IBM – COURSERA

DATA SCIENCE SPECIALIZATION

CAPSTONE PROJECT – FINAL REPORT

on

HOUSING SALES PRICE PREDICTION OF AMES, IOWA

Submitted by

VISHNU SIVADAS

20/05/2020

Contents

List of Figures

Chapter 1

INTRODUCTION

1.1 Background

Ames is a city in Story County, Iowa, United States, located approximately 30 miles (48 km) north of Des Moines in central Iowa. It is best known as the home of Iowa State University (ISU), with leading agriculture, design, engineering, and veterinary medicine colleges. Housing prices of this area depends on a lot of factors. For the people who are looking for buying a house or somebody who wants to sell a house, making a wild guess is difficult and often results in bad business decisions. In this project a model is created to tackle the same.

1.2 Business Problem

When we ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. There are a lot of features to be considered before one can set the price or start negotiating. The project aims in creating a model for predicting housing sales price for Ames, Iowa considering all the important features including the neighbourhood venues.

1.3 Target Audience

- House aspirants who can roughly estimate the value of a house based on its features and the average price.
- Real estate people and city planners who can decide what kind of venues to put around their products to maximize selling price.
- House sellers who can optimize their advertisements.

Chapter 2

DATA DESCRIPTION

2.1 Data Sources

Data sets are prepared from the following sources:

- The Ames Housing dataset is taken from Kaggle.com which was compiled by Dean De Cock for use in data science education. It consists of 79 explanatory variables describing various aspect of residential homes in Ames, Iowa.
- Foursquare API is used to get the most common venues of Ames, Iowa. There is a categorical variable 'Neighborhood' in Ames housing dataset. Using this variable and 'geopy' library in python, latitude and longitude of neighbourhoods' is found which in turn is used for finding nearby venues using Foursquare API.

2.2 Data Cleaning and Preparation

In the Ames housing dataset, each neighborhood is given a code, for example 'Blmngtn' was given for 'Bloomington Rd'. Using the code directly 'geocode' could not translate it onto the required latitude and longitude. Further data was given in 'kaggle.com' describing the neighborhood code into neighborhood name. Same was extracted and made into a Data frame which was passed into 'geocoder' for translation after concatenating 'Ames,Iowa'.

Still there were some locations whose translation could not be run by 'geocoder'. These were searched in the web and following actions were taken:

- If the name is different, decide which one to use after searching on the internet.

- If the neighborhood is missing from the geo data frame, add it's coordinate.
- If the neighborhood is made up, combine them into the larger neighborhood which exist in the geo data frame.

Figure 2.1: Neighborhood location data of Ames

[6]:

	Neighborhood	Neigh	Latitude	Longitude
0	Blmngtn	Bloomington Rd,Ames,Iowa	42.056049	-93.625519
1	Blueste	Bluestem,Ames,Iowa	42.011170	-93.645063
2	BrDale	North Grand mall,Ames,Iowa	42.049331	-93.622661
3	BrkSide	Brookside,Ames,Iowa	42.026770	-93.617055
4	ClearCr	Clear Creek,Ames,Iowa	41.787650	-93.267011
5	CollgCr	College Creek,Ames,Iowa	42.020616	-93.693098
6	Crawfor	Crawford,Ames,Iowa	42.028029	-93.607151
7	Edwards	Edwards,Ames,Iowa	42.025819	-93.668553
8	Gilbert	Zenorsville,Ames,Iowa	42.107206	-93.717999
9	IDOTRR	Ames,Ames,Iowa	42.027910	-93.644644
10	MeadowV	Meadow Village,Ames,Iowa	42.026770	-93.617055

The Foursquare API is used to explore the neighbourhoods' and segment them. The limit was set as 100 venue and the radius 1500 meter for each neighborhood from their given latitude and longitude information. Here is a head of the list Venues name, category, latitude and longitude information from Foursquare API.

Figure 2.2: Venue list sample of Ames

```
[14]: print(ames_venues.shape)
      ames_venues.head()
```

(886, 7)

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Blmngtn	42.056049	-93.625519	Ge-Angelo's Italian Restaurant	42.054871	-93.622739	Italian Restaurant
1	Blmngtn	42.056049	-93.625519	Bar	42.054446	-93.622896	Bar
2	Blmngtn	42.056049	-93.625519	Flame & Skewer	42.049287	-93.622321	American Restaurant
3	Blmngtn	42.056049	-93.625519	Anytime Fitness	42.054720	-93.622980	Gym / Fitness Center
4	Blmngtn	42.056049	-93.625519	Victoria's Secret PINK	42.049032	-93.622383	Lingerie Store

'One hot encoding' was done the 'Venue Category' and grouped by 'Neighborhood 'to make the required data set.

Figure 2.3: Neighborhood data set

	Neighborhood	Accessories Store	American Restaurant	Arcade	Art Museum	Asian Restaurant	Athletics & Sports	Auto Garage	Automotive Shop	Bakery	Bank	Bar	Baseball Field	Bed & Breakfast
0	Blmngtn	1	1	0	0	0	0	0	0	0	1	1	0	0
1	Blueste	0	1	1	0	1	0	0	0	1	0	6	0	0
2	BrDale	1	1	0	0	0	0	0	0	0	1	1	0	0
3	BrkSide	0	2	0	0	0	0	0	0	2	2	1	0	0
4	CollgCr	0	0	1	0	0	0	0	0	0	0	0	0	0
5	Crawfor	0	2	0	0	0	0	1	1	2	2	1	2	0

In the Ames Housing dataset, there are multiple features which have missing values and most of the features are object.

Figure 2.4: Missing values sample in Ames housing dataset

```
traindf.isnull().sum()
```

```
LotFrontage       486
LotArea             0
Street              0
Alley            2721
LotShape            0
LandContour         0
Utilities           2
LotConfig           0
LandSlope           0
Neighborhood        0
Condition1          0
Condition2          0
BldgType            0
HouseStyle          0
OverallQual         0
OverallCond         0
YearBuilt           0
YearRemodAdd        0
RoofStyle           0
RoofMatl            0
Exterior1st         1
Exterior2nd         1
MasVnrType         24
MasVnrArea         23
ExterQual           0
```

Some missing values are intentionally left blank in categorical type variables, for example: In the Alley feature, there are blank values meaning that there are no alley's in that specific house. 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.

'BsmtFinSF1', 'BsmtFinSF2', 'BsmtUnfSF', 'TotalBsmtSF', 'BsmtFullBath', 'BsmtFullBath', 'BsmtHalfBath', 'GarageCars' have one or two values. Those rows are respectfully dropped. 'LotFrontage' and 'MasVnrArea' feature has a lot of missing values. These columns are dropped. Some of the unimportant features like 'PoolQC','MSSubClass' are also dropped.

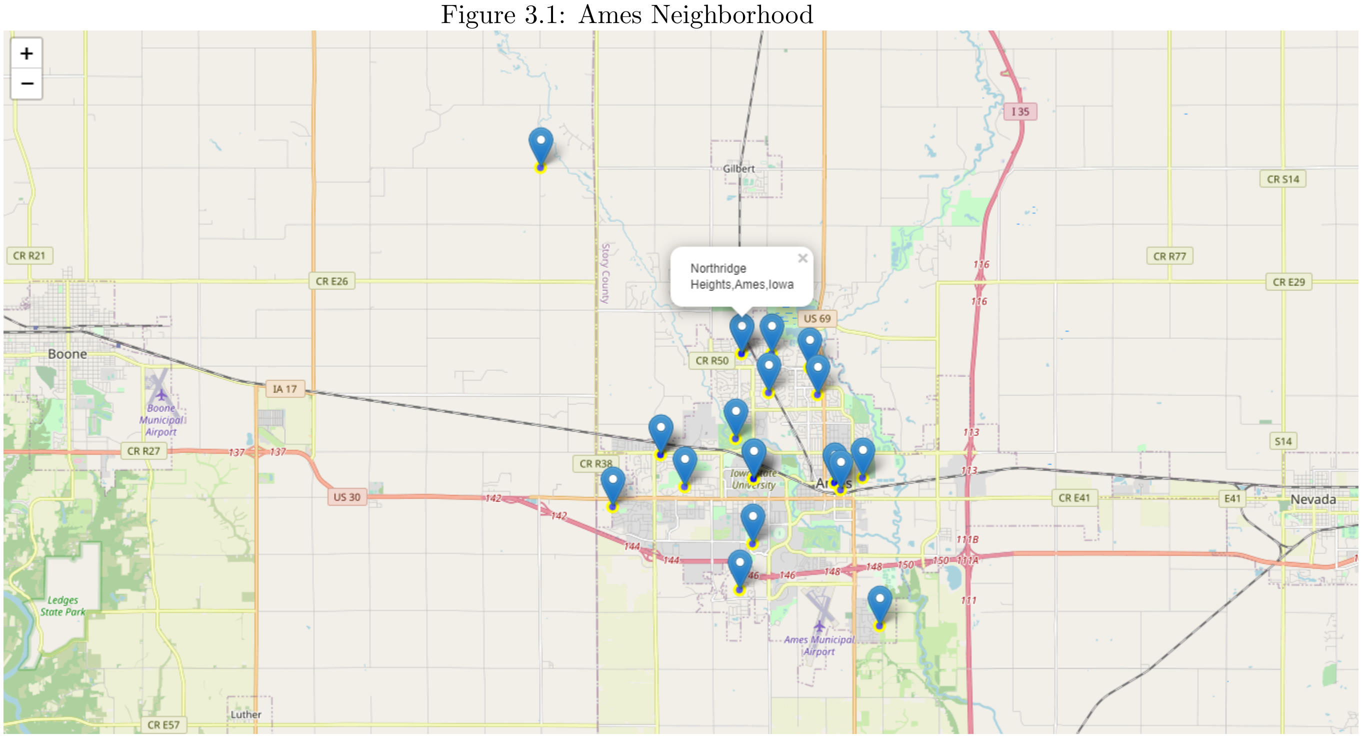
Chapter 3

METHODOLOGY

3.1 Exploratory Data Analysis

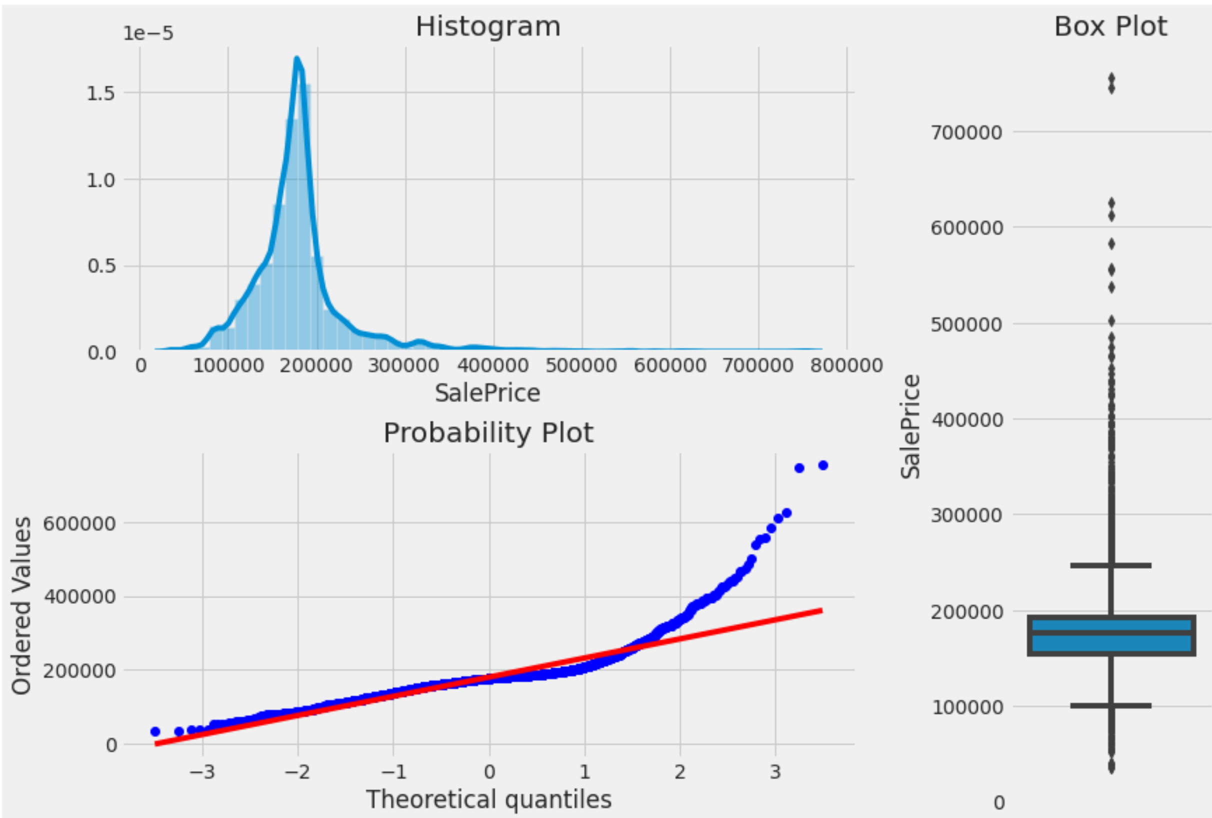
Folium library was used to visualise the neighborhood data of Ames city.

Figure 3.1: Ames Neighborhood



Target variable is ‘SalePrice’ which is available in the Ames housing data. Histogram, Boxplot and Q-Q plot is used to analyse ‘SalePrice’ as shown in Fig.

Figure 3.2: Histogram, Boxplot and Q-Q plot of SalePrice



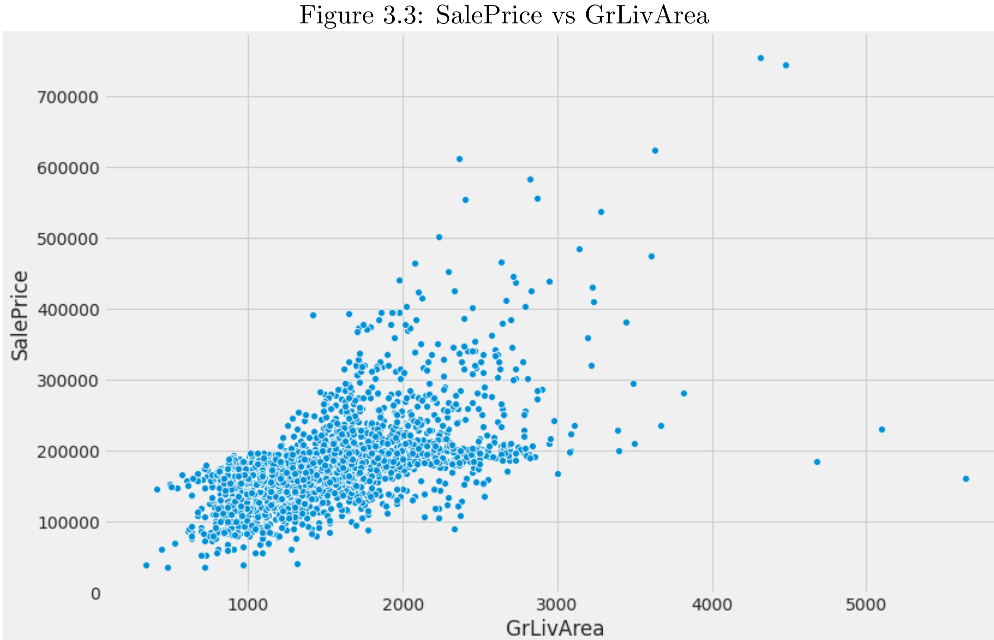
From the above figure it is observed that:

- Our target variable, SalePrice is not normally distributed.
- Our target variable is right-skewed.
- There are multiple outliers in the variable.

Correlation between other features and target variable 'SalePrice' is done and some features are plotted on a scatter plot for better understanding of the data.

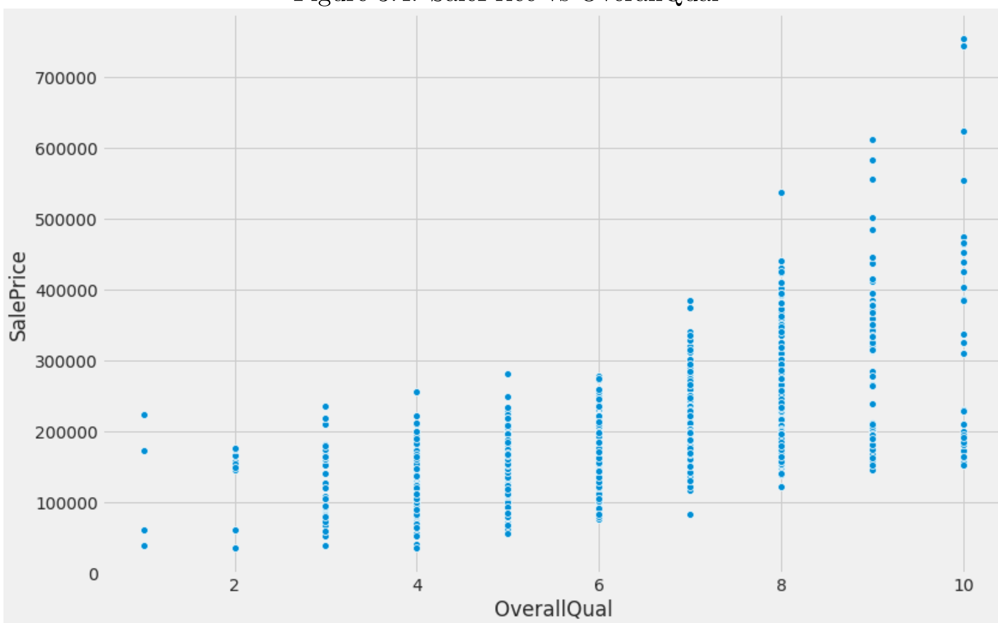
SalePrice vs GrLivArea

Figure 3.3: SalePrice vs GrLivArea



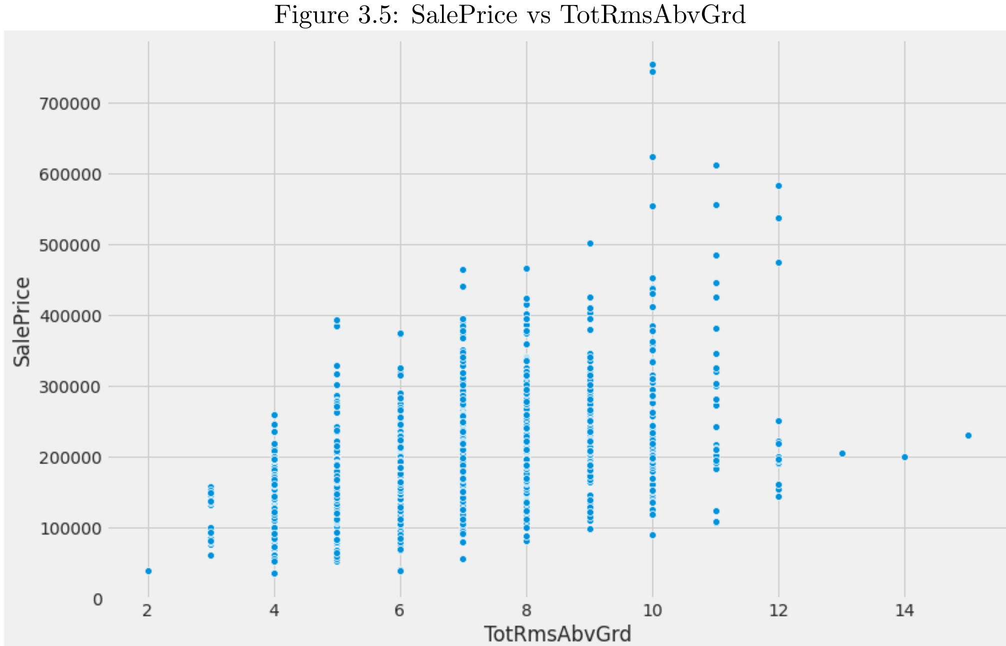
SalePrice vs OverallQual

Figure 3.4: SalePrice vs OverallQual



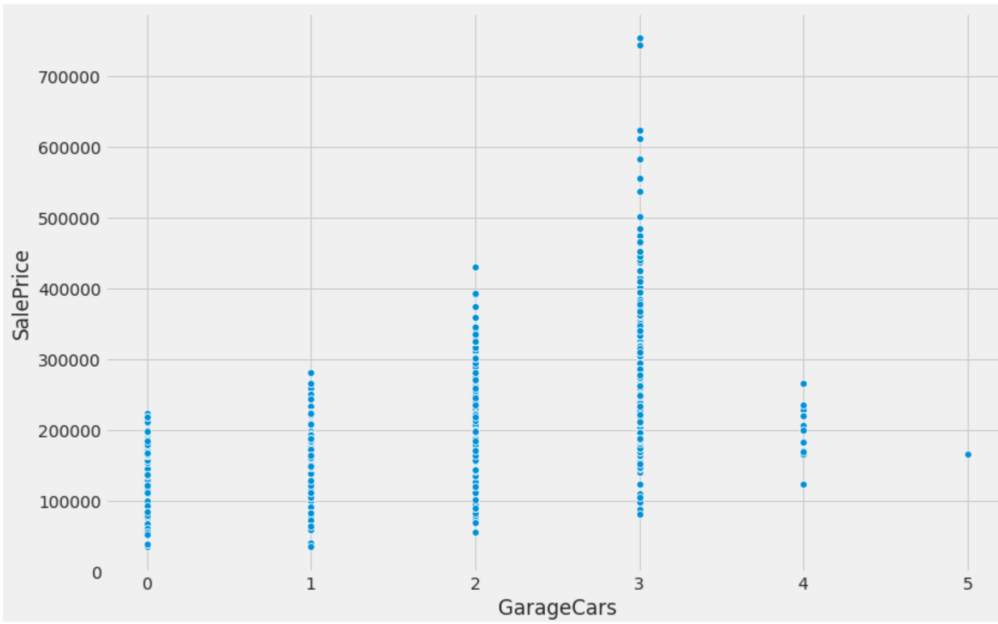
SalePrice vs TotRmsAbvGrd

Figure 3.5: SalePrice vs TotRmsAbvGrd



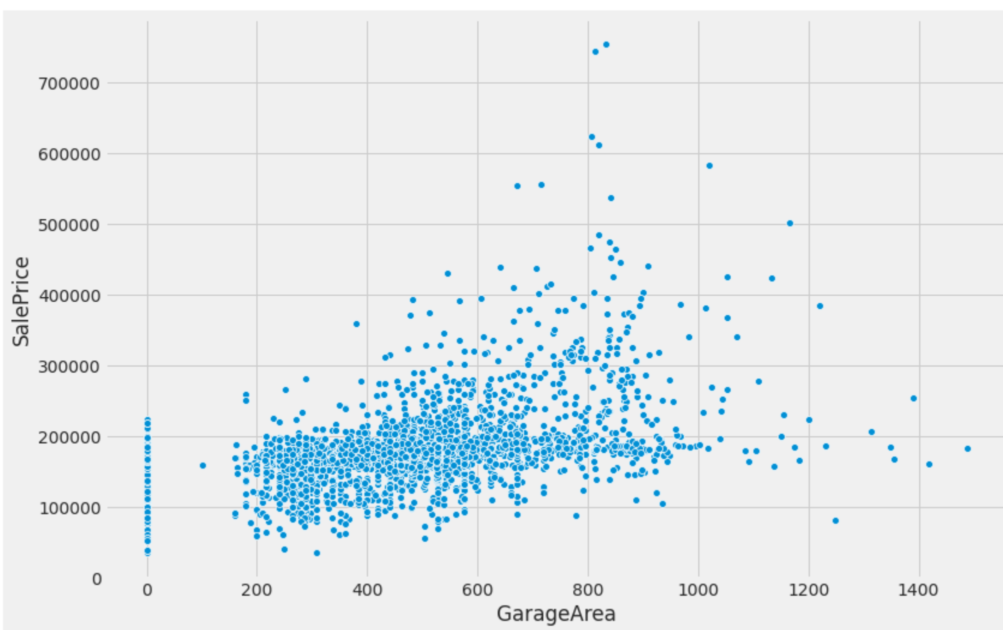
SalePrice vs GarageCars

Figure 3.6: SalePrice vs GarageCars



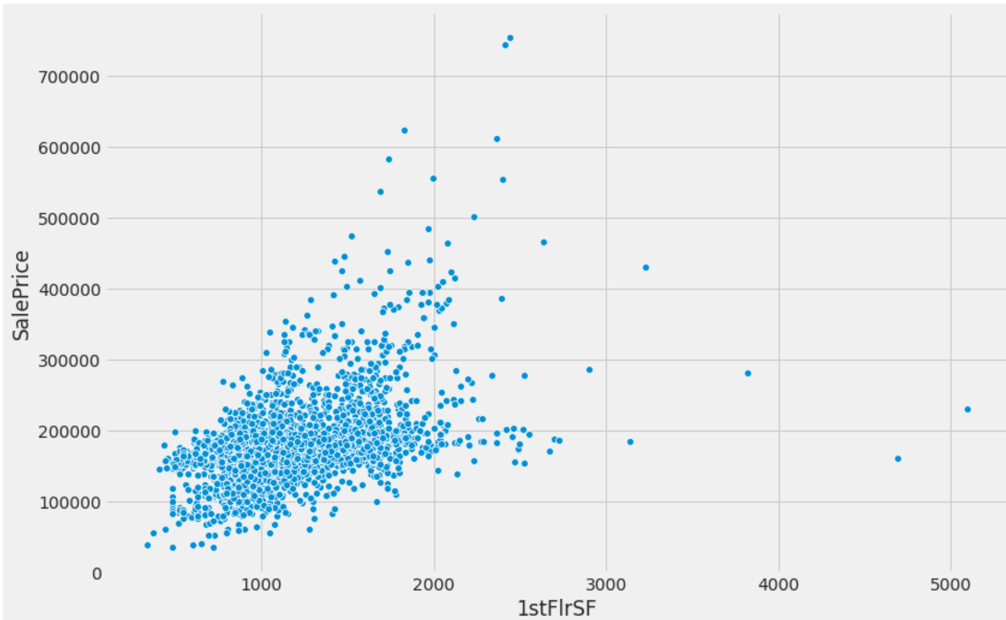
SalePrice vs GarageArea

Figure 3.7: SalePrice vs GarageArea



SalePrice vs 1stFlrSF

Figure 3.8: SalePrice vs 1stFlrSF

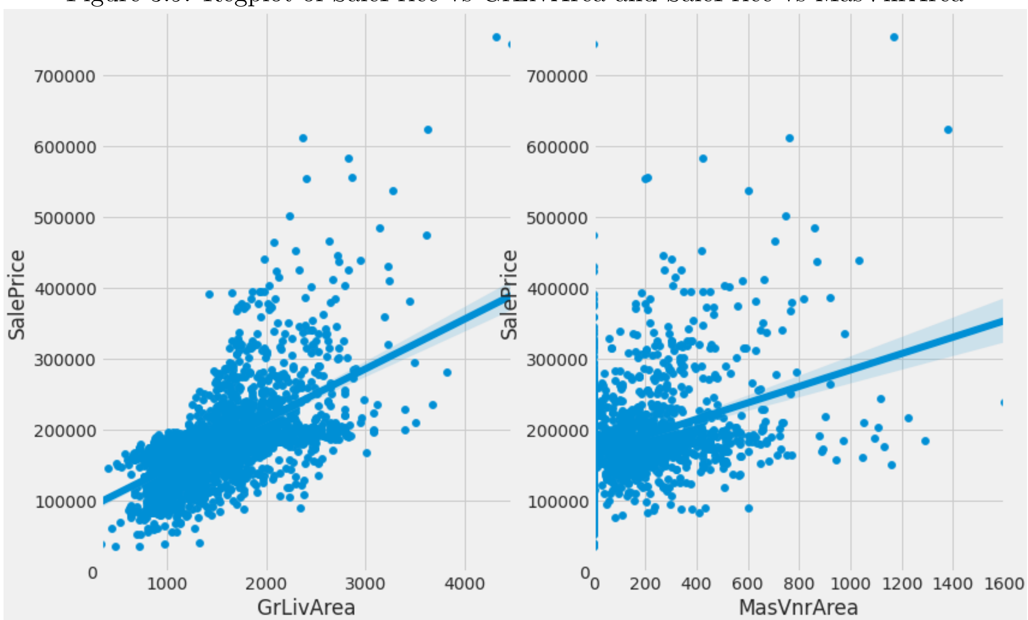


By analysing the scatter plots above, following observations were made:

- OverallQual is a categorical variable, and a scatter plot is not the best way to visualize categorical variables. However, there is an apparent relationship between the two features. The price of the houses increases with the overall quality.
- Our target variable shows an unequal level of variance across most predictor(independent) variables.
- There are many outliers in the scatter plots above.
- The two on the top-right edge of SalePrice vs. GrLivArea seems to follow a trend, which shows as the prices increased, so did the area.
- However, the two observations on the bottom right of the same chart do not follow any trends. These two values are to be removed.

To check the linearity of the variables, regplot was used from seaborn library.

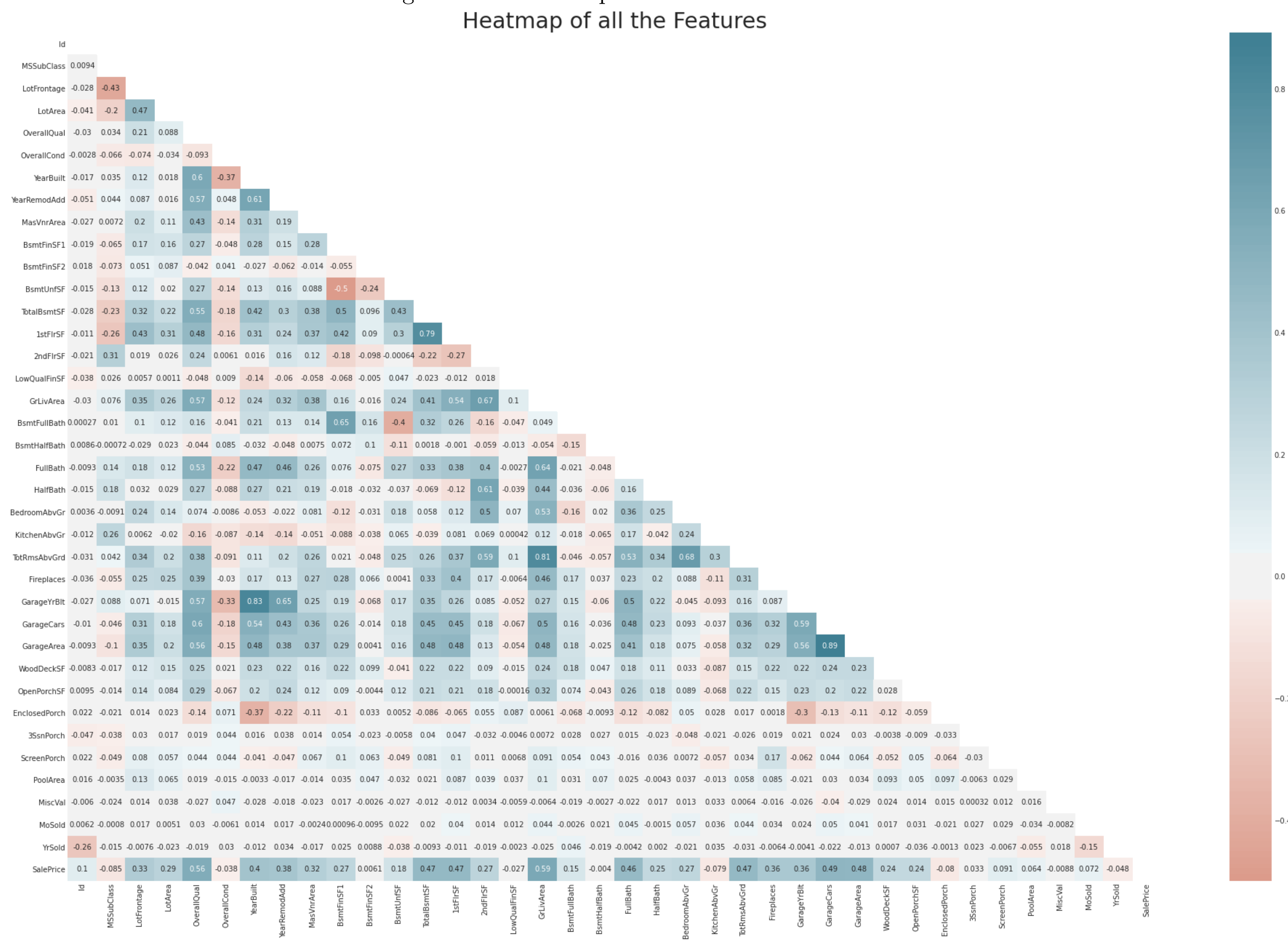
Figure 3.9: Regplot of SalePrice vs GrLivArea and SalePrice vs MasVnrArea



From the chart that there is a better linear relationship between SalePrice and GrLivArea than SalePrice and MasVnrArea. There are some outliers in the dataset.

To see the relationship between independent variables, heatmap is plotted using correlation coefficients.

Figure 3.10: Heatmap of all features



There is multicollinearity between various features which can be reduced using regularization regression models.

3.2 Feature Engineering

As seen in the explanatory data analysis section the target variable SalePrice is not normally distributed, right-skewed and has many outliers. For effective linear regression analysis this should be removed. The log transformation removes the normality of errors, which solves most of the other errors seen. Histogram, Boxplot and Q-Q plot of Saleprice below shows the errors have been removed. The error plot/Residual plot of SalePrice vs GrLivArea is plotted before and after log transformation. It is seen that variance dispersion with increasing GrLivArea is also decreased as in the figure.

Figure 3.11: Histogram, Boxplot and Q-Q plot of SalePrice after transformation

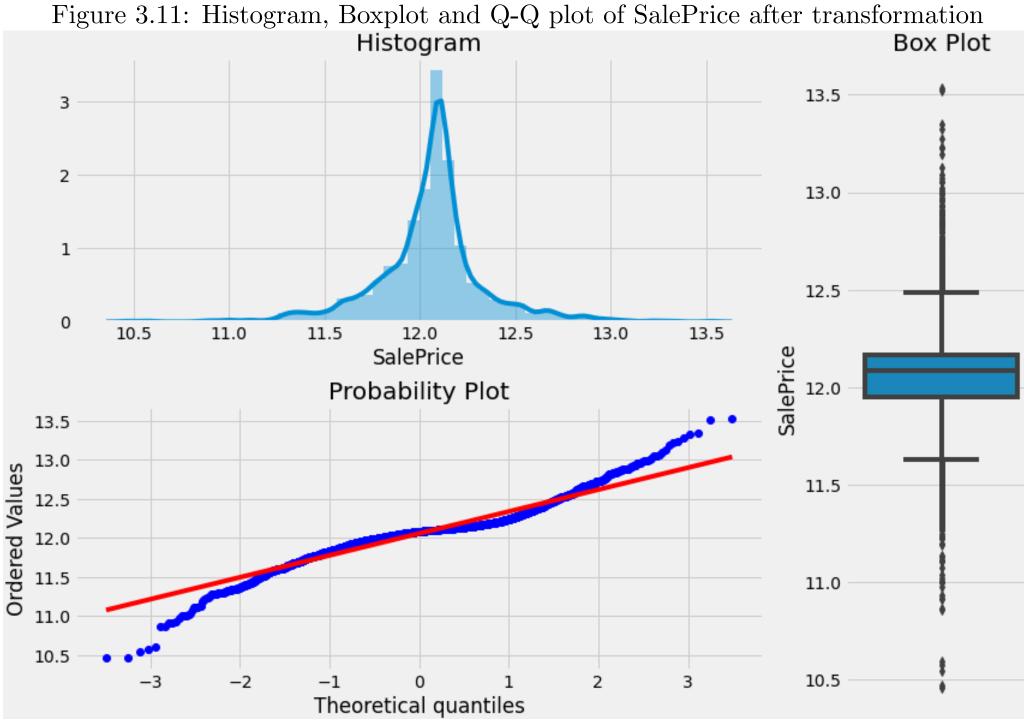
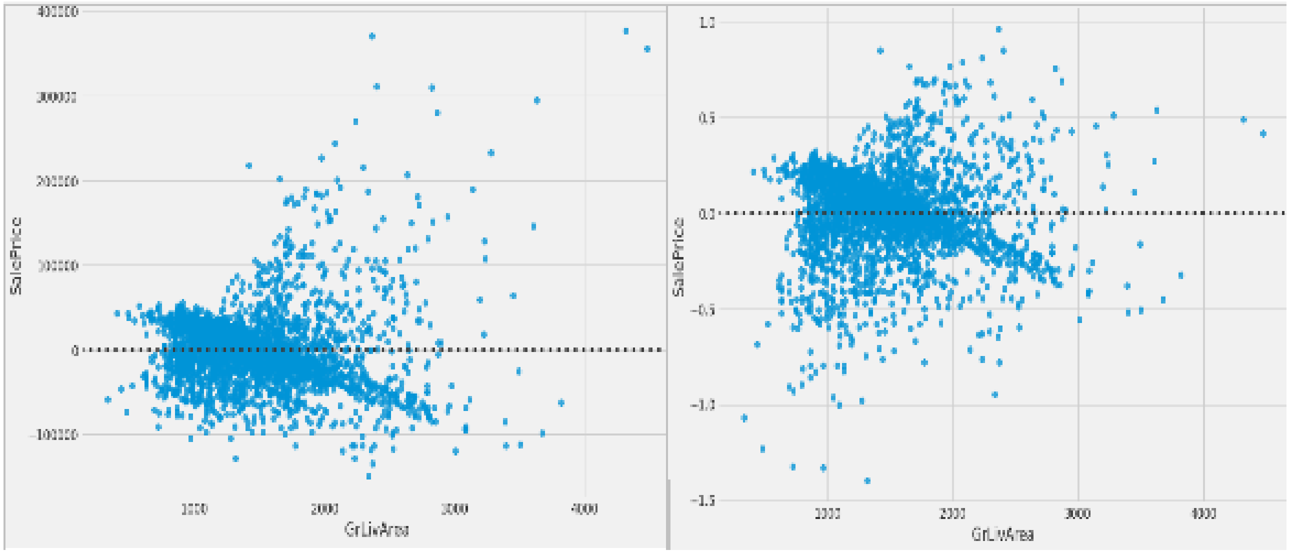
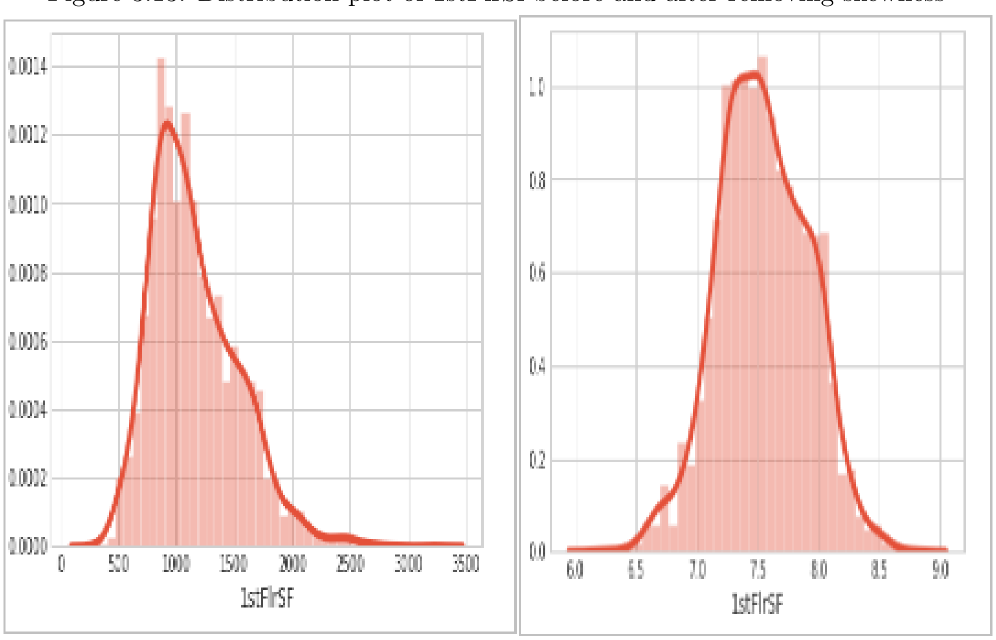


Figure 3.12: Error plot of SalePrice vs GrLivAr



On checking the skewness of features, many numerical variables MiscVal, PoolArea , LotArea, LowQualFinSF, 3SsnPorch, KitchenAbvGr, BsmtFinSF2, EnclosedPorch, ScreenPorch, BsmtHalfBath, OpenPorchSF, WoodDeckSF, GrLivArea, BsmtUnfSF, etc are seen to be skewed. Skewness is fixed of these features using box cox transformation. Distribution plot of 1stFlrSf is plotted to see the changes.

Figure 3.13: Distribution plot of 1stFlrSf before and after removing skewness



The Ames Housing data set and venue data are merged into a single one. All categorical variables are converted into dummy variables. Overfitted features are removed. Target variable is extracted as 'y' and features as X.

3.3 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,but 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.

Regularization methods work by penalizing the magnitude of the coefficients of features and at the same time minimizing the error between the predicted value and actual observed values. This minimization becomes a balance between the error and the size of the coefficients.

Ordinary least squared loss function minimizes the residual sum of the square to fit the data. Ridge regression adds penalty equivalent to the square of the magnitude of the coefficients. Lasso adds penalty equivalent to the absolute value of the sum of coefficients. This penalty is added to the least square loss function and replaces the squared sum of coefficients from Ridge. Elastic Net is the combination of both Ridge and Lasso. It adds both the sum of squared coefficients and the absolute sum of the coefficients with the ordinary least square function.

Support Vector Regression (SVR) uses the Support Vector Machine algorithm to predict a continuous variable. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. LightGBM is a gradient boosting framework that uses tree based learning algorithms. StackingCVRegressor extends the standard Stacking algorithm using Stacking prediction, and the predicted result is used as the input data of the 2-level classifier.

Chapter 4

RESULTS

Using Linear Regression we got very low R2 score 0.42

Figure 4.1: Results of linear regression

Using Multiple Linear Regression

```
[88] lr = LinearRegression(normalize=True, n_jobs=-1)
     lr.fit(X_train, y_train)
     y_pred = lr.predict(X_test)
```

```
[89] r2 = r2_score(y_test, y_pred) # r2 score
     mse = mean_squared_error(y_test, y_pred) # mse
     print("R2 score using Linear Regression:", r2)
     print("MSE using linear Regression:", mse)
```

```
R2 score using Linear Regression: 0.4242081819456648
MSE using linear Regression: 0.05467762478962617
```

Therefore advanced regularistion models such as Ridge, Lasso and Elasticnet was used. But R2 score didn't increase much. So 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.

Figure 4.2: Results of final model

```
[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
```

The model has good accuracy and a low error. 79 percentage R2 score means the model is able to explain 79 percentage of the response data around its mean. Hence the model for predicting housing sales price for Ames, Iowa considering all the important features including the neighbourhood venues.

Chapter 5

DISCUSSION

There is almost 37 percentage improvement in model by using advanced regression algorithms. But there is still variance that the model could not explain. There are 312 features but the data set contained only 2213 samples. More dataset can help training the model better. Some features are to be dropped to reduce the number of features in the data set. In the neighbourhood data some approximations are done to find the location details of nearby areas. 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 advanced methods need to be used for greater precision. More data, larger number of datasets, would help improve model performances significantly. .

Chapter 6

CONCLUSION

Following steps are followed in this 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 Recommendations for ways to improve

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.

Chapter 7

REFERENCE

[1] http://jse.amstat.org/v19n3/decock.pdf

[2] https://en.wikipedia.org/wiki/Ames,-Iowa

[3] https://stackabuse.com/multiple-linear-regression-with-python

[4] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.boxcox.html

[5] https://towardsdatascience.com/transforming-skewed-data-73da4c2d0d16

[6] https://data.library.virginia.edu/understanding-q-q-plots/