HOUSING & SCHOOLS ANALYSIS

CONTEXT: 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.

OBJECTIVE: Provide recommendations for areas of potential growth.

ABOUT THE DATASETS:

To begin this report, here's the data contained in each dataset.

housing.csv >> dataset of houses sold in 2019

neighborhood	name of the neighborhood, indicator of a general area in the city
beds	number of bedrooms in the unit
baths	number of bathrooms in the unit
sqft	unit square footage
lotsize	unit’s lot size
year	year that the unit was built
type	unit type
levels	how many floors are in the unit
cooling	whether or not the unit has cooling
heating	whether or not the unit has central heating
fireplace	whether or not the unit has a fireplace
elementary	unit’s assigned elementary school
middle	unit’s assigned middle school
high	unit’s assigned high school
soldprice	selling price of the home

schools.csv >> dataset of school rating

school	name of the high school
size	approximate student population size
rating	school rating on a 1 to 10 scale

1. Data summary, oddities, and outliers

a. What are all of the oddities and outliers in the dataset?

Looking at the unprocessed data, we notice a few oddities and outliers, particularly in the *housing.csv* dataset. This is the case:

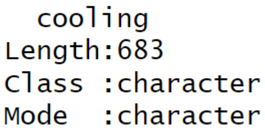
- Some columns are not categorized, which means that the different categories contained in these columns (the different blue, orange... neighborhoods, for example) are not recognized. It is the case for the following columns: *neighborhood*, *type*, *levels*, *cooling*, *heating*, *fireplace*, *elementary*, *middle* and *high*.
- Missing data (NA's) are indicated.
- Most numerical data show outliers, i.e. data that are extreme or far from other values.

b. How do you know?

Information on oddities and outliers is mainly provided in the summary and boxplots.

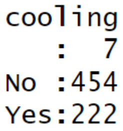
- Regarding the absence of categorization, this can be seen directly in the summary, as it does not display usable information for character data (neighborhood, type, etc.).

Before categorization ➔ *After categorization*



```
  cooling
Length:683
Class :character
Mode  :character
```

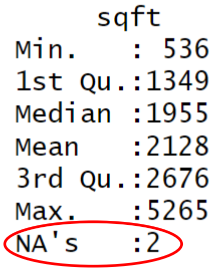
➔



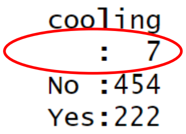
```
cooling
   :  7
No :454
Yes:222
```

- Missing data are noted NA's for numerical data directly in the summary too. In addition, once categorized, we deduce the absence of character data from unnamed values.

Numerical data Character data

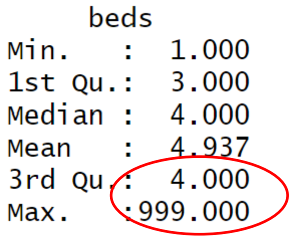


```
     sqft
Min.   : 536
1st Qu.:1349
Median :1955
Mean   :2128
3rd Qu.:2676
Max.   :5265
NA's   :2
```



```
cooling
   :  7
No :454
Yes:222
```

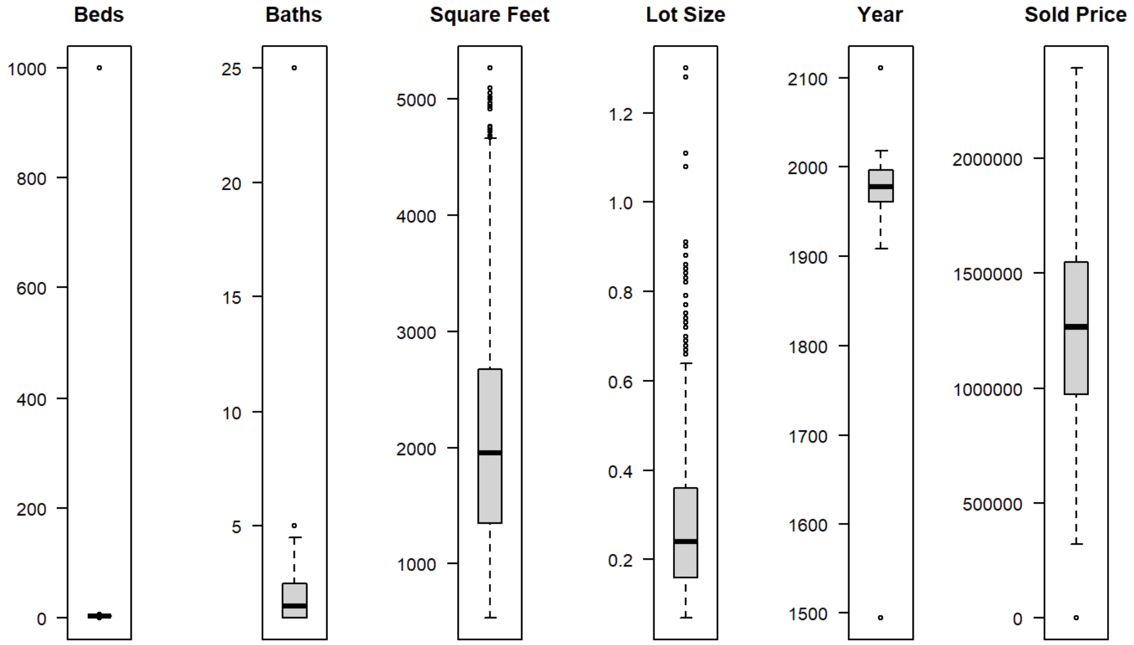
- Finally, outliers can be identified by locating large differences between the minimum value and the first quartile (Q1), or between the third quartile (Q3) and the maximum value. Using a boxplot, these outliers are more visible as they are indicated by points outside the box.



```
      beds
 Min.   :  1.000
 1st Qu.:  3.000
 Median :  4.000
 Mean   :  4.937
 3rd Qu.:  4.000
 Max.   :999.000
```

We notice a huge gap between the 3rd quartile and the maximum value in the beds column.
This is considered as an extreme outlier.

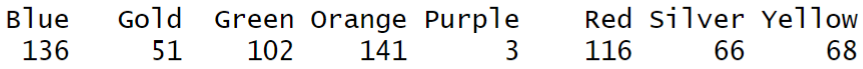
Let’s take a look at the following set of boxplots:



- We can deduce that the max value of the *beds* category must be removed or readjusted, as it is the only outlier in this category, as must the *baths* category and its max value of 25.
- The *sqft* category contains a few outliers at the top of the boxplot. The values remain fairly close, so their presence doesn't seem to be a cause for concern.
- The *lotsize* category has many outliers in the upper part of the boxplot, so 4 are quite far apart. A readjustment may be in order.
- The *year* category has 2 extreme values: the minimum value of 1495 identified above, and the maximum value for 2111. We could delete or readjust these values.
- Finally, *soldprice* contains only one outlier, the value 664.

Also, once that the columns *neighborhood*, *type*, *levels*, *cooling*, *heating*, *fireplace*, *elementary*, *middle* and *high* are categorized, we can analyze them:

- *neighborhood* is categorized by color and there are no NA's to report. However, we note that the purple category is under-represented: it contains 3 values, compared with 141 for orange, for example.



```
  Blue   Gold  Green Orange Purple    Red Silver Yellow
   136     51    102    141      3    116     66     68
```

- *levels* appears well-balanced, but has 6 NA's.

- *cooling*, *heating* and *fireplace* have 7, 7 and 6 NA's respectively.
- For the *elementary*, *middle* and *high* categories, there are no missing values. What's more, the values are well distributed overall.

c. How do you plan to address oddities and outliers?

There are several possible solutions to oddities and outliers.

- First, I categorized the columns concerned as seen above.
- Secondly, the *housing.csv* dataset contains 683 rows. Deleting some of the data should not affect future results. So deleting oddities and outliers might be an option.
- Finally, some oddities just seem to be typing errors when entering data. These values can therefore be readjusted.
- For missing values, you can create new values by predicting the most consistent value, or by using the mean or median value.

I have chosen to readjust outliers and use prediction to replace missing values.

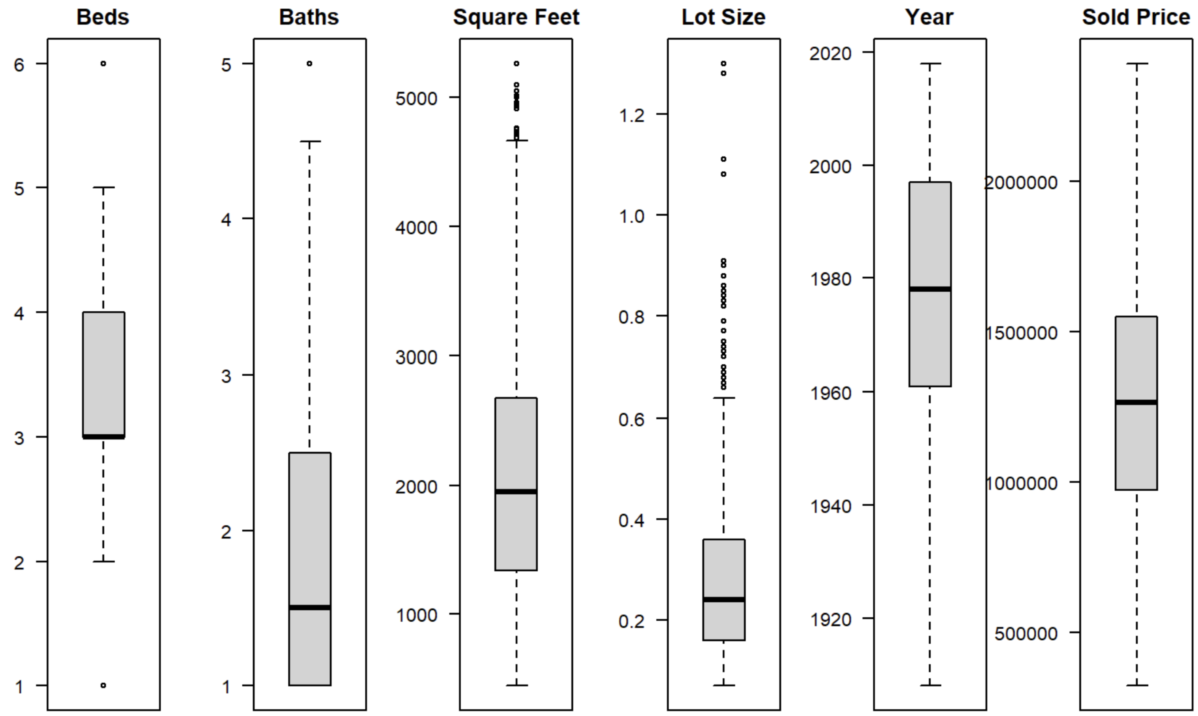
2. Data Cleaning

a. What did you change/remove from the original dataset? Why?

I chose not to remove any data from the datasets. Instead, I readjusted oddities and outliers, and changed some columns to make the analysis easier:

- Categorizing columns to get more information at first sight.
- Turning impossible outliers into more appropriate value thanks to logic and comparison with similar values:
  - 999 beds become 1 bed after comparison with similar houses.
  - 25 baths turn into 2.5 baths after comparison with similar houses and considering this outlier as a typing error.
  - Year 1495 become 1945 in comparison with similar houses and considering it as an error made while typing the number.
  - Year 2111 become 2011 as it is considered as a mistake made while typing the date.
  - Sold price of 664 turn into 664000. Indeed, in comparison with similar houses in the same neighborhood, soldprice should be between 423000 and 512000 or above, especially that this house is newer than the other 2 and it has cooling and heating. Therefore, price is probably missing zeros.
- Replacing values Yes and No by numerical value 1 and 0 to facilitate further analysis.
- Using prediction to change NA's into usable values: for each columns containing NA's, I created a prediction model.

Once all the modifications have been made, we obtain the following set of boxplots:



There are still a few outliers visible, but on further investigation, these values are not really isolated. If they appear as outliers, it's because their number is small in relation to the total number of data (683 lines).

b. Did you perform any merges?

The schools.csv dataset gives the size and rating of each of the schools listed in the housing.csv dataset. Therefore, I performed a merge on each school name and category: elementary, middle and high. This action adds a total of 6 columns to the housing dataset (2 per category of school for the size and the rating), providing all the information in a single dataset.

Here is the very last summary of the final dataset:

Summary Statistics	**Beds**	**Baths**	**Square Feet**	**Lotsize**	**Year**	**Levels**	**Soldprice ($)**
Minimum	1	1	445	0.07	1908	1	321 000
Median	3	1.5	1948	0.24	1978	1	1 267 000
Maximum	6	5	5265	1.30	2018	2	2 393 000

Summary Statistics	**Elementary School Size**	**Middle School Size**	**High School Size**	**Elementary School Rating**	**Middle School Rating**	**High School Rating**
Minimum	600	500	750	1	2	1
Median	750	700	1000	6	7	6
Maximum	900	900	1250	10	9	10

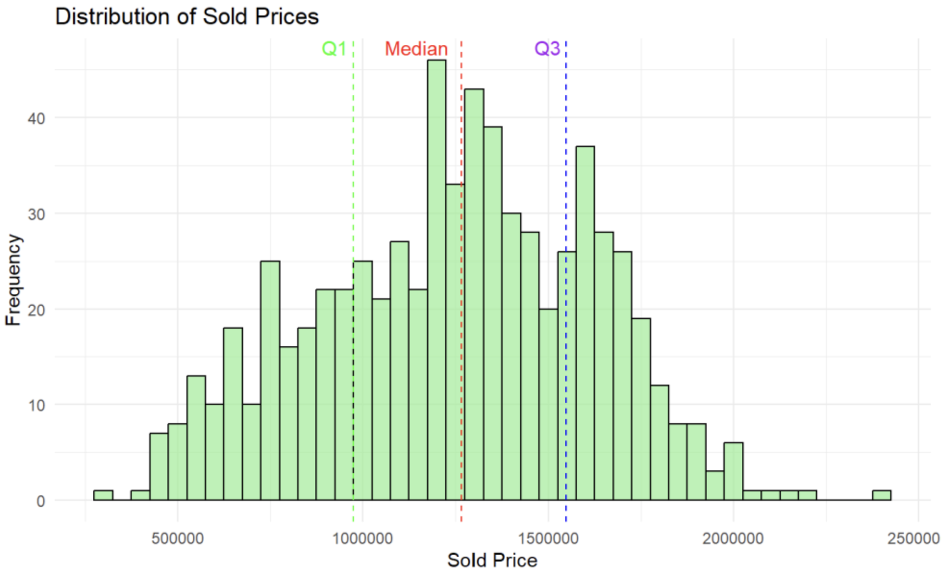
3. One-variable visuals

(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 so that someone can follow even without seeing the graph.)

a. Include at least one histogram

Let's start this series of graphs with 2 histograms giving us an indication of the distribution of sales prices and the number of square feet of the houses in the dataset.

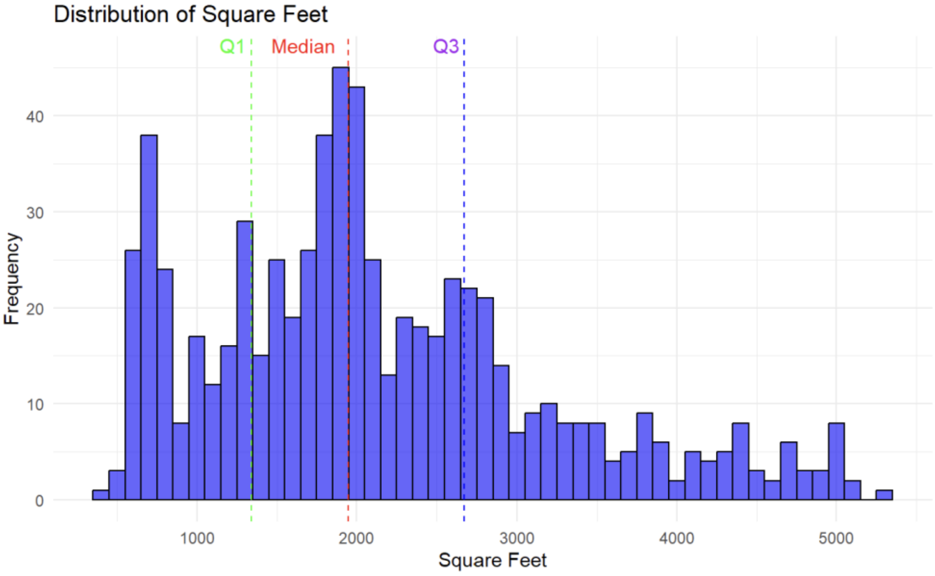
The first histogram shows the distribution of house sale prices. I've added the 1st quartile, median and 3rd quartile for greater precision.



The distribution is normal and well spread out overall, with a peak in values at the median. There's also a slight peak at the 3rd quartile. Finally, a few values are off-center below 500,000 and above 200,000, corresponding to the min and max values announced in the summary.

This suggests that most homes are sold between 100,000 and 175,000.

Now, let's take a look at which square feet are the most sold with the 2nd histogram showing the distribution of square feet. Like the previous graph, I've added the 1st quartile, median and 3rd quartile for greater precision.



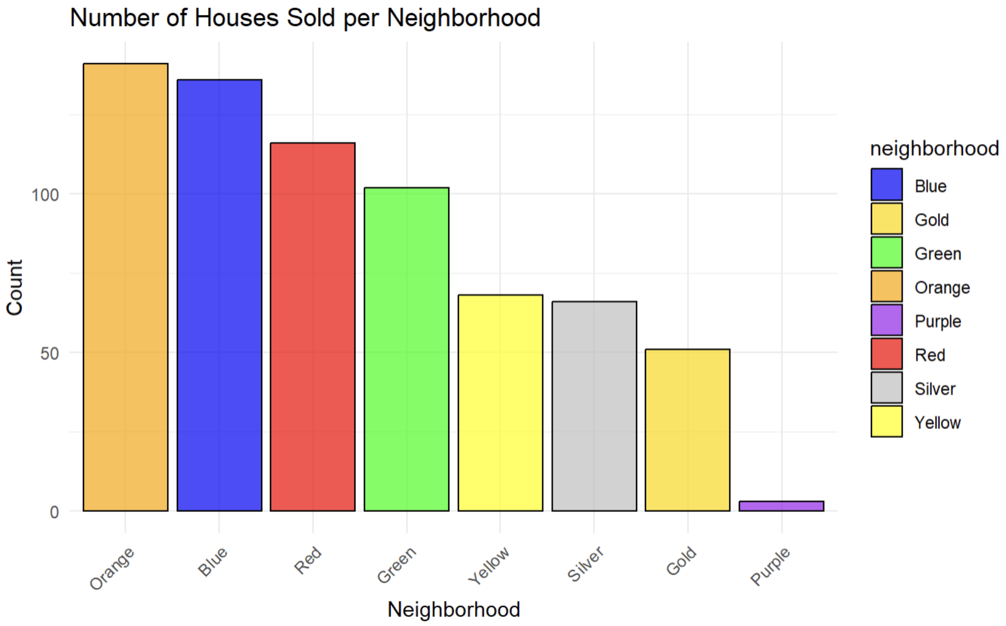
The distribution is a bit more balanced. Indeed, there's a peak between Q1 and Q3, but another peak is particularly visible in the lower part of the histogram around 750 square feet: not surprising, this corresponds to the 'condo' and 'condiminium' categories of houses, i.e. apartments. It's therefore normal to have a smaller set of surface areas than the rest.

The majority of homes sold are between 1,500 and 2,750 square feet for houses, and around 750 square feet for apartments.

b. Include at least one bar plot (of a different variable from the histogram)

Let's continue our analysis with 2 barplots. The first shows the number of homes sold per neighborhood.

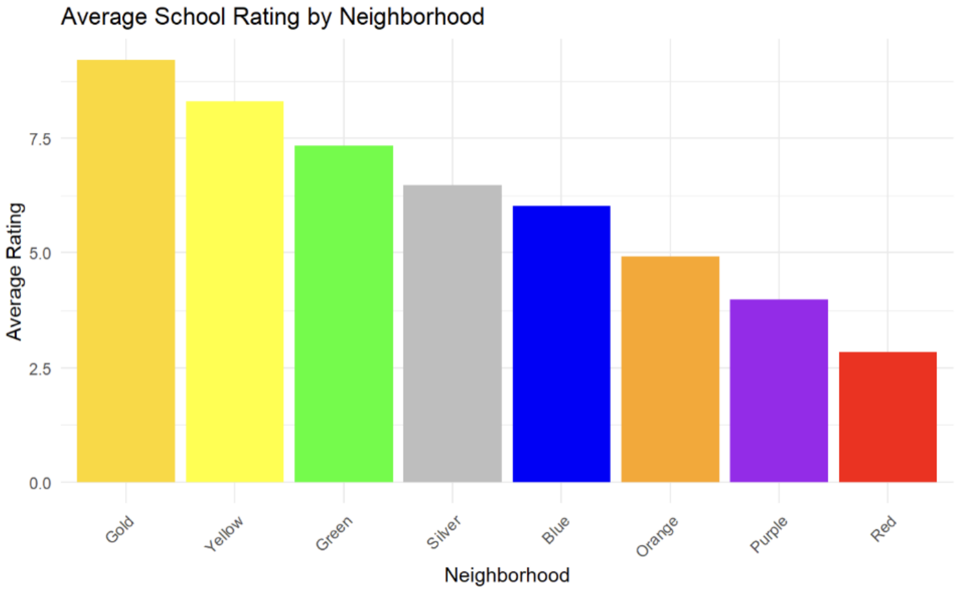
Each neighborhood is categorized by color. I've used the associated colors for ease of reading.



The Orange and Blue neighborhoods are fairly close in terms of counts, but Orange remains in the lead. Next come the Red, Green, Yellow, Silver and Gold neighborhoods. The Purple district is far behind, with only 3 homes sold as seen in the summary.

This means that the Orange, Blue and Red districts contain the most homes for sale, while the Purple district has virtually no homes for sale.

Now let's take a look at the second barplot on the Average School Rating by neighborhood. The color of each neighborhood is preserved for better lisibility and the data is sorted in descending order. The School Rating corresponds to the rating of the associated school on a scale of 0 to 10. Each neighborhood is assigned a set of schools: elementary, middle and high. We can therefore calculate the average rating by school group, and by neighborhood.



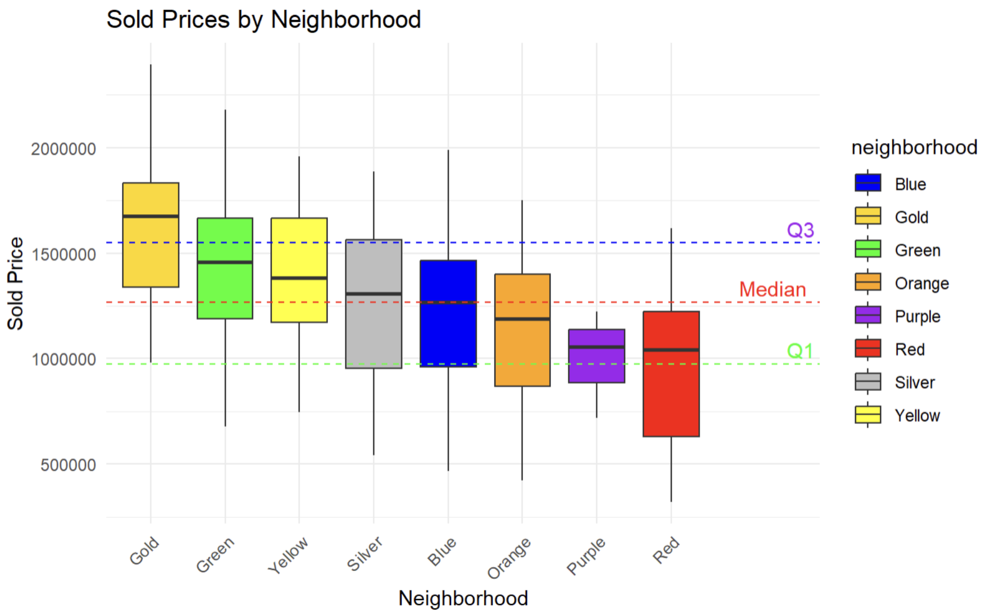
Gold leads the way with an average rating of about 8.5, followed by Yellow, Green, Silver, Blue, Orange, Purple and finally Red with an average rating of about 2.5.

We can deduce that the Gold, Yellow and Green districts can be highly prized, as they are associated with highly rated schools.

c. Include at least one box plot (of a different variable from the bar plot and histogram)

Let's move on to the boxplots.

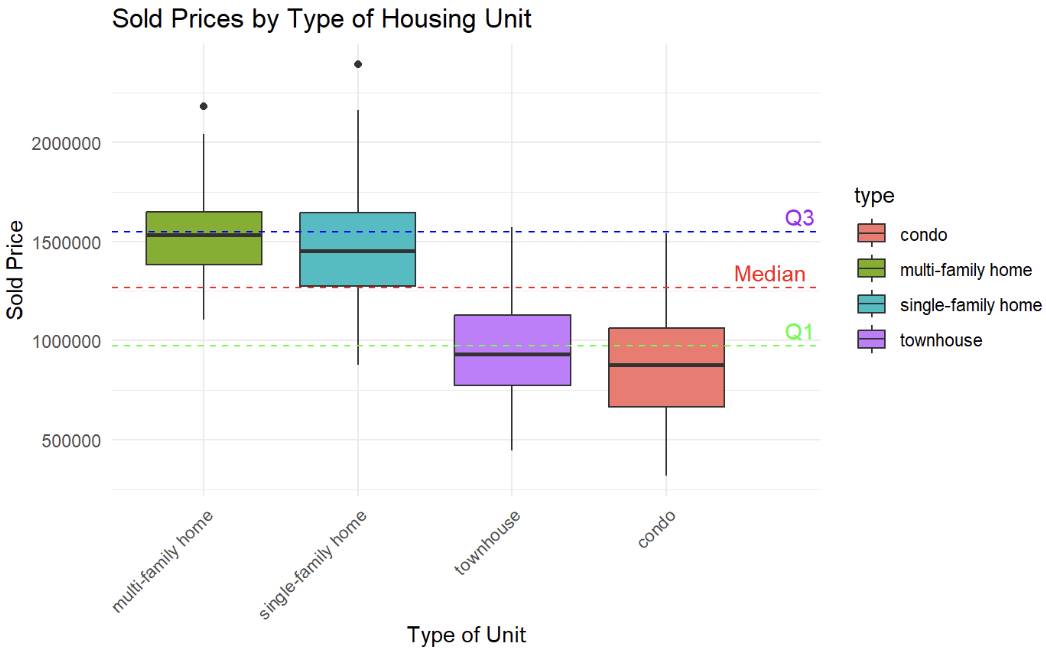
The first boxplot shows Sold Price by Neighborhood. Once again, the colors chosen correspond to the colors of the associated category. I've also added the 1st quartile, median and 3rd quartile of the sold price to better position the values. In addition, the data is sorted in descending order for easier readability.



Thus, we see that the Silver and Blue districts contain 50% or more of their values between the 1st and 3rd quartile of the sale price. The district stands out from the others by having more than 50% of its values above the 3rd quartile, or above 75% of other sold prices. The Green and Yellow districts are generally equivalent, with a similar 1st and 3rd quartiles, but a slightly lower median for the Yellow district. The Purple neighborhood is the one with the least extensive values, which can be explained by the few houses in this neighborhood (3). Finally, the Red district has almost 50% of its values below the 1st quartile of the sold price, which means below 75% of the other sold prices.

We can conclude that the Gold district is by far the most expensive and the red district is by far the cheapest. The Silver and Blue districts are generally within the norm.

The second boxplot represents the Sold Prices by Type of Housing Unit. In fact, there are several types of accommodation (condo, multi-family home, townhouse, etc.). It is interesting to know which type of housing is the most expensive. To facilitate the positioning of the values, I again added the 1st quartile, the median and the 3rd quartile. The values are also sorted in descending order.



From the outset, we notice that the two categories which dominate in terms of prices are multi-family home and single-family home, having part of their values above the 3rd quartile of the sold price. This type of house is undoubtedly the largest in order to accommodate large families. The other 2 categories have more than 50% of their values below the 1st quartile, or below 75% of the other sold prices.

It is no surprise that family homes are the most expensive and condo-type housing is the most affordable, as are townhouse-type housing.

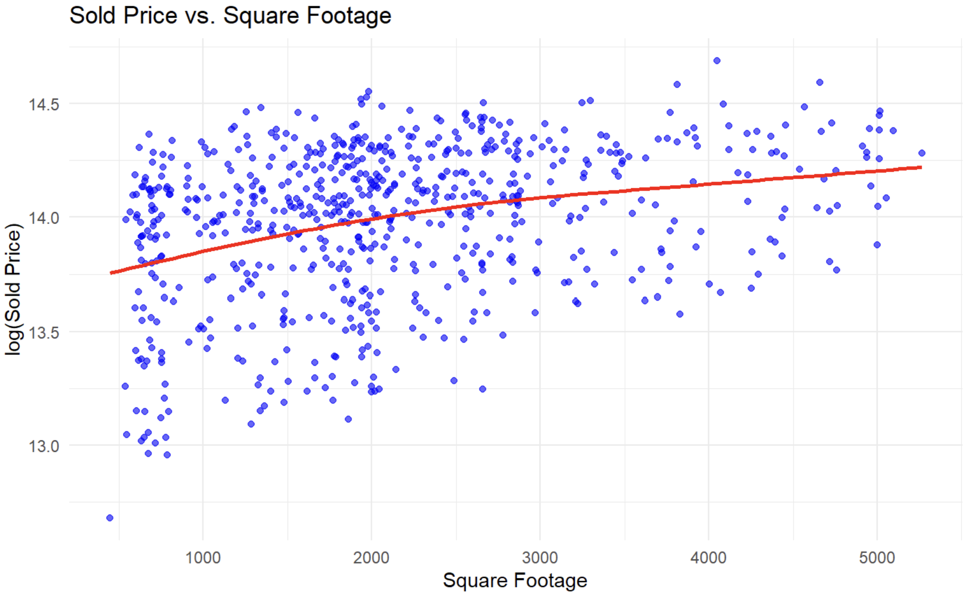
4. Two-variable visuals

(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.)

a. Include at least one scatter plot

Let’s move on to scatter plot.

The first scatter plot shows the relationship between the sold price of houses and their square footage. Each point represents a house, with the x-axis indicating the square footage and the y-axis indicating the sold price. This plot helps in understanding how house size influences its price and can reveal trends or patterns, such as whether larger houses tend to sell for higher prices.



I chose to use the logarithm to represent the sold price values. Indeed, the points are very scattered, the use of the logarithm makes it possible to bring the points together, in addition to avoiding a multitude of 0s on the y bar (due to high sold price). In addition, I drew a line to better visualize the trend of the points.

Thus, we observe a fairly significant dispersion of points with a slight trend emerging as seen with the red line. Two phases seem to be emerging:

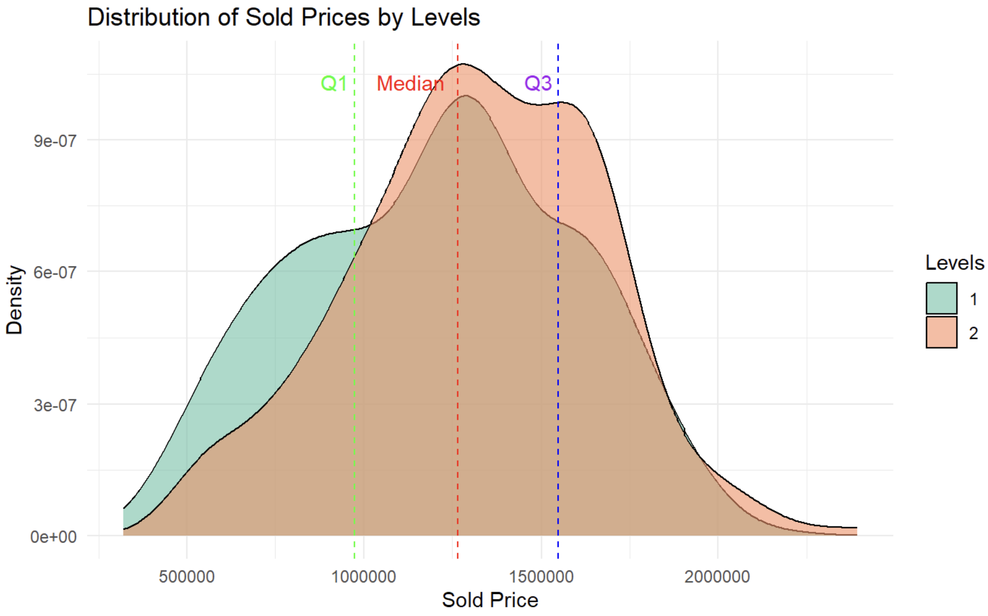
- In the first phase, the sale price increases almost linearly for an area between 500 and 2000 square feet, going from a logarithm equal to 13.75 to 14. This represents a sale price going from 936589 to 1202604, a sold price difference of 266015 for a square feet difference of 1500.
- In the second phase, the sale price increases almost linearly for an area varying from 2000 to 5000 square feet, going from a logarithm of 14 to approximately 14.2. This represents a sale price going from 1202604 to 1468864, which mean a sold price difference of 266260 for a square feet difference of 3000.

It would therefore seem that the more the area in square feet increases, the less the sold price increases, and the closer the prices become (wide dispersion at the beginning, and becomes finer with the increase in area)

b. Include at least one high density plot

We now come to the density plot.

The first plot shows the distribution of the sold price by level. To better position the values, I added the 1st quartile, the median and the 3rd quartile.



Only 2 levels appear: houses have only one or 2 floors. We see that whatever the number of levels, we observe a density peak around the median, as well as around the 3rd quartile.

We notice a greater density of 2-floors housing in the upper part of the sold price, and another peak in density just below the 1st quartile for 1-floors housing.

We deduce that 2-floors housing is generally more expensive than 1-floor housing.

5. Analysis

(The organization did not provide you with any specific topic to investigate. They are interested in what patterns you find. What are some interesting findings?)

a. Include at least one regression result

Now that we know more about the dataset and these trends, let's move on to the regression.

For this, I wanted to perform a multiple linear regression to predict the sold price of houses based on various predictors. Indeed, I wanted to know what factors really influence the sale price.

Here is my model:

- lm(soldprice ~ sqft + beds + baths + year + neighborhood + elementary_rating + middle_rating + high_rating, data = housingMerg)

The model predicts the soldprice based on sqft, beds, baths, year, neighborhood, and rating. The summary(lm_model) provides coefficients, standard errors, t-values, and p-values for each predictor, along with R-squared and adjusted R-squared values to assess the model's fit.

Here are the results:

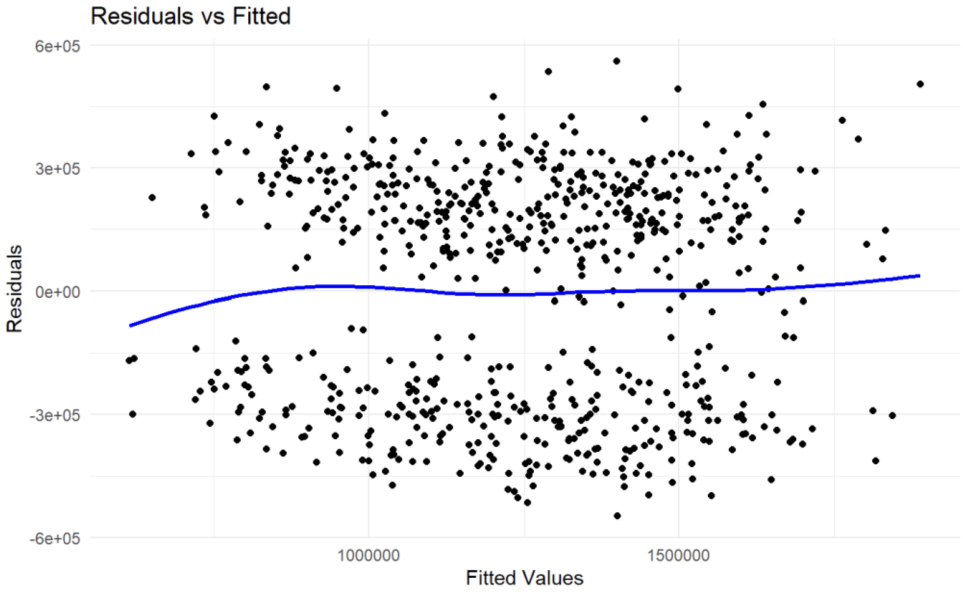
- **Statistically significant predictors** are Bedrooms, year built, middle school rating, and high school rating: they all have a p-value lower than 0.001. Bedrooms and high school rating are the most significant. Indeed:
  - An additional bedroom is associated with an increase in sold price by approximately $58530, holding other variables constant.
  - An increase in the high school rating by one point is associated with an increase in the sold price by $52,430, holding other variables constant.
- R-squared value is 0.4499 which indicates that approximately 44.99% of the variance in sold prices is explained by the model.
- P-value is lower than 2.2e-16, so the model is statistically significant overall, meaning that at least some of the predictors are significantly associated with the sold price.

Therefore, the analysis of the model indicates that while certain variables like the number of bedrooms, the year the house was built, and school ratings (middle and high school) have significant effects on the sold price, others such as square footage and the specific neighborhood do not, after accounting for other factors in the model.

This is quite a surprising result. Indeed, the surface area is generally a means of determining the price of housing and is commonly used to compare the price of housing depending on the location, and therefore the neighborhood.

Let's plot the residuals to check the fit of this model:

The residual plot helps to visually check if the residuals are randomly distributed, indicating a good model fit.



We observe 2 groups of points, one above 0, the other below 0, almost in symmetry. The blue line crosses the graph and generally runs along the x axis at y = 0, thus passing between the 2 groups of points.

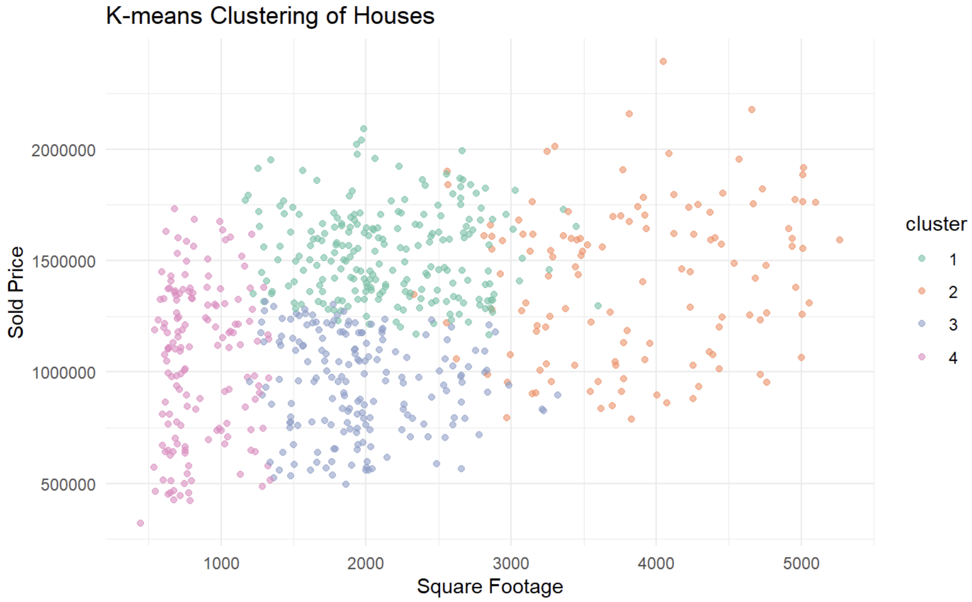
Here is what we can say about the graph:

- This symmetry suggests that the model does not systematically overpredict or underpredict the sold prices across the range of fitted values: the model's predictions are, on average, correct.
- Ideally, residuals should be randomly scattered around the horizontal axis without any distinct patterns or groups. The fact that we can see 2 groups of point may suggest that the model fits certain ranges of the data better than others.
- The fact that the line crosses the graph and runs close to the x-axis suggests that there is no major systematic error in the model: on average, the residuals are centered around zero and the model's predictions are not biased.

Therefore, while the model appears to be unbiased on average, the distinct grouping of residuals suggests there might be underlying complexities that need to be addressed to improve the model's fit.

b. Include at least one clustering result

The cluster_data contains selected features (sqft, beds, baths, soldprice) for clustering.



4 clusters are formed.

- Cluster 4 seems to take its values between 500 and 1500 square feet, and between 500000 and 1750000 sold price: this cluster likely represents smaller, possibly more affordable homes that still fall within a relatively high price range. These might be well-located or recently renovated smaller homes.
- Cluster 3 is located between 1500 and 3000 square feet, and between 500000 and 1250000 sold price: homes in this cluster are mid-sized and moderately priced. These could be average family homes in mid-range neighborhoods.
- Cluster 1 is just above 3, between 1250000 and 2000000: larger homes with a higher price range. These may represent more luxurious properties or homes in high-demand areas.
- Finally, cluster 2 is quite dispersed and is located between 3000 and 5000 square feet, and between 750000 and 200000 sold price: this cluster overlaps with Cluster 1 in terms of price but may represent a different segment due to the spread in square footage. These could be high-end properties that are diverse in their features and locations.

Therefore, this k-means clustering provides a meaningful segmentation of the housing data, revealing distinct groups of homes that differ in terms of square footage and sold price. This segmentation is valuable for market analysis, investment decisions, and strategic planning in the real estate sector.

6. Sensitivity Analysis

(The dataset has some missing data. Re-run your entire analysis filling in the missing data using a different method than your original approach (from 1 and 2). For example, if you used listwise deletion, try single imputation or multiple imputation. Do not use pairwise deletion.)

a. What method did you use to fill in the missing data this time around?

Let's remove all oddities and outliers from the dataset. With this method, we go from 683 row to 635, which means we removed 48 rows from *housing.csv*.

We merge the 2 datasets and here is the final summary:

Summary Statistics	Beds	Baths	Square Feet	Lotsize	Year	Levels	Soldprice ($)
Minimum	1	1	536	0.07	1908	1	423 000
Median	4	1.5	1961	0.24	1978	1	1 267 000
Maximum	6	5	5097	1.30	2018	2	2 393 000

Summary Statistics	Elementary School Size	Middle School Size	High School Size	Elementary School Rating	Middle School Rating	High School Rating
Minimum	600	500	750	1	2	1
Median	750	700	1000	6	7	6
Maximum	900	900	1250	10	9	10

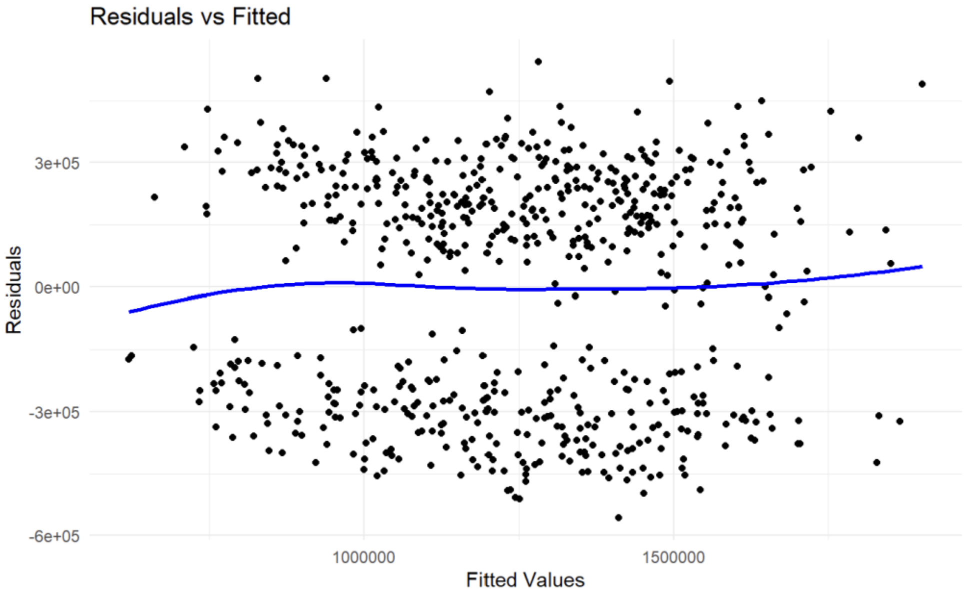
Some values of the summary statistics changes (values shown in orange). This is particularly the case for the square feet column for which all the values are different.

b. Compare and contrast your results from your original run of the analysis.

Let’s compare the model results:

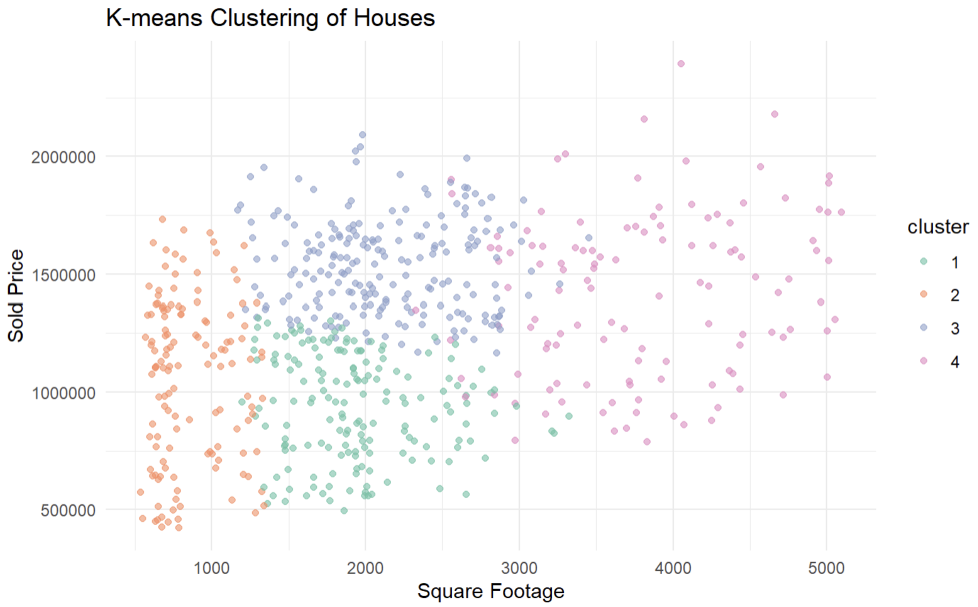
- **Statistically significant predictors** are still Bedrooms, year built, middle school rating, and high school rating: they all have a p-value lower than 0.001. But this time, the year and high school rating are the most significant, instead of bedrooms and high school rating. Here:
  - An additional year is associated with an increase in sold price by approximately $3353, holding other variables constant.
  - Value for high school rating does not change.
- R-squared value is 0.4497 and not 0.4499. The difference is not significant.
- The adjusted R-squared value is 0.4322 instead of 0.4384 which might be the only big difference between the 2 models.
- P-value is also lower than 2.2e-16, so the model is statistically significant overall, meaning that at least some of the predictors are significantly associated with the sold price.

Also by comparing the graphs, we do not see any major difference:



We get the exact same description as earlier.

Finally, K-means gives exactly the same observations:



It seems that, in this dataset, the loss of 48 rows does not affect the analysis results.

The similarity in results between deleting anomalies and readjusting values with predictions can be attributed to the inherent robustness of the dataset, the accuracy of prediction models, and the nature of the anomalies. Both methods aim to improve data quality, and when done correctly, they can lead to datasets that faithfully represent the underlying trends and patterns, resulting in similar analytical outcomes.