Project 1

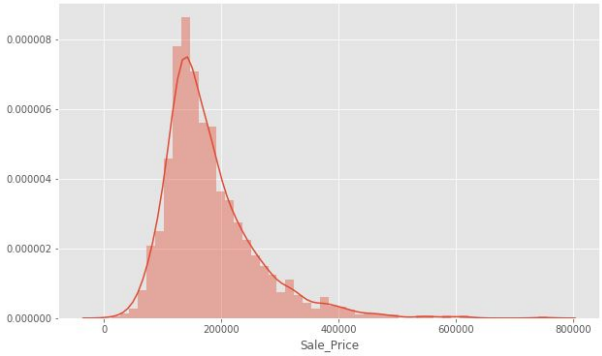
1. Language used Python
2. For feature evaluation and engineering I have used https://www.kaggle.com/leeclemmer/exploratory-data-analysis-of-housing-in-ames-iowa
3. All application depemdencies has been added to pipfile

Find categorical and numerical features

1. We divide the features into categorical and numerical features based on there types.
2. We also drop Longitude and Latitude from our analysis

Sales Price data analysis¶

From the distribution plot for Sales price we see the it is slightly positively skewed. We will do a log transformation for further analysis



Numerical feature data analysis

1. From the plots below we see that several of the numerical features are positively skewed and we will be log transforming these features
2. Couple of features looks like categorical, for example MS_SubClass. We will stringify them and treat as categorical.

Categorical data analysis

1. From plots below we see that several categorical features can be easily converted to numerical features by assigning some kind of rank to various categories. For example Bsmnt_Cond can be categorised as - a Poor-0 b Fair -1 c. Typlical-2 d. Good-3 e. Excellent-4
2. For Year built we will create ranges for the year and assign it ranking. The older the year built the lower the ranking

Fixing Null data

From the plot below we see that Garage_Yr_Blt has null data. We will mark null values with 0

Categorical data transformation

We will convert all categorical data by one hot encoding

Model for prediction¶

For prediction I have used two models from XGBoost package. For the first model I have used a range of values mainly for different sampling rates to find best sampling rate for a conservative model. Also I have used a depth of 8, which may actually result in overfitting For second model I have used parameters which are conservative. Also I have used a range of lower depths so as to avoid overfitting