MA717: Applied Regression and Experimental Data Analysis

Assignment template

04-12-2023

Task 1: Data reading and simple exploration (25%)

1.1. Read "College.csv" file into R with following command and use dim() and head() to check if you read the data correct. You should report the number of observations and the number of variables. (5 %)

```r
# To read 'College.csv'
mydata<-read.csv("College.csv", header=T, stringsAsFactors=TRUE)
# use dim()
dim(mydata)
```

```
## [1] 775  17
```

```r
# use head()
head(mydata)
```

```
##   Private Apps Accept Enroll F.Undergrad P.Undergrad Outstate Room.Board Books
## 1     Yes 1660   1232    721        2885         537     7440       3300   450
## 2     Yes 2186   1924    512        2683        1227    12280       6450   750
## 3     Yes 1428   1097    336        1036          99    11250       3750   400
## 4     Yes  417    349    137         510          63    12960       5450   450
## 5     Yes  193    146     55         249         869     7560       4120   800
## 6     Yes  587    479    158         678          41    13500       3335   500
##   Personal PhD Terminal S.F.Ratio perc.alumni Expend Grad.Rate Elite
## 1     2200  70       78      18.1          12   7041        60    No
## 2     1500  29       30      12.2          16  10527        56    No
## 3     1165  53       66      12.9          30   8735        54    No
## 4      875  92       97       7.7          37  19016        59   Yes
## 5     1500  76       72      11.9           2  10922        15    No
## 6      675  67       73       9.4          11   9727        55    No
```

1.2. Use your registration number as random seed, generate a random subset of College data with sample size 700, name this new data as mynewdata. Use summary() to output the summarized information about mynewdata. Please report the number of private and public university and the number of Elite university and non-Elite university in this new data. (12 %)

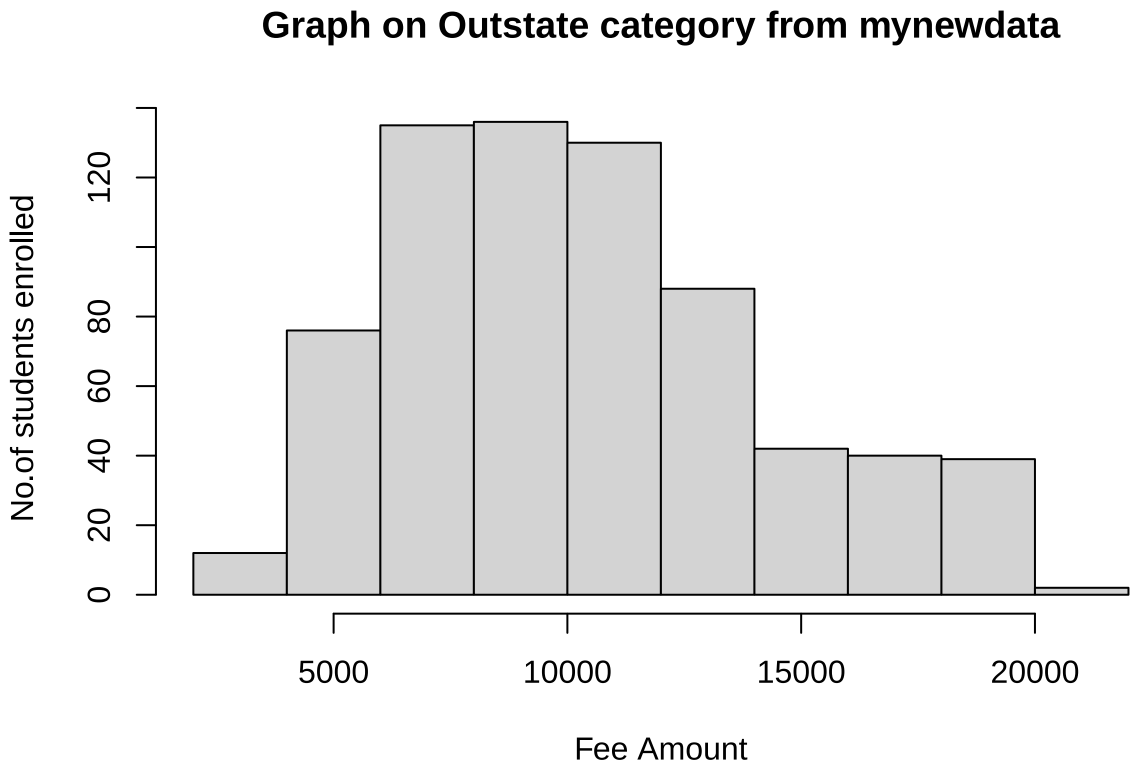
```r
# To set random seed
set.seed(2310158)
# To generate random subset of college data size 700
index<-sample(700)
mynewdata<-mydata[index, ]
# below code shows the summary of mynewdata
summary(mynewdata)
```

```
##  Private        Apps             Accept            Enroll
##  No :193   Min.   :   81.0   Min.   :   72.0   Min.   :  35.0
##  Yes:507   1st Qu.:  753.5   1st Qu.:  587.8   1st Qu.: 239.0
##            Median : 1557.0   Median : 1109.5   Median : 435.5
##            Mean   : 3051.4   Mean   : 2051.4   Mean   : 795.9
##            3rd Qu.: 3767.2   3rd Qu.: 2544.0   3rd Qu.: 913.0
##            Max.   :48094.0   Max.   :26330.0   Max.   :6392.0
##   F.Undergrad       P.Undergrad         Outstate       Room.Board
##  Min.   :  139.0   Min.   :    1.0   Min.   : 2340   Min.   :1780
##  1st Qu.:  980.2   1st Qu.:  100.8   1st Qu.: 7248   1st Qu.:3596
##  Median : 1711.5   Median :  379.5   Median : 9900   Median :4192
##  Mean   : 3795.7   Mean   :  889.5   Mean   :10378   Mean   :4357
##  3rd Qu.: 4292.2   3rd Qu.: 1017.2   3rd Qu.:12785   3rd Qu.:5026
##  Max.   :31643.0   Max.   :21836.0   Max.   :21700   Max.   :8124
##      Books           Personal          PhD             Terminal
##  Min.   :  96.0   Min.   : 250.0   Min.   :  8.00   Min.   : 24.00
##  1st Qu.: 450.0   1st Qu.: 878.8   1st Qu.: 62.00   1st Qu.: 71.00
##  Median : 513.5   Median :1200.0   Median : 75.00   Median : 82.00
##  Mean   : 549.8   Mean   :1340.3   Mean   : 72.75   Mean   : 79.74
##  3rd Qu.: 600.0   3rd Qu.:1692.5   3rd Qu.: 85.00   3rd Qu.: 92.00
##  Max.   :2340.0   Max.   :6800.0   Max.   :100.00   Max.   :100.00
##    S.F.Ratio      perc.alumni        Expend          Grad.Rate      Elite
##  Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00   No :633
##  1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6749   1st Qu.: 53.00   Yes: 67
##  Median :13.60   Median :21.00   Median : 8372   Median : 65.00
##  Mean   :14.14   Mean   :22.36   Mean   : 9549   Mean   : 65.06
##  3rd Qu.:16.52   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 77.00
##  Max.   :39.80   Max.   :63.00   Max.   :56233   Max.   :100.00
```

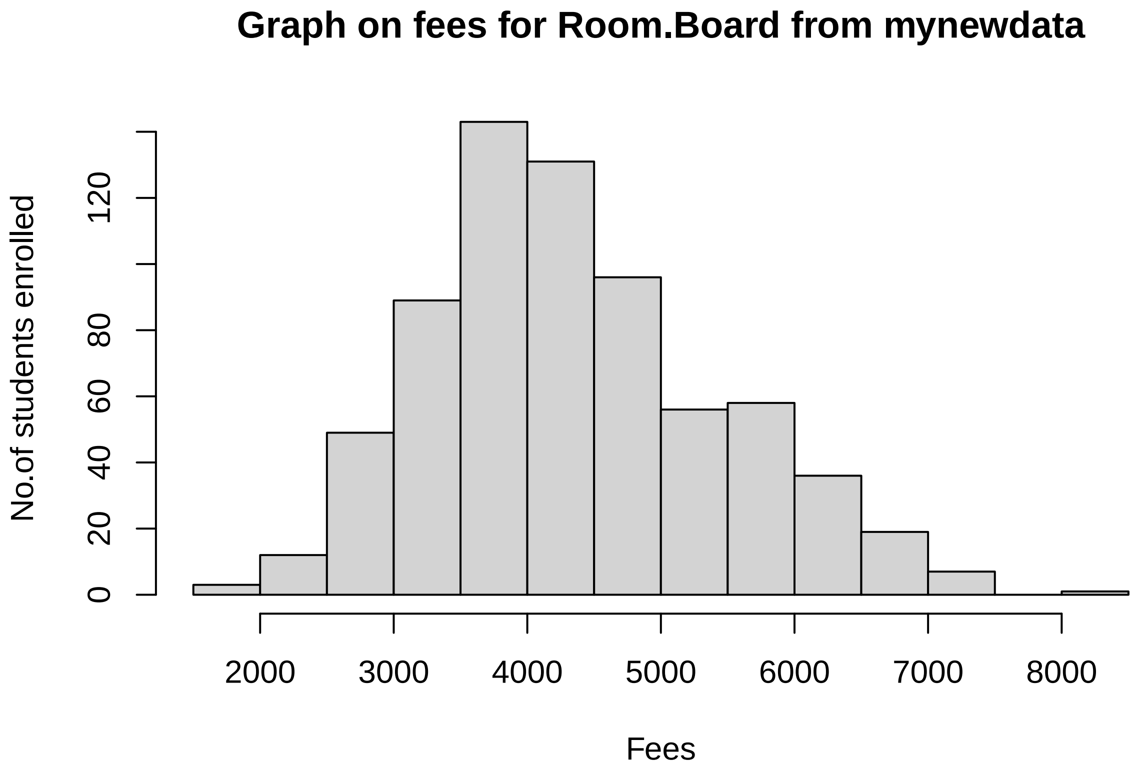
The above summary of my new data depicts the count of public, private, Elite and non-elite universities from the provided random data set sample size of 700. Therefore the inferences are as follows: No.of Private universities:507 No.of Public universities:193 No.of Elite universities:67 No.of non-elite universities:633

1.3. Use mynewdata, plot histogram plots of four variables "Outstate", "Room.Board", "Books" and "Personal". Give each plot a suitable title and label for x axis and y axis. (8%)

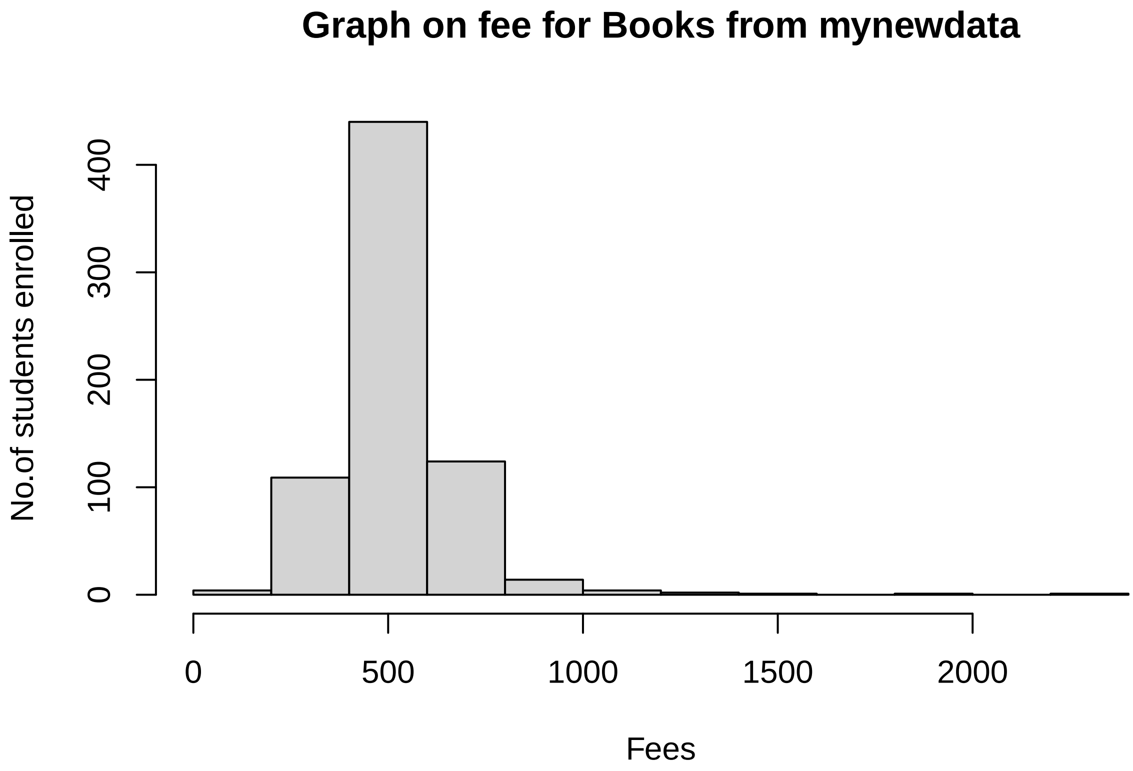
```r
# Histogram plots for Outstate category
hist(mynewdata$Outstate,xlab='Fee Amount',ylab='No.of students enrolled',
     main='Graph on Outstate category from mynewdata')
```



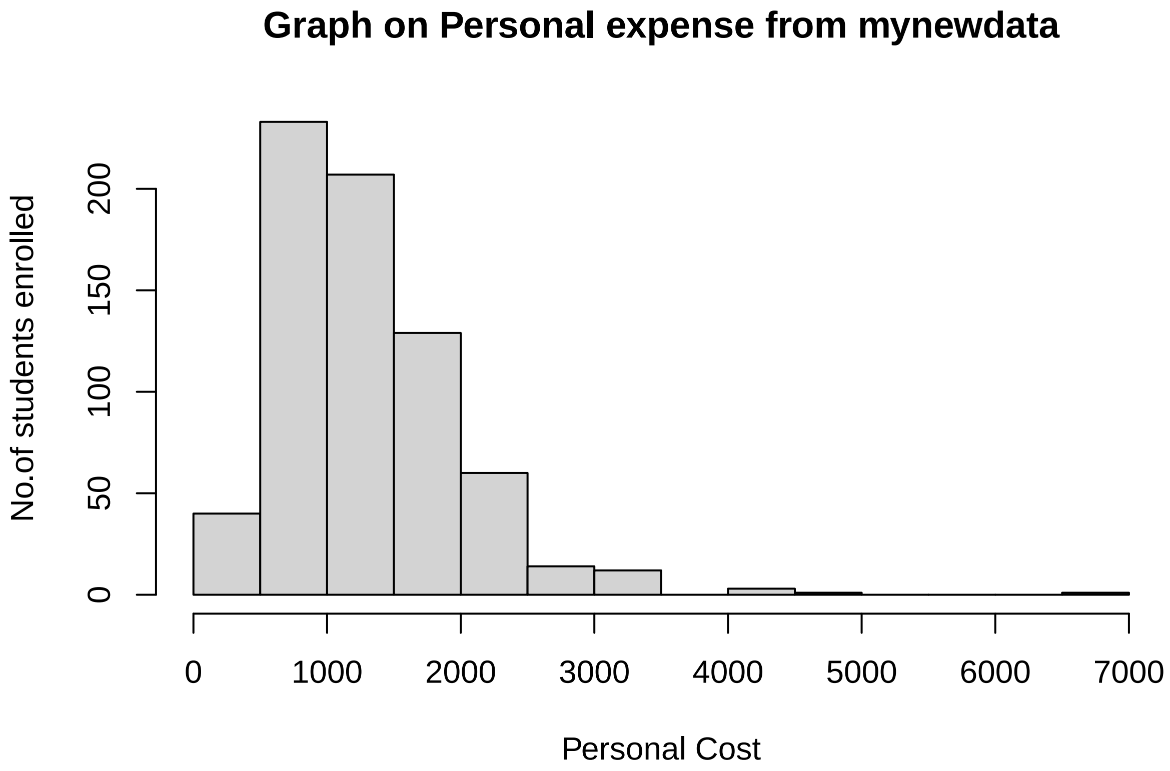
```r
# Histogram plots for Room.Board category
hist(mynewdata$Room.Board,xlab='Fees',ylab='No.of students enrolled',
     main='Graph on fees for Room.Board from mynewdata')
```



```r
# Histogram plots for fee category
hist(mynewdata$Books,xlab='Fees',ylab='No.of students enrolled',
     main='Graph on fee for Books from mynewdata')
```



```r
# Histogram plots for Personal category
hist(mynewdata$Personal,xlab='Personal Cost',ylab='No.of students enrolled',
     main='Graph on Personal expense from mynewdata')
```



Task 2: Linear regression (45%)

2.1. Use mynewdata, do a linear regression fitting when outcome is "Grad.Rate" and predictors are "Private" and "Elite". Show the R output and report what you have learned from this output (you need to discuss significance, adjusted R-squared and p-value of F-statistics). (6%).

```
#To set liner model regression with dataset mynewdata with predictors Private and Elite
mynewdata.regression<-lm(Grad.Rate~Private+Elite,data=mynewdata)
# To summarise the linear regression model.
summary(mynewdata.regression)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Private + Elite, data = mynewdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.492 -10.006   0.508  10.508  44.994
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   55.006      1.111  49.510   <2e-16 ***
## PrivateYes    11.486      1.298   8.846   <2e-16 ***
## EliteYes      18.140      1.972   9.198   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 15.32 on 697 degrees of freedom
## Multiple R-squared:  0.199,  Adjusted R-squared:  0.1967
## F-statistic: 86.57 on 2 and 697 DF,  p-value: < 2.2e-16
```

The above linear regression depicts the significance level of the Graduation rate concerning private and public universities. The coefficient shows that the intercept value is 55.006 estimating the Grad. rate with corresponds to Private and Elite universities. The coefficient of Private universities is 11.486 which is an increase with Grad. Rate compared to Non-Private universities. The coefficient of Elite Universities is 18.140 which is an increase with Grad. Rate compared to non-private universities.

The t-value estimates the standard deviation between estimate and zero. The high value of 49.510 describes the coefficients are highly statistically significant.

Pr(<|t|) determines the p-value for the t-test showing the coefficient significance level. The three coefficients (Grad. Rate, Private, Elite) showcase a high significance level as p-values are equal to zero which satisfies the null hypothesis. The *** symbol denotes the high level of significance with the predictors Private and Elite.

The residual standard deviation is 15.32 with a degree of freedom 697(700-3) representing the deviation amount of Grad. Rate with the predictors Private and Elite.

The Multi R-Squared 0.199 illustrates that 19.9 percent of the variance in response variable Grad. Rate with the predictor variables Private and Elite.

The adjusted R-square controls the predictor of the model to avoid overfitting which adjusts to 0.1967. There is not much difference between the Multi R squared and Adjust R-Square value which clearly shows that the model is not overfitting.

The F Statistics indicates 86.57 which helps to check the significance of the model. The p-value is zero which is much less than 86.57 which shows a high level of significance.

This infers that the variance of Grad. Rate depends on Private and Elite universities.

2.2 Use the linear regression fitting result in 2.1, calculate the confidence intervals for the coefficients. Also give the prediction interval of “Grad.Rate” for a new data with Private=“Yes” and Elite=“No”. (4%)

```r
# calculating the confidence interval for the coefficient for lm mynewdata.regression
confint(mynewdata.regression)
```

```
##                2.5 %   97.5 %
## (Intercept) 52.82481 57.18748
## PrivateYes   8.93666 14.03513
## EliteYes    14.26739 22.01171
```

```r
# To give the prediction interval of Grad.rate with new data(Private=Yes and Elite=No)
#Creating a new data point
new_datapoint <- data.frame(Private = rep("Yes", nrow(mynewdata)),
                            Elite = rep("No", nrow(mynewdata)))

# Predicting Grad.Rate with new data point with prediction interval
Expected_prediction <- predict(mynewdata.regression, newdata = new_datapoint,
                               interval = "prediction")
print(Expected_prediction)
```

```
##          fit       lwr       upr
## 1   66.49204 36.37287 96.61121
## 2   66.49204 36.37287 96.61121
## 3   66.49204 36.37287 96.61121
## 4   66.49204 36.37287 96.61121
## 5   66.49204 36.37287 96.61121
## 6   66.49204 36.37287 96.61121
## 7   66.49204 36.37287 96.61121
## 8   66.49204 36.37287 96.61121
## 9   66.49204 36.37287 96.61121
## 10  66.49204 36.37287 96.61121
## 11  66.49204 36.37287 96.61121
## 12  66.49204 36.37287 96.61121
## 13  66.49204 36.37287 96.61121
## 14  66.49204 36.37287 96.61121
## 15  66.49204 36.37287 96.61121
## 16  66.49204 36.37287 96.61121
## 17  66.49204 36.37287 96.61121
## 18  66.49204 36.37287 96.61121
## 19  66.49204 36.37287 96.61121
## 20  66.49204 36.37287 96.61121
## 21  66.49204 36.37287 96.61121
## 22  66.49204 36.37287 96.61121
## 23  66.49204 36.37287 96.61121
## 24  66.49204 36.37287 96.61121
## 25  66.49204 36.37287 96.61121
## 26  66.49204 36.37287 96.61121
## 27  66.49204 36.37287 96.61121
## 28  66.49204 36.37287 96.61121
## 29  66.49204 36.37287 96.61121
## 30  66.49204 36.37287 96.61121
## 31  66.49204 36.37287 96.61121
## 32  66.49204 36.37287 96.61121
## 33  66.49204 36.37287 96.61121
## 34  66.49204 36.37287 96.61121
## 35  66.49204 36.37287 96.61121
## 36  66.49204 36.37287 96.61121
## 37  66.49204 36.37287 96.61121
## 38  66.49204 36.37287 96.61121
## 39  66.49204 36.37287 96.61121
## 40  66.49204 36.37287 96.61121
## 41  66.49204 36.37287 96.61121
## 42  66.49204 36.37287 96.61121
## 43  66.49204 36.37287 96.61121
## 44  66.49204 36.37287 96.61121
## 45  66.49204 36.37287 96.61121
## 46  66.49204 36.37287 96.61121
## 47  66.49204 36.37287 96.61121
## 48  66.49204 36.37287 96.61121
## 49  66.49204 36.37287 96.61121
## 50  66.49204 36.37287 96.61121
## 51  66.49204 36.37287 96.61121
## 52  66.49204 36.37287 96.61121
## 53  66.49204 36.37287 96.61121
## 54  66.49204 36.37287 96.61121
## 55  66.49204 36.37287 96.61121
## 56  66.49204 36.37287 96.61121
## 57  66.49204 36.37287 96.61121
## 58  66.49204 36.37287 96.61121
## 59  66.49204 36.37287 96.61121
## 60  66.49204 36.37287 96.61121
## 61  66.49204 36.37287 96.61121
## 62  66.49204 36.37287 96.61121
## 63  66.49204 36.37287 96.61121
## 64  66.49204 36.37287 96.61121
## 65  66.49204 36.37287 96.61121
## 66  66.49204 36.37287 96.61121
## 67  66.49204 36.37287 96.61121
## 68  66.49204 36.37287 96.61121
## 69  66.49204 36.37287 96.61121
## 70  66.49204 36.37287 96.61121
## 71  66.49204 36.37287 96.61121
## 72  66.49204 36.37287 96.61121
## 73  66.49204 36.37287 96.61121
## 74  66.49204 36.37287 96.61121
## 75  66.49204 36.37287 96.61121
## 76  66.49204 36.37287 96.61121
## 77  66.49204 36.37287 96.61121
## 78  66.49204 36.37287 96.61121
## 79  66.49204 36.37287 96.61121
## 80  66.49204 36.37287 96.61121
## 81  66.49204 36.37287 96.61121
## 82  66.49204 36.37287 96.61121
## 83  66.49204 36.37287 96.61121
## 84  66.49204 36.37287 96.61121
## 85  66.49204 36.37287 96.61121
## 86  66.49204 36.37287 96.61121
## 87  66.49204 36.37287 96.61121
## 88  66.49204 36.37287 96.61121
## 89  66.49204 36.37287 96.61121
## 90  66.49204 36.37287 96.61121
## 91  66.49204 36.37287 96.61121
## 92  66.49204 36.37287 96.61121
## 93  66.49204 36.37287 96.61121
## 94  66.49204 36.37287 96.61121
## 95  66.49204 36.37287 96.61121
## 96  66.49204 36.37287 96.61121
## 97  66.49204 36.37287 96.61121
## 98  66.49204 36.37287 96.61121
## 99  66.49204 36.37287 96.61121
## 100 66.49204 36.37287 96.61121
## 101 66.49204 36.37287 96.61121
## 102 66.49204 36.37287 96.61121
## 103 66.49204 36.37287 96.61121
## 104 66.49204 36.37287 96.61121
## 105 66.49204 36.37287 96.61121
## 106 66.49204 36.37287 96.61121
## 107 66.49204 36.37287 96.61121
## 108 66.49204 36.37287 96.61121
## 109 66.49204 36.37287 96.61121
## 110 66.49204 36.37287 96.61121
## 111 66.49204 36.37287 96.61121
## 112 66.49204 36.37287 96.61121
## 113 66.49204 36.37287 96.61121
## 114 66.49204 36.37287 96.61121
## 115 66.49204 36.37287 96.61121
## 116 66.49204 36.37287 96.61121
## 117 66.49204 36.37287 96.61121
## 118 66.49204 36.37287 96.61121
## 119 66.49204 36.37287 96.61121
## 120 66.49204 36.37287 96.61121
## 121 66.49204 36.37287 96.61121
## 122 66.49204 36.37287 96.61121
## 123 66.49204 36.37287 96.61121
## 124 66.49204 36.37287 96.61121
## 125 66.49204 36.37287 96.61121
## 126 66.49204 36.37287 96.61121
## 127 66.49204 36.37287 96.61121
## 128 66.49204 36.37287 96.61121
## 129 66.49204 36.37287 96.61121
## 130 66.49204 36.37287 96.61121
## 131 66.49204 36.37287 96.61121
## 132 66.49204 36.37287 96.61121
## 133 66.49204 36.37287 96.61121
## 134 66.49204 36.37287 96.61121
## 135 66.49204 36.37287 96.61121
## 136 66.49204 36.37287 96.61121
## 137 66.49204 36.37287 96.61121
## 138 66.49204 36.37287 96.61121
## 139 66.49204 36.37287 96.61121
## 140 66.49204 36.37287 96.61121
## 141 66.49204 36.37287 96.61121
## 142 66.49204 36.37287 96.61121
## 143 66.49204 36.37287 96.61121
## 144 66.49204 36.37287 96.61121
## 145 66.49204 36.37287 96.61121
## 146 66.49204 36.37287 96.61121
## 147 66.49204 36.37287 96.61121
## 148 66.49204 36.37287 96.61121
## 149 66.49204 36.37287 96.61121
## 150 66.49204 36.37287 96.61121
## 151 66.49204 36.37287 96.61121
## 152 66.49204 36.37287 96.61121
## 153 66.49204 36.37287 96.61121
## 154 66.49204 36.37287 96.61121
## 155 66.49204 36.37287 96.61121
## 156 66.49204 36.37287 96.61121
## 157 66.49204 36.37287 96.61121
## 158 66.49204 36.37287 96.61121
## 159 66.49204 36.37287 96.61121
## 160 66.49204 36.37287 96.61121
## 161 66.49204 36.37287 96.61121
## 162 66.49204 36.37287 96.61121
## 163 66.49204 36.37287 96.61121
## 164 66.49204 36.37287 96.61121
## 165 66.49204 36.37287 96.61121
## 166 66.49204 36.37287 96.61121
## 167 66.49204 36.37287 96.61121
## 168 66.49204 36.37287 96.61121
## 169 66.49204 36.37287 96.61121
## 170 66.49204 36.37287 96.61121
## 171 66.49204 36.37287 96.61121
## 172 66.49204 36.37287 96.61121
## 173 66.49204 36.37287 96.61121
## 174 66.49204 36.37287 96.61121
## 175 66.49204 36.37287 96.61121
## 176 66.49204 36.37287 96.61121
## 177 66.49204 36.37287 96.61121
## 178 66.49204 36.37287 96.61121
## 179 66.49204 36.37287 96.61121
## 180 66.49204 36.37287 96.61121
## 181 66.49204 36.37287 96.61121
## 182 66.49204 36.37287 96.61121
## 183 66.49204 36.37287 96.61121
## 184 66.49204 36.37287 96.61121
## 185 66.49204 36.37287 96.61121
## 186 66.49204 36.37287 96.61121
## 187 66.49204 36.37287 96.61121
## 188 66.49204 36.37287 96.61121
## 189 66.49204 36.37287 96.61121
## 190 66.49204 36.37287 96.61121
## 191 66.49204 36.37287 96.61121
## 192 66.49204 36.37287 96.61121
## 193 66.49204 36.37287 96.61121
## 194 66.49204 36.37287 96.61121
## 195 66.49204 36.37287 96.61121
## 196 66.49204 36.37287 96.61121
## 197 66.49204 36.37287 96.61121
## 198 66.49204 36.37287 96.61121
## 199 66.49204 36.37287 96.61121
## 200 66.49204 36.37287 96.61121
## 201 66.49204 36.37287 96.61121
## 202 66.49204 36.37287 96.61121
## 203 66.49204 36.37287 96.61121
## 204 66.49204 36.37287 96.61121
## 205 66.49204 36.37287 96.61121
## 206 66.49204 36.37287 96.61121
## 207 66.49204 36.37287 96.61121
## 208 66.49204 36.37287 96.61121
## 209 66.49204 36.37287 96.61121
## 210 66.49204 36.37287 96.61121
## 211 66.49204 36.37287 96.61121
## 212 66.49204 36.37287 96.61121
## 213 66.49204 36.37287 96.61121
## 214 66.49204 36.37287 96.61121
## 215 66.49204 36.37287 96.61121
## 216 66.49204 36.37287 96.61121
## 217 66.49204 36.37287 96.61121
## 218 66.49204 36.37287 96.61121
## 219 66.49204 36.37287 96.61121
## 220 66.49204 36.37287 96.61121
## 221 66.49204 36.37287 96.61121
## 222 66.49204 36.37287 96.61121
## 223 66.49204 36.37287 96.61121
## 224 66.49204 36.37287 96.61121
## 225 66.49204 36.37287 96.61121
## 226 66.49204 36.37287 96.61121
## 227 66.49204 36.37287 96.61121
## 228 66.49204 36.37287 96.61121
## 229 66.49204 36.37287 96.61121
## 230 66.49204 36.37287 96.61121
## 231 66.49204 36.37287 96.61121
## 232 66.49204 36.37287 96.61121
## 233 66.49204 36.37287 96.61121
## 234 66.49204 36.37287 96.61121
## 235 66.49204 36.37287 96.61121
## 236 66.49204 36.37287 96.61121
## 237 66.49204 36.37287 96.61121
## 238 66.49204 36.37287 96.61121
## 239 66.49204 36.37287 96.61121
## 240 66.49204 36.37287 96.61121
## 241 66.49204 36.37287 96.61121
## 242 66.49204 36.37287 96.61121
## 243 66.49204 36.37287 96.61121
## 244 66.49204 36.37287 96.61121
## 245 66.49204 36.37287 96.61121
## 246 66.49204 36.37287 96.61121
## 247 66.49204 36.37287 96.61121
## 248 66.49204 36.37287 96.61121
## 249 66.49204 36.37287 96.61121
## 250 66.49204 36.37287 96.61121
## 251 66.49204 36.37287 96.61121
## 252 66.49204 36.37287 96.61121
## 253 66.49204 36.37287 96.61121
## 254 66.49204 36.37287 96.61121
## 255 66.49204 36.37287 96.61121
## 256 66.49204 36.37287 96.61121
## 257 66.49204 36.37287 96.61121
## 258 66.49204 36.37287 96.61121
## 259 66.49204 36.37287 96.61121
## 260 66.49204 36.37287 96.61121
## 261 66.49204 36.37287 96.61121
## 262 66.49204 36.37287 96.61121
## 263 66.49204 36.37287 96.61121
## 264 66.49204 36.37287 96.61121
## 265 66.49204 36.37287 96.61121
## 266 66.49204 36.37287 96.61121
## 267 66.49204 36.37287 96.61121
## 268 66.49204 36.37287 96.61121
## 269 66.49204 36.37287 96.61121
## 270 66.49204 36.37287 96.61121
## 271 66.49204 36.37287 96.61121
## 272 66.49204 36.37287 96.61121
## 273 66.49204 36.37287 96.61121
## 274 66.49204 36.37287 96.61121
## 275 66.49204 36.37287 96.61121
## 276 66.49204 36.37287 96.61121
## 277 66.49204 36.37287 96.61121
## 278 66.49204 36.37287 96.61121
## 279 66.49204 36.37287 96.61121
## 280 66.49204 36.37287 96.61121
## 281 66.49204 36.37287 96.61121
## 282 66.49204 36.37287 96.61121
## 283 66.49204 36.37287 96.61121
## 284 66.49204 36.37287 96.61121
## 285 66.49204 36.37287 96.61121
## 286 66.49204 36.37287 96.61121
## 287 66.49204 36.37287 96.61121
## 288 66.49204 36.37287 96.61121
## 289 66.49204 36.37287 96.61121
## 290 66.49204 36.37287 96.61121
## 291 66.49204 36.37287 96.61121
## 292 66.49204 36.37287 96.61121
## 293 66.49204 36.37287 96.61121
## 294 66.49204 36.37287 96.61121
## 295 66.49204 36.37287 96.61121
## 296 66.49204 36.37287 96.61121
## 297 66.49204 36.37287 96.61121
## 298 66.49204 36.37287 96.61121
## 299 66.49204 36.37287 96.61121
## 300 66.49204 36.37287 96.61121
## 301 66.49204 36.37287 96.61121
## 302 66.49204 36.37287 96.61121
## 303 66.49204 36.37287 96.61121
## 304 66.49204 36.37287 96.61121
## 305 66.49204 36.37287 96.61121
## 306 66.49204 36.37287 96.61121
## 307 66.49204 36.37287 96.61121
## 308 66.49204 36.37287 96.61121
## 309 66.49204 36.37287 96.61121
## 310 66.49204 36.37287 96.61121
## 311 66.49204 36.37287 96.61121
## 312 66.49204 36.37287 96.61121
## 313 66.49204 36.37287 96.61121
## 314 66.49204 36.37287 96.61121
## 315 66.49204 36.37287 96.61121
## 316 66.49204 36.37287 96.61121
## 317 66.49204 36.37287 96.61121
## 318 66.49204 36.37287 96.61121
## 319 66.49204 36.37287 96.61121
## 320 66.49204 36.37287 96.61121
## 321 66.49204 36.37287 96.61121
## 322 66.49204 36.37287 96.61121
## 323 66.49204 36.37287 96.61121
## 324 66.49204 36.37287 96.61121
## 325 66.49204 36.37287 96.61121
## 326 66.49204 36.37287 96.61121
## 327 66.49204 36.37287 96.61121
## 328 66.49204 36.37287 96.61121
## 329 66.49204 36.37287 96.61121
## 330 66.49204 36.37287 96.61121
## 331 66.49204 36.37287 96.61121
## 332 66.49204 36.37287 96.61121
## 333 66.49204 36.37287 96.61121
## 334 66.49204 36.37287 96.61121
## 335 66.49204 36.37287 96.61121
## 336 66.49204 36.37287 96.61121
## 337 66.49204 36.37287 96.61121
## 338 66.49204 36.37287 96.61121
## 339 66.49204 36.37287 96.61121
## 340 66.49204 36.37287 96.61121
## 341 66.49204 36.37287 96.61121
## 342 66.49204 36.37287 96.61121
## 343 66.49204 36.37287 96.61121
## 344 66.49204 36.37287 96.61121
## 345 66.49204 36.37287 96.61121
## 346 66.49204 36.37287 96.61121
## 347 66.49204 36.37287 96.61121
## 348 66.49204 36.37287 96.61121
## 349 66.49204 36.37287 96.61121
## 350 66.49204 36.37287 96.61121
## 351 66.49204 36.37287 96.61121
## 352 66.49204 36.37287 96.61121
## 353 66.49204 36.37287 96.61121
## 354 66.49204 36.37287 96.61121
## 355 66.49204 36.37287 96.61121
## 356 66.49204 36.37287 96.61121
## 357 66.49204 36.37287 96.61121
## 358 66.49204 36.37287 96.61121
## 359 66.49204 36.37287 96.61121
## 360 66.49204 36.37287 96.61121
## 361 66.49204 36.37287 96.61121
## 362 66.49204 36.37287 96.61121
## 363 66.49204 36.37287 96.61121
## 364 66.49204 36.37287 96.61121
## 365 66.49204 36.37287 96.61121
## 366 66.49204 36.37287 96.61121
## 367 66.49204 36.37287 96.61121
## 368 66.49204 36.37287 96.61121
## 369 66.49204 36.37287 96.61121
## 370 66.49204 36.37287 96.61121
## 371 66.49204 36.37287 96.61121
## 372 66.49204 36.37287 96.61121
## 373 66.49204 36.37287 96.61121
## 374 66.49204 36.37287 96.61121
## 375 66.49204 36.37287 96.61121
## 376 66.49204 36.37287 96.61121
## 377 66.49204 36.37287 96.61121
## 378 66.49204 36.37287 96.61121
## 379 66.49204 36.37287 96.61121
## 380 66.49204 36.37287 96.61121
## 381 66.49204 36.37287 96.61121
## 382 66.49204 36.37287 96.61121
## 383 66.49204 36.37287 96.61121
## 384 66.49204 36.37287 96.61121
## 385 66.49204 36.37287 96.61121
## 386 66.49204 36.37287 96.61121
## 387 66.49204 36.37287 96.61121
## 388 66.49204 36.37287 96.61121
## 389 66.49204 36.37287 96.61121
## 390 66.49204 36.37287 96.61121
## 391 66.49204 36.37287 96.61121
## 392 66.49204 36.37287 96.61121
## 393 66.49204 36.37287 96.61121
## 394 66.49204 36.37287 96.61121
## 395 66.49204 36.37287 96.61121
## 396 66.49204 36.37287 96.61121
## 397 66.49204 36.37287 96.61121
## 398 66.49204 36.37287 96.61121
## 399 66.49204 36.37287 96.61121
## 400 66.49204 36.37287 96.61121
## 401 66.49204 36.37287 96.61121
## 402 66.49204 36.37287 96.61121
## 403 66.49204 36.37287 96.61121
## 404 66.49204 36.37287 96.61121
## 405 66.49204 36.37287 96.61121
## 406 66.49204 36.37287 96.61121
## 407 66.49204 36.37287 96.61121
## 408 66.49204 36.37287 96.61121
## 409 66.49204 36.37287 96.61121
## 410 66.49204 36.37287 96.61121
## 411 66.49204 36.37287 96.61121
## 412 66.49204 36.37287 96.61121
## 413 66.49204 36.37287 96.61121
## 414 66.49204 36.37287 96.61121
## 415 66.49204 36.37287 96.61121
## 416 66.49204 36.37287 96.61121
## 417 66.49204 36.37287 96.61121
## 418 66.49204 36.37287 96.61121
## 419 66.49204 36.37287 96.61121
## 420 66.49204 36.37287 96.61121
## 421 66.49204 36.37287 96.61121
## 422 66.49204 36.37287 96.61121
## 423 66.49204 36.37287 96.61121
## 424 66.49204 36.37287 96.61121
## 425 66.49204 36.37287 96.61121
## 426 66.49204 36.37287 96.61121
## 427 66.49204 36.37287 96.61121
## 428 66.49204 36.37287 96.61121
## 429 66.49204 36.37287 96.61121
## 430 66.49204 36.37287 96.61121
## 431 66.49204 36.37287 96.61121
## 432 66.49204 36.37287 96.61121
## 433 66.49204 36.37287 96.61121
## 434 66.49204 36.37287 96.61121
## 435 66.49204 36.37287 96.61121
## 436 66.49204 36.37287 96.61121
## 437 66.49204 36.37287 96.61121
## 438 66.49204 36.37287 96.61121
## 439 66.49204 36.37287 96.61121
## 440 66.49204 36.37287 96.61121
## 441 66.49204 36.37287 96.61121
## 442 66.49204 36.37287 96.61121
## 443 66.49204 36.37287 96.61121
## 444 66.49204 36.37287 96.61121
## 445 66.49204 36.37287 96.61121
## 446 66.49204 36.37287 96.61121
## 447 66.49204 36.37287 96.61121
## 448 66.49204 36.37287 96.61121
## 449 66.49204 36.37287 96.61121
## 450 66.49204 36.37287 96.61121
## 451 66.49204 36.37287 96.61121
## 452 66.49204 36.37287 96.61121
## 453 66.49204 36.37287 96.61121
## 454 66.49204 36.37287 96.61121
## 455 66.49204 36.37287 96.61121
## 456 66.49204 36.37287 96.61121
## 457 66.49204 36.37287 96.61121
## 458 66.49204 36.37287 96.61121
## 459 66.49204 36.37287 96.61121
## 460 66.49204 36.37287 96.61121
## 461 66.49204 36.37287 96.61121
## 462 66.49204 36.37287 96.61121
## 463 66.49204 36.37287 96.61121
## 464 66.49204 36.37287 96.61121
## 465 66.49204 36.37287 96.61121
## 466 66.49204 36.37287 96.61121
## 467 66.49204 36.37287 96.61121
## 468 66.49204 36.37287 96.61121
## 469 66.49204 36.37287 96.61121
## 470 66.49204 36.37287 96.61121
## 471 66.49204 36.37287 96.61121
## 472 66.49204 36.37287 96.61121
## 473 66.49204 36.37287 96.61121
## 474 66.49204 36.37287 96.61121
## 475 66.49204 36.37287 96.61121
## 476 66.49204 36.37287 96.61121
## 477 66.49204 36.37287 96.61121
## 478 66.49204 36.37287 96.61121
## 479 66.49204 36.37287 96.61121
## 480 66.49204 36.37287 96.61121
## 481 66.49204 36.37287 96.61121
## 482 66.49204 36.37287 96.61121
## 483 66.49204 36.37287 96.61121
## 484 66.49204 36.37287 96.61121
## 485 66.49204 36.37287 96.61121
## 486 66.49204 36.37287 96.61121
## 487 66.49204 36.37287 96.61121
## 488 66.49204 36.37287 96.61121
## 489 66.49204 36.37287 96.61121
## 490 66.49204 36.37287 96.61121
## 491 66.49204 36.37287 96.61121
## 492 66.49204 36.37287 96.61121
## 493 66.49204 36.37287 96.61121
## 494 66.49204 36.37287 96.61121
## 495 66.49204 36.37287 96.61121
## 496 66.49204 36.37287 96.61121
## 497 66.49204 36.37287 96.61121
## 498 66.49204 36.37287 96.61121
## 499 66.49204 36.37287 96.61121
## 500 66.49204 36.37287 96.61121
## 501 66.49204 36.37287 96.61121
## 502 66.49204 36.37287 96.61121
## 503 66.49204 36.37287 96.61121
## 504 66.49204 36.37287 96.61121
## 505 66.49204 36.37287 96.61121
## 506 66.49204 36.37287 96.61121
## 507 66.49204 36.37287 96.61121
## 508 66.49204 36.37287 96.61121
## 509 66.49204 36.37287 96.61121
## 510 66.49204 36.37287 96.61121
## 511 66.49204 36.37287 96.61121
## 512 66.49204 36.37287 96.61121
## 513 66.49204 36.37287 96.61121
## 514 66.49204 36.37287 96.61121
## 515 66.49204 36.37287 96.61121
## 516 66.49204 36.37287 96.61121
## 517 66.49204 36.37287 96.61121
## 518 66.49204 36.37287 96.61121
## 519 66.49204 36.37287 96.61121
## 520 66.49204 36.37287 96.61121
## 521 66.49204 36.37287 96.61121
## 522 66.49204 36.37287 96.61121
## 523 66.49204 36.37287 96.61121
## 524 66.49204 36.37287 96.61121
## 525 66.49204 36.37287 96.61121
## 526 66.49204 36.37287 96.61121
## 527 66.49204 36.37287 96.61121
## 528 66.49204 36.37287 96.61121
## 529 66.49204 36.37287 96.61121
## 530 66.49204 36.37287 96.61121
## 531 66.49204 36.37287 96.61121
## 532 66.49204 36.37287 96.61121
## 533 66.49204 36.37287 96.61121
## 534 66.49204 36.37287 96.61121
## 535 66.49204 36.37287 96.61121
## 536 66.49204 36.37287 96.61121
## 537 66.49204 36.37287 96.61121
## 538 66.49204 36.37287 96.61121
## 539 66.49204 36.37287 96.61121
## 540 66.49204 36.37287 96.61121
## 541 66.49204 36.37287 96.61121
## 542 66.49204 36.37287 96.61121
## 543 66.49204 36.37287 96.61121
## 544 66.49204 36.37287 96.61121
## 545 66.49204 36.37287 96.61121
## 546 66.49204 36.37287 96.61121
## 547 66.49204 36.37287 96.61121
## 548 66.49204 36.37287 96.61121
## 549 66.49204 36.37287 96.61121
## 550 66.49204 36.37287 96.61121
## 551 66.49204 36.37287 96.61121
## 552 66.49204 36.37287 96.61121
## 553 66.49204 36.37287 96.61121
## 554 66.49204 36.37287 96.61121
## 555 66.49204 36.37287 96.61121
## 556 66.49204 36.37287 96.61121
## 557 66.49204 36.37287 96.61121
## 558 66.49204 36.37287 96.61121
## 559 66.49204 36.37287 96.61121
## 560 66.49204 36.37287 96.61121
## 561 66.49204 36.37287 96.61121
## 562 66.49204 36.37287 96.61121
## 563 66.49204 36.37287 96.61121
## 564 66.49204 36.37287 96.61121
## 565 66.49204 36.37287 96.61121
## 566 66.49204 36.37287 96.61121
## 567 66.49204 36.37287 96.61121
## 568 66.49204 36.37287 96.61121
## 569 66.49204 36.37287 96.61121
## 570 66.49204 36.37287 96.61121
## 571 66.49204 36.37287 96.61121
## 572 66.49204 36.37287 96.61121
## 573 66.49204 36.37287 96.61121
## 574 66.49204 36.37287 96.61121
## 575 66.49204 36.37287 96.61121
## 576 66.49204 36.37287 96.61121
## 577 66.49204 36.37287 96.61121
## 578 66.49204 36.37287 96.61121
## 579 66.49204 36.37287 96.61121
## 580 66.49204 36.37287 96.61121
## 581 66.49204 36.37287 96.61121
## 582 66.49204 36.37287 96.61121
## 583 66.49204 36.37287 96.61121
## 584 66.49204 36.37287 96.61121
## 585 66.49204 36.37287 96.61121
## 586 66.49204 36.37287 96.61121
## 587 66.49204 36.37287 96.61121
## 588 66.49204 36.37287 96.61121
## 589 66.49204 36.37287 96.61121
## 590 66.49204 36.37287 96.61121
## 591 66.49204 36.37287 96.61121
## 592 66.49204 36.37287 96.61121
## 593 66.49204 36.37287 96.61121
## 594 66.49204 36.37287 96.61121
## 595 66.49204 36.37287 96.61121
## 596 66.49204 36.37287 96.61121
## 597 66.49204 36.37287 96.61121
## 598 66.49204 36.37287 96.61121
## 599 66.49204 36.37287 96.61121
## 600 66.49204 36.37287 96.61121
## 601 66.49204 36.37287 96.61121
## 602 66.49204 36.37287 96.61121
## 603 66.49204 36.37287 96.61121
## 604 66.49204 36.37287 96.61121
## 605 66.49204 36.37287 96.61121
## 606 66.49204 36.37287 96.61121
## 607 66.49204 36.37287 96.61121
## 608 66.49204 36.37287 96.61121
## 609 66.49204 36.37287 96.61121
## 610 66.49204 36.37287 96.61121
## 611 66.49204 36.37287 96.61121
## 612 66.49204 36.37287 96.61121
## 613 66.49204 36.37287 96.61121
## 614 66.49204 36.37287 96.61121
## 615 66.49204 36.37287 96.61121
## 616 66.49204 36.37287 96.61121
## 617 66.49204 36.37287 96.61121
## 618 66.49204 36.37287 96.61121
## 619 66.49204 36.37287 96.61121
## 620 66.49204 36.37287 96.61121
## 621 66.49204 36.37287 96.61121
## 622 66.49204 36.37287 96.61121
## 623 66.49204 36.37287 96.61121
## 624 66.49204 36.37287 96.61121
## 625 66.49204 36.37287 96.61121
## 626 66.49204 36.37287 96.61121
## 627 66.49204 36.37287 96.61121
## 628 66.49204 36.37287 96.61121
## 629 66.49204 36.37287 96.61121
## 630 66.49204 36.37287 96.61121
## 631 66.49204 36.37287 96.61121
## 632 66.49204 36.37287 96.61121
## 633 66.49204 36.37287 96.61121
## 634 66.49204 36.37287 96.61121
## 635 66.49204 36.37287 96.61121
## 636 66.49204 36.37287 96.61121
## 637 66.49204 36.37287 96.61121
## 638 66.49204 36.37287 96.61121
## 639 66.49204 36.37287 96.61121
## 640 66.49204 36.37287 96.61121
## 641 66.49204 36.37287 96.61121
## 642 66.49204 36.37287 96.61121
## 643 66.49204 36.37287 96.61121
## 644 66.49204 36.37287 96.61121
## 645 66.49204 36.37287 96.61121
## 646 66.49204 36.37287 96.61121
## 647 66.49204 36.37287 96.61121
## 648 66.49204 36.37287 96.61121
## 649 66.49204 36.37287 96.61121
## 650 66.49204 36.37287 96.61121
## 651 66.49204 36.37287 96.61121
## 652 66.49204 36.37287 96.61121
## 653 66.49204 36.37287 96.61121
## 654 66.49204 36.37287 96.61121
## 655 66.49204 36.37287 96.61121
## 656 66.49204 36.37287 96.61121
## 657 66.49204 36.37287 96.61121
## 658 66.49204 36.37287 96.61121
## 659 66.49204 36.37287 96.61121
## 660 66.49204 36.37287 96.61121
## 661 66.49204 36.37287 96.61121
## 662 66.49204 36.37287 96.61121
## 663 66.49204 36.37287 96.61121
## 664 66.49204 36.37287 96.61121
## 665 66.49204 36.37287 96.61121
## 666 66.49204 36.37287 96.61121
## 667 66.49204 36.37287 96.61121
## 668 66.49204 36.37287 96.61121
## 669 66.49204 36.37287 96.61121
## 670 66.49204 36.37287 96.61121
## 671 66.49204 36.37287 96.61121
## 672 66.49204 36.37287 96.61121
## 673 66.49204 36.37287 96.61121
## 674 66.49204 36.37287 96.61121
## 675 66.49204 36.37287 96.61121
## 676 66.49204 36.37287 96.61121
## 677 66.49204 36.37287 96.61121
## 678 66.49204 36.37287 96.61121
## 679 66.49204 36.37287 96.61121
## 680 66.49204 36.37287 96.61121
## 681 66.49204 36.37287 96.61121
## 682 66.49204 36.37287 96.61121
## 683 66.49204 36.37287 96.61121
## 684 66.49204 36.37287 96.61121
## 685 66.49204 36.37287 96.61121
## 686 66.49204 36.37287 96.61121
## 687 66.49204 36.37287 96.61121
## 688 66.49204 36.37287 96.61121
## 689 66.49204 36.37287 96.61121
## 690 66.49204 36.37287 96.61121
## 691 66.49204 36.37287 96.61121
## 692 66.49204 36.37287 96.61121
## 693 66.49204 36.37287 96.61121
## 694 66.49204 36.37287 96.61121
## 695 66.49204 36.37287 96.61121
## 696 66.49204 36.37287 96.61121
## 697 66.49204 36.37287 96.61121
## 698 66.49204 36.37287 96.61121
## 699 66.49204 36.37287 96.61121
## 700 66.49204 36.37287 96.61121
```

```
#Note Output has printed the maximum lines,omitted 367 rows
#" [ reached getOption("max.print") -- omitted 367 rows ]"
```

2.3 Use mynewdata, do a multiple linear regression fitting when outcome is “Grad.Rate”, all other variables as predictors. Show the R output and report what you have learned from this output (you need to discuss significance, adjusted R-squared and p-value of F-statistics). Is linear regression model in 2.3 better than linear regression in 2.1? Use ANOVA to justify your conclusion. **(14%)**

```
#Calculating the multiple linear regression with Grad.Rate as response variable and all
#other variables as predictors.
multi_all<-lm(Grad.Rate~.,data= mynewdata)
# To summarise the multiple regression model.
summary(multi_all)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ ., data = mynewdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.961  -7.217  -0.687   7.313  53.298
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.9174965  4.9870577   6.801 2.27e-11 ***
## PrivateYes   2.9587921  1.8022409   1.642 0.101106
## Apps         0.0017520  0.0004405   3.977 7.73e-05 ***
## Accept      -0.0017153  0.0008588  -1.997 0.046191 *
## Enroll       0.0021302  0.0023232   0.917 0.359512
## F.Undergrad -0.0001418  0.0004055  -0.350 0.726673
## P.Undergrad -0.0017213  0.0003936  -4.373 1.42e-05 ***
## Outstate     0.0012869  0.0002429   5.298 1.58e-07 ***
## Room.Board   0.0019908  0.0006076   3.276 0.001105 **
## Books       -0.0011537  0.0030221  -0.382 0.702752
## Personal    -0.0015440  0.0008114  -1.903 0.057487 .
## PhD          0.1831051  0.0592415   3.091 0.002077 **
## Terminal    -0.0829992  0.0650230  -1.276 0.202227
## S.F.Ratio   -0.0277477  0.1665235  -0.167 0.867711
## perc.alumni  0.3449700  0.0507631   6.796 2.35e-11 ***
## Expend      -0.0006445  0.0001753  -3.675 0.000256 ***
## EliteYes     3.6753841  2.0872700   1.761 0.078710 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.73 on 683 degrees of freedom
## Multiple R-squared:  0.4586, Adjusted R-squared:  0.4459
## F-statistic: 36.16 on 16 and 683 DF,  p-value: < 2.2e-16
```

The output of the above multiple linear regression model has 16 predictor variables to predict Grad. Rate. The metrics of this model are as below:

Residuals(the difference between observed and predicted values)are -48.961 and 53.298. The intercept 33.9175 shows the actual value when the remaining predictors are zero. The Predictor PrivateYes tend to increase Grad. Rate approximately 2.9587 times, however not statistically significant. ELiteYes tend to increase Grad. Rate approximately 3.6754 with a 10 percent significance level. The remaining predictors almost fall near zero range to impact the Grad. Rate, however variables such as Apps, Accept, P.Undergrad, Outstate, Room.Board, PhD, perc. alumni, Expend shows a statistically significant level.

The predictors Enroll, F.Undergrad, Books, Personal, Terminal, S.F.Ratio, and EliteYes do not appear to have statistically significant effects on 'Grad. Rate' as their p-values (p > 0.05).

The R-squared value is 0.4586 which shows a 45.86 percentage of variance in Grad. Rate concerning predictors in this model and adjusted R-Square value is 0.4459 which is a slightly lower difference than R-squared value indicating limiting the model complexity.

The F-Statistics test values 36.16 and p-value < 2.2e-16 which are less than the F-Statistic values show that the model is statistically significant.

This concludes that this model has a 45.86 percent variance in Grad. Rate with that of all predictors included. Some predictors show no significant relationship with Grad. Rate, while some predictors show a higher significant association with Grad. Rate. This model helps to have better refinement of highly significant predictors to improve the accuracy of better fit.

```
# Using ANOVA to compare the linear regression model to Multiple regression model
anova(mynewdata.regression,multi_all)
```

```
## Analysis of Variance Table
##
## Model 1: Grad.Rate ~ Private + Elite
## Model 2: Grad.Rate ~ Private + Apps + Accept + Enroll + F.Undergrad +
##     P.Undergrad + Outstate + Room.Board + Books + Personal +
##     PhD + Terminal + S.F.Ratio + perc.alumni + Expend + Elite
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    697 163673
## 2    683 110622 14     53051 23.396 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above ANOVA table helps us to predict the better model for this analaysis.Model 1 analyse the response variable Grad.Rate with two predictors Private and Elite. where as Model 2 analyse the response variable with Grad.Rate with all predictors in the data set.

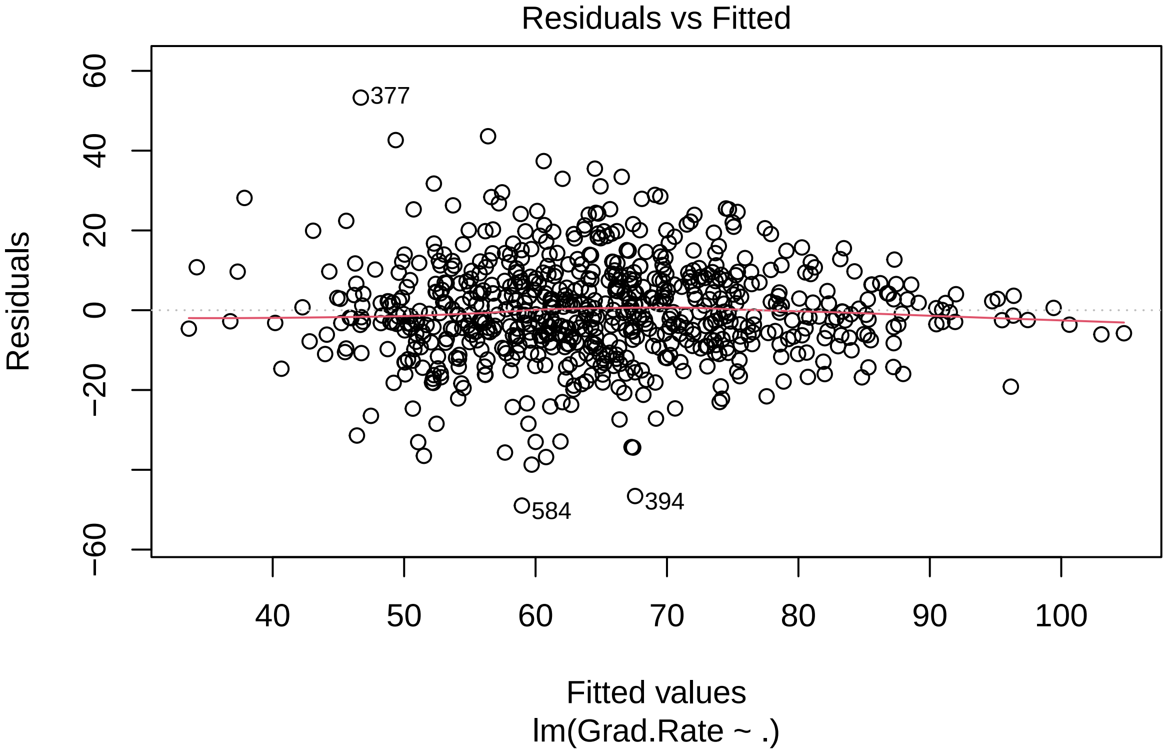
The Residual Degree of Freedom for model1 is 697 and model 2 is 683.The Residual Sum of Squares for model 1(163673) is higher than model 2(110622).Model 2 has sum of squares is 53051 which indicates the variability with multiple predictors when compared to model 1.

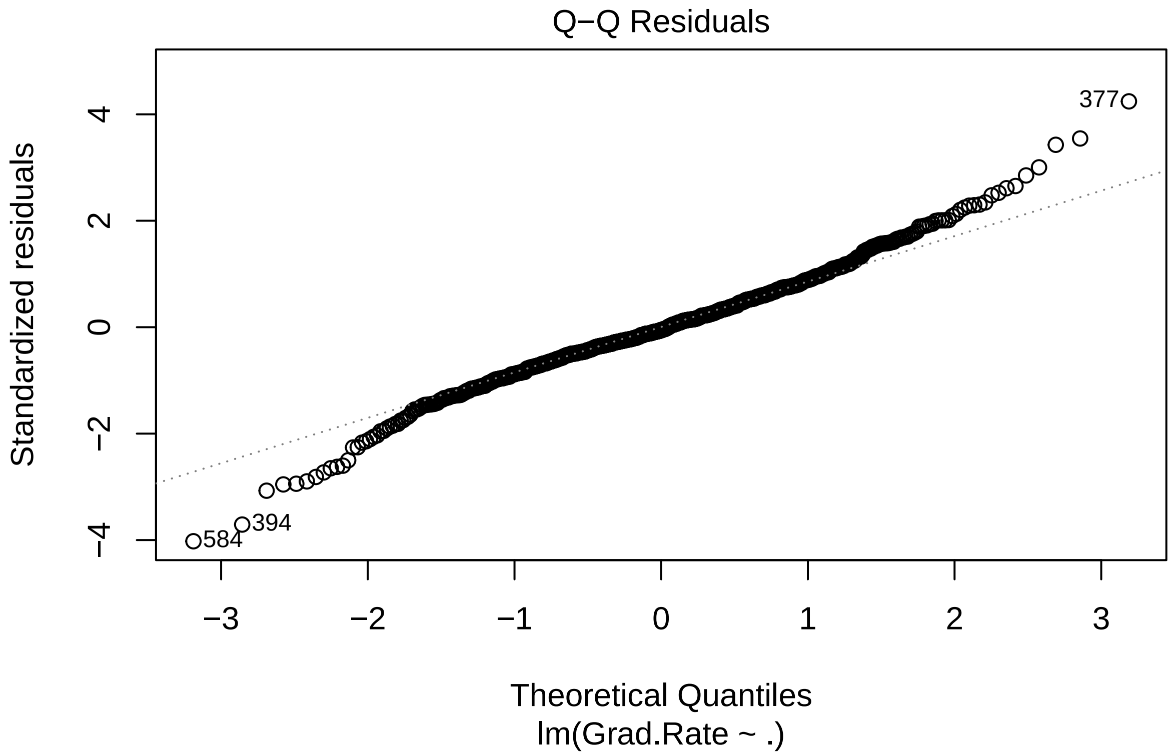
The F-Satistics is 23.396 shows significant improvement in fit of model 2 when compared to model 1. The p-value is very low which shows statistically high significant for model 2 when compared to model 1.

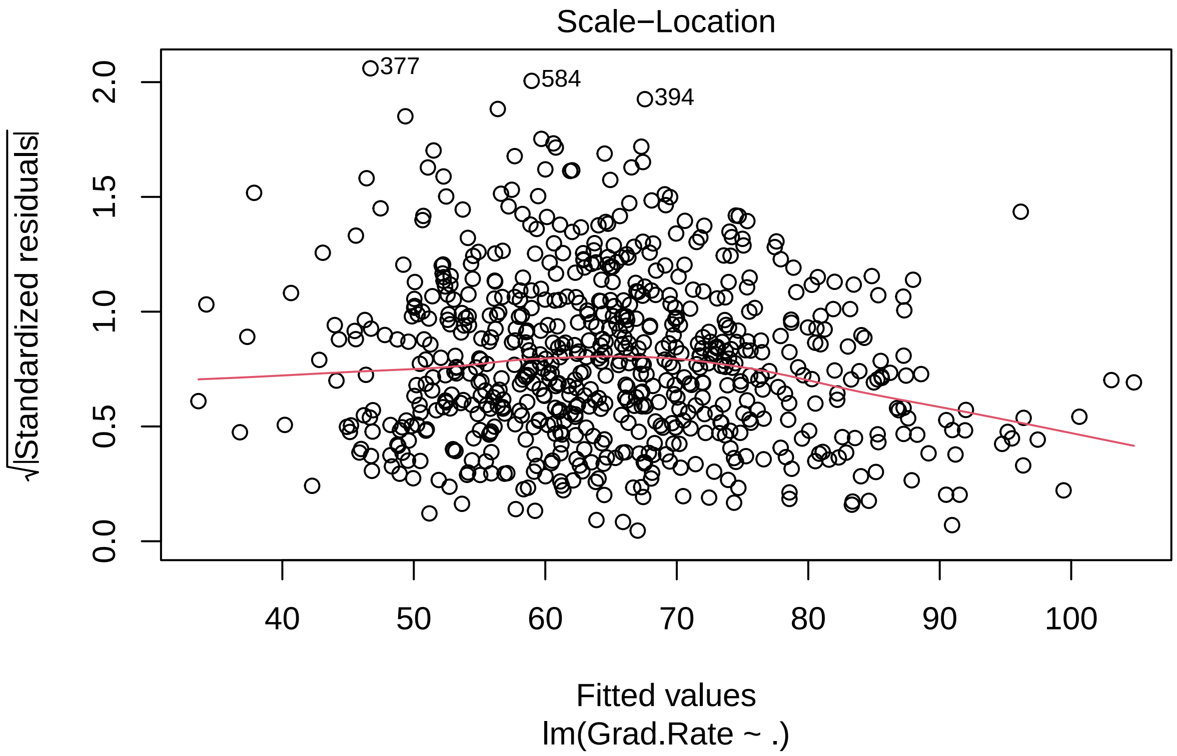
To conclude , this ANOVA table shows Model 2, with all predictors shows significantly better than model 1(where Private and Elite are the predictors). The lower RSS value of model 2 with more number of predictors shows that predictors contribute significantly on variation in the Grad. Rate. Hence, the model 2 provides a better fit than model 1.

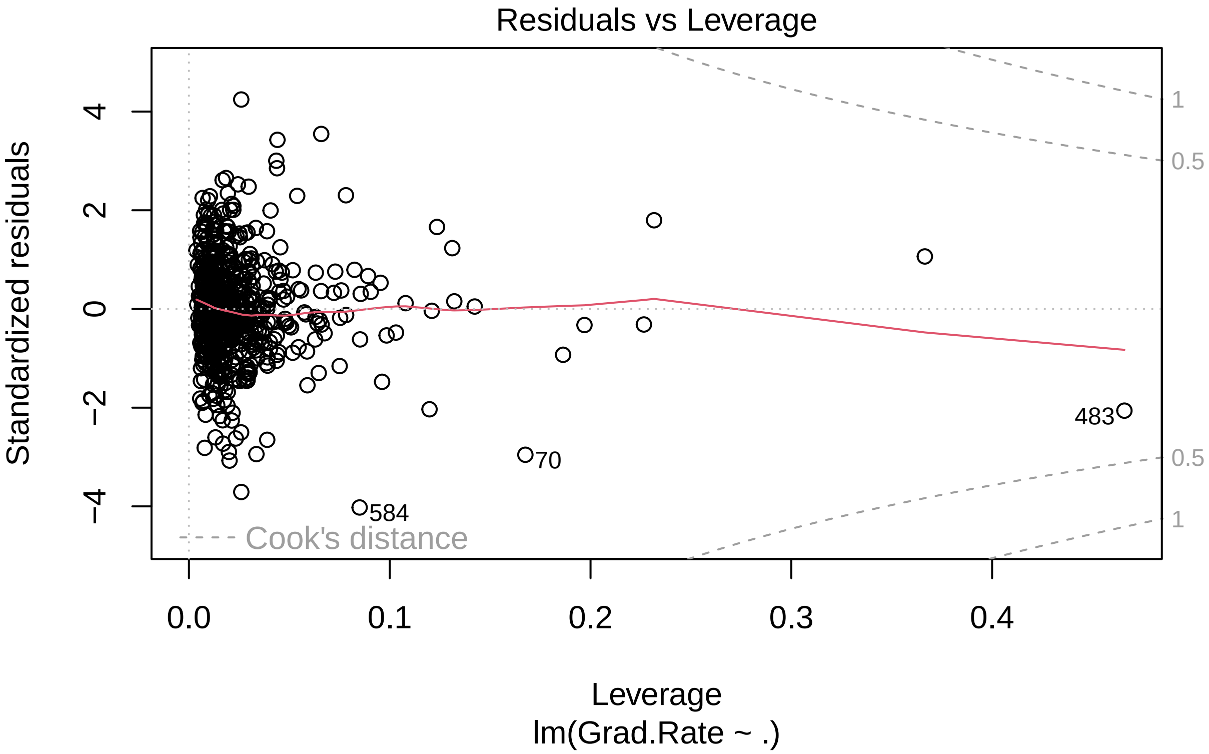
2.4. Use the diagnostic plots to look at the fitting of multiple linear regression in 2.3. Please comment what you have seen from those plots. (7%)

```
# To plot the fitting of multiple regression model
plot(multi_all)
```




Q−Q Residuals
Standardized residuals
−4
−2
0
2
4
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(Grad.Rate ~ .)
377
394
584


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
377
584
394
40
50
60
70
80
90
100
Fitted values
lm(Grad.Rate ~ .)



The above four graph plots are explained below:

1. Residuals Vs Fitted: The plots are scattered randomly around zero points. Also, the red line is almost near the zero point dotted lines which depicts that Grad. Rate is dependent on all other predictors.
2. Q-Q Residuals: This Quantile-Quantile point helps to check the normality of predictors. The points almost fall on the diagonal line, however towards the end, there is a slight variation which shows that variables are not fit on this model. By removing these the multiple regression fitting can be improved.
3. Scale-Location Plot: The plots are randomly scattered near the red line area. Some plots are scattered far away from the red line which represents that there are few unequal variables in residuals.
4. Residuals vs Leverage: This graph clearly shows the high influencing predictors among all other predictors concerning the leverage of Grad.Rate.There are a few plots that are scattered far away in the graph which clearly shows that some variables have low variance. The high value of residual and leverage will clearly show their influence on this multiple regression model.

By conclusion from the above graph, the different plots clearly show us the highly influential variables, and assumptions on the regression model and determine the reliability of the model.

2.5. Use mynewdata, do a variable selection to choose the best model. You should use plots to justify how do you choose your best model. Use the selected predictors of your best model with outcome "Grad.Rate", do a linear regression fitting and plot the diagnostic plots for this fitting. You can use either exhaustive, or forward, or backward selection method. (14%)

```r
# Using the library leaps
library(leaps)
```

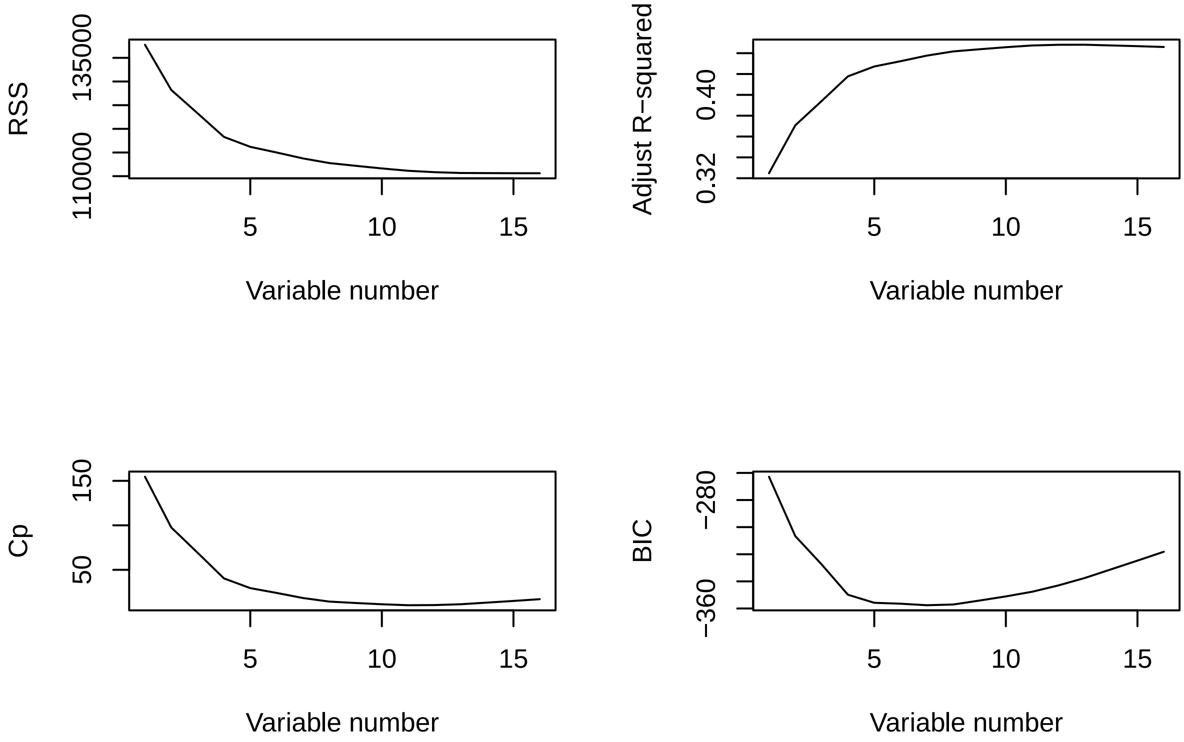
```
## Warning: package 'leaps' was built under R version 4.3.2
```

```r
# Set the regression subset to conduct exhaustive search through maximum 16 predictors
#to Graduation rates(Grad.Rate)
data_regsubset <- regsubsets(Grad.Rate~., data=mynewdata, nvmax=16)
# To summarise the regsubset under the label data_regsubset_sum
data_regsubset_sum <- summary(data_regsubset)
# To set the function for 2X2 plotting grid
par(mfrow=c(2,2))
#To set plotting to visualize the different relationship between the variable aiming to
#find a good fit model.
plot(data_regsubset_sum$rss, xlab="Variable number", ylab="RSS", type="l")

plot(data_regsubset_sum$adjr2, xlab="Variable number", ylab="Adjust R-squared", type="l")

plot(data_regsubset_sum$cp, xlab="Variable number", ylab="Cp", type="l")

plot(data_regsubset_sum$bic, xlab="Variable number", ylab="BIC", type="l")
```



```r
# To find the highest Adjusted R-Squared
which.max(data_regsubset_sum$adjr2)
```

```
## [1] 13
```

```r
# To find the lowest cp(Mallow's Cp)
which.min(data_regsubset_sum$cp)
```

```
## [1] 11
```

```r
# To find the lowest bic(Bayesian Information Criterion)
which.min(data_regsubset_sum$bic)
```

```
## [1] 7
```

```r
# To select the coefficient based on the specific criterion
coef(data_regsubset, 7)
```

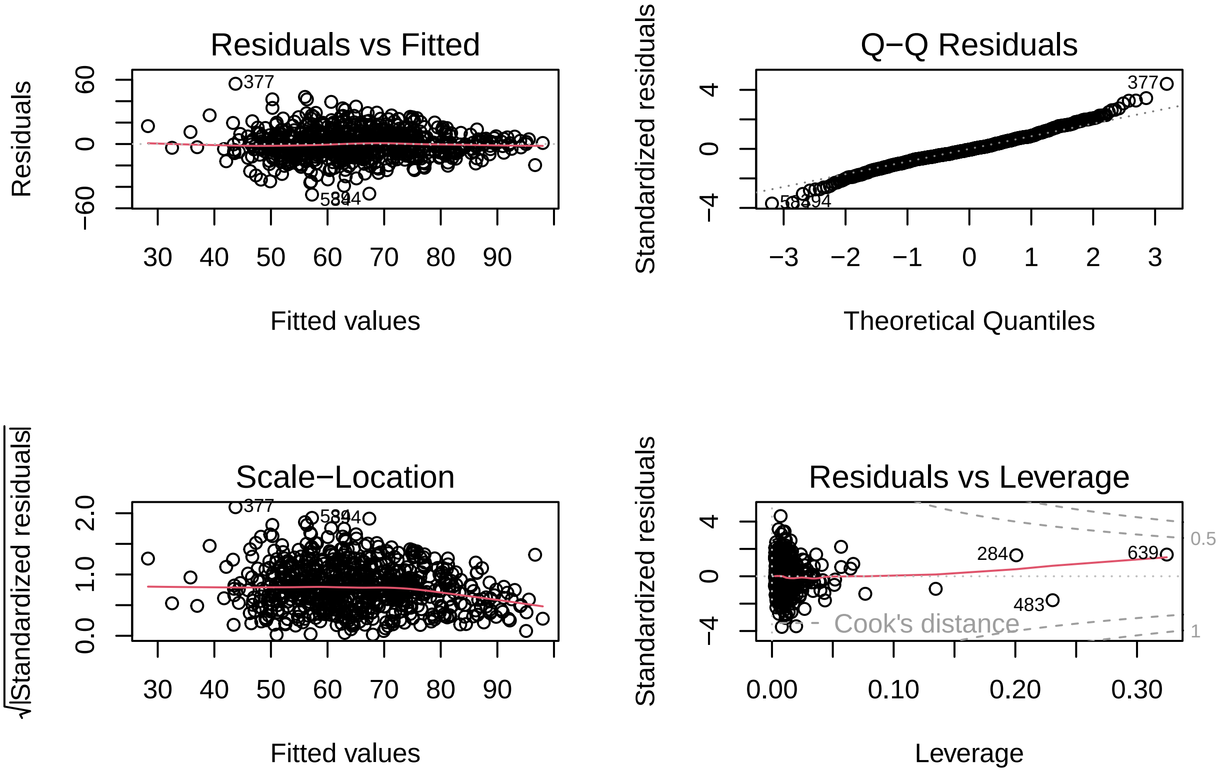
```
##   (Intercept)          Apps   P.Undergrad      Outstate    Room.Board 
## 28.7419969346  0.0009095431 -0.0021458594  0.0013998898  0.0021332326 
##           PhD   perc.alumni        Expend 
##  0.1011400592  0.3871144012 -0.0004591539
```

```r
# To set a fit of linear model with 7 predictors
lm_select_model = lm(Grad.Rate~Apps+P.Undergrad+Outstate+Room.Board+PhD+perc.alumni+Expend,
                     data = mynewdata)
# To summarise the lm_select_model
summary(lm_select_model)
```

```
## 
## Call:
## lm(formula = Grad.Rate ~ Apps + P.Undergrad + Outstate + Room.Board + 
##     PhD + perc.alumni + Expend, data = mynewdata)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -47.259  -7.352  -0.693   7.377  56.258 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 28.7419969  2.7410444  10.486  < 2e-16 ***
## Apps         0.0009095  0.0001483   6.135 1.43e-09 ***
## P.Undergrad -0.0021459  0.0003581  -5.992 3.34e-09 ***
## Outstate     0.0013999  0.0002151   6.509 1.45e-10 ***
## Room.Board   0.0021332  0.0005916   3.606 0.000333 ***
## PhD          0.1011401  0.0366389   2.760 0.005925 ** 
## perc.alumni  0.3871144  0.0492282   7.864 1.43e-14 ***
## Expend      -0.0004592  0.0001471  -3.121 0.001880 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 12.82 on 692 degrees of freedom
## Multiple R-squared:  0.4433, Adjusted R-squared:  0.4377
## F-statistic: 78.72 on 7 and 692 DF,  p-value: < 2.2e-16
```

```r
#To generate the diagnostic plots for the lm_select_model
plot(lm_select_model)
```



The exhaustive method of variable selection has been used here to check the different combinations of predictors to a maximum variables of 16. The regsubsets function helps to consider all possible combinations and evaluate each combination separately.

The best model based on different criteria is explained with a maximum value of sum$adjr2 and a minimum value of cp and bic.

The RSS graph plot shows that the number of variables is proportional to the residual sum of squares. The adjusted R-squared value shows the variance proportion between dependent and independent variables. As the no. of variables increases the adjusted R-Square value increases. The Cp graph indicates model fitness and helps to improve further fitness of the model. The BIC graph helps to avoid the overfit range in this model. The minimum value of BIC and Cp and the maximum value of adjusted R-square and RSS show the model fitting.

This model includes 7 predictors. The intercept value has been reduced to 28.74 when compared to the multiple regression model intercept of 33.91

Apps, Outstate, Room. Board, PhD, perc. alumni have positive effects on the graduation rate has a positive impact on Grad. Rate

P. Undergrad, Expend has negative effects on the graduation rate has a negative impact on Grad. Rate. All predictors except Room.Board, PhD, and Expend exhibit statistical significance at various levels (indicated by p-values).

Concerning model performance, the minimum and maximum residuals are -47.259 and 56.258, and the residual standard error is 12.82

Multiple R-squared is 0.4433, Adjusted R-squared is 0.4377 This model explains approximately 44.33% of the variance in the graduation rate.

The F-Statistic value of 78.72 with a very low p-value shows that the model is statistically significant.

The graphical representation of the plots in Residual Vs fitted shows that the plots are collectively near the red line. Few plots of variables show variance by falling away from the red line reducing the model fit value. The Q-Q residuals show a significant cluster of plots falling on the red line. The scale location graph shows random plotting of variables around the red line. The Residuals Vs leverage accumulates near the zero value. The scattered remaining minimal plots show the least impact on Grad. Rate.

Hence the predictors contribute to the variance in graduation rate.

Task 3: Open question (30%)

Use mynewdata, discuss and perform any step(s) that you think that can improve the fitting in Task 2. You need to illustrate your work by using the R codes, output and discussion.

```
# Transformation of Grad.Rate variable to make the distribution more symmetrical
#using square root
mynewdata$Grad.Rate = sqrt(mynewdata$Grad.Rate)
# Standardization of variable perc.alumni to make the mean as 0 and standard deviation
#as 1 using scale.
mynewdata$perc.alumni = scale(mynewdata$perc.alumni)
# Using logarithmic transformation to variable Apps,P.Undergrad, Outstate,PhD,Expend to
# improve the skewness and making relationship between the variable as more linear.
mynewdata$Apps = log(mynewdata$Apps)

mynewdata$P.Undergrad = log(mynewdata$P.Undergrad)

mynewdata$Outstate = log(mynewdata$Outstate)

mynewdata$PhD = log(mynewdata$PhD)

mynewdata$Expend = log(mynewdata$Expend)
# To set a regression model using the transformed variables
lm_trans = lm(Grad.Rate~Apps+P.Undergrad+Outstate+Room.Board+PhD+perc.alumni+Expend,
              data = mynewdata)
# To summarise the transformed regression model
summary(lm_trans)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Apps + P.Undergrad + Outstate + Room.Board +
##     PhD + perc.alumni + Expend, data = mynewdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8919 -0.4357  0.0075  0.5076  3.3793
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.224e-02  1.260e+00  -0.057 0.954300
## Apps         2.294e-01  3.925e-02   5.844 7.83e-09 ***
## P.Undergrad -7.061e-02  2.703e-02  -2.612 0.009205 **
## Outstate     1.017e+00  1.375e-01   7.393 4.16e-13 ***
## Room.Board   1.454e-04  4.065e-05   3.576 0.000373 ***
## PhD          2.416e-01  1.426e-01   1.694 0.090665 .
## perc.alumni  3.069e-01  4.151e-02   7.393 4.17e-13 ***
## Expend      -4.652e-01  1.295e-01  -3.592 0.000351 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8546 on 692 degrees of freedom
## Multiple R-squared:  0.4143, Adjusted R-squared:  0.4084
## F-statistic: 69.93 on 7 and 692 DF,  p-value: < 2.2e-16
```

To improve the fitting, a possible transformation of 7 variables has been performed. Sometimes transformation of the data set helps to improvise the better fit further.

Below are the comparison details of lm_select_model(same 7 variables without transformation) and lm_trans(same 7 variables with transformation)

The lm_select_model and lm_trans: both models have the same number of predictors. The difference lies in the values of residuals, intercepts, and resulting R-squared.

The lm_select_model depicts a higher residual value when compared to lm_trans which indicates that lm_select_model has more errors than that of lm_trans.

The residual standard error of lm_select_model is 12.82 which is higher than lm_trans model 0.8546 representing that lm_trans has a good fit.

The multiple R-squared of lm_trans is 0.4143 which is lower than lm_select_model 0.4433 indicating that lm_trans explains better variance with graduation rate.

This concludes that the lm_trans model has a better fit due to low residual standard error, higher R-squared, and better intercept value indicating a more accurate prediction of graduation rates based on the given predictors compared to lm_model_select.