**Assignment #1: Getting to Know Your Data**

**Introduction:** The data for this assignment is the Ames, Iowa housing data set with 2930 observations, **SalePrice** response variable and 80 (82- SalePrice- SID) predictors. This assignment requires to select 20/10 explanatory variables to predict the final price of each **typical** home using simple linear regression.

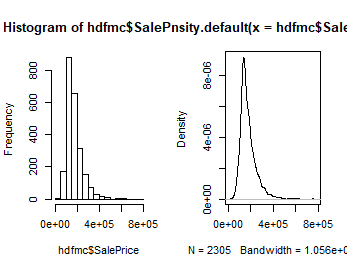
**Section 1: Sample Definition**

The problem statement is focused around definition of the typical home and would require drop all observation that considered to be **atypical.**

|  |  |  |  |
| --- | --- | --- | --- |
| Predictor | Drop | Counts | Waterfall |
| Zoning | Drop all non-residential areas (A,C,FL,I) | |  |  | | --- | --- | | **Row Labels** | **Count of SID** | | A (agr) | 2 | | C (all) | 25 | | FV | 139 | | I (all) | 2 | | **Grand Total** | **168** | | 2930 |
| Sale Condition (Nominal): Condition of sale | Drop all not normal | |  |  | | --- | --- | | **Row Labels** | **Count of SID** | | Abnorml | 174 | | AdjLand | 12 | | Alloca | 22 | | Family | 46 | | Partial | 203 | | **Grand Total** | **457** | | 2762 |
| SalePrice outliers and Abnormalities | boxplot.stats(hdfmc$SalePrice)  $`stats`  35000 129000 157000 203000 313000 | **<$35,000 – 0**  **>$313000 -118** | 2305 |
| Sample definition |  |  | 2187 |

**SalePrice analysis**

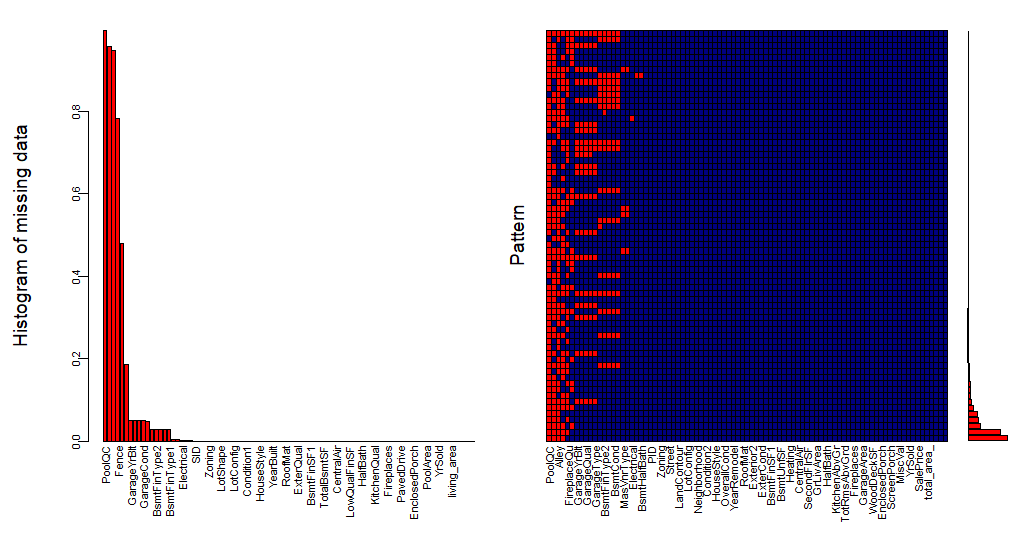
SalePrice Histogram and density shows that the distribution is shifted. Does this mean that it is not normal distribution?



**Section 2: Data Quality Check**

**Missing Value analysis**

Missing Value map



**Selected Fields**

Excluding SID, Zoning, Sale Condition (all “Normal” left), six fields with high parentage of missing data.

**selected\_columns**<-*c("PID","SubClass","LotArea","LandContour","Utilities","Neighborhood","BldgType","HouseStyle","OverallCond","YearBuilt","YearRemodel",“TotalBsmtSF", "Heating","CentralAir","Electrical","FirstFlrSF","SecondFlrSF","GrLivArea","YrSold","SalePrice")*

Splitting data into 3 groups **size, quality** and **calculated**

**propery\_size\_group** *<-c("LotArea","TotalBsmtSF","FirstFlrSF","SecondFlrSF","GrLivArea","SalePrice")*

**propery\_quality\_group***<-c("PID","SubClass","LandContour","Utilities","Neighborhood","BldgType","HouseStyle","OverallCond","YearBuilt","YearRemodel","Heating","CentralAir","Electrical","YrSold","SalePrice")*

**calucalted\_group***<-c(“living\_area,total\_area”,”ratio\_living\_to\_total”,”ratio\_lot\_to\_living”,”price\_per\_living\_SF”,”price\_per\_total\_SF”,log(SalePrice))*

**Compare with Spec**

Most of the variables fit the specification of the data dictionary with exception to Heating having some missing value

**Meaningless values**

Have not find any none-reasonable information

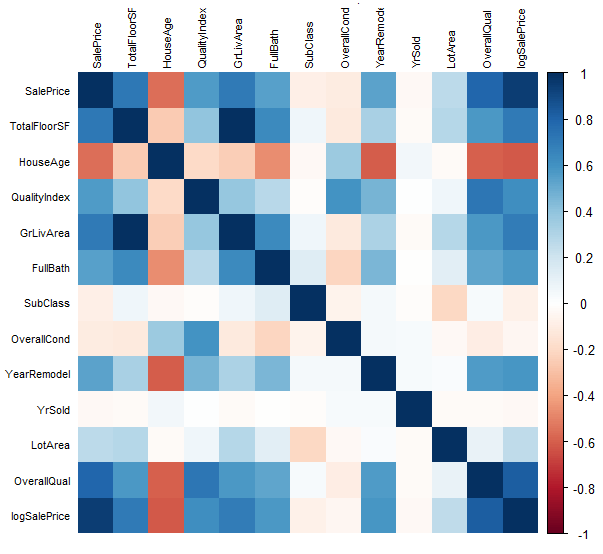
**Section 3: Initial Exploratory Data Analysis**

There are 2 categorical variables from the quality group which in my opinion would impact SalePrice

* Neighborhood
* Electrical

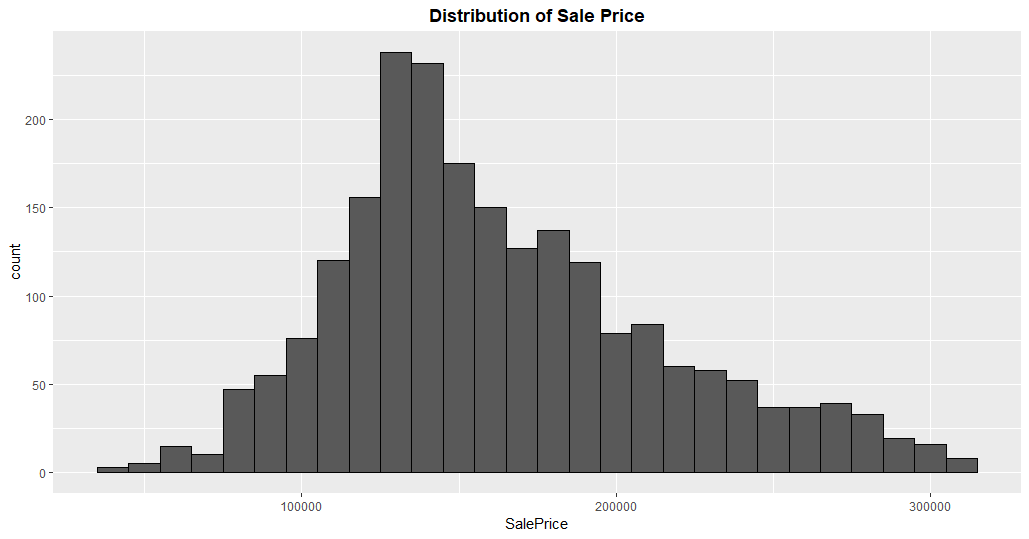
Based on the correlation heatmap the most effective quantitative predictors on the response SalePrice are

* TotalFloorSF – total living area
* QualityIndex
* GrLivingArea
* FullBath
* YearRemadeling
* LotArea
* OverolQuality
* Price\_per\_SF

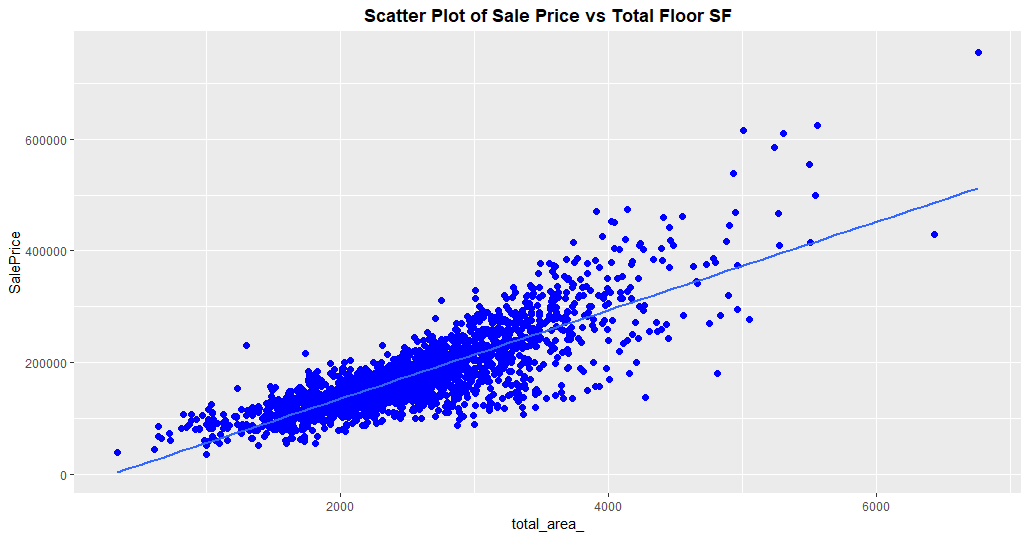


**Section 4: Exploratory Data Analysis for Modeling**

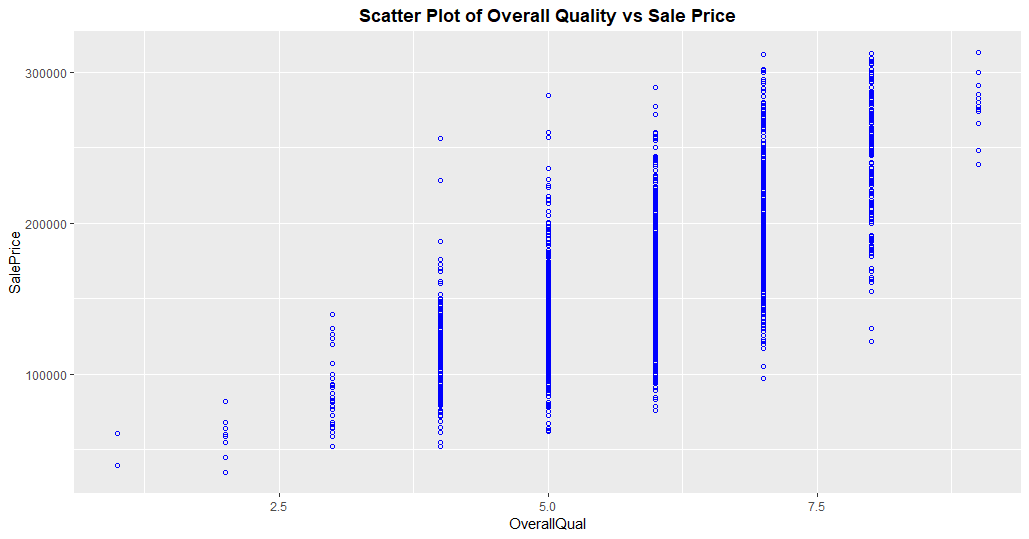
* Hypnotize: ***OverolQual , TotalFloorSF, Neighborhood*** most valued exploratory variables in order to predict ***SalePrice***
* **SalePrice**



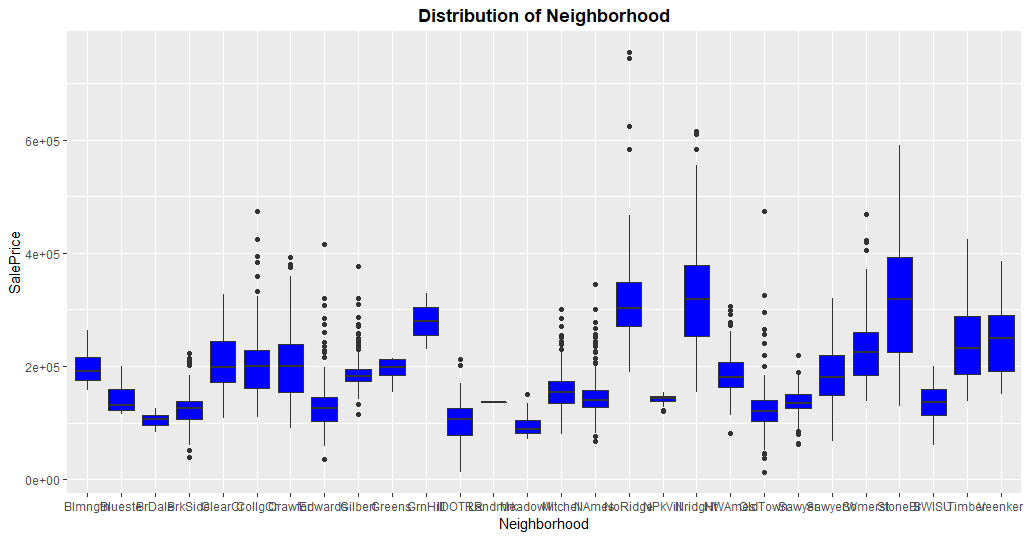
**TotalFloorSF**

****

**OverolQual**



**Neighboood**



**Section 5: Summary/Conclusions**

* SalePrice is shifted to the left and deviate from normal distribution. Removing outliers helped but there still a concern
* Total floor vs, Sale Price scatter plot shows response and fitted line variability growth with predictor suggesting possible transformation of the response variable [Everett]
* Total leaving area, Overall Quality condition and Neighborhood most valuable candidates for the linear regression model for SalePrice prediction
* PID is interesting and would require some additional research
* Would be interesting to understand same property growth year over year. That my help to understand an intercept [Everett]
* Intuition suggests there are some variable grouping and conversion especially for the nominal to ordinal
* Started to work on ratios such as leaving area to lot area , have not been able to finish

**Appendix**

Excluded Building Type, Heating, Central Air median and box size do not indicate impact

|  |  |  |
| --- | --- | --- |
|  |  |  |

Put your data quality check in an appendix in your assignment report.

(4) An Initial Exploratory Data Analysis

- **Pick ten variables from the twenty variables from your data quality check to explore in your initial exploratory data analysis.** Perform an initial exploratory data analysis. How do we perform an exploratory data analysis for continuous versus discrete (or categorical) data? Consider the use of scatterplots, scatterplot smoothers such as LOESS, and boxplots to produce relevant graphics when appropriate.

Now that we have a basic understanding, we can begin the modeling building process. Note that in the model building EDA we are particularly interested in the relationships between the response variable and the predictor variables.

Your initial exploratory data analysis should be at least one section in your assignment report. It might be wise to split your EDA into two sections in your report – one section for continuous variables and one section for discrete variables.

After we have performed the necessary prerequisite data work, we can then begin the modeling process. Every modeling process begins with an initial exploratory data analysis that is oriented for the problem at hand. Different statistical models require different types of exploratory analysis. In this assignment we will be developing an exploratory data analysis for a regression problem with a continuous response variable.

(5) An Initial Exploratory Data Analysis for Modeling

- What is the response variable in this problem? In addition to the raw response variable should we consider a transformation of the response variable? Consider SalePrice and log(SalePrice).

**- Pick three variables from the ten variables from your initial exploratory data analysis and explore their relationship with SalePrice and log(SalePrice).**

- Note that the correct EDA technique depends on the type of the predictor variable – discrete or continuous.

Items to discuss in this section of your report.

- Does your EDA suggest any potential difficulties or concerns for the model building process?

- Does your EDA suggest that there may be a need to consider transformations in the predictor variables at some point during the model building process?

These results should be in a separate section from your initial exploratory data analysis results.

(6) Summary/Conclusions:

Please summarize your findings in at least 1 paragraph in a conclusion / reflection section. You should, at a minimum, address the following questions:

* Does your EDA suggest any potential difficulties or concerns for the model building process?
* Does your EDA suggest that there may be a need to consider transformations in the predictor variables at some point during the model building process?