Assignment #1

Andrew Knight

Introduction:

Before we can begin to build statistical models, we will always need to get to know our data. The following report gives an overview of the first Exploratory Data Analysis of the Ames Housing Data. The goal is to review the dataset and learn as much as possible about the predictor variables before proceeding with the model building and regression analysis.

The sections that follow contain a basic survey of the data, a simple data quality check and initial EDA tasks. Relevant R output is provided where appropriate. This first assignment will become the basis for further data exploration in subsequent weeks which will contain the detailed analysis.

Data File: ames_housing_data.csv

Data Documentation: AmesHousingDataDocumentation.txt

Calculations File: Knight_Assignment1_code.R

Results:

Section 1: Sample Definition

This dataset includes the housing data from the Assessor's Office for Ames, IA. It contains a total 2930 samples each with 82 variables about houses, including most recent SalePrice. This is a snapshot of City of Ames Assessor's housing data between 2006 to 2010. The variables represent many aspects about each property. It contains qualitative and quantitative variables. Each variable definition gives data types and meaning for each categorical indicator.

The right data appears to be available and the variables are well defined as described in the companion documentation. The data do not appear to have missing values. Appropriate indicators like NA are used if that particular house attribute is not included; the numeric value 0 is used to indicate it does not exist for that numeric feature. This section will also cover qualification of records and drop conditions that will form the basis for sample data to be used in further analysis.

Given that our goal is to form a predictive model for sales price, the SalePrice variable seems like the best candidate for a response variable however a full investigation of the variables should still be done. We want to make sure the data are not skewed by property types that are dissimilar to the rest of the population. For example, we should look at outliers, especially on the high end to determine if they are the same type of residence. We should also choose predictor variable carefully and test hypotheses about which of them contribute significantly to

the price. Also, given that the description for the sale price data spans the years 2006 to 2010, we should consider when the selling price was obtained. When analyzing data as a subject matter expert, it is the job of the analyst to apply knowledge of the industry specifics to given data. Given knowledge that the housing crisis may have had some effect on the sale prices from 2008 through 2010, we may need to take that into consideration.

Section 2: Define the Sample Population

We will start with some assumptions about the housing market of Ames to assist with the properly staging the sample population. These assumptions include:

- Focus on residential (non-commercial, non-industrial, non-agricultural) properties. We should also focus on single family homes, condos or duplexes as opposed to multi-family apartment buildings. Unlike larger cities, most of the properties should exhibit low-density homes of this type however Section 2 should verify these assumptions. To keep early analysis simple, I am restricting the zoning to residential only, single family detached.

- Focus on standard sale conditions to avoid skewing estimates from things like foreclosure sales, short sales, partial or sales between family members. I also considered narrowing the sales types to exclude warranty deeds or court officer estate sale, however I opted to include all of these type for now because of lack of sufficient domain expertise. I don't want to start off affecting the data without really understanding each sale type and how it affects the overall price.

- Focus on home sale 'basics' of size, quality and location. As I perform the analysis I will be checking against common intuition that the size of the house and the quality of the house are primary drivers for the sale price. Location too is an important factor, however without local market knowledge about that provides a numeric ranking (ordinal) or historical price differentials between neighborhoods, this last piece may be difficult to incorporate into the analysis without further data.

Drop Conditions:

For now, I would choose to restrict to the sample to single family, residential zoned houses only. We would not want to include commercial property types or multi-family type houses in the same sample as single-family homes or condos. We would however want to include in our sample single family and condo units from each of the geographic areas within the city limits so as to not overfit our sample data. We should take a similar approach for other categorical variables which we choose to include. Arbitrarily excluding nominal conditional variables would negatively affect our model just like removing outliers without cause. Doing so may give the initial appearance of improving the model, but will mean the ultimate outcome is just plain wrong.

These Categories were Dropped:

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

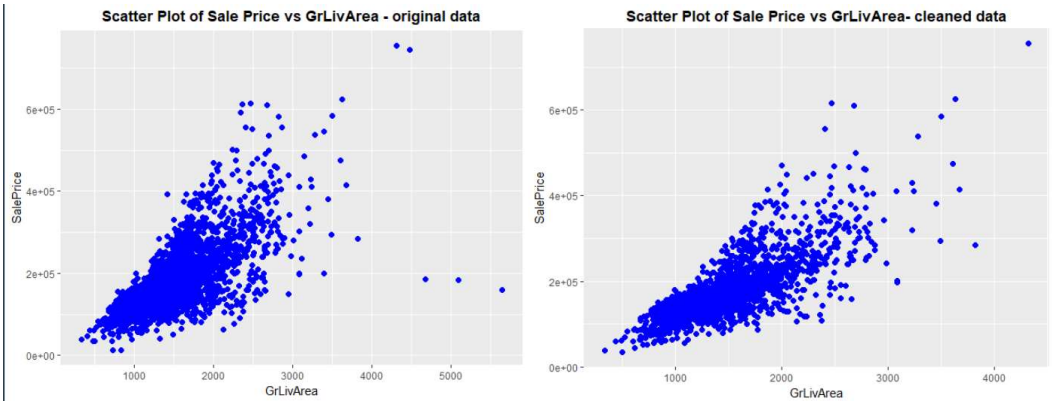
Remaining observations for the sample population.

Number of Rows: 1943, approximately 2/3 of the original observations remain.

Number of Cols: 84; Note: two new data fields were added, **TotalFloorSF** and **price_sqft**.

Section 3: Data Quality Check

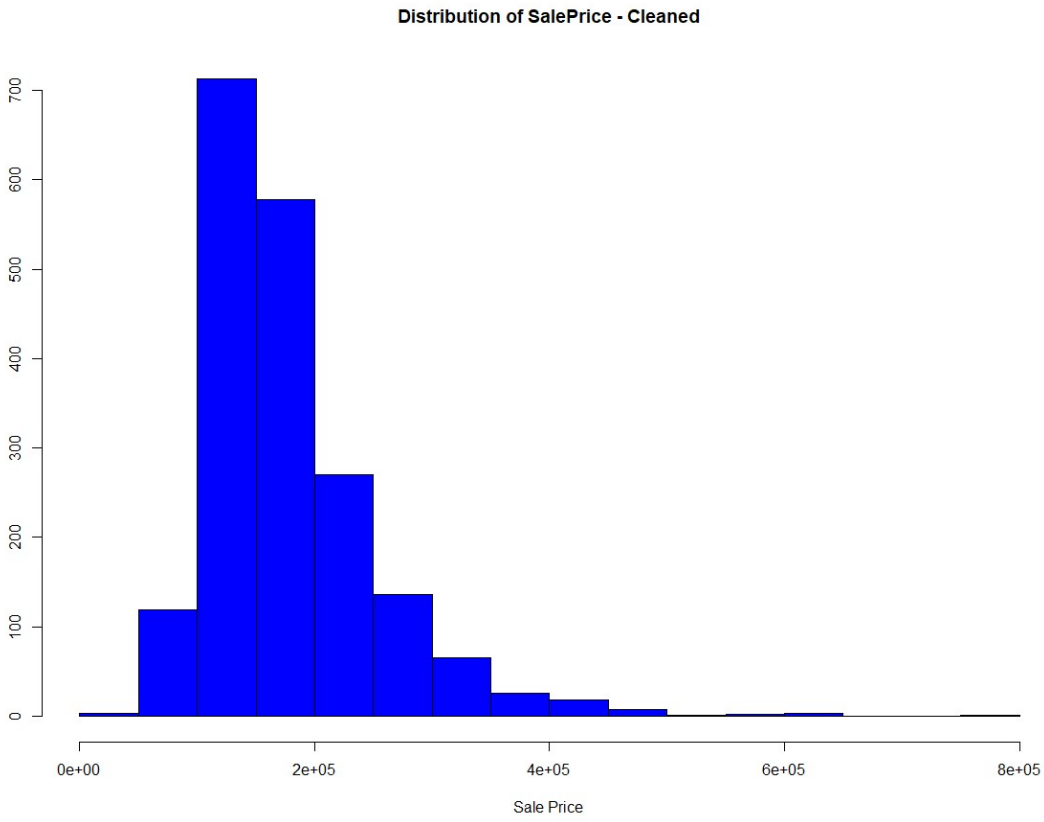
Outliers for predictor variables need to be evaluated closely before deciding which to remove. One example candidate for removing possible errors is in the mydata$GrLivArea which shows three values on the extreme end of the living area spectrum but towards the lower end of the sales price spectrum. These types of outliers require a closer look but perhaps should be removed. I noted that the outliers were not present in the cleaned data. Here is the before and after:



All value from the file are conveniently in a data frame format which can be subset for the various purposes. I created subsets for numerics, non-numerics, and by the number of predictor variables needed for each step.

Example 1: In this project you will be modeling the sales price of housing transactions. It should be obvious that none of these sales prices should be zero or negative. Observations with a zero or negative sales price should logically be considered to be errors.

I’ve verified that the minimum Sales price in this time range was $12,789 and the minimum price per square foot was $15.37, both of which seem reasonable. However, on the top end of the distribution, it may be necessary to remove some of the outliers. Here is the distribution of SalePrice for the cleaned data.



The twenty variables:

According to my initial strategy of focusing on common home value predictors, I considered variables that contributed to the 1. Size 2. Quality and 3. Location of the houses.

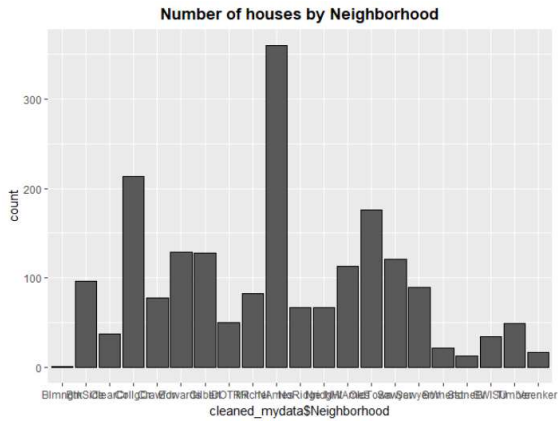
See Appendix A for the initial variable list of twenty.

Section 4: Initial Exploratory Data Analysis

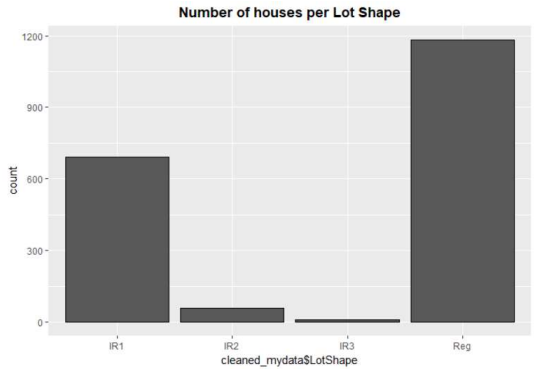
This has a mix of categorical (non-numeric), continuous and discrete variables, which I broke down into groups in code. I started looking at categorical variables like Neighborhood, LotShape, BldgType, KitchenQuality and even CentralAir to see if there were strong correlations overall quality, and thus price of the property.

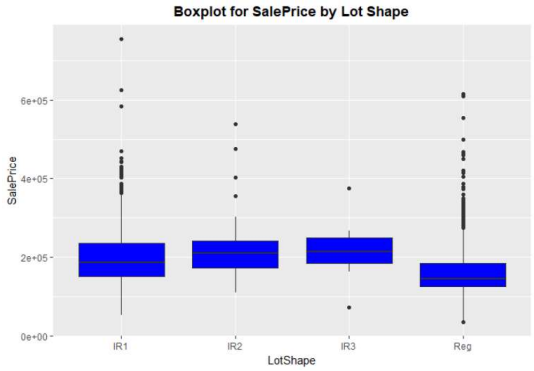
Here are some selected statistics & graphs tested:

- Neighborhoods plot showed correlations – more data may be needed to use this.

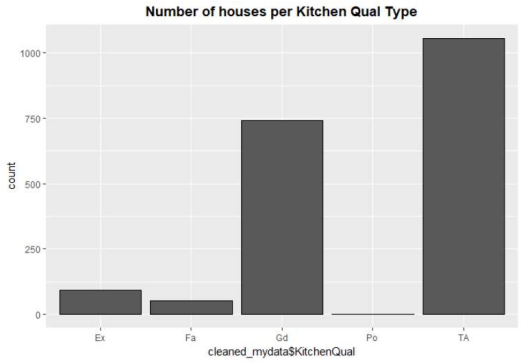


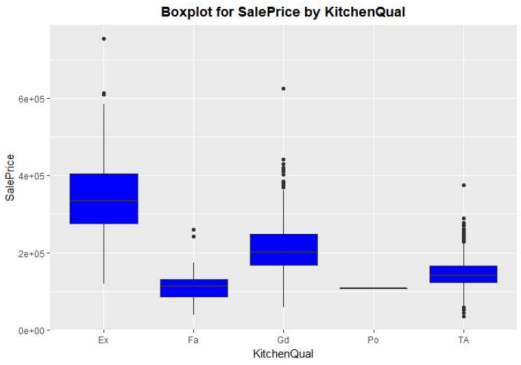
- LotShape: Regular shape homes make up 48%, IR1 make up 27% of sample (non-numeric).



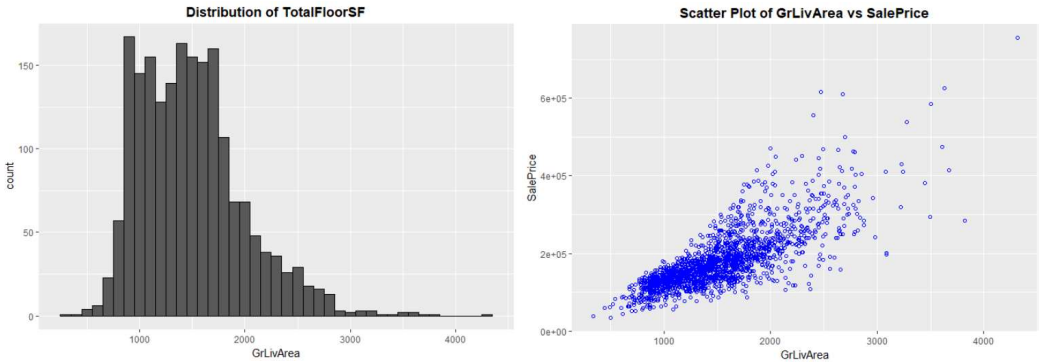


- KitchenQual: Rating for the quality of the kitchen in the home(non-numeric), 76% have a rating TA or better.

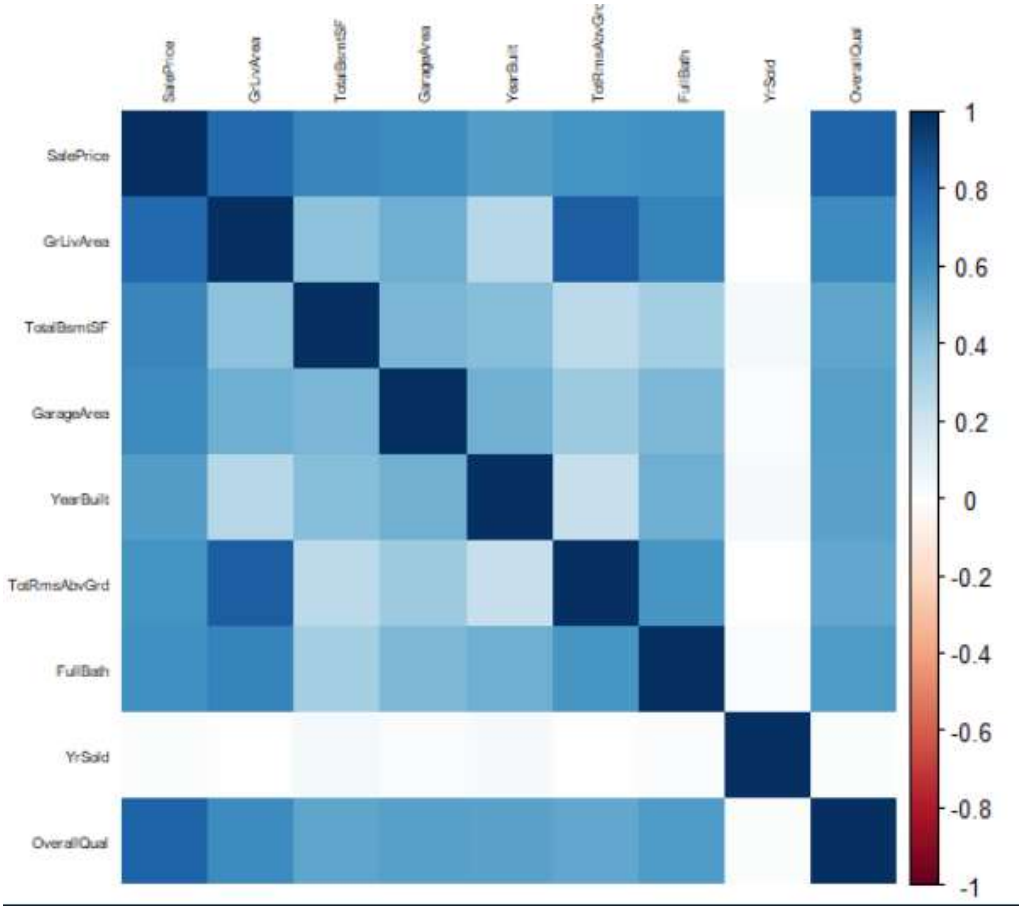




- GrLivArea: Total livable area SF above ground, approximately 1% of homes in cleaned data have more than 2850 total square feet above ground.



- Correlation matrix plot was done to confirm choices made based on correlation coefficient outputs when choosing variables for more narrow EDA in next section. Note that I have no variables which are negatively correlated.



The selected items above show differences in analysis for numerics (first and fourth) and non-numerics (second and third) in dealing with data. Insight can be gained from viewing both. Also, when performing bivariate analysis on continuous data (fourth) versus discrete (first) it is important to note limitations. First off, the discrete plots can offer some visual insight but will not give an accurate regression fit as the continuous. In my example, I've separated the two types for analysis. The str command or the dataset description can be used to identify variables.

When selecting from the initial twenty to narrow the list, I listed the min, median, mean, and max for each one that I selected to get an overview (for each numeric). All but one of the ten that I chose were numeric. See Appendix B below for the computed values.

I also did a basic correlation test to validate the positive correlation presented in the scatter plots. The Pearson Correlation Coefficient is listed for each numeric with an appropriate low, medium or high label. See table in Appendix B for these values.

Selected 10 variables and important attributes for Initial EDA in Appendix B.

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.

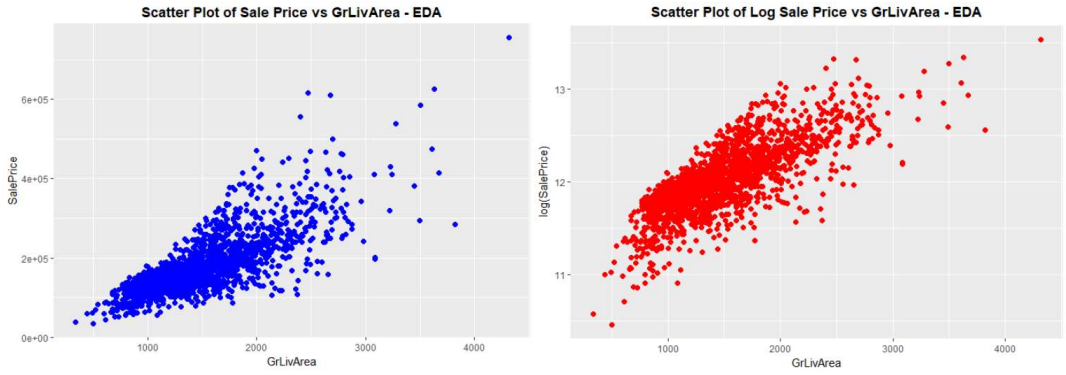
Section 5: Exploratory Data Analysis

After comparing possible response variables SalePrice and log(SalePrice) for the three variables under consideration, I choose to use the SalePrice rather than using at transformation to start.

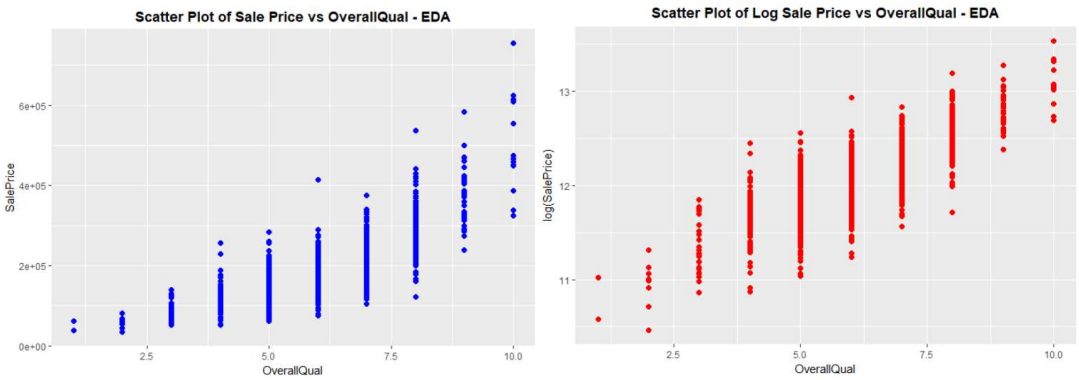
I ultimately chose to include the following three variables to complete the EDA: GrLivArea, TotalBsmtSF, and OverallQual due to their strength of relationship to sales price according to the correlation coefficient.

Comparing each for difference between SalePrice and log(SalePrice) for the two compelling input variables GrLivArea and OverallQual, we get:

1. GrLivArea

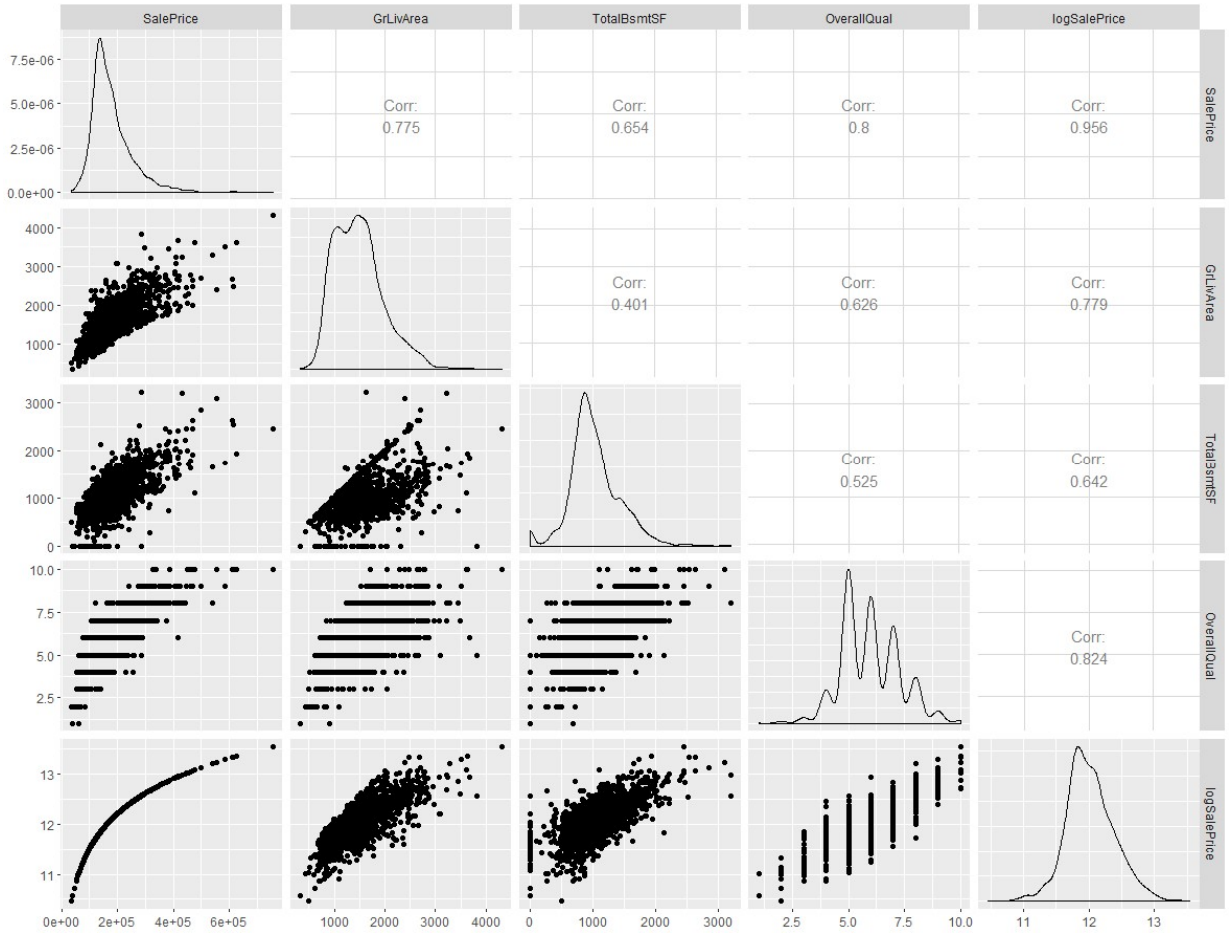


2. OverallQual



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 GrLivingArea for SalePrice and log(SalePrice) is 0.775 and 0.779 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 GrLivArea however the log transformation helps here too. On the whole the two predictor variables GrLivArea and OverallQual seem to give the best predictive performance.

Choosing the right model will depend on the appropriate selection of sample population. Also should a transformation actually be used, the model will have to account for that. It may be useful to consider other transformations to the cleaned data for the best model fit.

Conclusions:

In summary, this first EDA gives a broad overview of the housing data provided. However, more iterations are likely needed before we get a model with a high confidence of success. The relative high number of possible predictor variables in the data poses a challenge when deciding which to use. Also, given the disparity in the property and sale types, the EDA can be difficult. There are some NA and unknown values but this dataset is reasonably well structured. Trying to perform predictors for other geographic areas may not have the same depth of data. Other questions that could affect the validity of the analysis:

- Is the given time range sufficient for formulating a reliable model?
- If we want to use our final model in another city or state, is the data used representative of new area and is more data required?
- Should we be considering all property types in the Ames data? What is the predictive model being created for and where will it be used?

Other questions remain but the purpose of this first EDA has been to dive into the available data to understand it better before moving forward with analysis.

Appendix A: Twenty Variables for Data Quality Check

The list of twenty variable changed as I performed initial EDA.

Determine the list of twenty variables to test in the EDA, also include SalePrice as the response var

```
subdat <- subset(cleaned_mydata, select=c("SalePrice",
"TotalFloorSF","GrLivArea","BsmtFinSF1","TotalBsmtSF",
"LotArea","LotShape","LotFrontage","GarageArea","YearBuilt",
"BedroomAbvGr","TotRmsAbvGrd","FullBath","Neighborhood","CentralAir",
"YrSold","KitchenQual","OverallCond","OverallQual","YearRemodel"))

str(subdat)
```

Appendix B: Ten Variables for Initial EDA

The following are the 10 variables I selected for the Initial EDA.

Variable	Min	Median	Mean	Max
SalePrice	35000	160000	178464	755000
GrLivArea	334	1440	1491	4316
TotalBsmtSF	0	973	1032	3206
GarageArea	0	470	465	1488
YearBuilt	1872	1967	1967	2010
TotRmsAbvGrd	2	6	6	12
FullBath	0	1	2	3
YrSold	2006	2008	2008	2010
OverallQual	1	6	6	10
Neighborhood	Non-numeric	Non-numeric	Non-numeric	Non-numeric

Variable	Pearson Correlation Coeff to SalePrice	Strength of relationship
SalePrice	1	-
GrLivArea	0.775	High
TotalBsmtSF	0.654	High
GarageArea	0.636	Medium
YearBuilt	0.550	Medium
TotRmsAbvGrd	0.597	Medium
FullBath	0.602	Medium
YrSold	0.018	Low
OverallQual	0.800	High
Neighborhood	Non-numeric	Non-numeric

[end]