SAMPLE DELIVERABLE · WEB SCRAPING & DATA EXTRACTION

126,359 games, one clean dataset.

A public store catalogue scraped, cleaned, validated and turned into a dataset, a data dictionary and a short insight report.

Prepared by
Gergely Racz

Company
LeadFlow Automation

Web
leadflowautomation.net

Snapshot
31 Oct 2026

01 · THE JOB

What was scraped, and why

The brief behind this sample: *"Which kinds of games sell, and which niches are crowded?"* Answering it needs the whole catalogue, not a handful of bestsellers, so the job was to collect every game listed on a public online store and put it into one consistent table.

The result is one row per game with its title, release date, price, discount, review count, rating and up to seven tags. Public catalogue data only: no logins, no personal data, nothing behind a paywall.

126,359
unique games (after dedupe)

13
fields per row

~2 h
one full crawl, 2,500+ pages

429
distinct tags resolved to names

Scale at a glance

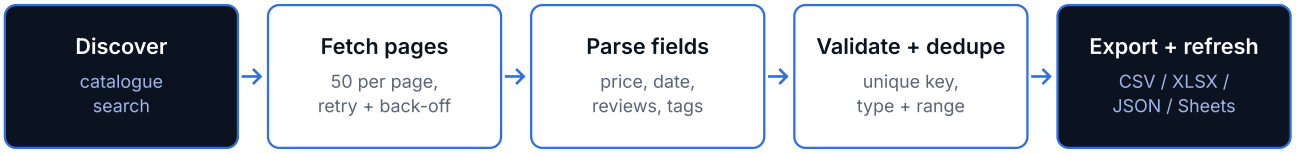
Measure	Value
Rows collected / unique	126,622 / 126,359
Release years covered	1998 to 2026 (23,617 titles released in 2026 so far)
Review volume captured	73.5 million user reviews across the catalogue
Paid / free titles	107,306 paid, 19,053 free
Refresh cadence	One full snapshot for this sample. The same pipeline runs daily or weekly with date-stamped snapshots, so you can track change over time.

Why a buyer cares

- A full census beats a sample: niche sizes, price bands and review concentration are measured, not guessed.
- Every row links back to its source page, so any number can be spot-checked in seconds.
- The same approach works for product catalogues, directories, job boards, listings, app stores and pricing pages.

02 · METHOD

How the data was collected



Built in **Python**. Where a site needs a real browser (scripts, logins you own, infinite scroll) the crawler uses **Playwright**; where a scheduled, hosted run is the better fit it is packaged as an **Apify** actor.

Crawling

- **Pagination.** The catalogue was read page by page (50 games per page, 2,500+ pages), with the position saved after every page so an interrupted run resumes instead of restarting.
- **Rate limits.** Polite, randomised delay between requests. When the site answered "too many requests" (84 times in this run) the crawler waited, doubled the wait on repeats, and retried the same page until it succeeded.
- **Retries.** Network errors, empty responses and unexpected pages are retried with back-off before a run is allowed to fail loudly.
- **Age gate.** Handled with the same consent the site asks a normal visitor for, so mature titles are not silently dropped from the census.

Cleaning & validation

- **Dedupe** on the store's own unique ID: 263 repeated rows removed.
- **Type conversion.** Prices from cents to dollars, review counts from text ("12,345 reviews") to integers, three date formats to ISO 8601.
- **Lookups.** Numeric tag IDs translated to readable names using the store's own tag table; unknown IDs are reported instead of silently dropped (there were none).
- **Honest blanks.** Anything the page does not show stays empty. Nothing is guessed or back-filled.
- **Range checks.** No future release dates, no negative prices, ratings within 0 to 100.

Compliance note. Only publicly visible catalogue pages, at a modest request rate, with no personal data collected. Site terms and robots rules are reviewed per target before a project starts.

03 · DATA DICTIONARY

Every field, explained

Also included as a sheet inside the Excel file. "Filled" is measured on the full 126,359-row dataset.

Field	Type	Meaning	Filled
appid	integer	Steam's unique numeric ID for the game; used as the primary key for deduplication.	100%
name	text	Game title as shown on the store page (HTML entities decoded).	99.99%
store_url	text	Canonical store link built from appid; lets you click through and spot-check any row.	100%
release_date	date (YYYY-MM-DD)	Release date normalised from three different on-page formats to ISO 8601.	99.99%
release_year	integer	Year part of release_date, for easy grouping and pivots.	99.99%
current_price_usd	decimal	Price in US dollars at scrape time, after any active discount. 0.00 = free.	100%
discount_pct	integer 0-100	Active discount at scrape time; 0 = full price.	100%
is_free	yes / no	Convenience flag: yes when current_price_usd is 0.	100%
review_count	integer	Number of user reviews shown. The store hides the count below 10 reviews, so 0 means fewer than 10.	100%
positive_review_pct	integer 0-100	Share of reviews that are positive. Blank when the store shows no rating.	56.8%
review_rating	text	Store rating label, from Overwhelmingly Positive to Overwhelmingly Negative. Blank when no rating is shown.	56.8%
primary_tag	text	First (most relevant) of the game's user-defined tags, translated from numeric ID to name.	99.98%
tags	text, ';'-separated	Up to 7 top tags in the store's own order of relevance, translated from numeric IDs to names.	99.98%

04 · QUALITY ASSURANCE

What was checked before delivery

Check	Result	Notes
Rows collected (raw)	126,622	Every row returned by the crawl
Duplicate rows removed	263 (0.21%)	Pages shift while the catalogue is being re-sorted; the same game can appear twice. Key: appid
Unique games kept	126,359	One row per game
Release date parsed	126,348 (99.99%)	3 date formats normalised to ISO; 11 rows have no date on the page and are left blank, not guessed
Name present	126,358 (99.99%)	1 row has no title on the page and is kept as-is
At least one tag	126,330 (99.98%)	All 429 tag IDs resolved to names; 0 unknown IDs
Full set of 7 tags	110,857 (87.7%)	Smaller or newer games carry fewer tags
Price present	126,359 (100%)	Integer cents converted to USD; 19,053 free titles flagged
Review rating present	71,778 (56.8%)	Not a scraper gap: the store hides ratings below 10 reviews
Release dates in the future	0	Checked against the snapshot date

Reading the numbers

Core fields (ID, name, date, price, tags) are **99.98% to 100% complete**. The one field below that, review rating at 56.8%, reflects how the store works: ratings only appear from 10 reviews, and 54,581 games (43.2%) are below that line. This is documented in the data dictionary so nobody mistakes it for missing data.

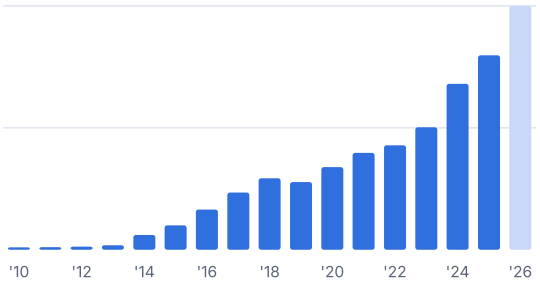
Spot checks

- Every row carries a store_url, so any value can be opened and compared with the live page in seconds.
- Totals reconciled: rows collected minus duplicates equals rows delivered.
- The excerpt supplied with this report is a random draw of 400 general-audience titles; the full file includes every game.

Four charts from the data

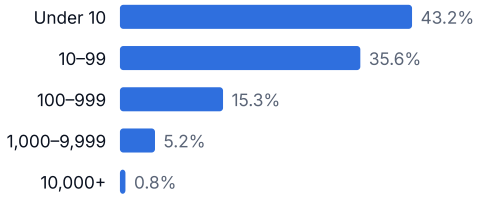
Releases per year

New games listed each year, 2010 to 2026. 2026 is a partial year (to 31 Oct) and already the biggest.



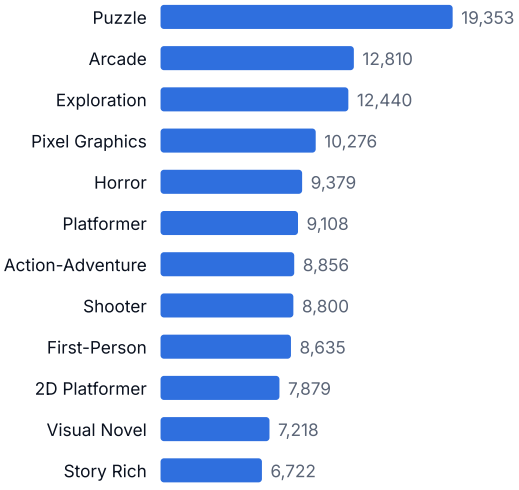
Most games stay unseen

Share of all 126,359 games by number of user reviews. 43% have fewer than 10.



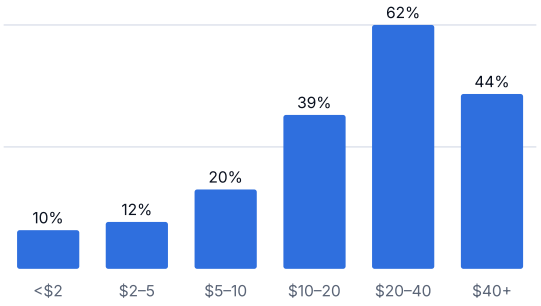
Top 12 tags

Games per tag. Broad labels (Indie, Action, Singleplayer...) and adult tags excluded.



Higher price, better odds

Share of paid games reaching 100+ reviews, by price band. Odds rise with price up to $40.



All four charts are computed from the full 126,359-row dataset, not from the 400-row excerpt.

06 · THE DATASET

What the table looks like

First 16 rows of the 400-row excerpt, 7 of 13 columns shown. Filters and a frozen header row are set up in the Excel file.

appid	name	release_date	current_price_...	review_c...	review_rating	primary_tag
240	Counter-Strike: Source	2004-11-01	9.99	45385	Overwhelmingly P...	Shooter
400	Portal	2007-10-10	9.99	85100	Overwhelmingly P...	Puzzle
11370	Nikopol: Secrets of the Immortals	2009-07-30	4.99	266	Mixed	Choose Your ...
25910	Supreme Ruler 2020 Gold	2009-07-09	9.99	207	Very Positive	Modern
35000	Mini Ninjas	2009-09-08	9.99	1469	Very Positive	Adventure
62100	Chime	2010-09-06	4.99	653	Very Positive	Casual
73010	Cities in Motion	2011-02-22	19.99	695	Mostly Positive	Simulation
218230	PlanetSide 2	2012-11-20	0.00	47567	Very Positive	Free to Play
219740	Don't Starve	2013-04-23	2.49	33113	Overwhelmingly P...	Survival
251310	Eleusis	2013-10-16	14.99	276	Mixed	Adventure
252730	YOU DON'T KNOW JACK Vol. 1 XL	2013-11-05	2.99	155	Very Positive	Trivia
259510	Shufflepuck Cantina Deluxe	2013-12-06	9.99	221	Very Positive	Sports
269850	Get Packed: Fully Loaded	2021-07-28	19.99	68	Mixed	Multiplayer
284140	Majestic Nights	2014-10-30	4.99	54	Mixed	Adventure
298220	Terraformer Expedition to Mars	2015-11-20	24.99	29	Mostly Negative	Simulation
300550	Shadowrun: Dragonfall - Director...	2014-09-18	14.99	4226	Very Positive	RPG

Headline findings (full data)

- **21.2%** of games reach 100 reviews; only **1,021** (0.8%) pass 10,000.
- The median paid game costs **$4.99**, unchanged in every release year since 2016.
- Releases grew from 6,919 in 2018 to **23,617 in 2026** (still counting).
- Puzzle is the most crowded tag, with **19,353** games.

Your own dataset, your format

Every project includes

- **The dataset** as CSV and Excel (formatted header, frozen pane, filters), plus JSON or a Google Sheet on request.
- **Data dictionary** with field types, meaning, examples and completeness.
- **QA summary**: rows in/out, duplicates removed, completeness per field, spot-check results.
- **Summary report** with charts and plain-language findings, like this one.
- **Optional scheduled refresh** (daily, weekly or monthly) delivered to a Sheet, a database or your inbox, with failure alerts.

A second example: Apify Store tracker

A scheduled tracker that snapshots the public Apify Store every day: **50,867 scraping tools** from 3,553 publishers in a single snapshot, with 30-day usage, ratings and pricing model. Together those tools were run more than **240 million times** in the 30 days before the snapshot, and the daily history makes week-on-week momentum measurable.

Also public on GitHub: a market-intelligence platform that turns app-store data into opportunity reports (github.com/zahor99/market-intelligence-platform).

Python	Playwright	Apify
crawlers, cleaning, validation	when a real browser is needed	hosted, scheduled runs

Next step

Send the site or catalogue you need data from and the fields you want. You get a scoped plan, a small sample to review, then the full run.

Gergely Racz · LeadFlow Automation · leadflowautomation.net