Data Visualization

Assignment Report- By Ananya Das(23262592) and Jaslyn Samantha Dsouza (23262425)

Abstract:

Video games have been around since the 1980's and have only continued to increase in popularity. People would even consider playing video games as a hobby! The quality of games has drastically improved, with various new genres coming on the rise. One such example would be Indie games. We wanted to see the growth of the various genres in gaming over the years and identify if there is a particular trend that is followed. Since arcade games were more significant in the 1980's than they are now, we wanted to see if we would be able to observe such shifts in game design through our visualisation.

Through our visualisation, we were able to see the growth of various genres over the years. We were also able to identify limitations in the dataset, especially in terms of game collection after 2020. While the data was still not the most clear, even after all the cleaning, we were able to identify the most common genre of video games (which was action) and were also able to trace out the reductions in production of games in certain genres.

1. Datasets:

We decided to use 6 datasets containing video game information to include as many games as possible. These datasets were all retrieved from kaggle but were extracted from different sources.

1. (Link: https://www.kaggle.com/datasets/ibriiee/video-games-sales-dataset-2022-updated-extra-feat)
   Our first dataset is the Video Game Sales Dataset Updated - Extra Feat taken from kaggle (which was taken from data world and made available on kaggle) It contains 16,720 records that include Name, Platform, Year, Genre, Publisher, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Global_Sales, Critic_Score, Critic_Count, User_Score, User_Count, Developer, Rating from 1980 to 2020. The data was obtained by scraping https://www.vgchartz.com/.

2. (Link: https://www.kaggle.com/datasets/arnabchaki/popular-video-games-1980-2023)
   Our next dataset is the Popular Video Games 1980 - 2023 (also taken from kaggle). It contains 1512 records that include Title, Release Date, Team, Rating, Times Listed, Number of Reviews, Genres, Summary, Reviews, Plays, Backlogs, Wishlist from 1980 to 2023. These records were obtained by scraping backloggd.

3. (Link: https://www.kaggle.com/datasets/jummyegg/rawg-game-dataset)
   Our third dataset is the Video Game dataset, that contains records scrapped from Rawg. It contains 474,418 records that include Id, Slug, Name, Metacritic, Released, TBA, Updated, Website, Rating, Rating_top, Playtime, Achievements_count, Ratings_count, Suggestions_count, Game_series_count, Reviews_count, Platforms, Developers, Genres, Publishers, ESRB_Rating, Added_status_yet, Added_status_owned, Added_status_beaten, Added_status_toplay, Added_status_dropped, Added_status_playing from 1962 to 2033. The data include in this set is more wide spread compared to the others gathered, as it contains a lot of mobile games and casino based games.

4. (Link: https://www.kaggle.com/datasets/lorentzyeung/imdb-video-games-dataset)
   This dataset extracts its information from IMDB. It contains 14,683 records that include Index, Title, Genre, User Rating, Number of Votes, Runtime, Year, Summary, Director, Starts, Certificate.

5. (Link: https://www.kaggle.com/datasets/thedevastator/global-video-game-sales-ratings)
   Our fifth dataset uses video game reviews and sales ratings taken from Metacritic. It contains 6875 records that include Name, Year_of_Release, Genre, Publisher, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Gloabl_Sales, Critic_Score, Critic_Count, User_Score, User_Count, Developer, Rating, Story Focus, Gameplay Focus, Series from 1985 to 2016.

6. (Link: https://www.kaggle.com/datasets/ashaheedq/video-games-sales-2019?select=vgsales-12-4-2019.csv)
   Our final dataset is the video game sales 2019 dataset that was also scraped from vgchartz.com. It contains 55,792 records that include Name, basename, Genre, ESRB_Rating, Platform, Publisher, Developer, VGChartz_Score, Critic_Score, User_Score, Total_Shipped, Global_Sales, NA_Sales, PAL_Sales, JP_Sales, Other_Sales, Year, Last_Update, URL, Status, Vgchartzscore, img_url.

   While some of the datasets were small and easy to open, the third dataset did face tiny issues in being opened in excel due to its size. Overall, we were able to collect 570,000 records from all 6 datasets. Because they come from relatively different backgrounds and are extracted during different time periods, we look towards the **variety** of big data in the collection of our datasets. While most records have similar structures (all of them being based off video game data), they do have different types of columnar values in each. Also, the slight difficulty in opening the 400,000 record dataset would also indicate the same being present if all the records were to be compiled into one csv, which presents the concept of **volume** in big data, but on a smaller scale.

   We decided to work with Python for cleaning and pre-processing the datasets. Then, we used Tableau to make our final visualisation.

2. Data Exploration, Processing, Cleaning and/or Integration:

Since we considered 6 datasets, we thought it would be time efficient if we split up the cleaning process for each. We decided to clean each individual dataset and get it to a relatively standard format, combine them all together, and clean it some more so that we would be left with a new clean dataset we would be able to work with. We wanted to make sure the final dataset doesn't look like it comes from various sources by ensuring uniform standards within the data.

We knew that we wanted to mainly work with the year and genres associated with each game. So most of the datasets followed the same sort of cleaning process wherein we dropped all the unwanted columns, duplicate rows (with same name, year and genre), and filled out null values. Since the date/year of release and the genres of the games were important for our visualisation, we wanted to stray away from just dropping the null values. Instead we tried identifying several ways to fill them with values. At the end of the day, manually filling them was the most accurate

way of filling the data. However, that was quite time consuming. It was an achievable tasks on the smaller datasets, but a huge pain for the bigger ones.

Our 3rd dataset (the one with 400,000 + records) had a lot of missing release date and genre values. We initially went about manually filling them by passing their correct values in if conditions. That is how we fixed around 50 records. Then we tried using various API'S (like the Steam API and the Rawg API). Since the data was scraped using the Rawg API'S, it was better at giving out some of the missing values. But it still left most of it NULL. We realise the main issue with the dataset is that it seems to contain a lot of play store games and casino based games, for which information about releases is not very public. So we tried our best to retain as much as possible. We dropped whatever we couldn't find values for.

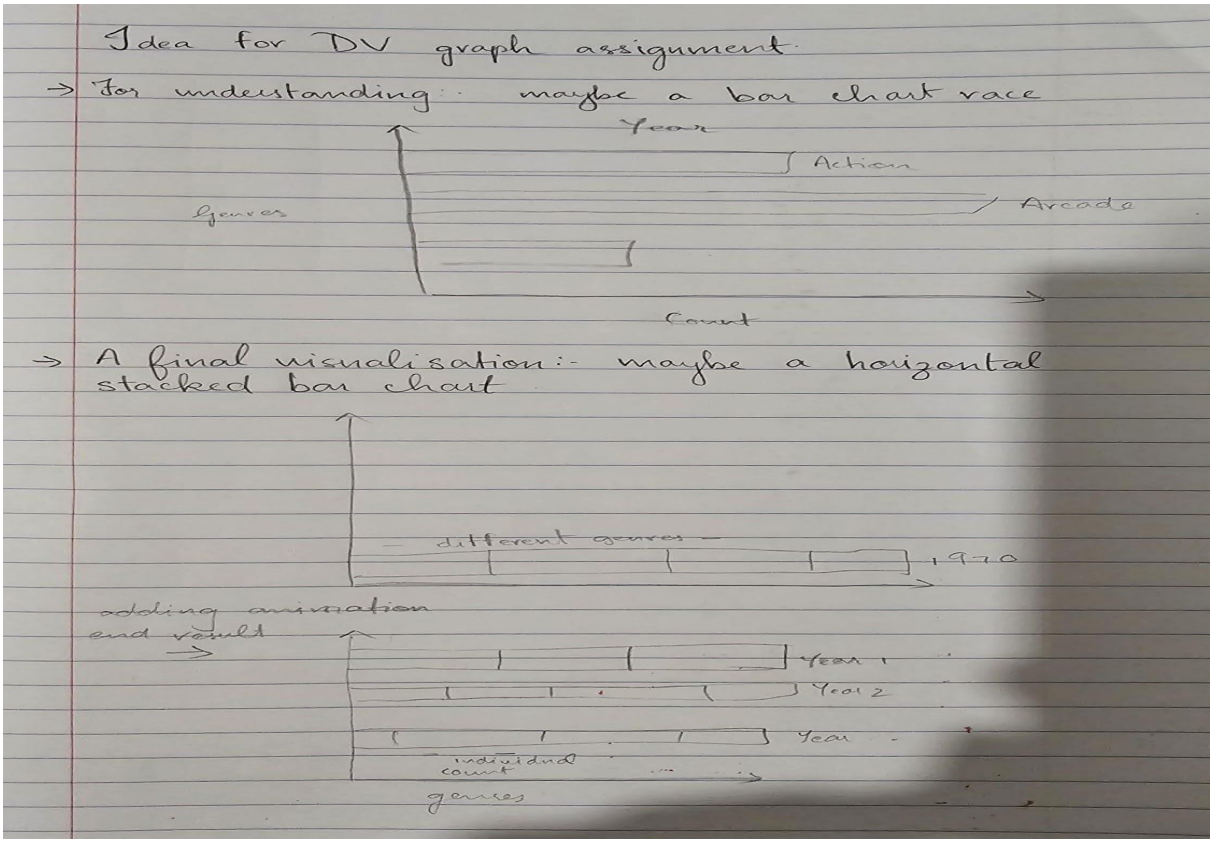
After cleaning all the datasets individually, we combine them together to form one uniform dataset. Before doing so, we make sure that all the datasets follow the same format. We also divide the records that contain a list of multiple genres into various records (for the same game).

At the end of the whole cleaning process, we obtained a dataset with 535,348 records (inclusive of the same game containing different genres). However, there was still a lot of bias in the data as it contained limited records from 2020 to 2024, thus not showing the true state of the current gaming industry.

3. Visualisation:

Our initial visualisation plan was to create a racing bar chart inorder to identify what genres were trending in what year and then for our final visualisation, create a stacked bar chart to show the distribution of genres across years. We planned to make it animated to show a continuous flow. However, since the last few years of the data were limited, the end of the animated chart wasn't giving out any useful information. This is why we decided to use the static version as our final visualisation.

Our initial sketch :



The racing bar chart was used in order to understand the data we will be working with a little better. It showed how the ‘Action’ genre was the most popular throughout all the years along with the rise and fall of other genres. It was able to help us understand how applying log on our actual graph (the stacked bar chart) wasn’t yielding accurate results because the action genre looked the smallest.

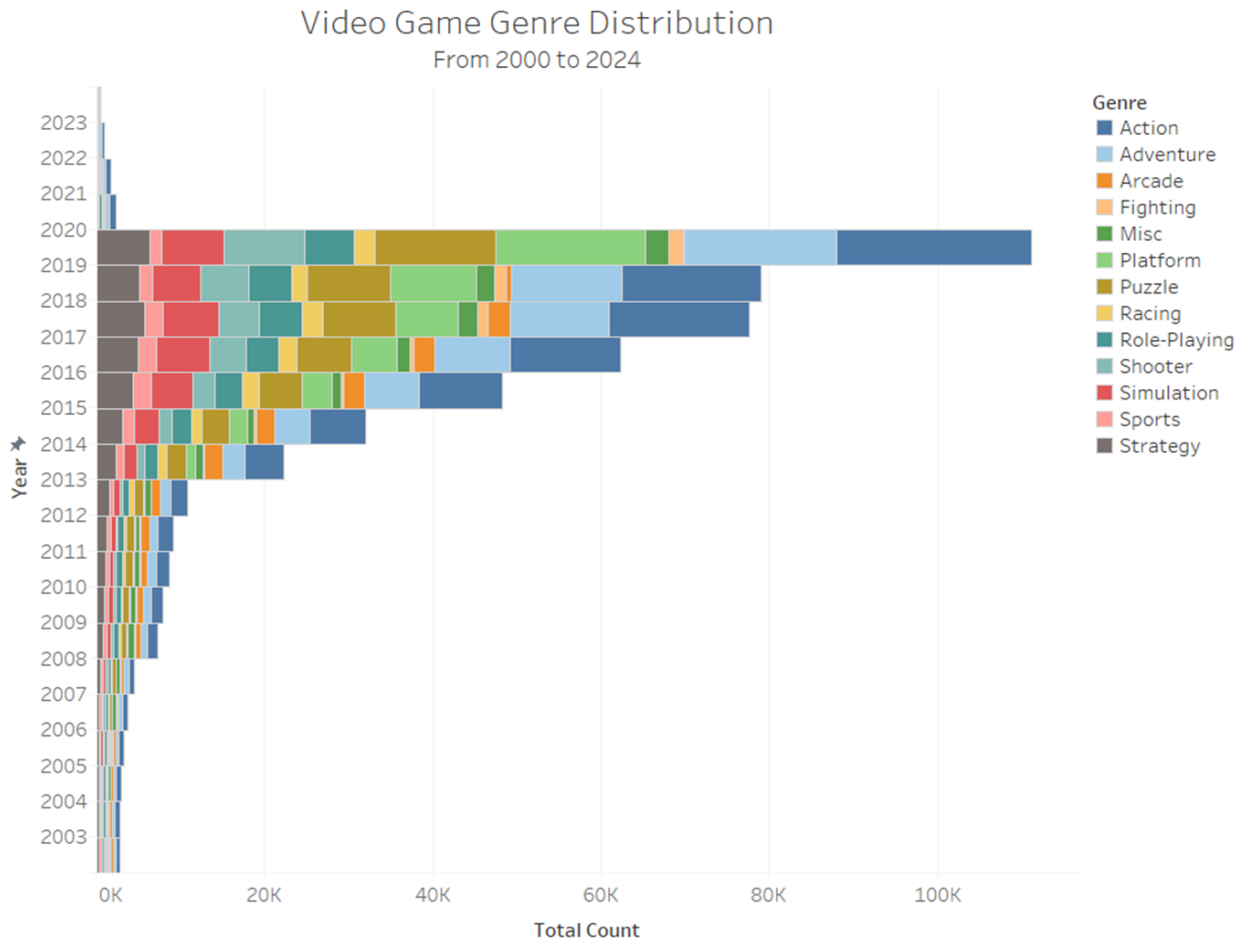
Keeping the basic nature of the data in mind after looking at the bar chart, we went forward with our initial idea of a stacked bar graph. Each bar represents a particular year, separated by various genres based on count. The x-axis denotes the total amount of records for that year while the y-axis represents the year. We decided on this chart because we thought it would be able to capture the complexities of genre and year within a singular chart.

On executing it with animation, we analysed that the data that we gathered for the year 2021 and above weren’ sufficient enough and hence showed relatively low values on the chart. Also due to a larger percentage of the dataset being centred around the 2012-2019 region, the other data points looked relatively small and

clustered. Since the animated chart would end on the last year (which is 2024) that didn't give out very useful information, we decided to recreate its static version.

The initial static version of the graph (containing data from 1950 to 2024) looked very crowded and not understandable. We decided to make the bars more legible by using logarithmic values, but that didn't work out as it gave out wrong information. For instance, it made the 'Strategy' genre look bigger than the 'Action' genre despite the data proving the opposite. So we decided to cut down the years and focus on creating a visualisation centred around the 2000's, as that is when technology and video games peaked.

Our final visualisation looks as follows:



While the data still looks crammed and doesn't capture the true statistics of the gaming industry in the current age, we still think the selected chart dictates a story of

(a) How the number of games have grown over the years,
(b) What genres are more popularly produced in within their time periods, and
(c) How the different genres have grown.

4. *Conclusion:*

To sum up, our examination of the video gaming scene from the 1980s to the present has provided valuable insights on the development of the sector. We overcome difficulties with data integration and cleaning by using a wide range of datasets to produce an extensive dataset with 535,348 records. We sought to provide a cohesive picture of gaming trends in spite of inherent biases and challenges.

Originally intended to be animated charts that would show genre trends over time in a dynamic manner, data restrictions for the last several years forced a change to a static visualisation focused on the critical 2000s. This choice gave us the opportunity to concentrate on a critical era that saw the pinnacle of gaming and technology development.Even though it is packed with information, the final static stacked bar chart tells an engaging story. It shows how the number of games has increased, identifies the most popular genres during particular periods, and depicts the changing trends in various gaming genres.

Although the dataset struggles with constraints from recent years, making it difficult to portray the entire state of the game business, the visualisations offer insightful information about past trends. The 'Action' genre's domination and the ebb and flow of other genres reflect the dynamic nature of the game industry.

Our endeavour is essentially a snapshot of the lengthy history of the gaming business. It highlights how crucial it is to have ongoing data updates and flexible visualisation techniques in order to keep up with the changing dynamics of the gaming industry. As new genres and technological advancements occur, our investigation establishes the foundation for future research, encouraging more studies into the intriguing world of video games.

5. *Tools used:*

1. Python:
   a)Pandas: For data manipulation and analysis.
   b)NumPy: Essential for numerical operations in Python.

2. Excel: For initial data exploration, cleaning, and simple visualizations,

3. Google Colab: A free, cloud-based version of Jupyter Notebooks that allows for easy collaboration and computation,

4. Tableau: A powerful data visualization tool that allows for the creation of interactive and shareable dashboards,

5. Rawg API: Used for retrieving video game data to incorporate into the visualizations.

6. Steam API: If utilized, it would have been for accessing data related to games available on the Steam platform,

6. *References*

1. Dataset was taken from kaggle
2. Reference for the api - RAWG Video Games Database API. RAWG, n.d. https://api.rawg.io/docs/

The video presentation is here
https://vimeo.com/890427944/ed9db059f2?share=copy