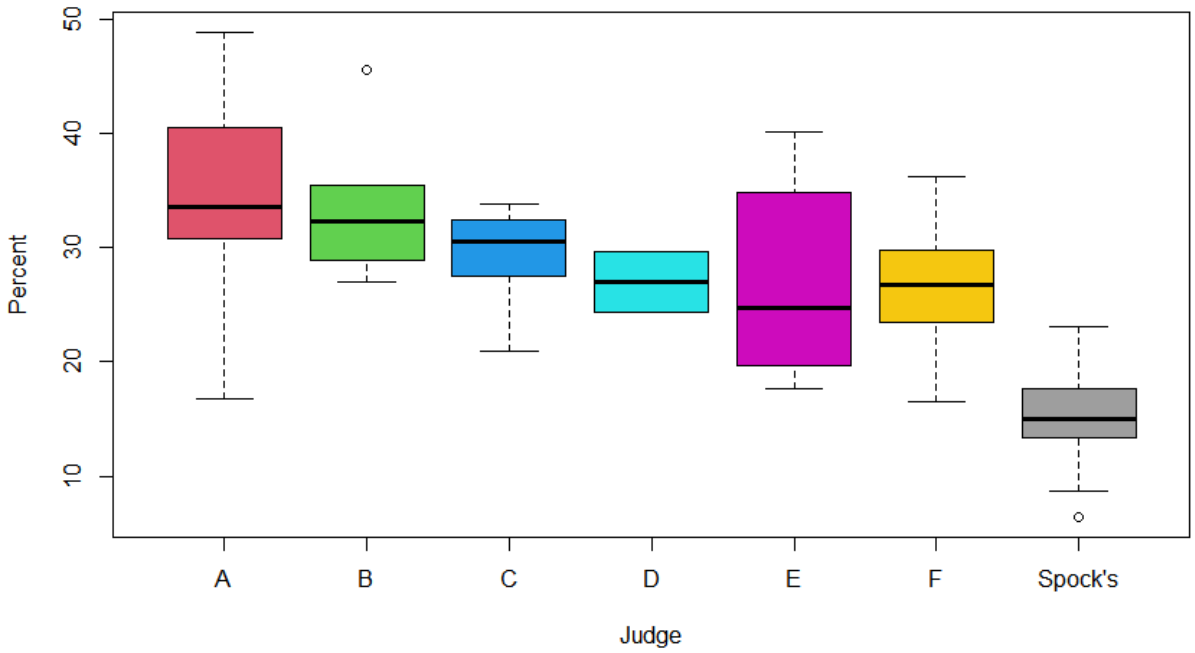
13. The present exercise is a case study on 300 juries selected at random for potential jury duty in Boston area. The actual case was on Benjamin Spock who was charged because of violating an Act by encouraging young men were to resist getting into military services for Vietnam. The defense in this case challenged jurors that there were no women in Spock jury. The data consists of 30 or more venires out of 300, selected on random, keeping the Spock venire nonrandom to analyze if there is really a difference in the Spock jury women with others jury women. Lets head to data for clear demonstration

```
> head(case0502)
  Percent   Judge
1     6.4 Spock's
2     8.7 Spock's
3    13.3 Spock's
4    13.6 Spock's
5    15.0 Spock's
6    15.2 Spock's
> table(case0502$Judge)

    A     B     C     D     E     F Spock's
    5     6     9     2     6     9     9
```

The above details give the details of the sample set containing the percentage of women in each venire. The Spock venire consist of 9 records where as F venire with 9 records, E venire with 6 records, D venire with 2 records, C venire with 9 records, B venire with 6 and A venire with 5, with a total record of 46 samples.

To get a big picture on the data, let’s make a boxplot

```
> boxplot(Percent ~ Judge, data = case0502, col = 2:8)
```



The boxplot shows that the Spock's venire is below the all other remaining venire's women percent. With this knowledge, let's get into the problem cases

a. To get the average percent of women from all 46 venires, let's calculate the mean of all the 6 group of venires. Below is the code to get individual means

```
> tapply(case0502$Percent, case0502$Judge, mean)
       A        B        C        D        E        F  Spock's
34.12000 33.61667 29.10000 27.00000 26.96667 26.80000 14.62222
```

By the above result, we can clearly distinguish that the percent of women in Spock's venire is low with a mean on 14.62% which is more below when compared to all other venire's women percent

The overall mean for all 46 venires is obtained by the below code

```
> mean(case0502$Percent)
[1] 26.58261
```

b. Let's see how many individual women percent records in Spock's venire is less than grand average from all 46 venires

From the above result obtained in a, the average percent of women from all 46 venires is 26.58. Below is the code to find how many records women percent in Spock's venire is less than this grand mean

```
> table(case0502$P[case0502$Judge == "Spock's"] < mean(case0502$Percent))


TRUE
   9
```

From the above we can say that all the 9 records in the Spock's venire is less than the grand average, which can be considered as one of the statistical evidences obtained from the case study

c. Now let's find the average of women percent in Spock's venire and see how many records in Spock's venire is less than this grand percent

```
> table(case0502$P[case0502$Judge == "Spock's"] < mean(case0502$P[case0502$Judge == "Spock's"]))


FALSE  TRUE
    5     4
```

From the above result, it is obtained that out of 9 records, 5 are above the average of Spock's venire and 4 in Spock's venire are less than the average women percent of Spock's venire

5.21

For the Spock's case study which was discussed under problem 5.13, before going into the details, in connection to the above obtained boxplot, let's get into the summary details

```
> tapply(case0502$Percent, case0502$Judge, summary)
$A
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.80   30.80   33.60   34.12   40.50   48.90
$B
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.00   29.68   32.35   33.62   34.80   45.60
$C
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   21.0    27.5    30.5    29.1    32.5    33.8
```

```
$D
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  24.30   25.65   27.00   27.00   28.35   29.70
$E
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  17.70   20.15   24.70   26.97   33.08   40.20
$F
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   16.5    23.5    26.7    26.8    29.8    36.2
$`Spock's`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.40   13.30   15.00   14.62   17.70   23.10
```

The summary details give that the means and medians are varied to all the venires but let’s find out if there is any real difference between the spread in the groups.

Let’s conduct a Levene’s test which is a robust test to pool variances in the population. This procedure obtains the usual one-way analysis of F-test on absolute values of the differences of observations from the group medians. The code and result of Levene’s test are as below

```
> leveneTest(Percent ~ Judge, data = case0502)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  6  1.2625 0.2969
      39
```

The test results give that after comparing the 46 records of women percent for 7 venires gives that there is no strong evidence with p value 0.297 that the spread of the data differs between the groups and gives no evidence that the variances are unequal.

5.18

This problem is based on a randomized experiment to estimate the effect of a fatty acid CPFA on the level of a protein in rat livers. The data consists of recording only one level of CPFA, taken in a day and control group of no CPFA was investigated each day along with another levels. Let's head to the data records

```
> head(ex0518)
  Protein Treatment  Day TrtDayGroup
1     154    CPFA50 Day1      Group1
2     177    CPFA50 Day1      Group1
3     174    CPFA50 Day1      Group1
4     164   CPFA150 Day2      Group2
5     192   CPFA150 Day2      Group2
6     159   CPFA150 Day2      Group2
```

Getting into the details of the records, we get

```
> table(ex0518$Treatment)

CPFA150 CPFA300 CPFA450  CPFA50 CPFA600 Control
      3       3       3       3       3      15
> table(ex0518$Day)

Day1 Day2 Day3 Day4 Day5
   6    6    6    6    6
> table(ex0518$TrtDayGroup)

 Group1 Group10  Group2  Group3  Group4  Group5  Group6  Group7  Group8
Group9
      3       3       3       3       3       3       3       3       3
3
```

From the above it can be noticed that typically there are 6 records (CPFA50, CPFA150, CPFA300, CPFA450, CPFA600 and control) taken 3 on CPFA protein level and 3 on control group level, totally 6 records per day for 5 days. From this, let's the summary details of the experiment

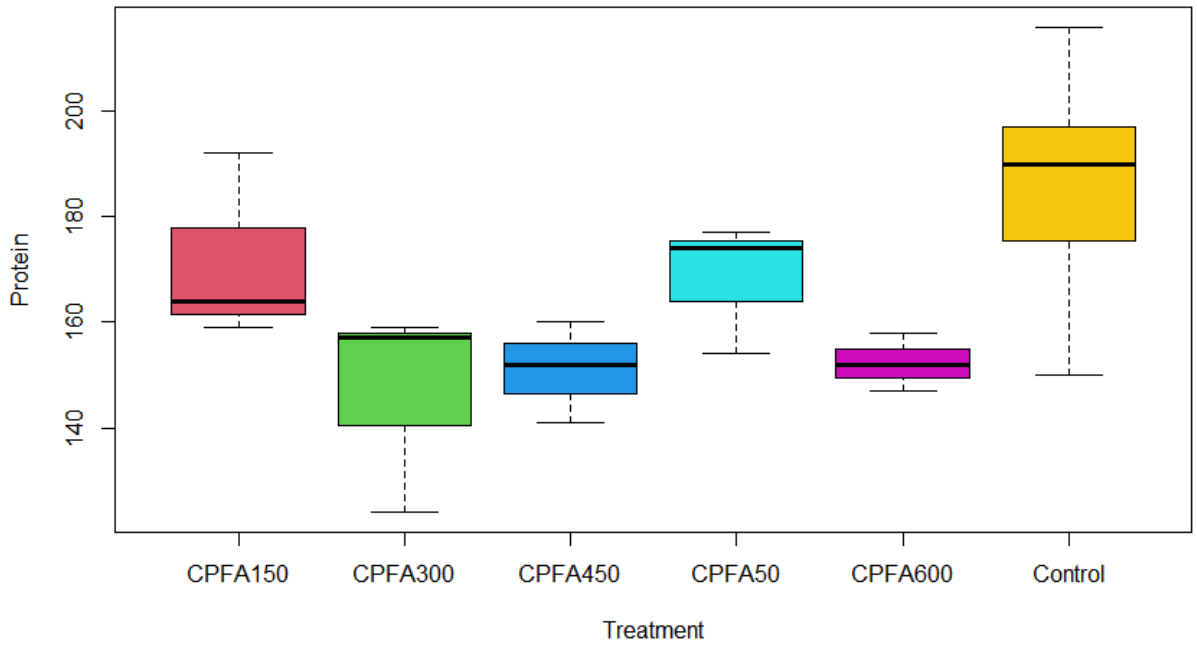
```
> tapply(ex0518$P, ex0518$Treatment, summary)
$CPFA150
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  159.0   161.5   164.0   171.7   178.0   192.0
$CPFA300
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  124.0   140.5   157.0   146.7   158.0   159.0
$CPFA450
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  141.0   146.5   152.0   151.0   156.0   160.0
$CPFA50
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  154.0   164.0   174.0   168.3   175.5   177.0
$CPFA600
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  147.0   149.5   152.0   152.3   155.0   158.0
$Control
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  150.0   175.5   190.0   185.6   197.0   216.0
> summary(ex0518)
    Protein        Treatment    Day     TrtDayGroup
 Min.   :124.0   CPFA150: 3   Day1:6   Group1 : 3
 1st Qu.:157.0   CPFA300: 3   Day2:6   Group10: 3
 Median :164.5   CPFA450: 3   Day3:6   Group2 : 3
 Mean   :171.8   CPFA50 : 3   Day4:6   Group3 : 3
 3rd Qu.:190.8   CPFA600: 3   Day5:6   Group4 : 3
 Max.   :216.0   Control:15            Group5 : 3
                                       (Other):12
```

We could clearly infer that the mean and medians of all the groups are clearly different but there are few similarities like – the mean and median of CPFA600 and CPFA450 are similar but let's see on how more these are different from each other and from other groups.

Let's see a boxplot on this data

```
> boxplot(Protein ~ Treatment, data = ex0518, col = 2:8)
```



5.25

This is a study on random sample containing annual income of 2584 Americans in 2005, selected for National Longitudinal Survey of Youth in 1979, who had paying jobs in 2005. The data contains the records of the number of years of education that one had completed by 2006 in categories - <12, 12, 13-15, 16 and >16. Let's head to the data by the below code

```
> #25
> head(ex0525)
  Subject  Educ Income2005
1       2    12       5500
2       6    16      65000
3       7    12      19000
4       8 13-15      36000
5       9 13-15      65000
```

```
6      13    16       8000
```

The below codes gives the summary of the data attributes

```
> summary(ex0525)
    Subject         Educ        Income2005
 Min.   :    2   12   :1020   Min.   :    63
 1st Qu.: 1586   13-15: 648   1st Qu.: 23000
 Median : 3108   16   : 406   Median : 38231
 Mean   : 3494   <12  : 136   Mean   : 49417
 3rd Qu.: 4636   >16  : 374   3rd Qu.: 61000
 Max.   :12140                Max.   :703637
```

The summary tells that the minimum and maximum value of the income field is spread on high scale and needs to be transformed to make a clear statistical analysis. By the below code, lets transform the income field to log scale and then analyze the summary

```
> income.log <- log(ex0525$Income2005)
> tapply(income.log, ex0525$Educ, summary)
$`12`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.704   9.902  10.342  10.227  10.779  12.924


$`13-15`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.061  10.086  10.545  10.391  10.968  12.458


$`16`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.298  10.373  10.942  10.797  11.396  13.160


$`<12`
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

5.858 9.547 10.065 9.899 10.519 11.513

$`>16`

Min. 1st Qu. Median Mean 3rd Qu. Max.

4.143 10.597 11.010 10.898 11.472 13.464

We could see that the mean and medians are close that no inference can be made whether any group is different or two are different from all the others. Moreover, the levels of the education field categories are also unordered. Let's order the groups by below commands

```
> educ.ordered <- factor(ex0525$Educ,
+                        levels = c("<12", "12", "13-15", "16", ">16") )
> levels(ex0525$Educ)
[1] "12"    "13-15" "16"    "<12"   ">16"
> levels(educ.ordered)
[1] "<12"   "12"    "13-15" "16"    ">16"
> table(ex0525$Educ == educ.ordered )
```

TRUE

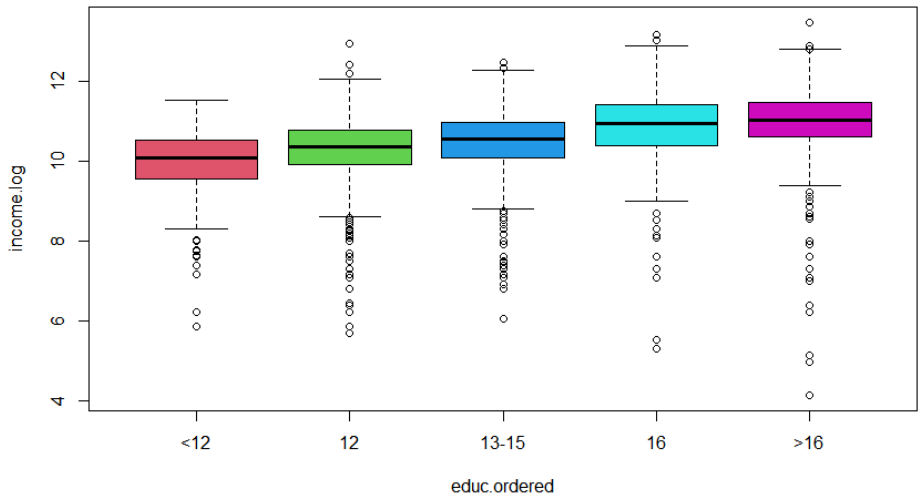
2584

The above commands show that the data is ordered in the table according to the order stored in the variable educ.ordered. So that now it will be easier for us to identify the groups in a particular fashion.

Now let's make a boxplot on this ordered and transformed data by the below command

```
> boxplot(income.log ~ educ.ordered, col = 2:8)
```

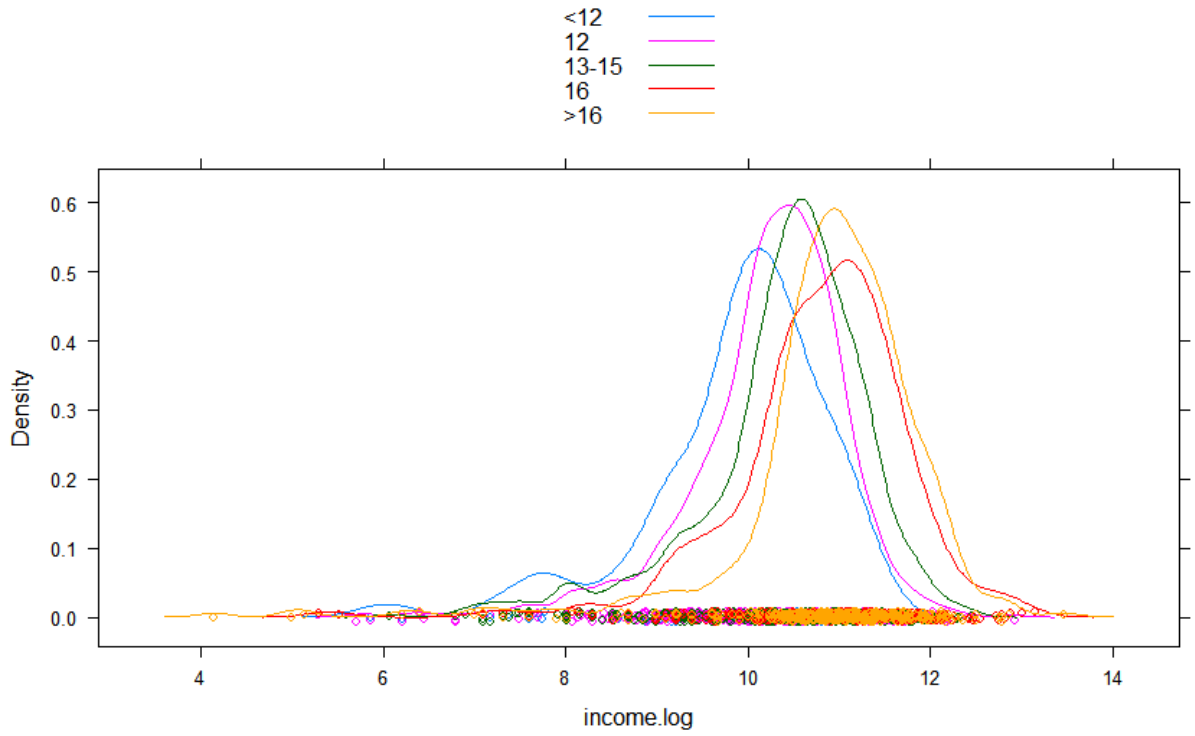
And the output is as below



The boxplot shows that there is a lot of spread in all the groups but the means are bit closer to distinguish if any group is different from other.

The density plot below shows the spread of all 2584 data points

> densityplot(income.log, groups = educ.ordered, auto.key = TRUE)



The plot also shows that the means are closer and gives no clear clue that if a group is different from other. Below is the code for pictorial result of grouped ANOVA

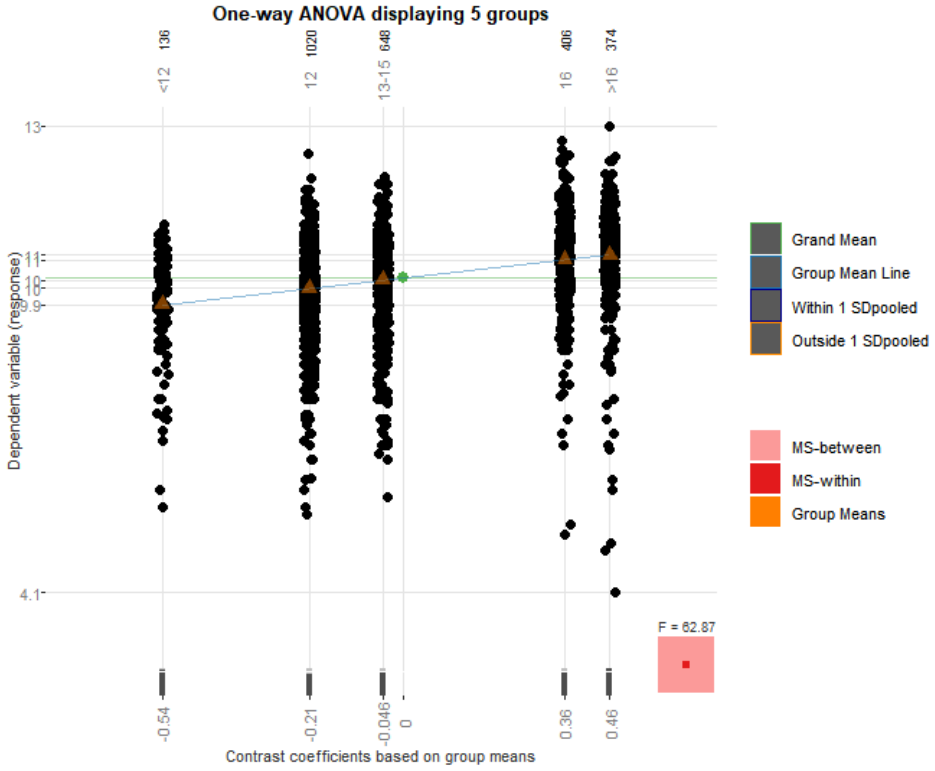
```
> granovagg.1w(income.log, group = educ.ordered)

By-group summary statistics for your input data (ordered by group means)

  group  group.mean  trimmed.mean  contrast  variance  standard.deviation
group.size

1   <12        9.90         10.04     -0.54      1.00                1.00
136

2    12       10.23         10.33     -0.21      0.73                0.85
1020

3 13-15       10.39         10.53     -0.05      0.86                0.93
648

4    16       10.80         10.89      0.36      0.92                0.96
406

5   >16       10.90         11.02      0.46      1.14                1.07
374

Below is a linear model summary of your input data

Call:

lm(formula = score ~ group, data = owp$data)

Residuals:

    Min      1Q  Median      3Q     Max

-6.7548 -0.3480  0.1208  0.5742  2.6967

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  9.89934    0.07977 124.092  < 2e-16 ***

group12      0.32787    0.08493   3.861 0.000116 ***

group13-15   0.49187    0.08775   5.606  2.3e-08 ***

group16      0.89775    0.09217   9.740  < 2e-16 ***

group>16     0.99856    0.09316  10.719  < 2e-16 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 0.9303 on 2579 degrees of freedom

Multiple R-squared: 0.08885, Adjusted R-squared: 0.08743

F-statistic: 62.87 on 4 and 2579 DF, p-value: < 2.2e-16

The above results show that there is strong evidence that at least one of the groups are statistically different form others with p value <0.0001.



The above plot shows the depiction of the results where the contrast difference of grouped mean and trimmed mead for group with year of education 16 and >16 are above 0 and grand mean while it is below 0 and grand mean for the remaining groups.

The below code is to find if there is any difference in the groups by ANOVA by drawing linear models.

> educ.lm <- lm(income.log ~ educ.ordered)

> summary(educ.lm)

Call:

lm(formula = income.log ~ educ.ordered)

Residuals:

Min 1Q Median 3Q Max

-6.7548 -0.3480 0.1208 0.5742 2.6967

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	9.89934	0.07977	124.092	< 2e-16	***
educ.ordered12	0.32787	0.08493	3.861	0.000116	***
educ.ordered13-15	0.49187	0.08775	5.606	2.3e-08	***
educ.ordered16	0.89775	0.09217	9.740	< 2e-16	***
educ.ordered>16	0.99856	0.09316	10.719	< 2e-16	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9303 on 2579 degrees of freedom

Multiple R-squared: 0.08885,   Adjusted R-squared: 0.08743

F-statistic: 62.87 on 4 and 2579 DF, p-value: < 2.2e-16

> anova(educ.lm)

Analysis of Variance Table

Response: income.log

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
educ.ordered	4	217.65	54.413	62.87	< 2.2e-16	***
Residuals	2579	2232.12	0.865			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

However, all the models gives same results that at least one group is different from other with strong evidence of $p < 0.0001$

Now lets do a pairwise comparison to know by what percent the mean or median for each category exceed the other lowest category. The below are the commands

Below is the comparison of the groups with educational years <12 with group of educational years 12

> # (<12) < 12

> fit.contrast(educ.lm, "educ.ordered", coef = c(-1,1,0,0,0),

```
+            conf.int = .95)

                    Estimate Std. Error  t value     Pr(>|t|)  lower CI  upper CI

educ.ordered c=( -1 1 0 0 0 ) 0.3278745 0.08492636 3.860692 0.0001158443 0.1613437 0.4944053

attr(,"class")

[1] "fit_contrast"
```

Since the log transformation has been applied on the income, lets back transform the results to get the results

```
> exp(fit.contrast(educ.lm, "educ.ordered", coef = c(-1,1,0,0,0),

+              conf.int = .95))

                    Estimate Std. Error t value Pr(>|t|) lower CI upper CI

educ.ordered c=( -1 1 0 0 0 ) 1.388015   1.088637 47.4982 1.000116 1.175089 1.639523

attr(,"class")

[1] "fit_contrast"
```

The above results show that the group with educational years <12 and 12 are different with median of the group with educational years <12 38.8% greater than median of the group with educational years 12 with a confidence interval of 17.5% as lower limit and 64% as upper limit.