# Analysis of Key Attributes Contributing to the House Price of Ames City, Iowa

Foundation of Data Science
December 10, 2023

# Table of Content

# Objectives

The primary objective of this report is to conduct a comprehensive analysis of the residential real estate market in Ames City, Iowa, with a particular focus on house pricing. By examining the available housing data, this study aims to identify the key variables that significantly influence the pricing of houses in this region. The analysis will include a variety of factors such as location, square footage, the number of bedrooms and bathrooms, the age of the property, and any recent renovations or upgrades. The report intends to provide valuable insights for construction firms by understanding the correlation between these factors and house prices. This report is committed to providing a clear, evidence-based understanding of what drives house prices in Ames City, ensuring that all conclusions drawn are based on robust data analysis and are presented with precision and clarity.

# Data Preparation

The dataset of our analysis is a machine-learning dataset sourced from Kaggle, a platform renowned for its extensive repository of datasets used for analytical and predictive modelling. To facilitate seamless collaboration and execution of our data analysis, our team has chosen to utilize Google Colab. Its robust collaboration features, coupled with an accessible cloud-based environment, make it an ideal platform for our project needs.

For accessibility and ease of integration, the dataset has been hosted on GitHub, allowing direct import into the Google Colab environment. This ensures that all team members have access to the data, enabling efficient iterative analysis processes.

Our dataset comprises seventy-nine distinct attributes, providing a rich source of variables to explore concerning the sale price of houses. Initial assessment of the data quality reveals a high degree of completeness, with seventy-three of these attributes maintaining non-null values in over 90% of the records. This high level of data integrity is instrumental in conducting a robust analysis.

Despite the overall data quality, we encounter a notable exception. The attribute representing lot frontage - commonly understood to have a strong correlation with sale price - exhibits only 82% non-null records. This presents a challenge as the completeness of this particular variable is crucial for a nuanced understanding of its impact on house pricing. Our next steps will involve addressing this gap, potentially through data imputation strategies, to ensure that the integrity of our analysis remains uncompromised.

Another challenge that emerges from our dataset is its time series nature, particularly concerning the fluctuating average house prices in tandem with the broader housing market trends. Such fluctuations can potentially introduce biases in our correlation analysis. For

instance, specific types of houses may appear to command higher prices simply because they are sold during periods of market upswing. This phenomenon could lead to erroneous conclusions about the true impact of certain house attributes on sales prices.

To mitigate this issue, our approach involves a methodical adjustment of the sales prices to account for market fluctuations over time. We will calculate the overall average monthly sale price across the dataset, creating a baseline for comparison. Then, we will compare this average with the individual average monthly sale prices to derive a ratio for each month. This ratio will effectively reflect the market's condition at the year and month of the sale.

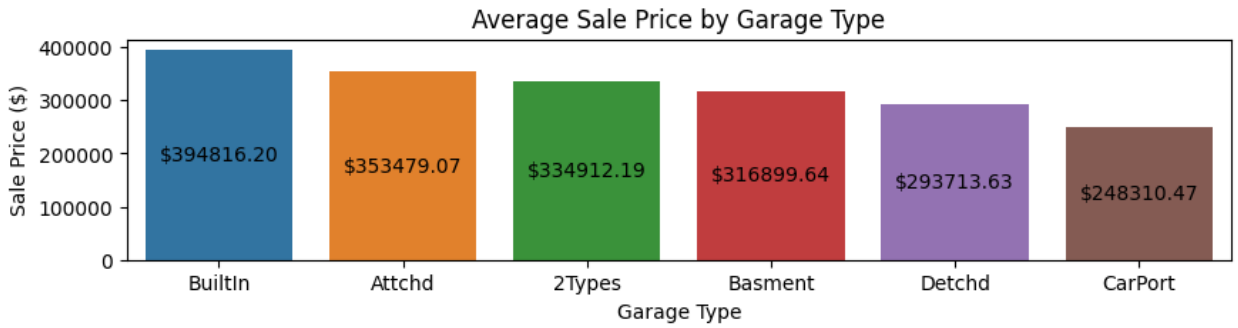
Subsequently, we will apply this ratio to the sale prices of individual properties, adjusting them based on the specific year and month of the sale. For more detail with this approach please refer to Appendix A. This method aims to normalize the prices, ensuring that our analysis accurately reflects the intrinsic value of the house attributes, rather than being influenced by the temporal market conditions.

An additional intricacy in our analysis stems from the impact of house age and recent renovations on sale prices. It is a well-established notion that newer houses, or those recently renovated, typically command higher market prices. This reality necessitates an adjustment in our analysis to account for the depreciation effect. Houses that are older or have yet to be recently updated are likely to sell for less, not necessarily due to a lack of intrinsic value, but due to natural wear and tear or outdated features. We took a similar approach in adjusting the larger market impact, the details can be found in Appendix B. By doing so, we intend to achieve a more reliable and accurate understanding of the factors that genuinely impact house prices in Ames City, Iowa.

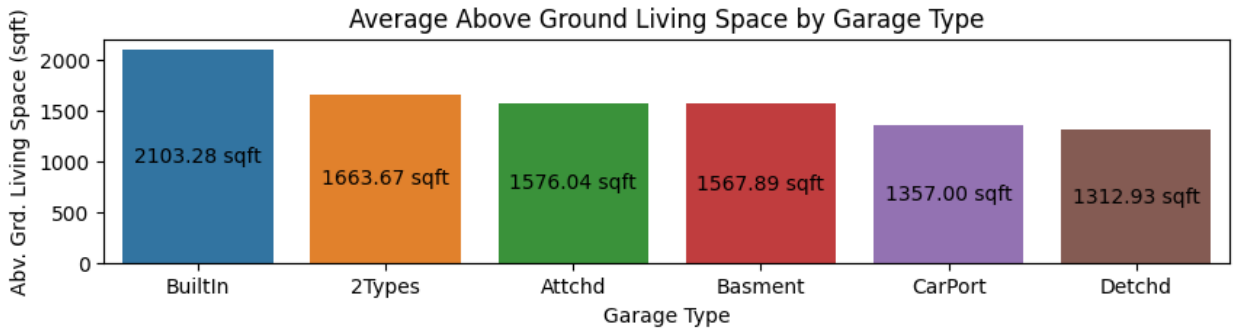
# Analysis

## Garage Type (GarageType)

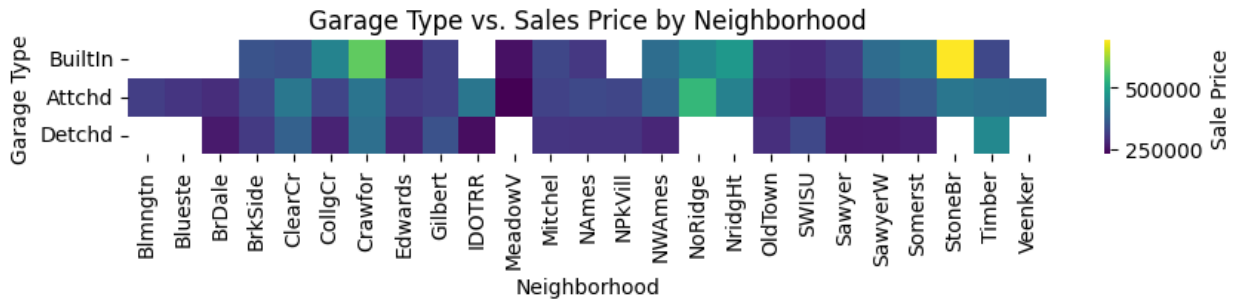
The Garage Type feature describes different types of garages, such as detached or attached, etc. To assess the correlation between garage type and sales price, we need to consider that 'GarageType' is a categorical variable while 'SalePrice' is numerical. Since correlation coefficients are typically used to measure the relationship between numerical variables, the direct correlation calculation does not apply to a categorical variable. However, we can still analyze the relationship by calculating the average sales price for each category of 'GarageType' to see if there are significant differences in sale prices among different garage types.




The chart above shows the average price for each garage type. We can see the type of garage has an impact on the sales price. With built-in garage (typically have a room above the garage) fetching the highest sales price, followed by attached. The reason the built-in garage type fetches a higher price could be related to the additional room above the garage, which means there is more living space. We can validate this assumption by checking the correlation between garage size and above-ground living area.



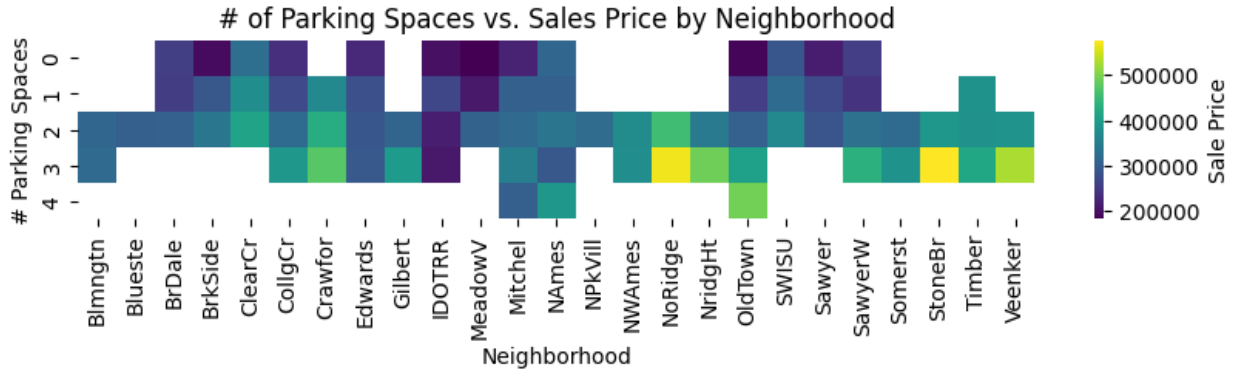

From the chart above we can see our assumption is correct. A property with a built-in garage on average has 527.24 square feet more above-ground living space compared to a property with an attached garage. What's interesting is that properties with two types of garages have more above-ground living space but fetch a lower sales price, one possible explanation is that properties with two types of garages are usually situated on a bigger lot, and bigger lots are usually located in suburban areas where property value is lower compare to city centre. However, after a quick check, there are only six properties in our 1460 records with two types of garages, it's not a significant enough number to draw a definitive conclusion. It's a similar situation for the basement garage and carport garage, they only have 19 and 9 records, respectively. Ignoring those garage types, plot the garage type against sales price grouped by neighbourhood, and we get the following heat map.



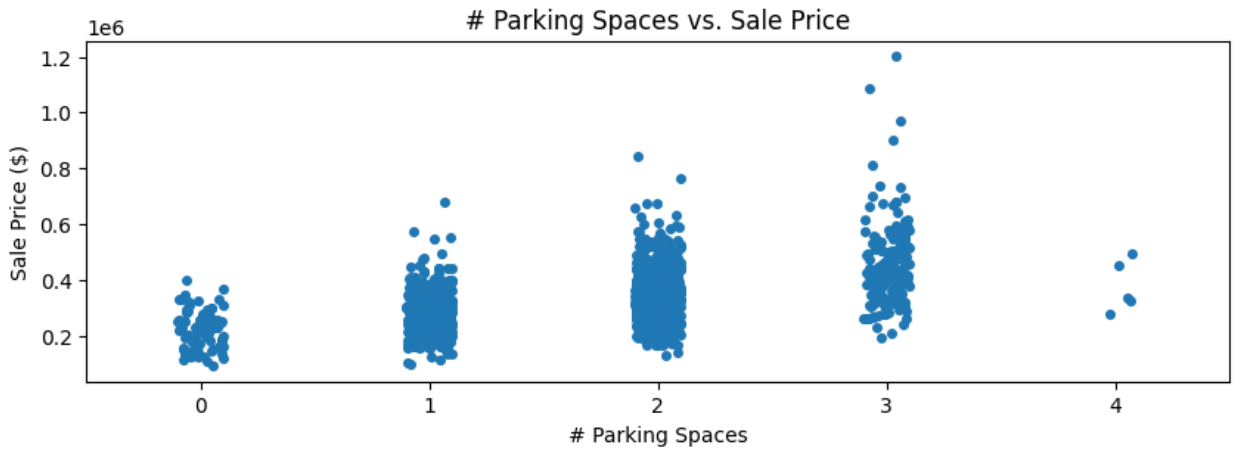

We can see that for the majority of the neighbourhood, the colour gets lighter as it gets closer to the top. It aligns with our early discovery, where built-in garages are usually associated with higher sales prices, followed by attached garages, and then detached garages.

## Number of Park Spaces in Garage (GarageCars)

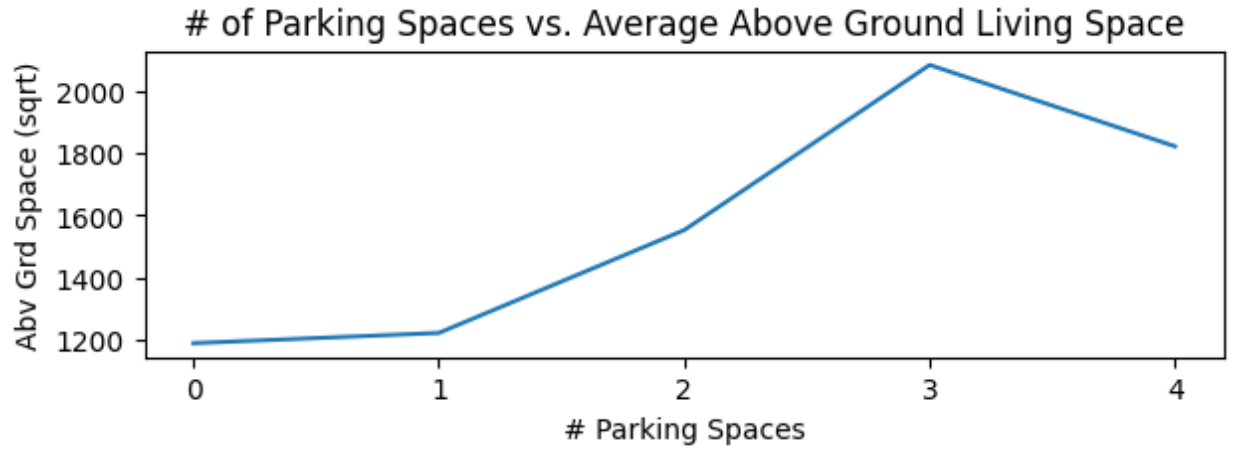
Garage Cars is the number of parking spaces in the garage. Firstly, we plot a heat map between garage cars and neighbourhoods, using sales price as heat value.



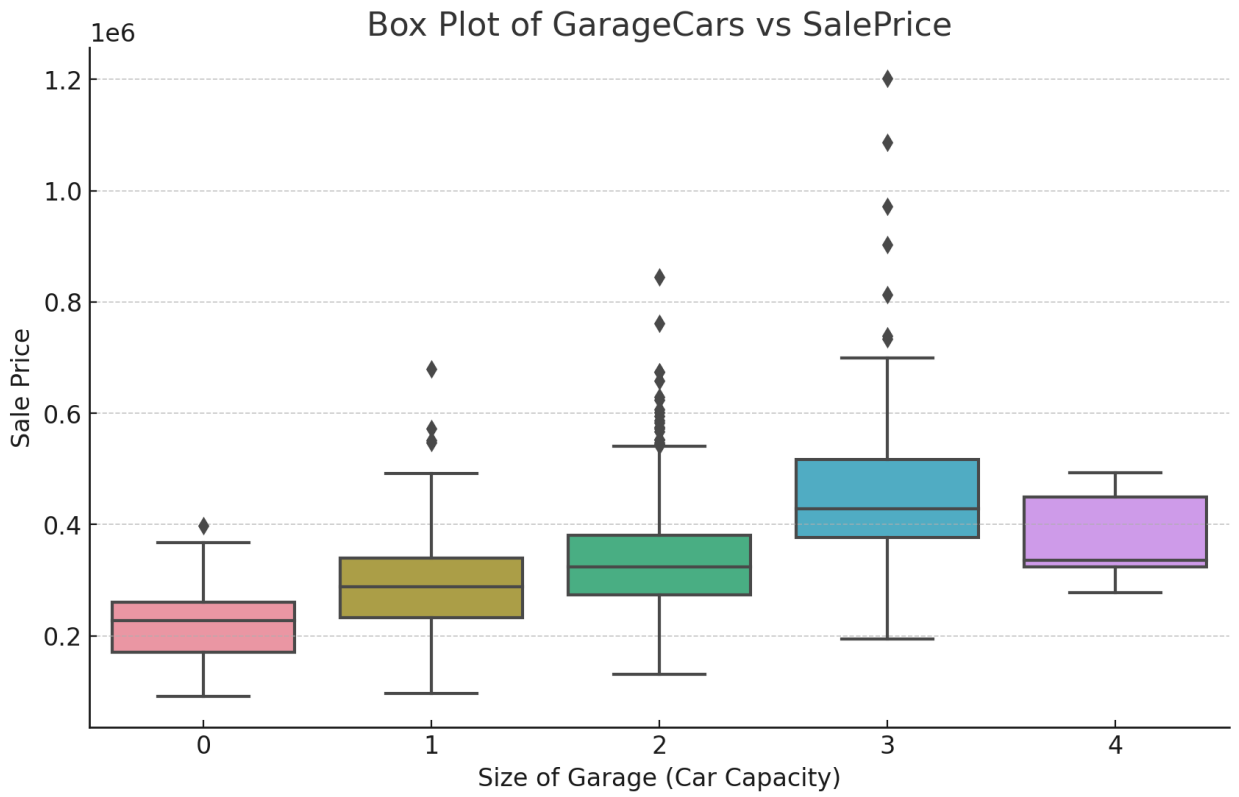

From the heat map, we can see that regardless of the neighbourhood, the sales price increases as the number of parking spaces in the garage increases. As a result, we can conclude that the effect of the number of parking spaces is agnostic across neighbourhoods. One interesting point to note here is that certain neighbourhoods only have two or more parking spaces, they generally have higher sales prices, and these are the more affluent neighbourhoods.
Next, we will plot the number of parking in the garage against the sales price.




There is not enough data for four-car garages to draw a concrete conclusion, but we can see a clear positive correlation - the house with more parking spaces in the garage would have a higher sales price. Naturally, properties with more parking spaces usually mean there are more above-ground living spaces. So the correlation between parking spaces could be an extension of the correlation between above-ground living space and sales price.



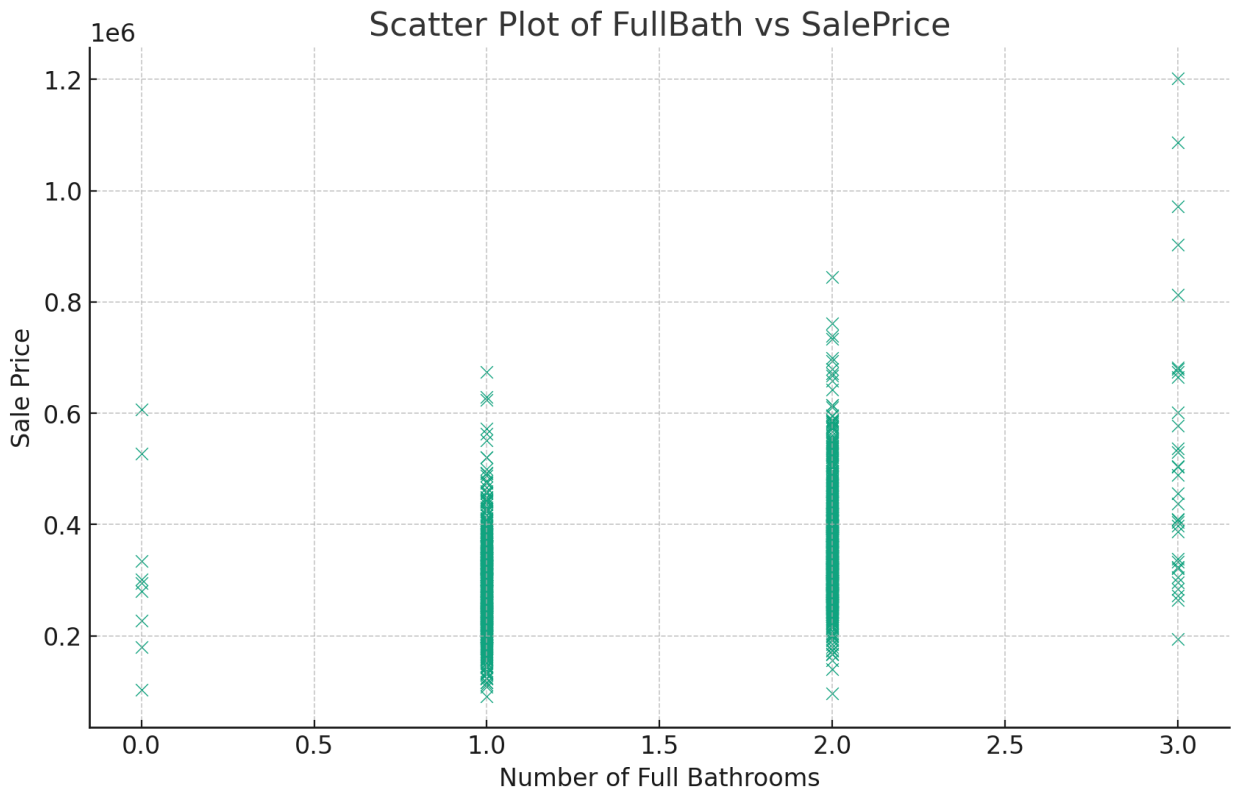

In the chart above, ignoring the 4 parking space data due to the small sample size, we can see more parking spaces mean more above-ground living space. And from the previous analysis, we know there is a very strong correlation between above-ground living space and property sales price. We can conclude that more parking space correlates to higher sales price because more parking space correlates to more above-ground living space, which then has a strong correlation with the sales price. We also want to consider sales volume, let's start by examining the relationship between 'GarageCars' and 'SalePrice', and then look at the distribution of 'GarageCars' in the dataset.




Houses with a 2-car garage are the most common with 824 properties, while only 181 properties have a 3-car garage. Based on this analysis, 2-car garages appear to be the optimal choice for achieving a good balance between a higher sale price and a reasonable transaction volume. This category is not only the most common in the dataset, suggesting high market activity, but also corresponds to a significant increase in median sale price compared to 1-car garages. While 3-car garages may offer higher sale prices, they are less common and cater to a more specific segment of the market, potentially limiting the volume of transactions.

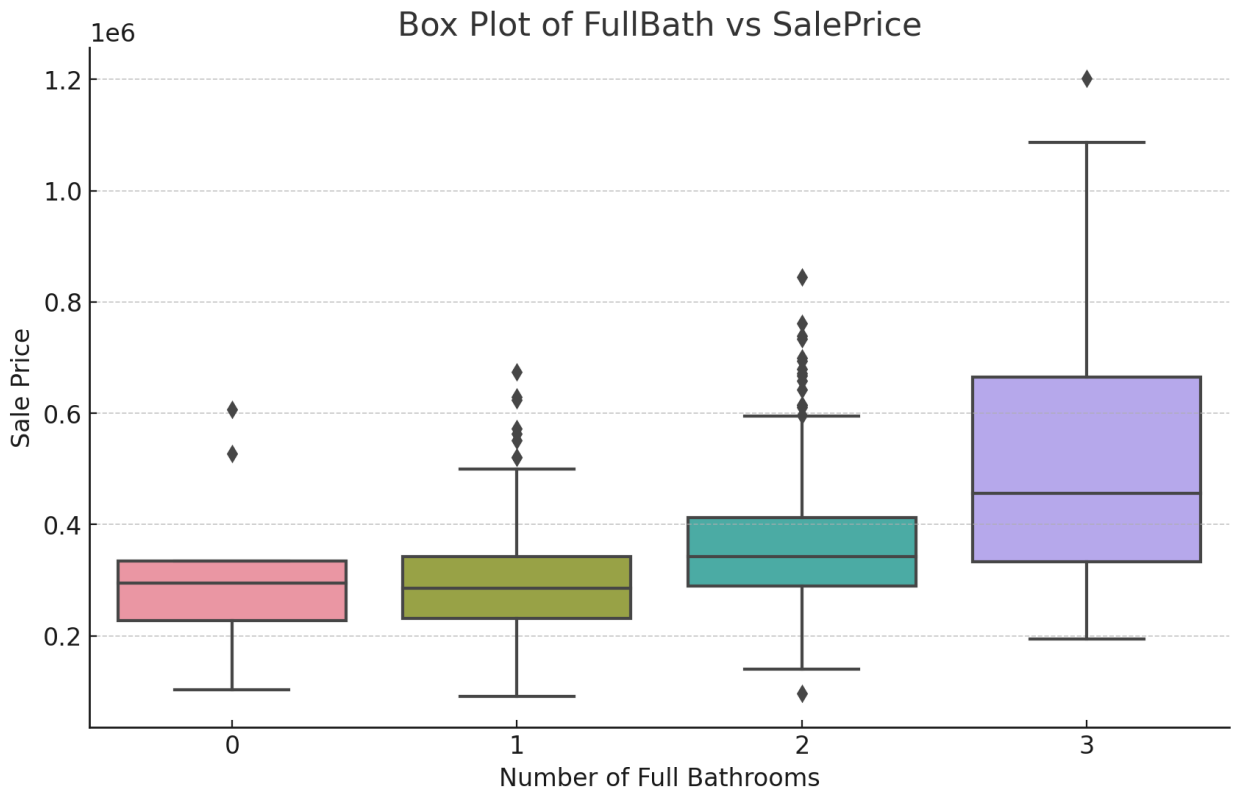
## Number of Full Baths (FullBath)

FullBath indicates the number of full bathrooms. The scatter plot of 'FullBath' vs 'SalePrice' and the correlation coefficient provide insights into their relationship.




The scatter plot shows a pattern where the sale price tends to increase with the number of full bathrooms. Each distinct vertical line represents properties with a specific number of full bathrooms. This pattern is expected, as 'FullBath' is a categorical variable with discrete values (0, 1, 2, 3). There is an upward trend, indicating that properties with more full bathrooms generally command higher sale prices. The correlation coefficient between 'FullBath' and 'SalePrice' is approximately 0.387. This value suggests a moderate positive correlation, indicating that an increase in the number of full bathrooms tends to be associated with an increase in the sale price of a property.

There is a moderate correlation between the number of full bathrooms ('FullBath') and the sale price ('SalePrice'). Properties with more full bathrooms typically have higher sale prices, reflecting the value that additional bathrooms add to a residential property. However, as 'FullBath' is a categorical variable, the relationship is not linear but shows a general trend of increasing value with additional bathrooms. But to determine the optimal number of full bathrooms, we also need to consider the transaction volume involves understanding how the number of full bathrooms in a house correlates with its sale price and the frequency of such configurations in the market.



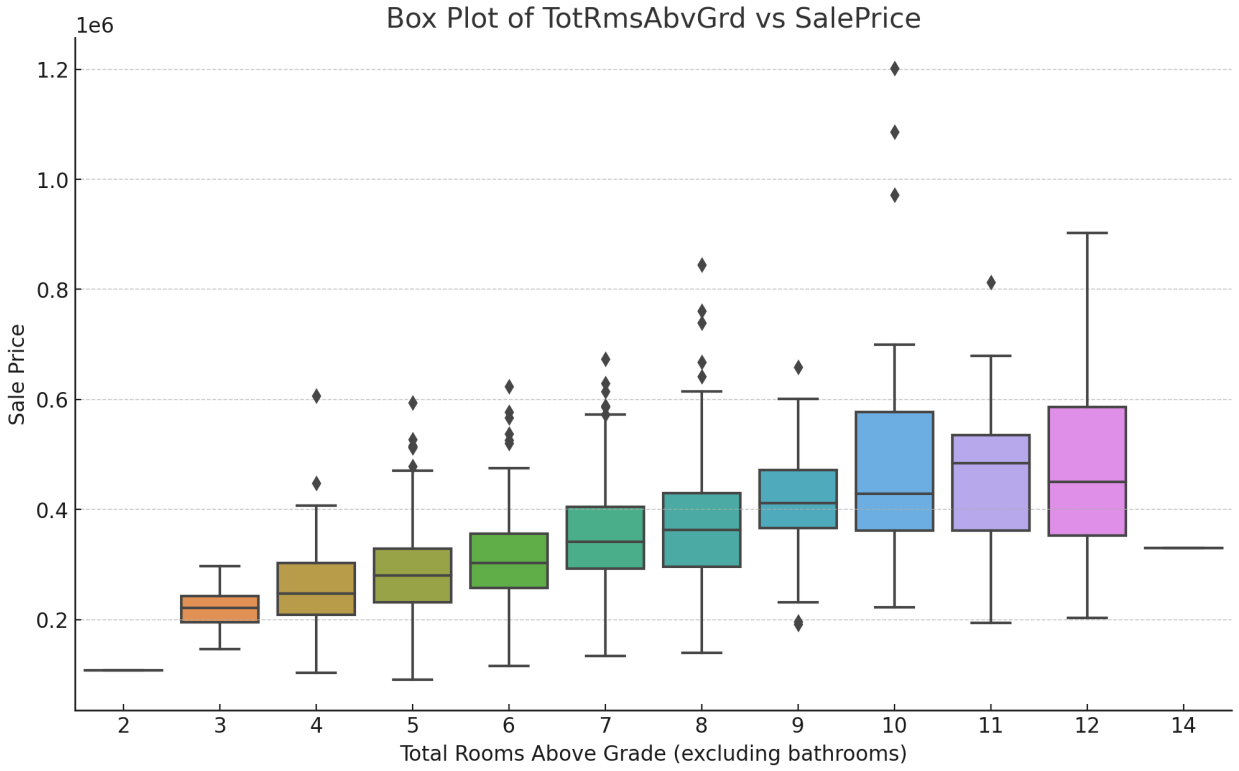

The box plot of 'FullBath' versus 'SalePrice' and the distribution of 'FullBath' provide valuable insights. The increase in sale price from 2 to 3 full bathrooms is noticeable, but it's also accompanied by a wider spread in prices and more outliers, indicating greater variability in sale prices for homes with 3 full bathrooms. 2 Full Bathrooms have 768 transactions (most common), while 3 Full Bathrooms have 33 transactions (less common, more variability). Based on the analysis, the optimal number of full bathrooms for achieving a good balance between a higher sale price and reasonable transaction volume appears to be 2 full bathrooms. This category is not only the most common in the dataset, suggesting a high volume of transactions, but also shows a significantly higher median sale price compared to homes with just 1 full bathroom.

While homes with 3 full bathrooms tend to have even higher sale prices, they are much less common and show more variability in sale prices, which might reflect a niche market with specific buyer preferences.

Thus, for a property aiming for a good market value and saleability, having 2 full bathrooms seems to be a practical and desirable feature in the current market. What is interesting is the connection between the number of full bathrooms and the total number of rooms above ground.

## Total Rooms Above Ground (TotRmsAbvGrd)

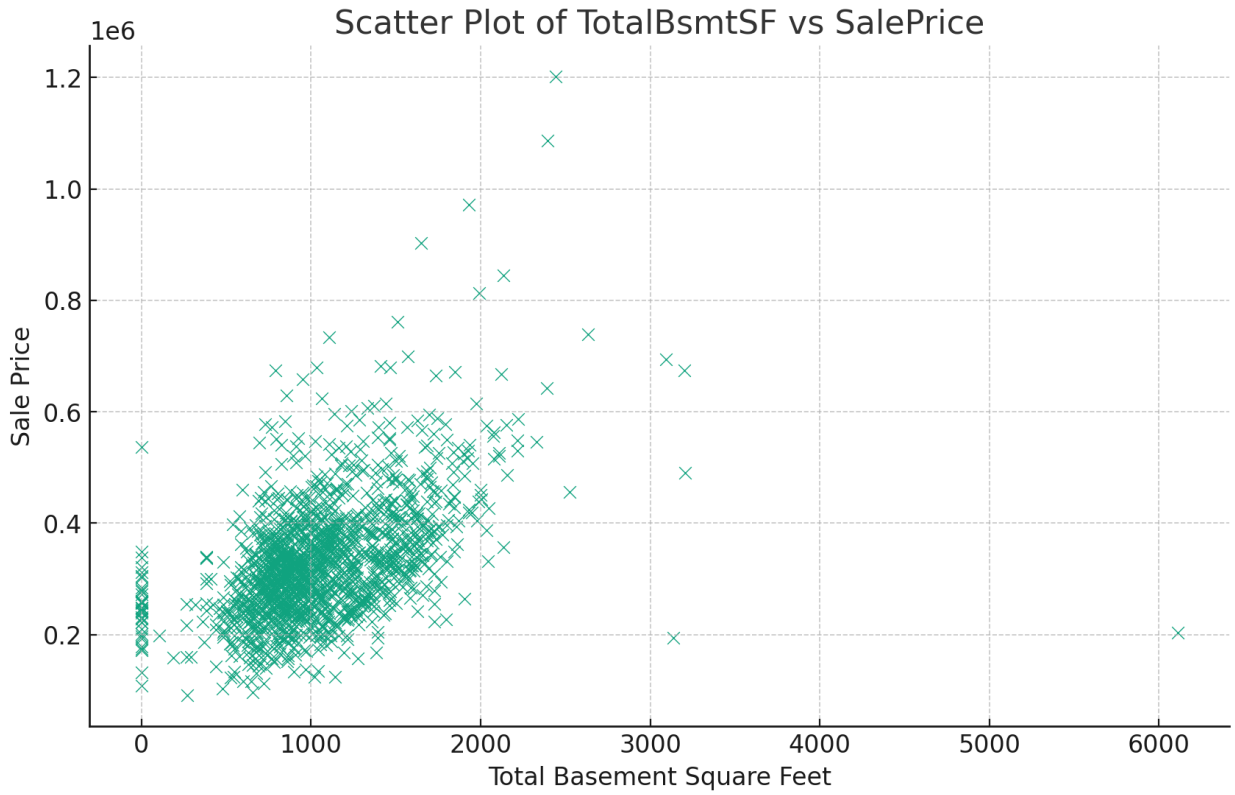
Let's examine the relationship between 'TotRmsAbvGrd' and 'SalePrice', and then analyze the distribution of 'TotRmsAbvGrd' in the dataset.




The box plot shows a general trend where 'SalePrice' tends to increase with the number of rooms above grade. Houses with a higher count of rooms typically have a higher median sale price, suggesting that more spacious homes with more rooms are valued higher in the market. Houses with 5 to 7 Rooms are more common, with 6 rooms being the most frequent. The frequency decreases as the room count increases past 8, with very high counts (like 11 or 12 rooms) being quite rare. Based on the analysis, the optimal number of rooms above grade for achieving a good balance between a higher sale price and a reasonable transaction volume appears to be 6 to 7 rooms. This range is not only common in the dataset, suggesting a high volume of transactions, but also shows a higher median sale price compared to homes with fewer rooms. Homes with 8 or more rooms, while fetching higher prices, are less common and might cater to a more specific, possibly niche market. In conclusion, properties with 6 to 7 total rooms above grade represent a desirable balance for a robust housing market, offering a blend of space and marketability. Another important factor that needs to be accounted for, when considering the total rooms of a house, is not just the rooms above ground, but the total square footage of the basement as well.

## Total Basement Square Footage (TotalBsmtSF)

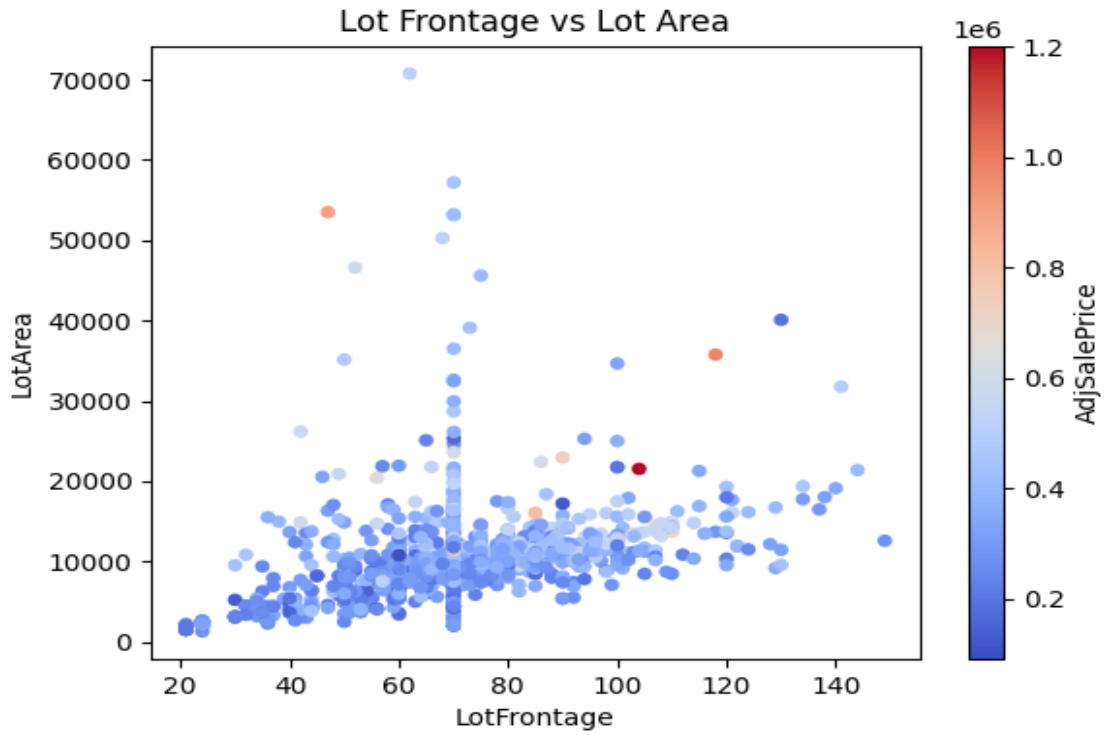
Total Basement Square Feet represents the total square feet of the basement. To understand the relationship of 'TotalBsmtSF' with the 'SalePrice,' we explored how variations in these property features typically affect the sale price.



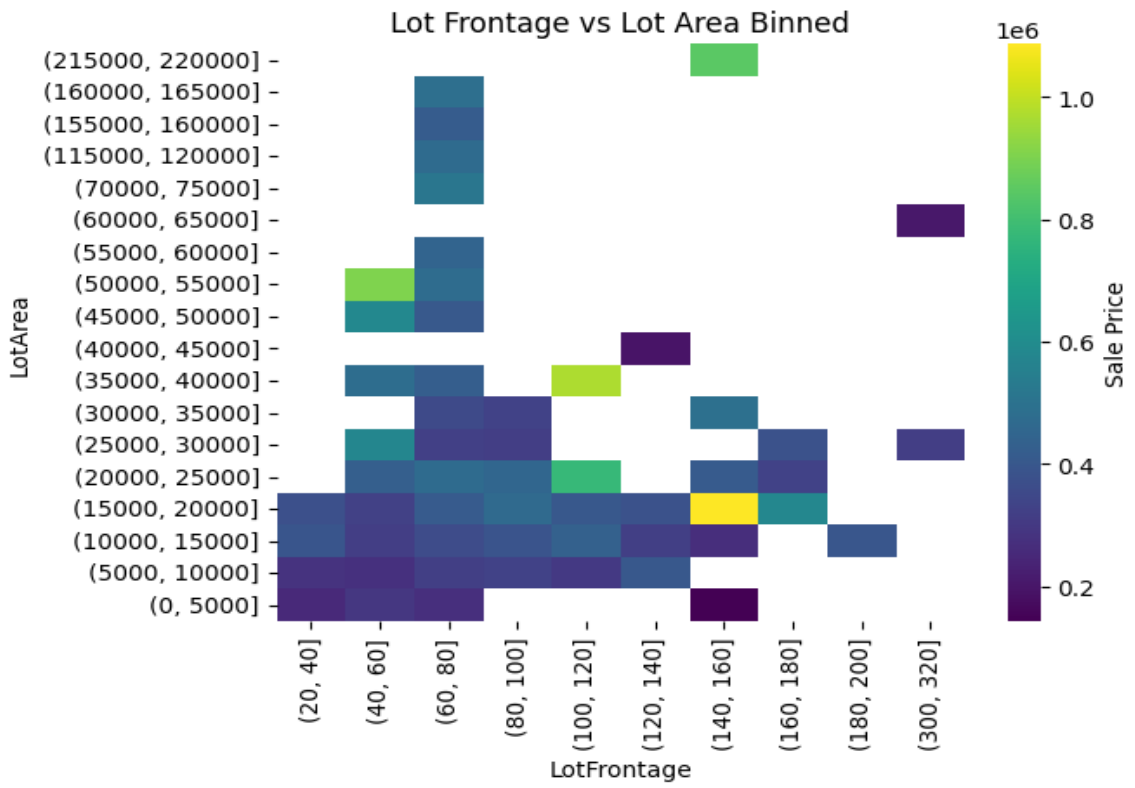

The scatter plot visually confirms a moderate positive correlation, showing that properties with larger basements generally have higher sale prices. The reason could be that a large basement could have the potential for additional living space, storage, or amenities like home gyms or used as an entertainment room. The plot also indicates a fairly linear relationship, with more variation in sale prices at higher basement sizes. The optimal range of 'TotalBsmtSF' for achieving the best sale price with a reasonable volume of transactions would likely be around the median to slightly above the 75th percentile. This range (about 992 to 1300 sq ft) represents a practical and desirable basement size for a majority of buyers. A more detailed analysis of GrLivArea can be found in Appendix D. A restriction on how big a house can be is often the lot size and area that the home is being built on, and it is interesting to see how these factors also play a role in the sales price.

## Lot Frontage and Lot Area (LotFrontage and LotArea)

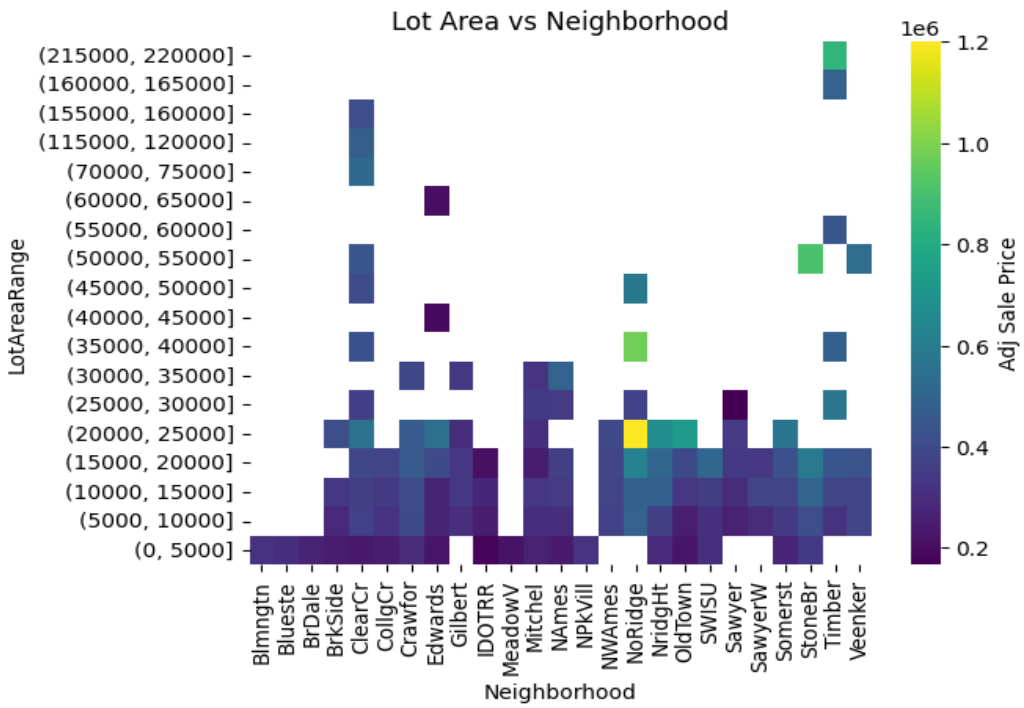
After the initial data preparation and analysis, two columns that were determined to be worth looking deeper into, with respect to Sales Price, were the Lot Frontage and Lot Area. Lot Frontage outlines the total length of the front of the lot, whereas Lot Area is the total area allocated for the specific lot. Both of these data columns were continuous and it would seem to reason that they should have a connection as the lot frontage is just one of the sides used to calculate the total lot area. Below is a comparison of both data sets, including a colour map for the associated sales price.



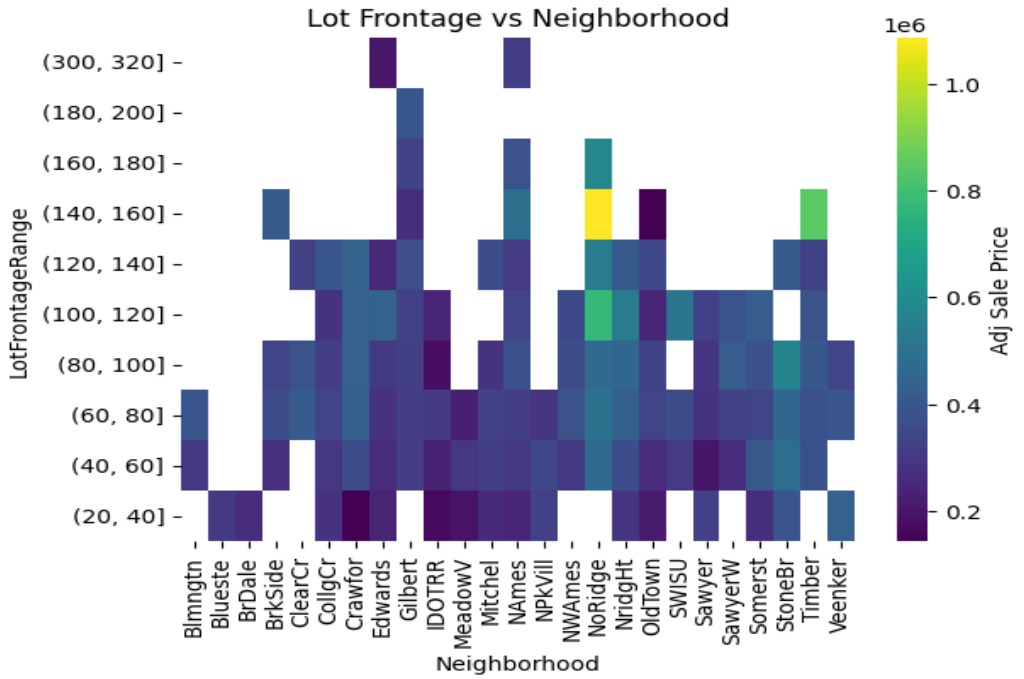

Barring a couple of outliers, the data shows what we expected, that the lot area is heavily connected to the lot frontage. What wasn't as expected was the associated sales prices, as can be seen by the colour map. Because both of these data sets are continuous, to get a better grasp on the associated sales price, the data was binned and compared.




Although both lot area and frontage are correlated, when looking at the associated price, it doesn't seem to line up as much as expected. If we look at the 140-160 lot frontage, it seems that the associated sales prices range from the very lowest to the highest, and not necessarily in the order of larger lot area. This would indicate that maybe these categories aren't as useful when trying to determine the drivers of sales price. This is further illustrated when looking at the lot frontage and lot area with respect to the different neighbourhoods. As can be seen in the plot below, there are only a couple of neighbourhoods which have a strong positive correlation with respect to increasing lot area and increasing sales price.



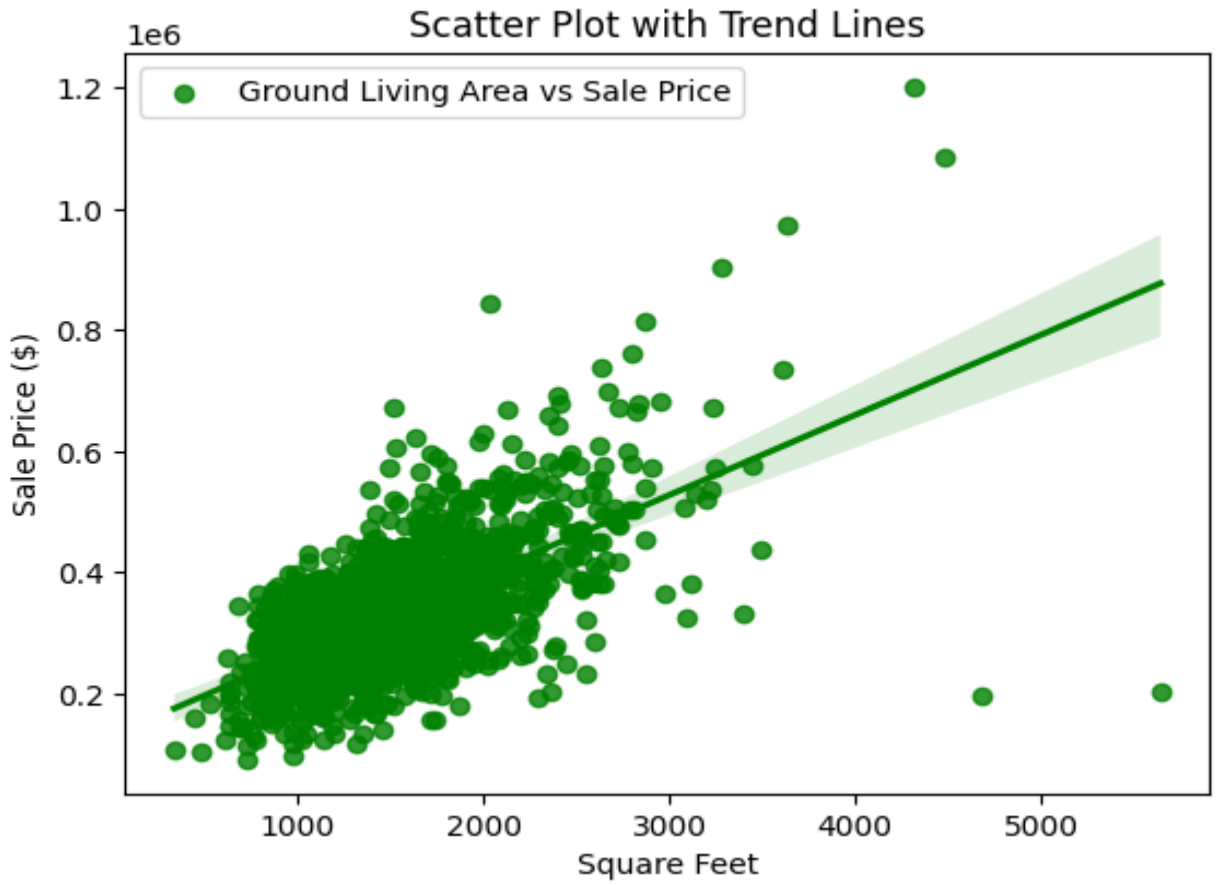

What this data shows is that only in a couple of specific neighbourhoods (ex. NoRidge and StoneBR) will the lot area, and even better still the lot frontage, be a good driving factor for a higher sales price.



What is interesting to note though, is that there is a strong correlation between the lot frontage, lot area, and ground living space (plots not shown). This makes sense as the ground floor living space is restricted by the size of the lot the house is built on, and further highlights the impact that the ground living area should have on the final sales price.

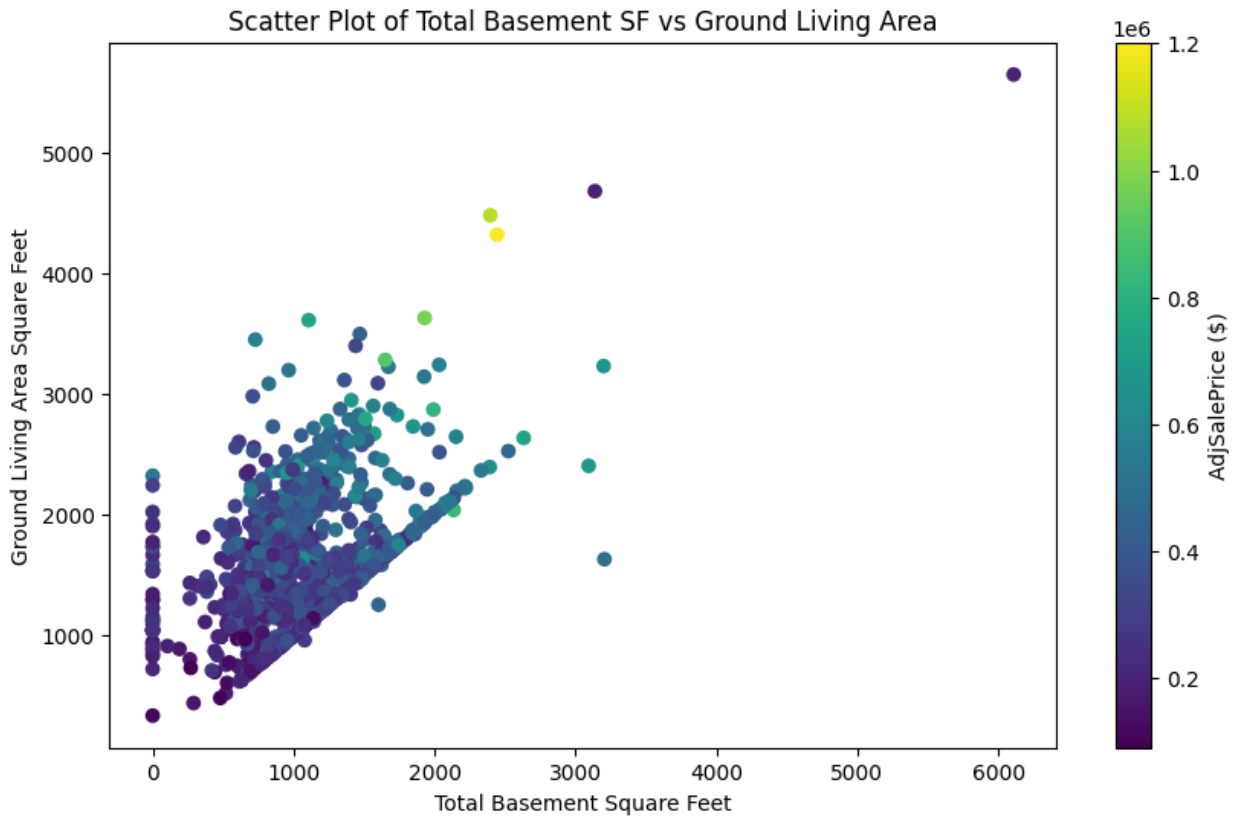
## Ground Living Area (GrLivArea)

There is a strong positive correlation between Ground Living Area Square Feet and Sale Price indicating that properties with larger living areas above the ground typically command higher prices. Larger living areas are often associated with luxury and therefore, a key selling point for families looking for a larger space or spacious homes. The functionality of the living area (i.e., layout, design, natural light) can also influence how this space impacts the sale price. The strong correlation suggests that market preferences lean towards larger living spaces, due to lifestyle trends favoring more open and spacious home designs.

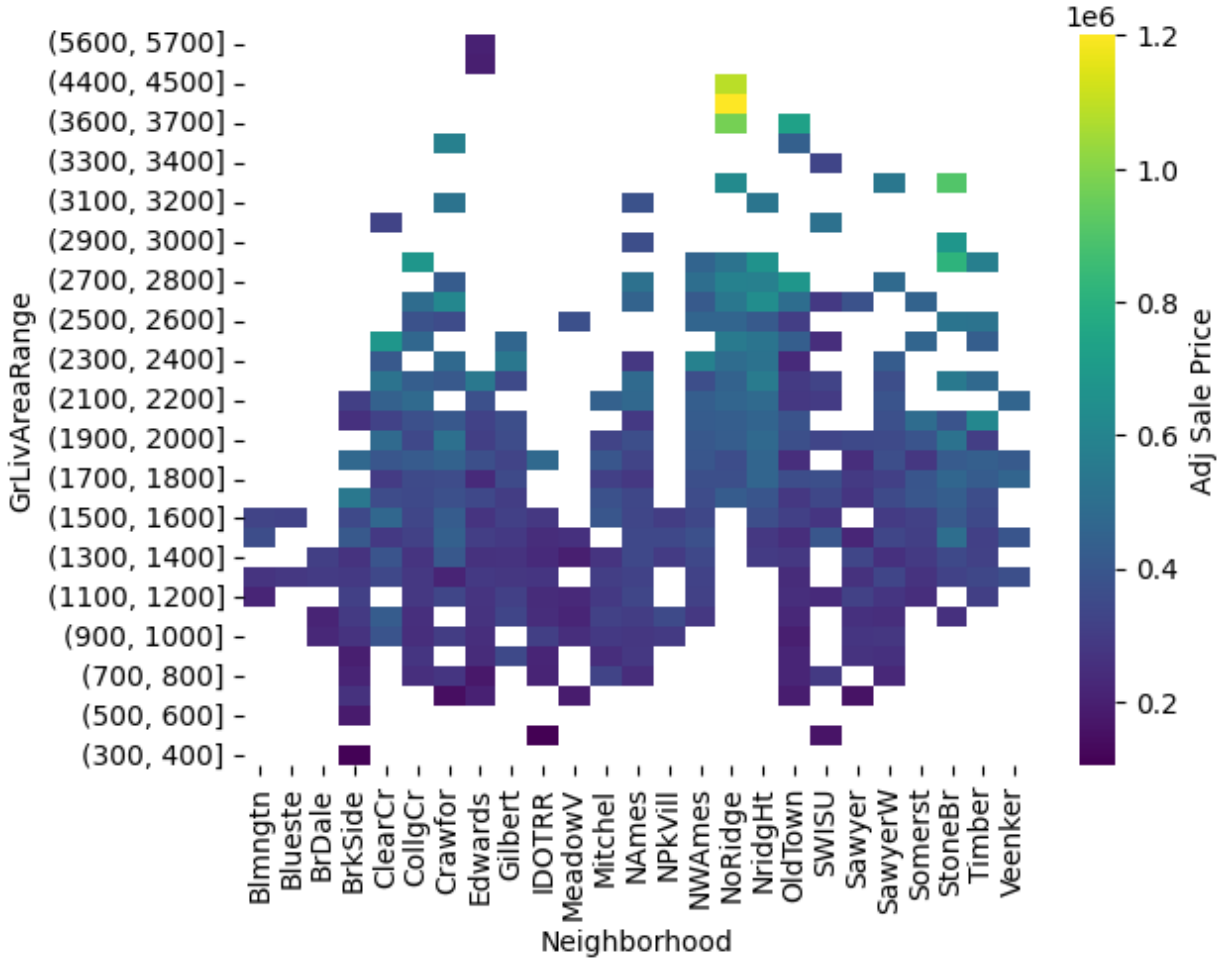



Total Basement Square Feet and Ground Living Area Square Feet highlight a positive correlation, though less pronounced than with Sale Price, suggesting some degree of association between

basement size and living area. In many cases, the basement area and living area above ground complement each other. A large, well-designed basement can compensate for a moderately sized living area and vice versa. Both variables contribute to the overall size and utility of the property, which is a significant factor in determining its market value.




A larger ground living area will fetch a higher price, but we also want to consider the volume. Due to the high price, property with larger ground area is far and between, which is evident in the low volume within our dataset. To find a good balance, we will consider the scatter plot of 'GrLivArea' against 'SalePrice', and then analyze the distribution of 'GrLivArea' to find the optimal range. The scatter plot of 'GrLivArea' vs 'SalePrice' reveals a positive correlation, indicating that larger living areas generally correspond to higher adjusted sale prices. However, this relationship seems to taper off beyond a certain point, suggesting that very large homes do not necessarily translate into proportionally higher sale prices. The optimal 'GrLivArea' range for achieving a high 'SalePrice' with reasonable transaction volume is likely around the median to the 75th percentile. This range (1464 to 1777 sq ft) is typical for many homes and likely to be in demand, ensuring a steady market activity. It is also interesting to note that the ground living area seemed to be agnostic of the neighbourhood chosen. As can be seen below, the larger the ground living area, the larger the sale price for most of the neighbourhoods.




# Conclusions

The comprehensive analysis of various features against 'SalePrice' (Sale Price) in residential properties uncovers nuanced insights into what drives home values. The selected features for this analysis were 'LotFrontage', 'LotArea', 'MSSubClass', 'BsmtQual', 'TotalBsmtSF', 'GrLivArea', 'FullBath', 'TotRmsAbvGrd', 'GarageType', and 'GarageCars'. Among these, the top five features that contribute significantly to a high 'SalePrice' are 'GrLivArea', 'TotalBsmtSF', 'FullBath', 'TotRmsAbvGrd', and 'GarageCars'.

- 'GrLivArea' (Above Grade Living Area): This feature consistently emerges as a crucial determinant of property value. The square footage of the living area above ground is a primary indicator of a home's size and utility. Larger living spaces typically correlate with higher property values, reflecting the premium placed on spaciousness. The optimal range is between 1464 to 1777 sq ft. This is typical for many homes and likely to be in demand, ensuring a steady market activity.

- 'TotalBsmtSF' (Total Basement Square Feet): The size of the basement area is significant, especially in markets where basements are used as additional living space. The optimal range for basement square footage is about 992 to 1300 sq ft. Well-designed and finished basements substantially increase the functional area of a house, contributing to a higher sale price.

- 'FullBath' (Number of Full Bathrooms): The number of full bathrooms in a house is a key feature of convenience and luxury. Properties with more full bathrooms tend to be more desirable, particularly in family-oriented or upscale neighbourhoods, directly impacting their sale price. The optimal number of full bathrooms is 2, this number is a good balance between saleability and sales price.

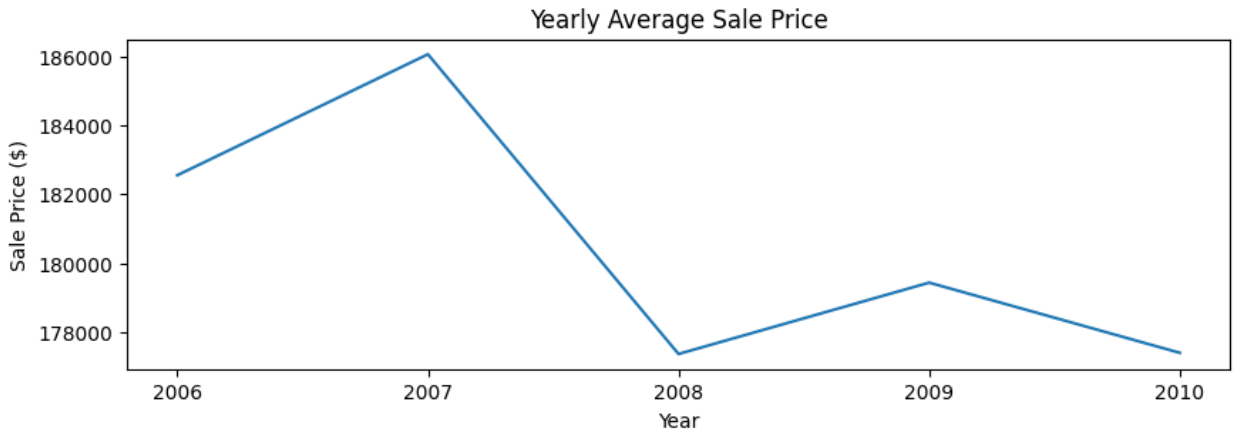
- 'TotRmsAbvGrd' (Total Rooms Above Grade): The total number of rooms, excluding bathrooms, is a measure of a home's capacity to accommodate various needs like bedrooms, offices, and entertainment spaces. In family-centric or affluent areas, the demand for more rooms is usually higher, thus impacting the sale price positively. The optimal number of rooms above grade for achieving a good balance between a higher sale price and a reasonable transaction volume appears to be 6 to 7 rooms.

- 'GarageCars' (Size of Garage in Car Capacity): In many regions, especially those with suburban or rural settings, the size of a garage, indicated by the number of cars it can hold, plays a significant role in home valuation. A larger garage capacity is often associated with higher property values. A 2-car garage strikes a practical balance in the current housing market, offering both desirable property features and strong marketability.

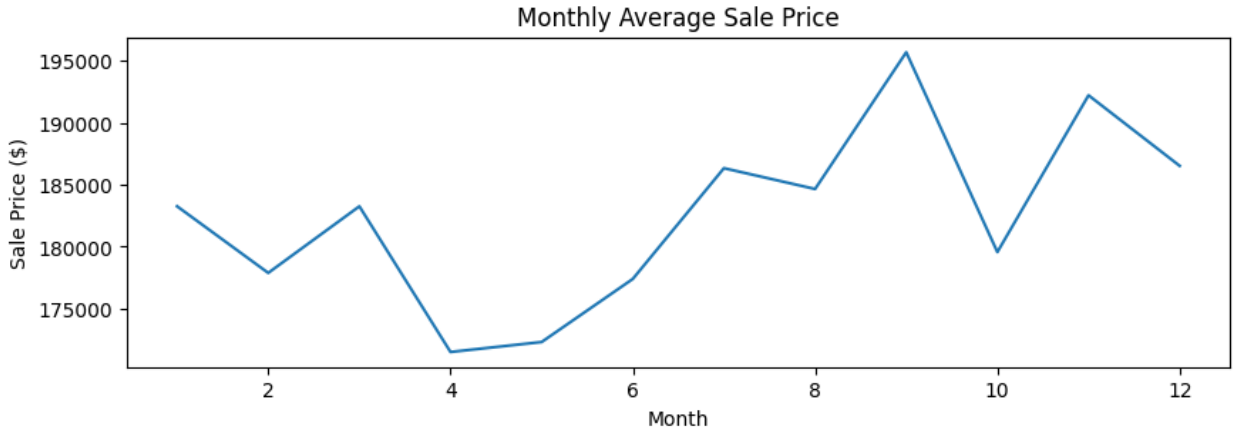
In summary, while each of these features individually influences the 'SalePrice', their impact is often interrelated and can vary by location and market conditions. Features like living area, basement space, and the number of bathrooms and rooms reflect both the functional and luxury aspects of a property, significantly affecting its perceived value. Meanwhile, practical aspects like garage size also play a crucial role, especially in areas where driving is the primary mode of transportation. This analysis underscores the multifaceted nature of property valuation, where a combination of practicality, luxury, and local market dynamics come into play.

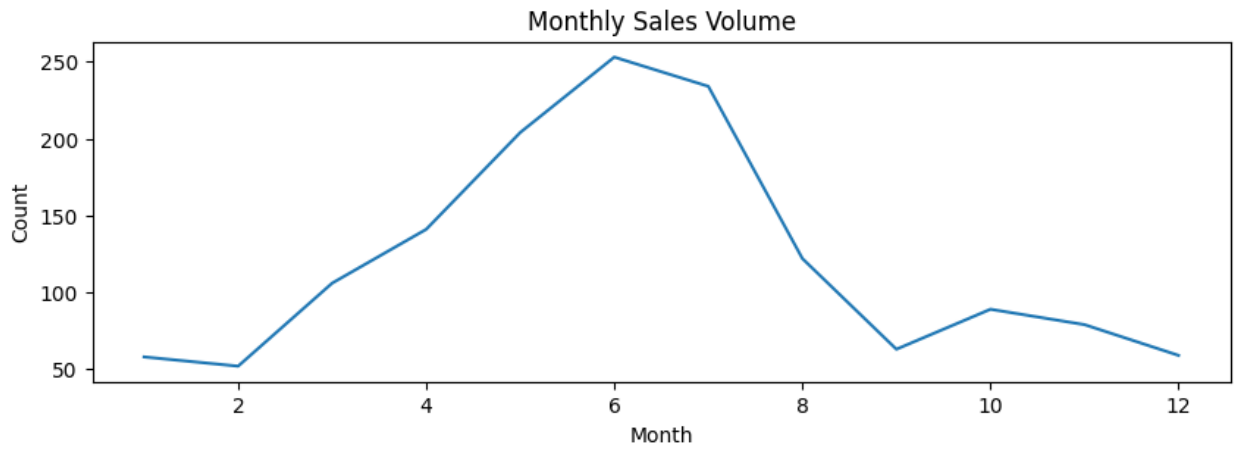
# Appendix A - Adjust sale price to remove temporal influence

The following diagram depicts the average annual sales price across the dataset.




We can see that the 2008 recession impacted the sales price. The following two diagrams depict the average monthly sales price and the monthly sales volume across the dataset.





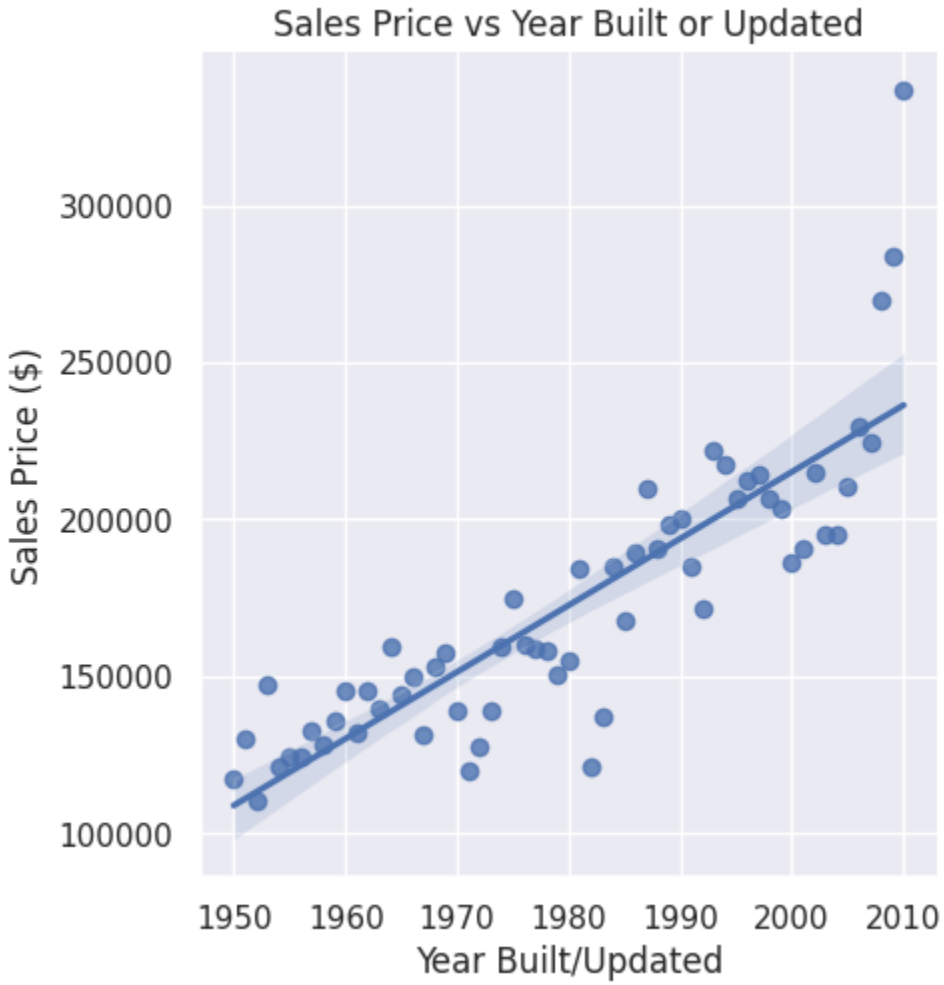

Besides the impact due to the greater economic condition, we are also observing seasonality with the data, where there are more sales during the summer month, the abundance of the supply contributes to the lower sales prices.

We decided to adjust the sales price. We calculated the average of the monthly average as the baseline. Then we calculate how the individual monthly average deviates from the overall average. The deviations are then used to adjust the sales prices of associated records. The code for the adjustment is shown below.

```
df = pd.read_csv('train.csv')

# build df_monthly_ratio, which is the average sale price for
# each year-month
df_monthly_ratio = pd.DataFrame(df.groupby(['YrSold',
'MoSold'])['SalePrice'].mean().mean() / df.groupby(['YrSold',
'MoSold'])['SalePrice'].mean())
df_monthly_ratio.rename(columns={'SalePrice': 'AdjRatio'}, inplace=True)

# merge df with df_monthly_ratio, so the adjustment ratio is
# populated for each row
df_adjusted = df.merge(df_monthly_ratio, how='left', on=['YrSold',
'MoSold'])

# add a new column and calculate the adjusted price
df_adjusted['AdjSalePrice'] = df_adjusted['SalePrice'] *
df_adjusted['AdjRatio']
```

# Appendix B - Adjust sale price to remove depreciation

The following diagram depicts the average sales price versus the year the house was built or last renovated. We can see there's a positive correlation, the newer the house the higher the price. However, this correlation is not very useful for a construction company to decide what features to include in their house.

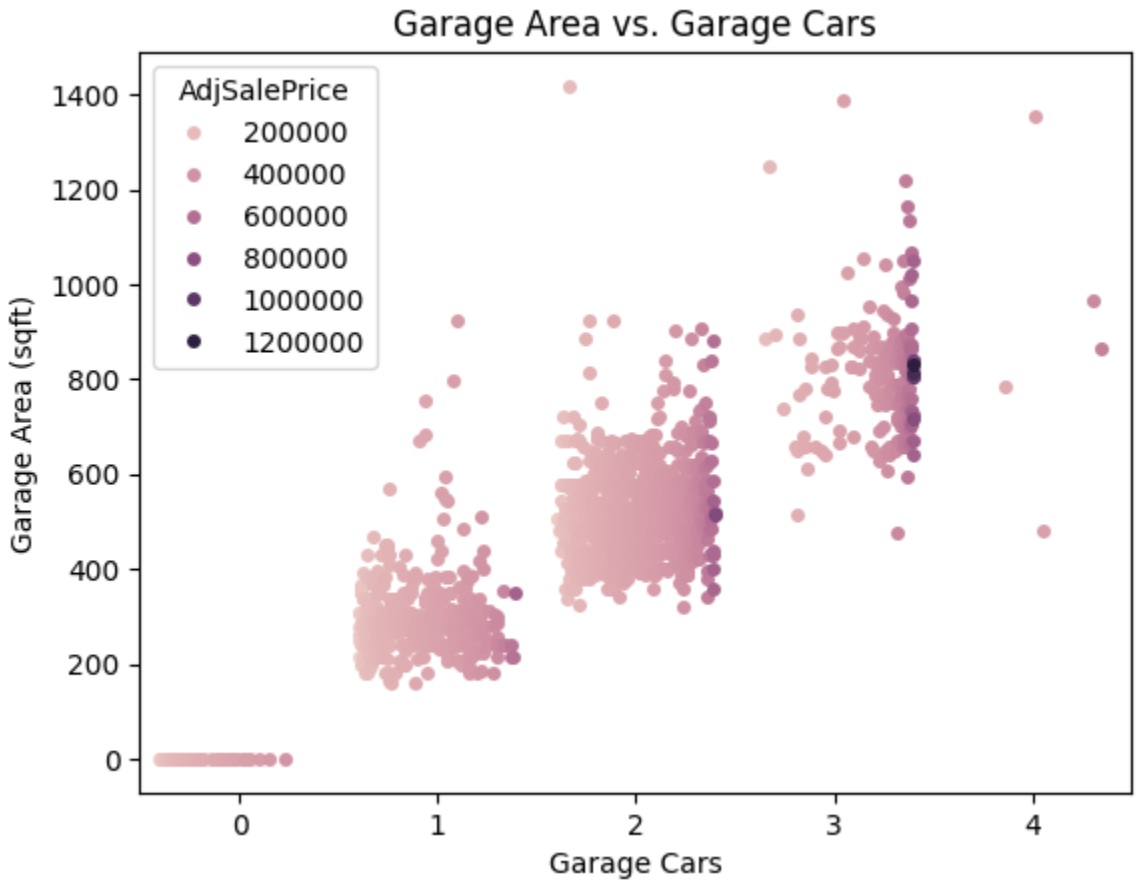



We took a similar approach to adjust the price as Appendix A.

```
# use the 2010 price as baseline
df_ratio                                                                =
pd.DataFrame(df.groupby('YrLastUpdated')['AdjSalePrice'].mean().loc[2010]
/ df.groupby('YrLastUpdated')['AdjSalePrice'].mean())

# calculate the adjustment ratio
df_ratio.rename(columns={'AdjSalePrice':            'AdjRatioYrUpdated'},
inplace=True)
```

```
# merge the adjustment ratio with the original data set
df_adjusted = df.merge(df_ratio, how='left', on=['YrLastUpdated'])

# adjust the price
df_adjusted['AdjSalePrice2']      =      df_adjusted['AdjSalePrice']      *
df_adjusted['AdjRatioYrUpdated']

# remove and rename columns
df_adjusted.drop(columns=['AdjRatio',    'AdjSalePrice',    'YrLastUpdated',
'AdjRatioYrUpdated'], inplace=True)
df_adjusted.rename(columns={'AdjSalePrice2':              'AdjSalePrice'},
inplace=True)
```

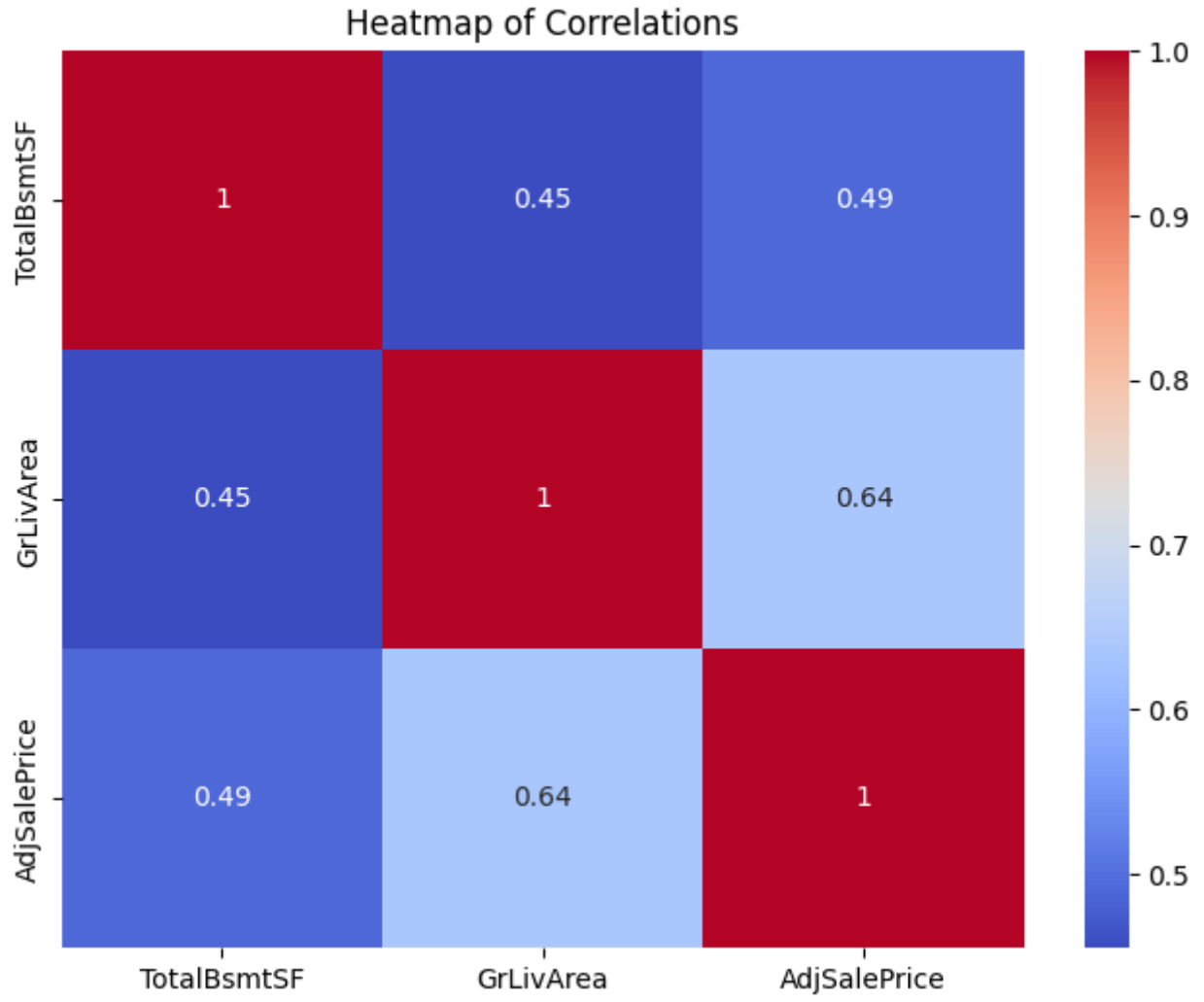
# Appendix C - GarageCars vs GarageArea

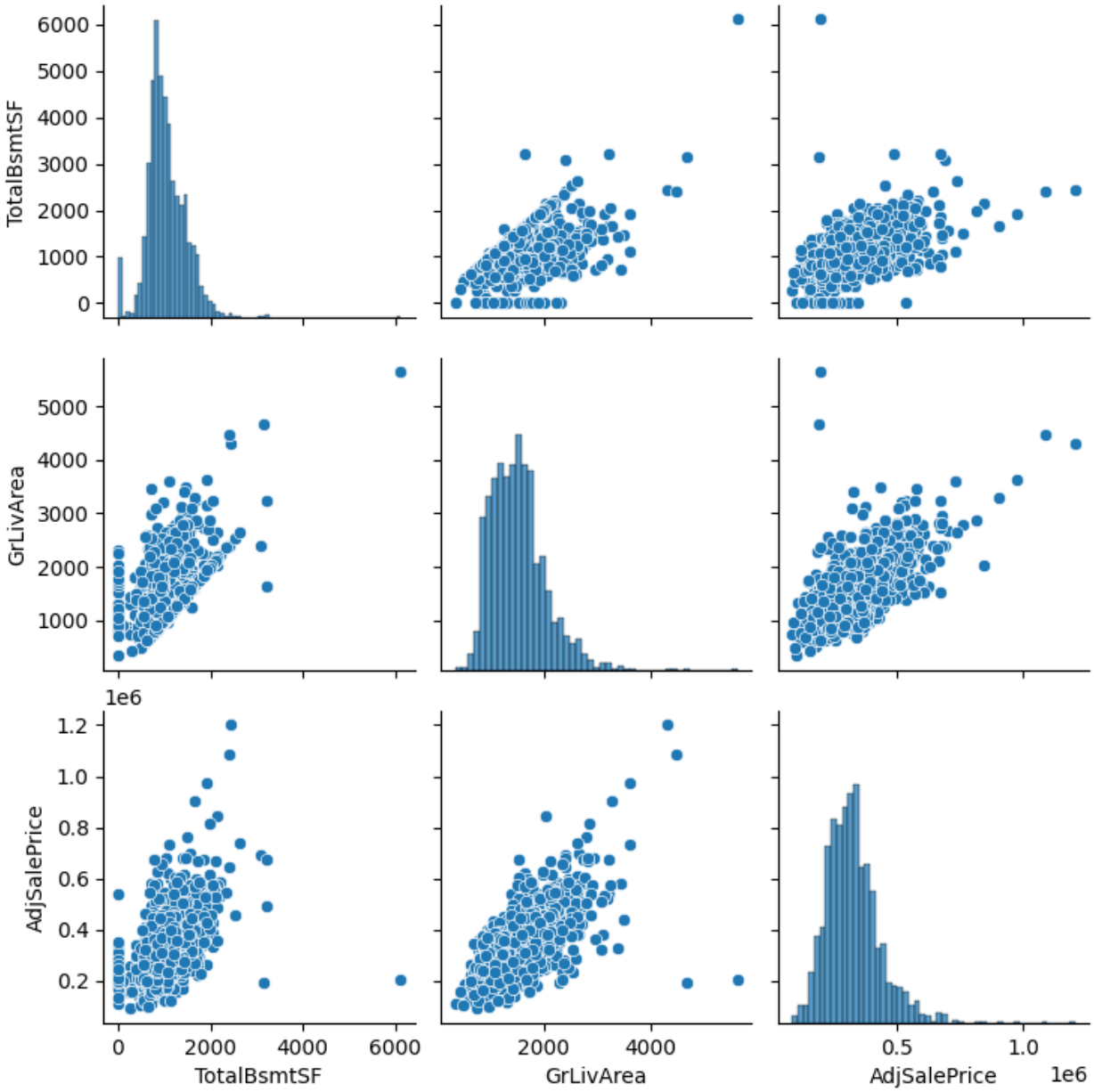



There are clear correlations between garage cars and garage areas; however, there seems to be a clearer correlation between garage cars and sale price compared to the garage area and sale price.

# Appendix D - TotalBsmtSF and GrLivArea Correlation Analysis

Total Basement Square Feet represents the total square feet of the basement area whereas the Ground Living Area Square Feet measures the living area square feet above the ground level. To understand the relationship of these variables with the sale price, we explored how variations in these property features typically affect the sale price. Total Basement Square Feet and Ground Living Area Square Feet are both continuous variables, so we explored a correlation analysis (i.e., correlation coefficient and pair plot) of each against the sale price. Below are the charts of the correlation analysis. Trend lines in these plots help in visualizing the strength and direction of the relationships.




TotalBsmtSF
GrLivArea
AdjSalePrice
1e6
TotalBsmtSF
GrLivArea
AdjSalePrice
1e6