DATA 596: Survival Analysis Project #2: Github Projects

Ellann Cohen & Endy Zarate Vasques, Spring 2025

What is Github? Open Source Software?

GitHub

- **Social code hosting**
- Repository hosting services provides features such as issue-tracking or pull request
- Support to help to evolve **Open Source Software** projects and manage the community behind the projects.

Open Source Software

- Computer software that is **free** and available for the **public to use, modify, or redistribute.**
- Often developed and maintained by **volunteers**

Goal

- Study project abandonment overtime
- Study the evolution of projects
- Raise the problem of sustainability in Open Source

Data & Methodology: Overview

Observational study

1127 open source github package repositories started in 2016, tracked through Oct 2021

- Four specific ecosystems
- Mix of user led and organization led projects
- Mix of number of users working on projects
- Must have been update once in 2016

Collection

- Used Github search API to gather projects
- Project analysis was done to create the other datasets we did not cover

Variables	Type	Unique values	Unit
Ecosystem	Factor	Laravel, NPM, R, Wordpress	-
Repo Type	Factor	Organization, User	-
User Group Size	Factor	1, 2, 3	-
Time	Response variable	0 - 70	months
Status	Response variable	1 (dead), 0 (alive/zombie)	-

Data & Methodology: Issues & Challenges

Dead or Alive?

- Dead: Abandoned with no human activity
- Zombie: Botted and non-coding

Possible issues

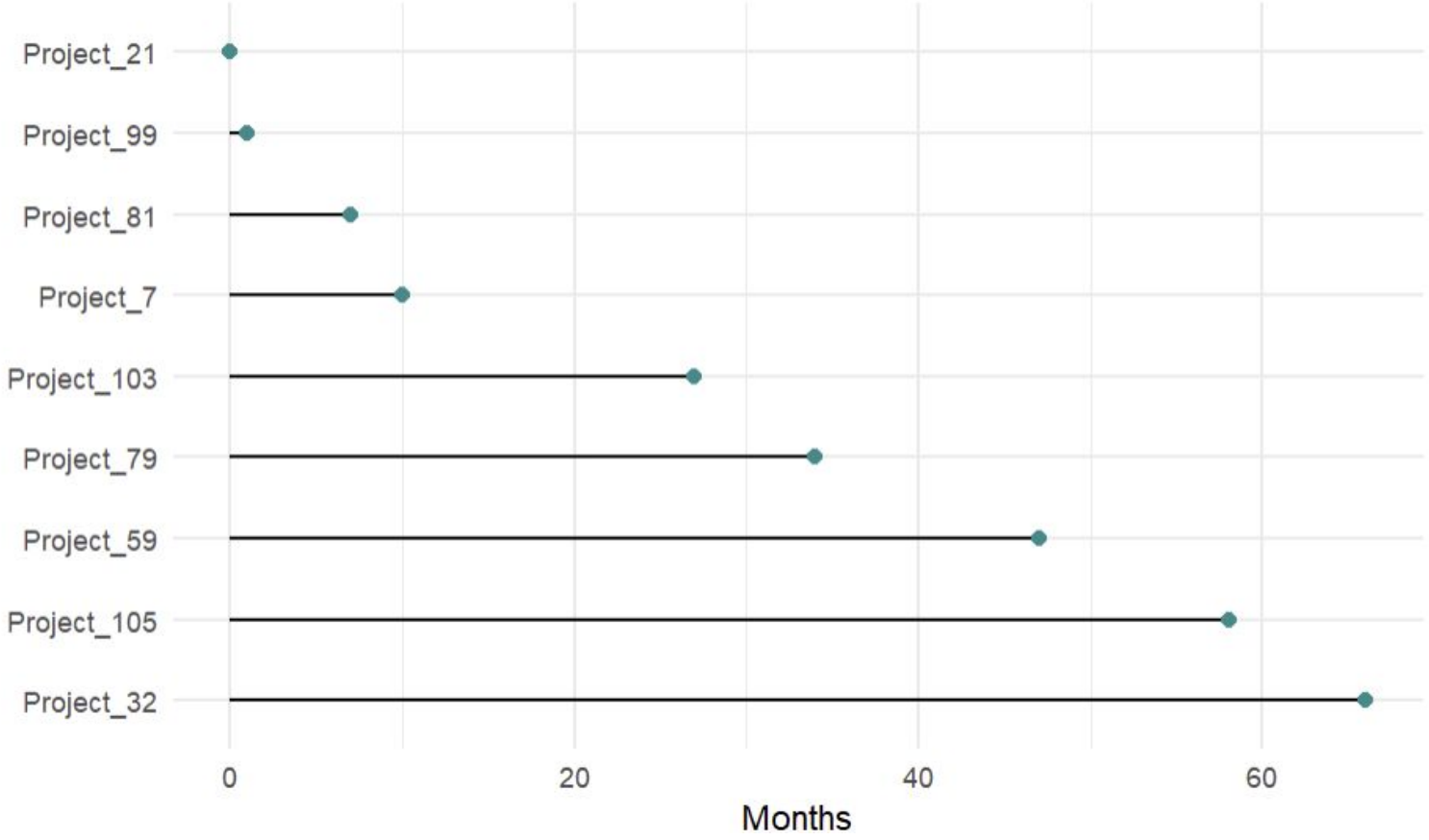
- Zombie projects - categorized as alive
- Only four ecosystems analyzed
- Some projects have a pre-planned end point
- Right censoring occurs
- Some projects did not survive beyond a month

Data & Methodology: Sample Sizes

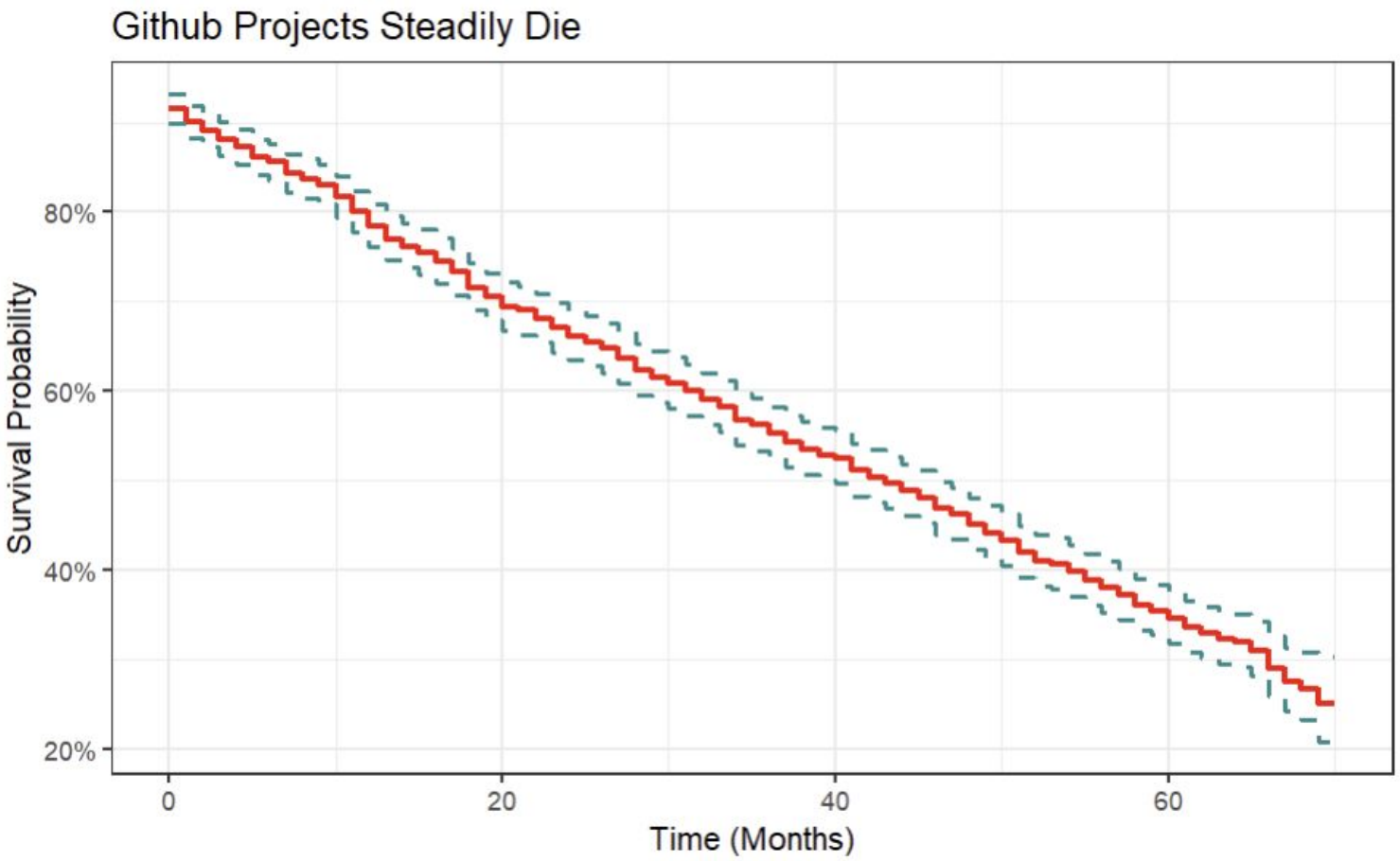
Ecosystem	Repo Type: Organization			Repo Type: User			TOTAL
	UserSize 1	UserSize 2	UserSize 3	UserSize 1	UserSize 2	UserSize 3	-
Laravel	9	21	13	30	21	14	***108***
NPM	11	31	31	62	95	50	***280***
R	24	60	38	36	28	13	***199***
WordPress	52	73	85	196	70	64	***540***
TOTAL	***96***	***185***	***167***	***324***	***214***	***141***	***1127***
	448			***679***			

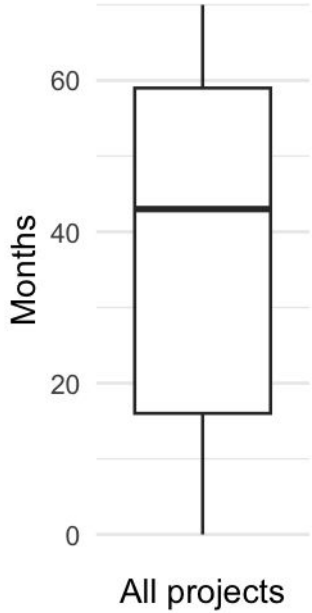
UserSize 1: Within smallest 25% of project user sizes (aka below min of interquartile range)
UserSize 2: Within middle 50% of project eser sizes (aka within interquartile range)
UserSize 3: Within top 25% of project user sizes (aka above max of interquartile range)

Data: Event Plot (for a few example projects)

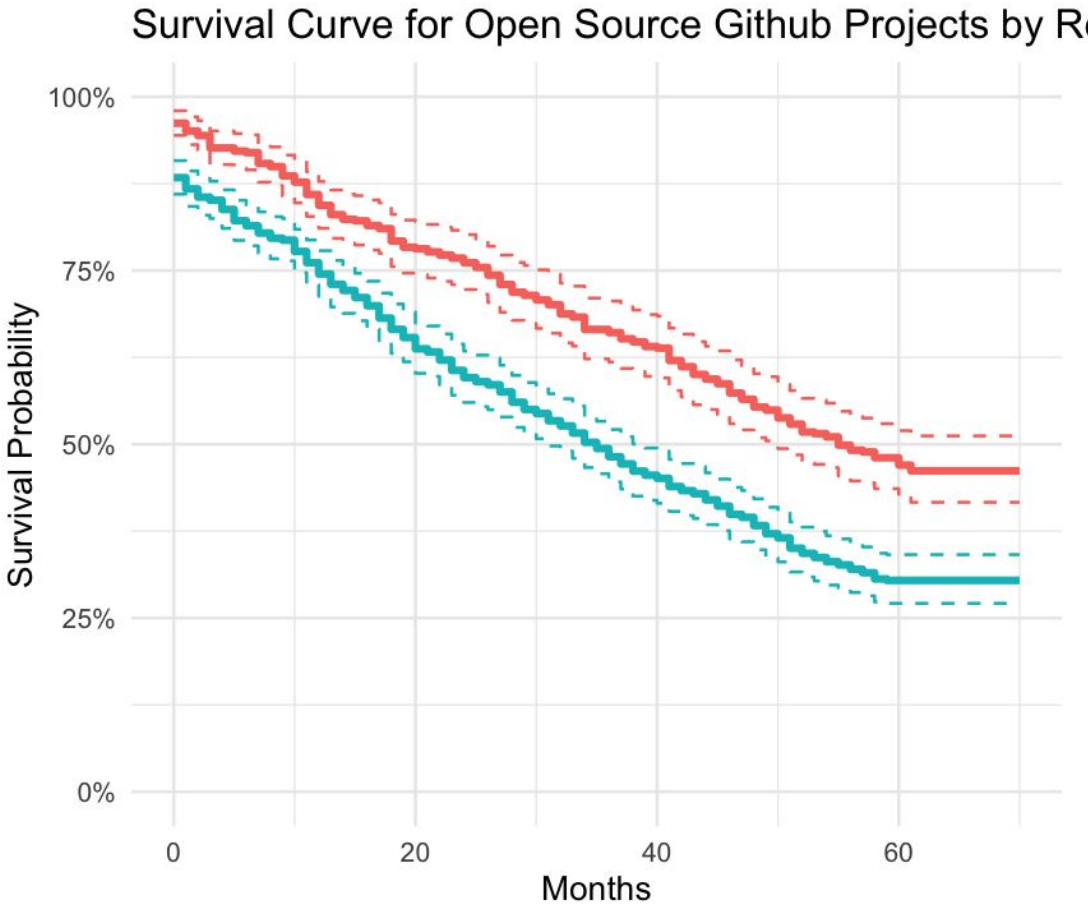


Survival Curves

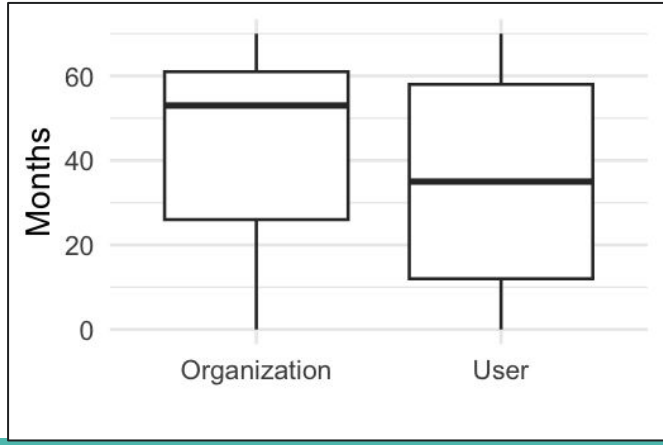


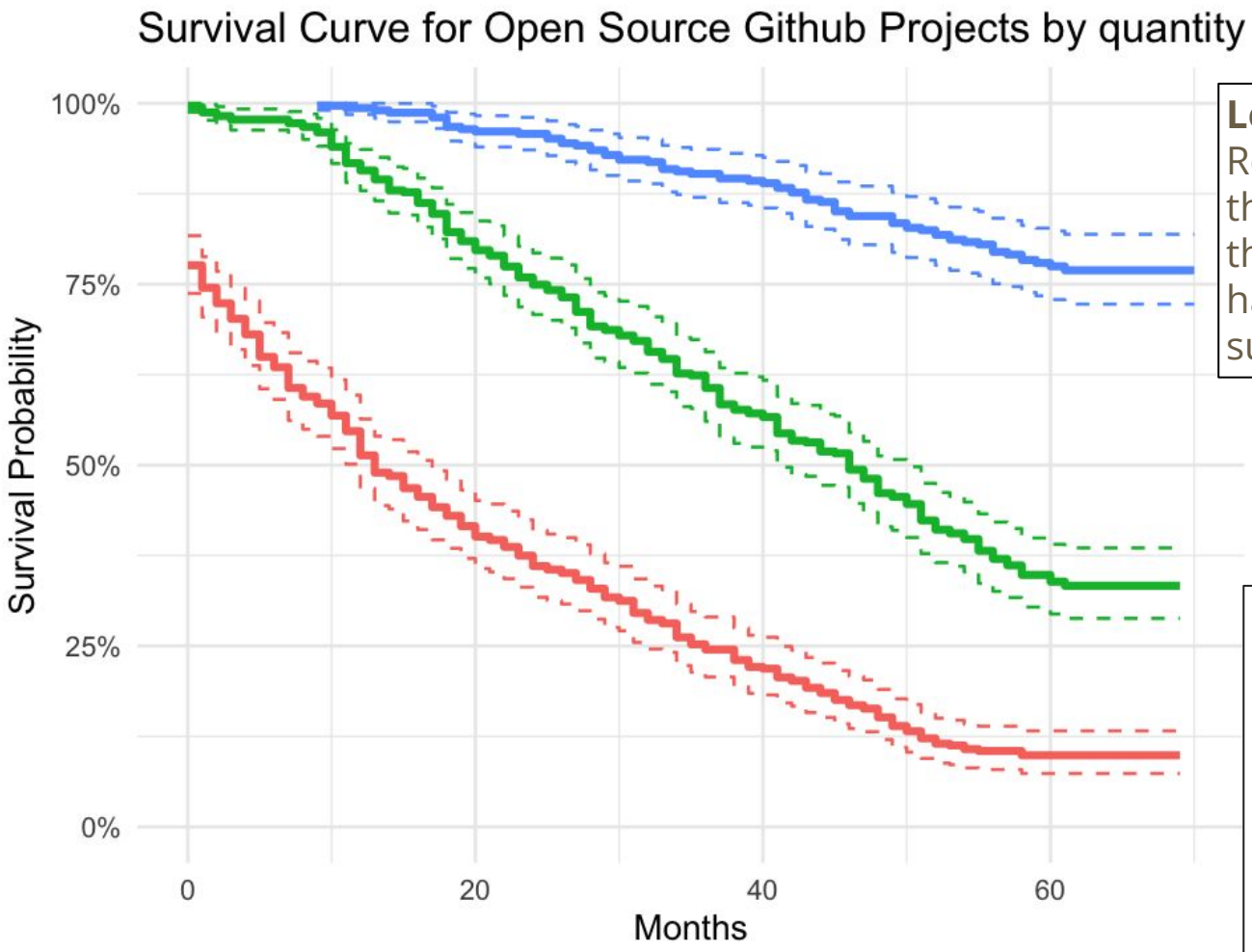


Max	70
3rd Q	59
Median	43
1st Q	16
Min	0

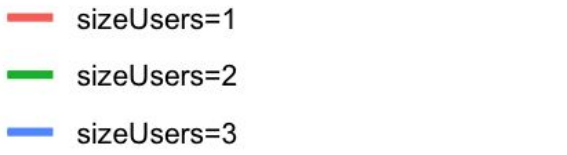


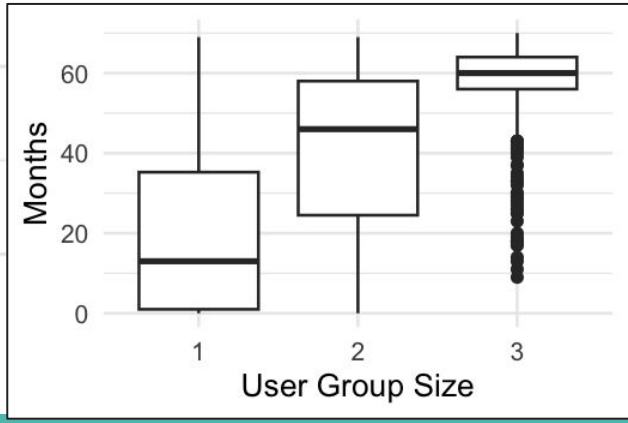
Log Rank Test: p = 1e-08
Reject null hypothesis. Thus, there is at least one time at which one repo type has statistically different survivability.

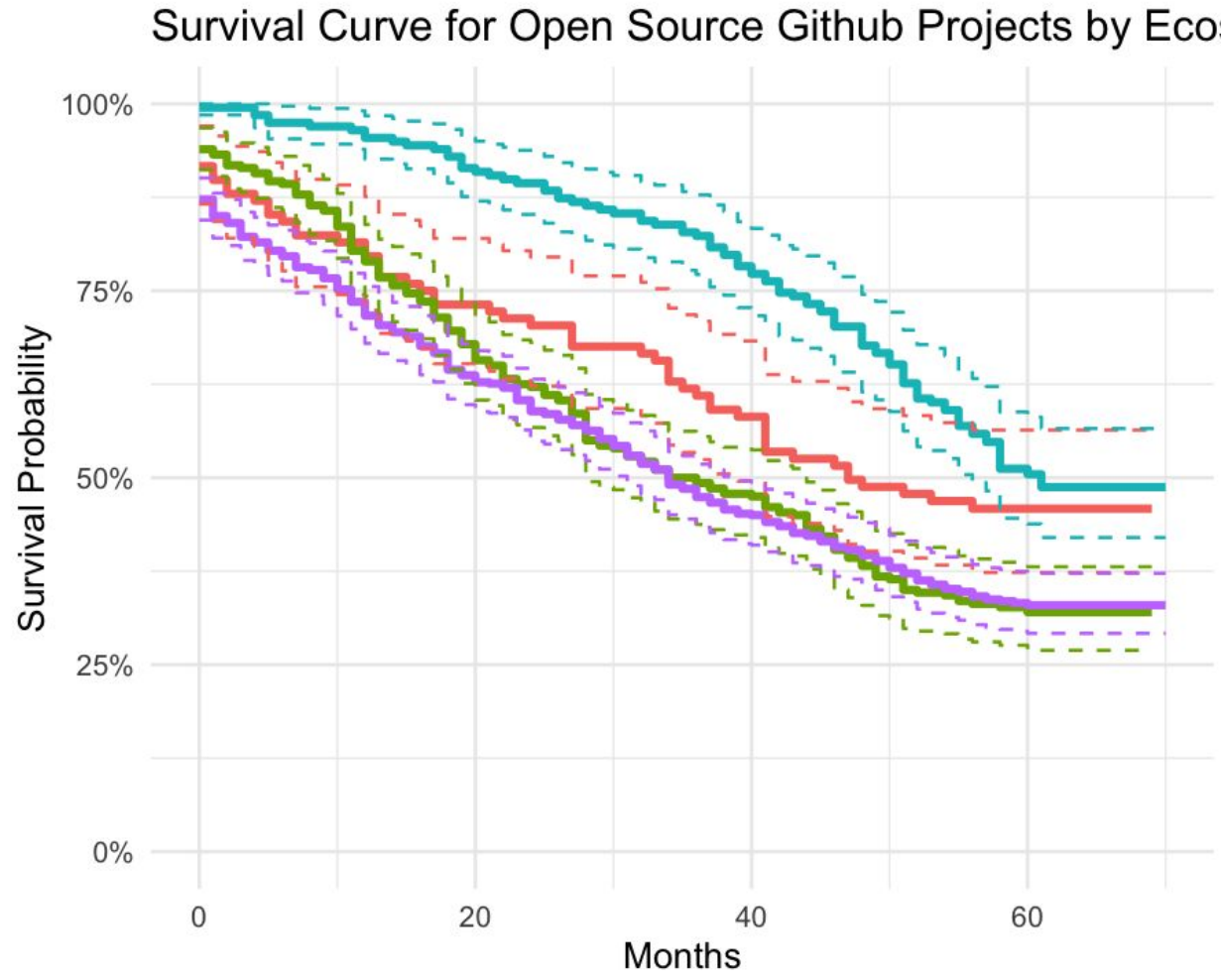




Log Rank Test: p = 1e-08
Reject null hypothesis. Thus, there is at least one time at which the qty of users on on projects has statistically different survivability.

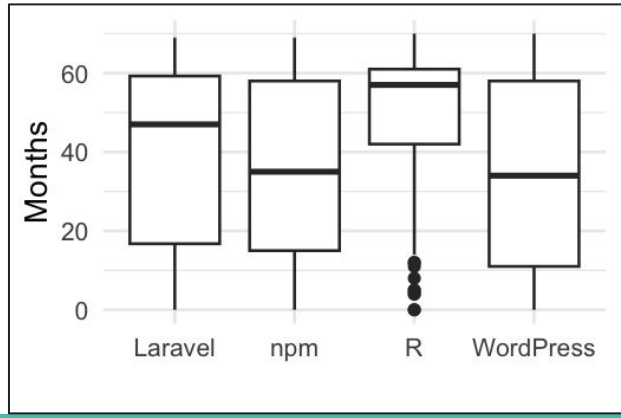




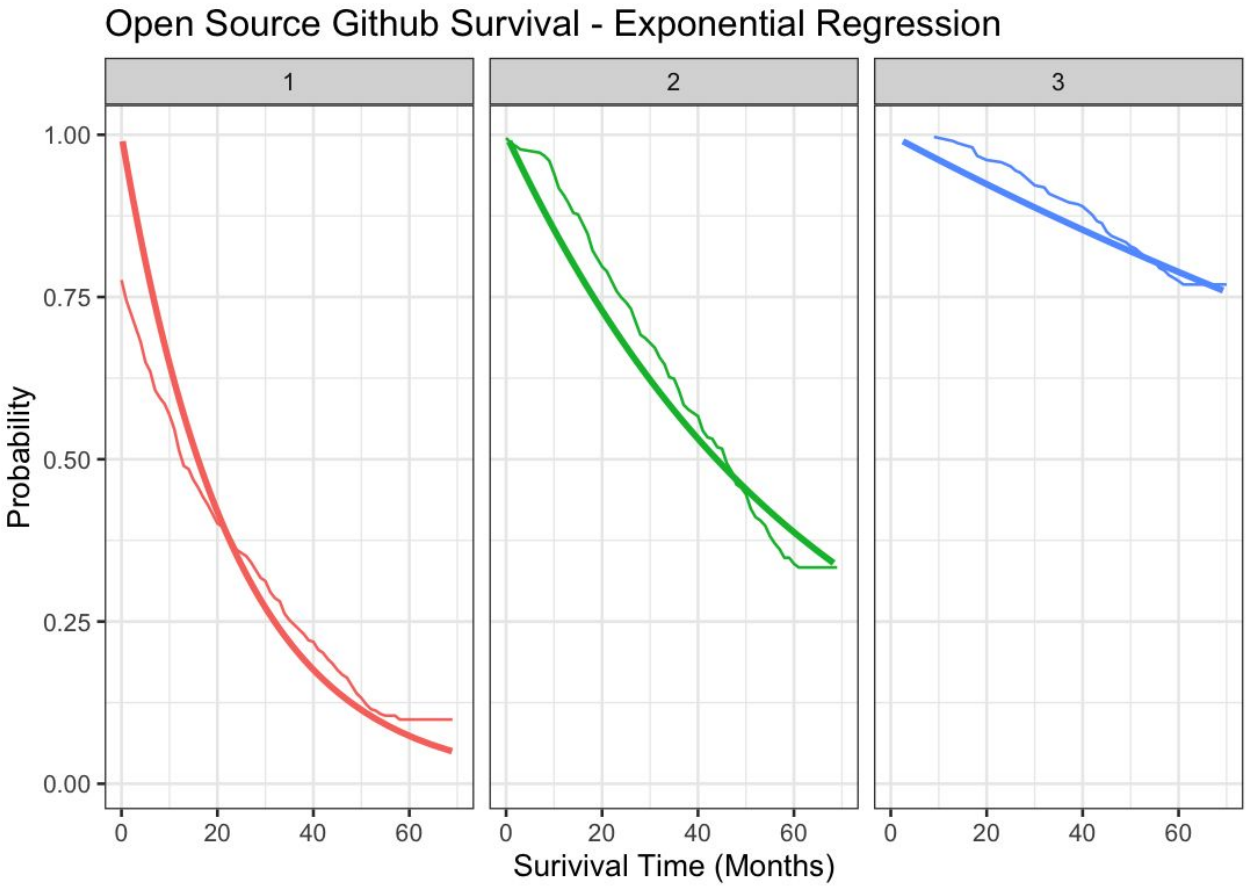


Log Rank Test: p = 4e-07
Reject null hypothesis. Thus, there is at least one time at which one ecosystem has statistically different survivability.

ecosystem=laravel
ecosystem=npm
ecosystem=r
ecosystem=wp

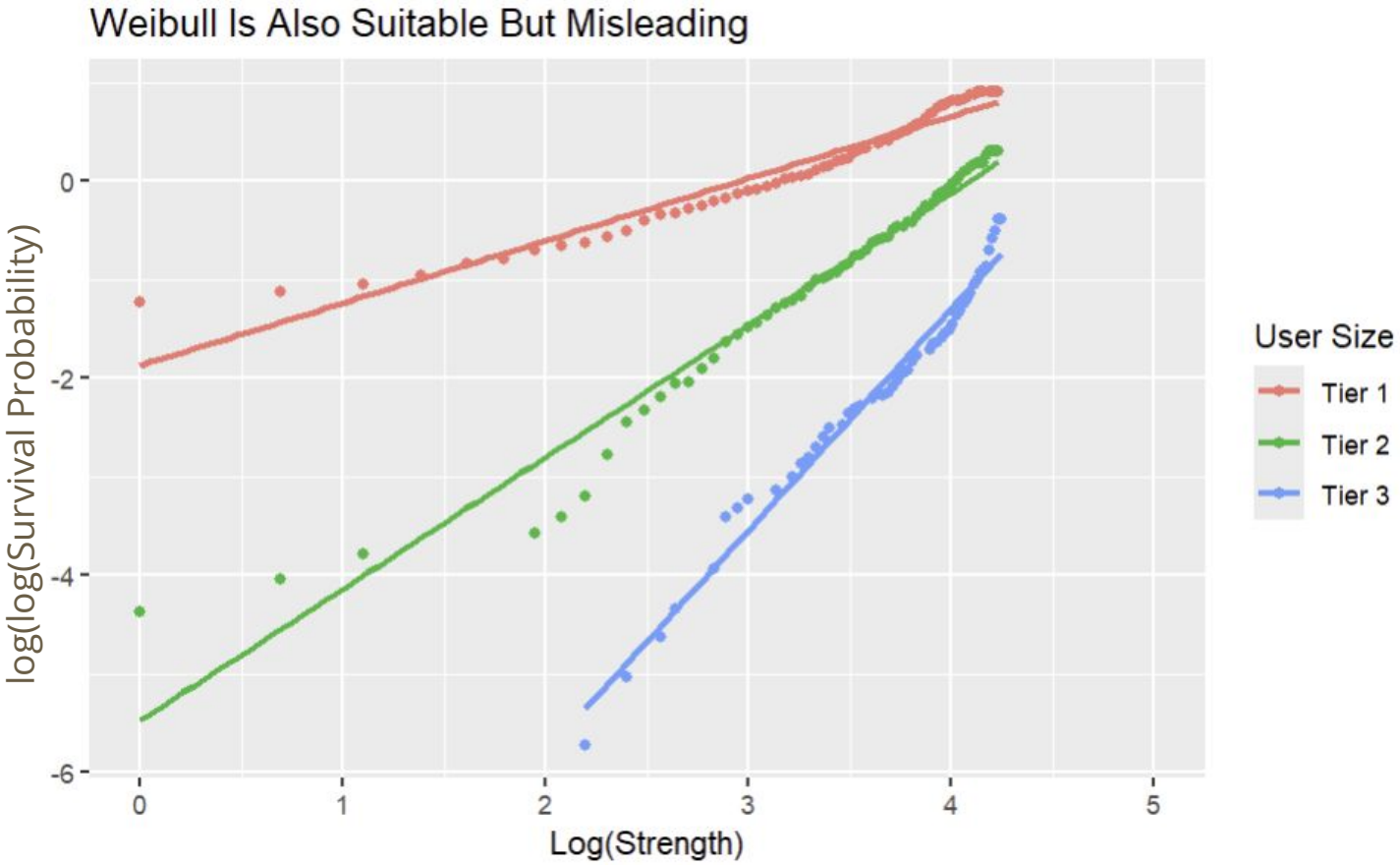


Fit of a Exponential Regression Model



Note: Weibull and Lognormal regressions models did not fit the data as well as the exponential regression.

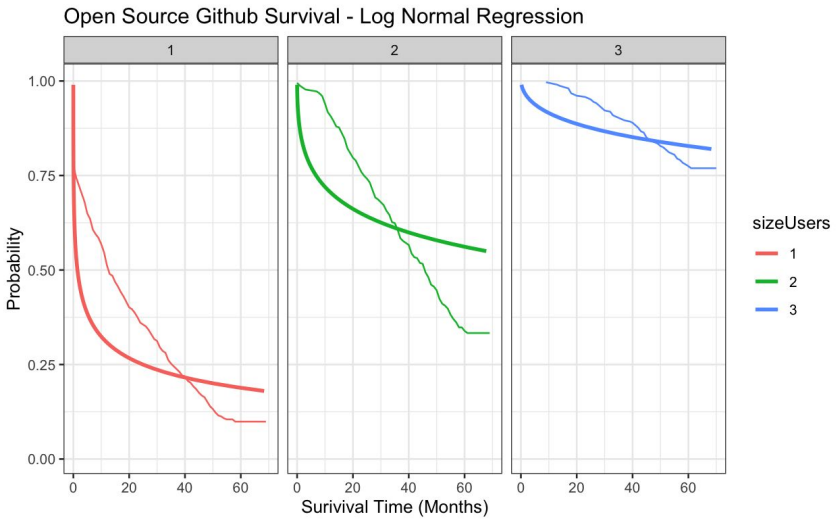
Why Not Bull?

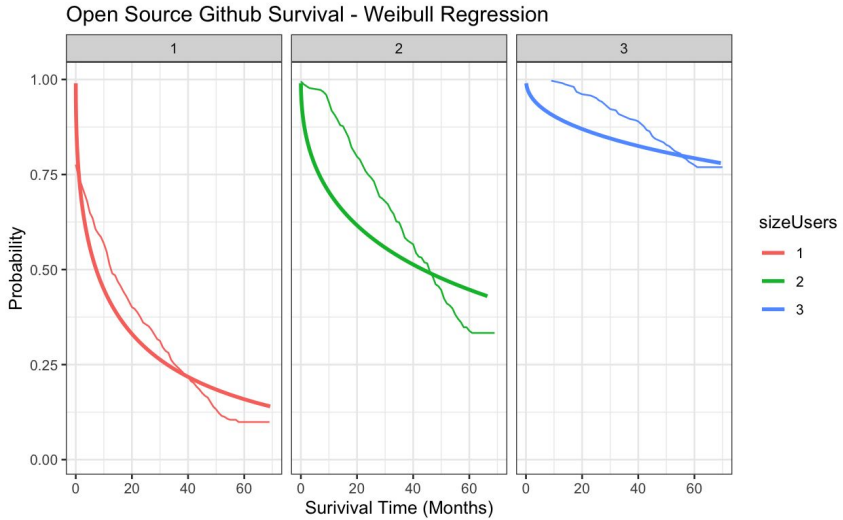

Weibull Is Also Suitable But Misleading
log(log(Survival Probability))
0
-2
-4
-6
0
1
2
3
4
5
Log(Strength)
User Size
Tier 1
Tier 2
Tier 3

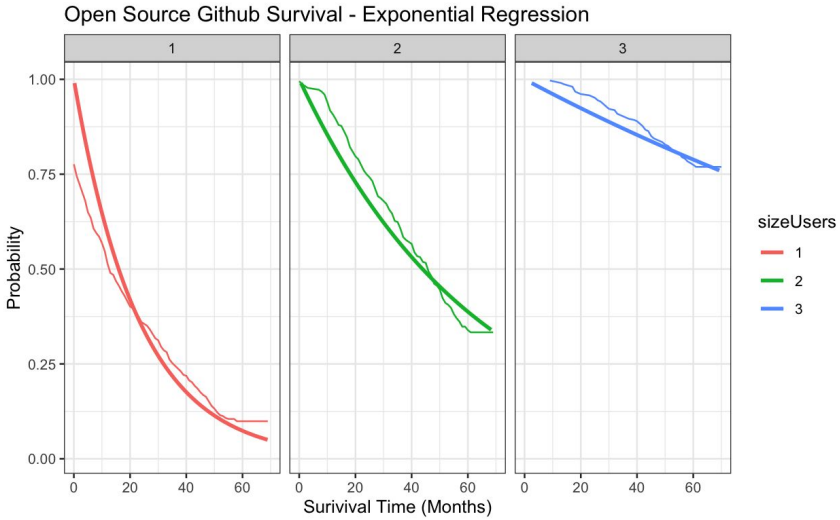
Regressions

```
[Weibull AIC]        6059.468
[Exponential AIC]    6694.187
[Lognormal AIC]      6368.114
```

The AIC results imply that the Weibull Regression would have the best fit and the exponential regression would have the worst fit, but visually the exponential has the best fit.







Conclusions

Open source projects are volatile.

Warning: Many organizations that heavily rely on Open Source projects without paying attention to their health nor contributing to their sustainability are at risk.

1. Projects **led by organizations** survive longer than projects led by individual users. (At least at one time)
2. Projects **supported by larger communities** survive longer than projects supported by smaller communities. (At least at one time)
3. Amongst R, npm, WordPress, and Laravel, **R projects** tend to survive the longest. (At least at one time)

Limitations of analysis:

- Did not analyze any confounding factors. A combination of any of the factors may have a greater impact than any on their own.
- No data about the number of coders developing for each ecosystem and the types of people that are prevalent in that ecosystem. For example, R is considered a very academic language while WordPress has a MUCH wider end use base and overall number of users. Does this make a difference?

Future Steps

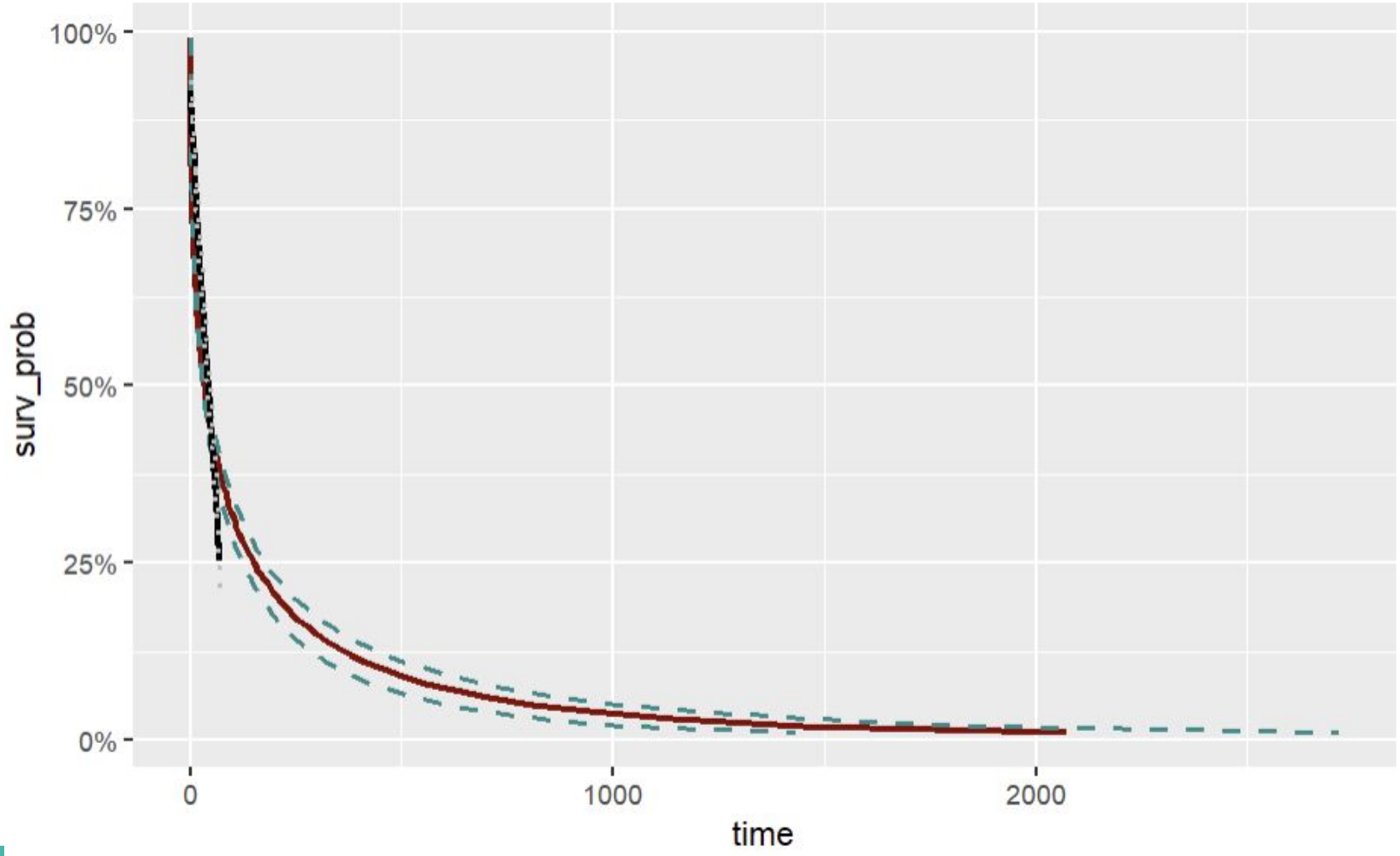
The dataset provided has 19 csv files. We only used 4 of them.

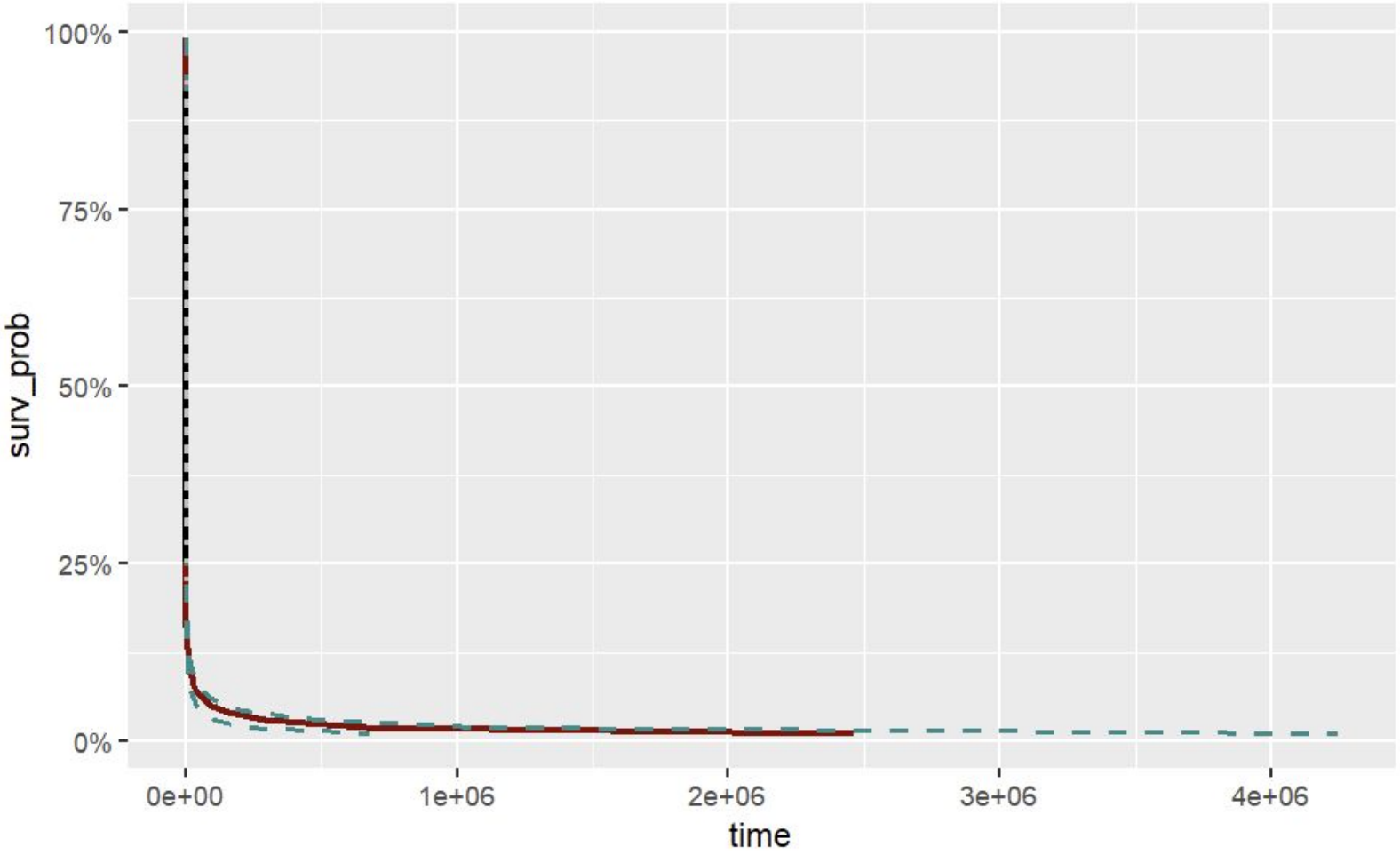
- Analyse the remaining datasets that accompanied the one we worked on.
  - The path projects took throughout the study
  - The effect bots had on survival time of a project
  - The survival rates of the projects that entered a zombie state. Did any come back to life?
  - The different effects of commits, pulls, issues, comments and reviews on the survival of a project

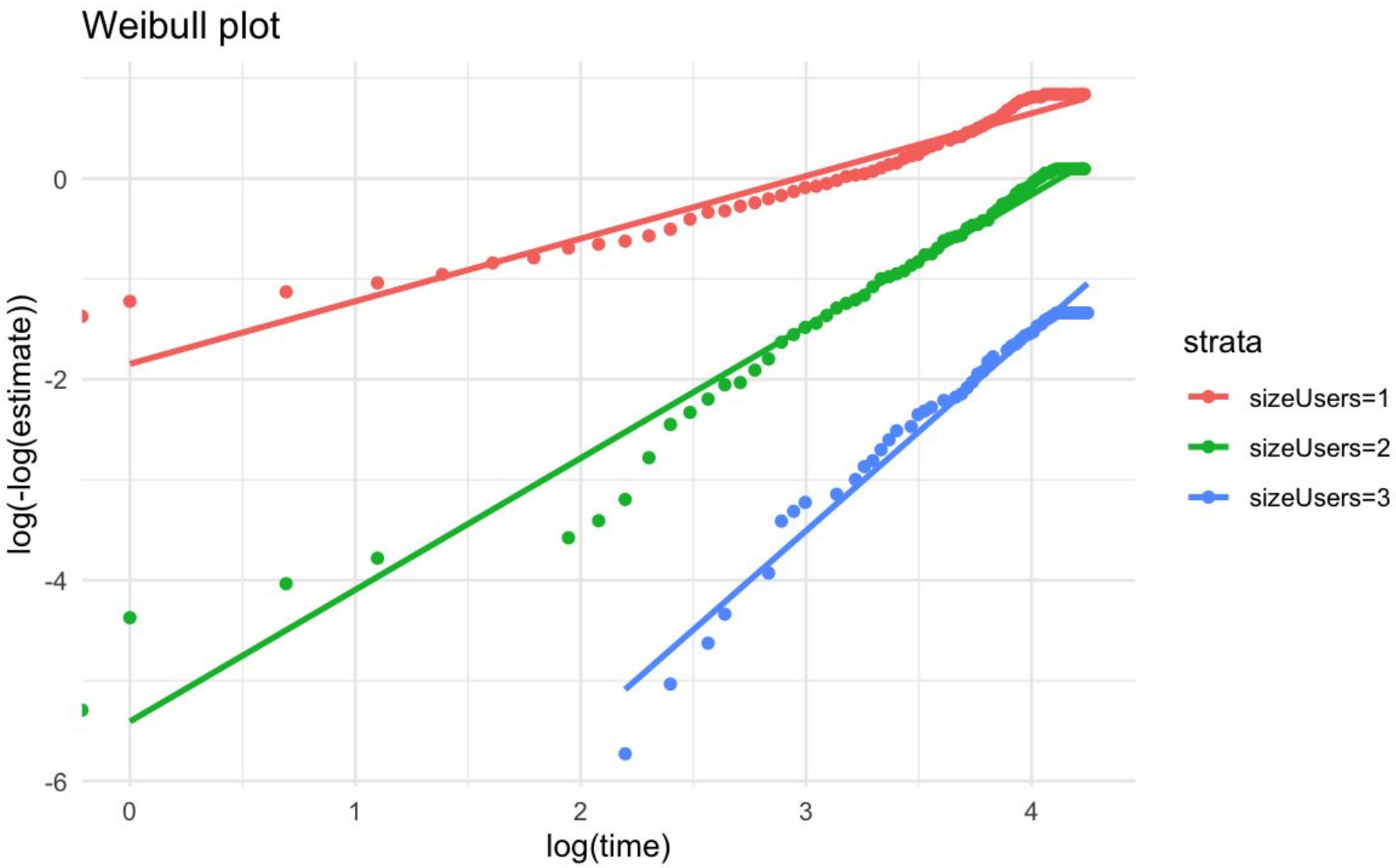
Backup

Who is presenting which slides

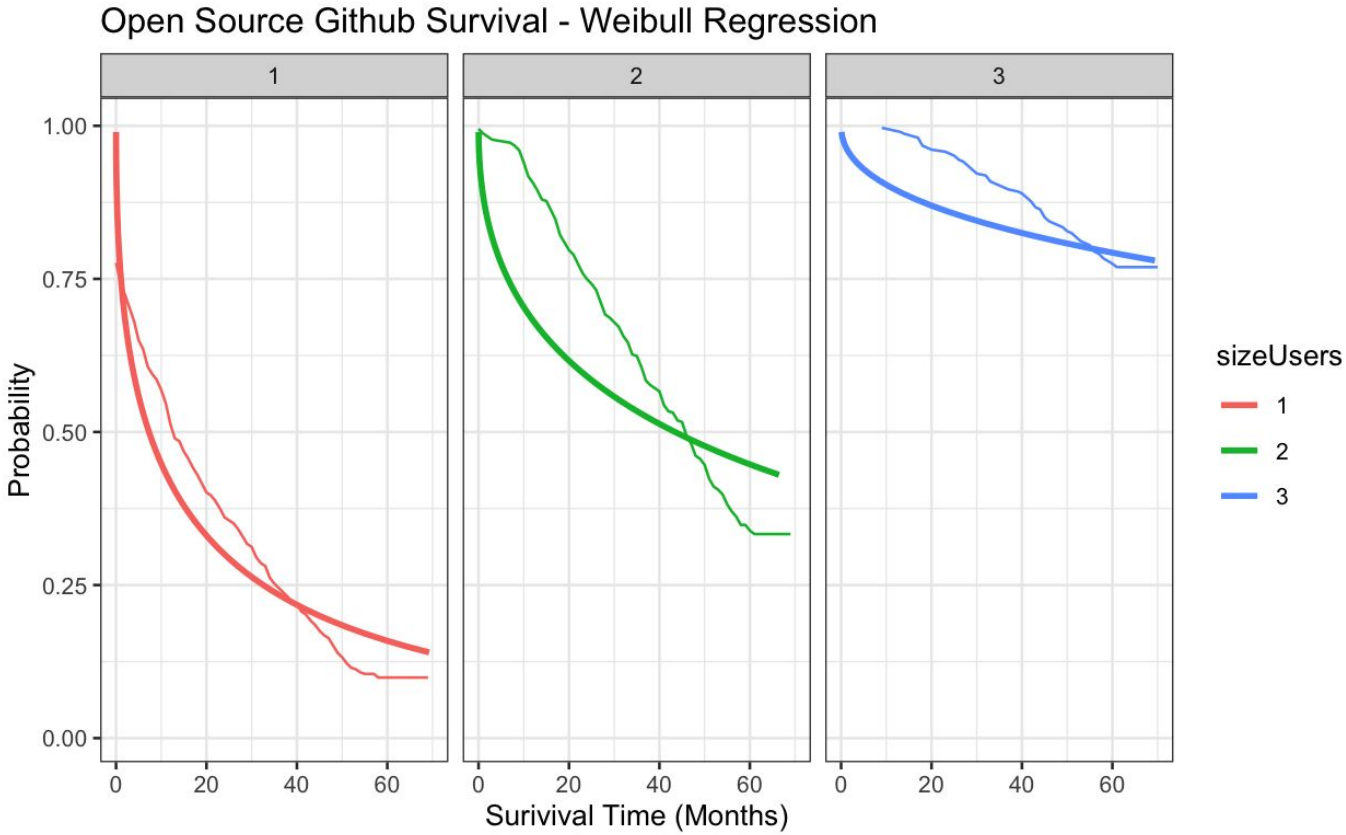
1. Title Slide - Endy
2. Intro - Endy
3. Data & Methodology 1 - Endy
4. Data & Methodology 2 - Endy
5. Sample sizes - Ellann
6. Event plot - Endy
7. KM plot Overall - Endy
8. KM plot 1 - Ellann
9. KM plot 2 - Ellann
10. KM plot 2 - Ellann
11. Regression - Endy
12. No Bull - Endy
13. Different regressions - Ellann
14. Conclusion - Ellann
15. Future Steps - Endy


100%
75%
50%
25%
0%
surv_prob
0
1000
2000
time

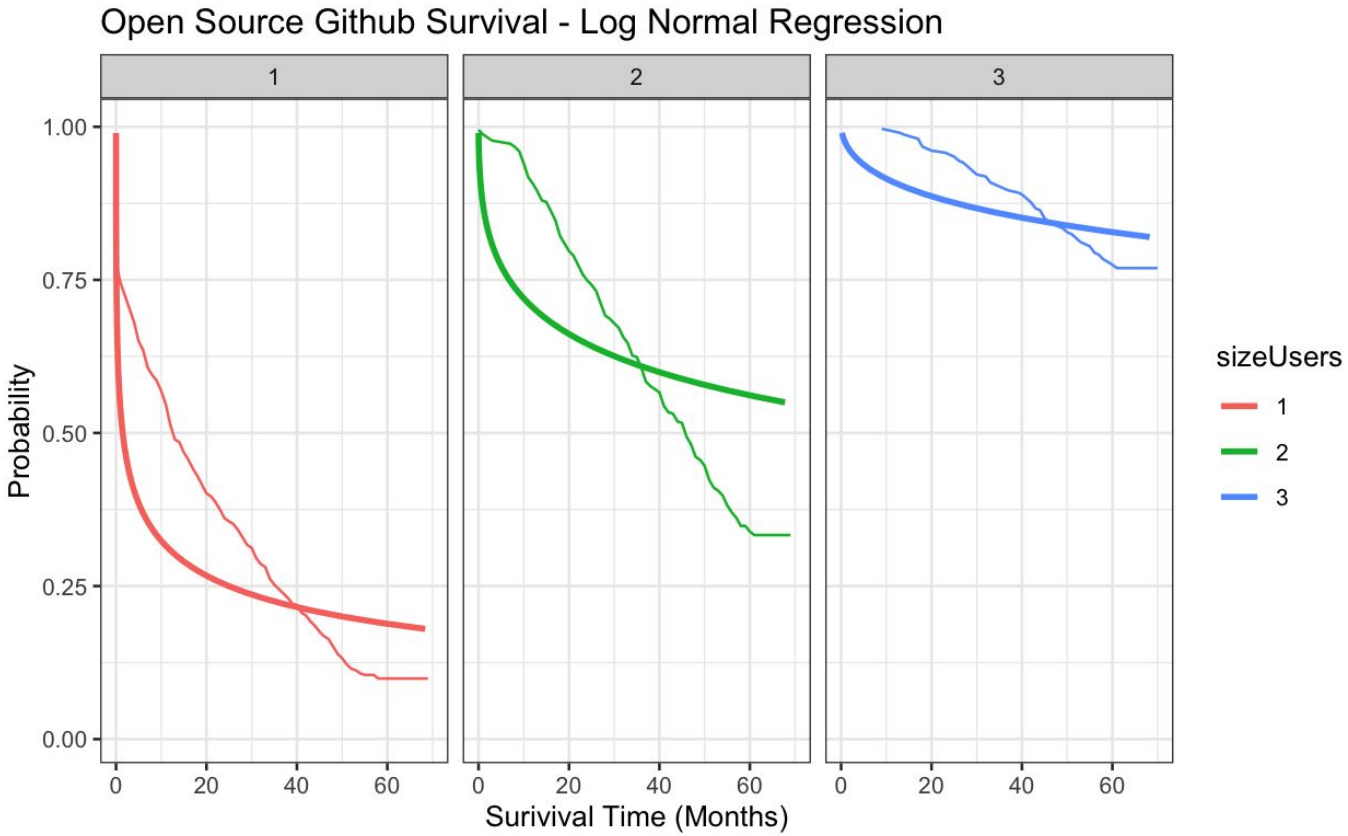

100%
75%
50%
25%
0%
surv_prob
0e+00
1e+06
2e+06
3e+06
4e+06
time


Weibull plot
log(-log(estimate))
0
-2
-4
-6
0
1
2
3
4
log(time)
strata
sizeUsers=1
sizeUsers=2
sizeUsers=3

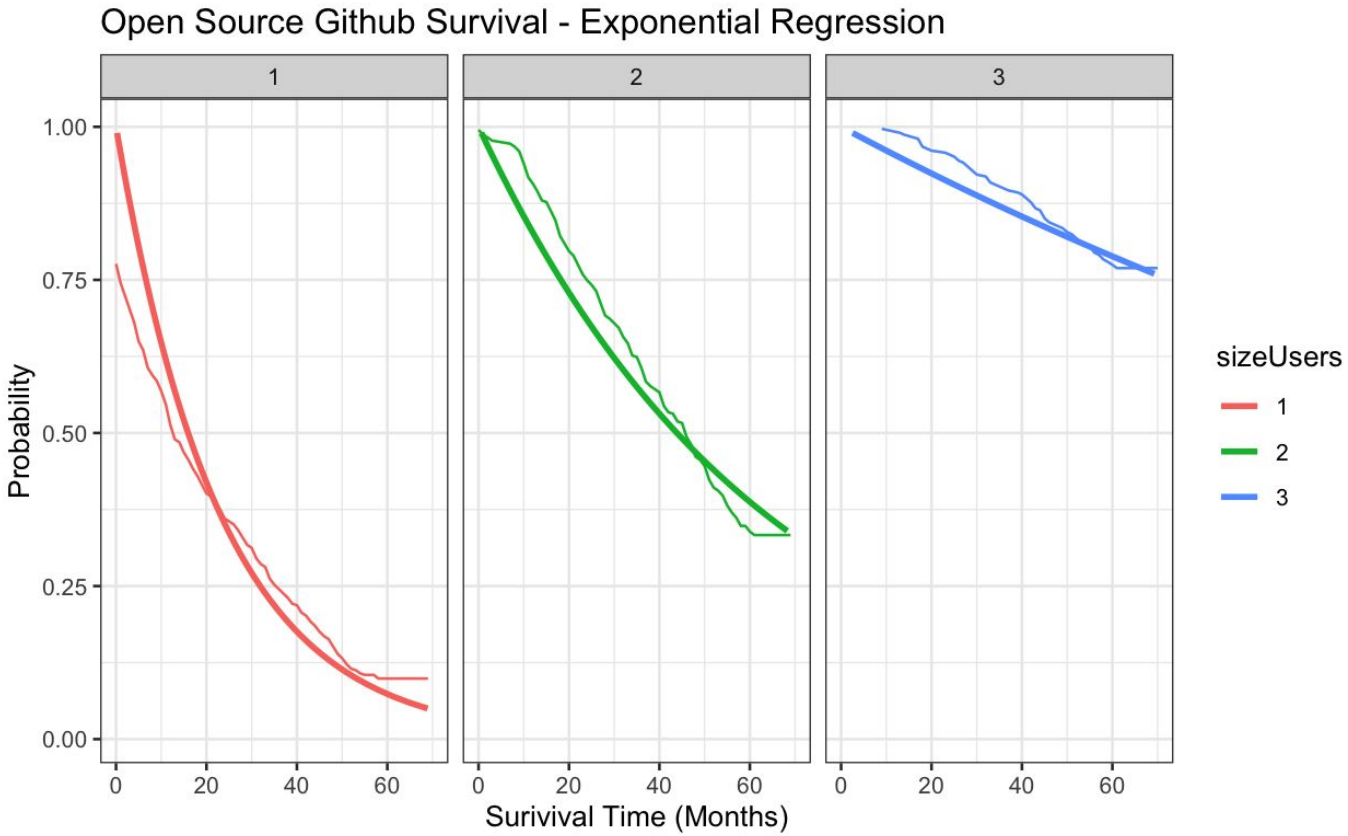
Weibull Regression



Lognormal Regression



Exponential Regression



Data & Methodology: Sample Sizes

Ecosystem	Repo Type: Organization			Repo Type: User			TOTAL
	UserSize 1	UserSize 2	UserSize 3	UserSize 1	UserSize 2	UserSize 3	-
Laravel	9	21	13	30	21	14	***108***
NPM	11	31	31	62	95	50	***280***
R	24	60	38	36	28	13	***199***
WordPress	52	73	85	196	70	64	***540***
TOTAL	***96***	***185***	***167***	***324***	***214***	***141***	***1127***

UserSize 1: Within smallest 25% of project user sizes (aka below min of interquartile range)
UserSize 2: Within middle 50% of project eser sizes (aka within interquartile range)
UserSize 3: Within top 25% of project user sizes (aka above max of interquartile range)