04-v2-nilsgandlau

Nils Gandlau

23 11 2019

Cleaning the data

For consistency, it's good practice to have a clear structure on how columns should be named.

```
# Function that cleans "dirty" column names
CleanColnames <- function(data){
  oldColnames <- names(data)

  # Remove whitespaces anywhere in column names
  cleanColnames <- sapply(oldColnames, function(colname){
    return(str_replace_all(colname, " ", ""))
  })

  # Each colname should start with a small letter
  cleanColnames <- sapply(cleanColnames, function(colname){
    firstLetter <- str_to_lower(str_sub(colname, 1, 1))
    restOfString <- str_sub(colname, 2, -1)
    return(paste0(firstLetter, restOfString))
  })

  setnames(data, old = oldColnames, new = cleanColnames)
}

CleanColnames(dtFull)
CleanColnames(dtPilot)
CleanColnames(dtRetentionFull)
CleanColnames(dtRetentionPilot)

# print example
names(dtRetentionPilot)
```

```
## [1] "purchaseOccasion"      "transitionProbability" "averageBasketSize"
```

Extract new features `year`, `month` and `day` from the `orderDate` column.

```
# Function that creates year, month, day columns from a single date column
CreateYearMonthDayColumns <- function(data, dateColumn, prefix){
  data[, tempColnameYear := lubridate::year(orderDate)]
  data[, tempColnameMonth := lubridate::month(orderDate)]
  data[, tempColnameDay := lubridate::day(orderDate)]

  newColnames <- sapply(c("Year", "Month", "Day"), function(x){
    return(paste0(prefix, x))
  })

  setnames(
    data,
    old = c("tempColnameYear", "tempColnameMonth", "tempColnameDay"),
    new = newColnames
  )
}

CreateYearMonthDayColumns(dtFull, dateColumn = "orderDate", prefix = "order")
CreateYearMonthDayColumns(dtPilot, dateColumn = "orderDate", prefix = "order")
```

Next, create a column `purchaseOccasion`, which represents the j'th purchase of the i'th customer.

```
dtFull[, purchaseOccasion := frank(orderId), by = .(userId)]
dtPilot[, purchaseOccasion := frank(orderId), by = .(userId)]
```

Next, create a column `orderPeriod`. A period is a single month. Since we have a total of 15 consecutive month in the data set, we will have 15 periods.

```
# Create a (monthly) orderPeriod column t = 1, ..., 15
CreatePeriodColumn <- function(data){
  data[, orderPeriod := rep(0, nrow(data))]
  data[orderYear == 2009, orderPeriod := orderMonth]
  data[orderYear == 2010, orderPeriod := orderMonth + 12]

  data[, orderQuarter := ifelse(
    orderPeriod %in% c(1:3), 1,
    ifelse(
      orderPeriod %in% c(4:6), 2,
      ifelse(
        orderPeriod %in% c(7:9), 3,
        ifelse(
          orderPeriod %in% 10:12, 4, 5
        )
      )
    )
  )]
}

CreatePeriodColumn(dtFull)
CreatePeriodColumn(dtPilot)
```

Analysis

To get a grasp on the development customer equity of the firm, we will investigate the following:

- number of customers
  - acquisition
  - retention rates
- revenue per customer & revenue per purchase

In the analysis, we use cohorts. For purposes of simplification and also due to data sparsity, we define quarterly cohorts: ie, a cohort is defined as those customers who have joined within a given quarter $Q1, \ldots, Q5$.

Since there are 15 months total in the data, we will have $15/3 = 5$ different cohorts $k = 1, \ldots, 5$.

Defining (Quarterly) Cohorts

First we define a column that divides customers into monthly cohorts, which we then use to compute our desired quarterly cohorts.

```
# Assigning customers to cohorts 1,...,15 (15 months => 15 cohorts)
CreateCohorts <- function(data){
  data[, cohortMonthly := min(orderPeriod), by = .(userId)]
  data[, cohortQuarterly := min(orderQuarter), by = .(userId)]
}

CreateCohorts(dtFull)
CreateCohorts(dtPilot)
```

Now we generate cohort-specific features such as

- number of customers for each cohort
- number of orders per customer for each cohort
- total revenue for each cohort
- mean `totalCharges` for each cohort
- standard deviation of `totalCharges` for each cohort

```
dtFullCohortQuarterly <- dtFull %>%
  mutate(cohort = as.character(cohortQuarterly)) %>%
  group_by(cohort) %>%
  summarise(
    nCustomers = n_distinct(userId),
    nOrders = n_distinct(orderId),
    nOrdersPerCustomer = round(nOrders / nCustomers, 2),
    cumulativeTotalCharges = round(sum(totalCharges),0),
    meanTotalCharges = round(mean(totalCharges),0),
    stdDevTotalCharges = round(sd(totalCharges),0)
  )

dtFullByCohortQuarterlyAndPurchase <- dtFull %>%
  mutate(cohort = as.character(cohortQuarterly)) %>%
  group_by(cohort, purchaseOccasion) %>%
  summarise(
    nCustomers = n_distinct(userId),
    nOrders = n_distinct(orderId),
    nOrdersPerCustomer = round(nOrders / nCustomers, 2),
    cumulativeTotalCharges = round(sum(totalCharges), 0),
    meanTotalCharges = round(mean(totalCharges), 0),
    stdDevTotalCharges = round(sd(totalCharges), 0)
  )
```

Descriptive Statistics of cohorts

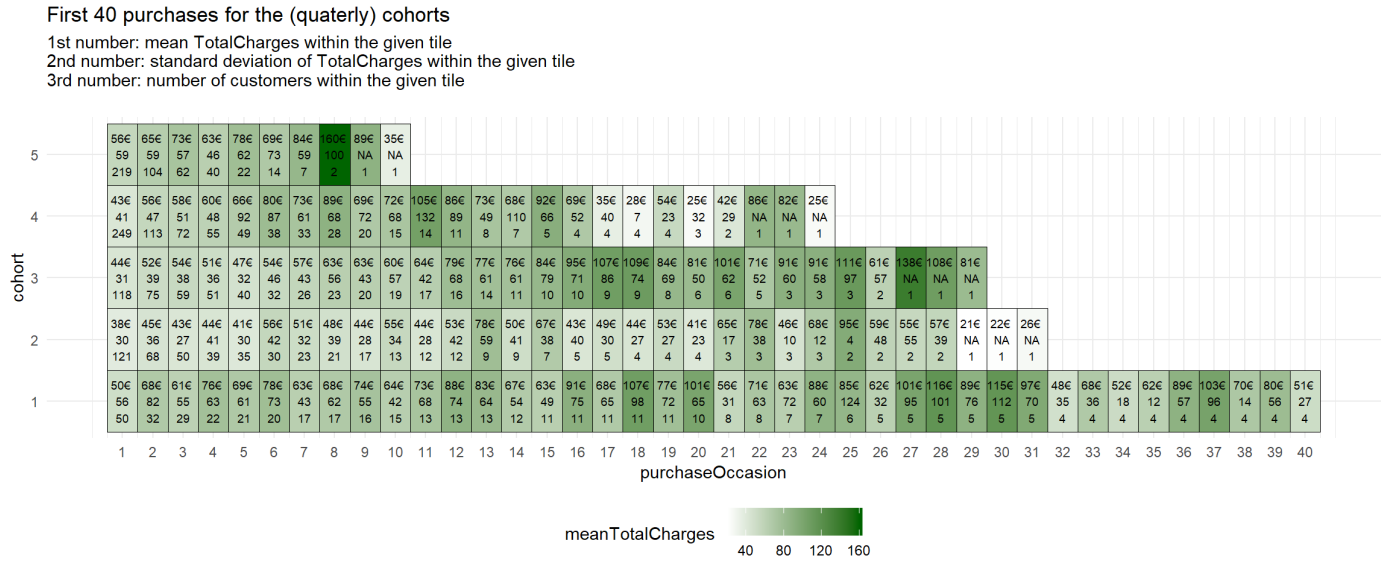
```
subtitle = "N = Number of customers
OPC = Orders Per Customers
cum€ = Total Revenue
mean€ = Mean Revenue per Customer
sd€ = Standard Deviation of Revenue Per Customer
"

ggplot(dtFullCohortQuarterly, aes(x = cohort, y = nCustomers)) +
  geom_col(alpha=0.8,  fill = "darkgreen") +
  geom_text(aes(label = paste0(
    "N=", nCustomers, "\n",
    "OPC=", nOrdersPerCustomer, "\n",
    "cum€=", round(cumulativeTotalCharges, 0), "\n",
    "mean€=", round(meanTotalCharges, 0), "\n",
    "sd€=", round(stdDevTotalCharges,0))
  ), vjust = -0.2, size = 3.5) +
  ylim(c(0, 400)) +
  theme_minimal() +
  ggtitle("Descriptive Statistics for Quarterly Cohorts", subtitle)
```

In the figure above you can observe the following:

- The total number of customers in a cohort tends to be growing. Cohort 1 had 50 customers, Cohort 2 had 121 customers, the newest cohort 5 has 219 customers.
- The number of orders a customers makes appears to be lower for newer cohorts. For example, whereas cohort 1 had 11.38 orders per customer, cohort 4 has only 2.98. However, this difference might also stem from the fact that one observes cohort 1 longer than cohort 4.
- The first cohort had the highest mean revenue per order (73€)

Overall, it appears to be that even though the number of customers increase cohort by cohort, the *quality* of the customer seems to decrease: customers from newer cohorts appear to (1) make less orders and (2) make less revenue per order.

```
# Average order-volume ($) by purchase occasion and cohort
subtitle = "1st number: mean TotalCharges within the given tile
2nd number: standard deviation of TotalCharges within the given tile
3rd number: number of customers within the given tile
"

ggplot(dtFullByCohortQuarterlyAndPurchase %>% filter(purchaseOccasion %in% 1:40),
       aes(x = purchaseOccasion, y = cohort, fill = meanTotalCharges)) +
  geom_tile(color = "black") +
  geom_text(aes(label = paste0(meanTotalCharges, "€ \n", stdDevTotalCharges, "\n", nCustomers)), size
 = 2.5) +
  scale_fill_gradient(low = "white", high = "darkgreen") +
  ggtitle("First 40 purchases for the (quaterly) cohorts", subtitle) +
  theme_minimal() +
  scale_x_continuous(breaks=1:40) +
  theme(legend.position = "bottom", legend.box = "horizontal")
```



In the figure above, one notices the following:

- Customers of cohort 1 make more purchases than those from other cohorts. That is, customers from cohort 1 appear to have the highest *transition probability* going from purchase to purchase: For example, 10% of cohort 1's customers make a 30th purchase. For cohort 5,4,3, no customer makes a 30th purchase.
- The overall trend seems to be, that newer customers make less purchases. Ie, the *transition probability* seems to be rapidly decreasing cohort by cohort.
- To compensate for making less purchases, do newer customers create more revenue per purchase? That's not too obvious, but rather no. Cohort 1's mean revenue per purchase is fairly steady and reaches from 48€ to 116€. One exception may be e.g. purchases 15 to 25 for cohort 3, where most purchases are in the region of approx. 90€.

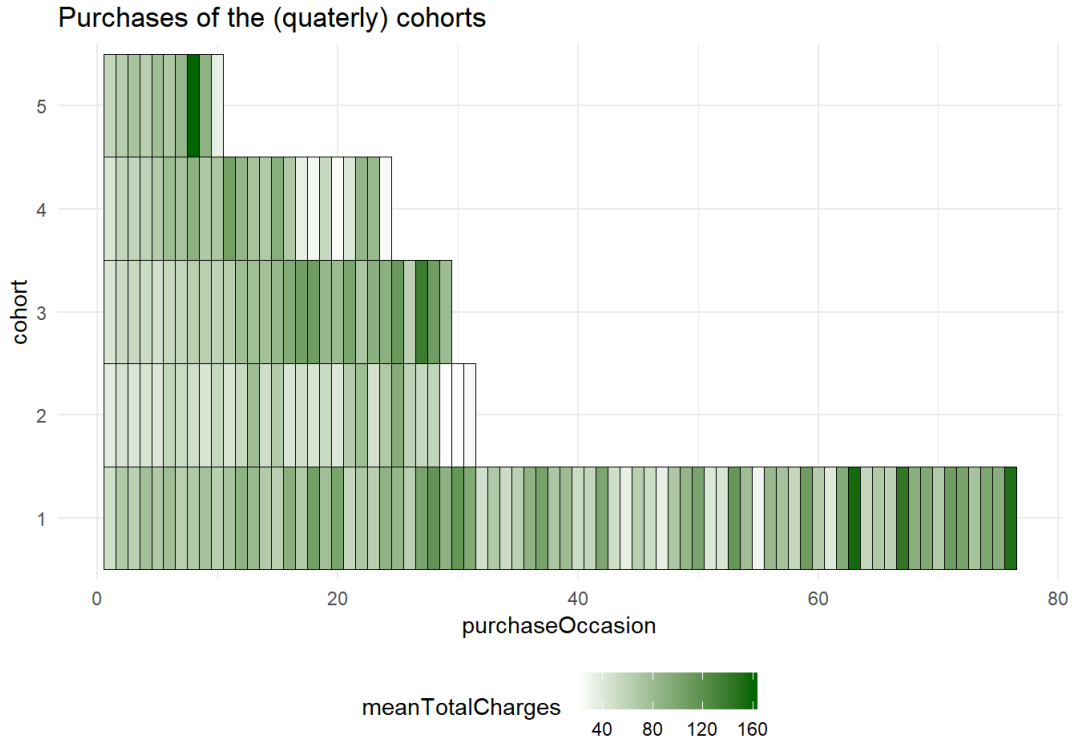
Overall, cohort 1 again performs best: Customers from cohort 1 made more purchases with steady and high revenue per purchase.

Nevertheless, when one ignores cohort 1, then the overall trend is positive in terms of revenue per purchase: E.g. cohort 2's purchases are mostly around 40-50 €, whereas cohort 5's purchases are around 60-70€.

However, there is a negative trend in terms of transition probabilities. Customers from newer cohorts make less purchases, having a drasticly less likeliness of making subsequent purchases. For example, even though there are 219 customers in cohort 5, only 10% make a 5th purchase, and only 0.5% of them make a 10th purchase. For cohort 1, the percentages where 42% and 30% respectively.

The same plot, just with all periods.

```
ggplot(dtFullByCohortQuarterlyAndPurchase,
       aes(x = purchaseOccasion, y = cohort, fill = meanTotalCharges)) +
  geom_tile(color = "black") +
  scale_fill_gradient(low = "white", high = "darkgreen") +
  ggtitle("Purchases of the (quaterly) cohorts") +
  theme_minimal() +
  theme(legend.position = "bottom", legend.box = "horizontal")
```

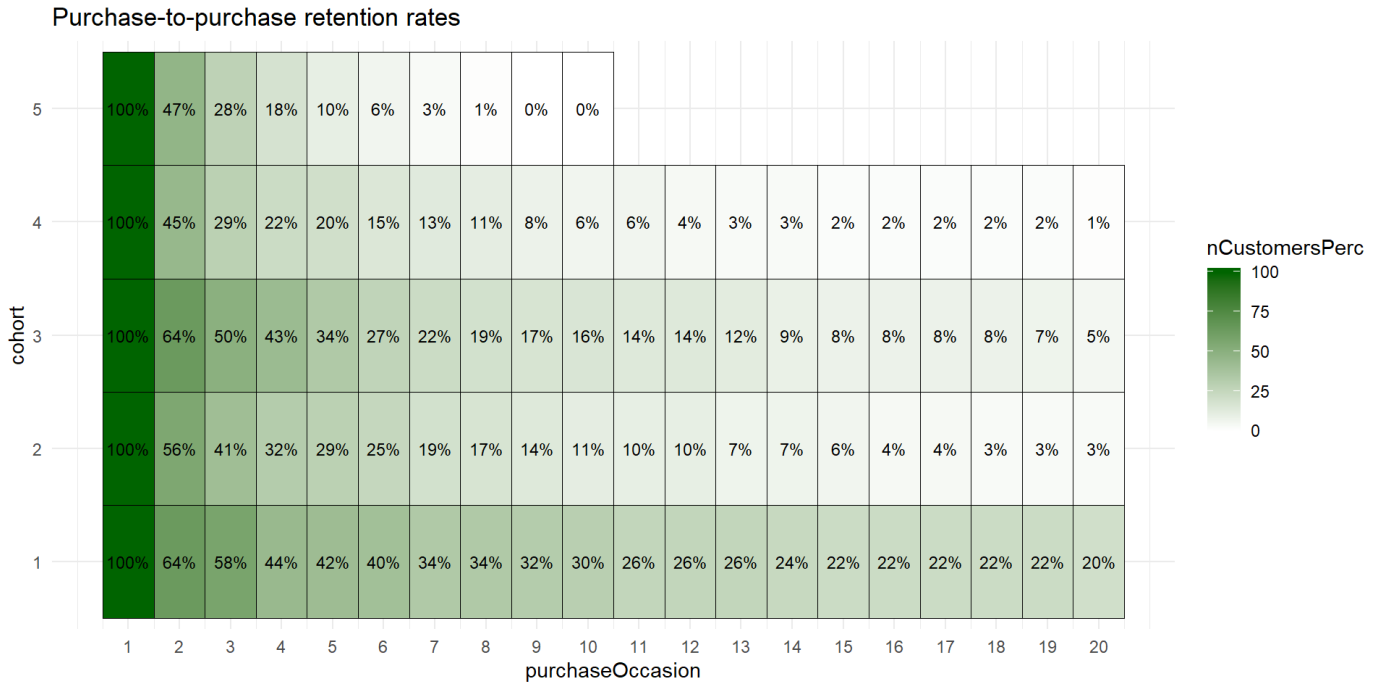


Purchases over time

```
dtPurchaseOverTime <- dtFull %>%
  mutate(cohort = as.character(cohortQuarterly)) %>%
  group_by(cohort, purchaseOccasion) %>%
  summarise(
    nCustomers = n_distinct(userId)
  )

# Normalize to 100%
nCustomers <- dtFull %>%
  mutate(cohort = as.character(cohortQuarterly)) %>%
  group_by(cohort) %>%
  summarise(totalNCustomers = n_distinct(userId))

# Join
dtPurchaseOverTime <- dtPurchaseOverTime %>%
  left_join(nCustomers) %>%
  arrange(purchaseOccasion) %>%
  ungroup() %>%
  mutate(cohort = as.factor(as.numeric(cohort))) %>%
  setDT()
```

```
## Joining, by = "cohort"
```

```
dtPurchaseOverTime[, nCustomersPerc := nCustomers / totalNCustomers]
dtPurchaseOverTime[, nCustomersPerc := round(nCustomersPerc * 100, 0)]

ggplot(dtPurchaseOverTime %>% filter(purchaseOccasion %in% 1:20), aes(x = purchaseOccasion, y = cohor
t, fill = nCustomersPerc)) +
  geom_tile(color="black") +
  scale_fill_gradient(low = "white", high = "darkgreen") +
  geom_text(aes(label = paste0(nCustomersPerc, "%")), size = 3) +
  ggtitle("Purchase-to-purchase retention rates") +
  scale_x_continuous(breaks = 1:20) +
  theme_minimal()
```



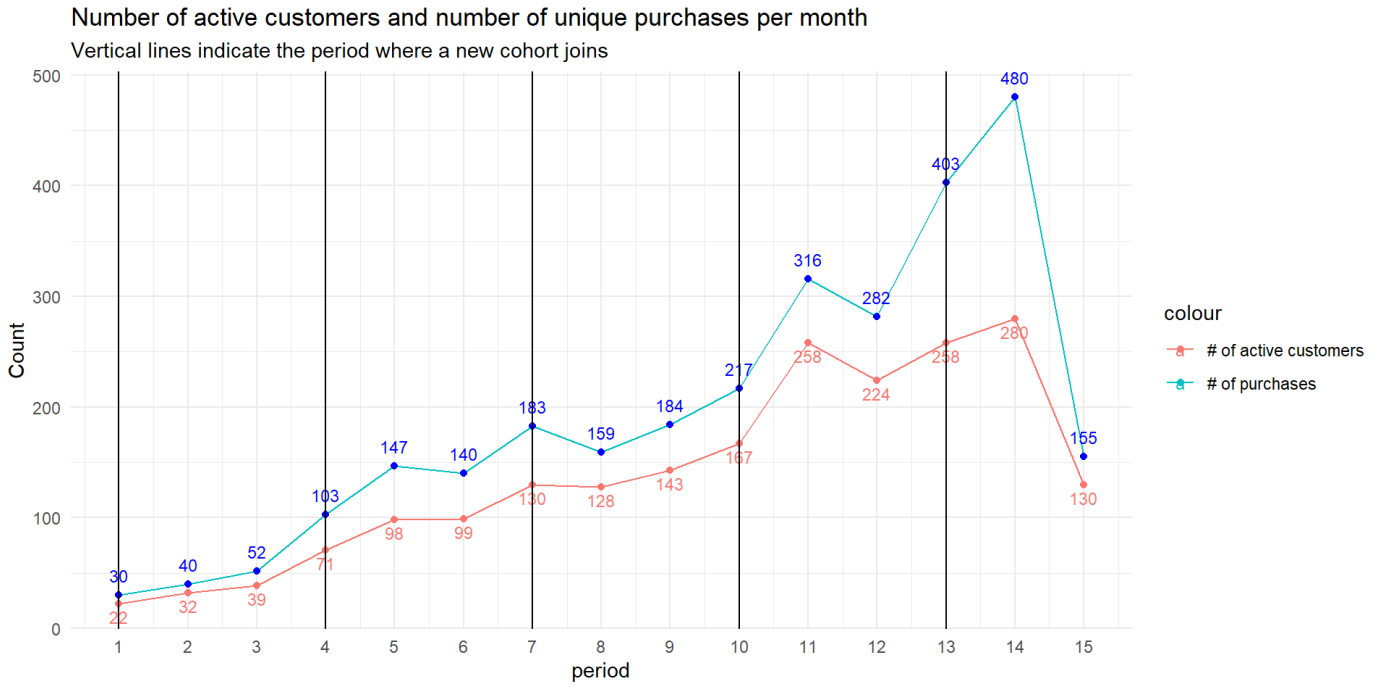
In this figure, you see a negative trend in terms of transition probabilities. Newer cohorts have worse transition probabilities. E.g., whereas 42% of customer from cohort 1 made a fifth purchase, only 10% of customers from cohort 2 make a fifth purchase.

Analysing number of customers

Number of customers

```
# Cumulative number of customers by month
dtFull[, firstPurchaseMonth := cohortMonthly, by = .(userId)]
dtFull[, lastPurchaseMonth := max(orderPeriod), by = .(userId)]

dtNCustomers <- rbindlist(lapply(1:15, function(period){
  # Compute the number of active users for the given period
  dtTemp <- dtFull[firstPurchaseMonth <= period & lastPurchaseMonth >= period, c("userId")]
  nActiveCustomers <- nrow(unique(dtTemp))

  # Compute the number of purchases for the given period
  dtTemp <- dtFull[orderPeriod == period, c("orderId")]
  nPurchases <- length(unique(dtTemp$orderId))

  return(data.table(
    period = period,
    nActiveCustomers = nActiveCustomers,
    nPurchases = nPurchases
  ))
}))
```

```
ggplot(dtNCustomers, aes(x = period, y = nActiveCustomers, color = "# of active customers")) +
  geom_line() +
  geom_point() +
  geom_text(aes(label = nActiveCustomers), vjust=1.5, size = 3) +
  geom_line(aes(x = period, y = nPurchases, color = "# of purchases")) +
  geom_point(aes(x = period, y = nPurchases), color = "blue") +
  geom_text(aes(y = nPurchases, label = nPurchases), color = "blue", vjust=-1, size = 3) +
  theme_minimal() +
  scale_x_continuous(breaks = 1:15) +
  ylab("Count") +
  geom_vline(xintercept = c(1, 4, 7, 10, 13)) +
  ggtitle(
    "Number of active customers and number of unique purchases per month",
    subtitle = "Vertical lines indicate the period where a new cohort joins")
```



The plot above shows the following:

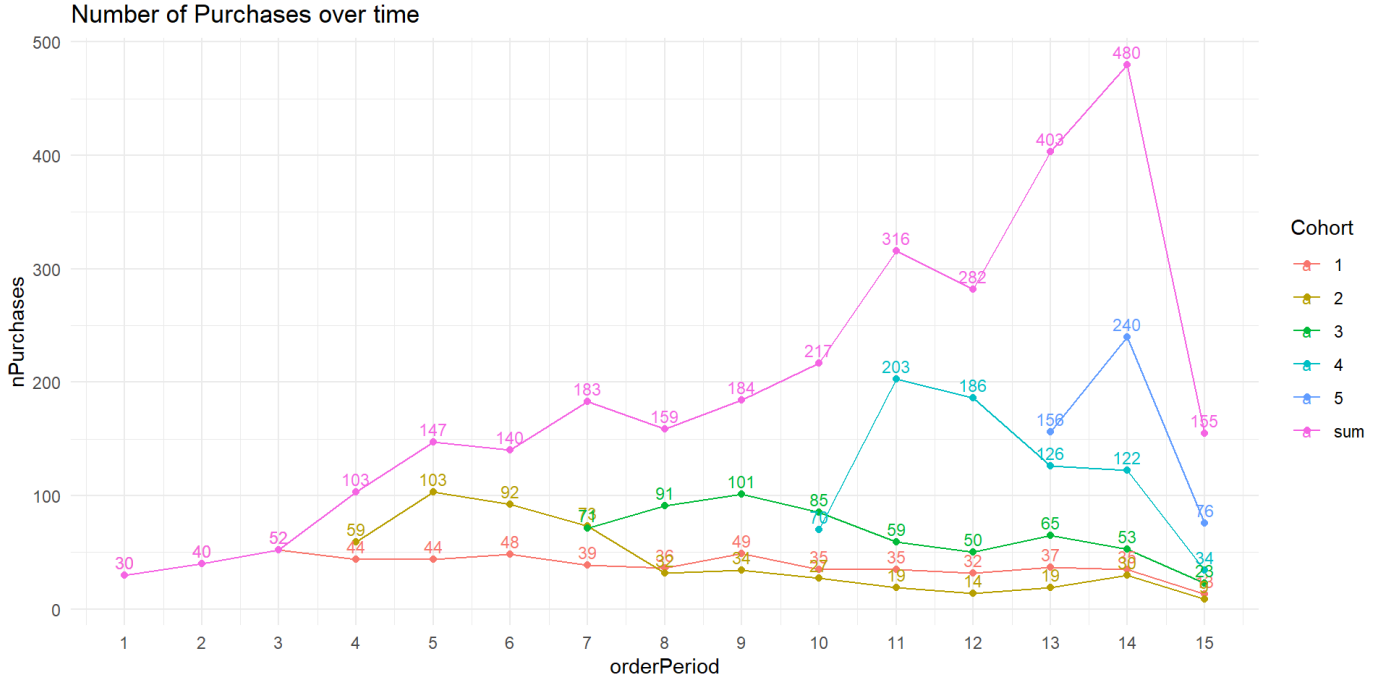
There is a growing trend both in terms of number of active customers and number of purchases. However, this positive development might be deceiving, since it aggregates over all cohorts. We will investigate how the purchase-curve decomposes into different the cohorts below.

Note: There are issues with this plot, which become clear when one sees how it's computed: Let $N(t)$ be the total number of active customers in period t. A customer is included in $N(t)$ if and only if his first purchase was *in or before* period t, and his last purchase that we have observed was *in or after* period t. So in some way, this data can be thought of being "right-censored": For a customer to be considered *active* in $t = 15$, she must have made an order in that period.

Number of purchases in each month/period by cohort

```
dtPurchases <- dtFull %>%
  mutate(cohort = as.character(cohortQuarterly)) %>%
  group_by(orderPeriod, cohort) %>%
  summarise(nPurchases = n()) %>%
  setDT()

sumPurchases <- dtFull %>%
  group_by(orderPeriod) %>%
  summarise(nPurchases = n()) %>%
  setDT()

sumPurchases[, cohort := "sum"]

dtPurchases <- rbind(dtPurchases, sumPurchases)
```

```
ggplot(dtPurchases, aes(x = orderPeriod, y = nPurchases, color = cohort)) +
  geom_line() +
  geom_point() +
  geom_text(aes(label = nPurchases), vjust = -0.5, size = 3) +
  scale_x_continuous(breaks = 1:15) +
  theme_minimal() +
  ggtitle(
    "Number of Purchases over time"
  ) +
  labs(color = "Cohort")
```



In the figure above, you see how the number of purchases is decomposed into the different cohorts.

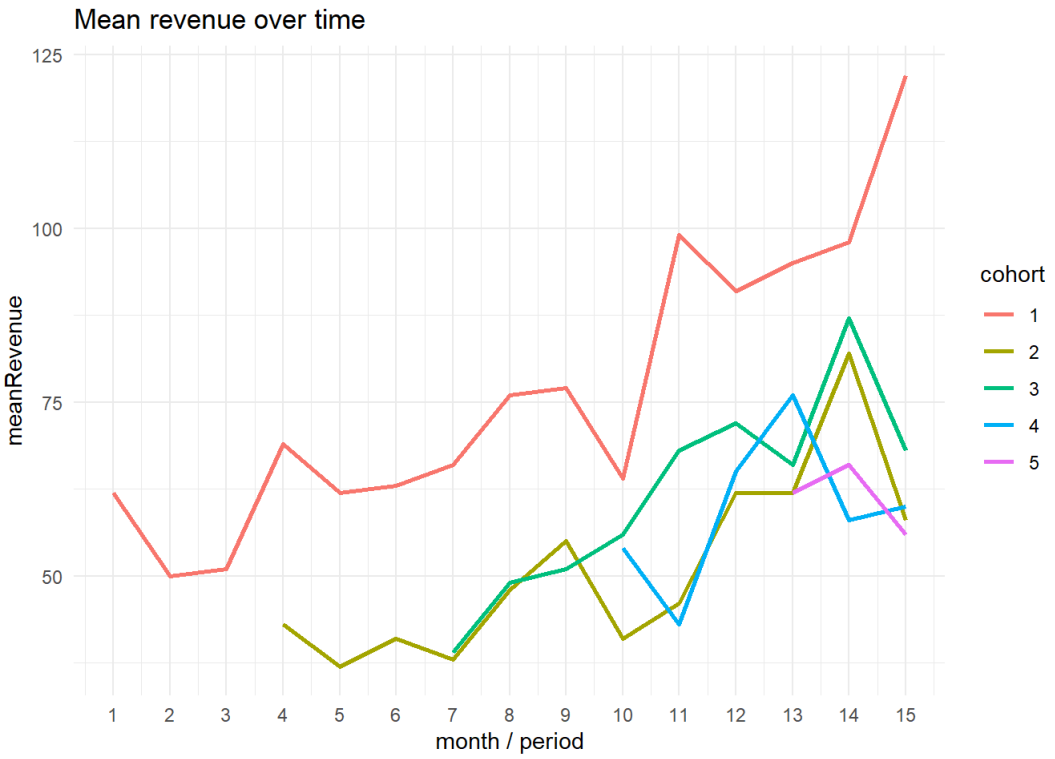
- Cohort 1 has a steady number of purchases over all 15 periods. The number of purchases are only slightly decreasing over time.
- Newest cohorts have a rapidly decreasing number of purchases over time, reinforcing the insight that we have gained when we analysed purchase-to-purchase transition probabilities.

In overall, the trend of newer cohorts having a steeply decreasing number-of-purchases-over-time trend is scary: Even though the firm is able to pour in more customers each period, those customers are only making very few purchases before "dying out".

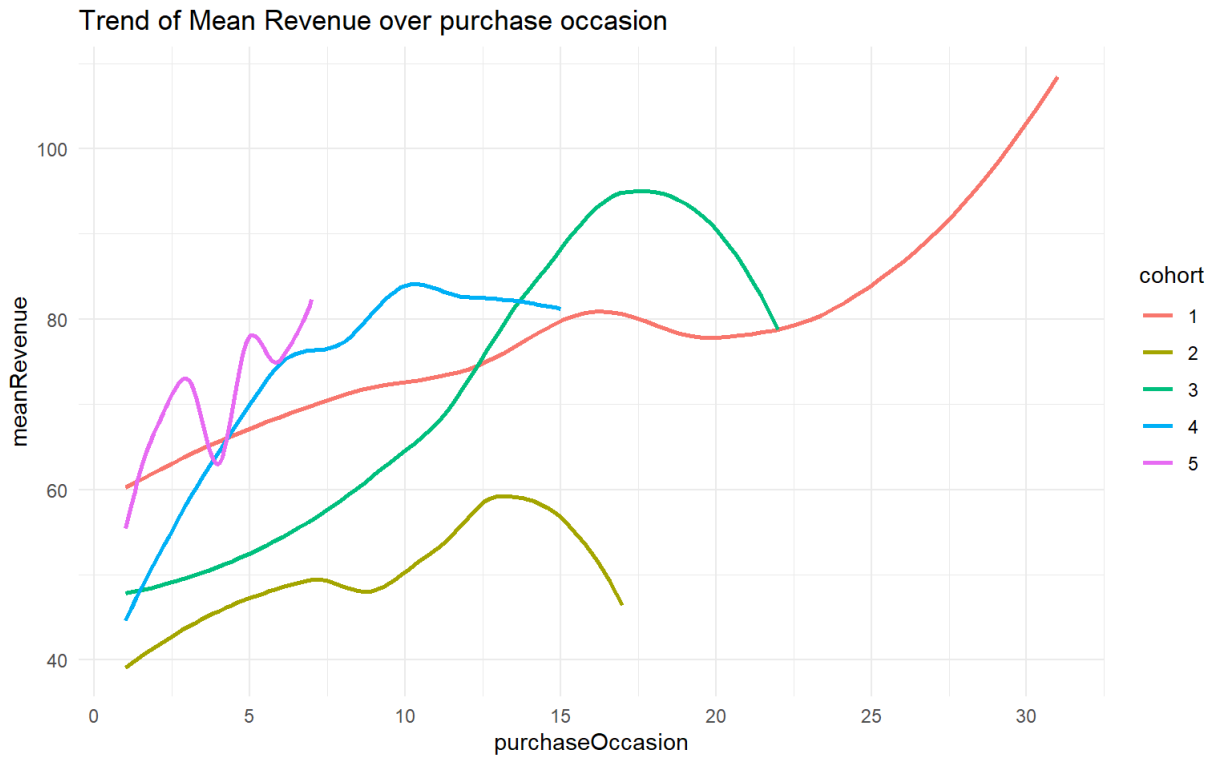
Average revenue per user by cohort

Finally, we analyse the average revenue on a cohort-level.

```
dtRev <- dtFull %>%
  mutate(cohort = as.character(cohortQuarterly)) %>%
  group_by(cohort, orderPeriod) %>%
  summarise(
    meanRevenue = round(mean(totalCharges), 0),
    nCustomers = n_distinct(userId)
  )

ggplot(dtRev, aes(x = orderPeriod, y = meanRevenue, color = cohort)) +
  geom_line(size = 1) +
  scale_x_continuous(breaks = 1:15) +
  theme_minimal() +
  ggtitle("Mean revenue over time") +
  xlab("month / period")
```




```
dtRev <- dtFull %>%
  mutate(cohort = as.character(cohortQuarterly)) %>%
  group_by(cohort, purchaseOccasion) %>%
  summarise(
    meanRevenue = round(mean(totalCharges), 0),
    nCustomers = n_distinct(userId)
  ) %>%
  filter(nCustomers >= 5)

ggplot(dtRev, aes(x = purchaseOccasion, y = meanRevenue, fill = cohort, color = cohort)) +
  stat_smooth(se = FALSE) +
  scale_x_continuous(breaks = seq(0, 70, 5)) +
  theme_minimal() +
  ggtitle("Trend of Mean Revenue over purchase occasion")
```



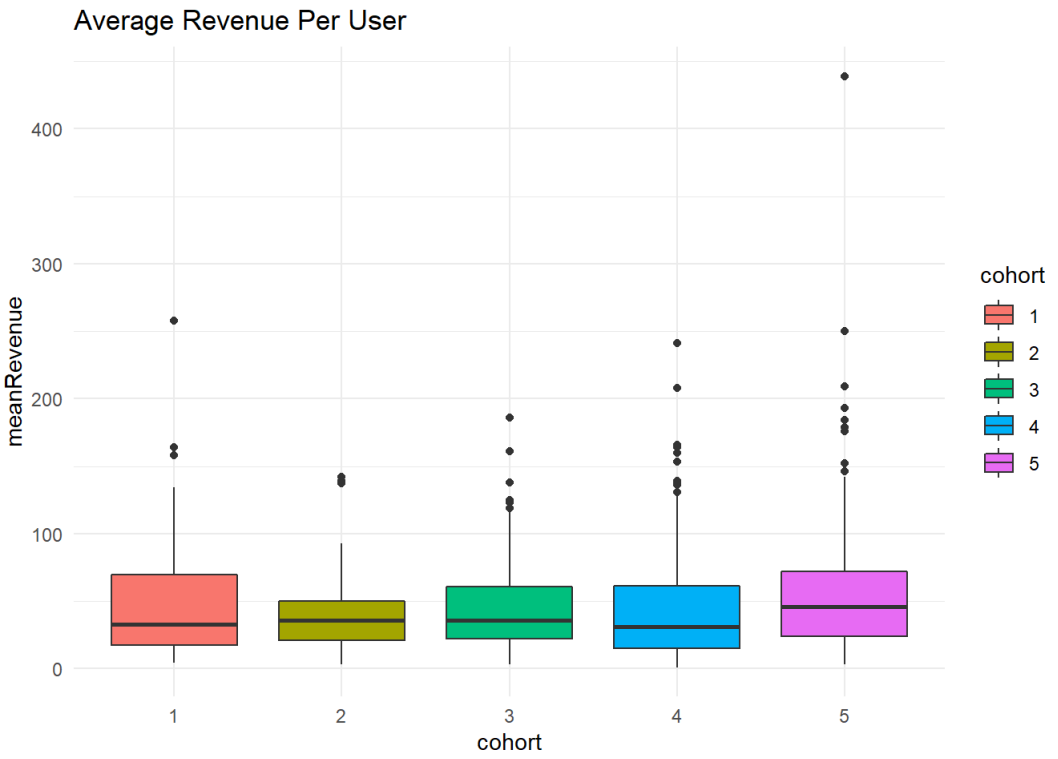
The figure above shows that within each cohort, there is an overall positive trend in the mean revenue per purchase: That is, on average, each additional purchase is larger (in terms of revenue) than the one before. This may be because as you become familiar with the service, you may order more.

From a business perspective, you may want to exploit this increasing trend – that is, you want to keep customers longer at the firm, such that they make more purchases and each time the next purchase is larger (in terms of revenue) than the one before.

However, we have seen that a customer from newer cohorts make only a few purchases before dying out. If the firm was able to hold customers for longer, it would be able to make use of the larger-orders-over-time effect.

Average revenue per user by cohort

```
dtARPU <- dtFull %>%
  mutate(cohort = as.character(cohortQuarterly)) %>%
  group_by(userId, cohort) %>%
  summarise(
    meanRevenue = round(mean(totalCharges), 0),
    sumRevenue = sum(totalCharges)
  )

ggplot(dtARPU, aes(x = cohort, y = meanRevenue, fill = cohort)) +
  geom_boxplot() +
  ggtitle("Average Revenue Per User") +
  theme_minimal()
```



Summary

The overall number of customers are growing. Each new cohort has more customers.

There is a highly negative trend in terms of purchase-to-purchase transition probabilities: Customers from newer cohorts make only a few purchases until dying out: Only 10% make a 5th purchase, only 0.1% make a 10th purchase. For the first cohort, these numbers were 42% and 30% respectively.

New cohorts have on average a slightly higher *mean revenue per order* than older ones. This is a positive trend. However, it is not exploited well enough because these customers make less purchases.

Note: Here, I just report my findings from the data. Due to time issues, I can't provide a more thorough analysis with CAC, comparisons with other companies, etc. Sorry!