1-DataPrep.R

arman

Sat May 20 13:05:09 2017

```r
# ***************************************************************************
#                    LOAD LIBRARY ----
# ***************************************************************************
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.3.3

##
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':
##
##     date
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:lubridate':
##
##     intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(car)
```

```
##
## Attaching package: 'car'

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

```r
library(Hmisc)   # describe
```

```
## Loading required package: lattice
```

```
## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     combine, src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```r
# ****************************************************************************
#                    PROCs ----
# ****************************************************************************
nweek <- function(x, format="%Y-%m-%d", origin){
  if(missing(origin)){
    as.integer(format(strptime(x, format=format), "%W"))
  }else{
    x <- as.Date(x, format=format)
    o <- as.Date(origin, format=format)
    w <- as.integer(format(strptime(x, format=format), "%w"))
    2 + as.integer(x - o - w) %/% 7
  }
}



# ****************************************************************************
#                    LOAD DATA ---- Transaction Data ----
# ****************************************************************************
# Make sure you are in current directory as in R-file is in. Should I do a commit?yes..

ce_data <- read.csv('./data/ConsumerElectronics.csv',stringsAsFactors = FALSE)

str(ce_data)
```

```
## 'data.frame':    1648824 obs. of  20 variables:
##  $ ï..fsn_id                      : chr  "ACCCX3S58G7B5F6P" "ACCCX3S58G7B5F6P" "ACCCX3S5AHMF55FV" "AC
##  $ order_date                     : chr  "2015-10-17 15:11:54" "2015-10-19 10:07:22" "2015-10-20 15:4
##  $ Year                           : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ Month                          : int  10 10 10 10 10 10 10 10 10 10 ...
##  $ order_id                       : num  3.42e+15 1.42e+15 2.42e+15 4.42e+15 4.42e+15 ...
##  $ order_item_id                  : num  3.42e+15 1.42e+15 2.42e+15 4.42e+15 4.42e+15 ...
##  $ gmv                            : num  6400 6900 1990 1690 1618 ...
##  $ units                          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ deliverybdays                  : chr  "\\N" "\\N" "\\N" "\\N" ...
##  $ deliverycdays                  : chr  "\\N" "\\N" "\\N" "\\N" ...
##  $ s1_fact.order_payment_type     : chr  "COD" "COD" "COD" "Prepaid" ...
##  $ sla                            : int  5 7 10 4 6 5 6 5 9 7 ...
##  $ cust_id                        : num  -1.01e+18 -8.99e+18 -1.04e+18 -7.60e+18 2.89e+18 ...
##  $ pincode                        : num  -7.79e+18 7.34e+18 -7.48e+18 -5.84e+18 5.35e+17 ...
##  $ product_analytic_super_category: chr  "CE" "CE" "CE" "CE" ...
##  $ product_analytic_category      : chr  "CameraAccessory" "CameraAccessory" "CameraAccessory" "Came
```

```
##  $ product_analytic_sub_category  : chr  "CameraAccessory" "CameraAccessory" "CameraAccessory" "Came
##  $ product_analytic_vertical      : chr  "CameraTripod" "CameraTripod" "CameraTripod" "CameraTripod"
##  $ product_mrp                    : int  7190 7190 2099 2099 2099 4044 4044 4044 4044 4044 ...
##  $ product_procurement_sla        : int  0 0 3 3 3 5 5 5 5 5 ...
```

```r
# ****************************************************************************
#                    DATA CLEANING ----
# ****************************************************************************

head(ce_data)
```

```
##          ï..fsn_id          order_date Year Month     order_id
## 1 ACCCX3S58G7B5F6P 2015-10-17 15:11:54 2015    10 3.419301e+15
## 2 ACCCX3S58G7B5F6P 2015-10-19 10:07:22 2015    10 1.420831e+15
## 3 ACCCX3S5AHMF55FV 2015-10-20 15:45:56 2015    10 2.421913e+15
## 4 ACCCX3S5AHMF55FV 2015-10-14 12:05:15 2015    10 4.416592e+15
## 5 ACCCX3S5AHMF55FV 2015-10-17 21:25:03 2015    10 4.419525e+15
## 6 ACCCX3S5JGAJETYR 2015-10-17 12:07:24 2015    10 3.419189e+15
##   order_item_id  gmv units deliverybdays deliverycdays
## 1  3.419301e+15 6400     1           \\N           \\N
## 2  1.420831e+15 6900     1           \\N           \\N
## 3  2.421913e+15 1990     1           \\N           \\N
## 4  4.416592e+15 1690     1           \\N           \\N
## 5  4.419525e+15 1618     1           \\N           \\N
## 6  3.419189e+15 3324     1           \\N           \\N
##   s1_fact.order_payment_type sla       cust_id       pincode
## 1                        COD   5 -1.012991e+18 -7.791756e+18
## 2                        COD   7 -8.990325e+18  7.335411e+18
## 3                        COD  10 -1.040443e+18 -7.477688e+18
## 4                    Prepaid   4 -7.604961e+18 -5.835932e+18
## 5                    Prepaid   6  2.894557e+18  5.347354e+17
## 6                    Prepaid   5 -7.641546e+18 -1.919053e+18
##   product_analytic_super_category product_analytic_category
## 1                              CE           CameraAccessory
## 2                              CE           CameraAccessory
## 3                              CE           CameraAccessory
## 4                              CE           CameraAccessory
## 5                              CE           CameraAccessory
## 6                              CE           CameraAccessory
##   product_analytic_sub_category product_analytic_vertical product_mrp
## 1               CameraAccessory              CameraTripod        7190
## 2               CameraAccessory              CameraTripod        7190
## 3               CameraAccessory              CameraTripod        2099
## 4               CameraAccessory              CameraTripod        2099
## 5               CameraAccessory              CameraTripod        2099
## 6               CameraAccessory              CameraTripod        4044
##   product_procurement_sla
## 1                       0
## 2                       0
## 3                       3
## 4                       3
## 5                       3
## 6                       5
```

```r
# . . . .   Outlier Treatment ----
# Remove orders before July'15 and after June'16
ce_data$order_date <- format(as.POSIXct(ce_data$order_date,format='%Y-%m-%d'),
                             format='%Y-%m-%d')
ce_data$order_date <- as.Date(ce_data$order_date, format = "%Y-%m-%d")

ce_data <- subset(ce_data, order_date > "2015-6-30" & order_date < "2016-7-1")

max(ce_data$product_mrp)
```

```
## [1] 299999
```

```r
#NA Values
sapply(ce_data, function(x) sum(is.na(x)))
```

```
##                       ï..fsn_id                      order_date
##                               0                               0
##                            Year                           Month
##                               0                               0
##                        order_id                   order_item_id
##                               0                               0
##                             gmv                           units
##                            4904                               0
##                   deliverybdays                   deliverycdays
##                               0                               0
##      s1_fact.order_payment_type                             sla
##                               0                               0
##                         cust_id                         pincode
##                            4904                            4904
## product_analytic_super_category       product_analytic_category
##                               0                               0
##   product_analytic_sub_category       product_analytic_vertical
##                               0                               0
##                     product_mrp         product_procurement_sla
##                               0                               0
```

```r
#Removed NA values from GMV
ce_data <- na.omit(ce_data)
ce_data <- subset(ce_data, product_mrp != 0)
warning()
```

```
## Warning:
```

```r
# Lets add a couple of variables to the CE data. List Price from GMV and Promotion which is
# the discount offered

#....List Price variable
ce_data$List_Price <- as.integer(ce_data$gmv / ce_data$units)

#....Promotion Variable
ce_data$Promotion <- as.numeric((ce_data$product_mrp - ce_data$List_Price) / ce_data$product_mrp)

#....Here we have created a Pricing categorical variable
ce_data$mrp_category[ce_data$product_mrp == 0] <- "Free"
ce_data$mrp_category[ce_data$product_mrp >= 150001] <- "Luxury"
ce_data$mrp_category[ce_data$product_mrp >= 80001 & ce_data$product_mrp <= 150000] <- "Premium"
```

```r
ce_data$mrp_category[ce_data$product_mrp >= 30001 & ce_data$product_mrp <= 80000] <- "Mid"
ce_data$mrp_category[ce_data$product_mrp > 0 & ce_data$product_mrp <= 30000] <- "Lower"

# ****************************************************************************
#                    FEATURE ENGINEERING ----
# ****************************************************************************

# create week,  week numbers start from min 'order date'
# . . . . Week Numbers ----
dates <- as.Date(
  gsub(" .*","",ce_data$order_date)
)
ce_data$week <- nweek(dates,origin = as.Date("2015-07-01"))


# . . . . Days, weeks, Month ----
# will compute Month, week, and no.of days per week (month, week)
#
dys <- seq(as.Date("2015-07-01"),as.Date("2016-06-30"),'days')
weekdays <- data.frame('days'=dys, Month = month(dys),
                       week = nweek(dys,origin = as.Date("2015-07-01")),
                       nweek = rep(1,length(dys)))
weekdays <- data.frame(weekdays %>% group_by(Month,week) %>% summarise(nweeks = sum(nweek)))
weekdays$fracDays <- weekdays$nweeks/7


# . . . . Strip Spaces ----
ce_data$product_analytic_vertical <- gsub(" +","",ce_data$product_analytic_vertical)




# ****************************************************************************
#                    LOAD DATA ---- Media & Inv Data ----
# ****************************************************************************
# . . . .   ProductList ----
productList_data      <-
  read.csv("./data/ProductList.csv", stringsAsFactors = FALSE,
           na.strings=c('\\N'))

# . . . .   Media Investment ----
mediaInvestment_data  <-
  read.csv("./data/MediaInvestment.csv", stringsAsFactors = FALSE)

# . . . .  Special Sale Event ----

specialSale_data      <-
  read.csv("./data/SpecialSale.csv", stringsAsFactors = FALSE)

# . . . .   Monthly NPS ----
monthlyNPS_data       <-
  read.csv("./data/MonthlyNPSscore.csv", stringsAsFactors = FALSE )
```

```r
# . . . .   Holiday List ----
holiday_list      <-
  read.csv("./data/HolidayList.csv", stringsAsFactors = FALSE)


# ************************************************************************
#                    DATA PREPARATION ----
# ************************************************************************
# . . . . . . .  Correct Data types ----
productList_data$Frequency <- as.integer(productList_data$Frequency)
```

```
## Warning: NAs introduced by coercion
```

```r
summary(productList_data)
```

```
##    Product            Frequency         Percent
##  Length:75          Min.   :     1   Min.   :  0.000
##  Class :character   1st Qu.:   386   1st Qu.:  0.000
##  Mode  :character   Median :  3889   Median :  0.200
##                     Mean   : 22281   Mean   :  2.671
##                     3rd Qu.: 20067   3rd Qu.:  1.450
##                     Max.   :287850   Max.   :100.000
##                     NA's   :1
```

```r
# . . . .   Media Investment ----
str(mediaInvestment_data)
```

```
## 'data.frame':    12 obs. of  12 variables:
##  $ Year             : int  2015 2015 2015 2015 2015 2015 2016 2016 2016 2016 ...
##  $ Month            : int  7 8 9 10 11 12 1 2 3 4 ...
##  $ Total.Investment : num  17.1 5.1 96.3 170.2 51.2 ...
##  $ TV               : num  0.2 0 3.9 6.1 4.2 5.4 4.4 2.6 9.3 5.2 ...
##  $ Digital          : num  2.5 1.3 1.4 12.6 1.3 3.1 0.5 1.9 2.1 0.9 ...
##  $ Sponsorship      : num  7.4 1.1 62.8 84.7 14.2 56.7 4.2 11.7 41.6 24.3 ...
##  $ Content.Marketing: num  0 0 0.6 3.4 0.2 1.1 0.9 0.6 0.4 0 ...
##  $ Online.marketing : num  1.3 0.1 16.4 24.4 19.6 22.5 22.9 19.9 18.4 16.5 ...
##  $ Affiliates       : num  0.5 0.1 5 7 6.6 6.8 7.4 6.5 6.2 5.7 ...
##  $ SEM              : num  5 2.5 6.2 31.9 5.2 11.2 4.2 4.9 5.2 4.2 ...
##  $ Radio            : num  NA NA NA NA NA NA 2.7 NA 0.9 NA ...
##  $ Other            : num  NA NA NA NA NA NA 27.1 NA 15.9 NA ...
```

```r
summary(mediaInvestment_data)
```

```
##       Year          Month       Total.Investment       TV
##  Min.   :2015   Min.   : 1.00   Min.   :  5.10   Min.   :0.000
##  1st Qu.:2015   1st Qu.: 3.75   1st Qu.: 46.77   1st Qu.:1.625
##  Median :2016   Median : 6.50   Median : 65.50   Median :4.050
##  Mean   :2016   Mean   : 6.50   Mean   : 70.55   Mean   :3.700
##  3rd Qu.:2016   3rd Qu.: 9.25   3rd Qu.: 97.22   3rd Qu.:5.250
##  Max.   :2016   Max.   :12.00   Max.   :170.20   Max.   :9.300
##
##     Digital        Sponsorship    Content.Marketing Online.marketing
##  Min.   : 0.500   Min.   : 1.10   Min.   :0.0000    Min.   : 0.10
##  1st Qu.: 1.200   1st Qu.:10.62   1st Qu.:0.0000    1st Qu.:14.30
##  Median : 1.400   Median :24.65   Median :0.5000    Median :19.00
##  Mean   : 2.483   Mean   :30.45   Mean   :0.6667    Mean   :16.14
```

```
##  3rd Qu.: 2.200   3rd Qu.:45.38   3rd Qu.:0.8250    3rd Qu.:22.60
##  Max.   :12.600   Max.   :84.70   Max.   :3.4000    Max.   :24.40
##
##    Affiliates         SEM             Radio           Other
##  Min.   :0.100   Min.   : 2.500   Min.   :0.900   Min.   : 5.00
##  1st Qu.:4.450   1st Qu.: 4.200   1st Qu.:1.000   1st Qu.:10.45
##  Median :6.350   Median : 5.100   Median :1.100   Median :15.90
##  Mean   :5.117   Mean   : 7.592   Mean   :1.567   Mean   :16.00
##  3rd Qu.:6.800   3rd Qu.: 6.375   3rd Qu.:1.900   3rd Qu.:21.50
##  Max.   :7.400   Max.   :31.900   Max.   :2.700   Max.   :27.10
##                                   NA's   :9       NA's   :9
```

```
# . . . . . . . .  Missing Values ----
mediaInvestment_data[is.na(mediaInvestment_data)] <- 0   # zero investment

# . . . . . . . .  Convert to weekly data ----
# convert montly spend to weekly
mediaInvestment_data <- cbind(Month=mediaInvestment_data[,c(2)],
                              mediaInvestment_data[,-c(1,2)]/4.30)
# Add weekly information
mediaInvestment_weekly <- merge(weekdays,mediaInvestment_data, by='Month', all.x = TRUE)

# Convert media Investment at weekly granularity
# pro-rate weekly investment as per the ratio of its days span over adjacent months
mediaInvestment_weekly <- data.frame(mediaInvestment_weekly %>% group_by(week) %>%
                                       summarise(TotalInvestment = sum(Total.Investment*fracDays),
                                                 TV = sum(TV*fracDays),
                                                 Digital=sum(Digital*fracDays),
                                                 Sponsorship = sum(Sponsorship*fracDays),
                                                 ContentMarketing = sum(Content.Marketing*fracDays),
                                                 OnlineMarketing = sum(Online.marketing*fracDays),
                                                 Affiliates = sum(Affiliates*fracDays),
                                                 SEM = sum(SEM*fracDays),
                                                 Radio = sum(Radio*fracDays),
                                                 Other = sum(Other*fracDays))
)

# . . . .   SPecialSale ----
str(specialSale_data)
```

```
## 'data.frame':    44 obs. of  2 variables:
##  $ Date      : chr  "7/18/2015" "7/19/2015" "8/15/2015" "8/16/2015" ...
##  $ Sales.Name: chr  "Eid & Rathayatra sale" "Eid & Rathayatra sale" "Independence Sale" "Independenc
```

```
specialSale_data$Date     <- as.Date(specialSale_data$Date, format = "%m/%d/%Y")
specialSale_data$week     <- nweek(specialSale_data$Date,origin = as.Date("2015-07-01"))

summary(specialSale_data)
```

```
##       Date             Sales.Name             week
##  Min.   :2015-07-18   Length:44          Min.   : 3.00
##  1st Qu.:2015-11-01   Class :character   1st Qu.:18.25
##  Median :2015-12-27   Mode  :character   Median :27.00
##  Mean   :2015-12-12                      Mean   :24.55
##  3rd Qu.:2016-02-01                      3rd Qu.:32.00
```

```
##  Max.   :2016-05-27                      Max.   :48.00
```

```r
sale_days <- as.data.frame(table(specialSale_data$week))
names(sale_days) <- c("week", "sale_days")

#Created sale days here and this will be used in the final merge, because we will be needing the number


# . . . .   HolidayList ----
holiday_list$Date     <- as.Date(holiday_list$Date, format = "%m/%d/%Y")
holiday_list$week     <- nweek(holiday_list$Date,origin = as.Date("2015-07-01"))

holiday_days <- as.data.frame(table(holiday_list$week))
names(holiday_days) <- c("week", "holidays")

#Added a new KPI holiday days here and this will be used in the final merge, this is different than sal


# . . . .   Monthly NPS ----
str(monthlyNPS_data)
```

```
## 'data.frame':    12 obs. of  2 variables:
##  $ Date: chr  "7/1/2015" "8/1/2015" "9/1/2015" "10/1/2015" ...
##  $ NPS : num  54.6 60 46.9 44.4 47 45.8 47.1 50.3 49 51.8 ...
```

```r
monthlyNPS_data$Date <- as.Date(monthlyNPS_data$Date, format = "%m/%d/%Y")
monthlyNPS_data$Month <- month(ymd(monthlyNPS_data$Date))
monthlyNPS_weekly   <- merge(weekdays, monthlyNPS_data, by='Month', all.x = TRUE)
# Average weekly NPS for the weeks span over adjacent months
monthlyNPS_weekly   <- monthlyNPS_weekly %>% group_by(., week) %>%
  summarise(., NPS = mean(NPS))

# ****************************************************************************
#                    CREATING ADSTOCK ----
# ****************************************************************************

#Creating a Dataset for just the sales/gmv data to be used as input for creating media Adstock
gmv_weekly <-  ce_data %>%
  group_by(week) %>%
  summarise(gmv=sum(gmv))
write.csv(gmv_weekly, file = "sales.csv",row.names=FALSE)


# . . . .   Media Adstock ----
media_adstk      <-
  read.csv("./data/media_adstock.csv", stringsAsFactors = FALSE)


# ****************************************************************************
#                    WEEKLY DATA AGGREGATION ----
# ****************************************************************************

#Weekly aggregation of ce_data

ce_data_weekly <-  ce_data %>%
```

```r
  group_by(mrp_category,
           product_analytic_sub_category,
           week) %>%
  summarise(gmv=sum(gmv),
            product_mrp=mean(product_mrp),
            list_price=mean(List_Price),
            units=sum(units),
            Promotion=mean(Promotion),
            sla=mean(sla),
            procurement_sla=mean(product_procurement_sla))


ce_data_weekly <- as.data.frame(ce_data_weekly)   # type cast to data.frame
summary(ce_data_weekly)
```

```
##  mrp_category       product_analytic_sub_category      week
##  Length:1144        Length:1144                   Min.   : 1.00
##  Class :character   Class :character              1st Qu.:17.00
##  Mode  :character   Mode  :character              Median :29.00
##                                                   Mean   :28.83
##                                                   3rd Qu.:42.00
##                                                   Max.   :53.00
##       gmv              product_mrp       list_price         units
##  Min.   :       131   Min.   :   499   Min.   :   129   Min.   :    1.0
##  1st Qu.:   137118   1st Qu.:  2268   1st Qu.:  1239   1st Qu.:    9.0
##  Median :   534200   Median :  7228   Median :  3184   Median :  195.5
##  Mean   :  3513803   Mean   : 28840   Mean   : 18759   Mean   : 1463.2
##  3rd Qu.:  4143242   3rd Qu.: 42156   3rd Qu.: 29103   3rd Qu.: 2009.8
##  Max.   :123984747   Max.   :299999   Max.   :224990   Max.   :52928.0
##    Promotion             sla         procurement_sla
##  Min.   :-0.06454   Min.   : 1.000   Min.   : -1.000
##  1st Qu.: 0.24046   1st Qu.: 4.660   1st Qu.:  2.414
##  Median : 0.37395   Median : 5.503   Median :  2.735
##  Mean   : 0.38982   Mean   : 5.536   Mean   :  5.817
##  3rd Qu.: 0.47658   3rd Qu.: 6.282   3rd Qu.:  3.500
##  Max.   : 0.98389   Max.   :15.000   Max.   :113.247
```

```r
#Converting dummy variables for mrp_category
fact1<- as.data.frame(model.matrix(~ce_data_weekly[, 1], data = ce_data_weekly))
fact1 <- fact1[,-1]
names(fact1) <- c("cat_luxury", "cat_mid", "cat_premium")
# Merge dummy variables with the main data frame
ce_data_weekly <- cbind(ce_data_weekly[,-10], fact1)
# ****************************************************************************
#                    MERGING DATA ----
# ****************************************************************************

# . . . .   Merge MediaInvestment & NPS ----
media_nps <- merge(media_adstk, monthlyNPS_weekly, by = 'week', all.x = TRUE)


# . . . .   Merge Sales & SaleDays
data <- merge(ce_data_weekly, sale_days, by = 'week', all.x = TRUE)
data$sale_days[is.na(data$sale_days)] <- 0 #no sale days in that week
```

```r
# . . . .   Merge data & holidays
data <- merge(data, holiday_days, by = 'week', all.x = TRUE)
data$holidays[is.na(data$holidays)] <- 0 #no sale days in that week

#. . . .   Merge Data & Media_NPS
data <- merge(data, media_nps, by = 'week', all.x = TRUE)

#Backing up the data
data_bkp <- data


#Removing mrp_category
#Keeping weeks, because I am thinking of using that to create our Training & Test Datasets
data <- data[,-c(2)]
quantile(data$product_mrp)
```

```
##         0%        25%        50%        75%       100%
##    499.000   2267.617   7228.424  42156.064 299999.000
```

```r
# ***************************************************************************
#           CREATE A NEW DATASET WITH ONLY THE IMP VARIABLES ----
# ***************************************************************************

camera_accessory_data <- subset(data, product_analytic_sub_category=="CameraAccessory")
home_audio_data       <- subset(data, product_analytic_sub_category=="HomeAudio")
gaming_accessory_data <- subset(data, product_analytic_sub_category=="GamingAccessory")

# . . . . Data Cleanup ----
# remove sub_category column
camera_accessory_data <- camera_accessory_data[,-2]
home_audio_data <- home_audio_data[,-2]
gaming_accessory_data <- gaming_accessory_data[,-2]


# . . . . Save Intrim Data ----
write.csv(data, file = "./intrim/eleckart.csv",row.names=FALSE)
write.csv(camera_accessory_data, file = './intrim/cameraAccessory.csv',row.names = FALSE)
write.csv(home_audio_data, file = './intrim/homeAudio.csv',row.names = FALSE)
write.csv(gaming_accessory_data, file = './intrim/gamingAccessory.csv',row.names = FALSE)
```