Case 4 Retail Relay (A)

Jonathan Ratschat / Franziska Bülck

14 11 2019

Preparing dataset

Load dataset

```
#install.packages("webshot")
#webshot::install_phantomjs()
```

```
#install.packages("readxl")
library(readxl)

Data <- read_excel("Retail-Relay-Full-Data.xls")
str(Data)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    2891 obs. of  7 variables:
##  $ OrderId     : num  262 278 294 301 302 321 333 341 353 373 ...
##  $ OrderDate   : POSIXct, format: "2009-01-11" "2009-01-20" ...
##  $ UserId      : num  47 47 47 47 47 47 47 47 47 47 ...
##  $ TotalCharges: num  50.7 26.6 38.7 53.4 14.3 ...
##  $ CommonId    : chr  "TRQKD" "4HH2S" "3TRDC" "NGAZJ" ...
##  $ PupId       : num  2 3 2 2 2 3 2 3 3 3 ...
##  $ PickupDate  : POSIXct, format: "2009-01-12" "2009-01-20" ...
```

```
summary(Data)
```

```
##     OrderId       OrderDate                       UserId
##  Min.   : 256   Min.   :2009-01-06 00:00:00   Min.   :    47
##  1st Qu.:1022   1st Qu.:2009-08-07 00:00:00   1st Qu.:  5534
##  Median :1778   Median :2009-11-15 00:00:00   Median : 42270
##  Mean   :1764   Mean   :2009-10-24 08:57:56   Mean   : 85587
##  3rd Qu.:2504   3rd Qu.:2010-01-26 00:00:00   3rd Qu.:132044
##  Max.   :3234   Max.   :2010-03-09 00:00:00   Max.   :396551
##   TotalCharges      CommonId             PupId
##  Min.   :  1.39   Length:2891        Min.   : 2.000
##  1st Qu.: 22.96   Class :character   1st Qu.: 4.000
##  Median : 44.81   Mode  :character   Median : 5.000
##  Mean   : 59.95                      Mean   : 6.848
##  3rd Qu.: 79.60                      3rd Qu.: 7.000
##  Max.   :690.98                      Max.   :20.000
##    PickupDate
##  Min.   :2009-01-06 00:00:00
##  1st Qu.:2009-08-07 00:00:00
##  Median :2009-11-16 00:00:00
##  Mean   :2009-10-25 06:48:56
##  3rd Qu.:2010-01-27 00:00:00
##  Max.   :2010-03-10 00:00:00
```

Transformation of variables

```
#Transform OrderID from num to factor
Data$OrderId <- as.factor(Data$OrderId)

#Tranform UserID from num to int
Data$UserId <- as.integer(Data$UserId)

#Transform PupID from numeric to factor
Data$PupId <- as.factor(Data$PupId)

str(Data)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    2891 obs. of  7 variables:
##  $ OrderId     : Factor w/ 2891 levels "256","257","258",..: 5 18 32 37 38 54 65 72 81 98 ...
##  $ OrderDate   : POSIXct, format: "2009-01-11" "2009-01-20" ...
##  $ UserId      : int  47 47 47 47 47 47 47 47 47 47 ...
##  $ TotalCharges: num  50.7 26.6 38.7 53.4 14.3 ...
##  $ CommonId    : chr  "TRQKD" "4HH2S" "3TRDC" "NGAZJ" ...
##  $ PupId       : Factor w/ 18 levels "2","3","4","5",..: 1 2 1 1 1 2 1 2 2 2 ...
##  $ PickupDate  : POSIXct, format: "2009-01-12" "2009-01-20" ...
```

Checking for duplicates

```
cat("The number of non-duplicate observations within the data set is",
    nrow(unique(Data)), "out of", "\n",
    nrow(Data),
    "indicating that there are",
    nrow(Data)-nrow(unique(Data)),
    "duplicates within the data set.","\n")
```

```
## The number of non-duplicate observations within the data set is 2891 out of
##  2891 indicating that there are 0 duplicates within the data set.
```

Creating the cohorts

```
library(lubridate)

# Getting the first transaction dates for each customer
join.date <- aggregate(OrderDate~UserId, Data, min, na.rm = TRUE)

# Changing the name of the column InvoiceDate to Join_Date
# since this is the first transaction date for each customer
colnames(join.date)[2] <- "JoinDate"

# Merge the Join date data to the Data data frame
Data <- merge(Data, join.date, by.x = "UserId",by.y = "UserId", all.x = TRUE)

# Creating the groups/Cohorts based on the join date month
Data$Cohort <- quarter(Data$JoinDate, with_year = TRUE, fiscal_start = 1)

#Remove join.date
rm(join.date)
```

```
str(Data)
```

```
## 'data.frame':    2891 obs. of  9 variables:
##  $ UserId      : int  47 47 47 47 47 47 47 47 47 47 ...
##  $ OrderId     : Factor w/ 2891 levels "256","257","258",..: 5 18 32 37 38 54 65 72 81 98 ...
##  $ OrderDate   : POSIXct, format: "2009-01-11" "2009-01-20" ...
##  $ TotalCharges: num  50.7 26.6 38.7 53.4 14.3 ...
##  $ CommonId    : chr  "TRQKD" "4HH2S" "3TRDC" "NGAZJ" ...
##  $ PupId       : Factor w/ 18 levels "2","3","4","5",..: 1 2 1 1 1 2 1 2 2 2 ...
##  $ PickupDate  : POSIXct, format: "2009-01-12" "2009-01-20" ...
##  $ JoinDate    : POSIXct, format: "2009-01-11 01:00:00" "2009-01-11 01:00:00" ...
##  $ Cohort      : num  2009 2009 2009 2009 2009 ...
```

```
head(Data)
```

```
##   UserId OrderId  OrderDate TotalCharges CommonId PupId PickupDate
## 1     47     262 2009-01-11        50.67    TRQKD     2 2009-01-12
## 2     47     278 2009-01-20        26.60    4HH2S     3 2009-01-20
## 3     47     294 2009-02-03        38.71    3TRDC     2 2009-02-04
## 4     47     301 2009-02-06        53.38    NGAZJ     2 2009-02-09
## 5     47     302 2009-02-06        14.28    FFYHD     2 2009-02-09
## 6     47     321 2009-02-17        29.50    HA5R3     3 2009-02-17
##              JoinDate Cohort
## 1 2009-01-11 01:00:00 2009.1
## 2 2009-01-11 01:00:00 2009.1
## 3 2009-01-11 01:00:00 2009.1
## 4 2009-01-11 01:00:00 2009.1
## 5 2009-01-11 01:00:00 2009.1
## 6 2009-01-11 01:00:00 2009.1
```

Cohort age

```
# Calculating the difference in days between the invoice date column by join date column
# There is no option for month, but getting the month from the days is simple division
Data$AgeByDay <- difftime(Data$OrderDate, Data$JoinDate, units = "days")

# Dividing the days by 30 to get the number of months
Data$AgeByMonth <- floor(Data$AgeByDay/90)
################################MAKE IT RIGHT GIRL!
```

QAU (Quarterly-Active-Users) Mixpanel

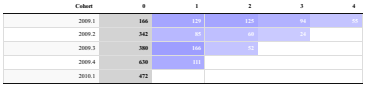
```
# Creating rows for each cohort group
# Creating columns for each value in the AgeByMonth column;0-14
# The default aggregation setup for dcast is, fun.aggregate = length
cohorts.wide <- reshape2::dcast(Data,Cohort~AgeByMonth,
                                value.var="UserId",
                                fun.aggregate = length)

# Cloning the output for retention and churn mixpanels
# to be used later
cw.retention <- cohorts.wide
cw.churn <- cohorts.wide
```

```r
# Creating 19 breaks and 20 rgb color values ranging from blue to white
breaks <- quantile(cohorts.wide[,2:6], probs = seq(.05, .95, .05), na.rm = TRUE)
colors <- sapply(round(seq(155, 80, length.out = length(breaks) + 1), 0),
                 function(x){ rgb(x,x,155, maxColorValue = 155) } )


# The Retention Mixpanel with counts
library(DT)

datatable(cohorts.wide,
              class = 'cell-border stripe',
             rownames = FALSE,
             options = list(
               ordering=F,
               dom = 't',
               pageLength = 20) ) %>%
             formatStyle("0",
                         backgroundColor = 'lightgrey',
                         fontWeight = 'bold') %>%
  formatStyle(names(cohorts.wide[c(-1,-2)]),fontWeight = 'bold',color = 'white', backgroundColor = styl
```



Cohort	0	1	2	3	4
2009.1	166	129	125	94	83
2009.2	342	85	60	54	
2009.3	380	166	52		
2009.4	630	111			
2010.1	472				

Retention Rate Mixpanel