



MULTI-STATE CHURN ANALYSIS

WITH A SUBSCRIPTION PRODUCT

GRADIENT

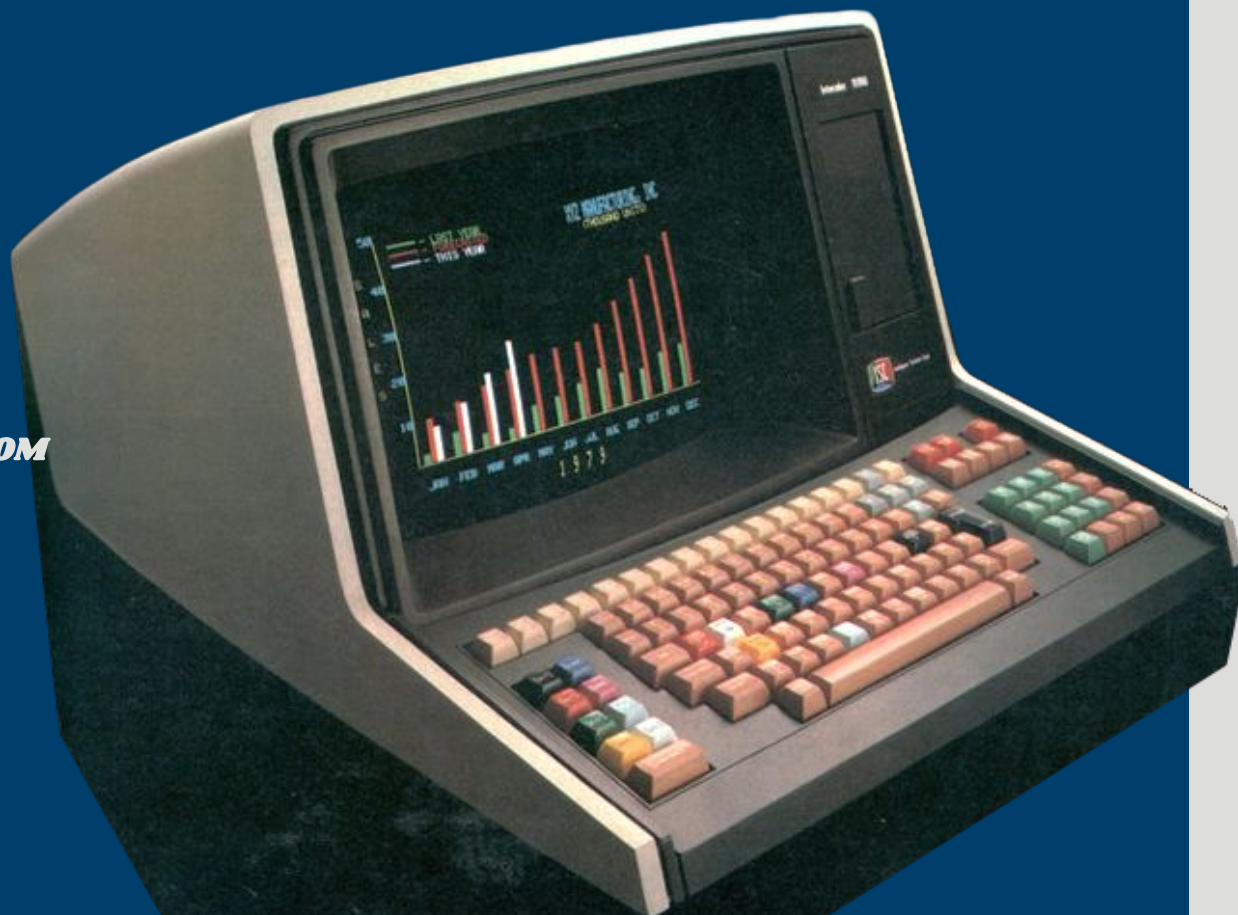
DEVELOPING INTELLIGENCE POWERED BY DATA

WHO IS THIS GUY?

MARCIN KOSIŃSKI

- ***WARSAW RUG***
- ***R BLOGGER*** *R-ADDICT.COM*
- ***WHYR2018.PL***

MARCIN@GRADIENTMETRICS.COM



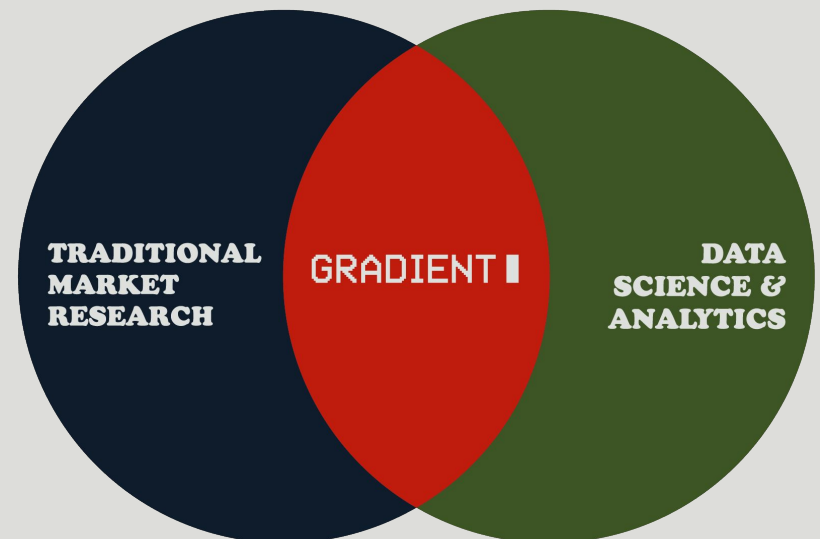
***Nice to
meet you!***



WE'RE GRADIENT:

A crew of quantitative marketers and technologists that gather hard data and build robust statistical models to guide organizations through their most difficult decisions.

We're confirmed data geeks, but word on the street is that we're easy to work with and pretty fun, too.



SURVIVAL ANALYSIS

DEFINITION & EXAMPLES

LET'S START TALKING

A branch of statistics for analyzing the **expected duration of time until** one or more **events** happen.

Examples

1. A death of the patient.
2. A deactivation of the service.
3. An accident on the road.
4. The device failure.
5. An employee leaving the company.
6. A customer cancelling subscription.



SURVIVAL ANALYSIS

QUESTIONS IT (MIGHT) ANSWER

LET'S START ASKING

What's the probability an event will (not) occur after a specific period of time?

Which characteristics indicate a reduced or increased risk of occurrence of an event?

What periods of time are most (or least) exposed to the risk of an event?



SURVIVAL ANALYSIS

CHALLENGES IT FACES

DEPENDENDING ON THE SCENARIO

Data

1. Censoring.
2. Interval data.
3. Observations may not be independent.
4. Time varying features.

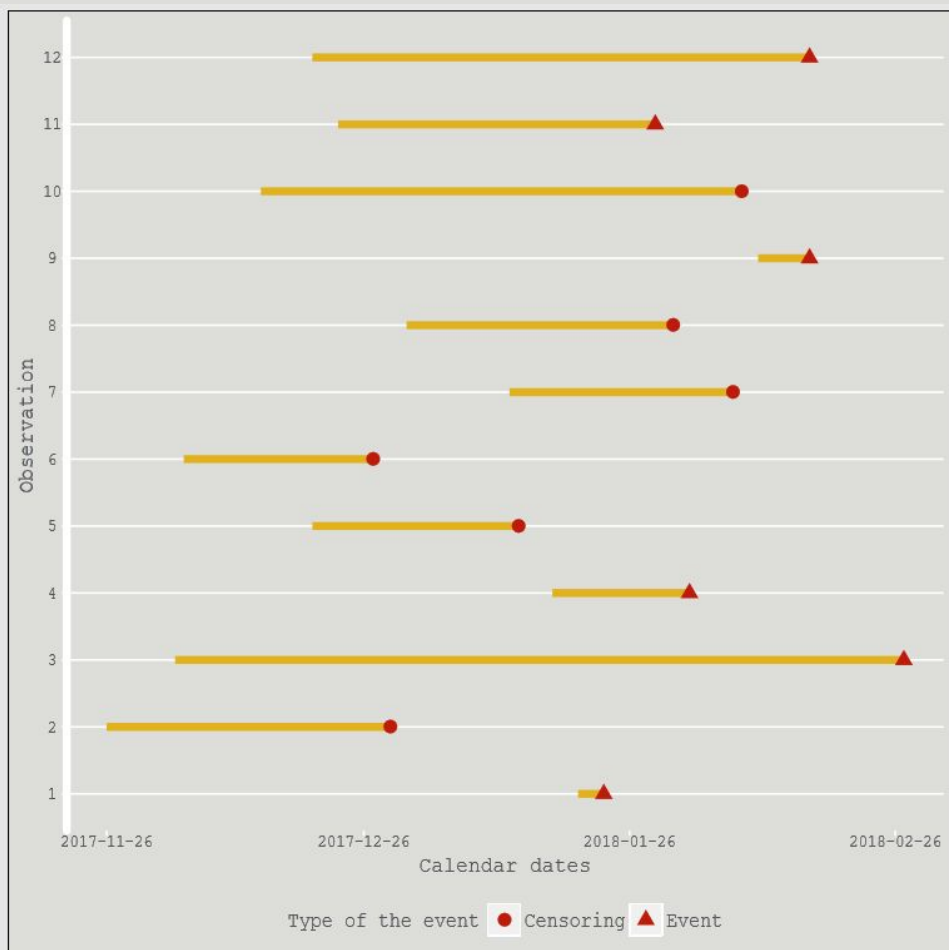
Events

1. Recurring events - one event might occur multiple times.
2. Competing risks - one of multiple events might occur.
3. A multi-state (cyclic/acyclic) nature of the process.



DATA STRUCTURE

SIMPLE CASE



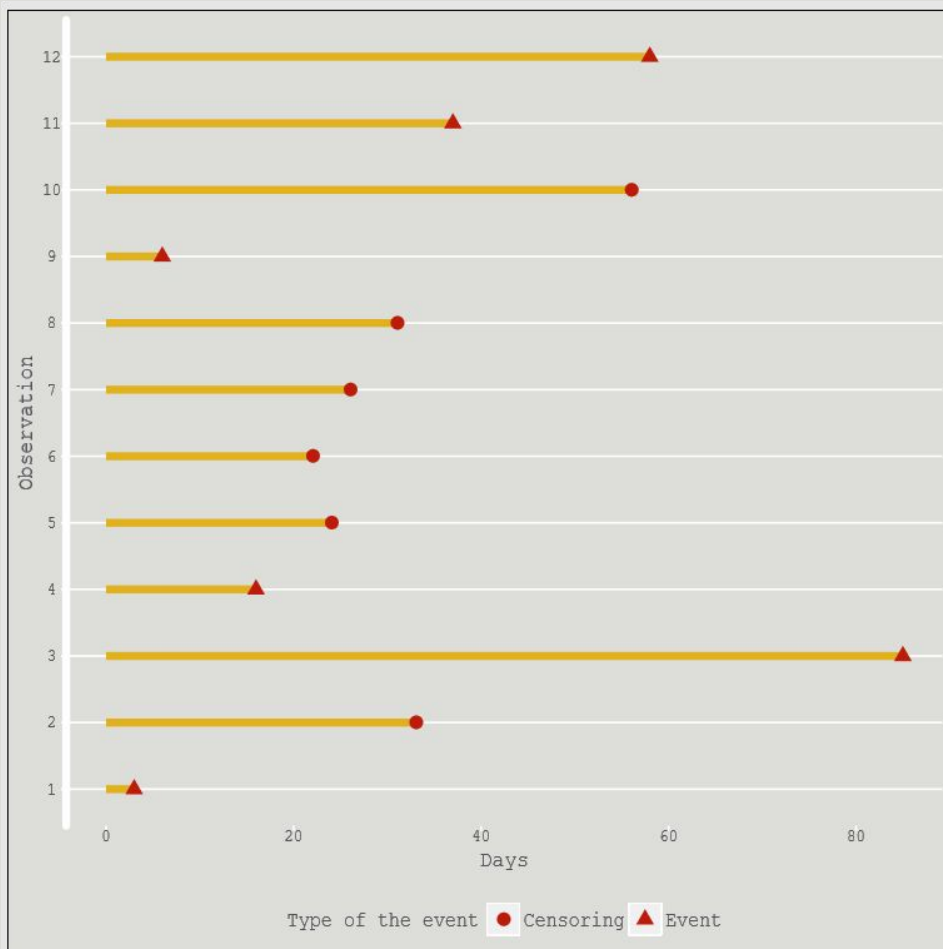
HEAD OF THE DATA

ID	Start Date	End Date	Status
1	2018-01-28	2018-02-22	Censoring
2	2017-12-16	2018-01-08	Event
3	2017-12-09	2018-01-06	Censoring
4	2018-01-16	2018-02-23	Censoring
5	2017-12-16	2018-02-11	Event
6	2018-02-18	2018-03-01	Event

Data **do not** correspond to the plot.

DATA STRUCTURE

SIMPLE CASE



HEAD OF THE DATA

ID	Time	Status
1	3 days	Event
2	33 days	Censoring
3	85 days	Event
4	16 days	Event
5	24 days	Censoring
6	22 days	Censoring

Data **do** correspond to the plot.

KAPLAN-MEIER ESTIMATES

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$

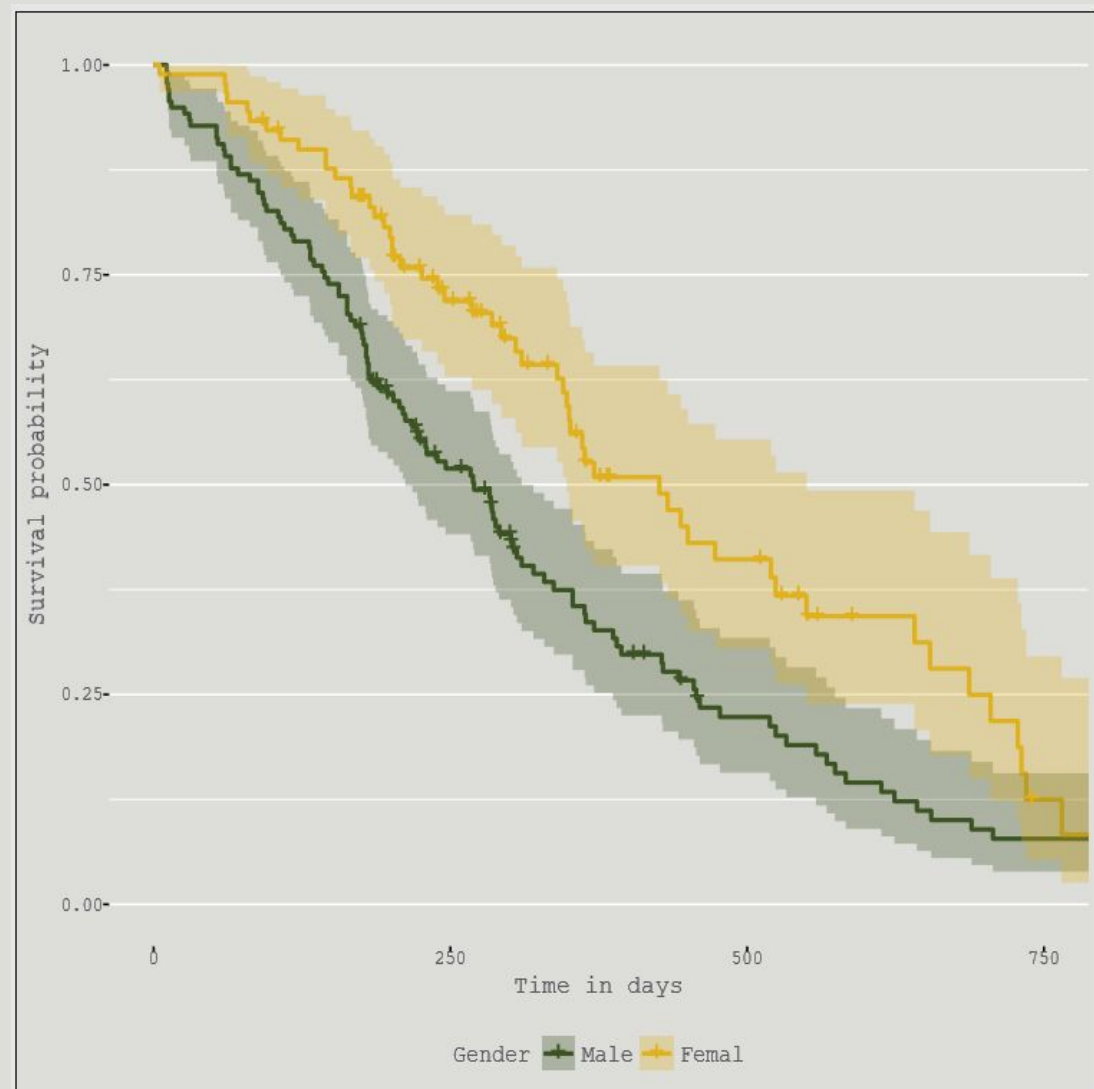
where

t_i - time of i-th event

n_i - number of observations
in a risk set at time t_i

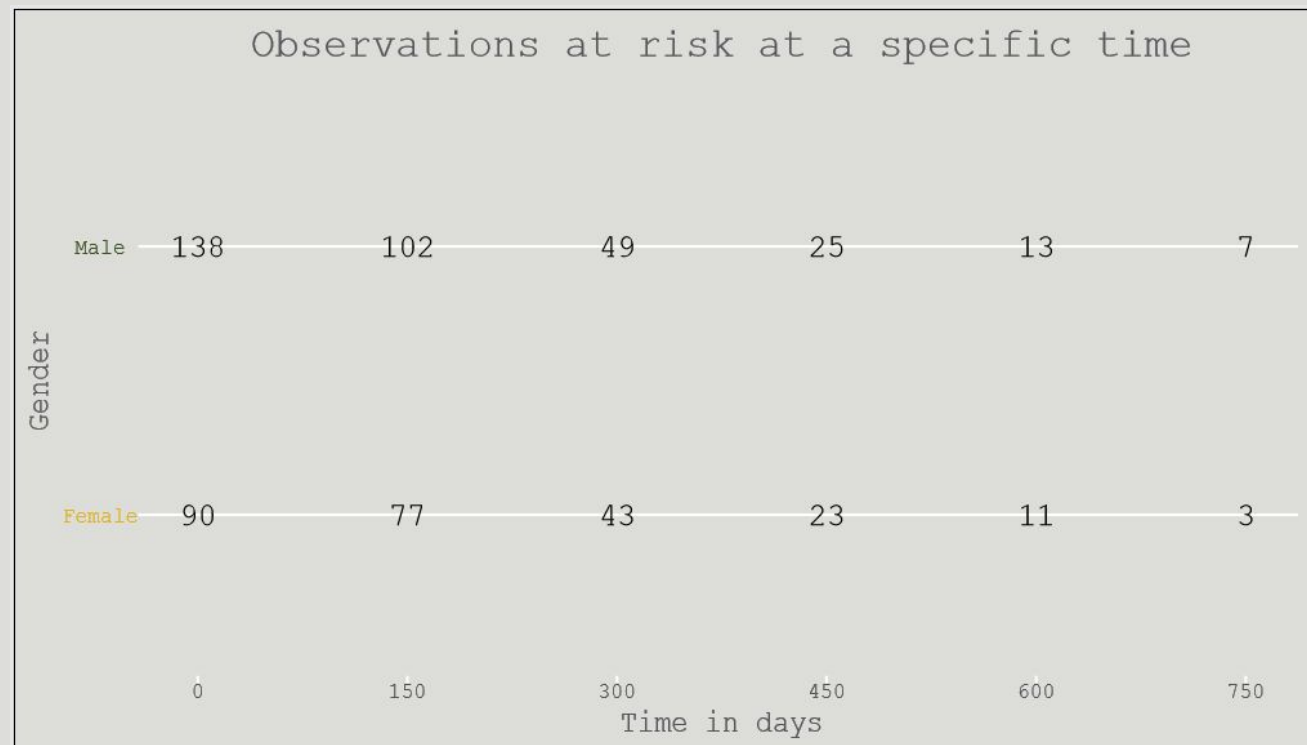
d_i - number of events at t_i

Log-rank test seeks for statistically significant differences between curves.

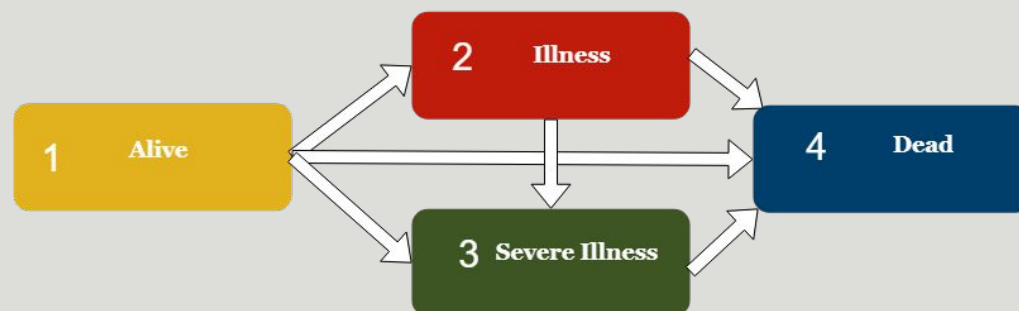
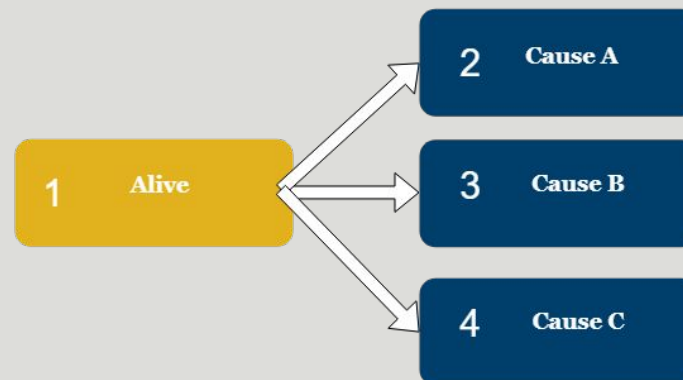
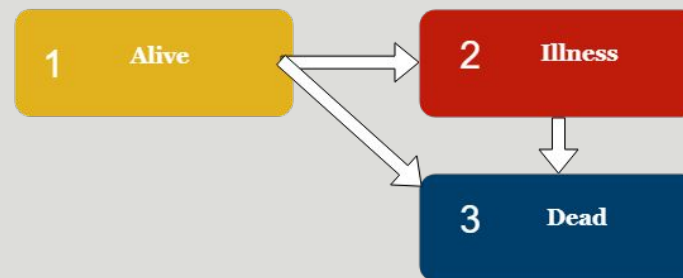
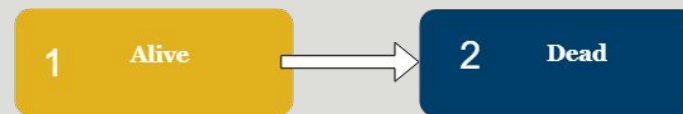
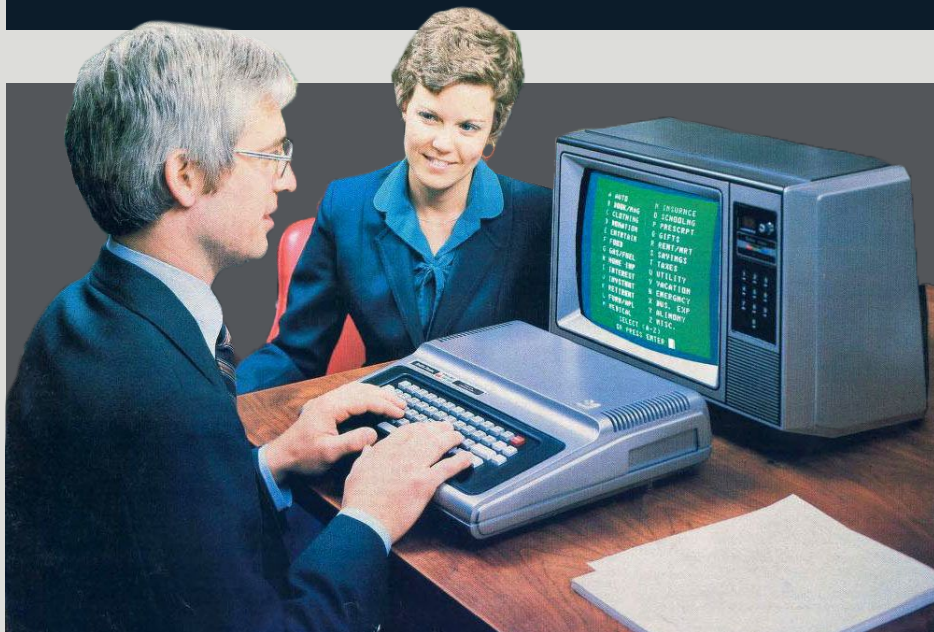


SURVIVORS AT A TIME

Useful when considering whether results at a specific time point are significant due to the sample size.

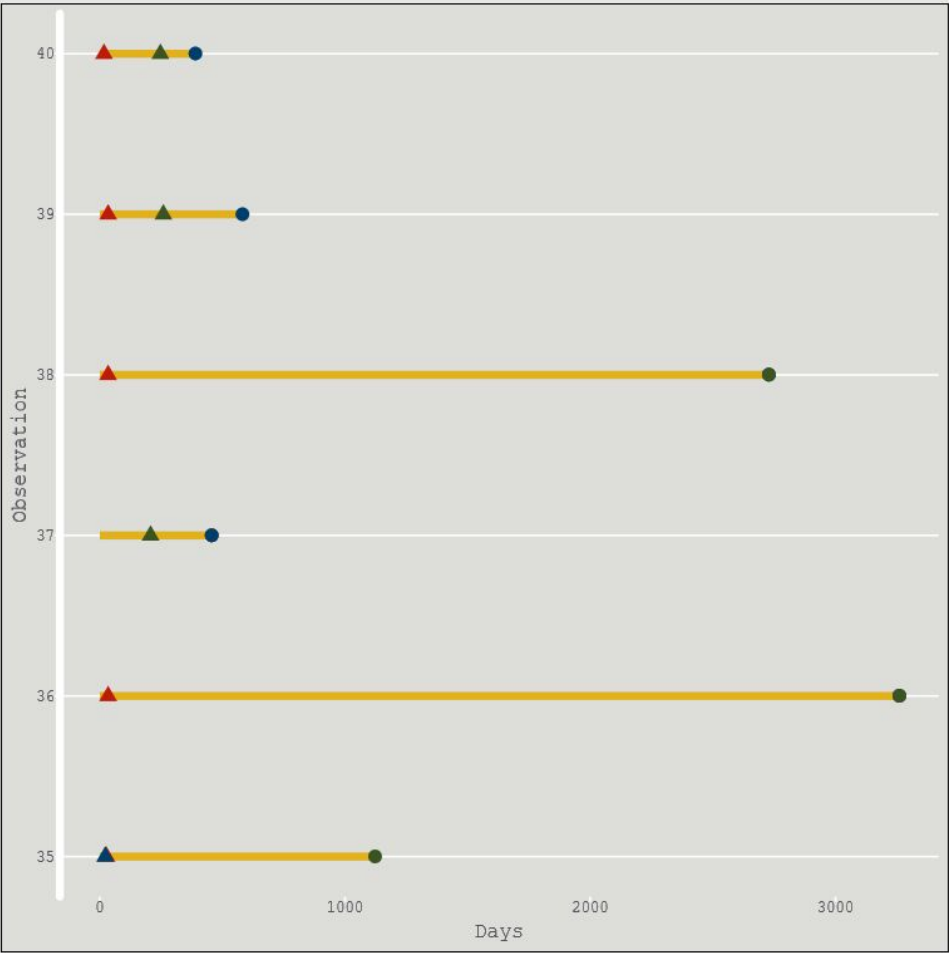


MULTI-STATE MODELS



DATA STRUCTURE

MULTI-STATE CASE

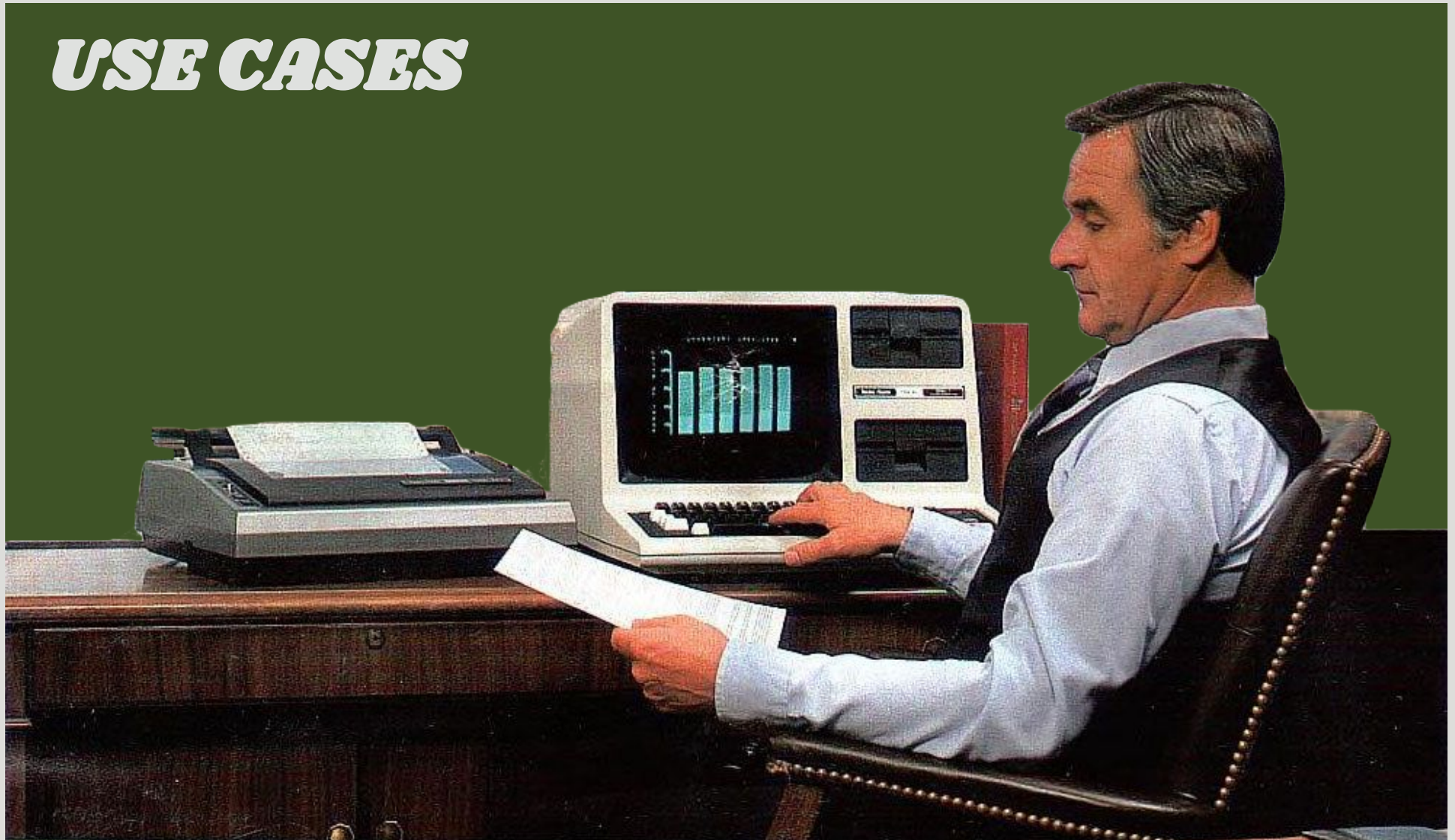


HEAD OF THE DATA

ID	Time 1	Event 1	Time 2	Event 2	Time 3	Event 3
1	22	1	995	0	995	0
2	29	1	12	1	422	1
3	1264	0	27	1	1264	0
4	50	1	42	1	84	1
5	22	1	1133	0	114	1
6	33	1	27	1	1427	0

Demonstrational data.

USE CASES



1 EVENT / COX PROPORTIONAL HAZARDS

COX METHODOLOGY OVERVIEW

1. Proportional hazards assumptions.
2. Functional form of continuous variables.
3. Independent observations.
4. Independent censoring from the mechanism that rules of event's times.
5. Non informative censoring
- does not give an information on parameters of the time distribution of events because it does not depend on them

```
coxph(Surv(futime, fustat) ~ age + ecog.ps + rx, data=ovarian)
```

NOTE

One can use accelerated failure time (AFT) models.

EXAMPLE COEFFICIENTS

variable	coef	exp(coef)
age	0.15	1.16
ecog.ps	0.10	1.11
rx	-0.81	0.44

DIAGNOSTIC PLOTS

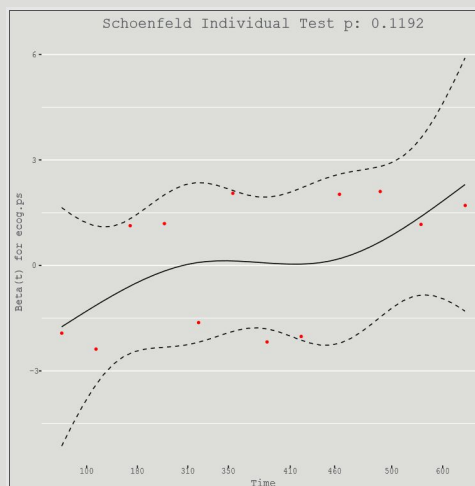
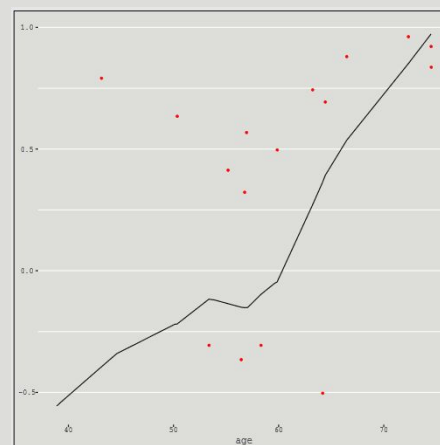


Fig. 1: Shoenfeld residuals.



OVARIAN DATA

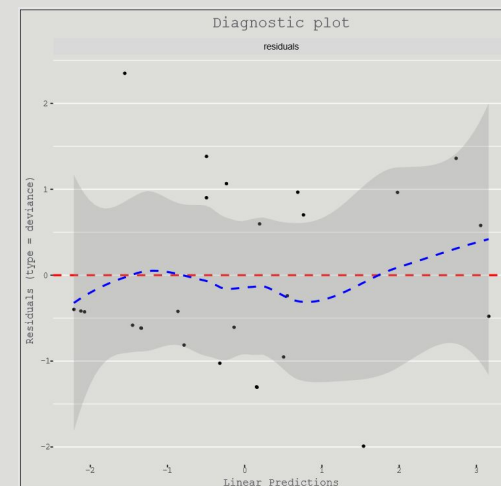


Fig. 2: Deviance residuals.

FUNCTIONS (survminer)

1. `ggcoxzph`
2. `ggcoxdiagnostics`
3. `ggcoxfunctional`

Fig. 3: Martingale residuals.

NEVENTS (ACYCLIC) MULTI-STATE MODEL

TRANSITION MATRIX

to

from	1	2	3	4	5
1	NA	1	2	NA	3
2	NA	NA	NA	4	5
3	NA	NA	NA	6	7
4	NA	NA	NA	NA	8
5	NA	NA	NA	NA	NA

NA = transition not possible

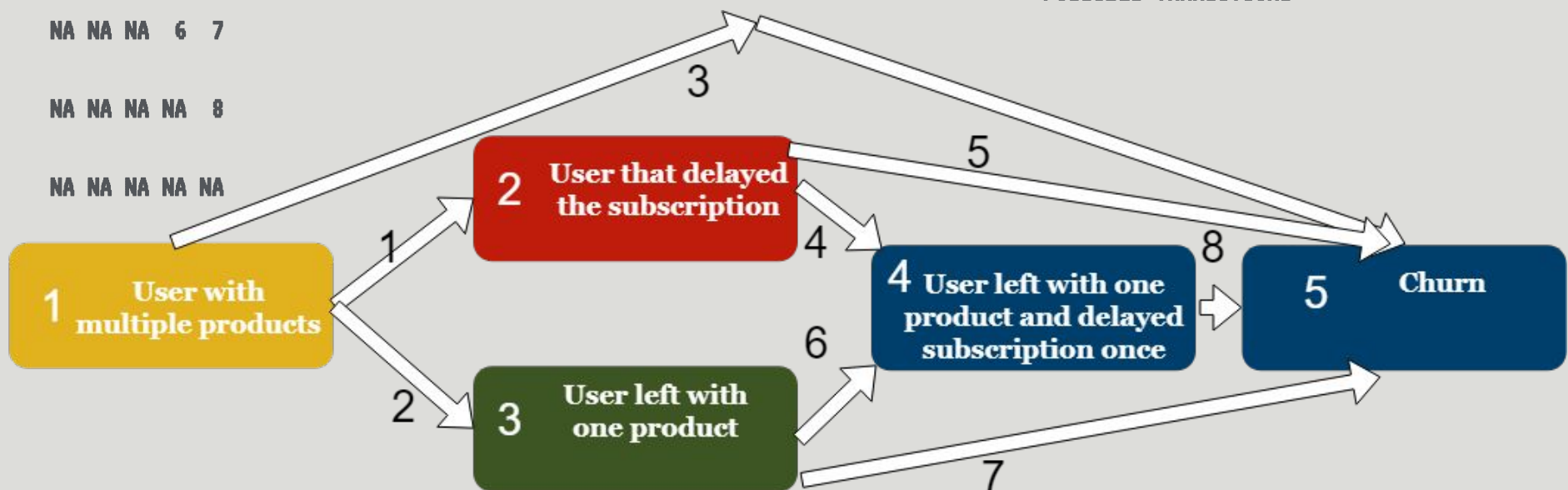
numbers in cells

=

names of transitions

The most complicated part is the proper data coding for the model's input.

POSSIBLE TRANSITIONS



NEVENTS (ACYCLIC) MULTI-STATE MODEL

SOME COEFFICIENTS

transition	age=>40	age=20-40	discount=yes	gender=female	year=2008-2012	year=2013-2017
1	-1.15	-0.77	-0.26	-0.72	0.80	0.94
2	-1.34	-0.72	-0.15	-0.58	0.39	0.31
3	-0.43	-0.04	0.08	-0.53	0.02	-0.11
4	-0.86	-0.66	-0.09	-0.22	0.13	0.23
5	0.14	-0.64	0.14	-0.24	-0.54	-0.63
6	-1.65	-1.23	0.24	-0.35	0.88	1.33
7	-0.82	-0.57	0.39	-0.57	-0.35	0.09

Reference level for

- age - below 20
- year - 2002-2007

NEVENTS (ACYCLIC) MULTI-STATE MODEL

PREDICTIONS OF THE STATE

Depending on the customer features, the predictions of being in a state after particular time are different.

Credits for modeling:

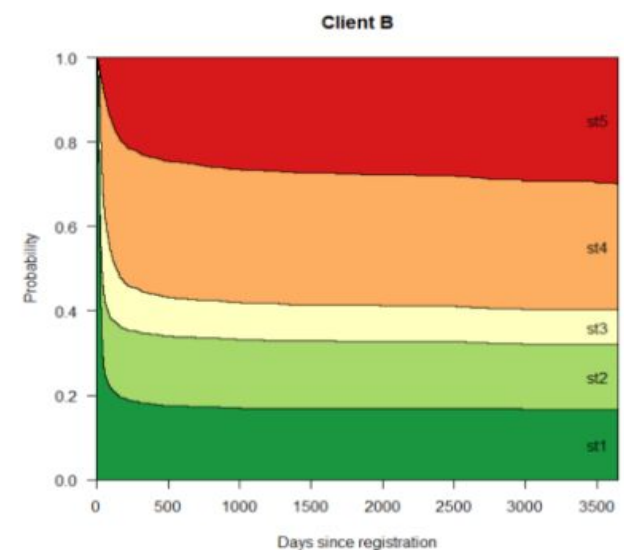
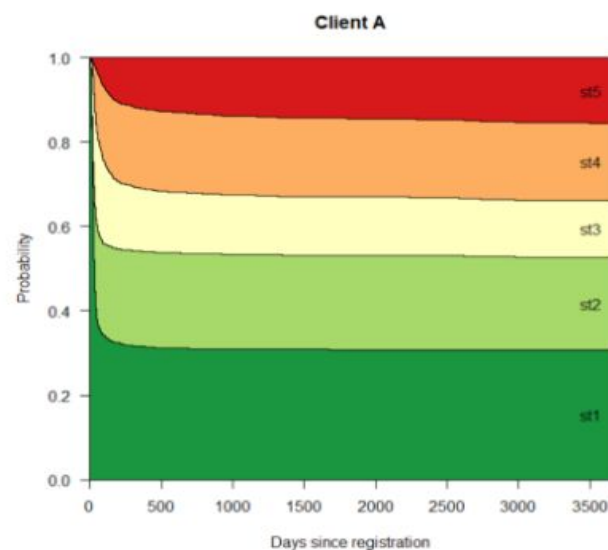
cran.r-project.org/package=mstate

Customer A

- Discount: Yes
- Gender: Female
- Joined: 2013-2017
- Age: Younger than 20

Customer B

- Discount: No
- Gender: Male
- Joined: 2002-2007
- Age: 20-40



NOTES

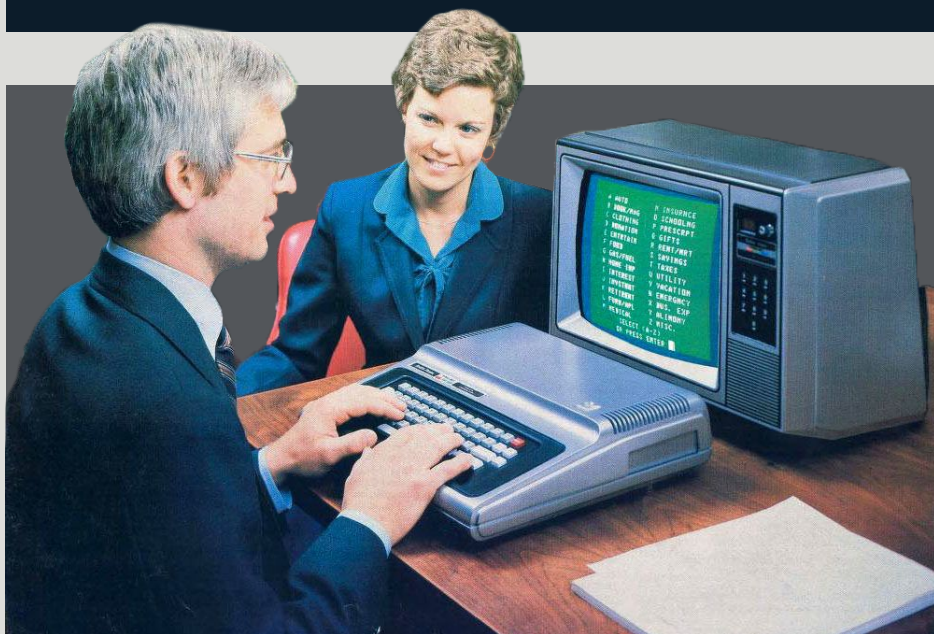


Model assumptions should be considered for every possible transition.

Time varying variables can be taken into the account when handling subscription based data.

Playing with cyclic models requires domain knowledge in (sub) Markov Chain field. ■

PLOTS BASED ON SURVMINER



Credits:

cran.r-project.org/package=survminer

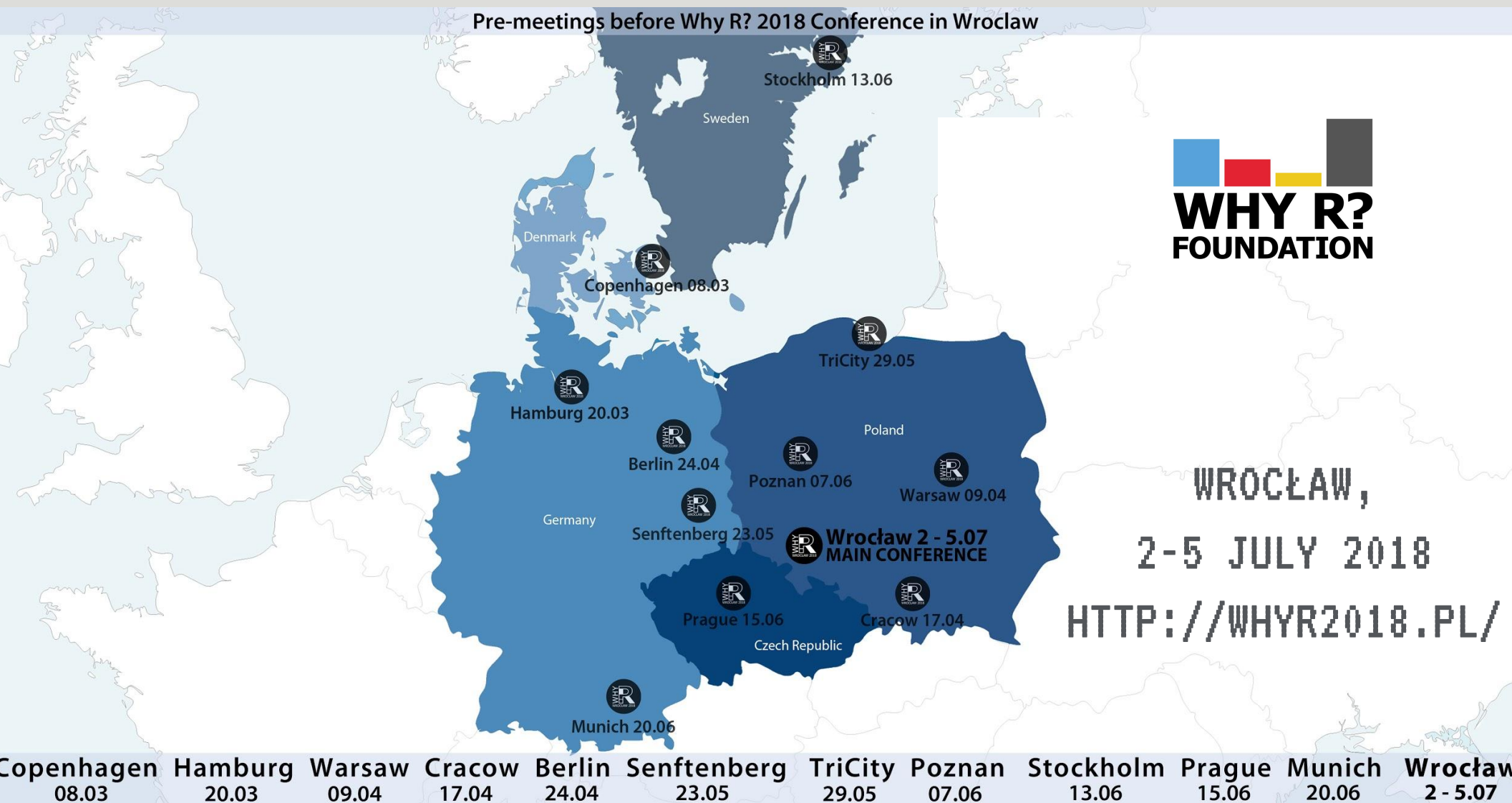
github.com/kassambara/survminer

www.ggplot2-exts.org/gallery/

stdha.com/english/rpkgs/survminer

GRADIENT ■

DID YOU LIKE THE TALK? JOIN US AT **WHY R? 2018 CONFERENCE.**



THANK YOU FOR THE ATTENTION

Presentation and codes

github.com/g6t/mchurn ■