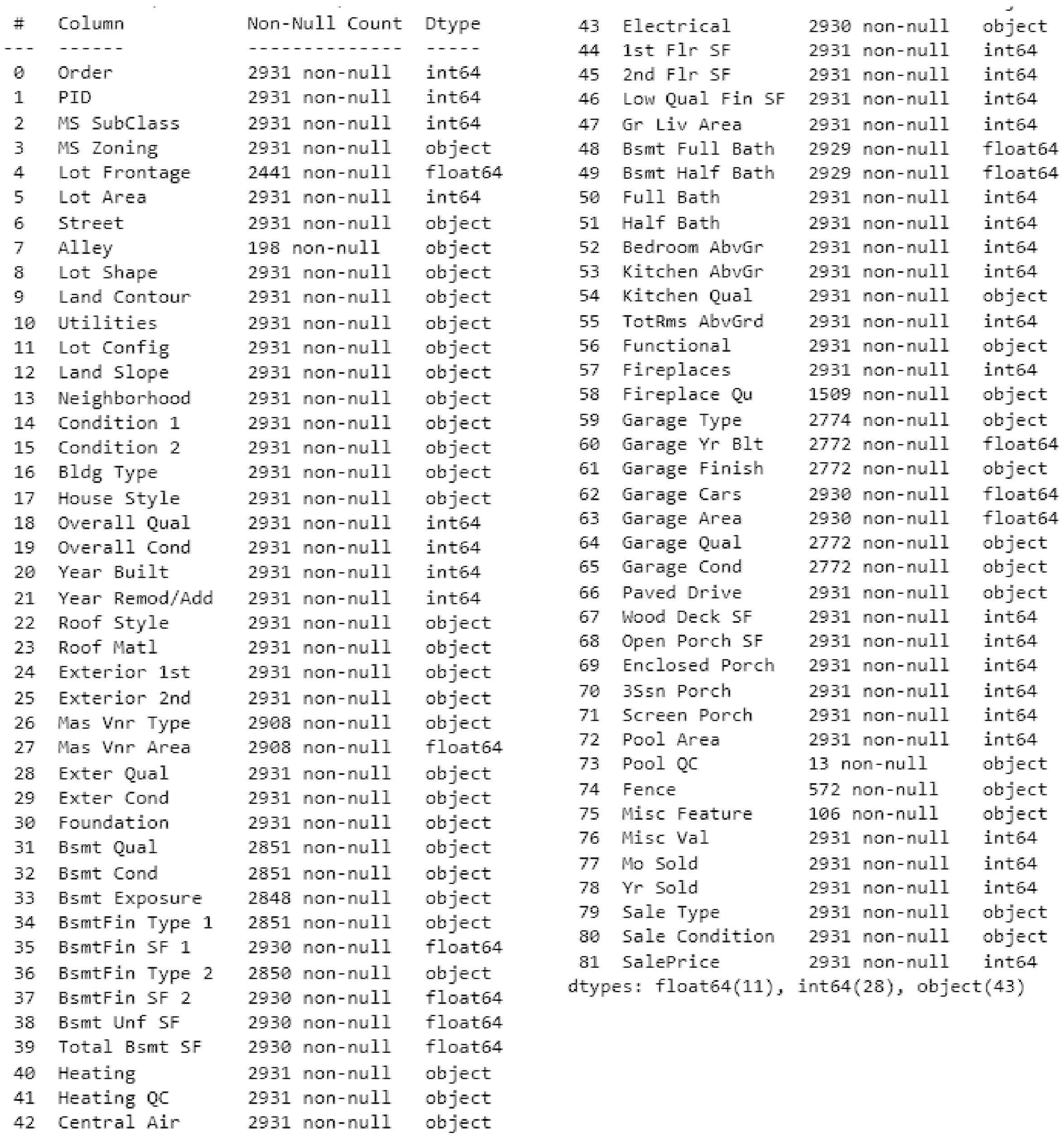
1. Introduction

We will use housing dataset for EDA report. Here, there is the link below for this dataset by using Python language and their different libraries.

“ https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBM-ML0232EN-SkillsNetwork/asset/Ames_Housing_Data1.tsv ”

Objectives for the report are to see how different variables are connected with each other, finding insights from dataset and preparing dataset for the machine learning techniques (regression, classification and so on).

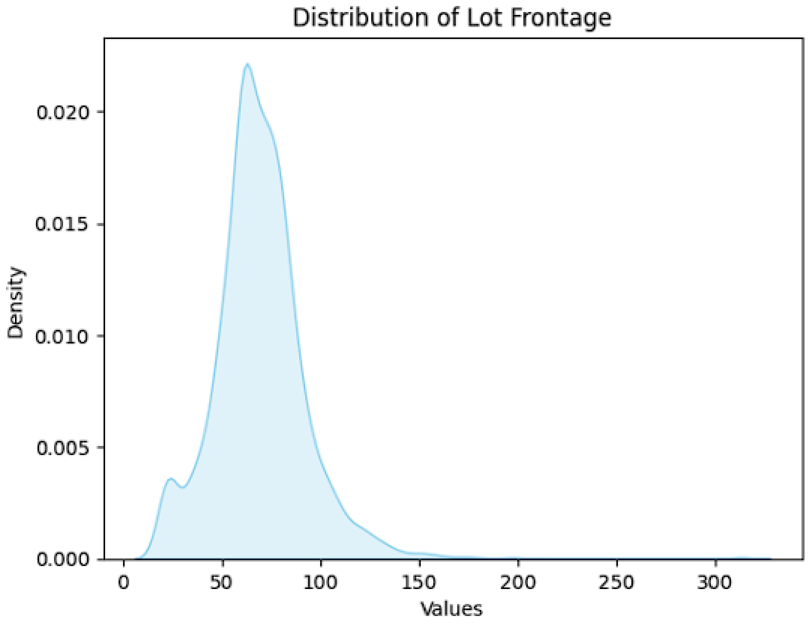
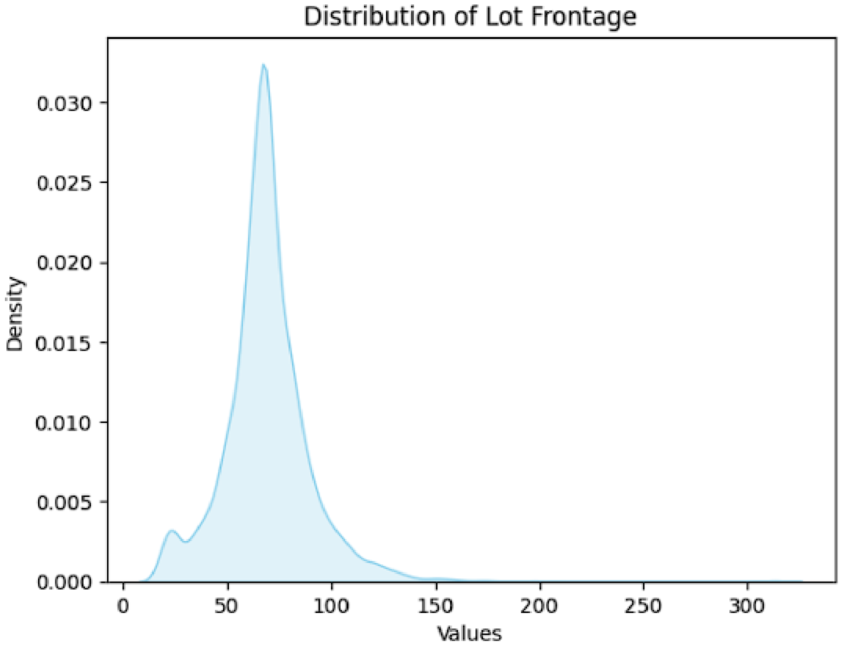
2. Data Description



```
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Order           2931 non-null   int64
 1   PID             2931 non-null   int64
 2   MS SubClass     2931 non-null   int64
 3   MS Zoning       2931 non-null   object
 4   Lot Frontage    2441 non-null   float64
 5   Lot Area        2931 non-null   int64
 6   Street          2931 non-null   object
 7   Alley           198 non-null    object
 8   Lot Shape       2931 non-null   object
 9   Land Contour    2931 non-null   object
 10  Utilities       2931 non-null   object
 11  Lot Config      2931 non-null   object
 12  Land Slope      2931 non-null   object
 13  Neighborhood    2931 non-null   object
 14  Condition 1     2931 non-null   object
 15  Condition 2     2931 non-null   object
 16  Bldg Type       2931 non-null   object
 17  House Style     2931 non-null   object
 18  Overall Qual    2931 non-null   int64
 19  Overall Cond    2931 non-null   int64
 20  Year Built      2931 non-null   int64
 21  Year Remod/Add  2931 non-null   int64
 22  Roof Style      2931 non-null   object
 23  Roof Matl       2931 non-null   object
 24  Exterior 1st    2931 non-null   object
 25  Exterior 2nd    2931 non-null   object
 26  Mas Vnr Type    2908 non-null   object
 27  Mas Vnr Area    2908 non-null   float64
 28  Exter Qual      2931 non-null   object
 29  Exter Cond      2931 non-null   object
 30  Foundation      2931 non-null   object
 31  Bsmt Qual       2851 non-null   object
 32  Bsmt Cond       2851 non-null   object
 33  Bsmt Exposure   2848 non-null   object
 34  BsmtFin Type 1  2851 non-null   object
 35  BsmtFin SF 1    2930 non-null   float64
 36  BsmtFin Type 2  2850 non-null   object
 37  BsmtFin SF 2    2930 non-null   float64
 38  Bsmt Unf SF     2930 non-null   float64
 39  Total Bsmt SF   2930 non-null   float64
 40  Heating         2931 non-null   object
 41  Heating QC      2931 non-null   object
 42  Central Air     2931 non-null   object
 43  Electrical      2930 non-null   object
 44  1st Flr SF      2931 non-null   int64
 45  2nd Flr SF      2931 non-null   int64
 46  Low Qual Fin SF 2931 non-null   int64
 47  Gr Liv Area     2931 non-null   int64
 48  Bsmt Full Bath  2929 non-null   float64
 49  Bsmt Half Bath  2929 non-null   float64
 50  Full Bath       2931 non-null   int64
 51  Half Bath       2931 non-null   int64
 52  Bedroom AbvGr   2931 non-null   int64
 53  Kitchen AbvGr   2931 non-null   int64
 54  Kitchen Qual    2931 non-null   object
 55  TotRms AbvGrd   2931 non-null   int64
 56  Functional      2931 non-null   object
 57  Fireplaces      2931 non-null   int64
 58  Fireplace Qu    1509 non-null   object
 59  Garage Type     2774 non-null   object
 60  Garage Yr Blt   2772 non-null   float64
 61  Garage Finish   2772 non-null   object
 62  Garage Cars     2930 non-null   float64
 63  Garage Area     2930 non-null   float64
 64  Garage Qual     2772 non-null   object
 65  Garage Cond     2772 non-null   object
 66  Paved Drive     2931 non-null   object
 67  Wood Deck SF    2931 non-null   int64
 68  Open Porch SF   2931 non-null   int64
 69  Enclosed Porch  2931 non-null   int64
 70  3Ssn Porch      2931 non-null   int64
 71  Screen Porch    2931 non-null   int64
 72  Pool Area       2931 non-null   int64
 73  Pool QC         13 non-null     object
 74  Fence           572 non-null    object
 75  Misc Feature    106 non-null    object
 76  Misc Val        2931 non-null   int64
 77  Mo Sold         2931 non-null   int64
 78  Yr Sold         2931 non-null   int64
 79  Sale Type       2931 non-null   object
 80  Sale Condition  2931 non-null   object
 81  SalePrice       2931 non-null   int64
dtypes: float64(11), int64(28), object(43)
```

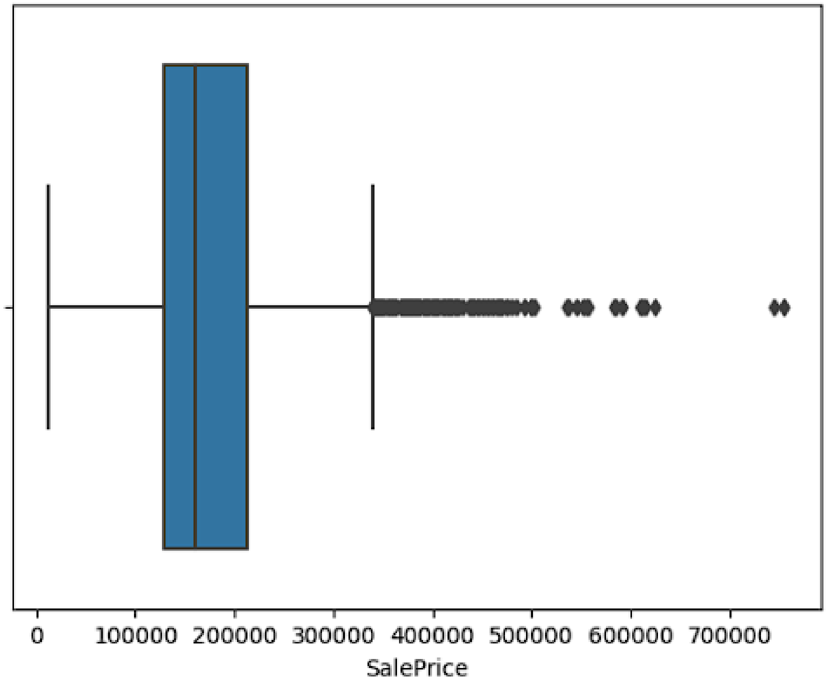
3. Data Cleaning

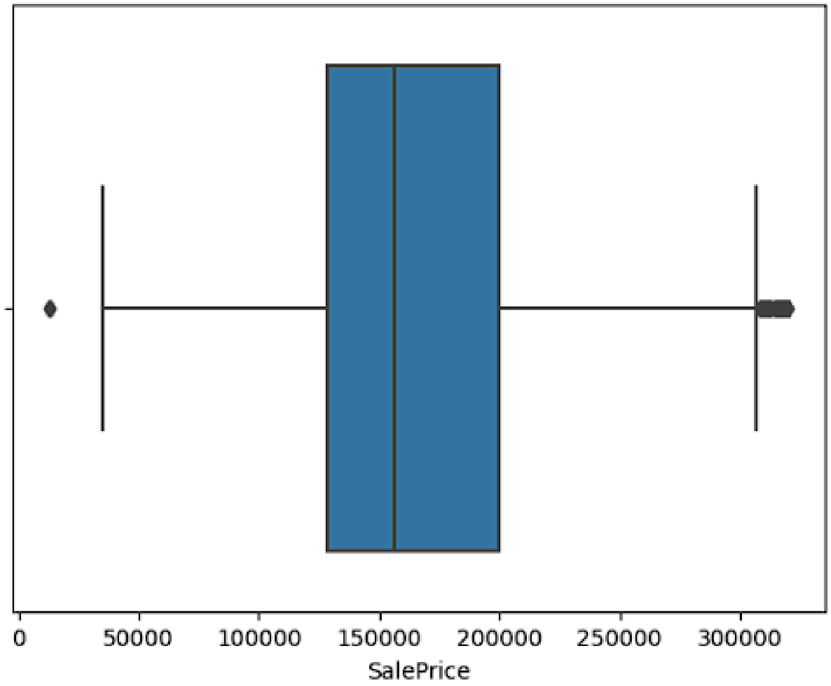
3.1 Replacing null values with medium

Image 3.1.1		Before replacing null value with medium
Image 3.1.2		After replacing null value with medium

3. Data Cleaning

3.2 Detecting and removing Outliers

<table>
<tr>
<td><b>Image 3.2.1</b></td>
<td>


</td>
<td>Before removing Outliers</td>
</tr>
<tr>
<td><b>Image 3.2.2</b></td>
<td>


</td>
<td>After removing Outlier</td>
</tr>
</table>

4. Exploratory Data Analysis

4.1 How variables correlated with each other

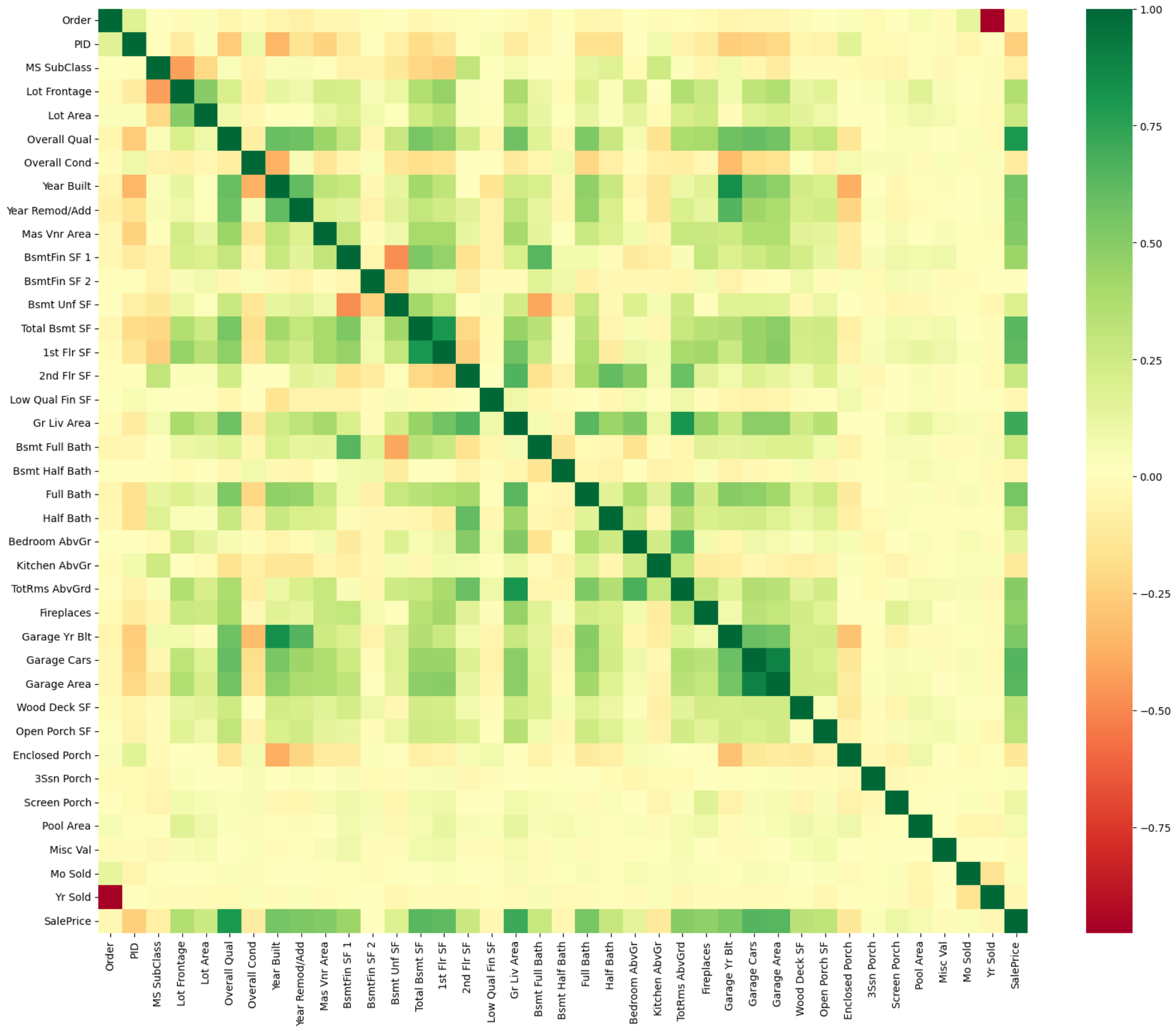


Image 4.1

4. Exploratory Data Analysis

4.2 Number of missing values

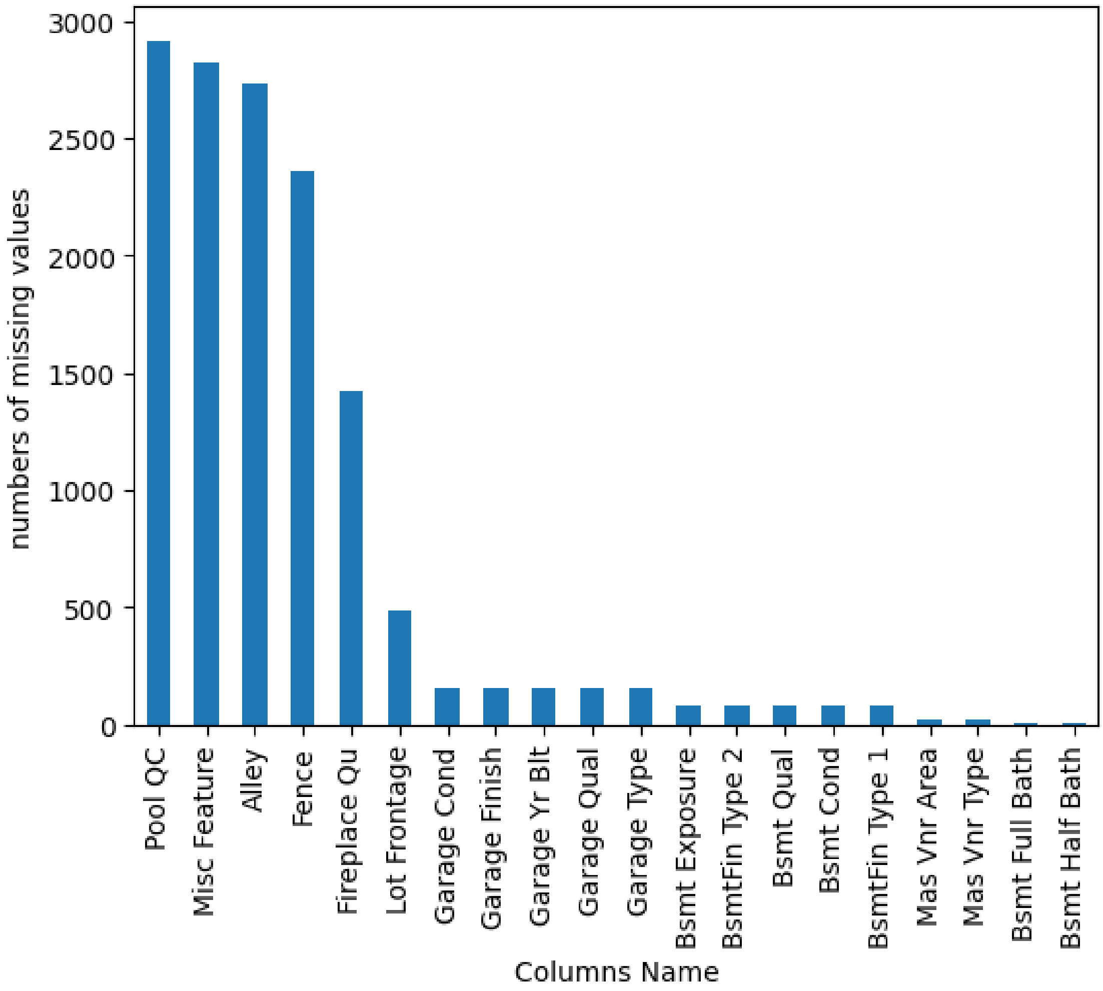
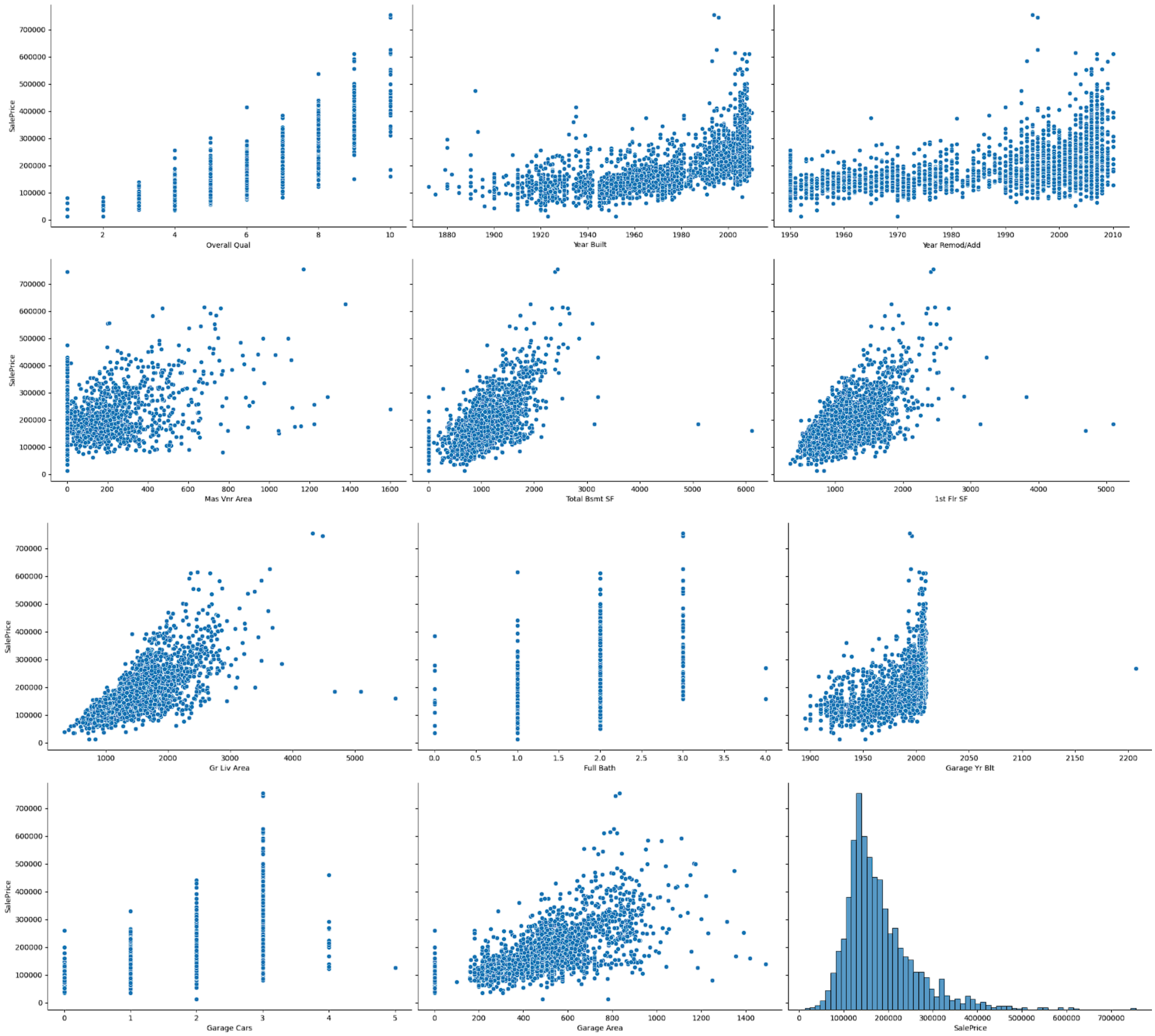


Image 4.2

5. Insights and Observations

5.1 'SalePrice' variable is highly correalated with these variables.

```
['Overall Qual', 'Year Built', 'Year Remod/Add', 'Mas Vnr Area', 'Total Bsmt SF',
'1st Flr SF', 'Gr Liv Area', 'Full Bath', 'Garage Yr Blt', 'Garage Cars', 'Garage
Area']
```



5. Insights and Observations

5.2 Checking Hypotheses

5.2.1 Hypotheses: There is a significant association between the house style and the presence of a garage.

Null Hypothesis (H0): The distribution of house styles is independent of the presence of a garage.

Alternative Hypothesis (H1): The distribution of house styles is associated with the presence of a garage.

After performing **Chi-Square Test**,

Chi-Square Statistic: 51.32719768881161

P-Value: 7.921714887486107e-09

Here the p value is less than 0.05, so we will reject null hypothesis. That means there is a significant association between the house style and the presence of a garage.

5. Insights and Observations

5.2 Checking Hypotheses

5.2.2 Hypotheses: There is a significant association between the Saleprice and Low Qual Fin SF.

Null Hypothesis (H0): The Saleprice of a house is dependent on the Low Qual Fin SF.

Alternative Hypothesis (H1): The Saleprice of house is dependent on the Low Qual Fin SF of the house.

After performing **Chi-Square Test**,

Chi-Square Statistic: 39.093421159520574

P-Value: 0.25166325236016723

Here the p value is more than 0.05, so we will accept null hypothesis. That means there is a significant association between the Saleprice and Low Qual Fin SF.

5. Insights and Observations

5.2 Checking Hypotheses

5.2.3 Hypotheses: There is a not any effect on sale price house whether house has garage or not.

Null Hypothesis (H0): The mean sale prices of houses with a garage are equal to the mean sale prices of houses without a garage.

Alternative Hypothesis (H1): The mean sale prices of houses with a garage are not equal to the mean sale prices of houses without a garage.

After performing **`t-test`**,

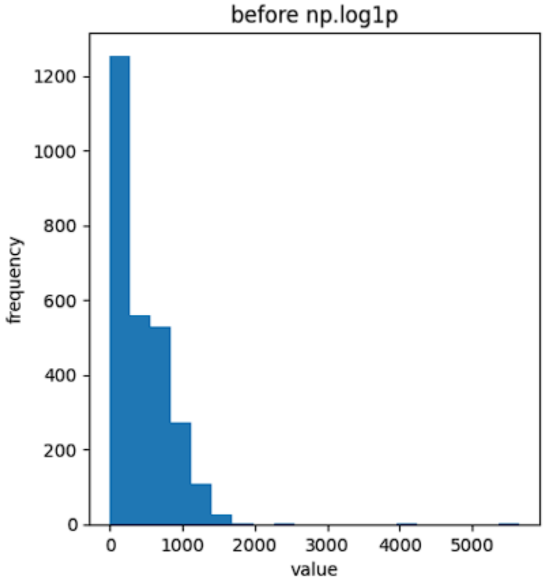
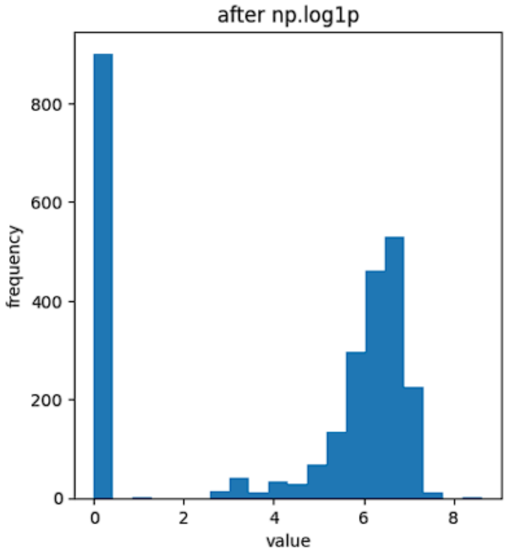
T-Statistic: 22.500369343146456

P-Value: 8.502481661331864e-58

Here the p value is less than 0.05, so we will reject null hypothesis. That means there is a significant association between the Saleprice and houses which have garages.

6. Possible Next Steps for Further Investigation

6.1 Removing skewness

Image 6.1.1	 	Before removing skewness
Image 6.1.2	 	After removing skewness

7. Conclusion

We have used housing data for EDA report. First, Data cleaning task has been done in which, replacing null value with appropriate value (mean, medium) and detecting and removing outlier task has been done. By using Heat map, analyze how variables are correlated with each other. At the end, I checked hypotheses and found some specific variables which are highly correlated with sale price that would be used for predicting sale price.