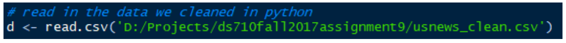
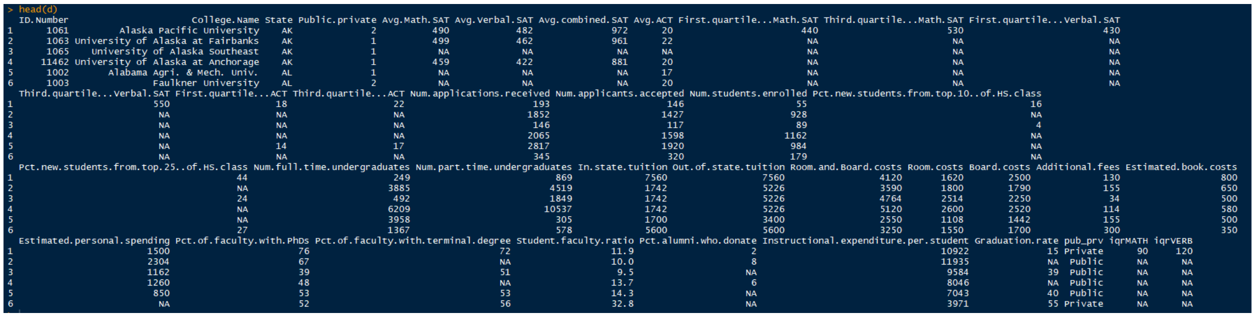
ADAM HENDEL
DS 710 - Homework 9
R assignment

1(a) Read the modified data into R. Check the first few values of each vector to ensure that they were read accurately.



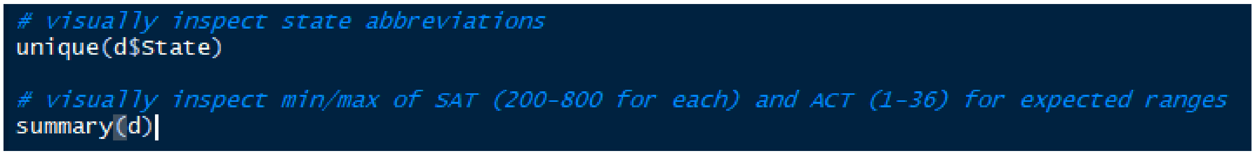
```
# read in the data we cleaned in python
d <- read.csv('D:/Projects/ds710fall2017assignment9/usnews_clean.csv')
```



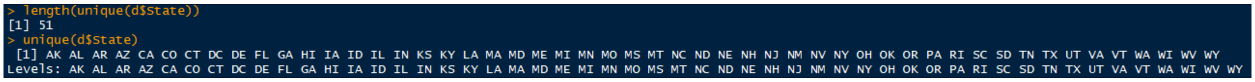
```
> head(d)
  ID.Number                      College.Name State Public.private Avg.Math.SAT Avg.Verbal.SAT Avg.combined.SAT Avg.ACT First.quartile...Math.SAT Third.quartile...Math.SAT First.quartile...Verbal.SAT
1      1061         Alaska Pacific University    AK              2          490            482              972      20                       440                       530                         430
2      1063 University of Alaska at Fairbanks    AK              1          499            462              961      22                        NA                        NA                          NA
3      1065    University of Alaska Southeast    AK              1           NA             NA               NA      NA                        NA                        NA                          NA
4     11462 University of Alaska at Anchorage    AK              1          459            422              881      20                        NA                        NA                          NA
5      1002       Alabama Agri. & Mech. Univ.    AL              1           NA             NA               NA      17                        NA                        NA                          NA
6      1003               Faulkner University    AL              2           NA             NA               NA      20                        NA                        NA                          NA
  Third.quartile...Verbal.SAT First.quartile...ACT Third.quartile...ACT Num.applications.received Num.applicants.accepted Num.students.enrolled Pct.new.students.from.top.10..of.HS.class
1                         550                   18                   22                       193                     146                    55                                        16
2                          NA                   NA                   NA                      1852                    1427                   928                                        NA
3                          NA                   NA                   NA                       146                     117                    89                                         4
4                          NA                   NA                   NA                      2065                    1598                  1162                                        NA
5                          NA                   14                   17                      2817                    1920                   984                                        NA
6                          NA                   NA                   NA                       345                     320                   179                                        NA
  Pct.new.students.from.top.25..of.HS.class Num.full.time.undergraduates Num.part.time.undergraduates In.state.tuition Out.of.state.tuition Room.and.Board.costs Room.costs Board.costs Additional.fees Estimated.book.costs
1                                        44                          249                          869             7560                 7560                 4120       1620        2500             130                  800
2                                        NA                         3885                         4519             1742                 5226                 3590       1800        1790             155                  650
3                                        24                          492                         1849             1742                 5226                 4764       2514        2250              34                  500
4                                        NA                         6209                        10537             1742                 5226                 5120       2600        2520             114                  580
5                                        NA                         3958                          305             1700                 3400                 2550       1108        1442             155                  500
6                                        27                         1367                          578             5600                 5600                 3250       1550        1700             300                  350
  Estimated.personal.spending Pct.of.faculty.with.PhDs Pct.of.faculty.with.terminal.degree Student.faculty.ratio Pct.alumni.who.donate Instructional.expenditure.per.student Graduation.rate pub_prv iqrMATH iqrVERB
1                        1500                       76                                  72                  11.9                     2                                 10922              15 Private      90     120
2                        2304                       67                                  NA                  10.0                     8                                 11935              NA  Public      NA      NA
3                        1162                       39                                  51                   9.5                    NA                                  9584              39  Public      NA      NA
4                        1260                       48                                  NA                  13.7                     6                                  8046              NA  Public      NA      NA
5                         850                       53                                  53                  14.3                    NA                                  7043              40  Public      NA      NA
6                          NA                       52                                  56                  32.8                    NA                                  3971              55 Private      NA      NA
```

New columns came in properly and NA values reported as expected.

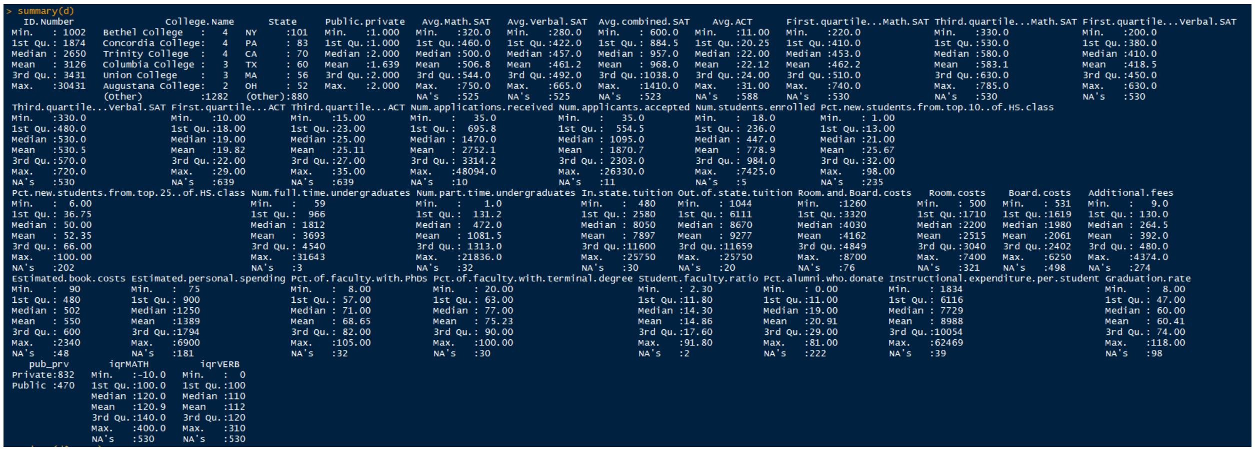
1(b) Examine the summary of each variable. Identify any unrealistic values and set them to missing. Write a sentence describing what you did, naming the colleges or universities affected. (For example, "Listed ages less than zero (ABC University, XYZ College) were converted to missing data.")



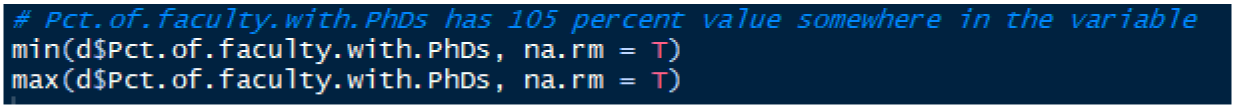
```
# visually inspect state abbreviations
unique(d$State)

# visually inspect min/max of SAT (200-800 for each) and ACT (1-36) for expected ranges
summary(d)
```



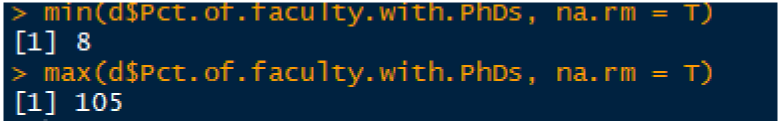
```
> length(unique(d$State))
[1] 51
> unique(d$State)
 [1] AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ NM NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV WY
Levels: AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ NM NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV WY
```



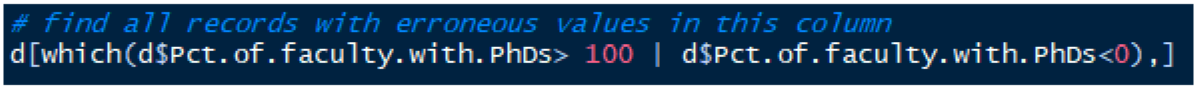
```
> summary(d)
   ID.Number                College.Name      State      Public.private   Avg.Math.SAT   Avg.Verbal.SAT  Avg.combined.SAT    Avg.ACT      First.quartile...Math.SAT Third.quartile...Math.SAT First.quartile...Verbal.SAT
 Min.   : 1002   Bethel College    :   4   NY     :101   Min.   :1.000   Min.   :320.0   Min.   :280.0   Min.   : 600.0   Min.   :11.00   Min.   :220.0             Min.   :330.0             Min.   :200.0
 1st Qu.: 1874   Concordia College:   4   PA     : 83   1st Qu.:1.000   1st Qu.:460.0   1st Qu.:422.0   1st Qu.: 884.5   1st Qu.:20.25   1st Qu.:410.0             1st Qu.:530.0             1st Qu.:380.0
 Median : 2650   Trinity College  :   4   CA     : 70   Median :2.000   Median :500.0   Median :457.0   Median : 957.0   Median :22.00   Median :453.0             Median :580.0             Median :410.0
 Mean   : 3126   Columbia College :   3   TX     : 60   Mean   :1.639   Mean   :506.8   Mean   :461.2   Mean   : 968.0   Mean   :22.12   Mean   :462.2             Mean   :583.1             Mean   :418.5
 3rd Qu.: 3431   Union College    :   3   MA     : 56   3rd Qu.:2.000   3rd Qu.:544.0   3rd Qu.:492.0   3rd Qu.:1038.0   3rd Qu.:24.00   3rd Qu.:510.0             3rd Qu.:630.0             3rd Qu.:450.0
 Max.   :30431   Augustana College:   2   OH     : 52   Max.   :2.000   Max.   :750.0   Max.   :665.0   Max.   :1410.0   Max.   :31.00   Max.   :740.0             Max.   :785.0             Max.   :630.0
                 (Other)          :1282   (Other):880                   NA's   :525     NA's   :525     NA's   :523      NA's   :588     NA's   :530               NA's   :530               NA's   :530
 Third.quartile...Verbal.SAT First.quartile...ACT Third.quartile...ACT Num.applications.received Num.applicants.accepted Num.students.enrolled Pct.new.students.from.top.10..of.HS.class
 Min.   :330.0               Min.   :10.00        Min.   :15.00        Min.   :   35.0           Min.   :   35.0         Min.   :  18.0        Min.   : 1.00
 1st Qu.:480.0               1st Qu.:18.00        1st Qu.:23.00        1st Qu.:  695.8           1st Qu.:  554.5         1st Qu.: 236.0        1st Qu.:13.00
 Median :530.0               Median :19.00        Median :25.00        Median : 1470.0           Median : 1095.0         Median : 447.0        Median :21.00
 Mean   :530.5               Mean   :19.82        Mean   :25.11        Mean   : 2752.1           Mean   : 1870.7         Mean   : 778.9        Mean   :25.67
 3rd Qu.:570.0               3rd Qu.:22.00        3rd Qu.:27.00        3rd Qu.: 3314.2           3rd Qu.: 2303.0         3rd Qu.: 984.0        3rd Qu.:32.00
 Max.   :720.0               Max.   :29.00        Max.   :35.00        Max.   :48094.0           Max.   :26330.0         Max.   :7425.0        Max.   :98.00
 NA's   :530                 NA's   :639          NA's   :639          NA's   :10                NA's   :11              NA's   :5             NA's   :235
 Pct.new.students.from.top.25..of.HS.class Num.full.time.undergraduates Num.part.time.undergraduates In.state.tuition Out.of.state.tuition Room.and.Board.costs   Room.costs    Board.costs    Additional.fees
 Min.   :  6.00                            Min.   :   59                Min.   :    1.0              Min.   :  480    Min.   : 1044        Min.   :1260         Min.   : 500   Min.   : 531   Min.   :   9.0
 1st Qu.: 36.75                            1st Qu.:  966                1st Qu.:  131.2              1st Qu.: 2580    1st Qu.: 6111        1st Qu.:3320         1st Qu.:1710   1st Qu.:1619   1st Qu.: 130.0
 Median : 50.00                            Median : 1812                Median :  472.0              Median : 8050    Median : 8670        Median :4030         Median :2200   Median :1980   Median : 264.5
 Mean   : 52.35                            Mean   : 3693                Mean   : 1081.5              Mean   : 7897    Mean   : 9277        Mean   :4162         Mean   :2515   Mean   :2061   Mean   : 392.0
 3rd Qu.: 66.00                            3rd Qu.: 4540                3rd Qu.: 1313.0              3rd Qu.:11600    3rd Qu.:11659        3rd Qu.:4849         3rd Qu.:3040   3rd Qu.:2402   3rd Qu.: 480.0
 Max.   :100.00                            Max.   :31643                Max.   :21836.0              Max.   :25750    Max.   :25750        Max.   :8700         Max.   :7400   Max.   :6250   Max.   :4374.0
 NA's   :202                               NA's   :3                    NA's   :32                   NA's   :30       NA's   :20           NA's   :76           NA's   :321    NA's   :498    NA's   :274
 Estimated.book.costs Estimated.personal.spending Pct.of.faculty.with.PhDs Pct.of.faculty.with.terminal.degree Student.faculty.ratio Pct.alumni.who.donate Instructional.expenditure.per.student Graduation.rate
 Min.   :  90         Min.   :  75                Min.   :  8.00           Min.   : 20.00                      Min.   : 2.30         Min.   : 0.00         Min.   : 1834                         Min.   :  8.00
 1st Qu.: 480         1st Qu.: 900                1st Qu.: 57.00           1st Qu.: 63.00                      1st Qu.:11.80         1st Qu.:11.00         1st Qu.: 6116                         1st Qu.: 47.00
 Median : 502         Median :1250                Median : 71.00           Median : 77.00                      Median :14.30         Median :19.00         Median : 7729                         Median : 60.00
 Mean   : 550         Mean   :1389                Mean   : 68.65           Mean   : 75.23                      Mean   :14.86         Mean   :20.91         Mean   : 8988                         Mean   : 60.41
 3rd Qu.: 600         3rd Qu.:1794                3rd Qu.: 82.00           3rd Qu.: 90.00                      3rd Qu.:17.60         3rd Qu.:29.00         3rd Qu.:10054                         3rd Qu.: 74.00
 Max.   :2340         Max.   :6900                Max.   :105.00           Max.   :100.00                      Max.   :91.80         Max.   :81.00         Max.   :62469                         Max.   :118.00
 NA's   :48           NA's   :181                 NA's   :32               NA's   :30                          NA's   :2             NA's   :222           NA's   :39                            NA's   :98
    pub_prv        iqrMATH          iqrVERB
 Private:832   Min.   :-10.0   Min.   :  0
 Public :470   1st Qu.:100.0   1st Qu.:100
               Median :120.0   Median :110
               Mean   :120.9   Mean   :112
               3rd Qu.:140.0   3rd Qu.:120
               Max.   :400.0   Max.   :310
               NA's   :530     NA's   :530
```



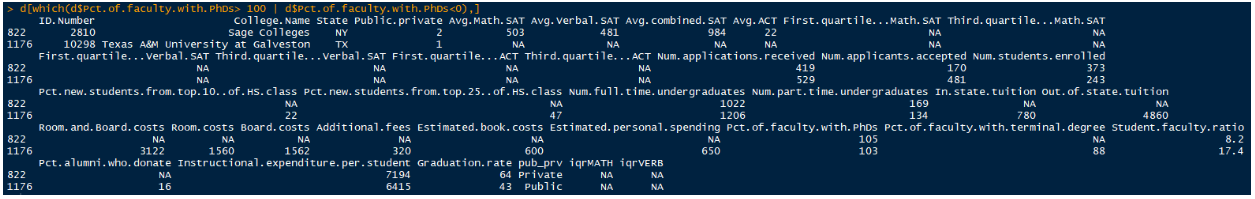
```
# Pct.of.faculty.with.PhDs has 105 percent value somewhere in the variable
min(d$Pct.of.faculty.with.PhDs, na.rm = T)
max(d$Pct.of.faculty.with.PhDs, na.rm = T)
```



```
> min(d$Pct.of.faculty.with.PhDs, na.rm = T)
[1] 8
> max(d$Pct.of.faculty.with.PhDs, na.rm = T)
[1] 105
```

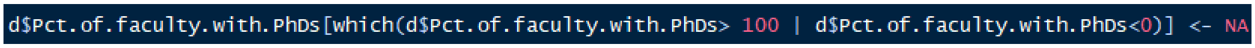


```
# find all records with erroneous values in this column
d[which(d$Pct.of.faculty.with.PhDs> 100 | d$Pct.of.faculty.with.PhDs<0),]
```



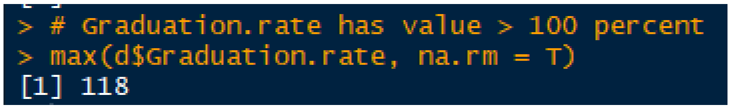
```
> d[which(d$Pct.of.faculty.with.PhDs> 100 | d$Pct.of.faculty.with.PhDs<0),]
     ID.Number                       College.Name State Public.private Avg.Math.SAT Avg.Verbal.SAT Avg.combined.SAT Avg.ACT First.quartile...Math.SAT Third.quartile...Math.SAT
822       2810                     Sage Colleges    NY              2          503            481              984      22                        NA                        NA
1176     10298 Texas A&M University at Galveston    TX              1           NA             NA               NA      NA                        NA                        NA
     First.quartile...Verbal.SAT Third.quartile...Verbal.SAT First.quartile...ACT Third.quartile...ACT Num.applications.received Num.applicants.accepted Num.students.enrolled
822                           NA                          NA                   NA                   NA                       419                     170                   373
1176                          NA                          NA                   NA                   NA                       529                     481                   243
     Pct.new.students.from.top.10..of.HS.class Pct.new.students.from.top.25..of.HS.class Num.full.time.undergraduates Num.part.time.undergraduates In.state.tuition Out.of.state.tuition
822                                         NA                                        NA                         1022                          169               NA                   NA
1176                                        22                                        47                         1206                          134              780                 4860
     Room.and.Board.costs Room.costs Board.costs Additional.fees Estimated.book.costs Estimated.personal.spending Pct.of.faculty.with.PhDs Pct.of.faculty.with.terminal.degree Student.faculty.ratio
822                    NA         NA          NA              NA                   NA                          NA                      105                                  NA                   8.2
1176                 3122       1560        1562             320                  600                         650                      103                                  88                  17.4
     Pct.alumni.who.donate Instructional.expenditure.per.student Graduation.rate pub_prv iqrMATH iqrVERB
822                     NA                                  7194              64 Private      NA      NA
1176                    16                                  6415              43  Public      NA      NA
```

# two schools have percentages >100 (Sage Colleges, Texas A&M) for this variable---set them to NA

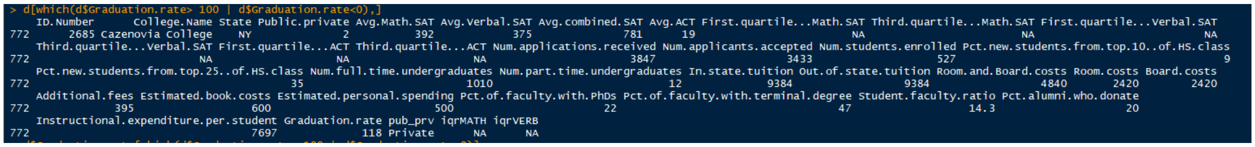


```
d$Pct.of.faculty.with.PhDs[which(d$Pct.of.faculty.with.PhDs> 100 | d$Pct.of.faculty.with.PhDs<0)] <- NA
```

# grad rates > 100 or < 0 are not possible

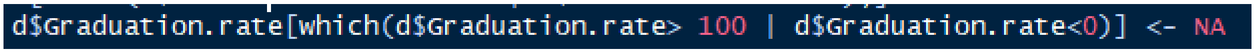


```
> # Graduation.rate has value > 100 percent
> max(d$Graduation.rate, na.rm = T)
[1] 118
```



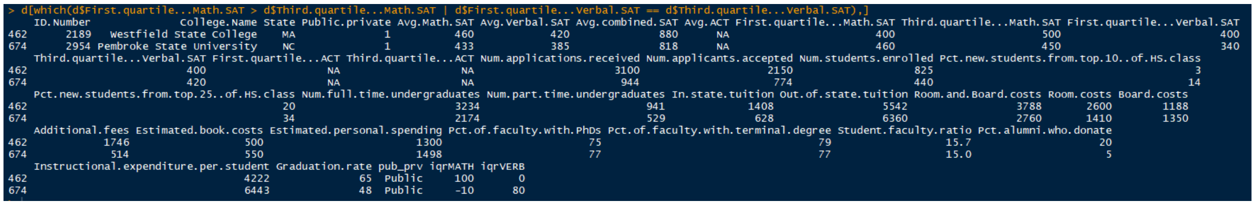
```
> d[which(d$Graduation.rate> 100 | d$Graduation.rate<0),]
    ID.Number      College.Name State Public.private Avg.Math.SAT Avg.Verbal.SAT Avg.combined.SAT Avg.ACT First.quartile...Math.SAT Third.quartile...Math.SAT First.quartile...Verbal.SAT
772      2685 Cazenovia College    NY              2          392            375              781      19                        NA                        NA                          NA
    Third.quartile...Verbal.SAT First.quartile...ACT Third.quartile...ACT Num.applications.received Num.applicants.accepted Num.students.enrolled Pct.new.students.from.top.10..of.HS.class
772                          NA                   NA                   NA                      3847                    3433                   527                                         9
    Pct.new.students.from.top.25..of.HS.class Num.full.time.undergraduates Num.part.time.undergraduates In.state.tuition Out.of.state.tuition Room.and.Board.costs Room.costs Board.costs
772                                        35                         1010                           12             9384                 9384                 4840       2420        2420
    Additional.fees Estimated.book.costs Estimated.personal.spending Pct.of.faculty.with.PhDs Pct.of.faculty.with.terminal.degree Student.faculty.ratio Pct.alumni.who.donate
772             395                  600                         500                       22                                  47                  14.3                    20
    Instructional.expenditure.per.student Graduation.rate pub_prv iqrMATH iqrVERB
772                                  7697             118 Private      NA      NA
```

Cazenovia College has 118 percent grad rate, set it to NA



```
d$Graduation.rate[which(d$Graduation.rate> 100 | d$Graduation.rate<0)] <- NA
```

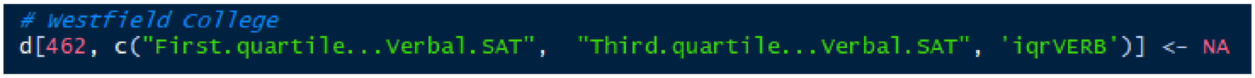
iqrMATH min value is reported at -10, so there is a record where 3QT > 1QT. Likewise, iqrVERB min is reported at 0, so 3QT==1QT in a record. We should take a look at these records.



```
> d[which(d$First.quartile...Math.SAT > d$Third.quartile...Math.SAT | d$First.quartile...Verbal.SAT == d$Third.quartile...Verbal.SAT),]
    ID.Number              College.Name State Public.private Avg.Math.SAT Avg.Verbal.SAT Avg.combined.SAT Avg.ACT First.quartile...Math.SAT Third.quartile...Math.SAT First.quartile...Verbal.SAT
462      2189   Westfield State College    MA              1          460            420              880      NA                       400                       500                         400
674      2954 Pembroke State University    NC              1          433            385              818      NA                       460                       450                         340
    Third.quartile...Verbal.SAT First.quartile...ACT Third.quartile...ACT Num.applications.received Num.applicants.accepted Num.students.enrolled Pct.new.students.from.top.10..of.HS.class
462                         400                   NA                   NA                      3100                    2150                   825                                         3
674                         420                   NA                   NA                       944                     774                   440                                        14
    Pct.new.students.from.top.25..of.HS.class Num.full.time.undergraduates Num.part.time.undergraduates In.state.tuition Out.of.state.tuition Room.and.Board.costs Room.costs Board.costs
462                                        20                         3234                          941             1408                 5542                 3788       2600        1188
674                                        34                         2174                          529              628                 6360                 2760       1410        1350
    Additional.fees Estimated.book.costs Estimated.personal.spending Pct.of.faculty.with.PhDs Pct.of.faculty.with.terminal.degree Student.faculty.ratio Pct.alumni.who.donate
462            1746                  500                        1300                       75                                  79                  15.7                    20
674             514                  550                        1498                       77                                  77                  15.0                     5
    Instructional.expenditure.per.student Graduation.rate pub_prv iqrMATH iqrVERB
462                                  4222              65  Public     100       0
674                                  6443              48  Public     -10      80
```

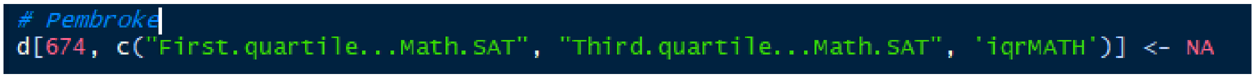
Westfield State College has Verbal 3QT == 1QT and Pembroke State University has Math 1QT>3QT.

Westfield-set Verbal 1QT, 3QT and IQR to NA



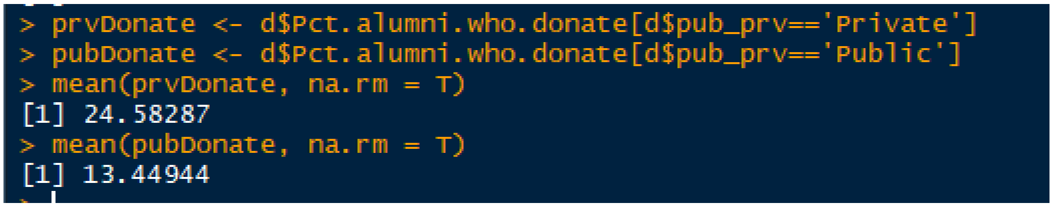
```
# Westfield College
d[462, c("First.quartile...Verbal.SAT",  "Third.quartile...Verbal.SAT", 'iqrVERB')] <- NA
```

Pembroke-set Math 1QT, 3QT and IQR to NA



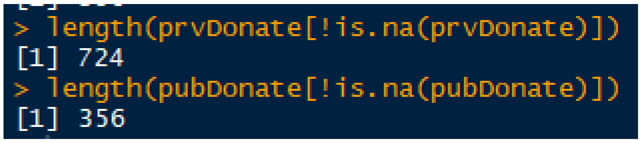
```
# Pembroke
d[674, c("First.quartile...Math.SAT", "Third.quartile...Math.SAT", 'iqrMATH')] <- NA
```

1(c) Find the mean percentage of alumni who donate, for private and public schools.

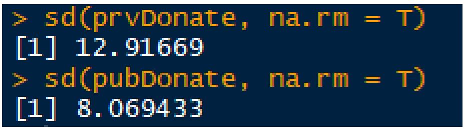


```
> prvDonate <- d$Pct.alumni.who.donate[d$pub_prv=='Private']
> pubDonate <- d$Pct.alumni.who.donate[d$pub_prv=='Public']
> mean(prvDonate, na.rm = T)
[1] 24.58287
> mean(pubDonate, na.rm = T)
[1] 13.44944
```

1(d) The two groups, public and private, have neither the same number of samples nor the same variance (as shown below)

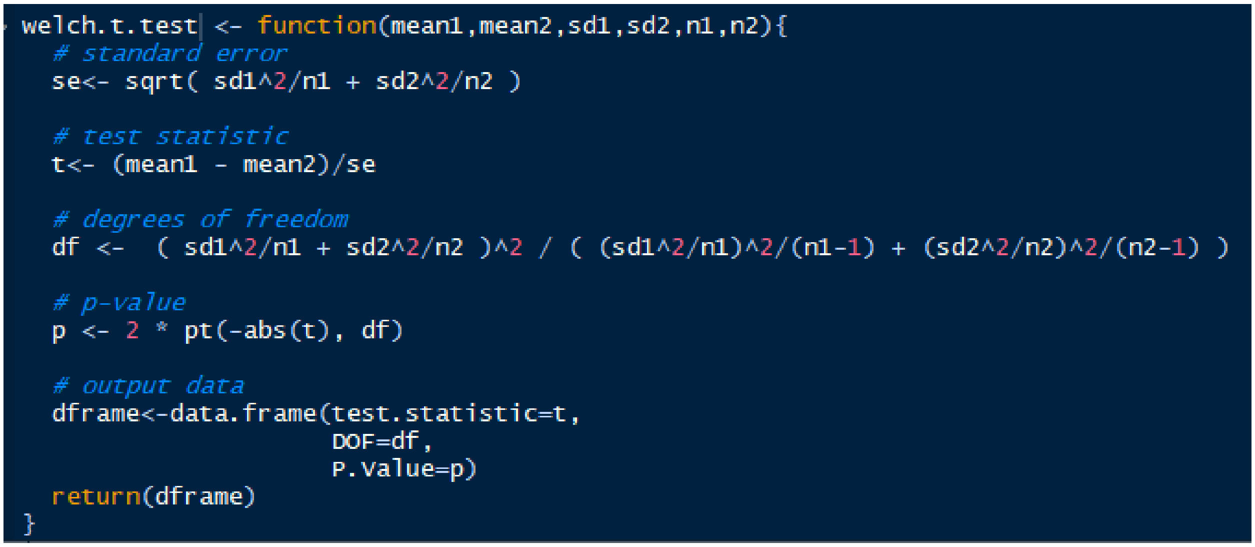


```
> length(prvDonate[!is.na(prvDonate)])
[1] 724
> length(pubDonate[!is.na(pubDonate)])
[1] 356
```

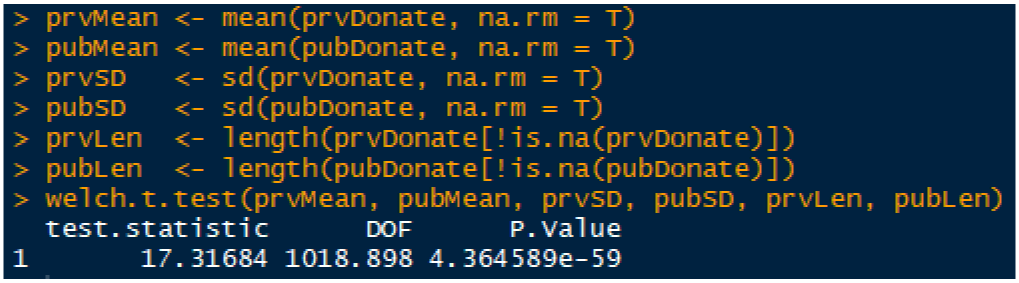


```
> sd(prvDonate, na.rm = T)
[1] 12.91669
> sd(pubDonate, na.rm = T)
[1] 8.069433
```

Thus, we will use Welch's t-test (same method from assignment 8)
https://en.wikipedia.org/wiki/Welch%27s_t-test:



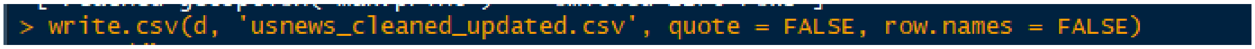
```
welch.t.test <- function(mean1,mean2,sd1,sd2,n1,n2){
  # standard error
  se<- sqrt( sd1^2/n1 + sd2^2/n2 )

  # test statistic
  t<- (mean1 - mean2)/se

  # degrees of freedom
  df <-  ( sd1^2/n1 + sd2^2/n2 )^2 / ( (sd1^2/n1)^2/(n1-1) + (sd2^2/n2)^2/(n2-1) )

  # p-value
  p <- 2 * pt(-abs(t), df)

  # output data
  dframe<-data.frame(test.statistic=t,
                     DOF=df,
                     P.Value=p)
  return(dframe)
}
```



```
> prvMean <- mean(prvDonate, na.rm = T)
> pubMean <- mean(pubDonate, na.rm = T)
> prvSD   <- sd(prvDonate, na.rm = T)
> pubSD   <- sd(pubDonate, na.rm = T)
> prvLen  <- length(prvDonate[!is.na(prvDonate)])
> pubLen  <- length(pubDonate[!is.na(pubDonate)])
> welch.t.test(prvMean, pubMean, prvSD, pubSD, prvLen, pubLen)
  test.statistic      DOF      P.Value
1       17.31684 1018.898 4.364589e-59
```

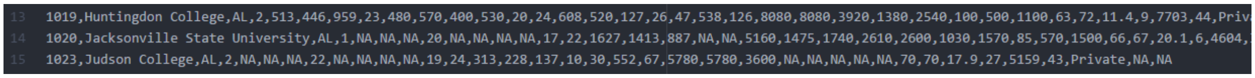
Given a *p-value* close to zero, there is enough evidence to reject the null hypothesis that the two means are equal. The percentage of alumni that donate back to the institution is greater in private schools than in public schools.

1(e) Write to CSV



```
> write.csv(d, 'usnews_cleaned_updated.csv', quote = FALSE, row.names = FALSE)
```

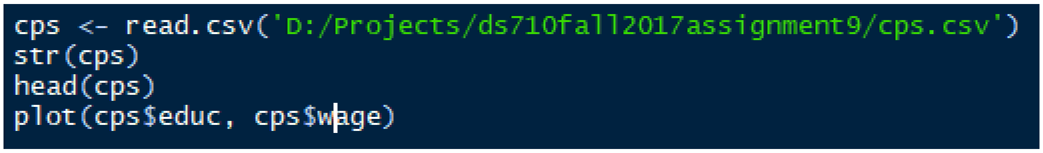
Confim w/ text editor (Atom) that quotations did not show up around entries.



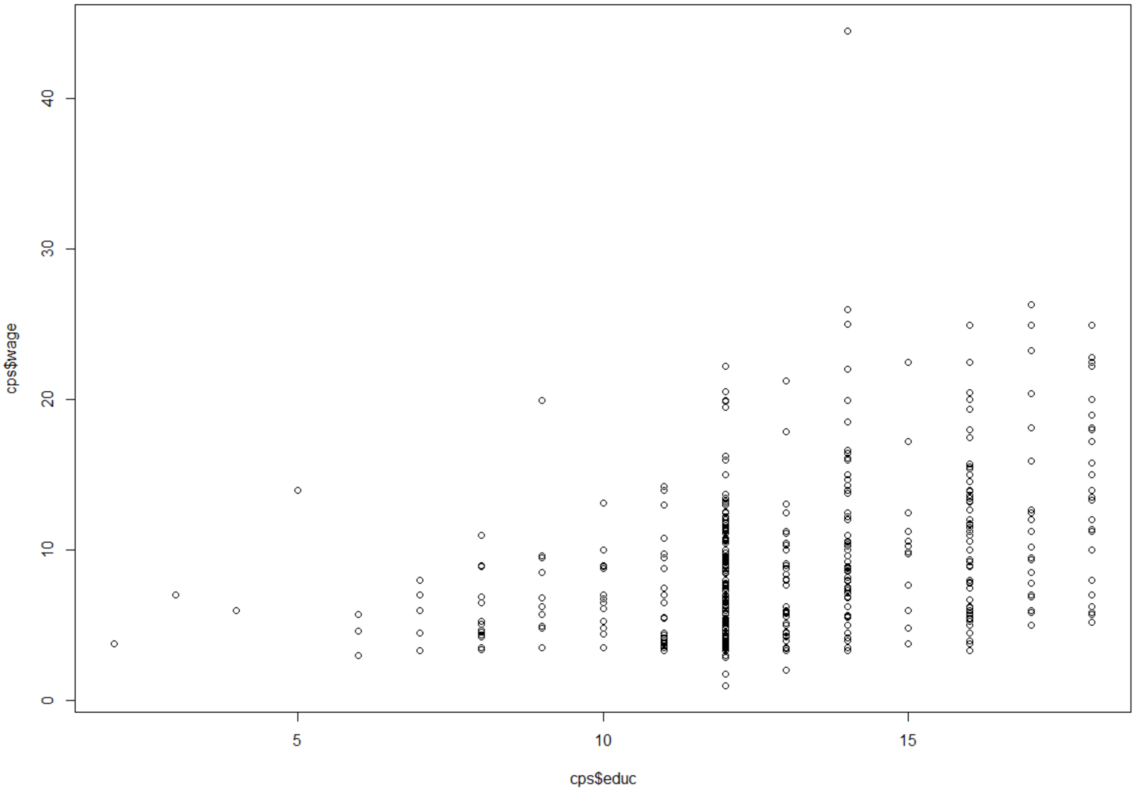
```
1019,Huntingdon College,AL,2,513,446,959,23,480,570,400,530,20,24,608,520,127,26,47,538,126,8080,8080,3920,1380,2540,100,500,1100,63,72,11.4,9,7703,44,Priv
1020,Jacksonville State University,AL,1,NA,NA,NA,20,NA,NA,NA,NA,17,22,1627,1413,887,NA,NA,5160,1475,1740,2610,2600,1030,1570,85,570,1500,66,67,20.1,6,4604,
1023,Judson College,AL,2,NA,NA,NA,22,NA,NA,NA,NA,19,24,313,228,137,10,30,552,67,5780,5780,3600,NA,NA,NA,NA,NA,70,70,17.9,27,5159,43,Private,NA,NA
```

2(a)

Read the data into R and plot wages versus education.

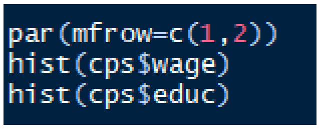


```
cps <- read.csv('D:/Projects/ds710fall2017assignment9/cps.csv')
str(cps)
head(cps)
plot(cps$educ, cps$wage)
```

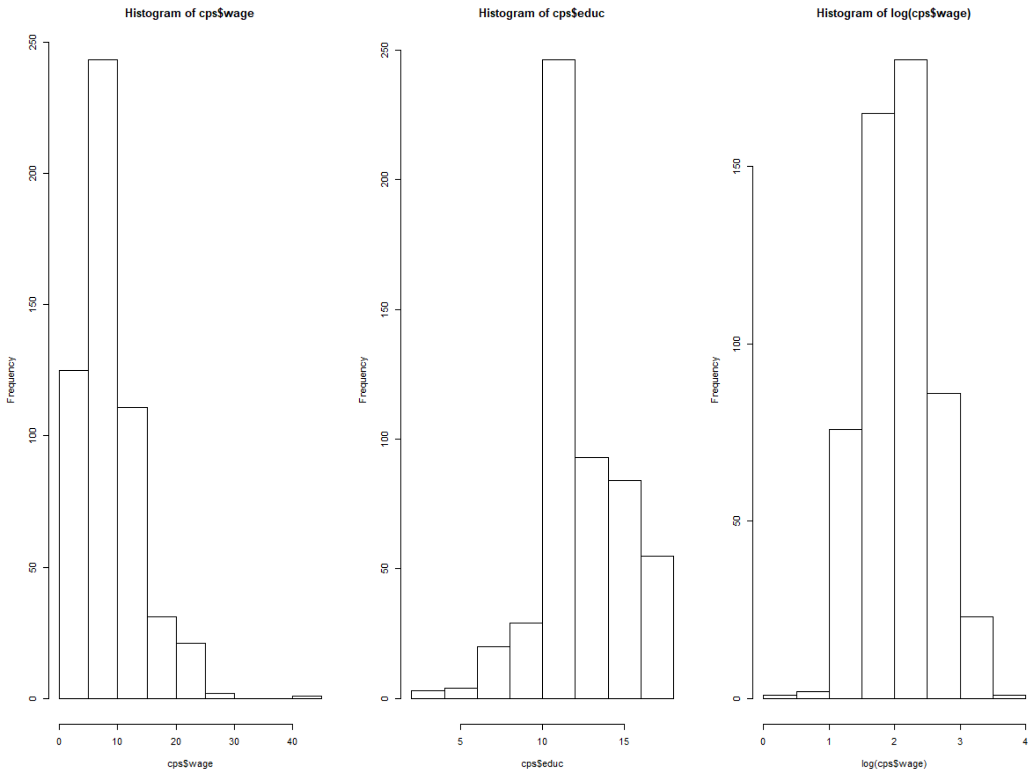


There appears to be a relationship between wages and education. Low education seems to have lower wages, but also lower observations. The min wage of higher education levels seems to be higher than lower educations as well. A linear regression might help explain the significant of the linear relationship.

We observe right skewness on the wage variable, thus a log transformation might help our analysis.

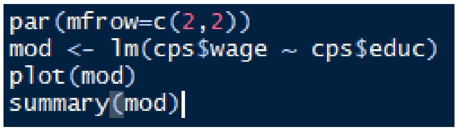


```
par(mfrow=c(1,2))
hist(cps$wage)
hist(cps$educ)
```

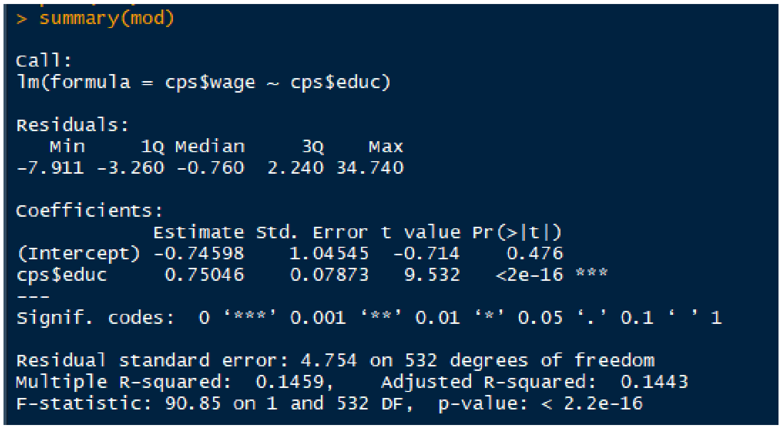


2 (b)

At first glance, it seems like the linear model could be a good start in the explanation of the relationship in our data (low p-value, but low adj R sqr)

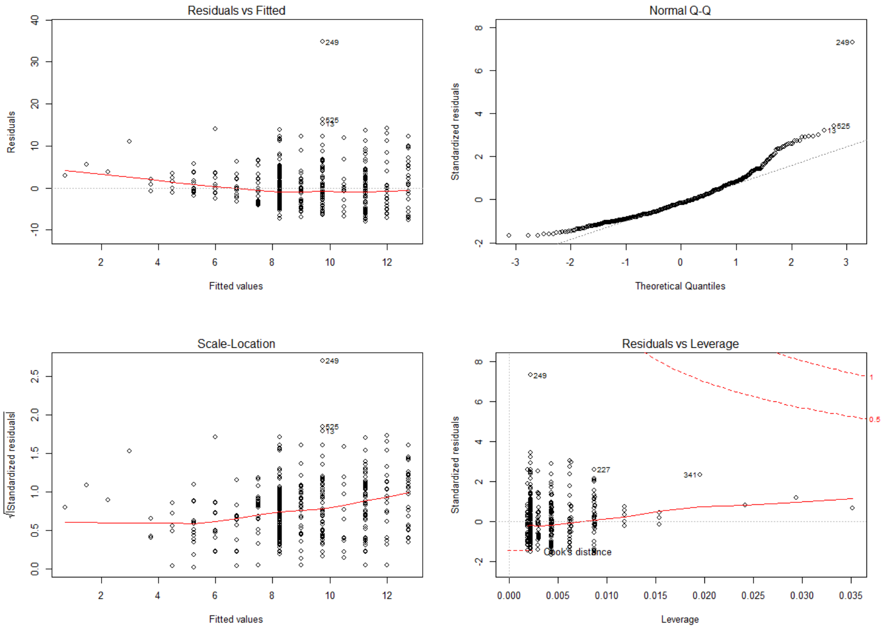


```
par(mfrow=c(2,2))
mod <- lm(cps$wage ~ cps$educ)
plot(mod)
summary(mod)
```

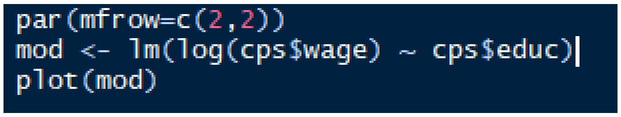


```
> summary(mod)

Call:
lm(formula = cps$wage ~ cps$educ)

Residuals:
   Min     1Q Median     3Q    Max 
-7.911 -3.260 -0.760  2.240 34.740 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.74598    1.04545  -0.714    0.476    
cps$educ     0.75046    0.07873   9.532   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.754 on 532 degrees of freedom
Multiple R-squared:  0.1459,    Adjusted R-squared:  0.1443 
F-statistic: 90.85 on 1 and 532 DF,  p-value: < 2.2e-16
```

When we examine diagnostic plots of the linear model we can see higher theoretical quantiles trail upwards on (QQ plot), thus a log transform of wage might help the residuals fall into a normal distribution.

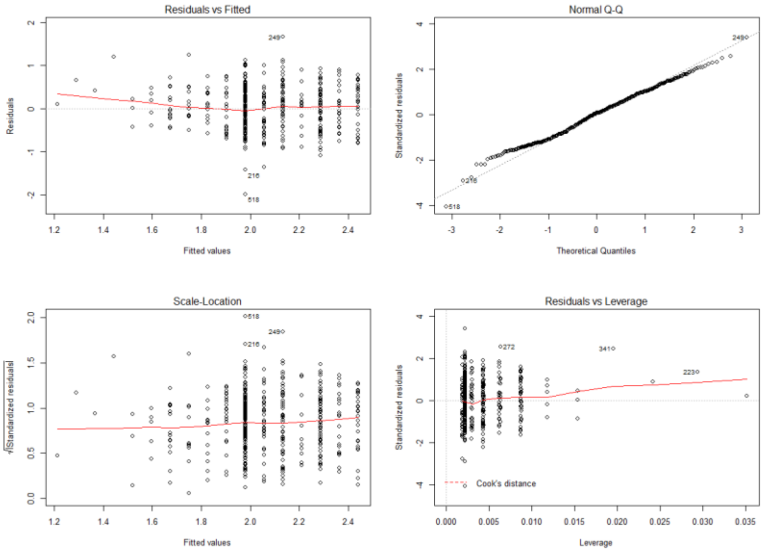


Build the Linear Model using a log transformation on wages

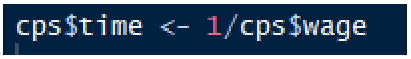


```
par(mfrow=c(2,2))
mod <- lm(log(cps$wage) ~ cps$educ)
plot(mod)
```

The residuals follow a normal distribution much better under a log transformation.

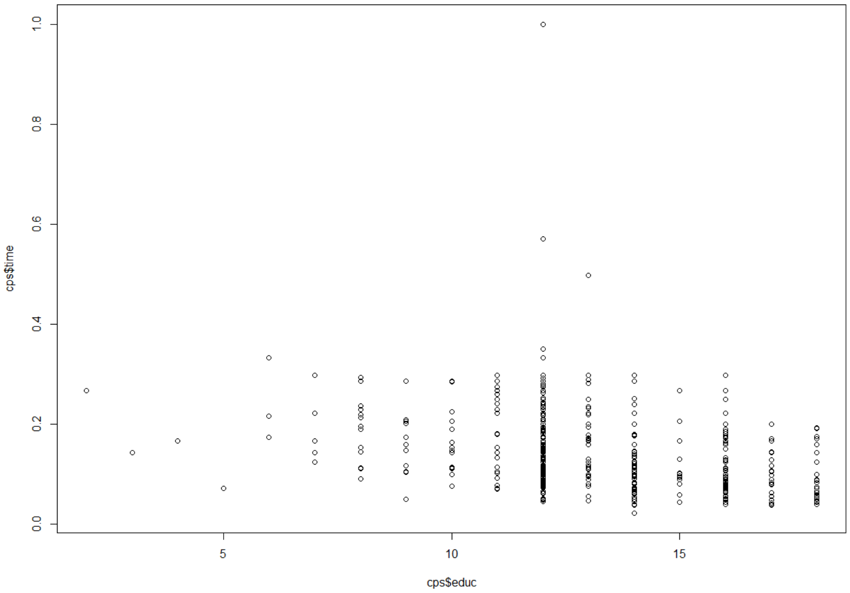


2(c) Create a new variable which is the inverse of wage. i.e. the amount of time (hours) it takes to earn a single dollar.

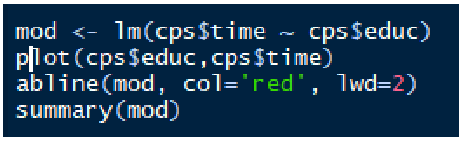


```
cps$time <- 1/cps$wage
```

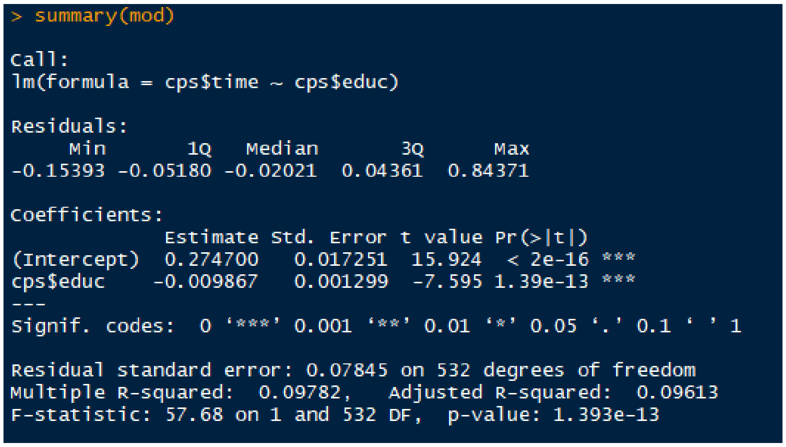
2(d) Linear regression is appropriate with these variables, since the dependent variable is continuous, and we have a numerical independent variable.



2(e) Perform linear regression:

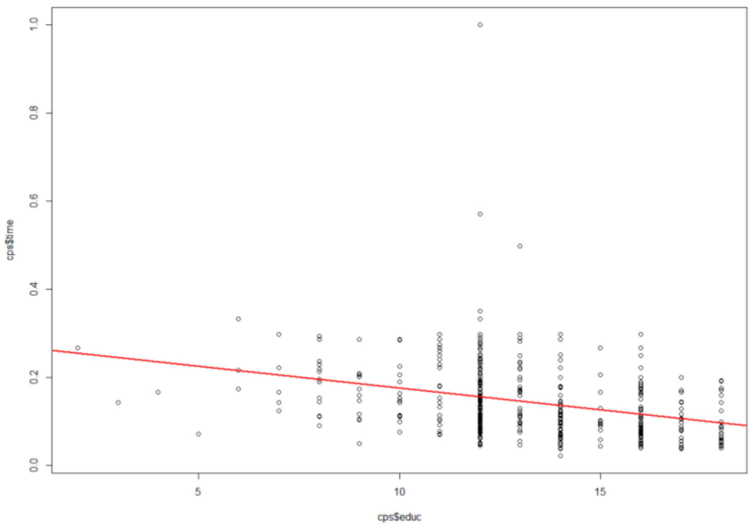


```
mod <- lm(cps$time ~ cps$educ)
plot(cps$educ,cps$time)
abline(mod, col='red', lwd=2)
summary(mod)
```

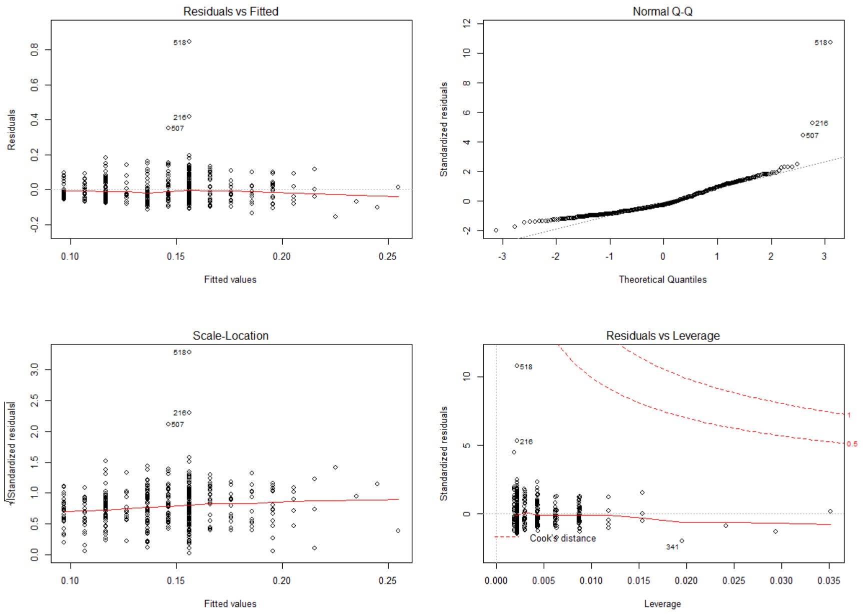


```
> summary(mod)

Call:
lm(formula = cps$time ~ cps$educ)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.15393 -0.05180 -0.02021  0.04361  0.84371 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.274700   0.017251  15.924  < 2e-16 ***
cps$educ    -0.009867   0.001299  -7.595 1.39e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.07845 on 532 degrees of freedom
Multiple R-squared:  0.09782,	Adjusted R-squared:  0.09613 
F-statistic: 57.68 on 1 and 532 DF,  p-value: 1.393e-13
```

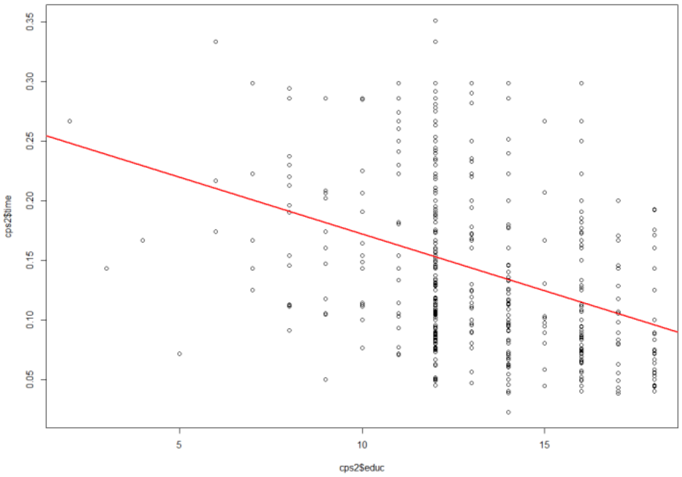
A linear model does describe the general relationship between amount of the hours it takes to earn one dollar and the years of educations. A p-value very close to zero provides enough evidence to reject the null hypothesis that the variable *time* is not related to the variable *education.* There is quite a bit of variability that the linear model does not account for though, as indicated by a low Adjusted R-squared. Generally, more years of education seem to be related to a shorter amount of time to earn one dollar. I am not exactly convinced by this data that simply attending grad school will equate to more earnings (though I am happy with my decision to attend grad school for many other reasons).

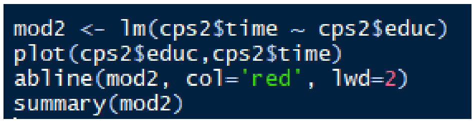


2(f) By inspection of the diagnostic plots we can see that records 216, 507 and 518 are outliers.

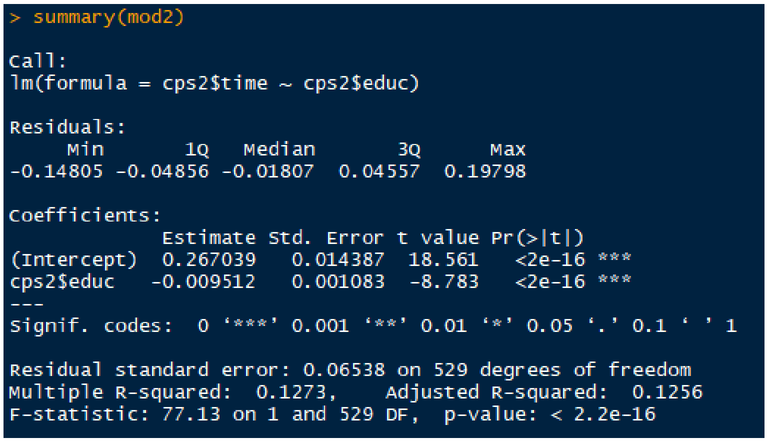


We can try removing these from the plot to see how our analysis changes.





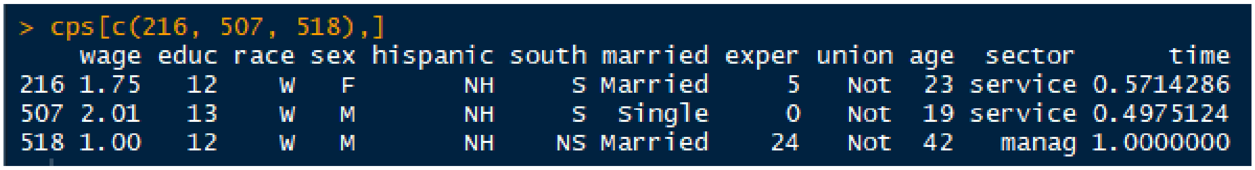
```
mod2 <- lm(cps2$time ~ cps2$educ)
plot(cps2$educ,cps2$time)
abline(mod2, col='red', lwd=2)
summary(mod2)
```



```
> summary(mod2)

Call:
lm(formula = cps2$time ~ cps2$educ)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.14805 -0.04856 -0.01807  0.04557  0.19798 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.267039   0.014387  18.561   <2e-16 ***
cps2$educ   -0.009512   0.001083  -8.783   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06538 on 529 degrees of freedom
Multiple R-squared:  0.1273,	Adjusted R-squared:  0.1256 
F-statistic: 77.13 on 1 and 529 DF,  p-value: < 2.2e-16
```

Removing the outliers further supports the relationship between years of education and time to earn a dollar. By removing the outliers from the analysis our Adjusted R-Squared increased to 0.13, which means there is still quite a bit of variance that the linear model is not explaining. The analysis doesn’t really address whether attending graduate school specifically increases wages over not attending grad school. To answer this question, I would focus the analysis more on earnings of individuals with >16 years of education. This could be framed as a decision to stop education at 16 years (i.e. bachelor degree) or continue to grad school to go >16 years of education.

I am curious to know more about the outliers that we removed:



```
> cps[c(216, 507, 518),]
    wage educ race sex hispanic south married exper union age  sector      time
216 1.75   12    W   F       NH     S Married     5   Not  23 service 0.5714286
507 2.01   13    W   M       NH     S  Single     0   Not  19 service 0.4975124
518 1.00   12    W   M       NH    NS Married    24   Not  42   manag 1.0000000
```

These were very low wage earners with 12-13 years of education, i.e. neither a bachelor nor master’s degree. While removing these outliers does improve the metrics of the linear model, it does not help address whether attending **grad school** (rather than stopping education at a bachelors degree) equates to increased earnings.