# DATA SCIENCE SPECIALIZATION

## Capstone Project

# HOUSING SALES PRICE PREDICTION OF AMES, IOWA

VISHNU SIVADAS

20-05-2020

# Contents

- Introduction
- Data description
- Data sources
  - Data cleaning and preparation
  - Methodology
- Exploratory Data Analysis
  - Training and fitting model
  - Results
- Discussions
- Conclusion

# Introduction

- Business Problem
  - Lots of factors to be considered before buying houses
  - Wild guess often result in bad business decisions
  - To create a model for predicting housing sales price for Ames, Iowa
  - All the important features including the neighbourhood venues to consider
- Target Audience
  - House aspirants who can roughly estimate the value of a house
  - Real estate people and city planners
  - House sellers who can optimize their advertisements.

# Data description

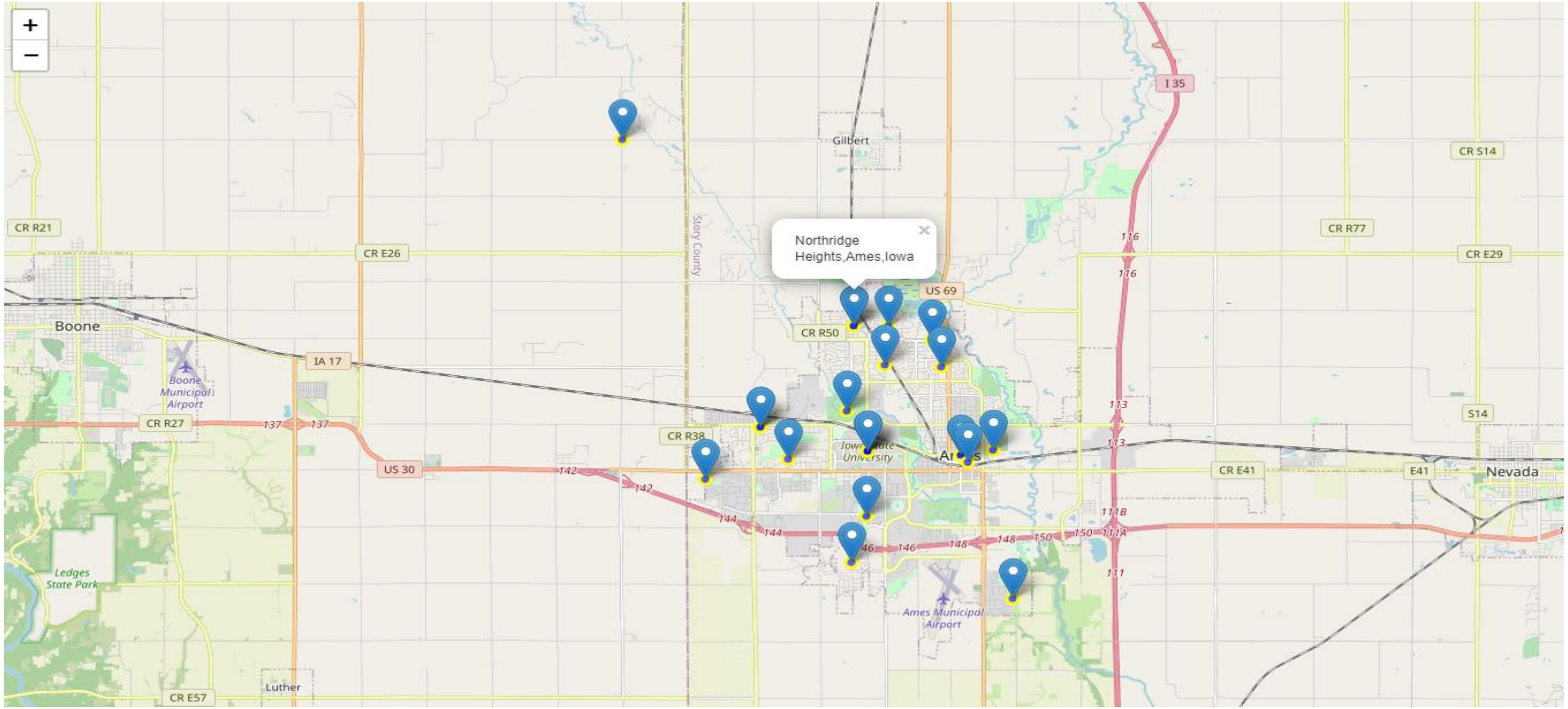
- Data Sources
  - The Ames Housing dataset is taken from Kaggle.com which was compiled by Dean De Cock
  - Foursquare API is used to get the most common venues of Ames, Iowa
  - Geopy library used to get location details of neighborhood
- Data Cleaning and preparation
  - Geopy library used to get location details of neighborhood
  - For each neighborhood, pass the obtained coordinates to FourSquare API
  - Apply one hot encoding to turn each venue type into a column with their occurrence as the value

# Data description

- Data Cleaning and preparation
  - In the Ames Housing dataset, there are multiple features which have missing values and most of the features are object.
  - Some missing values are intentionally left blank in categorical type variables. Those values are kept to 'None'.
  - The "OverallCond", "OverallQual" and "Zoning class" of the house are not numerical. They are converted into categorical variables.
  - Important years and months that should be categorical variables not numerical are converted.
  - Columns with one or two missing rows are dropped
  - Features with large number of missing features and unimportant are dropped

# Methodology

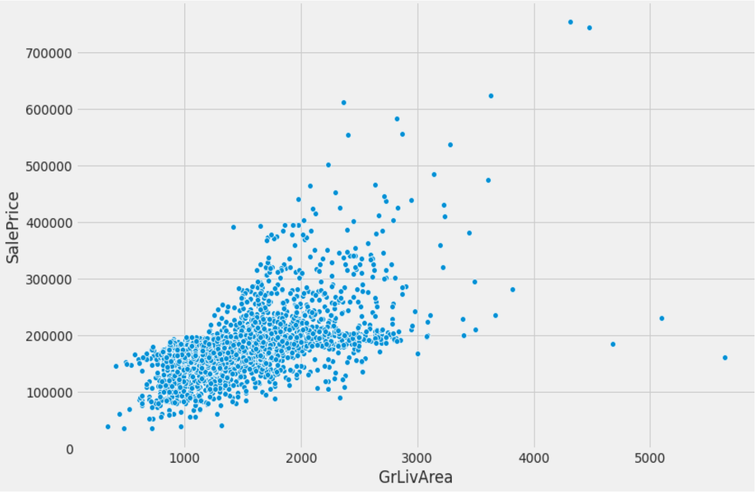
- Exploratory Data Analysis



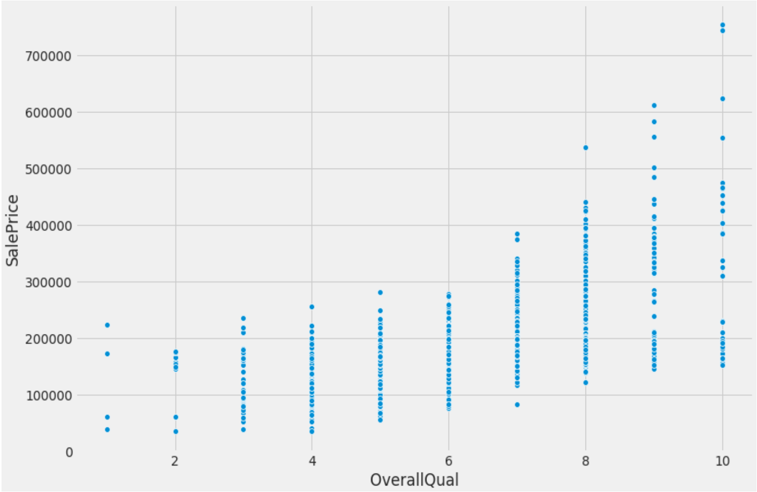
**Ames Neighborhood**

# Methodology

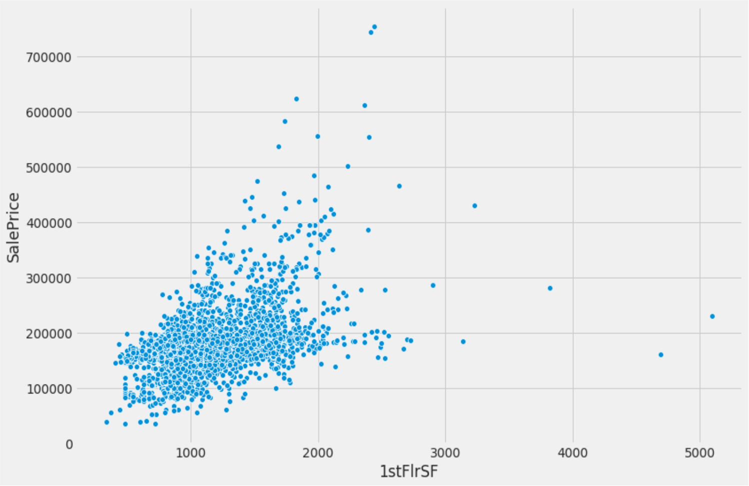
- Exploratory Data Analysis
  - Target variable, SalePrice is not normally distributed
  - Target variable is right-skewed and there are multiple outliers
  - Many outliers found in the scatter plots of features vs target variable
  - Target variable shows an unequal level of variance across most predictor variables.



SalePrice vs GrLivArea



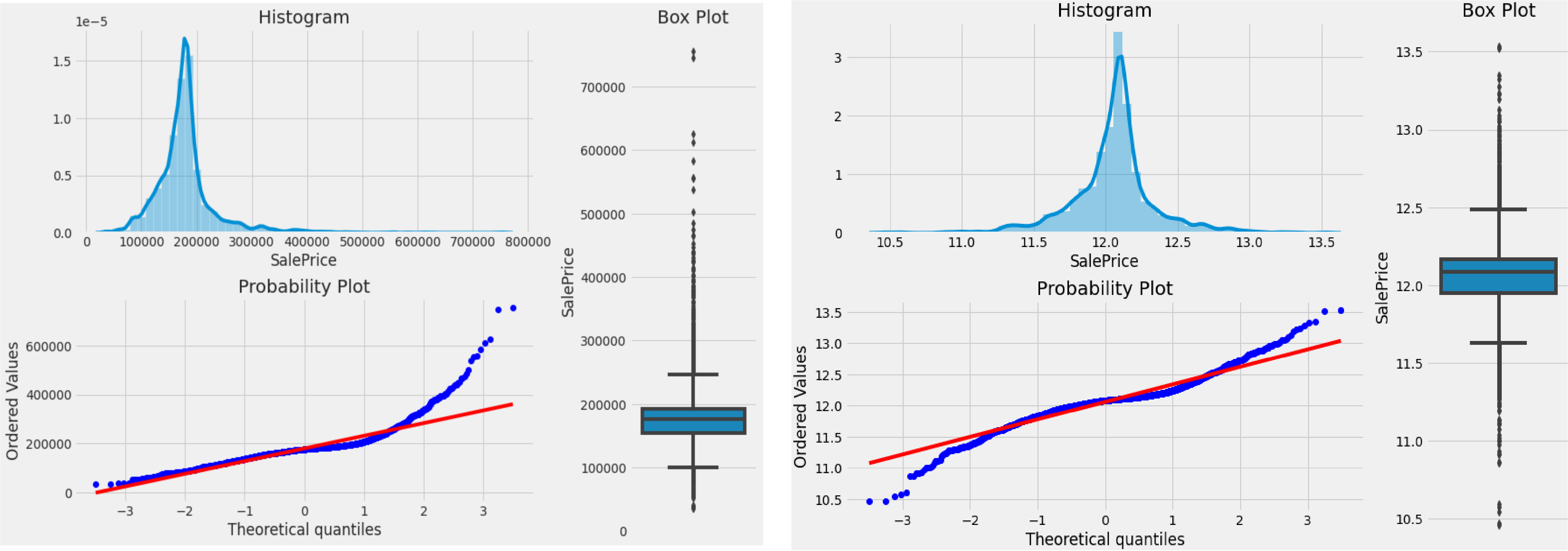
SalePrice vs OverallQual



SalePrice vs 1stFlrSF

# Methodology

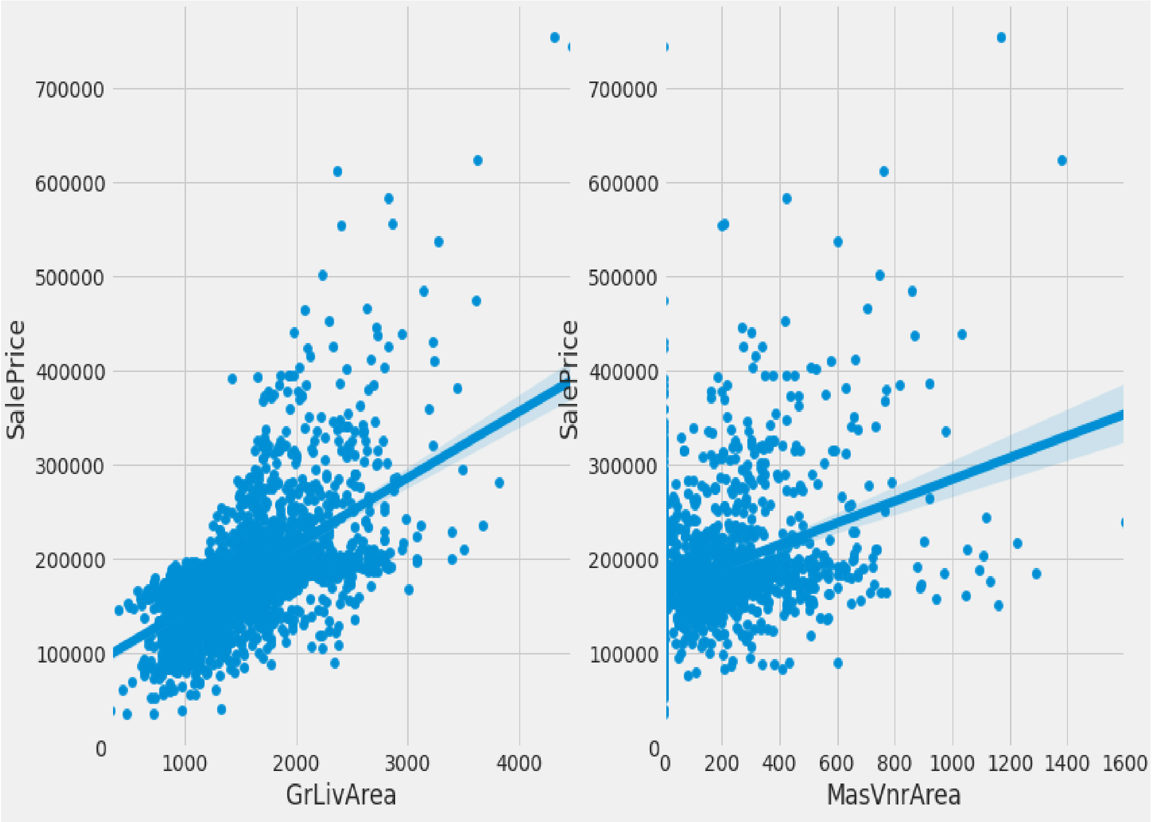
- Exploratory Data Analysis



**Histogram, Boxplot and Q-Q plot of SalePrice before and after log transformation**

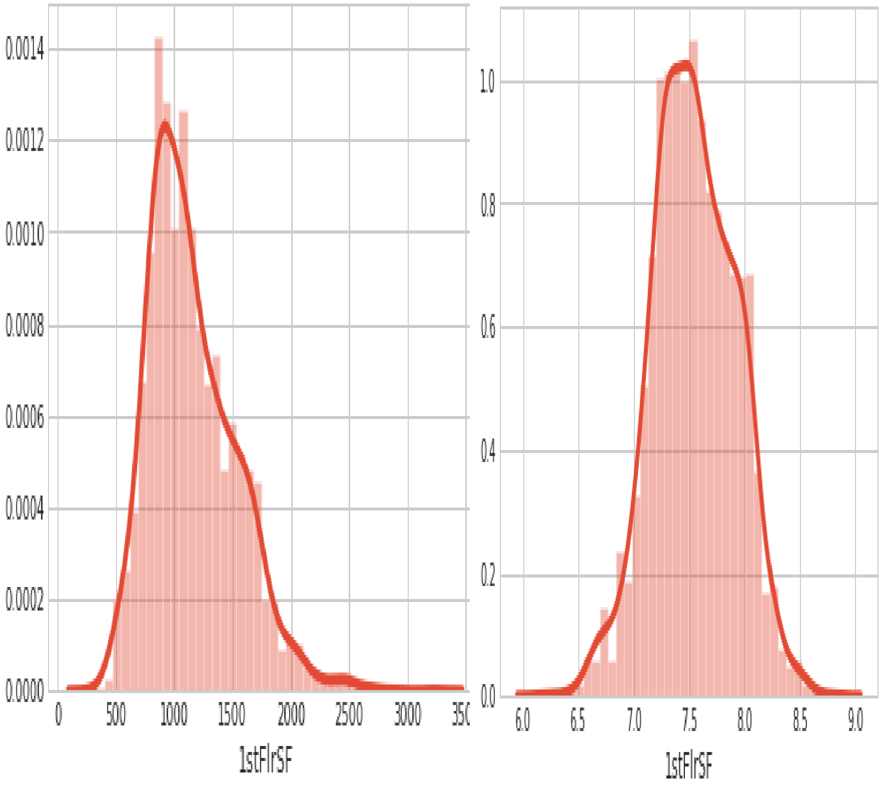
# Methodology

- Exploratory Data Analysis



**Regplot of SalePrice vs GrLivArea and**

**SalePrice vs MasVnrArea**



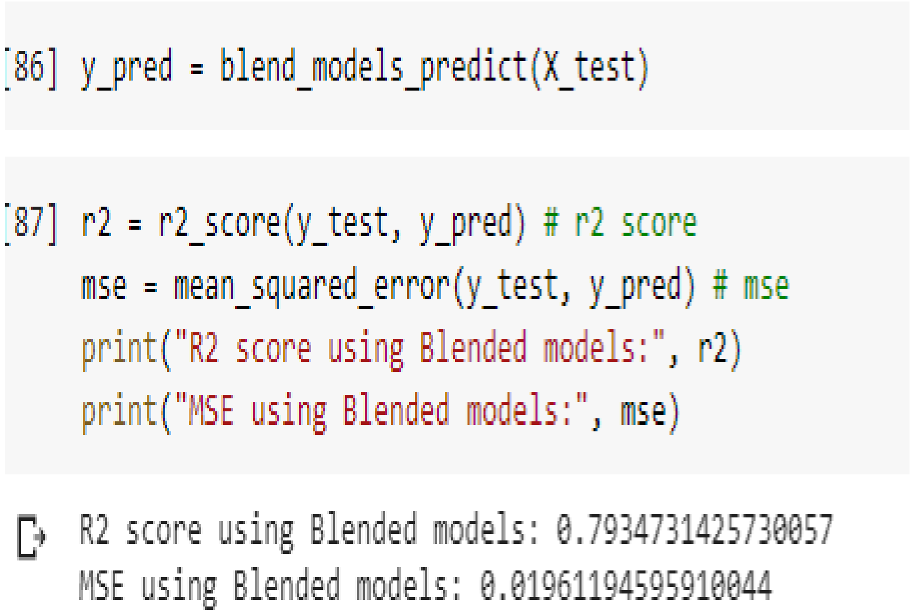
**Distribution plot of 1stFlrSf before and after removing skewness**

# Methodology

- Training and fitting model
  - The overall data is split into training and test data such as two-third of the data is training data using train_test_split function of scikit-learn library.
  - Machine learning linear regression models are used for training with R2 score and Mean Squared Error as evaluation metric
  - The model is trained using Linear Regression which uses Ordinary least squared method. But as we have seen there is multicollinearity between feature variables. Therefore advanced regularisation algorithms are used
  - When the advanced regression models Ridge, Lasso or ElasticNet was used individually, the result didn’t improve satisfactorily. Then a blended model of Ridge, Lasso, Elasticnet, SVR, XGBRegressor, LGBMRegressor and StackingCVRegressor was used with its individual weightage by trial and error to get good working model.

# Results

- Using Linear Regression we got very low R2 score 0.42
- A blended model of Ridge, Lasso, Elasticnet, SVR, XGBRegressor, LGBMRegressor and StackingCVRegressor to get a R2 score of 0.79 and Mean Squared Error of 0.019.
- The model has good accuracy and a low error.
- 79% R2 score means the model is able to explain 79% of the response data around its mean.
- Hence this model for predicting housing sales price for Ames, Iowa considering all the important features including the neighbourhood venues can be used for future predictions.



```
[86] y_pred = blend_models_predict(X_test)
```

```
[87] r2 = r2_score(y_test, y_pred) # r2 score
     mse = mean_squared_error(y_test, y_pred) # mse
     print("R2 score using Blended models:", r2)
     print("MSE using Blended models:", mse)
```

```
R2 score using Blended models: 0.7934731425730057
MSE using Blended models: 0.01961194595910044
```

**Results of final model**

# Discussions

- There is almost 37% improvement in model by using advanced regression algorithms.
- But there is still variance that the model could not explain. There are features and but the data set contained only samples.
- For linear regression problems, normal distribution, skewness and outliers play an important role in creating accuracy. These problems are solved to a great extent by transformation methods.
- More data, larger number of datasets, would help improve model performances significantly.

# Conclusion

- Following steps were followed in project:
  - Identifying business problem which is creating a model for predicting housing sales price for Ames, Iowa considering all the important features including the neighbourhood venues
  - Sourcing data  required for the project
  - Cleaning dataset
  - Analysing the data using various visualisation and statistical techniques,
  - feature engineering to optimise the model
  - Training and fitting the model, analysing the model and lastly
  - Recommendations for ways to improve the model
- House aspirants, Real estate people, city planners and house sellers, target audience of this project, can use the model to accurately predict housing sale price of Ames, Iowa.

# Thank You