VIDEO GAMES

FINAL PROJECT

Creating a snapshot of video game sales data from online sources

Arjun Remeshkumar Nair

Table of Contents

Executive Summary

As of 2021, video game sales worldwide had garnered almost US$191 billion annually, with a general trend indicating significant growth in the future (Wikipedia, 2023). In such an industry, data-driven decision-making is a necessity for market dominance. To cater to this need, this project aims to provide comprehensive information regarding the games and their sales, to both publishers and the general public. We utilize the VGChartz website to extract all available information and store it in an accessible and efficient structure. The final stored dataset comprises games ranked by their respective global and region-specific sales, along with the salient attributes such as the console, publisher or genre. After several steps of data wrangling and creating indexes on specific variables, we store the final dataset in a local Mongodb Database for future search and aggregation. Companies could build on top of this data to create a sales prediction model, conduct market and competitor research , and optimize their investment to balance the returns and reach a higher ROI. Additionally, decisions around new product introduction and development can be optimized using this data. For the general audience who are passionate about video games, they could refer to this script and download the dataset for personal exploration.

Project Background

<u>Introduction</u>

"Tetris, Mario Bros, Final Fantasy, SimCity, Overwatch, Pokemon Go, Fruit Ninja, NBA 2K, League of Legends, PUBG…" Among this list of games, no matter what age you are, what language you speak, you must be familiar with at least one of these. Video games have been around all of us for decades, and it definitely becomes part of our daily entertainment.

According to Investopedia, there are over 26 million gamers worldwide, which makes up 26% of the global population (Beattie, 2022). Compared with the movie and music industries, the video game industry is even larger than the combination of those two. With the growing number of gamers, the market value was predicted to be worth over $268 Billion by 2025 (Kirkcaldy, 2023). Despite this, the industry does not receive enough attention like the movie and music industries do.

In this project, we will dive into the video game industry. Combined with web-scraping skills, we will build a dataset, which can be available for everyone who is interested in exploring the history of video games, sales growth over the years, and consumer preferences. Our goal is to create a structured and well-defined dataset that not only provides publishers with qualified resources for better understanding the market and creating sales predictor models, but also serves as a resource for the game lovers to play and explore.

Gaming Equipment and Technology

In order to better understand our dataset, we will briefly introduce the *"Platform"* variable that will be included in our dataset. Platforms have five major segments, which can help analysts to look into a niche market of video games.

Console Gaming: This might be the most well-known one in the current digital gaming world. By definition, "a console game is a type of interactive multimedia software that uses a video game console to provide an interactive multimedia experience via a television or other display device. (Techopedia, 2018)" The game console usually will have handheld controllers. Xbox, Playstation series, Nintendo and Sega are the most popular consoles recently.

PC Gaming: It is also known as personal computer gaming, which defines the games played on computers. Quoted from the Built-in Website, PCs massively rise in gameplay with over 1.8 billion people choosing computers as a game tool by the end of 2022 (Daley, 2022).

Cloud Gaming: Cloud gaming is a relatively new type of video games. Without downloading and discs, people can play with it online anytime and anywhere they want, like on mobile phones, iPad, or console.

VR/AR Gaming: Virtual reality and augmented reality games bring players emerging into the virtual gaming world, so they could have better gaming experiences. Many big technology companies, like Meta, heavily invested in this area and started to launch the VR headset to move into the metaverse.

Mobile Gaming: Mobile gaming definitely has the largest group of players since people can easily play games on their phones without purchasing any other equipment. In 2027, mobile gaming players may reach over half of the worldwide population (Daley, 2022).

Game Genre

Game genres may be defined differently across different platforms. Based on the website we scraped, it separates the game genres into 12 types, and each genre has several examples.

1. Action (Example: God of War, Elden Ring, Spider Man)
2. Adventure (Example: The Last of Us, Star Wars Jedi)
3. Fighting (Example: Pokemon, Street Fighter, Dragon Ball Fighter)
4. Misc (Example: Just Dance, Guitar Hero series)
5. Platform (Example: Super Mario, Super Meat Boy, Hollow Knight)
6. Puzzle (Example: Braid, The Talos Principle, The Witness)
7. Racing (Example: Forza Horizon series, Wreckfest)

8. Role-playing (Example: Final Fantasy, Cyberpunk 2077, Persona 5 Royal)
9. Shooter (Example: Call of Duty series, ZED)
10. Simulation (Example: The Sims series, Animal Crossing)
11. Sports (Example: Nintendo Switch Sports, FIFA International Soccer)
12. Strategy (Example: Crusader Kings, Civilization)

Data Source and Web-Scraping Routine

Data Source

VGChartz (Video Game Chats) is a business intelligence and research firm (VGChartz, LinkedIn). Also being a video game industry research firm, they integrate sales data in both hardware and software, the latest news, and hot topics of games. Also, it's a social community that allows game lovers to leave reviews and exchange ideas.

For this project, we look at VGChartz's GameDB section *(Appendix 1)*. It is a game search page where people can search for any specific game that they are interested in, and also has lots of filters that can explore with it *(Appendix 2)*. Based on everyone's personal setting, the Result table will display all games and related information that you are looking for. Our team checked all filter boxes to include a comprehensive list of variables *(Appendix 3)*.

Web-Scraping Routine

Step1: Check if this website is scrapable.

To access the required data, we start by targeting the following page : VGChartz Game DB. We tried extracting the initial page to ensure that this website was able to retrieve all required information without any restrictions.

Step 2: Check if this website needs cookies.

Additionally, we found that no cookies were used in the script, since no session or user login was required to traverse through the pages.

Step 3: GET request to access all pages (Appendix 4)

The webpage has specific filters mentioned in the URL address. These filters are to ensure that we do not subset on any features and ensure that we retrieve all available features from the page. These filters include values for the 'Game', ' Developer', 'Publisher', 'Total Sales' etc. We use a GET request to traverse through all available pages. The last page number is extracted from the webpage and the script is looped across the range(1, last_page). This ensures that the code can be utilized even with the addition of new pages, without modification to the code.

Step 4: Download page content and write to file for data extraction (Appendix 5)

For each execution of the loop, the HTML content was downloaded and written to file. In case there are any changes, we could use the downloaded links to ensure the consistency. All these files are further read into a BeautifulSoup object and data extraction is performed.

Step 5: Find the body content of the data source (Appendix 6)

The element *'generalBody'* contains the table of information that we will extract. The relevant rows of the table are identified using the presence of game images as two different values.

Step 6: Data extraction and data design

The elements corresponding to each of these games are stored in a list and traversed through in a subsequent loop:

1. For each element, we observe that the first *'td'* element contains the rank of the game.
2. The *'td'* element with a predefined value of style is used to identify the Game Name and URL.

3. The subsequent *'td'* elements in the page provide the remaining details such as: Console, Publisher, Developer, VGChartz Score, Critic Score, User Score, Total Shipped, Total Sales, NA Sales, PAL Sales, JP Sales, Others Sales, Release Date and Last Updated Date.
    a. The URL of each game is opened and the webpage content is parsed using BeautifulSoup. We identify and extract any different releases of a game and store it as Releases, using the element *'li'* within *'gameBody'*. 'Releases' will be a list.
    b. The opened URL also contains the Genre of the game within the element *'p'* within *'h2'* within *'gameGenInfobox'*.

Step 7: Data storage and data process (Appendix 7)

All of the above data points are inserted into the Mongodb collection 'Overview' within the database 'Game_Sales'. For games without additional Releases, "NA" was inserted. A status update for data insertion is printed to the terminal for each page executed. Indices are created on 'Rank', 'Game Name', 'Genre', 'Console' and 'Publisher' fields.

Data Design Choices and Recommendations

Data Design Choices

Game_Sales.Overview			
Field	**Data Type**	**Unit**	**Description**
Rank	String	-	Rank of game based on overall sales
Name	String	-	Name of the game
URL	String	-	URL leading to the website of the game
Console	String	-	Console the game was released for

Genre	String	-	Genre of the game
Publisher	String	-	Publisher of the game
Developer	String	-	Developer of the game
Releases	String	-	Different game releases
VG_score	String	-	VGChartz score for the game
Critic_score	String	-	Critic score for the game
User_score	String	-	User score for the game
Total_shipped	String	Millions	Total shipped units of the game
Total_sales	String	Millions	Total sales of the game (US Dollars)
NA_sales	String	Millions	Sales in North America(US Dollars)
PAL_sales	String	Millions	Sales in Europe(US Dollars)
JP_sales	String	Millions	Sales in Japan(US Dollars)
Other_sales	String	Millions	Sales in other countries(US Dollars)
Release_date	String	-	Release date of the game
Last_update	String	-	Last update of game on VGChartz

We choose to store the web-scraped data in Mongodb due to the nature of lists existing as fields in the dataset. Mongodb is also a faster and more scalable alternative. The hierarchical structure supported by Mongodb provides an additional advantage.

The final dataset, named ‘Game_Sales’ within the collection ‘Overview’, consists of data extracted from 1,252 pages, with 50 results in each of the pages except the last, bringing the total number of games to 62,565. Among 62,565 documents, each has 19 features *(Appendix 8)*. We have not excluded any variables in our current extract since exclusion will not help us achieve our goal of providing data regarding all the features available on the webpage.

To give more details, the ‘Releases’ feature is made up of the different releases of the game, in the form of documents within a document *(Appendix 9)*. We have created indices on ‘Rank’, ‘Game Name’, ‘Console’, ‘Developer’ and ‘Publisher’, since these are the common fields that would be used for aggregation (*Appendix 10*).

As per business requirements and additional data availability, new collections can be created within our database to cater to specific groups of stakeholders or include additional features. With increase in data, we can implement the MapReduce function if required or create indices on more fields to improve query performance.

Recommendations for the data source:

1. Identifying and obtaining additional information relating to game specific attributes and sales of the game through various channels and across editions.
2. Obtaining more quality information regarding the reviews and ratings of games by critics and users alike across online social communities.
3. Improving the general quantity and quality of information.

Business Impact

Video games are closely linked to technology innovations, and as technology keeps improving, video games are upgrading at the same time. The pandemic in 2019 resulted in a surge of increasing numbers of game players, which changed part of people’s social and entertainment behaviors. Technology companies such as Google, Meta, Apple and Netflix, have all joined in this big trend and are competing for revenue streams (Beattie, 2022). Based on our final dataset scraped online, we can deeply research and generate data-driven strategies in three major areas.

Consumer Research

Conducting consumer research is important for companies to understand consumers' preferences. The more we get to know the target audience, the more precise strategies that we can apply to upgrade and promote video games.

Our final dataset includes columns of genre and release year that can reveal the trend of game genre change over decades, and how many games in a specific genre were released each year combined with the total shipped numbers. Also, instead of units, we can replace it with the total sales in dollars to analyze the historical popularity trend in games, which can represent customers' taste and preferences and help companies develop new games.

Industry Research

In the video game industry, fierce competition has always existed, with even large technology companies joining in and striving to compete on innovation and market share. At this point, our dataset could be utilized for industry and competitor analysis.

Within each game, we can obtain information about publishers, developers and consoles. Utilizing these data, we would know the game popularity (represented by sales or shipped units) among all aspects. These results can reveal competitive advantages in representative games for a developer or a publisher that customers perceive as superior value. Therefore, companies can change their strategy or look for suitable collaborators based on these data-driven insights.

Sales Prediction

Sales forecast is crucial for a company to effectively allocate its resources and come up with a cost-benefit analysis. It can be a decision-making tool to shape and determine the project path.

Based on the game attributes, we can predict the game sales before releasing. With the sales data in different geographic regions, forecasting can be divided into target markets as well.

Conclusion

This project aims to provide a structured and comprehensive dataset of video game sales and attributes, extracted from the VGChartz website, to assist companies in making data-driven decisions for market dominance and for game lovers to explore the industry's history, sales growth, and consumer preferences.

The whole web scraping process was done in Python with the BeautifulSoup library. GET requests allow us to access all web pages with different filter settings. By using the web page development tool to inspect elements and study the HTML structure, we successfully extracted all information and processed it by requiring various methods, like regular expressions. For user convenience on data searching, the final dataset is stored in MongoDB with indexes on specific variables .In conclusion, this project provides a valuable resource for both the industry and gaming enthusiasts to better understand the video game market.

Reference

Wikimedia Foundation. (2023, March 17). *Video game industry*. Wikipedia. Retrieved March 19, 2023, from https://en.wikipedia.org/wiki/Video_game_industry

What is a console game? - definition from Techopedia. Techopedia.com. (n.d.). Retrieved March 13, 2023, from https://www.techopedia.com/definition/756/console-game

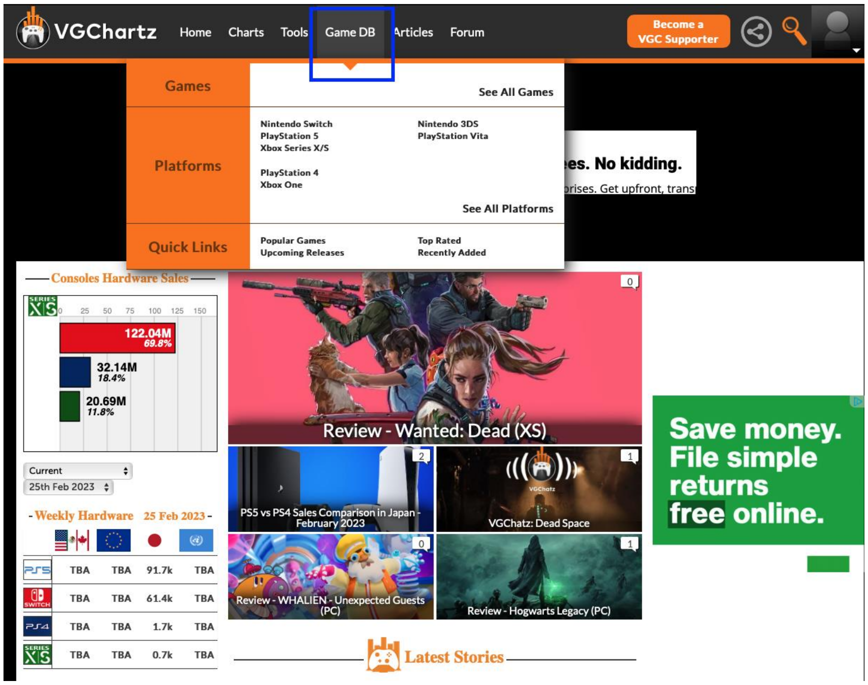
Beattie, A. (2022, December 19). *How the video game industry is changing*. Investopedia. Retrieved March 13, 2023, from https://www.investopedia.com/articles/investing/053115/how-video-game-industry-changing.asp

Daley, S. (2022, September 29). *Gaming*. BuiltIn. Retrieved March 13, 2023, from https://builtin.com/gaming

Kirkcaldy, A. (2023, January 12). *Video game industry statistics, trends and data in 2023*. WePC. Retrieved March 13, 2023, from https://www.wepc.com/news/video-game-statistics/

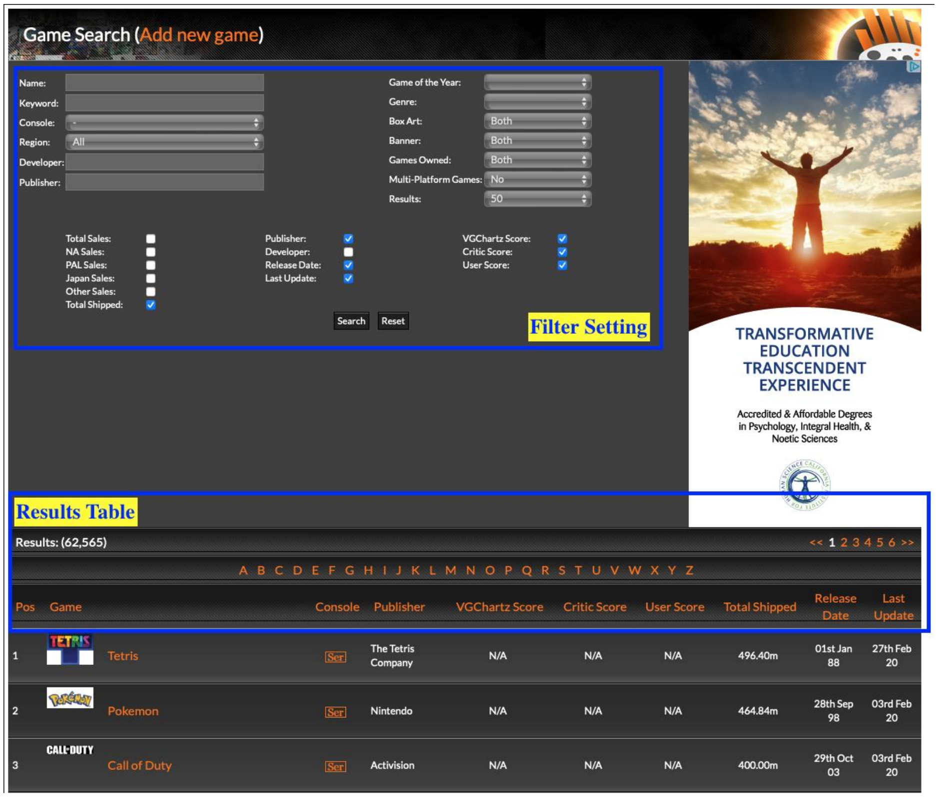
VGChartz. *LinkedIn Overview Page.* Retrieved March 14, 2023, from https://www.linkedin.com/company/vgchartz-ltd/about/

Appendix



PS5	TBA	TBA	91.7k	TBA
SWITCH	TBA	TBA	61.4k	TBA
PS4	TBA	TBA	1.7k	TBA
SERIES X|S	TBA	TBA	0.7k	TBA

Appendix 1


Game Search (Add new game)
Filter Setting
TRANSFORMATIVE EDUCATION TRANSCENDENT EXPERIENCE
Accredited & Affordable Degrees in Psychology, Integral Health, & Noetic Sciences
Results Table

Name:
Keyword:
Console: -
Region: All
Developer:
Publisher:

Game of the Year:
Genre:
Box Art: Both
Banner: Both
Games Owned: Both
Multi-Platform Games: No
Results: 50

Total Sales: □
NA Sales: □
PAL Sales: □
Japan Sales: □
Other Sales: □
Total Shipped: ☑

Publisher: ☑
Developer: □
Release Date: ☑
Last Update: ☑

VGChartz Score: ☑
Critic Score: ☑
User Score: ☑

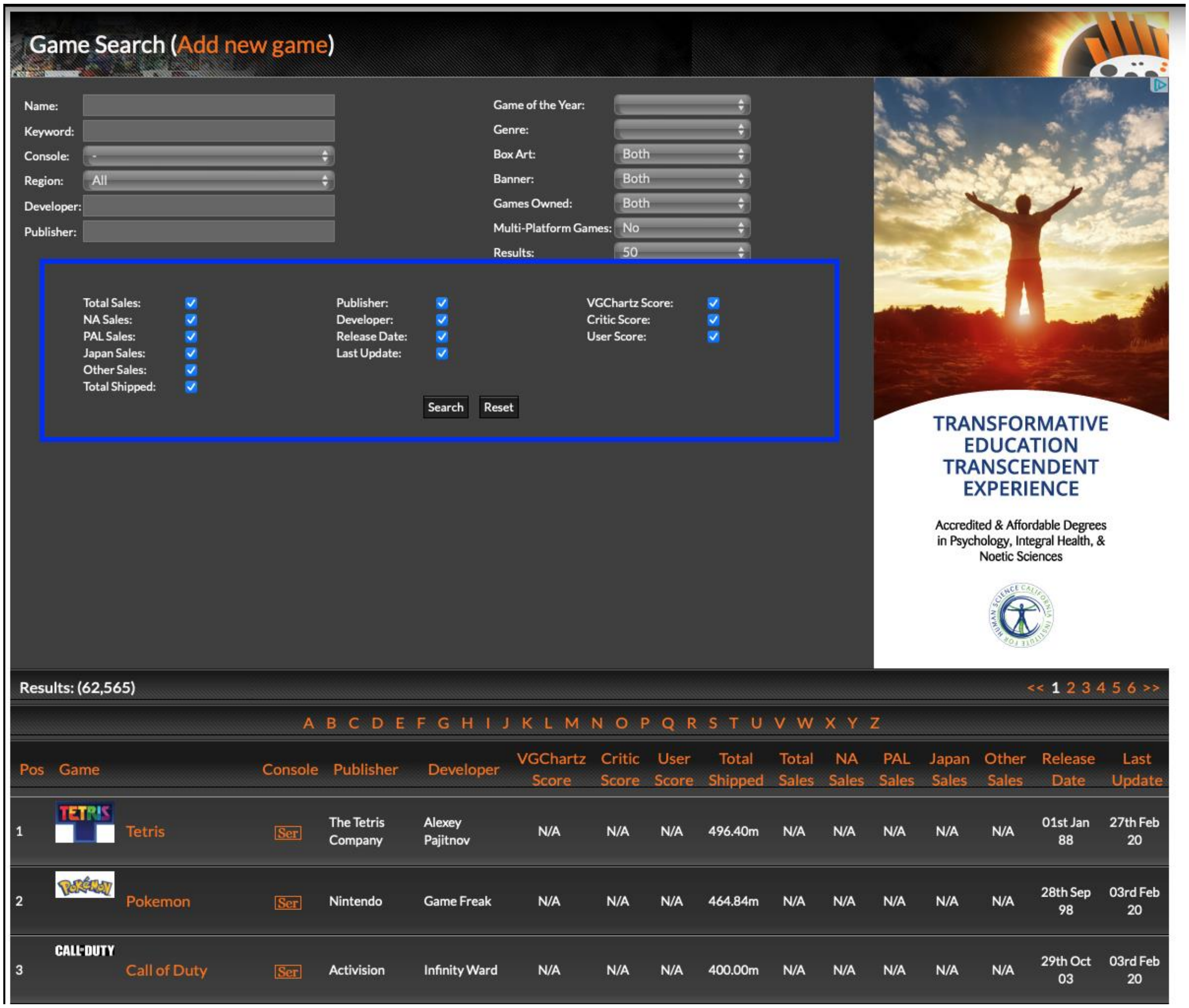
Search Reset

Results: (62,565) << 1 2 3 4 5 6 >>

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Pos	Game	Console	Publisher	VGChartz Score	Critic Score	User Score	Total Shipped	Release Date	Last Update
1	Tetris	Ser	The Tetris Company	N/A	N/A	N/A	496.40m	01st Jan 88	27th Feb 20
2	Pokemon	Ser	Nintendo	N/A	N/A	N/A	464.84m	28th Sep 98	03rd Feb 20
3	Call of Duty	Ser	Activision	N/A	N/A	N/A	400.00m	29th Oct 03	03rd Feb 20

Appendix 2



Game Search (Add new game)

Name:
Keyword:
Console: -
Region: All
Developer:
Publisher:

Game of the Year:
Genre:
Box Art: Both
Banner: Both
Games Owned: Both
Multi-Platform Games: No
Results: 50

Total Sales: ☑
NA Sales: ☑
PAL Sales: ☑
Japan Sales: ☑
Other Sales: ☑
Total Shipped: ☑

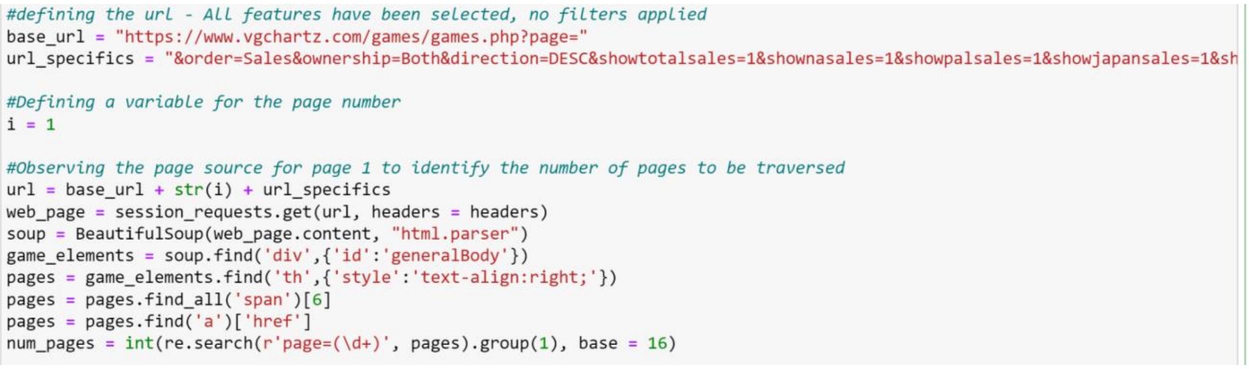
Publisher: ☑
Developer: ☑
Release Date: ☑
Last Update: ☑

VGChartz Score: ☑
Critic Score: ☑
User Score: ☑

Results: (62,565)

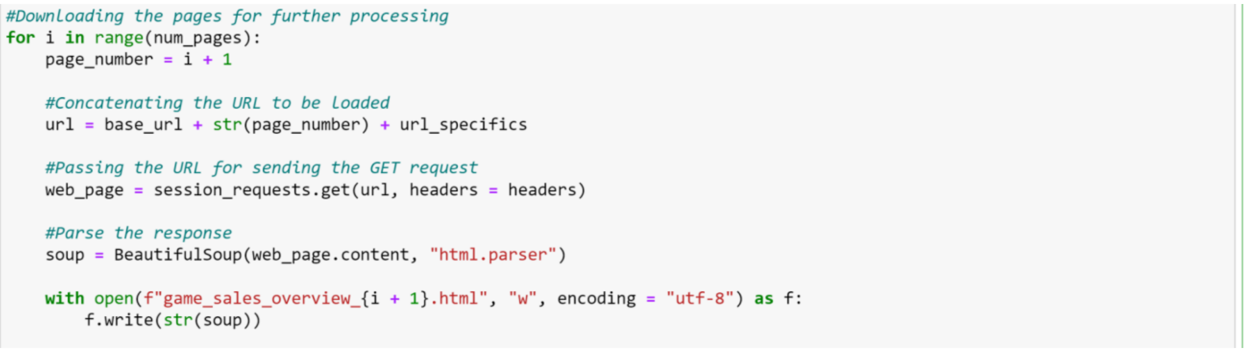
Pos	Game	Console	Publisher	Developer	VGChartz Score	Critic Score	User Score	Total Shipped	Total Sales	NA Sales	PAL Sales	Japan Sales	Other Sales	Release Date	Last Update
1	Tetris	Ser	The Tetris Company	Alexey Pajitnov	N/A	N/A	N/A	496.40m	N/A	N/A	N/A	N/A	N/A	01st Jan 88	27th Feb 20
2	Pokemon	Ser	Nintendo	Game Freak	N/A	N/A	N/A	464.84m	N/A	N/A	N/A	N/A	N/A	28th Sep 98	03rd Feb 20
3	Call of Duty	Ser	Activision	Infinity Ward	N/A	N/A	N/A	400.00m	N/A	N/A	N/A	N/A	N/A	29th Oct 03	03rd Feb 20

Appendix 3



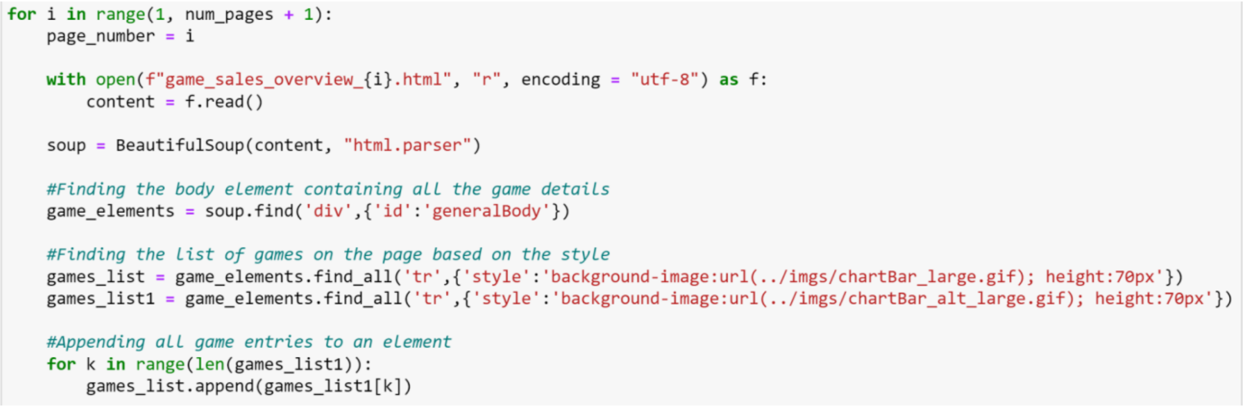
```
#defining the url - All features have been selected, no filters applied
base_url = "https://www.vgchartz.com/games/games.php?page="
url_specifics = "&order=Sales&ownership=Both&direction=DESC&showtotalsales=1&shownasales=1&showpalsales=1&showjapansales=1&sh

#Defining a variable for the page number
i = 1

#Observing the page source for page 1 to identify the number of pages to be traversed
url = base_url + str(i) + url_specifics
web_page = session_requests.get(url, headers = headers)
soup = BeautifulSoup(web_page.content, "html.parser")
game_elements = soup.find('div',{'id':'generalBody'})
pages = game_elements.find('th',{'style':'text-align:right;'})
pages = pages.find_all('span')[6]
pages = pages.find('a')['href']
num_pages = int(re.search(r'page=(\d+)', pages).group(1), base = 16)
```

Appendix 4



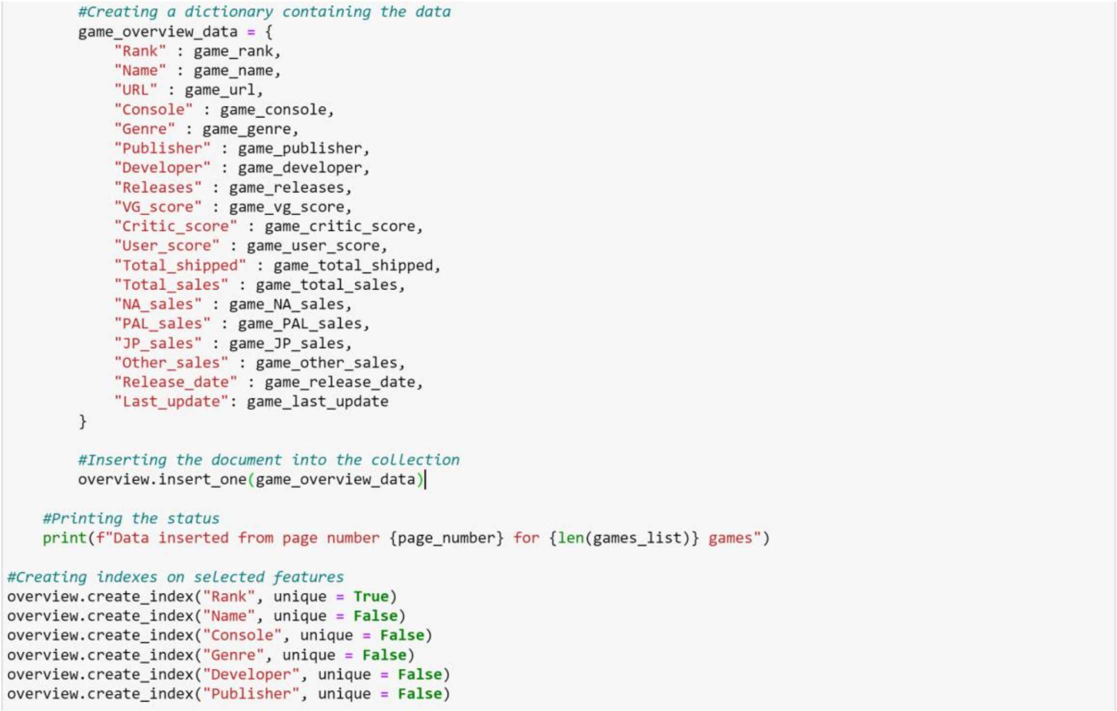
```
#Downloading the pages for further processing
for i in range(num_pages):
    page_number = i + 1

    #Concatenating the URL to be loaded
    url = base_url + str(page_number) + url_specifics

    #Passing the URL for sending the GET request
    web_page = session_requests.get(url, headers = headers)

    #Parse the response
    soup = BeautifulSoup(web_page.content, "html.parser")

    with open(f"game_sales_overview_{i + 1}.html", "w", encoding = "utf-8") as f:
        f.write(str(soup))
```

Appendix 5



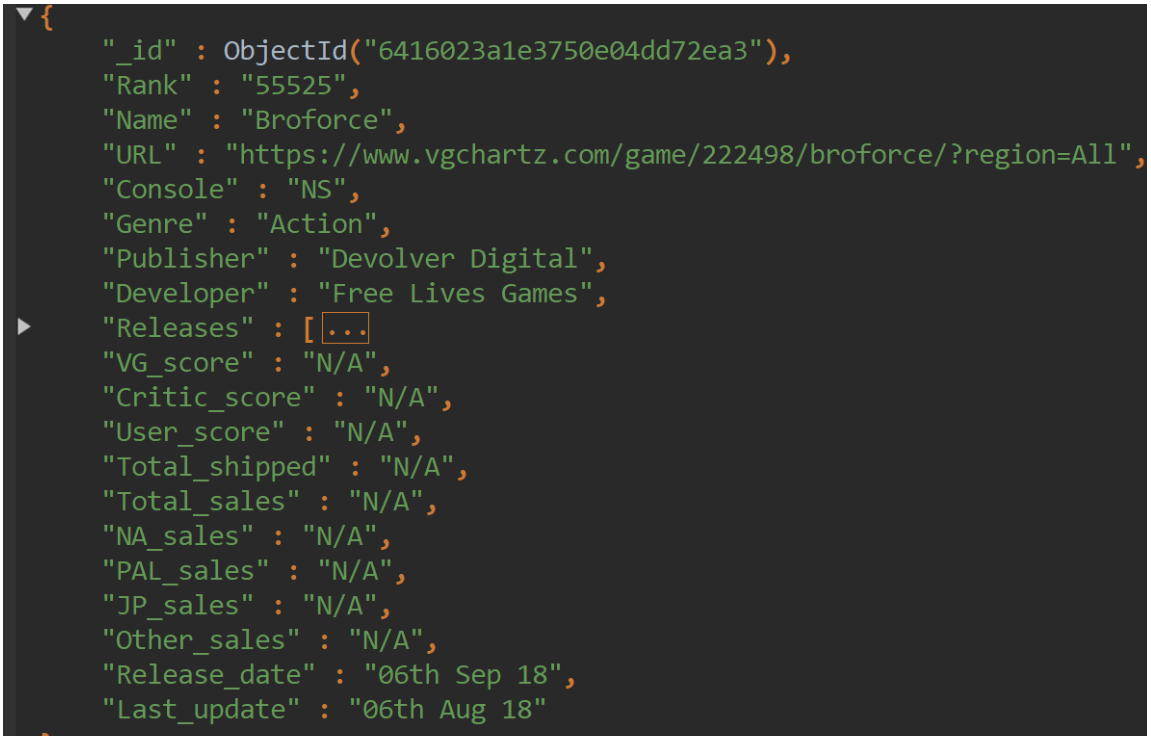
```
for i in range(1, num_pages + 1):
    page_number = i

    with open(f"game_sales_overview_{i}.html", "r", encoding = "utf-8") as f:
        content = f.read()

    soup = BeautifulSoup(content, "html.parser")

    #Finding the body element containing all the game details
    game_elements = soup.find('div',{'id':'generalBody'})

    #Finding the list of games on the page based on the style
    games_list = game_elements.find_all('tr',{'style':'background-image:url(../imgs/chartBar_large.gif); height:70px'})
    games_list1 = game_elements.find_all('tr',{'style':'background-image:url(../imgs/chartBar_alt_large.gif); height:70px'})

    #Appending all game entries to an element
    for k in range(len(games_list1)):
        games_list.append(games_list1[k])
```

Appendix 6



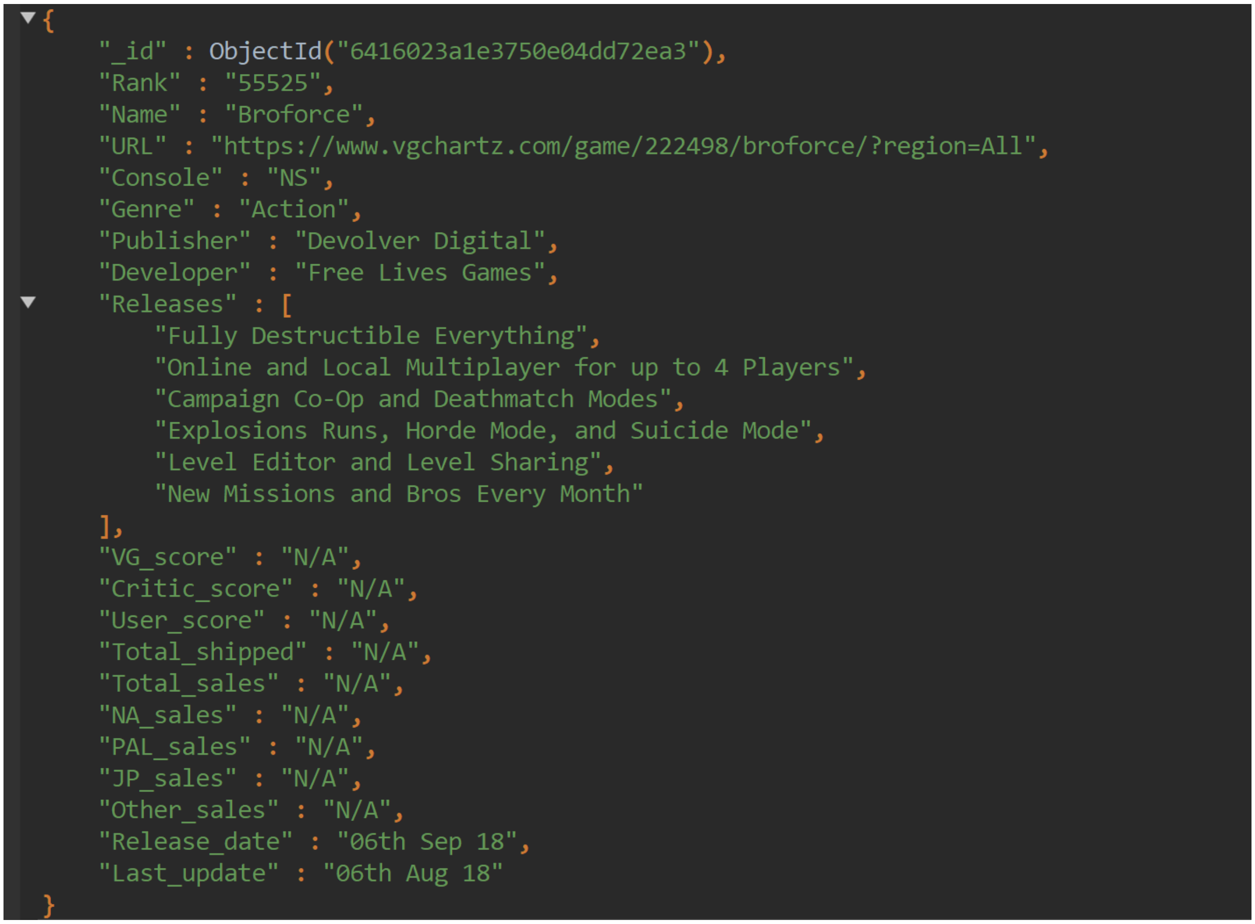
```
        #Creating a dictionary containing the data
        game_overview_data = {
            "Rank" : game_rank,
            "Name" : game_name,
            "URL" : game_url,
            "Console" : game_console,
            "Genre" : game_genre,
            "Publisher" : game_publisher,
            "Developer" : game_developer,
            "Releases" : game_releases,
            "VG_score" : game_vg_score,
            "Critic_score" : game_critic_score,
            "User_score" : game_user_score,
            "Total_shipped" : game_total_shipped,
            "Total_sales" : game_total_sales,
            "NA_sales" : game_NA_sales,
            "PAL_sales" : game_PAL_sales,
            "JP_sales" : game_JP_sales,
            "Other_sales" : game_other_sales,
            "Release_date" : game_release_date,
            "Last_update": game_last_update
        }

        #Inserting the document into the collection
        overview.insert_one(game_overview_data)

    #Printing the status
    print(f"Data inserted from page number {page_number} for {len(games_list)} games")

#Creating indexes on selected features
overview.create_index("Rank", unique = True)
overview.create_index("Name", unique = False)
overview.create_index("Console", unique = False)
overview.create_index("Genre", unique = False)
overview.create_index("Developer", unique = False)
overview.create_index("Publisher", unique = False)
```

Appendix 7



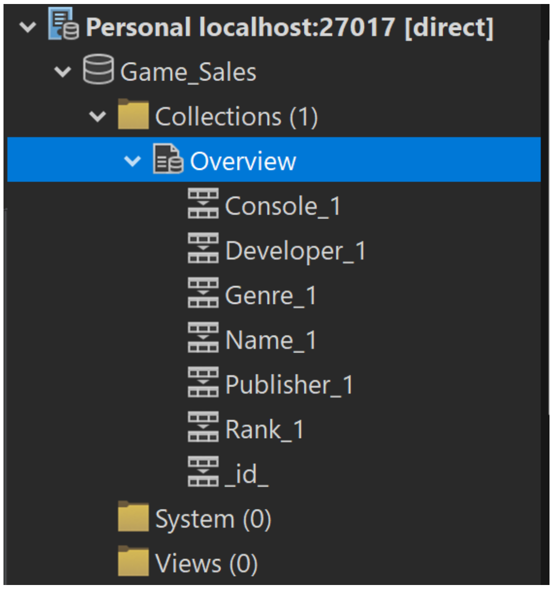
```
{
    "_id" : ObjectId("6416023a1e3750e04dd72ea3"),
    "Rank" : "55525",
    "Name" : "Broforce",
    "URL" : "https://www.vgchartz.com/game/222498/broforce/?region=All",
    "Console" : "NS",
    "Genre" : "Action",
    "Publisher" : "Devolver Digital",
    "Developer" : "Free Lives Games",
    "Releases" : [...
    "VG_score" : "N/A",
    "Critic_score" : "N/A",
    "User_score" : "N/A",
    "Total_shipped" : "N/A",
    "Total_sales" : "N/A",
    "NA_sales" : "N/A",
    "PAL_sales" : "N/A",
    "JP_sales" : "N/A",
    "Other_sales" : "N/A",
    "Release_date" : "06th Sep 18",
    "Last_update" : "06th Aug 18"
```

Appendix 8



```
{
    "_id" : ObjectId("6416023a1e3750e04dd72ea3"),
    "Rank" : "55525",
    "Name" : "Broforce",
    "URL" : "https://www.vgchartz.com/game/222498/broforce/?region=All",
    "Console" : "NS",
    "Genre" : "Action",
    "Publisher" : "Devolver Digital",
    "Developer" : "Free Lives Games",
    "Releases" : [
        "Fully Destructible Everything",
        "Online and Local Multiplayer for up to 4 Players",
        "Campaign Co-Op and Deathmatch Modes",
        "Explosions Runs, Horde Mode, and Suicide Mode",
        "Level Editor and Level Sharing",
        "New Missions and Bros Every Month"
    ],
    "VG_score" : "N/A",
    "Critic_score" : "N/A",
    "User_score" : "N/A",
    "Total_shipped" : "N/A",
    "Total_sales" : "N/A",
    "NA_sales" : "N/A",
    "PAL_sales" : "N/A",
    "JP_sales" : "N/A",
    "Other_sales" : "N/A",
    "Release_date" : "06th Sep 18",
    "Last_update" : "06th Aug 18"
}
```

Appendix 9



Appendix 10