The Challenge:
For this Assignment, in a group of 3, you need to collect data and then build a model to predict price of a player IPL 2018 auction.

Training Data:
For training you need to collect numbers and statistics for at least 40 players. Your target (dependent variable)should be price. Typical independent variables include but not limited to International T20 Matches Played, ODI, Matches Played, Strike rates – batting, bowling etc.

Testing Data:
Test your model for each match in IPL 2018 auction data.

Data collection :
Data is collected from 3 websites :

- Wikipedia
- Howstat
- ipltickets.net

Data from Wikipedia is collected manually which contains list of all the players base price and sold out price from 2008 to 2017. The sold out and base price data of 2018 is scraped from ipltickets.net using R. Auction data(prices) are converted from $(Dollar) to ₹(Rupee), converted into Net Present Value and then took the average of all the prices listing from 2008 to 2017.
The method of extracting data from websites is known as web scraping. Rvest is a package present in R language that is used to scrape data from html web pages. Web scraping packages can access the world wide web using Hypertext Transfer Protocol(HTTP) or through a web browser. Web pages are build using HTML or XHTML(text based mark-up languages). To extract particular type of data, HTML/CSS tags are required. Selector Gadget, an open source tool present in chrome extension, that makes CSS selector generation and discovery in a much easier way from complex websites.

Data Visualization :
Total of 48 players data is extracted from the website howstat and stored in excel file with each players data in each excel file. The whole extracted data is put together in one excel file with player names as rows and Variables in column which is done manually. There are 108 decision variables which are contributing to the output. All the categorical data(factors) is converted to numerical so that it can contribute to the model. For example, all the batsmen playing with left hand are assigned value of 1 and the batsmen playing with right hand are assigned a value of 2. Similar concept is applied to bowlers. All the non-filled values are given value 0 and all the NAs are converted to numerical.

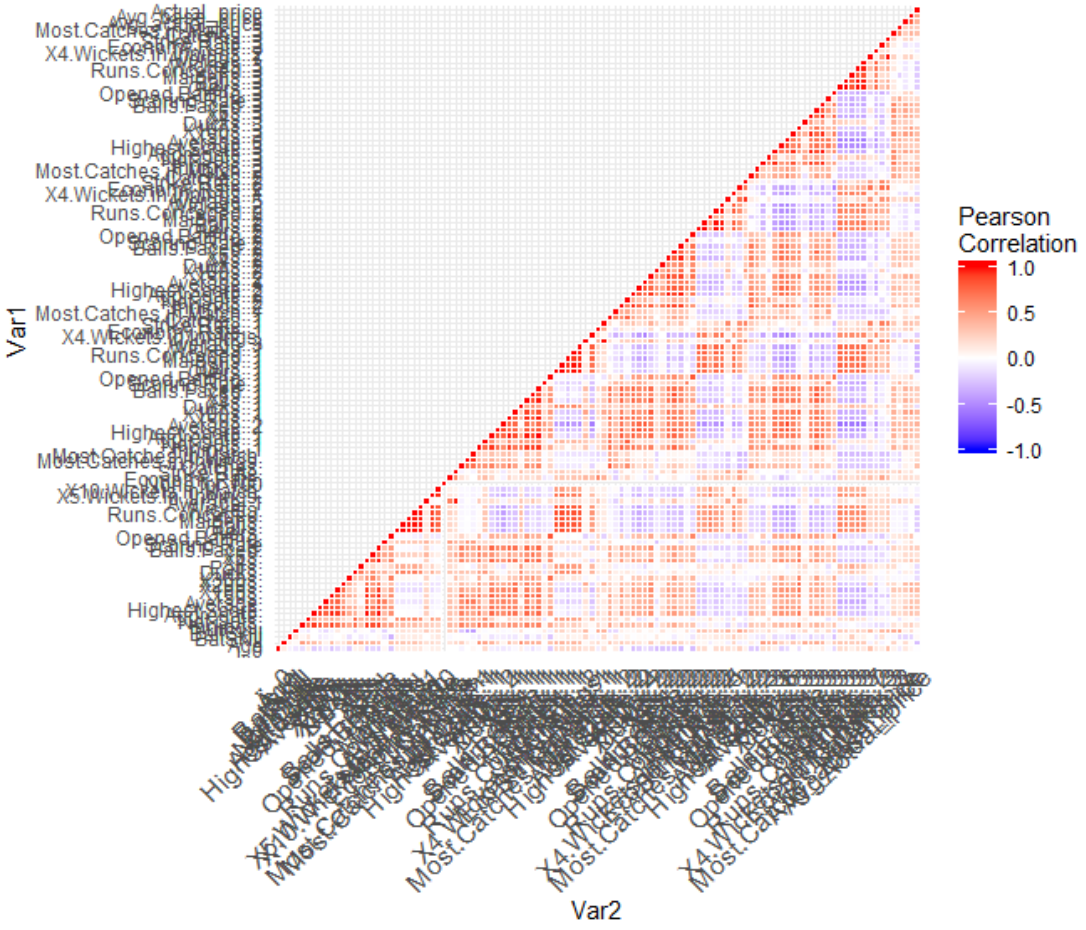
The 108 decision variables are PlayerName, DateofBirth, Age, BatSkill, BallSkill, Innings., Not.Outs., Aggregate., Highest.Score., Average., X50s., X100s., X200s., X300s., Ducks., Pairs., X4s., X6s., Balls.Faced., Scoring.Rate, Opened.Batting., Overs., Balls., Maidens., Runs.Conceded., Wickets., Average..1, X5.Wickets.in..Innings., X10.Wickets.in.Match., Best...Innings., Best...Match., None.for.100, Economy.Rate., Strike.Rate., Catches., Most.Catches.in.Innings., Most.Catches.in.Match., Innings..1, Not.Outs..1, Aggregate..1, Highest.Score..1, Average..2, X50s..1, X100s..1, Ducks..1, X4s..1, X6s..1, Balls.Faced..1, Scoring.Rate.1, Opened.Batting..1, Overs..1, Balls..1, Maidens..1, Runs.Conceded..1, Wickets..1, Average..3, X4.Wickets.in.Innings., Best., Economy.Rate..1, Strike.Rate..1, Catches..1, Most.Catches.in.Match..1, Innings..2, Not.Outs..2, Aggregate..2, Highest.Score..2, Average..4, X50s..2, X100s..2, Ducks..2, X4s..2, X6s..2, Balls.Faced..2, Scoring.Rate.2, Opened.Batting..2, Overs..2, Balls..2, Maidens..2, Runs.Conceded..2, Wickets..2, Average..5, X4.Wickets.in.Innings..1, Best..1, Economy.Rate..2, Strike.Rate..2, Catches..2, Most.Catches.in.Match..2, Innings..3, Not.Outs..3, Aggregate..3, Highest.Score..3, Average..6, X50s..3, X100s..3, Ducks..3, X4s..3, X6s..3, Balls.Faced..3, Scoring.Rate.3, Opened.Batting..3, Overs..3, Balls..3, Maidens..3, Runs.Conceded..3, Wickets..3, Average..7, X4.Wickets.in.Innings..2, Best..2,

Economy.Rate..3, Strike.Rate..3, Catches..3, Most.Catches.in.Match..3, Avg_actual_price, Avg_base_price, Actual_price.

The above data consists of information of the 48 players in all formats(Test, ODI, T20, IPL). Data of other league matches and domestic matches are not considered.

In the collected data, number of variables are more than the data collected which causes overfitting. So, to cut down the number of variables which were used for building the model, we searched for the dependencies and removed the unwanted columns. This can be determined from correlation matrix.

If the correlation between the two different variables is nearly equals one, then one variable is considered instead of two variables.



In the above figure, red colour indicates positive correlation and blue colour indicates negative correlation.

Considering correlation values which have values >0.95, this method is applied and the total number of variables are reduced to 89 variables. Then removing unwanted variables which contains almost all zeros or don't contribute to the output, It got reduced to 81 variables.

Building the Model, Results and Inferences :

After Collecting the Data and Pre processing it we have removed some unwanted variables even though fitting a linear model is not possible because :

1. **Number of Variables > Number of Observations**

so we cannot obtain adjusted R-square value for our model. To solve this problem we have to perform some feature selection methods such as Regularization

2. **Subset Selection**: identify a subset of predictors that strongly related to the response
    1. Best Subset Selection
    2. Stepwise Selection
       Forward Stepwise Selection ;Backward Stepwise Selection; Hybrid Approaches
    3. Choosing the Optimal Model
       Cp, AIC, BIC, and Adjusted R2 ; Validation and Cross-Validation
3. **Shrinkage**: fit a model using all predictors by shrinking estimated coefficients towards zero to reduce variance. :**LASSO and Ridge Regression**

4. **Dimension Reduction**: projecting the predictors(p) into a M dimensional subspace (M < p) :**PCR and PLS**

EXPLANATION:

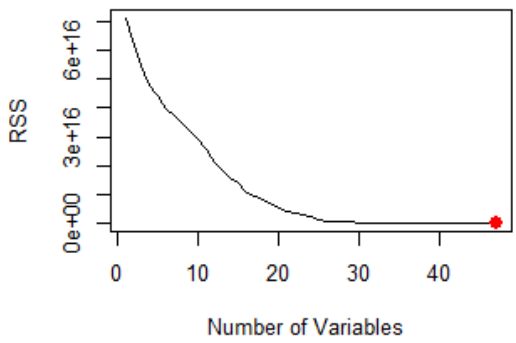
1. Best Subset Selection :
   There are total 81 variables after Data pre-processing and feature selection through Correlation Matrix.
   INFERENCES : so it does an exhaustive search of 2^{81} possibilities as it was too large and takes more time
2. Stepwise Selection
    - Forward Stepwise Selection :

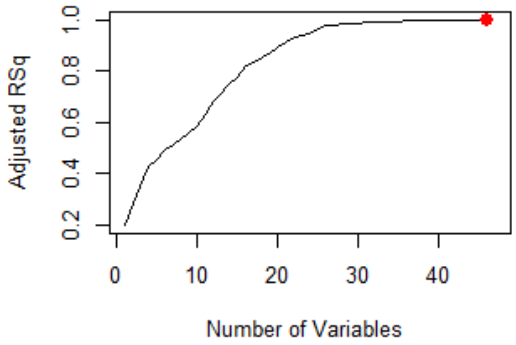
There are total 81 variables so somewhat forward selection is possible and sub-optimal compared to best-subset selection. I imported these four libraries MASS, ISLR, glmnet and leaps.after I fitted my forward sub selection method to my data. The **Results** of my R squared value are :

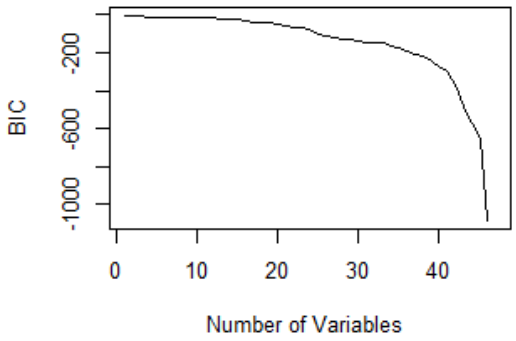
```
 [1] 0.2120351 0.3151020 0.3919882 0.4738682 0.5090502 0.5586925
 [7] 0.5843036 0.6134438 0.6451065 0.6739357 0.7167053 0.7637360
[13] 0.7907350 0.8247233 0.8447335 0.8803454 0.8961865 0.9079563
[19] 0.9199496 0.9392874 0.9520649 0.9625301 0.9677878 0.9743806
[25] 0.9853158 0.9897711 0.9919145 0.9934025 0.9945347 0.9956926
[31] 0.9962351 0.9968322 0.9972454 0.9979951 0.9984251 0.9990650
[37] 0.9993750 0.9995630 0.9997386 0.9998639 0.9999258 0.9999864
     [43] 0.9999988 0.9999998 1.0000000 1.0000000 1.0000000
```

INFERENCES : Around the value of 47 variable my R-square is one. Then I caluculated Cp ,AIC, BIC and Adjusted R-square Values and plotted the graphs shown below.

Results : The value for no. of variables to be selected with respect to RSS, Adj-Rsq ,AIC, BIC are 46,47,47,47







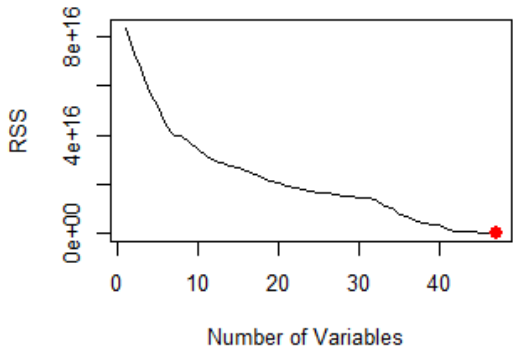
- Backward Stepwise Selection :

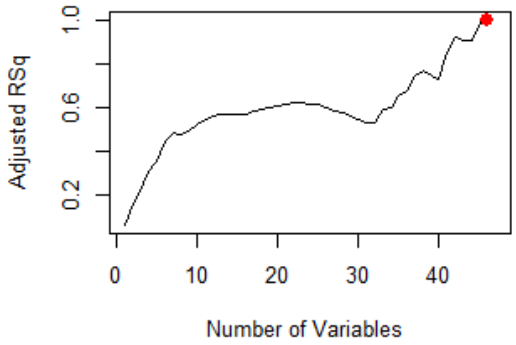
There are total 81 variables so somewhat forward selection is possible and sub-optimal compared to best-subset selection. I imported these four libraries MASS, ISLR, glmnet and leaps.after I fitted my forward sub selection method to my data. The **Results** of my R squared value are :

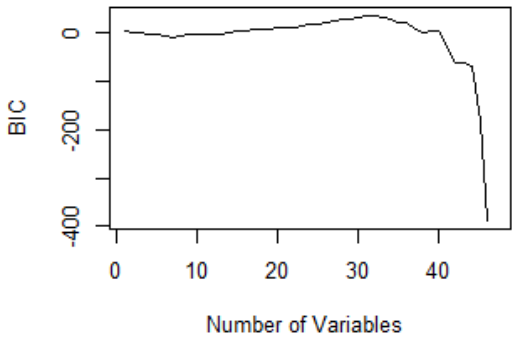
```
 [1] 0.07549438 0.19059802 0.26105340 0.36398809 0.42176749 0.51227930 0.55
       811232 0.56341333 0.58844882 0.62317512 0.64917184 0.67082237
[13] 0.68369024 0.69298204 0.70281065 0.71559091 0.73330893 0.74745732 0.7
       6090538 0.77357731 0.78598367 0.79638026 0.80515193 0.81194089
[25] 0.81721418 0.81941688 0.82305436 0.82819557 0.83225888 0.83615819 0.8
       4019996 0.85042134 0.87663834 0.88795410 0.91104193 0.92313541
[37] 0.94525115 0.95476606 0.95676101 0.95936791 0.98055137 0.99117147 0.9
             9182551 0.99373246 0.99944547 0.99999318 1.00000000
```

INFERENCES : Around the value of 47 variable my R-square is one. Then I caluculated Cp ,AIC, BIC and Adjusted R-square Values and plotted the graphs shown below.

Results : The value for no. of variables to be selected with respect to RSS, Adj-Rsq ,AIC, BIC are 46,47,47,47







The coefficients for Forward Selection are :

(Intercept)	Age	BatSkill	Not.Outs.	Average.
3.599193e+08	-1.313532e+07	1.896013e+06	-4.977653e+06	-9.893500e+05
X100s.	X200s.	Ducks.	X4s.	X5.Wickets.in..Innings.
7.016832e+05	-1.072329e+07	4.586512e+06	2.149855e+05	1.269615e+06
Catches.	Most.Catches.in.Match.	Innings..1	Aggregate..1	X100s..1
5.207062e+05	5.417648e+06	-7.384170e+05	2.102876e+04	1.630994e+06
X4s..1	X6s..1	Scoring.Rate.1	Balls..1	Wickets..1
-1.990923e+05	-3.019331e+04	3.585931e+05	-1.719196e+04	8.511551e+05
X4.Wickets.in.Innings.	Catches..1	Most.Catches.in.Match..1	Innings..2	Not.Outs..2
-4.418767e+04	1.972435e+05	-6.379551e+06	2.911570e+05	4.851188e+06
Highest.Score..2	X100s..2	Ducks..2	X4s..2	Scoring.Rate.2
7.525621e+04	2.430669e+07	-7.774533e+06	5.326568e+04	4.892544e+04
Balls..2	Maidens..2	X4.Wickets.in.Innings..1	Catches..2	Most.Catches.in.Match..2
5.226780e+03	-7.282122e+04	-1.346415e+03	-1.948616e+06	-2.508181e+07
Aggregate..3	X50s..3	X100s..3	Ducks..3	X4s..3
-9.231417e+04	-3.118612e+04	-2.430973e+06	2.599360e+06	7.042428e+05
X6s..3	Balls.Faced..3	Opened.Batting..3	X4.Wickets.in.Innings..2	Strike.Rate..3
1.877267e+05	6.159030e+04	-9.241382e+05	1.219648e+07	-2.247826e+05
Catches..3	Avg_actual_price	Avg_base_price		
1.339555e+06	1.004284e+00	-5.867239e-01		

The coefficients for backward Selection are :

(Intercept)	Age	BatSkill	BallSkill	Innings.	Not.Outs.
-1112595763.1	49630645.5	-142226860.6	81065344.7	-116579080.1	90891749.2
Aggregate.	Highest.Score.	Average.	X100s.	X200s.	X300s.
4344493.9	2450587.0	2805858.5	-306723931.9	-566805011.4	-970708324.7
Ducks.	X4s.	X6s.	Scoring.Rate	Balls.	Maidens.
192723139.9	-13738818.5	74242481.8	-4764862.3	247201.8	-11534232.2
Wickets.	Average..1	X5.Wickets.in..Innings.	Economy.Rate.	Strike.Rate.	Catches.
1672651.0	-36524155.6	30548414.9	-210736407.9	18248667.8	-9260092.4
Most.Catches.in.Innings.	Most.Catches.in.Match.	Innings..1	Aggregate..1	Highest.Score..1	X50s..1
642331497.0	-275506984.4	193150.3	-228504.2	-11607630.1	-39781153.3
X100s..1	X4s..1	X6s..1	Scoring.Rate.1	Balls..1	Wickets..1
82103701.9	2744258.9	-11700832.1	7273064.2	256372.4	4788908.2
X4.Wickets.in.Innings.	Economy.Rate..1	Strike.Rate..1	Catches..1	Most.Catches.in.Match..1	Innings..2
-161176571.4	77560974.4	-1057449.0	9315293.9	-192311522.5	-24313447.5
Not.Outs..2	Aggregate..2	Highest.Score..2	X100s..2	Ducks..2	X4s..2
-3093897.8	1369684.6	2969598.5	-387271042.1	-277657217.7	11424988.7

- Choosing the Optimal Model

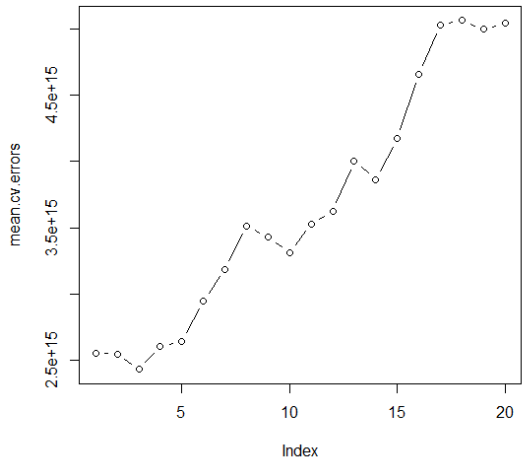
RESULTS:

Now we performed validation and cross validation Validation errors for Forward Model :

```
 [1] 1.888615e+15 3.037862e+15 3.037862e+15 3.426095e+15 3.424984e+15
 [6] 2.570951e+15 2.957942e+15 2.894986e+15 3.604158e+15 5.093671e+15
[11] 5.163444e+15 4.480094e+15 4.526668e+15 4.397623e+15 4.595657e+15
[16] 4.658251e+15 4.489039e+15 4.871431e+15 4.895542e+15 4.999805e+15
```

Cross Validation errors for Forward Model :

1	2	3	4	5
2.558191e+15	2.550505e+15	2.432548e+15	2.603644e+15	2.646736e+15
6	7	8	9	10
2.949860e+15	3.189391e+15	3.514536e+15	3.429724e+15	3.315093e+15
11	12	13	14	15
3.530362e+15	3.622608e+15	4.002320e+15	3.865699e+15	4.172183e+15
16	17	18	19	20
4.655979e+15	5.028398e+15	5.062948e+15	4.997043e+15	5.040219e+15



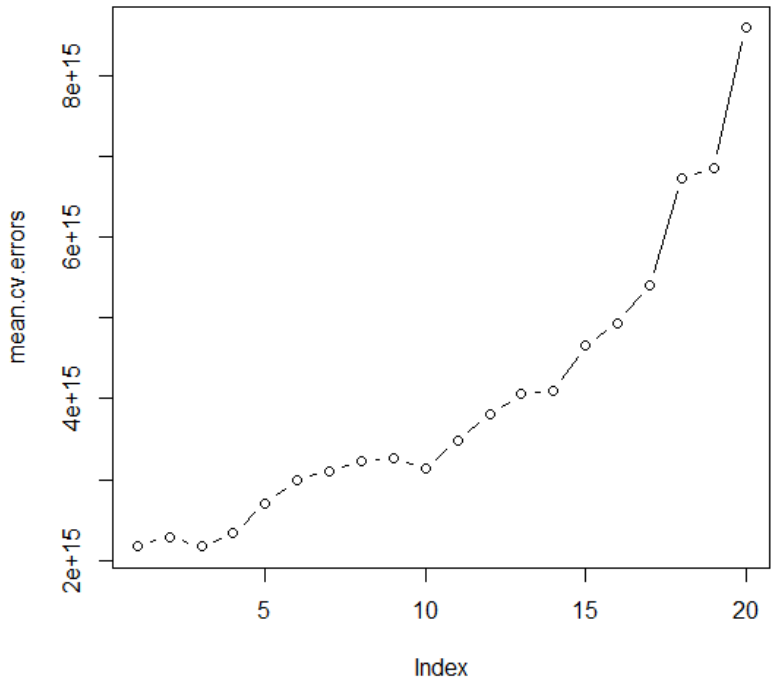
RESULTS:

Now we performed validation and cross validation Validation errors for Backward Model :

```
[1] 1.977368e+15 2.013289e+15 2.108028e+15 2.432386e+15 2.048173e+15 2.334
597e+15 2.243229e+15 2.617105e+15 2.848129e+15 4.978366e+15 5.016402e+15
[12] 4.939363e+15 3.653868e+15 4.698750e+15 5.188497e+15 9.626109e+15 1.00
3001e+16 1.252285e+16 1.252285e+16 3.566437e+16
```

Cross Validation errors for Backward Model :

```
           1            2            3            4            5
6            7            8            9           10           11
12
2.177451e+15 2.290430e+15 2.166357e+15 2.336188e+15 2.705388e+15 2.987881e
+15 3.108410e+15 3.221877e+15 3.269260e+15 3.140064e+15 3.484174e+15 3.801
982e+15
          13           14           15           16           17
18           19           20
4.055860e+15 4.099197e+15 4.656564e+15 4.941113e+15 5.403693e+15 6.728511e
+15 6.868748e+15 8.596826e+15
```

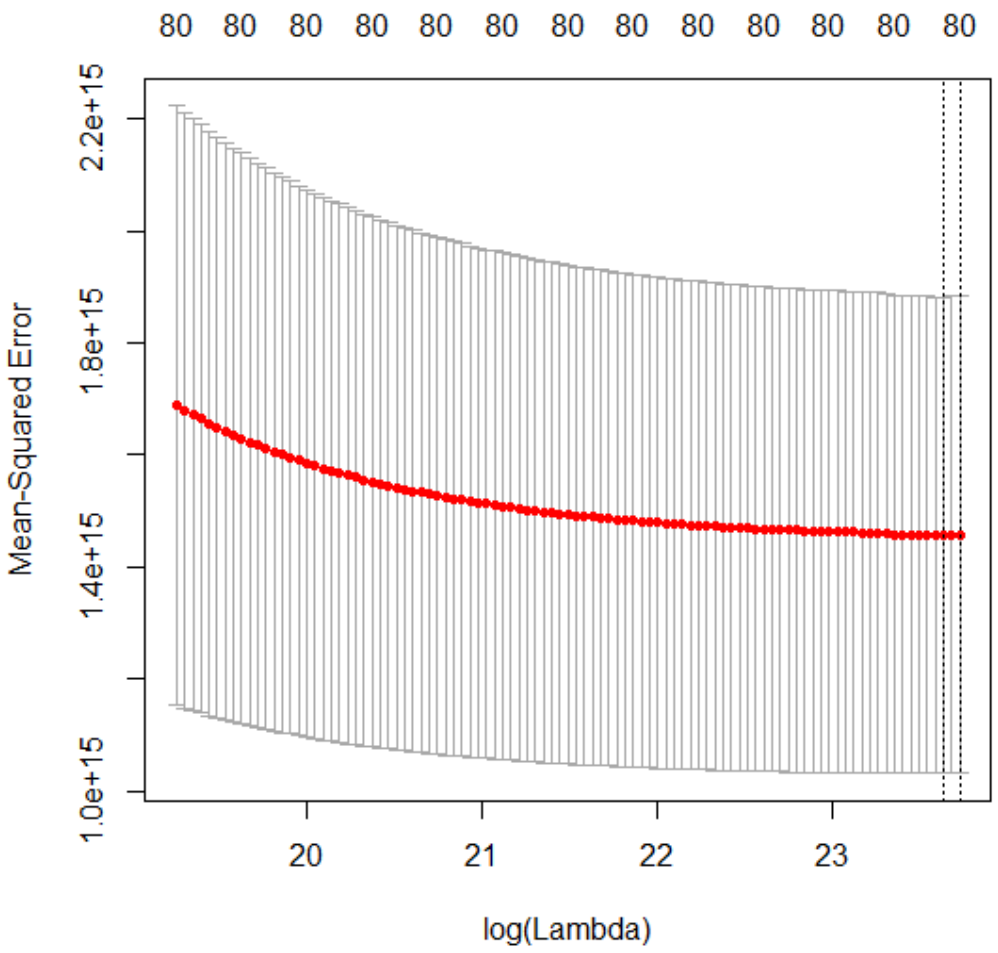


INFERENCES : The errors are in the range of e+15 because the predictor values taken are in crores (e+8) so in Normal terms if predictor values are in range of 0 to 1 then the errors are in range of e-2. The errors for both Models are around 0.217 %

3. Ridge Regression :
   Shrinks the regression coefficients by imposing a penalty on their size and L2 Norm.we split the data into 1:1 ratio of training and testing set.Then we will try to find the best lamda value

for regression using mean squared error.If the lamba =0 then ridge regression is nothing but normal regression

RESULTS AND INFERENCES :



From the graph best lamda value = 18335537823
As Lamda is Large , so almost all of the coefficients will be reduced to negligible .From the ridge prediction the mean square error is 0.24 % .The coefficients for prediction of best Lamda value :

(Intercept)	Age	BatSkill
5.729701e+07	-7.343799e+03	6.204467e+03
BallSkill	Innings.	Not.Outs.
5.525616e+02	1.860908e+02	-1.039347e+03
Aggregate.	Highest.Score.	Average.
1.214067e+01	3.193498e+02	1.719877e+03
X100s.	X200s.	X300s.
5.483216e+03	3.039583e+04	-5.049980e+04
Ducks.	X4s.	X6s.
-3.098365e+03	8.181040e+01	5.628185e+02
Scoring.Rate	Balls.	Maidens.
2.091047e+01	-3.809815e+00	-1.083982e+02
Wickets.	Average..1	X5.Wickets.in..Innings.
-2.147169e+02	2.455434e+02	-1.516524e+03
Economy.Rate.	Strike.Rate.	Catches.

```
            9.191338e+03            -5.294415e+00             7.119654e+02
Most.Catches.in.Innings.   Most.Catches.in.Match.               Innings..1
            1.312986e+04             8.084573e+03             1.954533e+02
            Aggregate..1         Highest.Score..1                  X50s..1
            1.023773e+01             4.710751e+02             1.703433e+03
                X100s..1                  X4s..1                   X6s..1
            4.925367e+03             8.530789e+01             3.712605e+02
          Scoring.Rate.1                Balls..1               Wickets..1
            3.886524e+02            -1.035934e+01            -4.084956e+02
  X4.Wickets.in.Innings.          Economy.Rate..1           Strike.Rate..1
           -5.956189e+03             2.653117e+03             9.025077e+02
              Catches..1 Most.Catches.in.Match..1               Innings..2
            5.437021e+02            -1.960974e+03             1.060070e+03
             Not.Outs..2             Aggregate..2         Highest.Score..2
            4.864185e+03             3.319876e+01             5.754113e+02
                X100s..2                 Ducks..2                  X4s..2
            3.455290e+03             2.059992e+03             3.584045e+02
                  X6s..2           Scoring.Rate.2                Balls..2
            4.969612e+02             3.341090e+02            -4.531072e+01
              Maidens..2               Wickets..2               Average..5
           -8.187025e+02            -1.055832e+03             9.690965e+02
X4.Wickets.in.Innings..1          Economy.Rate..2           Strike.Rate..2
           -3.339538e+04            -2.215802e+03             1.092168e+03
              Catches..2 Most.Catches.in.Match..2               Innings..3
            2.248174e+03             4.441328e+04             6.055946e+02
             Not.Outs..3             Aggregate..3         Highest.Score..3
            2.165473e+03             2.467565e+01             5.932971e+02
                 X50s..3                 X100s..3                 Ducks..3
            3.177591e+03             2.721089e+04             6.590851e+02
                  X4s..3                   X6s..3           Balls.Faced..3
            2.118648e+02             4.814622e+02             3.200464e+01
          Scoring.Rate.3        Opened.Batting..3                Balls..3
            9.022595e+02            -1.776393e+02            -2.341419e+01
              Maidens..3               Wickets..3 X4.Wickets.in.Innings..2
           -1.181649e+04            -4.473150e+02            -5.203384e+03
         Economy.Rate..3           Strike.Rate..3               Catches..3
           -2.452062e+03             2.975215e+02             1.551515e+03
Most.Catches.in.Match..3         Avg_actual_price           Avg_base_price
            3.491090e+04             1.278153e-03             1.539084e-03
```

4. LASSO regression

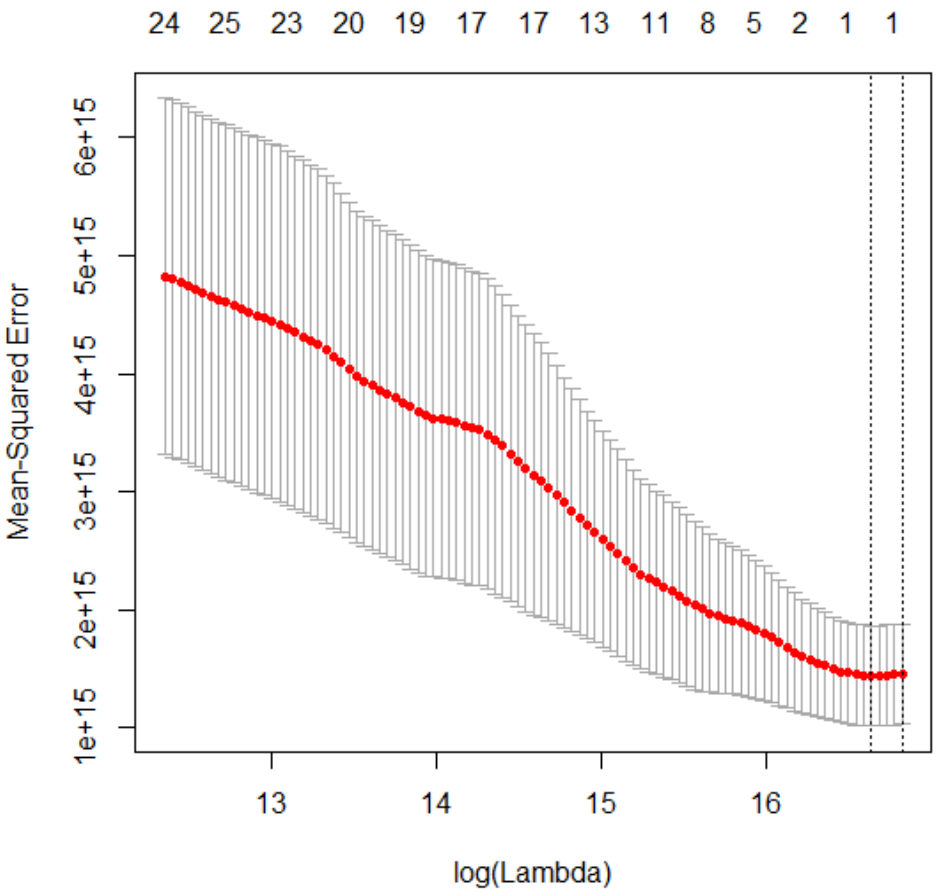
Produce simpler and more interpretable models with subset of predictors which is drawback of ridge regression Interpretation becomes difficult when p is quiet large and it uses L1 norm

RESULTS AND INFERENCES :

From the graph best lamda value = 16706659

As Lamda is Large , so almost all of the coefficients will be reduced to zero .From the LASSO prediction the mean square error is 0.24 % .The coefficients for prediction of best Lamda value :

```
  (Intercept) Most.Catches.in.Match..2           Avg_base_price
 5.524990e+07             3.999852e+05             1.086991e-01
```



5. Principal Components Regression :

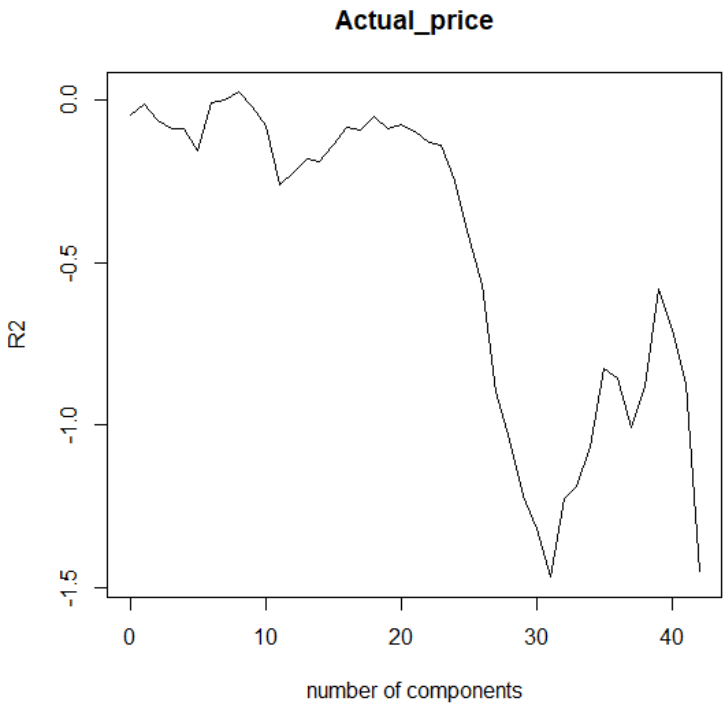
Transform the predictors and fit least squares model using transformed variables. If p is large relative to n, selecting a value of M « p significantly reduce variance of fitted coefficients
So we fitted the PCR, cross validated and observed our Adjusted R-Squared values as below table :

RESULTS :

```
Fit method: svdpc
Number of components considered: 42

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)   1 comps   2 comps   3 comps   4 comps   5 comps   6 comps   7 comps   8 comps   9 comps  10 comps  11 comps  12 comps  13 comps
CV       44296626  41662485  42372264  42795311  43157594  43749164  41703038  41882843  40331290  41201824  42613815  41854015  41956226  43590414
adjCV    44296626  41567436  42242955  42631157  42987235  43621637  41291213  41365813  40240877  40855860  42185149  41911595  40980499  42999642
       14 comps  15 comps  16 comps  17 comps  18 comps  19 comps  20 comps  21 comps  22 comps  23 comps  24 comps  25 comps  26 comps  27 comps  28 comps
CV     46335938  46974284  47985645  47170547  46937715  47385985  48482240  47984541  50691569  50683241  51721596  51951866  51683249  56124906  58211017
adjCV  45640184  45945453  46445736  46127014  45775911  46313056  47417391  47053573  49396524  49401692  50159357  50467696  50276824  54660583  56960496
       29 comps  30 comps  31 comps  32 comps  33 comps  34 comps  35 comps  36 comps  37 comps  38 comps  39 comps  40 comps  41 comps  42 comps
CV     59030501  61108522  61655974  61821332  55825092  54114843  53387044  53932612  52881907  54964991  53215542  53651538  56011315  62825602
adjCV  58571852  59152360  60270573  60012835  54514932  51858803  51110498  51714650  50940375  52910220  51569699  51796641  53695404  60157559

TRAINING: % variance explained
             1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
X              30.28    45.11    54.17    60.31    65.49    69.66    72.86    75.95    78.62     80.91     83.12     84.91     86.53     88.10     89.29
Actual_price   14.17    14.25    16.15    16.31    17.02    31.33    33.05    33.27    36.21     36.77     36.85     46.39     46.95     47.69     52.14
             16 comps  17 comps  18 comps  19 comps  20 comps  21 comps  22 comps  23 comps  24 comps  25 comps  26 comps  27 comps  28 comps  29 comps
X               90.40     91.48     92.31     93.12     93.84     94.46     95.06     95.62     96.14     96.61     97.05     97.41     97.74     98.06
Actual_price    55.46     55.47     57.00     58.63     59.55     60.05     62.36     62.67     64.06     64.13     64.40     64.40     64.65     64.81
             30 comps  31 comps  32 comps  33 comps  34 comps  35 comps  36 comps  37 comps  38 comps  39 comps  40 comps  41 comps  42 comps
X               98.32     98.57     98.77     98.95     99.11     99.26     99.38     99.49     99.58     99.66     99.73     99.80     99.85
Actual_price    75.48     75.79     80.71     81.98     89.42     90.56     90.65     90.79     91.93     91.93     93.40     96.03     96.12
```



INFERENCES :

The mean value of Error in PCR is around 0.17 so far the best And from the graph the first 30 principal components are chosen because around that value adjusted R square Value is Approximately one. Note that here graph is reverse because R2 is in negative terms

6. Partial Least Squares :

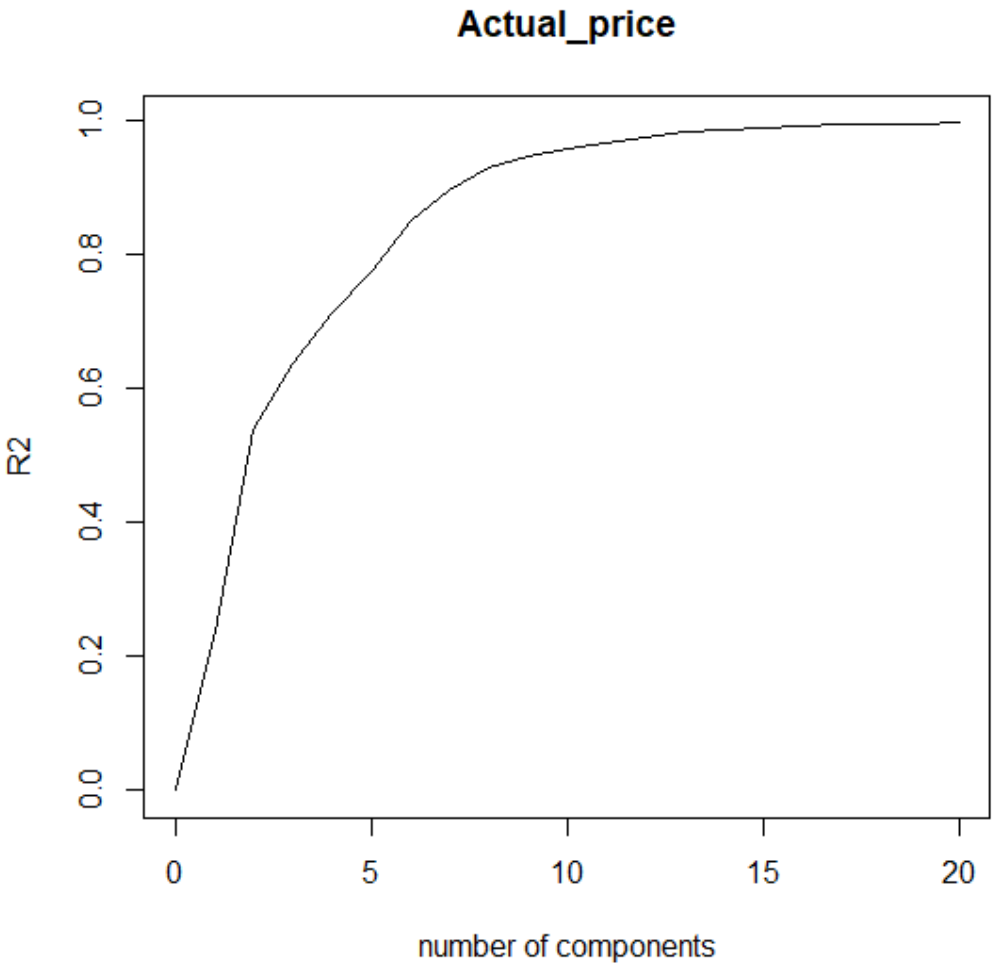
A supervised alternative to PCR ,find directions that help explain both the response and the predictors , place highest weight on the variables that are most strongly related to the response. So we fitted the PLS, cross validated and observed our Adjusted R-Squared values as below table :

RESULTS :

```
Data:   X dimension: 24 80
        Y dimension: 24 1
Fit method: kernelpls
Number of components considered: 20

VALIDATION: RMSEP
Cross-validated using 10 random segments.
      (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
CV       38884930  50688423  44090647  49885684  53447959  54012087  52672146  53697305  53969249  54634114  55133663  55394040  55456735  55500321
adjCV    38884930  48135138  43257195  48104620  50825785  51185267  49878580  50811028  51069746  51686348  52162685  52401619  52457995  52499497
      14 comps  15 comps  16 comps  17 comps  18 comps  19 comps  20 comps
CV    55541799  55577803  55595906  55596046  55594658  55594138  55594171
adjCV 52538679  52573004  52590155  52590190  52588860  52588362  52588394

TRAINING: % variance explained
             1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
X              9.878    39.98    50.83    59.41    64.57    69.53    73.76    77.57    79.85     81.86     83.39     85.11     86.72     89.71     91.37
Actual_price  50.214    59.46    78.13    90.10    94.54    97.34    98.50    99.00    99.40     99.69     99.90     99.97    100.00    100.00    100.00
             16 comps  17 comps  18 comps  19 comps  20 comps
X               92.83     94.58     96.06     97.05     98.19
Actual_price   100.00    100.00    100.00    100.00    100.00
```



INFERENCES :

The mean value of Error in PLS is around 0.16 so far the best. And from the graph the first 15 principal components are chosen because around that value adjusted R square Value is Approximately one.

Based on the Adjusted R-Squares and Mean Value of error We choose PLS is a best model for prediction and it makes sense because we have only 48 observations but 109 variables so from reducing the dimensions drastically and taking into consideration of all Variable the error has been better compared to other Models.