Sahil Rangwala
MSDS_422

ERD House Prices Write-Up

By importing the training data, it can be placed into a dataframe to hold all of the values. Performing basic descriptive statistics on the entire dataframe offers some insights regarding the differing potential independent variables as well as more information regarding the dependent variable distribution. A possible insight is that the standard deviation is very similar regarding Overall Quality and Overall Condition. This is understandable as both factors are related to one another thus focus should be only one of the features. Further analysis must be done to showcase the relationship regarding overall quality and SalePrice to distinguish the strength of this feature. It is important to note that the mean and median do not correspond to one another, however, more analysis regarding distribution of SalePrice is necessary to see if it follows a normal distribution. For reference, the mean is $180,921, and the median is $163,000.

The overall frequency of SalePrices is graphed to showcase the type of distribution that comes from the dependent variable. From graphing, it does not look like it follows a completely normal distribution thus further transformation is needed to fit a normal curve. Analyzing the dataset further, there exists many NaN's and null values within the different categorical and quantitative categories. After doing calculations to figure the # of nulls in each category, it is seen that PoolQC, MiscFeature, Alley, and Fence have the highest number of null values. This is seen in the table that displays the top 20 categories with missing values.

A way to combat these missing values is to replace the missing values by the mean of the feature. This is a common way to clean and modify a dataset that is polluted heavily with missing values. In this case, the features that have the highest chance of not containing a value are possible features a home will have. So by not having a value in a feature, it indicates that the home does not have that specific occurrence of that feature. For example, having an alley or pool is isolated only to certain homes which will directly affect the price of the home. Not having these features on the home also affects the home price thus replacing these missing values could accidentally skew our results by replacing it with the feature mean. In this case, the empty values will be dropped as there are over 50% missing values for PoolQC, MiscFeature, Alley, and Fence however this could be noted that only within Ames this house feature is not popular.

Feature engineering techniques can be used to help cut out categories that are related to one another and to remove redundancy. By removing redundancy in our dataset, it will help later on when trying to train the model as well as obtaining analysis of the dataset. Additionally, unique features could be created to help showcase other possible insights. Ratio between living area and lot area can be created to create a standard unit for living area comparisons amongst homes. Additionally, both quality and condition factors in the data set are correlated closely with one another and are separately correlated to SalePrice thus combining both factors into a score could help distinguish different homes based on build quality.

Sahil Rangwala
MSDS_422

The SalePrice column does not have any missing values, thus this data as a whole can be used to help create insights. A great way to see what are some possible independent variables is to see the correlation between differing variables and the dependent variable, SalePrice. The top three highest correlated variables are OverallQual with 0.81, GrLivArea with 0.73, and GarageCars with 0.69. Practically, this makes sense as the more quality home would indicate a pricier home. Additionally, more square footage in certain areas could drive up the price. It is interesting to see that the number of car garage has a higher correlation than in general square footage of the garage (with 0.649). The strength of these relationships can further be seen by plotting.

Plotting the correlations, these above strongly correlated variables are seen, as well as other ones such as YearBuilt and YearRemodeled. Practically speaking, this makes sense as the new home or finishing within the home would equate to a higher quality home that corresponds to a higher price. Additionally, looking specifically at OverallQual and YearBuilt, a slight correlation can be seen to back up our possible insight.

By doing calculations to distinguish the outlying data points, on the SalePrice dependent variable there were about 61 outliers that existed from our data. This was obtained by calculating quartile 1 and 3 and then determining the interquartile range. From this range, a norm calculation to indicate an outlier is to see if a data point falls in a region that is 1.5 * IQR more than quartile 3 range and less than quartile 1 range. After doing this calculation, 61 outliers are found. From graphing SalePrice, it is seen to be a rightward skewed graph. Additionally, the skewness value is 1.882876, offering an extremely positive skew. Due to this data and graph not following a normal distribution and the skewness value, different transformation methods could help tailor our data to obtain more insights.

Combined with both the outlier information and the graphing distribution, transformation processes can be applied to manipulate SalePrice to reduce skewness. Initial logarithmic transformation will possibly correct our skewness and reduce the overall outliers. After applying this transformation, the skewness value dropped significantly to 0.121347. Additionally, the graph follows more of a normal distribution than it did from before. Furthermore, the overall number of outliers also dropped by more than half as there exists only 28. By applying this logarithmic transformation, it led to SalePrice offering less outliers and a more normal distribution. Further analysis with min-max scaling and standard scaling is down below that as well. It is understandable that the Ames data is skewed without any manipulations as this area is a popular college town that will not offer as much home diversification compared to a bigger city. Increasing the data set to encompass more cities or towns could help have more of a holistic representation of home data within Iowa.