Data Analysis Project

Arijit Naskar & Spandan Ghoshal

Introduction

In this project, we analyze the "House Sales in King County, USA" dataset. This dataset contains house sale prices for King County, which includes Seattle during the time period May 2014 and May 2015. The dataset observations on 21613 many observations on the variables :-

price : Price is prediction target

bedrooms : Number of bedrooms

bathrooms : Number of bathrooms

sqft_living : Square footage of the home

sqft_lot : Square footage of the lot

floors : Total floors (levels) in house

waterfront : House which has a view to a waterfront

view : Has been viewed

condition : How good the condition is overall

grade : overall grade given to the housing unit, based on King County grading system

sqft_above : Square footage of house apart from basement

sqft_basement : Square footage of the basement

yr_built : Built Year

yr_renovated : Year when house was renovated

We applied various statistical tools to predict the price of a house. We divided the data analysis into two parts. Firstly we performed the exploratory analysis where we analyzed the data by drawing suitable diagrams to depict some dependence of price on different predictors and then made some crucial observations from there. Finally, to make conclusions based on statistical tools i.e testing of hypothesis, inference, estimation and various other regression methods, we answered to some of the observations we made earlier.

Importing the dataset

We import the dataset using R as follows :-

```
Data_House = read.table("E:\\Dekstop\\Project_Data_analysis_IITK\\kc_house_data_to_be_
head(Data_House)

    price bedrooms bathrooms sqft_living sqft_lot floors waterfront view
1  221900        3      1.00        1180     5650      1          0    0
2  538000        3      2.25        2570     7242      2          0    0
3  180000        2      1.00         770    10000      1          0    0
4  604000        4      3.00        1960     5000      1          0    0
5  510000        3      2.00        1680     8080      1          0    0
6 1230000        4      4.50        5420   101930      1          0    0
  condition grade sqft_above sqft_basement yr_built yr_renovated
1         3     7       1180             0     1955            0
2         3     7       2170           400     1951         1991
3         3     6        770             0     1933            0
4         5     7       1050           910     1965            0
5         3     8       1680             0     1987            0
6         3    11       3890          1530     2001            0
```

Checking presence of missing observations

We use the function is.na() to detect whether there is any missing observation in the given dataset or not as :-

```
sum(is.na(Data_House))

[1] 0
```

since there is no such missing observation, hence, we don't need to apply any data imputation techniques here and it makes the analysis little simpler.

Data type of each predictors

First we check the names of each and every variable present in teh dataset as :-

```
vec_names = names(Data_House)
vec_names

 [1] "price"         "bedrooms"      "bathrooms"     "sqft_living"
 [5] "sqft_lot"      "floors"        "waterfront"    "view"
 [9] "condition"     "grade"         "sqft_above"    "sqft_basement"
[13] "yr_built"      "yr_renovated"
```

To get an idea about each of them, we use the summary() function :-

```
data_summary = summary(Data_House)
data_summary
```

```
     price            bedrooms        bathrooms      sqft_living
 Min.   :  75000   Min.   : 0.000   Min.   :0.000   Min.   :  290
 1st Qu.: 321950   1st Qu.: 3.000   1st Qu.:1.750   1st Qu.: 1427
 Median : 450000   Median : 3.000   Median :2.250   Median : 1910
 Mean   : 540182   Mean   : 3.371   Mean   :2.115   Mean   : 2080
 3rd Qu.: 645000   3rd Qu.: 4.000   3rd Qu.:2.500   3rd Qu.: 2550
 Max.   :7700000   Max.   :33.000   Max.   :8.000   Max.   :13540
    sqft_lot           floors        waterfront           view
 Min.   :    520   Min.   :1.000   Min.   :0.000000   Min.   :0.0000
 1st Qu.:   5040   1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.0000
 Median :   7618   Median :1.500   Median :0.000000   Median :0.0000
 Mean   :  15107   Mean   :1.494   Mean   :0.007542   Mean   :0.2343
 3rd Qu.:  10688   3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.0000
 Max.   :1651359   Max.   :3.500   Max.   :1.000000   Max.   :4.0000
   condition         grade          sqft_above   sqft_basement
 Min.   :1.000   Min.   : 1.000   Min.   : 290   Min.   :   0.0
 1st Qu.:3.000   1st Qu.: 7.000   1st Qu.:1190   1st Qu.:   0.0
 Median :3.000   Median : 7.000   Median :1560   Median :   0.0
 Mean   :3.409   Mean   : 7.657   Mean   :1788   Mean   : 291.5
 3rd Qu.:4.000   3rd Qu.: 8.000   3rd Qu.:2210   3rd Qu.: 560.0
 Max.   :5.000   Max.   :13.000   Max.   :9410   Max.   :4820.0
    yr_built     yr_renovated
 Min.   :1900   Min.   :   0.0
 1st Qu.:1951   1st Qu.:   0.0
 Median :1975   Median :   0.0
 Mean   :1971   Mean   :  84.4
 3rd Qu.:1997   3rd Qu.:   0.0
 Max.   :2015   Max.   :2015.0
```

str() function also gives us some idea about the nature of each variable.

```
str(Data_House)
```

```
'data.frame': 21613 obs. of  14 variables:
 $ 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 ...
```

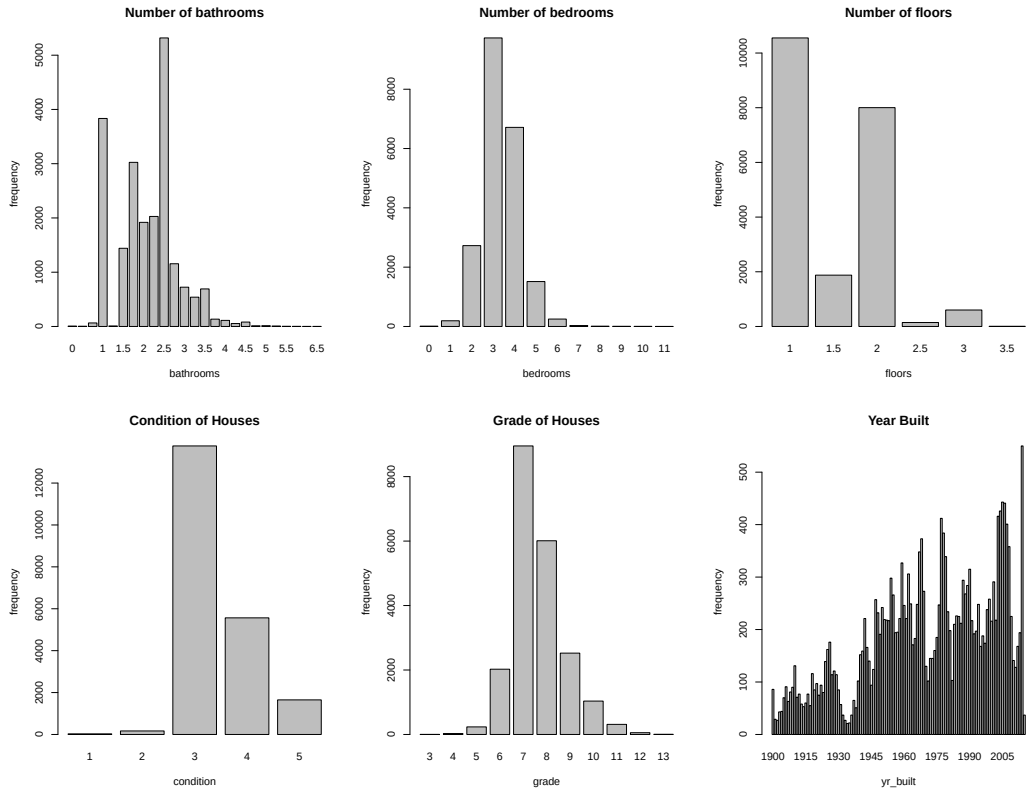
```
$ sqft_above   : int  1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...
$ sqft_basement: int  0 400 0 910 0 1530 0 0 730 0 ...
$ yr_built     : int  1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...
$ yr_renovated : int  0 1991 0 0 0 0 0 0 0 0 ...
```

Observation :-

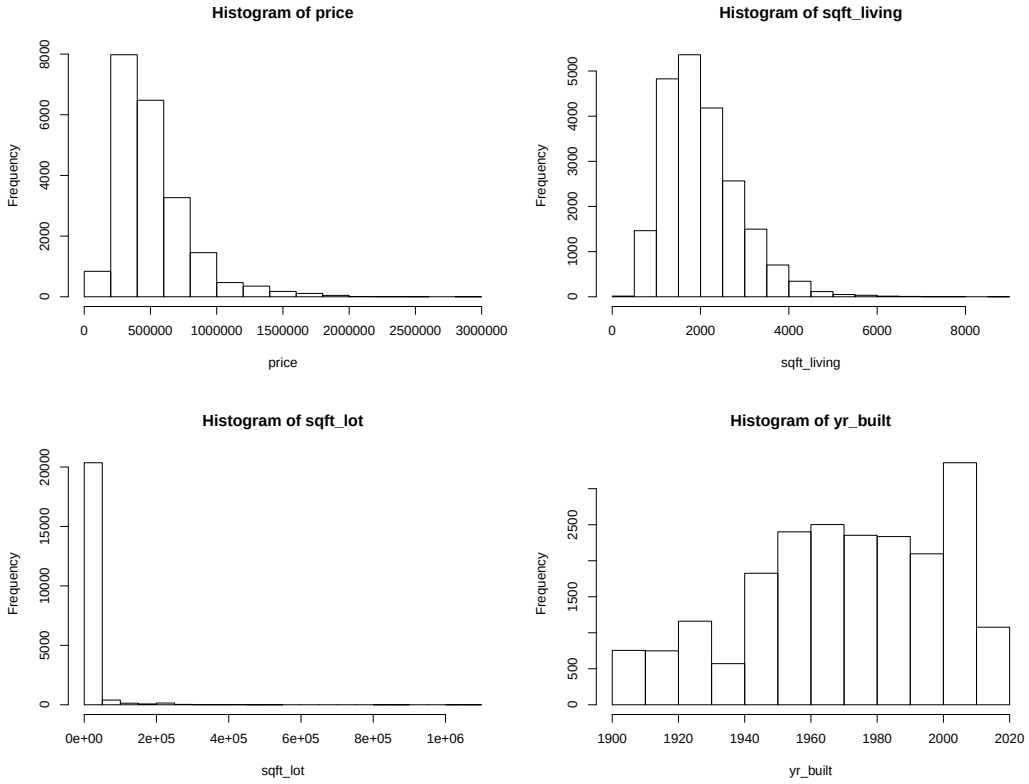
From the summary outputs for different variable, we can see that there might be potential outliers in the data as the ranges for some of the variables vary tremendously so checking presence of any outliers will be of great help in this case.

Barplot for different integer valued variables

We plot the bar diagrams for different count datas :-

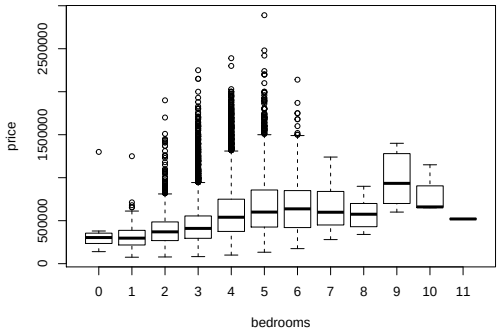


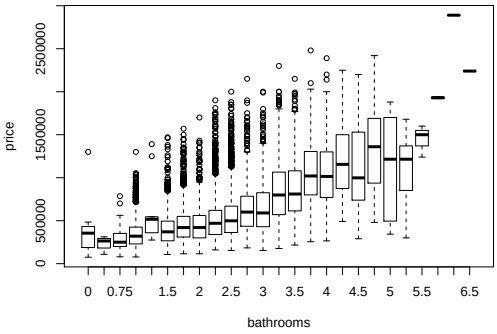
Histograms Plots

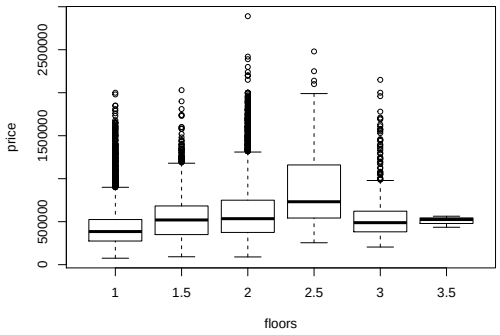


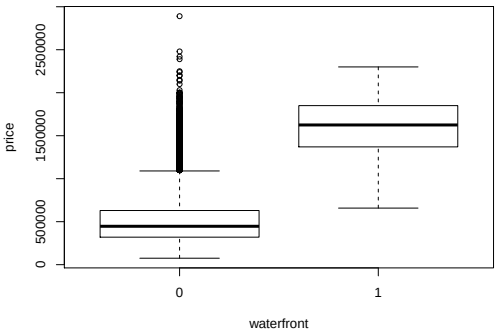
Boxplots

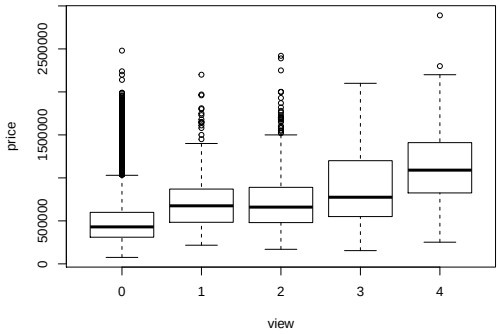
We plot the boxplots of price for different integer valued variables :-

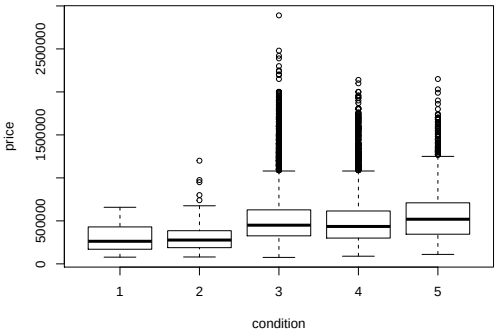

price
0
500000
1500000
2500000
0 1 2 3 4 5 6 7 8 9 10 11
bedrooms

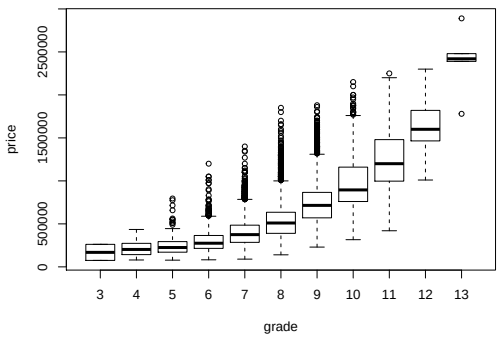

price
0
500000
1500000
2500000
0 0.75 1.5 2 2.5 3 3.5 4 4.5 5 5.5 6.5
bathrooms


price
0
500000
1500000
2500000
1 1.5 2 2.5 3 3.5
floors


price
0
500000
1500000
2500000
0 1
waterfront


price
0
500000
1500000
2500000
0 1 2 3 4
view


price
0
500000
1500000
2500000
1 2 3 4 5
condition


price
0
500000
1500000
2500000
3 4 5 6 7 8 9 10 11 12 13
grade

Observation :-

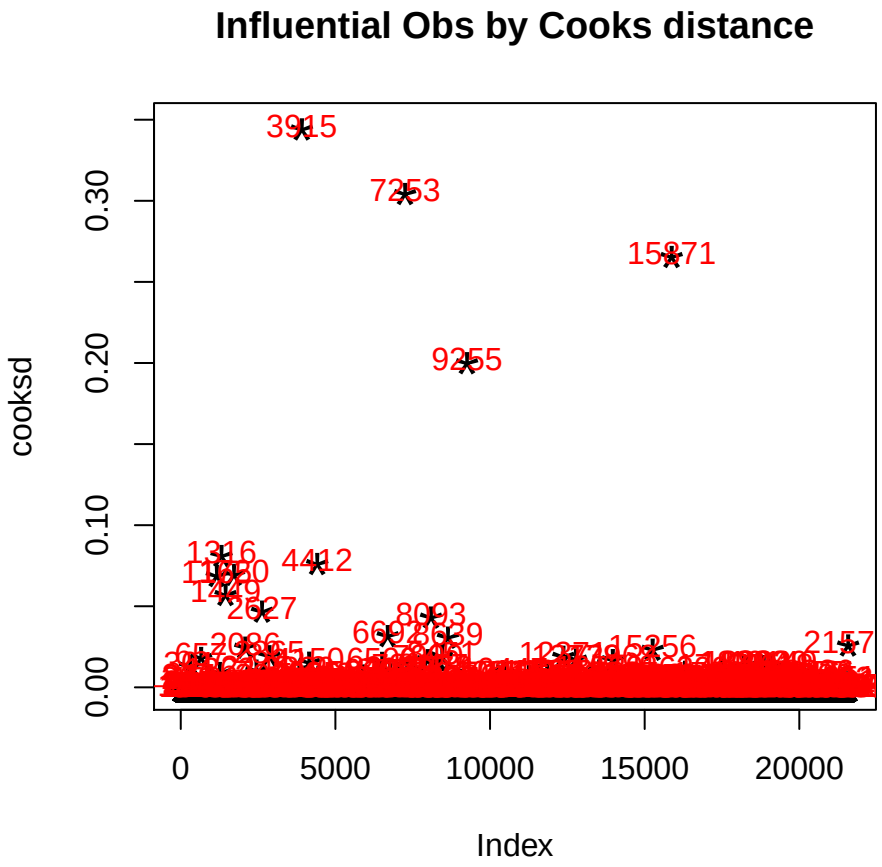
From the boxplots, we can see that in some of the plots, the price increases with increasing value of the other covariate like number of bedrooms, bathrooms, waterfront, view, condition, grade etc. We would perform ANOVA test to detect significant differences in price for different values of these covariates.

Confirmatory Data Analysis

Outlier Treatment

We use cook's distance for detecting any potential outlier in the data. We plot the cook's distance for each observation in a scatter plot w.r.t their indexes.

```
mod <- lm(price ~ ., data=Data_House)
cooksd <- cooks.distance(mod)
plot(cooksd, pch="*", cex=2, main="Influential Obs by Cooks distance")  # plot cook's
abline(h = 4*mean(cooksd, na.rm=T), col="red")  # add cutoff line
text(x=1:length(cooksd)+1, y=cooksd, labels=ifelse(cooksd>4*mean(cooksd, na.rm=T),nam
```

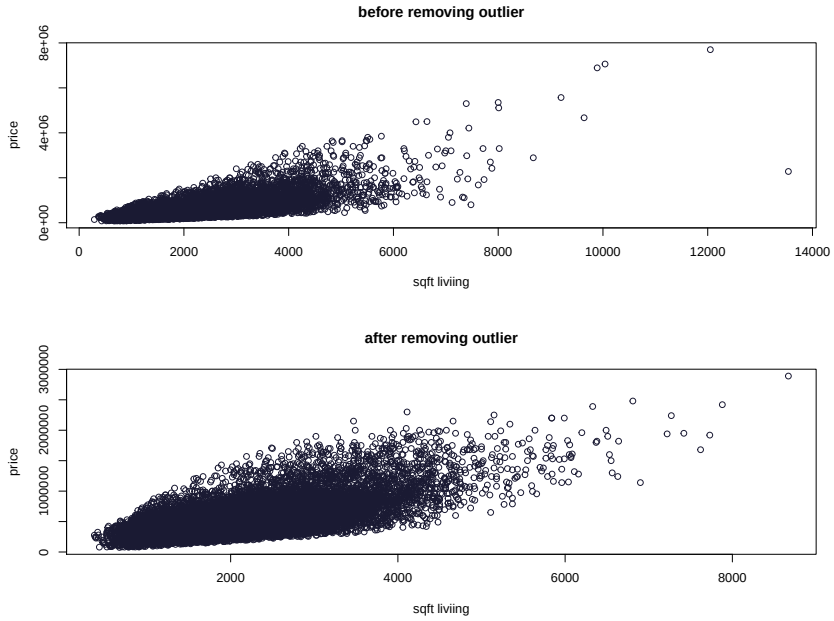


We detect potential by choosing those observations which are greater than 4 times average cook's distance. (some of the indexes of those outliers are given below)

```
ind_outlier = (cooksd>4*mean(cooksd, na.rm=T))
ind_out = which(ind_outlier == TRUE)
head(ind_out)

 22 116 154 231 240 247
 22 116 154 231 240 247
```

```
length(ind_out)

[1] 428

Data_House_1 = Data_House[-ind_out,]
dim(Data_House_1)

[1] 21185    14

attach(Data_House_1)
```

For comparison, we plot the scatter plots of price vs sqft living as shown :-



We should find the impact of the quantitative variables on the price of the houses separately.

Price vs Area of Basement

We obtain the impact of area of basement on the price of the houses.

```
lr=lm(price~sqft_basement)
summary(lr)


Call:
lm(formula = price ~ sqft_basement)

Residuals:
    Min      1Q  Median      3Q     Max
-474073 -181361  -62612  111493 1958639

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.614e+05  2.213e+03  208.47   <2e-16 ***
sqft_basement 1.877e+02  4.342e+00   43.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 268700 on 21183 degrees of freedom
Multiple R-squared:  0.08105,Adjusted R-squared:  0.081
F-statistic:  1868 on 1 and 21183 DF,  p-value: < 2.2e-16
```

The linear model is,

$$y_i = a + bx_i + e_i$$

where,y_i = Price of i th house and x = area of basement of the i th house, $i = 1, 2, .., 21613$.

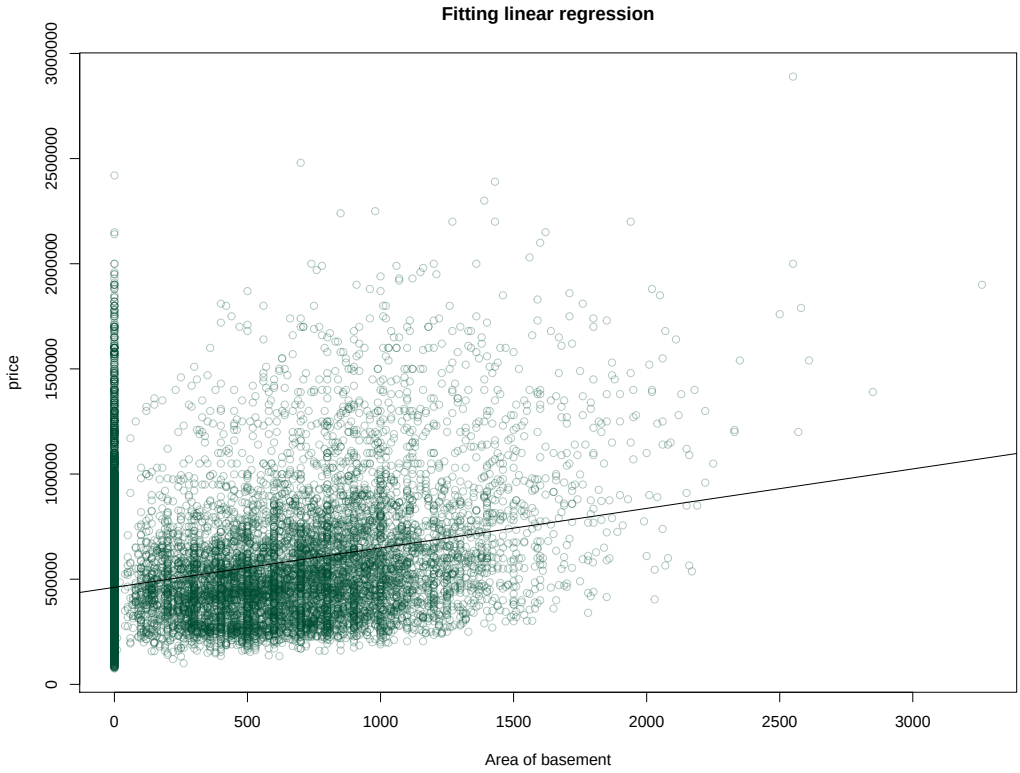
We set the null hypothesis,

$$H_0 : b = 0 \text{ vs } H_1 : b \neq 0$$

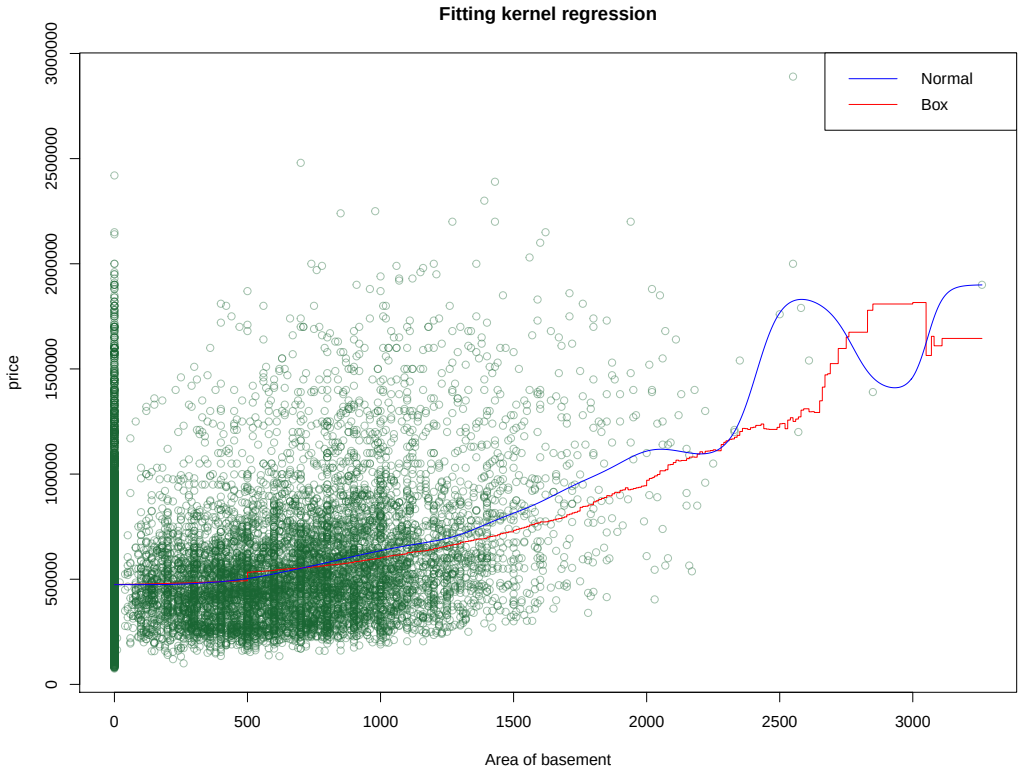
The p value of the hypothesis is less than 0.05.

Then we can conclude that **the area of basement of the houses have significant effect of the linear regression of price** .

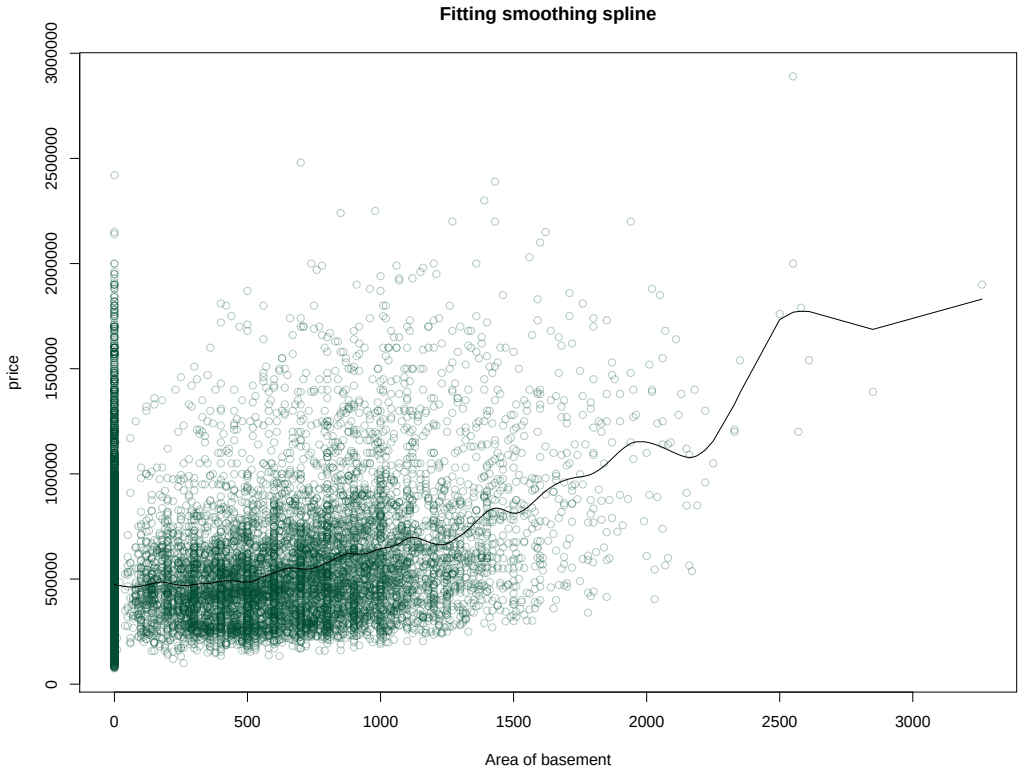
The fitted linear regression is as follows,



To get more precise relation between price and the area of the basements,we fitted both **Box kernel regression** and **Gaussian kernel regression.**



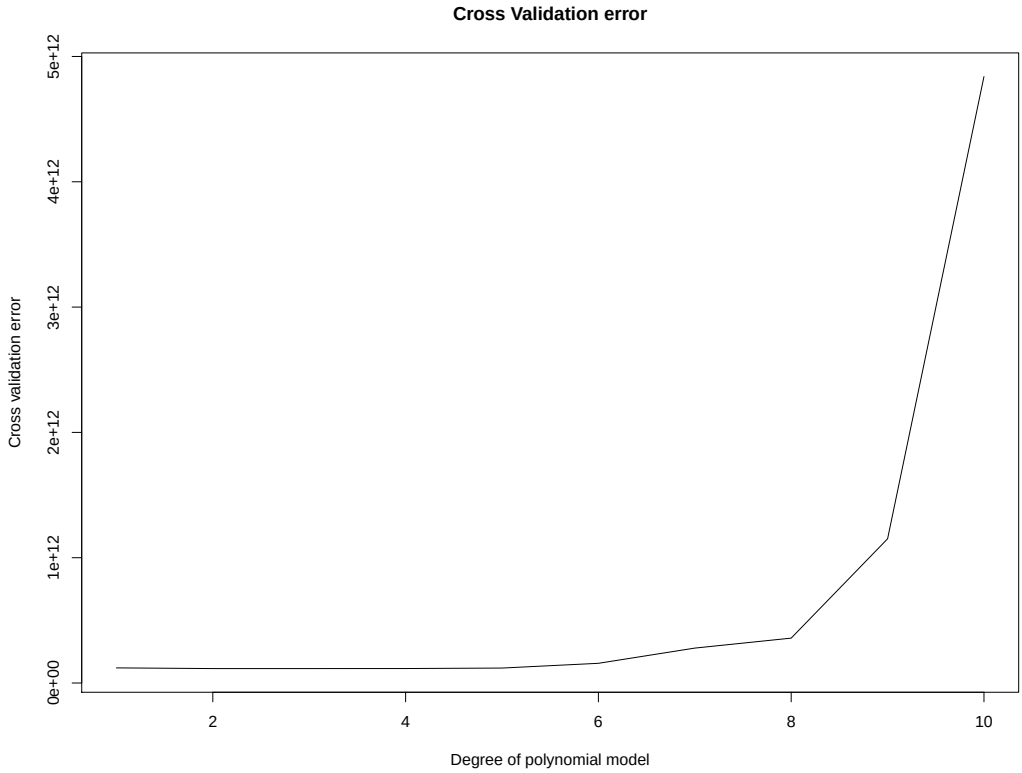
The fitted **spline regression** of the price on the area of basement using **smoothing spline,** is as follows,



We have the linear regression of the price on the area of basement. Further,we are interested in fitting a **polynomial regression** with appropriate degree.

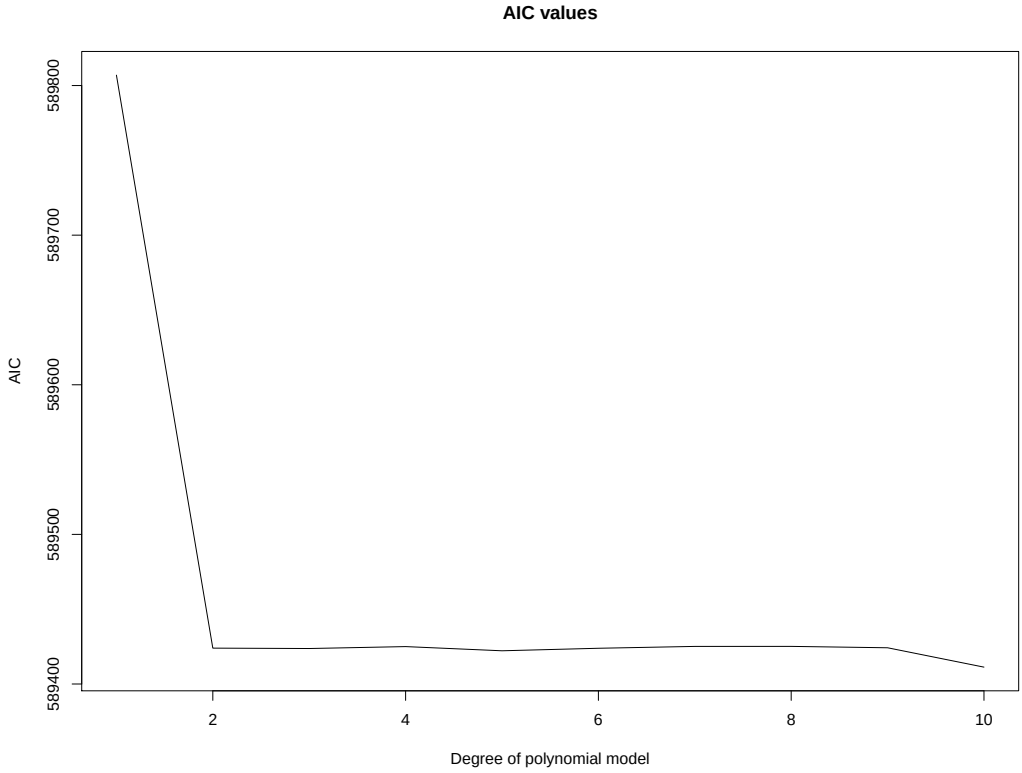
In order to determine the appropriate degree of the polynomial regression,we can fit the polynomial regression model with different degrees and then minimize the measure of optimism. We can obtain the measure of optimism using **Akaike information criterion(AIC), Bayesian information criterion(BIC)** and **Cross-validation** method**.**

The value of **Cross validation error** of the different degrees of polynomial model is as follows,



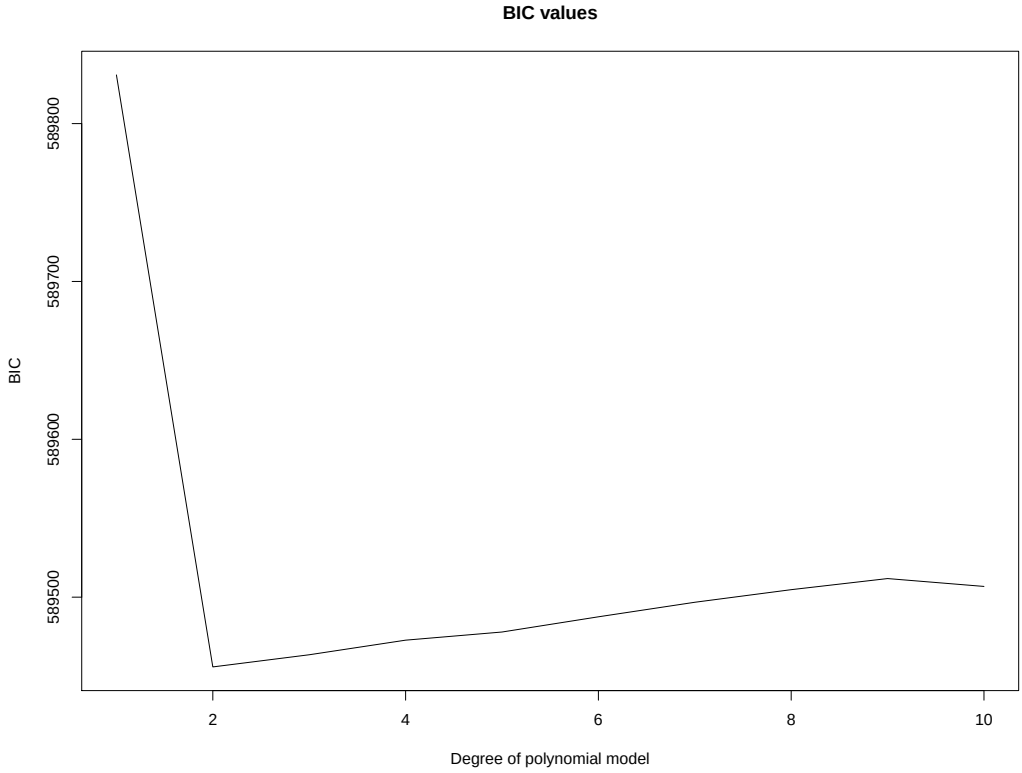
Cross Validation method suggest that we should fit a second degree **polynomial regression** of price on area of basement

The value of AIC of the different degrees of polynomial model is as follows,



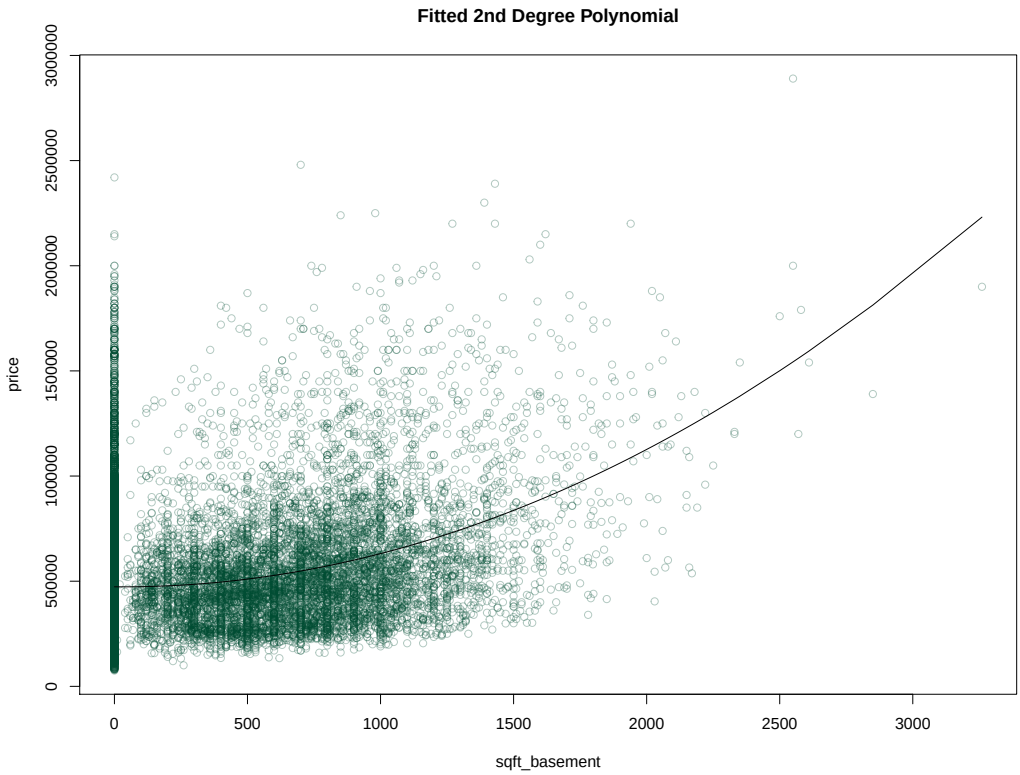
AIC method suggest that we should fit a ten degree **polynomial regression** of price on area of basement.

The value of **BIC** of the different degrees of polynomial model is as follows,



BIC method suggest that we should fit a second degree **polynomial regression** of price on area of basement.

Cross Validation method is the most accurate method among the given methods. Since, two of the methods suggest to fit a second degree polynomial hence, we should fit a second degree **polynomial regression** of price on area of basement.



Price vs Area Above

We obtain the impact of area of above on the price of the houses.

```
lr=lm(price~sqft_above)
summary(lr)

##
## Call:
## lm(formula = price ~ sqft_above)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -573911 -151298  -32904  104577 1582194
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.406e+05  3.791e+03   37.08   <2e-16 ***
## sqft_above  2.122e+02  1.967e+00  107.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 225200 on 21183 degrees of freedom
## Multiple R-squared:  0.3547,Adjusted R-squared:  0.3547
## F-statistic: 1.164e+04 on 1 and 21183 DF,  p-value: < 2.2e-16
```

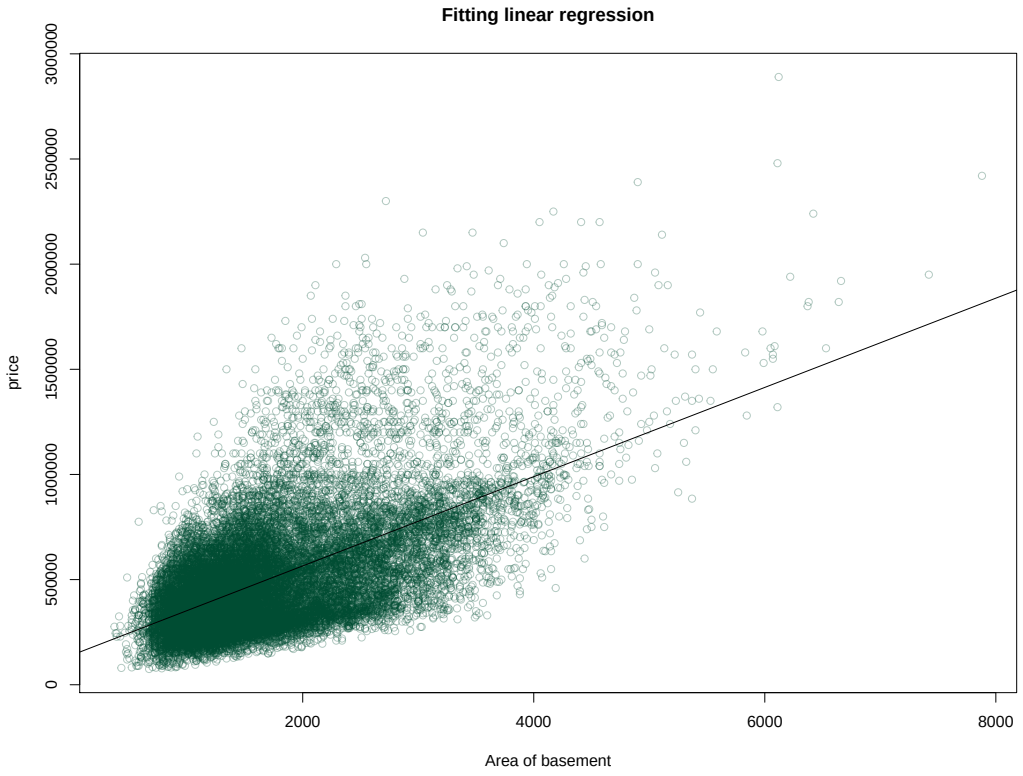
The linear model is,

$$y_i = a + bx_i + e_i$$

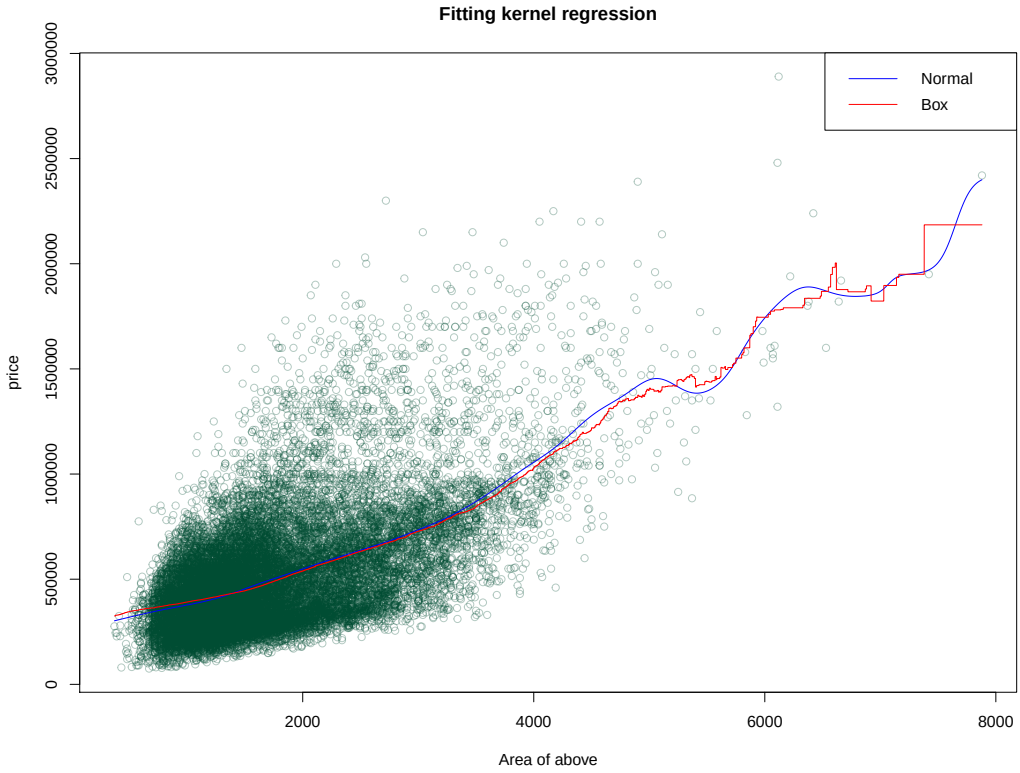
where,y_i = Price of i th house and x=area of above of the i th house, $i = 1, 2, .., 21613$.
We set the null hypothesis,

$$H_0 : b = 0 \text{ vs } H_1 : b \neq 0$$

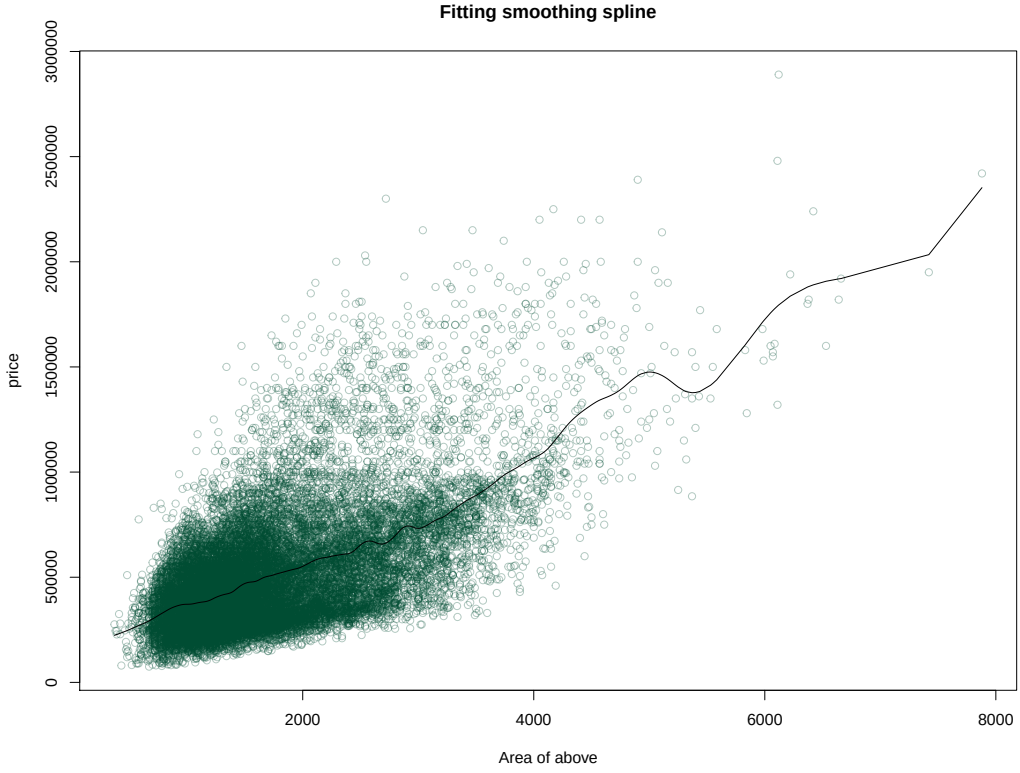
The p value of the hypothesis is less than 0.05.
Then we can conclude that **the area of above of the houses have significant effect of the linear regression of price** .
The fitted linear regression is as follows,



To get more precise relation between price and the area of the above,we fitted both **Box kernel regression** and **Gaussian kernel regression.**



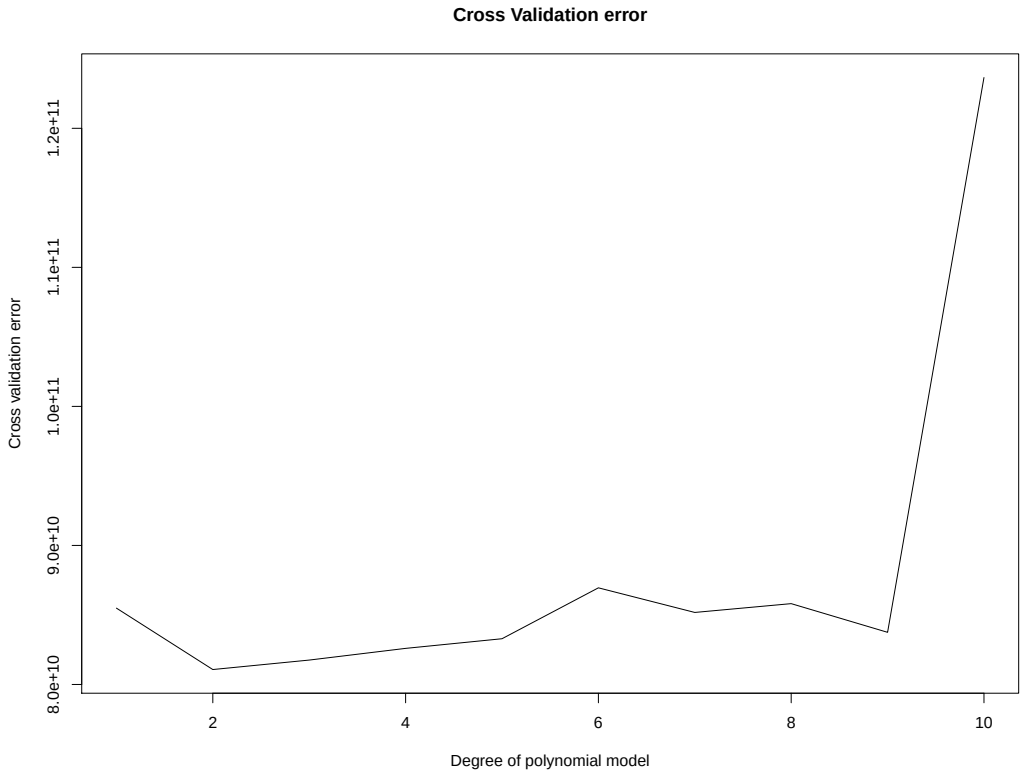
The fitted **spline regression** of the price on the area of the above using **smoothing spline,** is as follows,



We have the linear regression of the price on the area of above.Further,we are interested in fitting a **polynomial regression** with appropriate degree.

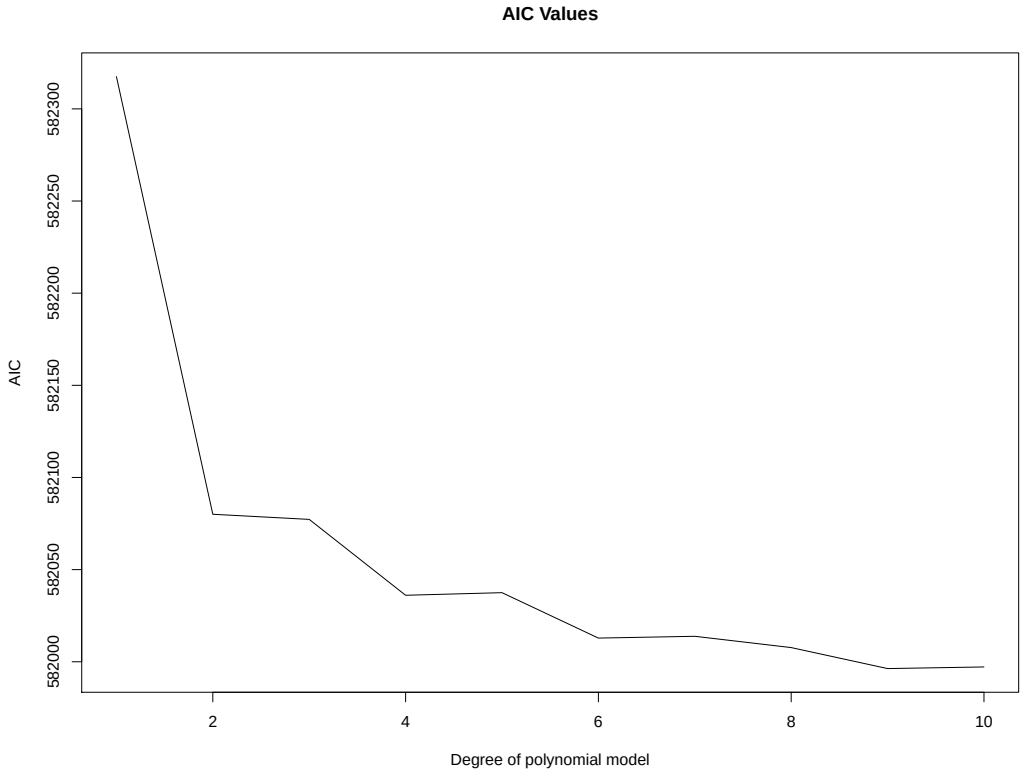
In order to determine the appropriate degree of the polynomial regression,we can fit the polynomial regression model with different degrees and then minimize the measure of optimism.We can obtain the measure of optimism using **Akaike information criterion(AIC),Bayesian information criterion(BIC)** and **Cross-validation** method**.**

The value of **Cross validation error** of the different degrees of polynomial model is as follows,



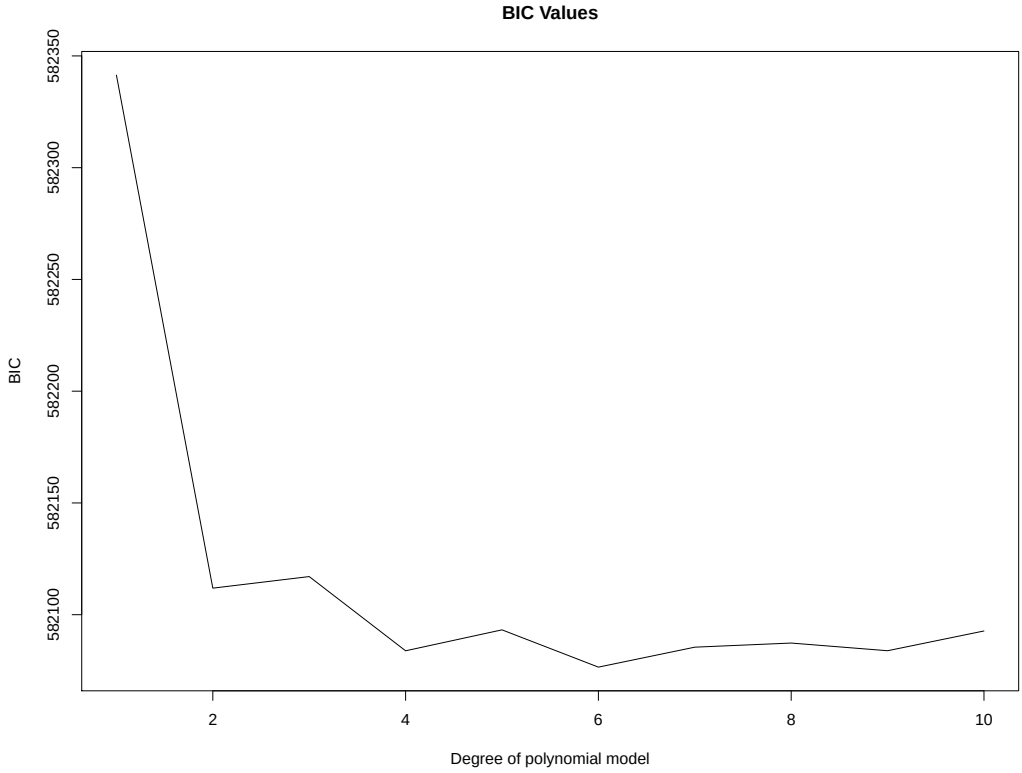
Cross Validation method suggest that we should fit a second degree **polynomial regression** of price on area of basement

The value of AIC of the different degrees of polynomial model is as follows,



AIC method suggest that we should fit a ten degree **polynomial regression** of price on area of basement.

The value of **BIC** of the different degrees of polynomial model is as follows,



BIC method suggest that we should fit a six degree **polynomial regression** of price on area of basement.

Cross Validation method is the most accurate method among the given methods.Thus,we should fit a two degree **polynomial regression** of price on area of basement.

Price vs Sqft Lot

We obtain the impact of area of lot on the price of the houses.

```
lr=lm(price~sqft_lot)
summary(lr)

##
## Call:
## lm(formula = price ~ sqft_lot)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -826007 -192869  -66823  116118 2335677
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.027e+05  2.078e+03  241.88   <2e-16 ***
## sqft_lot    8.060e-01  5.680e-02   14.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 279000 on 21183 degrees of freedom
## Multiple R-squared:  0.009417,Adjusted R-squared:  0.00937
## F-statistic: 201.4 on 1 and 21183 DF,  p-value: < 2.2e-16
```

The linear model is,

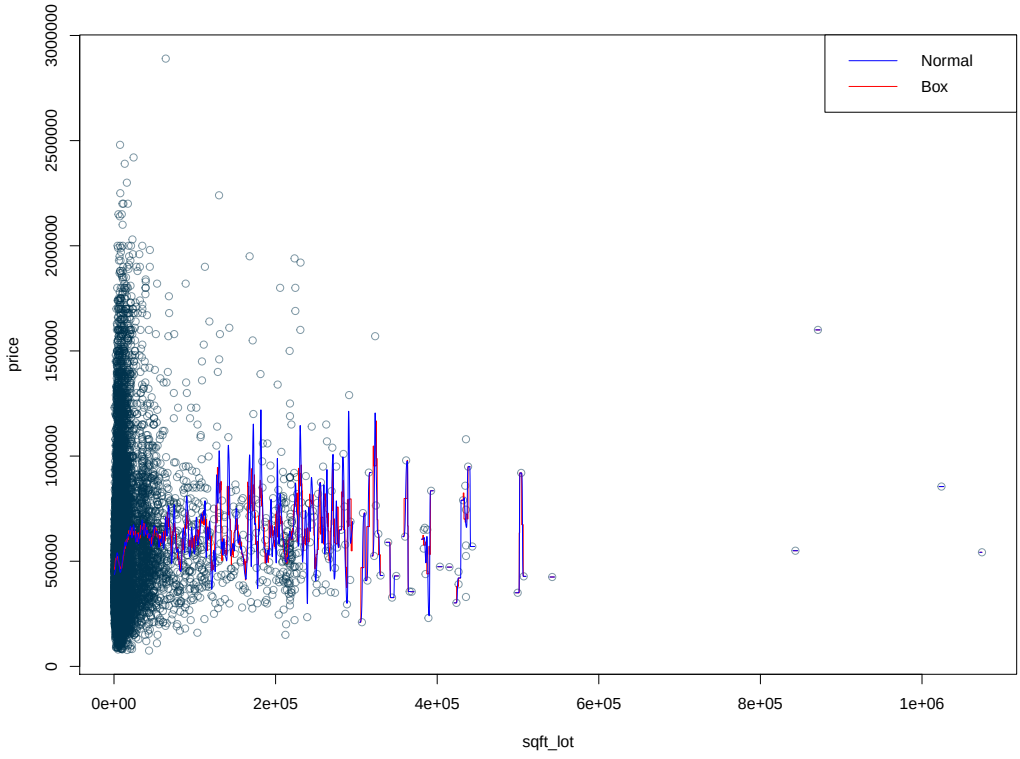
$$y_i = a + bx_i + e_i$$

where, y_i = Price of i th house and x = area of lot of the i th house, $i = 1, 2, .., 21613$.
We set the null hypothesis,

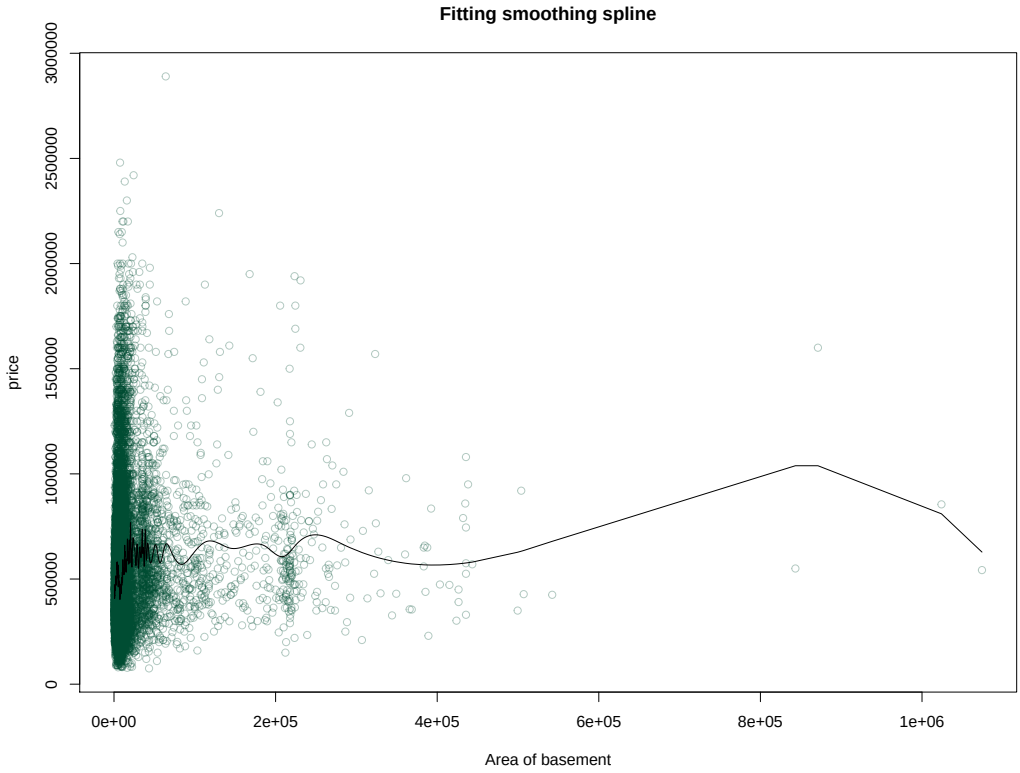
$$H_0 : b = 0 \text{ vs } H_1 : b \neq 0$$

The p value of the hypothesis is less than 0.05.
Then we can conclude that **the area of lot of the houses have significant effect of the linear regression of price** .
The fitted linear regression is as follows,

To get more precise relation between price and the area of the basements,we fitted both **Box kernel regression** and **Gaussian kernel regression.**



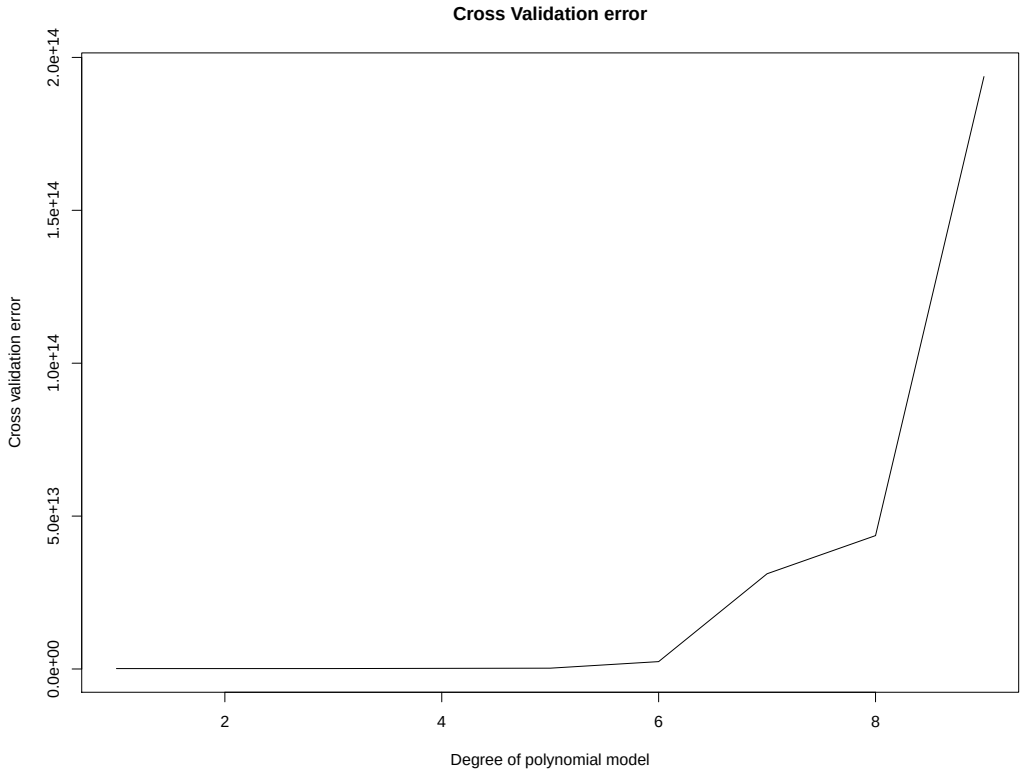
The fitted **spline regression** of the price on the area of basement using **smoothing spline,** is as follows,



We have the linear regression of the price on the area of basement.Further,we are interested in fitting a **polynomial regression** with appropriate degree.

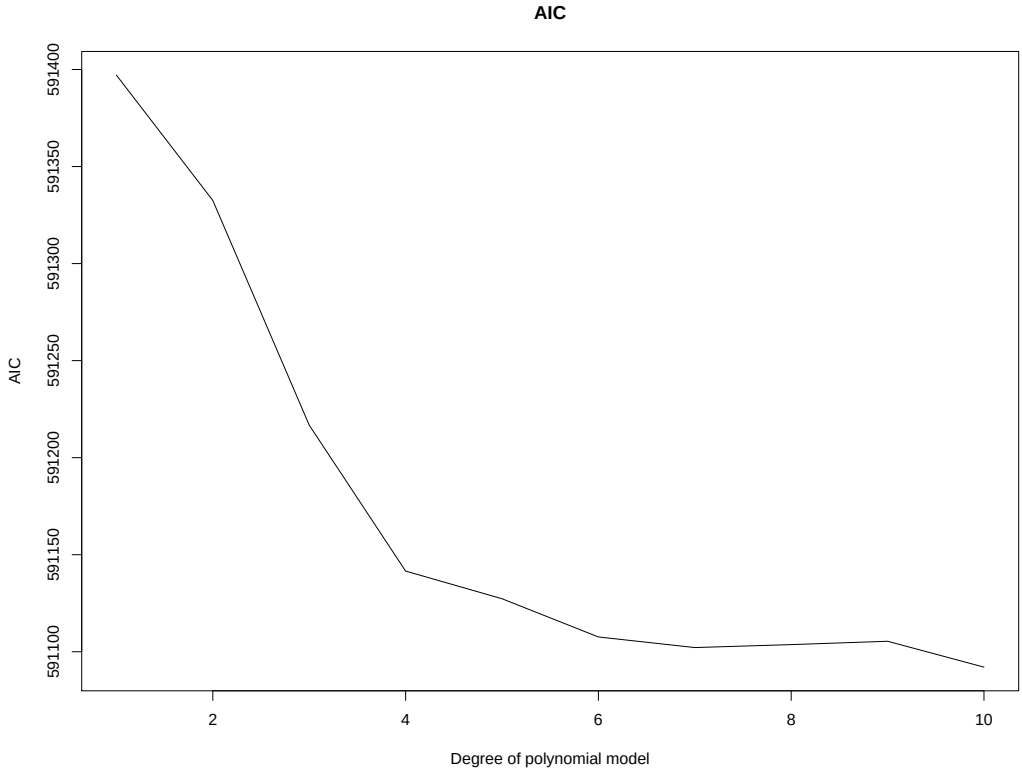
In order to determine the appropriate degree of the polynomial regression,we can fit the polynomial regression model with different degrees and then minimize the measure of optimism.We can obtain the measure of optimism using **Akaike information criterion(AIC),Bayesian information criterion(BIC)** and **Cross-validation** method**.**

The value of **Cross validation error** of the different degrees of polynomial model is as follows,



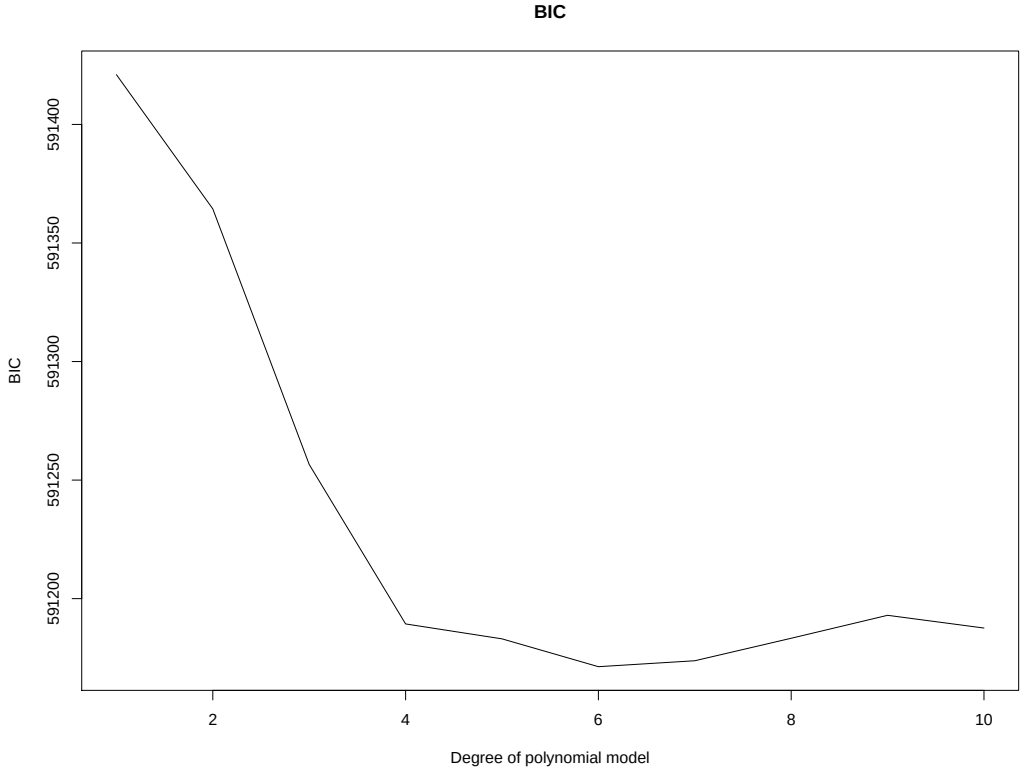
Cross Validation method suggest that we should fit a second degree **polynomial regression** of price on area of basement

The value of AIC of the different degrees of polynomial model is as follows,



AIC method suggest that we should fit a ten degree **polynomial regression** of price on area of basement.

The value of **BIC** of the different degrees of polynomial model is as follows,



BIC method suggest that we should fit a six degree **polynomial regression** of price on area of basement.
Cross Validation method is the most accurate method among the given methods.Thus,we should fit a six degree **polynomial regression** of price on area of basement.

Now,we should find the impact of the qualitative variables on the price of the house using **ANOVA** method.

We obtain the effect of number of bedrooms on the price of the house.

The **ANOVA** model is,

$$y_{ij} = \alpha + \mu_i + e_{ij}, i = 1, 2, .., n_i, j = 1, 2, .12$$

Here,y_{ij} = price of jth house with i number of bedrooms.
μ_i =additional effect of number of bedrooms on the price.
n_i =number of houses with i number of bedrooms
We set the null hypothesis,

$$H_0 : \mu_1 = \mu_1 = \mu_2 = .. = \mu_{12} = 0 \text{ vs } H_1 : H_0 \text{is not true}$$

The **ANOVA** table is given by,

```
summary(aov(price~as.factor(bedrooms)))

                       Df    Sum Sq   Mean Sq F value Pr(>F)
as.factor(bedrooms)    11 1.932e+14 1.757e+13   252.8 <2e-16 ***
Residuals           21173 1.471e+15 6.948e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is less than 0.05.

Thus,we can conclude that **number of bedrooms has significant effect on the price of the house.**

We obtain the effect of number of bathrooms on the price of the house.

The **ANOVA** table is given by,

```
summary(aov(price~as.factor(bathrooms)))

                        Df    Sum Sq   Mean Sq F value Pr(>F)
as.factor(bathrooms)    24 5.050e+14 2.104e+13     384 <2e-16 ***
Residuals            21160 1.159e+15 5.479e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is less than 0.05.

Thus,we can conclude that **number of bathrooms has significant effect on the price of the house.**

We obtain the effect of number of floors on the price of the house.

The **ANOVA** table is given by,

```
summary(aov(price~as.factor(floors)))

                     Df    Sum Sq   Mean Sq F value Pr(>F)
as.factor(floors)     5 1.671e+14 3.342e+13   472.8 <2e-16 ***
Residuals         21179 1.497e+15 7.069e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is less than 0.05.

Thus,we can conclude that **number of floors has significant effect on the price of the house.**

We obtain the effect of waterfront on the price of the house.

The **ANOVA** table is given by,

```
summary(aov(price~as.factor(waterfront)))

                         Df    Sum Sq   Mean Sq F value Pr(>F)
as.factor(waterfront)     1 4.549e+13 4.549e+13   595.3 <2e-16 ***
Residuals             21183 1.619e+15 7.642e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is less than 0.05.

Thus,we can conclude that **waterfront has significant effect on the price of the house.**

We obtain the effect of view on the price of the house.

The **ANOVA** table is given by,

```
summary(aov(price~as.factor(view)))

                   Df    Sum Sq   Mean Sq F value Pr(>F)
as.factor(view)     4 2.003e+14 5.009e+13   724.6 <2e-16 ***
Residuals       21180 1.464e+15 6.912e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is less than 0.05.

Thus,we can conclude that **view has significant effect on the price of the house.**

We obtain the effect of condtion on the price of the house.

The **ANOVA** table is given by,

```
summary(aov(price~as.factor(condition)))

                       Df    Sum Sq   Mean Sq F value Pr(>F)
as.factor(condition)    4 1.581e+13 3.953e+12   50.79 <2e-16 ***
Residuals           21180 1.649e+15 7.783e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is less than 0.05.

Thus,we can conclude that **condition has significant effect on the price of the house.**

We obtain the effect of grade on the price of the house.

The **ANOVA** table is given by,

```
summary(aov(price~as.factor(grade)))

                    Df    Sum Sq   Mean Sq F value Pr(>F)
as.factor(grade)    10 8.507e+14 8.507e+13    2214 <2e-16 ***
Residuals        21174 8.137e+14 3.843e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is less than 0.05.

Thus,we can conclude that **grade has significant effect on the price of the house.**

We obtain the effect of renovation indicator on the price of the house.

The **ANOVA** table is given by,

```
renovation_indicator=as.integer(yr_renovated>0)
summary(aov(price~as.factor(renovation_indicator)))

                                  Df    Sum Sq   Mean Sq F value Pr(>F)
as.factor(renovation_indicator)    1 1.554e+13 1.554e+13   199.7 <2e-16
Residuals                      21183 1.649e+15 7.783e+10

as.factor(renovation_indicator) ***
Residuals
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is less than 0.05.

Thus,we can conclude that **renovation indicator has significant effect on the price of the house.**

We obtain the effect of number of bathrooms on the price of the house.

The **ANOVA** table is given by,

```
summary(aov(price~as.factor(bedrooms)))

                     Df    Sum Sq   Mean Sq F value Pr(>F)
as.factor(bedrooms)  11 1.932e+14 1.757e+13   252.8 <2e-16 ***
Residuals         21173 1.471e+15 6.948e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is less than 0.05.

Thus,we can conclude that **number of bathrooms has significant effect on the price of the house.**