Pooja Deshpande

Assignment 1
MSDS – 410 Data Modeling for Supervised Learning, Summer 2020
Northwestern University

Contents:

1. Section I

Ames dataset consists of information about the properties sold in Ames, Iowa. Data has 82 columns, a mix of nominal, ordinal, discrete and continuous variables. Data describes the various aspects of the property and also gives information about the sale such as sale date and sale price.

The dataset consists of some important factors that determine the price of a property. It consists of 2930 observations. Some of the columns have NA (null) values which indicate the absence of a feature. For example: NA values in the Alley column indicate no alley access. Since a model is being built to predict the sale price of a typical home in Ames, Iowa, it is important to define a sample population which includes only the typical homes/sales in Ames.

SAMPLE POPULATION: The following table describes the drop conditions.

Col. Name	To Keep (Drop everything else)
SaleCondition	'Normal'
Zoning	'RL', 'RH', 'RM', 'FV'
HouseStyle	1Story', '2Story', '1.5Fin', 'SLvl', 'Sfoyer'
SubClass	20, 30, 50, 60, 70, 80, 85, 90, 120, 160, 190

Table 1: Drop conditions

- For the sample population, only the 'normal' sales have been considered. 'Abnormal', 'Partial', 'Family' sales were dropped.
- Only Residential zoning classifications were considered in the sample population. Agriculture, Commercial and industrial units were dropped.

- Unfinished 1.5 story houses, and 2.5 story houses were dropped from the sample population
- Out of 16 unique values in the SubClass column, 5 were dropped.

Drop conditions were defined based on the popular values in the dataset.

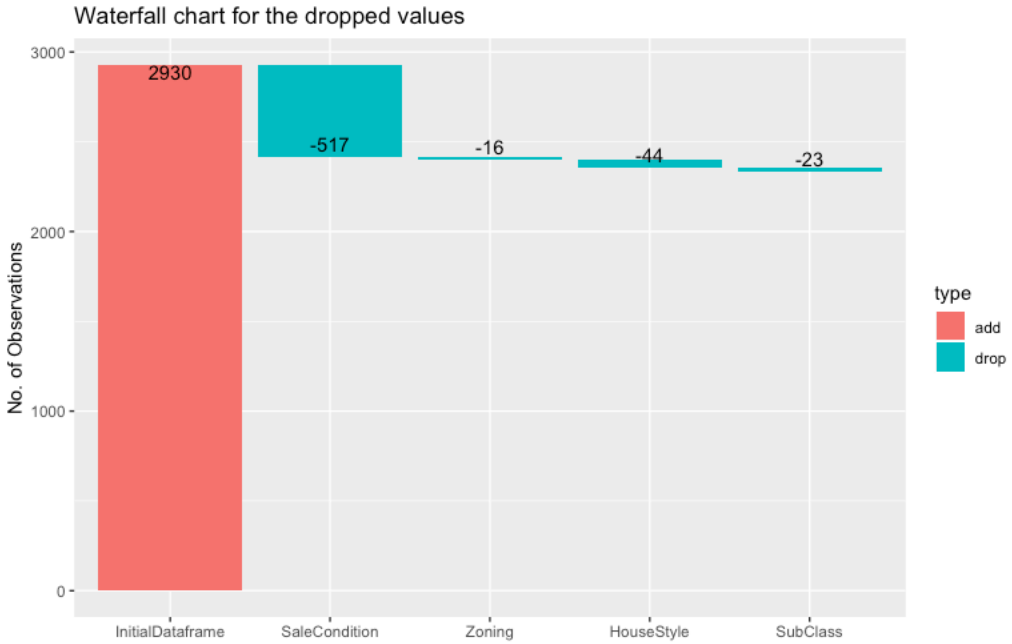

Fig 1: Waterfall chart

The final sample population consists of 82 columns and 2238 observations.

2. Section II

DATA QUALITY CHECK:

Data Quality check was performed on response variable, Sale Price and twenty predictors.

It was observed that there were no errors in the values of Sale Price, however, there were several outliers. Following are some univariate plots of Sale Price before and after the sample population was defined.

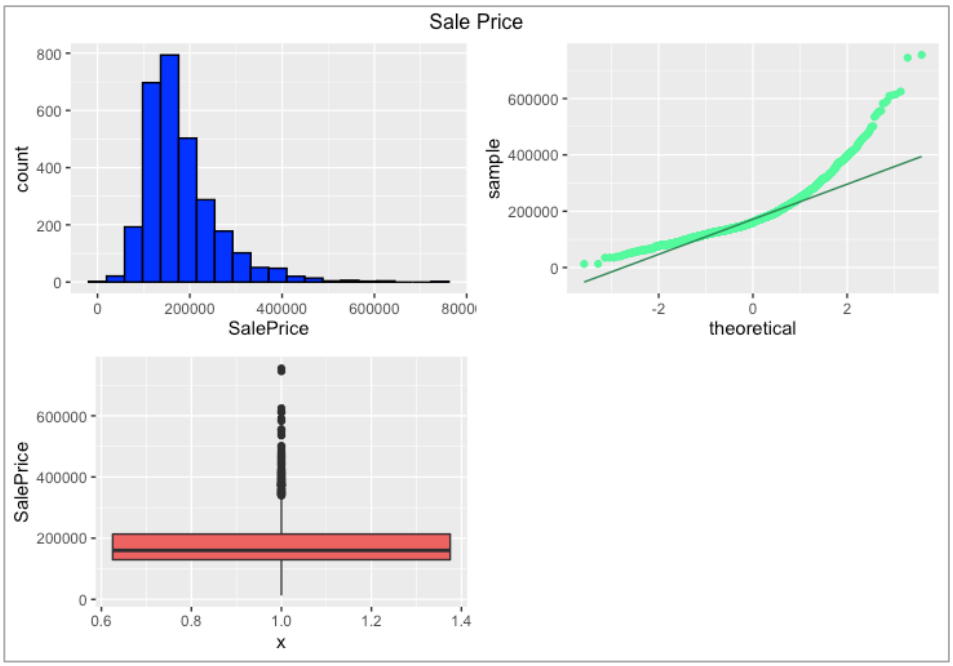

Fig 2: Sale Price before the sample extraction

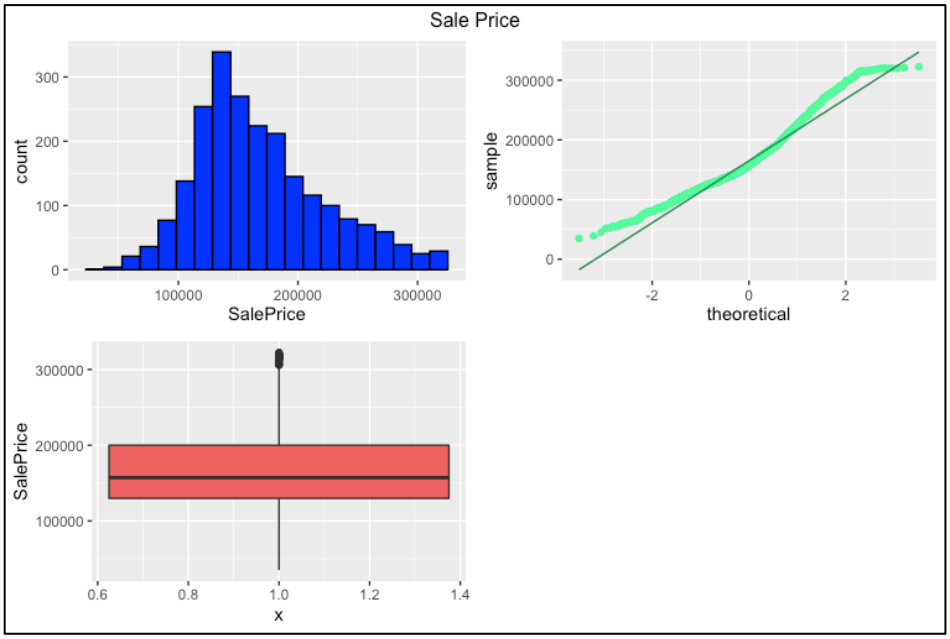

Fig 3: Sale Price after the sample extraction

As it is evident from the graphs, Sale price was positively skewed and had many outliers before the sample population was extracted. Values of the Sale price in the sampled dataset are more or less normally distributed with fewer outliers.

The following calculation also reveals that there are no extreme outliers in the final sample.

```
> summary(final_df$SalePrice)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  35000  129900  157250  168088  200000  322500
> q <- quantile(final_df$SalePrice,probs = c(0.25,0.75),names = FALSE)
> iqr <- IQR(final_df$SalePrice)
> extreme_outlier_bound <- q[2]+3.0*iqr
> extreme_outlier_bound
[1] 410300
> num_extreme_outliers <- dim(final_df[final_df[,"SalePrice"]>upper_bound,])[1]
> num_extreme_outliers
[1] 0
```

The following twenty predictors were selected, and data quality checks were performed on them.

	Column Name	Description	Data Quality
1	SubClass	Type of dwelling	Good, no erroneous values
2	Zoning	General zoning classification	No errors, Most of them are residential low density units
3	LotArea	Lot size in sqft	No errors, Highly positively skewed. Highly leptokurkic
4	Utilities	Types of utilities available	No errors, most homes have all public utilities
5	Neighborhood	Physical location within Ames city limits	No errors
6	Condition1	Proximity to various condition	No errors, most of them are in normal condition
7	BldgType	Type of dwelling	No errors, most are single family detached homes
8	HouseStyle	Style of dwelling	No errors, most are either one story or two story buildings
9	OverallQual	Overall quality and finish of the house	No errors, described on a scale of 1-10. Mean around 6
10	YearBuilt	Year when the home was built	No errors, values range from 1872 to 2010
11	ExterQual	Quality of the material in the exterior	No errors, most values describe 'Average' quality
12	ExterCond	Condition of the material in the exterior	No errors, most are in 'Average' condition
13	TotalBsmtSF	Total sqft of the basement area	No errors, Normal distributed
14	HeatingQC	Heating quality and condition	No errors
15	CentralAir	Central air conditioning	No errors, most homes have central AC
16	FirstFlrSF	First floor sqft	No errors, positive skewed
17	SecondFlrSF	Second floor sqft	No errors, positive skewed
18	TotRmsAbvGrd	Total rooms above grade(no bathrooms included)	No errors
19	KitchenQual	Quality of the kitchen	No errors, most in average or good condition
20	OverallCond	Overall condition of the house	No errors, mean around 5.6

Table 2: Results from data quality check

No errors/null values were found in any of the selected twenty variables. Some variables were found to be positively skewed. A more detailed summary and analysis of these variables can be found in the Appendix section.

3. Section III

INITIAL EDA

Continuous variables:

FirstFlrSF, SecondFlrSF, LotArea, TotalBsmtSF

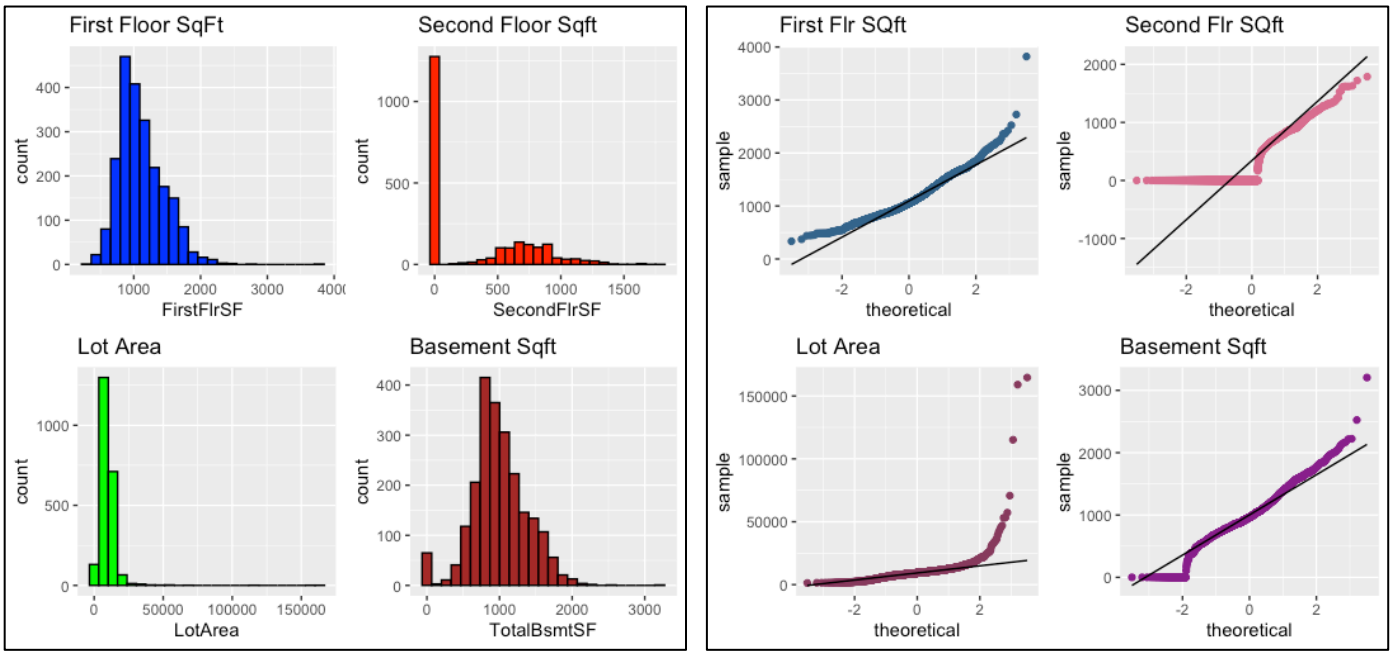

Fig 3,4: Histogram and qqplots of the continuous variables

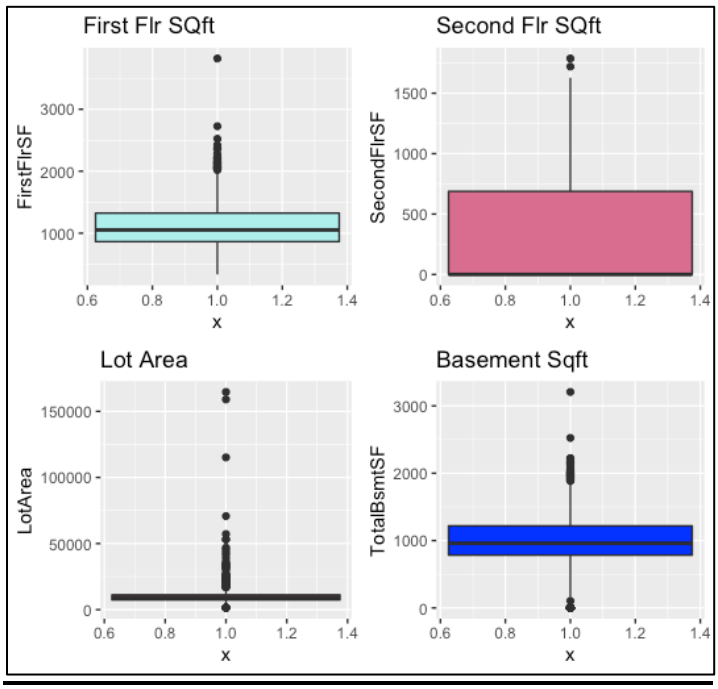

Fig 5: Boxplots of the continuous variables

- Based on the above plots, total basement square feet and first floor square feet and more or less normally distributed with some outliers. However, the total basement square feet is zero for many instances. This accounts for the homes that do not have a basement.
- Lot area looks more like an exponential distribution. It is highly skewed to the right and it is leptokurtic. The mean is greater than the median. Most of these homes have a basement, a garage and possibly large front/backyards. There are quite a few extreme outliers.
- SecondFlrSF has a large number of 0 values. That is because out of 2238 homes, only 678 are 2 storied buildings.

Following three new columns were created:

Age = Year built – Year sold

Total square feet = First floor + second floor + basement

Quality index = Overcall Condition * Overall quality

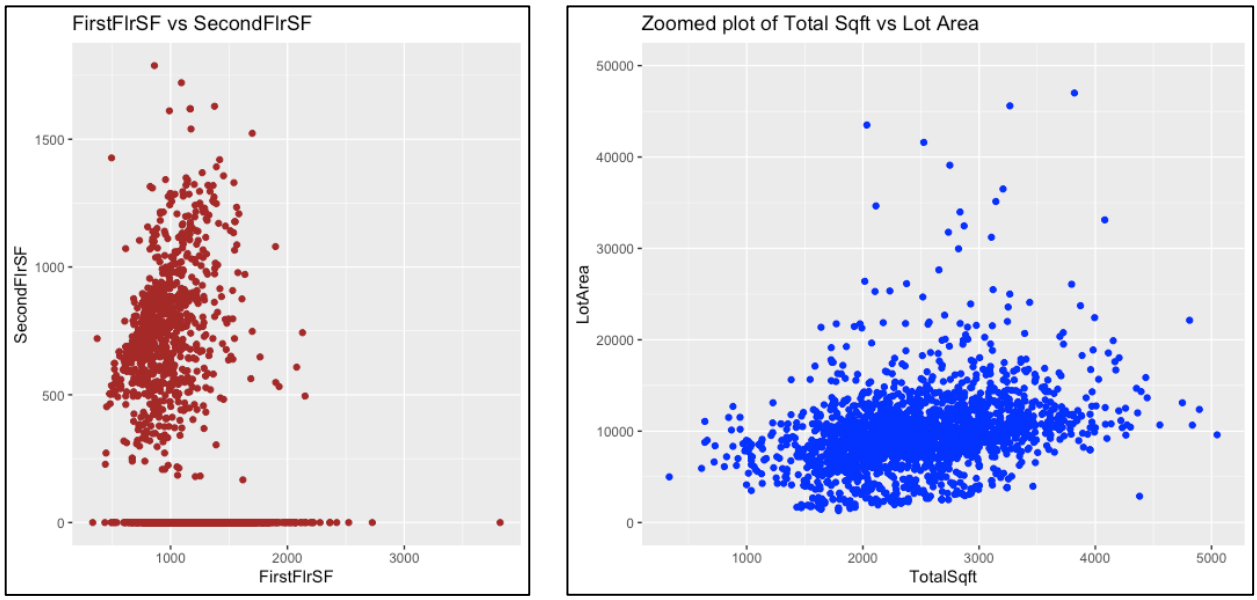

Fig 6, 7: Scatter plots of continuous variables

First Floor square feet is positively correlated with the second floor square feet. Total square feet is also correlated with the lot area.

<u>Discrete variables:</u>

Neighborhood, OverallCond, YearBuilt, SubClass, TotRmsAbvGrd, OverallQual

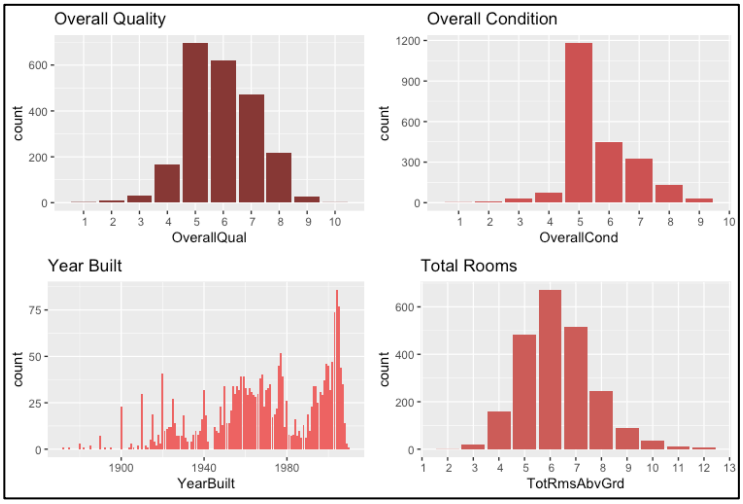

Fig 8: Bar plots of discrete variables

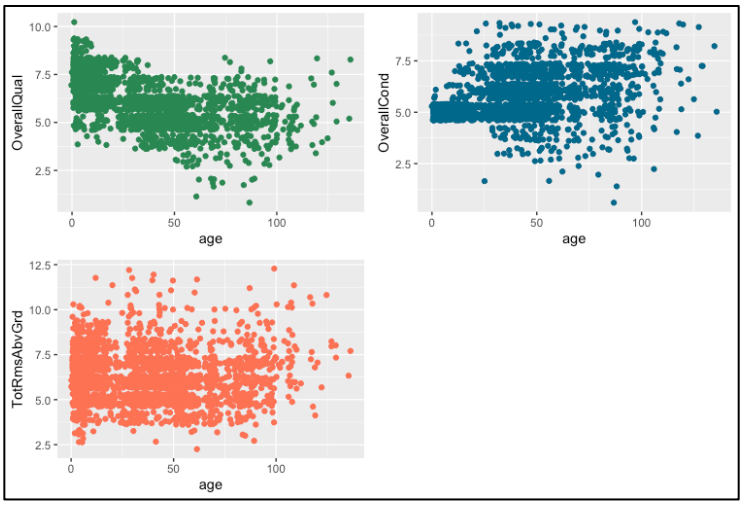

Fig 9: Jitter plots of discrete variables

- Most homes have average overall quality and condition. Mode is 5.
- Most of the homes listed in the dataset were built after 1980.
- Most homes have about 5 or 6 or 7 rooms above the basement.
- There is a decreasing trend with respect to age and the overall quality of homes, however, there is no trend in the plot with age and overall condition of the homes.

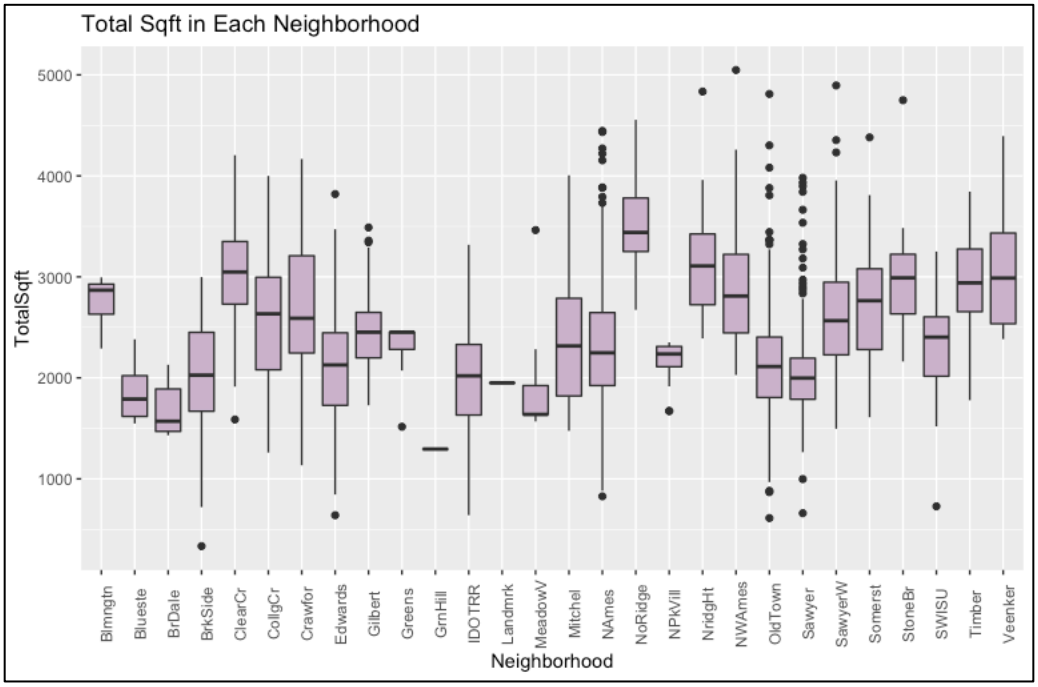

Fig 10: Total square ft per neighborhood

- North Ridge seems to have the largest homes. Home with the largest area is located in Northwest Ames.
- Green Hills and Landmark have only one home listed in the dataset.

4. Section IV

EDA FOR MODELING:

Four Variables:

TotalSqft, Neighborhood, Age, Qual_Index

These variables were selected the price of a home is largely dependent on these factors.

Following plots describe the relationship between the chosen predictors and the Sale price/Log of Sale price.

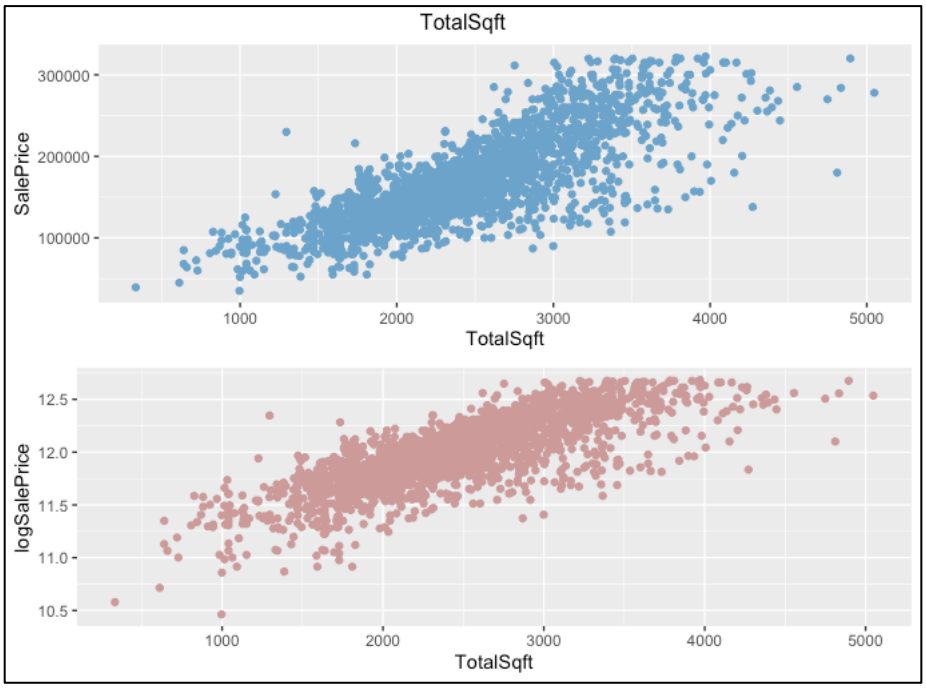

Fig 11: Total Sqft and Sale Price

- Total square feet and the sale price are positively correlated. As the square feet increases, the sale price increases. However, the variance of the data is reduced after the log transformation.

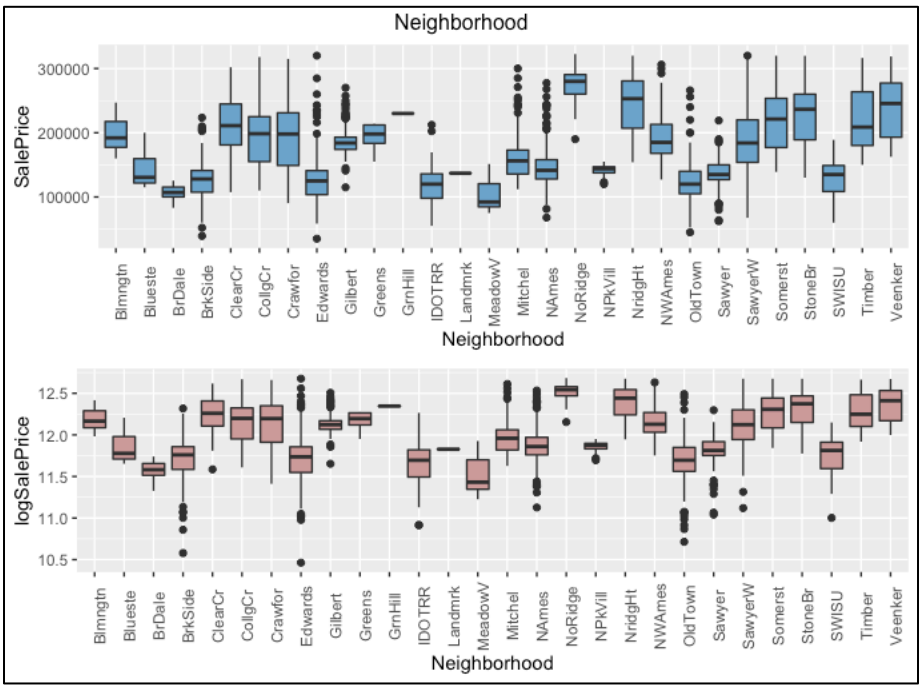

Fig 12: Neighborhood and sale price

- Some neighborhoods have higher sale prices than the others. Neighborhoods with lower home prices include Briar dale, Edwards, Meadow village. Neighborhoods with higher home prices are North Ride Heights, North Ridge and Veenker. The range of the log transformed sale price is lower.

Fig 13: Age and Sale price

- As expected, there is a negative correlation between sale price and age. However, there are certain outliers.

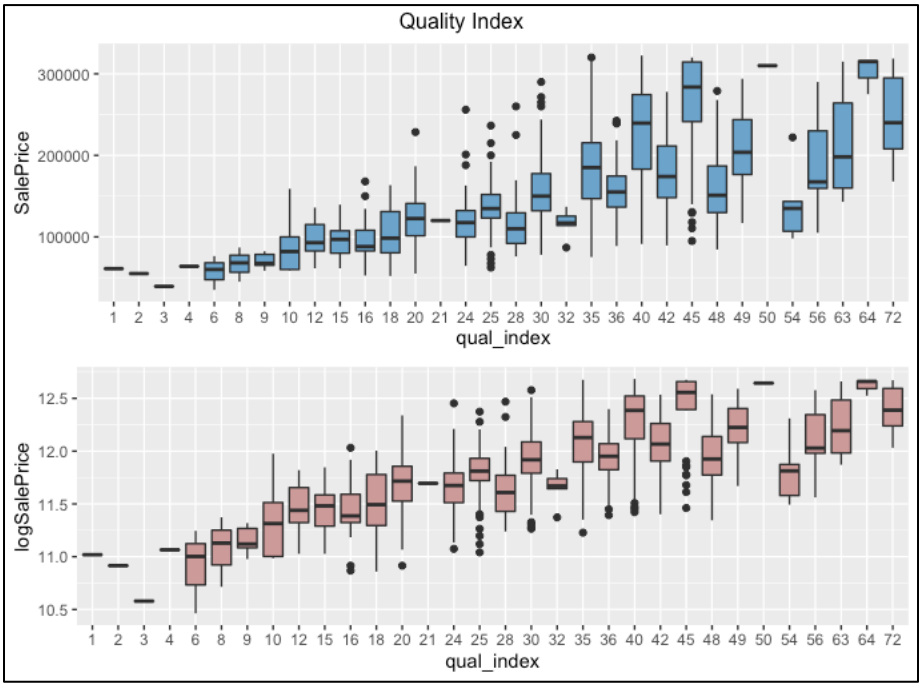

Fig 14: Quality index and sale price

- This was a surprising plot since one would expect the prices to increase as the quality index of a home increase. However, there are certain dips in the plot indicating that the positive correlation isn't consistent.

5. Section V

SUMMARY:

Ames dataset consisting of information regarding the homes sold in Ames, Iowa was chosen for this project. Sample population consisting of 'typical' homes in Ames was extracted. Data quality checks and EDA were performed. It was observed that the log transformation applied

on the sale price reduced the overall variance of the data. Log transformation would also help in satisfying the normal distribution assumption of regression analysis. Reducing the range of the data also helps in increasing the speed of the learning algorithm. Sale price is positively correlated with the total square feet of the house, negatively correlated with the age (except a few outliers), however, there was inconsistency in the correlation of the sale price with the quality index of the home.

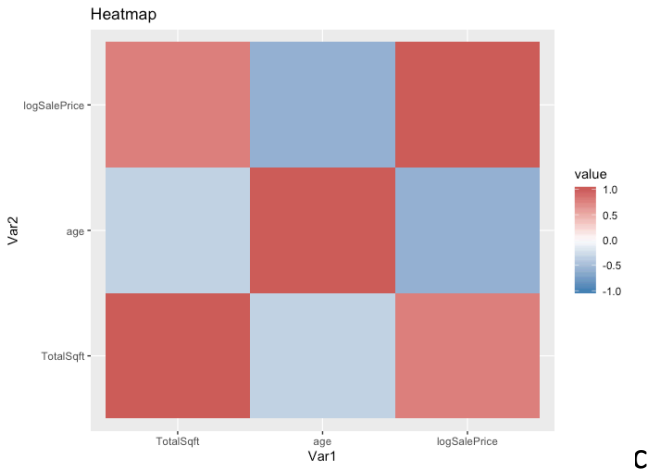

Fig 15: Correlation heatmap

The above heatmap indicates that there is very high positive correlation between Log Sale Price and total square feet. There is fairly good negative correlation between age and sale price. In addition to these two, neighborhood and quality index would be suitable to predict the sale price of a home in Ames, Iowa.

6. Appendix

Data quality analysis of twenty predictors:

```
SubClass
20     :824   60     :421   50     :241   120    :151   160    :115   30     :111   (Other):375

Zoning
A (agr):   0   C (all):   0   FV     :  90   I (all):   0   RH     :  19   RL     :1778   RM     : 351

LotArea
Min.   :  1300   1st Qu.:  7399   Median :  9306   Mean   :  9832   3rd Qu.: 11225   Max.   :164660
Standard Deviation : 6975.97
Skewness : 11.91899
Kurtosis : 232.313

Utilities
AllPub:2237   NoSeWa:   0   NoSewr:   1

Neighborhood
NAmes  : 389   CollgCr: 225   OldTown: 183   Edwards: 149   Sawyer : 135   Gilbert: 129   (Other):1028

Condition1
Norm   :1932   Feedr  : 122   Artery :  71   RRAn   :  34   PosN   :  30   RRAe   :  22   (Other):  27

BldgType
1Fam  :1854   2fmCon:  46   Duplex:  77   Twnhs :  84   TwnhsE: 177

HouseStyle
1Story :1133   2Story : 678   1.5Fin : 261   SLvl   : 109   SFoyer :  57   1.5Unf :   0   (Other):   0

OverallQual
Min.   : 1.000   1st Qu.: 5.000   Median : 6.000   Mean   : 5.924   3rd Qu.: 7.000   Max.   :10.000
Standard Deviation : 1.232465
Skewness : 0.03221511
Kurtosis : 3.006972
```

```
OverallCond
Min.   :1.000   1st Qu.:5.000   Median :5.000   Mean   :5.655   3rd Qu.:6.000   Max.   :9.000
Standard Deviation : 1.119123
Skewness : 0.6409474
Kurtosis : 3.881064

YearBuilt
Min.   :1872   1st Qu.:1954   Median :1971   Mean   :1970   3rd Qu.:1997   Max.   :2010
Standard Deviation : 28.60191
Skewness : -0.5762686
Kurtosis : 2.635137

ExterQual
Ex:  20   Fa:  23   Gd: 703   TA:1492

ExterCond
Ex:  10   Fa:  48   Gd: 250   Po:   2   TA:1928

TotalBsmtSF
Min.   :   0   1st Qu.: 784   Median : 960   Mean   :1001   3rd Qu.:1219   Max.   :3206
Standard Deviation : 377.5164
Skewness : 0.1115508
Kurtosis : 4.102482

HeatingQC
Ex:1071   Fa:  71   Gd: 400   Po:   0   TA: 696

CentralAir
N: 130   Y:2108
```

```
FirstFlrSF
Min.   : 334   1st Qu.: 864   Median :1052   Mean   :1113   3rd Qu.:1324   Max.   :3820
Standard Deviation : 338.2513
Skewness : 0.9086444
Kurtosis : 5.106887

SecondFlrSF
Min.   :   0.0   1st Qu.:   0.0   Median :   0.0   Mean   : 325.9   3rd Qu.: 688.0   Max.   :1788.0
Standard Deviation : 408.4801
Skewness : 0.7725305
Kurtosis : 2.287713

TotRmsAbvGrd
Min.   : 2.000   1st Qu.: 5.000   Median : 6.000   Mean   : 6.291   3rd Qu.: 7.000   Max.   :12.000
Standard Deviation : 1.424235
Skewness : 0.6057888
Kurtosis : 3.902624

KitchenQual
Ex:  63   Fa:  56   Gd: 883   Po:   1   TA:1235
```