DmV

Data Mining and Visualization

PREDICTIVE MODELLING

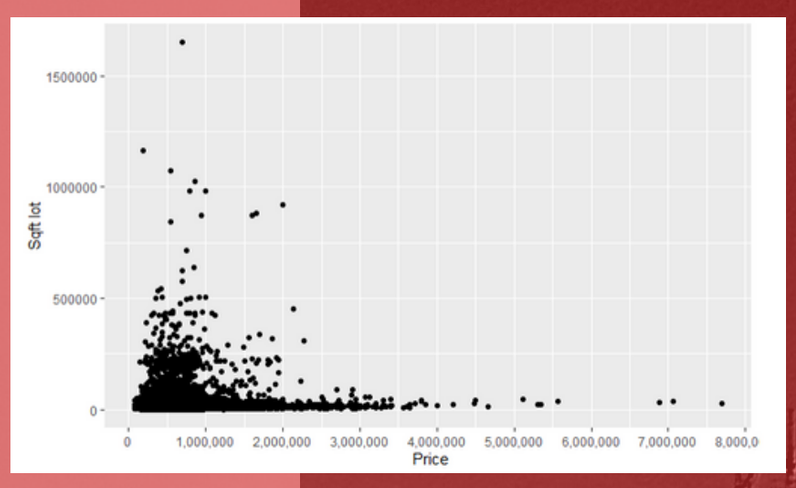
problem

- many house prices exceed the market desire
- to do the building effectively

reason why we choose the data of house sales

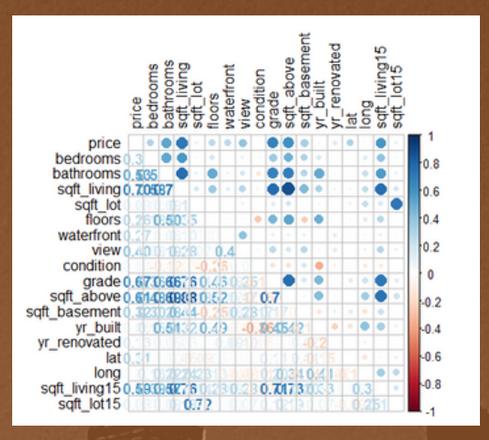
because we want to help people to get the best price to sell or even buy their house

plot



From this plot it is known that the land area and land price are not as significant as the building area

CORRELATIONS OF ALL NUMERIC VARIABLES



Highest correlation of price is seen with sqft_living,grade and sqft_above

ROW AND COLUMN OF DATA IS

```
[1] 21613    21
```

	id <dbl>	date <chr>	price <dbl>	bedrooms <int>	bathrooms <dbl>	sqft_living <int>
21608	2997800021	20150219T000000	475000	3	2.50	1310
21609	263000018	20140521T000000	360000	3	2.50	1530
21610	6600060120	20150223T000000	400000	4	2.50	2310
21611	1523300141	20140623T000000	402101	2	0.75	1020
21612	291310100	20150116T000000	400000	3	2.50	1600
21613	1523300157	20141015T000000	325000	2	0.75	1020

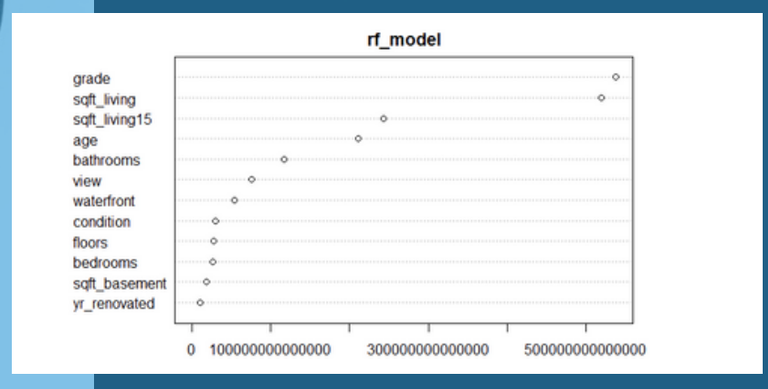
from the data plot we know that the majority of the selling price of the house is in the range of 200-250 million

summary

```
'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 ...
```

The data variables we have are 21, the data types are numeric and integer

rf_model

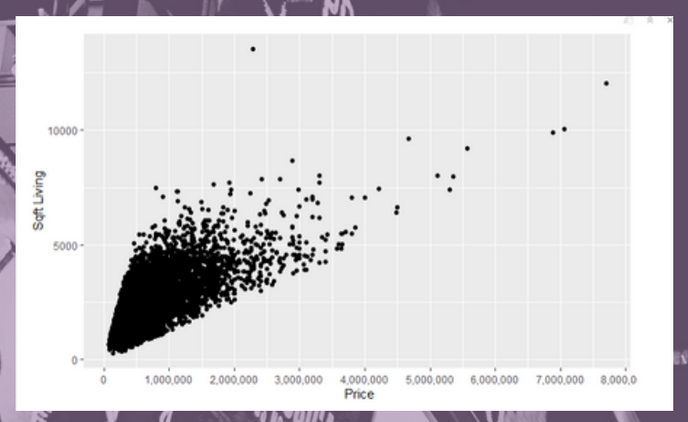


the model showcases the level of importance of each variable in explaining the variability in the response variable. The importance of variables is measured based on their contributions to the model's performance in predicting the response variable.

MISSING VALUE

```
            id          date         price      bedrooms     bathrooms
             0             0             0             0             0
   sqft_living      sqft_lot        floors    waterfront          view
             0             0             0             0             0
     condition         grade    sqft_above sqft_basement      yr_built
             0             0             0             0             0
  yr_renovated       zipcode           lat          long sqft_living15
             0             0             0             0             0
    sqft_lot15
             0
```

In this data, there's no missing value. it means that the data is complete, very profitable, because the analysis is more accurate and complete



square feet is very influential for the selling price of the house, the bigger it is, the higher the price

BUILD LINEAR REGRESSION MODEL

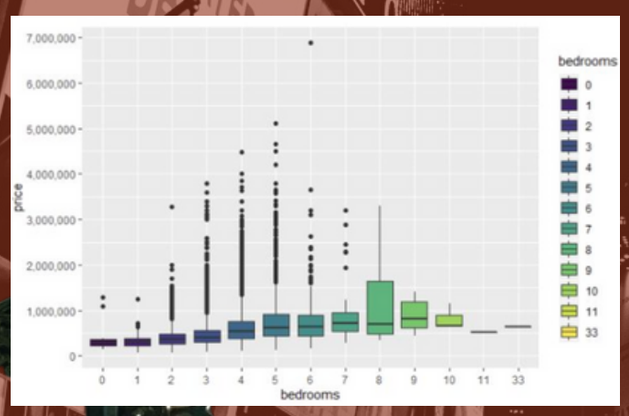
```
call:
lm(formula = price ~ sqft_living, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-1263979  -147372   -23668   106145  4346863

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) -46065.523   5307.196   -8.68 <0.0000000000000002 ***
sqft_living    282.092      2.338  120.67 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 263600 on 15201 degrees of freedom
Multiple R-squared:  0.4893,    Adjusted R-squared:  0.4892
F-statistic: 1.456e+04 on 1 and 15201 DF,  p-value: < 0.00000000000000022
```

Linear regression model on this single variable (sqft_living) alone explains almost 49% of variance in the data!



Bedrooms does not show a linear relationship with price, hence they should be kept as factor and hence a dummy variable will be created in linear regression for the same

KELOMPOK 4 : Data Science

1. **2602205736 - Salma Khaira Almuna - salma.almuna@binus.ac.id**
2. **2602205540 - Annaura Zyra Alifa W - annaura.windutomo@binus.ac.id**
3. **2602224230 - Johan - johan009@binus.ac.id**
4. **2602183760 - Athia Zahra - athia.zahra@binus.ac.id**
5. **2602170594 - Debby Syafira Wijaya - debby.wijaya@binus.ac.id**

BINUS UNIVERSITY