Data Quality Asessement of events data on ORBIT

By Hongxing Niu

2019-10-04

Preface

After you have acquired the data, you should do the Data Quality Assessement (DQA) before to move forward to Advanced Analytics.

The structure of this kernel is as follows

- **Preface**
- **Load data**
- **Data quality assessment using metrics of six dimensions: DQA1-DQA6**
- **Automated reporting**

Purpose and structure of this kernel

- **Give a systematic assessment of the quality of the raw data on ORBIT about events and Events** link
  - **1.Validity/Semantic Accurary** *Check if data conforms to the syntax (format, type, range) of its definition*
  - **2.Cardinality/uniqueness/duplicates** *Nothing will be recorded more than once based upon how that thing is identified*
  - **3.Completeness**
    a. **3.1. Overall completeness** *The overall proportion of stored data against the potential of "100% complete"*
    b. **3.2. Time series completeness** *The proportion of stored data against the potential of "100% complete" in the course of time*
  - **4.Accuracy**
    a. **4.1. Distribution of raw data** *Distribution analysis entails counting all the records associated with each value and dividing these by the total number of records to see what percentage of the data is associated with any specific value and how the percentages compare to each other*
    b. **4.2. Outlier detection** *An outlier is a value that lies in the tail of the statistical distribution of a set of data values(usually +-2 SD)*
  - **5.Consistency** *The absence of difference, when comparing two or more representations of a thing against a definition*
  - **6.Timeliness/availability** *The degree to which data represent reality from the required point in time*
- **Performs a data diagnosis and automatically generates a DQA report.**

This document introduces **Data Quality Assessment** methods. You will learn how to diagnose the quality of `tbl_df` data that inherits from data.frame and `data.frame` with functions provided by `dlookr`.

This package is in synergy with `dplyr`. Particularly in data exploration and data wrangle, it increases the efficiency of the `tidyverse` package group.

Table 1: A knitr kable

Id	UCB_Business_Unit___c	UCB_External_Id___c	Name
a0pG000000AATC1IAP	IMM	E-00062294	2016-09-17-Bonn-Sponsoring-Patienten-
a0p4A00000A0wciQAB	IMM	E-00106476	2018-04-25 Strategisches Beraterboard I
a0p4A00000EJaNIQA1	IMM	E-00122393	2018-11-27-Essen-Dermaxchange
a0p4A00000A0x9WQAR	IMM	E-00106624	2018-03-26München Round Table Prof.
a0p4A00000A0x9vQAB	IMM	E-00106628	2018-05-30 Hamburg Immunology for yo
a0pG000000BQxm3IAD	IMM	E-00075967	2017-10-07 Hamburg Weltrheumatag Pa
a0p4A000009z7rxQAA	IMM	E-00093724	2017-09-21-EssenProjektbesprechung
a0p4A000009zmEQQAY	IMM	E-00099660	2018-02-16-Würzburg-Sponsoring
a0p4A00000CldOHQAZ	IMM	E-00112519	2018-08-30-Halle-Sponsoring-Sommersch
a0pG000000ARWQnIAP	IMM	E-00065933	2016-11-03 Demmin Expertengespräch

Supported data structures

Data diagnosis supports the following data structures.

- data frame : data.frame class.
- data table : tbl_df class.
- **table of DBMS** : table of the DBMS through tbl_dbi.
- **Using dplyr backend for any DBI-compatible database.**

Load libraries

0. Load Data

```
[1] "Dimension of data:"
[1] 1632   24
```

0.Preview data

```
[1] "Preview of 'Events_nPVU_Germany' data frame"
```

1.DQA Dimension1/Rule1: Validity/Semantic Accurary

- Picking the data subset relevant to Epilepsy other than Parkinson's disease
- Cleansing up:
- Extract a substring (like "2017-09-21" from "2017-09-21T09:30:29.000+0000") both for Date and CreateDate;+Mis-format of "Call2_vod___r.CreatedDate"; +Wrongly formatted"Call2_vod___r.UCB_Owner_Full_N and "Call2_vod___r.Account_vod___r.UCB_Primary_Grand_Parent___c"

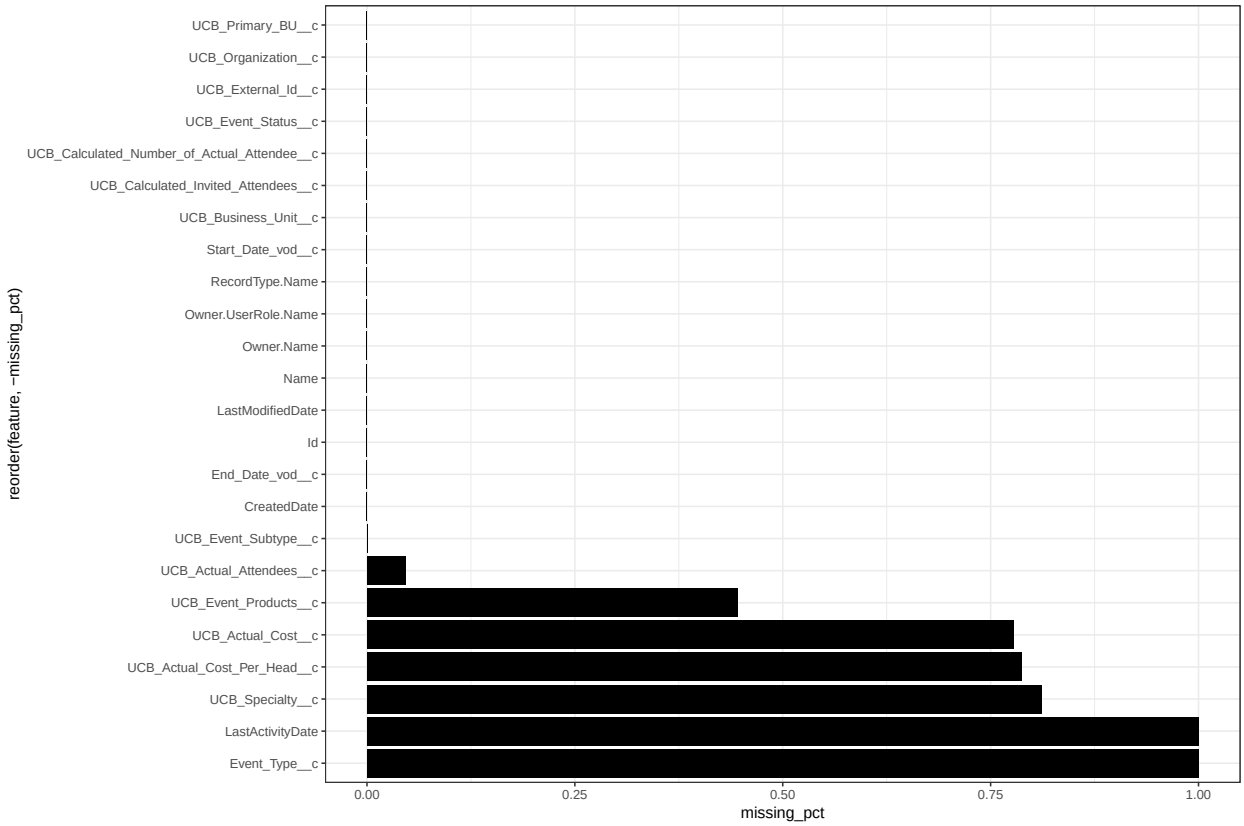
2. DQA Dimension2/Rule2: Cardinality/uniqueness/duplicates

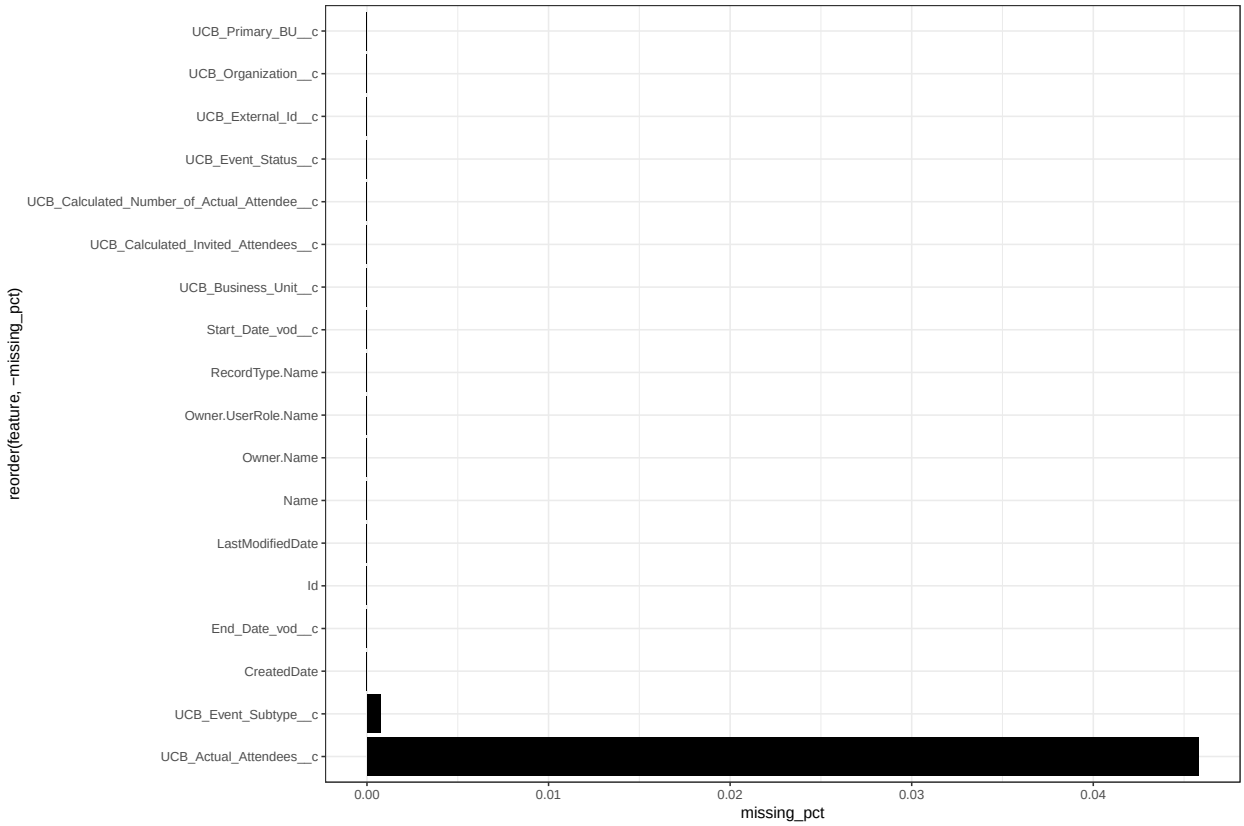
- Based on business understanding of ORBIT data, a hiddent feature "Discussion Priority" differentiates the duplicates. Many duplicates because of "Discussion Priority" behind the data (hinted by "Call2_vod___r.Detailed_Products_vod___c")

3. DQA Dimension3/Rule3: Handling missing data

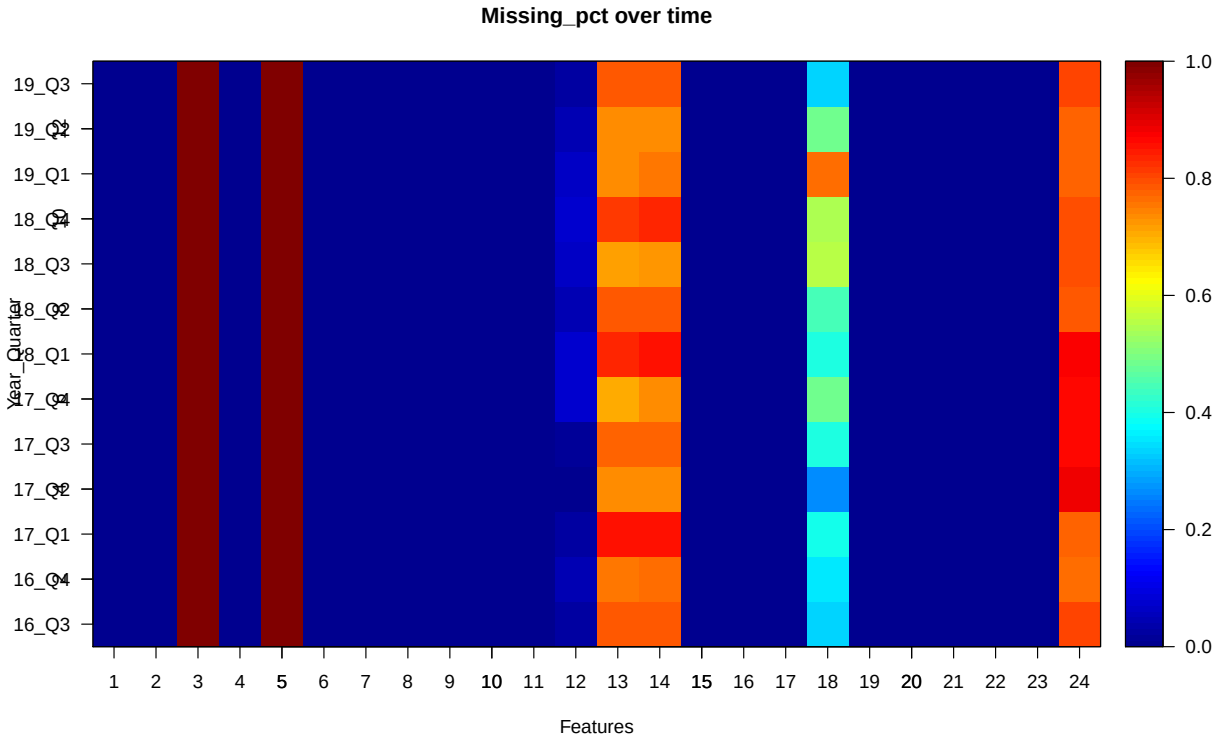
We want to get rid of any features that have >= 10% of the data missing. We're presuming there's little information in those features,and just adds noise.We plot all the features against the percent of missing values. We keep only those features with <10% missing values, and re-adjust the data frame, accordingly.

3.1 find percent missing values for each feature during all period





3.2 Handling historical missing data at at level of quarter, we use 3D plot, with time included to visualize the time series data select the columns to visualize



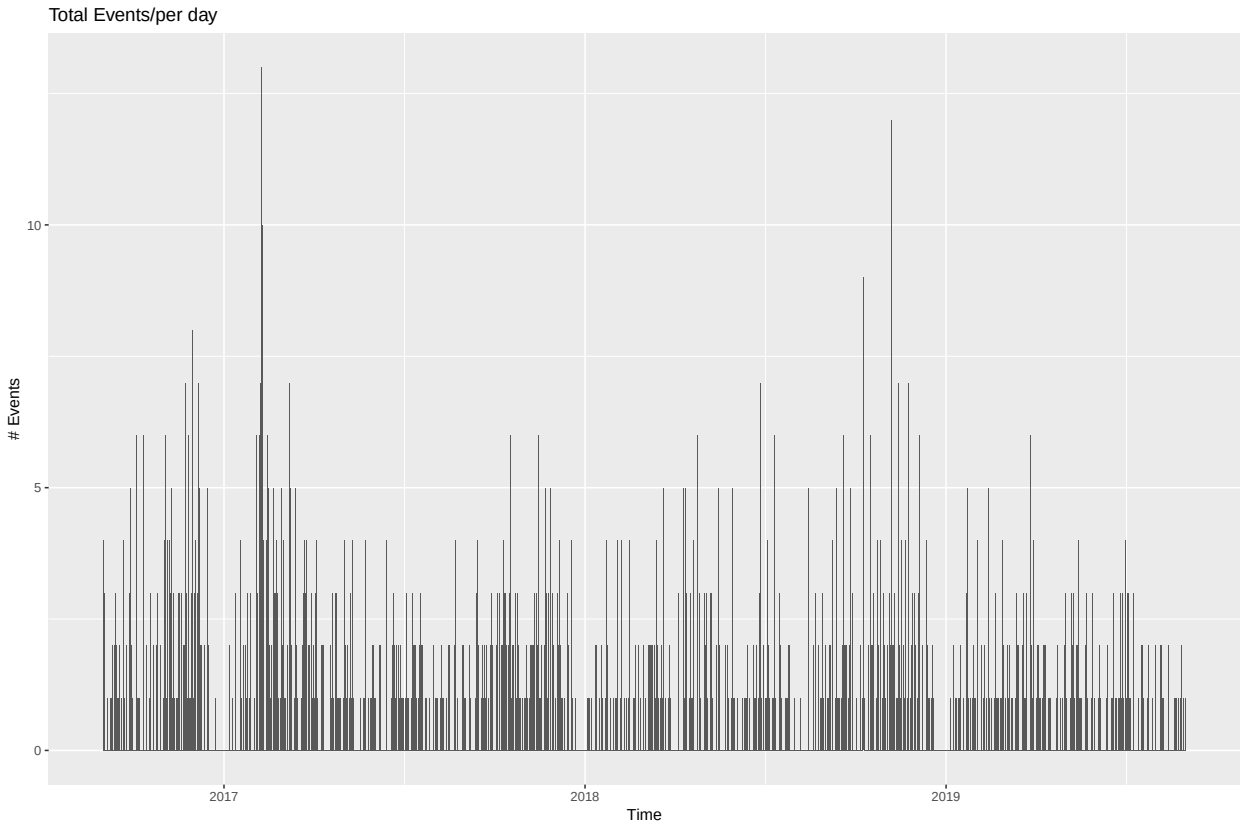

```
[1] "# Feature"
 [1] "CreatedDate"
 [2] "End_Date_vod__c"
 [3] "Event_Type__c"
 [4] "Id"
 [5] "LastActivityDate"
 [6] "LastModifiedDate"
 [7] "Name"
 [8] "Owner.Name"
 [9] "Owner.UserRole.Name"
[10] "RecordType.Name"
[11] "Start_Date_vod__c"
[12] "UCB_Actual_Attendees__c"
[13] "UCB_Actual_Cost__c"
[14] "UCB_Actual_Cost_Per_Head__c"
[15] "UCB_Business_Unit__c"
[16] "UCB_Calculated_Invited_Attendees__c"
[17] "UCB_Calculated_Number_of_Actual_Attendee__c"
[18] "UCB_Event_Products__c"
[19] "UCB_Event_Status__c"
[20] "UCB_Event_Subtype__c"
[21] "UCB_External_Id__c"
[22] "UCB_Organization__c"
[23] "UCB_Primary_BU__c"
[24] "UCB_Specialty__c"
```

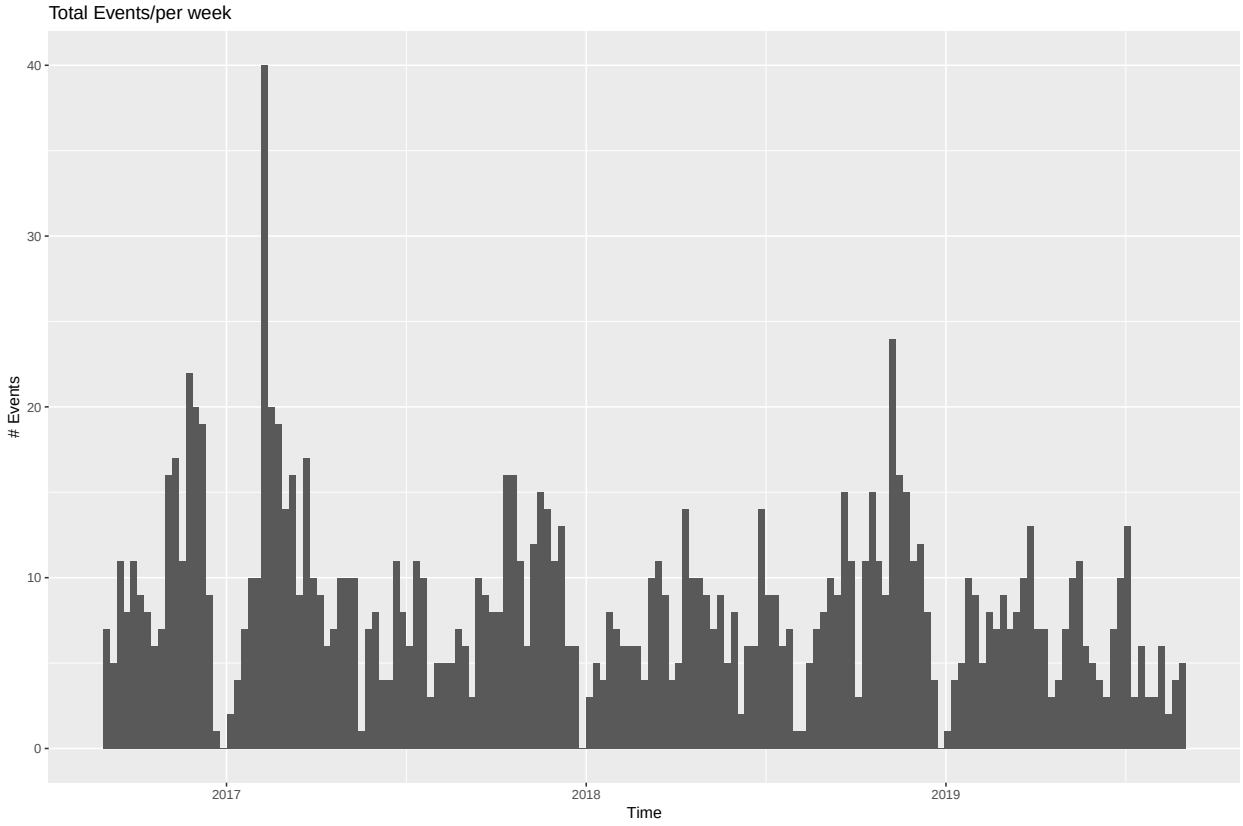
4. DQA Dimension4/Rule4: Distribution and Outlier Detection

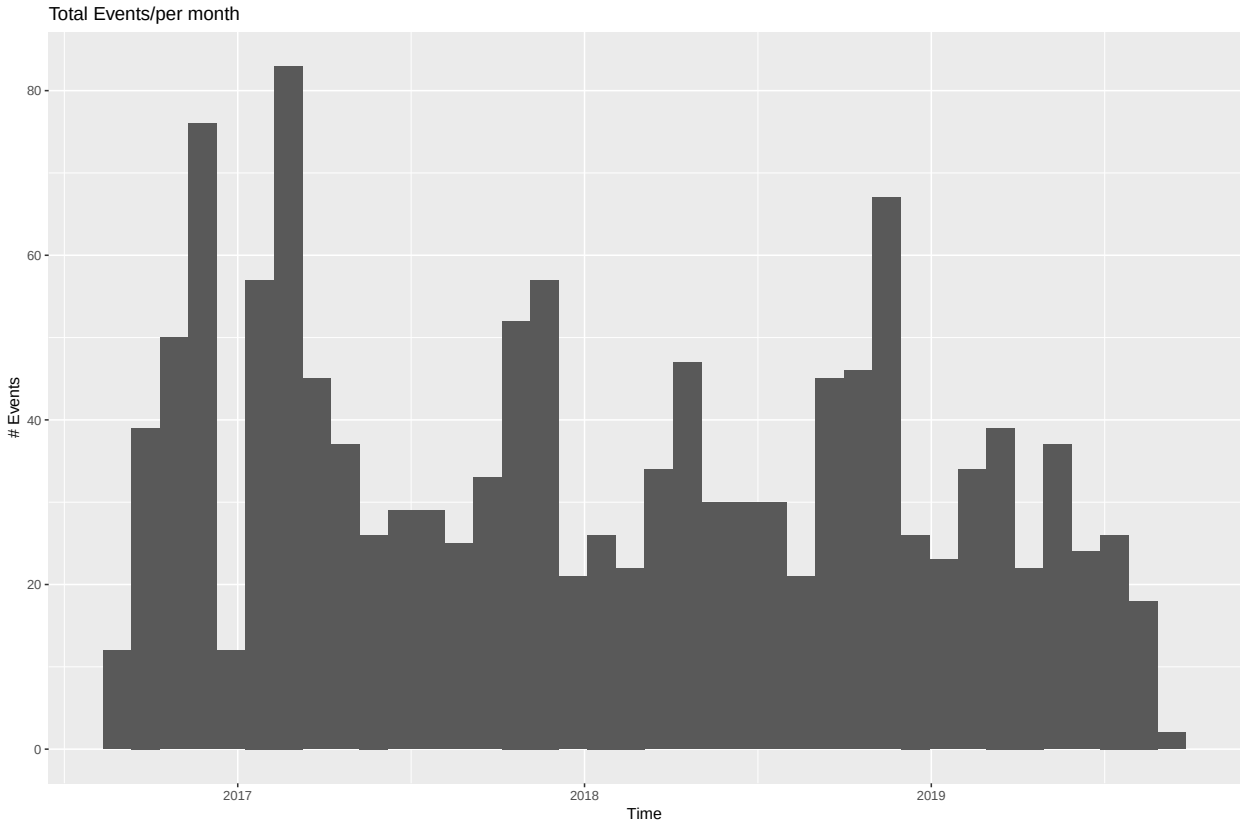
4.1 Distribution of raw data

4.1.1 Visualization and statistics

```
pdf
  2
```




Total Events/per week
40
30
20
10
0
Events
2017
2018
2019
Time


Total Events/per month
80
60
40
20
0
Events
2017
2018
2019
Time

4.1.2 Normality test on numeric variables using `normality()`

`normality()` performs a normality test on numerical data. `Shapiro-Wilk normality test` is performed. If the number of observations is larger than 5000, 5000 observations are extracted by random simple sampling and then tested.

The variables of `tbl_df` object returned by `normality()` are as follows.

- `statistic` : Statistics of the Shapiro-Wilk test
- `p_value` : p-value of the Shapiro-Wilk test
- `sample` : Number of sample observations performed Shapiro-Wilk test

We can use dplyr to sort non-normal distribution variables by p_value.:

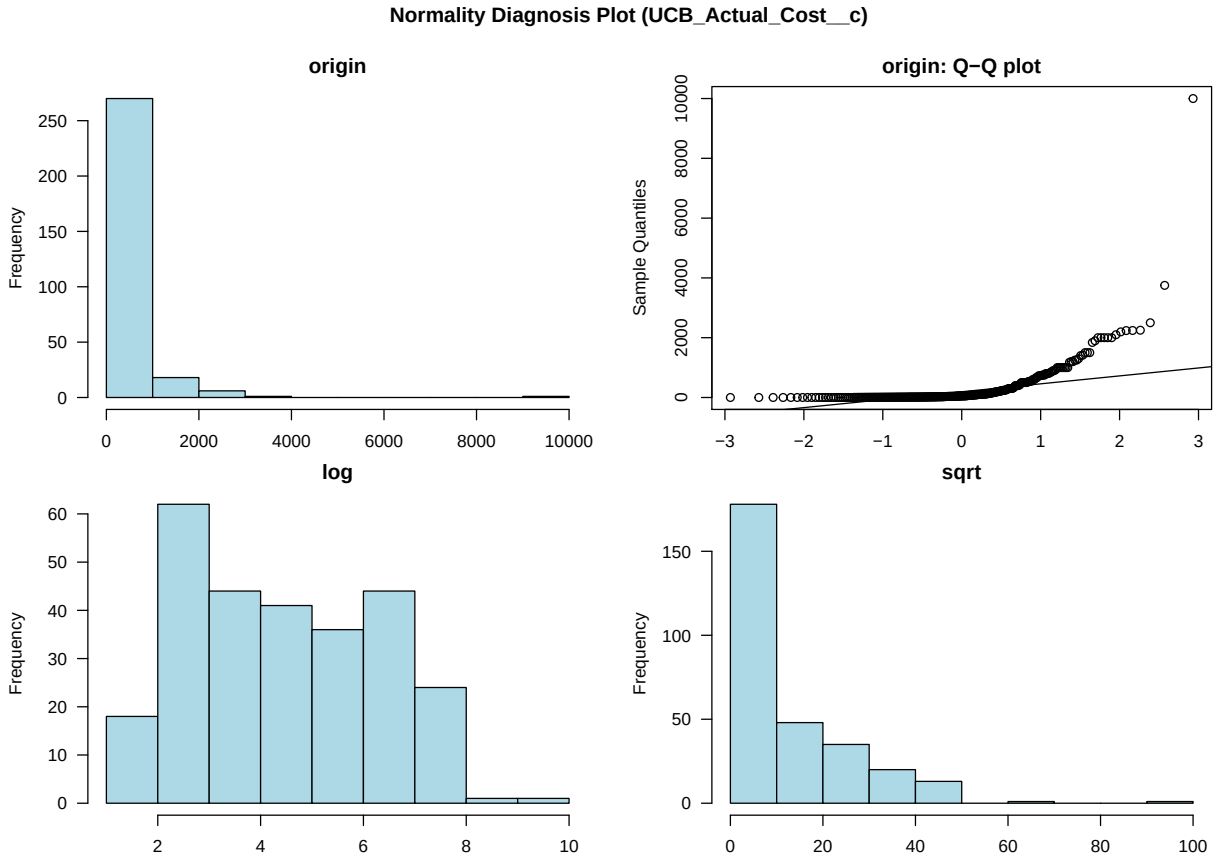
```
# A tibble: 5 x 4
  vars                                      statistic  p_value sample
  <chr>                                         <dbl>    <dbl>  <dbl>
1 UCB_Actual_Attendees__c                       0.108 1.34e-60   1332
2 UCB_Calculated_Invited_Attendees__c           0.671 6.14e-45   1332
3 UCB_Calculated_Number_of_Actual_Attendee__c   0.684 2.51e-44   1332
4 UCB_Actual_Cost_Per_Head__c                   0.197 4.97e-33   1332
# ... with 1 more row
```

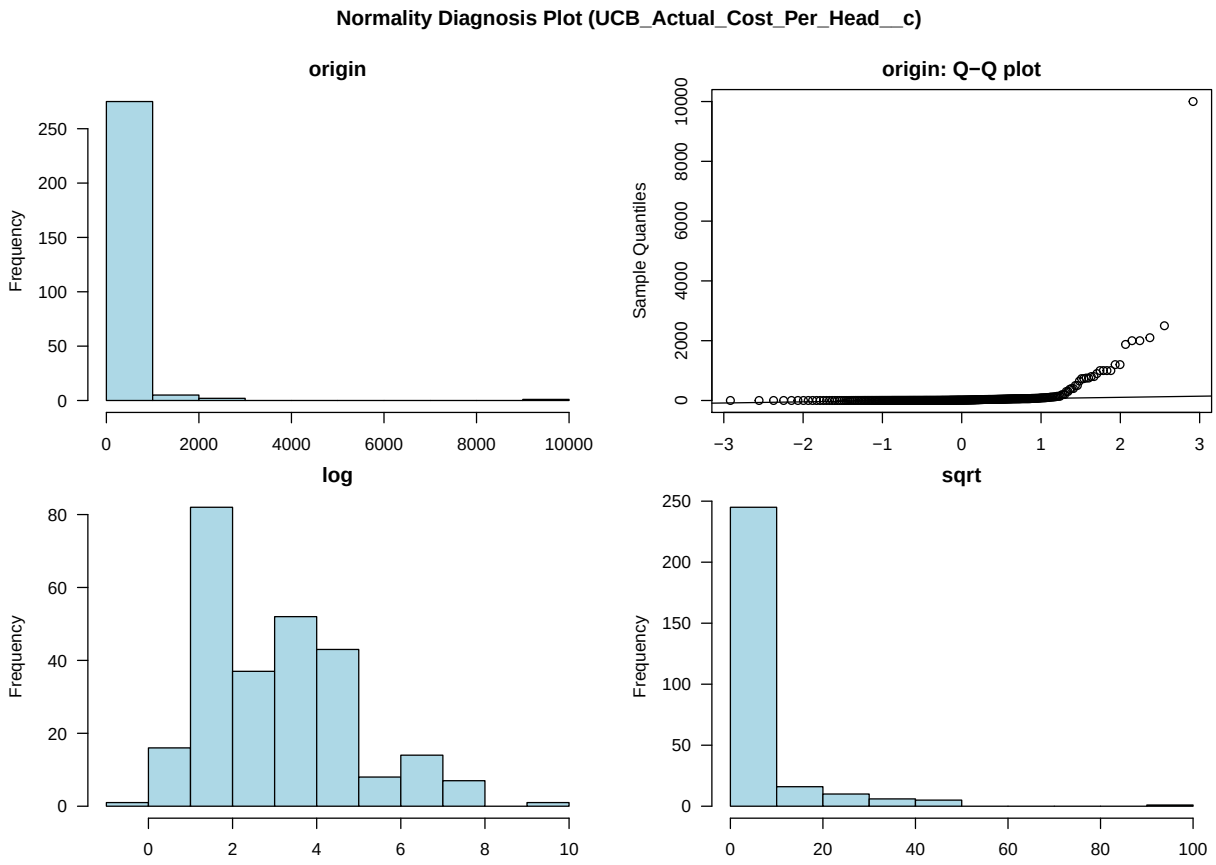
Normalization visualization of numerical variables using `plot_normality()` `plot_normality()` visualizes the normality of numeric data.

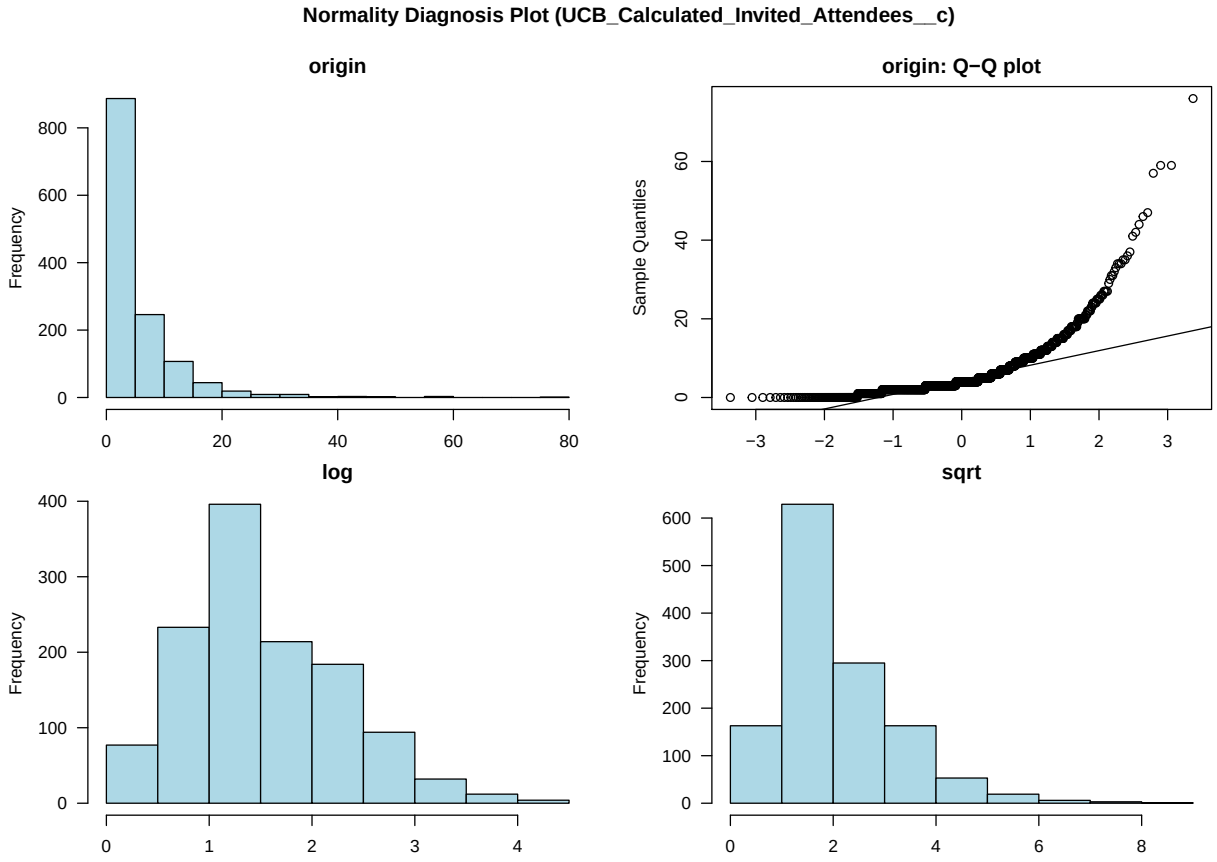
The information that `plot_normality()` visualizes is as follows.

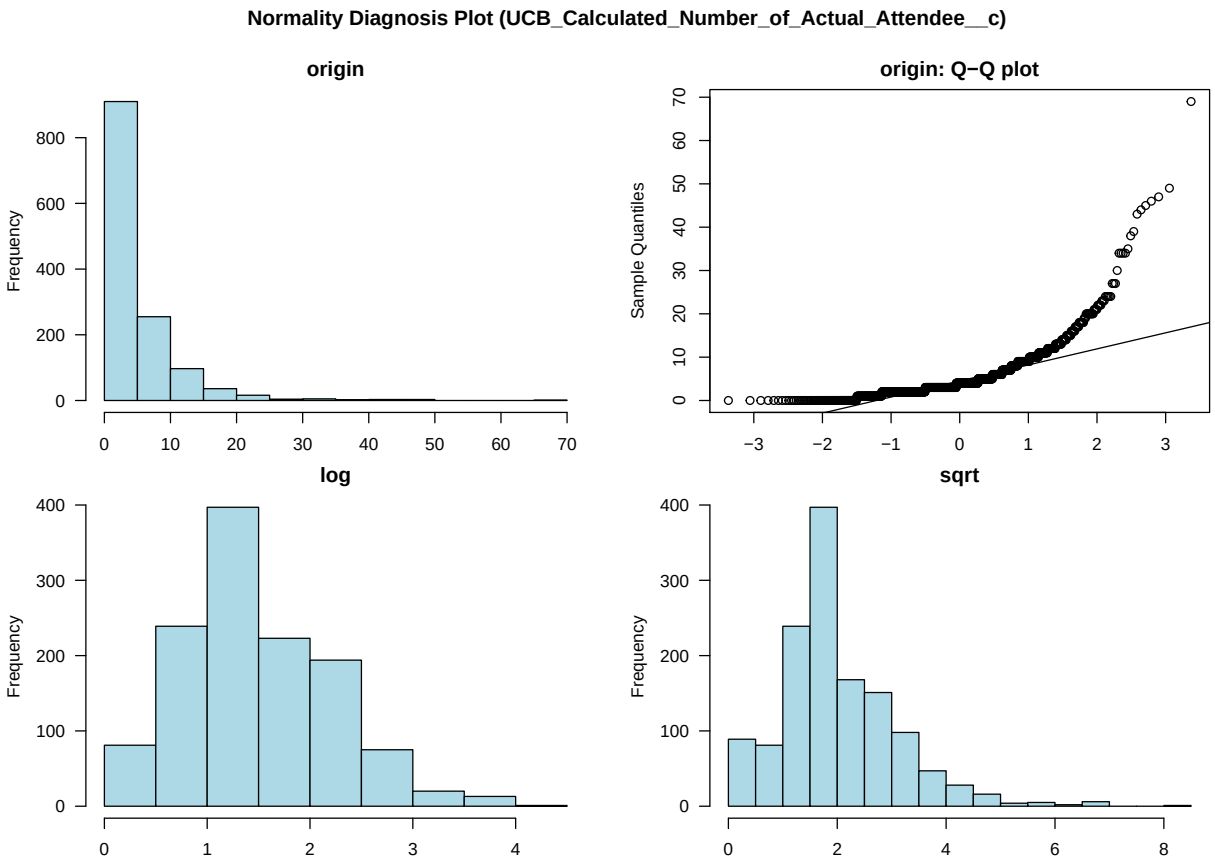
- `Histogram of original data`
- `Q-Q plot of original data`
- `histogram of log transformed data`
- `Histogram of square root transformed data`

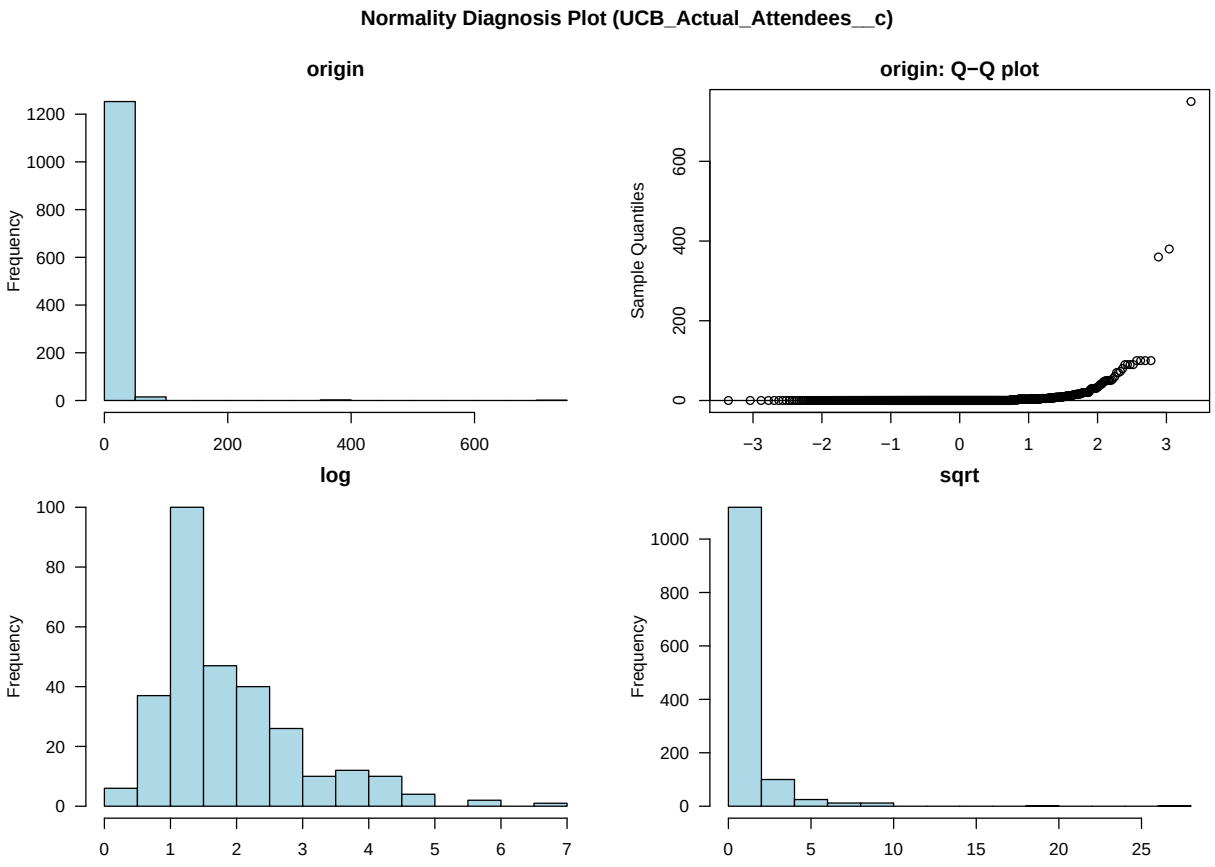
`plot_normality()` can also specify several variables like `normality()` function.


Normality Diagnosis Plot (UCB_Actual_Cost__c)
origin
Frequency
0
50
100
150
200
250
0
2000
4000
6000
8000
10000
origin: Q−Q plot
Sample Quantiles
0
2000
4000
6000
8000
10000
−3
−2
−1
0
1
2
3
log
Frequency
0
10
20
30
40
50
60
2
4
6
8
10
sqrt
Frequency
0
50
100
150
0
20
40
60
80
100


Normality Diagnosis Plot (UCB_Actual_Cost_Per_Head__c)
origin
Frequency
0
50
100
150
200
250
0
2000
4000
6000
8000
10000
origin: Q−Q plot
Sample Quantiles
0
2000
4000
6000
8000
10000
−3
−2
−1
0
1
2
3
log
Frequency
0
20
40
60
80
0
2
4
6
8
10
sqrt
Frequency
0
50
100
150
200
250
0
20
40
60
80
100


Normality Diagnosis Plot (UCB_Calculated_Invited_Attendees__c)
origin
Frequency
origin: Q−Q plot
Sample Quantiles
log
Frequency
sqrt
Frequency


Normality Diagnosis Plot (UCB_Calculated_Number_of_Actual_Attendee__c)
origin
Frequency
origin: Q−Q plot
Sample Quantiles
log
Frequency
sqrt
Frequency

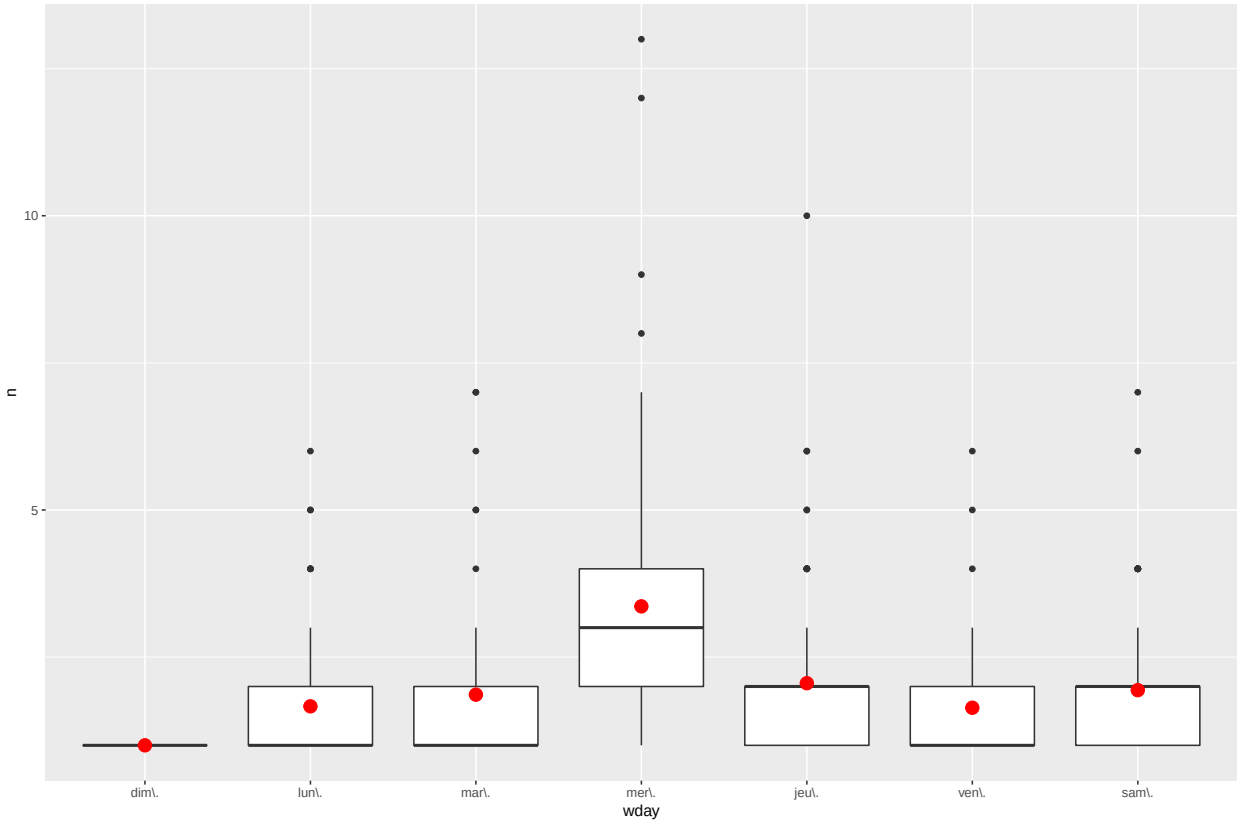


4.2 Detecting Outliers

day-of-week effect that dominates the subtler patterns we fit the model (lm) and display its predictions overlaid on the original data

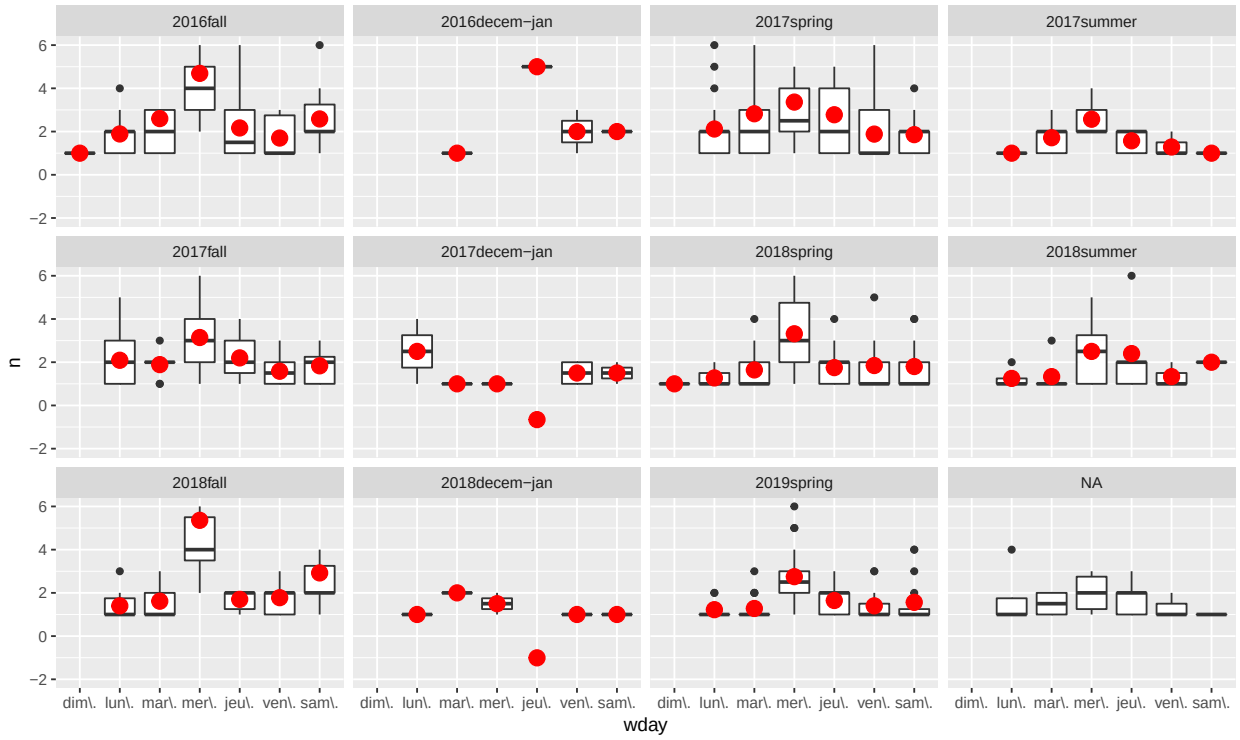
Let's figure out what data are getting wrong. First, I plotted the distribution of the data, and saw that it follows a normal distribution with mean = , std = To classify a data point as an outlier, I considered any data +/- 2 std above/below the mean is a significant error. Then, I created two data frames. One consisting of the "outliers", and another consisting of the "average". I plotted these on a graph (red = outliers; green = average), outlining the shape of the data. We observe something interesting.

```
pdf
  2
```



There is significant variation across the terms,so fitting a separate day-of-week effect for each term

```
pdf
  2
```

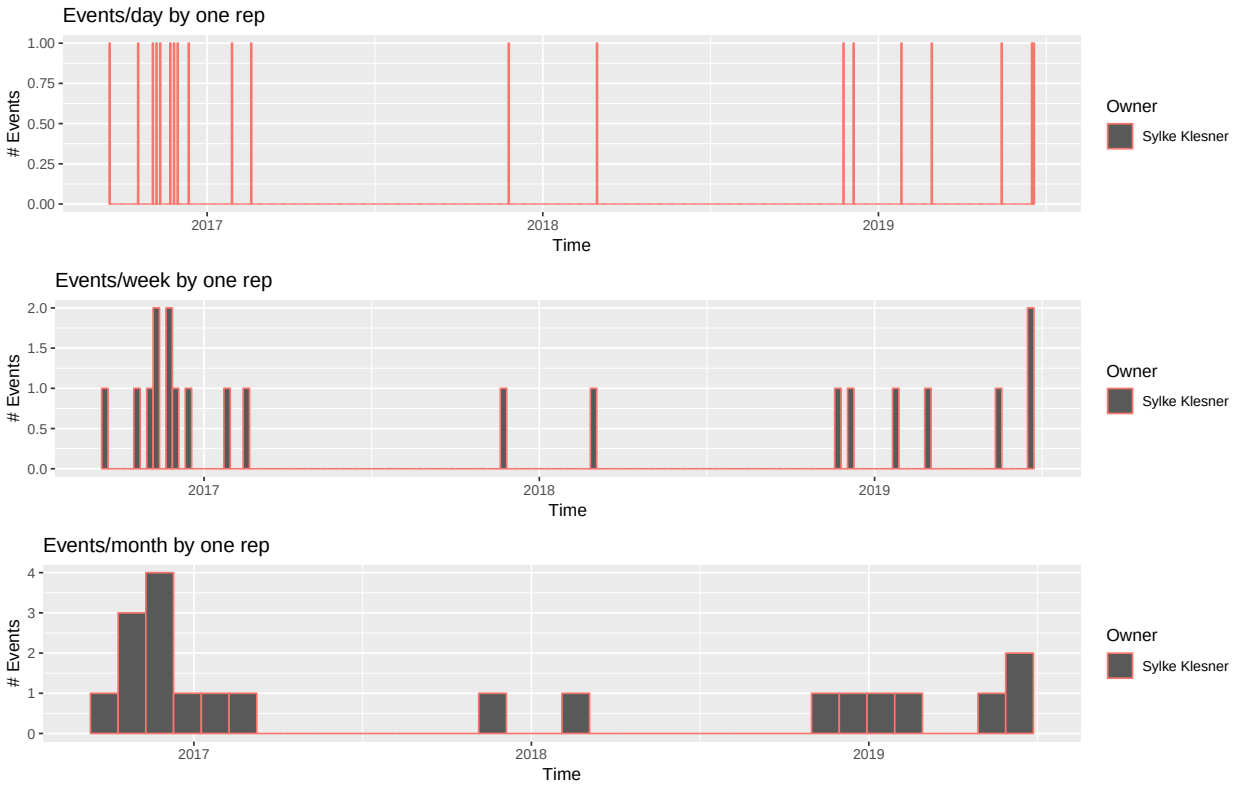


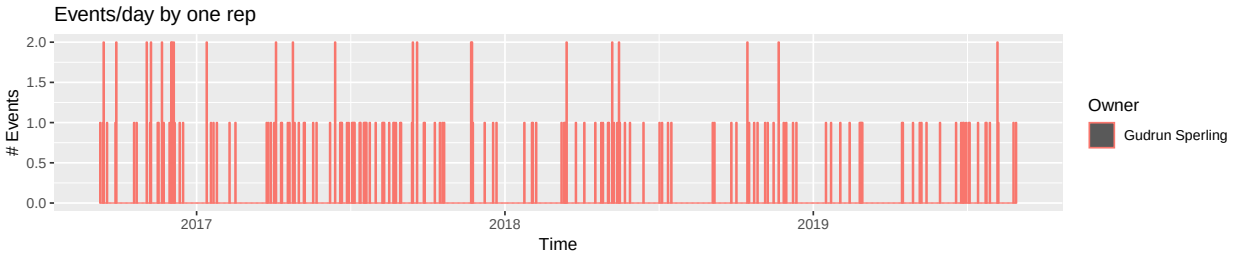
Understand how often each rep events on customers How many reps were involved in events

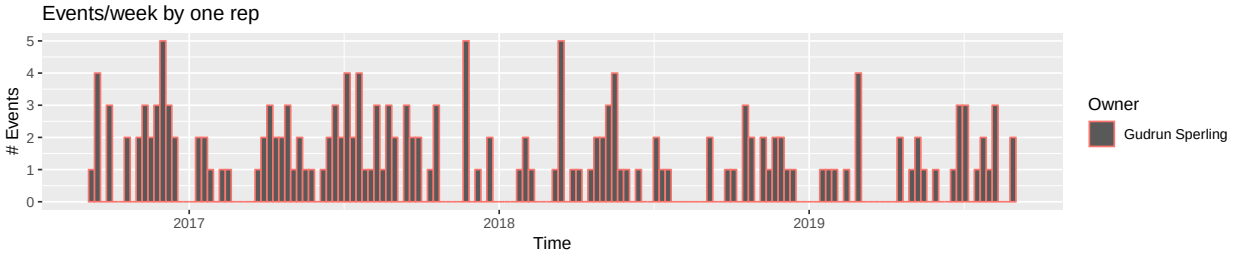
```
[1] List of reps who carried out events in 201608-201908
 [1] "Sylke Klesner"                  "Gudrun Sperling"
 [3] "Jutta Hoffmann-Nommensen"       "Astrid Hemschemeier"
 [5] "Constanze Erwin"                "Susanne Meinel-Schwebke"
 [7] "Nico Haag"                      "Lambrini Tontsidou"
 [9] "Bodo Stelle"                    "Nadine von Ryssel"
[11] "Friederike Hilger"              "Petra Opolka"
[13] "Angelika Gundlach"              "Sarah Sahlu"
[15] "Udo Lendl"                      "Stephanie Mueller-Varain"
[17] "Eike-Hans Zimmermann"           "Katrin Riedel"
[19] "Klaus Kalcum"                   "Rainer Krüger"
[21] "Balint Szilagyi"                "Petra Rösler"
[23] "Angela Hertel"                  "Lubica Vetrikova"
[25] "Ulrike Hermsdörfer-Vonalt"      "Alya Kokot"
[27] "Renate Bohnsack"                "Oliver Feil"
[29] "Jan Peetz"                      "Silvia Slazenger"
[31] "Caroline-Mascha Reifferscheid" "Sandra Mogk"
[33] "Jan Bretz"                      "Caroline Lehnart-Betz"
[35] "Heike Holländer"                "Sandra Schink"
[37] "Daniel Bura"                    "Heiko Sic"
[39] "Grit Brußig"                    "Michael Kemper"
[41] "Bianca Morgenstern"             "Viola Wernik"
[43] "Jan Helfrich"                   "Saskia Winter"
[45] "Fabian Paul"                    "Annina Getz"
[47] "Manh Dan Nguyen"                "Steffen Hiller"
[49] "Rüdiger Thees"                  "Florian Hinze"
[51] "Adrian Zander"                  "Tabea Thomas"
[53] "Wojciech Dombrowsky"            "Missanga van de Sand"
```

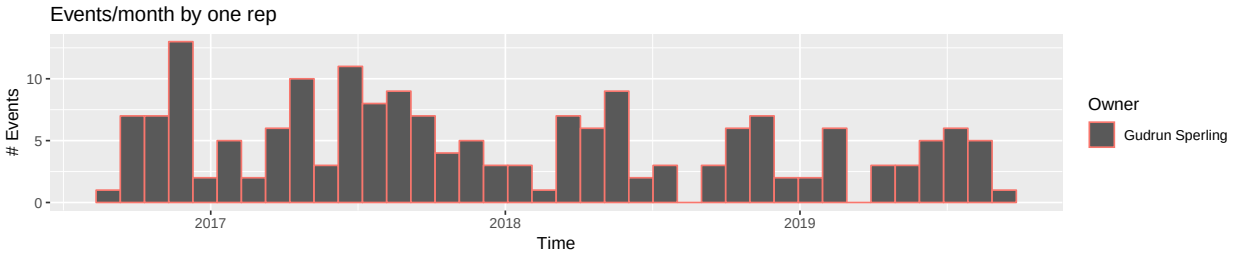
```
[55] "Joachim Hipp"
```

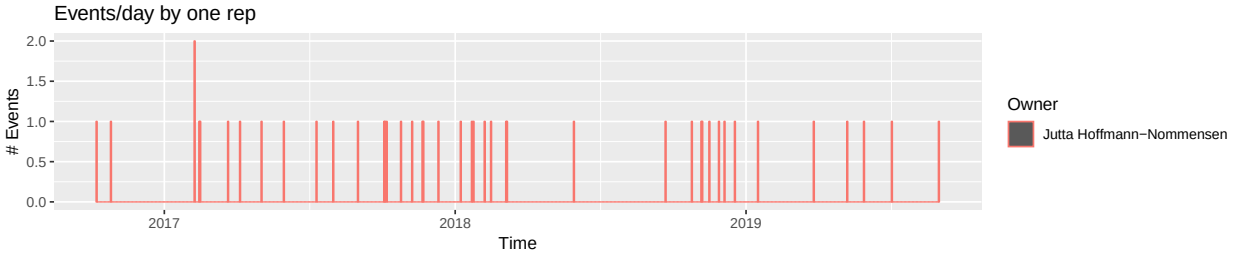
```
pdf
  2
```

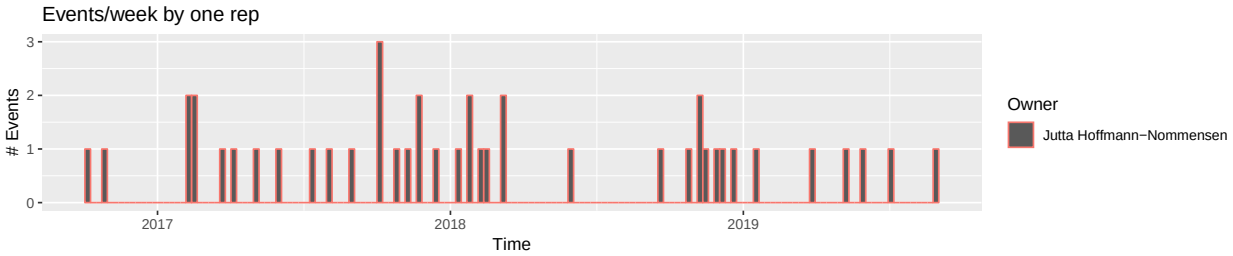


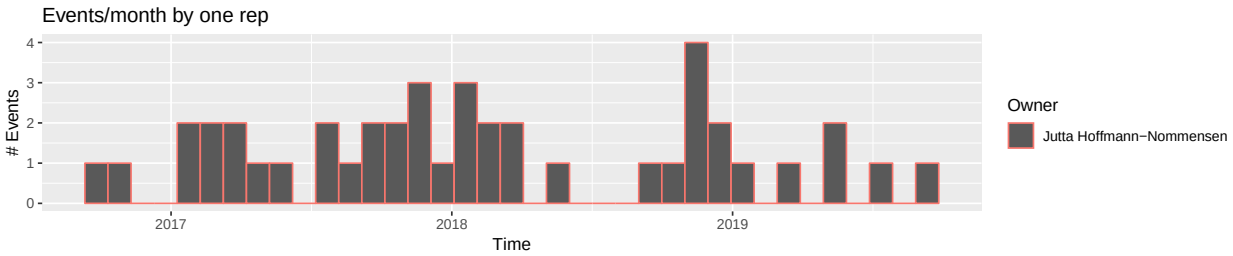

Events/day by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017
2018
2019
Time
Owner
Gudrun Sperling

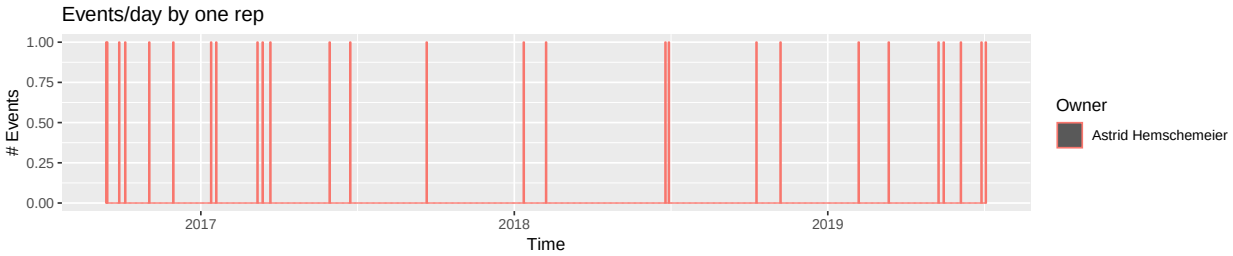

Events/week by one rep
Events
5
4
3
2
1
0
2017
2018
2019
Time
Owner
Gudrun Sperling

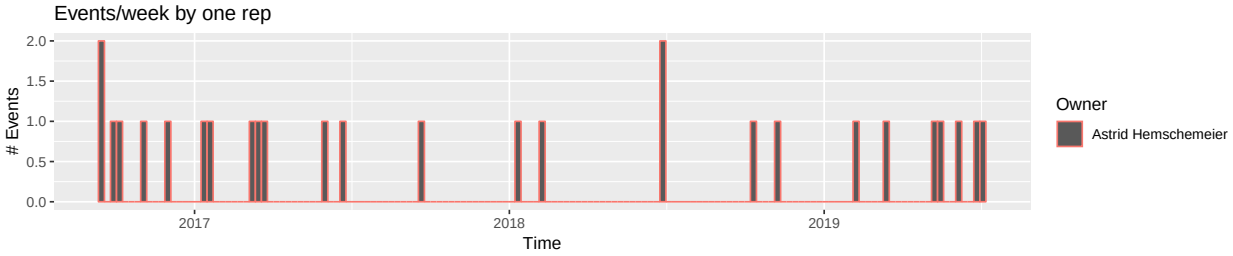

Events/month by one rep
Events
10
5
0
2017
2018
2019
Time
Owner
Gudrun Sperling

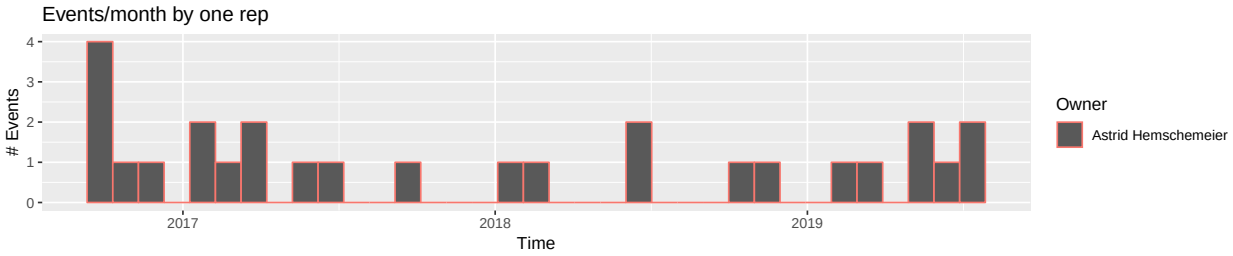

Events/day by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017
2018
2019
Time
Owner
Jutta Hoffmann−Nommensen

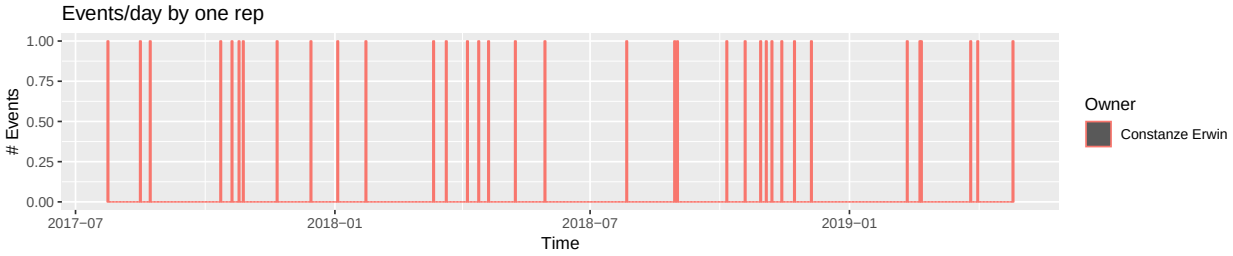

Events/week by one rep
Events
3
2
1
0
2017
2018
2019
Time
Owner
Jutta Hoffmann−Nommensen

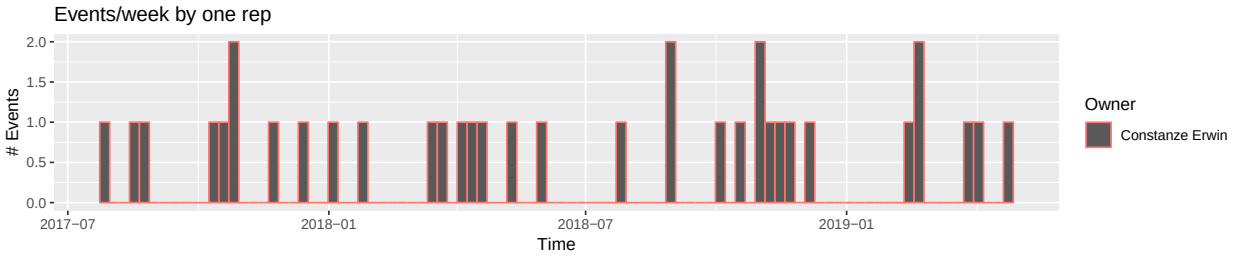

Events/month by one rep
Events
4
3
2
1
0
2017
2018
2019
Time
Owner
Jutta Hoffmann−Nommensen

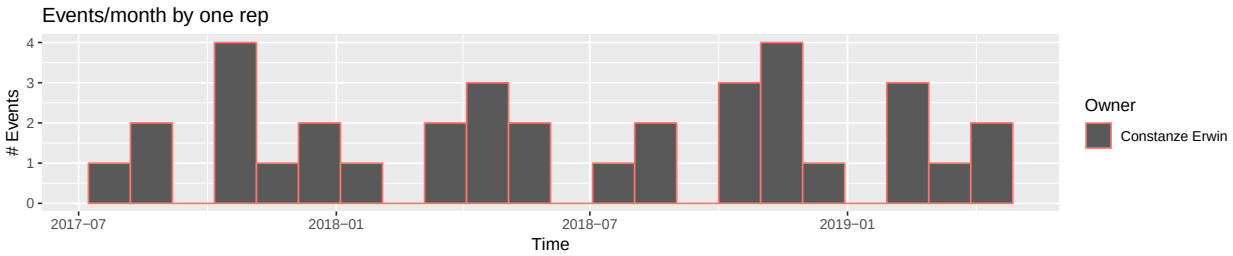

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017
2018
2019
Time
Owner
Astrid Hemschemeier

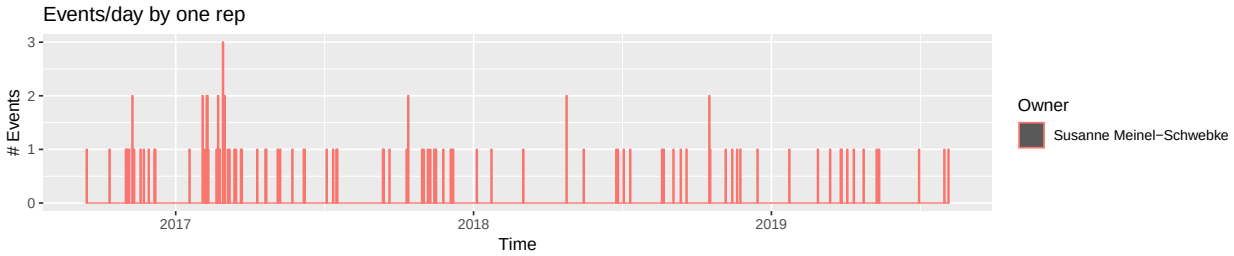

Events/week by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017
2018
2019
Time
Owner
Astrid Hemschemeier

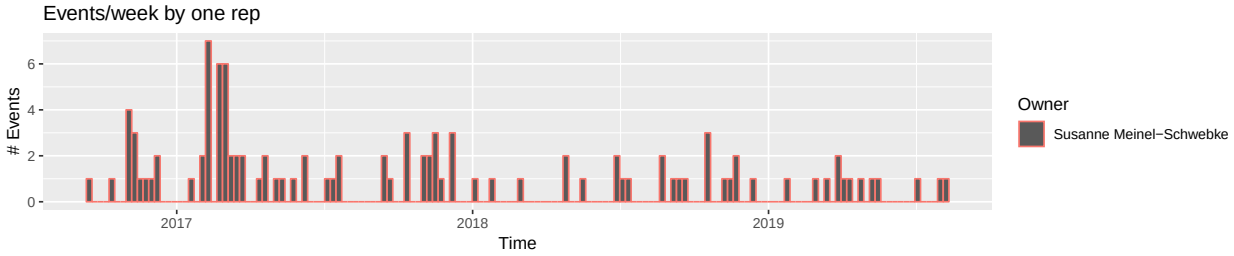

Events/month by one rep
Events
4
3
2
1
0
2017
2018
2019
Time
Owner
Astrid Hemschemeier

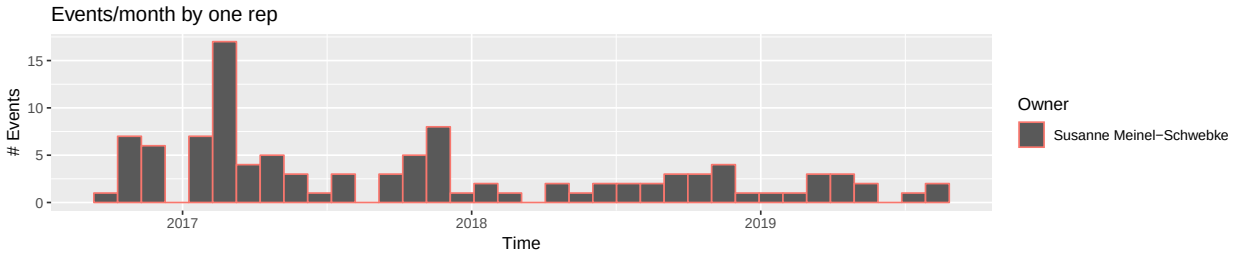

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017–07
2018–01
2018–07
2019–01
Time
Owner
Constanze Erwin

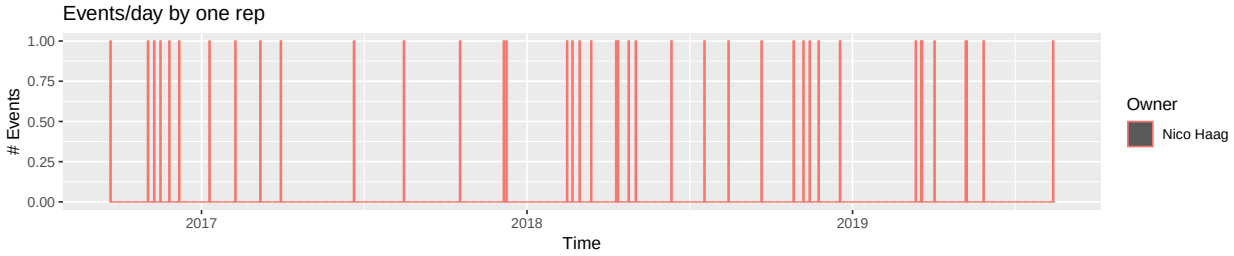

Events/week by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017–07
2018–01
2018–07
2019–01
Time
Owner
Constanze Erwin

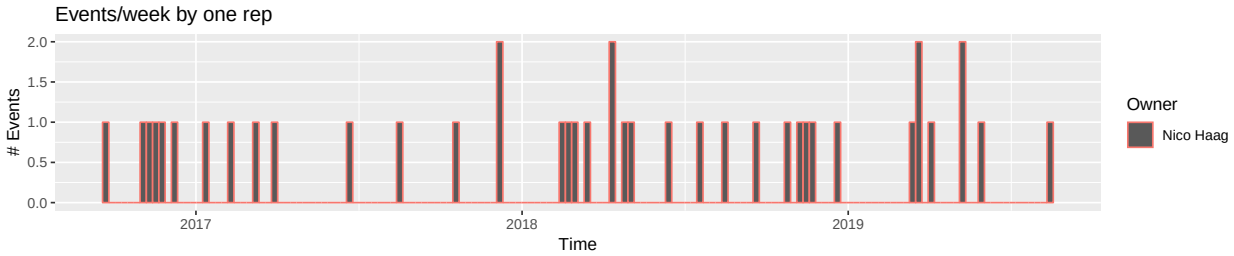

Events/month by one rep
Events
4
3
2
1
0
2017–07
2018–01
2018–07
2019–01
Time
Owner
Constanze Erwin

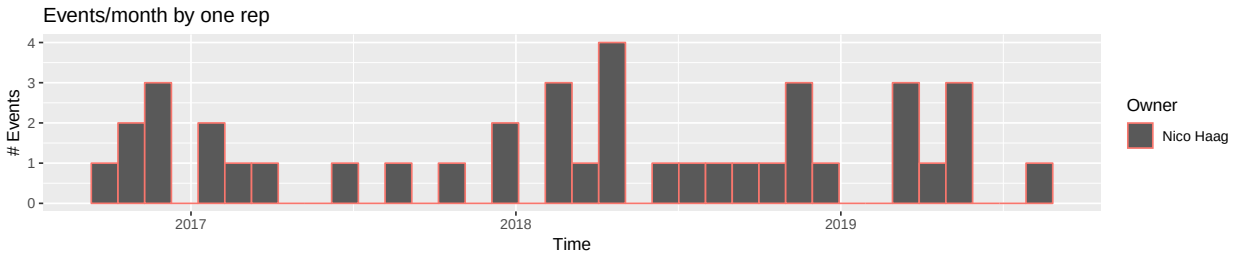

Events/day by one rep
Events
Time
2017
2018
2019
Owner
Susanne Meinel−Schwebke

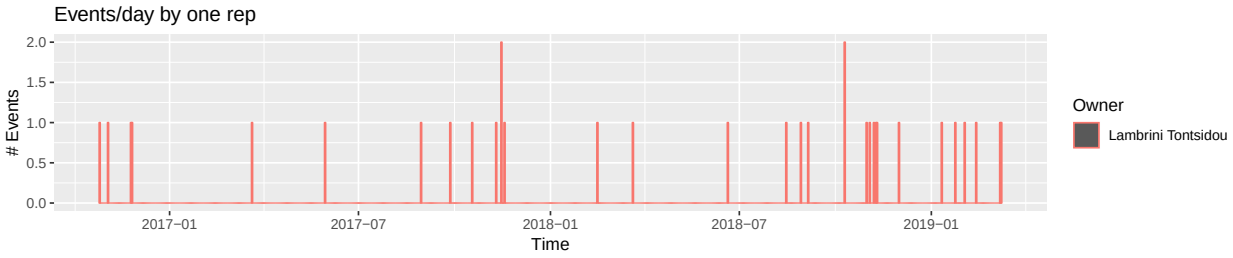

Events/week by one rep
Events
Time
2017
2018
2019
Owner
Susanne Meinel−Schwebke

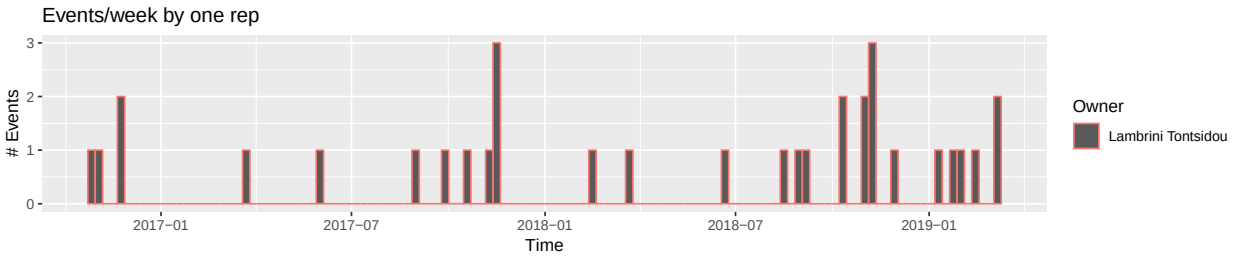

Events/month by one rep
Events
Time
2017
2018
2019
Owner
Susanne Meinel−Schwebke

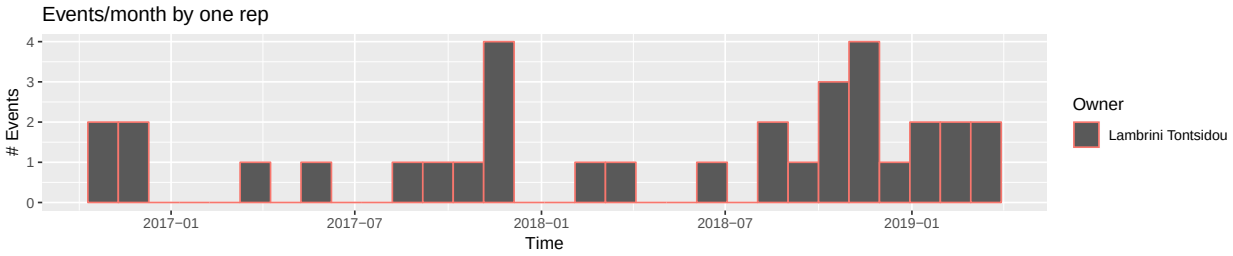

Events/day by one rep
Events
Time
2017
2018
2019
Owner
Nico Haag

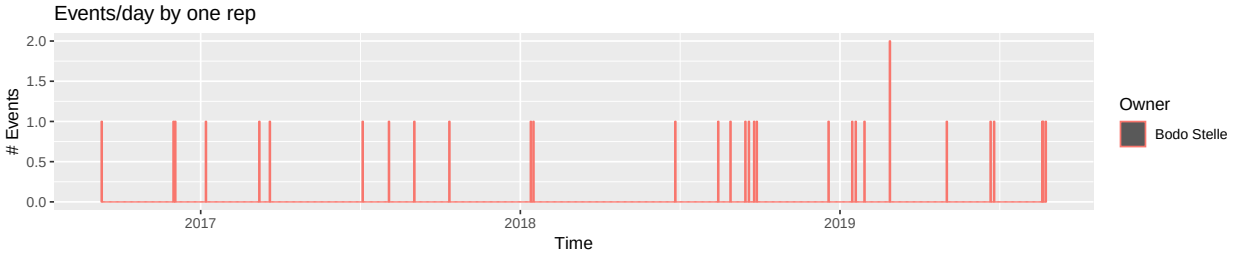

Events/week by one rep
Events
Time
2017
2018
2019
Owner
Nico Haag

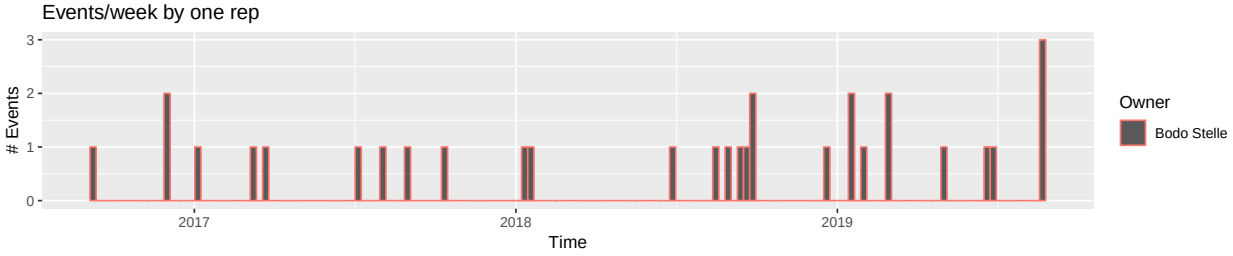

Events/month by one rep
Events
Time
2017
2018
2019
Owner
Nico Haag

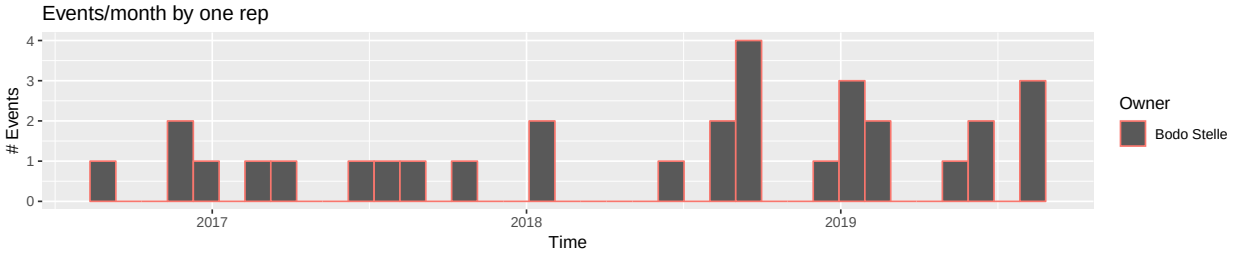

Events/day by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017–01
2017–07
2018–01
2018–07
2019–01
Time
Owner
Lambrini Tontsidou


Events/week by one rep
Events
3
2
1
0
2017–01
2017–07
2018–01
2018–07
2019–01
Time
Owner
Lambrini Tontsidou

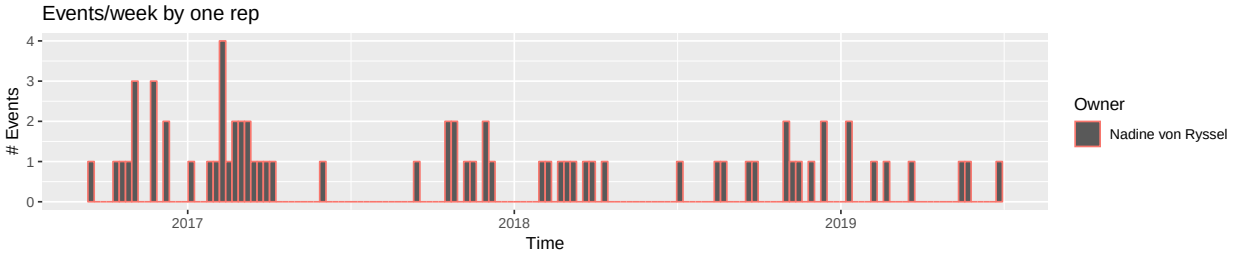

Events/month by one rep
Events
4
3
2
1
0
2017–01
2017–07
2018–01
2018–07
2019–01
Time
Owner
Lambrini Tontsidou

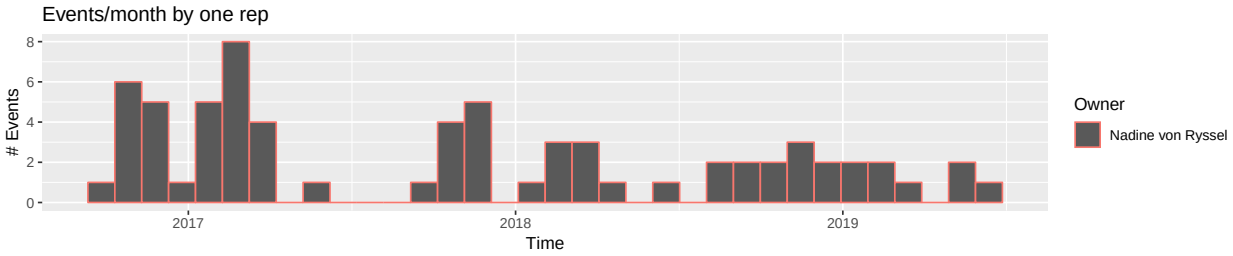

Events/day by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017
2018
2019
Time
Owner
Bodo Stelle

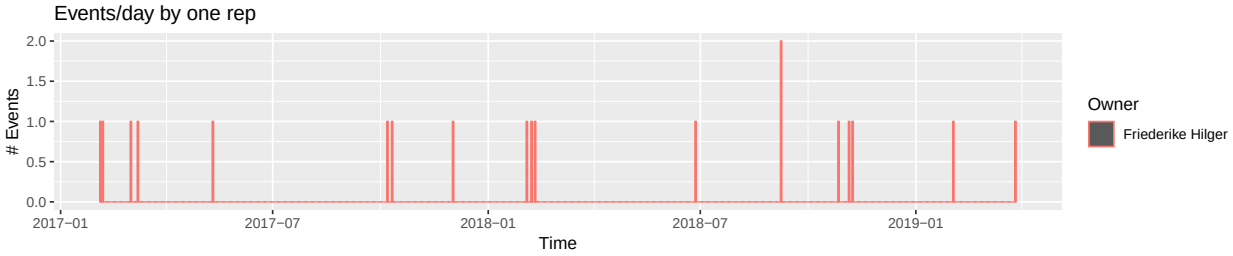

Events/week by one rep
Events
3
2
1
0
2017
2018
2019
Time
Owner
Bodo Stelle

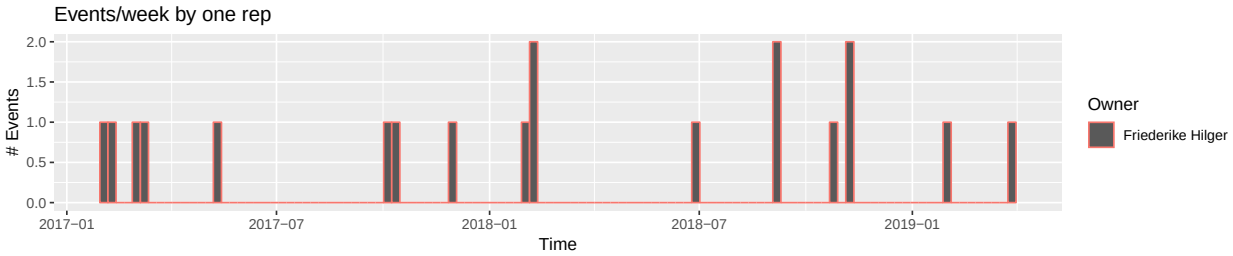

Events/month by one rep
Events
4
3
2
1
0
2017
2018
2019
Time
Owner
Bodo Stelle

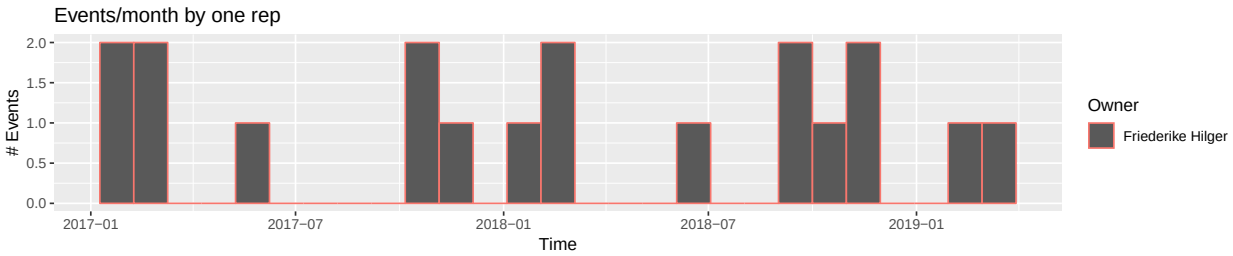

Events/day by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017
2018
2019
Time
Owner
Nadine von Ryssel

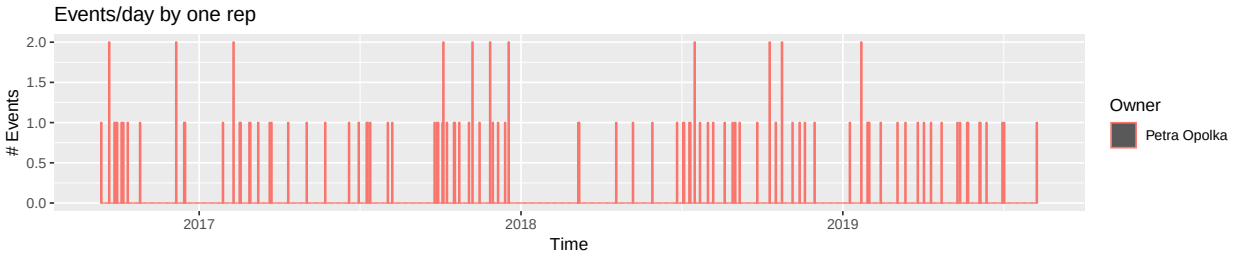

Events/week by one rep
Events
4
3
2
1
0
2017
2018
2019
Time
Owner
Nadine von Ryssel

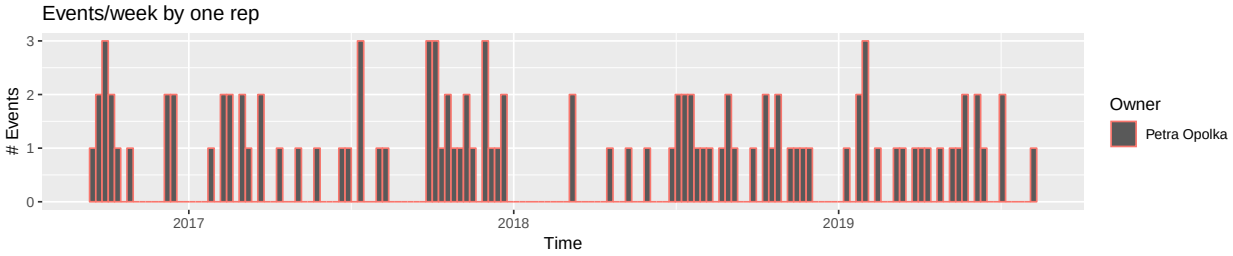

Events/month by one rep
Events
8
6
4
2
0
2017
2018
2019
Time
Owner
Nadine von Ryssel

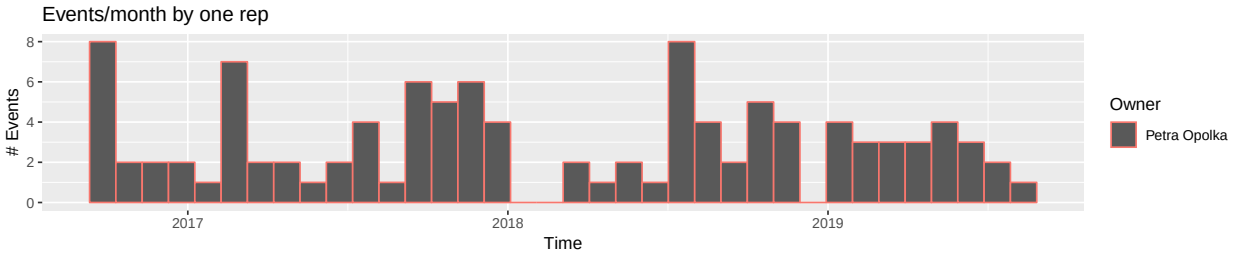

Events/day by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017–01
2017–07
2018–01
2018–07
2019–01
Time
Owner
Friederike Hilger

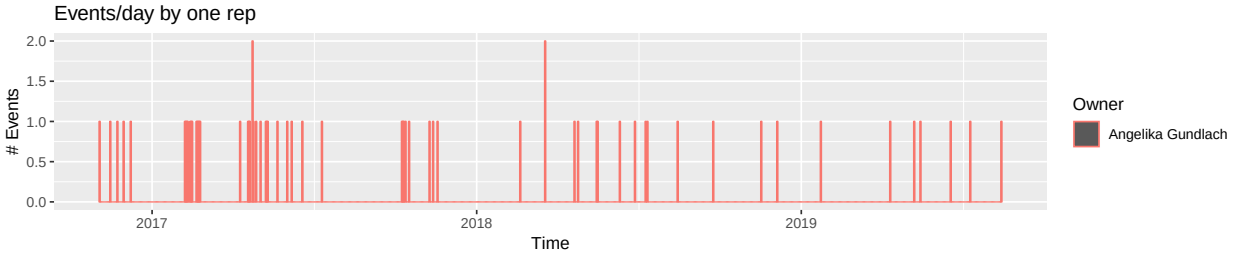

Events/week by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017–01
2017–07
2018–01
2018–07
2019–01
Time
Owner
Friederike Hilger

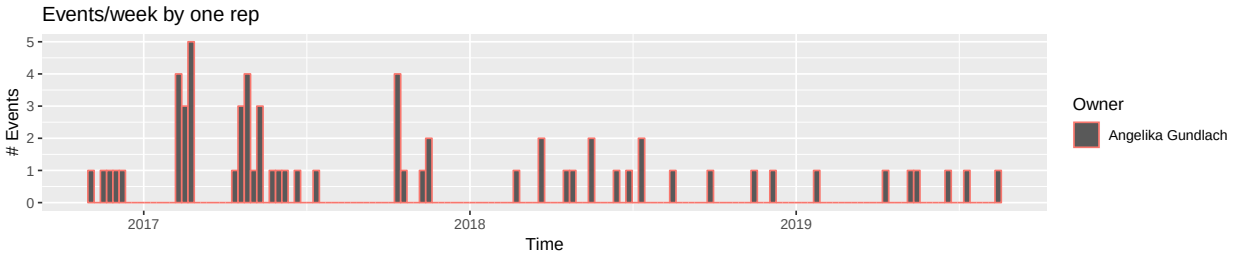

Events/month by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017–01
2017–07
2018–01
2018–07
2019–01
Time
Owner
Friederike Hilger

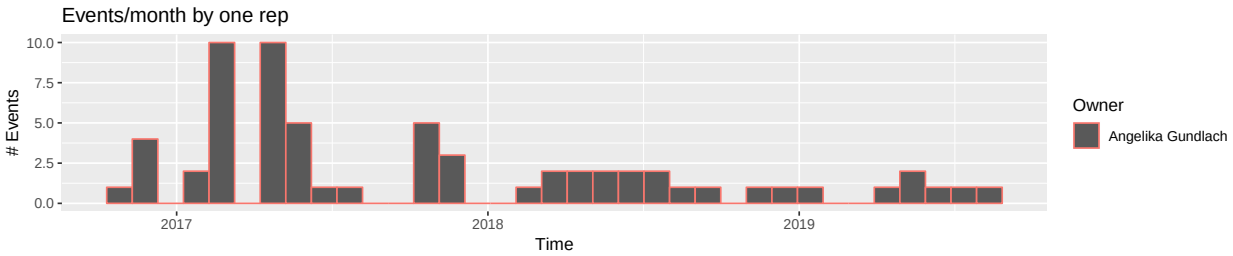

Events/day by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017
2018
2019
Time
Owner
Petra Opolka

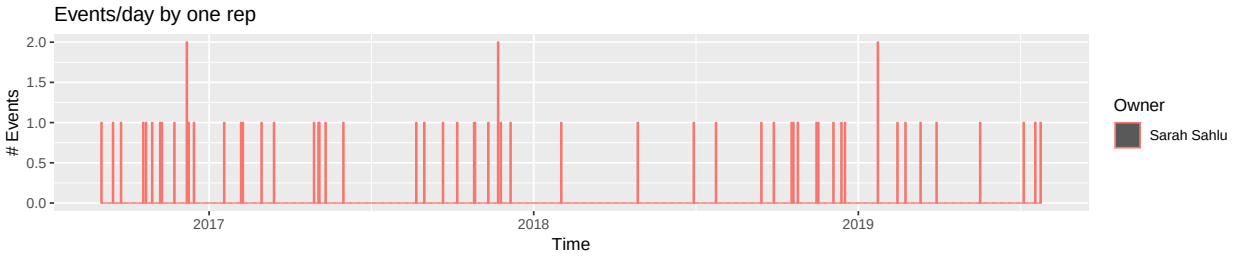

Events/week by one rep
Events
3
2
1
0
2017
2018
2019
Time
Owner
Petra Opolka

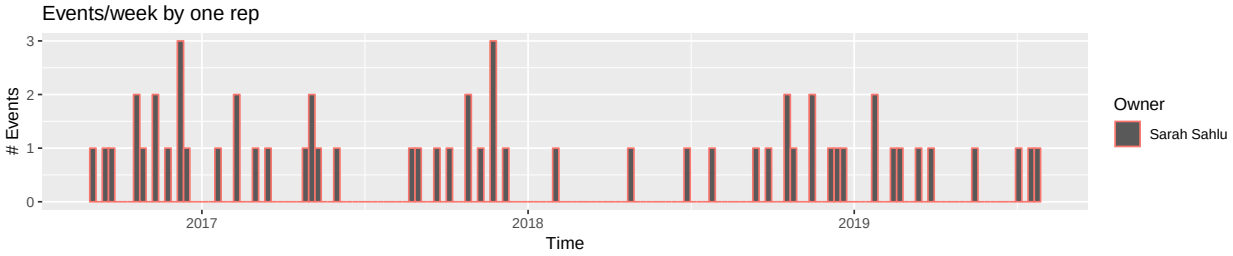

Events/month by one rep
Events
8
6
4
2
0
2017
2018
2019
Time
Owner
Petra Opolka

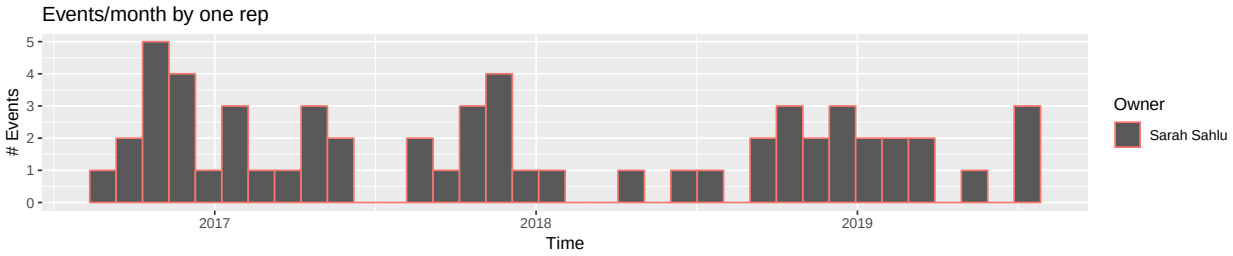

Events/day by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017
2018
2019
Time
Owner
Angelika Gundlach

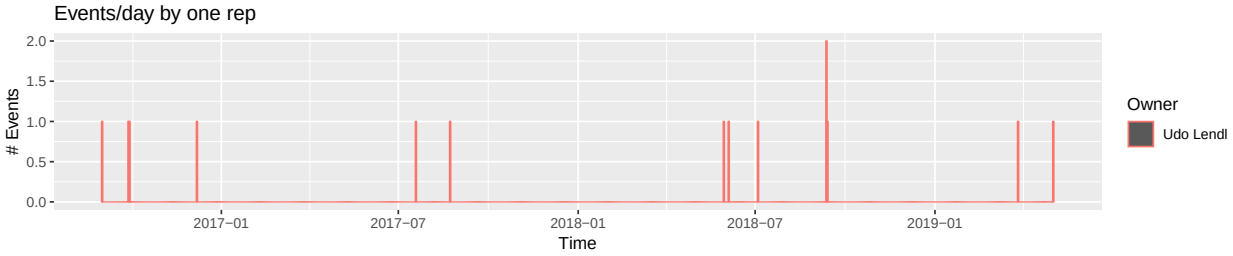

Events/week by one rep
Events
5
4
3
2
1
0
2017
2018
2019
Time
Owner
Angelika Gundlach

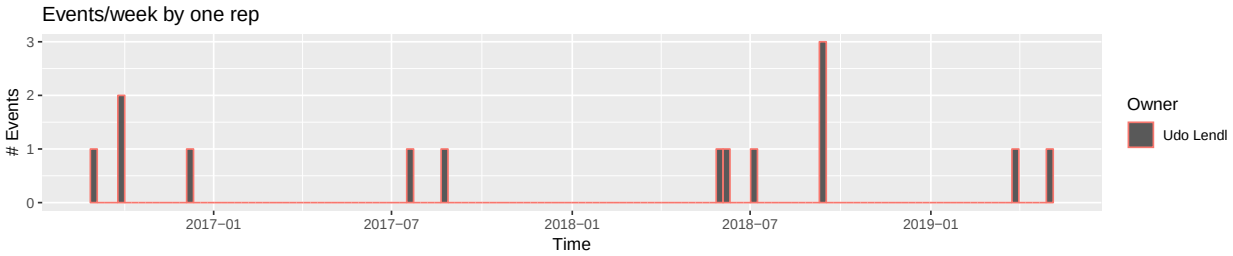

Events/month by one rep
Events
10.0
7.5
5.0
2.5
0.0
2017
2018
2019
Time
Owner
Angelika Gundlach

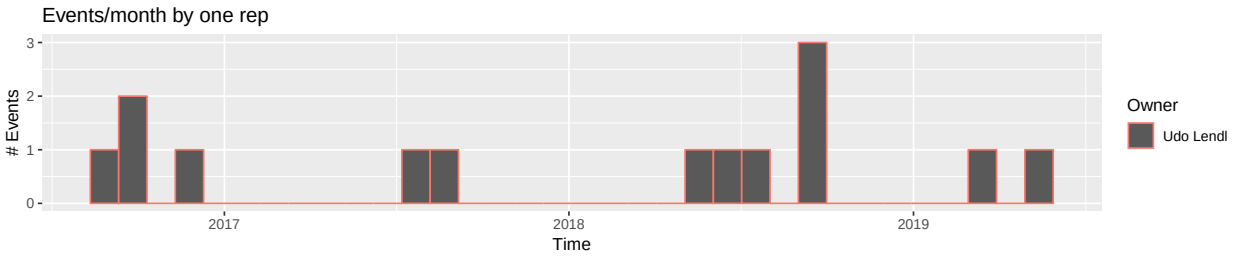

Events/day by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017
2018
2019
Time
Owner
Sarah Sahlu

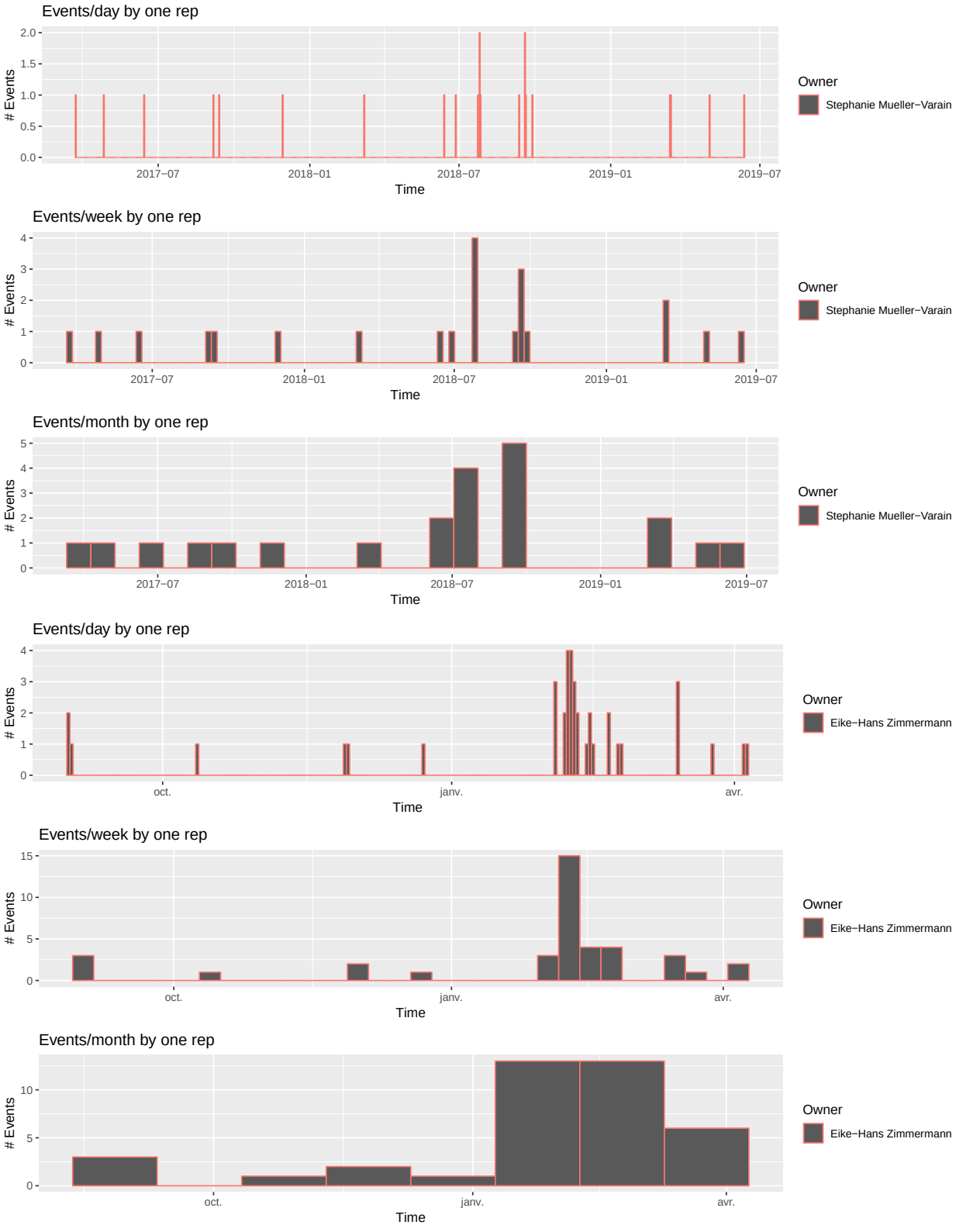

Events/week by one rep
Events
3
2
1
0
2017
2018
2019
Time
Owner
Sarah Sahlu

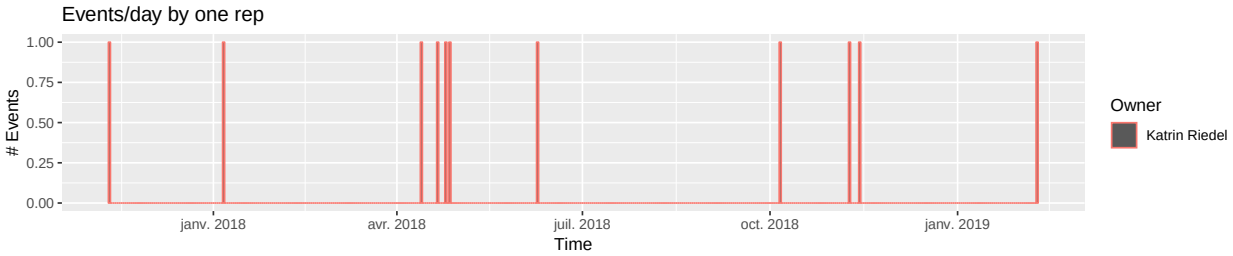

Events/month by one rep
Events
5
4
3
2
1
0
2017
2018
2019
Time
Owner
Sarah Sahlu

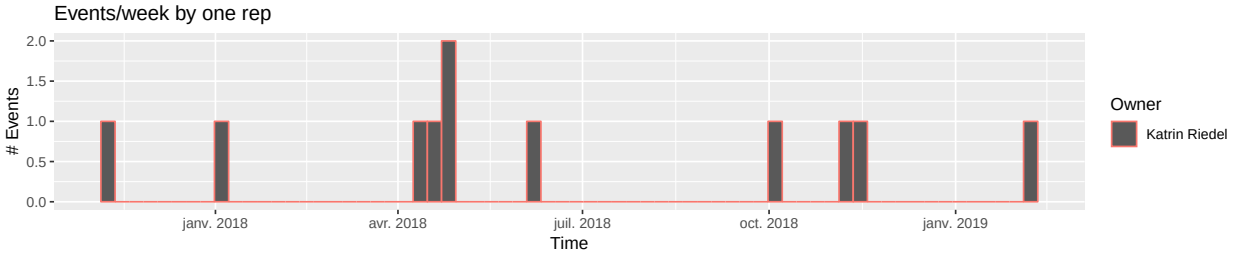

Events/day by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017–01
2017–07
2018–01
2018–07
2019–01
Time
Owner
Udo Lendl

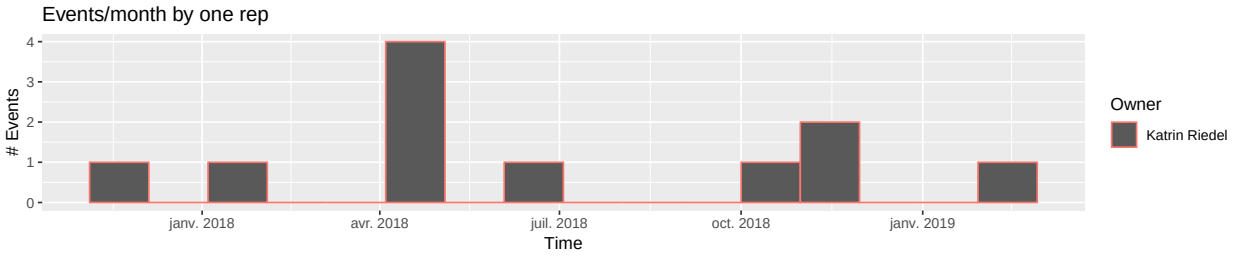

Events/week by one rep
Events
3
2
1
0
2017–01
2017–07
2018–01
2018–07
2019–01
Time
Owner
Udo Lendl

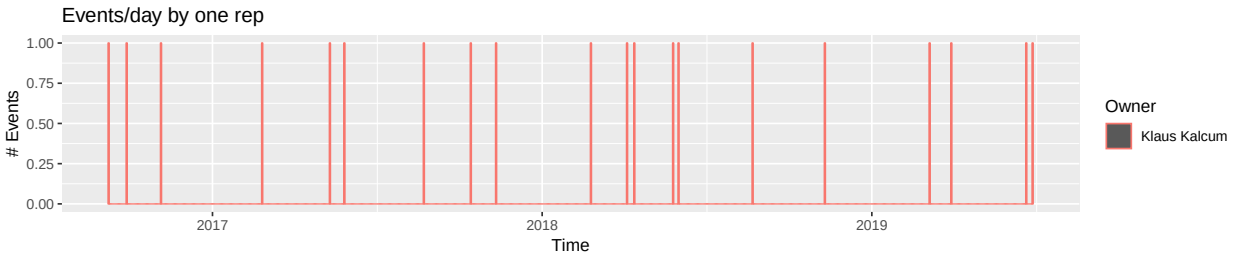

Events/month by one rep
Events
3
2
1
0
2017
2018
2019
Time
Owner
Udo Lendl

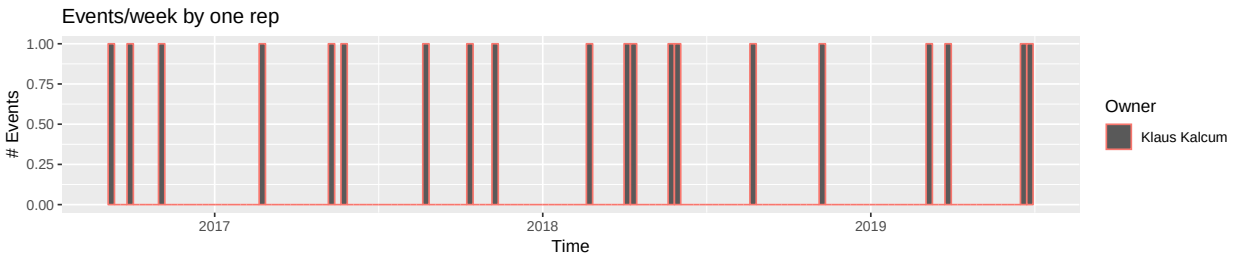

Events/day by one rep
Events
Time
2017–07
2018–01
2018–07
2019–01
2019–07
Owner
Stephanie Mueller−Varain
Events/week by one rep
Events
Time
2017–07
2018–01
2018–07
2019–01
2019–07
Owner
Stephanie Mueller−Varain
Events/month by one rep
Events
Time
2017–07
2018–01
2018–07
2019–01
2019–07
Owner
Stephanie Mueller−Varain
Events/day by one rep
Events
Time
oct.
janv.
avr.
Owner
Eike−Hans Zimmermann
Events/week by one rep
Events
Time
oct.
janv.
avr.
Owner
Eike−Hans Zimmermann
Events/month by one rep
Events
Time
oct.
janv.
avr.
Owner
Eike−Hans Zimmermann

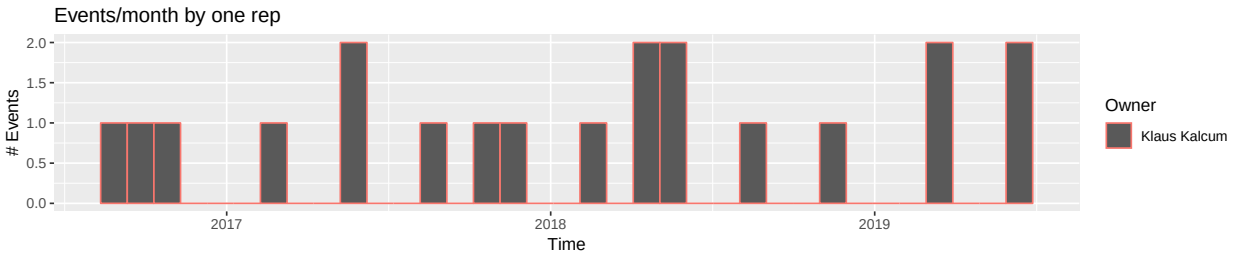

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
janv. 2018
avr. 2018
juil. 2018
oct. 2018
janv. 2019
Time
Owner
Katrin Riedel

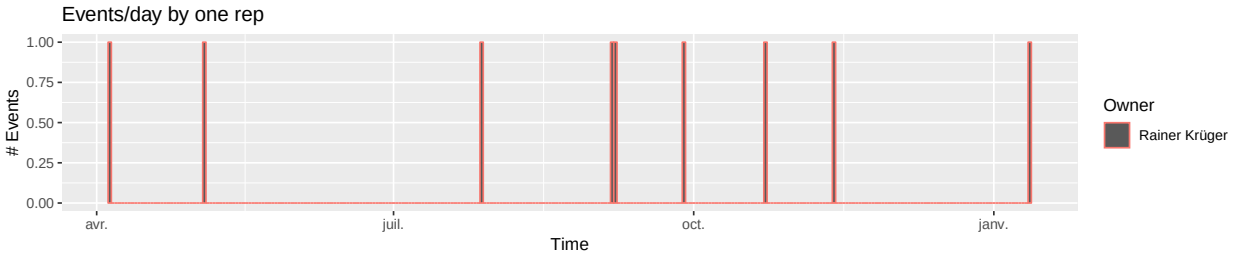

Events/week by one rep
Events
2.0
1.5
1.0
0.5
0.0
janv. 2018
avr. 2018
juil. 2018
oct. 2018
janv. 2019
Time
Owner
Katrin Riedel

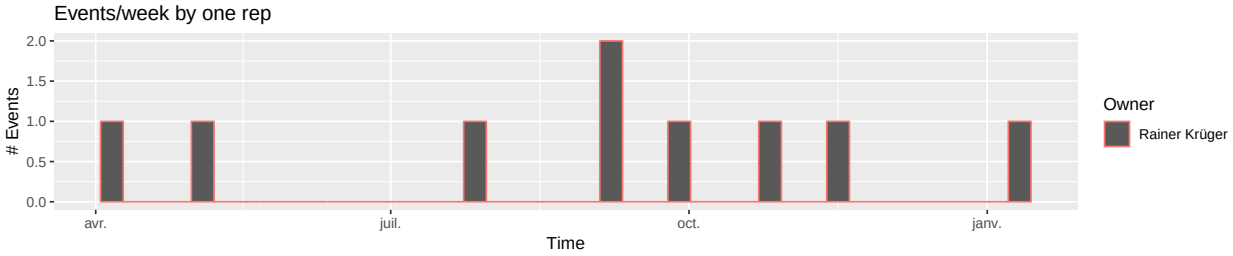

Events/month by one rep
Events
4
3
2
1
0
janv. 2018
avr. 2018
juil. 2018
oct. 2018
janv. 2019
Time
Owner
Katrin Riedel

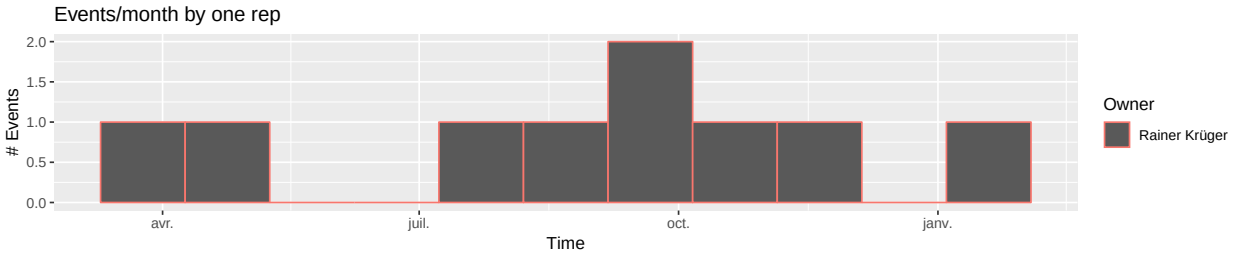

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017
2018
2019
Time
Owner
Klaus Kalcum

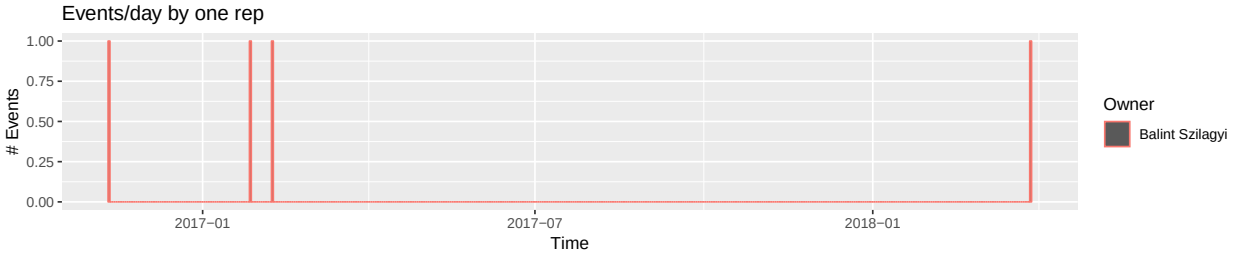

Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017
2018
2019
Time
Owner
Klaus Kalcum

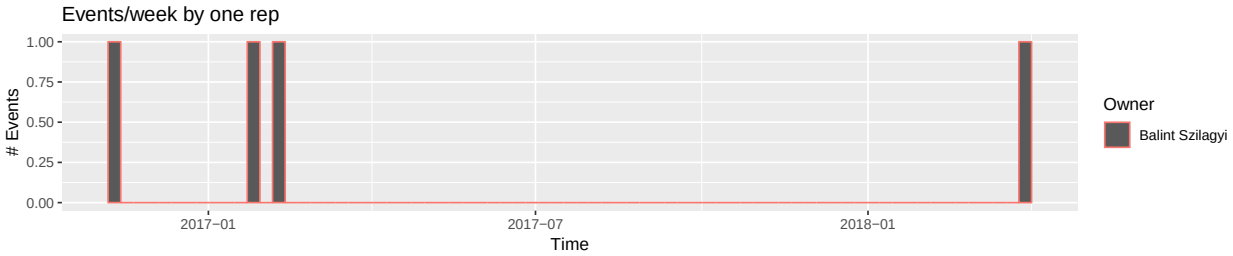

Events/month by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017
2018
2019
Time
Owner
Klaus Kalcum

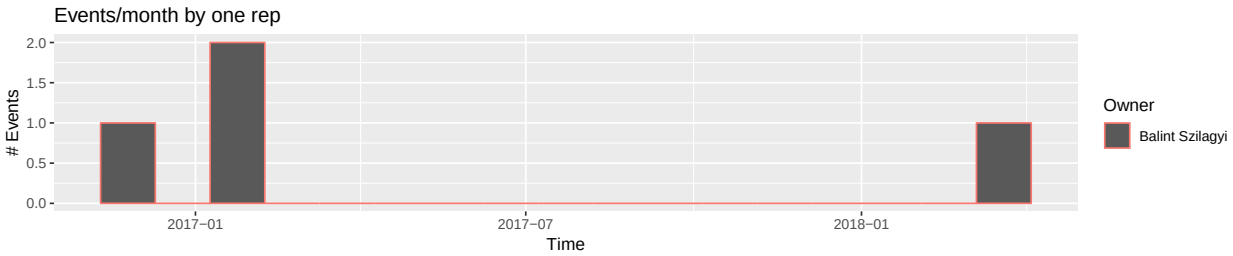

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
avr.
juil.
oct.
janv.
Time
Owner
Rainer Krüger

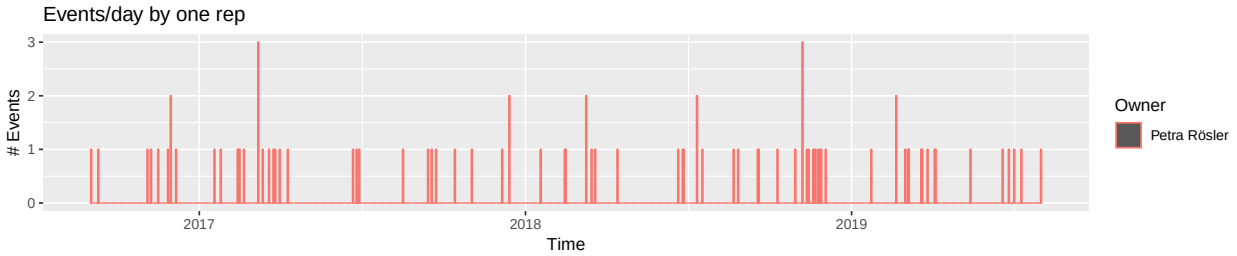

Events/week by one rep
Events
2.0
1.5
1.0
0.5
0.0
avr.
juil.
oct.
janv.
Time
Owner
Rainer Krüger

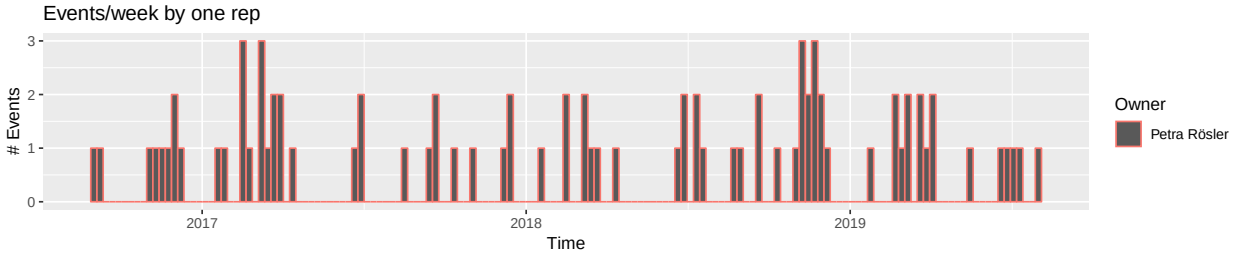

Events/month by one rep
Events
2.0
1.5
1.0
0.5
0.0
avr.
juil.
oct.
janv.
Time
Owner
Rainer Krüger

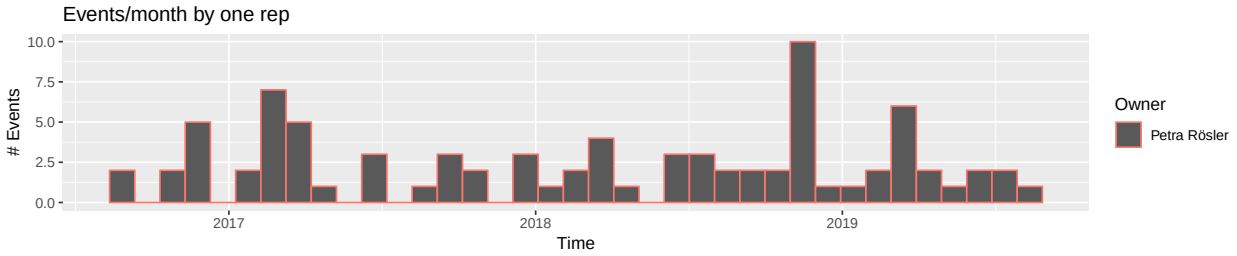

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017−01
2017−07
2018−01
Time
Owner
Balint Szilagyi

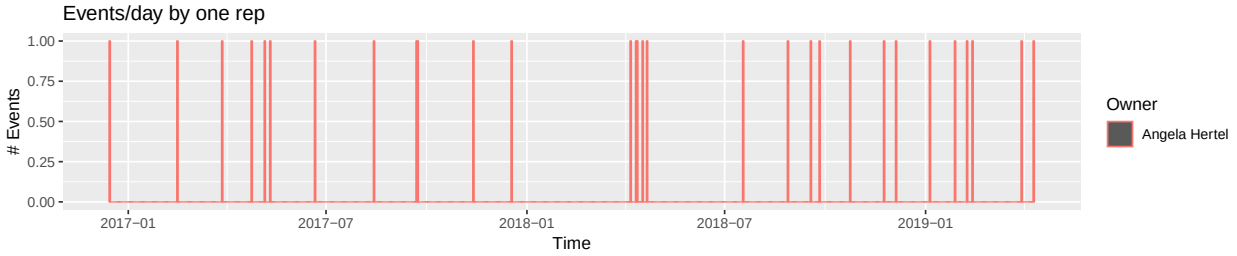

Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017−01
2017−07
2018−01
Time
Owner
Balint Szilagyi

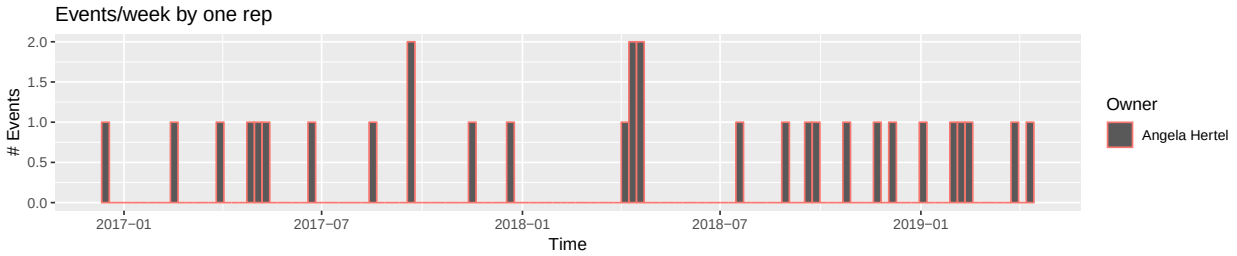

Events/month by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017−01
2017−07
2018−01
Time
Owner
Balint Szilagyi

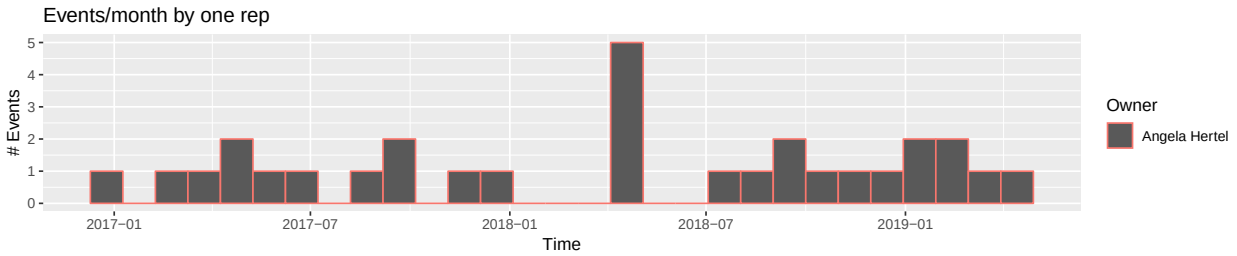

Events/day by one rep
Events
Time
2017
2018
2019
Owner
Petra Rösler

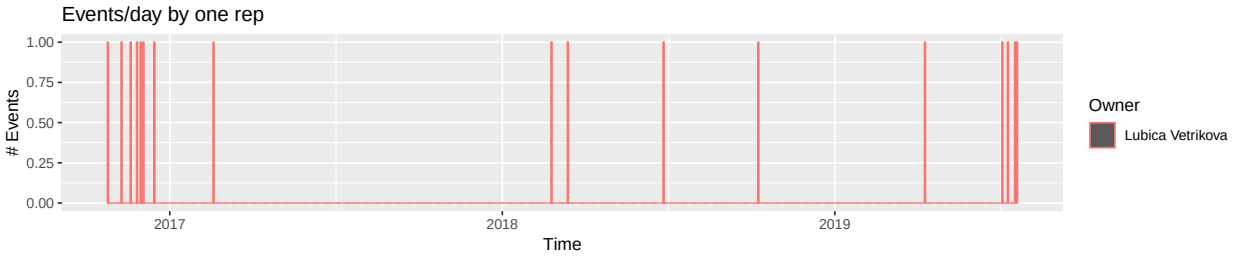

Events/week by one rep
Events
Time
2017
2018
2019
Owner
Petra Rösler

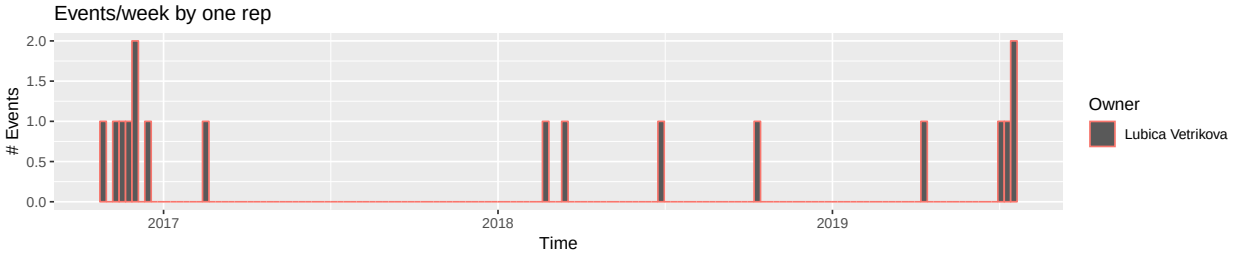

Events/month by one rep
Events
Time
2017
2018
2019
Owner
Petra Rösler

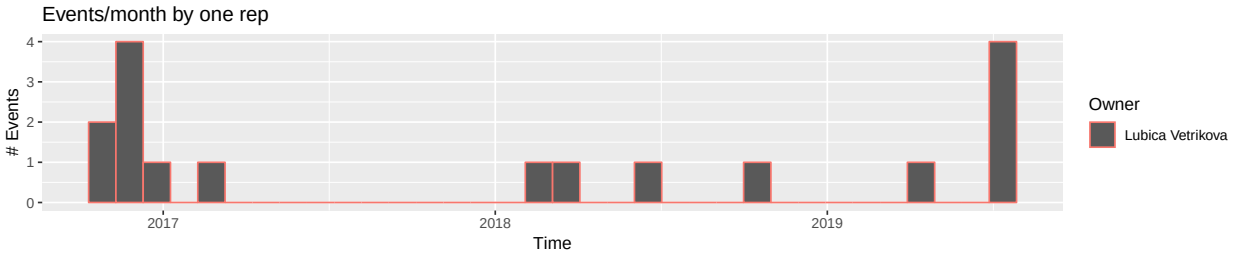

Events/day by one rep
Events
Time
2017–01
2017–07
2018–01
2018–07
2019–01
Owner
Angela Hertel

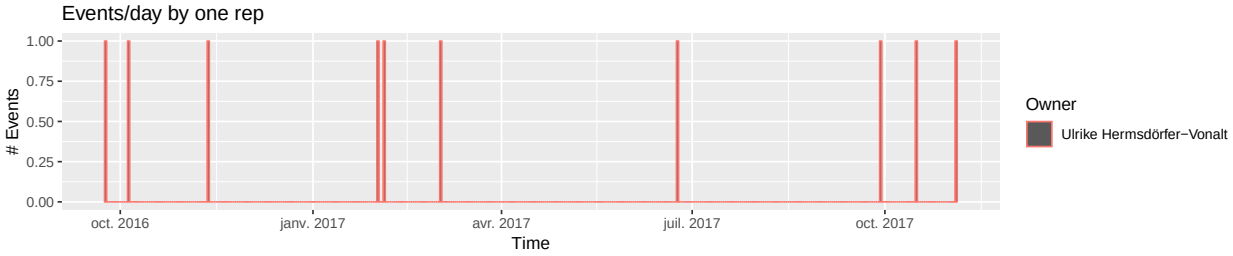

Events/week by one rep
Events
Time
2017–01
2017–07
2018–01
2018–07
2019–01
Owner
Angela Hertel

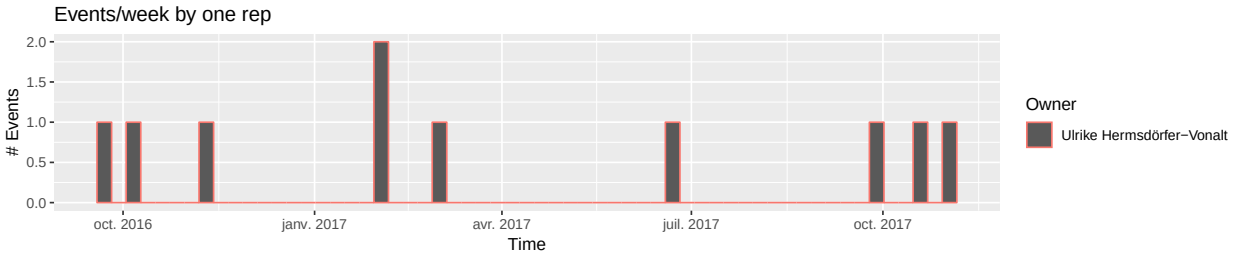

Events/month by one rep
Events
Time
2017–01
2017–07
2018–01
2018–07
2019–01
Owner
Angela Hertel

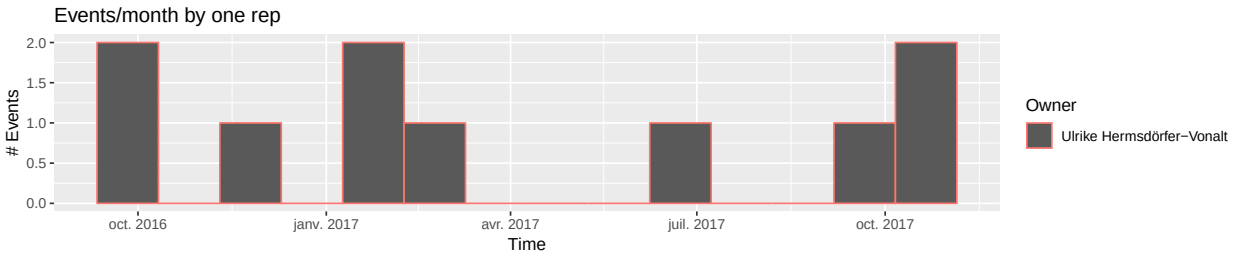

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017
2018
2019
Time
Owner
Lubica Vetrikova

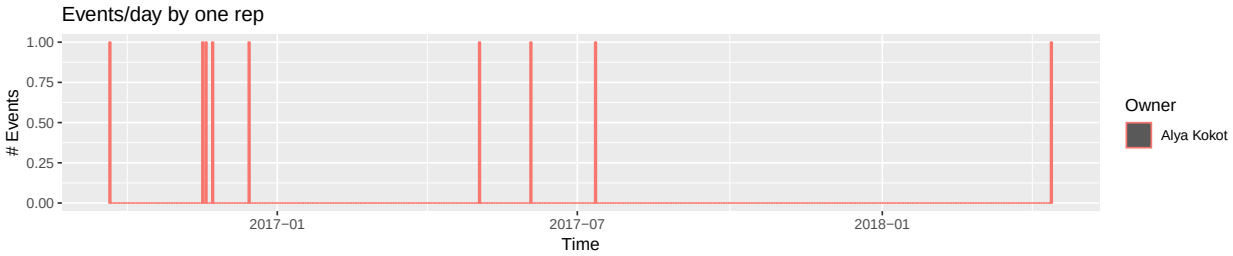

Events/week by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017
2018
2019
Time
Owner
Lubica Vetrikova

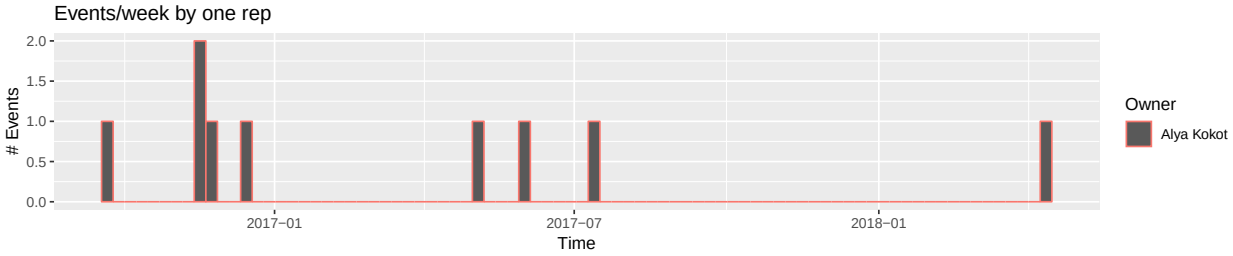

Events/month by one rep
Events
4
3
2
1
0
2017
2018
2019
Time
Owner
Lubica Vetrikova

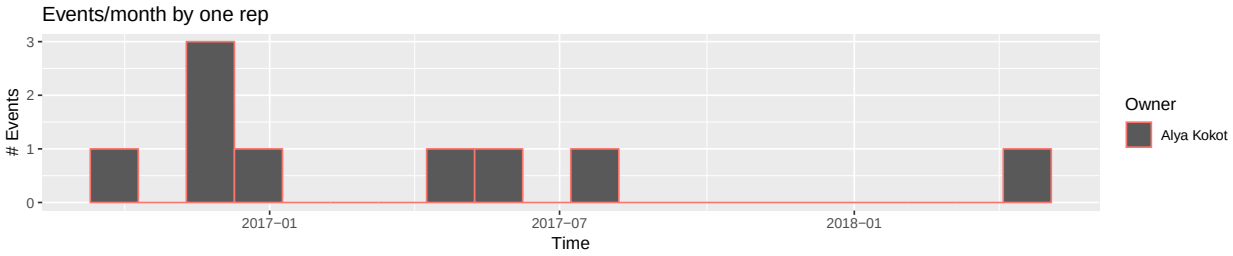

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
oct. 2016
janv. 2017
avr. 2017
juil. 2017
oct. 2017
Time
Owner
Ulrike Hermsdörfer−Vonalt

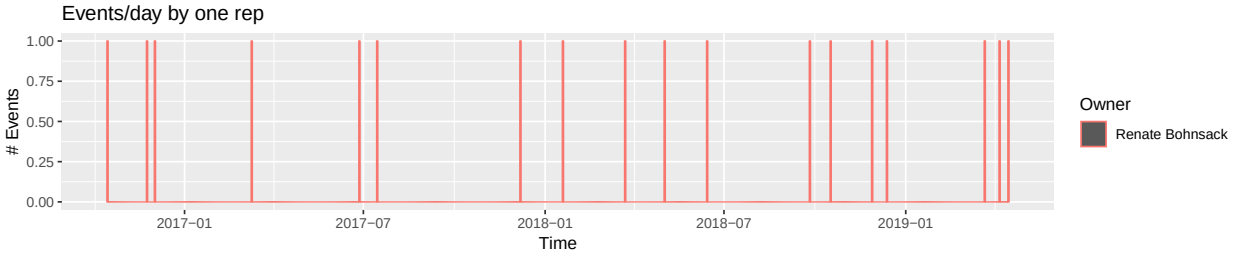

Events/week by one rep
Events
2.0
1.5
1.0
0.5
0.0
oct. 2016
janv. 2017
avr. 2017
juil. 2017
oct. 2017
Time
Owner
Ulrike Hermsdörfer−Vonalt

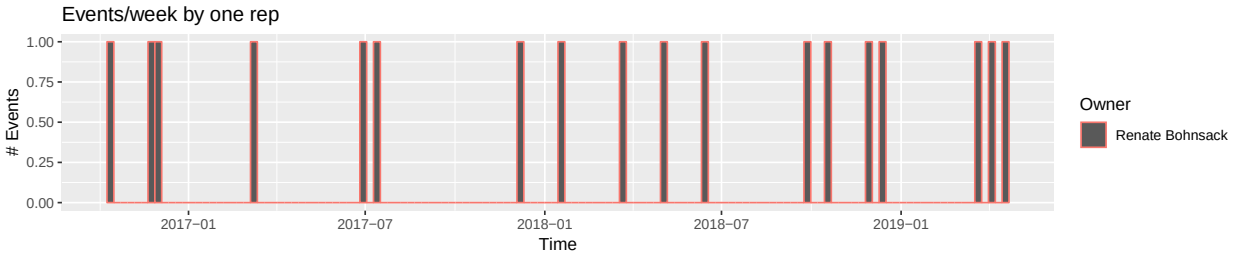

Events/month by one rep
Events
2.0
1.5
1.0
0.5
0.0
oct. 2016
janv. 2017
avr. 2017
juil. 2017
oct. 2017
Time
Owner
Ulrike Hermsdörfer−Vonalt

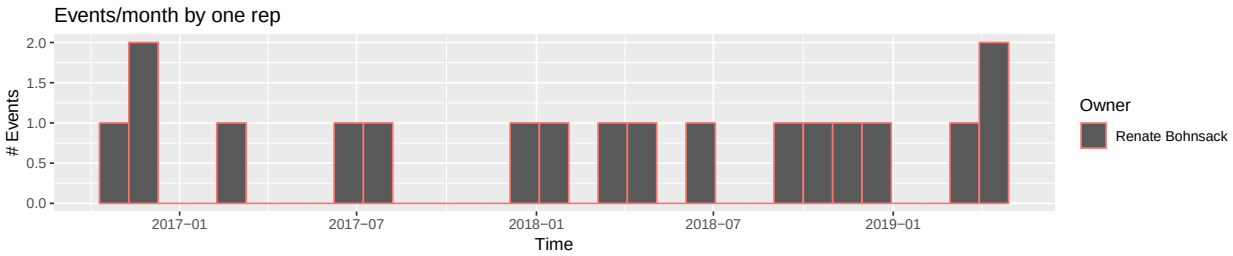

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017−01
2017−07
2018−01
Time
Owner
Alya Kokot

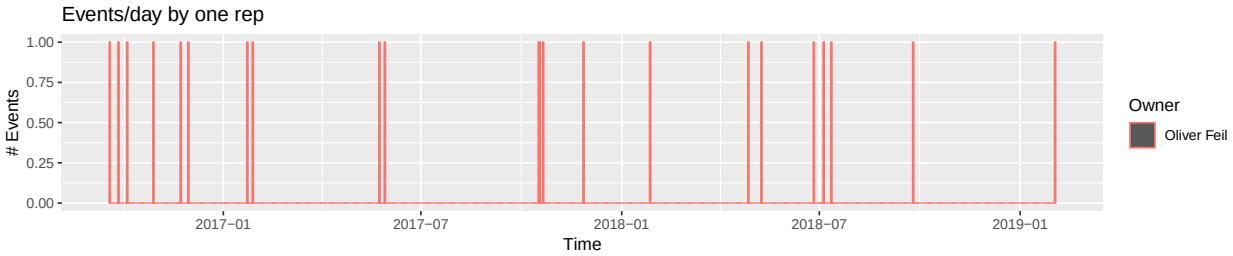

Events/week by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017−01
2017−07
2018−01
Time
Owner
Alya Kokot

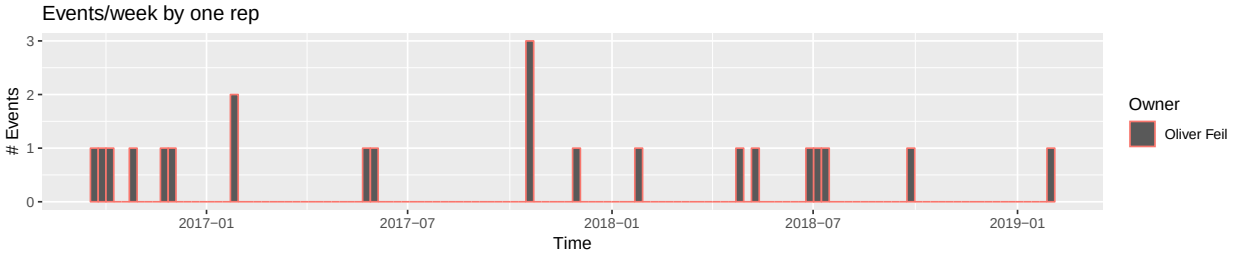

Events/month by one rep
Events
3
2
1
0
2017−01
2017−07
2018−01
Time
Owner
Alya Kokot

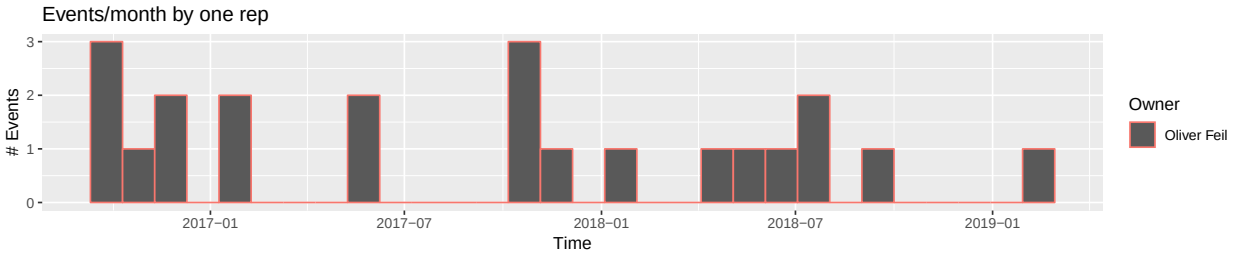

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017−01
2017−07
2018−01
2018−07
2019−01
Time
Owner
Renate Bohnsack

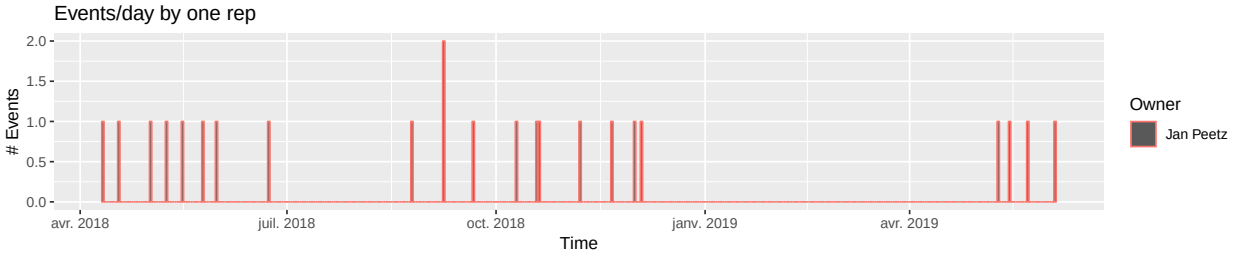

Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017−01
2017−07
2018−01
2018−07
2019−01
Time
Owner
Renate Bohnsack

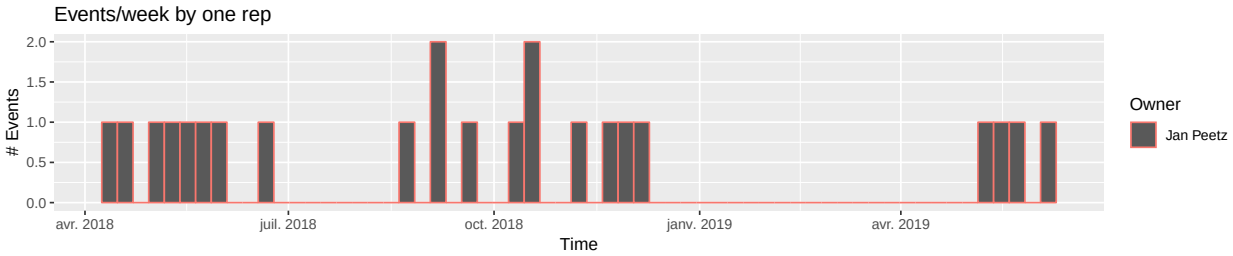

Events/month by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017−01
2017−07
2018−01
2018−07
2019−01
Time
Owner
Renate Bohnsack

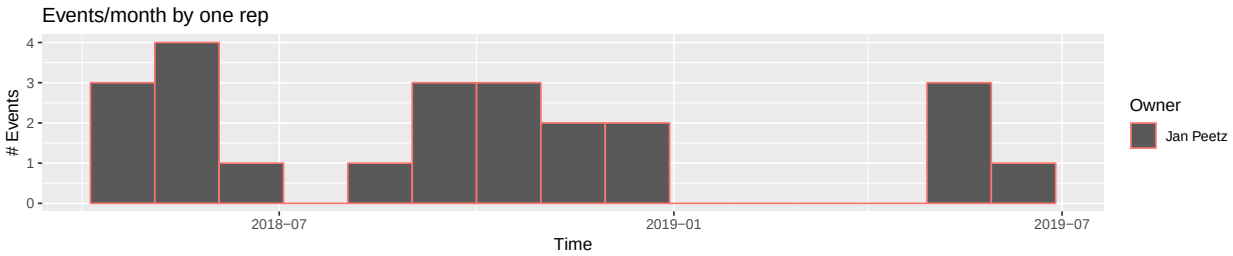

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017−01
2017−07
2018−01
2018−07
2019−01
Time
Owner
Oliver Feil

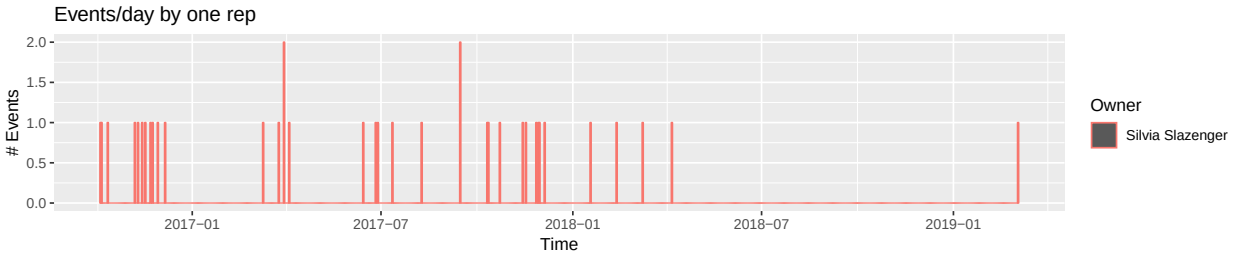

Events/week by one rep
Events
3
2
1
0
2017−01
2017−07
2018−01
2018−07
2019−01
Time
Owner
Oliver Feil

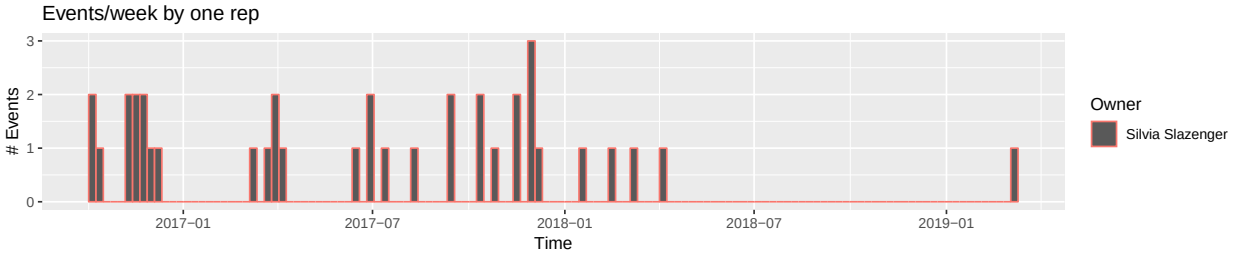

Events/month by one rep
Events
3
2
1
0
2017−01
2017−07
2018−01
2018−07
2019−01
Time
Owner
Oliver Feil

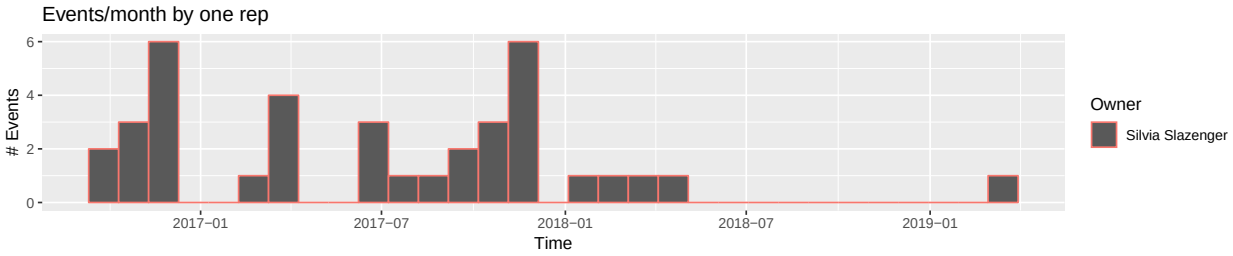

Events/day by one rep
Events
2.0
1.5
1.0
0.5
0.0
avr. 2018
juil. 2018
oct. 2018
janv. 2019
avr. 2019
Time
Owner
Jan Peetz

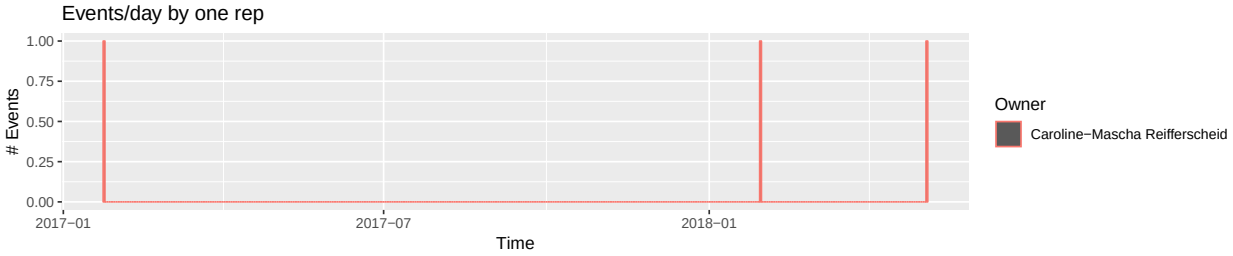

Events/week by one rep
Events
2.0
1.5
1.0
0.5
0.0
avr. 2018
juil. 2018
oct. 2018
janv. 2019
avr. 2019
Time
Owner
Jan Peetz

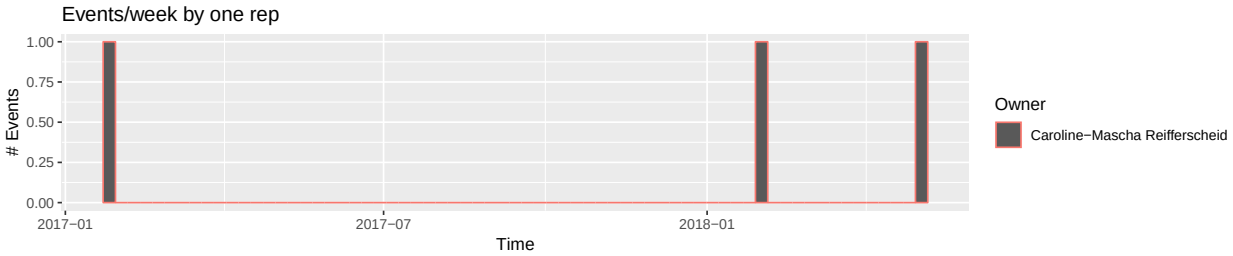

Events/month by one rep
Events
4
3
2
1
0
2018−07
2019−01
2019−07
Time
Owner
Jan Peetz

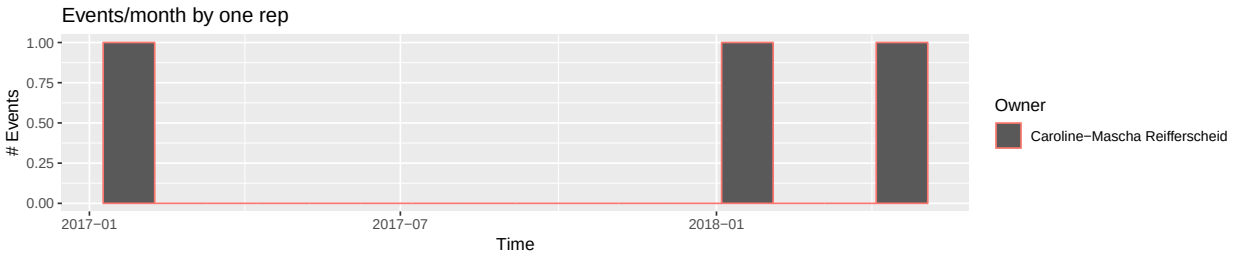

Events/day by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017–01
2017–07
2018–01
2018–07
2019–01
Time
Owner
Silvia Slazenger

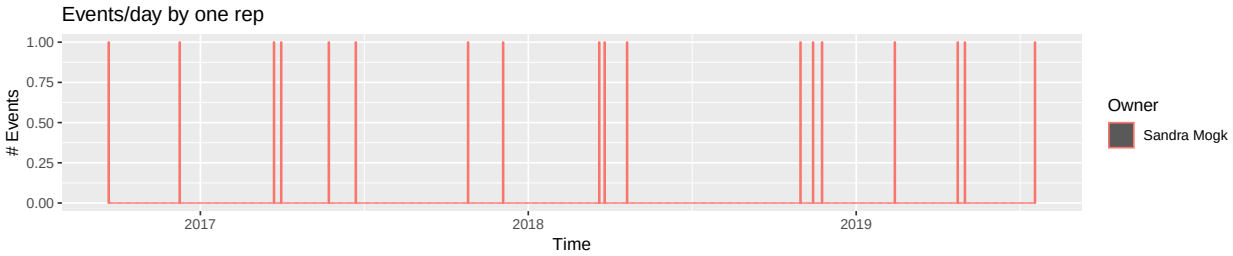

Events/week by one rep
Events
3
2
1
0
2017–01
2017–07
2018–01
2018–07
2019–01
Time
Owner
Silvia Slazenger

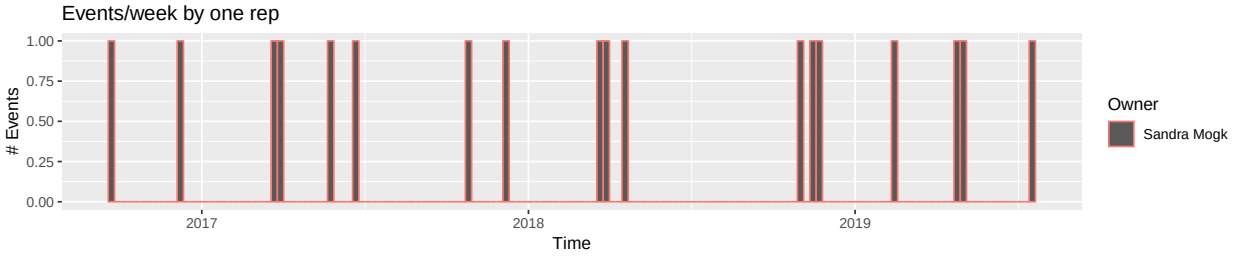

Events/month by one rep
Events
6
4
2
0
2017–01
2017–07
2018–01
2018–07
2019–01
Time
Owner
Silvia Slazenger

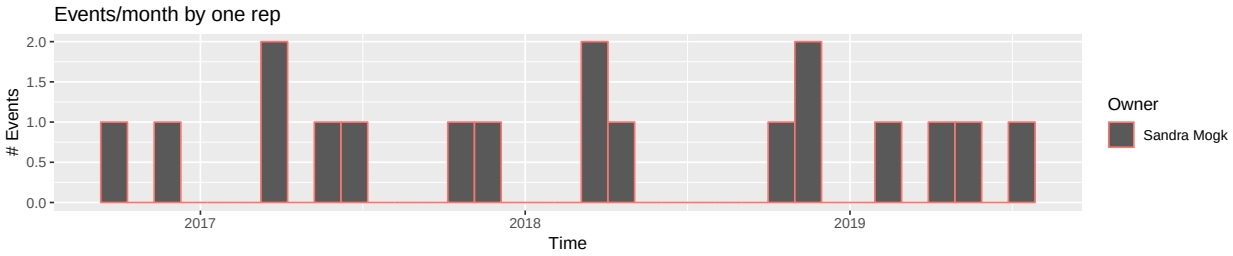

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017–01
2017–07
2018–01
Time
Owner
Caroline−Mascha Reifferscheid

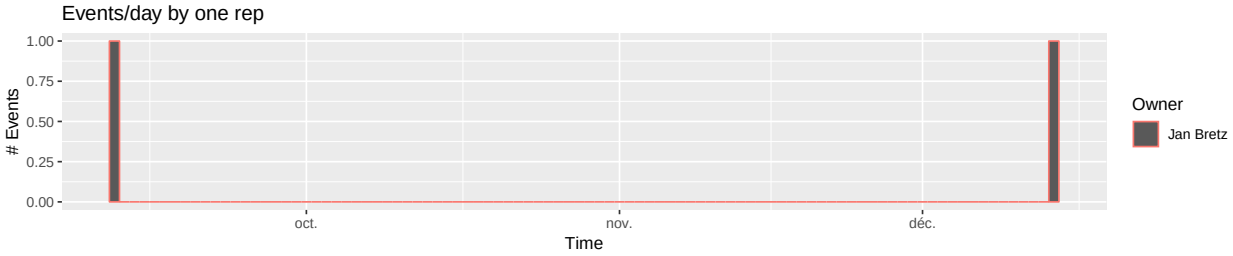

Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017–01
2017–07
2018–01
Time
Owner
Caroline−Mascha Reifferscheid

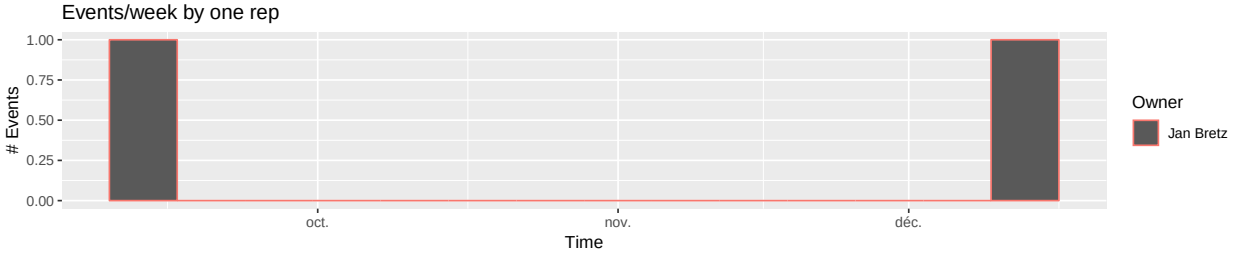

Events/month by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017–01
2017–07
2018–01
Time
Owner
Caroline−Mascha Reifferscheid

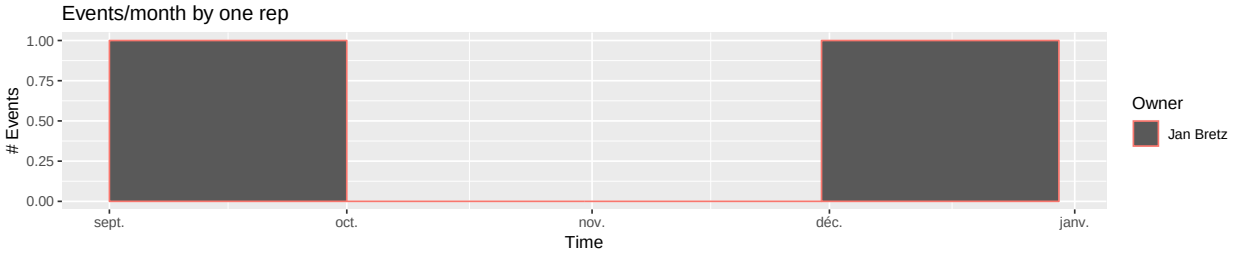

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017
2018
2019
Time
Owner
Sandra Mogk

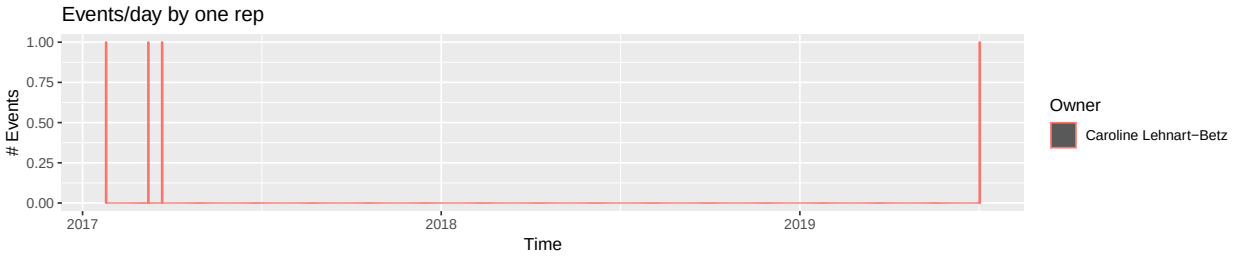

Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017
2018
2019
Time
Owner
Sandra Mogk

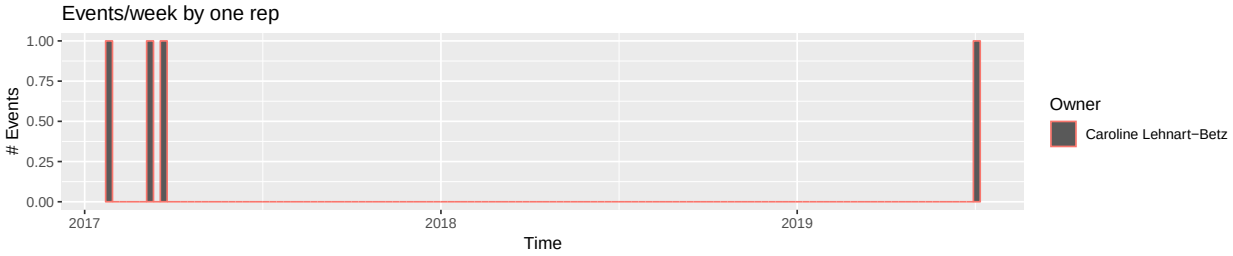

Events/month by one rep
Events
2.0
1.5
1.0
0.5
0.0
2017
2018
2019
Time
Owner
Sandra Mogk

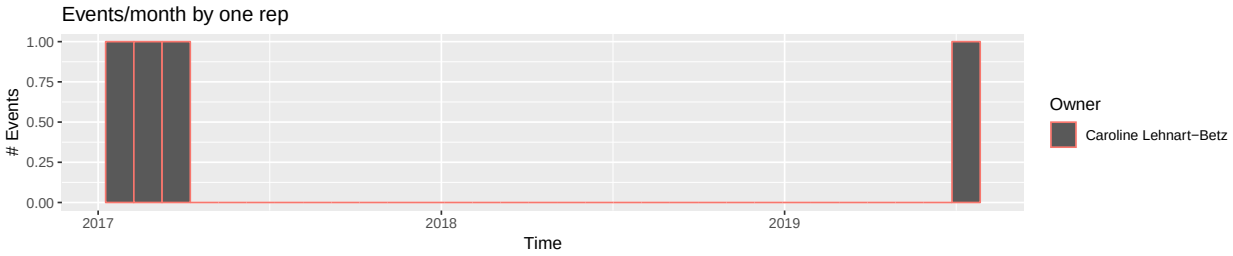

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
oct.
nov.
déc.
Time
Owner
Jan Bretz

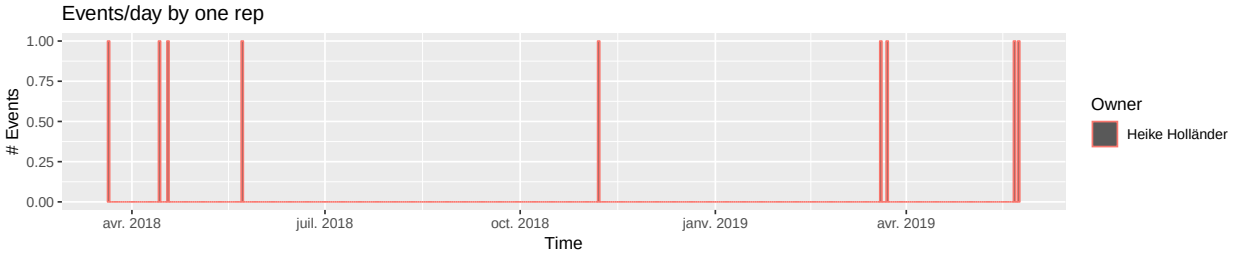

Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
oct.
nov.
déc.
Time
Owner
Jan Bretz

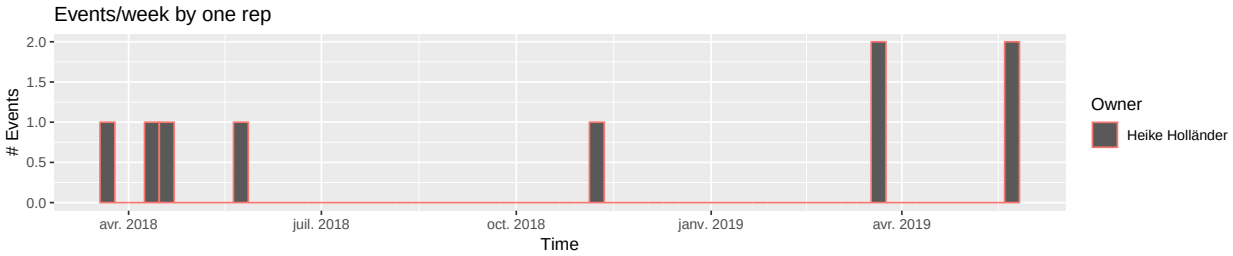

Events/month by one rep
Events
1.00
0.75
0.50
0.25
0.00
sept.
oct.
nov.
déc.
janv.
Time
Owner
Jan Bretz

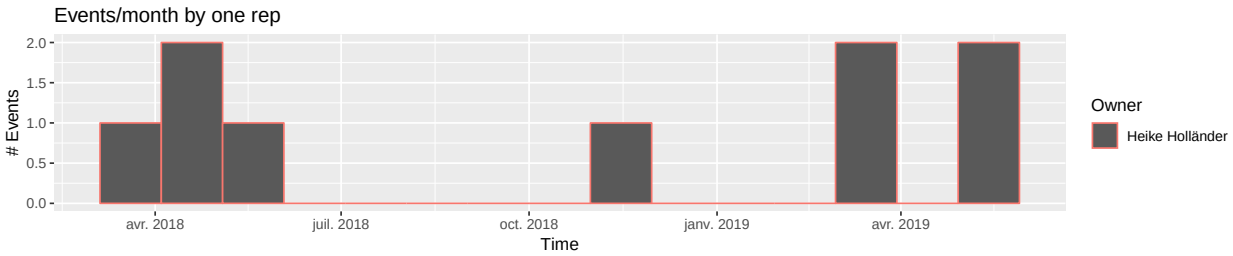

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017
2018
2019
Time
Owner
Caroline Lehnart−Betz

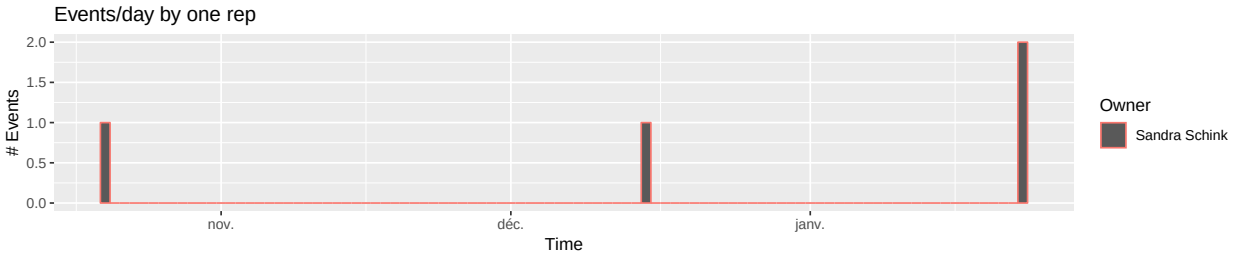

Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017
2018
2019
Time
Owner
Caroline Lehnart−Betz

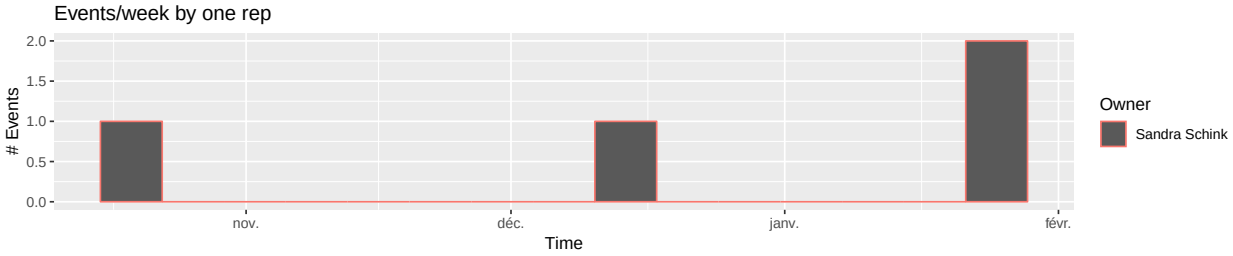

Events/month by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017
2018
2019
Time
Owner
Caroline Lehnart−Betz

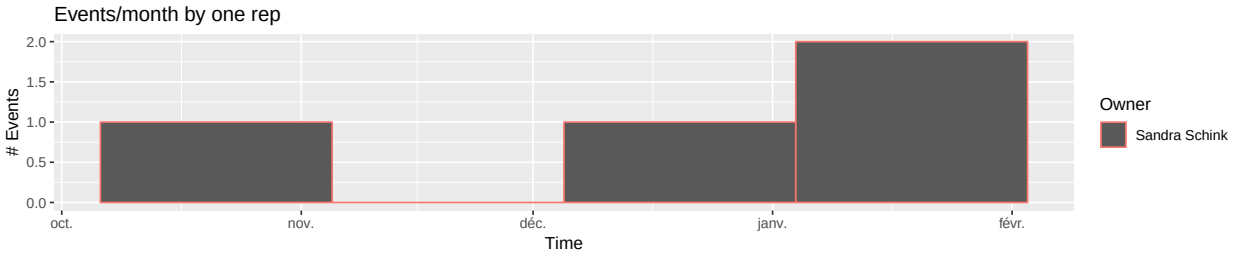

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
avr. 2018
juil. 2018
oct. 2018
janv. 2019
avr. 2019
Time
Owner
Heike Holländer

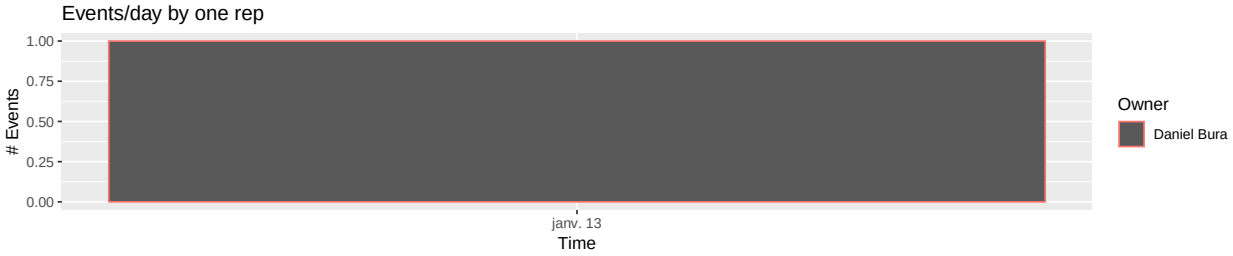

Events/week by one rep
Events
2.0
1.5
1.0
0.5
0.0
avr. 2018
juil. 2018
oct. 2018
janv. 2019
avr. 2019
Time
Owner
Heike Holländer

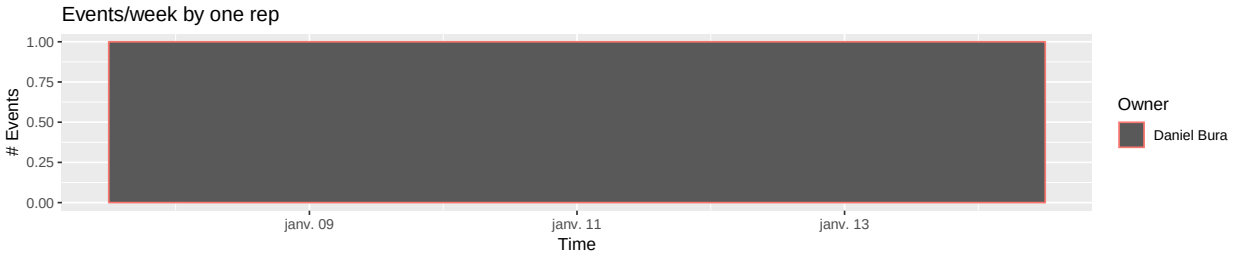

Events/month by one rep
Events
2.0
1.5
1.0
0.5
0.0
avr. 2018
juil. 2018
oct. 2018
janv. 2019
avr. 2019
Time
Owner
Heike Holländer

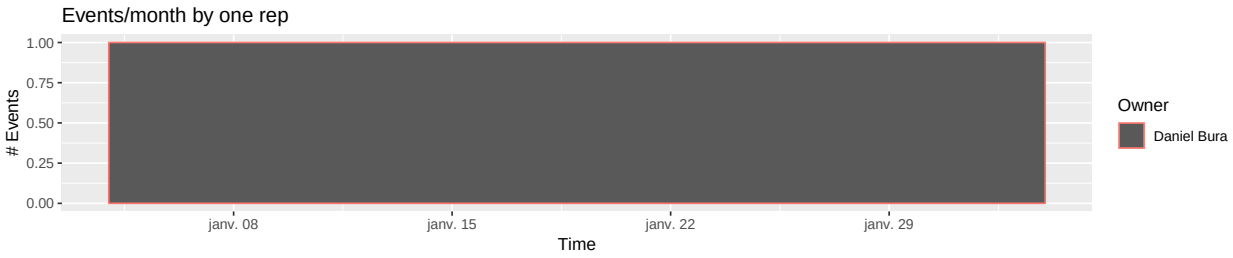

Events/day by one rep
Events
2.0
1.5
1.0
0.5
0.0
nov.
déc.
janv.
Time
Owner
Sandra Schink

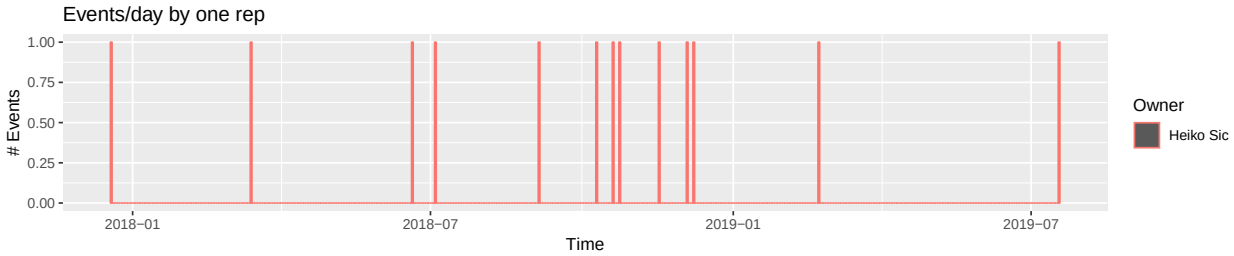

Events/week by one rep
Events
2.0
1.5
1.0
0.5
0.0
nov.
déc.
janv.
févr.
Time
Owner
Sandra Schink

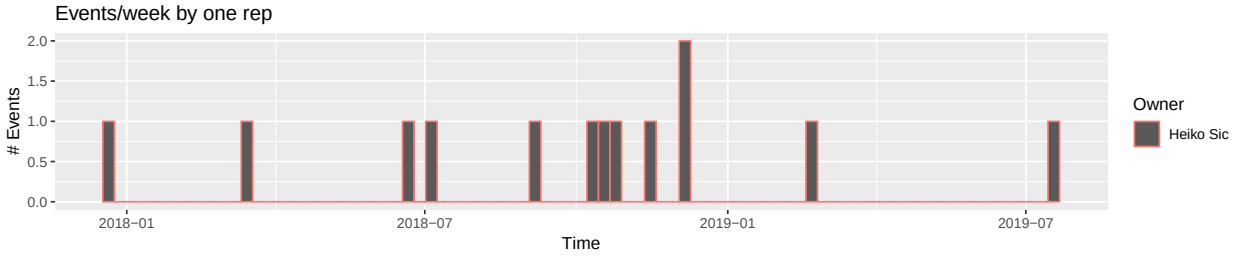

Events/month by one rep
Events
2.0
1.5
1.0
0.5
0.0
oct.
nov.
déc.
janv.
févr.
Time
Owner
Sandra Schink

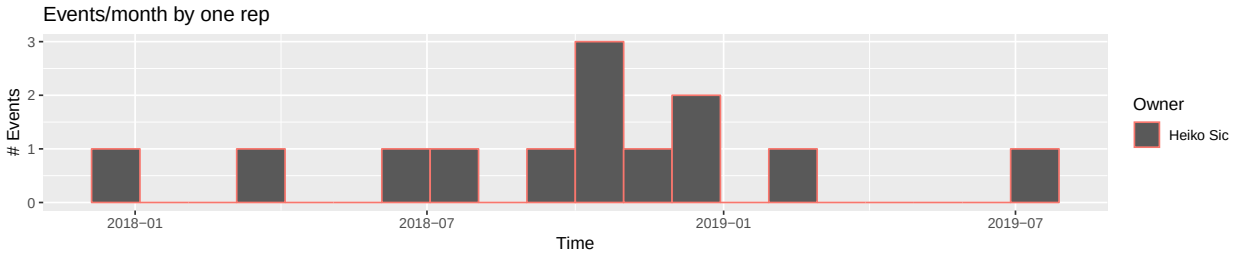

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
janv. 13
Time
Owner
Daniel Bura


Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
janv. 09
janv. 11
janv. 13
Time
Owner
Daniel Bura

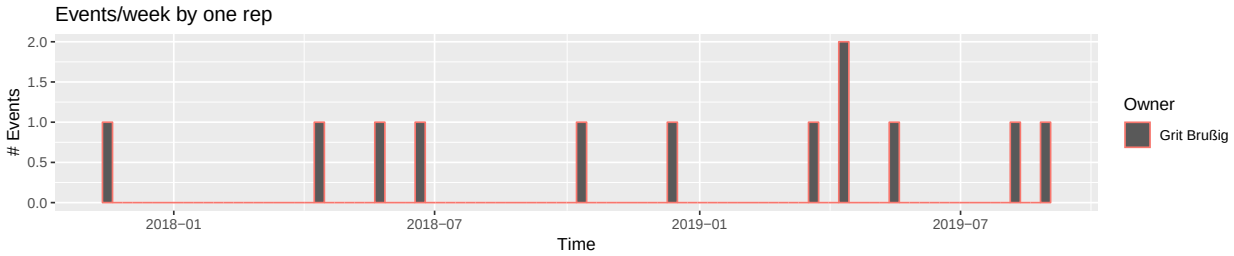

Events/month by one rep
Events
1.00
0.75
0.50
0.25
0.00
janv. 08
janv. 15
janv. 22
janv. 29
Time
Owner
Daniel Bura

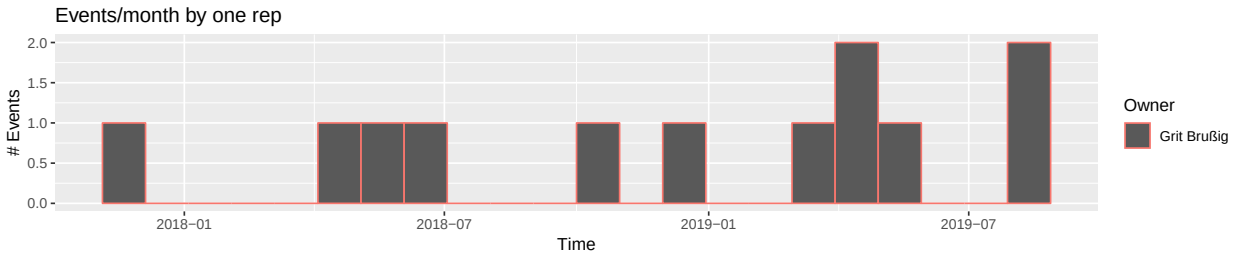

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
2018–01
2018–07
2019–01
2019–07
Time
Owner
Heiko Sic

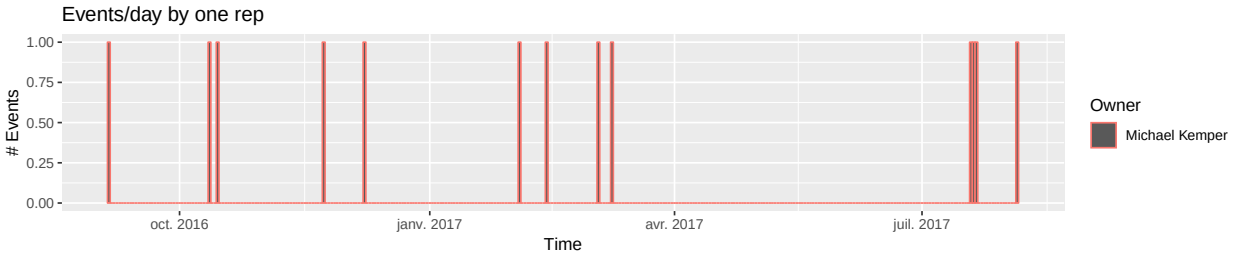

Events/week by one rep
Events
2.0
1.5
1.0
0.5
0.0
2018–01
2018–07
2019–01
2019–07
Time
Owner
Heiko Sic

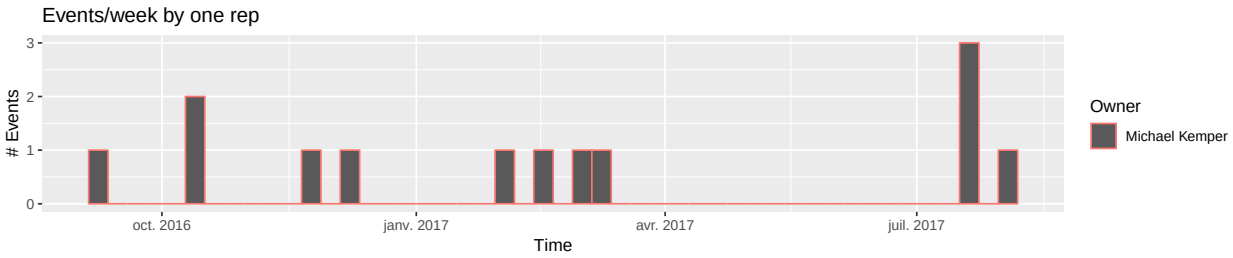

Events/month by one rep
Events
3
2
1
0
2018–01
2018–07
2019–01
2019–07
Time
Owner
Heiko Sic

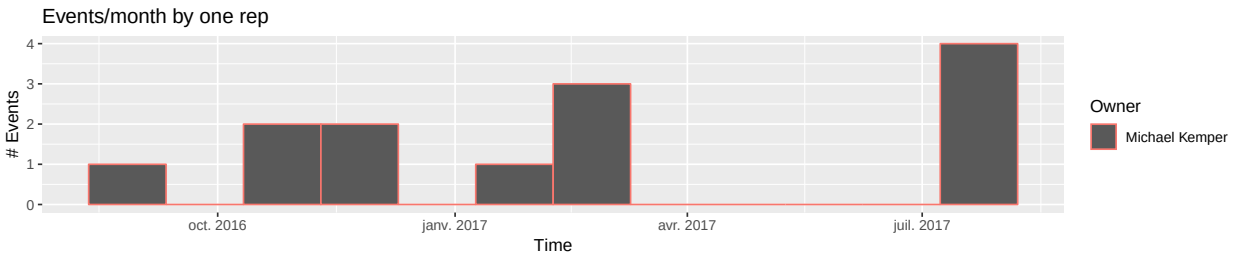

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
2018–01
2018–07
2019–01
2019–07
Time
Owner
Grit Brußig

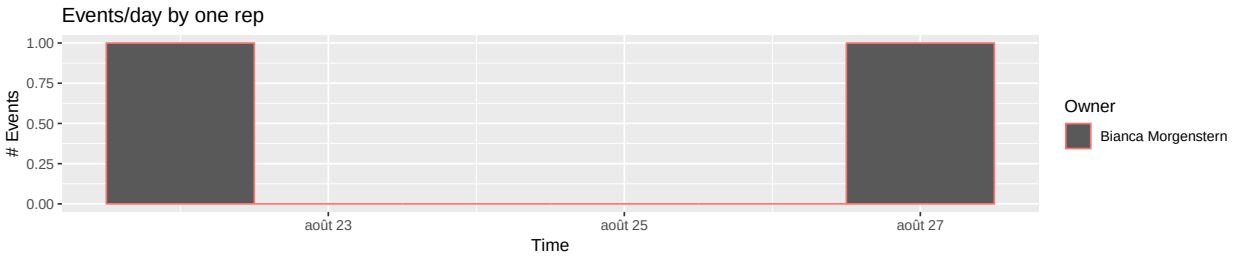

Events/week by one rep
Events
2.0
1.5
1.0
0.5
0.0
2018–01
2018–07
2019–01
2019–07
Time
Owner
Grit Brußig

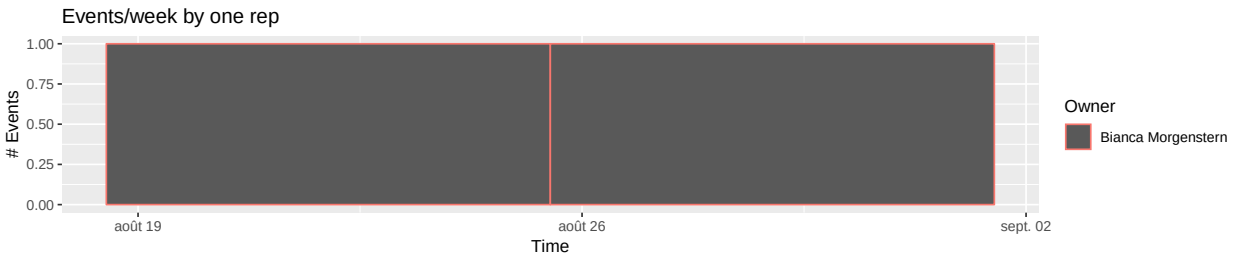

Events/month by one rep
Events
2.0
1.5
1.0
0.5
0.0
2018–01
2018–07
2019–01
2019–07
Time
Owner
Grit Brußig

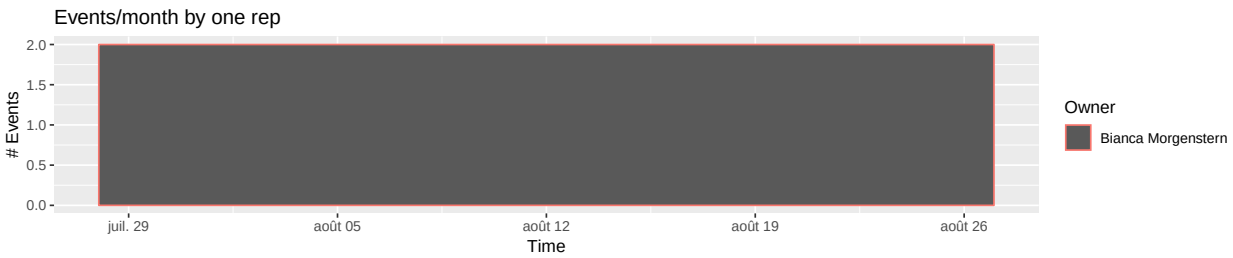

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
oct. 2016
janv. 2017
avr. 2017
juil. 2017
Time
Owner
Michael Kemper

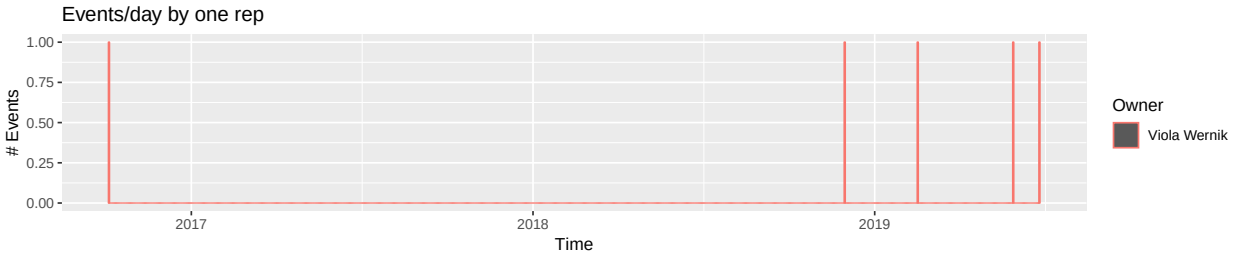

Events/week by one rep
Events
3
2
1
0
oct. 2016
janv. 2017
avr. 2017
juil. 2017
Time
Owner
Michael Kemper

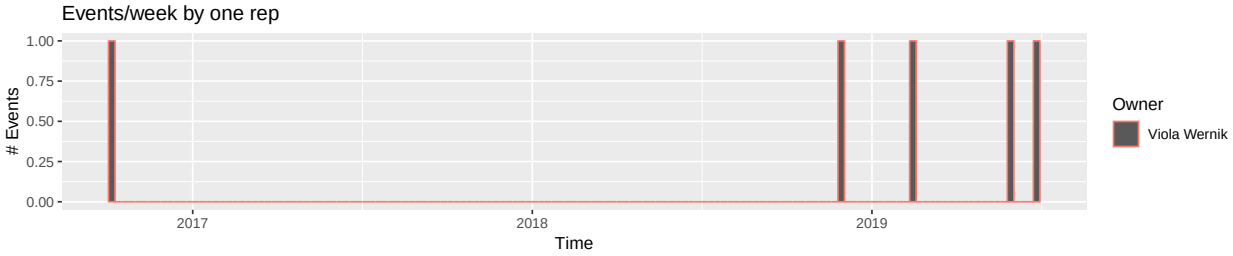

Events/month by one rep
Events
4
3
2
1
0
oct. 2016
janv. 2017
avr. 2017
juil. 2017
Time
Owner
Michael Kemper

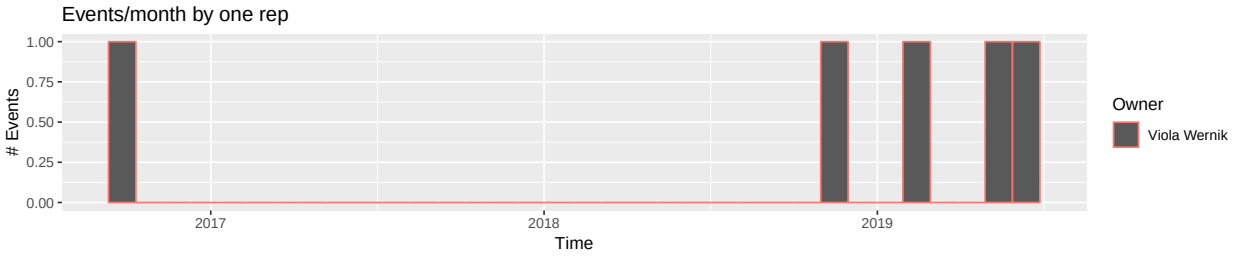

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
août 23
août 25
août 27
Time
Owner
Bianca Morgenstern

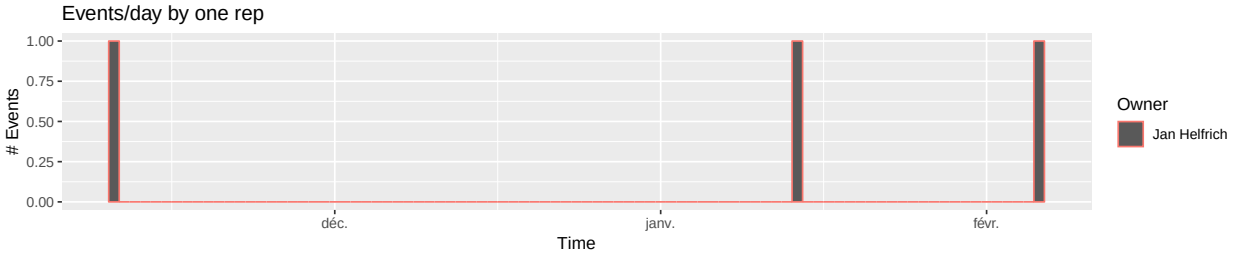

Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
août 19
août 26
sept. 02
Time
Owner
Bianca Morgenstern

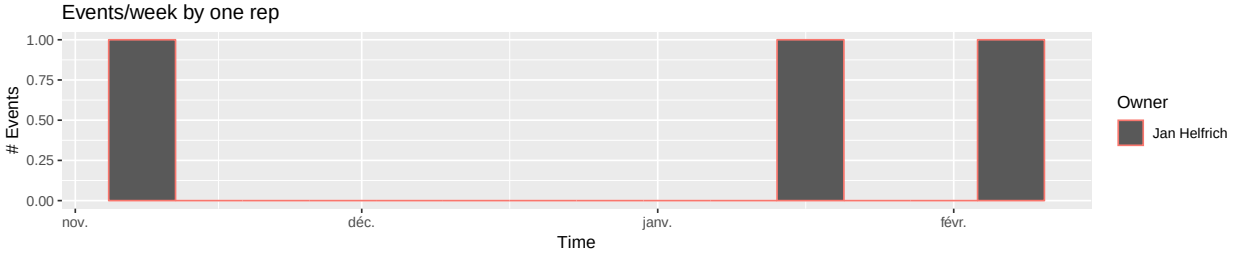

Events/month by one rep
Events
2.0
1.5
1.0
0.5
0.0
juil. 29
août 05
août 12
août 19
août 26
Time
Owner
Bianca Morgenstern

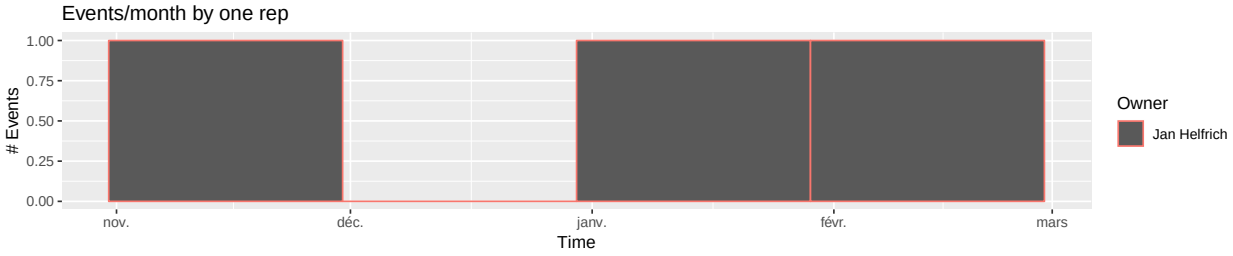

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017
2018
2019
Time
Owner
Viola Wernik

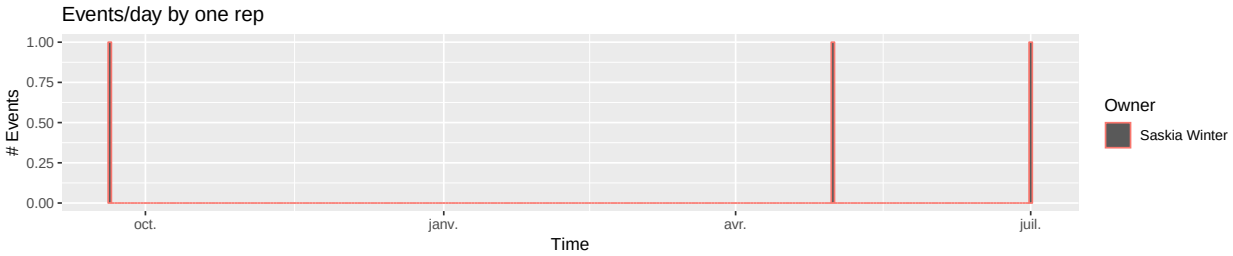

Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017
2018
2019
Time
Owner
Viola Wernik

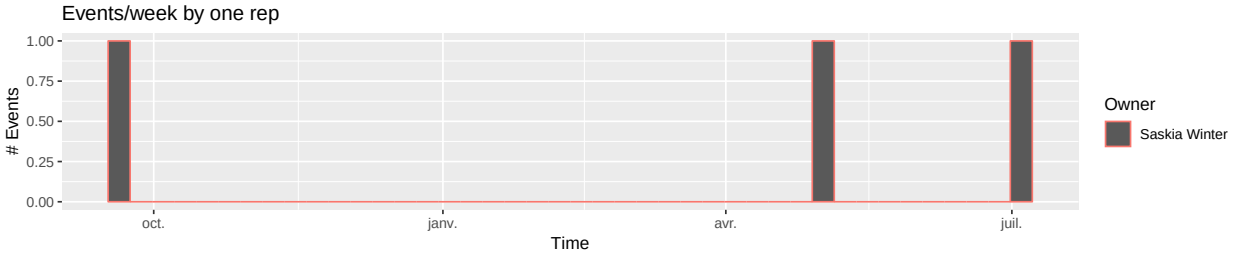

Events/month by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017
2018
2019
Time
Owner
Viola Wernik

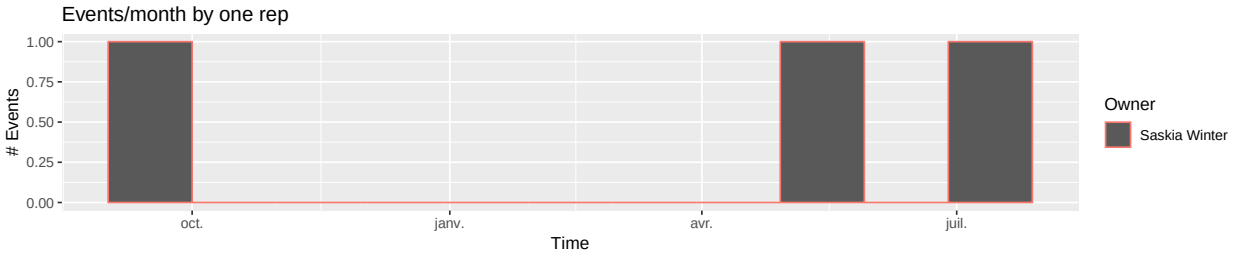

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
déc.
janv.
févr.
Time
Owner
Jan Helfrich

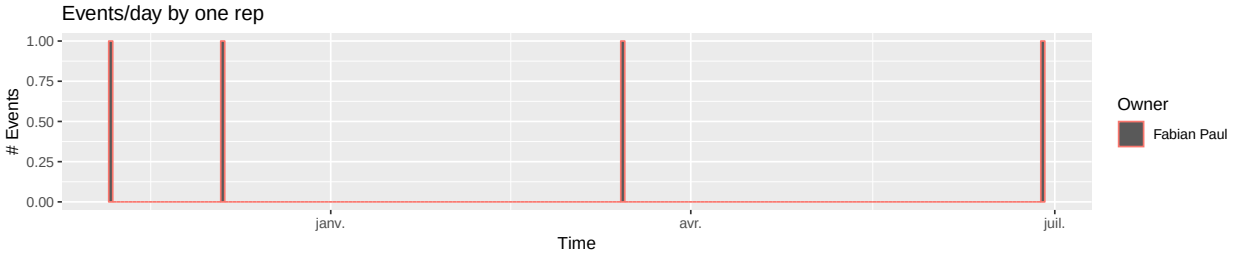

Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
nov.
déc.
janv.
févr.
Time
Owner
Jan Helfrich

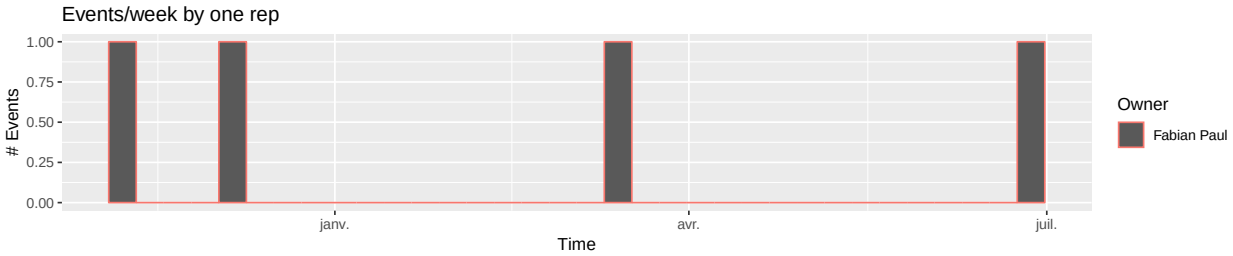

Events/month by one rep
Events
1.00
0.75
0.50
0.25
0.00
nov.
déc.
janv.
févr.
mars
Time
Owner
Jan Helfrich

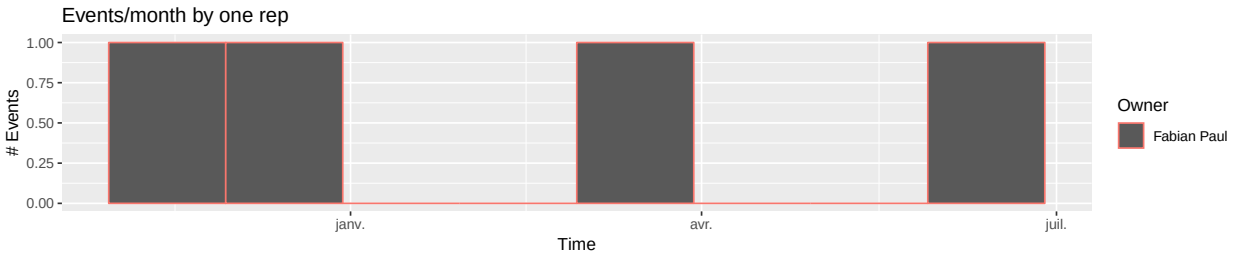

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
oct.
janv.
avr.
juil.
Time
Owner
Saskia Winter

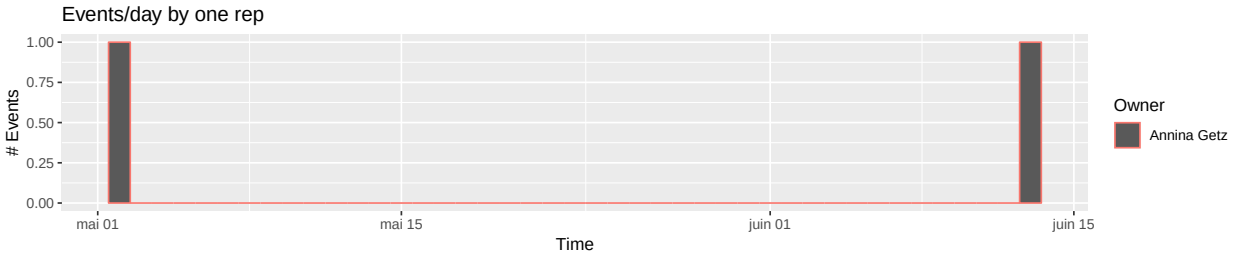

Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
oct.
janv.
avr.
juil.
Time
Owner
Saskia Winter

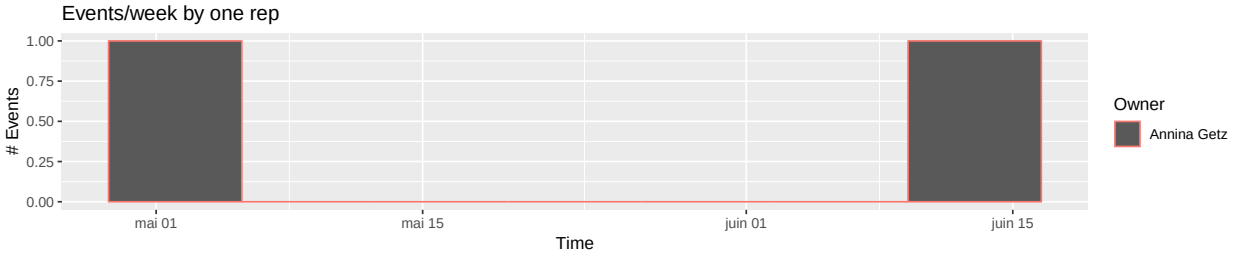

Events/month by one rep
Events
1.00
0.75
0.50
0.25
0.00
oct.
janv.
avr.
juil.
Time
Owner
Saskia Winter

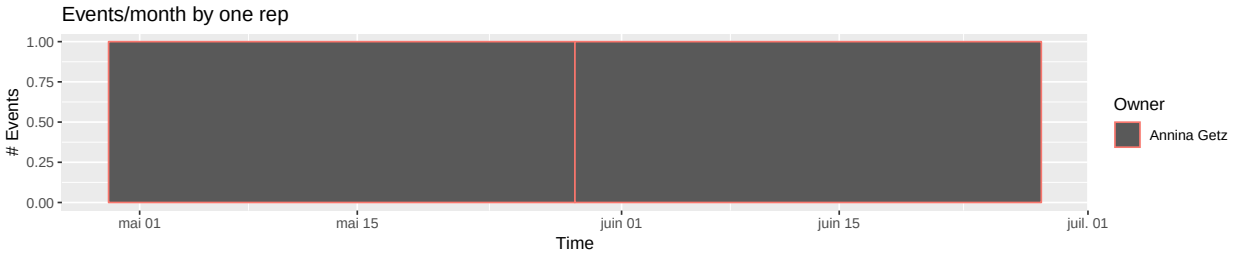

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
janv.
avr.
juil.
Time
Owner
Fabian Paul

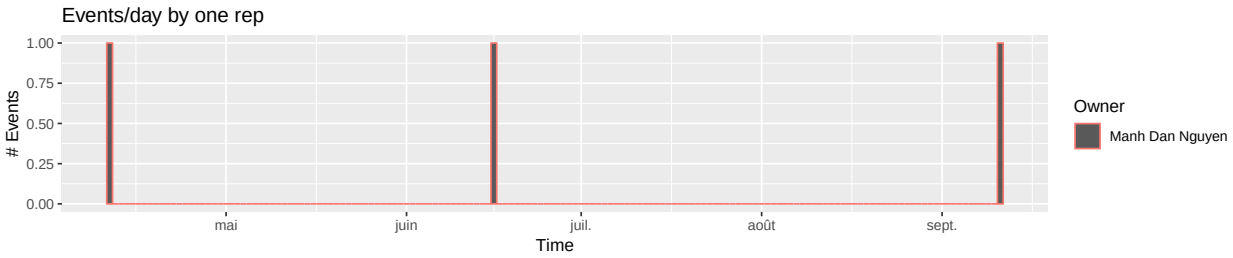

Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
janv.
avr.
juil.
Time
Owner
Fabian Paul

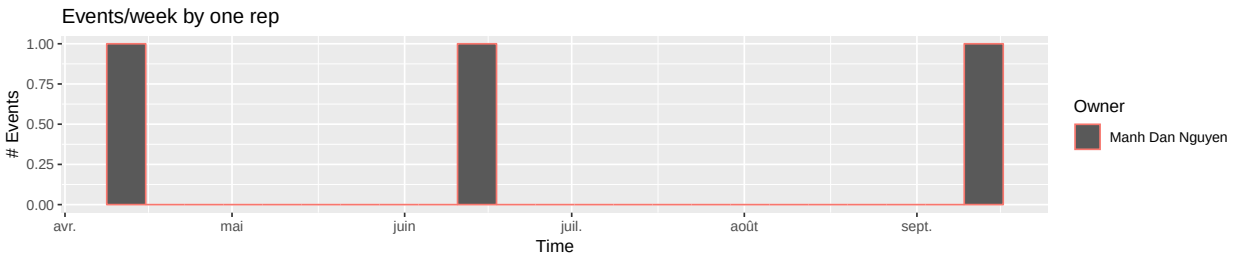

Events/month by one rep
Events
1.00
0.75
0.50
0.25
0.00
janv.
avr.
juil.
Time
Owner
Fabian Paul

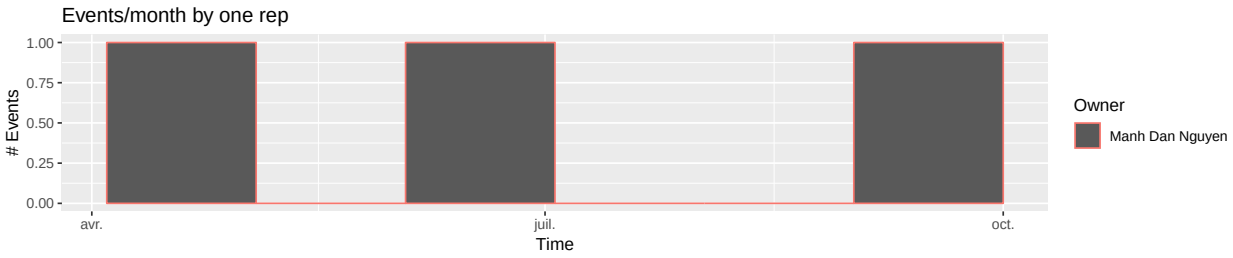

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
mai 01
mai 15
juin 01
juin 15
Time
Owner
Annina Getz

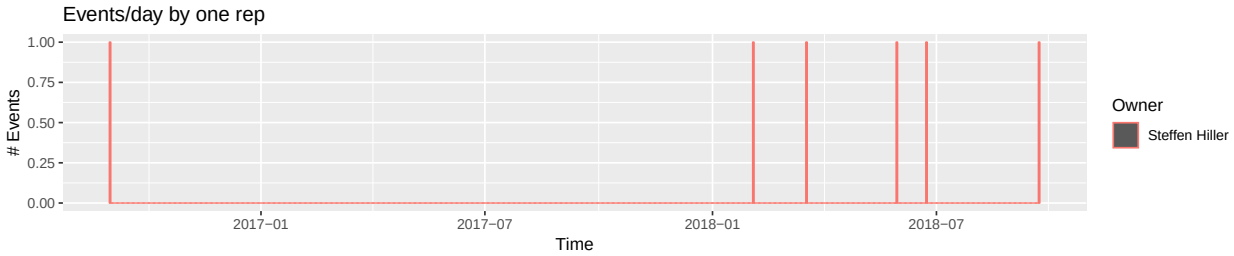

Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
mai 01
mai 15
juin 01
juin 15
Time
Owner
Annina Getz

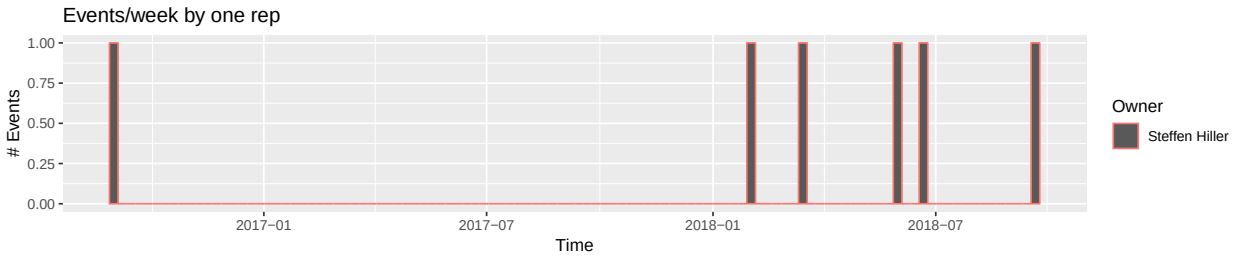

Events/month by one rep
Events
1.00
0.75
0.50
0.25
0.00
mai 01
mai 15
juin 01
juin 15
juil. 01
Time
Owner
Annina Getz

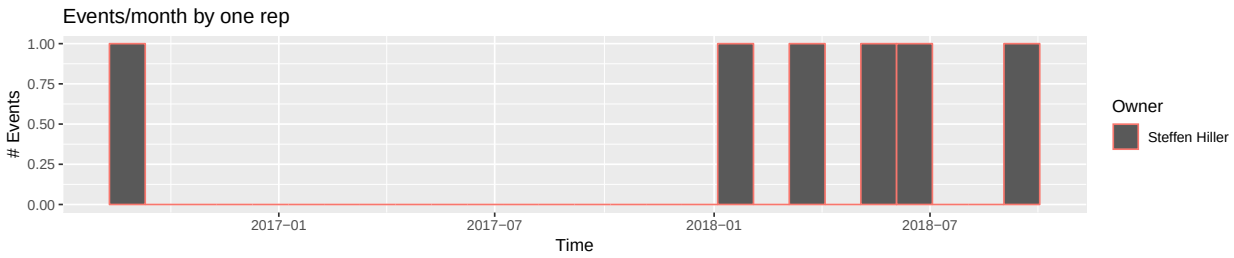

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
mai
juin
juil.
août
sept.
Time
Owner
Manh Dan Nguyen

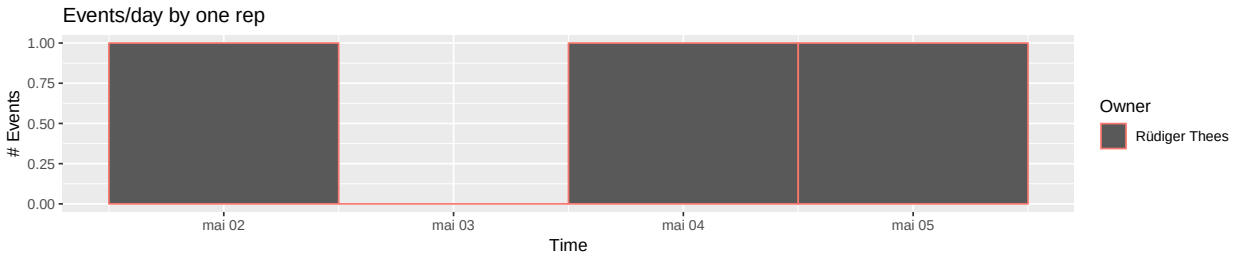

Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
avr.
mai
juin
juil.
août
sept.
Time
Owner
Manh Dan Nguyen

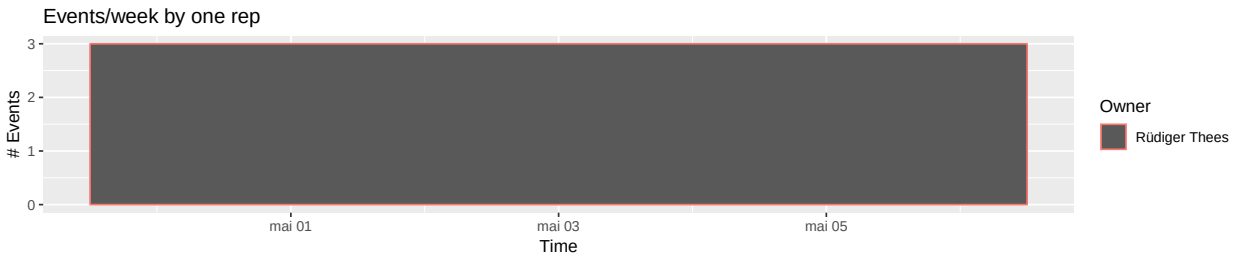

Events/month by one rep
Events
1.00
0.75
0.50
0.25
0.00
avr.
juil.
oct.
Time
Owner
Manh Dan Nguyen

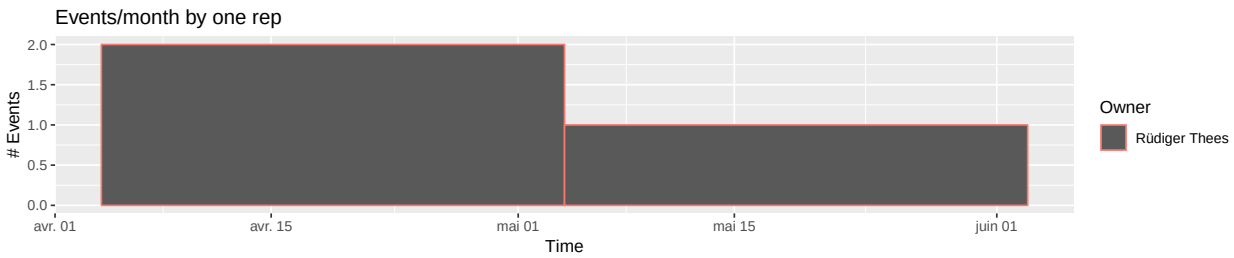

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017−01
2017−07
2018−01
2018−07
Time
Owner
Steffen Hiller

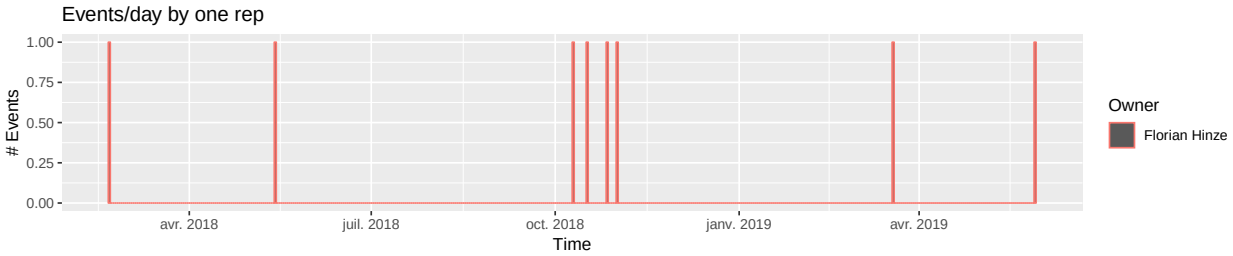

Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017−01
2017−07
2018−01
2018−07
Time
Owner
Steffen Hiller

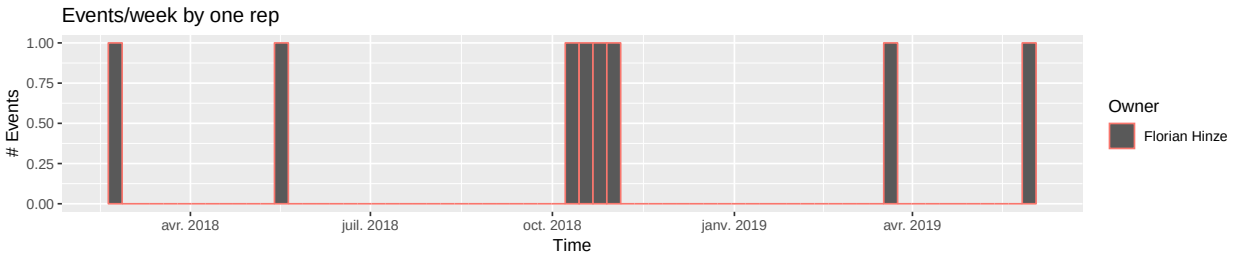

Events/month by one rep
Events
1.00
0.75
0.50
0.25
0.00
2017−01
2017−07
2018−01
2018−07
Time
Owner
Steffen Hiller

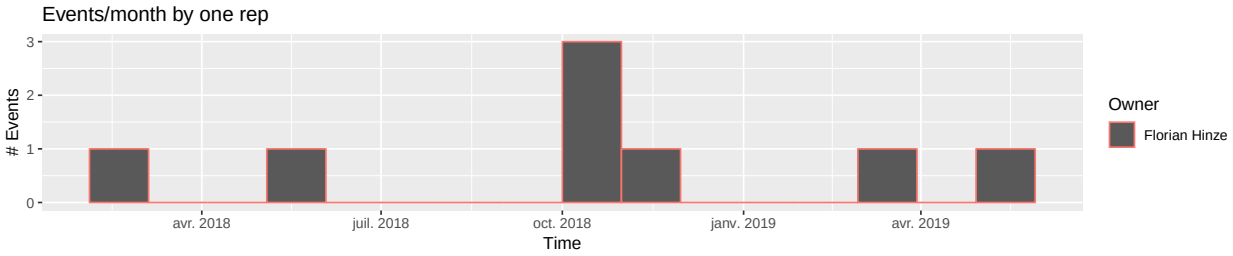

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
mai 02
mai 03
mai 04
mai 05
Time
Owner
Rüdiger Thees

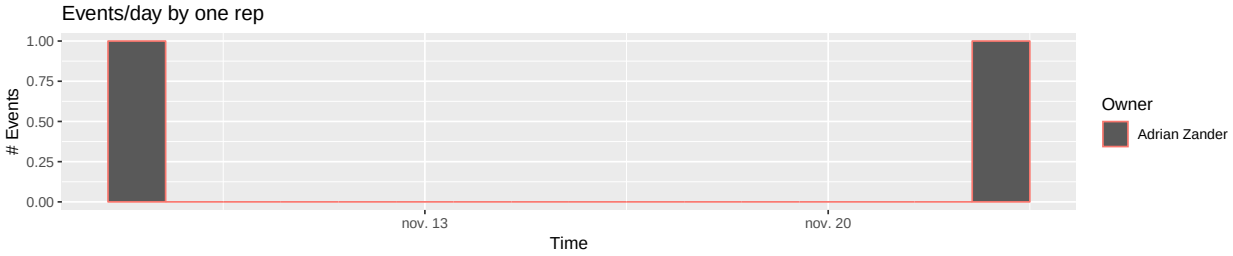

Events/week by one rep
Events
3
2
1
0
mai 01
mai 03
mai 05
Time
Owner
Rüdiger Thees

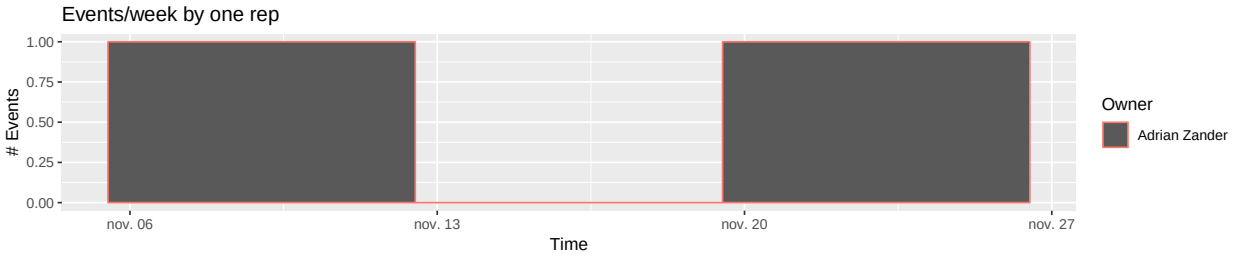

Events/month by one rep
Events
2.0
1.5
1.0
0.5
0.0
avr. 01
avr. 15
mai 01
mai 15
juin 01
Time
Owner
Rüdiger Thees


Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
avr. 2018
juil. 2018
oct. 2018
janv. 2019
avr. 2019
Time
Owner
Florian Hinze

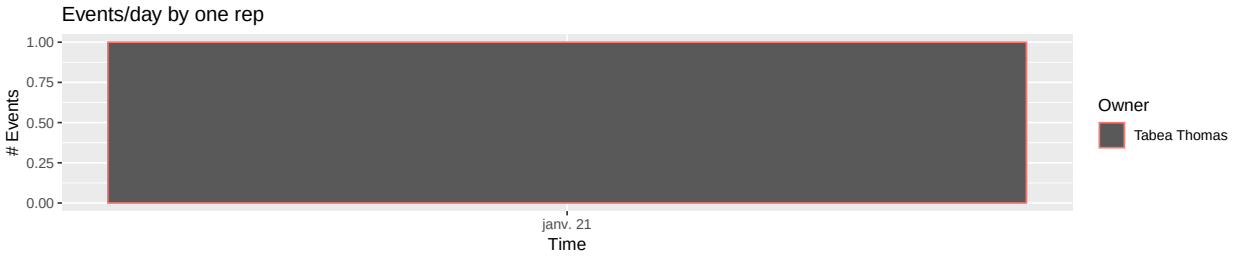

Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
avr. 2018
juil. 2018
oct. 2018
janv. 2019
avr. 2019
Time
Owner
Florian Hinze


Events/month by one rep
Events
3
2
1
0
avr. 2018
juil. 2018
oct. 2018
janv. 2019
avr. 2019
Time
Owner
Florian Hinze

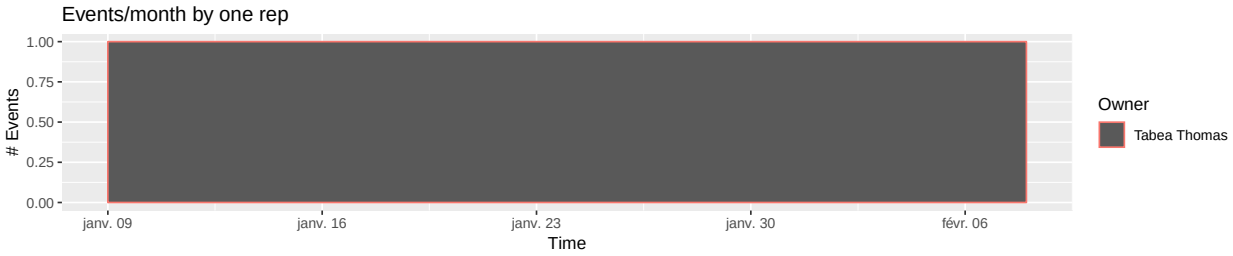

Events/day by one rep
Events
1.00
0.75
0.50
0.25
0.00
nov. 13
nov. 20
Time
Owner
Adrian Zander

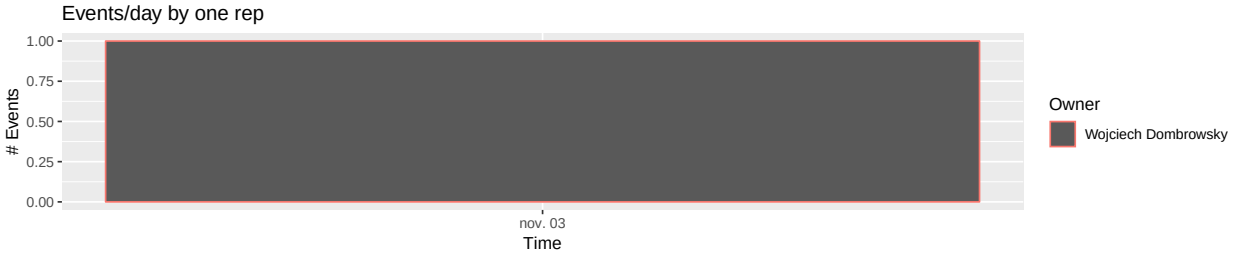

Events/week by one rep
Events
1.00
0.75
0.50
0.25
0.00
nov. 06
nov. 13
nov. 20
nov. 27
Time
Owner
Adrian Zander

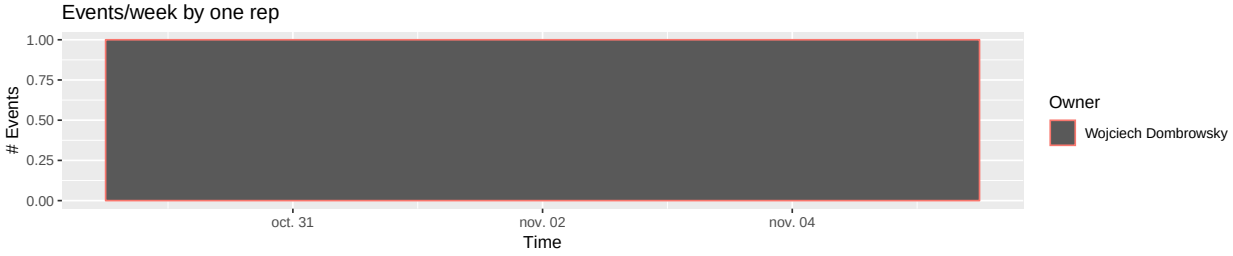

Events/month by one rep
Events
2.0
1.5
1.0
0.5
0.0
nov. 06
nov. 13
nov. 20
nov. 27
déc. 04
Time
Owner
Adrian Zander

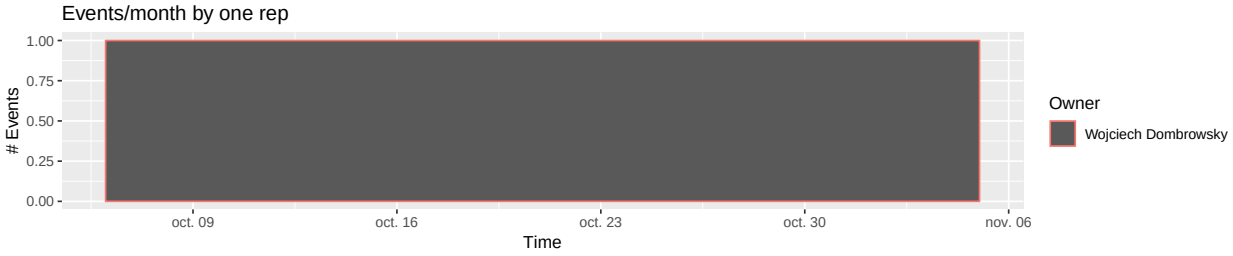

Events/day by one rep
1.00
0.75
0.50
0.25
0.00
Events
janv. 21
Time
Owner
Tabea Thomas

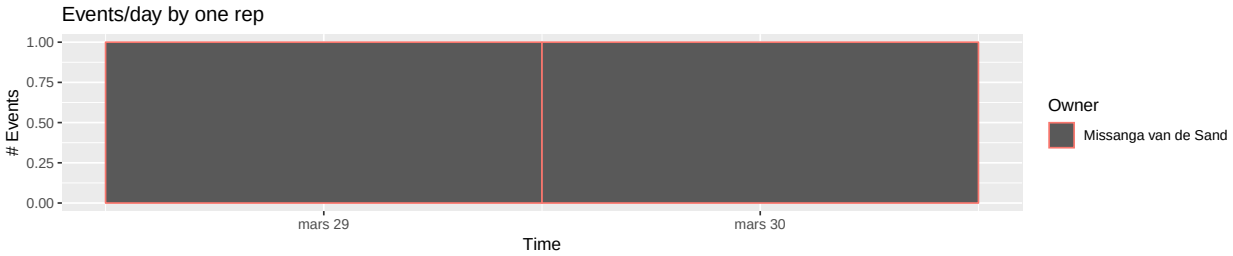

Events/week by one rep
1.00
0.75
0.50
0.25
0.00
Events
janv. 17
janv. 19
janv. 21
Time
Owner
Tabea Thomas

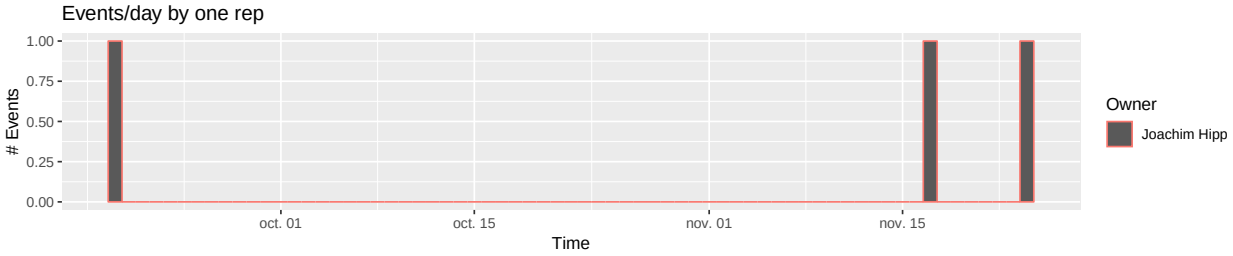

Events/month by one rep
1.00
0.75
0.50
0.25
0.00
Events
janv. 09
janv. 16
janv. 23
janv. 30
févr. 06
Time
Owner
Tabea Thomas

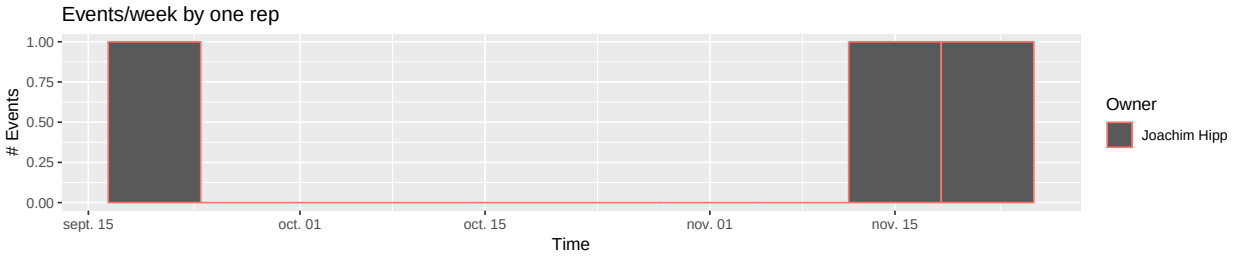

Events/day by one rep
1.00
0.75
0.50
0.25
0.00
Events
nov. 03
Time
Owner
Wojciech Dombrowsky

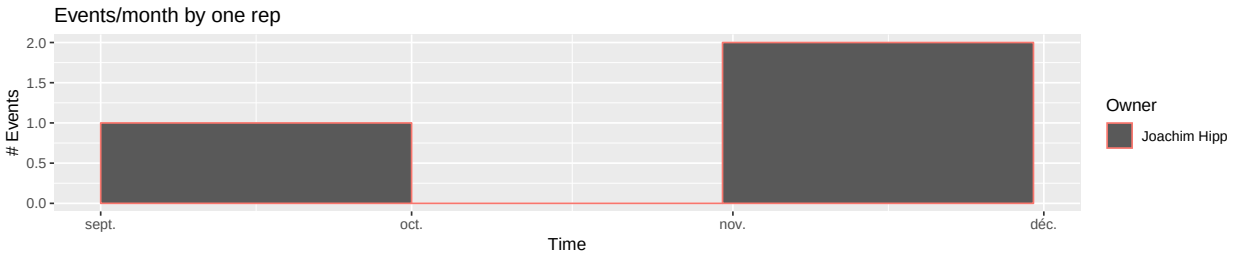

Events/week by one rep
1.00
0.75
0.50
0.25
0.00
Events
oct. 31
nov. 02
nov. 04
Time
Owner
Wojciech Dombrowsky


Events/month by one rep
1.00
0.75
0.50
0.25
0.00
Events
oct. 09
oct. 16
oct. 23
oct. 30
nov. 06
Time
Owner
Wojciech Dombrowsky





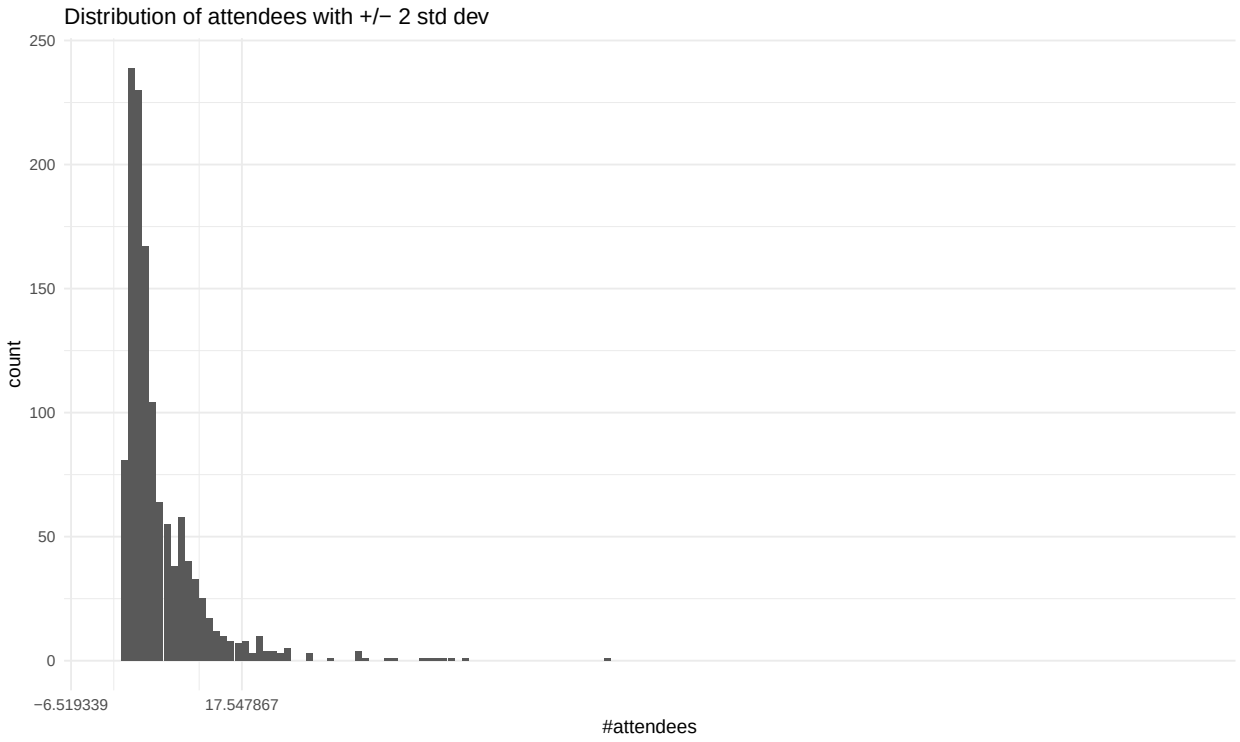




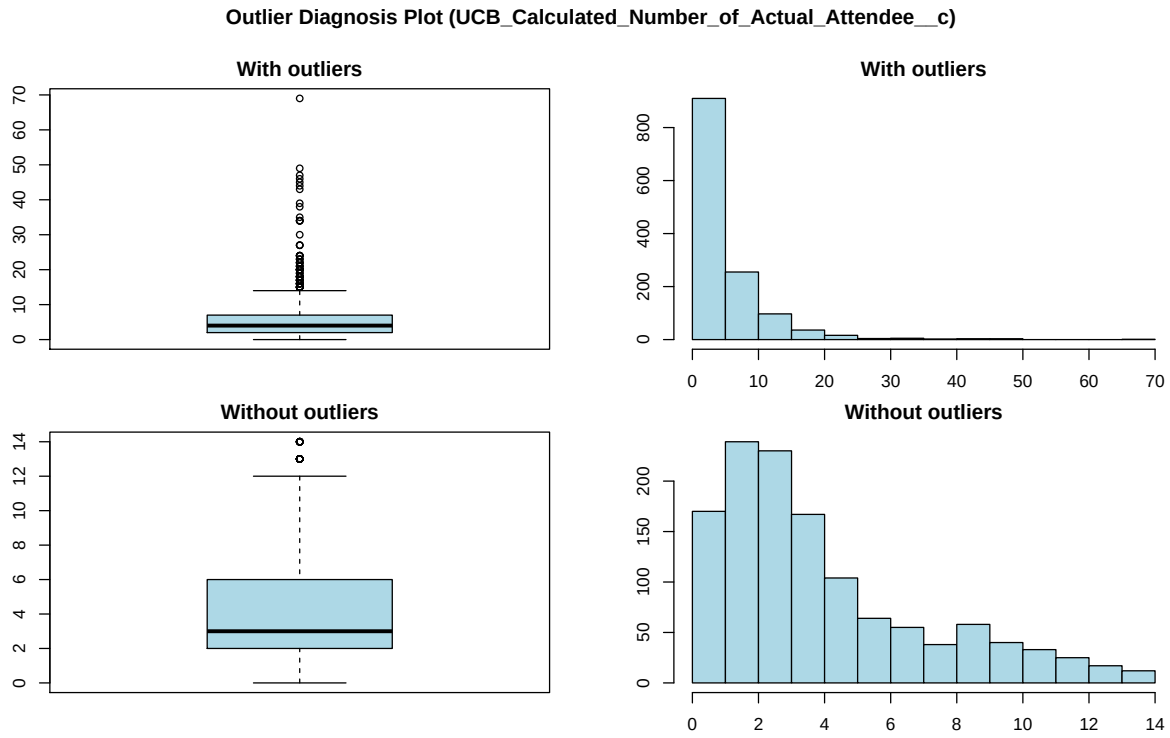
Distribution of attendees each event

```
[1] "mean:  5.51426426426426 std dev:  6.01680148079842"
```

```
pdf
  2
```

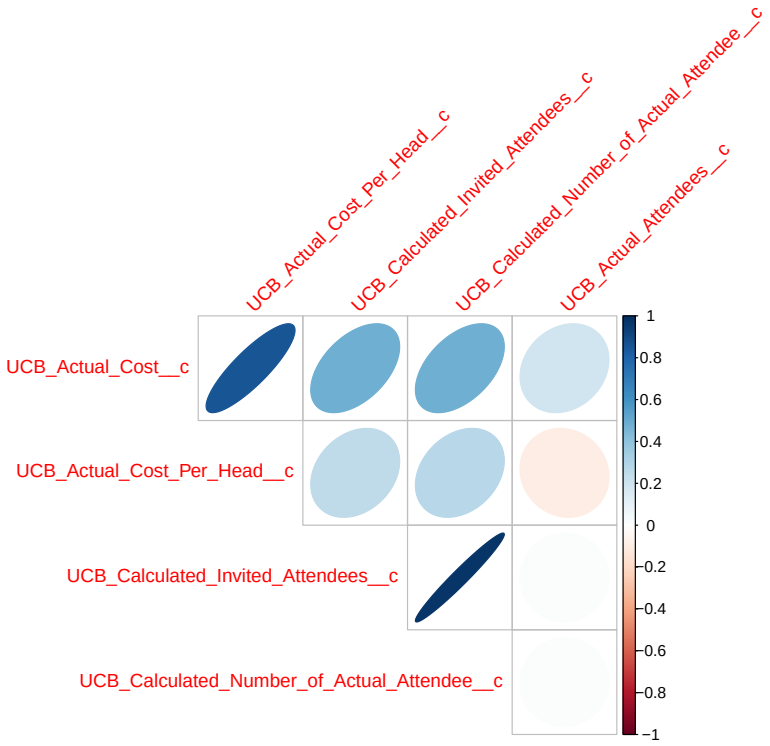

Distribution of attendees with +/− 2 std dev
count
250
200
150
100
50
0
−6.519339
17.547867
#attendees

```
pdf
  2
```


Outlier Diagnosis Plot (UCB_Calculated_Number_of_Actual_Attendee__c)
With outliers
With outliers
Without outliers
Without outliers
0 10 20 30 40 50 60 70
0 200 400 600 800
0 2 4 6 8 10 12 14
0 50 100 150 200

Bivariate data EDA. Calculation of `correlation coefficient` using `correlate()`

Visualization of the correlation matrix using `plot_correlate()` `plot_correlate()` visualizes the correlation matrix.



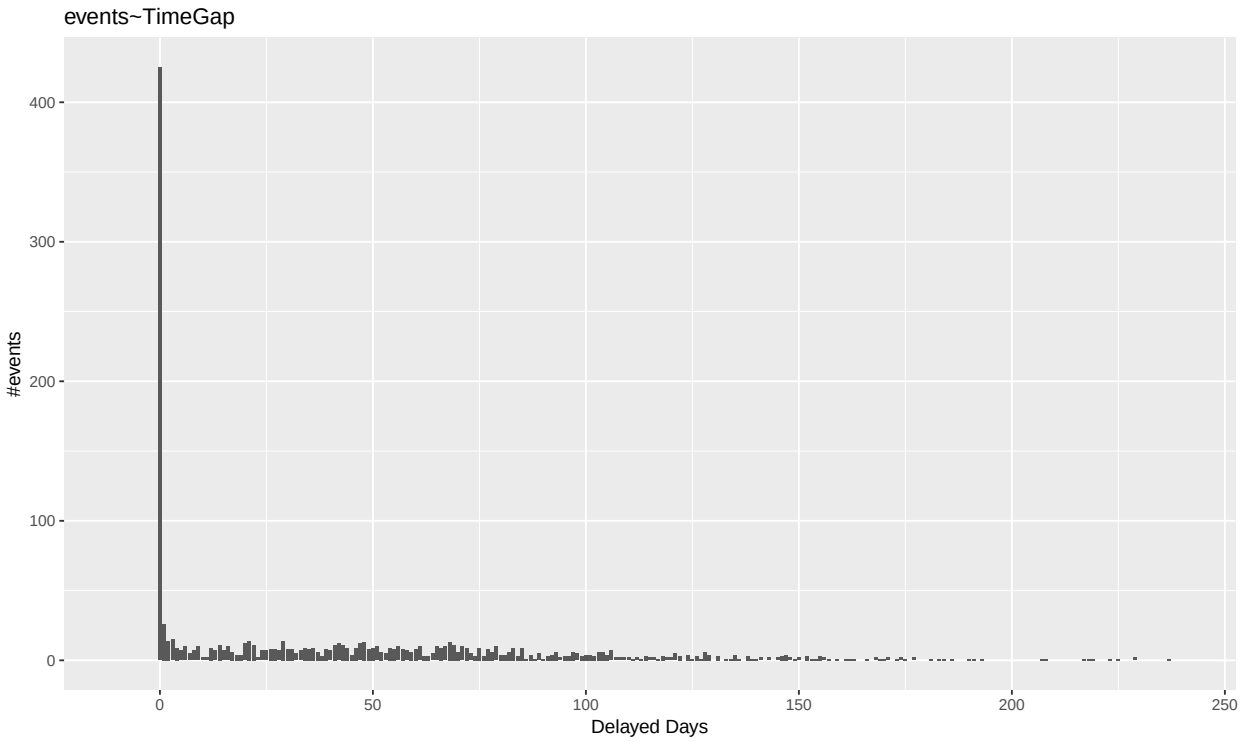
5. DQA Dimension5/Rule5: undestanding consistency of data from more than one source

Need other data sources for comparsion

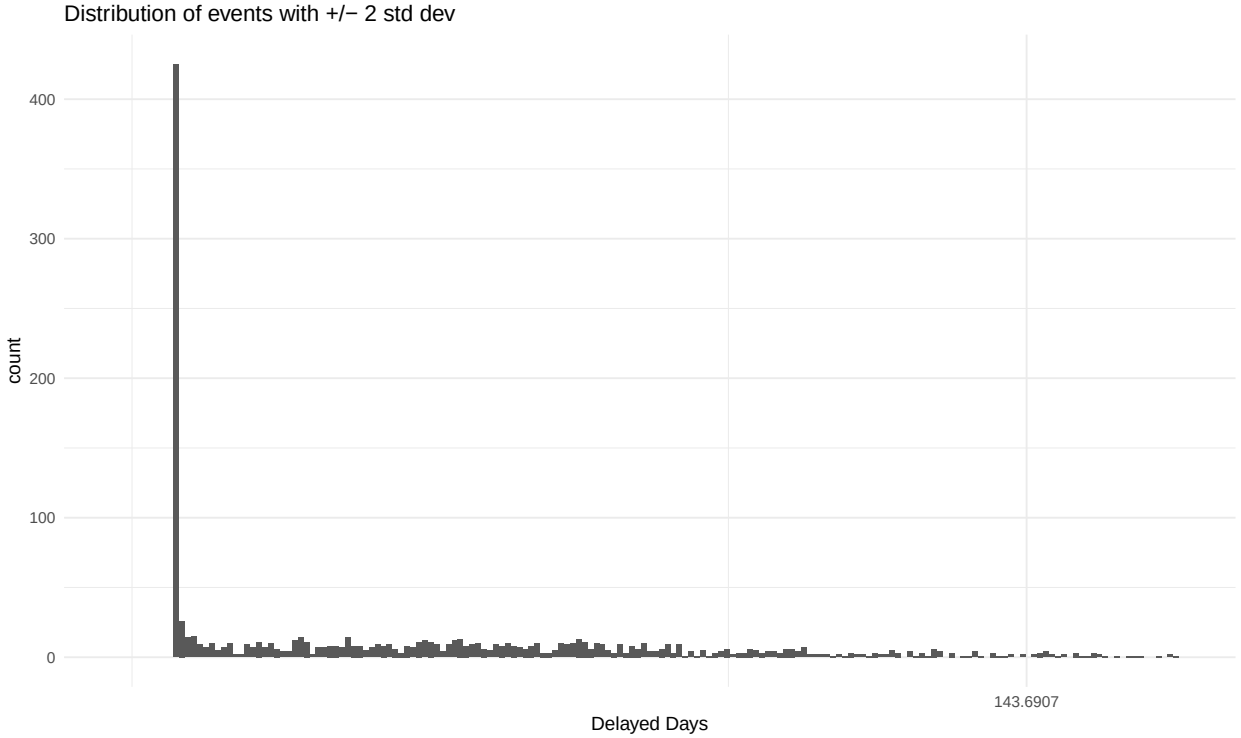
6. DQA Dimension6/Rule6: undestanding timeline, trend & seasonality

The degree to which data represent reality from the required point in time Time difference: Ind_6= max[(threshold-xx days)/threshold,0]

```
pdf
  2
```

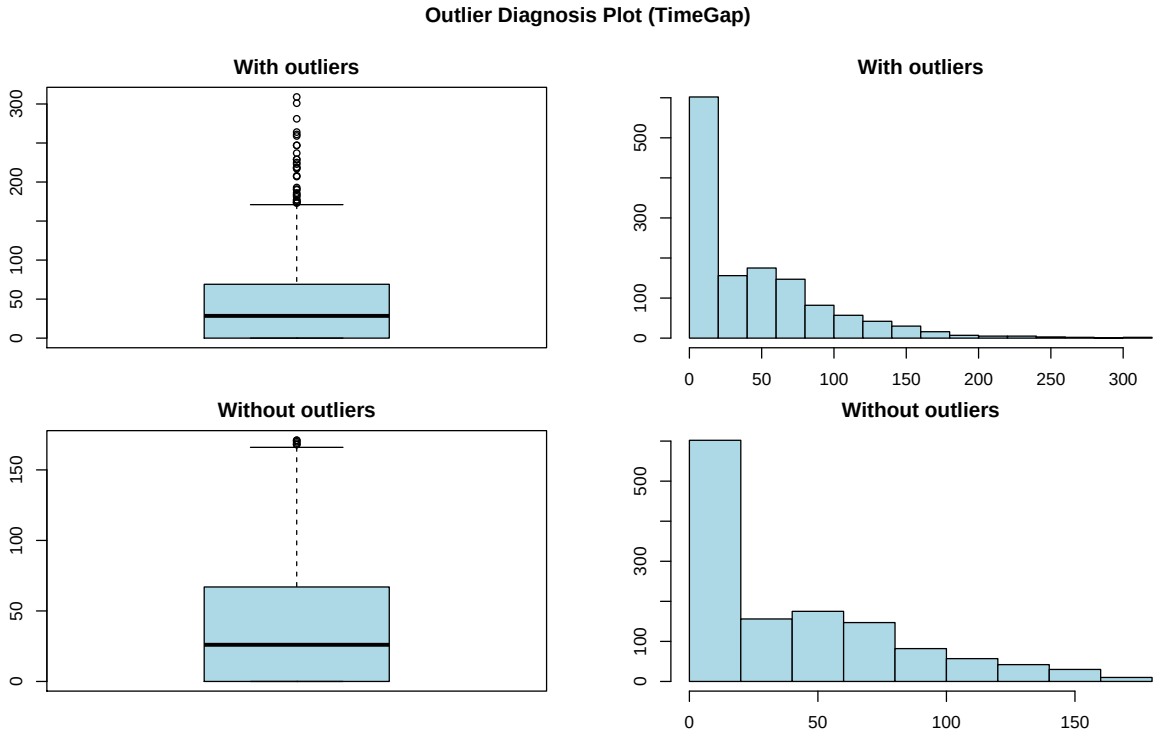



```
[1] "mean:  42.8843843843844 std dev:  50.4031544932102"
pdf
  2
```




```
pdf
```

2


Outlier Diagnosis Plot (TimeGap)
With outliers
0
50
100
200
300
With outliers
0
100
300
500
0
50
100
150
200
250
300
Without outliers
0
50
100
150
Without outliers
0
100
300
500
0
50
100
150