Assessment: Catch The Pink Flamingo Game

Oriname Agbi (22139986)

Module Lecturer: Muhammad Afzal

Date: 1st May 2023

Word count: 3680

Abstract

This research study delves into the vast potential of big data, specifically focusing on its application in the gaming industry. It explores the concept of big data, delineating its characteristics of Volume, Velocity, and Variety. This report analyses the game 'Catch The Pink Flamingo', developed by Eglence Inc., using big data tools and methodologies to extract valuable insights to enhance user experiences and game design. The study also considers the role of ethical practices in data storage and processing. In addition, it examines big data processing paradigms, notably batch processing, highlighting its functionality and application. The proposed solution uses Spark for data handling, employing various techniques such as Exploratory Data Analysis, Machine Learning, and Graph Analysis. The study yields practical recommendations to boost player engagement and revenue generation, based on demographic patterns, revenue sources, user behavior, and game dynamics. The report emphasizes the strategic advantages that innovative approaches to big data analysis can provide in a data-driven industry like gaming.

Contents

1 Introduction

In the rapidly evolving digital landscape, the concept of 'big data' has become a cornerstone, influencing every facet of our lives (Kitchin & McArdle 2016). The term 'big data' was first coined in the 1990s, marking the advent of an era where data sets became so large, diverse, and rapidly changing that they pushed the boundaries of conventional data management tools and techniques.

However, the vast potential of big data often lies concealed behind its inherent complexity and the necessity for innovative processing and analysis methodologies. Over the past few decades, big data has permeated a wide range of sectors. From health- care, where it aids in disease prediction and personalized medicine, to finance, where it informs investment strategies and risk management, big data has proven instrumental in driving insights and innovation. It has a particularly significant role in the technology and entertainment industries, where user-generated data forms a rich source of information.

The three Vs that encapsulate the concept of big data include Volume, Velocity, and Variety. 'Volume' refers to the enormous quantity of data that goes beyond the capacity of conventional storage and analytical tools. 'Velocity' denotes the ability to process and manage data in real time, a necessity in sectors such as banking, fraud detection, and healthcare. Lastly, 'Variety' alludes to the array of data types such as audio, video, structured, unstructured, and semi-structured data, necessitating adaptive storage and processing approaches.

This report focuses on the gaming industry, specifically the 'Catch The Pink Flamingo' game developed by Eglence Inc. With millions of players worldwide generating a plethora of data, the gaming industry presents a fertile ground for big data analysis. This analysis can enhance user experiences, optimize game design, and drive strategic decision-making.

The study aims to design and implement a comprehensive big data solution to harness the game's data effectively. This endeavor delves into the heart of big data processing paradigms, applying them to the rich dataset from 'Catch The Pink Flamingo'. We explore cutting-edge data exploration tools and visualization techniques, machine learning applications for classification and clustering, and graph analytics for in-depth insights.

Recognizing the critical role of ethics in this data-driven age, we also examine its implications on data storage and processing. The goal is to derive meaningful insights from the game's data, paving the way for improved gaming experiences for users and strategic advantages for the developers. This report strives to highlight how innovative approaches

to big data can uncover hidden patterns, reveal new knowledge, and open doors to a myriad of opportunities.

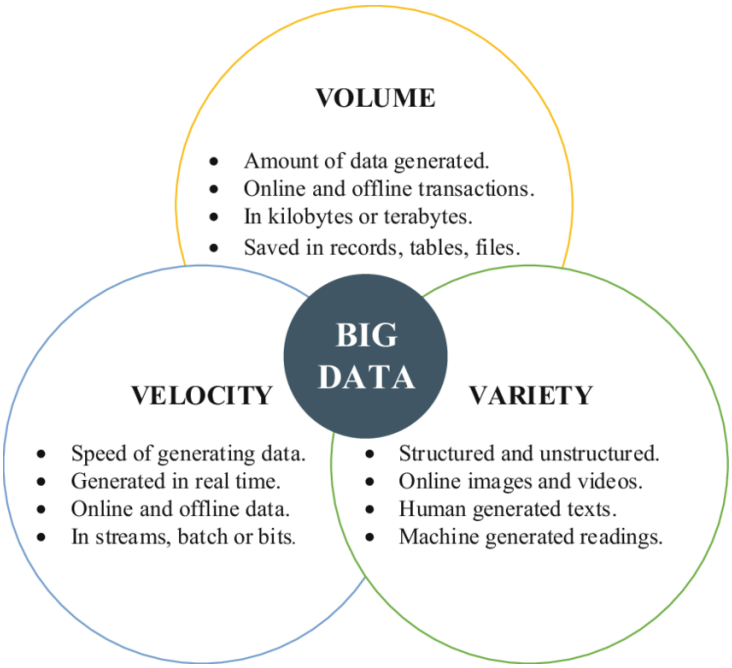

Figure 1: Three Vs of big data: volume, velocity, and variety

2 Big Data Processing Paradigm

2.1 Batch Processing Paradigm

Batch processing, a key technique in data processing, hinges on the Map-Reduce model, allowing concurrent computations across multiple data points—a crucial aspect in big data.

Hadoop, a prominent player, processes large data batches using two foundational elements: the Hadoop Distributed File System (HDFS) and the Map-Reduce model. HDFS ensures data resilience across nodes, while Map-Reduce facilitates processing capacity expansion—this horizontal scaling makes Hadoop a scalable solution for managing massive data sets without proportionate complexity increase (*Apache Hadoop* 2023).

Hadoop employs a two-step data processing approach: creating a static duplicate of the data set and running the desired operation on this duplicate, storing the outcome for future use.

Batch processing systems commonly utilize the Master-Slave model. Master nodes divide the larger task into manageable parts, delegate them to worker nodes, and integrate the results into the final solution (Dean & Ghemawat 2004).

However, batch processing systems, rigid and incapable of including newly arrived or

updated data after starting the process, can cause high latency. Nevertheless, they are effective where latency is less critical, like scientific calculations, graph analysis, and social network analytics. Now, let's consider some tools frequently used for batch processing:

Table 1: Batch Processing Technologies

Tools	Description
Batch Processing and Insights	
Apache Spark	A versatile tool known for its speed in big data processing, supports batch applications, and provides insights from large data sets.
Apache Flink	A combined stream and batch processing framework ideal for real-time data streams and big data processing.
Visualization	
Tableau	A data visualization tool that helps translate large datasets into insightful visuals.
Power BI	A business analytics tool by Microsoft offering interactive visualizations with self-service business intelligence capabilities.
Data Mining and ETL	
Pentaho	Offers a suite of open-source tools for data integration (ETL), OLAP services, reporting, data mining, and more.
Talend	A robust open-source data integration platform that offers data quality, ETL, and data management capabilities.

2.2 Real-time Processing Paradigm

Real-time processing is a cornerstone of big data management, addressing data velocity by offering immediate insights from incoming data streams, thereby significantly reducing latency compared to batch processing, which processes large data volumes at predetermined intervals(Casado & Younas 2014).

The real-time processing pipeline comprises three stages: ingestion, processing, and storage. Data, typically in semi-structured formats like JSON or XML, is ingested as a stream, processed in-memory for rapid insights, and subsequently stored for future use or immediate consumption.

Despite challenges, such as swiftly processing small data fragments without impeding the ingestion pipeline, real-time processing is crucial across various sectors. It enables high-frequency trading and instant credit card approvals in finance, personalized content delivery on social media platforms, immediate health alerts in healthcare, and real-time tracking and demand forecasting in logistics(Yaqoob et al. 2016).

Several tools support real-time processing in big data, including:

Table 2: Real-Time Processing Technologies

Tools	**Description**
Apache Kafka	A high-throughput distributed messaging system designed for real-time data processing. Handles trillions of events in a day.
Apache Storm	An open-source distributed real-time computation system. Used for real-time analytics, online machine learning, continuous computation, and more.
Apache Flink	An open-source stream processing framework providing high-throughput, low-latency data processing, and stateful computations over data streams.
Spark Streaming	An extension of the core Spark API enabling scalable, high-throughput, fault-tolerant stream processing of live data streams.
Google Cloud Dataflow	A fully-managed service for executing Apache Beam pipelines within the Google Cloud Platform ecosystem. Enables efficient, large-scale, distributed data processing tasks.
Amazon Kinesis	A platform to collect, process, and analyze real-time, streaming data to get timely insights and react quickly to new information.
Azure Stream Analytics	A real-time analytics and complex event-processing engine that analyzes and visualizes streaming data in real-time. Capable of processing millions of events per second.

2.3 Hybrid Processing Paradigm

Hybrid processing, which integrates various data processing methodologies, primarily real-time and batch processing, aims to enhance system performance, efficiency, and capability by leveraging the strengths of these methods. In the current data-centric era, hybrid processing systems proficiently manage both structured and unstructured data, catering to diverse data requirements.

A notable implementation of the hybrid processing model is the Lambda Architecture, introduced by Nathan Marz and James Warren (Marz & Warren 2015). This architecture comprises three critical layers: the batch layer, the speed layer (or real-time layer), and the serving layer.

The batch layer manages and processes historical data, performing comprehensive analysis on large data volumes, enabling businesses to extract valuable insights. This layer handles complex queries and computations, ensuring accurate results.

Contrastingly, the speed layer focuses on real-time data processing, aiming for immediate insights and responses. Optimized for low-latency operations, it provides real-time analytics and up-to-date information, facilitating prompt decision-making.

The serving layer, acting as a bridge between the batch and speed layers, combines their results to provide a comprehensive data view. It ensures the final output includes the latest insights from the speed layer and the thorough analysis from the batch layer, delivering a unified data representation.

The Lambda Architecture's layered structure allows organizations to handle both historical and real-time data effectively. By integrating the batch and speed layers through the serving layer, businesses gain a holistic approach to data processing, enhancing decision-making.

Hybrid processing systems' versatility, capable of handling immediate and batch data tasks, proves beneficial across various sectors. For instance, in finance, real-time processing handles instant transactions, while batch processing works through large historical data volumes for risk assessment and fraud detection. In healthcare, real-time data aids immediate patient monitoring, while batch processing identifies long-term health trends from historical data.

Technologies like Apache Kafka or Apache Flink manage real-time data streams, while Hadoop or Spark handle batch processing tasks. Frameworks like Apache Beam enable seamless integration of these processing models, allowing developers to construct data processing pipelines that effectively handle both bounded (batch) and unbounded (real-time) data..

Hybrid processing represents a strategic blend of real-time and batch processing. It combines the immediacy of real-time processing with the comprehensive analytical capabilities of batch processing, making it an increasingly relevant model in the face of growing data volumes and diverse processing needs.

2.4 Comparison of the three Paradigm

When comparing batch processing, real-time processing, and hybrid processing, we encounter three distinct paradigms in data processing.

Batch processing involves accumulating data over time and processing it in large chunks or "batches." It is ideal for handling substantial volumes of data and tasks that don't require immediate feedback. However, batch processing introduces latency since the results are not available until the entire batch is processed.

On the other hand, real-time processing operates on data immediately as it is produced, providing instantaneous or near-instantaneous insights. It is crucial for tasks that demand

immediate feedback or decision-making. However, real-time processing can be resource-intensive and challenging to manage, especially with high-velocity data streams.

Hybrid processing, a combination of batch and real-time processing, represents the best of both worlds. By utilizing a hybrid processing model, such as the Lambda Architecture, organizations can effectively manage and process massive amounts of data. The system employs batch processing for heavy computations on historical data while leveraging real-time processing for immediate insights on incoming data. This model offers a flexible, scalable, and efficient solution for complex big data scenarios, maximizing the strengths of both batch and real-time processing paradigms..

3 Exploratory Data Analysis

3.1 Flamingo Data Overview

For data acquisition, a total of seven datasets were gathered and examined to facilitate analysis. These datasets encompassed diverse information, such as ad clicks, buy clicks, game clicks, level events, team assignments, user sessions, and user details.

To prepare the data for analysis, several pre-processing and data transformation techniques were implemented. Initially, a new column called "age" was introduced by leveraging the "Date of birth" data from the user dataset. Subsequently, an "age-group" column was created to categorize users into eight distinct subgroups based on their ages. Additionally, the "country" column was transformed into a "continent" column to evaluate player demographics accurately. To consolidate the information from various sources, including team assignments and buy clicks, the datasets were merged into a unified dataframe..

3.2 Data Pre-Processing

Different pre-processing & data transformation techniques were employed to transform the data for the analysis. A new column named ”age” was added to ”users.csv” file. ”age” of the user was calculated using the ”Date of birth” column. Furthermore, ”age-group” column was added to segment the users into 8 subgroups. The ”country” column of ”user.csv” was converted into continent column. This step was used to check the demographics of the players. ”team.csv” dataset and ”buy-clicks.csv” files were joined together into one dataframe.

Table 3: Dataset Description

File Name	Description
`users.csv`	Holds information about the users playing the game
`ad-clicks.csv`	Contains information about the in-game ad clicks of the user
`buy-clicks.csv`	When a user makes an in-app purchase, a line is added in this file
`team.csv`	Holds information about all the teams in the game
`team-assignments.csv`	This file is updated each time a user starts or finishes a level
`user-session.csv`	Holds a record every time a user starts or finishes a game. Also starts a new session for each level cleared
`game-clicks.csv`	A record is added into this file when a user performs an in-game click
`combined-data.csv`	Holds the combined record of all user information, including team, team level, game clicks, buy ID, and average price

3.3 EDA Visualisation

3.3.1 Age Group of Players

Analyzing the visualization in Figure 3, we observe a stratification of users into eight distinct age brackets. A significant user concentration is evident within the 30-39 age demographic. This trend suggests that the game resonates strongly with individuals aged between 30 and 60, possibly due to its intuitive interface and user-friendly learning progression tailored for an older demographic. An intriguing outlier in the data is the complete absence of players younger than 20 years old. This insight is critical for understanding the game's user demographics and could inform future feature development and user engagement strategies..

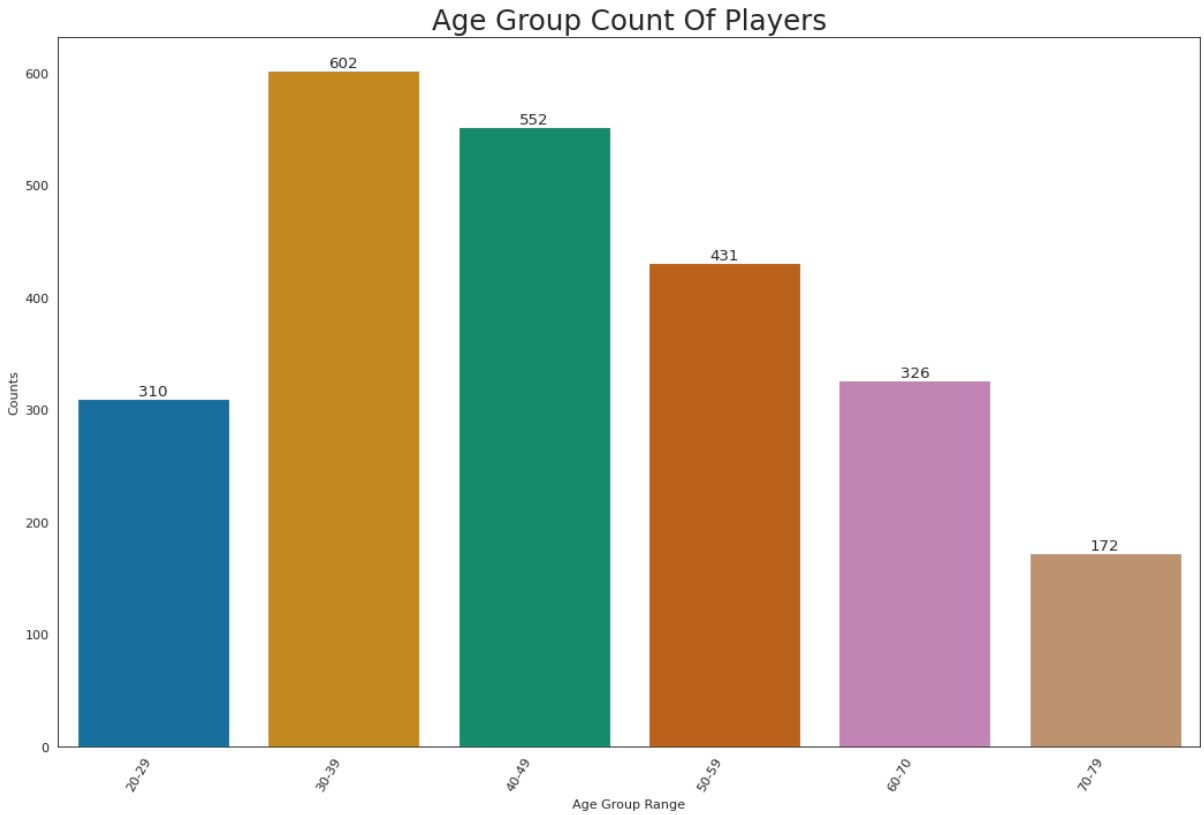

Figure 2: Age Group Range

3.3.2 Top Spending Teams Analysis

Figure 4, we can discern the spending patterns of various teams. Team 27 emerges as the highest spender, with a total transaction count exceeding 800. It's noteworthy that teams 35 and 64 exhibit similar spending behaviors, each with a transaction count surpassing 650, which is higher than that of team 53. To arrive at these insights, we merged the teams and buy-click datasets and aggregated the spending by teamID (Table 3). This analysis provides a clear picture of the spending habits of different teams, which could be instrumental in understanding their resource allocation strategies.

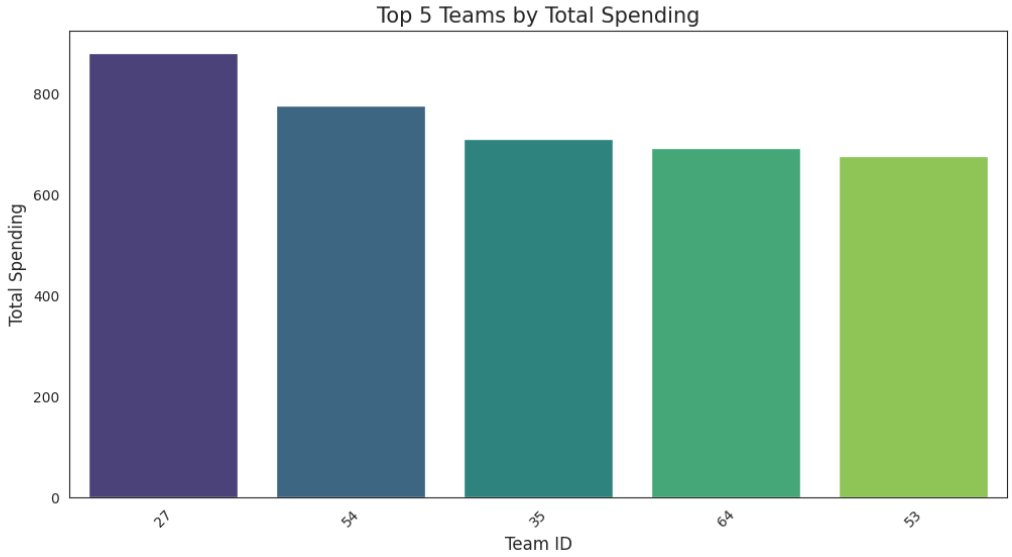

Figure 3: Top 5 Teams by Total Spending

3.3.3 Demographics With High Spending Analysis

Upon reviewing Figure 5., we observe that Africa and Asia are the regions with the highest in-game spending, accounting for 3889 and 3426 transactions respectively. As depicted in Figure 6, these transactions constitute over 40% of the total in-game expenditure. Interestingly, Antarctica registers the lowest in-game spending. This geographical analysis of in-game spending provides valuable insights into regional user behaviors and preferences, which could inform targeted marketing and development strategies.

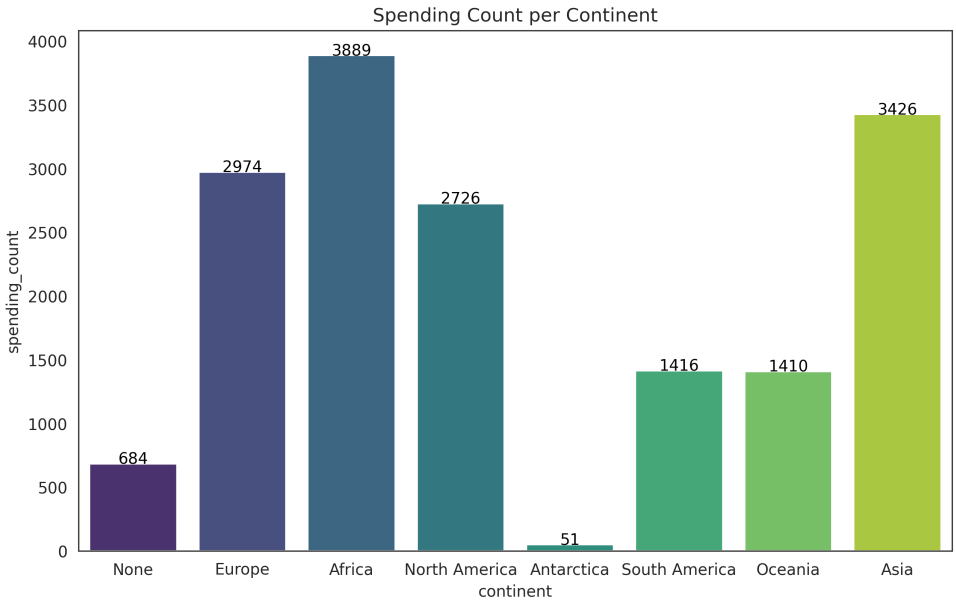

Figure 4: Spending count per Continent

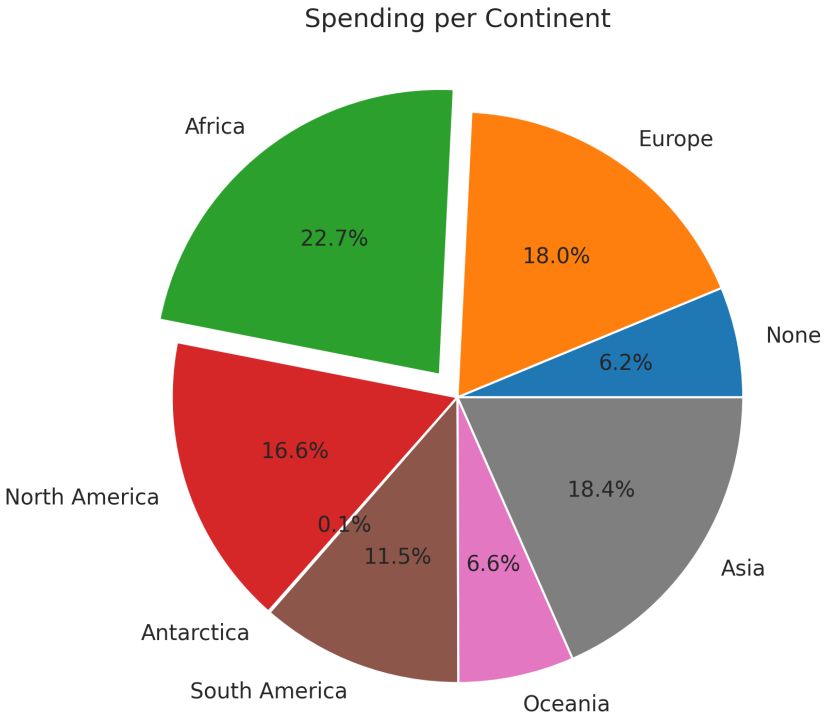

Figure 5: Spending count per continent

3.3.4 Highest Sold Product Analysis

Figure 7, we find that the most purchased item is item number 5, categorized under computers, with total sales exceeding 6 million. Conversely, the least purchased is item 0, falling under the sports category, with sales just above 300,000.

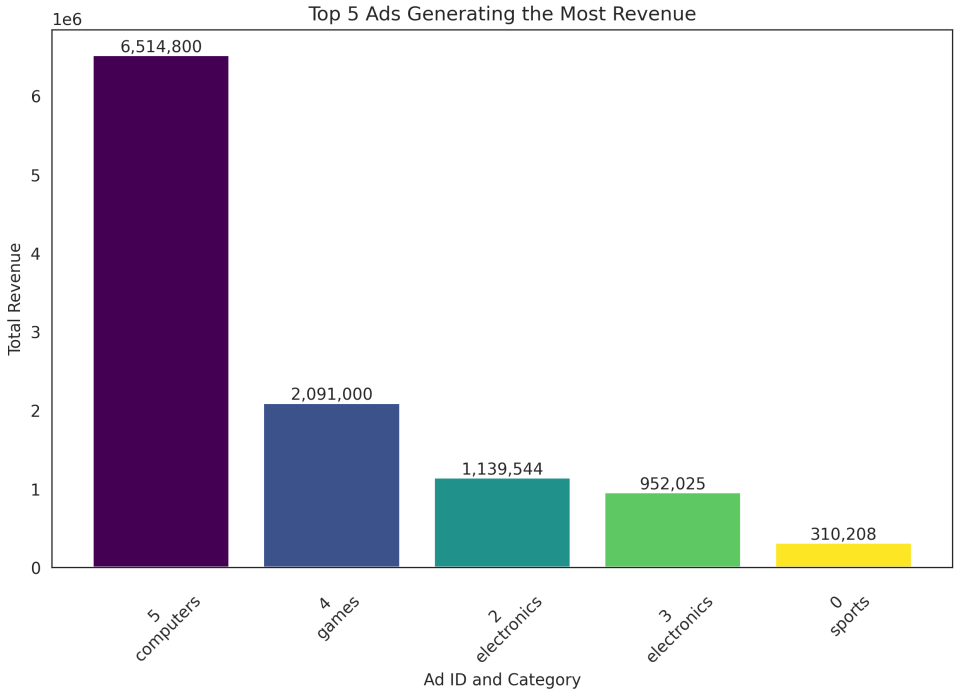

Figure 6: Top 5 Ads and its Category

3.3.5 Adclicks Analysis

Upon examining the treemap in Figure 8. of popular ad categories within the game, we observe that ads from the computers, games, and clothing categories garnered the most clicks from players, with counts of 2638, 2601, and 2340 respectively. In contrast, the automotive category registered the fewest clicks, with a count of 566. This analysis of ad click patterns provides valuable insights into player preferences, which could inform future ad placement and monetization strategies.

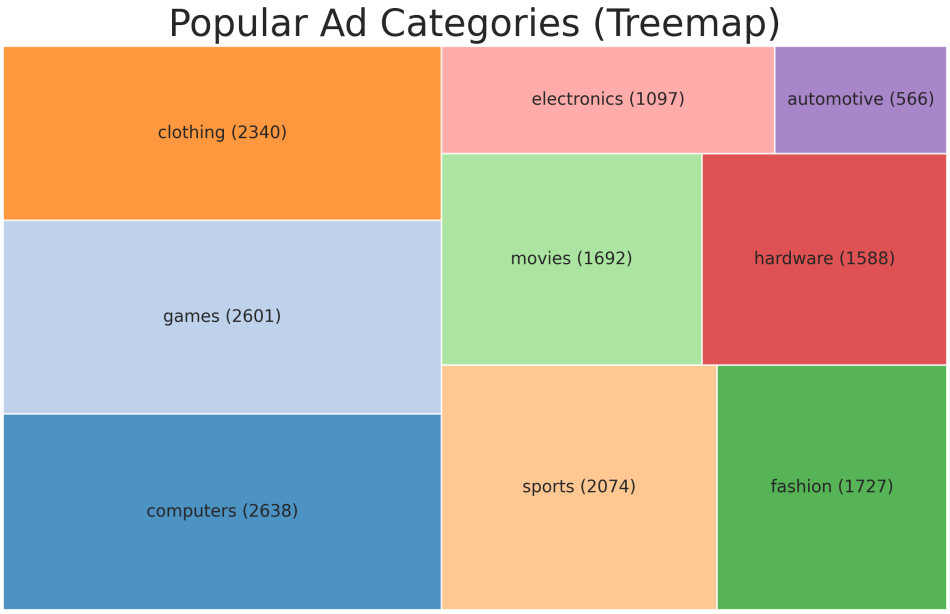

Figure 7: Ads Treemap

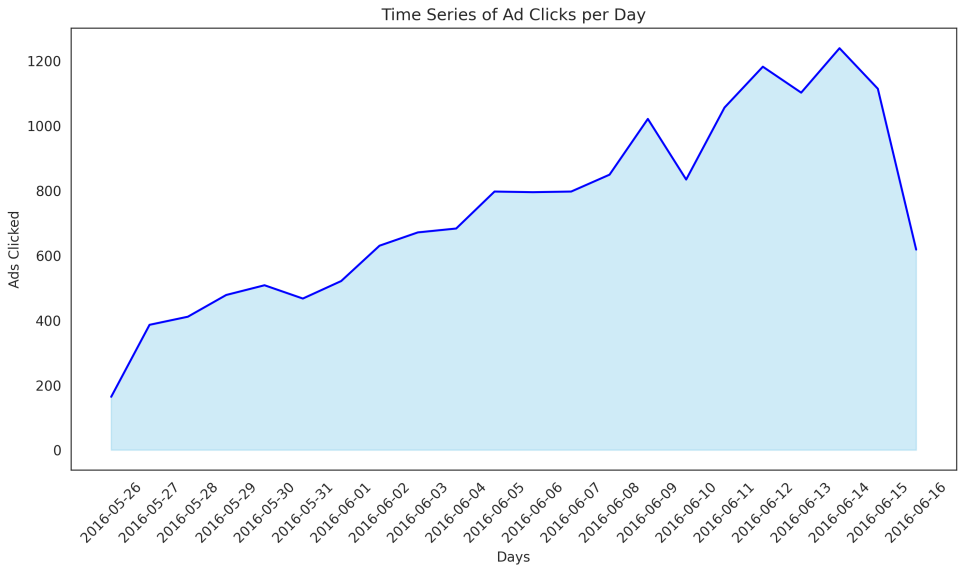

Figure 8: Timeseries of Ad clicks

3.3.6 Popular Device Platform

Interpreting Figure 9, we find that the iPhone emerges as the most favored platform among users, accounting for 41.9% of the game's total player base. Conversely, Mac registers the lowest usage at 3.9%. Interestingly, Windows holds a significant share of 13.4%, indicating its popularity among PC gamers. This platform usage analysis offers critical insights into player device preferences, which could guide platform-specific development and optimization strategies.

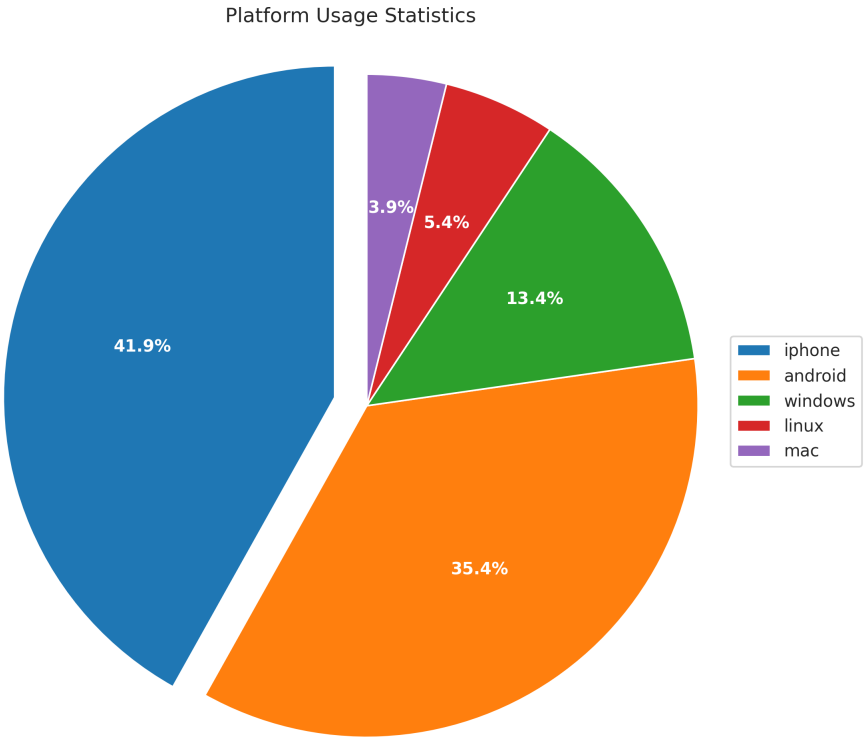

Figure 9: Ads clics per day

3.3.7 Hit Ratio Analysis

Figure 10 provides a comparative analysis of the total number of hits versus hit counts for each platform, along with their respective percentages. According to the figure, Mac registers the lowest hit count, while iPhone and Android platforms record the highest, both exceeding 11%. This hit count analysis provides valuable insights into platform engagement levels, which can inform platform-specific user engagement and retention strategies..

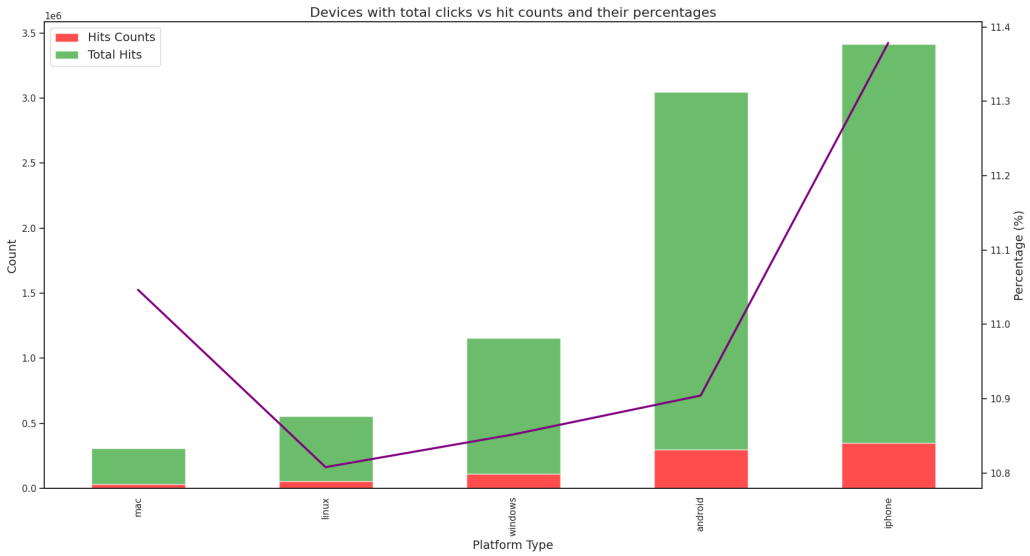

Figure 10: Hit Ratio

4 Machine Learning

Machine Learning is the field that enables machines to learn and make intelligent decisions by continuously refining themselves using complex mathematical models (Oussous et al. 2018). This realm has influenced society in monumental ways, impacting areas like computer vision, natural language processing (NLP), and the Internet of Things (IoT).

The importance of machine learning methodologies has been immensely felt in data-rich domains such as astronomy, music creation, and biology. Machine Learning has offered the tools to uncover knowledge and decode concealed patterns within these fields Machine Learning techniques were applied to the combined-data dataset to gain insights and segment data. The aim was to inform future business decisions and target users based on their gameplay abilities. "NULL" values in the "average price" column, representing users' average in-game spending, were replaced with "0". Two new categorical columns, "Spender Non Spender" and "Hitter and Non Hitter", were created for analysis. Players' total spending and hit count percentages were calculated and used to classify them into respective categories.

4.1 Classifcation Analysis

we also performed classification analysis on the combined-data dataset. We employed binary classification models, specifically Decision Tree and Logistic Regression, to predict the labels "Spender" and "Non Spender".

4.1.1 Decision Tree

We used a decision tree for binary classification into "Spender" and "Non Spender" categories based on attributes like "Total Hit Count", "Team Level", and "PlatformType". This prediction aids in designing game mechanics, in-game currencies, and marketing strategies. The model, fitted in PySpark, was trained on 80% of the data and tested on the remaining 20%. The model's accuracy, gauged using a confusion matrix, was 87.25%, correctly predicting 810 Non Spenders and 5 Spenders, with some misclassifications. The model is visualized in Figure 11.The model obtained from the training set was used on test data, and the accuracy of the model was gauged using the confusion matrix given in Table 4 below:

Table 4: Decision Tree Confusion Matrix of Spender

Spender	**Prediction**	**Count**
0 (True Negative)	0	810
1 (False Negative)	0	108
0 (False Positive)	1	11
1 (True Positive)	1	5

The accuracy of the model was calculated using

$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FN + FP}$$

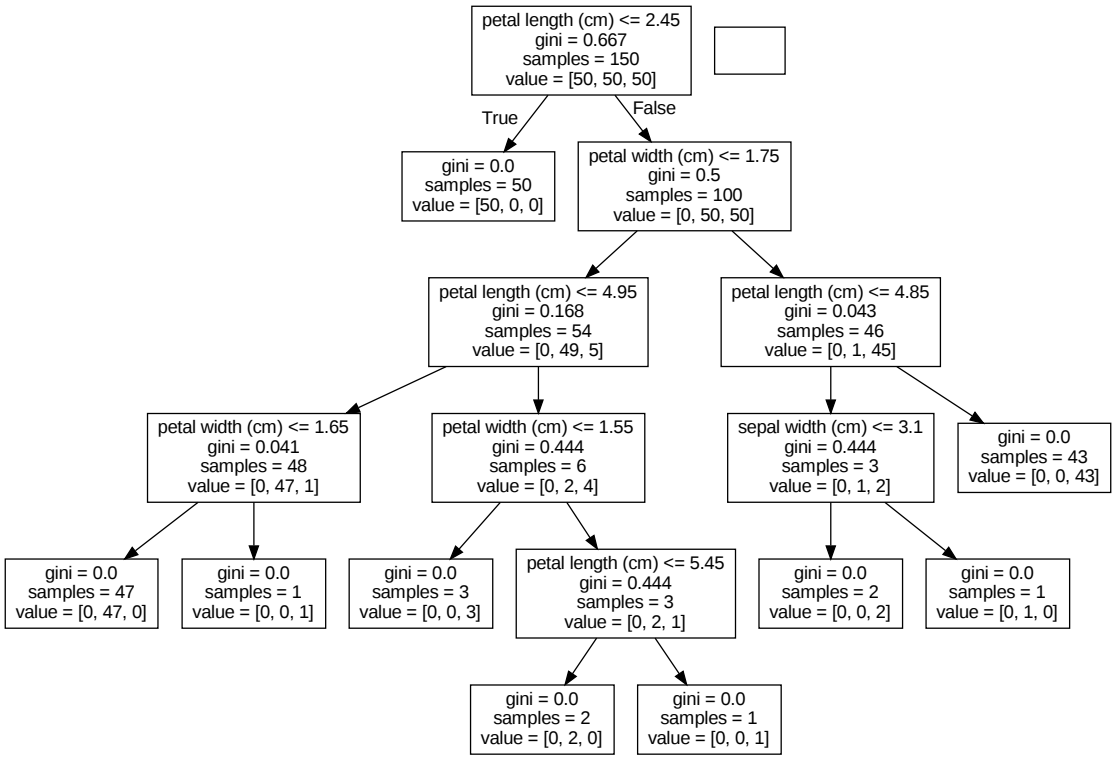

Figure 11: Decision tree

4.1.2 Logistic Regression

Logistic Regression is a common technique used for binary classification, distinguishing itself from decision trees. Rather than applying GINI impurity to ascertain feature significance for the root node, logistic regression leverages probabilities to classify samples based on the existing feature space.

The logistic regression approach uses probability to categorize samples. When a sample's probability exceeds 0.5

In this particular scenario, a logistic regression model was used to segregate players into two groups: "Spender" and "Non-Spender". The same attributes used in the decision tree model, specifically "Total Hit Count", "Team Level", and "Platform Type", were used.

The model was trained with a data division of 80% for training and 20% for testing. The model that resulted from the training data was subsequently tested on the testing data, and its accuracy was assessed through the provided confusion matrix.

Spender Prediction Count	**Predicted False (0)**	**Predicted True (1)**
0 (True Negative)	821	0
1 (False Negative)	112	0
1 (True Positive)	0	1

Logistic regression excels in identifying Non-Spenders, while the decision tree method is more effective at correctly classifying Spenders.

4.2 Clustering Analysis

4.2.1 K-Means Clustering

K-Means clustering, an unsupervised machine learning approach, is employed for structuring unlabelled data into homogenous groups by measuring the Euclidean distance from the centroid within the feature-space. The fundamental aim of K-means is to accentuate the similarity within each cluster, rendering it highly effective for large datasets due to its assurance of convergence within the feature space (Trevino 2019).

This algorithm is incredibly adaptable, suitable for a variety of grouping applications. However, a critical consideration in implementing K-means clustering is the selection of the optimal 'k' value, representative of the number of clusters. This value was discerned for our dataset by utilizing the Within Cluster Sum of Square Error (WSSE) method, often known as the elbow method.

The aptly named 'elbow method' visually interprets the point of inflection on the graph, akin to a human elbow, at a cluster size of five (refer to Figure 12). Thus, for this specific dataset, we selected k=5, resulting in five distinct clusters.

We focused our cluster analysis on the features of "platformType", "teamLevel", "hitter", "spender", "Average Price total", and "Average Hit Count", aiming to uncover trends among groups possessing different attributes, such as spender vs non-spender and hitter vs non-hitter. Interestingly, users on "iPhone" platforms seemed to spend more on in-game items.

These features yielded the cluster centers as shown in Table 6 (see Table 5 for features), with the order: "team Level", "platform Type indexed", "avg price total", "count hits total", "spender" and "hitter".

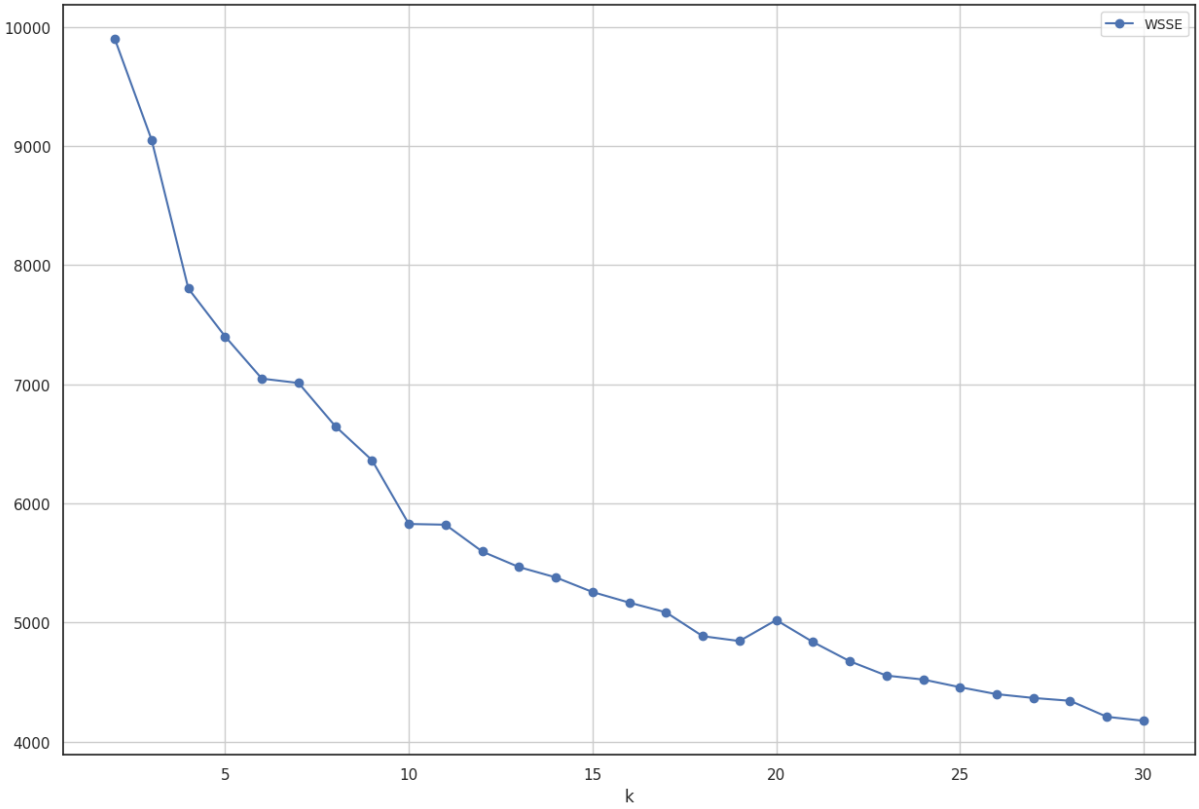

Figure 12: Elbow

Features	**Rationale**
platformType	PlatformType can be an important differentiator when it comes to game spending. For example, "iPhone" & "Mac" are expensive. So, the players using these platforms may not be averse to spending big on in-game items.
teamLevel	TeamLevel is also a discriminator. As, teams with high ranks may not require that much spending on items.
hitter	A player who is a strong hitter might not be that much interested in item purchase.
spender	Spender property is used to investigate how spending fares with "hitter" & "Team Level".
AveragetotalSpending	Average total spending per user is across all products is used.
Average hit Count	Average hit count per user is used to check individual hit count against the other properties.

Table 5: Features Used For K-Means Clustering

Cluster	Properties Cluster Center
Cluster 1	-0.15907806, -0.03636503, 0.56003647, -0.37215053, -0.21732124, -0.1600847, -0.17990985
Cluster 2	0.14750323, -0.45360234, 0.28719822, -0.13745186, 2.53035053, -0.0548085, 2.51408176
Cluster 3	2.3602386, 0.00424472, -0.20124819, -0.29208106, -0.24600105, -0.39919541, -0.26102093
Cluster 4	-0.75155886, 0.30198308, 0.3903432, 2.56166, -0.20978065, 1.79067607, -0.10433334
Cluster 5	-0.1950345, 0.11889969, -1.78521144, -0.35844995, -0.32858808, -0.45484407, -0.47903675

Table 6: Clusters

Cluster 1 includes players with a team level of -0.159 standard deviations below the mean, the highest spending (0.56), and low hit ranks (0.217). Cluster 2, with the lowest team level (0.147 below the mean), ranks fourth in spending (0.287) and has the highest hit count (2.53). Cluster 3, with the highest team level (2.36 above the mean), ranks third in spending (-0.201) and has a mixed player base with an average hit count (-0.246). Cluster 4, with the third-best team level (-0.75 below the mean), ranks lowest in spending (0.39) and has the third-lowest hit count (-0.209). Cluster 5, with the second-best team level (-0.195 above the mean), ranks second in spending (-1.785) and includes the best hitters (-0.479 above the mean). These clusters provide insights into player behaviors and preferences.

From our visual cluster comparison (Figure 13), intriguing patterns emerged. Lower-ranking teams, represented by the light and dark blue clusters, are major spenders, while high-ranking teams (green and red clusters) exhibit moderate spending and hit counts. This suggests a marketing focus towards beginner level players.

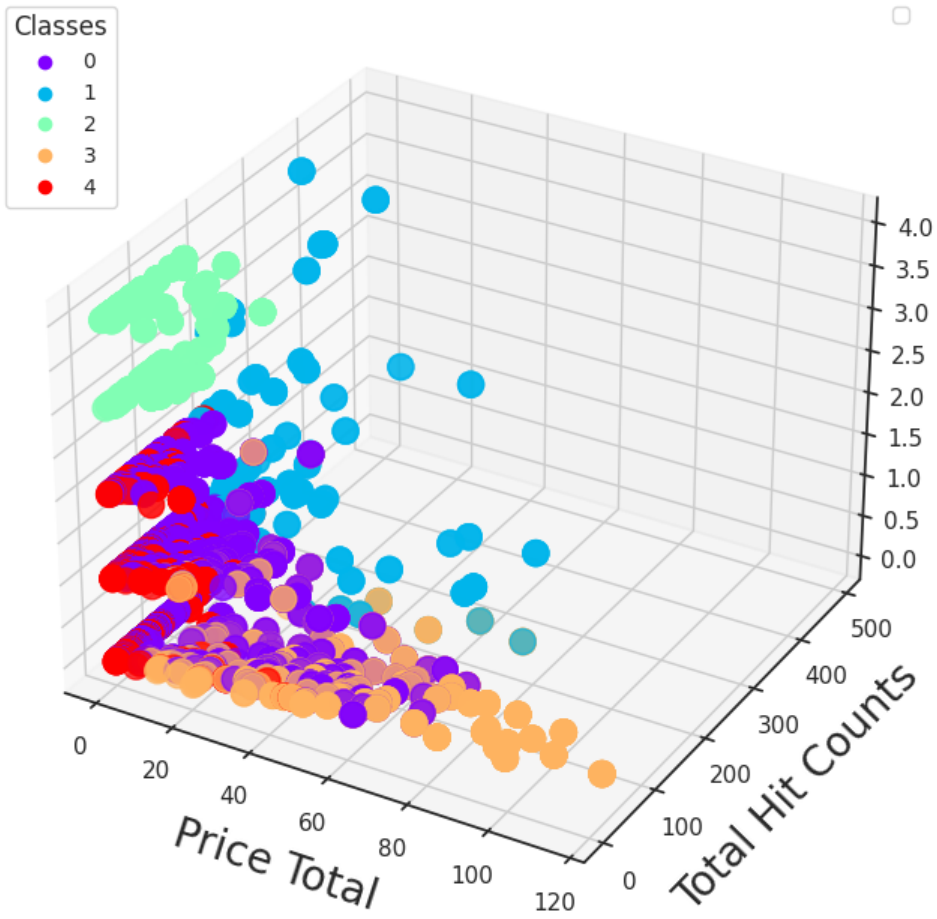

Figure 13: 3D Scatter Plot

5 Graph Analysis

5.1 Relationships

In this section, we delve into a comprehensive analysis of player engagement and interactions.Figure 14. illustrates the join and exit association pattern amongst players, providing a thorough understanding of the frequency of each player's involvement in the game and offering valuable insights into player retention.

Moving forward, we analyze the nature of player interaction through the lens of chat mentions in Figure 15. It highlights the extent of connectedness amongst players, acting as an indicator of the game's social aspect. Further, Figure 16 contributes to this narrative by representing the instances of chat responses, thus outlining the active dialogues within the gaming platform.

An interesting aspect of collective interaction comes to light through Figure 16 which emphasizes the most engaged Team Chat sessions. Figure 17, on the other hand, documents the longest conversation chains in the game. The maximum length observed is a chain

of 10 messages, thereby implying that prolonged chat exchanges are relatively infrequent amongst players. This finding could serve as a critical input for modifying marketing strategies, by diverting the focus from chat boxes to in-game sequences for advertising.

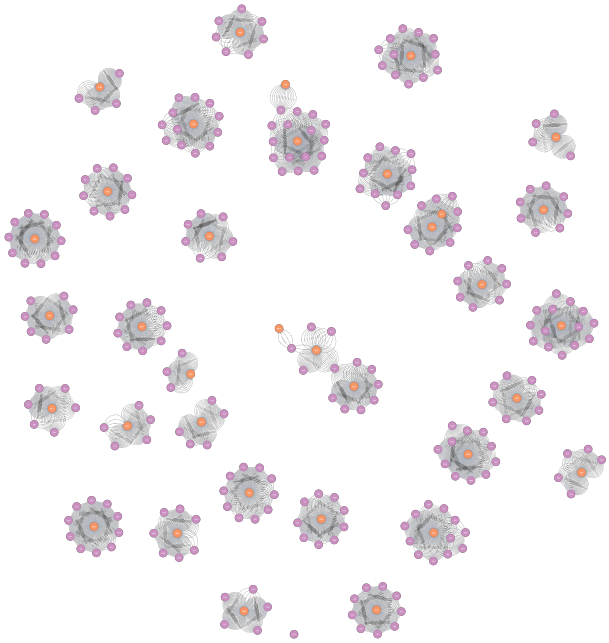

Figure 14: Join and Exit of Players

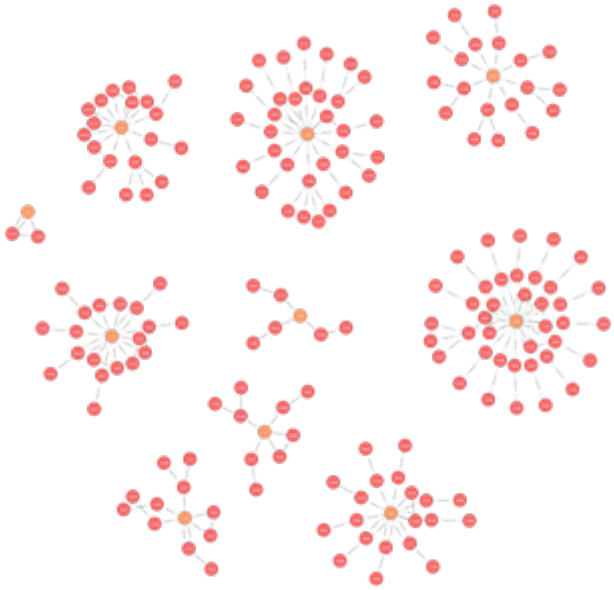

Figure 15: Caption

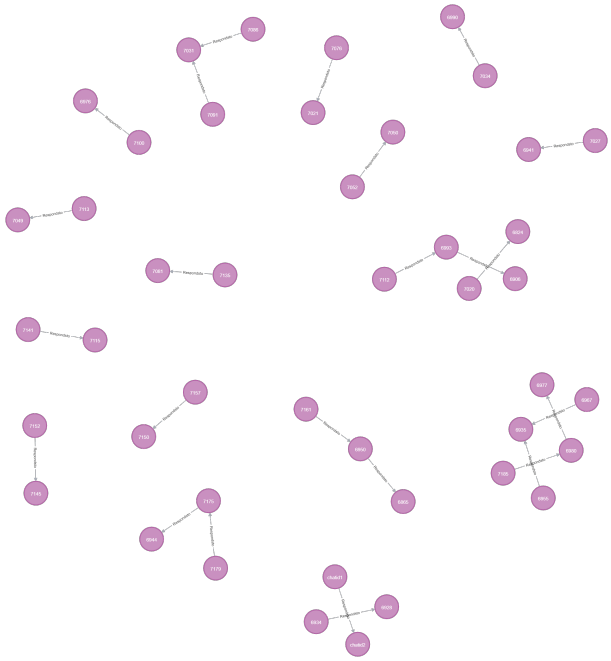

Figure 16: Caption

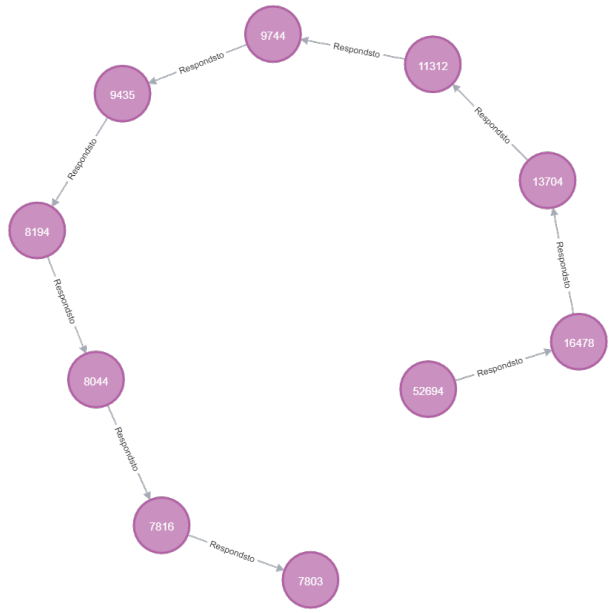

Figure 17: Caption

Finally, Table 8 compiles data on players who are frequently mentioned within chats. It reveals player "131" with the highest mentions at 53, followed by players "1204" and "621" each with 47 mentions. Additional players with significant mentions include "1428", "1506", and "283". This detailed enumeration can be instrumental in recognizing key influencers within the gaming community.

Team ID	Player Join Count
6792	100
6783	91
6925	87
6850	86
6791	81
6780	76
6809	72
6819	70
6790	67
6889	67

Table 7: Player Join Count per Team ID

Player ID	Mention Count
131	53
1204	47
621	47
1428	46
1506	46
283	42
674	42
1482	42
1450	42
1127	41

Table 8: Mention Count per Player ID

6 Big Data Ethics

Big data ethics is a branch of ethics that addresses the moral issues and responsibilities associated with the collection, analysis, and use of large datasets. It encompasses a wide range of ethical concerns, including privacy, consent, transparency, and data misuse.

Privacy is a significant concern in big data ethics. The collection and analysis of large amounts of personal data can lead to serious privacy violations if not properly managed (Mayer-Schönberger & Cukier 2013). Consent is another critical issue. Individuals should have the right to know and control how their data is being used.

Transparency in data processing is also crucial. Without it, there can be a lack of accountability and potential for unfair outcomes. Lastly, the misuse of data, such as for discriminatory practices or to manipulate behavior, is a pressing ethical concern.

Therefore, it's essential to establish ethical guidelines for big data practices to protect individuals' rights, promote fairness, and prevent misuse.

7 Conclusion, Findings and Recommendation

This report analyse the emerging trends in Big data. It's different processing paradigms, how they're merging together to answer varied solutions to the difficult problems faced during storage, processing & fetching of data. The report looks at the pressing ethical issues that arise from the storage & processing of big data. New laws that are likely to make a major shift in how Big data is handled. The report then develops a Big data solution using Spark for the "Catch the Pink Flamingo" game from the imaginary company named Eglence Inc. The solution performs EDA, Machine Learning & Graph analysis. The analysis has led to the following recommendations to improve the game.

7.1 Recommendation

The data analysis unveils certain recommendations to improve the gaming experience. Firstly, a significant dearth of players under the age of 19 has been noticed, while the majority are 30 or older. To broaden the game's appeal to a younger demographic, the incorporation of gameplay features and events catering to their interests is advisable.

In terms of revenue generation, most of it comes from Asia and Africa, even though Europe, North America, and South America boast stronger economies. By focusing more on these economically prosperous regions, the game's revenue could witness a substantial hike.

The revenue from in-game item purchases is predominantly derived from two products, out of a total of only six available items. It is proposed that Eglence Inc could expand their range of in-game items, which would not only diversify player options but also potentially enhance sales revenue.

With respect to differentiating spenders from non-spenders among users, it's suggested to employ an ensemble model comprising Logistic Regression and Decision Tree, which could yield a more precise and resilient classification.

The analysis indicates that lower-ranking teams are more inclined to purchase in-game items. Consequently, these players should be the prime focus of the advertising initiatives. Moreover, the most active teams, such as "6792" and "6783", should be targeted in advertising efforts.

Considering the longest recorded chat is limited to ten lines, it would be more beneficial to showcase advertisements before or after the gameplay, rather than in the chat box. This strategy resonates with the finding that lower rank teams and players are more susceptible to buying in-game items, implying that the development and marketing of these items should keep this audience segment in consideration.

Appendix

A Code

A.1 Code

References

Abadi, D. J., Carney, D., Çetintemel, U., Cherniack, M., Convey, C., Lee, S., Stonebraker, M., Tatbul, N. & Zdonik, S. (2003), 'Aurora: A new model and architecture for data

stream management', *The VLDB Journal* **12**(2), 120–139.
URL: *https://doi.org/10.1007/s00778-003-0095-z*

Akidau, T., Balikov, A., Bekiroglu, K., Chernyak, S., Haberman, J., Lax, R., McVeety, S., Mills, D., Nordstrom, P. & Whittle, S. (2013), Millwheel: Fault-tolerant stream processing at internet scale, *in* 'Proceedings of the VLDB Endowment', Vol. 6, p. 1033–1044.
URL: *https://dl.acm.org/doi/abs/10.14778/2536222.2536229*

Apache Hadoop (2023). Accessed: 2023-05-18.
URL: *https://en.wikipedia.org/wiki/Apache*$_H$*adoop*

Casado, R. & Younas, M. (2014), 'Emerging trends and technologies in big data processing', *Concurrency and Computation: Practice and Experience* **27**, n/a–n/a.

Casado, R. & Younas, M. (2015), 'Emerging trends and technologies in big data processing', *Concurrency and Computation: Practice and Experience* **27**(8), 2078–2091.
URL: *https://onlinelibrary.wiley.com/doi/full/10.1002/cpe.3398*

Chandio, A., Tziritas, N. & Xu, C.-Z. (2015), 'Big-data processing techniques and their challenges in transport domain', *ZTE Communications* **13**, 50–59.

Cheng, H., Rong, C. & Y., H. K. W. W. L. (2015), 'Secure big data storage and sharing scheme for cloud tenants', *China Communications* **12**(6), 106.
URL: *http://www.cic-chinacommunications.cn/EN/abstract/article*$_1$*27.shtml*

Davenport, T. H., Barth, P. & Bean, R. (2012), 'How'big data'is different', *MIT Sloan Management Review* **54**(1), 5.
URL: *https://www.hbs.edu/ris/Publication20Files/SMR-How-Big-Data-Is-Different*$_7$*82ad61f* − *8e5f* − *4b1e* − *b79f* − *83f33c903455.pdf*

Davis, K. (2012), *Ethics of Big Data: Balancing Risk and Innovation*, O'Reilly Media, Inc.

Dean, J. & Ghemawat, S. (2004), Mapreduce: Simplified data processing on large clusters, *in* 'Proceedings of the 6th Symposium on Operating System Design and Implementation', San Francisco, CA, pp. 137–150.

Dean, J. & Ghemawat, S. (2008), 'Mapreduce: simplified data processing on large clusters', *Communications of the ACM* **51**(1), 107–113.
URL: *https://dl.acm.org/doi/abs/10.1145/1327452.1327492*

Jain, P., Gyanchandani, M. & Khare, N. (2016), 'Big data privacy: a technological perspective and review', *Journal of Big Data* **3**.

Kitchin, R. & McArdle, G. (2016), 'What makes big data, big data? exploring the ontological characteristics of 26 datasets', *Big Data & Society* **3**(1), 1–10.

Li, N., Li, T. & Venkatasubramanian, S. (2007), t-closeness: Privacy beyond k-anonymity and l-diversity, *in* '2007 IEEE 23rd International Conference on Data Engineering', pp. 106–115.

Liu, X., Iftikhar, N. & Xie, X. (2014), Survey of real-time processing systems for big data, *in* 'Proceedings of the 18th International Database Engineering & Applications Symposium, IDEAS '14', Association for Computing Machinery, pp. 356–361.
URL: *https://doi.org/10.1145/2628194.2628251*

Madden, S. (2012), 'From databases to big data', *IEEE Internet Computing* **16**(3), 4–6.

Marz, N. & Warren, J. (2015), *Big Data: Principles and Best Practices of Scalable Real-time Data Systems*, 1st edn, Manning Publications Co., USA.

Matt, K. (2022), 'Taking a regulation-agnostic approach to privacy'.
URL: *https://hyperproof.io/resource/regulation-agnostic-approach-privacy/*

Mayer-Schönberger, V. & Cukier, K. (2013), *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Houghton Mifflin Harcourt.

Nisar, M. U., Fard, A. & Miller, J. A. (2013), Techniques for graph analytics on big data, *in* '2013 IEEE International Congress on Big Data', pp. 255–262.

Oussous, A., Benjelloun, F.-Z., Ait Lahcen, A. & Belfkih, S. (2018), 'Big data technologies: A survey', *Journal of King Saud University - Computer and Information Sciences* **30**(4), 431–448.
URL: *https://www.sciencedirect.com/science/article/pii/S1319157817300034*

Qiu, J., Wu, Q., Ding, G., Xu, Y. & Feng, S. (2016), 'A survey of machine learning for big data processing', *EURASIP Journal on Advances in Signal Processing* **2016**.

Sagiroglu, S. & Sinanc, D. (2013), Big data: A review, *in* '2013 International Conference on Collaboration Technologies and Systems (CTS)', pp. 42–47.

Shahrivari, S. (2014), 'Beyond batch processing: Towards real-time and streaming big data', *Computers* **3**(4), 117–129.
URL: *https://www.mdpi.com/2073-431X/3/4/117*

Trevino, A. (2019), 'Introduction to k-means clustering'.
URL: *https://blogs.oracle.com/ai-and-datascience/post/introduction-to-k-means-clustering*

Vuleta, B. (2022), 'How much data is created every day? [27 powerful stats]'.
URL: *https://seedscientific.com/how-much-data-is-created-every-day/*

White, T. (2015), *Hadoop: The Definitive Guide*, 4th edn, O'Reilly Media, Inc.

Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B. & Vasilakos, A. V. (2016), 'Big data: From beginning to future', *International Journal of Information Management* **36**(6, Part B), 1231–1247.
URL: *https://www.sciencedirect.com/science/article/pii/S0268401216304753*