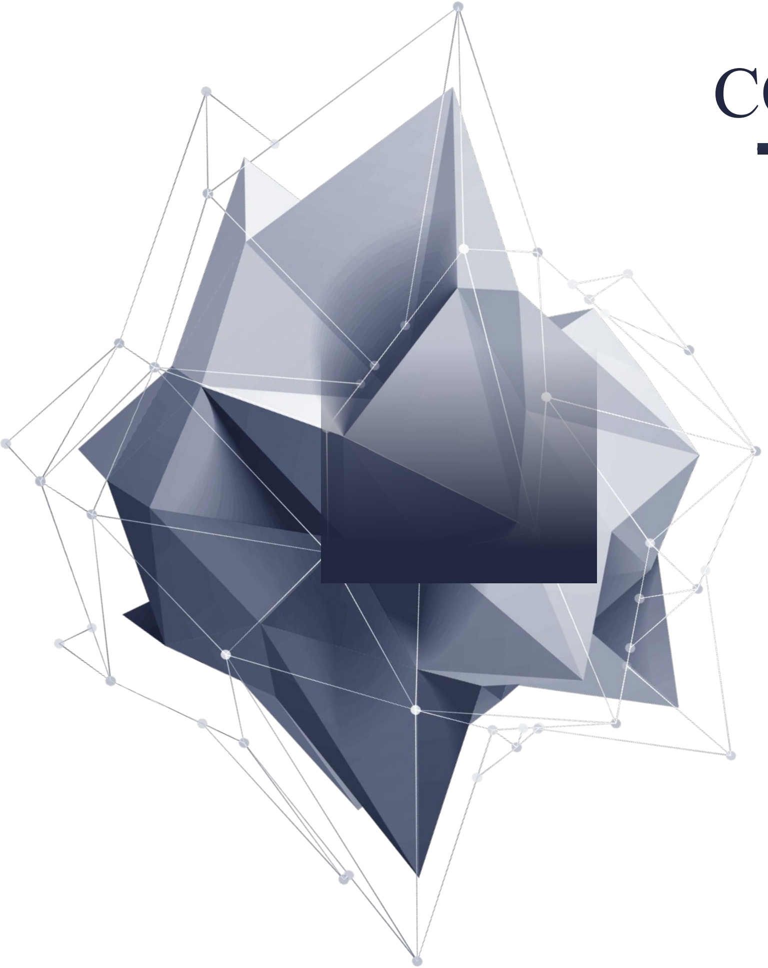
Data Mining Final Project

2014-15 House Price
in King County, WA Analysis

Presenter: Weike Zhou Yixuan Yang



CONTENTS

Introduction of Dataset

- **Background:**

  --House sales price for King county in Washington

  --Houses were sold between May 2014 and May 2015

  --Characteristics of house include number of bedrooms, bathrooms, building years, size of living area, etc

- **Data Details:**

  --*Source:* Kaggle Dataset

  --*Data Size:* The dataset contains House prices and characteristics with 21613 observations, 21 variables.

2014-15 House Price in King County, WA

```
'data.frame':	21613 obs. of  21 variables:
 $ id           : num  7129300520 6414100192 5631500400 2487200875
1954400510 ...
 $ 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 ...
```

Introduction of Dataset

- **Data Details:**

  --*Dependent variables:* Price.

  --*Independent variables:* Numeric *variables:* bedrooms, bathrooms, sqft_liv, sqft_lot, floors, sqft_above, sqft_basement, yr_built, yr_renovated,squft_liv15, squft_lot15

  Categorical *variables:* waterfront, view, grade, condition.

Waterfront	‘1’ if the property has a waterfront, ‘0’ if not.
View	How good the view of the property was (from 0 to 4)
Grade	Construction quality which refers to the types of materials used and the quality of workmanship. Higher grade, higher quality. Score from 4 to 12
Condition	Condition of the house, ranked from 1 to 5
squft_liv15	Average size of interior housing living space for the closest 15 houses, in square feet
yr_renovated	Year renovated. ‘0’ if never renovated (RISK)

Questions & Challenges

- **Questions:**

  Regression:

  --What factor will influence the price the most?

  --Which model could be best to predict the future price? Any business insights?

  Classification:

  --What factor will influence lower,median,high price the most?

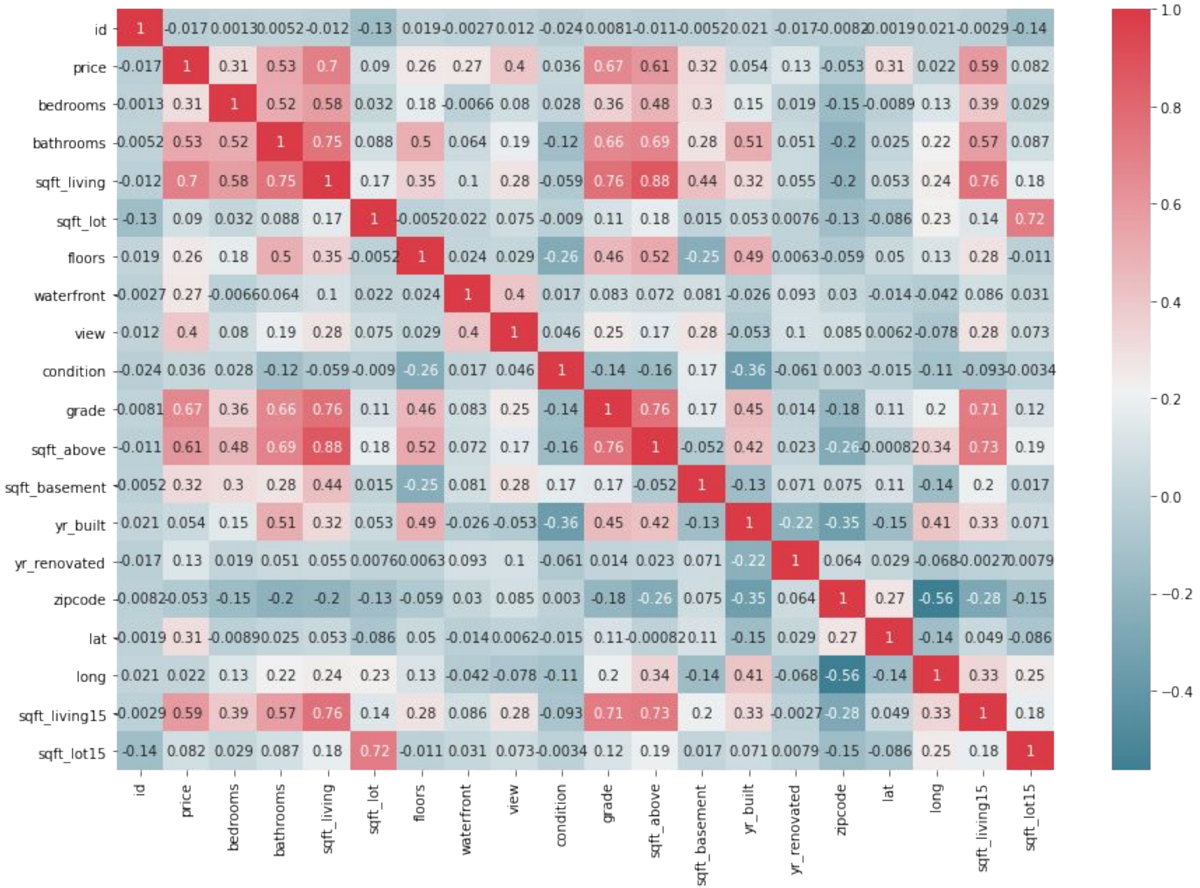
  --Which cluster size is the best one?

- **Challenges:**

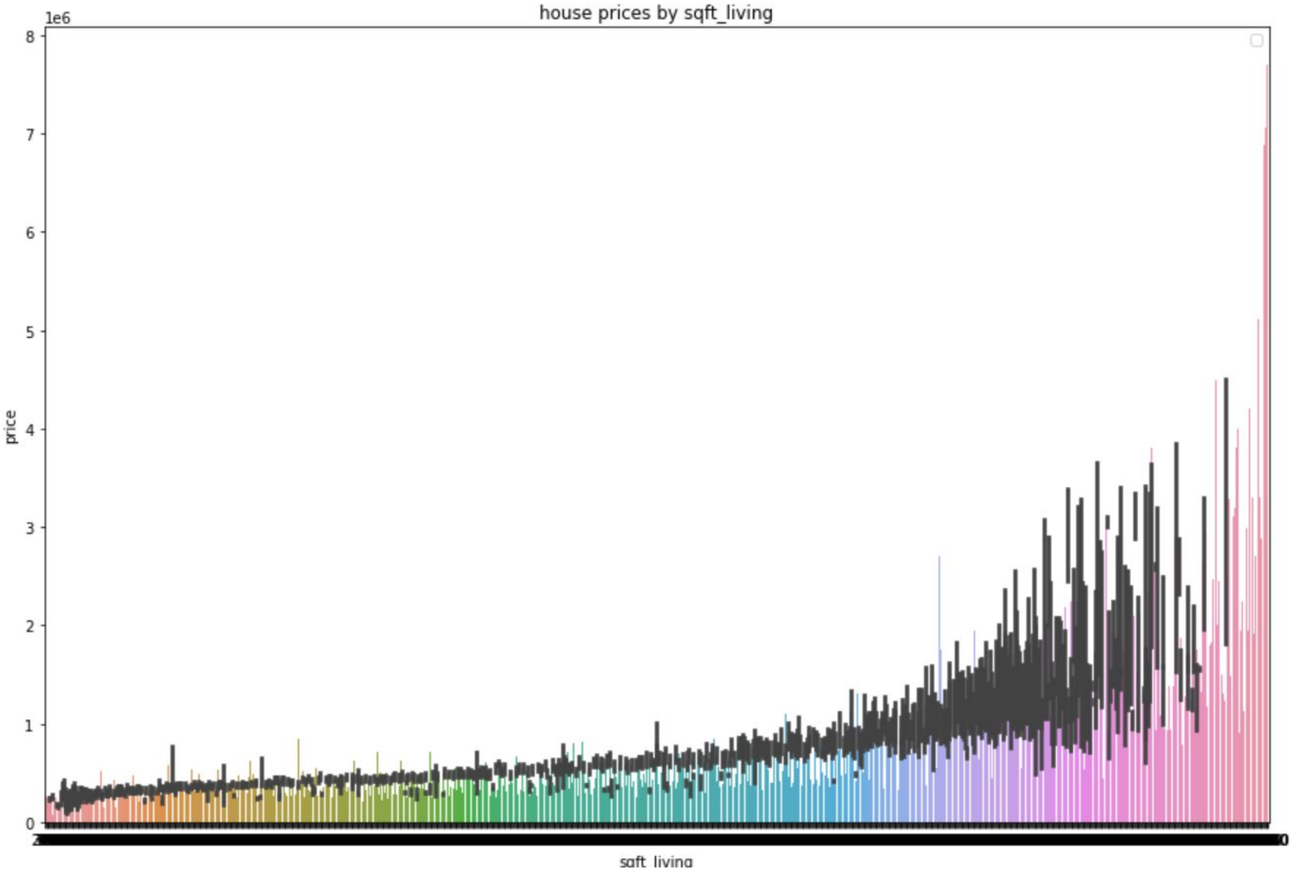
  --How to compare the models when some models are hard to interpret without coefficient results?

  --How to make our model more accurate? (drawback and advantages)

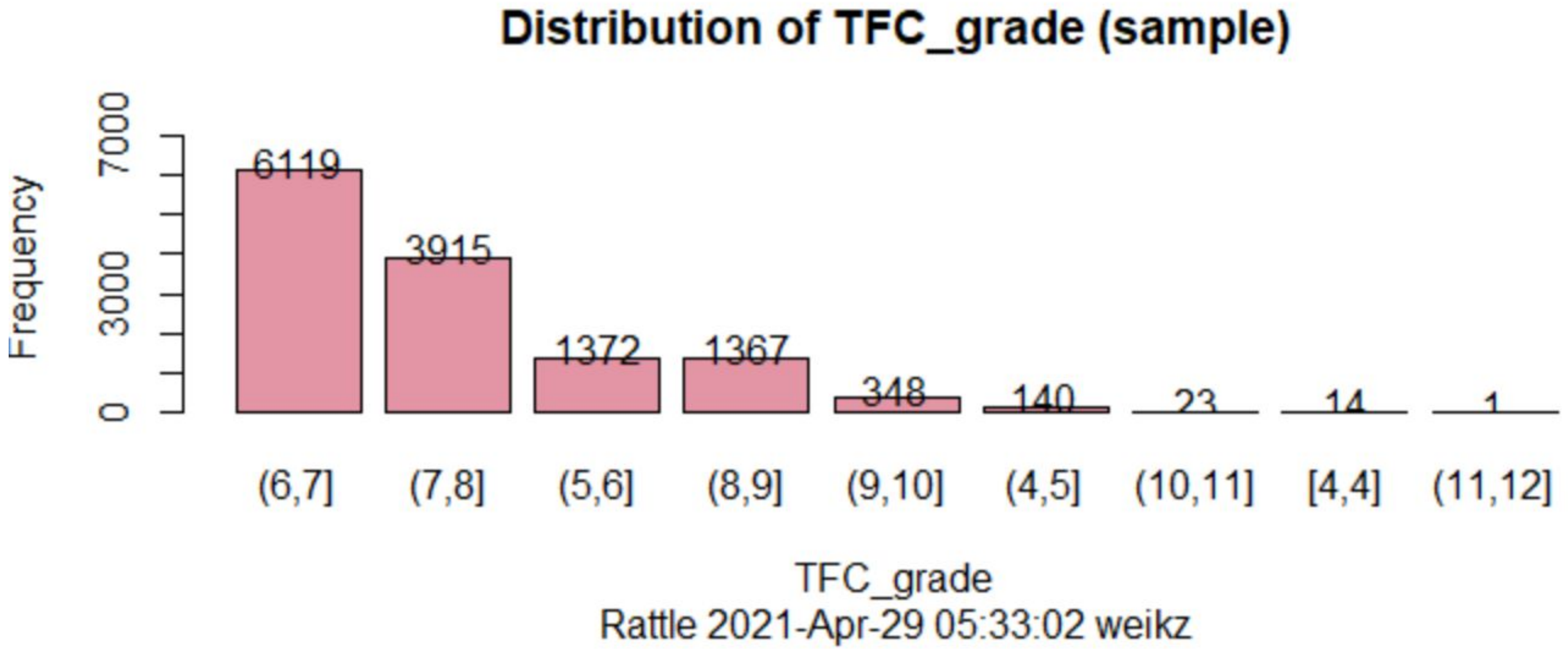
Data Exploration for **Correlation matrix**



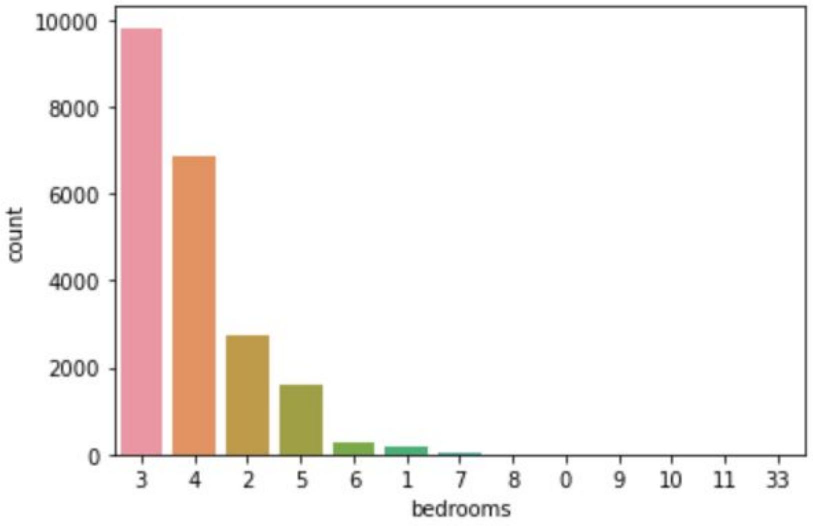
Data Exploration for **house_price Vs sqft_living**

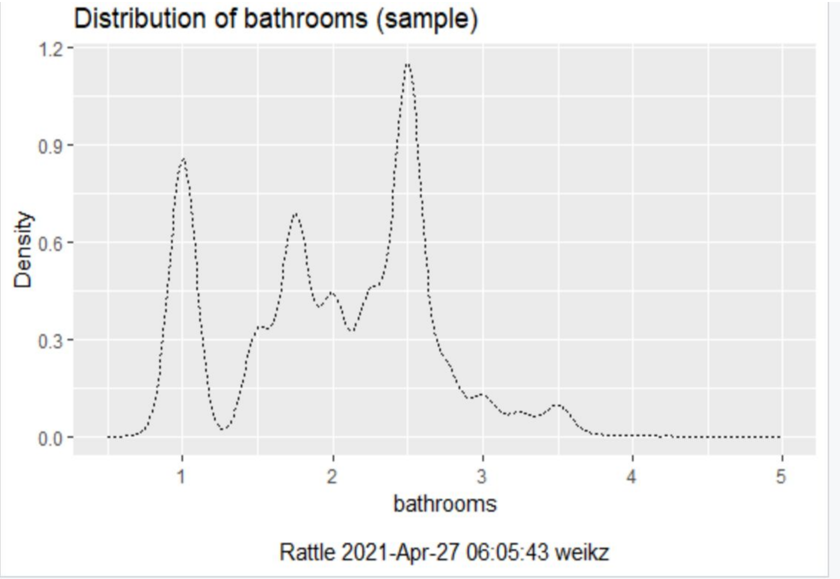


The distribution of price by grade

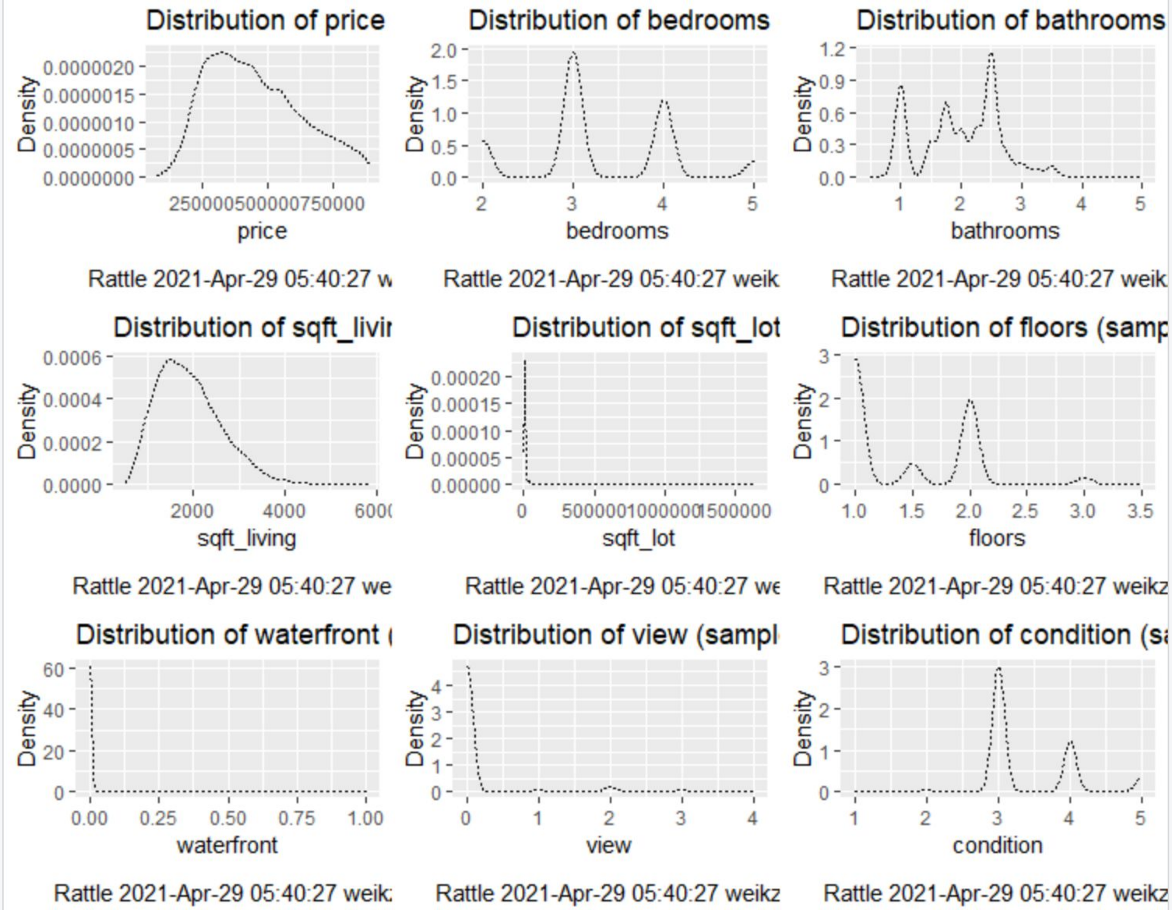


Data Exploration for **bedrooms VS bathrooms**

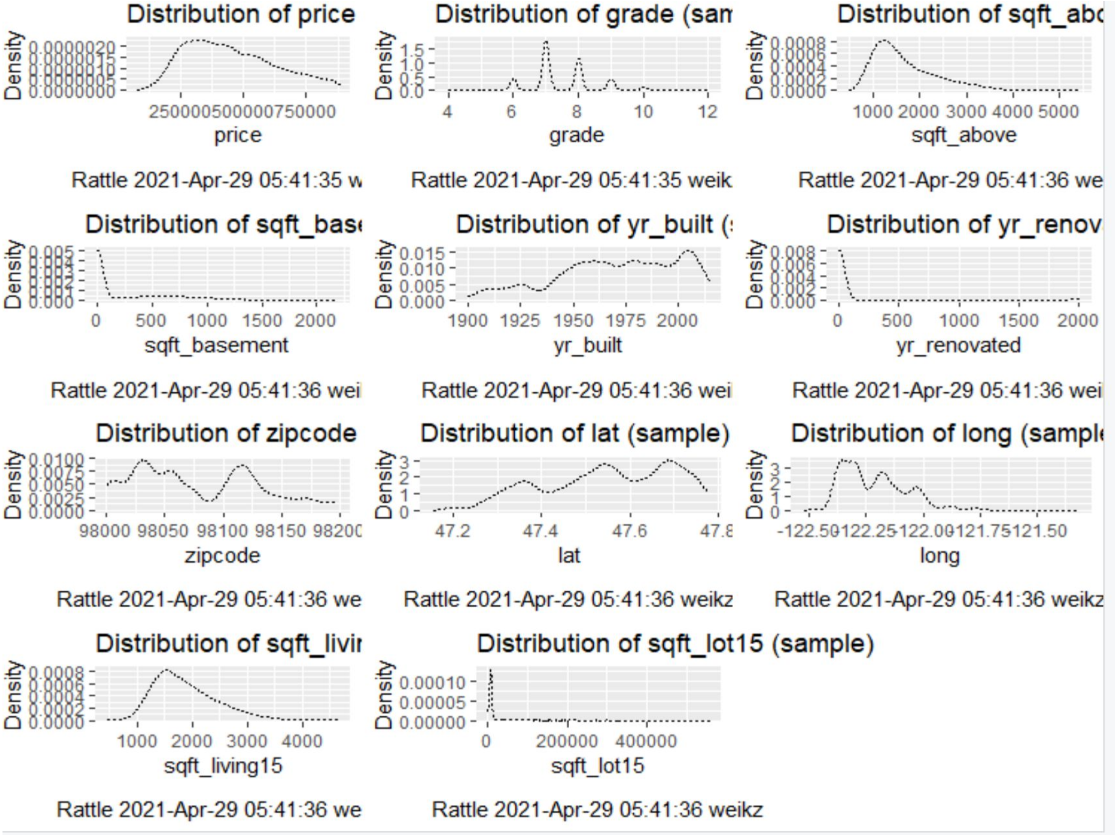




Data Exploration for **overview of all variables**

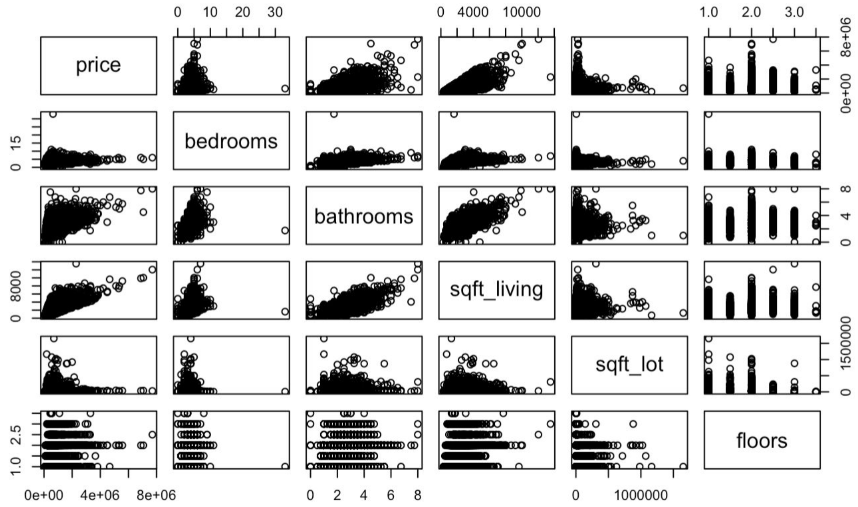


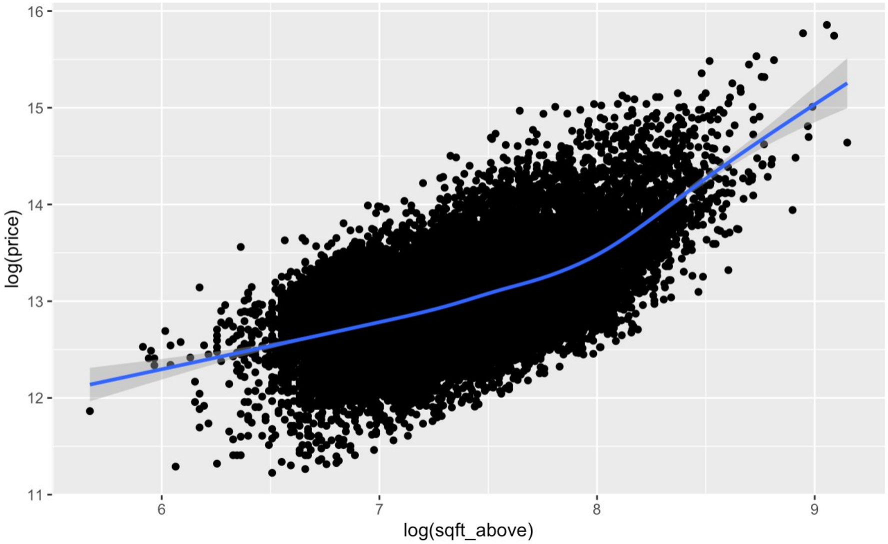
Data Exploration for **overview of all variables**



Data Exploration for **partial predictors' linear relationships**

- Drew the scatterplot about some variables. From the scatter plot below, we can see that there are some linear relationships implicitly
- Used log-log model to explore the linear relationship between price and sqft_above.





Data Processing

Data cleaning:

- Delete ID, date,NA…
- Filter dataset
- Check and convert data types
- Split the training and test dataset (70:30)
- waterfront, view, grade, condition.

```
# Filter the dataset
midrange_homes = data[(data['price'] < np.quantile(data['price'], 0.9))
                        & (data['bedrooms'].isin(range(2, 6)))]
```

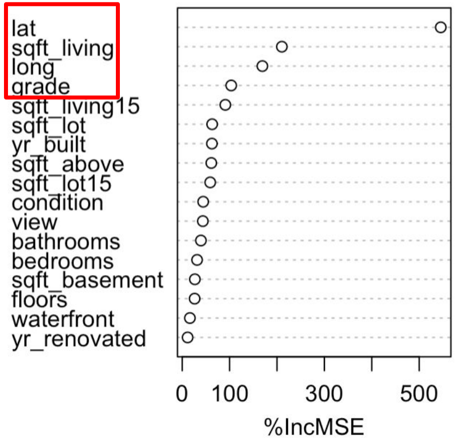
```
df$grade = factor(df$grade)
df$view = factor(df$view)
df$condition = factor(df$condition)
df$waterfront = factor(df$waterfront)
```

Modeling--regression

1. Random Forest

- mtry=17, test error=4,450,194,793
- The 1st 4 important variables: lat, sqft_living, long, grade

	%IncMSE	IncNodePurity
bedrooms	31.38267	1.432599e+12
bathrooms	39.46528	2.463375e+12
sqft_living	210.25284	1.081878e+14
sqft_lot	63.16172	7.121208e+12
floors	26.21764	6.225795e+11
waterfront	16.21966	6.895002e+11
view	43.52627	3.743517e+12
condition	44.23998	2.704850e+12
grade	103.41068	1.827374e+13
sqft_above	61.37143	6.632284e+12
sqft_basement	26.69226	2.041329e+12
yr_built	62.68930	7.709656e+12
yr_renovated	11.41584	7.156749e+11
lat	545.57275	1.770634e+14
long	169.14901	2.008848e+13
sqft_living15	90.97886	1.270739e+13
sqft_lot15	59.11461	6.061640e+12



- mtry=6, test error=4,669,091,136
- The 1st 4 important variables: lat, sqft_living, long, yr_built

	%IncMSE	IncNodePurity
bedrooms	19.476506	2.908050e+12
bathrooms	24.631939	6.915379e+12
sqft_living	60.552817	5.360374e+13
sqft_lot	40.796447	9.345874e+12
floors	17.162713	2.331588e+12
waterfront	14.084784	6.658717e+11
view	41.747572	4.196144e+12
condition	34.352632	3.261897e+12
grade	45.785388	5.225557e+13
sqft_above	34.784721	1.887886e+13
sqft_basement	27.490308	5.846785e+12
yr_built	58.582472	1.500194e+13
yr_renovated	8.706058	9.257194e+11
lat	281.000426	1.498791e+14
long	98.117964	1.641277e+13
sqft_living15	53.611592	2.890176e+13
sqft_lot15	44.808578	1.056674e+13

Modeling--regression

2. Multiple Linear Regression

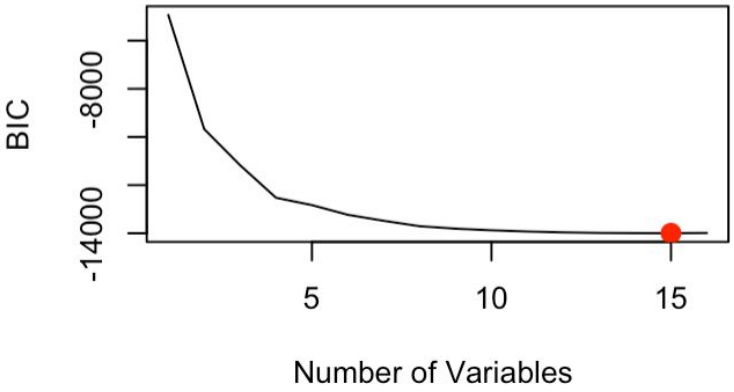
- After removing sqft_above, test error=52,188,694,666
- Bedrooms(-); Year built(-);
- Important coefficients: Bedrooms; grade
- Risk: Interaction terms, eg: condition and waterfront/view

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.514e+07  1.080e+06 -14.020  < 2e-16 ***
bedrooms     -7.111e+03  1.476e+03  -4.817 1.47e-06 ***
bathrooms     1.759e+04  2.302e+03   7.644 2.27e-14 ***
sqft_living   6.550e+01  2.909e+00  22.517  < 2e-16 ***
sqft_lot      1.580e-01  5.508e-02   2.869  0.00413 **
floors        1.799e+04  2.572e+03   6.996 2.77e-12 ***
waterfront    2.002e+05  2.372e+04   8.437  < 2e-16 ***
view          2.364e+04  1.715e+03  13.787  < 2e-16 ***
condition     2.676e+04  1.534e+03  17.441  < 2e-16 ***
grade         6.788e+04  1.531e+03  44.346  < 2e-16 ***
sqft_basement 4.704e-02  3.296e+00   0.014  0.98861
yr_built     -1.627e+03  4.935e+01 -32.980  < 2e-16 ***
yr_renovated  1.112e+01  2.641e+00   4.211 2.55e-05 ***
lat           5.113e+05  6.737e+03  75.887  < 2e-16 ***
long          5.196e+04  8.083e+03   6.429 1.33e-10 ***
sqft_living15 5.486e+01  2.664e+00  20.588  < 2e-16 ***
sqft_lot15   -4.914e+00  2.819e-01 -17.429  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 99480 on 12426 degrees of freedom
Multiple R-squared:  0.678,	Adjusted R-squared:  0.6776
F-statistic:  1635 on 16 and 12426 DF,  p-value: < 2.2e-16
```

Modeling--regression

3. Best Subset Selection

- 17→15 variables
- Test error=17,167,066,295
- Lose yr_built and sqft_above
- Important coefficients: sqft_living



(Intercept)	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront
-3.109016e+07	-5.897780e+03	-5.832878e+03	7.865466e+01	2.669444e-01	-3.448395e+03	1.829180e+05
view	condition	grade	yr_renovated	lat	long	sqft_living15
2.869096e+04	4.027556e+04	5.567426e+04	3.786431e+01	5.598731e+05	-3.418349e+04	5.576620e+01
sqft_lot15	sqft_basement					
-5.409317e+00	-4.136221e+00					

Modeling--regression

Fit shrunken model

4. Ridge Regression

- Best λ=10091.72
- 17→17 variables
- Test Error=9,964,802,522

(Intercept)	-1.711505e+07
bedrooms	-6.706726e+03
bathrooms	1.884217e+04
sqft_living	3.723874e+01
sqft_lot	1.597711e-01
floors	1.510872e+04
waterfront	1.699581e+05
view	2.404823e+04
condition	2.442202e+04
grade	6.196921e+04
sqft_above	3.033315e+01
sqft_basement	2.745417e+01
yr_built	-1.422785e+03
yr_renovated	1.351733e+01
lat	4.908777e+05
long	3.074496e+04
sqft_living15	5.393351e+01
sqft_lot15	-4.625121e+00

(-)

Important

5. Lasso Regression

- Best λ=217.4194
- 17→17 variables
- Test Error=9,927,462,627

(Intercept)	-1.601102e+07
bedrooms	-7.440098e+03
bathrooms	2.028151e+04
sqft_living	6.271046e+01
sqft_lot	1.685914e-01
floors	1.451453e+04
waterfront	1.801085e+05
view	2.386545e+04
condition	2.459468e+04
grade	6.826746e+04
sqft_above	3.447964e+00
sqft_basement	.
yr_built	-1.614595e+03
yr_renovated	1.097970e+01
lat	5.089835e+05
long	4.407038e+04
sqft_living15	5.402581e+01
sqft_lot15	-4.979224e+00

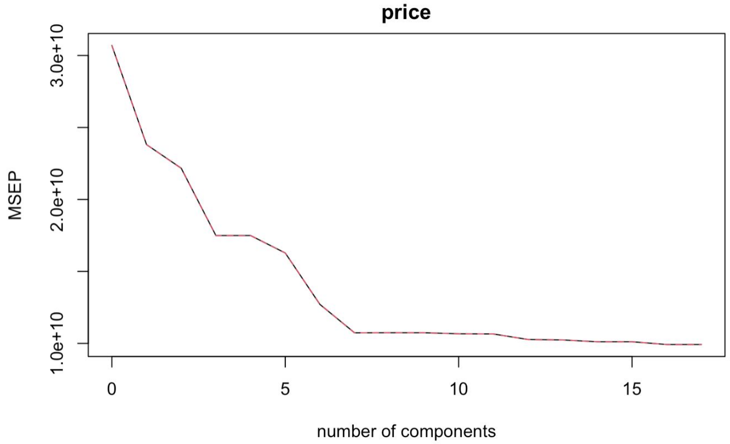
(-)

Important

Modeling--regression

6. Principal Components Regression

- Standardize each predictor; 10-fold CV
- Number of components considered: **17**
- Test Error=9,917,015,840



6*. Partial Least Squares

- The lowest CV error occurs when M = **9** partial least squares directions are used. (smallest adjCV)
- Test Error=10,299,700,761
- PLS searches for directions that explain variance in both the predictors and the response. It explains 67.97 % variance in Price.

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps	9 comps
X	26.33	37.26	45.46	54.98	59.06	63.59	66.76	72.07	76.36
price	45.29	64.92	67.26	67.63	67.92	67.96	67.97	67.97	67.97

Modeling--regression

Comparison of models

	Random Forest✔	Multiple Linear Regression	Best Subset Selection	Ridge Regression	Lasso Regression✔	PCR PLS
Test Error (MSE)	4,669,091,136	52,188,694,666	17,167,066,295	9,964,802,522	9,927,462,627	9,917,015,840 10,299,700,761
Advantages	-Importance rank; -low error rate(good prediction); -Simple model-fitting procedure	-Coefficients results; - +/- relationship; -Simple model-fitting procedure	-Coefficients results; -decrease variables(reduce complexity); -Simple model-fitting procedure	-Coefficients results; -standardize	-Coefficients results; -standardize; -low error rate	-Standardize; -Dimension reduction(PLS); -low error rate; -Simple model-fitting procedure
Disadvantages	-No coefficients; -no standardize;	-Interaction; -no standardize; -bad prediction	-So-so prediction	-Complex model-fitting procedure	-Complex model-fitting procedure	-No coefficients; -PCR not sparse

Conclusion

Regression:

- Random Forest:

  More important: lat, sqft_living, long, grade, sqft_living15

  Less important: sqft_basement, floors, waterfront
- Lasso: Price & bedrooms (-);Price & sqft_living (+); Price &sqft_living15(+)

- Business Insights: Higher **construction quality, size of living space, and average size of living space for the nearest 15 neighbors** will result in higher price.

  Sellers could highlight those in the ads, which may increase his/her house market value and gain competitive advantages.

  Buyers can forecast market trend and appraise the price based on the important factors.

Modeling--Define the response variables

- Develop a model to predict a house price low, median, or high
- Create a category variable, as the following picture
  (median price: 451033.600453) (70/15/15)

In [58]:
```
### The price range (mean: 451033.600453)
### lower:-nf~200000
### median:200000~6000000
## high: 600000~886000
```

In [61]:
```
bins = [-np.inf,200000,600000,np.inf]
labels=['low','medium','high']
item_price_range['Price Category'] = pd.cut(item_price_range['price'], bins=bins, labels=labels)
print (item_price_range)
```

```
          price Price Category
0      221900.0         medium
1      538000.0         medium
2      180000.0            low
3      604000.0           high
4      510000.0         medium
...         ...            ...
21608  360000.0         medium
21609  400000.0         medium
21610  402101.0         medium
21611  400000.0         medium
21612  325000.0         medium

[18999 rows x 2 columns]
```

Modeling--classification(linear regression)

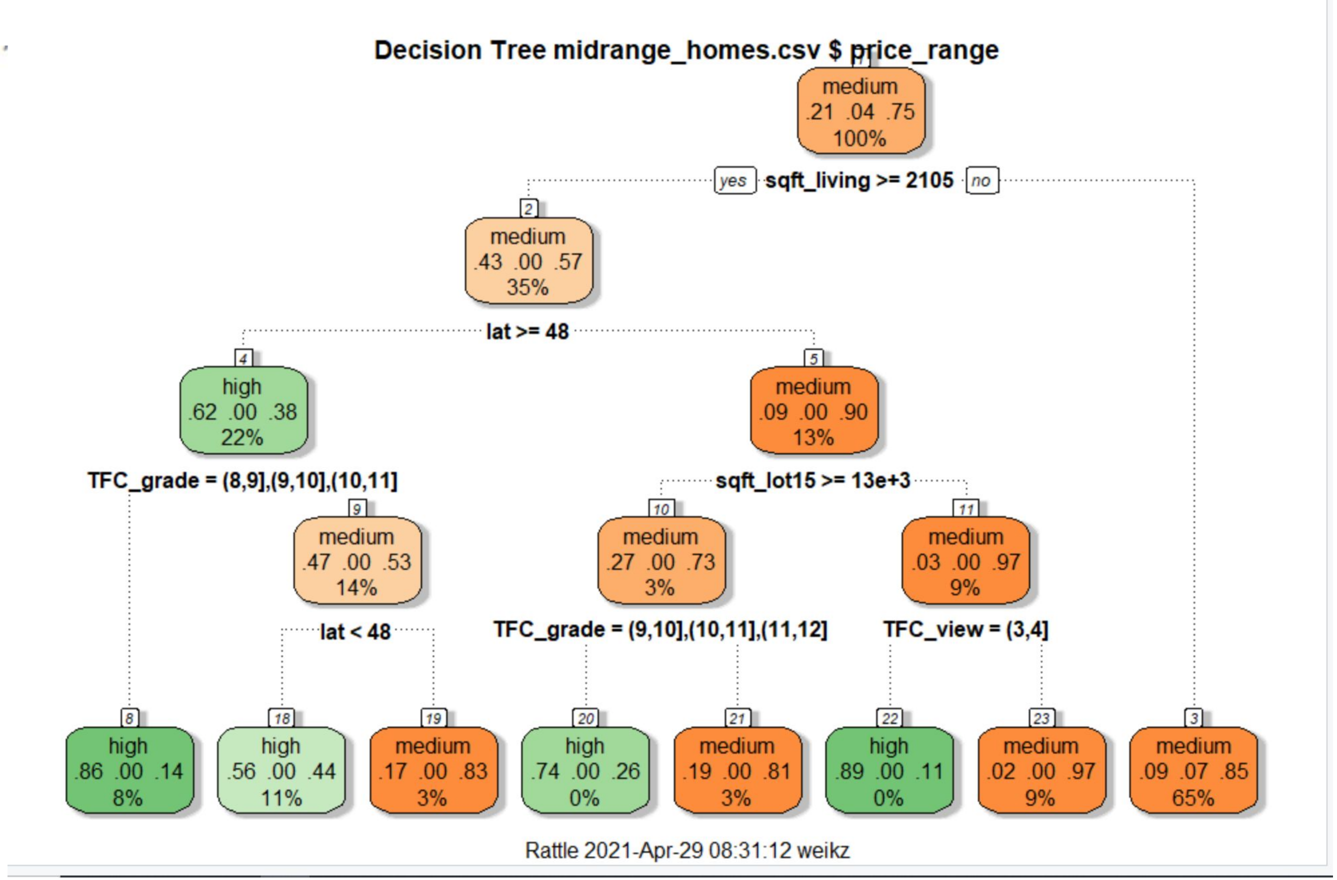
```
==== ANOVA ====
Analysis of Deviance Table (Type II tests)

Response: price_range
               LR Chisq Df Pr(>Chisq)
bedroms            17.12  2  0.0001913 ***
bathrooms          36.80  2  1.020e-08 ***
sqft_living         4.24  2  0.1199951
sqft_lot           17.88  2  0.0001312 ***
floors             55.87  2  7.368e-13 ***
sqft_above          5.00  2  0.0822394 .
sqft_basement       6.50  2  0.0387254 *
yr_built          419.20  2  < 2.2e-16 ***
yr_renovated       13.54  2  0.0011463 **
zipcode            27.27  2  1.197e-06 ***
lat              1510.93  2  < 2.2e-16 ***
long               10.30  2  0.0058130 **
sqft_living15     104.97  2  < 2.2e-16 ***
sqft_lot15         16.76  2  0.0002300 ***
TFC_waterfront     15.91  2  0.0003509 ***
TFC_view          100.90  8  < 2.2e-16 ***
TFC_condition     215.38  8  < 2.2e-16 ***
TFC_grade         799.95 16  < 2.2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
[1] "\n"
Time taken: 32.73 secs
```

```
Residual Deviance: 10455.33
AIC: 10579.33
Log likelihood: -5227.665 (62 df)
Pseudo R-Square: 0.54705700
```

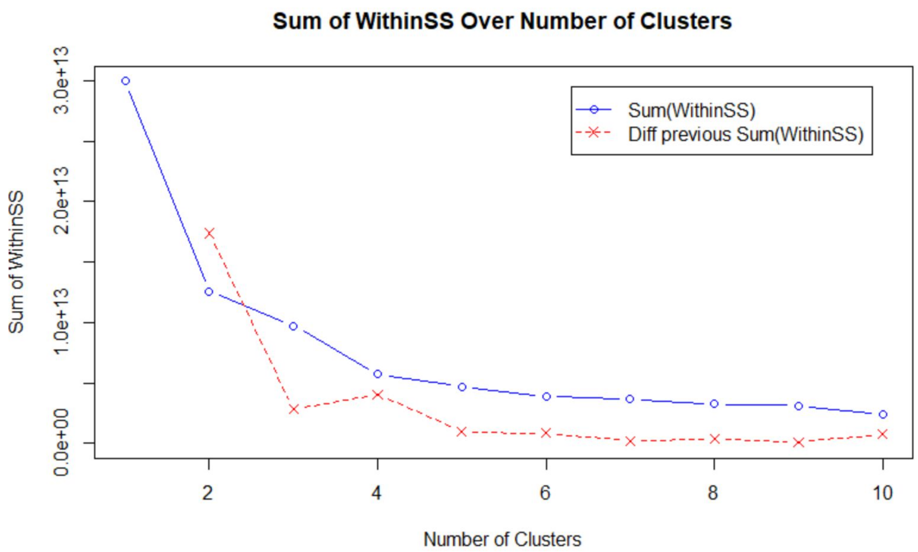
Modeling--classification(Random forest)

Modeling--classification(Decision tree)

	xerror	xstd
1	1.00000	0.015058
2	0.79650	0.013885
3	0.68025	0.013061
4	0.67421	0.013015
5	0.66395	0.012935
6	0.65640	0.012876
7	0.64130	0.012756
8	0.63345	0.012692
9	0.61987	0.012580
10	0.61111	0.012507
11	0.60930	0.012492
12	0.60779	0.012479
13	0.60568	0.012461
14	0.59390	0.012361
15	0.59300	0.012353
16	0.59028	0.012330
17	0.58907	0.012319
18	0.58756	0.012306
19	0.58545	0.012288
20	0.58122	0.012251
21	0.57790	0.012222
22	0.56884	0.012142
23	0.57307	0.012179



Modeling--classification(KNN)




```
Cluster sizes:

[1] "4203 427 3968 4701"

Data means:

     bedrooms     bathrooms
  0.428152493   0.335810545
  sqft_living      sqft_lot
  0.261844258   0.008253465
       floors    sqft_above
  0.185397398   0.236493493
sqft_basement      yr_built
  0.114728141   0.619460764
  yr_renovated      zipcode
  0.031786378   0.392773513
          lat          long
  0.637214823   0.253579026
sqft_living15    sqft_lot15
  0.337135452   0.020618840
```

	floors	sqft_above
1	0.07475613	0.1530376
2	0.16861827	0.2282345
3	0.01864919	0.1884912
4	0.42659009	0.3523762

	sqft_basement	yr_built
1	0.14972249	0.3873217
2	0.14322972	0.3422258
3	0.15729346	0.6130084
4	0.04492371	0.8576362

	yr_renovated	zipcode
1	0.0000000	0.6832073
2	0.9899931	0.4659830
3	0.0000000	0.1704075
4	0.0000000	0.3141509

	lat	long
1	0.7275900	0.1581885
2	0.6522842	0.2220055
3	0.5568037	0.3001399
4	0.6229179	0.3024314

Modeling--classification(Evaluation)

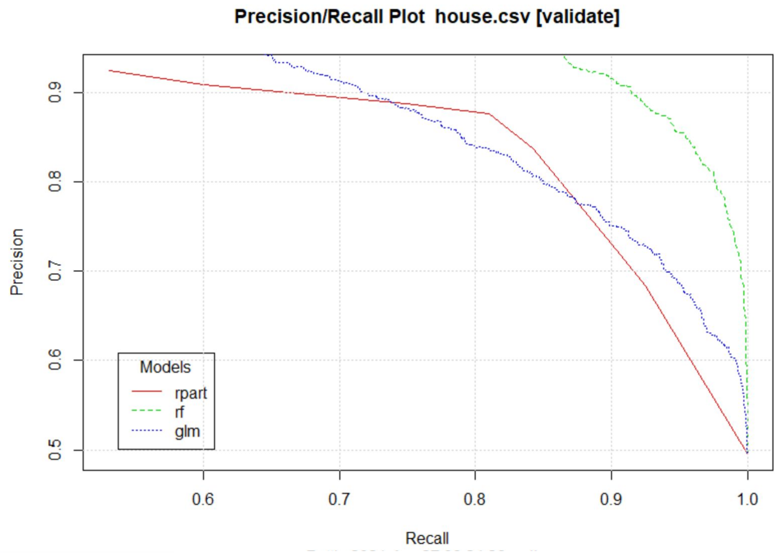
```
Error matrix for the Decision Tree model on midrange_homes.csv [validate]

        Predicted
Actual   high low medium Error
  high    443   0    175  28.3
  low       0  33     69  67.6
  medium  119  26   1984   6.8

Error matrix for the Decision Tree model on midrange_homes.csv [validate]

        Predicted
Actual   high low medium Error
  high   15.5 0.0    6.1  28.3
  low     0.0 1.2    2.4  67.6
  medium  4.2 0.9   69.6   6.8

Overall error: 13.7%, Averaged class error: 34.23333%
```

```
==========================================================================
Error matrix for the Linear model on midrange_homes.csv [validate]

        Predicted
Actual   high low medium Error
  high    365   0    253  40.9
  low       0  25     77  75.5
  medium  124  16   1989   6.6

Error matrix for the Linear model on midrange_homes.csv [validate]

        Predicted
Actual   high low medium Error
  high   12.8 0.0    8.9  40.9
  low     0.0 0.9    2.7  75.5
  medium  4.4 0.6   69.8   6.6

Overall error: 16.5%, Averaged class error: 41%
```

```
======================================================================
Error matrix for the Random Forest model on midrange_homes.csv [validate]

        Predicted
Actual   high low medium Error
  high    473   0    145  23.5
  low       0  36     66  64.7
  medium   76  23   2030   4.7

Error matrix for the Random Forest model on midrange_homes.csv [validate]

        Predicted
Actual   high low medium Error
  high   16.6 0.0    5.1  23.5
  low     0.0 1.3    2.3  64.7
  medium  2.7 0.8   71.3   4.7

Overall error: 10.8%, Averaged class error: 30.96667%
```

Modeling--classification(Future improvement)

- More range more error on predict price
- Only the low and high can predict price more accurately up 0.9.
- Drawback: The price around median is kinds of similar to each other, but it be define to low or high(that’s not good)




```
Area under the ROC curve for the rpart model on house.csv [validate] is 0.8864

Rattle timestamp: 2021-04-27 08:24:03 weikz
======================================================================
Area under the ROC curve for the rf model on house.csv [validate] is 0.9729

Rattle timestamp: 2021-04-27 08:24:03 weikz
======================================================================
Area under the ROC curve for the glm model on house.csv [validate] is 0.9153

Rattle timestamp: 2021-04-27 08:24:04 weikz
======================================================================
```

Question?