EDA on Mail Delivery Data

Divyang Vinubhai Hirpara

01/31/2023

*It will clear all the plots, the console and the workspace.* *It also sets the overall format for numbers.*

## PART 1:

#### 1

**Statement: “We have more customers than before, but our new customers are streaming less than before.”**

There are some questions raised for data analysis which are mentioned below.  
1. Who are the new customers? What are the age categories and gender?  
2. What is the most popular streaming content in the system?  
3. When compared to regular consumers, how do new customers stream on average?  
4. Was streaming content free before and now paid or partial pay for premium content?  
5. Is a decline in streaming activity specific to new customers, or is it a broader trend across the entire customer base?

#### 2

*Hint, Site A had 28 downloads on the first day, 29 on the second and so on*

site\_A\_DVH <- c(28, 29, 31, 28, 30, 30, 30, 32, 28, 33)  
site\_B\_DVH <- c(23, 19, 23, 33, 32, 27, 20, 24, 42, 32)  
site\_C\_DVH <- c(27, 26, 28, 25, 27, 27, 30, 30, 28, 26)  
  
cat("Downoads of Site A: ",site\_A\_DVH,'\n')

## Downoads of Site A: 28 29 31 28 30 30 30 32 28 33

cat("Downoads of Site B: ",site\_B\_DVH,"\n")

## Downoads of Site B: 23 19 23 33 32 27 20 24 42 32

cat("Downoads of Site C: ",site\_C\_DVH,"\n")

## Downoads of Site C: 27 26 28 25 27 27 30 30 28 26

###### A). Which site has the least downloads on a typical day?

Least\_dwnld\_site\_A\_DVH <- min(site\_A\_DVH)  
cat("The Least Download of site A is: ",Least\_dwnld\_site\_A\_DVH,"\n")

## The Least Download of site A is: 28

Days\_site\_A\_DVH <- which(site\_A\_DVH == Least\_dwnld\_site\_A\_DVH)  
cat("Downloads of site A were low on this days: ",Days\_site\_A\_DVH,"\n\n")

## Downloads of site A were low on this days: 1 4 9

Least\_dwnld\_site\_B\_DVH <- min(site\_B\_DVH)  
cat("The Least Download of site B is: ",Least\_dwnld\_site\_B\_DVH,"\n")

## The Least Download of site B is: 19

Days\_site\_B\_DVH <- which(site\_B\_DVH == Least\_dwnld\_site\_B\_DVH)  
cat("Downloads of site B were low on this day: ",Days\_site\_B\_DVH,"\n\n")

## Downloads of site B were low on this day: 2

Least\_dwnld\_site\_C\_DVH <- min(site\_C\_DVH)  
cat("The Least Download of site C is: ",Least\_dwnld\_site\_C\_DVH,"\n")

## The Least Download of site C is: 25

Days\_site\_C\_DVH <- which(site\_C\_DVH == Least\_dwnld\_site\_C\_DVH)  
cat("Downloads of site C were low on this day: ",Days\_site\_C\_DVH,"\n\n")

## Downloads of site C were low on this day: 4

**Interpretation**  
This problem is solved by finding minimum value of each site data set of site A, site B, site C.  
Site A: 28  
Site B: 19  
Site C: 25  
Therefore, Site B has least downloads on typical day with 19 downloads on day 2.

###### B). Which site has the most inconsistent usage?

cat("CV of Site A: ",sd(site\_A\_DVH)/mean(site\_A\_DVH),"\n")

## CV of Site A: 0.05782075

cat("CV of Site B: ",sd(site\_B\_DVH)/mean(site\_B\_DVH),"\n")

## CV of Site B: 0.2606765

cat("Cv of Site C: ",sd(site\_C\_DVH)/mean(site\_C\_DVH),"\n")

## Cv of Site C: 0.06009289

**Interpretation**  
To check the most inconsistent usage of the site, we need to check the Coefficient of Variation of each site. The high Coefficient of Variation of the site is considered the most inconsistent site usage.

CV = standard Deviation/mean

The above output clearly shows that Site B has a high Coefficient of Variation. Hence Site B has the most inconsistent usage.

## PART 2:

#### 1. Basic Manipulation

###### 1. Read in the text file and change to a data frame

data\_DVH <- read.table("PROG8430-23W-Assign01.txt",sep=",",header = TRUE)  
head(data\_DVH)

## Manufacturer Server DC SMBR SMBT Conn  
## 1 Lled MG9696 Waterloo 102479 43473 6625  
## 2 Ovonel RX8838 Waterloo 103678 62534 7580  
## 3 Lled MB3406 Cambridge 102003 35916 5957  
## 4 Lled MB3406 Kitchener 98889 40245 6120  
## 5 Highway-Passenger DF6726 Cambridge 104907 25422 5839  
## 6 Highway-Passenger DF6726 Kitchener 102659 53168 7076

**Interpretation**  
Text file shall be read with ‘read.table’ function in R.  
Text file is comma separated hence, sep=” ,” is used to identify a rows and column.  
header=TRUE is used due to the text file is generated with header in first line.  
By default, 6 records are displayed with ‘head’ function as shown above.  
There are total 6 columns with manufacturer, server, dc of Character data type and smbr, smbt, conn of integer datatype.

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

#Append data\_DVH initials to column names  
colnames(data\_DVH) <- paste(colnames(data\_DVH), "DVH", sep = "\_")  
head(data\_DVH,10)

## Manufacturer\_DVH Server\_DVH DC\_DVH SMBR\_DVH SMBT\_DVH Conn\_DVH  
## 1 Lled MG9696 Waterloo 102479 43473 6625  
## 2 Ovonel RX8838 Waterloo 103678 62534 7580  
## 3 Lled MB3406 Cambridge 102003 35916 5957  
## 4 Lled MB3406 Kitchener 98889 40245 6120  
## 5 Highway-Passenger DF6726 Cambridge 104907 25422 5839  
## 6 Highway-Passenger DF6726 Kitchener 102659 53168 7076  
## 7 Highway-Passenger DF6726 Elmira 106037 59596 7258  
## 8 Lled MB3406 Waterloo 101077 64132 7391  
## 9 Lled MG9696 Cambridge 101662 42928 6608  
## 10 Ovonel RQ8547 Kitchener 90592 61989 6671

**Interpretation**  
Every column are replaced with initials.  
Manufacturer –> Manufacturer\_DH  
Server –> Server\_DH  
DC –> DC\_DH  
SMBR –> SMBR\_DH  
SMBT –> SMBT\_DH  
Conn –> Conn\_DH

###### 3. Change each character variable to a factor variable.

data\_DVH$Manufacturer\_DVH <- as.factor(data\_DVH$Manufacturer\_DVH)  
data\_DVH$Server\_DVH <- as.factor(data\_DVH$Server\_DVH)  
data\_DVH$DC\_DVH <- as.factor(data\_DVH$DC\_DVH)  
  
head(data\_DVH,10)

## Manufacturer\_DVH Server\_DVH DC\_DVH SMBR\_DVH SMBT\_DVH Conn\_DVH  
## 1 Lled MG9696 Waterloo 102479 43473 6625  
## 2 Ovonel RX8838 Waterloo 103678 62534 7580  
## 3 Lled MB3406 Cambridge 102003 35916 5957  
## 4 Lled MB3406 Kitchener 98889 40245 6120  
## 5 Highway-Passenger DF6726 Cambridge 104907 25422 5839  
## 6 Highway-Passenger DF6726 Kitchener 102659 53168 7076  
## 7 Highway-Passenger DF6726 Elmira 106037 59596 7258  
## 8 Lled MB3406 Waterloo 101077 64132 7391  
## 9 Lled MG9696 Cambridge 101662 42928 6608  
## 10 Ovonel RQ8547 Kitchener 90592 61989 6671

**Interpretation**  
as.factor(variable): used to change data type into factor value.  
From above change of char variables to factor variables.  
Character type (chr) changed to –> (fctr)

###### 4. What are the dimensions of the dataset (rows and columns)?

dim(data\_DVH)

## [1] 90000 6

**Interpretation**  
In order to find number of rows and column in dataset, dim() function is used.  
There are 90000 row and 6 column are present in table.

#### 2. Summarizing Data

###### 1. Means and Standard Deviations

1. Calculate the mean and standard deviation for Server Message Blocks Received.

mean\_SMBR\_DVH <- mean(data\_DVH$SMBR\_DVH)  
print(paste0("Mean Of Server Message Blocks Received: ",mean\_SMBR\_DVH))

## [1] "Mean Of Server Message Blocks Received: 100017.478544444"

sd\_SMBR\_DVH <- sd(data\_DVH$SMBR\_DVH)  
print(paste0("Standard Deviation Of Server Message Blocks Received: ",sd\_SMBR\_DVH))

## [1] "Standard Deviation Of Server Message Blocks Received: 10002.4583223398"

**Interpretation**  
mean() function is used to calculate average of values.  
sd() function is used to calculate standard deviation of values.  
MEAN of SMBR: 100017.18  
SD of SMBR: 10002.46

1. Use the results above to calculate the coefficient of variation (rounded to 3 decimal places).

cv\_SMBR\_DVH <- sd\_SMBR\_DVH/mean\_SMBR\_DVH  
cv\_SMBR\_DVH <- round(cv\_SMBR\_DVH,3)  
print(paste0("coefficient of Variation Of Server Message Blocks Received: ",cv\_SMBR\_DVH))

## [1] "coefficient of Variation Of Server Message Blocks Received: 0.1"

**Interpretation**  
Coefficient Variation is calculated by dividing mean of SMBR from standard deviation of SMBR.Value of mean and sd is taken from previous answer.  
round() function returns floating point number is rounded to desire number. Here answer is rounded with 3 as per instruction.  
CV of SMBR: 0.1

1. Calculate the mean and standard deviation for Serve Message Blocks Transmitted.

mean\_SMBT\_DVH <- mean(data\_DVH$SMBT\_DVH)  
print(paste0("Mean Of Server Message Blocks Trasmitted: ",mean\_SMBT\_DVH))

## [1] "Mean Of Server Message Blocks Trasmitted: 49966.0049333333"

sd\_SMBT\_DVH <- sd(data\_DVH$SMBT\_DVH)  
print(paste0("Standard Deviation Of Server Message Blocks Transmitted: ",sd\_SMBT\_DVH))

## [1] "Standard Deviation Of Server Message Blocks Transmitted: 10024.435354702"

**Interpretation**  
MEAN of SMBT: 49966.0049  
SD of SMBT: 10024.4353

1. Also calculate the coefficient of variation (rounded to 3 decimal places).

cv\_SMBT\_DVH <- round(sd\_SMBT\_DVH/mean\_SMBT\_DVH,3)  
print(paste0("coefficient of Variation Of Server Message Blocks Transmitted: ",cv\_SMBT\_DVH))

## [1] "coefficient of Variation Of Server Message Blocks Transmitted: 0.201"

**Interpretation**  
CV of SMBT: 0.201

1. Does the SMBT or SMBR have more variation?

**Interpretation**  
Yes, SMBT(0.201) has more variation compared to SMBR(0.1). from the above results of SMBT and SMBR.

###### 2. Calculate the 45th percentile of the number of Server Message Blocks Transmitted. This calculation should be rounded to the nearest whole number (no decimal places).

percentile45\_smbt\_DVH <- quantile(data\_DVH$SMBT\_DVH, 0.45)  
print(paste0("The 45th percentile of the number of Server Message Blocks Transmitted: ",round(percentile45\_smbt\_DVH)))

## [1] "The 45th percentile of the number of Server Message Blocks Transmitted: 48741"

**Interpretation**  
quantile() is used to find 45th percentile of smbt.  
The 45th percentile of the number of Server Message Blocks Transmitted: 48741

####2. Organizing Data ######1. Summary Table  
a. Create a table showing the average Server Message Blocks Transmitted by Manufacturer. This should be rounded to two decimal places.

ServerManufacture\_DVH <- aggregate(data\_DVH$SMBT\_DVH,   
 by=list(data\_DVH$Manufacturer\_DVH),   
 FUN=mean,   
 na.rm=TRUE)  
colnames(ServerManufacture\_DVH) <- c("Manufacture","Average SMBT")  
ServerManufacture\_DVH

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

**Interpretation**  
aggregate() function is used with appropriate arguments to find average smbt by manufacture. where, fun=mean represents average of smbt, by=list(data\_DVH$Manufacturer\_DVH) represents average by manufacture.

1. Which Manufacturer’s Servers have, on average, transmitted the most server message blocks? Which manufacturer is it?

**Interpretation**  
From the previous result, Lled has most transmitted server message which is 50008.12.

###### 2. Cross Tabulation

1. Create a table counting all Servers by Data Centre

ServerDC\_DVH <- table(data\_DVH$Server\_DVH,data\_DVH$DC\_DVH)  
ServerDC\_DVH

##   
## 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

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

Note: 1 is set as margin parameter  
1 –> value divide by row sum 2 –> Value divide by Column sum

Percent\_ServerDC\_DVH <- round(prop.table(ServerDC\_DVH,2),3)  
Percent\_ServerDC\_DVH

##   
## Bridgeport Cambridge Elmira Kitchener Waterloo  
## DF6726 0.332 0.323 0.329 0.329 0.324  
## DJ3756 0.007 0.006 0.007 0.007 0.006  
## MB3406 0.245 0.253 0.254 0.251 0.253  
## MG9696 0.080 0.083 0.080 0.081 0.082  
## RQ8547 0.233 0.234 0.233 0.234 0.236  
## RX8838 0.103 0.101 0.097 0.098 0.099

1. What percentage of servers at Elmira are MG9696?

Percent\_ServerDC\_DVH[,'Elmira']['MG9696']

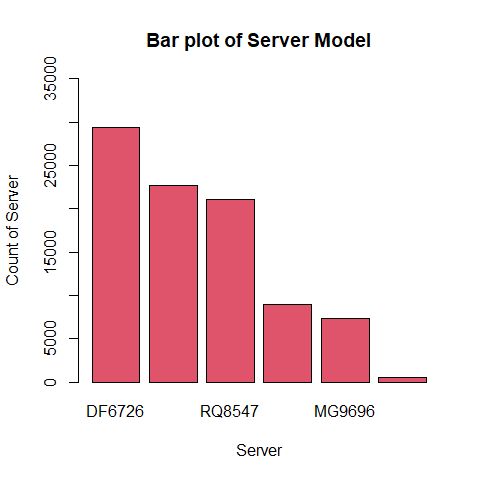
## MG9696   
## 0.08

**Interpretation**  
Percentage of Server MG9696 at Elmira is 8%.

###### 3. Bar Plot

1. Create a bar plot of count of Servers Models.
2. The plot should be:
   1. Rank ordered by highest count of Server Model.
   2. Properly labeled (title, x-axis, etc)
   3. The bars should have a different colour than the one shown in class.

Server\_DVH <- table(data\_DVH$Server\_DVH)  
Server\_DVH <- Server\_DVH[order(Server\_DVH,decreasing = TRUE)]  
barplot(Server\_DVH,  
 main="Bar plot of Server Model",  
 ylab="Count of Server",  
 xlab = "Server",  
 col = 2,  
 ylim =c(0,35000)  
 )



**Interpretation**  
From above Bar chart, count of server varies from 0 to 35000. All the server are displayed as ascending order. There are total 6 servers available in bar char with Red color.

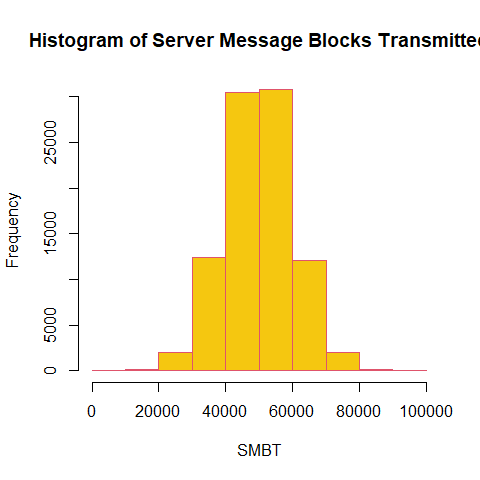
1. Based on the bar plot, (approximately) how many of Server RX8838 are there?

**Interpretation**  
Based on the Bar Plot, there are around 9000 RX8838 server shown.

###### 4. Histogram

1. Create a histogram of Server Message Blocks Transmitted.
2. The plot should be properly labeled and a unique colour and have 10 breaks.

hist(data\_DVH$SMBT\_DVH,  
 col=7,  
 border = 2,  
 main="Histogram of Server Message Blocks Transmitted",  
 xlab = "SMBT",  
 breaks = 10  
 )

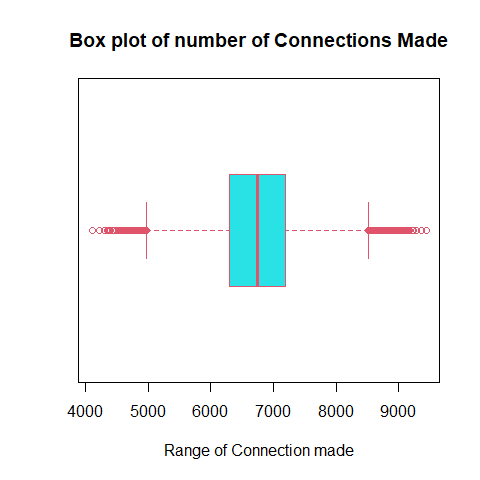


1. Which range of SMBT is the most common?  
   *Ans. From the above graph, most common range of SMBT is between 40000 to 60000.*

###### 5. Box Plot

1. Create a horizontal box plot of number of Connections Made.
2. The plot should be properly labeled and a unique colour.

boxplot(data\_DVH$Conn\_DVH,  
 col=5,  
 border = 2,  
 horizontal = TRUE,  
 main = "Box plot of number of Connections Made",  
 xlab = "Range of Connection made"  
 )



1. Based on the box plot, approximately how many servers made fewer than 6160 connections?

sum(data\_DVH$Conn\_DVH < 6160)

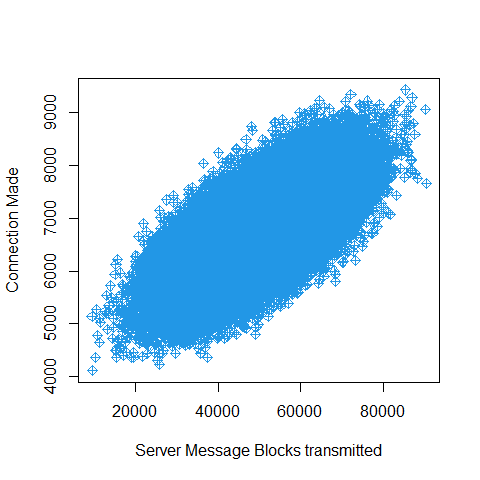
## [1] 16437

**Interpretation**  
Number of Server fewer than 6160 is 16437.

###### 6. Scatter Plot x

1. Create a scatter plot comparing Server Message Blocks Transmitted and Connections Made.
2. The plot should be properly labeled with a marker type different than the one demonstrated in class.

plot(data\_DVH$SMBT\_DVH,data\_DVH$Conn\_DVH,  
 pch=9,  
 xlab = "Server Message Blocks transmitted",  
 ylab = "Connection Made",  
 col=4  
 )



1. Does there appear to be an association between Server Message Blocks Transmitted and Connections Made?

**Interpretation**  
As from plot, SMBT are positively increasing with connection made. There is a association between SMBT and connection made.