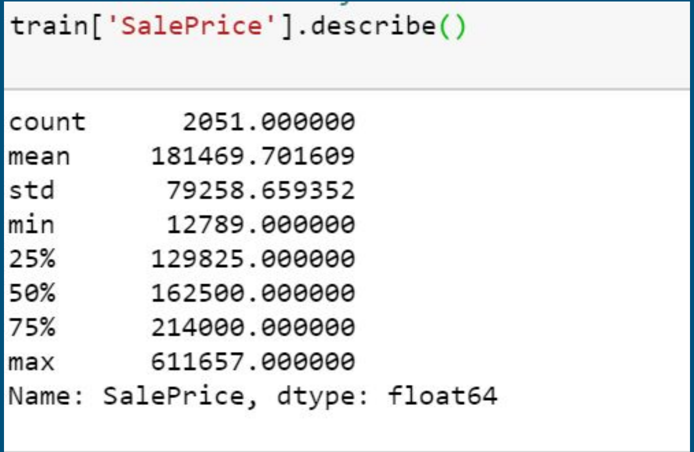
Regression and Classification with Ames Housing Data

A report by Sapna M.K

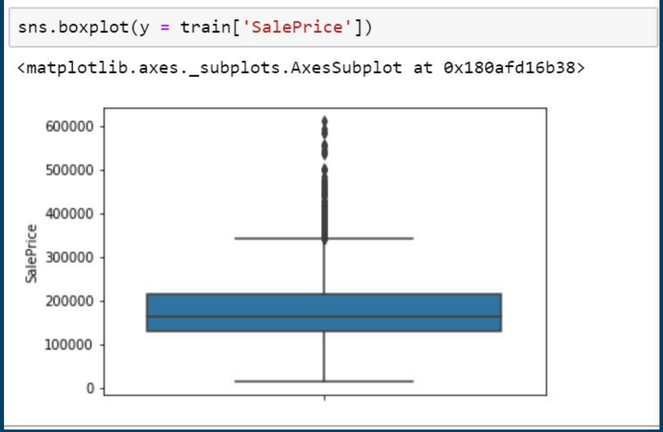
Initial EDA showed that the number of features had missing values which need to managed appropriately before modelling.

A mean and median of the Sale Price is 180,000 and 160,000 respectively, indicating a positive skew which can be confirmed by both histogram and box plot. The 75th percentile was also considerably less than the maximum sale price



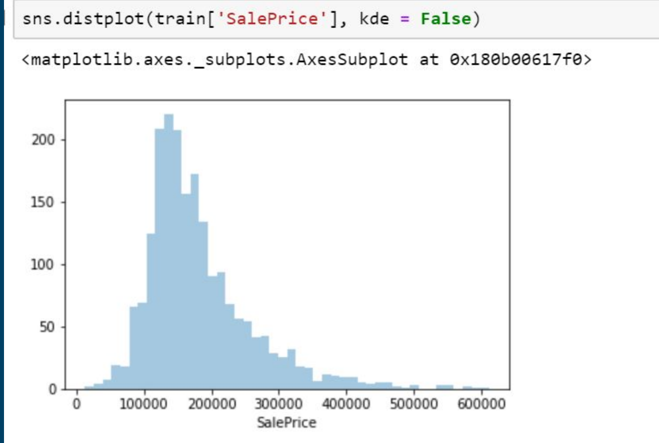
```
train['SalePrice'].describe()
```

```
count      2051.000000
mean     181469.701609
std       79258.659352
min       12789.000000
25%      129825.000000
50%      162500.000000
75%      214000.000000
max      611657.000000
Name: SalePrice, dtype: float64
```



```
sns.boxplot(y = train['SalePrice'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x180afd16b38>
```



```
sns.distplot(train['SalePrice'], kde = False)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x180b00617f0>
```

Cleaning the Train dataset.

To clean the null values

For Continous data, with outliers , the null values are filled with Median of the particular column.

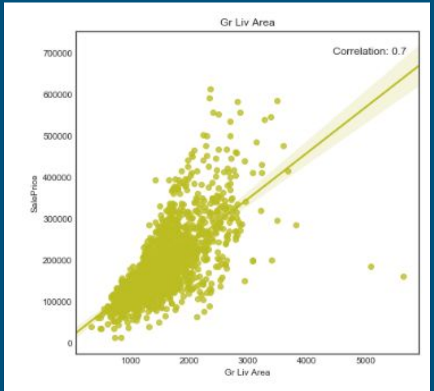
WHereas, the categorical features, we can fill the missing values with the most common value that is the mode of the column

To clean the object data types

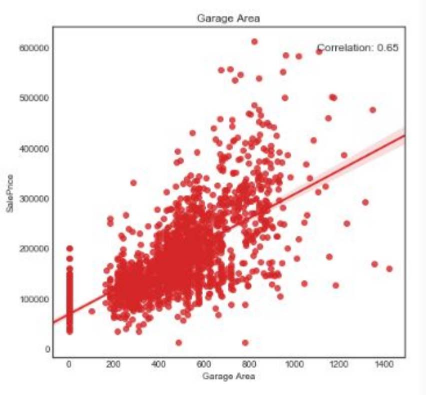
Converting the categorical column into a one-hot encoded matrix using pandas get_udummies method.

Whereas the ordinal data are grouped them in the dictionary using mapping method.

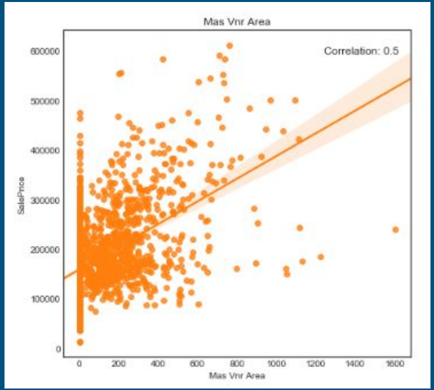
Cleaning and EDA of numerical, fixed features to identify features for inclusion in model



From the above scatter plot, we can prefer there is high correlation of Sale Price and Ground Living Area.



Can see positive corelation between Garage Area and SAle Price.



Can see positive corelation between Masonary Veneer Area and SAle Price.

Train and Evaluate a linear regression model to predict Sale prices using fixed features

Linear Regression Model Score : 85.355%

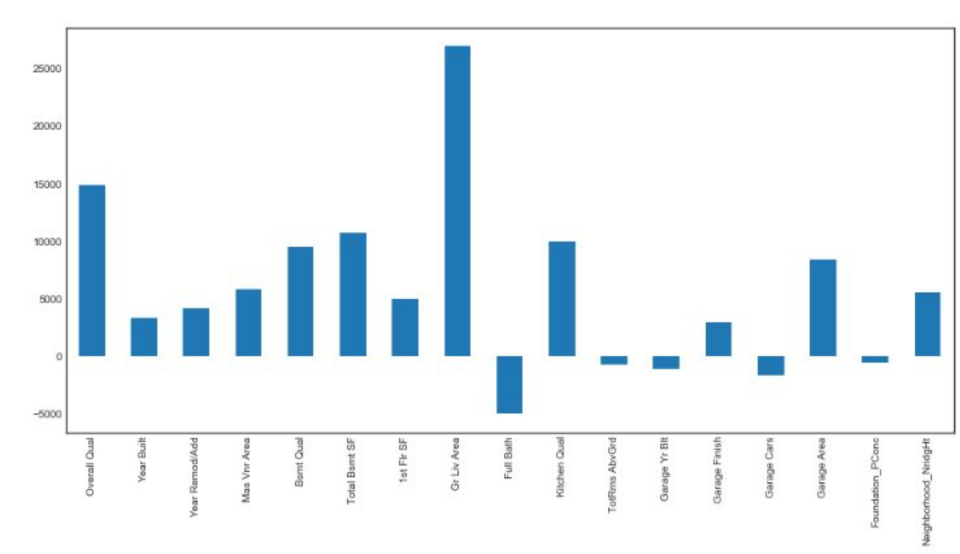
Ridge Regression Model Score : 85.367%

Lasso Regression Mode lScore: 85.355%

After Instantiating the Linear, Ridge and Lasso Regression model, we did fit on the training data and transform on both train and test data

The Ridge Regression model has

Shows the relationship between the Sale Price and the co-efficients of the selected features.



Solution

From the above histogram, we can infer that the sale price is positively co-related with the overall quality of the house which includes kitchen and Basement. So, if the builder work towards this feature, the sale price will increase.

https://docs.google.com/presentation/d/1K-zXStBLQy7ze4V3Wsz-lAC_l-KFWj5VGWHa8MRlSDY/edit?usp=sharing