Capstone Project 1 - Milestone Report

Project Scope

The scope of this project is to predict home prices in Ames, Iowa. This project is based on a competition on Kaggle, **House Prices: Advanced Regression Techniques**. The goal of this project is to understand and apply the right feature engineering and also practice regression techniques like random forest and gradient boosting.

What is The Problem?

Use the Ames, Iowa housing data set to successfully predict home prices using machine learning.

Who is the Client?

There are several clients.

a) Homebuyers/sellers - with the ability to predict home prices, both home buyers and home sellers are able to optimize their selling or bidding prices.
b) State - with the ability to predict home prices, the state is able to predict the property tax revenue.

Data Set and Data Wrangling

Where is the data from and what's in the data?

The training and test data is provided on Kaggle. The files include data such as sale price, lot area, lot shape, neighborhood, basement condition, year built, heating, central air etc.

In total, there are 83 features in the data set - 40 of them are numerical and 43 are categorical.

The train data set has 1460 rows of data and the test data set has 1459 rows of data.

Missing Data

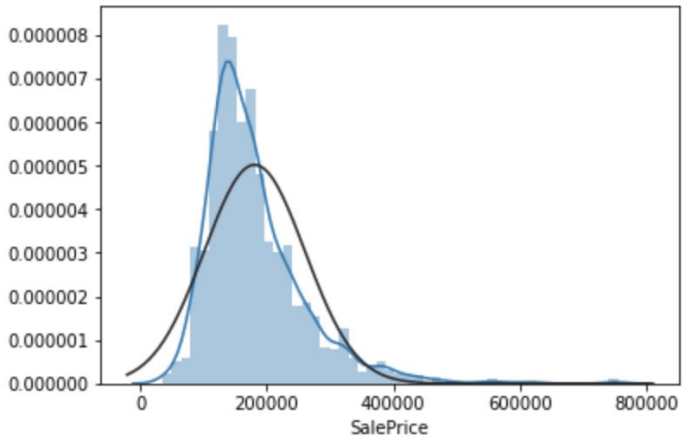
For categorical features that have missing data, fill the missing data with “None”. For numerical features that have missing data, fill the missing data with the mean value of the column. Apply these rules to both train and test data sets.

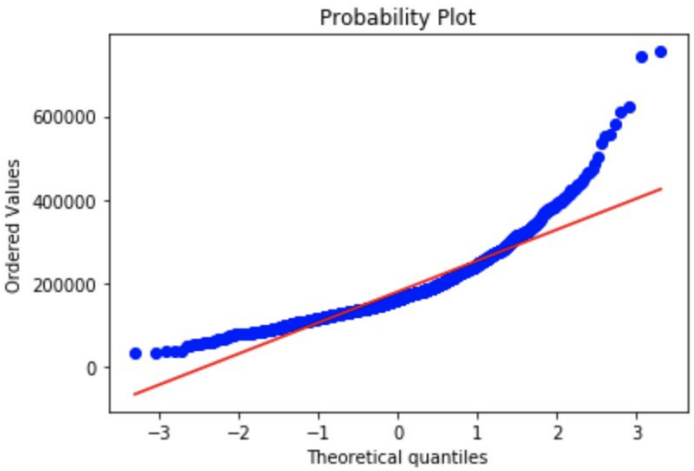
Exploratory Data Analysis

What is the Distribution of Target Feature - SalePrice?

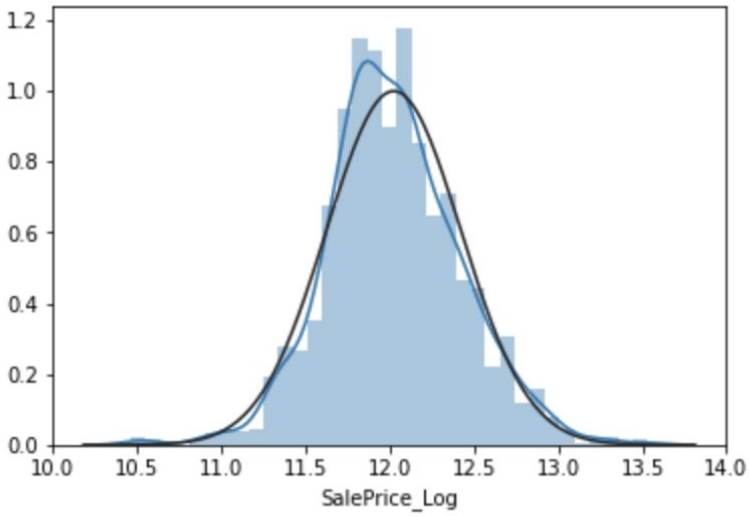
By charting the SalePrice in a histogram against a Normal Probability Plot, we can see that SalePrice is not a normal distribution. The peakness and skewness of SalePrice do not follow the diagonal line of a Normal Probability Plot.

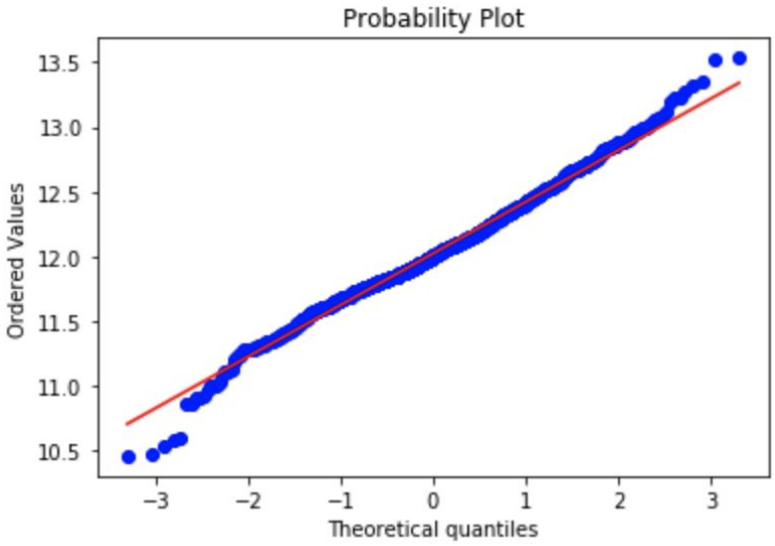
```
sns.distplot(df_train['SalePrice'], fit=norm);
fig = plt.figure()
res = stats.probplot(df_train['SalePrice'], plot=plt)
```





Since certain ML regression models assume normal distribution, we need to make a log transformation to SalePrice. Once that's done, we plot the Log SalePrice histogram against the Normal Probability Plot again to make sure it looks like a normal distribution. Based on the charts below, the log SalePrice is now a normal distribution.

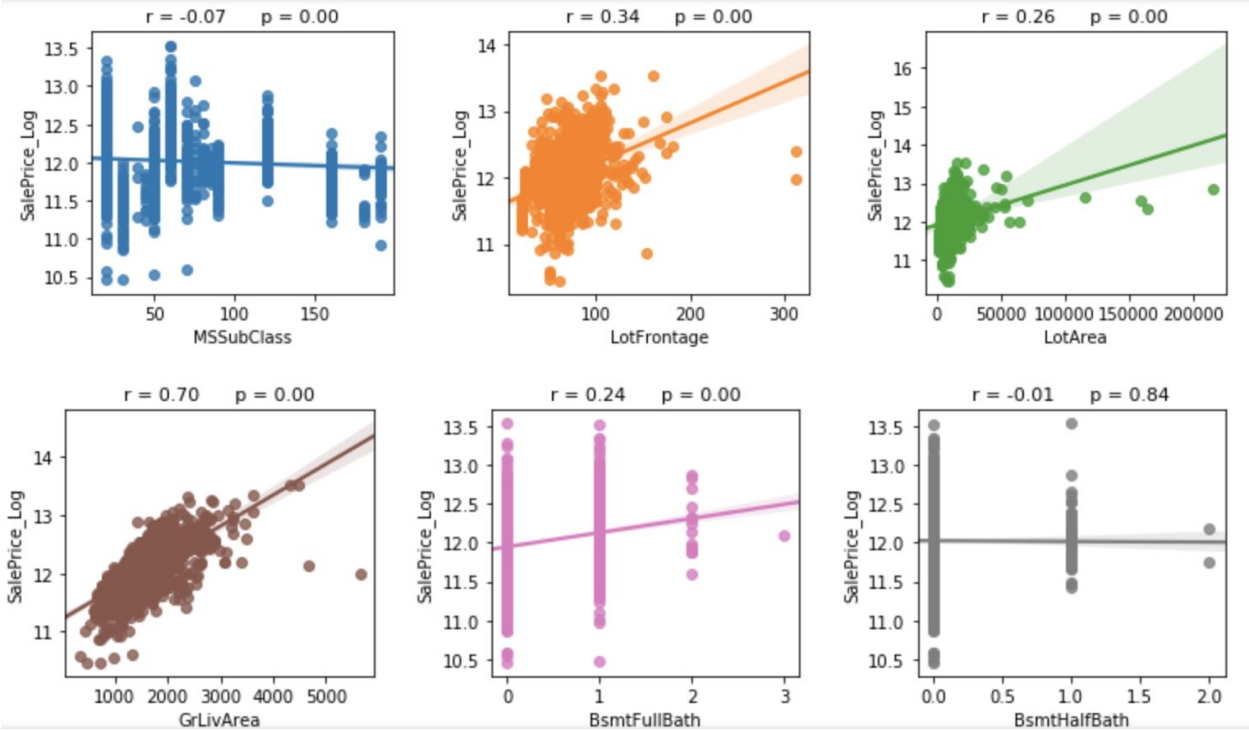




Relationships Between Numerical Features and SalePrice

Visualize The Relationships

Utilize scatter plots to visualize the relationship between SalePrice and each of the numerical features. Use the Pearson correlation coefficient to measure the linear relationship between two datasets. An **r score** is between -1 and +1 with 0 implying no correlation. Correlations of -1 or +1 imply an exact linear relationship. Positive correlations imply that as x increases, so does y. Negative correlations imply that as x increases, y decreases.

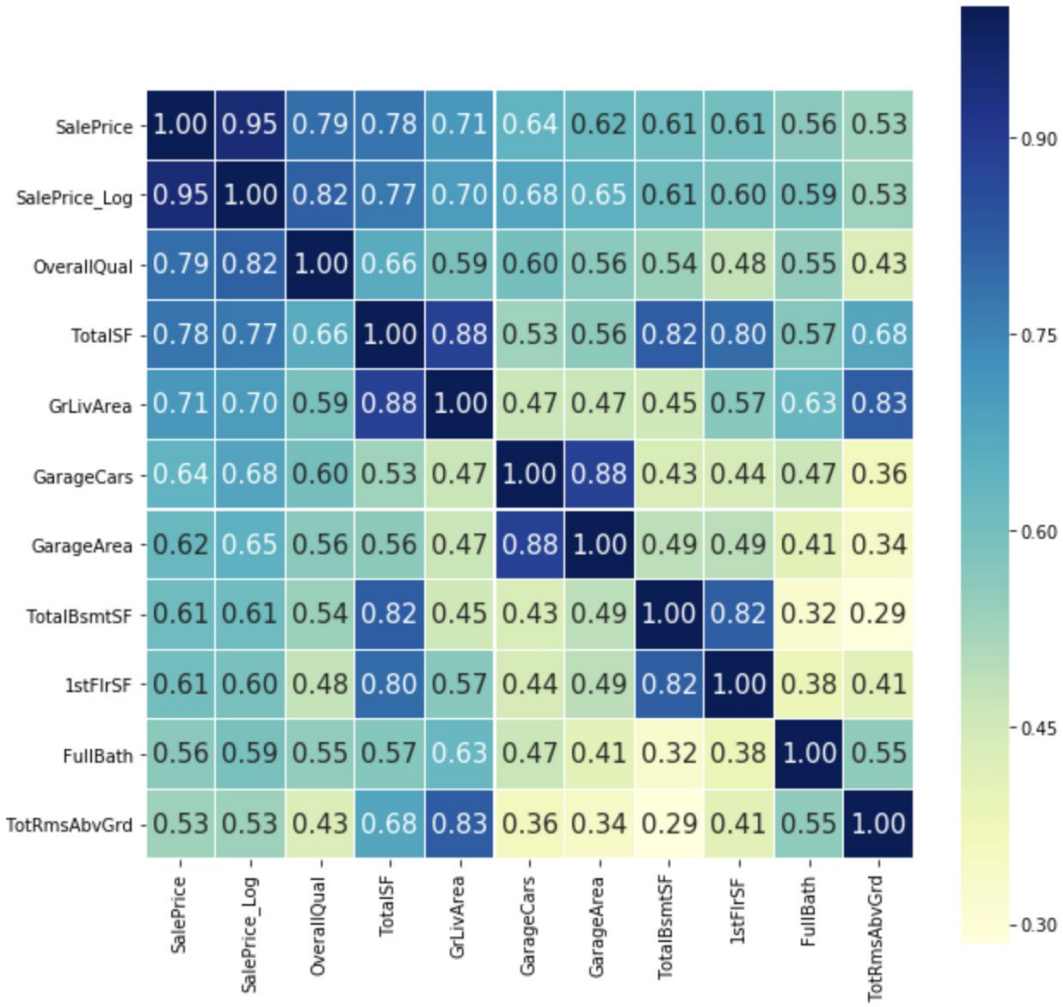


Top Numerical Features with the Highest Correlation to SalePrice

Visualization is great, however it's hard to select the top correlated features when there are so many features. Compute pairwise correlation of columns using DataFrame.corr and it returns a matrix. With this, we can easily select the top 10 features with the highest correlations to SalePrice and plot a heatmap using Seaborn.

```
#Method 1 to get the top 10 correlated variables that are related to SalePrice,
k = 11
cols = corr_matrix.nlargest(k, 'SalePrice')['SalePrice'].index
cols
```

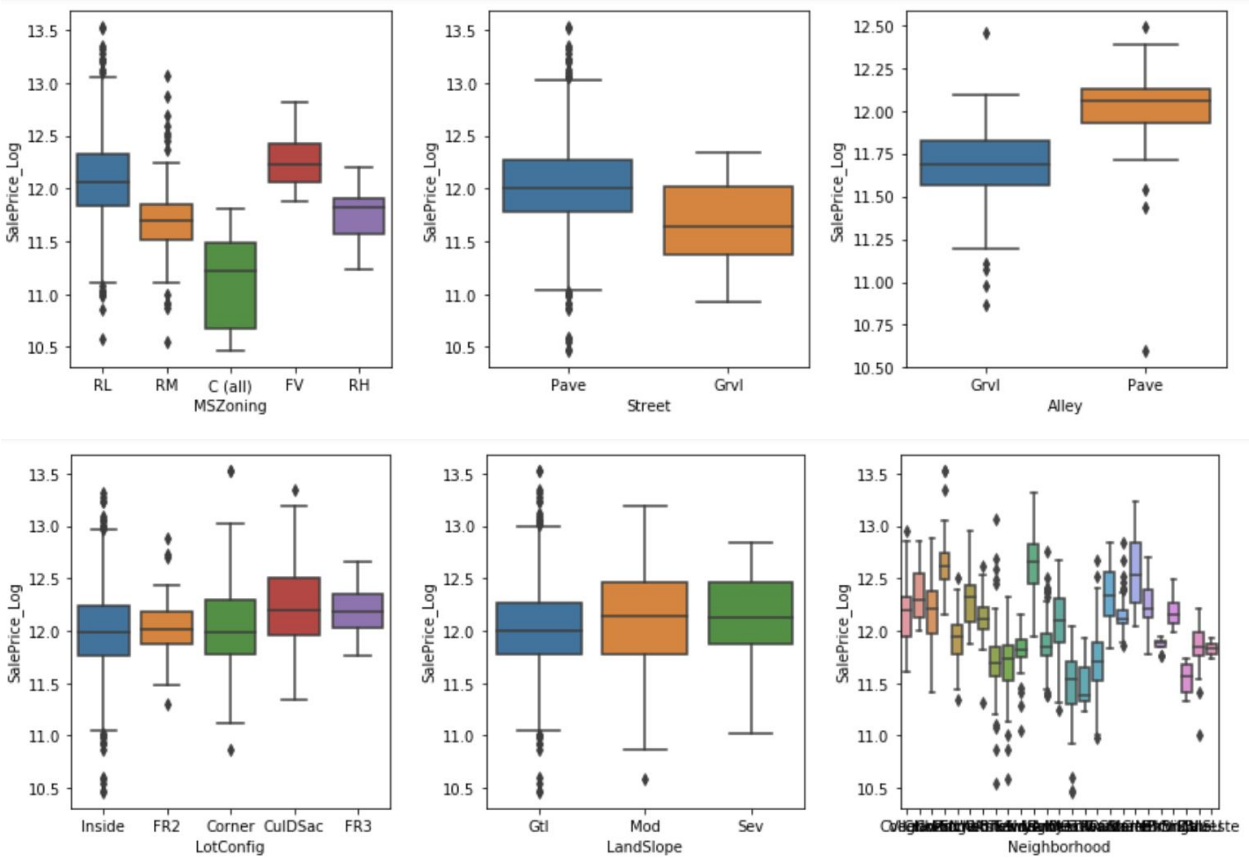
```
Index(['SalePrice', 'SalePrice_Log', 'OverallQual', 'TotalSF', 'GrLivArea',
       'GarageCars', 'GarageArea', 'TotalBsmtSF', '1stFlrSF', 'FullBath',
       'TotRmsAbvGrd'],
      dtype='object')
```



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

Relationships Between Categorical Features and SalePrice

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 SalePrice log are - 'MSZoning', 'Neighborhood', 'Condition2', 'RoofMatl', 'Exterior1st', 'Exterior2nd', 'ExterQual', 'BsmtQual', 'BsmtCond', 'Heating', 'CentralAir', 'KitchenQual', 'GarageType', 'GarageQual', 'PoolQC', 'MiscFeature', 'SaleType'