DATA542: Data Wrangling - Project Milestone 1

Preethi Vezhavendan, Mohammad Zaed Iqbal Khan

Research Question 1: To what extent do the different Autonomous Coding Agents differ in efficiency, as measured by the time from PR creation to PR closing and the number of iteration cycles required (commits, reviews, and change requests)?

Methodology: The Agents' efficiency will be evaluated by measuring the full lifecycle of their pull requests, as well as the number of review iterations required to complete each one. The timeline will be calculated by taking the difference between the **created_at** and **closed_at** timestamps for each pull request, grouped by **id** and **agent**. Pull requests without a **closed_at** timestamp will be excluded. The number of iterations will be determined by merging the **pull_request** and **pr_reviews** tables, filtering for closed pull requests, and counting the number of review records associated with each pull request ID. Descriptive statistics and comparative visualizations will be done to identify patterns, differences, and trends across agents.

Research Question 2: How does the type of reviewer - human, bot, or a combination of both - impact the efficiency of AI-generated pull requests, as measured by the time from PR creation to closure and the number of iteration cycles required for completion?

Methodology: To evaluate how reviewer type affects the efficiency of AI-generated pull requests (PRs), we will first calculate iteration cycles by grouping **pr_timeline_df** by **pr_id** to count the iteration cycles per PR. We will then filter **pr_df** for closed PRs, slicing and converting **created_at** and **closed_at** to datetime objects to compute the lifecycle in hours for each PR. Iteration cycles and lifecycle hours will be merged with reviewer information from **pr_reviews_df**, retaining the **user_type** of reviewers. The resulting dataframe, cleaned for duplicates and missing values, will include, for each PR, its iteration cycles, lifecycle duration, and reviewer type. Efficiency will be analyzed by comparing distributions of lifecycle hours and iteration cycles across three reviewer categories - User, Bot, and Both (combined human and bot reviewers) - using **boxplots** and **descriptive statistics**.

Research Question 3: How does the tone of reviewer feedback - positive, negative, or neutral - affect the efficiency of AI-generated pull requests, as measured by the time from PR creation to closure and the number of iteration cycles required for completion?

Methodology: To investigate how the tone of reviewer feedback influences the efficiency of AI-generated pull requests (PRs), we will leverage the merged dataframe from RQ2, which contains **pr_id**, **iteration_cycles**, and **lifecycle_hours**, with the latter two serving as measures of efficiency. We will extract the **pr_id** and **body** attributes from **pr_reviews_df** and merge them with the efficiency dataframe, where body contains the textual review messages. Sentiment analysis will be performed on each review message using NLTK's SentimentIntensityAnalyzer to classify the feedback as positive, neutral, or negative. Efficiency will then be analyzed by comparing the distributions of lifecycle hours and iteration cycles across these sentiment categories using **box plots** and **descriptive statistics.**