Project scope update:

I slightly refined the scope. Originally I planned to use three sources that all contained overlapping information. I replaced one source with a sales dataset to enable a cross-dataset comparison (ratings vs. sales). The final goal is unchanged: analyze game ratings and predict Steam's positive rating ratio (positive_reviews / total_reviews). I standardized genres across sources (a canonical genre_std), filtered to 2000–2020, removed NSFW/non-game rows, and required at least 10 reviews for Steam when computing ratios and running the regression. For full pipeline details (collection, cleaning, and reproduction), see **README.md.**

Data sources

Collected in code (src/data_collection.py) and saved into data/ (but not committed).

- **RAWG API** – unfiltered sample (10k) of games with fields: name, rating, released, platforms, genres. Used for Questions #1–2 (genre/platform analysis).
- **Steam (Kaggle via API)** – positive_ratings, negative_ratings, price, average_playtime, owners, platforms, publisher, categories, release_date, etc. I compute positive_rating_ratio = positive / (positive+negative). Used for Questions #3–5 and as the **target** in the prediction model.
- **Video Game Sales (Kaggle via API)** – global & regional sales + Year, Platform, Genre. Used for Question #6 (ratings vs. sales, by canonical genre).

APIs used.

- **RAWG** REST API (/api/games), authenticated with RAWG_API_KEY from .env.
- **Kaggle API** for programmatic downloads; credentials provided via env vars or a local kaggle.json (both excluded from Git).

Preprocessing highlights (src/data_cleaning_preprocessing.py).

- Canonical genre_std mapping (e.g., rpg→role-playing; platformer→platform; indie/casual→misc).
- Year extraction (released → year) and windowing to 2000–2020.
- Save cleaned CSVs: rawg_clean.csv, steam_clean.csv, vgsales_clean.csv.

Issues / difficulties

- **Rate limits/latency** when collecting RAWG; solved by paging sequentially, light sleeps, and caching CSVs so re-runs skip collection.
- **Schema/label mismatch** (genres differ across sources); solved with the canonical genre_std and lower-casing/stripping.
- **Model performance**: baseline linear regression has modest R2R^2R2. Next steps: try log-transforms (price, owners), add review-volume as a feature, test regularized models, and consider interaction terms (genre×platform, year trends).
- **Testing**: added a lightweight src/tests.py that (1) smoke-tests RAWG, (2) checks file existence & schemas, (3) validates cleaned year ranges, and (4) verifies results/ is writable.