Assignment #1
Oscar Hernandez

Introduction:
The purpose of this report is to summarize the data exploration process of the Ames, Iowa housing data set. The housing data set contains 2930 observations spread across 83 variables which are either continuous, discrete, ordinal or nominal. The variables are representative of housing characteristics such as bathroom square footage, size of garage, and exterior quality just to name a few. As part of any data analysis project, it is necessary to get a firm understanding of the type of data that is available and determine which parts of the data will aid in our final goal. Our final goal is to build a linear regression model to predict the value of a "typical" home, which we will define in the subsequent section of this report. Therefore, it is prudent to being exploring the housing data to determine if the right data exists to answer our problem.

Data Survey and Sample Population:
The first topic that needs to be addressed is determining what data is available to us. The housing data that is provided represents different characteristics of a home. Our final goal is to build a linear regression model to predict the value of a "typical" home, therefore we would need different quantitative and qualitative measurements of a home to complete our task. The housing data provides us the data that we need and can also help us address problems such as which variables has more predictive influence over the price of a home, determining how many variables allow for a more accurate price prediction and which variables are not helpful with regards to home valuations.
As mentioned, it is necessary to define a "typical" home. The "typical" home will represent our population of interest and will ultimately become our sample data. Therefore, we will create "drop conditions" from the original housing data. The "drop conditions" that have been chosen to represent our "typical" home will be placed in a waterfall which is displayed blow.

Drop Conditions

01: Not Residential	168
02: Not SFR	440
03: Lot Area Greater than 20,000 Square Feet	81
04: Poor House Condition	11
05: Built Pre-1950	533
06: Poor Exterior Condition	152
07: Does not have all utilities	1
08: Home Non Functional	5
09: Non-Normal Sale	267
10: Sale Price Greater than $400,000	26
99: Eligible Sample	1246

The "drop conditions" listed above represent items that are not representative of a "typical home" in Ames, Iowa. For example, the decision was made to exclude homes priced over $400,000 which resulted in 26 non-eligible observations. Furthermore, it makes sense to remove observations where the property sale took place in a non-residential zone (168 observations) and the property was not a single-family home which is indicated by the "SFR" tag (440 observations). The waterfall created provides us with a sample of 1246 eligible observations from the original 2930 in the housing data. Prior to completing any exploratory data analysis (EDA), it is necessary to complete a data quality check on the eligible sample that was created, which will be done next.

Data Quality Check:
Our main concern with regards to the data quality check is determining how many missing values are present and ensuring that we have valid values in our variables (i.e. negative square footage). To complete our data quality check, we will choose twenty variables which are will list them below

Table of 20 Variables

[1] "SalePrice" "LotArea" "LotFrontage" "MasVnrArea" "GarageYrBlt" "OverallCond" "TotalBsmtSF"
[8] "GrLivArea" "GarageArea" "PoolArea" "ScreenPorch" "OpenPorchSF" "WoodDeckSF" "GarageCars"
[15] "OverallQual" "TotRmsAbvGrd" "Fireplaces" "EnclosedPorch" "MiscVal" "MoSold"

As mentioned, one of the concerns is making sure that our variables have correct values within them. Specifically, we need to confirm that our discrete and ordinal variables only include whole numbers, do not have negative numbers and do not include observations that are out of range according to the data dictionary provided. Creating tables (included in the Appendix) allows to easily check all those conditions.

Our discrete and ordinal variables all passed the data quality check. Specifically, the "MoSold" variable did not include any observations where negative numbers, 0 or greater than 12 were present which is required because observation values need to map with actual month values (e.g. should only have 1-12 as valid values). Furthermore, "GarageYrBlt" included values that correctly mapped to actual years and were within range which is necessary considering that we cannot have any values such as 2023.4 in our variable. A table of the remaining discrete and ordinal variables are listed in the Appendix which all display valid values such whole numbers, non-negative numbers and are within the data dictionary range. For illustrative purposes, we will provide the table of "MoSold."

MoSold Table

1	2	3	4	5	6	7	8	9	10	11	12
49	61	96	119	171	247	195	83	63	60	62	40

Lastly, our table of 20 variables also included continuous variables. Using the summary() R function allows us to check that the continuous variables do not contain any negative values. The reason that negative values are problematic is because the continuous variables are measurements of physical areas or monetary value. Therefore, negative values do not make sense. Specifically, one of the most important variables that was included in the data quality check is "SalePrice" which is the sale price of the property. By using the summary() function, we can see that there are not any observations that include negative numbers and we also confirm that the max observation value is $387,000. The output confirms our earlier "drop condition" filter. Moreover, there may be signs of problematic variables such as "MasVnrArea", "PoolArea", "ScreenPorch", "EnclosedPorch" and "MiscVal." All those variables have a median value of 0, which may result in their lack of usefulness when building the final linear regression model. Lastly, we noticed that "LotFrontage", "GarageYrBlt" and "MasVnrArea" include missing values which we need to be aware of since the decision could be made to delete those observations. The Appendix includes the entire summary() output but for illustrative purposes, we will provide the output for "LotArea" below.

LotArea Summary Output

```
    LotArea
 Min.   : 2887
 1st Qu.: 8749
 Median :10016
 Mean   :10486
 3rd Qu.:11828
 Max.   :19900
```

Exploratory Data Analysis – Intro:

The final component of this report is concerned with conducting exploratory data analysis on 10 variables the were a part of the 20 variables used in the data quality check. This section is devoted to highlighting any distribution patterns and relationships amongst the variables. The variables that will make up our EDA are listed in the table below and are broken down by whether they are a continuous or categorical variable. Our EDA will be split between those two types of variables because different methods are preferred when analyzing different types of variables. For example, using a boxplot to analyze the distribution of a discrete variable doesn't make much sense. However, creating a density plot for a continuous

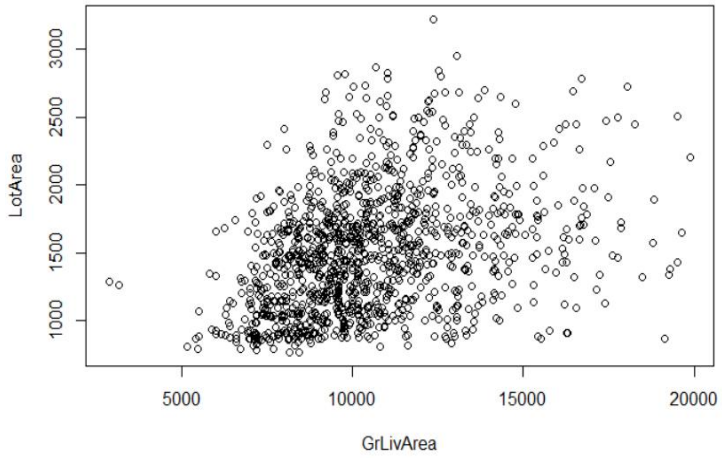
EDA Variables

Continuous	Discrete
LotArea, GrLivArea, LotFrontage, TotalBsmtSF, GarageArea, MasVnrArea	OverallCond, GarageCars, Fireplaces, TotRmsAbvGrd

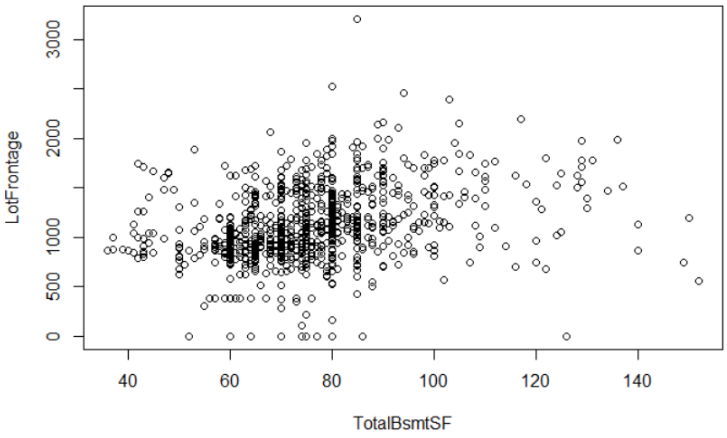
Exploratory Data Analysis – Continuous Variables:

In this section, we will utilize scatterplots to explore the correlation amongst our continuous variables. The following scatterplots illustrate the relationships between the selected continuous variables as noted in the captions.

LotArea vs. GrLivArea Plot



LotFrontage vs. Total Plot

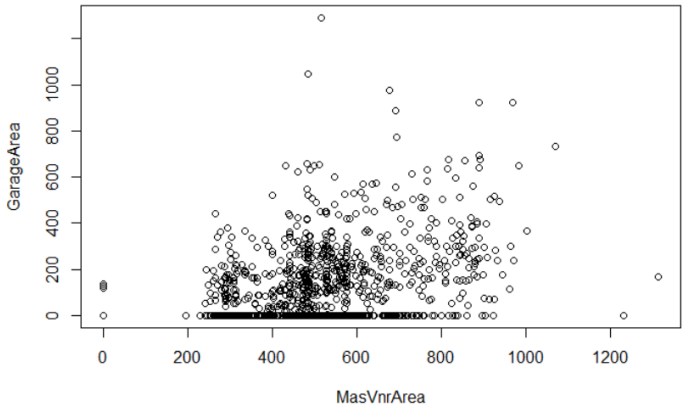


By looking at the scatterplots to the left, we can see very limited evidence that the continuous variables included show signs of correlation.

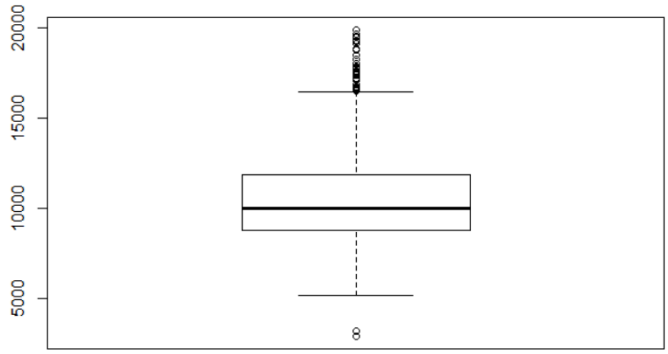
The first scatterplot becomes wider as the data points progress upwards and towards the right, which indicates that for larger data points, the relationship between the variables becomes weaker. In other words, the larger the above grade living area becomes, the less correlated it is as the "LotArea" increases.

Our second scatterplot shows a different pattern and we can also see signs of overplotting as well between the 60-80 "TotalBsmtSF" data points. In this particular scatterplot, there is also very little evidence to believe that the two variables are correlated.

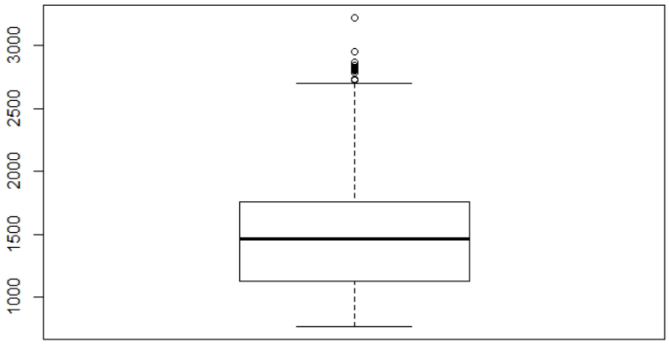
GarageArea vs. MasVnrArea



GarageArea vs. MasVnrArea



GarageArea vs. MasVnrArea



The scatterplot between "GarageArea" and "MasVnrArea" shows the weakest evidence of correlation and linearity. This particular plot suffers from the most overplotting because of the "MasVnrArea." There are many observations that have 0 as their value.

Finally, we can look at a couple boxplots to determine any evidence of normality within our continuous variables.

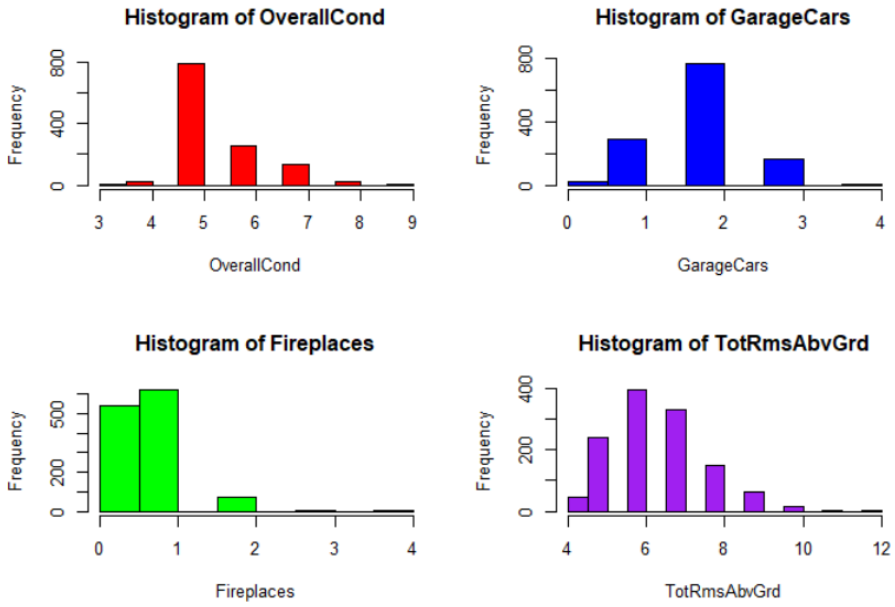
Our first boxplot shows evidence of outliers on both sides of plot. In particular, the "LotArea" variable appears to have an abundance of outliers on the top half of the plot. The distribution of this variable indicates it has a positive skew.

Our final boxplot of the "GrLivArea" variable also shows evidence of outliers. These outliers are concentrated on the top half of the plot. Therefore, we can also say that the "GrLivArea" distribution may have a positive skew as well.

Overall, it doesn't appear that these two variables are following a normal distribution.

Exploratory Data Analysis – Categorical Variables:
We only have four categorical variables that need to be reviewed and the decision was made to use a histogram to get a better sense of the data.
We can see the frequencies of each of these categorical variables which provide some insights into a "typical" home's characteristics. One of the most notable highlights is the lack of fireplaces in "typical" homes. Based on the histogram below, we can see that most of these homes had one or no fireplaces. Furthermore, another noteworthy highlight is that "typical" homes will most likely have garages that fit at most two cars. Overall, these histograms provide us useful information in our assessment of "typical" homes.

Histograms of Categorical Variables



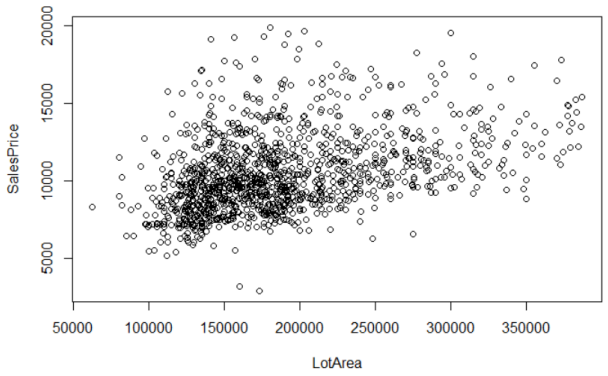
Initial Exploratory Data Analysis for Modeling:
The final section of this report is concerned reviewing the relationship between certain variables and the response variable that is at the crux of our statistical problem – "SalePrice." Furthermore, we will consider the log transformation of "SalePrice" in our assessment. Transformation of data is useful because it can make highly skewed distributions less skewed. Essentially, our goal is to achieve normality prior to any actual modeling.

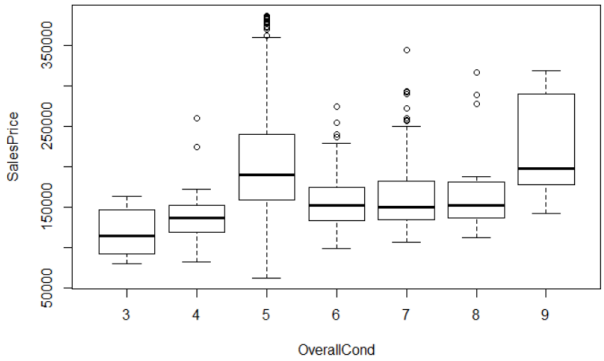
After reviewing the SalePrice plots, we can determine there may be some difficulties for the model building process with regards to certain variables. For example, it doesn't appear that "LotArea" will be useful predictor variable based on the non-linear relationship that it displays with "SalePrice." The same can be said even after transformation. Moreover, this non-linear relationship brings up the notion that it may be necessary to transform the predictor variable during the model building process. As noted earlier, "LotArea" displayed a positive skew, so if it were to be used it in the predictive model, we may need to do a log transformation to make it less skewed.

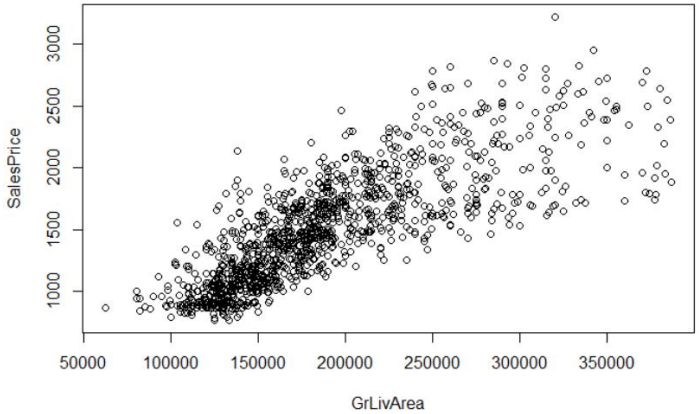
Furthermore, a discrete variable "OverallCond" doesn't appear to be useful as a predictor variable. We can see that as the overall condition of the home becomes larger, the median price of the home doesn't appear to change much. We would think that a home with an excellent condition rating would have a far higher selling price than a home with an average rating.

Lastly, we do see some good evidence of linearity with regards to the "GrLivArea" variable which makes it our best candidate so far in terms of using it as a predictor variable in our model. On the following page, all the plots used in the EDA for modeling section are displayed.

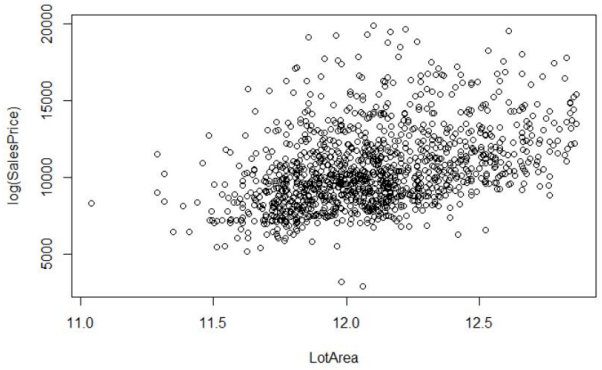
SalePrice Plots

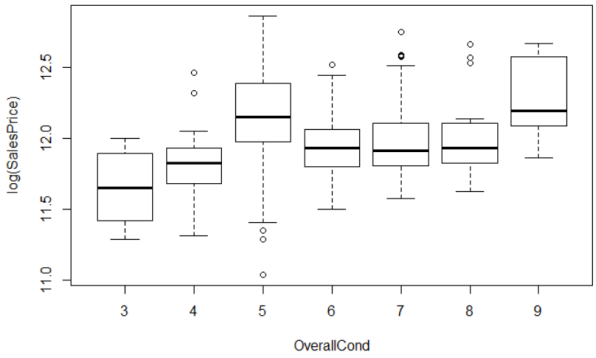


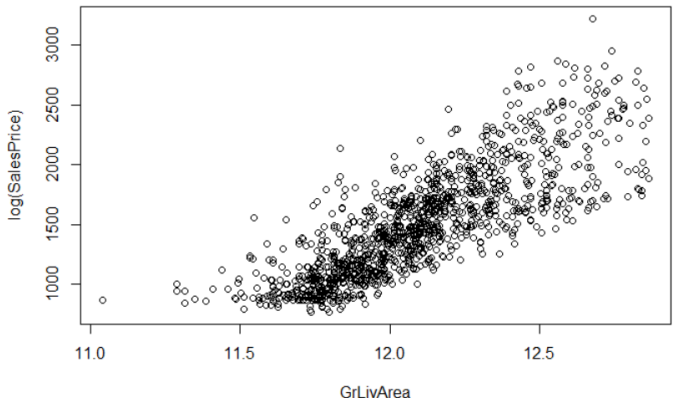




Log Transformation of SalePrice Plots







Conclusions:

The goal of this report was to explore the Ames, Iowa housing data and gain a base-level understanding of the data while keeping in mind that the final task will be to develop a linear regression model that predicts a “typical” home value. To develop this model, it is necessary to determine if we have the right data at hand, which our analysis has concluded that we do since we were able to create a sample population based on criteria we deemed characteristic of a “typical” home. Furthermore, reviewing the quality of the data provided us insight into how “clean” the data was which is especially important because having invalid data will make our model weaker, if not useless. We can conclude that the variables we reviewed passed the data quality check. Lastly, it was necessary to conduct EDA on continuous and categorical variables to check if linearity relationships existing between variables, check normality assumptions and determine if any potential difficulties may arise during the model building process. Overall, it can be concluded that there shouldn’t be any major issues that arise during the model building processes and that there exist some potentially useful predictor variables such as GrLivArea.

Appendix:

Data Quality Check Tables (Discrete and Ordinal variables)

```
> table(qc_20$OverallCond)

  3   4   5   6   7   8   9
  8  23 795 254 134  26   6
> table(qc_20$OverallQual)

  3   4   5   6   7   8   9  10
  3  59 362 320 302 163  35   2
> table(qc_20$GarageYrBlt)

1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973
  19   13    8   15   22   11   28   25   21   26   21   20   23   18   23   17   26   27   28   19   11   12    8    6
1974 1975 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997
  24   13   19   39   22   10   11    8    4    8   10    5    4    7   11    7   18   11   14   33   22   20   28   26
1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009
  40   31   30   27   44   51   56   50   25   30   17    4
> table(qc_20$GarageCars)

  0   1   2   3   4
 20 286 772 166   2
> table(qc_20$Fireplaces)

  0   1   2   3   4
538 626  75   6   1
> table(qc_20$MoSold)

  1   2   3   4   5   6   7   8   9  10  11  12
 49  61  96 119 171 247 195  83  63  60  62  40
```

Data Quality Summary Output

```
   SalePrice        LotArea        LotFrontage      MasVnrArea      GarageYrBlt    OverallCond    TotalBsmtSF
Min.   : 62383   Min.   : 2887   Min.   : 36.00   Min.   :   0.0   Min.   :1950   Min.   :3.00   Min.   :   0
1st Qu.:142000   1st Qu.: 8749   1st Qu.: 65.00   1st Qu.:   0.0   1st Qu.:1965   1st Qu.:5.00   1st Qu.: 864
Median :174000   Median :10016   Median : 74.00   Median :   0.0   Median :1984   Median :5.00   Median :1054
Mean   :187269   Mean   :10486   Mean   : 75.79   Mean   : 113.3   Mean   :1983   Mean   :5.47   Mean   :1100
3rd Qu.:217000   3rd Qu.:11828   3rd Qu.: 82.00   3rd Qu.: 188.0   3rd Qu.:2000   3rd Qu.:6.00   3rd Qu.:1313
Max.   :387000   Max.   :19900   Max.   :152.00   Max.   :1290.0   Max.   :2009   Max.   :9.00   Max.   :3206
                                 NA's   :273      NA's   :9        NA's   :20
   GrLivArea      GarageArea       PoolArea        ScreenPorch      OpenPorchSF      WoodDeckSF       GarageCars
Min.   : 768   Min.   :   0.0   Min.   :  0.000   Min.   :  0.00   Min.   :  0.00   Min.   :   0.0   Min.   :0.000
1st Qu.:1126   1st Qu.: 396.0   1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:   0.0   1st Qu.:2.000
Median :1462   Median : 484.0   Median :  0.000   Median :  0.00   Median : 35.00   Median :  84.0   Median :2.000
Mean   :1499   Mean   : 501.3   Mean   :  2.519   Mean   : 18.11   Mean   : 47.66   Mean   : 109.2   Mean   :1.875
3rd Qu.:1760   3rd Qu.: 578.0   3rd Qu.:  0.000   3rd Qu.:  0.00   3rd Qu.: 72.00   3rd Qu.: 192.0   3rd Qu.:2.000
Max.   :3222   Max.   :1314.0   Max.   :800.000   Max.   :576.00   Max.   :324.00   Max.   :1424.0   Max.   :4.000

 OverallQual      TotRmsAbvGrd      Fireplaces      EnclosedPorch       MiscVal            MoSold
Min.   : 3.000   Min.   : 4.000   Min.   :0.0000   Min.   :  0.00   Min.   :    0.00   Min.   : 1.000
1st Qu.: 5.000   1st Qu.: 6.000   1st Qu.:0.0000   1st Qu.:  0.00   1st Qu.:    0.00   1st Qu.: 4.000
Median : 6.000   Median : 6.000   Median :1.0000   Median :  0.00   Median :    0.00   Median : 6.000
Mean   : 6.202   Mean   : 6.471   Mean   :0.6404   Mean   : 13.22   Mean   :   54.22   Mean   : 6.124
3rd Qu.: 7.000   3rd Qu.: 7.000   3rd Qu.:1.0000   3rd Qu.:  0.00   3rd Qu.:    0.00   3rd Qu.: 7.000
Max.   :10.000   Max.   :12.000   Max.   :4.0000   Max.   :368.00   Max.   :12500.00   Max.   :12.000
```

Code:

```
#Read in our Ames Housing Data
ames_df <- read.csv('ames_housing_data.csv', header=TRUE, stringsAsFactors = FALSE)

#Create waterfall of drop conditions
ames_df$dropCondition <- ifelse(!ames_df$Zoning %in% c('RH','RL','RP','RM'), '01: Not Residential',
 ifelse(ames_df$BldgType!='1Fam','02: Not SFR',
 ifelse(ames_df$LotArea>20000,'03: Lot Area Greater than 20,000 Square Feet',
 ifelse(ames_df$OverallCond<3,'04: Poor House Condition',
 ifelse(ames_df$YearBuilt<1950,'05: Built Pre-1950',
 ifelse(!ames_df$ExterCond %in% c('Ex', 'GD', 'TA', 'Fa'),'06: Poor Exterior Condition',
 ifelse(ames_df$Utilities!='AllPub','07: Does not have all utilities',
 ifelse(!ames_df$Functional %in% c('Typ', 'Min1', 'Min2', 'Mod'),'08: Home Non Functional',
 ifelse(ames_df$SaleCondition!='Normal','09: Non-Normal Sale',
 ifelse(ames_df$SalePrice>400000, '10: Sale Price Greater than $400,000',
 '99: Eligible Sample'))))))))))

#Create a table summarizing Drop Conditions
table(ames_df$dropCondition)

#Save table as 'waterfall'
waterfall <- table(ames_df$dropCondition)

#Display the table as a column matrix
as.matrix(waterfall, 11,1)

# Eliminate all observations that are not part of the eligible sample population
eligible_population <- subset(ames_df,dropCondition=='99: Eligible Sample')

# Check that all remaining observations are eligible
table(eligible_population$dropCondition)

#####Data Quality Check Section#####
#Create a subset of the 20 variables we want to complete a data quality check on
qc_20 <- subset(eligible_population, select = c(SalePrice, LotArea, LotFrontage, MasVnrArea,
GarageYrBlt, OverallCond, TotalBsmtSF,GrLivArea,GarageArea, PoolArea,ScreenPorch,OpenPorchSF,
WoodDeckSF, GarageCars, OverallQual, TotRmsAbvGrd,Fireplaces, EnclosedPorch, MiscVal, MoSold))

#Create a table displaying the names of our variables
colnames(qc_20)
quality_check20_table <- colnames(qc_20)
quality_check20_table

#Print out summary of 20 variables in qc_20
summary(qc_20)

#We have a combination of continuous, discrete and ordinal variables
#For our ordinal and discrete variables, we want to create tables
table(qc_20$OverallCond)
table(qc_20$OverallQual)
table(qc_20$GarageYrBlt)
table(qc_20$GarageCars)
table(qc_20$TotRmsAbvGrd)
```

```
table(qc_20$Fireplaces)
table(qc_20$MoSold)

###Initial Exploratory Data Analysis of continuous variables###
plot( qc_20$LotArea, qc_20$GrLivArea, ylab='LotArea', xlab='GrLivArea')
plot( qc_20$LotFrontage, qc_20$TotalBsmtSF, ylab='LotFrontage', xlab='TotalBsmtSF')
plot( qc_20$GarageArea, qc_20$MasVnrArea, ylab='GarageArea', xlab='MasVnrArea')

boxplot(qc_20$LotArea, xlab='LotArea')
boxplot(qc_20$GrLivArea, xlab='GrLivArea')

###Initial Exploratory Data Analysis of categorical variables###
par(mfrow=c(2,2))
hist(qc_20$OverallCond, main = "Histogram of OverallCond", xlab= "OverallCond", col = 'red')
hist(qc_20$GarageCars, main="Histogram of GarageCars", xlab= "GarageCars", col='blue')
hist(qc_20$Fireplaces, main = "Histogram of Fireplaces", xlab="Fireplaces", col='green')
hist(qc_20$TotRmsAbvGrd, main="Histogram of TotRmsAbvGrd", xlab= "TotRmsAbvGrd", col='purple')
par(mfrow=c(1,1))

##Initial Exploratory Data Analysis for Modeling with SalePrice###
plot(qc_20$SalePrice, qc_20$LotArea,ylab='SalesPrice', xlab='LotArea')
boxplot(qc_20$SalePrice~qc_20$OverallCond, data=qc_20, ylab='SalesPrice', xlab='OverallCond')
plot(qc_20$SalePrice, qc_20$GrLivArea,ylab='SalesPrice', xlab='GrLivArea')

##Initial Exploratory Data Analysis for Modeling with log(SalePrice)###
plot(log(qc_20$SalePrice), qc_20$LotArea,ylab='log(SalesPrice)', xlab='LotArea')
boxplot(log(qc_20$SalePrice)~qc_20$OverallCond, data=qc_20, ylab='log(SalesPrice)',
xlab='OverallCond')
plot(log(qc_20$SalePrice), qc_20$GrLivArea,ylab='log(SalesPrice)', xlab='GrLivArea')
```