Bemba Munkhchuluun /100405853/

SUMMARY OF TRAIN.CSV

This assignment utilized the Train.csv from Kaggle, which comprises 81 columns and 1,460 rows.

I perform exploratory data analysis (EDA) on this dataset.

1. DATA PREPARATION

Download the Library -> Read csv -> Describe dataset -> Recognize the missing values

1. Despite the dataset's multitude of variables, I have chosen 'SalePrice' as the dependent variable. Key statistics for 'SalePrice' are as follows: mean = $180,921, standard deviation = $79,443, minimum = $34,900, and maximum = $755,000.
2. Since there are missing values in 19 variables, I have decided to exclude these variables from my analysis

2. ANALYZING NUMERICAL VARIABLES

Draw a Histogram and boxplot for SalePrice -> Correlation matrix -> Sort all correlation coefficients related to SalePrice -> Draw pairplot for comparing relationship between variables

1. Based on the histogram (Figure #1) and boxplot (Figure #2), we can observe that the distribution of 'SalePrice' is right-skewed and contains outliers, which have an impact on both the variation and the mean.

Figure #1

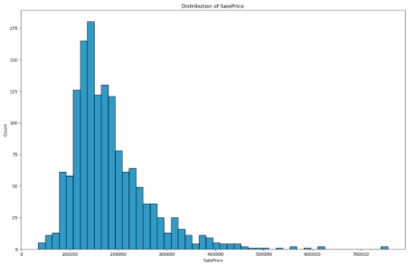
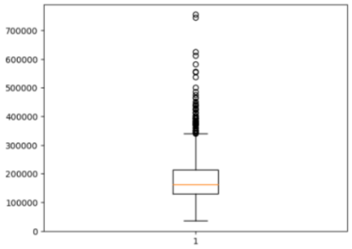
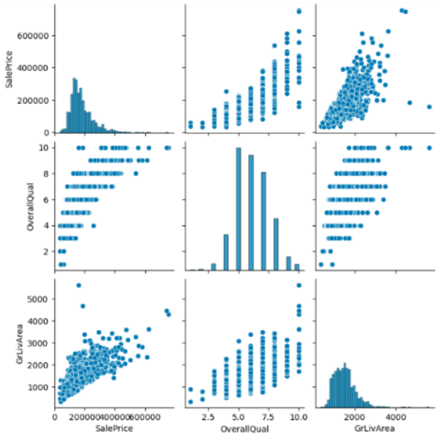


Figure #2



2. For the 'SalePrice' as a dependent variable we can see that there are strong positive correlations, such as with 'OverallQual'(Material and finish quality) and 'GrLivArea' (Ground living area), indicate factors that positively impact. Conversely, 'KitchenAbvGr'(Number of kitchens) and 'EnclosedPorch'(Area), suggest factors that negatively influence sale prices.
3. Pairplot: (Figure #3)
   a. Both the overall quality of a property ('OverallQual') and the size of the living area ('GrLivArea') are positively associated with higher sale prices ('SalePrice')
   b. Outliers in 'GrLivArea' vs. 'SalePrice' might be worth investigating further.
   c. 'SalePrice' is positively skewed, with a tail extending towards higher prices. This suggests that more properties have sale prices on the lower end, while relatively fewer have very high sale prices.
   d. The scatterplots show that 'SalePrice' relationships with 'OverallQual' and 'GrLivArea' could potentially be modeled with linear regression. This suggests that these variables might be good candidates for predicting sale prices.

Figure #3

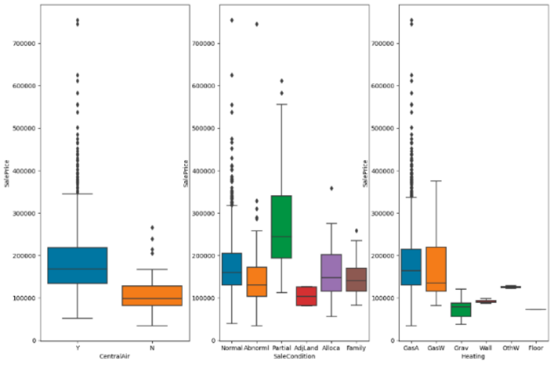


3. ANALYZING CATEGORICAL VARIABLES (Figure #4)

I have created a subplot to compare the similarity between two categorical variables, 'CentralAir' and 'SaleCondition'.

The highest sale prices are typically associated with properties that have central air conditioning (CentralAir: Yes) (SalePrice AVG~180K) or are in a 'Partial' sale condition (SalePrice AVG~240K). We can observe a few outliers in each category, except for the 'SaleCondition' category, which is 'AdLand'.

Figure #4


SalePrice
Y
N
CentralAir
SalePrice
Normal
Abnorml
Partial
AdjLand
Alloca
Family
SaleCondition
SalePrice
GasA
GasW
Grav
Wall
OthW
Floor
Heating