1_DataPreparation.R

atchirc

Mon Apr 10 00:14:30 2017

```
# ******************************************************************************
#             MARKET MIX  MODELLING
#
#       PGDDA ( IIIT Bangalore )
#       April 2017
#       AtchiReddy (atchireddi@gmail.com)
#
#       DATA CLEANING & DATA PREPARATION
#
# ******************************************************************************
```

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

# ******************************************************************************
#                   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 ----
# ******************************************************************************
ce_data <- read.csv('../input/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 ...
```

```
atchircUtils::naSummary(ce_data)
```

```
##                              Vars  NAS     class    perNAS
## 1                       ï..fsn_id    0 character 0.0000000
## 2                      order_date    0 character 0.0000000
## 3                            Year    0   integer 0.0000000
## 4                           Month    0   integer 0.0000000
## 5                        order_id    0   numeric 0.0000000
## 6                   order_item_id    0   numeric 0.0000000
## 8                           units    0   integer 0.0000000
## 9                   deliverybdays    0 character 0.0000000
## 10                  deliverycdays    0 character 0.0000000
## 11     s1_fact.order_payment_type    0 character 0.0000000
## 12                            sla    0   integer 0.0000000
## 15 product_analytic_super_category    0 character 0.0000000
## 16      product_analytic_category    0 character 0.0000000
## 17  product_analytic_sub_category    0 character 0.0000000
## 18      product_analytic_vertical    0 character 0.0000000
## 19                    product_mrp    0   integer 0.0000000
## 20       product_procurement_sla    0   integer 0.0000000
## 7                            gmv 4904   numeric 0.2974241
## 13                       cust_id 4904   numeric 0.2974241
## 14                       pincode 4904   numeric 0.2974241
```

```
# ***************************************************************************
#                   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
# . . . .   Missing Values ----
ce_data <- ce_data[,-c(9,10)]   # Omit 'deliverybday' & 'deliverycdays'

ce_data <- na.omit(ce_data)   # 4904 missing values, can be ignored


# . . . .   Outlier Treatment ----
# Remove orders before July'15 and after June'16
ce_data <- ce_data[ce_data$order_date>as.Date('2015-6-30'),]
ce_data <- ce_data[ce_data$order_date<as.Date('2016-7-1'),]

# . . . .   Correct Data Types ----

# 'order_id', 'order_item_id', 'cust_id', 'pincode' are qualitative data
#  having numeric values, let's convert them to character type

ce_data <- cbind(ce_data[,-c(5,6,11,12)],
          sapply(ce_data[,c(5,6,11,12)],as.character) )   # operate on interested columns

# gmv & mrp make non-zero
ce_data$gmv <- ce_data$gmv+1
ce_data$product_mrp <- ce_data$product_mrp+1
```

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

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

# replace spaces
ce_data$product_analytic_vertical <- gsub(" +","",ce_data$product_analytic_vertical)

# compute discount gmv
ce_data$discount_gmv <- as.integer(ce_data$gmv/ce_data$units)

# discount
ce_data$discount  <- 100.0-(ce_data$discount_gmv*100/ce_data$product_mrp)



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

# Drop 'fsn_id', 'order_data', 'Year', 'Month', 'sl_fact.order_type',
#  'order_id', 'order_item_id', 'cust_id', 'pincode',

ce_data <- ce_data[,-c(1,2,3,7,9,15,16,17,18)]

str(ce_data)
```

```
## 'data.frame':    1643311 obs. of  12 variables:
##  $ Month                        : int  10 10 10 10 10 10 10 10 10 10 ...
##  $ gmv                          : num  6401 6901 1991 1691 1619 ...
##  $ units                        : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ sla                          : int  5 7 10 4 6 5 6 5 7 8 ...
##  $ product_analytic_category    : chr  "CameraAccessory" "CameraAccessory" "CameraAccessory" "Camera
##  $ product_analytic_sub_category: chr  "CameraAccessory" "CameraAccessory" "CameraAccessory" "Camera
##  $ product_analytic_vertical    : chr  "CameraTripod" "CameraTripod" "CameraTripod" "CameraTripod" .
##  $ product_mrp                  : num  7191 7191 2100 2100 2100 ...
##  $ product_procurement_sla      : int  0 0 3 3 3 5 5 5 5 5 ...
##  $ week                         : num  16 17 17 16 16 16 16 16 18 17 ...
##  $ discount_gmv                 : int  6401 6901 1991 1691 1619 3325 3696 3696 3696 3696 ...
##  $ discount                     : num  10.99 4.03 5.19 19.48 22.9 ...
```

```
ce_data_weekly <-  ce_data %>%
                     group_by(product_analytic_category,
                              product_analytic_sub_category,
                              product_analytic_vertical,
                              Month,
                              week) %>%
                     summarize(gmv=sum(gmv),
```

```
                                product_mrp=mean(product_mrp),
                                units=sum(units),
                                sla=mean(sla),
                                procurement_sla=mean(product_procurement_sla))

str(ce_data_weekly)
```

```
## Classes 'grouped_df', 'tbl_df', 'tbl' and 'data.frame':  3394 obs. of  10 variables:
##  $ product_analytic_category    : chr  "Camera" "Camera" "Camera" "Camera" ...
##  $ product_analytic_sub_category: chr  "Camera" "Camera" "Camera" "Camera" ...
##  $ product_analytic_vertical    : chr  "Camcorders" "Camcorders" "Camcorders" "Camcorders" ...
##  $ Month                        : int  1 1 1 1 1 1 2 2 2 2 ...
##  $ week                         : num  27 28 29 30 31 32 32 33 34 35 ...
##  $ gmv                          : num  121058 474609 235794 550487 551529 ...
##  $ product_mrp                  : num  10996 37531 20083 30335 31328 ...
##  $ units                        : int  4 13 11 19 18 4 18 19 18 12 ...
##  $ sla                          : num  7 6.54 5.27 6.11 6.17 ...
##  $ procurement_sla              : num  -0.5 1.462 0.545 1.632 0.667 ...
##  - attr(*, "vars")=List of 4
##   ..$ : symbol product_analytic_category
##   ..$ : symbol product_analytic_sub_category
##   ..$ : symbol product_analytic_vertical
##   ..$ : symbol Month
##  - attr(*, "drop")= logi TRUE
```

```
# ************************************************************************
#                   DATA PREPARATION ----
# ************************************************************************

# # Create subset for categories 'CameraAccessory', 'HomeAudio', 'GamingAccesory'
# camera_accessory_data <- subset(ce_data, product_analytic_sub_category=="CameraAccessory")
# home_audio_data       <- subset(ce_data, product_analytic_sub_category=="HomeAudio")
# gaming_accessory_data <- subset(ce_data, product_analytic_sub_category=="GamingAccessory")
#
#
# # ************************************************************************
# #                    Save CLEAN DATA ----
# # ************************************************************************
#
write.csv(ce_data, '../intrim/ConsumeElectronics.csv')
# write.csv(camera_accessory_data, '../intrim/CameraAccesory.csv')
# write.csv(home_audio_data, '../intrim/HomeAudio.csv')
# write.csv(gaming_accessory_data, '../intrim/GamingAccessory.csv')




# Observations :
#     1. why -ve values in  'Cust_id' and 'pincode'
#     2. Order_id/cust_id/pincode has any naming convention
#     3. fsn_id has any naming convention
#     4. what is NPS score
#     5. should special sale days be marked in the dataset
#     6. which day to be considered start of week
```

```
#     7. Few More Insights in product list Tab
#     8. Elaboration on Media Investment
#     9. product details are given in order dataset,
#           why aditional documentation,
#    10. How to ratio NPS & media spend to weekly
#    11. gmv vs mrp vs units. ( is gmv gt/lt mrp)
#    12. product_mrp is zero..??
#
#
#
#
# Data Augmentation :
#     1. Derive day
#     2. Derive week
#     3. Derive Month
#     4. Mark Special Sale Dates
#     5.
```

‘‘‘