Yamileth Hercules
100385215
CPSC-4800
November 6, 2022

House Prices Data Analysis Report

Introduction

House Prices dataset has 1840 rows and 81 variables or columns.
The objective of this report is to analyze if the Sale Price is associated with other variables, whether numerical and/or categorical.

Missing Values

- 19 Columns have missing values.

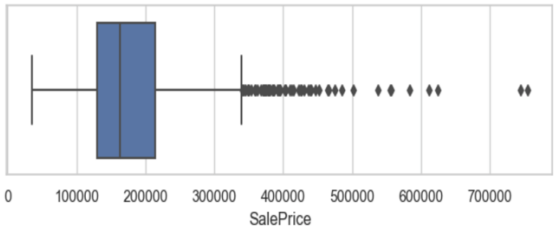
The top 5 columns that have missing values of more than 40% are:

Column	Frequency	Percentage% Missing values
PoolQC	1453	99.52
MiscFeature	1406	96.30
Alley	1369	93.76
Fence	1179	80.75
FireplaceQu	690	47.26

The above values are going to be ignored because the percentage of missing values is high. The rest of the variables were replaced by the subcategory with maximum frequency.

Summary Statistics

- There are 81 variables in dataframe, 36 of them are numerical variables, the rest are categorical.
-



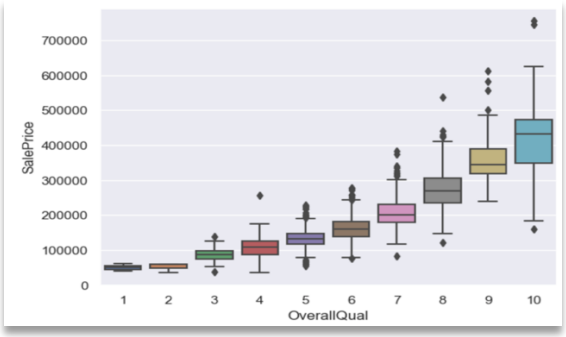
- Houses price ranges from $34,900(min) to $75,500(max). Sale price distribution of houses show maximum sales are between prices $34,179 - $214,925.
- The average price of houses is $180,921.19
- From 81 variables, 37 of them are numerical and the rest are categorical.
- There are 9 numerical variables that have a linear relation with SalePrice.

Hypothesis 1:
Determine if SalePrice is associated with at least 4 numerical variables in the dataset

There are 9 categorical variables that have a quite strong association with Sale price (mentioned in Jupyter Notebook).

- The top 4 relevant with SalePrice are:
  1. OverallQual
  2. GrLivArea
  2. GarageCars
  3. GarageArea

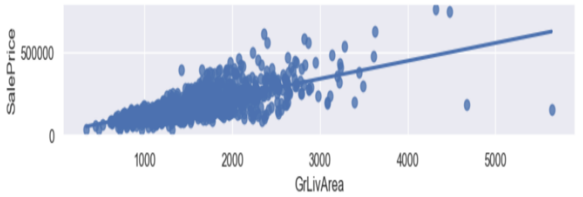
Ordinal Variable

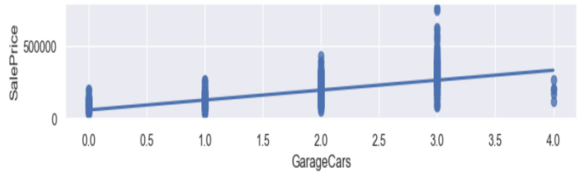


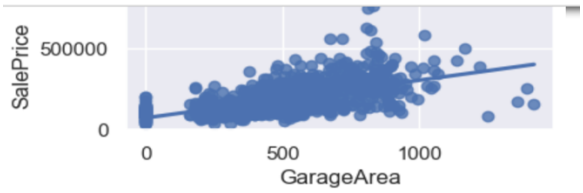
- OverallQual rates the overall material and finish of the house from poor quality to Excellent. OverallQual has a linear correlation with SalePrice even though it is ordinal categorical

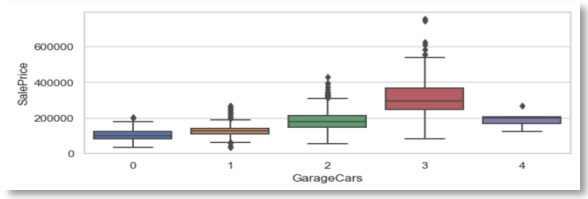
Yamileth Hercules
100385215
CPSC-4800
November 6, 2022

variable. The median increased when the number increased.







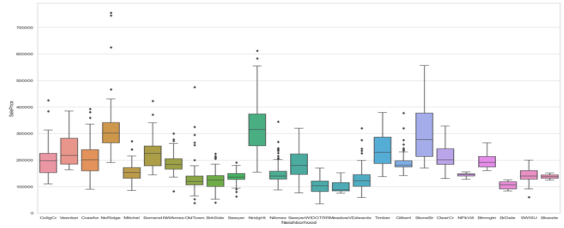


- After 3 garage spaces, the SalePrice begins to decrease.
- The variables mentioned above follow a positive association of linear tendency form, from moderate to strong relation (corr > 0.6) between those variables and SalePrice. It means Saleprice increases if the value is higher.

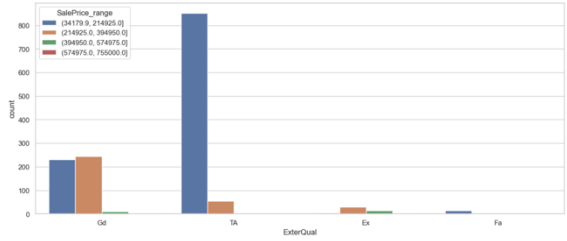
Hypothesis 2:
Determine if SalePrice is associated with at least 4 categorical variables in the dataset

- The top 4 categorical variables that have a higher association with SalePrice that are relevant are:

1.Neighborhood
2.ExterQual
3.KitchenQual
4.BsmtQual



- The distributions of neighborhood shows some extreme values in differents neighborhoods, some Neighborhood make SalePrice bigger than others. The top 3 Neighborhoods with the maximum median values are:
  1. NridgHt 315000.0
  2. NoRidge 301500.0
  3. StoneBr 278000.0



- For different ranking of evaluation, the ranges of prices change. The variables ExterQual, KitchenQual and BsmtQual follow that tendency.

Conclusions

The sale price is related to both numerical variables and categorical variables. Pearson's correlation analysis and Chi squared analysis demonstrated the correlation between them.
To analyze categorical variables, ranges were generated for the SalePrice in order to filter the most relevant ones.