STEAM GAMES ANALYSIS

Overview

Steam is a digital distribution platform developed by Valve Corporation for purchasing and playing video games. It was launched in September 2003 and has since become the largest digital distribution platform for PC gaming.

About this dataset

The dataset "All Steam Special and defined metadata" is a comprehensive collection of data encompassing all games available on the Steam platform, along with their corresponding metadata. It serves as a valuable resource for researchers, developers, and gaming enthusiasts interested in exploring and analysing the vast Steam gaming ecosystem.

The dataset includes information about each game, such as its title, release date, developer, publisher, genre, user reviews, ratings, and system requirements. It covers a wide range of game genres, including action, adventure, strategy, role-playing, simulation, sports, and more, providing a diverse and extensive representation of the Steam game library.

Business Problem

In recent years, various games on the Steam platform have experienced high rates of negative reviews and low user engagement. This has resulted in fewer active players and ultimately lower revenues for game developers and publishers. Addressing these issues is critical for enhancing game visibility, user satisfaction, and profitability. Therefore, identifying and mitigating the factors contributing to negative reviews and low engagement is essential for improving the success and performance of Steam games.

Objectives

The primary objective is to analyse the factors influencing the success of games on Steam, with a focus on improving user engagement, reducing negative reviews, and increasing the estimated number of owners. To achieve this, we will identify key factors that influence game success, such as user reviews, pricing strategies, release dates, downloadable content (DLC), and the influence of developers and publishers. By enhancing user engagement, we aim to improve the overall gameplay experience, optimize content updates, and evaluate the impact of developer interaction with the gaming community. To reduce negative reviews, we will identify common complaints, address technical issues, and improve customer support. Additionally, we will explore marketing and promotion strategies, analyze genre popularity, and leverage sales and discounts to increase the estimated number of owners. Through strategic recommendations and data-driven decision-making, we aim to provide comprehensive business advice that enhances the overall performance and profitability of games on the platform. This includes maximizing user retention, optimizing revenue streams, and fostering long-term growth.

Research Questions

What are the variables that affect the success of Steam games?

How can we reduce negative reviews and improve user engagement for Steam games?

How can game developers and publishers be assisted in making pricing and promotional decisions?

Hypotheses

Higher prices lead to more negative reviews and lower user engagement.

Games with longer wait times for updates or bug fixes receive more negative reviews.

The majority of users prefer games from well-known developers or publishers, leading to higher engagement and positive reviews.

Tools and Technologies

Data Cleaning and Preparation

1. **Excel:**

Purpose: Initial data cleaning and preparation.

Usage: Removing duplicates, handling missing values, and basic data formatting.

2. **SQL:**

Purpose: Data cleaning, transformation, and validation.

Usage: Writing queries to clean data, perform transformations, and cross-verify findings from Python.

3. **Python:**

Purpose: Advanced data analysis, visualization, and calculations.

Libraries:

pandas: Data manipulation, analysis, and Data Visualization. matplotlib,

seaborn: Data visualization.

Usage: Conducting complex calculations, generating visualizations, and performing statistical analysis.

Dashboard Creation and Reporting

Data Source

This dataset is compiled from the Kaggle data repository.

Link for the dataset used for analysis

https://www.kaggle.com/datasets/mexwell/steamgames

Follow the below link for complete SQL and Python code

https://github.com/Nihar-Padhi/Steam-Game-Analysis

List of column types of the data set

Numeric – AppID, Estimated_owners, PeakCC, Required_age, Price, DLC_count, Metacritic_score, User_score, Positive, Negative, Achievements, Recommendations.

Mixed – Name, Developers, Publishers.

Categorical – Categories, Genres, Operating_system, Supported_language.

Date & time – Release_date, Average_playtime_forever, Average_playtime_two_weeks, Median_playtime_forever, Median playtime two weeks.

Information regarding the columns used for the analysis

1. AppID:

Description: A unique identifier for each game on the Steam platform.

2. Name:

Description: The title of the game.

3. PeakCC (Peak Concurrent Players):

Description: The highest number of players playing the game simultaneously at any given time.

4. Required_age:

Description: The minimum age requirement to play the game, is often due to content rating (e.g., 18 for mature content).

5. Price:

Description: The cost of the game in the default currency (usually USD).6. DLC_count: The number of downloadable content (DLC) items available for the game.

7. Metacritic_score:

Description: The average score given to the game by professional critics, aggregated by Metacritic.

8. User_score:

Description: The average score given to the game by users on Steam.

9. Positive:

Description: The number of positive user reviews for the game on Steam.

10. Negative:

Description: The number of negative user reviews for the game on Steam.

11. Achievements:

Description: The number of achievements that players can earn in the game.

12. Recommendations:

Description: The number of users who have recommended the game.

13. Average_playtime_forever:

Description: The average total playtime (in minutes) of the game across all users who own it.

14. Average_playtime_two_weeks:

Description: The average playtime (in minutes) of the game over the last two weeks across all users who own it.

15. Median_playtime_forever:

Description: The median total playtime (in minutes) of the game across all users who own it.

16. Median_playtime_two_weeks:

Description: The median playtime (in minutes) of the game over the last two weeks across all users who own it.

17. Estimated_owners:

Description: An estimate of how many users own the game.

18. Developers:

Description: The company or individual(s) who developed the game.

19. Publishers:

Description: The company or individual(s) responsible for publishing the game.

20. Categories:

Description: The various features or modes of the game, such as single-player, multiplayer, co-op, etc.

21. Genres:

Description: The genre(s) of the game, such as action, adventure, RPG, etc.

32.Operating_system:

Description: The operating system(s) that the game supports (e.g., Windows, macOS, Linux). 33.

Supported_language:

Description: The languages that the game supports, including interface, audio, and subtitles.

34. Release_date:

Description: The date when the game was officially released on Steam.

Exploratory Data Analysis

Creating a new table for the Genres, Language, Category, and Operating_system columns as they contain multiple values in a single cell. On importing the Data set to the My SQL for analysis this column will violate the **First normal form [1NF]** rule every cell should have atomic value.

Genres column

By using Excel's split columns under the data tab using ", "as the delimiter value splitting the values of each cell only containing the single value.

In the new table, there are a maximum of 16 genres after splitting there are 16 individual columns present.

There are too many null values in the new table. To tackle this problem and to reduce null values, keep only the first 3 columns with a minimum number of null values and remove the remaining columns.

□ Removing the other columns as there are maximum null values.

Dim_Genres Table

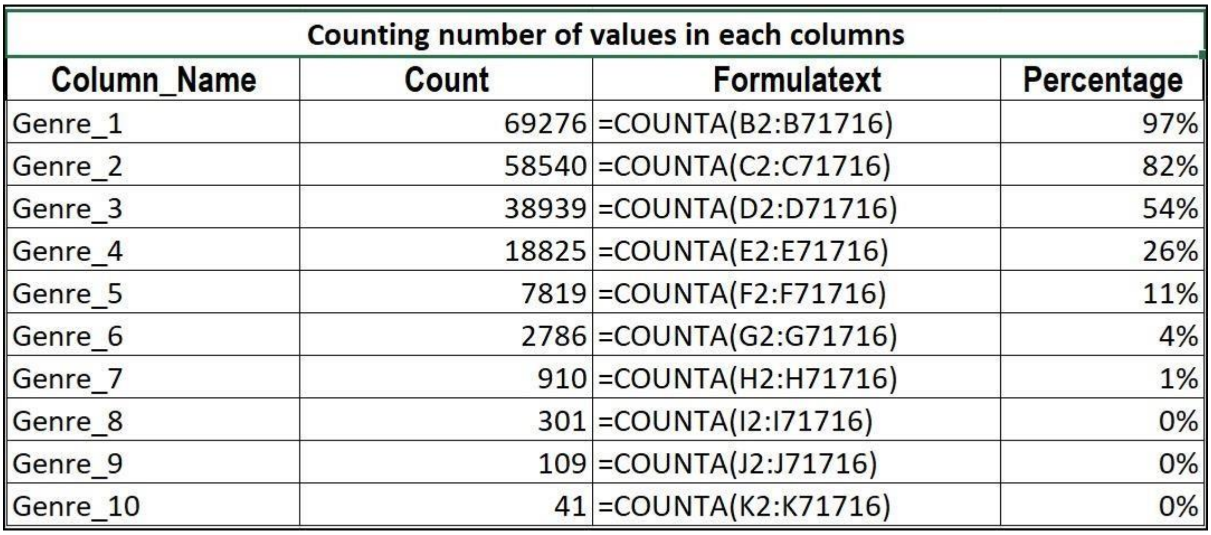
1. AppID

Description: Unique identifier for each title name. Data Type: Integer Example: 12345.

2. genre_1, genre_2, genre_3

Description: " The genre(s) of the game, such as action, adventure, RPG, etc

The fig 1.1 represents the number of values present in the fourth column



Counting number of values in each columns			
Column_Name	**Count**	**Formulatext**	**Percentage**
Genre_1	69276	=COUNTA(B2:B71716)	97%
Genre_2	58540	=COUNTA(C2:C71716)	82%
Genre_3	38939	=COUNTA(D2:D71716)	54%
Genre_4	18825	=COUNTA(E2:E71716)	26%
Genre_5	7819	=COUNTA(F2:F71716)	11%
Genre_6	2786	=COUNTA(G2:G71716)	4%
Genre_7	910	=COUNTA(H2:H71716)	1%
Genre_8	301	=COUNTA(I2:I71716)	0%
Genre_9	109	=COUNTA(J2:J71716)	0%
Genre_10	41	=COUNTA(K2:K71716)	0%

The above figure shows how many cells filled the Genres in each column have values. The columns with less than 50 % of values are removed.

Dim_Language Table

1. AppID

Description: Unique identifier for each title name. Data Type: Integer Example: 12345.

2. Language_1, Language_2, Language_3, Language_4

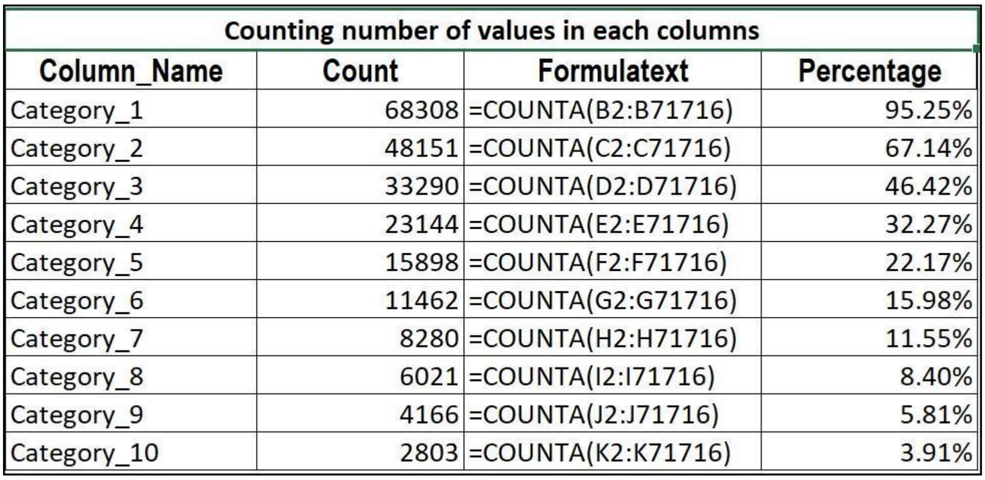
Description: The languages that the game supports, including interface, audio, and subtitles.

Dim_Category:

1. AppID

Description: Unique identifier for each title name. Data Type: Integer Example: 12345.

The fig 1.2 represents the number of values present in the fourth column



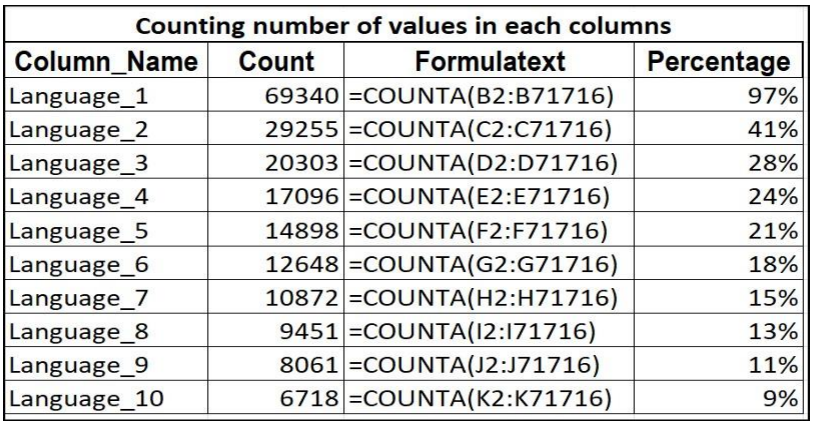
Counting number of values in each columns			
Column_Name	**Count**	**Formulatext**	**Percentage**
Category_1	68308	=COUNTA(B2:B71716)	95.25%
Category_2	48151	=COUNTA(C2:C71716)	67.14%
Category_3	33290	=COUNTA(D2:D71716)	46.42%
Category_4	23144	=COUNTA(E2:E71716)	32.27%
Category_5	15898	=COUNTA(F2:F71716)	22.17%
Category_6	11462	=COUNTA(G2:G71716)	15.98%
Category_7	8280	=COUNTA(H2:H71716)	11.55%
Category_8	6021	=COUNTA(I2:I71716)	8.40%
Category_9	4166	=COUNTA(J2:J71716)	5.81%
Category_10	2803	=COUNTA(K2:K71716)	3.91%

The above figure shows how many cells filled the Category in each column have values. The columns with less than 50 % of values are removed.

3. Category_1, Category_2, Category_3

Description: The various features or modes of the game, such as single-player, multiplayer, co-op etc.

The fig 1.3 represents the number of values present in the fourth column



Counting number of values in each columns

Column_Name	Count	Formulatext	Percentage
Language_1	69340	=COUNTA(B2:B71716)	97%
Language_2	29255	=COUNTA(C2:C71716)	41%
Language_3	20303	=COUNTA(D2:D71716)	28%
Language_4	17096	=COUNTA(E2:E71716)	24%
Language_5	14898	=COUNTA(F2:F71716)	21%
Language_6	12648	=COUNTA(G2:G71716)	18%
Language_7	10872	=COUNTA(H2:H71716)	15%
Language_8	9451	=COUNTA(I2:I71716)	13%
Language_9	8061	=COUNTA(J2:J71716)	11%
Language_10	6718	=COUNTA(K2:K71716)	9%

The above figure shows how many cells filled the Language in each column have values. The columns with less than 50 % of values are removed.

Finding the missing values or null from all the tables:

Select the complete title column to find the missing values from the individual values.

Home tab > Find & Search > Go to special > Blank

(OR)

Use the Excel formula =Countblank(range) for each column from all tables.

- The columns that need fixing are the title name has 1 missing name for Appid 396420 removing the row as there is no replacement for the value from all tables.
- In the table fact_Steamgames columns Developers and Publishers have null values and replace them as “Unknown”.
- The Dim_Geners tables Geners_1 has 2440 null values replacing it with “Others The Dim_Language table Language_1 has 2376 null values replacing it with “Others”.
- The Dim_Category table Category_1 has 3408 null values replacing it will “Others”.

Changing the format of the Release_date column

The Release_date column under Fact_Steamgames has the format of mmm dd-yyyy changing the format to “yyyy-mm-dd”.

 This was achieved by customer formatting and Ifs condition to convert the mmm to mm value.

Using Python to find the outliers present in the data set

Outliers:

Outliers are data points that deviate significantly from the majority of a dataset, often indicating variability, measurement error, or unusual occurrences. Detecting outliers can help identify errors, anomalies, or unique insights, and can be achieved using statistical methods like the IQR (Interquartile Range) method or Z-score analysis.

Detecting the outliers with the help of graphs like Boxplot, Histogram, density plot, or kde plot.

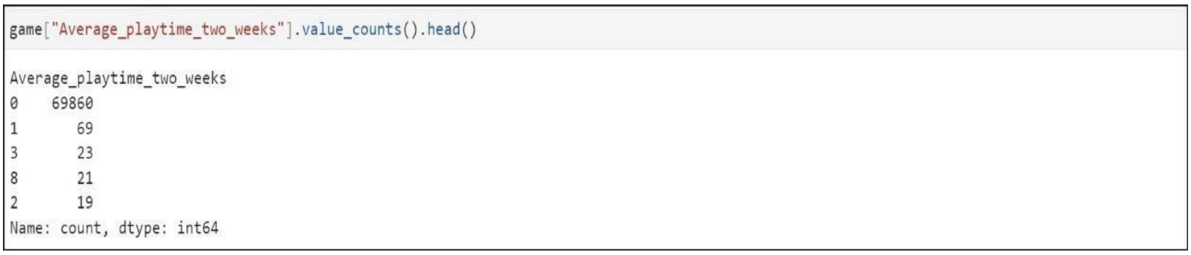
By checking the Skewness is between -0.5 to 0.5 which tells about the distribution.

There are too many 0 values which are unavailable or missing values making the column distribution highly Skewed and also incorrect.

Removing the columns that are more than 50% of 0 values will affect the quality of the column and distribution.

✓ The price column is an exception for outliers since some games are priced at 100 times the average price. Cross-verifying this with the website confirms that the prices are accurate. Therefore, these prices are retained despite the significant difference. List of the columns that are being removed:

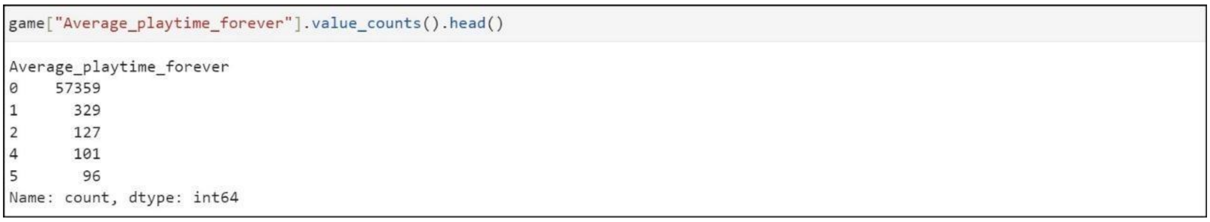
Average_playtime_two_weeks



```
game["Average_playtime_two_weeks"].value_counts().head()

Average_playtime_two_weeks
0    69860
1       69
3       23
8       21
2       19
Name: count, dtype: int64
```

Given that the column for “Average playtime over two weeks” has 69,860 entries with a value of 0, representing less than 1% of the data has values, it is clear that this column does not provide significant information for analysis. Therefore, it would be reasonable to consider removing or disregarding this column to avoid noise or misleading insights in your analysis. Average _Playtime_forever

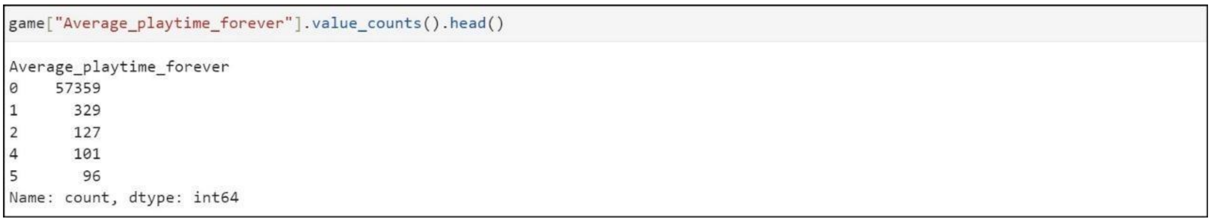


```
game["Average_playtime_forever"].value_counts().head()

Average_playtime_forever
0    57359
1      329
2      127
4      101
5       96
Name: count, dtype: int64
```

The column "Average playtime for two weeks" has 57,359 entries with a value of 0, which is not logically accurate and will negatively impact any analysis involving this column. Therefore, it is advisable to exclude or disregard this column to ensure the integrity of the analysis.

Median_playtime_two_weeks

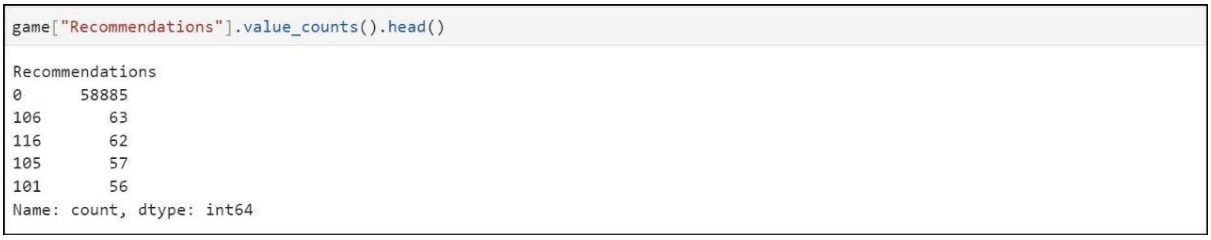


```
game["Average_playtime_forever"].value_counts().head()

Average_playtime_forever
0    57359
1      329
2      127
4      101
5       96
Name: count, dtype: int64
```

Similar to the other column, "Median playtime for two weeks" has 57,359 entries with a value of 0. This will not contribute to the analysis and will compromise the quality of the data. Therefore, it is advisable to exclude or disregard this column to maintain the integrity of the analysis

Recommendation

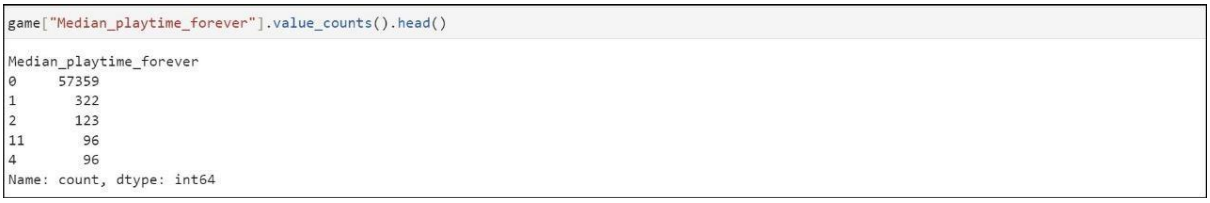


```
game["Recommendations"].value_counts().head()

Recommendations
0      58885
106       63
116       62
105       57
101       56
Name: count, dtype: int64
```

The column has very few data entries for the recommendations made by players. Upon verifying with the website, it is evident that 58,885 values are 0, which indicates that the data is not accurate. This will negatively impact the analysis.

Median_playtime_forever



```
game["Median_playtime_forever"].value_counts().head()

Median_playtime_forever
0     57359
1       322
2       123
11       96
4        96
Name: count, dtype: int64
```

This is the final column that needs to be removed. Similar to the previous columns, "Median playtime forever" contains inaccurate data, with 57,356 out of 70,000 entries being 0. This will not aid the analysis and acts as an outlier.

This column with ‘0’ values acts as an outlier, compromising the column's quality and distorting the mean value of the data. There are also missing and incorrect values. To maintain data integrity, it is necessary to remove these columns before proceeding further with exploratory data analysis (EDA) using Python.

EDA Compilation:

During the exploratory data analysis (EDA) stage, we iteratively cleaned the dataset and created new columns and tables to enhance our understanding. Using descriptive statistics and visualizations, we identified significant patterns and trends, particularly examining relationships between game prices, user reviews, and engagement levels. This preliminary analysis provided valuable insights into the dataset, laying the foundation for more detailed analysis and modelling in subsequent stages of the project.

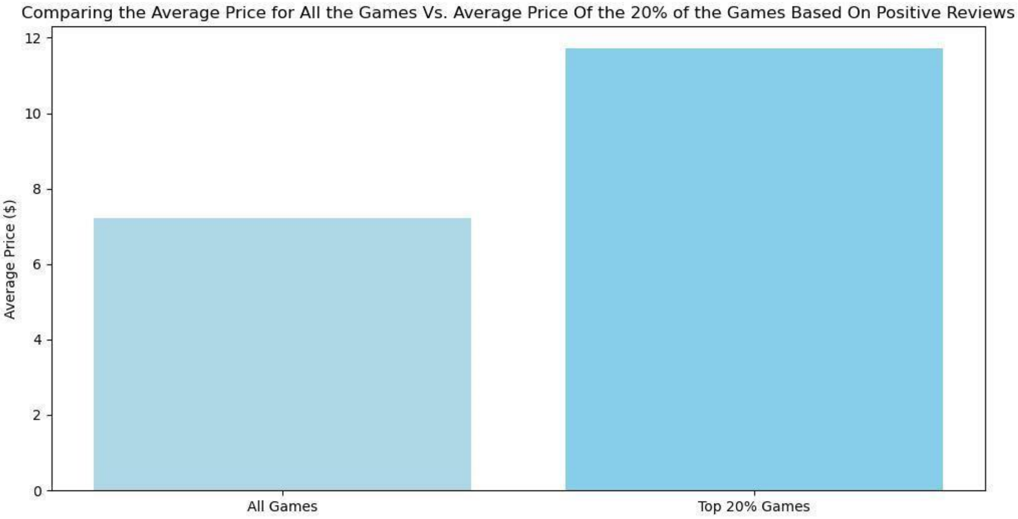
NOTE: EDA is an iteration process for any analysis while plotting the graphs or when finding the solutions for the problems additional cleaning and removing the data has to be done which may not be done in the initial stages.

Analysis and Findings:

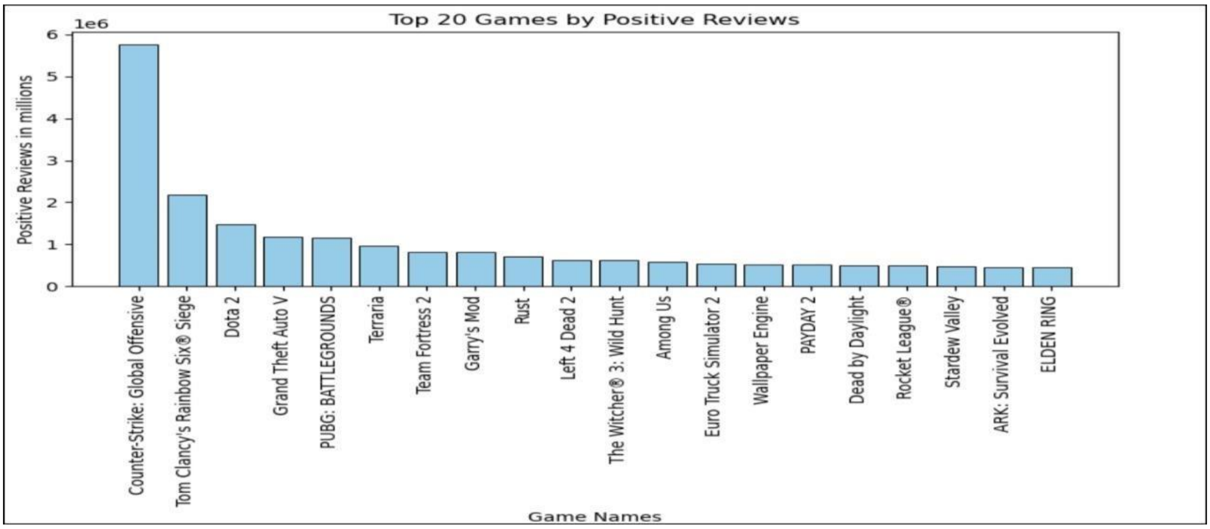
- **Positive review analysis**

The following graph shows the positive reviews based on different aspects like Price, DLC, Number of users, and Supported OS.

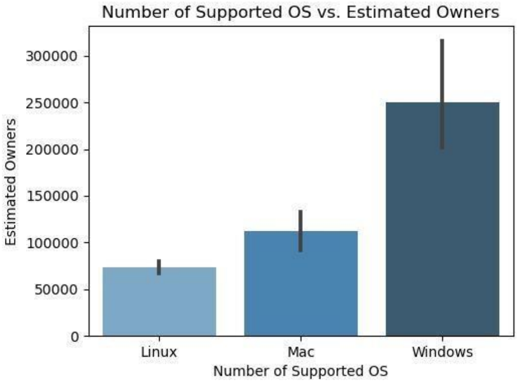
The accompanying bar graph shows the Top 20% of Games and other games based on Positive reviews and their total average price. There are still a significant number of positive reviews for the Top 20% of games based on the total average price. This shows that the users are more interested in buying the game even if the price is high and all give a positive review, which has significantly impacted the game's positive review.

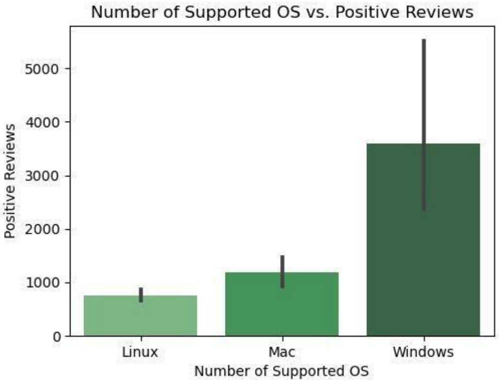


This bar graph demonstrates the top 20 most popular games based on positive reviews. This graph clearly shows how the positive reviews have drastically dropped from the highest being the ‘Counter strike: Global offensive’ and last being the ‘ELDEN RING’ with the difference of 5 million reviews showing how much positive reviews are unevenly distributed.



The bar graphs illustrate the usage distribution based on operating systems (OS) and the number of positive reviews. From the graphs, it is evident that Steam is predominantly popular among Windows users. Moreover, games on Windows receive a significantly higher number of positive reviews, around 3800, which is a notable figure.

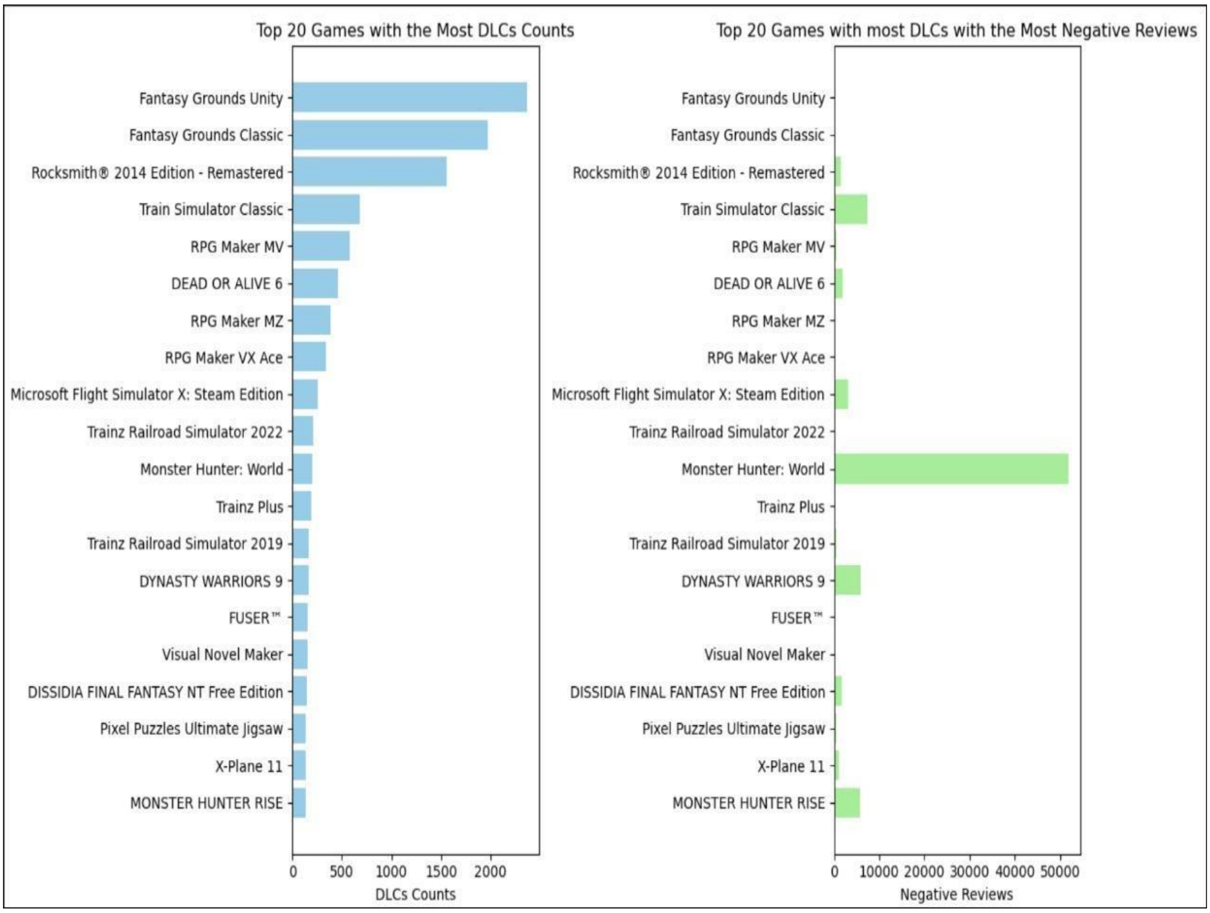




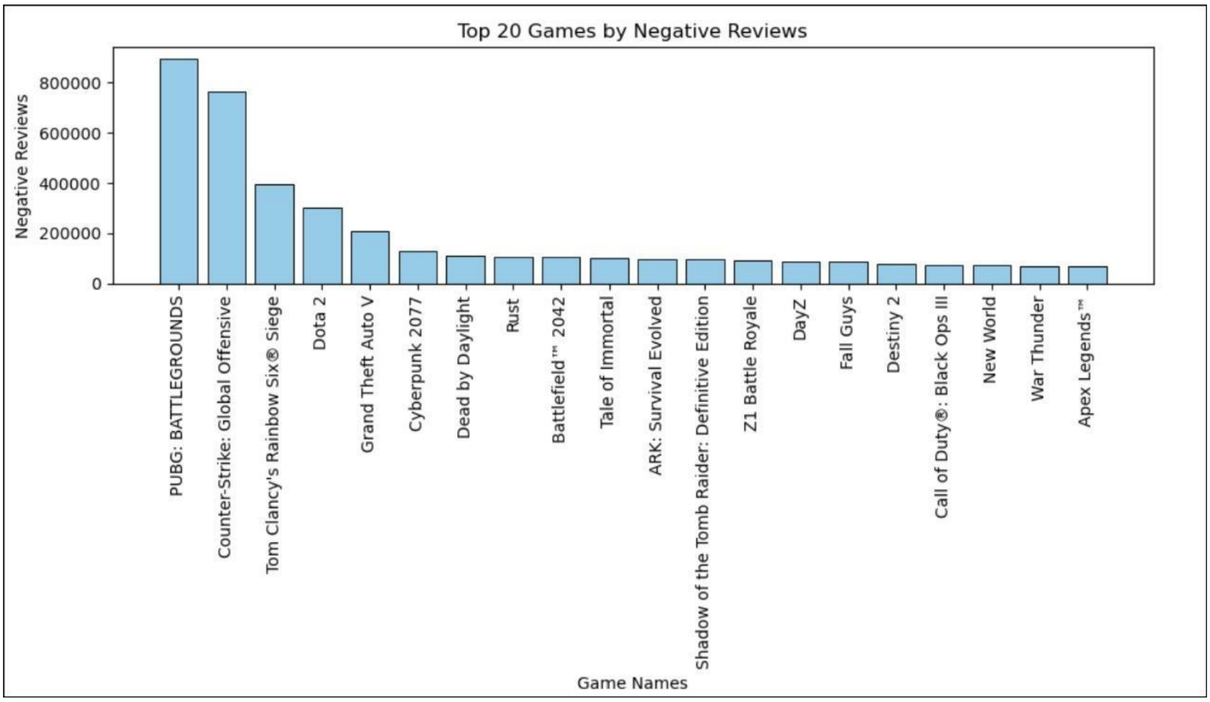
- **Negative review analysis**

The following graph shows the positive reviews based on different aspects like the DLC, Number of users, and Developers.

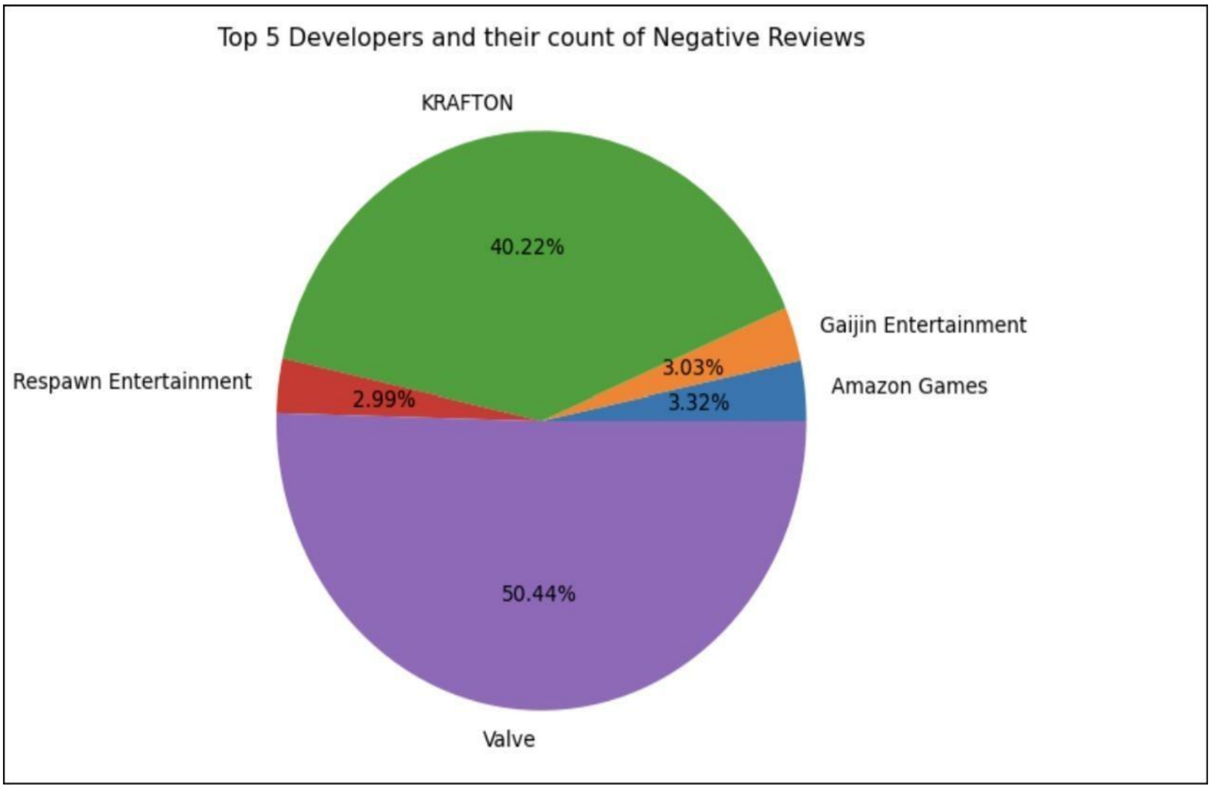
The graphs represent the top 20 games with the highest DLC and negative reviews which clearly shows that there is no relationship between the DLC counts and the number of the negative review.



This bar graph demonstrates the top 20 most popular games based on Negative reviews. The games that have the highest positive reviews are also been in the negative reviews should the reviews depend on the game's popularity.

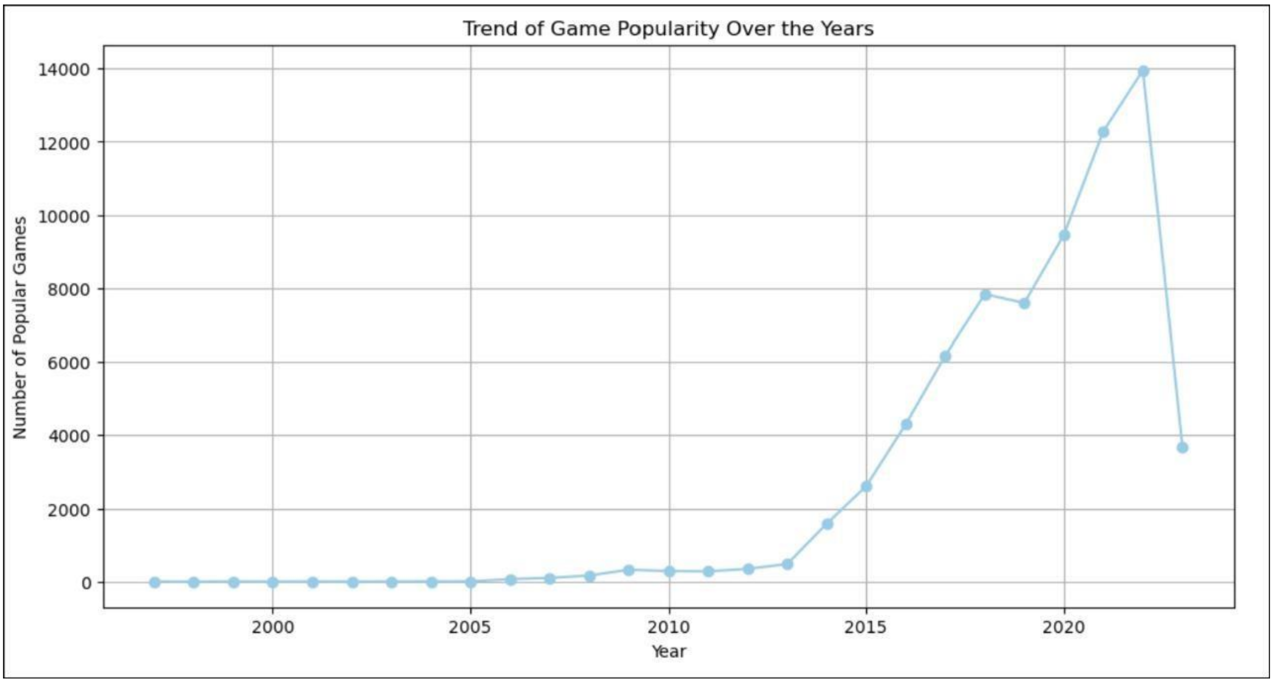


The pie chart represents the developers with the most negative reviews with the 'KRAFTON' and 'Valve' having 90% of the negative reviews among the top 5.

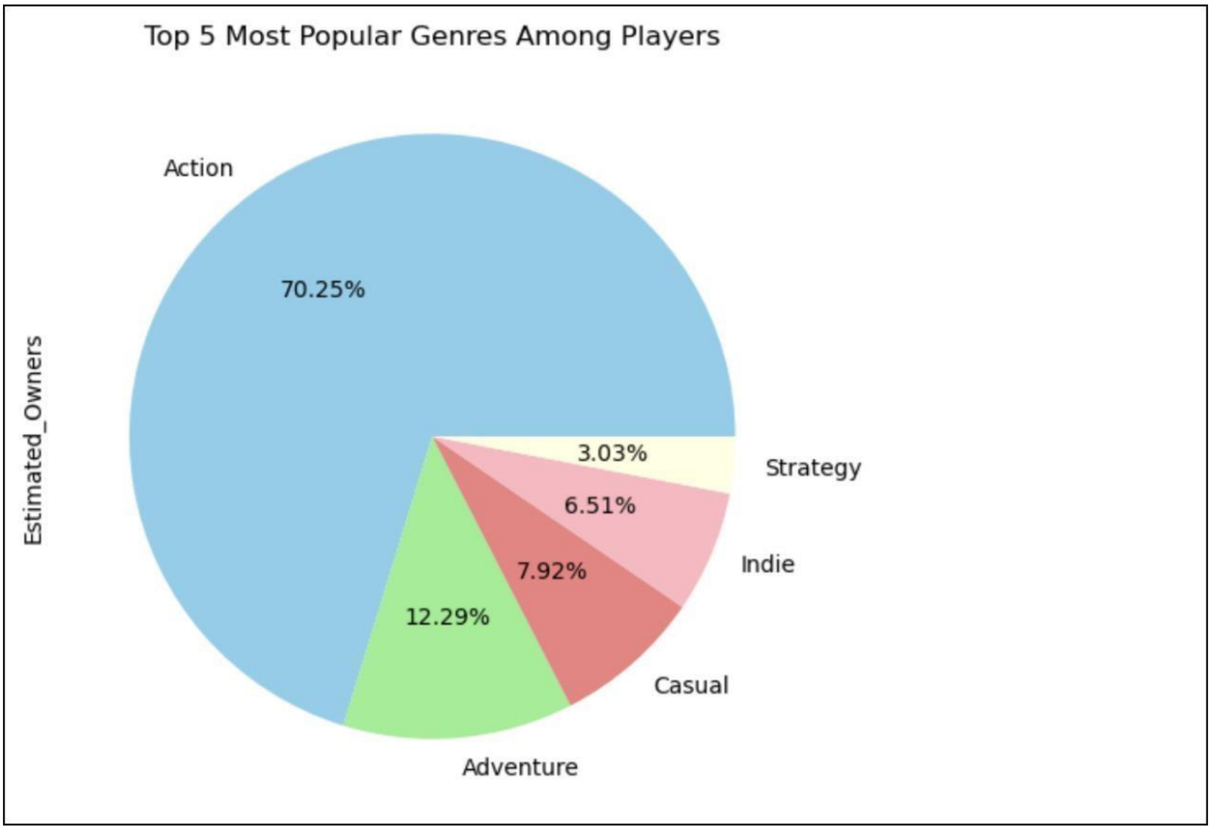


General analysis

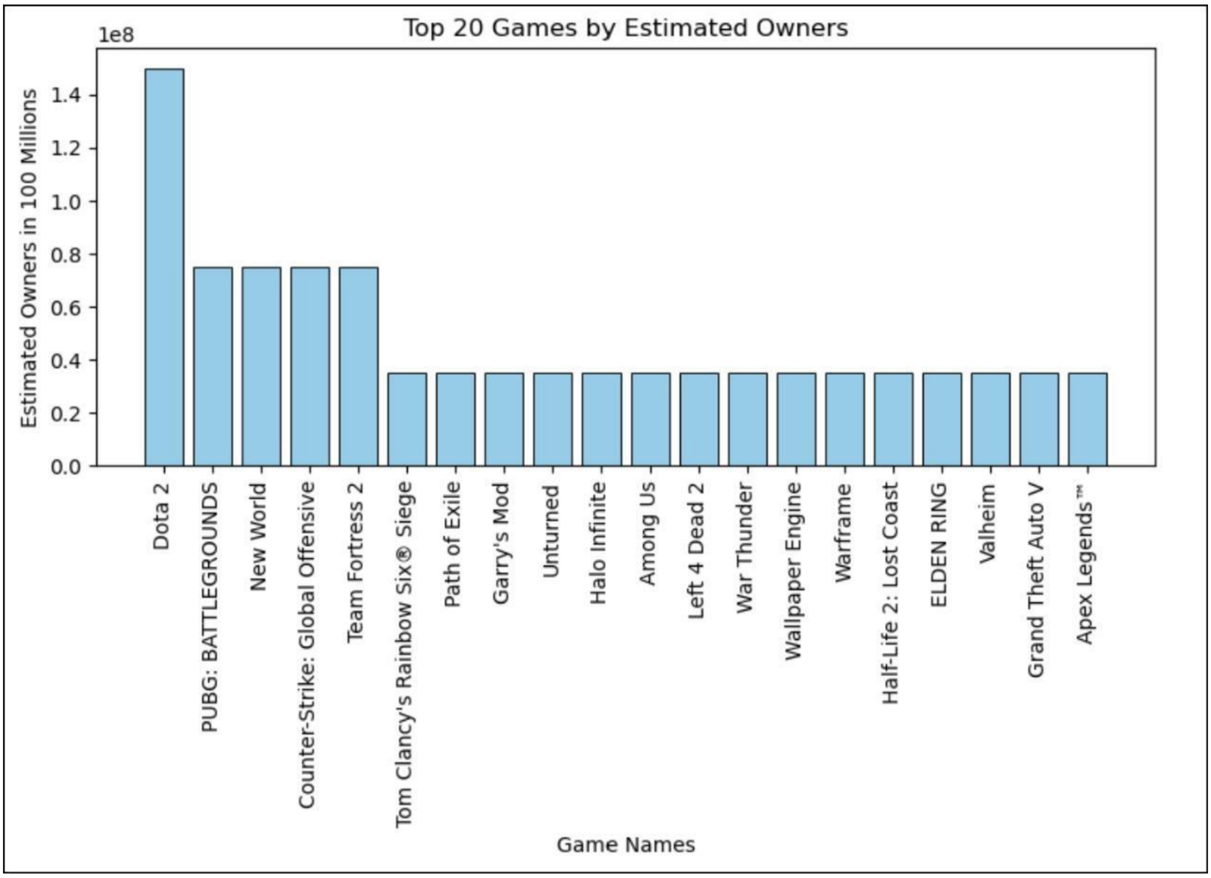
The line graph represents the trend of the popular games, it shows a upward trend from the year 2015 to 2022 after that there is a big leap in the popular games.



The pie chart represents the distribution of the player base on the different genres which shows that the most popular genre is action covering around 70% of the user base that concluded that most of the user base likes the action games.



The bar graph represents the top 20 titles with the most active user base with Dota 2 being a multiplayer online battle arena (MOBA) with an active player base of 140 million users owing the game title.



Finding the answers to some of the common questions:

Solution for the questions using Python

1. What are the Highest Priced and Lowest Priced Games?

Highest Price And The Game

```
steam_games[steam_games["Price"]>0]["Price"].max() steam_games[steam_games["Price"]==999]
```

Lowest Price And The Game

```
steam_games[steam_games["Price"]>0]["Price"].min() steam_games[steam_games["Price"]==0.37]
```

Free Games

```
steam_games[steam_games["Price"]==0]["AppID"].count()
```

- Findings

- We have 12409 games that are free of cost and out of 71716 games 59305 games are paid games from which the average Price is around 8.73.

And the highest Priced Games is Ascent Free Roaming VR Experience with a price tag of 999 Developed and Published by Fury Games for Windows only on 2019-12-27.

And the lowest Priced Game is IN THE BUILDING: CATS 3 with a price tag of only 0.37, Developed and Published by Laush Dmitriy Sergeevich and Laush Studio Respectively For only Windows on 2023-01-05.

2. Which Developer has the most successful Game?

- Popular Developer

```
steam_games.groupby("Developers")[["Estimated_Owners","Positive"]].sum().sort_values(by="Estimated_Owners",ascending=False)
```

- Most Popular Game

```
steam_games[steam_games["Developers"]=="Valve"].sort_values(by="Estimated_Owners", ascending=False)
```

- Findings

The Biggest Developers Valve with over 573 million Users Globally. And their DOTA-2 Game is very popular world wide as it is having 150 million users released for all the OS, a multiplayer Action game which was released on 2013-07-09. They are dominating the Top of the Chart as the Amazon games and as well as Krafton Games are also in the top 3 list still they are having only around 75 million users each.

3. Which Publisher has the most Game?

-Publisher

------------- steam_games.groupby("Publishers")[["Estimated_Owners","Positive"]].sum().sort_values(by="Estimated_Owners",ascending=False)

- Findings

- The Biggest Publishers Valve with over 629.87 million Users Globally. And in the Top 5 there are also Ubisoft, Electronic Arts, Sega, and 2k with each having nearly 120-160 million users worldwide.

4. Top Most Popular Genres?

Top Genres

----

steam_games.groupby("Genre")["Estimated_Owners"].sum().sort_values(ascending=False). head(5)

- Findings

- Most of the games are based on the action genre but there are also adventure and Casual games that are being developed widely. Action games are the most popular with a huge player base of 4.45 billion gamers at the top of the chart but Adventure, casual, and Indie games are also very popular with 779,502,412 million players worldwide respectively.

5. Top most Least Popular genres?

- Least Popular Genres

----

steam_games.groupby("Genre")["Estimated_Owners"].sum().sort_values(ascending=True). head(5)

-Findings

- The games based on genres like Photo Editing, Accounting, Video Production, and Web Publishing are not that popular as they have only like 300k to 500k users worldwide.

Cross-verifying the answers using SQL

1. What are the Highest Priced and Lowest Priced Games?

- Highest Price Game

SELECT * FROM STEAM_GAMES_DATA

where price = (select max(price) from steam_games_data);

- Lowest Price Game

```
SELECT * FROM STEAM_GAMES_DATA where price = (select

min(price) from steam_games_data where price <> 0);
```

- Findings

- We have 12409 games that are free of cost and out of 71716 games 59305 games are paid games from which the average Price is around 8.73.

The highest-priced game is Ascent Free Roaming VR Experience with a price tag of 999 Developed and Published by Fury Games for Windows only on 2019-12-27.

The lowest-priced game is IN THE BUILDING: CATS 3 with a price tag of only 0.37, Developed and Published by Laush Dmitriy Sergeevich and Laush Studio Respectively For only Windows on 2023-0105.

2. Which Developer has the most successful Game?

-Developer Name

```
select Developers,sum(estimated_owners) as Users from steam_games_data group by

developers  order by 2 DESC; - Game Name
```

```
select * from steam_games_data where Estimated_owners = ( select max (estimated_owners)

from steam_games_data); - Findings
```

- The Biggest Developer Valve with over 573 million Users Globally. Their DOTA-2 Game is very popular worldwide as it has 150 million users released for all OS, a multiplayer Action game which was released on 2013-07-09.

They are dominating the Top of the Chart as the Amazon games and as well as Krafton Games are also in the top 3 list still they have only around 75 million users each.

3. Which Publisher has the most Game?

- Publisher

```
select top 5 Publisher, sum (Estimated_Owners) as Users fr om steam_games_data group by

Publisher order by 2 DESC;
```

- Findings

- The Biggest Publishers Valve with over 629.87 million Users Globally. And in the Top 5 there are also Ubisoft, Electronic Arts, Sega, and 2k with each having nearly 120-160 million users worldwide.

4. Top Most Popular Genres?

- Top 5 Genres

```
SELECT top 5
```

```
  Genres,
  SUM (CAST (Estimated_Owners AS bigint)) AS Users
FROM steam_games_data
GROUP BY Genres
ORDER BY Users DESC;
```

- Findings

- Most of the games are based on the action genre but there are also adventure and Casual games that are being developed widely. Action games are the most popular with a huge fan base of 4.45 billion gamers at the top of the chart but Adventure, casual, and Indie games are also very popular with 779,502,412 million players worldwide respectively.

5. Top most Least Popular genres?

- Least Popular Genres

```
SELECT top 5 Genres, SUM (CAST (Estimated_Owners AS bigint)) AS Users
FROM steam_games_data
GROUP BY Genres
ORDER BY Users ASC;
```

-Findings

- The games based on genres like Photo Editing, Accounting, Video Production, and Web Publishing are not that popular as they have only 300k to 500k users worldwide.

Suggestions:

- Developers should focus on improving graphics and mechanics rather than being overly concerned with pricing, as negative reviews are not primarily based on price.
- Given that the majority of players are on the Windows platform, prioritize addressing negative feedback specifically from Windows users.
- Action games consistently receive high positive reviews, indicating a strong player preference. Developers should prioritize the development of action games.
- Larger companies like Valve and Krafton account for a significant portion (70%) of negative reviews. They should prioritize understanding player concerns and providing effective solutions.
- A few games dominate the market (e.g., Dota 2, Counter-Strike, Minecraft, PUBG) due to their strong developer support. Steam should support smaller developers by fostering collaboration and development opportunities.
- There is no correlation between negative reviews and the availability of downloadable content (DLC). Developers should focus on enhancing the base game.
- Establishing a platform for enhanced user feedback and addressing user issues will improve game quality.

- Optimizing software to perform well on lower-end devices can expand the player base.
- Sequel games tend to attract more players. Developers should consider creating sequels to capitalize on this trend.

Conclusion:

The gaming market is trending towards future technological advancements, such as the introduction of virtual reality, which is revolutionizing the industry. By prioritizing improvements in-game mechanics, addressing platform-specific issues, developing action games, supporting smaller developers, and optimizing software, Steam can significantly enhance overall user satisfaction and game quality. These efforts will lead to a more positive gaming experience and broader market reach.

The author - Nihar Padhi