Class: Exploratory data analysis and visualization

Assignment: Final

Author: Kangjun Zou

12/15/2023

Introduction

The goal of this project is to assist your organization to discover housing market patterns and areas of potential growth in those neighborhoods within the provided data. I am going to work you through the data structure, data cleaning, data merges, exploratory visuals and models we train to further our understanding of how different factors/variables affect housing prices. How missing data is dealt has always been a controversial area. We adopted 2 different approaches on missing data. I am going to walk you through both approaches later in the report to compare and contrast the results.

Data description

There are 2 datasets we adopted for this analysis:

1. housing.csv – dataset of houses sold in 2019. There are 683 observations and 15 variables.

- 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

Here down below is the number of missing data existing in each variable in the data set (42 missing data in total):

neighborhood	beds	baths	sqft	lotsize	year
0	0	0	2	20	0
type	levels	cooling	heating	fireplace	elementary
0	0	7	7	6	0
middle	high	soldprice			
0	0	0			

Here down below are the unique values existing in each categorical variable:

neighborhood: Blue, Gold, Green, Orange, Purple, Red, Silver, Yellow

type: condo, multi-family home, single-family home, townhouse, town house, condominium

levels: 1, 2, ?

cooling: No, Yes

heating: No, Yes

fireplace: No, Yes

elementary (23 in total): Birman Elementary, Bobcat Elementary, Caracal Elementary, Cheetah Elementary, Cougar Elementary, Jaguar Elementary, Kodkod Elementary, Korat Elementary, Leopard Elementary, Lion Elementary, Lynx Elementary, Margay Elementary, Munchkin Elementary, Ocelot Elementary, Ocicat Elementary, Oncilla Elementary, Panther Elementary, Puma Elementary, Savannah Elementary, Serval Elementary, Sphynx Elementary, Tiger Elementary, Wildcat Elementary

middle (15 in total): Bear Middle, Coyote Middle, Culpeo Middle, Dhole Middle, Epicyon Middle, Fox Middle, Hound Middle, Husky Middle, Jackal Middle, Panda Middle, Raccoon Middle, Sloth Middle, Vulpini Middle, Wolf Middle, Zorro Middle

high (14 in total): Alpine High, Atlas High, Avalanche High Blizzard High, Channel High, Crevasse High, Delta High, Glacier High, Moraine High, Mountain High, Ravine High, River High, Summit High, Valley High

2. schools.csv – dataset of school rating. There are 52 observations and 3 variables.

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

Here down below are the unique values existing in each categorical variable:

school (52 in total): Birman Elementary, Bobcat Elementary, Caracal Elementary, Cheetah Elementary, Cougar Elementary, Jaguar Elementary, Kodkod Elementary, Korat Elementary, Leopard Elementary, Lion Elementary, Lynx Elementary, Margay Elementary, Munchkin Elementary, Ocelot Elementary, Ocicat Elementary, Oncilla Elementary, Panther Elementary, Puma Elementary, Savannah Elementary, Serval Elementary, Sphynx Elementary, Tiger Elementary, Wildcat Elementary, Bear Middle, Coyote Middle, Culpeo Middle, Dhole Middle, Epicyon Middle, Fox Middle, Hound Middle, Husky Middle, Jackal

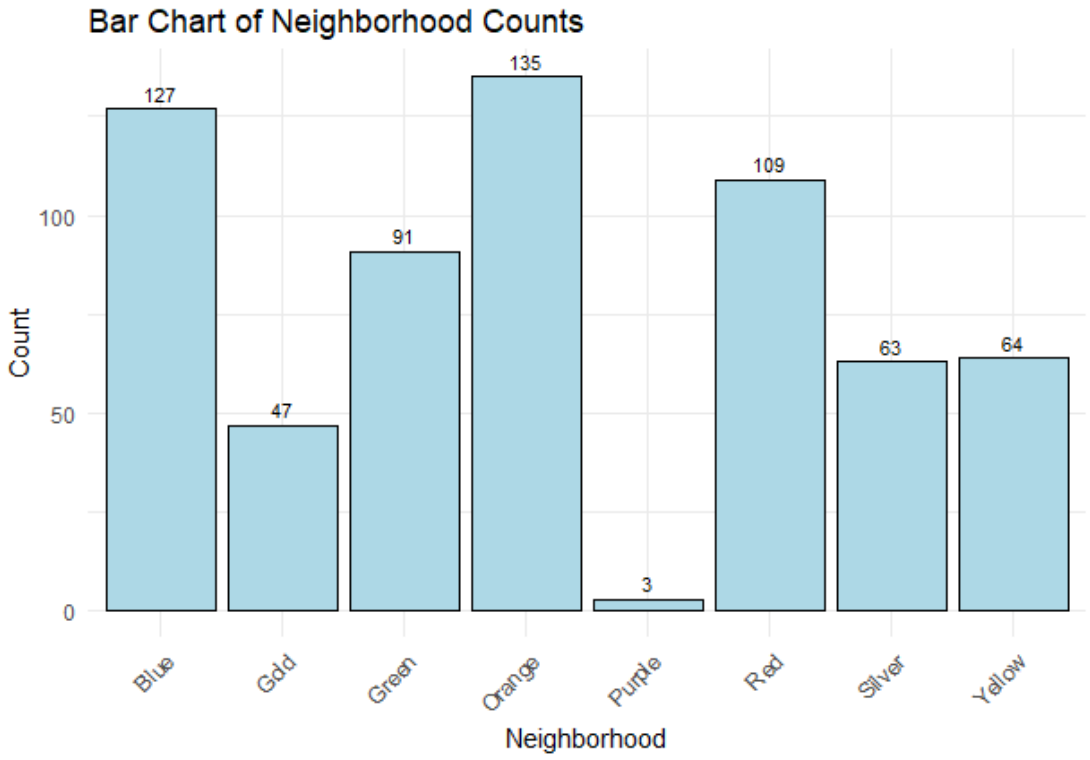
Middle, Panda Middle, Raccoon Middle, Sloth Middle, Vulpini Middle, Wolf Middle, Zorro Middle , Alpine High, Atlas High, Avalanche High Blizzard High, Channel High, Crevasse High, Delta High, Glacier High, Moraine High, Mountain High, Ravine High, River High, Summit High, Valley High

Data cleaning: Approach #1

Housing.csv

Categorical variables:

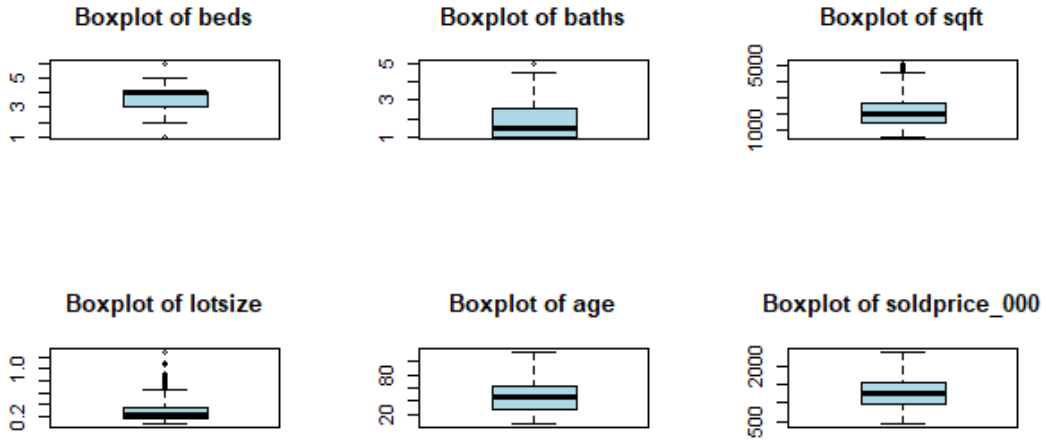
1. All rows that contain missing data (42 rows in total) were removed.
2. In variable "type", "town house" was replaced by "townhouse".
3. In variable "type", "condominium" was replaced by "condo".
4. All rows where "levels" is "?" were removed.
5. 3 houses from neighborhood purple were removed, which wouldn't be representative of the neighborhood and would have skewed the analysis if kept.



New variables:

6. A new variable, "age", was engineered from "year", representing how old the house is. "year" was removed after.
7. A new variable, "soldprice_000" was engineered from "soldprice", representing the sold price in thousands. "soldprice" was removed after.

Numeric variable



8. The row where age equal to -88 was removed.
9. The row where age equal to 528 was removed.
10. The row where sold price equal to 0.664 was removed.
11. The row where baths equal to 25 was removed.

Outliers removed above do not make sense in reality. Thus removed. Other outliers remained because they can be easily justified.

schools.csv

This data set is clean.

Merge

The cleaned housing.csv was merged with school.csv by school.

2 new variables were engineered:

mean_size: the average size of schools around the house

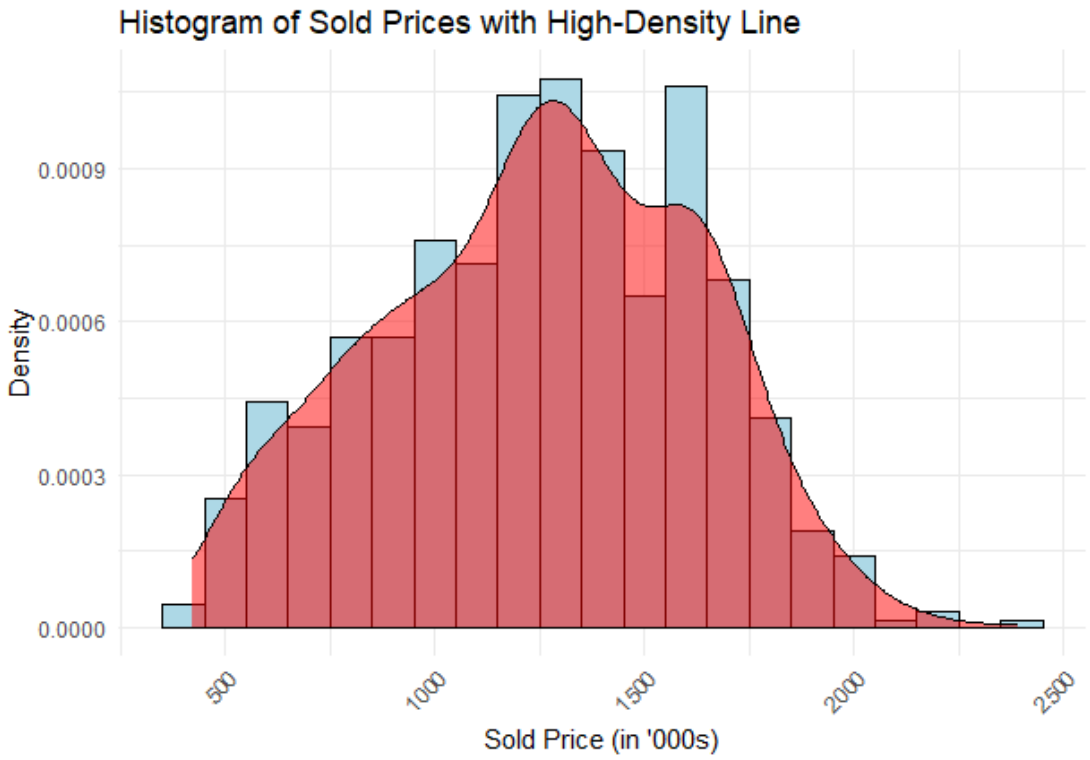
mean_rating: the average rating of schools around the house

Here down below are the selected feature kept in the merged data set:

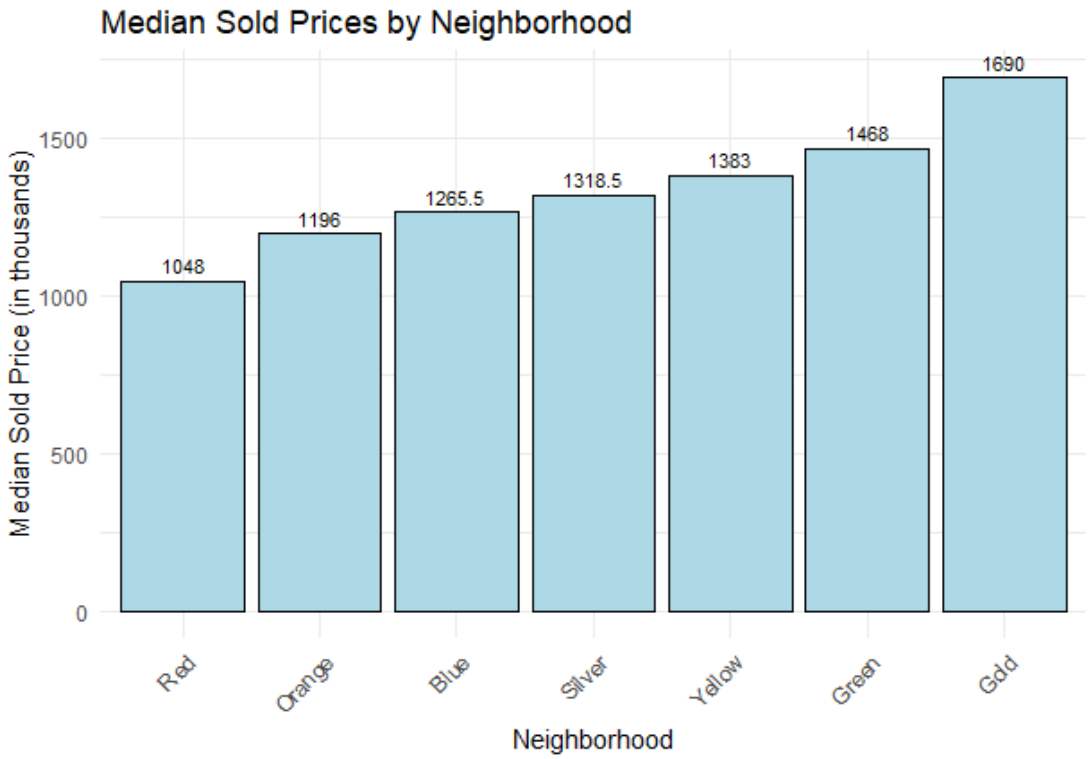
"neighborhood", "beds", "baths", "sqft", "lotsize", "type", "levels", "cooling", "heating", "fireplace", "age", "soldprice_000", "mean_size", "mean_rating"

Here, we conclude the data cleaning approach 1.

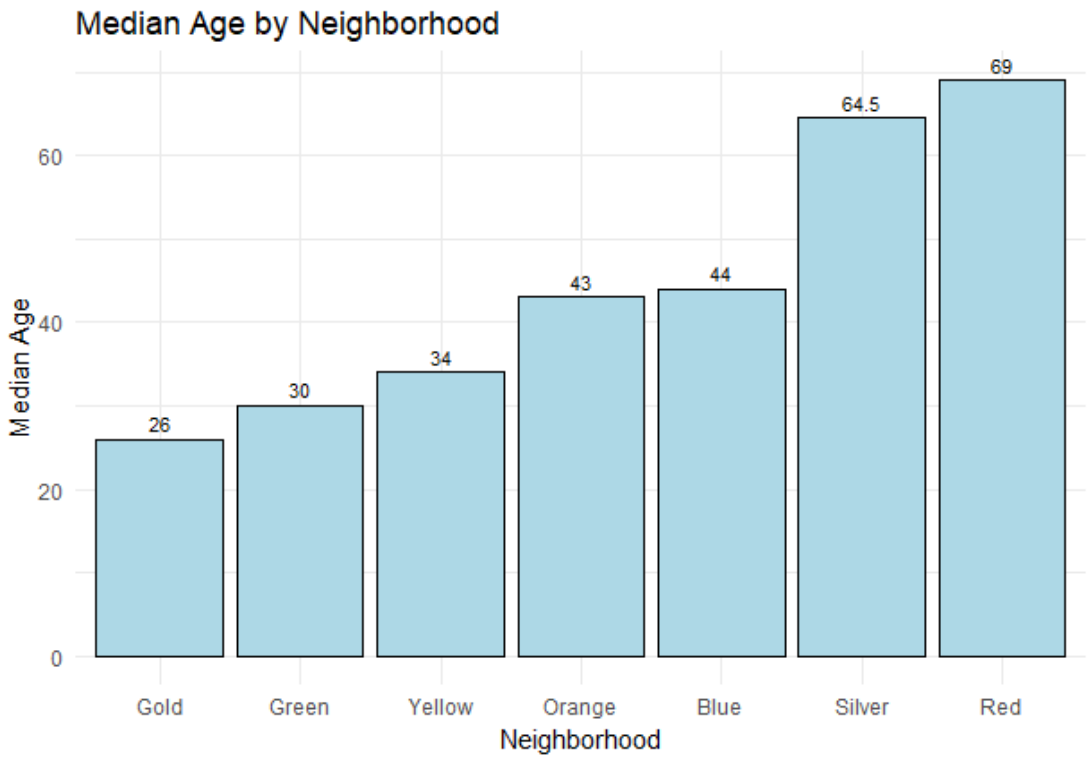
Data exploratory analysis and visualization



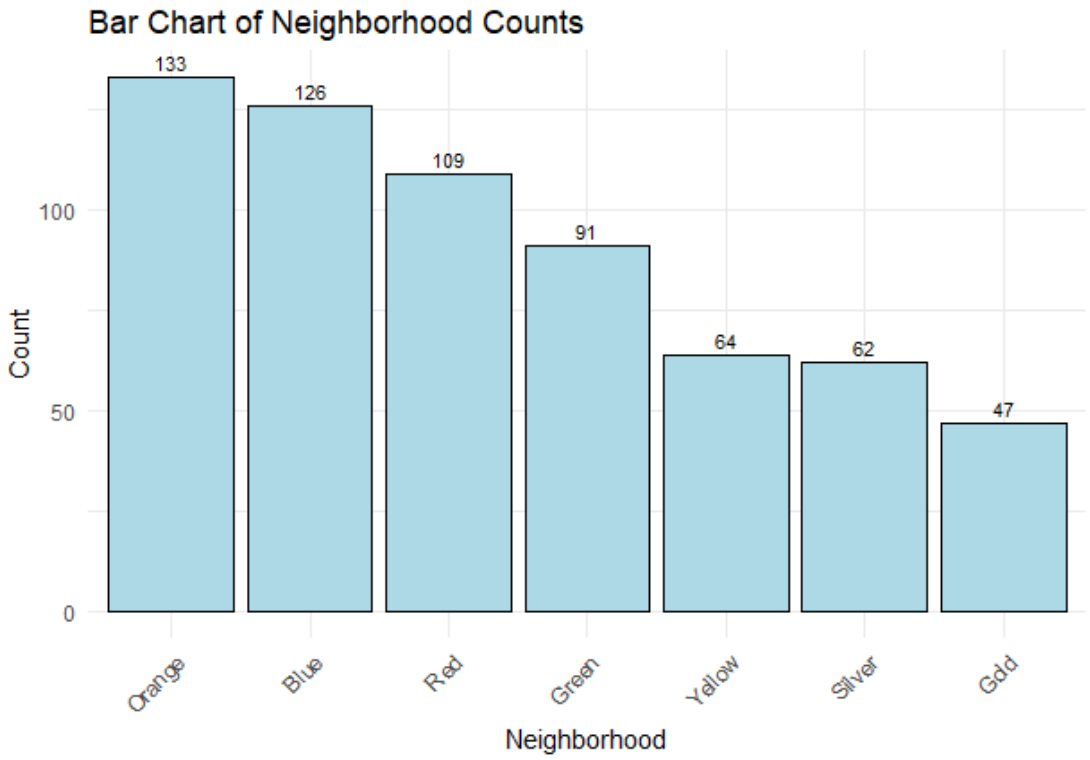
Housing prices are almost normally distributed. Mean and median of housing prices are $1,250,000 and $1,270,000 respectively, which are very close to each other.



Houses in neighborhood gold have the highest median price, then neighborhood green and yellow. Neighborhood red has the lowest median house price.

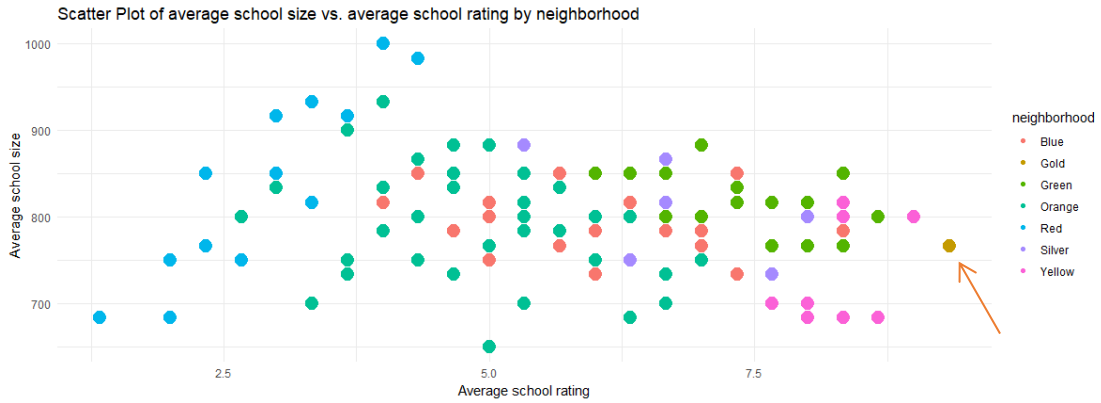


Neighborhood gold is also the youngest neighborhood among all neighborhoods, then neighborhood green and yellow. Neighborhood red is the oldest one. There is a clear pattern here: the most expensive 3 neighborhoods are also the youngest 3 neighborhoods in the exact same order. The cheapest neighborhood is also the oldest.



Neighborhood gold is also the smallest (47) compared to the other neighborhoods, then

silver (62) and yellow (64). Neighborhood orange is the biggest neighborhood (133), then blue (126) and red (109).

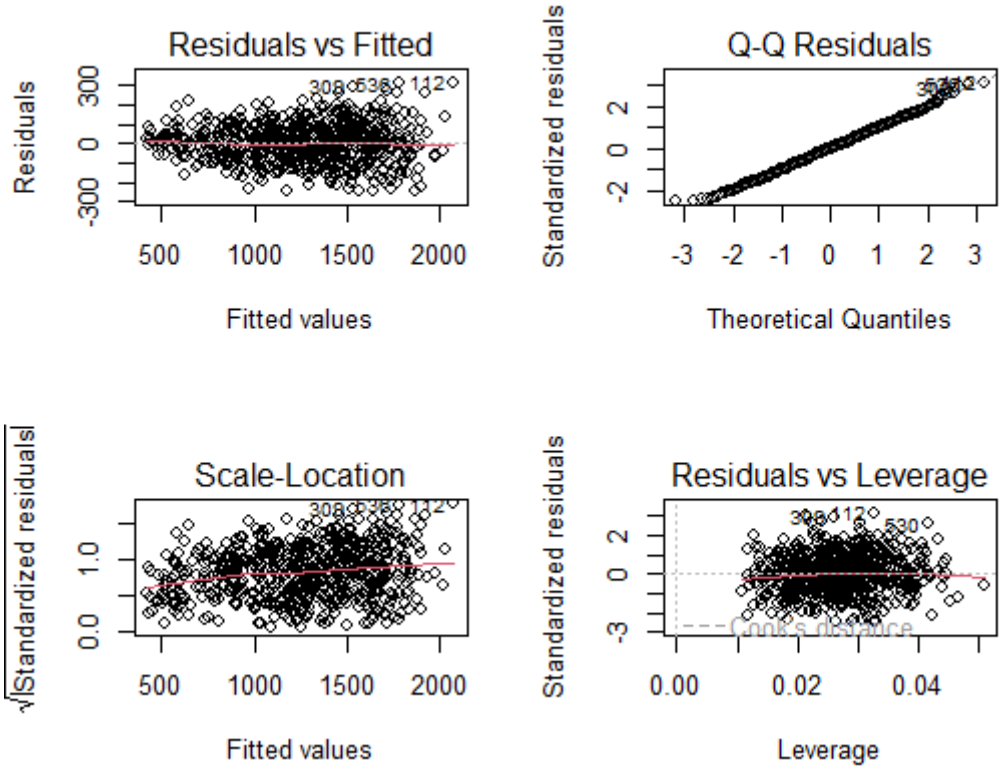


Neighborhood gold is the point on the far right of this graph, representing the highest average school rating (at least 9, on a scale from 0-10) and with average school size somewhere in the middle since this area is newly developed.

If we take a look at all 4 graphs above together, the neighborhood gold is the most newly developed, has the highest housing prices and probably has the most room for growth as well with only 47 existing houses. Good schools would also attract population. It is very clear at this point that gold is the neighborhood that has the most potential growth.

So, we have narrowed down which area has the most potential growth, which is the neighborhood gold. Now, let's see if we can build a regression model to effectively predict housing prices.

Regression model



This looks like a very good regression model. All diagnosis plots look very good with no clear patents.

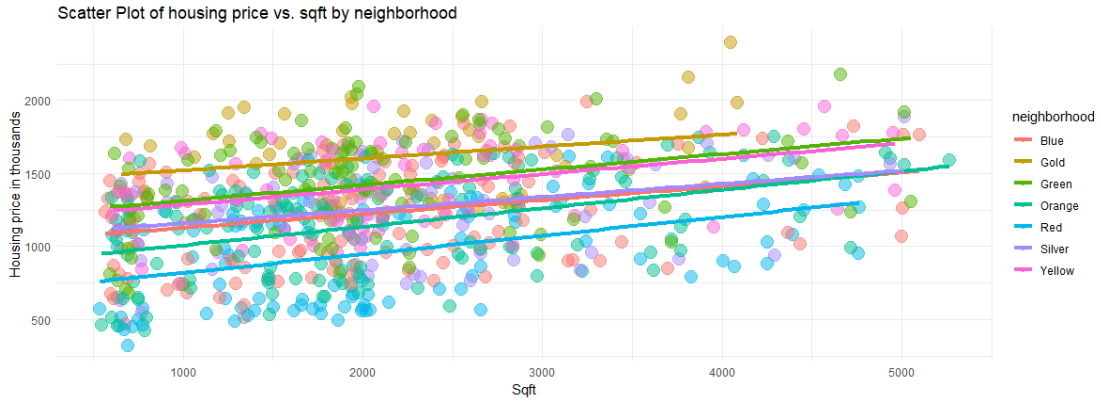
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              474.56258   62.94805   7.539 1.71e-13 ***
neighborhoodGold          56.96537   23.53037   2.421  0.01577 *
neighborhoodGreen         39.05205   15.74592   2.480  0.01340 *
neighborhoodOrange       -23.98558   13.84271  -1.733  0.08365 .
neighborhoodRed           49.23231   21.67282   2.272  0.02345 *
neighborhoodSilver        13.79861   16.76197   0.823  0.41071
neighborhoodYellow       -40.03577   19.73643  -2.029  0.04294 *
beds                      59.46093    7.47980   7.950 8.99e-15 ***
sqft                       0.03316    0.01033   3.211  0.00139 **
typemulti-family home    563.02557   16.08452  35.004  < 2e-16 ***
typesingle-family home   577.69257   10.53339  54.844  < 2e-16 ***
typetownhouse             66.66279   12.90047   5.167 3.21e-07 ***
heatingYes                31.42286   11.55765   2.719  0.00674 **
fireplaceYes              39.60857    9.16249   4.323 1.80e-05 ***
age                       -3.14316    0.24348 -12.909  < 2e-16 ***
mean_size                 -0.20165    0.06802  -2.964  0.00315 **
mean_rating               70.17570    4.75501  14.758  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

All variables listed above are significant predictors to the housing price. Negative

coefficients mean that they have negative impact on housing price, for example, age and average school size. The older the house is, the less it is worth. The more populated the school is, the less houses around it are worth. On the other hand, positive coefficients mean they have a positive impact on housing price, for example, the number of beds and whether it has a fireplace. The more beds there are, the more expensive those houses are. If it has a fireplace, the house is going to worth more too.

With adjusted R-squared equal to 0.93, 93% of the data is explained by the model, which is really good.

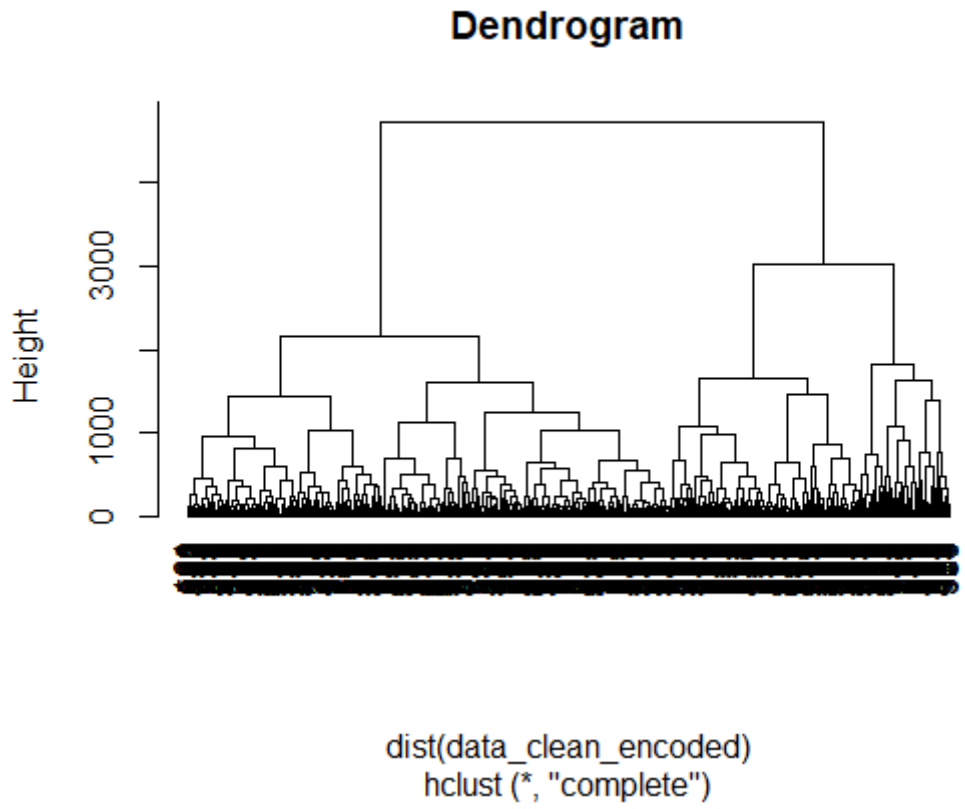
Kmeans clustering



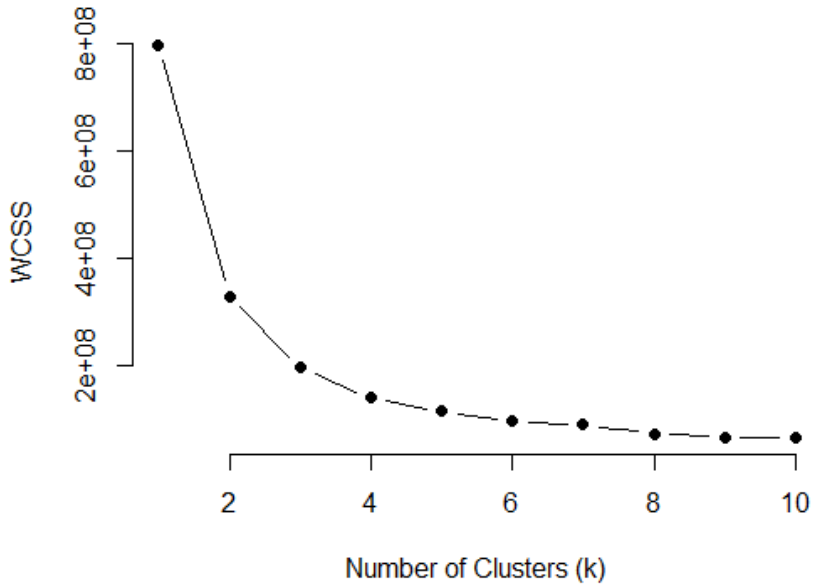
We cannot visually show the clusters because there are just too many dimensions. But here above is a two-dimensional example (sqft vs housing price) with regression lines running through each neighborhood group. Visually, we can see there might be potentially 4 clusters:

1. Gold
2. Green and Yellow
3. Sliver, Blue and Orange
4. Red

Let's reconfirm with the dendrogram and the elbow method.



The dendrogram represents the hierarchical clustering structure of the data. It looks like the ideal number of clusters is also 4.



The elbow method is used to determine the optimal number of clusters by looking at the Within-Cluster-Sum-of-Squares (WCSS). The plot shows the WCSS for a range of cluster numbers (k). The "elbow" point is where the reduction in WCSS starts to slow down, and adding more clusters doesn't significantly decrease WCSS.

It looks like the elbow occurs around k=4, which supports the choice made based on the dendrogram and the scatter plot.

Sensitivity test

Data cleaning

Approach #2

Instead of removing all missing data, let's try to impute them.

1. 2 Nas in Sqft were from neighborhood red with 1 bed and 1 bath. One subset that meet the same criteria was created to calculate the mean, which was used to impute the missing data.
2. 20 Nas in lotsize were imputed by running a regression line between sqft and lotsize.
3. Nas in cooling, heating and fireplace were replaced by "No". Most houses don't have cooling, heating or fireplace. So more likely than not, those houses won't have it either.

All other data cleaning steps are the same as approach #1.

Summary

All results are identical, including all the graphs, regression and clustering results. So, all conclusions we reached with data cleaning approach #1 remind valid: Neighborhood Gold is the area with the most potential growth. Further research needs to be conducted to discover which type of housing should be built and with what features.