Assignment #1
Prabhat Thakur

Introduction:

Building effective statistical model always requires good understanding of the experimental data in hand. The aim of this assignment is to review the Ames Housing Data using Exploratory Data Analysis (EDA) techniques and learn about the different independent variables, also called predictor variables before moving to model building and regression analysis.

This report discusses different steps; (1) data survey, (2) data quality check, and (3) initial exploratory data analysis tasks performed to analyze housing dataset and associated findings. R programming language has been used for this exercise.

Data File: ames_housing_data.csv (Ames, Iowa housing data set posted in Canvas).

Results:

Section 1: Data Survey

The dataset contains information from the Ames Assessor's Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010. The data contains 2930 obs. of 82 variables of quantitative and qualitative types which includes 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables (and 2 additional observation identifiers). The supporting data dictionary document AmesHousingDataDocumentation.txt provides definition of each variable, data types and description of categorical variable values.

Form the first look, dataset appears to contain lot of useful attributes for houses and doesn't appear to have missing values. 'NA' for categorical values and 0 for numeric values has been used for missing or not application housing attribute.

Our end goal is to build linear regression models to predict sales price of a home. The dataset contains SalePrice (Continuous) variable for all 2930 observations and will act as our response variable. There are no missing values in SalePrice variable for all observations however a detailed investigation of the SalePrice and other variables should still be performed to make sure the data is not skewed, if there are outliers then we should find proper justification for such values so that informed decision can be taken to exclude those observations from the data population. Property types, Age and Sale Condition (considering the financial recession during 2008) can contribute to outliers in SalePrice. Choosing right predictor variables which contribute significantly to the house price is very important for building efficient predictive regression mode.

Section 2: Sample Definition

For defining our sample population we will begin with identifying housing attributes which can be excluded from the population to make it homogeneous and keep similar type properties or sales conditions. We will make some assumptions to identify atypical housing property from Ames housing market to extract the population of interest (typical homes).

We will analyze sales condition variable and exclude observations other than Normal sales conditions. Out of 2930 observations 2413 represents Normal sales conditions which is about 82% of total population. It seems to be a good option to only include Normal sales conditions in our sample population, abnormal sales like trade, foreclosure, short sale and Sale between family members may not be as correlated with other variables as Normal sales.

```
> summary (mydata$SaleCondition)
Abnorml AdjLand  Alloca  Family  Normal Partial 
    190      12      24      46    2413     245 
```

Another variable to analyze is Bldg Type (Nominal): Type of dwelling. We will restrict our analysis to Single family homes only which are about 83% of total population.

```
> summary (mydata$BldgType)
  1Fam 2fmCon Duplex  Twnhs TwnhsE 
  2425     62    109    101    233 
```

MS Zoning (Nominal): Identifies the general zoning classification of the sale, seems to be another variable which we can look into. There are 4 other classification other than Residential like Agriculture, Commercial, Floating Village Residential, and Industrial. For this assignment I will focus only to property sales in Residential zoning.

```
> summary (mydata$Zoning)
A (agr) C (all)      FV I (all)      RH      RL      RM 
      2      25     139       2      27    2273     462 
```

Drop Conditions:

Based on above discussed points, I would choose to restrict sample dataset to single family, residential zoned houses with normal sales condition only. We would not want to include commercial property types or multi-family type houses in the same sample as single-family homes. Removing too many attributes randomly may result in over fitting the model. We will want our sample population to be a good representative of Ames housing market.

These Categories were Dropped:

- All observations with **SaleCondition** not equal to “Normal”
- All observations with **BldgType** not equal to “1Fam”
- All observations with **Zoning** not equal to “RH”, “RL”, “RP”, “RM”

Remaining observations for the sample population.
Number of observations: 1943, approximately 66% of the original observations remain.

Note: Four new variables **TotalFloorSF** (total floor square feet), **HouseAge** (Age of House), **price_sqft** (price / sqft) and **QualityIndex** were derived which could be useful predictor.

```
mydata$TotalFloorSF <- mydata$FirstFlrSF + mydata$SecondFlrSF
mydata$HouseAge <- mydata$YrSold - mydata$YearBuilt
mydata$QualityIndex <- mydata$OverallQual * mydata$OverallCond
mydata$price_sqft <- mydata$SalePrice/mydata$TotalFloorSF
```

```
> length(mydata_sample)
[1] 86
> nrow(mydata_sample)
[1] 1943
> nrow(mydata_sample) / nrow(mydata)
[1] 0.6631399
```

Section 3: Data Quality Check

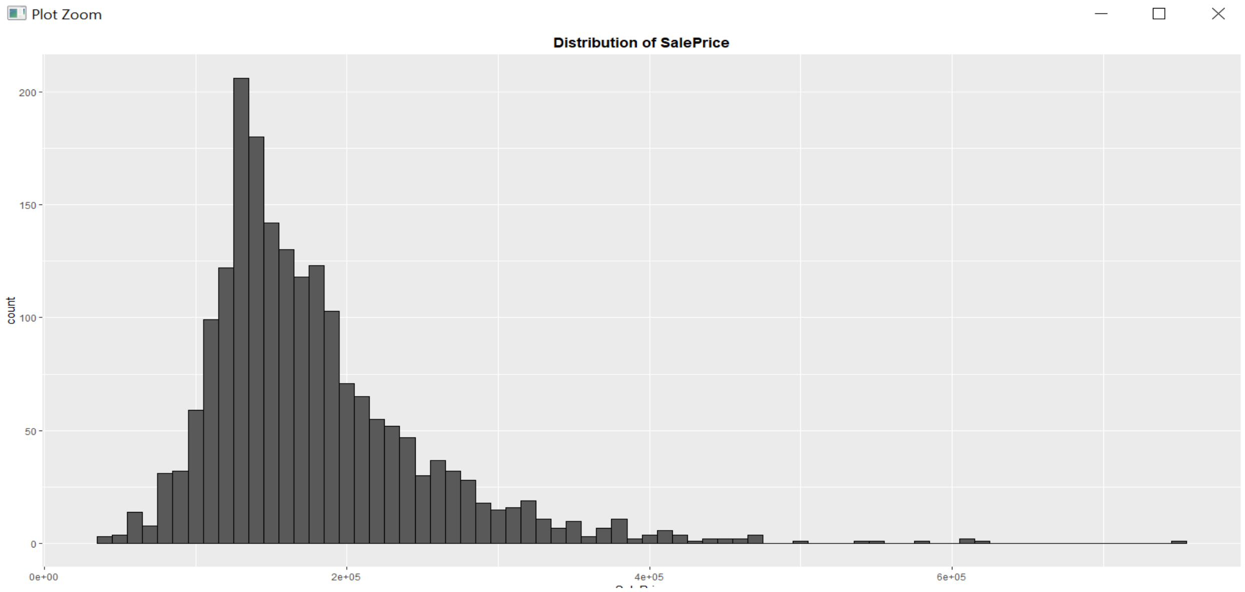
Data quality check is performed on sample data of 1943 observations we created in section 2. In order to check the quality of the data, a subset of 20 variables have been selected including SalePrice as the response variable as well as TotalFloorSF, HouseAge, price_sqft calculated variables from our sample population. Since sample already represents single family, residential zoned houses with normal sales condition, I have not included SaleCondition, BldgType and Zoning variables in selected 20 variables. See appendix A for list of variable.

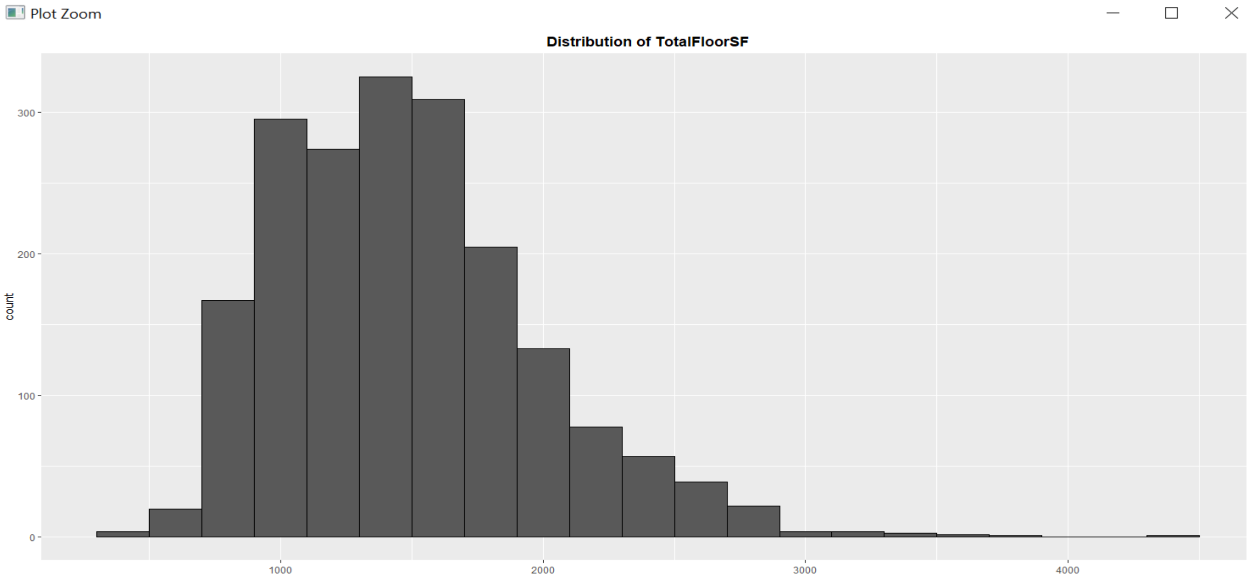
The selected variables include numeric (continuous and discrete) as well as categorical (nominal or ordinal) types and are listed below.

Continuous: "SalePrice", "TotalFloorSF", "price_sqft", "LotArea", "BsmtFinSF1", "TotalBsmtSF", "GarageArea", "PoolArea"
Discrete: "FullBath", "YrSold","HouseAge", "OverallCond", "OverallQual", "YearRemodel"
Categorical: "LotShape", "BsmtCond", "Neighborhood", "HouseStyle", "KitchenQual", "GarageQual

From analyzing SalePrice, TotalFloorSF and price_sqft variables, Sale price seems to have some outliers which could be due to large property in the sample dataset. There are couple outliers in TotalFloorSF and LotArea as well. A closer look is required on these outliers but for this given dataset it seems to be safe to remove them from analysis.

Here is the distribution of SalePrice and Total Floor SF for the sample dataset.





Section 4: Initial Exploratory Data Analysis

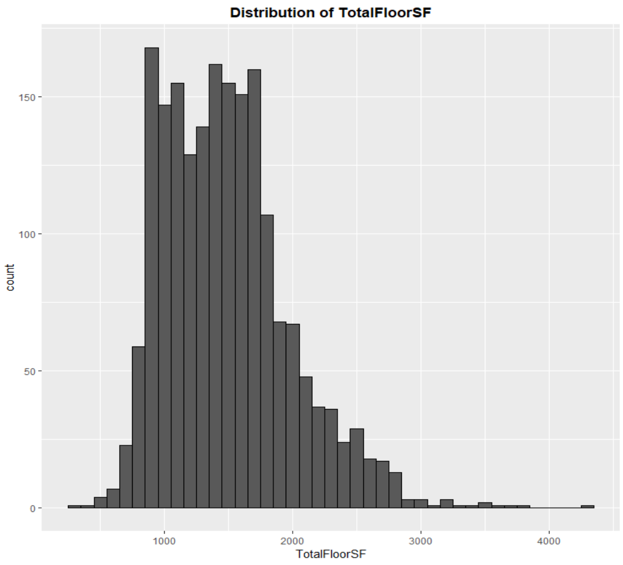
After we have performed the necessary prerequisite data work, we can then begin the modeling process. Every modeling process begins with an initial exploratory data analysis that is oriented for the problem at hand. Different statistical models require different types of exploratory analysis. In this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

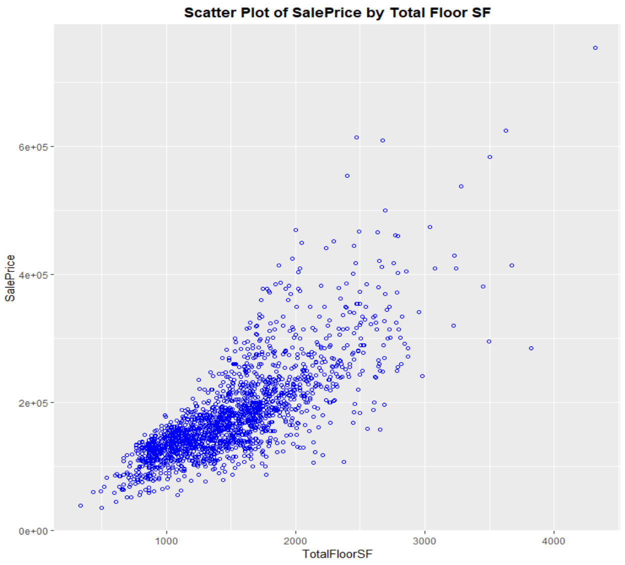
In this section, we will select 10 variables and perform initial exploratory data analysis to examine relationships between the response variable (SalePrice) and the predictor variables.

Following are the 10 variables I have selected for initial EDA : "SalePrice","TotalFloorSF","price_sqft","LotArea","LotShape","TotalBsmtSF", "Neighborhood", "KitchenQual", "OverallQual", "GarageArea"

Below are some selected plots and graphs to explain the relationship.

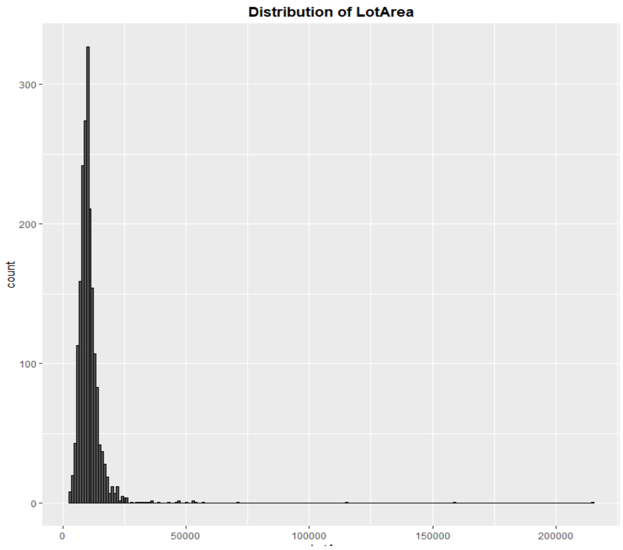
1. "SalePrice" vs "TotalFloorSF"

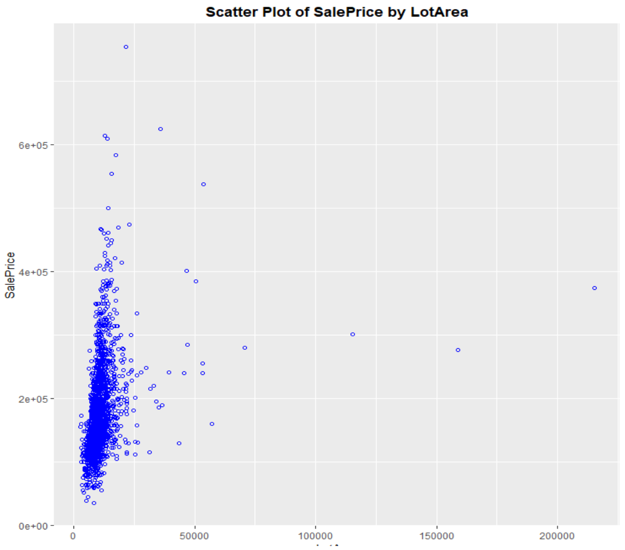




We can see a general positive relationship between sales price and total floor SF area. However, the plot indicates couple outliers towards the lower side of right end.

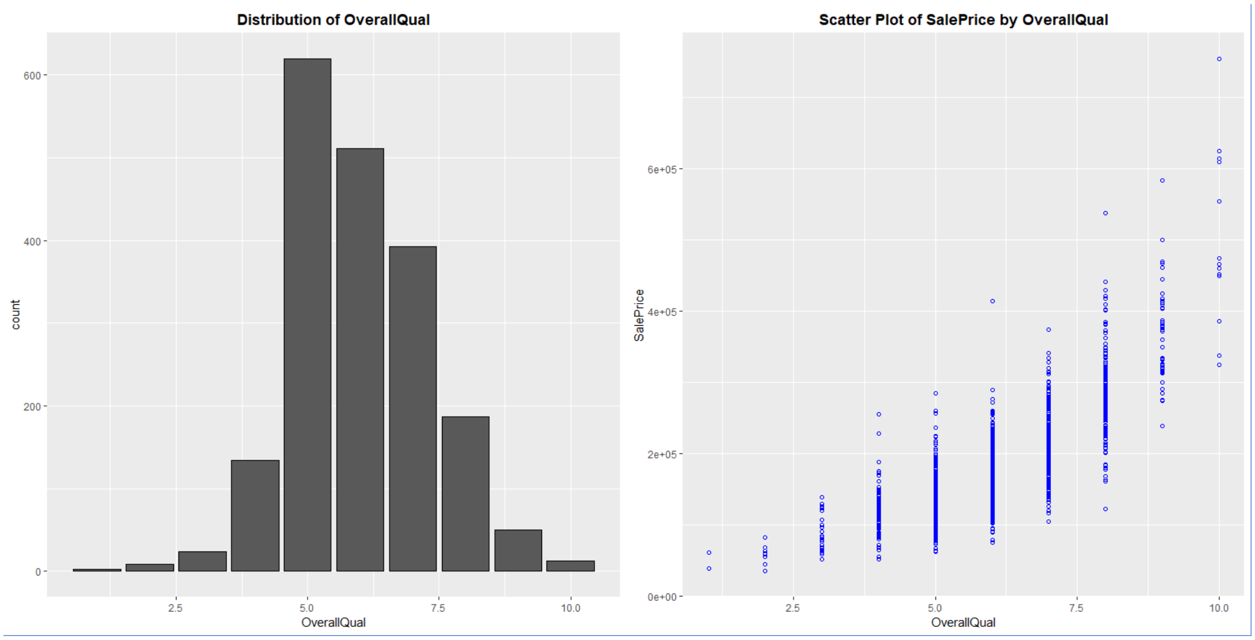
2. "SalePrice" vs " LotArea "





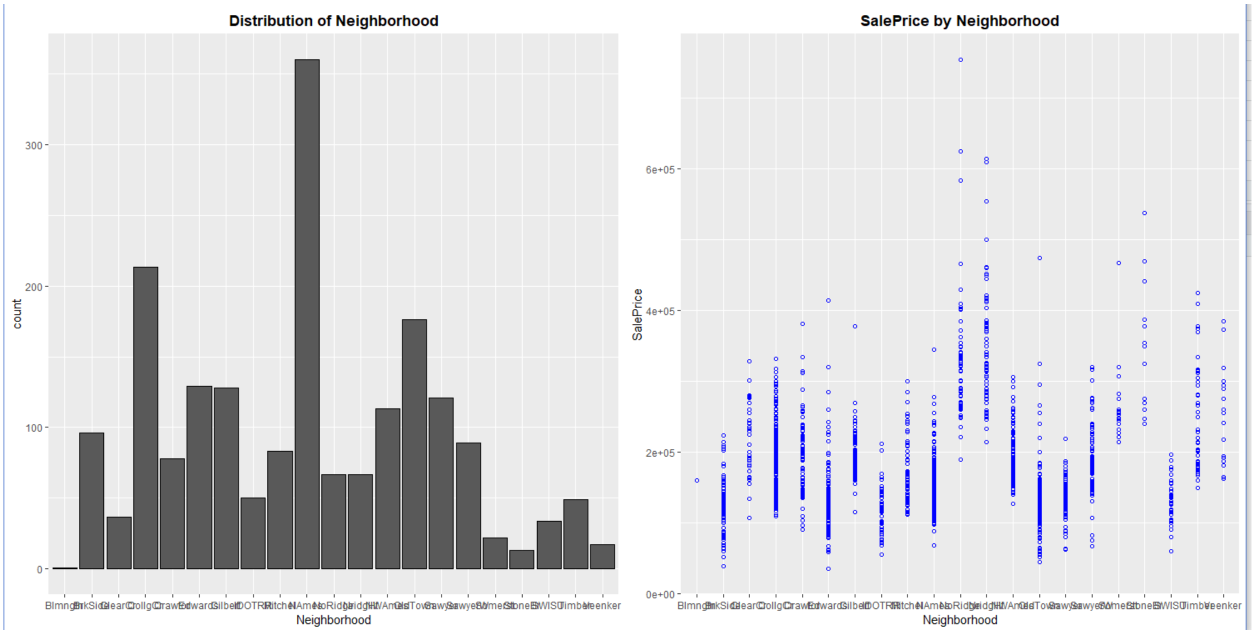
From the second scatter plot below, we can again see a general positive relationship between sales price and LotArea. However, the plot indicates few outliers.

3. "SalePrice" vs " OverallQual "



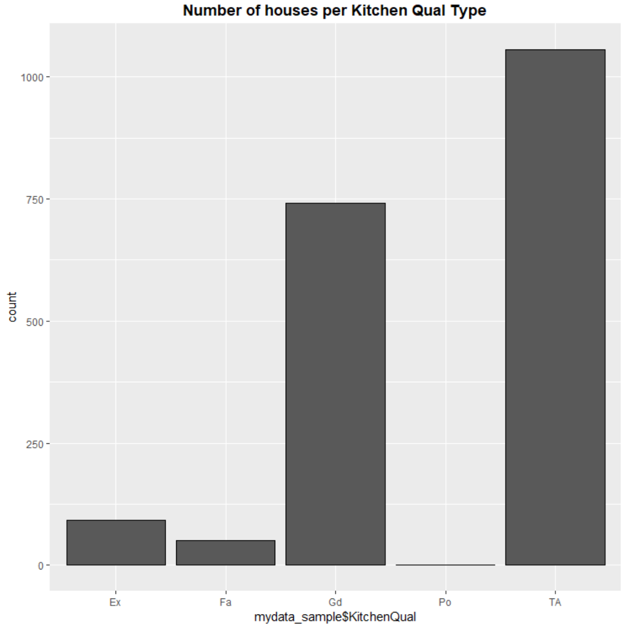
Most of the houses are in good condition and we can see positive relationship between sales price and overall quality ratings.

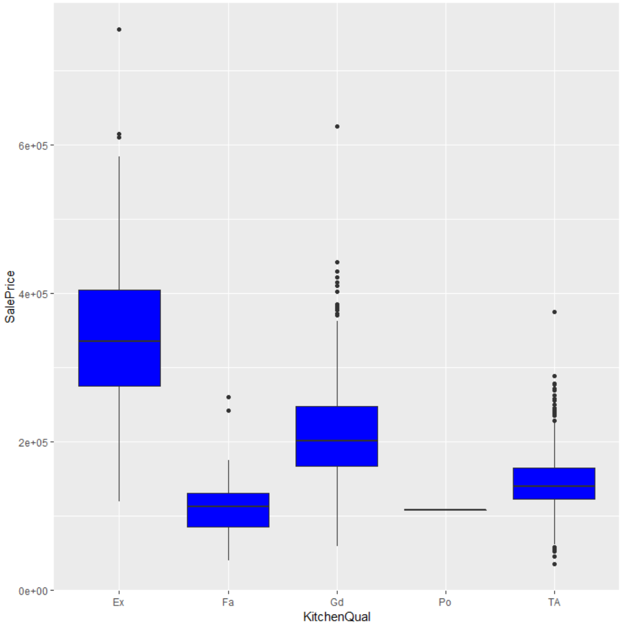
4. "SalePrice" vs " Neighborhood "



We have a good distribution of houses from different neighborhoods. There seems to be some relationship between sale price and neighborhood but without some kind of ranking on neighborhood it cannot be a used as useful predicator.

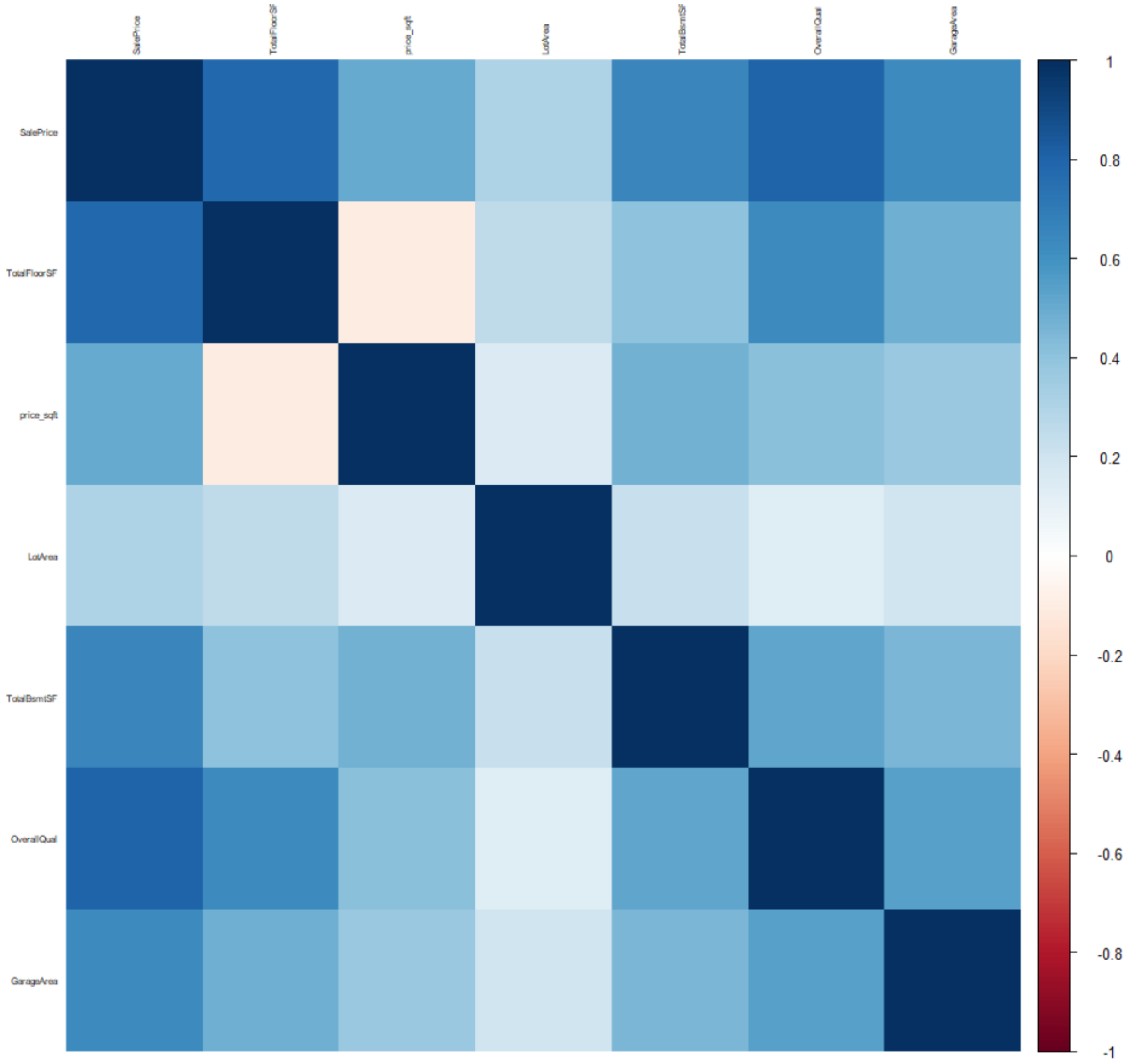
5. "SalePrice" vs " KitchenQual "





From the box plot above we can see positive relationship between sales price and kitchen quality ratings. 97% of houses has Kitchen Quality rating TA or higher.

I also created Correlation matrix plot to see which variables are strongly related to SalePrice response variable. It will be helpful in next section to narrow down predictor variables by selecting the best explanatory variables of SalePrice.



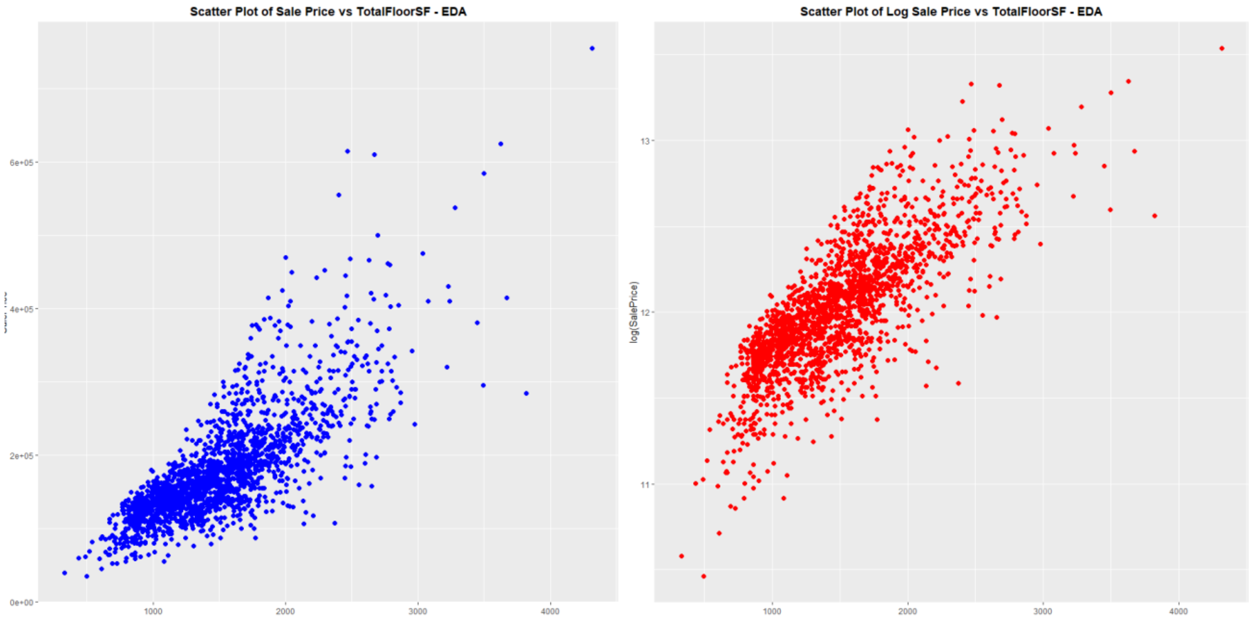
The selected numeric variables are positively correlated to SalePrice (none of them are negatively correlated). TotalFoorSF, OverallQual, TotalBsmtSF and GarageArea seems to be a good predictor of SalePrice. For the selected 10 variables, LotArea is least positively related to SalePrice.

Section 5: Initial Exploratory Data Analysis for Modeling

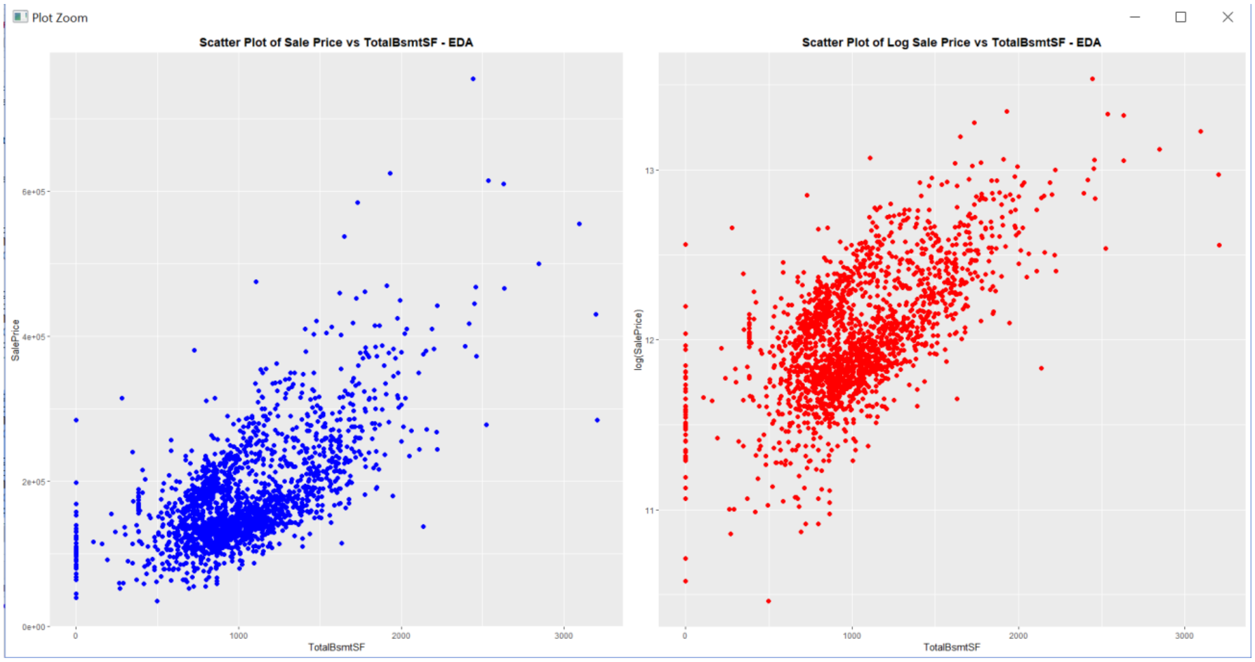
Using above Correlation matrix plot and correlation coefficient of each of the numeric variables with respect to response variable (SalePrice) in this problem, I decided to use following three variables to complete the EDA: TotalFoorSF,TotalBsmtSF, and OverallQual due to their strong positive relationship to sales price.

Before we consider a transformation of the response variable such as log(SalePrice), we will first explore above three variables relationship with SalePrice and log(SalePrice)

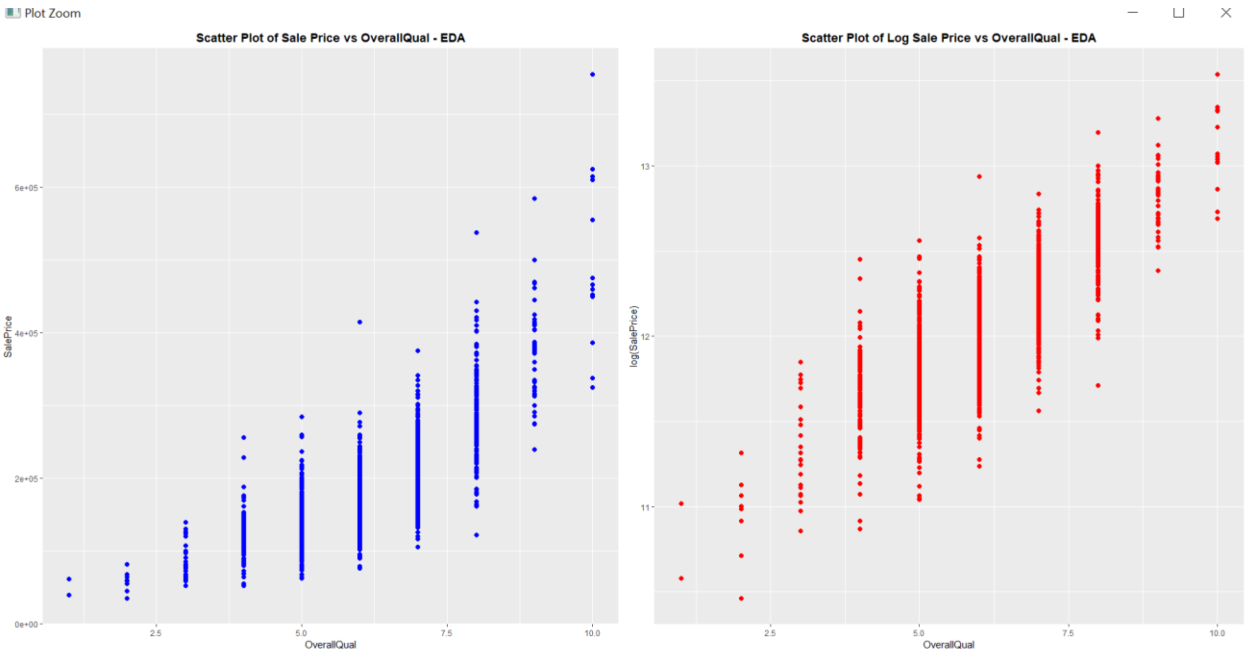
1. Relationship with TotalFoorSF



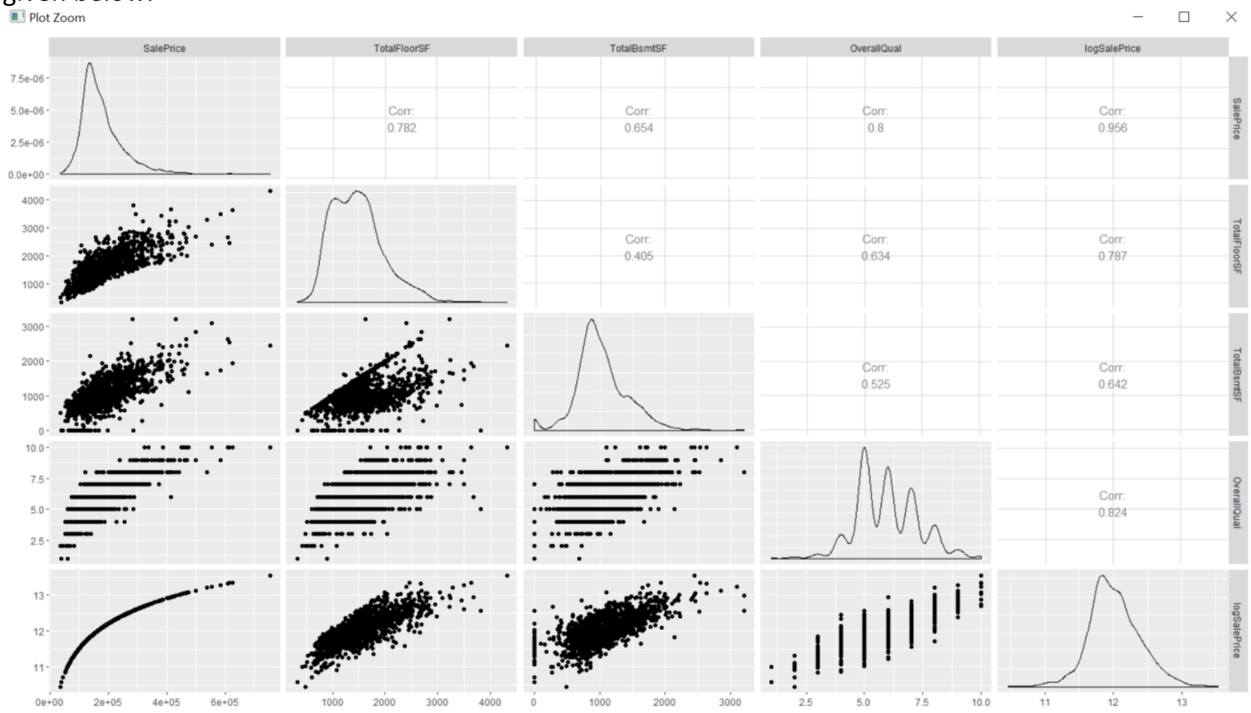
2. Relationship with TotalBsmtSF



3. Relationship with OverallQual



Correlation Plots:
The output from the ggpairs plot for the three input variables and both possible response variables is given below.



It turns out that using the log of the SalePrice may be a slightly better fit. Comparing the Correlation Coefficient for TotalFoorSF for SalePrice and log(SalePrice) is 0.782 and 0.787 respectively. Likewise the values for OverallQual is 0.8 and 0.824 respectively. However, TotalBsmtSF shows a slightly weaker correlation for log(SalePrice). TotalBsmtSF also shows the lowest coefficient values of the three inputs.

The discrete plots for Overall Quality make it slightly more difficult to view trends than the continuous variable TotalFoorSF however the log transformation helps here too. On the whole the two predictor variables TotalFoorSF and OverallQual seem to give the best predictive performance. Our model should also account for transformed value of SalePrice should it be used while predicting house sales prices.

Potential difficulties:
Above EDA analysis indicates existence of outliers in our variables. Future examination of those observation is necessary to filter out such records.
Another issue is homoscedasticity and assumption violations. SalePrice scatter plots indicates heteroscedasticity with variation increasing with sale price. Though this violation can be alleviated by transforming the response variable (sale price), the resulting equation yields difficult to interpret fitted values (selling price in log or square root dollars).

Conclusions:
This assignment give a broad overview of the EDA process in model building and regression analysis. There are some outliers, NA and missing values in our dataset but it is reasonably well structured. We performed first iteration of isolating variables which will likely have superior explanatory power over house prices. More iteration using different combination of potential predictive variables are needed to come up with a efficient predictive model. The relative high number of possible predictor variables in the data poses a challenge when deciding which to use.

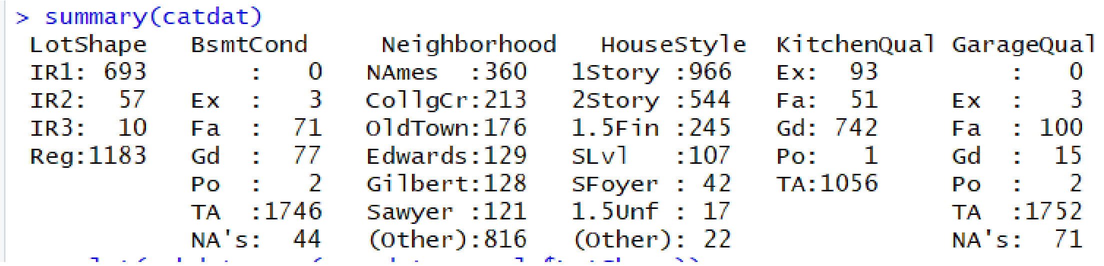
Appendix A: 20 Variables for Data Quality Check

#List of 20 variables to test in the EDA including SalePrice as the response variable

```
> subdat <- subset(mydata,
select=c("SalePrice","TotalFloorSF","HouseAge","price_sqft","LotArea","LotSha
pe","FullBath","BsmtCond","BsmtFinSF1","TotalBsmtSF","Neighborhood","HouseSty
le","YrSold","KitchenQual","OverallCond","OverallQual","YearRemodel","GarageA
rea","GarageQual","PoolArea"))

> names(subdat)
 [1] "SalePrice"    "TotalFloorSF" "HouseAge"     "price_sqft"   "LotArea"
"LotShape"     "FullBath"     "BsmtCond"     "BsmtFinSF1"   "TotalBsmtSF"
[11] "Neighborhood" "HouseStyle"   "YrSold"       "KitchenQual"
"OverallCond"  "OverallQual"  "YearRemodel"  "GarageArea"   "GarageQual"
"PoolArea"
```

#I first looked at categorical variables using summary statistics and bar charts.



```
> summary(catdat)
 LotShape    BsmtCond      Neighborhood   HouseStyle   KitchenQual GarageQual
 IR1: 693        :   0   NAmes  :360   1Story :966   Ex:  93          :   0
 IR2:  57    Ex  :   3   CollgCr:213   2Story :544   Fa:  51     Ex   :   3
 IR3:  10    Fa  :  71   OldTown:176   1.5Fin :245   Gd: 742     Fa   : 100
 Reg:1183    Gd  :  77   Edwards:129   SLvl   :107   Po:   1     Gd   :  15
             Po  :   2   Gilbert:128   SFoyer : 42   TA:1056     Po   :   2
             TA  :1746   Sawyer :121   1.5Unf : 17               TA   :1752
             NA's:  44   (Other):816   (Other): 22               NA's:  71
```

a. About 61 properties have regular LotShape.

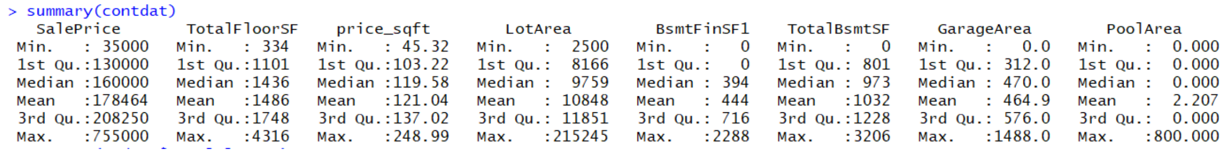
b. Value for Basement condition is either missing or basement doesn't exists in 44 properties; which seems to be ok.

c. Value for Garage Quality is either missing or garage doesn't exists in 71 properties; which seems to be ok.

d. Sample data has good representation of different Neighborhood and HouseStyle.

e. 97% of houses has Kitchen Quality rating TA or higher.

#Summary statistics for **Continues Numeric variables**.

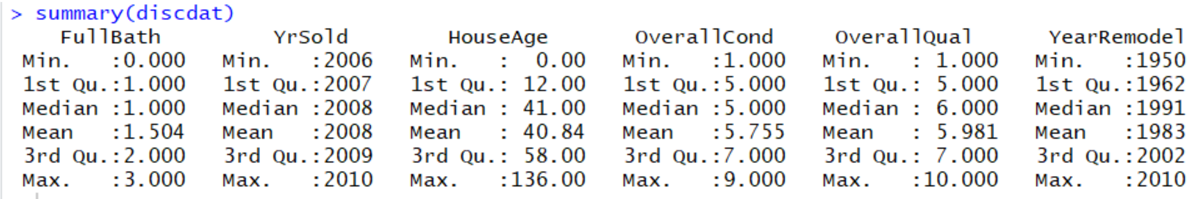


```
> summary(contdat)
   SalePrice       TotalFloorSF    price_sqft        LotArea          BsmtFinSF1      TotalBsmtSF       GarageArea        PoolArea
 Min.   : 35000   Min.   : 334   Min.   : 45.32   Min.   :  2500   Min.   :   0   Min.   :   0   Min.   :   0.0   Min.   :  0.000
 1st Qu.:130000   1st Qu.:1101   1st Qu.:103.22   1st Qu.:  8166   1st Qu.:   0   1st Qu.: 801   1st Qu.: 312.0   1st Qu.:  0.000
 Median :160000   Median :1436   Median :119.58   Median :  9759   Median : 394   Median : 973   Median : 470.0   Median :  0.000
 Mean   :178464   Mean   :1486   Mean   :121.04   Mean   : 10848   Mean   : 444   Mean   :1032   Mean   : 464.9   Mean   :  2.207
 3rd Qu.:208250   3rd Qu.:1748   3rd Qu.:137.02   3rd Qu.: 11851   3rd Qu.: 716   3rd Qu.:1228   3rd Qu.: 576.0   3rd Qu.:  0.000
 Max.   :755000   Max.   :4316   Max.   :248.99   Max.   :215245   Max.   :2288   Max.   :3206   Max.   :1488.0   Max.   :800.000
```

a. Only 13 (0.7%) observations with more than 3500 total floor SF area.

b. Only 4 observations with more than 600,000 sale price.

#Summary statistics for **Discrete Numeric variables**.



```
> summary(discdat)
    FullBath         YrSold        HouseAge        OverallCond     OverallQual      YearRemodel
 Min.   :0.000   Min.   :2006   Min.   :  0.00   Min.   :1.000   Min.   : 1.000   Min.   :1950
 1st Qu.:1.000   1st Qu.:2007   1st Qu.: 12.00   1st Qu.:5.000   1st Qu.: 5.000   1st Qu.:1962
 Median :1.000   Median :2008   Median : 41.00   Median :5.000   Median : 6.000   Median :1991
 Mean   :1.504   Mean   :2008   Mean   : 40.84   Mean   :5.755   Mean   : 5.981   Mean   :1983
 3rd Qu.:2.000   3rd Qu.:2009   3rd Qu.: 58.00   3rd Qu.:7.000   3rd Qu.: 7.000   3rd Qu.:2002
 Max.   :3.000   Max.   :2010   Max.   :136.00   Max.   :9.000   Max.   :10.000   Max.   :2010
```