ETL (Extraction, Transformation, Load)

Project Documentation

Javier Alejandro Vergara Zorrilla

Dayanna Vanessa Suarez Mazuera (2221224)

Juan Camilo Buitrago Gonzalez (2221314)

David Felipe Martinez (2205286)

Universidad Autónoma de Occidente

Group 4

February 8, 2024

This document content a resume what was done in our project, we will explain in detail every step in the process and explain what did we do, what was the porpoise of the investigation and the conclusions we developed.

This is the dataset which was used for the research:

- **Name:** all_video_games(cleaned).csv
- **Number of Columns:** 9 (Title, Release Date, Developer, Publisher, Genres, Product Rating, User Score, User Ratings Score, Platforms Info)
- **Number of Rows:** 14.035
- **Kaggle Link:** https://www.kaggle.com/datasets/beridzeg45/video-games

Migration to the Database

PostgreSQL was chosen as our relational database, due to great capacity to handle larger volumes of data, the multiples advanced tools and the high opinion and reputation among users.

The first objective, is migrate the data to our database PostgreSQL; Within the “src” there are two important files: “config.py” houses the connection with the PostgreSQL where there’s a function called “config” which read the information about the PostgreSQL connection.

This function uses the python class “ConfigParser” for parse the file in an INI format, then, the necessary connection parameters will be extracted and returned as a dictionary. If the requested information is not founded an exception will be raised.

The second file, 'main.py', establishes a connection with the PostgreSQL database using the information provided by the “config.py” file. It creates a table named 'games' in the database, confirms the transaction, and then closes the connection, also, if there’s an existing table with that name, it will be removed and recreated again with the information from the dataset. For each row in the dataset, the SQL command will be formatted.

Once all the rows have been added with the information, the changes are saved and the connection gets close.

EDA and Analysis

Once the dataset has been sent to the database and recovered with Python, we proceed to develop the EDA (Exploratory Data Analysis), the objective of the EDA is to understand the information in the dataset, transform it if it's necessary, and based on this cleaning process draw conclusions.

The file for the EDA process is located in the folder “src” too.

First of all, the libraries must be imported, for our EDA, we will need:

- **Pandas:** This library will allow us read, understand and transform the dataset that we specific it.
- **Matplotlib and Seaborn:** This library will be useful to create graphics that will show the inf information transformed in order to get easier for us to obtain the conclusions that we need.
- **Psycopg2:** psycopg2 will provide the tools that we need to create the connection with the database.
- **StringIO:** StringIO provides an in-memory file-like object that can be used for reading from or writing to strings as if they were files

This was the case with EDA process:

Initial Data Exploration

The initial examination of the dataset is performed to understand its structure, variables, and overall content. Methods such as “head()” were used to visualize the first few rows, “info()” to obtain information about data types and null values, and “describe()” to obtain descriptive statistics for numeric variables.

Data cleaning and manipulation

Data cleaning tasks were performed to ensure data quality and consistency. This included removing null or duplicate values using methods such as “dropna()” and “drop_duplicates()”, correcting data types using “astype()” and “to_datetime()”, and normalizing data where necessary.

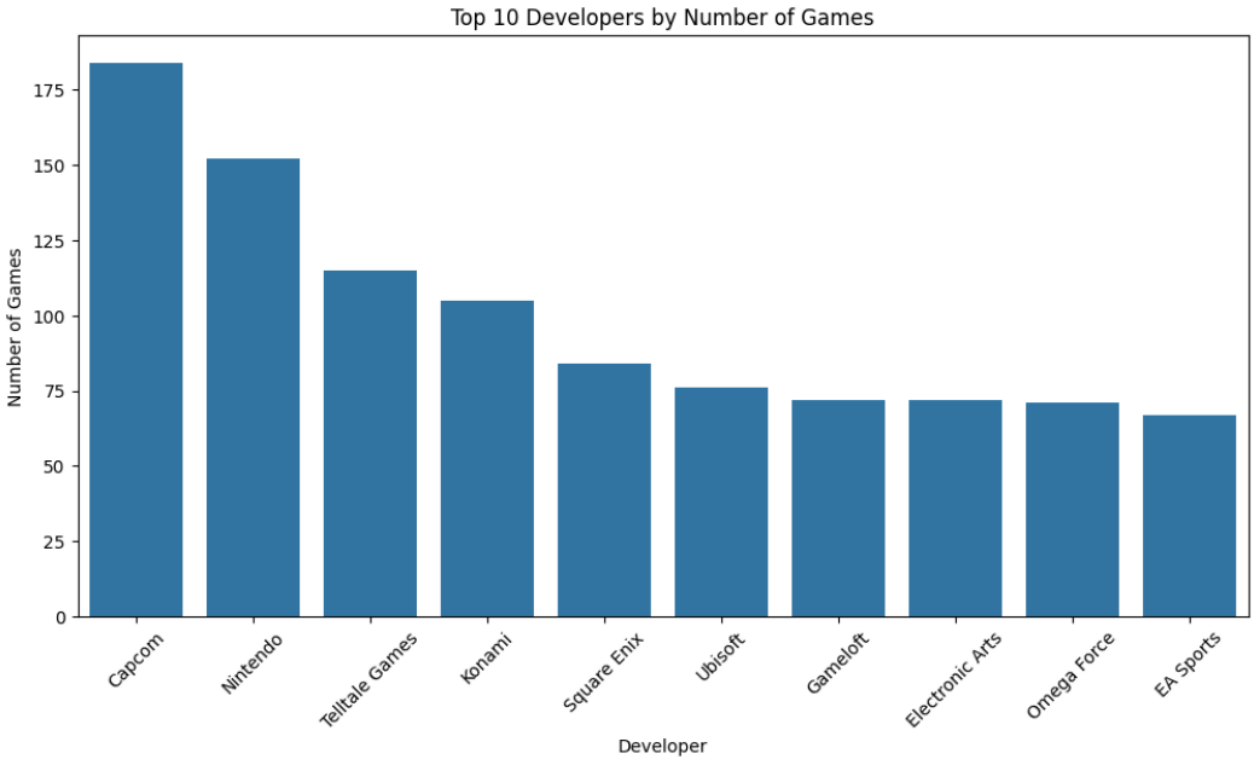
Data visualization

Various data visualization techniques were used to explore relationships between variables and to identify potential patterns or trends. These included creating histograms with “hist()”, scatter plots with scatterplot(), and bar charts with “countplot()” from the seaborn library.

Once the cleaning process has been completed, a new table in the database called “games_cleaned” is created, the data in the cleaned data frame is indexed in the new table using the method “copy_from()” of psycopg2.

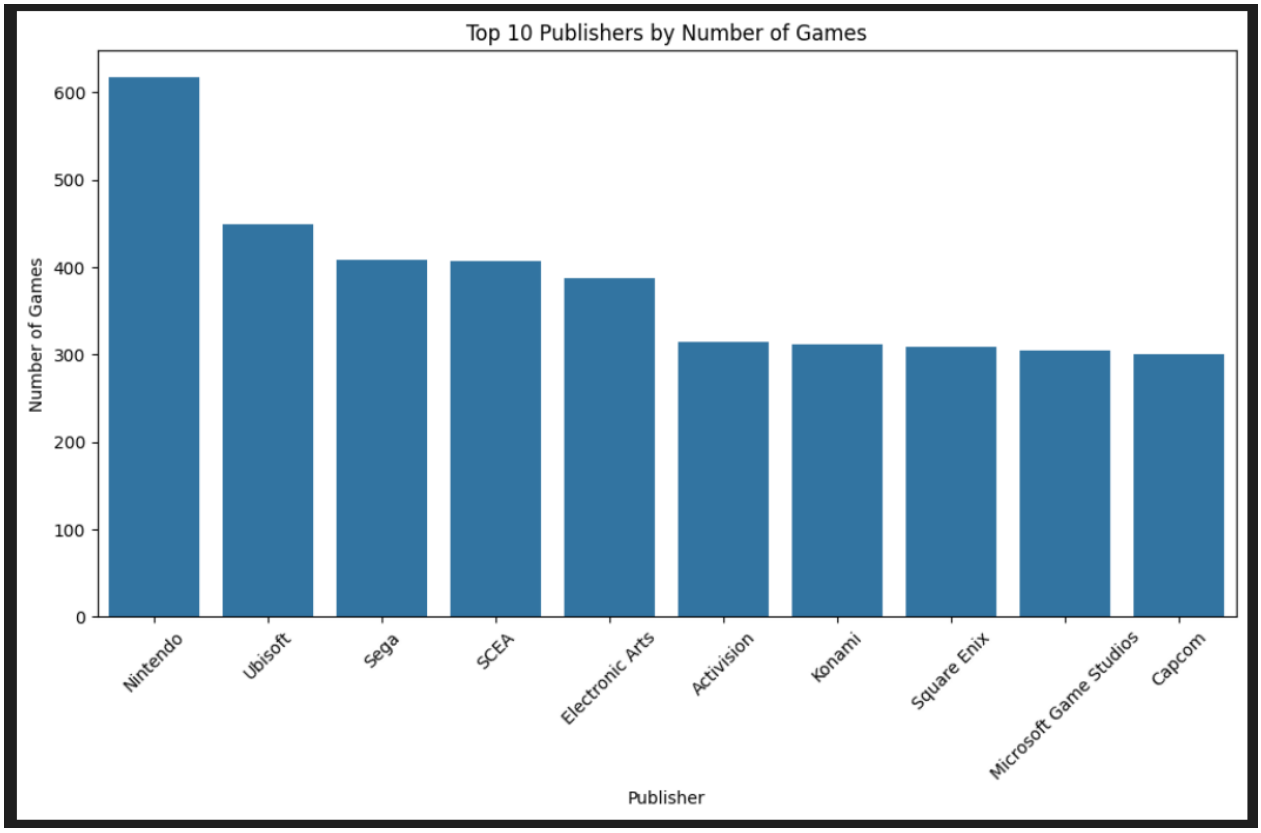
Graphics

1. **Top 10 Developers by Number of Games**



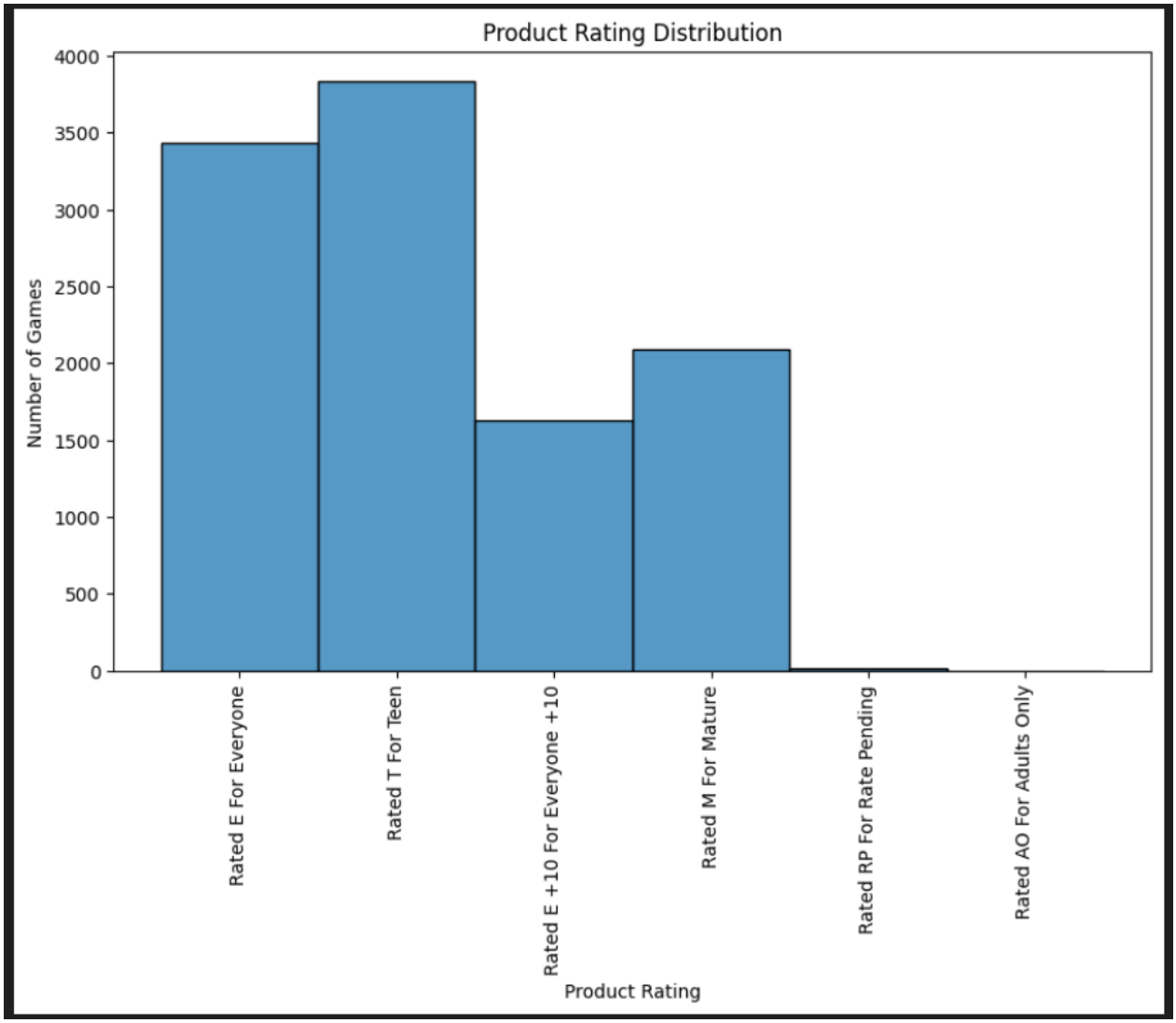
In terms of game production, Capcom leads the list with the highest number of games developed, indicating significant activity in the industry. Nintendo follows closely in second place, followed by Telltale Games in a prominent position. Konami and Square Enix occupy intermediate positions with moderate numbers of games in development. Ubisoft, Electronic Arts, and Omega Force are at a lower level in terms of game production, while Gameloft and EA Sports have the fewest developed games on the list. This distribution highlights the differences in production activity between the companies mentioned.

2. **Top 10 Publishers by Number of Games**



Looking at the number of games published by each publisher in the video game industry, a clear hierarchy emerges that reflects the influence and prominence of certain companies in the market. Nintendo and Ubisoft stand out as the clear leaders in this regard, demonstrating their broad influence and diversity within the industry. Further down the list are companies such as Sega and SCEA, which are not on the same level as Nintendo and Ubisoft, but still play a significant role in the market. Companies like Konami, Square Game Studios, and Capcom rank lower in terms of published games, but still have a significant impact on the market and remain familiar names to video game enthusiasts.

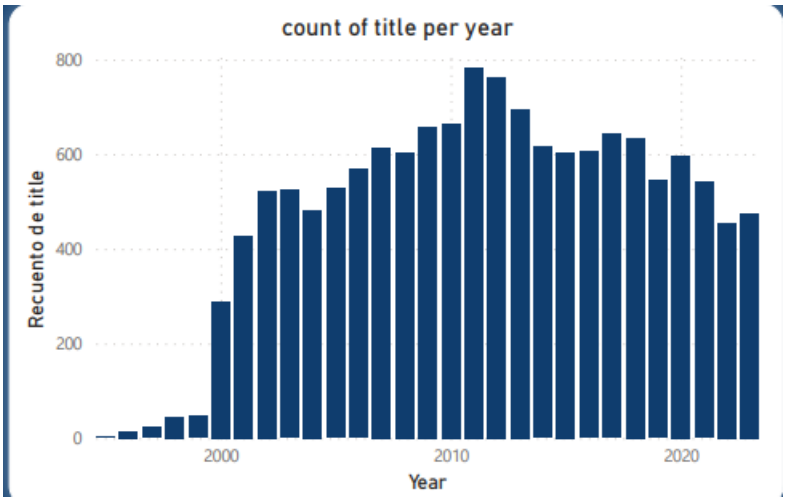
3. **Product Rating Distribution.**



Before analyze the different game classifications represented in the chart, the "Rated T for Teens" category stands out with nearly 4,000 games, indicating a clear preference for teen-oriented content. The "Rated E for Everyone" category has a solid presence with almost 3,500 games, indicating a significant demand for games suitable for all ages. The "Rated E +10 for Everyone +10" category has just over 1,500 games, indicating a particular focus on pre-teens. On the other hand, "Rated M for Mature" has about 2150

games, indicating a considerable supply of games for adult audiences. However, other less common classifications, such as Rated RP for Rate Pending and Rated AO for Adults Only, have little presence in the market, which could indicate a more rigorous rating process or limited demand for content that is still under review or that is intended exclusively for adult audiences. This analysis highlights the diversity and relevance of the various ratings in the video game industry, providing a comprehensive view of audience preferences and available offerings.

4. **Count of title per Year:**



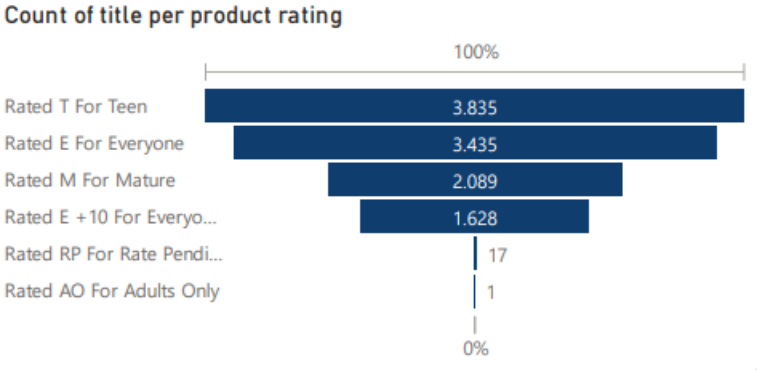
- This bar graph shows the number of video game titles released each year from 1995 to 2023.

- There is a notable increase in annual releases from 2000 to around 2011, where it reaches its peak.

- After the peak in 2011, there is a gradual decrease in annual releases.

Conclusion: The video game industry had significant growth in the first decade of the 21st century but has been slowly decreasing in terms of new titles released annually since then. It would be interesting to investigate the reasons behind this trend. Perhaps it could be due to market saturation, changes in consumer preferences, or the rise of mobile and indie games.

5. Count of title per product rating:

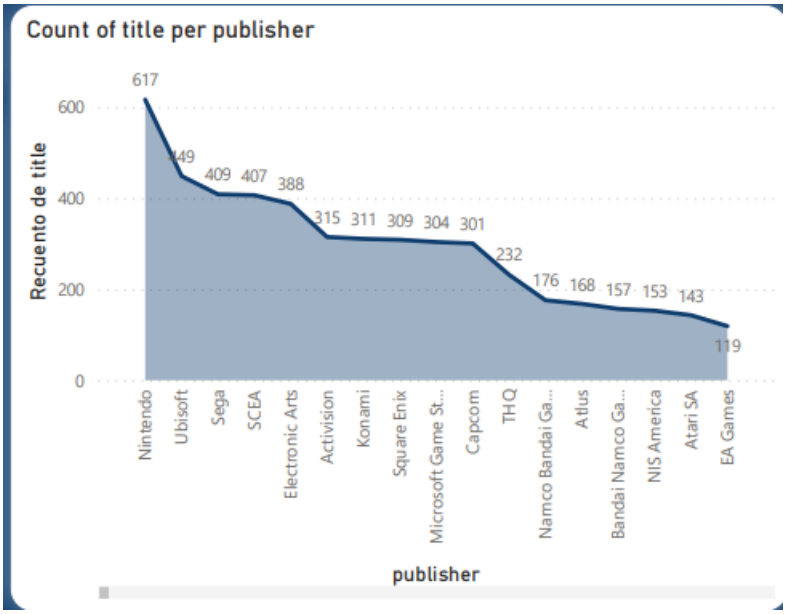


It shows the number of games rated according to their rating: Teen, Everyone, Mature, and others.

Most games are rated as “Teen” or “Everyone”, with a smaller amount rated as “Mature”.

Conclusion: Developers tend to create games that are accessible to a wider audience (teenagers and everyone), possibly to maximize sales.

6. Count of title per publisher:



The “Count of title per publisher” graph is a bar chart that represents the number of video game titles published by different publishers. Each bar corresponds to a publisher, and the height of the bar indicates the number of titles that publisher has released.

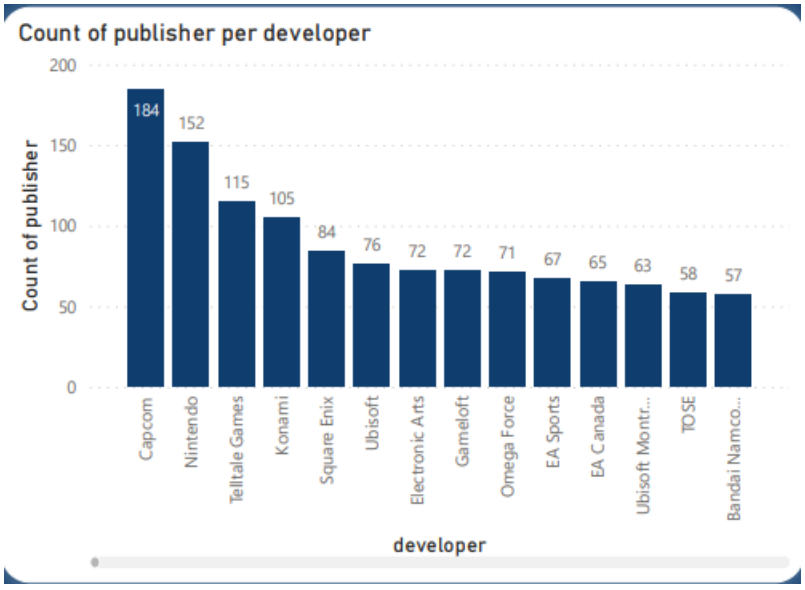
From the graph, we can see that the publishers are Nintendo, Ubisoft, SEGA, Sony, Electronic Arts, Activision, Square Enix, Microsoft Game Studios, Capcom, THQ, Namco Bandai Games, Konami Digital Entertainment, NIS America Inc., Atlus, and EA Games, etc.

Nintendo has the highest count with 617 titles, indicating that they have a wide range of games under their belt.

EA Games has the lowest visible count with 119 titles, suggesting that they might focus on fewer, high-quality titles.

Conclusion: This graph shows the distribution of game titles among different publishers and gives us an insight into their production rate. It’s clear that some publishers are more prolific than others, which could be due to various factors such as their size, resources, and business strategy. It’s also worth noting that a higher count of titles does not necessarily mean higher quality or success in the market.

7. Count of publisher per developer



The graph is titled “Count of publisher per developer”. It’s a bar chart that represents the number of publishers associated with each developer. The x-axis represents different developers, and the y-axis shows the count of publishers.

From the graph, we can see that:

Capcom has the highest number of publishers at 184.

Nintendo follows with 152 publishers.

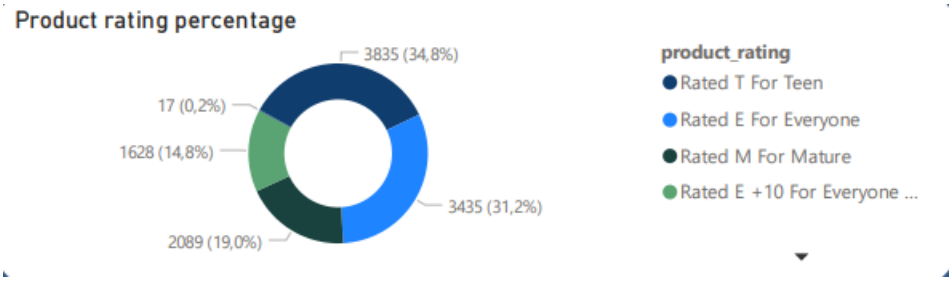
Konami has 115, Square Enix has 105, and Ubisoft has 84.

Electronic Arts and EA Sports have similar counts at around 70.

Other developers like Omega Force, Ubisoft Montreal, TOSE, and Bandai Namco have counts ranging from 57 to 72.

Conclusion: The graph shows the distribution of publishers among different developers. Capcom and Nintendo are leading in terms of diversity in publishers. There is a significant drop from Nintendo to Konami. The rest of the developers have relatively close counts of publishers, indicating a competitive environment with no significant dominance by any single developer beyond Capcom and Nintendo. This could reflect the business strategies, market presence, and partnerships of these developers.

8. **Product rating percentage**



The graph is a donut chart titled “Product rating percentage”. It represents the distribution of product ratings. Each segment of the donut chart corresponds to a different product rating category, and the size of each segment represents the percentage of products in that category.

From the graph, we can see that:

Rated T For Teen: This is the largest segment, representing 34.8% (3385 products) of the total. These are games that are suitable for ages 13 and up.

Rated E10+ For Everyone 10 and Older: This is the second largest segment, representing 31.2% (3435 products) of the total. These games are suitable for ages 10 and up.

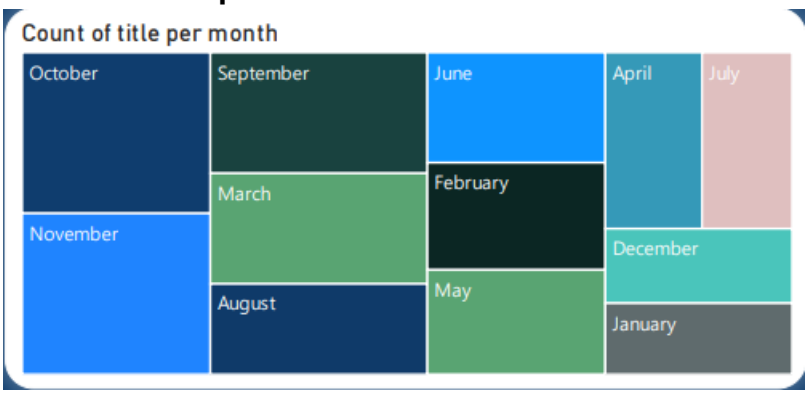
Rated E For Everyone: This segment represents 19% (2089 products) of the total. These games are suitable for all ages.

Rated M For Mature: This segment represents 14.8% (1628 products) of the total. These games are suitable for ages 17 and up.

Unspecified Category: This is the smallest segment, representing only 0.2% (17 products) of the total.

Conclusion: Most of the products are rated either T (Teen) or E10+ (Everyone 10 and Older), indicating that they are suitable for teenagers and pre-teens. The smallest segment is the unspecified category, which suggests that most products have a specified rating. This distribution could reflect the target demographics of the video game industry, with a focus on teenagers and pre-teens.

9. **Count of title per month**



The graph is a treemap titled “Count of title per month”. It represents the distribution of video game titles released each month. Each segment of the treemap corresponds to a different month, and the size of each segment represents the count of titles released in that month.

From the graph, we can see that:

October: This is one of the larger segments, indicating that a significant number of titles are released in October.

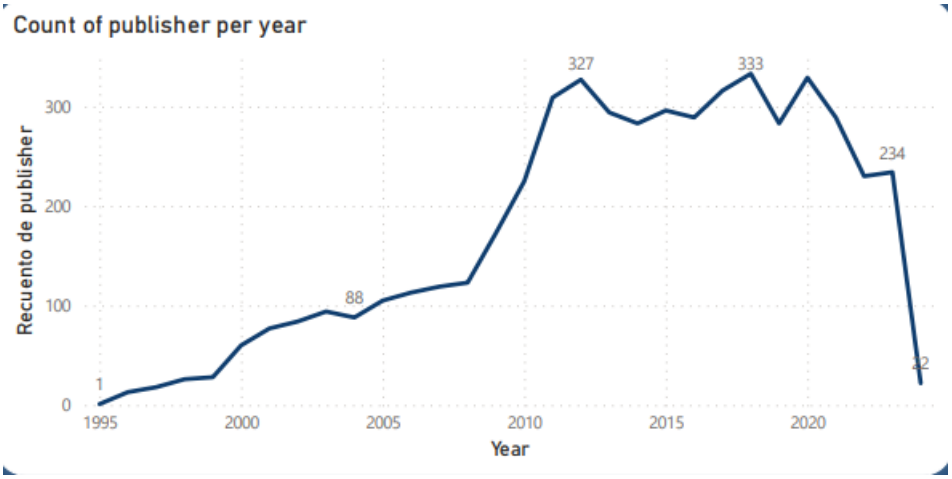
November: This is another large segment, suggesting that November is also a popular month for game releases.

September and March: These months have medium-sized segments, indicating a moderate number of game releases.

The rest of the months (August, June, February, May, April, December, January, and July) have smaller segments, indicating fewer game releases.

Conclusion: Most video game titles are released in October and November, which could be due to the holiday season when demand for video games is high. The months with the fewest releases are spread throughout the year. This distribution could reflect the strategic planning of the video game industry to maximize sales.

10. Count of publisher per year



The graph is a line chart titled “Count of publisher per year”. It represents the number of publishers each year. The x-axis represents the years from 1995 to 2023, and the y-axis represents the count of publishers.

From the graph, we can see that:

The count of publishers has generally increased from 1995 to around 2018.

There is a slight drop after 2018, but the count remains relatively high.

Conclusion: The graph shows a trend of increasing publishers over the years, with a slight drop after 2018. This could be due to various factors such as market saturation, changes in the publishing industry, or other factors. Despite the slight drop, the count remains relatively high, indicating a healthy and competitive publishing landscape in the video game industry up to 2023.