T1-24-25 DAS 732-Data Visualization Assignment 3: Twitch Dataset Analysis Using Visual Analytics Framework

Udayagiri Narayana Srimanth
IMT2022052
Udayagiri.Srimanth@iiitb.ac.in

Khadarabad Tahir Mohammed
IMT2022100
Khadarabad.Mohammed@iiitb.ac.in

Pradyun Devarakonda
IMT2022525
Devarakonda.Pradyun@iiitb.ac.in

***Index Terms*—Twitch, Streamers, Views, Followers, Games, Game Genres, Watch Time, Stream Time**

I. Introduction

A. Methodology

In this study, we analyze the "Top 1000 Streamers on Twitch" dataset, encompassing data from August 2019 to August 2020, using a workflow derived from the Visual Analytics framework proposed by Keim et al. (2008) [3]. This framework integrates data preprocessing, visualization, and interactive analysis to derive actionable insights effectively. The methodology begins with data cleaning, normalization, and filtering to ensure accuracy and relevance.

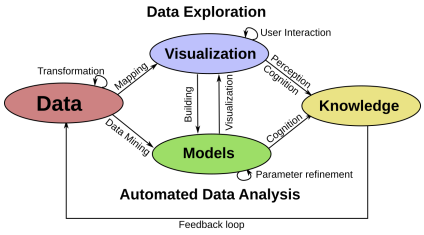

Figure 1: The visual analytics workflow, based on Keim et al. [3] Image Courtesy: Wikimedia Commons

B. Datasets

Top 1000 Streamers on Twitch (August 2019 - 2020):

Our core dataset has analytics data of the Twitch's top 1000 streamers (based on the number of followers) from August 2020. It has a total of 11 columns.

1) Channel Name: Display name of the channel
2) Watch Time(Minutes): Total watch time on the channel in that particular year
3) Stream time(minutes): Total time streamed on the channel in that particular year
4) Peak viewers: Max viewers reached by a stream on the channel in that particular year
5) Average viewers: Average viewers across streams on the channel in that particular year
6) Followers: The number of followers for each channel
7) Followers gained: The number of followers gained for each channel in that particular year
8) Views gained: The number of views gained for each channel in that particular year
9) Partnered: Whether the channel is a Twitch Partner or not
10) Mature: Whether the channel is marked as mature content or not
11) Language: The primary language in which the streamer's content is broadcast.

Of the 11 columns, 2 columns (Channel Name and Language) are of string data type, 2 columns (Partnered and Mature) are of boolean data type, and the remaining 7 columns are of integer data type.

Top Games on Twitch (2016 - 2023):

Our complementary dataset analytics data of the

top 200 games on twitch on each month from 2016 to 2023. It has a total of 12 columns.

1) Rank: Rank in the month
2) Game: Name of the game or category
3) Month: Month in question
4) Year: Year in question
5) Hours_watched: Hours watched on Twitch
6) Hours_streamed: Hours streamed on Twitch
7) Peak_viewers: Maximum viewers at one instant
8) Peak_channels: Maximum channels at one instant
9) Streamers: Amount of streamers who streamed the game
10) Avg_viewers: Average amount of viewers
11) Avg_channels: Average amount of channels
12) Avg_viewer_ratio: Average amount of viewers per channel

Of the 12 columns, 1 column (Game) is of string data type, 1 column (Avg_viewer_ratio) is of float data type, and the remaining 10 columns are of integer data type.

II. Data Manipulation and Inferences

A. Prior Visualizations and Inferences

This section summarizes the visualizations and key inferences derived from Assignment-1. These insights serve as foundational observations for the analyses in Assignment-3 and are presented as an appendix for reference.

Visualization Methodology in A1:

1) **Exploratory Analysis of Dataset Features** Initial exploration focused on understanding the dataset's composition by examining the distribution of numerical values across integer columns and analyzing categorical columns to determine the proportion of streamers across various categories. This provided a comprehensive overview of the dataset's structure, enabling us to identify trends and anomalies within the data.
2) **Trend Identification via Regression Analysis** Line and scatter plots were created for numerical variables, often complemented with regression lines to uncover correlations or trends between two or more variables. This visual approach helped identify patterns, such as relationships between follower count, stream duration, and viewer retention.
3) **Category-Wise Numerical Analysis** Numerical values were further analyzed within categorical groupings, such as streamer language or genre preferences, to assess audience behavior and content trends. This analysis offered insights into the underlying structure and dynamics of the data related to audience engagement and streamer activity.
4) **Hypothesis Formation and Explanatory Insights** Based on observed patterns, hypotheses were proposed to explain specific behaviors or data trends. These explanations were supported by evidence from the visualizations, enriching our understanding of the dataset and highlighting the interplay between various factors influencing Twitch streamers and their audiences.

Relevant Inferences from A1:

1) **Follower Counts and Viewer Engagement** Streamers with higher follower counts exhibited consistently higher average watch times and viewer counts. This trend likely stems from a combination of factors, including viewer loyalty, which ensures consistent attendance, and the engaging nature of content produced by these streamers. Larger follower bases may also attract additional viewers through recommendations or social proof, further amplifying these metrics.
2) **Stream Time and Audience Retention** Increased stream time does not necessarily lead to higher follower counts or average viewership. This finding suggests that while prolonged streaming might offer greater exposure, it does not guarantee audience growth. External factors such as content type, time zones, and audience retention appear to play more significant roles. However, stream time does correlate with a steady increase in total watch time, indicating that audience engagement remains relatively stable regardless of session length.
3) **Maturity and Audience Evolution** For

streamers with substantial follower counts, content maturity becomes less of a determinant for growth. While maintaining a family-friendly image might be beneficial during the initial growth phase, loyal audiences tend to remain engaged even if the streamer transitions to mature content over time. This underscores the importance of cultivating a strong community rather than solely focusing on content restrictions.

4) **Diminishing Follower Growth** As streamers accumulate larger audiences, follower growth rates tend to slow. While the general trend shows a downward trajectory, data reveals that growth stagnation can occur at any follower count, regardless of the streamer's size or popularity. This suggests a saturation effect, where streamers must employ novel strategies to sustain growth beyond certain thresholds.
5) **Viewer Behavior and Watch Time** Channels with larger follower counts do not necessarily see significantly higher per-viewer watch times. Larger channels often cater to broader, more transient audiences, leading to shorter engagement durations per viewer. In contrast, smaller channels may cultivate highly engaged, loyal communities that contribute to longer watch times. This highlights the potential value of niche content and personalized engagement for smaller streamers.
6) **Stream Duration and Viewership** Merely increasing the number of hours streamed does not directly correlate with increased viewership. This observation underscores the limitations of relying solely on stream duration as a growth strategy. Instead, factors such as content quality, audience engagement, and promotional efforts play more critical roles in attracting and retaining viewers. This insight emphasizes the importance of strategic planning over sheer volume.
7) **Language and Viewer Dynamics** Viewer engagement on Twitch varies significantly based on the language of the stream. Language shapes audience preferences, engagement levels, and community dynamics, reflecting cultural and regional factors. Streams in non-English languages often attract highly localized, loyal audiences, while English-language streams tend to reach more diverse and transient viewers. This finding highlights the importance of linguistic and cultural considerations when strategizing audience growth.

B. Data Preprocessing

The dataset used for this analysis was of high quality, requiring minimal preprocessing. It contained no missing or null values, and no corrupted entries were identified, ensuring the integrity of the data for subsequent analyses.

During an exploratory review, we identified 44 channels that did not represent individual streamers but were instead associated with Esports leagues, tournaments, or organizations. To ensure the relevance and consistency of our analysis, we categorized channels into two distinct groups:

- **Personalities**: Individual streamers who broadcast their content on Twitch.
- **Esports**: Channels representing Esports tournaments, leagues, or game-specific organizations.

The channel *Fextralife* was excluded from both categories. This decision aligns with findings that the disproportionately high watch times reported for this channel likely result from embedded streams on external websites rather than organic audience engagement on Twitch. This exclusion is consistent with the methodology employed in Assignment-1, where *Fextralife* was also omitted from visualizations to prevent distortion of results.

This categorization ensured that our analysis focused on meaningful comparisons between individual streamers and Esports-related entities, while outliers like *Fextralife* were systematically excluded to maintain the validity and reliability of our findings.

C. Visual Analytics Workflow

This is the visual analytics workflow for sections II, III and IV.

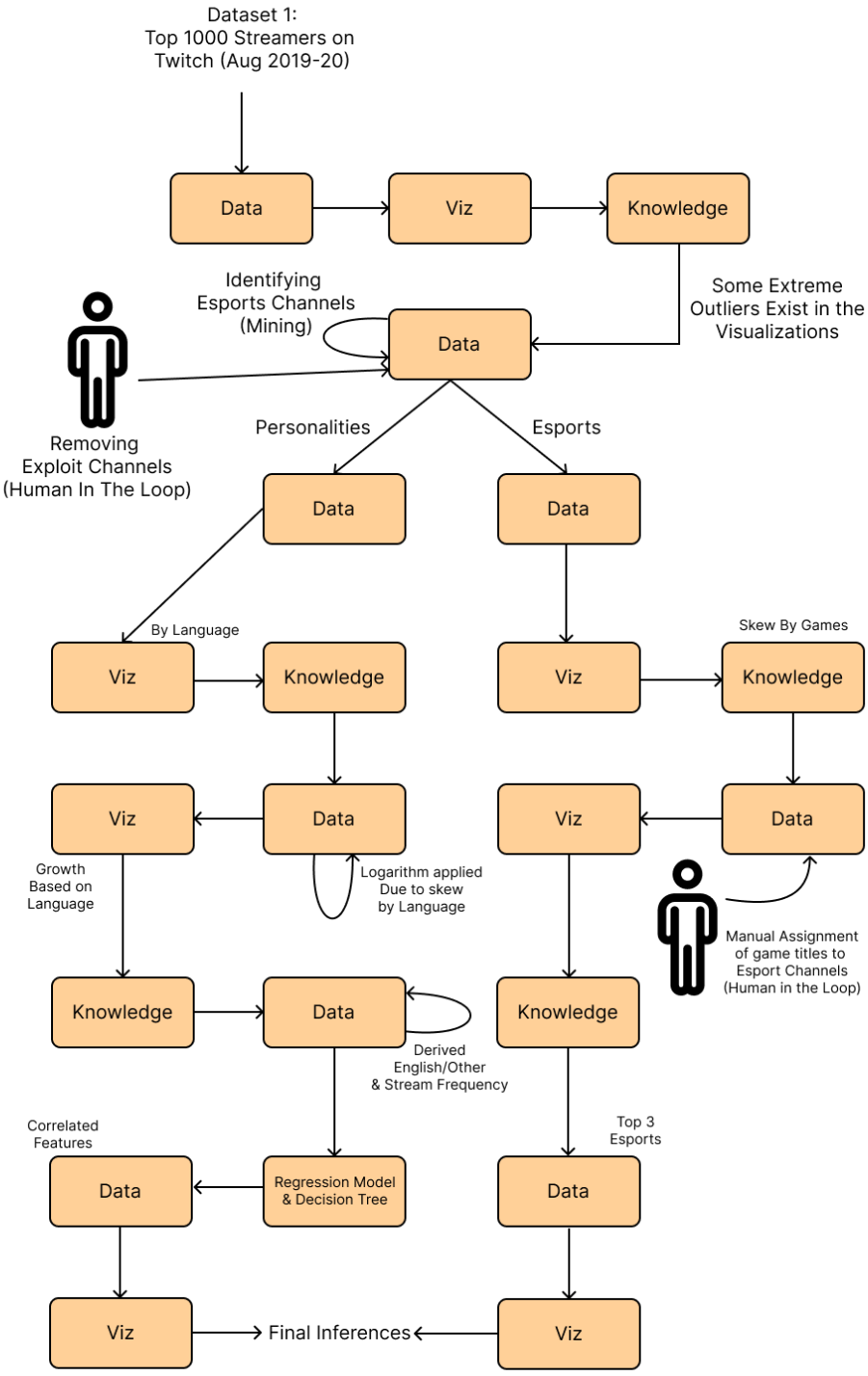

Figure 2: Visual Analytics Workflow for Sections II, III and IV

III. Streaming Personality Analysis

In this section, we examine the **Personalities** category, focusing on how language influences the behavior and reach of streamers on Twitch. By analyzing these factors mathematically, we aim to validate or refine the hypotheses and inferences made in Assignment-1 and explore new insights that further build on the findings from previous analyses. This approach continues the flow of the data pipeline by incorporating the insights derived from different models.

Twitch, being a global platform, exhibits diverse language distribution among its user base. However, it is well-established that **English** dominates in terms of both users and content. According to several sources that track web traffic by region, such as Backlinko [1] and SimilarWeb [6], the largest share of Twitch users (over 20%) are based in the United States. This figure does not account for other English-speaking countries, including Canada, the United Kingdom, Australia, and several others, which further amplify the dominance of English-speaking users on the platform.

To contextualize this dominance, we refer to Figure 3, which presents data from Google Trends showing the global interest in the search term "Twitch" from August 2019 to 2020. From the visualization, it is evident that Twitch enjoys significant engagement in regions such as Portugal, Spain, Finland, Denmark, and Argentina. These regions, though smaller in comparison to English-speaking countries, exhibited notable interest in the platform, indicating a growing and diverse audience.

Further analysis of language distribution on Twitch is provided by the word clouds shown in Figures 4 and 5. The first word cloud visualizes the number of channels per language, highlighting the prevalence of English-language streams. While the English language clearly dominates, we observe significant participation from other languages, particularly **Spanish** and **Portuguese**. These languages, while smaller in number, are seeing an increase in both streamers and viewers, indicating strong momentum for growth in these regions.

The second word cloud shows the number of followers gained by channels in different languages.

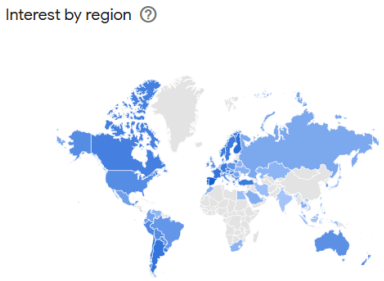

Figure 3: Interest by Region, for the search term "twitch" from August 2019 to 2020. Image Courtesy: Google Trends

This visualization reinforces the trend seen in the previous word cloud, where English-language channels lead, but Spanish and Portuguese channels exhibit promising growth. This is consistent with the data observed in other regions, suggesting that non-English-speaking audiences are becoming increasingly engaged on Twitch.

Figure 4: Word Cloud showing Language by Number of Channels

Figure 5: Word Cloud showing Language by Number of Followers Gained

Figure 6, examines the **Maximum Peak Viewership by Language**, offering insights into the

audience engagement on Twitch based on language. By observing this bar chart, we can infer the scale of audience participation during major Twitch events. English-language streams maintain the highest peak viewership, which is consistent with the large user base in English-speaking countries. However, the increasing participation of Spanish and Portuguese streams during peak events suggests an expanding global audience for these languages, which may indicate the growing influence of non-English content on the platform.

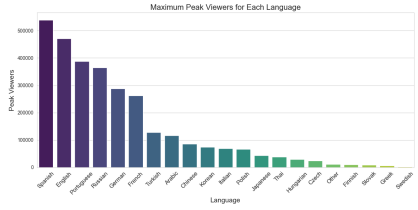

Figure 6: Bar Chart showing Maximum Peak Viewership by Language

Events Behind the Top Three Peaks in Viewership by Language from August 2019 - 2020

***1. Shroud (English)**:*

- **Peak Viewers**: 471,281
- **Event**: *Shroud's Return to Twitch*
- After Mixer shut down in mid-2020, Shroud returned to Twitch on August 12, 2020. His first live stream on Twitch after the hiatus was highly anticipated, leading to a massive influx of viewers.
- Shroud is a popular gaming personality known for his gameplay in titles like *Valorant*, *PUBG*, and *Counter-Strike: Global Offensive*, which further fueled the excitement.

***2. Gaules (Portuguese)**:*

- **Peak Viewers**: 387,315
- **Event**: *CS:GO Major Streams and Regional Popularity*
- Gaules is a Brazilian streamer renowned for broadcasting *Counter-Strike: Global Offensive* matches. His peak occurred during a major CS:GO tournament in 2020, such as the ESL One: Road to Rio.
- He provided Portuguese-language commentary, appealing to Brazil's massive gaming audience. His unique style and engaging community also contributed to the high viewership.

***3. TheGrefg (Spanish)**:*

- **Peak Viewers**: 538,444
- **Event**: *Fortnite and Announcement Hype*
- TheGrefg is a prominent Fortnite streamer and content creator. His peak in 2020 is linked to his announcement of receiving a Fortnite Icon Series skin, which was highly anticipated within the community.
- Fortnite remains immensely popular in Spanish-speaking countries, and TheGrefg's influence as a creator further amplified the event.

Each of these events highlights a combination of individual streamer influence, game-specific excitement, and major platform announcements or tournaments that drove record-breaking viewership in 2020.

Through the analysis of language distribution, viewer engagement, and peak viewership, we can observe clear trends in the geographical and linguistic shifts in Twitch's audience. English-language content continues to dominate, but other languages such as Spanish and Portuguese are gaining significant traction, both in terms of channels and followers gained. These insights offer valuable implications for streamers, marketers, and content creators, as they can tailor their strategies to target emerging audiences in different linguistic regions.

What Streamers Should Aim for Growth of Certain Metrics

With an understanding of audience demographics and their shifting trends, it is possible to identify which features correlate with other statistics or growth metrics. This section explores such correlations and presents insights derived from visualizations and statistical models, providing actionable information for streamers seeking to optimize their strategies.

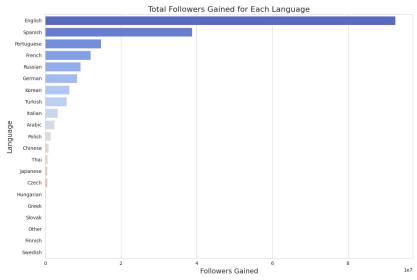

Figure 7: Bar Chart showing Number of Followers Gained by Language

Figure 7 represents the number of followers gained by language. While the dominance of English-language channels is evident, the large disparity in follower counts makes it challenging to interpret proportions for other languages effectively. To address this issue, we applied a logarithmic transformation to the data, allowing for better representation and analysis of follower growth across languages with smaller but emerging communities.

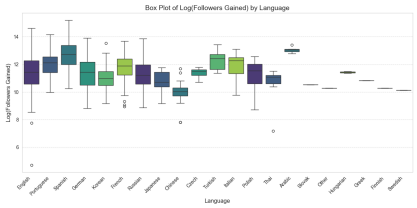

Figure 8: Box Plot of Logarithm of Followers Gained by Language

Figure 8 presents a box plot of the logarithm of followers gained by language. This visualization reveals that many non-English languages, including Spanish, Portuguese, and Arabic, demonstrate a higher median logarithmic growth in followers compared to their overall follower counts. This indicates significant growth within smaller linguistic communities. While English-language channels gained the most followers overall, the distribution is spread across a much larger number of streamers, with considerable variance in individual follower growth. In contrast, the more connected nature of these smaller linguistic communities allows channels to achieve consistent and meaningful follower increases.

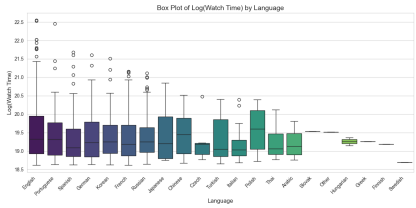

Figure 9: Box Plot of Logarithm of Aggregate Watch Time by Language

The box plot in Figure 9 illustrates the logarithm of aggregate watch time by language. Interestingly, the distribution of watch times for English-language channels is similar to that of other languages, suggesting that the average streamer, irrespective of language, engages audiences for comparable durations. This could reflect global viewing patterns, where audiences spend similar amounts of time on Twitch regardless of their geographic or linguistic background. Alternatively, it may indicate a standard quality of content across streamers in these languages.

However, the presence of significant outliers (represented by circles above the fourth quartile) in the English-language distribution skews the overall watch time. These outliers are typically top-tier streamers with substantially higher engagement metrics. Factors contributing to this disparity may include:

- The size of the audience pool for English-language content, which is inherently larger.
- The loyalty of viewers, who are more likely to repeatedly watch content from top streamers.
- The broader appeal of high-quality content, which attracts diverse audiences.
- A cycling effect within larger audiences, where viewers frequently enter and exit streams, boosting aggregate watch times.

The data highlights the growing influence of non-English communities on Twitch. These languages, despite smaller overall numbers, demonstrate strong engagement and consistent growth. For streamers, this suggests that focusing on building loyal

audiences within specific linguistic niches could yield substantial benefits. Meanwhile, for English-language streamers, differentiation through quality content and strategic audience engagement remains critical to standing out in an increasingly competitive space.

If we examine the correlation matrix between all the features for Personality streamers, we can gain valuable insights into the relationships between various attributes.

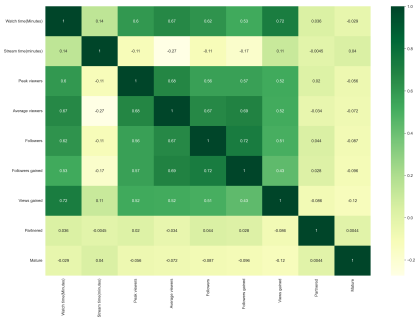

Figure 10: Correlation Matrix for the Features of Personality Streamers

The dataset lacked a feature that explicitly captured the consistency of streamers, such as how many days in a month or year they streamed. To address this gap, we engineered a new feature called **stream frequency**. A streamer is classified as "frequent" if their total stream time exceeds the average stream time of the top 1000 streamers. Conversely, those who do not meet this criterion are categorized as "not frequent." This feature provides a proxy for assessing a streamer's consistency in streaming activity.

Features Strongly Correlating with Followers:

We employed an Ordinary Least Squares (OLS) multiple linear regression model to investigate the relationship between follower count (or log-transformed follower count) and several independent variables, including **Language**, **Frequency of Streaming**, and **Maturity**. This model was chosen for its interpretability and robustness in quantifying relationships between the dependent variable and multiple predictors.

Key reasons for selecting the OLS regression approach include:

- **Interpretability**: The regression coefficients provide a straightforward means of understanding the impact of each predictor variable while controlling for others.
- **Hypothesis Testing**: The p-values associated with the predictors allow for hypothesis testing, enabling the assessment of statistical significance for each feature's contribution to the dependent variable.

The results of the regression analysis revealed statistically significant relationships (p-values ¡ 0.05) between follower count and the following predictors:

1) **Language**: Whether the streamer uses English or a non-English language.
2) **Frequency of Streaming**: Whether the streamer is classified as frequent or not based on their streaming activity.
3) **Maturity**: Whether the streamer's content is marked as mature or not.

These findings suggest that language, consistency in streaming activity, and maturity designation significantly influence a streamer's follower growth on Twitch. This highlights the importance of these features in understanding audience engagement and streamer success.

Features Strongly Correlating with Followers Gained

To identify the features most strongly influencing the growth of followers, we utilized a **Decision Tree Regressor**. This method allowed us to quantify the relative importance of various features in predicting the number of followers gained by streamers.

The analysis, visualized in Figure 11, identified the following four features as the most impactful:

1) **Followers**: The number of followers a streamer already has is the strongest predictor of followers gained. This is intuitive, as an existing follower base often defines the rate and order of growth, creating a compounding effect.
2) **Spanish**: As previously established, the significant growth of Spanish-speaking streamers

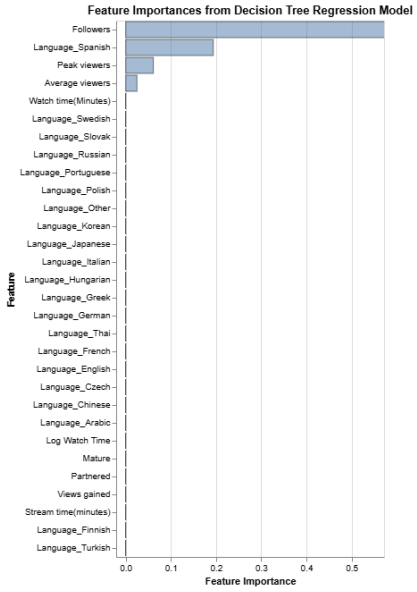

Figure 11: Feature Importances from the Decision Tree Regressor for Followers Gained

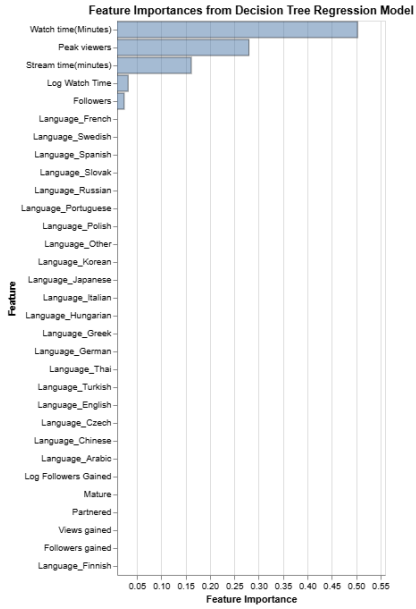

Figure 12: Feature Importances from the Decision Tree Regressor for Average Viewers

during the analyzed period highlights its importance. This feature underscores the expansion of Spanish-language content on Twitch as a driver for follower acquisition.

3) **Peak Viewers**: This metric reflects a streamer's reach and their ability to attract viewers beyond their regular audience. High peak viewership often signals successful promotional efforts or event-based streaming.
4) **Average Viewers**: A critical measure of engagement and audience size, average viewers indicate the consistency of a streamer's content in retaining their audience over time.

Features Strongly Correlating with Average Viewers

To determine the features most impactful on average viewership, we employed a **Decision Tree Regressor**. This analysis provides insight into the factors that influence how engaged a streamer's audience is over a typical stream.

As shown in Figure 12, the five most impactful features are:

1) **Watch Time**: This reflects the aggregate time viewers are willing to spend on a streamer's channel, directly correlating with audience engagement.
2) **Peak Viewers**: A key indicator of reach, peak viewers measure the ability of a streamer to attract attention during critical moments, such as events or promotions.
3) **Stream Time**: While stream time plays a role in engagement, it is slightly negatively correlated with average viewership (as observed in the correlation matrix, Figure 10), suggesting that longer streams do not necessarily equate to higher average viewership.
4) **Logarithm of Watch Time**: Similar to raw watch time, this metric highlights the level of engagement but adjusts for the exponential growth seen in large channels.
5) **Followers**: The size of a streamer's audience base strongly influences their average viewership, as it dictates the potential pool of regular and casual viewers.

These findings provide actionable insights for

streamers aiming to grow their follower base or increase average viewership. Emphasizing content quality, engagement, and strategic audience targeting in emerging linguistic markets, such as Spanish, can lead to sustained growth.

IV. Esports Channels' Analysis

In this section, we examine the **Esports** category, which includes channels officially associated with game companies, leagues, or tournaments. These channels are distinct from individual streamers in their purpose and audience. The 43 channels categorized under esports include:

Call of Duty (callofduty), CapcomFighters, dota2mc_ru, dota2ti, dota2ti_ru, DreamHackCS, DreamHackDota2_RU, DreamLeague, EAMaddenNFL, EASPORTSFIFA, ESAMarathon, ESL_CSGO, ESL_CSGO_FR, ESL_CSGOb, ESL_DOTA2, ESL_SC2, LCK, LCK_Korea, LCS, NBA2KLeague, OverwatchLeague, PG_Esports, PUBG, PlayHearthstone, Riot Games (riotgames), RiotGamesBrazil, RiotGamesJP, RiotGamesOCE, RiotGamesRU, RiotGamesTurkish, RocketLeague, StarCraft, StarLadder5, StarLadder_cs_en, Twitch, TwitchRivals, UCCleague, Warcraft, WePlayEsport_EN, WePlayEsport_RU, btscsgo, btssmash, primeleague.

This analysis focuses on the language demographics of their audiences and engagement metrics with these streams.

Correlation Analysis

We first analyze the correlation matrix for the features of these esports channels.

Language Distribution

The language distribution, as shown in Figure 14, reveals that the vast majority of esports channels use English, followed by Russian. Other languages, including Korean, account for a very small number of channels. This is noteworthy, as South Korea is often regarded as the "home of esports." However, based on data from 2020, Twitch seems to have limited penetration in the broader South Korean esports scene.

The distribution of average viewers by language for esports channels is visualized using a violin plot

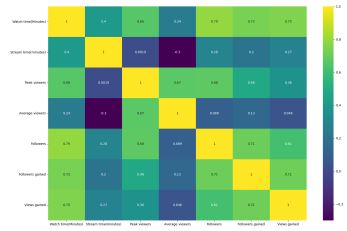

Figure 13: Correlation Matrix for the Features of Esports Channels

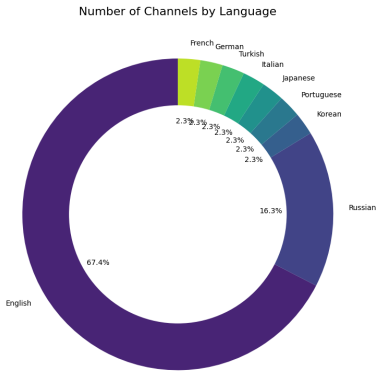

Figure 14: Donut Chart of Number of Esports Channels by Language

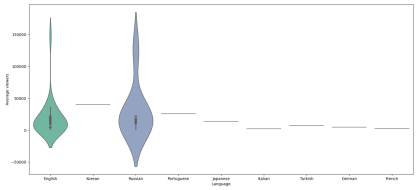

Figure 15: Violin Plot of the Distribution of Average Viewers by Language

in Figure 15. The plot reveals several key insights into the audience dynamics for different languages:

English esports channels exhibit a relatively compact distribution, characterized by a lower spread of average viewers. However, there are notable outliers above the central distribution. This pattern may be attributed to the significantly larger number of English-language channels, which could lead to greater variance in individual channel performance.

Russian esports channels, on the other hand, display wider spread of average viewers. The distribution indicates a higher density of average viewership concentrated around its central peak, suggesting a robust engagement for Russian-language channels. The other languages are displayed as dashes becasue they only have one channel each.

Average Viewers vs. Followers

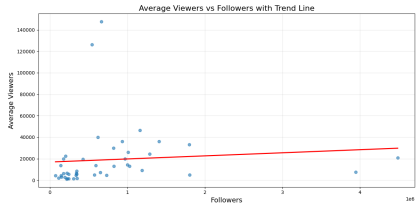

Figure 16: Scatter Plot of Average Viewers vs Followers

The scatter plot in Figure 16 shows that the average viewership trend line increases very gradually with follower counts for esports channels. The line moves steadily from approximately 20,000 average viewers to 30,000 average viewers, with most channels clustering around the line. Two notable exceptions are *dota2ti* and *dota2ti_ru*, which exhibit significantly higher average viewership, at just above 140,000 and 120,000, respectively. These outliers highlight the immense popularity of certain Dota 2 tournaments in 2020. Overall, this analysis indicates that the general audience size for esports channels hovered around the 20,000 average viewership mark during the analyzed period.

Average Viewers vs. Stream Time

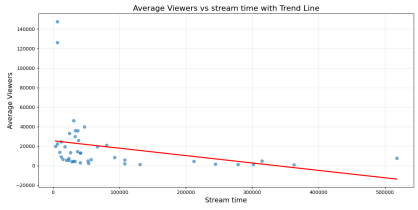

Figure 17: Scatter Plot of Average Viewers vs Stream Time

Figure 17 depicts the relationship between average viewership and stream time. The trend line shows a slight downward slope, indicating that longer stream times are not necessarily associated with higher average viewership. A prominent cluster of channels lies around the 20,000 average viewers mark and a stream time of approximately 50,000 minutes. This suggests that esports channels often follow a pattern of streaming a fixed duration of highly engaging content, rather than relying on extended streaming hours to grow viewership.

PairPlot Visualization

Figure 18: PairPlot Visualization for the Features of the Esports Channels

Figure 18 presents a pairplot visualization for the features of the esports channels. This comprehensive scatterplot matrix allows us to observe pairwise relationships between multiple features simultaneously. It serves as a valuable tool for identifying potential correlations, patterns, or trends among the features of the dataset. By visualizing these relationships, we can uncover new insights or validate existing hypotheses regarding the interaction of variables in the esports domain.

Esports Channels by Game

Since the majority of the esports channels are directly associated with game developers or championships held periodically for specific games, we manually mapped each channel to its respective game. For channels not specific to a particular game, such as the official Twitch channel or Twitch Rivals, we categorized them under the label *Various*.

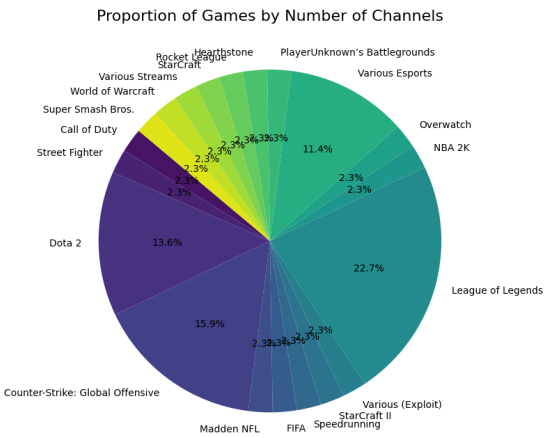

Figure 19: Pie Chart of Games by Number of Esports Channels

Some notable games represented in this data, seen in the pie chart 19 including *Call of Duty*, *Overwatch*, and *Rocket League* etc. each of which commands a significant following in its own right. However, to conduct a more in-depth analysis, we focused on three games with substantial representation across multiple esports channels. These games are:

1) League of Legends
2) Dota 2
3) Counter-Strike: Global Offensive (CS:GO)

1. League of Legends

League of Legends (abbreviated as LoL) is a multiplayer online battle arena (MOBA) game developed and published by Riot Games. Since its release in 2009, it has become one of the most popular esports titles globally, known for its strategic depth, high skill ceiling, and competitive scene. League of Legends esports features a robust structure of regional leagues and international tournaments, drawing millions of viewers annually.

The **World Championship (Worlds)** is the pinnacle of League of Legends esports, held annually to determine the best team in the world. It features the top teams from various regional leagues competing for the coveted Summoner's Cup and a multi-million-dollar prize pool, attracting a massive global audience.

The **League Championship Series (LCS)** is the premier professional League of Legends competition in North America. It serves as a platform for the region's top teams to compete and qualify for Worlds, showcasing North America's talent and strategies.

The **League of Legends Champions Korea (LCK)** represents the South Korean professional scene, often regarded as one of the most competitive and dominant regions in League of Legends history. South Korean teams are known for their discipline, innovation, and consistent performance on the global stage.

The **Prime League** is the regional professional league for German-speaking countries, including Germany, Austria, and Switzerland. It focuses on fostering local talent and serves as a gateway for teams aspiring to compete on the international stage.

As illustrated in Figure 20, the highest average viewership across League of Legends esports channels is observed on the LCS channel. This result aligns with expectations, as Twitch usage is most

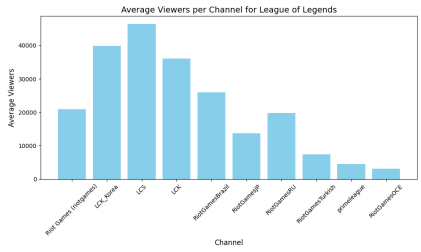

Figure 20: Average Viewers of League of Legends Esports Channels

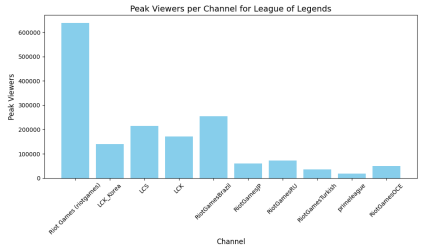

Figure 21: Peak Viewership of League of Legends Esports Channels

prominent in the United States and Canada. Given that the LCS represents the North American region, it naturally attracts a substantial local audience. The LCK and LCK_Korea channels also report high average viewership, reflecting the region's reputation as the most competitive in League of Legends esports. Additionally, these channels draw a significant number of international fans who follow South Korean teams for their strategic prowess and consistent global performance.

Although the *World Championship (Worlds)* is widely regarded as the most-watched esports event over the past decade, its average viewership on the *riotgames* Twitch channel appears lower than anticipated. This discrepancy can be attributed to regional viewing preferences; North American and European audiences may not consistently tune in to matches featuring Chinese teams or other non-Western competitors. Moreover, China, one of the largest esports markets, primarily uses alternative streaming platforms to watch Worlds, reducing the Twitch viewership figures.

Despite the lower average viewership, the Finals of Worlds attract unparalleled attention, as shown in Figure 21. The peak viewership during this event significantly surpasses that of any other channel. The second-highest peak viewership is observed on RiotGamesBrazil, which streams Worlds with Portuguese commentary, catering to the substantial esports audience in Brazil.

2. Dota 2

Dota 2 is a multiplayer online battle arena (MOBA) game developed and published by Valve Corporation. Known for its complex gameplay, steep learning curve, and strong emphasis on strategy and teamwork, Dota 2 has cultivated a passionate global player base and one of the most vibrant esports ecosystems. Its competitive scene is renowned for its high-stakes tournaments and impressive prize pools, drawing millions of viewers worldwide.

The International (TI) is the crown jewel of Dota 2 esports and one of the most prestigious tournaments in gaming history. Hosted annually by Valve, TI brings together the best teams from around the globe to compete for the largest prize pools in esports, which often exceed tens of millions of dollars. The tournament's unique crowdfunding model, driven by in-game Battle Pass sales, contributes to its massive prize pool.

ESL tournaments are a cornerstone of the Dota 2 competitive calendar, offering high-quality production and competitive gameplay. These events often serve as significant milestones for teams preparing for The International, showcasing regional and international talent.

DreamLeague is another prominent tournament series, known for its mix of regional and international competition. Organized in collaboration with ESL, DreamLeague provides an engaging platform for professional teams to compete while delivering high-quality entertainment to fans.

DreamHack, a global gaming festival, frequently includes Dota 2 tournaments as part of its lineup. DreamHack events combine competitive Dota 2 gameplay with a festival-like atmosphere, attracting casual and hardcore fans alike, and serving as a key stepping stone for aspiring professional players.

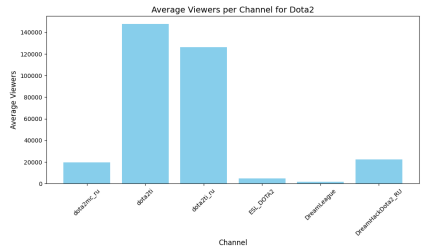

Figure 22: Average Viewers of Dota 2 Esports Channels

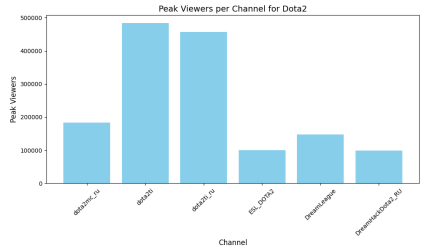

Figure 23: Peak Viewership of Dota 2 Esports Channels

The exceptionally high viewership for The International 2020 (TI10) can be attributed to several key factors. Firstly, the COVID-19 pandemic led to an increase in online entertainment consumption, as people were confined to their homes, boosting digital esports viewership. TI10, being the most prestigious esports tournament, had a significant fanbase eagerly awaiting the event after its delay. Additionally, the record-breaking prize pool, funded by the community through in-game purchases, generated further excitement. This context makes it logical that the highest two channels for viewership were dota2ti and dota2ti_ru, which consistently topped both the average viewership plot 22 and peak viewership plot 26, as these channels directly streamed the highly anticipated International event and catered to massive, diverse fanbases.

3. Counter-Strike: Global Offensive (CS:GO)

Counter-Strike: Global Offensive (CS:GO) is a highly competitive first-person shooter game developed by Valve and Hidden Path Entertainment. It has been a staple of the esports scene since its release in 2012, known for its strategic gameplay, team-based mechanics, and tactical depth. CS:GO features a highly competitive environment, with professional players and teams from around the world competing in various tournaments, making it one of the most-watched esports games globally.

The **ESL Pro League** is one of the most prestigious and long-standing CS:GO tournaments. It brings together the top teams from around the world to compete in a league format, culminating in a grand final event. Known for its high production value and global reach, ESL Pro League has become a cornerstone of the CS:GO competitive calendar, attracting millions of viewers.

StarLadder is another prominent organizer of CS:GO tournaments, known for hosting major events such as the StarLadder Major. StarLadder tournaments are celebrated for their high-quality production, fan engagement, and competitive integrity, providing a platform for top-tier teams to showcase their skills in intense competition.

DreamHack is a global esports festival and gaming convention that also hosts some of the most iconic CS:GO tournaments. DreamHack events are known for their unique atmosphere, combining competitive gaming with a festival-like environment. The DreamHack Masters series is particularly well-regarded for its competitive matches and large prize pools, making it one of the most anticipated events in the CS:GO esports scene.

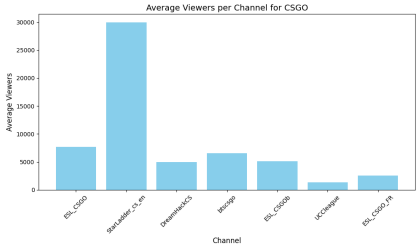

Figure 24: Average Viewers of CS:GO Esports Channels

As seen in the average viewership plot (Figure 24), both **ESL_CSGO** and **DreamHackCS** experience relatively low average viewership compared to other esports events. This can be attributed to

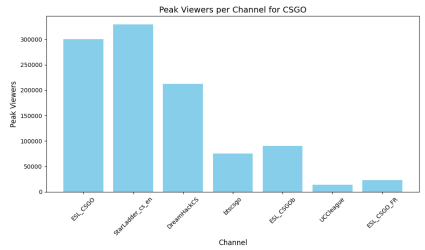

Figure 25: Peak Viewership of CS:GO Esports Channels

the extended length of these tournaments, which typically span multiple days or weeks with games occurring at various times. The early stages of these events, featuring less high-profile matchups, tend to attract fewer viewers, which brings down the overall average. Additionally, the timing of matches, often scheduled across different time zones, can result in off-peak hours where viewership drops. However, both tournaments see significant spikes in viewership during critical moments, such as finals or elimination rounds, where the competition reaches its peak, as demonstrated by the higher peak viewership numbers in Figure 25.

In contrast, **StarLadder** consistently shows both high average and peak viewership across its events, as reflected in the viewership plots. The StarLadder tournaments, particularly the **StarLadder Major**, have earned a reputation for being some of the most prestigious events in the CS:GO scene, attracting top-tier teams and a dedicated global fanbase. This sustained popularity and strong engagement throughout the event contribute to StarLadder's consistently high average viewership. Additionally, the tournament's compact format and fewer fluctuations in match timing likely contribute to maintaining steady viewership, with peak viewership reaching impressive heights, especially during the finals, showcasing its broad appeal and consistent viewer retention.

V. Games Dataset

After looking the Esports data by game, we wanted to further our analysis and give all streamers the same treatment. We had to expand with another dataset. Here is the visual analytics workflow for section V and VI.

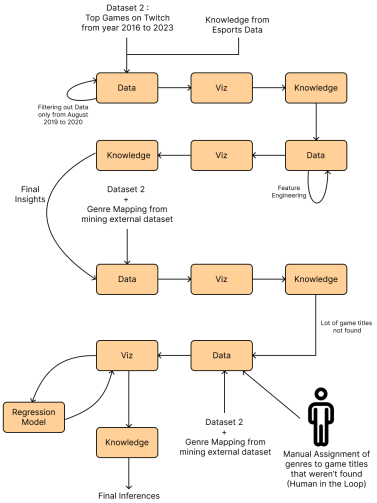

Figure 26: Visual Analytics Workflow for Sections V and VI

A. Data Preprocessing

To further our analysis, we utilized the dataset titled *Top Games on Twitch (2016–2023)*, which contains detailed information on Twitch game statistics spanning from 2016 to 2023. To align with the scope of our study, which focuses on data from August 1, 2019, to August 31, 2020, we applied the following data preprocessing steps:

1) **Temporal Filtering:** We filtered the dataset to include only data within the specified time range of August 1, 2019 to August 31, 2020. This step ensured that our analysis would be relevant and specific to the chosen timeframe.
2) **Null Value Handling:** To ensure data integrity, we conducted a thorough check for null values across all columns. The dataset was found to be free of null values.
3) **Data Type Conversion:** The `Month` and `Year` columns were converted from their original format to integer data types. This conversion facilitated numerical computations and streamlined further analysis.

4) **Duplicate Detection:** A duplicate check was performed across all rows to identify any redundant entries. No duplicate records were found, ensuring the dataset's uniqueness and reliability.
5) **Feature Engineering:** To enhance our analysis, we introduced two derived features that provide additional insights:
   - **Viewer-to-Streamer Ratio:** Calculated as the ratio of `Avg_Viewers` to `Streamers`, this feature highlights the audience engagement relative to the number of content creators.
   - **Hours per Channel:** Derived as the ratio of `Hours_Watched` to `Peak_Channels`, this metric offers insights into the distribution of viewer activity across channels.
6) **Saving the Preprocessed Data:** After completing the preprocessing steps, the refined dataset was saved as a new file. This preprocessed file serves as the foundation for our subsequent analysis and visualization, ensuring consistency and reproducibility of results.

These preprocessing steps were vital in preparing the dataset for a focused and meaningful exploration of Twitch game trends during the selected timeframe. By addressing data integrity and augmenting the dataset with derived features, we ensured a robust foundation for generating valuable insights.

B. General analysis on the dataset:

The correlation heatmap which is shown in Fig27 reveals key relationships between the numerical features in the dataset. A strong positive correlation is observed between Hours Streamed and Streamers (0.96), indicating that games with more streamers tend to have higher streaming hours. Similarly, Hours Watched shows a high correlation with Average Viewers (1.00), which is expected as more viewers directly translate to increased hours watched. Peak Channels and Average Channels are also closely correlated with Hours Streamed and Streamers, reflecting the interdependence of streaming activity and channel counts. However, the Avg

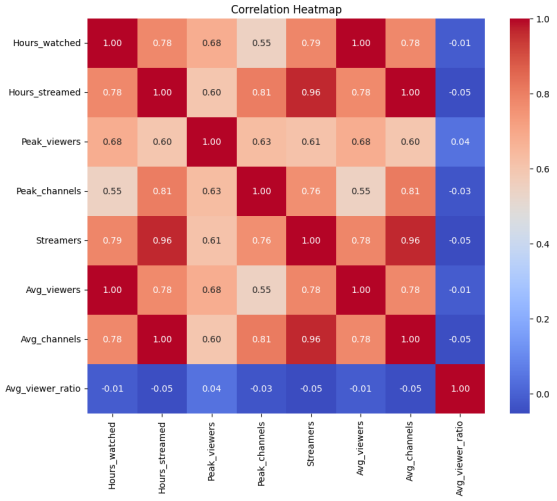

Figure 27: Correlation Matrix for the numerical Features

Viewer Ratio has minimal or no significant correlation with most features, suggesting that viewer-to-streamer dynamics operate independently of aggregate metrics like hours watched or streamed. This highlights that some games might have high engagement per streamer irrespective of their overall popularity. These correlations help identify critical factors driving Twitch game trends, providing a foundation for further analysis.

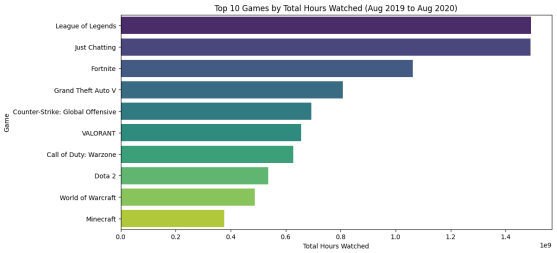

Figure 28: Top 10 Games based on Hours Watched.

From Figure 28, it is evident that *League of Legends* and *Just Chatting* are the most-watched categories, with approximately equal total watch times. These are followed by other popular titles, including *Fortnite*, *Grand Theft Auto V*, *Counter-Strike: Global Offensive (CS:GO)*, *Valorant*, *Call*

of Duty (COD), *Dota 2*, *World of Warcraft*, and *Minecraft*, each contributing significantly to the overall hours watched during the analyzed period.

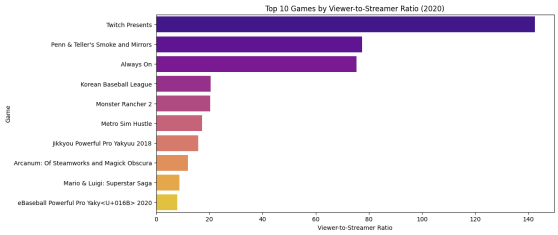

Figure 29: Top Ten Games by Viewer-To-Streamer Ratio

From Fig29 we can say that "Twitch Presents" dominates the list with the highest ratio, indicating its exceptional popularity. These are streams curated by Twitch, sometimes during in person conventions like TwitchCon, so they are designed to be highly engaging and catered to a broad audience. Similarly, "Penn & Teller's Smoke and Mirrors" and "Always On" also show high ratios, where these are also categories where an event catering to a large audience was streamed. The rest of the games on the list could also represent games a popular streamer decided to stream a handful of times, leading to a huge increase in the viewer-to-streamer ratio, or special events that captured audience attention.

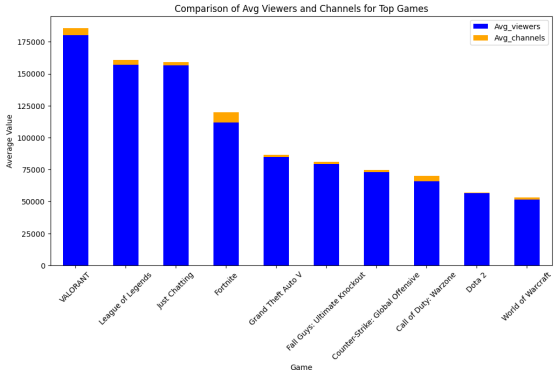

Figure 30: Showing How Avg Viewers and Avg Channels Vary for Top Games.

The bar chart in Fig30 the average viewers and average channels for top games, highlighting the balance between viewership and content creators. Games like VALORANT, League of Legends, and the Just Chatting category have the highest average viewers, suggesting their widespread popularity and appeal to large audiences. However, the relatively small orange bars representing average channels indicate that even with many content creators, viewer engagement remains high.

In contrast, games like Dota 2 and World of Warcraft have lower average viewers and channels, indicating a smaller but possibly dedicated audience base. The chart also reveals games such as Fortnite and VALORANT, which maintain a notable balance between average viewers and average channels, reflecting their consistent performance in attracting both streamers and audiences.

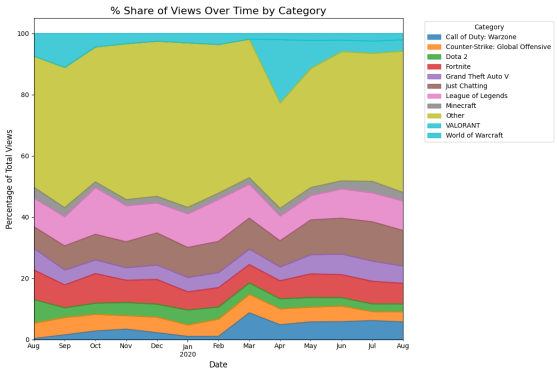

Figure 31: Percentage Share of Viewers over time by Top 10 Most Viewed Channels & Others

This stacked area chart in Fig31 illustrates the percentage share of total views across various categories over time. The "Other" category consistently dominates, indicating significant interest in games or activities outside the major listed categories. Notable trends include the steady viewership of "League of Legends" and "Just Chatting," suggesting their enduring popularity. Emerging categories such as "VALORANT" demonstrate gradual growth, likely tied to their release dates or major updates. Meanwhile, established categories like "Fortnite" and "World of Warcraft" maintain a stable yet lesser share of views. The visualization highlights the evolving preferences of viewers and

the competitive nature of online content.

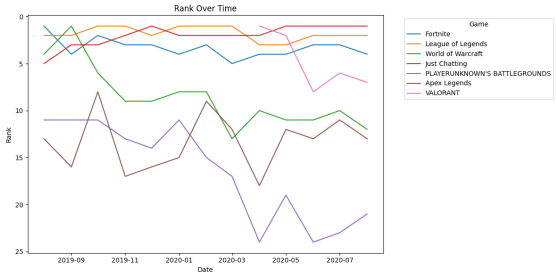

Figure 32: Rank of Popular Games over time

This line graph in Fig32 visualizes the rank of different games or categories over time based on their performance metrics. ”League of Legends” and ”Fortnite” consistently occupy top ranks, reflecting their sustained popularity among audiences. The introduction of ”VALORANT” is a standout trend, as it climbs the ranks rapidly post-launch, highlighting its immediate impact. Conversely, games like ”PLAYERUNKNOWN’S BATTLEGROUNDS” and ”Apex Legends” exhibit fluctuating ranks, suggesting variable levels of viewer engagement. This visualization provides a clear picture of how competition and new entrants influence the dynamics of category rankings over time.

The series of graphs in Fig33 analyze the performance of the top 10 games on streaming platforms . The graphs include hours watched, hours streamed, number of streamers, average viewers, and average viewer ratio.

Hours Watched: ”Fortnite” and ”League of Legends” consistently dominated hours watched, with a noticeable spike for ”Just Chatting” during early 2020, reflecting its rise in popularity as a non-gaming category. However, the popularity of games like ”Valorant” surged dramatically upon its release, showing a sharp peak during the beta testing phase around mid-2020.

Hours Streamed: ”Fortnite” and ”League of Legends” remained dominant in streaming activity, but ”Valorant” experienced a notable spike during its launch. Other games maintained relatively stable trends, though the pandemic seems to have generally increased streaming activity across most titles.

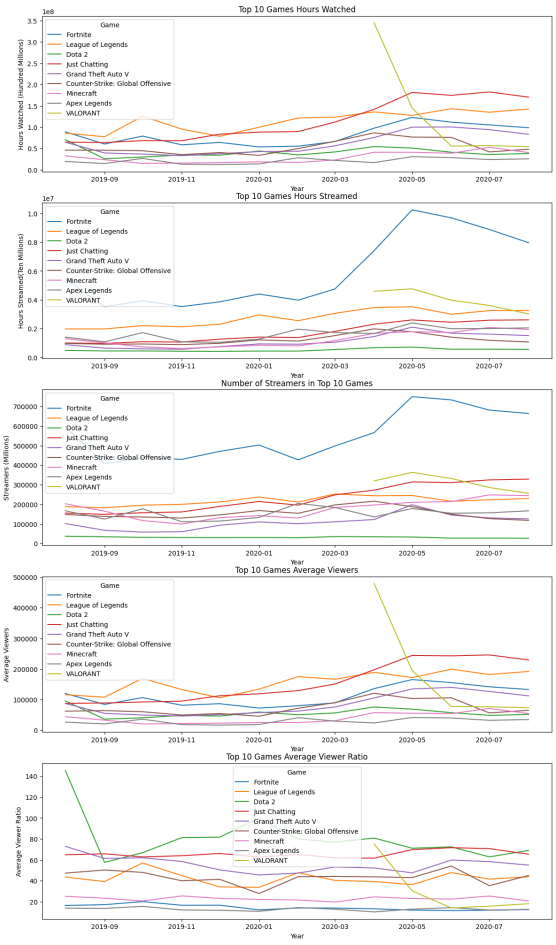

Figure 33: Ananlysis on Top 10 games

Number of Streamers: ”Fortnite” retained its lead in streamer count, while ”League of Legends” and ”Just Chatting” steadily grew. The rapid surge in streamers for ”Valorant” during its launch period in 2020 indicates significant interest and hype among content creators.

Average Viewers: The average viewer count aligns closely with trends in hours watched, emphasizing the draw of games like ”League of Legends” and ”Just Chatting.” ”Valorant” showed an explosive yet temporary rise in its average viewers, driven by strategic marketing and limited beta access.

Average Viewer Ratio: The viewer-to-streamer

ratio provides insights into the engagement levels per streamer. ”League of Legends” and ”Just Chatting” maintained strong ratios, indicating a consistent audience engagement. Meanwhile, ”Valorant” exhibited extreme fluctuations, underscoring the impact of promotional events.

Overall, the data reflects a dynamic and evolving streaming landscape influenced by major game releases, content shifts, and external factors like the COVID-19 pandemic, which boosted online engagement during this period.

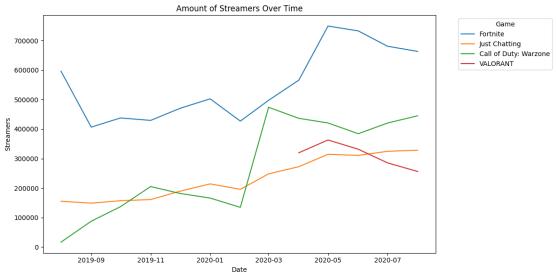

Figure 34: Number of Streamers for popular games over time

The line graph in Fig34 shows the number of streamers participating in specific categories over time. ”Fortnite” initially shows the highest number of streamers, although it experiences fluctuations, suggesting changing interest among content creators. The rise of ”Call of Duty: Warzone” streamers is particularly noteworthy, as it aligns with the game’s release and growing popularity. Similarly, the ”Just Chatting” category displays a steady upward trend, reflecting the increasing appeal of non-gaming, personal interaction-based content. ”VALORANT,” on the other hand, experiences a rapid increase in streamers following its launch, demonstrating how new releases can capture creators’ attention.

This line graph in Fig35 complemented by a shaded area representing variability, shows the viewer-to-streamer ratio over time. Peaks in the ratio indicate moments when viewer demand exceeded the number of streamers, which could coincide with significant events like game releases, tournaments, or updates. For instance, the spike

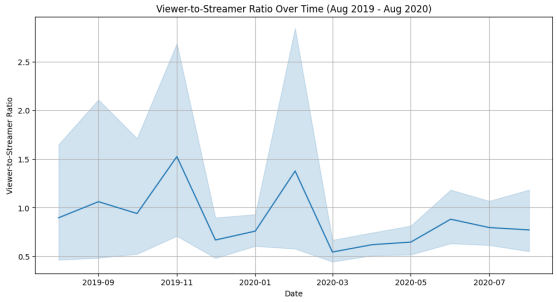

Figure 35: viewer to streamer ratio over time

around late 2019 suggests heightened interest in a specific game or category during that period. Toward mid-2020, the ratio stabilizes, reflecting a balance between the growth of the streaming community and audience engagement. This graph underscores how key events and overall trends shape the viewer-streamer dynamics.

VI. Categorical Analysis of Games

In this section, we do a further analysis on games and game data from the preprocessed Top games on Twitch 2016 - 2023 dataset([4]).

We’ve divided the flow into 3 phases:

1) Phase 1: Data Mining and Visualizations
2) Phase 2: Re-merging with a Human in the Loop and visualizations
3) Phase 3: Regression Model

A. Phase 1: Data Mining and Visualizations

To better categorize our games, we had to think of game genres. We extracted the genres for the games from the Popular Video Games 1980 - 2023([2]) dataset.

The games could have one or more of the genres below:

- **Adventure**: Games focusing on exploration, storytelling, and puzzle-solving in immersive environments. Examples: *Fortnite, World of Warcraft, Grand Theft Auto V, Minecraft, Remnant: From the Ashes*
- **Arcade**: Fast-paced games designed for reflex-based challenges and high-score competition. Examples: *Street Fighter V, Mario Kart 8, Cuphead, Brawlhalla, Streets of Rage 4*

- **Brawler**: Action-packed games emphasizing melee combat and intense battles. Examples: *Path of Exile, For Honor, Sekiro: Shadows Die Twice, Astral Chain, God of War*
- **Card & Board Game**: Games based on traditional card or board mechanics with digital enhancements. Examples: *Hearthstone, Tabletop Simulator*
- **Fighting**: Games centered on one-on-one combat and complex move sets. Examples: *Super Smash Bros. Ultimate, Tekken 7, Street Fighter V, Mortal Kombat 11, Dragon Ball FighterZ*
- **Indie**: Games developed by smaller, independent studios, often with unique gameplay or storytelling. Examples: *Dead by Daylight, Rocket League, Subnautica, RimWorld, Telling Lies*
- **MOBA (Multiplayer Online Battle Arena)**: Team-based games with strategic objectives and player roles. Examples: *League of Legends, Dota 2*
- **Platform**: Games emphasizing precision movement and jumping through obstacle-laden environments. Examples: *Super Mario Maker 2, Super Smash Bros. Ultimate, Super Mario 64, Roblox, Terraria*
- **Point-and-Click**: Narrative-driven games involving puzzle-solving through interactions with objects and characters. Examples: *Telling Lies, Danganronpa: Trigger Happy Havoc, Danganronpa 2: Goodbye Despair*
- **Puzzle**: Games focusing on logic, problem-solving, and mental challenges. Examples: *The Forest, Detroit: Become Human, The Legend of Zelda: A Link to the Past, Untitled Goose Game, Catherine: Full Body*
- **RPG (Role-Playing Game)**: Games where players assume roles of characters in expansive worlds, often with character progression. Examples: *Fortnite, League of Legends, World of Warcraft, Remnant: From the Ashes, Destiny 2*
- **Racing**: Games focused on vehicular competition or challenges. Examples: *Rocket League, Grand Theft Auto: San Andreas, Mario Kart 8, Need for Speed: Heat, Grand Theft Auto IV*
- **Real-Time Strategy**: Strategy games where decisions must be made dynamically in real-time. Examples: *RimWorld, Europa Universalis IV, Clash of Clans*
- **Shooter**: Games emphasizing ranged combat with firearms or similar mechanics. Examples: *Fortnite, Overwatch, Apex Legends, Tom Clancy's Rainbow Six Siege, Counter-Strike: Global Offensive*
- **Simulator**: Games simulating real-world or fictional activities with immersive mechanics. Examples: *Minecraft, Sea of Thieves, The Sims 4, Roblox, RimWorld*
- **Sport**: Games based on real-world or fantasy sports, often emphasizing competition. Examples: *Rocket League, Need for Speed: Heat*
- **Strategy**: Games involving long-term planning and resource management to achieve objectives. Examples: *Fortnite, League of Legends, Overwatch, Hearthstone, Dead by Daylight*
- **Tactical**: Games with strategic depth requiring careful planning and execution of moves. Examples: *Tom Clancy's Rainbow Six Siege, Fire Emblem: Three Houses, Bloons TD 6, Plague Inc: Evolved*
- **Turn-Based Strategy**: Games where players take turns to execute their moves strategically. Examples: *Hearthstone, Fire Emblem: Three Houses, Final Fantasy VII, Darkest Dungeon, Paper Mario: The Thousand-Year Door*
- **Visual Novel**: Story-rich games with interactive narrative experiences, often featuring branching storylines. Examples: *Phoenix Wright: Ace Attorney Trilogy, Danganronpa: Trigger Happy Havoc, Danganronpa 2: Goodbye Despair, Persona 4 Golden*
- **Other**: Diverse games that don't fit into specific genres or represent general categories. Examples: *Just Chatting, Art, Music*

For games in the dataset [4] where we couldn't find exact word mappings for the titles, such as:

- Old School RuneScape
- World of Tanks
- Black Desert Online
- Dota Underlords

We assigned a 'Not_Found' genre in the merged

dataset for these titles.

Visualizations for the merged dataset: We plotted an interactive parallel coordinates plot for key metrics(Hours_watched, Hours_streamed, Peak_viewers, Avg_viewers, Avg_channels, Hours_per_Channel, Avg_viewer_ratio, and Viewer_Streamer_Ratio)(36)

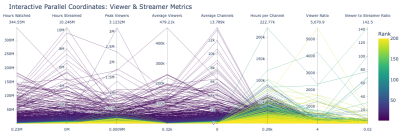

Figure 36: Interactive Parallel Coordinates plot for key Metrics(Hours_watched, Hours_streamed, Peak_viewers, Avg_viewers, Avg_channels, Hours_per_Channel, Avg_viewer_ratio, and Viewer_Streamer_Ratio)

To better understand correlations between the axes, we reordered the axes(Figure.37):

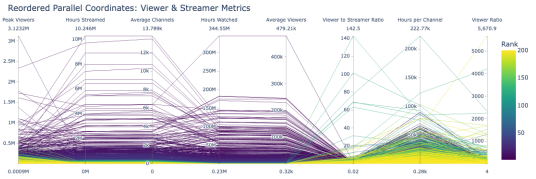

Figure 37: Interactive Parallel Coordinates plot 36 with reordered axes

We also plotted a SPLOM for numerical features(Figure38).

Inferences from the SPLOM(Figure38):

- Hours_watched: This feature shows a positive correlation with the number of viewers, as indicated by the scatter plot in row 1, column 4. The data points are more concentrated at lower hour values, with a wider spread as the hours watched increases.
- Peak_viewers: This feature demonstrates a wide range of values, with some observations having significantly higher peak viewership compared to others. This suggests the dataset contains a mix of content that attracts varying levels of concurrent viewers.
- Avg_viewers: Similar to Peak_viewers, the Avg_viewers feature also shows a wide range of values, with some content attracting significantly more viewers on average than others.

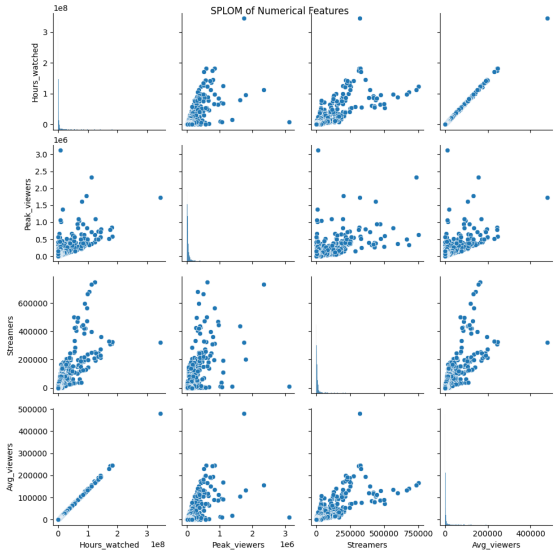

Figure 38: SPLOM of numerical features

To utilize the now extracted 'Genres' feature, we plotted the Total Hours Watched per Game Genre in the time period of our interest(August 2019 to August 2020)(Figure.39)

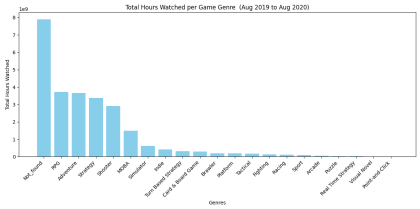

Figure 39: Total Hours Watched per Game Genre (Aug 2019 to Aug 2020)

B. Phase 2: Re-merging with a Human in the Loop and visualizations

From Figure.39 we see that an overwhelming amount of watch-time can be seen for the

Not_found category of titles. In fact there were 350+ titles to which we didn't find genre mappings in Phase 1.

To address this issue, we manually mapped every 'Not_found' title to a genre from the list of genres that were previously assigned to other titles in Phase 1.

For the game titles we couldn't find exact word mappings in Phase 1, we manually assigned them genres after looking those up on the web.

An example mapping we did in a CSV file is:

- Old School RuneScape,RPG
- World of Tanks,Strategy
- Black Desert Online,RPG
- Dota Underlords,Real Time Strategy

For the non-game list of titles, we mapped them to the 'Other' genre. For example we made the mappings:

- Just Chatting,Other
- Art,Other
- Music,Other

Visualizations for the merged and manually mapped dataset: The distribution of this new data is visualized as an interactive pie chart(Figure.40).

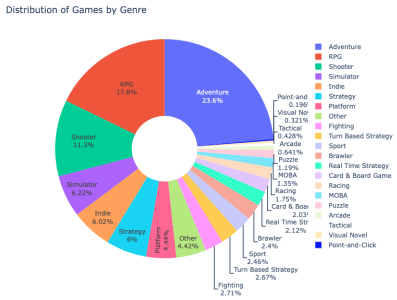

Figure 40: Distribution of Games by Genre

We see that Adventure(23.6%), RPG(17.8%), and Shooter(11.3%) are three major genres in the dataset.

We then plotted an interactive treemap, which displays the top 3 games per genre by total watch time(Figure.41).

We plotted a Genre-enabled interactive Sankey plot(Figure.42) which is the counterpart of the parallel coordinates plot made in Phase 1(Figure.37).

The Sankey plot contains a list of flows. Each flow is denoted by a color and is specific to a genre.

We plotted an interactive bubble chart to draw out the relationship between average viewers, streamers, and hours watched per genre(Figure.43). This bubble chart can simultaneously display the popularity (hours watched), engagement (average viewers), and activity level (streamers).

In the interactive Plotly bubble chart, each data point (bubble) is sized, positioned, and colored based on the values from the dataset as follows:

1) Positioning (x and y coordinates): The x-axis represents the number of Streamers (horizontal position of the bubble). The y-axis represents the Avg_viewers (vertical position of the bubble). Each data point's position is determined by the corresponding values in the Streamers and Avg_viewers columns for each genre.
2) Sizing (bubble size): Bubble size is based on the Hours_watched column, normalized to ensure the sizes are visually distinct and fit well in the plot. The size is computed using:

$$\text{bubble\_data['Bubble\_size']} = \frac{\text{bubble\_data['Hours\_watched']}}{\max(\text{bubble\_data['Hours\_watched']})} \times 1000$$

This normalization maps the Hours_watched values to a consistent scale (in this case, a maximum size of 1000), so that larger values result in bigger bubbles, but all bubbles remain proportional to their respective values.
3) Coloring: The color of each bubble corresponds to the Hours_watched column value. A continuous color scale (Viridis) is used, where: Smaller Hours_watched values are represented with lighter colors. Larger Hours_watched values are represented with darker, richer colors. This provides a visual cue about the magnitude of Hours_watched.
4) Hover Information: When you hover over a bubble, it shows the following details(Figure.44): Genre: The name of the genre. Number of Streamers: Value from the Streamers column. Average Viewers: Value from the Avg_viewers column. Total Hours

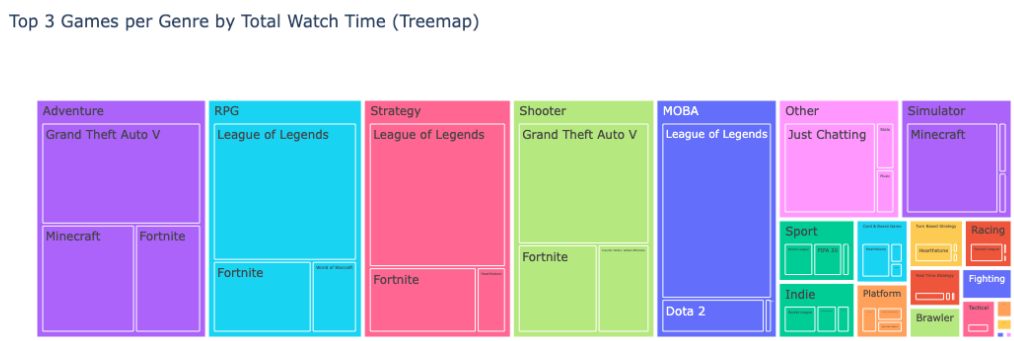

Figure 41: Top 3 watched games per genre

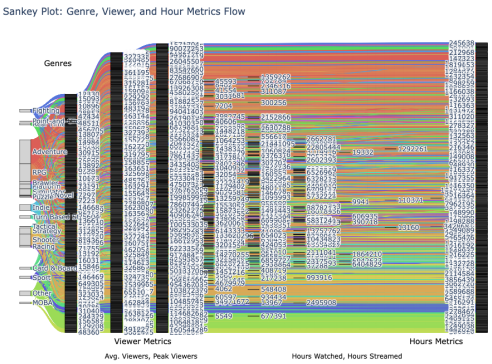

Figure 42: Sankey Plot: Genre, Viewer, and Hour Metrics Flow

Watched: Value from the Hours_watched column.

We made an interactive treemap to analyze the genre popularity based on total number of hours watched(Figure.45).

- Shooter genre is the most popular, with over 7,510,296,876 hours watched.
- Adventure genre is the second most popular, with over 7,362,196,945 hours watched.
- RPG genre is the third most popular, with over 6,069,310,399 hours watched.
- MOBA genre is the fourth most popular, with over 3,621,505,097 hours watched.
- The "Other" genre category includes a significant number of hours watched, at over 2,286,524,080.
- Simulator and Sport genres have the next highest number of hours watched, at around 1,878,178,392 and 680,854,509 respectively.
- Visual Novel and Point and Click genres have the lowest number of hours watched, at around 10,802,396 and 7,712,619 respectively.

We made a juxtaposed view of two time series visualizations composed of interactive stacked area and line charts depicting month-wise total and average watch time by genre(Figure.46).

Inferences from the Time Series Data

- **Total Hours Watched Over Time by Genre (Left Chart)**
  - **Overall Trend:**
    - Total hours watched across all genres collectively peaked around mid-2020, suggesting a significant event or seasonality (e.g., the COVID-19 pandemic) increased viewership.
    - A noticeable dip occurred between late 2019 and early 2020, reflecting a temporary reduction in total viewership for all genres.
  - **Genre-Specific Insights:**

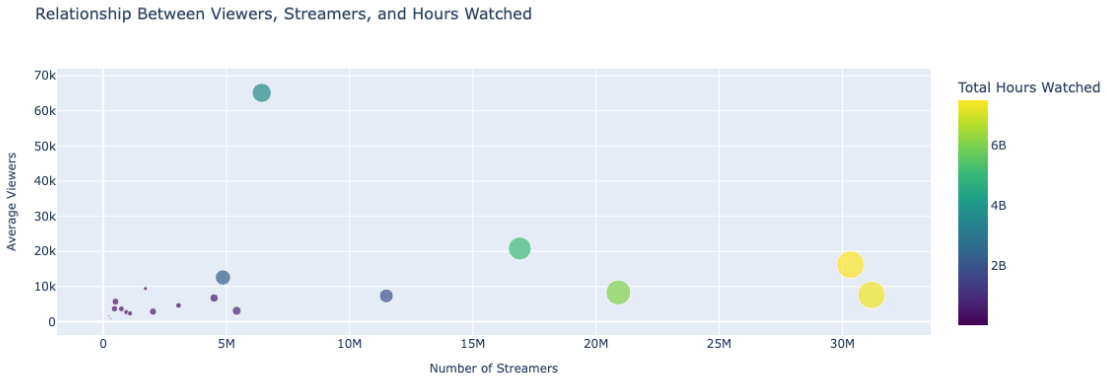

Figure 43: Relationship Between Viewers, Streamers, and Hours Watched(Per Genre)

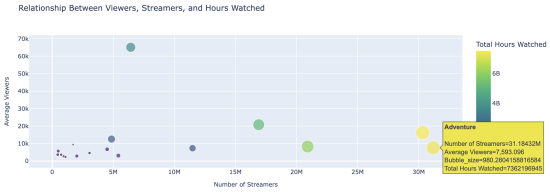

Figure 44: Relationship Between Viewers, Streamers, and Hours Watched(Per Genre): Hovered

    - ∗ Dominant genres, such as RPG, Shooter, and MOBA, contributed significantly more to the total hours watched.
    - ∗ Niche genres like *Visual Novel* and *Turn-Based Strategy* contributed relatively less, reflecting smaller audience sizes.
  - **Growth and Decline:**
    - ∗ The recovery after early 2020 was uniform across genres, but dominant genres experienced faster growth, widening the contribution gap.
- **Average Watch Time Per Game Over Time by Genre (Right Chart)**
  - **Volatility:**
    - ∗ The average watch time per game is highly volatile compared to total hours watched.
    - ∗ Certain genres, such as *Visual Novel* and *Turn-Based Strategy*, exhibit extreme spikes, indicating periodic surges in popularity due to specific games.
  - **Distinct Spikes:**
    - ∗ A sharp spike for *Visual Novel* in late 2019 and for *Turn-Based Strategy* in early 2020 suggests viral games or events driving temporary high viewership.
  - **Uniformity Across Most Genres:**
    - ∗ Stable average watch time for genres like *MOBA*, *Shooter*, and *RPG* indicates consistent audience engagement.
  - **Correlation Between Metrics:**
    - ∗ Genres with lower total hours watched exhibit higher variability in average watch time, possibly due to smaller but more dedicated communities.
- **Combined Inferences**
  - Dominant genres shape the overall trend of total hours watched, while smaller genres add variability to average watch time.
  - Periodic spikes in both charts align with genre-specific events or game releases.

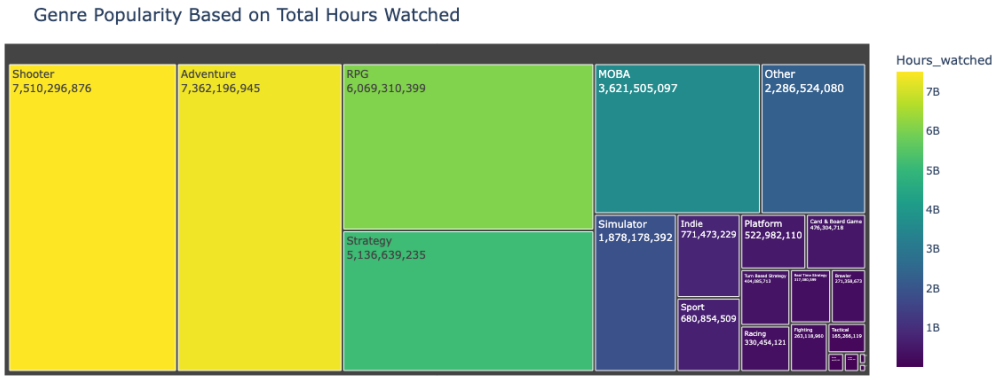

Figure 45: Genre Popularity Based on Total Hours Watched

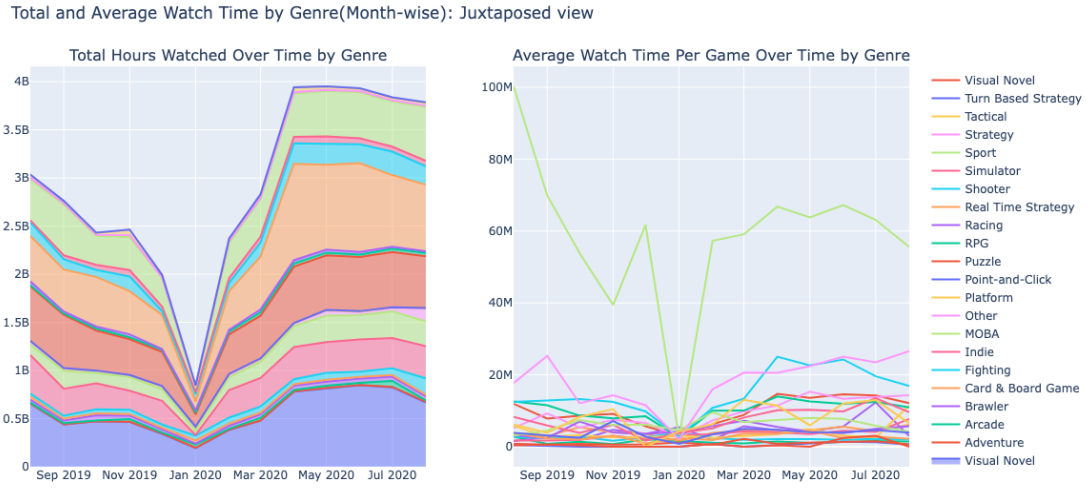

Figure 46: Total and Average Watch Time by Genre(Month-wise): Juxtaposed view

- Consistent dips in late 2019 and rises through mid-2020 suggest a cyclical or event-driven seasonality.

- **Hypotheses**

1) The increase in total hours watched during mid-2020 correlates with the global COVID-19 pandemic, which led to more people staying indoors and engaging in gaming.
2) Spikes in average watch time for *Visual Novel* and *Turn-Based Strategy* are driven by the release of specific popular games or events in those genres.
3) Dominant genres like RPG and Shooter consistently attract more viewers due to their broad appeal and frequent content updates.
4) The seasonal dip in late 2019 corresponds

to a period of reduced gaming activity, possibly due to holiday travel or fewer major game releases.

5) Smaller genres with niche audiences have higher variability in average watch time, indicating the impact of individual games on the metrics.

The broad classification of games by genres gave us a way of making set visualizations to visualize intersections, as each game can be classified into multiple genres.

We first plotted a Venn Diagram to see if any intersections exist between the top 3 most popular genres(popularity based on game count per genre)(Figure.47).

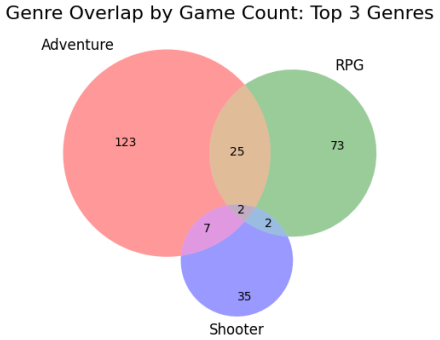

Figure 47: Genre Overlap by Game Count of Top 3 Genres: Adventure, RPG, and Shooter

There were significant intersections between the top 3 genres. To visualize all the intersections present in the data, we made an UpSet Plot([5])(Figure.48).

Phase 3: Regression Model

To analyze the effectiveness of our genre mappings, we created regression models that predict the Hours_watched based on the other features.

Preprocessing the Data: Data preprocessing is critical to ensure the model can accurately interpret the features and handle any inconsistencies in the dataset. The following steps were performed:

- **Feature Selection:** We removed columns such as ‘Game‘ and ‘Date‘ as they do not directly contribute to the regression analysis and are non-numeric in nature. The target variable ‘Hours_watched‘ was also separated from the feature set.
- **Categorical Features Encoding:** The categorical columns ‘Genres‘ and ‘Month‘ were encoded using one-hot encoding to convert them into numerical representations. This allows the model to interpret the categorical data effectively.
- **Handling Missing Values:** Missing numerical values were imputed using the mean of the respective column, while missing categorical values were imputed using the most frequent value.
- **Scaling Numerical Features:** Standard scaling was applied to numerical features to standardize their range and improve the model’s performance by ensuring all features contribute equally.

Model Creation and Training: We utilized two regression models to predict ‘Hours_watched‘:

- **Random Forest Regressor:** A pipeline was created to preprocess the data and fit a Random Forest Regressor. Random Forest is an ensemble learning method that uses multiple decision trees to improve prediction accuracy. The pipeline ensured seamless preprocessing and training.
- **Linear Regression:** As a simpler baseline model, a Linear Regression model was also implemented. One-hot encoding was directly applied to the categorical features, and the model was trained on the processed data.

Model Evaluation: The performance of both models was evaluated using the following metrics:

- **R-squared (R^2):** Indicates the proportion of variance in the target variable explained by the model. Higher values signify better performance.
- **Mean Squared Error (MSE):** Measures the average squared difference between predicted and actual values, with lower values indicating better predictions.
- **Root Mean Squared Error (RMSE):** The square root of MSE, providing an interpretable

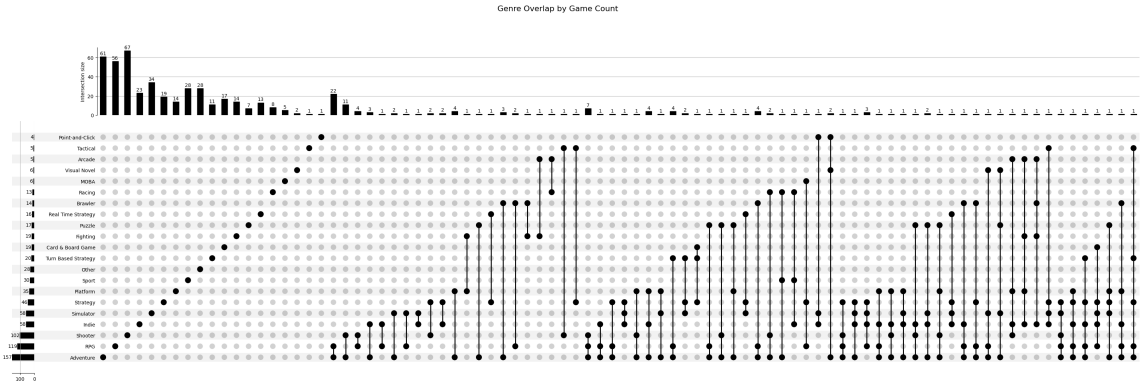

Figure 48: Genre Overlap by Game Count: Upset Plot([5])

error metric in the same units as the target variable.

The Random Forest Regressor demonstrated high performance with satisfactory R^2, MSE, and RMSE values, making hyperparameter tuning unnecessary in this instance. Similarly, the Linear Regression model achieved exceptional results, demonstrating near-perfect predictive capability.

Findings and Next Steps: Both models provided robust predictions for the 'Hours_watched' metric. The high R^2 value of the Linear Regression model (0.9999) indicates that the dataset is well-structured and features are highly predictive of the target variable. While Random Forest offers flexibility and handles non-linear relationships effectively, its complexity might not be necessary given the excellent performance of the simpler Linear Regression model.

Future work could explore additional features, fine-grained hyperparameter tuning, or other regression algorithms to further optimize the predictions or generalize the model to other datasets.

As of the time of writing this report, we have not performed hyperparameter tuning as we got a satisfactory R^2 score without it.

An R^2 value close to 1 suggests that the model explains nearly all of the variance in the target variable (Hours_watched), which indicates a very good fit.

1) Model Coefficients Analysis: The coefficients of a regression model represent the impact of each feature on the predicted outcome. In our case, the model's coefficients indicate the effect of each feature, including categorical variables like genres, on the `Hours_watched`. Higher magnitude coefficients imply stronger relationships with the target variable, while lower magnitude coefficients suggest weaker relationships.

The model coefficients for each feature are as follows:

- `Genres_MOBA`: 59578.06
- `Genres_Strategy`: 8958.99
- `Genres_Simulator`: 6124.15
- `Genres_RPG`: 5568.92
- `Genres_Shooter`: 4929.28
- `Genres_Adventure`: 3218.64
- `Avg_viewers`: 732.75
- `Month`: 518.20
- `Genres_Indie`: 446.38
- `Peak_channels`: 33.46
- `Hours_streamed`: 24.53
- `Streamers`: 1.11
- `Hours_per_Channel`: 0.98
- `Peak_viewers`: -0.15
- `Avg_viewer_ratio`: -16.87
- `Genres_Tactical`: -993.60
- `Viewer_Streamer_Ratio`: -1081.70
- `Genres_Brawler`: -1582.44
- `Genres_Platform`: -2775.91
- `Genres_Puzzle`: -2931.49
- `Genres_Fighting`: -3041.06
- `Genres_Turn Based Strategy`: -

5309.95

- `Genres_Racing`: -5487.97
- `Genres_Visual Novel`: -5529.26
- `Genres_Arcade`: -6064.93
- `Genres_Sport`: -7474.54
- `Genres_Other`: -8293.85
- `Genres_Real Time Strategy`: -9306.35
- `Genres_Card & Board Game`: -11279.06
- `Avg_channels`: -18161.17
- `Genres_Point-and-Click`: -18754.00

These coefficients can be visualized in(Figure.49).

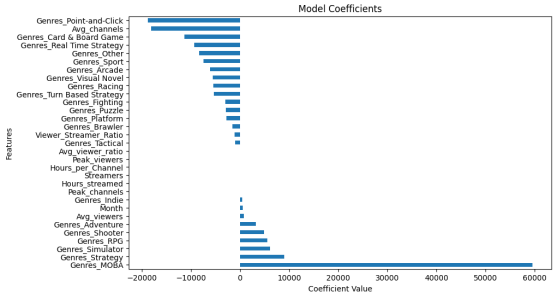

Figure 49: Model Coefficient Data

As we can see, the highest coefficients are associated with the genre features, such as `Genres_MOBA`, followed by `Genres_Strategy` and `Genres_Simulator`. These genres have a significant influence on `Hours_watched`, highlighting the importance of genre-specific features in predicting viewership time.

Visualizing the Coefficients

Visualizing the model's coefficients helps in understanding the relative importance of each feature. We use a horizontal bar chart to compare the magnitude of the coefficients for all features. Features with larger coefficients have a stronger impact on the target variable.

Examining the Influence of Categorical Features (Genres)

Since the `Genres` variable was one-hot encoded into multiple binary features, each genre has its own coefficient. By examining the coefficients corresponding to the genre features, we can assess the relative importance of each genre in determining `Hours_watched`.

We visualized the genre-specific coefficients in the following bar chart, which highlights the genres that have the most significant influence on the target variable(Figure.50).

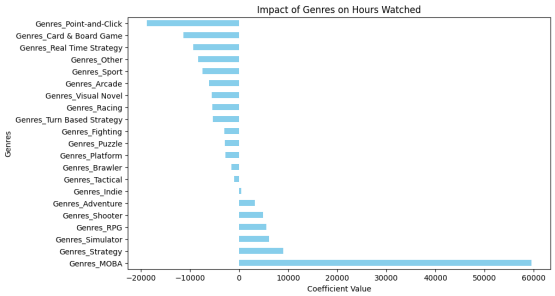

Figure 50: Impact of Genres on Hours Watched

Visualizing the Correlation Matrix

A correlation matrix is a useful tool for identifying relationships between features and understanding the potential multicollinearity issues in the dataset. High correlation between features may lead to redundancy and affect the stability of the model.

We calculated the correlation matrix for the features in the training dataset, including the target variable `Hours_watched`. The following heatmap shows the correlation between features(Figure.51).

The heatmap visually represents how strongly the features are correlated with each other and with `Hours_watched`. This analysis ensures that no features are highly correlated, which could potentially affect model interpretability.

Model Performance Evaluation

To assess the performance of the regression model, we evaluated several key metrics:

- **R^2**: 0.9999
- **Mean Squared Error (MSE)**: 58002674592.81
- **Root Mean Squared Error (RMSE)**: 240837.44

The R^2 score of 0.9999 indicates that the model explains almost all the variance in the target variable, suggesting excellent predictive accuracy. The

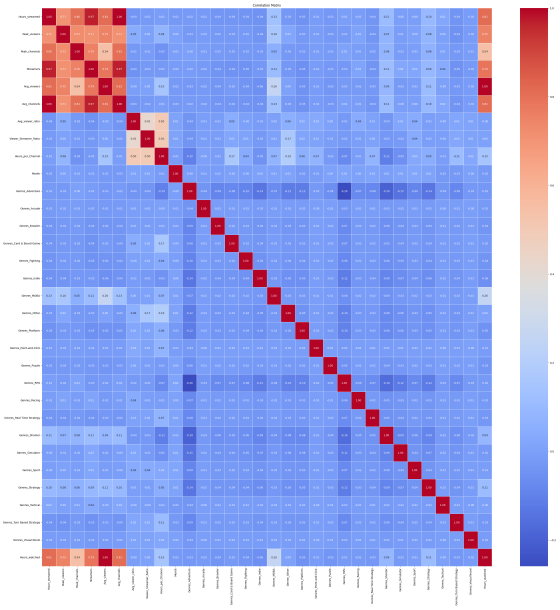

Figure 51: Correlation matrix between features

MSE and RMSE provide additional context regarding the magnitude of errors in the model's predictions, with lower values indicating better performance. The values indicate that the Mean Squared Error(MSE) and the Root Mean Squared Error(RMSE) are high(in contrast to the good R^2 score) suggesting that future improvements could involve minimizing them. At the time of writing this report, we end Phase 3 solely because of the R^2 score obtained and significant contribution of genres as model coefficients.

Phase 3: Conclusion

In this section, we interpreted the regression model by examining the coefficients, visualizing the feature importance, analyzing the correlation matrix, and evaluating the model's performance. The model achieved a near-perfect R^2 score, demonstrating that it effectively predicts `Hours_watched` based on the provided features.

The genre features, particularly `Genres_MOBA`, had the most significant impact on the target variable, underscoring the importance of genre-specific data. Furthermore, the analysis of the correlation matrix confirmed that there were no major issues with multicollinearity, ensuring that the model's coefficients are reliable. Viz Overall, incorporating `Genres` as one-hot encoded features gave a great model performance, leading to a highly accurate prediction of `Hours_watched`.

VII. Contributions

- **Narayana Udayagiri**:
  - Section 2: Data Manipulation and Inferences
  - Section 3: Streaming Personality Analysis
- **Tahir Khadarabad**:
  - Section 4: Esports Channels' Analysis
  - Section 5: Games Dataset
- **Pradyun Devarakonda**:
  - Section 6: Categorical Analysis of Games

References

1. Backlinko, Twitch User Statistics (2024), https://backlinko.com/twitch-users (2024). Accessed: 2024-12-11.
2. Chaki, A.: Popular Video Games (1980 - 2023), https://www.kaggle.com/datasets/arnabchaki/popular-video-games-1980-2023 (2023). Accessed: 2024-12-11.
3. Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., Melançon, G.: Visual analytics: Definition, process, and challenges. Springer (2008)
4. Kirsh, R.: Evolution of Top Games on Twitch, https://www.kaggle.com/datasets/rankirsh/evolution-of-top-games-on-twitch/data (2024). Accessed: 2024-12-11.
5. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., Pfister, H.: UpSet: visualization of intersecting sets. IEEE transactions on visualization and computer graphics **20**(12), 1983–1992 (2014)
6. SimilarWeb, Twitch.tv - Audience Geography and Analytics, https://www.similarweb.com/website/twitch.tv/#geography (2024). Accessed: 2024-12-11.

Appendix: A1 - Analysis of the Twitch Dataset

Udayagiri Narayana Srimanth
IMT2022052
Udayagiri.Srimanth@iiitb.ac.in

Khadarabad Tahir Mohammed
IMT2022100
Khadarabad.Mohammed@iiitb.ac.in

Pradyun Devarakonda
IMT2022525
Devarakonda.Pradyun@iiitb.ac.in

***Index Terms*—Twitch, Streamers, Views, Followers**

I. Introduction

We worked on analyzing data and creating visualizations of the Top Streamers on Twitch dataset.The dataset contains data on Twitch's top 1000 streamers(based on the number of followers) from August 2020. It has a total of 11 columns.

1) Channel Name: Display name of the channel
2) Watch Time(Minutes): Total watch time on the channel in that particular year
3) Stream time(minutes): Total time streamed on the channel in that particular year
4) Peak viewers: Max viewers reached by a stream on the channel in that particular year
5) Average viewers: Average viewers across streams on the channel in that particular year
6) Followers: The number of followers for each channel
7) Followers gained: The number of followers gained for each channel in that particular year
8) Views gained: The number of views gained for each channel in that particular year
9) Partnered: Whether the channel is a Twitch Partner or not
10) Mature: Whether the channel is marked as mature content or not
11) Language: The primary language in which the streamer's content is broadcast.

To effectively categorize streamers(based on the number of followers) in the dataset, we divided them into four categories:

- Category 1 - 0 to 100k followers
- Category 2 - 100k to 500k followers
- Category 3 - 500k to 1M followers
- Category 4 - 1M to 10M followers

This categorization serves two main purposes:

- To segment the wide range of follower counts, as the numbers were too large to consider collectively.
- To reflect social perceptions of popularity, where 100k, 500k, 1M, and 10M followers are commonly regarded as significant milestones for streamers.

Of the 11 columns, 2 columns (Channel Name and Language) are of string data type, 2 columns (Partnered and Mature) are of boolean data type, and the remaining 7 columns are of integer data type.

II. Task division

We divided our work into three major tasks:

- T1: Overview of Key Metrics
- T2: Growth Trends and Audience Engagement
- T3: Language-Specific Insights

As a first step, we created a correlation matrix of all 10 columns to analyze the relationships between the variables.

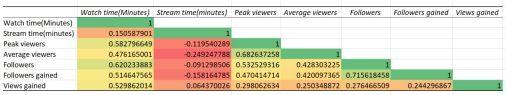

Fig. 1. The Correlation matrix of the Dataset

III. T1: Overview of Key Metrics

This section provides analysis of key performance metrics for the top Twitch streamers, including follower counts, total views, and watch time. By examining these metrics, we aim to understand the correlation between them. The insights gained offer a foundational overview of streaming trends and audience interaction on the platform.

A. Observation 1

The majority of the top 1000 streamers on Twitch fall within Category 2 (100,000 to 500,000 followers), comprising more than 55% of the dataset. This is followed by Category 1 (1,000,000 to 10,000,000 followers), Category 3 (500,000 to 1,000,000 followers), and Category 4 (0 to 100,000 followers).

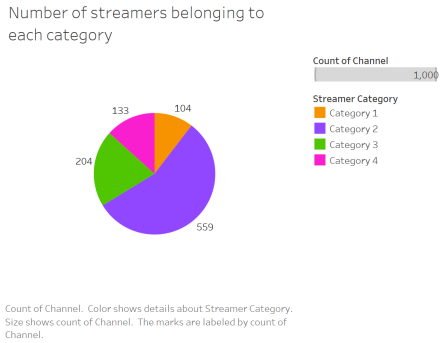

Fig. 2. Streamers of each category

The distribution of these categories is depicted through a pie chart (Fig.2), with each colour corresponding to one of the

categories. The size of each segment illustrates the proportion of the respective category, and the colour scheme was selected to reflect the aesthetic used by Twitch in their annual Twitch Recaps.

At the time of data collection in August 2020, Tfue was the top streamer, with over 8.9 million followers, while the streamer ranked 1000th, voicetv, had 3,660 followers. The range of follower counts within the top 1000 highlights significant variability in audience size. This dataset allows for the examination of correlations between variables such as follower count, watch time, and stream time.

Given that the dataset is of the top 1000 streamers, any trends observed can be used to analyze how the average viewer interacts with users who stream professionally because it excludes the vast majority of accounts, operated by viewers, where stream times and follower counts are virtually zero. This prevents skewing the data, and allows us to apply our insights to the platform as a whole.

B. Hypothesis 1

High-performing streamers are more likely to be Twitch Partners. Successful streamers on Twitch are heavily engaged in the platform's revenue streams, which are primarily derived from advertisements and channel subscriptions. To access these revenue sources, streamers must achieve partnership status, which provides them with a share of the platform's earnings, a direct point of contact at Twitch, better channel customization and special promotional opportunities.

This hypothesis is validated by the data, it shows that 97.8% of the top 1000 streamers are partnered, supporting the conclusion that partnership is nearly universal among this group.

C. Observation 2

We observe that streamers with higher follower counts are more likely to maintain higher average watch time and viewer counts(Fig.3, Fig.4, Fig.5). This phenomenon is logical, as streamers who consistently attract large audiences tend to cultivate a loyal viewer base. These viewers are more inclined to return for subsequent streams, contributing to sustained high viewership and prolonged watch times. Streamers with larger followings often produce content that is more engaging, diverse, or of higher quality, which in turn fosters viewer retention. This feedback loop of increased audience engagement and viewer retention likely reinforces the correlation between high follower counts and elevated average watch time and viewer numbers.

The figure(Fig.3) is a line chart in which average watch time is plotted against the number of followers, excluding Category 4 (1,000,000 to 10,000,000 followers). The exclusion of the top 204 streamers in Category 4 was necessary, as their inclusion would have resulted in a disproportionately steep linear regression slope. This provides a more accurate representation of the correlation between these two variables. The colour scheme was chosen based on Twitch's colour palette for visual appeal.

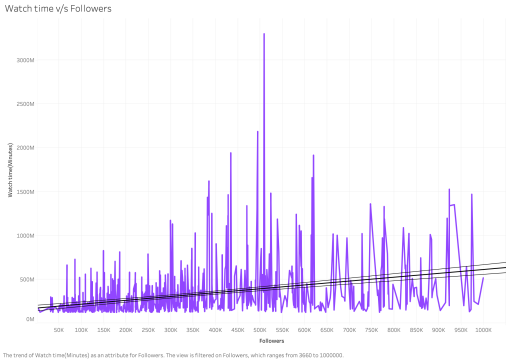

Fig. 3. Watchtime vs Followers

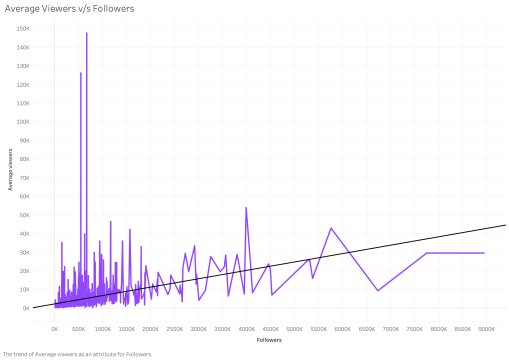

Fig. 4. Average Viewers vs Followers

The figure(Fig.4) is a line plot, where every point on it represents average viewership at a specific follower count. As anticipated, the trend line shows a positive correlation between follower count and average viewership. The colour scheme was similarly selected based on Twitch's colour palette for coherence and visual consistency across figures.

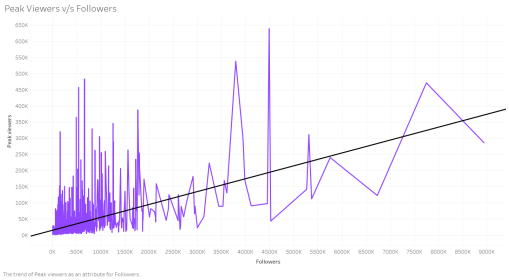

Fig. 5. Peak Viewers vs Followers

The graph(Fig.5) presents a line plot in which each data point represents the peak viewership of a channel at a given follower count. The trend line illustrates a positive correlation between follower count and peak viewership. The colour scheme was chosen to align with Twitch's colour palette.

The graph (Fig.6) illustrates that peak views exhibit a higher initial value and increase more rapidly with follower count compared to average views. This observation is supported by the trend lines, where the slope and intercepts of the peak

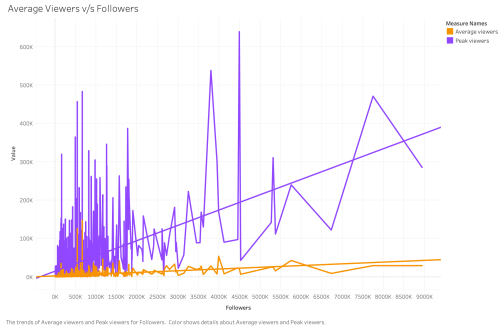

Fig. 6. Peak Viewers and Average Viewers vs Followers

views trend line (purple) are significantly greater than those of the average views trend line (orange).

D. Observation 3

The impact of stream time across various metrics reveals nuanced patterns in the behaviors of streamers and their audience. As streamers increase their follower counts, the average amount of time they stream tends to decrease(Fig.7).

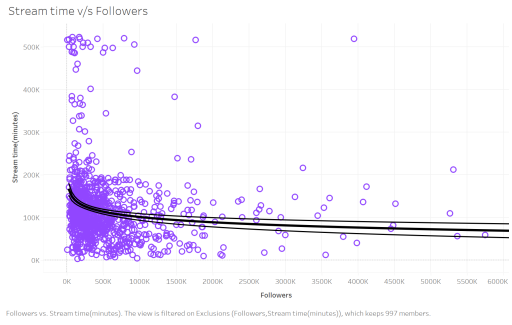

Fig. 7. Stream time vs Followers

Streamers with fewer than 500,000 followers tend to have significantly higher stream times, while those with greater follower counts exhibit shorter stream durations. This trend suggests that streamers with higher follower counts have likely already invested substantial time in growing their audience and now adopt a more comfortable streaming schedule, relying on their established viewership base.

The figure(Fig.7) is a scatter plot, with each point representing the number of hours streamed by a streamer relative to their follower count. A logarithmic curve best fits the data, indicating that streamers with lower follower counts typically invest more hours than those with a larger audience. The colours are aligned with Twitch's colour palette.

We observe that average viewership tends to decrease as stream time increases(Fig.8). This outcome is expected, as extremely long streams, such as those lasting 15 hours, are unlikely to retain viewers for the entire duration. Instead, viewers tend to rotate in and out, depending on time zones and the content being streamed during particular segments, resulting in fluctuating viewership and lower overall averages. The scatter plot(Fig.8) represents average viewership

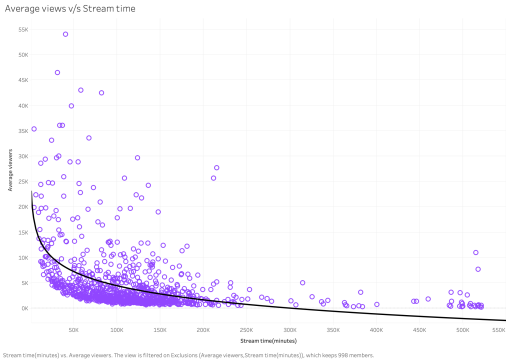

Fig. 8. Average views v/s Stream time

for different stream duration, where each point corresponds to a specific time streamed. A logarithmic trend line best fits the data, showing an initial sharp decline in viewership, which stabilizes after a certain threshold. Two outliers: dota2ti and dota2ti_ru were removed due to their low stream times but unusually high viewership. These channels stream "The International", a yearly Dota 2 tournament that is rapidly growing in viewership every year. The colour scheme follows Twitch's colour palette.-

Lastly, we observe a gradual increase in average watch time as stream duration lengthens(Fig.9). Watch time is a key metric for advertisers, who prioritize platforms where viewers are highly engaged for longer periods. This trend highlights why streamers may choose to extend their stream times in an effort to enhance monetization potential through advertising revenue.

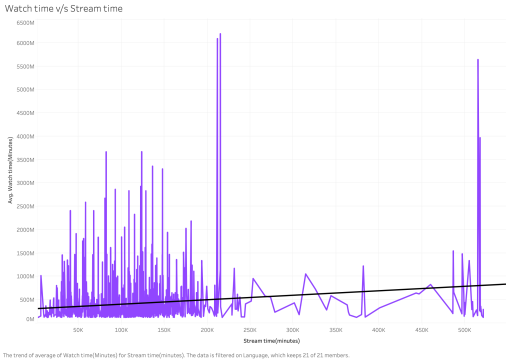

Fig. 9. Average Watch time v/s Stream time

The line plot(Fig.9) illustrates how average watch time evolves with increasing stream duration. The linear regression line shows a slight upward trend, indicating a slow but steady increase in watch time as stream length grows. The colours are based on Twitch's colour palette.

E. Observation 4

According to Twitch, streamers are labeled as mature if their content includes themes such as mature-rated games, sexual content, drug use, violence, excessive profanity, or gambling. Out of the top 1000 streamers, 230 are categorized as mature.

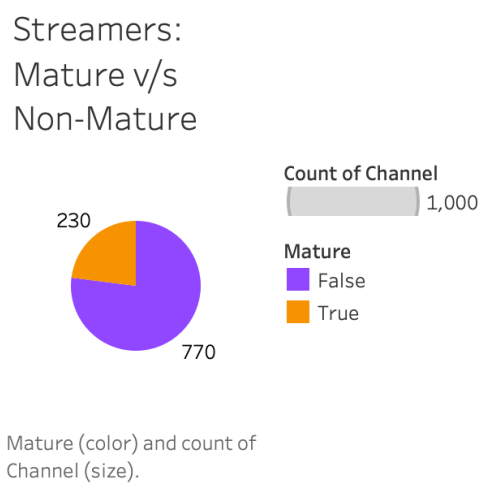

Fig. 10. Number of streamers: mature vs. non mature

The accompanying figure(Fig.10) is a pie chart illustrating the proportion of mature versus non-mature streamers, with colour coding representing each category. The size of the segments corresponds to the number of streamers in each category, and the colour scheme is consistent with Twitch's official palette.

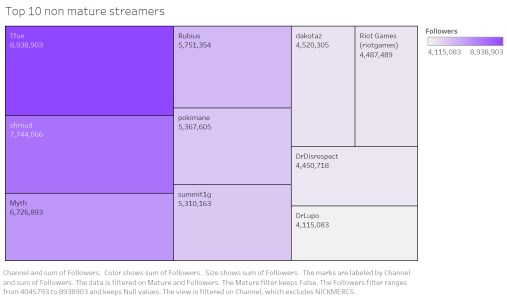

Fig. 11. Top 10 Non-mature streamers(by followers)

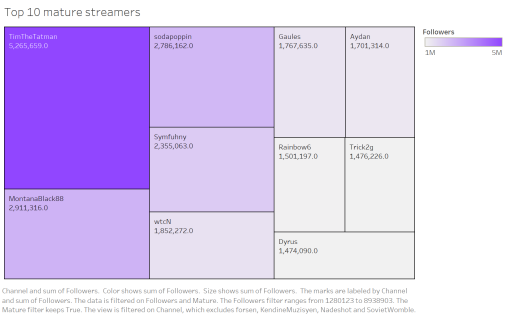

Fig. 12. Top 10 Mature streamers(by followers)

When analyzing the top streamers in both mature and non-mature categories, we find that non-mature streamers such as Tfue, Myth, Rubius, and Pokimane, initially built their following through family-friendly games like Fortnite. Although some later transitioned to games like Valorant or CS:GO, they maintained family-friendly content to retain their younger audience base.

Conversely, top mature streamers primarily focus on mature-rated games such as Call of Duty and Rainbow Six Siege, attracting older audiences. Some of these streamers also stream Valorant or CS:GO, but their target demographic is not younger viewers, which allows them more flexibility in content maturity.

Further analysis of average viewers versus followers(Fig.13) shows that around the 800,000-follower mark, the trend lines for mature and non-mature content intersect. Below this threshold, mature streamers tend to have fewer average viewers compared to their non-mature counterparts. However, once this threshold is surpassed, content maturity has little impact, and average viewership increases across the board for both categories. The accompanying scatter plot represents

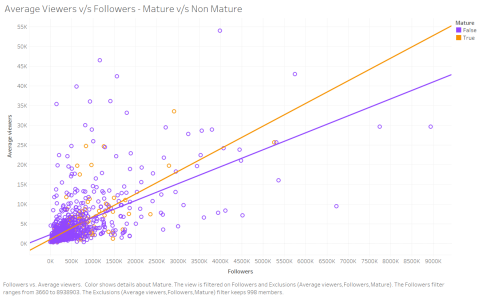

Fig. 13. Average views vs. Followers: mature vs. non mature

streamers' follower counts and their corresponding average viewer counts, with separate colours indicating mature and non-mature categories. Trend lines for both groups are shown, with two outliers (dota2ti and dota2ti_ru) excluded to prevent skewing. The colour palette adheres to Twitch's colours.

IV. T2: Growth Trends and Audience Engagement

This section examines growth trends and audience engagement metrics to uncover patterns and insights into streamer performance and viewer behavior.

A. Observation 5

The regression line in the plot of the percentage of followers(Fig.14) gained versus the number of followers indicates that smaller channels are more likely to experience a higher percentage increase in followers compared to larger channels. This suggests a saturation effect, where follower growth becomes more gradual as a streamer accumulates a larger audience. Although the general trend is downward, the distribution of data points reveals that growth stagnation is possible at any follower count. This suggests that follower gain is not necessarily guaranteed for larger channels, just as smaller streamers may still experience rapid growth.

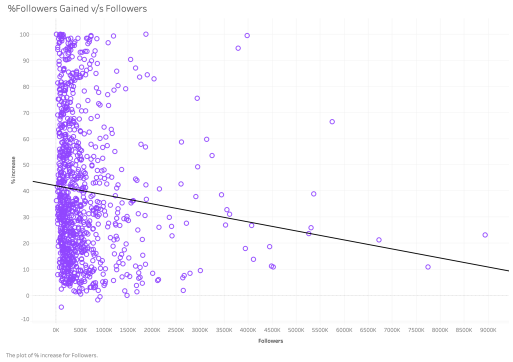

Fig. 14. Percentage Followers Gained vs Followers

The graph is a scatter plot, where each point represents a streamer's follower count and the percentage increase in followers over a year. The colour palette adheres to Twitch's official branding.

Contrary to expectations, the trend for average followers gained versus stream time(Fig.15) is also downward. This suggests that streaming for longer durations does not necessarily correlate with an increase in followers. From prior analysis, we know that smaller streamers tend to stream for longer hours. However, follower growth depends more on expanding the viewer base than on simply increasing stream time.

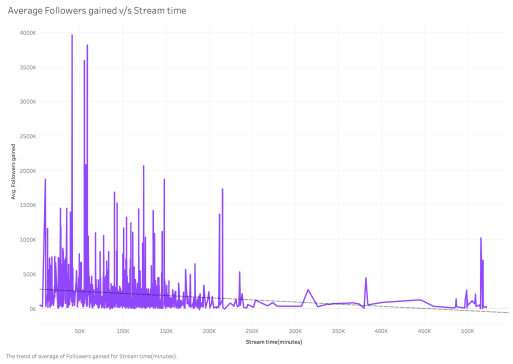

Fig. 15. Average Followers Gained vs Stream Time

This line chart represents the average number of followers gained for specific stream times. The colour scheme is based on Twitch's palette.

B. Hypothesis 2

We hypothesize that viewers do not spend significantly more time on channels with higher follower counts. This may be because larger channels often have a broader and more transient audience, with viewers cycling in and out more frequently. Additionally, smaller channels might cultivate a more engaged and loyal viewer base, leading to longer watch times per viewer. Factors such as highly interactive content or personal connections with the audience in smaller channels could contribute to this trend, whereas larger streamers may experience more passive engagement due to the size and diversity of their audience.

In fact, the data(Fig.16) shows that average watch time per viewer decreases steadily as follower numbers increase. We can also see the distribution at any number of followers is quite wide, meaning engaging content is agnostic to follower counts.

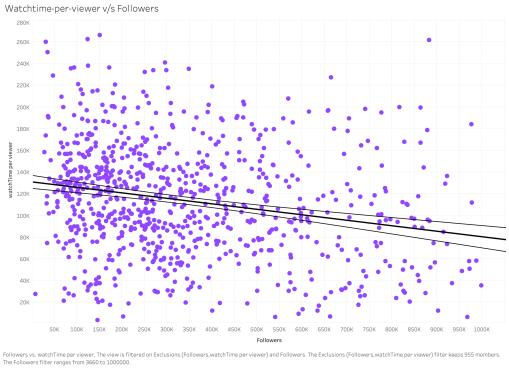

Fig. 16. Average Watch time per viewer in a Year vs Followers

Average watch time per viewer was calculated by dividing the average watch time by the average number of viewers. This data was plotted on a scatter plot, with the dots representing individual channels. The colour scheme, based on the Twitch palette, enhances visual clarity. Forty-four outliers with exceptionally high per viewer watch times were removed from the dataset to avoid skewing results. These outliers could be due to several factors like:

- Exceptionally engaging or interactive content
- Viewer incentives, such as Twitch Drops
- Alternative strategies that, while compliant with Twitch's Terms of Service, do not reflect typical viewer behavior

Similar to Observation 2, Category 4 streamers were excluded to prevent the plot from being skewed. The trend observed is likely to be consistent across the platform.

C. Observation 6

The plot examining the relationship between average views gained and stream time(Fig.17) reveals little to no significant correlation between the two variables. This suggests that merely increasing the number of hours a streamer broadcasts does not directly result in an increase in viewership. This insight highlights the limitation of relying solely on stream duration as a strategy for growth, implying that factors other than stream time, such as content quality, audience engagement, and external promotion play a more critical role in driving viewership. The graph is a line plot, with the line representing the average views gained corresponding to specific stream durations. One outlier, Fextralife, was excluded due to its anomalous data: 670,137,548 views gained with only 147,885 minutes streamed. This channel has faced accusations of inflating its viewership by embedding its Twitch stream within its external website, which features wikis, guides, and reviews for popular games. This contributed to an inflated view count that skews statistical analysis. The colour palette for the plot aligns with Twitch's colours.

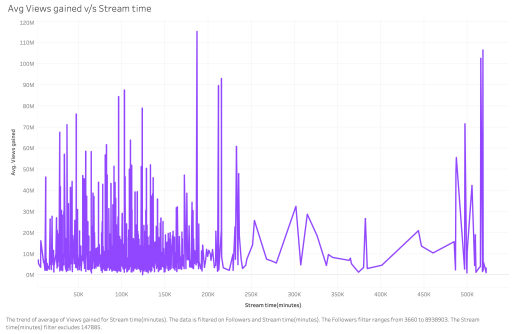

Fig. 17. Average Views Gained vs Stream time

V. T3: Language-Specific Insights

The content viewers engage with on Twitch varies significantly depending on the language of the stream. Language plays a crucial role in shaping audience preferences, engagement levels, and community dynamics. This section explores language-specific insights, highlighting key trends and differences in viewer behavior across various language groups on the platform.

A. Observation 7

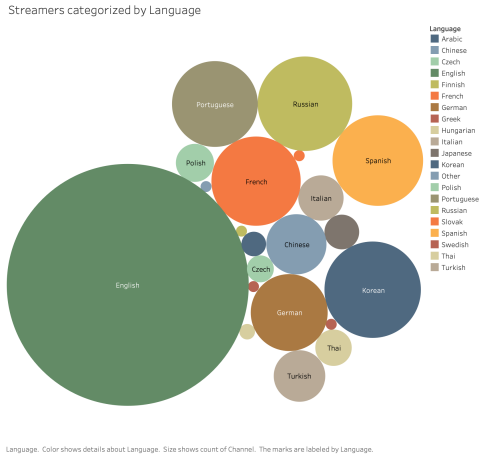

Fig. 18. Streamers: Categorized by Language

The majority of streamers, accounting for 48.5%, conducted their broadcasts in English, followed by Korean at 7.7%, Russian at 7.4%, and other languages as depicted in the diagram (Fig.18). To visually represent the distribution of languages, we employed a bubble chart, where the size of each bubble corresponds to the proportion of streamers using that language. The colour scheme for the chart was selected from the Tableau colour palette to ensure clarity and visual consistency.

B. Observation 8

Let us examine the top ten streamers by followers within each language category to better understand viewing preferences across different linguistic demographics.

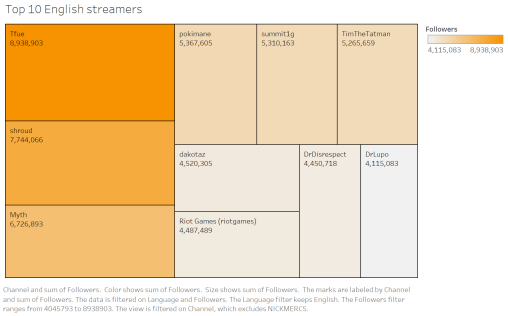

Fig. 19. Top 10 Streamers: English

In the English-speaking community, the leading channels predominantly feature streamers who gained prominence through Fortnite. Additionally, League of Legends enjoys significant popularity, with streamers of Valorant and CS:GO also ranking highly.

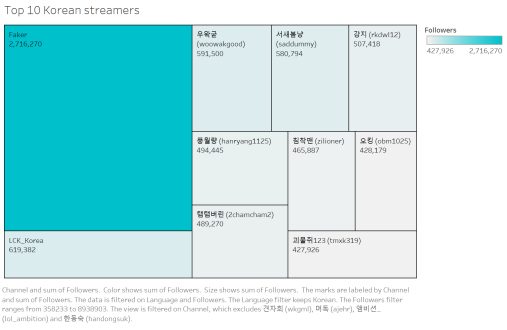

Fig. 20. Top 10 Streamers: Korean

In the Korean market, Faker, a renowned professional League of Legends player, holds a commanding lead in follower count. The top channels also include Just Chatting streamers, many of whom engage with League of Legends content.

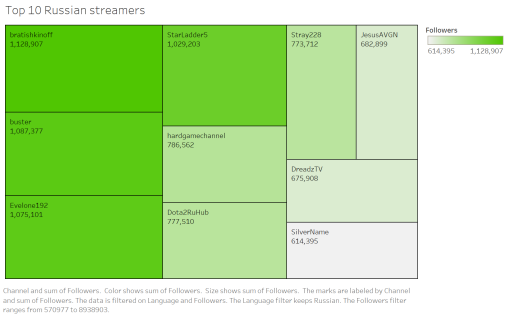

Fig. 21. Top 10 Streamers: Russian

For the Russian audience, Counter-Strike streamers dominate the top ranks, with Just Chatting content also holding substantial viewer interest. Dota 2 remains a significant draw within this demographic.

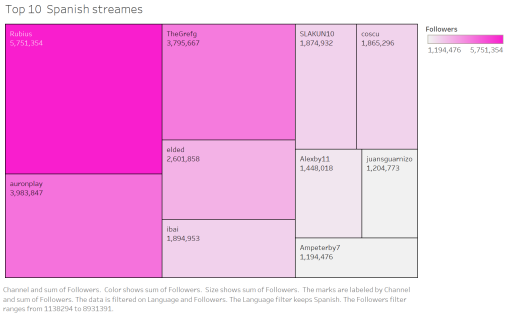

Fig. 22. Top 10 Streamers: Spanish

In the Spanish-speaking community, Fortnite and Minecraft are the most popular among the top streamers. Additionally, content related to League of Legends and Just Chatting also garners considerable attention.

C. Observation 9

We plotted the 'loyalty' of followers of different content(Fig.23,24). The loyalty of followers is higher if they have a higher average number of views for the same number of followers. This loyalty could be compared by comparing the slopes(and the intercepts) of the followers vs average views.

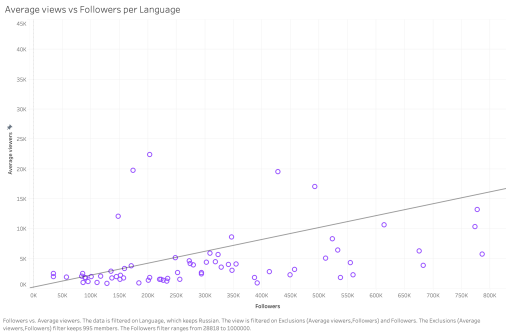

Fig. 23. Average Views vs Followers: Russian

The graph(Fig.23) illustrates that Russian streamers exhibit a steeper slope compared to their counterparts in other languages, suggesting a greater degree of viewer loyalty and engagement within the Russian-speaking audience.

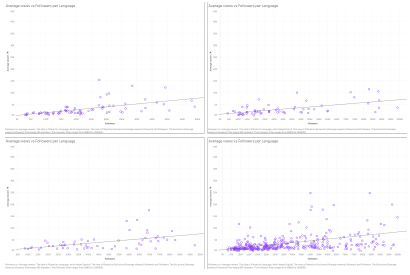

Fig. 24. Top Left - Korean, Top Right - French, Bottom Left - Spanish, Bottom Right - English

D. Observation 10

In the top 5 most popular languages(based on the number of streamers) Fig.25 shows that English content has more mature streamers than any other content, followed by French, Spanish, Russian, and Korean.

Korean content has fewer mature streamers than any other content.

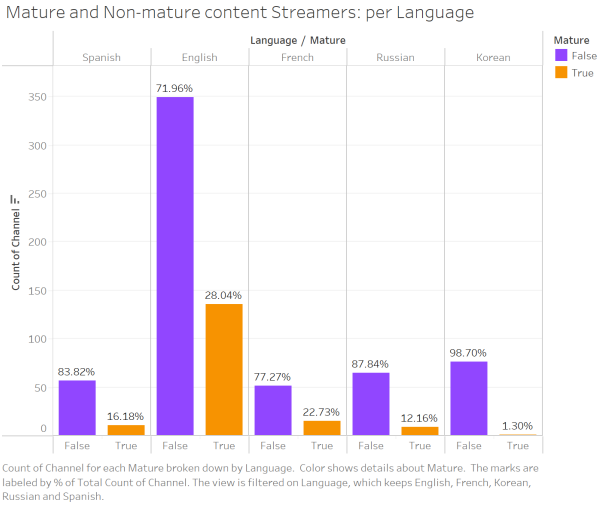

Fig. 25. Percentage Of Mature vs Non-mature streamers: categorized by language

E. Observation 11

From(Fig.26) we can tell that Spanish followers grew the most (concerning the percentage followers gained), followed by Italian followers and Portuguese followers.

Fig.26 also depicts the percentage followers gained concerning mature and non mature content per language.

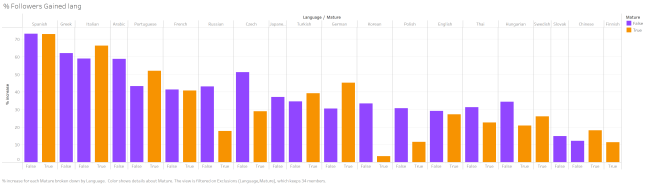

Fig. 26. Percentage Followers gained: categorized by Language

VI. Visualizations Used

We've used plenty of visualizations in the proper context. The visualizations used are:

1) Scattered Plots
2) Line Plots
3) Tree Maps
4) Pie charts
5) Bar Graphs
6) Bubble Graphs

VII. Contributions

The division of tasks was collaboratively decided, and responsibilities were allocated as follows:

1) Information and General Formatting: This task was undertaken collaboratively by all members to ensure consistency and coherence throughout the document.
2) T1 & T2: Tahir and Pradyun jointly conducted the analysis and visualizations for these sections. Our exploration of the data made more sense to do collaboratively, as they were closely linked.
3) T3: Narayana focused on this task independently. The insights from this section were integrated with the rest of the project to form a cohesive final report.