Mapeamento do Futuro: Analisando as Tendências Globais dos Jogos

**INTRODUÇÃO**

Desde os primórdios dos videogames, quando simples pontos luminosos em telas monocromáticas simulavam jogos de tênis, a indústria de games passou por uma evolução vertiginosa. Dos *arcades* dos anos 80, com seus clássicos inesquecíveis, até os consoles de última geração e a era dos jogos online, a jornada tem sido marcada por inovações tecnológicas e mudanças nos hábitos dos jogadores.

Com o avanço da tecnologia, os jogos se tornaram cada vez mais complexos e imersivos, envolvendo milhões de jogadores em todo o mundo. Paralelamente a esse crescimento, as empresas do setor perceberam a necessidade de compreender melhor seu público e as tendências do mercado. A coleta e análise de dados se tornaram essenciais para tomar decisões estratégicas, desenvolver novos produtos e personalizar experiências.

Neste contexto, o presente estudo propõe utilizar a metodologia *CRISP-DM* (*Cross-Industry Standard Process for Data Mining*) para mapear o futuro dos jogos, analisando as tendências globais do setor. Através de uma abordagem estruturada e sistemática, buscamos identificar padrões, prever comportamentos e gerar insights valiosos para empresas, desenvolvedores e jogadores.

**CRISP-DM**

**Entendimento do negócio:**

Analisar e seguir tendencias tem sido parte do cotidiano corporativo a muitos anos, entender como o negócio está girando e como se aproveitar de uma oportunidade são cruciais nas competitivas disputas. Para a indústria de *games* não tem sido diferente, entender seu público, o que agrada e o que chateia é de extrema importância, para isto, buscamos entender as tendencias do mercado de jogos para possíveis insights para investimentos futuros de jogos.

Buscaremos entender através dos dados, como as relações de venda são afetadas ou por gênero, por distribuidora, ano de lançamento ou regiões de maior sucesso.

**Entendimento dos Dados:**   
Para entendermos esse mercado, vamos primeiramente analisar nosso *dataset* para entendermos as possíveis variáveis relevantes e explorar os dados para identificar as métricas de sucesso que serão utilizadas para avaliar os resultados.

**Explorando o *dataset*:**

* **Ferramentas:**

Utilizaremos a biblioteca *pandas* em *python* para a visualização dos tipos da coluna, aliada à bliblioteca *Python*-*Docx* para gerar as tabelas.

O dicionário de dados irá nos ajudar a entender as variáveis disponíveis e seus tipos, nos dando um panorama das métricas que poderemos utilizar para explorar os sucessos globais.  
No dataset *vgsales.csv,* o dicionário de dados possui:

**Dicionário:**

Utilizamos a biblioteca pandas em python para a visualização dos tipos da coluna aliada à bliblioteca Python-Docx para gerar as tabelas.

|  |  |  |
| --- | --- | --- |
| Variável | Tipo | Descrição |
| Rank | int64 | Posição do jogo em um ranking geral. |
| Name | object | Nome do jogo. |
| Platform | object | Plataforma em que o jogo foi lançado (neste caso, 2600). |
| Year | int64 | Ano de lançamento. |
| Genre | object | Gênero do jogo. |
| Publisher | object | Editora do jogo. |
| NA\_Sales | float64 | Vendas na América do Norte. |
| EU\_Sales | float64 | Vendas na Europa. |
| JP\_Sales | float64 | Vendas no Japão. |
| Other\_Sales | float64 | Vendas em outras regiões. |
| Global\_Sales | float64 | Vendas globais |

A variável Rank, mostra a posição do jogo em relação à sua venda global, que por sua vez, tem seu valor como a soma de todas as vendas em cada região registrada.

As variáveis Plataform, Genre e Publisher podem nos esclarecer em como essas variáveis podem influenciar na venda, seja pelo gênero do momento, ou a plataforma mais atualizada.

**Prepração de dados:**

Agora vamos utilizar a biblioteca pandas para a identificação de nulos por coluna

**Valores Nulos por Coluna:**

|  |  |
| --- | --- |
| Coluna | Quantidade de Nulos |
| Rank | 0 |
| Name | 0 |
| Platform | 0 |
| Year | 0 |
| Genre | 0 |
| Publisher | 36 |
| NA\_Sales | 0 |
| EU\_Sales | 0 |
| JP\_Sales | 0 |
| Other\_Sales | 0 |
| Global\_Sales | 0 |

Nos deparamos com 36 valores nulos em nosso dataframe, quando checamos manualmente os dados, podemos ver que as distribuidoras estão devidamente registradas, mas ainda sim segue como nullo, neste caso, utilizaremos a biblioteca pandas para excluir as linhas que possuem valores nulos para que nosso dataframe fique o mais ‘saudavel’ possível.

Na verificação de outliers no nosso dataset, foram avaliados através da técnica z-score alguns outliers, em busca das explicações do mesmo, utilizamos o pandas para copiar os dados onde o z-score ultrapassa o valor 3 e salvamos em um dataset diferente, onde avaliamos que os outliers representam fenômenos reais, sendo assim, decidimos não remover os outliers para obter uma melhor precisão no nosso modelo.

**Modelagem:**

Para a clusterização do nosso dataset, precisamos entender o numero de clusters que serão utilizados no nosso projeto, para isso, utilizamos o método do cotovelo, onde podemos ter uma inferência na quantidade de de cluster. O método foi aplicado através de um algoritmo em python, tendo nos retornado o uso de 2 cluesters.

Após isso, aplicamos o algoritmo k-means ao dataframe limpo, chamado vgsales\_limpo.csv, onde nos retornou um novo csv ‘vgsales\_com\_clusters.csv’ onde podemos identificar as duas categorias criadas, 0 e 1.