COM SCI-X 450.2 Exploratory Data Analysis and Visualization Final Assignment

Name: Chi-Lin Hung SID: X2072273

1. Data Overview

Data summary, oddities, and outliers

- **housing.csv**

Neighborhood		Statistic	Beds	Baths	Sqft	Lot Size	Year	Sold Price
Orange	139	Minimum	1.000	1.000	536	0.0700	1495	664
Blue	135	1st Quartile	3.000	1.000	1349	0.1600	1961	975250
Red	116	Median	4.000	1.500	1952	0.2400	1978	1266500
Green	102	Mean	4.937	2.001	2128	0.2887	1977	1245208
Yellow	68	3rd Quartile	4.000	2.500	2675	0.3600	1997	1551250
Silver	66	Maximum	999.00	25.00	5265	1.3000	2018	2393000
Others	54	NA's			2	19		

High School		Middle School		Elementary School	
Channel	69	Wolf	63	Leopard	47
Delta	66	Fox	56	Kodkod	37
Alpine	58	Hound	56	Savannah	37
Crevasse	58	Coyote	53	Puma	36
Ravine	52	Sloth	51	Ocicat	35
Avalanche	51	Raccoon	48	Cheetah	34
Others	326	Others	353	Others	454

Levels		heating	
1	422	No	530
2	255	Yes	146
?	6		7
cooling		**fireplace**	
No	454	No	459
Yes	222	Yes	218
	7		6

Property Type	
Condo	150
Multi-family Home	69
Single-family Home	331
Town House	130

- **school.csv**

Statistic	Size (Students)	Rating
Minimum	500.0	1.000
1st Quartile	700.0	4.000
Median	750.0	6.500
Mean	784.6	6.115
3rd Quartile	862.5	8.000
Maximum	1250.0	10.000
NA's		

Elementary School	
length	52
type	character

School dataset has complete data, and does not have oddities or outliers.

After conducting a statistical summarization of the housing data, I found lots of missing data and that some units have irrational number of bedrooms or bathrooms. One house has 999 bedrooms, and the other has 25 bathrooms. I then look closer into the data on the excel sheet and find out that the 999-bedroom unit has only 1 bathroom and 753 square and that the 25-bathroom unit has 4 bedrooms. I reckon that 999 is probably a typo, whereas the 25-bathroom unit might either be a typo or this building used to be a bathhouse. Since the 999 bedrooms unit is obviously insane, I elide this row of data while implementing analysis later on. As for the 25-bedroom unit, though it seems to be a possible design, it is still an extreme outlier, so I will also elide it. For the NAs, I'll firstly ignore them for plotting to grasp a more precise tendency, and then implement listwise deletion for statistical analysis.

In addition, two units include abnormal built years. One unit was built in 1495. If the year is correct, this unit should be a monument or cultural heritage, but it on sale. I will keep the data as a marketing highlight for the area, but omit it when plotting and analyzing so viewers will have better picture of the data. The other unit is built in 2111, in 88 years, which is impossible, so I will elide this.

In terms of sold price, there is also a weird price in orange neighborhood, which is sold only $664. As the unit is of a regular size with normal equipment and is relatively new, the price is quite doubtful. However, since other invisible uninhabitable factors such as supernatural events cannot be ruled out, I will keep this row of data.

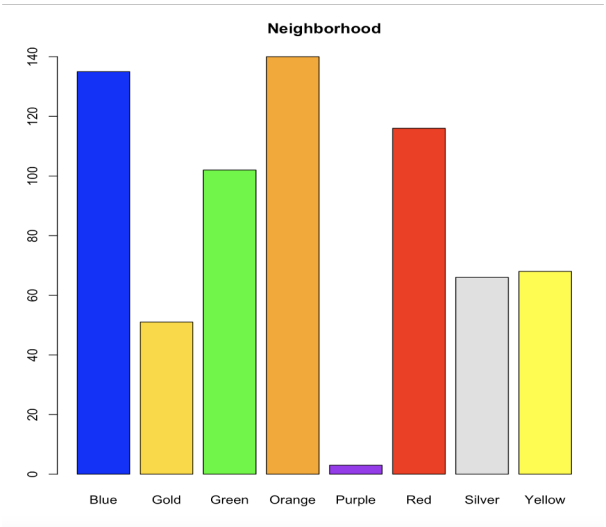
For unit type, there are two different "town house" and "condo". I then unify the spelling. Therefore, there should be 4 types of unit in total, instead of 6. Lastly, for NA, I will first implement listwise deletion.

2. Data Cleaning

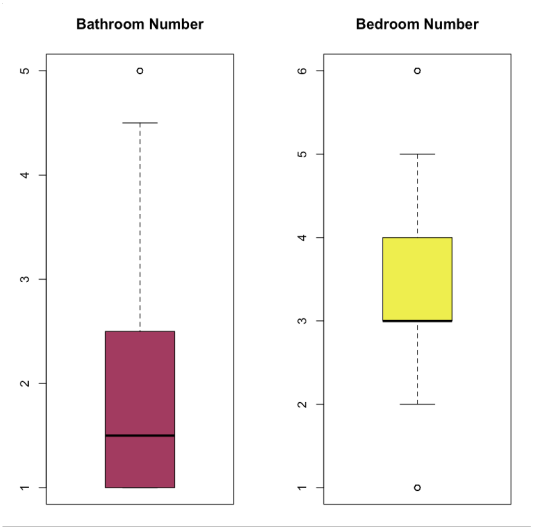
Dataset modification or merge

a. Remove the rows (units) with irrational figures: with 999 bedrooms, with 25 bathrooms, built in 2011 & 1495
b. Unify the spelling of “town house” and “condo” in type
c. Merged the housing and school datasets using ***left_join()*** function. I merged each of the three stages seperately.

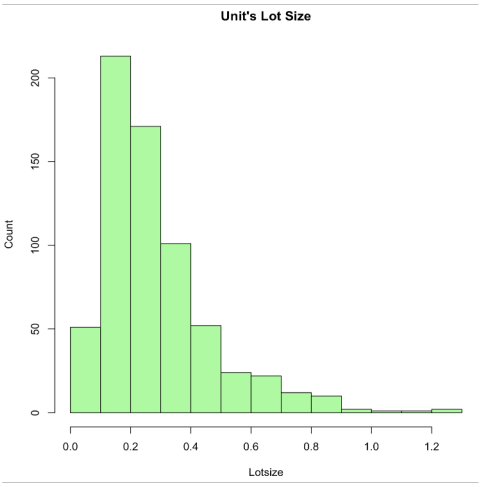
3. One-Variable Visuals



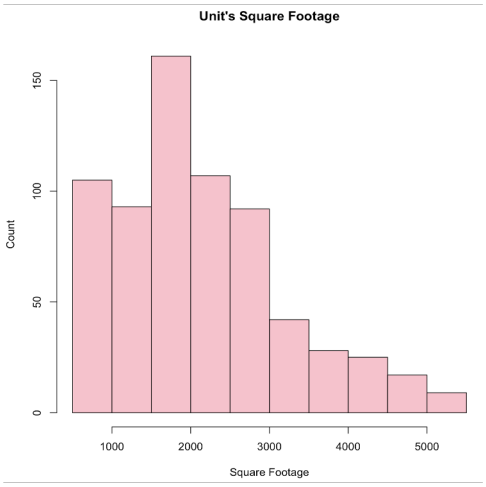
The bar plot shows the number of units in each neighborhood. There are 8 neighborhood in total. Orange and Blue ones have the most units, both more than 130. Red and Green comes in the 3rd and 4th place, with also more than 100 units. Silver and Gold ones have just over 60 units. Gold unit has around 50 units. Purple has the least, with only 3 units.



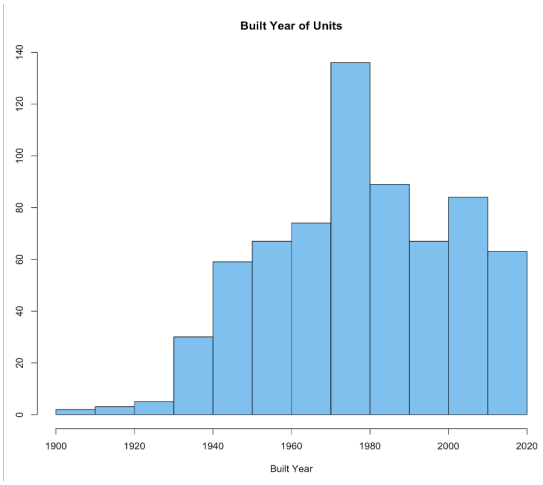
From the boxplots, we see that most units are equiped with 1 to 2.5 bathrooms and 3 to 4 bedrooms. On average there are 1.5 bathrooms and 3 bedrooms in a unit.



The green histogram illustrates the units’ lot size. Most of the units are at around 0.1 to 0.4. The mode is at around 0.2 lot size. Trendily speaking, the lot size data is right skewed. The larger the size, the fewer the unit. We see that the largest units have more than 1.2 in size.



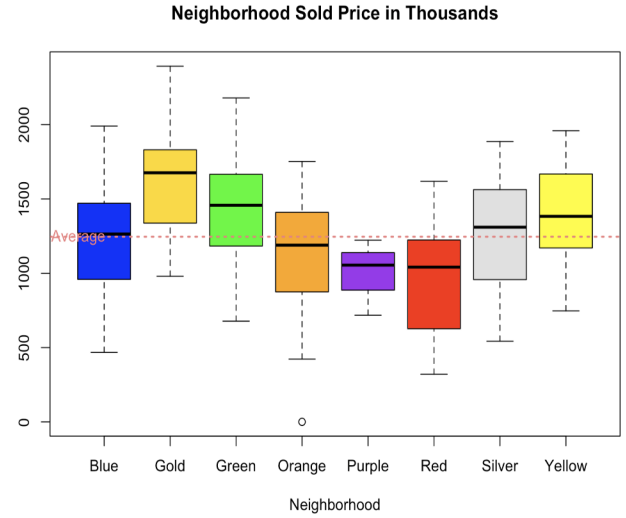
The pink histogram shows the units’ square footage. Most of the units are less than 3000. The mode is at around 1500 to 2000. There are still a fare amount of units size between 3000 and 5000, which account for more than 100, around one-fifth of the total unit count. Approximately 10 units are more than 5000 square footage in size.



The blue histogram illustrates the units’ built years. Excluding the extreme and irrational outliers, we see that most units are built between 1940 and 2020 in the target areas. From 1900s onwards, the number of units built increase every decade, reaching a peak in 1970s. Nearly 140 units were built within this decade. After 1980, around 70 units were built on average per decade.

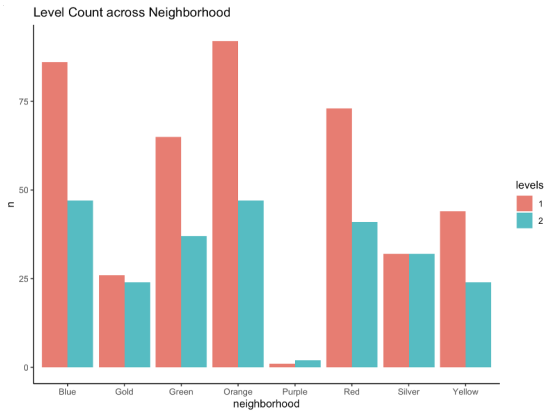
4. Two or More Variables Visuals

a. Neighborhood features

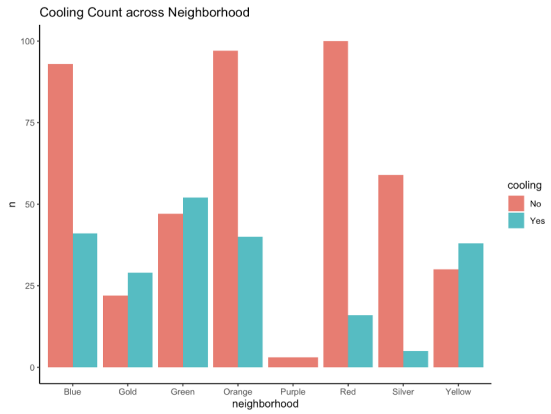


The boxplot presents the relation between unit sold price and their neighborhood. The Gold neighborhood is the most expensive, over 75% of unit is above 1.4 million, whereas the cheapest on would be red neighborhood. Green, Yellow, and Blue are relatively expensive, with over 50% units above market price.

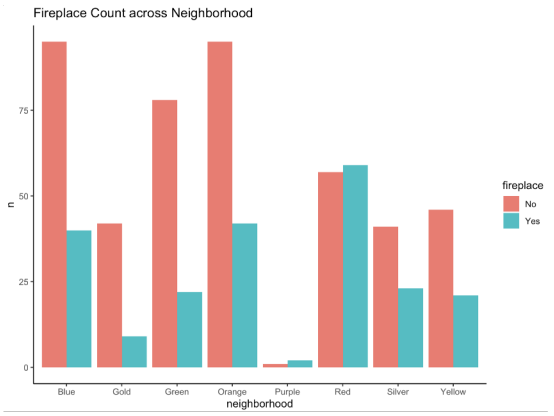
This high density plot shows the built year of units in each neighborhood. Blue and Orange neighborhoods have units built across 60 years from 1940 to 2022. Gold, Green, and Yellow are newer neighborhoods where units are all built after 1970. Red and Silver are neighborhoods with older units. All units were built before 1980s. Red is the only one with units built earlier than 1940.



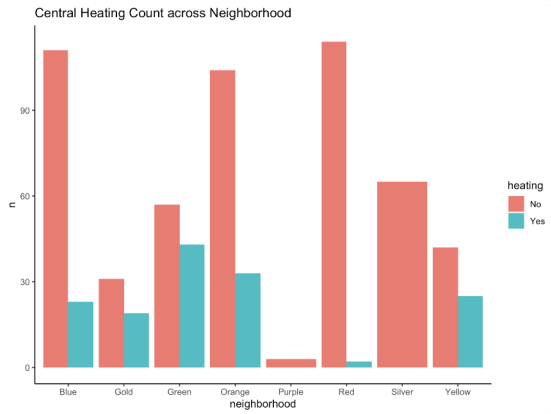
The bar plot shows the levels of units in the neighborhoods. Relatively higher percentage of Blue, Green, Orange, Red, and Yellow ones have only one level in a unit. Gold, Silver, and Purple have equal or similar amount of one- and two-level units.



The bar plot shows the amount of units with cooling in all neighborhoods. More units in Gold, Green, and Yellow neighborhoods install cooling.



The bar plot shows the amount of units with fireplace in all neighborhoods. Higher percentage of most neighborhoods doesn't have fireplace, except for Red and Purple ones.



The bar plot shows the amount of units with heating in all neighborhoods. All units in Silver and purple and almost all Red units do not have heating. Meanwhile, the plot suggests that heating is less installed across neighborhoods.

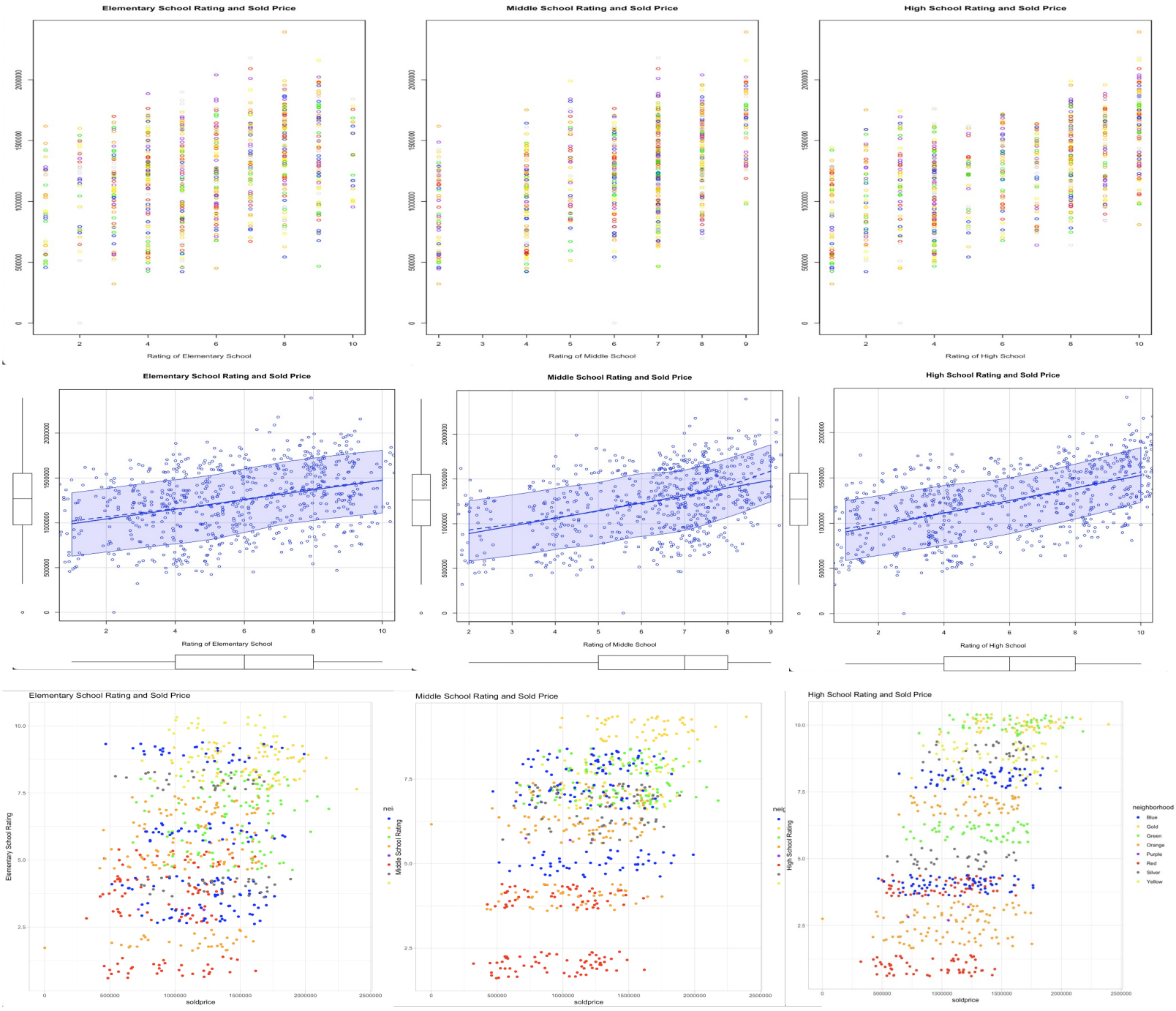
b. What are the determinant or relevant features of sold price on the basis of different neighborhood?

In this subsection, we discuss the relation between price and other features.

First of all, the heatmap shows the relation among units' neighborhood, types, and their sold price. The shades of color represent the price of the unit. The darker the color, the higher the price. The number in the blocks are the count of the type of units corresponding to a particular neighborhood.

In total, there are four types of units. It is shown that multi-family home and single-family home are the most expensive type across all neighborhoods. There's no multi-family home in the Red, Purple, and Orange neighborhood. This is probably the reason why these neighborhood are comparatively cheaper. The Gold, Green, and Yellow neighborhoods units are the most expensive regardless of the type. Every neighborhood has condo and single-family home. Only Purple neighborhood does not have townhouse.

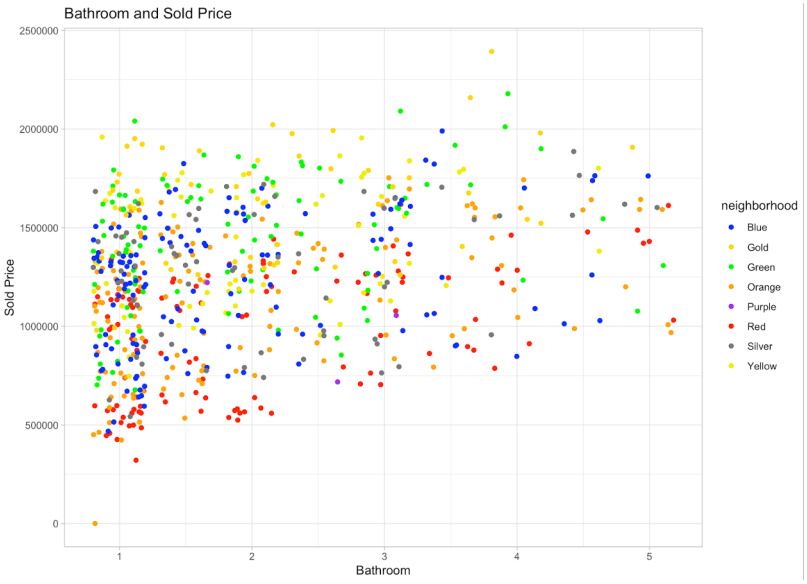
School Rating and Unit Sold Price



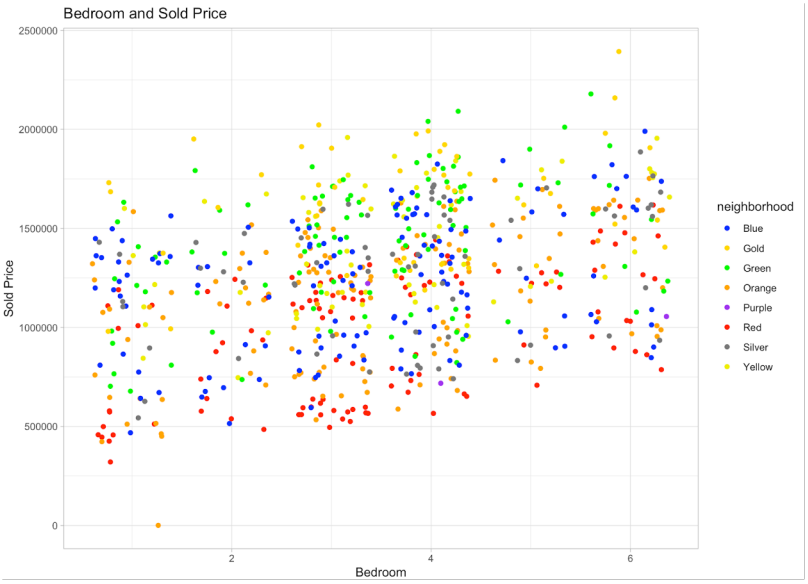
This page present the school rating in relation to the units sold price and their relations to neighborhood names.

Different types of high density plots are demonstrated to provide details with variety. The first row of density plots and special plots suggest that the price may relate to the school rating the unit assigned to. When a unit is assigned to a higher rating school, it is more likely to have higher price. From the seller's perspective, it is more likely to raise the price of a unit if it is near a high-rating school. From the buyers' perspective, it is still possible to find a more affordable house near good schools.

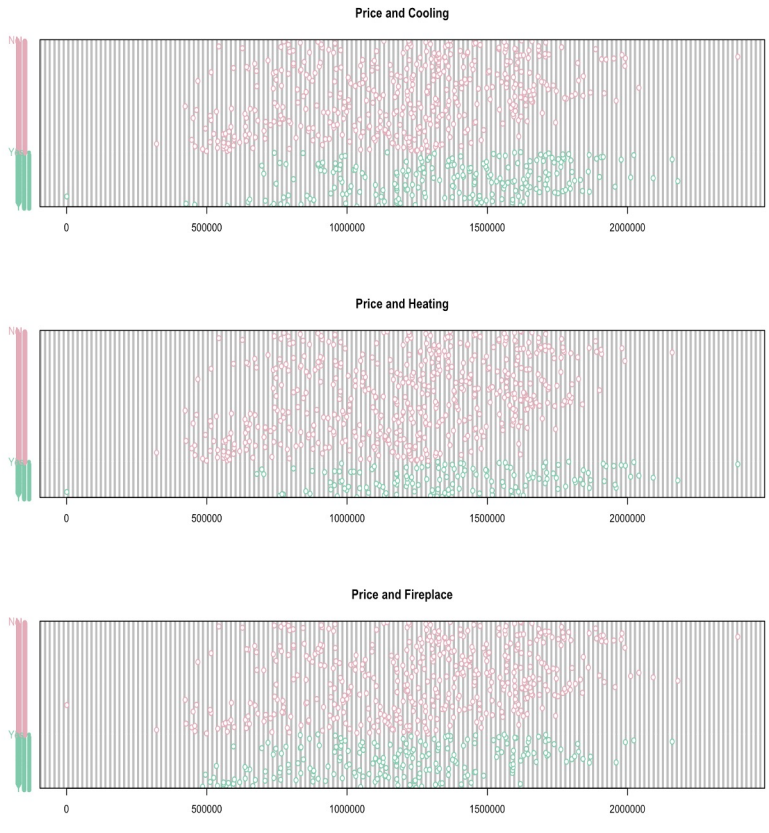
To have a closer look into the relation with neighborhoods, the third row of jitter plot provides clearer view. As can be seen, the Gold, Yellow and Silver neighborhood are close to very high-rating elementary schools, and relatively high ranked middle and high schools. Green neighborhood is close to very good middle school, but may not necessarily be close to top elementary or high school. Approximately half of the units in the Blue neighborhood are close to good schools. Orange neighborhood are close to mediocre schools. Lastly, the Red neighborhood is close to low-rating school of all three stages.



This scatter plot shows the relation between the number of bathroom and unit sold price, colored by neighborhood. From the distribution, we can see that units with more bathroom more likely correspond with higher price, but not necessary, indicating that price may also be affected by other factors.



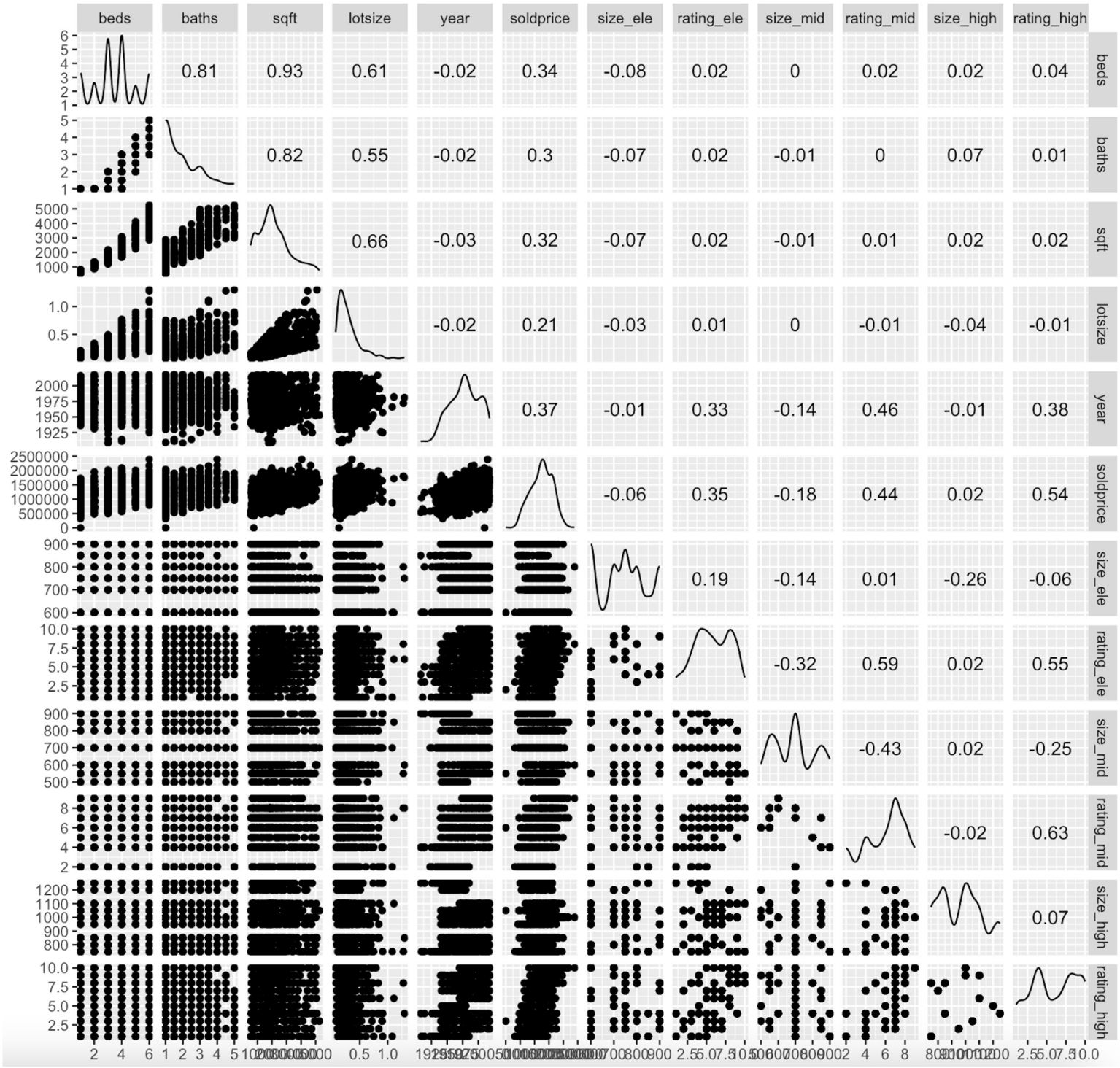
This scatter plot shows the relation between the number of bedroom and unit sold price, colored by neighborhood. Similar to the previous plot, more bedrooms relates to higher price, while it is not necessary the case and the variation in price may not be too much.



The dot charts illustrate the relations between sold price and whether cooling, heating, or fireplace is installed in a unit. It seems that the relation is not significant. I'd suggest that whether cooling, heating, or fireplace is in a unit does not affect the sold price.

5. Analysis

Overview



The scatter matrix provides the relations among variables. There are certainly some highly related variables. Some relations make sense, such as bedrooms and bathrooms and sizes and rooms. Some does not, like school ratings among stages. What I want to highlight is the relation between price and the rest variables. It seems that except for schools sizes, most features seem to relate to price to a certain degree.

Regression

In this section, I am digging into the features that affect the up and down of the sold price of unit.
I use the relevant features mentioned earlier and add neighborhood and unit type as independent variables to create my first model.

	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	-5615000.00	407900.00	-13.765	< 2e-16	***
neighborhoodGold	62160.00	22510.00	2.761	0.00593	**
neighborhoodGreen	25930.00	14630.00	1.772	0.07680	
neighborhoodOrange	-8730.00	13330.00	-0.655	0.51271	
neighborhoodPurple	20150.00	57440.00	0.351	0.72584	
neighborhoodRed	25660.00	22980.00	1.117	0.26456	
neighborhoodSilver	-12710.00	15600.00	-0.815	0.41552	
neighborhoodYellow	-6596.00	18180.00	-0.363	0.71692	
beds	59350.00	7164.00	8.284	7.01e-16	***
baths	10550.00	6362.00	1.658	0.09777	.
sqft	21.08	10.52	2.005	0.04543	*
lotsize	26070.00	27400.00	0.951	0.34182	
year	2958.00	204.60	14.459	<2e-16	***
typemulti-family home	565800.00	14880.00	38.016	<2e-16	***
typesingle-family home	578000.00	9797.00	N/A	<2e-16	***
typetown house	84240.00	12030.00	7.006	6.25e-12	***
rating_ele	6516.00	2293.00	2.841	0.00464	**
rating_mid	11410.00	4103.00	2.781	0.00558	**
rating_high	43700.00	2133.00	20.488	< 2e-16	***

Table 1. *Model 1 Result*

From the result, we see not all variables are significant. Only the three school ratings, year, bedrooms, square footage, three types of units, and the Gold neighborhood (p < 2.2e-16, R-square = .9356)

	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	-5315000.00	1037000.00	-5.127	3.89e-07	***
beds	75290.00	20340.00	3.701	0.000233	***
sqft	12.30	27.93	0.44	0.659907	
year	2957.00	532.70	5.552	4.13e-08	***
rating_ele	-24.72	5923.00	-0.004	0.996672	
rating_mid	20960.00	8185.00	2.561	0.010648	*
rating_high	49650.00	5088.00	N/A	9.7592e-16	***

Table 3. *Model 3 Result*

The second outcome shows that all features except for some of the types are significant. I then wonder what the result would be like if taking out the unit type, so here's the outcome without type as the independent variable. It turns out type is an indispensable variable as the explanatory power drops almost 50% (p < 2.2e-16, R-square = .4342).

	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	-5724000.0	360200.000	-15.891	< 2e-16	***
beds	64020.0	7071.000	9.054	< 2e-16	***
sqft	27.0	9.715	2.780	0.00560	**
year	3015.0	185.000	16.295	<2e-16	***
typemulti-family home	568700.0	14960.000	38.013	< 2e-16	***
typesingle-family home	579700.0	9882.000	58.664	<2e-16	***
typetown house	88380.0	12090.000	7.310	7.9e-13	***
rating_ele	6557.0	2063.000	3.178	0.00155	**
rating_mid	10260.0	2856.000	3.594	0.00035	***
rating_high	45260.0	1778.000	25.452	<2e-16	***

Table 2. *Model 2 Result*

I then conducted another test using only the significant variables. Here, I also elide neighborhood because I construe that only one-eighth of a feature being significant may not make much difference. The result confirms my theory (p < 2.2e-16, R-square = .9321). Not much difference was found in the explanatory power.

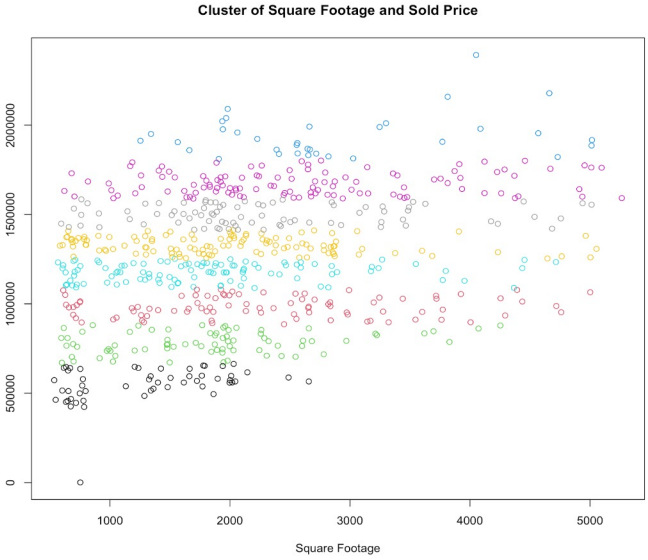
	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	437684.7	34613.7	12.645	< 2e-16	***
typemulti-family home	541803.0	89387.4	6.061	2.29e-09	***
typesingle-family home	681332.0	40532.9	16.809	< 2e-16	***
typetown house	196200.9	46033.7	4.262	2.32e-05	***
rating_high	74550.8	5299.0	14.069	< 2e-16	***
typemulti-family home:rating_high	-120.2	11893.4	-0.010	0.9919	
typesingle-family home:rating_high	-14742.5	6200.5	-2.378	0.0177	*
typetown house:rating_high	-15176.3	7200.7	-2.108	0.0354	*

Table 4. *Model 4 Result*

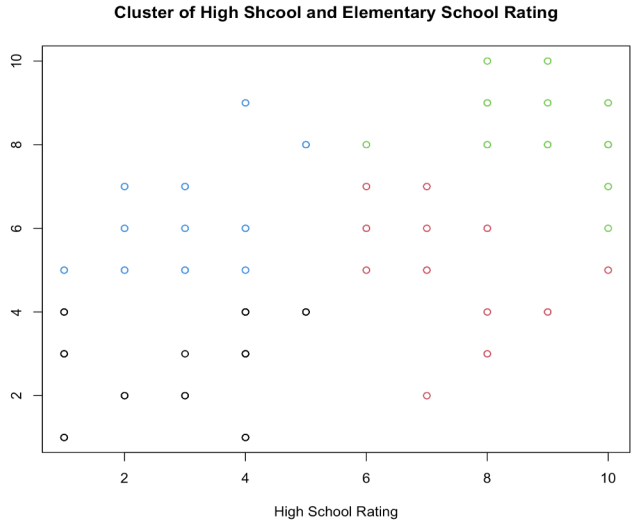
From the previous outcomes, I find that not only unit type but also high school rating has always been very crucial variables. Thus, I conduct an interaction regression model. The result suggests that single-family home or town house type together with high school rating are significant in terms of explaining the sold price (p < 2.2e-16, R-square = .7926).
For single-family home type with high school rating, the equation looks like: 437684.7 + 681332.0*typesingle-family home + 74550.8*rating_high – 14742.5*typesingle-family home:rating_high.
For town house type with high school rating, the equation would be: 437684.7 + 196200.9*typetown house + 74550.8*rating_high – 15176.3*typetown house:rating_high.

Cluster

As a way to form groups, this cluster plot is based on the high school rating and the sold price. The number of the expected cluster is eight, which is the same as the amount of neighborhood. We can see that the groups here are formed based more on price rather than the high school rating. That is, units of similar price are categorized into the same cluster.



In this cluster plot, I apply the square footage and the sold price as the basis. The cluster count is still 8. Again, the groups are formed based more on price rather than the high school rating. Similar prices units are clustered into the same group.



Here I choose two discrete features for prediction. The features are high school and elementary school ratings, and the cluster is defined to be 4. We see that the data are categorized into high-high, high-low, low-high, and low-low in terms of rating.

6. Sensitivity Analysis

In the previous analysis, I implemented listwise deletion (i.e. delete the rows with NA). For sensitivity analysis, I am going to fill the categorical NA (i.e., "?", or "") values with the majority category and the numerical variables with the mean. Since the NA values comprise merely a very small portion. In this analysis, I will focus only on the regression analysis. The extreme outliers are still removed.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -5.615e+06  4.079e+05 -13.765  < 2e-16 ***
neighborhoodGold         6.216e+04  2.251e+04   2.761  0.00593 **
neighborhoodGreen        2.593e+04  1.463e+04   1.772  0.07680 .
neighborhoodOrange      -8.730e+03  1.333e+04  -0.655  0.51271
neighborhoodPurple       2.015e+04  5.744e+04   0.351  0.72584
neighborhoodRed          2.566e+04  2.298e+04   1.117  0.26456
neighborhoodSilver      -1.271e+04  1.560e+04  -0.815  0.41552
neighborhoodYellow      -6.596e+03  1.818e+04  -0.363  0.71692
beds                     5.935e+04  7.164e+03   8.284 7.01e-16 ***
baths                    1.055e+04  6.362e+03   1.658  0.09777 .
sqft                     2.108e+01  1.052e+01   2.005  0.04543 *
lotsize                  2.607e+04  2.740e+04   0.951  0.34182
year                     2.958e+03  2.046e+02  14.459  < 2e-16 ***
typemulti-family home    5.658e+05  1.488e+04  38.016  < 2e-16 ***
typesingle-family home   5.780e+05  9.797e+03  59.000  < 2e-16 ***
typetown house           8.424e+04  1.203e+04   7.006 6.25e-12 ***
rating_ele               6.516e+03  2.293e+03   2.841  0.00464 **
rating_mid               1.141e+04  4.103e+03   2.781  0.00558 **
rating_high              4.370e+04  2.133e+03  20.488  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 96520 on 639 degrees of freedom
  (21 observations deleted due to missingness)
Multiple R-squared:  0.9356,    Adjusted R-squared:  0.9338
F-statistic: 515.9 on 18 and 639 DF,  p-value: < 2.2e-16
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-980909 -286779  104394  233693  560843

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.174e+06  1.026e+06  -5.044 5.88e-07 ***
beds         6.874e+04  1.993e+04   3.449 0.000598 ***
sqft         2.290e+01  2.739e+01   0.836 0.403381
year         2.885e+03  5.269e+02   5.475 6.18e-08 ***
rating_ele   9.602e+02  5.848e+03   0.164 0.869634
rating_mid   1.976e+04  8.068e+03   2.449 0.014562 *
rating_high  5.027e+04  5.019e+03  10.016  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 283100 on 670 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.4349,    Adjusted R-squared:  0.4299
F-statistic: 85.95 on 6 and 670 DF,  p-value: < 2.2e-16
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-628849  -61302   -3777   62716  369344

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -5.721e+06  3.597e+05 -15.905  < 2e-16 ***
beds                     6.287e+04  6.987e+03   8.999  < 2e-16 ***
sqft                     2.870e+01  9.610e+00   2.986 0.002930 **
year                     3.011e+03  1.847e+02  16.306  < 2e-16 ***
typemulti-family home    5.615e+05  1.469e+04  38.214  < 2e-16 ***
typesingle-family home   5.797e+05  9.825e+03  59.000  < 2e-16 ***
typetown house           8.916e+04  1.199e+04   7.434 3.24e-13 ***
rating_ele               6.938e+03  2.053e+03   3.379 0.000769 ***
rating_mid               1.065e+04  2.839e+03   3.752 0.000190 ***
rating_high              4.526e+04  1.770e+03  25.573  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 99220 on 667 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.9309,    Adjusted R-squared:  0.93
F-statistic: 998.8 on 9 and 667 DF,  p-value: < 2.2e-16
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-654736 -119931   -8023  103311  676519

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                         426834      33334  12.805  < 2e-16 ***
typemulti-family home               562361      88389   6.362 3.68e-10 ***
typesingle-family home              690247      39283  17.571  < 2e-16 ***
typetown house                      205536      44966   4.571 5.78e-06 ***
rating_high                          76189       5137  14.832  < 2e-16 ***
typemulti-family home:rating_high    -3695      11771  -0.314   0.7537
typesingle-family home:rating_high  -16249       6038  -2.691   0.0073 **
typetown house:rating_high          -16281       7043  -2.312   0.0211 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 171800 on 671 degrees of freedom
Multiple R-squared:  0.7934,    Adjusted R-squared:  0.7913
F-statistic: 368.2 on 7 and 671 DF,  p-value: < 2.2e-16
```

The result doesn't show much difference. The first model R-square remains 0.9356. The second model drops 0.0012, which is 0.9309. The third and fourth drop less than 0.001. The significant variables remain significant. The correlations decrease very slightly. This means the results aren't actually affected. This is probably because the amount of NA aren't many, or the imputations suit the dataset.