CMP 7203: Big Data Management

Final Assessment

Introduction to Big Data

Big Data Ecosystem: Catch The Pink Flamingo

Sachin

Student Id: 23235298

BIRMINGHAM CITY University

Faculty of Computing, Engineering, and The Built Environment

School of Computing and Digital Technology

Table of Contents

1. Introduction

In the 21st century, big data grows rapidly with its significance and importance in the market, although research goes up to 8Vs and 10Vs, the Primary focus was always on 5Vs as discussed below and visualized in Figure 1.

1.1 Vs in Big Data

1. **Volume**: Volume refers to a large amount of data and expands continuously, more data means we can pull out more insights.
2. **Velocity**: Velocity refers to the speed of data being fetched, whether it is real-time or batch process.
3. **Variety**: Variety of data means how diverse your data is, structured, unstructured, or semi-structured along with the different sources it is coming from.
4. **Veracity**: Veracity refers to the quality of the data, if the quality is not good analysis will lead to misinformational insight.
5. **Value**: Value refers to the data which are having some useful information, it doesn't have dummy data or information that is not useful for analysis.
   (Saeed, 2021)

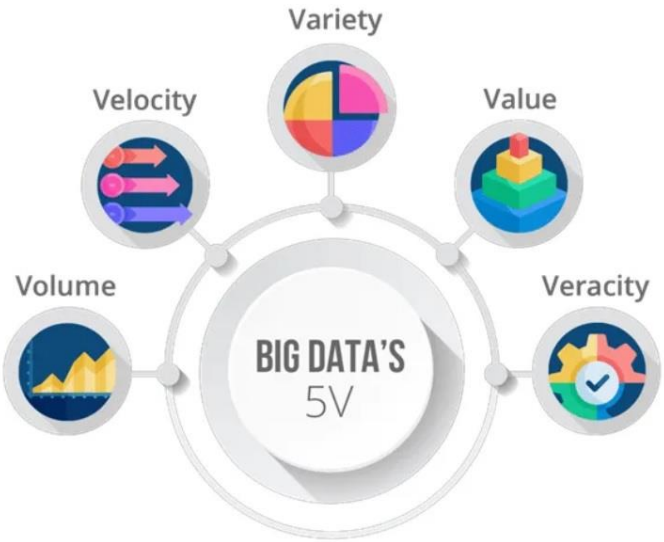

Figure 1: Big Data's Five Vs

Image Source: Google Images

1.2 Big Data Processing Paradigm

Mainly there are three popular paradigms as discussed below, Batch Processing, Real-Time Processing and Hybrid Processing

1.2.1 Batch Processing

Let's understand with an example, Previously when data was not that big, processors processed the data smoothly, and efficiently on time, After some time when data became big data the problem started for the processor, and then the data was divided into groups to perform separately on, This way processing becomes faster and efficient with big data as shown in Figure 2.

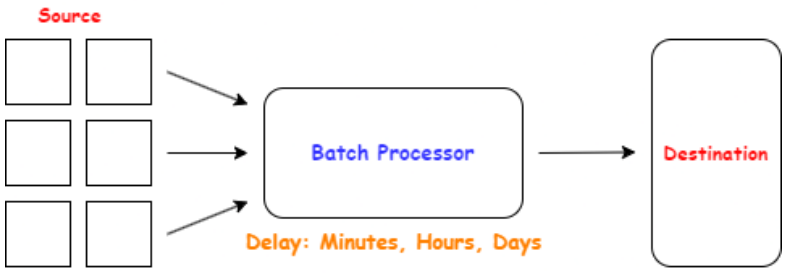

Figure 2: Batch Processing

1.2.2 Real-time Processing

Real-time processing analyses live data unlike batch processing, there are many applications or use cases in the world that require real-time processing for example air ticket booking systems, online streaming platforms, online game streaming, etc. In this process data is being processed in real-time to take action on the output effectively, this is the best example of velocity from the 5Vs of big data as shown in Figure 3.

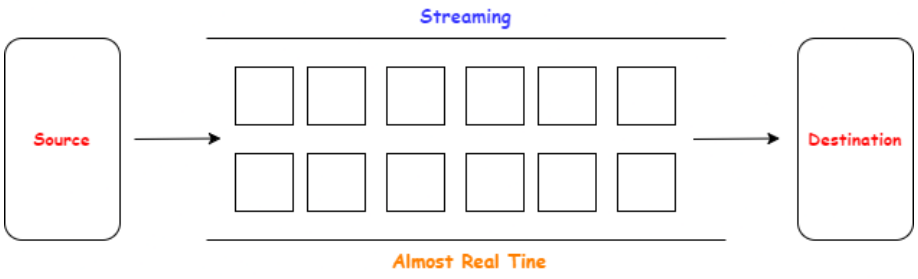

Figure 3: Real-Time Processing

1.2.3 Hybrid Model

Many applications or use cases require a combination of batch processing and real-time processing, this architecture is known as lambda architecture which consists of batch layer, serving layer, and speed layer as shown in Figure 4.

The batch layer, also known as batch processing handles the main dataset that remains unchanged and stored in a distributed file system. These data are processed offline in big batches. The batch layer is responsible for computing the dataset and generating results.

The Serving layer is responsible for managing results generated by the batch layer. It stored batch views in a datastore from where it can be easily queried whenever it is required.

The speed layer, dedicated to real-time processing and handling incoming data streams with low latency, updates data in real-time when it comes. The speed layer handles batch data and real-time data simultaneously.

To obtain optimal results it is better to query both batch and real-time, this process is synchronized and has the potential to solve several complex data problems

(Casado, 2015)

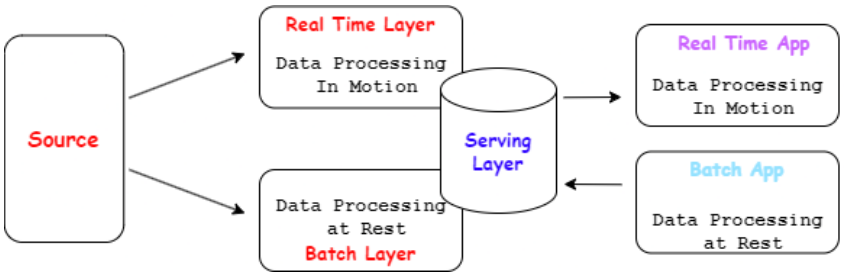

Figure 4: Hybrid Processing

2. Different organizations and use cases using big data paradigms.

Companies Names	Batch Processing	Real-Time Processing	Hybrid Processing
Amazon	Inventory management and order fulfillment	Real-time recommendation engine for personalized products.	Fraud detection and real-time inventory updates.
Google	Search indexing, and data analytics.	YouTube processes live video streams.	AdWords platform for ad campaign optimization.
Facebook	Data warehousing and analytics.	Real-time notifications and news feed.	User engagement analytics.
Netflix	Content recommendation algorithms.	Real-time video streaming service.	Personalized content delivery.
Uber	Financial reporting, driver payouts.	Real-time ride tracking and surge pricing.	Route optimization.
Twitter	Trend analysis, User activity metrics.	Real-time tweet delivery and notifications.	Sentiment analysis.
Airbnb	Pricing optimization	Real-time booking system and notifications.	Dynamic pricing
Salesforce	Data synchronization, reporting	Real-time sales alerts and lead scoring	Customer insights
Spotify	Music recommendation algorithms	Real-time music streaming service	Personalized playlists
Microsoft	Data warehousing, large-scale data transformations	Real-time data ingestion and processing (Azure Stream Analytics)	Event-driven applications and analytics.
IBM	Payroll and Billing	Real-time data analytics (IBM Streams)	Data integration and analytics solutions
Oracle	Data loading, ETL, reporting (Oracle Database)	Real-time data processing (Oracle Stream Analytics)	Complex event processing and real-time analytics
Apple	Batch processing for app store analytics	Real-time notifications and updates for app users	Hybrid approach for personalized recommendations
Tesla	Batch processing for vehicle diagnostics and software updates.	Real-time monitoring of vehicle performance and safety.	Hybrid approach for autonomous driving algorithms.

Table 1: Organizations use big data paradigms

3. Future role of big data in the era of LLMs

In the era of rapidly changing Artificial Intelligence, Big data, and large language models combined can change the AI-driven applications in the future. When a large language model is on the way to growing day by day which means training on datasets then big data concepts become so important in LLM. This frontier promises to open new ways in natural language processing, code comprehension, etc, let's explore the potential of big data in the era of large language models (LLMs)

3.1 BitNet b1.58 and the Dawn of Efficient 1-Bit Large Language Models

BitNet b1.58, a 1-bit large language model, provides full efficiency performance with latency, memory, and energy usage. It shows the shift towards transition cost-effective and scalable large language models (LLMs) which handle extensive data with high computational cost. Organizations can boost natural language processing tasks, streamline model deployment, and unlock data-driven insights using 1-bit LLMs which integrate AI with big data analytics and promote innovative applications (Ma, 2024).

3.2 The Influence of Big Data on Large Language Models (LLMs)

Big data shapes the large language models significantly which enables adaptability with extensive dataset training. Even with challenges like bias mitigation and scalability, big data appreciates new applications like multimodal understanding and performance enhancement in low-resource languages, which influence the potential of large language models and provide a way of innovation and progress in the domain of AI-driven language understanding and generation (Zhou, 2024).

3.3 Transforming AI Mathematics: The Fusion of Big Data and LLM

Big data and large language models are integrated to enhance Human-Computer Mutual Assistance (HCMA) which includes a combination of large language models and Mathematical sequencing and positioning systems (MSPS). The motive of this process is to enhance machine learning's opaque nature to be transparent to improve AI mathematics. Big data is used to train large language models which helps to generate new texts. This helps to enhance large language models' natural processing tasks and in the way of artificial intelligence success (Zou, 2023).

3.4 Assessing Large Language Models' Code Reading Abilities in Big Data

In big data, large language model code comprehension is assessed for metamorphic testing which shows the strengths and limitations. Using LLM as different inputs like modified code snippets, shows the weaknesses that enhance the utility in pivotal coding scenarios. Using Metamorphic relations, several test cases challenge large language models to generate and understand the codes to ensure big data maintains robustness in a big data environment. This research focuses on navigating LLMs in big data complexity and gives the power to comprehend dynamic codes in data landscapes in real-world applications. (Li, 2023).

4. Introduction to the Big Data Ecosystem

A big data ecosystem has been established using the Catch the Pink Flamingo dataset. Starting with the Exploratory Data Analysis (EDA) to understand the characteristics of the dataset. Machine Learning models have been used for classification and clustering. For classification techniques Decision Trees and Support Vector Machines have been used to effectively group the data and for clustering techniques K-Means and Gaussian Mixture Models have been used to reveal the hidden patterns in the dataset. In the end, to understand interconnections and relationships better, neo4j has been used for graph analytics. The motive behind this strategy is to find meaningful insights from the Catch the Pink Flamingo dataset.

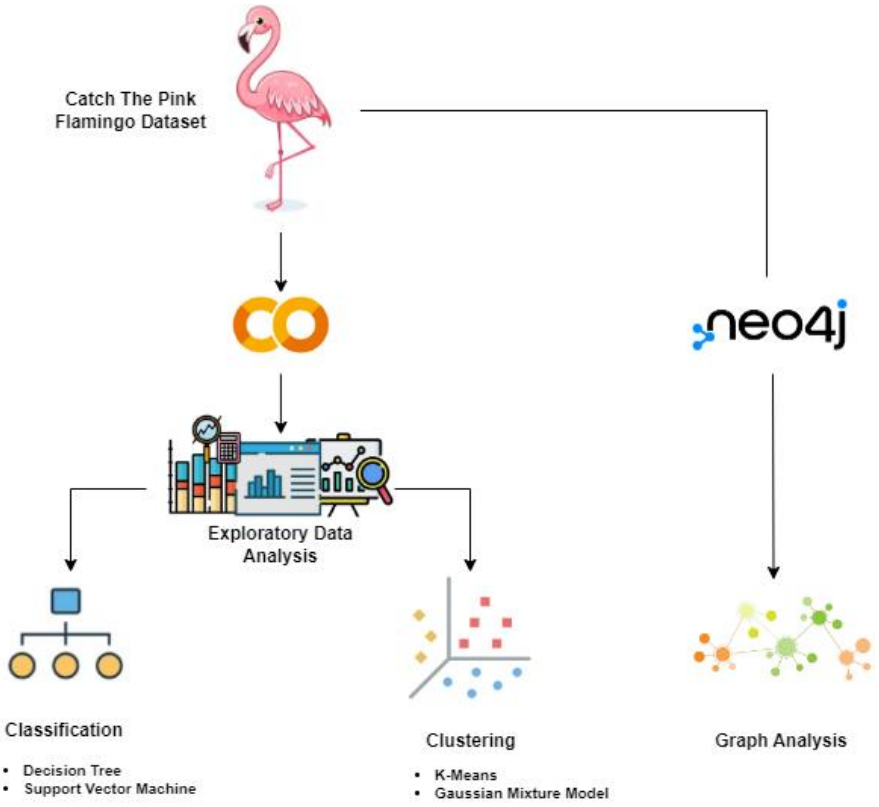

Figure 5: Complete Flow Chart of Big Data Ecosystem

5. Catch The Pink Flamingo

Catch the Pink Flamingo is a multiplayer game developed by Eglence Inc. where players have to capture pink flamingos from different locations on the map. Players progress through the levels where with increasing levels they have to face bigger missions and maps. This game promotes teamwork as well where players can create or join teams. Each level introduces new missions and maps, players have to adapt and make strategies to win. Players can communicate through chat boards or social media platforms like Twitter. Level 1 is like a tutorial that explains the foundation of the gameplay, The Player will collect points for the correct catch of the flamingo while points are deducted for the wrong catching. The dynamic gameplay and interactive features of this game made it engaging.

5.1 Data Description

Out of the 15 dataset files, 9 will be used for exploratory data analysis (EDA), labelled *ad-clicks.csv, buy-clicks.csv, users.csv, team.csv, team- assignments.csv, level-events.csv, user-session.csv, game-clicks.csv, and combined_ data.csv* as discussed in Table1.

The remaining 6 files, labelled *chat_create_team_chat.csv, chat_item_team_chat.csv, chat_join_team_chat.csv, chat_leave_team_chat.csv, chat_mention_team_chat.csv, and chat_respond_team_chat.csv* will be used for graph analysis as discussed in Table2.

This dataset description table 1 & 2 is sourced from: https://eagronin.github.io/capstone-acquire/

File Name	Description	Fields
ad-clicks.csv	A line is added to this file when a player clicks on an advertisement in the Flamingo app.	**timestamp:** When the click occurred. **txId:** A unique id (within ad-clicks.log) for the click. **userSessionid:** The id of the user session for the user who made the click. **teamid:** The current team id of the user who made the click. **userid:** The user id of the user who made the click. **adId:** The id of the ad clicked on. **adCategory:** The category/type of ad clicked on.
buy-clicks.csv	A line is added to this file when a player makes an in-app purchase in the Flamingo app.	**timestamp:** When the purchase was made. **txId:** A unique id (within buy-clicks.log) for the purchase. **userSessionId:** The id of the user session for the user who made the purchase. **team:** The current team id of the user who made the purchase. **userId:** The user id of the user who made the purchase. **buyId:** The id of the item purchased. **price:** The price of the item purchased.
users.csv	This file contains a line for each user playing the game.	**timestamp:** When user first played the game. **userId:** The user id assigned to the user.

		nick: The nickname chosen by the user. **twitter:** The twitter handle of the user. **dob:** The date of birth of the user. **country:** The two-letter country code where the user lives.
team.csv	This file contains a line for each team terminated in the game.	**teamId:** The id of the team **name:** The name of the team. **teamCreationTime:** The timestamp when the team was created. **teamEndTime:** The timestamp when the last member left the team. **strength:** A measure of team strength, roughly corresponding to the success of a team. **currentLevel:** The current level of the team.
team- assignments.csv	A line is added to this file each time a user joins a team. A user can be in at most a single team at a time.	**timestamp:** When the user joined the team. **team:** The id of the team. **userId:** The id of the user. **assignmentId:** A unique id for this assignment.
level-events.csv	A line is added to this file each time a team starts or finishes a level in the game	**timestamp:** When the event occurred. **eventId:** a unique id for the event. **teamId:** the id of the team. **teamLevel:** the level started or completed. **eventType:** the type of event, either start or end.
user- session.csv	Each line in this file describes a user session, which denotes when a user starts and stops playing the game. Additionally, when a team goes to the next level in the game, the session is ended for each user in the team and a new one started.	**timestamp:** a timestamp denoting when the event occurred. **userSessionId:** a unique id for the session. **userId:** the current user's ID. **teamId:** the current user's team. **assignmentId:** the team assignment id for the user to the team.

		sessionType: whether the event is the start or end of a session. **teamLevel:** the level of the team during this session. **platformType:** the type of platform of the user during this session.
game-clicks.csv	A line is added to this file each time a user performs a click in the game.	**timestamp:** when the click occurred. **clickId:** a unique id for the click. **userId:** the id of the user performing the click. **userSessionId:** the id of the session of the user when the click is performed. **isHit:** denotes if the click was on a flamingo (value is 1) or missed the flamingo (value is 0). **teamId:** the id of the team of the user. **teamLevel:** the current level of the team of the user.
combined_ data.csv	Combines data from 3 of the log files: user-session.csv, buy-clicks.csv, and game-clicks.csv.	**userid**: User ID **userSessionid:** User session ID **team_level:** User's team level **platformType:** Platform used by user **count_gameclicks:** Total number of game clicks for user session **count_hits:** Total number of game hits for user session **count_buyid:** Total number of purchases for user session **avg_price:** Average purchase price for user session

Table 2: Game Dataset Description

File Name	Description	Fields
chat_create_team_chat.csv	A line is added to this file when a player creates a new chat with their team.	userid, teamid, TeamChatSessionID, timestamp
chat_item_team_chat.csv	Creates nodes labeled ChatItems.	userid, teamchatsessionid, chatitemid, timestamp
chat_join_team_chat.csv	Creates an edge labeled “Joins” from User to TeamChatSession.	userid, TeamChatSessionID, timestamp
chat_leave_team_chat.csv	Creates an edge labeled “Leaves” from User to TeamChatSession.	userid, teamchatsessionid, timestamp
chat_mention_team_chat.csv	Creates an edge labeled “Mentioned”.	ChatItem, userid, timestamp
chat_respond_team_chat.csv	A line is added to this file when player with chatid2 responds to a chat post by another player with chatid1.	chatid1, chatid2, timestamp

Table 3: Chat Dataset Description

6. Exploratory Data Analysis

Exploratory Data Analysis means looking into a dataset closely to pull out some patterns and unique insights from it, to ensure that our guesses are right, and to summarize using numbers and graphs as discussed in (Good, 1983).

6.1 Ad-Clicks

6.1.1 Are there any missing values?

Figure 2 shows that there are no missing values present in the ad-clicks dataset.

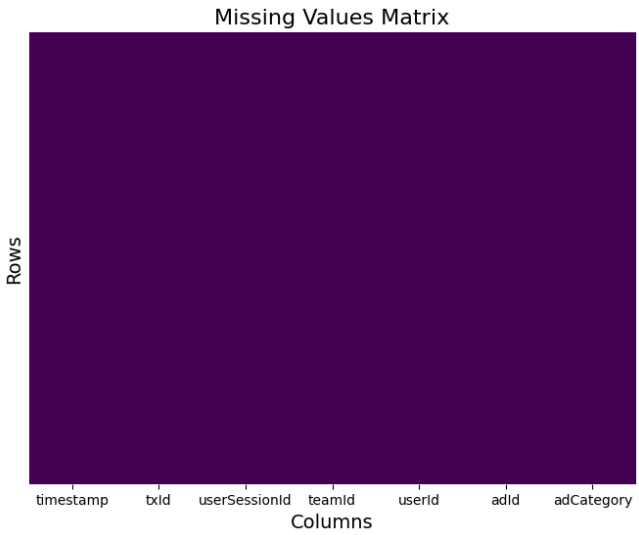

Figure 6: Missing Values In ad-clicks Dataset

6.1.2 Ad Click Trends Over Time?

The time series plot in Figure 3 depicts ad clicks over time. It shows the ups and downs of the user engagement with the advertisement. It will help to understand ad placement to optimize the application.

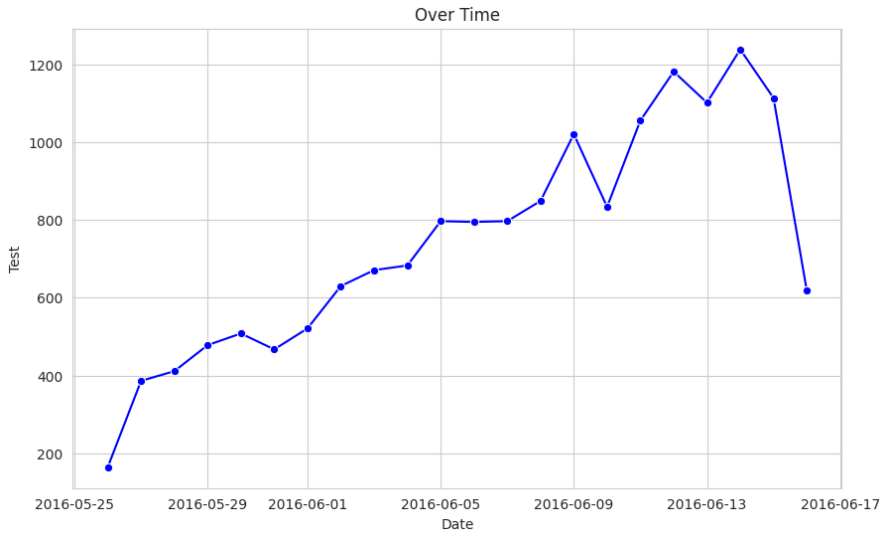

Figure 7: Depicting the daily counts of ad clicks over time.

6.1.3 Which Teams Dominate Ad Clicks?

The largest block in Figure 4 of the Treemap is team 64, which consists of a 684 ad-clicks count. This Treemap shows the distribution of ad click count between the top 10 teams.

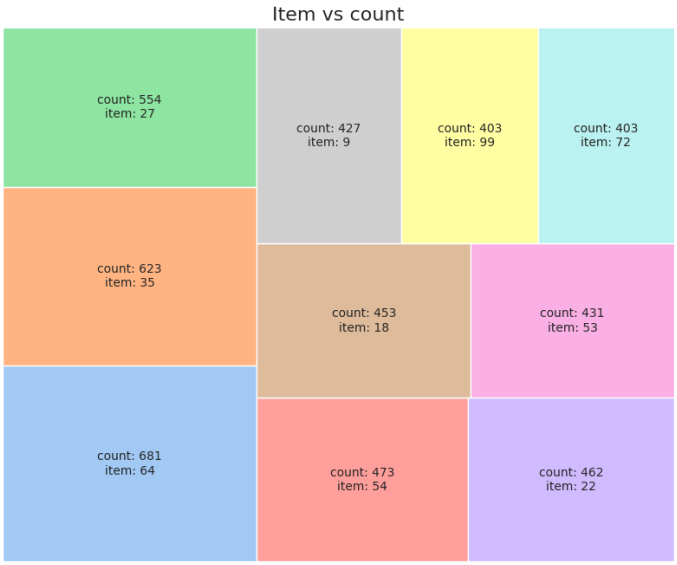

Figure 8: Ad click distribution among top teams

6.1.4 Distribution of Ad Categories Across Ad-Ids

The scatter plot in Figure 5 illustrates the notable ad-ids concentration in the 'Computers' and 'Games' categories. Which shows that ads in these categories are frequent.

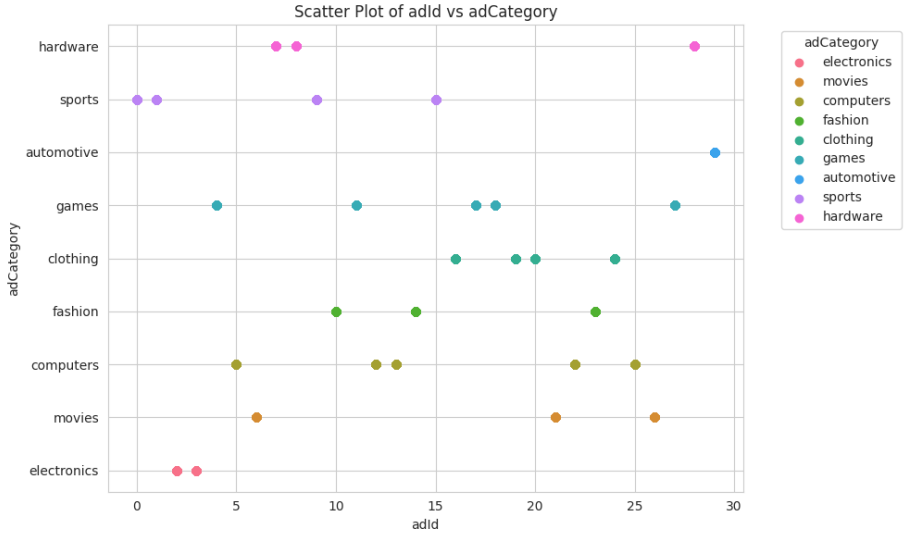

Figure 9: Distribution of ad Categories

6.1.5 Distribution of Session Durations

In Figure 6, the Histogram depicts the distribution of session durations. Each bar represents a range of session duration and different colours represent several frequencies.

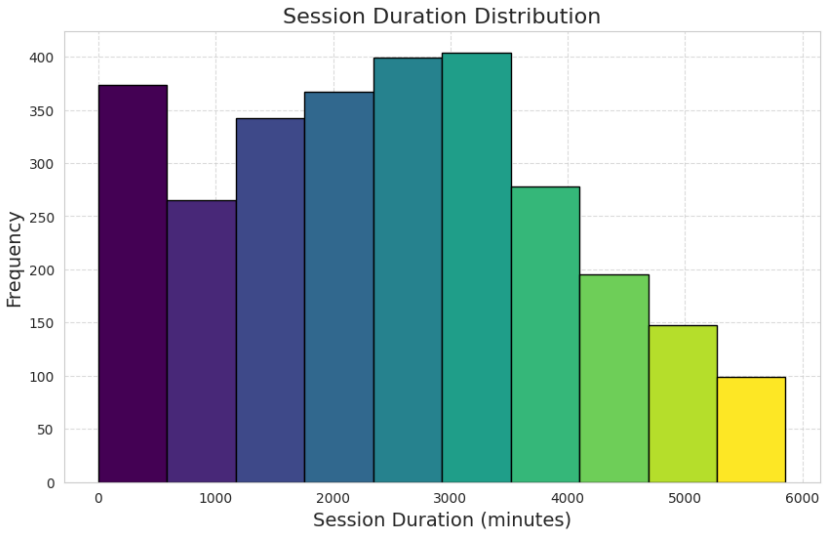

Figure 10: Session Duration Distribution

6.1.6 Distribution of Ad Activity

The bar plot in Figure 7 shows that ad-id 20 attracts the most activity which surpasses 600 clicks. This insight shows the most engaging advertisement.

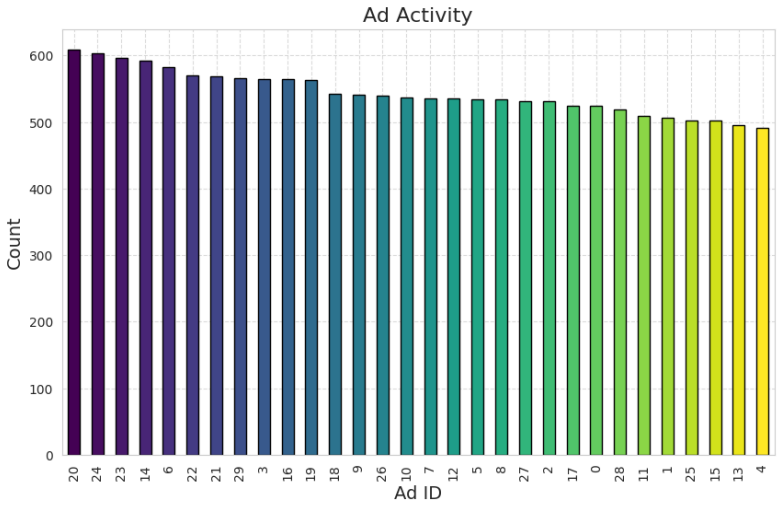

Figure 11: Ad Activity Distribution

6.1.7 Distribution of Ad Categories

The bar plot in Figure 8 illustrates the distribution of ad categories. Each bar represents a unique ad category. It shows the diversity of ad types. It is clear that 'Computers' and 'Games' categories are prominent and having more than 2500 counts.

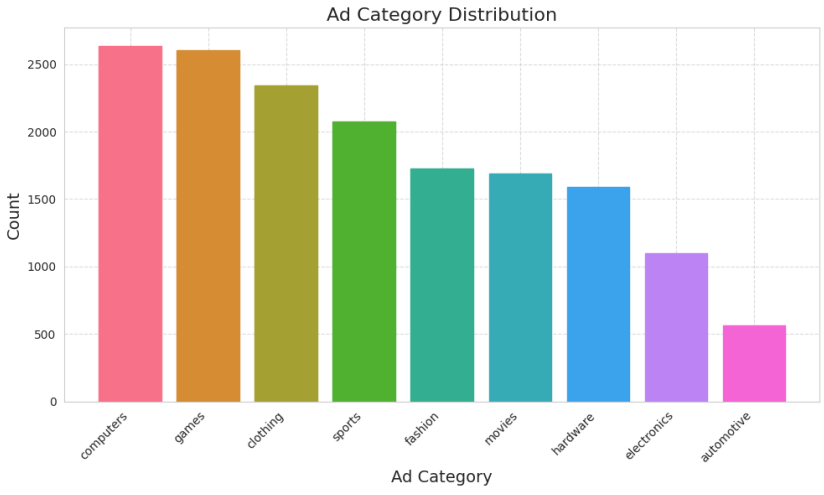

Figure 12: Ad Category Distribution

6.2 Buy-Clicks

6.2.1 Are there any missing values?

As shown in Figure 9, There are no missing values present in the Buy-Clicks dataset.

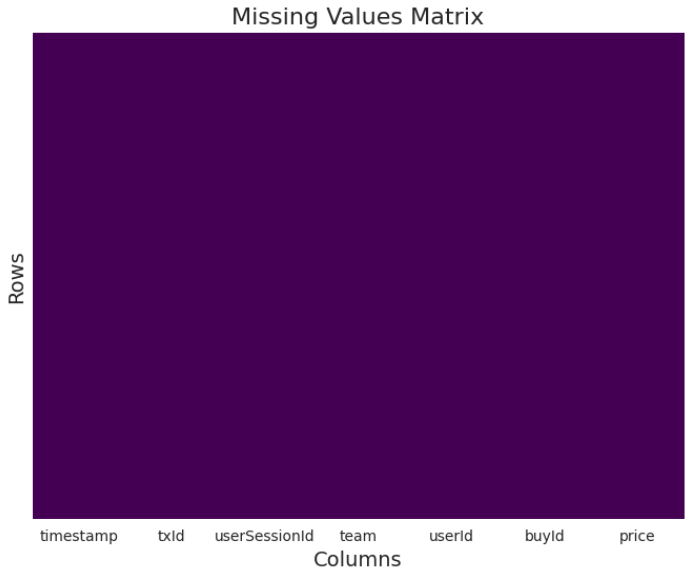

Figure 13: Missing Values in Buy-Clicks Dataset

6.2.2 Daily Counts of In-App Purchases Over Time

The time series plot in Figure 10 talks about daily in-app purchase counts over time. It helps to assess user engagement in purchase activity.

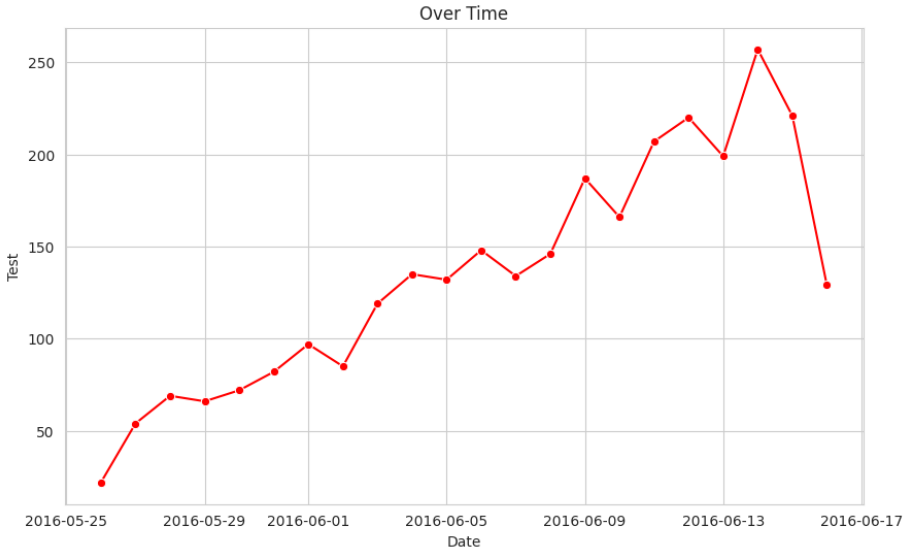

Figure 14: Depicting the daily counts of buy-clicks over time.

6.2.3 Top 20 Teams with the Highest Purchase Counts

The histogram in Figure 11 depicts the top 20 teams having the most purchase counts. This shows the purchasing behavior of different teams. Team 27 has more than 100 purchase count and 64 is almost 100.

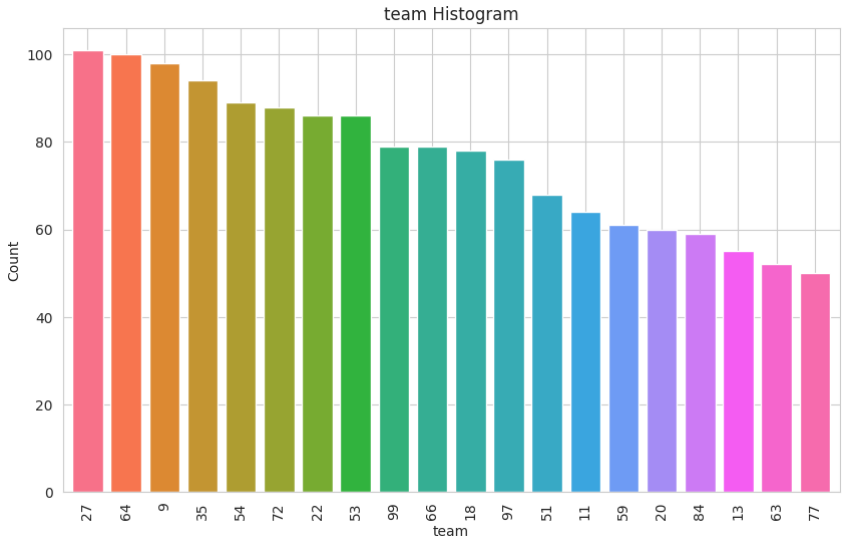

Figure 15: Top team's purchase counts distribution

6.2.4 Correlation Analysis

The correlation plot in Figure 12 shows the relationship between team, buy id, and price. No significant relationship can be seen between the team and other variables while buy ID and price have a strong positive correlation.

Figure 16: Correlation Analysis: Team, Buy ID, and Price

6.3 Game-Clicks

6.3.1 Are there any missing values?

There are no missing values present in the Game-Clicks dataset as shown in Figure 13.

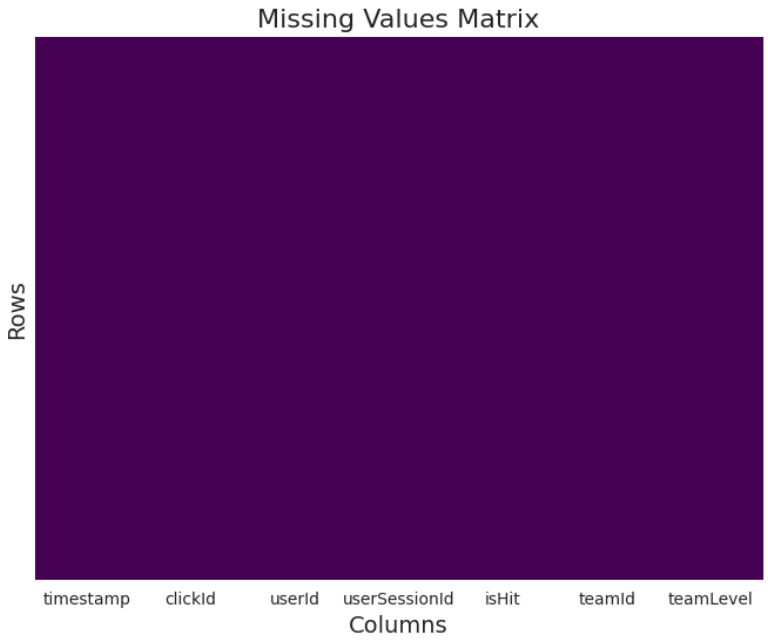

Figure 17: Missing Values in Game-Clicks Dataset

6.3.2 Daily Click Counts in the Game Over Time

In Figure 14, the Time series plot shows the daily click count in the game over time.

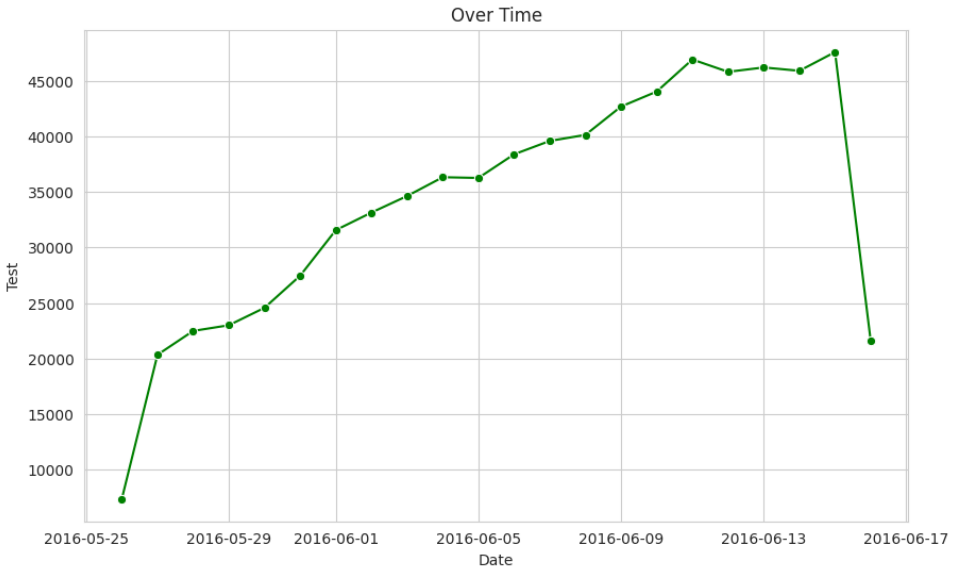

Figure 18: Daily click counts in the game, depicted over time.

6.4 Level Events

6.4.1 Are there any missing values?

There are no missing values present In the Level Events dataset as shown in Figure 15.

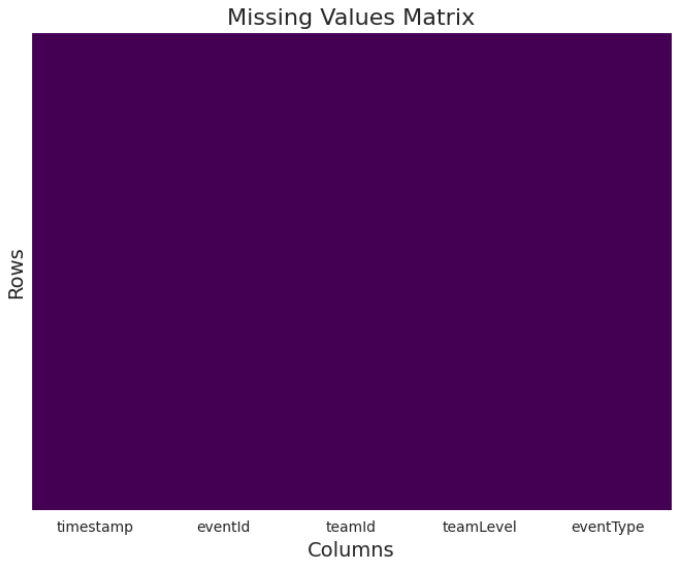

Figure 19: Missing Values in Level Events Dataset

6.4.2 Daily Level Events Over Time

Figure 16 shows the time series plot of daily-level events over time. It shows the user progression and activity engagement.

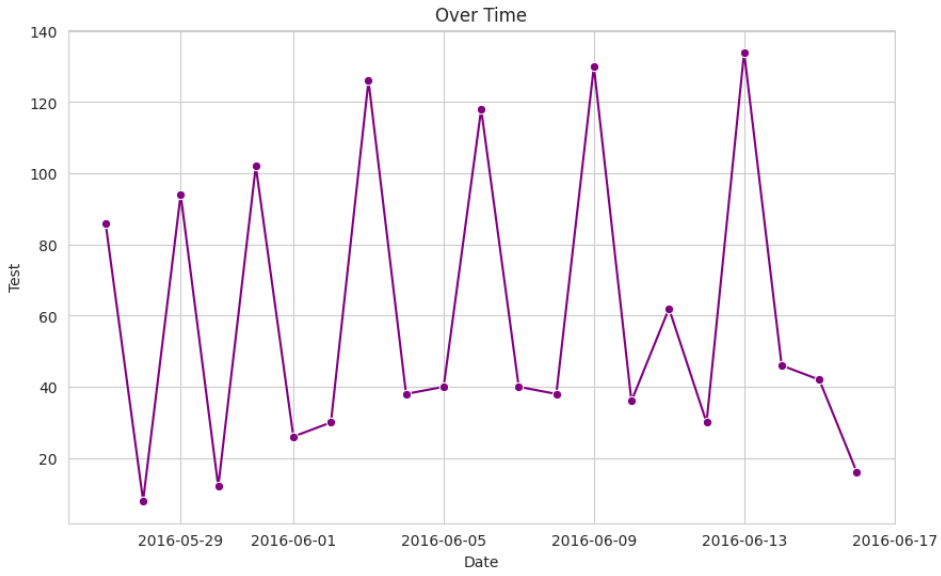

Figure 20: Daily-level events depicted over time in the game.

6.5 Team Assignment

6.5.1 Are there any missing values?

No missing values are present in the Team Assignment dataset as shown in Figure 17.

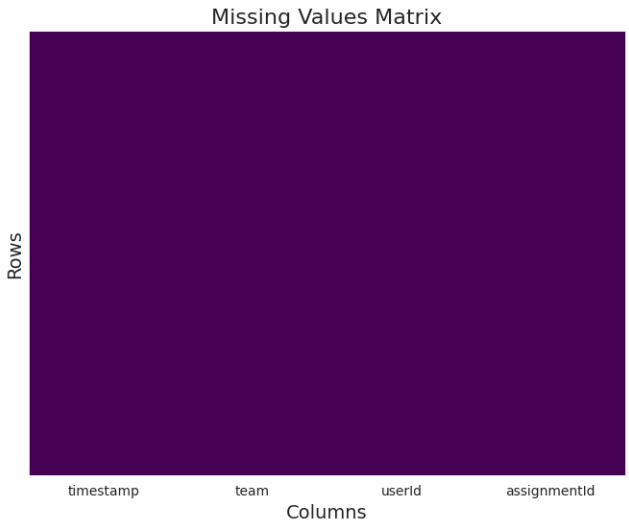

Figure 21: Missing Values in Team Assignments Dataset

6.5.2 Daily Team Assignments Over Time

The time series plot shown in Figure 18 depicts the daily team assignment count over time. It shows the user engagement with team-related activities.

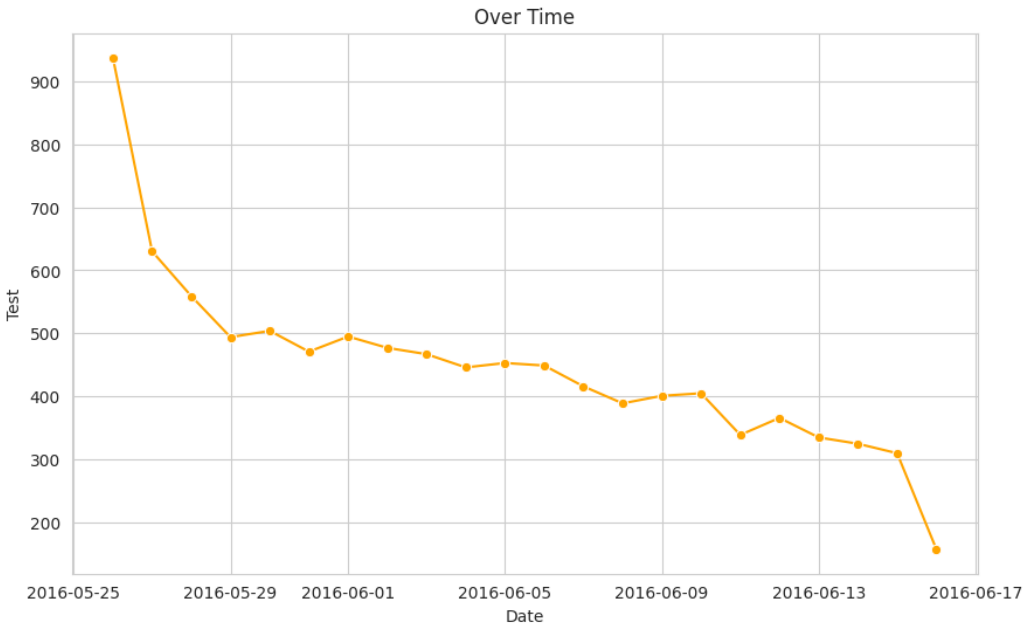

Figure 22: Daily team assignments depicted over time in the game

6.6 Team

4.6.1 Are there any missing values?

There are no missing values present in the Team dataset as shown in Figure 19.

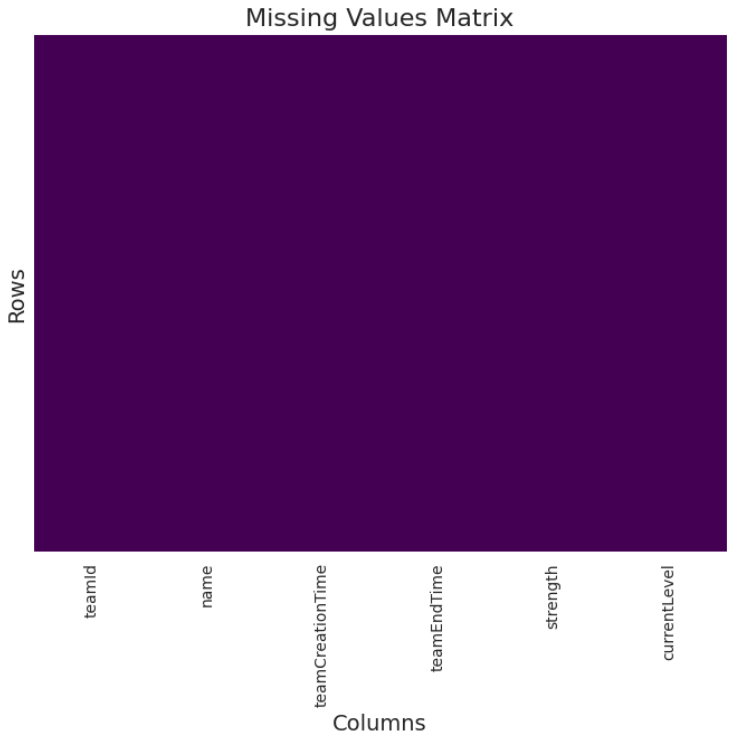

Figure 23: Missing Values in Team Dataset

4.6.2 Daily Team Creation Over Time

The time series plot shown in Figure 20 illustrates the daily team creation count in the game over time.

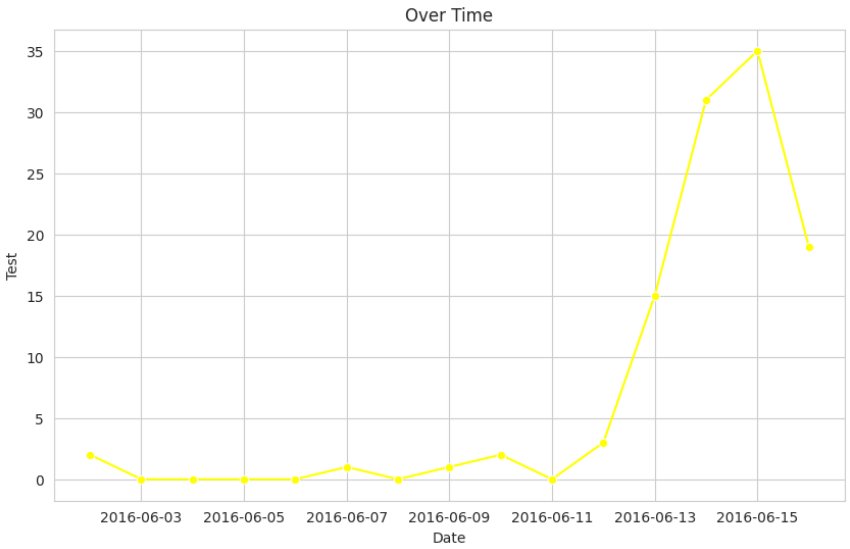

Figure 24: Daily Team Creation Over Time

4.6.3 Top 20 Teams by Strength

Figure 21 shows the strength analysis. Interestingly, some powerful teams are not big. Team 9 which lies on ranks 3 on spending and strength despite being 4th largest in terms of team member

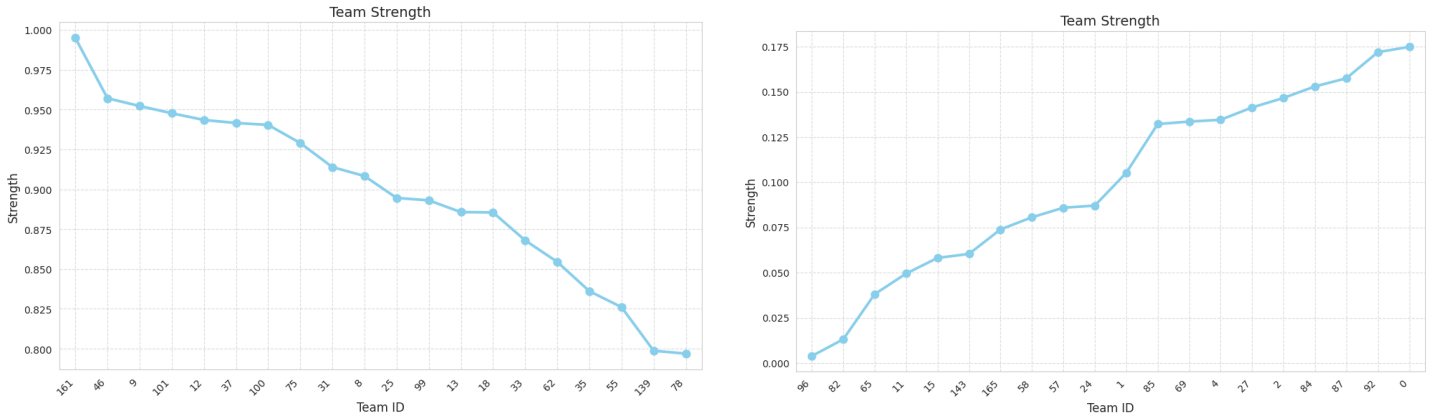

Figure 25: Team Strengths

6.7.1 Are there any missing values?

There are no missing values present in the Teams dataset as shown in Figure 22.

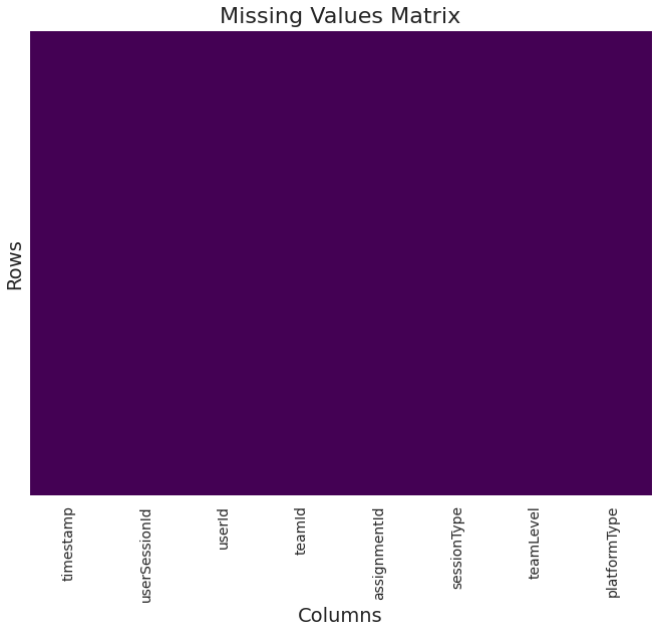

Figure 26: Missing Values in User Session Dataset

6.7.2 Daily User Session Trends

In Figure 23, the Time series plot shows the fluctuations of user sessions over time. Patterns suggest that there could be possible seasonal trends.

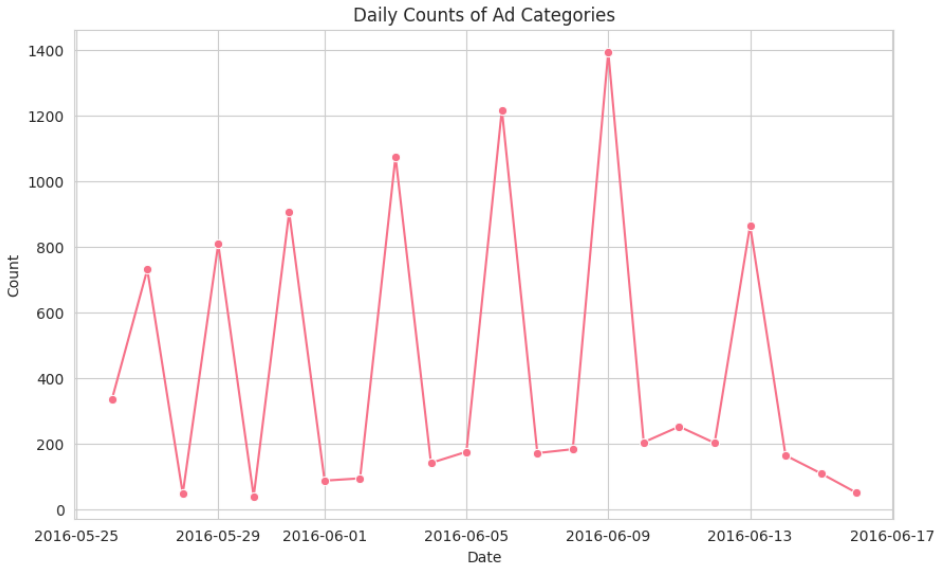

Figure 27: Daily User Session Over Time

6.7.3 Distribution of Session Types by Platform

Figure 24 depicts the stacked histogram showing session-type frequencies on different platforms. Mac users have maximum start or end sessions which shows a significant user base. This insight can help the marketing team to target Mac users.

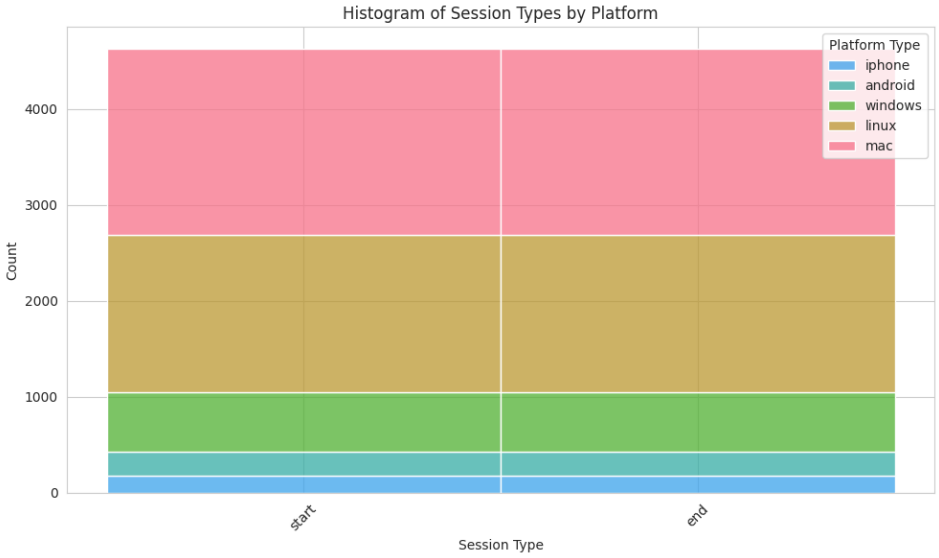

Figure 28: Session Types by Platform

6.8 Users

6.8.1 Are there any missing values?

Figure 25 shows that a few data are missing in a country column of the user dataset.

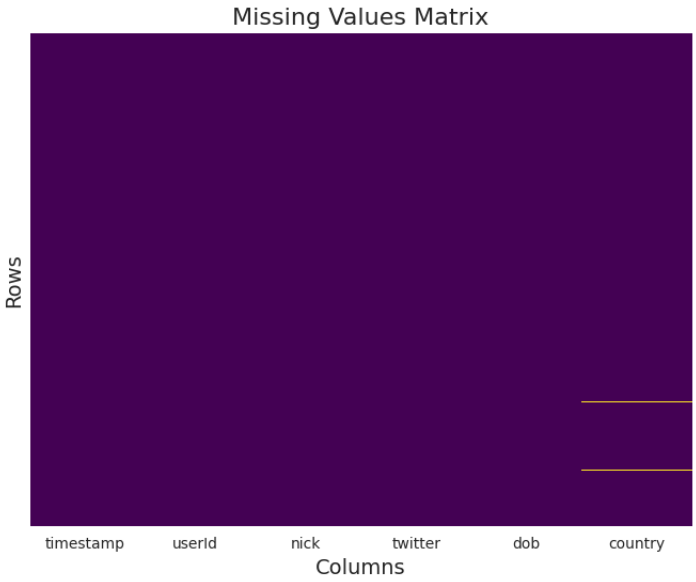

Figure 29: Missing Values in Users Dataset

6.8.2 Distribution of Player Birth Years

The histogram in Figure 26 illustrates the distribution of player ages in the game community. The birth year 1990 has a high frequency.

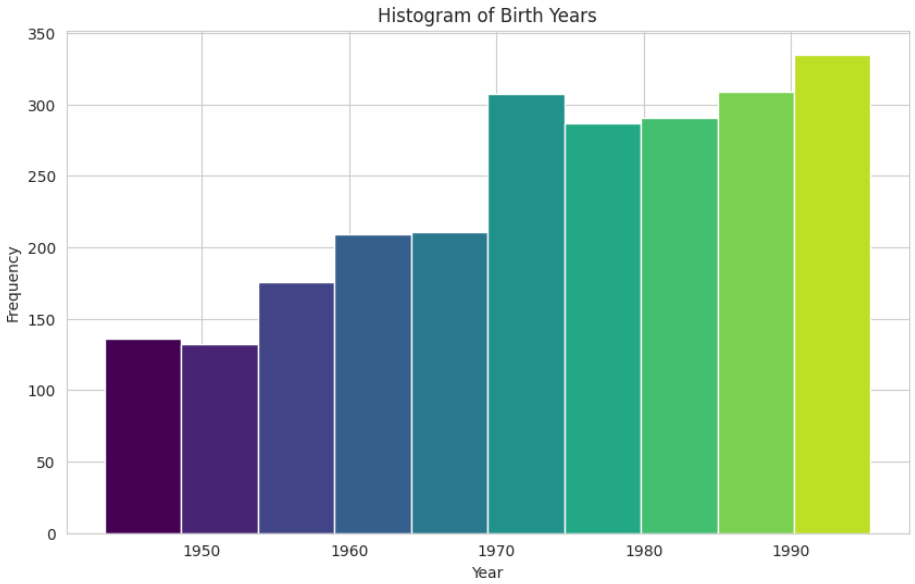

Figure 30: Distribution of Player Birth Years

6.9 Combined Data

6.9.1 Are there any missing values?

Figure 27 shows that many values from column count_buyid and avg_price are missing in the Combined dataset.

Figure 31: Missing Values in Combined Dataset

6.9.2 Players' Preferred Devices?

In Figure 28, the Pie chart illustrates the distribution of platform types. IPhone has 41.9% and Android has 35.4% showing that the iPhone is popular among players.

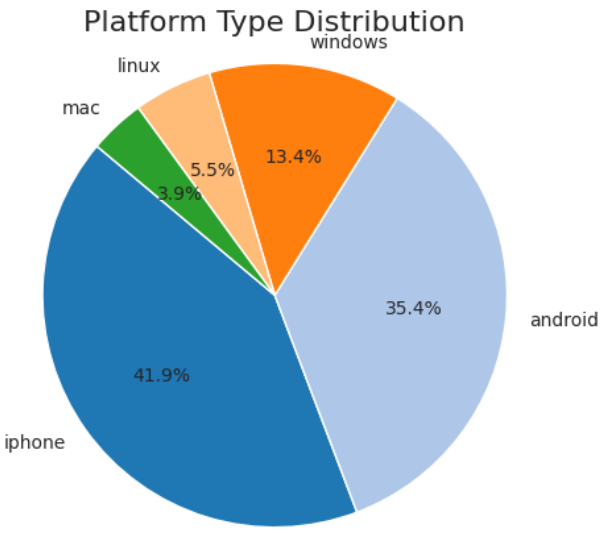

Figure 32: Platform Usage Distribution

6.9.3 Game Clicks vs. Hits on Different Platforms

Figure 29 shows that Linux users show maximum Game clicks and hit count in comparison to other platforms.

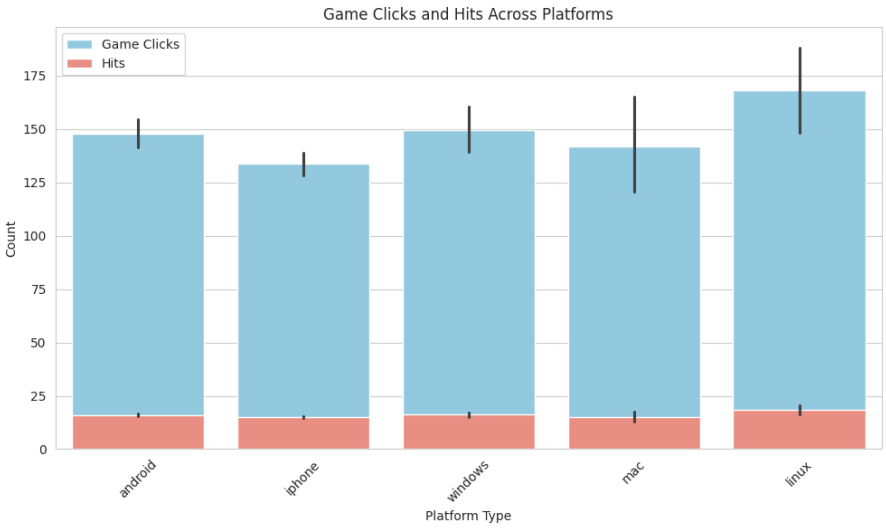

Figure 33: Comparative Analysis of Game Clicks and Hits Across Platforms

6.9.4 Analysing Average Price Distribution Across Platform Types

The histogram in Figure 30 shows the distribution of different platforms for varying average prices. It can help in analyzing pricing strategies and consumer preferences.

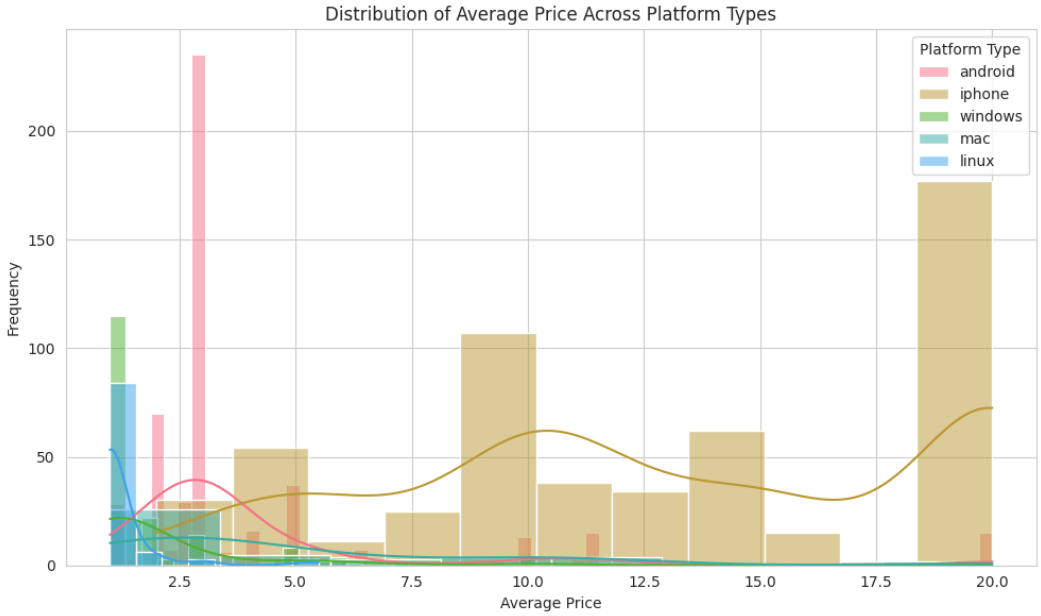

Figure 34: Distribution of Average Prices

6.9.5 Comparative Analysis of Team Levels Across Platform Types

Figure 31 depicts the distribution of team levels by platform types, The iPhone platform for team level 6 has having highest count. It shows the correlation between platform and player progression.

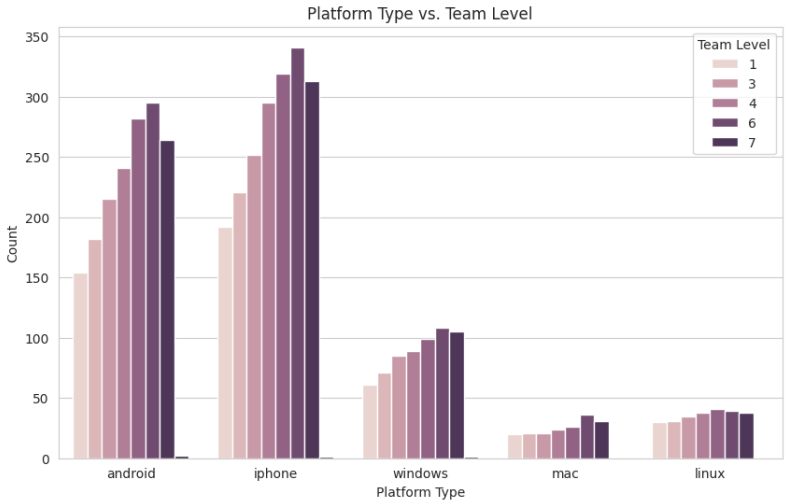

Figure 35: Distribution of Team Levels by Platform Type

6.9.6 Analysing Conversion Rates Across Platform Types

The Bar plot in Figure 32 shows that the Windows platform has the highest conversion rate indicating the possibility of in-game purchases.

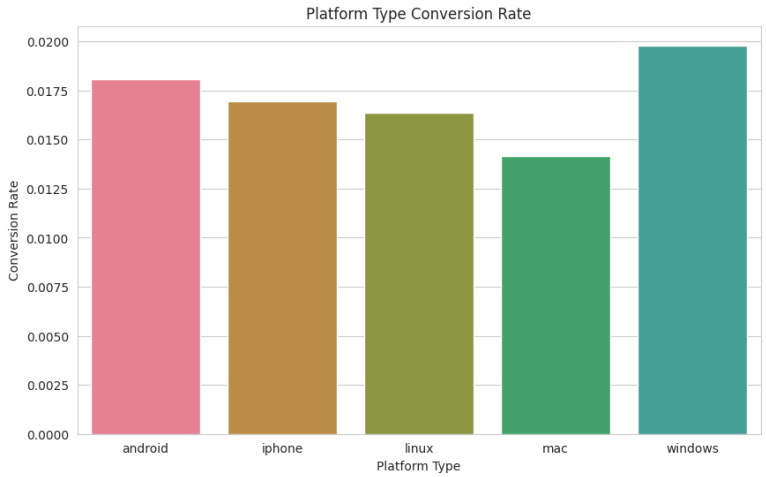

Figure 36: Comparison of Conversion Rates

7. Machine Learning Modelling

7.1 Classification

Classification in machine learning means to teach computers to understand and classify. We show an example to the model which is known as training data and then we test the model by showing a new example which is known as test data. When the model got trained successfully use this model for classification techniques (Kotsiantis, 2007).

7.1.1 Decision Tree

A decision tree is like a flow chart that helps in decision-making and gives choices and potential results based on factors like probabilities and cost. It uses the if-else rule to define the categories which looks like a tree having a root, branches, and leaves (Song, 2015).

We can see the confusion matrix of the decision tree in Figure 33 and the classification report in Table 3 which shows the model evaluation report having 1.0 for accuracy, precision, recall, and F1-Score. It indicates that the model has strong predictive capability supported by distributed instances across all test data.

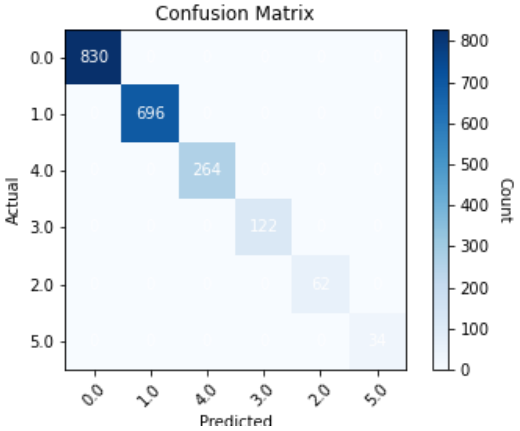

Figure 37: Decision Tree Confusion Matrix

	accuracy	precision	recall	f1-score	support
0.0	1.0	1.0	1.0	1.0	830
1.0	1.0	1.0	1.0	1.0	696
4.0	1.0	1.0	1.0	1.0	62
3.0	1.0	1.0	1.0	1.0	122
2.0	1.0	1.0	1.0	1.0	264
5.0	1.0	1.0	1.0	1.0	34

Table 4: Decision Tree Classification Report

7.1.2 Support Vector Machine

Support Vector Machine is like a smart tool that learns from the examples. We train this model with different example sets and it learns accordingly. Once the model learns effectively, we can use the model for classification (Cervantes, 2020).

The Confusion Matrix of the Support Vector Machine can be seen in Figure 34 and the classification report in Table 4 which shows moderate performance with accuracy, precision, recall, and an F1-Score of approximately 0.58. The support section talks about the effect of potential data imbalance.

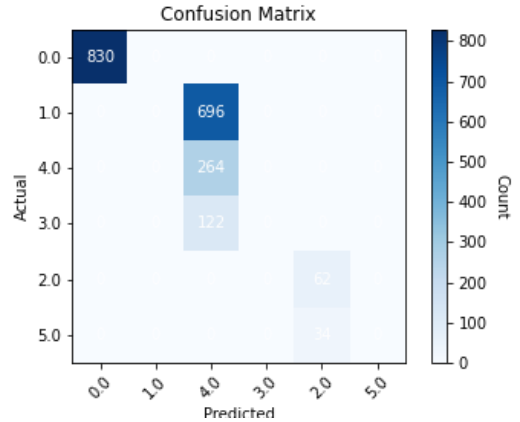

Figure 38: SVM Confusion Matrix

	accuracy	precision	recall	f1-score	support
0.0	0.575697	0.465366	0.575697	0.465366	830
1.0	0.575697	0.465366	0.575697	0.465366	696
4.0	0.575697	0.465366	0.575697	0.465366	62
3.0	0.575697	0.465366	0.575697	0.465366	122
2.0	0.575697	0.465366	0.575697	0.465366	264
5.0	0.575697	0.465366	0.575697	0.465366	34

Table 5: SVM Classification Report

7.2 Clustering

Clustering organizes data points based on similarity without predefined labels. It's unsupervised learning, focusing on finding patterns and grouping similar points together. Methods like distance measurement determine group membership, aiming to create clusters where points are more alike (Ezugwu, 2022).

7.2.1 K-Means

Clustering organizes the data points on the basis of similarity without any pre-defined labels. It is an unsupervised learning technique that groups similar points. Methods like distance measurement determine group membership which helps to create clusters where points are most similar (Sinaga, 2020).

Clustering performance looks better with a smaller number of clusters. As shown in Figure 35 Silhouette Score is increased up to 0.95 approximately when the cluster ranges from 10 to 20 but when the number of clusters increases Silhouette Score starts decreasing, leading to low-quality clustering. So, for this dataset maybe choosing a smaller number of clusters will be beneficial.

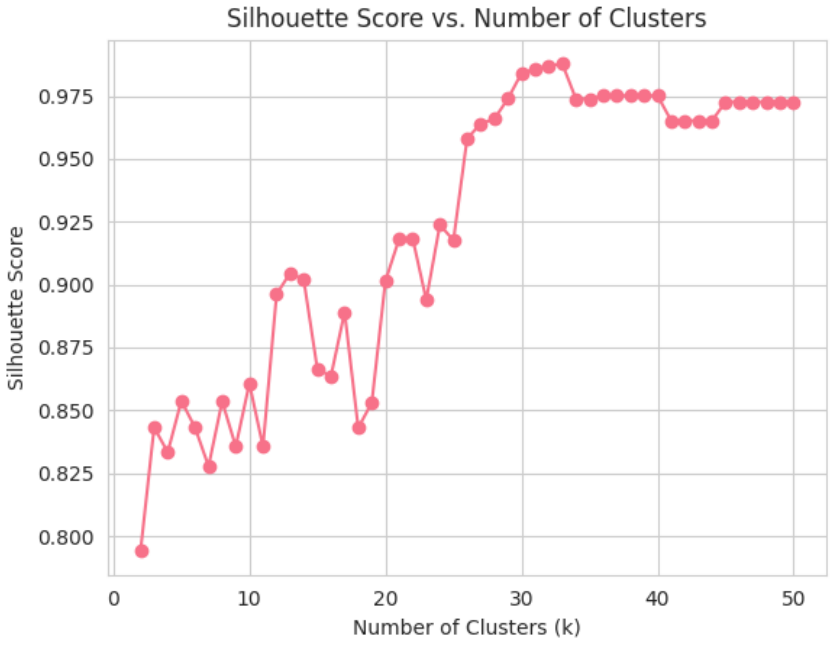

Figure 39: K-Means's Silhouette score vs number of clusters

7.2.2 Gaussian Mixture Models (GMMs)

A Gaussian Mixture Model (GMM) divides the data points into clusters or estimates data density assuming a mixed Gaussian distribution. It assigns each point to each cluster which understands the membership of multiple clusters. GMM has been used in machine learning for pattern recognition (Wang, 2019).

Figure 36 depicts the Silhouette Score vs Number of clusters which assesses the effectiveness of clustering showing higher silhouette scores give better matches within the clusters. Variability increases with a greater number of clusters which suggests potential over-lifting. Optimal cluster selection is used to balance performance and stability.

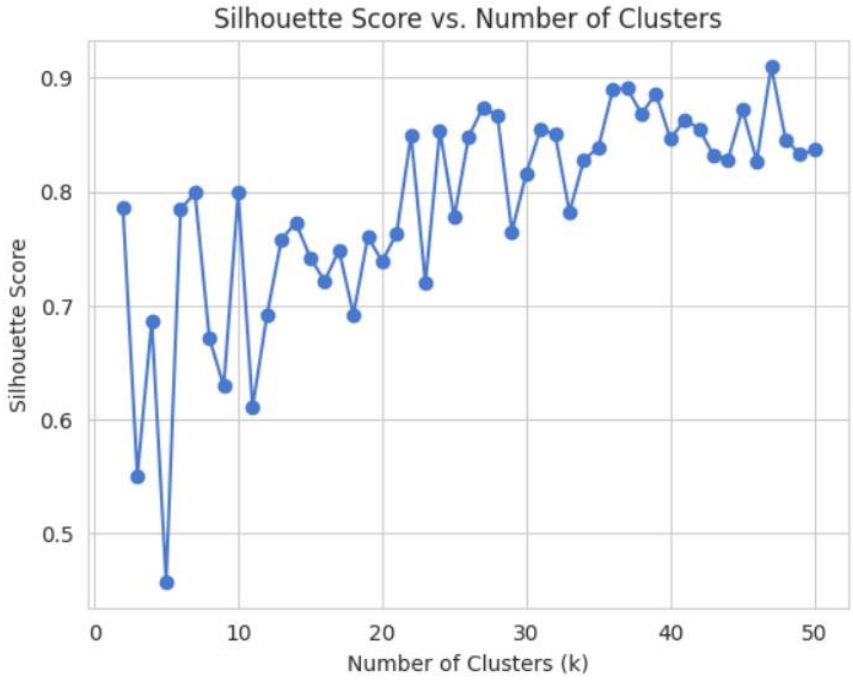

Figure 40: GMM's Silhouette score vs number of clusters

8. Graph Analysis

In multiplayer online games graph analysis is used to plot the actions of players a player joins which type of team, what is the behavior of the player in games, and the different gameplay styles of a player. With the help of graph analytics player's behavior evolution can be tracked as how socially connected a player is. Overall to understand the relationship between the game attributes and players (Lee, 2018).

8.1 Chat Items Created in Team Chat Sessions

Figure 37 Graph depicts interactions of team chat sessions where a chat item has been created. Nodes represent users, chat items, and team chat sessions on the other hand edges indicate which user creates a chat item and which chat item is associated with which team.

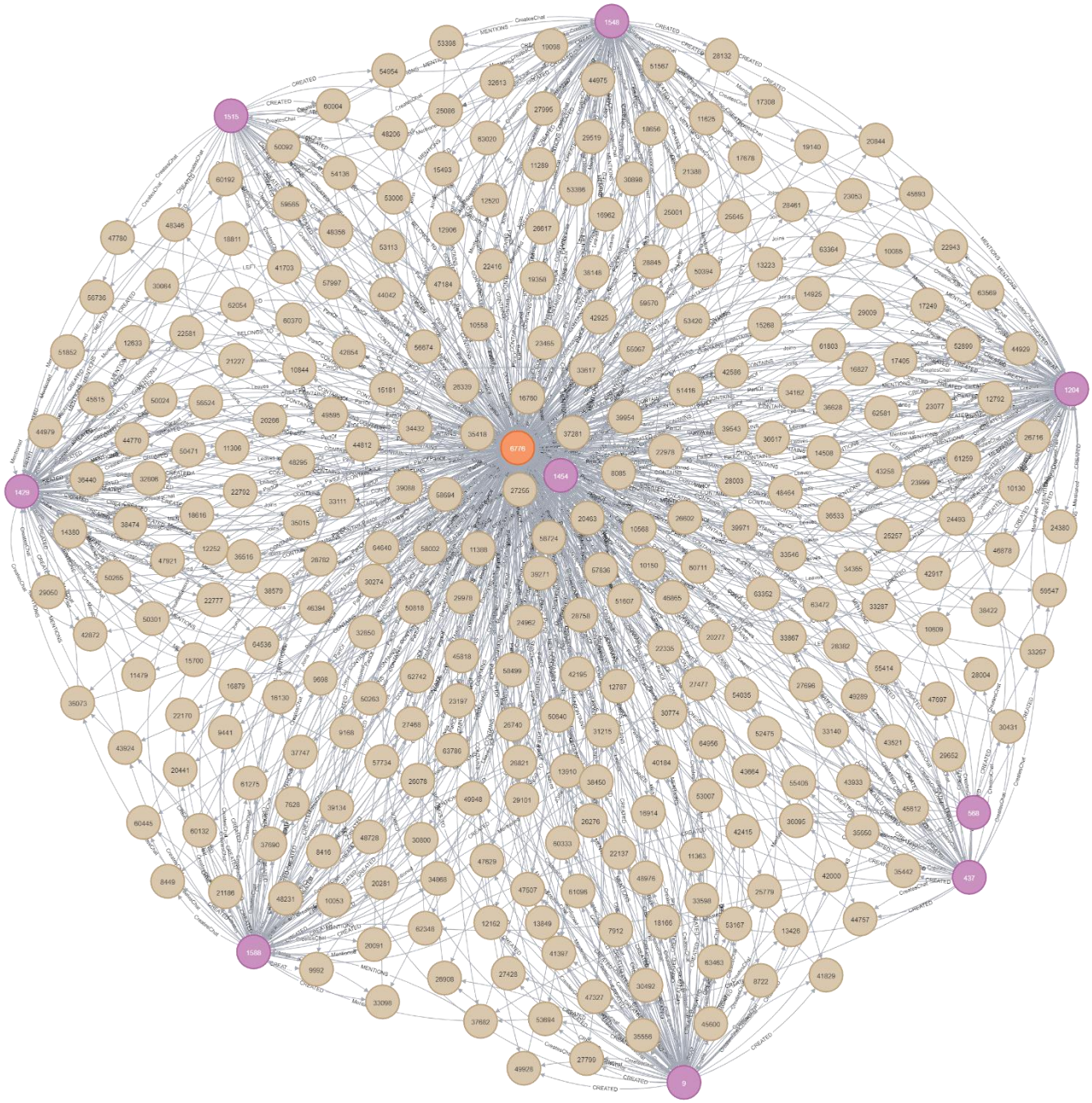

Figure 41: User-Team Chat Interactions

8.2 Team Chat Sessions Created by Users

Figure 38 illustrates the graph of users who created chat sessions within their respective teams. Nodes represent users, team chat sessions, and teams while edges represent which user created which session and which session belongs to which team.

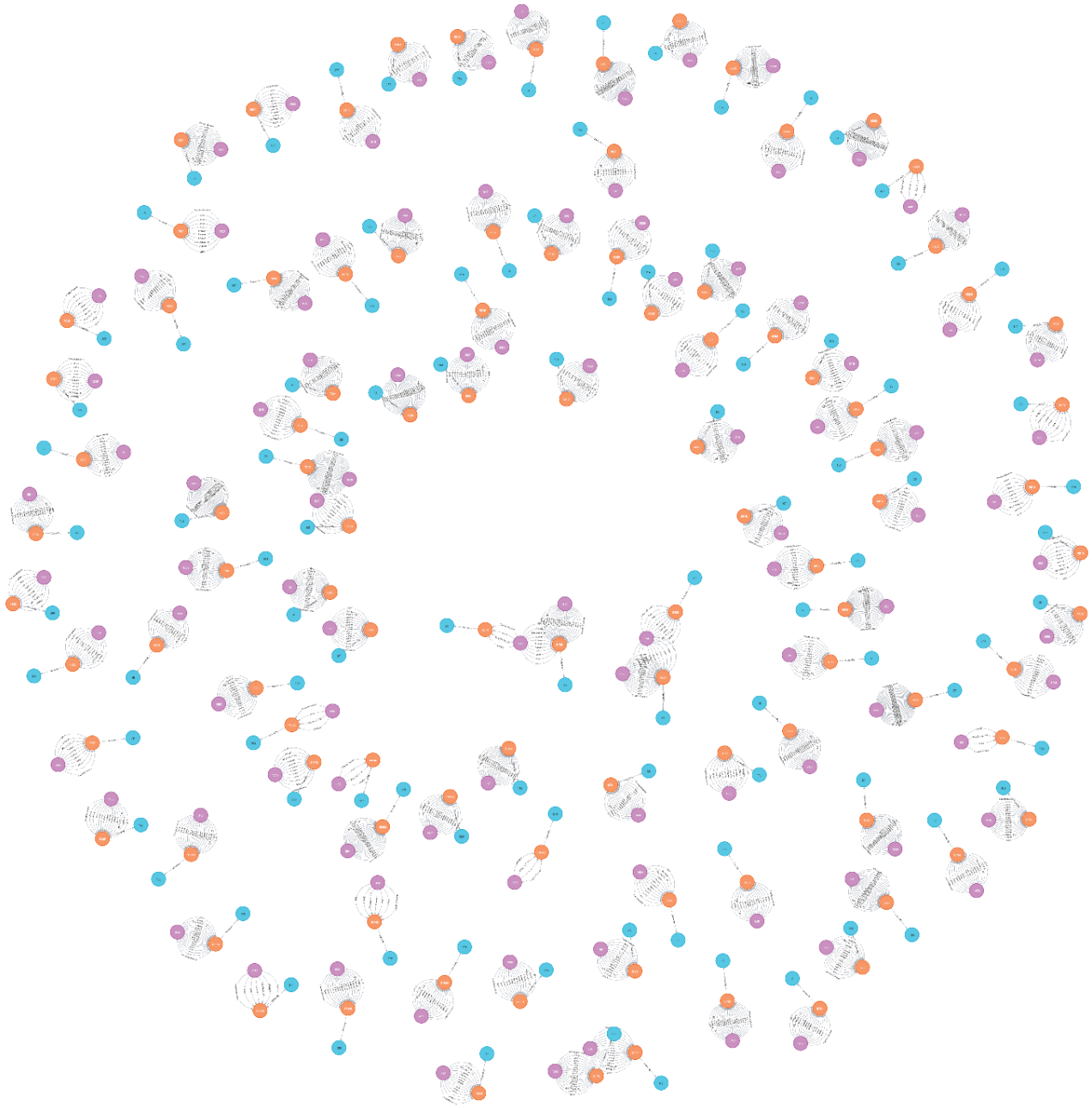

Figure 42: Creation of team chat sessions by users within their teams.

8.3 Users Joining Team Chat Sessions

Figure 39 shows a graph of users joining team chat sessions over time. Nodes represent team chat sessions and users while edges represent users joining specific sessions on specific timestamps.

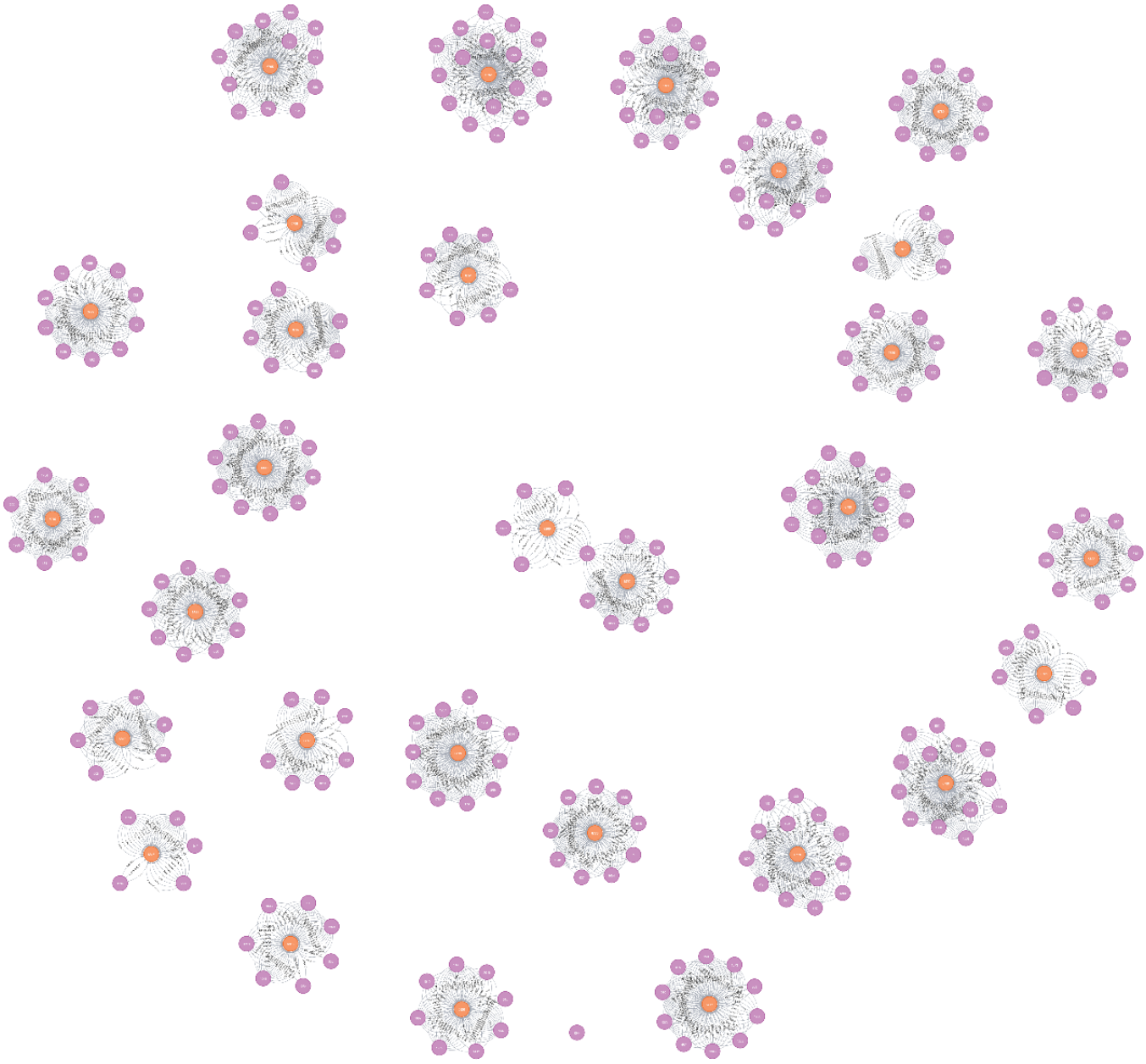

Figure 43: Users joining team chat sessions

8.4 Users Leaving Team Chat Sessions

Figure 40 shows the graph of users leaving team chat sessions on a specific timestamp. Nodes represent users and team chat sessions while edges represent users leaving specific sessions on specific timestamps.

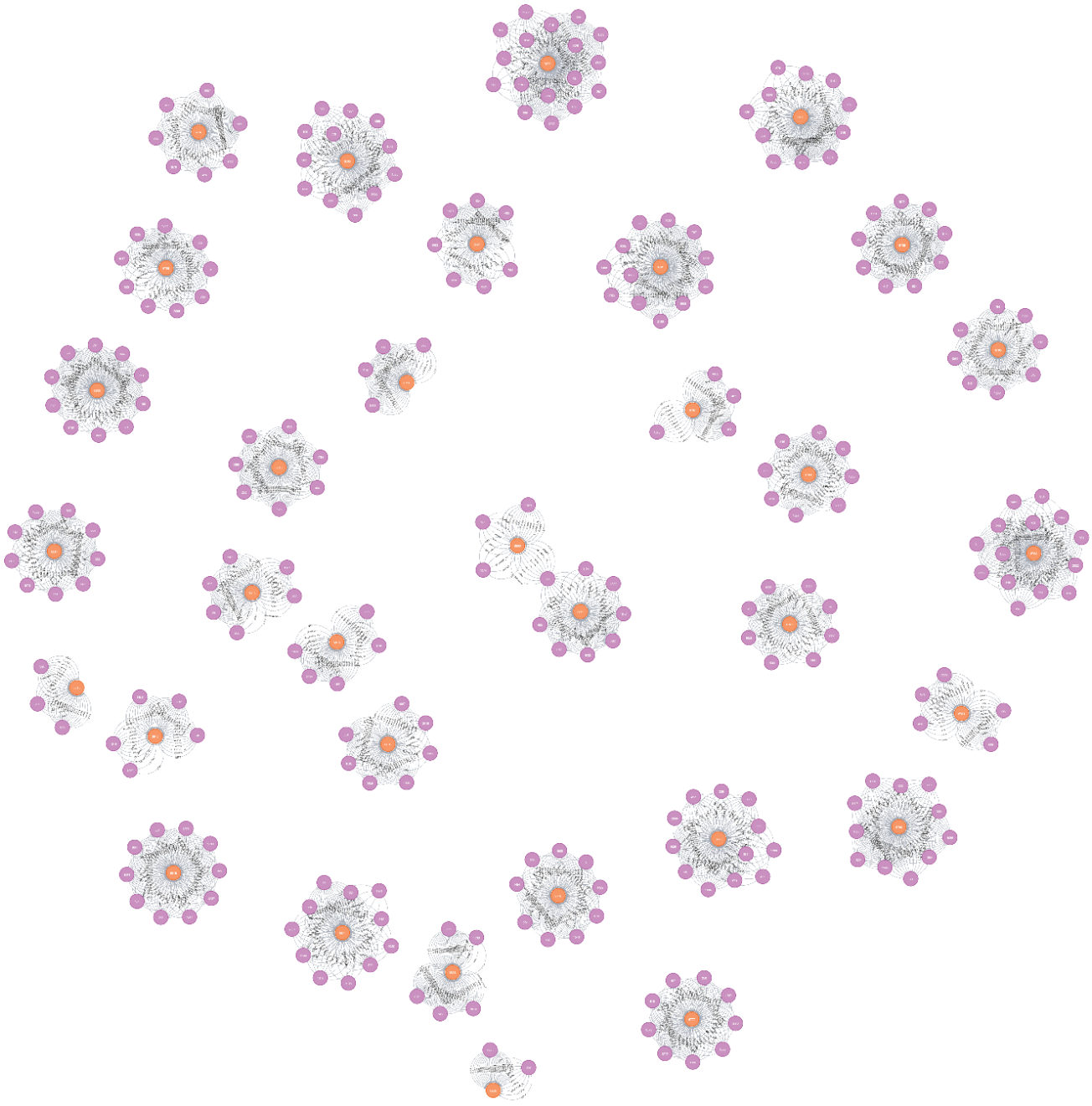

Figure 44: Users leaving team chat sessions

8.5 User Mentions in Team Chat

Figure 41 illustrates the graph having instances where users are mentioned in chat items with timestamps. Nodes represent chat items and users while edges represent mentions with timestamps.

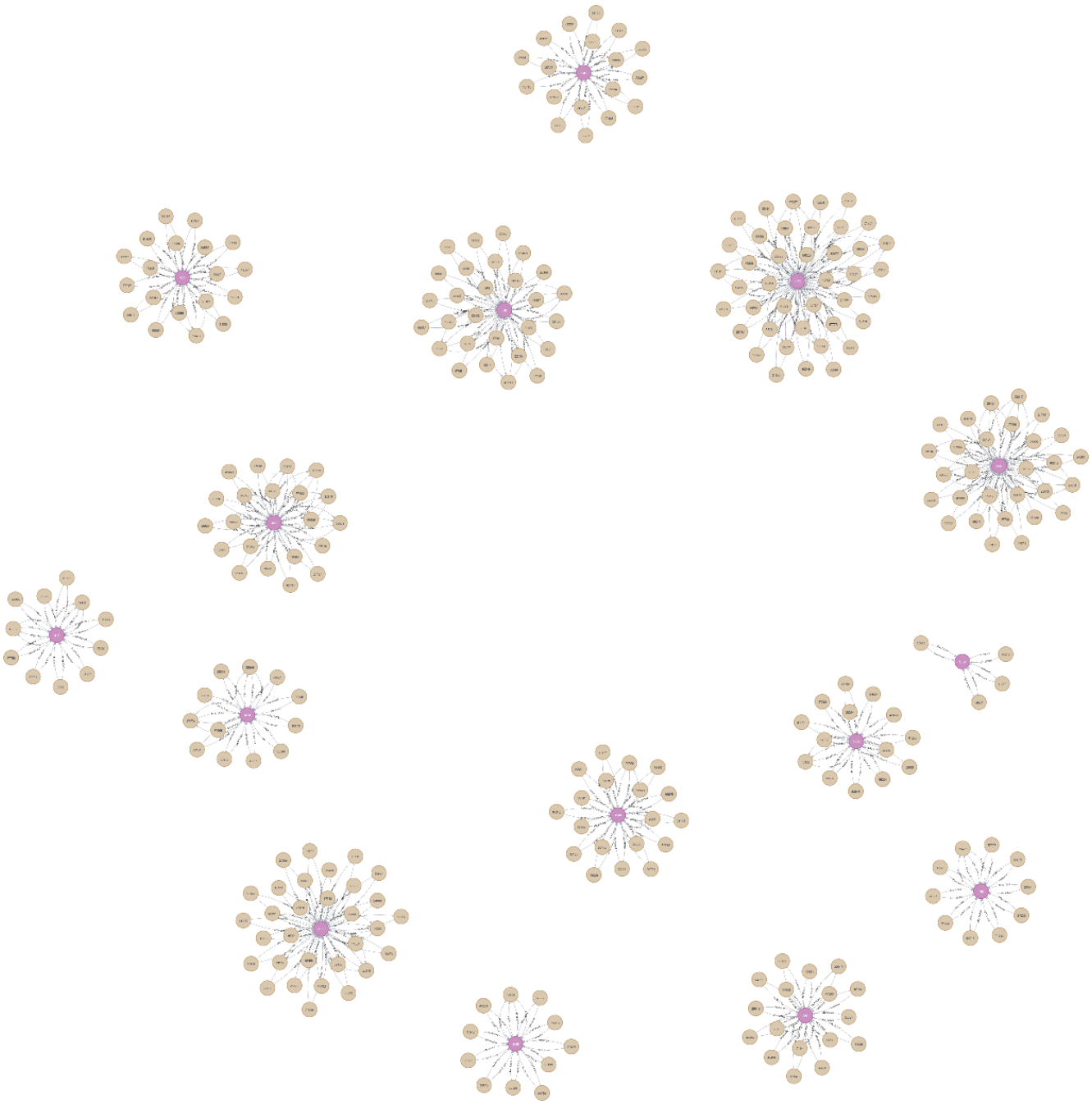

Figure 45: User mentions in team chat items

8.6 Responses in Team Chat

Figure 42 depicts the graph of users' responses of the chat item in a team chat with timestamps. Nodes represent chat items and users while edges represent responses with timestamps.

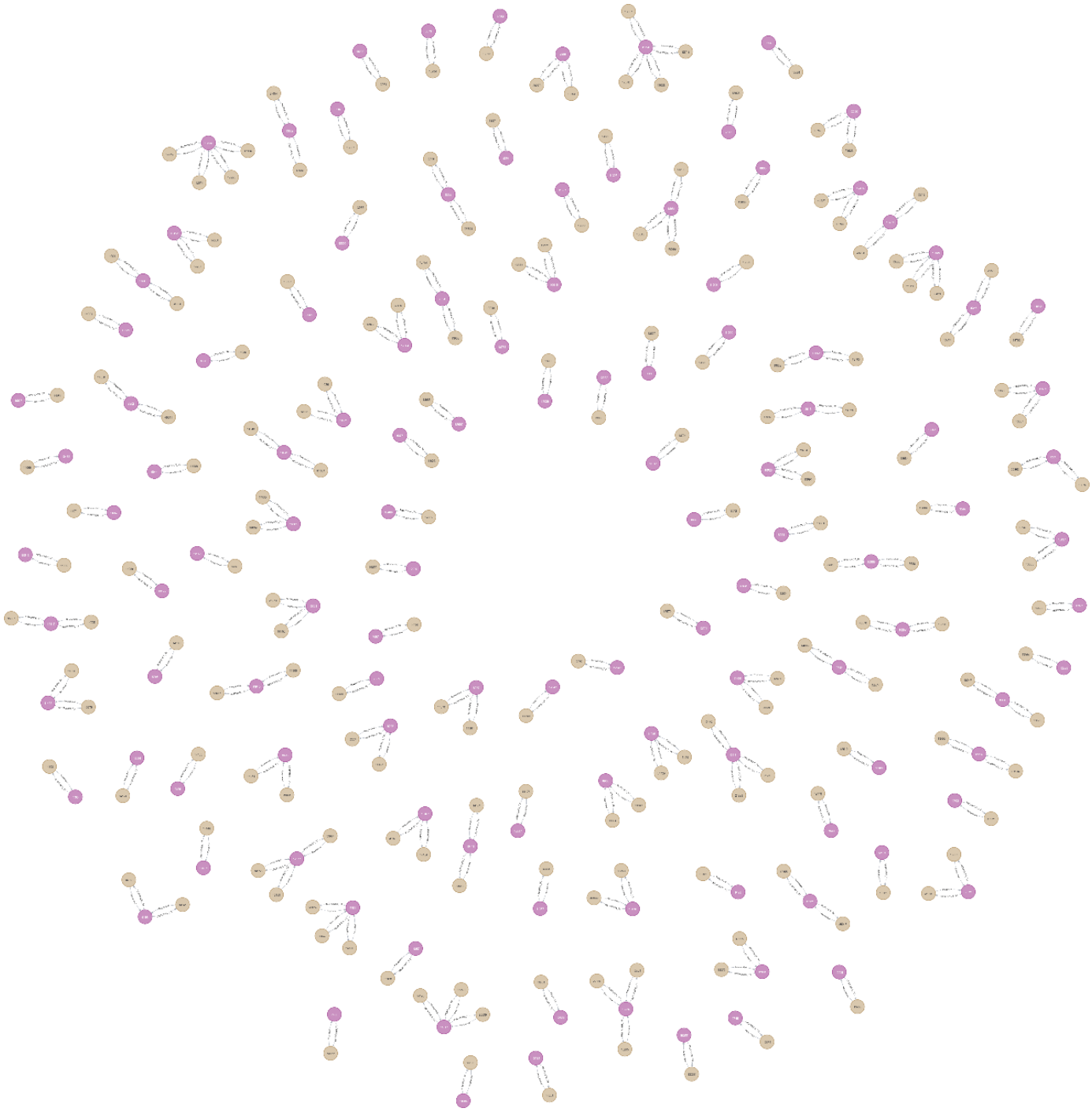

Figure 46: Responses made by users to chat items in team chat

9. Role of Ethics

Aspect	Description	Reference	Applicable?
Data Quality	Data Quality is very important for data analysis, if the quality is not good then insights can be irrelevant, decisions can be biased, and can lead to unethical consequences. While collecting data ethics should be on focus which includes privacy and consent. Transparent regarding data sources, purpose and it's usage. Protect the sensitive personal information.	(Ferretti, 2022) (Omer Osman Abdelrahman, 2022)	Yes
Machine Learning	Equity is important in machine learning selection, for example, the model can be trained only on an available dataset, if new data comes it won't recognize it. And if available data is biased, then it is unfair for a particular community. This is important to keep in mind that the model should handle data with fairness. Many machine learning models are not transparent, making it difficult to understand their decision-making process. Which leads to difficulty in tracking if the model makes a biased decision. That's why it is suggested to measure accountability.	(Piano, 2020) (Howe III, 2020) (Fletcher, 2022)	Yes

Table 6: Role of Ethics

10.Conclusion

"Catch The Pink Flamingo" showed the efficacy of Big Data analytics. Deep Exploratory Data Analysis (EDA), Using different machine learning models for classification and clustering, and Innovative knowledge graphs using Neo4j represented important and meaningful insights. EDA provided an important foundation while machine learning models presented a nuanced understanding of data and its structure. Neo4j explored hidden relationships and communities with graph analytics. Combinedly these approaches represent the extensive understanding of the dataset which helps to lead in the direction of actionable intelligence. This project is capable of exploring advanced analytics techniques in big data ecosystems.

11.Limitations and Recommendations

Despite the success of the project "Catch the Pink Flamingo" there are some possibilities to improve in the future.

Firstly, Data quality is very important for the analysis. In this project, there was an issue of data type conversion because of formatting issues, and missing values. For reliable and representative datasets robust quality and preprocessing is very important for accurate analysis and modelling.

Secondly, Algorithm selection affects the model performance. In the future, it is required to focus on an exhaustive model search or try new techniques.

Additionally, Scalability and computational requirements were also challenges for graph analytics, especially with big datasets like Catch the Pink Flamingo. Optimizing algorithms or distributed computing frameworks can resolve this issue and complex graphs can be analysed.

With some of these limitations, some recommendation for the future is data quality assurance and investing in preprocessing, experimenting with diverse machine learning algorithms, use scalable graph analytics methods.

12.Source Code

GitHub Link: https://shorturl.at/kwFH8

13.References

Casado, R. &. Y. M., 2015. Emerging trends and technologies in big data processing. Concurrency and Computation: Practice and Experience. Volume 27(8), pp. 2078-2091.

Cervantes, J. G.-L. F. R.-M. L. &. L. A., 2020. A comprehensive survey on support vector machine classification: Applications, challenges and trends.. Volume 408, pp. 189-215.

Ezugwu, A. E. I. A. M. O. O. O. A. L. A. J. O. E. C. I. &. A. A. A., 2022. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges,.

Ferretti, A. I. M. V. M. R. H. S. &. V. E., 2022. The challenges of big data for research ethics committees: A qualitative Swiss study.. *Journal of Empirical Research on Human Research Ethics.,* Volume 17(1-2), pp. 129-143.

Fletcher, J. &. K. A., 2022. Ethical Principles for Web Machine Learning..

Gogtay, N. J. &. T. U. M., 2017. Principles of correlation analysis.. *Journal of the Association of Physicians of India,* pp. 78-81.

Good, I. J., 1983. The philosophy of exploratory data analysis. pp. 283-295.

Howe III, E. G. &. E. F., 2020. Ethical challenges posed by big data. Innovations in clinical neuroscience. Volume 17, pp. 10-12.

Kotsiantis, S. B. Z. I. &. P. P., 2007. Supervised machine learning: A review of classification techniques.. Volume 160(1), pp. 3-24.

Lee, E. W. J. K. H. &. K. H. K., 2018. No silk road for online gamers! using social network analysis to unveil black markets in online games.. *In Proceedings of the 2018 World Wide Web Conference,* pp. 1825-1834.

Li, Z. L. Z. X. K. &. L. X., 2023. Evaluating LLM's Code Reading Abilities in Big Data Contexts using Metamorphic Testing.. *9th International Conference on Big Data and Information Analytics (BigDIA),* pp. 232-239.

Ma, S. W. H. M. L. W. L. W. W. H. S. .. &. W. F., 2024. The Era of 1-bit LLMs: All Large Language Models are in 1.58 Bits. pp. 2402-17764.

Omer Osman Abdelrahman, F., 2022. The Leading Role of University Libraries on the Ethics of Providing Big Data and the Analysis and the Extent of its Impact on Scientific Research: A survey of the libraries of the universities of the Kordofan sector.. *ARID International Journal of Informetrics..*

Piano, S. L., 2020. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward.. Volume Humanities and Social Sciences Communications volume 7(9), pp. 1-7.

Saeed, N. &. H. L., 2021. Big data characteristics (V's) in industry. *Iraqi Journal of Industrial Research,* Volume 8, pp. 1-9.

Sinaga, K. P. &. Y. M. S., 2020. Unsupervised K-means clustering algorithm.. *IEEE.*

Song, Y. Y. &. Y. L. U., 2015. Decision tree methods: applications for classification and prediction.. *Shanghai archives of psychiatry,* Volume 27(2), p. 130.

Wang, Z. D. C. C. R. M. &. F. B., 2019. Comparison of K-means and GMM methods for contextual clustering in HSM.. pp. 154-159.

Zhou, Y. G. C. W. X. C. Y. &. W. Y., 2024. A Survey on Data Augmentation in Large Model Era.

Zou, X., 2023 . New Opportunities for AI Innovation with Big Data: Indirect Docking between GLPS and LLM. *6th International Conference on Artificial Intelligence and Big Data (ICAIBD), IEEE,* pp. 444-450.