House Prices Factors Analysis

----Take King County for Example

Ziqi Chen, Chien-Chieh Hu, Yu Li, Vivian Lin, Kefang Tian, Yi Zheng

1. Introduction

In this project, we use a variety of research methods to find out the relationship between the price of houses in Kings County and various factors. Our goal is to discover what factors are influencing the price of houses and some features int eh house sale market in Kings County.

1.1 Dataset description

This dataset contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. It has 21 variables including 19 features of the houses plus ID and price column, along with 21613 observations.

1.1.1 Source of data

Our dataset is from Kaggle (www.kaggle.com), which was founded in 2010 as a platform for predictive modelling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models.

1.1.2 Codebook

Names	Description
id	Unique ID for each home sold, 10 digit numbers
date	Date of the house sale, YYYYMMDD
price	The price of the house, in USD, with 2 decimals
bedrooms	The numbers of the bedrooms the house has
bathrooms	Number of bathrooms, where. 5 accounts for a room with a toilet but no shower
sqft_living	The total living area in a house, in square feet
sqft_above	The area above the ground, in square feet
sqft_basement	The area of basement, in square feet
yr_built	The year which the house were built, YYYY
yr_renovated	The year which the house were renovated, YYYY, 0 as the house hasn’t been renovated
zipcode	Zip code where the house in, 5 digit numbers
lat	The latitude of the house
long	The longitude of the house
sqft_lot	Square footage of the land space
floors	The numbers of floors the house has
waterfront	Whether the apartment was overlooking the waterfront or not
view	An index from 0 to 4 of how good the view of a house
conditions	The condition rating of the house range from 1 to 5

grade	An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high quality level of construction and design.
sqft_living15	The interior housing living space for the nearest 15 neighbors, in square feet.
sqft_lot15	The land lots of the nearest 15 neighbors, in square feet.

1.2 Purpose of the analysis of these data

Intuitively, we'd like to know what factors such as rooms, area and condition are significantly affect the price of a house and how they affect the price when they come together.

Also, when putting every house points in our dataset on the map (Figure 1-1), we find that the high price houses (red points, more than $2000k) and low price (black points, less than $200k) houses are to some degree concentrated. We try to testify this observation.

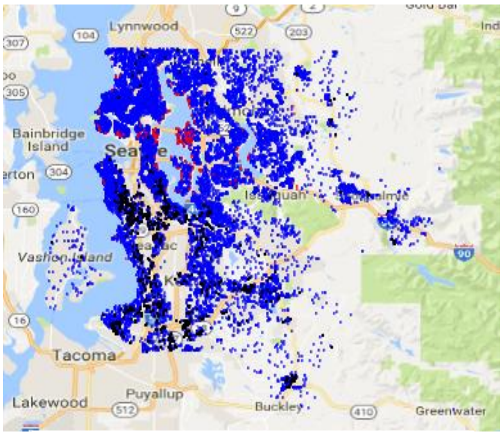


Figure 1-1

2. Statistical Analysis

2.1 ANOVA

First, let's analyze now the relationship between the independent variables available in the dataset and the dependent variable that we are trying to predict (i.e., price). These analyses should provide some interesting insights for our regression models.

We'll be using ANOVA and correlations coefficients (e.g., Pearson, Spearman) to explore potential associations between the variables.

a) Categorical variables

A one-way analysis of variance (ANOVA) is used when you have a categorical independent variable (with two or more categories) and a normally distributed interval dependent variable and you wish to test for differences in the means of the dependent variable broken down by the level of the independent variable.

First, we will try to assess if having a waterfront is related to a higher house value. Waterfront is a dichotomous variable with underlying continuous distribution (having a waterfront is better that not having a waterfront). From the results, we can see that the mean of the dependent variable differs significantly among the levels of program type. Only 163 of the houses with "waterfront" have a significant higher price. And r-square is only 7.10% which is relatively low correlation with the price.

Let's move on to the other variables. All the other variables are all significantly differ among the levels of the program type. "View" has an r-squared 16% which is modest correlated to the price. "Condition" has an r-squared 0.69% which is extremely low with the price. We will consider eliminate it from the regression model. And "grade" has an r-squared 51.97% which seems to be the best predictor among these categorical variables.

b) Continuous variables

Since these variables are measured on a continuous scale, we can use Pearson's coefficient r to measure the strength and direction of the relationship between price and other variables.

There is a clear linear association between the price and sqft_living (r = 0.7), indicating a strong positive relationship. sqft_living should be a good predicator of house price. Bedroom, bathroom, sqft_above and sqft_basement all have very high correlation with the price which we think they should be good predicators of house price. However the correlation between sqft_living and sqft_above, sqft_basement is also very high. We can use only sqft_above and sqft_basement instead of use all of the three.

Sqft_living 15 also have very high correlation with the price(r = 0.585). We are not sure if the relationship with house price is actually due to the average square footage of the 15th closest houses. This is because of the high correlation between sqft_living15 and sqft_living. To assess the true relationship between price and sqft_living15, we can use the Partial Correlation test. The correlation can assess the association between two continuous variables while controlling for the effect of other continuous variables called covariates. We will test the relationship between price and sqft_living15 using sqft_living as covariate. When controlling for sqft_living, the relationship between price and sqft_living15 disappeared (corr = 0.117). Sqft_lot, yr_built, yr_renovated and sqft_lot15 are poorly correlated to the price.

2.2 Regression Analysis

Next, we'd like to see what variables when they are put together affect the prices of houses in King County using Regression Models, which have good explanation ability.

Intuitively, you should not pay or owe anything to no-existing house, which means when the number of rooms, the square feet of living or basement and so on are all 0, the house price should be 0. So, we choose a regression through origin model through the following analysis.

Besides the attributes of the house itself, such as the number of bedrooms and bathrooms, the square feet of living and lot area and so on, we consider some features that may also affect the price. One is the transaction date, which could demonstrate whether the house prices experience an obvious trend (increasing or decreasing) during the period our dataset is set. The other is about the location. Since Seattle is the biggest city in King County and most population, we'd like to see if there is significant difference between in Seattle or not.

Firstly, we use all the variables those are thought useful to explain the house price then use backward stepwise to find the best model. We get the regression model as following:

Price = -0.04*days_since_2014.5.1 – 66.5*bedrooms + 31*bathrooms + 0.29*sqft_above + 0.23*sqft_basement – 0.0003*sqft_lot – 56.1*floors + 571*waterfront – 31.9*condition + 2.78*grade + 57.8*view + 1.36*age_built + 118.5*seattle_or_not;

RSE = 228 and Adjusted R-squared = 0.8781.

The performance of this model is not bad and the parameters are all significant. But we still find some parameters of the variables are negative such as "bedrooms" and "floors", which are against the common sense. An explanation of this is the collinearity: the larger the area of living a house has, the more bedrooms it has. The correlation coefficient of "bedrooms" and "sqft_above" is 0.49 and the correlation coefficient of "floors" and "sqft_above" is 0.52.

Also, we tend to believe the outliers play an important role in the model as Figure2-1 shows, the residuals belong to different distributions, especially the residuals at both ends. Back to the price data itself, it indicates some prices are extremely high so that we could not regard all the houses as a whole. With reality and the data structure, we take our focus on the data in which prices are between the first quartile (322000) and third quartile (645000). The houses in the range are the main body of original dataset which we define as middle-class houses.

a) Middle-class house

We get a much better model in this dataset:

Price = 14.8*bathrooms + 0.024*sqft_above + 0.026*sqft_basement + 0.00013*sqft_lot + 7.9*condition + 41*grade + 0.84*age_built + seattle_or_not;

RSE = 80.5 and Adjusted R-squared = 0.9708.

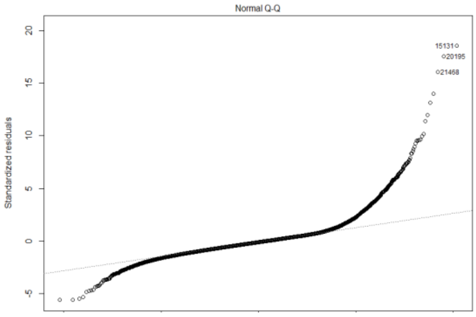


Figure 2-1

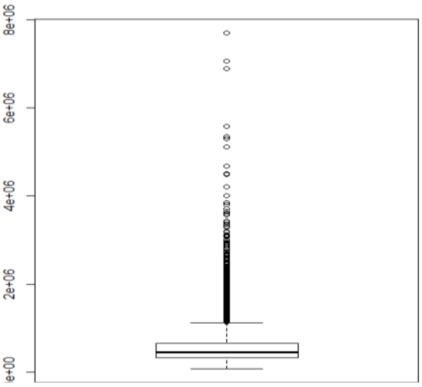


Figure 2-2

b) Luxury house

For comparison, we try to explore the high-price houses. We take the houses whose prices are more than 2000000. We get:

Price = 0.17*sqft_above + 0.19*sqft_basement + 435*waterfront + 159.6*grade;

RSE = 518 and Adjusted R-squared = 0.9645.

The RSE increases sharply which means the model performance of luxury house is worse than the one of middle-class house.

But by comparison, there are still some interesting insights: people who buy middle-class house regard the practical and functional features like "condition" and "bathrooms" more important, while the luxury house buyers pay more attention on the enjoyment who like to live by waterfront.

2.3 Zip code Dummy

We also try to extract information from the variable "zip code", which represents the location of every house in our dataset. From Figure 2-3, we can see obviously different cities have different price levels.

Also, according to the "zip code" and their location, we divide the King County into 6 areas and regard this area variable as dummy variable in our regression model. We get better model performance and find different areas have different prices.

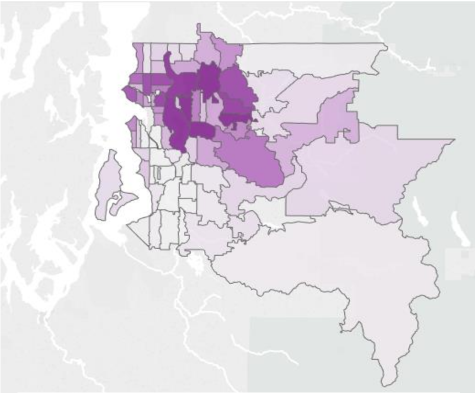


Figure 2-3

3. Conclusions

3.1 Answers of the purposes

We find out that for common price houses, bathroom, sqft_above, sqft_basement, sqft_lot, condition, grade, age_built etc can affect the price, which are more related to “practical” and “functional”. On the other hand, for luxury house waterfront becomes a significant factor to affect the price, which may mean that rich men focus more on enjoyment.

Also, we find that the prices of houses in different cities are significantly different, which the location is also an important factor of the house price.

3.2 Software

We use R and Minitab in our statistical analyses. Both are excellent statistical software, which are easy to use and give detailed regression results. (See Appendix)

Appendix:

a) Parts of Minitab results:

One-way ANOVA: price versus waterfront

```
Source         DF           SS           MS        F      P
waterfront      1  2.06679E+14  2.06679E+14  1650.46  0.000
Error       21611  2.70624E+15  1.25225E+11
Total       21612  2.91292E+15

S = 353871   R-Sq = 7.10%   R-Sq(adj) = 7.09%


Level      N     Mean    StDev
0      21450   531564   341600
1        163  1661876  1120372

       Individual 95% CIs For Mean Based on Pooled StDev
Level     --+---------+---------+---------+-------
0         *
1                                     (*-)
          --+---------+---------+---------+-------
       600000    900000   1200000   1500000

Pooled StDev = 353871
```

One-way ANOVA: price versus view

```
Source     DF           SS           MS        F      P
view        4  4.90079E+14  1.22520E+14  1092.69  0.000
Error   21608  2.42284E+15  1.12127E+11
Total   21612  2.91292E+15

S = 334854   R-Sq = 16.82%   R-Sq(adj) = 16.81%


                               Individual 95% CIs For Mean Based on
                               Pooled StDev
Level      N     Mean   StDev   ----+---------+---------+---------+-----
0      19489   496564  287133   (*
1        332   812281  510950             (*)
2        963   792401  510105             *)
3        510   971965  612692                  (*)
4        319  1463711  952210                                   (*)
                                ----+---------+---------+---------+-----
                               600000    900000   1200000   1500000

Pooled StDev = 334854
```

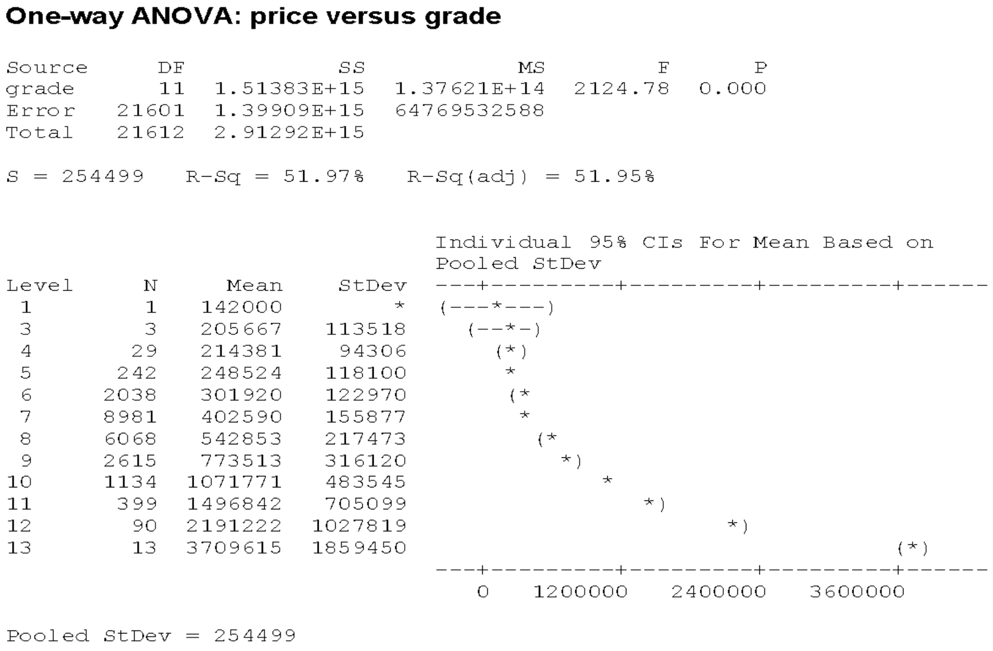
One-way ANOVA: price versus condition

```
Source        DF           SS           MS      F      P
condition      4  2.00346E+13  5.00866E+12  37.41  0.000
Error      21608  2.89288E+15  1.33880E+11
Total      21612  2.91292E+15

S = 365896   R-Sq = 0.69%   R-Sq(adj) = 0.67%


                              Individual 95% CIs For Mean Based on Pooled StDev
Level      N    Mean   StDev    ---+---------+---------+---------+------
1         30  334432  271173    (----------*----------)
2        172  327287  245418          (---*----)
3      14031  542013  364449                                *)
4       5679  521200  358516                              *)
5       1701  612418  410972                                     (*)
                                ---+---------+---------+---------+------
                              240000    360000    480000    600000

Pooled StDev = 365896
```



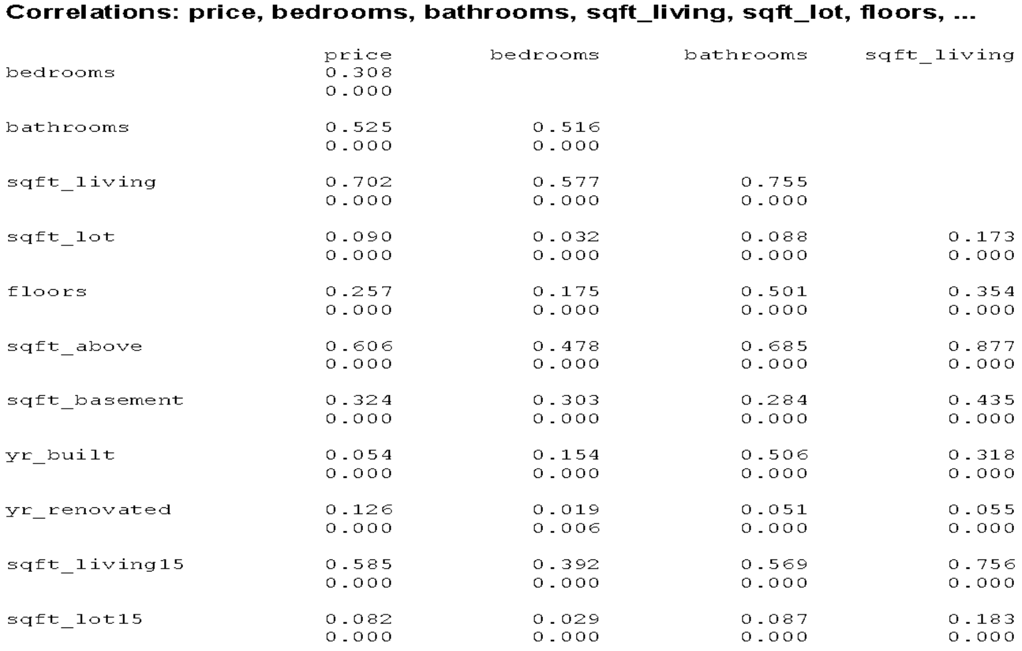
One-way ANOVA: price versus grade

```
Source      DF           SS           MS        F      P
grade       11  1.51383E+15  1.37621E+14  2124.78  0.000
Error    21601  1.39909E+15  64769532588
Total    21612  2.91292E+15

S = 254499   R-Sq = 51.97%   R-Sq(adj) = 51.95%

                                Individual 95% CIs For Mean Based on
                                Pooled StDev
Level     N     Mean    StDev   ---+---------+---------+---------+------
 1        1   142000        *   (---*---)
 3        3   205667   113518     (--*-)
 4       29   214381    94306       (*)
 5      242   248524   118100        *
 6     2038   301920   122970        (*
 7     8981   402590   155877         *
 8     6068   542853   217473          (*
 9     2615   773513   316120            *)
10     1134  1071771   483545               *
11      399  1496842   705099                  *)
12       90  2191222  1027819                        *)
13       13  3709615  1859450                                    (*)
                                ---+---------+---------+---------+------
                                   0   1200000   2400000   3600000

Pooled StDev = 254499
```



Correlations: price, bedrooms, bathrooms, sqft_living, sqft_lot, floors, ...

```
                        price     bedrooms    bathrooms  sqft_living
bedrooms                0.308
                        0.000

bathrooms               0.525        0.516
                        0.000        0.000

sqft_living             0.702        0.577        0.755
                        0.000        0.000        0.000

sqft_lot                0.090        0.032        0.088        0.173
                        0.000        0.000        0.000        0.000

floors                  0.257        0.175        0.501        0.354
                        0.000        0.000        0.000        0.000

sqft_above              0.606        0.478        0.685        0.877
                        0.000        0.000        0.000        0.000

sqft_basement           0.324        0.303        0.284        0.435
                        0.000        0.000        0.000        0.000

yr_built                0.054        0.154        0.506        0.318
                        0.000        0.000        0.000        0.000

yr_renovated            0.126        0.019        0.051        0.055
                        0.000        0.006        0.000        0.000

sqft_living15           0.585        0.392        0.569        0.756
                        0.000        0.000        0.000        0.000

sqft_lot15              0.082        0.029        0.087        0.183
                        0.000        0.000        0.000        0.000
```

b) R code

```
#Read data
getwd()
dir_path <-"C:/Users/yizheng/Documents/575"
setwd(dir_path)
rm(list=ls())
infile<-"kc_house_data_clean.csv"
kc_house=read.csv(infile, header = TRUE, sep = ",")
#Data Overview on map
library("RgoogleMaps")
bb=qbbox(lat=kc_house$lat,lon = kc_house$long)
```

```
MyMap=GetMap.bbox(bb$lonR,bb$latR,destfile = "King County.png",GRAYSCALE = FALSE)
tmp=PlotOnStaticMap(MyMap,lat=kc_house$lat,lon=kc_house$long ,cex=0.25,pch=20,col=rgb(0,0,1,0.5),add=FALSE)
tmp=PlotOnStaticMap(MyMap,lat=kc_house$lat[which(kc_house$price>2000000)],lon=kc_house$long[which(kc_house$price>2000000)],cex=0.25,pch=20,col=rgb(1,0,0,0.5),add=TRUE)
tmp=PlotOnStaticMap(MyMap,lat=kc_house$lat[which(kc_house$price<200000)],lon=kc_house$long[which(kc_house$price<200000)],cex=0.25,pch=20,col=rgb("black"),add=TRUE)
#Regression Models
test_kc<-lm(price/1000~-1+days_since_2014.5.1+bedrooms+bathrooms+sqft_above+sqft_basement+sqft_lot+floors+waterfront+condition+grade+view+age_built+age_renovated+seattle_or_not,data=kc_house)
step(test_kc)
plot(test_kc, which=1:4)
test_kc_coll<-lm(price/1000~-1+days_since_2014.5.1+bathrooms+sqft_above+sqft_basement+sqft_lot+waterfront+condition+grade+view+age_built+seattle_or_not,data=kc_house)
step(test_kc_coll)
#Common houses
kc_common_house<-kc_house[which((kc_house$price<=645000)&(kc_house$price>=322000)),]
test_kc_coll<-lm(price/1000~-1+days_since_2014.5.1+bathrooms+sqft_above+sqft_basement+sqft_lot+waterfront+condition+grade+view+age_built+seattle_or_not,data=kc_common_house)
test_kc_coll1<-lm(price/1000~-1+bathrooms+sqft_above+sqft_basement+sqft_lot+condition+grade+age_built+seattle_or_not,data=kc_common_house)
#Luxury houses
kc_luxury_house<-kc_house[which((kc_house$price>=2000000)&(kc_house$price<6000000)),]
test_kc_coll<-lm(price/1000~-1+days_since_2014.5.1+bathrooms+sqft_above+sqft_basement+sqft_lot+waterfront+condition+grade+view+age_built+seattle_or_not,data=kc_luxury_house)
test_kc_coll1<-lm(price/1000~-1+sqft_above+sqft_basement+waterfront+grade,data=kc_luxury_house)
#Zip code dummy
library(datasets)
set.seed(1234)
km_zip=kmeans(kc_house$zipcode,centers=6,nstart = 100)
kc_house$zipcodelab[km_zip$cluster==1]=1
kc_house$zipcodelab[km_zip$cluster==2]=2
kc_house$zipcodelab[km_zip$cluster==3]=3
kc_house$zipcodelab[km_zip$cluster==4]=4
kc_house$zipcodelab[km_zip$cluster==5]=5
kc_house$zipcodelab[km_zip$cluster==6]=6
kc_house$zipcodelab<- factor(kc_house$zipcodelab)
test_kc_dummy<-lm(price/1000~-1+days_since_2014.5.1+bedrooms+bathrooms+sqft_above+sqft_basement+sqft_lot+floors+waterfront+condition+grade+view+age_built+age_renovated+zipcodelab,data=kc_house)
```