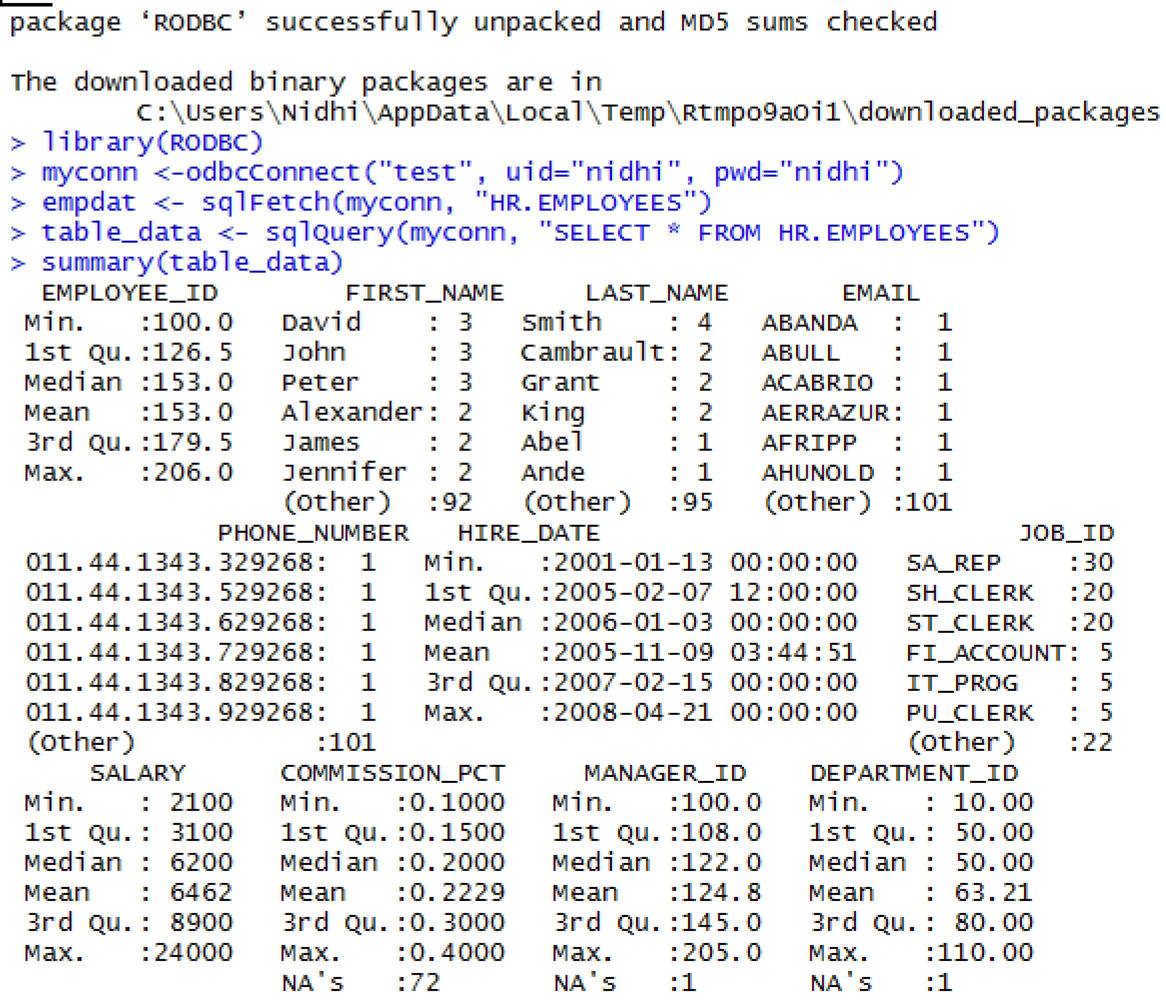
1 a) <u>Create an R data frame set based on the full contents of the HR.Employees table in Oracle XE</u>

```
install.packages("RODBC")
library(RODBC)
myconn <-odbcConnect("test", uid="nidhi", pwd="nidhi")
empdat <- sqlFetch(myconn, "HR.EMPLOYEES")
table_data <- sqlQuery(myconn, "SELECT * FROM HR.EMPLOYEES")
summary(table_data)
```

<u>Output:</u>



```
package ‘RODBC’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Nidhi\AppData\Local\Temp\Rtmpo9aOi1\downloaded_packages
> library(RODBC)
> myconn <-odbcConnect("test", uid="nidhi", pwd="nidhi")
> empdat <- sqlFetch(myconn, "HR.EMPLOYEES")
> table_data <- sqlQuery(myconn, "SELECT * FROM HR.EMPLOYEES")
> summary(table_data)
  EMPLOYEE_ID          FIRST_NAME      LAST_NAME         EMAIL    
 Min.   :100.0   David     : 3   Smith     : 4   ABANDA  :  1  
 1st Qu.:126.5   John      : 3   Cambrault: 2   ABULL   :  1  
 Median :153.0   Peter     : 3   Grant     : 2   ACABRIO :  1  
 Mean   :153.0   Alexander: 2   King      : 2   AERRAZUR:  1  
 3rd Qu.:179.5   James     : 2   Abel      : 1   AFRIPP  :  1  
 Max.   :206.0   Jennifer  : 2   Ande      : 1   AHUNOLD :  1  
                 (Other)   :92   (Other)   :95   (Other) :101  
            PHONE_NUMBER    HIRE_DATE                          JOB_ID  
 011.44.1343.329268:  1   Min.   :2001-01-13 00:00:00   SA_REP    :30  
 011.44.1343.529268:  1   1st Qu.:2005-02-07 12:00:00   SH_CLERK  :20  
 011.44.1343.629268:  1   Median :2006-01-03 00:00:00   ST_CLERK  :20  
 011.44.1343.729268:  1   Mean   :2005-11-09 03:44:51   FI_ACCOUNT: 5  
 011.44.1343.829268:  1   3rd Qu.:2007-02-15 00:00:00   IT_PROG   : 5  
 011.44.1343.929268:  1   Max.   :2008-04-21 00:00:00   PU_CLERK  : 5  
 (Other)           :101                                 (Other)   :22  
     SALARY      COMMISSION_PCT     MANAGER_ID     DEPARTMENT_ID   
 Min.   : 2100   Min.   :0.1000   Min.   :100.0   Min.   : 10.00  
 1st Qu.: 3100   1st Qu.:0.1500   1st Qu.:108.0   1st Qu.: 50.00  
 Median : 6200   Median :0.2000   Median :122.0   Median : 50.00  
 Mean   : 6462   Mean   :0.2229   Mean   :124.8   Mean   : 63.21  
 3rd Qu.: 8900   3rd Qu.:0.3000   3rd Qu.:145.0   3rd Qu.: 80.00  
 Max.   :24000   Max.   :0.4000   Max.   :205.0   Max.   :110.00  
                 NA's   :72       NA's   :1       NA's   :1       
```

b) Use R to update all employees with the last name of TAYLOR to the last name of SMITH

Commands Used

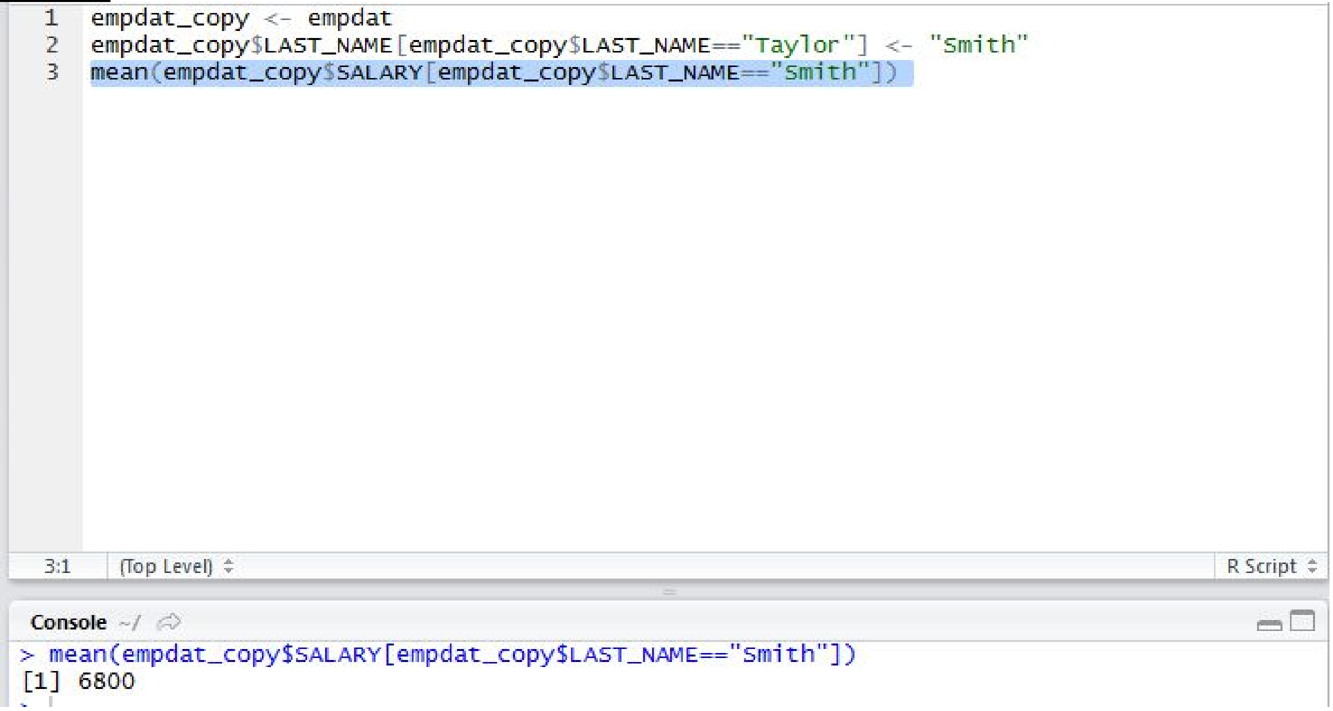
empdat <- sqlFetch(myconn, "HR.EMPLOYEES")
empdat_copy <- empdat
empdat_copy$LAST_NAME[empdat_copy$LAST_NAME=="Taylor"] <- "Smith"

c) Then, use R to determine the average salary of employees with the last name SMITH, using the data frame you updated in the previous step

Commands Used

mean(empdat_copy$SALARY[empdat_copy$LAST_NAME=="Smith"])

Output:



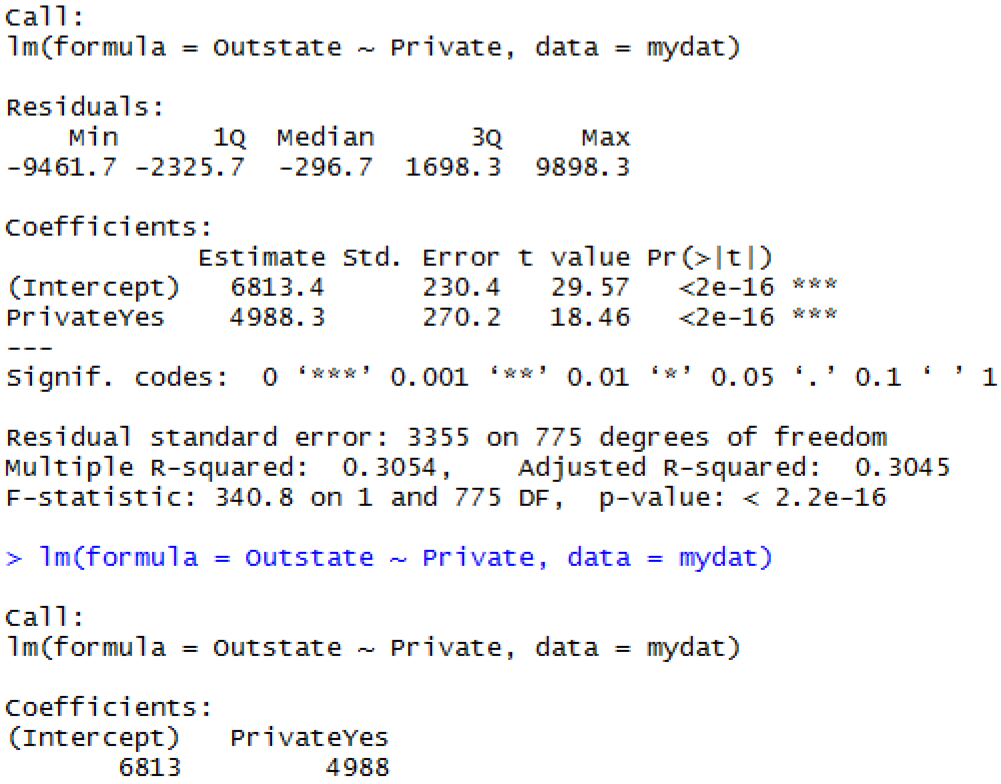
```
empdat_copy <- empdat
empdat_copy$LAST_NAME[empdat_copy$LAST_NAME=="Taylor"] <- "Smith"
mean(empdat_copy$SALARY[empdat_copy$LAST_NAME=="Smith"])
```

```
> mean(empdat_copy$SALARY[empdat_copy$LAST_NAME=="Smith"])
[1] 6800
```

2 a) How much does being a private school raise the out of state tuition costs? (Consider only those two variables in your analysis.) Is the effect statistically significant?

<u>Commands used for linear Regression Analysis</u>:

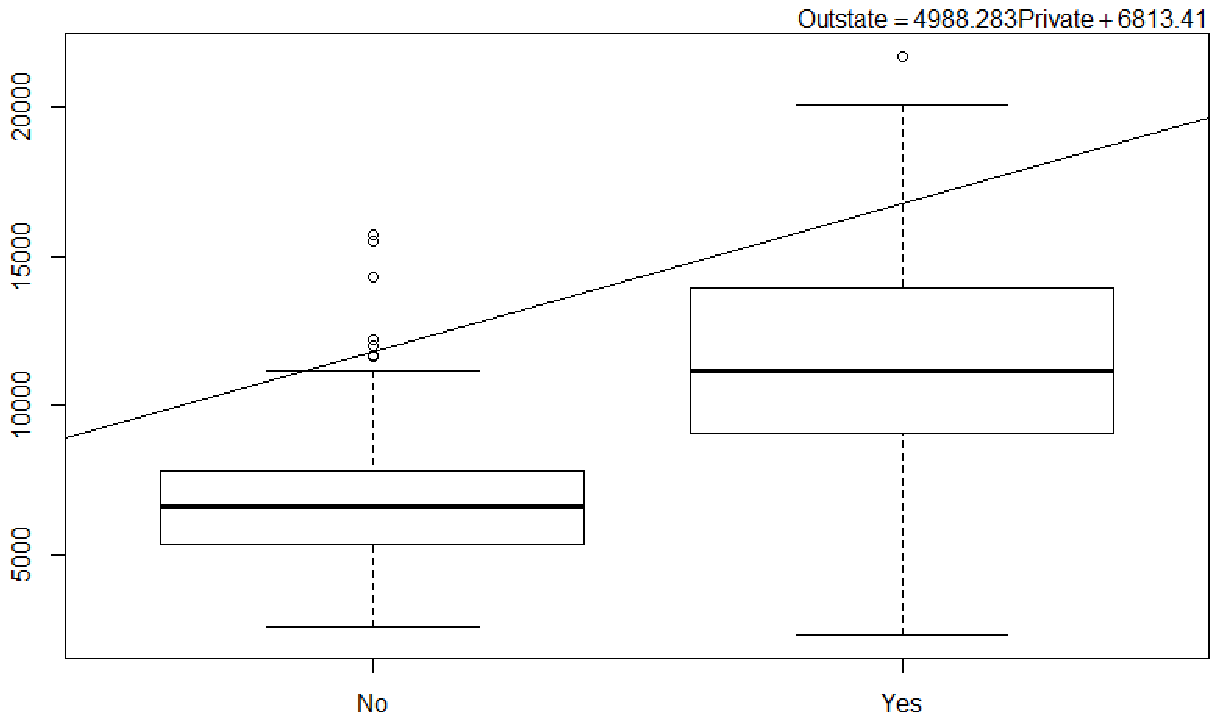
```
install.packages("ISLR")
library(ISLR)

attach(College)
mydat = data.frame(Private, Outstate)
lm.fit = lm(data = mydat, formula=Outstate ~ Private)
summary(lm.fit)
lm(formula = Outstate ~ Private, data = mydat)
plot(Private, Outstate)
abline(lm.fit)
lm_coef <- round(coef(lm.fit), 3)
mtext(bquote(Outstate == .(lm_coef[2])*Private + .(lm_coef[1])),
+  adj=1, padj=0) # display equation
```



```
Call:
lm(formula = Outstate ~ Private, data = mydat)

Residuals:
    Min      1Q  Median      3Q     Max 
-9461.7 -2325.7  -296.7  1698.3  9898.3 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   6813.4      230.4   29.57   <2e-16 ***
PrivateYes    4988.3      270.2   18.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3355 on 775 degrees of freedom
Multiple R-squared:  0.3054,    Adjusted R-squared:  0.3045 
F-statistic: 340.8 on 1 and 775 DF,  p-value: < 2.2e-16

> lm(formula = Outstate ~ Private, data = mydat)

Call:
lm(formula = Outstate ~ Private, data = mydat)

Coefficients:
(Intercept)   PrivateYes  
       6813         4988  
```

Output on console:

Plot Obtained:



The median of Out of state tuition for Private schools is approximately 12,000 where as the median of Out of state tuition for non-Private schools is approximately 7000. So Private Schools raised the out of state tuition fee by approximately 5000.

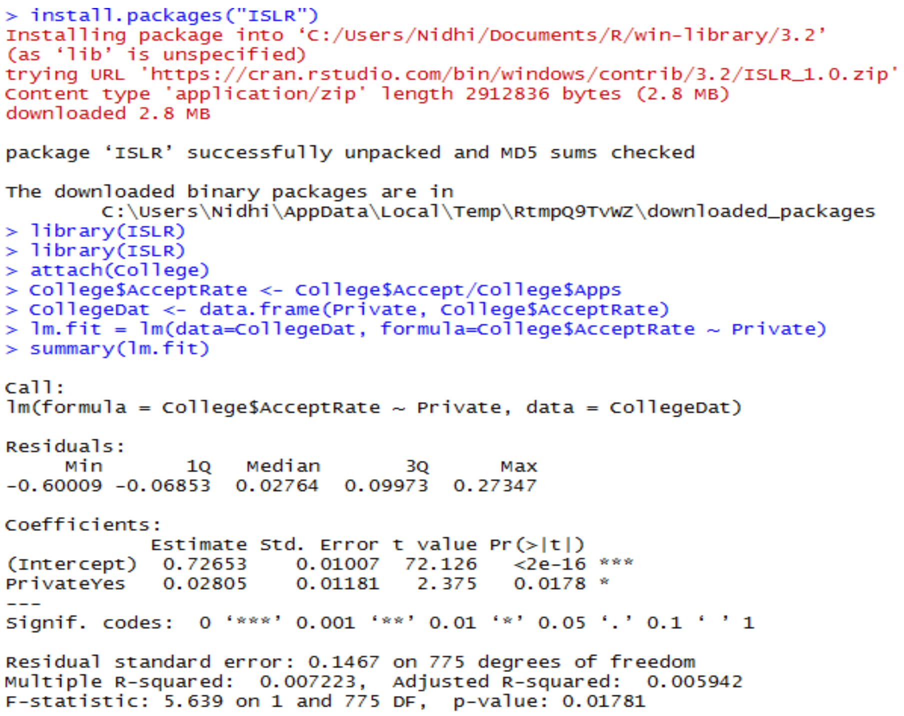
P-value – tells us the likelihood that a variable has a real relationship to what is being measured (measure of statistical significance). P-values range from 0 to 1 – lower values mean that relationship is more likely to be real. P-values of 0.05 and 0.01 usually considered important thresholds. As P-value < 2.2 e-16, therefore the effect is statistically significant.

2.b. Create a column that calculates the acceptance rate as follows:
`College$AcceptRate <- College$Accept/College$Apps`
Do private colleges have a higher or lower acceptance rate than public colleges? (Consider only those two variables in your analysis.) Is the effect statistically significant?

Command used:

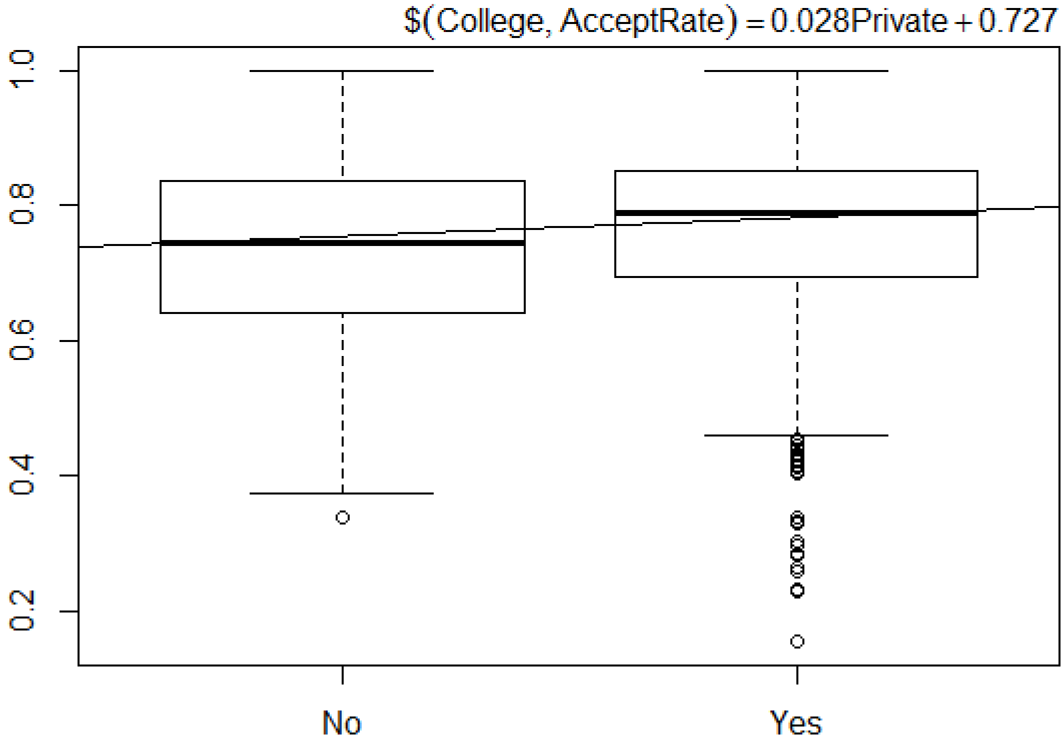
```
install.packages("ISLR")
library(ISLR)
attach(College)
College$AcceptRate <- College$Accept/College$Apps
CollegeDat <- data.frame(Private, College$AcceptRate)
lm.fit = lm(data=CollegeDat, formula=College$AcceptRate ~ Private)
summary(lm.fit)
plot(Private, College$AcceptRate)
abline(lm.fit)
lm_coef <-  round(coef(lm.fit), 3)
mtext(bquote(College$AcceptRate == .(lm_coef[2])*Private + .(lm_coef[1])), adj=1, padj=0) # display equation
```

Output on console:



```
> install.packages("ISLR")
Installing package into ‘C:/Users/Nidhi/Documents/R/win-library/3.2’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.2/ISLR_1.0.zip'
Content type 'application/zip' length 2912836 bytes (2.8 MB)
downloaded 2.8 MB

package ‘ISLR’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Nidhi\AppData\Local\Temp\RtmpQ9TvWZ\downloaded_packages
> library(ISLR)
> library(ISLR)
> attach(College)
> College$AcceptRate <- College$Accept/College$Apps
> CollegeDat <- data.frame(Private, College$AcceptRate)
> lm.fit = lm(data=CollegeDat, formula=College$AcceptRate ~ Private)
> summary(lm.fit)

Call:
lm(formula = College$AcceptRate ~ Private, data = CollegeDat)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.60009 -0.06853  0.02764  0.09973  0.27347 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.72653    0.01007  72.126   <2e-16 ***
PrivateYes   0.02805    0.01181   2.375   0.0178 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1467 on 775 degrees of freedom
Multiple R-squared:  0.007223,	Adjusted R-squared:  0.005942 
F-statistic: 5.639 on 1 and 775 DF,  p-value: 0.01781
```

Plot Obtained:



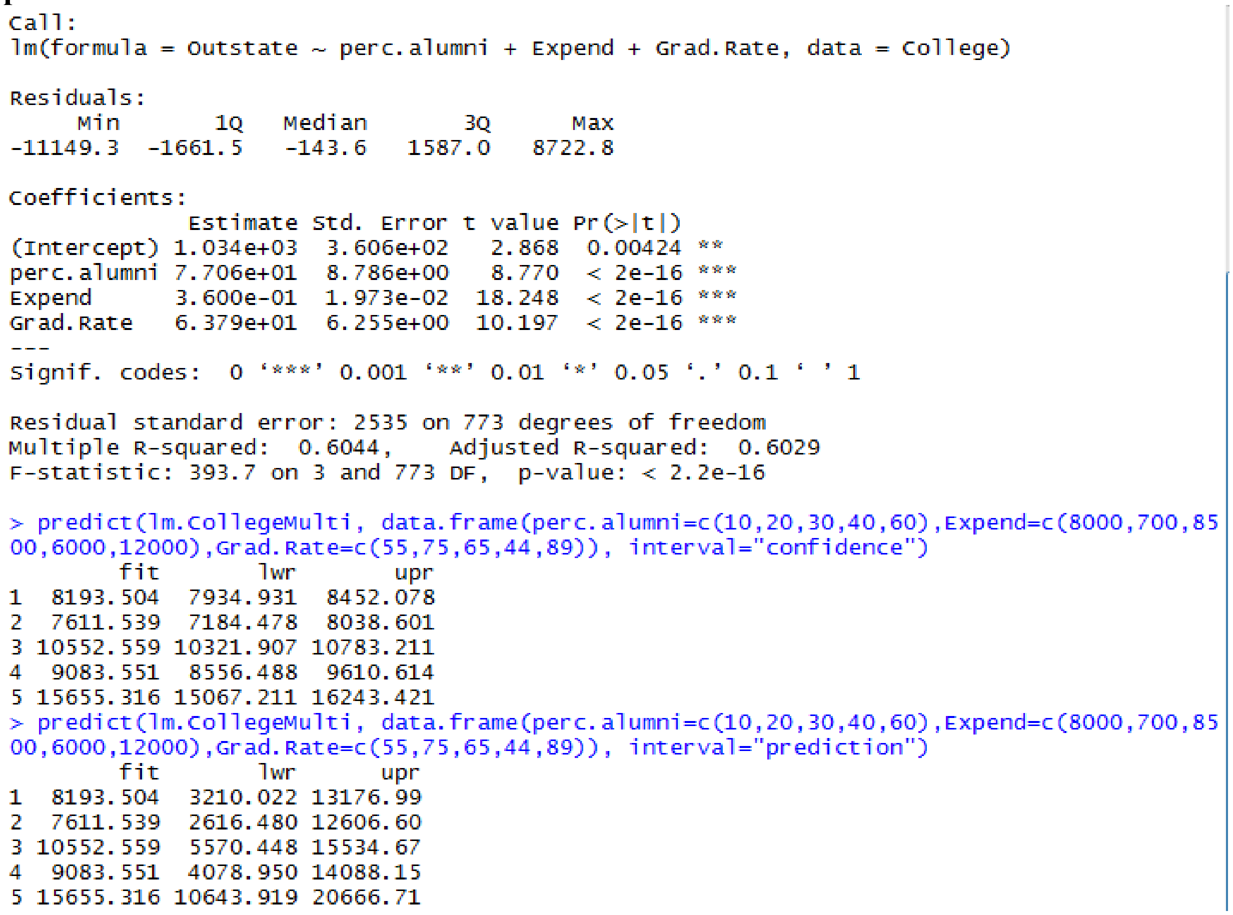
Median value for acceptance rate of private colleges is approximately 0.8 where as Median value for acceptance rate of private colleges is approximately 0.75. So private collges have higher acceptance rate compared to public colleges. As P-value is 0.01, the effect is statistically Significant.

2c. Build a model that predicts out of state tuition based on the percent of alumni who donate (perc.alumni), instructional expenditure per student (Expend), and graduation rate (Grad.rate). Do those three variables account for at least 70% of the variation in out of state tuition?

Commands Used:

```
install.packages("ISLR")
library(ISLR)
lm.CollegeMulti = lm(Outstate~perc.alumni+Expend+Grad.Rate,data=College)
summary(lm.CollegeMulti)
predict(lm.CollegeMulti,
data.frame(perc.alumni=c(10,20,30,40,60),Expend=c(8000,700,8500,6000,12000),Grad.Rate=c(55,75,
65,44,89)), interval="confidence")
predict(lm.CollegeMulti,
data.frame(perc.alumni=c(10,20,30,40,60),Expend=c(8000,700,8500,6000,12000),Grad.Rate=c(55,75,
65,44,89)), interval="prediction")
```

Output On Console:



```
Call:
lm(formula = Outstate ~ perc.alumni + Expend + Grad.Rate, data = College)

Residuals:
     Min       1Q   Median       3Q      Max 
-11149.3  -1661.5   -143.6   1587.0   8722.8 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.034e+03  3.606e+02   2.868  0.00424 ** 
perc.alumni 7.706e+01  8.786e+00   8.770  < 2e-16 ***
Expend      3.600e-01  1.973e-02  18.248  < 2e-16 ***
Grad.Rate   6.379e+01  6.255e+00  10.197  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2535 on 773 degrees of freedom
Multiple R-squared:  0.6044,	Adjusted R-squared:  0.6029 
F-statistic: 393.7 on 3 and 773 DF,  p-value: < 2.2e-16

> predict(lm.CollegeMulti, data.frame(perc.alumni=c(10,20,30,40,60),Expend=c(8000,700,85
00,6000,12000),Grad.Rate=c(55,75,65,44,89)), interval="confidence")
        fit       lwr       upr
1  8193.504  7934.931  8452.078
2  7611.539  7184.478  8038.601
3 10552.559 10321.907 10783.211
4  9083.551  8556.488  9610.614
5 15655.316 15067.211 16243.421
> predict(lm.CollegeMulti, data.frame(perc.alumni=c(10,20,30,40,60),Expend=c(8000,700,85
00,6000,12000),Grad.Rate=c(55,75,65,44,89)), interval="prediction")
        fit       lwr      upr
1  8193.504  3210.022 13176.99
2  7611.539  2616.480 12606.60
3 10552.559  5570.448 15534.67
4  9083.551  4078.950 14088.15
5 15655.316 10643.919 20666.71
```

We assess a regression model using R2 statistic. It Tells us proportion of the variability in Y that can be explained by X Value between 0 and 1. Higher values mean that the model is more predictive. “Good” value relative to domain. Usually want to look at “Adjusted R2” which penalizes more complex models and small sample sizes.

As R squared value is 0.6044, so it accounts for 60% Variation.