Github - Análise dos projetos mais famosos

Felipe Arruda, Cecília Soares, Diogo Ferreira, João Calil

Contextualização: Github

- Maior rede colaborativa de projetos usando o controle de versionamento GIT.

- Para a base de dados, foram utilizados os 1843 projetos (repositórios) com maior numero de **Estrelas**.

Objetivo do Trabalho

- Identificar como os dados dos projetos mais famosos se relacionam
- Estabelecer possíveis padrões na utilização da ferramenta
- Entender melhor o comportamento dos seus usuários.

Base de Dados - Variáveis

- Name
- Owner Type
- Created at (Year)/(Month)/(Day)
- Last Updated at (Year)/(Month)/(Day)
- Size (kB)
- Stars
- Language
- Has Wiki
- Forks
- Open Issues
- Num. Cmts. (Dia da Semana)
- Total Commits

Base de Dados - Variáveis (Esclarecimentos)

Owner Type: O dono do projeto. Pode ser um usuário (user) ou uma organização (organization).

Stars: Quando um usuário gosta de um projeto, ele pode marcá-lo com uma estrela; assim, projetos mais populares têm mais estrelas.

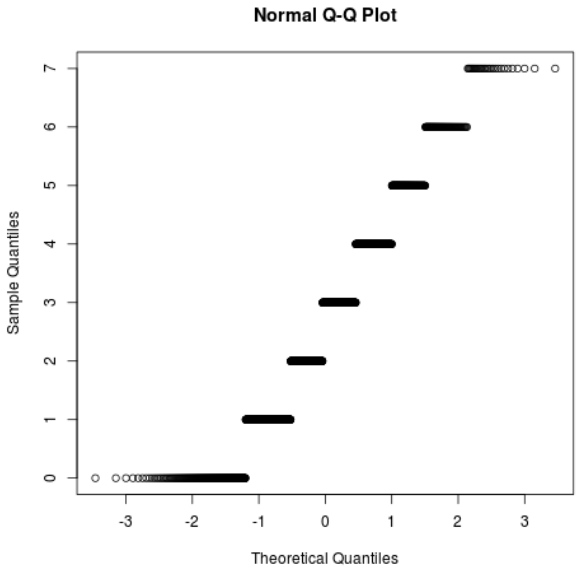
Language: Linguagem principal usada no projeto; Caso não haja linguagem, o campo é assinalado como “none” (nenhuma).

Fork: Uma vertente de um projeto; normalmente é criado quando alguém quer colaborar em um projeto.

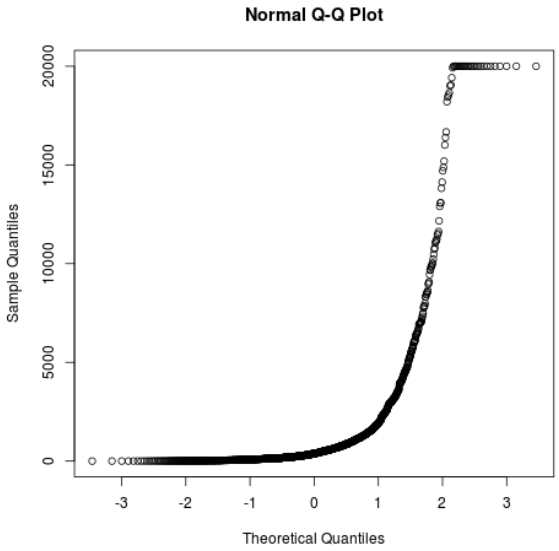
Open Issues: Bugs/melhorias/dúvidas em aberto

Commit: Uma alteração no projeto/repositório. Ex.: Uma mudança em um código ou em uma imagem, uma nova página em um livro, dentre outros.

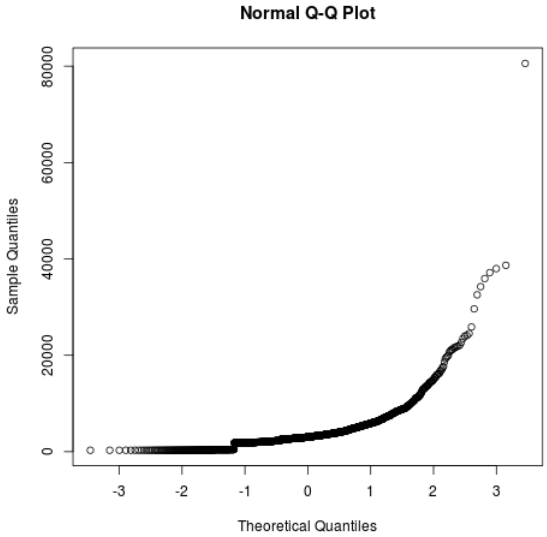
Testes de Normalidade: QQNorm



Age



Total Commits

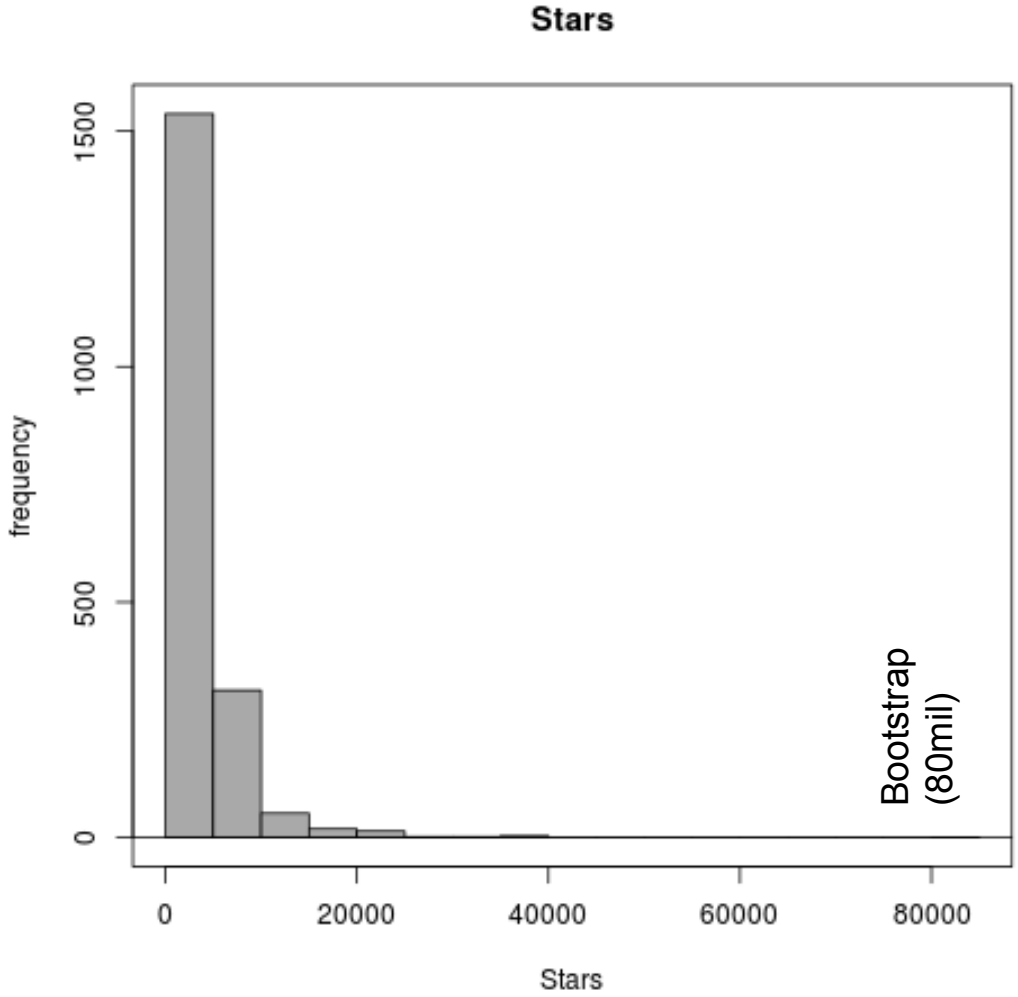


Stars

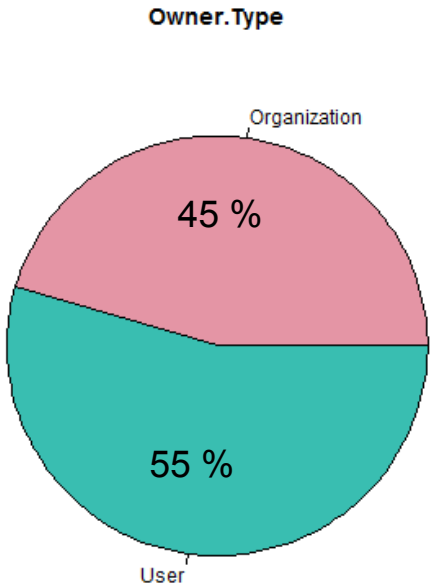
Testes de Normalidade: Shapiro-Wilk

Variável	p.value
Size	1.334735e-68
Stars	2.152607e-54
Watchers	2.152607e-54
Forks	1.833763e-64
Open.Issues	5.945727e-61
Num..Cmts..Dom	4.849562e-60
Num..Cmts..Seg	5.006677e-60
Num..Cmts..Ter	3.93423e-60
Num..Cmts..Qua	2.055576e-60
Num..Cmts..Qui	3.409515e-60
Num..Cmts..Sex	4.873476e-60
Num..Cmts..Sab	1.52919e-60
Total.Commits	8.222452e-60
Age	2.458655e-25

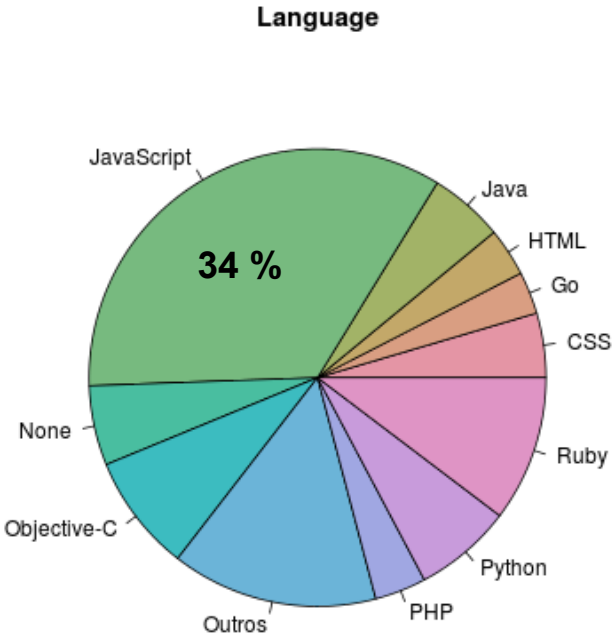
Análise - Stars


Stars
frequency
1500
1000
500
0
0
20000
40000
60000
80000
Stars
Bootstrap
(80mil)

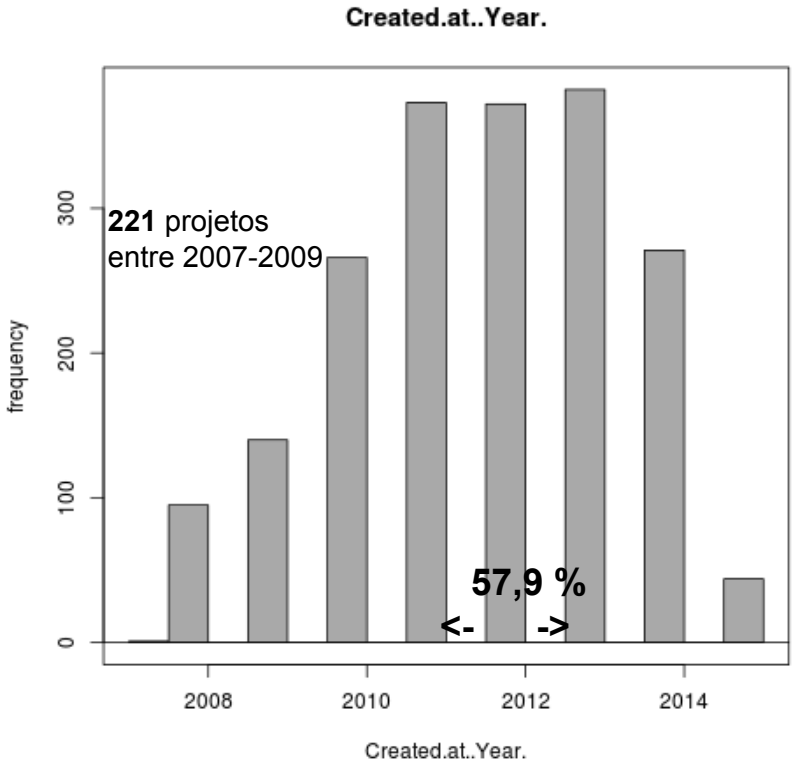
Análise - Owner Type

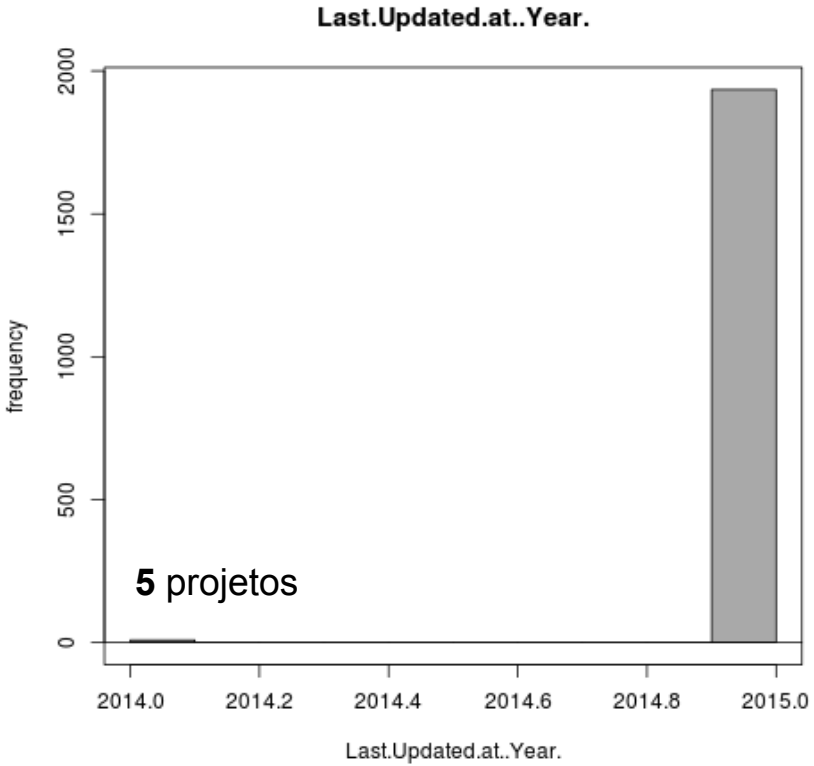


Análise - Languages



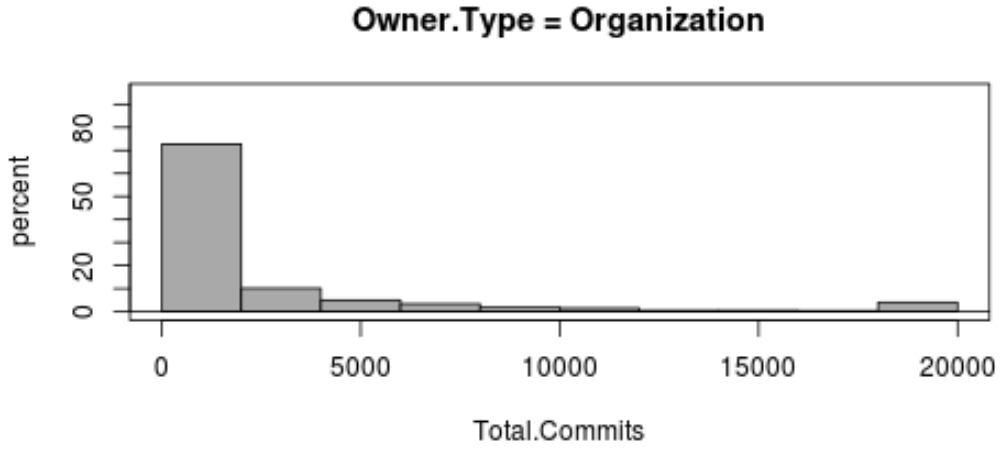
Data de Criação e Data de Atualização

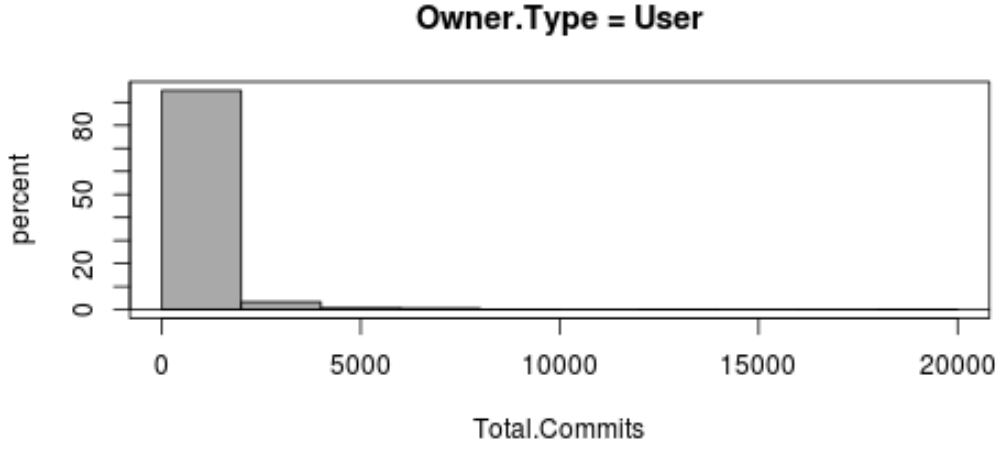




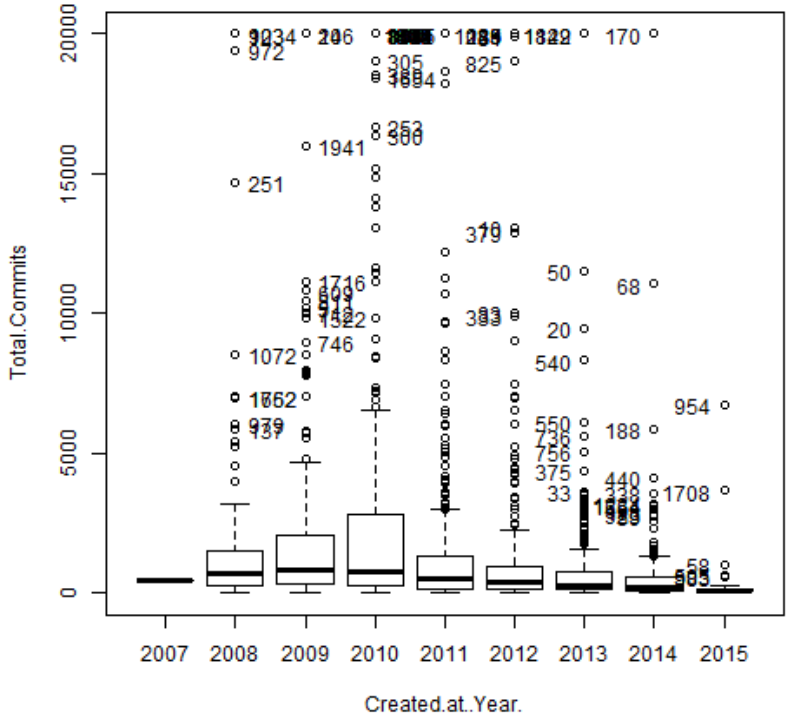
Correlações:
Total de Commits x Owner Type

Tipo de Usuário	Média (Total de Commits)	CV (Total de Commits)
Organização	2570	1.74
Usuário	555	2.08

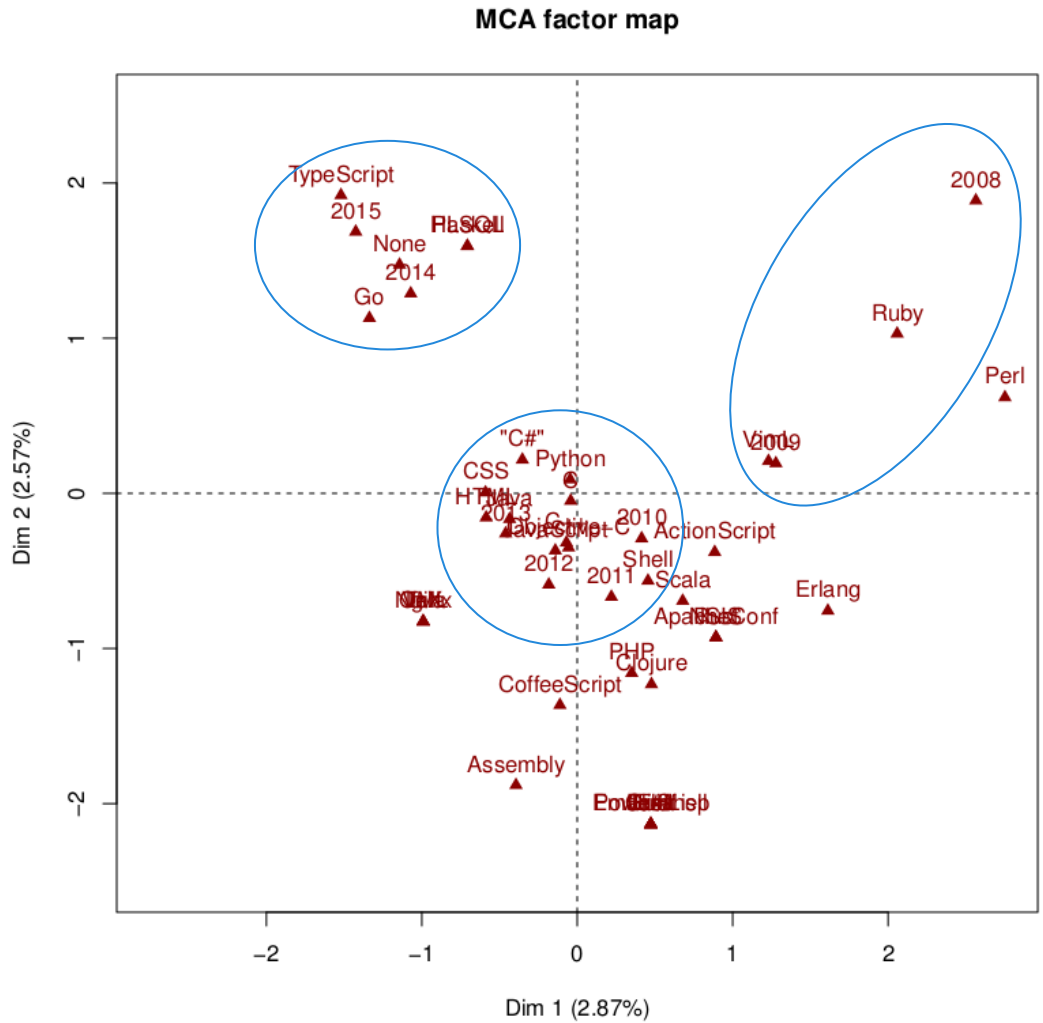

Owner.Type = Organization
percent
0
20
50
80
0
5000
10000
15000
20000
Total.Commits


Owner.Type = User
percent
0
20
50
80
0
5000
10000
15000
20000
Total.Commits

Data de Criação vs Total de Commits

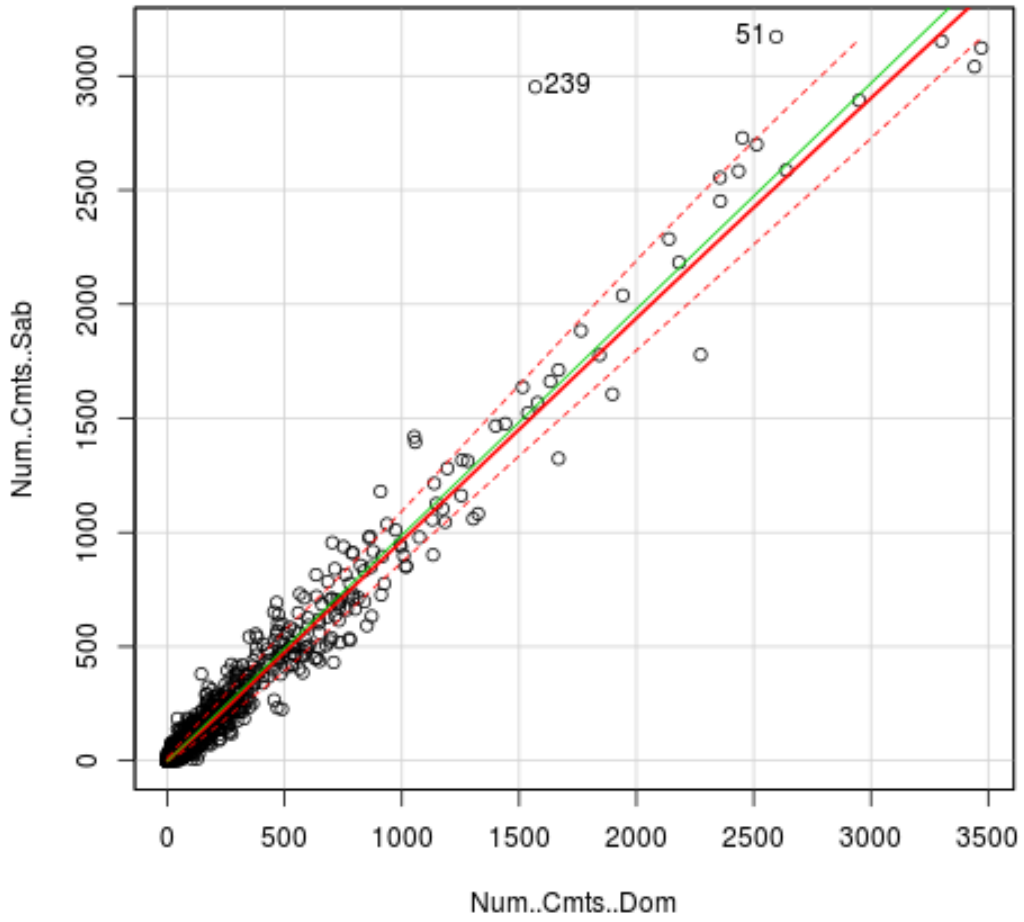


Correlação:
Linguagem e Ano de Criação


MCA factor map
TypeScript
2015
None
2014
Go
2008
Ruby
Perl
"C#"
Python
CSS
2010
ActionScript
2012
Shell
2011
Scala
Erlang
PHP
Clojure
CoffeeScript
Assembly
Dim 2 (2.57%)
Dim 1 (2.87%)
2
1
0
−1
−2
−2
−1
0
1
2

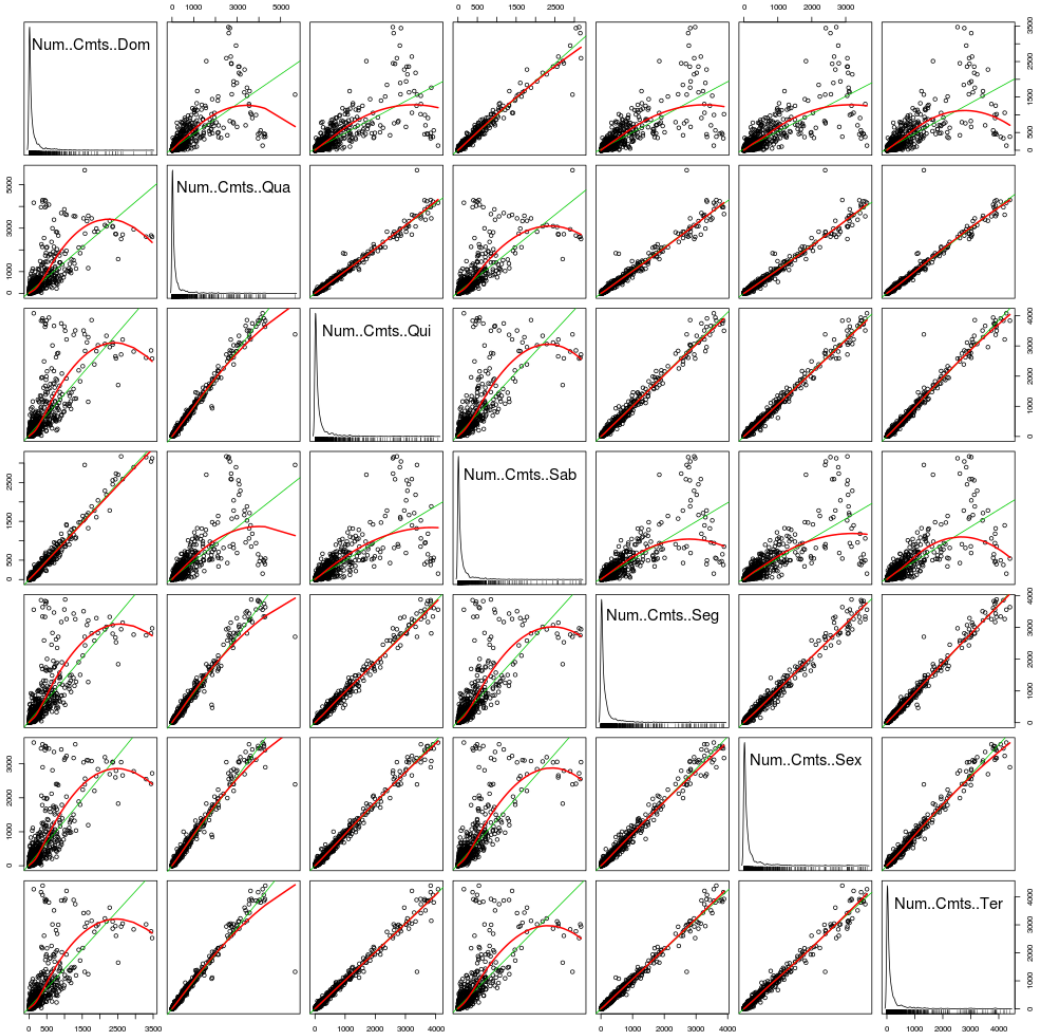
Número de commits por dia da semana

- Sábado x Domingo
- Dias úteis x Dias úteis


51
239
Num..Cmts..Sab
Num..Cmts..Dom
0
500
1000
1500
2000
2500
3000
3500

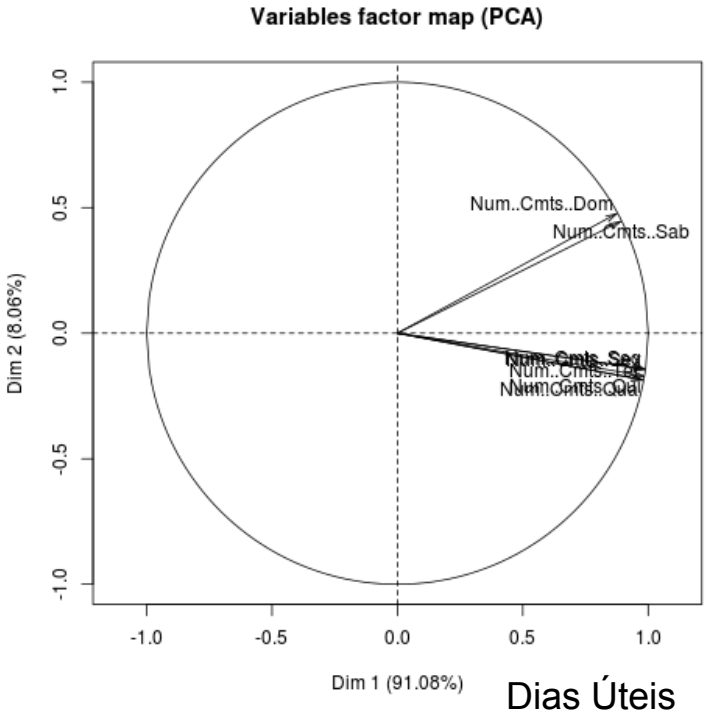
Matriz de Correlação

	Num.. Cmts..Dom	Num.. Cmts..Qua	Num.. Cmts..Qui	Num.. Cmts..Sab	Num.. Cmts..Seg	Num.. Cmts..Sex	Num.. Cmts..Ter
Num.. Cmts..Dom	1.0000000	0.7609127	0.7774553	**0.9814794**	0.7955426	0.7939936	0.7700988
Num.. Cmts..Qua	0.7609127	1.0000000	**0.9909539**	0.7904662	**0.9837513**	**0.9841419**	**0.9800930**
Num.. Cmts..Qui	0.7774553	0.9909539	1.0000000	0.8002759	**0.9918765**	**0.9937085**	**0.9908228**
Num.. Cmts..Sab	0.9814794	0.7904662	0.8002759	1.0000000	0.8105627	0.8155841	0.7828633
Num.. Cmts..Seg	0.7955426	0.9837513	0.9918765	0.8105627	1.0000000	**0.9908058**	0.9917477
Num.. Cmts..Sex	0.7939936	0.9841419	0.9937085	0.8155841	**0.9908058**	1.0000000	**0.9904784**
Num.. Cmts..Ter	0.7700988	0.9800930	0.9908228	0.7828633	**0.9917477**	0.9904784	1.0000000


Num..Cmts..Dom
Num..Cmts..Qua
Num..Cmts..Qui
Num..Cmts..Sab
Num..Cmts..Seg
Num..Cmts..Sex
Num..Cmts..Ter

Análise Fatorial:
Final de Semana e Dias úteis

Final de Semana



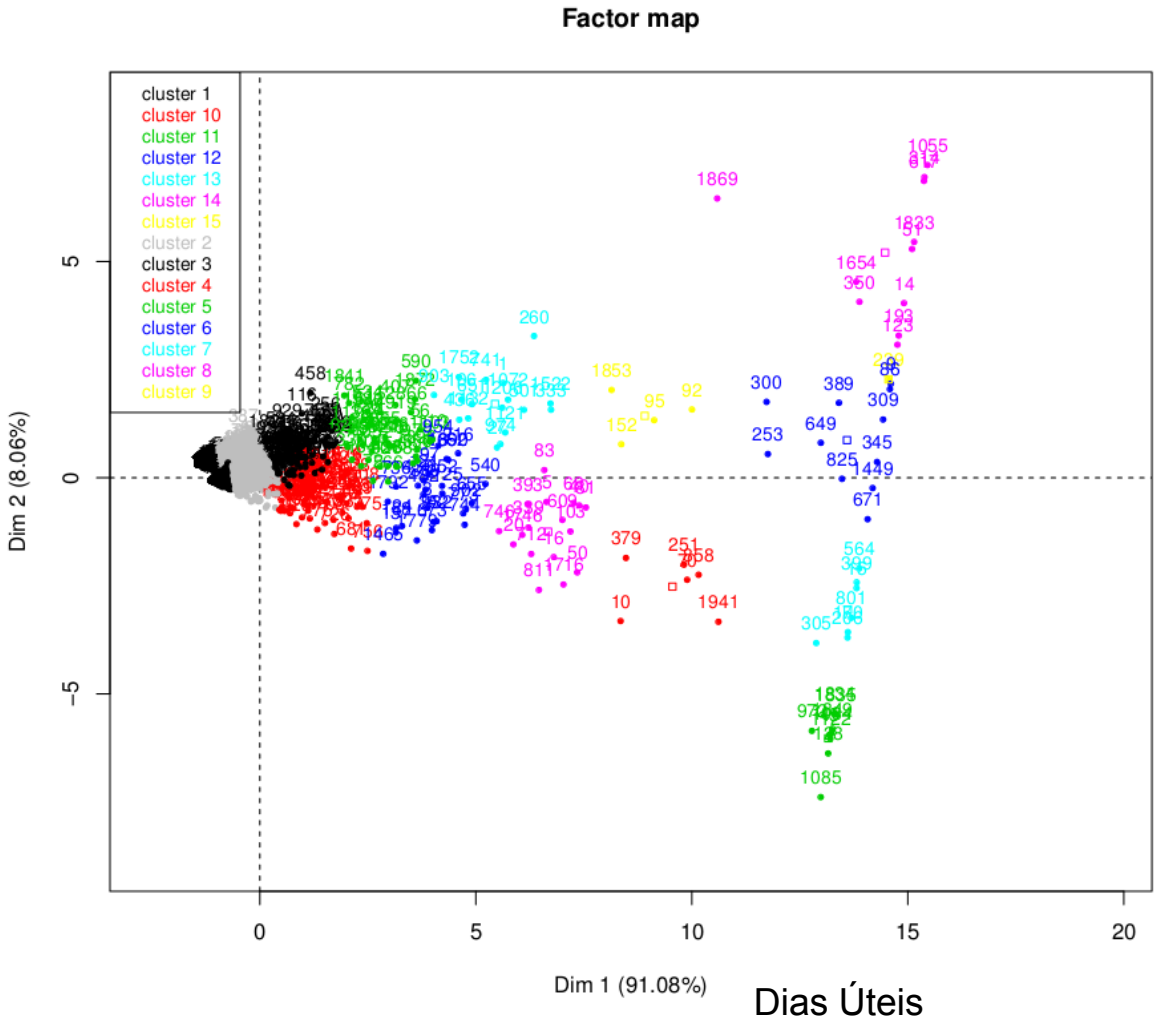
Dias Úteis

Análise Fatorial:
Final de Semana e Dias úteis

	Factor1 (Dias Úteis)	Factor2 (Finais de Semana)
Num..Cmts..Dom	0.462	**0.869**
Num..Cmts..Qua	**0.903**	0.405
Num..Cmts..Qui	**0.907**	0.415
Num..Cmts..Sab	0.482	**0.873**
Num..Cmts..Seg	**0.895**	0.435
Num..Cmts..Sex	**0.894**	0.440
Num..Cmts..Ter	**0.913**	0.394
Proportion Var	0.645	0.342
Cumulative Var	0.645	**0.987**

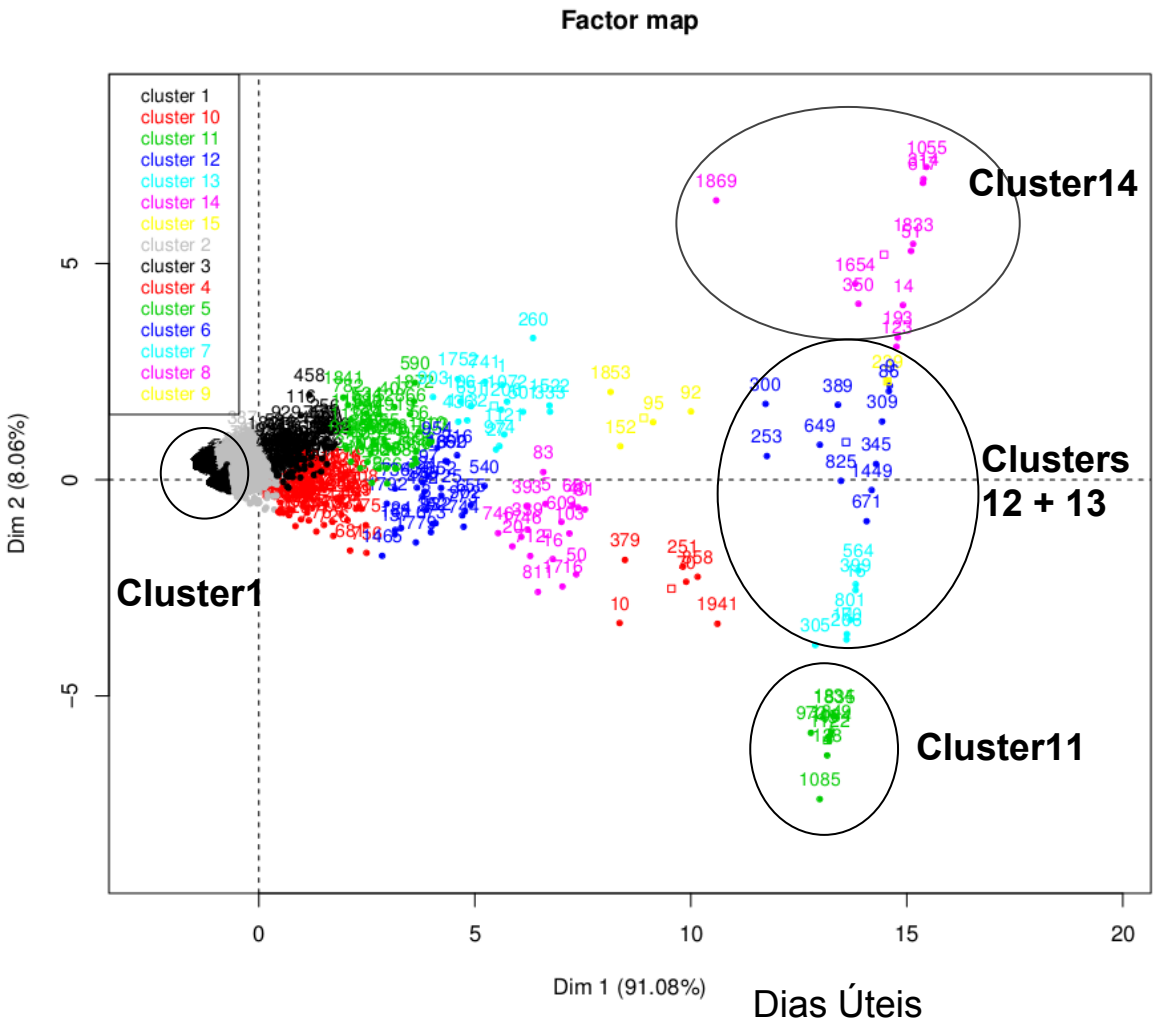
Análise de Cluster:
Commits por dia da semana

Final de Semana



Dias Úteis

Final de Semana



- **Cluster 1:** Poucos Commits;
- **Cluster 14:** muitos Commits no Final de Semana;
- **Cluster 11:** muitos commits em dias úteis;
- **Cluster 12 e 13:** muitos commits tanto nos finais de semana quanto nos dias úteis.

Análise de Clusters: Commits por dia da semana

Cluster	Média de Forks	Média de Stars	Média de Total.Commits
Cluster 1	574	3223	227
Cluster 11	1297	4478	19935
Cluster 14	**3327**	6949	19166
Cluster 12 e 13	3051	**7387**	19325

Conclusão

Foi possível:

- Determinar correlações entre os dados observados
- Entender melhor o comportamento dos usuários do site Github
- Ter uma ideia melhor de quais fatores podem levar um projeto a ser um projeto de renome.

Mais informações

Projeto: https://github.com/arruda/estatistica-github

Licença de Uso: MIT