Linear Regression Project

Swarup

6 September 2019

Introduction

The company is trying to decide whether to focus their efforts on their mobile app experience or their website. Following is predict is analysis for this company

Read the File as Ecommerce.Customers

```
Ecommerce.Customers=read.csv(file.choose())
head(Ecommerce.Customers)
```

```
##                          Email
## 1     mstephenson@fernandez.com
## 2             hduke@hotmail.com
## 3              pallen@yahoo.com
## 4       riverarebecca@gmail.com
## 5 mstephens@davidson-herman.com
## 6        alvareznancy@lucas.biz
##                                          Address
## 1               835 Frank Tunnel\nWrightmouth, MI 82180-9605
## 2             4547 Archer Common\nDiazchester, CA 06566-8576
## 3 24645 Valerie Unions Suite 582\nCobbborough, DC 99414-7564
## 4           1414 David Throughway\nPort Jason, OH 22070-1220
## 5    14023 Rodriguez Passage\nPort Jacobville, PR 37242-1057
## 6    645 Martha Park Apt. 611\nJeffreychester, MN 67218-7250
##             Avatar Avg..Session.Length Time.on.App Time.on.Website
## 1           Violet            34.49727    12.65565        39.57767
## 2        DarkGreen            31.92627    11.10946        37.26896
## 3           Bisque            33.00091    11.33028        37.11060
## 4      SaddleBrown            34.30556    13.71751        36.72128
## 5 MediumAquaMarine            33.33067    12.79519        37.53665
## 6      FloralWhite            33.87104    12.02693        34.47688
##   Length.of.Membership Yearly.Amount.Spent
## 1             4.082621            587.9511
## 2             2.664034            392.2049
## 3             4.104543            487.5475
## 4             3.120179            581.8523
## 5             4.446308            599.4061
## 6             5.493507            637.1024
```

Convert into Dataframe

```
Ecommerce.Customers=data.frame(Ecommerce.Customers)
```

Analyzing the Summary Statistics

```
summary(Ecommerce.Customers[c(-1,-2,-3)])
```

```
##  Avg..Session.Length  Time.on.App     Time.on.Website Length.of.Membership
##  Min.   :29.53       Min.   : 8.508   Min.   :33.91   Min.   :0.2699
```

```
##  1st Qu.:32.34       1st Qu.:11.388   1st Qu.:36.35   1st Qu.:2.9304
##  Median :33.08       Median :11.983   Median :37.07   Median :3.5340
##  Mean   :33.05       Mean   :12.052   Mean   :37.06   Mean   :3.5335
##  3rd Qu.:33.71       3rd Qu.:12.754   3rd Qu.:37.72   3rd Qu.:4.1265
##  Max.   :36.14       Max.   :15.127   Max.   :40.01   Max.   :6.9227
##  Yearly.Amount.Spent
##  Min.   :256.7
##  1st Qu.:445.0
##  Median :498.9
##  Mean   :499.3
##  3rd Qu.:549.3
##  Max.   :765.5
```

```
data.class(Ecommerce.Customers)
```

```
## [1] "data.frame"
```

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 3.5.3
```

```
describe(Ecommerce.Customers)
```

```
##                       vars   n   mean     sd median trimmed    mad    min
## Email*                   1 500 250.50 144.48 250.50  250.50 185.32   1.00
## Address*                 2 500 250.50 144.48 250.50  250.50 185.32   1.00
## Avatar*                  3 500  69.74  41.03  70.50   69.73  54.11   1.00
## Avg..Session.Length      4 500  33.05   0.99  33.08   33.06   1.03  29.53
## Time.on.App              5 500  12.05   0.99  11.98   12.06   1.01   8.51
## Time.on.Website          6 500  37.06   1.01  37.07   37.05   1.03  33.91
## Length.of.Membership     7 500   3.53   1.00   3.53    3.55   0.89   0.27
## Yearly.Amount.Spent      8 500 499.31  79.31 498.89  499.07  76.82 256.67
##                         max  range  skew kurtosis   se
## Email*               500.00 499.00  0.00    -1.21 6.46
## Address*             500.00 499.00  0.00    -1.21 6.46
## Avatar*              138.00 137.00  0.00    -1.27 1.83
## Avg..Session.Length   36.14   6.61 -0.03    -0.01 0.04
## Time.on.App           15.13   6.62 -0.09     0.10 0.04
## Time.on.Website       40.01   6.09  0.01    -0.12 0.05
## Length.of.Membership   6.92   6.65 -0.11     0.32 0.04
## Yearly.Amount.Spent  765.52 508.85  0.03     0.43 3.55
```

Univariate Analysis

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 3.5.3
```

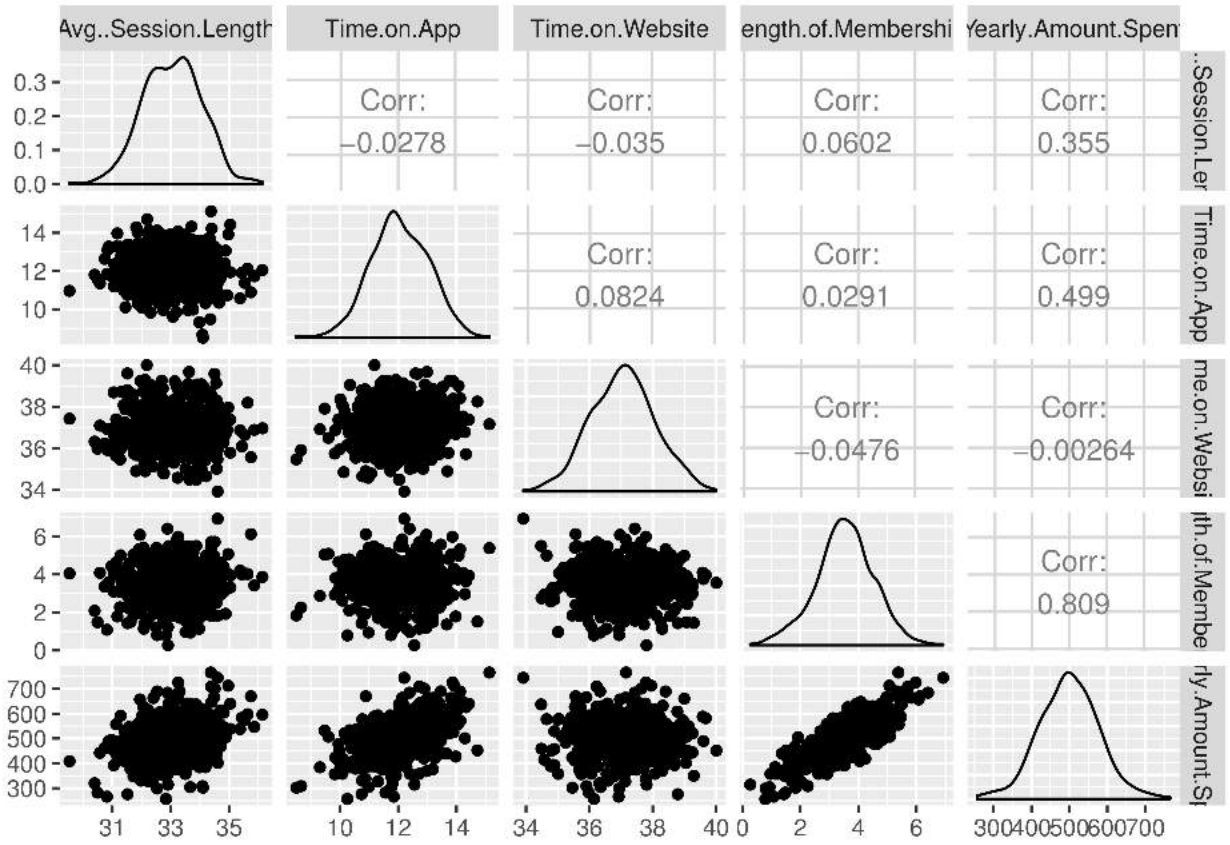
```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
ggpairs(Ecommerce.Customers[c(-1,-2,-3)])
```



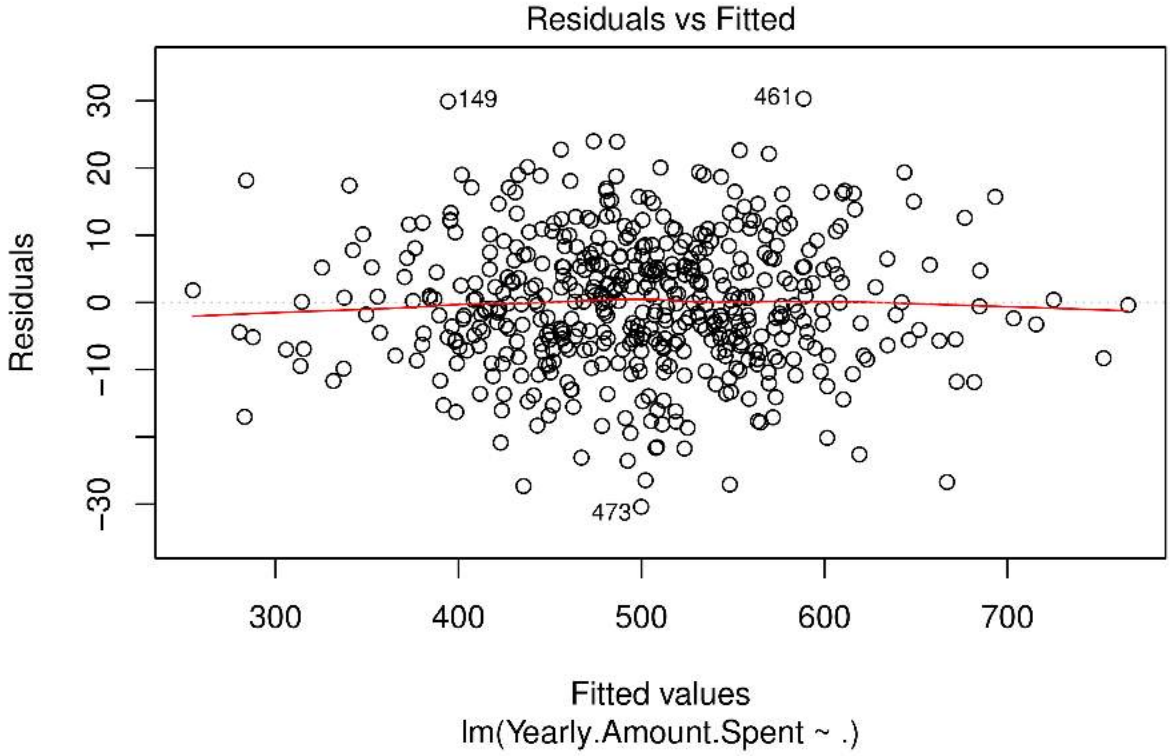
####Dummy Regression Model

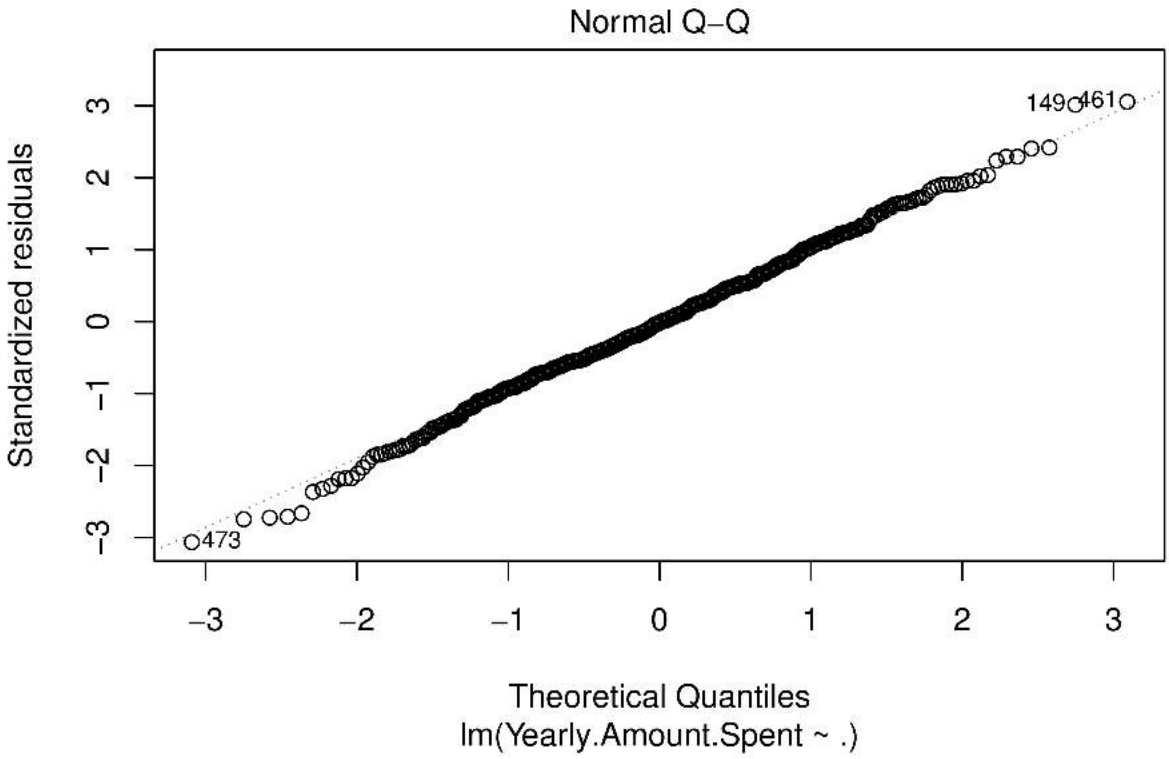
```
model0=lm(Yearly.Amount.Spent~.,data = Ecommerce.Customers[c(-1,-2,-3)])
summary(model0)
```

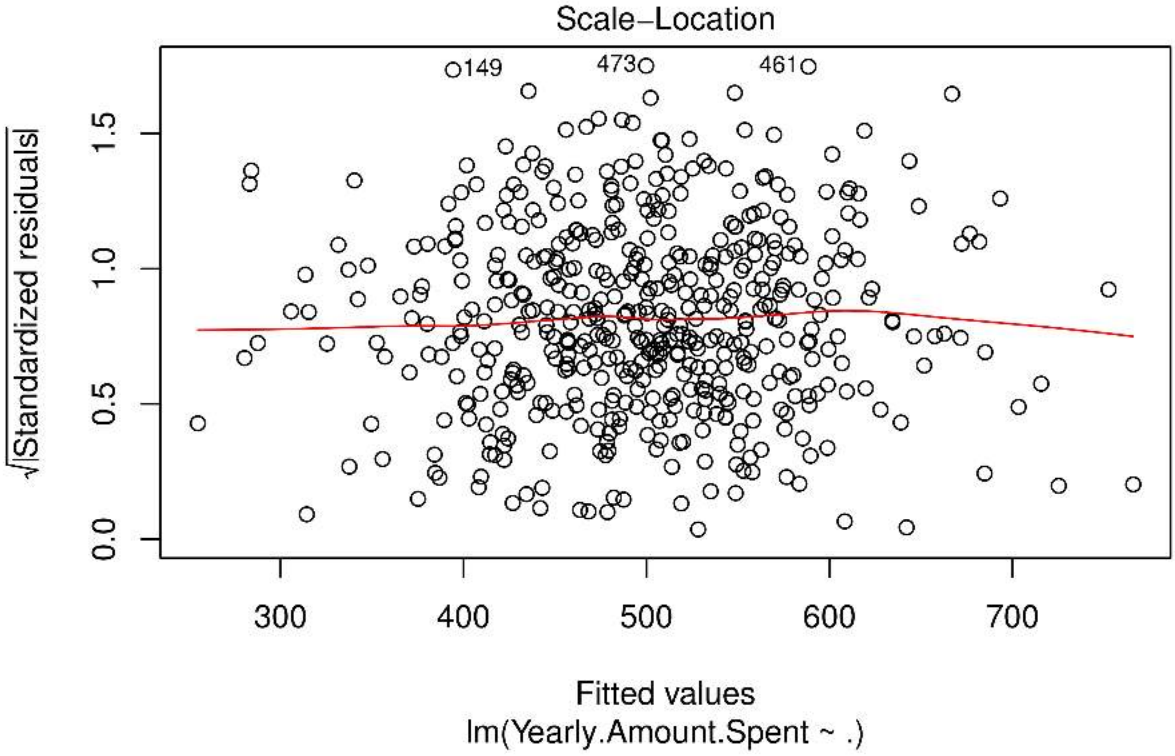
```
##
## Call:
## lm(formula = Yearly.Amount.Spent ~ ., data = Ecommerce.Customers[c(-1,
##     -2, -3)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.4059  -6.2191  -0.1364   6.6048  30.3085
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1051.5943    22.9925 -45.736   <2e-16 ***
## Avg..Session.Length     25.7343     0.4510  57.057   <2e-16 ***
## Time.on.App             38.7092     0.4510  85.828   <2e-16 ***
## Time.on.Website          0.4367     0.4441   0.983    0.326
## Length.of.Membership    61.5773     0.4483 137.346   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.973 on 495 degrees of freedom
```

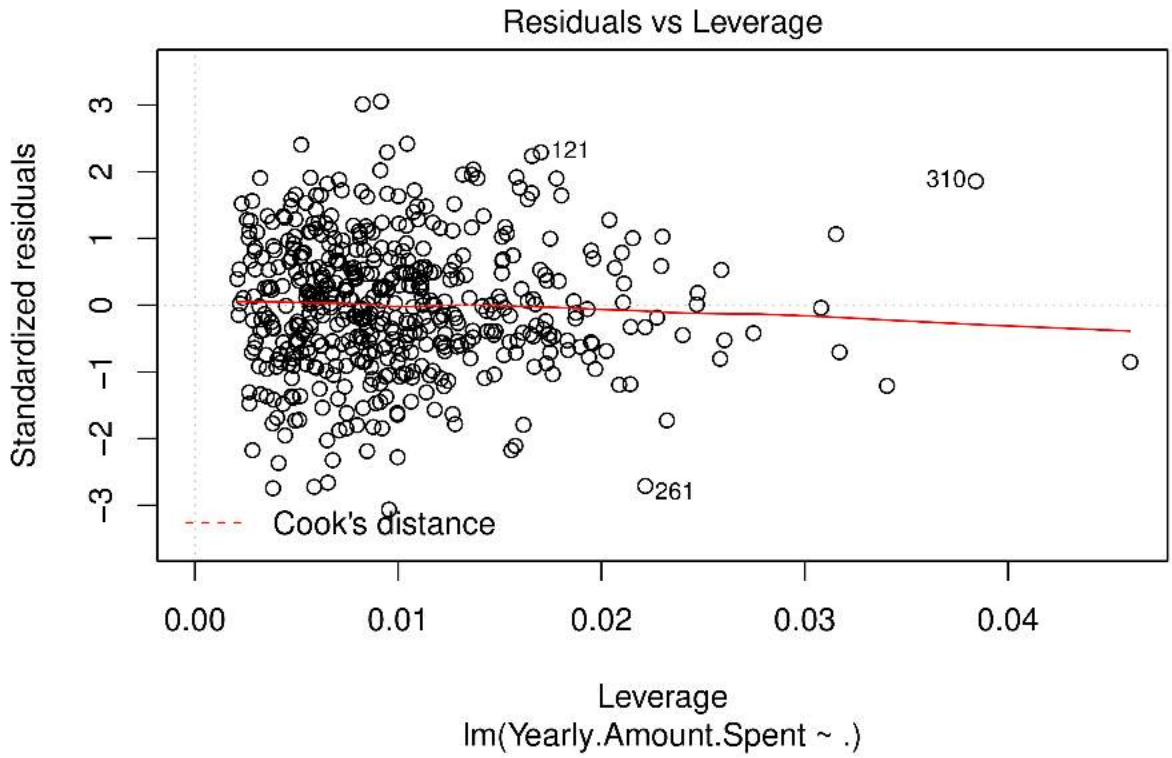
```
## Multiple R-squared:  0.9843, Adjusted R-squared:  0.9842
## F-statistic:  7766 on 4 and 495 DF,  p-value: < 2.2e-16
```

```
abline(plot(model0))
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Yearly.Amount.Spent ~ .)
473
149
461
−3
−2
−1
0
1
2
3


Scale−Location
149
473
461
√|Standardized residuals|
0.0
0.5
1.0
1.5
300
400
500
600
700
Fitted values
lm(Yearly.Amount.Spent ~ .)



** Comments :- Avg..Session.Length,Time.on.App,Length.of.Membership are significant variables **

** Tested using Corrplot ** #### Correlation Plot

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.5.3
```

```
## corrplot 0.84 loaded
```

```
corrplot(cor(Ecommerce.Customers[c(-1,-2,-3)]))
```

Tolerance 1-R2 VIF VIF =1/(1-Tolerance)

Observations:-

- ** High Adjusted R2 value 98.35% of yearly amount spent in the test data is explained by Avg..Session.Length + Time.on.App + Length.of.Membership Tolerace is high 1-R2 (0.0165 Certainly High Multicollinearity should be 0.1-0.01) VIF =1/(1-Tolerance) (1/(1-0.0165)) =1.016 (Moderately Multicollinearity) Range should be 10-100)**

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.5.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 3.5.2
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
vif(model0)
```

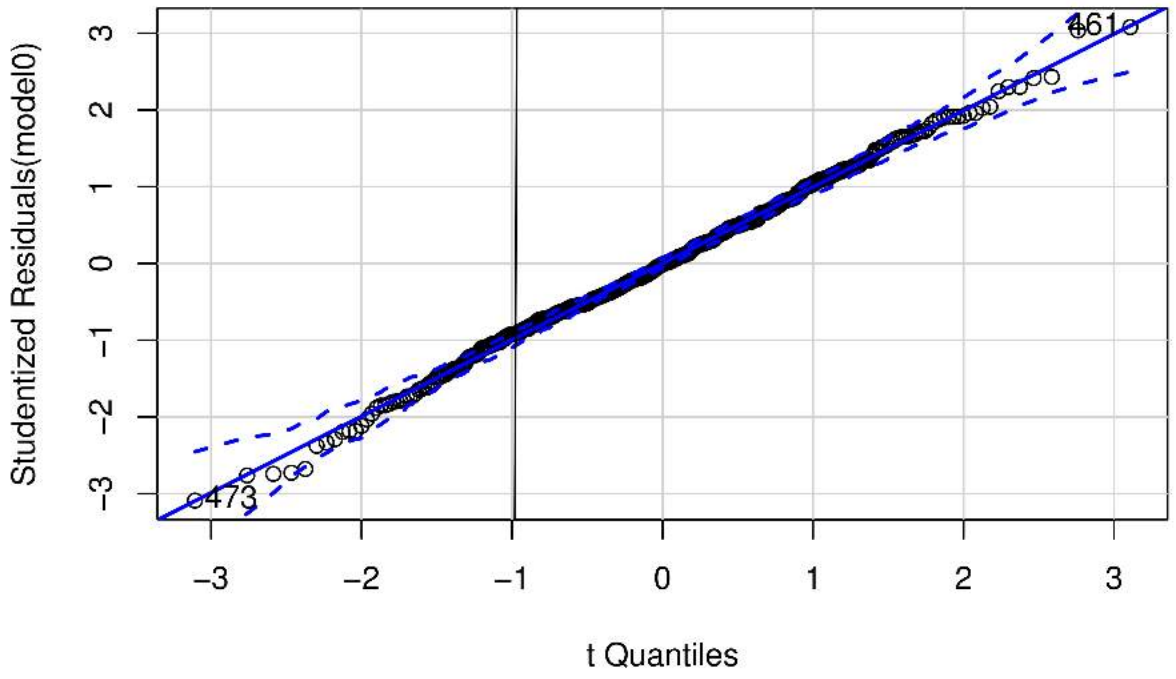
```
##  Avg..Session.Length          Time.on.App      Time.on.Website
##             1.005422             1.008684             1.010275
## Length.of.Membership
##             1.006949
```

** VIF 1-5 Moderately Correlated **

Multivariate Normality (Tested using qq Plot and Shapiro Test)

** qqPlot **

```
abline(qqPlot(model0))
```



Shapiro Test Ultimate Test for Normality

```
shapiro.test(resid(model0))
```

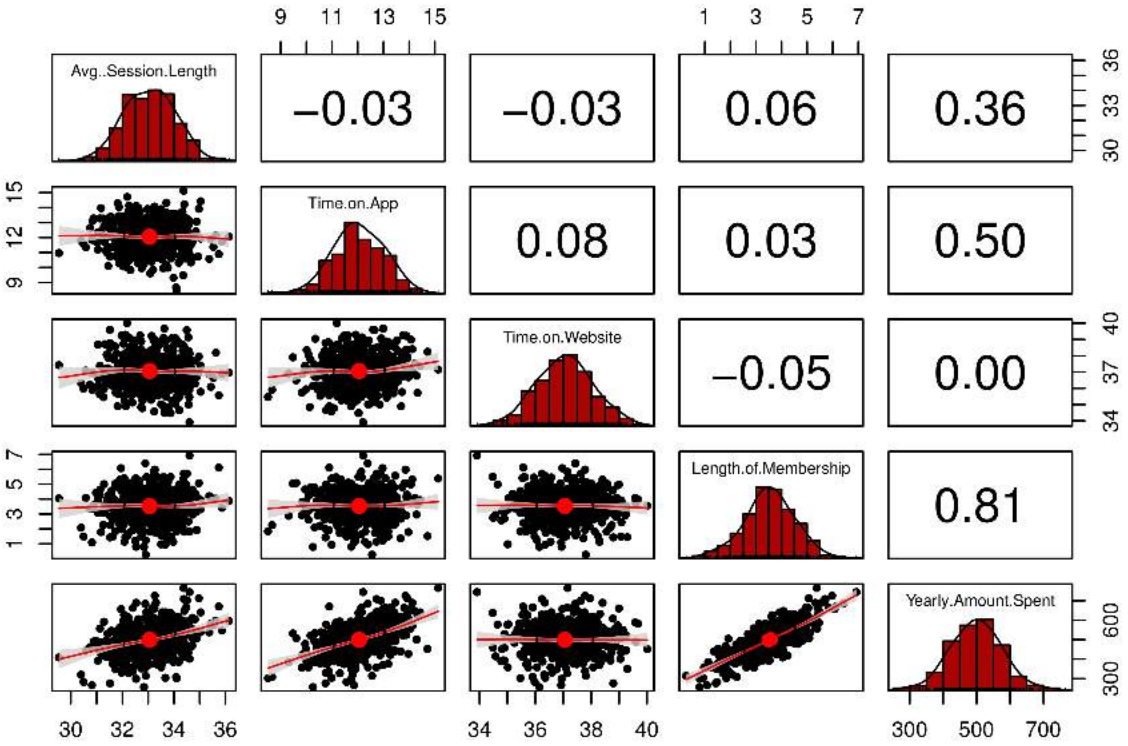
```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model0)
## W = 0.99782, p-value = 0.7716
```

Hypothesis for Shapiro Test

- ** H0= Data is Normally Distributed **
- ** H1=Data is Not Normally Distributed As p is not low than 0.05 hence we fail to reject the Null Hypothesis **

Create Bivariate Scatter Plots

```
pairs.panels(Ecommerce.Customers[c(-1,-2,-3)],method = "pearson",hist.col="#aa0505",density = TRUE,ellip
```



Observations * **Length of Membership vs Yearly Amount Spent** ** Iligh Correlation * Time on App Vs Yearly Amount Spent** ** Moderately correlated * Avg Session Length vs Yearly Amount Spent** ** Low Correlation **

```
library(corrplot)
corrplot(cor(Ecommerce.Customers[c(-1,-2,-3)]))
```

Checking for AutoCorrelation ** GAUSS-MARKOV DOUBT**

** DURBIN-WATSON TEST Hypothesis Testing for Auto Correlation * H0:- There is no auto correlation * H1:- There is auto correlation among the dependent variables **

```
durbinWatsonTest(model0)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.05328995       1.88665   0.222
##  Alternative hypothesis: rho != 0
```

Observation

- ** As p is not less 0.05 hence we fail to reject the null hypothesis **
- ** Which means there is no auto correlation **
- ** D-W Statistic is 1.88 which is close to 2 {Range 0-4 and close to 2 means no autocorrelation} **

TEST FOR HOMOSCEDASTICITY GOLD FIELD QUANT TEST

```
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 3.5.3
## Loading required package: zoo
## Warning: package 'zoo' was built under R version 3.5.3
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
```

```
##
##     as.Date, as.Date.numeric
```

```
gqtest(model0)
```

```
##
##  Goldfeld-Quandt test
##
## data:  model0
## GQ = 1.0753, df1 = 245, df2 = 245, p-value = 0.2851
## alternative hypothesis: variance increases from segment 1 to 2
```

** Hypothesis Testing for GOLDFIELD QUANT Test * H0:- The residuals are equal across the regression line (Homoscedasticity) * H1:- The residuals are not equal the regression line (HetroScedasticity) **

Observation

- **As p is not less than 0.05 then we fail to reject the null hypothesis**
- **Test for Homoscedasticity which means residuals have equal variance across the regression line**

Based off this plot what looks to be the most correlated feature with Yearly Amount Spent?

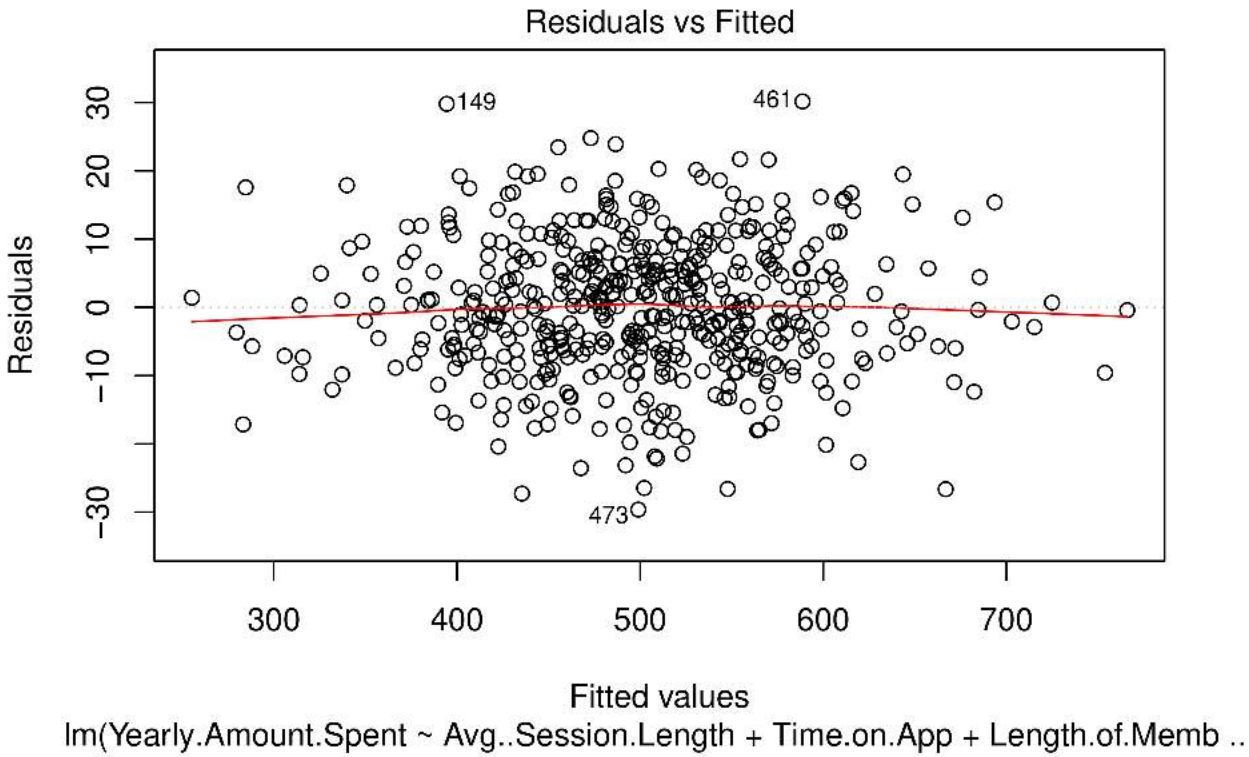
Length of Membership vs Yearly Amount Spent

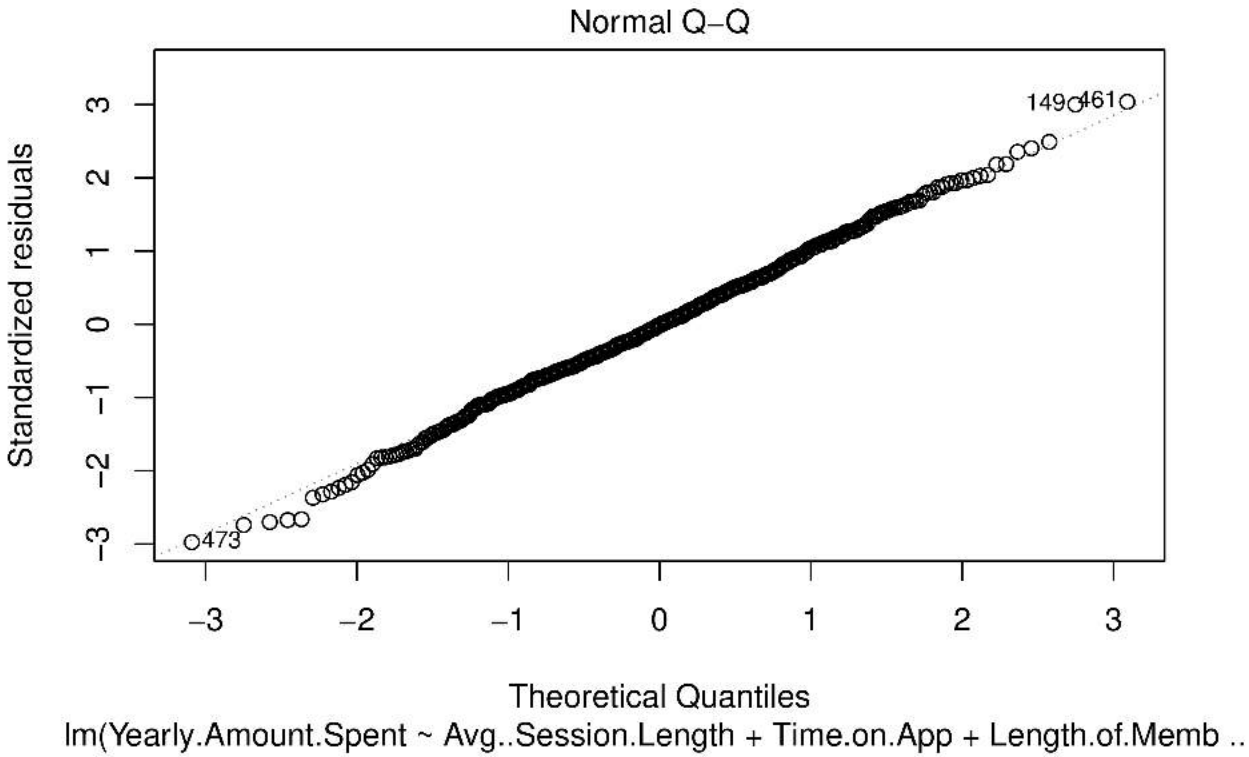
Using Step AIC

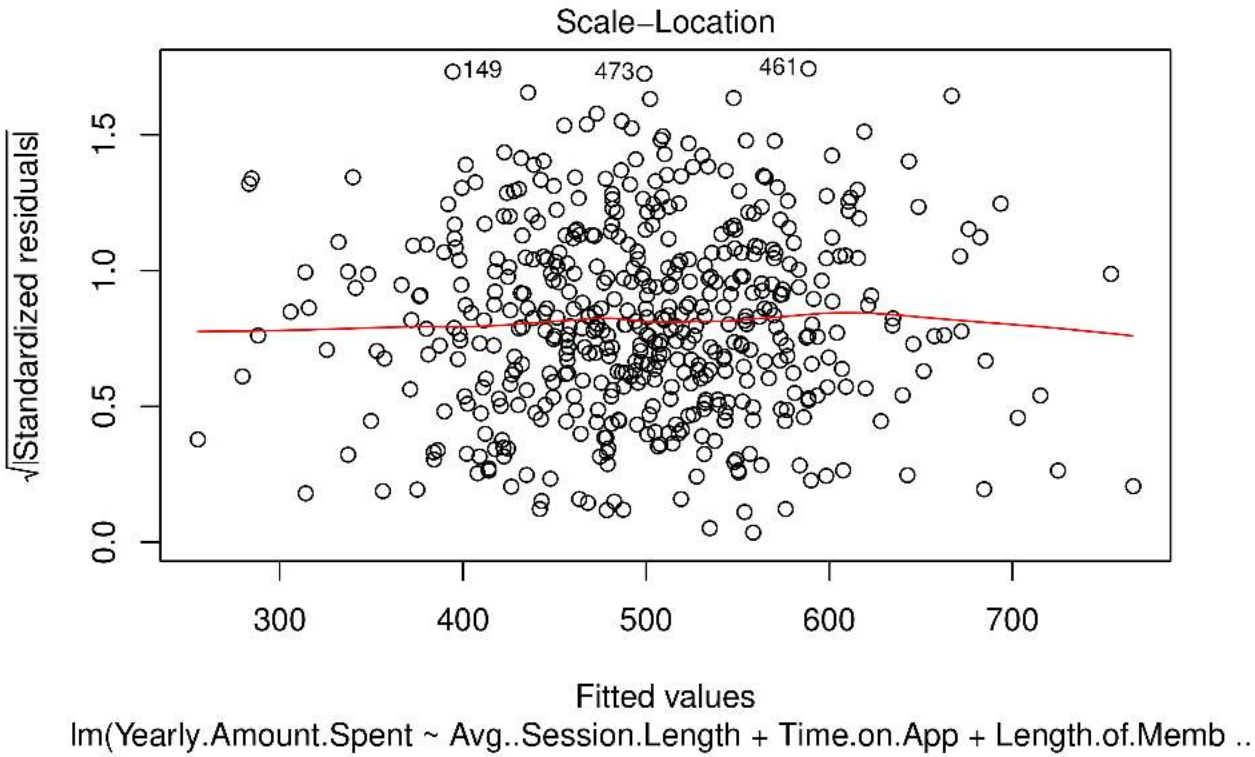
```
library(MASS)
attach(Ecommerce.Customers)
model01=stepAIC(model0)
```

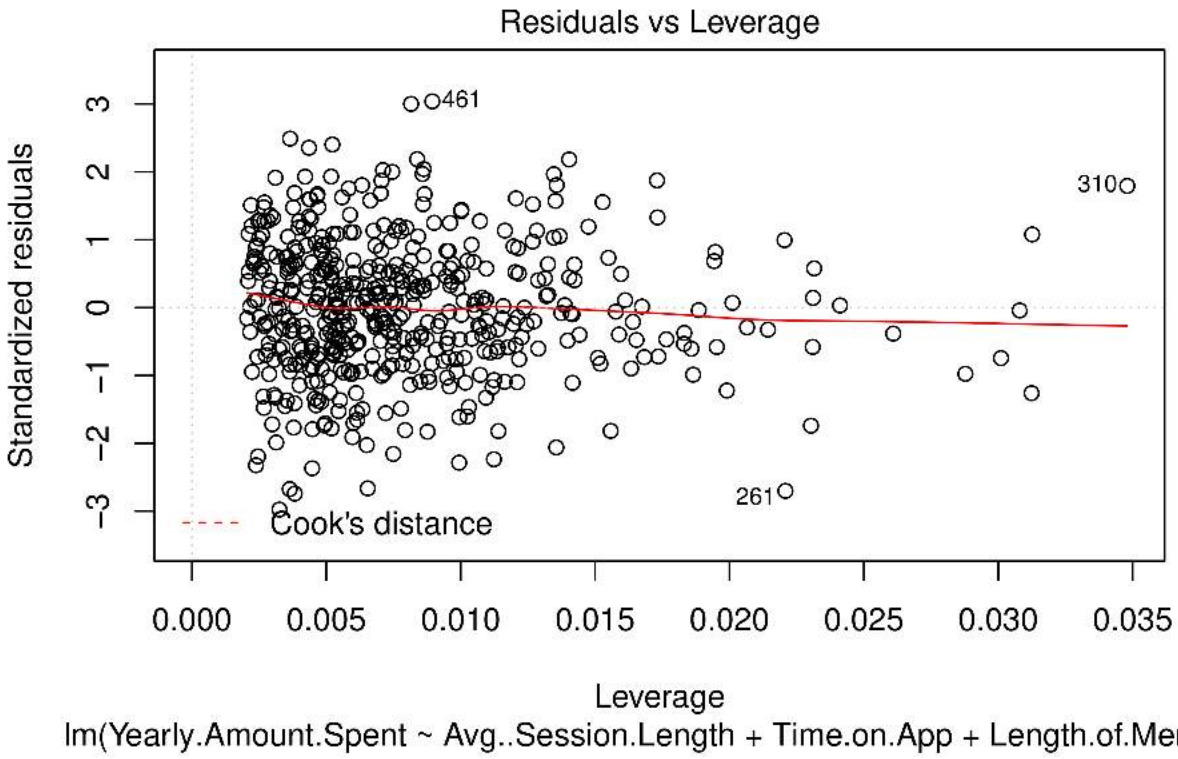
```
## Start:  AIC=2304.88
## Yearly.Amount.Spent ~ Avg..Session.Length + Time.on.App + Time.on.Website +
##     Length.of.Membership
##
##                        Df Sum of Sq     RSS    AIC
## - Time.on.Website       1        96   49332 2303.9
## <none>                               49236 2304.9
## - Avg..Session.Length   1    323812  373047 3315.4
## - Time.on.App           1    732713  781948 3685.5
## - Length.of.Membership  1   1876320 1925555 4136.1
##
## Step:  AIC=2303.86
## Yearly.Amount.Spent ~ Avg..Session.Length + Time.on.App + Length.of.Membership
##
##                        Df Sum of Sq     RSS    AIC
## <none>                               49332 2303.9
## - Avg..Session.Length   1    323767  373099 3313.5
## - Time.on.App           1    739204  788535 3687.7
## - Length.of.Membership  1   1879405 1928736 4134.9
```

```
plot(model01)
```


Residuals vs Fitted
149
461
473
Residuals
30
20
10
0
−10
−30
300
400
500
600
700
Fitted values
lm(Yearly.Amount.Spent ~ Avg..Session.Length + Time.on.App + Length.of.Memb ..


Normal Q–Q
Standardized residuals
Theoretical Quantiles
lm(Yearly.Amount.Spent ~ Avg..Session.Length + Time.on.App + Length.of.Memb ..
149
461
473
−3
−2
−1
0
1
2
3


Scale−Location
149
473
461
√|Standardized residuals|
0.0
0.5
1.0
1.5
300
400
500
600
700
Fitted values
lm(Yearly.Amount.Spent ~ Avg..Session.Length + Time.on.App + Length.of.Memb ..




```
summary(model01)
```

```
##
## Call:
## lm(formula = Yearly.Amount.Spent ~ Avg..Session.Length + Time.on.App +
##     Length.of.Membership, data = Ecommerce.Customers[c(-1, -2,
##     -3)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.628  -6.378  -0.135   6.351  30.169
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1035.3396    15.9829  -64.78   <2e-16 ***
## Avg..Session.Length     25.7210     0.4508   57.05   <2e-16 ***
## Time.on.App             38.7460     0.4494   86.21   <2e-16 ***
## Length.of.Membership    61.5560     0.4478  137.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.973 on 496 degrees of freedom
## Multiple R-squared:  0.9843, Adjusted R-squared:  0.9842
## F-statistic: 1.036e+04 on 3 and 496 DF,  p-value: < 2.2e-16
```

** Final Model contains :-Avg..Session.Length + Time.on.App + Length.of.Membership Adjusted R2 is high for Model01 It means Average session Length Time on App and Length of the membership is able to

cxplain** #### Anova Plot

```
anovafull=aov(model01)
summary(anovafull)
```

```
##                       Df  Sum Sq Mean Sq F value Pr(>F)
## Avg..Session.Length    1  395805  395805    3980 <2e-16 ***
## Time.on.App            1  814585  814585    8190 <2e-16 ***
## Length.of.Membership   1 1879405 1879405   18896 <2e-16 ***
## Residuals            496   49332      99
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anovafull
```

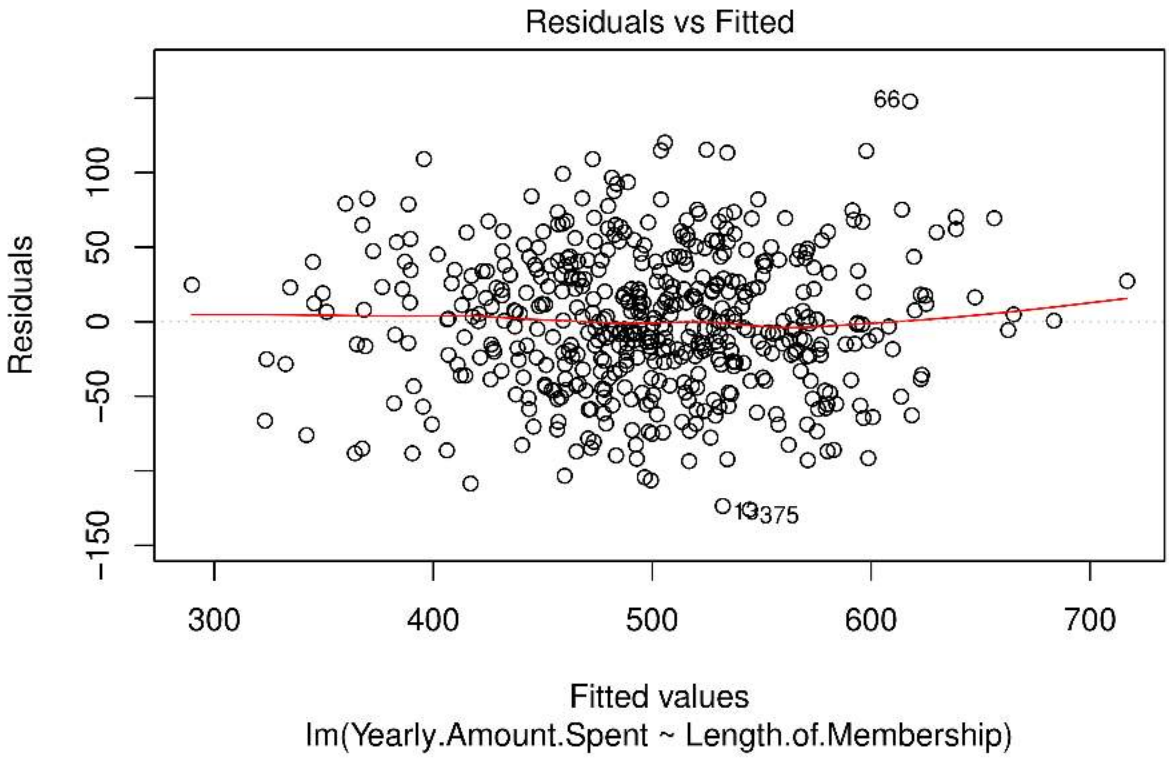
```
## Call:
##    aov(formula = model01)
##
## Terms:
##                 Avg..Session.Length Time.on.App Length.of.Membership
## Sum of Squares             395805.2    814585.0            1879404.5
## Deg. of Freedom                   1           1                    1
##                 Residuals
## Sum of Squares    49331.7
## Deg. of Freedom       496
##
## Residual standard error: 9.972918
## Estimated effects may be unbalanced
```

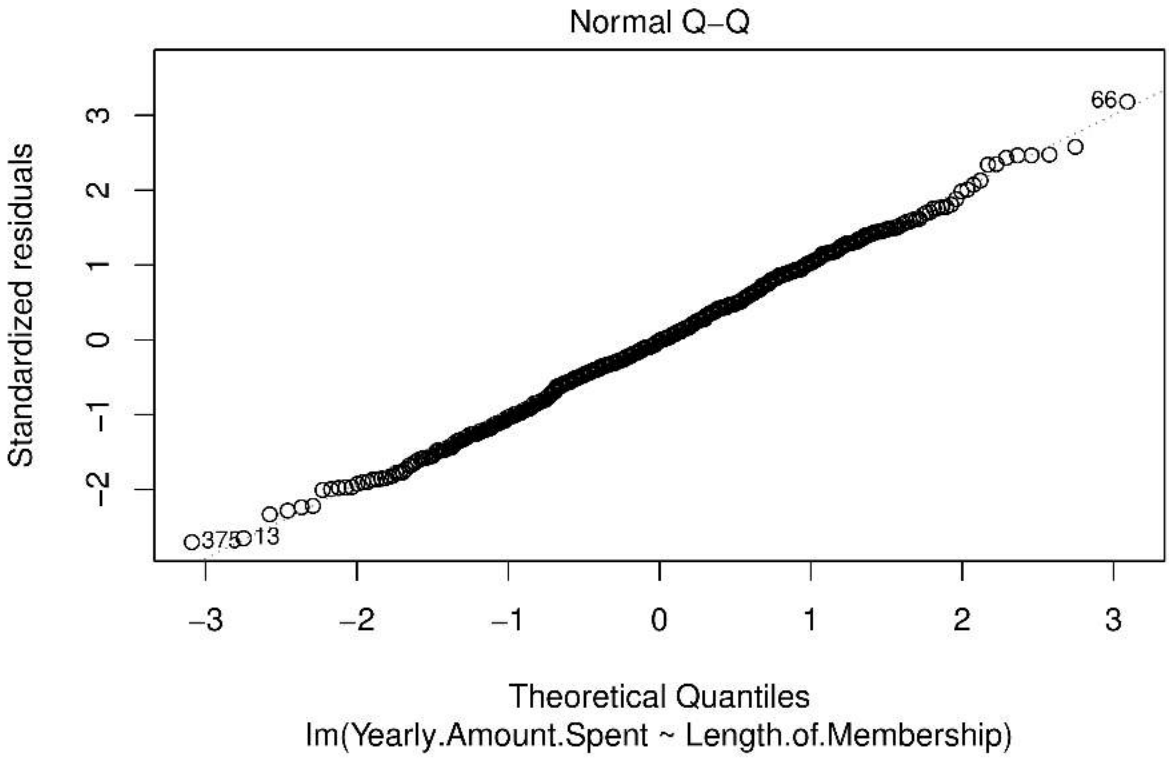
Create a linear model plot of Yearly Amount Spent vs. Length of Membership.

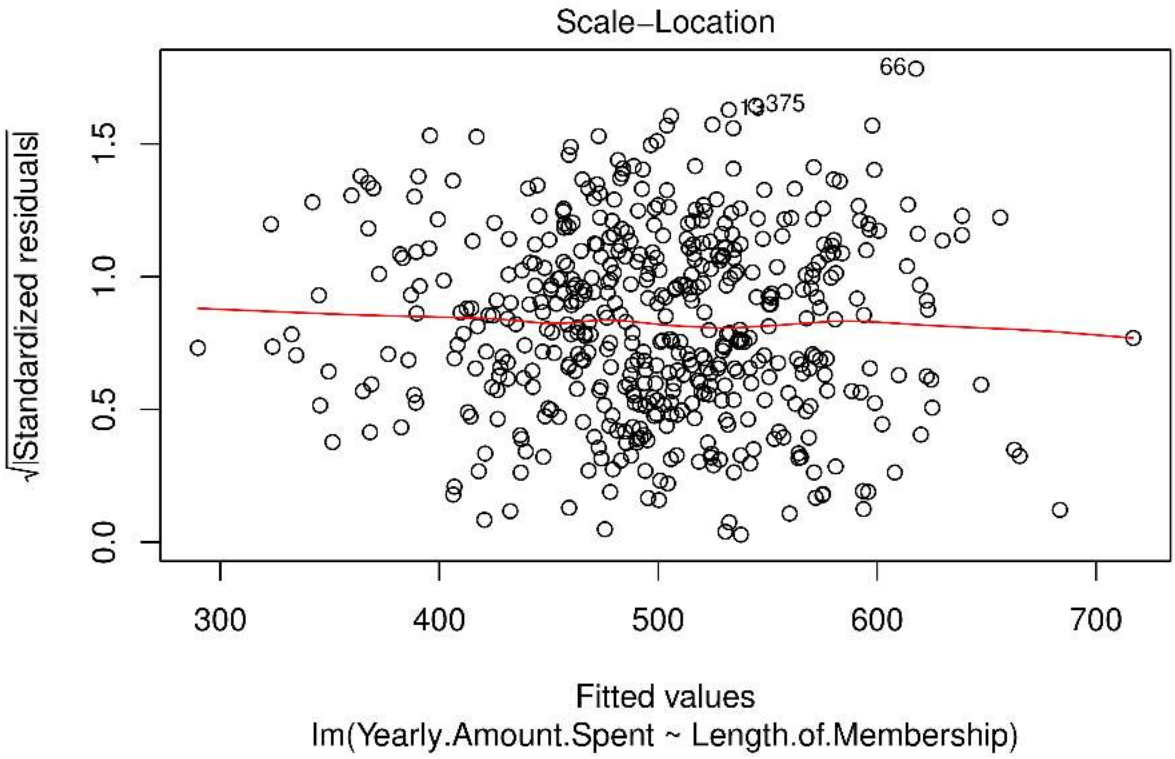
```
attach(Ecommerce.Customers)
```

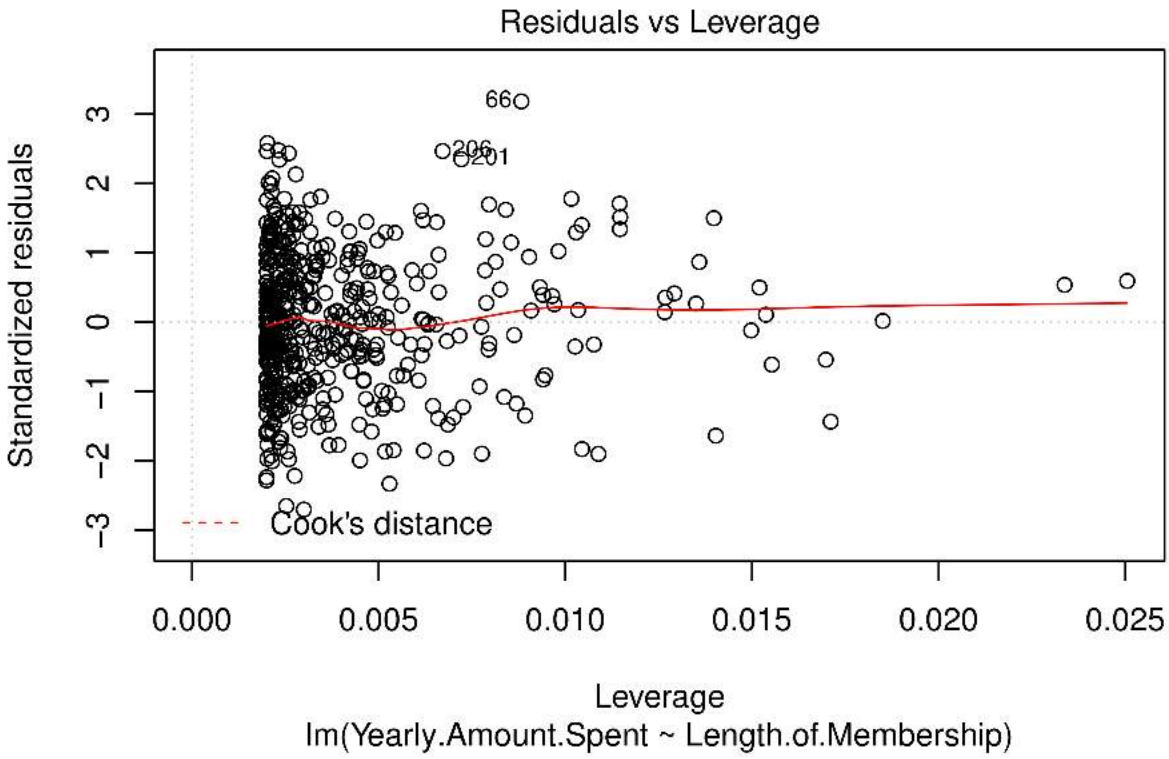
```
## The following objects are masked from Ecommerce.Customers (pos = 3):
##
##     Address, Avatar, Avg..Session.Length, Email,
##     Length.of.Membership, Time.on.App, Time.on.Website,
##     Yearly.Amount.Spent
```

```
model1=lm(Yearly.Amount.Spent~Length.of.Membership,data = Ecommerce.Customers)
abline(plot(model1))
```


Residuals vs Fitted
Residuals
Fitted values
lm(Yearly.Amount.Spent ~ Length.of.Membership)
66
375
−150
−50
0
50
100
300
400
500
600
700


Normal Q–Q
Standardized residuals
Theoretical Quantiles
lm(Yearly.Amount.Spent ~ Length.of.Membership)
66
375
13


Scale−Location
66
13375
√|Standardized residuals|
0.0
0.5
1.0
1.5
300
400
500
600
700
Fitted values
lm(Yearly.Amount.Spent ~ Length.of.Membership)




```
summary(model1)
```

```
##
## Call:
## lm(formula = Yearly.Amount.Spent ~ Length.of.Membership, data = Ecommerce.Customers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -125.975  -29.032   -0.494   33.033  147.777
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           272.400      7.675   35.49   <2e-16 ***
## Length.of.Membership   64.219      2.090   30.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46.66 on 498 degrees of freedom
## Multiple R-squared:  0.6546, Adjusted R-squared:  0.6539
## F-statistic: 943.9 on 1 and 498 DF,  p-value: < 2.2e-16
```

Interpretation of Plots:-

** Adjusted R2 is low 0.6539 which means Length of Membership is able to explain 65.39% of the variance in Yearly Amount Spent Residual vs Fitted Plot :- Test the assumption of (1) Linearlity and (2) Equal Variance Homoscedasticity #### **A "good" residuals vs. fitted plot should be relatively shapeless without clear patterns in the data, no obvious outliers, and be generally symmetrically**

distributed around the 0 line without particularly large residuals. Q-Q Plot (For Dependent Variable MULTIVARIATE NORMALITY) :- The normal qq plot helps us determine if our dependent variable is normally distributed by plotting quantiles (i.e. percentiles) from our distribution against a theoretical distribution. If our data is normally distributed, it will be plotted in a generally straight line on the qq plot. Scale-Location Plot(For Residuals) :-The Scale-Location plot shows whether our residuals are spread equally along the predictor range, i.e. homoscedastic. We want the line on this plot to be horizontal with randomly spread points on the plot. Residual Vs Leverage Plot :-The Residuals vs. Leverage plots helps you identify influential data points on your model. Outliers can be influential, though they don't necessarily have to it and some points within a normal range in your model could be very influential. The points we're looking for(or not looking for) are values in the upper right or lower right corners, which are outside the red dashed Cook's distance line. These are points that would be influential in the model and removing them would likely noticeably alter the regression results. **

Analysis:-

** All Plots Okay **

Splitting the data to Training and Testing Data

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.5.3
```

```
## Loading required package: lattice
```

```
set.seed(421)
data1=createDataPartition(Ecommerce.Customers$Yearly.Amount.Spent,p=0.7,list = FALSE)
traindata=Ecommerce.Customers[data1,]
testdata=Ecommerce.Customers[-data1,]
```

Model using Average.session.Length, TimeonApp and Lengthofthemembership

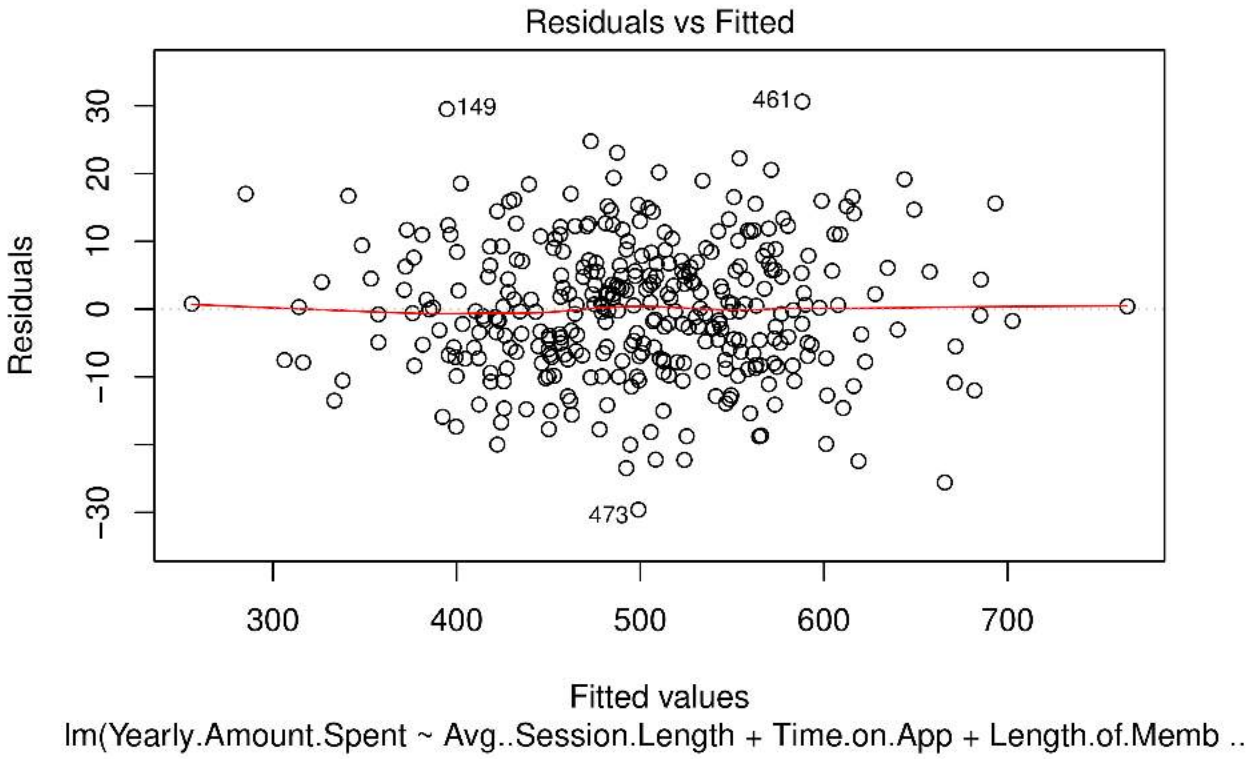
Hypothesis of Linear Regression :- * ** H0:- There is no relatioship between Yearly Amount Spent Vs Average.session.Length, TimeonApp and Lengthofthemembership * H1:- There is a Linear relationship beween Yearly Amount Spent Vs Average.session.Length, TimeonApp and Lengthofthemembership **

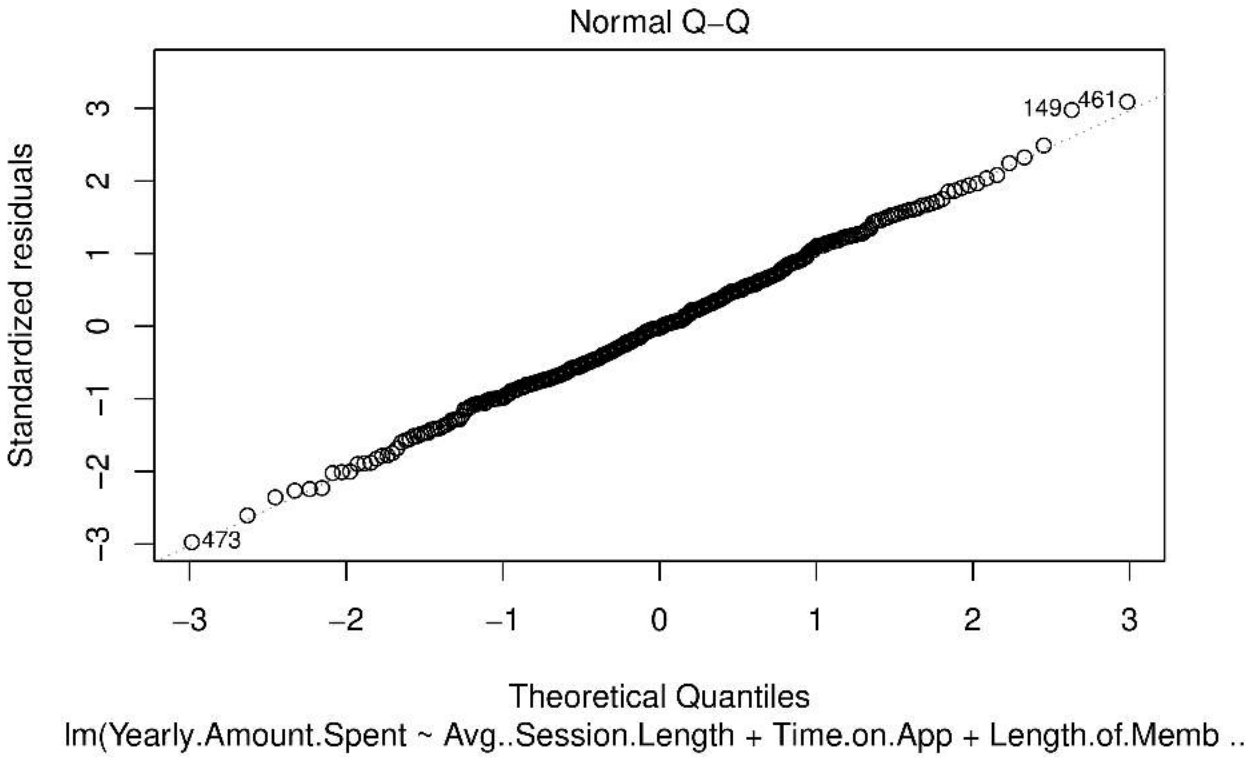
```
model3=lm(Yearly.Amount.Spent~Avg..Session.Length+Time.on.App+Length.of.Membership,data = traindata)
summary(model3)
```

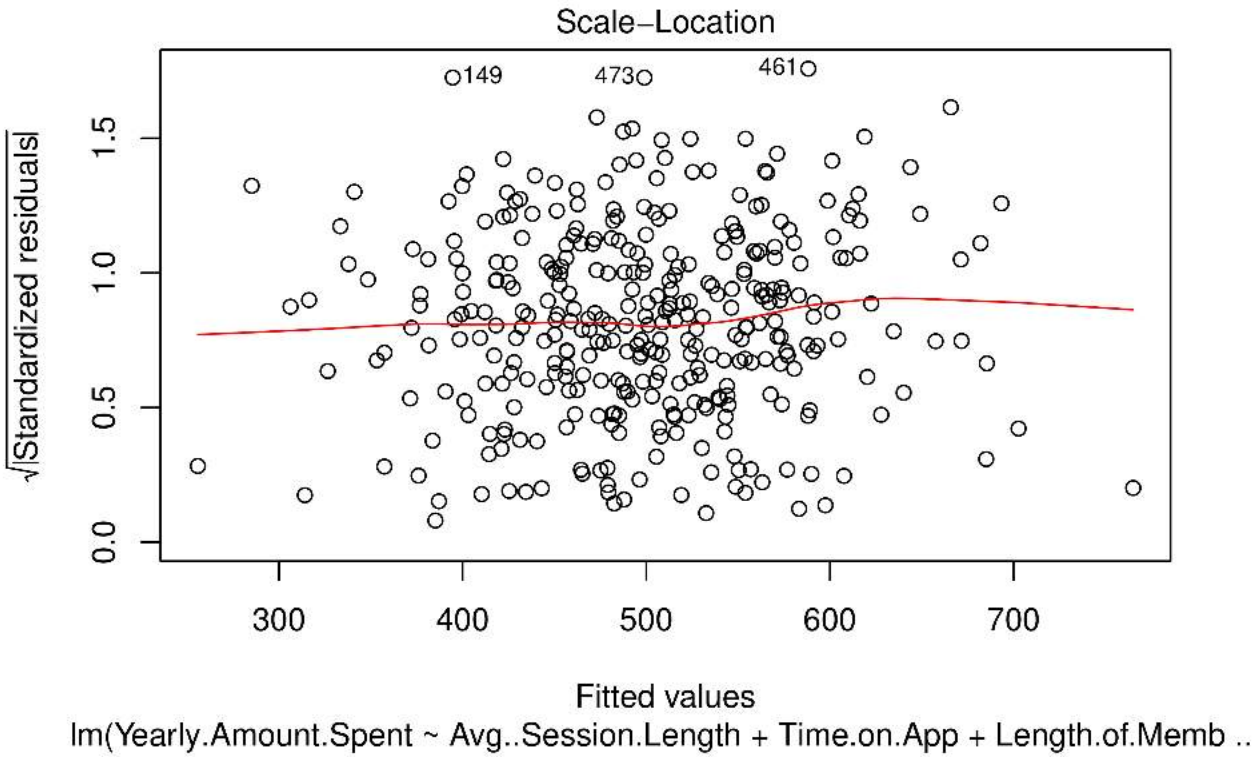
```
##
## Call:
## lm(formula = Yearly.Amount.Spent ~ Avg..Session.Length + Time.on.App +
##     Length.of.Membership, data = traindata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.6126  -6.9429  -0.2278   6.4231  30.6177
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1017.9761    19.4886  -52.23   <2e-16 ***
## Avg..Session.Length     25.2820     0.5495   46.01   <2e-16 ***
## Time.on.App             38.5136     0.5533   69.61   <2e-16 ***
## Length.of.Membership    61.6365     0.5396  114.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

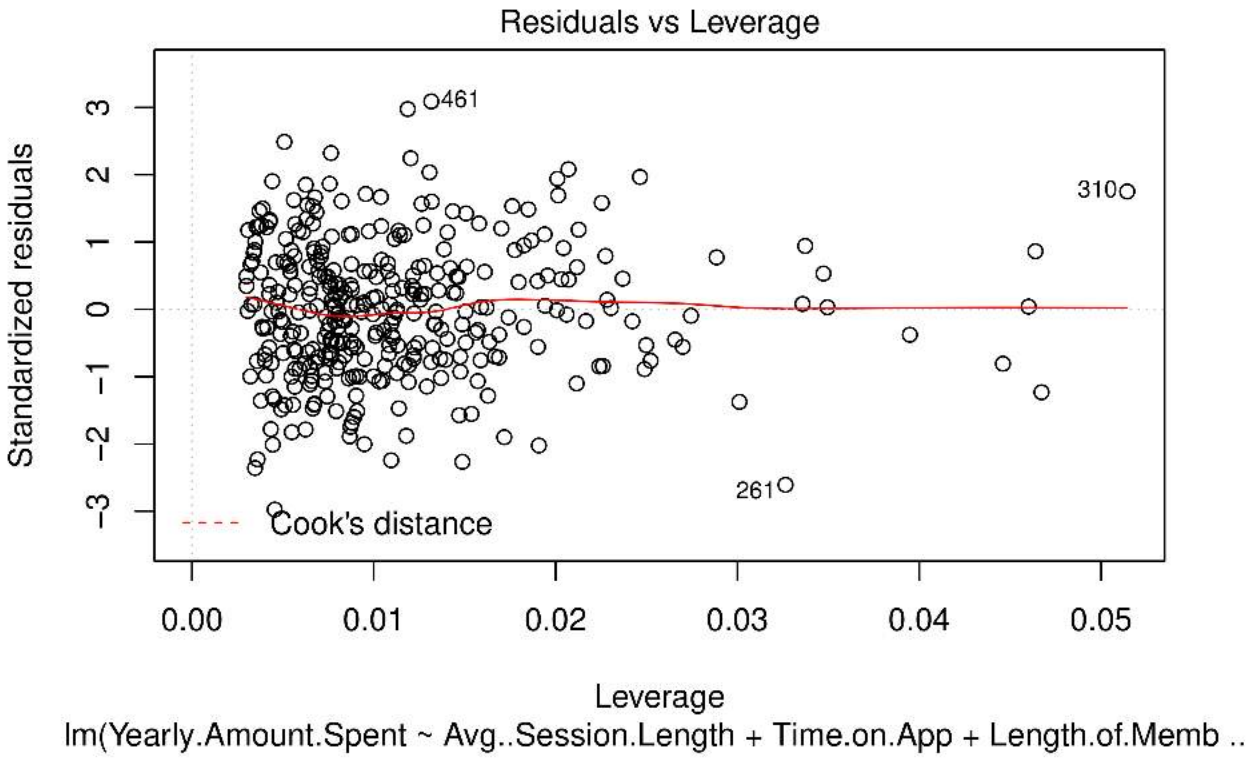
```
## Residual standard error: 9.977 on 348 degrees of freedom
## Multiple R-squared:  0.9836, Adjusted R-squared:  0.9835
## F-statistic:  6957 on 3 and 348 DF,  p-value: < 2.2e-16
```

```
abline(plot(model3))
```




Normal Q–Q
Standardized residuals
Theoretical Quantiles
lm(Yearly.Amount.Spent ~ Avg..Session.Length + Time.on.App + Length.of.Memb ..
473
149
461
−3
−2
−1
0
1
2
3


Scale−Location
149
473
461
√|Standardized residuals|
0.0
0.5
1.0
1.5
300
400
500
600
700
Fitted values
lm(Yearly.Amount.Spent ~ Avg..Session.Length + Time.on.App + Length.of.Memb ..



Creating Alternate Model

```
model001=stepAIC(model0,direction = "backward")
```

```
## Start:  AIC=2304.88
## Yearly.Amount.Spent ~ Avg..Session.Length + Time.on.App + Time.on.Website +
##     Length.of.Membership
##
##                        Df Sum of Sq     RSS    AIC
## - Time.on.Website       1        96   49332 2303.9
## <none>                                49236 2304.9
## - Avg..Session.Length   1    323812  373047 3315.4
## - Time.on.App           1    732713  781948 3685.5
## - Length.of.Membership  1   1876320 1925555 4136.1
##
## Step:  AIC=2303.86
## Yearly.Amount.Spent ~ Avg..Session.Length + Time.on.App + Length.of.Membership
##
##                        Df Sum of Sq     RSS    AIC
## <none>                                49332 2303.9
## - Avg..Session.Length   1    323767  373099 3313.5
## - Time.on.App           1    739204  788535 3687.7
## - Length.of.Membership  1   1879405 1928736 4134.9
```

```
model002=stepAIC(model0,direction = "both")
```

```
## Start:  AIC=2304.88
```

```
## Yearly.Amount.Spent ~ Avg..Session.Length + Time.on.App + Time.on.Website +
##     Length.of.Membership
##
##                        Df Sum of Sq     RSS    AIC
## - Time.on.Website       1        96   49332 2303.9
## <none>                                49236 2304.9
## - Avg..Session.Length   1    323812  373047 3315.4
## - Time.on.App           1    732713  781948 3685.5
## - Length.of.Membership  1   1876320 1925555 4136.1
##
## Step:  AIC=2303.86
## Yearly.Amount.Spent ~ Avg..Session.Length + Time.on.App + Length.of.Membership
##
##                        Df Sum of Sq     RSS    AIC
## <none>                                49332 2303.9
## + Time.on.Website       1        96   49236 2304.9
## - Avg..Session.Length   1    323767  373099 3313.5
## - Time.on.App           1    739204  788535 3687.7
## - Length.of.Membership  1   1879405 1928736 4134.9
```

```
library(car)
```

Model using only Length of Membership as Independent Variable

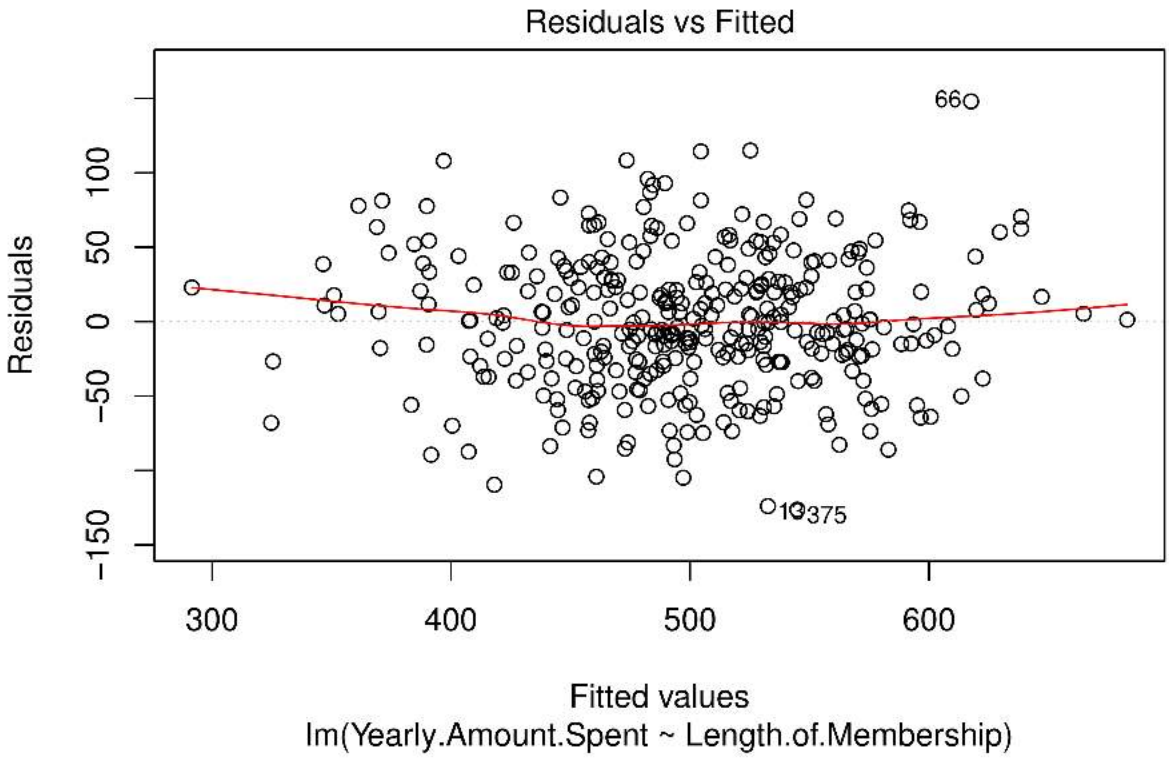
Create Linear Model on Training Data

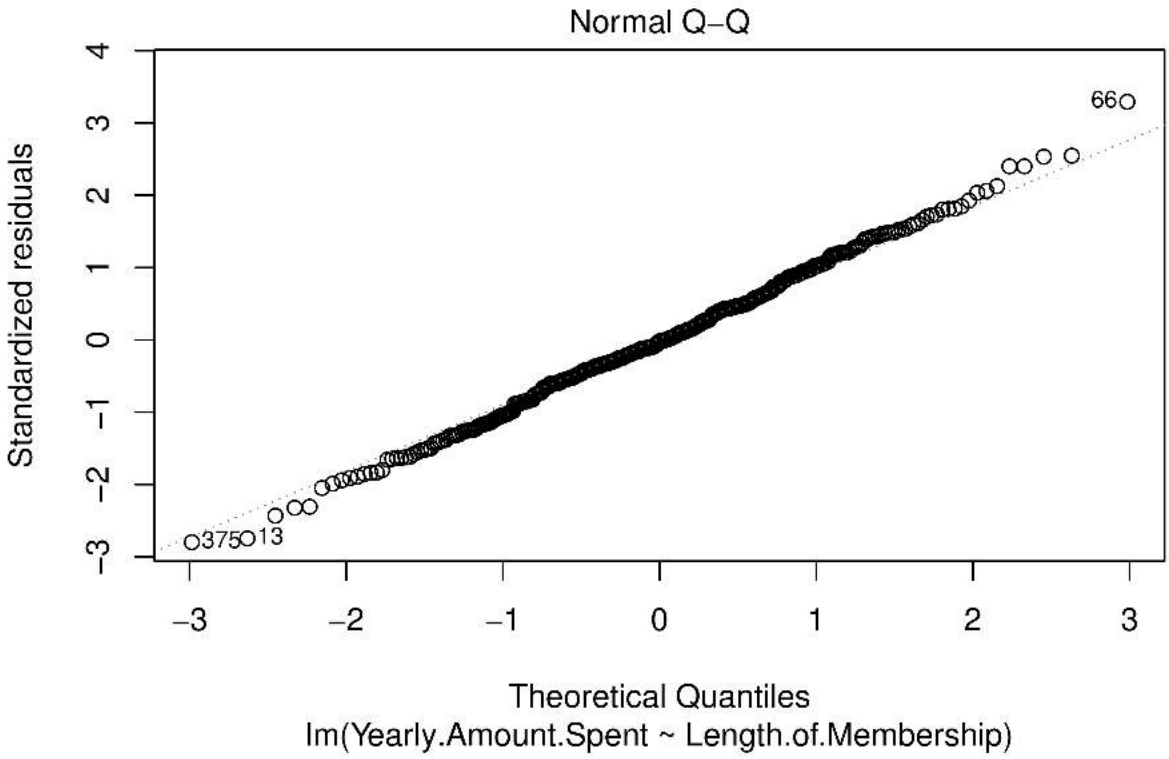
Hypothesis of Linear Regression :- * ** H0:- There is no relatioship between Yearly Amount Spent Vs Length of Membership * H1:- There is a Linear relationship beween Yearly Amount Spent Vs Length of Membership **

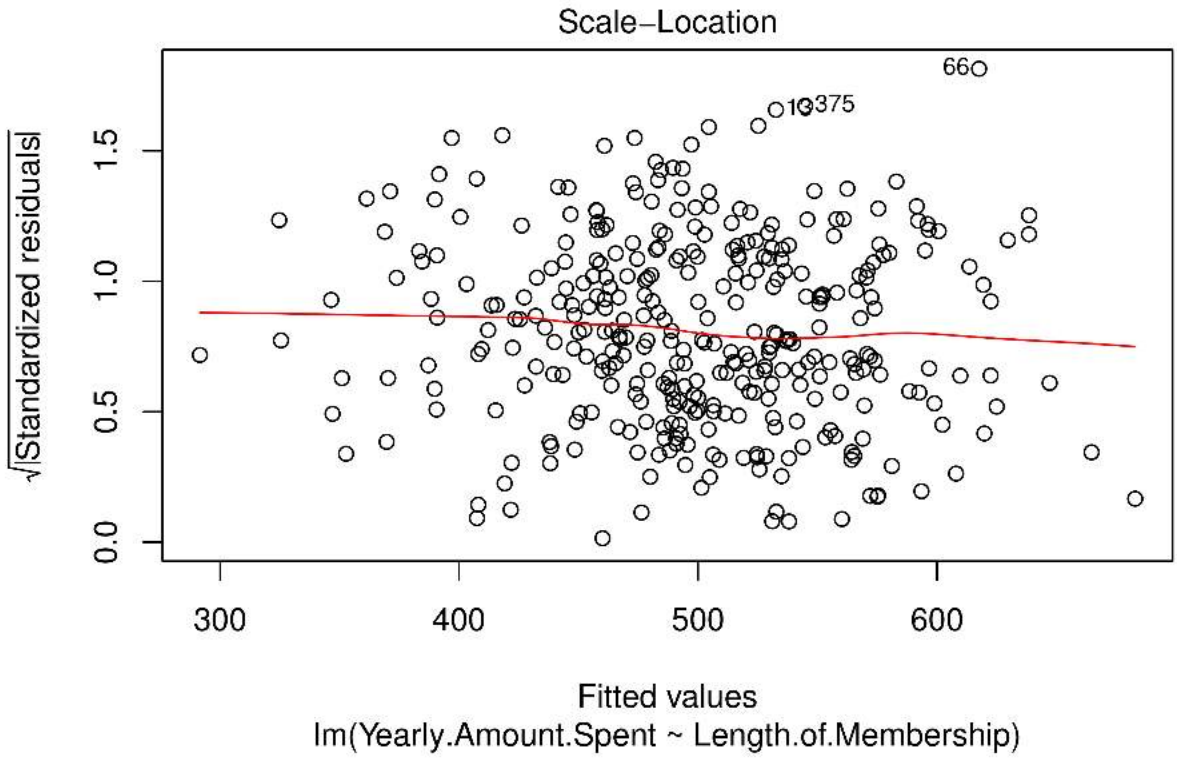
```
model2=lm(Yearly.Amount.Spent~Length.of.Membership,data=traindata)
summary(model2)
```

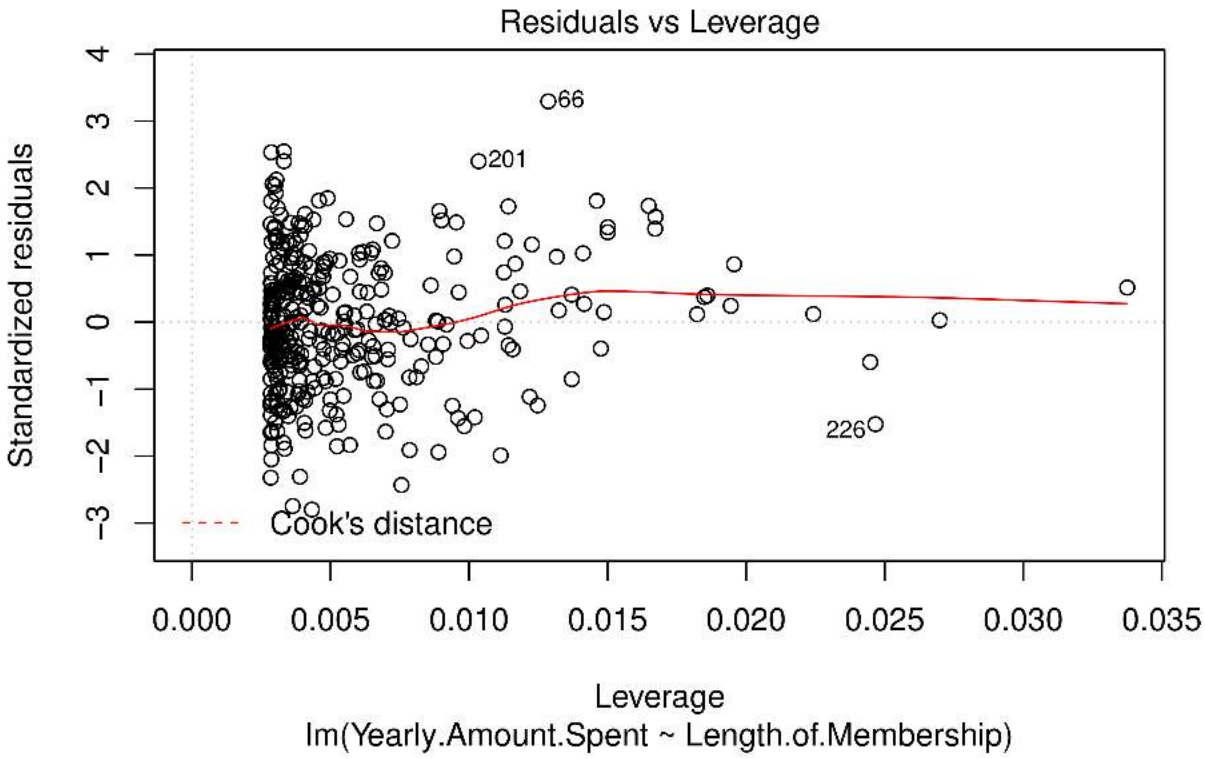
```
##
## Call:
## lm(formula = Yearly.Amount.Spent ~ Length.of.Membership, data = traindata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -126.256  -27.002   -1.058   28.731  147.934
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         274.314      8.940   30.68   <2e-16 ***
## Length.of.Membership   63.833      2.443   26.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45.22 on 350 degrees of freedom
## Multiple R-squared:  0.6611, Adjusted R-squared:  0.6602
## F-statistic: 682.9 on 1 and 350 DF,  p-value: < 2.2e-16
```

```
abline(plot(model2))
```


Residuals vs Fitted
Residuals
−150
−50
0
50
100
300
400
500
600
Fitted values
lm(Yearly.Amount.Spent ~ Length.of.Membership)
66
375


Normal Q–Q
Standardized residuals
Theoretical Quantiles
lm(Yearly.Amount.Spent ~ Length.of.Membership)
66
375
13


Scale−Location
66
375
√|Standardized residuals|
0.0
0.5
1.0
1.5
300
400
500
600
Fitted values
lm(Yearly.Amount.Spent ~ Length.of.Membership)



Predit the Test Data (only Length.of.Membership)

```
predict2=predict(model2,testdata)
output2=cbind(predict2,testdata$Yearly.Amount.Spent)
summary(predict2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   334.1   465.5   502.7   501.2   535.0   716.2
```

```
output2
```

```
##     predict2
## 10  478.7599 427.1994
## 11  432.7872 492.6060
## 15  433.5491 470.4527
## 18  428.8422 407.7045
## 20  533.7679 605.0610
## 23  417.6293 436.5156
## 30  577.0488 554.7221
## 33  551.5464 588.7126
## 35  537.6002 507.4418
## 36  507.2602 521.8836
## 37  392.2378 347.7769
## 39  571.1706 478.1703
## 42  570.8915 501.8744
## 45  418.4935 448.2298
## 46  508.1424 549.8606
## 49  478.7423 479.7319
```

```
## 50  478.4786 416.3584
## 51  655.8314 725.5848
## 54  520.4218 451.4574
## 59  495.9383 496.6507
## 62  598.6648 507.2126
## 63  580.8822 613.5993
## 65  495.3242 540.2634
## 69  451.7401 408.6202
## 70  501.0916 451.5757
## 71  440.4267 444.9666
## 74  520.1733 534.7772
## 76  473.5104 478.7194
## 84  396.4967 338.3199
## 91  432.3742 449.0703
## 92  537.6640 611.0000
## 95  541.9881 514.0890
## 97  534.8692 521.1430
## 99  524.0116 507.3901
## 104 463.8761 492.1051
## 108 466.1168 378.3309
## 109 520.3581 570.4517
## 113 500.5105 424.7626
## 115 624.5025 642.1016
## 118 593.7886 593.0772
## 127 536.4174 516.8316
## 128 479.7467 468.4457
## 137 523.0801 529.2301
## 138 433.3657 433.0488
## 141 527.0234 448.9333
## 142 481.8836 472.9922
## 144 366.4583 350.0582
## 151 443.6102 426.7752
## 153 618.4573 555.8926
## 154 662.2074 657.0199
## 155 521.0211 595.8038
## 156 462.0741 503.9784
## 158 716.2127 744.2219
## 160 515.9190 528.2238
## 162 336.1443 357.5914
## 164 503.1669 490.2066
## 165 562.8634 550.0476
## 172 425.0301 439.8913
## 175 508.5330 465.1766
## 176 383.8378 373.8857
## 177 579.0228 532.7175
## 180 494.0795 501.1002
## 186 505.4141 485.9231
## 191 570.0196 612.3852
## 200 488.8971 467.5019
## 204 471.4802 392.4974
## 206 597.7384 712.3963
## 208 577.3304 562.0820
## 209 471.6757 412.0129
## 213 519.6798 536.1309
```

```
## 214 460.0058 558.4273
## 218 584.0873 528.9336
## 221 509.1308 519.3730
## 223 522.8104 502.4098
## 224 554.5078 604.3348
## 235 442.4140 493.1802
## 243 429.8732 451.6286
## 244 507.7516 490.6004
## 246 428.3331 409.0705
## 248 534.6121 647.6195
## 252 484.1791 393.8574
## 262 494.9370 514.0098
## 268 377.9423 399.9839
## 272 593.9451 628.0478
## 277 517.3066 423.3083
## 281 471.3253 511.9799
## 285 503.7500 463.5914
## 286 468.9544 471.6029
## 287 506.3062 626.0187
## 299 623.0039 587.5748
## 300 368.8905 282.4712
## 301 515.0565 473.9499
## 302 459.3222 489.9081
## 303 502.1383 541.9722
## 304 343.5466 266.0863
## 316 530.1806 584.1059
## 320 534.1105 596.5167
## 322 480.4607 542.4125
## 329 466.0028 422.3687
## 331 534.6462 442.0644
## 332 509.8538 533.0401
## 333 414.1175 424.2028
## 336 488.1076 443.4419
## 337 499.8285 478.6009
## 341 463.5479 501.1225
## 343 521.5044 486.0834
## 352 581.2453 533.3966
## 357 580.3512 640.5841
## 361 410.8792 444.5761
## 366 595.9134 594.2745
## 371 579.3720 521.2408
## 378 485.6120 538.9420
## 382 487.8955 547.1907
## 383 479.5808 410.6029
## 384 516.2777 583.9778
## 387 468.7849 550.8134
## 395 515.1553 557.6083
## 398 496.9667 547.7100
## 400 454.9432 408.2169
## 402 494.6509 506.3759
## 413 466.6460 444.0538
## 414 580.1467 493.1813
## 416 365.5205 275.9184
## 419 528.5280 475.7251
```

```
## 420 510.5759 483.5432
## 426 571.3214 574.4157
## 429 512.9375 556.2981
## 431 567.6617 556.1864
## 432 416.2789 475.0716
## 433 548.0626 486.9471
## 435 334.0937 304.1356
## 439 499.9706 392.9923
## 441 500.8609 499.1402
## 451 518.5424 475.0154
## 452 427.6707 436.7206
## 454 505.9870 478.1831
## 458 495.9448 534.7715
## 459 542.3088 537.9158
## 465 613.9136 689.2357
## 466 474.1468 543.1326
## 471 467.2616 424.7288
## 478 535.3498 487.5555
## 485 451.9468 462.6565
## 491 476.2248 510.4014
## 495 451.0591 510.6618
## 496 513.4704 573.8474
## 498 590.8168 551.6201
## 500 448.9084 497.7786
```

Predit the Test Data (using Average.session.Length, TimeonApp and Lengthofthemembership)

```
predict3=predict(model3,testdata )
output3=cbind(predict3,testdata$Yearly.Amount.Spent)
output3
```

```
##     predict3
## 10  441.8571 427.1994
## 11  508.1742 492.6060
## 15  461.0153 470.4527
## 18  411.4963 407.7045
## 20  595.9816 605.0610
## 23  432.8445 436.5156
## 30  572.1265 554.7221
## 33  578.7756 588.7126
## 35  514.1126 507.4418
## 36  516.5460 521.8836
## 37  350.6034 347.7769
## 39  480.0344 478.1703
## 42  512.7689 501.8744
## 45  448.0465 448.2298
## 46  550.1735 549.8606
## 49  473.6154 479.7319
## 50  418.5962 416.3584
## 51  724.9205 725.5848
## 54  457.7729 451.4574
## 59  482.0840 496.6507
## 62  501.4642 507.2126
## 63  621.5497 613.5993
```

```
## 65  535.3026 540.2634
## 69  408.7009 408.6202
## 70  445.4202 451.5757
## 71  438.4528 444.9666
## 74  542.5234 534.7772
## 76  455.9795 478.7194
## 84  338.0543 338.3199
## 91  438.6813 449.0703
## 92  606.7504 611.0000
## 95  510.1324 514.0890
## 97  547.8420 521.1430
## 99  502.3868 507.3901
## 104 495.5312 492.1051
## 108 390.6941 378.3309
## 109 555.8821 570.4517
## 113 443.3329 424.7626
## 115 643.0927 642.1016
## 118 577.9139 593.0772
## 127 521.2895 516.8316
## 128 468.1455 468.4457
## 137 519.0515 529.2301
## 138 443.7926 433.0488
## 141 463.4121 448.9333
## 142 466.0112 472.9922
## 144 341.9834 350.0582
## 151 426.3697 426.7752
## 153 559.3112 555.8926
## 154 663.5683 657.0199
## 155 598.0214 595.8038
## 156 502.0153 503.9784
## 158 753.7012 744.2219
## 160 527.4893 528.2238
## 162 366.3128 357.5914
## 164 495.1720 490.2066
## 165 553.1496 550.0476
## 172 450.9126 439.8913
## 175 471.6574 465.1766
## 176 380.1735 373.8857
## 177 538.4631 532.7175
## 180 504.9265 501.1002
## 186 501.1507 485.9231
## 191 608.8314 612.3852
## 200 473.3600 467.5019
## 204 388.6786 392.4974
## 206 714.4572 712.3963
## 208 544.0214 562.0820
## 209 410.4926 412.0129
## 213 537.6871 536.1309
## 214 557.7750 558.4273
## 218 533.6764 528.9336
## 221 517.7632 519.3730
## 223 517.8863 502.4098
## 224 599.3561 604.3348
## 235 510.8772 493.1802
```

```
## 243 431.7222 451.6286
## 244 483.2882 490.6004
## 246 395.9394 409.0705
## 248 650.8254 647.6195
## 252 402.3165 393.8574
## 262 520.2607 514.0098
## 268 402.4221 399.9839
## 272 634.8380 628.0478
## 277 435.0118 423.3083
## 281 507.5070 511.9799
## 285 445.2292 463.5914
## 286 466.8354 471.6029
## 287 609.9729 626.0187
## 299 598.8538 587.5748
## 300 289.5871 282.4712
## 301 492.0325 473.9499
## 302 503.5095 489.9081
## 303 551.5538 541.9722
## 304 284.4081 266.0863
## 316 580.7510 584.1059
## 320 593.5194 596.5167
## 322 531.9004 542.4125
## 329 420.0833 422.3687
## 331 443.5464 442.0644
## 332 524.3591 533.0401
## 333 407.0031 424.2028
## 336 448.1119 443.4419
## 337 477.8060 478.6009
## 341 519.4099 501.1225
## 343 477.8231 486.0834
## 352 530.7603 533.3966
## 357 645.7157 640.5841
## 361 442.5381 444.5761
## 366 589.3243 594.2745
## 371 520.1354 521.2408
## 378 531.5615 538.9420
## 382 535.7787 547.1907
## 383 412.7455 410.6029
## 384 585.6893 583.9778
## 387 530.5515 550.8134
## 395 569.3409 557.6083
## 398 554.3064 547.7100
## 400 436.0912 408.2169
## 402 495.8205 506.3759
## 413 467.6771 444.0538
## 414 498.5803 493.1813
## 416 280.7423 275.9184
## 419 503.1077 475.7251
## 420 478.8161 483.5432
## 426 576.8537 574.4157
## 429 558.7136 556.2981
## 431 547.4542 556.1864
## 432 479.2831 475.0716
## 433 510.2759 486.9471
```

```
## 435 314.0059 304.1356
## 439 399.6642 392.9923
## 441 501.0040 499.1402
## 451 473.7135 475.0154
## 452 430.5494 436.7206
## 454 479.6881 478.1831
## 458 532.2415 534.7715
## 459 542.2499 537.9158
## 465 676.3210 689.2357
## 466 534.2564 543.1326
## 471 433.7179 424.7288
## 478 495.3045 487.5555
## 485 457.6343 462.6565
## 491 501.2629 510.4014
## 495 512.0710 510.6618
## 496 575.7447 573.8474
## 498 555.8924 551.6201
## 500 481.3089 497.7786
```

Evaluating the Model

- **Root Mean Squared Error (RMSE)**
- **R-square and/or Adjusted R-squared**
- **Residual Plots** ** Let's evaluate our model performance by calculating the residual sum of squares and the explained variance score (R^2).**

Calculate the Mean Absolute Error, Mean Squared Error, and the Root Mean Squared Error.

** FOR Predicted values of Test Data (containing only Length.of.Membership) **

```
RMSE(predict2,testdata$Yearly.Amount.Spent)
```

```
## [1] 49.91243
```

```
range(testdata$Yearly.Amount.Spent)
```

```
## [1] 266.0863 744.2219
```

** Mean Squared Error **

```
library(MLmetrics)
```

```
## Warning: package 'MLmetrics' was built under R version 3.5.3
```

```
##
## Attaching package: 'MLmetrics'
```

```
## The following objects are masked from 'package:caret':
##
##     MAE, RMSE
```

```
## The following object is masked from 'package:psych':
##
##     AUC
```

```
## The following object is masked from 'package:base':
##
##     Recall
```

```
MSE(predict2,testdata$Yearly.Amount.Spent)
```

```
## [1] 2491.25
```

** FOR Predicted values of Test Data(using Average.session.Length, TimeonApp and Lengthofthemembership) **

```
RMSE(predict3,testdata$Yearly.Amount.Spent)
```

```
## [1] 10.02303
```

```
range(testdata$Yearly.Amount.Spent)
```

```
## [1] 266.0863 744.2219
```

** Mean Squared Error **

```
MSE(predict3,testdata$Yearly.Amount.Spent)
```

```
## [1] 100.4612
```

Conclusion

** We still want to figure out the answer to the original question, do we focus our efforst on mobile app or website development? * Model 3 is more accurate (Average.session.Length, TimeonApp and Lengthofthemembership)**

```
summary(model3)
```

```
##
## Call:
## lm(formula = Yearly.Amount.Spent ~ Avg..Session.Length + Time.on.App +
##     Length.of.Membership, data = traindata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.6126  -6.9429  -0.2278   6.4231  30.6177
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1017.9761    19.4886  -52.23   <2e-16 ***
## Avg..Session.Length     25.2820     0.5495   46.01   <2e-16 ***
## Time.on.App             38.5136     0.5533   69.61   <2e-16 ***
## Length.of.Membership    61.6365     0.5396  114.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.977 on 348 degrees of freedom
## Multiple R-squared:  0.9836, Adjusted R-squared:  0.9835
## F-statistic:  6957 on 3 and 348 DF,  p-value: < 2.2e-16
```

```
summary(model0)
```

```
##
## Call:
## lm(formula = Yearly.Amount.Spent ~ ., data = Ecommerce.Customers[c(-1,
##     -2, -3)])
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -30.4059  -6.2191  -0.1364   6.6048  30.3085
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1051.5943    22.9925 -45.736   <2e-16 ***
## Avg..Session.Length     25.7343     0.4510  57.057   <2e-16 ***
## Time.on.App             38.7092     0.4510  85.828   <2e-16 ***
## Time.on.Website          0.4367     0.4441   0.983    0.326
## Length.of.Membership    61.5773     0.4483 137.346   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.973 on 495 degrees of freedom
## Multiple R-squared:  0.9843, Adjusted R-squared:  0.9842
## F-statistic:  7766 on 4 and 495 DF,  p-value: < 2.2e-16
```

CONCLUSIONS

- ** From regression Model Time.on.Website is not significant as the p value is not less that 0.05**
- ** Also from the coefficient of Time.on.Website the time coefficent is 0.4367 when annual spent is 0 **
- ** Time on app is a significant variable as per the regression model**
- ** Also the coefficent for Time on App is 38.70, when annual spent is 0 **