Ames Housing Data Analysis
Assignment #1
John Moderwell
MSDS 410

Introduction:

The Ames Housing data set consists of information from the Ames Assessor's Office and is used to compute the value of individual residential properties sold in the city of Ames, Iowa from 2006 to 2010. The purpose of this report is to investigate the data and report the findings of the initial exploratory data analysis. This report is broken down into five sections. Section 1 provides a definition and description of the data that will be used in the analysis. Section 2 provides the results from the data quality check of the twenty variables selected. Section 3 is the initial exploratory data analysis which takes a closer look at ten selected variables and reports the results. Section 4 builds upon Section 3 and provides an exploratory data analysis for modeling purposes and discusses the 3 variables selected for this analysis. Section 5 provides a summary of the results and discusses any general observations about the EDA and considerations for future reports. This section also reflects on how the decisions made in the EDA and its results will be used to build a multivariate regression model to predict the sale price of houses as the course progresses.

Sample Definition:

The Ames Housing dataset is comprised of 82 columns including 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables. There are also 2 columns (SID and PID) used as observation identifiers. These are not relevant for the purposes of this assignment and were removed. There were also 5 additional variables created. This gives a total of 85 variables used for 2930 residential properties.

It was observed that multiple building types were included in the original data set. It does not make sense to compare one building type to another as each type has a different perception of value in the eyes of a buyer and this is reflected in the different zoning types associated with each building.

The original dataset consists of 5 building types and 8 zoning classifications :

	A (agr)	C (all)	FV	I (all)	RH	RL	RM
1Fam	2	22	77	2	12	2009	301
2fmCon	0	3	0	0	4	30	25
Duplex	0	0	0	0	4	92	13
Twnhs	0	0	19	0	1	28	53
TwnhsE	0	0	43	0	6	114	70

As the purpose of this assignment is to predict the sale price of a home, it does not make sense to include buildings falling under the zoning classifications industrial (I), agriculture (A) or

commercial (C). Residential zoning types including residential high density (RH), residential low density (RL), residential low density park (RP), residential medium density (RM) and floating village residential will be used for further analysis. For building type, we are looking to analyze houses that represent the typical “house” that a family would buy. Duplexes (Duplx) and townhouses (TwnhsE and TwnhsI) are not good representations and are removed leaving single family detached (1Fam) and two family conversion houses (2FmCon) left for analysis. After these conditions are met, there are 2399 observations remaining. Here is a breakdown of the distribution of homes in the four zoning classifications by building type.

	A (agr)	C (all)	FV	I (all)	RH	RL	RM
1Fam	0	0	77	0	12	2004	301
2fmCon	0	0	0	0	0	0	0
Duplex	0	0	0	0	0	0	0
Twnhs	0	0	0	0	0	0	0
TwnhsE	0	0	0	0	0	0	0

It is evident that none of the properties fall under the two family conversion building type so this can be removed from consideration. While investigating the data further, it was also observed that the majority of properties ranged from 250 total sq ft. to 4000 sq ft. There were 5 properties with over 4000 total sq ft. which were identified as outliers and removed from the sample. This leaves us with a dataset of 85 variables and 2395 properties.

Data Quality Check:

After performing an exploratory analysis of the 85 variables, the assignment requires that 20 variables be selected based on their ability to help predict sale price. Using knowledge as a potential home buyer and descriptive information of the variables provided in the Ames housing documentation text, potential variables are narrowed down to about 35 variables. From here, correlations between sale price and numerical variables are studied and visualized with correlation matrices and scatter plots. For categorical variables, relationships are studied and visualized using box plots. Below is a list of the variables selected for further analysis:

```
 [1] "LotArea"      "HouseAge"     "TotalFloorSF"  "SaleCondition" "PavedDrive"   "GarageArea"   "GarageFinish"
 [8] "FireplaceQu"  "KitchenQual"  "TotRmsAbvGrd"  "OverallQual"   "YearBuilt"    "BsmtQual"     "TotalBsmtSF"
[15] "FirstFlrSF"   "GrLivArea"    "FullBath"      "GarageCars"    "SecondFlrSF"  "CentralAir"
```

Ames Housing Data Analysis
Assignment #1
John Moderwell
MSDS 410

Below are the results of the data quality check showing the variable, variable type, missing value counts/percentages, number of unique values and the ratio of unique observations to total observations. This table was created using the diagnose() function from the dlookr package.

	variables	types	missing_count	missing_percent	unique_count	unique_rate
1	SalePrice	integer	0	0.00000000	912	0.3809523810
2	LotArea	integer	0	0.00000000	1651	0.6896407686
3	HouseAge	integer	0	0.00000000	125	0.0522138680
4	TotalFloorSF	integer	0	0.00000000	1192	0.4979114453
5	SaleCondition	factor	0	0.00000000	6	0.0025062657
6	PavedDrive	factor	0	0.00000000	3	0.0012531328
7	GarageArea	integer	1	0.04177109	585	0.2443609023
8	GarageFinish	factor	89	3.71762740	5	0.0020885547
9	FireplaceQu	factor	1090	45.53049290	6	0.0025062657
10	KitchenQual	factor	0	0.00000000	5	0.0020885547
11	TotRmsAbvGrd	integer	0	0.00000000	11	0.0045948204
12	OverallQual	integer	0	0.00000000	10	0.0041771094
13	YearBuilt	integer	0	0.00000000	115	0.0480367586
14	BsmtQual	factor	45	1.87969925	6	0.0025062657
15	TotalBsmtSF	integer	1	0.04177109	967	0.4039264829
16	FirstFlrSF	integer	0	0.00000000	990	0.4135338346
17	GrLivArea	integer	0	0.00000000	1193	0.4983291562
18	FullBath	integer	0	0.00000000	4	0.0016708438
19	GarageCars	integer	1	0.04177109	7	0.0029239766
20	SecondFlrSF	integer	0	0.00000000	586	0.2447786132
21	CentralAir	factor	0	0.00000000	2	0.0008354219

From this table, it is initially apparent that there are missing values in GarageArea, GarageFinish, FireplaceQu, BsmtQual, TotalBsmtSF and GarageCars. This will be further explored. The selected variables are then separated by data type: integer and factor. Below are summaries of the 13 integer variables and the 7 factor variables. The summary of the integer variables shows the minimum, the 1st quartile value, the median, the mean, the 3rd quartile value and the maximum value. The summary of the factor variables shows the frequency counts for each class.

Ames Housing Data Analysis
Assignment #1
John Moderwell
MSDS 410

Integer Variables Summary:

```
   SalePrice        LotArea          HouseAge       TotalFloorSF    GarageArea      TotRmsAbvGrd     OverallQual
Min.   : 12789   Min.   :  2500   Min.   :  0.00   Min.   : 334   Min.   :   0.0   Min.   : 2.000   Min.   : 1.000
1st Qu.:131425   1st Qu.:  8214   1st Qu.:  8.00   1st Qu.:1127   1st Qu.: 325.0   1st Qu.: 6.000   1st Qu.: 5.000
Median :165200   Median :  9806   Median : 38.00   Median :1464   Median : 480.0   Median : 6.000   Median : 6.000
Mean   :185572   Mean   : 10886   Mean   : 37.71   Mean   :1511   Mean   : 483.1   Mean   : 6.514   Mean   : 6.112
3rd Qu.:220000   3rd Qu.: 11952   3rd Qu.: 56.00   3rd Qu.:1780   3rd Qu.: 588.0   3rd Qu.: 7.000   3rd Qu.: 7.000
Max.   :625000   Max.   :215245   Max.   :136.00   Max.   :3820   Max.   :1488.0   Max.   :12.000   Max.   :10.000
                                                                  NA's   :1
   YearBuilt     TotalBsmtSF     FirstFlrSF     GrLivArea      FullBath       GarageCars     SecondFlrSF
Min.   :1872   Min.   :   0   Min.   : 334   Min.   : 334   Min.   :0.000   Min.   :0.000   Min.   :   0.0
1st Qu.:1951   1st Qu.: 816   1st Qu.: 892   1st Qu.:1132   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:   0.0
Median :1970   Median : 992   Median :1086   Median :1468   Median :2.000   Median :2.000   Median :   0.0
Mean   :1970   Mean   :1061   Mean   :1167   Mean   :1516   Mean   :1.544   Mean   :1.789   Mean   : 343.6
3rd Qu.:2000   3rd Qu.:1281   3rd Qu.:1386   3rd Qu.:1786   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.: 728.0
Max.   :2010   Max.   :3206   Max.   :3820   Max.   :3820   Max.   :3.000   Max.   :5.000   Max.   :1862.0
               NA's   :1                                                    NA's   :1
```

By looking at this information, it is apparent that there are three missing values, one in GarageArea, one in TotalBsmtSF, and one in GarageCars. These observations will be removed. Other than that, the data seems to be fairly complete.

Factor Variables Summary:

```
SaleCondition  PavedDrive GarageFinish FireplaceQu KitchenQual BsmtQual    CentralAir
Abnorml: 147   N: 160         :   1    Ex  :  39   Ex: 169         :   1   N: 128
AdjLand:   7   P:  59     Fin : 594    Fa  :  66   Fa:  54     Ex  : 210   Y:2266
Alloca :   8   Y:2175     RFn : 687    Gd  : 646   Gd: 966     Fa  :  78
Family :  41              Unf :1023    Po  :  40   Po:   1     Gd  : 946
Normal :1987              NA's:  89    TA  : 513   TA:1204     Po  :   0
Partial: 204                           NA's:1090               TA  :1114
                                                               NA's:  45
```

While the initial data quality check shows the NA values as missing values, further investigation shows that this is not the case. The NA's actually indicate that the physical feature is not present in the property. Therefore, these values will be kept and the data seems accurate and complete.

In terms of outliers, there does not appear to be any values that should not be included in future analysis other than the 5 outliers greater than 4000 total sq ft that were removed earlier.

Initial Exploratory Data Analysis:

Since Sale Price is the only response variable, it is important to get a good idea of how the data is shaped. Here is bar plot of Sale Price:

Ames Housing Data Analysis
Assignment #1
John Moderwell
MSDS 410

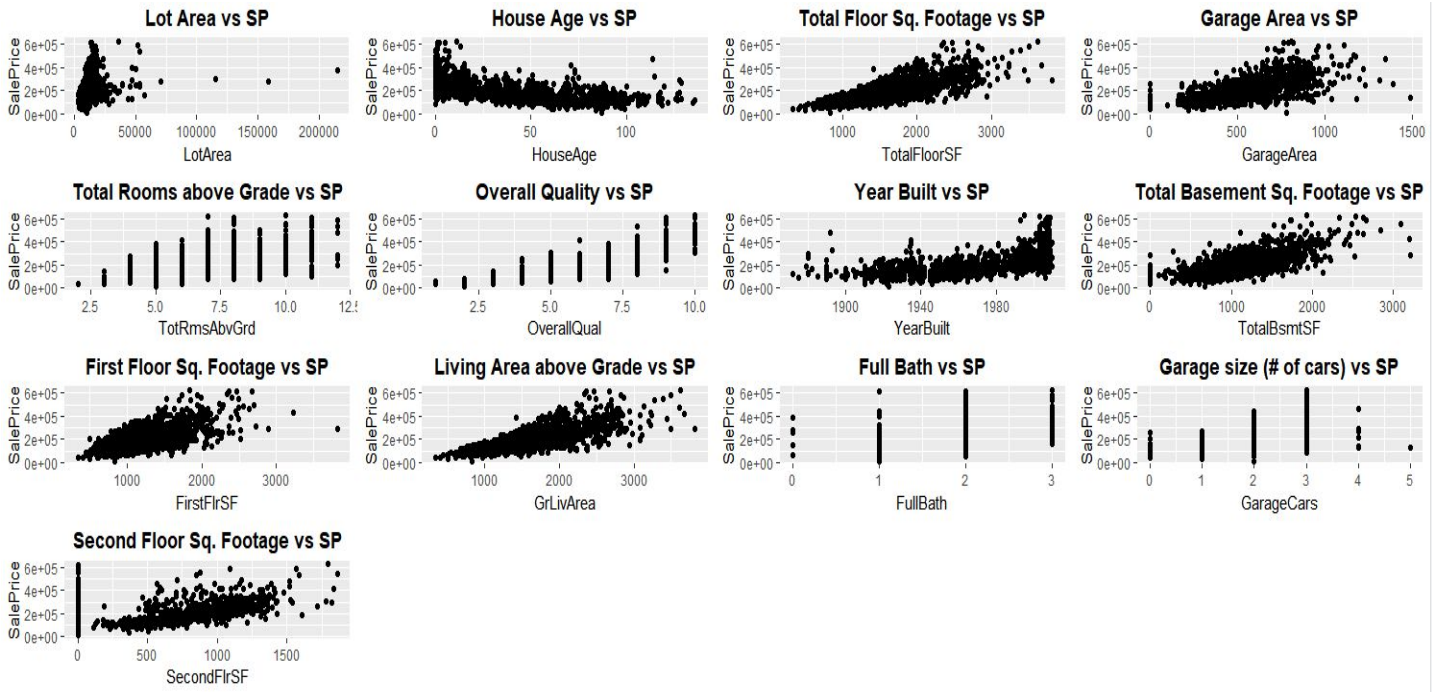
It is evident that the distribution is mostly normal with some positive skewness (meaning there is a tail extending to the right). Skewness indicates the symmetry of the distribution whereas kurtosis is a measure determining if the diistribution is heavy tailed or light-tailed in relation to a normal distribution The exact value of the skewness was calculated to be 1.60. The kurtosis was calculated to be 3.71 which indicates that the tails are slightly heavier than a normal distribution (kurtosis = 3).

The assignment requires that 10 variables be selected from the 20 variables for further data analysis and possible use in modeling. These variables were selected based on the correlation strength in relation to the response variable (Sale Price). The strength of the relationship was the biggest factor in choosing variables and both strongly positive and strongly negative correlations were desired. This was especially true for numerical variables. In addition, plots visualizing these relationships were important for categorical variables as well as using prior knowledge about the data set and common sense.

Ames Housing Data Analysis
Assignment #1
John Moderwell
MSDS 410

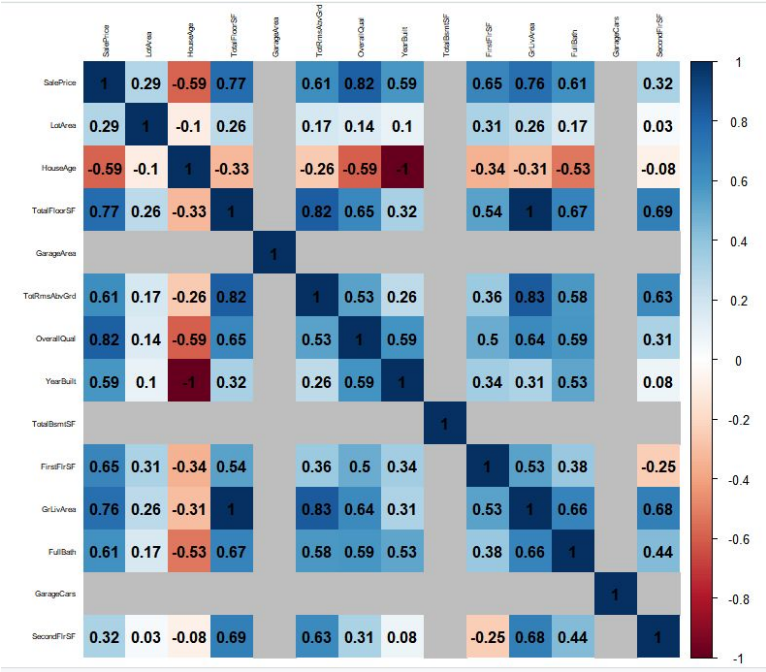
Numerical Variables EDA:

This graphic visualizes the relationship between explanatory numerical variables and the response variable Sale Price using scatterplots. There seems to be especially strong correlations in variables TotFloorSF, TotalBsmtSF, GrLivArea and OverallQual. These relationships will be explored further by calculating the correlations.



Ames Housing Data Analysis
Assignment #1
John Moderwell
MSDS 410

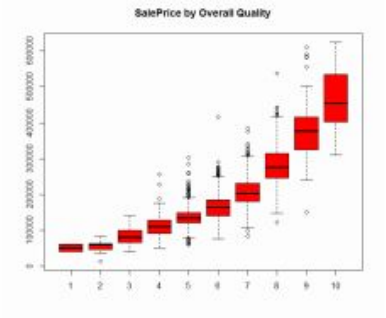
This graphic visualizes a correlation matrix between the numerical variables:



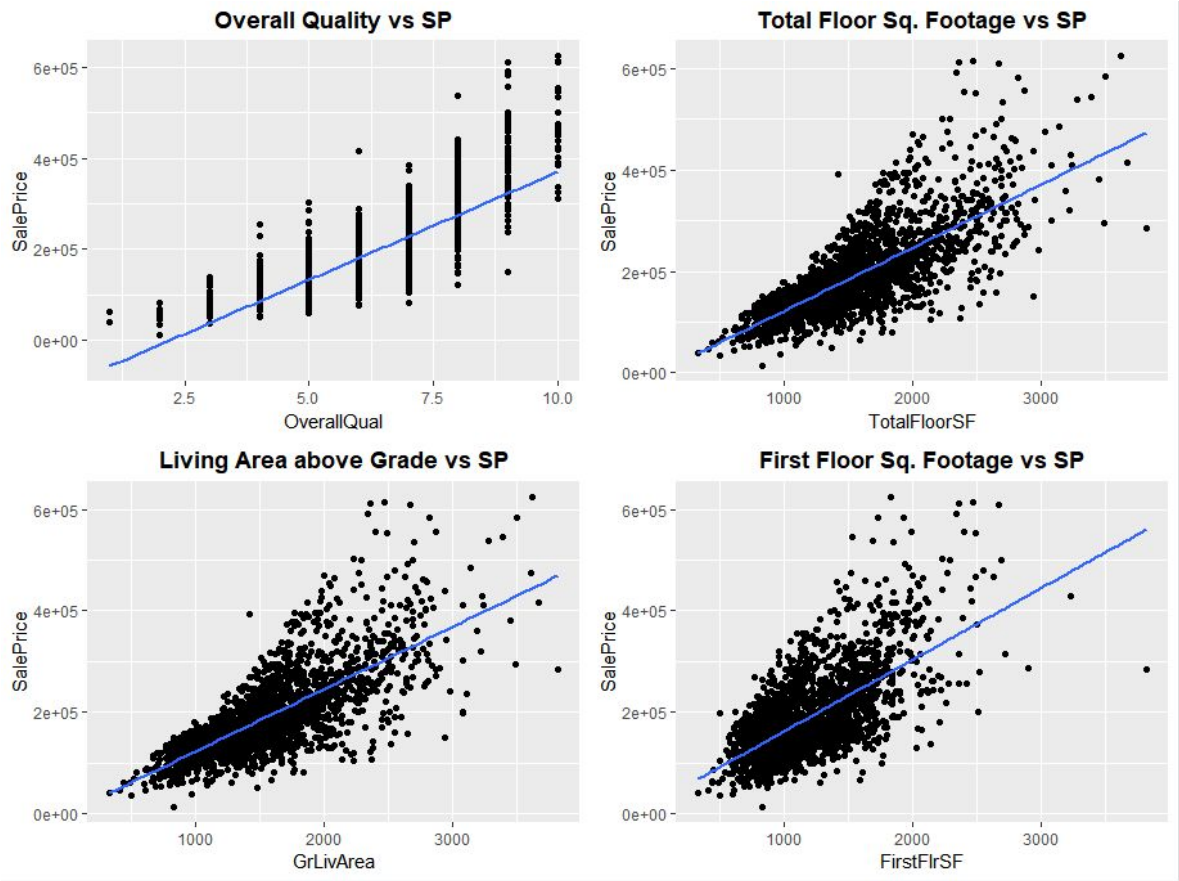
It is evident that OverallQual (.82), TotalFloorSF (.77) GrLivArea (.76), and FirstFlrSF (.65) have the highest correlations with Sale Price. FullBath and TotRmsAbvGrd (.61 each) also have some correlation but are not as strong. The only negative relationship seen is with HousingAge (-.59) but does not seem strong enough to warrant further consideration.

Although OverallQual has the highest correlation with Sale Price, it is an ordinal variable that is based on subjective findings. The boxplot below shows that there is significant overlap between the classes which makes OverallQual less than ideal as a predictor variable..

Ames Housing Data Analysis
Assignment #1
John Moderwell
MSDS 410



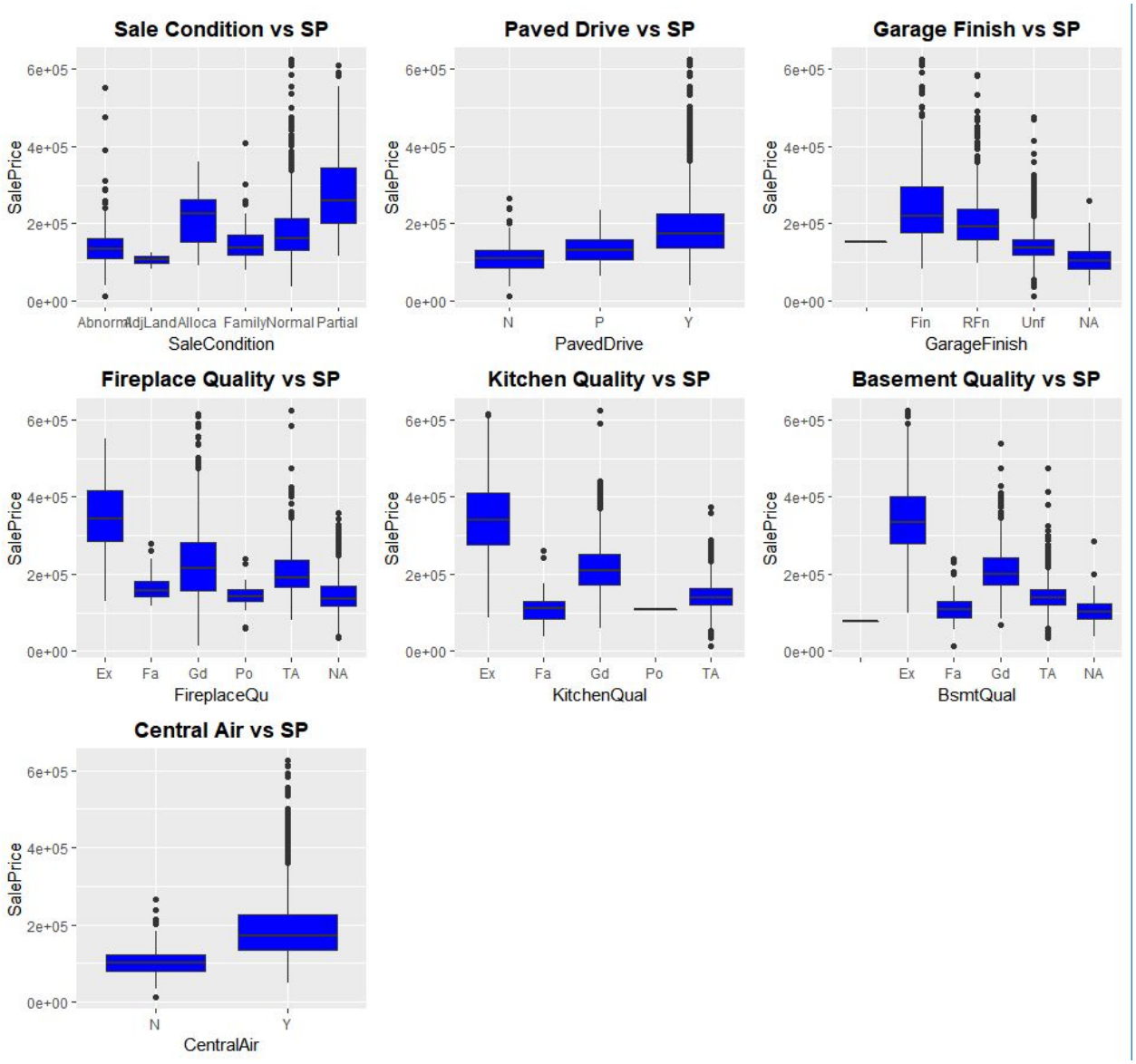
Variables OverallQual, TotalFloorSF, GrLivArea and FirstFlrSF will be visualized individually and plotted against Sale Price to get a better idea of the relationship. A line of best fit will also be included for each plot.



Based on these plots and strength of correlation, variables FirstFlrSF and GrLivArea will be used for further analysis.

Ames Housing Data Analysis
Assignment #1
John Moderwell
MSDS 410

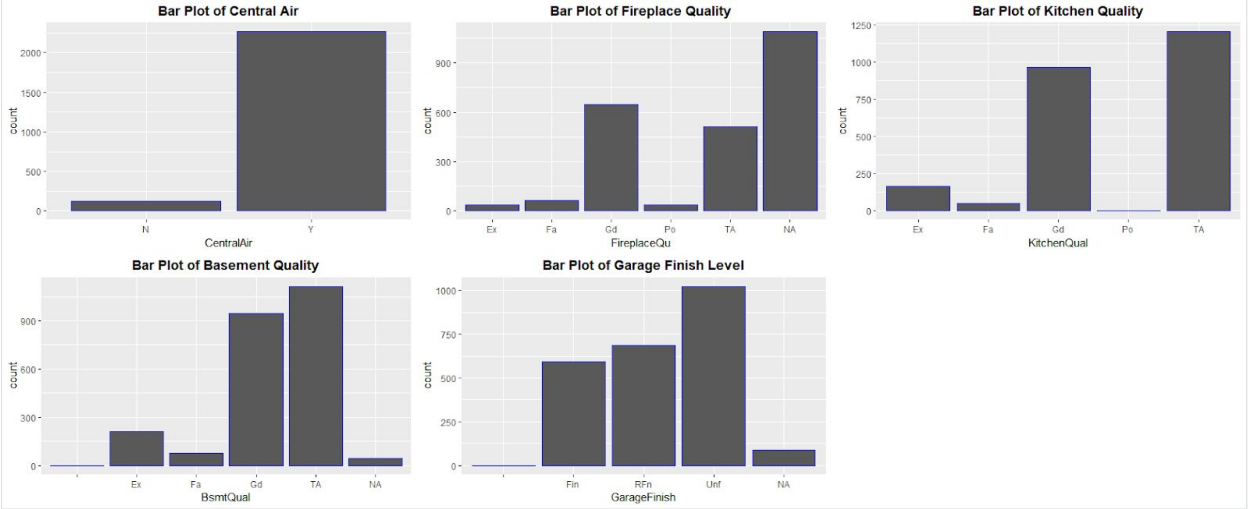
Categorical Variables EDA:



From these boxplots of categorical variables, there is some indication that the presence of a paved driveway has a positive effect on sale price. Similarly, the existence of finished garages and central air as well as high quality fireplaces, kitchens and basements correlate with higher prices. However, correlation is not causation and they may not be the driving factors for the

Ames Housing Data Analysis
Assignment #1
John Moderwell
MSDS 410

increases in price. Variables CentralAir, FireplaceQu, KitchenQual, BsmtQual and GarageFinish need to be studied further.

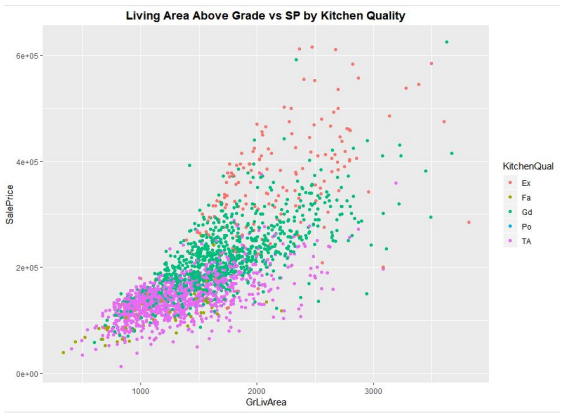


After studying these bar plots, FireplaceQu does not appear to be a good predictor variable as over 1000 properties do not have a fireplace. KitchenQual, BsmtQual and GarageFinish have the potential to be good predictors of sale price. As a potential home buyer, I would say that kitchen quality and basement quality are better drivers of price because they are features that are used more heavily. Between BsmtQual and KitchenQual, KitchenQual appears to have a lesser amount of overlap between each class (Excellent, Good, Typical/Average, Fair and Poor). This indicates that kitchen quality has a direct effect on price. Therefore, KitchenQual will be used for further analysis.

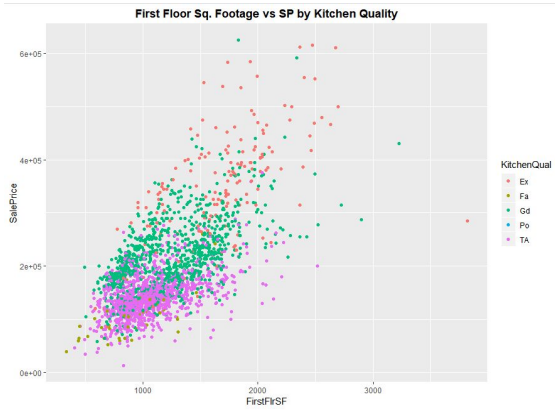
Ames Housing Data Analysis
Assignment #1
John Moderwell
MSDS 410

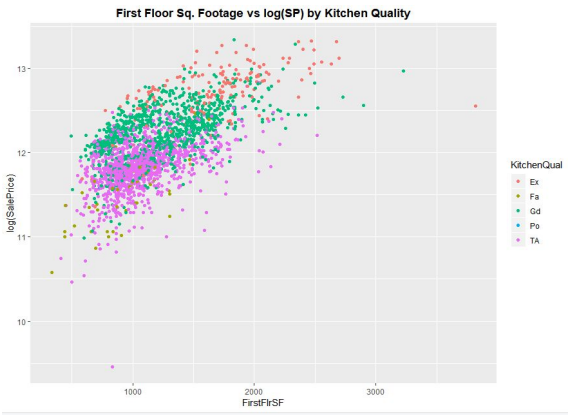
Exploratory Data Analysis for Modeling:

From the initial exploratory data analysis, quantitative variables FirstFlrSF and GrLivArea have been selected along with the qualitative variable KitchenQual. The relationships between these variables will be explored further using scatter plots. Below are scatter plots showing the amount of above grade living area versus sale price and log(sale price) grouped by quality of the kitchen:



Here are scatter plots of total first floor sq. footage versus sale price and log(sale price) grouped by kitchen quality:





Ames Housing Data Analysis
Assignment #1
John Moderwell
MSDS 410

There appears to be a strong relationship between these variables, especially between first floor area and kitchen quality as there seems to be minimal overlap between each class of kitchen. There is also a strong positive correlation in relation to sale price, although properties with “Excellent” kitchens do not exhibit as much correlation with the amount of above grade living area and the total first floor area.

It is also apparent that a log transformation of the response variable has a moderate effect on reducing variation. This indicates that further transformations may be used in the future.

Summary/Conclusions:

- Exploratory data analysis is critical to building an accurate regression model. While much of this process is a science, there is also an art to it. Common sense and the use of prior knowledge about the data set should be utilized as it is not always clear what is the best decision to make.
- It is important that initial results are not taken at face value. It is important to investigate further to see if the results are really accurate. This was especially true in the selection of variables because the initial analysis may not be representative of what is actually having an effect and what is not.
- Data quality checks are critical aspects of this process. It is very important that the data being analyzed is complete and not full of errors. The sample extracted for analysis needs to be representative of the goals of the project. For example, the goal of this assignment is to predict the value of “typical houses” therefore industrial, commercial and agricultural buildings should not be included.
- Variable transformations show promise in reducing variation and making the data more linear but further analysis of the effect on predictor variables is required.

Ames Housing Data Analysis
Assignment #1
John Moderwell
MSDS 410

Appendix:

Data Quality Check R Code:

Generate report of missing values in variables

```
require(dlookr)
dataqualitycheck <- diagnose(newdata)
Dataqualitycheck
```

##Divide Variables into two types: Factor and Integers

```
newdataintegers <- newdata[,sapply(newdata, is.integer)]

newdatafactors <- newdata[, sapply(newdata,is.factor)]
```

##Remove missing values

```
newdatafactors <- na.omit(newdatafactors)
```

##Summarize each variable and check for possible outliers

```
newdataintegerssummary <- summary(newdataintegers)
newdataintegerssummary

newdatafactorssummary <- summary(newdatafactors)
newdatafactorssum
```