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 consequence (Schreier, 2018).

As a result of the scarce availability of video game industry data, available research is consequently influenced. Research observing how video game 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 video game dataset through the use of Steam API. Steam API enables the collection of various datapoints pertaining to users and video games available its online retail store. This project specifically will focus on the collection of video game datapoints in an attempt to create a dataset that includes a comprehensive list of games and related data.

It is important to note several limitations in using the 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 utilise 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.

The team finds several benefits in making such a dataset available. From a managerial perspective, this 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 benefit as well. Researchers can modify and couple this dataset with others that have parameters they would like to analyse. 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 behaviours 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.

<https://variety.com/2019/gaming/news/video-games-300-billion-industry-2025-report-1203202672/>

<https://www.trtworld.com/life/video-game-industry-winners-in-the-pandemic-44119>

<https://newsroom.accenture.com/news/global-gaming-industry-value-now-exceeds-300-billion-new-accenture-report-finds.htm>

<https://www.gamesindustry.biz/articles/2017-02-15-why-is-the-games-industry-still-struggling-with-digital-data>

<https://kotaku.com/some-reasons-why-the-games-industry-is-so-secretive-an-1825929139>

https://newsletter.gamediscover.co/p/how-that-game-sold-on-steam-using?utm\_source=url