Data Wrangling

The Data

For my initial capstone project ("House Prices: Advanced Regression Techniques"), the primary data source is the Kaggle.

Analyzing the Dataset

Initial data shape: 1460 rows, 81 columns
dtypes: float64(3), int64(35), object(43)
Null Values: Yes, some of the qualitative fields have null values (e.g. LotFrontage, Alley, MasVnrType, MasVnrArea, etc.). Fill the null values with empty strings (for now).

Outliers: Drop the outliers

- What kind of cleaning steps did you perform?

Now that I have got a general idea about my data set, it's also a good idea to take a closer look at the data itself. With the help of the head() and tail() functions of the Pandas library, I can easily check out the first and last lines of your DataFrame, respectively. Let us look at some sample data

```
train.describe()
train.head()
train.tail()
train.shape
pd.isnull(train).any()
```

- How did you deal with missing values, if any?

```
# Visualising missing values
categorical_features = train.select_dtypes(include=[np.object])
Categorical_features.columns
```

```
df = train
#for back fill
df.fillna(method='bfill',inplace=True)
df = train
#for front fill
df.fillna(method='ffill',inplace=True)


train['LotFrontage'].fillna(0, inplace=True)
```

- Were there outliers, and how did you handle them?

```
#Drop observations where GrLivArea is greater than 4000
sq.fttrain.drop(train[train.GrLivArea>4000].index, inplace = True)
train.reset_index(drop = True, inplace = True)

#Drop observations where TotlaBsmtSF is greater than 3000 sq.ft
train.drop(train[train.TotalBsmtSF>3000].index, inplace = True)
train.reset_index(drop = True, inplace = True)

#Drop observations where YearBulit is less than 1893 sq.ft
train.drop(train[train.YearBuilt<1900].index, inplace = True)
train.reset_index(drop = True, inplace = True)
```