Analyzing Housing Datasets: another way to recognize people's buying behavior

Rian Dwi Putra
School of Computing

National College of Ireland

Dublin, Ireland

x22108637@student.ncirl.ie

I. MOTIVATION

The house is one of the physical infrastructures that can support survival of humans in their social status in society. House is a place where the process of education, self-maturation, socialize with the environment so that humans themselves can become a person who has good personality and behavior, and functions as a place to live or stay and a means of family development.

Whereas housing is a group of houses that function as an environment where living or residential environment that is equipped with infrastructure and facilities environment.

Housing demand system that occurs in society is always related to several things that must be understood as follows:

1. Housing needs are basic and basic needs that are objective and the same for everyone. Where the notion of 'need' here is related to the problem of meeting the basic human needs of the house as a place to live and shelter.

2. Demand for housing is more subjective, depending on taste and level of economic ability. Because everyone has different tastes and economic abilities. With these differences, there will be various variations of housing needs.

Then the demand for housing will be influenced by factors such as the social, economic and cultural conditions of the community itself.

From the description above, that the influence of the concept of housing, location and the price adjustment to the purchase decision is not yet known with certainty. For this reason, research will be conducted on home sales prices and characteristics for Seattle and King County, WA (May 2014 - 2015).

The reason the author took this study is to find out how far where the effect of housing concept, facilities and price adjustment affects purchasing decisions on the housing estate.

II. RESEARCH QUESTIONS

Based on dataset given, following are the questions of this study need to be answered:
1. What variable has the most impact to the sales of Housing?
2. What variable has the least impact to the sales of Housing?
3.Mention every dependent variable for Housing Dataset Scenario.
4.With which categorical variable, the price column has a close relationship?

III. DATA SOURCE DESCRIPTION

The experiment presented in this study focus on the housing dataset which is from:

https://geodacenter.github.io/data-and-lab/KingCounty-HouseSales2015/

Table 1. Table Format for Housing Dataset

Variable	Description
ID	Number of Identification
Date	Date which this house has been sold
Price	Price
Bedrooms	How many bedrooms in this house
Bathrooms	How many bathrooms in this house
Sqft_liv	**Living area size**
Sqft_lot	Lot size
Floors	**Floors which had by this house**
Waterfront	a part of house that borders a body of water.
view	How is the rating of the house was
condition	the state of something with regard to its appearance, quality, or working order.
grade	a particular level of rank, quality, proficiency, intensity, or value.
sqft_above	How long it was from above the ground
sqft_basmt	How long it was from below the ground
yr_built	The time this house was built
yr_renov	The time this house was renovated
zipcode	zip code
lat	Latitude

long	Longitude
squft_liv15	
squft_lot15	
Shape_leng	Polygon length
Shape_Area	Polygon area

IV. DATA PREPROCESSING

Data preprocessing is one step in the process of data mining and data analysis. In this process, raw data is retrieved and prepared into a format that computers can understand and analyze. This needs to be done because raw data in the real world, whether in the form of text, images, or videos, is messy. So, it will damage the computer to process it. This process will transform data into a format that is easier and more effective to process. In addition, this study also cannot process raw data, so this process is particularly important to do to simplify the next process, namely data analysis. Preprocessing itself involves data validation and imputation. The validation is to assess the level of completeness and accuracy of the filtered data.

1. Reading Data

How to read a file with a .csv extension on R is to use the read.csv command, followed by the folder address (path).

```
mydata=read.csv("kc_house_data.csv")

head(mydata,10)
           id             date   price bedrooms bathrooms sqft_living sqft_lot floors
 7129300520 20141013T000000  221900        3      1.00        1180     5650      1
 6414100192 20141209T000000  538000        3      2.25        2570     7242      2
 5631500400 20150225T000000  180000        2      1.00         770    10000      1
 2487200875 20141209T000000  604000        4      3.00        1960     5000      1
 1954400510 20150218T000000  510000        3      2.00        1680     8080      1
 7237550310 20140512T000000 1225000        4      4.50        5420   101930      1
 1321400060 20140627T000000  257500        3      2.25        1715     6819      2
 2008000270 20150115T000000  291850        3      1.50        1060     9711      1
 2414600126 20150415T000000  229500        3      1.00        1780     7470      1
 3793500160 20150312T000000  323000        3      2.50        1890     6560      2
 waterfront view condition grade sqft_above sqft_basement yr_built yr_renovated zipcode
          0    0         3     7       1180             0     1955            0   98178
          0    0         3     7       2170           400     1951         1991   98125
          0    0         3     6        770             0     1933            0   98028
          0    0         5     7       1050           910     1965            0   98136
          0    0         3     8       1680             0     1987            0   98074
          0    0         3    11       3890          1530     2001            0   98053
          0    0         3     7       1715             0     1995            0   98003
          0    0         3     7       1060             0     1963            0   98198
          0    0         3     7       1050           730     1960            0   98146
          0    0         3     7       1890             0     2003            0   98038
     lat     long sqft_living15 sqft_lot15
 47.5112 -122.257          1340       5650
 47.7210 -122.319          1690       7639
 47.7379 -122.233          2720       8062
 47.5208 -122.393          1360       5000
 47.6168 -122.045          1800       7503
 47.6561 -122.005          4760     101930
 47.3097 -122.327          2238       6819
 47.4095 -122.315          1650       9711
```

Figure 1. Reading data

2. Viewing the structure of Data.

Viewing the data structure can be done using str() function

```
> str(mydata)
'data.frame':   21613 obs. of  21 variables:
 $ id           : num  7.13e+09 6.41e+09 5.63e+09 2.49e+09 1.95e+09 ...
 $ date         : chr  "20141013T000000" "20141209T000000" "20150225T000000" "20141209T000000" ...
 $ price        : num  221900 538000 180000 604000 510000 ...
 $ bedrooms     : int  3 3 2 4 3 4 3 3 3 3 ...
 $ bathrooms    : num  1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...
 $ sqft_living  : int  1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...
 $ sqft_lot     : int  5650 7242 10000 5000 8080 101930 6819 9711 7470 6560 ...
 $ floors       : num  1 2 1 1 1 1 2 1 1 2 ...
 $ waterfront   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ view         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ condition    : int  3 3 3 5 3 3 3 3 3 3 ...
 $ grade        : int  7 7 6 7 8 11 7 7 7 7 ...
 $ sqft_above   : int  1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...
 $ sqft_basement: int  0 400 0 910 0 1530 0 0 730 0 ...
 $ yr_built     : int  1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...
 $ yr_renovated : int  0 1991 0 0 0 0 0 0 0 0 ...
 $ zipcode      : int  98178 98125 98028 98136 98074 98053 98003 98198 98146 98038 ...
 $ lat          : num  47.5 47.7 47.7 47.5 47.6 ...
 $ long         : num  -122 -122 -122 -122 -122 ...
 $ sqft_living15: int  1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ...
 $ sqft_lot15   : int  5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 ...
```

Figure 2. Structure of data

3. Viewing the summary of Data

Viewing the summary can be done by using summary() function.

```
> summary(mydata)
       id                 date                price             bedrooms        bathrooms      
 Min.   :1.000e+06   Length:21613       Min.   :  75000   Min.   : 0.000   Min.   :0.000  
 1st Qu.:2.123e+09   Class :character   1st Qu.: 321950   1st Qu.: 3.000   1st Qu.:1.750  
 Median :3.905e+09   Mode  :character   Median : 450000   Median : 3.000   Median :2.250  
 Mean   :4.580e+09                      Mean   : 540088   Mean   : 3.371   Mean   :2.115  
 3rd Qu.:7.309e+09                      3rd Qu.: 645000   3rd Qu.: 4.000   3rd Qu.:2.500  
 Max.   :9.900e+09                      Max.   :7700000   Max.   :33.000   Max.   :8.000  
   sqft_living       sqft_lot           floors        waterfront           view          condition   
 Min.   :  290   Min.   :    520   Min.   :1.000   Min.   :0.000000   Min.   :0.0000   Min.   :1.000  
 1st Qu.: 1427   1st Qu.:   5040   1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000  
 Median : 1910   Median :   7618   Median :1.500   Median :0.000000   Median :0.0000   Median :3.000  
 Mean   : 2080   Mean   :  15107   Mean   :1.494   Mean   :0.007542   Mean   :0.2343   Mean   :3.409  
 3rd Qu.: 2550   3rd Qu.:  10688   3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000  
 Max.   :13540   Max.   :1651359   Max.   :3.500   Max.   :1.000000   Max.   :4.0000   Max.   :5.000  
     grade          sqft_above    sqft_basement       yr_built     yr_renovated       zipcode     
 Min.   : 1.000   Min.   : 290   Min.   :   0.0   Min.   :1900   Min.   :   0.0   Min.   :98001  
 1st Qu.: 7.000   1st Qu.:1190   1st Qu.:   0.0   1st Qu.:1951   1st Qu.:   0.0   1st Qu.:98033  
 Median : 7.000   Median :1560   Median :   0.0   Median :1975   Median :   0.0   Median :98065  
 Mean   : 7.657   Mean   :1788   Mean   : 291.5   Mean   :1971   Mean   :  84.4   Mean   :98078  
 3rd Qu.: 8.000   3rd Qu.:2210   3rd Qu.: 560.0   3rd Qu.:1997   3rd Qu.:   0.0   3rd Qu.:98118  
 Max.   :13.000   Max.   :9410   Max.   :4820.0   Max.   :2015   Max.   :2015.0   Max.   :98199  
      lat             long        sqft_living15    sqft_lot15     
 Min.   :47.16   Min.   :-122.5   Min.   : 399   Min.   :   651  
 1st Qu.:47.47   1st Qu.:-122.3   1st Qu.:1490   1st Qu.:  5100  
 Median :47.57   Median :-122.2   Median :1840   Median :  7620  
 Mean   :47.56   Mean   :-122.2   Mean   :1987   Mean   : 12768  
 3rd Qu.:47.68   3rd Qu.:-122.1   3rd Qu.:2360   3rd Qu.: 10083  
 Max.   :47.78   Max.   :-121.3   Max.   :6210   Max.   :871200  
```

Figure 3. Summary of data

4. Check for missing values in Data

```
> NA_values=data.frame(no_of_na_values=colSums(is.na(mydata)))
>
> head(NA_values,21)
               no_of_na_values
id                           0
date                         0
price                        0
bedrooms                     0
bathrooms                    0
sqft_living                  0
sqft_lot                     0
floors                       0
waterfront                   0
view                         0
condition                    0
grade                        0
sqft_above                   0
sqft_basement                0
yr_built                     0
yr_renovated                 0
zipcode                      0
lat                          0
long                         0
sqft_living15                0
sqft_lot15                   0
>
```

Figure 4. missing value

V. MODEL BUILDING PROCESS

1. Dividing Data into Train and Test Set

When we need to perform analysis in R, we often divide the dataset into two parts: train data and test data. Train data is used to build the model, while test data is used to test the model, we build to see how accurate the model we build is. Some literature uses the concept of 70:30 or 80:20 for the composition of train data and test data.

```
set.seed(123)
sample = sample.split(mydata,SplitRatio = 0.8)
train_data =subset(mydata,sample ==TRUE)
test_data=subset(mydata, sample==FALSE)
```

Figure 5. Splitting data

2. Exploratory Data Analysis

Exploratory Data Analysis is an initial investigative test process that aims to identify patterns, find anomalies, test hypotheses and check assumptions. By doing EDA, users will be greatly assisted in detecting errors from the start, being able to identify outliers, knowing the relationships between data and being able to explore important factors from the data. The EDA process is

very useful in the process of statistical analysis . However, if necessary, exploratory data analysis is very helpful in analyzing and discovering about the properties of data which can later be useful in selecting the right statistical model.

Thus, in exploratory data analysis, it is the nature of the observed data that will determine the appropriate statistical analysis model (or improvement of the planned analysis)..

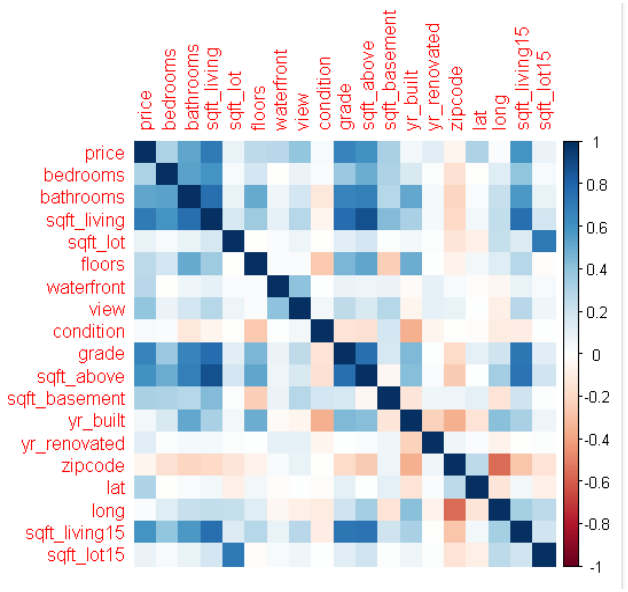

Figure 6. Corrplot

Based on the corrplot above, price is interrelated with bedroom, sqft_above, bathroom, sqft_living 15, Sqft_living, sqft_above , view , grade, sqft_basement, lat. So, we can draw scatter plots (determining the relationship between the variables with price) and boxplot (determining relationship between price and another categorical variables)

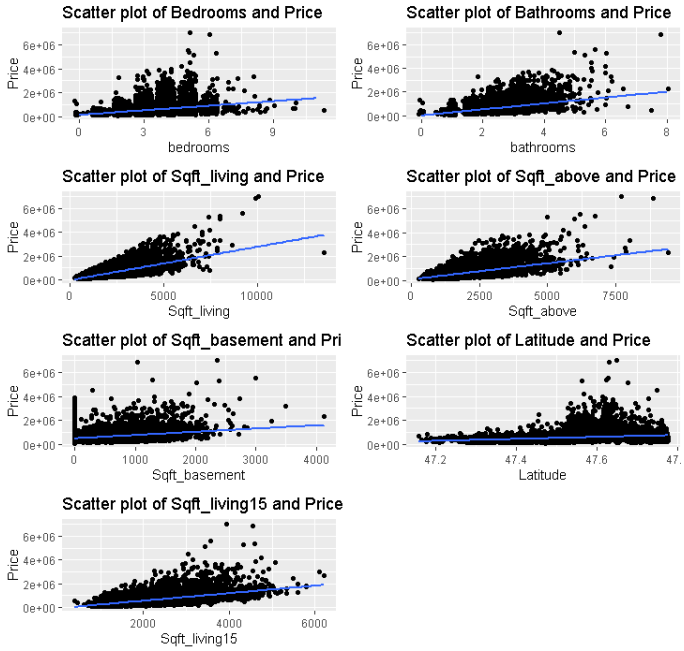

Figure 7. Scatter plot

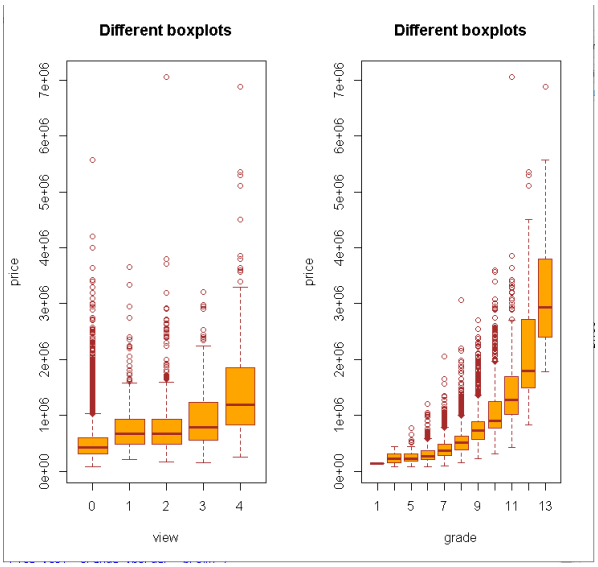

Figure 8. Box plot

The Scatter Diagram functions to test how strong the relationship is between 2 (two) variables and determines the type of relationship of the 2 (two) variables whether it is a positive relationship, a negative relationship or no relationship at all. A scatterplot graph or which has another name as a scatterplot graph is a graphic diagram that is built from two X and Y axes (X variables and Y variables). The values of this pair of variables are described as dots.

Furthermore, we can draw ggplot, allows you to create graphs that simultaneously represent both univariate and multivariate numeric and categorical data. The groupings can be represented by color, symbol, point size and thickness.

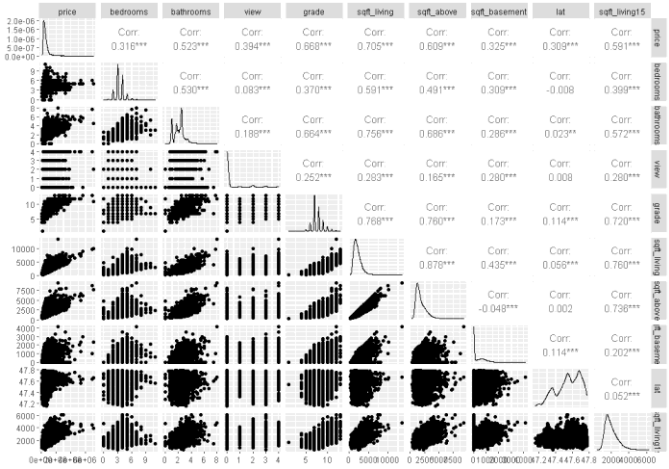

Figure 9. gg plot

And checking the outliers is the last step for EDA

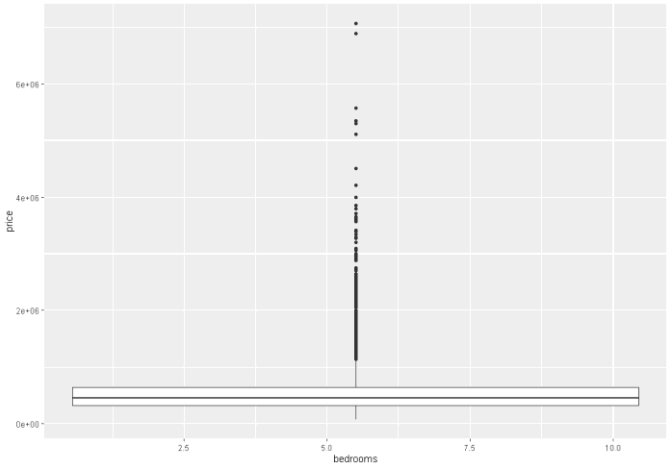

Figure 10. checking for outliers

Outliers are data points whose values are significantly different from a certain population. While this definition may seem simple, determining which data points are outliers is quite subjective, depending on the study and the extent of the data collected. For better understanding, this study will compare the data with and without outlier.

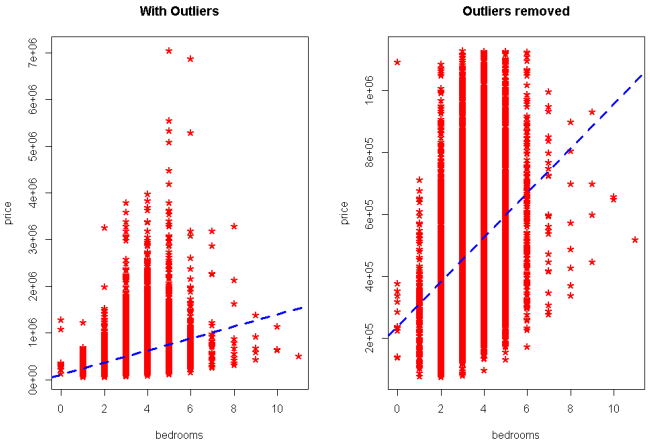

Figure 11. with and without outliers

3. Modeling

All the variables will be chosen for the full model based on the corrplot. And to determine the relationship, a linear model will be a good fit.

```
Call:
lm(formula = price ~ bedrooms + bathrooms + sqft_living + sqft_lot +
    floors + waterfront + view + condition + grade + sqft_above +
    sqft_basement + yr_built + yr_renovated + zipcode + lat +
    long + sqft_living15 + sqft_lot15, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1297877  -100037    -8975    78030  4079859

Coefficients: (1 not defined because of singularities)
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.326e+06  3.395e+06   1.569  0.11667
bedrooms      -3.999e+04  2.255e+03 -17.736  < 2e-16 ***
bathrooms      4.062e+04  3.753e+03  10.824  < 2e-16 ***
sqft_living    1.540e+02  5.082e+00  30.291  < 2e-16 ***
sqft_lot       1.690e-01  5.617e-02   3.009  0.00263 **
floors         8.448e+03  4.163e+03   2.030  0.04241 *
waterfront     5.952e+05  1.941e+04  30.656  < 2e-16 ***
view           4.872e+04  2.450e+03  19.886  < 2e-16 ***
condition      2.720e+04  2.709e+03  10.038  < 2e-16 ***
grade          9.395e+04  2.496e+03  37.647  < 2e-16 ***
sqft_above     3.167e+01  5.037e+00   6.287 3.31e-10 ***
sqft_basement         NA         NA      NA       NA
yr_built      -2.634e+03  8.387e+01 -31.405  < 2e-16 ***
yr_renovated   1.794e+01  4.217e+00   4.254 2.11e-05 ***
zipcode       -5.617e+02  3.820e+01 -14.706  < 2e-16 ***
lat            6.033e+05  1.234e+04  48.896  < 2e-16 ***
long          -2.094e+05  1.526e+04 -13.721  < 2e-16 ***
sqft_living15  2.435e+01  3.987e+00   6.109 1.03e-09 ***
sqft_lot15    -4.302e-01  8.432e-02  -5.101 3.41e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 202500 on 16450 degrees of freedom
Multiple R-squared:  0.7007,    Adjusted R-squared:  0.7004
F-statistic:  2265 on 17 and 16450 DF,  p-value: < 2.2e-16
```

Figure 12. building the model

Because the relationships between those variables are quite good in the value of R-squared (0.7004), so this study decide to continue with this model, excluding the sqft_lot, floors and sqft_basement variable.

VII. Diagnostic and Assumptions Checking

One method for assumptions checking is a measure of influence called Cook's Distance, which is formula to detect the magnitude of the effect on all regression coefficient estimates. By doing this, we can plot the Cook's Distance.

```
cooksd <- cooks.distance(model)
mean(cooksd)
```

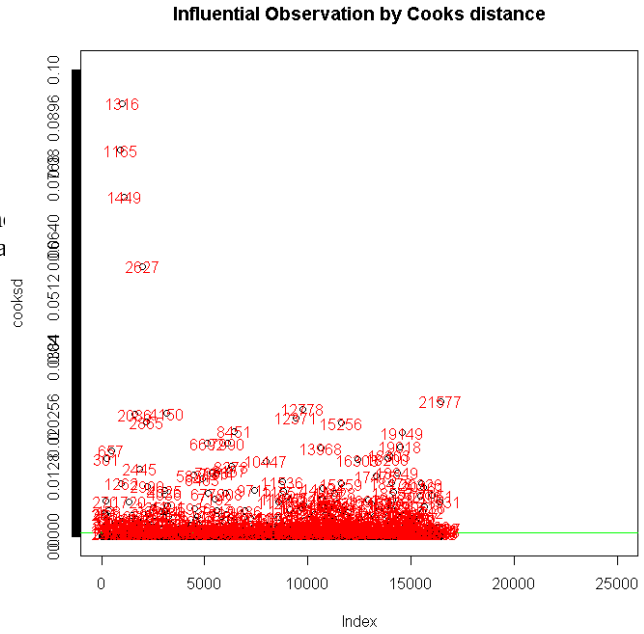

Figure 13. Cook's distance

By using Cook's Distance, we will know the point which will influence the most to our data.

```
influential <- as.numeric(names(cooksd)[(cooksd > 4*mean(cooksd, na.rm=T))])
head(train_data[influential, ])
```

Figure 14. the most influential point

Then, we will remove the influential outliers.

```
> influential_data=train_data[influential, ]
>
> influencial_outliers=inner_join(outliers_data,influential_data)
Joining, by = c("id", "date", "price", "bedrooms", "bathrooms", "sqft_living", "sqft_lot", "floors",
"waterfront", "view", "condition", "grade", "sqft_above", "sqft_basement", "yr_built", "yr_renovated",
"zipcode", "lat", "long", "sqft_living15", "sqft_lot15", "sale_date_year", "age", "reno")
>
```

Figure 15. removal of the influential point

Now, we have observation that not only works as outliers but also as influential data, so we need to keep these observations.

VI. MODEL SUMMARY

Finally, we will compare the accuracy before and after the model building process.

```
> pred=model$fitted.values
>
> tally_table=data.frame(actual=train_data$price, predicted=pred)
>
> mape=mean(abs(tally_table$actual-tally_table$predicted)/tally_table$actua
> accuracy=1-mape
> accuracy
[1] 0.7428314
```

Figure 16. accuracy before model building process

```
> pred=model2$fitted.values
>
> tally_table=data.frame(actual=train_data2$price, predicted=pred)
>
> mape=mean(abs(tally_table$actual-tally_table$predicted)/tally_table$actual
> accuracy=1-mape
> accuracy
[1] 0.7980394
>
```

Figure 17. accuracy after model building process

Based on the accuracy measured, we can conclude that our model can predict price until 79%.

And answering the research question, we can recapitulate that :

1. What variable has the most impact to the sales of Housing?

All of variable except sqft_lot, floors and sqft_basement. Because without these variables, we can get the optimum the value of R-squared (0.7004).

2. What variable has the least impact to the sales of Housing?

Similar to question number 1, sqft_lot, floors and sqft_basement have the least impact to the model. So, we can exclude it from the model.

3.Mention every dependent variable for Housing Dataset Scenario.

,sqft_living, sqft_lot, floors, view, condition, grade, sqft_above, sqft_basement, floors, waterfront yr_built, yr_renovated, zipcode, lat, long, sqft_living15, sqft_lot15, bedrooms, bathrooms are dependant variable to the price (as the independent one).

4.With which categorical variable, the price column has a close relationship?

Based on previous analysis, view and grade are the categorical variable which has closest relationship to the price.

Apart from those questions, there are other things we can interpret from the dataset. Here they are:

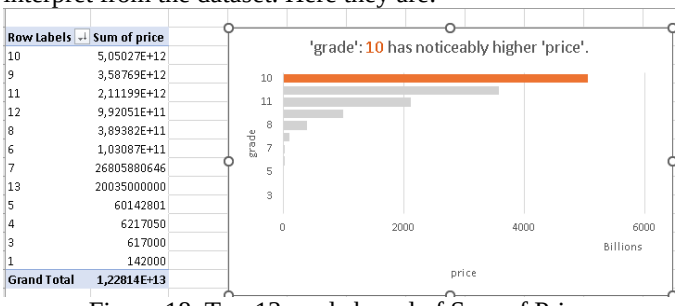

Figure 18. Top 13 grade based of Sum of Price

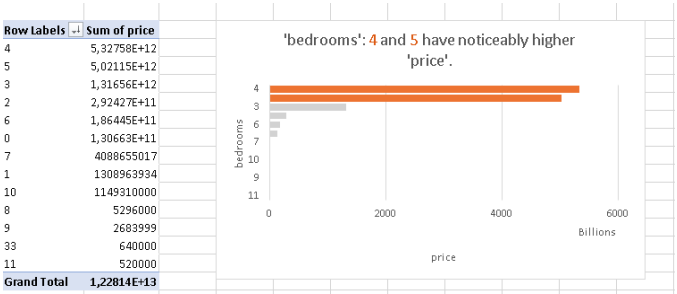

Figure 19. Top 13 bedrooms based of Sum of Price

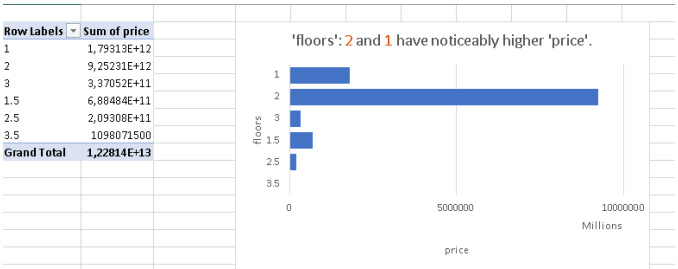

Figure 20. Sum of price based on floors

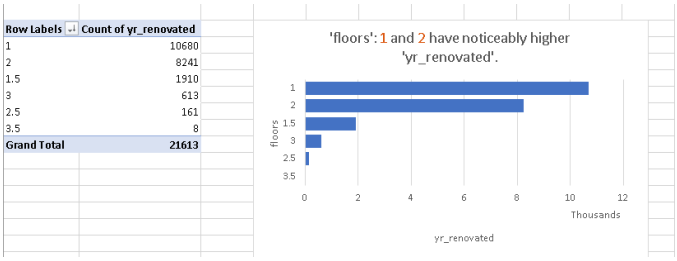

Figure 20. how many renovation has been done for every type of floors

VI. REFERENCES

[1] A. C. Goodman and T. G. Thibodeau, “Housing market segmentation,” Journal of housing economics, vol. 7, no. 2, pp. 121–143, 1998.
[2] C. Garriga, R. Manuelli, and A. Peralta-Alva, “A macroeconomic model of price swings in the housing market,” American Economic Review, vol. 109, no. 6, pp. 2036–2072, 2019.
[3] E. Mast, The effect of new market-rate housing construction on the low-income housing market. Upjohn Institute WP, 2019.
[4] G. D. Jud and D. T. Winkler, “The dynamics of metropolitan housing prices,” The journal of real estate research, vol. 23, pp. 29–46, 2002.
[5] J. A. Kahn, “What drives housing prices?,” in FRB of New York Staff Report, 2008.