Universidade Federal Fluminense

GET00126 - ANÁLISE MULTIVARIADA I

Análise de Agrupamento de Jogos da Steam

Thiago Senra

Julho 2025

Resumo

Este trabalho tem como objetivo aplicar técnicas de análise de agrupamentos para segmentar jogos da plataforma Steam com base em variáveis como gênero, avaliações, tempo de jogo, preço, popularidade e compatibilidade com sistemas operacionais. A base utilizada contém informações de 27.075 jogos lançados até o ano de 2019.

Após etapas de pré-processamento e padronização, foi adotado o método hierárquico aglomerativo com métrica de Gower, adequada para dados mistos. A análise de silhueta indicou a formação ideal de quatro clusters, que foram posteriormente caracterizados.

Os resultados revelaram grupos distintos de jogos, como títulos populares e bem distribuídos em termos de gênero e suporte (Cluster 1), jogos de ação com conteúdo intenso e compatibilidade restrita (Cluster 2), jogos estratégicos voltados a públicos específicos (Cluster 3), e jogos *indie* e *casual* com baixa popularidade, mas alto número de conquistas (Cluster 4).

A análise evidenciou o potencial das técnicas multivariadas na identificação de padrões em grandes bases de dados, com aplicações práticas em marketing, recomendação de conteúdo e estudos de mercado no setor de jogos digitais.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

1.1 Motivação

Com o crescimento exponencial da indústria de jogos digitais, plataformas como a Steam reúnem milhares de títulos com características variadas, como gênero, popularidade, tempo médio de jogo e avaliação dos usuários. Nesse contexto, identificar padrões e agrupar jogos com características semelhantes pode oferecer insights relevantes para áreas como marketing, recomendação personalizada e desenvolvimento de novos produtos.

1.2 Objetivos

O presente trabalho tem como objetivo realizar uma análise de agrupamentos com base em um conjunto de dados contendo informações sobre 27.075 jogos disponíveis na Steam. Por meio de técnicas de pré-processamento e da aplicação do método hierárquico com a distância de Gower, buscou-se formar grupos (clusters) de jogos que compartilham perfis similares. A análise permitiu não apenas uma compreensão mais clara da diversidade presente na plataforma, mas também a caracterização de diferentes perfis de jogos com base em variáveis como gênero, avaliações, tempo de jogo e suporte a plataformas.

1.3 Contextualização Acadêmica

Este trabalho foi desenvolvido no âmbito da disciplina GET00126 – Análise Multivariada I, cursada no semestre 2025/1, e oferecida pelo Departamento de Estatística da Universidade Federal Fluminense, sob a supervisão da professora Jéssica Kubrusly.

2 Materiais e Métodos

2.1 Base de Dados

A base de dados utilizada neste estudo foi extraída da plataforma Steam e contém informações de 27.075 jogos lançados até o ano de 2019. As variáveis disponíveis abrangem aspectos como nome, data de lançamento, desenvolvedora, preço, número de conquistas, avaliações positivas e negativas, tempo de jogo, gêneros, plataformas compatíveis, entre outros.

Os dados estão disponíveis publicamente em: `https://www.kaggle.com/datasets/nikdavis/steam-store-games`.

2.2 Pré-processamento dos Dados

Diversas etapas de preparação dos dados foram necessárias antes da aplicação do método de clusterização:

- Transformação da variável *release_date* para *release_year*;
- Criação da variável categórica *popularity*, a partir da variável *owners*, com cinco níveis: Muito Baixa, Baixa, Média, Alta e Muito Alta;
- Seleção dos gêneros com frequência maior ou igual a 150 e transformação das variáveis *genres* e *platforms* em dummies;
- Padronização das variáveis numéricas;
- Seleção final das variáveis: gêneros (16 categorias), plataformas (Linux, Mac), popularidade, inglês (indicador), conquistas, ano de lançamento, avaliações positivas e negativas, tempo médio e mediano de jogo, e preço.

2.3 Método de Agrupamento

A análise de agrupamentos foi realizada com base em variáveis numéricas e categóricas simultaneamente. Para isso, foi utilizada a função `daisy()` com a métrica de Gower, que permite calcular dissimilaridades entre dados mistos, considerando automaticamente escalas, variáveis binárias e fatoriais.

O método de aglomeração utilizado foi o hierárquico aglomerativo, com ligação *Ward.D2*, que visa minimizar a variância intra-grupos, promovendo a formação de clusters compactos e homogêneos.

2.3.1 Determinação do Número de Clusters

Para definir o número ideal de agrupamentos, foi utilizado o dendrograma resultante da aplicação do método hierárquico. A inspeção visual do dendrograma indicou um corte natural na altura que separa os dados em quatro grupos bem definidos, com boa separação entre as ramificações principais.

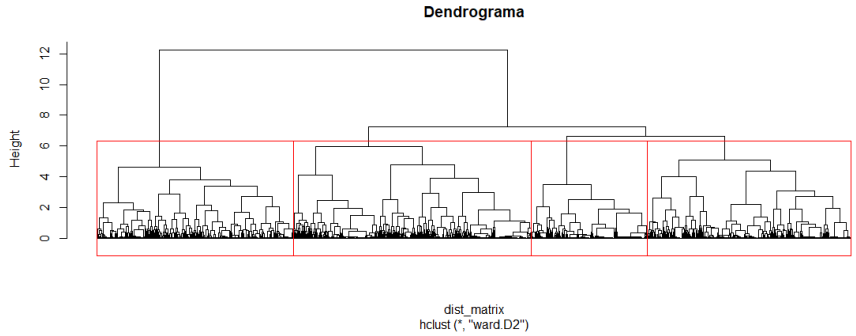

Figura 1: Dendrograma gerado pelo método hierárquico com ligação *Ward.D2*.

Além disso, foi calculado o gráfico de silhueta para diferentes valores de k (número de clusters). O valor médio da silhueta foi maximizado em $k = 2$, mas por conta da baixa representatividade, optamos por seguir com a maximização em $k = 4$, pois esta quantidade de grupos proporciona a melhor combinação entre coesão interna e separação entre grupos.

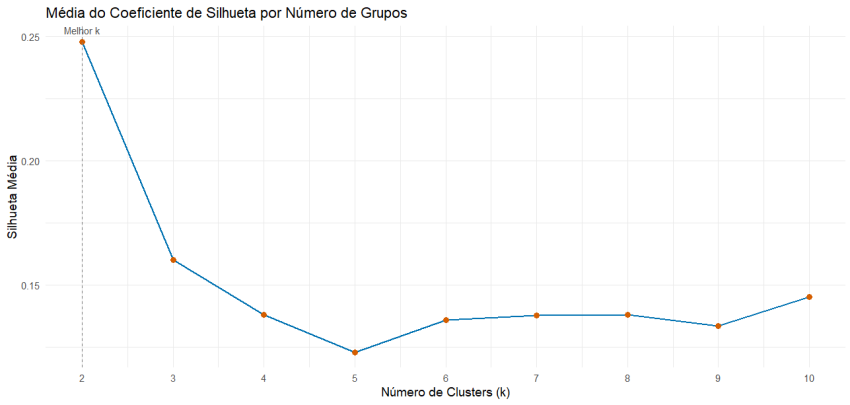

Figura 2: Gráfico de silhueta para diferentes valores de k..

Com base nesses critérios, optou-se pela segmentação da base de dados em quatro clusters, posteriormente analisados e caracterizados segundo seus perfis médios.

3 Análise Descritiva

Antes da aplicação da técnica de clusterização, foi realizada uma análise descritiva das variáveis selecionadas com o objetivo de compreender a distribuição dos jogos em relação a características fundamentais.

3.1 Gêneros dos Jogos

Os gêneros mais frequentes na base de dados foram *Indie*, *Action*, *Casual* e *Adventure*, refletindo a diversidade e popularidade desses estilos na plataforma Steam. Jogos classificados como *Strategy*, *Simulation* e *RPG* também apresentaram representatividade significativa.

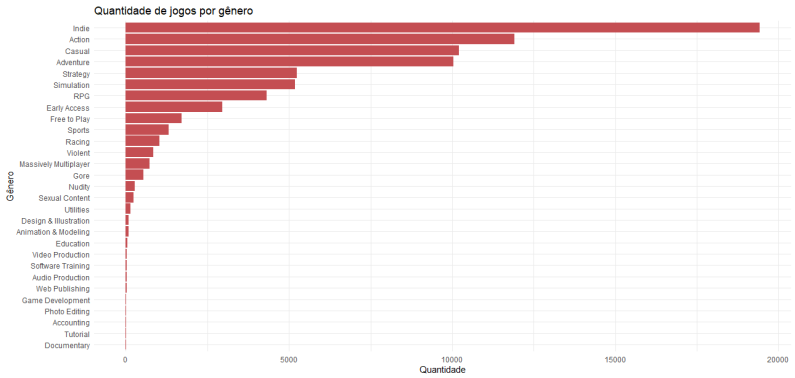

Figura 3: Distribuição dos gêneros dos jogos.

3.2 Compatibilidade com Plataformas

A maioria dos jogos oferece suporte para o sistema operacional Windows. Em menor escala, observou-se suporte para as plataformas Mac e Linux. A distribuição desigual reflete a predominância do Windows como ambiente de jogos digitais.

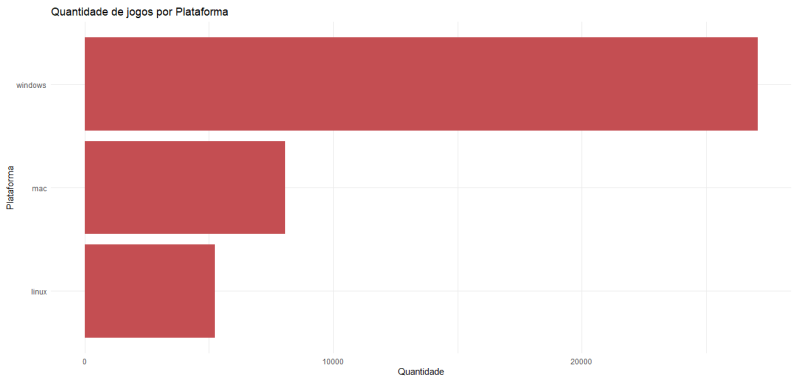

Figura 4: Distribuição das plataformas dos jogos.

3.3 Ano de Lançamento

A distribuição dos lançamentos mostra crescimento expressivo ao longo do tempo, especialmente a partir de 2014, indicando o aumento da acessibilidade e da produção de jogos independentes nos últimos anos do período analisado.

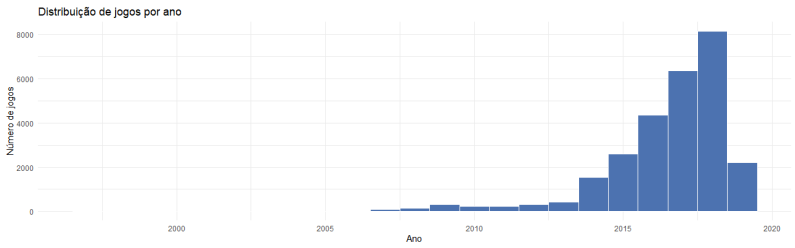

Figura 5: Distribuição dos anos de lançamento dos jogos.

3.4 Popularidade

A variável *popularity*, criada a partir da estimativa de donos (*owners*), mostra que a maior parte dos jogos se concentra nas categorias de popularidade Muito Baixa e Baixa. Apenas uma pequena fração dos títulos alcança níveis de popularidade considerados Muito Alta, o que sugere uma concentração de títulos com alcance limitado na plataforma.

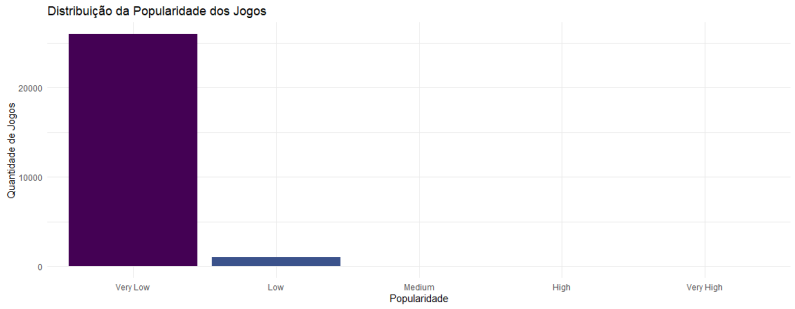

Figura 6: Distribuição da popularidade dos jogos.

3.5 Tempo de Jogo

O tempo de jogo foi analisado por meio de duas métricas: o tempo médio e o tempo mediano jogado por usuário. A maior parte dos jogos apresenta tempos reduzidos, indicando um padrão comum de jogos rápidos ou com pouco engajamento prolongado por parte dos jogadores.

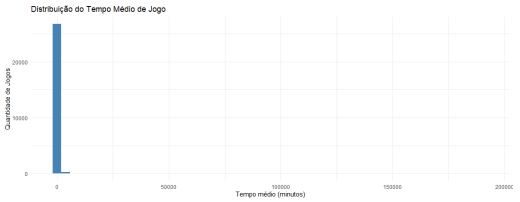

Figura 7: Distribuição do tempo médio de jogo.

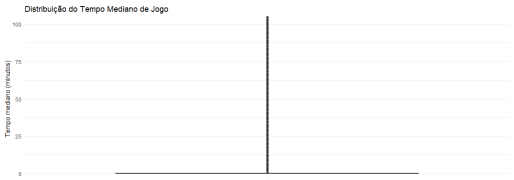

Figura 8: Distribuição do tempo mediano de jogo.

3.6 Avaliações

As avaliações dos jogos são divididas em positivas e negativas. Nota-se uma grande assimetria: enquanto alguns jogos acumulam dezenas de milhares de avaliações, a maioria apresenta baixa visibilidade ou engajamento dos usuários.

Figura 9: Distribuição das avaliações positivas dos jogos

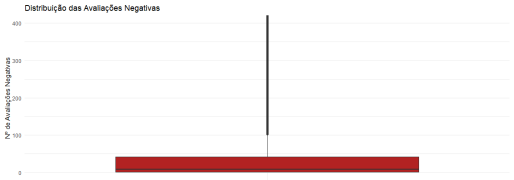

Figura 10: Distribuição das avaliações negativas dos jogos.

3.7 Preço

A análise dos preços revelou que a maioria dos jogos possui valores acessíveis ou são gratuitos. Títulos com preços muito elevados são raros e possivelmente refletem edições especiais ou conteúdo adicional incluso.

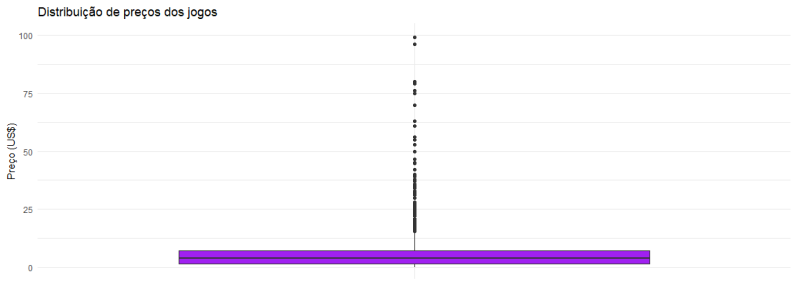

Figura 11: Distribuição dos preços dos jogos.

4 Resultados

Nesta seção são apresentados os resultados obtidos com a aplicação do método de agrupamento hierárquico. As análises foram divididas em duas partes: variáveis numéricas e variáveis categóricas. Em seguida, é feita a caracterização dos clusters com base na combinação desses resultados.

4.1 Análise das Variáveis Numéricas

Os boxplots a seguir comparam a distribuição das variáveis quantitativas entre os quatro clusters formados. As variáveis selecionadas foram aquelas que apresentaram maior poder de discriminação entre os grupos, como número de conquistas, tempo médio e mediano de jogo, avaliações positivas e negativas, preço e ano de lançamento.

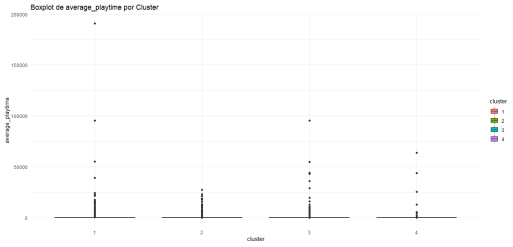

Figura 12: Distribuição do tempo médio de jogo por cluster.

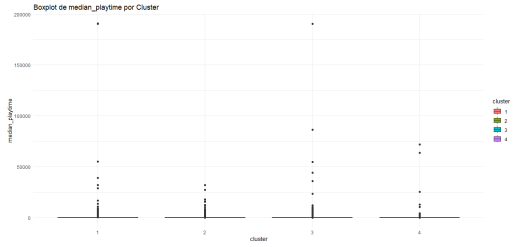

Figura 13: Distribuição da mediana de tempo de jogo por cluster.

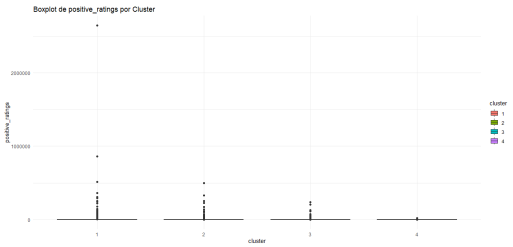

Figura 14: Distribuição das avaliações positivas por cluster.

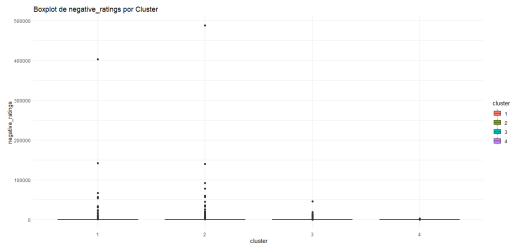

Figura 15: Distribuição das avaliações negativas por cluster.

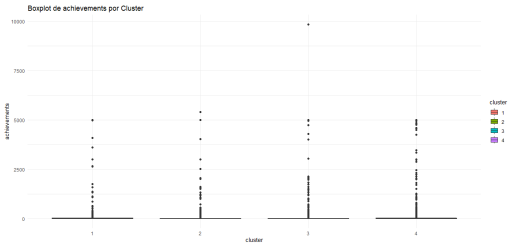

Figura 16: Distribuição do número de conquistas por cluster.

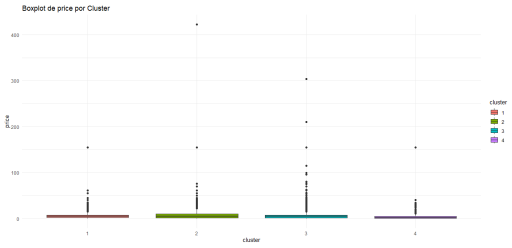

Figura 17: Distribuição do preço de jogo por cluster.

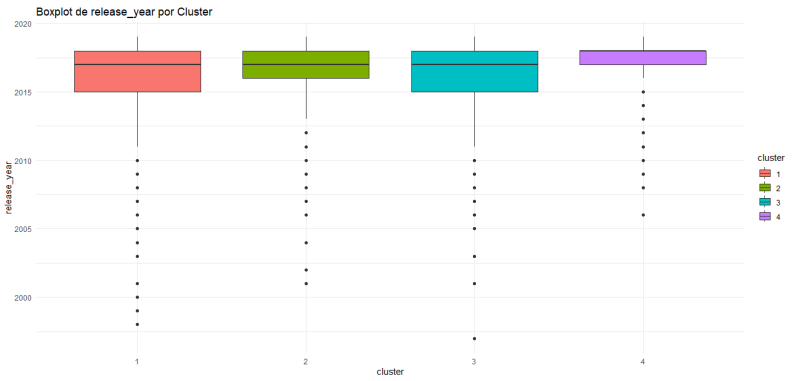

Figura 18: Distribuição do ano de lançamento por cluster.

- **Conquistas:** O Cluster 4 apresenta o maior número médio de conquistas por jogo, superando significativamente os demais grupos. Já o Cluster 1 também possui valor relativamente alto, enquanto o Cluster 2 apresenta a menor média nesse aspecto.

- **Tempo médio e mediano de jogo:** O Cluster 3 se destaca por apresentar o maior tempo médio e mediano de jogo, o que sugere títulos mais longos e com maior engajamento. O Cluster 4, por outro lado, concentra jogos de curta duração.

- **Avaliações positivas e negativas:** O Cluster 1 possui os maiores valores médios de avaliações positivas, seguido pelo Cluster 2. O Cluster 2, entretanto, apresenta também a maior média de avaliações negativas. O Cluster 4 tem os menores valores em ambas as métricas.

- **Preço:** Os preços são relativamente baixos em todos os clusters, mas o Cluster 1 apresenta os maiores valores médios. O Cluster 4 concentra a maioria dos jogos gratuitos ou de baixo custo.

4.2 Análise das Variáveis Categóricas

As figuras a seguir apresentam a proporção de categorias como gêneros, plataformas compatíveis e faixas de popularidade entre os clusters. Essas variáveis ajudam a entender o perfil dos jogos de cada grupo além dos valores numéricos.

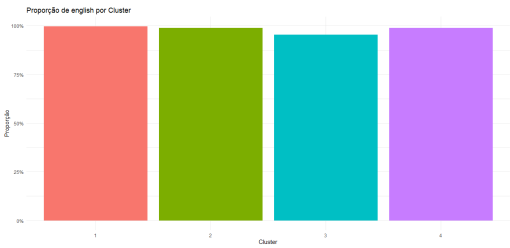

Figura 19: Proporção de idioma inglês por cluster.

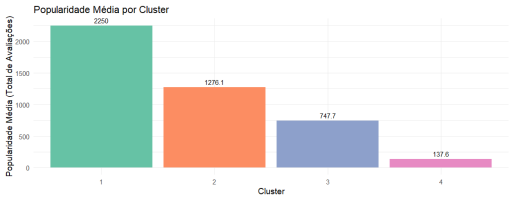

Figura 20: popularidade média dos jogos por cluster.

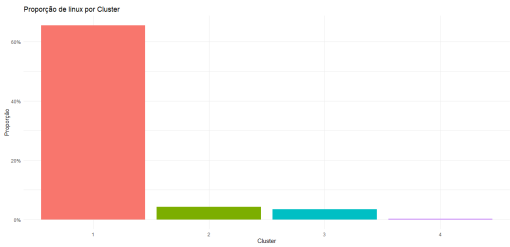

Figura 21: Proporção de suporte linux por cluster.

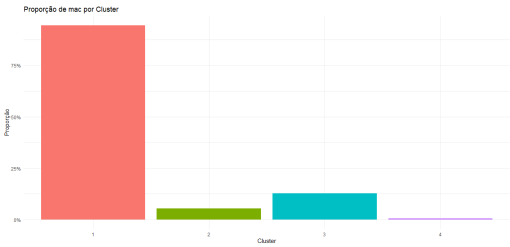

Figura 22: Proporção de suporte mac por cluster.

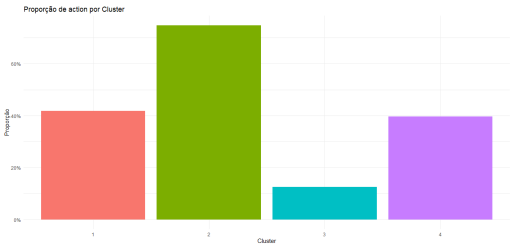

Figura 23: proporção de gênero ação por cluster.

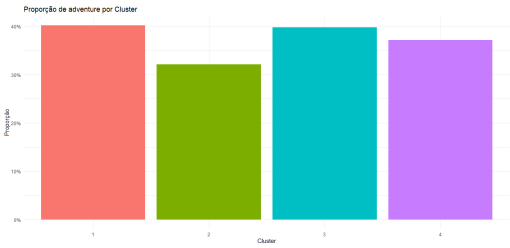

Figura 24: proporção de gênero aventura por cluster

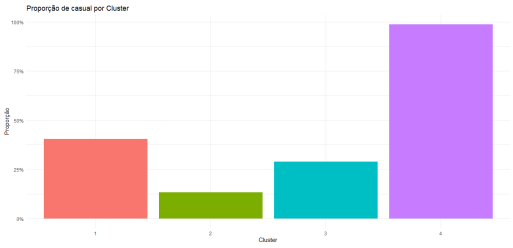

Figura 25: proporção de gênero casual por cluster.

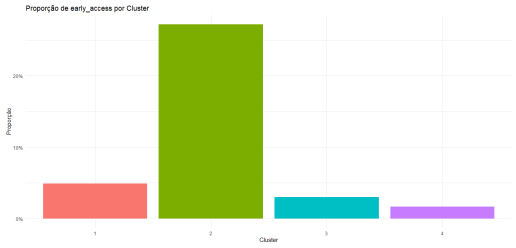

Figura 26: proporção de gênero acesso antecipado por cluster

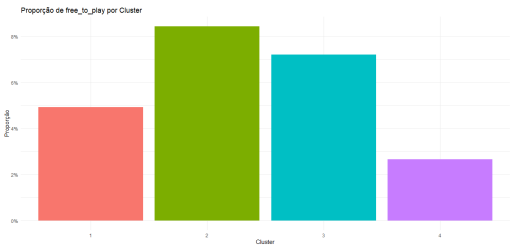

Figura 27: proporção de gênero free to play por cluster.

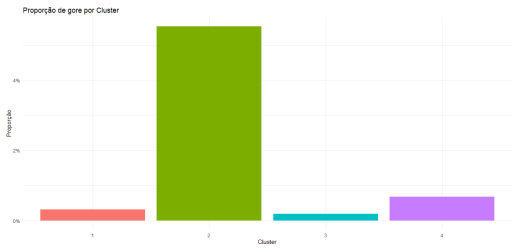

Figura 28: proporção de gênero gore por cluster

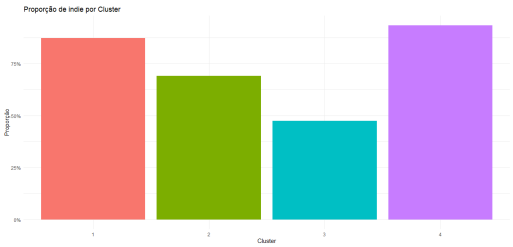

Figura 29: proporção de gênero indie por cluster.

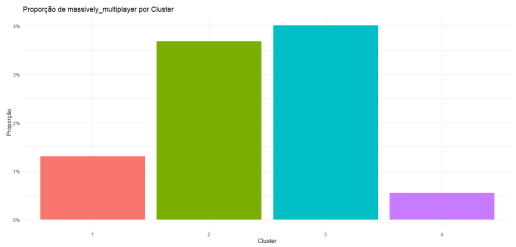

Figura 30: proporção de gênero multiplayer massivo por cluster

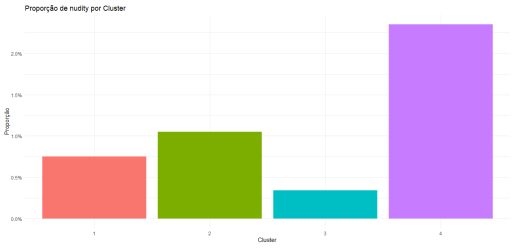

Figura 31: proporção de gênero nudez por cluster.

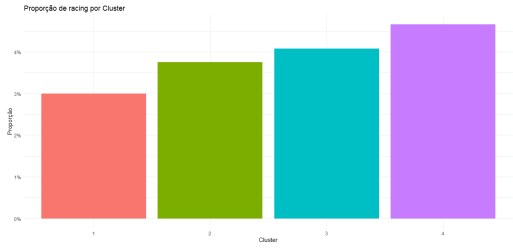

Figura 32: proporção de gênero corrida por cluster

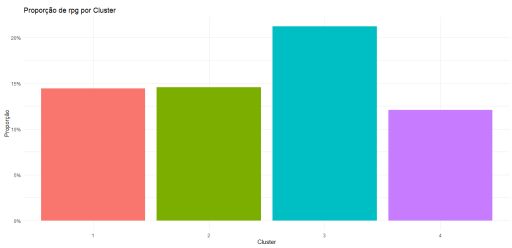

Figura 33: proporção de gênero rpg por cluster.

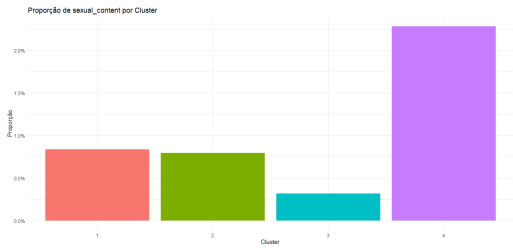

Figura 34: proporção de gênero conteúdo sexual por cluster

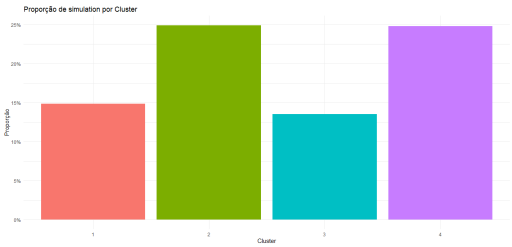

Figura 35: proporção de gênero simulação por cluster.

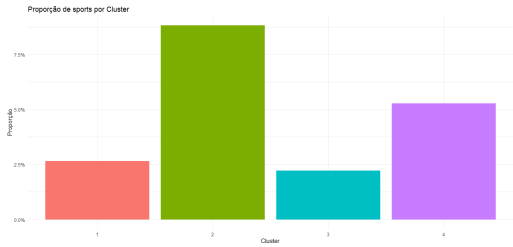

Figura 36: proporção de gênero esportes por cluster

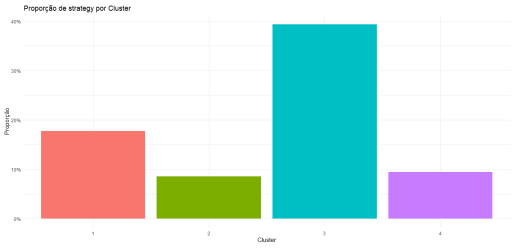

Figura 37: proporção de gênero estratégia por cluster.

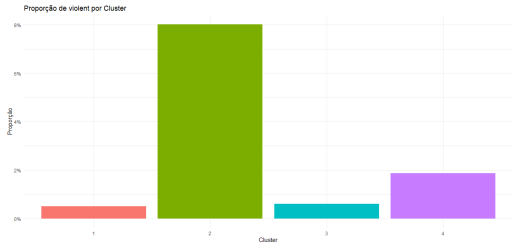

Figura 38: proporção de gênero violência por cluster

- **Gêneros:** O Cluster 1 apresentou uma distribuição mais equilibrada dos gêneros, enquanto o Cluster 2 concentra a maior parte dos jogos de ação (*action*). O Cluster 3 se destaca em jogos de estratégia e aventura, enquanto o Cluster 4 é quase inteiramente composto por jogos *casual* e *indie*, refletindo títulos mais simples e acessíveis.
- **Popularidade:** A maioria dos jogos nos Clusters 3 e 4 apresenta popularidade Muito Baixa ou Baixa. O Cluster 1 concentra os jogos mais populares, seguido pelo Cluster 2 com popularidade moderada.
- **Compatibilidade com Plataformas:** O Cluster 1 apresenta amplo suporte para Mac e Linux, diferentemente dos Clusters 2, 3 e especialmente o 4, que praticamente não oferecem compatibilidade além do Windows.

4.3 Caracterização dos Clusters

Com base nas análises numéricas e categóricas apresentadas, é possível descrever os principais perfis de cada grupo identificado:

Tabela 1: Resumo das características principais dos clusters.

Cluster	Descrição
1	Jogos populares, com forte presença de títulos *indie*, distribuição equilibrada entre os gêneros *action*, *adventure* e *casual*, e alto suporte a plataformas como Mac e Linux. Apresenta valores medianos relativamente altos em avaliações e tempo de jogo.
2	Grupo com maior concentração de jogos de ação, conteúdo intenso (ex: *gore*, *violent*), baixa compatibilidade com Mac/Linux e popularidade intermediária. É o grupo com maior mediana de avaliações negativas.
3	Dominado por jogos estratégicos e de aventura, com segundo maior tempo médio de jogo e boas médias de conquistas. Público mais nichado e engajado.
4	Predominância de jogos *casual* e *indie*, com baixíssima popularidade e compatibilidade, mas com o maior número de conquistas por jogo, indicando um design voltado a metas e desafios.

5 Conclusão

Este estudo teve como objetivo aplicar técnicas de análise de agrupamentos para segmentar jogos da plataforma Steam com base em características diversas, incluindo avaliações, tempo de jogo, gêneros, preço, compatibilidade com plataformas e popularidade. Após o pré-processamento dos dados e a padronização das variáveis, foi utilizado o método hierárquico aglomerativo com distância de Gower, apropriado para dados mistos, resultando na formação de quatro clusters bem definidos.

A análise descritiva dos grupos revelou padrões interessantes. O Cluster 1 representa jogos mais populares, com bom suporte a múltiplas plataformas e diversidade de gêneros. O Cluster 2 concentra jogos de ação com maior incidência de conteúdo intenso, enquanto o Cluster 3 agrupa jogos estratégicos e de aventura voltados a um público mais nichado. Já o Cluster 4 é formado majoritariamente por jogos *casual* e *indie*, com baixa visibilidade, mas alto número médio de conquistas.

Os resultados obtidos demonstram o potencial da análise de cluster para identificar perfis distintos de produtos em grandes bases de dados. No contexto da indústria de jogos, tais segmentações podem ser aplicadas em estratégias de marketing, sistemas de recomendação, curadoria de conteúdo e análise de mercado.

Como trabalhos futuros, seria possível aplicar outros métodos de agrupamento (como K-prototypes), realizar uma análise mais aprofundada com apenas jogos criados por grandes empresas, ou ainda expandir a base de dados com variáveis adicionais relacionadas ao comportamento dos usuários.