#Following the CRISP-DM methodology:

 # 1. Business understanding- Business Problem:
**
#"Need to develop a Market mix model to observe the actual impact of different marketing variables over the last year.
Using the understanding, we have to recommend the optimal budget allocation for different marketing levers for the next year"
#
"To create market mix models for three product sub-categories - camera accessory, home audio and gaming accessory.
**Also, the models have to be built at a weekly level"
#
2. Data understanding
****
-Explain the data, sources, and results from exploratory analysis

3. Data Preparation
****
-Briefly explain the data preparation process
#
4. Data Modeling
****
Planned to develop 5 models
#
5. Model Evaluation
****
#
Insights
-Explain the final conclusion or insights derived from the model
#
##
 # Data Cleaning and Data PREPERATION STEPS FOLLOWED #
##

1. Identify the columns having "na"
Omitted the records having "na"

2. Formatted the order_date to match our defined format for further analysis

3. Removed the records that are not in our analysis period "July 15 to June 16" - Outlier identification

4. Handled the missing values

5. Calcuated the list price of each units from Order level data

6. Calcuated the discount of each units from Order level data

```
# 7. Framed a holiday date for analysis Period from Media Data spreadsheet

# 8. Translate the transaction data and Media monthly level data to weekly granular level

# 9. Extracted the records those are only from the holiday period from Order level data

# 10. Extracted the records based on the three sub-category per the Project SCOPE
   #"Camera_accessory", "home_audio" & "gaming_accessory" from Order level data

# 11. Frame the final Dataset for the model building by merging the Media and Order level Dataset at
weekly level
   # from  Order level data and the Media Data and other info spreadsheet

##############################################################################################
#                                        LOAD THE REQUIRED LIBRARIERS                        #
##############################################################################################
library("gdata")
library("ggplot2")
#install.packages("magrittr")
library(magrittr)
library(lubridate)
library(stringr)
library(dplyr)
library(MASS)
library(car)
library(GGally)
library(scales)
library(DataExplorer)
#library("gdata")

##############################################################################################
#                                        LOAD THE DATASETS                                   #
##############################################################################################

#Load the Order Level Data
capstone <- read.csv("ConsumerElectronics.csv", stringsAsFactors = F)
capstonebk <- capstone

#View(capstone)
cap1 <- capstone %>% filter(product_analytic_sub_category == 'CameraAccessory')
names(cap1)
unique(cap1$product_analytic_vertical)

#Load the Media Investment information
mediainvestment <- read.xls("Media data and other information.xlsx", sheet = 2, header = TRUE, skip =
1)
head(mediainvestment)
```

```
#Processing the Month column to be double digit for further processing
for (i in 1:length(mediainvestment$Month)) {
 if(nchar(mediainvestment[i,2])=="1")
 {
  mediainvestment$Month[i] <- assign(paste0(mediainvestment$Month[i]), paste0('0',
mediainvestment[i,2]))
 }
 else{
  print("No change")
 }
}

#Further processing the Media investment data
mediainvestment$year_month <- paste(mediainvestment$Year , mediainvestment$Month , sep = '_')
mediainvestment$Year <- as.numeric(mediainvestment$Year)
mediainvestment$Month <- as.numeric(mediainvestment$Month)
mediainvestment$year_month <- factor(mediainvestment$year_month, levels =
mediainvestment$year_month[order(mediainvestment$Year , mediainvestment$Month)])


#############################################################################################
#                                    DATA UNDERSTANDING                                     #
#############################################################################################

##  From the below plot we can say that expense is lowest for august and expense is high in
Sep,Oct,Dec,Mar
ggplot(mediainvestment, aes(x = year_month, y = Total.Investment)) + geom_bar(stat = "identity")


##Load the Product List information
productlist <- read.xls("Media data and other information.xlsx", sheet = 1, header = TRUE, skip = 0)
head(productlist)
head(productlist)

##Load the Monthly NPS Score information
npsscore <- read.xls("Media data and other information.xlsx", sheet = 4, header = TRUE)


#############################################################################################
#                                    DATA PROCESSINGTS                                      #
#############################################################################################
Process the Monthly NPS Score information from spreadsheet for further analysis
npsscore <- npsscore[2:13]
npsscore <- as.data.frame(t(npsscore))
npsscore$Month <- c(seq(7,12,1),seq(1,6,1))
colnames(npsscore)[1] <- "NPS"

#ggplot(productlist, aes(x = X, y = Frequency)) + geom_bar(stat = "identity")
```

#Process the Media Investment information to be converted to daywise

```
mediainvestment_daywise <- mediainvestment
mediainvestment_daywise$Total.Investment <-
(mediainvestment_daywise$Total.Investment*10000000)/30
mediainvestment_daywise$TV <- (mediainvestment_daywise$TV*10000000)/30
mediainvestment_daywise$Digital<- (mediainvestment_daywise$Digital*10000000)/30
mediainvestment_daywise$Sponsorship<- (mediainvestment_daywise$Sponsorship*10000000)/30
mediainvestment_daywise$Content.Marketing<-
(mediainvestment_daywise$Content.Marketing*10000000)/30
mediainvestment_daywise$Online.marketing<-
(mediainvestment_daywise$Online.marketing*10000000)/30

mediainvestment_daywise$X.Affiliates<- (mediainvestment_daywise$X.Affiliates*10000000)/30
mediainvestment_daywise$SEM<- (mediainvestment_daywise$SEM*10000000)/30
mediainvestment_daywise$Radio<- (mediainvestment_daywise$Radio*10000000)/30
mediainvestment_daywise$Other<- (mediainvestment_daywise$Other*10000000)/30
```

#Data imputation for NA values

```
mediainvestment_daywise[which(is.na(mediainvestment_daywise$Radio)), "Radio"] <- 0
mediainvestment_daywise[which(is.na(mediainvestment_daywise$Other)), "Other"] <- 0
```

#Extract the Date as "Year-Month-Day" from the Order data provided Data format "Year-Month-Day Hr:mn:ss"

#This mandatory for the code to train the given date format is YYYY-mm-dd. This can be used to fetch the date data per the analysis requirement as "Yr" "YEAR" etc

```
capstone$Odate <- as.Date(capstone$order_date,format = "%Y-%m-%d")

#Add a new column "yrmonth" using  "%Y_%m"
capstone$yrmonth <- format(capstone$Odate,"%Y_%m")

capstone_bk <- capstone
#capstone <- capstone_bk

dim(capstone)
head(capstone)
names(capstone)
str(capstone)
summary(capstone)

#Process the weeknumber to have the continuity from July 2015 to June 2016 as "1-53"
#View(capstone$weeknumber)
capstone$weeknumber <- week(capstone$order_date)
capstone$weeknumber <- ifelse(capstone$Year ==
2016,capstone$weeknumber+53,capstone$weeknumber)
```

#As we have for our analysis week 26 that is June 2015 is considerered to be the first week, updating week 26 as the first week.

```
capstone$weeknumber <- ifelse(capstone$weeknumber > 26, capstone$weeknumber-26,
capstone$weeknumber)

unique(capstone$weeknumber)

#Check if we have week number beyond 54 weeks. If we have more than 54 weeknumber, then
#we need to process the outlier dates i.e dates outside of the analysis period "July 2015 to June 2016"
min(capstone$weeknumber)   #1
max(capstone$weeknumber)   #57

#------------------Format the columns "order_id","order_item_id","cust_id" & "pincode"------------------#
#capstone$order_id <-format(capstone$order_id, scientific = FALSE)
#capstone$order_item_id <-format(capstone$order_item_id, scientific = FALSE)
#capstone$cust_id <-format(capstone$cust_id, scientific = FALSE)
capstone$pincode <-format(capstone$pincode, scientific = FALSE)

#------------------Extract the non-negative from cust_id" & "pincode"------------------#
## extract the numbers
regexp <- "[[:digit:]]+"

#process string
capstone$pincode <- str_extract(capstone$pincode, regexp)
capstone$cust_id <- str_extract(capstone$cust_id, regexp)

#------------------To check the Total Number of COLUMNS having NA Values------------------#
#sum(is.na(capstone))
options(repr.plot.width=8, repr.plot.height=3)

# Visualize the missing values using the DataExplorer package
plot_missing(capstone)

# To find the list of columns having NA
colnames(capstone)[colSums(is.na(capstone)) > 0]
#"gmv"     "cust_id" "pincode"

#------------------Extract the Dataframe WITHOUT NA's------------------#

capstonewithna <- capstone %>% filter(is.na(capstone$pincode))
nrow(capstonewithna)      #4904

#Omit the NA's
capstonewithoutna <- capstone %>% na.omit()
#DF without NA's
nrow(capstonewithoutna)   #1643920

paste("Orignal record count is :",nrow(capstone), "Record count after omitting NA's :",
nrow(capstonewithoutna),
```

```
                                                    "Record count with NA is ",(nrow(capstone) -
nrow(capstonewithoutna)))

# "Orignal record count is : 1648824 Record count after omitting NA's : 1643920 Record count with NA is
4904"

#------------------Extract the Date and Day columns------------------#


capstonewithoutna$Odate <- as.Date(capstonewithoutna$order_date,format = "%Y-%m-%d")    #Is
mandatory for the code to train the given date format is YYYY-mm-dd,

#Extract the Date in %Y_%m" format
capstonewithoutna$yrmonth <- format(capstonewithoutna$Odate,"%Y_%m")
unique(capstonewithoutna$yrmonth)

uniqueyrmonth <- unique(capstonewithoutna$yrmonth)
length(uniqueyrmonth)

#Extract the Date in "%Y-%m-%d" format
capstonewithoutna$yrmonthdate <- format(capstonewithoutna$Odate,"%Y_%m_%d")
head(capstonewithoutna$yrmonthdate)

uniqueyrmonthdate <- unique(capstonewithoutna$yrmonthdate)
length(uniqueyrmonthdate)

#-----------Extract the ORDER DATA only from "July 2015 to June 2016" per the Scope of the project

#Excluding the Orders outside of "July 2015 to June 2016"  => 2015_05, 2015_06, ,2016_07
#expectedyrmonth <-
c('2015_07','2015_08','2015_09','2015_10','2015_11','2015_12','2016_01','2016_02','2016_03','2016_0
4','2016_05','2016_06')
expectedyrmonth <-
c("2015_07","2015_08","2015_09","2015_10","2015_11","2015_12","2016_01","2016_02","2016_03","
2016_04","2016_05","2016_06")

capstonewithoutna_outsideperiod <- capstonewithoutna
capstonewithoutna_outsideperiod <-
subset(capstonewithoutna_outsideperiod,!(capstonewithoutna$yrmonth %in%
c("2015_07","2015_08","2015_09","2015_10","2015_11","2015_12","2016_01","2016_02","2016_03","
2016_04","2016_05","2016_06")))
nrow(capstonewithoutna_outsideperiod)  #609

#Found we have ORDER DATA even from 2016_07" "2015_06" "2015_05" which is out of the Scope of
this project.
unique(capstonewithoutna_outsideperiod$yrmonth)

capstonewithoutna_withinperiod <- capstonewithoutna
```

```
capstonewithoutna_withinperiod <-
subset(capstonewithoutna_withinperiod,capstonewithoutna$yrmonth %in%
c("2015_07","2015_08","2015_09","2015_10","2015_11","2015_12","2016_01","2016_02","2016_03","
2016_04","2016_05","2016_06"))
nrow(capstonewithoutna_withinperiod)    #1643311

nrow(capstonewithoutna) - nrow(capstonewithoutna_outsideperiod)   #1643311  matches with above
count.

paste("we had ",length(unique(capstonewithoutna_withinperiod$yrmonthdate)),"unique number of
Transactional days")


capstonewithoutna_withinperiod_bk<-capstonewithoutna_withinperiod

#After processing the ORDER Data to be within the analysis period "July 2015 to June 2016" . Check the
WeekNumbers to be within 54,
min(capstonewithoutna_withinperiod$weeknumber)
max(capstonewithoutna_withinperiod$weeknumber)

#Note: When we do diff between 1st july 2015 and 30th June 2016, we get => 52 weeks 2 days(366
days)
#so HERE we have 53rd week as the max.


#----------Calculate Unit price from the ORDER Level Data----------#
capstonewithoutna_withinperiod$list_price <-
capstonewithoutna_withinperiod$gmv/capstonewithoutna_withinperiod$units

head(capstonewithoutna_withinperiod)

#----------Process the MRP & GMV if it is less than 0 from the ORDER Level Data----------#
capstonewithoutna_withinperiod_bk <- capstonewithoutna_withinperiod
capstonewithoutna_withinperiod <- capstonewithoutna_withinperiod_bk

nrow(capstonewithoutna_withinperiod)
capstonewithoutna_withinperiod <-
subset(capstonewithoutna_withinperiod,!capstonewithoutna_withinperiod$product_mrp <= 0)

nrow(capstonewithoutna_withinperiod)
length(unique(capstonewithoutna_withinperiod$order_id))
length(unique(capstonewithoutna_withinperiod$order_item_id))

#1643311-1638021

capstonewithoutna_withinperiod <-
subset(capstonewithoutna_withinperiod,!capstonewithoutna_withinperiod$gmv <= 0)
nrow(capstonewithoutna_withinperiod)
```

```
#1638021-1637036

#----------Process the calculated unit price, if Unit price greater than MRP----------#
#Adding a column "moreListPrice" with 1 OR 0, if the List Price is greater than the MRP(1) or Not(0)
capstonewithoutna_withinperiod$moreListPrice <- ifelse(capstonewithoutna_withinperiod$list_price >
capstonewithoutna_withinperiod$product_mrp,1,0)

capstonewithoutna_withinperiod <-
subset(capstonewithoutna_withinperiod,!capstonewithoutna_withinperiod$list_price >
capstonewithoutna_withinperiod$product_mrp)
listpricemorethanMRP <-
subset(capstonewithoutna_withinperiod,capstonewithoutna_withinperiod$list_price >
capstonewithoutna_withinperiod$product_mrp)
#View(listpricemorethanMRP)  #33632

nrow(capstonewithoutna_withinperiod)   #1603404
nrow(listpricemorethanMRP)
paste("Total % of OUTLIER in List price is :",
format((nrow(listpricemorethanMRP)/nrow(capstonewithoutna_withinperiod))*100,digits =2),"%")


#----------Calculate unit discount----------#
capstonewithoutna_withinperiod$discount <- (capstonewithoutna_withinperiod$product_mrp -
capstonewithoutna_withinperiod$list_price) / capstonewithoutna_withinperiod$product_mrp
head(capstonewithoutna_withinperiod)


discountinNegative <-
subset(capstonewithoutna_withinperiod,capstonewithoutna_withinperiod$discount < 0)
nrow(discountinNegative)   #33632
#View(capstonewithoutna_withinperiod)

unique(capstonewithoutna_withinperiod$discount)

#----------Calculate if we have any negative values for deliverybdays and deliverycdays----------#
table(capstonewithoutna_withinperiod$sla)

unique(capstonewithoutna_withinperiod$deliverybdays)
deliverybdays_positive <- subset(capstonewithoutna_withinperiod,
capstonewithoutna_withinperiod$deliverybdays > 0)

nrow(deliverybdays_positive)   #333453
unique(deliverybdays_positive$deliverybdays)

setdiff(unique(capstonewithoutna_withinperiod$deliverybdays),unique(deliverybdays_positive$delivery
bdays))
# "\\N" "0"   "-71" "-72" "-53" "-40" "-39" "-41" "-14" "-45" "-56" "-73" "-75" "-12" "-22" "-44" "-42" "-43"
"-74" "-13" "-98" "-46" "-76" "-10" "-77"
```

```
unique(capstonewithoutna_withinperiod$deliverycdays)
deliverycdays_positive <- subset(capstonewithoutna_withinperiod,
capstonewithoutna_withinperiod$deliverycdays > 20)

nrow(deliverycdays_positive)   #335274
unique(deliverycdays_positive$deliverycdays)

setdiff(unique(capstonewithoutna_withinperiod$deliverycdays),unique(deliverycdays_positive$delivery
cdays))
# "\\N"   "0"    "-832"  "-840"  "-628"  "-836"  "-834"  "-476"  "-466"  "-848"  "-844"  "-482"  "-837"  "-
849"  "-835"  "-16"   "-532"  "-655"  "-859"  "-879"  "-14"   "-25"   "-516"  "-876"  "-492"  "-510"
# "-512"  "-871"  "-875"  "-153"  "-5345" "-884"  "-115"  "-544"  "-893"  "-11"   "-898"  "-908"
```

#----------Calculate if the Order is placed during the Sale Day----------#

```
#Sales Calendar
#2015
#Eid & Rathayatra sale (18-19th July)
#Independence Sale (15-17th Aug)
#Rakshabandhan Sale (28-30th Aug)
#Daussera sale (17-15th Oct)
#Big Diwali Sale (7-14th Nov)
#Christmas & New Year Sale (25th Dec'15 - 3rd Jan'16)

#Translate the above Sale period into a Salecalendar vector for processing
SaleCalendar_2015 <-
c('2015_07_18','2015_07_19','2015_08_15','2015_08_16','2015_08_17','2015_08_28','2015_08_29','20
15_08_30','2015_10_15','2015_10_16','2015_10_17','2015_11_07','2015_11_08',

'2015_11_09','2015_11_10','2015_11_11','2015_11_12','2015_11_13','2015_11_14','2015_12_25','2015
_12_26','2015_12_27','2015_12_28','2015_12_29','2015_12_30',
                '2015_12_31','2016_01_01','2016_01_02','2016_01_03')

#Note: we consider jan03,jan02,jan01 of 2016 as in sale calendar of 2015 itself.

#capstonewithoutna_withinperiod_2015saleday <- capstonewithoutna_withinperiod %>%
filter(capstonewithoutna_withinperiod$yrmonthdate %in% SaleCalendar_2015)
#nrow(capstonewithoutna_withinperiod_2015saleday)    #189719

##2016
#Republic Day (20-22 Jan)
#BED (1-2 Feb)
#FHSD (20-21 Feb)
#Valentine's Day (14-15 Feb)
#BSD-5 (7-9 Mar)
#Pacman (25-27 May)
```

```
SaleCalendar_2016 <-
c('2016_01_20','2016_01_21','2016_01_22','2016_02_01','2016_02_02','2016_02_14','2016_02_15','20
16_02_20','2016_02_21','2016_03_07','2016_03_08','2016_03_09','2016_03_25','2016_03_27')

#capstonewithoutna_withinperiod_2016saleday <- capstonewithoutna_withinperiod %>%
filter(capstonewithoutna_withinperiod$yrmonthdate %in% SaleCalendar_2016)
#nrow(capstonewithoutna_withinperiod_2016saleday)     #86135


TotalSalecalendar2015_2016 <- c(SaleCalendar_2015,SaleCalendar_2016)
TotalSalecalendar2015_2016

#Adding a column with 1 OR 0, if the Transaction is within the sale period(1) or Not(0)
capstonewithoutna_withinperiod$Saleday <- ifelse(capstonewithoutna_withinperiod$yrmonthdate
%in% TotalSalecalendar2015_2016,1,0)
head(capstonewithoutna_withinperiod)
#filter(capstonewithoutna_withinperiod, capstonewithoutna_withinperiod$Saleday ==1) %>% nrow()


capstonewithoutna_withinperiod_Totalsaleday <-
subset(capstonewithoutna_withinperiod,capstonewithoutna_withinperiod$yrmonthdate %in%
TotalSalecalendar2015_2016)
nrow(capstonewithoutna_withinperiod_Totalsaleday)     #275854

capstonewithoutna_withinperiod_Not4mTotalsaleday <-
subset(capstonewithoutna_withinperiod,!capstonewithoutna_withinperiod$yrmonthdate %in%
TotalSalecalendar2015_2016)
nrow(capstonewithoutna_withinperiod_Not4mTotalsaleday)  #1367457

paste("Percentage of Sale day contribution to the total is :",
format((nrow(capstonewithoutna_withinperiod_Totalsaleday))/(nrow(capstonewithoutna_withinperiod
_Not4mTotalsaleday))*100,digits =2),"%")


#275854+1367457 #1643311
#275854/1367457 #.2


#----------Media Investment period----------#
#2015_7,2015_8,2015_9,2015_10,2015_11,2015_12,2016_1,2016_2,2016_3,2016_4,2016_5,2016_6
#same as yrmonth
```

```
#############################################################################################
#               Analysis on the selected 3 sub-categories                                   #
#############################################################################################

#------------------Analysis Category - 3 product sub-categories
################"camera accessory, home audio and gaming accessory"#############

analysisCategory <- c('CameraAccessory','HomeAudio','GamingAccessory')
#capstonewithoutna_withinperiod_b317 <- capstonewithoutna_withinperiod
#capstonewithoutna_withinperiod <- capstonewithoutna_withinperiod_Filter_grpbyOdate_media

names(capstonewithoutna_withinperiod)
unique(capstonewithoutna_withinperiod$product_analytic_sub_category)

analysisCategory <- c('CameraAccessory','HomeAudio','GamingAccessory')

#------------------Filtering for "CameraAccessory" based on Order level data------------------#
camera_accessory <-
subset(capstonewithoutna_withinperiod,capstonewithoutna_withinperiod$product_analytic_sub_categ
ory == 'CameraAccessory')
#View(camera_accessory)
nrow(camera_accessory)
str(camera_accessory)
paste("Total Camera_accessory contribution of the total is :",
format((nrow(camera_accessory))/(nrow(capstonewithoutna_withinperiod))*100,digits =2),"%")

#------------------Filtering for "HomeAudio based" on Order level data------------------#
home_audio <-
subset(capstonewithoutna_withinperiod,capstonewithoutna_withinperiod$product_analytic_sub_categ
ory == 'HomeAudio')
head(home_audio)
nrow(home_audio)

paste("Total home_audio contribution of the total is :",
format((nrow(home_audio))/(nrow(capstonewithoutna_withinperiod))*100,digits =2),"%")

#------------------Filtering for "GamingAccessory" based on Order level data------------------#
gaming_accessory <-
subset(capstonewithoutna_withinperiod,capstonewithoutna_withinperiod$product_analytic_sub_categ
ory == 'GamingAccessory')
head(gaming_accessory)
nrow(gaming_accessory)

paste("Total gaming_accessory contribution of the total is :",
format((nrow(gaming_accessory))/(nrow(capstonewithoutna_withinperiod))*100,digits =2),"%")

capstonewithoutna_withinperiod_bkp <- capstonewithoutna_withinperiod
```

```
#capstonewithoutna_withinperiod <- capstonewithoutna_withinperiod_bkp

##Merge the Order Level Data with the Media data and other information data for Model Building

#----------MERGE NPS Score with the Main Product list----------#

#library(ggplot2)
#ggplot(capstonewithoutna_withinperiod, aes(x=s1_fact.order_payment_type,y=gmv)) + geom_point()

ncol(capstonewithoutna_withinperiod)  #After dropping the below variables, we have
names(capstonewithoutna_withinperiod)
#View(capstonewithoutna_withinperiod)

capstonewithoutna_withinperiod_376 <- capstonewithoutna_withinperiod
#capstonewithoutna_withinperiod <- capstonewithoutna_withinperiod_425
#capstonewithoutna_withinperiod <- capstonewithoutna_withinperiod_final1


#Here we are merging using "yrmonth" of electronics master sheet with "Month" of Npsscore sheet.
head(npsscore,12)
# for (i in 1:length(npsscore$Month)) {
#   if(nchar(npsscore[i,2])=="1")
#   {
#     npsscore$Month[i] <- assign(paste0(npsscore$Month[i]), paste0('0', npsscore[i,2]))
#   }
#   else{
#     print("No change")
#   }
# }

capstonewithoutna_withinperiod_npsscore <-
merge(capstonewithoutna_withinperiod,npsscore,by.x="Month",by.y="Month" )
names(capstonewithoutna_withinperiod_npsscore)

unique(capstonewithoutna_withinperiod$Month)
unique(capstonewithoutna_withinperiod_npsscore$yrmonth)
#unique(productlist$X)

#Here we are merging using "product_analytic_vertical" as the productlist is having more than 75
unique component those are listed in "product_analytic_vertical" column in electronics master sheet
head(productlist)
capstonewithoutna_withinperiod_bk05 <- capstonewithoutna_withinperiod
capstonewithoutna_withinperiod_npsscore_ProductList <-
merge(capstonewithoutna_withinperiod_npsscore,productlist,by.x="product_analytic_vertical",by.y="X
" )
head(capstonewithoutna_withinperiod_npsscore_ProductList)

#unique(capstonewithoutna_withinperiod_npsscore_ProductList$product_analytic_vertical)
```

```
capstonewithoutna_withinperiod_bk408 <- capstonewithoutna_withinperiod
#capstonewithoutna_withinperiod <- capstonewithoutna_withinperiod_bk408
capstonewithoutna_withinperiod <- capstonewithoutna_withinperiod_npsscore_ProductList
capstonewithoutna_withinperiod_411 <- capstonewithoutna_withinperiod
#----

#----------Categorical variable treatment----------#
capstonewithoutna_withinperiod_b412 <- capstonewithoutna_withinperiod
names(capstonewithoutna_withinperiod)
capstonewithoutna_withinperiod$s1_fact.order_payment_type <-
factor(capstonewithoutna_withinperiod$s1_fact.order_payment_type)
capstonewithoutna_withinperiod$moreListPrice <-
factor(capstonewithoutna_withinperiod$moreListPrice)
capstonewithoutna_withinperiod$Saleday <- factor(capstonewithoutna_withinperiod$Saleday)

#----------DUMMY VARIABLE CREATION----------#

#1 s1_fact.order_payment_type
#2 moreListPrice
#3 SaleDay

# The creation of dummy variables to convert a categorical variable into a numeric variable is an
important step of data preparation.
unique(capstonewithoutna_withinperiod$s1_fact.order_payment_type)
#"COD"     "Prepaid"
unique(capstonewithoutna_withinperiod$moreListPrice)
#unique(capstonewithoutna_withinperiod$SaleDay)

# 1. One simple way to convert "s1_fact.order_payment_type" variable to numeric is to replace the
levels- COD and Prepaid with 1 and 0 is:
levels(capstonewithoutna_withinperiod$s1_fact.order_payment_type)<-c(1,0)
# 2,3 Already have SaleDay, moreListPrice as categorical
capstonewithoutna_withinperiod$s1_fact.order_payment_type
# Now store the numeric values in the same variable
capstonewithoutna_withinperiod$s1_fact.order_payment_type <-
as.numeric(levels(capstonewithoutna_withinperiod$s1_fact.order_payment_type))[capstonewithoutna
_withinperiod$s1_fact.order_payment_type]
capstonewithoutna_withinperiod$moreListPrice <-
as.numeric(capstonewithoutna_withinperiod$moreListPrice)
capstonewithoutna_withinperiod$Saleday <- as.numeric(capstonewithoutna_withinperiod$Saleday)
#capstonewithoutna_withinperiod$Frequency <-
as.numeric(capstonewithoutna_withinperiod$Frequency)
#head(View(capstonewithoutna_withinperiod))
colnames(capstonewithoutna_withinperiod)
# Check the summary of those variable (We can find the min, max, median)
summary(capstonewithoutna_withinperiod$s1_fact.order_payment_type)
summary(capstonewithoutna_withinperiod$moreListPrice)
```

```
summary(capstonewithoutna_withinperiod$Saleday)

analysisCategory
names(capstonewithoutna_withinperiod)
summary(factor(capstonewithoutna_withinperiod_bk408$product_analytic_sub_category))


################################################################################################
#                              DATA PREPARATIONF FOR THE FINAL MODEL PREPARATION                #
################################################################################################

#Creating the Dummy Variables for the 3 sub-categories
#Add the Dummy variables to the 3 sub-categories individual Dataset
#Aggregate the 3 sub-categories individual dataset using the "Weeknumber"
 #capstonewithoutna_withinperiod_Filter_grpbyweeknumber_GA
 #capstonewithoutna_withinperiod_Filter_grpbyweeknumber_HA
 #capstonewithoutna_withinperiod_Filter_grpbyweeknumber_CA

#Preparing the MediaInvest data for merging with the Order level data
#Aggregate the Media investment dataset using the "Weeknumber"
#Merge the Media investment dataset with 3 sub-categories dataset


#----------Creating the Dummy variable for the CameraAccessory subcategory
"product_analytic_vertical" unique values into Columns----------#

## Create a dataset only for Home audio , camera accessory and gaming accessories
#capstonewithoutna_withinperiod_3subcategory <- filter ( capstonewithoutna_withinperiod
,capstonewithoutna_withinperiod$product_analytic_sub_category %in% analysisCategory )
capstonewithoutna_withinperiod_CameraAccessory <- filter ( capstonewithoutna_withinperiod
,capstonewithoutna_withinperiod$product_analytic_sub_category == 'CameraAccessory' )

capstonewithoutna_withinperiod_CA <- capstonewithoutna_withinperiod_CameraAccessory
capstonewithoutna_withinperiod_CA$product_analytic_vertical <-
factor(capstonewithoutna_withinperiod_CA$product_analytic_vertical)
levels(capstonewithoutna_withinperiod_CA$product_analytic_vertical)
dummy_subcat<-data.frame(model.matrix(~product_analytic_vertical,data =
capstonewithoutna_withinperiod_CA))
dummy_subcat<-dummy_subcat[,-1]
capstonewithoutna_withinperiod_CA1<-capstonewithoutna_withinperiod_CA
capstonewithoutna_withinperiod_CA1bk <- capstonewithoutna_withinperiod_CA1
capstonewithoutna_withinperiod_CA1<-capstonewithoutna_withinperiod_CA1[,-1]
capstonewithoutna_withinperiod_CA1<-cbind(capstonewithoutna_withinperiod_CA1,dummy_subcat)
colnames(capstonewithoutna_withinperiod_CA1)

names(capstonewithoutna_withinperiod_CA1)
str(capstonewithoutna_withinperiod_CA1)
```

```
capstonewithoutna_withinperiod_472 <- capstonewithoutna_withinperiod
capstonewithoutna_withinperiod_CA <- capstonewithoutna_withinperiod_CA1
```

#-Creating the Dummy variable for the HomeAudio subcategory "product_analytic_vertical" unique values into Columns--#

```
capstonewithoutna_withinperiod_HomeAudio <- filter ( capstonewithoutna_withinperiod
,capstonewithoutna_withinperiod$product_analytic_sub_category == 'HomeAudio' )

capstonewithoutna_withinperiod_HA <- capstonewithoutna_withinperiod_HomeAudio
capstonewithoutna_withinperiod_HA$product_analytic_vertical <-
factor(capstonewithoutna_withinperiod_HA$product_analytic_vertical)
levels(capstonewithoutna_withinperiod_HA$product_analytic_vertical)
dummy_subcat<-data.frame(model.matrix(~product_analytic_vertical,data =
capstonewithoutna_withinperiod_HA))
dummy_subcat<-dummy_subcat[,-1]
capstonewithoutna_withinperiod_HA1<-capstonewithoutna_withinperiod_HA
capstonewithoutna_withinperiod_HA1bk <- capstonewithoutna_withinperiod_HA1
capstonewithoutna_withinperiod_HA1<-capstonewithoutna_withinperiod_HA1[,-1]
capstonewithoutna_withinperiod_HA1<-cbind(capstonewithoutna_withinperiod_HA1,dummy_subcat)
colnames(capstonewithoutna_withinperiod_HA1)

names(capstonewithoutna_withinperiod_HA1)
str(capstonewithoutna_withinperiod_HA1)

capstonewithoutna_withinperiod_493 <- capstonewithoutna_withinperiod
capstonewithoutna_withinperiod_HA <- capstonewithoutna_withinperiod_HA1

#------------------Creating the Dummy variable for the GamingAccessory subcategory
"product_analytic_vertical" unique values into Columns------------------#

capstonewithoutna_withinperiod_GamingAccessory <- filter ( capstonewithoutna_withinperiod
,capstonewithoutna_withinperiod$product_analytic_sub_category == 'GamingAccessory' )

capstonewithoutna_withinperiod_GA <- capstonewithoutna_withinperiod_GamingAccessory
capstonewithoutna_withinperiod_GA$product_analytic_vertical <-
factor(capstonewithoutna_withinperiod_GA$product_analytic_vertical)
levels(capstonewithoutna_withinperiod_GA$product_analytic_vertical)
dummy_subcat<-data.frame(model.matrix(~product_analytic_vertical,data =
capstonewithoutna_withinperiod_GA))
dummy_subcat<-dummy_subcat[,-1]
capstonewithoutna_withinperiod_GA1<-capstonewithoutna_withinperiod_GA
capstonewithoutna_withinperiod_GA1bk <- capstonewithoutna_withinperiod_GA1
capstonewithoutna_withinperiod_GA1<-capstonewithoutna_withinperiod_GA1[,-1]
capstonewithoutna_withinperiod_GA1<-cbind(capstonewithoutna_withinperiod_GA1,dummy_subcat)
colnames(capstonewithoutna_withinperiod_GA1)

names(capstonewithoutna_withinperiod_GA1)
```

```
str(capstonewithoutna_withinperiod_GA1)

capstonewithoutna_withinperiod_513 <- capstonewithoutna_withinperiod
capstonewithoutna_withinperiod_GA <- capstonewithoutna_withinperiod_GA1

#----------Added the Dummy variables for the Home audio , camera accessory and gaming accessories from product_analytic_vertical
#View(capstonewithoutna_withinperiod_CA)
#View(capstonewithoutna_withinperiod_HA)
#View(capstonewithoutna_withinperiod_GA)


#----------Data Aggregation----------#

#----------Aggregate the Home audio,camera accessory and gaming accessories dataset using the "Weeknumber"
names(capstonewithoutna_withinperiod_CA)
unique(capstonewithoutna_withinperiod$product_analytic_sub_category)

capstonewithoutna_withinperiod_Filter_grpbyweeknumber_CA <- capstonewithoutna_withinperiod_CA %>%
        group_by(product_analytic_sub_category,weeknumber) %>% summarise(gmv=sum(gmv), product_mrp=mean(product_mrp),units=sum(units),

discount=mean(discount),NPS=mean(NPS),Frequency=mean(Frequency),Percent=mean(Percent),CameraAccessory=sum(product_analytic_verticalCameraAccessory),

CameraBag=sum(product_analytic_verticalCameraBag),CameraBattery=sum(product_analytic_verticalCameraBattery),CameraBatteryCharger=sum(product_analytic_verticalCameraBatteryCharger),

CameraBatteryGrip=sum(product_analytic_verticalCameraBatteryGrip),CameraEyeCup=sum(product_analytic_verticalCameraEyeCup),CameraFilmRolls=sum(product_analytic_verticalCameraFilmRolls),

CameraHousing=sum(product_analytic_verticalCameraHousing),CameraMicrophone=sum(product_analytic_verticalCameraMicrophone),CameraMount=sum(product_analytic_verticalCameraMount),

CameraRemoteControl=sum(product_analytic_verticalCameraRemoteControl),CameraTripod=sum(product_analytic_verticalCameraTripod),ExtensionTube=sum(product_analytic_verticalExtensionTube),

Filter=sum(product_analytic_verticalFilter),Flash=sum(product_analytic_verticalFlash),FlashShoeAdapter=sum(product_analytic_verticalFlashShoeAdapter),Lens=sum(product_analytic_verticalLens)

,ReflectorUmbrella=sum(product_analytic_verticalReflectorUmbrella),Softbox=sum(product_analytic_verticalSoftbox),Strap=sum(product_analytic_verticalStrap),

Teleconverter=sum(product_analytic_verticalTeleconverter),Telescope=sum(product_analytic_verticalTelescope),PaymentType=sum(s1_fact.order_payment_type),IsmoreListPrice=sum(moreListPrice),isSaleday=sum(Saleday))
```

```
names(capstonewithoutna_withinperiod_HA)
capstonewithoutna_withinperiod_Filter_grpbyweeknumber_HA <-
capstonewithoutna_withinperiod_HA %>%
            group_by(product_analytic_sub_category,weeknumber) %>%
summarise(gmv=sum(gmv),product_mrp=mean(product_mrp),units=sum(units),discount=mean(discoun
t),NPS=mean(NPS),Frequency=mean(Frequency),

Percent=mean(Percent),Dock=sum(product_analytic_verticalDock),DockingStation=sum(product_analyti
c_verticalDockingStation),FMRadio=sum(product_analytic_verticalFMRadio),

HiFiSystem=sum(product_analytic_verticalHiFiSystem),HomeAudioSpeaker=sum(product_analytic_verti
calHomeAudioSpeaker),KaraokePlayer=sum(product_analytic_verticalKaraokePlayer),

SlingBox=sum(product_analytic_verticalSlingBox),SoundMixer=sum(product_analytic_verticalSoundMix
er),VoiceRecorder=sum(product_analytic_verticalVoiceRecorder),

PaymentType=sum(s1_fact.order_payment_type),IsmoreListPrice=sum(moreListPrice),isSaleday=sum(Sa
leday))



names(capstonewithoutna_withinperiod_GA)
capstonewithoutna_withinperiod_Filter_grpbyweeknumber_GA <-
capstonewithoutna_withinperiod_GA %>%
  group_by(product_analytic_sub_category,weeknumber) %>% summarise(gmv=sum(gmv),
product_mrp=mean(product_mrp),units=sum(units),discount=mean(discount),NPS=mean(NPS),Frequen
cy=mean(Frequency),Percent=mean(Percent),

GameControlMount=sum(product_analytic_verticalGameControlMount),GamePad=sum(product_analyt
ic_verticalGamePad),

GamingAccessoryKit=sum(product_analytic_verticalGamingAccessoryKit),GamingAdapter=sum(product
_analytic_verticalGamingAdapter),

GamingChargingStation=sum(product_analytic_verticalGamingChargingStation),GamingHeadset=sum(pr
oduct_analytic_verticalGamingHeadset),

GamingKeyboard=sum(product_analytic_verticalGamingKeyboard),GamingMemoryCard=sum(product_
analytic_verticalGamingMemoryCard),

GamingMouse=sum(product_analytic_verticalGamingMouse),GamingMousePad=sum(product_analytic
_verticalGamingMousePad),

GamingSpeaker=sum(product_analytic_verticalGamingSpeaker),JoystickGamingWheel=sum(product_an
alytic_verticalJoystickGamingWheel),
```

```
MotionController=sum(product_analytic_verticalMotionController),TVOutCableAccessory=sum(product
_analytic_verticalTVOutCableAccessory),

PaymentType=sum(s1_fact.order_payment_type),IsmoreListPrice=sum(moreListPrice),isSaleday=sum(Sa
leday))

View(capstonewithoutna_withinperiod_Filter_grpbyweeknumber_GA)
View(capstonewithoutna_withinperiod_Filter_grpbyweeknumber_HA)
View(capstonewithoutna_withinperiod_Filter_grpbyweeknumber_CA)
```

#----------Preparing the MediaInvest data for merging with the Order level data Dataset extracted for "Home audio,camera accessory and gaming accessories" above

```
#Preparation for the MediaInvest Merge
mediainvestment <- read.xls("Media data and other information.xlsx", sheet = 2, header = TRUE, skip =
1)
daysinmonths<-c(31,30,30,31,30,31,31,29,31,30,31,30)
mediainvestment$daysinmonths<-daysinmonths

getMediaidx<-function(monthnum){
 for (i in 1:12)
 {

   if (mediainvestment[i,2]==monthnum)
   {
     return(i)
   }

 }

}

stdate<-as.Date(ymd("20150701"))
dfdayinvestments <- data.frame(matrix(ncol = 12, nrow = 0))
x <- c("yymm","date",  "weeknumber","TV","Digital","Sponsorship","Content Marketing","Online
marketing","Affiliates","SEM","Radio","Other"
)
colnames(dfdayinvestments) <- x
stdate<-ymd("20150701")
weeknum=1
i=1
while (i <=366) {

 for (j in 1:7)
 {
```

```
   yymm<-year(as.Date(stdate))
   mm<-month(as.Date(stdate))
   if (mm<10) {
    stryymm<-paste(toString(yymm),"_","0",toString(mm),sep="")
   }
   else {
    stryymm<-paste(toString(yymm),"_",toString(mm),sep="")
   }
   idx<-getMediaidx(mm)
   TVinvst<-mediainvestment[idx,4]/mediainvestment[idx,13]
   Digitalinvst<-mediainvestment[idx,5]/mediainvestment[idx,13]
   Sponsorinvst<-mediainvestment[idx,6]/mediainvestment[idx,13]
   CMinvst<-mediainvestment[idx,7]/mediainvestment[idx,13]
   OMinvst<-mediainvestment[idx,8]/mediainvestment[idx,13]
   Affinvst<-mediainvestment[idx,9]/mediainvestment[idx,13]
   SEMinvst<-mediainvestment[idx,10]/mediainvestment[idx,13]
   Radioinvst<-mediainvestment[idx,11]/mediainvestment[idx,13]
   Otherinvst<-mediainvestment[idx,12]/mediainvestment[idx,13]
   dfdayinvestments[i,]<-
c(stryymm,stdate,weeknum,TVinvst,Digitalinvst,Sponsorinvst,CMinvst,OMinvst,Affinvst,SEMinvst,Radioi
nvst,Otherinvst)
   stdate<-stdate+days(1)
   i=i+1
   if (i>366){
    break
   }
  }
  weeknum=weeknum+1
}
dfdayinvestments1<-dfdayinvestments[1:366,]
#as.Date(as.integer(max(dfdayinvestments1$date)))
names(dfdayinvestments1)

#----------Aggregate the Media Invest dataset using the "Weeknumber"

dfdayinvestmentsweekly<-
dfdayinvestments1%>%dplyr::group_by(weeknumber)%>%dplyr::summarise(TV=sum(as.numeric(TV)),D
igital=sum(as.numeric(Digital)),Sponsorship=sum(as.numeric(Sponsorship)),
                                                    Content_Marketing=sum(as.numeric(`Content
Marketing`)),Online_marketing=sum(as.numeric(`Online
marketing`)),Affiliates=sum(as.numeric(Affiliates)),

SEM=sum(as.numeric(SEM)),Radio=sum(as.numeric(Radio)),Other=sum(as.numeric(Other)))%>%dplyr::a
rrange(as.integer(weeknumber))

dfdayinvestmentsweekly[which(is.na(dfdayinvestmentsweekly$Radio)), "Radio"] <- 0
dfdayinvestmentsweekly[which(is.na(dfdayinvestmentsweekly$Other)), "Other"] <- 0
```

```
View(dfdayinvestmentsweekly)


#MediaInvest Merge below with three Sub-categories
'CameraAccessory','HomeAudio','GamingAccessory'

#------------------MERGE the Media Dataset with the Extracted Order Level dataset for the 3 Sub-categories----------
#CameraAccessory
capstonewithoutna_withinperiod_Filter_grpbyOdate_media_CA <-
merge(capstonewithoutna_withinperiod_Filter_grpbyweeknumber_CA,dfdayinvestmentsweekly,by.x="
weeknumber",by.y="weeknumber")
#HomeAudio
capstonewithoutna_withinperiod_Filter_grpbyOdate_media_HA <-
merge(capstonewithoutna_withinperiod_Filter_grpbyweeknumber_HA,dfdayinvestmentsweekly,by.x="
weeknumber",by.y="weeknumber")
#GamingAccessory
capstonewithoutna_withinperiod_Filter_grpbyOdate_media_GA <-
merge(capstonewithoutna_withinperiod_Filter_grpbyweeknumber_GA,dfdayinvestmentsweekly,by.x="
weeknumber",by.y="weeknumber")

#------------------To view the merged DATASET for the 3 sub-categories
#View(capstonewithoutna_withinperiod_Filter_grpbyOdate_media_CA)
#View(capstonewithoutna_withinperiod_Filter_grpbyOdate_media_HA)
#View(capstonewithoutna_withinperiod_Filter_grpbyOdate_media_GA)

# names(capstonewithoutna_withinperiod_Filter_grpbyOdate_media_CA)
# names(capstonewithoutna_withinperiod_Filter_grpbyOdate_media_HA)
# names(capstonewithoutna_withinperiod_Filter_grpbyOdate_media_GA)

#------------------To verify the Total Media investment made for each of the Media verticals
sum((capstonewithoutna_withinperiod_Filter_grpbyOdate_media_CA$TV)/3)
sum((capstonewithoutna_withinperiod_Filter_grpbyOdate_media_CA$Digital)/3)
sum((capstonewithoutna_withinperiod_Filter_grpbyOdate_media_CA$Sponsorship)/3)
sum((capstonewithoutna_withinperiod_Filter_grpbyOdate_media_CA$Content_Marketing)/3)
sum((capstonewithoutna_withinperiod_Filter_grpbyOdate_media_CA$Online_marketing)/3)
sum((capstonewithoutna_withinperiod_Filter_grpbyOdate_media_CA$Affiliates)/3)
sum((capstonewithoutna_withinperiod_Filter_grpbyOdate_media_CA$SEM)/3)

names(capstonewithoutna_withinperiod_Filter_grpbyOdate_media_CA)
unique(capstonewithoutna_withinperiod_Filter_grpbyOdate_media_CA$product_analytic_sub_categor
y)
#-----------------------------------------------------------------------------------------------------
```