1_DataPreparation.R

atchirc

Sat Apr 08 08:18:01 2017

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

```
# ***************************************************************************
#                   LOAD LIBRARY ----
# ***************************************************************************



# 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
```

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

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

# . . . .   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(5,6,17,18)],
           sapply(ce_data[,c(5,6,17,18)],as.character) )   # operate on interested columns


# ***************************************************************************
#                   Feature Engineering ----
# ***************************************************************************


# ***************************************************************************
#                   Save CLEAN DATA ----
# ***************************************************************************




# 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



# Data Augmentation :
#     1. Derive day
#     2. Derive week
#     3. Derive Month
#     4. Mark Special Sale Dates
#     5.
```

“‘