Rhyme Take-Home Challenge

Goal

Clean and integrate user and event data, producing a usable dataset for analytics.

Steps

1. Email Normalization: Trim, lowercase, and Unicode-normalize emails (remove \u00A0, \u200B, homoglyphs).
2. Name Splitting: Extract first and last names while discarding prefixes and suffixes (e.g., "Dr.", "Jr.").
3. Email Matching: Joined events with users using normalized emails.
4. Unmatched Event Logging: Identified and logged events without a user match.
5. Duplicate Detection: Reported duplicate emails from users.
6. Event Summary: Generated a CSV showing number of events per user.
7. Logging: Used Python logging for key steps.

Output Files

- joined_events.csv: Final joined and cleaned dataset
- unmatched_event_ids.csv: Events that didn't match a user
- duplicate_emails.csv: Detected duplicate user emails
- user_event_counts.csv: Number of events per user

Notes

- Invalid or malformed emails (e.g. missing domains) were not fixed, only normalized.

- Events with no matching user were excluded from the final dataset but logged separately.