**Discovering Hidden Patterns in Game Genres and Player Behavior**

**Team DataCore Members**

**Benjamin Grob -** Computer Science - [grobb@mail.gvsu.edu](mailto:grobb@mail.gvsu.edu)

**Christopher Khajira -** Computer Science - [khajirac@mail.gvsu.edu](mailto:khajirac@mail.gvsu.edu)

**David Kiboi -** Data Science - [kiboid@mail.gvsu.edu](mailto:kiboid@mail.gvsu.edu)

**Dorin Jerotich -** Data Science - [jeroticd@mail.gvsu.edu](mailto:jeroticd@mail.gvsu.edu)

**Overview of the Project**

We are examining the domain of Gaming and its trends on the popular digital storefront “STEAM”. Our team seeks to find various associations between games and their genres. To do this we will use clustering techniques to analyze what genres co-occur the most. By doing this we will be able to create visual aids that show what combinations of genres are popular among steam users and if certain combinations work better than others. If we are successful in this task we will be able to identify the most and least successful genres and market trends.

Our approach involves two primary methods. First, we will apply association rule mining to identify which genres and tags commonly co-occur, for instance, discovering that “Multiplayer” often appears alongside “Survival” and “Open World.” Second, we will use clustering to group games with similar behavioral fingerprints based on features such as playtime, price and ratings, allowing us to uncover natural groupings within the dataset.

Our motivation comes from the broader question of what defines engagement in digital gaming ecosystems. By identifying recurring genre and behavior patterns, our project can highlight how certain combinations succeed together and reveal the unique correlations that define player preferences and the market trends.

### **Related Work**

There have been a number of small projects conducted on the Steam marketplace. This comes as no surprise, as Steam is the most popular digital storefront for video games and had an estimated revenue of over $10 Billion in 2024. Annie Chien published the article “What 15 Years of Steam Data Tells Us About Gaming Industry Trends” that analyses what marketplace and genre trends succeed in today's gaming market. Our analysis will differ from Chiens’ slightly as we have a greater focus on the popularity of different combinations of genres rather than the monetary success of each genre on its own. Similarly Alden Yuan published an analysis that looked into the success of games across different genres. In Yuan’s analysis however, only English Windows games were included in the data set. Currently, our team doesn’t have a reason to limit our data to English windows games, so we will capture a larger percentage of the gaming industry.

Our project differs from previous analyses in a few key ways. First, we focus on pattern discovery rather than prediction by using association rule mining and clustering to uncover co-occurring genres and group structures, instead of measuring success through sales or revenue. Second, we include all available genres and tags, capturing the full diversity of Steam’s global and multi-platform market. Finally, we focus on the relationships between attributes such as how gameplay length, price, and player reviews interact across different genres to reveal meaningful trends in gaming behavior.

### **Data Plan**

We will use the Steam Store Games Dataset, which is publicly available on Kaggle at<https://www.kaggle.com/datasets/nikdavis/steam-store-games>. This dataset was compiled by Nik Davis using publicly accessible information scraped from Steam’s official store pages and metadata provided by SteamSpy’s API. It contains over 27,000 game entries, offering a comprehensive view of the gaming marketplace.

Each entry includes rich metadata such as the title, developer, publisher, genres, tags, price, release date, average and median playtime, number of reviews, and percentage of positive ratings. These features make it well-suited for data mining tasks, particularly those involving behavioral and categorical pattern discovery. The dataset was created for educational and analytical purposes.

### **Preprocessing Steps**

Before conducting our analysis, we will perform several preprocessing steps to ensure that the dataset is clean and suitable for data mining. The process will begin with data cleaning, where we will remove duplicate records, handle null values and correct any malformed or inconsistent entries. Next, during feature engineering, we will transform the genre and tag fields into binary indicator columns (for instance, “Action = 1” and “Strategy = 0”) to facilitate association rule mining. We will also normalize continuous variables such as price, playtime, and number of reviews to ensure that all features contribute proportionally during clustering analysis. Also, we will encode categorical variables to make them compatible with numerical analysis techniques.

### **Implementation plan**

Our project will implement a complete data mining pipeline that begins with data acquisition and preparation, where we will load the dataset from Kaggle and perform data cleaning and feature engineering using Pandas and NumPy. After preparing the data, we will conduct exploratory data analysis to generate descriptive statistics and create initial visualizations using Matplotlib and Seaborn to better understand the distribution and characteristics of the dataset. Next, we will perform association rule mining by applying algorithms to identify frequent genre and tag combinations among games. Following this, we will conduct clustering analysis to discover groups of similar games. This will involve normalizing numeric features such as playtime, price and ratings, then applying both K-Means and Hierarchical Clustering using Scikit-learn.

### **Evaluation plan**

We plan to evaluate our data mining algorithms through a combination of quantitative metrics and qualitative assessment adapted to the objective of each technique. For association rule mining, success rate will be based on standard metrics like support, confidence and lift, concentrating on rules with high lift values which indicate a meaningful relationship between items. In clustering, we will evaluate the quality of clusters with a silhouette score to measure both cohesion and separation. Since our project is oriented towards pattern discovery rather than prediction, we do not need a test set for the specific purpose of accuracy measurement.

We are going to confirm instead by comparing the results from different techniques, for example, if association rules suggest relationships then we can check whether the factors indeed appear together. Clustering results will as well be compared to certain other standard methods for instance using a hierarchical clustering algorithm.

### **Plan for group collaboration**

Our team will collaborate through a structured and organized workflow designed to maintain consistency and efficiency throughout the project. We will hold weekly in-person check-ins every Tuesday after class to review progress and address any challenges, supplemented by Zoom meetings for additional coordination as needed. To ensure proper version control and collaboration, all notebooks, code, and visualizations will be managed through our shared GitHub repository, allowing for smooth integration. Each team member will be responsible for specific components of the project: Benjamin will handle data preprocessing and cleaning, Christopher will focus on association rule mining and visualization, David will lead clustering and evaluation metric analysis, and Dorin will be responsible for report writing and presentation graphics..

### **Timeline**

* Week 6 - Team setup, dataset exploration, and cleaning
* Week 7 - Feature engineering and tag transformation
* Week 8 - Implement association rule mining
* Week 9 - Perform clustering analysis and validate results
* Week 10 - Conduct correlation analysis and finalize visualizations
* Week 11 - Integrate findings, polish report, and rehearse presentation
* Week 12 - Submit final report and GitHub repository

### **References**

1. Davis, Nik. Steam Store Games Dataset. Kaggle, 2024.  
   <https://www.kaggle.com/datasets/nikdavis/steam-store-games>
2. Chien, Annie. What 15 Years of Steam Data Tells Us About Gaming Industry Trends. 2024.  
   <https://anniechien.substack.com/p/what-15-years-of-steam-data-tells-88a>
3. Yuan, Alden. Exploratory Data Analysis on Steam Store Games Dataset. 2023.  
   <https://aldenyuan.medium.com/exploratory-data-analysis-on-steam-store-games-dataset-f4ff06da44eb>
4. Han, Jiawei, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques. 4th ed., Elsevier, 2022. <https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques>