Universidade de Coimbra

Departamento de Engenharia Informática

UNIVERSIDADE D
COIMBRA

Visualização de Dados

Projeto 2 - Visualização de um dataset referente a vendas de videojogos

Mafalda Duarte e Rodrigo Santos

2023/2024

I - Autores

- Mafalda Duarte - 2021236492 - uc2021236492@student.uc.pt
- Rodrigo Santos - 2021236556 - uc2021236556 @student.uc.pt

II - Introdução

Este projeto tem como objetivo analisar e extrair informação de um dataset que contém dados sobre a venda de videojogos em diferentes regiões. Este conjunto tem 11 variáveis que permitem obter uma ampla visão do desempenho de diferentes jogos ao longo do tempo. O nosso foco é compreender padrões/tendências, através da visualização dos dados, na indústria dos videojogos de maneira a sermos capazes de concluir quais variáveis podem influenciar o sucesso de um jogo numa determinada região.

Esta análise pode ser benéfica para profissionais da indústria pois pode auxiliar na tomada de decisão sobre plataformas, regiões e géneros de jogos para futuros lançamentos. Perceber as preferências específicas de cada região poderá ser uma forma de criar melhores estratégias de vendas e melhorar a comercialização de jogos. Não só os profissionais diretamente ligados ao desenvolvimento de jogos e da sua comercialização podem beneficiar do nosso estudo, criadores de conteúdo em plataformas como Youtube e Twitch, podem entender quais géneros de jogos interessam ao público da sua região e assim conseguir aumentar os seus visualizadores e consequentemente o seu lucro.

De maneira geral tentámos responder a diversas perguntas como qual o país mais consumista relativamente a esta indústria, qual ou quais géneros têm mais sucesso de forma geral e em cada região especificamente, quais os anos de maior sucesso da indústria, qual a plataforma com mais jogos e quais os géneros e muitas outras que podem ser úteis para planear um lançamento de um jogo.

III - Trabalhos Relacionados

Neste projeto tentámos diversificar os gráficos utilizados. Além das bases que já tínhamos do projeto 1, fomos também à procura de outros trabalhos relacionados com o nosso para termos uma perspetiva mais ampla daquilo que poderíamos fazer de maneira a enriquecer o nosso projeto.

Para a realização do nosso tree map (Fig.7), tivemos como inspiração aquele criado no notebook da referência **[1]** e decidimos alterar as cores para uma melhor visualização dos géneros de jogos lançados para cada plataforma. Quanto ao gráfico animado (Fig.4), tirámos inspiração de **[6],** mudámos o intervalo de anos para [1980, 2020] e os publishers. Fomos também ao projeto 1 de ambos, para tentar encontrar mais visualizações e daí surgiu o nosso bubble chart (Fig.11), que consegue mostrar muita informação, sendo no nosso caso os 50 jogos com mais vendas globais por ano, em que o tamanho das bolhas está relacionado com o número de plataformas para o qual aquele jogo foi lançado e a cor por género. Por último, utilizámos

como referência o heatmap de **[5]**, visto que é uma abordagem diferente do heatmap que fizemos com plotly express (Fig.8).

IV - Dados

O dataset utilizado foi retirado do Kaggle e contém uma lista com videojogos que tenham vendido mais de 100 000 cópias. O conjunto de dados inicial era constituído por 16598 linhas, cada uma correspondente a um jogo diferente, e 11 variáveis que nos dão diferentes informações desde o rank do jogo em termos de vendas globais, até a detalhes como o nome (Name), plataforma de lançamento (Platform), ano em que foi lançado (Year), género (Genre), editora (Publisher), número de vendas em regiões específicas, América do Norte (NA_Sales), Europa (EU_Sales), Japão (JP_Sales) e outras (Other_Sales) e número de vendas globais (Global_Sales). Vamos trabalhar com 12 géneros, 31 plataformas e 583 editoras diferentes.

Das 11 variáveis 4 são categóricas e 7 são numéricas. O conjunto de dados não tem duplicados, no entanto contém 307 linhas com valores em falta, ou seja, 1.85% do número total de linhas. Destes, 271 valores estão em falta na coluna 'Year' e os restantes 58 na coluna 'Publisher'.

Como a percentagem de linhas com valores em falta está perto dos 2%, decidimos substituir alguns de maneira a diminuir essa percentagem. Fomos a algumas das linhas com valores em falta e fomos, de acordo com pesquisa, completar alguns dos valores até a percentagem ficar abaixo de 1%. Substituímos 145 valores da coluna 'Year' e 11 da coluna 'Publisher' e retirámos as restantes linhas que continham valores em falta. No final ficámos com um conjunto de dados com 16433 linhas. Por fim notámos que os dados da coluna 'Year' estavam em float, mas como se trata de uma variável referente aos anos que são sempre números inteiros decidimos mudar o tipo dessa coluna para int.

V - Design

Durante a criação das nossas visualizações de dados tivemos em atenção os princípios de design de maneira a assegurarmos que a nossa informação era mostrada de forma simples e clara. Criámos para o nosso projeto um total de 12 gráficos (Figuras 1 a 12).

Utilizámos legendas simples e diretas em todas as representações e procurámos escolher gráficos que permitissem uma interpretação intuitiva. Tentámos utilizar sempre as mesmas cores para representar as mesmas variáveis e tivemos por base, em grande parte, paletas de cores já existentes fornecidas pela biblioteca plotly. Por exemplo, para a representação de cada género utilizamos a paleta 'Set3' que podemos observar na Fig.13, procuramos também utilizar cores fáceis de diferenciar entre si. As visualizações apresentadas permitem em grande parte identificar padrões gerais de maneira a beneficiar um maior número de pessoas. Tivemos o cuidado de tratar os dados inicialmente para garantir que a informação mostrada não apresenta falhas.

VI - Implementação

Começámos por importar as bibliotecas que iriamos utilizar, sendo estas: pandas, numpy, plotly (graph_objects, express e subplots) e seaborn. De uma maneira geral, implementámos métodos idênticos para todas as visualizações. Manipulámos o dataframe com os dados que pretendíamos utilizar com ".groupby" e assim criámos um novo dataframe para praticamente todas as nossas visualizações. De seguida, para as visualizações em que foi necessário, recorremos ao comando sort para obtermos o top dos melhores valores e apenas nos focarmos neste top. Posteriormente, criámos a função com parâmetro "df" e, com a ajuda das bibliotecas anteriormente importadas, obtivemos as visualizações que pretendíamos e alterámos a sua aparência, tal como a cor de fundo dos plots e o tipo e a cor da letra tanto para os títulos como para as legendas dos eixos. Por último, chamámos a função e substituímos o parâmetro pelo dataframe que pretendíamos utilizar, que é sempre aquele que criámos anteriormente na mesma célula da função.

VII - Reflexão

Conseguimos implementar um conjunto de visualizações que de forma simples conseguiram responder a várias questões, não só as mencionadas na introdução. Conseguimos mostrar que a América do Norte é a região com mais gastos nesta indústria (Fig.1.), podemos ver que os géneros 'Action', 'Sports', 'Misc' e 'Shooter' são os favoritos dessa região (Fig.2.). As vendas dos géneros 'Action' e 'Sports' são mais elevadas globalmente, no entanto no Japão são os jogos de 'Role-Playing' que lideram vendas como podemos concluir através da visualização 2 (Fig.2.). Podemos observar na Fig.8. com mais clareza que a predominância do género 'Action' se tornou mais evidente por volta de 1995 e que géneros como 'Strategy' e 'Adventure' não são os mais comprados.

Para uma visão geral de alguma da informação que acabámos de referir podemos analisar a Fig.10 que mostra com mais detalhe o número de vendas do género em cada região em separado. Aqui fica ainda mais evidente a supremacia da América do Norte na compra de jogos que apenas perde na categoria de 'Role-Playing' para o Japão.

Para obtermos um conjunto de informações mais gerais fizemos representações de quase todas as nossas variáveis, conseguimos mostrar quais as editoras melhores sucedidas Fig.3. e a forma como as suas vendas foram mudando de ano para ano Fig.4. (este é um plot animado mas, como não é possível colocar no relatório em movimento, criámos a Fig.5. que mostra exatamente a mesma informação mas de forma estática). Achámos pertinente utilizar o gráfico animado pois é possível ir ao ano que queremos analisar e visualizar as informações de forma mais clara.

Gráficos como a Fig.6. e Fig.11. contêm bastante informação. O objetivo da Fig.6. é entendermos se o género mais publicado é aquele com maior número de vendas e se o jogo mais vendido tinha o mesmo género que algum dos anteriores. Esta não foi a nossa melhor visualização pois, para conseguirmos legendar na figura o que era cada barra decidimos colocar padrões, mas estes não são bem visíveis nas barras de menor dimensão assim como as cores. De maneira geral para as barras maiores funciona e conseguimos fazer a análise. Já na Fig.11, conseguimos fazer uma visualização que mostra quatro variáveis de maneira bastante

perceptível.

Existe ainda muita informação que podemos retirar de todos os gráficos sendo a referida apenas alguma.

Com base no projeto 1, tentámos também adaptar o chloropleth para os nossos dados mas chegámos à conclusão que não se enquadra bem com os mesmos e não iria proporcionar uma visualização útil.

Em geral fomos capazes de criar visualizações que cumprem o seu propósito e seguir os princípios aprendidos. Não só conseguimos obter perceções gerais de regiões, géneros e plataformas como também analisámos em particular as vendas dos top 50 jogos com mais vendas globais e da distribuição das suas vendas por regiões (Fig.12.).

Em suma, mantivemos a consistência em elementos visuais e conseguimos apresentar gráficos esteticamente agradáveis, conseguimos implementar um gráfico interativo ou seja cuja informação que mostra é personalizável, tivemos em atenção cores de títulos, tipos de letra e fundos dos plots. Por fim, e o mais importante, conseguimos responder a todas as questões úteis em relação à indústria como já tínhamos mencionado anteriormente.

VIII - Figuras

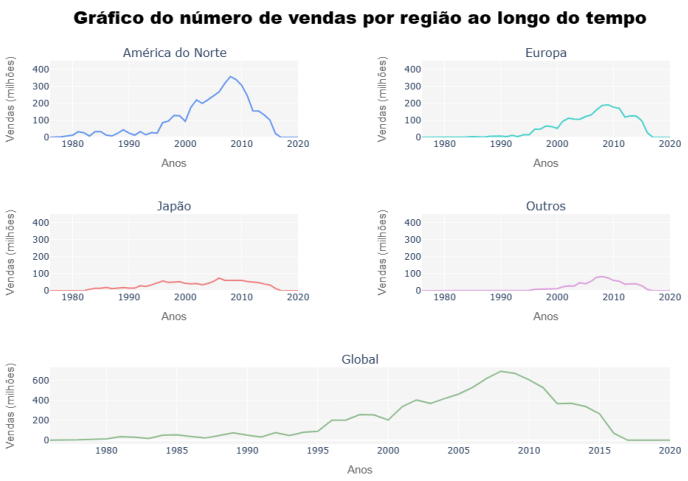

Fig.1- Visualização 1

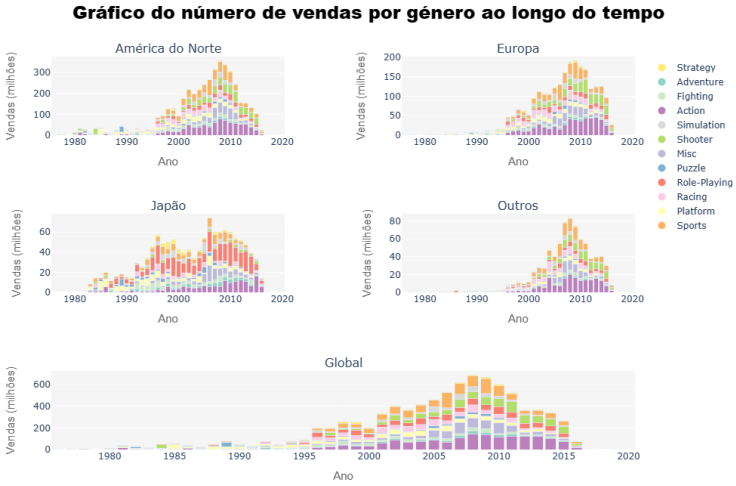

Fig.2- Visualização 2

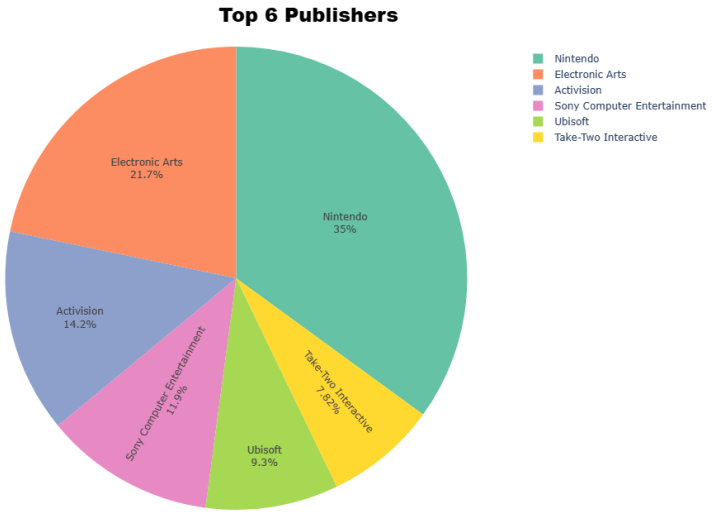

Fig.3- Visualização 3

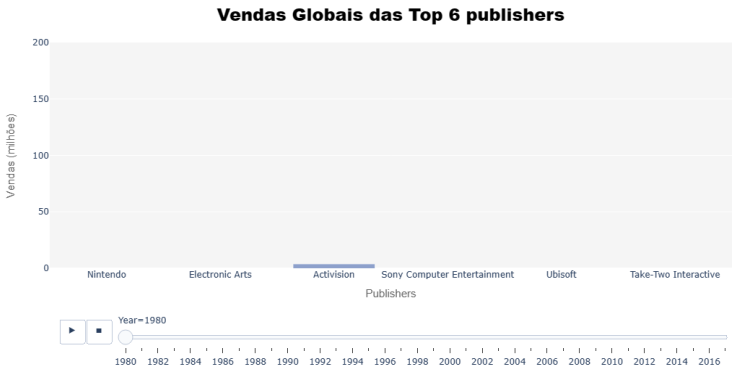

Fig.4- Visualização 4

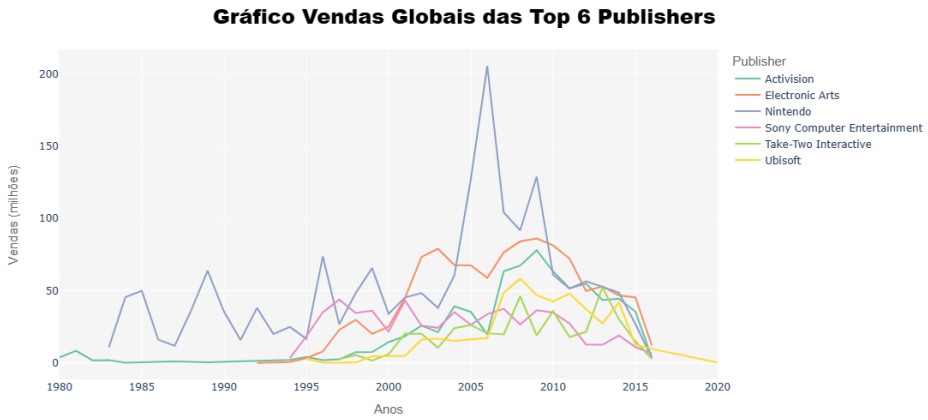

Fig.5 - Visualização 5

Fig.6 - Visualização 6

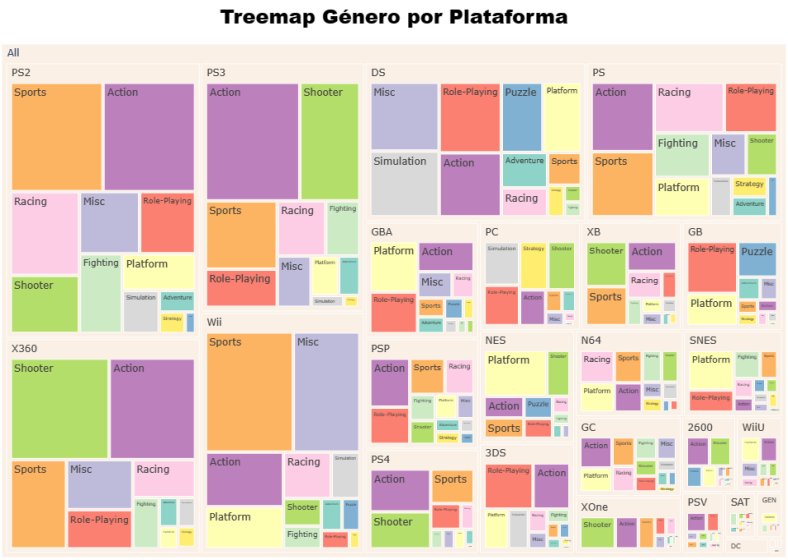

Fig.7- Visualização 7

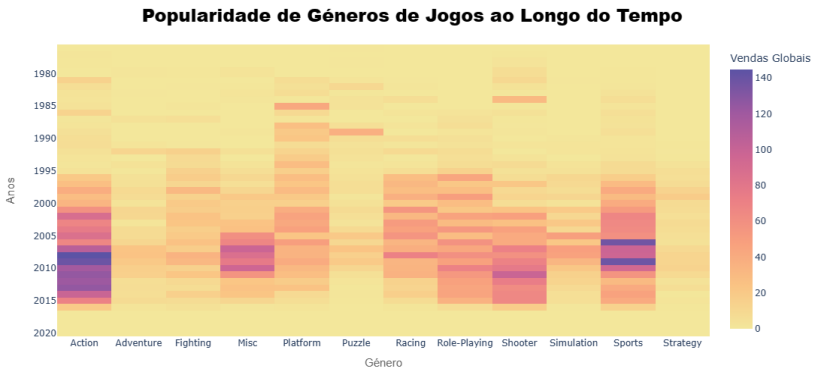

Fig.8- Visualização 8

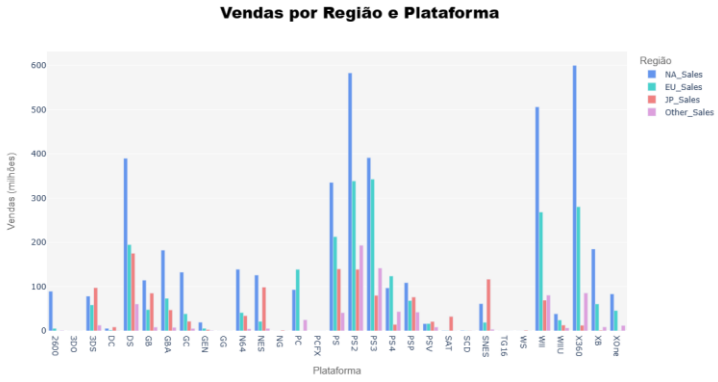

Fig.9- Visualização 9

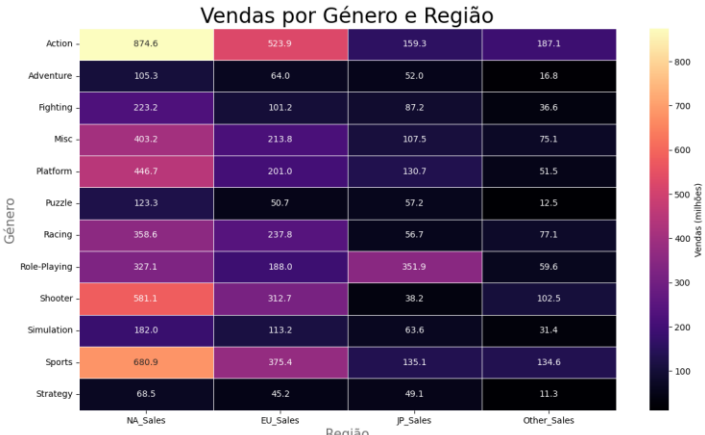

Fig.10 - Visualização 10

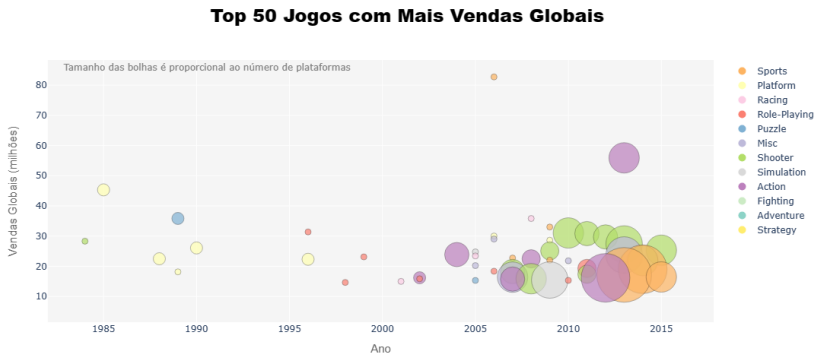

Fig.11 - Visualização 11

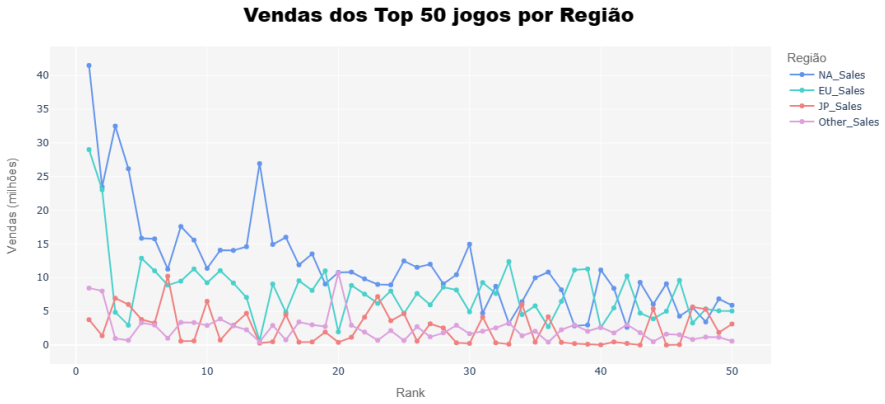

Fig.12 - Visualização 12

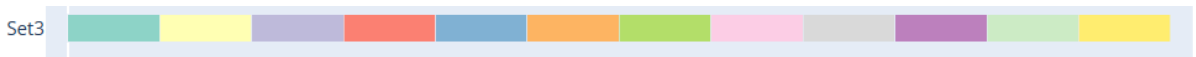

Fig.13 - Exemplo de Paleta utilizada (Set3)

IX - Referências

[1] Video Game Sales (kaggle.com)

[2] Video Games Sales | Kaggle

[3] 2. Video Games sales Project | Kaggle

[4] Video Game Sales Analysis and Visualization. | Kaggle

[5] Akmal Video Games Sales | Kaggle

[6] Video Game Sales with Plotly | Kaggle

[7] plotly.graph_objects.Figure — 5.18.0 documentation

[8] Plotly Python Graphing Library

[9] Discrete colors in Python (plotly.com)

[10] List of named colors — Matplotlib 3.8.2 documentation

[11] plotly.subplots.make_subplots — 5.18.0 documentation

[12] Subplots in Python (plotly.com)

[13] Pie charts in Python (plotly.com)

[14] Bubble charts in Python (plotly.com)

[15] Patterns, hatching, texture in Python (plotly.com)

[16] Text and annotations in Python (plotly.com)

[17] plotly.graph_objects.Bar — 5.18.0 documentation

[18] Intro to animations in Python (plotly.com)

[19] Treemap charts in Python (plotly.com)

[20] Creating annotated heatmaps — Matplotlib 3.8.2 documentation

[21] Heat map in matplotlib | PYTHON CHARTS (python-charts.com)

[22] Heatmaps in Python (plotly.com)

[23] Bar charts in Python (plotly.com)

[24] seaborn.heatmap — seaborn 0.13.0 documentation

[25] Choosing Colormaps — Matplotlib 3.8.2 documentation