########AS Group Assignment##########

#The null hypothesis for ANOVA is that the mean

#(average value of the dependent variable) is the same for all groups.

# The alternative #or research hypothesis is that the average is not the

# same for all groups.

set.seed(123)

setwd("G:/Grreat lakes/Advance stat/Group Assignment")

mydata<- read.csv("Expectations Evaluation.csv")



# Converting Y2 and Y1 in numeric variable

mydata$Y2<- as.numeric(mydata$Y2)

mydata$Y1<- as.numeric(mydata$Y1)

#glm(Y1~.,data=mydata)

# Outlier Analysis - Varaiable

outlier\_upper=function(x){

q = quantile(x)

names(q) = NULL

q1 = q[2]

q3 = q[4]

QR = q3-q1

return(q3+1.5\*QR);

}

outlier\_lower=function(x){

q = quantile(x)

names(q) = NULL

q1 = q[2]

q3 = q[4]

QR = q3-q1

return(q1-1.5\*QR);

}

**# outlier limits validation** ------------------

Y2\_upper = outlier\_upper(mydata$Y2)

Y2\_lower = outlier\_lower(mydata$Y2)

Y1\_upper = outlier\_upper(mydata$Y1)

Y1\_lower = outlier\_lower(mydata$Y1)

**# Outlier data**

mydata[mydata$Y2>Y2\_upper | mydata$Y2<Y2\_lower , ]

mydata[mydata$Y1>Y1\_upper | mydata$Y1<Y1\_lower , ]

mydata = subset( mydata, mydata$Y2<=Y2\_upper & mydata$Y2>=Y2\_lower)

mydata = subset( mydata, mydata$Y1<=Y1\_upper & mydata$Y1>=Y1\_lower)

nrow(mydata)

#mydata$Expectatations.Manipulation=ifelse(mydata$Expectatations.Manipulation=="h",7,1)

HI<-table(mydata$Expectatations.Manipulation,mydata$Quality.Manipulation)

#mean(HI[,1])

#mean(HI[,2])

# **Null Hypothesis:** There is no difference in the evaluation of the magazine and the Quality Manipulation

aov(Y1~Quality.Manipulation,mydata)->ft

summary(ft)

**Output**:

**summary(ft)**

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

**Quality.Manipulation 1 43.3 43.33 19.65 1.55e-05 \*\*\***

**Residuals 196 432.1 2.20**

**---**

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

**Result**: Since P value is less than 0.05 , we will reject the Null Hypothesis.

Conclusion: There is no difference in the evaluation of the magazine and the Quality Manipulation evaluation.

TukeyHSD(ft)

**Tukey multiple comparisons of means**

**95% family-wise confidence level**

**Fit: aov(formula = Y1 ~ Quality.Manipulation, data = mydata)**

**$Quality.Manipulation**

**diff lwr upr p adj**

**g-b 0.9363311 0.51978 1.352882 1.55e-05**

# Analysis on Y1 and Quality Manipulation to see if there is any difference between these two.

tab<-table(mydata$Y1,mydata$Quality.Manipulation)

tab

b g

1 1 0

2 12 3

3 18 8

4 22 18

5 27 21

6 15 20

7 6 23

8 2 2

mn\_b<-1+24+54+88+135+90+42+16

mn\_g<-6+24+72+105+120+161+16

mn\_b

output : 420

mn\_g

output: 504

Conclusion : We can see that there is much difference in mean of factor of Quality Manipulation across all the value of Y1.

# Applying Anova to check if there is difference between Actual Evaluation of magazine and Expectation Manipulation.

#Null : There is no difference between Actual Evaluation of magazine and Expectation Manipulation.

aov(Y1~Expectatations.Manipulation,mydata)->ft

summary(ft)

Output:

**summary(ft)**

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

**Expectatations.Manipulation 1 7.9 7.935 3.327 0.0697 .**

**Residuals 196 467.5 2.385**

**---**

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

**Result:**

Since P value is more than 0.05, we will accept the Null Hypothesis.

Conclusion : There is no difference between Actual Evaluation of magazine and Expectation Manipulation.

TukeyHSD(ft)

**Tukey multiple comparisons of means**

**95% family-wise confidence level**

**Fit: aov(formula = Y1 ~ Expectatations.Manipulation, data = mydata)**

**$Expectatations.Manipulation**

**diff lwr upr p adj**

**l-h 0.4004082 -0.03253337 0.8333497 0.0696833**

**#** Analysis on Y1 and Expected Manipulation to see if there is any difference between these two.

tab<-table(mydata$Y1,mydata$Expectatations.Manipulation)

Output :

|  |
| --- |
| **Tab**    **h l**  **1 1 0**  **2 10 5**  **3 16 10**  **4 18 22**  **5 25 23**  **6 15 20**  **7 14 15**  **8 1 3** |
|  |
| |  | | --- | | **>** | |

mn\_h<- 1+20+48+72+125+90+98+8

mn\_l<-10+30+88+115+120+105+24

mn\_h

Output : 462

mn\_l

Output : 492

Conclusion: We can see that there is no much difference in mean of factor of Expected Manipulation across all the value of Y1.

aov(Y1~Expectatations.Manipulation+Quality.Manipulation,mydata)->ft

summary(ft)

Output:

summary(ft)

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

Expectatations.Manipulation 1 7.9 7.94 3.655 0.0574 .

Quality.Manipulation 1 44.1 44.11 20.317 1.13e-05 \*\*\*

Residuals 195 423.4 2.17

---

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

**Question 2:**

# To see the interaction effect of Expectatations.Manipulation and Quality.Manipulation on Y1.

# Null Hypothesis: There is no difference between effects of Expectatations.Manipulation and Quality.Manipulation on Y1.

aov(Y1~Expectatations.Manipulation\*Quality.Manipulation,mydata)->ft

summary(ft)

Output:

summary(ft)

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

Expectatations.Manipulation 1 7.9 7.94 3.722 0.0552 .

Quality.Manipulation 1 44.1 44.11 20.691 9.5e-06 \*\*\*

Expectatations.Manipulation:Quality.Manipulation 1 9.8 9.79 4.592 0.0334 \*

Residuals 194 413.6 2.13

---

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

Result: Since P value is less than 0.05 Quality.Manipulation, we can assume that this variable is significant and also interaction effect is less than 0.05 we will reject the Null Hypothesis.

TukeyHSD(ft)

TukeyHSD(ft)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = Y1 ~ Expectatations.Manipulation \* Quality.Manipulation, data = mydata)

$Expectatations.Manipulation

diff lwr upr p adj

l-h 0.4004082 -0.008929968 0.8097463 0.05516

$Quality.Manipulation

diff lwr upr p adj

g-b 0.9445971 0.5349453 1.354249 9.5e-06

$`Expectatations.Manipulation:Quality.Manipulation`

diff lwr upr p adj

l:b-h:b -0.007164404 -0.7528735 0.7385447 0.9999944

h:g-h:b 0.505002001 -0.2519258 1.2619298 0.3115137

l:g-h:b 1.388320546 0.6189077 2.1577334 0.0000323

h:g-l:b 0.512166405 -0.2411866 1.2655194 0.2950111

l:g-l:b 1.395484950 0.6295886 2.1613813 0.0000264

l:g-h:g 0.883318545 0.1064949 1.6601422 0.0187571

**Conclusion:**

**The yellow marked result are playing a significant role in varying the Y1.**

**Replacing “h” with 7 and converting into integer to Expectatations.Manipulation column to find out the difference between expectation manipulation and Quality Manipulation.**

**Null Hypothesis: There is no difference between Expected Manipulation and Quality Manipulation of magazine.**

mydata$Expectatations.Manipulation=ifelse(mydata$Expectatations.Manipulation=="h",7,1)

aov(Expectatations.Manipulation~Quality.Manipulation,mydata)->ft

summary(ft)

summary(ft)

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

Quality.Manipulation 1 0.8 0.758 0.083 0.773

Residuals 196 1781.1 9.087

**Conclusion : we will not reject the Null Hypothesis.**

**Using Lm Command**

mydata$Expectatations.Manipulation=ifelse(mydata$Expectatations.Manipulation==7,"h",1)

**Taking training data and testing data**

mydata$random <- runif(nrow(mydata), 0, 1);

mydata <- mydata[order(mydata$random),]

mydata.dev <- mydata[which(mydata$random <= 0.75),]

mydata.val <- mydata[which(mydata$random > 0.75),]

c(nrow(mydata.dev), nrow(mydata.val))

Output:

[1] 151 47

ft1<- lm(Y1~Quality.Manipulation,data=mydata.dev)

#ft1<- lm(Y1~.,data=mydata.dev)

str(mydata.val$Quality.Manipulation)

summary(ft1)

**summary(ft1)**

**Call:**

**lm(formula = Y1 ~ Quality.Manipulation, data = mydata.dev)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-3.395 -1.240 -0.240 0.760 3.605**

**Coefficients:**

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

**(Intercept) 4.3947 0.1717 25.596 < 2e-16 \*\*\***

**Quality.Manipulationg 0.8453 0.2436 3.469 0.000682 \*\*\***

**---**

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

**Residual standard error: 1.497 on 149 degrees of freedom**

**Multiple R-squared: 0.07475, Adjusted R-squared: 0.06854**

**F-statistic: 12.04 on 1 and 149 DF, p-value: 0.0006821**

**Conclusion : We will reject the Null Hypothesis.**

Quality.Manipulation<-data.frame(Quality.Manipulation= mydata.val$Quality.Manipulation)

str(Quality.Manipulation)

predict(ft1 , c(Quality.Manipulation))

**predict(ft1 , c(Quality.Manipulation))**

**1 2 3 4 5 6 7 8 9 10**

**4.394737 4.394737 4.394737 4.394737 5.240000 4.394737 4.394737 4.394737 4.394737 4.394737**

**11 12 13 14 15 16 17 18 19 20**

**4.394737 4.394737 4.394737 5.240000 5.240000 5.240000 5.240000 4.394737 5.240000 5.240000**

**21 22 23 24 25 26 27 28 29 30**

**4.394737 4.394737 4.394737 5.240000 5.240000 4.394737 4.394737 4.394737 4.394737 5.240000**

**31 32 33 34 35 36 37 38 39 40**

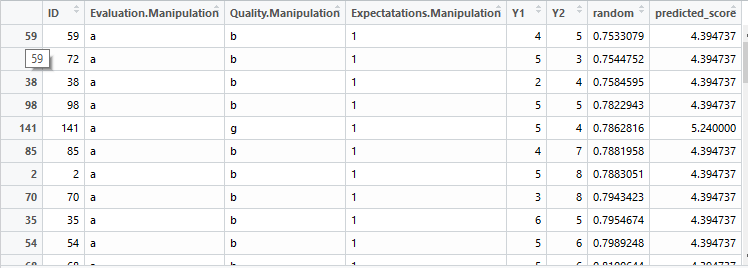
**4.394737 5.240000 5.240000 5.240000 5.240000 4.394737 5.240000 5.240000 5.240000 4.394737**

**41 42 43 44 45 46 47**

**4.394737 4.394737 5.240000 5.240000 4.394737 5.240000 4.394737**

mydata.val$predicted\_score<-predict(ft1 , c(Quality.Manipulation))

View(mydata.val)



mydata$Expectatations.Manipulation=ifelse(mydata$Expectatations.Manipulation=="h",7,1)

ft1<- lm(Expectatations.Manipulation~Quality.Manipulation,data=mydata)

summary(ft1)

**Call:**

**lm(formula = Expectatations.Manipulation ~ Quality.Manipulation,**

**data = mydata)**

**Residuals:**

**Min 1Q Median 3Q Max**

**-3.095 -2.971 2.905 3.029 3.029**

**Coefficients:**

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

**(Intercept) 3.9709 0.2970 13.369 <2e-16 \*\*\***

**Quality.Manipulationg 0.1239 0.4288 0.289 0.773**

**---**

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

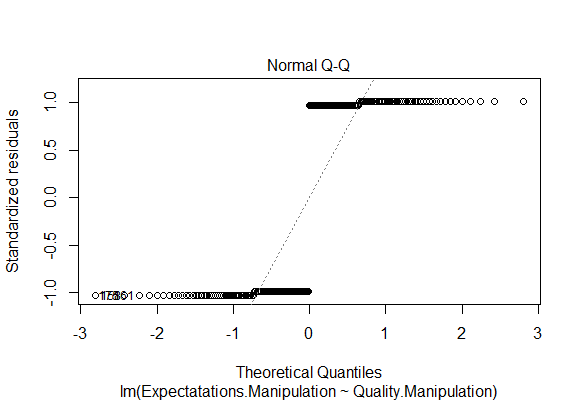
**Residual standard error: 3.014 on 196 degrees of freedom**

**Multiple R-squared: 0.0004255, Adjusted R-squared: -0.004674**

**F-statistic: 0.08344 on 1 and 196 DF, p-value: 0.773**

**Conclusion : We will not reject the Null Hypothesis.**

plot(ft)



Some more analysis

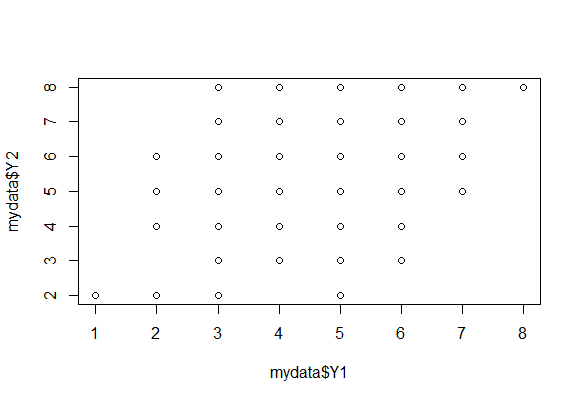
Relation between Y1 and Y2 (how are they correlated)

cor(mydata$Y1,mydata$Y2)

Output :

[1] 0.4697632 🡪 Highly Correlated

plot(mydata$Y1,mydata$Y2)



table(mydata$Y1,mydata$Y2)

**table(mydata$Y1,mydata$Y2)**

**2 3 4 5 6 7 8**

**1 1 0 0 0 0 0 0**

**2 6 0 1 6 2 0 0**

**3 1 4 2 5 7 4 3**

**4 0 2 7 10 14 3 4**

**5 1 1 6 13 17 7 3**

**6 0 2 3 7 9 8 6**

**7 0 0 0 1 9 15 4**

**8 0 0 0 0 0 0 4**