DSC 465 Final Project Report
Driving Factors of Housing Prices

1. Introduction:

The overall objective of this analysis is to explore housing sale data and investigate how the key property and land factors impact the sales price of houses in Ames, Iowa between the years of 2006 and 2010. Through our analysis we will determine how different features, qualities, and conditions correlate to the sales price of the property. These features include variables describing lot side, internal features and quality, external features and quality and the type of building and neighborhood its located in. We investigated how key features like Year Built, Neighborhood, Condition of the property and Quality of materials, Basement and Basement Quality, Garage and Garage Quality, and Overall Quality, Dwelling Type, Lot Area, Building Type are correlated to Sales Price. Our data was compiled by Truman State University for use in data science education and was available on Kaggle.com as part of a competition to predict sales prices using machine learning algorithms most accurately. Since the competition was designed to allow participants to train and test a model this analysis utilizes 1460 records from the training data set which contained the Sale Price whereas the unused test data set did not include Sale Price.

2. Data Overview:

The full data set included 2930 observations with 81 total variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) which are involved in assessing housing prices in Ames, Iowa. Of these key measures we focused on how Year Built, Neighborhood, Condition of the property and Quality of materials, Basement and Basement Quality, Garage and Garage Quality, and Building Type, and Lot Area are correlated to Sales Price.

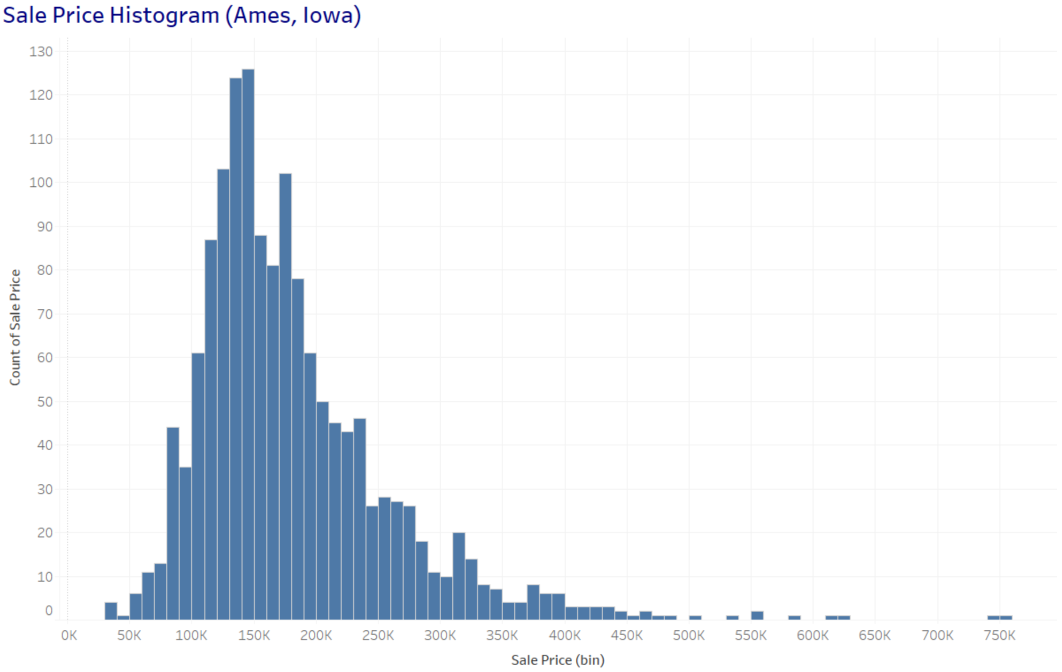
However, the dataset utilized for this project did impose some limitations on our analysis and visualizations. The primary limitation was that the test dataset provided in the competition did not contain values for the Sale Price variable making those 1460 records unusable in this study and cutting our overall housing market population in half. Additionally, while almost all columns are complete and matched appropriately with the respective house id record, the CentralAir field was not accurately mapped and thus could not be used for this project.

Ames Iowa, Dataset URL:
https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview.

3. Exploratory Analysis

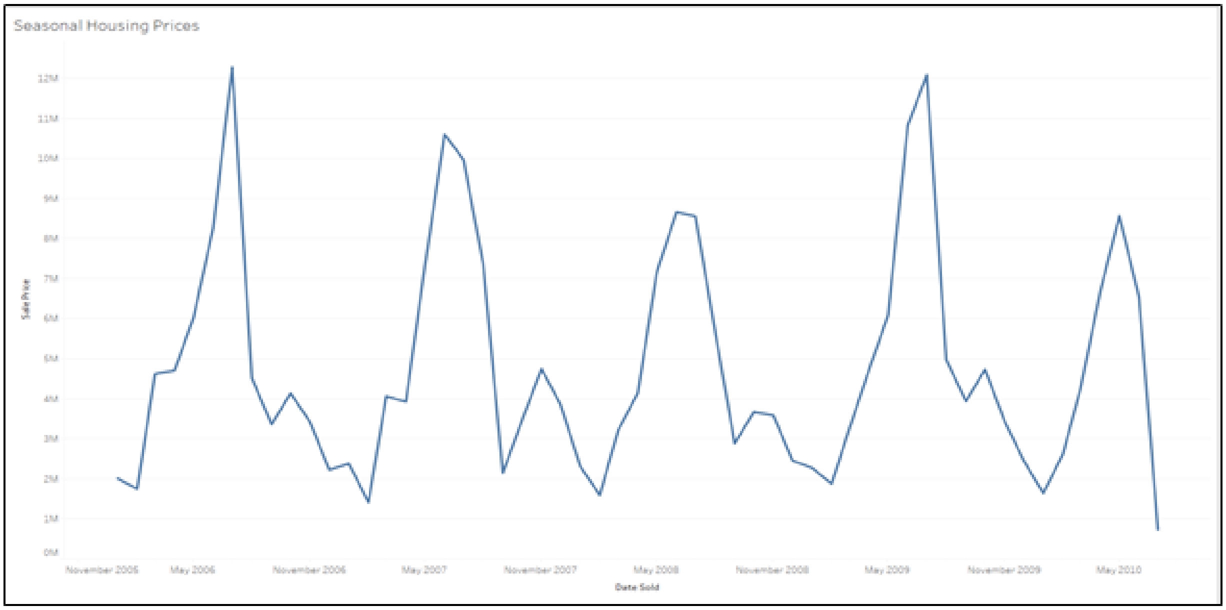
Data Exploration of Ames, Iowa housing dataset features using summary statistics.
Exploratory Visualization: Please see the link below for all the exploratory visualization.
https://documentcloud.adobe.com/link/review?uri=urn:aaid:scds:US:158279e9-aabd-4de4-b8dad9ba3d8abd47

In addition to the exploratory analysis done in Adobe, we can also see the distribution of sales prices across the dataset. As expected, the data shows a normal distribution with only some outliers for higher sales price.



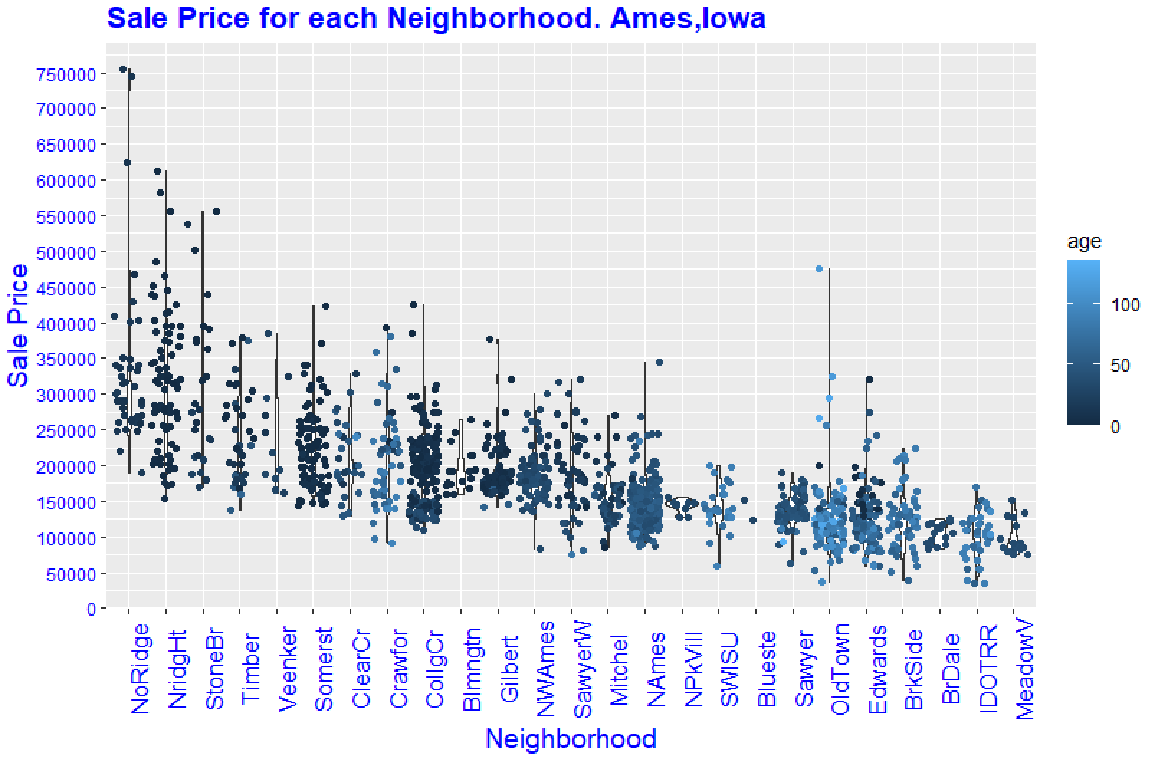
The trend of count of Sale Price for Sale Price (bin). The view is filtered on Sale Price (bin), which includes everything.

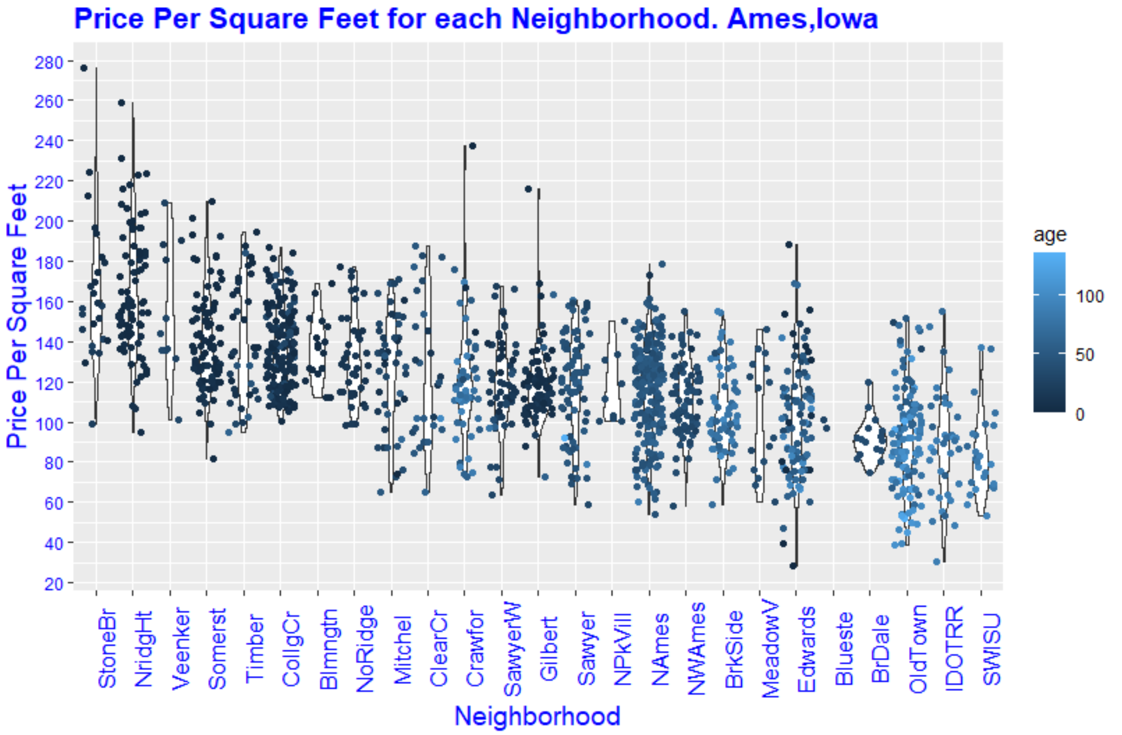
An initial easy plot below shows the seasonality of the sale price. There is a clear pattern showing increased prices in the summer months. This is shown in more detail in a later section.



4. Visualizations

Impact of location, square feet (above ground living area) and age of the properties on sale price.

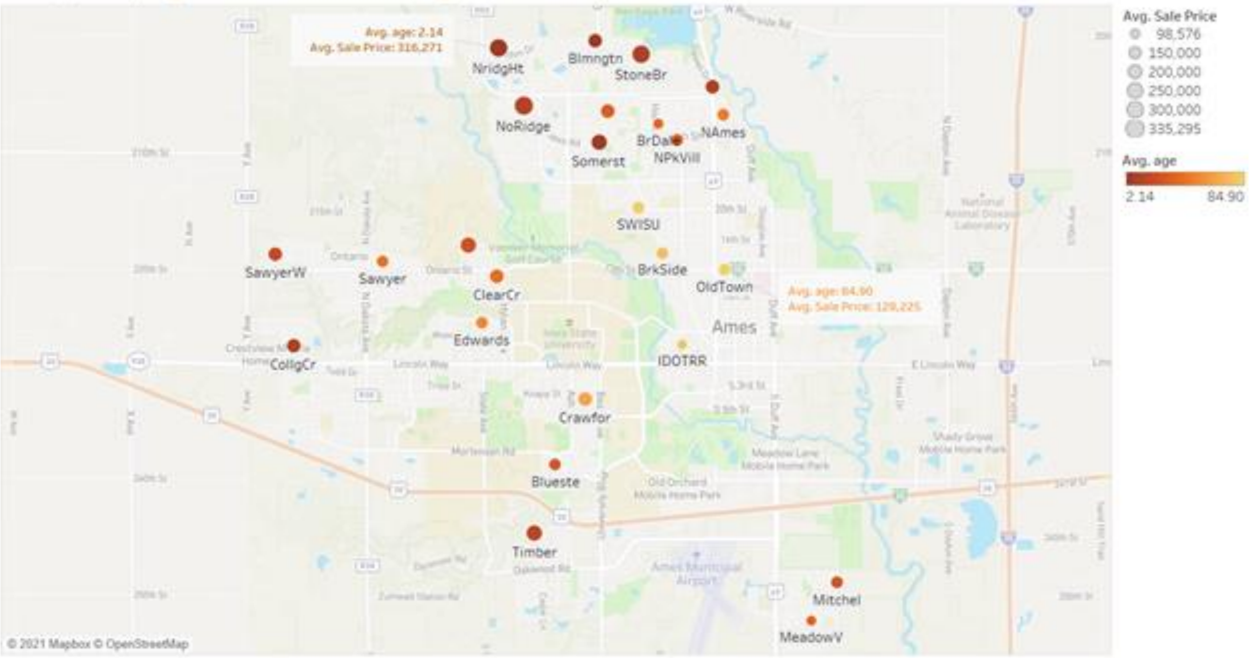




The above two visualizations allow us to analyze which neighborhoods have higher prices per square foot, sale price and allows us to see if there is a clear relationship between the age of the properties in the neighborhoods and their prices. As we can see, there appears to be some correlation between the age of the home and the price of the property with older homes having lower prices. Visualizations also allow us to see which neighborhoods consist of generally older or newer properties. We can see that Stone Brook (StoneBr) and Northridge Heights (NridgHt) have the newest properties and highest price per square feet and sale price. Old Town (OldTown) has some of the oldest properties and lower prices per square feet and sale prices. Many other attributes such as type of building, basement quality, basement condition, lot area, etc. affect the price of the properties. Hence, we see a wide distribution of values in each neighborhood. Some of these attributes we will further explore in this project.

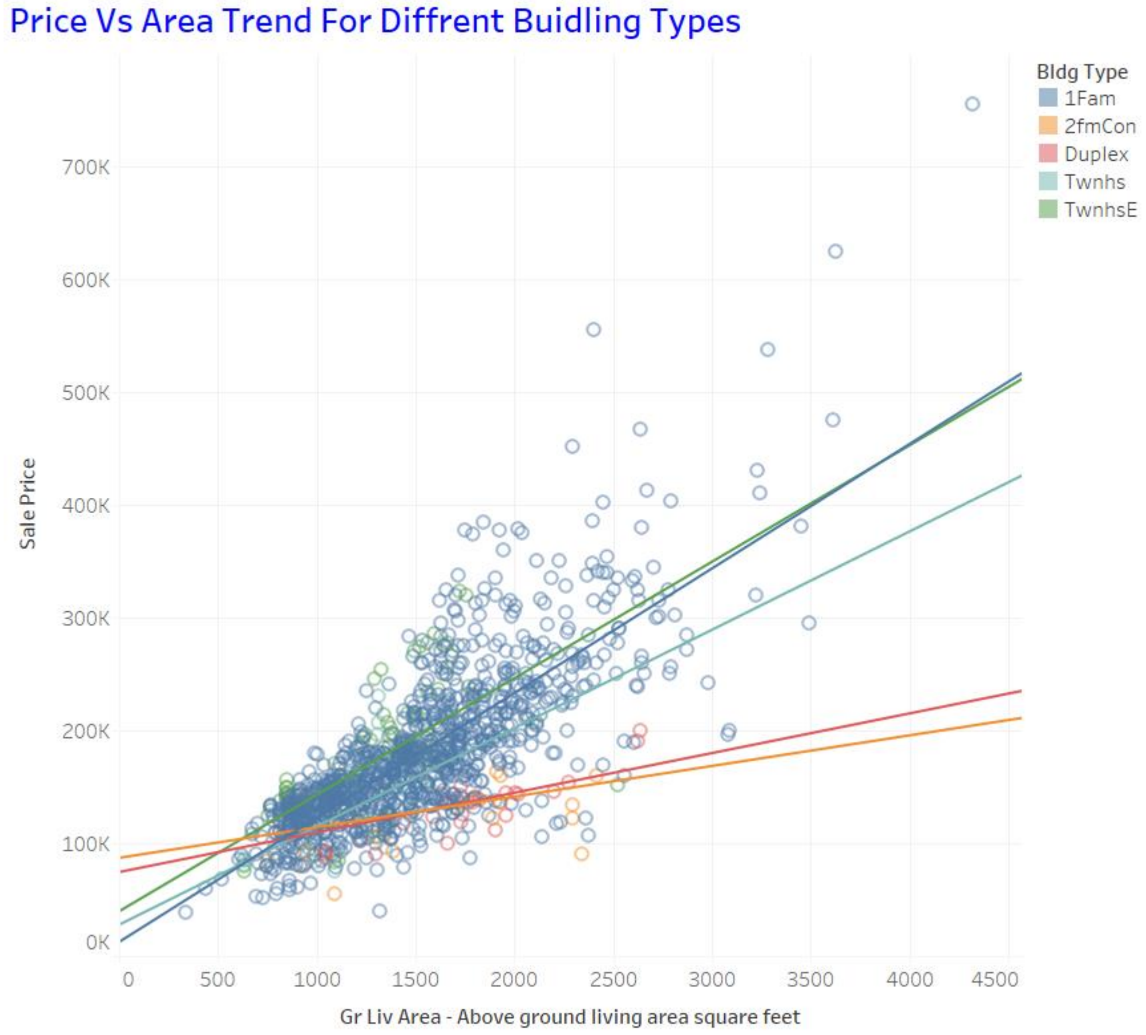
We can visualize from the below map, where each of the neighborhoods are in context to Iowa State University. Looking at the map we can see the housing construction has radiated outward from the university. A ring of old construction surrounds the university area. The very old homes are just east of the university in the old town area. Newer construction seems to radiate outward from the university campus, towards north and west.

Ames,Iowa (Neighborhood - Sale Price - Age of the home)



Map based on Longitude and Latitude. Color shows average of age. Size shows average of Sale Price. The marks are labeled by Neighborhood. The data is filtered on Bldg Type, which keeps 1Fam, 2fmCon, Duplex, Twnhs and TwnhsE.

Sale Price Vs. Square feet (above ground living area) for different building types

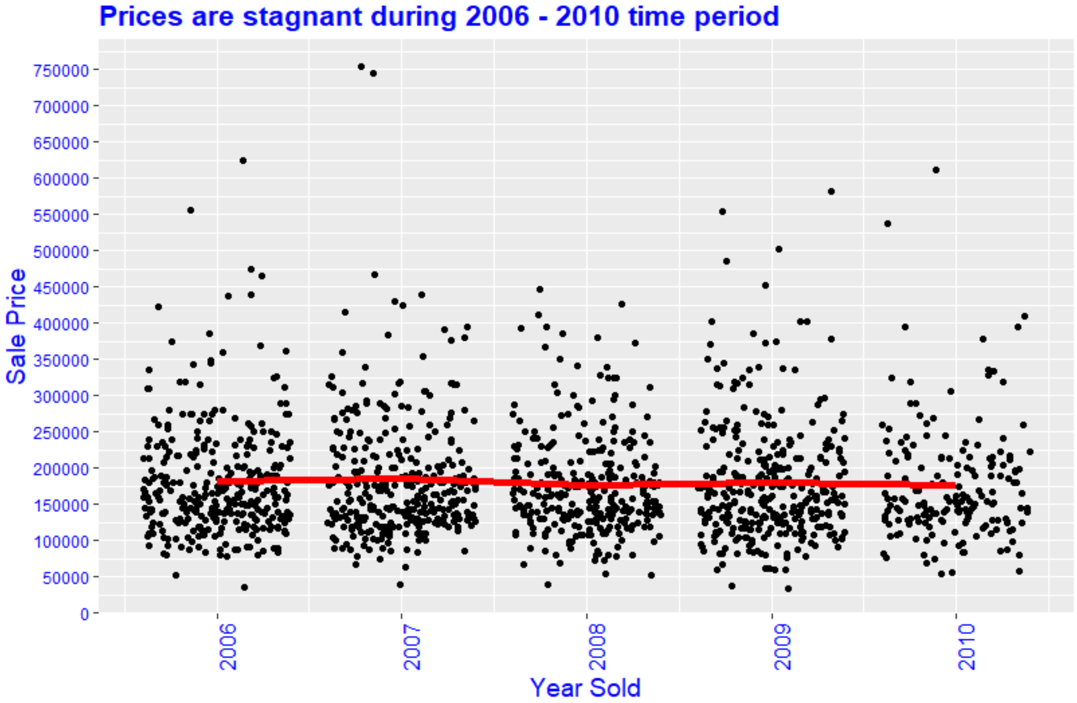


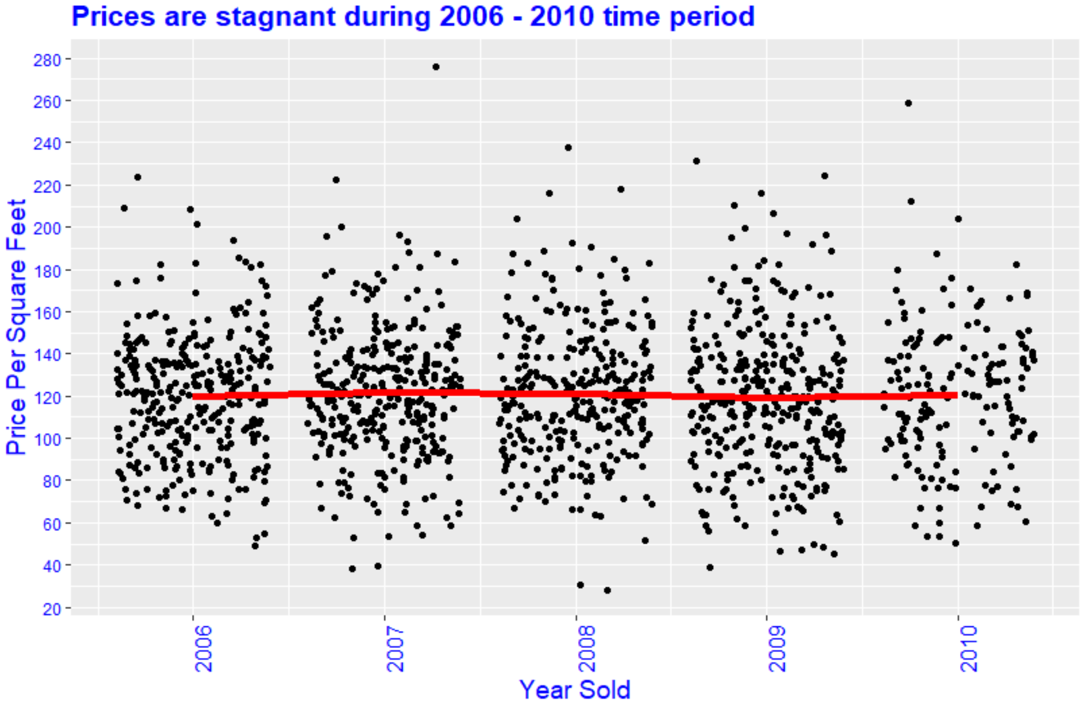
Gr Liv Area vs. Sale Price. Color shows details about Bldg Type. The data is filtered on Sale Condition, which keeps Normal. The view is filtered on Bldg Type, which keeps 1Fam, 2fmCon, Duplex, Twnhs and TwnhsE.

We can see from the above chart that the above ground living area (square feet) has a strong positive linear relationship to the sale price. In the chart above, we see that the rate of price increase differs across various building types. The slope of the line represents the rate of increase. Single-family and townhouse end units (TwnhsE) have the steepest slope, followed by Townhouse Inside Unit (Twnhs). Duplex and Two-family conversions which are originally built as one-family dwellings have the lowest slope.

Average sale price analysis for the 2006 to 2010 time

Our dataset covers housing sales data from 2006 – 2010. So, it represents the main time frame of the housing collapse in the country. In the chart below, the red line represents the mean price. As we can see from the chart, we see no growth in prices during this time. Given that timing, it is surprising that we do not see prices dropping in 2009-2010.





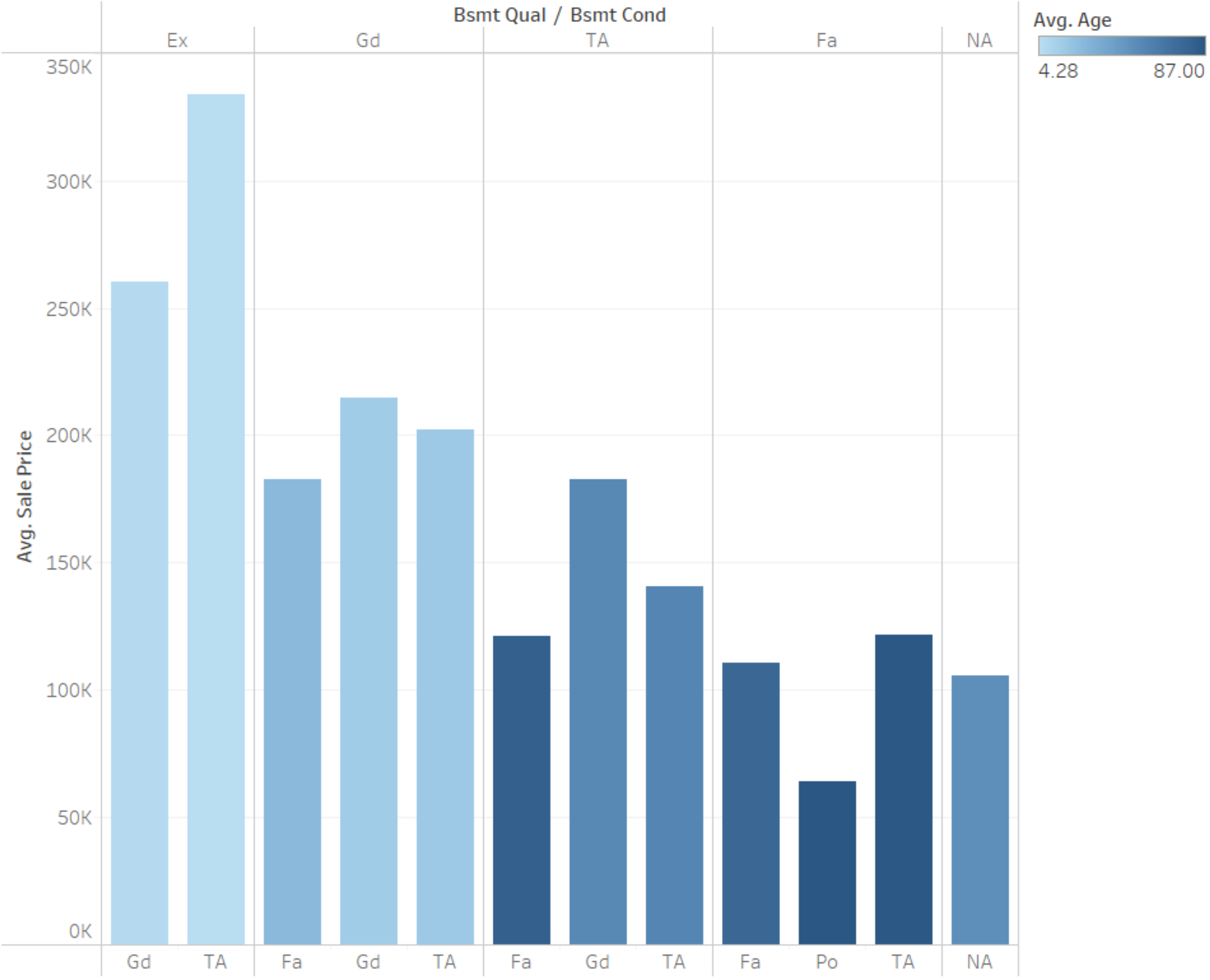
Impact of basement quality and basement condition on sale price.

Avg. Sale Price by quality of the basement.
(BsmtQual: Evaluates the height of the basement)

Ex	Excellent (100+ inches)
Gd	Good (90-99 inches)
TA	Typical (80-89 inches)
Fa	Fair (70-79 inches)
Po	Poor (<70 inches)
NA	No Basement

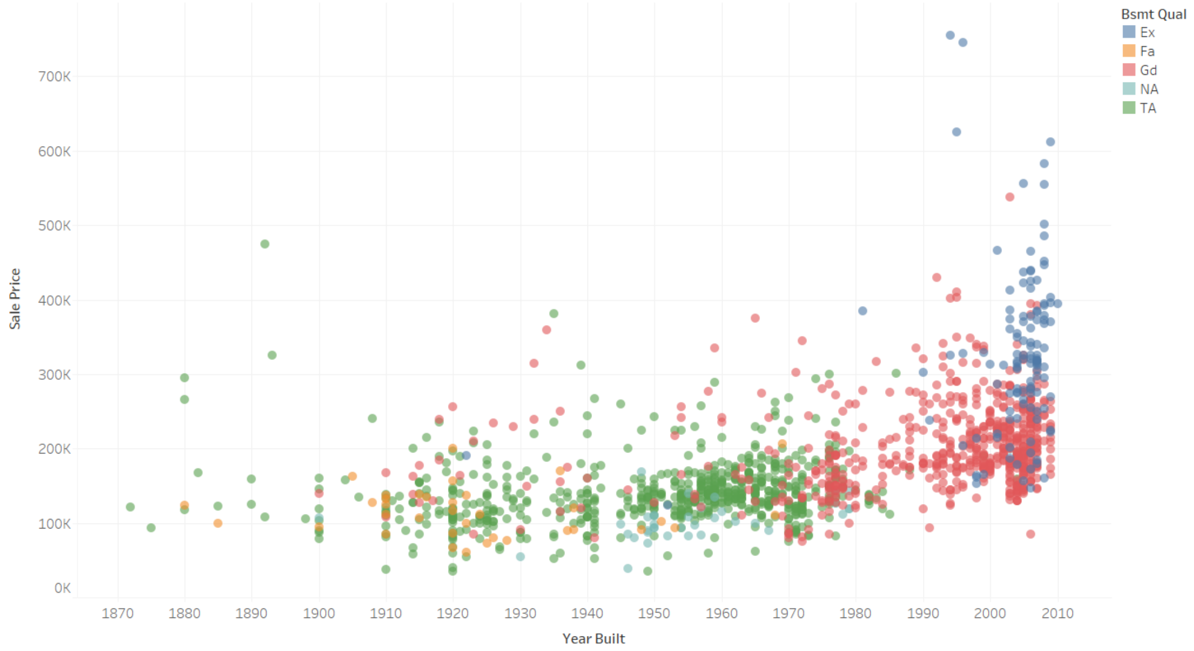
BsmtCond: Evaluates the general condition of the basement

Ex	Excellent
Gd	Good
TA	Typical - slight dampness allowed
Fa	Fair - dampness or some cracking or settling
Po	Poor - Severe cracking, settling, or wetness
NA	No Basement



Average of Sale Price for each Bsmt Cond broken down by Bsmt Qual. Color shows average of Age.

Year Built Vs. Sale Price (Impact of Basement Quality)

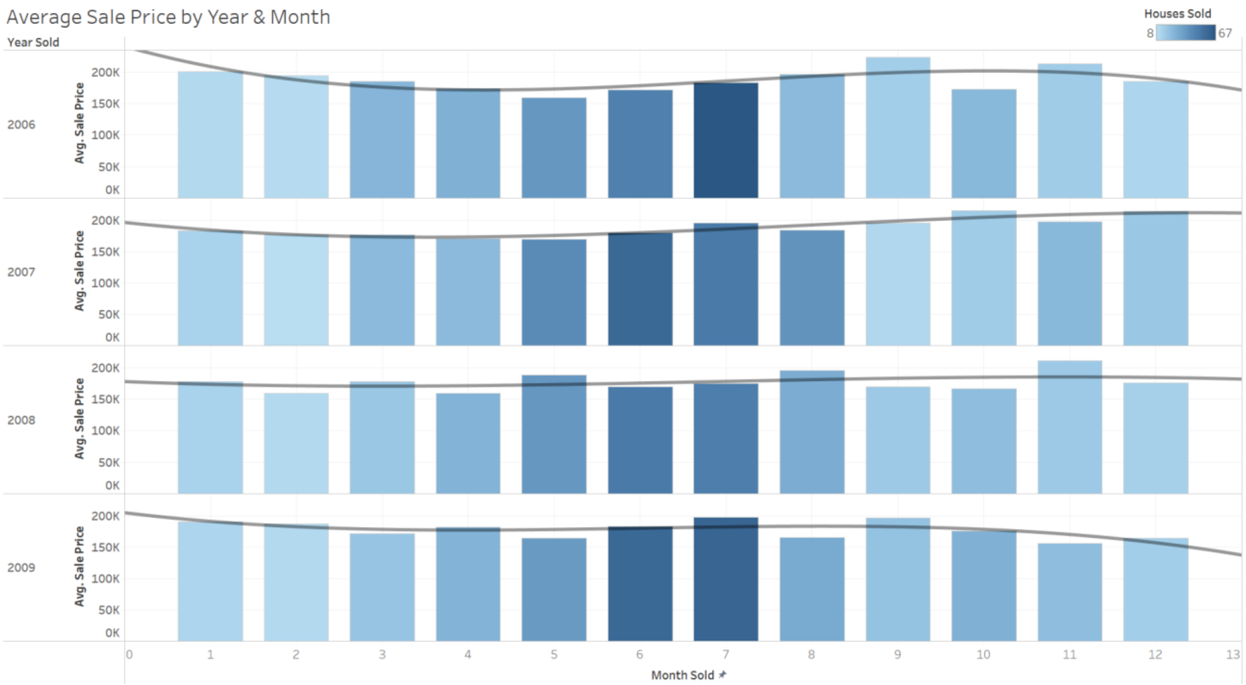


The plot of Sale Price for Year Built. Color shows details about Bsmt Qual. The view is filtered on Bsmt Qual and Year Built. The Bsmt Qual filter keeps Ex, Fa, Gd, NA and TA. The Year Built filter ranges from 1872 to 2010.

From the above two charts, we can see the impact of basement quality which evaluates the height of the basement on sale price. As we can see from both the charts that newer homes are built with 100+ inches of the basement, with good to the typical condition of the basement, and generally have a higher sale price.

We can see from the year-built chart from 1870 up to about 1970 homes had fair (70 - 79 inches) to typical height (80 - 89 inches) for the basement with few exceptions. From 1970 up to 2000 many houses were made with good quality (90 - 99 inches) basement. From 2000 onwards we see houses are generally built with good (90 - 99 inches) to the excellent quality basement (100+).

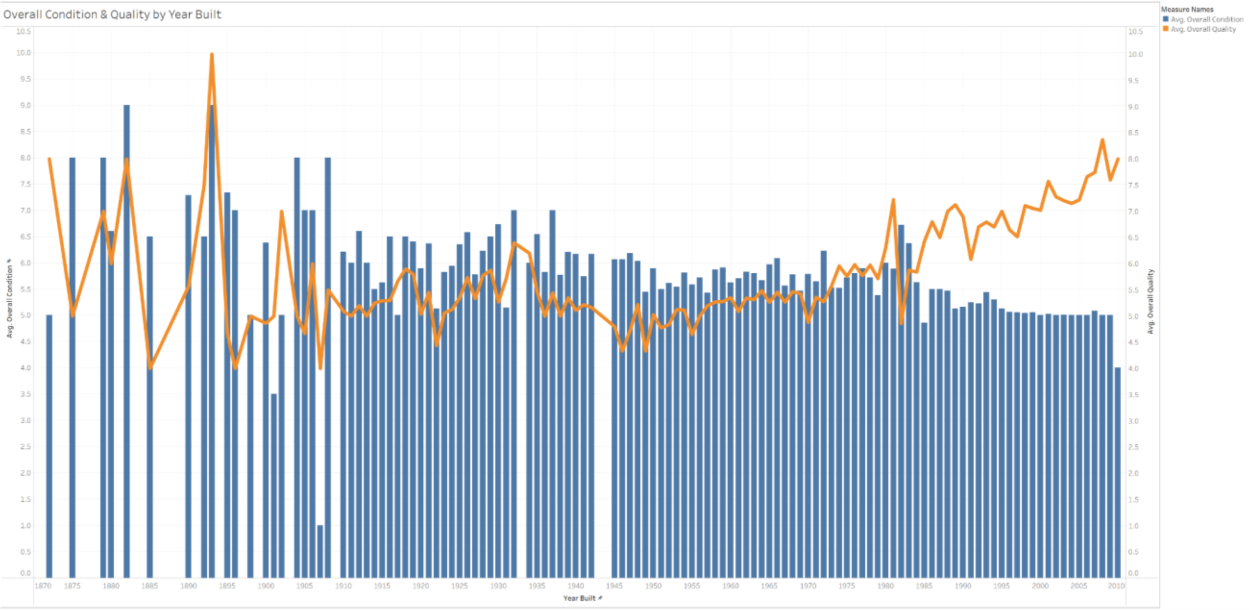
How Does Seasonality Impact Sale Price?



Using the frequency of sales by month from 2006 - 2009 we can track how seasonality impacts sales prices and the number of houses sold in Ames, Iowa. Using the bar's color to analyze the frequency of home sales by month, you can see that the first few months of each year yield fewer home sales, but higher sale prices. The number of houses sold then increases in spring months, March, and April, and really trends upwards through the summer months, May - August (months 5-8). Inversely, we see the sale prices of properties sold in the spring and summer drop due to the number of houses becoming available for purchase at this time. This would appropriately reflect the supply and demand curve in the housing market during warmer months as many people list their properties for sale. As Fall approaches we can see sale price trend upwards as supply and demand shifts again. High sale frequency in June and July likely leaves fewer homes available for purchase as families with kids and college students begin their school year, holidays approach, and home showings become less frequent due to weather growing colder. Finally, we see December as one of the slowest months for home sales in addition to reporting some of the lowest sale prices all year. This is likely the result of there being a smaller number of houses available for purchase and fewer buyers looking to purchase and relocate around the holidays. Once the year has passed, frequency and sale price then shift back to higher levels at the start of the new year.

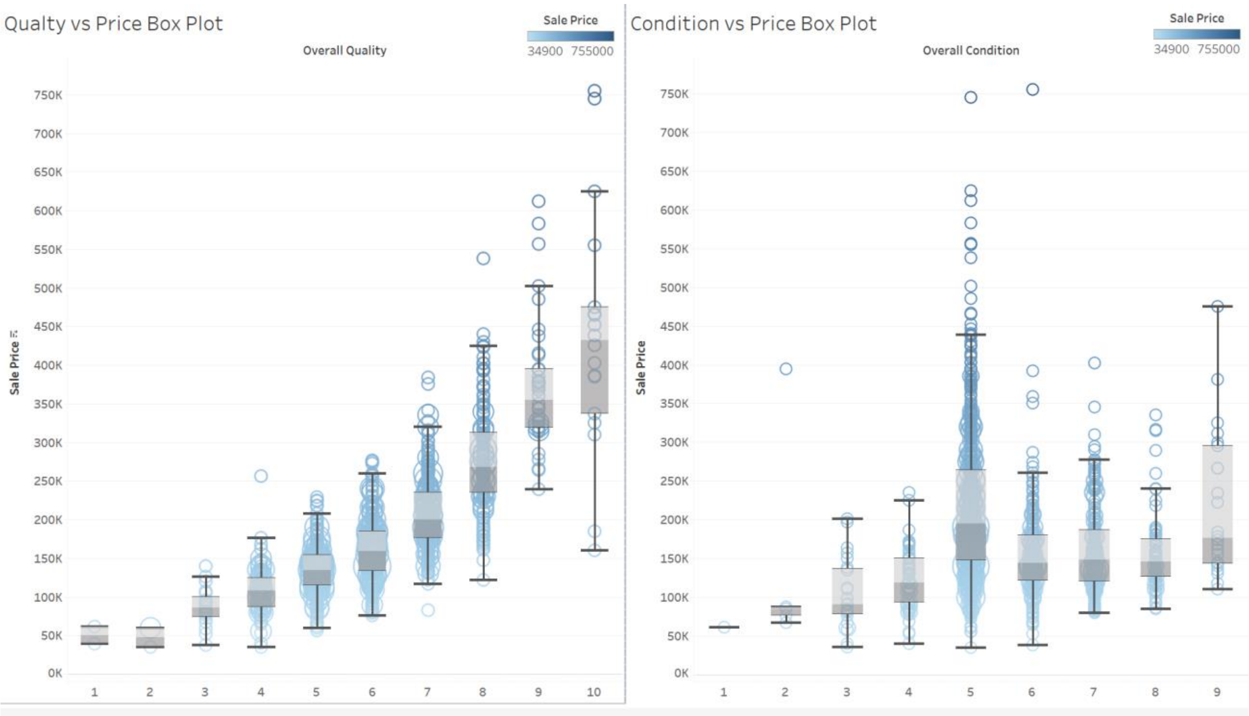
Taking the above visualization into account, it is best to list and sell a property in the first couple months of the year or near the end of the summer. This would allow the seller to benefit from fewer houses on the market at these times when buyers are looking to move after major holiday seasons, before the upcoming school year begins, and before major weather changes. On the other side, buyers can look to pay the best price in April or May or wait until the winter holiday season to purchase their homes.

How does the Quality (of Materials) and Condition vary by Year Built?



The Ames, Iowa dataset included houses that were built as early as 1872 and when the Quality of Materials used, and the Condition of the property are graphed over time we can see how these variables have changed over the last one-hundred and thirty years. The trend observed here is that the quality of material used when building properties was generally rated below the overall condition of those buildings for roughly one hundred years. It wasn't until the mid-1980's that the quality of materials started to exceed the condition of those properties. This phenomenon is interesting because not only would one expect the quality of materials to increase over time, (possibly due to innovation in home building or economic factors), but one would also generally expect the condition of houses built more recently to be rated higher than those from 50 years before. While individuals purchasing newer properties can expect a higher quality home, the building may not be as well-kept and visually appealing, so what exactly are these buyer's paying for? This leads us to our next question.

How does the Quality of Materials and the Current Condition of the Property Impact Sale Price?

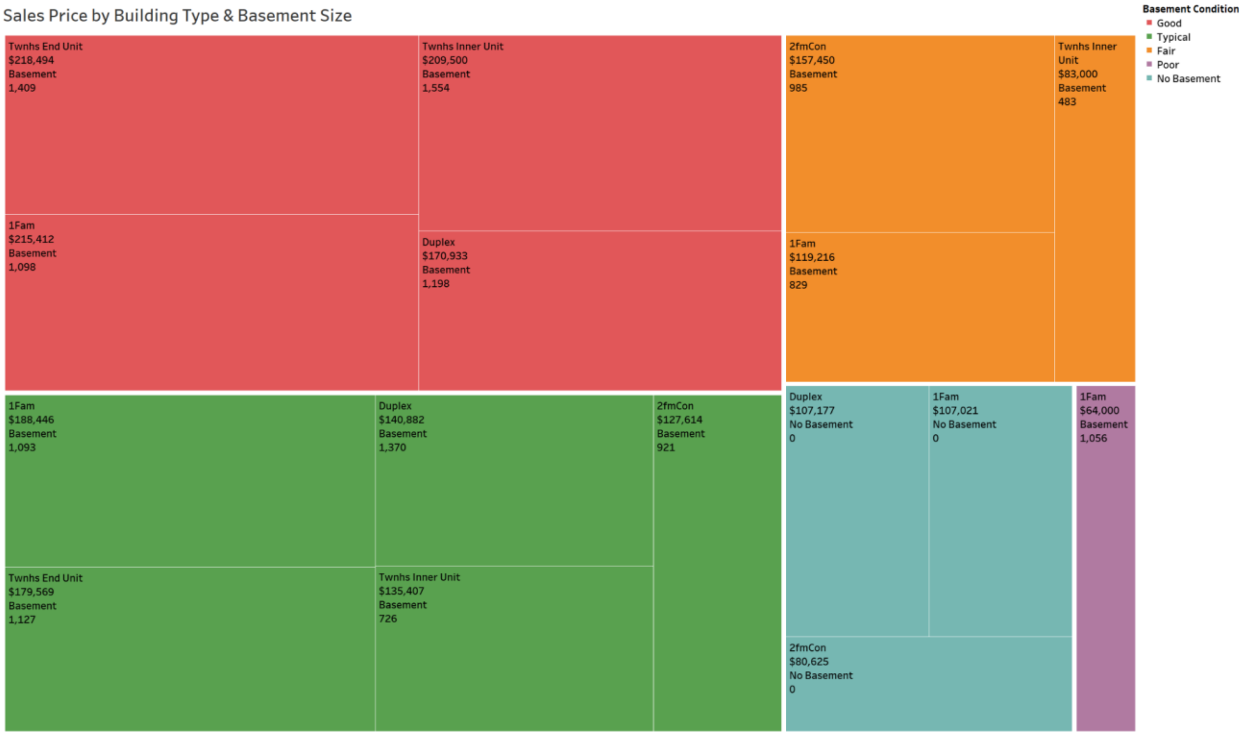


Quality: Based on the width and density of the circles representing a home sale within each Quality indicator, we see that most of the properties sold in Ames, Iowa between 2006 - 2009 had a Quality rating between 5 and 8. This allows us to identify an industry standard for the quality level of materials used when these properties were built. It is also clear that for these Quality levels, Sale Price is densely grouped around the distribution's median indicating not much deviation in Sale Price. Appropriately, Sale Price for these properties' trends upward as Quality increases, however very few houses sold in these years had material quality ratings higher than 8. For those properties that had Quality ratings higher than 8, there is a larger deviation from the median sale price, with some sales far exceeding the boxplots' upper quartile range representing outliers.

Condition: Similarly, for the current condition of the property sold we can identify that most of the properties had condition ratings of 5. While most of the sale price distributions are tightly grouped between the median and upper quartile of each boxplot, when condition is rated 5 or higher there are several outliers sold well beyond each boxplots' upper quartile range. Surprisingly, as the condition improves beyond a rating of 5 there is no increase in sales price. The median of each boxplot for conditions higher than 5 are lower than that of a rating of 5. This is to say that houses in far better condition did not actually have higher sale prices.

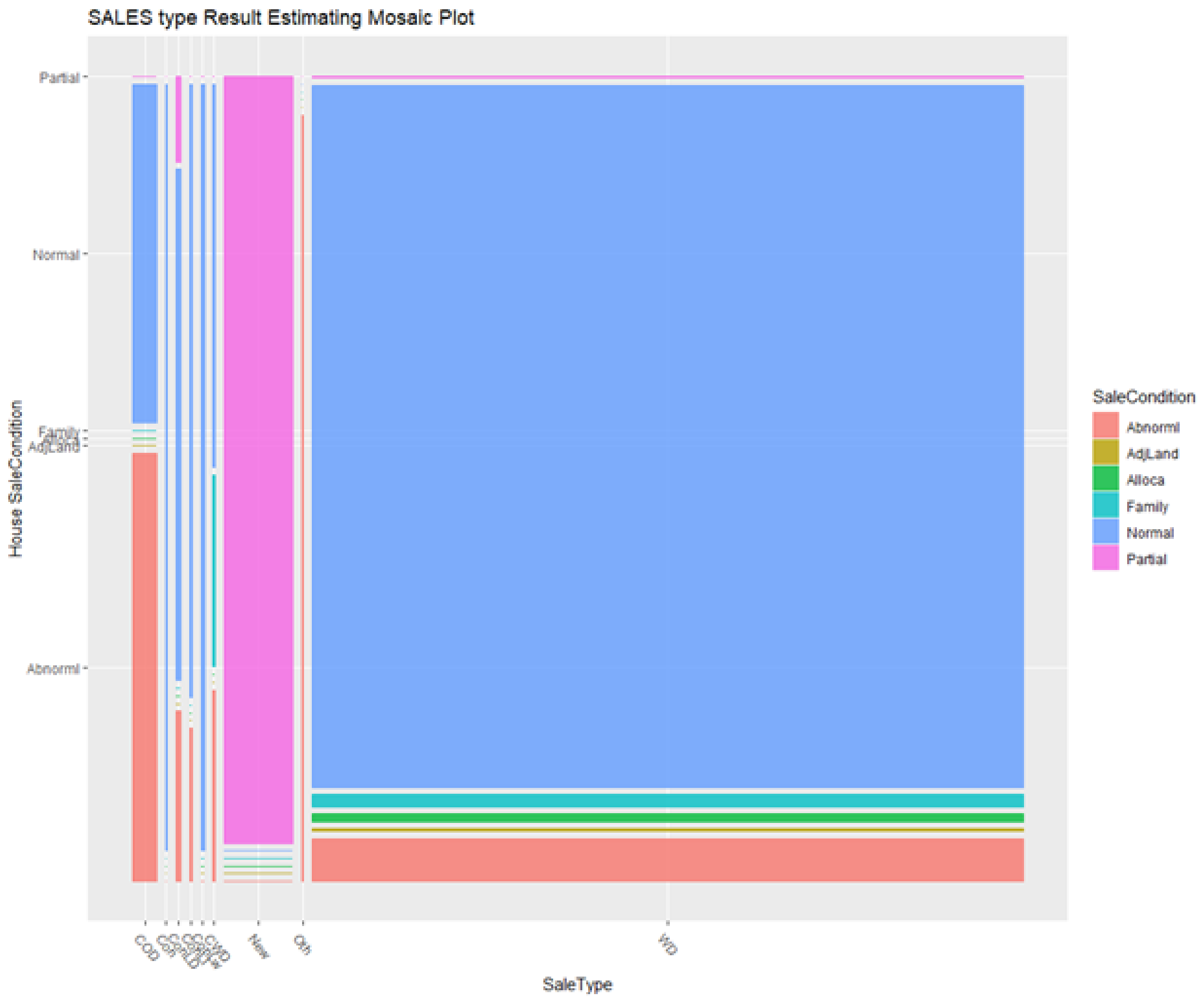
Overall, it is easy to identify the positive correlation that exists between the quality of materials used to build the property and its sale price. However, the condition of the property had very little impact on the sale price in this dataset when the condition rating exceeded 5. As such, there are diminishing returns for sale price regarding the effort one can put into improving the condition of existing homes. Based on other analysis done in this project it is best to investigate other ways to increase the sale price of a property than solely focusing on improving the condition of the home.

Impact of Basements and their Condition on Sale Price by Building Type:

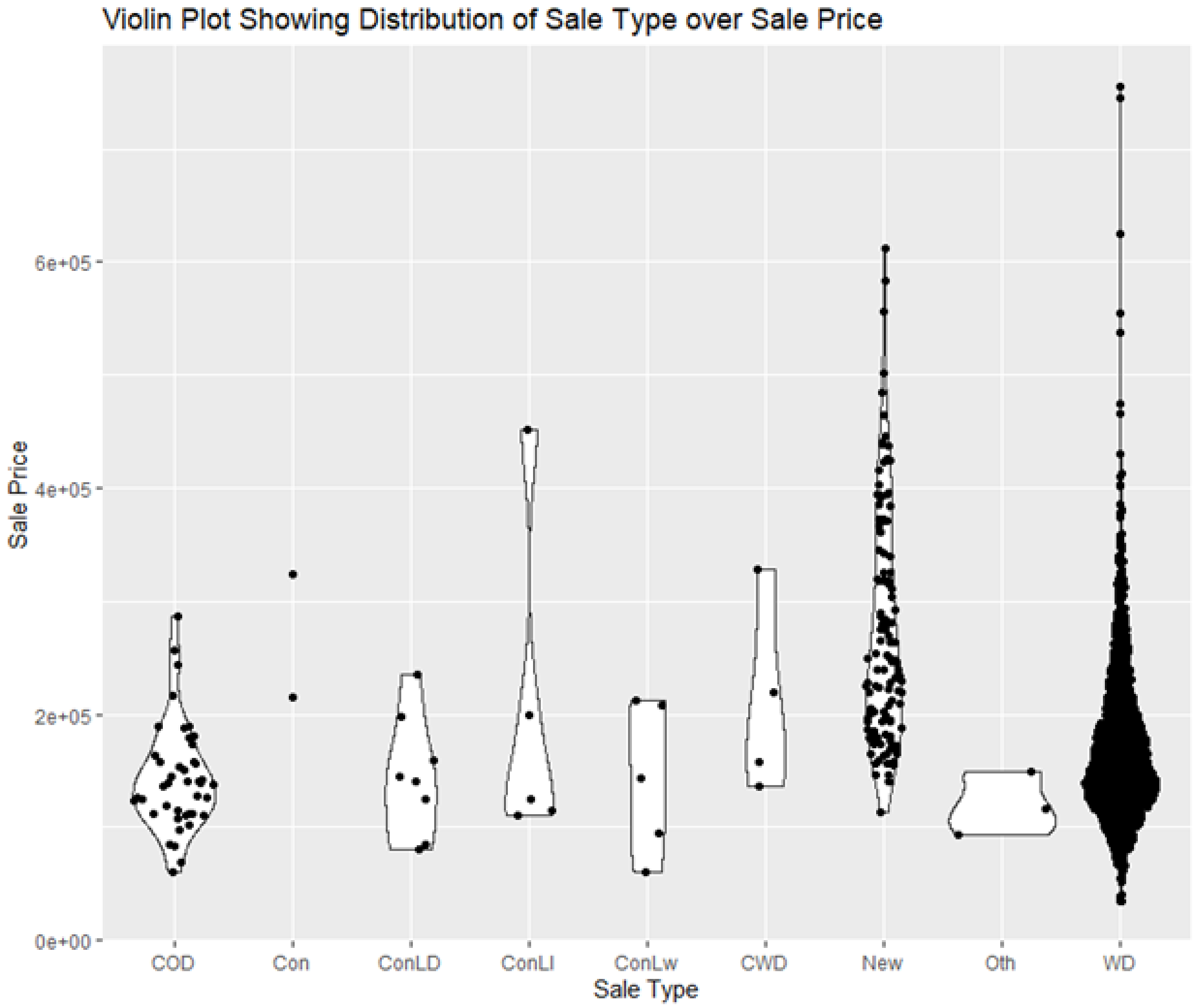


Basements can be an important piece of a home that provide residents additional space for storage or even a finished living space. This visualization shows the average sales price of properties with and without basements, the basements square feet, and condition at time of sale. It is not surprising to observe that properties with large basements that are in good condition were the ones with the higher sales price. However, it was interesting to find that properties with no basement sold for far more than properties that had a basement in poor condition. This is observed by the Single-Family Homes (1Fam) in the bottom right with no basement that sold on average for $107,021 versus the Single-Family Homes with a basement in poor condition that sold for nearly $43,000 less. The range in sale price between these two property groups could be attributed to concerns around the property’s foundation and any repair expenses needed post purchase.

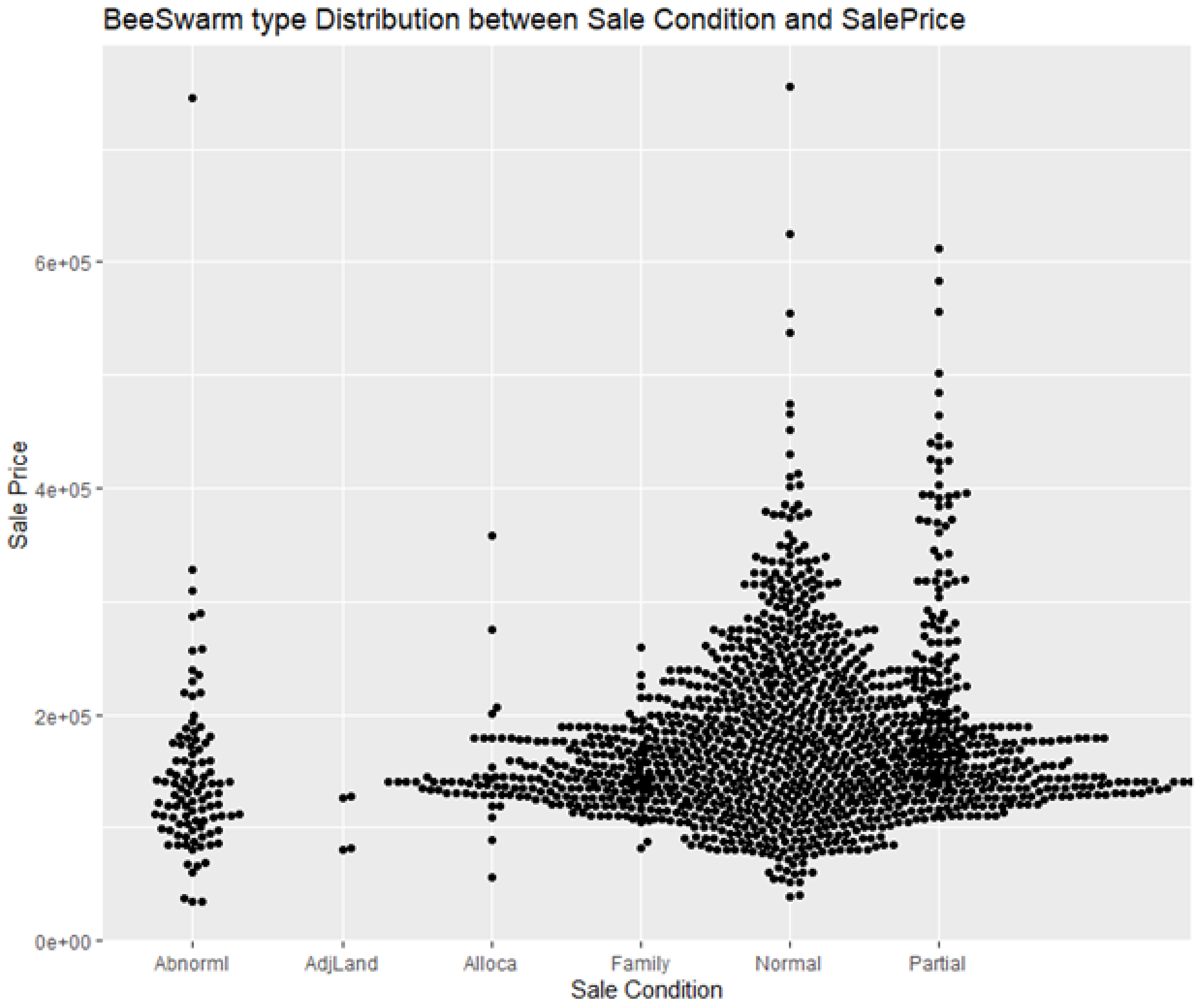
How do different conditions of houses have different types of sales?



The above graph displays a Mosaic plot which gives us the proportion of a categorical variable in terms of another categorical variable. Here the 2 categorical variables are Sale Type and House Sale condition. The description of types of variables are displayed at the bottom left corner. The most visible part is the blue one which denotes that most of WD which means Warranty Deed - Conventional sales are that of normal House Condition. If we consider the sale type as New, we can see that most of the new houses are partially sold.



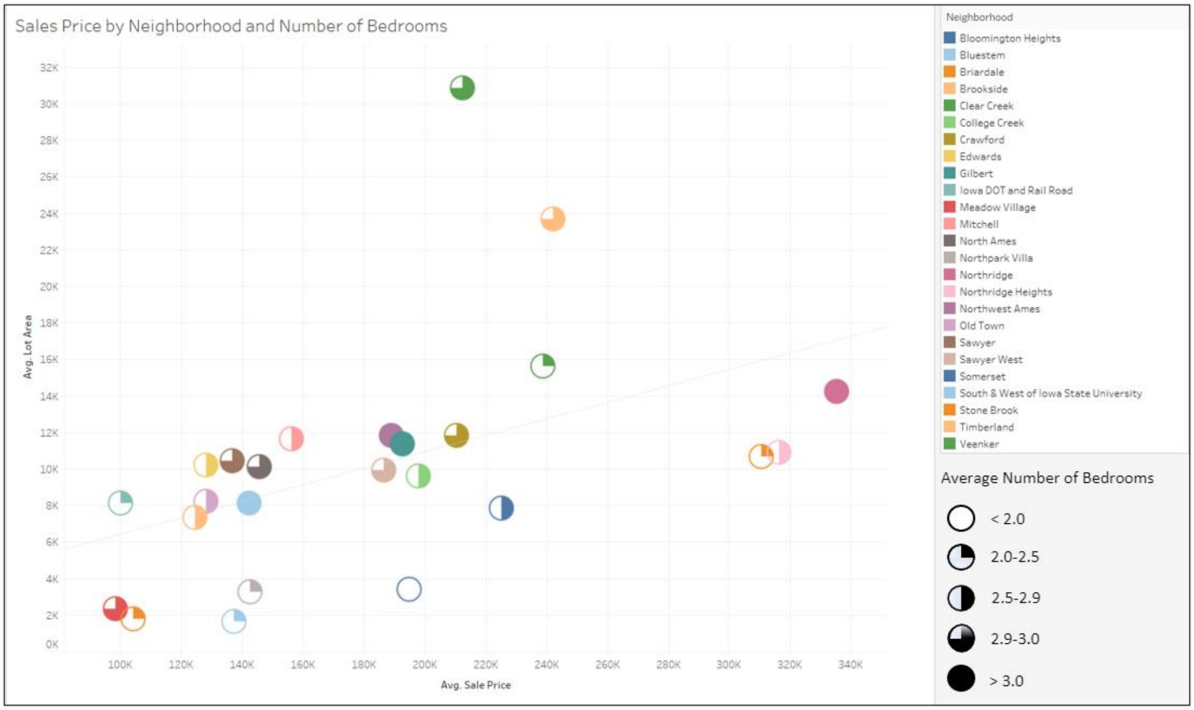
In this second graph a violin plot is displayed over one categorical variable which is Sale type and another numerical variable which is Sale Price. Here we can observe that the Warranty Deed- conventional type sale has happened the most as there are so many data points present as compared to others. Moreover, the bulge in the violin graph of WD category shows that the mode value of the prices is about $130K. Also, there are some outliers which are also displayed at the top of the violin figure.



In this 3rd graph a Beeswarm type distribution is displayed where you can see that the houses having sale conditions as normal are the highest houses sold as there are so many data points. Also, we can see that most of the houses having sale condition as normal are sold around $130K. Moreover, houses with adjoining land purchase are least sold houses as there are very few data points in Adj.Land category.

Number of Bedrooms, Neighborhood, and Area

How Number of Bedrooms Factor In



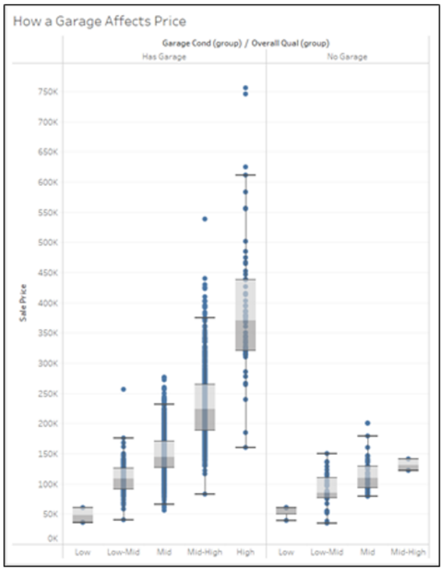
The above graph shows the average price and lot are plotted for each neighborhood within the Ames area. It shows that, for the most part, as lot area increases so does Sale Price. Some specific neighborhoods have larger houses that are not as pricey – Clear Creek and Brookside. These areas could be less desirable, driving the price down per square foot. Additionally, some smaller houses are a lot more expensive in some areas. The expensive areas look to be Northridge, Northridge Heights, and Stone Brook. The size of each mark indicates average number of bedrooms. This does not seem to trend by neighborhood.

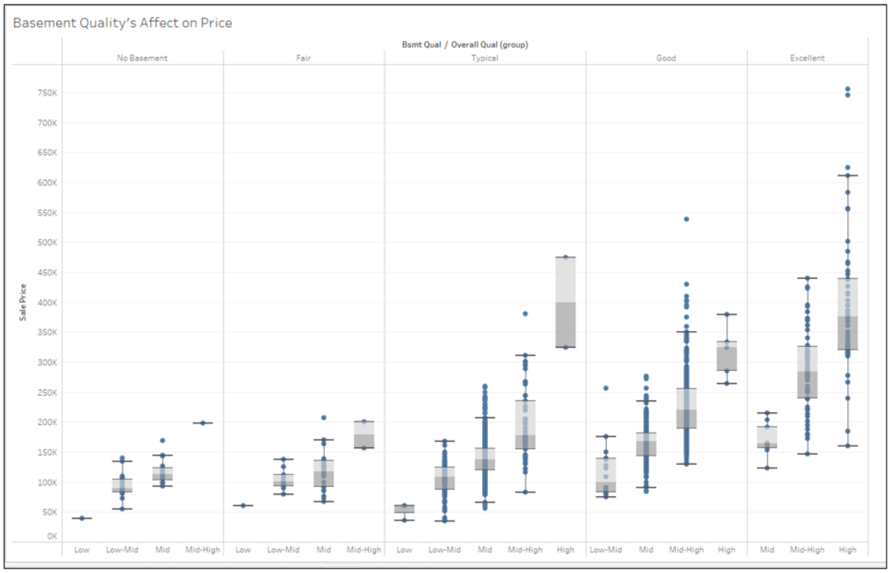
The glyph on the graph represents the average number of bedrooms for that lot size, price, and neighborhood. It does not appear that the number of bedrooms is necessarily consistently going up as does price and lot area. This makes sense, as the quality and location show to matter more.

To make this graphic, Tableau was used, and an aggregate scatter plot was created. Color represents neighborhood and glyph represent the number of bedrooms.

Miscellaneous Features

Miscellaneous Features influence Price


How a Garage Affects Price


Basement Quality's Affect on Price

The miscellaneous features analyzed above are basement and garage.

Garage: The garage graphic (left) shows that there is a significant difference in prices when there is a garage vs. when there is not. The right side of the graphic shows that different distribution of sales price when there is a garage and, on the right, it shows the distribution of sales price when there is not a garage. These are also divided out by overall property quality.

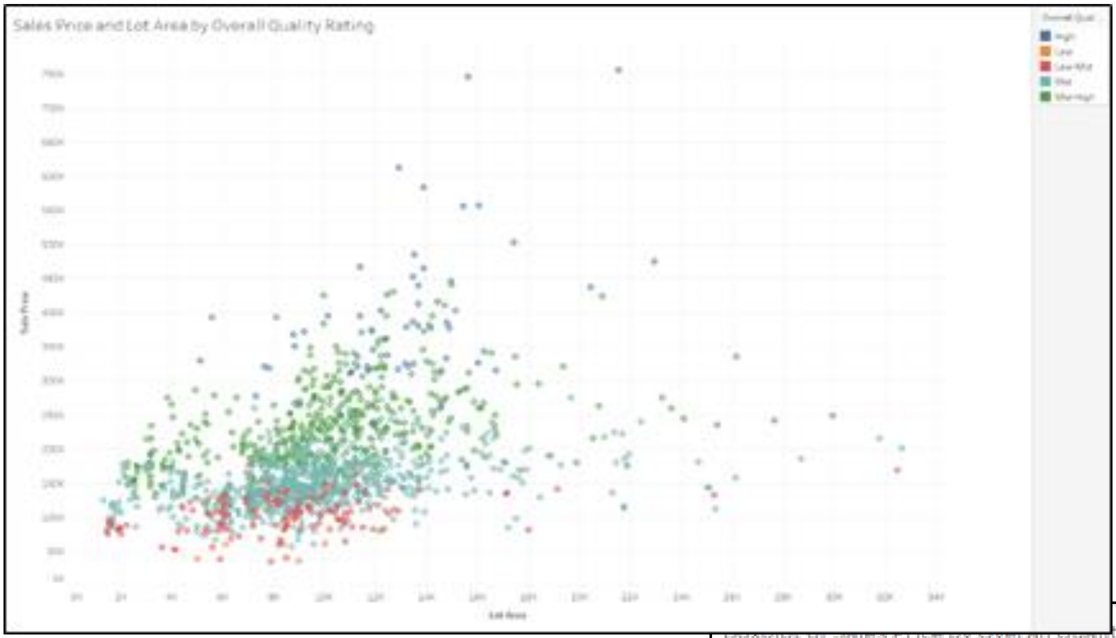
Garages are sometimes nonnegotiable features for certain individuals. Especially up in Iowa, where the winters can bring snow, many people must desire garages. If someone has a car and needs to drive to work each day, it is crucial. This can be used to sellers' advantage and if they are not, this analysis shows they could be losing money.

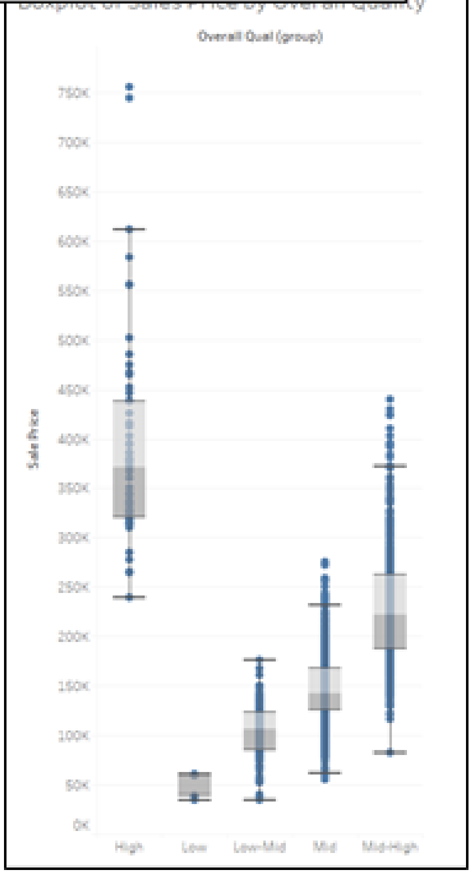
Basements: Basements add additional space to a home, as stated earlier in the report. It is not enough just to have a basement though; the quality also matters. The boxplot to the right above is sales price divided out by overall quality as well as quality of garage.

There are a couple of outliers and some spots where there is less data, for example there are not many high-quality homes with just "typical" basements. In fact, there are not many high-quality homes with no, or fair basements either. This shows that basements are a feature that are sought after and for a listing to be high quality, a basement is probably included.

This is interesting in a place like Iowa vs. somewhere like Chicago, New York City, or another large city because in big cities there are not as many properties with basements. Having more space is more of a luxury in a city whereas it is more common in places like Ames Iowa.

Overall Quality Influences Price



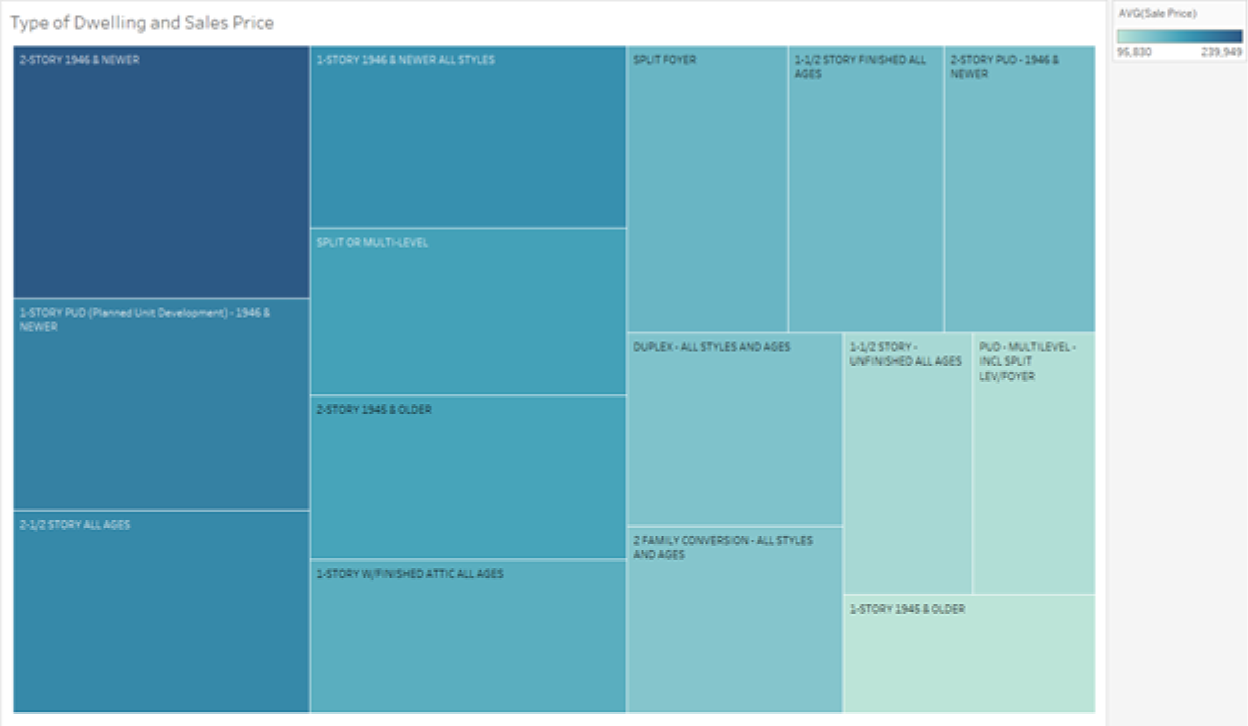


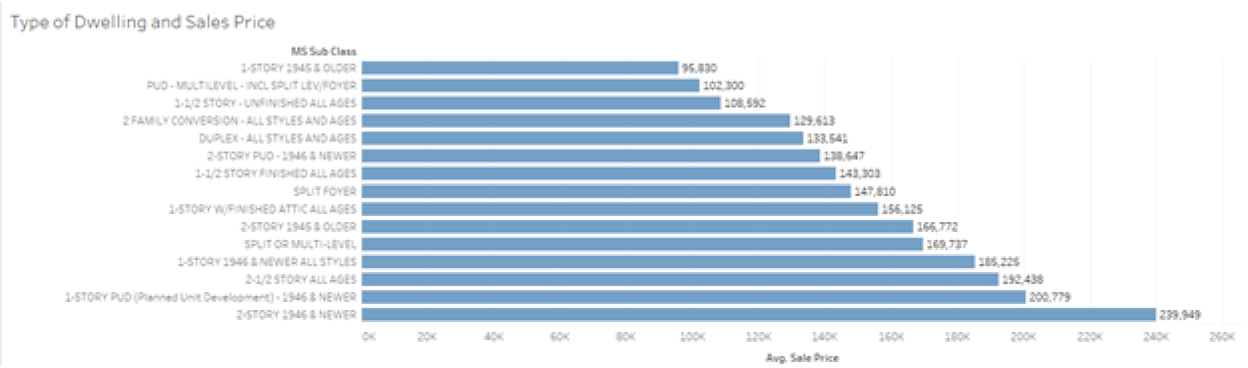
The figure above shows a scatterplot with an overlay of overall quality ratings grouped from Low to High. The trend shows that lot area and quality are somewhat correlated, but the overall quality appears to be clustered in groups and changed based on sale price (or sales price is dependent on quality). This is shown through the flat size of the clustering. There are some very large places that have a low overall quality, but they remain in a similar threshold of others in that quality from a price perspective.

The boxplot to the right shows a better depiction of how the distribution of sales price is laid out depending on the overall quality of the home. This graph and the above graph have the same overall takeaway, aside from the fact that the boxplot does not show how lot size is affected.

These plots were both created in tableau. The first plot is a scatterplot, and the colors are defining the overall quality score. The second plot is a boxplot.

How the Type of Dwelling Affects Sales Price





The above graphics show how the dwelling type has an influence on price. The dwelling types are listed below:

20	1-STORY 1946 & NEWER ALL STYLES
30	1-STORY 1945 & OLDER
40	1-STORY W/FINISHED ATTIC ALL AGES
45	1-1/2 STORY - UNFINISHED ALL AGES
50	1-1/2 STORY FINISHED ALL AGES
60	2-STORY 1946 & NEWER
70	2-STORY 1945 & OLDER
75	2-1/2 STORY ALL AGES
80	SPLIT OR MULTI-LEVEL
85	SPLIT FOYER
90	DUPLEX - ALL STYLES AND AGES

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

The graphics show that the most expensive dwelling on average is the 2-story 1946 and newer dwellings. 3-Story 1945 and older dwellings are the least expensive listings.

Conclusion:

In conclusion, using the Ames, Iowa Housing data set, our team was able to identify multiple. factors that highly influenced the sale price of homes. We were also able to identify overall trends building quality and condition, the neighborhoods and dwelling types that are most desirable, the seasonality in Ames Iowa and how that affects sales, as well as several additional takeaways. The benefit of an analysis like this allows home buyers to decide what price range they should be looking for as well as what sellers could possibly get for their listing.