Project 2 – Ames Housing Data and Kaggle Challenge

Oscar Goh
GA DSI 22

Problem statement

To build the best regression model, using the Ames Housing Dataset, to predict the price of a house at sale

Datasets

Train set:

- 2051 observations
- 81 features

Test set:

- 879 observations
- 80 features

```
# Check Loaded datasets
traindf.shape
```

```
(2051, 81)
```

```
testdf.shape
```

```
(879, 80)
```

Exploratory Data Analysis

- Check for null values
- Explore ways to handle null values

```
# Null count in percent sorted in descending order
null_percent = round((null_sum / 2051), 3)
null_percent.sort_values(ascending = False)
```

```
Pool QC           0.996
Misc Feature      0.968
Alley             0.932
Fence             0.805
Fireplace Qu      0.488
Lot Frontage      0.161
Garage Yr Blt     0.056
Garage Cond       0.056
Garage Qual       0.056
Garage Finish     0.056
Garage Type       0.055
Bsmt Exposure     0.028
```

Data cleaning

- Fix null values
- Check and process categorical features
- Check and process numerical features

Data Cleaning

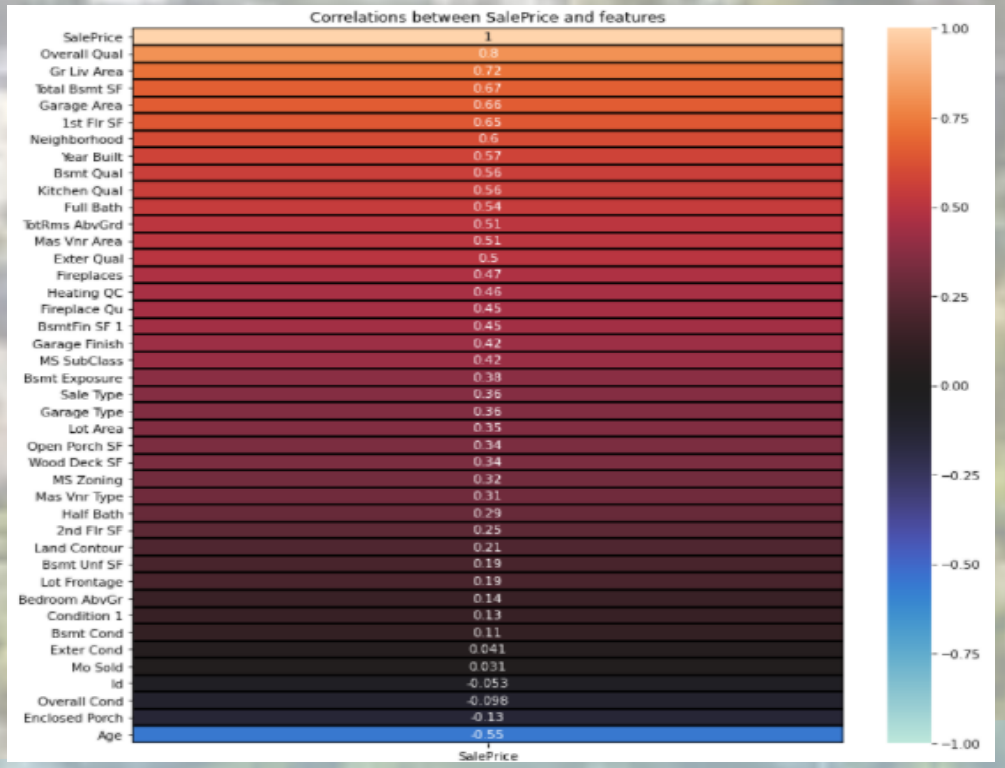
Summary of fixing null values

```
# Check categorical features
objectdf = traindf.select_dtypes(include=['object'])
for column in objectdf.columns:
    print(column)
    print(objectdf[column].value_counts(dropna = False))
    print('\n')
```

```
#Check numerical features
numberdf = traindf.select_dtypes(include=['number'])
for column in numberdf.columns:
    print(column)
    print(numberdf[column].value_counts(dropna = False))
    print('\n')
```

Exploratory Visualizations

- Scatterplot
- Histogram
- Boxplot
- Heatmap
- Remove outliers
- Feature engineering



Pre-processing

- Assign X and y
- Train test split
- Scaling

```
# Performing scaling
ss = StandardScaler()
train_scaled = ss.fit_transform(trainnumdf)
```

```
# Splitting into train and test sets
X_train, X_holdout, y_train, y_holdout = train_test_split(X, y, random_state = 69)
```

```
X = finaltraindf[['Overall Qual', 'Gr Liv Area', 'Garage Area', 'Total Bsmt SF', '1st Flr SF', 'Age', 'Full Bath',
                  'Mas Vnr Area', 'TotRms AbvGrd', 'Fireplaces', 'BsmtFin SF 1', 'Neighborhood', 'Year Built',
                  'Bsmt Qual', 'Kitchen Qual', 'Exter Qual','Heating QC', 'Fireplace Qu','Garage Finish', 'MS SubClass']]
y = finaltraindf['SalePrice']
```

Modeling

- Linear regression
- Ridge regression
- Lasso regression

Model	Train Score (RMSE)	Test Score (RMSE)
Linear Regression	25388.43	27181.15
Ridge Regression	25393.6	**27091.35**
Lasso Regression	25392.61	27145.27

Conclusion & Summary

- Top 5 features

- Bottom 5 features

As observed from the coefficients from ridge model, the 5 features that add the most value to a home are

1. Exter Qual: Exterior material quality
2. Kitchen Qual: Kitchen quality
3. Gr Liv Area: Above grade (ground) living area square feet
4. Neighborhood: Physical locations within Ames city limits
5. Overall Qual: Overall material and finish quality

The features that hurt the value of the home the most are

1. Age: Age of house
2. Full Bath: Full bathrooms above grade
3. TotRms AbvGrd: Total rooms above grade (does not include bathrooms)
4. Year Built: Original construction date
5. Bsmt Qual: Height of the basement