1. **Before the final assignment, I read the file:**

Read two csv datasets from the file:

housing <- read.csv("housing.csv")

schools <- read.csv("schools.csv")

Merging the above two variables

housing <- merge (housing, schools, by.x = "elementary", by.y = "school")

Also, creating a file to store images

if(!file.exists(paste(getwd(), '/Image', sep='' )))

dir.create("./Image")

1. **Data summary, oddities, and outliers**

I use the “summary” code to roughly check the data. However, because half of the column are characters, I cannot check if there is some oddities or outliers until I convert character columns to factors. for ex:

housing$neighborhood <- as.factor(housing$neighborhood)

Then, re-check the variables by “summary” again.

It clearly showed me that there are three distinct types of abnormalities: **outliers**, **oddities**, and **empty values** (Figure 1).

A screenshot of a computer

Description automatically generated with low confidence

**Figure 1**

When I first check the outliers, there are two methods that I chose.

1. Using scattering: (Figure 2)

Calendar

Description automatically generated

**Figure 2**

2. Using box plot: (Figure 3)

Chart, box and whisker chart

Description automatically generated

**Figure 3**

Finally, I summarized what I should not do in session 2. (Table 1)

|  |  |
| --- | --- |
| **Column names** | **What I should do** |
| **beds** | Remove outliers |
| **baths** | Remove outliers |
| **sqft** | Remove or replace NA's |
| **lotsize** | Remove or replace NA's |
| **year** | Remove lowest (1495) and impossible data (2111) |
| **type** | Renamed miswritten words |
| **levels** | Remove or replace “?” |
| **cooling** | Remove or replace empty |
| **heating** | Remove or replace empty |
| **fireplace** | Remove or replace empty |

**Table 1**

1. **Data cleaning**

The “beds,” the “baths,” and the “year” variables tell me that each variable has one salient outlier. As a result, I decide to remove the spike because it is meaningless, or the data is wrong. (Even if it is rational, no repetition of data can hardly assist me in analyzing data)

housingclean <- housing[housing$beds < 100, ]

housingclean <- housingclean[housingclean$baths < 15, ]

housingclean <- housingclean[housingclean$year > 1600 & housingclean$year < 2100, ]

Except for that, in the “type” column, I found some typing issues. It is obvious that some parameters “townhouse” was miswritten into “town house.” Therefore, I renamed one of them. Besides, because factor variables will cause residues (the original parameter item names keep, but the summary table is right), I convert the variables to characters, then roll back to factors to avoid the issues.

housing$type <- as.factor(housing$type)

housingclean$type[housingclean$typ == 'townhouse'] <-'town house'

housing$type <- as.character(housing$type)

housing$type <- as.factor(housing$type)

Then, I can choose to remove the “?”, “NA,” “empty” row or replace “?”, “NA,” and “empty” with a given data. In this session, I want to remove them directly. However, because the “lotsize” column has 20 “NA”, I cannot just delete them. As a result, I should check these 20 rows if those rows are different from others in other columns.

I immediately check them by box plots. In (Figure 4), The “True” box means the “lotsize” column’s data is filled with the data, while “False” means the “lotsize” is missing. It tells me that these 20 rows are normal: All of them are similar to the background dataset. So, I can remove them directly.

housingcleanNA <- housingclean[!is.na(housingclean$sqft),]

housingcleanNA <- housingcleanNA[!is.na(housingcleanNA$lotsize),]

housingcleanNA <- housingcleanNA[which(housingcleanNA$levels !='?'),]

housingcleanNA <- housingcleanNA[which(housingcleanNA$cooling !=''),]

housingcleanNA <- housingcleanNA[which(housingcleanNA$heating !=''),]

housingcleanNA <- housingcleanNA[which(housingcleanNA$fireplace !=''),]

Box and whisker chart

Description automatically generated

**Figure 4**

And when I turn it to the character variable and roll it back to the factor variable, the summary table is clear finally (Figure 5).

**A close-up of a document

Description automatically generated with medium confidence**

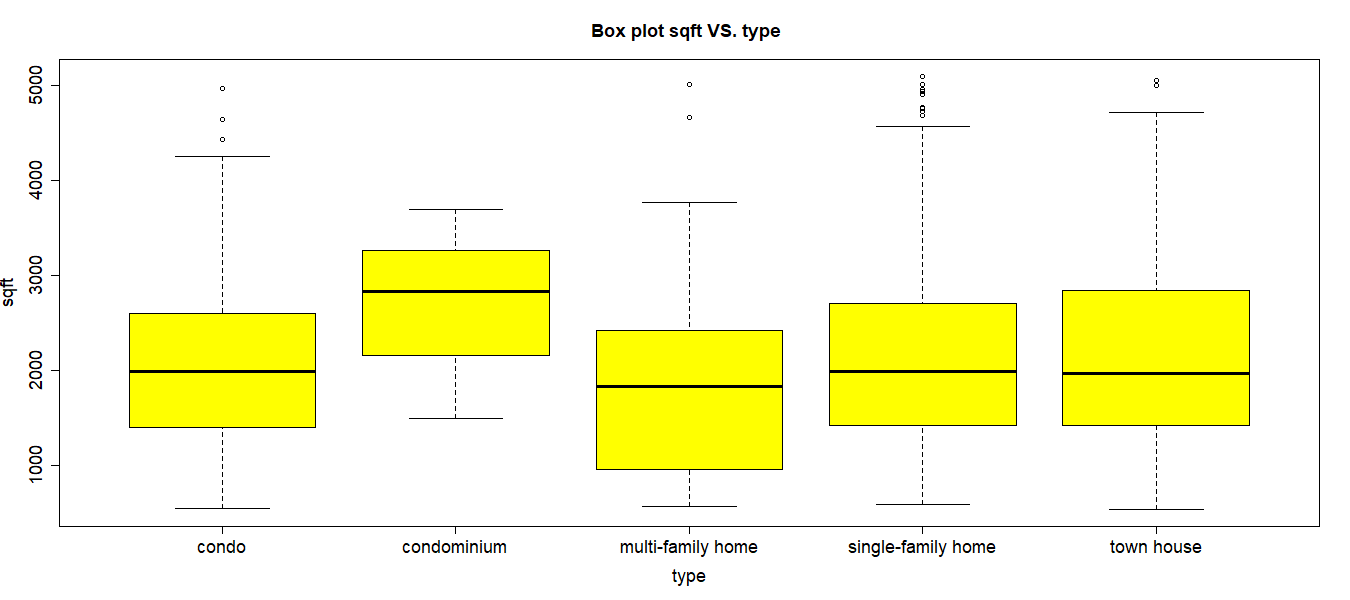
**Figure 5**

1. **One-variable visuals**

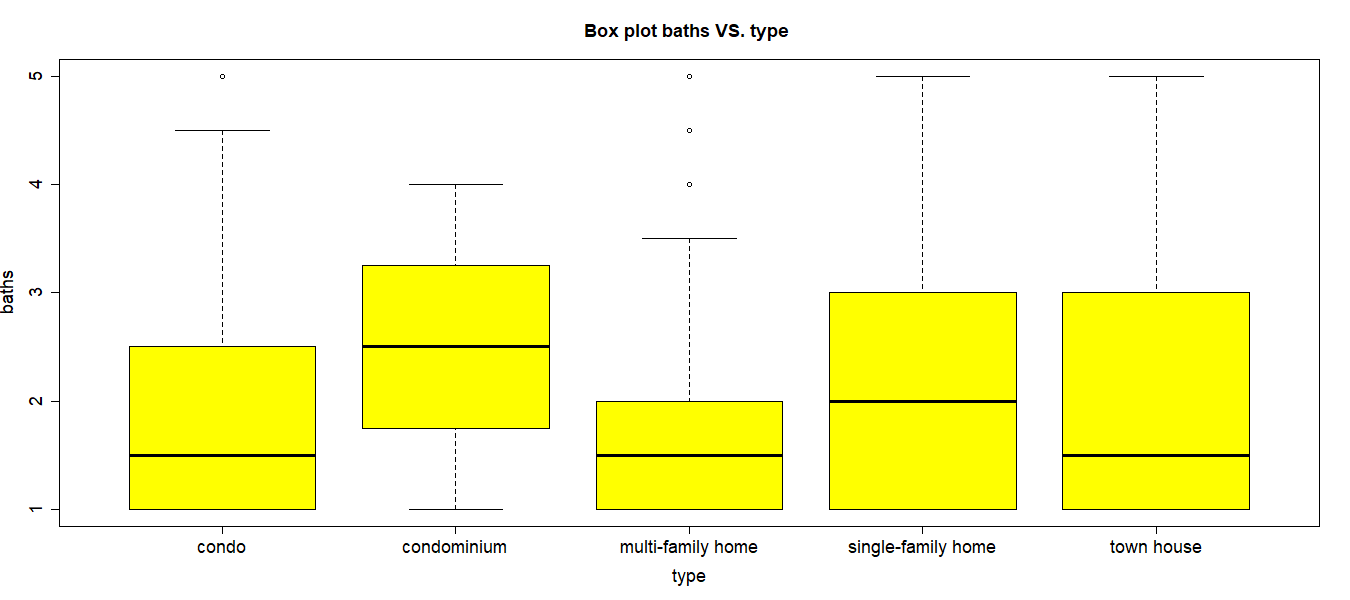
P.S: Before starting my visualization, I try to build my favorite function. However, I still did not complete all of them. Several functions such as bar plots and complex histograms are waiting to be finished (I want my favorite function can print the title, the xlab, the ylab, directly generating multi-graph in one curve... etc, but it will cost me too much time). I will put them in the follow-up.

To begin with, I first check the “type” column, because it cost me a lot of effort to manage previous issues.

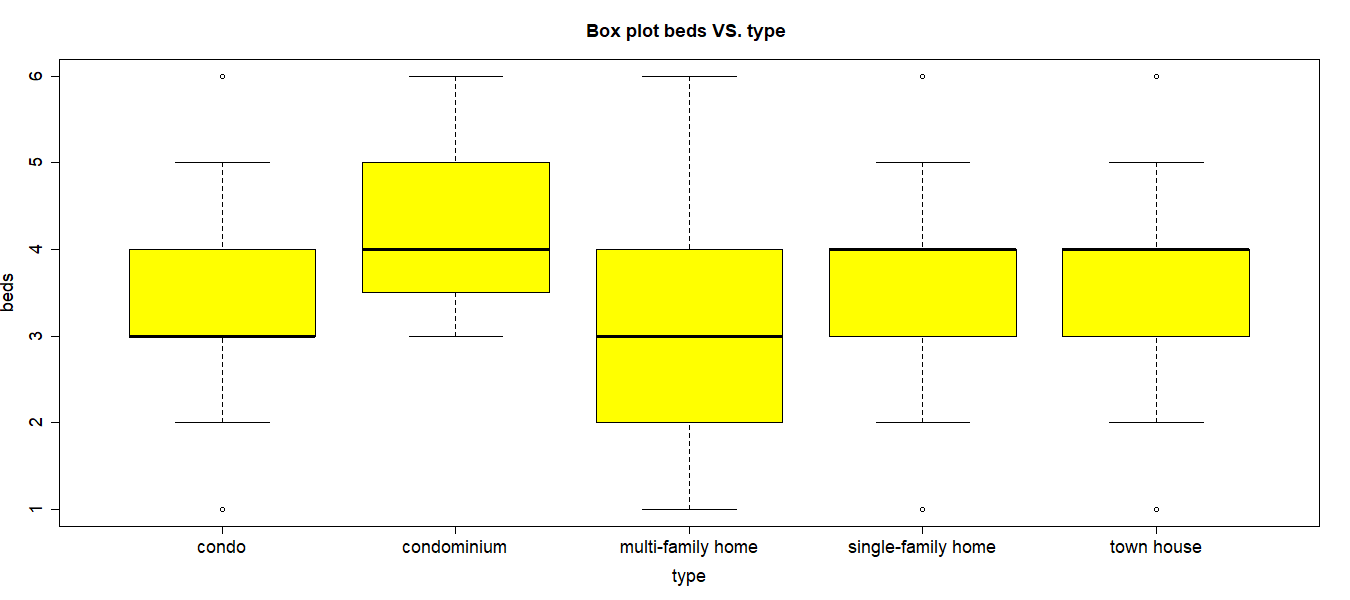
To remind, the “condominium” bar has only three row data, so even though its behavior is different from others, I will not spend time discussing it. When I first check the bar plot, most of them looks similar (Figure 6~9), except for the Figure 10, the “multi-family home” and the “single-family home” looks comparative higher than others. Maybe it gives me a good hint.



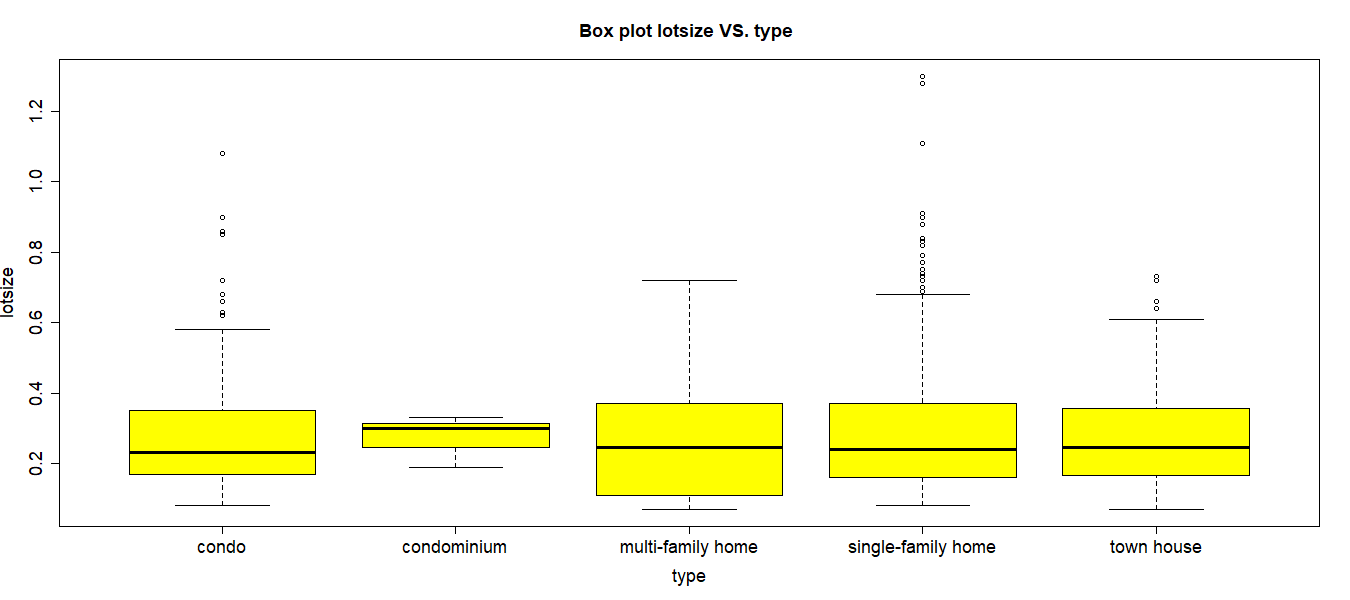
**Figure 6**



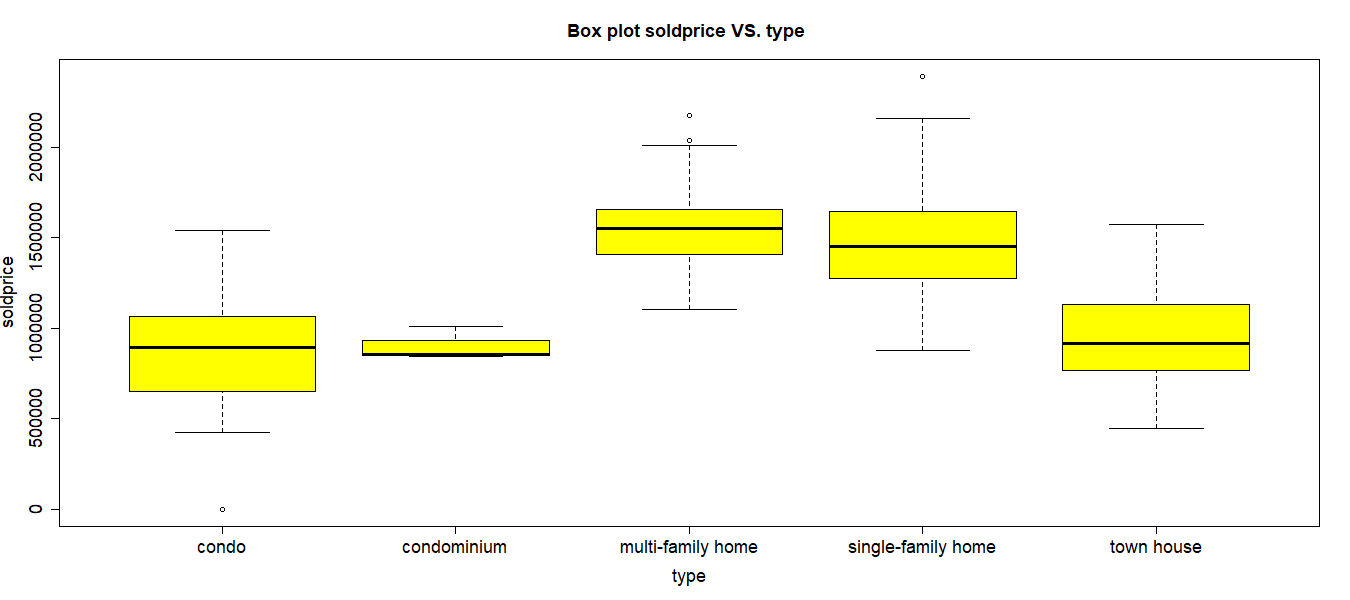
**Figure 7**



**Figure 8**

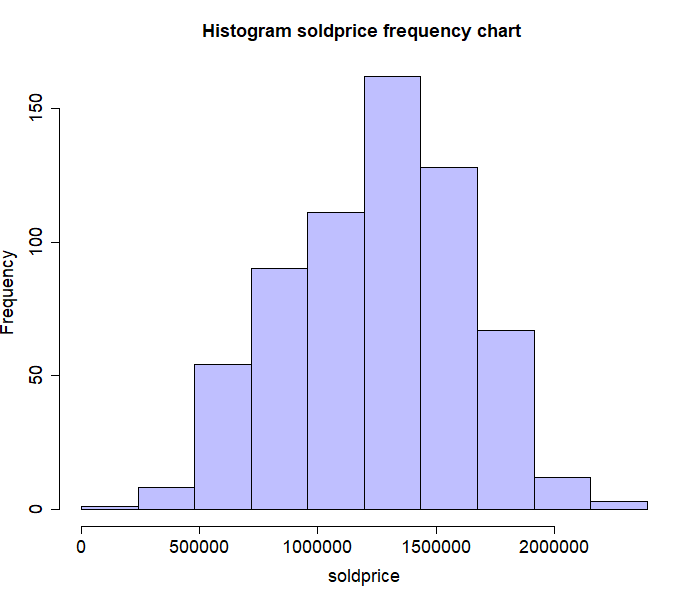


**Figure 9**



**Figure 10**

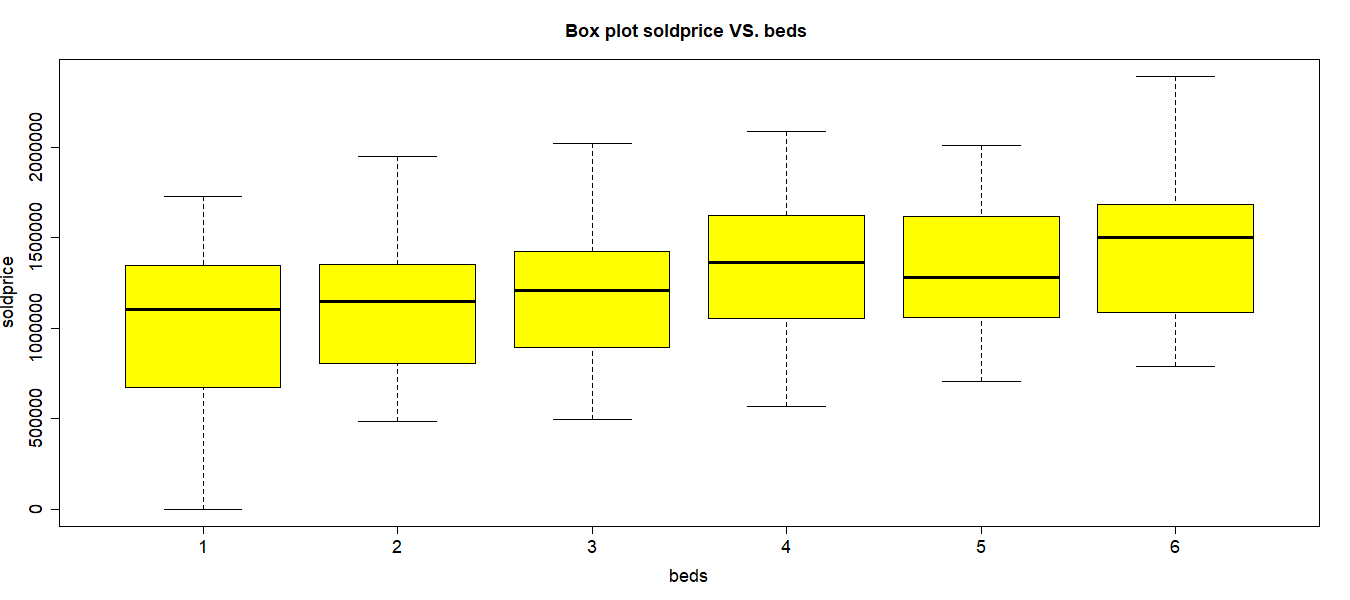
Then, I check the sold price’s distribution if there is normal or abnormal. The histogram shows that it is an evenly distributed (Figure 11).



**Figure 11**

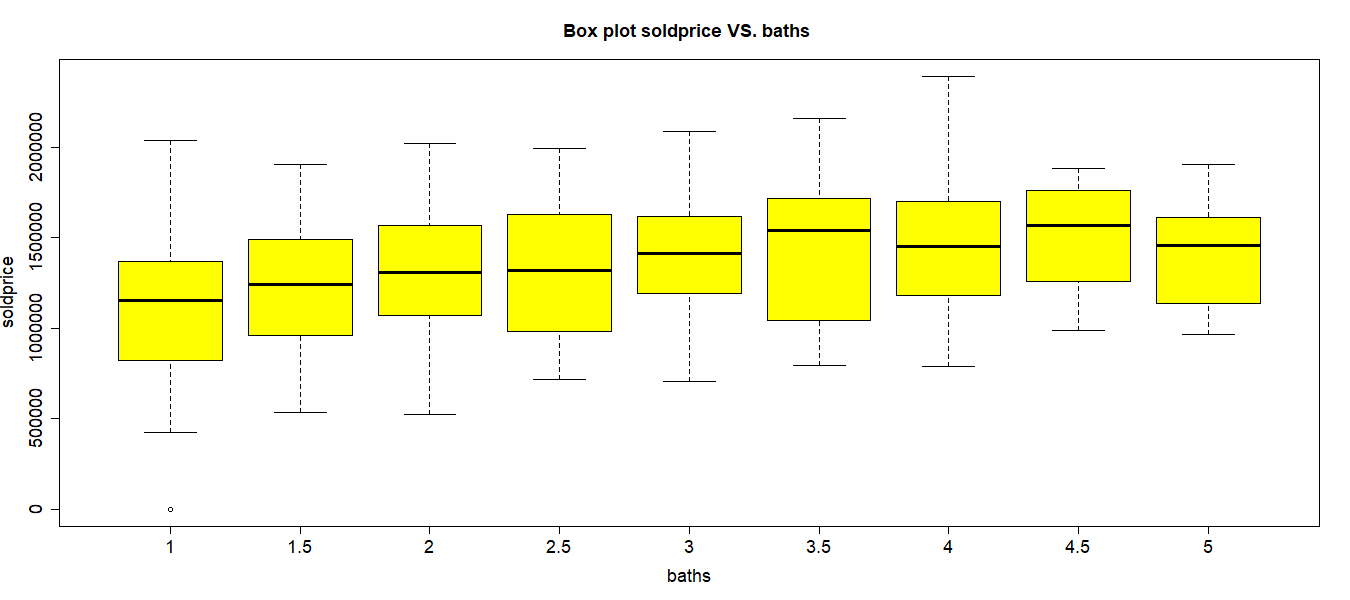
And then, I want to check which character variables can makes higher sold price.

**The “beds” VS. the “sold price”:** slightly increase (Figure 12).



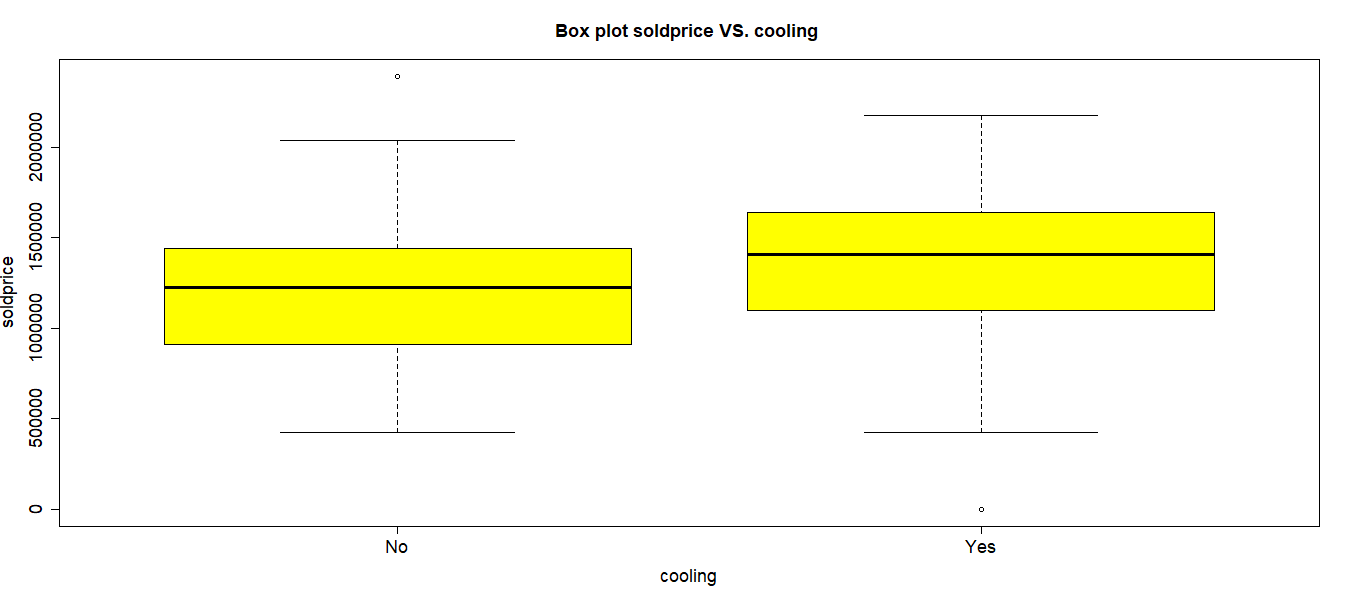
**Figure 12**

**The “baths” VS. the “sold price”:** slightly increase (Figure 13).



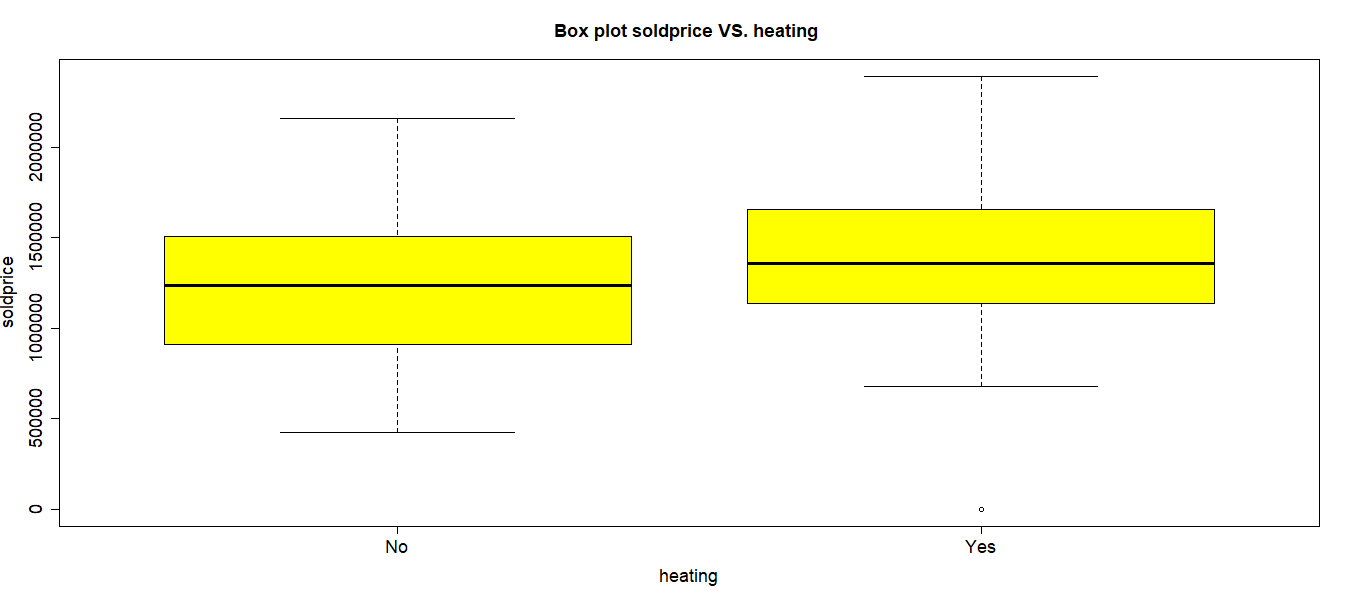
**Figure 13**

**The “cooling” VS. the “sold price”:** “Yes” has a higher sold price (Figure 14).

6

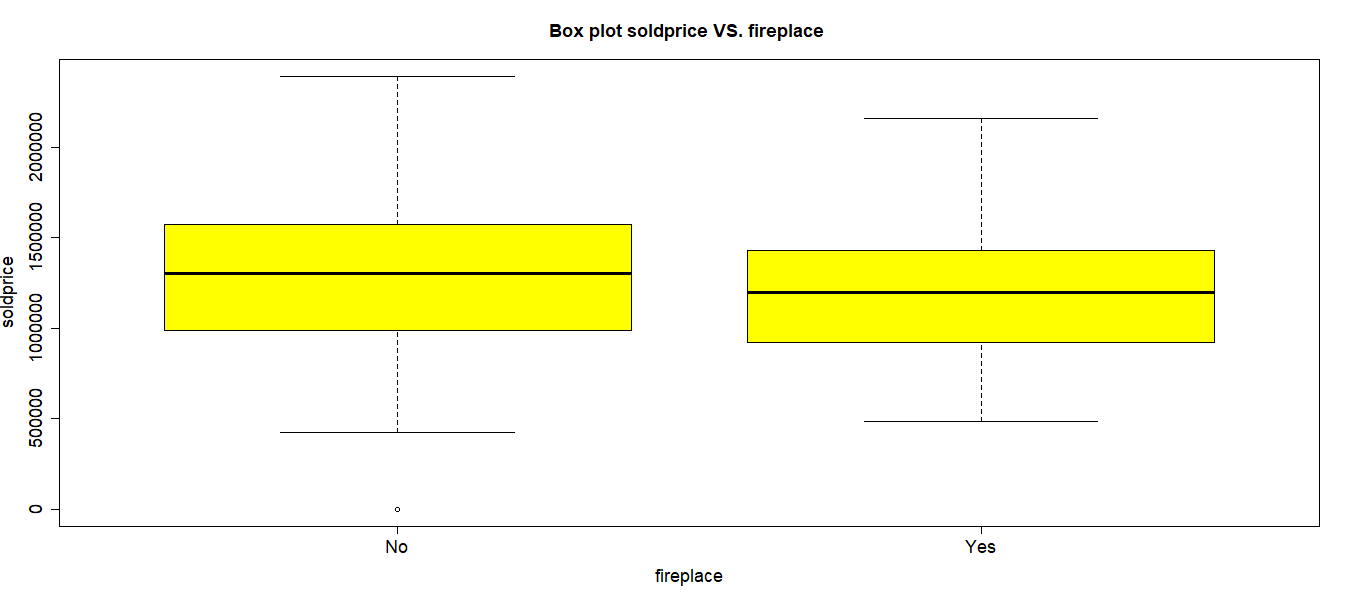
**Figure 14**

**The “heating” VS. the “sold price”:** “Yes” has a slightly higher sold price (Figure 15).



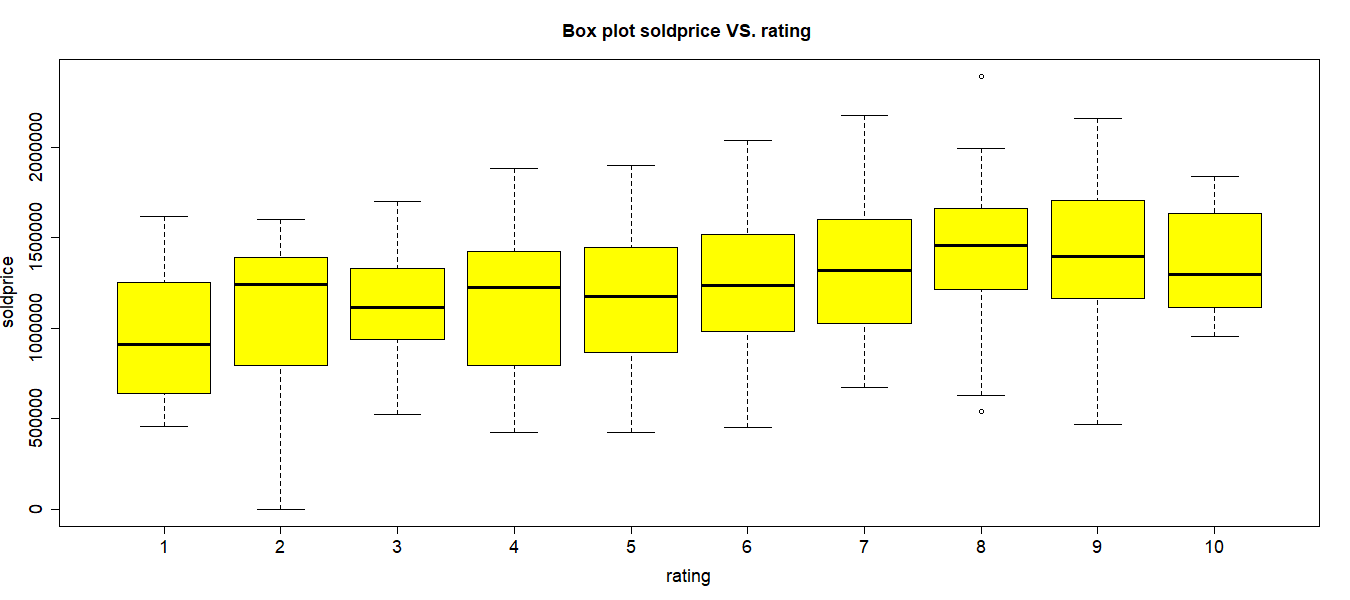
**Figure 15**

**The “fireplace” VS. the “sold price”:** “Yes” has a **lower** sold price (Figure 16).



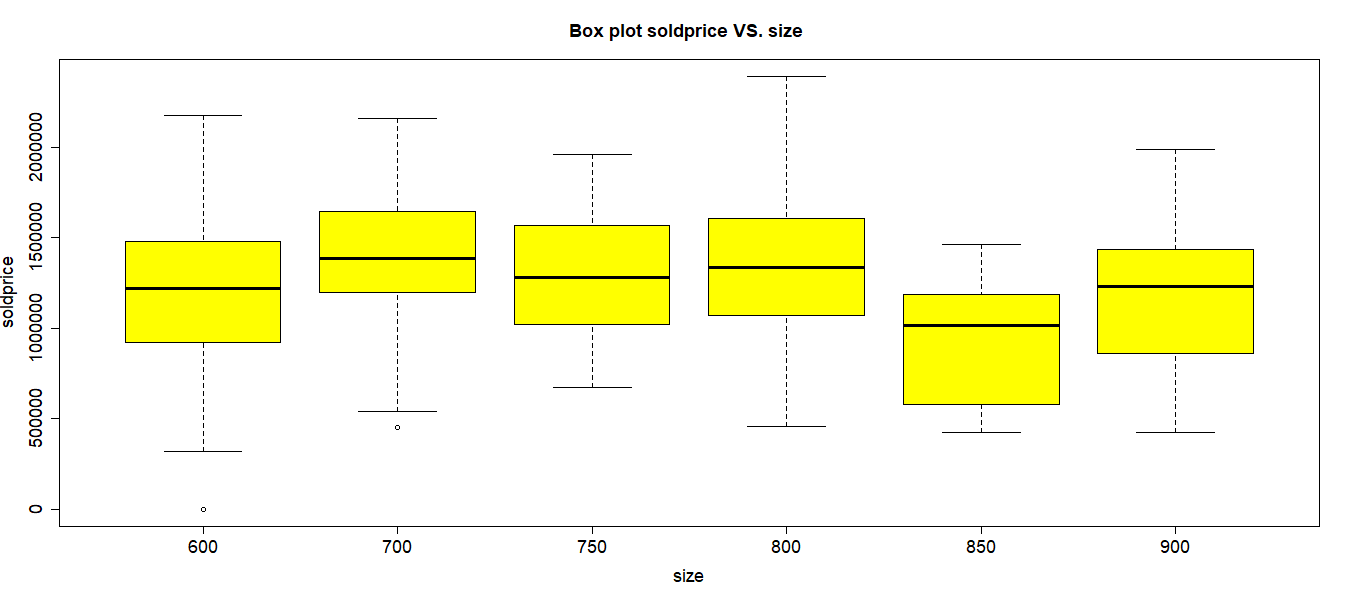
**Figure 16**

**The “rating” VS. the “sold price”:** high rating causes higher sold price (Figure 17).



**Figure 17**

However, the “school size” VS. the “sold price” has not trend (Figure 18).



**Figure 18**

**The “levels” VS. the “sold price”:** No obvious trend (Figure 19).

**Chart, diagram, box and whisker chart

Description automatically generated**

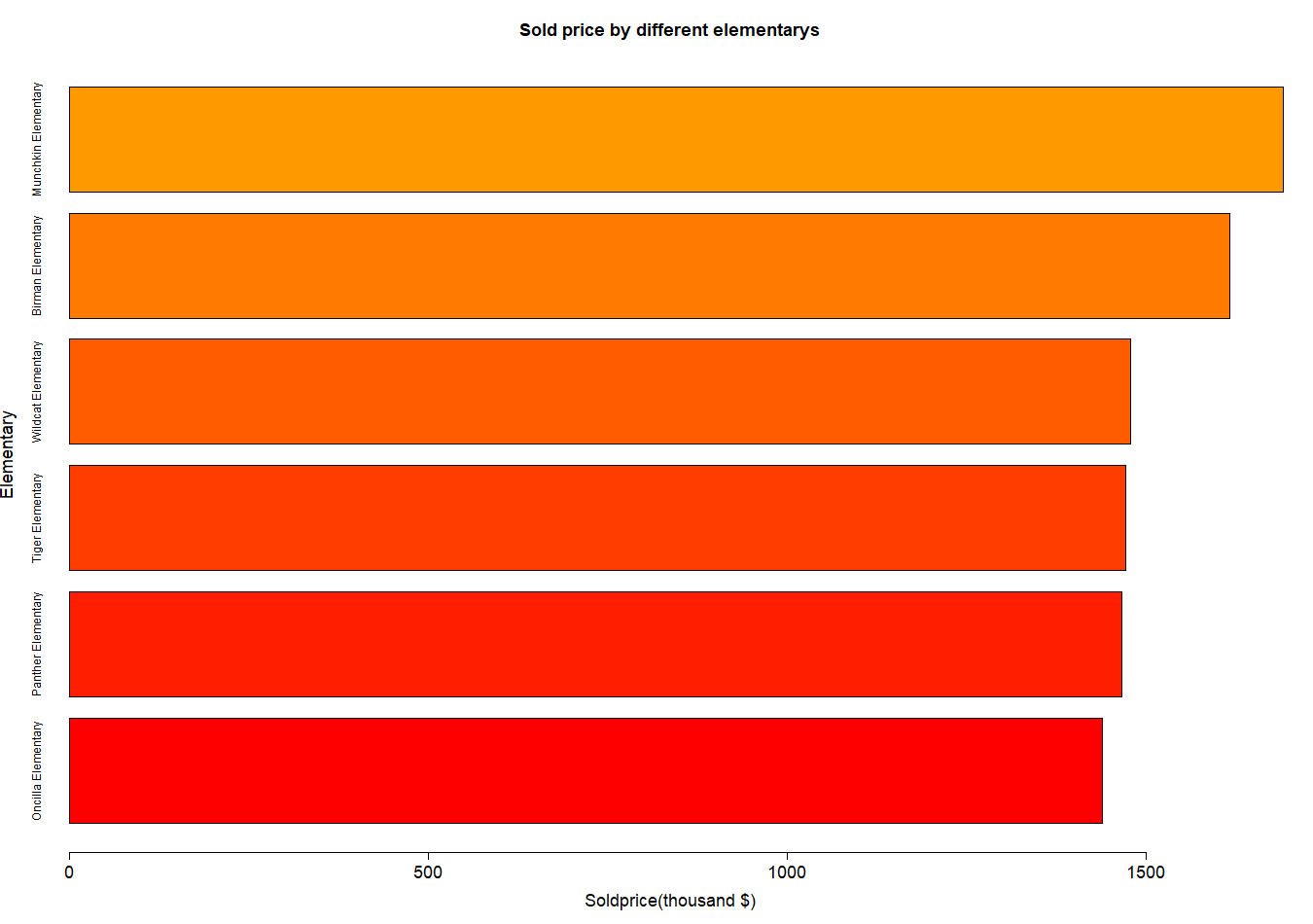
**Figure 19**

If I check by some characters, such as “elementary” (Figure 20),



**Figure 20**

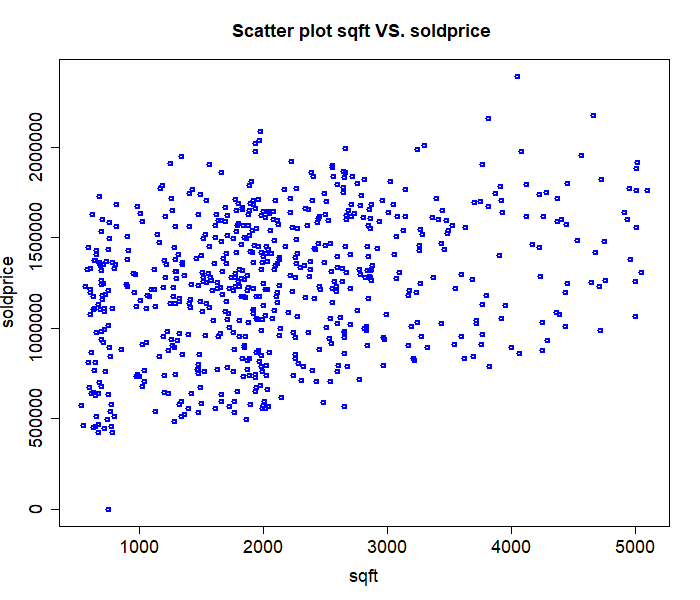
Besides, the two highest sold price of schools was “Munchkin elementary” and “Birman elementary” (Figure 21).



**Figure 21 The top 5 bar plot**

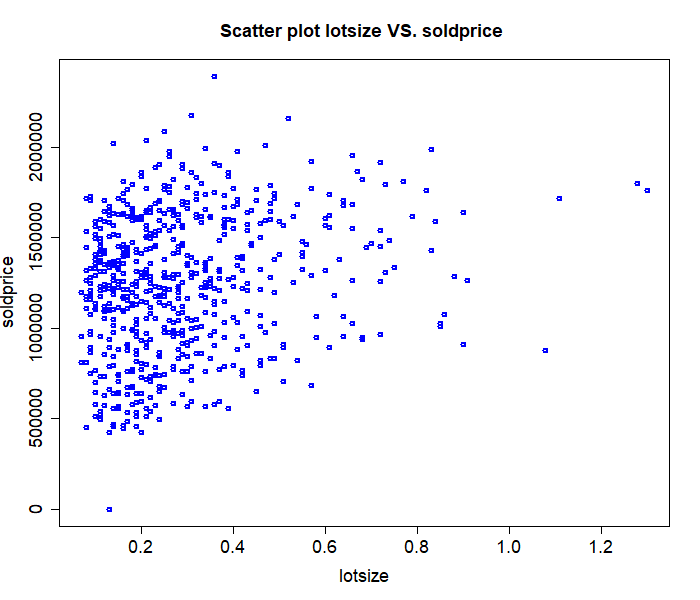
1. **Two-variable visuals**

**The “sqft” VS. the “sold price”**: slightly increase (Figure 22).



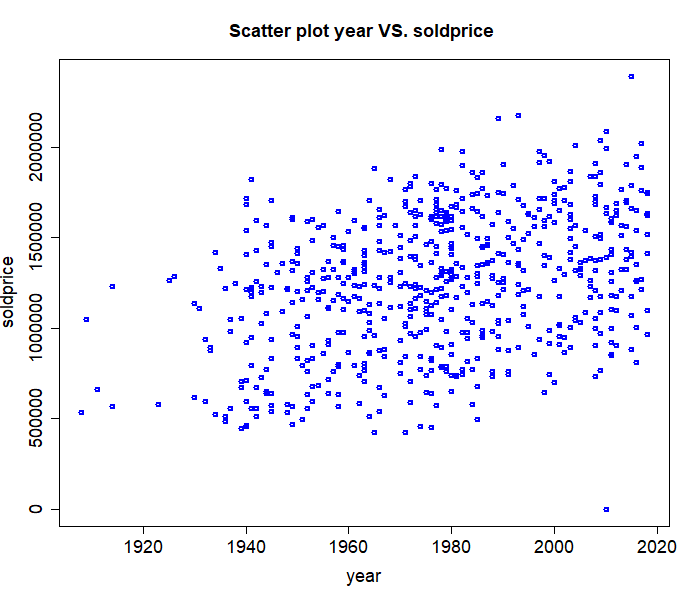
**Figure 22**

**The “lotsize” VS. the “sold price”**: slightly increase (Figure 23).



**Figure 23**

**The “year” VS. the “sold price”**: slightly increase trend (Figure 24).



**Figure 24**

The density plot tells me that distinct types of houses are all evenly distributed (Figure 25). Again, the “condominium” only has 3-row data, so I will miss the result. That is to say, even though the “condominium” has the highest peak, it does not mean anything. Besides, it meets previous expectations that the “multi-family home” and the “single-family home” both have a higher sold price.

Chart, histogram

Description automatically generated

**Figure 25**

Finally, I combined session 3 and session 4, making a note of what I observed in Table2.

|  |  |
| --- | --- |
| **Variables** | **Descriptions (causes a higher sold price)** |
| **type** | multi-family home, single-family home |
| **sqft** | slightly positive |
| **beds** | slightly positive |
| **baths** | slightly positive |
| **sqft** | slightly positive |
| **lotsize** | slightly positive |
| **cooling** | “Yes” has a higher sold price. |
| **Heating** | “Yes” has a slightly higher sold price. |
| **Fireplace** | **“No”** has a higher sold price. |
| **Rating** | higher rate has higher sold price. |
| **Size** | No trend |
| **levels** | “Yes” has a slightly higher sold price. |
| **Elementary** | Top 2: Munchkin and Birman |

**Table 2**

1. **Analysis**

I drew variable-variable scatter plot to check if there has a correlation (Figure 26). Table

Description automatically generated with medium confidence

**Figure 26**

According to session 4 and the above graph, the four variables: “beds”, “baths”, “sqft”, and “lotsize” have highly correlated with one another, but all of their correlation with the sold price are not strong enough (Figure 27~29).

Chart, bubble chart

Description automatically generated

**Figure 27**

Calendar

Description automatically generated

**Figure 28**

A picture containing table

Description automatically generated

**Figure 29**

If I put all numeric variables into a correlation comparison, I will get the same result (Table3).

**Chart, table

Description automatically generated**

**Table 3**

When I fit a linear regression with all numeric variables and binary variables (m1). However, there are some variables that have high P-value (Table 4), and the overall adjusted R-squared is low: 0.305 (<0.5) (Figure 30).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **m1** | **Estimate** | **Std. Error** | **t value** | **Pr(>|t|)** |
| **(Intercept)** | -7818633.89 | 1357866.39 | -5.76 | **0** |
| **beds** | 73420.51 | 24656.97 | 2.98 | **0** |
| **baths** | 23608.83 | 21117.03 | 1.12 | **0.26** |
| **sqft** | -13.44 | 35.38 | -0.38 | **0.7** |
| **lotsize** | 68349.31 | 122007.35 | 0.56 | **0.58** |
| **year** | 4392.97 | 701.36 | 6.26 | **0** |
| **levels** | 20840.28 | 37314.34 | 0.56 | **0.58** |
| **cooling** | 16965.43 | 29689.46 | 0.57 | **0.57** |
| **heating** | 20763.55 | 35462.37 | 0.59 | **0.56** |
| **fireplace** | 3617.48 | 28176.62 | 0.13 | **0.9** |
| **size** | -300.13 | 127.24 | -2.36 | **0.02** |
| **rating** | 39173.97 | 5634.62 | 6.95 | **0** |

**Table 4**

**Table

Description automatically generated**

**Figure 30**

As a result, I remove all the variables with P-value higher than 0.05, then re-run the linear regression (m3) (Table 5). it looks better but R-squared still low, even though the adjusted R-squared is still low: 0.3103 (<0.5).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **m3** | **Estimate** | **Std. Error** | **t value** | **Pr(>|t|)** |
| **(Intercept)** | -8347502.58 | 1107699.4 | -7.54 | **0** |
| **beds** | 85874.05 | 8544.05 | 10.05 | **0** |
| **year** | 4705.04 | 559.91 | 8.4 | **0** |
| **size** | -321.71 | 125.26 | -2.57 | **0.01** |
| **rating** | 39792.95 | 5580.06 | 7.13 | **0** |

**Table 5**

Finally, I put “type” into variables and re-run it (m5) (Table 5). Therefore, the P-value is still ok, and the adjusted R-squared: 0.8103 is good (Figure 31).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **m5 (soldprice)** | **Estimate** | **Std. Error** | **t value** | **Pr(>|t|)** |
| **(Intercept)** | -7979187.59 | 524769.56 | -15.21 | **0** |
| **beds** | 84546.13 | 4067.38 | 20.79 | **0** |
| **year** | 4321.42 | 265.38 | 16.28 | **0** |
| **size** | -263.43 | 59.57 | -4.42 | **0** |
| **rating** | 38418.78 | 2670.86 | 14.38 | **0** |
| **typecondominium** | **-60988.69** | **87056.41** | **-0.7** | **0.48** |
| **typemulti-family home** | 608230.4 | 23004.24 | 26.44 | **0** |
| **typesingle-family home** | 585116.23 | 15193.09 | 38.51 | **0** |
| **typetown house** | 66354.18 | 18541.07 | 3.58 | **0** |

**Table 6**

Text

Description automatically generated

**Figure 31**

If I use the previous independent variables to fit the soldprice^0.5, the result and the adjust R-squared: 0.8303 is also good (m6).

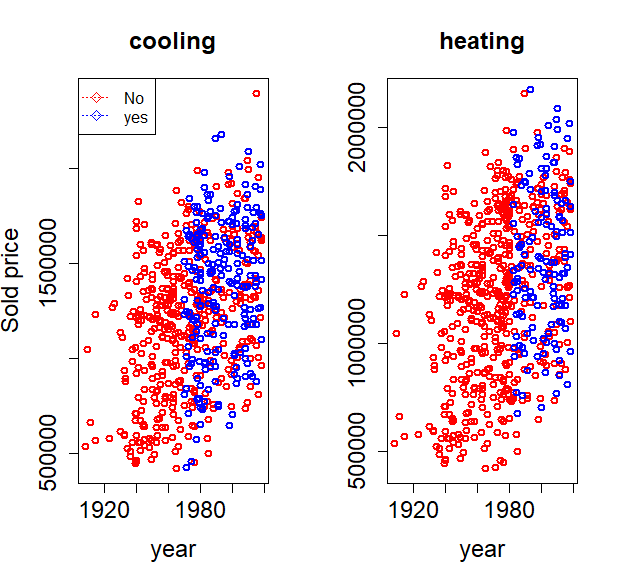
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **m6 (soldprice^0.5)** | **Estimate** | **Std. Error** | **t value** | **Pr(>|t|)** |
| **(Intercept)** | -3181.88 | 262.7 | -12.11 | **0** |
| **beds** | 40.5 | 2.04 | 19.89 | **0** |
| **year** | 2 | 0.13 | 15.02 | **0** |
| **size** | -0.11 | 0.03 | -3.77 | **0** |
| **rating** | 18.26 | 1.34 | 13.65 | **0** |
| **typecondominium** | **-17.01** | **43.58** | **-0.39** | **0.7** |
| **typemulti-family home** | 288.73 | 11.52 | 25.07 | **0** |
| **typesingle-family home** | 281.26 | 7.61 | 36.98 | **0** |
| **typetown house** | 39.61 | 9.28 | 4.27 | **0** |

**Table 7**

By the way, if I remove the independent “year” only and re-run m1, the independent variables “cooling” and “heater” will have a different result (m21) (Table 8). It means “cooling” and “heating” have some multilinearity to “year.” That is to say, I can change “year” to “cooling” + “heater.” If I drew the “cooling” and the “heating” VS. the “year” scatter plot grouped by “Yes” or “No” (Figure 32), we can find that “Yes” data clustered on the right side, meaning that year dominate the “cooling” or the “heating” variable.



**Table 8**



**Figure 32**

Then, I execute the “cluster” code check if I can predict the “type” variable by all the variables I found at m5. If I can predict the result, the data can be called the supervised data.

However, the result (Figure 33~35) tells me my dream is only a dream. The cluster could hardly return the result I want them to. Hence, I consider that is because in the scatter plot of soldprice VS. beds (Figure 33.a), soldprice VS. year (Figure 34.a), beds VS. year (Figure 35.a) with several types of houses, there are no cluster formed. As a result, the prediction of cluster is impossible to be found.

**Graphical user interface, text, application

Description automatically generated**

**Figure 33 (a) Figure 33 (b)**

**Chart, scatter chart

Description automatically generated**

**Figure 34 (a) Figure 34 (b)**

**Chart

Description automatically generated**

**Figure 35 (a) Figure 35 (b)**

When I draw a 3D plot between those three variables, the result is the same (Figure 36). Hence, it is difficult to predict a model just by the eyes.

**Chart, scatter chart

Description automatically generated**

**Figure 36**

1. **Sensitivity Analysis**

In the previous session, I remove all the missing data. But in this chapter, I try to fill them in with different values.

I arranged Table 9 again and noted the columns that have the missing data only (Table 4). To remind, there are two different types of missing data, binary and int (or float). So, in the binary data, I tried to replace the empty data with the mode (median), while in the int or float column, I gave them the predicted data by calculating a linear regression model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Column names** | **Character** | **What I should do** | **count** |
| 1 | sqft | int | Remove or replace NA's | 2 |
| 2 | lotsize | float | Remove or replace NA's | 20 |
| 3 | levels | binary | Remove or replace “?” | 6 |
| 4 | cooling | binary | Remove or replace empty | 7 |
| 5 | heating | binary | Remove or replace empty | 7 |
| 6 | fireplace | binary | Remove or replace empty | 6 |

**Table 9**

1. Replacing with the mode (median) (the row 3~6):

housingcleanpred$levels[which(housingcleanpred$levels =='?')] <- "1"

housingcleanpred$levels <- as.character(housingcleanpred$levels)

housingcleanpred$levels <- as.factor(housingcleanpred$levels)

housingcleanpred$cooling <- as.character(housingcleanpred$cooling)

housingcleanpred$cooling[which(housingcleanpred$cooling =='')] <- 'No'

housingcleanpred$cooling <- as.factor(housingcleanpred$cooling)

housingcleanpred$heating <- as.character(housingcleanpred$heating)

housingcleanpred$heating[which(housingcleanpred$heating =='')] <- 'No'

housingcleanpred$heating <- as.factor(housingcleanpred$heating)

housingcleanpred$fireplace <- as.character(housingcleanpred$fireplace)

housingcleanpred$fireplace[which(housingcleanpred$fireplace =='')] <- 'No'

housingcleanpred$fireplace <- as.factor(housingcleanpred$fireplace)

1. Replacing with the predicted value:

I focus on different variables to predict their own linear regression, then use the model to predict the value and fill in the empty. I first executed a linear regression model by sqft VS. other variables (m1\_ReplaceTOmode), then picked up the variables with P-values that are lower than 0.05 and re-run linear regression (m2\_ReplaceTOmode). Finally, I predict the missing “sqft” data by using m2\_ReplaceTOmode.

Sqft:

m1\_ReplaceTOmode <- lm (sqft~beds + baths + lotsize + levels + cooling + heating + fireplace + soldprice + size + rating + type, data = housingcleanpred)

m2\_ReplaceTOmode <- lm (sqft~beds + baths + lotsize + levels, data = housingcleanpred)

summary(m2\_ReplaceTOmode)

housingcleanpred$sqft[is.na(housingcleanpred$sqft)] <- predict(m2\_ReplaceTOmode, list(

beds = housingcleanpred$beds[is.na(housingcleanpred$sqft)],

baths = housingcleanpred$baths[is.na(housingcleanpred$sqft)],

lotsize = housingcleanpred$lotsize[is.na(housingcleanpred$sqft)],

levels = housingcleanpred$levels[is.na(housingcleanpred$sqft)]

))

In the missing “lotszie” data, I use the same way to create a linear regression model (m4\_ReplaceTOmode), then use them to create the data I want.

Lotsize:

m3\_ReplaceTOmode <- lm (lotsize~beds + baths + sqft + lotsize + levels + cooling + heating + fireplace + soldprice + size + rating + type, data = housingcleanpred)

m4\_ReplaceTOmode <- lm (lotsize~beds + sqft + levels, data = housingcleanpred)

summary(m4\_ReplaceTOmode)

housingcleanpred$lotsize[is.na(housingcleanpred$lotsize)] <- predict(m4\_ReplaceTOmode, list(

beds = housingcleanpred$beds[is.na(housingcleanpred$lotsize)],

sqft = housingcleanpred$sqft[is.na(housingcleanpred$lotsize)],

levels = housingcleanpred$levels[is.na(housingcleanpred$lotsize)]

))

As a result, when I check the “summary” code, the result is rational (Figure 37).

A screenshot of a computer

Description automatically generated with low confidence

**Figure 37**

When I executed the code to check a correlation (Table 10), the correlation tells me it does not change a lot. When compared to Table 3.



**Table 10**

If I check the correlation bias (session 5’ correlation: Table 3 – the above’s correlation: Table 10), it seems the correlation does not have an apparent shift (Table 11).



**Table 11 Bias correlation = Table 10 – Table 3**

When I check the “lotsize” VS. the “soldprice” scatter plot (Figure 38), even though I added 44 new data into the scatter plot (Figure 23), the result looks similar, and when I check the related correlation, it is almost the same (Table 12).

Chart, scatter chart

Description automatically generated

**Figure 23 Figure 38**

|  |  |  |
| --- | --- | --- |
| **Original** | **heating** | **soldprice** |
| **heating** | 1.00 | 0.20 |
| **soldprice** | 0.20 | 1.00 |
| **New** | **heating** | **soldprice** |
| **heating** | 1.00 | 0.19 |
| **soldprice** | 0.19 | 1.00 |

**Table 12**

Therefore, when I compare the linear regression between the two different methods of managing the missing data, the original one (m5) and the new one (m5\_after) are similar. That is to say, there is the same order of intercept and value in different variables. Besides, the order of the P-value is similar, too. The result tells me that the new treatment will not impact my final conclusion (Table 13).



**Table 13**

In conclusion, there are several things I found in this final assignment:

* In the “type” variable, the “multi-family home” and “single-family home” can be sold at a higher price.
* House that has a fireplace will decrease its sold price (The “soldprice” with the “fireplace” is “No” > the “soldprice” with the “fireplace” is “Yes.”
* The “year” variable and the “cooling,” and “heating” is multilinearity each other. The scatter plot tells me that since the 1970s, some houses have started to install heating and cooling system. Therefore, when I run the linear regression, the “cooling and the “heating” can be replaced by the “year.”
* There are two linear regression lines I built:

1. Soldprice = -7979187.59 + 84546.13\*beds + 4321.42\*year - 263.43\*size + 38418.78\*rating - 60988.69\*type(condominium) + 608230.4\*type(multi-family home) + 585116.23\*type(single-family home) + 66354.18\*type(town house)
2. (Soldprice)^0.5 = -3181.88 + 40.5beds + 2\*year - 0.11\*size + 18.26\*rating - 17.01\*type(condominium) + 288.73\*type(multi-family home) + 281.26\*type(single-family home) +39.61\*type(town house)

* There are two treatments I used toward the missing data. Both of them are similar and have equivalent results.

~The End~