Assignment1 - Exploratory Data Analysis with R

Geedhu Kizhakepura Velayudhan - Student_ID 8899510

2023-06-04

PART A

QUESTION 1

(4 points) You are working Streaming Service. The following statement is made by your manager. Based on the examples and discussion in Lecture 1, transform it into a question that can be answered with data analytics. Make sure you discuss the logic and reasoning you use to transform it and what problems you might define. (NOTE – This question is worth 4 marks, so answer appropriately). We have more customers than before, but our new customers are streaming less than before

Answer:

Logic and Reasoning:

To transform the manager's statement into questions, we need to focus on two elements mentioned:

1. the increase in the number of customers.
2. the decrease in streaming activity among new customers.

Problems to define:

1. What do you think is the possible cause of less streaming among new customers?
2. From when did you started noticing about the "streaming less activity" among new customers.
3. Have we established clear criteria for categorizing customers as either new or existing? For example, we could define "new customers" as those who signed up within the past three months and "existing customers" as those who have been subscribed for more than three months.
4. Did you collect any data about customer's streaming activities of both new and existing customers like time stamps of when customers start and stop streaming, the titles they watch, and the duration of each streaming session.If any already collected may we have the access to the data?
5. By what percentage would you expect the streaming activity to be increased?

QUESTION 2

2. (4 points) Consider the following three arrays of data. Each array is data for one file sharing site. The numbers in the array represent the number of downloads for each site in a day (for example, Site A had 28 downloads on the first day, 29 on the second and so on). Site A: (28 29 31 28 30 30 30 32 28 33) Site B: (23 19 23 33 32 27 20 24 42 32) Site C: (27 26 28 25 27 27 30 30 28 26) Based on the data provided, and using the skills learned in this class, answer the following questions. Make sure to provide evidence for your answers.

a) (2 points) Which site has the least downloads on a typical day?
b) (2 points) Which site has the most inconsistent usage?

Answer:

```
#Assigning variable to each site

sitea<- c(28, 29, 31, 28, 30, 30, 30, 32, 28, 33)
siteb<- c(23, 19, 23, 33, 32, 27, 20, 24, 42, 32)
sitec<- c(27, 26, 28, 25, 27, 27, 30, 30, 28, 26)


# a)
# By calculating mean of each site and then comparing them will provide the site has
# the site that has the least downloads on a typical day.

meanA<- mean(sitea)
meanB<- mean(siteb)
meanC<- mean(sitec)

# As Mean of Site C is the least, the Site C has the least downloads on a typical day.

# b)
# By calculating the variance of each site, we will get to know which site has the most inconsistent us

varA<- var(sitea)
varB<- var(siteb)
varC<- var(sitec)

#As Variance of Site B is the highest, the Site B has the most inconsistent usage.
```

a) (2 points) Which site has the least downloads on a typical day?

Answer: As Mean of Site C is the least, the Site C has the least downloads on a typical day.

b) (2 points) Which site has the most inconsistent usage?

Answer: As Variance of Site B is the highest, the Site B has the most inconsistent usage.

PART 2

1. BASIC MANIPULATION

```
###BASIC SETUP
# Clean workspace
rm(list=ls())
options(scipen=9)

getwd()
```

```
## [1] "C:/Users/Geedhu/Downloads"
```

```
# 1. Read data and converted to dataframe names 'data'

#Kindly uncomment the below to read data
data <- read.table(file = "Assign01.txt", header = TRUE, sep = ",")

# 2. Append your initials to all variables in the data frame

new_names <- paste0(names(data), "_kv")
colnames(data) <- new_names

#Check if initials are added or not
colnames(data)
```

```
## [1] "Manufacturer_kv" "Server_kv"       "DC_kv"           "SMBR_kv"
## [5] "SMBT_kv"         "Conn_kv"
```

```
#3. Change each character variable to a factor variable
data$Manufacturer_kv<- factor(data$Manufacturer_kv)
data$Server_kv <- factor(data$Server_kv)
data$DC_kv<-data$DC_kv
#4. Dimensions of the dataset (rows and columns)?

dimensions <- dim(data)
num_rows<- dimensions[1]
num_cols<- dimensions[2]
#Print dimensions,rows and columns
cat("Dimensions of dataset: ",dimensions)
```

```
## Dimensions of dataset:  90000 6
```

```
cat("Number of Rows: ",num_rows)
```

```
## Number of Rows:  90000
```

```
cat("Number of Columns: ",num_cols)
```

```
## Number of Columns:  6
```

2. SUMMARIZING DATA

```
# a. Calculate the mean and standard deviation for Server Message Blocks Received.

data$SMBR_kv<- as.numeric(data$SMBR_kv)
Mean_SMBR<- mean(data$SMBR_kv)
var_SMBR<- var(data$SMBR_kv)
sd_SMBR<- sd(data$SMBR_kv)
```

```r
#Print mean, variance and standard deviation of SMBR

cat("Mean of SMBR: ",Mean_SMBR)
```

1.Means and Standard Deviations

```
## Mean of SMBR:  100017.5
```

```r
cat("Variance of SMBR: ",var_SMBR)
```

```
## Variance of SMBR:  100049172
```

```r
cat("Standard deviation of SMBR: ",sd_SMBR)
```

```
## Standard deviation of SMBR:  10002.46
```

```r
# b. Use the results above to calculate the coefficient of variation (rounded to 3 decimal places).

cv_SMBR<- round((sd_SMBR/Mean_SMBR),3)

cat("Coefficient of Variation: ",cv_SMBR)
```

```
## Coefficient of Variation:  0.1
```

```r
# c. Calculate the mean and standard devia,on for Server Message Blocks Transmitted.

data$SMBT_kv<- as.numeric(data$SMBT_kv)
mean_SMBT<- mean(data$SMBT_kv)
var_SMBT<- var(data$SMBT_kv)
sd_SMBT<- sd(data$SMBT_kv)

cat("Mean of SMBT: ",mean_SMBT)
```

```
## Mean of SMBT:  49966
```

```r
cat("Variance of SMBT: ",var_SMBT)
```

```
## Variance of SMBT:  100489304
```

```r
cat("Standard deviation of SMBT: ",sd_SMBT)
```

```
## Standard deviation of SMBT:  10024.44
```

```r
# d. Also calculate the coefficient of variation (rounded to 3 decimal places).
cv_SMBT<- round((sd_SMBT/mean_SMBT),3)

cat("Coefficient of Variation of SMBT is ",cv_SMBT)
```

```
## Coefficient of Variation of SMBT is  0.201
```

e. Does the SMBT or SMBR have more variation?

Answer : SMBT has more variation as CV is higher for SMBT than SMBR.

```r
num_values <- length(data$SMBT_kv)
Percentile_SMBT <- round((0.45*(num_values + 1)),0)

cat("45th percentile is: ",Percentile_SMBT)
```

2. Calculate the 45th percen,le of the number of Server Message Blocks TransmiSed. This calcula,on should be rounded to the nearest whole number (no decimal places).

```
## 45th percentile is:  40500
```

ORGANIZING DATA

1.Summary Table

a) Create a table showing the average Server Message Blocks Transmi5ed by Manufacturer. This should be rounded to two decimal places.

```r
# Creating the summary table
summary_table <- aggregate( data$SMBT_kv ~ data$Manufacturer_kv ,FUN = function(x) round(mean(x), 2))

# Naming the columns
colnames(summary_table) <- c("Manufacturer", "Average SMBT")

#Print the summary table
summary_table
```

```
##        Manufacturer Average SMBT
## 1 Highway-Passenger     49916.14
## 2              Lled     50008.12
## 3            Ovonel     49973.76
```

b) Which Manufacturer's Servers have, on average, transmitted the most server message blocks?

Answer: Lled is the Manufacturer who transmitted the most server message blocks.

2.Cross Tabulation

a) Create a table counting all Servers by Data Centre.

```r
#Create the cross table
cross_table <- table(data$Server_kv, data$DC_kv)
# Print the cross table
cross_table
```

```
##
##          Bridgeport Cambridge Elmira Kitchener Waterloo
##   DF6726       2971      4385   5869      7363     8827
##   DJ3756         60        87    118       157      163
```

```
##    MB3406       2188      3433   4534      5634     6882
##    MG9696        719      1128   1435      1810     2237
##    RQ8547       2082      3184   4161      5248     6421
##    RX8838        925      1365   1734      2191     2689
```

b) Change the table to show the percentage of each Server in each Data Centre.This should be rounded to three decimal places.

```
#Create Percentable table
percentage_table <- round(prop.table(cross_table, margin = 2) * 100,3)
# Print Percentage table
percentage_table
```

```
##
##          Bridgeport Cambridge Elmira Kitchener Waterloo
##    DF6726     33.214    32.285 32.878    32.866   32.430
##    DJ3756      0.671     0.641  0.661     0.701    0.599
##    MB3406     24.461    25.276 25.399    25.148   25.284
##    MG9696      8.038     8.305  8.039     8.079    8.219
##    RQ8547     23.276    23.443 23.310    23.425   23.590
##    RX8838     10.341    10.050  9.714     9.780    9.879
```

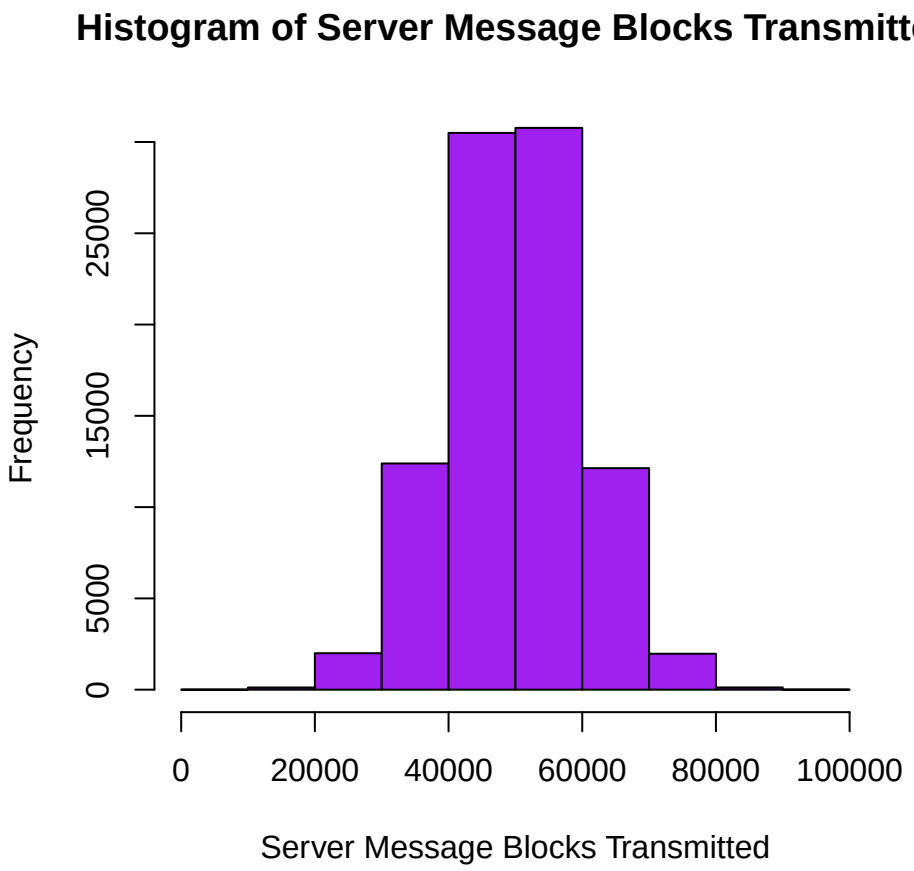
c) What percentage of servers at Elmira are MG9696?

```
servers_at_elmira <- percentage_table["MG9696","Elmira"]
```

3.Histogram

a) Create a histogram of Server Message Blocks Transmitted.
b) The plot should be properly labelled and a unique colour and have 10 breaks. For both a and b , below is the code:

```
hist(data$SMBT_kv,
     xlab = "Server Message Blocks Transmitted",
     ylab = "Frequency",
     main = "Histogram of Server Message Blocks Transmitted",
     col = "purple",
     breaks = 10)
```



c) Which range of SMBT is the most common?

Answer: The most comman range in SMBT is 50000 to 60000.