Project Report

Ali Alwahdani & Taka Oya

Ames, Iowa Housing Market Analysis

---

1. Introduction

The real estate market in Ames, Iowa, presents a unique opportunity to understand the dynamics that influence property values. In this study, we delve into the complex interplay of factors that determine these values, guided by a comprehensive dataset from Kaggle detailing residential property sales from 2006 to 2010, and supplemented by current market data from Amesrealestate.com. Our investigation revolves around three pivotal research questions: Firstly, how has the distribution of sale prices changed from the period of 2006-2010 compared to the present market? Secondly, what are the distinguishing characteristics of houses in the top and bottom 25% price brackets in the Kaggle dataset? Lastly, we seek to understand the relationship between property size (square footage) and sale price across both datasets. These questions aim to shed light on the evolving nature of real estate values in Ames, offering valuable insights for a range of stakeholders.

2. Data - ***Due to length, data dictionary is provided at the end of the report***

The data utilized in this study comprises two primary sources. The Kaggle dataset provides an extensive overview of individual residential property sales in Ames from 2006 to 2010, capturing a variety of details such as architectural features, conditions, and amenities. The current data, on the other hand, offers a contemporary perspective of the market, enabling us to compare past and present trends.

2.1 Kaggle Data

In our analysis of the Ames real estate market, we performed extensive data cleaning and feature engineering on the Kaggle dataset. The dataset, titled "wrangling_project_kaggle.csv", was initially loaded into our R environment. Our first step involved a detailed examination of the data structure, which was crucial for understanding the types and nature of data we were dealing with.

The cleaning process began by iterating through each column in the dataframe. For numeric columns, we ensured that all non-NA values were correctly formatted as integers where applicable, and as numeric types otherwise. Character columns were converted to factors to facilitate categorical analysis. This step was crucial for ensuring data integrity and appropriateness for our analytical methods.

Our feature engineering process focused on simplifying and extracting meaningful insights from the dataset. We created new columns that better represented the data in a more analyzable form. This transformation allowed us to analyze more relevant features of the properties.

After creating these new columns, we nullified the original columns from which they were derived to avoid redundancy in our analysis. Finally, we confirmed the changes by reviewing the structure and the first few rows of the modified dataframe. The cleaned and enhanced dataset was then saved as "kaggle_data_CLEANED.csv", ready for further analysis in our study. The final dataframe contains 1,460 observations of 78 variables.

2.2 Ames Real Estate Website Data

We employed web scraping techniques to gather current data from the Ames Real Estate website. This process was crucial to acquire up-to-date information on property sales in Ames, enhancing our analysis with contemporary market insights. We initiated our data collection by setting up a user agent to ensure successful scraping and to mimic a typical browser request, thereby avoiding potential access issues.

Our scraping loop was designed to navigate through multiple pages of the website, systematically collecting key property data such as sale prices, bedrooms, bathrooms, and square footage. We extracted this information from the HTML content of each page using the xml2 library in R, ensuring accurate and comprehensive data retrieval. After scraping, we combined the data into a structured dataframe named “ames_current”.

Post-scraping, we undertook a thorough cleaning process to refine the data. This involved removing non-numeric characters from the sale prices, bedrooms, bathrooms, and square footage columns, and trimming any whitespace. We then converted these columns to integer data types to facilitate numerical analysis. This cleaning process was vital for maintaining data integrity and ensuring its usability for statistical analysis.

Additionally, we calculated a new variable, “Price Per SqFt”, by dividing the sale price by the square footage, and rounded it to the nearest whole number. This metric provided a standardized measure to compare properties of different sizes and values.

Finally, we verified the structure of the cleaned dataset and saved it as 'ames_current.csv'. This file represents a snapshot of the current real estate market in Ames, serving as an important component of our comparative analysis with historical data from the Kaggle dataset. This dataframe contains 65 observations of 6 variables.

2.3 Data Merging

The last step in preparing data for analysis was our merging process. While merging was important, most of the analysis was done with separate dataframes as it prevented errors when comparing across time periods yet was used to confirm findings often during the final analysis and give us a comprehensive view of summary statistics when examining both time periods combined.

We began by loading the pre-processed 'ames_current.csv' file, which contained the current real estate data, and removed any unnecessary columns, such as the automatically generated 'X' column from the CSV read process. Next, we imported the cleaned Kaggle dataset from 'kaggle_data_CLEANED.csv', focusing on selected columns that were critical for our analysis, namely 'SalePrice', 'TotalRoomsExclBaths', 'TotalBathrooms', 'TotalSqFt', and 'PricePerSqFt'.

To ensure consistency across datasets, we rounded the 'PricePerSqFt' values in the Kaggle data to the nearest whole number, aligning it with the format in the Ames current data. We then proceeded to rename the columns in the Kaggle dataset to match those in the Ames current dataset. This step was crucial for seamless vertical merging, as it ensured that the corresponding columns from both datasets were correctly aligned.

After confirming that there were no inconsistencies in the column names between the two datasets, we merged them using the rbind function in R. This function appended the rows of the Kaggle data beneath those of the Ames current data, creating a comprehensive dataset that included both historical and current market information.

The final merged dataset, named 'all_data_merged', represented a complete picture of the Ames real estate market, spanning from 2006 to the present. We then exported this unified dataset to a CSV file, 'all_data_merged.csv', making it readily available for in-depth analysis and interpretation in subsequent stages of our study.

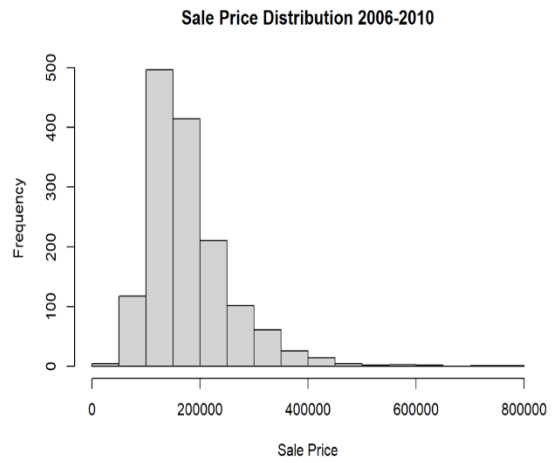
3. Analysis

3.1 Comparative Analysis of Property Sale Price Distribution: 2006-2010 vs. Current Market

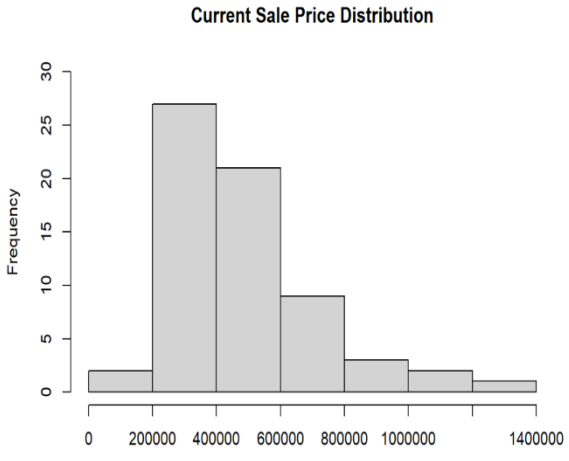
To investigate how the distribution of sale prices in the Ames real estate market has evolved from 2006-2010 to the present, we employed statistical analysis techniques, focusing on identifying shifts in market trends over time. The analysis began with the calculation of correlation coefficients to explore the relationship between historical and current sale prices.

We constructed histograms to compare the frequency distribution of sale prices from both datasets. As shown in Figure 2, the histogram revealed a broadening of the price range in the current market, with an increased number of properties sold at higher price points. This expansion suggests a diversification in the types of properties sold, possibly reflecting a growing market and changes in consumer preferences.

To quantify the shift in sale prices, we computed the mean and median values for both periods. The mean sale price showed a notable increase, underscoring the overall rise in property values. However, it was important to consider the potential impact of outliers on the mean. Therefore, the median, a more robust measure in the presence of outliers, was also considered, which reinforced the findings indicated by the mean.

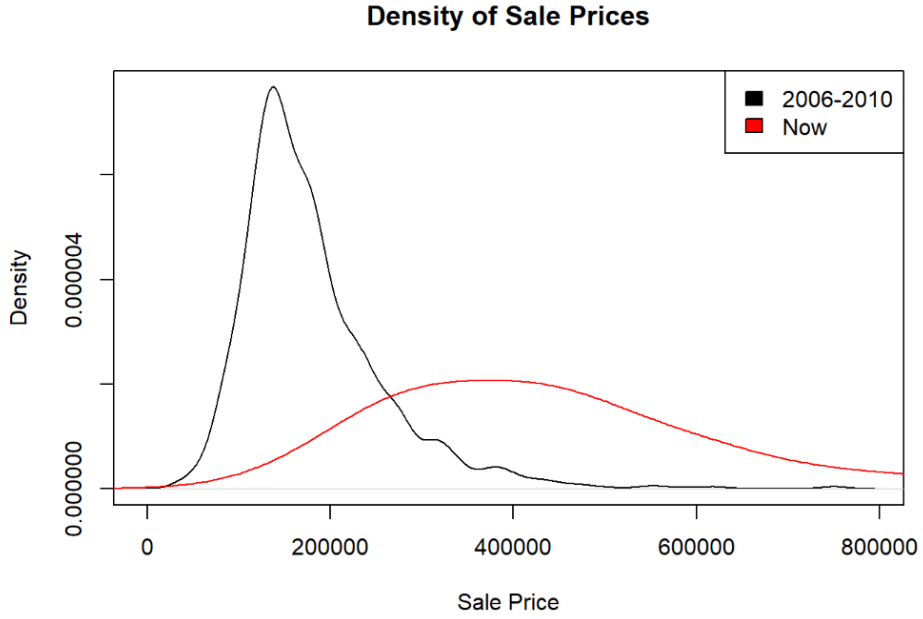



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34900  129975  163000  180921  214000  755000
```




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 144900  314900  428900  478321  575000 1339000
```

Furthermore, this density chart underscores a significant change in the Ames property market. The initial steep peak for the 2006-2010 period indicates a market dominated by mid-range properties with less variance in sale prices. Such a pattern often points to a robust middle-market where standard family homes predominate without a significant presence of either low-cost or luxury housing.



In contrast, the flatter and broader current market distribution suggests that Ames has experienced economic growth or demographic shifts that have introduced a wider array of property values. This could be indicative of several trends: an increase in high-value property transactions, perhaps due to an influx of affluent residents or a boom in luxury home constructions; an expansion of the market to include more entry-level homes catering to first-time homebuyers; or even a reaction to changing market demands, such as a growing preference for larger homes or those with more amenities.

This observed market broadening could also reflect national economic trends, such as inflation or increased lending activity that has made home ownership accessible to a wider segment of the population. For market analysts and prospective buyers, the implications are clear: the Ames real estate market has diversified significantly, offering a wider range of investment opportunities and housing options.

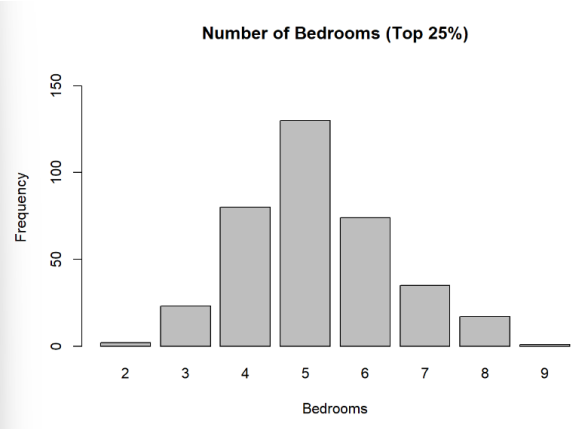
In conclusion, our analysis revealed a distinct evolution in the Ames real estate market, characterized by an increase in median sale prices and a broadening of the price range. These findings are crucial for understanding the current market dynamics and can inform future real estate investments and policy decisions in Ames.

3.2 Characteristics of Houses in Different Price Brackets (2006-2010)

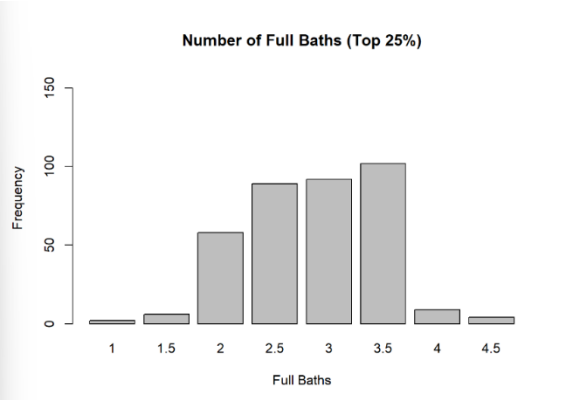
Our investigation into the characteristics that distinguish the top and bottom quartiles in housing prices has yielded significant insights. By employing a series of bar plots and box plots, we have uncovered key distinctions in property features that correlate with housing prices.

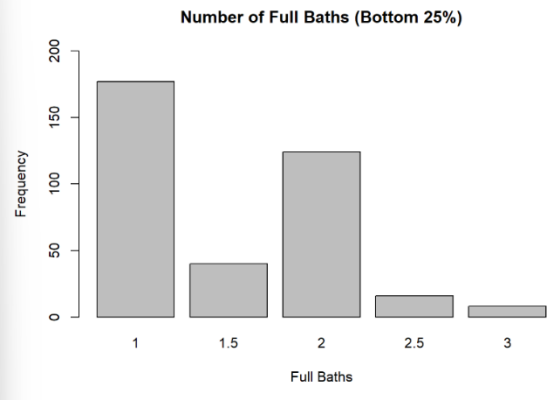
Number of Bedrooms and Bathrooms:

The bar plots depicting the number of bedrooms for both the top and bottom 25% reveal a marked difference in distribution. The top quartile tends to feature homes with a greater number of bedrooms, peaking at 5 bedrooms, which may indicate a preference for larger families or higher value placed on extra space. Conversely, the bottom quartile peaks at 4 bedrooms, suggesting smaller household sizes or more economical space usage.



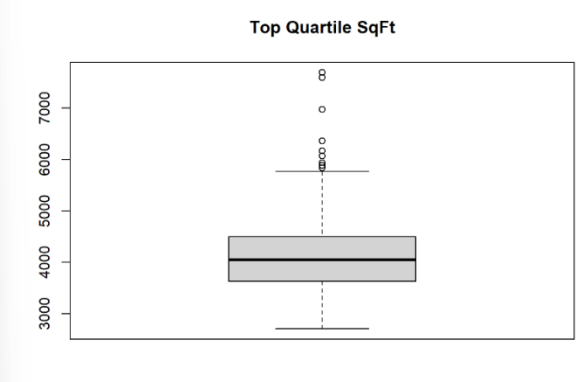
Full bathrooms follow a similar trend, with the top quartile peaking at 3.5 full baths, while the bottom quartile shows a peak at fewer baths. This indicates that homes with more bathrooms tend to be valued higher, a sign of luxury or convenience that contributes to a higher sale price.

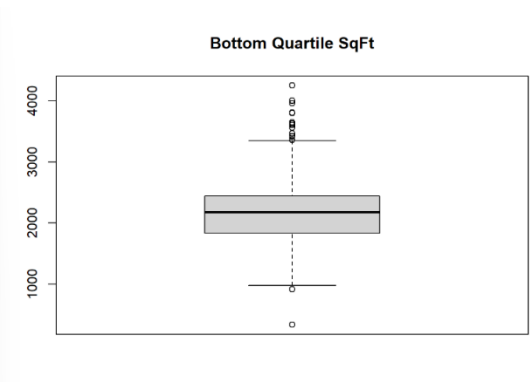




Square Footage:

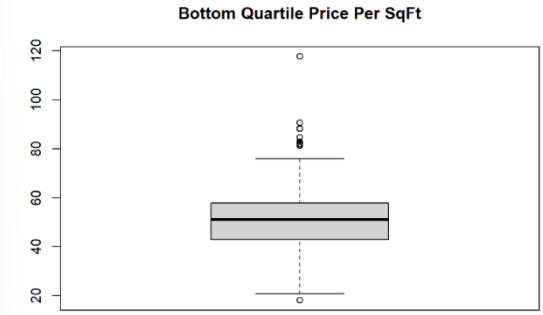
The boxplots for square footage clearly illustrate a higher median square footage in the top quartile, with a notably wider quartile range and outliers indicating a diversity in larger property sizes. This suggests a strong correlation between larger homes and higher market values. In contrast, the bottom quartile has a tighter interquartile range and lower median square footage, reinforcing the trend that smaller homes fall into the lower price brackets.





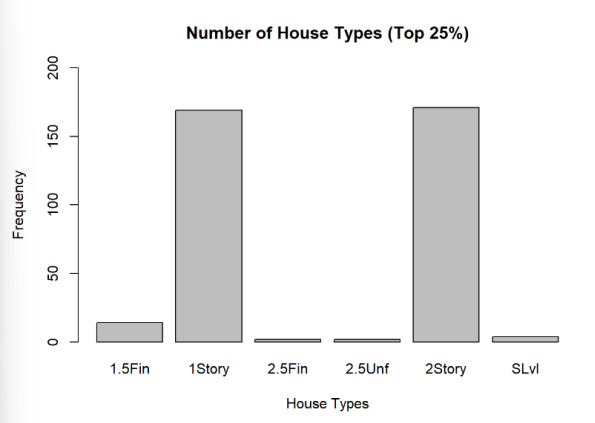
Price per Square Foot:

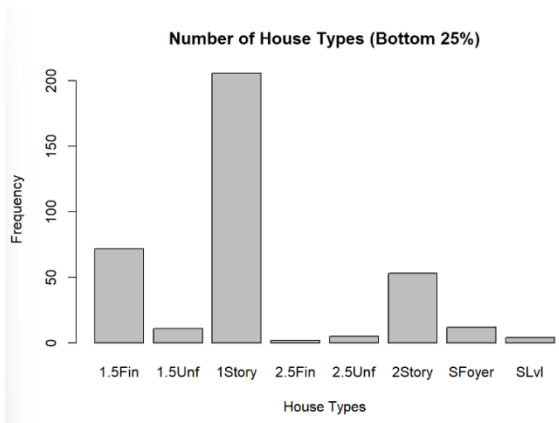
The distribution of price per square foot in the top quartile is presented with a narrower interquartile range compared to the bottom quartile. Notably, the median price per square foot is higher in the top quartile. This implies that bigger homes are also more expensive when normalizing for size than smaller homes. So, for a buyer, you tend to pay more for the size of the home itself as well as more per square foot of the house as well.



House Types:

The variety of house types within each quartile shows distinct preferences or availability in the housing market. The top quartile exhibits a higher frequency of two-story houses, which are often more expensive due to their size and architectural complexity. On the other hand, the bottom quartile shows a slightly greater diversity in house types, with single-story homes being most common, potentially reflecting a market segment that prioritizes affordability and accessibility.





Correlation Analysis:

The correlation matrices reveal that square footage holds the highest correlation with sale price in both the top and bottom quartiles. This underscores the universal importance of size in determining house value across different market segments. Other factors like the number of bedrooms and bathrooms also show significant correlations but to a lesser extent, which implies that while they contribute to property value, the total size of the property is a more decisive factor in the sale price.

```
##                    SalePrice
## TotalRoomsExclBaths 0.3813126
## TotalBathrooms      0.3788223
## TotalSqFt           0.7741334
## PricePerSqFt        0.7123877
## SalePrice           1.0000000
```

Top Quartile

```
##                    SalePrice
## TotalRoomsExclBaths 0.1929584
## TotalBathrooms      0.2590147
## TotalSqFt           0.4586787
## PricePerSqFt        0.2211916
## SalePrice           1.0000000
```

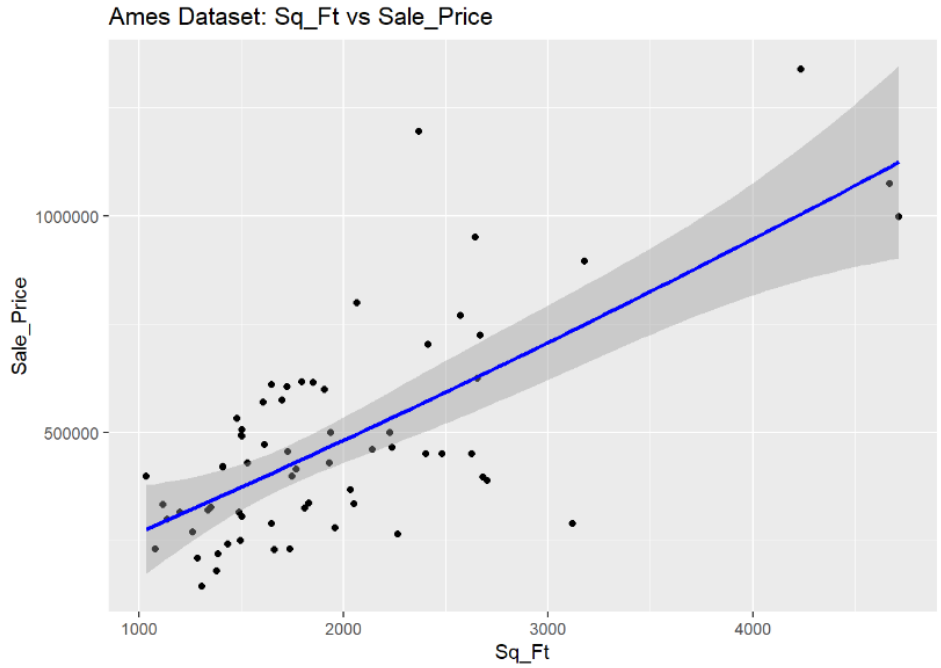
Bottom Quartile

In summary, the analysis of the characteristics of houses in the top and bottom price brackets reveals a consistent pattern where larger property size is strongly associated with higher sale prices. The number of bedrooms and bathrooms also contributes to value but is secondary to the overall square footage. These findings can inform potential buyers, sellers, and investors about which property features are most valued in the housing market. Additionally, the higher price per square foot in smaller homes might reflect a premium on space efficiency or location desirability in the lower quartile. Future analysis might delve into the reasons behind the higher price per square foot in the bottom quartile, including examining the role of location, condition, and market trends.

3.3 Further Analysis of Square Footage to Sale Price

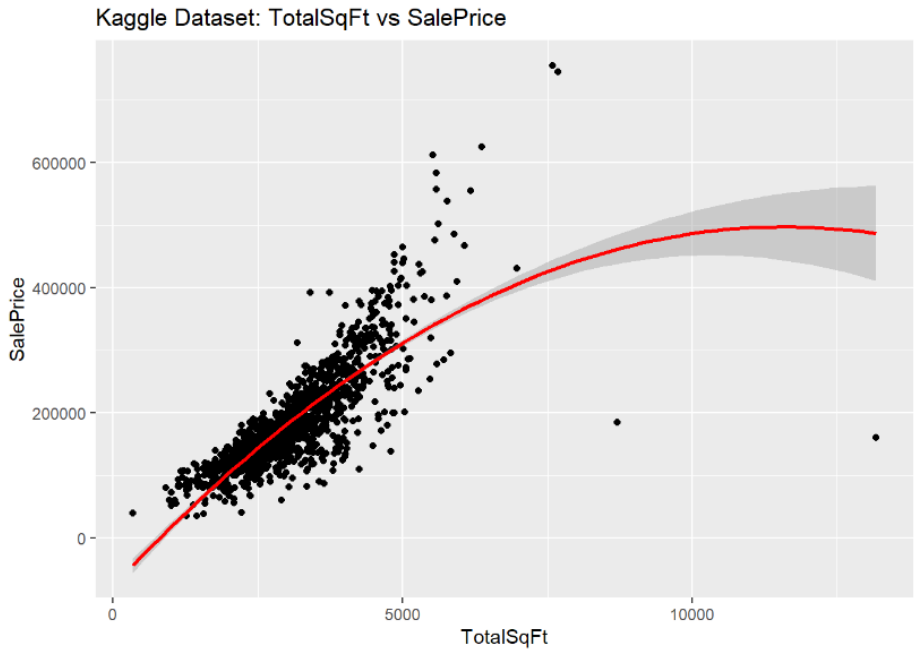
Considering the correlation analysis that highlighted square footage as having the highest correlation with sale price, our investigation for Question 3 was particularly focused on examining this relationship further. The Ames dataset represents current data, whereas the Kaggle dataset encompasses historical data from 2006-2010, providing a comparative view of market trends over time.

The scatter plot from the Ames dataset displays a positive linear relationship between square footage (Sq_Ft) and sale price (Sale_Price). The fitted line slopes upward, reinforcing the well-understood market principle that larger properties generally fetch higher prices. The narrow confidence interval suggests a strong and stable linear correlation, which indicates that property size is a consistent and significant determinant of housing price in the current market.



The scatter plot for the historical Kaggle dataset reveals a more complex, non-linear relationship. The fitted curve suggests that the impact of square footage on sale price increases at a decreasing rate for larger properties. This could be indicative of a market where additional square footage beyond a certain threshold does not proportionally increase property value—a phenomenon often attributed to diminishing returns.

Notably, the plot shows several outliers, particularly for larger properties, which skew the results. These outliers may represent luxury or unique properties that do not follow general market trends. The broader confidence band around the curve implies greater variability in sale prices for properties of similar sizes, reflecting the less predictable nature of the market during the 2006-2010 period.



The comparison of current and historical data underscores square footage as a pivotal factor in real estate valuation. Over the years, the relationship between property size and sale price has become more linear (given the outliers), suggesting a market that values square footage in a more standardized manner. The presence of outliers in the historical data calls for further examination to understand the unique attributes of properties that deviate from typical market patterns. This analysis is important for stakeholders in the housing market to make educated decisions, and it also sets the stage for future investigations that might incorporate additional variables such as location, property condition, and economic indicators to provide a holistic view of real estate valuation dynamics.

4. Conclusion

In our analysis of the housing market through the lens of property size and sale price, we delved into three main research questions using two distinct datasets: the current Ames dataset and historical data from Kaggle spanning 2006-2010. Our findings are as follows:

Sale Price Distribution:

The comparison of sale price distribution between the two time periods demonstrated a clear shift towards higher prices in the current market, with a notable increase in the range and median of sale prices.

Characteristics of Houses in Different Price Brackets:

Our exploration into the features of houses in the top and bottom price quartiles revealed that larger houses with more bedrooms and bathrooms tend to occupy the top quartile, whereas smaller, more economical houses fall into the bottom quartile. This characteristic held true across both datasets.

Property Size vs. Sale Price Relationship:

Square footage emerged as the predominant factor influencing sale price in both datasets, with a stronger linear correlation in the current market. The historical data suggested a more complex, non-linear relationship, with outliers indicating that the largest properties did not always conform to the general pricing trends.

Our analysis, while thorough, is not without limitations. The most significant constraint was the disparity in dataset sizes—the current Ames dataset was substantially smaller than the historical Kaggle dataset, which may have introduced some bias or overrepresentation in our comparisons. Additionally, the presence of outliers, particularly in the Kaggle dataset, suggests that exceptional properties do not follow the general trends and may skew the overall analysis. These anomalies call for a deeper investigation to understand their unique market positions.

Further work in this area could involve a deeper examination of additional factors that influence housing prices, such as location, property condition, and economic indicators. Moreover, expanding the current timeframe dataset to match the comprehensiveness of the historical dataset could provide a more balanced view of market dynamics. The incorporation of geographic data could also enhance our understanding of regional market variations. This project sets the stage for a more in-depth and diversified exploration of real estate valuation and its driving forces.

DATA DICTIONARY

Kaggle

Column	Type	Description
Id	Numeric	Unique identifier for each property
MSSubClass	Numeric	Type of dwelling involved in the sale
MSZoning	Text	General zoning classification
LotFrontage	Numeric	Linear feet of street connected to property
LotArea	Numeric	Lot size in square feet
Street	Text	Type of road access to the property
Alley	Text	Type of alley access to property
LotShape	Text	General shape of the property
LandContour	Text	Flatness of the property
Utilities	Text	Type of utilities available
LotConfig	Text	Lot configuration
LandSlope	Text	Slope of property
Neighborhood	Text	Physical locations within Ames city limits
Condition1	Text	Proximity to main road or railroad
Condition2	Text	Proximity to main road or railroad (if a second is present)
BldgType	Text	Type of dwelling
HouseStyle	Text	Style of dwelling

OverallQual	Numeric	Overall material and finish quality
OverallCond	Numeric	Overall condition rating
YearBuilt	Numeric	Original construction date
YearRemodAdd	Date	Remodel date
RoofStyle	Text	Type of roof
RoofMatl	Text	Roof material
Exterior1st	Text	Exterior covering on house
Exterior2nd	Text	Exterior covering on house (if more than one material)
MasVnrType	Text	Masonry veneer type
MasVnrArea	Numeric	Masonry veneer area in square feet
ExterQual	Text	Exterior material quality
ExterCond	Text	Present condition of the material on the exterior
Foundation	Text	Type of foundation
BsmtQual	Text	Height of the basement
BsmtCond	Text	General condition of the basement
BsmtExposure	Text	Walkout or garden level basement walls
BsmtFinType1	Text	Quality of basement finished area
BsmtFinSF1	Numeric	Type 1 finished square feet
BsmtFinType2	Text`	Quality of second finished area (if present)
BsmtFinSF2	Numeric	Type 2 finished square feet
BsmtUnfSF	Numeric	Unfinished square feet of basement area
TotalBsmtSF	Numeric	Total square feet of basement area
Heating	Text	Type of heating
HeatingQC	Text	Heating quality and condition
CentralAir	Text	Central air conditioning
Electrical	Text	Electrical system
1stFlrSF	Numeric	First Floor square feet
2ndFlrSF	Numeric	Second floor square feet
LowQualFinSF	Numeric	Low-quality finished square feet (all floors)
GrLivArea	Numeric	Above grade (ground) living area square feet
BsmtFullBath	Numeric	Basement full bathrooms
BsmtHalfBath	Numeric	Basement half bathrooms
FullBath	Numeric	Full bathrooms above grade
HalfBath	Numeric	Half baths above grade
BedroomAbvGr	Numeric	Number of bedrooms above basement level
KitchenAbvGr	Numeric	Number of kitchens

Current Ames

Column Name	Data Type	Description	Example Values
Sale_Price	Integer	The final sale price of the home.	1339000, 1195000
Bedrooms	Integer	Number of bedrooms in the home.	6, 7
Bathroom	Float	Number of bathrooms in the home.	5, 4
Sq_Ft	Integer	Total square footage of the house.	4237, 2368
Price_Per_SqFt	Float	Sale price of the home per square foot.	316, 505