cesae digital

Centro para o Desenvolvimento de Competências Digitais

Projeto de Programação em R

Formando: David Nogueira

Curso: Data Analyst

Data: 09/04/2024

Índice

Descrição do projeto 2

Questões do tipo BI (business intelligence) 3

Questões do tipo BA (business analytics) 5

Conclusão 11

Bibliografia/Siteografia 12

Descrição do projeto

O projeto de programação em R tem como objetivo realizar uma análise abrangente dos dados de jogos disponíveis na plataforma Steam no período de 2013 a 2023. A Steam é uma das maiores plataformas de distribuição digital de jogos e programas para PC, oferecendo uma ampla variedade de títulos de diversos generos.

Os dados analisados consistem em um conjunto de informações relevantes sobre os jogos, incluindo o nome do jogo, data de lançamento, preço, número de avaliações positivas e negativas, identificação de categorias (app_id), número máximo de utilizadores.

O objetivo principal do projeto é explorar esses dados para identificar insights interessantes sobre o mercado de jogos da Steam ao longo dos anos. Algumas das análises planejadas incluem:

O projeto tem a visualização dos dados por meio de gráficos e tabelas para facilitar a interpretação e comunicação dos resultados. Além disso, foram utilizadas técnicas de pré-processamento e limpeza de dados (excel, powerBI e R) para garantir a qualidade e integridade das análises.

Por fim, o objetivo é fornecer insights valiosos para desenvolvedores de jogos e programas, analistas de mercado e entusiastas da indústria de jogos, contribuindo para uma compreensão mais profunda do cenário dos jogos na plataforma Steam ao longo dos anos.

Questões do tipo BI (business intelligence)

Qual o jogo mais vendido ao longo dos anos (2013-2023)?

Essa questão é fundamental para entender a popularidade e o sucesso comercial dos jogos na plataforma Steam ao longo do tempo. Identificar o jogo mais vendido pode fornecer insights valiosos sobre as preferências dos jogadores e as tendências de mercado. Além disso, essa informação pode ser útil para desenvolvedores de jogos e investidores ao avaliar o potencial de lucro de novos projetos.

Como o meu dataset não tem uma coluna de sales, eu verifiquei o número de owners máximo, pois cada jogador teve de adquirir o jogo. No entanto há jogos grátis (price = 0), mas que são muito rentáveis por terem transações dentro do jogo. Por isso, dividi esta pergunta com duas análises, uma sem jogos grátis e outra com jogos grátis.

```
> jogo_mais_vendido_sem_gratis
           name release_date price positive negative  app_id max_owners
45369 New World         2021 39.99   154914    73900 1063730  100000000
```

O jogo mais vendido não incluído os grátis é o New World, com um total de **cem milhões** (100000000) **de jogadores**. Consultando a tabela auxiliar verifica-se que app_id corresponde a um jogo MMO (Massively Multiplayer Online) **onde há cooperação entre jogadores**, mas também **jogador vs jogador** (PVP) e onde é possível efetuar **transações dentro do jogo com dinheiro**. O **preço de $39.99 está dentro do intervalo esperado**, pois a maioria dos jogos de desenvolvedoras conhecidas têm um preço de lançamento de $69.99 para jogos de franquias já famosos entre os jogadores. O **feedback dos jogadores** deste jogo verifica-se que **é praticamente positivo** (154914) comparativamente ao negativo (73900). Ao pesquisar na steam foi verificado que a razão do **feedback negativo foi devido às transações dentro do jogo desequilibrarem as lutas que ocorrem entre os jogadores** [1].

```
> jogo_mais_vendido_com_gratis
        name release_date price positive negative app_id max_owners
45483 Dota 2         2013     0  1477153   300437    570  200000000
```

O jogo mais vendido incluído os grátis é o Dota 2, com um total de **duzentos milhões** (200000000) **de jogadores**. Consultando a tabela auxiliar verifica-se que app_id corresponde a um jogo **onde há cooperação entre jogadores** e onde é possível efetuar **transações dentro do jogo com dinheiro**. O jogo é **grátis**, mas tem transações dentro do jogo, e sendo que é um **jogo de 2013 já tem uma comunidade bem definida que paga para ter conteúdo exclusivo dentro do jogo**. O **feedback dos jogadores** deste jogo verifica-se que **é muito positivo** (1477153) comparativamente ao negativo (300437). Ao pesquisar na steam foi verificado que a razão do **feedback negativo foi devido a ser um jogo difícil de aprender e por haver na comunidade vários jogadores com comportamentos incorretos** [2].

Qual o maior preço de jogos na plataforma?

Saber qual é o jogo com o maior preço na plataforma Steam é importante para entender a amplitude dos preços dos jogos disponíveis. Isso pode indicar se há uma procura por jogos premium e se os jogadores estão dispostos a pagar mais por determinados títulos. Além disso, essa informação pode ser útil para consumidores que desejam tomar decisões informadas sobre suas compras de jogos.

```
> maior_preco
                       name release_date price positive negative app_id max_owners
32856 Aartform Curvy 3D 3.0         2013 299.9       32       13 253670      20000
```

O maior preço na plataforma steam é de $299.9 e é do Aartform Curvy 3D 3.0. Consultando a tabela auxiliar verifica-se que app_id corresponde a jogos de um único jogador (singleplayer), como tal fui pesquisar e averiguei que **não é um jogo, mas um programa de escultura para artistas 3D**. Verifica-se que é este software já tem mais de 10 anos (**2013**) e mesmo assim tem poucos utilizadores (max_owners=20000). O **feedback é praticamente positivo, mas é pequeno** (32 positivos e 13 negativos). A razão do **feedback negativo é do preço elevado para a quantidade de ferramentas que se podem utilizar no programa** [3].

Qual o jogo com feedback mais positivo?

Identificar o jogo com o feedback mais positivo pode ser útil para os jogadores que procuram experiências de jogo de alta qualidade. Além disso, para desenvolvedores de jogos, entender o que os jogadores valorizam e apreciam em um jogo pode ajudar a orientar o desenvolvimento de futuros projetos e melhorar a qualidade geral dos jogos lançados. Esta informação também pode ser valiosa para editores e investidores ao avaliar o potencial de sucesso de um determinado jogo.

```
> mais_feedback_positivo
       name release_date price positive negative app_id max_owners
45483 Dota 2         2013     0  1477153   300437    570  200000000
```

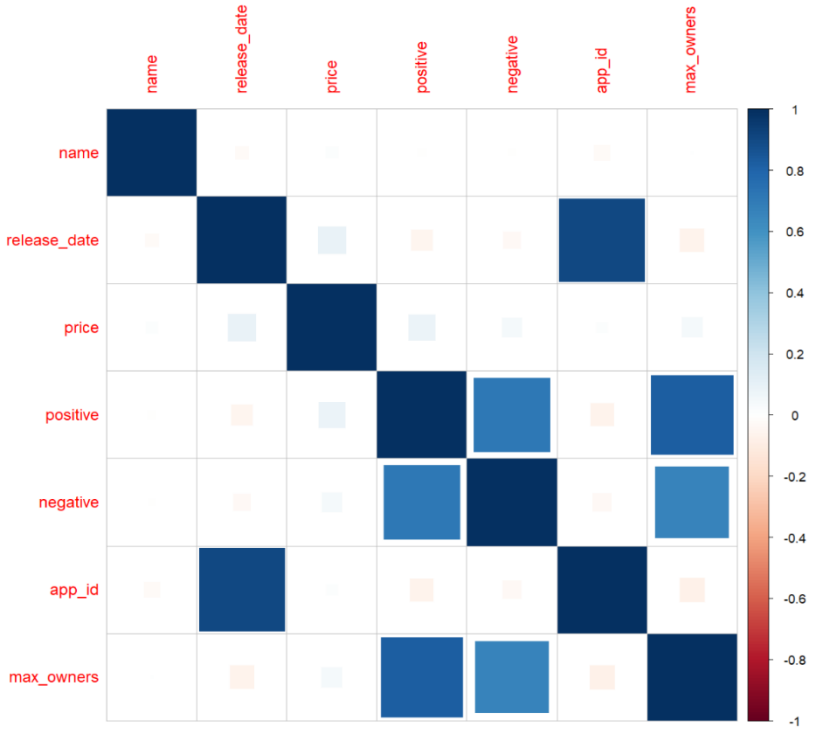
O jogo com mais feedback positivo é o Dota 2, com quase um milhão e meio de feedback positivo. O Dota 2 já foi referenciado neste projeto pois também é jogo que tem maior número de utilizadores (max_owners). Este feedback muito positivo é devido a ser um jogo com mais de 10 anos onde já tem uma comunidade construída e passado todo este tempo ainda continua a ter atualizações frequentes. Ao ser um **jogo grátis** faz com que muitos **jogadores tenham a possibilidade de experimentar**, gostarem da experiência e **contribuírem para o feedback** [2].

Questões do tipo BA (business analytics)

Nota: na limpeza dos dados foram removidos dois fatores (mínimo de utilizadores e tempo para concluir o jogo) devido a encontrarem-se incompletos e por serem irrelevantes para as analises a efetuar.

Para ter uma **visão geral dos dados** efetuei uma tabela e um gráfico de correlações recorrendo ao método "square":

	name	release_date	price	positive	negative	app_id	max_owners
name	1.0000000000	-0.02096750	0.01472169	-0.005848503	-0.003741659	-0.02871208	0.0004475138
release_date	-0.0209675036	1.00000000	0.09592200	-0.055765131	-0.032913497	0.90415680	-0.0679578638
price	0.0147216862	0.09592200	1.00000000	0.081686196	0.046096420	0.01237417	0.0498768125
positive	-0.0058485029	-0.05576513	0.08168620	1.000000000	0.714050656	-0.06448060	0.8261974524
negative	-0.0037416593	-0.03291350	0.04609642	0.714050656	1.000000000	-0.03841487	0.6601004334
app_id	-0.0287120840	0.90415680	0.01237417	-0.064480599	-0.038414875	1.00000000	-0.0720966870
max_owners	0.0004475138	-0.06795786	0.04987681	0.826197452	0.660100433	-0.07209669	1.0000000000



Analisando a tabela e o gráfico de **correlações de dados** verifica-se que:

1. **Data de lançamento (release_date) com as categorias dos jogos (app_id)** apresentam a maior correlação que é de 0,904, num intervalo de 0 a 1.
2. **Número máximo de utilizadores (max_owners) com o feedback positivo (positive)** apresentam uma correlação de 0,826
3. **Feedback negativo (negative) com o feedback positivo (positive)** apresentam uma correlação de 0,714
4. **Número máximo de utilizadores (max_owners) com o feedback negativo (negative)** apresentam uma correlação de 0,660

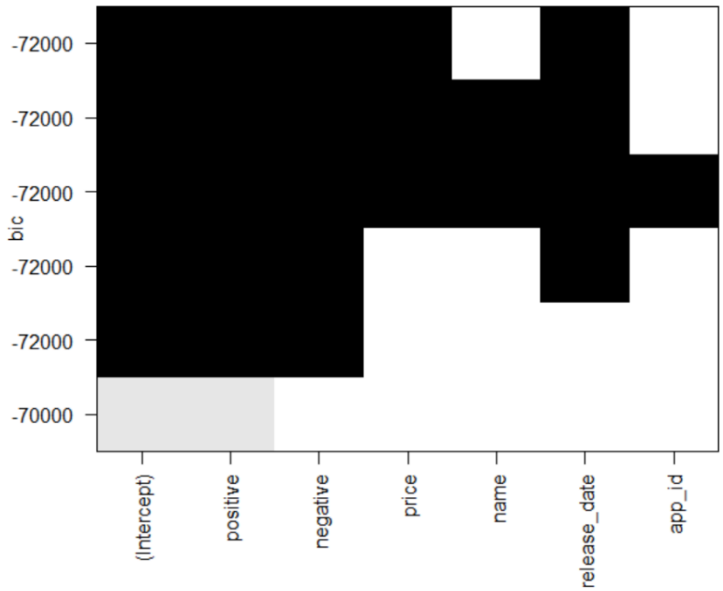
Quais são os fatores que mais afetam as vendas de um jogo?

Essa questão é crucial para entender os principais impulsionadores por trás das vendas de jogos na plataforma Steam. Ao analisar diversos fatores como preço, avaliações dos usuários, data de lançamento, genero do jogo, entre outros, é possível identificar quais desses fatores têm o maior impacto nas vendas. Isso pode ajudar desenvolvedores, editores e investidores a tomar decisões estratégicas, como definir preços adequados, planejar campanhas de marketing e direcionar o desenvolvimento de novos jogos.

Nota: No **dataset não existe uma coluna de vendas(sales)**. Para se **verificar as vendas será utilizado o número máximo de utilizadores (max_owners)**, pois esses jogadores tiveram de adquirir os jogos para poderem jogá-los.

Para a análise sobre os fatores que mais afetam as vendas do jogo, foi averiguado os campos que têm maior correlação com o campo "max_onwers" através de uma tabela e os melhores subsets através de um gráfico.

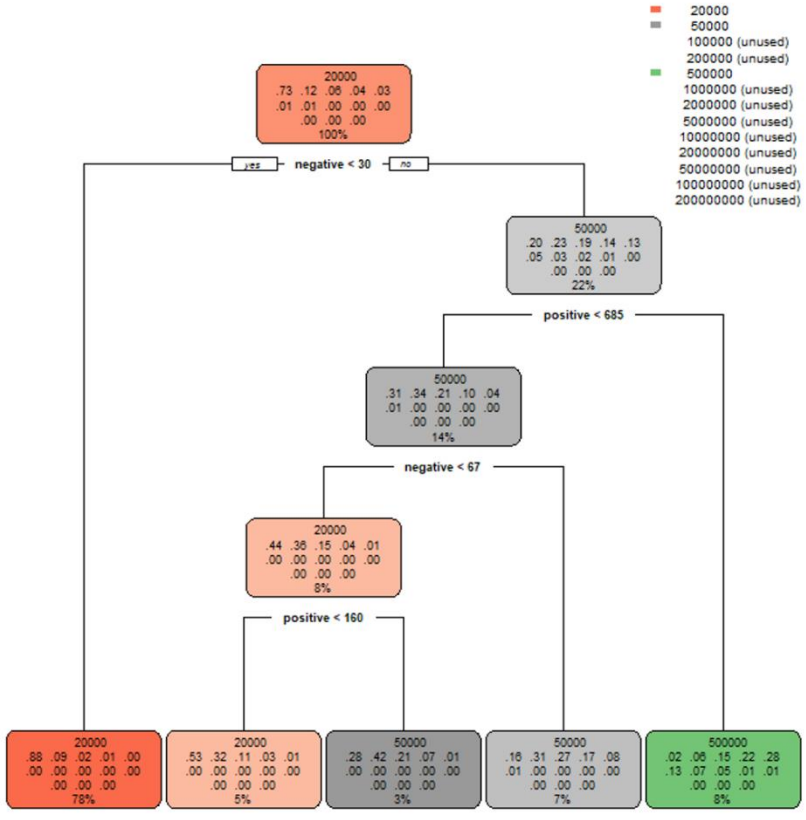
max_owners	positive	negative	price	name	release_date	app_id
1.0000000000	0.8261974524	0.6601004334	0.0498768125	0.0004475138	-0.0679578638	-0.0720966870



Com tabela e o gráfico verifica-se que os campos de **maior correlação com o campo max_owners são o feedback positivo e negativo**, portanto será verificado posteriormente os fatores que podem afetar esse feedback. Posteriormente os campos com **alguma importância é a data de lançamento(release_date) e o preço (price)**.

As **razões para o valor alto de correlação entre o feedback positivo e o número de utilizadores**, seguida pela correlação com o feedback negativo:

- **Qualidade do Jogo**: Jogos com uma receção positiva dos jogadores tendem a atrair mais utilizadores. Se um jogo é bem recebido pela comunidade de jogadores devido à sua jogabilidade, gráficos, enredo ou outros elementos, é mais provável que atraia uma base de utilizadores maior.
- **Recomendações e Avaliações**: Os jogadores muitas vezes tomam decisões de compra com base em avaliações e recomendações de outros jogadores. Se um jogo recebe feedback positivo de outros jogadores, ele é mais propenso a atrair novos utilizadores que confiam nessas avaliações.
- **Retenção de Utilizadores**: Jogos com feedback positivo têm maior probabilidade de reter utilizadores existentes por mais tempo. A satisfação dos jogadores com o jogo os mantém entretidos por mais tempo, resultando em uma base de utilizadores maior ao longo do tempo.
- **Buzz Marketing (Boca a Boca)**: Jogos com feedback positivo são mais propensos a gerar marketing boca a boca positivo, onde os jogadores recomendam o jogo a amigos e familiares. Isso pode levar a um aumento no número de utilizadores conforme mais pessoas se tornam conscientes e interessadas no jogo.



Embora haja uma correlação entre o feedback positivo e o número de utilizadores, outros fatores não incluídos na análise também podem influenciar o número de utilizadores. Além disso, a correlação entre o feedback negativo e o número de utilizadores pode indicar que **jogos com baixo feedback negativo também tendem a atrair mais utilizadores** (como se pode verificar na arvore de decisão que 78% dos utilizadores optam por um jogo quando este tem um feedback negativo baixo).

De acordo com um conjunto de fatores, o jogo terá feedback positivo?

Esta questão busca entender quais são os fatores-chave que influenciam o feedback positivo dos utilizadores sobre um jogo. Ao analisar uma combinação de fatores como preço, avaliações dos utilizadores, genero do jogo, duração do jogo, entre outros, é possível determinar quais características levam a um feedback mais positivo. Isso pode ser valioso para desenvolvedores de jogos ao orientar o processo de desenvolvimento para criar jogos que atendam às expectativas e preferências dos jogadores, resultando em uma melhor receção e sucesso no mercado.

Para a análise sobre o conjunto de fatores que levem a previsão de um jogo ter feedback positivo, foi averiguado os campos que têm maior correlação com o campo "positive" através de uma tabela.

positive	max_owners	negative	price	name	release_date	app_id
1.000000000	0.826197452	0.714050656	0.081686196	-0.005848503	-0.055765131	-0.064480599

Com esta tabela de correlações confirma-se mais uma vez a informação anteriormente verificada, em que o **feedback positivo tem uma forte correlação com o campo max_owners e o feedback negativo**. Posteriormente o campo com **alguma importância é o preço (price)**.

Posteriormente foi efetuado um modelo de previsão, em que se obteve o R^2:

```
Call:
lm(formula = positive ~ max_owners + negative + price, data = trainx)

Residuals:
    Min      1Q  Median      3Q     Max
-429060    -227      -3     119  468900

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.653e+02  4.572e+01  -5.802 6.61e-09 ***
max_owners   3.581e-03  4.070e-05  87.994  < 2e-16 ***
negative     3.041e+00  2.572e-02 118.240  < 2e-16 ***
price        3.495e+01  3.679e+00   9.500  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7115 on 39996 degrees of freedom
Multiple R-squared:  0.754,     Adjusted R-squared:  0.754
F-statistic: 4.087e+04 on 3 and 39996 DF,  p-value: < 2.2e-16
```

Um **coeficiente de determinação (R^2) de 0.754 indica que aproximadamente 75.4% da variabilidade na variável de resposta (positive, ou seja, o feedback positivo dos jogadores) pode ser explicada pelo modelo**. Isso significa que o modelo é capaz de capturar e explicar uma percentagem significativa da variabilidade observada no feedback positivo dos jogadores em relação aos jogos.

Um R^2 de 0.754 é considerado bastante alto e sugere que o modelo tem uma capacidade razoavelmente boa de prever o feedback positivo com base nos dados fornecidos. Isso pode ser interpretado como uma indicação de que os fatores considerados no modelo (como feedback negativo, preço, etc.) têm uma forte relação com o feedback positivo dos jogadores.

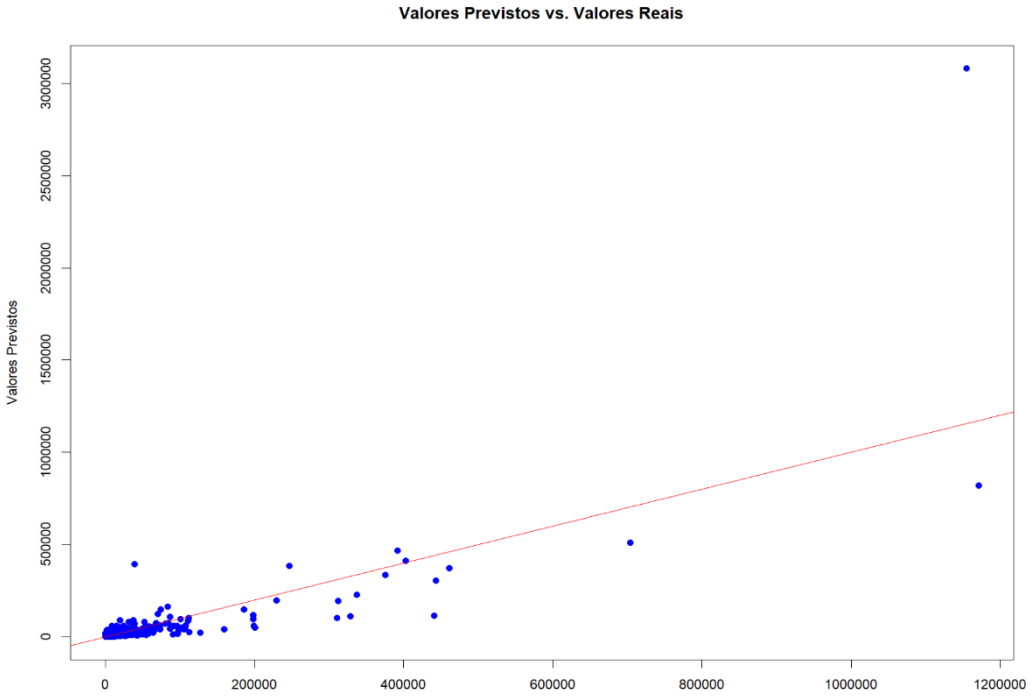
No entanto, é importante ressaltar que o R^2 por si só não diz nada sobre a validade ou a qualidade do modelo. Portanto, é sempre crucial realizar uma análise mais aprofundada, incluindo a avaliação de outras métricas de desempenho do modelo e a consideração de possíveis limitações nos dados e na modelagem. Como tal, fiz uma tabela e um gráfico de dispersão com os dados previstos comparativamente aos dados reais:

Tabela com os valores reais:

```
> summary(steamgames)
      name         release_date      price             positive          negative          app_id          max_owners
 Min.   :    1   Min.   :2013   Min.   :  0.000   Min.   :      0   Min.   :     0.0   Min.   :    570   Min.   :    20000
 1st Qu.:14484   1st Qu.:2018   1st Qu.:  1.990   1st Qu.:      4   1st Qu.:     1.0   1st Qu.: 676023   1st Qu.:    20000
 Median :29593   Median :2020   Median :  4.990   Median :     16   Median :     4.0   Median :1095830   Median :    20000
 Mean   :29602   Mean   :2019   Mean   :  7.815   Mean   :   1046   Mean   :   193.5   Mean   :1165656   Mean   :   135329
 3rd Qu.:44690   3rd Qu.:2021   3rd Qu.:  9.990   3rd Qu.:     80   3rd Qu.:    24.0   3rd Qu.:1579013   3rd Qu.:    50000
 Max.   :59798   Max.   :2023   Max.   :299.900   Max.   :1477153   Max.   :895978.0   Max.   :2690780   Max.   :200000000
```

Tabela com os valores previstos:

```
     price             positive          negative          max_owners          predictions
 Min.   :  0.000   Min.   :      0   Min.   :     0.0   Min.   :    20000   Min.   :  -193.6
 1st Qu.:  1.990   1st Qu.:      4   1st Qu.:     1.0   1st Qu.:    20000   1st Qu.:   -75.4
 Median :  4.990   Median :     16   Median :     4.0   Median :    20000   Median :    82.7
 Mean   :  7.853   Mean   :   1058   Mean   :   225.6   Mean   :   129946   Mean   :  1160.5
 3rd Qu.:  9.990   3rd Qu.:     81   3rd Qu.:    24.0   3rd Qu.:    50000   3rd Qu.:   366.8
 Max.   :299.900   Max.   :1171197   Max.   :895978.0   Max.   :100000000   Max.   :3082297.0
```



Comparando os valores previstos do feedback positive com os valores reais no conjunto de dados original verifica-se que os valores são próximos. O valor mínimo do real e do previsto em ambos é 0. O valor máximo do feedback positivo real é 1477153 enquanto o previsto é 1171197, havendo então uma diferença de 305956 entre o máximo real e o previsto. A média do valor real é de 1046 enquanto o valor previsto é de 1058.

Com análise do gráfico de dispersão averigua-se que os valores previstos e os valores reais estão próximos. Mas também há alguns valores estão mais dispersos e não tão condensados.

Portanto, este modelo é bom para determinar previsões de feedback positivo de jogos, tendo só em consideração que o valor máximo pode não ser tão próximo ao real, apresentando valores abaixo do real.

Conclui-se então que **um jogo terá feedback positivo se**:

- **Tiver um grande número de jogadores**, pois o jogo irá receber mais feedback positivo. Isso se o jogo ficar popular e houver satisfação nos utilizadores com a experiência geral do jogo.
- **Receber menos feedback negativo.** Isso sugere uma relação inversa entre a satisfação dos usuários e as críticas negativas.
- **Preço corresponder às expectativas.** Jogos com preços mais altos podem ser percebidos como oferecendo maior valor e, portanto, receber mais feedback positivo, se corresponderem às expectativas dos utilizadores.

Conclusão

Com a realização deste projeto de programação em R sobre os jogos da plataforma Steam no período de 2013 a 2023, foi possível obter insights valiosos sobre o comportamento dos jogos e dos utilizadores. As perguntas selecionadas para análise foram fundamentais para compreender diversos aspetos do mercado de jogos online.

Iniciamos identificando o jogo mais vendido ao longo dos anos, considerando tanto jogos pagos quanto gratuitos. Em seguida, foi determinado o maior preço de jogos na plataforma e identificar o jogo com o feedback mais positivo, o que nos permitiu compreender melhor as preferências dos jogadores.

Uma das análises mais importantes foi a identificação dos fatores que mais afetam as vendas de um jogo. Ao explorar a correlação entre diferentes variáveis, como feedback positivo, feedback negativo, preço e outros, pudemos identificar padrões significativos que influenciam o sucesso comercial de um jogo na plataforma.

Uma questão adicional levantada foi se, com base em um conjunto de fatores, seria possível prever se um jogo teria feedback positivo.

Apesar das dificuldades encontradas, como a necessidade de limpeza prévia do dataset (mais de 60 mil dados limpos e normalizados em excel, power BI e R) e a seleção de modelos adequados, como a utilização do rpart para construção da árvore de decisão (em vez do C50) e a análise do R^2 e gráfico de dispersão (sem utilizar o modelo roc) para avaliar a qualidade do modelo de previsão, conseguimos chegar a conclusões relevantes.

Em suma, este projeto proporcionou uma visão abrangente do mercado de jogos da Steam, demonstrando a importância da análise de dados na compreensão e tomada de decisões no setor de entretenimento digital. As descobertas e metodologias empregadas neste projeto podem servir de base para análises futuras e para aprimorar estratégias de desenvolvimento e comercialização de jogos na plataforma.

Bibliografia/Siteografia

[1] Valve Corporation, *New World*, consultada em 06 de Abril de 2024: https://store.steampowered.com/app/1063730/New_World/

[2] Valve Corporation, *Dota 2*, consultada em 06 de Abril de 2024: https://store.steampowered.com/app/570/Dota_2/

[3] Valve Corporation, *AartfromCurvy 3D 3.0*, consultada em 06 de Abril de 2024: https://store.steampowered.com/app/253670/Aartform_Curvy_3D_30/