BUDT 704: DATA PROCESSING AND ANALYSIS IN PYTHON

"**Exploratory Data Analysis in the Video Game Streaming Industry**"

Data Science Research Report
By

Tanmay Sakharkar
Prachi Patankar
Gautam Bhatia
Gandhali Kokate
Rajat Marathe
Atharva Tilak

INTRODUCTION

The gaming industry has seen a rise in revenue in recent years. As per estimates, the gaming industry is set to take over the film and music industry in the coming years. Many related industries have also shown growth. One such industry is the Online streaming industry. Many new emerging companies such as Twitch, YouTube Gaming, Facebook Gaming, etc have shown great growth potential.

Twitch is an online video streaming platform on which content creators or "streamers" can live stream content which includes a wide variety of categories ranging from video games to cooking.

We have decided to use data from Twitch, specifically from years 2016-2023, to analyze and detect trends over 7 years.

This report aims to analyze the data from Twitch, an online video streaming platform, to find trends and anomalies in the viewership of video games.

We aim to answer the following research questions:
1. What is the total hours streamed for the top 10 games in descending order?
2. What is the average number of viewers for the top 10 Games?
3. Which genre has the highest peak viewers and the corresponding peak viewers and peak viewers ratio?
4. Games with the highest average viewers each month and the number of streamers for these games?
5. Which year had the highest total hours watched and the corresponding total hours watched?
6. Identify the top 5 games with the highest average viewer ratio.
7. Which game had the highest rank across all years?
8. Which games had the highest rank across all years?
9. What was the trend for the average viewership over the years?
10. What are the top 5 games by total hours watched across all years?

BACKGROUND

Twitch as a company was founded in the year 2011. It started as a video streaming platform which mainly focused on live streaming gameplay from video games which users would broadcast on the website which other users could watch. Twitch started growing with the introduction of streaming video game tournaments for audiences worldwide. Over the years, Twitch expanded into diversifying the content available to stream. This included the introduction of "IRL" streams in which users could stream their daily lives in a video log-style format. The popularity of Twitch increased around 2013 and 2014 when big gaming companies started broadcasting their tournaments on

Twitch. Besides the tournament scene, Twitch's influence also grew due to the rise in popularity of its content creators.

Our report deals with the analysis of data from Twitch from 2016 to 2023. This data will help us understand which games helped Twitch grow the most over 8 years. We will also be able to highlight the periods in which viewership was significantly affected by global events and other related phenomena. We would also use this data to predict the future trends in the video game streaming industry based on the popularity of games over the years. We can predict which genre of gaming will have a deep impact on the streaming industry.

METHODOLOGY

By exploring the data, we can highlight meaningful insights by processing various relevant queries and presenting the outcomes through visualizations.

1) Data Procurement
2) Data Cleaning
3) Analyzing the data to find trends
4) Highlighting key insights and recommendations
5) Forecast viewership trends

Forecasting may require procuring additional data related to future games which may or may not be released. The forecasting may not be accurate completely considering the unpredictable nature of the video game industry itself.

DELIVERABLES

Since the video game industry is a rapidly growing market, the streaming industry will also be growing at an exponential rate. To stay with the market trends, we will be delivering a final presentation and project report highlighting key insights related to market trends, category trends and most importantly, viewership analysis. Our analysis will be mainly directed towards providing insights to marketing and operations teams at Twitch who can utilize the findings from this report to understand the viewership trends on their platform.

DATASET SELECTION AND EXPLANATION

[Twitch Data | Kaggle](#)

For analysis, we have taken data from the past 8 years as these were the years when Twitch had respectable competitors in the market. The data was procured from Kaggle.

The Data can be divided into two categories mainly, data related to general viewership on the platform and data related to each particular game over the years.

In the data, we can see the rank of the game in a particular month, total viewership of that particular game for the respective months, peak viewership, peak number of

channels streaming that particular game, total hours streamed in a month, total hours watched by viewers in a month and average viewership for that particular game.

The overall data includes the total number of viewers on the platform for each respective month and year.

DATA DICTIONARY
Rank - Rank of a game in a month based on hours watched
Game - Name of the Game
Month - Months of each year
Year - Year from 2017 to 2023
Hours_watched - Total number of hours watched of a particular game
Hours_streamed - Total number of hours streamed of a particular game
Peak_viewers - Peak viewers of a particular game
Peak_channels - Peak channels streaming a particular game
Streamers - Number of streamers of a particular game
Avg_viewers - Average viewers in a month
Avg_channels - Average channels streaming a game
Avg_viewer_ratio - Ratio of the average number of viewers for each channel

DATA CLEANING
The data procured contained Null Values and the names of games contained special characters and missing values which required adjustment to be understandable and legible for analysis. We started by importing the dataset and checking if values were imported properly. For this analysis, we have only taken the Top 100 ranked games for each month. We start by checking for null values in the data. We replace the special character styling with the appropriate English alphabet.

Specifically, we followed the following steps for cleaning the dataset:

1. Filtering Top 100 Ranks.
2. Dropping Null Values.
3. Checking Special Characters in the 'Game' Column.
4. Replacing 'Pokémon' with 'Pokemon'.
5. Replacing 'Ragnarök' with 'Ragnarok'.
6. Removing 'Just Chatting' from the Data.
7. Removing 'Special Events' from the Data.

DATA ANALYSIS AND MODELING
Business Problem 1 - What are the total hours streamed for the top 10 games in descending order?

To solve this problem, we start by grouping based on games and aggregating by summing the total number of hours streamed for that particular game and displaying the top 10 games.

```
Game
Fortnite                            382065012
League of Legends                   195677443
Apex Legends                        153747049
VALORANT                            140340258
Call of Duty: Warzone               128379086
Minecraft                           107219350
Grand Theft Auto V                  104498883
Counter-Strike: Global Offensive     80709385
World of Warcraft                    79131092
Dead by Daylight                     64976497
```

Interpretation -
Fortnite was streamed the most (382 million hours) on the platform followed by League of Legends (195 million hours).

Business Problem 2 - What is the average number of viewers for the top 10 Games?
To solve this problem, we first group the data based on the game, aggregate the average viewership, and display the top 10 games.

```
Game
League of Legends                   147281.989247
VALORANT                            122537.000000
Fortnite                            100524.733333
Grand Theft Auto V                   87926.107527
Counter-Strike                       72571.250000
Counter-Strike: Global Offensive     66604.696629
Call of Duty: Warzone                66177.060000
Dota 2                               58036.860215
Apex Legends                         57694.446429
Diablo IV                            52237.166667
Name: Avg_viewers, dtype: float64
```

Interpretation -
League of Legends has the highest average viewership.

Business Problem 3 - Which game has the highest peak viewers and the corresponding peak viewers and peak viewers ratio?

We solve this problem by locating the index of the game with the highest peak viewers. The peak viewers ratio is then calculated by dividing the peak viewership for that channel by the average number of channels streaming that game.

The game with the highest peak viewers is League of Legends with 3145012 viewers and a peak viewers ratio of 1221.36

Business Problem 4 - What games with the highest average viewers each month and what is the number of streamers for these games?

We solve this problem by following these steps:

1. Loop through unique years in 'twitch_game_data' and filter data for the current year.
2. Nested loop through unique months in the current year and filter data for the current month.
3. Find rows with maximum average viewers.
4. Append data to lists initialized earlier.
5. Create DataFrame from lists.

	Year	Month	Game	Average Viewers	Number of Streamers
0	2016	1	League of Legends	127021	129172
1	2016	2	League of Legends	134035	117996
2	2016	3	League of Legends	127206	117734
3	2016	4	League of Legends	122933	113251
4	2016	5	League of Legends	108585	118593
5	2016	6	League of Legends	113141	102620
6	2016	7	League of Legends	104807	105041
7	2016	8	League of Legends	109939	105795
8	2016	9	League of Legends	112746	110946
9	2016	10	League of Legends	140383	111071
10	2016	11	League of Legends	106718	103085
11	2016	12	League of Legends	109487	36118
12	2017	1	League of Legends	133690	129423
13	2017	2	League of Legends	148160	114317
14	2017	3	League of Legends	122254	96112
15	2017	4	League of Legends	115233	97370
16	2017	5	League of Legends	118343	121711
17	2017	6	League of Legends	117478	116484
18	2017	7	League of Legends	114044	125411
19	2017	8	PLAYERUNKNOWN'S BATTLEGROUNDS	99339	176534
20	2017	9	League of Legends	103507	118820
21	2017	10	League of Legends	127429	125262
22	2017	11	League of Legends	115323	130486
23	2017	12	League of Legends	92539	137669
24	2018	1	League of Legends	119818	165074
25	2018	2	League of Legends	120888	145155
26	2018	3	Fortnite	159342	741423
27	2018	4	Fortnite	178935	877646
28	2018	5	Fortnite	194619	930299
29	2018	6	Fortnite	176439	847488
30	2018	7	Fortnite	204692	945248
31	2018	8	Fortnite	176899	974227
32	2018	9	Fortnite	147879	900461
33	2018	10	Fortnite	150908	822907
34	2018	11	Fortnite	149921	800661
53	2020	6	League of Legends	199291	215394
54	2020	7	League of Legends	181861	223201
55	2020	8	League of Legends	191859	229635
56	2020	9	Among Us	204634	410661
57	2020	10	League of Legends	230973	246211
58	2020	11	League of Legends	168407	254028
59	2020	12	Fortnite	146288	667652
60	2021	1	Rust	256350	91592
61	2021	2	League of Legends	223876	268439
62	2021	3	Grand Theft Auto V	248422	186339
63	2021	4	Grand Theft Auto V	333016	211141
64	2021	5	Grand Theft Auto V	341075	197818
65	2021	6	Grand Theft Auto V	237277	165667
66	2021	7	Grand Theft Auto V	191842	196492
67	2021	8	Grand Theft Auto V	194561	179725
68	2021	9	Grand Theft Auto V	201009	139075
69	2021	10	League of Legends	280719	185819
70	2021	11	Grand Theft Auto V	180262	123262
71	2021	12	Grand Theft Auto V	157493	148924
72	2022	1	League of Legends	248024	321715
73	2022	2	Lost Ark	217560	196480
74	2022	3	League of Legends	186490	242170
75	2022	4	League of Legends	180413	239721
76	2022	5	League of Legends	214655	237227
77	2022	6	League of Legends	165064	222695
78	2022	7	League of Legends	163239	229473
79	2022	8	League of Legends	179386	222853
80	2022	9	VALORANT	176847	424130
81	2022	10	League of Legends	214867	204092
82	2022	11	Grand Theft Auto V	138326	196976
83	2022	12	Grand Theft Auto V	130478	231641
84	2023	1	League of Legends	170247	231482
85	2023	2	League of Legends	168252	206265
86	2023	3	League of Legends	165663	214209
87	2023	4	Grand Theft Auto V	152483	236273
88	2023	5	League of Legends	168867	198215

35	2018	12	Fortnite	159008	1013029		88	2023	5	League of Legends	168867	198215
36	2019	1	Fortnite	170259	1013016		89	2023	6	Diablo IV	220420	222226
37	2019	2	Apex Legends	183444	663781		90	2023	7	Grand Theft Auto V	189542	295498
38	2019	3	League of Legends	146004	172418		91	2023	8	Grand Theft Auto V	140018	261648
39	2019	4	Fortnite	139698	884315		92	2023	9	Grand Theft Auto V	152697	205932
40	2019	5	Fortnite	144317	846724
41	2019	6	Fortnite	128998	676020
42	2019	7	Fortnite	121815	619076
43	2019	8	Fortnite	119856	596217
44	2019	9	World of Warcraft	124925	126563
45	2019	10	League of Legends	169123	195507
46	2019	11	League of Legends	132735	200004
47	2019	12	League of Legends	105759	211766
48	2020	1	League of Legends	134129	237426
49	2020	2	League of Legends	175015	211492
50	2020	3	League of Legends	166352	252555
51	2020	4	VALORANT	479209	319709
52	2020	5	VALORANT	194859	363116

Interpretation -

Games and Average Viewership:

- League of Legends: This game dominates the leaderboard, appearing multiple times throughout the months with the highest average viewership. In August 2023, it had the highest average viewership with 140,383 viewers.
- Grand Theft Auto V: This game also ranks highly, appearing several times on the leaderboard. Its highest average viewership was in December 2022 with 157,483 viewers.
- Fortnite: This popular battle royale game makes a few appearances on the leaderboard, with its highest average viewership in October 2023 at 102,920 viewers.
- VALORANT: This competitive shooter appeared on the leaderboard in August 2023 with an average viewership of 74,142.
- World of Warcraft: This long-standing MMORPG appears on the leaderboard in October and November 2023 with an average viewership of around 200,000.

Streamers:

The number of streamers for each game is also shown on the leaderboard. While League of Legends has the highest average viewership, it doesn't necessarily have the highest number of streamers. For example, in August 2023, Fortnite had more streamers than League of Legends (66,700 vs. 41,066).

Business Problem 5 - Which year had the highest total hours watched and the corresponding total hours watched?

We solve this problem by following these steps:

1. Group by year and sum hours watched.
2. Find the year with the maximum total hours watched.
3. Find the maximum total hours watched value.

```
The year with the highest total hours watched is: 2021 with a total of 17653847814 hours
```

Business Problem 6 - What are the top 5 games with the highest average viewer ratio?

For this problem, we use the 'nlargest' method to select the top 5 rows from the twitch_game_data DataFrame based on the values in the 'Avg_viewer_ratio' column. It sorts the DataFrame in descending order based on the 'Avg_viewer_ratio' and then selects the top 5 rows.

```
                                   Game  Avg_viewer_ratio
6039                  Doctor Who Pinball           4586.72
5831                  Doctor Who Pinball           4070.44
4482  Penn & Teller's Smoke and Mirrors           3695.19
2695              IMDb Live Viewing Party           3607.03
4436                      Twitch Presents           3418.93
```

Business Problem 7 - What game had the highest rank across all years?

We use the 'idxmin' method to find the index of the row with the minimum value in the 'Rank' column of the twitch_game_data DataFrame. The loc method is then used to retrieve the entire row based on this index.

After identifying the row with the minimum rank, we extract the value in the 'Game' column from that row.

```
The game with the highest rank across all years is: League of Legends
```

Business Problem 8 - What games had the highest rank across all years?

We group the twitch_game_data DataFrame by the 'Year' column. For each group (year), the apply method is used with a lambda function to find the row with the minimum value in the 'Rank' column using idxmin(). The result is a DataFrame with one row per year, representing the game with the highest rank for each year.
After identifying the top-ranked game for each year, we specify that only the 'Rank' and 'Game' columns should be included in the final result.

```
          Rank                 Game
Year
2016        1    League of Legends
2017        1    League of Legends
2018        1    League of Legends
2019        1             Fortnite
2020        1    League of Legends
2021        2                 Rust
2022        2    League of Legends
2023        2    League of Legends
```

Business Problem 9 - Descending order of average viewership for months

We group the twitch_game_data DataFrame by the 'Month' column and calculate the mean (average) value of 'Avg_viewers' for each month. After that, we sort the results in descending order of average value.

```
Month
4      13753.087231
5      13421.506329
2      13072.107459
8      12998.274684
6      12952.726582
1      12803.581013
3      12703.587863
9      12665.074778
10     12623.210145
7      12538.484197
11     12288.063768
12     11815.305797
```

Business Problem 10 - What is the trend for average viewership over time?

We generate a line chart to visualize the trend of average viewership over time using the matplotlib library.
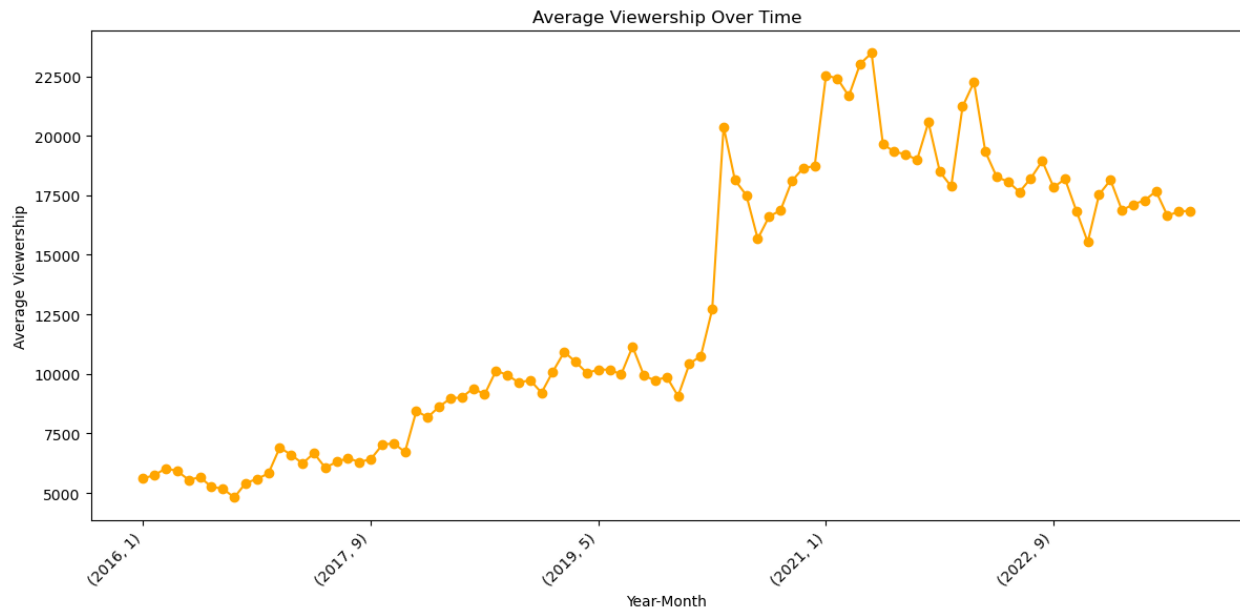
Interpretation -

- Overall upward trend: The average viewership for all Twitch games has generally increased over time. The line slopes upward from 2016 to 2023, indicating a steady growth in the average number of viewers watching Twitch streams.
- Fluctuations within the trend: While the overall trend is upward, there are also noticeable fluctuations within the data. There are periods of both growth and decline in average viewership, suggesting that the growth is not always linear.
- Seasonal patterns: There appears to be a seasonal pattern in the data, with peaks in viewership during the summer months (June-August) and dips in the winter months (December-February). This could be due to factors like school vacations, holidays, and lockdowns during the COVID-19 pandemic (2020-2021).
- Specific events: The data might also be influenced by specific events that generate excitement and viewership around certain games or streamers. For example, the release of a major new game or a popular esports tournament could lead to a temporary spike in average viewership.

Business Problem 11 - What are the top games with the highest peak viewership

We generate a bar chart to visualize the top games with the highest peak viewership. This line uses the 'nlargest' method to select the top 10 rows from the twitch_game_data DataFrame based on the values in the 'Peak_viewers' column. It sorts the DataFrame in descending order based on the 'Peak_viewers' and then selects the top 10 rows.
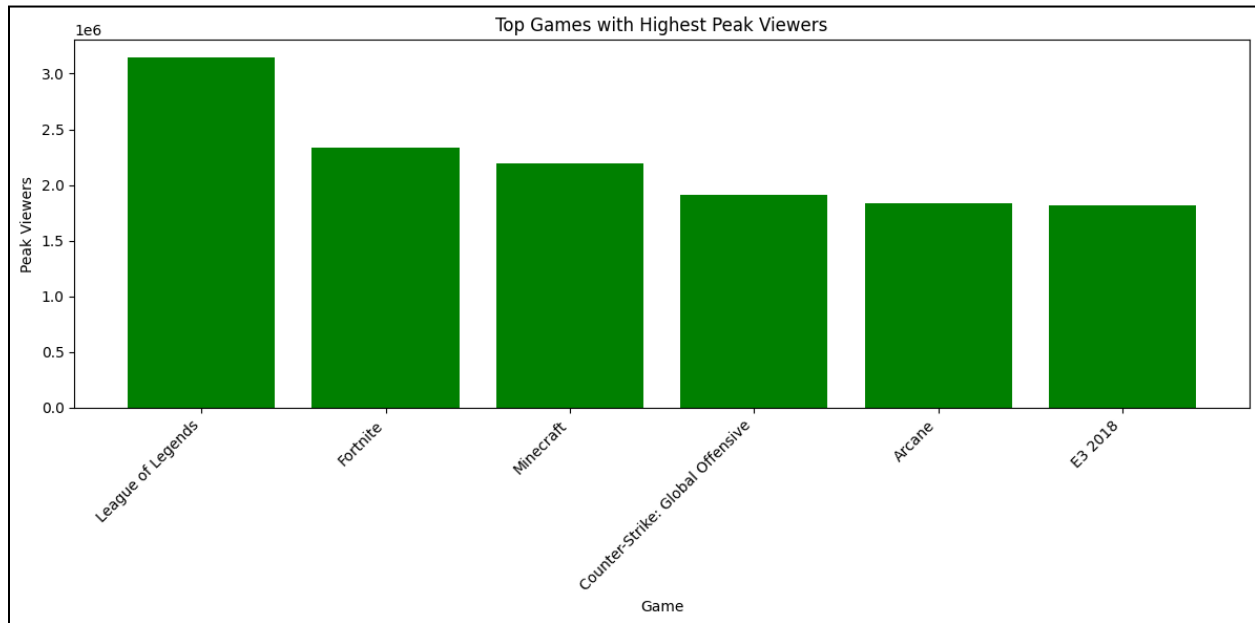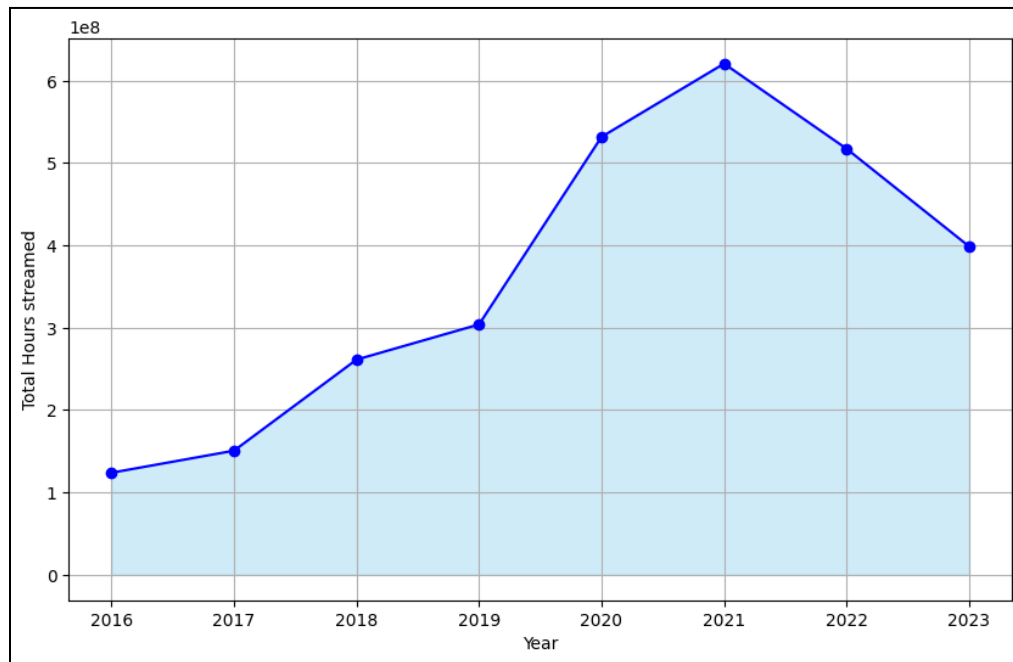

Interpretation -

● League of Legends dominates the list, with a peak viewership of over 3 million viewers. This is likely due to its large and dedicated fanbase, as well as its popularity in the esports scene.
● Fortnite and Minecraft are both popular battle royale and sandbox games, respectively, and their high peak viewership numbers reflect their broad appeal to a large audience.
● Counter-Strike: Global Offensive is a long-standing competitive shooter with a passionate community, and its peak viewership numbers demonstrate its continued popularity on Twitch.
● It's worth noting that peak viewership can be influenced by various factors beyond the game itself, such as streamer popularity, event hype, and platform-wide promotions.

Business Problem 12 - What is the trend of total hours streamed over time?

An area plot to visualize the total hours streamed over the years.
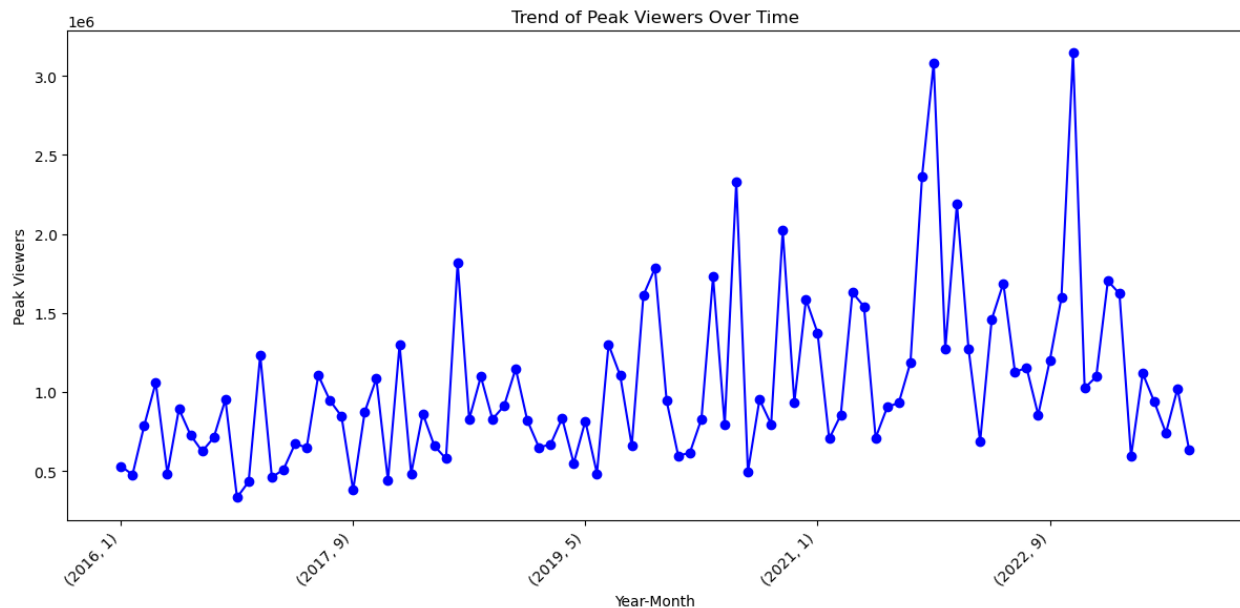

Interpretation -

- Growing popularity of Twitch: The increasing popularity of Twitch as a platform for gaming, entertainment, and social interaction is likely a major driver of the growth in total hours streamed.
- More streamers: The number of streamers on Twitch has also been steadily increasing, which naturally contributes to the overall amount of content being streamed.
- Longer streams: Individual streamers may also be broadcasting for longer durations on average, further contributing to the rise in total hours streamed.

Business Problem 13 - What is the trend of peak viewership over time?

A line chart to visualize the trend of peak viewership over time.

Interpretation -

- High peak viewership: Both 2021 and 2022 stand out as having some of the highest peak viewership numbers in the entire dataset.
- Possible influence of COVID-19: The timing of these high peaks coincides with the COVID-19 pandemic, which led to increased lockdowns and stay-at-home restrictions in many parts of the world. This could have contributed to a rise in Twitch viewership as people sought entertainment and social interaction during periods of isolation.

   Other possible reasons for high peak viewership:

- Major events: The peaks in the data likely coincide with major events or milestones on Twitch, such as popular esports tournaments, streamer collaborations, or platform-wide promotions. These events can generate temporary spikes in viewership but may not be sustainable in the long term.
- Game popularity fluctuations: The popularity of certain games can also influence peak viewership. Games with large and dedicated fanbases or those experiencing periods of high activity (e.g., new releases, major updates) may contribute to peaks in viewership.

● Streamer viewership variation: Individual streamer viewership can also fluctuate significantly over time. Streamers with large followings may not always achieve the same peak viewership numbers consistently, impacting the overall trend.

Business Problem 14 - What is the distribution of total hours watched by year?
A pie chart to illustrate the distribution of total hours watched by year.

Distribution of Total Hours Watched by Year
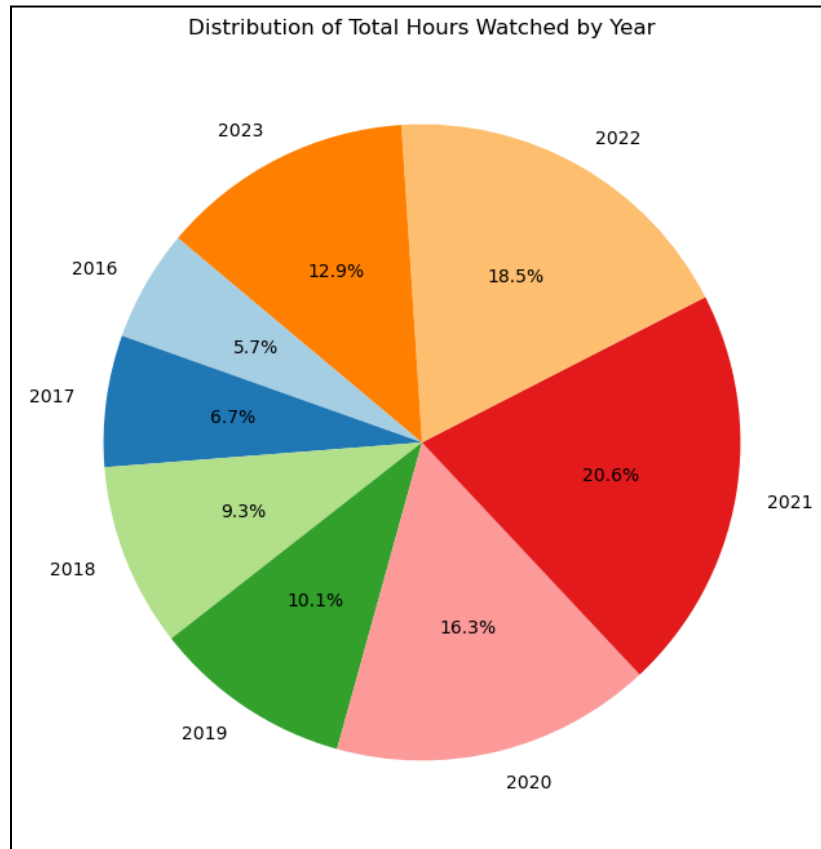
Interpretation -

● The pie chart clearly shows that total hours watched on Twitch has increased steadily over the years, with the most recent years contributing the largest share.
● The prominent slices for 2022 and 2023 highlight the significant growth in viewership.
● This trend suggests that Twitch is becoming increasingly popular and attracting a larger audience over time.

Business Problem 15 - What is the correlation between different columns?

Visualization of the correlation matrix of numeric variables in the twitch_game_data DataFrame using a heatmap. The color intensity in the heatmap represents the strength and direction of the correlation between pairs of numeric variables. The annotations provide the actual correlation coefficients for interpretation.
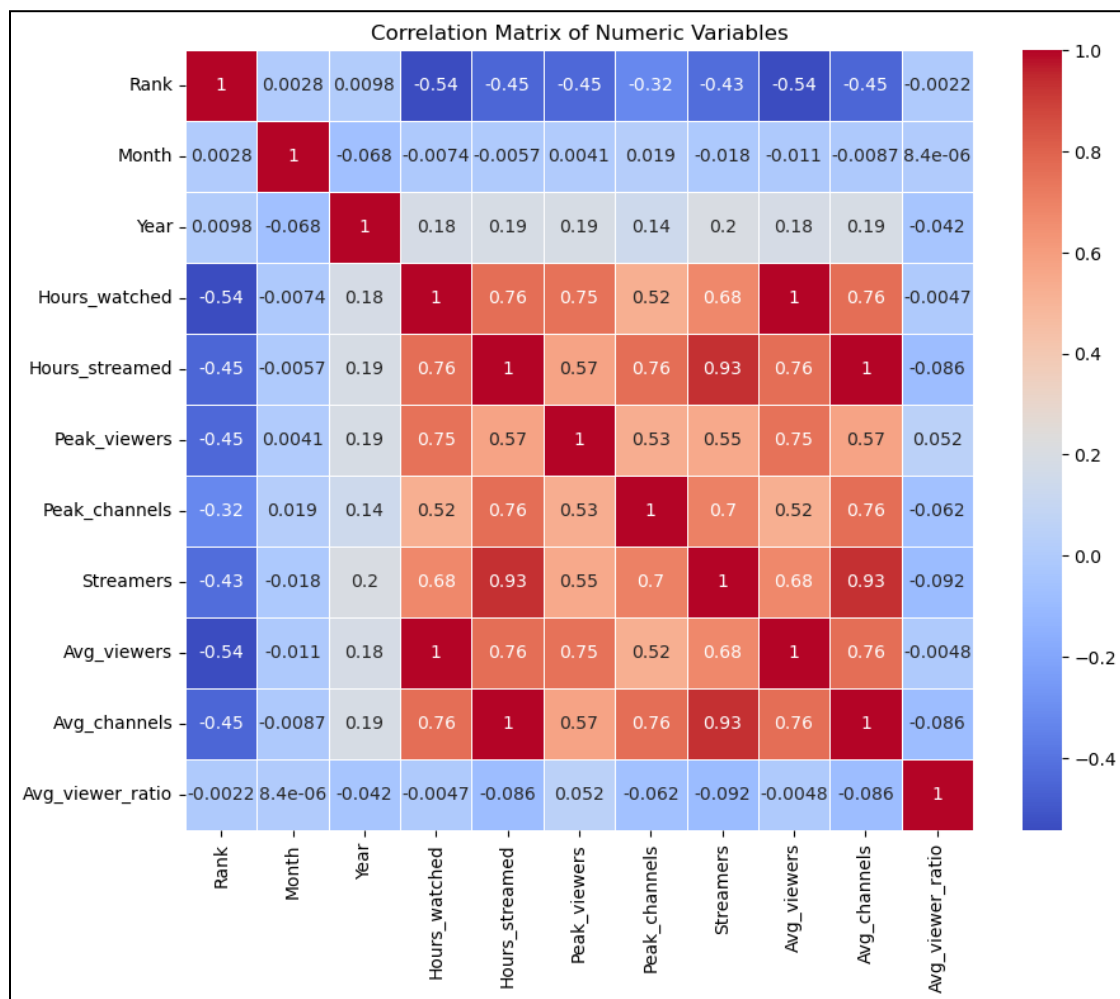
Interpretation -

Strong Positive Correlations:

- Hours_watched & Hours_streamed (0.76): This indicates a strong positive correlation, meaning that games with higher average viewership hours tend to have more hours streamed overall. This suggests that viewers are engaged with games they watch for longer periods, leading to higher total stream durations.
- Hours_watched & Peak viewers (0.75): This is another strong positive correlation, implying that games with higher average viewership hours also tend

to have higher peak viewership numbers. This suggests that longer viewership durations can lead to larger audiences tuning in at peak times.
- Peak viewers & Peak_channels (0.53): This moderate-to-strong positive correlation indicates that games with higher peak viewership numbers tend to have more concurrent channels broadcasting them. This suggests that popular games attract more streamers vying for a share of the large audience.

Moderate Positive Correlations:

- Rank & Hours_watched (0.54): This moderate positive correlation implies that games ranked higher in terms of average viewership tend to have more hours watched overall. This suggests that higher-ranked games are more popular and attract larger audiences, leading to longer watch times.
- Rank & Peak viewers (0.45): This moderate positive correlation indicates that higher-ranked games also tend to have higher peak viewership numbers. This reinforces the notion that popularity translates to larger audiences at peak times.
- Hours_watched & Avg_viewers (0.68): This strong positive correlation suggests that games with higher average viewership hours tend to have higher average viewers per stream. This implies that viewers stick around longer for streams of games they enjoy, leading to higher average audience sizes.

Negative Correlations:

- Rank & Peak_channels (-0.32): This moderate negative correlation implies that games ranked higher in terms of average viewership tend to have fewer concurrent channels at peak viewership times. This could be due to several reasons, such as:
- Domination by large streamers: A few large streamers might dominate the viewership for these games, leaving less room for smaller channels to compete at peak times.
- Genre specialization: Games in specific genres might attract dedicated communities with fewer overall streamers, leading to lower channel counts despite high viewership.
- Rank & Avg_channels (-0.45): This moderate negative correlation reinforces the previous point, suggesting that higher-ranked games also tend to have fewer channels broadcasting them on average. This further indicates potential dominance by large streamers or genre specialization.

Business Problem 16 - What is the distribution of total hours watched by the top 5 games?

A pie chart that visualizes the distribution of total hours watched by the top 5 games in the twitch_game_data DataFrame. Each wedge in the pie represents a game, and the size of each wedge corresponds to the proportion of total hours watched for that game.

Distribution of Total Hours Watched by Top 5 Games
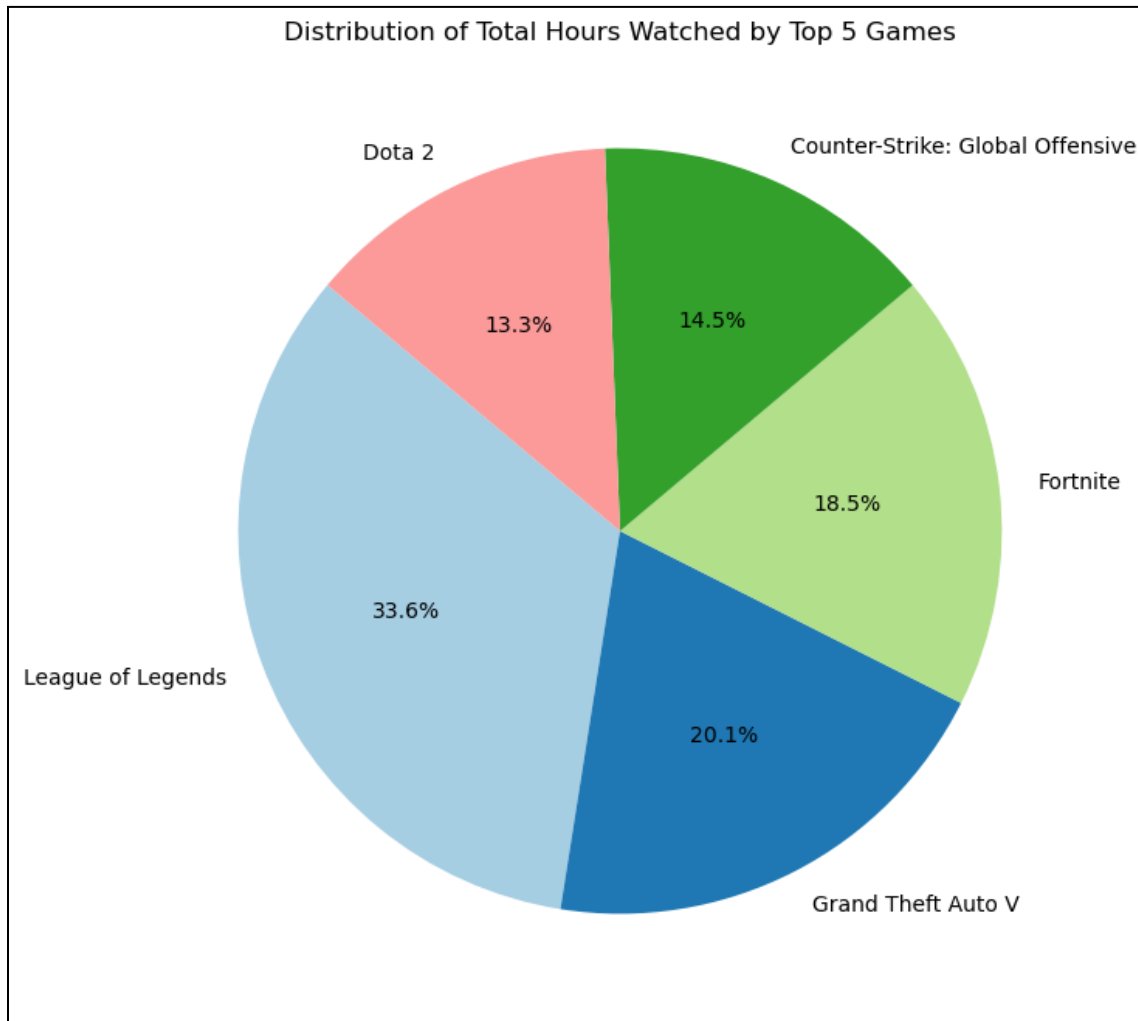
Interpretation -

- League of Legends stands out with the largest slice, representing 33.6% of the total hours watched. This indicates that League of Legends is by far the most popular game among the top 5, attracting significantly more viewership than the other games.
- The remaining four games, Counter-Strike: Global Offensive (15.3%), Fortnite (18.5%), Grand Theft Auto V (20.1%), and Dota 2 (13.3%), share the remaining portion of the total hours watched. Their slice sizes suggest a range of popularity

within the top 5, with Grand Theft Auto V and Fortnite attracting slightly more viewership than Counter-Strike and Dota 2.

Model Building:

We implement a linear regression model using the scikit-learn library in Python to predict the number of 'Streams' in the future. Here's a breakdown of our implementation along with the plot:

1. Read Data and Define Features (X) and Target Variable (y):
   - We use the dataset (Twitch_global_data.csv) and read it into a Pandas DataFrame (twitch_global_data).
   - We separate the features (X) by excluding the 'Streams' column, and the target variable (y) is set to the 'Streams' column.

2. Split the Data into Training and Testing Sets:
   80% of the data is used for training (X_train, y_train), and 20% is used for testing (X_test, y_test).

3. Model Creation, Training and Prediction:
   - A linear regression model is created using the LinearRegression class from scikit-learn.
   - The model is trained on the training data using the fit method.
   - Predictions are made on the test set using the trained model (predict method).

4. Evaluate the Model:
   Model performance is evaluated using mean squared error (mse), root mean squared error (rmse), and R-squared (r2) scores.

```
Mean Squared Error: 2372684372989.27
Root Mean Squared Error: 1540352.03
R-squared: 0.94
Coefficients: [ 5.17654810e+05 -1.11130812e+05  6.29552605e-03 -3.19579534e-01
 -1.53366957e-02 -2.15795772e+01  1.82447957e+02 -7.56651905e+05]
Intercept: -1018980274.5847639
```

Interpretation:
   - The low values of MSE and RMSE indicate that the model has relatively small prediction errors.

- The high R-squared value (close to 1) suggests that the model explains a large proportion of the variance in the target variable.
- The coefficients provide insights into the impact of each feature on the target variable. Positive coefficients indicate a positive relationship, while negative coefficients indicate a negative relationship.
- The intercept value is meaningful in the context of the specific data and features used in the model.

Overall, the model appears to perform well based on the provided evaluation metrics

5. Create a Pair Plot with Regression Line:

- Features and the target variable are combined into a single DataFrame (data).
- A pair plot is created using Seaborn, including scatter plots for each pair of features and a regression line. The diagonal plots show kernel density estimates.
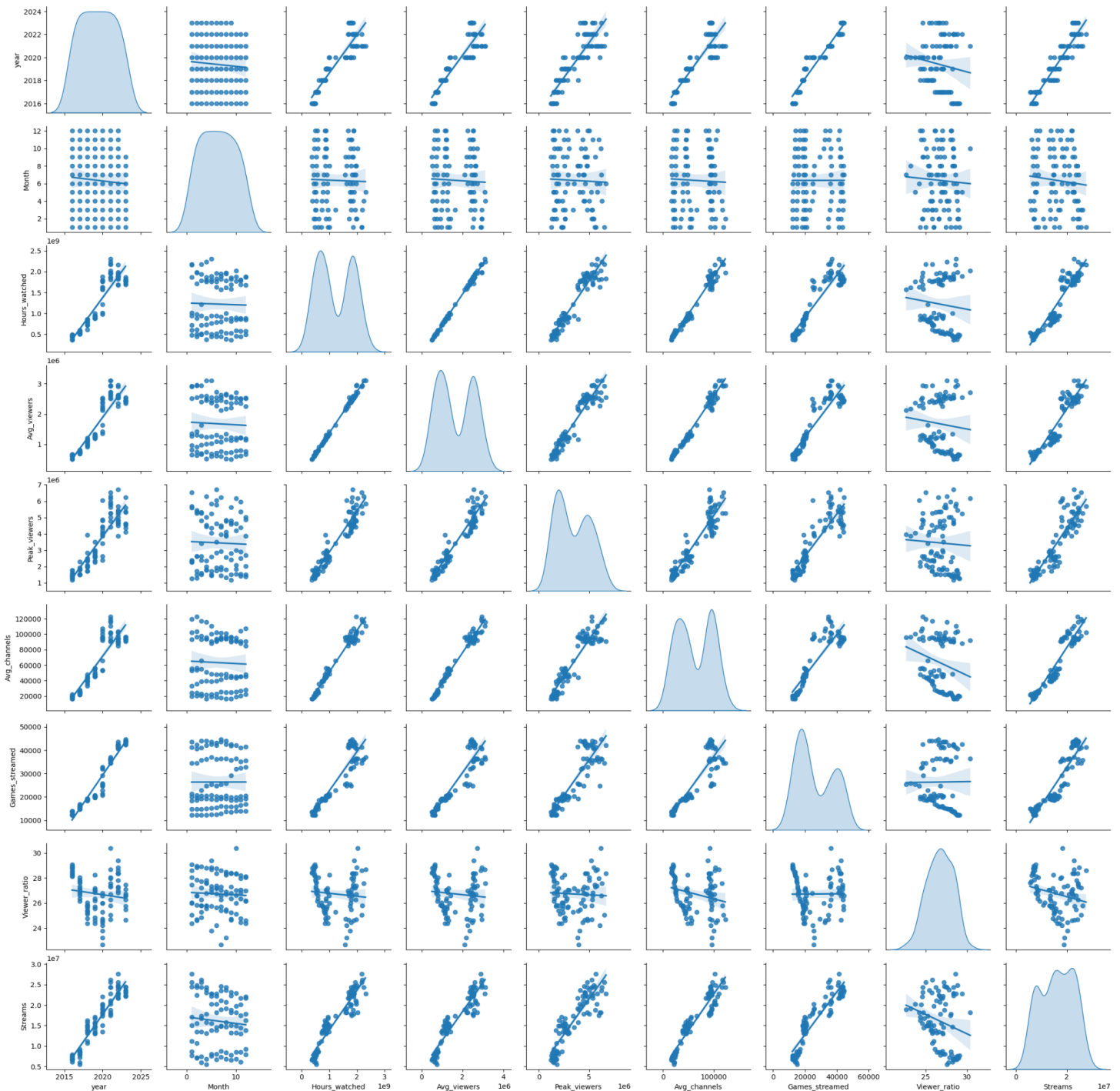
STRATEGY RECOMMENDATIONS
Based on the analysis, we came up with the following recommendations for Twitch.

1) League of Legends and Fortnite seem to be the dominating games on the platform and Twitch should focus on building relationships with the respective parent companies for these games.

2) Twitch experienced a significant increase in its total viewership around 2021 and has been growing steadily, they should focus on improving their results to increase revenues.
3) The total viewership has started to decline from 2021 to 2023 so Twitch should focus on its efforts to increase viewership.

DATA LIMITATIONS AND NEXT STEPS

The data is not categorized enough to perform clustering and is not meaningful enough to run a regression analysis to predict viewership trends in the coming years, although we did manage to create a model to predict the number of streams. It is recommended to collect more descriptive data to perform a better regression analysis in the future.

The dataset also included many Null values which were eliminated to perform meaningful analysis. Twitch is recommended to keep its data clean and not report any missing values in the future to prevent any skewness in future reports.

REFERENCES

1) Statista. (n.d.). *Games live streaming - Worldwide | Statista market Forecast.*

   https://www.statista.com/outlook/dmo/digital-media/video-games/games-live-streaming/

   worldwide

2) *Game Streaming Market Insights.* (n.d.).

   https://www.mordorintelligence.com/industry-reports/game-streaming-market