Assignment3

Reshmi

2022-09-02

#PROBLEM 1  
  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.8 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.1  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

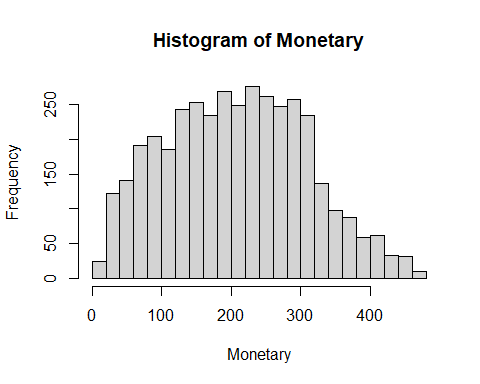
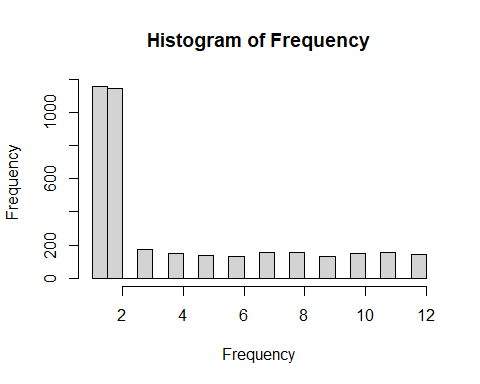
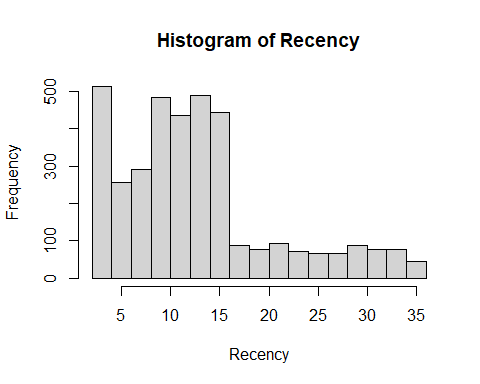
df <- read.csv('CharlesBookClubDataset.csv')  
summary(df)

## X Seq. ID. Gender   
## Min. : 0.0 Min. : 1 Min. : 25 Min. :0.0000   
## 1st Qu.: 999.8 1st Qu.:1001 1st Qu.: 8253 1st Qu.:0.0000   
## Median :1999.5 Median :2000 Median :16581 Median :1.0000   
## Mean :1999.5 Mean :2000 Mean :16595 Mean :0.7045   
## 3rd Qu.:2999.2 3rd Qu.:3000 3rd Qu.:24838 3rd Qu.:1.0000   
## Max. :3999.0 Max. :4000 Max. :32977 Max. :1.0000   
##   
## M R F FirstPurch   
## Min. : 15.0 Min. : 2.00 Min. : 1.000 Min. : 2.00   
## 1st Qu.:130.0 1st Qu.: 8.00 1st Qu.: 1.000 1st Qu.:12.00   
## Median :208.0 Median :12.00 Median : 2.000 Median :20.00   
## Mean :208.2 Mean :13.43 Mean : 3.831 Mean :26.51   
## 3rd Qu.:283.0 3rd Qu.:16.00 3rd Qu.: 6.000 3rd Qu.:36.00   
## Max. :479.0 Max. :36.00 Max. :12.000 Max. :99.00   
## NA's :93 NA's :342 NA's :218   
## ChildBks YouthBks CookBks DoItYBks   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.0000 Median :0.0000 Median :0.0000 Median :0.0000   
## Mean :0.6398 Mean :0.3048 Mean :0.7312 Mean :0.3508   
## 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :7.0000 Max. :5.0000 Max. :7.0000 Max. :5.0000   
##   
## RefBks ArtBks GeogBks ItalCook   
## Min. :0.0000 Min. :0.000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.0000 Median :0.000 Median :0.0000 Median :0.0000   
## Mean :0.2562 Mean :0.289 Mean :0.3875 Mean :0.1252   
## 3rd Qu.:0.0000 3rd Qu.:0.000 3rd Qu.:1.0000 3rd Qu.:0.0000   
## Max. :4.0000 Max. :5.000 Max. :6.0000 Max. :3.0000   
##   
## ItalAtlas ItalArt Florence Related.Purchase  
## Min. :0.0000 Min. :0.00000 Min. :0.0000 Min. :0.000   
## 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.0000 1st Qu.:0.000   
## Median :0.0000 Median :0.00000 Median :0.0000 Median :0.000   
## Mean :0.0375 Mean :0.04575 Mean :0.0845 Mean :0.885   
## 3rd Qu.:0.0000 3rd Qu.:0.00000 3rd Qu.:0.0000 3rd Qu.:1.000   
## Max. :2.0000 Max. :2.00000 Max. :1.0000 Max. :8.000   
##   
## Yes\_Florence No\_Florence Name Phone\_No.   
## Min. :0.0000 Min. :0.0000 Length:4000 Length:4000   
## 1st Qu.:0.0000 1st Qu.:1.0000 Class :character Class :character   
## Median :0.0000 Median :1.0000 Mode :character Mode :character   
## Mean :0.0845 Mean :0.9155   
## 3rd Qu.:0.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000   
##   
## Address Job   
## Length:4000 Length:4000   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##   
##

#to get number of missing values in each column  
colSums(is.na(df))

## X Seq. ID. Gender   
## 0 0 0 0   
## M R F FirstPurch   
## 93 342 218 0   
## ChildBks YouthBks CookBks DoItYBks   
## 0 0 0 0   
## RefBks ArtBks GeogBks ItalCook   
## 0 0 0 0   
## ItalAtlas ItalArt Florence Related.Purchase   
## 0 0 0 0   
## Yes\_Florence No\_Florence Name Phone\_No.   
## 0 0 0 0   
## Address Job   
## 0 0

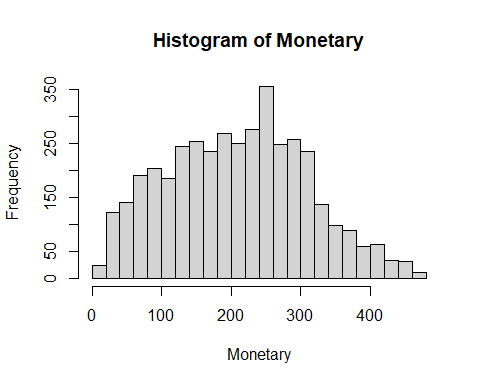
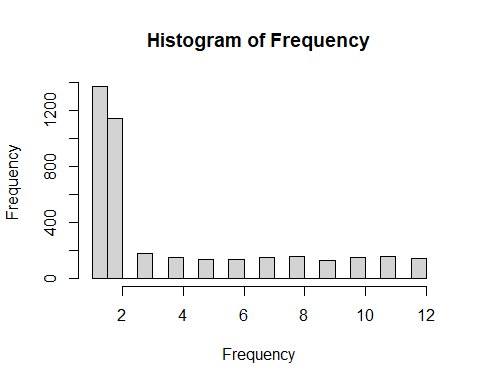
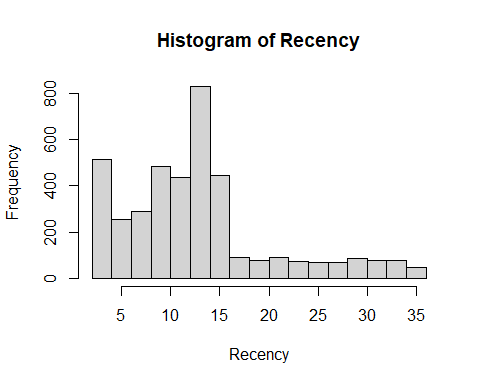
#PROBLEM 2  
  
plotHistogram <- function()   
{  
 Recency <- df$R  
 Frequency <- df$F  
 Monetary <- df$M  
 hist(Recency, breaks=20)  
 hist(Frequency, breaks=20)  
 hist(Monetary, breaks=20)  
}  
plotHistogram()



#since the graphs show that they are positively skewed, missing values can be imputed with the mode  
  
findMode <- function(x)   
{  
 mode <- names(which.max(table(x)))  
 if(is.numeric(x))   
 return(as.numeric(mode))  
 mode  
}  
df$R[is.na(df$R)] <- findMode(df$R)  
df$F[is.na(df$F)] <- findMode(df$F)  
df$M[is.na(df$M)] <- findMode(df$M)  
  
colSums(is.na(df)) #checking if NaN values are filled

## X Seq. ID. Gender   
## 0 0 0 0   
## M R F FirstPurch   
## 0 0 0 0   
## ChildBks YouthBks CookBks DoItYBks   
## 0 0 0 0   
## RefBks ArtBks GeogBks ItalCook   
## 0 0 0 0   
## ItalAtlas ItalArt Florence Related.Purchase   
## 0 0 0 0   
## Yes\_Florence No\_Florence Name Phone\_No.   
## 0 0 0 0   
## Address Job   
## 0 0

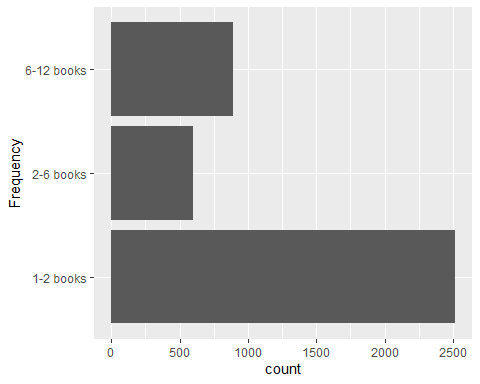
plotHistogram()



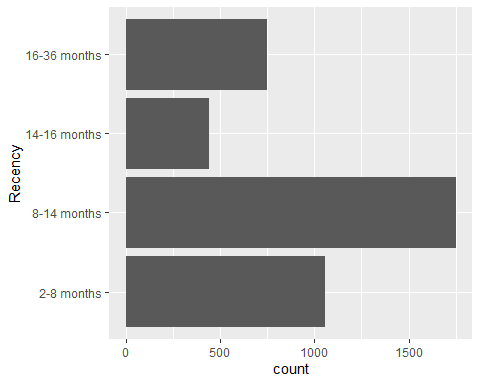
#PROBLEM 3  
  
#4 bins are used for Recency, 5 bins for Monetary and 3 bins for Frequency as every bin is to have the same number of observations because it's based on the quantiles which is optimal for binning at the breakpoints  
  
df <- df %>% mutate(Rcode=cut(df$R,breaks=unique(quantile(df$R,probs=seq.int(0,1,by=1/4))),include.lowest=TRUE),  
   
Mcode=cut(df$M,breaks=unique(quantile(df$M,probs=seq.int(0,1,by=1/5))),include.lowest=TRUE),  
  
Fcode=cut(df$F,breaks=unique(quantile(df$F,probs=seq.int(0,1,by=1/4))),include.lowest=TRUE))  
  
#new columns  
levels(df$Mcode) <- c('$15-$112','$112-$181','$181-$242','$242-$296','$296-$479')  
levels(df$Rcode) <- c('2-8 months','8-14 months','14-16 months','16-36 months')  
levels(df$Fcode) <- c('1-2 books','2-6 books','6-12 books')  
summary(df[c('Mcode', 'Rcode', 'Fcode')])

## Mcode Rcode Fcode   
## $15-$112 :801 2-8 months :1059 1-2 books :2515   
## $112-$181:808 8-14 months :1749 2-6 books : 596   
## $181-$242:800 14-16 months: 443 6-12 books: 889   
## $242-$296:791 16-36 months: 749   
## $296-$479:800

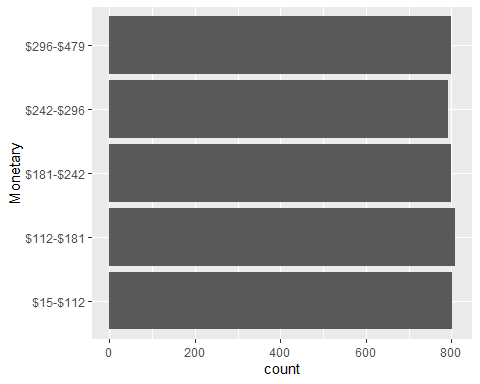
#PROBLEM 4  
  
# bar graphs  
ggplot(df, aes(x = Fcode)) +geom\_bar() +coord\_flip () +labs(x = "Frequency")



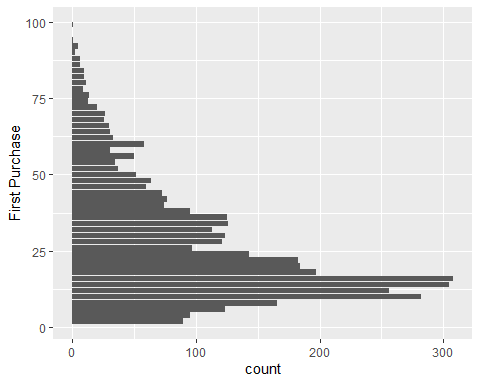
ggplot(df, aes(x = Rcode)) +geom\_bar() +coord\_flip () +labs(x = "Recency")



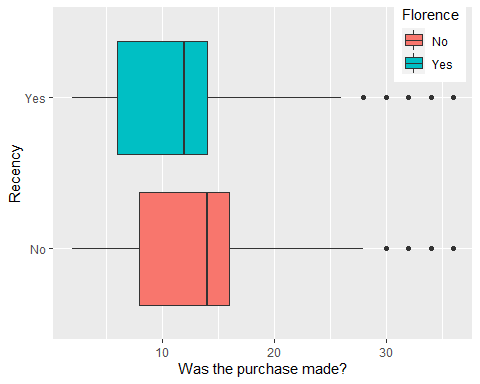
ggplot(df, aes(x = Mcode)) +geom\_bar() +coord\_flip () +labs(x = "Monetary")



ggplot(df, aes(x = FirstPurch)) +geom\_bar() +coord\_flip () +labs(x = "First Purchase")



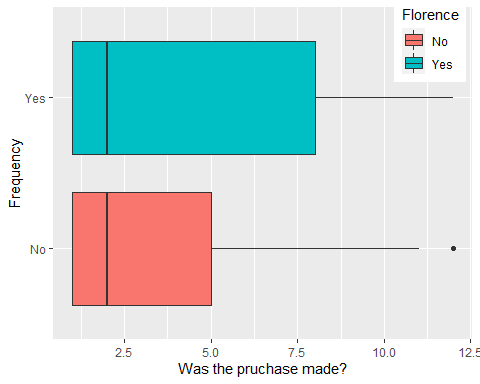
df$Florence <- factor(df$Florence,labels = c("No", "Yes")) #categorical feature  
  
#box plots  
ggplot(df, aes\_string(x="Florence", y="R", fill="Florence")) +geom\_boxplot() +coord\_flip() +labs(x="Recency", y="Was the purchase made?") +theme(legend.position=c(0.9, 0.9))



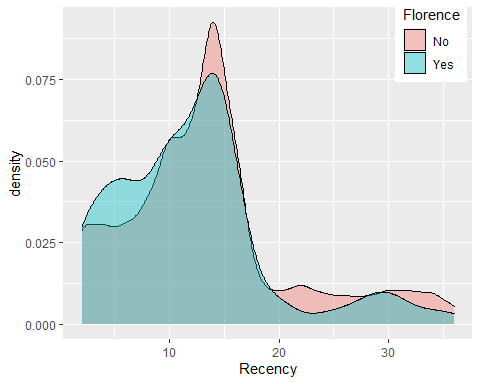
ggplot(df, aes\_string(x="Florence", y="M", fill="Florence"))+geom\_boxplot() +coord\_flip() +labs(x = "Monetary", y = "Was the pruchase made?") +theme(legend.position=c(0.9, 0.9))



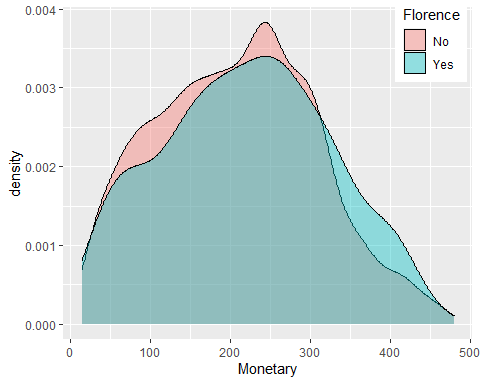
ggplot(df, aes\_string(x="Florence", y="F", fill="Florence")) +geom\_boxplot() +coord\_flip() +labs(x="Frequency", y="Was the pruchase made?") +theme(legend.position=c(0.9, 0.9))



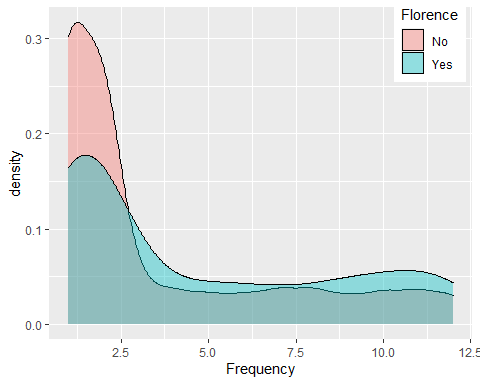
#density plots  
ggplot(df, aes\_string(x="R", fill="Florence")) +geom\_density(alpha=0.4) +labs(x="Recency") +theme(legend.position=c(0.9, 0.9))



ggplot(df, aes\_string(x="M", fill="Florence")) +geom\_density(alpha=0.4) +labs(x="Monetary") +theme(legend.position=c(0.9, 0.9))



ggplot(df, aes\_string(x="F", fill="Florence")) +geom\_density(alpha=0.4) +labs(x="Frequency") +theme(legend.position=c(0.9, 0.9))



#PART 2 - ANOVA

#PROBLEM 1  
  
library(ggpubr)  
library(dplyr)  
library(ggplot2)  
library(ggpubr)  
library(broom)  
library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

## The following object is masked from 'package:purrr':  
##   
## some

data <- read.csv('Scenario 1.csv')  
  
#1)Scully can use Fisher's test, i.e., One-way anova as there's only a single independent variable  
#2)Scully would have used the aov() fucntion  
  
oneway <- aov(No.of.items ~ POI, data)  
summary(oneway)

## Df Sum Sq Mean Sq F value Pr(>F)  
## POI 4 127 31.75 1.025 0.393  
## Residuals 995 30827 30.98

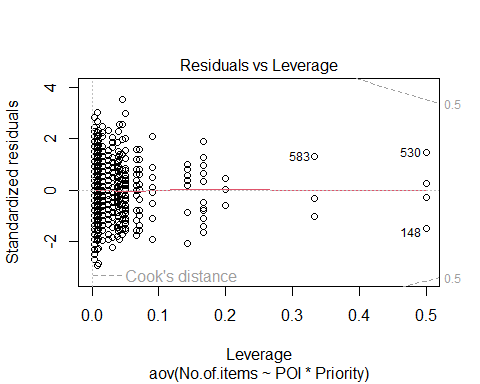
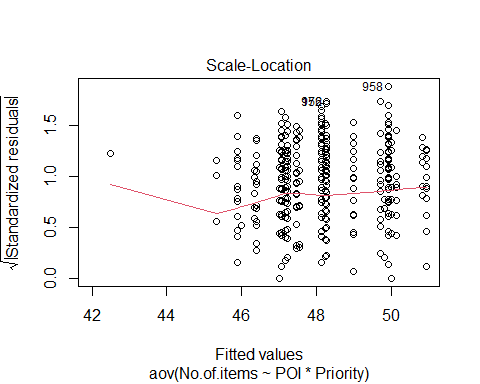
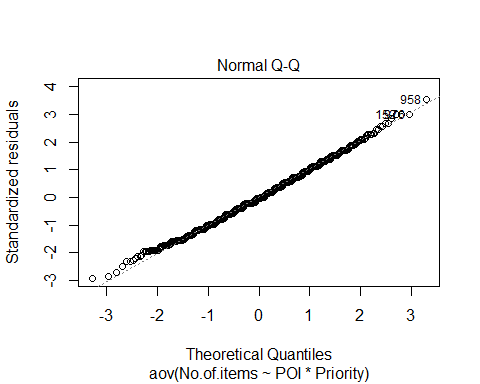
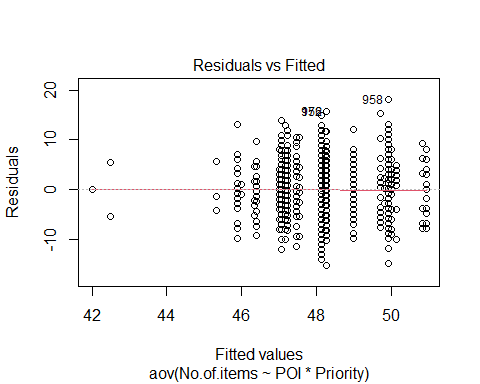
#3)based on the hypothesis of one way anova, since p value > 0.05 there is no relation between the person of interest and the average number of evidence collected against them

#PROBLEM 2  
data <- read.csv('Scenario 2.csv')  
#1)Scully can use two-way anova as there are two categorical variables on which it is dependent  
#2)Scully would have used the aov() fucntion  
twoway <- aov(No.of.items ~ POI \* Priority, data)  
summary(twoway)

## Df Sum Sq Mean Sq F value Pr(>F)   
## POI 4 317 79.29 2.880 0.0218 \*   
## Priority 4 690 172.53 6.268 5.57e-05 \*\*\*  
## POI:Priority 16 347 21.66 0.787 0.7019   
## Residuals 975 26839 27.53   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#3) Based on the hypotheses of two way ANOVA test, since p value < 0.05 there is maybe a relation between the person of interest and the avg number of evidence collected against them and between the Priority and the avg number of evidence collected against them. Since p value > 0.05 there is no interaction between the Priority and person of interest.  
  
#4) Scully should take note of the assumptions made: homogeneity of variance  
plot(twoway)

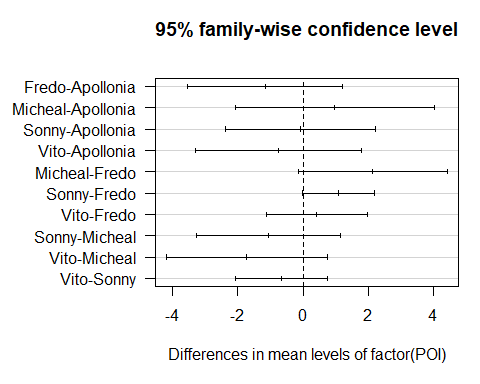
## Warning: not plotting observations with leverage one:  
## 883



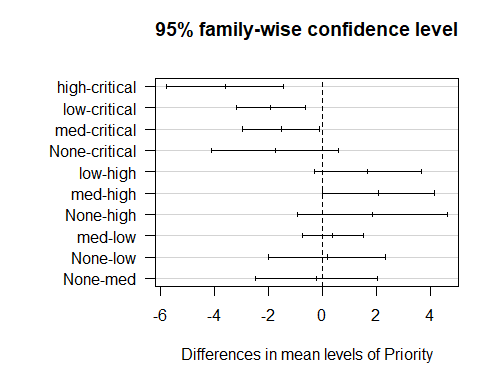
# categorical variables cannot be compared with F statistic and they are normally-distributed dependent variable  
   
  
#PROBLEM 3  
tukey.twoway <- TukeyHSD(aov(formula = No.of.items ~ factor(POI) + Priority, data ))  
tukey.twoway

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = No.of.items ~ factor(POI) + Priority, data = data)  
##   
## $`factor(POI)`  
## diff lwr upr p adj  
## Fredo-Apollonia -1.16541489 -3.5237413613 1.1929116 0.6595601  
## Micheal-Apollonia 0.97971301 -2.0408102508 4.0002363 0.9019675  
## Sonny-Apollonia -0.07537018 -2.3461092406 2.1953689 0.9999847  
## Vito-Apollonia -0.74221825 -3.2602501872 1.7758137 0.9289544  
## Micheal-Fredo 2.14512791 -0.1256015196 4.4158573 0.0745263  
## Sonny-Fredo 1.09004471 -0.0003255081 2.1804149 0.0501114  
## Vito-Fredo 0.42319665 -1.1173020367 1.9636953 0.9443143  
## Sonny-Micheal -1.05508319 -3.2347079983 1.1245416 0.6770113  
## Vito-Micheal -1.72193126 -4.1581154204 0.7142529 0.3011838  
## Vito-Sonny -0.66684807 -2.0695912176 0.7358951 0.6918200  
##   
## $Priority  
## diff lwr upr p adj  
## high-critical -3.4375139 -5.5986591 -1.2763687 0.0001482  
## low-critical -1.9212087 -3.1984071 -0.6440102 0.0004101  
## med-critical -1.5518320 -2.9734111 -0.1302530 0.0243448  
## None-critical -1.7809723 -4.1189883 0.5570436 0.2287821  
## low-high 1.5163053 -0.4599513 3.4925618 0.2221932  
## med-high 1.8856819 -0.1868144 3.9581782 0.0944859  
## None-high 1.6565416 -1.1256648 4.4387481 0.4802869  
## med-low 0.3693766 -0.7513055 1.4900587 0.8966415  
## None-low 0.1402363 -2.0280257 2.3084984 0.9997817  
## None-med -0.2291403 -2.4854672 2.0271866 0.9987008

par(mar=c(5,8,4,1)+.1)  
tukey.plot.test<-TukeyHSD(aov(formula = No.of.items ~ factor(POI), data))  
plot(tukey.plot.test, las = 1)



par(mar=c(5,8,4,1)+.1)  
tukey.plot.test<-TukeyHSD(aov(formula = No.of.items ~ Priority, data))  
plot(tukey.plot.test, las = 1)



#No pairs of POI have a statistically significant difference in mean no of evidence generated which implies that no difference in the average no of Evidence items discovered when compared with any two POI. Critical priority has a different mean which implies that assigning it to work on the cases generate different no of evidence items compared to the rest of the priorities.