EDA Customer Purchase Journey Prediction

SumbarajuAditya

1/8/2022

```
library(reshape2)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(grid)
library(gridExtra)
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:reshape2':
##
##     dcast, melt
```

```
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 4.0.5
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
library(xts)
```

```
## Warning: package 'xts' was built under R version 4.0.5
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.0.5
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
##
## Attaching package: 'xts'
```

```
## The following objects are masked from 'package:data.table':
##
##     first, last
```

Including Plots

Loading dataset from four input files

```
df.TFDNov <- read.csv("C:/BU/DSC680/project1/data/D11", header = T, sep = ";",
    stringsAsFactors = FALSE)
df.TFDDec <- read.csv("C:/BU/DSC680/project1/data/D12", header = T, sep = ";",
    stringsAsFactors = FALSE)
df.TFDJan <- read.csv("C:/BU/DSC680/project1/data/D01", header = T, sep = ";",
    stringsAsFactors = FALSE)
df.TFDFeb <- read.csv("C:/BU/DSC680/project1/data/D02", header = T, sep = ";",
    stringsAsFactors = FALSE)
DNov_cols <- colnames(df.TFDNov)
DDec_cols <- colnames(df.TFDDec)
DJan_cols <- colnames(df.TFDJan)
DFeb_cols <- colnames(df.TFDFeb)
identical(DNov_cols, DDec_cols)
```

```
## [1] TRUE
```

```
identical(DNov_cols, DJan_cols)
```

```
## [1] TRUE
```

```
identical(DNov_cols, DFeb_cols)
```

```
## [1] TRUE
```

```
df.TFD0 <- rbind(df.TFDJan, df.TFDFeb)
df.TFD1 <- rbind(df.TFDNov, df.TFDDec)
df.TFD_FullSet <- rbind(df.TFD0, df.TFD1)
```

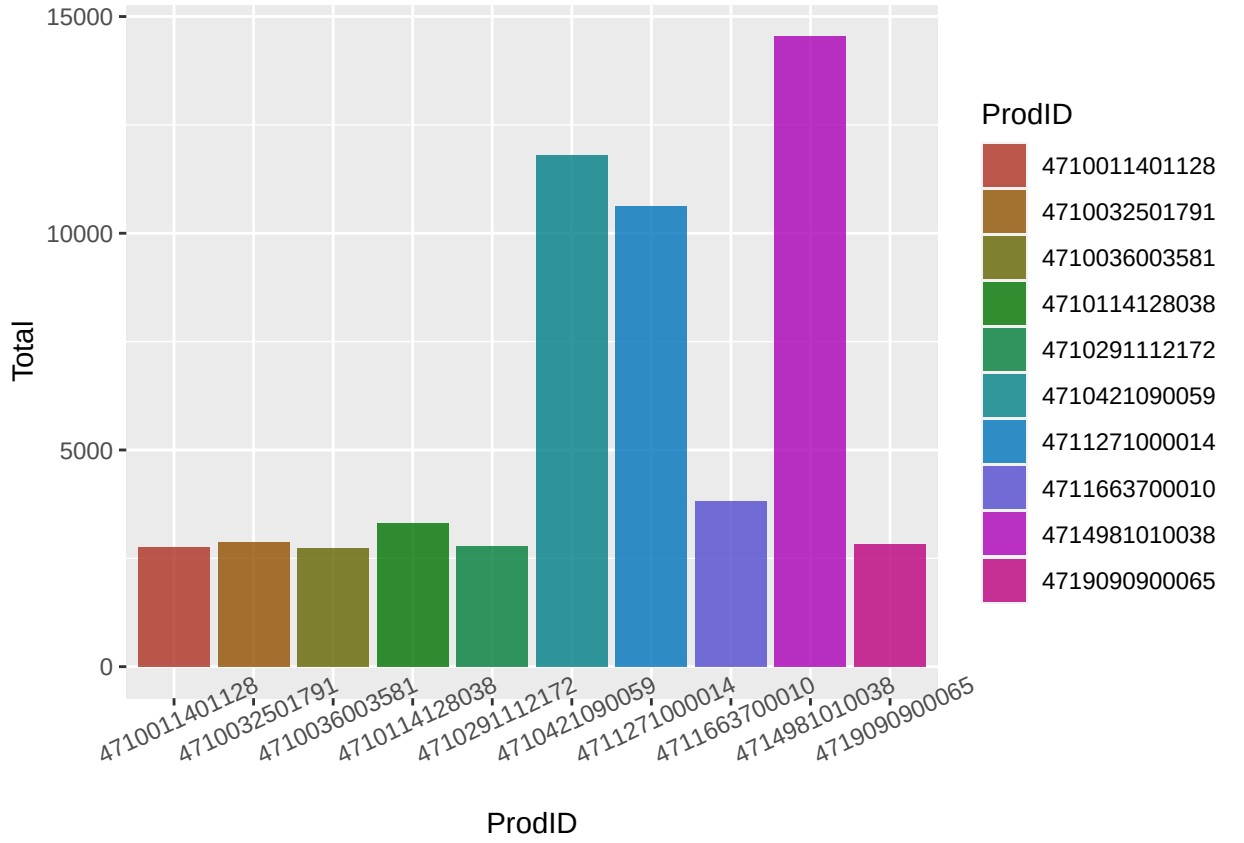
```
TFcols <- c("DateTime", "CustID", "Age_cat", "ResArea",
    "ProdSub", "ProdID", "Cost", "Asset", "SalesPrice")
colnames(df.TFD_FullSet) <- TFcols
dt.TF <- as.data.table(df.TFD_FullSet)
dt.TF$DateTime <- as.POSIXct(dt.TF$DateTime)
dt.TF$ProdID <- as.factor(dt.TF$ProdID)
df.TFD_FullSet$DateTime <- as.POSIXlt(df.TFD_FullSet$DateTime)
df.TFD_FullSet$Age_cat = as.factor(df.TFD_FullSet$Age_cat)
df.TFD_FullSet$ResArea = as.factor(df.TFD_FullSet$ResArea)
df.TFD_FullSet$CustID = as.factor(df.TFD_FullSet$CustID)
df.TFD_FullSet$ProdSub = as.factor(df.TFD_FullSet$ProdSub)
```

top selling products

```
dt.TF_prodid <- dt.TF[, list(TotAmount = sum(Cost)),
    by = list(DateTime, ProdID, ResArea)]
topPr <- dt.TF_prodid[, list(Total = sum(TotAmount)),
    by = .(ProdID)]
topPr <- topPr[order(-Total)]
head(topPr)
```

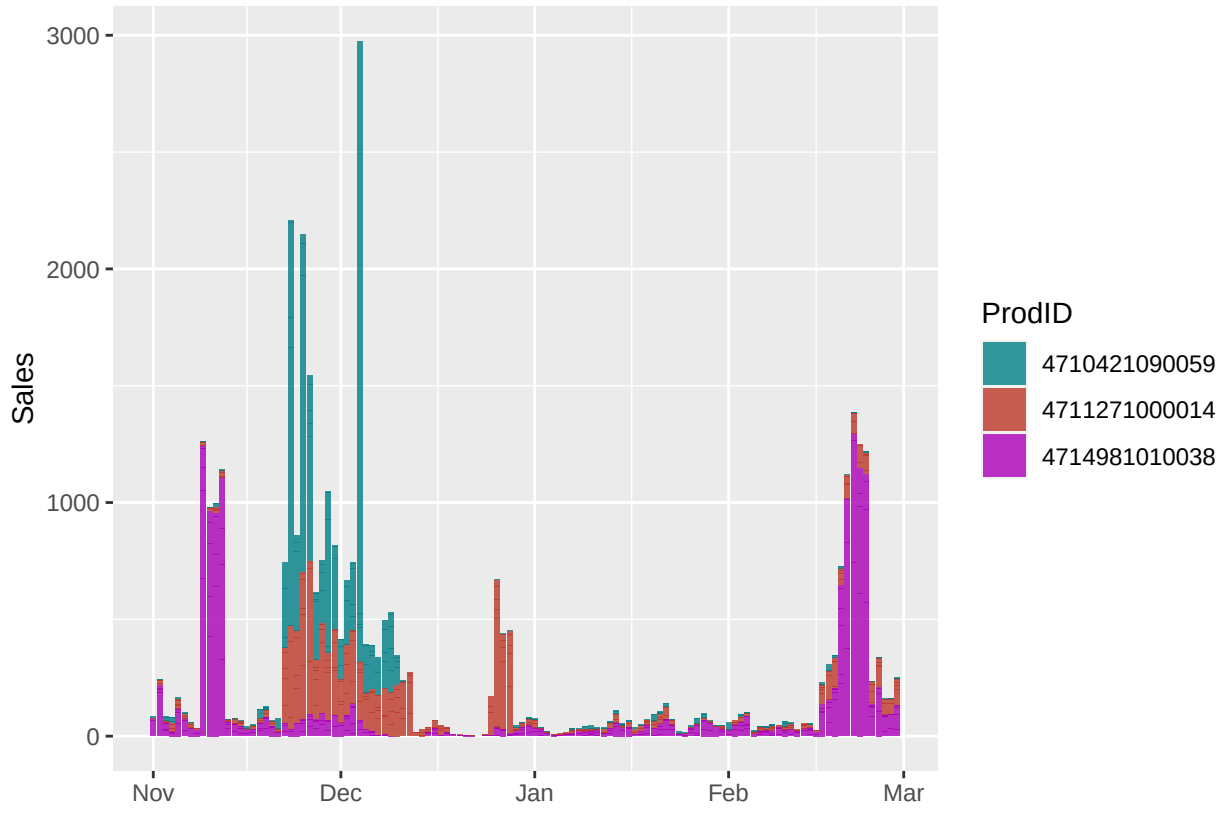
```
##           ProdID Total
## 1: 4714981010038 14537
## 2: 4710421090059 11790
## 3: 4711271000014 10615
## 4: 4711663700010  3810
## 5: 4710114128038  3322
## 6: 4710032501791  2865
```

```
ggplot(topPr[Total >= 2700, ]) + geom_bar(aes(x = ProdID,
    y = Total, fill = ProdID), stat = "identity", alpha = 0.8) +
    theme(axis.text.x = element_text(angle = 25)) +
    scale_fill_hue(l = 40)
```

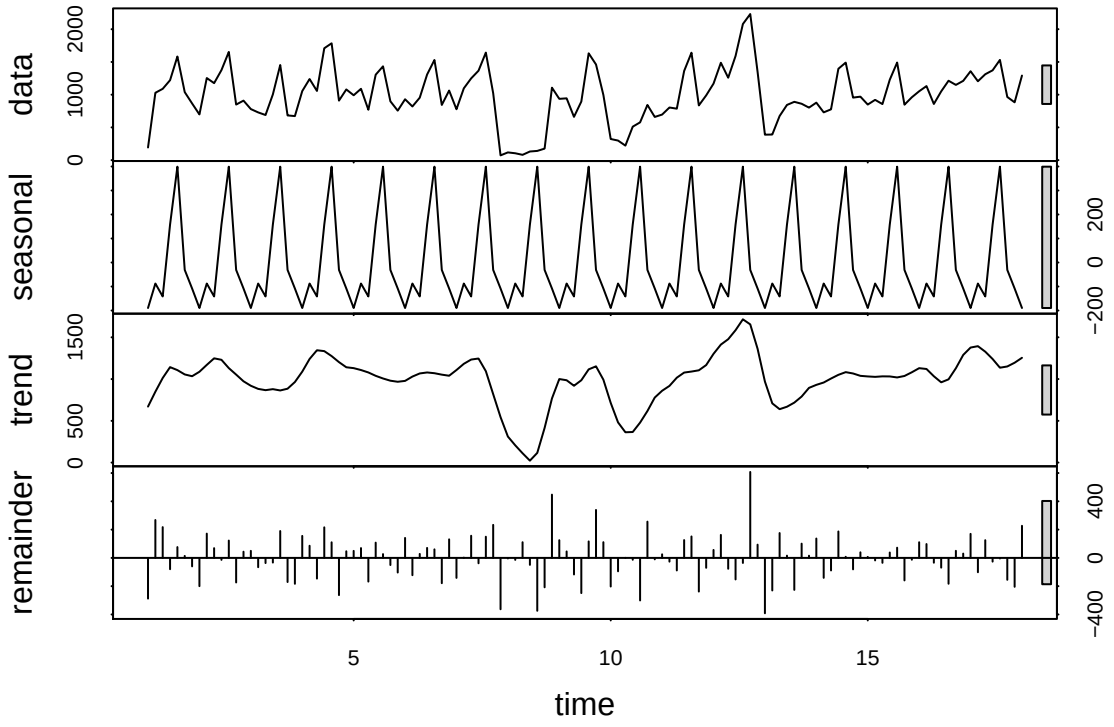


Number Products Sold per Day

```
dt.TF_prodid <- dt.TF[, list(TAmount = sum(Cost)),
    by = list(DateTime, ProdID, ResArea)]
dt.TF_prodidtop3 <- subset(dt.TF_prodid, ProdID %in%
    c(4714981010038, 4710421090059, 4711271000014))
dt.TF_prodid_nores <- dt.TF[, list(TAmount = sum(Cost)),
    by = list(DateTime, ProdID)]
dt.TF_prodid1 <- subset(dt.TF_prodid_nores, ProdID ==
    4714981010038)
dt.TF_prodid2 <- subset(dt.TF_prodid_nores, ProdID ==
    4710421090059)
dt.TF_prodid3 <- subset(dt.TF_prodid_nores, ProdID ==
    4711271000014)
ggplot(dt.TF_prodidtop3[, list(TAmount), by = list(DateTime,
    ProdID)]) + geom_bar(aes(x = DateTime, y = TAmount,
    fill = ProdID), stat = "identity", alpha = 0.8) +
    scale_fill_manual(values = c("#007F85", "#BD3828",
        "#AB00B6")) + labs(y = "Sales", x = "")
```

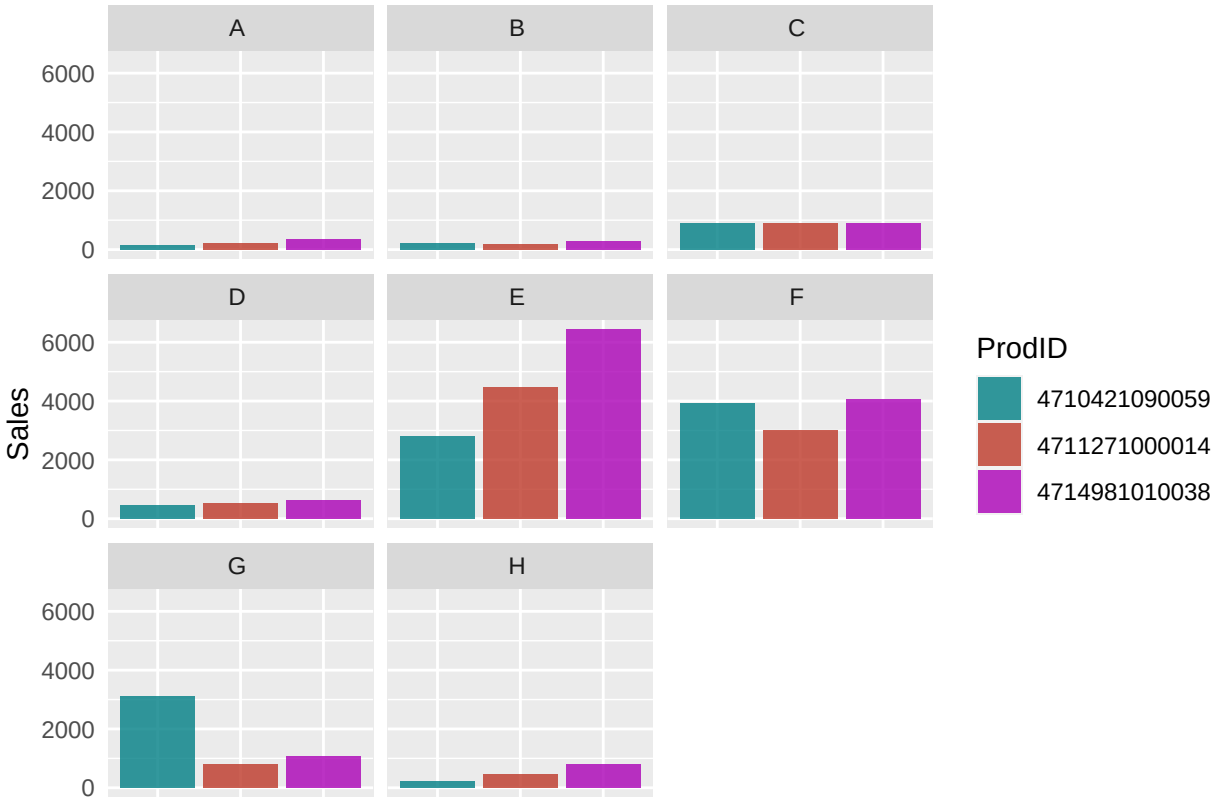



```
dt.TF_trans <- unique(dt.TF[, list(DateTime, CustID,
    ResArea)])
df.trans <- dt.TF_trans[, list(num_trans = length(CustID)),
    by = DateTime]
ts_trans <- xts(df.trans$num_trans, as.Date(df.trans$DateTime))
attr(ts_trans, "frequency") <- 7
trans_decom <- stl(as.ts(ts_trans), s.window = "periodic",
    t.window = 7)
plot(trans_decom)
```



Sales per Region observations Splitting the data into regions shows the difference in the number of transactions in each Region. 1. Region E being the busiest region with the highest number of sales. 2. Regions A and B can be seen to have low numbers of sales. Most Regions show a similar sales profile for all three top products 3. Region G shows the highest number of sales for Product 4710421090059 where as this product is recorded least sales in other regions.

```
ggplot(dt.TF_prodidtop3[, list(num_trans = sum(TAmount)),
    by = list(ProdID, ResArea)]) + geom_bar(aes(x = ProdID,
    y = num_trans, fill = ProdID), stat = "identity",
    alpha = 0.8) + scale_fill_manual(values = c("#007F85",
    "#BD3828", "#AB00B6")) + facet_wrap(~ResArea) +
    labs(y = "Sales", x = "") + theme(axis.ticks = element_blank(),
    axis.text.x = element_blank())
```


A
B
C
D
E
F
G
H
Sales
6000
4000
2000
0
ProdID
4710421090059
4711271000014
4714981010038