Dataset Metadata — final_dataset_6000.csv

Data source

This dataset was compiled from the **SQuaD: The Software Quality Dataset (CSV)** (Version 1), hosted on Fairdata Etsin. The local dataset used in this project was created by extracting and merging three tables from SQuaD: **GITHUB_METRICS**, **PROCESS_METRICS**, and **RELEASES**.

Dataset landing page: https://etsin.fairdata.fi/dataset/2209b041-b35b-4863-8798-3204186b6b11

Citation: Robredo, M., Esposito, M., Taibi, D., Peñaloza, R., & Lenarduzzi, V. (2025). SQuaD: The Software Quality Dataset (CSV) (Version 1). University of Oulu. https://doi.org/10.23729/fd-c528d131-2c8c-3e61-91f1-a075931e73dc

Field descriptions

Brief descriptions of each attribute in **final_dataset_6000.csv**.

Field	Type	Description
id	int64	Row identifier from the dataset pipeline (if present).
n_releases	int64	Number of releases observed for the project (or up to this release, depending on mining).
release	object	Release tag name for the release (e.g., v1.2.3).
total_LOC	float64	Total lines of code (LOC) for the project (snapshot/approximation from mining pipeline).
t_lines	int64	Total lines changed in the release window (aggregate change volume).
lines_added	int64	Lines of code added in the release window (change volume).
max_lines_added	int64	Maximum lines added among commits in the release window.
avg_lines_added	float64	Average lines added per commit in the release window.
weighted_age	float64	Age-related metric weighted by change history (proxy for code maturity/recency).
n_fix	int64	Number of fix-related commits/changes in the release window (bug-fix activity).
n_auth	float64	Number of unique authors contributing in the release window.
churn	int64	Code churn in the release window (added + deleted, or similar change instability measure).
max_churn	int64	Maximum churn among commits in the release window.
avg_churn	float64	Average churn per commit in the release window.
max_change_set	int64	Maximum change set size among commits (largest commit footprint).
avg_change_set	float64	Average change set size per commit.
age	int64	Age of the project or component at the time of the release (days).
date_parsed	object	Release date parsed as a datetime; used to order releases chronologically within a project.
id_repo	float64	GitHub internal repository ID at time of collection.
full_name	object	GitHub repository full name (owner/repo).
owner	object	GitHub organization/user owning the repository.
repo_name	object	Repository short name.
updated_at	object	Repository last updated timestamp from GitHub metadata.
pushed_at	object	Repository last push timestamp from GitHub metadata.
size	float64	Repository size reported by GitHub (KB).

Field	Type	Description
language	object	Primary language as reported in GitHub metadata at time of data collection.
forks_count	float64	Number of forks reported by GitHub at time of collection.
stargazers_count	float64	Number of stars reported by GitHub at time of collection.
watchers_count	float64	Number of watchers/subscribers reported by GitHub at time of collection.
has_issues	object	Whether GitHub Issues is enabled for the repository (boolean).
archived	object	Whether repository is archived (boolean).
disabled	object	Whether repository is disabled (boolean).
fork	object	Whether repository is a fork (boolean).
num_contributors	float64	Number of contributors (as mined/queried by the dataset pipeline).
sbom_flag	object	Whether repository has an SBOM-related flag in the dataset pipeline (boolean/indicator).
num_commits	float64	Total number of commits in repository (as mined/queried).
num_buggy_commits	float64	Number of commits labeled buggy/defect-inducing (from dataset pipeline).
lifetime_days	float64	Repository lifetime in days (from first to last observed activity).
num_languages	float64	Number of programming languages detected in repository.
num_stars	float64	Stars count (duplicate/alternative star metric from pipeline).
main_language	object	Main language label used by the dataset pipeline (may differ from GitHub 'language').
num_files	float64	Number of files detected in repository snapshot.
num_methods	float64	Number of methods/functions detected (static analysis metric).
num_buggy_files	float64	Number of files labeled buggy/defect-inducing (pipeline).
num_buggy_methods	float64	Number of methods labeled buggy/defect-inducing (pipeline).
num_bugs	float64	Number of bugs/defects associated with the release or mined window (from dataset pipeline).
defect_prone	int64	Target (classification): 1 if release labeled defect-prone, else 0 (derived from fix activity).
effort_next_lines_added	float64	Target (regression): next release effort proxy = lines_added in the subsequent release.
effort_next_n_auth	float64	Aux target: number of authors in the subsequent release (optional effort proxy).