PES University, Bangalore

Established under the Karnataka Act No. 16 of 2013

UE20CS312 - Data Analytics

Worksheet 1a - Part 2: EDA with R | ANOVA

SOLUTION SET

Harshith Mohan Kumar - harshithmohankumar@pesu.pes.edu

Yashas Kadambi - yashasks@pesu.pes.edu

Nishanth M S - nishanthmsathish.23@gmail.com

Anushka Hebbar - anushkahebbar@pesu.pes.edu

Prerequisites

To download the data required for this worksheet, visit this Github link. This worksheet has two parts, the first focuses on the basics of dealing with data and exploratory data analysis using R. The second deals with ANOVA. To help guide you through the worksheet, here are a few resources:

- Revise how to deal with DataFrames in R here.
- This online book has everything you need to get started with visualizations in R.
- Check out this resource for an excellent deep-dive of visualizations using the `ggplot2` library **(optional)**.
- The following are resources to learn about ANOVA:
  - Anova in Python
  - Anova in R

Part I. Exploratory Data Analysis with R

Book Club Marketing Dataset

Charles Book Club (CBC) is a book club that has an active database of 500,000 subscribers. The organization sends out monthly mailings to its database of members with the latest promotional offerings. Its marketing team would like to see if customer data can be used to reduce the cost of marketing activities to improve the profitability of their marketing operations. For an initial pilot of a predictive analytics solution, CBC decided to focus on its strongest customers and run a marketing test for a new book release of *'The Art History of Florence'*.

The dataset provided consists of information about customer purchases CBC has as its disposal after conducting the marketing test. Use the `CharlesBookClubDataset.csv` for Part I of the worksheet. This data was adapted from a famous business database called the 'Charles Book Club', dealt with in more detail in a case study from the 'Data Mining for Business Analytics' book.

Data Dictionary

```
ID#: Customer Identification number
Gender: Male, Female
M: Monetary - Total money spent on books
```

```
R: Recency - Months since last purchase
F: Frequency - Total number of purchases
FirstPurch: Months since first purchase
ChildBks: Number of purchases from category of child books |
YouthBks: Number of purchases from category of youth books
CookBks: Number of purchases from category of cook books
DoItYBks: Number of purchases from category of DIY books
RefBks: Number of purchases from category of reference books
ArtBks: Number of purchases from category of art books
GeoBks: Number of purchases from category of geography books
ItalCook: Number of purchases of book title 'Secrets of Italian Cooking'
ItalAtlas: Number of purchases of book title 'Historical Atlas of Italy'
ItalArt: Number of purchases of book title 'Italian Art'
Related  Purchase: Number of related books purchased
Florence: = 1 if 'Art History of Florence' was purchased; = 0 if not
```

Loading the Dataset

Use the following commands to load the dataset from CSV format and get a high-level overview of its fields:

```
library(tidyverse)
cbc_df <- read_csv(path_to_csv)
head(cbc_df)
```

Points

The problems for this part of the worksheet are for a total of 8 points, with a non-uniform weightage.

- *Problem 1* : 1 point
- *Problem 2* : 2 points
- *Problem 3* : 2 points
- *Problem 4.1* : 1 point
- *Problem 4.2* : 1 point
- *Problem 4.3* : 1 point

Problems

Problem 1 (1 point)

Generate an understanding of the dataset via a summary of its features. Find the count, missing count, minimum, 1st quartile, median, mean, 3rd quartile, max and standard deviation of all relevant columns. Separately, print the total number of missing values in each column.

Solution 1

```
library(tidyverse)
cbc_df <- read.csv('CharlesBookClubDataset.csv')
summary(cbc_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   :0.1253
##  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 print the number of missing values in each column,

```r
# Number of missing values in each column
colSums(is.na(cbc_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
```

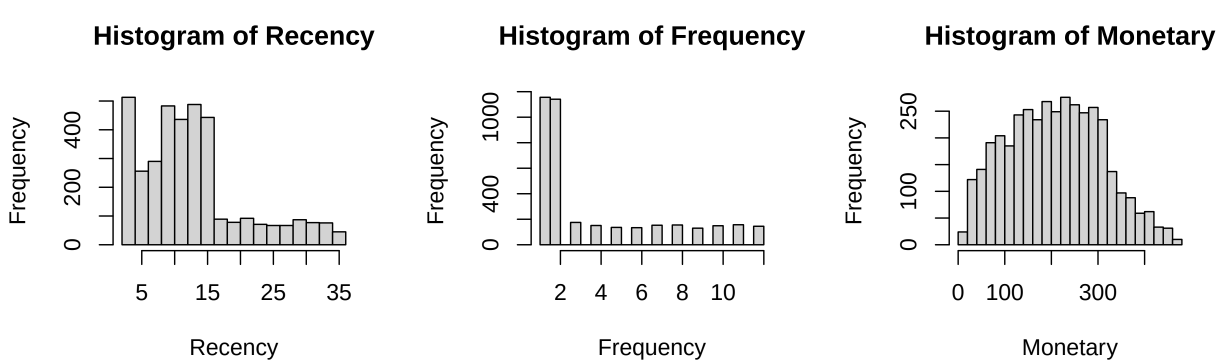
Problem 2 (2 points)

Replace missing values within the Recency, Frequency, and Monetary features with suitable values. Explain your reasoning behind the method of substitution used. *Hint:* Try plotting the distribution of the values in each feature using the `hist` function. Think about how to best deal with data imputation. Also, plot the distribution of feature values after imputation.

Solution 2

To figure out which measure of central tendency is to be used to impute missing values, plot the distribution of the feature values.

```r
# Function to plot the distribution of necessary features
plot_hist_20_bins <- function() {
  Recency <- cbc_df$R
  Frequency <- cbc_df$F
  Monetary <- cbc_df$M
  hist(Recency, breaks=20)
  hist(Frequency, breaks=20)
  hist(Monetary, breaks=20)
}

plot_hist_20_bins()
```



The output depicts positively skewed distributions; we'd be better off imputing missing values with the mode of the feature values.

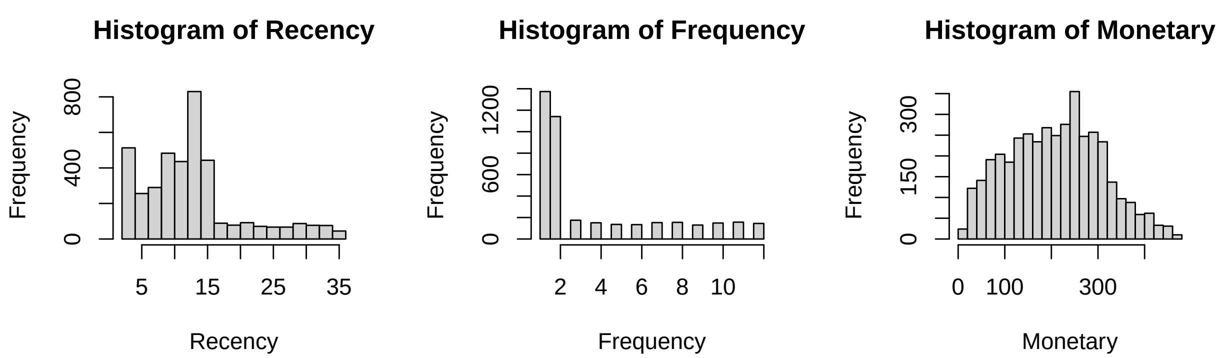
```r
# Function to calculate the mode of a feature
get_mode <- function(x) {
     mode0 <- names(which.max(table(x)))
     if(is.numeric(x)) return(as.numeric(mode0))
     mode0
}

# Apply to all necessary columns
cbc_df$R[is.na(cbc_df$R)] <- get_mode(cbc_df$R)
cbc_df$F[is.na(cbc_df$F)] <- get_mode(cbc_df$F)
cbc_df$M[is.na(cbc_df$M)] <- get_mode(cbc_df$M)
```

To check our results,

```r
# Number of missing values
colSums(is.na(cbc_df))
```

```
##                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
```

```r
# Plot histograms after imputation
plot_hist_20_bins()
```



Problem 3 (2 points)

Discretize the continuous values of Monetary, Recency, and Frequency into appropriate bins, and create three new columns `Mcode`, `Rcode` and `Fcode` respectively, for the discretized values. Explicitly mention the number of bins used and explain the choice for the bin size. Print out the summary of the newly created columns. *Hint:* Use the `cut` function to break on preset breakpoints. What are the most optimum breakpoints you can choose? Try to think of a statistical function that provides these breakpoints for optimum binning.

Solution 3

Create bins based on quantiles in every feature. This performs binning by setting every bin to have the same number of observations. For Recency, use 4 bins; setup 5 bins for Monetary and 3 bins for Frequency.

```
# Create new features that are a result of binning the previous ones
cbc_df <- cbc_df %>% mutate(Rcode=cut(cbc_df$R,
                                      breaks=unique(
                                      quantile(cbc_df$R,
                                      probs=seq.int(0,1,by=1/4))),
                                      include.lowest=TRUE),
                            Mcode=cut(cbc_df$M,
                                      breaks=unique(
                                      quantile(cbc_df$M,
                                      probs=seq.int(0,1,by=1/5))),
                                      include.lowest=TRUE),
                            Fcode=cut(cbc_df$F,
                                      breaks=unique(
                                      quantile(cbc_df$F,
                                      probs=seq.int(0,1,by=1/4))),
                                      include.lowest=TRUE))

# Set the level strings
levels(cbc_df$Mcode) <- c('$15-$112', '$112-$181', '$181-$242', '$242-$296', '$296-$479')
levels(cbc_df$Rcode) <- c('2-8 months', '8-14 months', '14-16 months', '16-36 months')
levels(cbc_df$Fcode) <- c('1-2 books', '2-6 books', '6-12 books')

summary(cbc_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
```

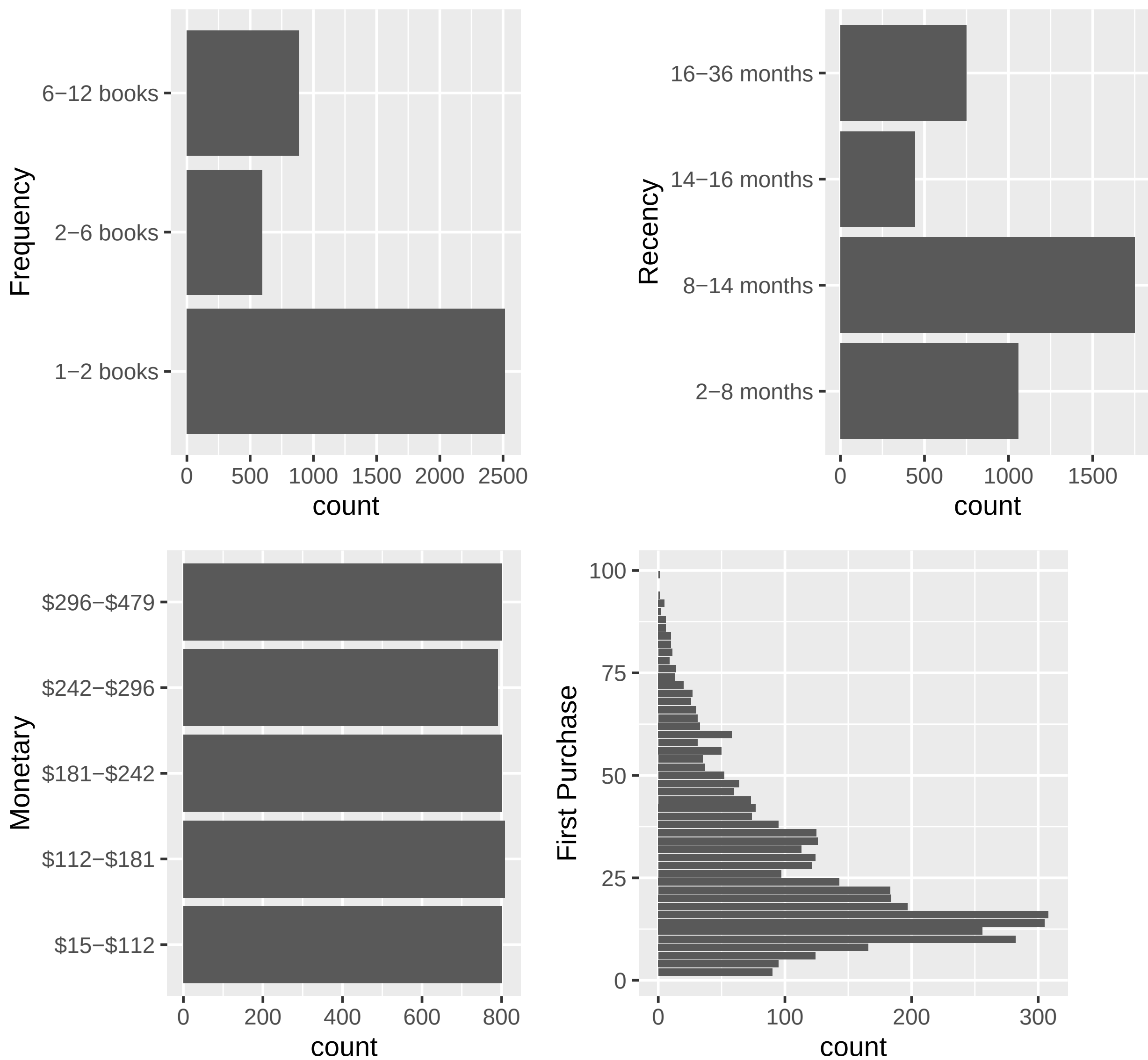
Problem 4

The marketing team heavily relies on the RFM variables of the recency of last purchase, total number of purchases, and total money spent on purchases to gauge the health of the members of the book club. Increases in either the frequency of purchases or monetary spend and decreases in time since last purchase across the customer base, will intuitively lead to more sales for the business.

4.1 Bar Graphs (1 point) Create and visualize histograms for the discretized Recency, Frequency, Monetary features. Also create one for the `FirstPurch` feature.

Solution 4.1 Code for bar graphs:

```
# Plot bar graphs
ggplot(cbc_df, aes(x = Fcode)) + geom_bar() + coord_flip () + labs(x = "Frequency")
ggplot(cbc_df, aes(x = Rcode)) + geom_bar() + coord_flip () + labs(x = "Recency")
ggplot(cbc_df, aes(x = Mcode)) + geom_bar() + coord_flip () + labs(x = "Monetary")
ggplot(cbc_df, aes(x = FirstPurch)) + geom_bar() + coord_flip () + labs(x = "First Purchase")
```



4.2 Box Plot (1 point) Transform the `Florence` variable into a categorical feature that can take up the values `True` or `False`. Create and visualize horizontal box plots for the original Recency, Frequence, Monetary and `FirstPurch` features against the `Florence` variable. *Hint:* To transform `Florence`, use the concept of factors in R and set the labels `True` and `False`.

Solution 4.2 Code for box plots:

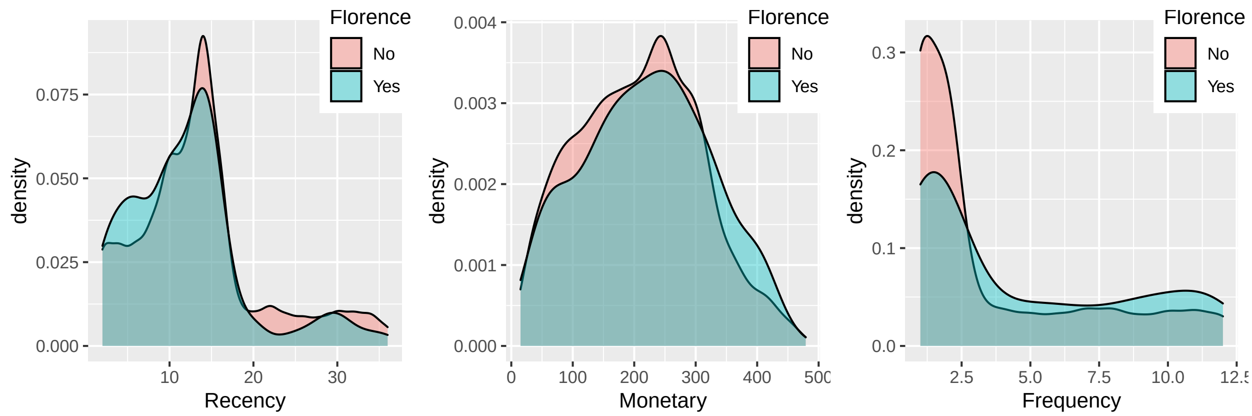
```
# Create a categorical feature for Florence
cbc_df$Florence <- factor(cbc_df$Florence, labels = c("No", "Yes"))

# Plot box plots
ggplot(cbc_df, aes_string(x = "Florence", y = "R", fill = "Florence")) +
geom_boxplot() +
coord_flip() +
labs(x = "Recency", y = "Did the customer make a purchase?") +
theme(legend.position = c(0.9, 0.9))
```

```r
ggplot(cbc_df, aes_string(x = "Florence", y = "M", fill = "Florence")) +
geom_boxplot() +
coord_flip() +
labs(x = "Monetary", y = "Did the customer make a purchase?") +
theme(legend.position = c(0.9, 0.9))

ggplot(cbc_df, aes_string(x = "Florence", y = "F", fill = "Florence")) +
geom_boxplot() +
coord_flip() +
labs(x = "Frequency", y = "Did the customer make a purchase?") +
theme(legend.position = c(0.9, 0.9))
```

4.3 Density Plot (1 point) Create and visualize a density plot for Recency, Frequency, Monetary and `FirstPurch` features.

Solution 4.3 Code for density plots:

```r
# Plot density plots
ggplot(cbc_df, aes_string(x = "R", fill = "Florence")) +
geom_density(alpha = 0.4) +
labs(x = "Recency") +
theme(legend.position = c(0.9, 0.9))

ggplot(cbc_df, aes_string(x = "M", fill = "Florence")) +
geom_density(alpha = 0.4) +
labs(x = "Monetary") +
theme(legend.position = c(0.9, 0.9))

ggplot(cbc_df, aes_string(x = "F", fill = "Florence")) +
geom_density(alpha = 0.4) +
labs(x = "Frequency") +
theme(legend.position = c(0.9, 0.9))
```



Part II. ANOVA

An Analysis of Variance Test, or ANOVA, can be thought of as a generalization of the t-tests for more than 2 groups. The independent t-test is used to compare the means of a condition between two groups. ANOVA is used when we want to compare the means of a condition between more than two groups. ANOVA tests if there is a difference in the mean somewhere in the model (testing if there was an overall effect), but it does not tell us where the difference is (if there is one). To find where the difference is between the groups, we have to conduct post-hoc tests.

To perform any tests, we first need to define the null and alternate hypothesis:

- **Null Hypothesis:** There is *no significant difference* among the groups.
- **Alternate Hypothesis:** There is a *significant difference* among the groups.

Points

The problems for this part of the worksheet are for a total of 6 points, with a non-uniform weightage.

- *Problem 1* : 2 points
- *Problem 2* : 3 points
- *Problem 3* : 1 point

Scenario 1

It's a brand new day in the 99th precinct of the New York Police Department. Lieutenant Terrance has had enough of Hitchcock and Scully being useless paper pushers and wanted to assign them work to help the investigations; they were assigned the duty of gaining insights from the different types of objects in the evidence log of an ongoing investigation focused on the New York Mafia.

Problems

Problem 1 (2 points)

Captain Holt provided a file containing the names of a few `People of Interest` and the number of items logged at various evidence lockers of various precincts pertaining to them. He also instructs Peralta and Diaz to look into the file as he was told it should contain more information.

Scully decided to use ANOVA.

For this problem, use the data file named `Scenario 1.csv` in the data repository. Load the following libraries before moving on and read the dataset,

```
library(ggpubr)
library(dplyr)
library(ggplot2)
library(ggpubr)
library(broom)
library(car)

data <- read.csv('Scenario 1.csv')
```

```
##        POI No.of.items
## 1    Sonny           38
## 2    Fredo           51
## 3  Micheal           41
## 4    Fredo           51
## 5    Fredo           58
## 6    Sonny           41
## 7    Sonny           51
## 8    Sonny           44
## 9    Fredo           52
## 10   Fredo           47
```

1. Consider the dataset. Which type of ANOVA can Scully use? (Justify why the particular test)
2. What function(s) could have been used by Scully for ANOVA if he uses the R programming language?
3. What does the output of this/these functions tell Scully? (Specify hypotheses and what each column in the summary of the output means considering 5% significance)

```
library(ggpubr)
library(dplyr)
library(ggplot2)
library(ggpubr)
library(broom)
library(car)
```

Solution 1

1. One-way Anova [Fisher's test]
2. `aov()`

```
scene.1.file <- read.csv('Scenario 1.csv')
one.way <- aov(No.of.items ~ POI, data = scene.1.file)
summary(one.way)
```

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

3. Description of each column. Hypotheses of One-way ANOVA test and 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 (3 points)

Peralta and Diaz find a member of the family, a certain Frank Pentangeli, through Doug Judy. They discovered that the *famiglia* had altered this file resulting in invalid results. The original file was then recovered by the squad and was sent to Scully and Hitchcock for analysis. To their surprise they discovered that the file also had additional column of which gives the priority.

The dataset has three columns:

- First column has the **Person of Interest(POI)** in the Mafia
- Second column has the number of evidence items collected in particular evidence locker (evidence lockers are present across the city and many precincts have multiple squads working on the mafia, so one POI has multiple entries).
- Third column gives the **Priority** given to collect the evidence by a particular squad with respect to a POI.

Read the dataset before moving on. For this problem, use the data file named `Scenario 2.csv` in the data repository.

```
data <- read.csv('Scenario 2.csv')
```

1. Consider the data. Which type of ANOVA can Scully use? (Justify why the particular test)
2. What function(s) could have been used by Scully for the ANOVA if he uses the R programming language?
3. What does the output of this/these functions tell Scully? (Specify hypotheses and what each column in the summary of the output means considering 5% significance)
4. Hitchcock thinks that Scully has missed a task which completes the ANOVA test. What should Scully have thought of? *Hint:* Philosophically, a hypothesis is a proposition made as a basis for reasoning, without any assumption of its truth.

Solution 2

1. Two-way Anova
2. `aov()`

```
scene.2.file <- read.csv('Scenario 2.csv')
two.way <- aov(No.of.items ~ POI * Priority, data = scene.2.file)
summary(two.way)
```

```
##                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. Description of each column. Hypotheses of Two way ANOVA test.

- since p value < 0.05 there is maybe a relation between the person of interest and the average number of evidence collected against them.

- since p value < 0.05 there is maybe a relation between the Priority and the average number of evidence collected against them.

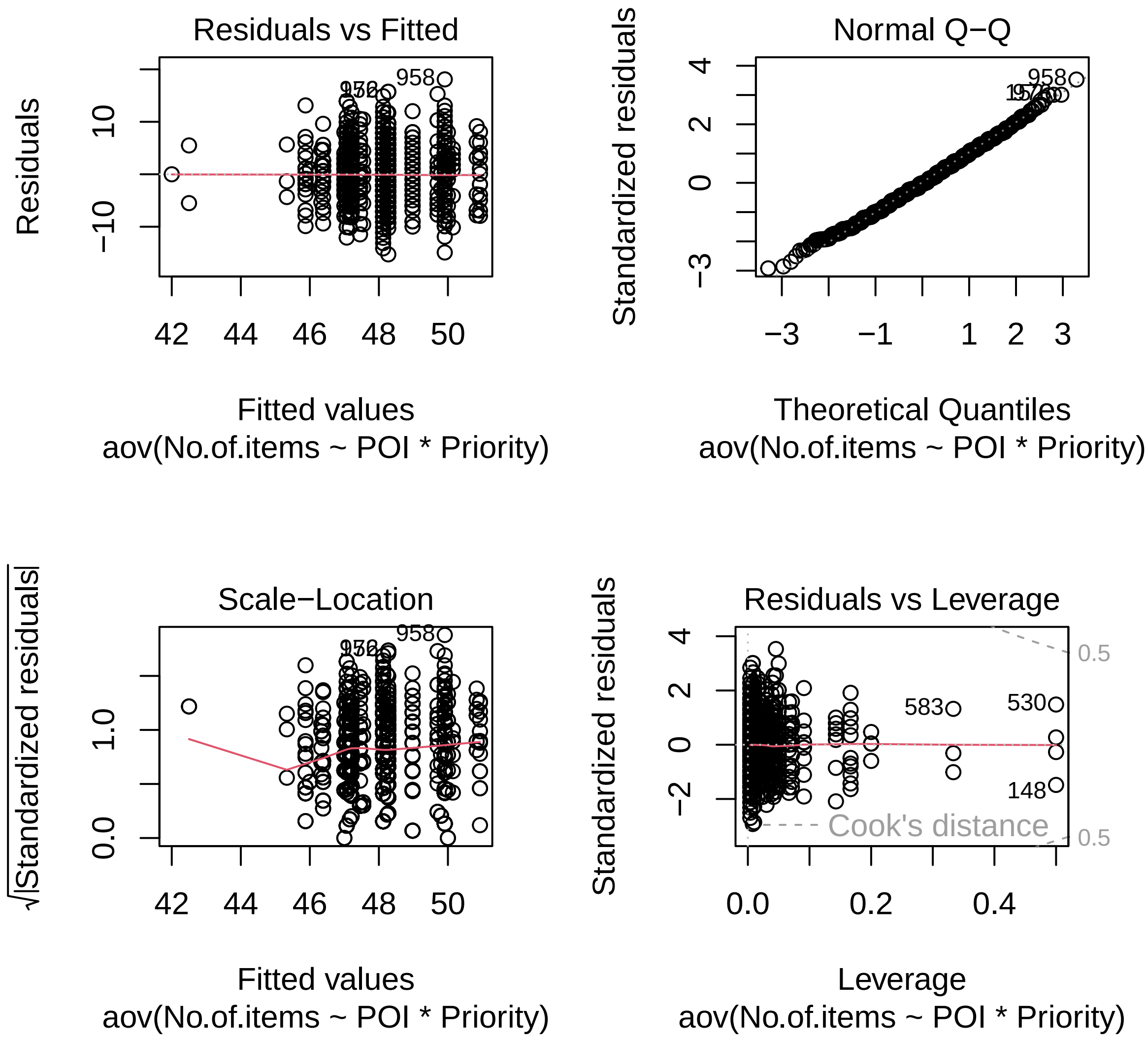
- Categorical variables cannot be compared with F Statistic and can only be ensured to be independent variables by experimental design. (**Wrong answer** since p value > 0.05 there is no interaction between the Priority and person of Interest.)

4. 3 assumptions of 2 way ANOVA are:

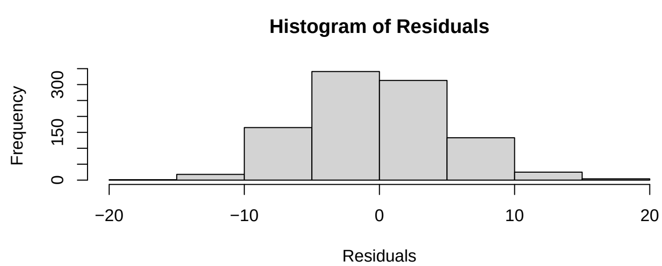
   - Homogeneity of variance (homoscedasticity) [Any one graph with brief explanation on why]
   - Normally-distributed dependent variable
   - Independence of observations: Categorical variables cannot be compared with F Statistic and can only be ensured to be independent variables by experimental design. (**Wrong answer** since p value > 0.05 there is no interaction between the Priority and person of Interest.)

```
plot(two.way)
```

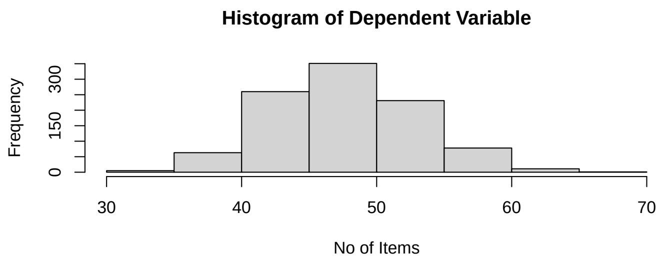
```
## Warning: not plotting observations with leverage one:
##   883
```




```
hist(two.way$residuals,main="Histogram of Residuals",xlab="Residuals")
```



```r
hist(scene.2.file$No.of.items,main="Histogram of Dependent Variable",xlab="No of Items")
```



Problem 3 (1 point)

Hitchcock also wanted to compare the number of items collected for each pair of Person of Interest and priority. He decided to follow the common practice of doing a **Tukey's HSD** . The Tukey's Honestly-Significant-Difference[TukeyHSD] test lets us see which groups are different from one another.

What insights did Hitchcock gain after doing the Tukey's HSD? (The `TukeyHSD` function can be used to do this test and the output of this function can be represented graphically using the `plot` function.)

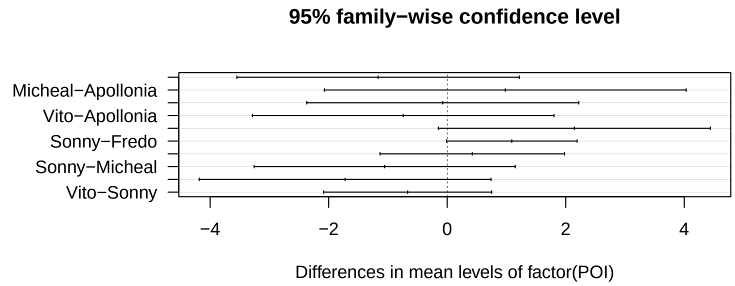
Solution 3

```r
tukey.two.way<-TukeyHSD(aov(formula = No.of.items ~ factor(POI) + Priority, data = scene.2.file))

tukey.two.way
```

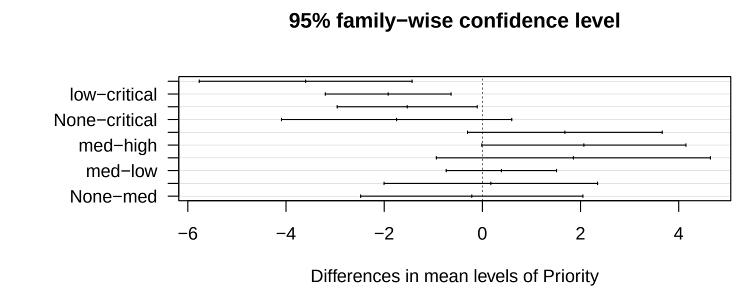
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = No.of.items ~ factor(POI) + Priority, data = scene.2.file)
##
## $`factor(POI)`
##                           diff           lwr       upr     p adj
## Fredo-Apollonia   -1.16541489 -3.5237413613 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
```

```r
par(mar=c(5,8,4,1)+.1)
tukey.plot.test<-TukeyHSD(aov(formula = No.of.items ~ factor(POI), data = scene.2.file))
plot(tukey.plot.test, las = 1)
```




```r
par(mar=c(5,8,4,1)+.1)
tukey.plot.test<-TukeyHSD(aov(formula = No.of.items ~ Priority, data = scene.2.file))
plot(tukey.plot.test, las = 1)
```



Reading the Graph: Any group which doesn't contain 0 in the confidence interval.

Here it can be seen that critical priority has a different mean compared to the other classes. This says that having a critical Priority assigned to working on the cases generate different no of evidence items compared to the rest of the priorities.

Also it can been seen that there is no pairs of POI has a statistically significant difference in mean no of evidence generated. In other words, there is no difference in the average no of Evidence items discovered when compared with any two POI.