Scott Jue
MSDS 410
1/08/2023

Assignment #1: Getting to Know Your Data - Exploratory Data Analysis (EDA)

Introduction

The purpose of this exercise is to explore the Ames Housing dataset so that we can have a better understanding of the data prior to developing a linear regression model. Through this assignment we will gain a high-level overview of the variables contained in the dataset, as well as, assess the quality and limitations of the data. Additionally, we will be able to perform initial exploration on potential areas of interest to for regression modeling purposes.

The data we will be using comes from the Ames Assessor's Office and is used for computing assessed values for individual residential properties sold in Ames, Iowa from 2006 to 2010. In addition to the sale price data, the dataset contains many of the variables with data describing property characteristics. This presents an opportunity to develop a linear regression model that can accurately predict the future sale price of home properties in Ames, Iowa. However, prior to modeling we must first define the sample population as well as perform exploratory analysis on different independent variables to determine their quality and usefulness as potential predictors for the model. The objective of this model is to provide estimates of home values (i.e. predicted sales price) for typical homes in Ames, Iowa.

Results

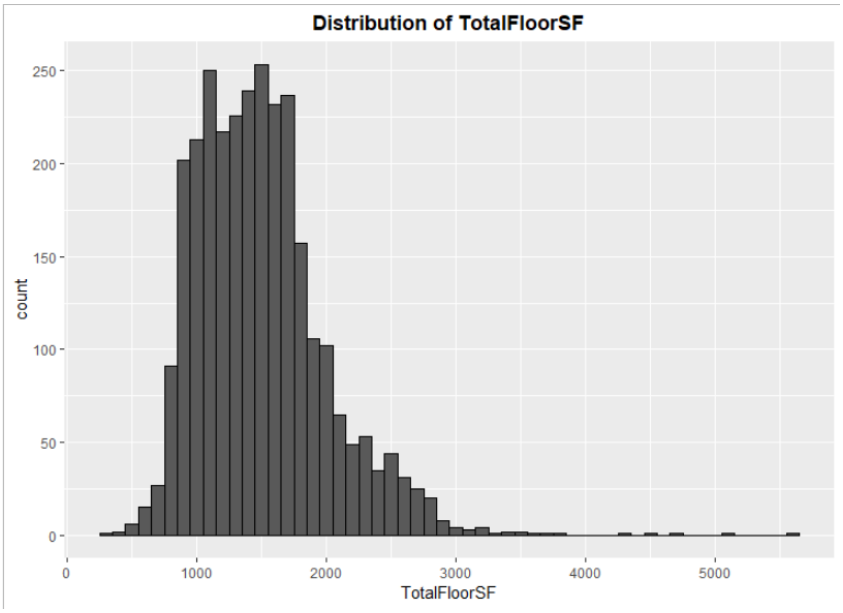
Task #1 Data Survey:

The Ames Housing dataset consists of 2,930 observations and 82 variables which represent residential property sales in Ames, Iowa from 2006 to 2010. For the types of variables, we have 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables (and 2 additional observation identifiers). There are several variables that are commonly known to influence home values present in the dataset such as home size (square footage), condition, age, neighborhood, etc. So, it appears we have the right type of data to construct a linear regression model that predicts the sale price of a typical home in Ames, IA. However, we have to be careful in the application of the regression model as this data is from 2006 to 2010; therefore, the analysis will only be relevant to that time period and for only homes around the Ames, IA area. As such, using a regression model developed using this data may have decreased accuracy if using input data from years outside of this time frame and/or housing data outside of Ames, IA.

Additionally, the variables present in the dataset also present an opportunity to create new variables using the existing data in the original dataset. By doing this can capture additional property characteristics by combining data from other variables. For example, I have created TotalFloorSF by adding the first-floor area and second-floor area, HouseAge by finding the difference between the year sold and year built, QualityIndex by multiplying overall quality with overall condition, logSalePrice by performing a log transformation on the sale price variable, and price_sqft by dividing the sale price by the TotalFloorSF variable. These new variables have been added to the dataset for use in future analysis.

While reading the data documentation, I noticed that it is suggested to exclude 5 observations from the original dataset. This is due to 3 of these observations being partial sales which cause them to

be outliers. And the other 2 observations are also considered to be outliers due to the their unusually large size compared to the rest of the houses in the dataset. To remove these 5 outliers, it is recommended to remove any observations with more than 4,000 sq. ft. When looking at the below histogram, we can see that there are 5 observations with TotalFloorSF greater than 4,000 sq. ft. These will be the observations we will be excluding from our analysis in order for these outliers to not influence our model.



Task #2 Define the Sample Population:

There are several subsets of residential property sales transactions within the original dataset, so it is important to isolate a sample population from the dataset. The objective of model is to predict the home values for typical homes in Ames, Iowa, therefore a sample of the original dataset was created by excluding certain observations. Below is the waterfall of filter conditions that were used to create a subset of the original dataset that more accurately represents a "typical" home.

1) Remove suggest outliers:
   Filter to only include observations where TotalFloorSF < 4000
2) Remove non-residential zones:
   Filter to only include observations where Zoning = "RH", "RL", "RP", or "RM"
3) Remove observations with abnormal utilities:
   Filter to only include observations where Utilities = "AllPub"
4) Remove observations non-single-family homes
   Filter to only include observations where BldgType = "1Fam"
5) Remove observations without a garage:
   Filter to only include observations where GarageCars >= 1
6) Remove observations with abnormal sales:
   Filter to only include observations where SaleCondition = "Normal"

After these filter conditions have been applied, we are left with a sample population of 1915 observations that approximates a sale of "typical" home in Ames, Iowa which is now defined as a normal

sales transaction of a single-family home less than 4,000 sq. ft. in a residential zone with all public utilities having at least a 1 car garage.

Task #3 Data Quality Check:

20 independent discrete and continuous variables were selected for further analysis to assess the quality of the data.

SalePrice	TotalFloorSF	HouseAge	QualityIndex	price_sqft
LotArea	BsmtinSF1	OverallQual	BedroomAbvGr	TotalBsmtSF
FullBath	GarageCars	TotRmsAbvGr	GarageArea	GrLivArea
Fireplaces	PoolArea	OpenPorchSF	ScreenPorch	EnclosedPorch

The table in Appendix A shows the max, min, mean, median, and standard deviations of these 20 variables. Several variables have a minimum value of 0 which could be valid for some variables such as PoolArea if the home did not have a pool or Fireplaces if there wasn't a fireplace. However, observations with variables such as BedroomAbvGr and FullBath that have a value of 0 would appear to be outliers and probably should be removed as they are not representative of a "typical" home. The table in Appendix B shows that there are no missing/null values in our sample population and therefore, we need to assume that the observations with a value of 0 for BedroomAbvGr and FullBath are valid outliers.

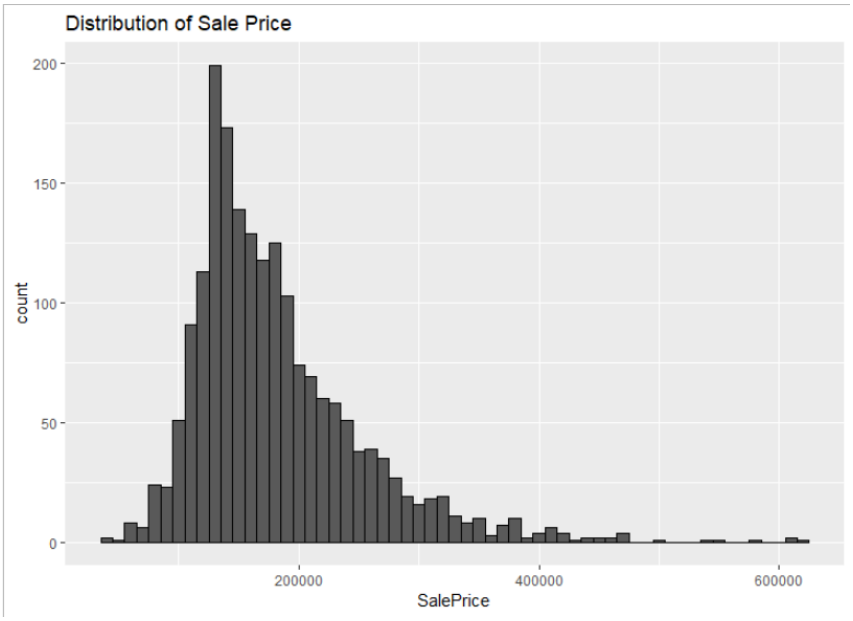
Next, I explored the amount of observations with 0 value variables to determine if any of the variables have a large portion of values with 0 and therefore, may not be meaningful enough to use for our analysis of "typical" homes. The table in Appendix C sows that the variables BsmtFinSF1, Fireplaces, OpenPorchSF, PoolArea, and ScreenPorch have a significant amount of 0 value observations. For this reason, I will remove these variables from the data subset that will be used for our analysis. Additionally, we can also see that are 4 observations with 0 BedroomAbvGr and 4 observations with 0 FullBath. Therefore, we can consider these outliers and I will also remove these observations from the subset.

Task #4 Initial Exploratory Data Analysis

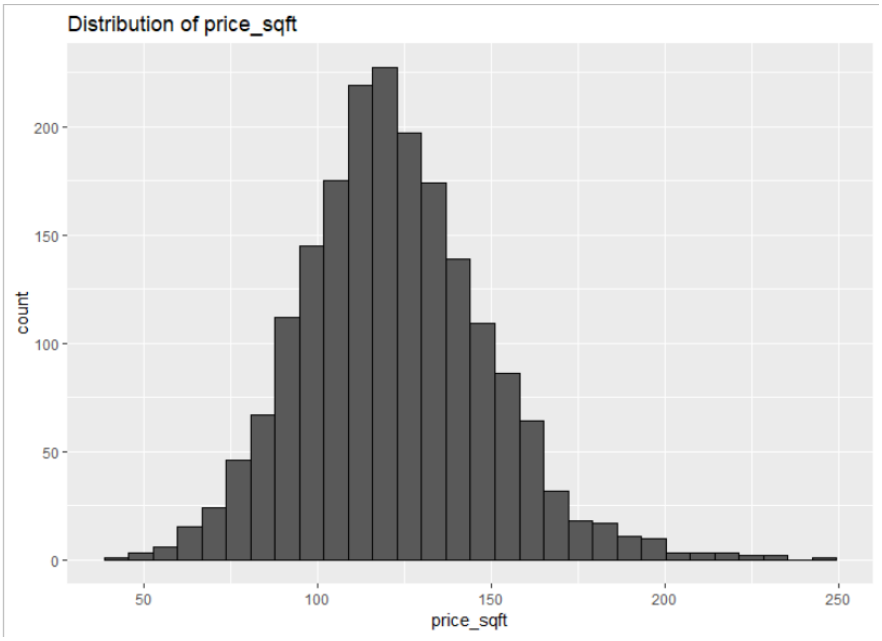
The initial exploratory analysis consists of looking at the relationships between the variables within the dataset. Therefore, I have selected the below 10 variables to explore further to gain more insight on how these variables might relate to each other.

TotalFloorSF	HouseAge	TotRmsAbvGrd	LotArea	OverallQual
TotalBsmtSF	FullBath	GarageCars	GrLivArea	Neighborhood

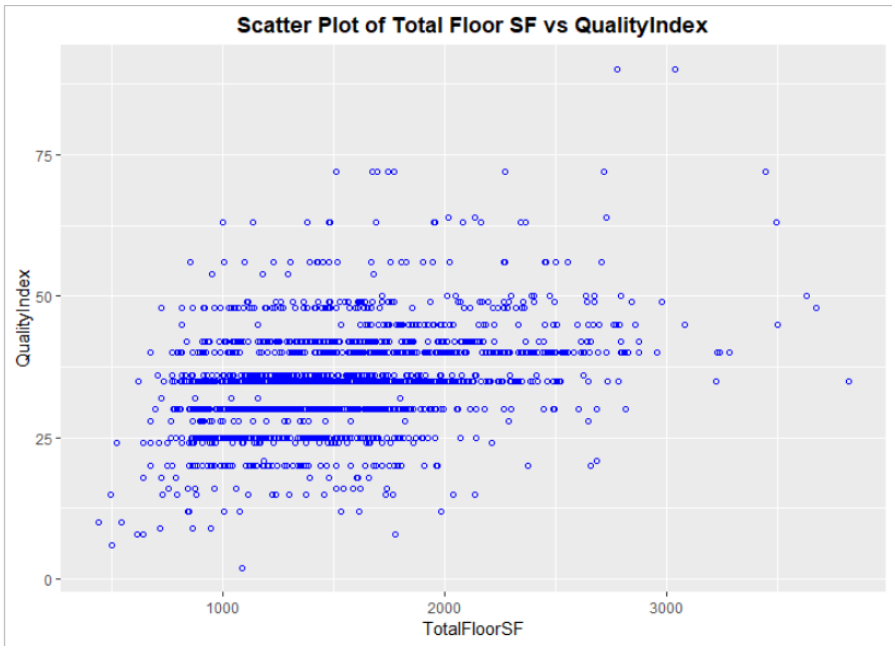
Before exploring these 10 variables, we fill first begin by looking the distribution of the response variable, SalePrice. Below is a histogram of sale prices which shows a right-skewed distribution and a mean of $182,172 and median of $165,000.



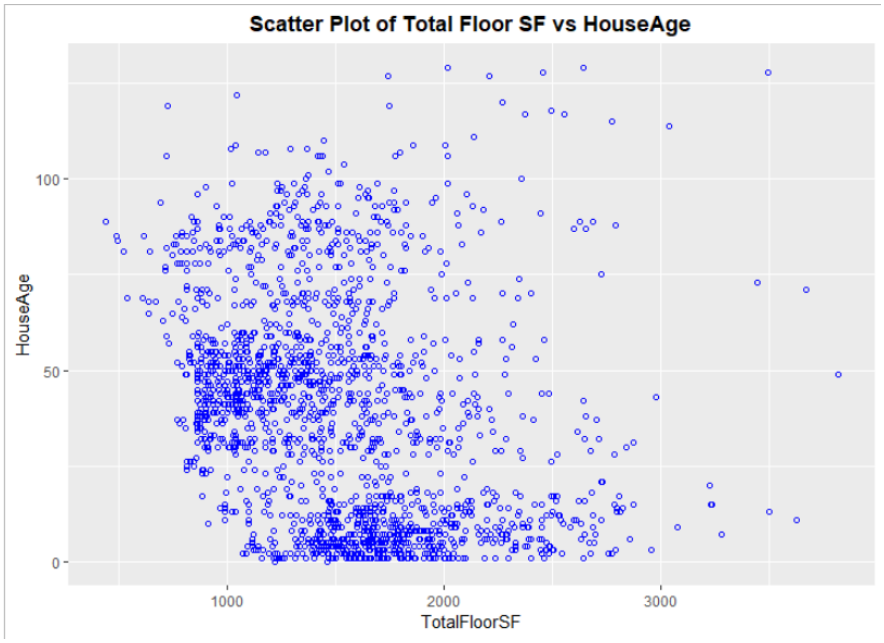
The below histogram shows a normal distribution of price per square foot for typical homes in Ames, IA during this time period. We can also see that the majority of these transactions had a sales price roughly equivalent to $125 per square foot.



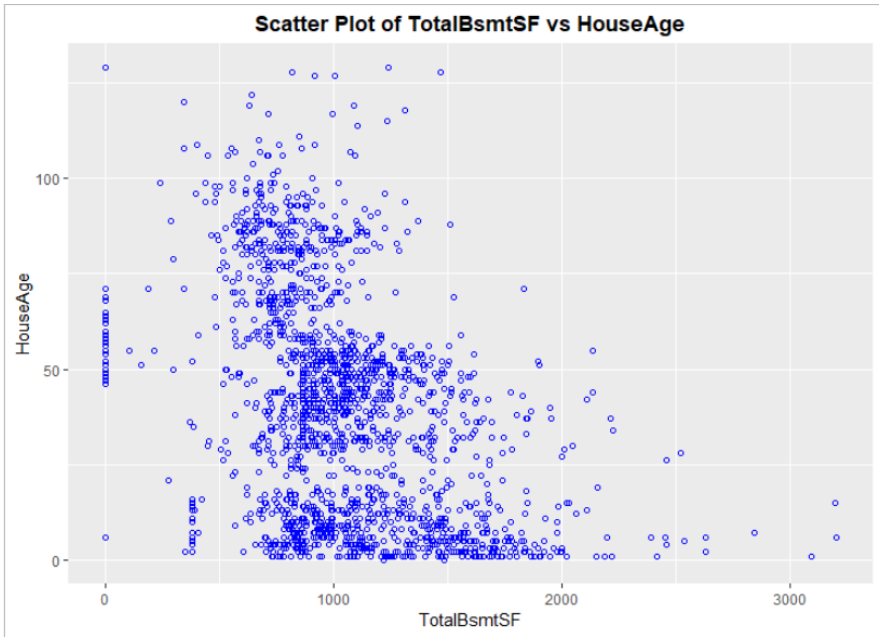
Then we will look at relationships between the different variables. First, we will look at the relationship between TotalFloorSF and QualityIndex in the scatterplot below. From this we can determine that homes with smaller TotalFloorSF tend to have a lower QualityIndex as compared to homes with larger TotalFloorSF



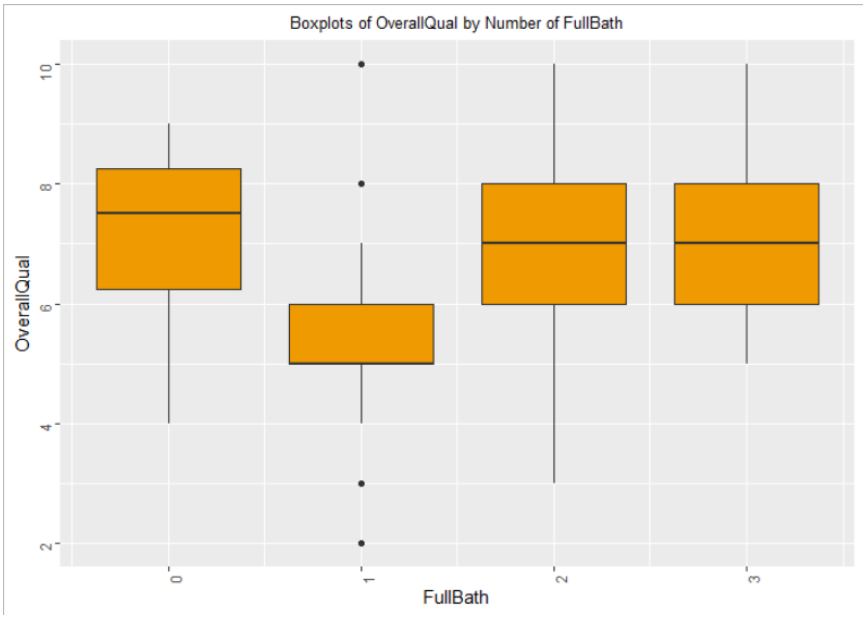
The next scatterplot shows TotalFloorSF vs HouseAge. We can see that newer homes tend to be larger than older homes; however, that is not always the case as a couple of the largest homes are over 50 years hold. Therefore, there it can be concluded that a strong relationship between the age of the house and the TotalFloorSF does not exist.



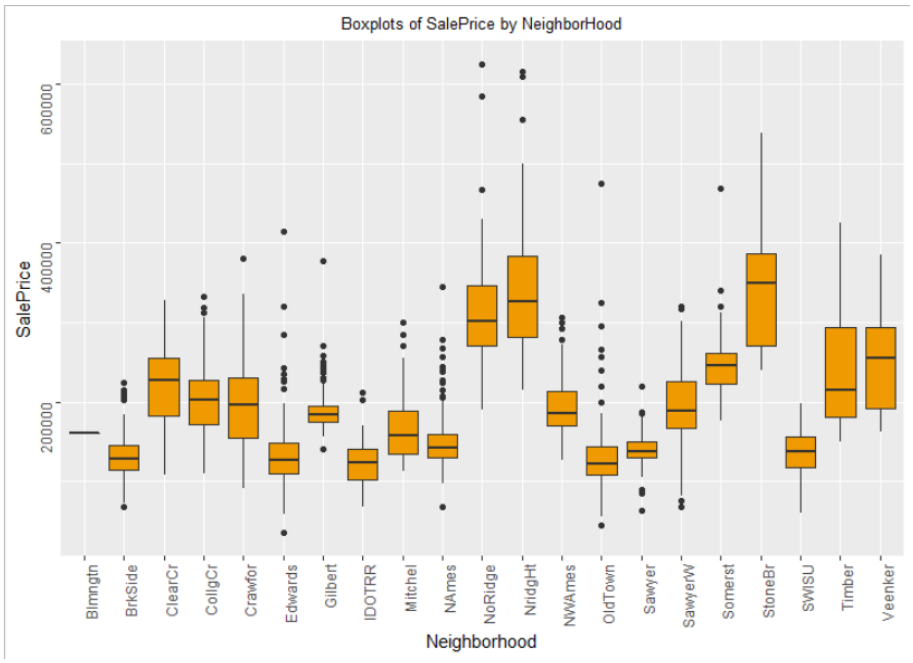
Next, we will look at the relationship between HouseAge and TotalBsmtSF which can be seen in the scatterplot below. Here we can see that the larger basements are associated with newer homes and that homes with 0 basement are were generally around 50 years old or more.



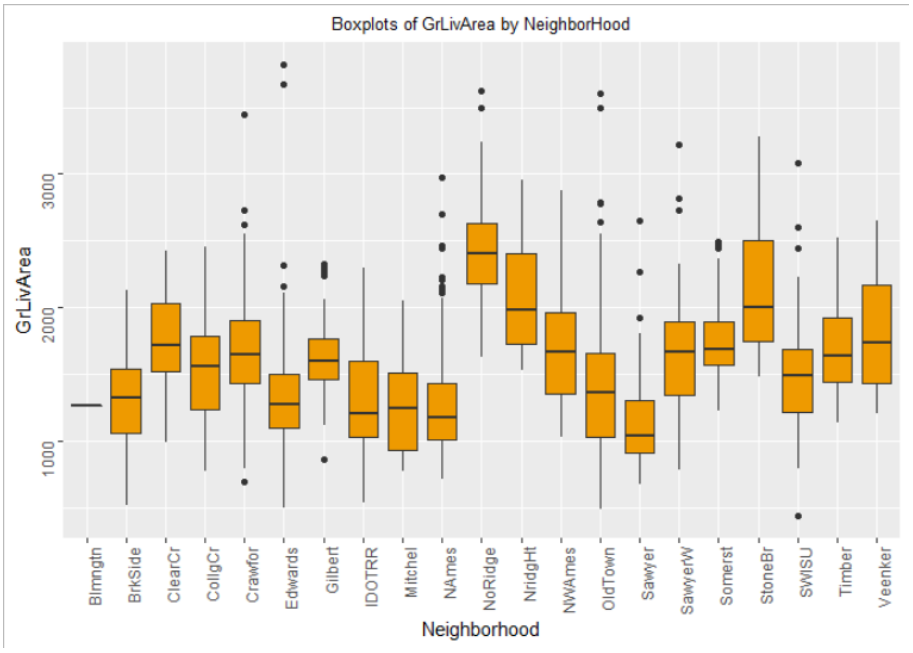
There does not appear to be any meaningful relationships between OverallQual and the number of FullBaths as shown by the below boxplots.



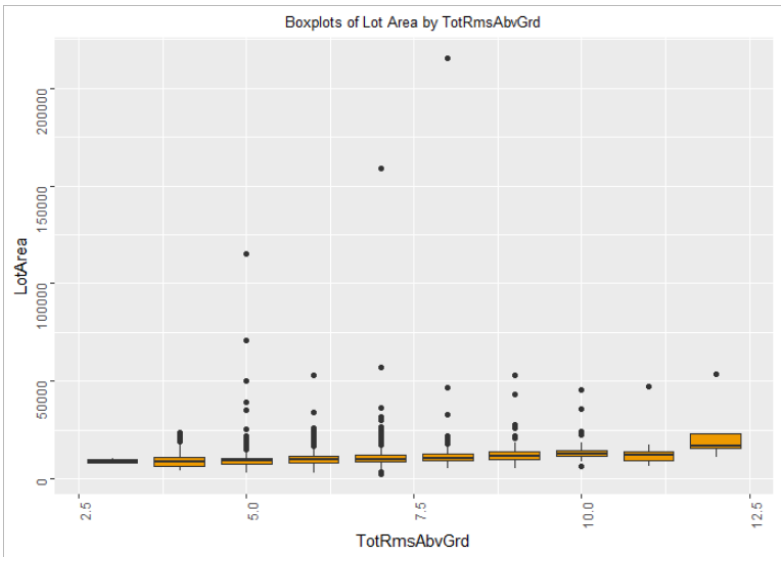
The next set of variables we will look at are Neighborhood in relation to SalePrice, which can be seen in the side-by-side box plot below. We can see that some neighborhoods have a higher average price than other neighborhoods. Northridge, Northridge Heights, and Stone Brook neighborhoods are the three neighborhoods that have the highest average home sale price.



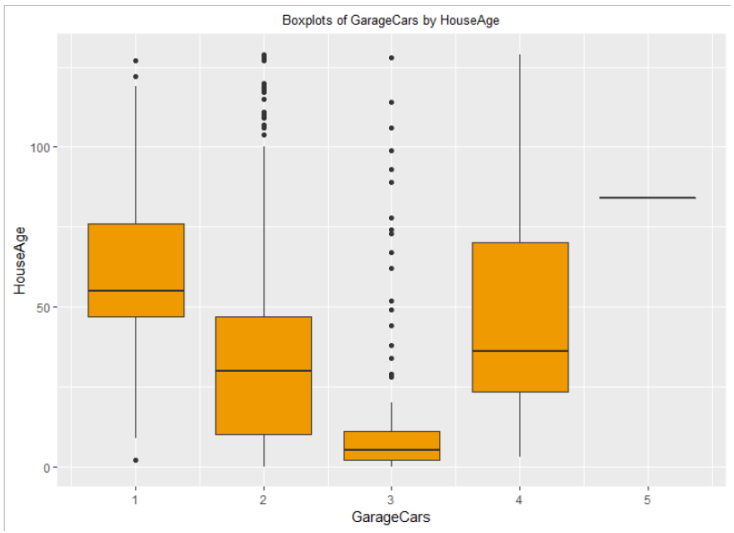
Additionally, I also looked at the relationship between Neighborhoods and GrLivArea which shows Northridge, Northridge Heights, and Stone Brook as having larger homes. This would make sense as these also are the higher priced homes as previously discussed.



Another relationship that I analyzed was LotArea by TotRmsAbvGrd. I was interested in seeing if the homes with bigger lots had more rooms. However, there does not seem to be a strong linear relationship between these variables as can be seen in the below plot. The homes with the three largest lots have roughly the average amount of rooms.



Lastly, we will look at the relationship of HouseAge and GarageCars. We can see in the below side-by-side boxplots that there does not appear to a strong relationship between these two variables. However, we can see that newer homes tend to have a 3-car garage while there was a bit more variability in the older homes.



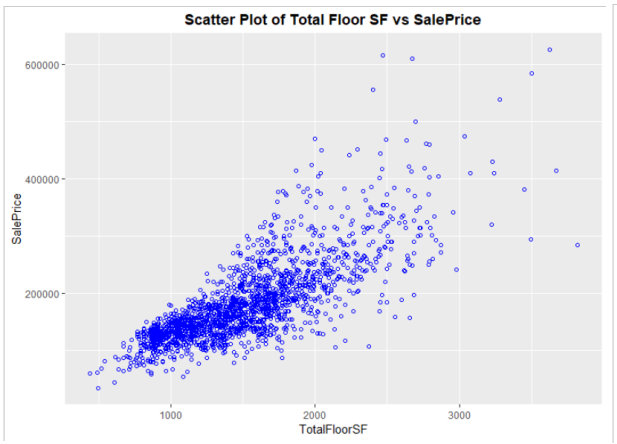
Task #5 Initial Exploratory Data Analysis for Modeling

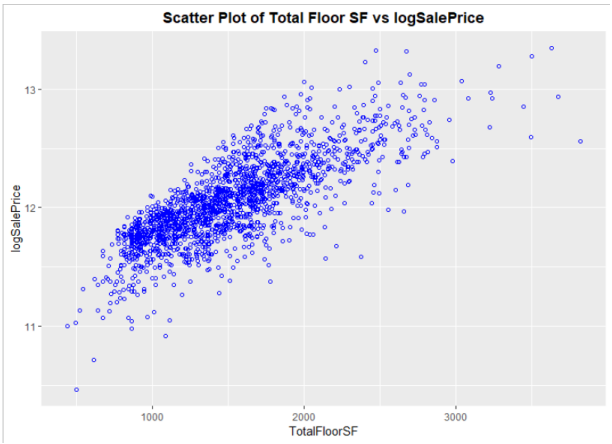
For regression modeling, the response variable for the Ames Housing data is SalePrice as our objective is to predict the sale price for the houses in Ames, Iowa. A potential concern discovered in the previous section is that there are some variables that share a relationship with others which could lead to covariance. Therefore, I will select 3 independent variables to explore the relationship with our response variable to see if they would appropriate to use as predictor variables in our linear regression model. These variables will be TotalFloorSF, TotRmsAbvGrd, and QualityIndex, which appear to show a positive correlation to the response variable.

Additionally, it was also discovered that the distribution of the SalePrice data is right-skewed, so in order to create a more normalized dataset we will also explore performing a logarithmic

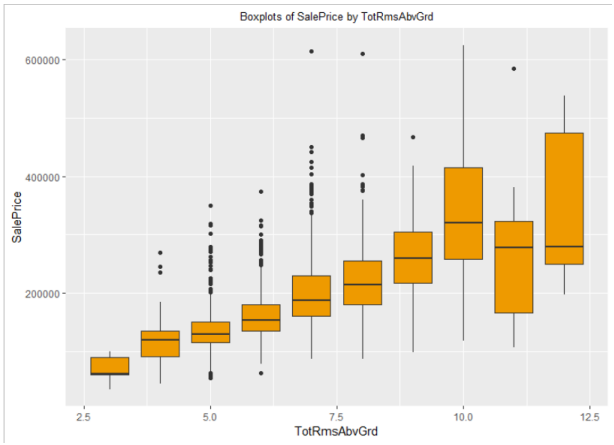
transformation on the SalePrice variable. Below are scatterplots and side-by-side boxplots that show the relationships between these 3 predictor variables vs SalePrice and logSalePrice.

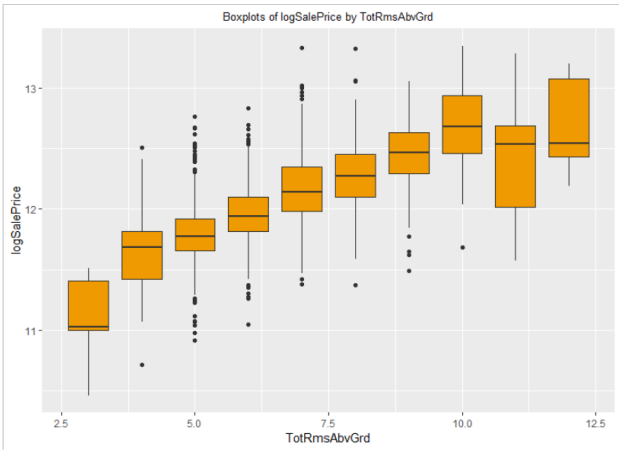
TotalFloorSF:



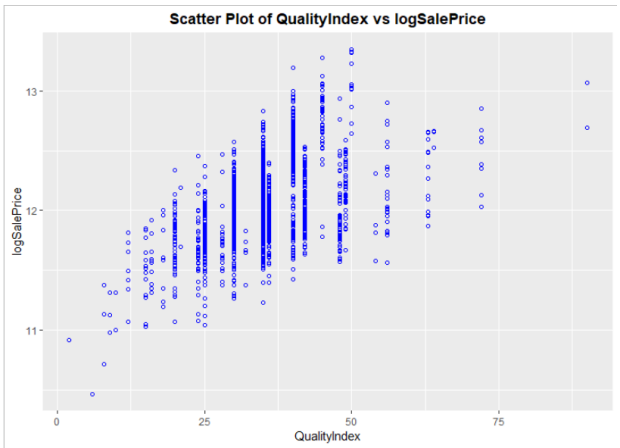


TotRmsAbvGrd:





QualityIndex:

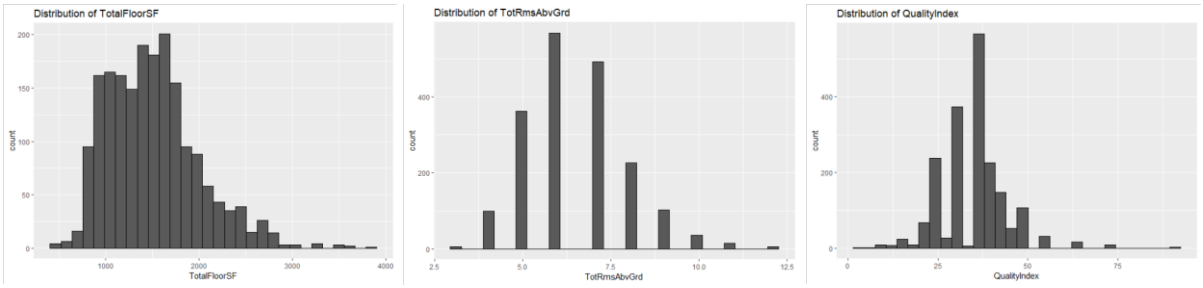


Correlation Coefficients:

	SalePrice	logSalePrice
TotalFloorSF	0.7740293	0.7826206
TotRmsAbvGrd	0.595299	0.6151584
QualityIndex	0.5089803	0.5382911

We can see that for all three variables there appears to be a positive correlation to SalesPrice. Larger homes, as denoted by home with more TotalFloorSF and more TotRmsAbvGrd, are correlated to higher sales price. Additionally, A higher QualityIndex score correlates to having a higher sales price as well. Furthermore, we can also see that using the logSalePrice variables results in less variability. Therefore, for modeling purposes we should plan to use logSalePrice as the response variable.

Below are histograms plots that show the distribution of the predictor variables. We can see that the distribution of these variables is roughly approximates a normal distribution, outside of a few outliers. Therefore, it does not appear that we should perform any transformations on these predictor variables.



Summary/Conclusions

During this EDA process of the Ames Housing Sales dataset, I was able to get a high-level overview of the different variables and how they relate to each other as well as the desired response variable, SalePrice. Additionally, through the data quality checks I was able to isolate poor quality data and observations that might be irrelevant to our analysis of "typical" homes. In completing the assigned tasks of this assignment, I was able to gain insight into the dataset so that I could make an informed selection for three predictor variables than can potentially be used to create a linear regression model for predicting the sale price. For modeling, there are a few areas of concern. First, the distribution of the response variable SalePrice is right-skewed. For this reason, I have performed a logarithmic transformation on this variable and created a new variable logSalePrice that alleviates the skewness. For the three potential predictor variables, these show to have a normal distribution; therefore, I do not recommend performing any transformations on these variables. Additionally, I have found that there are several outliers in the dataset when looking at the different variables. These are especially apparent in when looking at some of the boxplots. Given more time, I would have liked to investigate these outliers further to determine if they should be removed from this analysis and model building. In conclusion, the biggest takeaway I had from this assignment, is that we need to be careful to ensure that we are choosing a sample population with observations that accurately represent a "typical" home in Ames, IA during between 2006 and 2010. This requires us to understand that dataset so that we know what kinds of observations we are looking for in order to build an accurate model.

Appendix A: Summary Table of 20 Variables

key	Max	Mean	Med	Min	SD
BedroomAbvGr	5	2.93	3	0	1
BsmtFinSF1	2288	450.86	404	0	421
EnclosedPorch	1012	24.92	0	0	68
Fireplaces	4	0.65	1	0	1
FullBath	3	1.53	2	0	1
GarageArea	1488	486.55	480	100	182
GarageCars	5	1.81	2	1	1
GrLivArea	3820	1511.16	1466	438	488
HouseAge	129	39.03	39	0	29
logSalePrice	13	12.05	12	10	0
LotArea	215245	10899.28	9786	2500	7860
OpenPorchSF	570	46.99	26	0	65
OverallQual	10	6.06	6	2	1
PoolArea	800	2.24	0	0	35
price_sqft	249	122.11	120	45	27
QualityIndex	90	34.60	35	2	9
SalePrice	625000	182349.84	165000	35000	71543
ScreenPorch	576	18.48	0	0	61
TotalBsmtSF	3206	1044.65	985	0	397
TotalFloorSF	3820	1506.73	1464	438	486
TotRmsAbvGrd	12	6.48	6	3	1

Appendix B: Variables with null value observation

	key	NA Observations
1	BedroomAbvGr	0
2	BsmtFinSF1	0
3	Fireplaces	0
4	FullBath	0
5	GarageArea	0
6	GarageCars	0
7	GrLivArea	0
8	HouseAge	0
9	logSalePrice	0
10	LotArea	0
11	OpenPorchSF	0
12	OverallQual	0
13	PoolArea	0
14	price_sqft	0
15	QualityIndex	0
16	SalePrice	0
17	ScreenPorch	0
18	TotalBsmtSF	0
19	TotalFloorSF	0
20	TotRmsAbvGrd	0

Appendix C: Variables with zero value observations

Variable	Zero Obs	Zero Obs Percent
BedroomAbvGr	4	0%
BsmtFinSF1	548	30%
Fireplaces	845	45%
FullBath	4	0%
GarageArea	0	0%
GarageCars	0	0%
GrLivArea	0	0%
HouseAge	2	0%
logSalePrice	0	0%
LotArea	0	0%
OpenPorchSF	866	45%
OverallQual	0	0%
PoolArea	1906	99%
price_sqft	0	0%
QualityIndex	0	0%
SalePrice	0	0%
ScreenPorch	1727	90%
TotalBsmtSF	39	3%
TotalFloorSF	0	0%
TotRmsAbvGrd	0	0%