Analyzing Data Patterns on IGDB Gaming Dataset

“Group 15”

Xiaohai Zheng, Jiangrong Liu, Zhuoxin Liu, Suraj P N, Zixi Chen

> *Video games: Not just play, but a window into technology, culture, and creativity*

> *A cultural shift: From arcade cabinets to digital ecosystems*

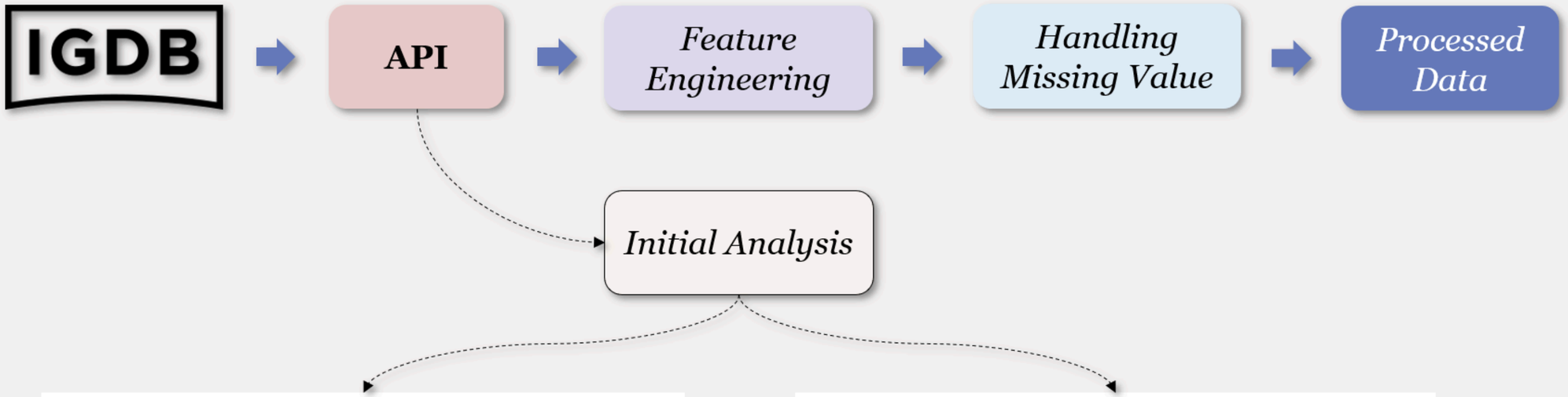
> *IGDB: At the crossroads of gaming's past, present, and future*

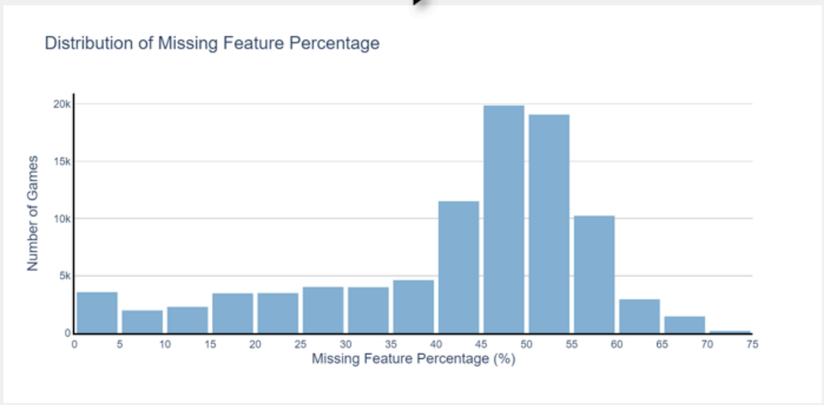
IGDB - Internet Game DataBase

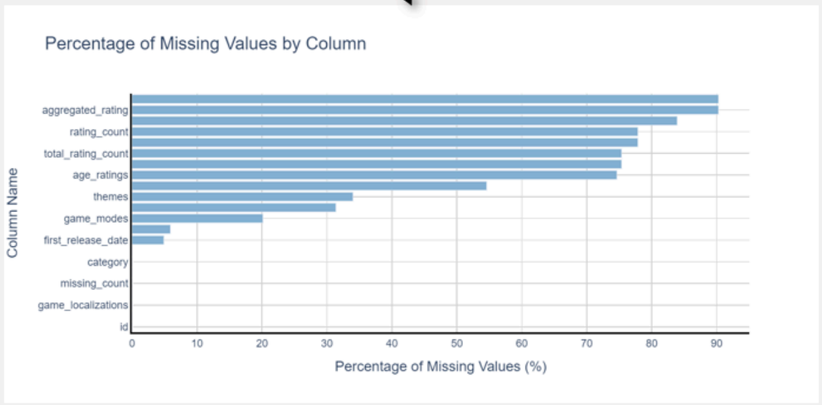
- *Data diversity:*
  - *Cataloging over 255,000 games, over 90,000 PC games*
  - *Games Released between: 1970-2024*
  - *Various features (game genre, rating, engine, etc.)*

- Data *accessibility: open user-friendly API*

Data Preprocessing Pipeline







For detailed data preprocessing and feature engineering process please refer to the GitHub repo.

Contents

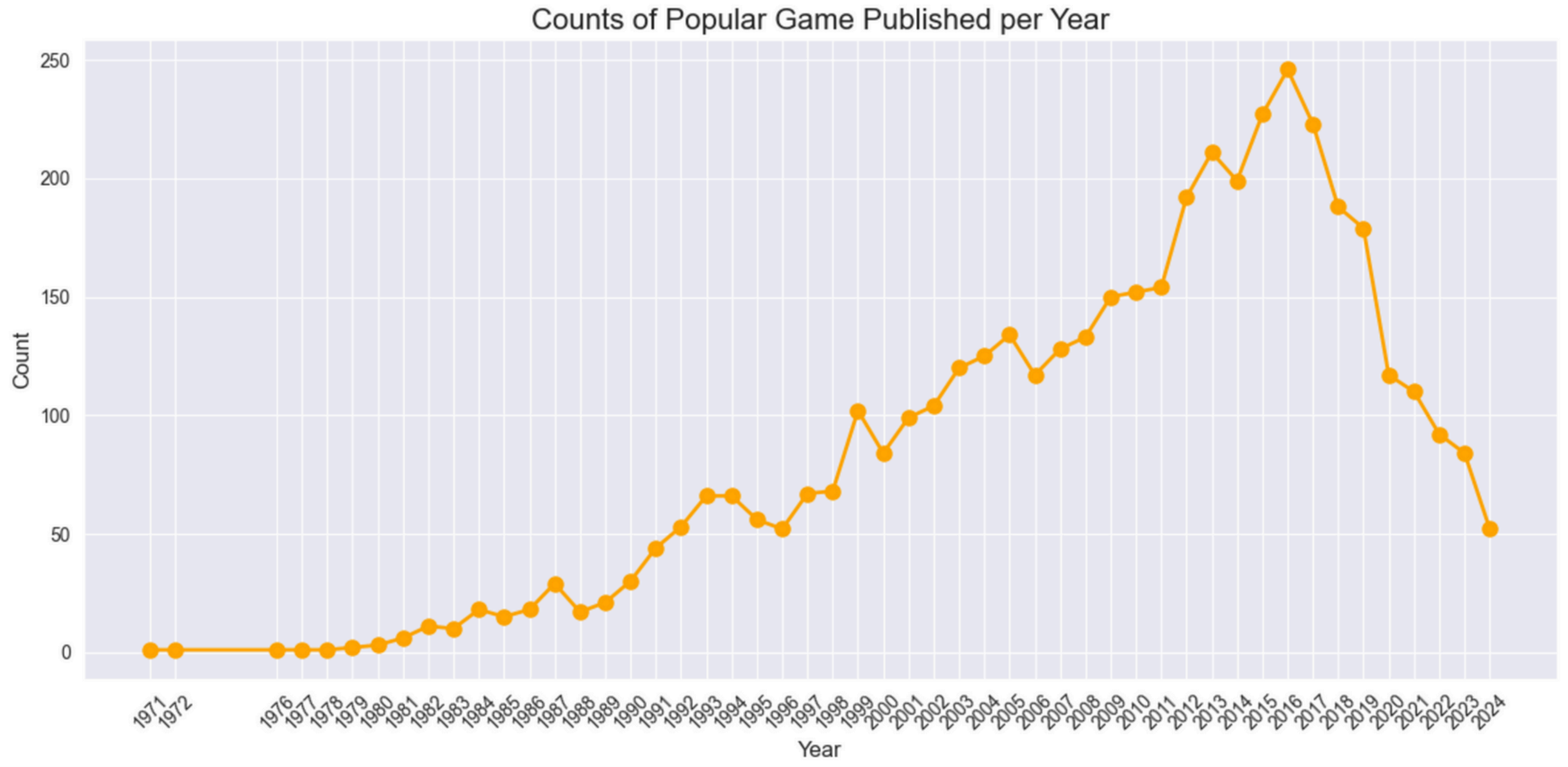
Game Development & Localization Strategy

With our Analysis, game companies can:

1. Identify Emerging Trends
2. Optimize Localization Efforts
3. Make Informed Decisions on Game Engine choice

Part 1

“Game Popularity Trends: The 2010s—A Golden Era for Top Titles”



Trending Genres: Most Popular Genres Over Time

Understand which genres have become more popular over time

***Analysis Steps**:*

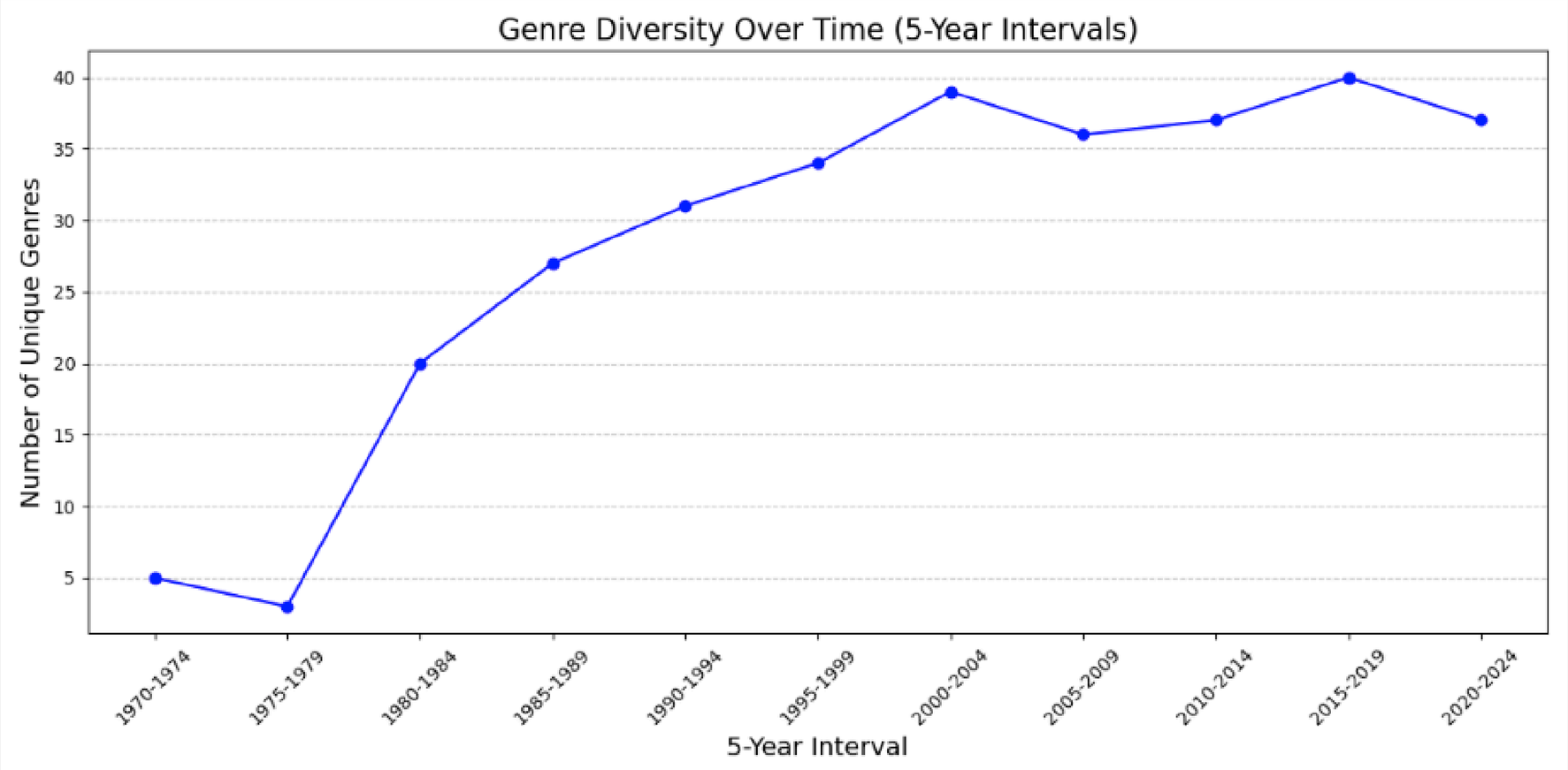
- *Analyze release_year and release_month to identify the rise or decline of specific genres over time*
- *Visualize the number of games released in each genre by year/decade*
- *Correlate the genre popularity with the overall aggregated rating or player rating*

***Visualizations**:*

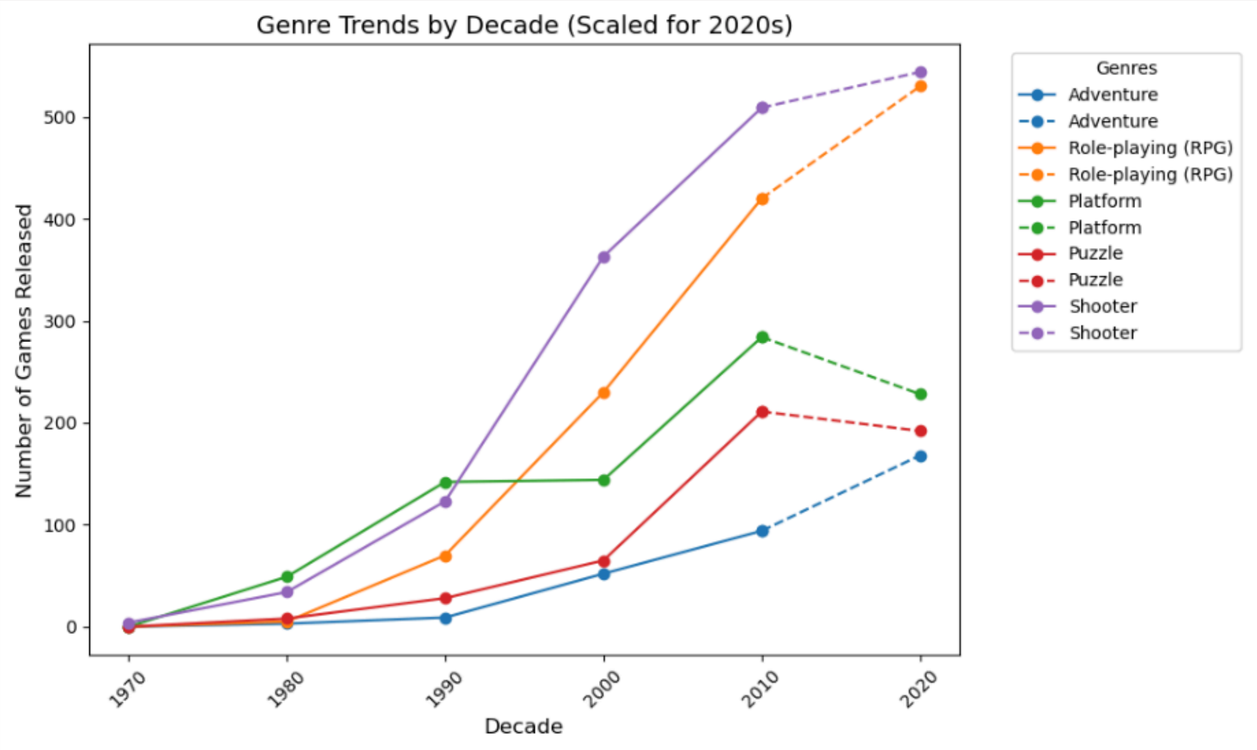
1. *Genre Diversity Over Time (5-Year Intervals)*
2. *Game Trends by Decade*
3. *Game Popularity by Average Rating across Decades (Top 5 Genres)*

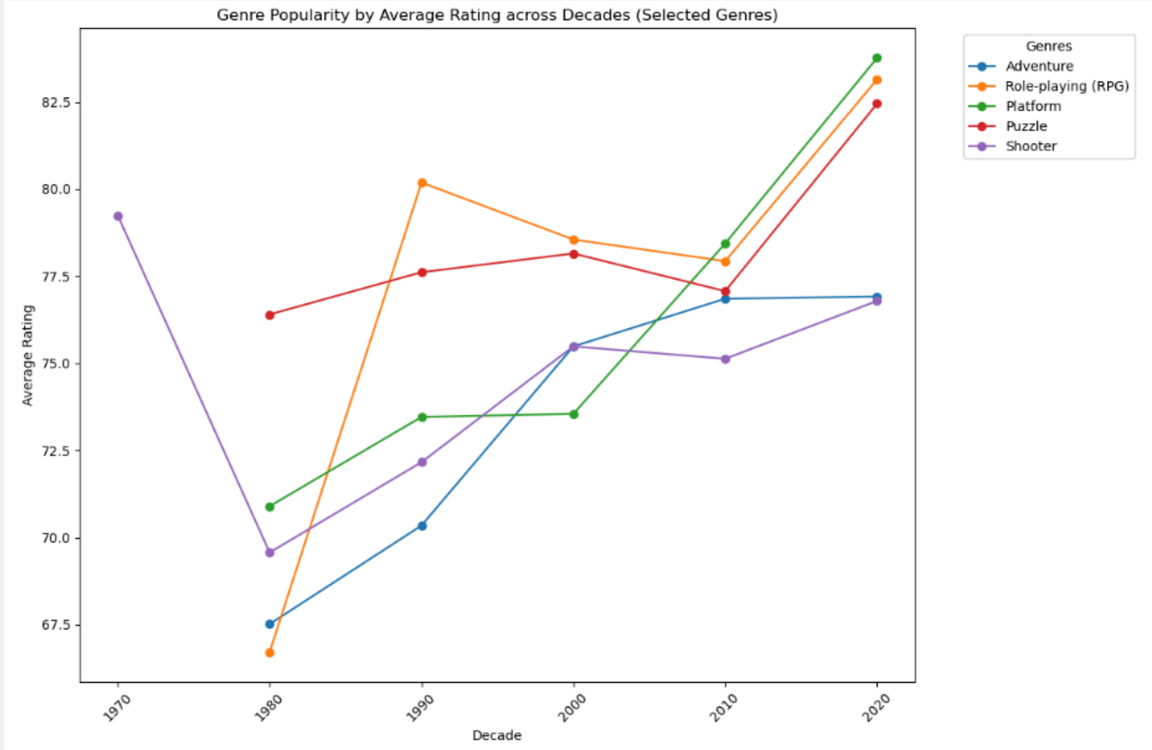
***Features Used**: Genres, Rating, Rating Count, Initial Release Date*

“Rising Genre Diversity Over Time: Highlighting the Growing Importance of Innovation and Variety in Gaming”



“RPGs Surpass Shooters in Popularity, with Rising Ratings for RPG and Platform Games Over Decades”





Data Equally Scaled across genres for 2020s as data is available only until 2024

Game Engine Preferences by Genre

Identify which game engines are most commonly used for specific genres and correlate with popularity

Analysis Steps:

- *Analyze the game_engines column and group by genre to find the most commonly used engines*
- *Count the number of games developed using each engine and analyze the relationship with game popularity (ratings, release success)*
- *Co-relate the genre popularity with the overall aggregated rating or player rating*

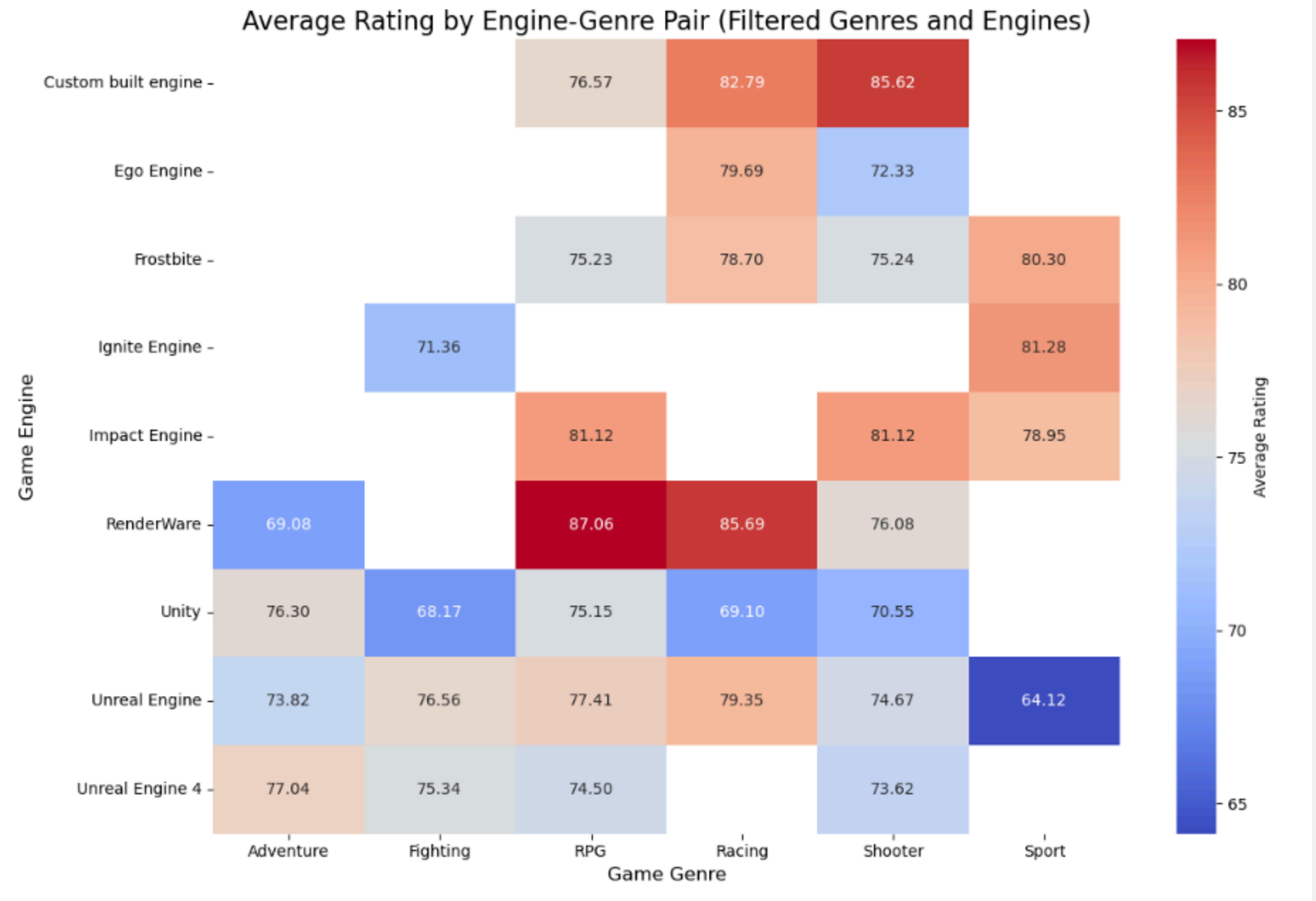
Visualizations*:*

1. *Heatmap of Average rating by Engine-Genre pair*
2. *Bubble Chart of Engine-Genre Popularity by Game Count*

Features Used*: Genres, Rating, Rating Count, Game Engines*

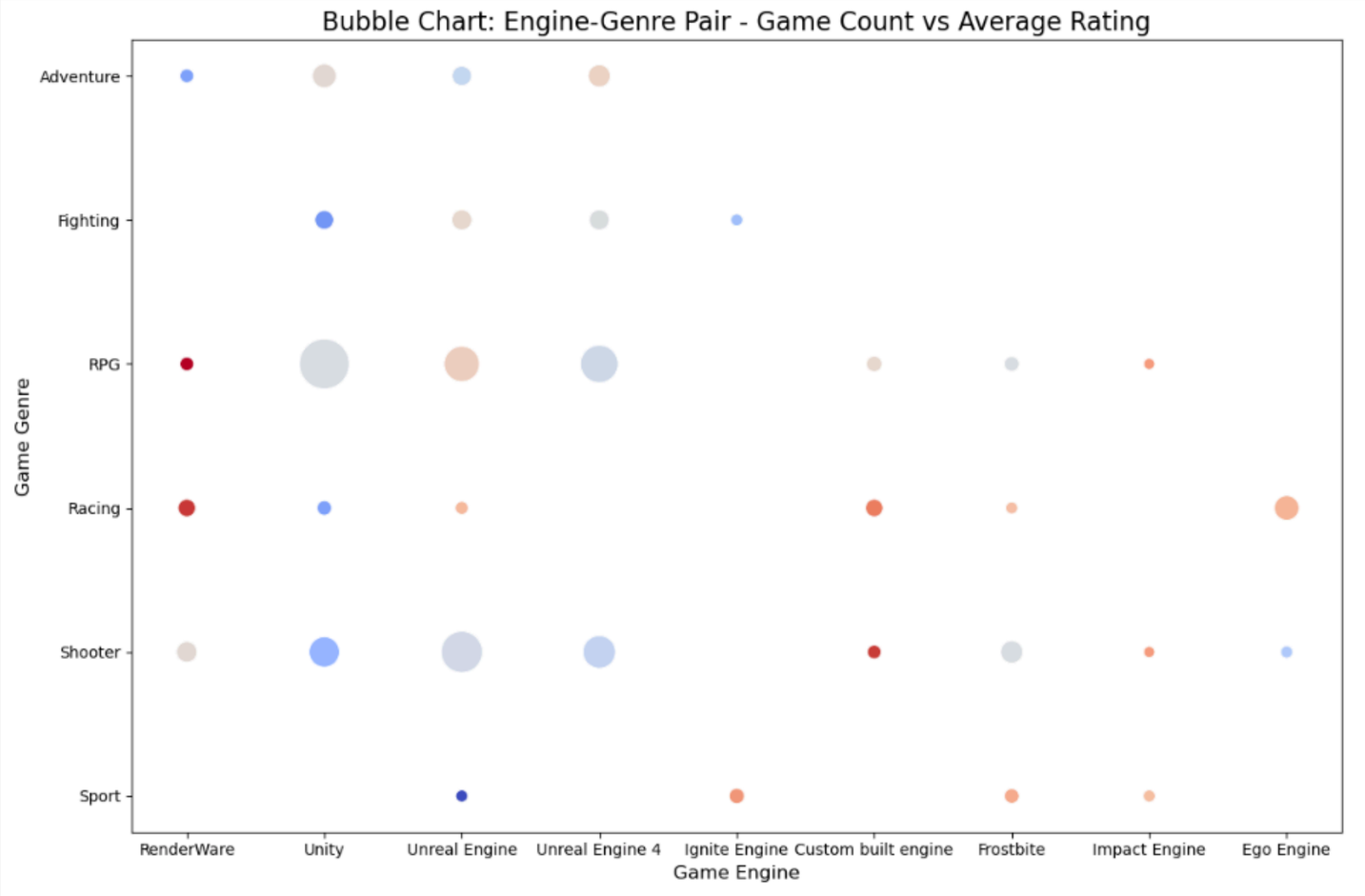
Engine-Genre Ratings Heatmap

“RenderWare Excels with RPG and Racing Genres”



Engine-Genre Popularity

“Unity Dominates RPG Game Development”



Company-Engine & Company-Localization Relationship

Explore the Relationship Between Companies, Game Engines, and Localization.

Analysis Steps:

- Identify which game engines are most commonly used by each company
- Correlate company size with engine popularity
- Assess how companies localize games to different regions
- Compare the aggregated ratings of games based on their localization categories (None, Minimal, Extensive)

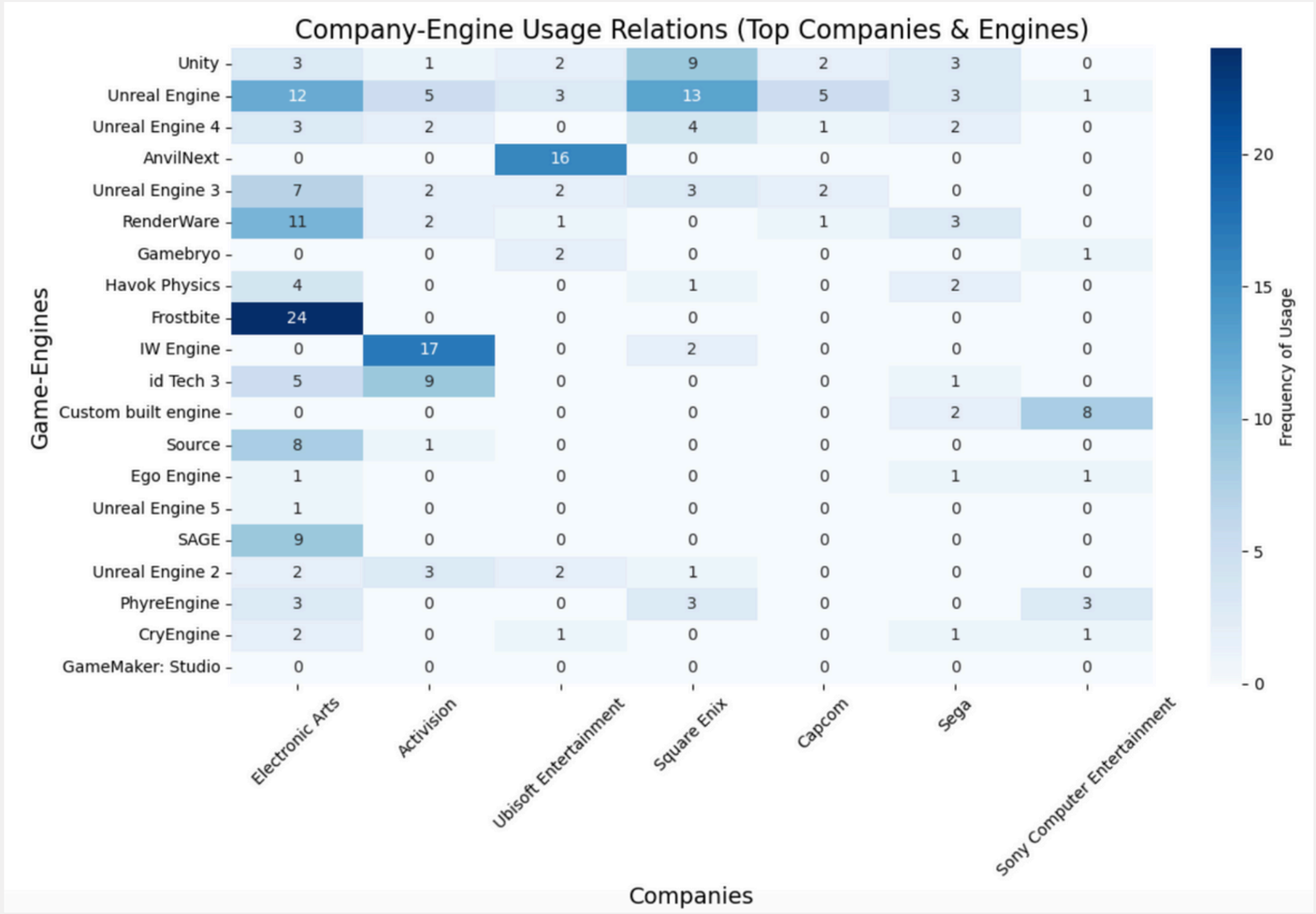
Visualizations*:*

1. *Heatmap of Company-Engine Usage*
2. *Distribution of Ratings by Localization Categories*

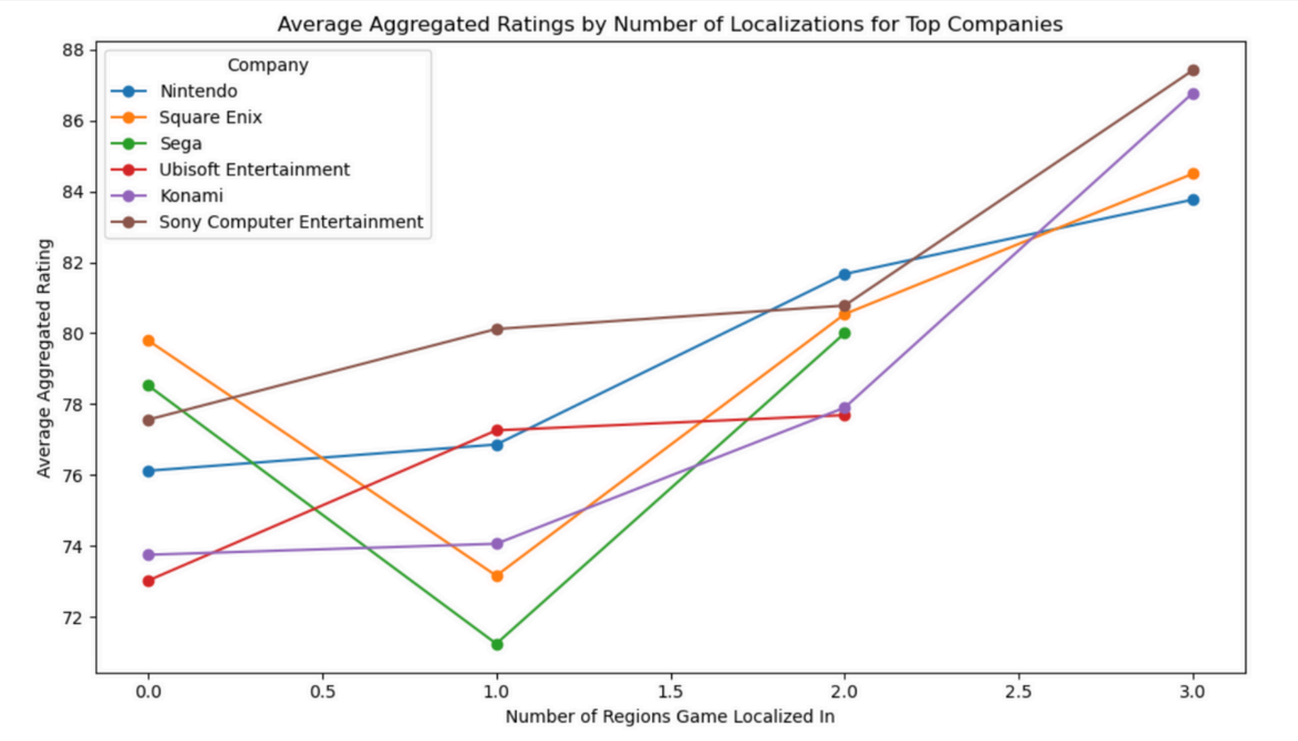
Features Used*: Genres, Rating, Rating Count, Game Engines, Company, Localization, Languages*

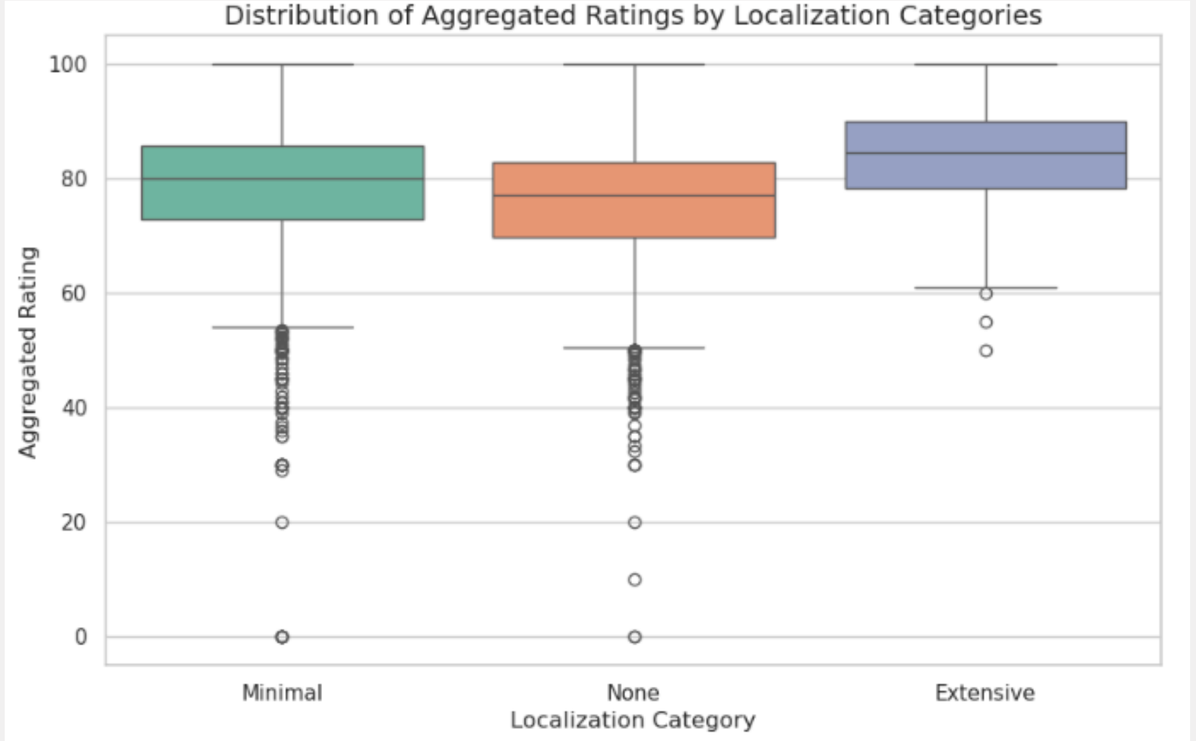
Company-Engine Usage

“Electronic Arts (Frostbite) & Activision (IW Engine) Dominate”



“Localization Drives Higher Company Ratings”





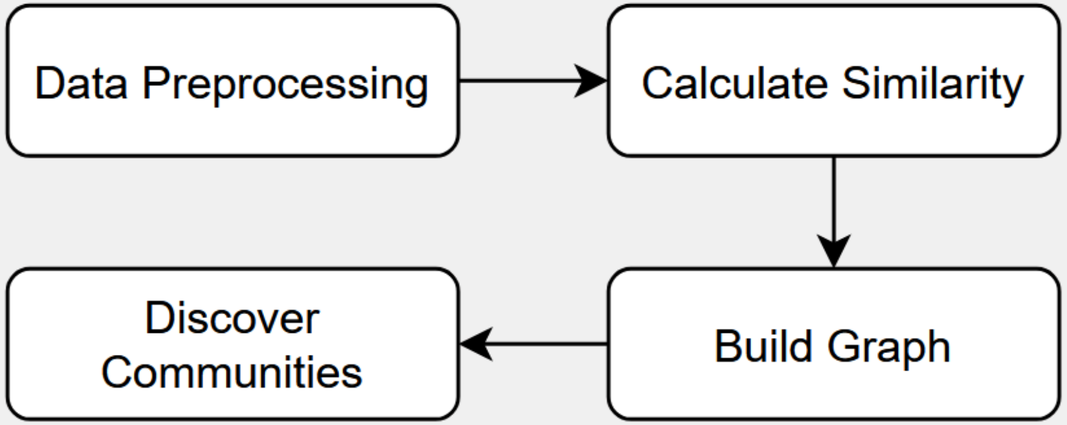
Minimal: 1-2 Localizations
None: No Localization
Extensive 3+ Localizations

Clustering & Community Insights

Can we divide the dataset into clusters consisting of similar games and find relation & insights?

Part 2

Analysis Workflow



	genres	keywords	themes
0	5, 10, 31	3, 21, 22, 25, 30, 57, 64, 72, …	1, 27, 33, 3
1	12, 31	129, 151, 537, 592, 623, 770, 8…	1, 17, 38
2	5, 8, 9, 31	575, 592, 962, 1158, 1181, 1293…	1, 18, 27
3	12, 31	22, 96, 129, 151, 159, 211, 221…	1, 17, 23, 3
4	5, 8, 9	129, 137, 558, 575, 603, 962, 1…	1, 18, 27
5	5, 10, 31	21, 57, 58, 72, 109, 129, 155, …	1, 23, 38
6	5	3, 5, 429, 605, 660, 1089, 1158…	1, 18, 19

id	name
1	Action
17	Fantasy
18	Science fiction
19	Horror
20	Thriller
21	Survival
22	Historical
23	Stealth
27	Comedy
28	Business
31	Drama
32	Non-fiction
33	Sandbox
34	Educational
35	Kids
38	Open world
39	Warfare
40	Party
41	4X (explore, expand, exploit, and exterminate)
42	Erotic
43	Mystery
44	Romance

Determine similarties

- *Three matrices with the same weight*
- *Text processing using list: abc ->[’a’, ‘b’, ‘c’]*
- *Dice-Sørensen similarity*
- *Large data: divided into bathces*

$$\mathrm{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|}$$

Dice-Sørensen: Unordered; Value [0,1]; intuitively captures overlap; efficent

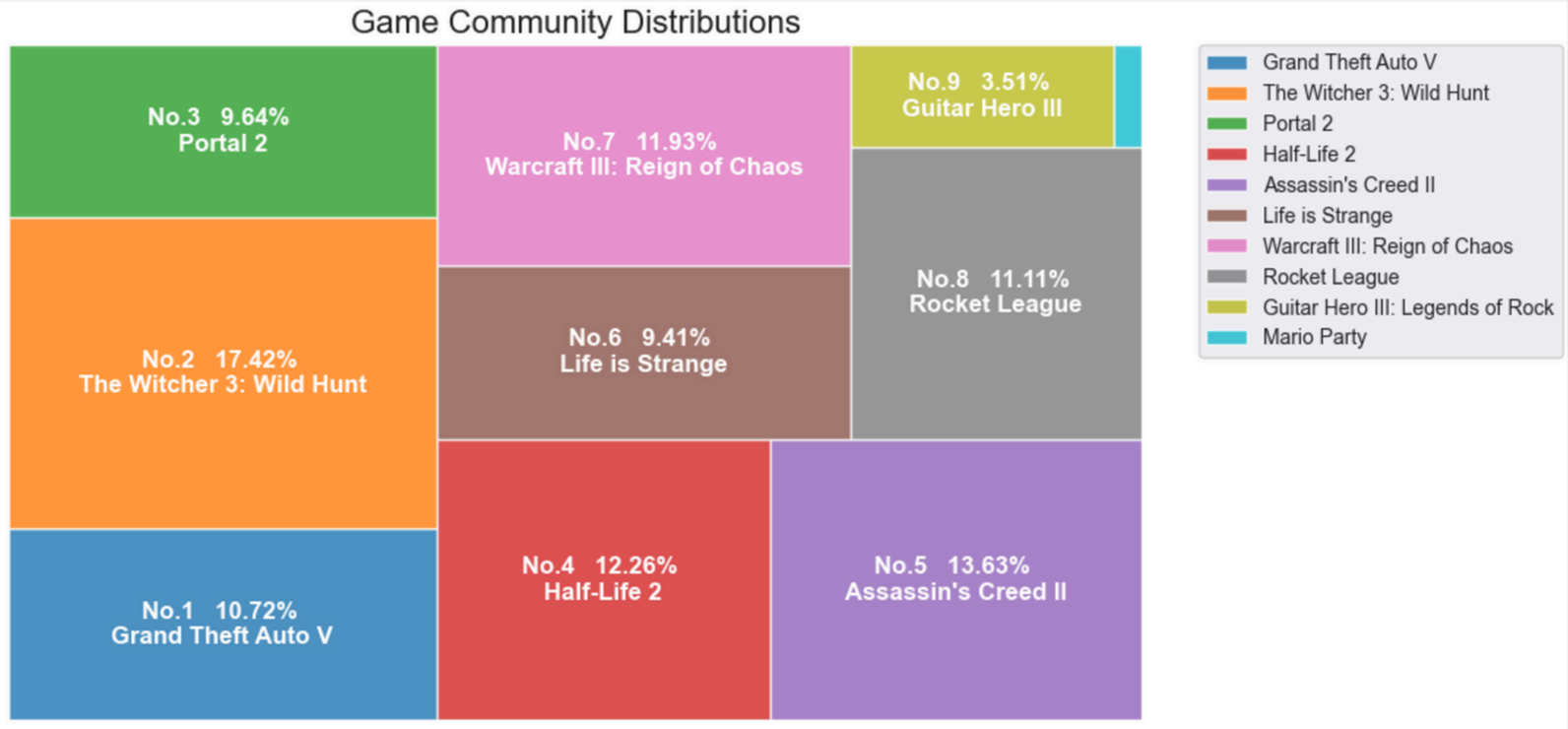
Graph Construction

- *Selecting thresholds: link nodes with similarity >= 0.5*
- *Implemented different kinds of modularity algorithms, chose Louvain*
- *Only keep communities with size >= 5*
- *10 communities in total. 16 - 759 games each*

	Algorithm: Louvain
Approach	*Modularity optimization via local greedy moves and aggregation*
Efficiency	*Highly efficient, scalable to large graphs* *O(nlogn)*
Output	*A flat partitioning of communities*
Strength	*Scales well to large graphs, produces high-modularity partitions*

Results & Analysis

- *Clusters: use the most rated game inside to represent the community*
- *Games have been successfully divided into communities*

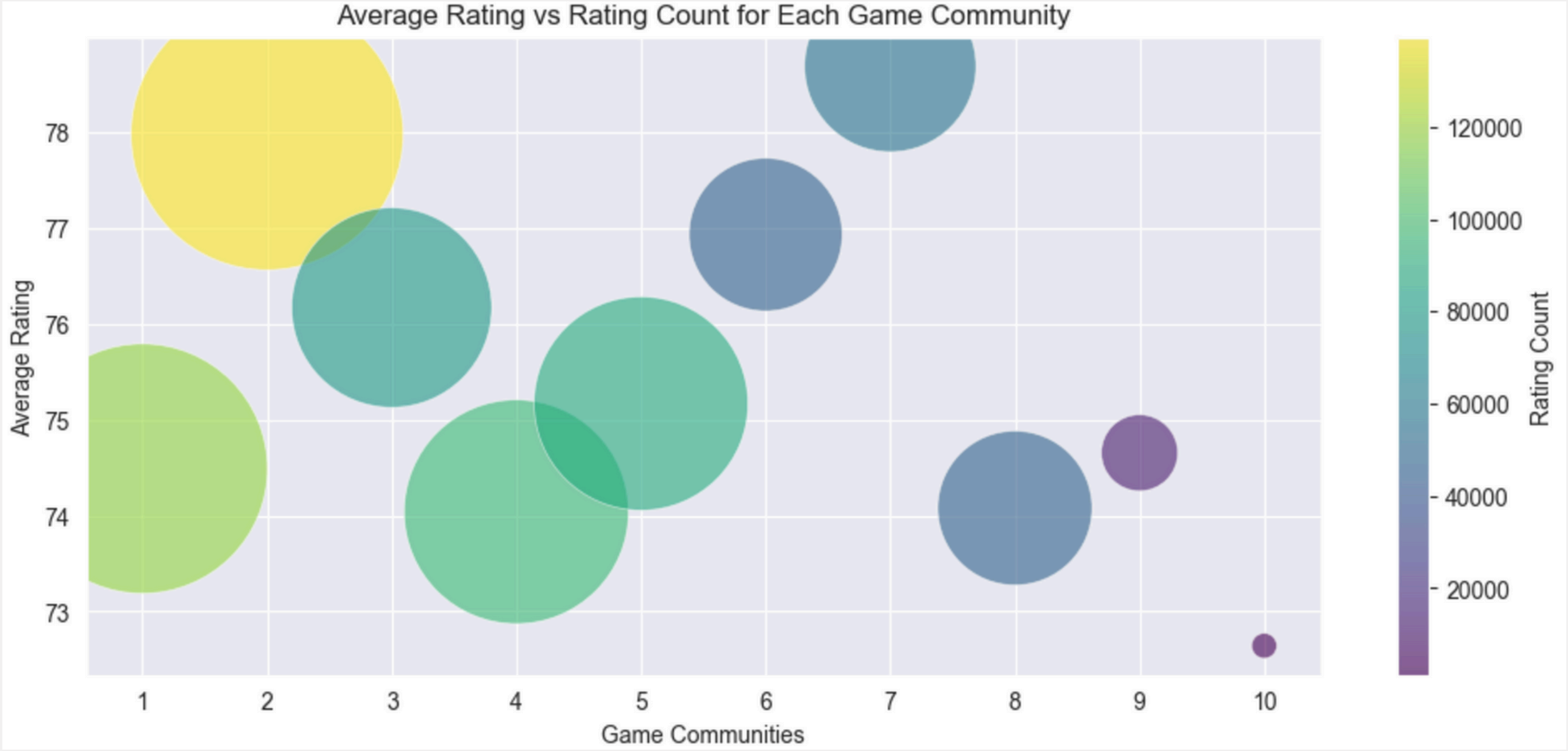


Visualization

No.1 10.72%
Grand Theft Auto V

No.6 9.41%
Life is Strange

Visualization



Game Rating Prediction

Can we predict the rating of a game based on its features?

Part 3

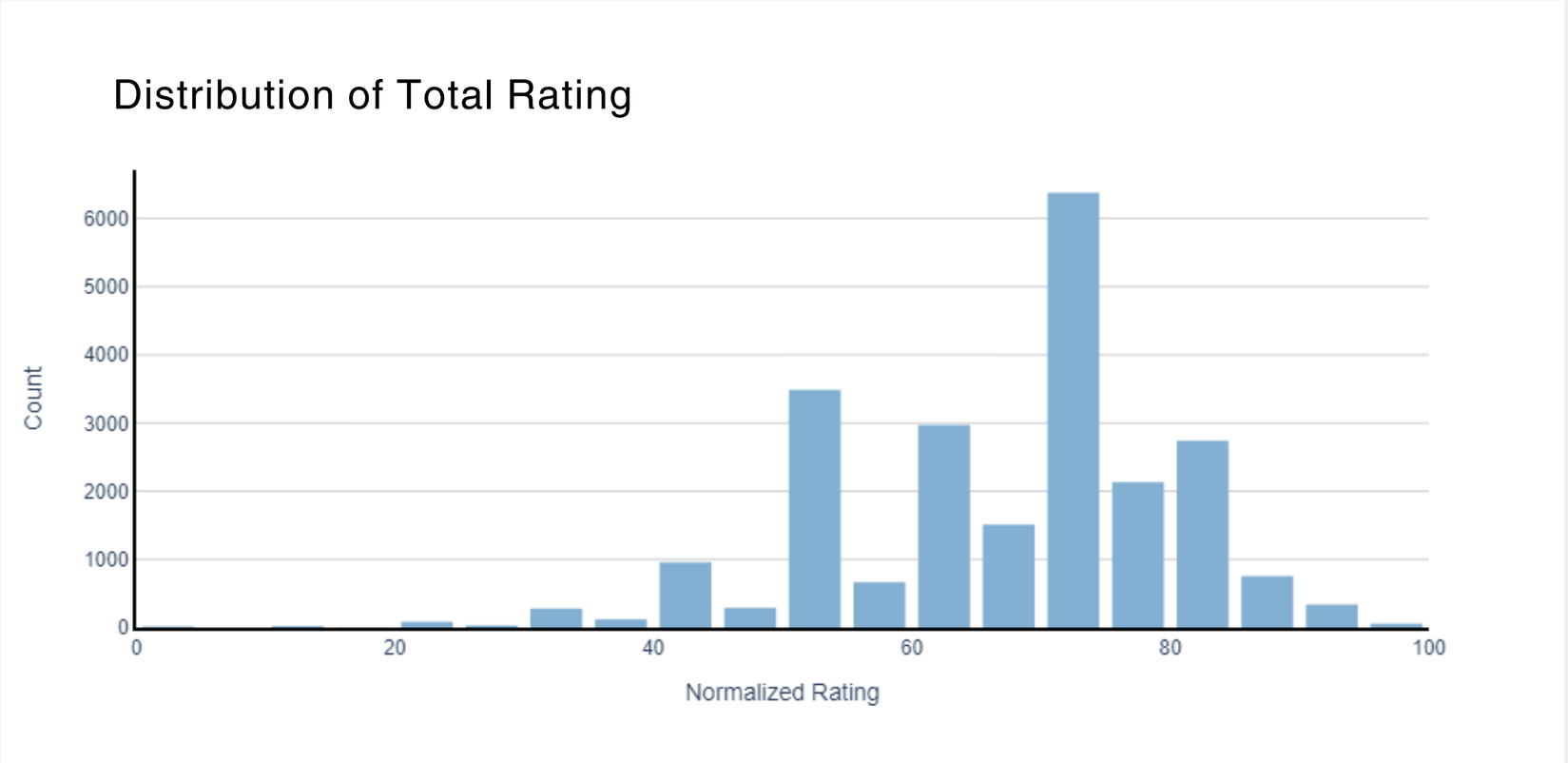
Game Rating Prediction

Data Sample:

- *75 features*
- *3-Class Classification*
- *Training (Cross Validation)*
  Testing

Evaluation Metrics:

- *F1-Score*
- *ROC AUC Score*
- *Accuracy*



Game rating is affected by many random factors inside games and in reality. Direct regression is not practical yet not necessary. Therefore we label each game data into three classes based on the distribution.

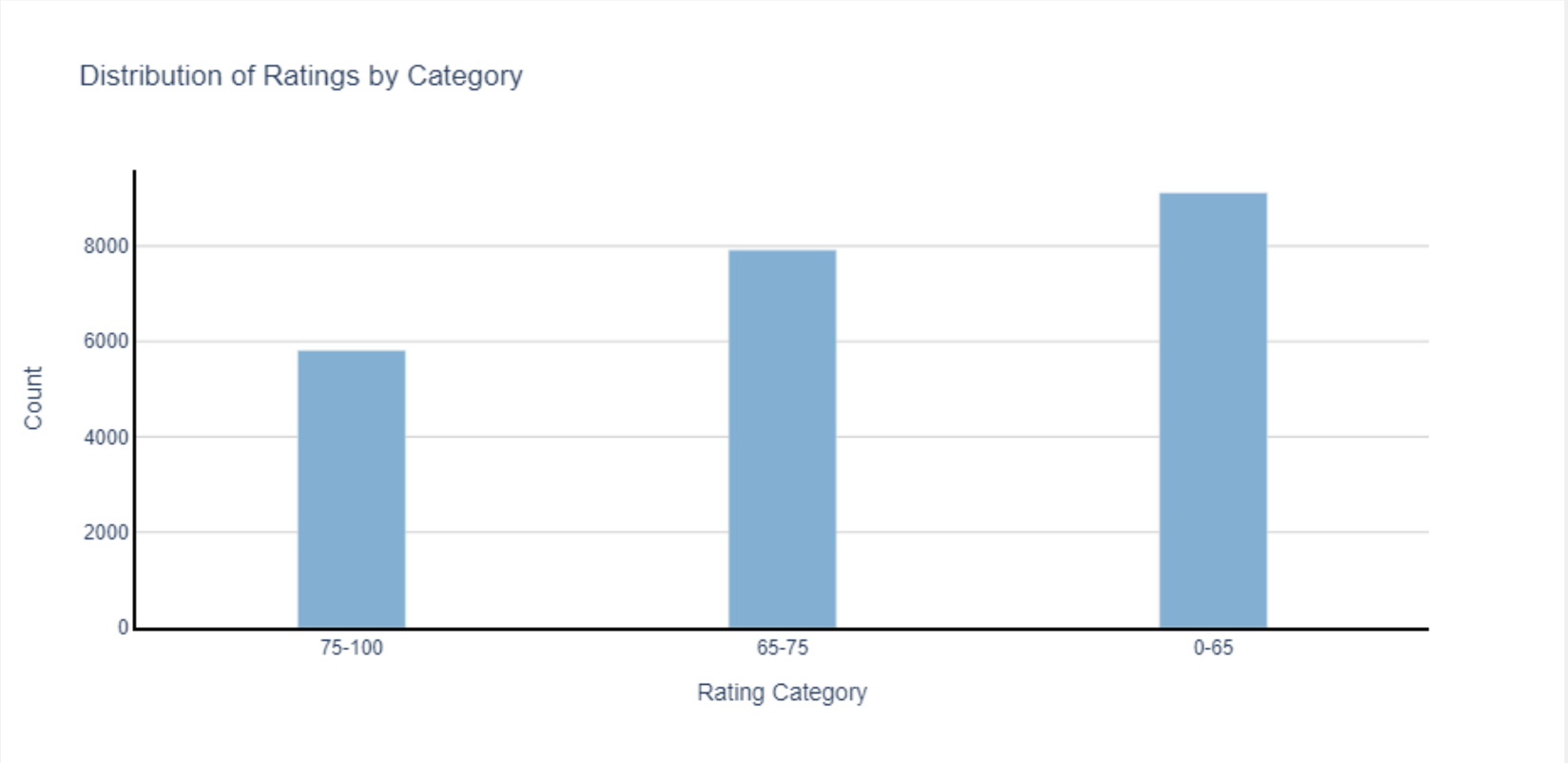
Game Rating Prediction

Data Sample:

- *75 features*
- *3-Class Classification*
- *Training (Cross Validation)*
  Testing

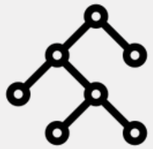
Evaluation Metrics:

- *F1-Score*
- *ROC AUC Score*
- *Accuracy*



Game rating is affected by many random factors inside games and in reality. Direct regression is not practical yet not necessary. Therefore we label each game data into three classes based on the distribution.

ML Methods



Random Forest

Advantages

- *Robust to overfitting*
- *Handles missing data and outliers well*

Disadvantages

- *Limited optimization*

XGBoost

Advantages

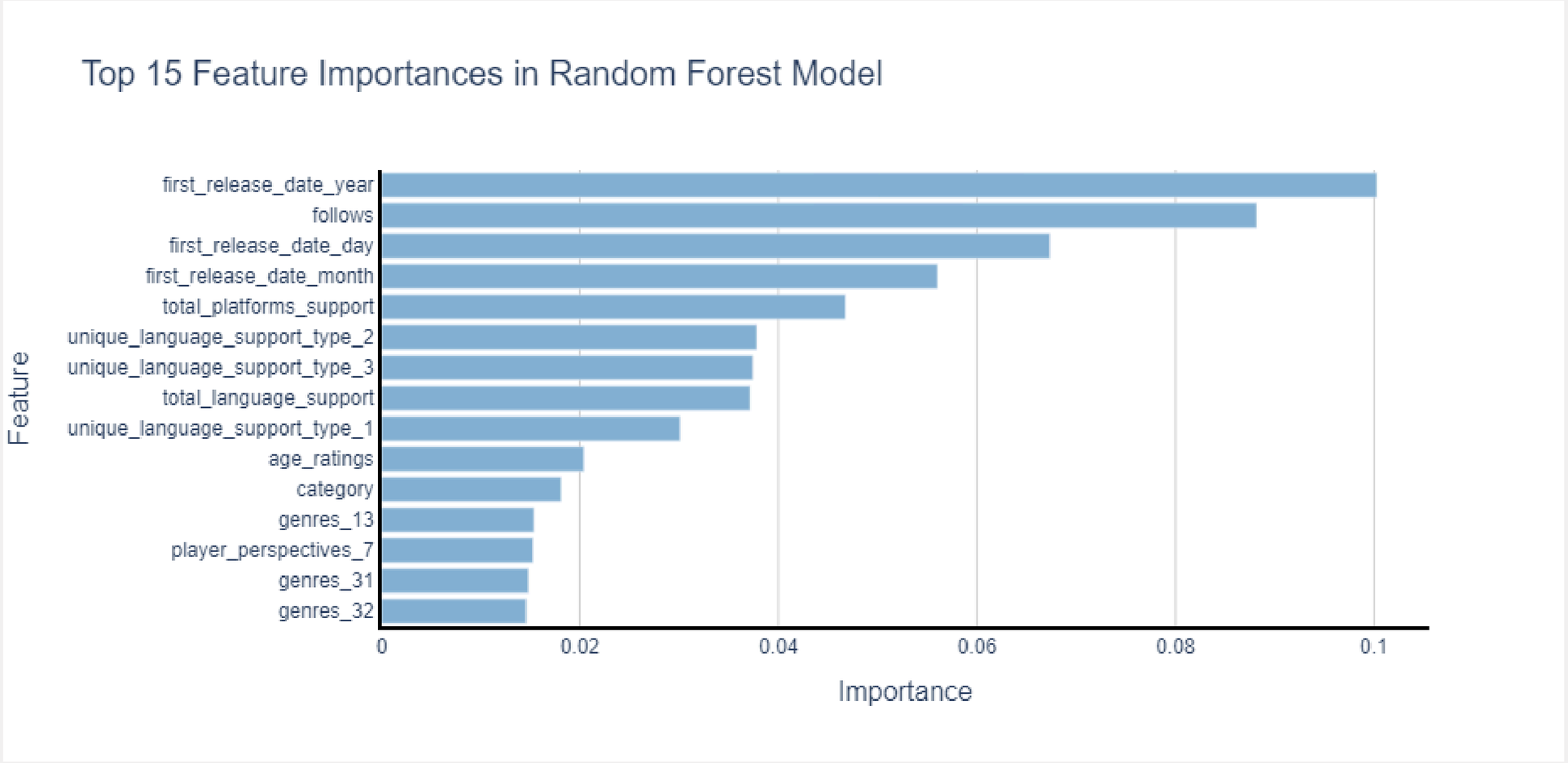
- *High performance and accuracy*
- *Feature importance for feature selection*

Disadvantages

- *Less interpretable*
- *Longer training time for large datasets*

The results of these methods will be shown in the following slides.

Feature Importance



The detail mapping of the feature names to actual features are shown in GitHub.

Results Summary

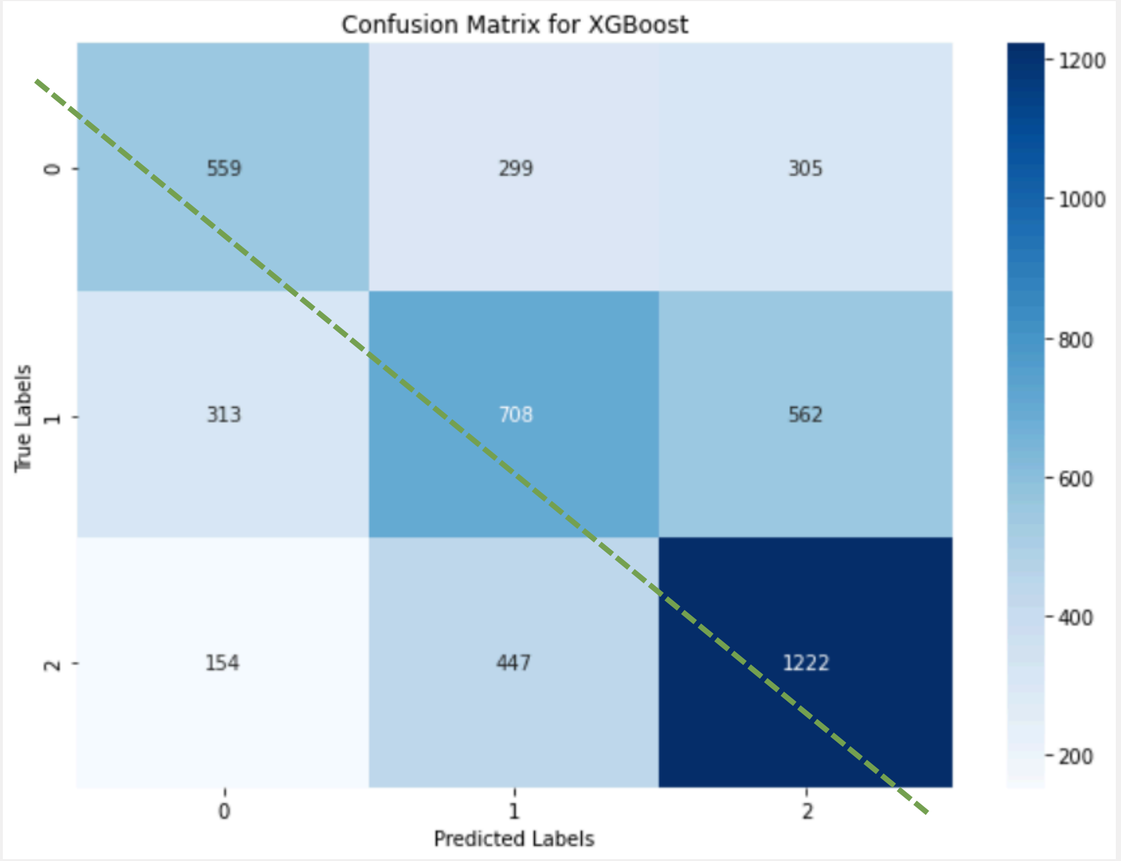
Three-class classification results.

Model	*Accuracy (%)*	*ROC AUC Score*	*F1 Score*
Random Forest	*53.86*	*0.6437*	*0.5334*
XGBoost	*54.48*	*0.6487*	*0.5408*

XGBoost surpass Random Forest.

Confusion Matrix

- *Strong ability to identify Class 3 (0-65) instances*
- *Class 2 (65-75) exhibits a balanced performance but also notable misclassifications into Classes 1 and 3.*
- *Class 1 (75-100) has the lowest number of correct predictions, indicating a potential area for model improvement.*



Thank You