Julia Rodd Assignment 1

Contents

Introduction

This assignment analyzes housing data from Ames, IA from 2006 - 2010 with a primary purpose of understanding which variables would be good candidates for predicting the sale price of a home. Exporatory data analysis techniques are used to aid in the understanding of this data as well as to inform which observations (houses) should be excluded. Commentary is provided throughout this assignment on trends. Ultimately, analysis is performed on a subset of the Ames housing data with some initial conclusions on key variables and additional information needed. The R ggplot and dplyr packages are the two primary packages used in this assignment.

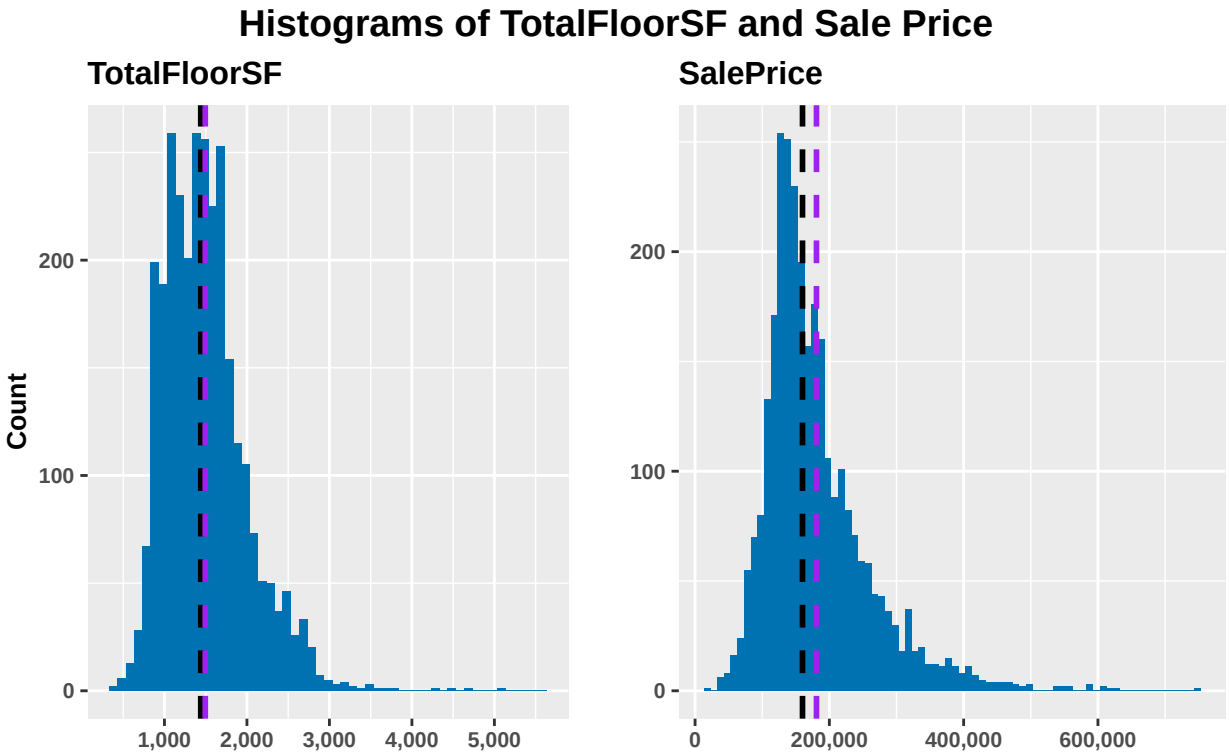
Results

Section 1: Data Survey

Within the Ames data set, we have 2,930 observations, which represent individual homes that were sold between 2006 - 2010 in Ames, IA. More importantly, we have 82 variables in our data set. These 82 variables represent characteristics of these individual homes. Some examples of these variables include: sale price, month and year sold, the type of home, number of bathrooms, and details on several characteristics of the home, such as the size (in square feet) and condition of the basement. The variables in the Ames data set are of mixed type: there are both numeric and factor variables.

Upon initial inspection of the data based on the readings, this data set appears rich and to contain needed variables to predict/fit the sale price of a home. However, we will need to perform some initial exploratory analysis to better understand what kind of data we are working with and what considerations we may need to make in creating a sample data set for modeling.

We start our analysis by examining histograms of TotalFloorSF and SalePrice.



NOTE: Median values are displayed as the dashed black line and Mean values are displayed as the purple dashed line

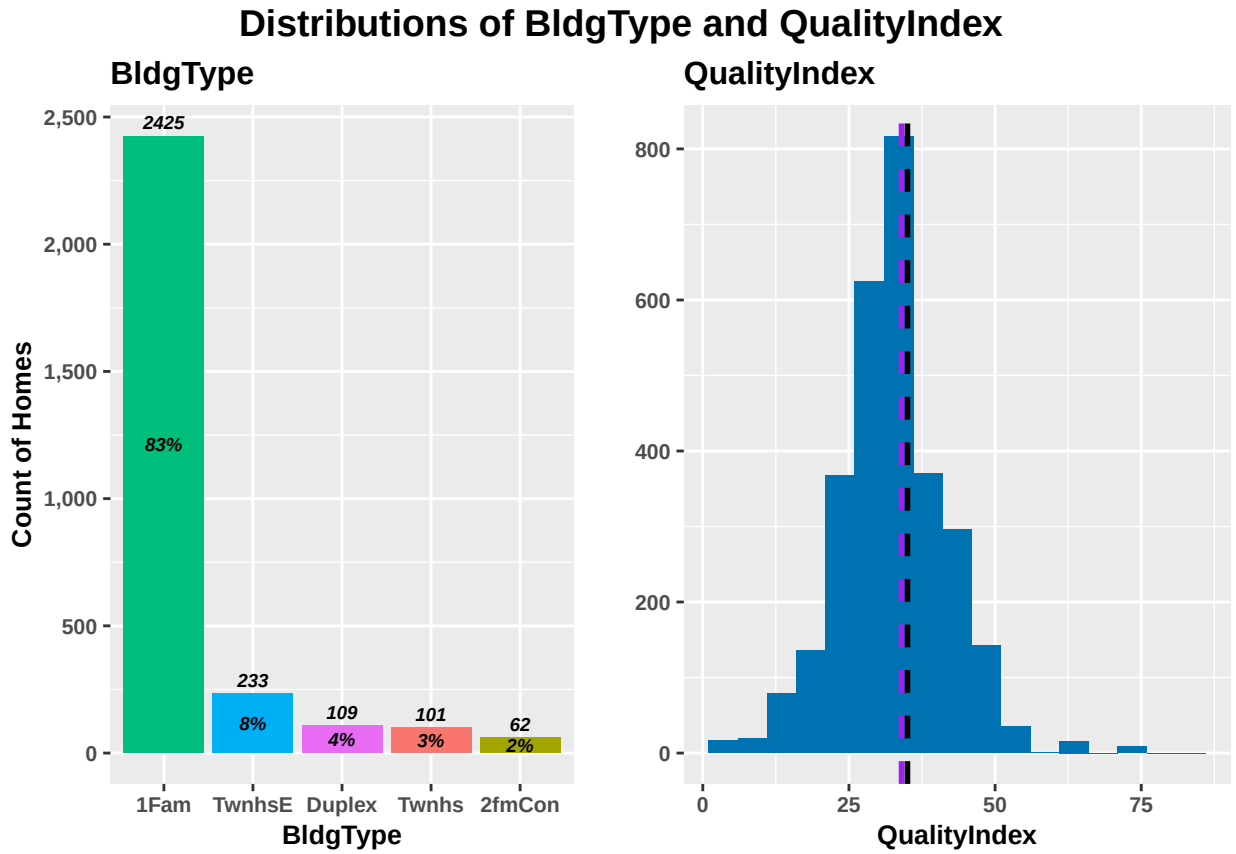
From the histograms above, we can see that there is a wide range of total square feet and the prices of homes sold. For instance, there are five houses that have a TotalFloorSF $>$ 4,000 and 6 houses that have a SalePrice $>=$ 600,000.

In using our summary statistics for TotalFloorSF, we can calculate the IQR and identify if we have any extreme outliers. In fact, we have 8 extreme outliers (TotalFloorSF $>$ 3600).

Similarly, using the IQR of SalePrice, we identify that there are 26 houses that are considered extreme outliers by SalePrice (SalePrice $>$ 465500). In fact, three of these homes also have a TotalFloorSF $>$ 3600.

In determining how to handle outliers, it is always best to have appropriate context for these data points so as not to remove inherent variation in the data. We will continue to explore relationships of variables in our data set and will keep in mind these outliers that we have identified, as they may be candidates for defining our sample data set. For now, we will include these outliers in our analysis.

Below, we examine distributions of two other variables: BldgType and QualityIndex.



We can readily see that most of the homes in our data set are single family detached homes (1Fam). We have much fewer homes of other building types.

This observation of distribution by building type is extremely important, as it raises the question if all building types should be included in our sample data set. It is also unknown if this distribution by building type is representative of Ames, IA. For example, there may be fewer non single family homes, and all of those were included in our study, but only a few single family homes are included. Regardless of the representativeness of this data, it is a known fact that generally the type of building impacts the sale price of a home. BldgType will be important variable to explore further to see if we should be limiting the types of buildings in our sample data set.

In viewing the distribution of QualityIndex, we can see that we have a wide range of values. Even still, the median (purple dashed line) and mean (black dashed line) values are fairly similar. Using a similar appraoch to above, we can calculate the IQR to determine if we have any extreme outliers. In fact, we have 11 homes that fit this criteria (QualityIndex > 70).

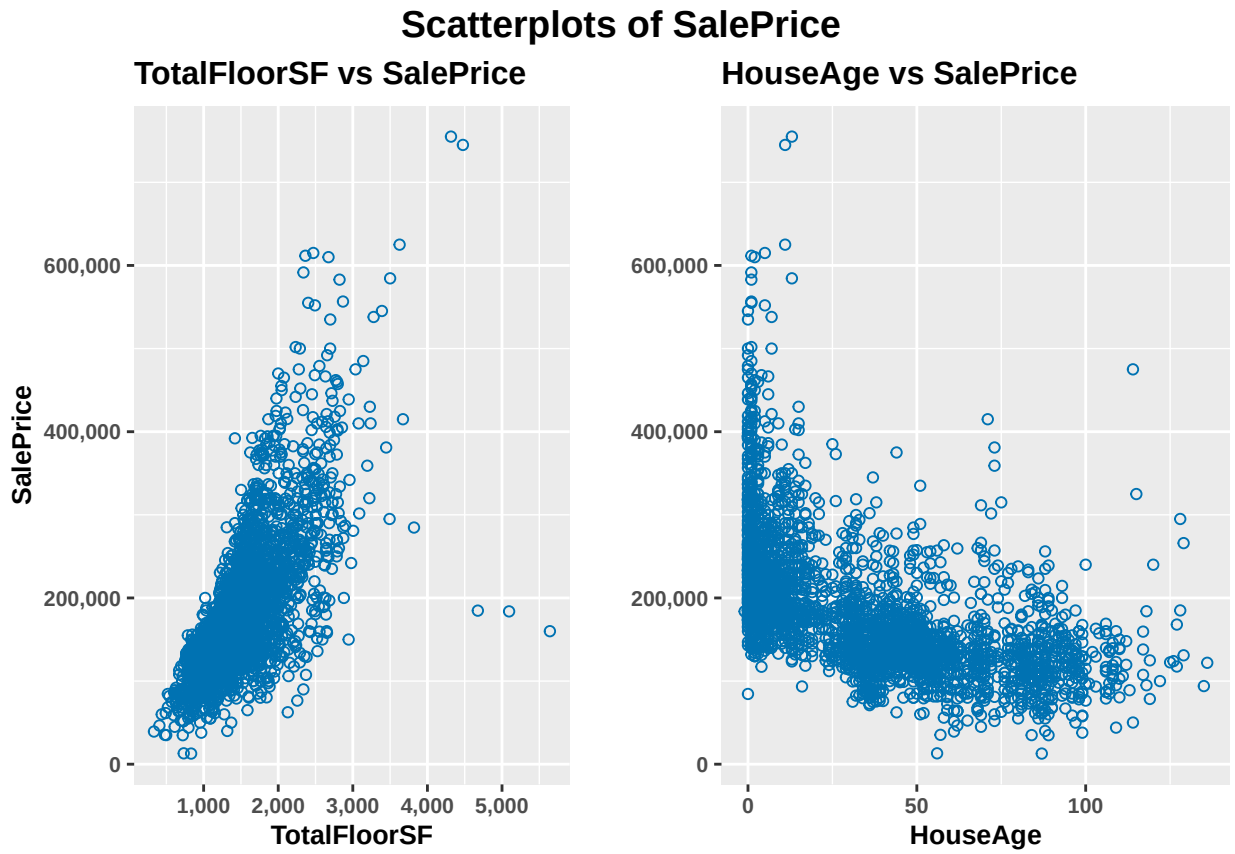
We will want to examine the relationships these variables have with one another and with other variables in our data set to determine the appropriate scope for our sample data set.

Section 2: Define the Sample Data Population

We will build upon the visuals and commentary from Section 1 by highlighting relationships between variables and offering further considerations for creating our sample data set.

For instance, we will want to understand how the numeric characteristics of the homes (e.g., total square feet, house age, lot area, etc.) vary by the factor variables (e.g., quality index, lot shape, etc.). This point is even more important because we are working with an observational data set, and we will want to incorporate environment variables into our analysis.

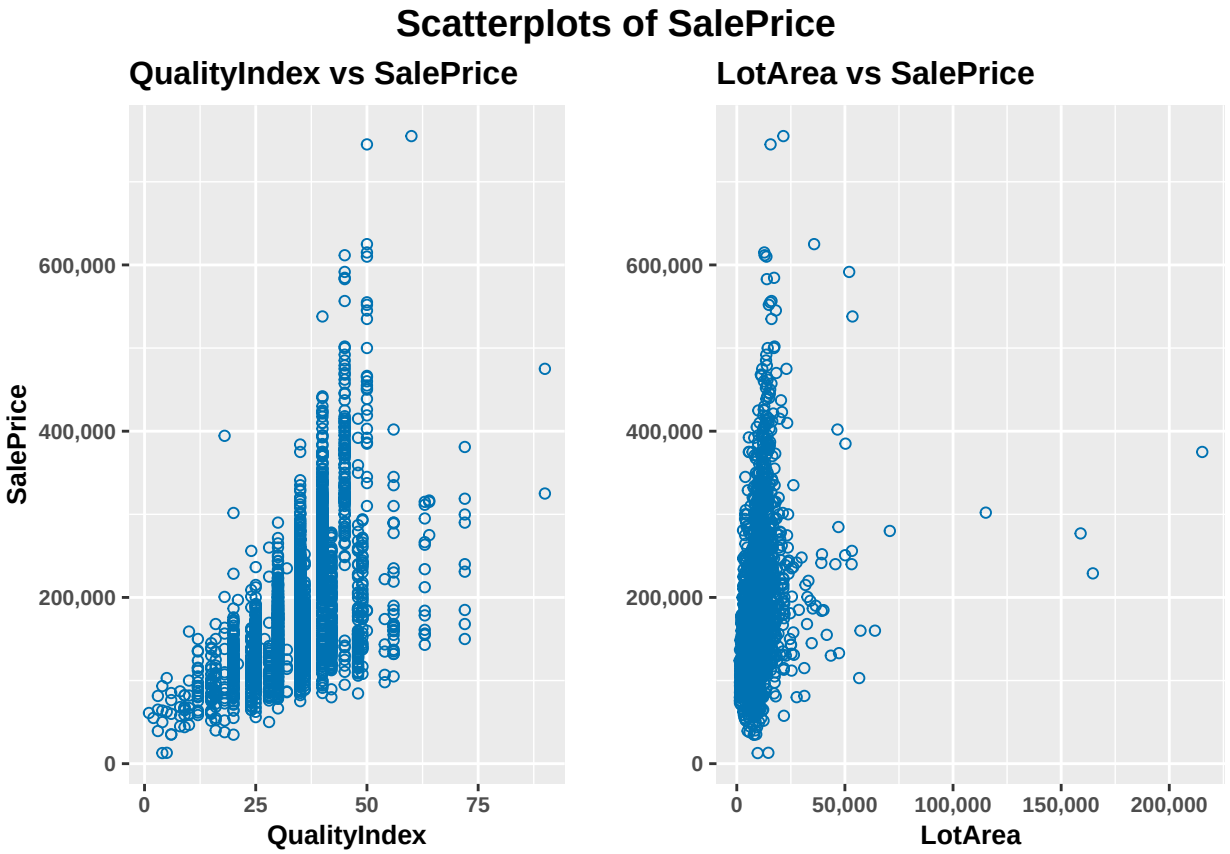
In this section, we will take a broad approach and explore various relationships with select variables. We begin by examining the relationships of SalePrice and TotalFloorSF as well as SalePrice and HouseAge.



First, we can readily see that we there is a wedge-shaped pattern between SalePrice and TotalFloorSF, also known as heteroscedasticity. This means that when TotalFloorSF doubles then SalePrice does not necessarily double, as there is increased variability in SalePrice in the upper range. This trend is extremely important for two reasons. One, it means that we will need to consider a transformation of SalePrice, since heteroscedasticity is a violation of a linear regression model. Second, this trend alludes to the fact that TotalFloorSF on its own does not fully explain all of the variability in SalePrice.

The scatterplot of SalePrice and HouseAge shows a negative and fairly linear relationship. That is, as HouseAge increases then SalePrice decreases. There appear to be some houses that depart from the normal trend in the 75+ HouseAge range.

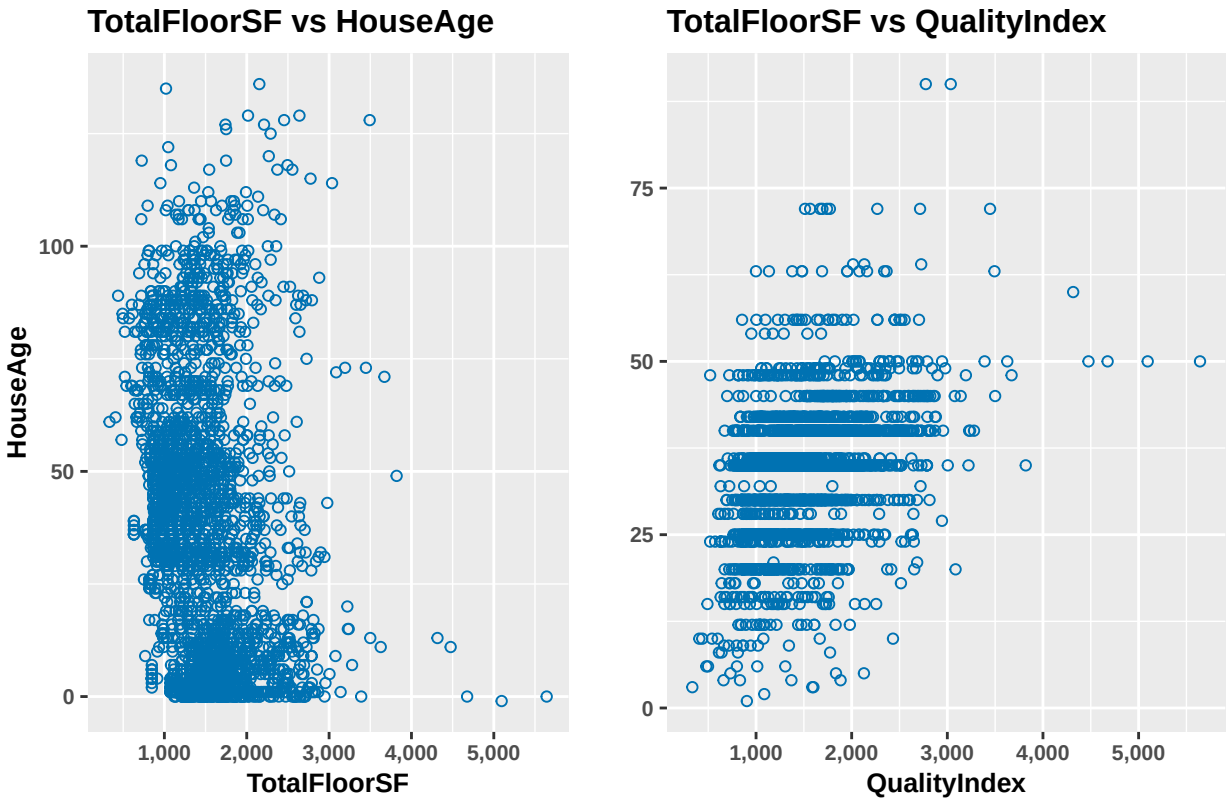
Next, we examine the relationships of SalePrice with QualityIndex and LotArea.



We have a moderately strong linear relationship between QualityIndex and SalePrice. We can see, similar to above, that there is greater variability in SalePrice as QualityIndex increases, further supporting that a transformation of SalePrice will be needed.

The relationship between LotArea and SalePrice does not appear to be linear, as most observations (homes) have lot areas that are less than 50,000. This trend makes common sense, as fewer homes should have large lot areas.

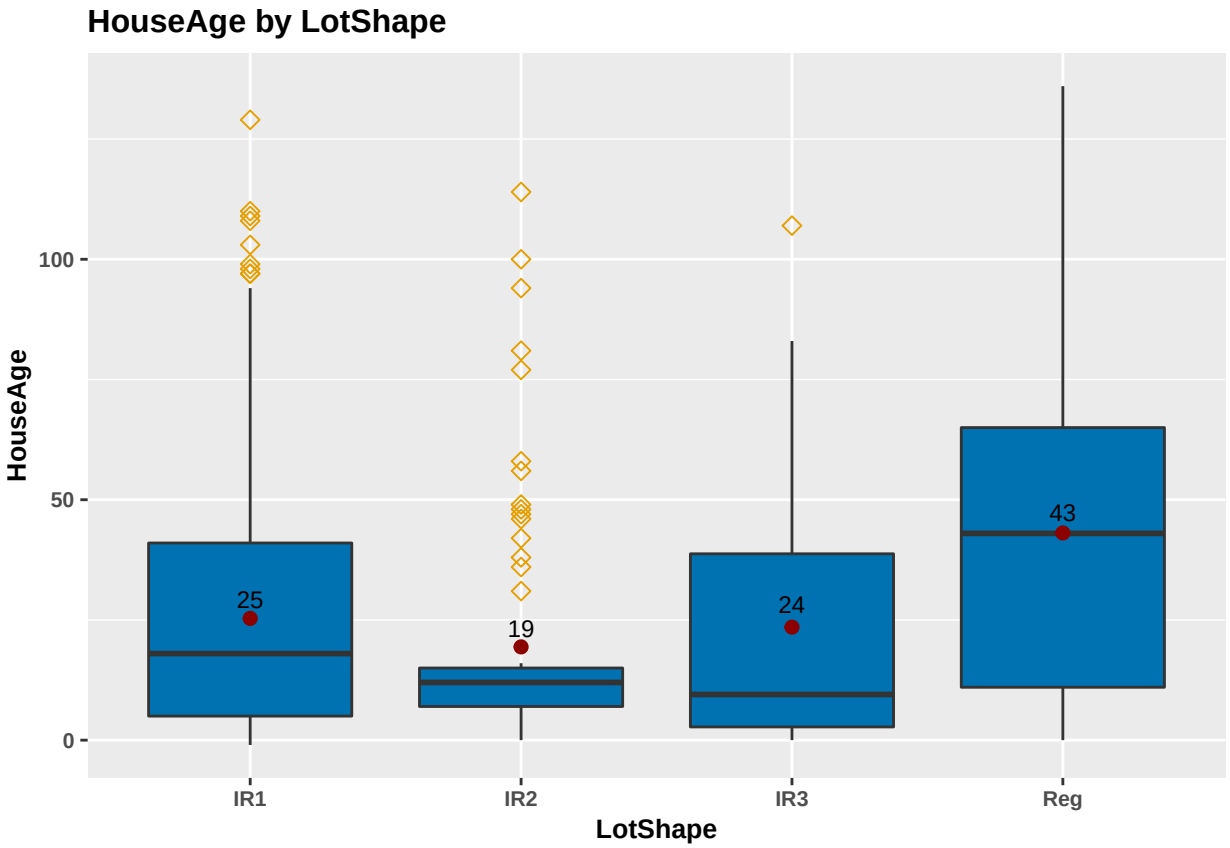
We now turn our attention to examining the relationships between TotalFloorSF and HouseAge in addition to QualityIndex.



TotalFloorSF and HouseAge have a weak negative linear relationship. This observation is good, however, as TotalFloorSF and HouseAge each have fairly strong linear relationships with SalePrice.

TotalFloorSF and QualityIndex have a moderately strong positive linear relationship. We can see that there is more variability in TotalFloorSF as QualityIndex increases, but the variability is not as great as the wedge-shaped patterns revealed by SalePrice and TotalFloorSF.

Lastly, we can analyze HouseAge by LotShape.



We can see that the age of a home varies by lot shape. We can see that Regular lot shapes are associated with older homes while non Regular lot shapes are associated with newer homes. One hypothesis that helps to explain this relationship might be that the type of building is confined to specific lot shapes that influence the age of the home.

Given all of this information, we are ready to provide some commentary on our sample data scope.

Sample Data Scope

Based on the visuals presented thus far, we have two approaches in forming our sample: 1) be narrow, or 2) be broad. For instance, we saw above that the key numeric characteristics of a house, sale price, house age, and total square feet, vary by building type. One approach in forming our sample could be to just focus on single family homes (which are also most prevalent in our data set) and remove these other building types. This approach would offer potentially greater confidence in translating this model to other single family homes but it would not enable generalizability to all homes in Ames. In contrast, if we were to be broad by incorporating all building types into our sample, we might have less accuracy in predicting the sale price of a home because of variability in our sample and with different building types. This model, even with this less accuracy, could be more generalizable to all Ames homes. There are pros and cons to both approaches.

For our purposes, we will elect to be broad and create a model incorporating all building types. However, we will utilize two drop conditions to narrow our data set:

1. We will remove houses with total square feets >= 4,000, and
2. We will remove houses with a sale price >= 600,000

The rationale behind these drop conditions is that we are wanting to predict the sale price of a home based on characteristics of that home for 'typical' Ames houses. Houses with these large square feets and/or large

sale prices are atypical and may negatively impact the regression model for other Ames houses. Therefore, we will remove these 9 houses and will proceed forward analyzing 2,921 houses.

It is also worth noting that we will continue to validate these drop conditions. As we move forward with our analysis, we may find that we need to reconsider our drop conditions and exclude some other data points.

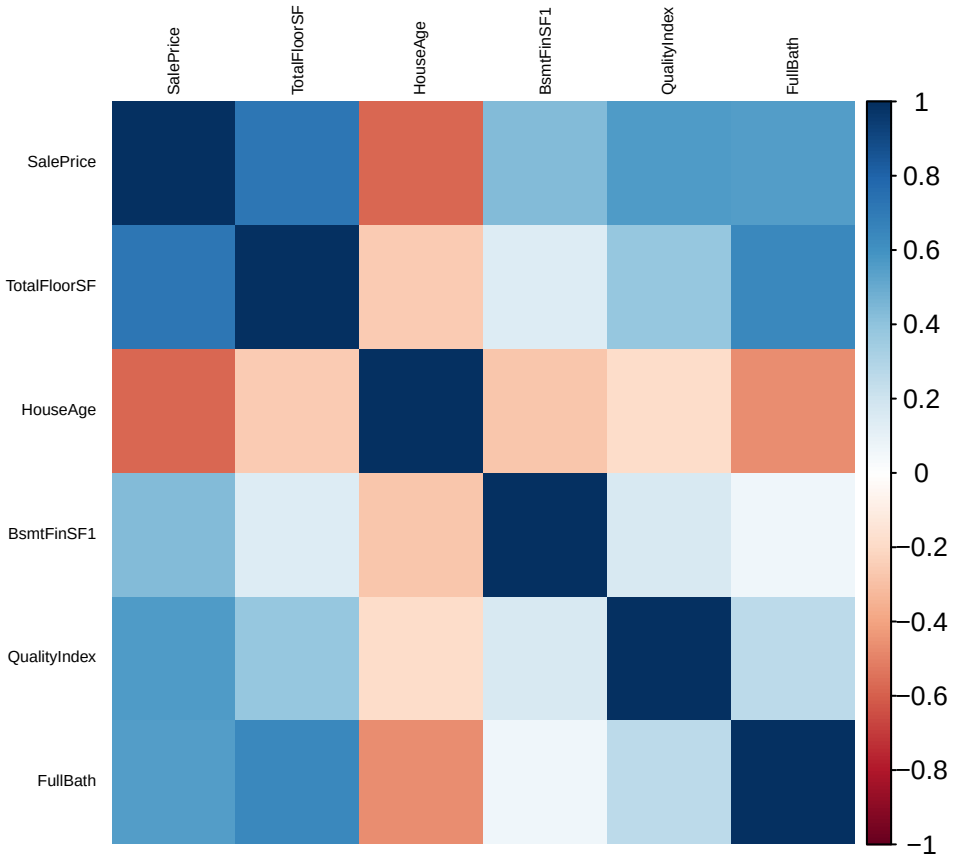
Section 3: An Initial Exploratory Data Analysis

For our exploratory data analysis, we will examine the following 10 variables. These 10 variables were chosen based on their relationship with SalePrice (shown above) or from experience on some factors that may impact the sale price of a home.

1. SalePrice
2. TotalFloorSF
3. HouseAge
4. BsmtFinSF1
5. LotShape
6. QualityIndex
7. BldgType
8. HouseStyle
9. Neighborhood
10. FullBath

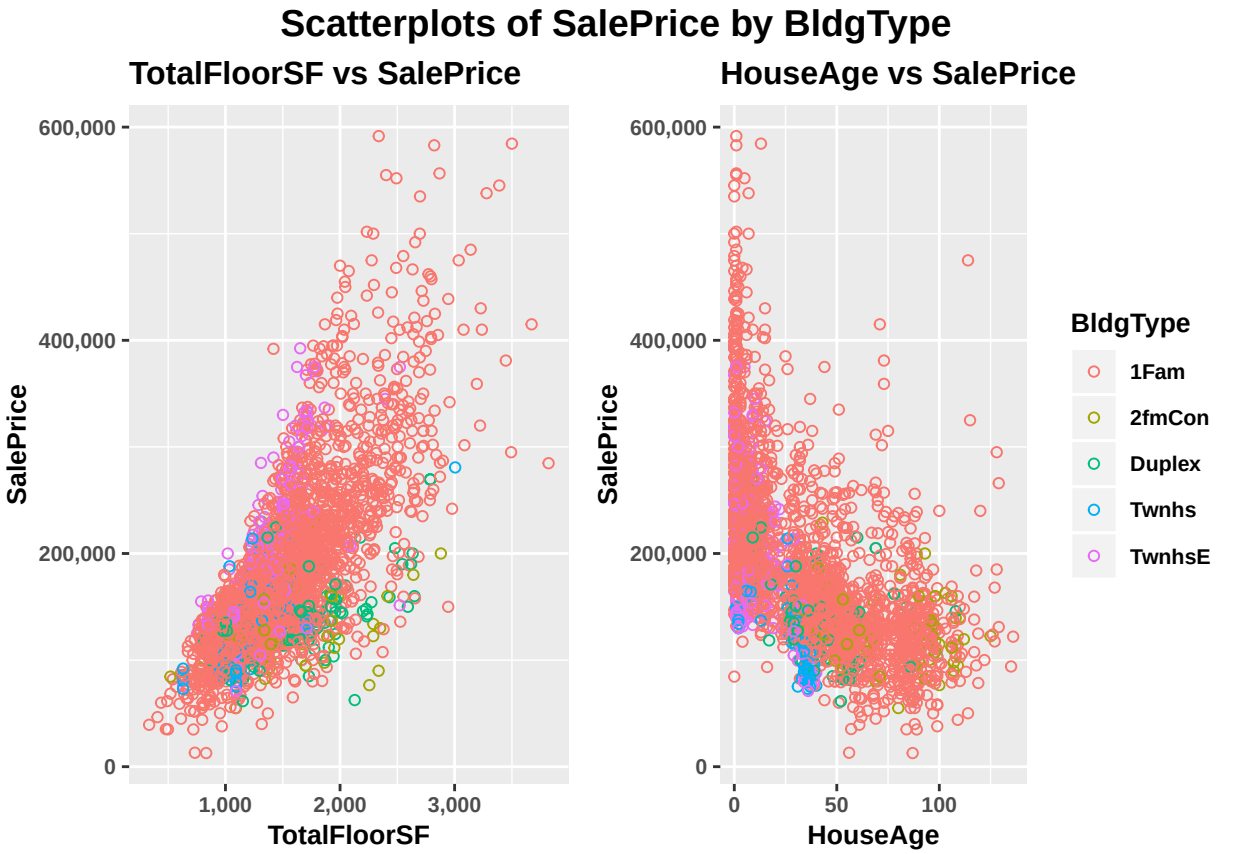
In this section, we will focus on the relationships of variables with SalePrice, as SalePrice is our response variable for our regression model.

We start our analysis by focusing on our numeric variables. We generate a correlation plot to better understand the strength of the linear relationships between these variables, as we already have some understanding of the shapes of these relationships from prior sections.



We see strong linear relationships between SalePrice and TotalFloorSF (positive) as well as between SalePrice and HouseAge (negative). QualityIndex and SalePrice seem to have a moderately strong positive linear relationship. For the remaining relationships, there is nothing of note as those linear relationships are either weak or less than .50.

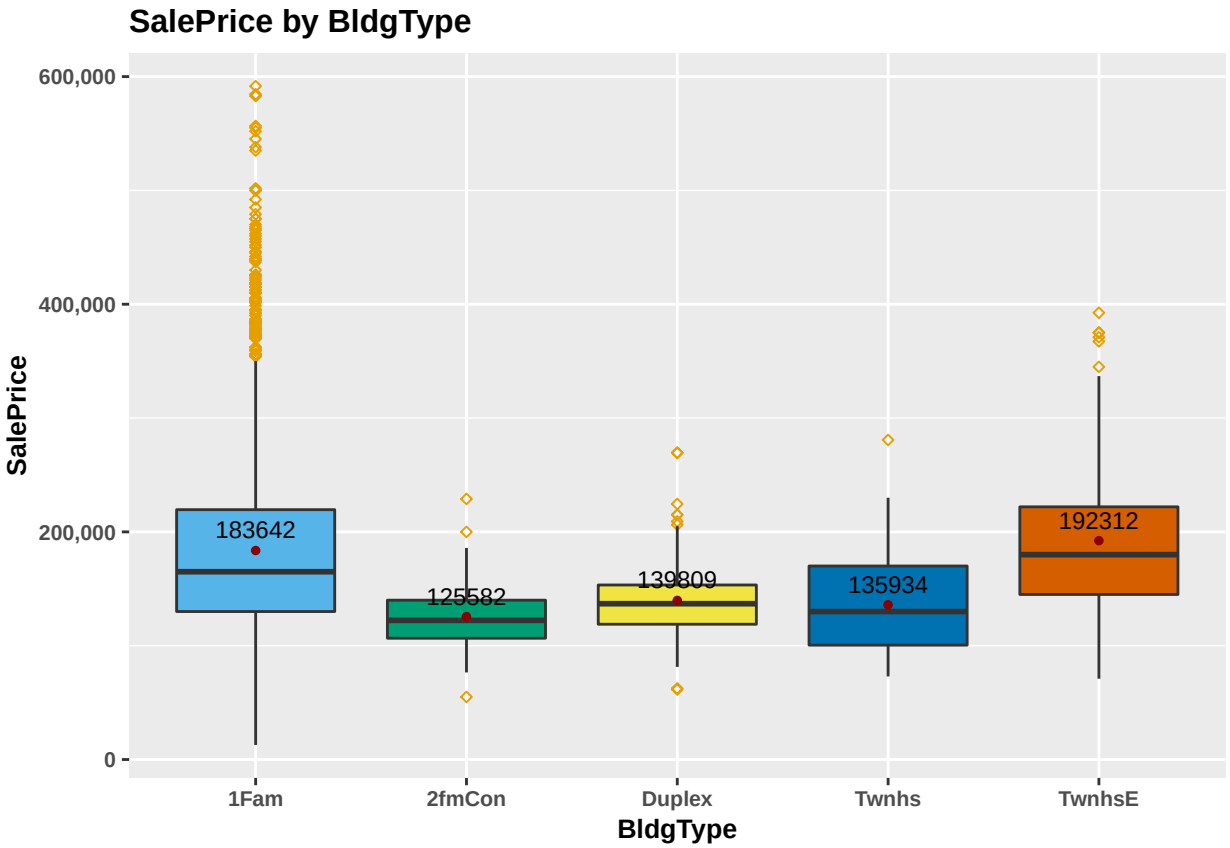
Now, we turn our attention to focus on how our numeric variables vary by our factor variables. To begin, we examine the relationships of SalePrice and TotalFloorSF as well as SalePrice and HouseAge, as these two relationships were the most strong in the correlation matrix.



We can see that there are differences in these relationships by building type. For instance, in the scatterplot of SalePrice and TotalFloorSF we can see that the purples are primarily on the upper left of the plot, while the greens and golds are on the bottom of the plot. This pattern tells us that Townhome end units have a smaller range for TotalFloorSF but are associated with higher sale prices. Additionally, Duplex and Two Family conversion homes are typically larger, but their SalePrice is on the lower end of the range.

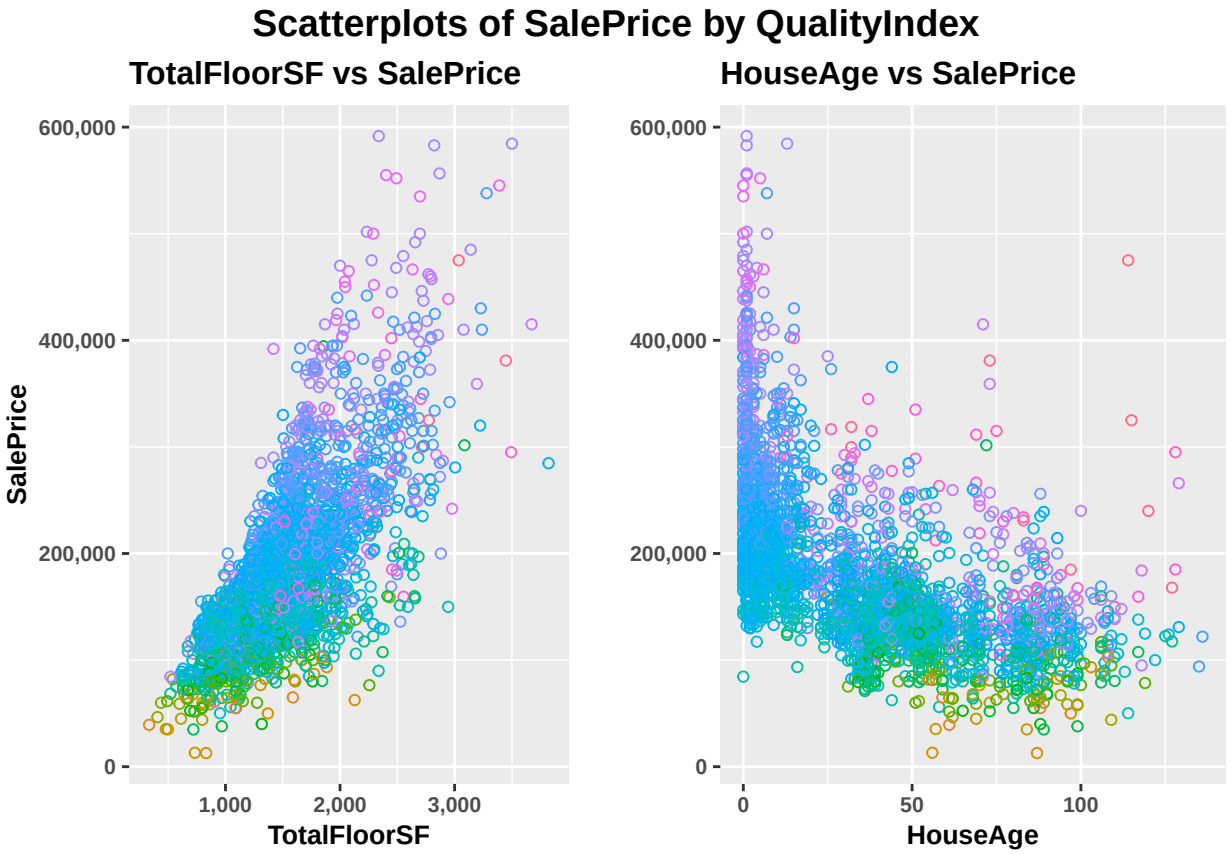
Moreover, in examining the scatterplot of SalePrice and HouseAge, we can see that the purples are almost in a vertical line, indicating that Townhome end units are younger in age. We also see the presence of blue dots at the bottom left, which also shows us that Townhome inside units are also younger in age. For every other building type, there appears to be variability in the relationship between SalePrice and HouseAge.

We analyze boxplots of SalePrice by BldgType for a different look at this relationship.



We can see that there are some slight differences in SalePrice by BldgType. The middle building types all have similar median and mean values as well as ranges. However, the Single Family homes and Townhome end units have similar median and mean values as well as ranges. It might be worthwhile to collapse these building types once we get to the modeling stage. It is also worth noting that Single Family homes have the most outliers out of all building types.

Next, we explore how SalePrice and its relationships with TotalFloorSF and HouseAge vary by QualityIndex, as a factor.

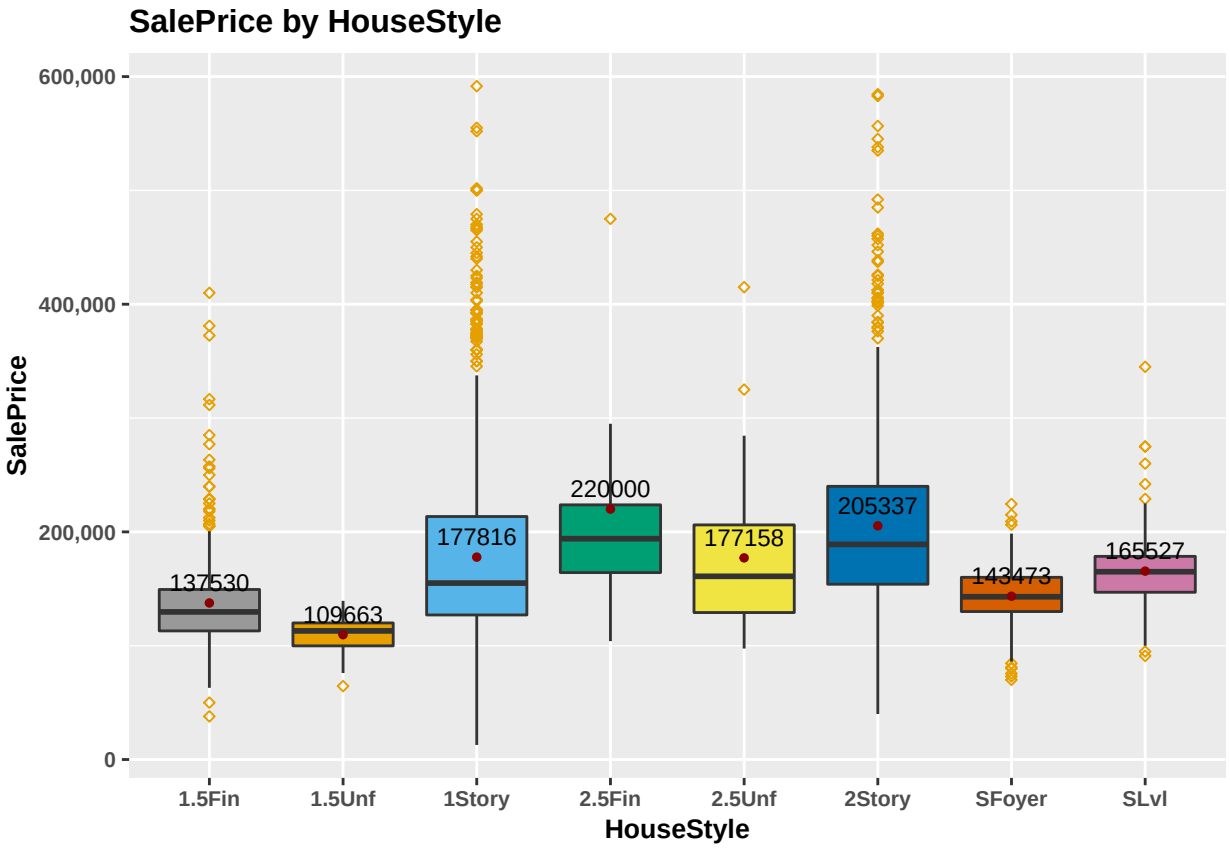


For readability purposes, the QualityIndex scale is not displayed. Below is a rough guideline on colors:

- Red and yellow colors denote scores < 10;
- Green denotes scores from 10 - 25;
- Blue denotes scores from 25 - 40;
- Purple denotes scores from 40 - 50;
- Pink denotes scores > 50

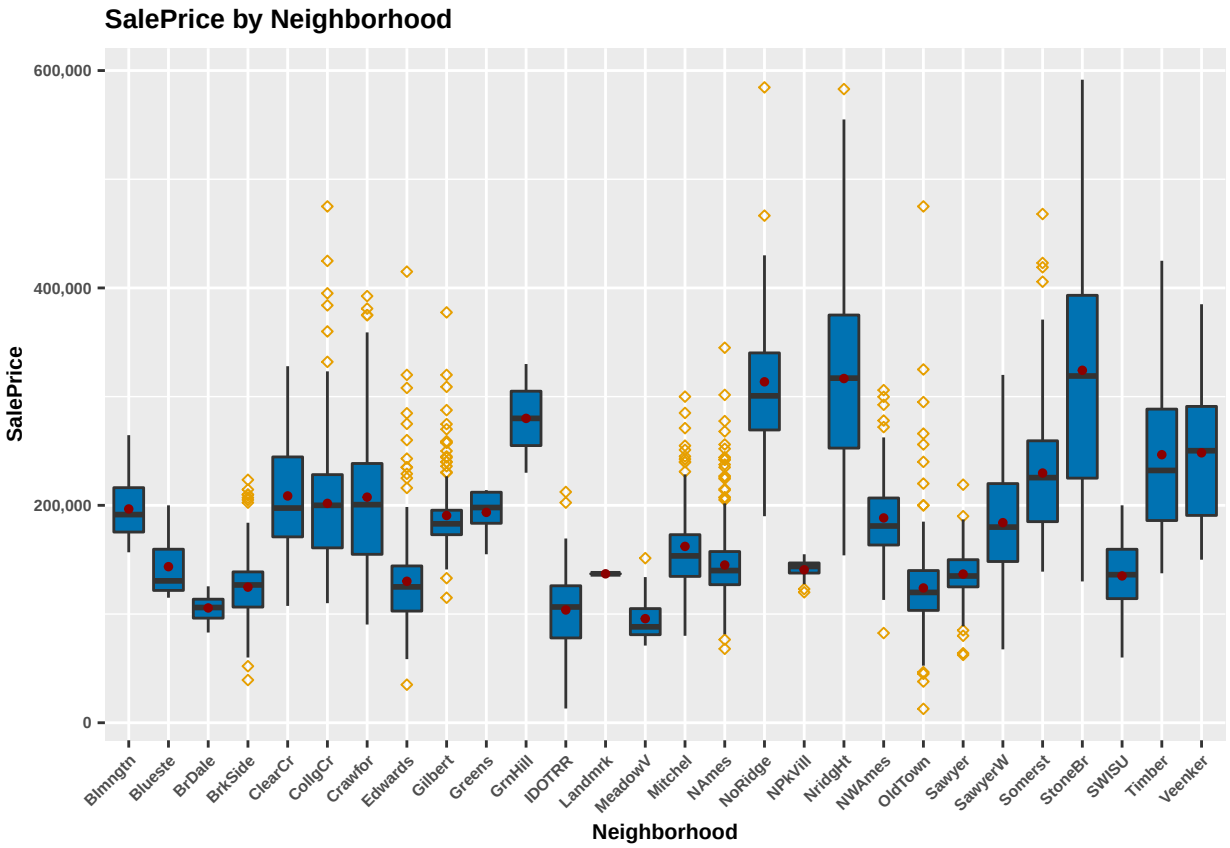
We can see that overall these relationships do vary by QualityIndex as there is clear layering of colors. We can see that the higher quality (i.e., purples, pinks) is associated with more TotalFloorSF and a higher SalePrice. However, we can also see that higher quality has a wide range of HouseAge, as the purple and pink dots are all along the upper ridge of the scatterplot. One hypothesis for this trend is there might have been some renovation or additions on these older homes.

We continue our analysis by analyzing boxplots of SalePrice by our factor variables, starting with HouseStyle.



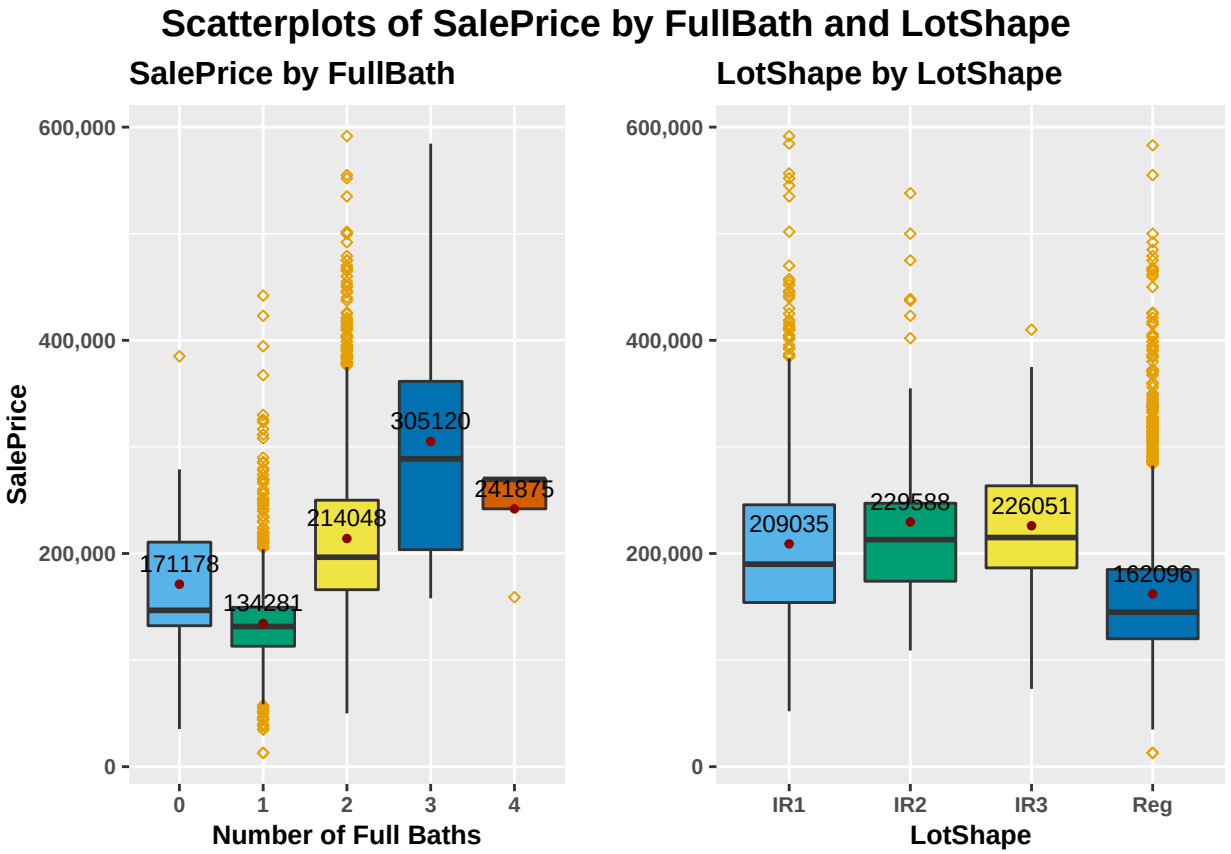
While there are differences in SalePrice by HouseStyle and many outliers within each style, overall, there are not large differences in median and median SalePrice values by HouseStyle. The largest differences are between the 1.5Fin/1.5Unf house styles compared to others. The one story and two story house styles are all comparable, as are the SFoyer and SLvl styles. There may be opportunity when modeling to collapse these house style factor levels into fewer levels but this would have to be explored further.

We continue our analysis by viewing boxplots of SalePrice by Neighborhood.



Similar to the boxplots of SalePrice by HouseStyle, we can see that there is variability in SalePrice by Neighborhood. The StoneBr, NridgeHt, and NoRidge neighborhoods all boast sale prices in the upper ranges. We can even see that the StoneBr neighborhood has a much wider range in SalePrice compared to other neighborhoods. In contrast, the MeadowV and BrDale neighborhoods have lower sale prices.

Below, we explore the relationship of SalePrice by the number of fullbaths and lot shape.



We can see that there are differences in SalePrice by FullBath. What is surprising is that homes with 0 full baths have a higher median and mean SalePrice than homes with 1 full bath. However, there is a difference in SalePrice when homes have 2-4 full baths. The range in SalePrice is widest when a home has 3 full baths.

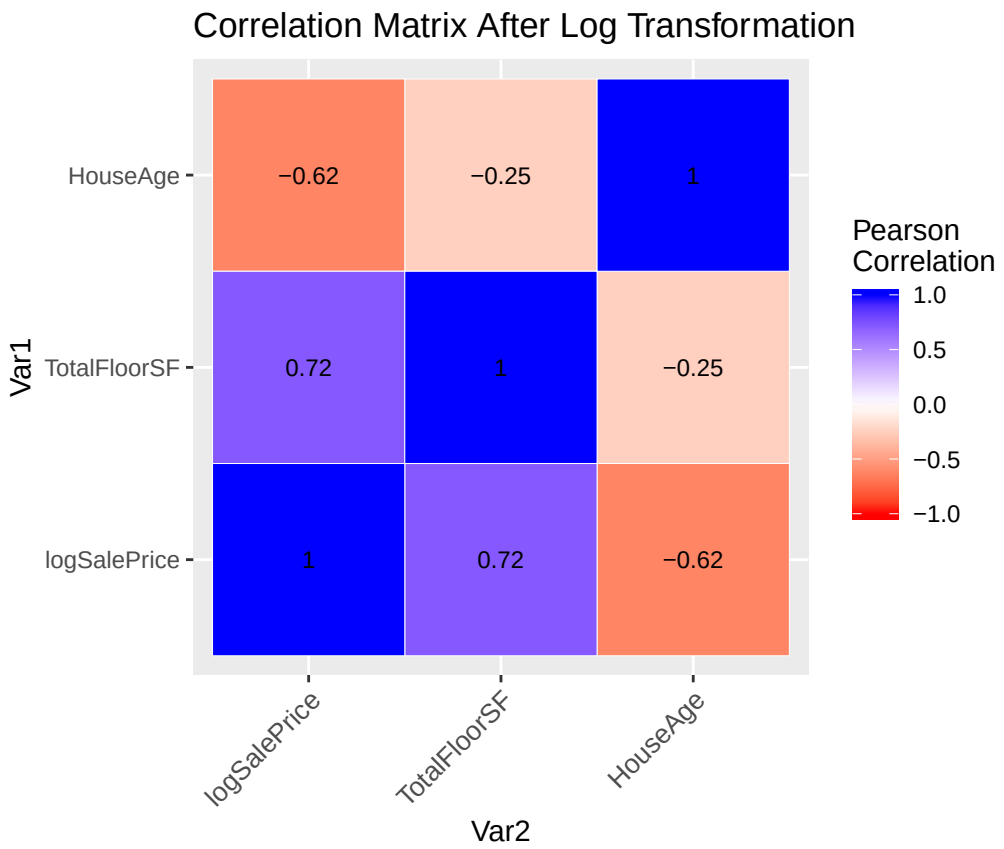
Overall, we see a difference in SalePrice when the LotShape is Regular vs Not Regular, as the median/mean values are different. Moreover, the range of SalePrice for Regular vs Not Regular lot shapes do not overlap. Even still, all lot shapes have outliers that require a greater understanding in how best to handle.

Section 4: An Initial Exploratory Data Analysis for Modeling

In this section, we explore the relationships of TotalFloorSF, HouseAge, and BldgType with SalePrice and logSalePrice. We select these variables as both TotalFloorSF and HouseAge have fairly strong linear relationships with SalePrice and SalePrice varied by BldgType.

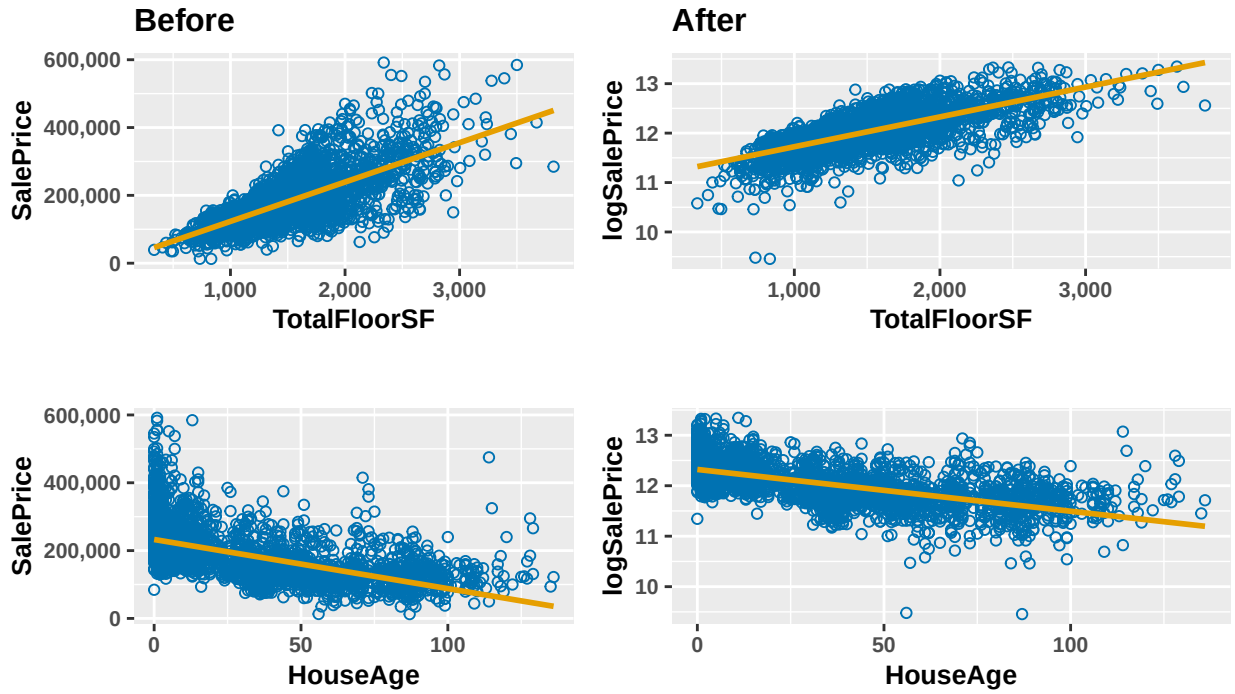
Since we showed visuals with SalePrice in prior sections, we will focus more on the relationships these three variables have with logSalePrice. Visuals of SalePrice will be shown to provide additional context where appropriate.

We begin by examining a correlation matrix with logSalePrice.



We can see that applying the log transformation to SalePrice has not negatively impacted any of the correlations. In fact, the relationship between logSalePrice and HouseAge is improved (before = -.57, after = -.62).

We next explore the shape of the relationships between logSalePrice, TotalFloorSF, and HouseAge.

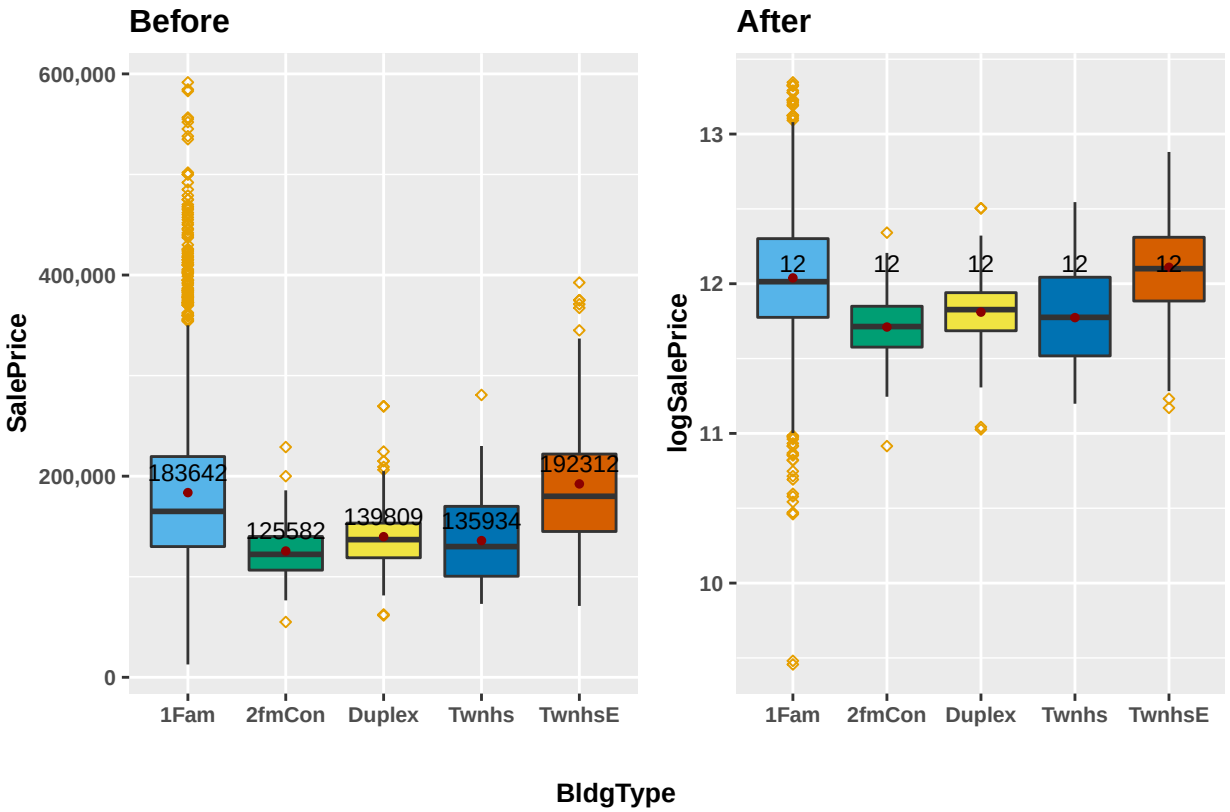


By having the plots before the log transformation of SalePrice for comparison, we can see that the fit of the linear regression lines is improved after the log transformation. For instance, before the transformation, the regression lines are not fully in the middle of the data cloud. There is greater variability above these lines for both plots, which denotes that the regression lines are not appropriately accounting for all the variability in these relationships.

In contrast, after the log transformation, the linear regression lines go through the middle of the data cloud, and there is less variability above and below the lines. There still appears to be some potential outliers that have a logSalePrice < 10 that might need to be addressed.

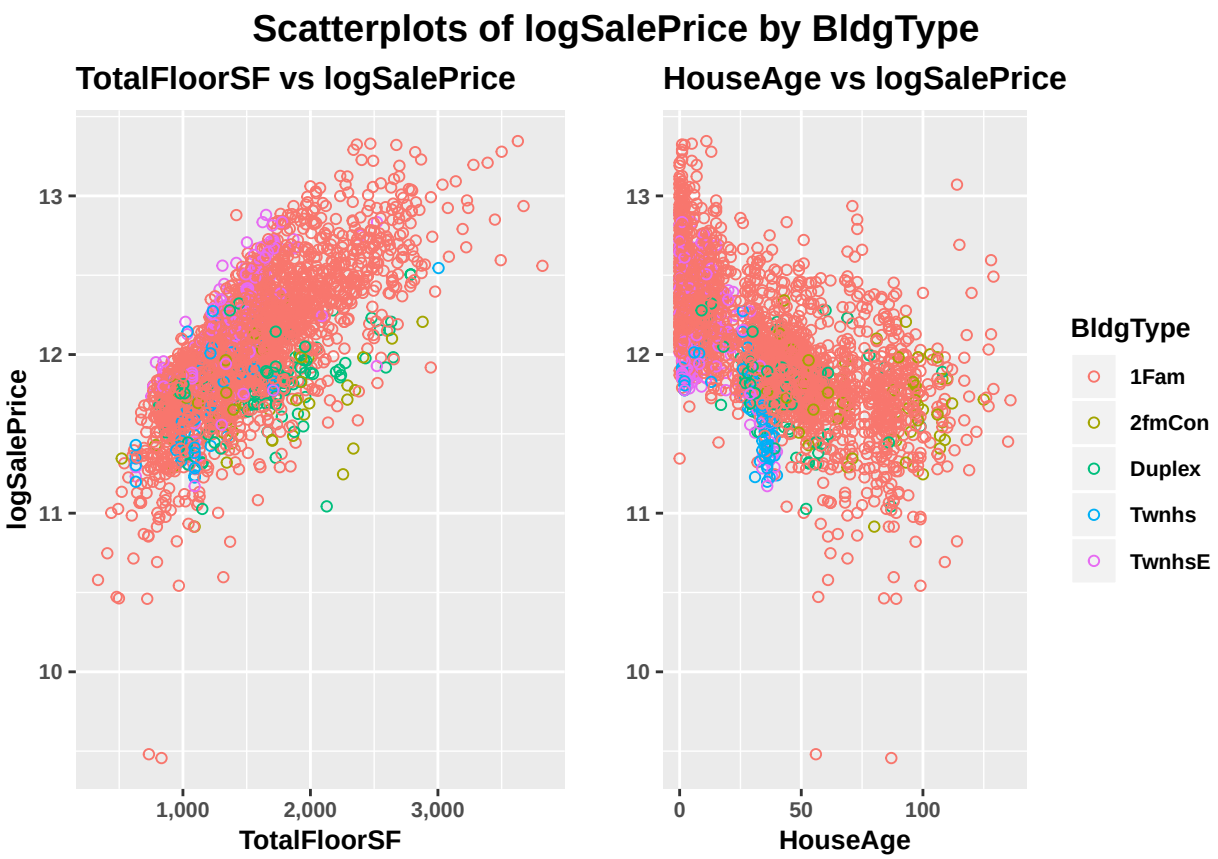
These comparisons suggest that keeping an eye out for wedge-shaped patterns with relationships is a key part of the exploratory data analysis process; it also has the potential to impact which variables are included in the regression model.

We continue our analysis by examining logSalePrice by BldgType.



We can see that the log transformation of SalePrice has improved the spreads of SalePrice by BldgType. The difference in SalePrice by BldgType still remains after the log transformation but the outliers are more evenly distributed to the lower and upper ends, instead of being concentrated toward the upper limits.

Lastly, we will examine the interaction of SalePrice, TotalFloorSF, and HouseAge by BldgType.



The trends shared in Section 2 still hold true.

We see a concentration of Townhome end units in the upper left of the scatterplot with TotalFloorSF and a concentration of Townhomes in the lower left of the scatterplot with HouseAge. This pattern suggests that Townhown end units are higher in price and generally Townhomes are younger in age.

We also see a concentration of Duplex and Two Family conversion homes in the lower right of the scatterplot by TotalFloorSF. Duplex homes are tied to lower ages, but there appears to be a wider age range For Two Family conversion homes.

Single Family homes follow a linear relationship and demonstrate a spread in logSalePrice and HouseAge.

Summary and Conclusions

In this assignment, we analyzed data from the Ames housing data set. We focused our analysis on exploring the relationships of both numeric and factor variables by our predictor variable, SalePrice.

Through our exploratory data analysis, we defined typical Ames houses as houses that had less than 4,000 total square feet and sale prices under 600,000. We saw that SalePrice and TotalFloorSF had a wedge-shaped pattern, which suggested that SalePrice required a transformation. We also saw that SalePrice varied by the building type, number of full baths, and neighborhood and that some of these factor variables could possibily be collapsed into fewer factors if incorporated into a regression model. This last point requires validation but is an idea for continued analysis.

After applying a log transformation to SalePrice, we could see that we still had fairly strong linear relationships with our key numeric variables, TotalFloorSF and HouseAge, and it differed by BldgType.

There are some considerations for improving our analysis and increasing the accuracy of predicting sale price. For one, it would be helpful to have an understanding of the inventory of Ames houses by building type. This

sample mostly has single family homes, but other building types are included to aid in the generalizability of this analysis. However, it is unclear if the distribution by building types is truly representative of the Ames housing market.

Moreover, through the analysis, it is readily apparent that there are several outliers (SalePrice $>$ 465500). Because it is unknown if these homes truly reflect variability in the data/Ames housing market or if these homes are errors/contain true differences from other houses, these homes are included in the analysis. Having additional context would help validate that these outliers are being handled appropriately. For our purposes, we are including most of these outliers in our data sample so as not to introduce unncessary bias.

Lastly, we may consider to incorporate when the home was built, as it could be important to better understand how these variables differ when dealing with a new construction home. Generally, houses that are newer are higher in price. This might be a more effective than building type, but that analysis is beyond the scope of this assignment.

Appendix

Task 3: A Data Quality Check

For our data quality check, we select the following 20 variables:

1. SalePrice
2. TotalFloorSF
3. PriceSqFt
4. HouseAge
5. LotArea
6. LotShape
7. QualityIndex
8. ExterQual
9. BldgType
10. HouseStyle
11. Neighborhood
12. YearBuilt
13. FullBath
14. BsmtFinSF1
15. BsmtFinType1
16. Fireplaces
17. GarageType
18. GarageFinish
19. PoolArea
20. PoolQC

Data quality is an important first step to perform before any analysis. In this step, we look for values that do not align to the data dictionary, missing values, and values outside of the range.

Here is a summary of some of our findings:

Comments on Data Values

- There is a home with a HouseAge of -1, which indicates that it was sold before it was built. There are also many homes that have a HouseAge of 0, which indicates new construction homes.
- There is 1 home with a missing value for BsmtFinSF1. This value would have to be imputed before moving forward. One would assume that a missing value denotes a 0.

- There are 157 NA's for GarageType which indicates "No Garage." There are also 157 NA's for GarageFinish, which validates that there are no missing values for Garage.
- PoolArea has 0 values. However, in this context, the 0 is associated to "No Pool" so when viewing summary measures for PoolArea, the 0 is included. This observation is something to be aware of when working with PoolArea. These 0 values would want to be removed so better summary measures could be obtained for when a pool is present.

Comments on Data Representation

- There are only 16 homes with a LotShape of IR3.
- There are 11 homes with a QualityIndex > 70 (which is more than 3 * IQR).
- There are only 8 homes that have a HouseStyle of 2.5Fin and 19 homes that have a HouseTyle of 1.5Unf.
- There is only 1 home from Landmrk, 2 homes from GrnHill, 8 homes from Greens, and 10 homes from Blueste neighborhoods.

All of these points suggest that there may be some concerns with the representativeness of this data. Because there are so few values that fit the above criteria, it is unclear if the conclusions drawn using these homes are truly valid.

We also examined if there were any columns with all missing values, and we got a result of 0.