STEAM REVIEW ANALYSIS

Presented by -

Rindhuja Johnson
Shiva Kumar Goud Mucharla
Sujay Deevela

Introduction

Steam

- Online Game Platform: Available for use in almost 29 languages
- Utilizes HDFS for big data storage and retrieval for analysis

Project Overview

Extended EDA on the Steam Reviews Dataset 2021

Game Recommendation with Collaborative Filtering

Project Outcome

- Understand game popularity trend, gamer language preferences and gaming patterns
- Implement a recommendation system model for games using ALS.

Objectives

1 Games Analysis

- Determine the Most Popular Games
- Determine the Most recommended games

2 Demographic Analysis

- Determine the Most Popular Language
- Determine the languages used for Popular games

3 Time Pattern Analysis

- Analysis how the popularity for games changed across the months
- Analyze the trend in popularity of 5 most reviewed games in each quarter of 2021

4 Author/Review Analysis

- Determine any relation between number of games reviewed and total reviews by a single gamer.
- Analyze the play time of the gamers at the time of review, after review, and forever.
- Analyze any pattern in true/false recommendations based on when the player reviews.

5 Game Recommendation

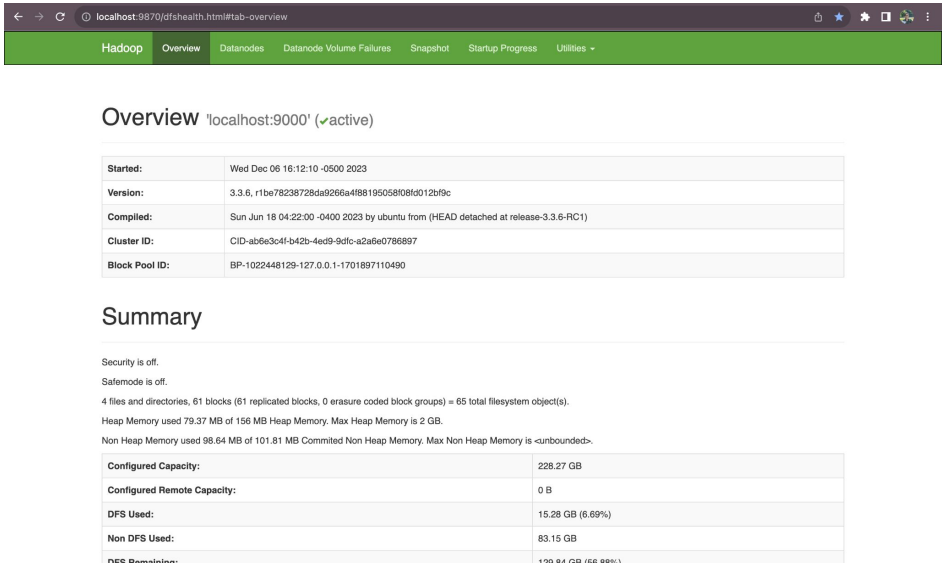
- Use Alternating Least Squares (ALS) algorithm from Spark MLLib
- Obtain at least 5 recommendation games for each game.

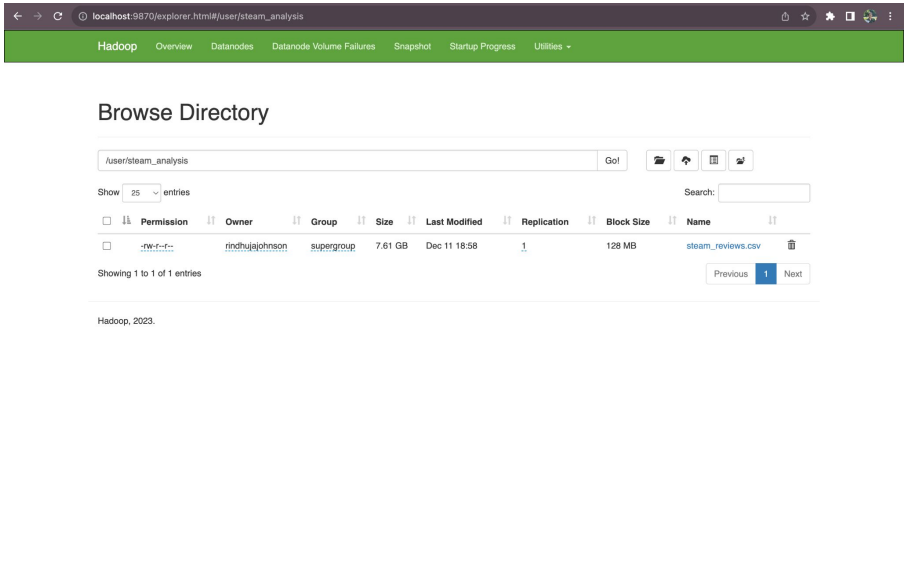
Data Sources and Collection

- Data Source: Kaggle - Steam Reviews 2021
- Comprises approximately 8 GB of data
- 17 million rows and 23 columns.
- The dataset provides detailed information about individual game reviews, including attributes such as the review text, author details, game information, and more.

The data was downloaded to the local, moved to the HDFS directory, and extracted using Pyspark session

Data Storage in HDFS

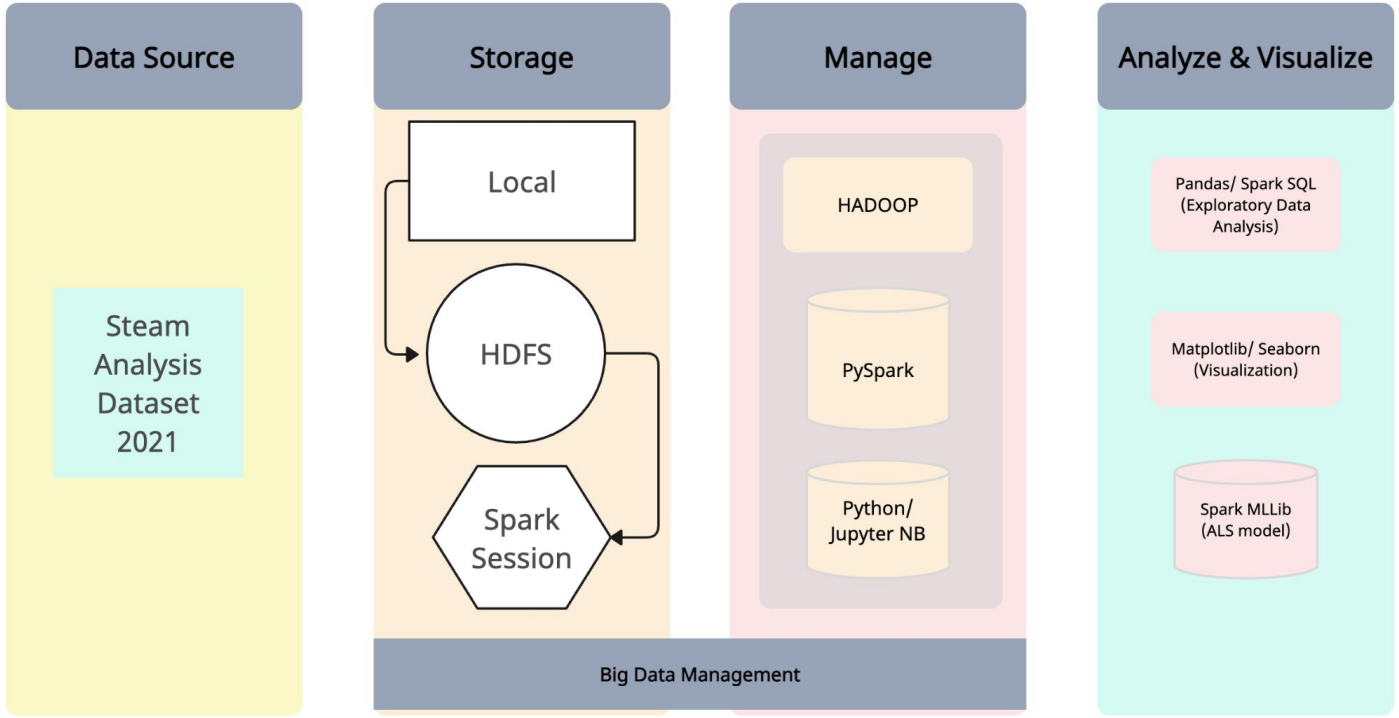




Tools and Technologies

- Hadoop Distributed File System (HDFS): Used for storing the large dataset.
- PySpark: Used for data extraction, cleaning and pre-processing.
- Spark SQL: Used for extraction, processing, and analysis of the data set.
- Pandas: Employed for exploratory data analysis.
- Matplotlib and Seaborn: Utilized for data visualization to derive meaningful insights.
- Spark MLLib: Used the ALS collaborative filtering algorithm for recommending games.

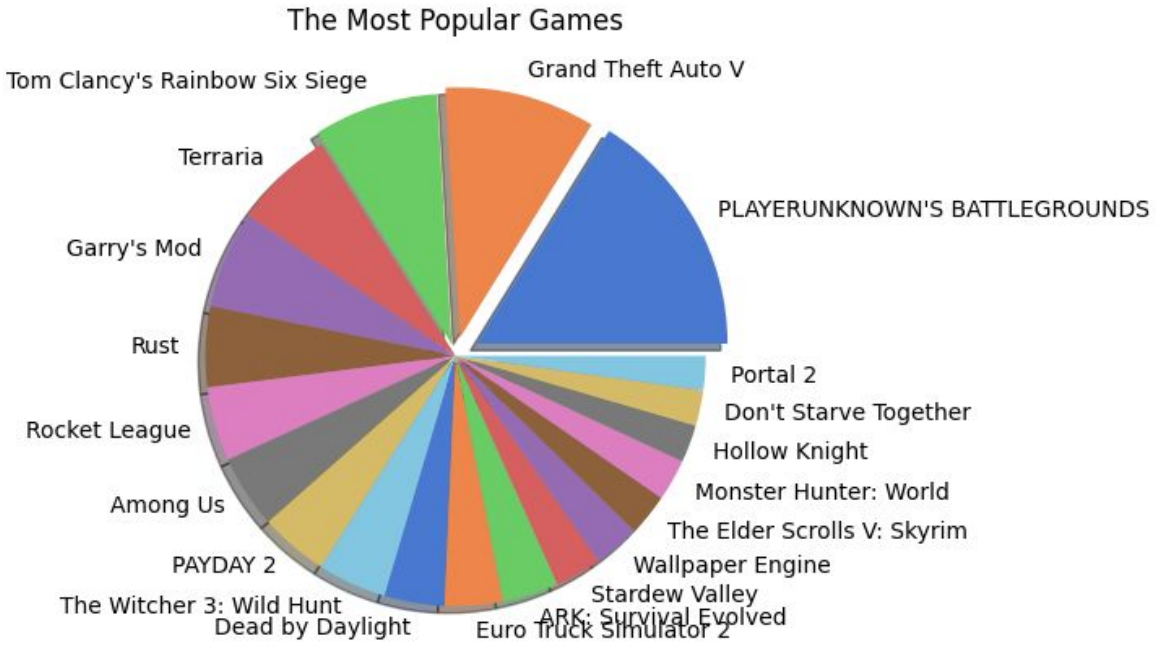
Data Stack Diagram



Analysis & Insights

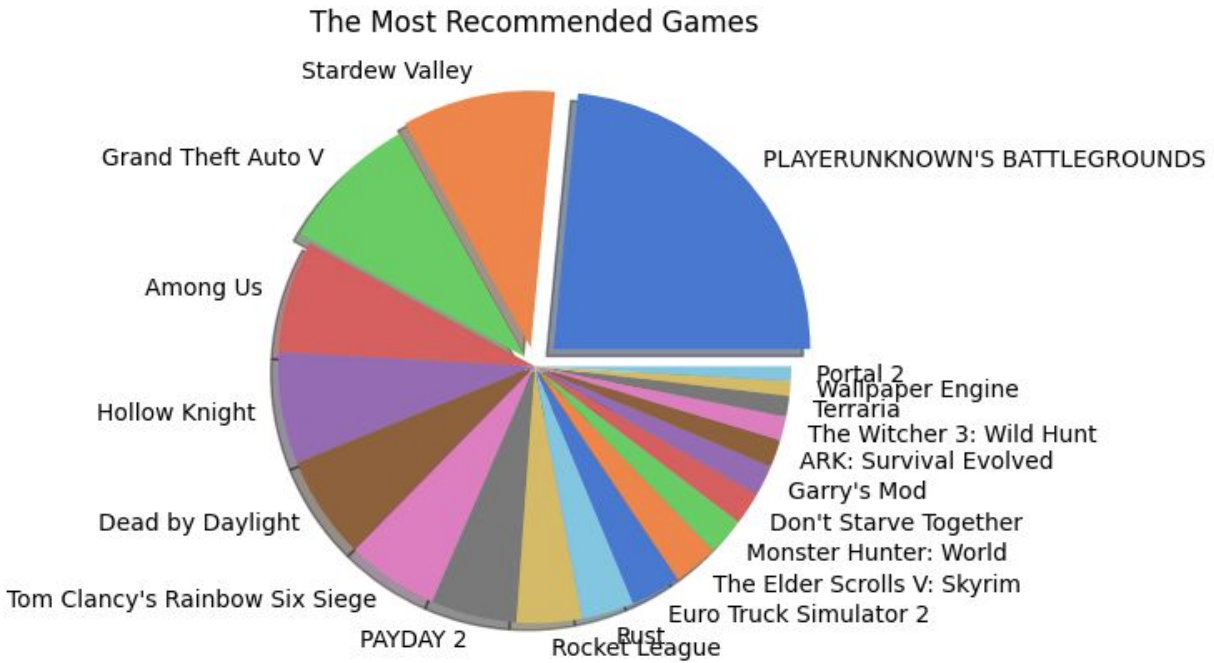
- Most popular games needed **not** be the most recommended ones!
- English is the most popular review language, however, Schinese is highly used by the gamers of the most popular games.
- There was a hike in the number of reviews received from October to December in the year of 2021.
- Most of the reviewers seem to have stopped playing the game shortly after giving the review. When the average of total time played and time played at the time of review is 202 hours and 167 hours respectively, the time played after review is only 35 hours!

Which games are most reviewed by gamers?



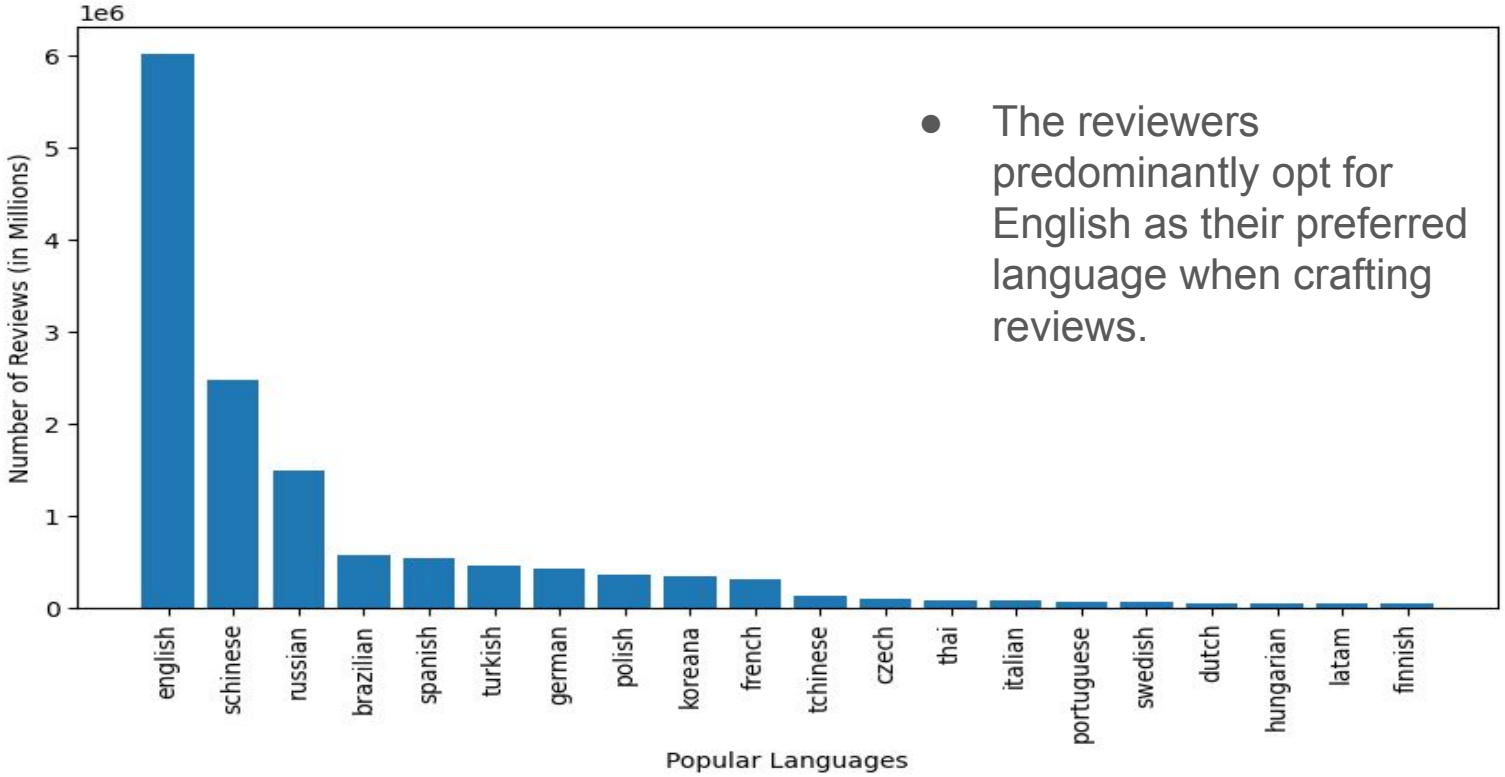
- PUBG stands out as the most popular game in comparison to all others in the gaming landscape.
- Grand Theft Auto (GTA V) holds the position of being the second most popular game in the gaming industry.

Does more reviews mean a highly recommended one?



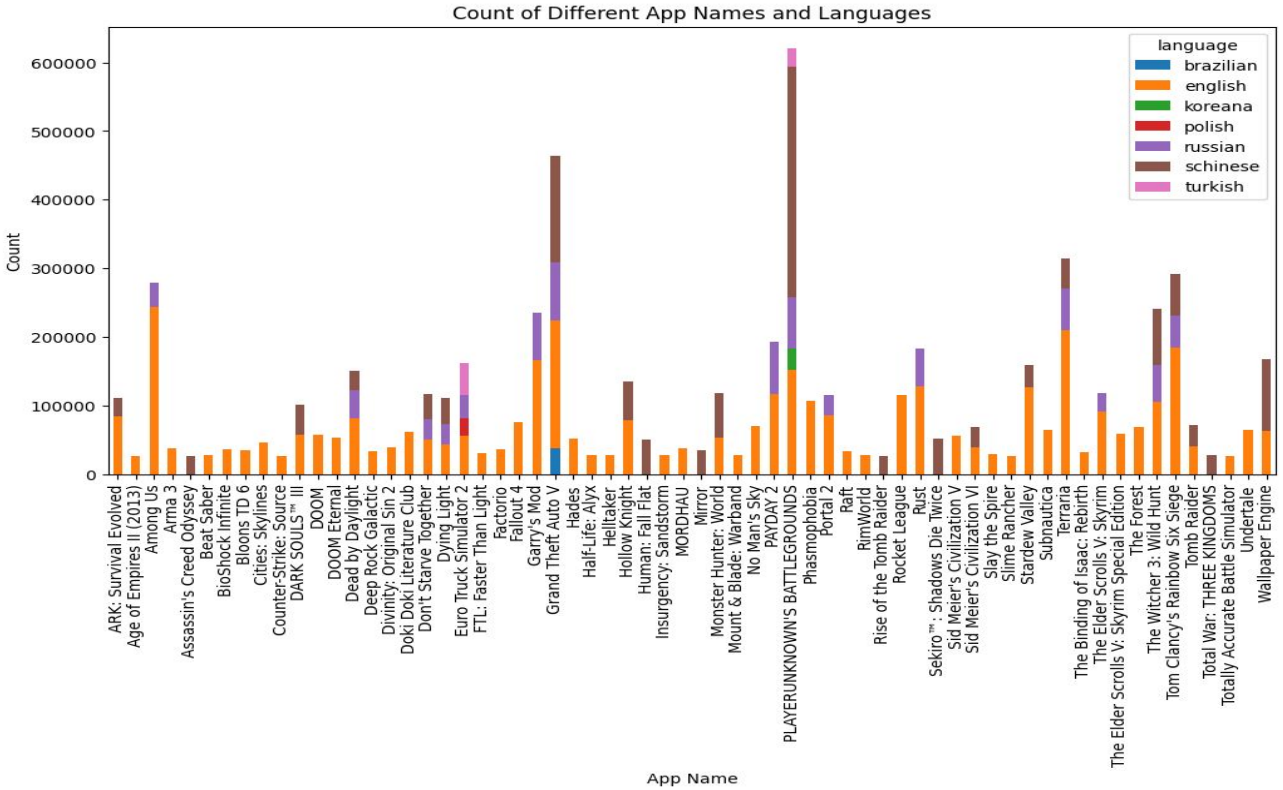
- According to our model prediction, PUBG emerges as the most recommended game among all reviewers..

Which language does most players use?



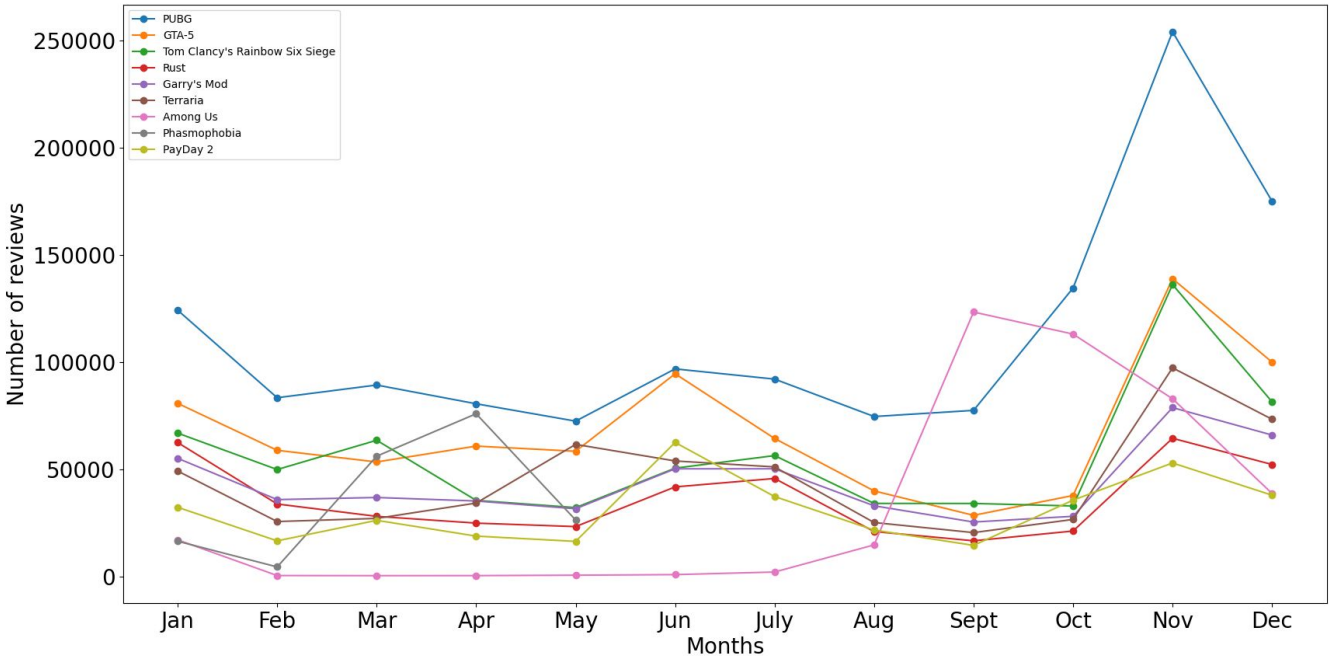
- The reviewers predominantly opt for English as their preferred language when crafting reviews.

Is there any linguistic trend in gaming - games, or review style?



- The predominant language used in the reviews for the most popular game is Schinese.
- A higher percentage of PUBG players use Schinese compared to other languages.

How did the reviewing trend for games vary throughout the year of 2021?

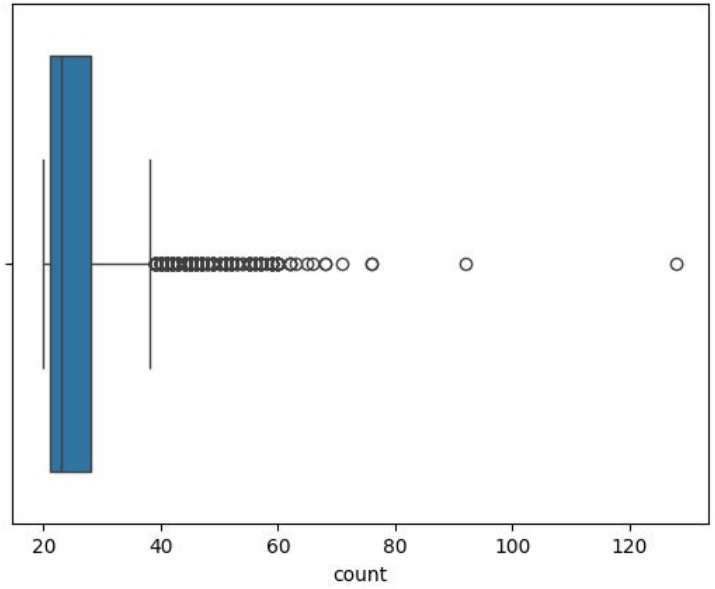


- During the months of October to December, PUBG experienced the highest player count compared to other months.

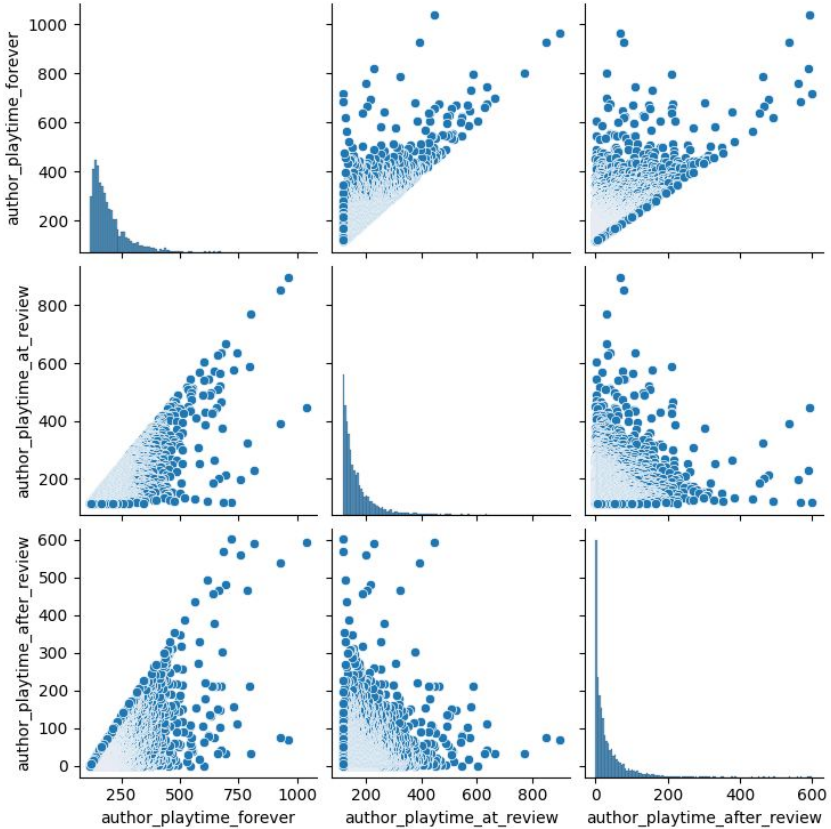
Does a player review more than one game?

	author_steamid	count
0	76561198315585536	128
1	76561198239163744	90
2	76561198112669681	76
3	76561198057221241	70
4	76561198038757354	70
...	...	...
4995	76561198334598369	19
4996	76561198043176189	19
4997	76561198338467392	19
4998	76561198116417799	19
4999	76561198338888728	19

- The highest number of games reviewed by an author is 128 and has an average of around 25.



When do the players usually review the game?



- The three variables - playtime_at_review, playtimw_after_review, and playtime_forever - give the insight that most of the players reviewed the game and did not continue playing for long

Game Recommendation using ALS from Spark ML Library

- Alternating Least Squares algorithm: A collaborative filtering model in Spark ML library
- For each game, we found 5 recommended games which can be provided to new users
- The recommendations are based on the gaming history of existing players.

author_index	recommendations
26	[{545, 5.110069}, {23, 5.0741754}, {32, 4.9596233}, {16, 4.921337}, {236, 4.593187}]
27	[{124, 7.9713182}, {61, 7.3510895}, {659, 7.1857305}, {519, 7.1357174}, {480, 7.0634856}]
28	[{252, 6.4423847}, {212, 6.064964}, {528, 5.785183}, {519, 5.609013}, {108, 5.532746}]
31	[{140, 6.260532}, {108, 5.997504}, {211, 5.6662126}, {558, 5.578244}, {116, 5.475158}]
34	[{631, 5.137752}, {0, 4.973231}, {7, 4.961696}, {4, 4.9556036}, {321, 4.827304}]
44	[{428, 6.420355}, {320, 5.9340463}, {101, 5.7874656}, {845, 5.6964426}, {225, 5.6908174}]
53	[{702, 5.6680846}, {324, 5.264774}, {690, 5.1133165}, {5, 5.10114}, {77, 5.0023475}]
65	[{507, 6.2150884}, {664, 5.956748}, {396, 5.612644}, {4, 5.311983}, {627, 5.2957177}]
76	[{165, 8.520896}, {487, 8.438303}, {421, 8.228063}, {481, 7.7405367}, {141, 7.6824822}]
78	[{415, 7.122621}, {512, 6.273131}, {267, 5.9899096}, {74, 5.974411}, {625, 5.9524136}]
81	[{350, 5.63751}, {816, 5.3708744}, {70, 5.2583594}, {241, 5.21456}, {340, 5.192349}]
85	[{222, 5.295117}, {828, 5.1678305}, {682, 5.153618}, {685, 5.0253687}, {1, 4.961582}]
101	[{342, 1.322423}, {826, 1.3036016}, {442, 1.2901685}, {481, 1.2434942}, {165, 1.2193744}]
103	[{48, 5.0112233}, {139, 4.9997683}, {453, 4.9245925}, {656, 4.8900204}, {514, 4.794793}]
108	[{141, 5.4130588}, {23, 5.0113544}, {13, 4.987937}, {643, 4.5259447}, {366, 4.405514}]
115	[{631, 1.0593572}, {718, 1.0053402}, {2, 0.99875695}, {0, 0.9948738}, {311, 0.9258937}]
126	[{351, 5.8417525}, {217, 5.4630923}, {841, 5.160208}, {535, 5.1304216}, {557, 5.0728765}]
133	[{627, 6.684775}, {531, 6.2489}, {38, 6.0673504}, {139, 6.037232}, {379, 6.010695}]
137	[{645, 7.115007}, {570, 7.019359}, {364, 6.9159355}, {194, 6.6930475}, {33, 6.5604224}]
148	[{1, 4.995995}, {3, 4.98124}, {828, 4.61103}, {779, 4.403008}, {524, 4.1038485}]

only showing top 20 rows

Future Work

- Implement advanced natural language processing (NLP) techniques for deeper sentiment analysis.
- Explore deep learning models for enhanced game recommendations.
- Extend the analysis to include user interactions, such as likes and comments on reviews.

Challenges

- Big Data Processing
- Data Quality and Cleaning
- Complexity of ALS Model
- Scalability

References

Steam Reviews
https://www.kaggle.com/code/gonzafrancoandres/steam-reviews

Review Analysis by PySpark
https://www.kaggle.com/code/iplori/review-analysis-by-pyspark

Install Hadoop on MacOS (Macbook M1)
https://codewitharjun.medium.com/install-hadoop-on-macos-efe7c860c3ed

Through comprehensive analysis across these categories, the Steam Review Analysis project seeks to provide valuable insights for game developers, platform administrators, and researchers interested in understanding user sentiments and behaviors within the Steam gaming community.