**MET CS 555- ASSIGNMENT 5**  
BU ID: U55-32-1699

**1. Reading the data into a csv file and then extracting the data in R:**  
  
> setwd("C:/Users/Lenovo/Downloads")

> groups <- read.csv("assgn5.csv")

> View(groups)

Summarizing the data:  
  
> library(psych)

> describe(groups)

vars n mean sd median trimmed mad min max range skew kurto se

group\* 1 45 2.00 0.832 2.00 1.48 1 3 2 0.00 -1.57 0.12

iq 2 45 39.33 6.91 39 39.43 7.41 24 52 28 -0.15 -0.65 1.03

age 3 45 25.98 10.64 20 25.11 5.93 14 46 32 0.67 -1.24 1.59

**Number of students in each group:**  
  
> table(groups$group)

Chemistry student Math student Physics student

15 15 15

Therefore, each group has an equal distribution of 15 students each.

Summary with respect to group and age:  
> table(groups$group,groups$age)

14 15 16 17 18 19 20 22 23 24 28 32 33 36 38 39 40 41 42 44 46

Chemistry student 0 0 0 0 0 0 0 0 0 0 0 1 1 1 2 1 1 3 1 2 2

Math student 0 0 1 0 2 3 3 2 2 1 1 0 0 0 0 0 0 0 0 0 0

Physics student 1 2 3 3 2 2 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0

It is evident from the above results that chemistry students consist of the older population and the physics students comprise of the younger population, whereas the math students lie in between the two groups.

> aggregate(groups$age, by=list(groups$group), summary)

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

1 Chemistry student 32.000 38.00000 41.00000 40.06667 43.00000 46.00000

2 Math student 16.000 19.00000 20.00000 20.73333 22.50000 28.00000

3 Physics student 14.000 16.00000 17.00000 17.13333 18.50000 20.00000

|  |
| --- |
| > boxplot(groups$age~groups$group, main="Age by group", xlab="group",ylab="age",  ylim=c(14, 46)) |
|  |
| |  | | --- | |  | |

**Summary with respect to group and iq:**

> table(groups$group,groups$iq)

24 25 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 49 51 52

Chemistry student 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 3 3 1 1 2

Math student 1 0 1 0 0 0 0 0 1 2 1 3 1 1 1 0 0 2 1 0 0 0 0 0

Physics student 0 1 1 1 1 1 2 3 0 1 0 0 2 0 1 1 0 0 0 0 0 0 0 0

This data above revelas that the chemistry students have a higher level of IQ than the physics and maths students. The physics students have the least range of IQ and the maths students have an intermediate level of IQ.

> aggregate(groups$iq, by=list(groups$group), summary)

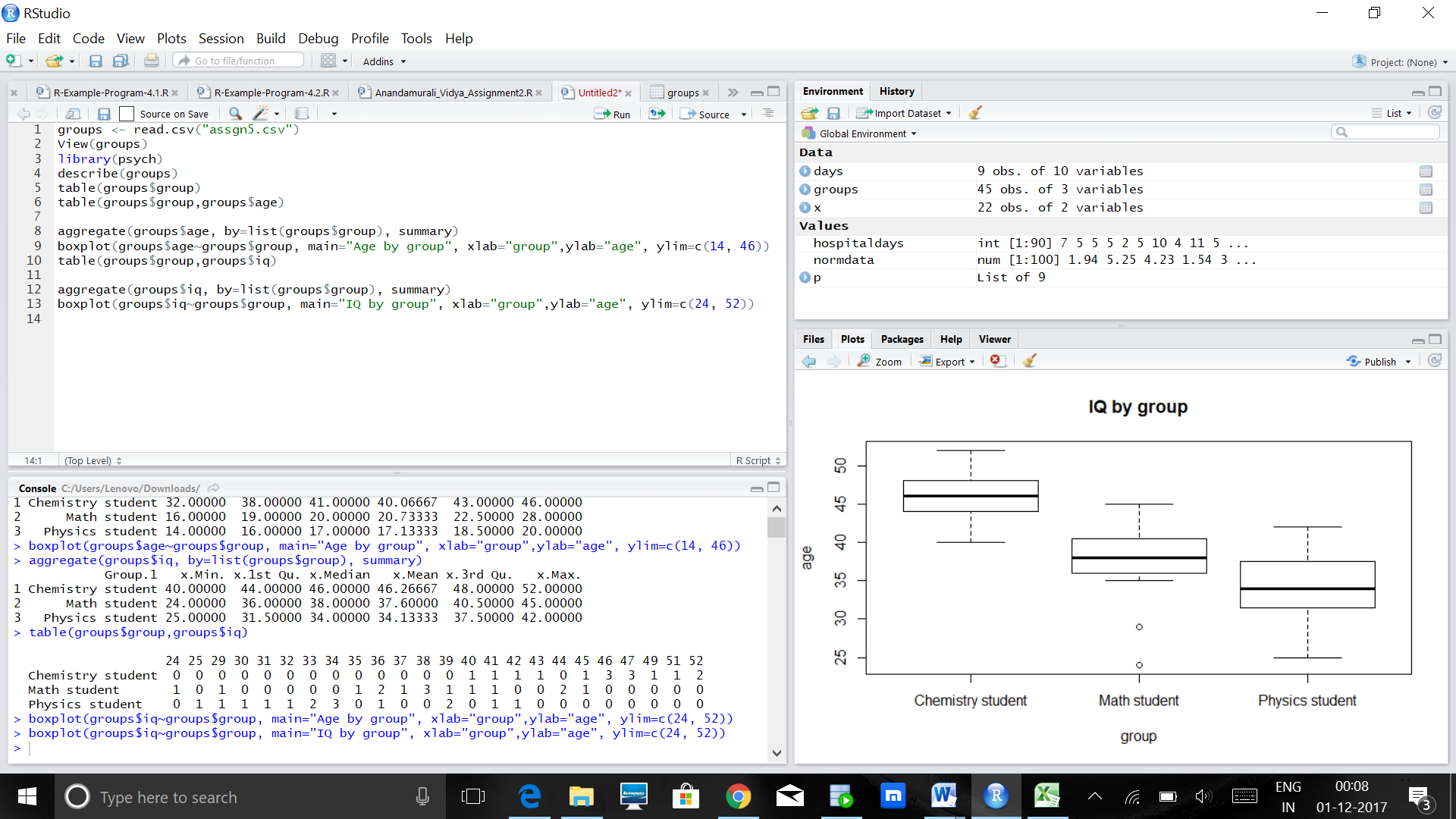
Group.1 x.Min. x.1st Qu. x.Median x.Mean x.3rd Qu. x.Max.

1 Chemistry student 40.000 44.00000 46.00000 46.26667 48.00000 52.00000

2 Math student 24.000 36.00000 38.00000 37.60000 40.50000 45.00000

3 Physics student 25.000 31.50000 34.00000 34.13333 37.50000 42.00000

|  |
| --- |
| > boxplot(groups$iq~groups$group, main="IQ by group", xlab="group",ylab="age",  ylim=c(24, 52)) |
|  |
| |  | | --- | |  | |



**2. 5 step procedure:**  
step 1:  
H0 : All the underlying population means are equal  
H1 : Not all of the underlying population means are equal  
alpha = 0.05

Step 2:  
Select the appropriate test statistic  
F=MSB/MSW

Step 3:  
State the decision rule  
Determine the appropriate value from the F-distribution with k−1=3−1=2, n−k= 45−3=42 degrees of freedom and associated with a right hand tail probability of α=0.05.

Using the software: > qf(.95, df1=2, df2=42)

[1] 3.219942

Therefore, F statistic is 3.2199 ; We reject H0, if F > or = 3.2199.

Step 4:   
> m<- lm(groups$age~ groups$group + groups$iq)

> anova(m)

Analysis of Variance Table

Response: groups$age

Df Sum Sq Mean Sq F value Pr(>F)

groups$group 2 4563.4 2281.69 256.9626 < 2e-16 \*\*\*

groups$iq 1 57.5 57.54 6.4804 0.01476 \*

Residuals 41 364.1 8.88

---

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

In this step, group and IQ together have a significance towards the age factor in the data set.  
By comparing the F values obtained from the above test and the value obtained in our previous step, we can infer that F value is always greater than 3.2199 for both the pairs.

Step 5:  
Since the F value obtained using the anova table produces a higher value, the null hypothesis can be rejected and it can be inferred that there is difference in values of the three different groups given.

Now that there is a difference in the group values, a pairwise comparison must be performed to find out where exactly the differences exist.

**Pairwise Comparisons :**  
Chemistry versus Physics-  
1. Set up hypothesis:  
H0 : µ(chemistry)= µ(Physics)  
H1 : µ(Chemistry) ≠ µ(Physics)  
α = 0.05

2. Selecting the appropriate T test:  
t = x¯Chemistry−x¯Physics/ root(sp^2(1/nChemistryt+1/nPhysics))

3. Decision Rule:  
n-k = 45-3 = 42,  
α/2 = 0.05/2 = 0.025  
t = 2.018082  
> qt(0.975,42)

[1] 2.018082

Decision Rule: Reject H0 if t ≥2.018

4. Computing the test statistic:  
t = 46.2 – 34.1 / root(8.88(1/15 + 1/15))  
 = 12.1/1.08  
 = 11.11

5. We reject H0 since, 11.11 > 2.018, and hence we have proof that mean of both the chemistry and physics group vary significantly.

Tukey Comparison Tests using the software:  
  
With respect to the age parameter:  
> m5<- aov(formula = groups$age ~groups$group + groups$iq,data = groups)

> TukeyHSD(m5)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = groups$age ~ groups$group + groups$iq, data = groups)

$`groups$group`

diff lwr upr p adj

Math student-Chemistry student -19.33333 -21.979175 -16.6874917 0.00000

Physics student-Chemistry student -22.93333 -25.579175 -20.2874917 0.00000

Physics student-Math student -3.60000 -6.245842 -0.9541584 0.00544

With respect to the IQ parameter:  
> m5<- aov(formula = groups$iq ~groups$group + groups$age,data = groups)

> TukeyHSD(m5)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = groups$iq ~ groups$group + groups$age, data = groups)

$`groups$group`

diff lwr upr p adj

Math student-Chemistry student -8.666667 -12.588432 -4.7449018 0.0000099

Physics student-Chemistry student-12.133333 -16.055098 -8.2115684 0.000000

Physics student-Math student -3.466667 -7.388432 0.4550982 0.0924098

Therefore, the above two methods can also be used to find out the specifications about the pairwise comparisons of the test to be performed.

**3. Creating dummy variables and re running the anova test:**There are three groups in the above problem and hence we need k-1 dummy variables i.e 3-1 = 2 dummy variables.

**Group group2 group3**  
Chemistry 0 0  
Physics 1 0  
Math 0 1

Creating the dummy variables:  
groups$g0 <- ifelse(groups$group =='Chemistry student', 1, 0)

groups$g1 <- ifelse(groups$group =='Physics student', 1, 0)

groups$g2 <- ifelse(groups$group =='Math student', 1, 0)

Analysing the variables:

> m2 <- lm(groups$iq ~ groups$g1+ groups$g2, data = groups)

> summary(m2)

Call:

lm(formula = groups$iq ~ groups$g1 + groups$g2, data = groups)

Residuals:

Min 1Q Median 3Q Max

-13.6000 -2.1333 -0.1333 2.7333 7.8667

Coefficients:

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

(Intercept) 46.267 1.213 38.157 < 2e-16 \*\*\*

groups$g1 -12.133 1.715 -7.076 1.13e-08 \*\*\*

groups$g2 -8.667 1.715 -5.054 8.93e-06 \*\*\*

---

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

Residual standard error: 4.696 on 42 degrees of freedom

Multiple R-squared: 0.5585, Adjusted R-squared: 0.5375

F-statistic: 26.57 on 2 and 42 DF, p-value: 3.496e-08

> anova(m2)

Analysis of Variance Table

Response: groups$iq

Df Sum Sq Mean Sq F value Pr(>F)

groups$g1 1 608.40 608.40 27.587 4.676e-06 \*\*\*

groups$g2 1 563.33 563.33 25.543 8.931e-06 \*\*\*

Residuals 42 926.27 22.05

---

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

> m3 <- lm(groups$age~ groups$g1+ groups$g2)

> summary(m3)

Call:

lm(formula = groups$age ~ groups$g1 + groups$g2)

Residuals:

Min 1Q Median 3Q Max

-8.0667 -1.7333 -0.1333 1.8667 7.2667

Coefficients:

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

(Intercept) 40.0667 0.8181 48.98 <2e-16 \*\*\*

groups$g1 -22.9333 1.1569 -19.82 <2e-16 \*\*\*

groups$g2 -19.3333 1.1569 -16.71 <2e-16 \*\*\*

---

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

Residual standard error: 3.168 on 42 degrees of freedom

Multiple R-squared: 0.9154, Adjusted R-squared: 0.9114

F-statistic: 227.3 on 2 and 42 DF, p-value: < 2.2e-16

> anova(m3)

Analysis of Variance Table

Response: groups$age

Df Sum Sq Mean Sq F value Pr(>F)

groups$g1 1 1760.0 1760.04 175.34 < 2.2e-16 \*\*\*

groups$g2 1 2803.3 2803.33 279.27 < 2.2e-16 \*\*\*

Residuals 42 421.6 10.04

---

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

Equation : (with chemistry as the reference group for IQ)  
y = 46.2 - 12.13Groupmaths - 8.66Groupphysics + e

(with chemistry as the reference group for age)  
y = 40.06 – 22.9Groupmaths -19.33Groupphysics +e

**As we can see, by comparing the two models, even before and after adding the dummy variables the results are the same.**

The Tukey comparison for the above can be given as :  
> m4<- aov(formula =groups$age ~groups$group,data = groups)

> TukeyHSD(m4)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = groups$age ~ groups$group, data = groups)

$`groups$group`

diff lwr upr p adj

Math student-Chemistry student -19.33333 -22.144010 -16.5226572 0.00000

Physics student-Chemistry student -22.93333 -25.744010 -20.1226572 0.00000

Physics student-Math student -3.60000 -6.410676 -0.7893238 0.00917

The above can show the variation in the above group comparisons.

**4. One way Anova adjusting for age:**> library(car)

> Anova(lm(groups$iq~groups$group+groups$age), type=3)

Anova Table (Type III tests)

Response: groups$iq

Sum Sq Df F value Pr(>F)

(Intercept) 152.74 1 7.8294 0.007797 \*\*

groups$group 21.89 2 0.5610 0.574969

groups$age 126.42 1 6.4804 0.014763 \*

Residuals 799.84 41

---

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

In step 2, we observed that the values where significant when we considered the age factor for both the IQ as well as the group to which they belonged. Now, as we do the Ancova test, we realise that the age factor alone produces the significance as opposed to the conclusion we arrived at in step 2. Therefore now, we can say that, the age is significantly influenced by the IQ levels of an individual.

**Mean square Analysis:**  
> library(lsmeans)

> lsmeans(lm(groups$iq~groups$group+groups$age), pairwise~groups$group, adjust = "Tukey")

$lsmeans

groups$group lsmean SE df lower.CL upper.CL

Chemistry student 32.96513 1.229289 41 30.48253 35.44773

Math student 34.06032 1.140782 41 31.75646 36.36418

Physics student 32.96513 1.229289 41 30.48253 35.44773

Confidence level used: 0.95

$contrasts

contrast estimate SE df t.ratio p.value

Chemistry student - Math student -1.095193 0.4302201 41 -2.546 0.0384

Chemistry student - Physics student 0.000000 0.0000000 41 NaN NaN

Math student - Physics student 1.095193 0.4302201 41 2.546 0.0384

P value adjustment: tukey method for comparing a family of 3 estimates

The above data has unbalanced data points which is rectified by the mean square model and produces a better data set to analyse. The p value is significant only for the groups, chemistry –math and math – physics and this is because : “Two LS means that share one or more of the same grouping symbols are not significantly different at the stated value of alpha, after applying the multiplicity adjustment (in this case Tukey’s HSD).”  
The least square means are the adjusted means that are generated at the 95% confidence interval. The lsmean provides a starting point to the dataset. The combinations of the chemistry and math, chemistry and physics and math and physics sometimes show two LS means that share the same average value at the same alpha level at the stated Tukey HSD level, and hence both don’t remain significant in the model.