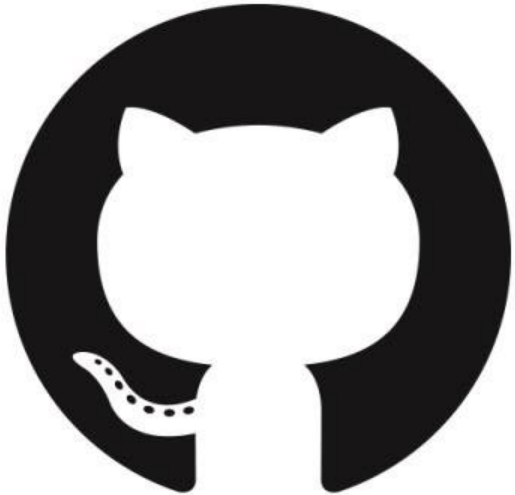
What makes a successful GitHub project: Delving into the behavior of contributors

By

Yanqing ZHOU, Ranjani AV



GitHub

Data introduction

How does github work? What do these variables mean?

Github

- Introduction
  - What is Git?
  - What is GitHub?
  - Advantages

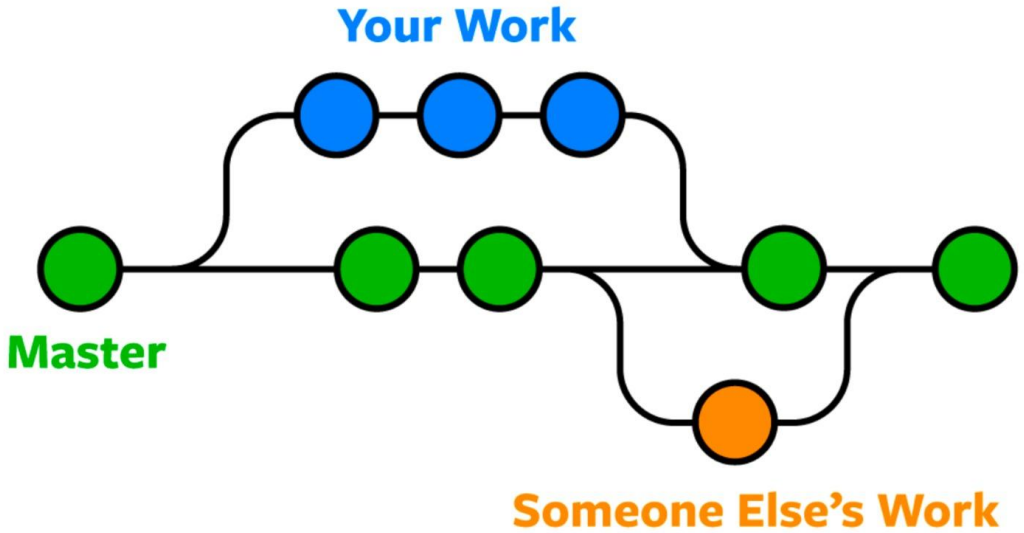
Features:

- Stars
- Branches
- Commits
- Contributors

Github Introduction

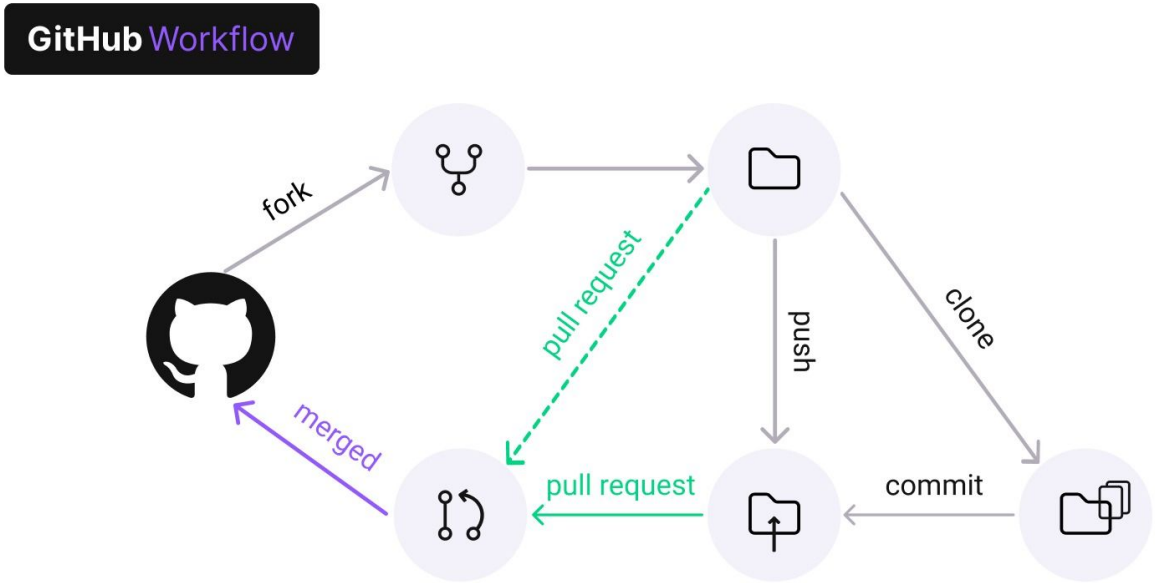
Features:

- Stars
- Branches
- Commits
- Contributors



McDonald, N., & Goggins, S. (2013). Performance and participation in open source software on github. In *CHI'13 extended abstracts on human factors in computing systems* (pp. 139-144).

Github Workflow



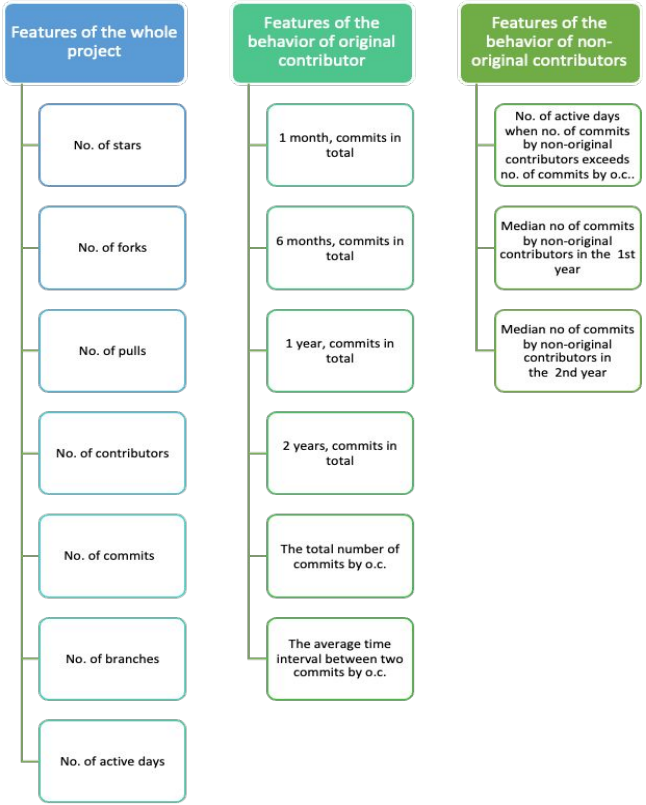
The problem or challenge

Behaviour of github projects based on the contributor’s actions.

Data Processing

What features are we interested in?

Features

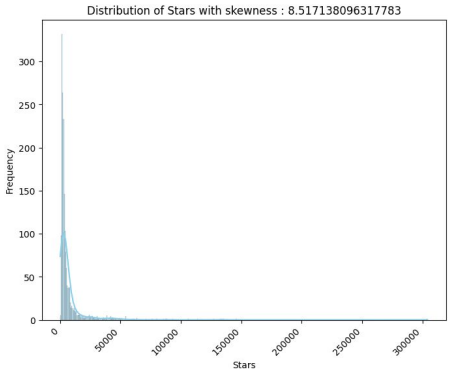

Features of the whole project
No. of stars
No. of forks
No. of pulls
No. of contributors
No. of commits
No. of branches
No. of active days
Features of the behavior of original contributor
1 month, commits in total
6 months, commits in total
1 year, commits in total
2 years, commits in total
The total number of commits by o.c.
The average time interval between two commits by o.c.
Features of the behavior of non-original contributors
No. of active days when no. of commits by non-original contributors exceeds no. of commits by o.c..
Median no of commits by non-original contributors in the 1st year
Median no of commits by non-original contributors in the 2nd year

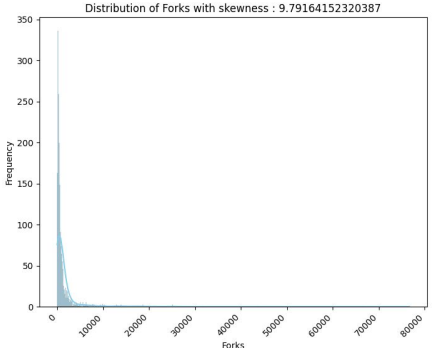
Data distribution and transformation

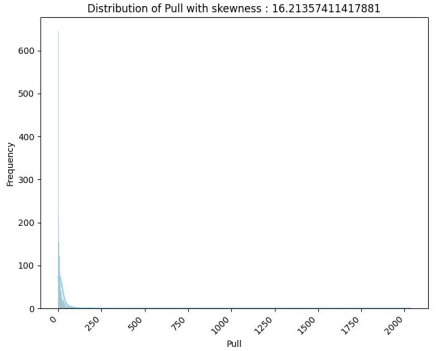
What is the distribution of the features that we extract?

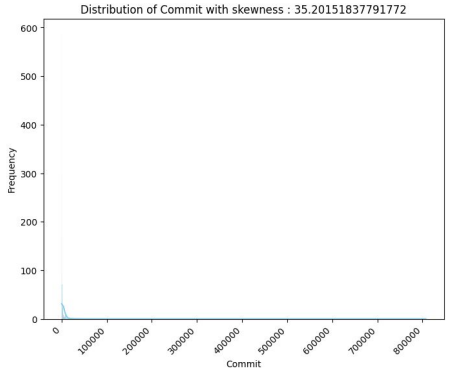
A really skewed dataset

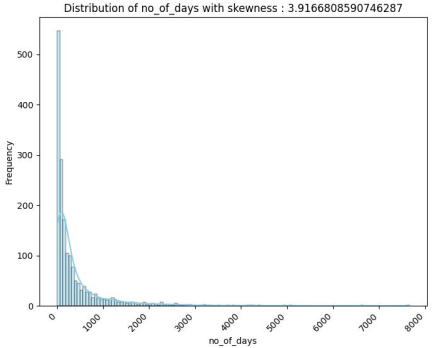
Features of the whole project

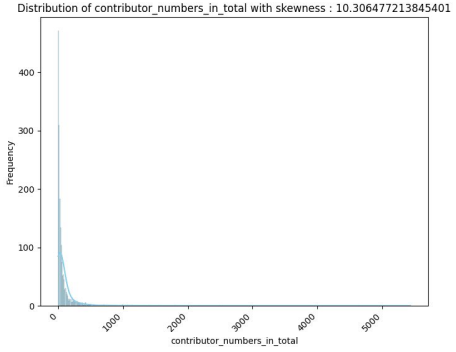






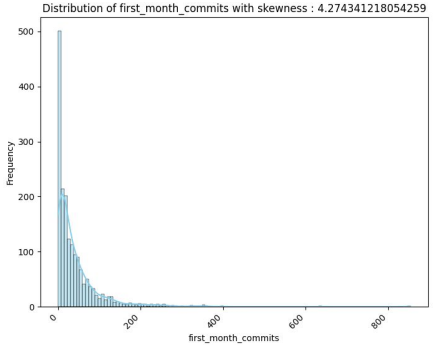


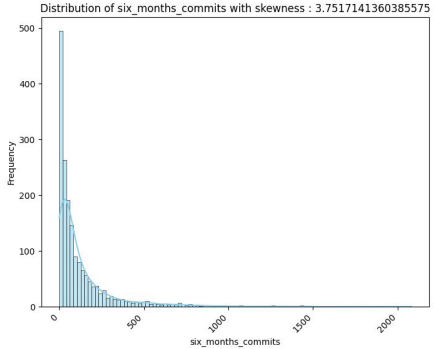


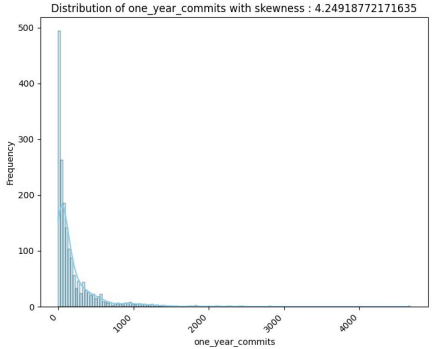


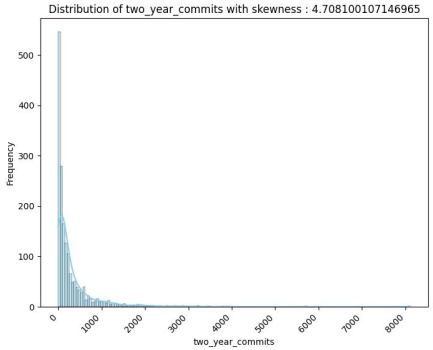
A really skewed dataset

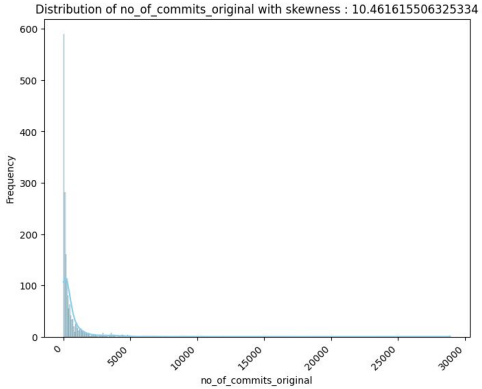
Features of the behaviour of original contributor

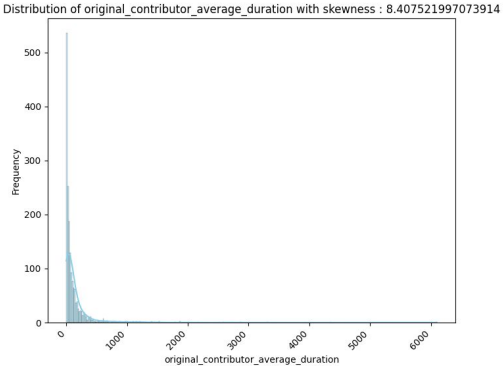






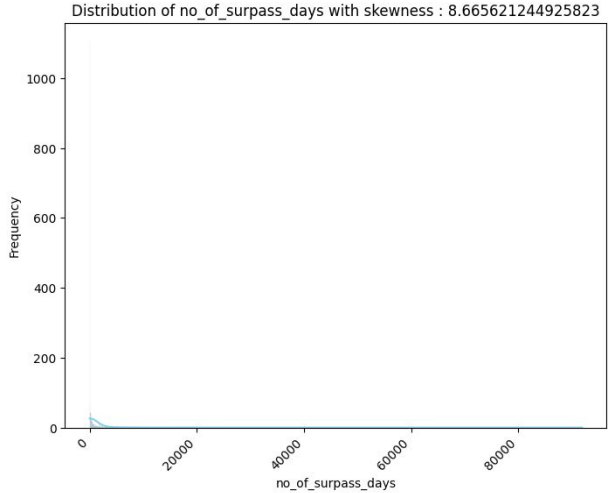


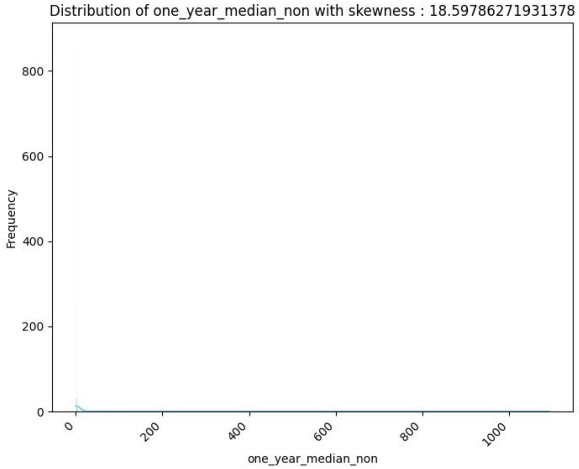


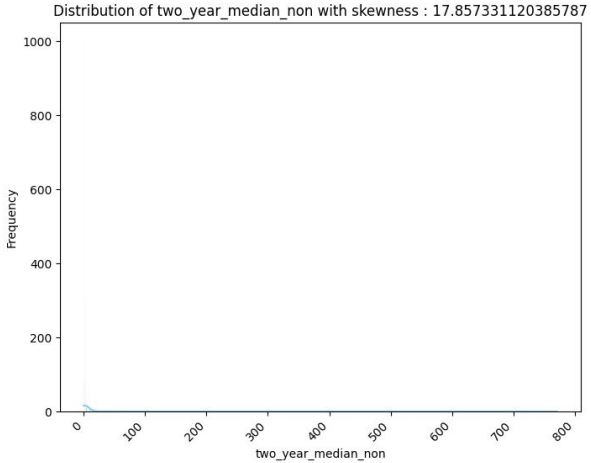


A really skewed dataset

Features of the behaviour of non-original contributors



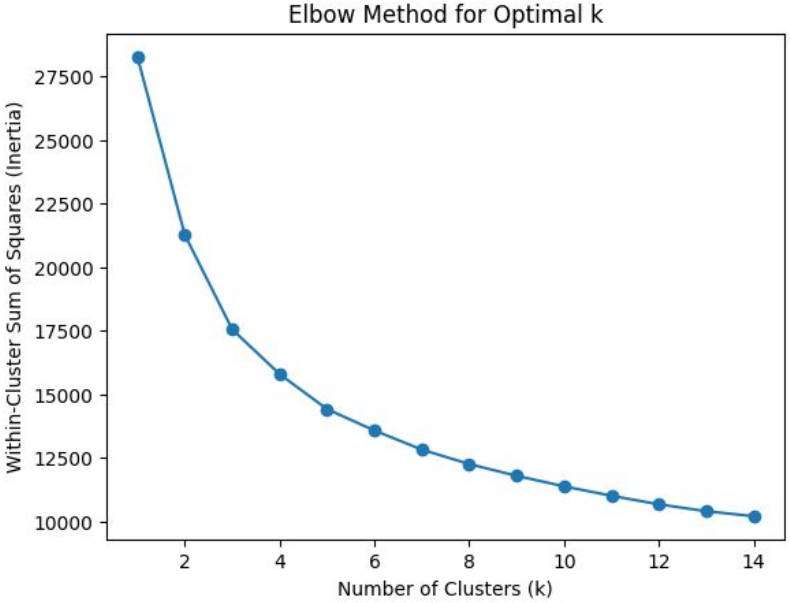


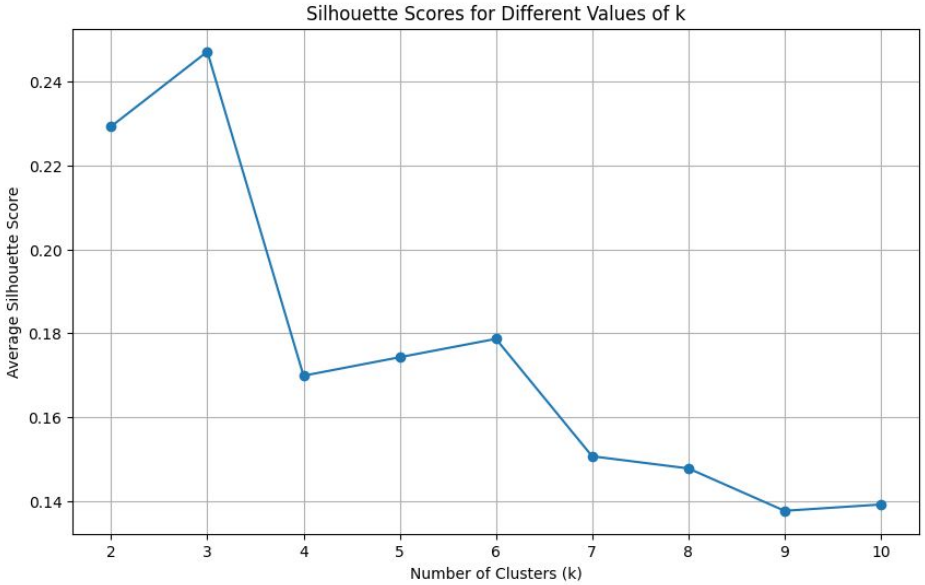


Clustering : K-means

How did we choose the optimal K? What results have we got?

Elbow method & Silhouette Scores





Then we choose k=3 for clustering

Cluster 0

- Observations : 256
- Moderate popularity (Stars)
- Not-so-active original contributor
  - Lower values for 1-month, 6-month, one-year, 2-year commits
  - Lower values for commits by original contributor
- Relatively active non-original contributors
  - Relatively high values for the median number of commits pushed by non-original contributors

	Mean	Median	Std
Stars	5004.730469	2701	3573.105174
Forks	943.90625	500.5	705.3833609
Pull	16.765625	5	9.181161869
Commit	2085.457031	391.5	710.9644362
Branches	9.1953125	3	5.141094654
contributor_numbers_in_total	69.64453125	25	22.8346179
first_month_commits	2.609375	1	37.93900416
six_months_commits	3.48046875	2	87.26835557
one_year_commits	3.95703125	2	128.1123408
two_year_commits	4.984375	2	184.556109
no_of_commits_original	27.21484375	2	331.589359
one_year_median_non	5.599609375	1	9.679818506
two_year_median_non	3.607421875	1.25	2.320202468
original_contributor_average_duration	147.5247933	0.983888889	332.3655242
non_original_contributor_average_duration	703.9824068	144.2180903	1146.294703
no_of_surpass_days	1506.265625	227.5	232.0393335
no_of_days	329.0390625	96	124.4701172

Cluster 1

- Observations : 567
- Higher popularity (Stars)
- Active original contributor
  - Higher values for 1-month, 6-month, one-year, 2-year commits
  - Higher values for commits by original contributor
- Active non-original contributors
  - Higher values for the median number of commits pushed by non-original contributors
  - Frequent commits by non-original contributors

	Mean	Median	Std
Stars	13991.37213	5707	23298.30792
Forks	3105.111111	1066	6380.55443
Pull	40.13227513	13	118.8698778
Commit	7129.181658	2226	35276.9239
Branches	47.99647266	9	430.2480507
contributor_numbers_in_total	228.4021164	101	480.8009445
first_month_commits	69.34215168	47	78.52851233
six_months_commits	264.345679	181	274.6240805
one_year_commits	443.8659612	316	455.6816891
two_year_commits	718.6631393	484	764.0135082
no_of_commits_original	1394.520282	762	2323.615376
one_year_median_non	12.93915344	1.5	63.0559745
two_year_median_non	9.989417989	1	52.22286667
original_contributor_average_duration	23.15771385	16.61999153	23.47463495
non_original_contributor_average_duration	62.6714286	35.15614037	96.15118588
no_of_surpass_days	3380.299824	312	9040.033202
no_of_days	866.5220459	589	810.276877

Cluster 2

- Observations : 944
- Lower popularity (Stars)
- Relatively active original contributor
  - Moderate values for 1-month, 6-month, one-year, 2-year commits
  - Moderate values for commits by original contributor
- Not-so-active non-original contributors
  - Lower values for the median number of commits pushed by non-original contributors
  - Lower values for surpass days

	Mean	Median	Std
Stars	3368.275424	2442.5	3573.105174
Forks	531.7065678	339.5	705.3833609
Pull	5.996822034	3	9.181161869
Commit	359.9025424	215.5	710.9644362
Branches	3.683262712	2	5.141094654
contributor_numbers_in_total	21.76271186	15	22.8346179
first_month_commits	32.68961864	20	37.93900416
six_months_commits	73.47139831	46	87.26835557
one_year_commits	106.4883475	67.5	128.1123408
two_year_commits	147.5741525	93	184.556109
no_of_commits_original	208.5264831	115	331.589359
one_year_median_non	1.914194915	1	9.679818506
two_year_median_non	1.59904661	1	2.320202468
original_contributor_average_duration	196.6258501	100.3423253	332.3655242
non_original_contributor_average_duration	641.4945338	298.3761806	1146.294703
no_of_surpass_days	59.79131356	0	232.0393335
no_of_days	107.9894068	69	124.4701172

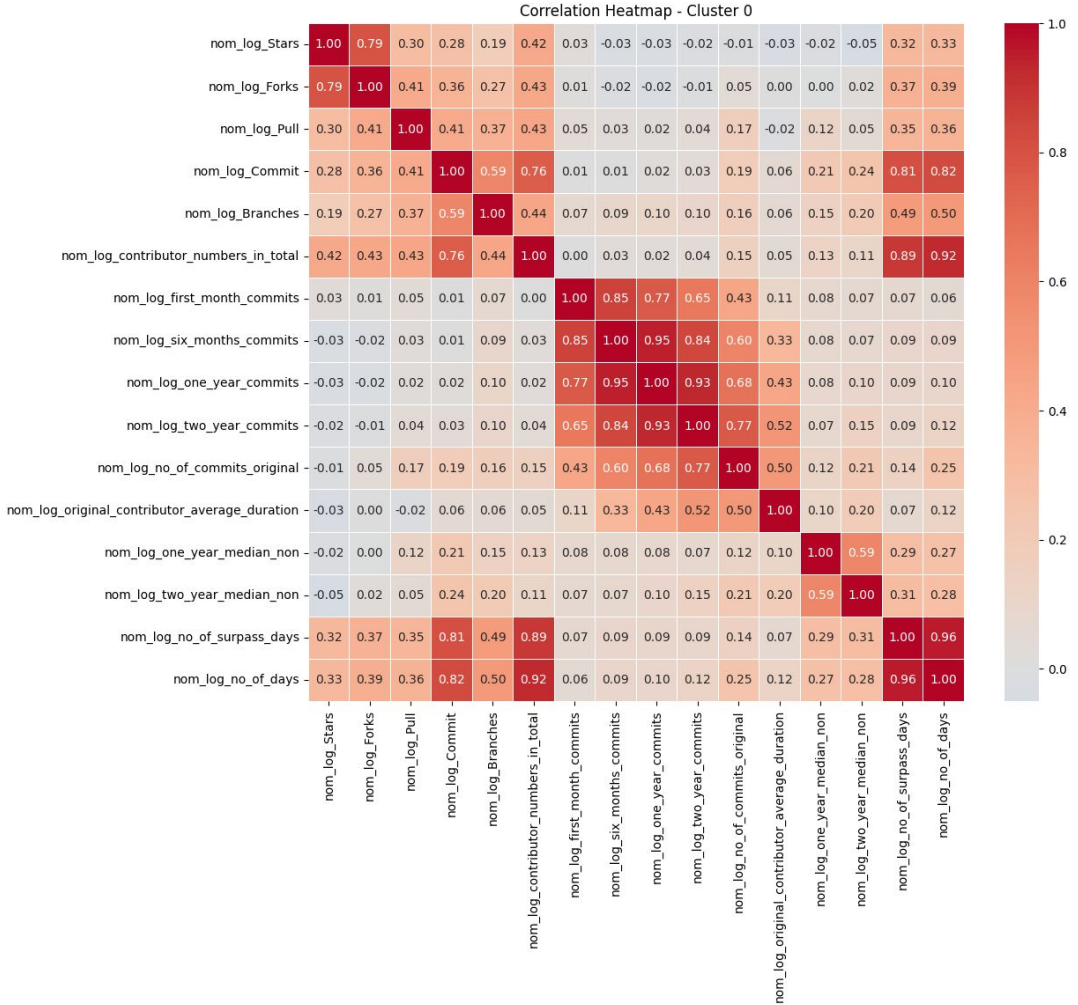
Heatmaps
Interpretation

Brighter colors are used to represent larger values, while cooler or darker colors represent smaller values.

Total number of contributors is strongly correlated to number of commits.

Total number of contributors is weakly correlated to number of commits in the first six months by the original contributor.

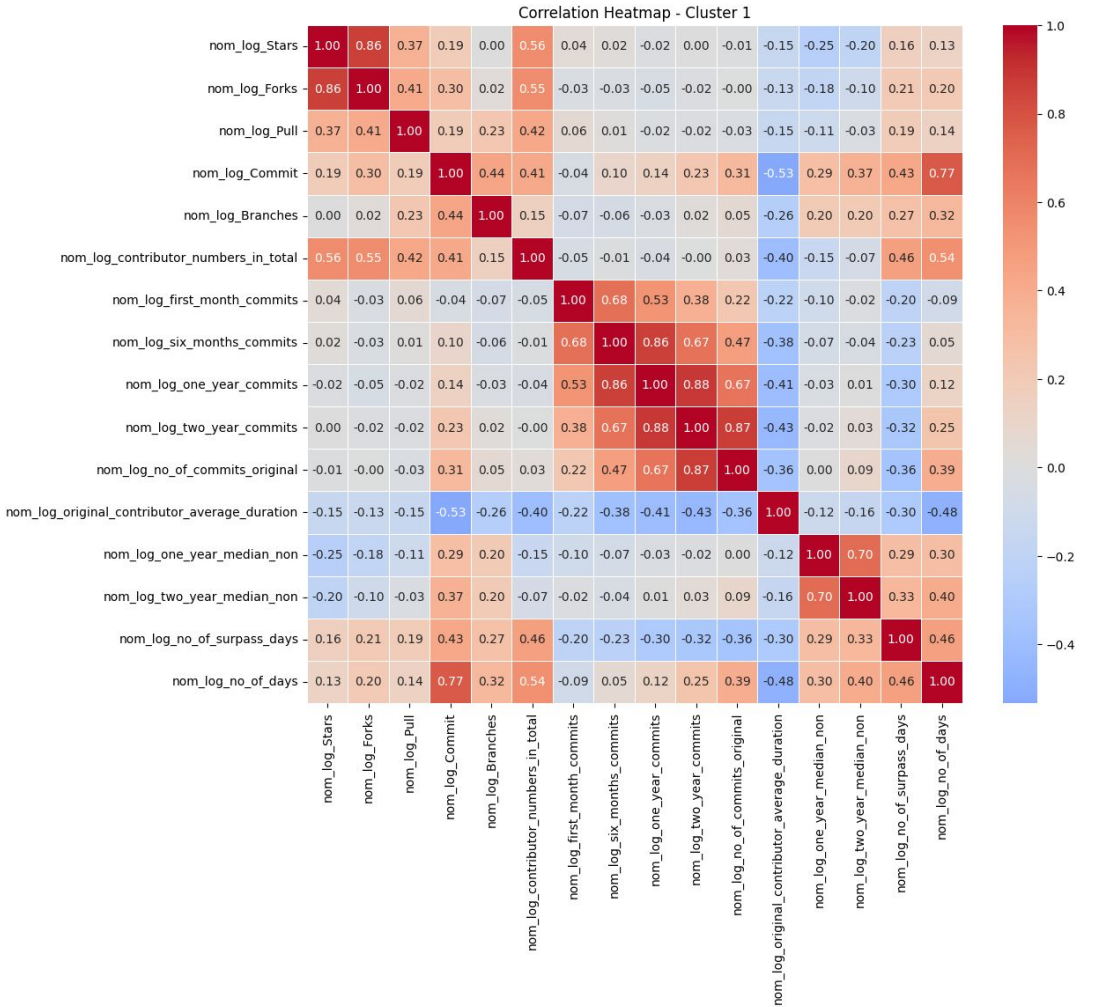
Number of commits in first year is strongly correlated to the average duration or original contributor



Total number of active days is strongly correlated to number of commits.

Total number of branches is weakly correlated to number of commits in the first year by the original contributor.

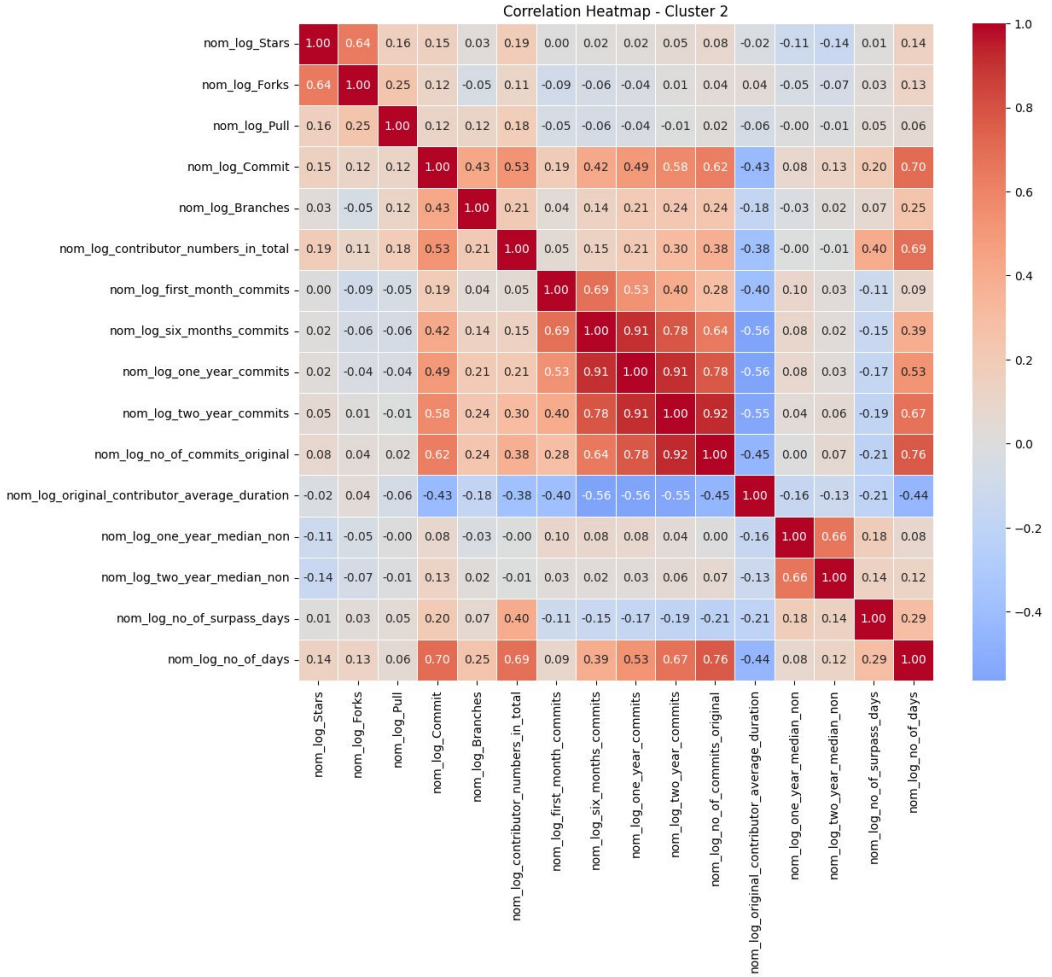
Total number of commits is strongly correlated to the average duration or original contributor



Total number of days is strongly correlated to number of commits.

Number of commits in first year is weakly correlated to total number of commits.

The heatmap shows that the behaviour is opposite to what we had earlier assumed that the behaviour of original contributor affected the popularity of a project.



Linear Regression

Features to select for Linear Regression

- Cluster 0:
  - No of contributors
  - No. of commits original
- Cluster 1:
  - No of contributors
  - OC average duration
  - Surpass days
- Cluster 2:
  - No. of contributors
  - 1 yr commits

Look out for:

- Heteroscedasticity
- Autocorrelation
- MSE

Features to select for Linear Regression

- Coefficients for Cluster 0:
  - No of contributors 0.176
  - No. of commits original 0.022
- Coefficients for Cluster 1:
  - No of contributors 0.96
  - OC average duration 0.15
  - Surpass days -0.12
- Coefficients for Cluster 2:
  - No. of contributors 0.36
  - 1 yr commits -0.11

- Mean Squared Error for Cluster 0: 0.47922558111941005
- P-values = [0.00, 0.09]
- Mean Squared Error for Cluster 1: 0.8454827333969253
- P-values = [0.00, 0.03, 0.01]
- Mean Squared Error for Cluster 2: 1.5257554081734643
- P-values = [0.00, 0.04]

Coefficients for Cluster 0: [0.18080723 0.00429922]

OLS Regression Results

Dep. Variable:	nom_log_Stars	R-squared (uncentered):	0.168
Model:	OLS	Adj. R-squared (uncentered):	0.166
Method:	Least Squares	F-statistic:	95.53
Date:	Mon, 20 Nov 2023	Prob (F-statistic):	1.60e-38
Time:	01:13:39	Log-Likelihood:	-994.64
No. Observations:	950	AIC:	1993.
Df Residuals:	948	BIC:	2003.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
nom_log_contributor_numbers_in_total	0.3677	0.027	13.452	0.000	0.314	0.421
nom_log_no_of_commits_original	-0.0744	0.045	-1.653	0.099	-0.163	0.014

Omnibus:	105.285	Durbin-Watson:	0.951
Prob(Omnibus):	0.000	Jarque-Bera (JB):	718.800
Skew:	0.213	Prob(JB):	8.21e-157
Kurtosis:	7.240	Cond. No.	1.86

Notes:

[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.

[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Coefficients for Cluster 1: [0.9678353 0.13718928 -0.10643942]

OLS Regression Results

Dep. Variable:	nom_log_Stars	R-squared (uncentered):	0.502
Model:	OLS	Adj. R-squared (uncentered):	0.499
Method:	Least Squares	F-statistic:	187.7
Date:	Mon, 20 Nov 2023	Prob (F-statistic):	3.56e-84
Time:	01:13:39	Log-Likelihood:	-753.74
No. Observations:	562	AIC:	1513.
Df Residuals:	559	BIC:	1526.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025
nom_log_contributor_numbers_in_total	0.9194	0.052	17.812	0.000	0.818
nom_log_original_contributor_average_duration	0.1741	0.082	2.129	0.034	0.014
nom_log_no_of_surpass_days	-0.0999	0.039	-2.587	0.010	-0.176

Omnibus:	3.279	Durbin-Watson:	0.851
Prob(Omnibus):	0.194	Jarque-Bera (JB):	3.222
Skew:	0.146	Prob(JB):	0.200
Kurtosis:	2.771	Cond. No.	3.65

Notes:

[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.

[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Coefficients for Cluster 2: [0.40063161 -0.08754915]

OLS Regression Results

Dep. Variable:	nom_log_Stars	R-squared (uncentered):	0.193
Model:	OLS	Adj. R-squared (uncentered):	0.187
Method:	Least Squares	F-statistic:	30.32
Date:	Mon, 20 Nov 2023	Prob (F-statistic):	1.57e-12
Time:	01:13:39	Log-Likelihood:	-336.39
No. Observations:	255	AIC:	676.8
Df Residuals:	253	BIC:	683.9
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
nom_log_contributor_numbers_in_total	0.4014	0.055	7.259	0.000	0.292	0.510
nom_log_one_year_commits	0.0609	0.030	2.000	0.047	0.001	0.121

Omnibus:	98.667	Durbin-Watson:	1.022
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1532.482
Skew:	-1.074	Prob(JB):	0.00
Kurtosis:	14.816	Cond. No.	1.84

Notes:

[1] R^2 is computed without centering (uncentered) since the model does not contain a constant.

[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Results for Linear Regression

- Lower values of MSE indicate a better fit of the model to the data.
- Cluster 0 has the lowest MSE, suggesting that the linear regression model performs relatively well on this cluster compared to the others.
- Lower the p-values, more significant the variable.
- Overall, we need a better model.

Thank you for your attention