HOUSES PRICE ANALYSIS

USING R PROGRAMMING

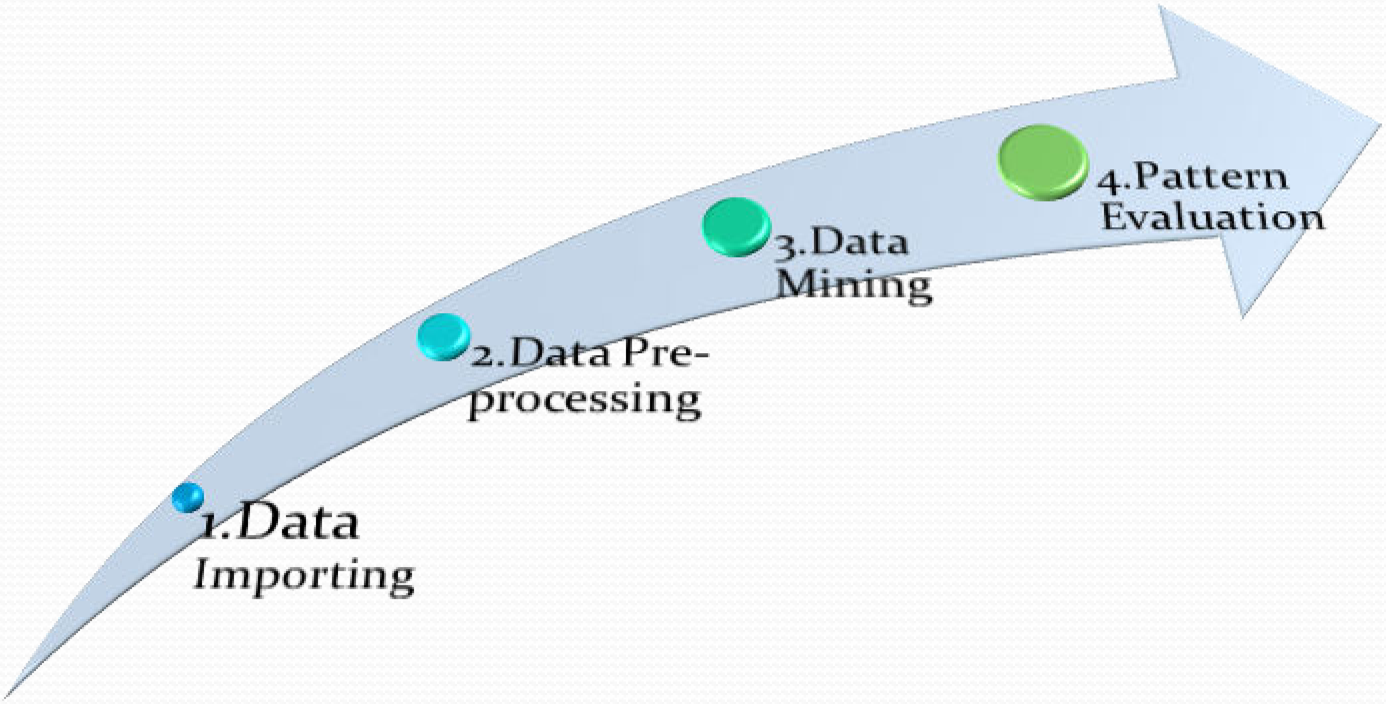
PROBLEM STATEMENT

We have the “House_for_Sale” dataset, which constitues of entries such as price of houses, lot size, number of rooms, living area etc.

- We are supposed to:

  understand the data-set and design a model which will help in predicting the prices of houses.

Tasks to be performed



Tasks to be Performed

Data Importing	Data-Pre processing	Data Mining	Pattern Evaluation
• Import the "House for sale" dataset	• Understand the structure of data and find correlation between different entities	• Use Linear regression to predict the rates of houses	• Evaluate which model is better for dataset

1.Data Cleaning: R codes

```
read.csv("C:/Users/user/Desktop/houses.csv")->houses
str(houses)
#data cleaning
library(dplyr)
houses%>%select(c(-1,-2))->houses
houses
```

```
> read.csv("C:/Users/user/Desktop/houses.csv")->houses
> str(houses)
'data.frame':	1728 obs. of  16 variables:
 $ X.1          : int  1 2 3 4 5 6 7 8 9 10 ...
 $ X            : int  1 2 3 4 5 6 7 8 9 10 ...
 $ price        : int  132500 181115 109000 155000 86060 120000 153000 170000 90000 1
00 ...
 $ lot_size     : num  0.09 0.92 0.19 0.41 0.11 0.68 0.4 1.21 0.83 1.94 ...
 $ waterfront   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ age          : int  42 0 133 13 0 31 33 23 36 4 ...
```

The CSV file consists of houses data set. First two columns x.1 and x are numberings. So we need to remove the columns to create a net data set.

Pre-processing: Using R code

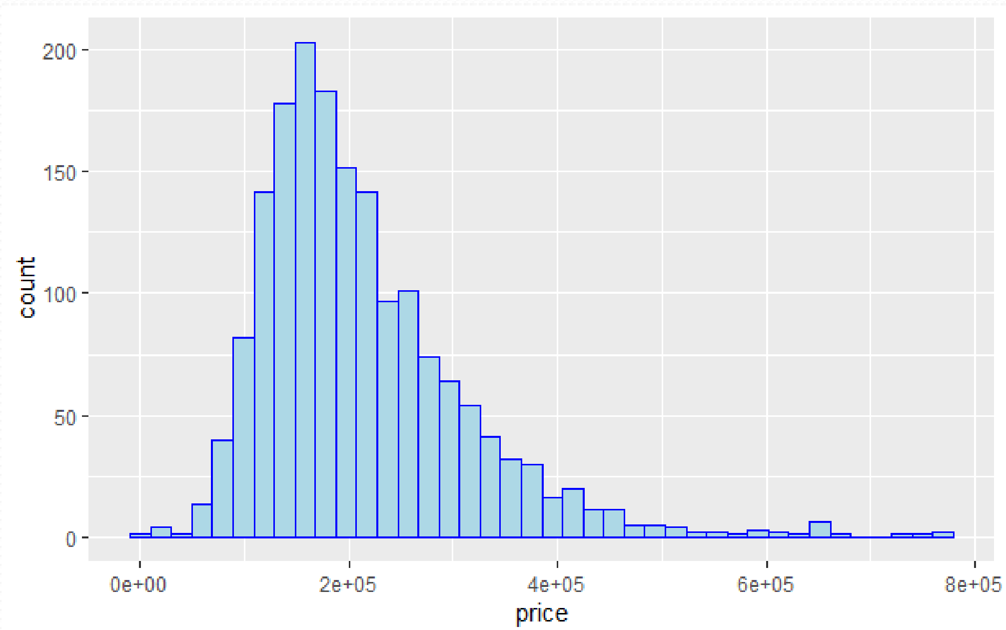
```
houses$air_cond<-factor(houses$air_cond,labels=c("NO","YES"))
houses$construction<-factor(houses$construction, labels =c("NO","YES"))
houses$waterfront<-factor(houses$waterfront, labels = c("NO","YES"))
houses$fuel<-factor(houses$heat, labels=c("Gas","Electric","Oil"))
houses$sewer<-factor(houses$sewer, labels = c("None","Privet","Public"))
houses$heat<-factor(houses$heat, labels=c("Hot Air","Hot water","Electric"))
```

```
> houses$air_cond<-factor(houses$air_cond,labels=c("NO","YES"))
> houses$construction<-factor(houses$construction, labels =c("NO","YES"))
> houses$waterfront<-factor(houses$waterfront, labels = c("NO","YES"))
> houses$fuel<-factor(houses$heat, labels=c("Gas","Electric","Oil"))
> houses$sewer<-factor(houses$sewer, labels = c("None","Privet","Public"))
> houses$heat<-factor(houses$heat, labels=c("Hot Air","Hot water","Electric"))
> houses
   price lot_size waterfront age land_value construction air_cond     fuel
1 132500     0.09         NO  42      50000           NO       NO      Oil
2 181115     0.92         NO   0      22300           NO       NO Electric
3 109000     0.19         NO 133       7300           NO       NO Electric
4 155000     0.41         NO  13      18700           NO       NO      Gas
5  86060     0.11         NO   0      15000          YES      YES      Gas
6 120000     0.68         NO  31      14000           NO       NO      Gas
7 153000     0.40         NO  33      23300           NO       NO Electric
8 170000     1.21         NO  23      14600           NO       NO      Gas
```

Now if house has waterfront we can change 1 to "Yes" and if it does not have water front we can change 0 to "NO". Similarly we can change values to yes and no for construction. For Fuel type instead of 2,3,4 we can use electric , gas and air. And so on.

Visualization: Using R code

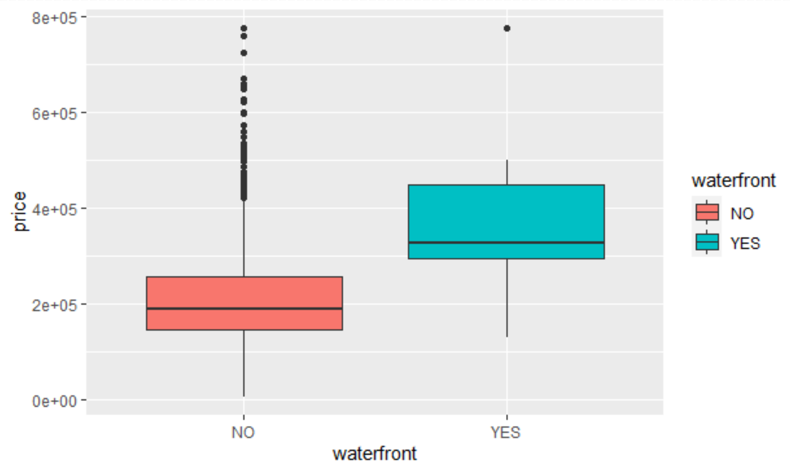
```
•library(ggplot2)
•ggplot(data=houses,aes(x=price))+geom_histogram(bins=40)
•ggplot(data=houses,aes(x=price))+geom_histogram(bins=40,fill=
"lightblue",col="blue")
```

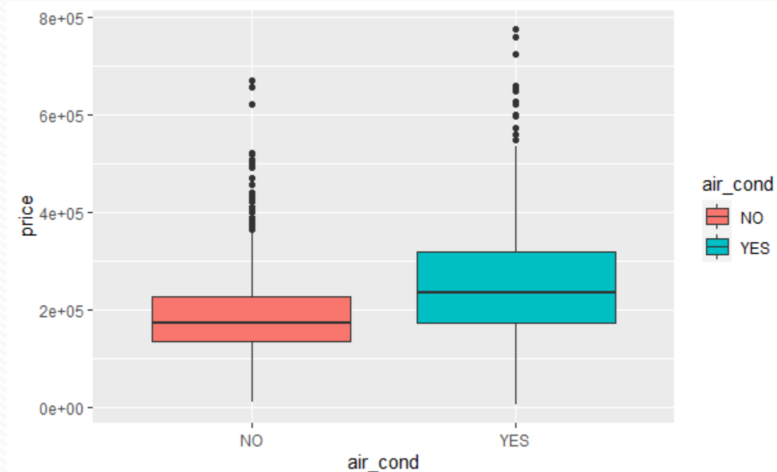


Distribution of Price. From histogram we can say avg. Price of a house is 2 lakhs and maximum price will be around 7.5 lakhs.

Visualization: Using R code

- ggplot(data=houses,aes(y=price,x=waterfront,fill=waterfront))+geom_boxplot()
- ggplot(data=houses,aes(y=price,x=air_cond,fill=air_cond))+geom_boxplot()

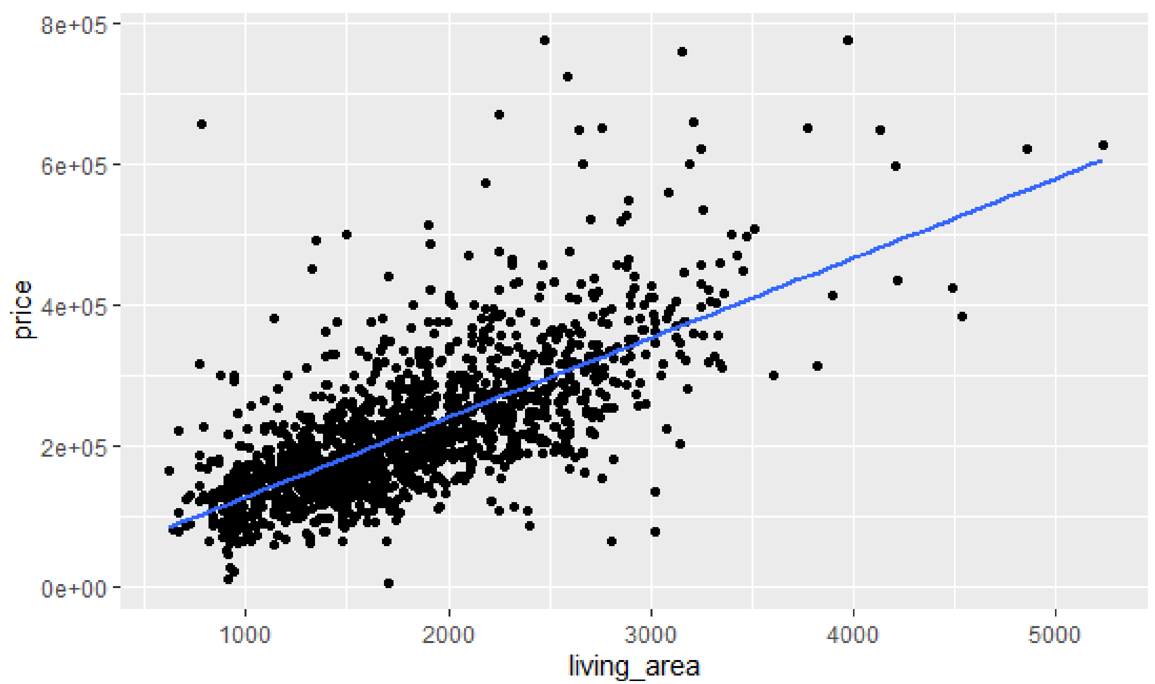




Water front has two categories so it gives two colours by default. From box plot it is clear that if a house has a water front it has high price. Same as before we can see house with air conditioning gas high value.

Visualization: Using R code

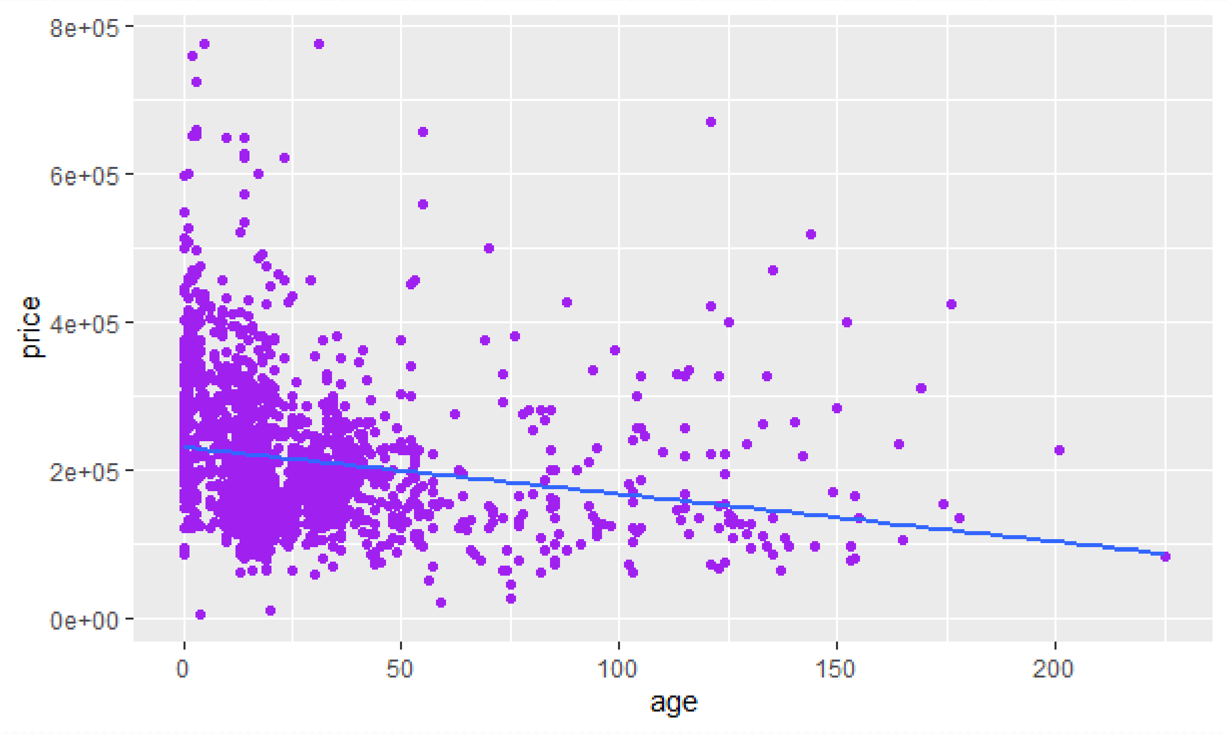
```
•ggplot(houses,aes(x=living_area,y=price))+geom_point()+geom_smooth(method = "lm",se=F)
```



To see how price varies with living area we use scatter plot and a line. We van see that if area of living area increases price also increases, almost a linear relationship.

Visualization: Using R code

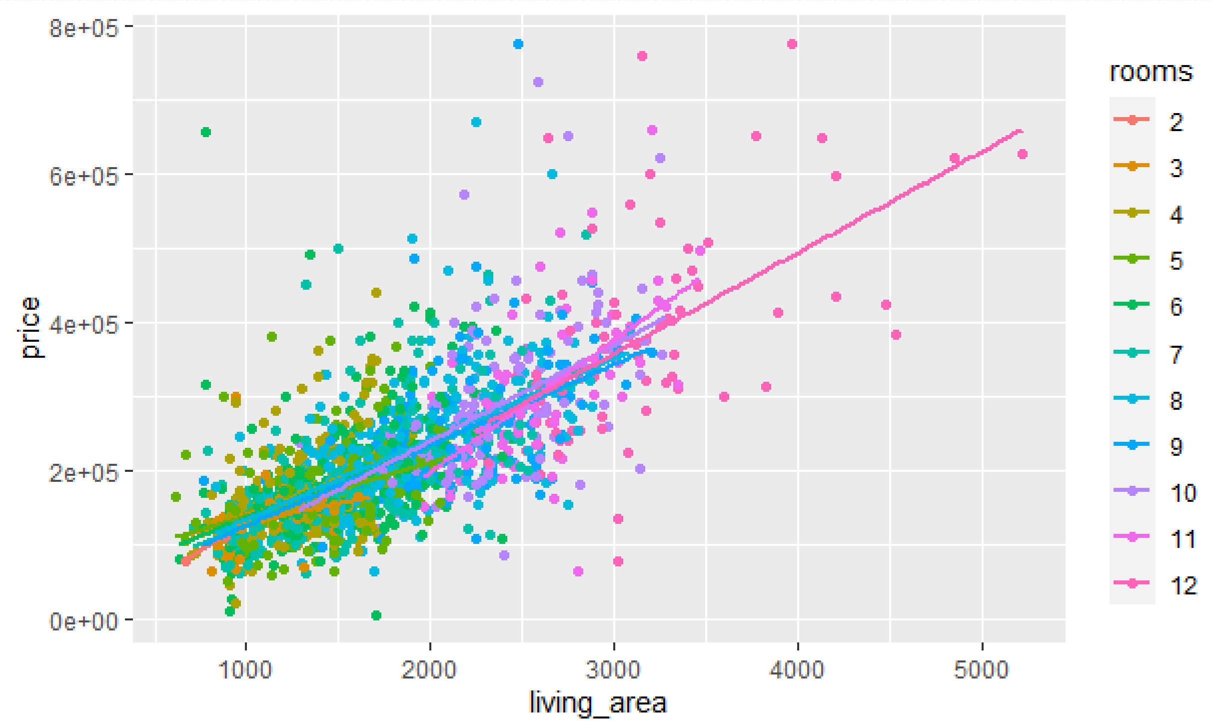
•ggplot(houses,aes(x=age,y=price))+geom_point(col="purple")+geom_smooth(method = "lm",se=F)



Price vs age of the house: we can see this is inverse relationship. If house is old price is low. And if house is new then price is high.

Visualization: Using R code

•ggplot(houses,aes(x=living_area,y=price,col=factor(rooms)))+geom_point()+geom_smooth(method ="lm",se=F)+labs(col="rooms")



For this graph y axis is price and x axis is living area and colours are determines by number of rooms. From graph wee see if a house has 4-6 rooms price will be 1.5 lakh to 4.5 lakh.

Visualization: Using R code

```
•ggplot(houses,aes(x=age,y=price,col=factor(rooms)))+geom
_point()+geom_smooth(method
="lm",se=F)+labs(col="rooms")
```

Here x axis represent age , y axis price and colours are rooms.

Splitting data: R code

```
•library(caTools)
•sample.split(houses$price,SplitRatio= 0.65)->split_index
•train<-subset(houses,split_index==T)
•test<-subset(houses,split_index==F)
•nrow(train)
•nrow(test)
```

We need to split our data between training and testing data set with split ratio 0.65. We do this because it helps us to measure the accuracy of the model. We build our model based on training set and test its accuracy by testing set.

Building First 1st Model : R code

- mod1<-lm(price~.,data=train)
- predict(mod1,test)->result
- compare_result<-cbind(actual=test$price,predicted=result)
- as.data.frame(compare_result)->compare_result
- error<-compare_result$actual-compare_result$predicted
- cbind(compare_result,error)->compare_result
- sqrt(mean(compare_result$error^2))->rmse1
- rmse1

```
> compare_result<-cbind(actual=test$price,predicted=result)
> compare_result<-cbind(actual=test$price,predicted=result)
> as.data.frame(compare_result)->compare_result
> error<-compare_result$actual-compare_result$predicted
> cbind(compare_result,error)->compare_result
> sqrt(mean(compare_result$error^2))->rmse1
> rmse1
[1] 54313.59
```

Here root mean square error is 54313.59

Analysis of ANOVA Table:

summary(mod1)

```
Call:
lm(formula = price ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-232603  -35502   -5204   28127  456021

Coefficients: (2 not defined because of singularities)
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -6.067e+03  2.402e+04  -0.253   0.8007
lot_size        6.111e+03  2.509e+03   2.435   0.0150 *
waterfrontYES   1.293e+05  1.853e+04   6.977 4.87e-12 ***
age            -1.446e+02  6.692e+01  -2.161   0.0309 *
land_value      9.643e-01  5.467e-02  17.638  < 2e-16 ***
constructionYES -4.950e+04 7.918e+03  -6.252 5.53e-10 ***
air_condYES     1.226e+04  4.109e+03   2.984   0.0029 **
fuelElectric   -9.595e+03  5.053e+03  -1.899   0.0578 .
fueloil        -1.071e+04  4.888e+03  -2.192   0.0286 *
```

Stars tell us how much impact one independent variable has on dependent variable. Greater no of stars means greater impact. For example if a house has a water front it or newly constructed will have a great impact on price of the house. But heat, sewer , fuel they don't have any significant effect on price house. Also Value of adjusted R^2 indicates the accuracy. More closer the value is compared to 1 more accuracy the model has. Here value is 0.656

Building First 2nd Model : R code

- mod2<-lm(price~.-fireplaces-sewer-fuel,data=train)
- predict(mod2,test)->result2
- compare_result2<-cbind(actual=test$price,predicted=result2)
- compare_result2
- as.data.frame(compare_result2)->compare_result2
- error<-compare_result2$actual-compare_result2$predicted
- cbind(compare_result2,error)->compare_result2
- sqrt(mean(compare_result2$error^2))->rmse2
- rmse2

```
> mod2<-lm(price~.-fireplaces-sewer-fuel,data=train)
> predict(mod2,test)->result2
> compare_result2<-cbind(actual=test$price,predicted=result2)
> as.data.frame(compare_result2)->compare_result2
> error<-compare_result2$actual-compare_result2$predicted
> cbind(compare_result2,error)->compare_result2
> sqrt(mean(compare_result2$error^2))->rmse2
> rmse2
[1] 54213.83
> 
```

For this model we are excluding fireplaces, fuel and sewer as they have insignificance impact on price. We need to compare this model with previous model to see which model is better. Here RMSE is 54213.83

Analysis of ANOVA Table:

summary(mod2)

```
> summary(mod2)

Call:
lm(formula = price ~ . - fireplaces - sewer - fuel, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-233128  -35516   -5134   28943  455607 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      4.291e+03  7.336e+03   0.585  0.55873    
lot_size         6.699e+03  2.305e+03   2.906  0.00373 ** 
waterfrontYES    1.300e+05  1.849e+04   7.033 3.30e-12 ***
age             -1.391e+02  6.645e+01  -2.093  0.03651 *  
land_value       9.603e-01  5.422e-02  17.712  < 2e-16 ***
constructionYES -4.915e+04  7.863e+03  -6.251 5.56e-10 ***
air_condYES      1.190e+04  4.050e+03   2.937  0.00337 ** 
heatHot water   -9.814e+03  5.021e+03  -1.954  0.05086 .  
heatElectric    -1.082e+04  4.862e+03  -2.225  0.02628 *  
living_area      6.135e+01  5.235e+00  11.720  < 2e-16 ***
bathrooms        2.717e+04  3.986e+03   6.815 1.45e-11 ***
rooms            2.329e+03  1.077e+03   2.163  0.03073 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59850 on 1261 degrees of freedom
Multiple R-squared:  0.6601,	Adjusted R-squared:  0.6571 
F-statistic: 222.6 on 11 and 1261 DF,  p-value: < 2.2e-16
```

Wee see that Adjusted R square value increases to 0.6571 from 0.656. So this model is better than previous one.

Conclusion

- Wee see that Root Mean square error for first model is 54313.59 and for second one is 54213.83. AS we see that error for second model is lesser than first one so model two is better than model one in this scenario.