Northwestern University
Master of Science in Data Science

By: Ali Gowani

Table of Contents

Section 1: Overview & Target Definition

For us to understand and then define the Ames data, we need to see how the data looks like at a high level. By evaluating it at a 30,000-foot level, we can get a sense of the data: what type of features does it contain (categorical, ordinal, numerical, etc.), is the dataset complete or is it missing values, are there extreme outliers that we should consider, does it have the variables we need in order to predict the appropriate response, and etc.

The Ames dataset is provided by the Ames (Iowa) Assessor's Office, which was used to generate the assessed values for individual residential properties sold in Ames between 2006 and 2010. In total, there are 2,930 observations and 80 features (plus 2 Identifiers). From this, we can gather that there are enough observations for us to model our data in order to predict SalePrice as our response variable. However, we would need to reduce the number of features so that our model is not overfitting and filter our data to ensure we are predicting the SalesPrice for the appropriate "property".

Figure 1: Ames Dataset: Data Types

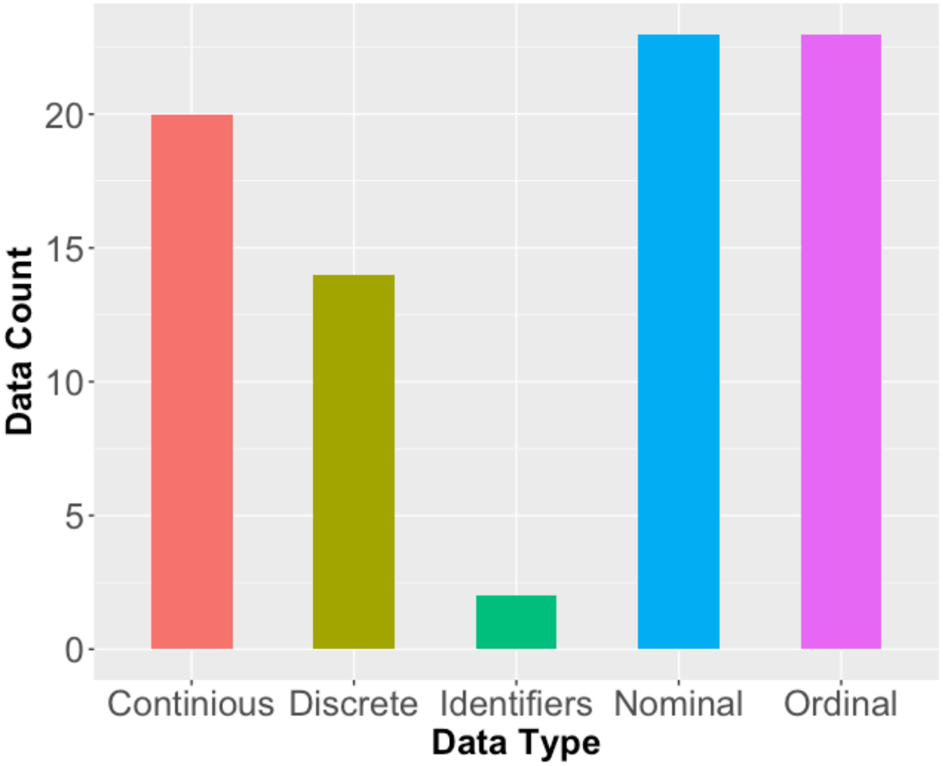
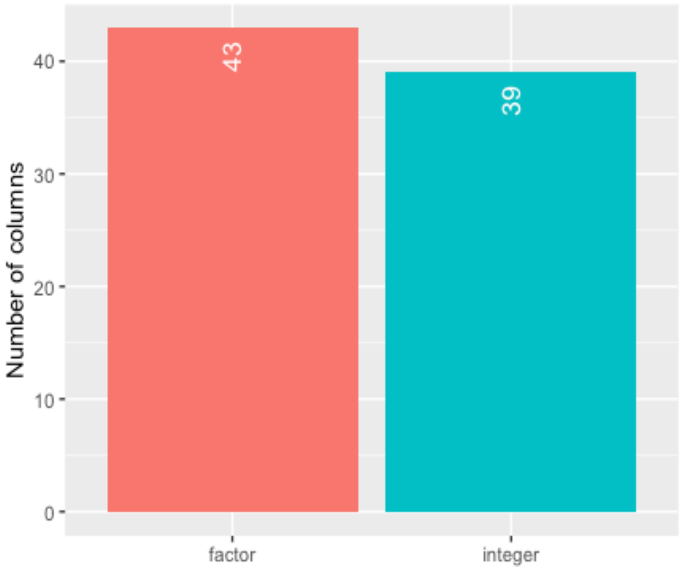


Figure 2: mydata Column (82) Types



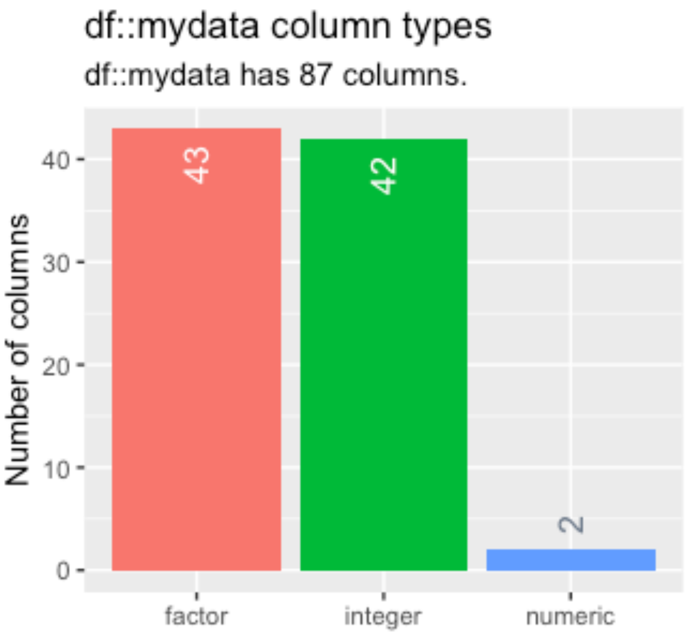
The above charts show the various types of data and this is important as it will help us analyze the data. When graphing the data to better understand it, we should be mindful of the different types of data that will be well suited for visualization, for example, between discrete and continuous variables.

Before, we define our population, we were given code to add additional features that may help our model. These features would bring clarity for us to properly build our model. The dataset now contains these additional features:

- TotalFloorSF, HouseAge, QualityIndex, logSalePrice, price_sqft

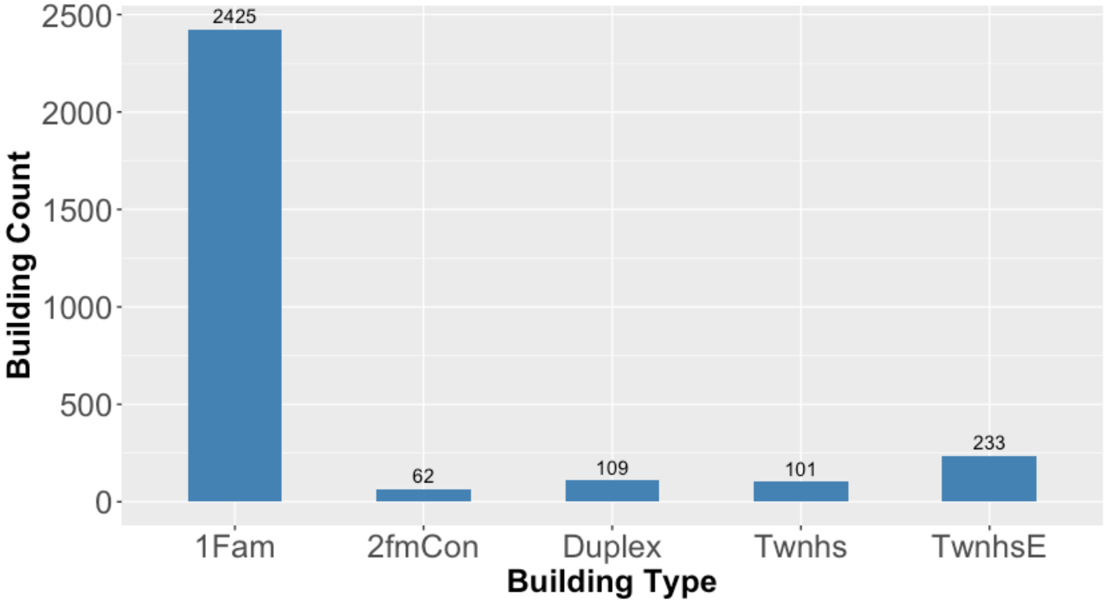
I also added TotalBath as part of our features as it did not make sense to have 2 variables (e.g.: FullBath and HalfBath) that would provide similar information and it might be beneficial to look at them together as we trim the number of features.

Figure 3: Additional Features Column (87) Types



We can now define the population we want to sample. In this case, our response variable is SalePrice but SalePrice for what? A feature that can significantly impact SalePrice variable is Building Type (BldgType). For example, it is widely known to use square feet as predictor of SalePrice, assuming, other factors are held equal, but this can be misleading if we are comparing the square feet of a townhouse versus a single-family home. A townhouse can be 3,000 square feet similar to a 3,000 square feet single family house, however, the value of a townhouse would be lower as it may mean that there is a shared wall with another house, there may not be covered parking, there may be shared areas on the property, etc. If we are going to predict the price of a house, then we need to ensure we are comparing similar types of houses and in this case, we should only look at single family homes.

Figure 4: Ames Dataset Building Types



By remove all Building Types except, 1Fam, then we are left with 2,425 observations from our original dataset of 2,930. While we want to maintain as many observations as possible, we do not want our data response variables to be swayed with irrelevant data.

In addition, we see that the distribution of extremes SalePrice can be problematic. We know that houses under $50,000 will not help our model and homes over $500,000 will not help either so we can filter those observations.

We have now defined our target population. If we select observations of homes that are BldgType of 1Fam and homes whose SalePrice is not less than $50,000 and not over $500,000 then we can build a model to predict our target variable of SalePrice. By filtering the for 1Fam and the range of SalePrice, this leaves us with 2,397 observations. We can now move forward with doing a Data Quality Check. We will look at how “clean” is our data and to reduce the number of features for us to model.

Section 2: Data Quality Check

Before starting the Data Quality Check, we need to reduce our features and select features that we feel would be important and add meaningful value to our model. While the features may vary from each model, I have selected the following 20 features that I feel would be important and perhaps, if I were a buyer then I would look at these features for my target (SalePrice). I also looked at whether the features were complete or had missing data. If features had missing data, then I selected to not include it as part of my selection of 20 features.

20 Features from Filtered Data:

SalePrice, TotalFloorSF, HouseAge, QualityIndex, TotalBath, SubClass, LotArea, Utilities, Neighborhood, BldgType, HouseStyle, ExterCond, Exterior1, ExterQual, Foundation, BsmtQual, KitchenQual, Functional, GarageType, Condition1, PoolArea

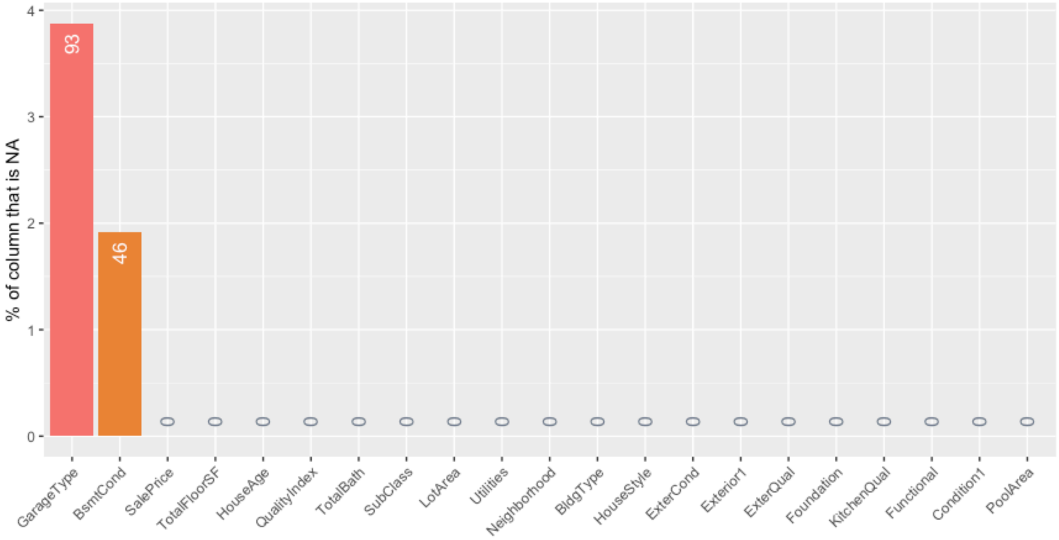
I used SummaryTools to evaluate the dataset for distribution of values, frequency of values, simple statistics, ordinal and nominal values, etc. [*Reference Appendix A: Data Quality Check (20 Variables)*]. By going through each feature, we can see that we do not have any duplicate observations. This is important as at times, when receiving data or when merging various datasets, we can have entire rows that are duplicates or portions of rows that are duplicates and we must decide which row or observation we should keep. In this case, we do not have any duplicate rows, so we do not need to take any actions to address this.

However, we do have missing data (GarageType and BsmtCond). The missing data is quite nominal; under 5% for both features. We can impute the values based on some logic (e.g.: mean, median, etc.) or make an assumption as to what value should go into each observation using other factors (e.g.: age of the house, etc.). A common approach is the perhaps put “NA” for GarageType which would indicate “No Garage” and we can use “NA” for “No Basement.” This would be a conservative approach as that house may actually have a Garage or a Basement, but it just was not recorded.

Based on additional research, I was able to find that a rule of thumb is that if less than 5% of the observations are missing then the missing data can simply be deleted without any significant ramifications (Harrell, 2001).

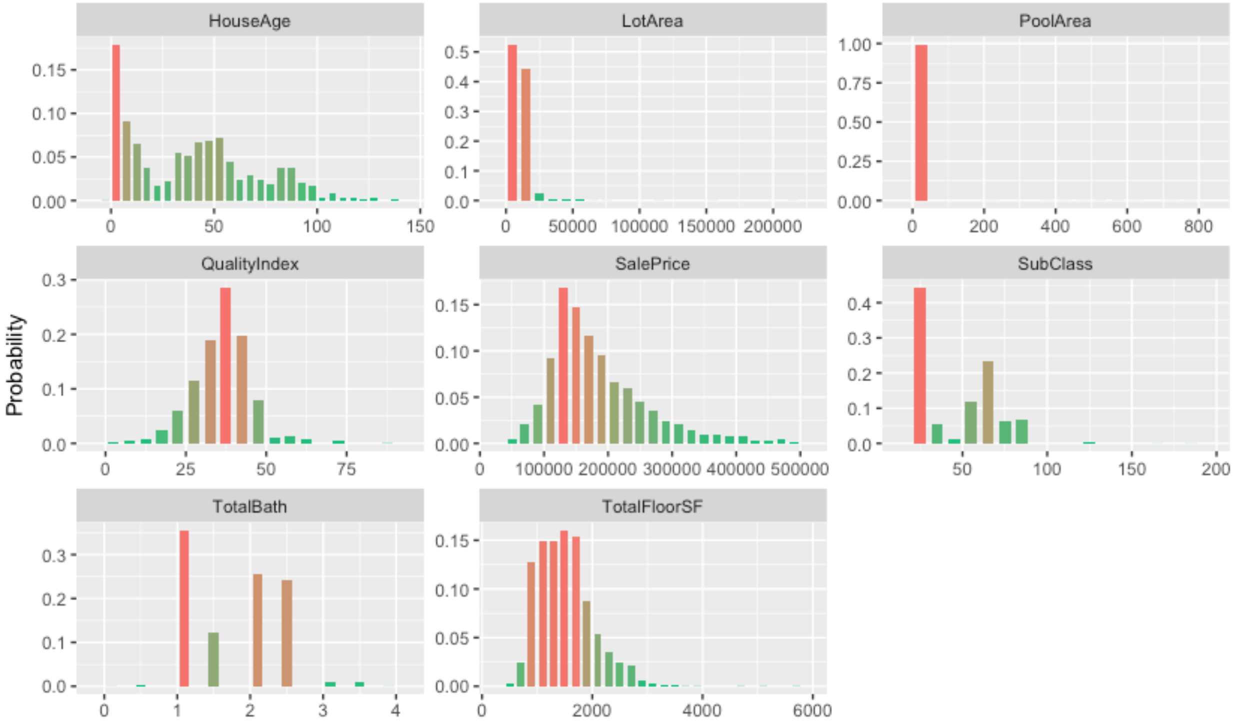
When it comes to imputing values and addressing null and NAs, I am new to this area so may have to research more as to how to handle such scenarios. Especially, how to properly address each type of feature: numerical, ordinal or categorical.

Figure 5: Prevalence of NAs in Dataset



As we continue to explore the shape of the data with 20 variables of our sample population, we see some interesting visuals that may guide our decision on which variables to remove. For our numeric variables, we see interesting things with PoolArea, SubClass, TotalBath and LotArea.

Figure 6: Histograms of Numeric Columns in Defined and Selected Dataset



Pool Area: this seemed like a good variable to select as a pool usually, tends to be attractive option to sell the house though, it doesn't always relate to increase in SalePrice. We selected Pool Area, but it seems the pools are all less than 800 square feet. In addition, it seems there are only 13 pools out of the 2,425 observations. This could mean that not many people have pools in their homes, or the dataset doesn't have the size of the all the pools captured.

Sub Class: while most of the houses fall under the SubClass of 100, we have few outliers that are not too much of a concern but may skew our data.

- 120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER
- 150 1-1/2 STORY PUD - ALL AGES
- 160 2-STORY PUD - 1946 & NEWER
- 180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
- 190 2 FAMILY CONVERSION - ALL STYLES AND AGES

SubClass	Freq
20	1077
30	139
40	6
45	18
50	287
60	574
70	128
75	23
80	118
85	48
120	5
160	1
190	1
<NA>	0
Total	2425

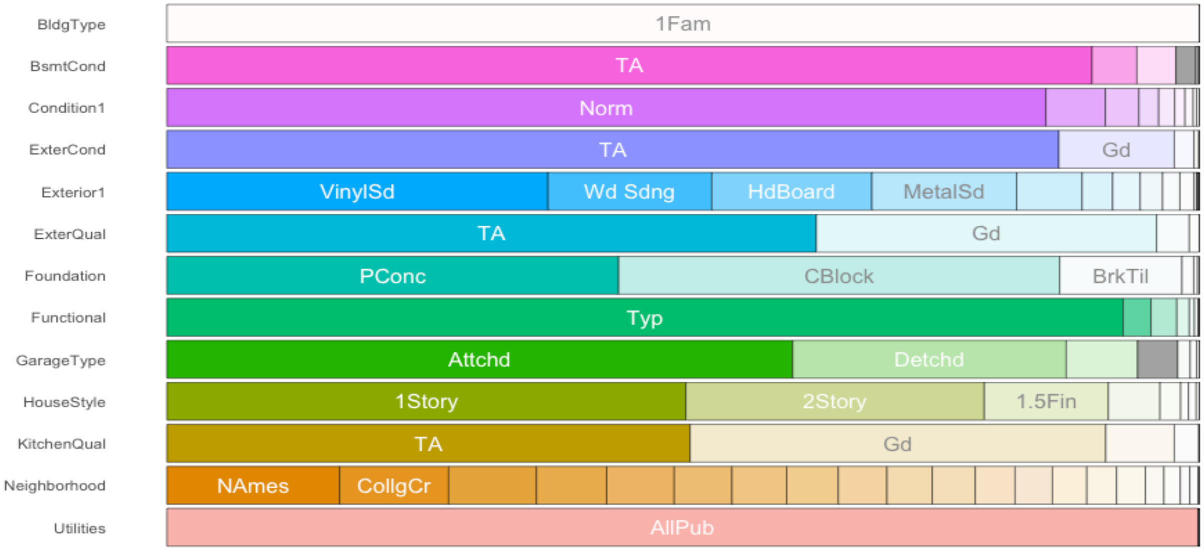
Total Bath: we combined full and half bath to provide a sum of Total Bath. This initially sounded like a good idea but after seeing the graph above, it makes me wonder whether 2 Half Bathrooms are same as 1 Full Bathroom. A Full Bathroom usually consists of a shower so I would say that 2 Half Bathrooms are not the same has 1 Full Bathroom. Also, when looking at the raw dataset, the Half Bathroom column had whole numbers. In seeing houses in the market, a Half Bathroom is denoted by .5, so in calculating the Total Bath variable, I did divide Half Bathroom by half to ensure we were properly capturing the .5. In either case, I think it would have been better to have kept these two variables separate.

- mydata$TotalBath <- mydata$FullBath + (mydata$HalfBath / 2)

Lot Area: I initially thought that this would be a good variable but looking at the histogram for this variable, it really doesn't provide a unique perspective to our model. We have few extreme outliers but if we are looking at the SalePrice of a house, the Lot Area probably won't factor too much into the model.

Let's look at the categorial variables from our selection of the 20 that we selected. The BldgType is 1Fam as we had used it for our sample definition. However, there are few variables that jump out that may not provide meaningful information: BsmtCond, Functional and Utilities.

Figure 7: Frequency of Categorical Features



BsmtCond: most of the values in this variable are TA (Typical), so it does not really add unique information between the different observations. In addition, there is missing information. We would have to really consider whether this variable is adding a new dimension to our model for it to be robust in predicting SalePrice.

Functional: I thought this was an interesting variable as it refers to the functionality or livability of the home. Does the home warrant any deductions, is it severely damaged or should it be salvaged? What we should have considered is that cities have ordinances that usually do not allow homes to go too long without repairs so we should expect most homes to fall under Typical Functionality. Therefore, this may not provide any additional unique information to our model.

Utilities: I initially thought that utilities are a must and all houses may not have all utilities. Looking back at the dictionary, it shows that options for No Sewer, while exists are extremely unlikely, at least in today's day and age. This variable does not seem to add any meaningful information.

Section 3: An Initial Exploratory Data Analysis

After conducting the data quality check with the 20 variables and getting a better understanding of the type of meaningful information each may add to build a robust model, I have selected the following 10 variables for us to do a deeper analysis.

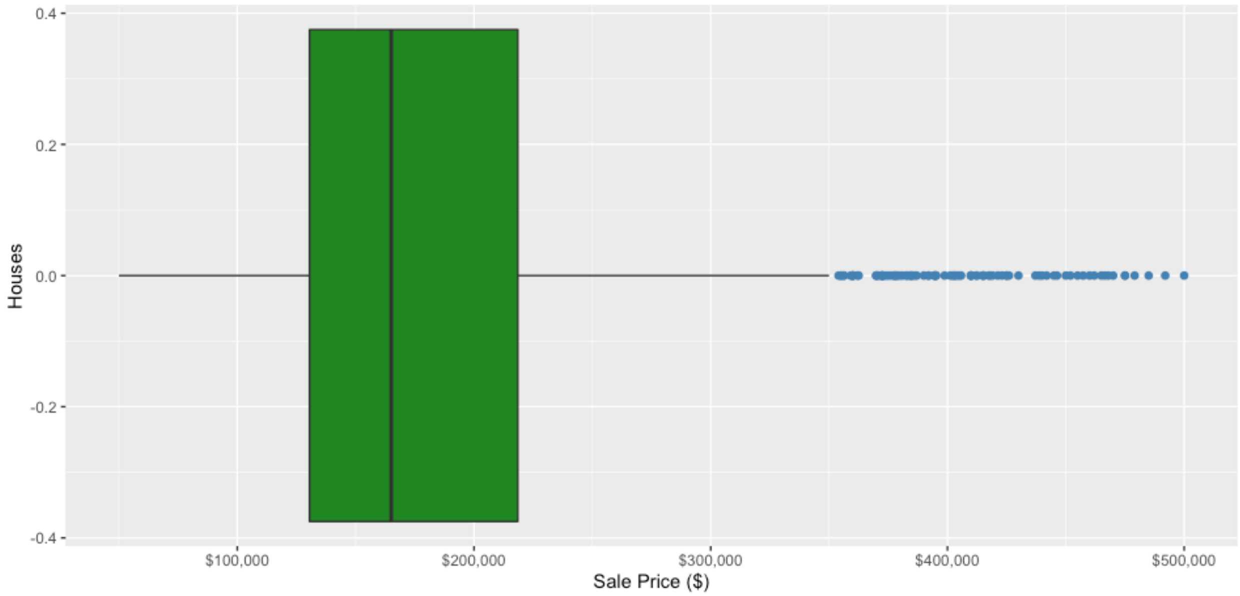
10 Features from Filtered Data:

SalePrice, TotalFloorSF, HouseAge, QualityIndex, Neighborhood, BldgType, HouseStyle, Exterior1, Foundation, GarageType, Condition1

As we have narrowed our features from 20 to 10 and we have already done a high-level exploration of the type of features, our focus should shift to see how these 10 variables interact with our target variable: SalePrice. We will use this information to begin our modeling process.

Before we start to compare variables to SalePrice, let's look at how SalePrice is distributed and whether there are any outliers.

Figure 8: Box Plot of SalePrice

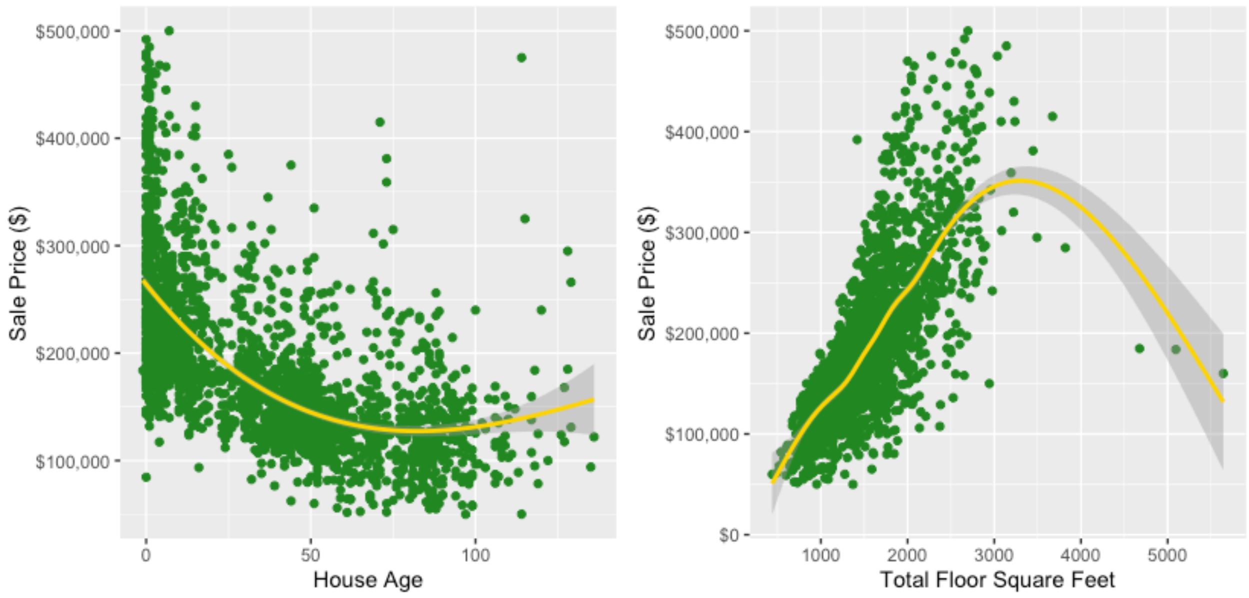


We can see that majority of homes fall between $150,000 and $225,000 and then there is a significant gap before we start seeing outliers around the $350,000 mark. This is even after we removed observations that were more extreme, where SalePrice was greater than $500,000. If we had to hypothesize as to why there are outliers, it could mean that new construction for homes have been started and are in high demand. New homes usually cost more than existing homes with similar features. It could also mean that there is a significant disparity between average, middle-class citizens and those of high-class.

Let’s continue to explore SalePrice compared to other features in our dataset. A common feature evaluated when buyers consider SalePrice is the age of the house. We would typically expect a house to get its peak SalePrice when it is new, assuming the demand for houses in the area is held equal. However, as the house gets older, we would expect the price of the house to stay at or below the cost it would take to build a new home with same features. This is because of the wear and tear of the house. The graph below shows what is to be expected, older houses have a SalePrice that is lower. However, we should keep in mind that it could also be that the older houses are smaller so perhaps their SalePrice would be lower. It is important to look at the other chart below, where we compare the total size of the house (Total Floor Square Feet) to SalePrice.

In general, we should see bigger houses would have a higher SalePrice and this holds true based on the graph below. However, we do see some points that don’t hold true. We see large houses that are greater than 4,000 sq. ft. but are priced below $200,000. Perhaps, these are the houses that were built many years ago and need a lot of repairs. It could also be that these houses have significant issues (e.g.: foundation issues, weather damage, termites, etc.).

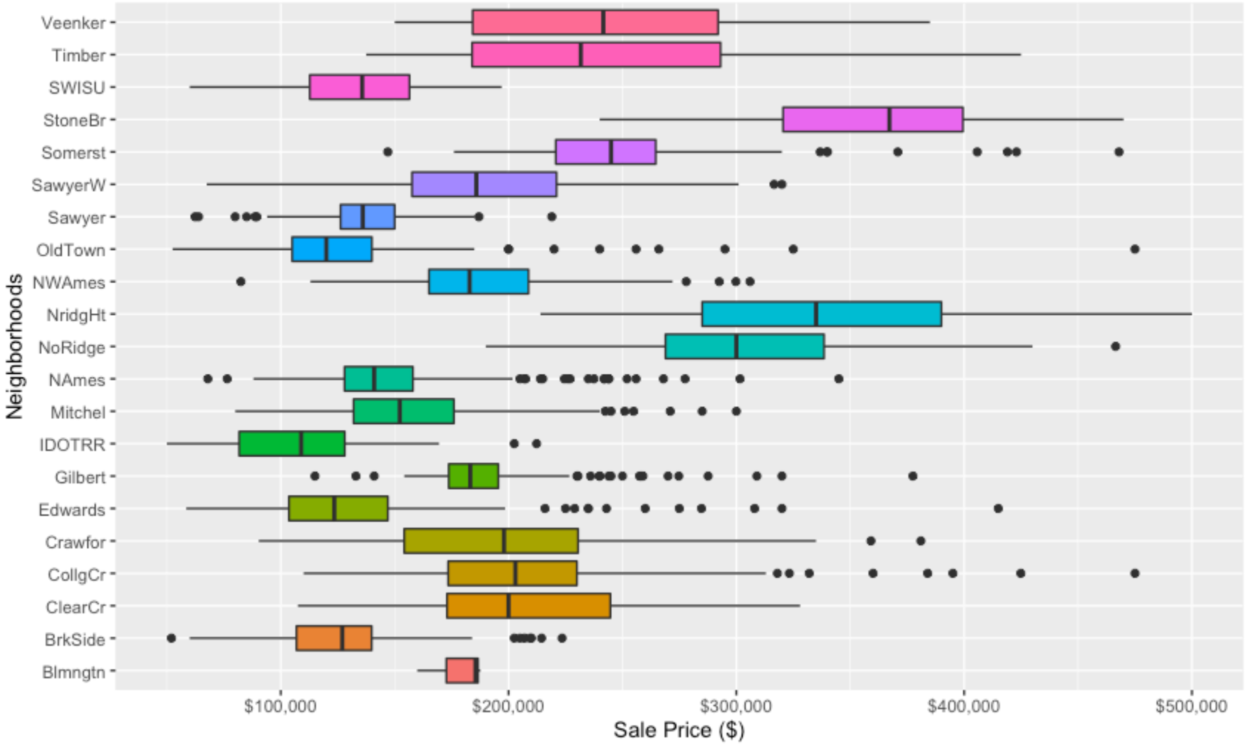
Figure 9: Comparing SalePrice to House Age and Total Floor Sq. Ft.



Let’s continue to evaluate our features to SalePrice but this time, let’s focus on two categorical features: Neighborhood and Home Style. A real estate friend once told me, that the three most important factors in selling a home: “location, location, location.” Needless to say, Neighborhood is one of the key categorical features that we should evaluate. Homes of the same age, features and condition can have significantly different SalePrice because of its location. Sometimes this has to do with school districts and crime.

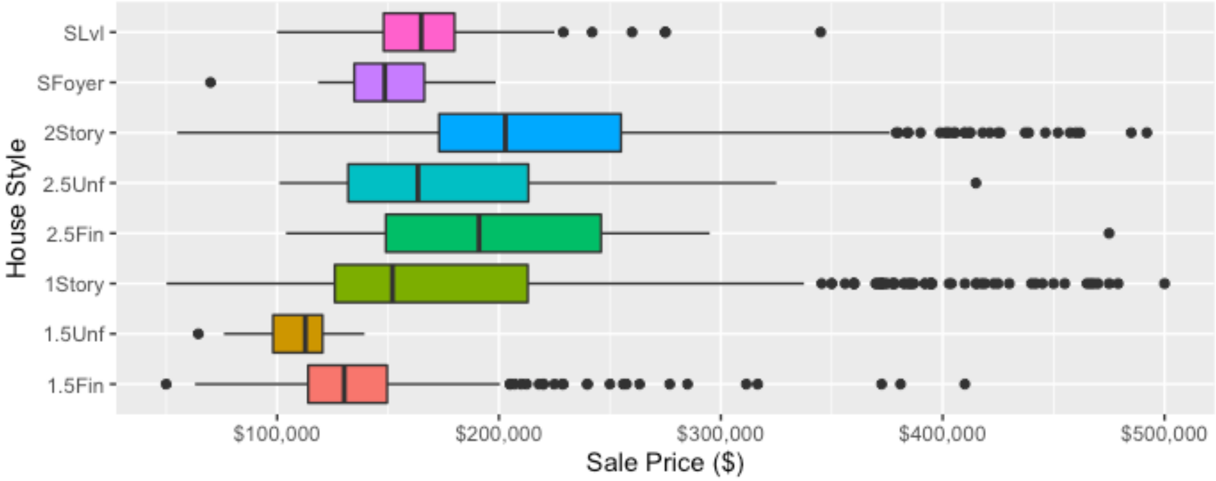
What is interesting is neighborhoods such as, Veenker, SWISU and Timber have a wide range of SalePrices but no outliers. Then there are others such as, Gilbert and Edwards that have many outliers at high SalePrice. We can see how Neighborhood would be an important Feature in our model as it can help determine the range of SalePrice just by being in one neighborhood versus another.

Figure 10: Box Plot Comparing Neighborhoods to SalePrice



I am not sure whether comparing House Style to SalePrice add any more information to what we already know. We know there are SalePrices of houses in extreme but seeing this on various House Style does not provide any new information. It would be interesting if we saw more extreme values on bigger houses such as, with 2 Story houses but it seems there are extremes for all types of typical House Styles (1 Story and 2 Story). We even see extreme SalePrices for 1.5 Story Houses where the second story is not yet finished.

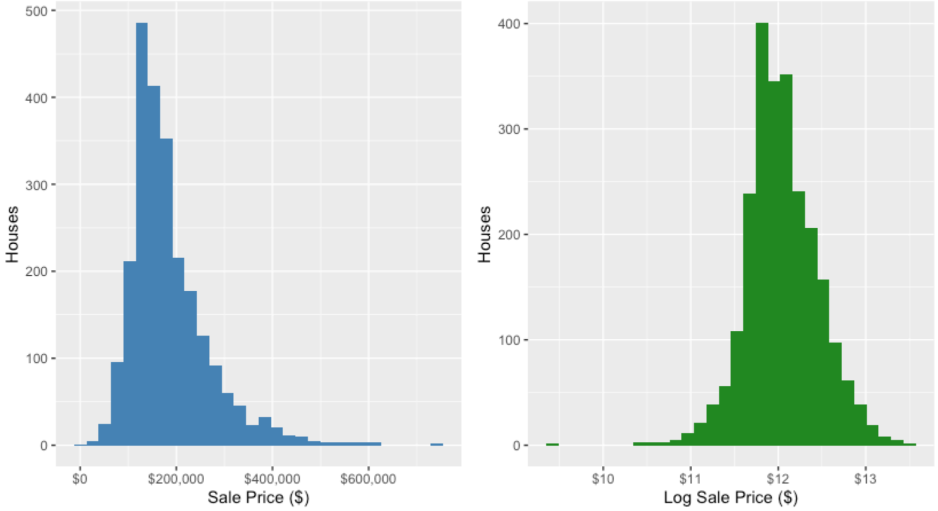
Figure 11: Box Plot Comparing House Style to SalePrice



Section 4: Exploratory Data Analysis for Modeling

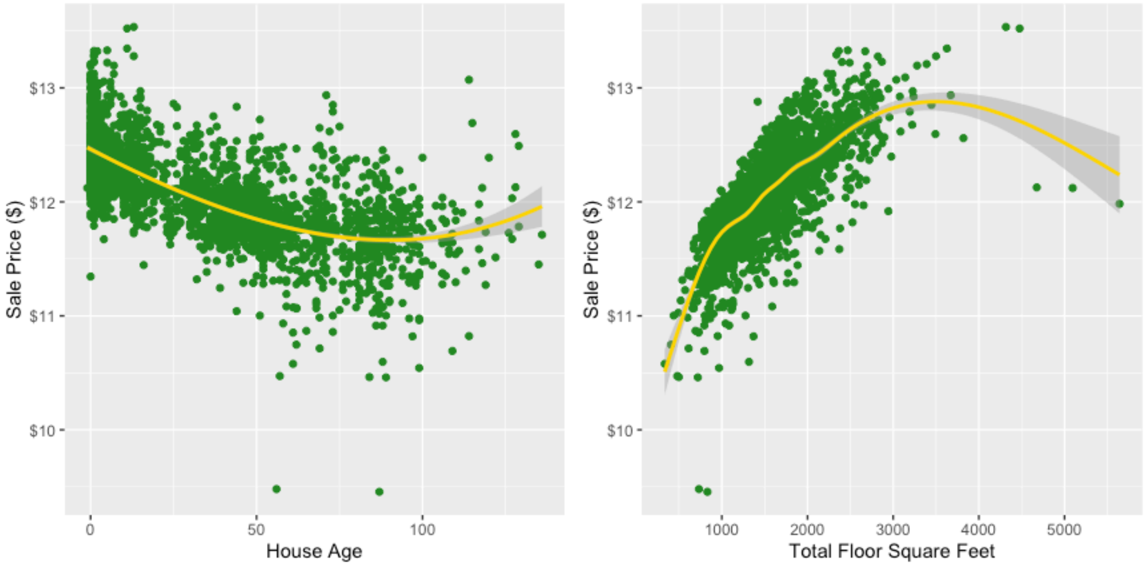
Our response variable in this problem is SalePrice. Because of a wide disparity of SalePrices, we should consider transforming the response variable. This will allow us to obtain residuals to be approximately symmetrically distributed. We can see in the blue histogram on the left that it is skewed towards the left and has some outliers at the high side of the SalePrice. This can be problematic when houses have extreme values to the left (low SalePrice) and right (high SalePrice). It can skew our results and our model may not be as effective. When we convert the SalePrice to Log SalePrice, the distribution of the SalePrice is on a consistent scale and has more of a symmetrical distribution.

Figure 12: SalePrice and Log SalePrice Comparison



Comparing the House Age and Total Floor Square Feet to the Log SalePrice, we can see that the data is distributed more symmetrically and may produce a more robust model to predict our target variable: SalePrice.

Figure 13: Comparing Log SalePrice to House Age and Total Floor Sq. Ft.



The two figures below show us comparing the Log SalePrice to variables (i.e.: Neighborhoods and House Style) we compared previously with just the SalePrice. Unlike the previous figures with the SalePrice, where most of the outliers were to the right (higher SalePrice), it is more symmetrically distributed. This was even more apparent when looking at House Style and Log SalePrice. The outliers seem to be even distributed on the low end and high end of the SalePrice between the various house styles. Transforming SalePrice to a Logarithmic SalePrice should help build a better model.

Figure 14: Box Plot Comparing Neighborhoods to Log SalePrice

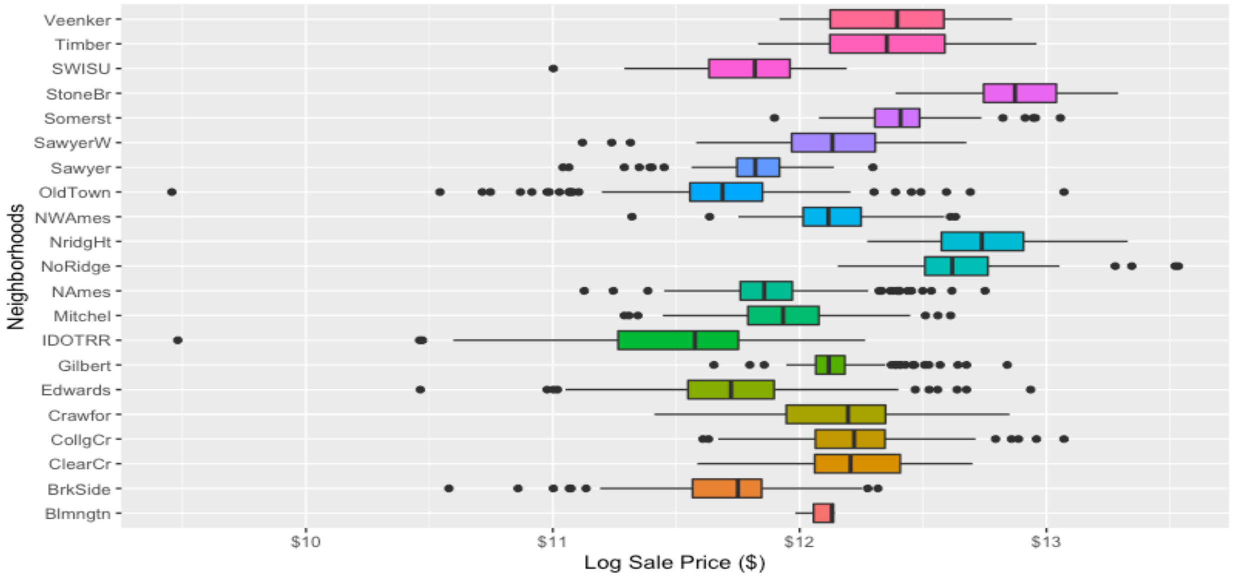
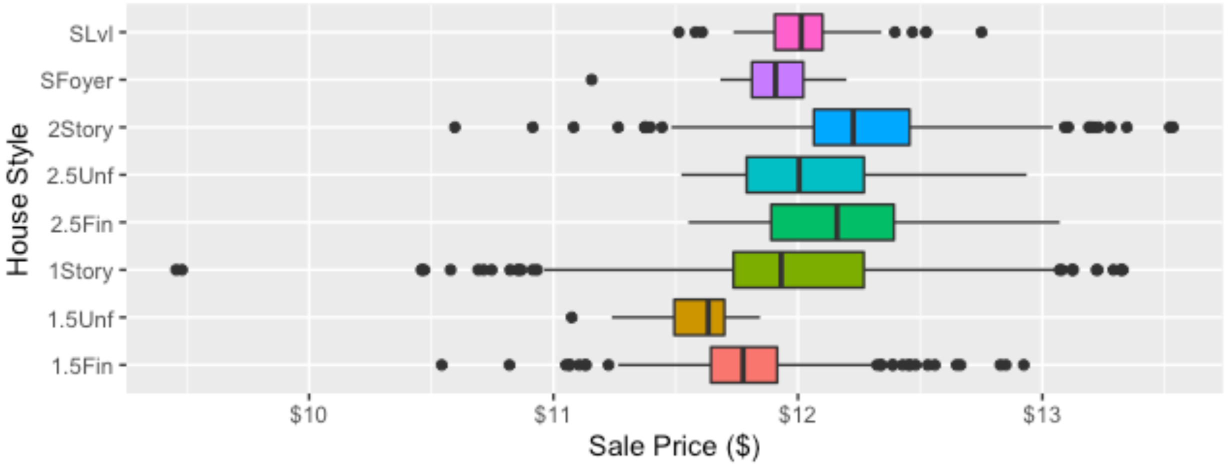


Figure 15: Box Plot Comparing House Style to Log SalePrice



Looking at the skewness of our numeric variables, it would be more appropriate to log transform all numeric variables as oppose to picking and selecting only certain variables.

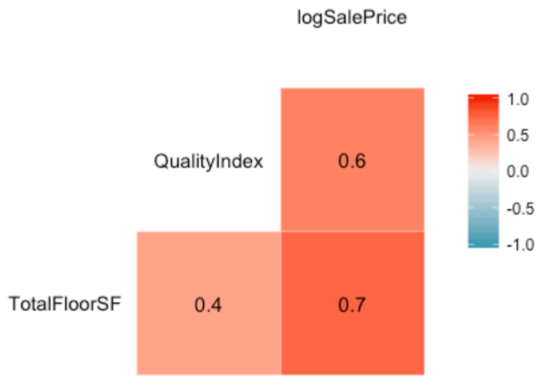
3 Features from Filtered Data:

TotalFloorSF, Neighborhood, QualityIndex, logSalePrice

I selected the 3 features above as I think they provide a unique value to each observation. However, I think it is going to be quite challenging for our model to perform with low error or smaller range of error if we only select few features as it may not be able to generalize well.

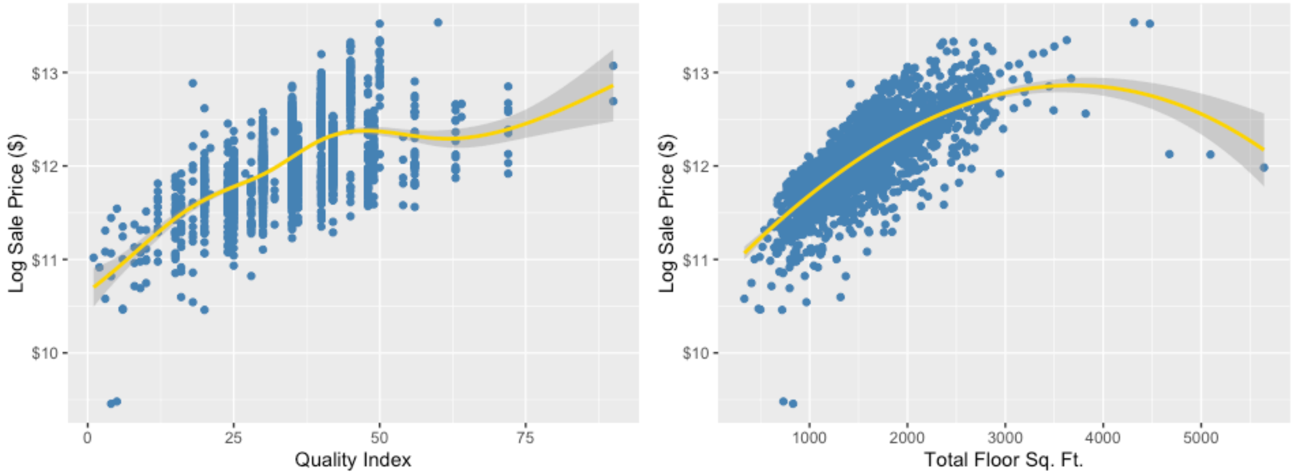
The correlation between QualityIndex and TotalFloorSF compared to Log SalePrice seems to be high relatively speaking as we only have 3 numeric variables.

Figure 16: Correlation Plot Between Floor Size, Quality and Log SalePrice



The 2 scatter plots below show how the Log SalePrice interacts with QualityIndex and TotalFloorSF variables. It validates our assumption that higher quality of the houses and larger houses warrant a higher price. Though, we do see some variation in Log SalePrice towards the higher QualityIndex and TotalFloorSF. This could be explained that perhaps the neighborhood of the house is not in a desirable location or it could also be anomalies. If we can do further investigation, then I would suggest we should remove these observations as they are introducing a bias to our model with these outliers.

Figure 17: Scatter Plot Comparing Quality Index and Total Floor Sq. Ft. to Log SalePrice



Section 5: Summary, Conclusion and Reflection

To summarize, we were asked to evaluate the Ames housing dataset by conducting an exploratory data analysis and by using the SalePrice as our target variable. While, we were not told to actual execute a model to determine the accuracy and error of it, we were told to select a subset of the variables. We original had 82 variables and then added few more variables to help with certain things such as quality and condition of the house to totaling the number of bathrooms. This increased the number of variables in our dataset to 87.

There were features of the dataset that had missing values and then to determine, what we should do with those values. Should we remove those observations or impute them into the missing data. We also had to deal with outliers and in our case of SalePrice, there were many extreme outliers. We hypothesized about them to determine if we could address them in a meaningful manner. Our approach to define our target population was to remove observations where the SalePrice fell below $50,000 and about $500,000. This would help mitigate observations that were just too extreme to help us build a sensible model. In addition, our focus was for Single Family houses, so we removed other types of houses (i.e.: Two-family Conversion, Duplex, Townhouse End Unit and Townhouse Inside Unit).

We selected 20 variables, plus our target variable to start getting a better sense of where we were headed. This I thought was a difficult exercise to reduce the number of features in our dataset and then to reduce it even further to 10. And then to go even further and reduce it to 3 variables!

An area that I would have liked to focus on for this exercise is to evaluate the various metrics on the condition and the quality of the house. There were features for quality of the roof, basement, garage, interior, exterior, etc. It would have been worthwhile to see if we had to select 1 of these measures then which would be ideal.

An area of concern for handling the model building process would be what to do with missing data. In general, I am hesitant to delete anything, but I guess in modeling, we have to take some chances. Another area of concern would be whether we have the right type of features to truly build a good model or at least a model that minimizes our range for error. We were given many features but could we be missing features not necessarily about the house itself but more so of the buyer or the economy. SalePrice is not just a function of the quality of the house and what features it has but also a function of what the buyer is willing to pay for it. In an expanding economy, a buyer may pay more for the house whereas, in a recession, the buyer may pay less for the house.

In reflection, I am getting a deeper appreciation to first study and understand the data before trying to analyze it. For example, we may have a variable for the condition of the house, but do we understand what each option for the condition really means. Our dataset shows there are 20 unique neighborhoods in Ames, Iowa but do we understand what makes each neighborhood different. Going a bit further, we may know the types of houses in these neighborhoods, but do we know who lives there and the type of people? Understanding the nuances of our data will help us weave a better story and our model will be more robust.

Another area that I found quite interesting and something that I learned is that when selecting our features, we want each of them to bring something new or different into our model. During some of the analysis that I had to redo, I learned that the age of the house and the quality of it have a high negative correlation. As the age of the house increases the quality index (the overall quality and condition) of the house will decrease. This is not an absolute rule but has a high correlation. If the features provide the same type of information, then they may not be independent of each other. If this is the case, then it may be better to remove one of the features.

Lastly, I learned that I need to develop a better framework on how I approach modeling. I had to go back many times to the dictionary to better understand the variables and the type of information in them. I also had to think through what types of variables could be omitted because another variable that I had selected was providing similar information. I spent many more hours that I had originally planned, and I had planned for over 20 hours! Learning modeling, especially, with no prior background is going to be challenging but I’m for it!

Appendix

A: Data Quality Check (20 Variables and SalePrice)

Dimensions: 2397 x 21

Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
1	SalePrice [integer]	Mean (sd) : 182637.1 (75106.2) min < med < max: 50000 < 165000 < 500000 IQR (CV) : 88000 (0.4)	902 distinct values	2397 (100%)	0 (0%)
2	TotalFloorSF [integer]	Mean (sd) : 1506.2 (502.5) min < med < max: 438 < 1458 < 5642 IQR (CV) : 649 (0.3)	1181 distinct values	2397 (100%)	0 (0%)
3	HouseAge [integer]	Mean (sd) : 38.2 (30.7) min < med < max: -1 < 38 < 136 IQR (CV) : 49 (0.8)	127 distinct values	2397 (100%)	0 (0%)
4	QualityIndex [integer]	Mean (sd) : 34.2 (9.2) min < med < max: 1 < 35 < 90 IQR (CV) : 10 (0.3)	35 distinct values	2397 (100%)	0 (0%)
5	TotalBath [numeric]	Mean (sd) : 1.7 (0.6) min < med < max: 0 < 2 < 4 IQR (CV) : 1.5 (0.4)	0.00 : 1 (0.0%) 0.50 : 4 (0.2%) 1.00 : 854 (35.6%) 1.50 : 291 (12.1%) 2.00 : 614 (25.6%) 2.50 : 581 (24.2%) 3.00 : 26 (1.1%) 3.50 : 25 (1.0%) 4.00 : 1 (0.0%)	2397 (100%)	0 (0%)
6	SubClass [integer]	Mean (sd) : 41.6 (22.2) min < med < max: 20 < 30 < 190 IQR (CV) : 40 (0.5)	13 distinct values	2397 (100%)	0 (0%)
7	LotArea [integer]	Mean (sd) : 10885.9 (7484) min < med < max: 2500 < 9800 < 215245 IQR (CV) : 3720 (0.7)	1650 distinct values	2397 (100%)	0 (0%)
8	Utilities [factor]	1. AllPub 2. NoSeWa 3. NoSewr	2394 (99.9%) 1 (0.0%) 2 (0.1%)	2397 (100%)	0 (0%)
9	Neighborhood [factor]	1. Blmngtn 2. Blueste 3. BrDale 4. BrkSide 5. ClearCr 6. CollgCr 7. Crawfor 8. Edwards 9. Gilbert 10. Greens [18 others]	3 (0.1%) 0 (0.0%) 0 (0.0%) 105 (4.4%) 43 (1.8%)	2397 (100%)	0 (0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
			253 (10.6%) 87 (3.6%) 157 (6.5%) 163 (6.8%) 0 (0.0%) 1586 (66.2%)		
10	BldgType [factor]	1. 1Fam 2. 2fmCon 3. Duplex 4. Twnhs 5. TwnhsE	2397 (100.0%) 0 (0.0%) 0 (0.0%) 0 (0.0%) 0 (0.0%)	2397 (100%)	0 (0%)
11	HouseStyle [factor]	1. 1.5Fin 2. 1.5Unf 3. 1Story 4. 2.5Fin 5. 2.5Unf 6. 2Story 7. SFoyer 8. SLvl	288 (12.0%) 18 (0.8%) 1205 (50.3%) 7 (0.3%) 18 (0.8%) 693 (28.9%) 48 (2.0%) 120 (5.0%)	2397 (100%)	0 (0%)
12	ExterCond [factor]	1. Ex 2. Fa 3. Gd 4. Po 5. TA	11 (0.5%) 46 (1.9%) 268 (11.2%) 1 (0.0%) 2071 (86.4%)	2397 (100%)	0 (0%)
13	Exterior1 [factor]	1. AsbShng 2. AsphShn 3. BrkComm 4. BrkFace 5. CBlock 6. CemntBd 7. HdBoard 8. ImStucc 9. MetalSd 10. Plywood [6 others]	33 (1.4%) 0 (0.0%) 6 (0.3%) 71 (3.0%) 2 (0.1%) 64 (2.7%) 372 (15.5%) 1 (0.0%) 336 (14.0%) 152 (6.3%) 1360 (56.7%)	2397 (100%)	0 (0%)
14	ExterQual [factor]	1. Ex 2. Fa 3. Gd 4. TA	76 (3.2%) 23 (1.0%) 790 (33.0%) 1508 (62.9%)	2397 (100%)	0 (0%)
15	Foundation [factor]	1. BrkTil 2. CBlock 3. PConc 4. Slab 5. Stone 6. Wood	284 (11.8%)	2397 (100%)	0 (0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
			1024 (42.7%) 1049 (43.8%) 27 (1.1%) 9 (0.4%) 4 (0.2%)		
16	BsmtCond [factor]	1. (Empty string) 2. Ex 3. Fa 4. Gd 5. Po 6. TA	1 (0.0%) 3 (0.1%) 91 (3.9%) 104 (4.4%) 4 (0.2%) 2148 (91.4%)	2351 (98.08%)	46 (1.92%)
17	KitchenQual [factor]	1. Ex 2. Fa 3. Gd 4. Po 5. TA	159 (6.6%) 57 (2.4%) 965 (40.3%) 1 (0.0%) 1215 (50.7%)	2397 (100%)	0 (0%)
18	Functional [factor]	1. Maj1 2. Maj2 3. Min1 4. Min2 5. Mod 6. Sal 7. Sev 8. Typ	12 (0.5%) 9 (0.4%) 60 (2.5%) 65 (2.7%) 29 (1.2%) 1 (0.0%) 1 (0.0%) 2220 (92.6%)	2397 (100%)	0 (0%)
19	GarageType [factor]	1. 2Types 2. Attchd 3. Basment 4. BuiltIn 5. CarPort 6. Detchd	15 (0.7%) 1453 (63.1%) 29 (1.3%) 166 (7.2%) 6 (0.3%) 635 (27.6%)	2304 (96.12%)	93 (3.88%)
20	Condition1 [factor]	1. Artery 2. Feedr 3. Norm 4. PosA 5. PosN 6. RRAe 7. RRAn 8. RRNe 9. RRNn	78 (3.2%) 138 (5.8%) 2041 (85.2%) 19 (0.8%) 37 (1.5%) 24 (1.0%) 46 (1.9%) 6 (0.2%) 8 (0.3%)	2397 (100%)	0 (0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
21	PoolArea [integer]	Mean (sd) : 2.5 (37.6) min < med < max: 0 < 0 < 800 IQR (CV) : 0 (15.3)	12 distinct values	2397 (100%)	0 (0%)

B: References

1. Harrell, Frank E. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer, 2001.