СОРЕВНОВАНИЕ DATA FUSION

| 15 ФЕВРАЛЯ – 5 АПРЕЛЯ 2024 |

Meetup #1

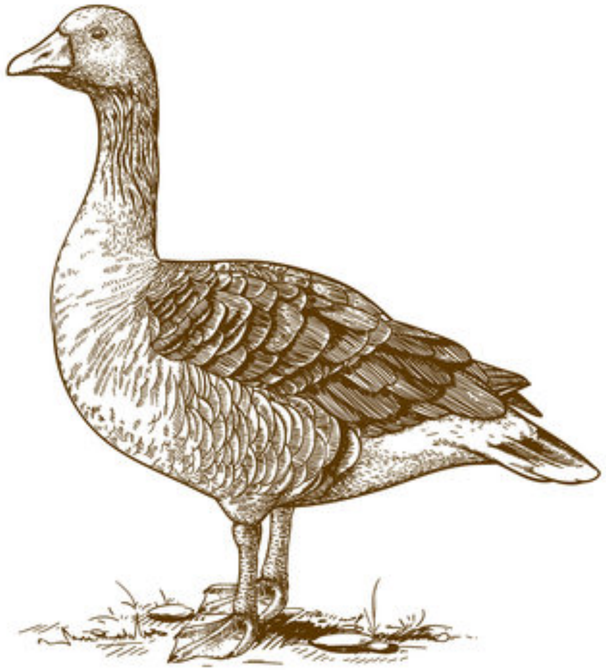
29.02.2024

Задача 2

Предсказание оттока

Часть 0

О таблетках и гусях



Часть 1

Что происходит в табличных данных?

Данные

Для решения задачи “Отток” предлагается несколько групп данных и материалов:

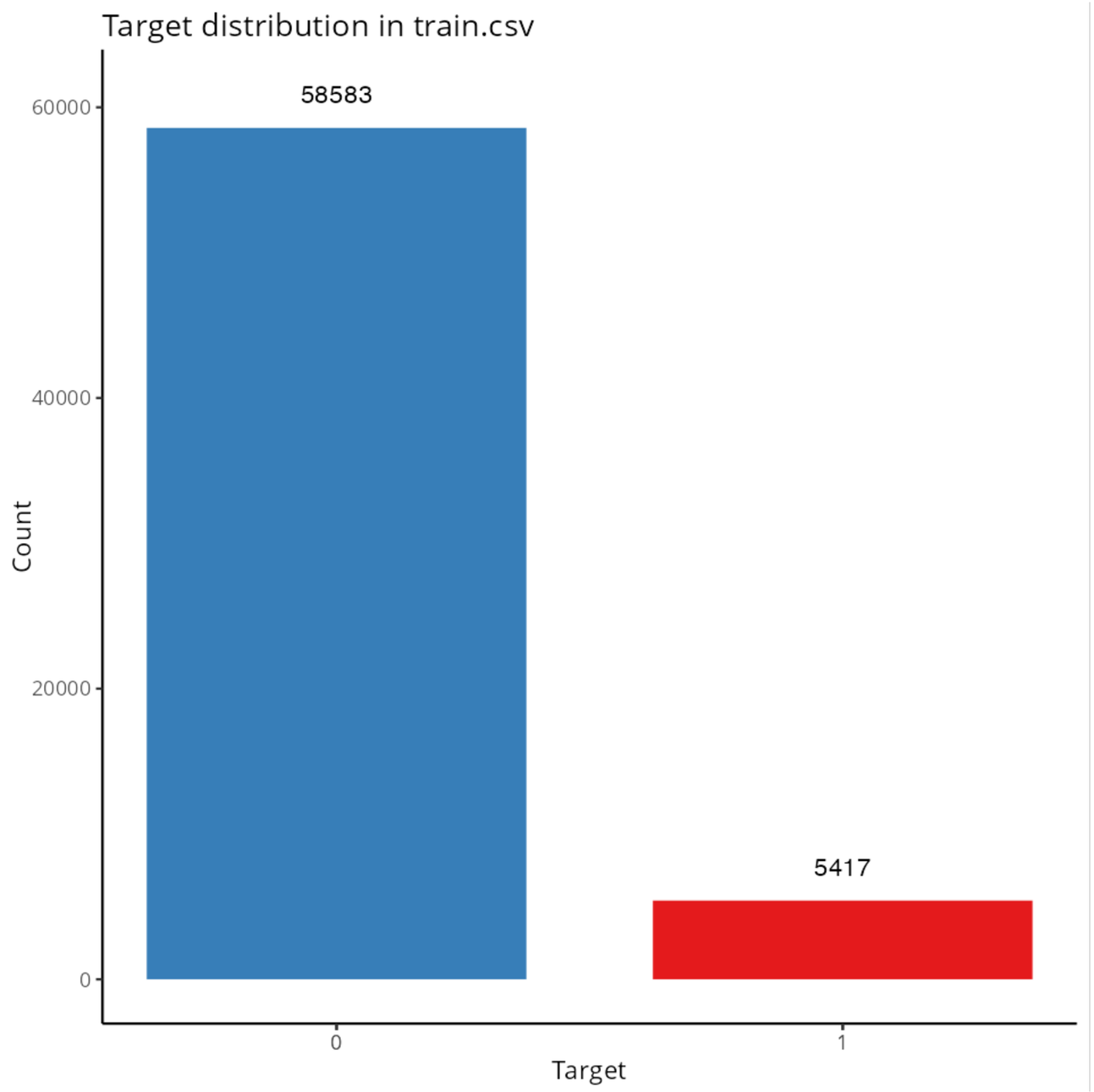
1. Табличные клиентские данные:

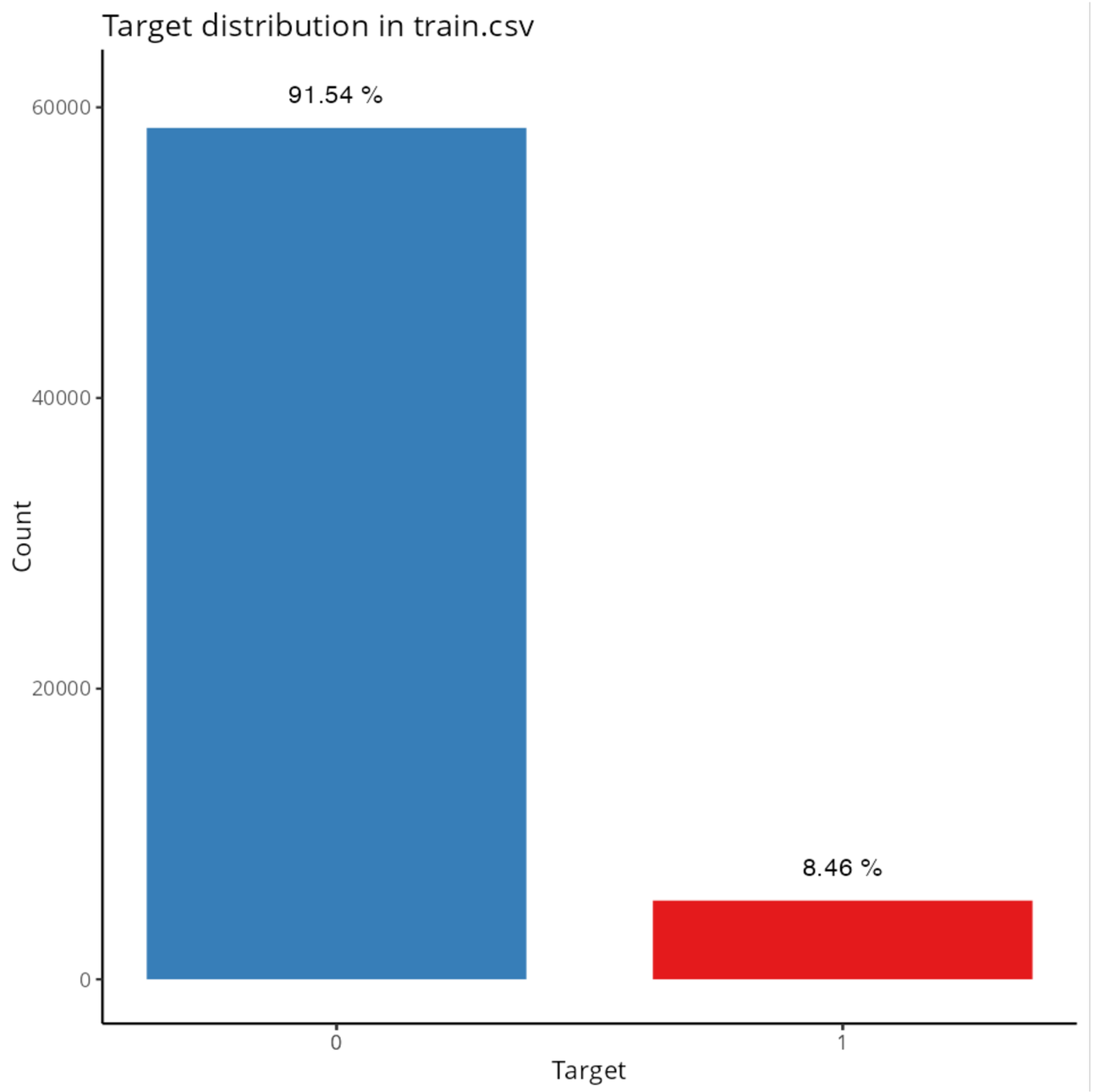
Участникам доступны несколько наборов данных и артефактов:

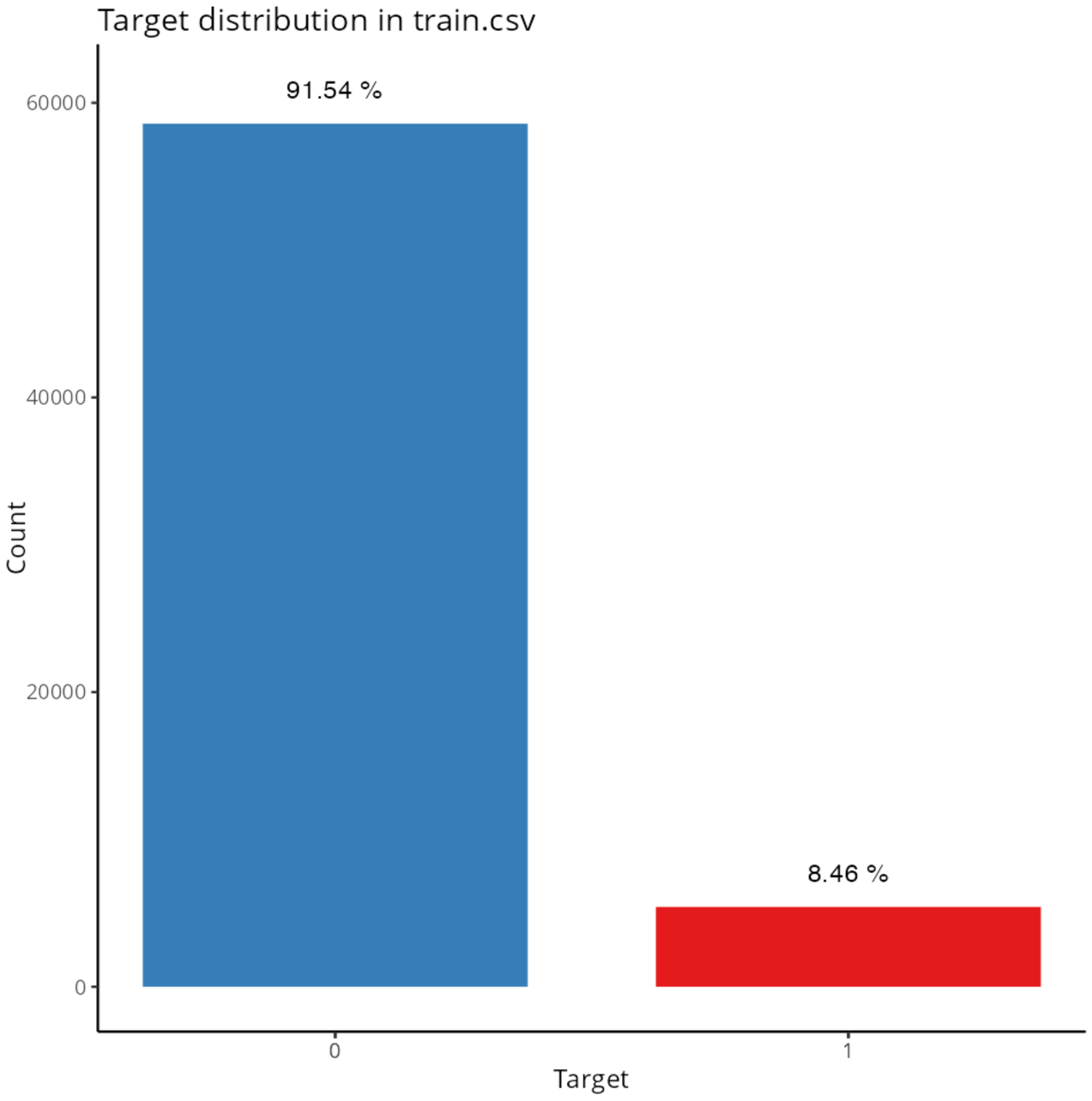
1. Базовая информация про все 96,000 клиентов в табличном `.csv` формате: `clients.csv` (**2.5 MB**)
2. Тренировочные данные по целевой переменной и времени последней транзакции для 64,000 клиентов: `train.csv` (**745 KB**)
3. Информация об отчетах, в рамках которых клиенты сгруппированы по времени: `report_dates.csv` (**288 KB**)

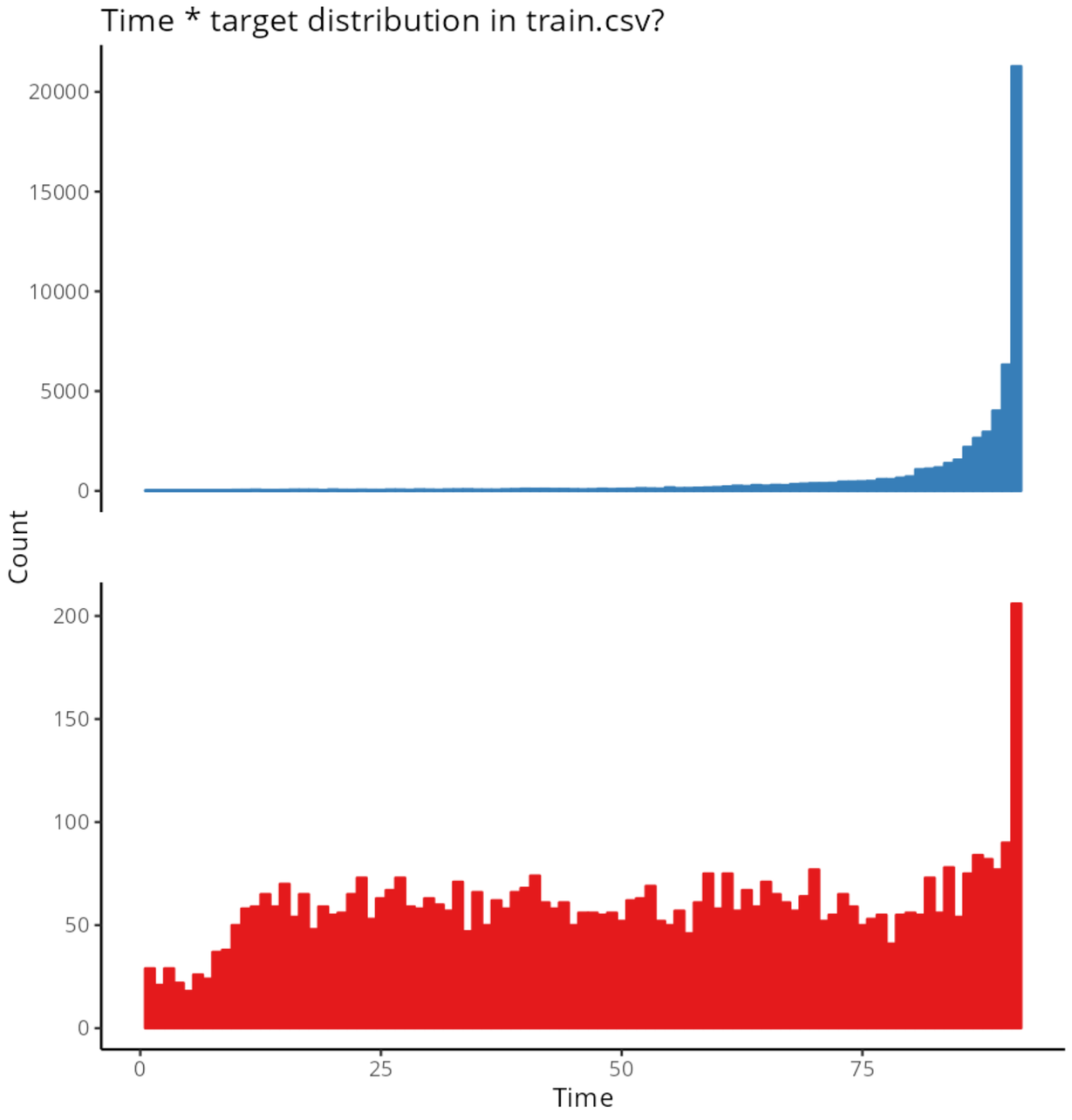
2. Данные клиентских транзакций:

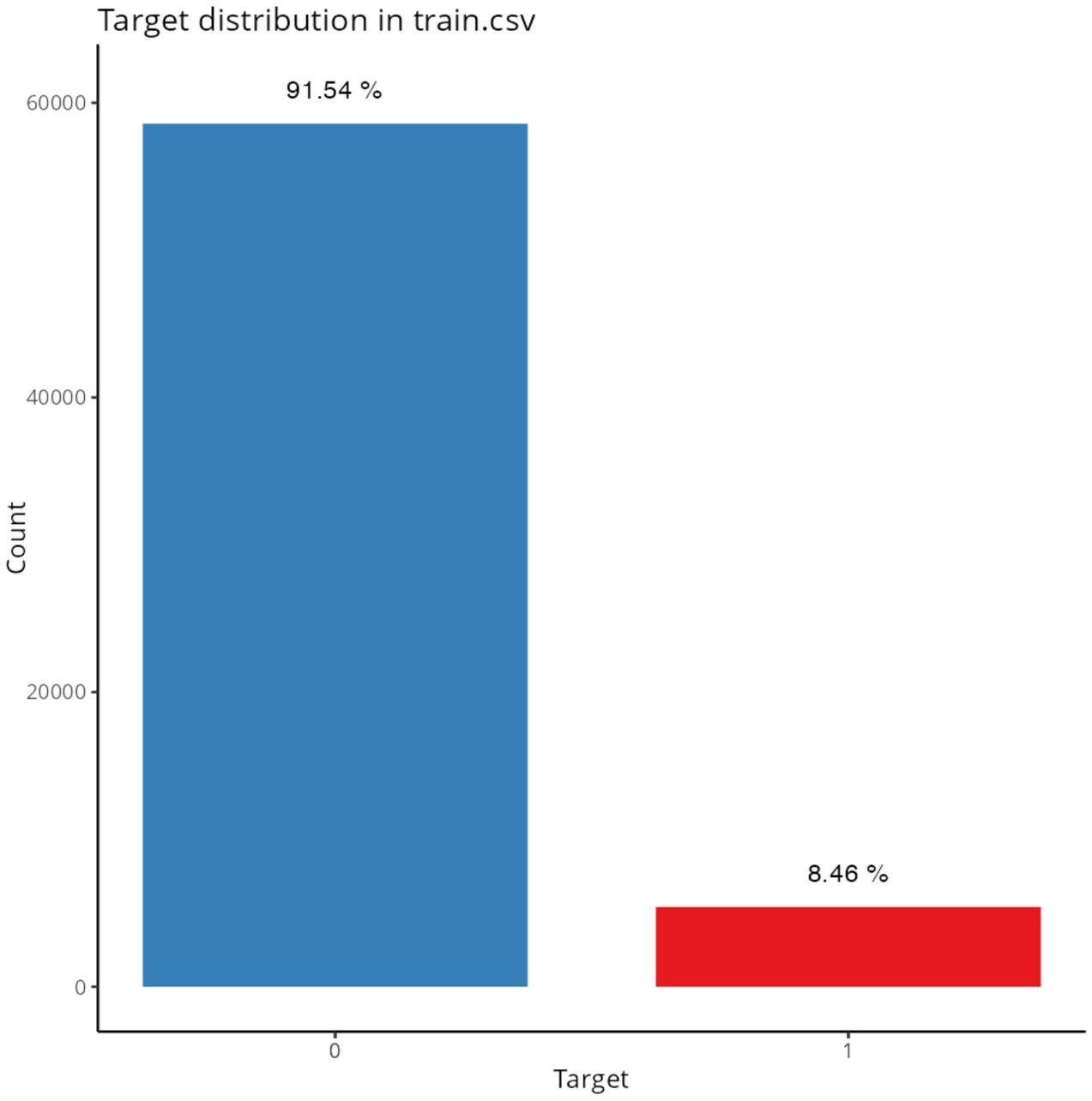
1. Клиентские транзакции для всех 96,000 клиентов (~13M) в табличном `.csv` формате: `transactions.csv.zip` (**197 MB**)

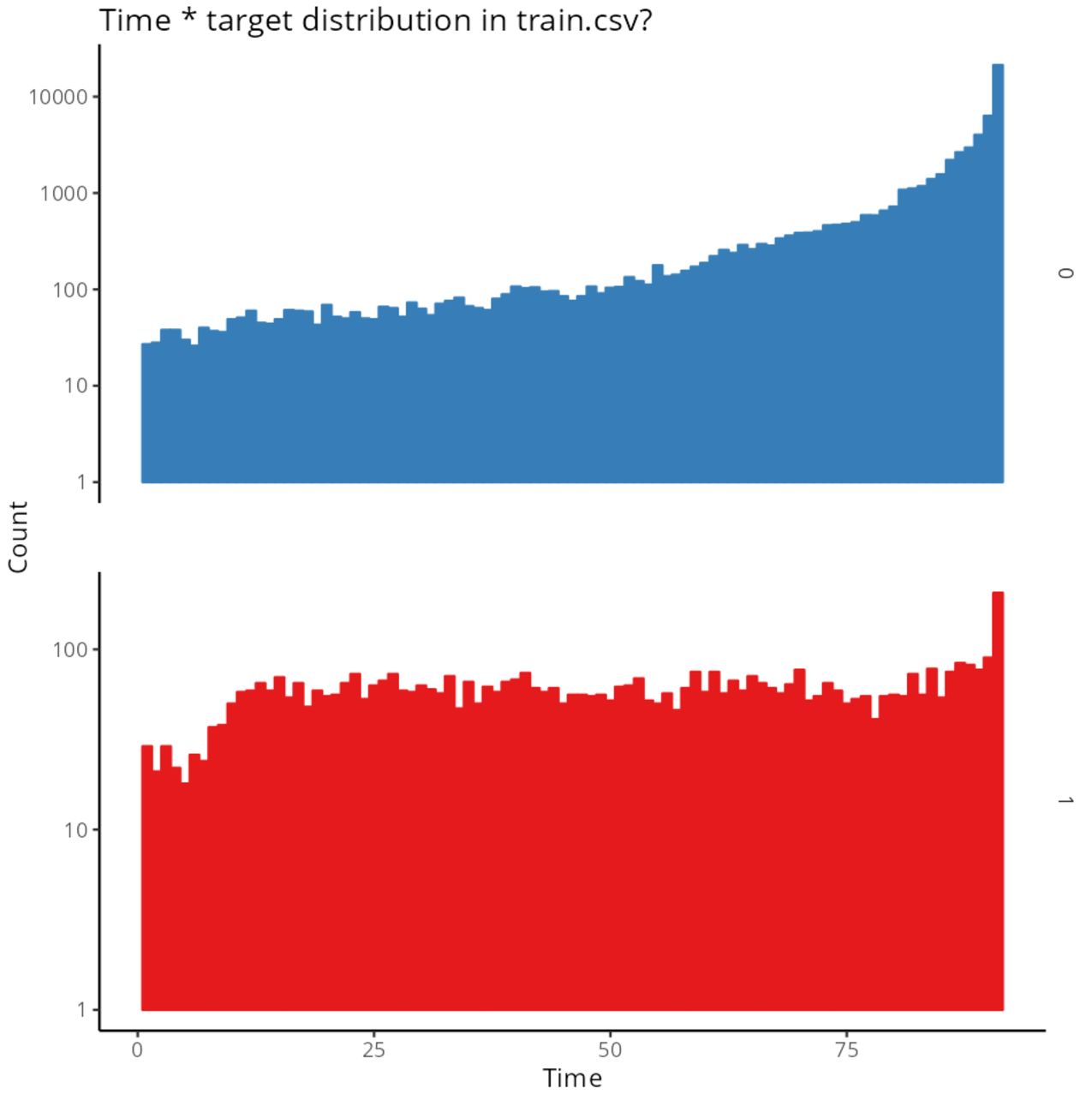

Target distribution in train.csv
58583
5417
60000
40000
20000
0
Count
0
1
Target

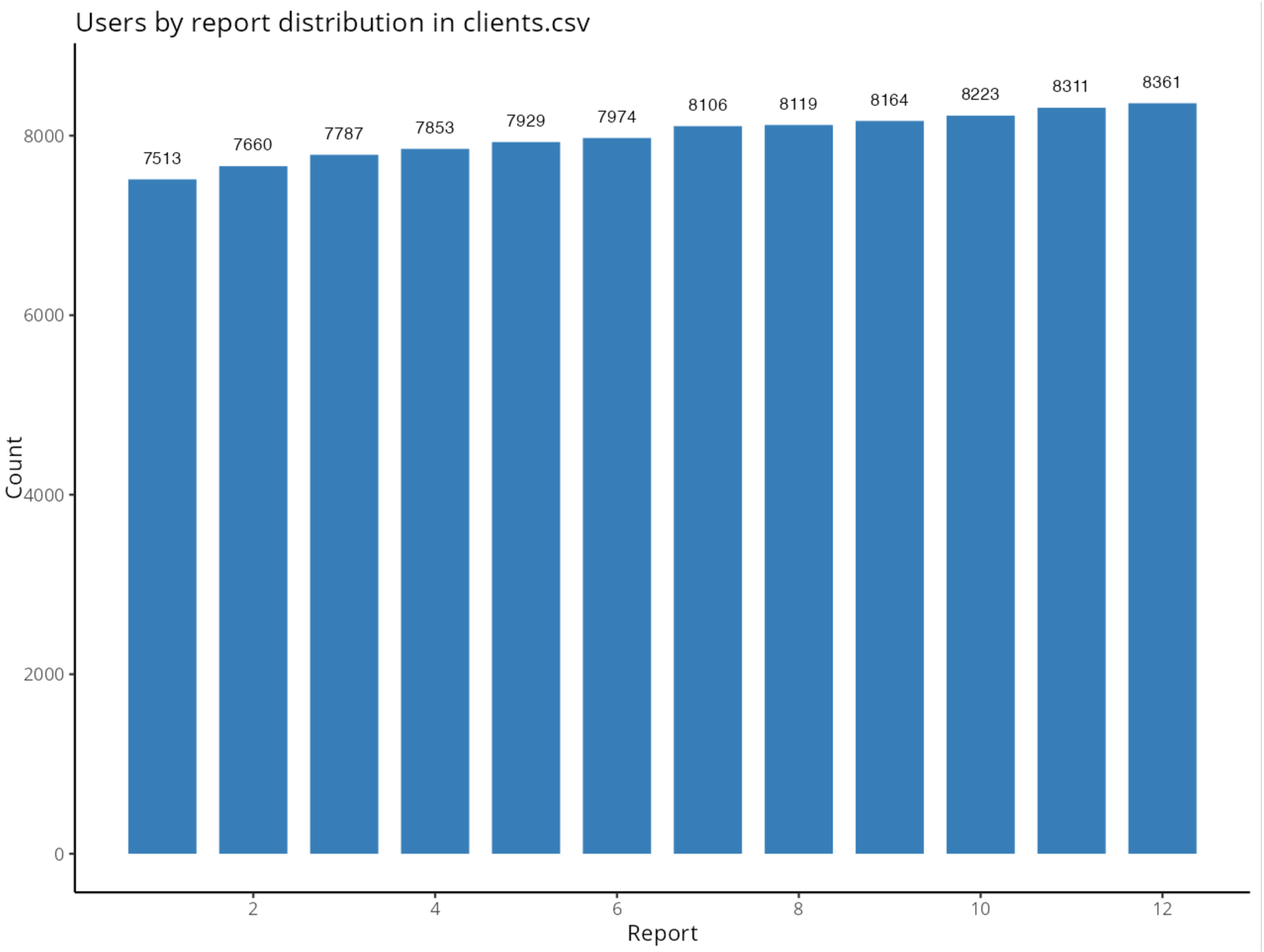

Target distribution in train.csv
91.54 %
8.46 %
60000
40000
20000
0
Count
0
1
Target

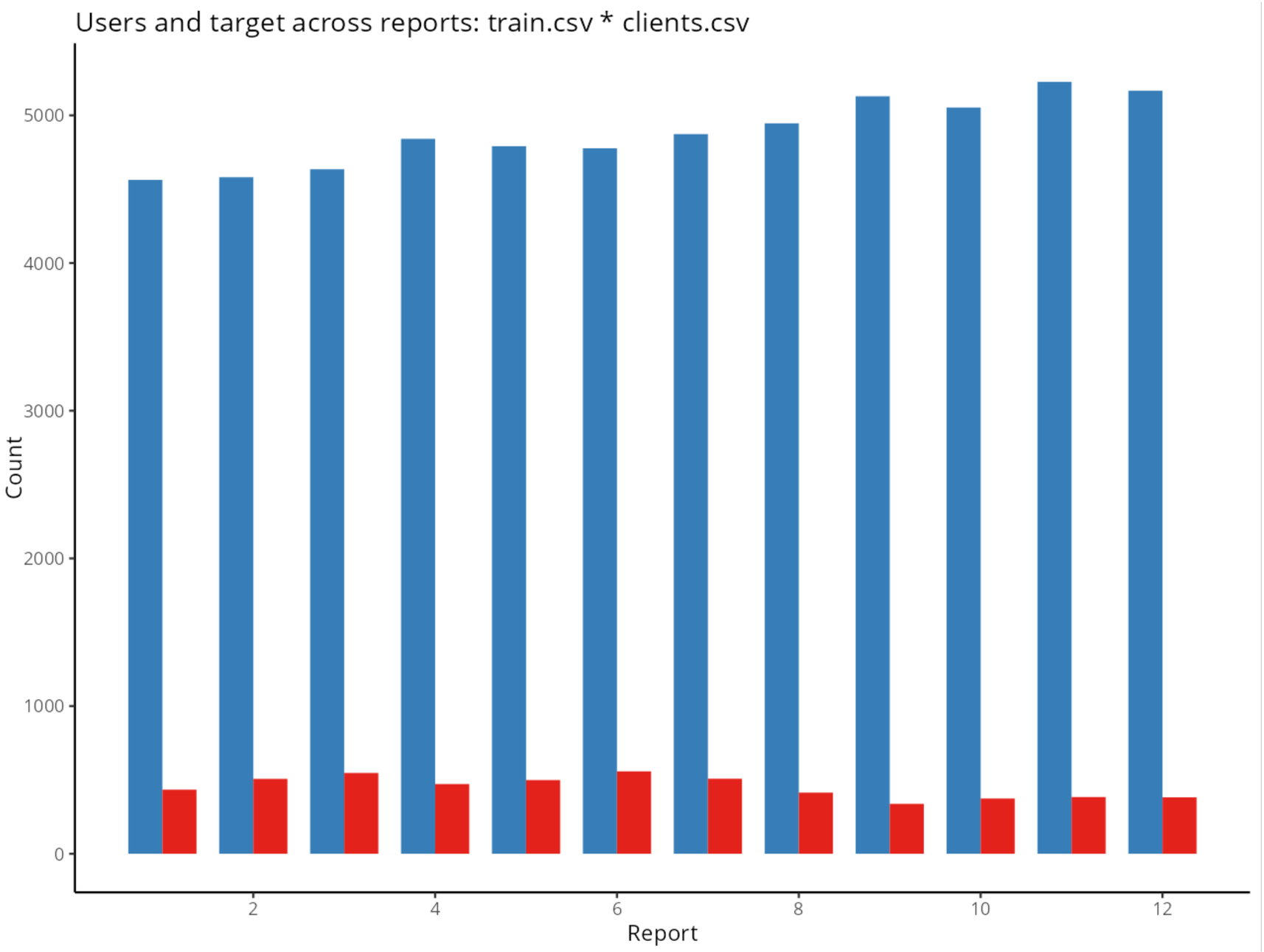

Target distribution in train.csv
91.54 %
8.46 %
60000
40000
20000
0
Count
0
1
Target


Time * target distribution in train.csv?
20000
15000
10000
5000
0
200
150
100
50
0
Count
0
25
50
75
Time
0
1


Target distribution in train.csv
91.54 %
8.46 %
60000
40000
20000
0
Count
0
1
Target


Time * target distribution in train.csv?
10000
1000
100
10
1
100
10
1
Count
0
1
0
25
50
75
Time


Users by report distribution in clients.csv
7513
7660
7787
7853
7929
7974
8106
8119
8164
8223
8311
8361
8000
6000
4000
2000
0
Count
2
4
6
8
10
12
Report


Users and target across reports: train.csv * clients.csv
5000
4000
3000
2000
1000
0
Count
2
4
6
8
10
12
Report

Часть 2

Что происходит в транзакционных данных?

Данные

Для решения задачи “Отток” предлагается несколько групп данных и материалов:

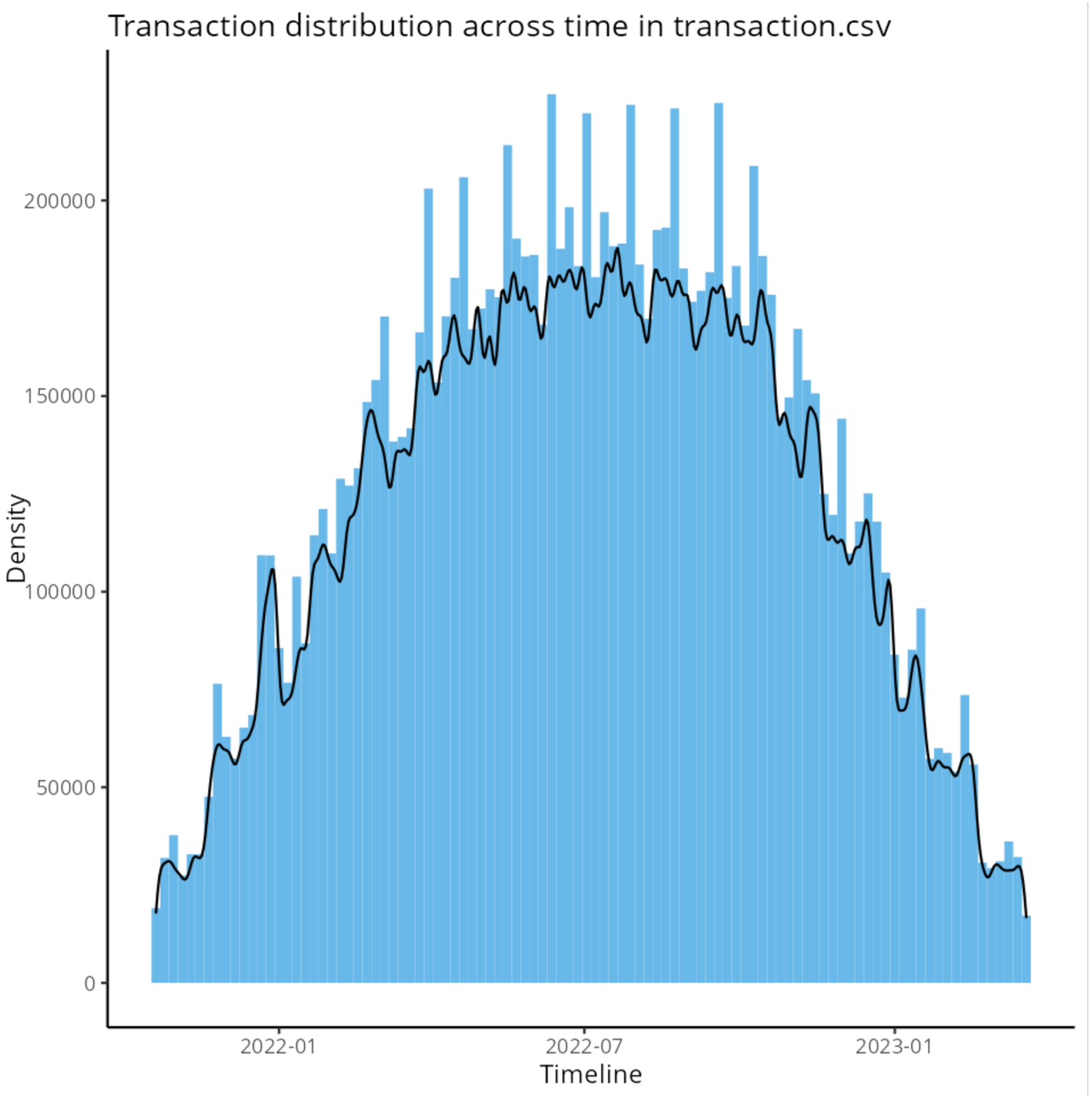
1. Табличные клиентские данные:

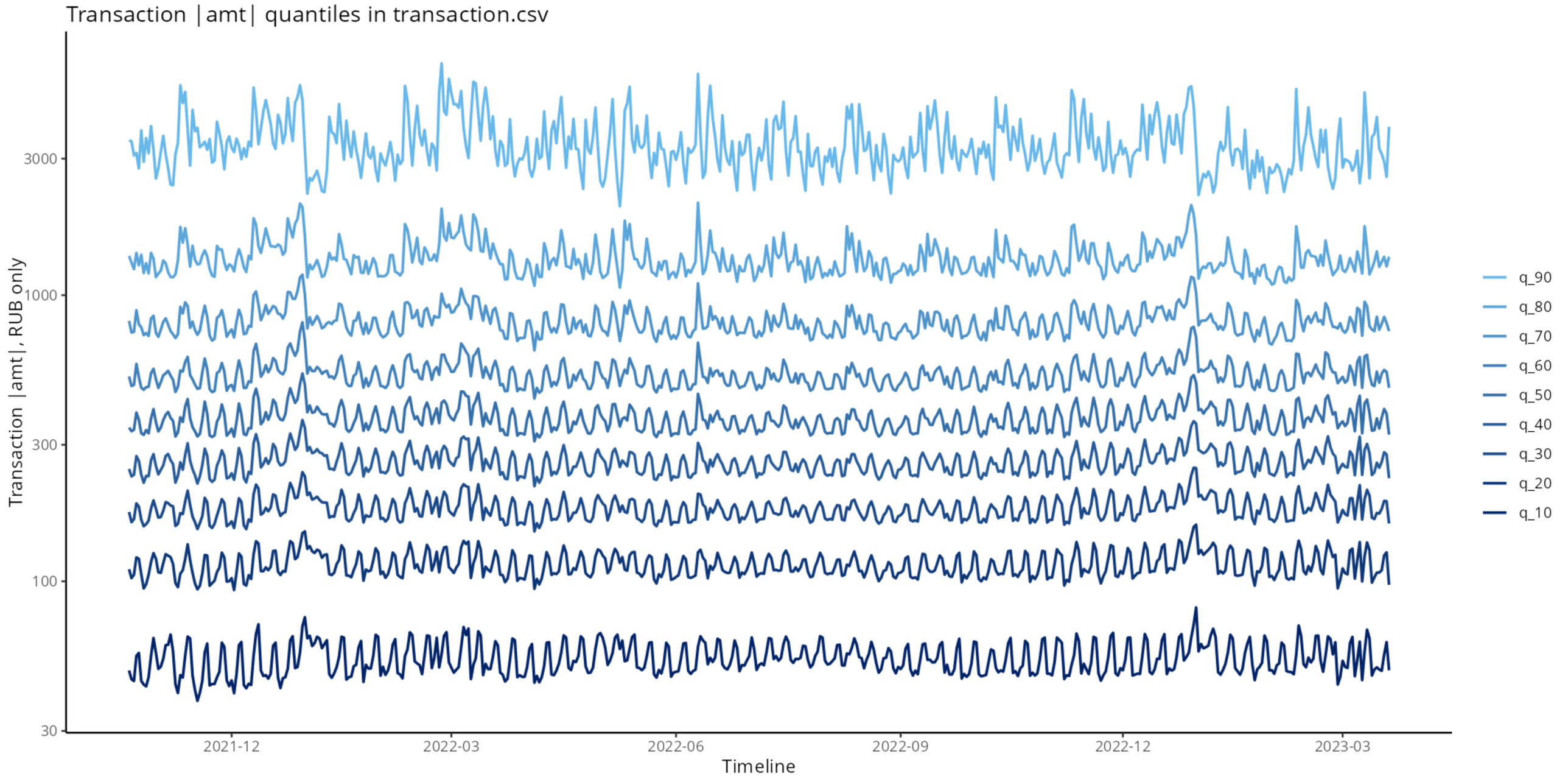
Участникам доступны несколько наборов данных и артефактов:

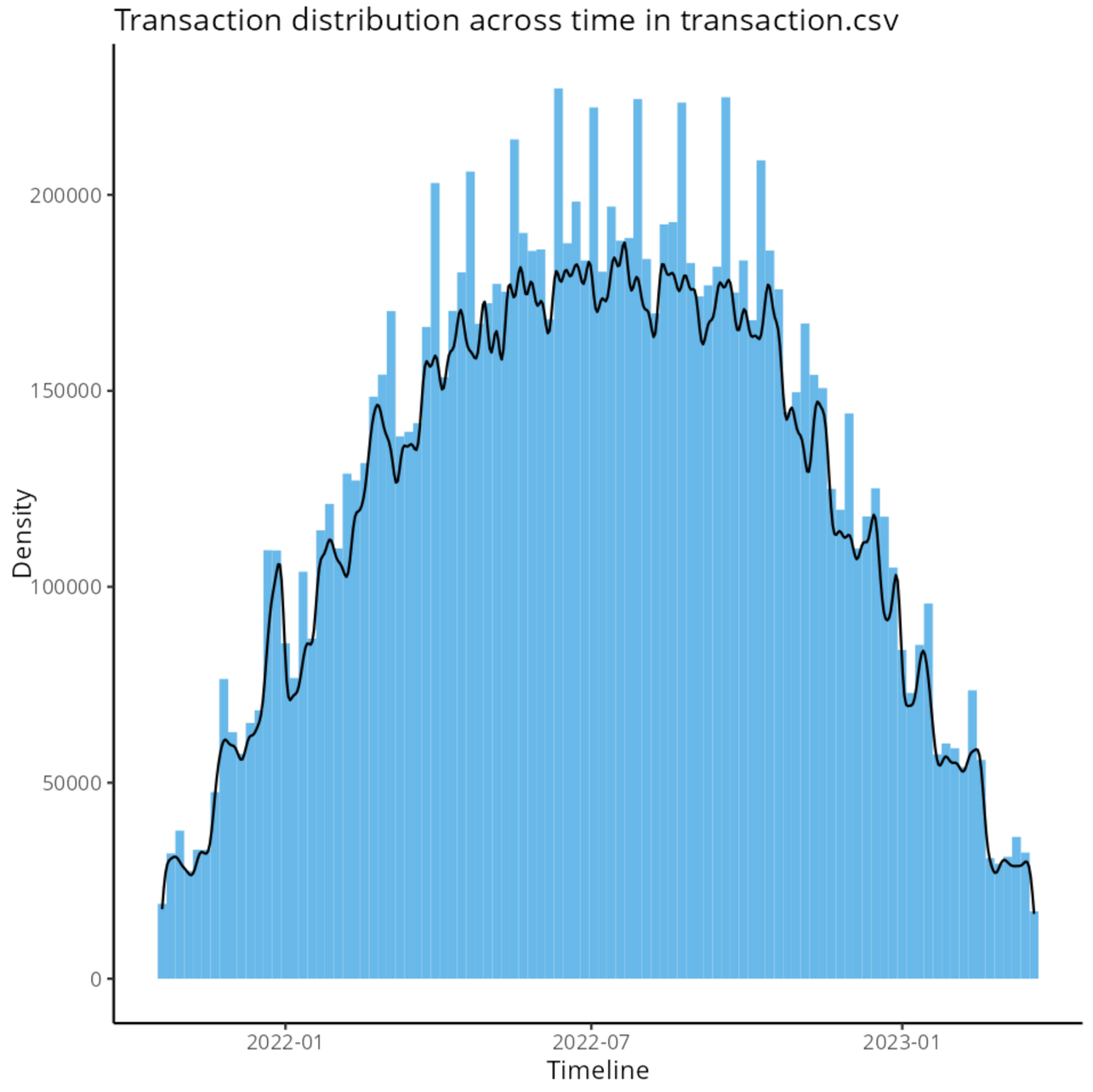
1. Базовая информация про все 96,000 клиентов в табличном `.csv` формате: `clients.csv` (**2.5 MB**)
2. Тренировочные данные по целевой переменной и времени последней транзакции для 64,000 клиентов: `train.csv` (**745 KB**)
3. Информация об отчетах, в рамках которых клиенты сгруппированы по времени: `report_dates.csv` (**288 KB**)

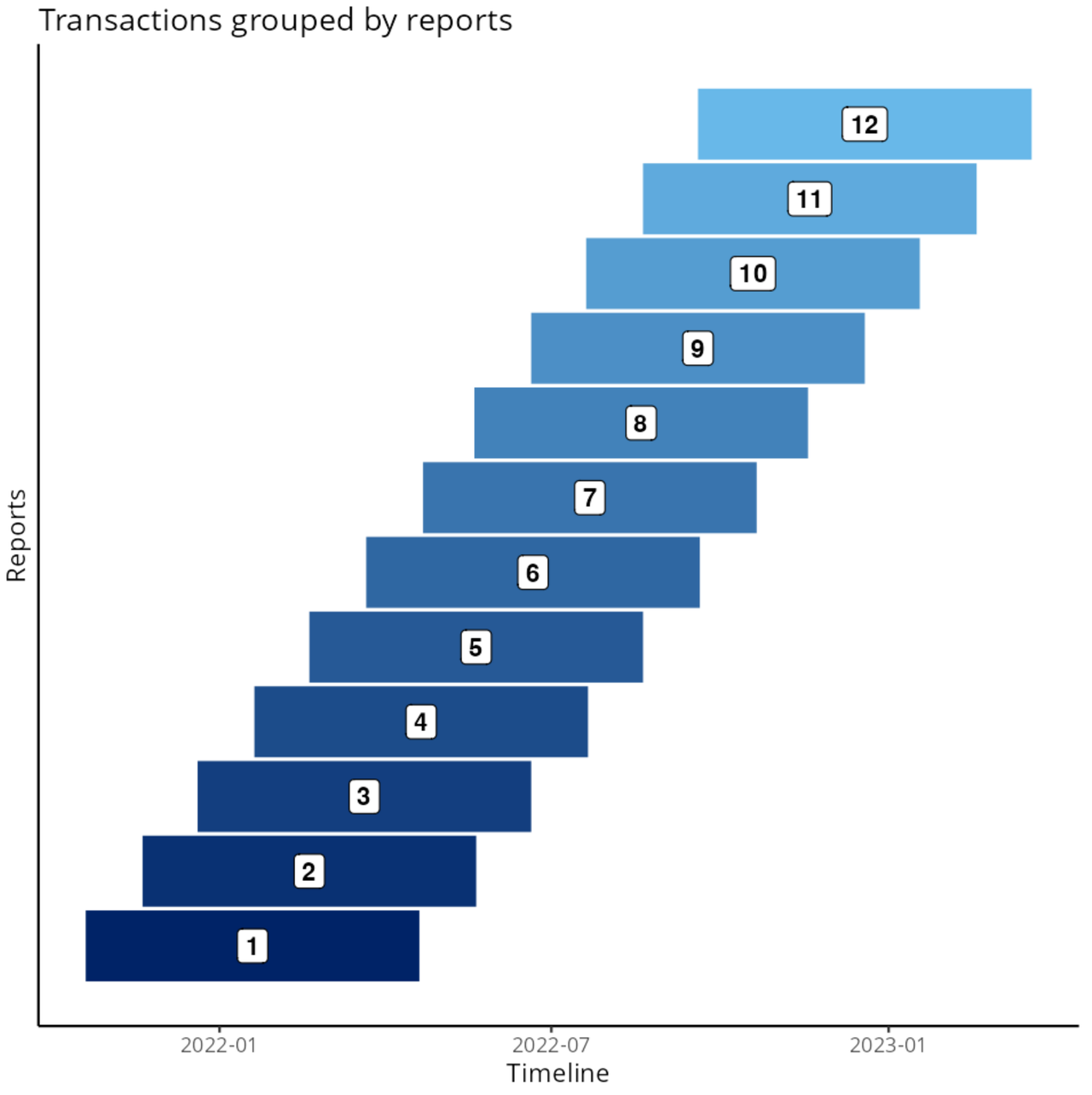
2. Данные клиентских транзакций:

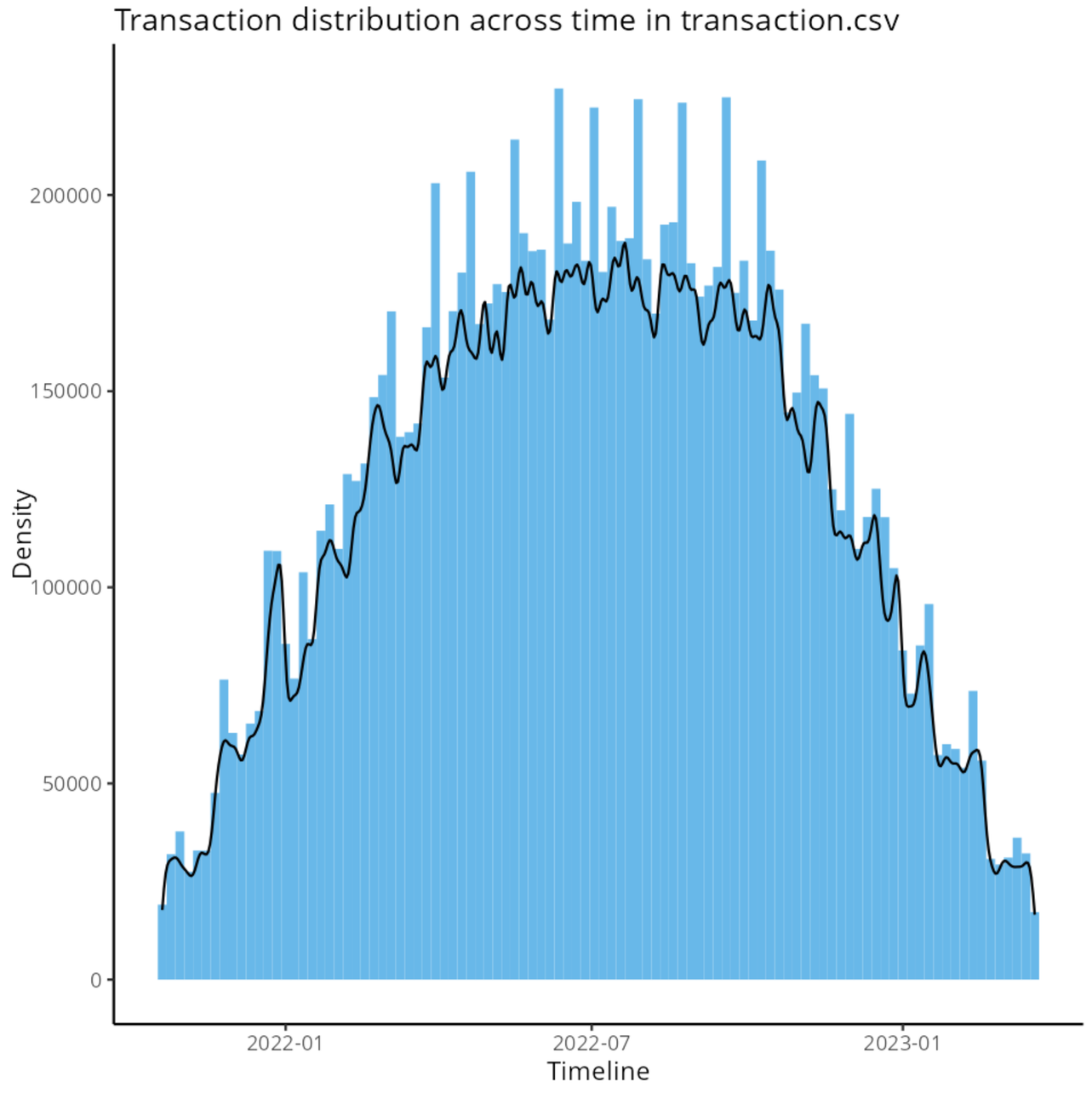
1. Клиентские транзакции для всех 96,000 клиентов (~13M) в табличном `.csv` формате: `transactions.csv.zip` (**197 MB**)

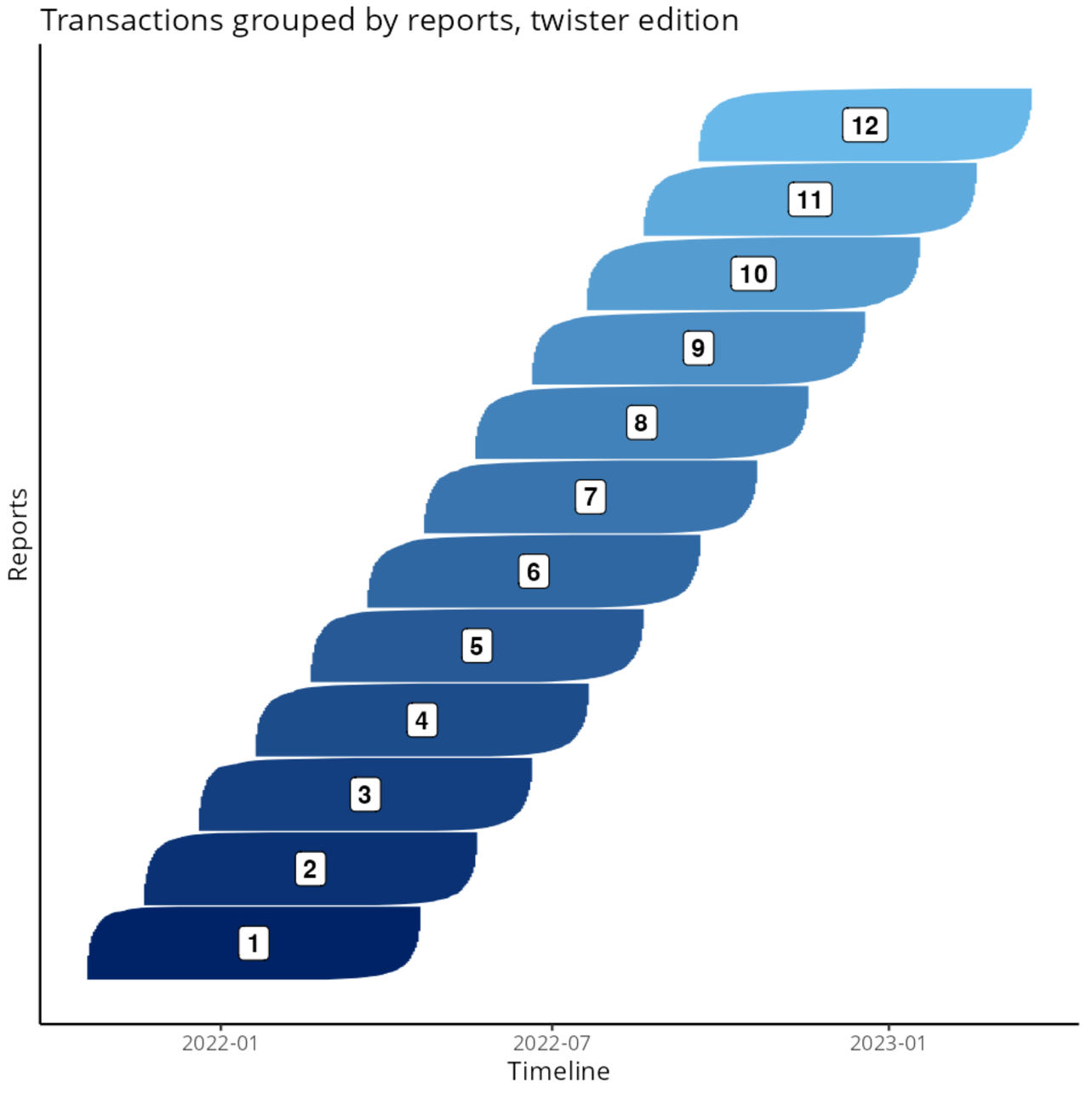

Transaction distribution across time in transaction.csv
200000
150000
100000
50000
0
Density
2022-01
2022-07
2023-01
Timeline

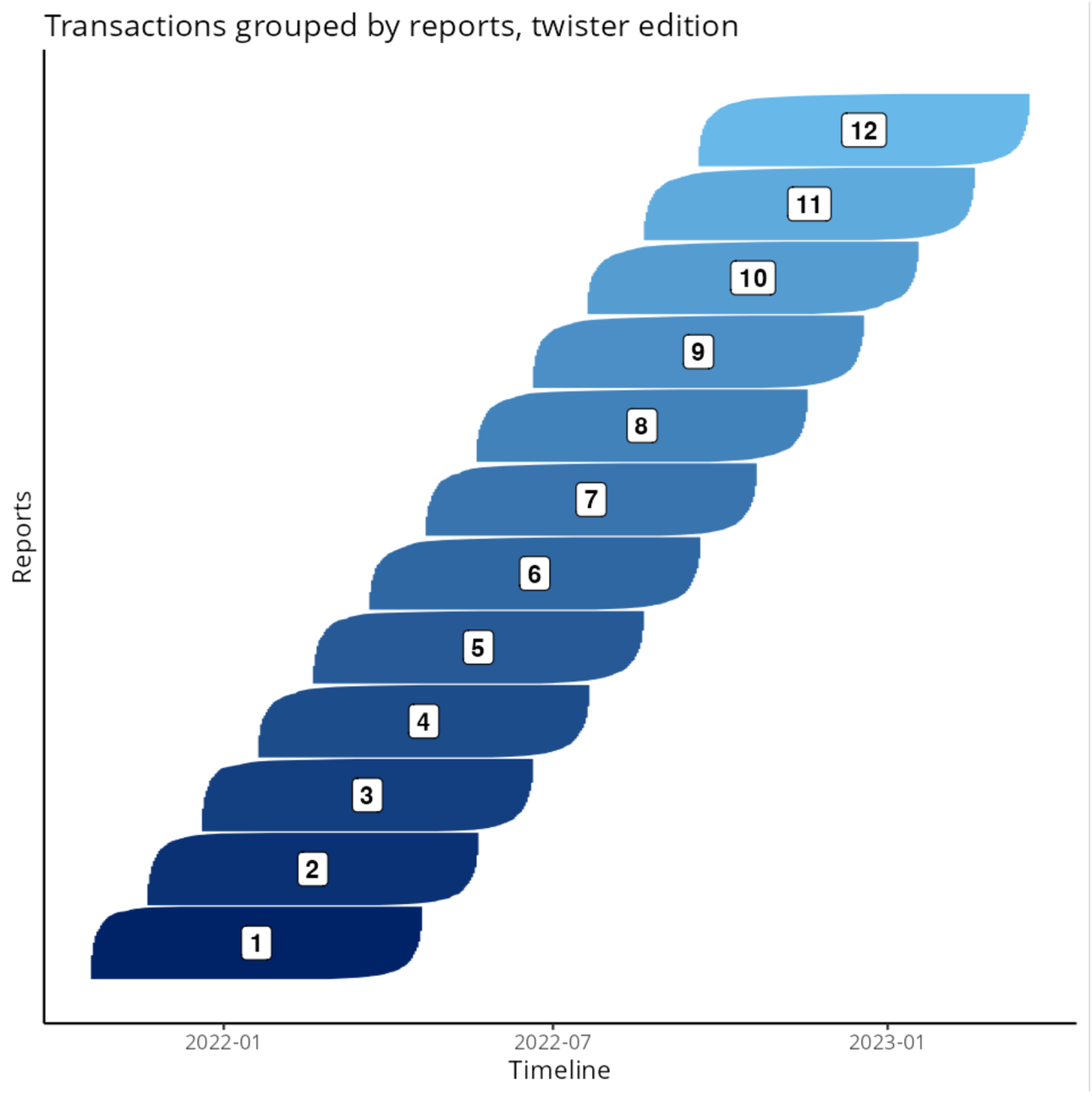

Transaction |amt| quantiles in transaction.csv
Transaction |amt|, RUB only
3000
1000
300
100
30
2021-12
2022-03
2022-06
2022-09
2022-12
2023-03
Timeline
q_90
q_80
q_70
q_60
q_50
q_40
q_30
q_20
q_10

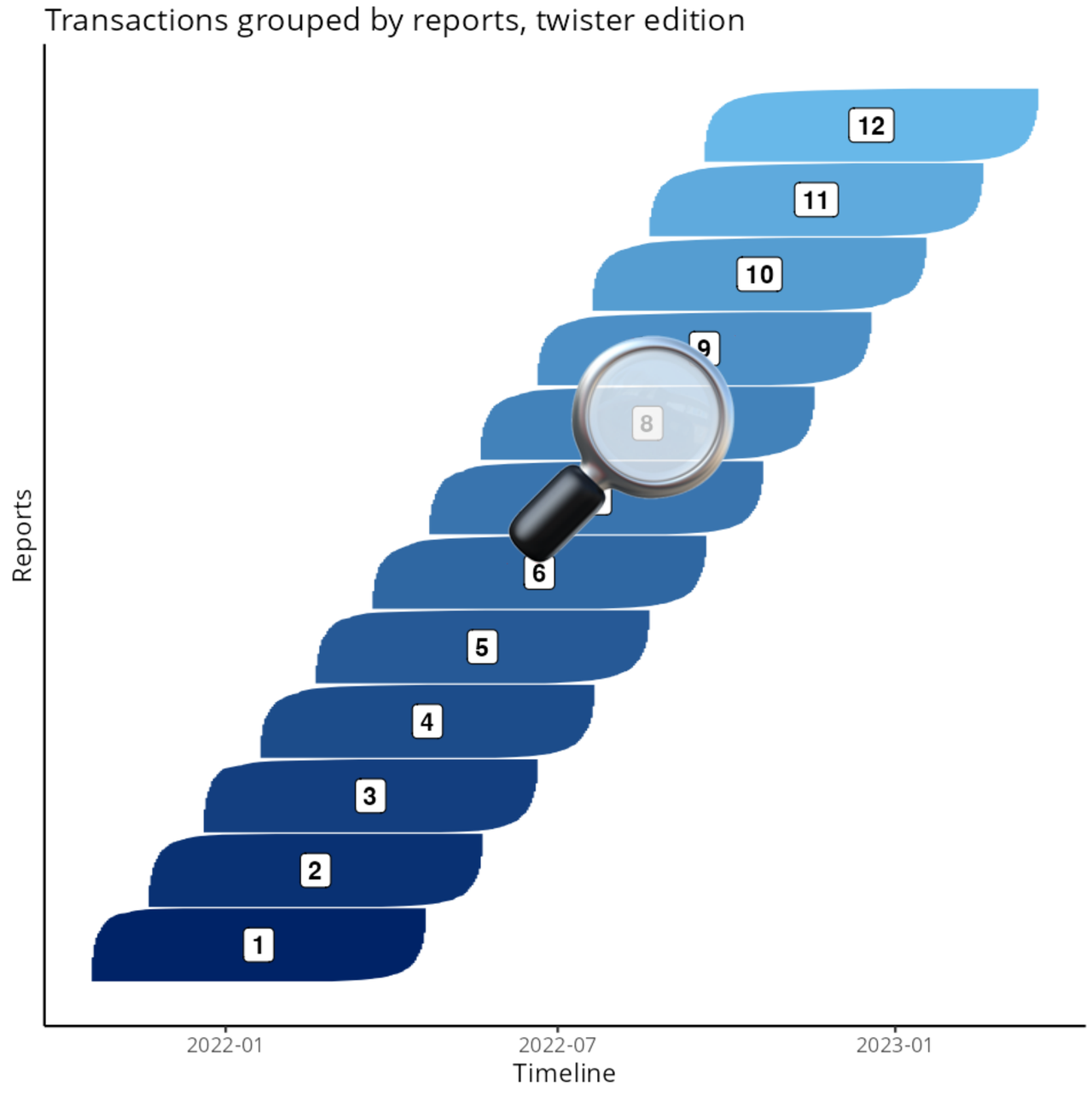

Transaction distribution across time in transaction.csv
Density
200000
150000
100000
50000
0
2022-01
2022-07
2023-01
Timeline

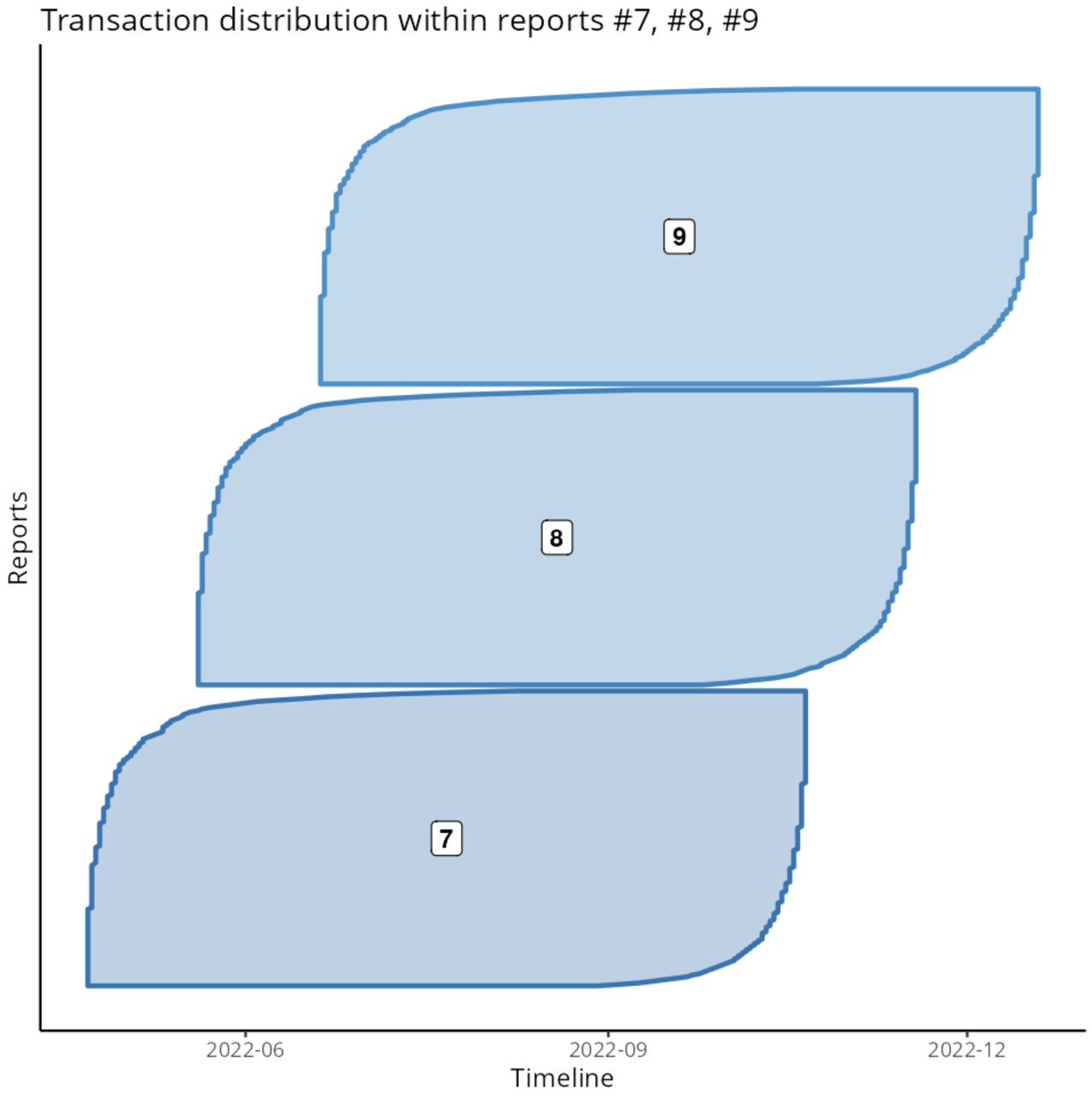

Transactions grouped by reports
Reports
12
11
10
9
8
7
6
5
4
3
2
1
2022-01
2022-07
2023-01
Timeline

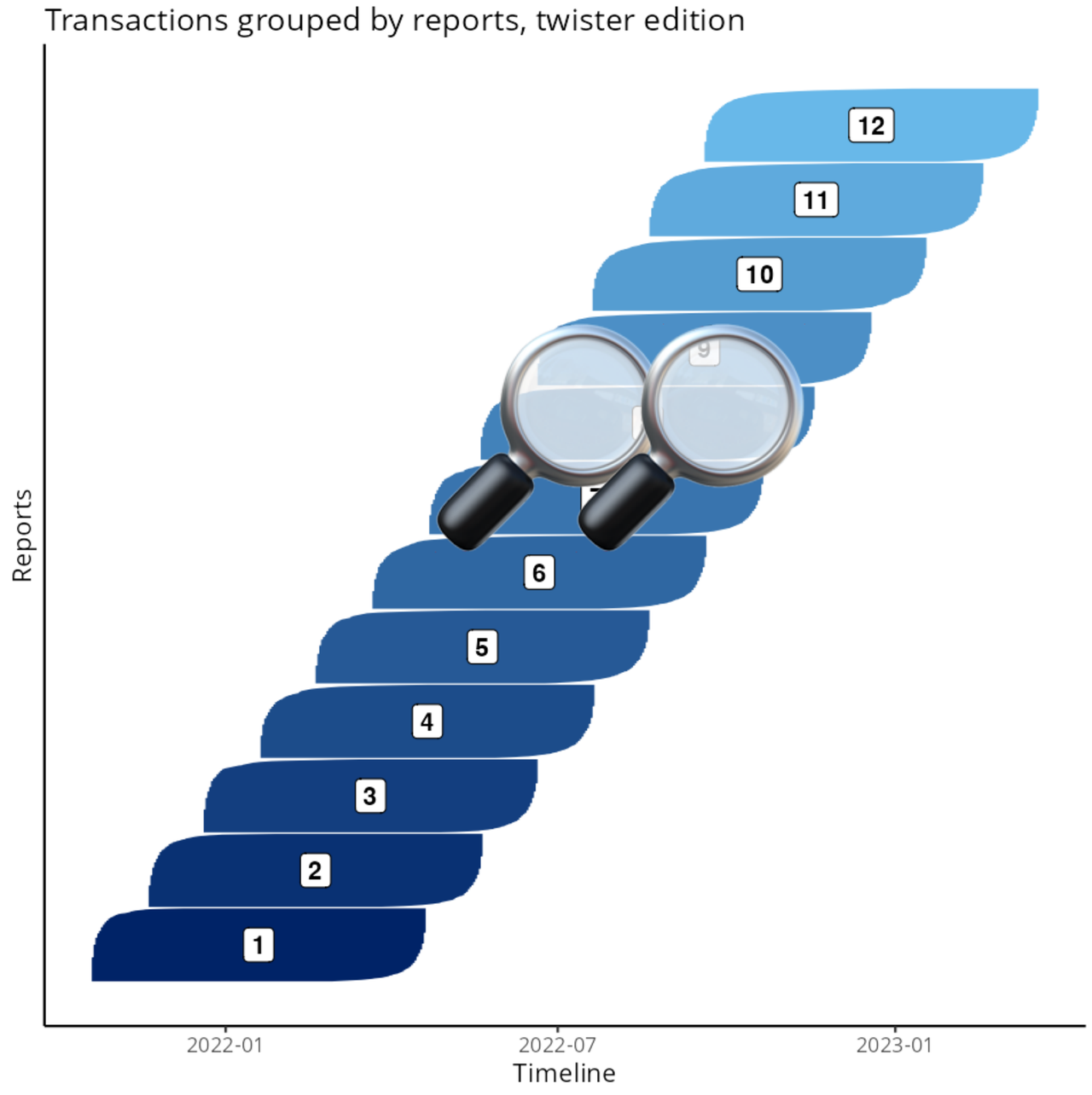

Transaction distribution across time in transaction.csv
Density
200000
150000
100000
50000
0
2022-01
2022-07
2023-01
Timeline

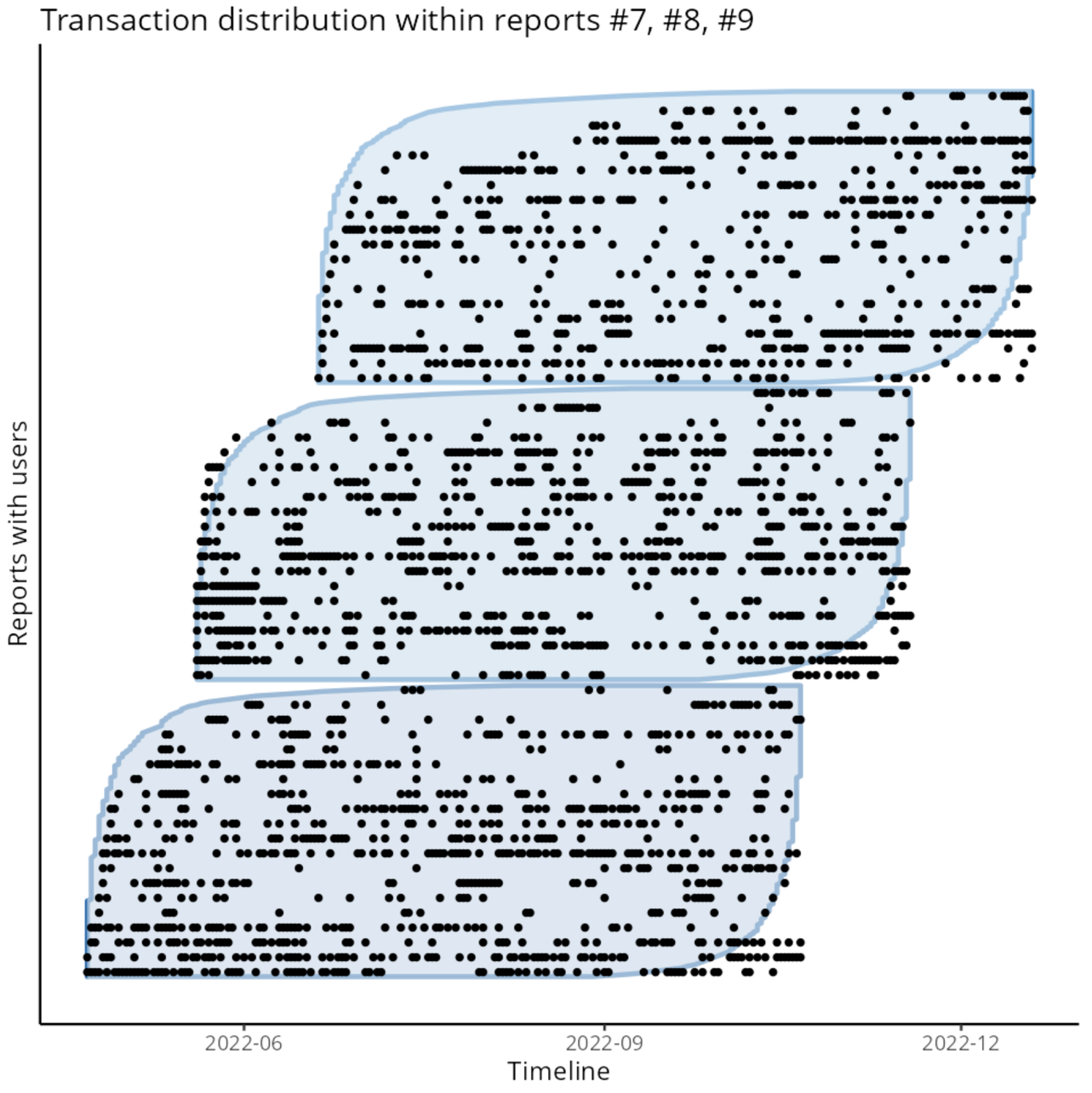

Transactions grouped by reports, twister edition
Reports
12
11
10
9
8
7
6
5
4
3
2
1
2022-01
2022-07
2023-01
Timeline


Transactions grouped by reports, twister edition
Reports
12
11
10
9
8
7
6
5
4
3
2
1
2022-01
2022-07
2023-01
Timeline


Transactions grouped by reports, twister edition
12
11
10
9
8
6
5
4
3
2
1
Reports
2022-01
2022-07
2023-01
Timeline

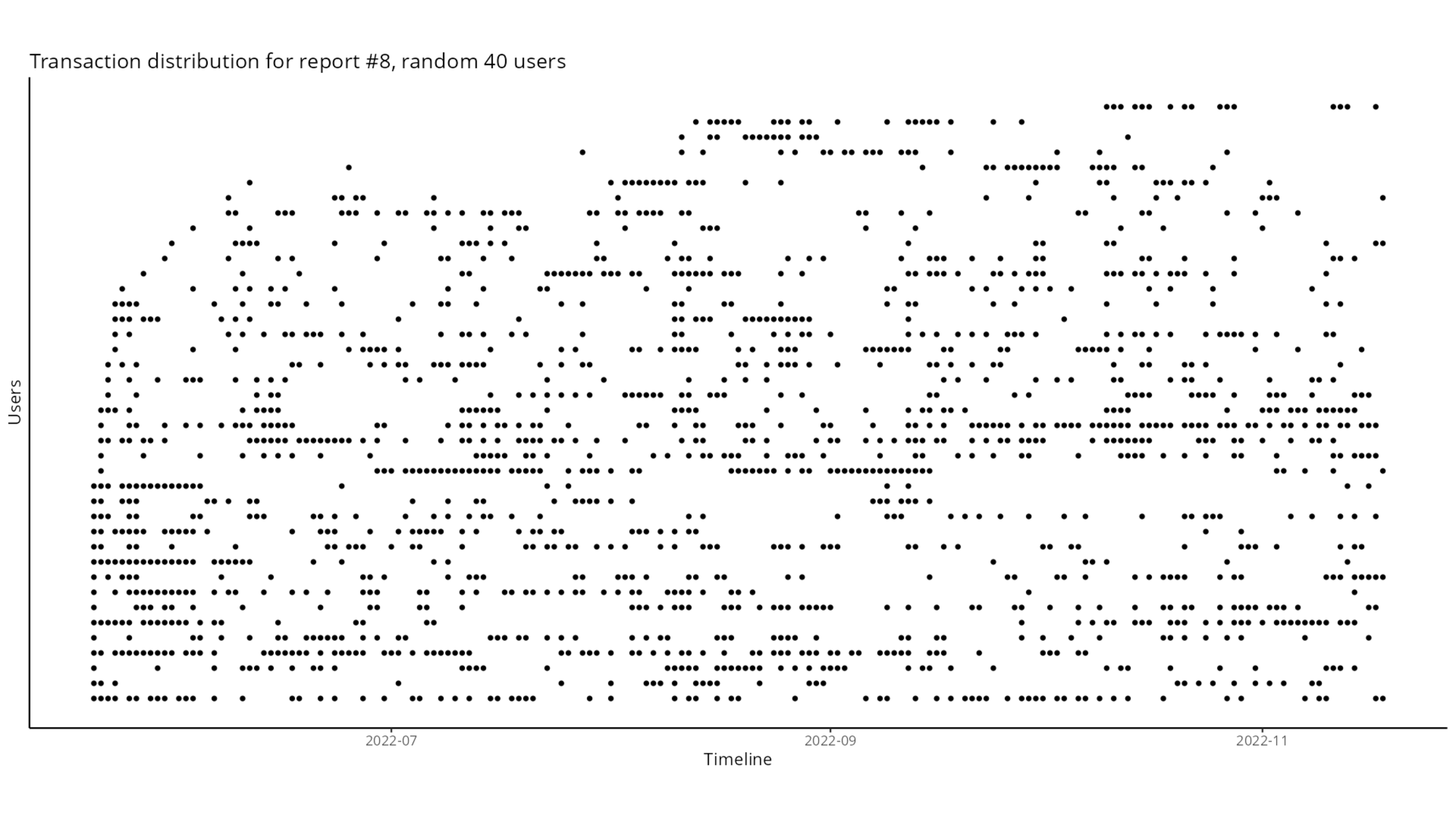

Transaction distribution within reports #7, #8, #9
9
8
7
Reports
2022-06
2022-09
2022-12
Timeline

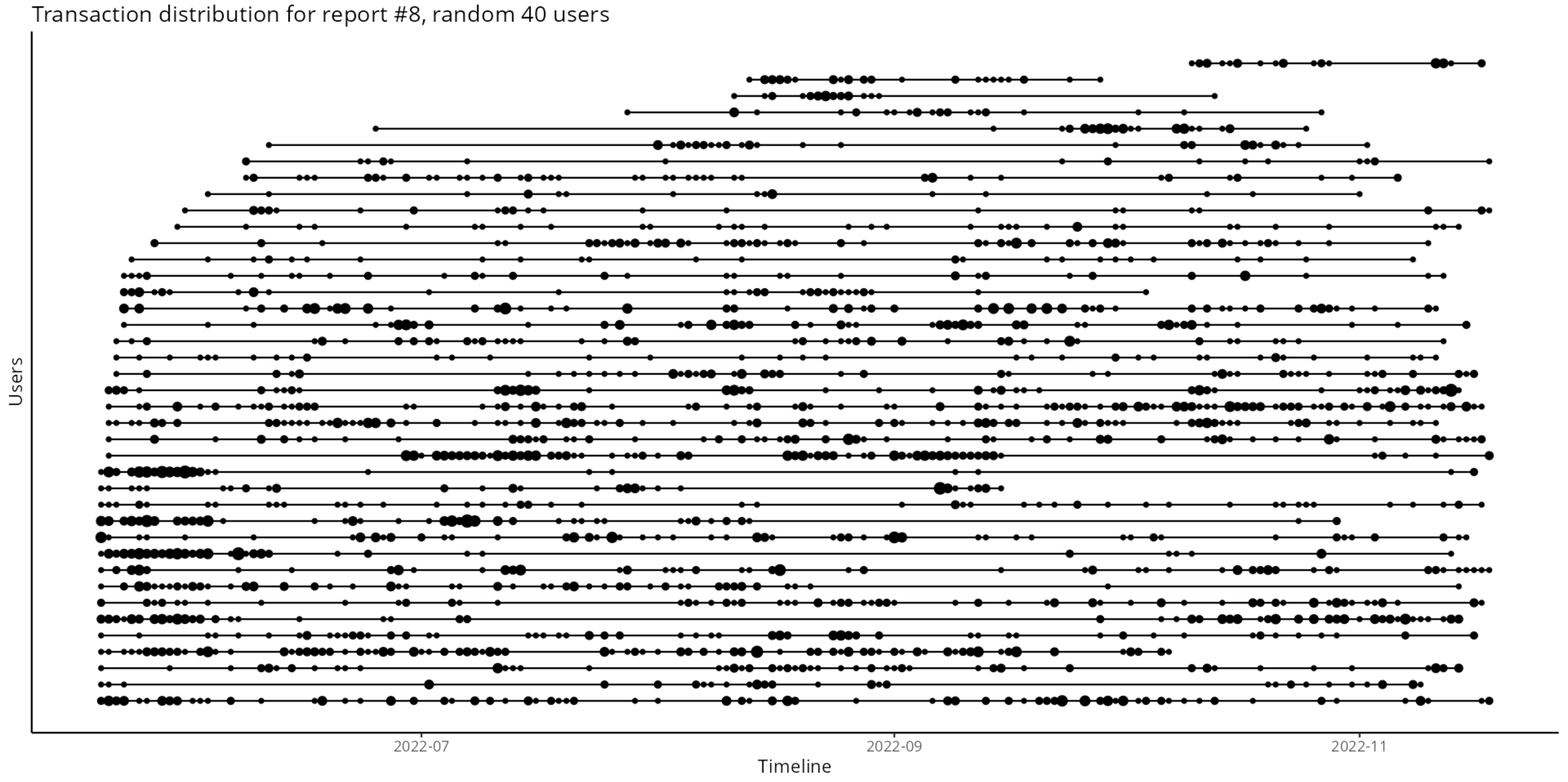

Transactions grouped by reports, twister edition
12
11
10
9
6
5
4
3
2
1
Reports
2022-01
2022-07
2023-01
Timeline

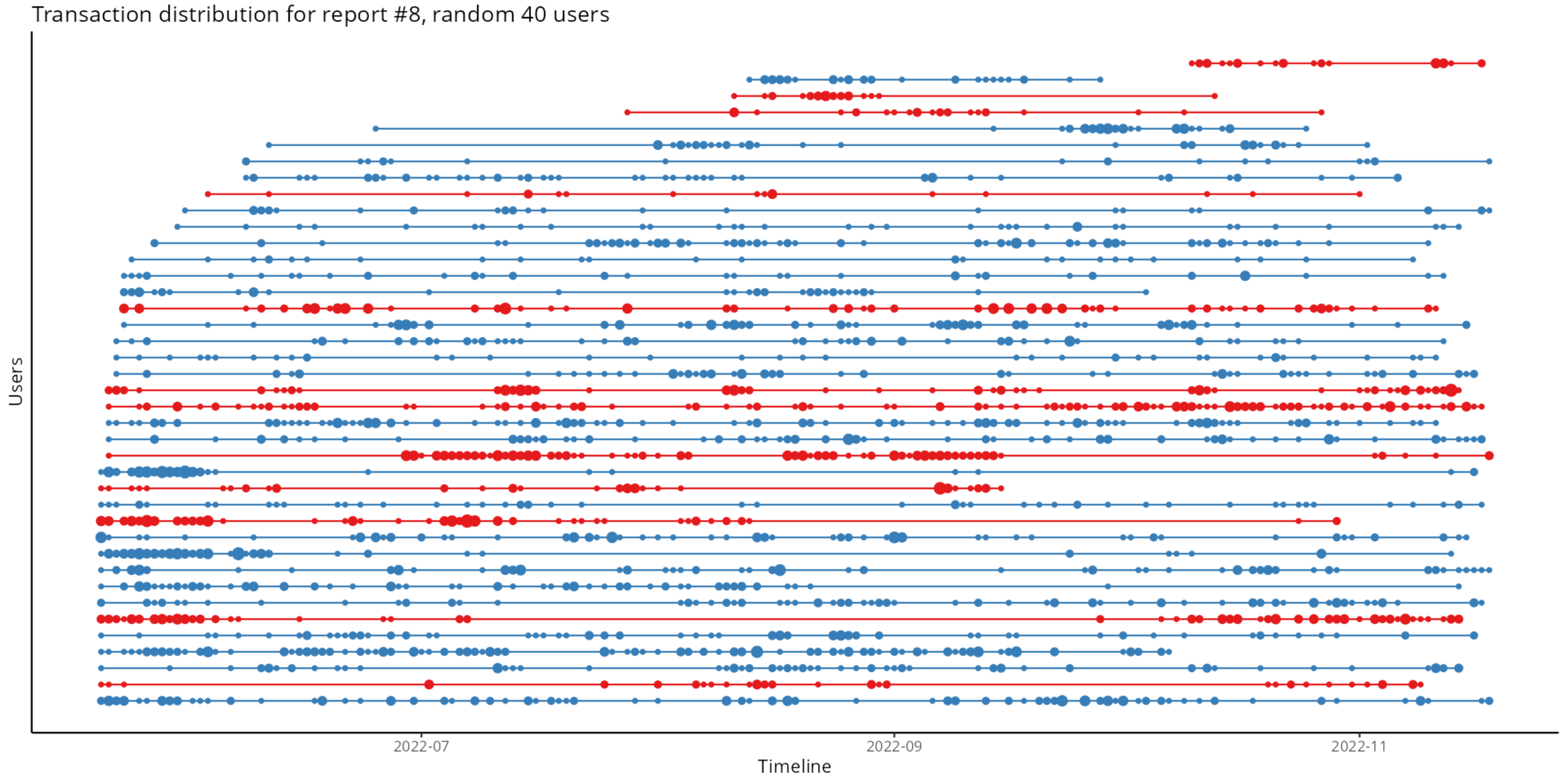

Transaction distribution within reports #7, #8, #9
Reports with users
2022-06
2022-09
2022-12
Timeline

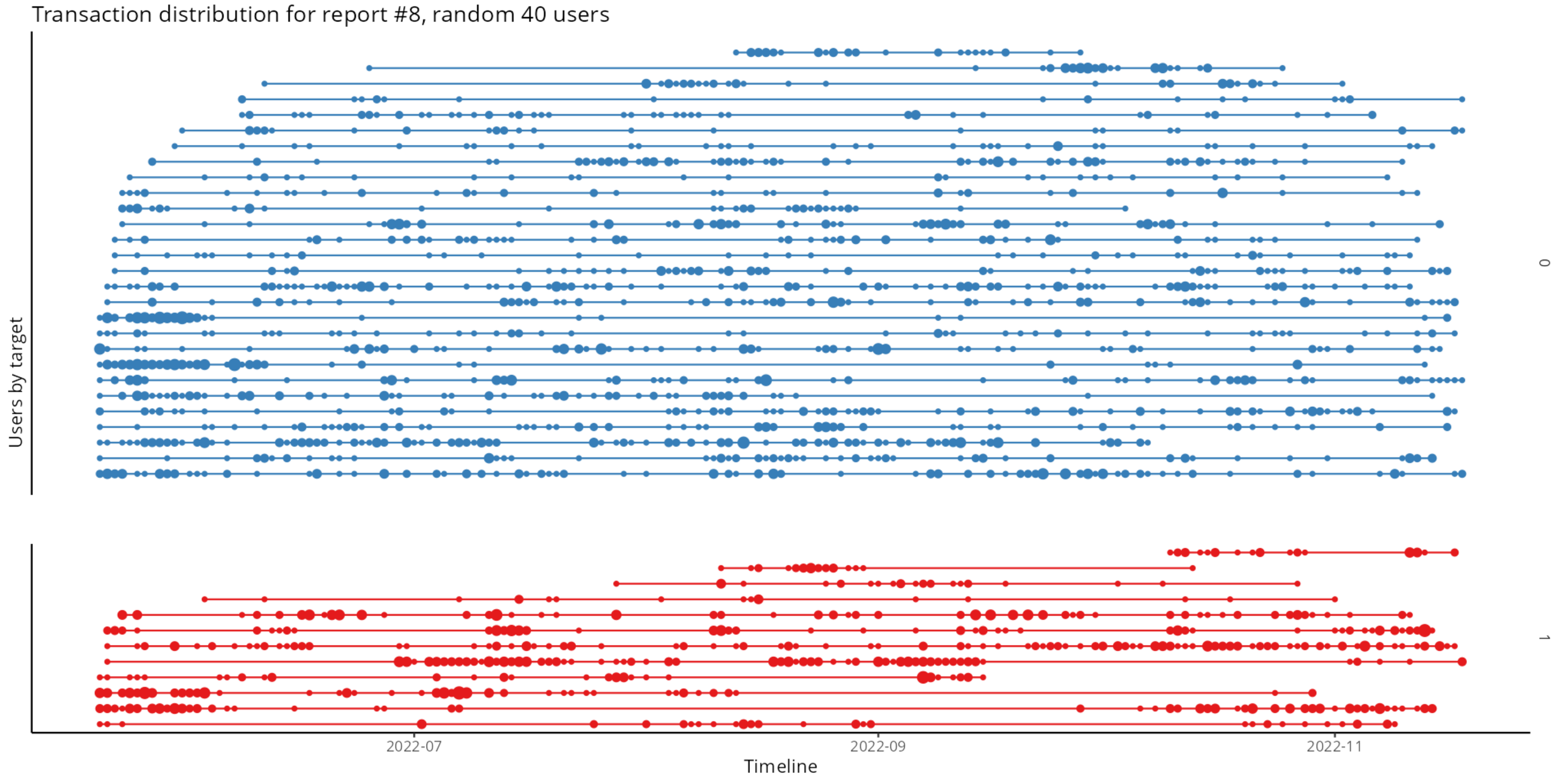
Часть 3

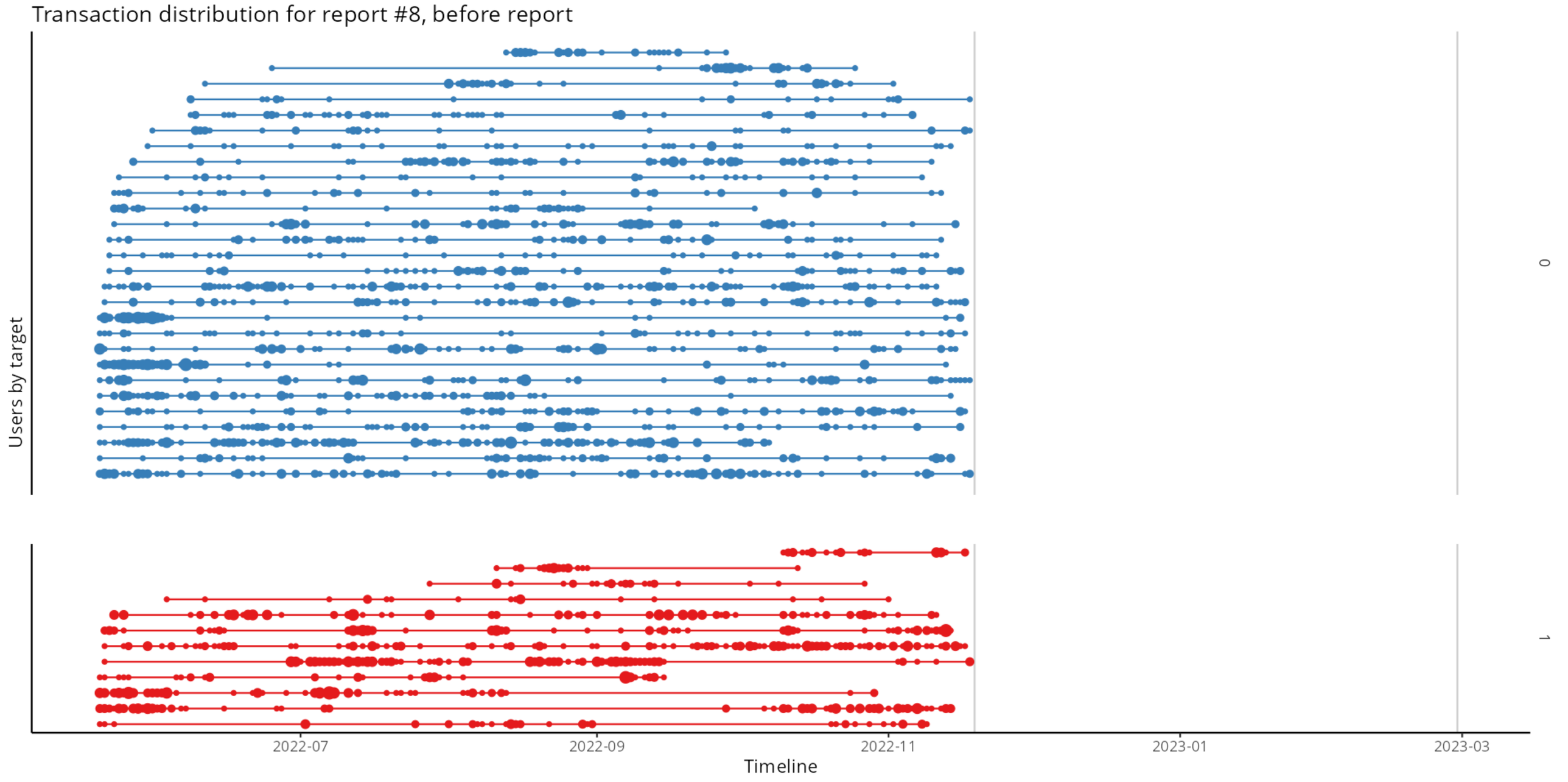
Что на самом деле происходит в транзакциях? ~~Что предсказываем?~~

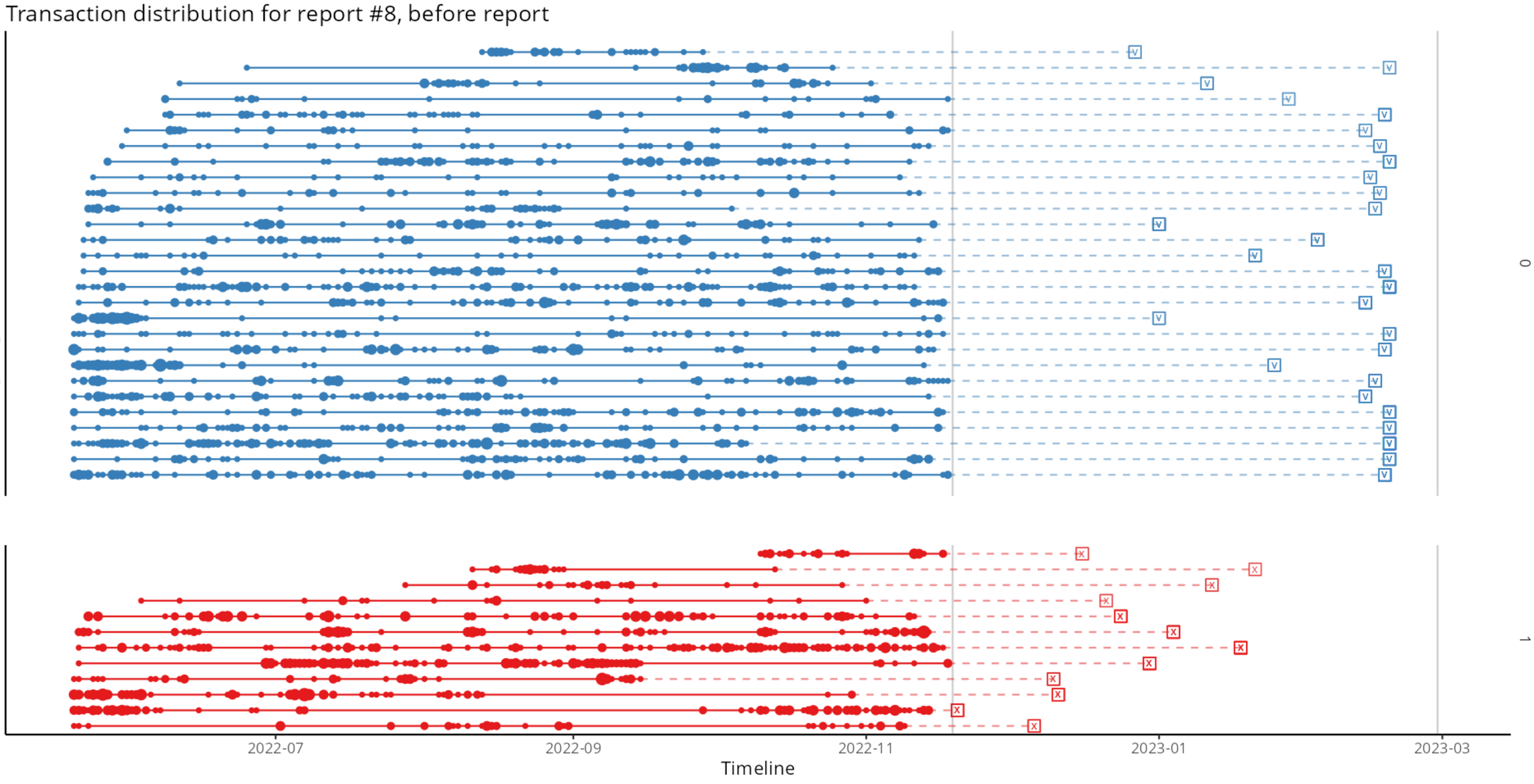

Transaction distribution for report #8, random 40 users
Users
2022-07
2022-09
2022-11
Timeline

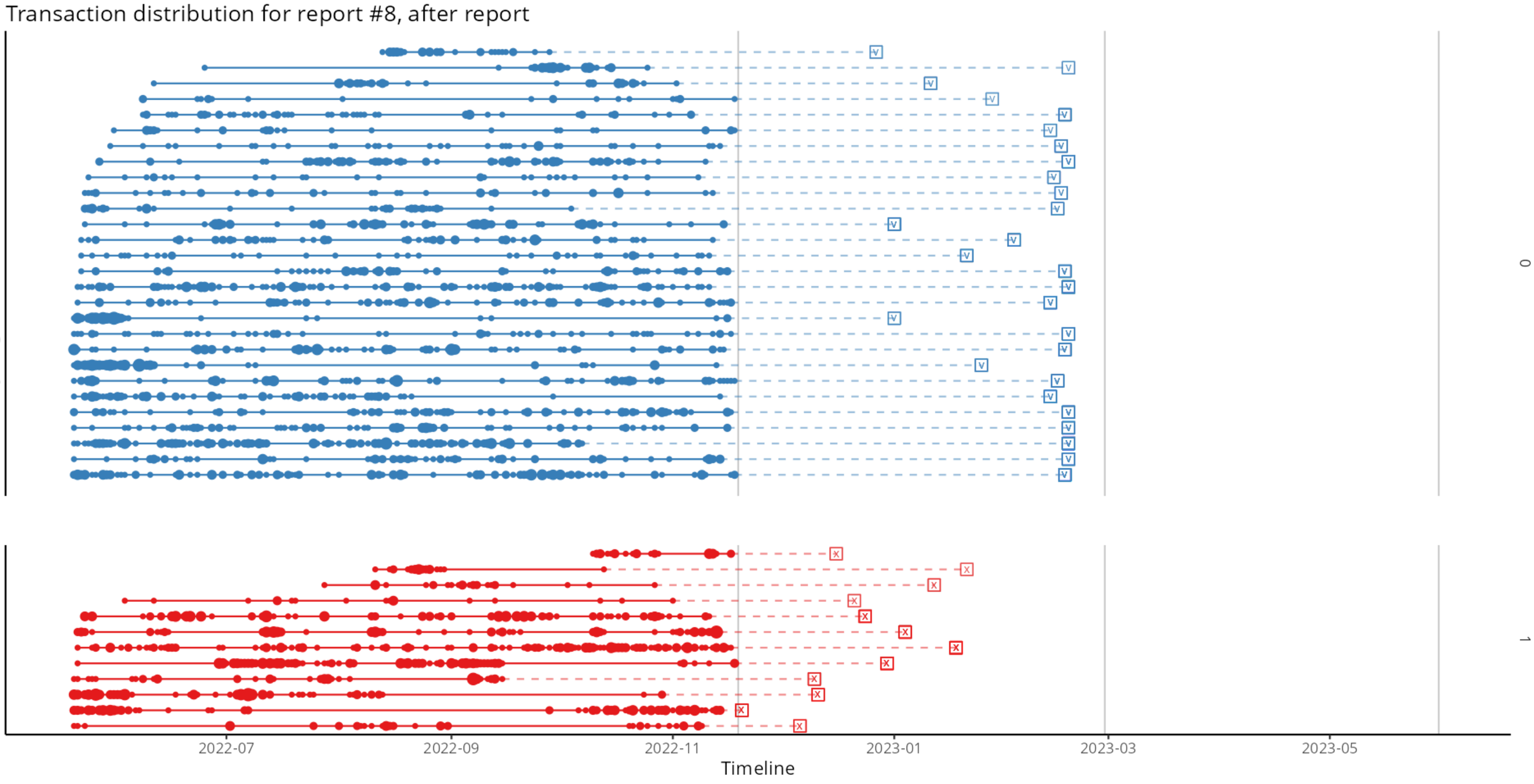

Transaction distribution for report #8, random 40 users
Users
2022-07
2022-09
2022-11
Timeline

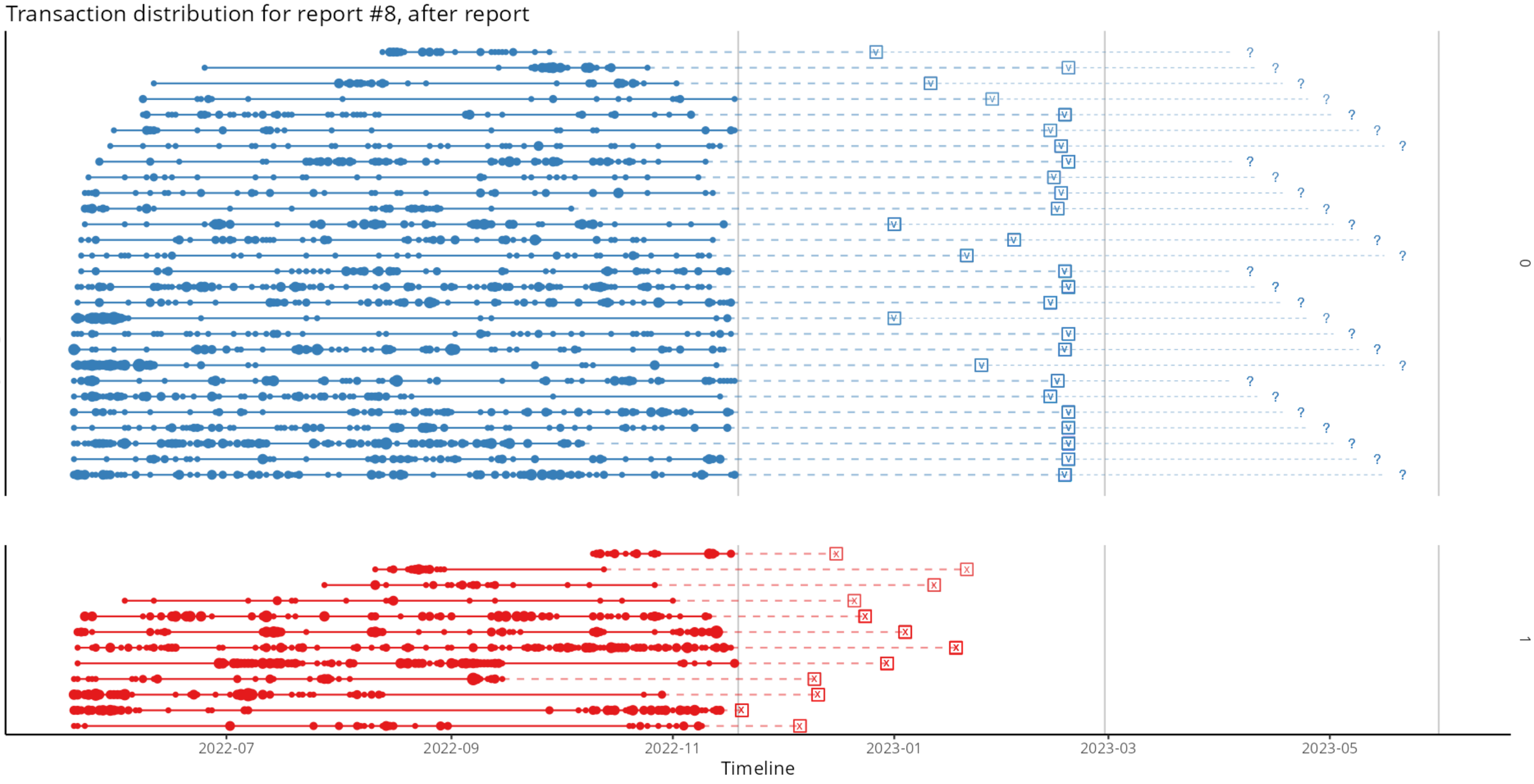

Transaction distribution for report #8, random 40 users
Users
2022-07
2022-09
2022-11
Timeline


Transaction distribution for report #8, random 40 users
Users by target
0
1
2022-07
2022-09
2022-11
Timeline


Transaction distribution for report #8, before report
Users by target
0
1
2022-07
2022-09
2022-11
2023-01
2023-03
Timeline

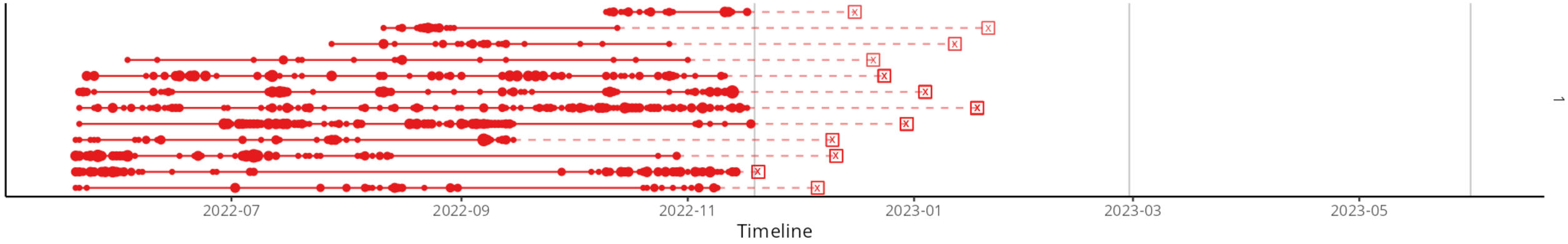

Transaction distribution for report #8, before report
Users by target
0
1
2022-07
2022-09
2022-11
2023-01
2023-03
Timeline


Transaction distribution for report #8, after report
Users by target
0
1
2022-07
2022-09
2022-11
2023-01
2023-03
2023-05
Timeline


Transaction distribution for report #8, after report
Users by target
0
1
2022-07
2022-09
2022-11
2023-01
2023-03
2023-05
Timeline

Исходные метки оттока, используемые в этом соревновании, получены именно с помощью этого процесса. Как не трудно видеть, это определение подразумевает, что последняя транзакция клиента должна была состояться до даты отчета. В таком случае, можно считать что у нас есть 2 определения оттока:

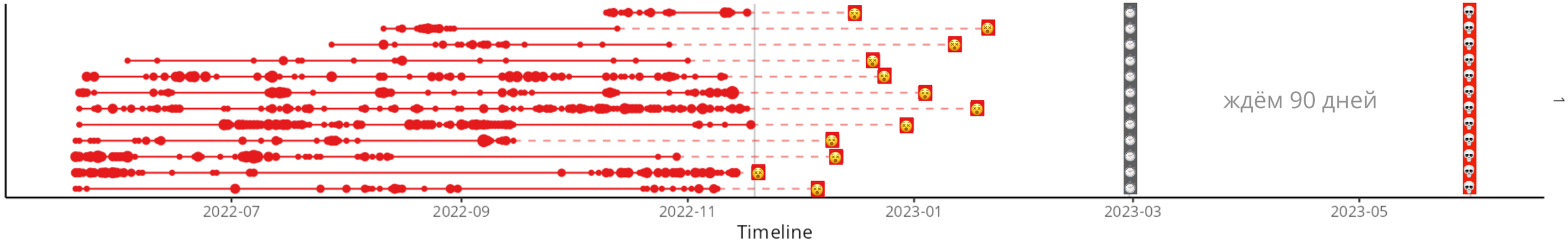
- **Документальный** отток: вызревшие метки, по которым срабатывает бизнес-правило (через 60-90 дней после отчета).
- **Клинический** отток: событие, когда у клиента “остановился пульс” транзакций, и которое наступило не раньше даты последней транзакции (до отчета). А спустя 60-90 дней отсутствия транзакций, по этому клиенту запишут статус документального оттока.

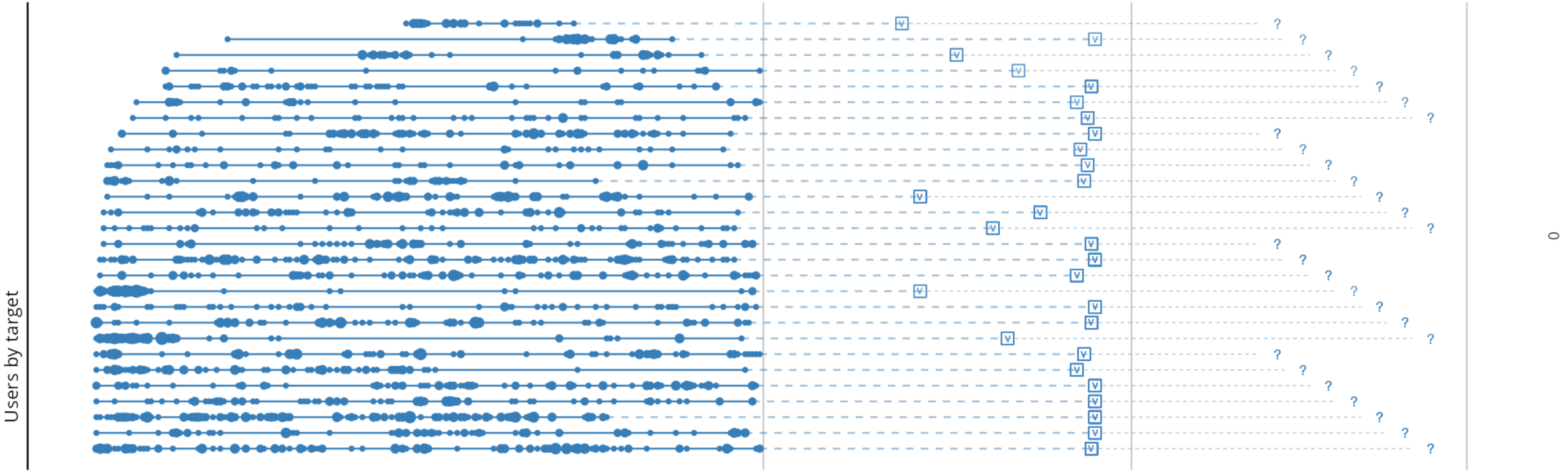


“клинический” отток

дата отчета

“документальный” отток





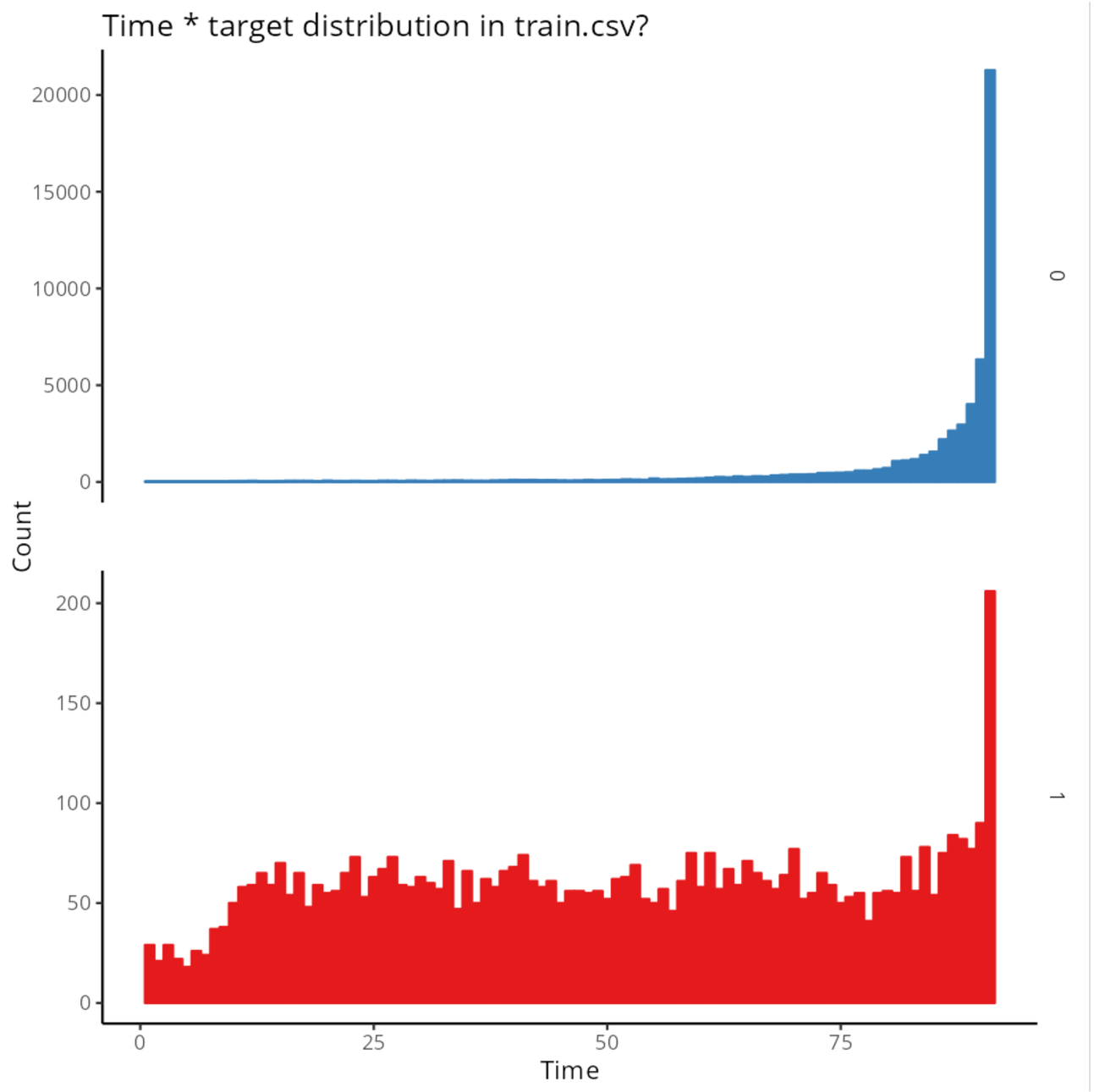
последние транзакции до отчета

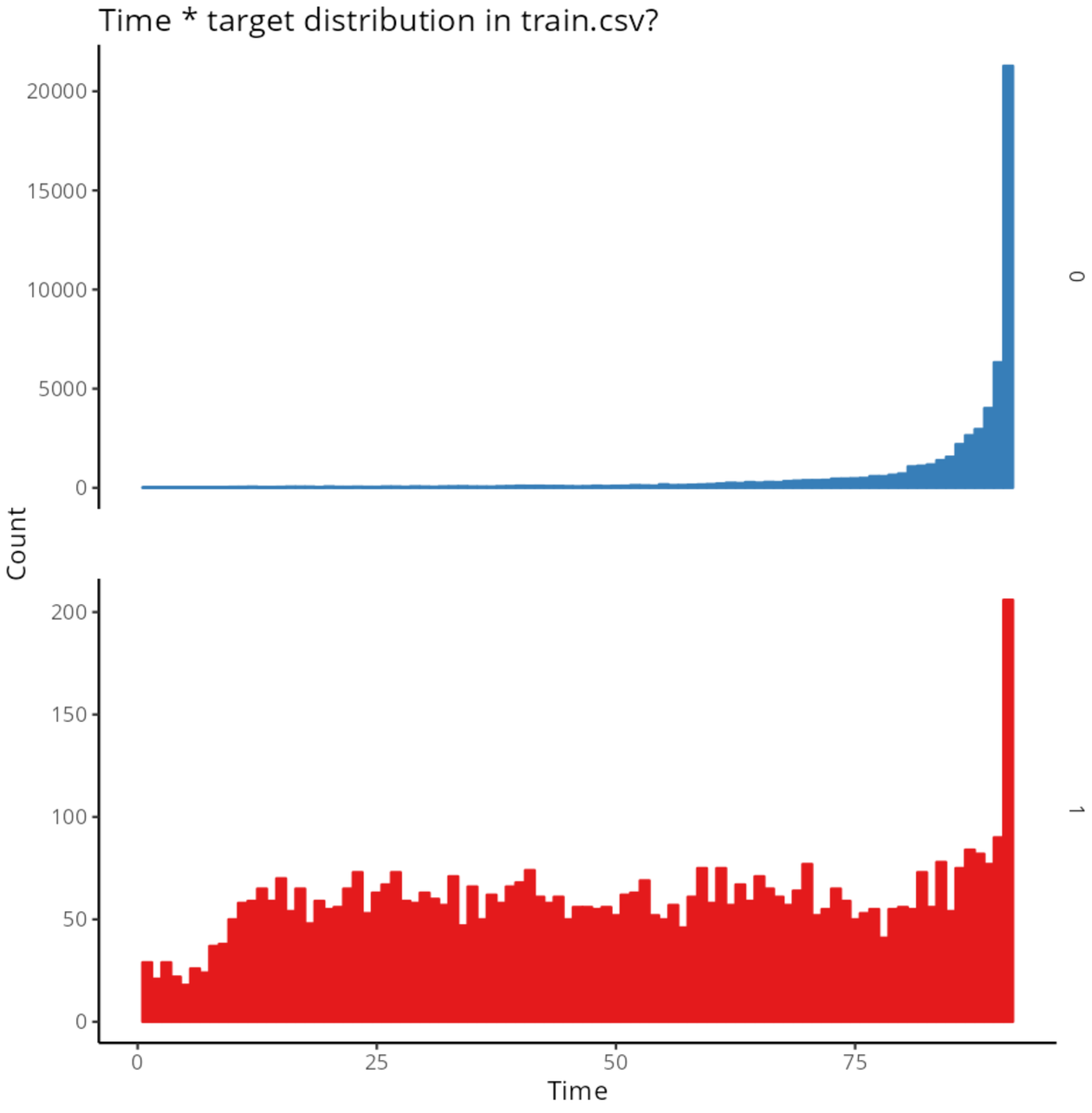
дата отчета

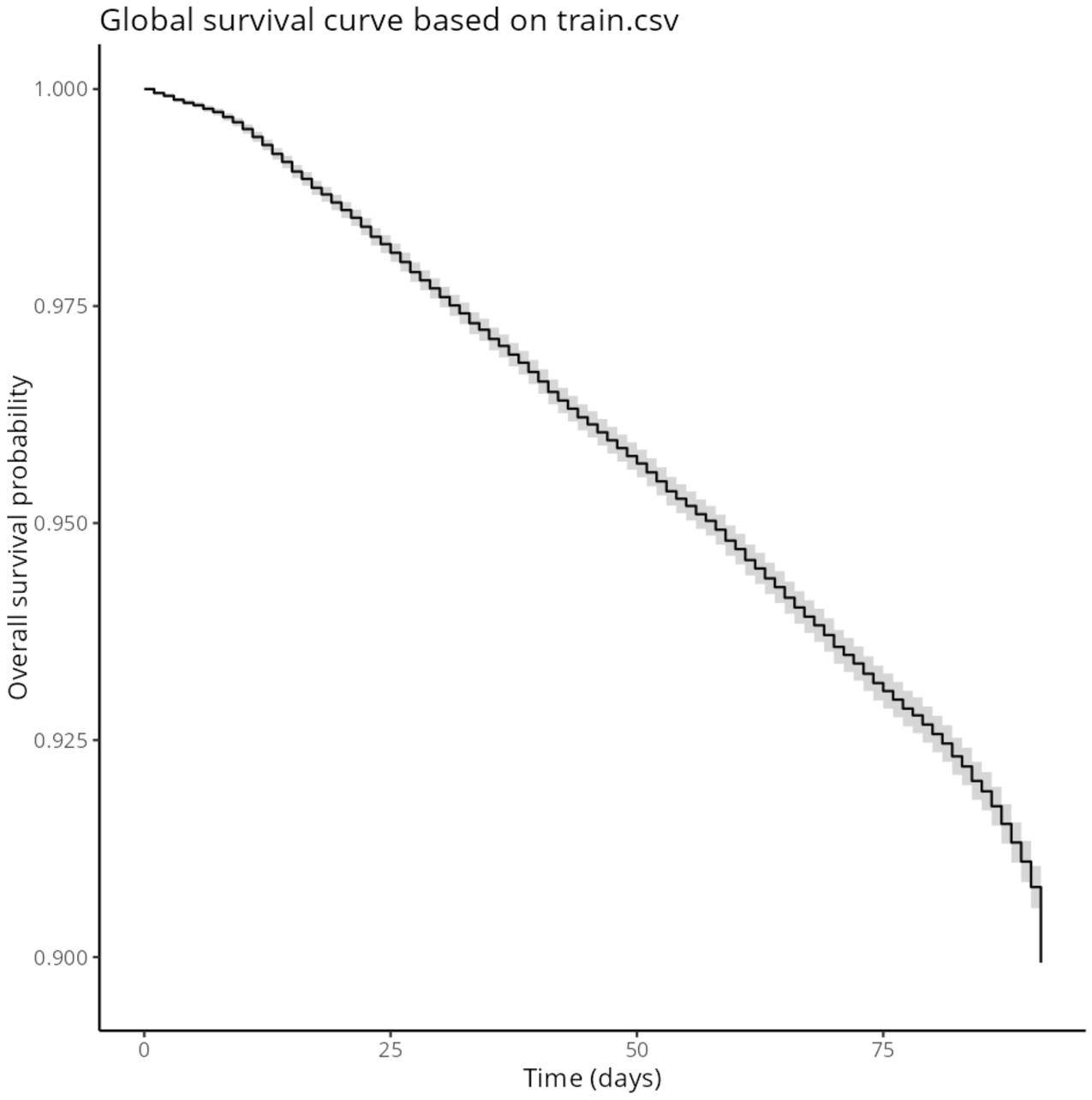
цензурированные данные

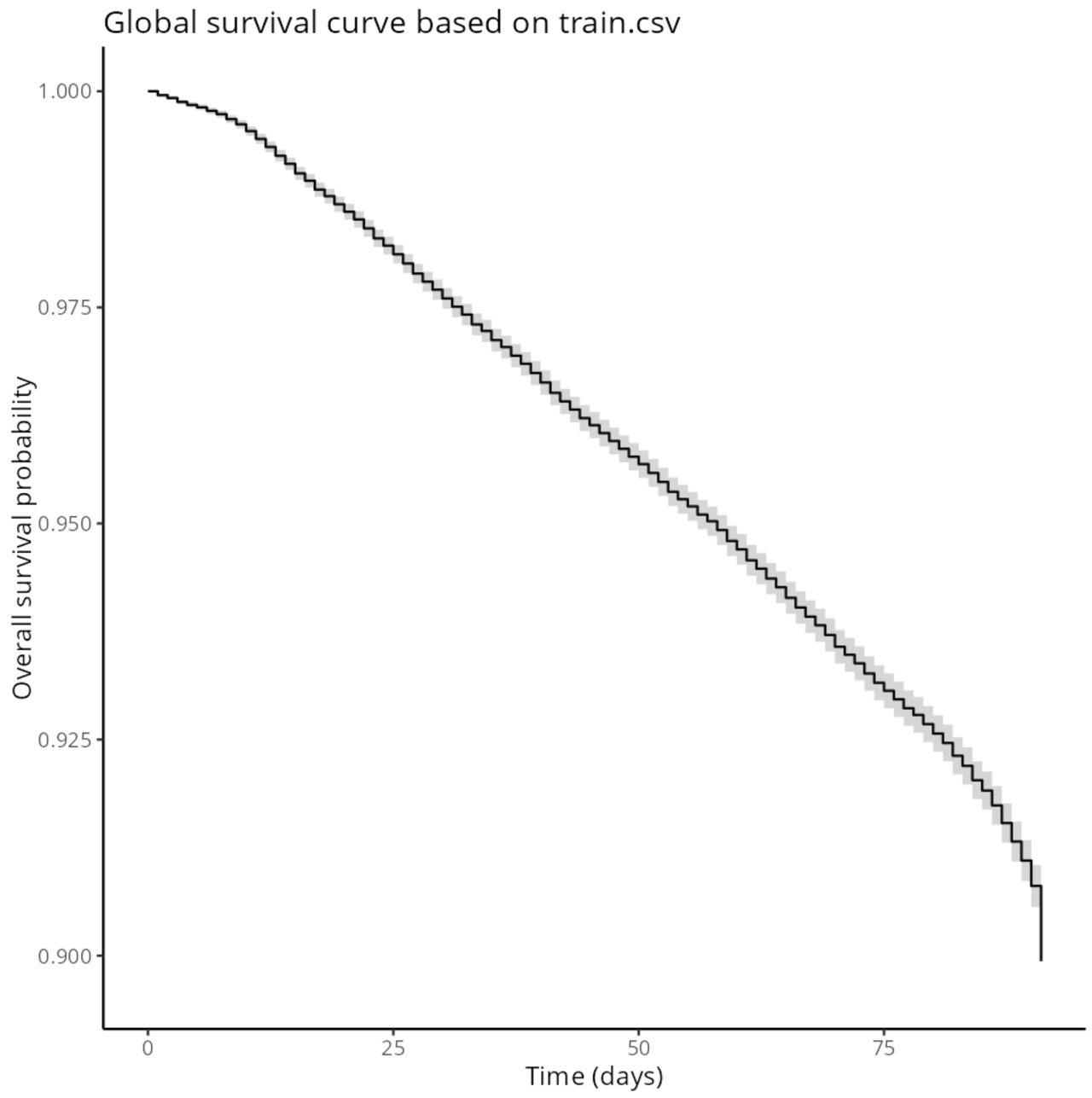
Часть 4

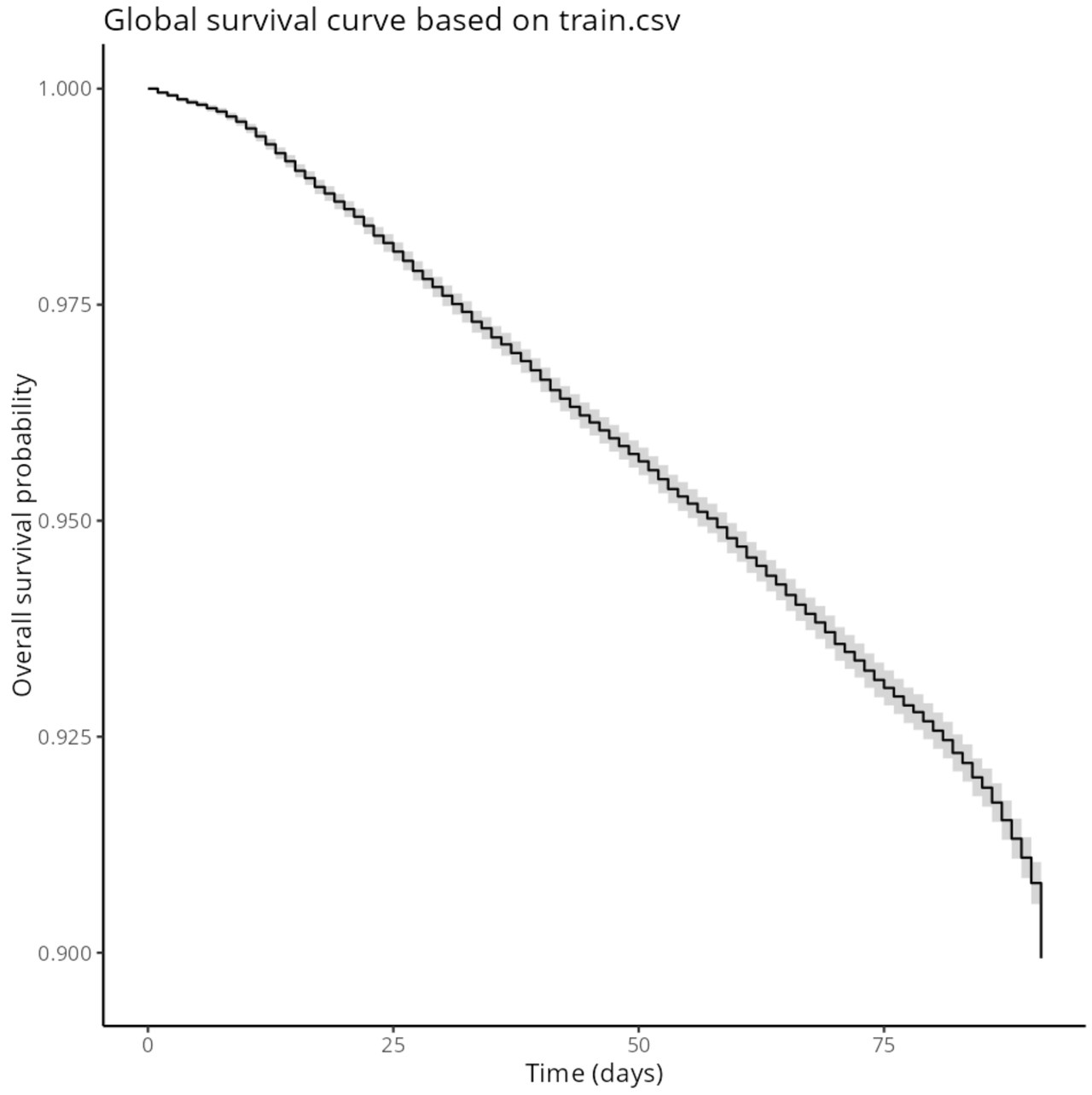
Что еще мы можем предсказывать?

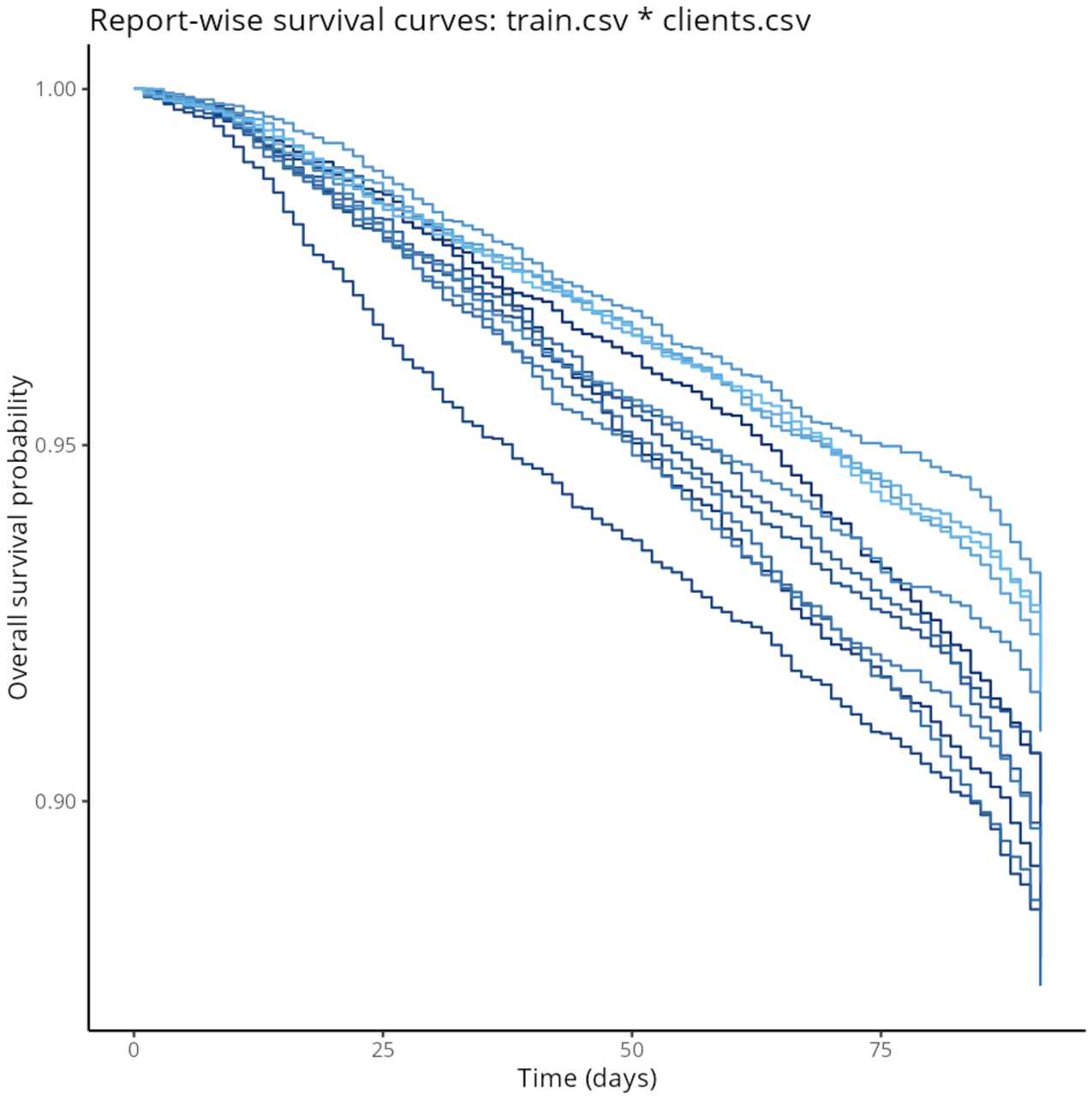

Time * target distribution in train.csv?
20000
15000
10000
5000
0
0
200
150
100
50
0
Count
0
25
50
75
Time
0
1

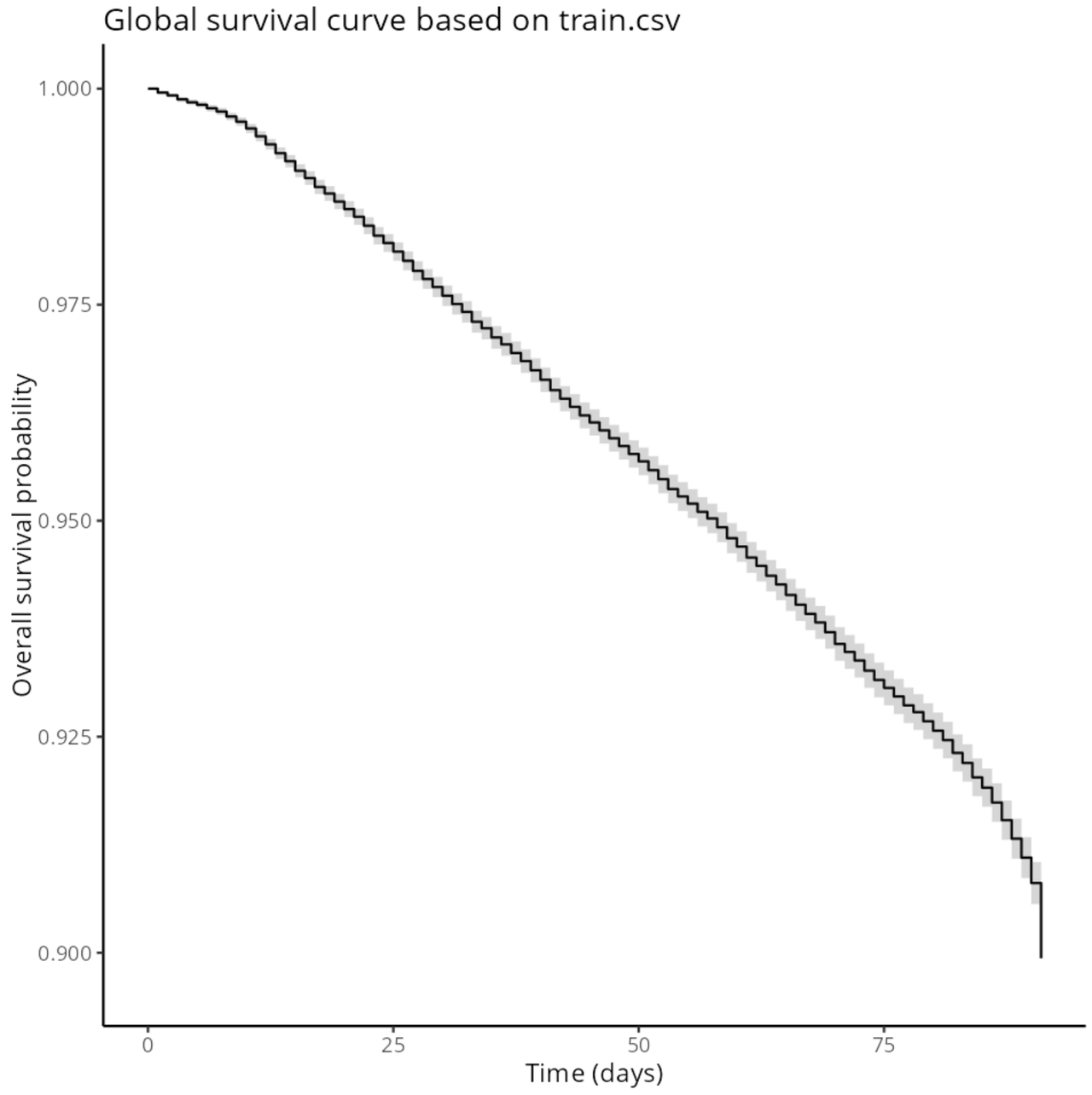

Time * target distribution in train.csv?
20000
15000
10000
5000
0
Count
0
200
150
100
50
0
1
0
25
50
75
Time

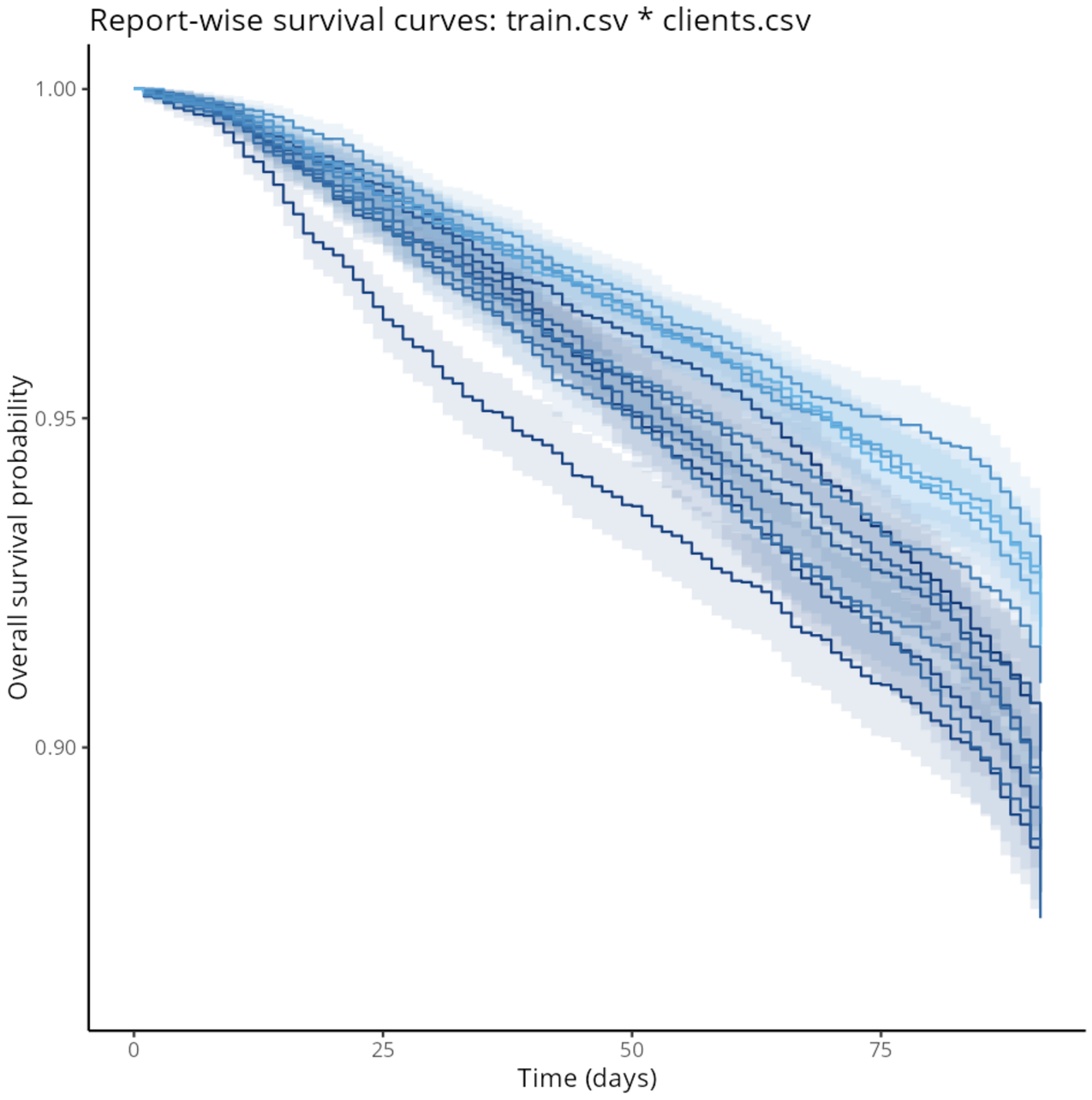

Global survival curve based on train.csv
1.000
0.975
0.950
0.925
0.900
Overall survival probability
0
25
50
75
Time (days)

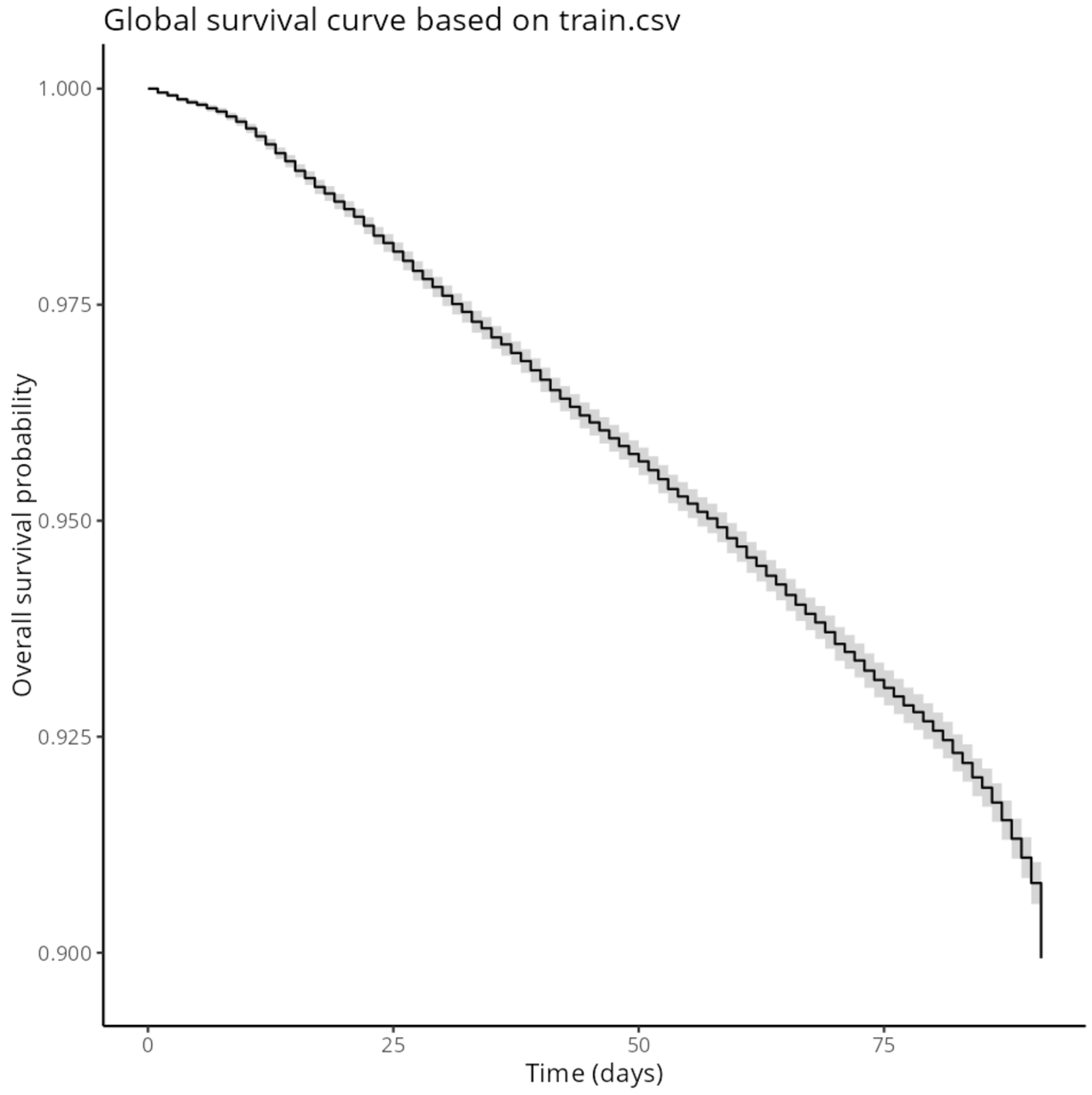

Global survival curve based on train.csv
1.000
0.975
0.950
0.925
0.900
Overall survival probability
0
25
50
75
Time (days)

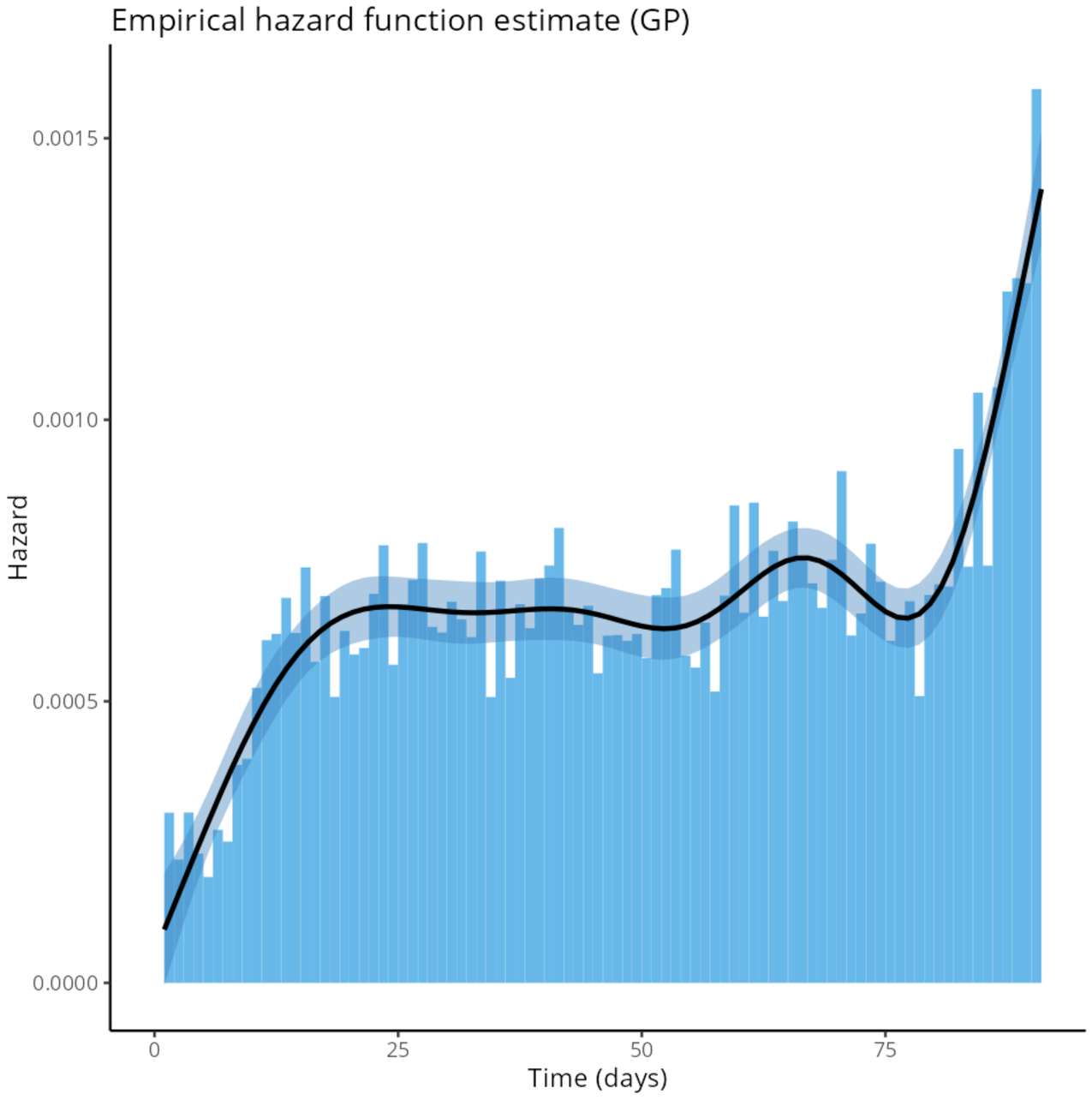

Global survival curve based on train.csv
1.000
0.975
0.950
0.925
0.900
Overall survival probability
0
25
50
75
Time (days)


Report-wise survival curves: train.csv * clients.csv
1.00
0.95
0.90
Overall survival probability
0
25
50
75
Time (days)


Global survival curve based on train.csv
1.000
0.975
0.950
0.925
0.900
Overall survival probability
0
25
50
75
Time (days)


Report-wise survival curves: train.csv * clients.csv
1.00
0.95
0.90
Overall survival probability
0
25
50
75
Time (days)

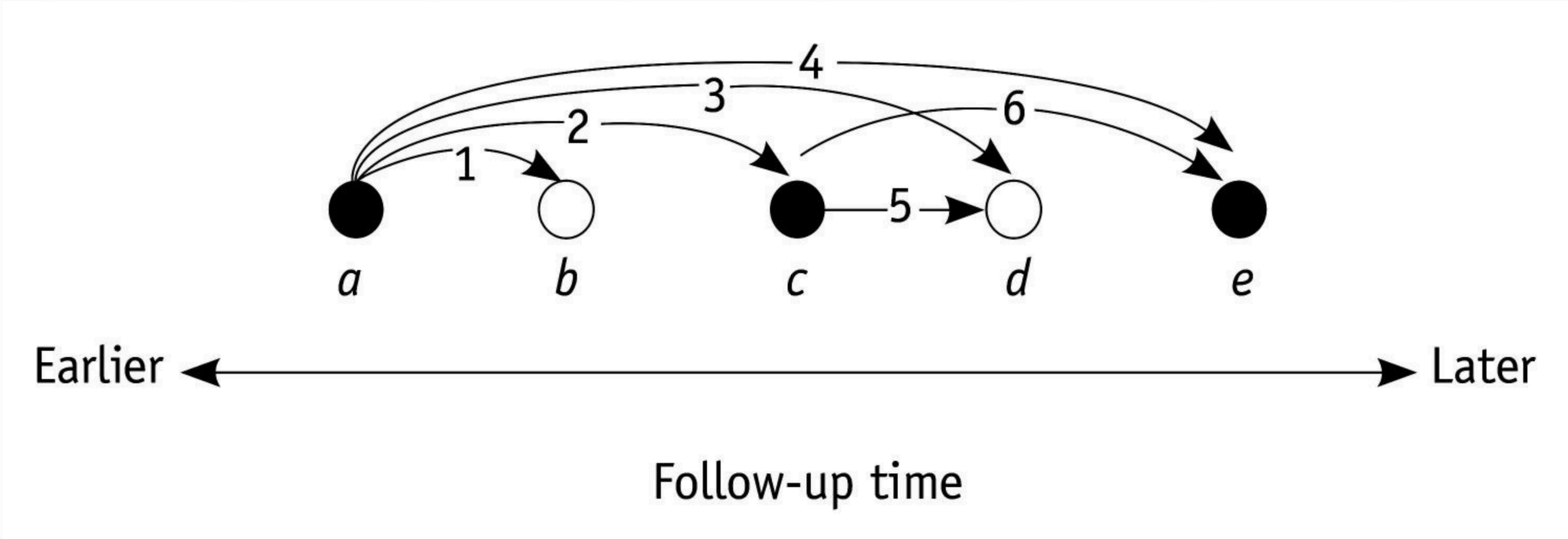

Global survival curve based on train.csv
1.000
0.975
0.950
0.925
0.900
Overall survival probability
0
25
50
75
Time (days)


Empirical hazard function estimate (GP)
0.0015
0.0010
0.0005
0.0000
Hazard
0
25
50
75
Time (days)

Часть 5

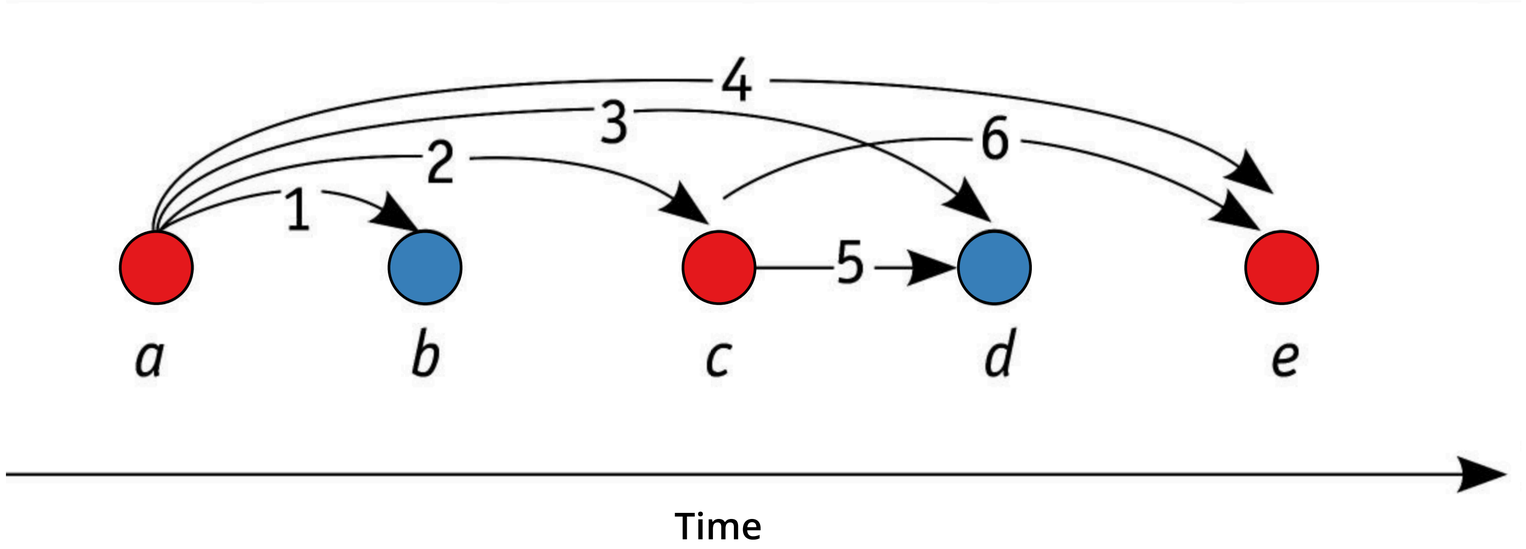
Про метрику

Concordance Index



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8484151/

Concordance Index



Событие наступило (отток)

Событие цензурировано

- 5 клиентов => 10 возможных пар для сравнения
- Только 6 пар из 10 можем сравнивать из-за цензурирования
- Наглядно: **b** и **d** не сравнимы так как оба цензурированы

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8484151/

Часть 6

Куда копать?

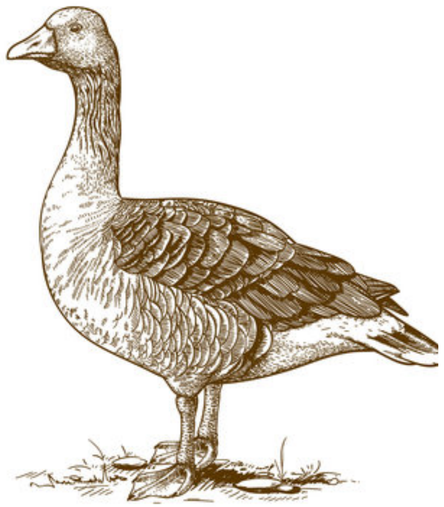
Статьи:

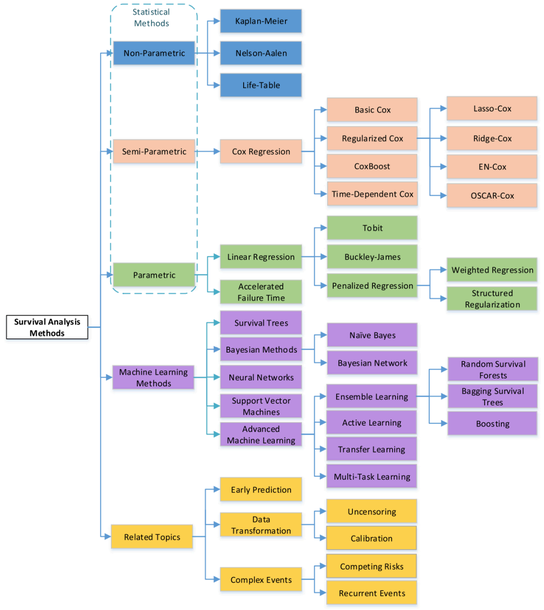
- **[Arxiv, 2023]** *Deep Learning for Survival Analysis: A Review* — качественный обзор современных нейросетевых архитектур и постановок в Time-to-Event задачах.
- **[NCBI, 2021]** *Review of Statistical Methods for Evaluating the Performance of Survival or Other Time-to-Event Prediction Models (from Conventional to Deep Learning Approaches)* — обзор метрик, используемых в Time-to-Event задачах.
- **[ACM, 2019]** *Machine Learning for Survival Analysis: A Survey* — обзор применения машинного обучения в Time-to-Event задачах, может использоваться как пособие для начинающих.
- **[Git, 2016]** *WTTE-RNN - Less hacky churn prediction* — легендарный пост-статья про проблемы задачи предсказания оттока, и как к задаче стоило бы подходить. **(выбор редакции ♥)**
- **[Chalmers Thesis, 2017]** *WTTE-RNN : Weibull Time To Event Recurrent Neural Network A model for sequential prediction of time-to-event in the case of discrete or continuous censored data, recurrent events or time-varying covariates* — "сопроводительный" диссер, по материалам которого была написана статья выше.

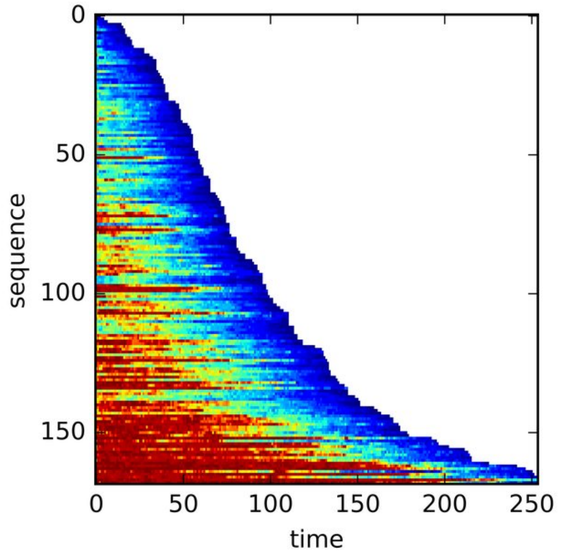
Библиотеки:

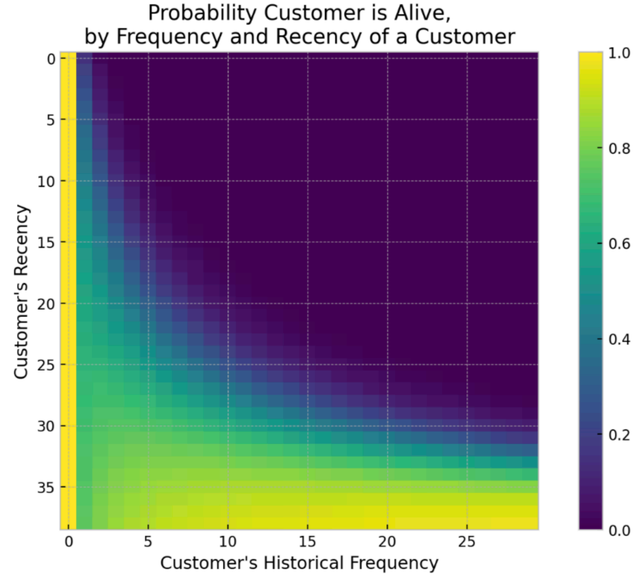
- **Survival analysis with Catboost** — почти не секретный туториал по тому, как запускать Cox и AFT модели на Catboost
- **scikit-survival** — библиотека по Survival / Time-to-Event задача на Python, на основе scikit-learn
- **auton-survival** — актуальная библиотека по Survival / Time-to-Event анализу на Python, включая нейросетевые модели от CMU
- **lifelines** — общая библиотека по классическому Survival Analysis на Python с основным статистическим инструментарием
- **(давно не обновлялась) PySurvival** — библиотека по Survival / Time-to-Event задачам на Python. В том числе с туториалом по оттоку (churn)
- **(давно не обновлялась) PyCox** — библиотека по нейросетевым моделям для Time-to-Event задач на основе pytorch

Куда копать?









1. Включить “гуся”

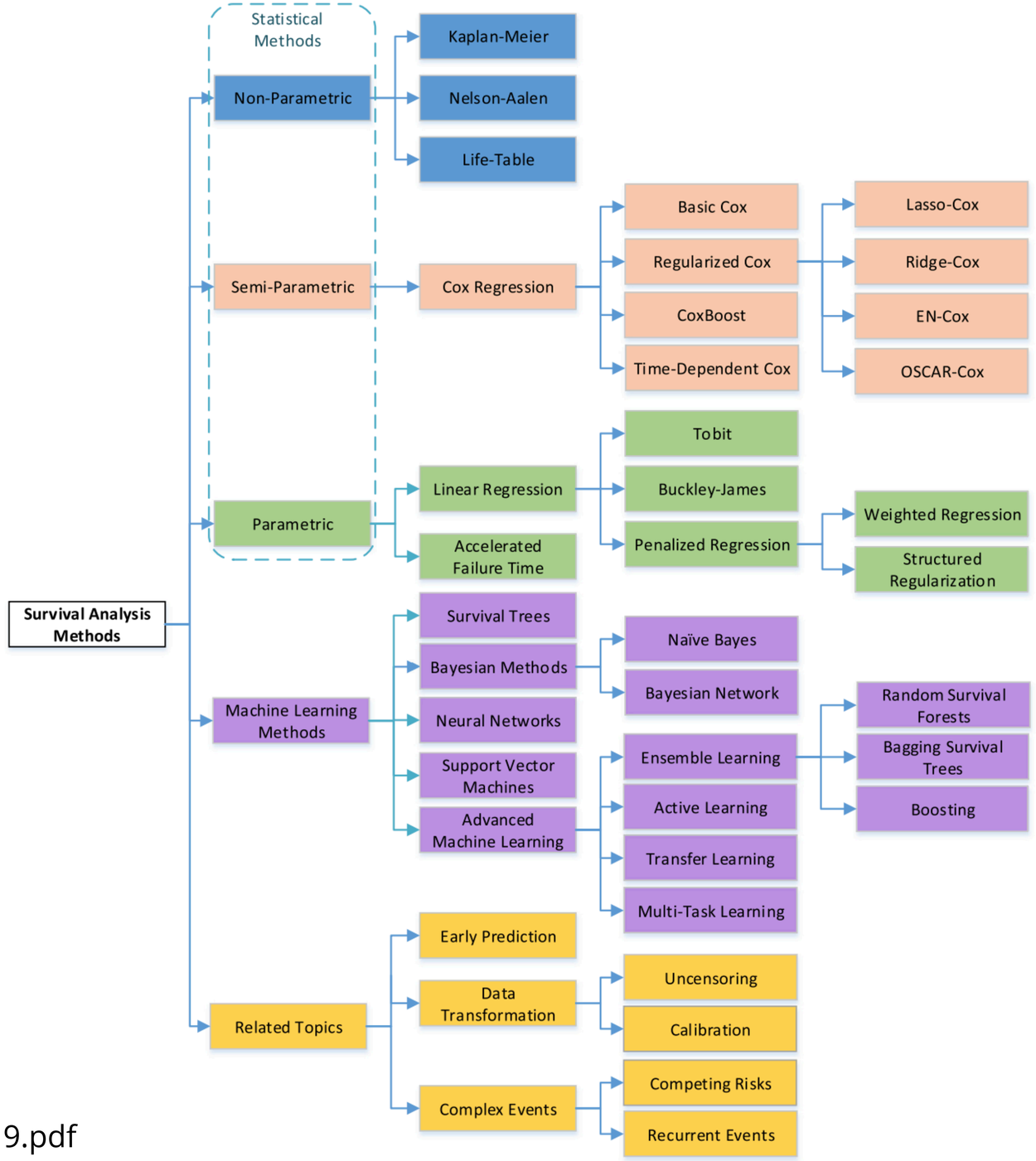
2. Survival ML

3. Survival DL

4. Bayes++ BTYD


В ТРЯСИНЕ ЗАБОТ
Включить гуся
202 m
18
1
0

Survival ML

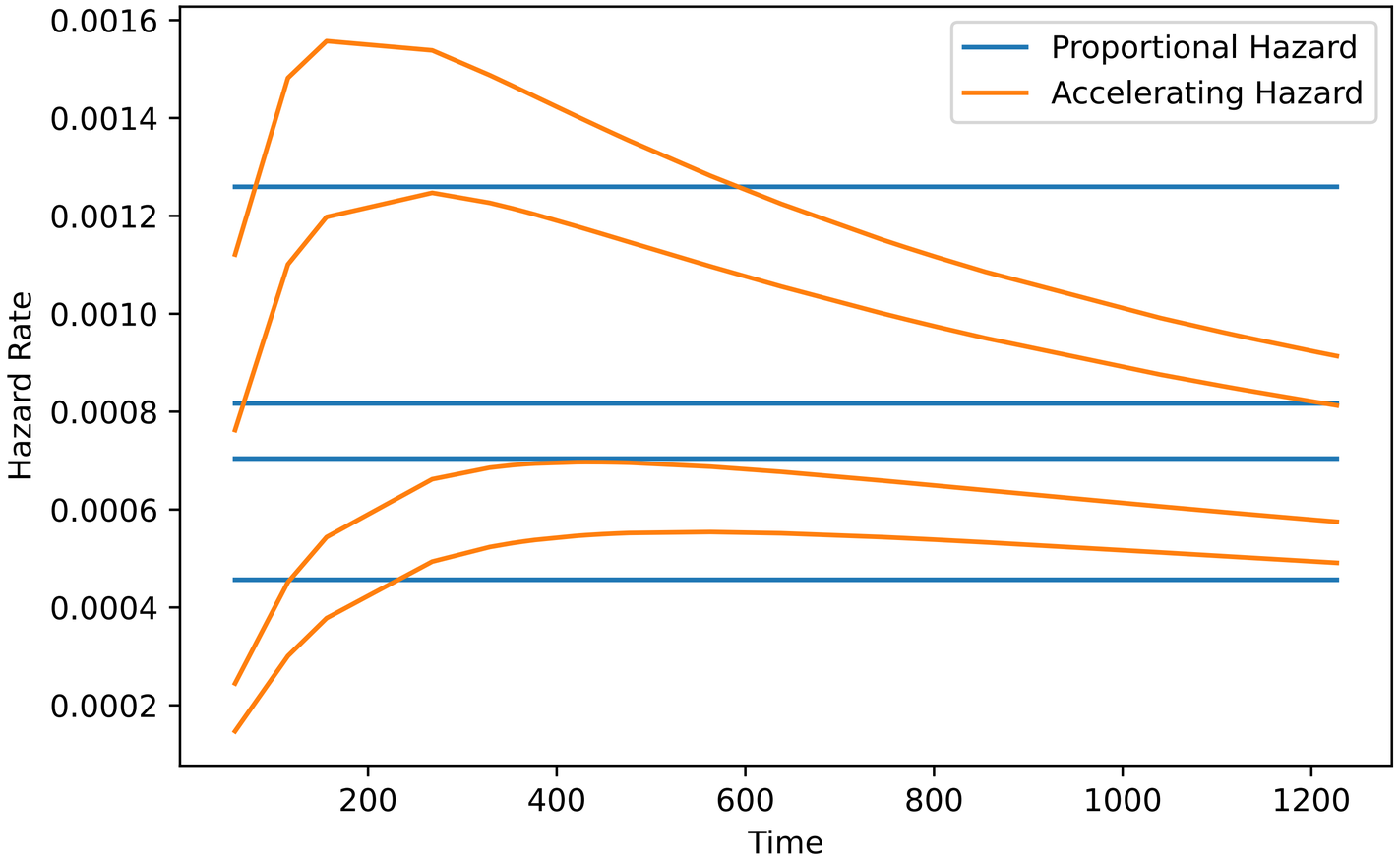


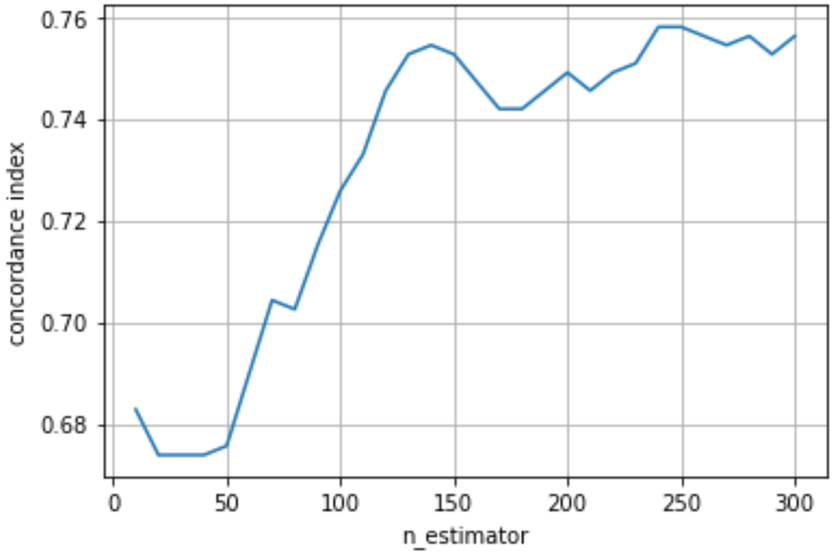
https://dmkd.cs.vt.edu/papers/CSUR19.pdf

Survival ML

https://dmkd.cs.vt.edu/papers/CSUR19.pdf

Survival ML



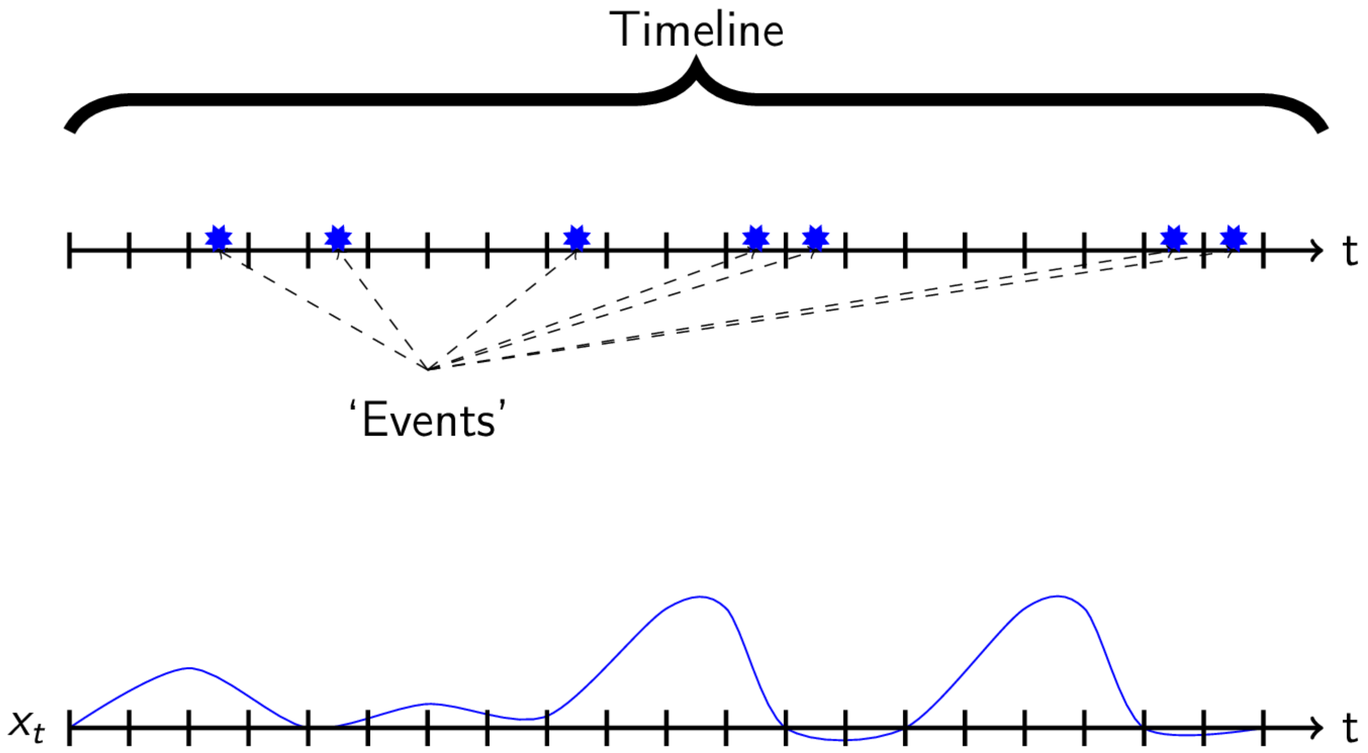


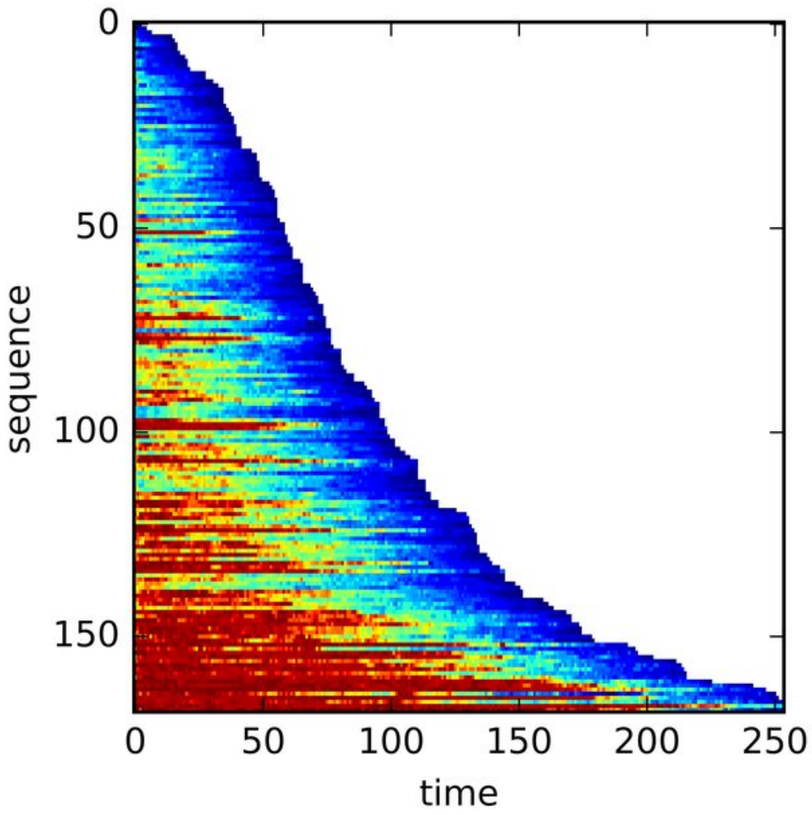
https://scikit-survival.readthedocs.io/en/stable/user_guide/boosting.html

Survival DL

WTTE-RNN - Less hacky churn prediction

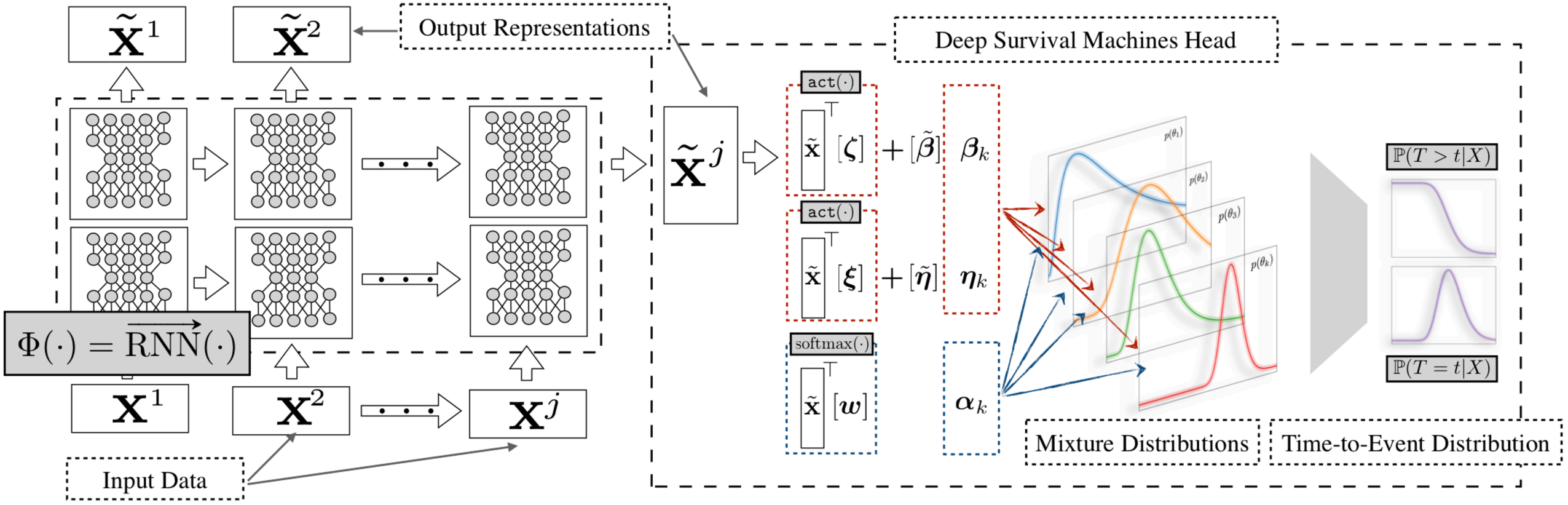
22 Dec 2016





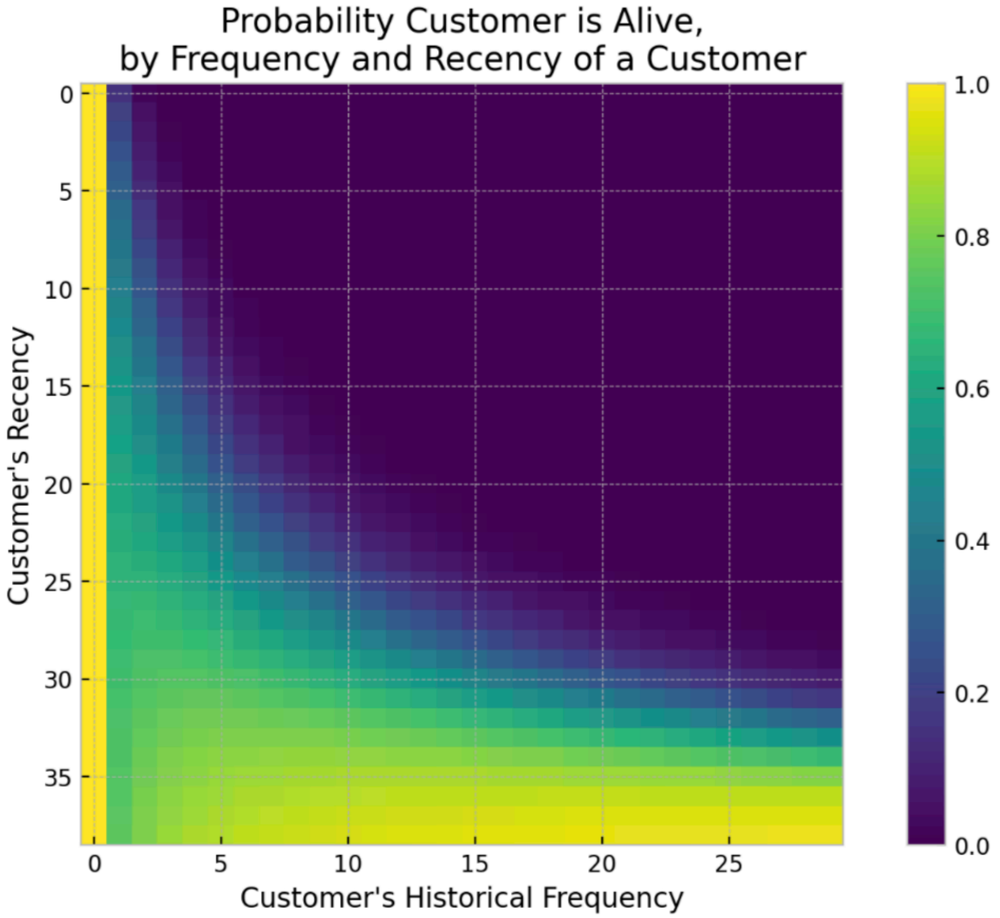
https://ragulpr.github.io/2016/12/22/WTTE-RNN-Hackless-churn-modeling/

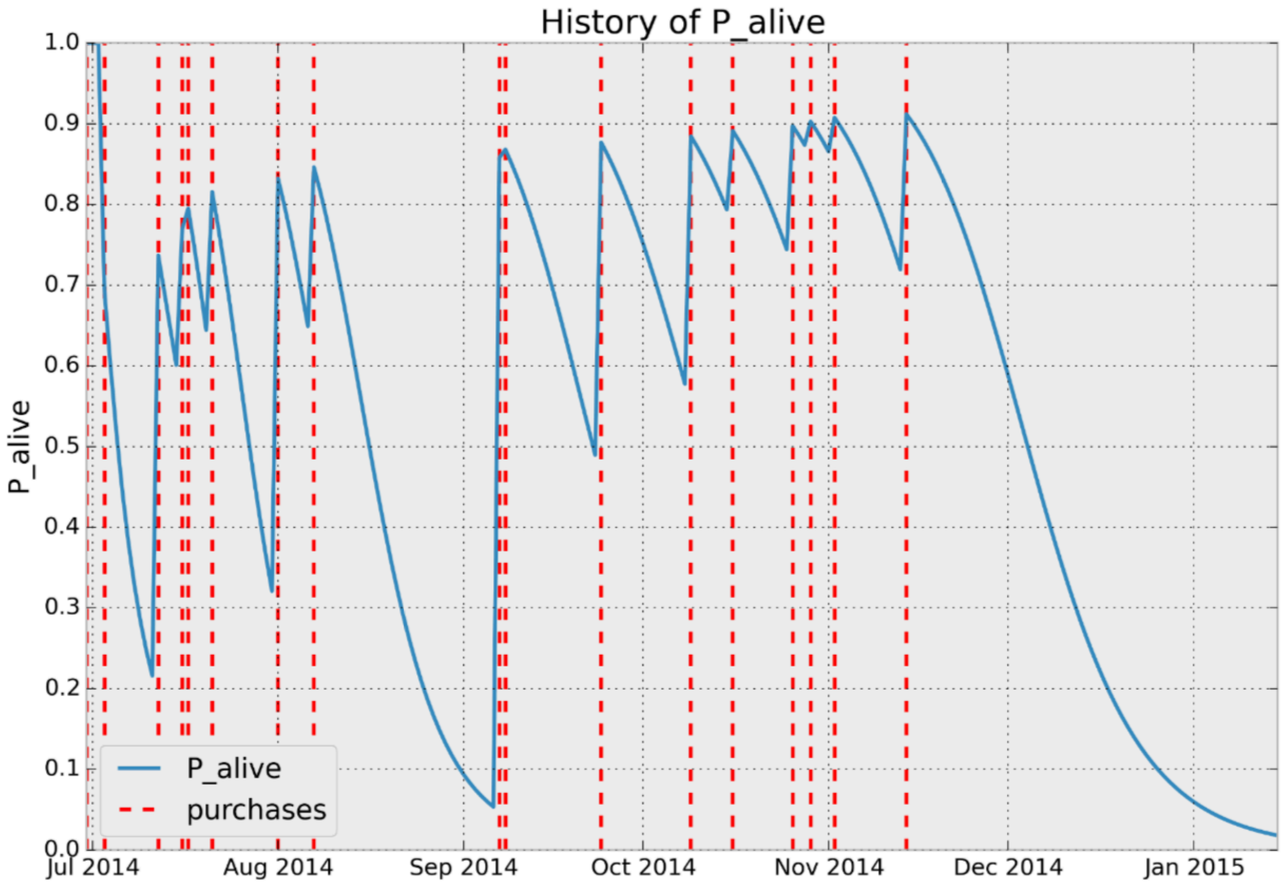
Survival DL



https://autonlab.org/auton-survival/models/dsm/index.html#auton_survival.models.dsm.DeepSurvivalMachines

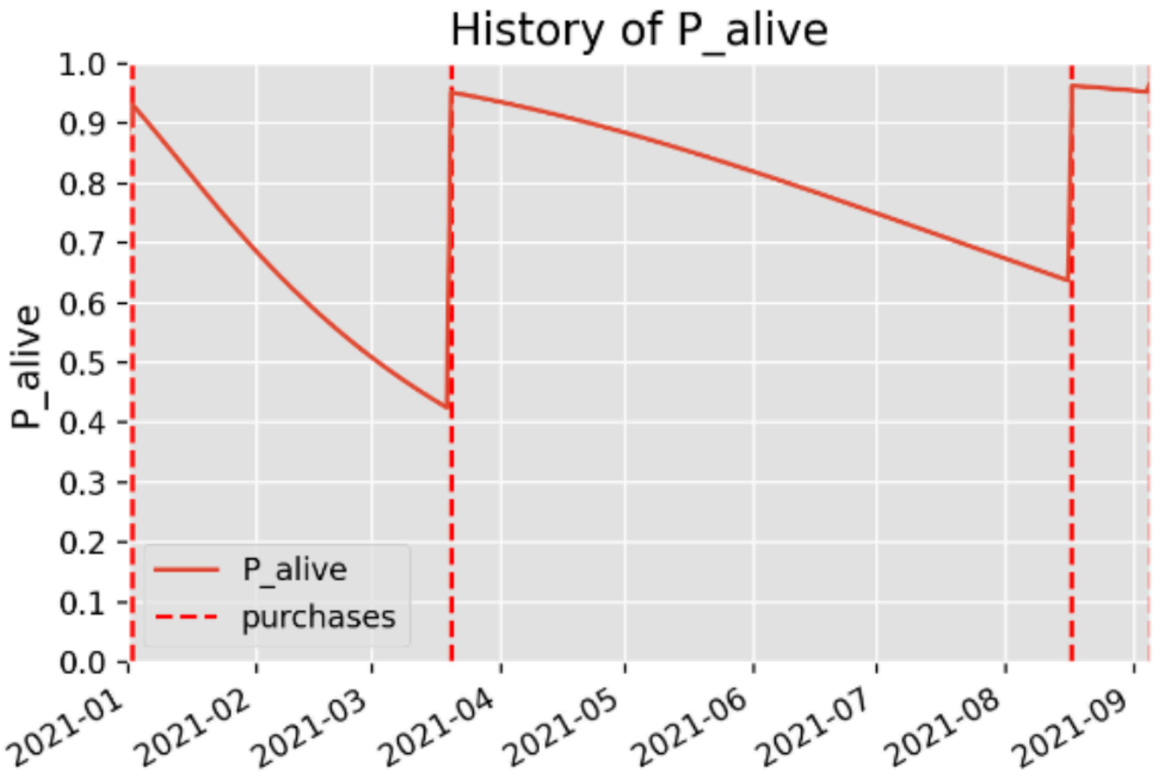
Buy Till You Die (BTYD)





https://juanitorduz.github.io/html/berlin_bayesians_btyd.html#/probability-of-alive

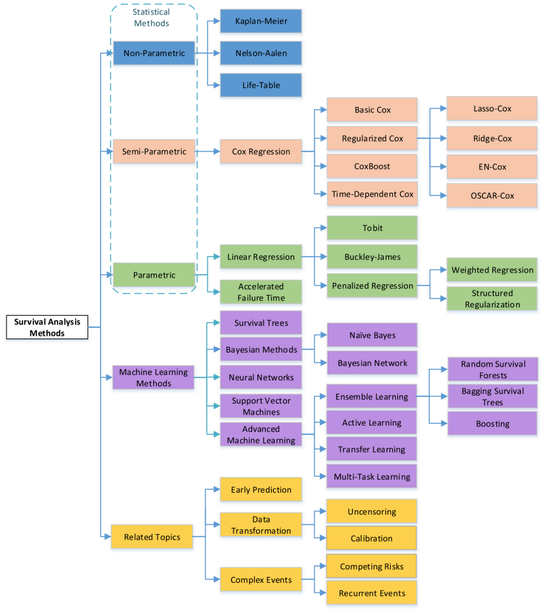
Buy Till You Die (BTYD)



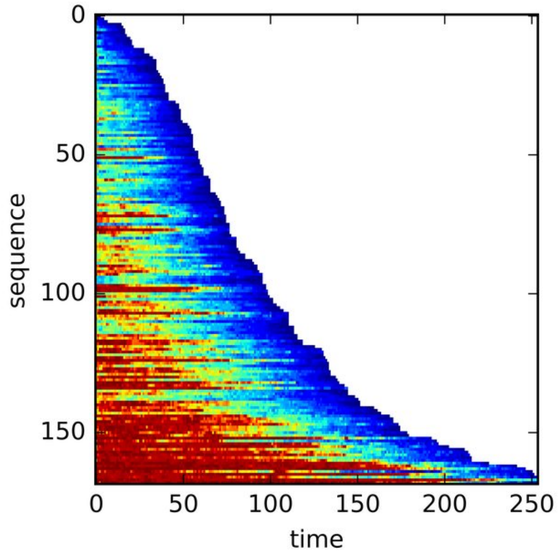
https://juanitorduz.github.io/html/berlin_bayesians_btyd.html#/probability-of-alive-1

Куда копать?

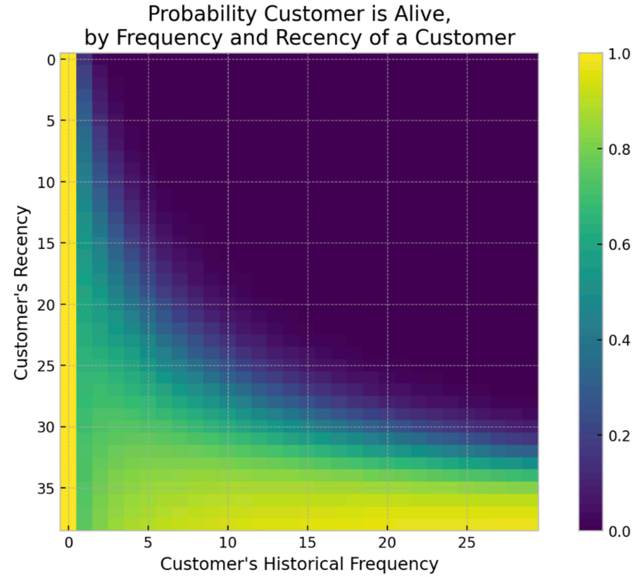
1. Включить “гуся”



2. Survival ML



3. Survival DL



4. Bayes++ BTYD

Спасибо!

Трек соревнований на ODS:
Data Fusion Contest 2024

Telegram:
Data Fusion Contest 2024 Chat

Q&A time


Наташ, делаешь задачу?
На западе уже все решили
и первую
и вторую задачу