Include title page for documentation!

# 1. Motivation[[1]](#footnote-1)

**Datasheets for Datasets -> pick a good name for dataset  
*Adapted from:* *Gebru, Morgenstern, Vecchione, Vaughan,   
Wallach, Daumeé, and Crawford. (2018). Datasheets for Datasets.***

**Derive**

***1.1*** *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

In 2019, GlobalData reports indicated that the video game industry’s value would exceed $300 billion by 2025, becoming one of the world’s biggest industries worldwide (Lanier, 2019). The Covid-19 pandemic was as unpredictable as it was beneficial for the video game industry, with MarketWatch analyses reporting massive financial gains due to the accommodating pandemic conditions (TrtWorld, 2021). As a result, Accenture reports that as of mid-2021 the video game industry had already exceeded its forecasted $300 billion value expected for 2025 (Nolibois, 2021).

These are all impressive financial achievements that show the general public’s interest in video games as an entertainment media. However, despite this remarkable growth, the video game industry at large is highly secretive with its information. This is not limited only to publisher or developer studios, but even game distribution platforms such as Steam, PlayStation Store and others. This has been a noted problem since 2017, leaving third party groups as major procurers of data and estimations (Dring, 2017). This indicates a clear issue for any academic or industry participant that would like to investigate statistical phenomenons within the industry as the data they may require is hard to find or non-existent. Interestingly (and not necessarily related to this project), the industry’s secrecy is also pointed inwards towards employees and developers with negative consequences (Schreier, 2018).

As a result of the scarce availability of video game industry data, available research is consequently influenced. Research observing how video games as media affects us mentally or cognitively is far more prevalent and varied than statistical studies done within the industry. Furthermore, statistical studies performed within the industry may very well rarely use complete datasets (archival data) due to their scarcity,

opting instead for the usage of surveys and other instances of field data.

To aid against the lack of available data on the video game industry, this team proposes the creation of a dataset through the use of Steam API. Steam API enables the collection of various data points pertaining to users and video games available in its online retail store. This project specifically will focus on the collection of software datapoints as provided by Steam in an attempt to create a dataset that includes a comprehensive list of software available for sale.

It is important to note several limitations in using Steam and its API. To begin with, while Steam is the leading gaming PC online store, it is by no means the only available one (i.e. Epic Store, GOG.com, etc.). Furthermore, Steam is only available on PC, other gaming platforms such as PlayStation or Nintendo Switch utilize different online stores to facilitate game sales. This can be highly relevant for some games, as they may be platform or store exclusives. Lastly, as Steam is an online store exclusively, this data will not cover physical retail sales. These are all important caveats to understand before using this dataset, as assuming similarity between game stores (either offline or online for each different platforms) is quite possibly erroneous.

Motivation for selected data source (5%) -> why did we choose steam and not another website?

***1.2*** *Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?*

The dataset is created by five students (team 8) of Tilburg University. It is created for the Online Data Collection and Management course of the master Marketing Analytics. The instructor for this course is Hannes Datta.

***1.3*** *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

No company nor person funded the creation of this dataset. Therefore, this team is be completely independent in creating the dataset and performing statistical analysis.

# 2. Composition

***2.1*** *What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

Each instance in the dataset represents a software product from the Steam website. Steam has the following software products: games, movies, music, tools, downloadable content (DLC) and demos. The different instances do not have a direct interaction nor correlation with each other. However, each instance of downloadable content (DLC) and demos is always related to a base game.

***2.2*** *How many instances are there in total (of each type, if appropriate)?*

The dataset consists of around 138.000 instances on March 10 2022. New products are often added on the Steam website, so it is most likely that there will be more instances by now.

|  |  |  |  |
| --- | --- | --- | --- |
| **Product type** | **Instances** | **Average price** | **Freemium** |
| Games | Count (number / percentage) | Average | Count |
| Movies |  |  |  |
| Music |  |  |  |
| Tools |  |  |  |
| DLC |  |  |  |
| Demo |  |  |  |
| **Total** |  |  |  |

***2.3*** *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

The dataset contains all the instances that when retrieved give a success is true value and therefore provide a “data” object. Failure results in no “data” object. Which is possibly caused by a piece of software being unavailable, either due to it has not been released/public yet or being deleted.

***2.4*** *What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a description.*

The data consists of unprocessed text. Every app\_id contains of several text, like name of the game, required age to play the game. Also, a description is given about the game and the requirements in terms of RAM to play the game. Even the price and possible discount is retrieved.

Table 1 dataframe\_raw

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Type of data** |
| type | The type of instance, like game, DLC, movies, music, tools and demos. | String |
| name | Name of the piece of software. | String |
| steam\_appid | A unique number that is associated with each instance. | Numeric |
| required\_age | The required age to use the instance | Numeric |
| is\_free | Indication if the product is free or not. | Bolean |
| detailed\_description | Description as displayed in the Steam store page. | String |
| about\_the\_game | Information about the game. | String |
| short\_description | Shorter description about the software. | String |
| supported\_languages | Languages supported by the software. | String |
| header\_image | Link with the image of the software in JPG. | String |
| website | Link with the website of the developers of the sotware. | String |
| mac\_requirements | Shows if the piece of software is compatible for a Mac system. It also shows which requirements are needed. | empty? |
| linux\_requirements | Shows if the piece of software/game is compatible for a Linux system. It also shows which requirements are needed. | empty? |
| developers | The developers of the software. | String |
| publishers | The name of the publisher business. It is possible that the developer and publisher are the same studio. | String |
| packages | Some games have packages. The variable “packages” shows the number of the package. | Number |
| background | This contains a link with the background image in JPG. | String |
| background\_raw | This contains a link with the background image in JPG. | String |
| fullgame.appid | If the product type is a DLC or demo, it will show the app\_id of the corresponding full game. | Numeric |
| fullgame.name | If the product type is a DLC or demo, it will show the name of the corresponding full game. | String |
| pc\_requirements.minimum | The minimum pc requirements the software need. | String |
| pc\_requirements.recommended | The recommended pc requirements for the software. | String |
| price\_overview.currency | The currency of the price of the software. | String |
| price\_overview.initial | Initial price of the software. | Numeric |
| price\_overview.final | Final price of the software. | Numeric |
| price\_overview.discount\_percent | Discount on the software in percentage. | Numeric |
| price\_overview.initial\_formatted | The initial price in a certain format, for example 3,29€. | String |
| price\_overview.final\_formatted | The final price in a certain format. | String |
| platforms.windows | True means that the software is available for windows, false means it isn’t. | Bolean |
| platforms.mac | True means that the software is available for mac, false means it isn’t. | Bolean |
| platforms.linux | True means that the software is available for linux, false means it isn’t. | Bolean |
| release\_date.coming\_soon | True means that the software isn’t released yet, false means it is released. | Bolean |
| release\_date.date | The release date of a piece of software, as displayed on Steam. | String |
| support\_info.url | This contains the URL, when support is needed for the software. | String |
| support\_info.email | This contains the email, when support is needed | String |
| content\_descriptors.ids | The id number of content descriptors. | Numeric |
| content\_descriptors.notes | Notes on the content of the software, for example violence. | String |
| collection\_details.created\_by | Details of who collected the data. | String |
| collection\_details.created\_at | When the data was collected. | Date? |
| legal\_notice | Information about copyright. | String |
| pc\_requirements | Pc requirements for the software. | Empty? |
| achievements.total | The total achievement of a software. | Numeric |
| mac\_requirements.minimum | Minimum mac requirements for the software. | String |
| mac\_requirements.recommended | Recommended mac requirements for the software. | String |
| linux\_requirements.minimum | Minimum Linux requirements for the software. | String |
| linux\_requirements.recommended | Recommended Linux requirements for the software. | String |
| Demos? |  |  |
| controller\_support |  | String |
| recommendations.total | Number of people who recommended the software. | Numeric |

Table 2 dataframe\_package\_groups

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Type of data** |
| name | Value that takes the form “default”. | String |
| title | Title of the package. | String |
| description |  |  |
| selection\_text | The following text: “Select a purchase option”. | Unprocessed text |
| save\_text |  |  |
| display\_type | Zero means that the software is not included in a publicly available package.  One means that it is. | Numeric |
| is\_recurring\_subscription | False means that the software is an one time purchase. True means it is a subscription. | Bolean |
| subs | Information about the package, like package id. | NESTED |
| steam\_appid | The app\_id to which the information above refers to. | Numeric |

Table 3 dataframe\_categories

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Type of data** |
| id | ID number of the category | Numeric |
| description | The category of the product, like single player, partial controller support etc. | String |
| steam\_appid | The app\_id to which the information above refers to. | Numeric |

Table 4 dataframe\_genres

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Type of data** |
| id | The id number of a genre of the software. | Numeric |
| description | The genre of the software, like action, adventure, indie, etc. | String |
| steam\_appid | The app\_id to which the information above refers to. | Numeric |

Table 5 dataframe\_screenshots

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Type of data** |
| id | The id refers to the order of the images displayed from the store for a certain software. | Numeric |
| path\_thumbnail | The link to the image of the software in JPG. | String |
| path\_full | The full size image of the thumbnail on the same row. | String |
| steam\_appid | The app\_id to which the information above refers to. | Numeric |

Table 6 dataframe\_movies

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Type of data** |
| id | The id refers to the order of the videos displayed from the store for a certain software. | Numeric |
| name | The name of the video. | String |
| thumbnail | The thumbnail of the video. | String |
| highlight | Highlight represents whether the video is shown in the software store page. True means that the video is shown. False means that the video isn’t shown. | Bolean |
| webm.480 | Link to the video in webm.480. | String |
| webm.max | Link to the video in webm.max. | String |
| mp4.480 | Link to the video in mp4.480. | String |
| mp4.max | Link to the video in mp4.max. | String |
| steam\_appid | The app\_id to which the information above refers to. | Numeric |

Table 7 dataframe\_achievements.highlighted

|  |  |  |
| --- | --- | --- |
| **Variable** | **Description** | **Type of data** |
| total | The number of total achievements of a software. | Numeric |
| highlighted | The name of the Steam achievements and a link to an image of the achievement. | NESTED |
| steam\_appid | The app\_id to which the information above refers to. | Numeric |

***2.5*** *Is there a label or target associated with each instance? If so, please provide a description.*

The app\_id is the unique identifier that is associated with each instance of software.

***2.6*** *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

If any information is missing from individual instances, it is because the information was unavailable.

***2.7*** *Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

Instances that are DLCs or demos are dependent on a base game. These DLCs or demos are identifiable as they have unique variables named “fullgames.appid” and “fullgame.name”.

***2.8*** *Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

The raw dataset has not been split to give the opportunity for the open domain to enhance their analyses. One recommended split would be on the variable “type” to focus analyses on specific categories of software within the dataset i.e. game.

***2.9*** *Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

The dataset is not self-contained. The dataset rely on external resources; the URL-link of the Steam site.

In addition, the Metacritic score available for some software is solely dependent on Metacritic.com. This is a review aggregate website that summarises the overall average player score for certain software -> not mentioned in the dataframes.

# The data collected is to a large extent self contained within the Steam API. The dataset contains thumbnails that are hosted on a external cloud storage (Akamai). These are fairly constant trough time unless the owner of the software changes the thumbnails or deletes the software. Furthermore there are no licenses fees associated collecting this data from the API. Since their is no archive within the API data collected is a snapshot.

***2.10*** *Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.*

The dataset does not contain data that might be considered confidential. The data in the dataset is also publicly available from their website. It is possible to use the Steam APIs to collect data about the users. This project does not utilize user data.

***2.11*** *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

In our knowledge no content is offensive, insulting, threatening, or otherwise cause anxiety, if viewed directly. However, however steam does include age restriction (‘required\_age’) for some of its software.

***2.12*** *Does the dataset relate to people? If not, you may skip the remaining questions in this section.*

The dataset does not relate to people. Therefore, the questions 2.12 till 2.15 are not relevant.

***2.13*** *Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

-

***2.14*** *Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.*

-

***2.15*** *Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

-

# 3. Collection Process

***3.1*** *How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

The data associated with each instance acquired, was directly observable, such as raw text and the rating of a product.

***3.2*** *What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?*

For the data collection, the Steam API is used. This was validated by the fact that team 8 got an official API key.

More explanation

***3.3*** *If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?*

The dataset was not a sample of a larger set.

***3.4*** *Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?*

The students of the master Marketing Analytics at Tilburg University were involved in the data collection process. There was no monetary compensation involved. Their work was compensated by acquiring knowledge about API’s and how to work with Python.

***3.5*** *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

Because an API is used, the data is real time collected and updated every time someone run the code.

***3.6*** *Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

There were no ethical review processes conducted.

***3.7*** *Does the dataset relate to people? If not, you may skip the remaining questions in this section.*

The dataset does not relate to people. Therefore, the questions 3.8 till 3.12 are not relevant.

***3.8*** *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?*

-

***3.9*** *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or other- wise reproduce, the exact language of the notification itself.*

-

***3.10*** *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

-

***3.11*** *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

-

***3.12*** *Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

-

# 4. Preprocessing, cleaning, labeling

***4.1*** *Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.*

Preprocessing: For the preprocessing part, we first collected all the app\_ids. After that we got the JSON file with all the data.

Cleaning:

we do clean, because the final CSV file will only contain the important columns (variables).

Labeling:

it might be labeling when we change the column names from the JSON file (e.g. removing the underscores).

***4.2*** *Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the “raw” data.*

The raw data is saved in a JSON file. PROVIDE A LINK.

***4.3*** *Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.*

Other than Python and the Steam API, no other software is used. PROVIDE A LINK.

# 5. Uses

***5.1*** *Has the dataset been used for any tasks already? If so, please provide a description.*

The dataset has not been used for any tasks yet.

***5.2*** *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

There is no repository available yet for papers and systems that use the dataset scraped in this project.

***5.3*** *What (other) tasks could the dataset be used for?*

The dataset could be used for several tasks. From a managerial perspective, the dataset can help provide a detailed general landscape of Steam as a store (i.e. general pricing outlooks, genre statistics, DLC availability, etc.). It can help prospective entrants facilitate planning – for instance, how would a game released without a publisher succeed with reviews as reference, what is the general pricing trend in that case? Apart from managerial use, there are academic benefits as well. Researchers can modify and couple this dataset with others that have parameters they would like to analyze. Studies about how the amount of articles on a video game interact with the number or content of reviews, or how the genre of a game may showcase differences in reviewing behaviors may now become a possibility as the daunting task of acquiring data is reduced significantly through the availability of this dataset.

Another interesting benefit is the potential improvement of already existing estimation formulas and tools or the creation of new ones. For example, the Boxleiter method allows the estimation of video game units sold (and in turn video game revenue) by using the number of reviews a game has on Steam and multiplying it by a number that differs based on release year, genre and others (Carless, 2020). This alone is invaluable in an industry where sales figures are hard to acquire for analyses. Nonetheless, making more data available publicly can only serve to improve and correct such methods and our understanding of the phenomena around them. Suppose that more data availability could even lead to accurate estimates of marketing spending on individual games, downloadable content and remastered games – this is the ideal towards which we hope to contribute.

***5.4*** *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

There is nothing about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses.

***5.5*** *Are there tasks for which the dataset should not be used? If so, please provide a description.*

The dataset should not be used for commercial ends, because it is prohibited by the company Steam.

1. \* https://arxiv.org/abs/1803.09010 [↑](#footnote-ref-1)