Case 6 - Olist

Jonathan Ratschat, Franziska Bülck

06.12.2019

# Event Log Definitions

Event Log Process mining assumes the existence of an event log where each event refers to a case, an activity/event, and a point in time. An event log can be seen as a collection of cases and a case can be seen as a trace/sequence of events.

Case A case is a coherent process iteration.

Trace A trace is a sequence of activity labels.

# Preparation of dataset

#install.packages("shiny")  
library(shiny)

## Importing dataset

#load event log data  
load("Olist\_Event\_Log\_2019.RData")

#change data types

event\_log$voucher <- as.factor(event\_log$voucher)  
event\_log$customer\_state <- as.factor(event\_log$customer\_state)

## Exploring Data

str(event\_log)

## Classes 'eventlog', 'tbl\_df', 'tbl' and 'data.frame': 144986 obs. of 30 variables:  
## $ order\_id : chr "000aed2e25dbad2f9ddb70584c5a2ded" "000aed2e25dbad2f9ddb70584c5a2ded" "000aed2e25dbad2f9ddb70584c5a2ded" "000aed2e25dbad2f9ddb70584c5a2ded" ...  
## $ activity : Factor w/ 6 levels "approve","customer provides review",..: 5 1 3 4 6 2 5 1 3 4 ...  
## $ timestamp : POSIXct, format: "2018-05-11 20:33:38" "2018-05-11 20:57:03" ...  
## $ review\_score : num 1 1 1 1 1 1 3 3 3 3 ...  
## $ boleto : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ credit\_card : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ customer\_id : Factor w/ 24120 levels "000161a058600d5901f007fab4c27140",..: 24118 24118 24118 24118 24118 24118 14306 14306 14306 14306 ...  
## $ debit\_card : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ nmb\_pay\_types : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ not\_defined : num 0 0 0 0 0 0 0 0 0 0 ...  
## $ order\_estimated\_delivery\_date: POSIXct, format: "2018-05-22" "2018-05-22" ...  
## $ total\_installments : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ total\_pay\_value : num 153 153 153 153 153 ...  
## $ voucher : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ customer\_state : Factor w/ 27 levels "AC","AL","AM",..: 26 26 26 26 26 26 26 26 26 26 ...  
## $ nmb\_items : num 1 1 1 1 1 1 1 1 1 1 ...  
## $ unique\_items : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ nmb\_sellers : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ max\_shipping\_date : POSIXct, format: "2018-05-16 20:57:03" "2018-05-16 20:57:03" ...  
## $ sum\_items\_value : num 144 144 144 144 144 144 47.9 47.9 47.9 47.9 ...  
## $ nmb\_categories : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ avg\_desc\_length : num 116 116 116 116 116 116 735 735 735 735 ...  
## $ avg\_ph\_quality : num 1 1 1 1 1 1 5 5 5 5 ...  
## $ max\_weight : num 468 468 468 468 468 468 800 800 800 800 ...  
## $ max\_length : num 23 23 23 23 23 23 20 20 20 20 ...  
## $ max\_height : num 18 18 18 18 18 18 20 20 20 20 ...  
## $ max\_width : num 19 19 19 19 19 19 20 20 20 20 ...  
## $ status : Factor w/ 1 level "complete": 1 1 1 1 1 1 1 1 1 1 ...  
## $ activity\_instance : chr "1" "2" "3" "4" ...  
## $ .order : int 1 2 3 4 5 6 7 8 9 10 ...  
## - attr(\*, "case\_id")= chr "order\_id"  
## - attr(\*, "activity\_id")= chr "activity"  
## - attr(\*, "activity\_instance\_id")= chr "activity\_instance"  
## - attr(\*, "lifecycle\_id")= chr "status"  
## - attr(\*, "resource\_id")= chr "customer\_id"  
## - attr(\*, "timestamp")= chr "timestamp"

summary(event\_log)

## order\_id activity   
## Length:144986 approve :24120   
## Class :character customer provides review :24253   
## Mode :character deliever\_carrier :24120   
## deliever\_customer :24120   
## purchase :24120   
## sent out review survey to customer:24253   
##   
## timestamp review\_score boleto   
## Min. :2016-10-04 09:16:33 Min. :1.000 Min. :0.0000   
## 1st Qu.:2017-09-21 12:09:55 1st Qu.:4.000 1st Qu.:0.0000   
## Median :2018-01-29 15:18:45 Median :5.000 Median :0.0000   
## Mean :2018-01-09 08:40:18 Mean :4.158 Mean :0.1969   
## 3rd Qu.:2018-05-10 23:30:58 3rd Qu.:5.000 3rd Qu.:0.0000   
## Max. :2018-10-26 21:36:41 Max. :5.000 Max. :1.0000   
## NA's :6   
## credit\_card customer\_id debit\_card   
## Min. :0.0000 9e29cde4ddb42f9330bcf3c9bc23d177: 10 Min. :0.00000   
## 1st Qu.:1.0000 0033823ee55671ac5317d423291333c2: 8 1st Qu.:0.00000   
## Median :1.0000 060feb3bc91804b596b106fa63834799: 8 Median :0.00000   
## Mean :0.7732 080f252f36da3da6eb5b04407af199f7: 8 Mean :0.01528   
## 3rd Qu.:1.0000 08a113f40963196742d5538a95776e34: 8 3rd Qu.:0.00000   
## Max. :1.0000 090c8832aab02cc3a2a256a01b7c1068: 8 Max. :1.00000   
## (Other) :144936   
## nmb\_pay\_types not\_defined order\_estimated\_delivery\_date total\_installments  
## Min. :1.000 Min. :0 Min. :2016-11-24 00:00:00 Min. : 1.000   
## 1st Qu.:1.000 1st Qu.:0 1st Qu.:2017-10-05 00:00:00 1st Qu.: 1.000   
## Median :1.000 Median :0 Median :2018-02-16 00:00:00 Median : 2.000   
## Mean :1.022 Mean :0 Mean :2018-01-25 19:54:50 Mean : 2.974   
## 3rd Qu.:1.000 3rd Qu.:0 3rd Qu.:2018-05-28 00:00:00 3rd Qu.: 4.000   
## Max. :2.000 Max. :0 Max. :2018-10-23 00:00:00 Max. :25.000   
##   
## total\_pay\_value voucher customer\_state nmb\_items unique\_items   
## Min. : 10.07 0:139700 SP :60922 Min. : 1.000 Min. :1.000   
## 1st Qu.: 61.78 1: 5286 RJ :17876 1st Qu.: 1.000 1st Qu.:1.000   
## Median : 105.28 MG :17718 Median : 1.000 Median :1.000   
## Mean : 158.49 RS : 8178 Mean : 1.141 Mean :1.038   
## 3rd Qu.: 176.16 PR : 7232 3rd Qu.: 1.000 3rd Qu.:1.000   
## Max. :4950.34 SC : 5368 Max. :15.000 Max. :7.000   
## (Other):27692   
## nmb\_sellers max\_shipping\_date sum\_items\_value   
## Min. :1.000 Min. :2016-10-08 13:26:12 Min. : 0.85   
## 1st Qu.:1.000 1st Qu.:2017-09-21 06:36:43 1st Qu.: 45.90   
## Median :1.000 Median :2018-01-26 14:28:44 Median : 86.90   
## Mean :1.013 Mean :2018-01-08 14:15:40 Mean : 135.75   
## 3rd Qu.:1.000 3rd Qu.:2018-05-10 16:15:19 3rd Qu.: 149.90   
## Max. :5.000 Max. :2018-09-18 21:10:15 Max. :4799.00   
##   
## nmb\_categories avg\_desc\_length avg\_ph\_quality max\_weight   
## Min. :1.000 Min. : 4.0 Min. : 1.000 Min. : 0   
## 1st Qu.:1.000 1st Qu.: 354.0 1st Qu.: 1.000 1st Qu.: 300   
## Median :1.000 Median : 610.0 Median : 2.000 Median : 700   
## Mean :1.008 Mean : 795.3 Mean : 2.241 Mean : 2123   
## 3rd Qu.:1.000 3rd Qu.: 999.0 3rd Qu.: 3.000 3rd Qu.: 1825   
## Max. :3.000 Max. :3988.0 Max. :20.000 Max. :30000   
## NA's :2164 NA's :2164 NA's :12   
## max\_length max\_height max\_width status   
## Min. : 7.00 Min. : 2.00 Min. : 7.00 complete:144986   
## 1st Qu.: 18.00 1st Qu.: 8.00 1st Qu.: 15.00   
## Median : 25.00 Median : 13.00 Median : 20.00   
## Mean : 30.18 Mean : 16.63 Mean : 23.06   
## 3rd Qu.: 38.00 3rd Qu.: 20.00 3rd Qu.: 30.00   
## Max. :105.00 Max. :105.00 Max. :118.00   
## NA's :12 NA's :12 NA's :12   
## activity\_instance .order   
## Length:144986 Min. : 1   
## Class :character 1st Qu.: 36247   
## Mode :character Median : 72494   
## Mean : 72494   
## 3rd Qu.:108740   
## Max. :144986   
##

Findings:

* 144,986 orders
* Missing data -timestamp: 6 -max\_width: 12 -max\_weight: 12 -max\_height: 12 -avg\_ph\_quality: 2164

# Analysis

## 1. Descriptive Analysis

#install.packages("bupaR")  
library(bupaR)

### a. How many casesdoes the event log contrain

n\_cases(event\_log)

## [1] 24120

### b. How many events does the event log contain

n\_events(event\_log)

## [1] 144986

### c. How many unique activities are executed throughout the process

#install.packages("plyr")  
library(plyr)  
  
n\_activities(event\_log)

## [1] 6

activity\_labels(event\_log)

## [1] purchase approve   
## [3] deliever\_carrier deliever\_customer   
## [5] sent out review survey to customer customer provides review   
## 6 Levels: approve customer provides review ... sent out review survey to customer

Six activities are executed throughout the process. The activities are purchase, approve, deliever\_carrier, deliever\_customer, sent out review survey to customer and customer provides review.

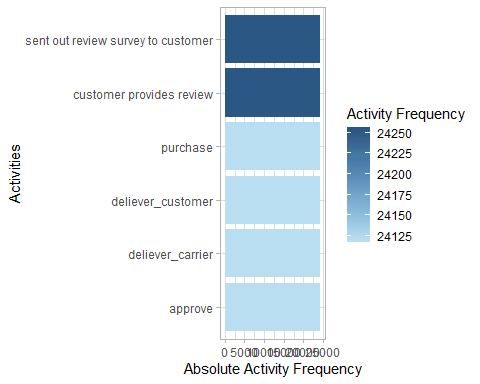
### d. Summary of the event log and visualization of interesting variables

summary(event\_log)

## Number of events: 144986  
## Number of cases: 24120  
## Number of traces: 31  
## Number of distinct activities: 6  
## Average trace length: 6.011028  
##   
## Start eventlog: NA  
## End eventlog: NA

## order\_id activity   
## Length:144986 approve :24120   
## Class :character customer provides review :24253   
## Mode :character deliever\_carrier :24120   
## deliever\_customer :24120   
## purchase :24120   
## sent out review survey to customer:24253   
##   
## timestamp review\_score boleto   
## Min. :2016-10-04 09:16:33 Min. :1.000 Min. :0.0000   
## 1st Qu.:2017-09-21 12:09:55 1st Qu.:4.000 1st Qu.:0.0000   
## Median :2018-01-29 15:18:45 Median :5.000 Median :0.0000   
## Mean :2018-01-09 08:40:18 Mean :4.158 Mean :0.1969   
## 3rd Qu.:2018-05-10 23:30:58 3rd Qu.:5.000 3rd Qu.:0.0000   
## Max. :2018-10-26 21:36:41 Max. :5.000 Max. :1.0000   
## NA's :6   
## credit\_card customer\_id debit\_card   
## Min. :0.0000 9e29cde4ddb42f9330bcf3c9bc23d177: 10 Min. :0.00000   
## 1st Qu.:1.0000 0033823ee55671ac5317d423291333c2: 8 1st Qu.:0.00000   
## Median :1.0000 060feb3bc91804b596b106fa63834799: 8 Median :0.00000   
## Mean :0.7732 080f252f36da3da6eb5b04407af199f7: 8 Mean :0.01528   
## 3rd Qu.:1.0000 08a113f40963196742d5538a95776e34: 8 3rd Qu.:0.00000   
## Max. :1.0000 090c8832aab02cc3a2a256a01b7c1068: 8 Max. :1.00000   
## (Other) :144936   
## nmb\_pay\_types not\_defined order\_estimated\_delivery\_date total\_installments  
## Min. :1.000 Min. :0 Min. :2016-11-24 00:00:00 Min. : 1.000   
## 1st Qu.:1.000 1st Qu.:0 1st Qu.:2017-10-05 00:00:00 1st Qu.: 1.000   
## Median :1.000 Median :0 Median :2018-02-16 00:00:00 Median : 2.000   
## Mean :1.022 Mean :0 Mean :2018-01-25 19:54:50 Mean : 2.974   
## 3rd Qu.:1.000 3rd Qu.:0 3rd Qu.:2018-05-28 00:00:00 3rd Qu.: 4.000   
## Max. :2.000 Max. :0 Max. :2018-10-23 00:00:00 Max. :25.000   
##   
## total\_pay\_value voucher customer\_state nmb\_items unique\_items   
## Min. : 10.07 0:139700 SP :60922 Min. : 1.000 Min. :1.000   
## 1st Qu.: 61.78 1: 5286 RJ :17876 1st Qu.: 1.000 1st Qu.:1.000   
## Median : 105.28 MG :17718 Median : 1.000 Median :1.000   
## Mean : 158.49 RS : 8178 Mean : 1.141 Mean :1.038   
## 3rd Qu.: 176.16 PR : 7232 3rd Qu.: 1.000 3rd Qu.:1.000   
## Max. :4950.34 SC : 5368 Max. :15.000 Max. :7.000   
## (Other):27692   
## nmb\_sellers max\_shipping\_date sum\_items\_value   
## Min. :1.000 Min. :2016-10-08 13:26:12 Min. : 0.85   
## 1st Qu.:1.000 1st Qu.:2017-09-21 06:36:43 1st Qu.: 45.90   
## Median :1.000 Median :2018-01-26 14:28:44 Median : 86.90   
## Mean :1.013 Mean :2018-01-08 14:15:40 Mean : 135.75   
## 3rd Qu.:1.000 3rd Qu.:2018-05-10 16:15:19 3rd Qu.: 149.90   
## Max. :5.000 Max. :2018-09-18 21:10:15 Max. :4799.00   
##   
## nmb\_categories avg\_desc\_length avg\_ph\_quality max\_weight   
## Min. :1.000 Min. : 4.0 Min. : 1.000 Min. : 0   
## 1st Qu.:1.000 1st Qu.: 354.0 1st Qu.: 1.000 1st Qu.: 300   
## Median :1.000 Median : 610.0 Median : 2.000 Median : 700   
## Mean :1.008 Mean : 795.3 Mean : 2.241 Mean : 2123   
## 3rd Qu.:1.000 3rd Qu.: 999.0 3rd Qu.: 3.000 3rd Qu.: 1825   
## Max. :3.000 Max. :3988.0 Max. :20.000 Max. :30000   
## NA's :2164 NA's :2164 NA's :12   
## max\_length max\_height max\_width status   
## Min. : 7.00 Min. : 2.00 Min. : 7.00 complete:144986   
## 1st Qu.: 18.00 1st Qu.: 8.00 1st Qu.: 15.00   
## Median : 25.00 Median : 13.00 Median : 20.00   
## Mean : 30.18 Mean : 16.63 Mean : 23.06   
## 3rd Qu.: 38.00 3rd Qu.: 20.00 3rd Qu.: 30.00   
## Max. :105.00 Max. :105.00 Max. :118.00   
## NA's :12 NA's :12 NA's :12   
## activity\_instance .order   
## Length:144986 Min. : 1   
## Class :character 1st Qu.: 36247   
## Mode :character Median : 72494   
## Mean : 72494   
## 3rd Qu.:108740   
## Max. :144986   
##

#install.packages("ggplot2")  
library(ggplot2)  
  
#install.packages("edeaR")  
library(edeaR)  
  
ActivityFrequency <- event\_log %>% activity\_frequency(level = "activity")   
  
plot(ActivityFrequency)



#install.packages("processmonitR")  
library(processmonitR)  
  
#activity\_dashboard(event\_log)

## 2. Control-Flow Perspective and Structure of the Process

### a. How many different variants of the processs are in the event log

n\_traces(event\_log)

## [1] 31

number\_of\_traces(event\_log)

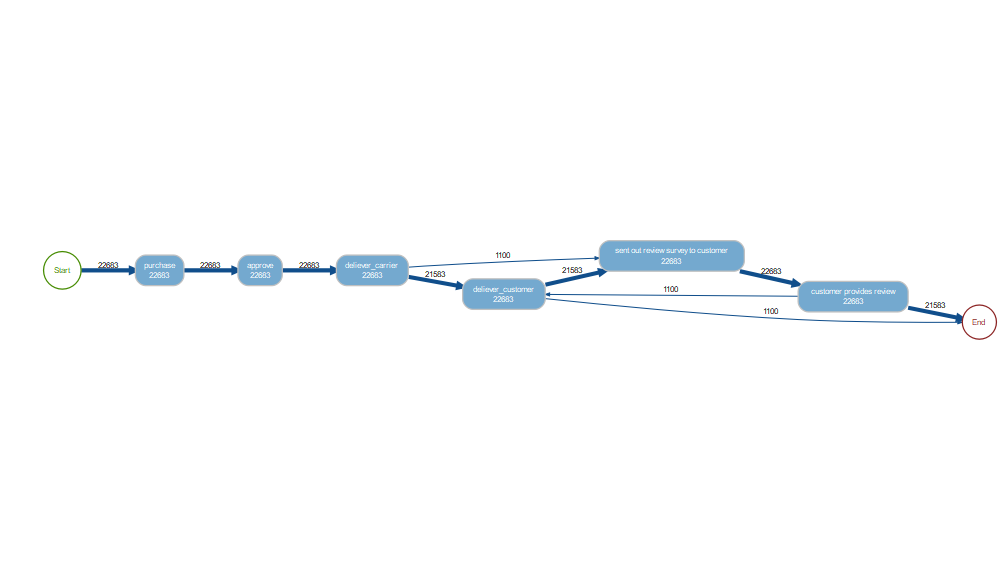
## # A tibble: 1 x 2  
## absolute average\_coverage  
## <dbl> <dbl>  
## 1 31 778.

trace\_list(event\_log)

## # A tibble: 31 x 3  
## trace absolute\_frequen~ relative\_frequen~  
## <chr> <int> <dbl>  
## 1 purchase,approve,deliever\_carrier,deliev~ 21583 0.895   
## 2 purchase,approve,deliever\_carrier,sent o~ 1100 0.0456   
## 3 purchase,approve,deliever\_carrier,sent o~ 883 0.0366   
## 4 purchase,approve,deliever\_customer,sent ~ 2 0.0000829  
## 5 purchase,deliever\_carrier,deliever\_custo~ 5 0.000207   
## 6 purchase,approve,sent out review survey ~ 47 0.00195   
## 7 purchase,approve,deliever\_customer,sent ~ 4 0.000166   
## 8 purchase,approve,deliever\_carrier,deliev~ 28 0.00116   
## 9 purchase,approve,deliever\_carrier,deliev~ 54 0.00224   
## 10 purchase,approve,deliever\_carrier,sent o~ 24 0.000995   
## # ... with 21 more rows

### b. What is the most frequent variant of the process

#Visual solution with a process map showing the process variants with 90% or more presence (thus the most frequent)  
event\_log %>%  
 filter\_trace\_frequency(perc = 0.9) %>%  
 process\_map()



#Alternative solution  
TraceVariantFreq <- count(traces(event\_log))  
  
which.max(TraceVariantFreq$absolute\_frequency)

## [1] 3

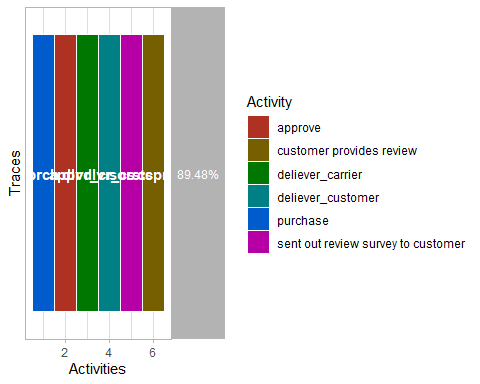
TraceVariantFreq$trace [3]

## [1] "purchase,approve,deliever\_carrier,deliever\_customer,sent out review survey to customer,customer provides review"

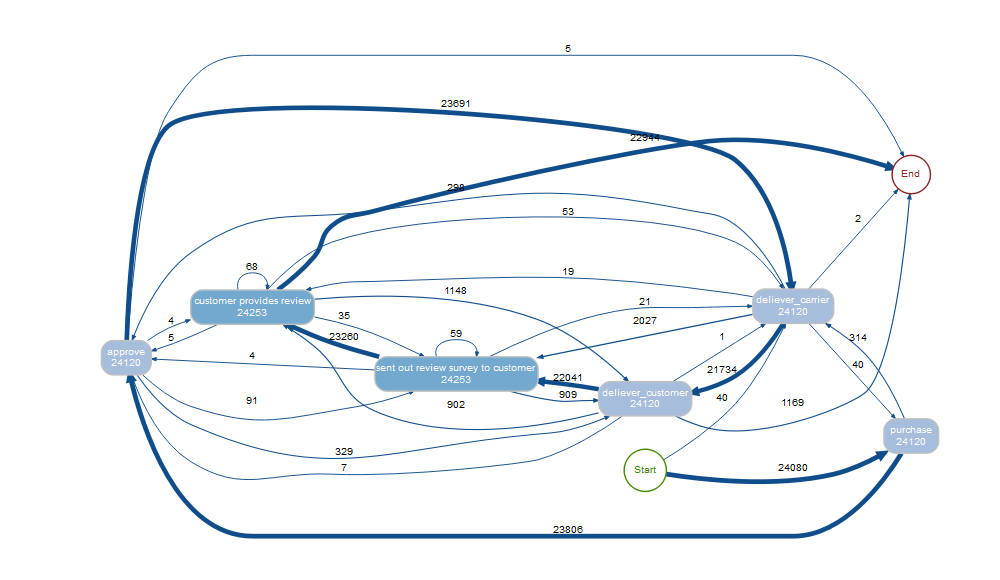
Solution: “purchase,approve,deliever\_carrier,deliever\_customer,sent out review survey to customer,customer provides review”

### c. Visualization of the process

#install.packages("processmapR")  
library(processmapR)  
  
trace\_explorer(event\_log)



process\_map(event\_log)



### d. How many cases, absolute and relative follow the most frequent process variant

Most frequent process variant: “purchase,approve,deliever\_carrier,deliever\_customer,sent out review survey to customer,customer provides review”

max(TraceVariantFreq$absolute\_frequency)

## [1] 21583

### e. How long are the traces

trace\_length(event\_log)

## min q1 median mean q3 max st\_dev   
## 6.0000000 6.0000000 6.0000000 6.0110282 6.0000000 10.0000000 0.1492225   
## iqr   
## 0.0000000

The traces have a range between 6 and 10 activities. The mean and the median are at 6 activities.

### f. What are the typical start and end avtivities

start\_activities(event\_log, level="activity")

## # A tibble: 2 x 4  
## activity absolute relative cum\_sum  
## <fct> <int> <dbl> <dbl>  
## 1 purchase 24080 0.998 0.998  
## 2 deliever\_carrier 40 0.00166 1

end\_activities(event\_log, level="activity")

## # A tibble: 4 x 4  
## activity absolute relative cum\_sum  
## <fct> <int> <dbl> <dbl>  
## 1 customer provides review 22944 0.951 0.951  
## 2 deliever\_customer 1169 0.0485 1.000  
## 3 approve 5 0.000207 1.000  
## 4 deliever\_carrier 2 0.0000829 1

The typical start activity is “purchase”. The typical end activity is “customer provides review”.

### g. What are the deviations from the most frequent variant of the process

TraceVariantFreq$trace [7]

## [1] "purchase,approve,deliever\_carrier,sent out review survey to customer,customer provides review,deliever\_customer"

Looking at the TraceVariantFreq Subset, we can see that the second most common process variant is “purchase,approve,deliever\_carrier,sent out review survey to customer,customer provides review,deliever\_customer”.

In this process order, the customer provides his feedback before he actually receives the order. All other steps are the same as in the most frequent process variant. So the difference to the most frequent process variant is, that the feedback does not consider the received product.

## 3. Performance and Time Perspective

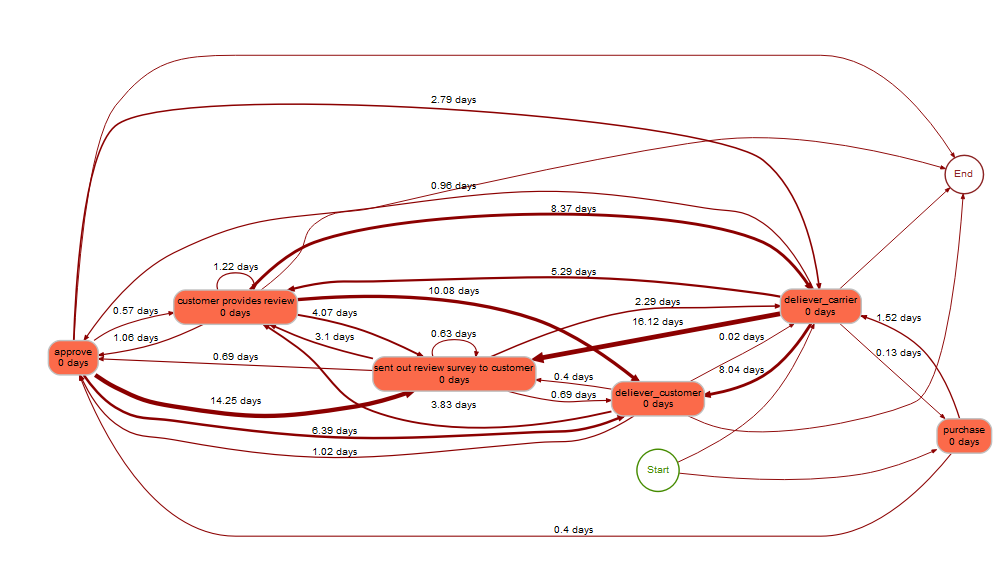
### a. What are the minimum, maximum and mean throughput time

throughput\_time(event\_log)

## min q1 median mean q3 max st\_dev   
## 2.107350 9.160880 13.216545 15.946221 19.160576 522.969039 6.000000   
## iqr <NA>   
## NA 9.999696   
## attr(,"units")  
## [1] "days"

### b. Performance Map

process\_map(event\_log, (performance(mean, "days")))

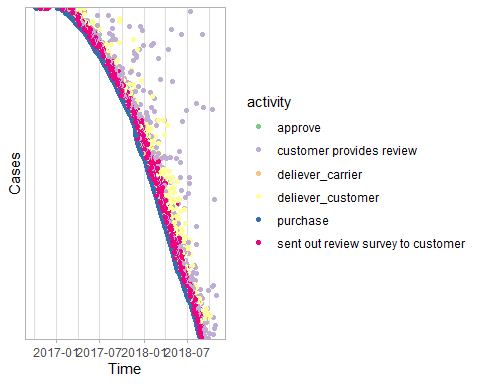


Process Map Interpretaion: The process maps shows the amount of days that it takes from one process step to the next. In the most frequent process variant, the start activity is “purchase”. It takes 0.4 days to the next step “approve”, then 2.79 days for “deliver carrier”, then 8.04 days to “deliver customer”, then 0.4 days to “send out review survey to customer” and finally 3.1 days until the “customer provides feedback”.

If Olist wants to save time and deliver orders faster, time could be saved when going from “approve” to “deliver carrier” and from “deliver carrier” to “deliver customer”. Especially the last mentioned takes quite long.

### c. Dotted Chart

dotted\_chart(event\_log, x= "absolute", y= "start")



Interpretation of the dotted chart:

The dotted chart shows the cases (orders from different customers) dependent of the time. The dots represent different activities. In this chart all processes start at the same time (but it is the process time not the calendar time!). The vertical lines show activities that happened at the same time. From the dotted chart it can be derrived why some processed take longer than others.

In this dotted chart we can see that the delivery of the customer and the feedback provision by the customer take quite long at times. It may shorten the order process if the delivery can be fastened and the customers could be encouraged to provide their feedback faster.

## 4. Satisfaction Analysis: Investigation of drivers of customer satisfaction

### a. Regression Analysis with review score as DV

#regression analysis with avg\_desc\_length,avg\_ph\_quality, customer\_state, nmb\_sellers and voucher as IVs.  
mod1 <- lm(review\_score ~ avg\_desc\_length + avg\_ph\_quality + customer\_state + nmb\_sellers + voucher + total\_pay\_value, data = event\_log)  
  
summary(mod1)

##   
## Call:  
## lm(formula = review\_score ~ avg\_desc\_length + avg\_ph\_quality +   
## customer\_state + nmb\_sellers + voucher + total\_pay\_value,   
## data = event\_log)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.4486 -0.2664 0.7241 0.8035 3.1925   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.581e+00 1.167e-01 47.829 < 2e-16 \*\*\*  
## avg\_desc\_length 4.837e-05 5.301e-06 9.126 < 2e-16 \*\*\*  
## avg\_ph\_quality 7.071e-03 1.967e-03 3.595 0.000325 \*\*\*  
## customer\_stateAL -6.376e-01 1.262e-01 -5.051 4.40e-07 \*\*\*  
## customer\_stateAM -2.751e-01 1.398e-01 -1.967 0.049138 \*   
## customer\_stateAP -1.444e-01 1.754e-01 -0.823 0.410481   
## customer\_stateBA -4.767e-01 1.147e-01 -4.157 3.23e-05 \*\*\*  
## customer\_stateCE -4.764e-01 1.170e-01 -4.071 4.68e-05 \*\*\*  
## customer\_stateDF -2.830e-01 1.158e-01 -2.444 0.014532 \*   
## customer\_stateES -2.962e-01 1.156e-01 -2.561 0.010427 \*   
## customer\_stateGO -3.705e-01 1.156e-01 -3.204 0.001354 \*\*   
## customer\_stateMA -5.629e-01 1.194e-01 -4.714 2.43e-06 \*\*\*  
## customer\_stateMG -2.340e-01 1.136e-01 -2.060 0.039448 \*   
## customer\_stateMS -2.813e-01 1.196e-01 -2.351 0.018705 \*   
## customer\_stateMT -1.975e-01 1.181e-01 -1.672 0.094482 .   
## customer\_statePA -4.781e-01 1.179e-01 -4.056 5.00e-05 \*\*\*  
## customer\_statePB -3.599e-01 1.222e-01 -2.946 0.003217 \*\*   
## customer\_statePE -3.412e-01 1.163e-01 -2.935 0.003338 \*\*   
## customer\_statePI -3.365e-01 1.224e-01 -2.750 0.005967 \*\*   
## customer\_statePR -1.860e-01 1.142e-01 -1.629 0.103401   
## customer\_stateRJ -4.347e-01 1.136e-01 -3.826 0.000131 \*\*\*  
## customer\_stateRN -1.036e-01 1.233e-01 -0.840 0.400837   
## customer\_stateRO -5.553e-02 1.302e-01 -0.426 0.669816   
## customer\_stateRR 1.111e-01 1.994e-01 0.557 0.577262   
## customer\_stateRS -2.395e-01 1.141e-01 -2.099 0.035783 \*   
## customer\_stateSC -2.946e-01 1.146e-01 -2.572 0.010116 \*   
## customer\_stateSE -5.867e-01 1.256e-01 -4.671 3.00e-06 \*\*\*  
## customer\_stateSP -1.749e-01 1.133e-01 -1.544 0.122711   
## customer\_stateTO -3.079e-02 1.302e-01 -0.236 0.813117   
## nmb\_sellers -1.169e+00 2.764e-02 -42.279 < 2e-16 \*\*\*  
## voucher1 1.745e-02 1.799e-02 0.970 0.332024   
## total\_pay\_value -2.115e-04 1.696e-05 -12.471 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.271 on 142790 degrees of freedom  
## (2164 observations deleted due to missingness)  
## Multiple R-squared: 0.02213, Adjusted R-squared: 0.02192   
## F-statistic: 104.2 on 31 and 142790 DF, p-value: < 2.2e-16

The average product descriptions, average photo quality, and number of different sellers have an highly significant influence on the review score. Whether the customer used a voucher or not has no impact. The place of residence a customer lives in sometimes has a significant influence, depending on the federal state he lives in.

The regression however does not explain a large amount of the realisation of the review score as the R² value is low.

There is a mistake present in our current regression model since one order takes on average six rows in the event\_log. The transformation in the wide-format would have solved this problem, but following error message appears:

#install.packages("dplyr")  
library(dplyr)  
library(tidyr)  
  
#Try to change event\_log from long-format to wide-format. Error when entering code:  
#event\_log\_spread <- spread(event\_log %>% dplyr::select(order\_id, activity, timestamp, review\_score),   
# activity, timestamp, drop = FALSE)  
  
#Fehler: kann Vektor der Größe 86.7 GB nicht allozieren

### b. Add a variable indicating if an order followed the most common process variant

### c. Add a variable containing the throughput time

The throughput time is considered to be the time between start activity and end activity.

-> Event Log includes timestamp for activties

#Create data matrix consisting of   
throughput <- durations(event\_log, units = "days")  
  
#Merge with event\_log  
event\_log <- merge(event\_log, throughput, by.x = "order\_id", by.y = "order\_id")  
  
#Regression model  
mod2 <- lm(review\_score ~ avg\_desc\_length + avg\_ph\_quality + customer\_state + nmb\_sellers + voucher + total\_pay\_value + duration\_in\_days, data = event\_log)  
  
summary(mod2)

##   
## Call:  
## lm(formula = review\_score ~ avg\_desc\_length + avg\_ph\_quality +   
## customer\_state + nmb\_sellers + voucher + total\_pay\_value +   
## duration\_in\_days, data = event\_log)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.5507 -0.3279 0.6375 0.7852 10.1415   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.086e+00 1.147e-01 53.054 < 2e-16 \*\*\*  
## avg\_desc\_length 4.512e-05 5.202e-06 8.673 < 2e-16 \*\*\*  
## avg\_ph\_quality 2.605e-03 1.931e-03 1.349 0.177438   
## customer\_stateAL -6.046e-01 1.239e-01 -4.880 1.06e-06 \*\*\*  
## customer\_stateAM -1.616e-01 1.372e-01 -1.178 0.238925   
## customer\_stateAP 1.295e-01 1.722e-01 0.752 0.451901   
## customer\_stateBA -5.110e-01 1.126e-01 -4.540 5.62e-06 \*\*\*  
## customer\_stateCE -4.693e-01 1.148e-01 -4.088 4.36e-05 \*\*\*  
## customer\_stateDF -4.200e-01 1.137e-01 -3.695 0.000220 \*\*\*  
## customer\_stateES -4.127e-01 1.135e-01 -3.637 0.000276 \*\*\*  
## customer\_stateGO -4.734e-01 1.135e-01 -4.172 3.02e-05 \*\*\*  
## customer\_stateMA -5.358e-01 1.172e-01 -4.571 4.86e-06 \*\*\*  
## customer\_stateMG -3.920e-01 1.115e-01 -3.515 0.000440 \*\*\*  
## customer\_stateMS -3.639e-01 1.174e-01 -3.100 0.001934 \*\*   
## customer\_stateMT -2.593e-01 1.159e-01 -2.237 0.025303 \*   
## customer\_statePA -4.453e-01 1.157e-01 -3.850 0.000118 \*\*\*  
## customer\_statePB -3.855e-01 1.199e-01 -3.216 0.001301 \*\*   
## customer\_statePE -3.862e-01 1.141e-01 -3.385 0.000711 \*\*\*  
## customer\_statePI -3.799e-01 1.201e-01 -3.164 0.001557 \*\*   
## customer\_statePR -3.507e-01 1.121e-01 -3.128 0.001760 \*\*   
## customer\_stateRJ -5.442e-01 1.115e-01 -4.880 1.06e-06 \*\*\*  
## customer\_stateRN -1.287e-01 1.210e-01 -1.064 0.287480   
## customer\_stateRO -3.391e-02 1.278e-01 -0.265 0.790745   
## customer\_stateRR 1.233e-01 1.957e-01 0.630 0.528686   
## customer\_stateRS -3.360e-01 1.120e-01 -3.000 0.002700 \*\*   
## customer\_stateSC -4.223e-01 1.124e-01 -3.756 0.000173 \*\*\*  
## customer\_stateSE -5.139e-01 1.233e-01 -4.169 3.06e-05 \*\*\*  
## customer\_stateSP -3.958e-01 1.113e-01 -3.557 0.000375 \*\*\*  
## customer\_stateTO -8.228e-02 1.278e-01 -0.644 0.519746   
## nmb\_sellers -1.220e+00 2.714e-02 -44.958 < 2e-16 \*\*\*  
## voucher1 2.886e-02 1.766e-02 1.634 0.102163   
## total\_pay\_value -1.765e-04 1.665e-05 -10.598 < 2e-16 \*\*\*  
## duration\_in\_days -1.838e-02 2.478e-04 -74.179 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.247 on 142753 degrees of freedom  
## (2200 observations deleted due to missingness)  
## Multiple R-squared: 0.05843, Adjusted R-squared: 0.05821   
## F-statistic: 276.8 on 32 and 142753 DF, p-value: < 2.2e-16

Duration\_in\_days have an highly significant influence on the review\_score.

### d. Additional variables based on the process analysis

None