Assignment #1
Syamala Srinivasan

Introduction:

This assignment involves many of the initial steps of conducting an analysis. Understanding the data-set and ensuring that it is clean of any errors is essential. Another fundamental factor of beginning an analysis is ensuring that the data utilized is the correct data needed. In any data-set there will be populations of data that are fundamentally different from the others and must be removed before analysis is conducted to ensure that similar populations of data are being analyzed and results are correct.

The overall objective of the Ames assignment is to build predictive model for sale price of a home. There are many factors within the Ames data-set that can be used as predictive variable to this equation such as lot size, overall quality and year built. In this assignment variables such as these and their relationship to the sale price of homes in Ames will be analyzed to determine while variables have the strongest relationship with sale price.

Results:

Section 1: Sample Definition

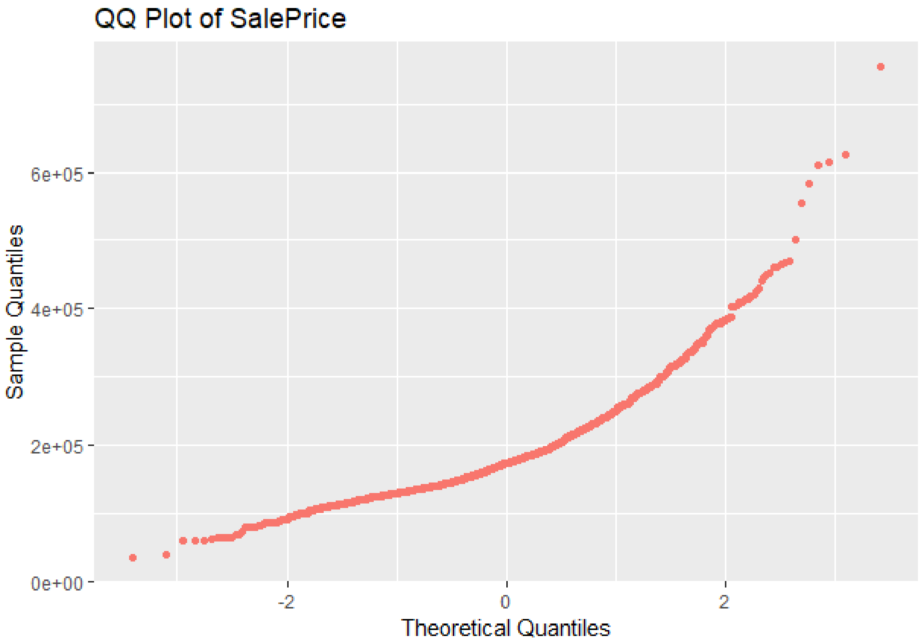
Task 1:

The data-set to be analyzed contains housing data detailing the various figures and variables nearly 3,000 residential property sales in Ames, Iowa that occurred between 2006 and 2010. This large data-set includes 87 variables including 44 numeric variables and 43 factors. The data contained in this data-set represents individual aspects of each residential property sale and specifics on the home itself.

The dataset includes a large amount of descriptive qualities for each home. These variables will be extremely valuable to conducting linear regression models. Assessing the relationships between these variables and the sale price of the homes will lend to detailed linear regression models and outputs.

There are some observation that should be removed. While there are many outliers in the dataset regarding sale price, lot area, total square footage, etc. these outliers cannot be removed without careful understanding. Additionally the data must be cleaned to ensure that it does not contain any incorrect data or data that may skew the results to an incorrect conclusion. Unclean data may include negative values, blank values, and data that can be determined to be entered incorrectly.

Caution will need to be applied when building a regression model for SalePrice. There are many attributes that we could value such as number of rooms, lot size, etc. however there may be a more complex relationship that would need to be uncovered.



The above QQ graph of the sale price of Ames homes illustrates that the sale price is not normally distributed. This is further confirmed when looking at the skewness. The skewness of sale price is 1.82. When utilizing the log of the sale price instead, the sale price is significantly closer to a normal distribution with a skewness of .27.

Task 2:

The objective of this task is to provide estimates of home values for typical homes in Ames. A "typical" home in Ames would logically only include homes. Essentially this definition of "home" excludes buildings identified as "PUDs", Duplexes, Agriculture, Commercial lots, etc. The removal of these populations ensures that future trends and analysis are comparing similar sample populations.

The following populations have been removed from the data-set:

Variable Name	**Populations Dropping**	**Drop Count**	**Cumulative Drop**
Zoning	A, C, FV, I, RH, RP, RM	657	657
Building Type	2fmCon, Duplex, Twnhs, TwnhsE	505	921
Functionality	Maj1, Maj2, Min1, Min2, Mod, Sal, Sev	202	1063
SubClass	90,120,150,160,180	448	1068
SaleCondition	Abnorma, AdjLand, Alloca, Family, Partial	517	1381

This results in a total population of 1,549 typical Ames homes.

Section 2: Data Quality Check

Task 3:

A data quality check was undertaken to ensure that homes that indicated a discrepancy were highlighted. These discrepancies include simultaneous mentions of having one and two stories, remodel dates earlier than initial build dates, blank values, etc. Each value in the data set was also checked to ensure that there were no blank or negative values. The twenty variables selected consist of multiple data types that are believed to have a relationship with sale price.

The following variables will be used for the data quality check:

Variable Name	**Variable Type**	**Description**
SalePrice	Continuous	The price of the home
HouseStyle	Nominal	Style of dwelling
TotalFloorSF	Continuous	The first floor square footage plus the second floor square footage
Overall Quality	Ordinal	Rates the overall material and finish of the house
Overall Condition	Ordinal	Rates the overall condition of the house
Year Built	Discrete	Original construction date
Year Remod/Add	Discrete	Remodel date
2nd Flr SF	Continuous	Second floor square feet
Garage Finish	Ordinal	Interior finish of garage
Garage Cars	Discrete	Size of garage in car capacity
SubClass	Nominal	Type of dwelling
TotRmsAbvGrd	Discrete	Total room above ground
TotalBsmtSF	Continuous	Total square feet of basement area
MaVnrType	Nominal	Masonry veneer type
MasVrArea	Continuous	Masonry veneer area in square feet
LotFrontage	Continuous	Linear feet of street connected to property
LotArea	Continuous	Lot size in square feet
Fireplaces	Discrete	Number of fireplaces
MicFeature	Nominal	Miscellaneous feature not covered in categories
MoSold	Discrete	Month Sold (MM)

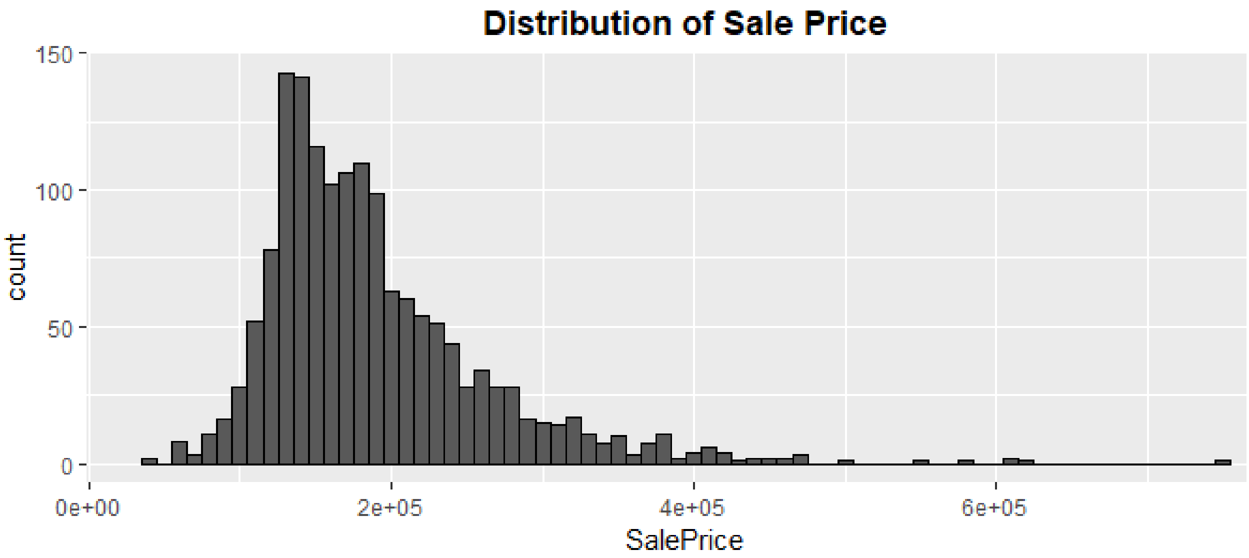
The results of the data quality check are located in the appendix.

Section 3: Initial Exploratory Data Analysis

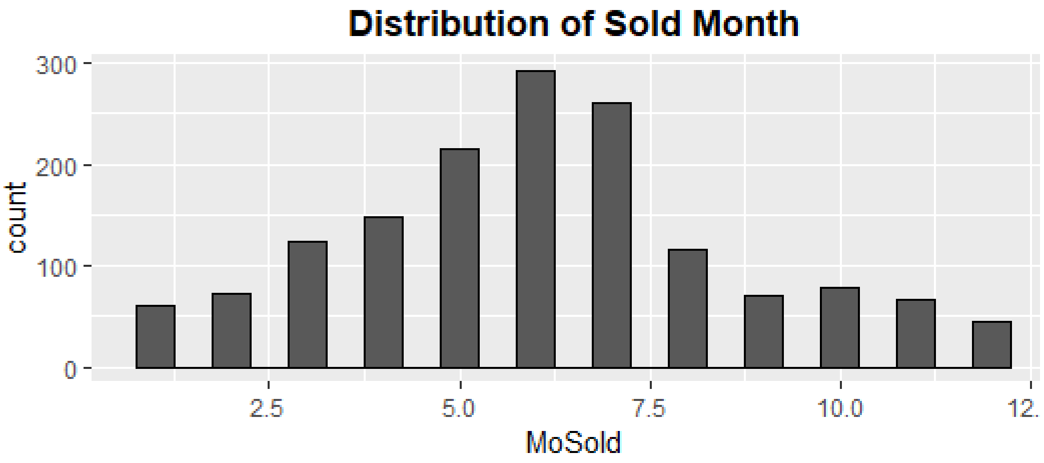
Task 4:

The 10 variables selected for an initial exploratory data analysis include:

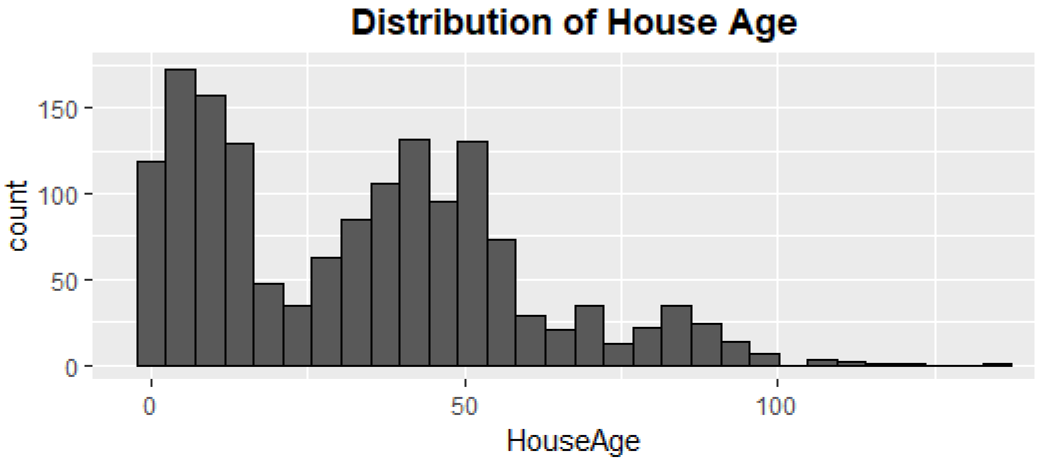
Variable Name	**Variable Type**	**Description**	**Range**	**Mean**
SalePrice	Continuous	The price of the home	$35,000 – $755,000	$189,002
HouseStyle	Nominal	Style of dwelling	N/A	
TotalFloorSF	Continuous	The first floor square footage plus the second floor square footage	334 – 4,316	1,503
Overall Quality	Ordinal	Rates the overall material and finish of the house	1-10	6
Overall Condition	Ordinal	Rates the overall condition of the house	2-9	6
Year Built	Discrete	Original construction date	1875 - 2010	1975
Year Remod/Add	Discrete	Remodel date	1950 - 2010	1985
Garage Cars	Discrete	Size of garage in car capacity	0-4	2
Exterior Quality	Ordinal	Evaluates the present condition of the material on the exterior	N/A	N/A
TotalBsmtSF	Continuous	Total square feet of basement area	0 - 3,206	1,082



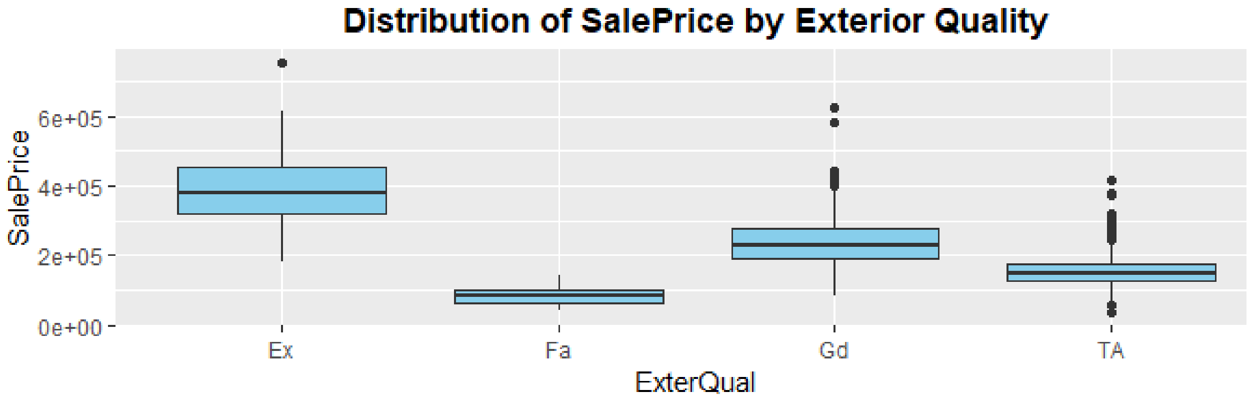
The average price of a home that sold in Ames in the timeframe sampled is $189,002. There are a few outliers that exist when viewing the distribution of prices. When viewing homes that sold for more than $400,000 there were a few details that stood out. 81% of these outlier homes are located in only 2 neighborhoods, Northridge and Northridge Heights. All homes located in Northridge Heights sold for between $200,000 and $700,000 while homes sold in Northridge average between $200,000 and $400,000. The outliers noted in the above graph also appear to be outliers for Northridge.



Roughly 40% of all sales in Ames occurred within June and July. This is consistent across all years. There is no indication that homes sell for more or less during these months. In fact, homes only sold for 3% greater than average during June and 1% less during July.

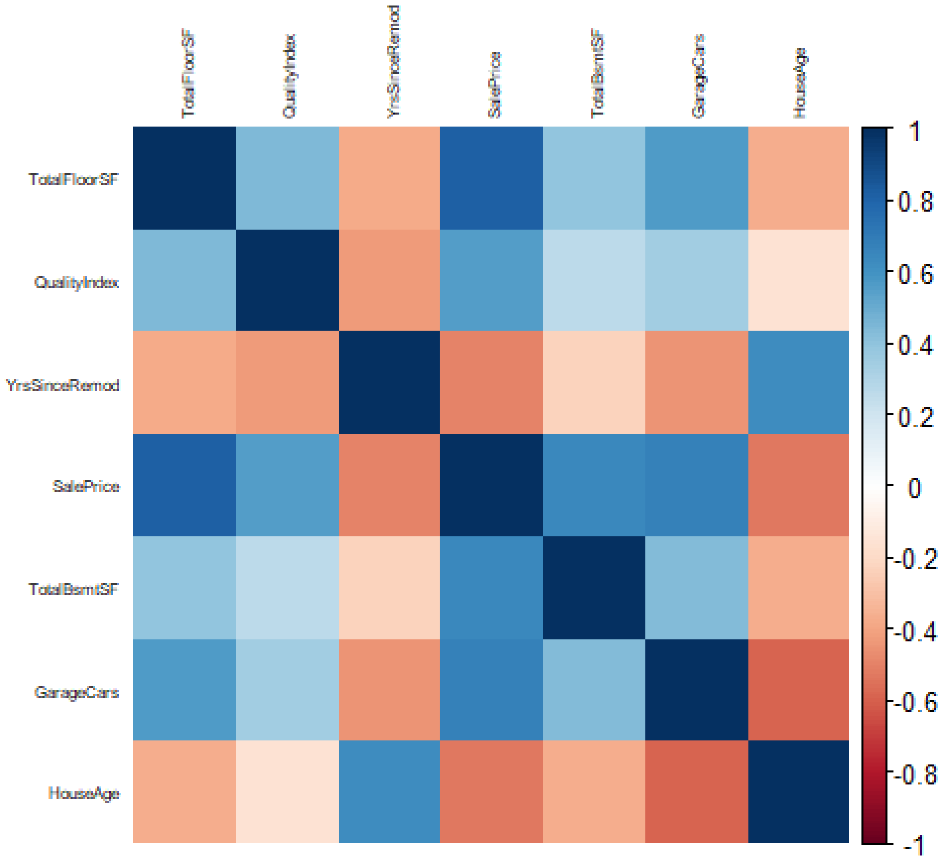


Homes selling in Ames are relatively new. If this sample population is taken to be a reflection of the population of homes in Ames, it can be determined that there have been two recent spikes in homes in Ames in the past, one roughly 50 years ago and one ongoing currently.



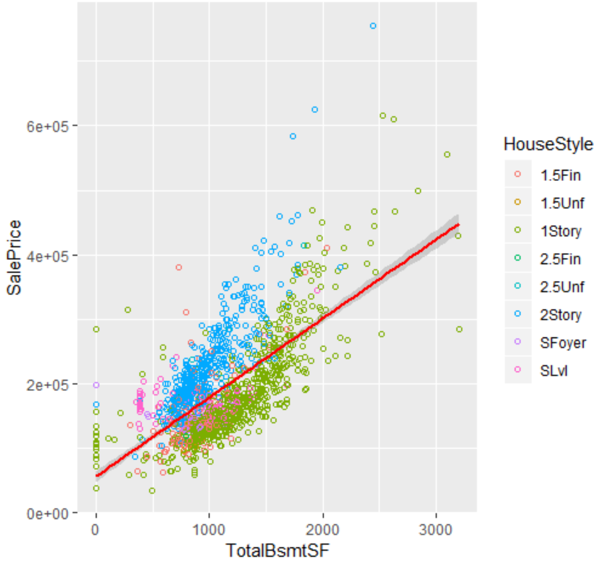
The above graph evaluates how the various qualities of the exterior of the house affects the sale price of a house. Curb appeal is a very important part of selling a house and it would be expected that having a higher quality exterior would lend to a more expensive sale. Utilizing the present attributes to attempt to determine what qualifies as “excellent”, it is notable that almost 50% of excellent homes have a stone masonry veneer type while only 12% of other homes have this veneer. Furthermore, 54% of excellent homes have between 200 and 500 square feet of masonry veneer while other homes typically have less than 300 square feet. Lastly, exterior quality also appears to also be driven from the roof type. Excellent homes are roughly 70% hip roofs while 76% of other homes are gable roofs. Therefore, it can be determined that the quality of the exterior of a home, and specifically the type of veneer, directly relate to the sale price of a home.

Correlation of Home Attributes to Sale Price



When the above chart is utilized to analyze the affect of variables to the sale price of the home many conclusions can be drawn. Of the figures compared, there are two figures that have a negative relationship with the price of a home. These attributes are the age of the home and the years between remodeling and selling. The moderate relationship (correlation coefficient of -.53) between house age and sale price and (correlation coefficient of -.50) between years since remodel and sale price, indicate that older or the longer since a remodel, the lower the price of the home will be. This point is further proven by viewing the relationship between the Quality Index and the years since the last remodel. The correlation coefficient between Quality Index and years since remodeling is -.42. This indicates a moderate negative relationship, the longer since a home has been remodeled the lower the quality of the home.

Total Basement Square Feet by Sale Price and House Style

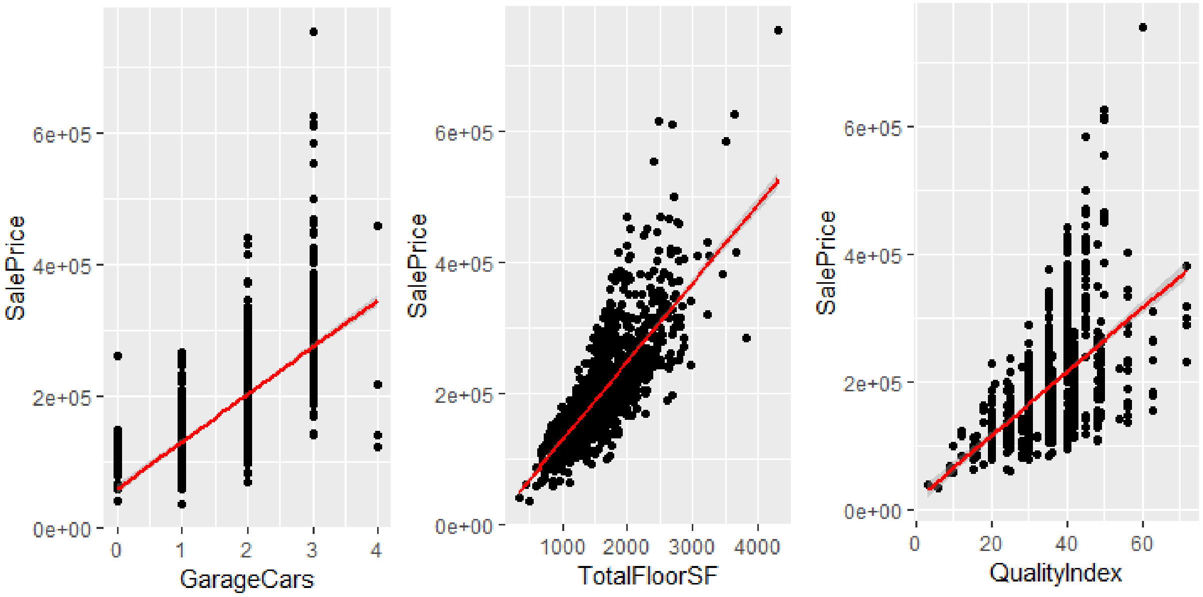


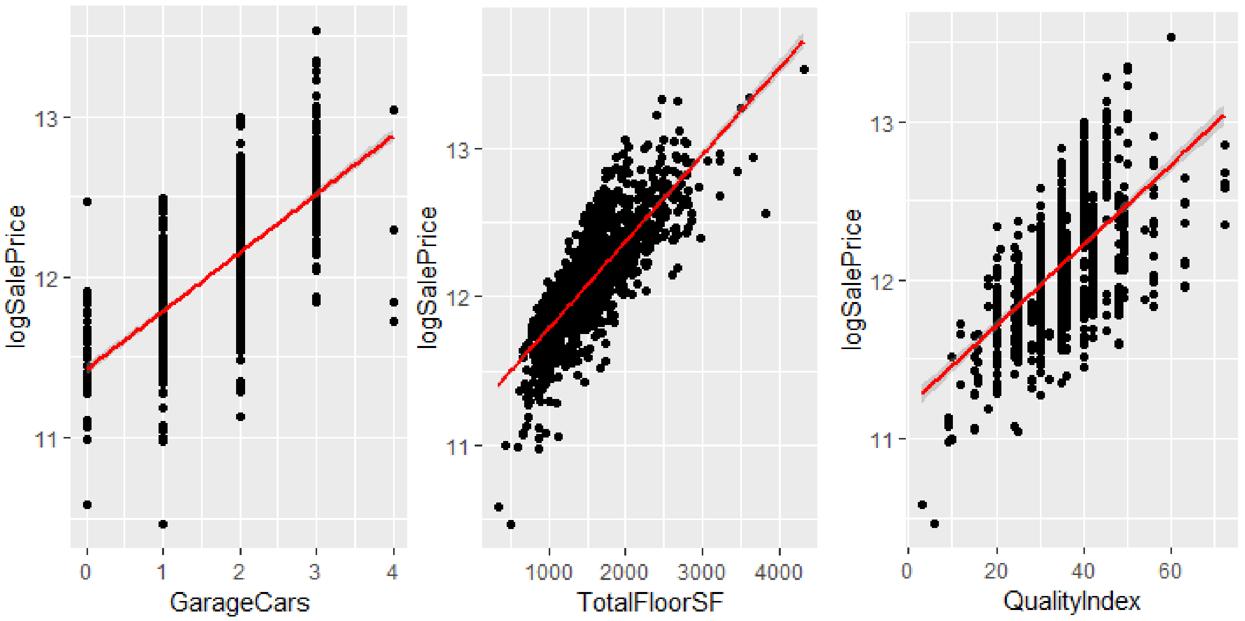
The chart on the left illustrates the price of homes compared to the total square footage of the basement and the style of the home. As evident by the regression line, there is a positive relationship among basement size and the price of a home. The correlation coefficient between sale price and total basement square feet is .65, this is a high positive trend that contains a few outliers. It can also be further proven with this graph that homes of similar size sell for more if there is a second story. The 2 story and 1 story homes appear to be split along the regression line. This indicates that homes with the same basement size sell for more if they have 2 stories compared to 1.

Section 4: Exploratory Data Analysis for Modeling

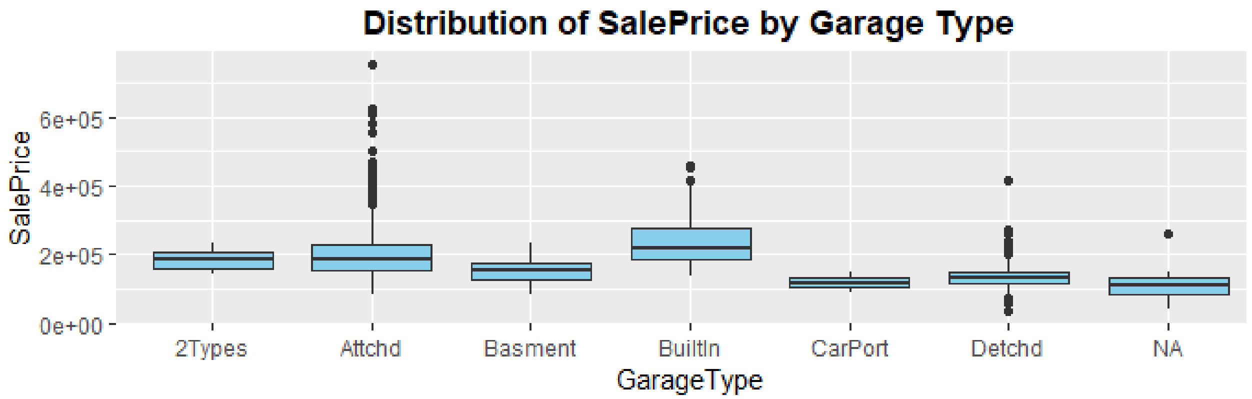
Task 5:

The three variables selected for this section will be GarageCars, TotalFloorSF and Quality Index.





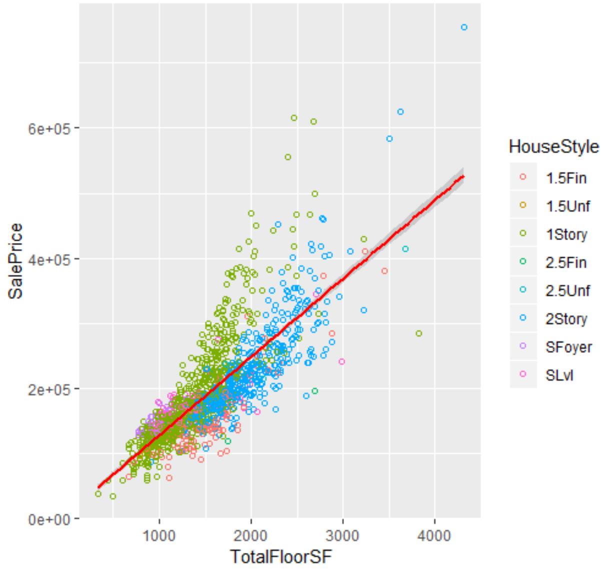
Detailed above are three predictor variables suspected have a strong relationship with sale price. Log transformation is utilized in this section to assist with confirming to normality and to assist with visualizing the large span of sale prices.



There is a strong positive correlation between the car capacity of a garage and the sale price of the home and an even larger correlation when utilizing the log of the sale price. These correlations are .68 and .71 respectively. As mentioned previously there are other aspects of the garage that influence the sale price in addition to the car capacity. For example, while a detached garage is less valuable to a potential buyer, a detached garage that will fit all your cars has an increased value. These two variables combined play a large role in estimating the price of a home. The size of the garage in terms of car capacity has a high positive relationship with the sale price of a house. In fact, houses with one garage tend to sell for an average of $25,000 more than homes with no garage. The type of garage is also a factor to the price of a home. Homes with detached garages sell for $40,000 less on average than homes with other garage styles while homes with an attached garage sell for $70,000 more on average than homes with other garage styles. The largest difference in price is evident when viewing the Tukey HSD of the garage types. The largest difference exists between car ports and Built-in garages. This difference totals to a $120,237 smaller sale for homes with car ports.

Utilizing the log of the sale price presents a slightly less normally distributed relationship between sale price and total square footage of a home, in fact the skewness only increases from .81 to .83.

Distribution of Floor Square Footage to Sale Price by House Style



The chart on the left illustrates the relationship between sale price, floor space, and the style of the house. It is evident by the regression line that there is a positive relationship between sale price and total floor space. Also evident on this graph while the number of floors has an impact on price, it is not as drastic as floor space. Until total floor square feet reaches roughly 2,000 square feet, the 2 story homes sold for the same or less than one story homes of the same size. This makes practical sense as a small home with multiple stories could present a complicated home to navigate and may have a tougher time appealing to a buyer resulting in a lower price.

There is a moderate correlation between the quality index and the sale price of a home. The correlation coefficient is .56 but after conducting the log transformation of sale price the correlation coefficient is increased to .59. The graphs above illustrate the relationship between quality and price. In the log transformed graph it can be deciphered that homes lower on the quality index scale rarely are priced above the regression line.

Conclusions:

There are many factors to be considered when attempting to predict the price of a home. Unfortunately, one of the largest and most fundamental aspects is the preference and requirements by the potential homebuyers that may result in a higher or lower price and this attribute is difficult to quantify. While this does pose a potential problem, it is still possible to attempt to predict the sale price of a home based solely off the features of the home.

The EDA does suggest that there may need to be transformations in the predictor variables during the building process as a few transformations have already been made. Some of the transformations already undertaken include categorizing finished/unfinished homes, calculating the years between remodel and selling, creating the quality index and total square footage.

At this point in the analysis, the data-set has been reviewed to ensure that it contains, correct, accurate and sufficient data. The data set has been enhanced by removing unnecessary datapoints that would have created false assumptions based on incorrect outputs. Initial estimations on the significance of various predictor values to the sale price of a home have been identified and reviewed.

Two real-estate slogans were proven to me while conducting this analysis. These include "location, location, location" and "curb appeal". The location of the home had a large influence on the price. There were certain neighborhoods where even the lowest sold homes were significantly more expensive than homes of similar size in other neighborhoods. Equally so, the quality and condition of the exterior of the home adds thousands of dollars to the sale price of a home.

Appendix:

Data Quality Check:

Variable Name	**Quality Checks**	**Poor Quality Data**
SalePrice	Homes for sold for $0 or negative values	None
HouseStyle	Homestyle showing 1 story while SecondFlrSF has a value	233, 2000, 2006
TotalFloorSF	Homes with negative or 0 square footage	None
Year Built	Remodel Year is earlier than Build Year	851
Year Remod/Add	Remodel Year is earlier than Build Year	851
2nd Flr SF	Homestyle showing 1 story while SecondFlrSF has a value	233, 2000, 2006
Year Sold	The data appears to be grabbed in July of 2010 which makes 2010 incomplete. Any annual analysis would need to exclude 2010	1 - 341
Garage Cars	No Garage but number of cars capacity	None
SubClass	SubClass indicates 1 story, House style indicates 2 story	2735
TotRmsAbv Grd	0 or negative values	None

TotalBsmtSF	BsmtFinSF + BsmtFinSF2 + BsmtUnfSF != TotalBsmtSF	None
MasVnrType	!= None but MasVnrArea is 0	1641,1741,1786
MasVrArea	= None but MasVrArea is >0	364, 442, 1914
LotFrontage	Is blank	12,24,25,56,58,59,75,80,87,89,109,111,113,114,119,123,124,141,145,160,193,209,222,228,230,233,234,243,258,261,265,290,313,326,335,346,349,358,359,361,363,364,374,375,377,378,383,388,392,394,396,397,420,476,479,481,482,484,485,486,491,492,493,498,501,504,506,551,557,558,559,560,565,575,581,582,583,584,585,587,590,598,599,603,609,610,611,616,625,630,633,674,730,733,748,774,775,777,780,784,785,786,795,833,834,854,857,858,864,865,866,867,869,872,874,902,921,937,939,954,956,964,966,983,984,988,991,997,1006,1007,1008,1016,1020,1023,1025,1033,1038,1050,1090,1091,1096,1102,1105,1146,1151,1158,1182,1183,1187,1190,1199,1200,1202,1203,1204,1207,1218,1219,1220,1229,1235,1248,1264,1265,1331,1355,1377,1380,1383,1384,1385,1388,1389,1392,1393,1398,1420,1422,1429,1430,1435,1436,1445,1447,1448,1449,1456,1457,1460,1461,1462,1463,1495,1531,1535,1538,1539,1542,1564,1565,1566,1573,1585,1599,1616,1617,1618,1622,1623,1627,1628,1629,1630,1644,1645,1651,1652,1660,1664,1665,1670,1671,1712,1752,1754,1755,1758,1763,1764,1773,1812,1814,1817,1822,1826,1827,1828,1829,1833,1856,1866,1876,1882,1887,1888,1895,1896,1897,1912,1928,1942,1945,1964,2054,2057,2060,2064,2067,2072,2073,2075,2114,2115,2116,2117,2118,2120,2124,2132,2146,2147,2153,2176,2217,2223,2224,2225,2227,2229,2248,2249,2253,2254,2267,2268,2271,2273,2292,2294,2301,2306,2308,2312,2313,2314,2324,2325,2338,2344,2345,2346,2347,2353,2360,2362,2403,2432,2437,2438,2439,2441,2447,2452,2454,2483,2491,2493,2494,2495,2496,2498,2502,2503,2518,2526,2528,2538,2541,2542,2547,2553,2567,2577,2618,2671,2709,2713,2715,2717,2720,2724,2726,2736,2747,2765,2766,2772,2791,2793,2794,2796,2798,2860,2872,2894,2895,2898,2899,2927
LotArea	Negative or 0	None
Fireplaces	No fireplace but contains fireplace quality	None
MiscFeature	MiscFeature is NA but contains a value	None
MoSold	Not between 1-12	None
Overall Quality	Not between 1-10	None
Overall Condition	Not between 1-10	None

○