QUALITY REPORT

Trusted Zone

Introduction and Purpose

The Trusted Zone is the cleaned, standardized, and validated data. Data arriving from the Formatted Zone undergoes processes before being stored here, for quality assurance.

The goal of these processes is to make the data consistent, and ensure we can consume the data with confidence, knowing that fundamental structural, format, and content issues have been addressed.

This report details the specific processes implemented for JSON and image data, the rationale behind them, and an analysis of remaining limitations.

Integrity Checks

Firstly, we check that **`game_id`** is unique within both **`steam`** and **`steamspy`** datasets:

```
Steam: Confirmed. All 100 game_ids are unique.
SteamSpy: Confirmed. All 100 game_ids are unique.
```

We get the number of **`game_id`**'s present in both datasets, **`game_id`**'s present uniquely in Steam, and present uniquely in SteamSpy:

```
Number of games in both datasets: 100
Number of games only in Steam: 0
Number of games only in SteamSpy: 0
```

Now we set the list of **most relevant attributes** to be analyzed:

STEAM_IMPORTANT_FIELDS = [
  "name", "release_date", "required_age", "price", "dlc_count",
  "detailed_description", "about_the_game", "header_image", "windows", "mac", "linux",
  "metacritic_score", "achievements", "recommendations", "developers",
  "publishers", "categories", "genres"
]

STEAMSPY_IMPORTANT_FIELDS = [
  "positive", "negative", "estimated_owners", "average_playtime_forever",
  "median_playtime_forever", "tags"
]

And we check missing values within the defined relevant fields:

```
Checking 100 entries in Steam for missing important fields...
Removed 3 entries (3.0%) from Steam due to missing important fields.
97 entries remain in Steam.
Breakdown of missing fields causing removal in Steam:
  - Field 'about_the_game' was missing 1 times.
  - Field 'release_date' was missing 1 times.
  - Field 'genres' was missing 1 times.

Checking 100 entries in SteamSpy for missing important fields...
Removed 0 entries (0.0%) from SteamSpy due to missing important fields.
100 entries remain in SteamSpy.
```

We can see that 3 instances have one missing value in the following fields: **`about_the_game`**, **`release_date`**, **`genres`**.

We validate the data type for most relevant fields. We have defined a dictionary for each dataset that maps each field with its proper data type:

```
Total number of type mismatches found in Steam: 0
Total number of type mismatches found in SteamSpy: 0
```

Numerical Fields Analysis

Firstly, we carry out a descriptive analysis of numerical fields:

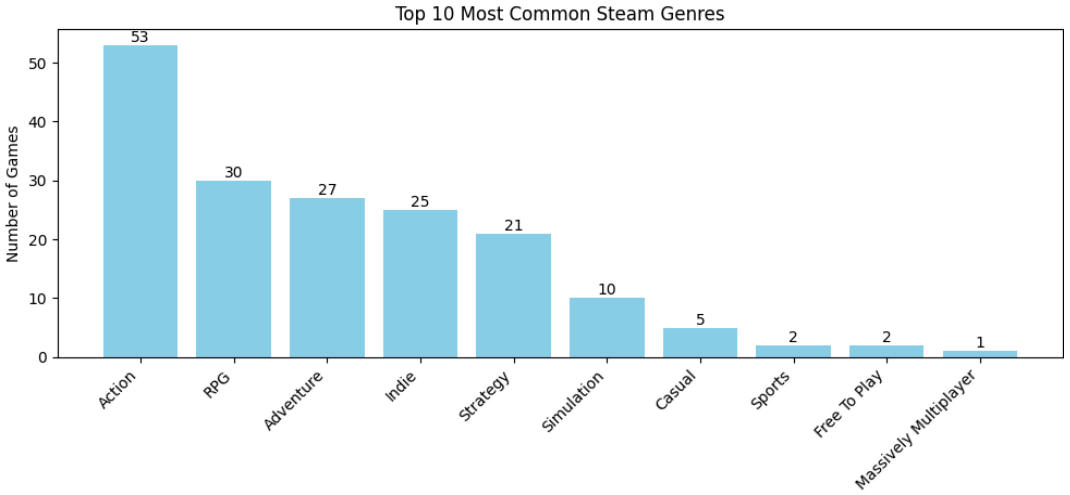
Campo	Count	Min	Max	Mean	Median	Stdev
price	100	0.0	69.99	19.0263	19.99	17.21070624
metacritic_score	100	90	97	91.91	91.0	1.75289083
dlc_count	100	0	245	4.51	0.0	24.71146424
achievements	100	0	520	39.02	20.0	69.59826713
required_age	100	0	18	5.95	0.0	8.15057165
recommendations	100	0	1843282	86285.75	19482.0	228192.8615
positive	100	101	1739980	99508.06	19553.5	245687.9726
negative	100	2	250576	8319.33	1538.5	28933.39045
average_playtime_forever	100	0	22164	1996.09	933.0	3205.23717
median_playtime_forever	100	0	5793	756.52	362.0	1138.671577
peak_ccu	100	0	67851	2680.21	106.0	9160.139638

We can see that many fields present significant right asymmetry, where the **mean is much higher than the median**. For example:

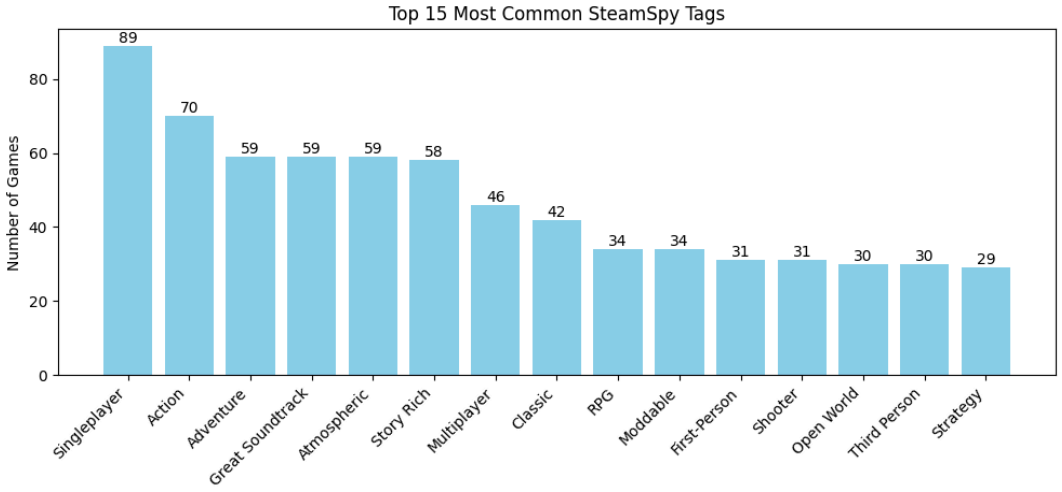
- **dlc count** (median 0 and mean 4.4), that means most of the games have no DLCs, but few have many.
- **required age** (median 0, mean 5.6),
- **recommendations**, **positive**, **negative ratings**, and **peak ccu** have high maximums and standard deviations, which could mean that few very popular games dominate these metrics.
- **price** distribution seems to be less asymmetric with more similar values of mean and median.
- **metacritic score** is ranged only from 90 to 97 due to the filter applied to keep only the 100 highest rated games.

Content Analysis

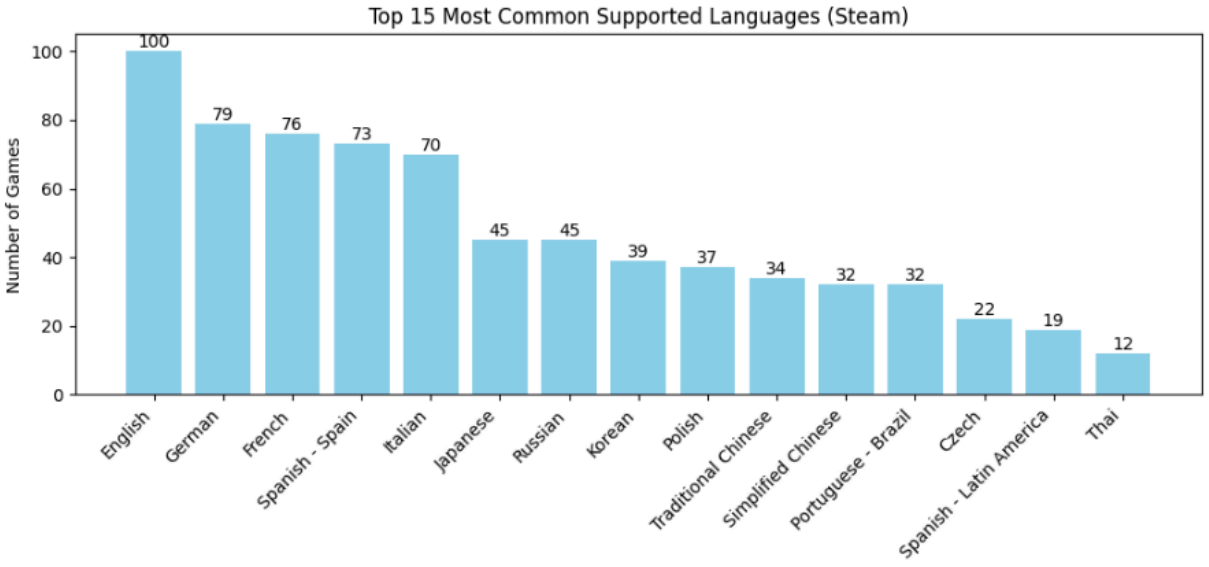
First of all, we check the **frequency distribution** of genres, tags, supported languages, and description lengths.



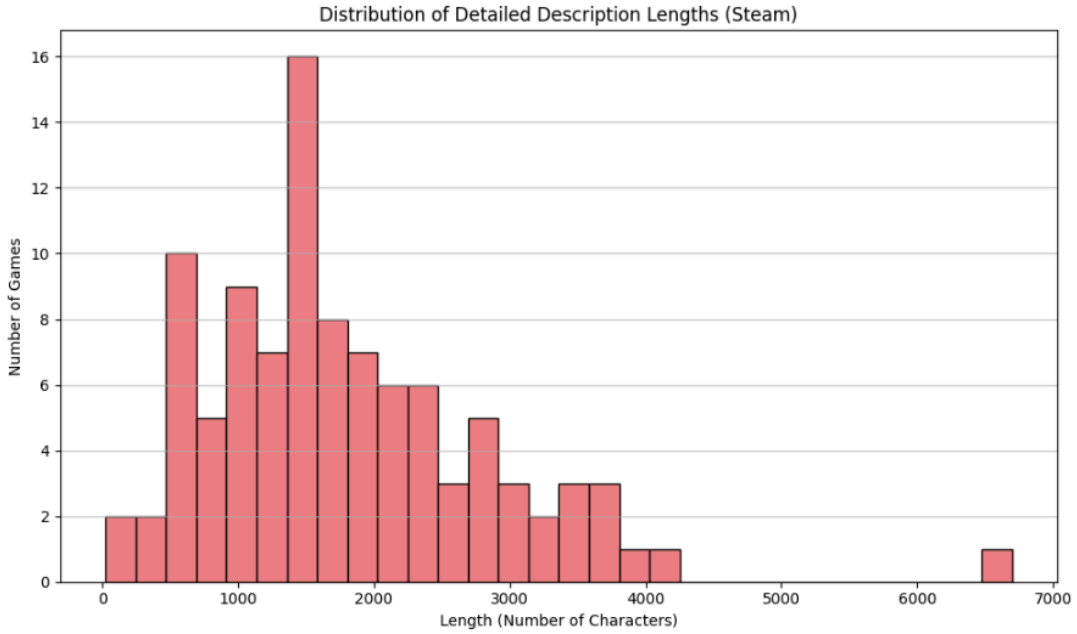
The distribution is dominated by the **Action** genre, **RPG**, **Adventure** , **Indie**, and **Strategy** are also common (20-30 games each). Remaining genres appear much less frequently, indicating a strong focus on action/adventure/RPG within the dataset (100 highest-rated games).



Singleplayer is the most frequent tag (almost 90 games), followed by **Action** (70 games). **Adventure**, **Great Soundtrack**, **Atmospheric**, and **Story Rich** are also common (around 60 games each). The distribution here is less asymmetric than genres.



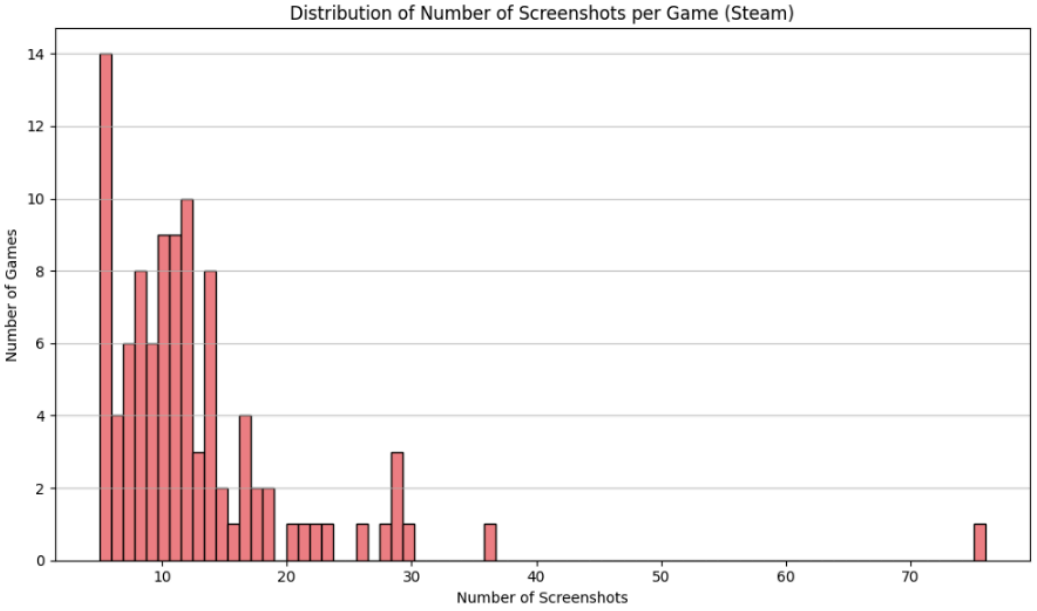
English is the most supported language, present in all 100 games analyzed in the dataset. **German**, **French**, **Spanish (Spain)**, and **Italian** are also very common (70-80 games). This shows a strong focus on European languages after English.



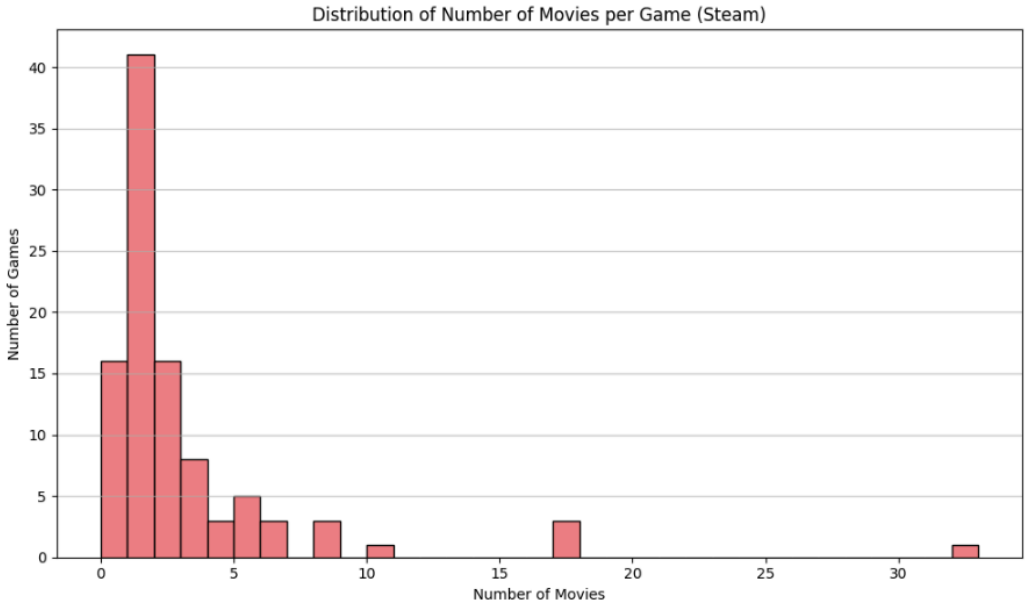
The histogram shows a **right-assymetryc distribution**. Most game descriptions have between 500 and 2,500 characters. There are fewer games with very short descriptions (probably because these descriptions are not preprocessed and they include information about updates, advertisements, or other irrelevant text).

Finally, we check missing and average number of screenshots and videos in the dataset, where we can see that **16% of the top 100 games do not include a video**.

Now, we analyze the distribution of screenshots and videos per game:



The histogram shows a **right-assymetric distribution** of screenshots per game. The majority of the games have a relatively small number of screenshots (5-15). There appears to be a significant outlier game with approximately 75 screenshots.



The distribution shows a similar behaviour, a **right-assymetric distribution,** where the majority of the games have **0, 1, or 2 videos**. In this case there is also an outlier with around **32 videos**.