One Way Anova Test for Student's Performance

Sweta Swarupa

Contents

1 Introduction

This dataset consists of student test score data for subjects including math, reading, and writing. In this analysis I am going to conduct one-way ANOVA test and analysis of Variance Post-Hoc test to determine the impact of the categorical variables ('gender', 'race/ethnicity', 'parental level of education', 'lunch', 'test preparation course') on student's math, reading, and writing test scores.

2 Reading file

```
students <- read.csv("C:/ANOVA/StudentsPerformance.csv")
str(students)
## 'data.frame':    1000 obs. of  8 variables:
##  $ gender                     : chr  "female" "female" "female" "male" ...
##  $ race.ethnicity             : chr  "group B" "group C" "group B" "group A" ...
##  $ parental.level.of.education: chr  "bachelor's degree" "some college" "master's degree" "associate's
##  $ lunch                      : chr  "standard" "standard" "standard" "free/reduced" ...
##  $ test.preparation.course    : chr  "none" "completed" "none" "none" ...
```

```
##  $ math.score                : int  72 69 90 47 76 71 88 40 64 38 ...
##  $ reading.score             : int  72 90 95 57 78 83 95 43 64 60 ...
##  $ writing.score             : int  74 88 93 44 75 78 92 39 67 50 ...
```

There are 1000 observations and 8 variables. There are 5 categorical variables and 3 different student scores - math, reading and writing scores.

3 Checking Categorical Variables

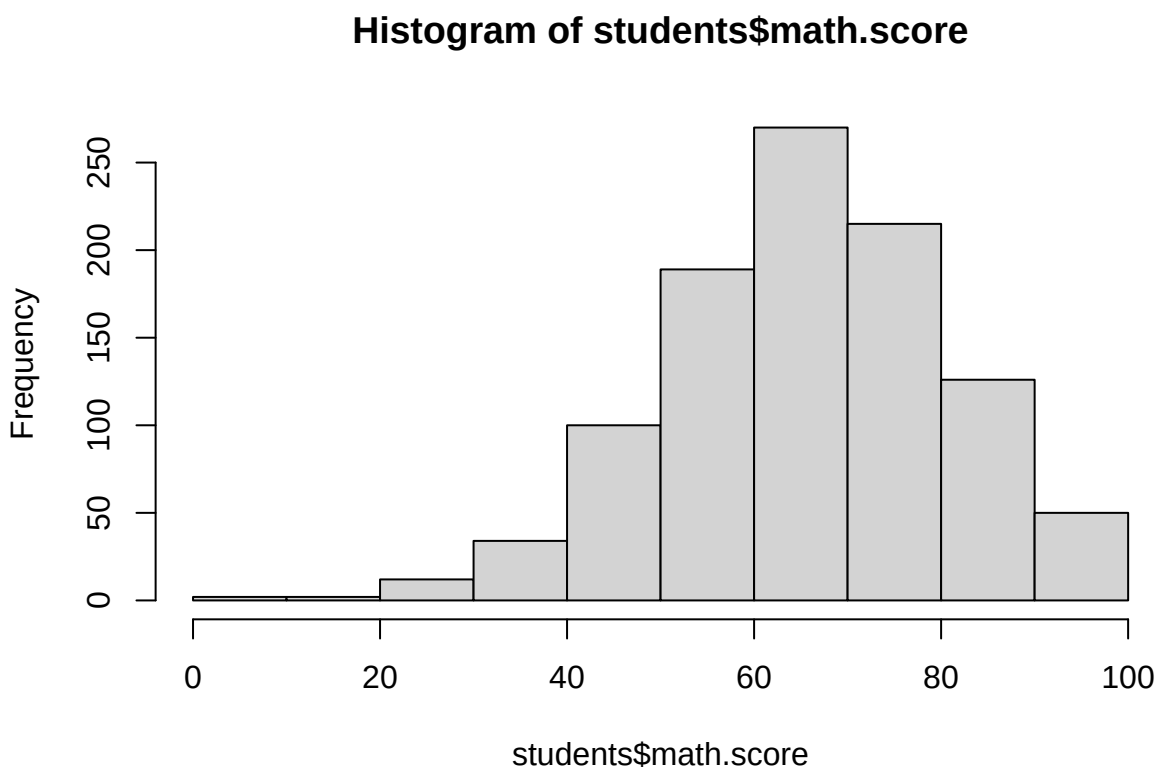
```
table(students$gender)
##
## female   male
##    518    482
table(students$race.ethnicity)
##
## group A group B group C group D group E
##      89     190     319     262     140
table(students$parental.level.of.education)
##
## associate's degree  bachelor's degree        high school    master's degree
##                222                118                196                 59
##       some college   some high school
##                226                179
table(students$lunch)
##
## free/reduced     standard
##          355          645
table(students$test.preparation.course )
##
## completed      none
##       358       642
```

One of the limitations of a one-way ANOVA is that it compares three or more than three categorical groups to establish whether there is a difference between them. Within each group there should be three or more observations to compare means of the samples.

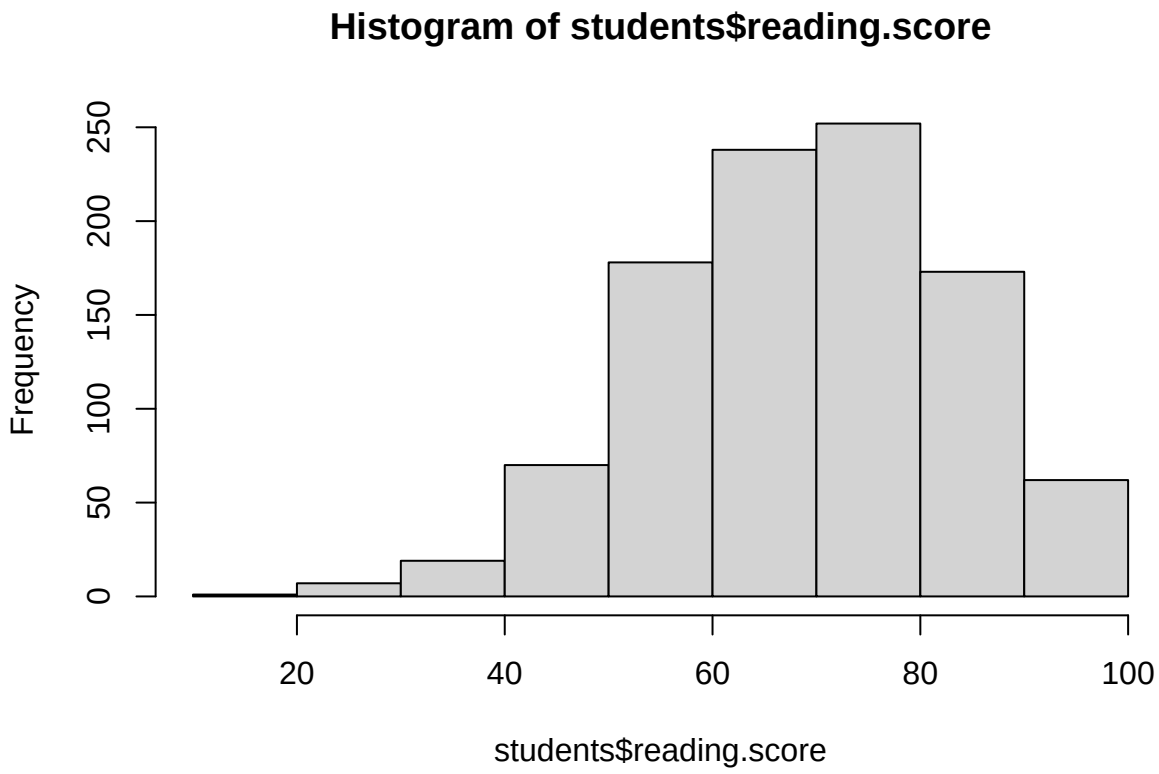
Since the variables gender, lunch and test preparation course have only 2 groups, we will be doing one-way ANOVA tests for race/ethnicity and parental level of education.

4 Ploting Histograms for Math, Reading and Writing scores

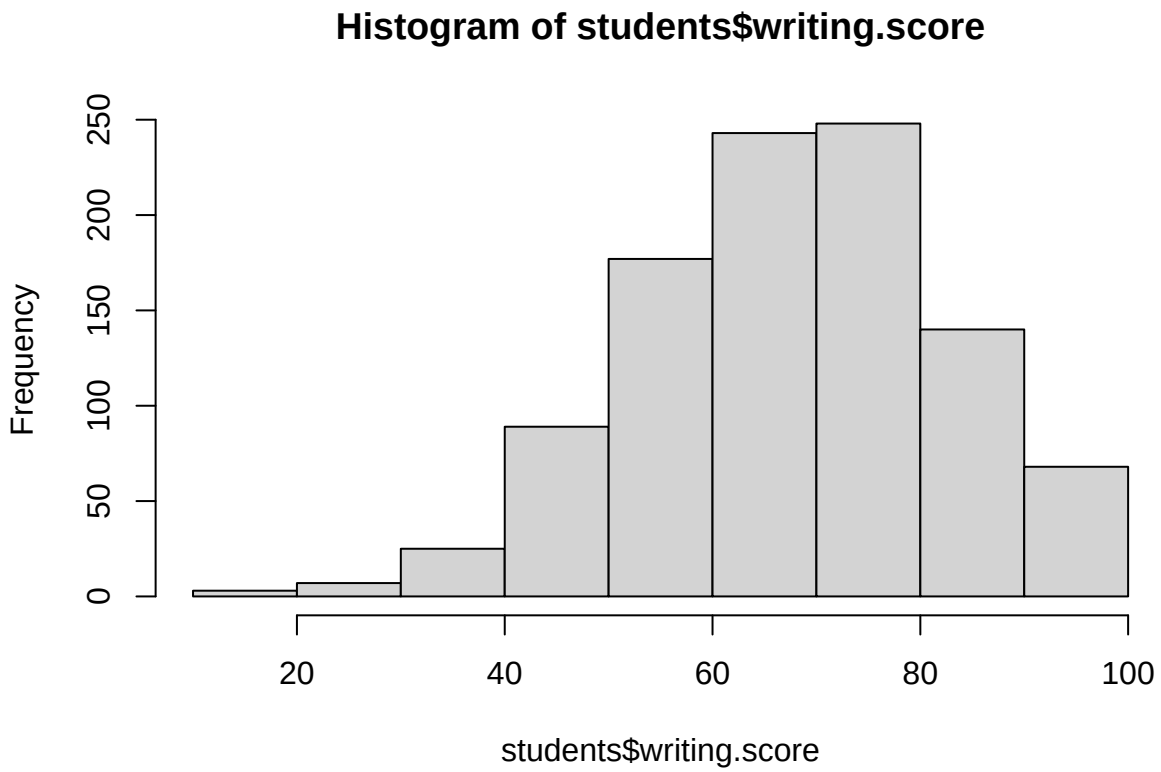
```
hist(students$math.score)
```




```
hist(students$reading.score)
```




```
hist(students$writing.score)
```



All three test scores have normal distribution.

5 One-Way ANOVA Test

One-Way ANOVA hypothesis:

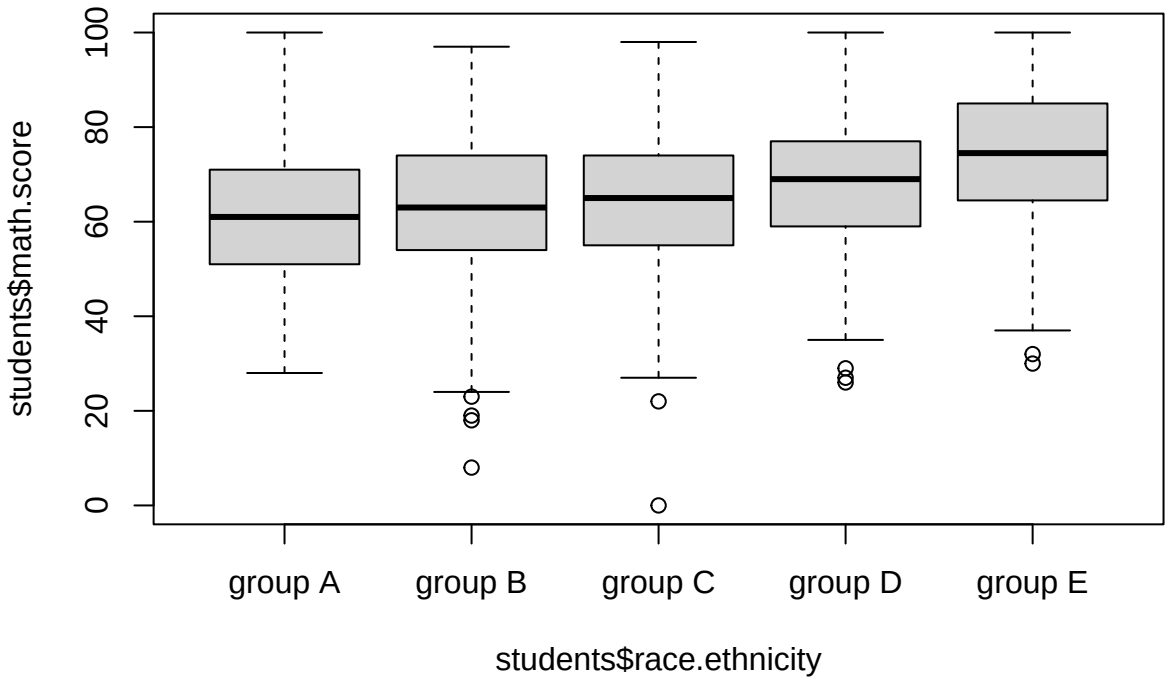
Null hypothesis (H0): There is no difference between groups and have equal means. Alternative hypothesis (H1): There is a difference between the means of three groups.

One-Way Anova assumptions:

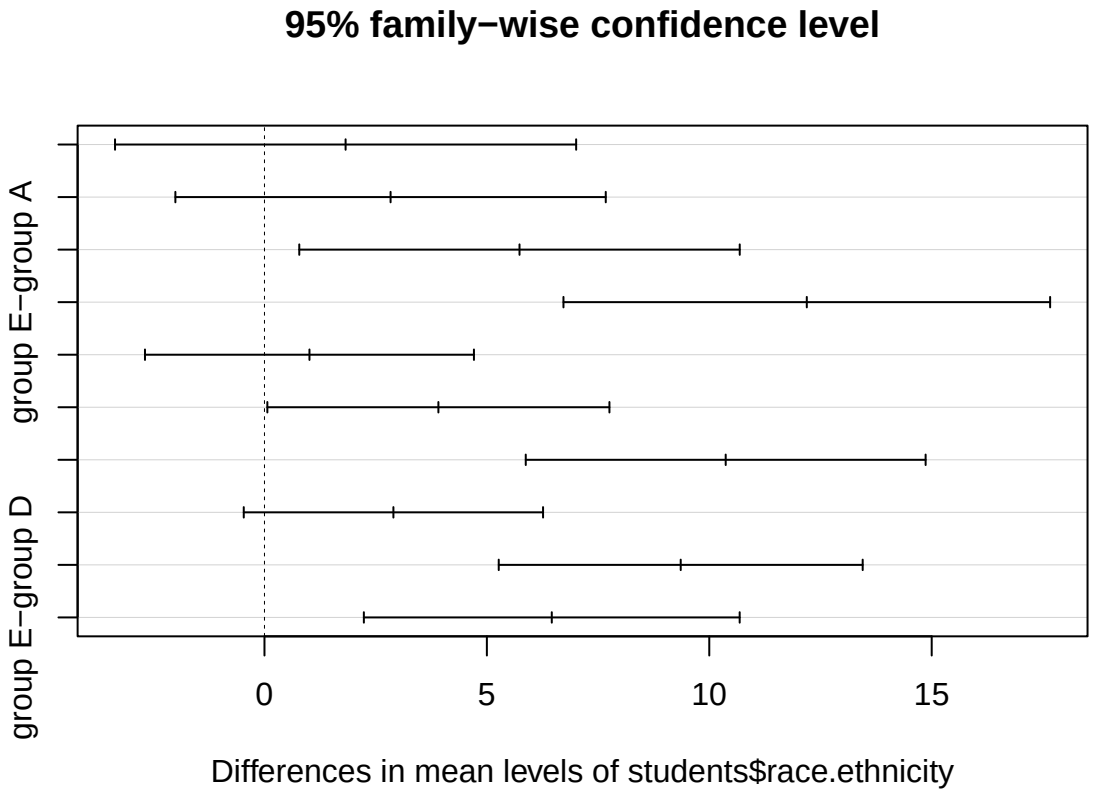
Normality: Each sample is taken from a normally distributed population Sample independence: Each sample has been drawn independently of the other samples Variance equality: The variance of data in the different groups should be the same Dependent variable: Should be continuous Hypothesis: Using a 95% confidence internal

5.1 Anova for Ethnicity on Math Score

```
boxplot(students$math.score ~ students$race.ethnicity, data = students)
```



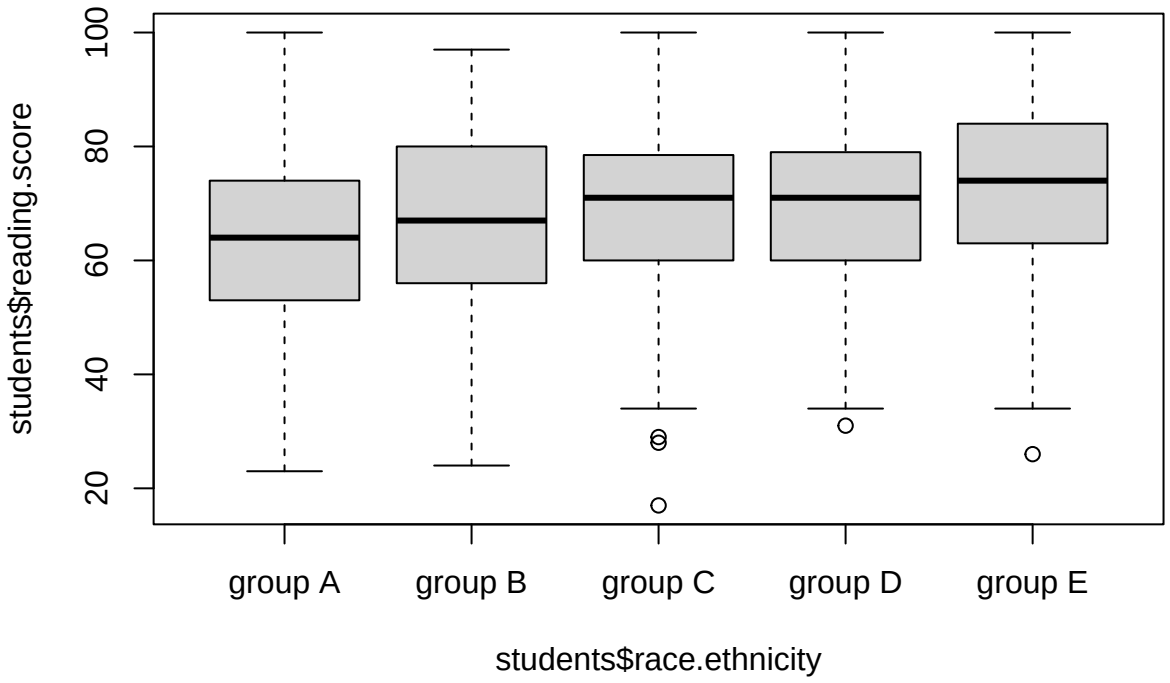

```
anova.em <- aov(students$math.score ~ students$race.ethnicity, data = students)
summary(anova.em)
##                          Df Sum Sq Mean Sq F value   Pr(>F)
## students$race.ethnicity   4  12729    3182   14.59 1.37e-11 ***
## Residuals               995 216960     218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(anova.em)
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = students$math.score ~ students$race.ethnicity, data = students)
##
## $`students$race.ethnicity`
##                       diff         lwr       upr     p adj
## group B-group A  1.823418 -3.35997818  7.006814 0.8723586
## group C-group A  2.834736 -2.00279565  7.672268 0.4968040
## group D-group A  5.733382  0.78239222 10.684372 0.0138238
## group E-group A 12.192215  6.72151591 17.662914 0.0000000
## group C-group B  1.011318 -2.68671543  4.709352 0.9451894
## group D-group B  3.909964  0.06470228  7.755225 0.0440476
## group E-group B 10.368797  5.87410158 14.863492 0.0000000
## group D-group C  2.898646 -0.46589828  6.263189 0.1289617
## group E-group C  9.357479  5.26646348 13.448494 0.0000000
## group E-group D  6.458833  2.23426347 10.683403 0.0003084
plot(TukeyHSD(anova.em))
```



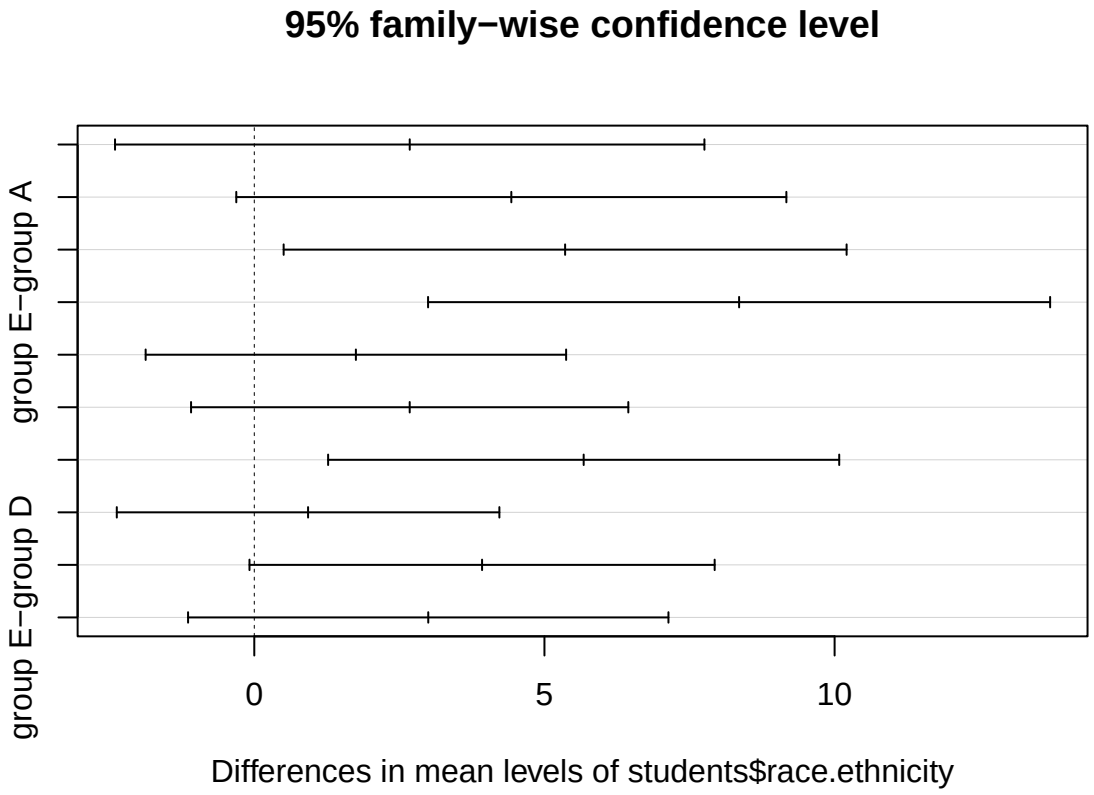
By just looking at the box plot, we can tell that median of group E is higher than other groups. The p value is lower than .001. Hence, we can reject the null hypothesis. The tukey score also shows that the means are very different for group E and A, Group E and B, group E and C.

5.2 Anova for Ethnicity on Reading Score

```
boxplot(students$reading.score ~ students$race.ethnicity, data = students)
```



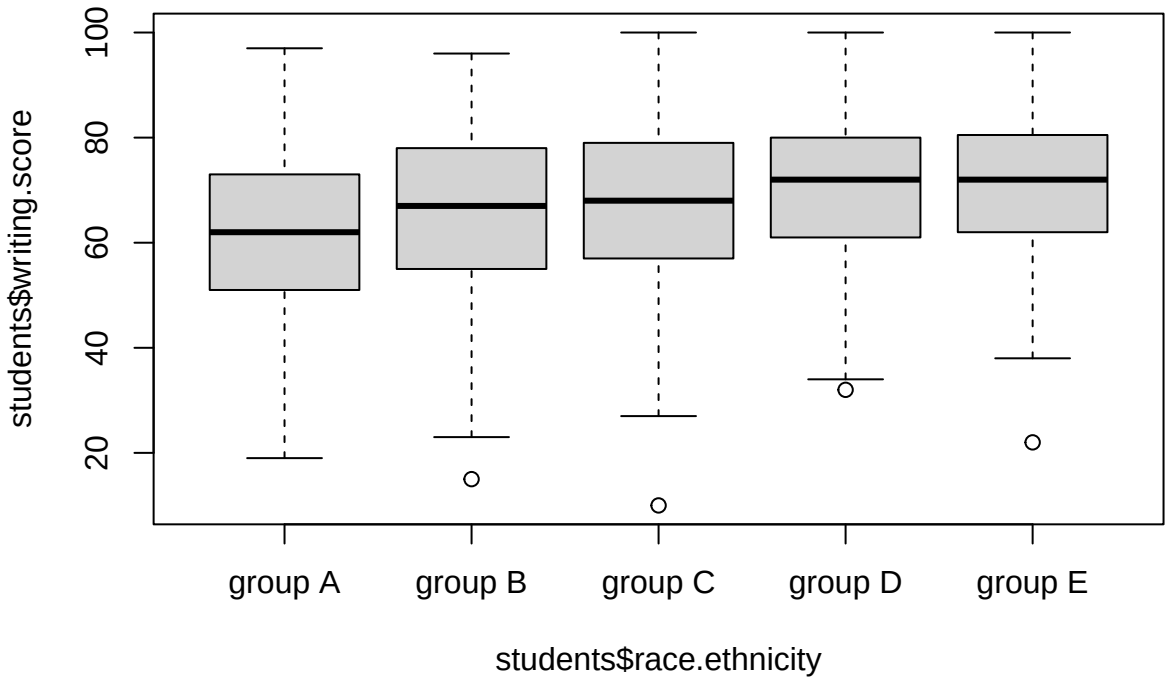

```
anova.er <- aov(students$reading.score ~ students$race.ethnicity, data = students)
summary(anova.er)
##                          Df Sum Sq Mean Sq F value   Pr(>F)
## students$race.ethnicity   4   4706  1176.6   5.622 0.000178 ***
## Residuals               995 208246   209.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(anova.er)
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = students$reading.score ~ students$race.ethnicity, data = students)
##
## $`students$race.ethnicity`
##                         diff         lwr       upr     p adj
## group B-group A 2.6784743 -2.39976087  7.756709 0.6009584
## group C-group A 4.4292910 -0.31009682  9.168679 0.0799351
## group D-group A 5.3563770  0.50583339 10.206921 0.0219169
## group E-group A 8.3544141  2.99470490 13.714123 0.0002170
## group C-group B 1.7508167 -1.87219101  5.373824 0.6784186
## group D-group B 2.6779028 -1.08934583  6.445151 0.2954782
## group E-group B 5.6759398  1.27243316 10.079447 0.0040770
## group D-group C 0.9270861 -2.36919766  4.223370 0.9395630
## group E-group C 3.9251232 -0.08289327  7.933140 0.0582314
## group E-group D 2.9980371 -1.14082421  7.136898 0.2767422
plot(TukeyHSD(anova.er))
```



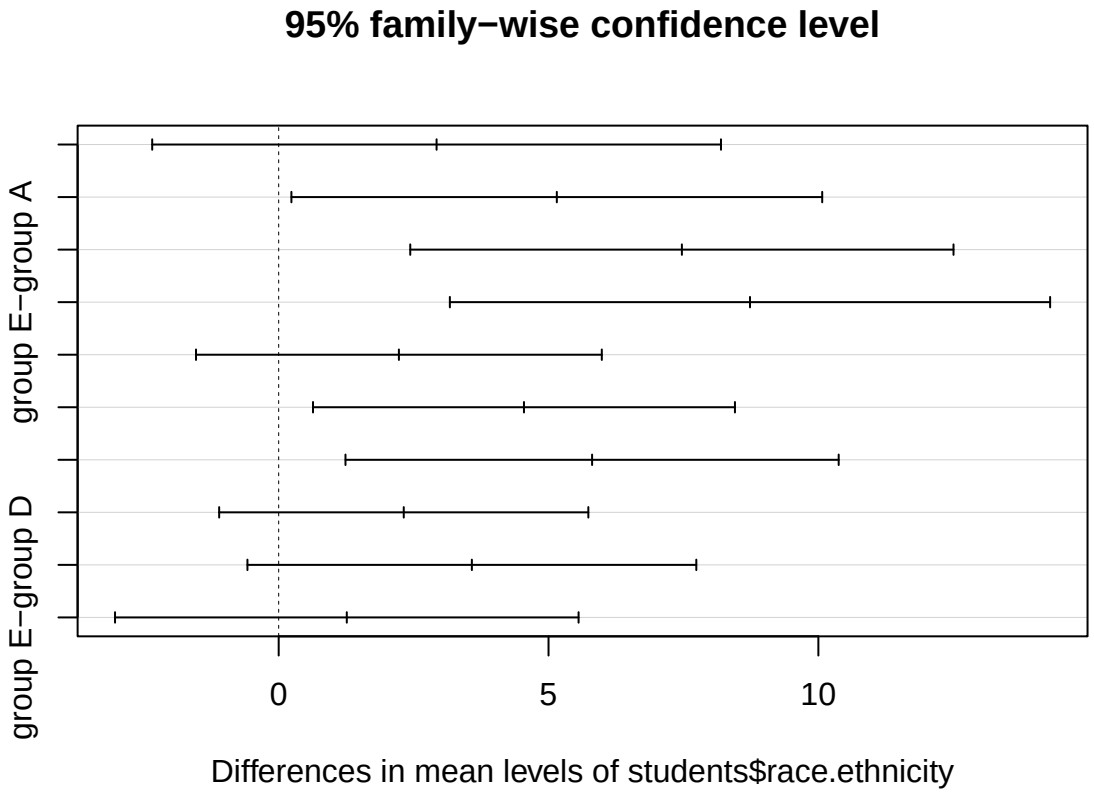
Similar to math score, p value of reading score is also less that .001. The difference in mean between the groups are less for reading compared to maths. The means are very different for group E and A, Group E and B, group D and A. Group C and D looks very similar to each other.

5.3 Anova for Ethnicity on Writing Score

```
boxplot(students$writing.score ~ students$race.ethnicity, data = students)
```



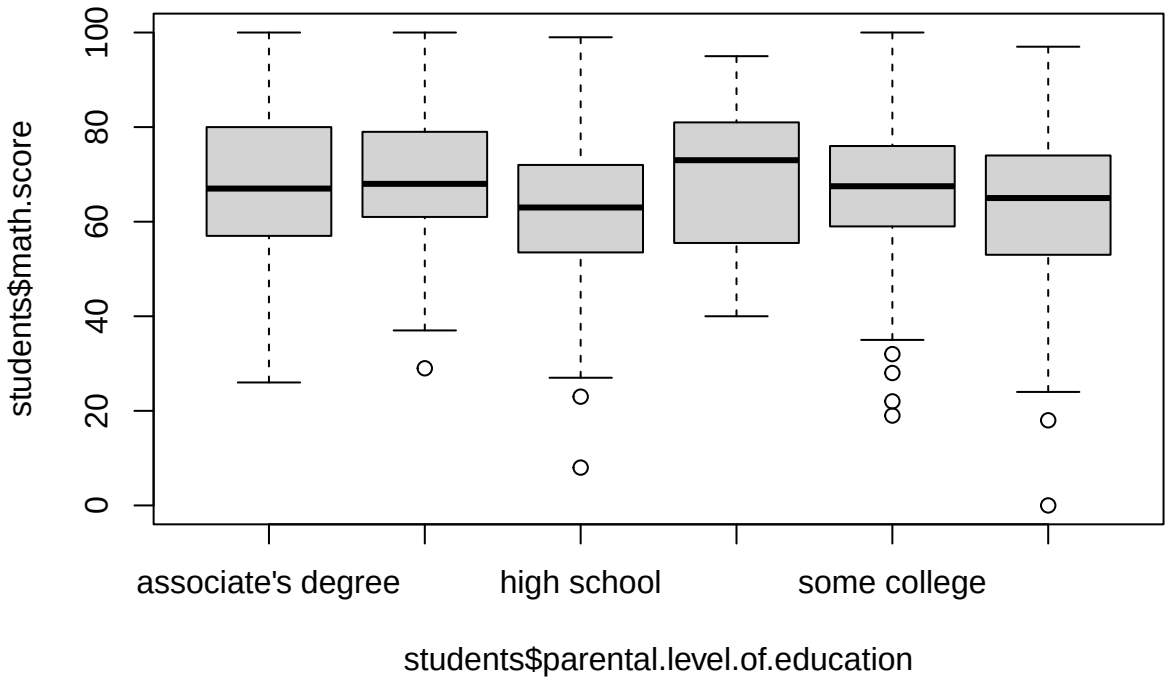

```
anova.ew <- aov(students$writing.score ~ students$race.ethnicity, data = students)
summary(anova.ew)
##                          Df Sum Sq Mean Sq F value  Pr(>F)
## students$race.ethnicity   4   6456  1614.0   7.162 1.1e-05 ***
## Residuals               995 224221   225.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(anova.ew)
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = students$writing.score ~ students$race.ethnicity, data = students)
##
## $`students$race.ethnicity`
##                       diff        lwr       upr     p adj
## group B-group A 2.925843 -2.3435724  8.195258 0.5513495
## group C-group A 5.153429  0.2356178 10.071240 0.0346280
## group D-group A 7.470881  2.4377292 12.504033 0.0005145
## group E-group A 8.732986  3.1714998 14.294471 0.0001892
## group C-group B 2.227586 -1.5318166  5.986989 0.4853085
## group D-group B 4.545038  0.6359643  8.454112 0.0132671
## group E-group B 5.807143  1.2378577 10.376428 0.0048638
## group D-group C 2.317452 -1.1029267  5.737831 0.3445476
## group E-group C 3.579557 -0.5793492  7.738463 0.1296283
## group E-group D 1.262105 -3.0325720  5.556781 0.9296838
plot(TukeyHSD(anova.ew))
```



The p value is less than .001. Group B and C looks very similar. Group D and E looks very similar. Mean values between group E and A and group D and A is very high.

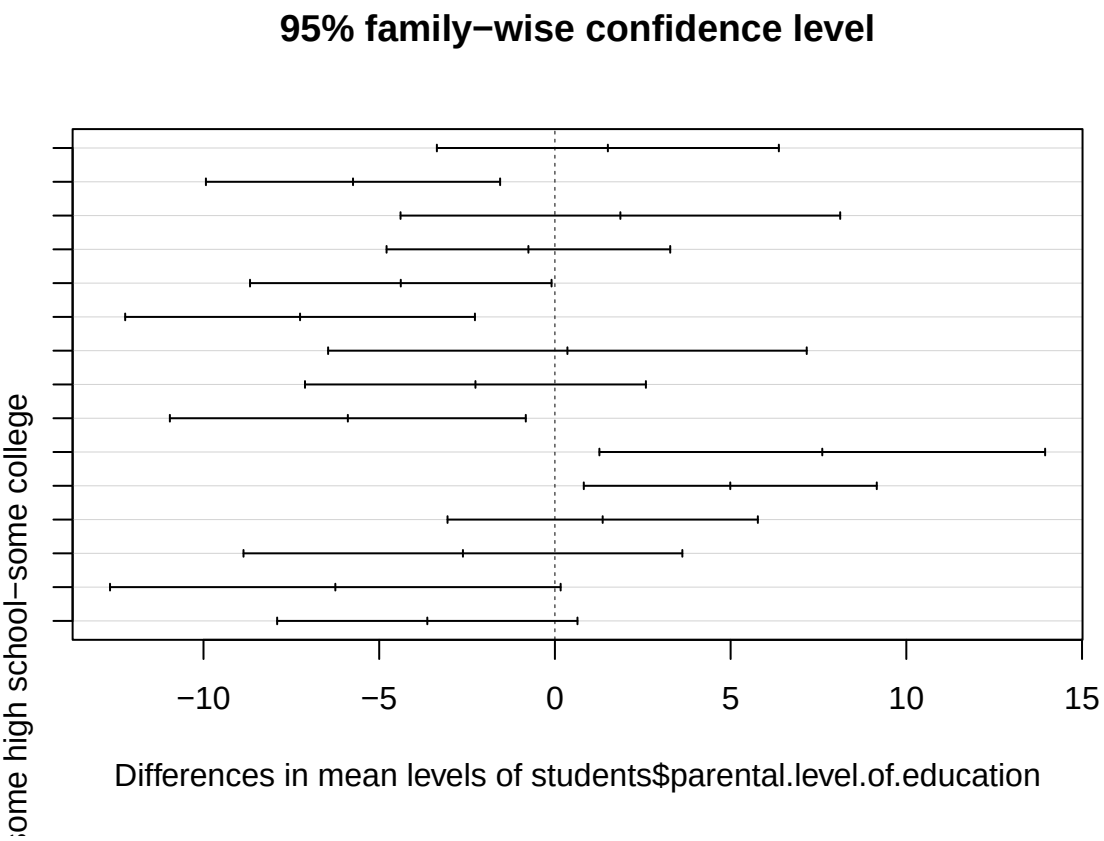
5.4 Anova for Parental Education on Math Score

```
boxplot(students$math.score ~ students$parental.level.of.education, data = students)
```




```
anova.pm <- aov(students$math.score ~ students$parental.level.of.education, data = students)
summary(anova.pm)
##                                   Df Sum Sq Mean Sq F value   Pr(>F)
## students$parental.level.of.education   5   7296  1459.1   6.522 5.59e-06 ***
## Residuals                            994 222394   223.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(anova.pm)
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = students$math.score ~ students$parental.level.of.education, data = students)
##
## $`students$parental.level.of.education`
##                                           diff        lwr         upr
## bachelor's degree-associate's degree  1.5069476  -3.358681  6.37257671
## high school-associate's degree       -5.7451278  -9.931142 -1.55911404
## master's degree-associate's degree    1.8628798  -4.392702  8.11846150
## some college-associate's degree      -0.7545643  -4.790324  3.28119570
## some high school-associate's degree  -4.3856762  -8.675962 -0.09539047
## high school-bachelor's degree        -7.2520754 -12.228447 -2.27570365
## master's degree-bachelor's degree     0.3559322  -6.453904  7.16576824
## some college-bachelor's degree       -2.2615119  -7.112174  2.58915025
## some high school-bachelor's degree   -5.8926238 -10.957021 -0.82822687
## master's degree-high school           7.6080076   1.265908 13.95010753
## some college-high school              4.9905635   0.821956  9.15917095
## some high school-high school          1.3594516  -3.056030  5.77493356
```

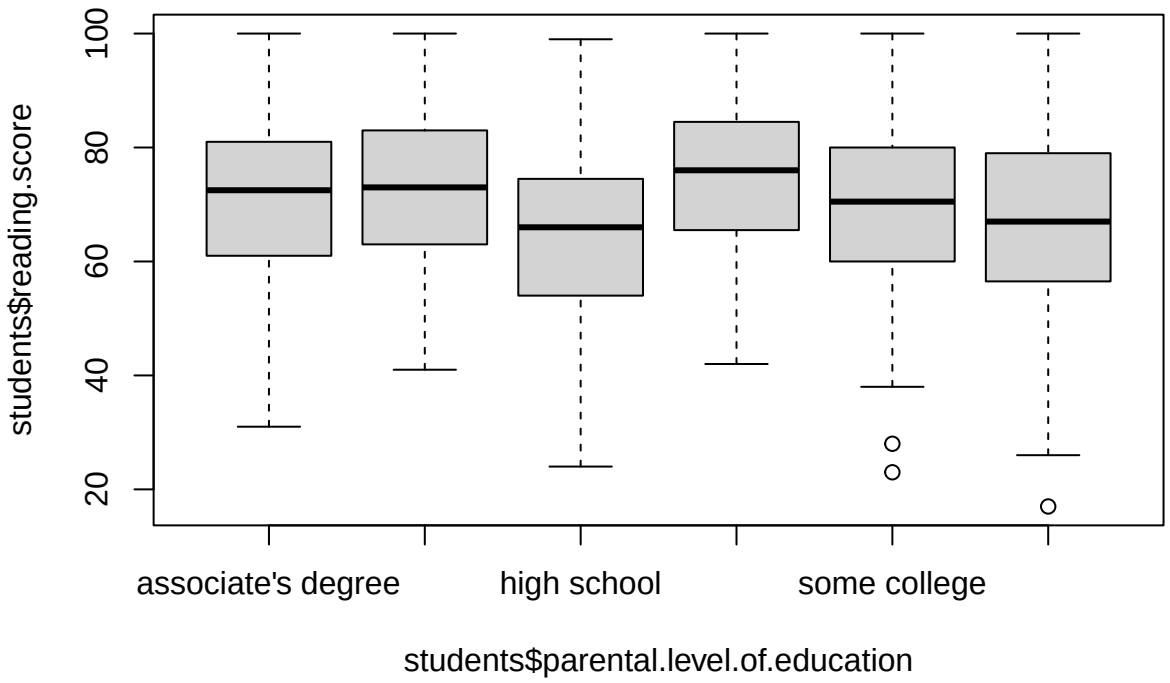
```
## some college-master's degree          -2.6174441  -8.861392  3.62650328
## some high school-master's degree      -6.2485560 -12.659958  0.16284570
## some high school-some college         -3.6311119  -7.904416  0.64219230
##                                           p adj
## bachelor's degree-associate's degree 0.9502834
## high school-associate's degree       0.0013308
## master's degree-associate's degree   0.9578456
## some college-associate's degree      0.9947937
## some high school-associate's degree  0.0417546
## high school-bachelor's degree        0.0004918
## master's degree-bachelor's degree    0.9999897
## some college-bachelor's degree       0.7676188
## some high school-bachelor's degree   0.0118857
## master's degree-high school          0.0083719
## some college-high school             0.0085748
## some high school-high school         0.9514996
## some college-master's degree         0.8384321
## some high school-master's degree     0.0610615
## some high school-some college        0.1481784
plot(TukeyHSD(anova.pm))
```



By just looking at the box plot we can say that the median for parents with high school education is lower than parents with higher education. The p value is less than .001 so we can reject the null hypothesis. The tukey score also shows that the mean between high school-bachelor's degree, some high school-master's degree, some high school-bachelor's degree and high school-associate's degree is very high.

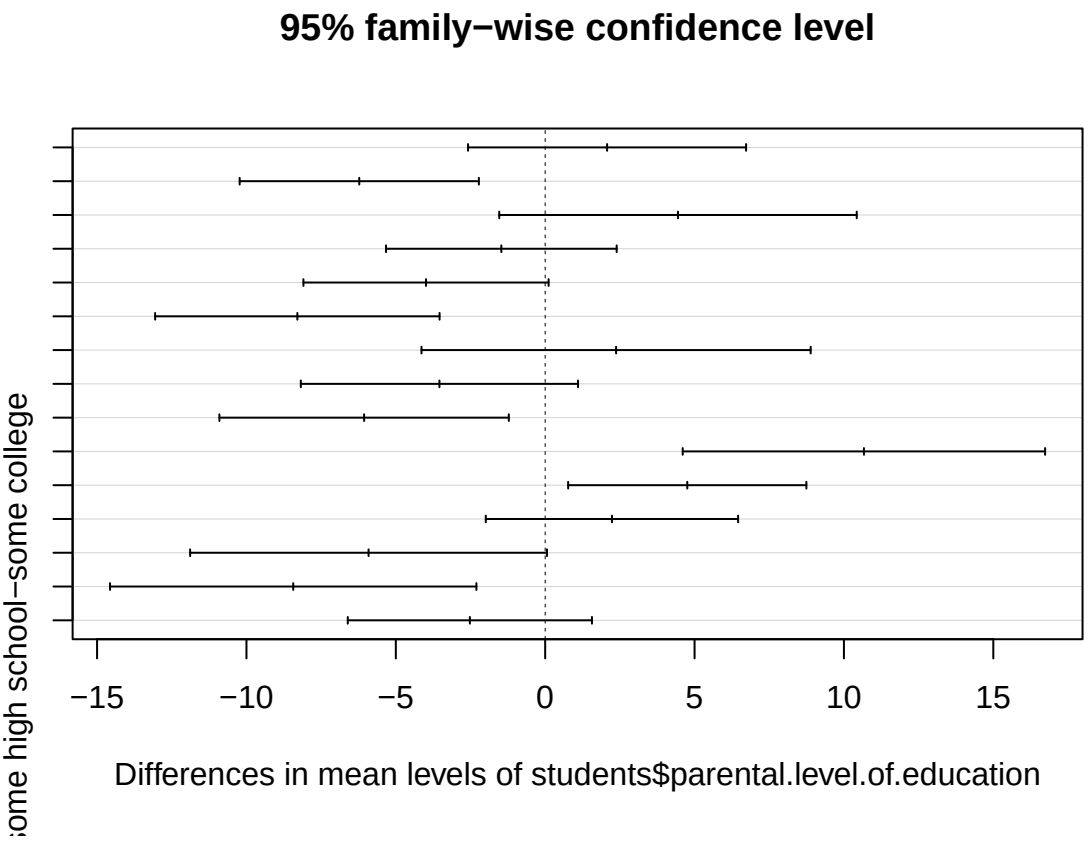
5.5 Anova for Parental Education on Reading Score

```
boxplot(students$reading.score ~ students$parental.level.of.education, data = students)
```




```
anova.pr <- aov(students$reading.score ~ students$parental.level.of.education, data = students)
summary(anova.pr)
##                                    Df Sum Sq Mean Sq F value   Pr(>F)
## students$parental.level.of.education   5   9506  1901.3   9.289 1.17e-08 ***
## Residuals                            994 203446   204.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(anova.pr)
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = students$reading.score ~ students$parental.level.of.education, data = students)
##
## $`students$parental.level.of.education`
##                                         diff         lwr         upr
## bachelor's degree-associate's degree  2.072072  -2.5816716  6.72581573
## high school-associate's degree       -6.223846 -10.2275701 -2.22012251
## master's degree-associate's degree    4.444953  -1.5382140 10.42812087
## some college-associate's degree      -1.467751  -5.3277641  2.39226226
## some high school-associate's degree  -3.989380  -8.0928354  0.11407454
## high school-bachelor's degree        -8.295918 -13.0555821 -3.53625460
## master's degree-bachelor's degree     2.372881  -4.1404041  8.88616683
```

```
## some college-bachelor's degree       -3.539823  -8.1792515  1.09960551
## some high school-bachelor's degree   -6.061453 -10.9053082 -1.21759683
## master's degree-high school          10.668800   4.6028817 16.73471777
## some college-high school              4.756095   0.7690199  8.74317086
## some high school-high school          2.234466  -1.9887334  6.45766511
## some college-master's degree         -5.912704 -11.8847442  0.05933545
## some high school-master's degree     -8.434334 -14.5665358 -2.30213195
## some high school-some college        -2.521630  -6.6088425  1.56558345
##                                          p adj
## bachelor's degree-associate's degree 0.8006901
## high school-associate's degree       0.0001463
## master's degree-associate's degree   0.2770595
## some college-associate's degree      0.8871557
## some high school-associate's degree  0.0622049
## high school-bachelor's degree        0.0000113
## master's degree-bachelor's degree    0.9042540
## some college-bachelor's degree       0.2486973
## some high school-bachelor's degree   0.0049602
## master's degree-high school          0.0000090
## some college-high school             0.0089453
## some high school-high school         0.6574517
## some college-master's degree         0.0541073
## some high school-master's degree     0.0012867
## some high school-some college        0.4912798
plot(TukeyHSD(anova.pr))
```

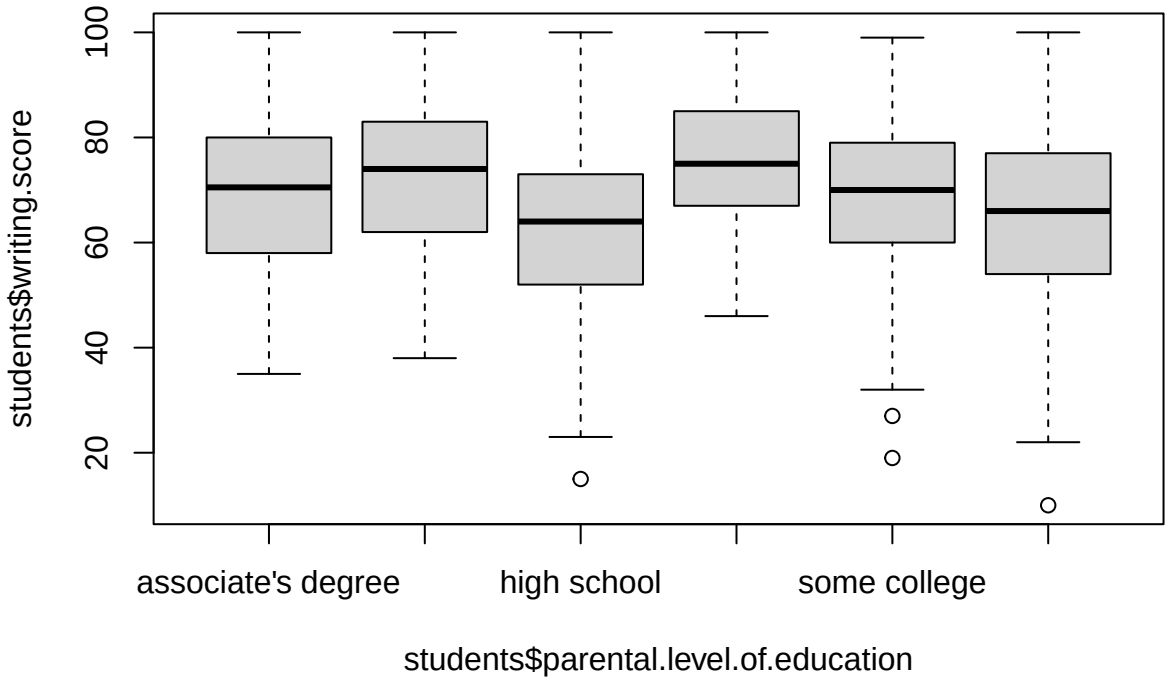


Similar to math score, p value for reading score is less than .001, so we can reject the null hypothesis.The tukey score shows that the mean between master's degree-high school, some high school-master's degree, high school-bachelor's

degree, high school-associate's degree is very high.

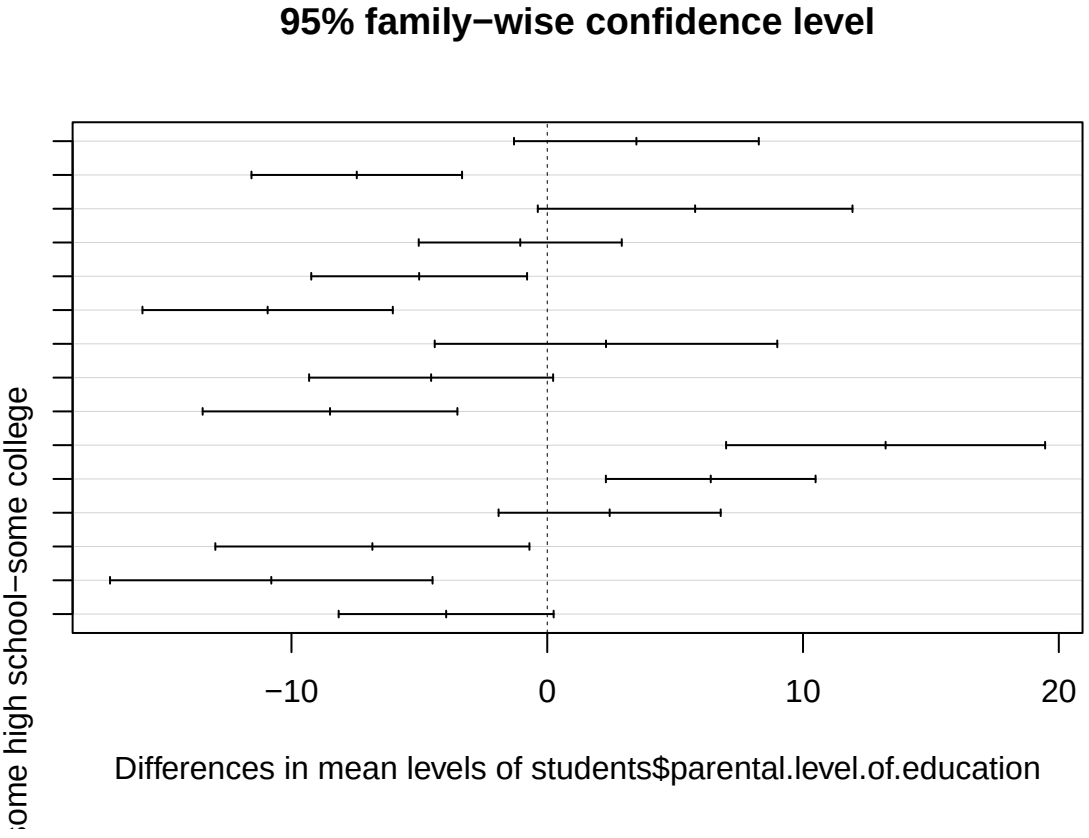
5.6 Anova for Parental Education on Writing Score

```
boxplot(students$writing.score ~ students$parental.level.of.education, data = students)
```




```
anova.pw <- aov(students$writing.score ~ students$parental.level.of.education, data = students)
summary(anova.pw)
##                                      Df Sum Sq Mean Sq F value   Pr(>F)
## students$parental.level.of.education   5  15623  3124.6   14.44 1.12e-13 ***
## Residuals                            994 215054   216.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(anova.pw)
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = students$writing.score ~ students$parental.level.of.education, data = students)
##
## $`students$parental.level.of.education`
##                                           diff         lwr        upr
## bachelor's degree-associate's degree   3.484960  -1.2997057  8.2696248
## high school-associate's degree        -7.447417 -11.5637755 -3.3310582
## master's degree-associate's degree     5.781570  -0.3699194 11.9330588
## some college-associate's degree       -1.055688  -5.0242936  2.9129167
```

```
## some high school-associate's degree   -5.008128  -9.2270238 -0.7892327
## high school-bachelor's degree        -10.932376 -15.8259415 -6.0388112
## master's degree-bachelor's degree      2.296610  -4.3999105  8.9931308
## some college-bachelor's degree        -4.540648  -9.3105953  0.2292994
## some high school-bachelor's degree    -8.493088 -13.4732134 -3.5129622
## master's degree-high school           13.228987   6.9924189 19.4655542
## some college-high school               6.391728   2.2924864 10.4909704
## some high school-high school           2.439289  -1.9027200  6.7812971
## some college-master's degree          -6.837258 -12.9773065 -0.6972098
## some high school-master's degree     -10.789698 -17.0944142 -4.4849817
## some high school-some college         -3.952440  -8.1546364  0.2497568
##                                          p adj
## bachelor's degree-associate's degree 0.2987656
## high school-associate's degree       0.0000043
## master's degree-associate's degree   0.0794141
## some college-associate's degree      0.9740854
## some high school-associate's degree  0.0094688
## high school-bachelor's degree        0.0000000
## master's degree-bachelor's degree    0.9245528
## some college-bachelor's degree       0.0725881
## some high school-bachelor's degree   0.0000192
## master's degree-high school          0.0000000
## some college-high school             0.0001376
## some high school-high school         0.5960379
## some college-master's degree         0.0189417
## some high school-master's degree     0.0000177
## some high school-some college        0.0790042
plot(TukeyHSD(anova.pw))
```



Again p value is less than .001. There is huge difference in the mean values for different groups.

6 Discussion

Race/Ethnicity and Parental level of Education was statistically tested against the exam scores using a 1-Way ANOVA test. This test allows us to accurately confirm that both Race/Ethnicity and Parental level of Education has an impact on students test scores. Using a 95% confidence interval, we achieved p-values less than 0.001 for each category of data. This allows us to reject our null hypothesis and summarize that the both the categorical data has a significant impact on the reading, writing, and math scores.