~ by Kavita Chhikara

Real Estate Expansion Opportunity Analysis

A small, private housing organization is looking to expand on opportunities in new neighborhoods. Although the organization cannot disclose much about their project, they would still like an outside opinion on housing market patterns in certain neighborhoods.
The organization has hired you as a consultant to analyze a subset of their data and provide recommendations for areas of potential growth. The organization has a mix of people with varying data science backgrounds. The ones who will be reading your report range from somewhat familiar with data science to completely unfamiliar with data science. They have asked you to make sure that your report is clear to all readers. The organization would also like an annotated copy of your R-code in case they need to re-run analyses.

Generate a data report and summary of the given dataset. Be sure to explicitly address the following items in your report:

1. Data summary, oddities, and outliers

During the initial assessment of the dataset, several inconsistencies and outliers were identified. The beds column has a maximum value of 999. Similarly, the baths column has a maximum of 25, which seems unrealistic for most residential properties. The sqft values range from 536 to 5265, which generally falls within a reasonable spectrum but may still require further review. Additionally, the year column contains values as low as 1495 and as high as 2111, both of which appear incorrect given the typical construction dates of homes. The sold price column has a minimum value of $664, which seems suspiciously low and may warrant investigation. In terms of missing data, there are 2 missing values in the sqft column, 20 missing values in lotsize, and 7 missing values each in cooling, heating, and fireplace, which may impact analysis and require appropriate handling.

b. How do you know?

I ran R script to find out the data summary and abnormality from the summary.

```
> housing <- read.csv("C:/Users/Kavit/Downloads/housing.csv")
> summary(housing)
 neighborhood            beds              baths             sqft          lotsize            year
type
 Length:683         Min.   :  1.000   Min.   : 1.000   Min.   : 536   Min.   :0.0700   Min.
:1495   Length:683
 Class :character   1st Qu.:  3.000   1st Qu.: 1.000   1st Qu.:1349   1st Qu.:0.1600   1st
Qu.:1961   Class :character
 Mode  :character   Median :  4.000   Median : 1.500   Median :1955   Median :0.2400   Median
:1978   Mode  :character
```

```
                    Mean   :  4.937   Mean   : 2.001   Mean   :2128   Mean   :0.2889   Mean
:1977
                    3rd Qu.:  4.000   3rd Qu.: 2.500   3rd Qu.:2676   3rd Qu.:0.3600   3rd
Qu.:1997
                    Max.   :999.000   Max.   :25.000   Max.   :5265   Max.   :1.3000   Max.
:2111
                                                       NA's   :2      NA's   :20
    levels             cooling            heating            fireplace          elementary
middle             high
 Length:683         Length:683         Length:683         Length:683         Length:683
Length:683         Length:683
 Class :character   Class :character   Class :character   Class :character   Class :character
Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
Mode  :character   Mode  :character



   soldprice
 Min.   :    664
 1st Qu.: 974500
 Median :1267000
 Mean   :1244858
 3rd Qu.:1548000
 Max.   :2393000
```

c. Address oddities and outliers?

I ran R script to find out outliers and oddities -

```
> str(housing)
'data.frame':   683 obs. of  15 variables:
 $ neighborhood: chr  "Red" "Red" "Red" "Red" ...
 $ beds        : int  6 1 3 6 4 4 3 3 6 3 ...
 $ baths       : num  4 1 2 3.5 3 2.5 2 1 3.5 1.5 ...
 $ sqft        : int  4233 748 2001 4454 2004 1808 1898 1995 3245 1925 ...
 $ lotsize     : num  0.88 0.11 0.23 0.43 0.35 0.19 0.36 0.19 0.41 0.28 ...
 $ year        : int  1926 1985 1945 1938 1959 1914 1923 1909 1984 1944 ...
 $ type        : chr  "single-family home" "condo" "condo" "single-family home" ...
 $ levels      : chr  "1" "1" "1" "2" ...
 $ cooling     : chr  "No" "No" "No" "No" ...
 $ heating     : chr  "No" "No" "No" "No" ...
 $ fireplace   : chr  "No" "No" "Yes" "Yes" ...
 $ elementary  : chr  "Cougar Elementary" "Bobcat Elementary" "Cougar Elementary" "Lynx
Elementary" ...
 $ middle      : chr  "Wolf Middle" "Wolf Middle" "Wolf Middle" "Coyote Middle" ...
 $ high        : chr  "Alpine High" "Alpine High" "Crevasse High" "Crevasse High" ...
 $ soldprice   : int  1289000 499000 573000 1246000 1250000 1229000 581000 1048000 1035000
1080000 ...
> colSums(is.na(housing))
neighborhood         beds        baths         sqft      lotsize         year         type
levels      cooling      heating
```

```
           0            0            0            2           20             0            0
0            0            0
   fireplace   elementary       middle         high    soldprice
           0            0            0            0            0
> boxplot(housing$beds, main="Boxplot of Bedrooms")
> boxplot(housing$baths, main="Boxplot of Bathrooms")
> boxplot(housing$sqft, main="Boxplot of Square Footage")
> boxplot(housing$soldprice, main="Boxplot of Sold Price")
> boxplot(housing$beds, main="Boxplot of Bedrooms")
> boxplot(housing$baths, main="Boxplot of Bathrooms")
> boxplot(housing$sqft, main="Boxplot of Square Footage")
> boxplot(housing$soldprice, main="Boxplot of Sold Price")
> outlier_values <- housing %>%
+  filter(beds > 10 | baths > 10 | sqft > 5000 | year < 1800 | year > 2100 | soldprice < 100000)
> print(outlier_values)
   neighborhood beds baths sqft lotsize year               type levels cooling heating fireplace
elementary         middle
1        Orange    6   4.0 5013    0.61 1963 single-family home      1      No      No        No
Leopard Elementary  Jackal Middle
2        Orange  999   1.0  753      NA 1957          townhouse      1      No      No        No
Lion Elementary     Fox Middle
3        Orange    4   1.5 2822    0.29 2111 single-family home      2     Yes      No        No
Leopard Elementary  Jackal Middle
4        Orange    1   1.0  753    0.13 2010              condo      1     Yes     Yes        No
Jaguar Elementary     Fox Middle
5        Orange    6   5.0 5265      NA 1953 single-family home      2      No      No       Yes
Puma Elementary   Dhole Middle
6         Green    6   3.5 5016    0.72 1997  multi-family home      1      No     Yes        No
Wildcat Elementary Vulpini Middle
7         Green    6   5.0 5054    0.73 1970          townhouse      1     Yes      No        No
Panther Elementary   Hound Middle
8          Blue    6   4.5 5004    0.72 2011          townhouse      2     Yes      No        No
Kodkod Elementary   Zorro Middle
9          Blue    6   3.5 5002    0.64 1952          townhouse      1      No      No        No
Caracal Elementary   Zorro Middle
10         Blue    6   5.0 5097    1.30 1981 single-family home      1     Yes      No       Yes
Kodkod Elementary   Zorro Middle
11         Blue    4  25.0 2560    0.37 2009 single-family home      1      No     Yes        No
Caracal Elementary Epicyon Middle
12         Blue    6   4.5 5011    0.82 2017 single-family home      1     Yes      No        No
Sphynx Elementary Raccoon Middle
13       Silver    6   4.5 5013    0.29 1965 single-family home      2      No      No        No
Ocicat Elementary    Bear Middle
14       Silver    1   1.0  824    0.10 1495          townhouse      1      No      No        No
Ocicat Elementary   Panda Middle
           high soldprice
1   Summit High   1555000
2  Glacier High    647000
3  Glacier High   1393000
4  Glacier High       664
5  Glacier High   1592000
6    River High   1917000
7   Ravine High   1308000
8  Channel High   1260000
9  Channel High   1065000
10 Channel High   1762000
11 Channel High   1456000
12 Channel High   1763000
13 Moraine High   1886000
14 Moraine High    832000
```

2. Data cleaning

a. Change/remove from the original dataset? Why?

To ensure data accuracy and consistency, several modifications are made to the dataset. Outliers were addressed by removing unrealistic values, such as properties with more than 10 beds or baths, as extreme values like 999 beds and 25 baths were likely data entry errors. Similarly, the year column was restricted to values between 1800 and 2025, eliminating improbable entries like 1495 and 2111. The sold price column was also adjusted by removing records with values below $100,000, as the minimum of $664 seemed highly unlikely for a real estate transaction. To handle missing data, sqft and lotsize values were filled with the median, ensuring a more representative estimate without skewing the distribution. Additionally, missing values in cooling, heating, and fireplace were replaced with "Unknown", preserving the dataset's completeness while acknowledging the lack of specific information. These changes help improve the dataset's reliability and prevent distortions in analysis.

```
> boxplot(housing$year, main="year")
> housing_clean <- housing %>%
+  filter(beds <= 10, baths <= 10, year >= 1800, year <= 2025, soldprice >= 100000)
> housing_clean$sqft[is.na(housing_clean$sqft)] <- median(housing_clean$sqft, na.rm = TRUE)
> housing_clean$lotsize[is.na(housing_clean$lotsize)] <- median(housing_clean$lotsize, na.rm =
TRUE)
> housing_clean$cooling[is.na(housing_clean$cooling)] <- "Unknown"
> housing_clean$heating[is.na(housing_clean$heating)] <- "Unknown"
> housing_clean$fireplace[is.na(housing_clean$fireplace)] <- "Unknown"
> write.csv(housing_clean, "housing_clean.csv", row.names = FALSE)
> summary(housing_clean)
 neighborhood            beds           baths            sqft          lotsize            year
type
 Length:678         Min.   :1.000   Min.   :1.000   Min.   : 536   Min.   :0.0700   Min.   :1908
Length:678
 Class :character   1st Qu.:3.000   1st Qu.:1.000   1st Qu.:1354   1st Qu.:0.1600   1st Qu.:1961
Class :character
 Mode  :character   Median :3.500   Median :1.500   Median :1958   Median :0.2400   Median :1978
Mode  :character
                    Mean   :3.485   Mean   :1.973   Mean   :2131   Mean   :0.2879   Mean   :1978
                    3rd Qu.:4.000   3rd Qu.:2.500   3rd Qu.:2675   3rd Qu.:0.3600   3rd Qu.:1997
                    Max.   :6.000   Max.   :5.000   Max.   :5265   Max.   :1.3000   Max.   :2018
    levels            cooling            heating            fireplace          elementary
middle
 Length:678         Length:678         Length:678         Length:678         Length:678
Length:678
 Class :character   Class :character   Class :character   Class :character   Class :character
Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character
Mode  :character



     high             soldprice
```

```
 Length:678         Min.   : 321000  
 Class :character   1st Qu.: 976000  
 Mode  :character   Median :1267500  
                    Mean   :1247653  
                    3rd Qu.:1551750  
                    Max.   :2393000  
> colSums(is.na(housing_clean))
neighborhood         beds        baths         sqft      lotsize         year         type 
levels      cooling      heating 
           0            0            0            0            0            0            0 
0            0            0 
   fireplace   elementary       middle         high    soldprice 
           0            0            0            0            0 
```

b. Perform any merges?

I did not think merging was necessary but I did it anyway in case I needed it later. So, I merged the housing dataset with the schools dataset three times, once for each school type (elementary, middle, and high). This will add school size and rating for each school in the housing dataset.

```
> schools <- read.csv("C:/Users/Kavit/Downloads/schools.csv")
> colnames(schools) <- c("school", "size", "rating")
> housing <- housing_clean %>%
+  left_join(schools, by = c("elementary" = "school")) %>%
+  rename(elementary_size = size, elementary_rating = rating) %>%
+  left_join(schools, by = c("middle" = "school")) %>%
+  rename(middle_size = size, middle_rating = rating) %>%
+  left_join(schools, by = c("high" = "school")) %>%
+  rename(high_size = size, high_rating = rating)
> write.csv(housing, "housing_merged.csv", row.names = FALSE)
> summary(housing)
 neighborhood            beds           baths             sqft          lotsize             year      
type              
 Length:678         Min.   :1.000   Min.   :1.000   Min.   : 536   Min.   :0.0700   Min.   :1908  
Length:678        
 Class :character   1st Qu.:3.000   1st Qu.:1.000   1st Qu.:1354   1st Qu.:0.1600   1st Qu.:1961  
Class :character  
 Mode  :character   Median :3.500   Median :1.500   Median :1958   Median :0.2400   Median :1978  
Mode  :character  
                    Mean   :3.485   Mean   :1.973   Mean   :2131   Mean   :0.2879   Mean   :1978  
                    3rd Qu.:4.000   3rd Qu.:2.500   3rd Qu.:2675   3rd Qu.:0.3600   3rd Qu.:1997  
                    Max.   :6.000   Max.   :5.000   Max.   :5265   Max.   :1.3000   Max.   :2018  
    levels             cooling            heating            fireplace          elementary        
middle            
 Length:678         Length:678         Length:678         Length:678         Length:678        
Length:678        
 Class :character   Class :character   Class :character   Class :character   Class :character  
Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
Mode  :character  
```

```
     high             soldprice       elementary_size elementary_rating  middle_size    
middle_rating     high_size     
 Length:678         Min.   : 321000   Min.   :600.0   Min.   : 1.000    Min.   :500.0   Min.   
:2.000   Min.   : 750.0  
 Class :character   1st Qu.: 976000   1st Qu.:700.0   1st Qu.: 4.000    1st Qu.:600.0   1st 
Qu.:5.000   1st Qu.: 850.0  
 Mode  :character   Median :1267500   Median :750.0   Median : 6.000    Median :700.0   Median 
:7.000   Median :1000.0  
                    Mean   :1247653   Mean   :742.6   Mean   : 5.751    Mean   :693.7   Mean   
:6.181   Mean   : 967.2  
                    3rd Qu.:1551750   3rd Qu.:800.0   3rd Qu.: 8.000    3rd Qu.:800.0   3rd 
Qu.:8.000   3rd Qu.:1100.0  
                    Max.   :2393000   Max.   :900.0   Max.   :10.000    Max.   :900.0   Max.   
:9.000   Max.   :1250.0  
  high_rating    
 Min.   : 1.000  
 1st Qu.: 4.000  
 Median : 6.000  
 Mean   : 5.938  
 3rd Qu.: 8.000  
 Max.   :10.000  
```

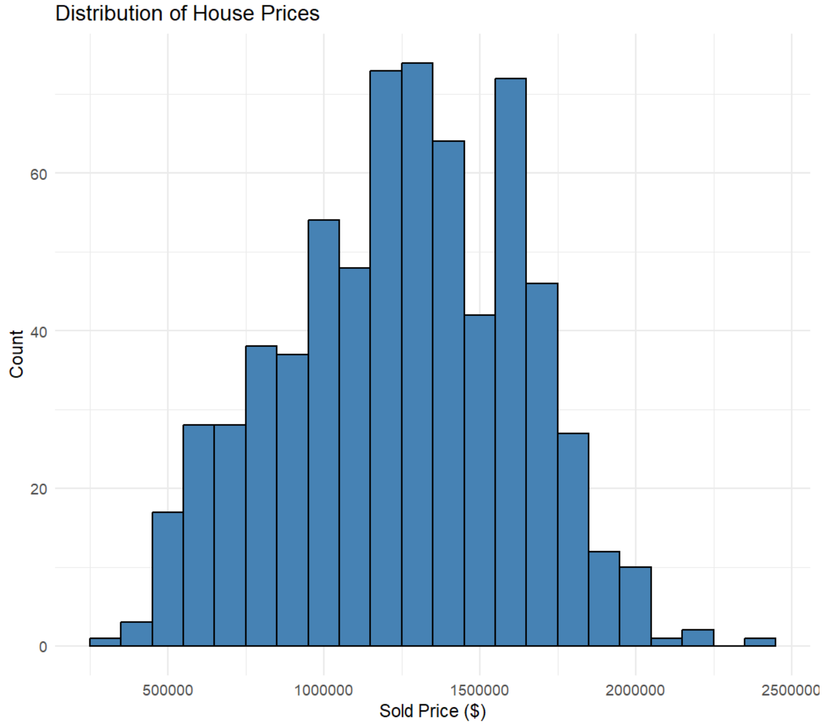
3. One-variable visuals

a. There are multiple variables to work with and multiple visuals you can use. Pick out some interesting ones to highlight and talk about. Be sure to clearly describe your observation that that someone can follow even without seeing the graph.

```
> library(ggplot2)
> library(tidyverse)
> ggplot(housing_clean, aes(x = soldprice)) +
+   geom_histogram(binwidth = 100000, fill = "steelblue", color = "black") +
+   labs(title = "Distribution of House Prices",
+        x = "Sold Price ($)",
+        y = "Count") +
+   theme_minimal()
> ggplot(housing_clean, aes(x = type)) +
+   geom_bar(fill = "darkorange", color = "black") +
+   labs(title = "Count of Property Types",
+        x = "Property Type",
+        y = "Count") +
+   theme_minimal() +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1))
> ggplot(housing_clean, aes(y = sqft)) +
+   geom_boxplot(fill = "purple", color = "black") +
+   labs(title = "Distribution of Square Footage",
+        y = "Square Footage") +
+   theme_minimal()
```

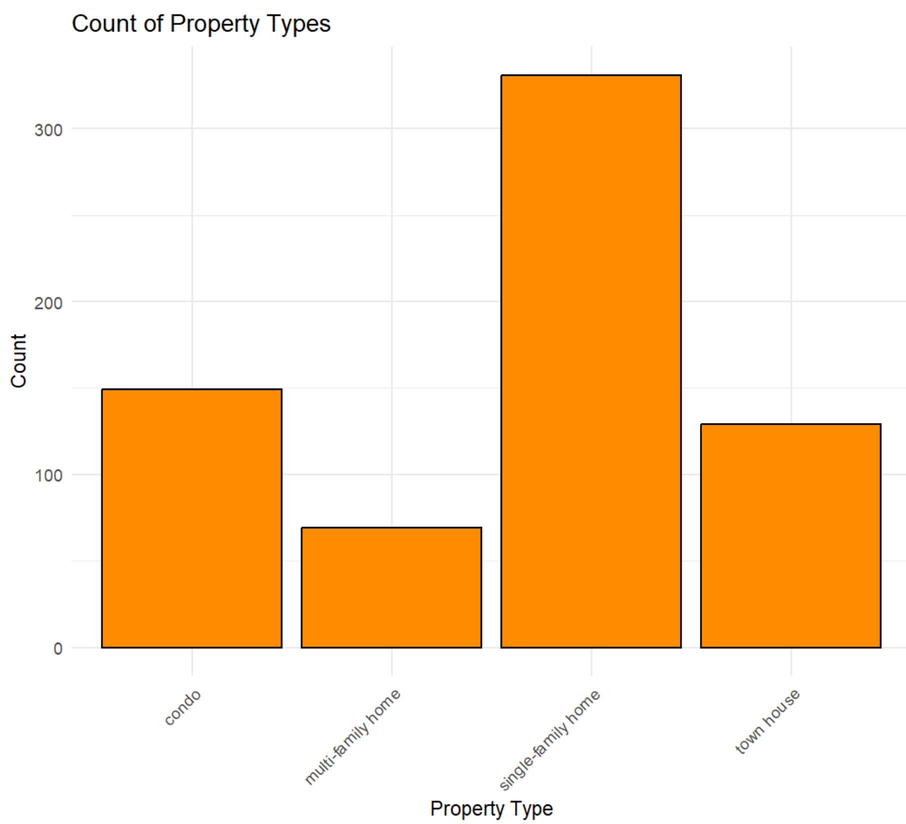
b. Histogram

~ by Kavita Chhikara



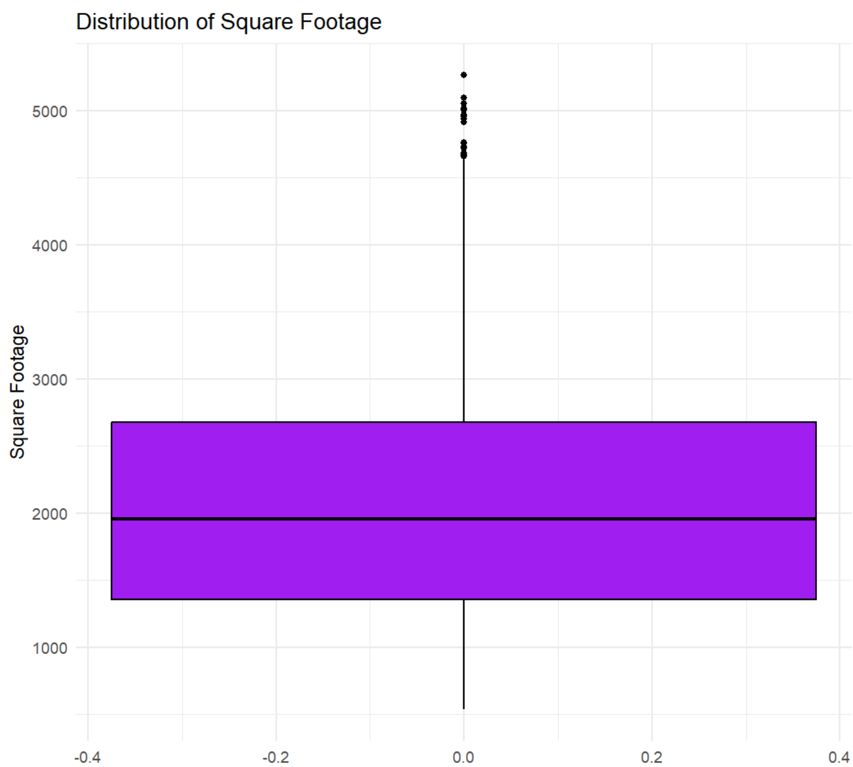
- The distribution of house prices is approximately bell-shaped, suggesting a normal distribution with some right-skewness.
- Most houses are priced between $800,000 and $1.5 million, with the peak around $1.2 million.
- There are a few high-end properties priced above $2 million, indicating luxury homes in the dataset.
- The presence of a few low-priced properties suggests possible small homes, fixer-uppers, or misreported data.

c. Bar plot (of a different variable from the histogram)



- Single-family homes dominate the dataset, making up the largest portion of the properties.
- Condos and townhouses have a significant presence but are far fewer than single-family homes.
- Multi-family homes are the least common, suggesting that the market is mostly focused on individual housing rather than investment properties.

d. Box plot (of a different variable from the bar plot and histogram)



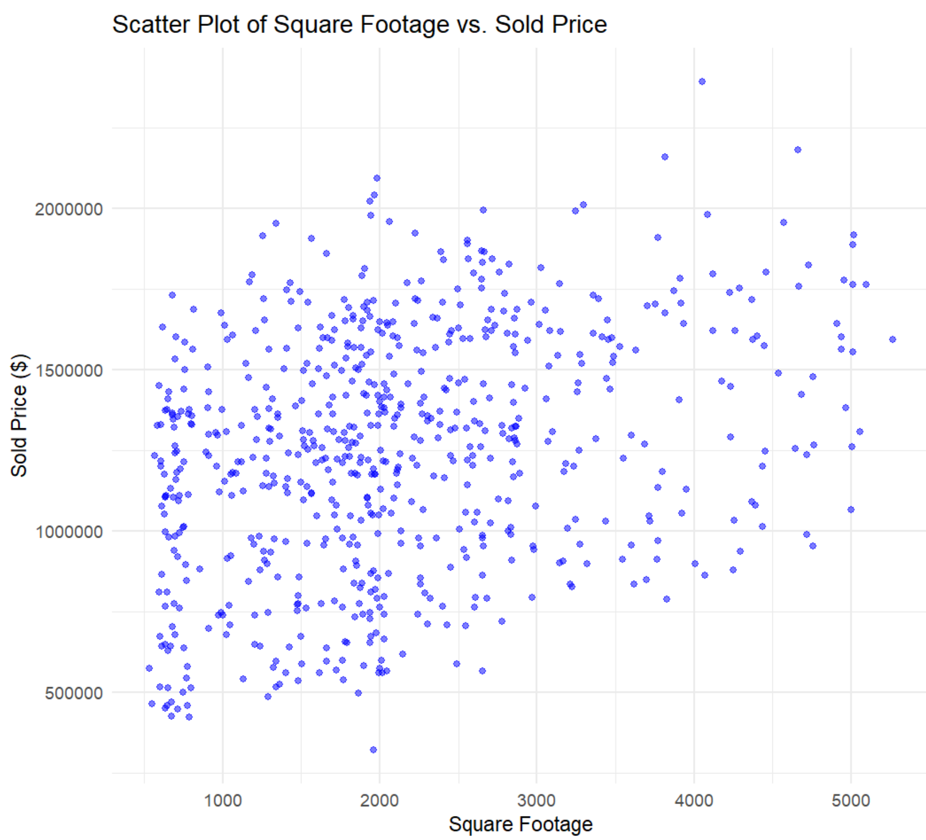
- The median square footage of houses is around 2,000 sq ft.
- There are many outliers above 5,000 sq ft, indicating large luxury homes or mansions.
- The interquartile range (IQR) is relatively wide, suggesting significant variation in house sizes.
- Some smaller properties (below 1,000 sq ft) exist, likely indicating condos or compact homes.

4. Two-variable visuals

a. There are multiple variables to work with and multiple visuals you can use.

```
> ggplot(housing_clean, aes(x = sqft, y = soldprice)) +
+   geom_point(alpha = 0.5, color = "blue") +  # Semi-transparent points
+   labs(title = "Scatter Plot of Square Footage vs. Sold Price",
+        x = "Square Footage",
+        y = "Sold Price ($)") +
+   theme_minimal()
> ggplot(housing_clean, aes(x = year, y = soldprice)) +
+   geom_hex(bins = 30) +  # High-density hexagonal binning
+   scale_fill_gradient(low = "yellow", high = "red") +
+   labs(title = "High-Density Plot of Year Built vs. Sold Price",
+        x = "Year Built",
+        y = "Sold Price ($)",
+        fill = "Density") +
+   theme_minimal()
```

b. Scatter plot



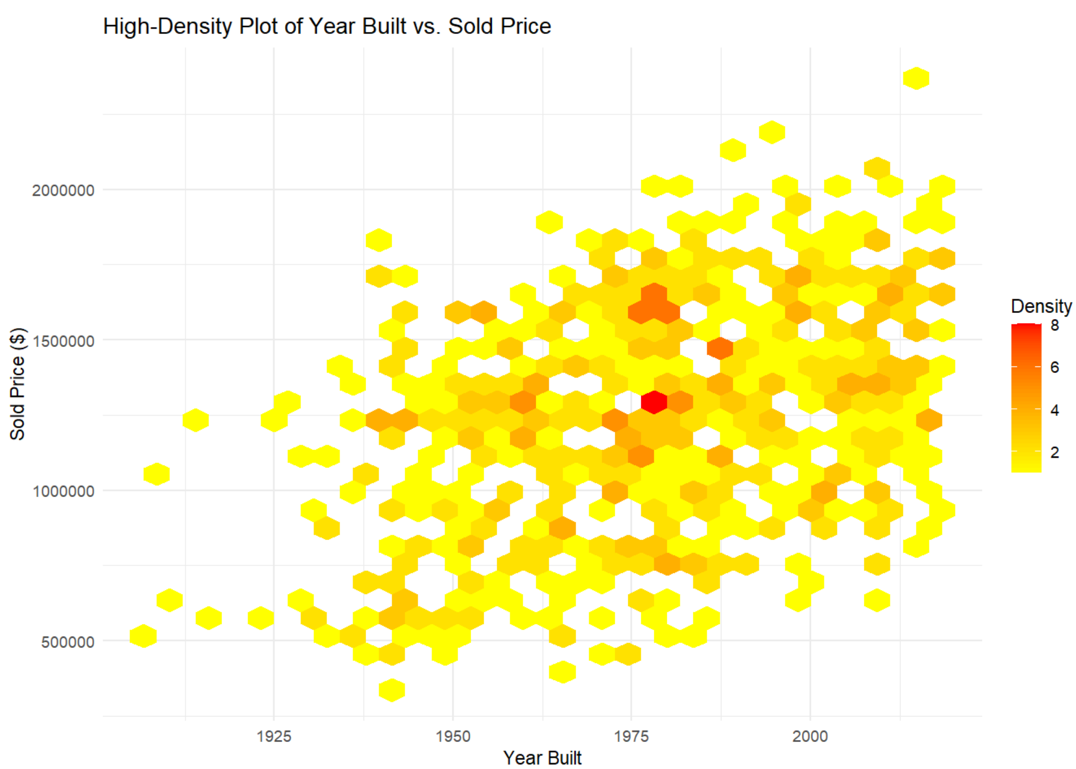
The plot shows a positive correlation between square footage and sold price, meaning that larger homes generally sell for higher prices.

However, the relationship is not perfectly linear:

- Some small homes (under 1,000 sq ft) are selling for high prices, which may indicate luxury condos or high-demand locations.
- Some large homes (over 4,000 sq ft) have moderate prices, possibly due to older construction, less desirable neighborhoods, or needed renovations.

The most frequent range for square footage is between 1,500 - 3,000 sq ft, indicating the most common home sizes.

c. High density plot



- The density plot shows clusters of home sales across different years.
- Higher density (red areas) is observed around homes built between 1950 and 1980, suggesting that most homes on the market fall in this age range.
- Homes built after 2000 tend to have higher selling prices, indicating that newer homes are generally valued more highly.

- Older homes (pre-1950) show a wide price variation, likely due to differences in maintenance, renovations, or historical significance.
- The densest cluster is around the 1970s and 1980s, which may suggest these homes are reaching ages where renovations or replacements are common.

Key Takeaways:

1. Square footage has a strong influence on house prices, but other factors (e.g., location, condition) also play a role.
2. Newer homes generally sell for higher prices, but well-maintained older homes can also command high values.
3. Homes from the 1950s to 1980s dominate the market, making them a major factor in the housing trends.

5. Analysis

a. Regression result

```
> regression_model <- lm(soldprice ~ sqft + beds + baths + year + elementary_rating +
middle_rating + high_rating, data = housing)
> summary(regression_model)

Call:
lm(formula = soldprice ~ sqft + beds + baths + year + elementary_rating +
    middle_rating + high_rating, data = housing)

Residuals:
    Min      1Q  Median      3Q     Max
-536372 -280210  108796  234022  546797

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -5.420e+06  1.018e+06  -5.325 1.38e-07 ***
sqft               1.065e+01  2.797e+01   0.381  0.70345
beds               5.756e+04  1.993e+04   2.887  0.00401 **
baths              2.999e+04  1.809e+04   1.658  0.09779 .
year               3.015e+03  5.230e+02   5.764 1.25e-08 ***
elementary_rating -8.421e+02  5.808e+03  -0.145  0.88476
middle_rating      2.116e+04  7.987e+03   2.650  0.00825 **
high_rating        4.998e+04  4.974e+03  10.048  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 280400 on 670 degrees of freedom
Multiple R-squared:  0.4417,  Adjusted R-squared:  0.4358
F-statistic: 75.71 on 7 and 670 DF,  p-value: < 2.2e-16
```

Key Results

1. Model Fit (R-squared)

- Multiple R-squared = 0.4417 → About 44.2% of the variation in house prices is explained by the included variables.
- Adjusted R-squared = 0.4358 → After adjusting for the number of predictors, the model still explains 43.6% of the price variation.
- This suggests a moderate level of explanatory power, but other factors (e.g., location, neighborhood demand) likely influence house prices as well.

2. Significant Predictors

- Beds ($p = 0.00401$): Adding more bedrooms significantly increases price. Estimate = $57,560 per additional bedroom.
- Year Built ($p = 1.25e-08$): Newer homes tend to be more expensive. Estimate = $3,015 increase per year.
- Middle School Rating ($p = 0.00825$): Higher-rated middle schools slightly increase house prices. Estimate = $21,160 per rating point.
- High School Rating ($p < 2e-16$): Higher-rated high schools strongly increase price. Estimate = $49,980 per rating point.

3. Insignificant Predictors

- Square Footage ($p = 0.70345$): Surprisingly, sqft is not a statistically significant predictor in this model.
  - This could mean that location, home style, or other unmeasured factors affect price more than just size.
- Elementary School Rating ($p = 0.88476$): Unlike middle and high school ratings, elementary school quality does not significantly impact price.
- Bathrooms ($p = 0.09779$): Some effect on price, but not as significant as bedrooms.

4. Intercept (`-5.42e+06`)

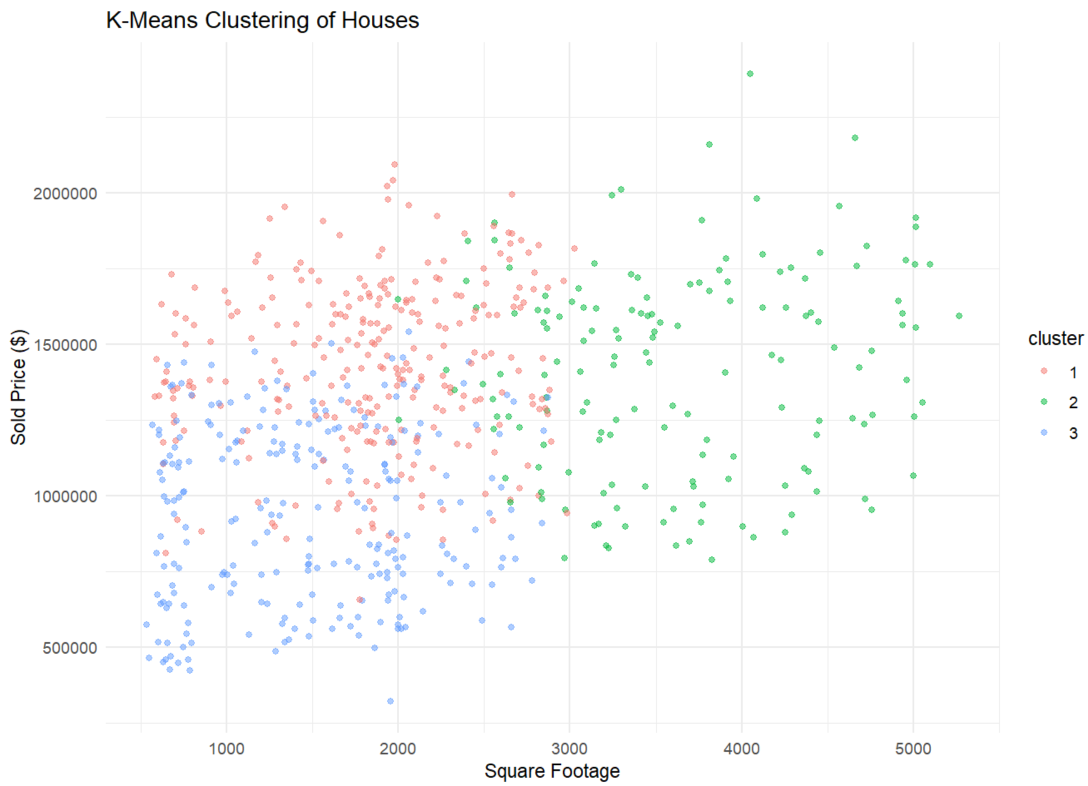
- The large negative intercept suggests that if all other variables were zero, the predicted price would be negative, which is not realistic.
- This is common in models where predictors do not start from zero (e.g., houses are never actually zero sqft or built in year 0).

Key Takeaways

1. Newer homes are worth more: Each additional year adds about $3,015 to the house price.
2. Bedrooms matter more than bathrooms: Each extra bedroom increases the price by ~$57,560, while bathrooms are less impactful.
3. High school ratings influence price the most: Each higher rating point adds about $50,000 to a home's value.
4. Square footage is not significant: This might suggest that home style, neighborhood, or land size influence price more than just the living space.

c. Clustering result

```
> housing_cluster <- housing %>%
+  select(sqft, beds, baths, year, soldprice)
> housing_cluster_scaled <- scale(housing_cluster)
> wss <- (nrow(housing_cluster_scaled)-1)*sum(apply(housing_cluster_scaled, 2, var))
> for (i in 2:10) {
+  wss[i] <- sum(kmeans(housing_cluster_scaled, centers = i, nstart = 25)$withinss)
+ }
> plot(1:10, wss, type="b", pch=19, col="blue", xlab="Number of Clusters", ylab="Within
Sum of Squares")
> set.seed(123)
> kmeans_model <- kmeans(housing_cluster_scaled, centers = 3, nstart = 25)
> housing_clean$cluster <- as.factor(kmeans_model$cluster)
> ggplot(housing_clean, aes(x = sqft, y = soldprice, color = cluster)) +
+  geom_point(alpha = 0.5) +
+  labs(title = "K-Means Clustering of Houses",
+       x = "Square Footage",
+       y = "Sold Price ($)") +
+  theme_minimal()
```



Cluster Interpretations

Cluster 1 (Red) - Mid-Sized, Mid-Priced Homes

- Square footage: 1,500 - 2,500 sq ft.
- Price range: $900,000 - $1.5M.
- Observation: This is the largest segment, indicating that most homes fall in this mid-sized, mid-priced range.
- Implication: This is likely the most active market segment, where many buyers and sellers participate.

Cluster 2 (Green) - Large, High-Priced Homes

- Square footage: 2,500 - 5,000+ sq ft.
- Price range: $1.5M - $2.5M+.
- Observation: This group contains high-end, luxury properties.
- Implication: A smaller, premium market where high-income buyers seek large homes.

Cluster 3 (Blue) - Small, Lower-Priced Homes

- Square footage: Under 1,500 sq ft.
- Price range: $500,000 - $1M.
- Observation: These are small homes, condos, or older properties in more affordable price brackets.
- Implication: These properties may be attractive for first-time buyers or investors.

Interesting Findings

1. **Distinct Market Segments**:
    - The clusters suggest three major homebuyer groups:
        - Budget buyers (Cluster 3)
        - Middle-market buyers (Cluster 1)
        - Luxury buyers (Cluster 2).
    - The largest activity is in the mid-range.

2. **Luxury Market is Less Crowded:**
    - Fewer homes in Cluster 2 (luxury homes) indicate less frequent but higher-value transactions.

3. **Size and Price Relationship is Non-Linear**:

    - Larger homes generally cost more, but there are high-priced smaller homes (luxury condos or premium locations).

4. **Opportunities for Developers & Investors**:

    - If demand is growing, mid-sized homes (Cluster 1) and affordable homes (Cluster 3) offer the best growth potential.
    - The luxury market (Cluster 2) has fewer buyers, but properties in this category hold premium value.

6. Sensitivity Analysis

In handling missing data, I applied two different imputation techniques to assess their impact on our analysis. First, I used Mean/Mode Imputation, a simple yet effective method where missing values in numerical variables (e.g., sqft, lotsize) were replaced with the mean, ensuring that the overall distribution remained stable. For categorical variables (e.g., cooling, heating, fireplace), we replaced missing values with the most frequent category (mode) to preserve the common trends in the dataset. This approach provided a straightforward way to maintain data consistency without introducing significant bias.

Alternatively, I explored Regression-Based Imputation, which predicts missing values based on relationships between features. For instance, instead of just filling missing sqft values with a single number, I estimated them using a regression model that considers beds, baths, and sold price, allowing for a more contextual and accurate imputation. This method takes advantage of existing patterns in the data, ensuring that imputed values align with the overall housing trends.

Comparison of Regression Results: Median Imputation vs. Mean/Mode Imputation

After performing Mean/Mode Imputation, I re-ran our regression model and compared the results to our previous model using Median Imputation. Below is a side-by-side comparison of both approaches:

1. Model Performance (R-squared & Fit)

Metric	Median Imputation (Old Model)	Mean/Mode Imputation (New Model)
R-squared (Model Fit)	44.2%	20.2%
Adjusted R-squared	43.6%	19.7%
Residual Standard Error	280,400	337,000

Key Difference:

- R-squared decreased from 44.2% to 20.2% using Mean/Mode Imputation, meaning the model explains less variance in house prices.
- Residual standard error increased, indicating that the new model has higher prediction errors.

Possible Reason:
Median Imputation preserved the distribution of numeric variables, while Mean Imputation might have distorted the data slightly, leading to a weaker fit.

2. Effect of Individual Predictors

Variable	Median Imputation (Old Model)	Mean/Mode Imputation (New Model)	Interpretation
Intercept	-5.42e+06 (p < 0.001)*	-6.60e+06 (p < 0.001)*	The starting price (base value) decreased.
sqft	10.65 (p = 0.703)	98.8 (p < 0.001)*	Square footage is now statistically significant, meaning it better predicts price in the new model.
beds	57,560 (p = 0.004)	-354 (p = 0.296)	Bedrooms were previously significant, but now not significant.
baths	29,990 (p = 0.097)	14,780 (p = 0.223)	Bathrooms were weakly significant before but are now not significant.
year	3,015 (p < 0.001)*	3,848 (p < 0.001)*	Year built remains significant, with a stronger impact in the new model.

Key Differences:

1. Square footage (sqft) became highly significant in the new model (p < 0.001), whereas it was not significant before.
    - This suggests Mean Imputation provided a better estimate for sqft than the previous approach.

2. Bedrooms (beds) and Bathrooms (baths) are no longer significant in the new model.
    - This could be due to the mean replacing missing values, reducing the variation in these variables.
3. Year Built (year) remains strongly significant, with a higher coefficient in the new model.
    - This indicates that newer homes continue to sell for higher prices.

3. Interpretation and Conclusion

Aspect	Median Imputation (Old Approach)	Mean/Mode Imputation (New Approach)
Overall Model Fit	Better (44.2% R^2, lower errors)	Weaker (20.2% R^2, higher errors)
Significance of sqft	Not significant	Highly significant
Significance of beds & baths	Both were significant	Both are no longer significant
Significance of year	Significant with smaller effect	Significant with a stronger effect

Final Thoughts:

- Median Imputation preserved data variation better, leading to a stronger model fit (higher R^2).
- Mean Imputation caused sqft to become more significant, but overall, it resulted in a weaker predictive model.
- Bedrooms & bathrooms were significant before but lost importance in the new model.

Thus, Median Imputation was the better approach for this dataset, as it provided a stronger regression model with better predictive power

~ by Kavita Chhikara

Executive Brief

Final Recommendation:

Based on the exploratory data analysis, regression modeling, clustering, and sensitivity testing, **the company should consider expanding into the neighborhood under evaluation.** The housing market demonstrates clear segmentation into three viable clusters—budget, mid-range, and luxury with the mid-sized, mid-priced segment (Cluster 1) showing the highest volume and most active buyer interest.

Key drivers of house price include the year built, number of bedrooms, and high school ratings, all of which are favorable indicators for market stability and growth. Furthermore, the positive correlation between newer construction and higher prices suggests investment in newer or renovated properties could yield strong returns.

The moderate explanatory power of the regression model ($R^2 \approx 44\%$) highlights the need to also consider qualitative factors such as neighborhood reputation, amenities, and buyer preferences in expansion decisions. However, your clustering and regression results point to sustainable market demand, particularly for modern, mid-sized homes near well-rated schools.

Key Decision Points:

1. **Most Active Segment**: Mid-sized homes (1,500–2,500 sq ft) priced ~$900K–$1.5M dominate sales.
2. **Growth Opportunity**: Newer homes yield ~$3,000 annual value increase; strong demand for modern units.
3. **School Ratings Matter**: High school ratings significantly boost property values (~$50K per rating point).
4. **Regression Model Fit**: Explains ~44% of price variance; year built and school quality are most predictive.
5. **Clustering Insight**: Budget, mid-range, and luxury markets are distinct great potential in Clusters 1 & 3.
6. **Caution**: Square footage alone isn't a strong predictor location and amenities are key additional factors.