Capstone Project 1 - Notes

Capstone Project 1 - Notes **1**

Kaggle Project **2**

Exploratory Data Analysis **2**

Data Wrangling Exercise **2**

What kind of cleaning steps did you perform? 2

Analyze 2

How did you deal with missing values, if any? 2

Apply Inferential Statistics **3**

Distribution of the Target Variable SalePrice 3

Relation of Features to Target Variable SalePrice 3

Numerical Features 3

Categorical Features 3

Kaggle Project

[House Prices: Advanced Regression Techniques]

Exploratory Data Analysis

Data Wrangling Exercise

What kind of cleaning steps did you perform?

Analyze

1) Load the test and training data set in data frames.
2) Analyze Sale Price with describe(). Check for min, max and missing values.
3) Analyze the correlation between Sale Price and the rest of the variables.
4) Use scatter plots to determine the relationship between Sale Price and numerical variables.
5) Convert categorical variables to numerical values to calculation their correlation to Sale Price.
6) Use box plot to determine the relationship between Sale Price and categorical variables.
7) Use a correlation matrix and heatmap to determine highly correlated variables.

How did you deal with missing values, if any?

1) Identify all the variables with missing values
2) For numerical variables, fill the missing values with mean of the variable.
3) For categorical variables, fill all missing values with ‘None.

Apply Inferential Statistics

Distribution of the Target Variable SalePrice

Charted a histogram against a Normal Probability Plot to check if SalePrice is normally distributed.
Conclusion: The conclusion was the peakness and skewness do not follow the diagonal line.
Action: Some ML regression models assume normal distribution. Hence we need to make a log transformation to SalePrice.

Relation of Features to Target Variable SalePrice

There are a total of **40** numerical features and **43** categorical features.

Numerical Features

We can use scatter plots to plot out the relationship between each of the numerical features and SalePrice. However, the best way is to calculate **Pearson product correlation coefficients** and display the correlation score in a matrix with a colored-heatmap.

With this matrix, we can conclude that the highly correlated variables to SalePrice are - 'OverallQual', 'TotalSF', 'GrLivArea', 'GarageCars', 'TotalBsmtSF', '1stFlrSF', 'FullBath', 'TotRmsAbvGrd', 'YearBuilt'

Categorical Features

We utilize a box plots to visualize the correlations between SalePrice and all the Categorical features.

Based on the box plots, we can identify the categorical features that have strong correlation to sale price log are - 'MSZoning', 'Neighborhood', 'Condition2', 'RoofMatl', 'Exterior1st', 'Exterior2nd', 'ExterQual', 'BsmtQual', 'BsmtCond', 'Heating', 'CentralAir', 'KitchenQual', 'GarageType', 'GarageQual', 'PoolQC', 'MiscFeature', 'SaleType'