

PHÂN TÍCH RATINGS TRÊN NỀN TẢNG STEAM GAME

START MENU SIGN IN

LINK SOURCE:
[HTTPS://STORE.STEAMPOWERED.COM/SEARCH/?
CATEGORY1=998&ndl=1](https://store.steampowered.com/search/?CATEGORY1=998&ndl=1)



GIỚI THIỆU CHỦ ĐỀ VÀ ĐỘNG LỰC

ĐỒ ÁN NÀY TẬP TRUNG VÀO VIỆC NGHIÊN CỨU VÀ PHÂN TÍCH ĐÁNH GIÁ GAME TRÊN STEAM, MỘT NỀN TẢNG PHÂN PHỐI GAME TRỰC TUYẾN HÀNG ĐẦU.

SỬ DỤNG CÁC KỸ THUẬT KHOA HỌC DỮ LIỆU, ĐỒ ÁN SẼ KHÁM PHÁ CÁC XU HƯỚNG TRONG NGÀNH CÔNG NGHIỆP GAME, SỰ TƯƠNG TÁC CỦA CỘNG ĐỒNG NGƯỜI CHƠI, VÀ NHỮNG YẾU TỐ QUYẾT ĐỊNH ĐẲNG CẤP CỦA MỘT TRÒ CHƠI TRÊN NỀN TẢNG NÀY.



AI

[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)

MENU

01

07

12



QUY TRÌNH

◆ CÁC CHỦ ĐỀ



Crawl



Tiền xử lí



Kỹ thuật đặc trưng



EDA



Mô hình hóa



Triển khai mô hình



Kết luận

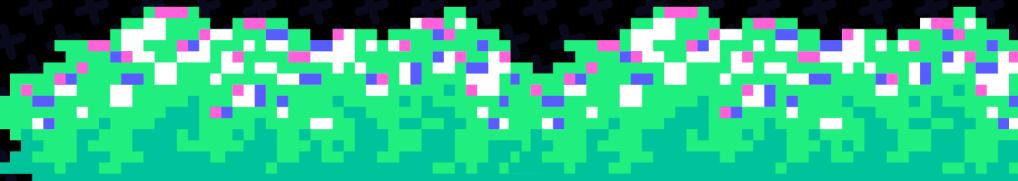
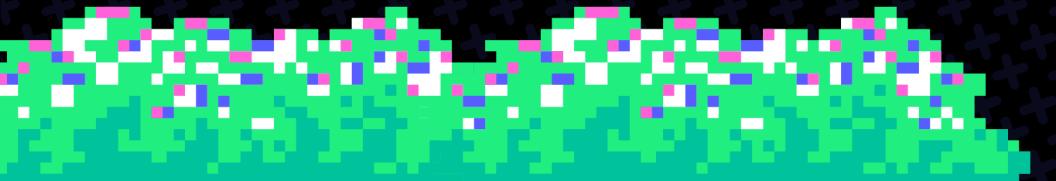
SIGN IN



[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)

CRAWL

BEAUTIFUL SOUP



SIGN IN

CRAWL

[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)



Cào từng link game

Match your search. 4,465 titles have been excluded based on your preferences. However, none of these titles would appear on the first page of results.



Baldur's Gate 3



3 Aug, 2023

-10%

~~990.000,00đ~~
891.000,00đ



Lethal Company



24 Oct, 2023



14



Counter-Strike 2



22 Aug, 2012



Free



ELDEN RING



25 Feb, 2022



-40%

~~898.000,00đ~~
539.000,00đ



Cyberpunk 2077



10 Dec, 2020



-50%

~~990.000,00đ~~
495.000,00đ



PUBG: BATTLEGROUNDS



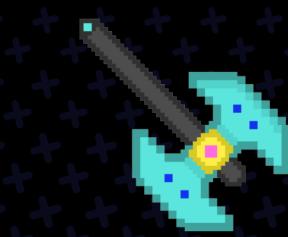
21 Dec, 2017



Free

~~Free~~

Cào giá game và
%Discount



SIGN IN

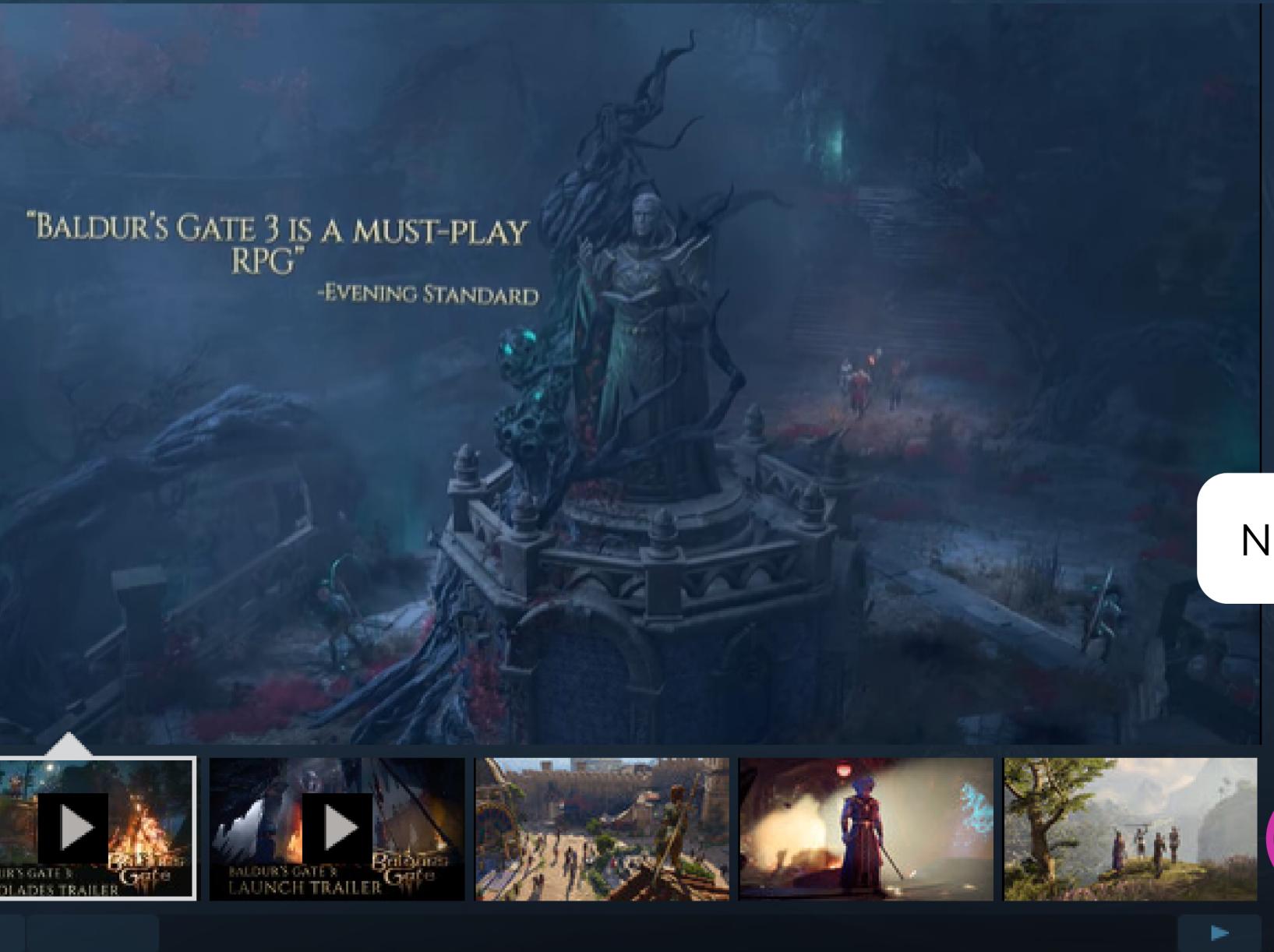
CRAWL

[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)

All Games > RPG Games > Baldur's Gate 3

Baldur's Gate 3

Community Hub



Baldur's Gate 3 is a story-rich, party-based RPG set in the universe of Dungeons & Dragons, where your choices shape a tale of fellowship and betrayal, and the lure of absolute power.

Ngày xuất bản

Overwhelmingly Positive (31,774)
Overwhelmingly Positive (465,455)

RELEASE DATE: 3 Aug, 2023

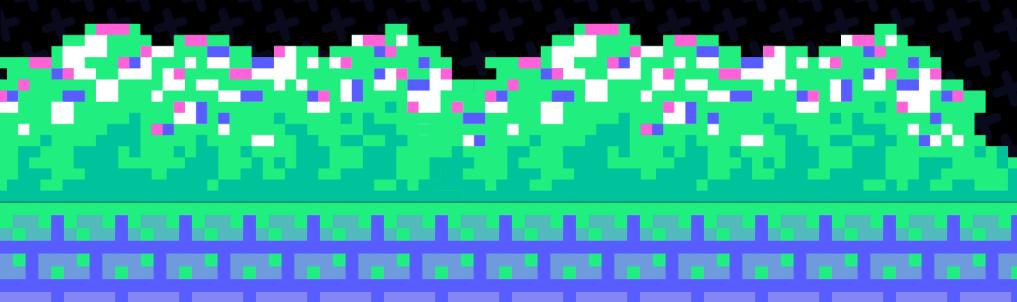
DEVELOPER: Larian Studios

PUBLISHER: Larian Studios

Tags

Popular user-defined tags for this product:

RPG Choices Matter Story Rich Adventure +



SIGN IN



Ngôn ngữ



Languages:	Interface	Full Audio	Subtitles
English	✓	✓	✓
French	✓		✓
German	✓		✓
Spanish - Spain	✓		✓
Polish	✓		✓

See all 15 supported languages

CRAWL

[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)



CUSTOMER REVIEWS

Overall Reviews:

Overwhelmingly Positive (465,459)

REVIEW TYPE

PURCHASE TYPE

- All (521,147)
- Positive (504,212)
- Negative (16,935)

Số review tích cực, tiêu
cực, và tổng



TITLE: Baldur's Gate 3

GENRE: Adventure, RPG, Strategy

DEVELOPER: Larian Studios

PUBLISHER: Larian Studios

FRANCHISE: Baldur's Gate

RELEASE DATE: 3 Aug, 2023

EARLY ACCESS RELEASE DATE: 7 Oct, 2020



Tên, Thể loại, Nhà phát triển,
Nhà phát hành, Thương hiệu



SIGN IN

[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)

TIỀN XỬ LÍ

PANDAS
NUMPY

MATPLOTLIB, SEABORN



MENU



TIỀN XỬ LÍ



Mỗi cột có ý nghĩa gì?



.shape
Dữ liệu có 9583 dòng và 15 cột

- Title: Tên trò chơi
- Genre: Thể loại trò chơi
- Tags: Nhãn trò chơi
- withDLC: game có downloadabel content hay không?
- isMature: game có nội dung không phù hợp cho trẻ nhỏ không?
- Franchise: tên thương hiệu
- ReleaseDate: Ngày phát hành
- Developer: Nhà phát triển
- Publisher: Nhà phát hành
- Languages: Ngôn ngữ được hỗ trợ
- PositiveReviews: Số đánh giá tích cực
- TotalReviews: Tổng số đánh giá của người chơi
- NegativeReviews: Số đánh giá tiêu cực
- OriginalPrice: Giá bán gốc
- DiscountPercent: Giảm giá áp dụng trên giá gốc(%)

[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)

MENU



TIỀN XỬ LÍ

Tạo các cột cần thiết

Tính lại giá trị ở cột OriginalPrice sau khi đã áp dụng giảm giá

- Chuyển tất cả giá trị với đơn vị tiền tệ "\$" sang "VND"
- Loại bỏ kí tự hiển thị đơn vị tiền tệ
- Đưa giá trị "Free" thành 0 ở cột OriginalPrice

=> Tạo thêm cột mới DiscountedPrice

df['DiscountedPrice']

```
0      891000.0
1          0.0
2     142000.0
3     495000.0
4     538800.0
      ...
9578      0.0
9579    33750.0
9580      0.0
9581    87500.0
9582    73500.0
```

Name: DiscountedPrice, Length: 9583, dtype: float64

[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)

MENU



TIỀN XỬ LÍ

 Xử lý missing value

- Bỏ cột **Franchise** chứa nhiều dữ liệu bị thiếu (hơn 50%)
- Loại bỏ các game có Title là NaN
- Kiểm tra dữ liệu ở cột Release Date, nếu giá trị là Nan, đồng nghĩa với trò chơi điện tử này chưa được phát hành chính thức nên xóa khỏi dataframe.
- Các cột còn lại fill theo mode

Title	0.542628
Genre	0.709590
Tags	0.542628
withDLC	0.000000
isMature	0.000000
Franchise	52.248774
ReleaseDate	0.594803
Developer	0.605238
Publisher	0.782636
Languages	0.542628
PositiveReviews	0.678284
TotalReviews	0.678284
NegativeReviews	0.678284
OriginalPrice	0.000000
DiscountPercent	14.223103
dtype:	float64

[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)

MENU



TIỀN XỬ LÍ



Kiểm tra tính toàn vẹn

Have to update the value in Positive or Negative review using Total and another

- Cột TotalReview khác tổng PositiveReview + NegativeReview

=> Vì vậy tính lại cột TotalReview

[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)

MENU



TIỀN XỬ LÍ



Kiểm tra trùng lặp

Drop duplicated rows, but keep first occurrence

- Dữ liệu có sự trùng lặp

=> Sử dụng `df.drop_duplicates(keep = 'first', inplace = True)`

[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)

MENU



TIỀN XỬ LÍ

- ❖ Xử lý các cột số (numeric columns):
 - withDLC
 - isMature
 - PositiveReviews
 - TotalReviews
 - NegativeReviews
 - OriginalPrice
 - DiscountPercent
 - discountedPrice

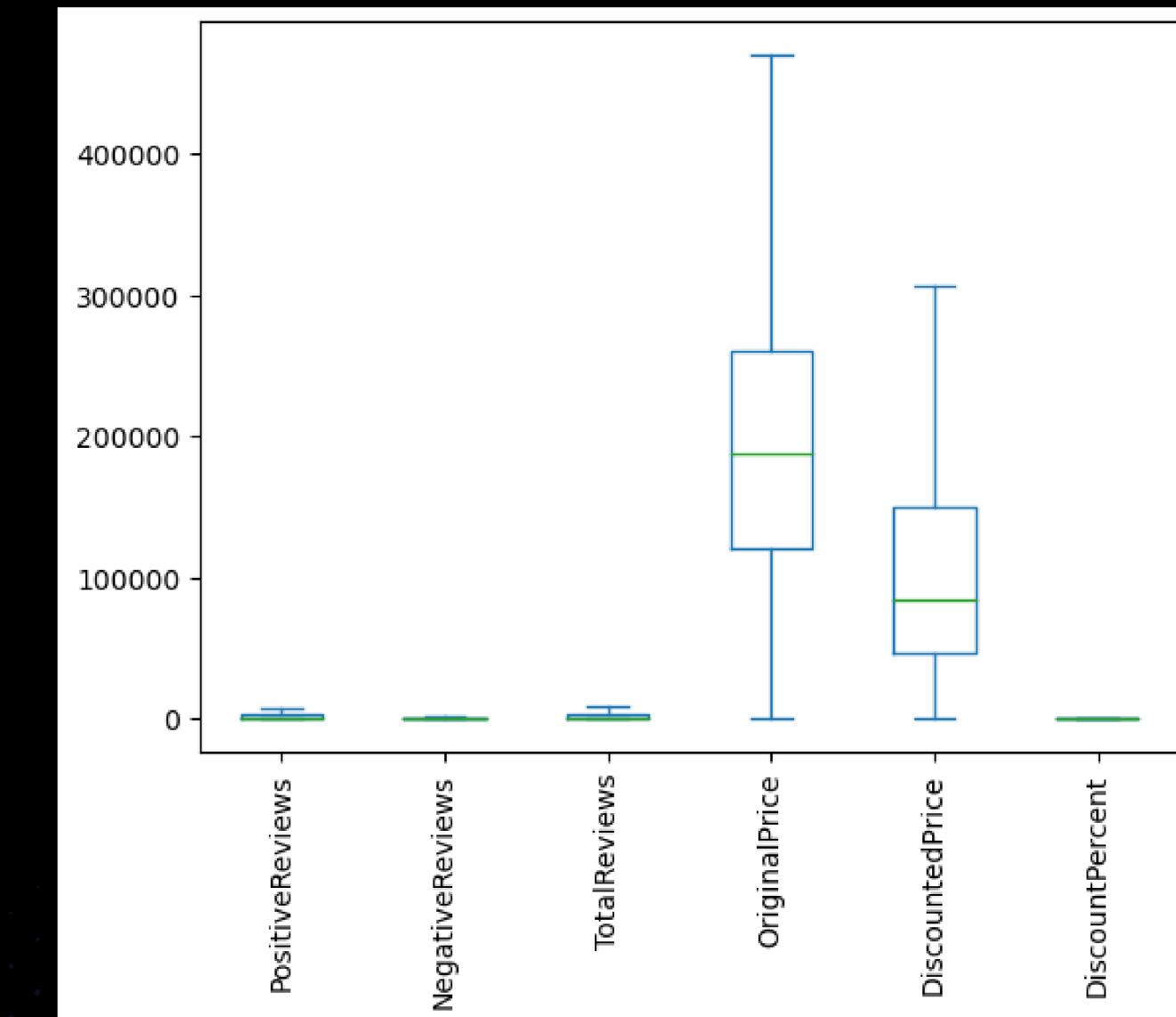
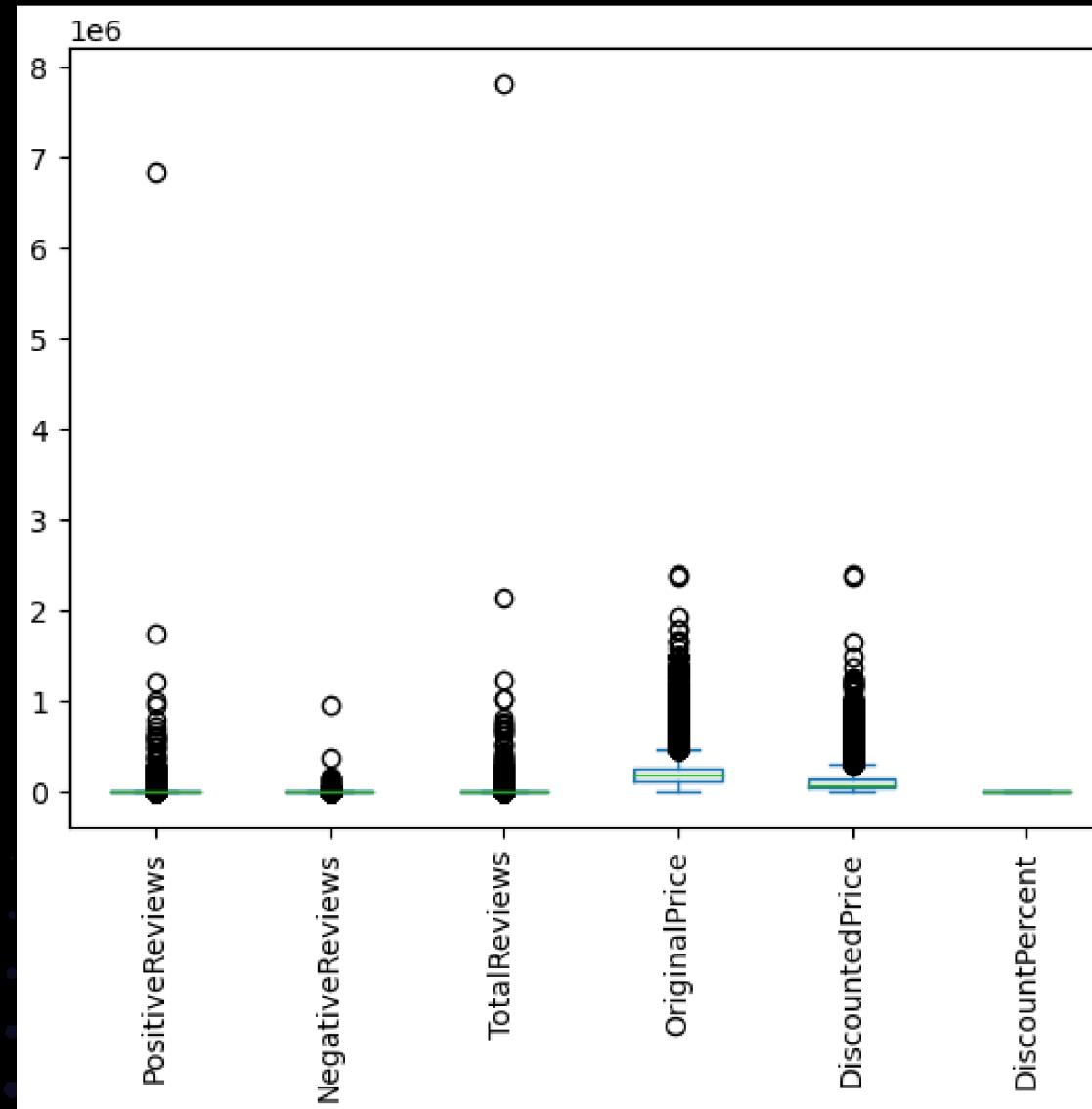
	withDLC	isMature	PositiveReviews	TotalReviews	NegativeReviews	OriginalPrice	DiscountPercent	DiscountedPrice
count	9269.000000	9269.000000	9.269000e+03	9.269000e+03	9269.000000	9.269000e+03	9269.000000	9.269000e+03
mean	0.449563	0.218254	8.497340e+03	9.664048e+03	1166.707628	2.366783e+05	0.444999	1.210667e+05
std	0.497476	0.413083	8.242057e+04	9.349746e+04	11853.418491	2.129635e+05	0.270401	1.361701e+05
min	0.000000	0.000000	6.000000e+00	9.000000e+00	0.000000	0.000000e+00	0.000000	0.000000e+00
25%	0.000000	0.000000	2.140000e+02	2.550000e+02	27.000000	1.200000e+05	0.250000	4.625000e+04
50%	0.000000	0.000000	7.590000e+02	9.010000e+02	111.000000	1.880000e+05	0.500000	8.460000e+04
75%	1.000000	0.000000	3.060000e+03	3.605000e+03	470.000000	2.600000e+05	0.700000	1.505000e+05
max	1.000000	1.000000	6.845621e+06	7.809836e+06	964215.000000	2.400000e+06	0.980000	2.400000e+06

MENU



TIỀN XỬ LÍ

♦ Outliers



[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)

MENU



TIỀN XỬ LÍ

❖ Xử lí các cột categorical:

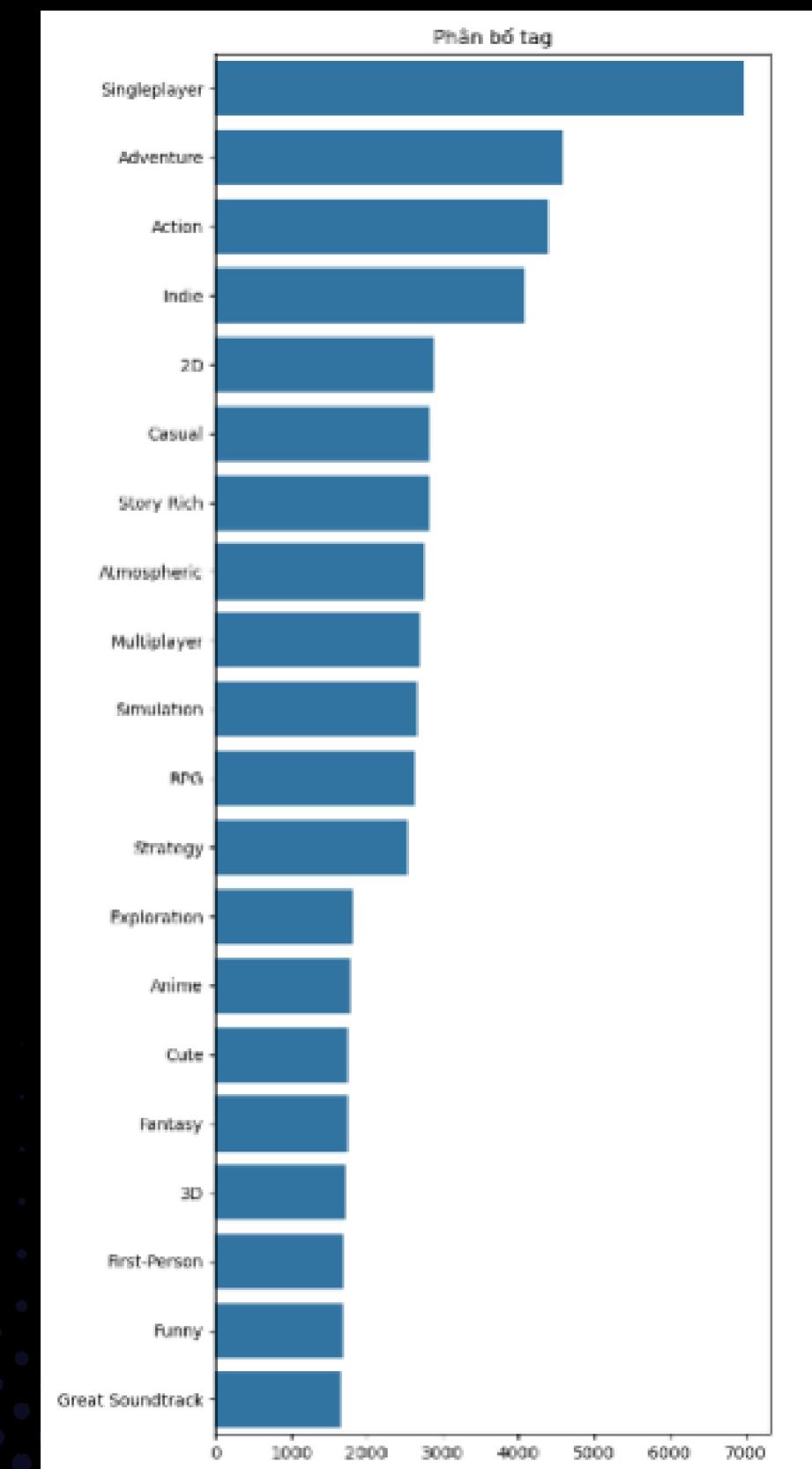
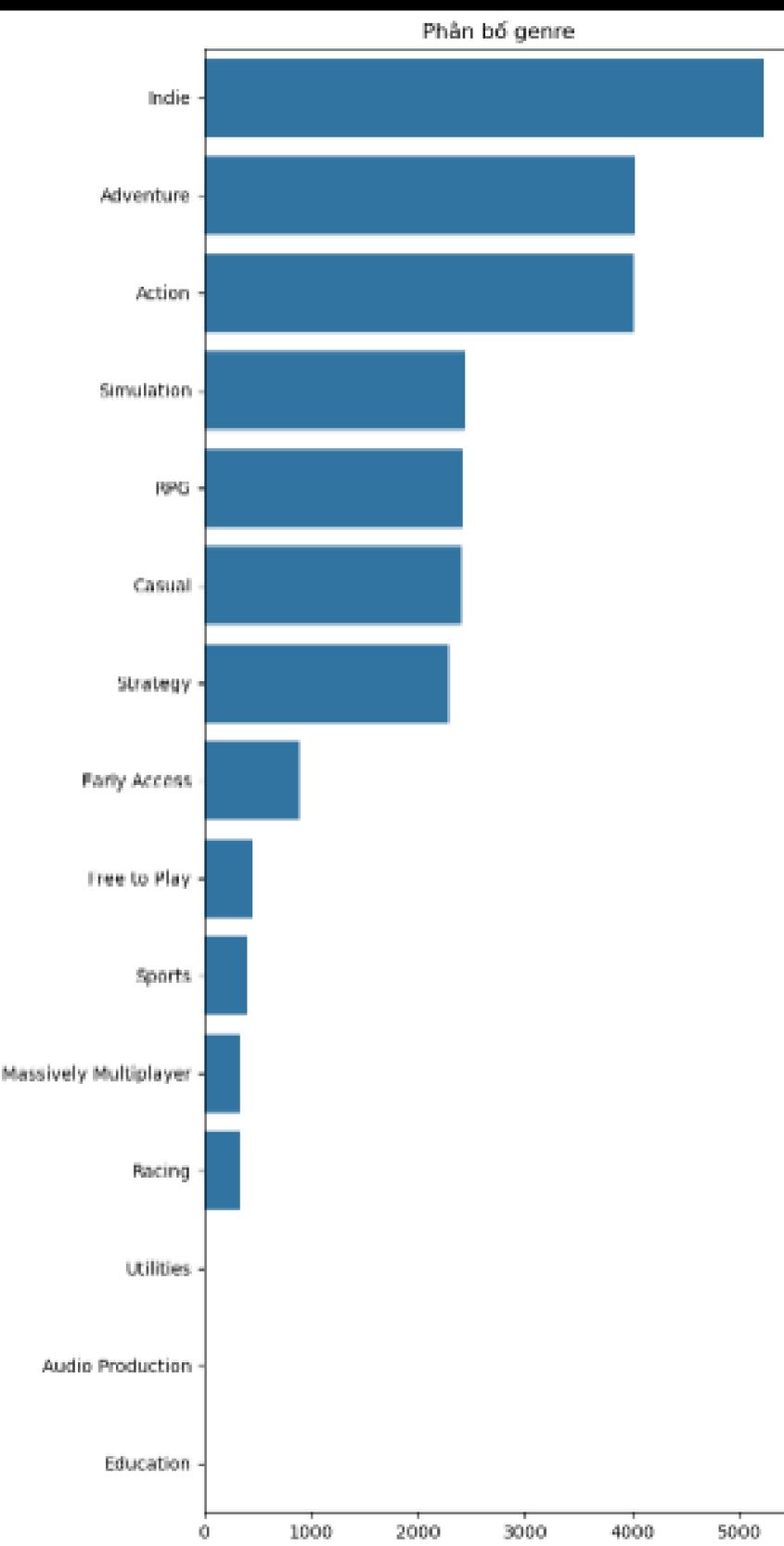
- Genre
- Tags
- Developer
- Publisher
- Languages

MENU



TIỀN XỬ LÍ

➔ Phân tích các cột Genre và Tags



[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)

MENU



FEATURE ENGINEERING



Phân tích cột datetime

Quá trình chuyển đổi thông tin thời gian thành dạng số để máy tính có thể hiểu được và sử dụng trong các mô hình học máy. Số hóa các cột datetime là vì:

- Thuận tiện cho mô hình
- Tạo thêm feature mới (Feature Engineering)
- Tính Liên tục

MENU



FEATURE ENGINEERING



Tạo cột tính điểm cho dữ liệu

Sử dụng công thức tính điểm dựa trên trang web chính thức của SteamDB:

$$\text{Rating} = \text{ReviewScore} - (\text{ReviewScore} - 0.5) * 2^{-\log_{10}(\text{TotalReviews}+1)}$$

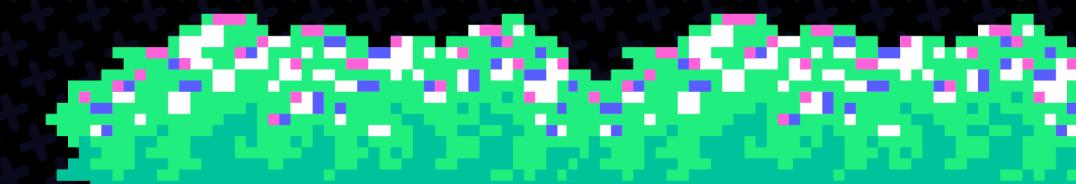
Trong đó:

- Total Reviews = Positive Reviews + Negative Reviews (tổng đánh giá của game)
- Review Score = Positive Reviews / Total Reviews (tỉ lệ đánh giá tích cực / tổng đánh giá)

SIGN IN

DATA EXPLORATORY & QUESTIONS

[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)

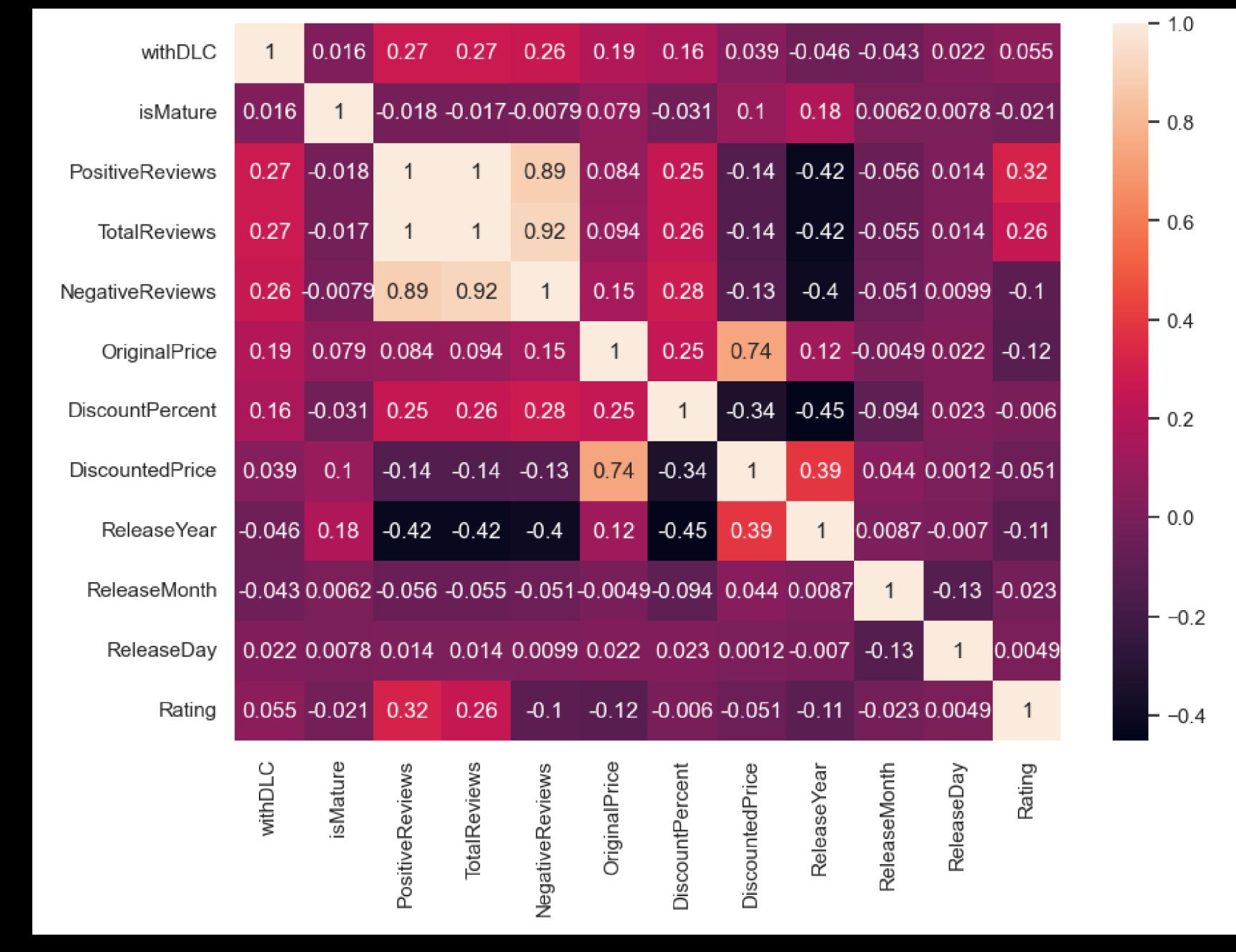


MENU



DATA EXPLORATORY

➡ Kiểm tra tính tương quan dữ liệu



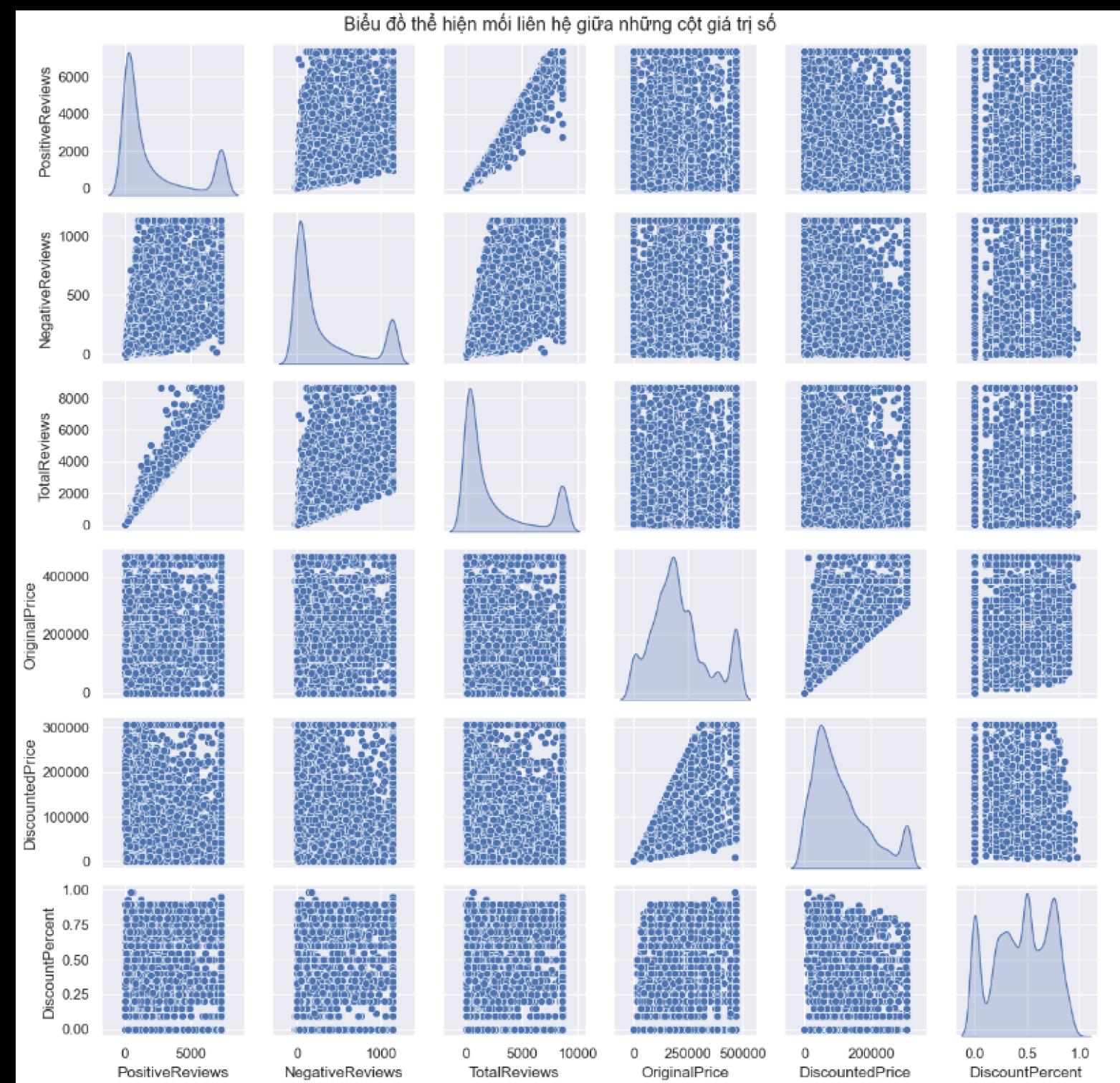
[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)

MENU



DATA EXPLORATORY

➡ Kiểm tra mối liên hệ giữa các cột số



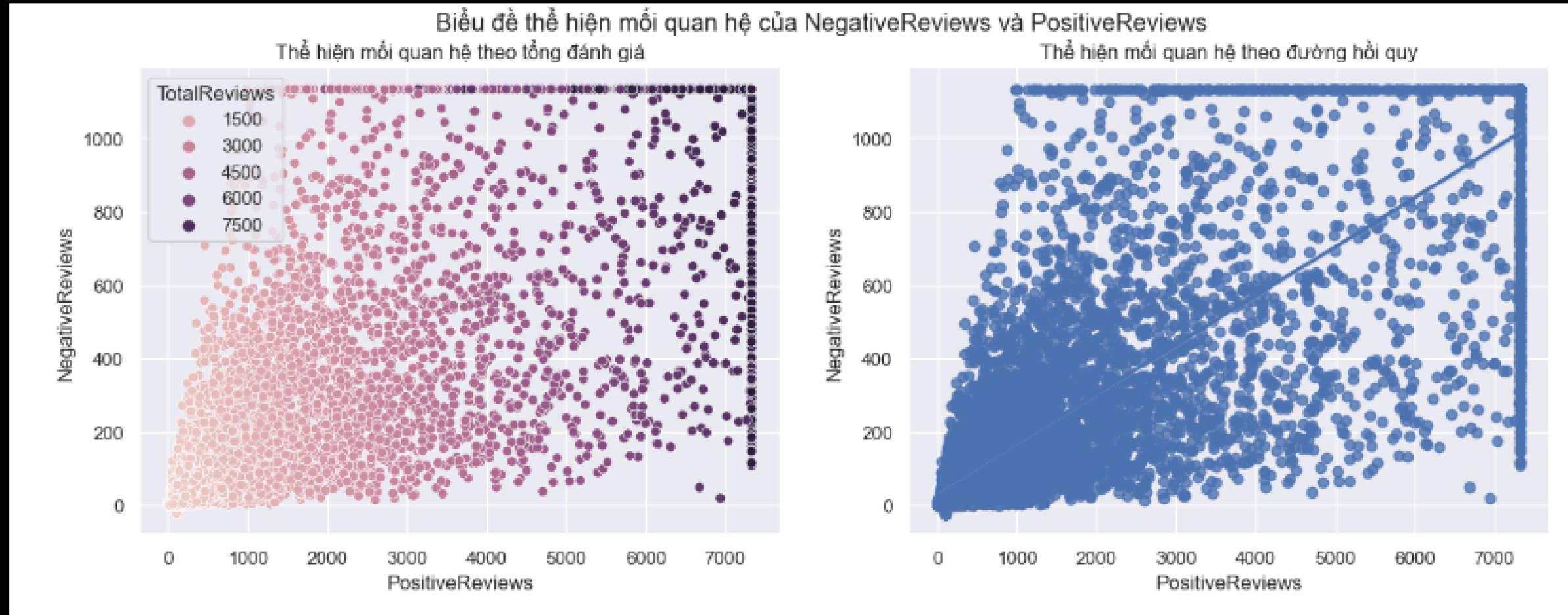
[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)

MENU



DATA EXPLORATORY

➔ PositiveReviews và NegativeReviews tăng giảm như nào dựa vào số lượng TotalReviews?



Khi TotalReviews ít thì các điểm dữ liệu giữa NegativeReviews và PositiveReviews nằm sát và chỉ chít lên nhau, chứng tỏ trong khoảng này nó có mối tương quan mạnh.

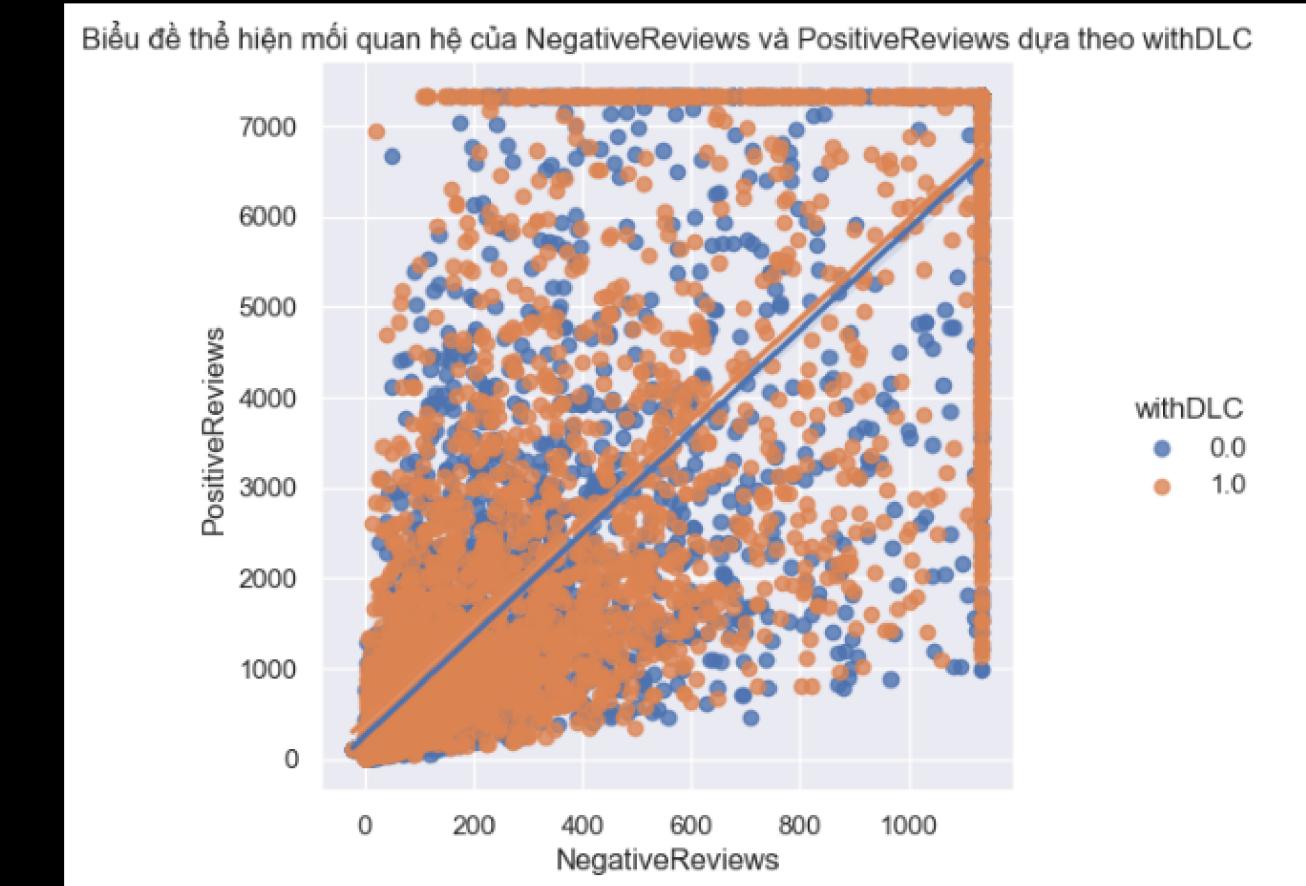
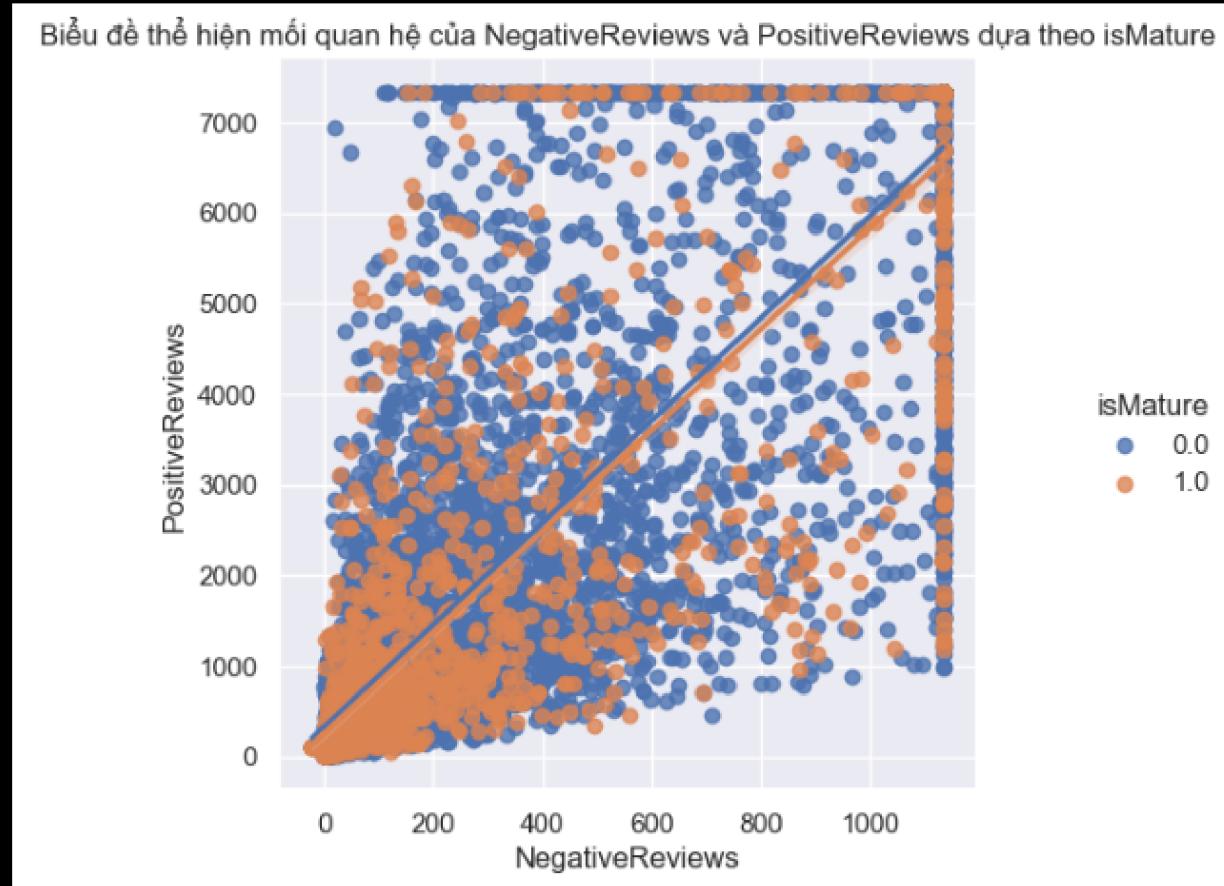
Khi TotalReviews ngày càng lớn dần, thì khoảng cách các điểm dữ liệu ngày càng nầm rời rạc và cách xa đường hồi quy, lượng PositiveReviews ngày càng tăng trong khi NegativeReviews ngày càng giảm.

MENU



DATA EXPLORATORY

➔ Mỗi quan hệ giữa PositiveReviews và NegativeReviews dựa theo isMature và withDLC



Ta thấy hai đường hồi quy phân loại theo withDLC và isMature trên mỗi hình sát nhau và gần như là một, do đó những game có DLC hay game có Mature content không ảnh hưởng quá nhiều khi người chơi đánh giá game.

MENU



QUESTIONS

QUESTIONS 1: Đối với những game có DLC thì giá tiền gốc có nhiều hơn những game không kèm theo DLC hay không và Original Price cao thấp có mối quan hệ như thế nào với PositiveReviews?

Bước 1: Đầu tiên, ta chia thành hai df (withDLC, withoutDLC)

```
df_with_dlc = cleaned_df[cleaned_df['withDLC'] == 1]
```

```
df_without_dlc = cleaned_df[cleaned_df['withDLC'] == 0]
```

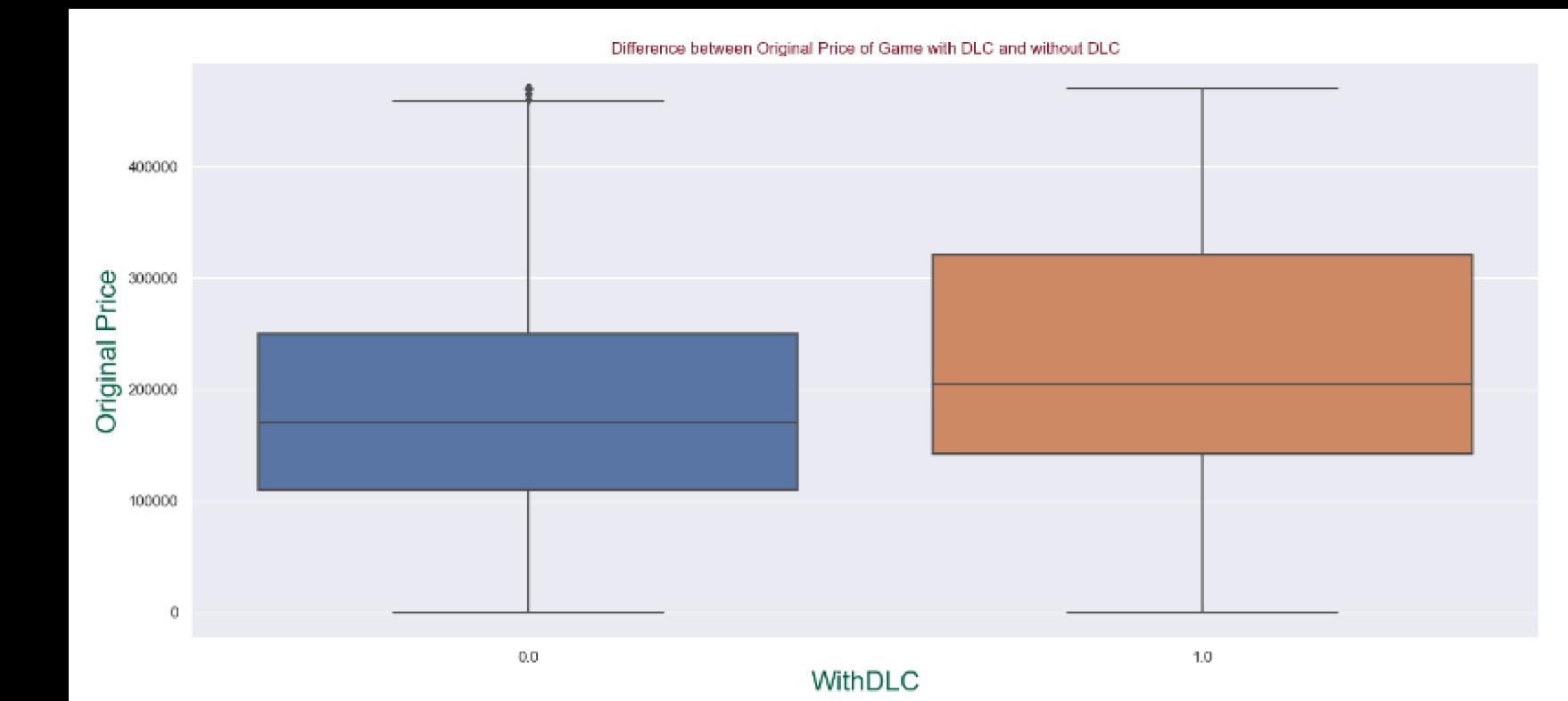
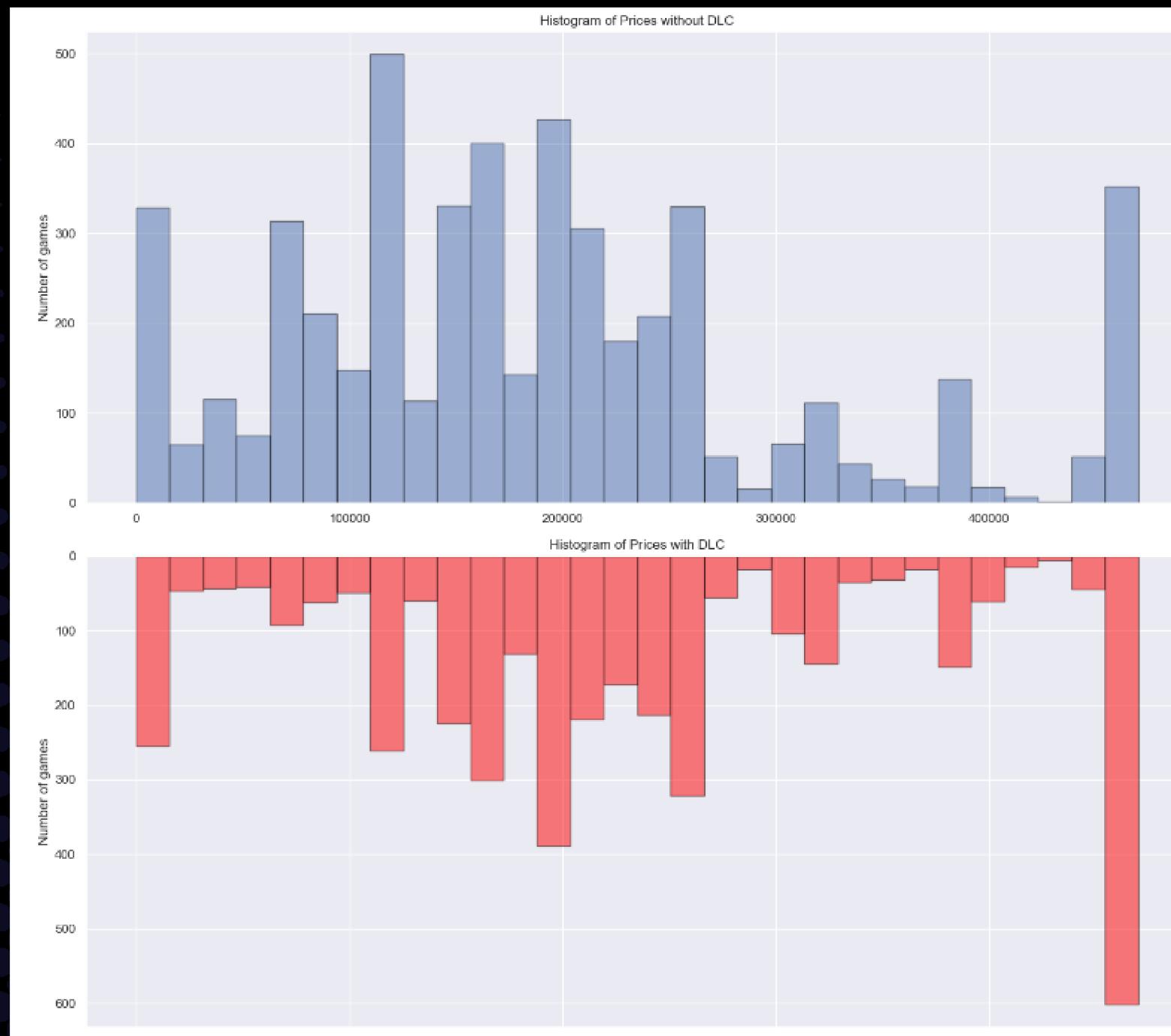
Bước 2: Vẽ histogram và boxplot

MENU



QUESTIONS

QUESTIONS 1: Đối với những game có DLC thì giá tiền gốc có nhiều hơn những game không kèm theo DLC hay không và Original Price cao thấp có mối quan hệ như thế nào với PositiveReviews?



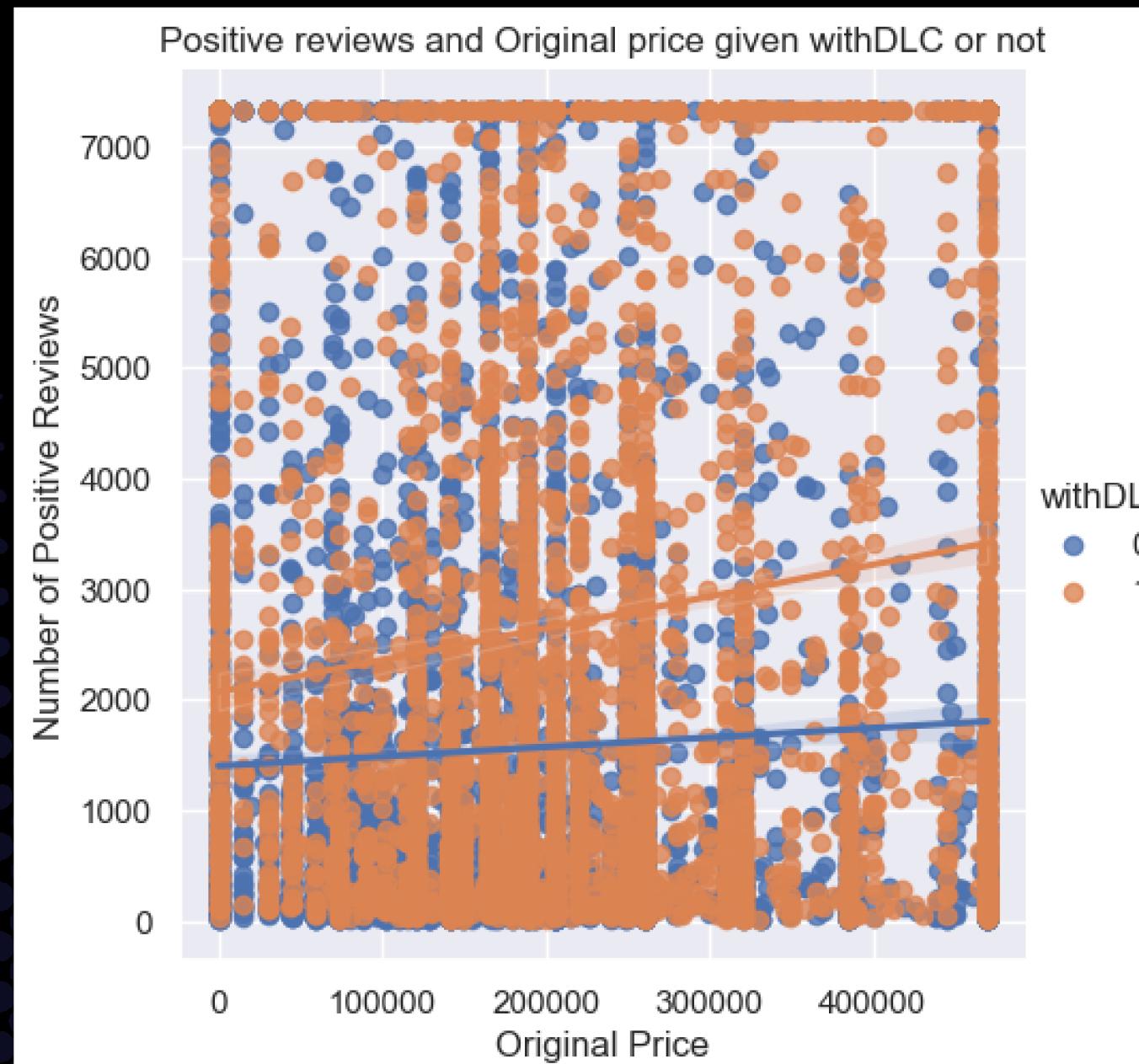
- Trong khoảng giá từ 0 đến 250.000 thì số lượng game có kèm DLC ít hơn nhiều các game không có kèm DLC
- Từ mốc giá 250.000 trở đi thì ngược lại, ta nhận thấy sự nổi bật hơn hẳn các game có kèm DLC với các game không có kèm DLC và sự khác biệt này càng rõ ràng khi giá tiền càng tăng

MENU



QUESTIONS

QUESTIONS 1: Đối với những game có DLC thì giá tiền gốc có nhiều hơn những game không kèm theo DLC hay không và Original Price cao thấp có mối quan hệ như thế nào với PositiveReviews?



- Giá tiền gốc của game có kèm theo downloadable content có xu hướng cao hơn những game không có mặc dù game không có DLC có số lượng nhiều hơn
- Sự không liên quan mật thiết gì giữa withDLC và PositiveReviews.

=> Các nhà phát hành game có thể phát triển game cùng với DLC để có thể tăng giá thành sản phẩm với mục đích thu lợi nhuận cao hơn mà không ảnh hưởng tiêu cực đến đánh giá của người dùng về game

MENU



QUESTIONS

QUESTIONS 2: Mỗi thể loại game thường được giảm giá như thế nào?

Bước 1: Đầu tiên, ta lấy cột Genre và DiscountPercent để lấy thông tin cần thiết cho câu hỏi:

```
genre_df = cleaned_df[['Genre', 'DiscountPercent']]
```

Bước 2: Vì Genre là một cột đặc biệt, vì mỗi sample có thể có nhiều thể loại, do đó để thể hiện rõ những thể loại nào đang phát triển, ta sẽ chia các Genre trong mỗi sample đó thành từng sample mới:

```
explode_genre_df = genre_df.assign(Genre=genre_df['Genre'].str.split(', ')).explode('Genre')
```

```
explode_genre_df = explode_genre_df.reset_index(drop=True)
```

Bước 3: Tính trung bình DiscountPercent của từng Genre:

```
mean_discount_df = explode_genre_df.groupby('Genre')['DiscountPercent'].mean().reset_index()
```

```
mean_discount_df = mean_discount_df[mean_discount_df['Genre'] != 'Free to Play']
```

```
mean_discount_df.sort_values(by=['DiscountPercent'], inplace=True)
```

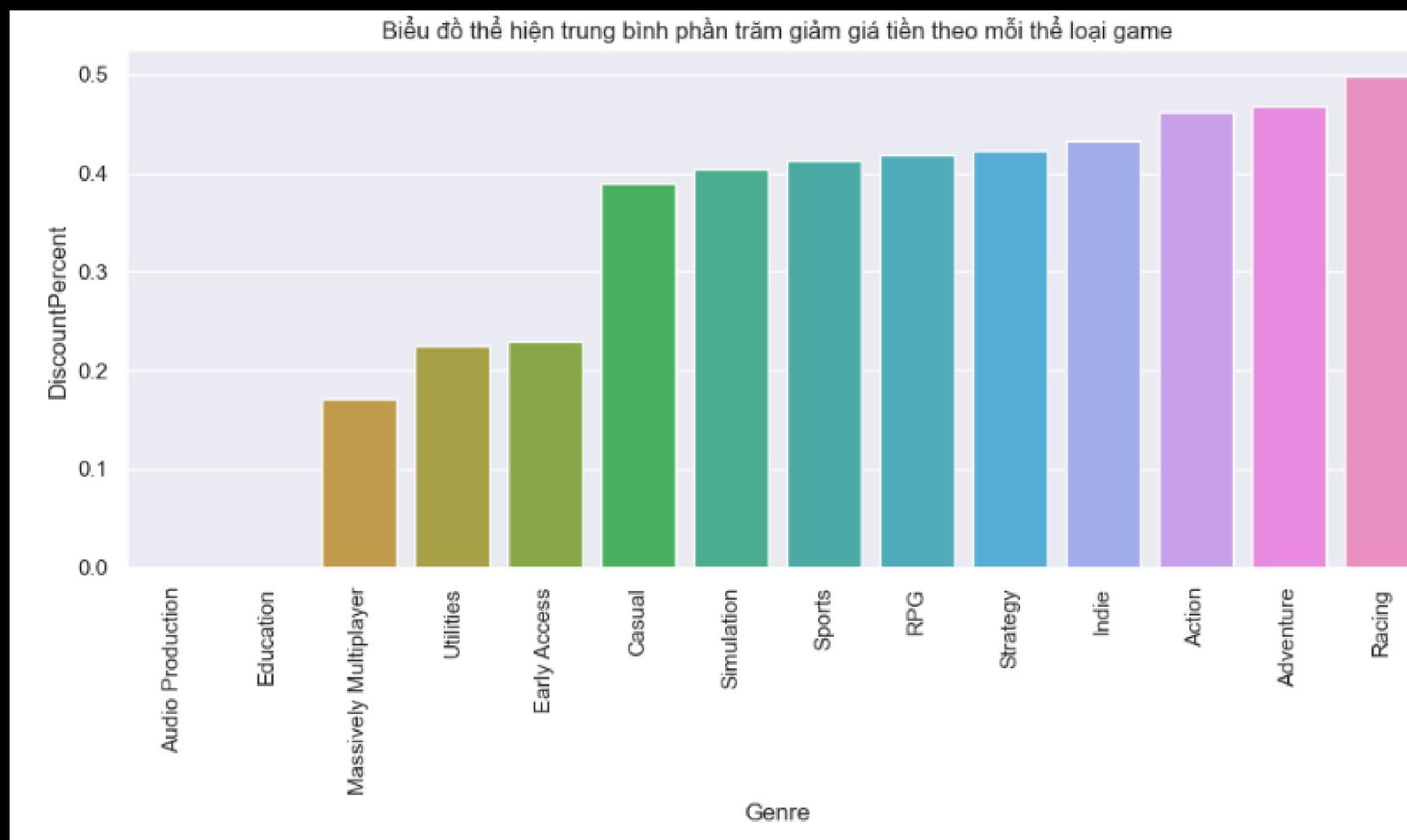
MENU



QUESTIONS

QUESTIONS 2: Mỗi thể loại game thường được giảm giá như thế nào?:

Trực quan hóa dữ liệu lên biểu đồ:



Giáo dục và Toolkit hỗ trợ phát triển game (Education, Utilities):

Discount thấp (dưới 20%).

Không phải là game giải trí đơn thuần, mang tính chất hỗ trợ và giáo dục.

Massive Multiplayer:

Discount thấp (xấp xỉ 20%).

Thường là game online đa người chơi, được nhiều game thủ đón nhận.

Có khả năng do tính cạnh tranh cao và sự thu hút của game đa người chơi, giảm giá không cao.

Giải trí cao (Various Entertainment Genres):

Discount trung bình lớn (khoảng 40%).

Đa dạng và là những thể loại được nhà phát triển tập trung nhiều nhất.

Có thể do sự cạnh tranh cao, mong muốn nhanh chóng tiếp cận người chơi, dẫn đến lượng discount cao.

Nhà phát triển nổi tiếng và lâu đời:

Có khả năng giữ được discount trung bình lớn.

Đối mặt với áp lực cạnh tranh từ những game mới nổi.

Kết luận:

Game giải trí thuần túy có discount trung bình khá cân bằng, không lệch nhiều.

Giảm giá không ảnh hưởng quá nhiều đến tính tồn tại và xu hướng phát triển của thể loại game trong tương lai.

Đánh giá mang tính chủ quan và không áp dụng hoàn toàn cho mọi trường hợp.

MENU



QUESTIONS

QUESTIONS 3: Những game được nhiều sự quan tâm đánh giá từ người chơi, có khuyến mãi như thế nào?

Ở câu hỏi 2, ta đã tìm hiểu mỗi thể loại game giảm giá nhiều hay ít ảnh hưởng như thế nào đến xu hướng phát triển của game đó. Vậy những game có khuyến mãi lớn có được người dùng quan tâm nhiều hay không?

Phân loại đánh giá theo chuẩn của Steam system:

0% - 19%: Negative

20% - 39%: Mostly Negative

40% - 69%: Mixed

70% - 79%: Mostly Positive

80% - 100%: Positive

MENU



QUESTIONS

QUESTIONS 3: Những game được nhiều sự quan tâm đánh giá từ người chơi, có khuyến mãi như thế nào?

Do data cào vè khá chênh lệch về số lượng game được làm theo từng năm, nên ta sẽ lấy 2 năm gần nhất là 2022 và 2023 để đối chiếu và so sánh.

Bước 1: Lựa chọn các cột cần thiết để trả lời câu hỏi (**Rating**, **DiscountPercent**)

Bước 2: Tạo một cột **Conclusion** lưu kết luận về đánh giá của game đó dựa vào phân loại đánh giá

Bước 3: Ta tính giá trị trung bình của DiscountPercent cho năm 2022 và 2023

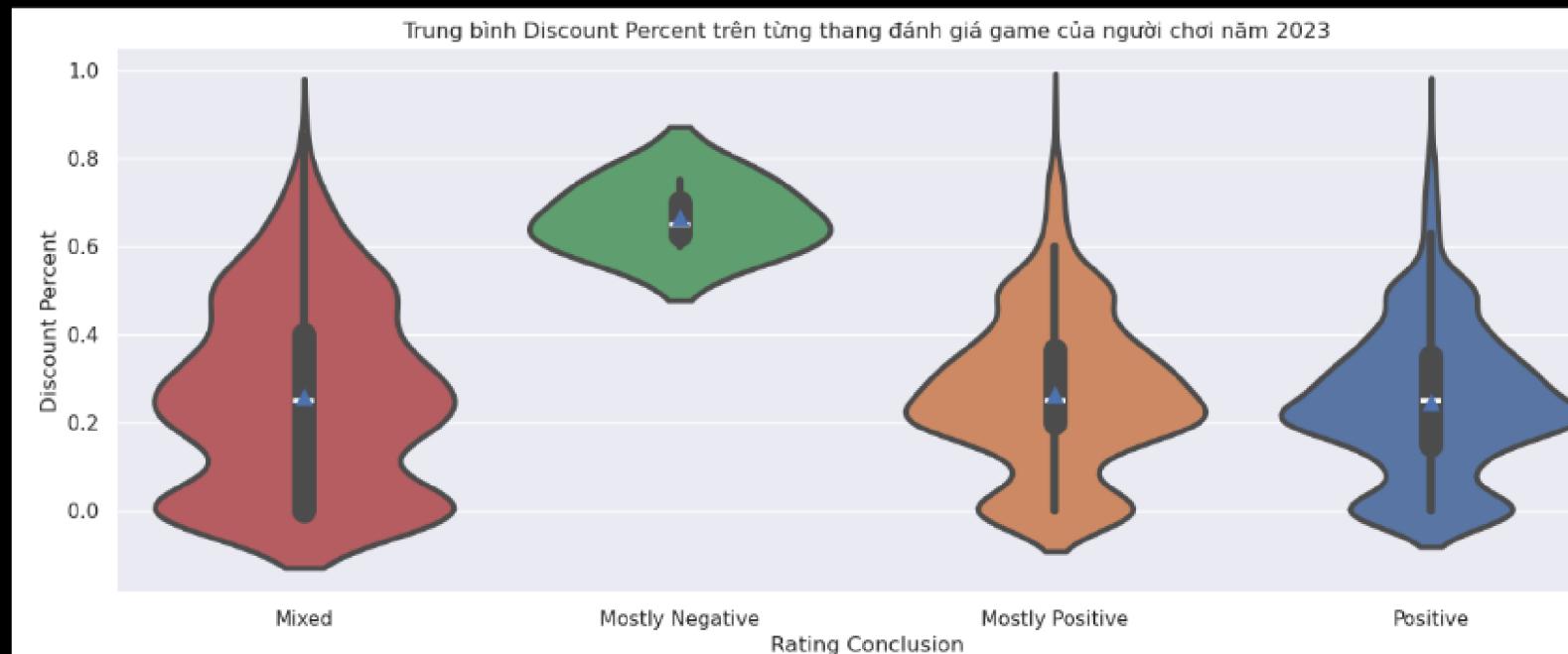
```
mean_2022 = discount_2022.groupby(['Conclusion'])['DiscountPercent'].mean()  
mean_2022 = mean_2022.to_frame().rename(columns= {'DiscountPercent': 'MeanDiscountPercent'}).reset_index()  
mean_2023 = discount_2023.groupby(['Conclusion'])['DiscountPercent'].mean()  
mean_2023 = mean_2023.to_frame().rename(columns= {'DiscountPercent': 'MeanDiscountPercent'}).reset_index()
```

MENU



QUESTIONS

QUESTIONS 3: Những game được nhiều sự quan tâm đánh giá từ người chơi, có khuyến mãi như thế nào?



Năm 2023:

- Giá trị trung bình DiscountPercent xấp xỉ 25% cho mỗi loại đánh giá.
- Mostly Negative có phân bố lợn cợn, tập trung ở khoảng discount 60%-80%.
- Sự tập trung của người chơi đánh giá nhiều nhất và giá trị trung bình discount tương đồng.

Năm 2022:

- Giá trị trung bình DiscountPercent khoảng 40%, trùng với median của từng loại đánh giá.
- Phân bố đánh giá của người chơi đa dạng hơn, không tập trung vào một số DiscountPercent cụ thể.
- Mostly Negative có phân bố lợn cợn, tập trung ở khoảng discount 50%.

Lí do Mostly Negative lợn cợn trong cả 2 năm:

- Có thể do lượng dữ liệu của các game được đánh giá Mostly Negative ít hơn so với các loại đánh giá khác.
- Tạm thời xét những đánh giá trung bình trở lên để có cái nhìn tích cực.

Dự đoán:

- Sự khác biệt trong phân bố của từng mức đánh giá giữa năm 2023 và 2022 có thể giúp nhà phát triển xây dựng chiến lược giảm giá phù hợp.
- Có thể xác định vùng giá trung bình nào thu hút nhiều người chơi và đồng thời duy trì lợi nhuận cho game.

MENU



QUESTIONS

QUESTIONS 4: Thể loại game nào được các nhà phát triển ưu chuộng và mang lại Rating cao?

Bước 1: Tách từng thể loại game và sử dụng explode() để chuyển thành từng dòng với từng thể loại.

```
copy_df['Genre'] = copy_df['Genre'].str.split(', ')
```

```
genre_counts = copy_df.explode('Genre')['Genre'].value_counts()
```

```
average_ratings_genre = copy_df.explode('Genre').groupby('Genre')['Rating'].mean()
```

Bước 2: Tính trung bình rating từng thể loại.

```
average_ratings_genre = copy_df.explode('Genre').groupby('Genre')['Rating'].mean()
```

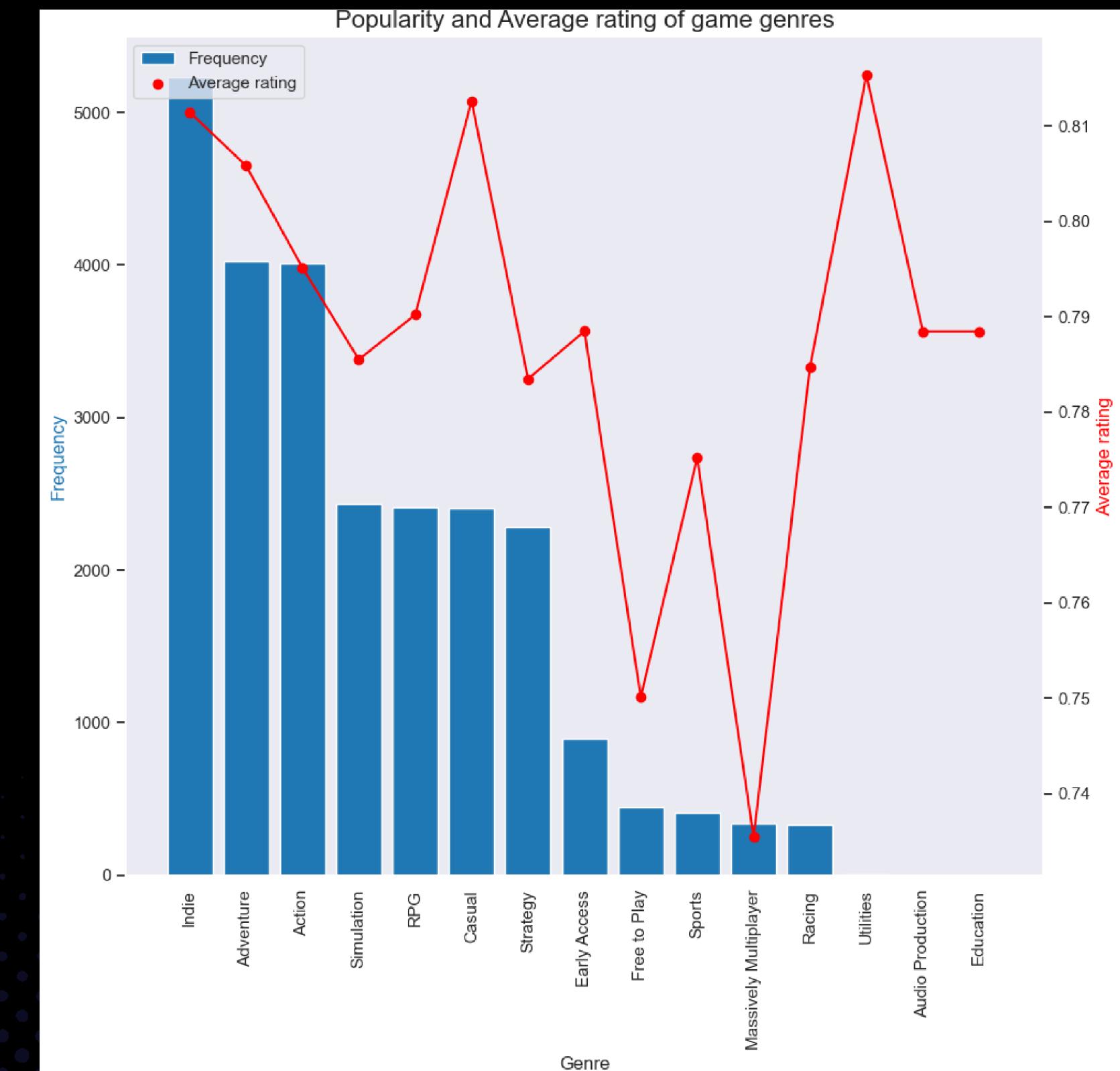
MENU



QUESTIONS

QUESTIONS 4: Thể loại game nào được các nhà phát triển ưa chuộng và mang lại Rating cao?

- Thể loại Indie rất phổ biến và có xếp hạng trung bình cao là 0,81. Adventure, Action cũng được ưa chuộng với đánh giá tích cực.
- Massively Multiplayer ít phổ biến hơn và có xếp hạng trung bình thấp hơn là 0,735.
- Casual, RPG, Sports, Racing đều có xếp hạng tích cực và tương đối phổ biến.
- Các game Utilities, Audio Production và Education được đánh giá cao nhưng chưa được phát triển rộng rãi.
- Các nhà phát triển, phát hành game nên đẩy mạnh gia công các game thuộc thể loại Massively Multiplayer
- Các thể loại game như Audio Production, Education cần được chú trọng hơn khi mang lại nhiều phản hồi tích cực, đây có thể là tiềm năng có thể khai thác để nâng cao lợi nhuận đáng kể



MENU



QUESTIONS

QUESTIONS 5: Thị trường phát triển game tiềm năng trong những năm gần đây?

Bước 1: Tách từng ngôn ngữ và sử dụng explode() để chuyển thành từng dòng.

```
df_copy['Languages'] = df_copy['Languages'].str.split(', ')
data_exploded = df_copy.explode('Languages')
```

Bước 2: Gộp các game theo từng ngôn ngữ và đếm dựa trên số lượng TotalReview

Vì không có trường ‘Số lượng tải’ nên ta dựa vào tổng Review để phân tích một cách tương đối

```
lang_review_df = data_exploded.groupby('Languages')['TotalReviews'].sum()
```

MENU



QUESTIONS

QUESTIONS 5: Thị trường phát triển game tiềm năng trong những năm gần đây?



♦ Xem xét sự phân bố ngôn ngữ

Nhận thấy ngôn ngữ English là ngôn ngữ phổ biến, nếu ta đánh giá thị trường dựa trên ngôn ngữ game hỗ trợ chỉ đúng một cách tương đối. Vì vậy ta xem xét thử nếu 'English' xuất hiện ở tất cả các game, ta sẽ loại ra để có cái nhìn khách quan hơn.

MENU



QUESTIONS

QUESTIONS 5: Thị trường phát triển game tiềm năng trong những năm gần đây?

Languages	French	German	Italian	Japanese	Korean	Portuguese - Brazil	Russian	Simplified Chinese	Spanish - Spain
Release Year									
2017	1376751.0	1364842.0	1109393.0	1126290.0	795820.0	920303.0	1202959.0	1097694.0	1298717.0
2018	1356440.0	1374569.0	1056818.0	1212342.0	965788.0	961112.0	1258591.0	1375362.0	1273457.0
2019	1477097.0	1446988.0	1017363.0	1411214.0	1092173.0	1078032.0	1338698.0	1513451.0	1380774.0
2020	1664230.0	1631824.0	1267389.0	1591162.0	1275277.0	1217418.0	1498210.0	1774130.0	1488875.0
2021	1840388.0	1837308.0	1342580.0	1777277.0	1475429.0	1435718.0	1693168.0	2012130.0	1650419.0
2022	1599787.0	1631390.0	1228257.0	1533338.0	1294873.0	1140198.0	1356542.0	1741597.0	1455878.0
2023	1583463.0	1605980.0	1177595.0	1527569.0	1293649.0	1299165.0	1254018.0	1726551.0	1442680.0

Bước 2:
Lọc lấy top 10
ngôn ngữ xuất
hiện nhiều nhất
và tổng số
TotalReview các
game của từng
ngôn ngữ.

MENU



QUESTIONS

QUESTIONS 5: Thị trường phát triển game tiềm năng trong những năm gần đây?



- Thị trường game của top 10 ngôn ngữ đều phát triển trong những năm gần đây, đạt đỉnh vào năm 2021, giảm nhẹ trong giai đoạn 2021-2022 và phát triển đều từ 2022-2023.
- Ngôn ngữ Trung giản thể dẫn đầu từ năm 2019.
- Ngôn ngữ Nhật, Đức, Pháp cũng phát triển mạnh. Tuy nhiên, số lượng game hỗ trợ tiếng Nga giảm trong giai đoạn 2022-2023.
- Ngôn ngữ Bồ Đào Nha tăng mạnh, cho thấy thị trường game Brazil có tiềm năng lớn.

SIGN IN

★FEATURE ENGINEERING★

SKLEARN

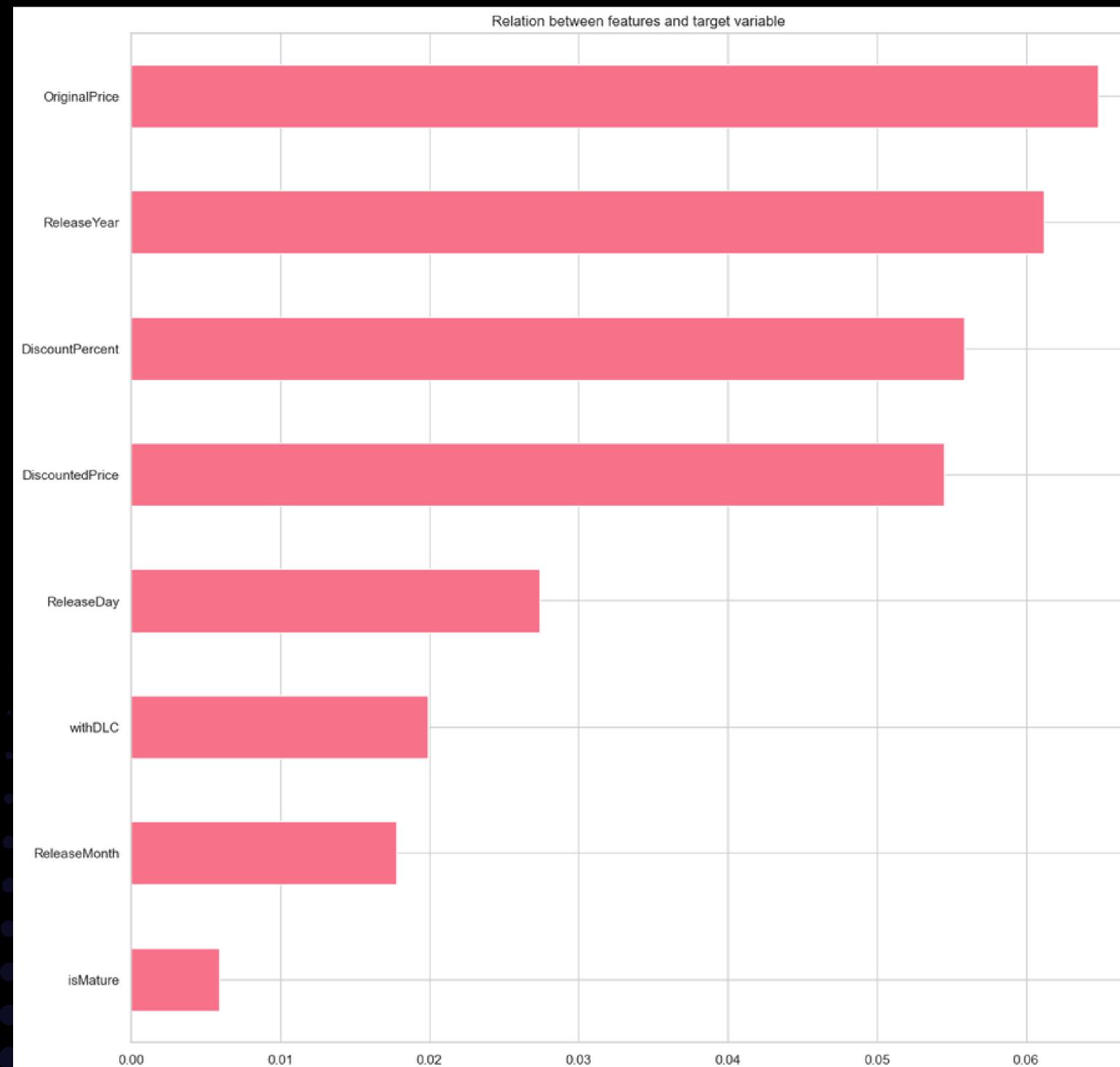
[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)



MENU



MUTUAL INFORMATION



➡ Để có thể nhìn nhận rõ mối quan hệ giữa các cột lý tưởng để làm feature cho model và cột target Rating

Ta nhận thấy cột 'Title' chỉ đơn giản là các giá trị thể hiện tên của các trò chơi điện tử trong tập dữ liệu, ta có thể suy đoán nó không mang lại ý nghĩa hay mối quan hệ gì với cột target Rating

MENU



CATEGORICAL FEATURES



Các cột: Genre, Developer,
Publisher, Languages

Genre, Languages: One Hot Encoding.

Developer, Publisher: TargetEncoder. Lý do không sử dụng One hot encoding do lo ngại việc quá nhiều cột ảnh hưởng đến hiệu năng của mô hình.

MENU



NUMERICAL FEATURES



Các cột: withDLC, isMature, PositiveReviews, TotalReviews, NegativeReviews, OriginalPrice, DiscountPercent, DiscountedPrice, ReleaseYear, ReleaseMonth, ReleaseDay

MinMaxScaler

(scale các giá trị về miền giá trị trong khoảng 0 đến 1)

SIGN IN



[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)

DATA MODELING

SKLEARN



PLAYER 1



ĐẶT VĂN ĐỀ

MỤC TIÊU CỦA MÔ HÌNH ĐƯỢC PHÁT TRIỂN SẼ GIÚP ĐÁNH GIÁ XẾP HẠNG ($0 \rightarrow 1$) CỦA CÁC GAME NHẰM:

- GIÚP NGƯỜI CHƠI CÓ CÁI NHÌN TRƯỚC VỀ CHẤT LƯỢNG CỦA MỘT TRÒ CHƠI MỚI DỰA TRÊN CÁC YẾU TỐ NHƯ THỂ LOẠI, NHÀ PHÁT TRIỂN, VÀ CÁC YẾU TỐ KHÁC.
- GIÚP NHÀ PHÁT TRIỂN HIỂU ĐƯỢC PHẢN HỒI TIỀM NĂNG TỪ CỘNG ĐỒNG NGƯỜI CHƠI, GIÚP HỌ CẢI THIỆN TRÒ CHƠI HOẶC ĐIỀU CHỈNH CHIẾN LƯỢC TIẾP THỊ.

♦ DỰ ĐOÁN SCORE RATING CỦA MỘT GAME VỚI THÔNG TIN ĐẦU VÀO CỤ THỂ

[BACK TO AGENDA PAGE](#)

🗡️ 01 ⚪ 07 ⭐ 12



INPUT



CÁC FEATURE: GENRE, DEVELOPER,
PUBLISHER, LANGUAGES, WITHDLC, ISMATURE,
ORIGINALPRICE, DISCOUNTPERCENT,
DISCOUNTEDPRICE, RELEASEYEAR,
RELEASEMONTH, RELEASEDAY

OUTPUT



RATING.

MENU



DATA MODELING

◀ Tạo pipeline cho quá trình huấn luyện mô hình

Pipeline xử lý cho các cột category

```
target_encode_transformer = Pipeline(steps= [('TargetEncoder', TargetEncoder())])
onehot_encode_transformer = Pipeline( steps=[('OneHotEncoder', OneHotEncoder(handle_unknown='ignore'))])
```

Pipeline xử lý cho các cột numeric

```
numeric_transfomer = Pipeline(steps= [('scaler', MinMaxScaler())])
```

Pipeline cho quá trình tiền xử lí

```
preprocessor = ColumnTransformer(
    transformers=[('one_hot_transformer', onehot_encode_transformer, [cat_cols[0], cat_cols[3]]),
    target_encode_transformer, list(cat_cols[1:2])), ('numeric_transfomer', numeric_transfomer, list(numeric_cols)),],
    remainder='passthrough')
```

MENU



DATA MODELING



Cross-validation để đánh giá và so sánh mô hình

Ta so sánh hiệu năng từng mô hình dựa vào trung bình điểm lỗi MSE (mean squared error) trên từng fold của cross-validation

```
cv = KFold(n_splits = 5, shuffle = True, random_state = 100)
def Cross_val_scores(model, X_val, y_val):
    scores = -1 * cross_val_score(model, X_val, y_val, cv = cv, scoring = 'neg_mean_squared_error', error_score='raise')
    return scores.mean()
```

MENU



DATA MODELING



Lựa chọn những mô hình để so sánh và phân chia dữ liệu

- Linear Regression
- Decision Tree
- Random Forest
- Gradient Boosting

Lưu cột **Rating** (cột đích) vào dữ liệu cần tìm y và loại khỏi dữ liệu huấn luyện X

Chia tập dữ liệu thành các tập huấn luyện và các tập dự đoán
`X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 42)`

MENU



DATA MODELING

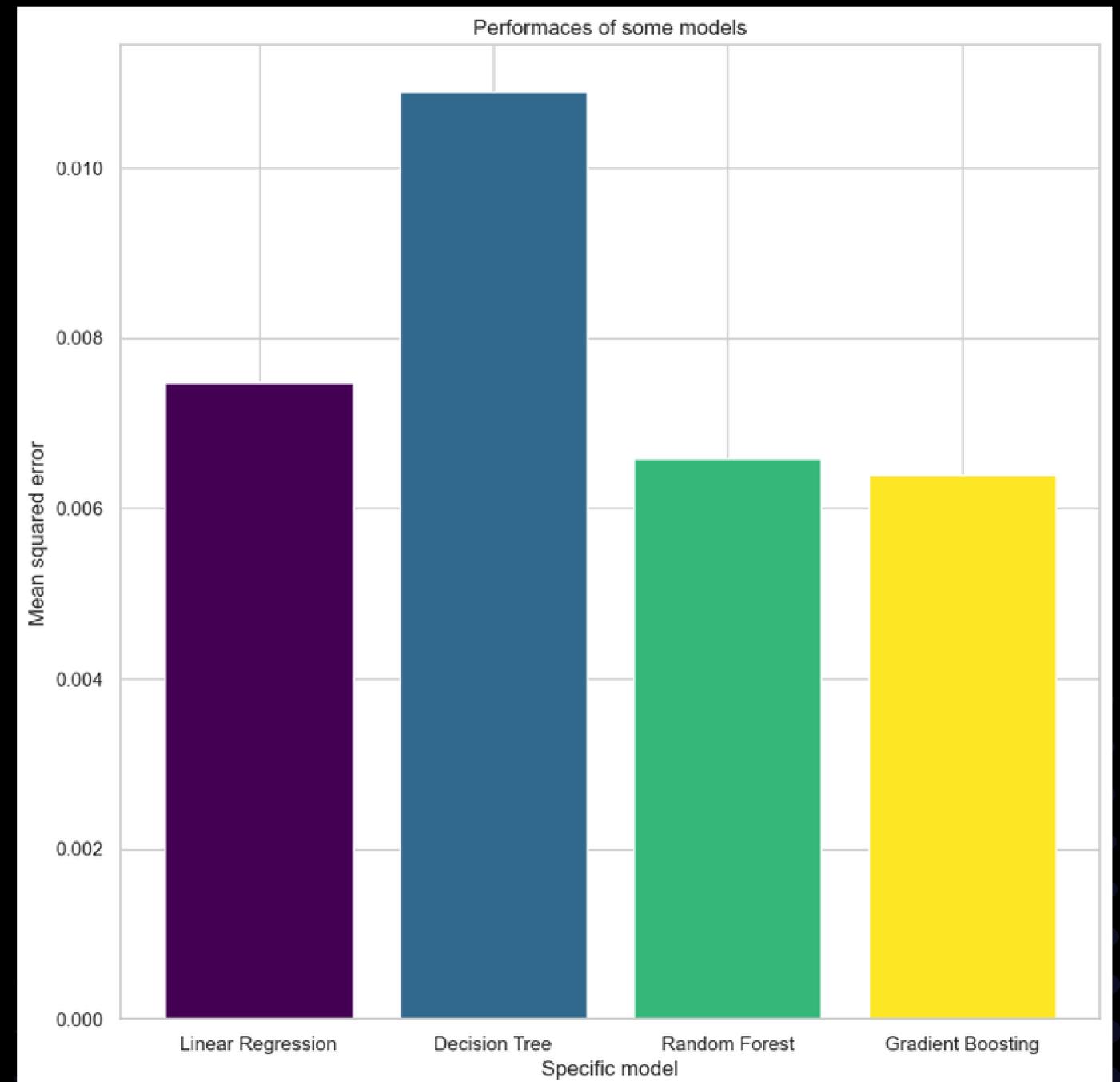


Kết quả huấn luyện dữ liệu sau khi thực hiện các quy trình

Linear Regression cross-validated score: 0.007475939091895395
Decision Tree cross-validated score: 0.010896071515283347
Random Forest cross-validated score: 0.006588402152583326
Gradient Boosting cross-validated score: 0.006392564481637252

Dựa vào biểu đồ, ta thấy MSE trên từng model khá là chênh lệch, với MSE cao nhất là `Decision Tree` với khoảng 0.01 (lệch nhất trong 4 mô hình), 3 mô hình còn lại thì cho ra kết quả tương đối ngang nhau và thấp nhất là `Gradient Boosting` với MSE hơn 0.006 một chút.

◆ Gradient Boosting





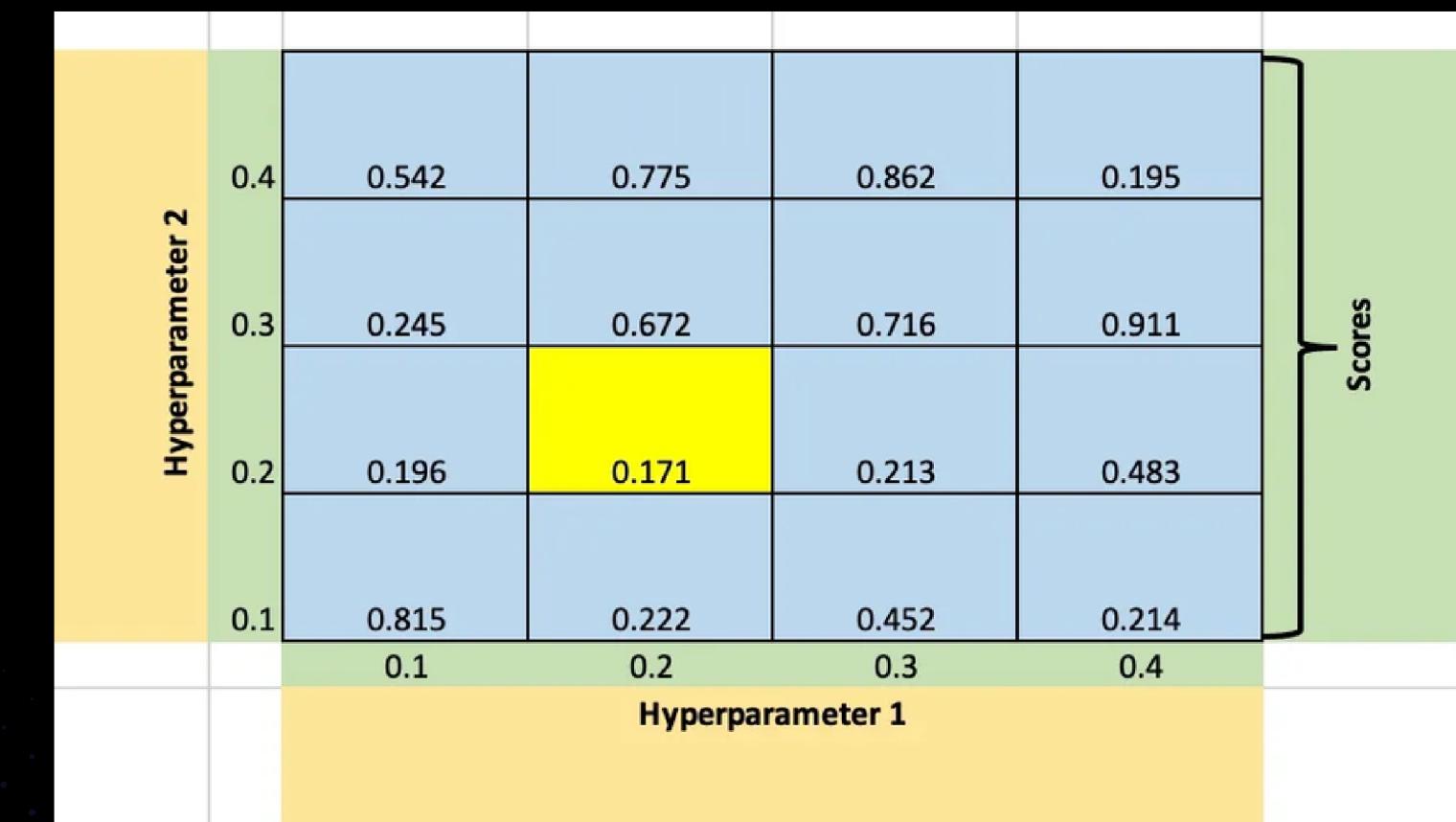
DATA MODELING



Fine tuning tìm ra bộ tham số tốt nhất cho model

GRIDSEARCHCV

- Mô hình sẽ tìm kiếm một kết hợp siêu tham số tốt nhất dựa trên Mọi phép thử trong không gian grid search.
- Tuy nhiên rất tốn kém về chi phí tính toán nhưng bù lại giá trị tìm được đảm bảo là tốt nhất trong không gian grid search.



MENU



FINE TUNING



Fine tuning tìm ra bộ tham số tốt nhất cho model

- Tạo lưới parameter để lấy mẫu
- n_estimators = [50, 100, 150]
max_depth = [3, 4, 5]
max_leaf_nodes = [None, 5, 10, 20]
learning_rate = [0.01, 0.1, 0.2]

- Tạo model sử dụng GradientBoostingRegressor
 - Sử dụng K-Fold Cross Validation với k=10
- GridSearchCV(estimator = xg, param_grid = random, cv = cv, n_jobs = -1)

MENU



FINE TUNING



Fine tuning tìm ra bộ tham số tốt nhất cho model

- Sau khi fit model, ta được tham số tốt nhất:
 - 'GradientBoostingRegressor__learning_rate': 0.1,
 - 'GradientBoostingRegressor__max_depth': 5,
 - 'GradientBoostingRegressor__max_leaf_nodes': None,
 - 'GradientBoostingRegressor__n_estimators': 150

Số điểm tốt nhất mà mô hình đạt được: 0.1564286139996527

MENU

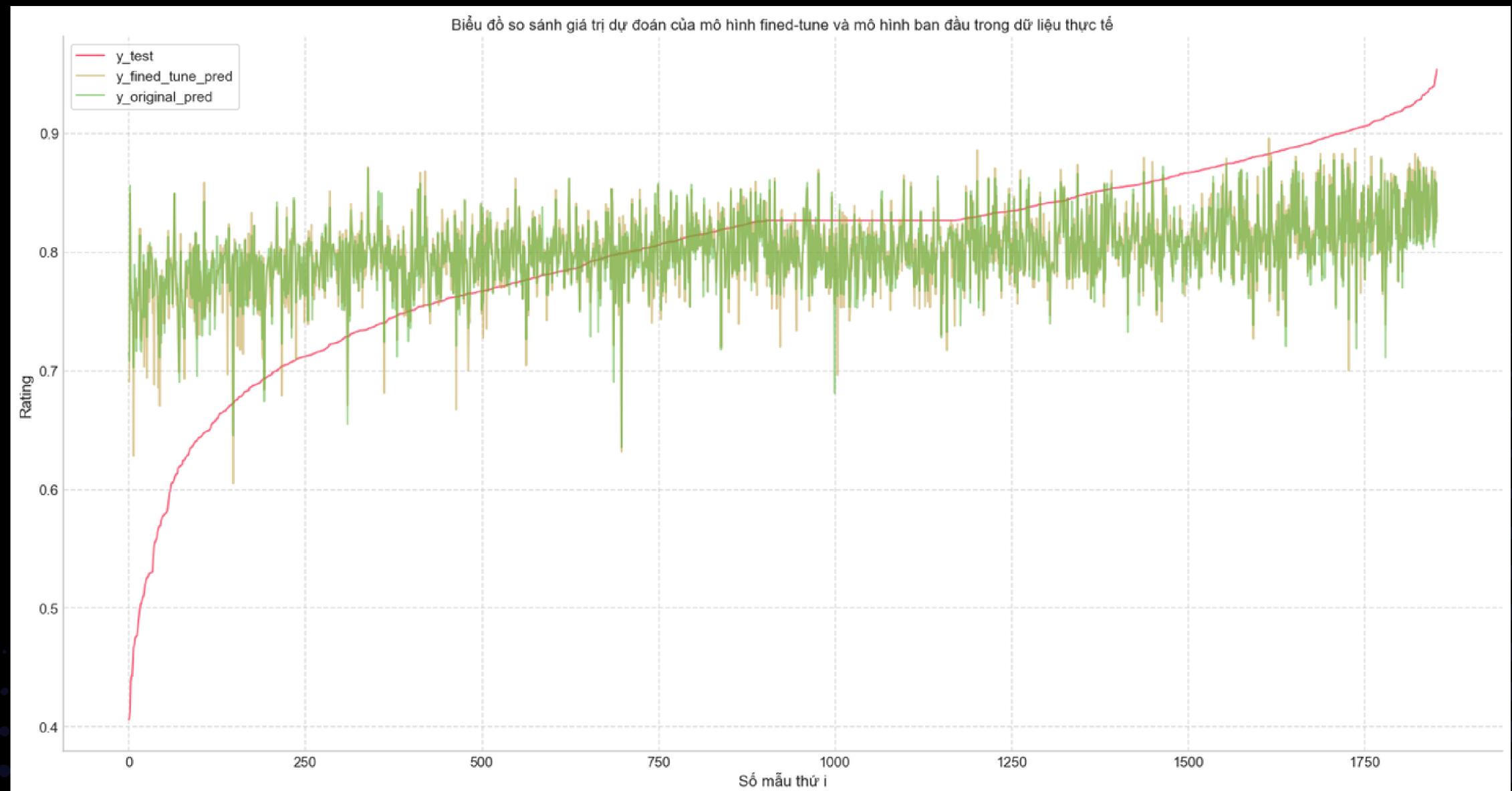


EVALUATION

- ❖ Với bộ tham số đã được tinh chỉnh, ta phỏng đoán độ chính xác của mô hình sau khi xây dựng.

Train Accuracy: 0.615
Test Accuracy: 0.211

Hiệu năng của mô hình:
Accuracy = 92.35%.



Biểu đồ so sánh giá trị dự đoán của mô hình fined-tune và mô hình ban đầu trong dữ liệu thực tế

SIGN IN

[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)

DEPLOY

TRIỂN KHAI MÔ HÌNH

FLASK

ENGINER



MENU



DEPLOY

Giao diện nhập thông tin của game đang muốn xem đánh giá

Gồm một số thông tin cần điền như sau:

- Genre
- isMature
- withDLC
- Publisher
- Developer
- OriginalPrice
- DiscountPercent
- Languages
- ReleaseYear
- ReleaseMonth
- ReleaseDay

Sau khi nhập đầy đủ thông tin, ta nhấn [PREDICT!](#)

LET'S RATE YOUR GAME!

Genre:	Action, Adventure, Indie	isMature:	0
withDLC:	1	Developer:	BANDAI
Publisher:	NAMCO	OriginalPrice:	90000
ReleaseYear:	2022	DiscountPercent:	0.5
ReleaseMonth:	3	Languages:	English
ReleaseDay:	12		

[Predict!](#)

STEAM®

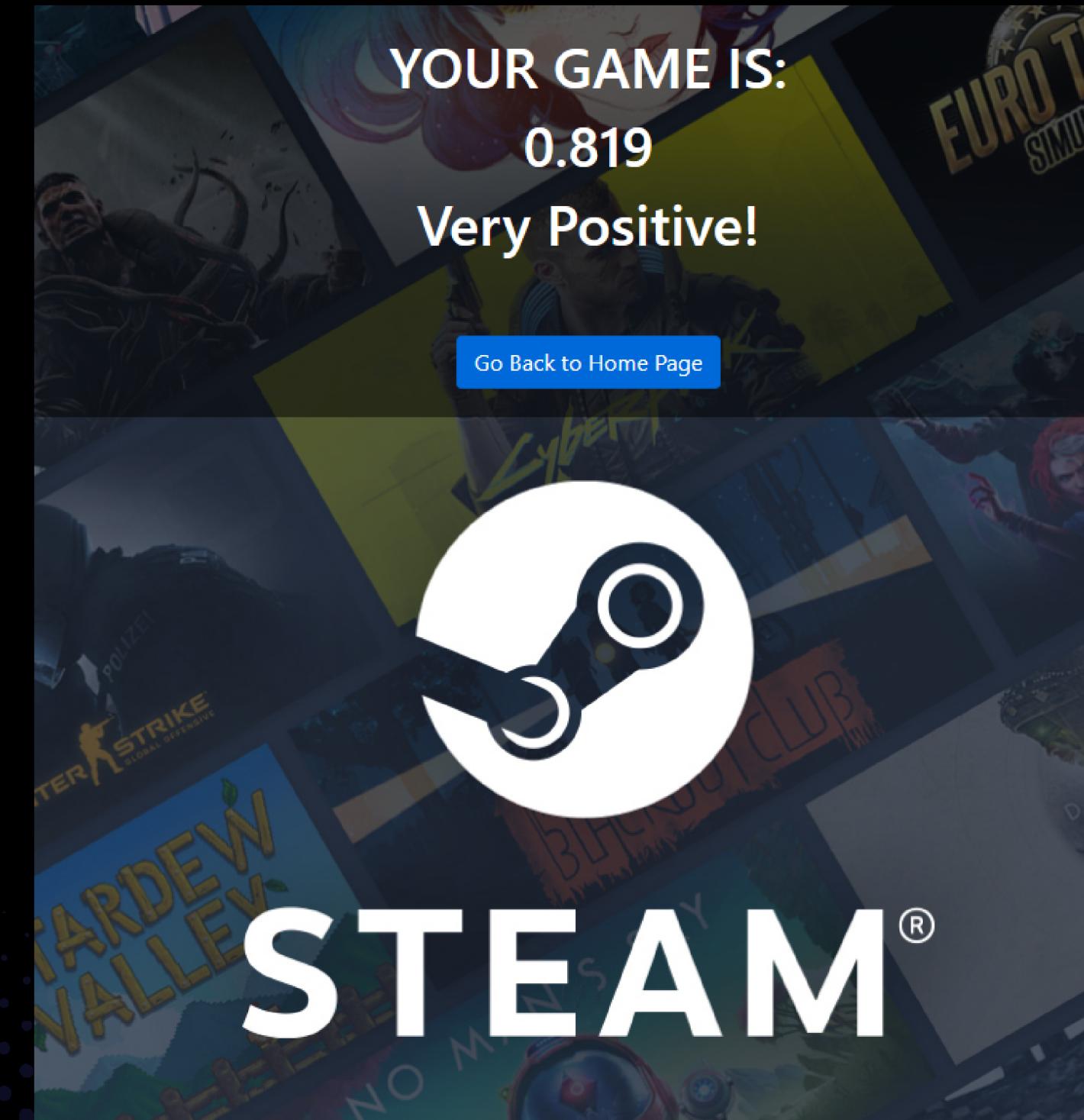
MENU



DEPLOY



Giao diện hiển thị kết quả chấm điểm:



SIGN IN



[QUAY LẠI TRANG CHƯƠNG TRÌNH](#)



TỰ ĐÁNH GIÁ



MENU



TỰ ĐÁNH GIÁ

👉 Từng thành viên đã gặp những khó khăn gì khi làm đồ án?

Khó khăn chung:

- Quá nhiều kiến thức mới phải tự học và nghiên cứu.
- Quản lý thời gian để vừa học và nghiên cứu cho đồ án, vừa ôn thi cuối kì và làm những đồ án môn khác.

Dù thời gian có hạn, nhưng đây là môn nền tảng quan trọng trong chuyên ngành và đi làm sau này, nên đây là đồ án bỏ nhiều thời gian và công sức nhất trong học kì, luôn cố gắng hoàn thiện đồ án tốt nhất có thể!

MENU



TỰ ĐÁNH GIÁ

➔ Từng thành viên đã gặp những khó khăn gì khi làm đồ án?

➔ Trần Đình Nhật Trí

- Khó khăn trong việc merge code trên jupyter notebook do vấn đề về phần mềm và môi trường
- Một số kĩ năng vẫn chưa thể làm tốt khi mô hình hóa dữ liệu (encoding,...)
- Đắn đo trong việc lựa chọn và đưa ra câu hỏi hay và nặng ý nghĩa trong đồ án, liệu câu hỏi này có phù hợp hay không?
- Khó khăn trong việc trình bày, mô tả quá trình và tổ chức code

➔ Nguyễn Thùy Uyên

- Cào dữ liệu mất nhiều thời gian và phải liên tục chỉnh sửa code do kết nối mạng và thay đổi bộ dữ liệu
- Chưa thể tìm hiểu kĩ và hiểu rõ về các kiến thức ML, DL
- Đắn đo trong việc lựa chọn và đưa ra câu hỏi
- Tinh chỉnh siêu tham số chưa phù hợp dẫn đến mất thời gian

➔ Nguyễn Ngọc Gia Minh

- Merge code dễ gây conflict vì tổ chức file không tương đồng, file notebook khó để merge
- Tìm hiểu và khai thác được các câu hỏi thật sự mang ý nghĩa (cần kết hợp các cột dữ liệu ra sao, lựa chọn biểu đồ như thế nào, các bước giải quyết, nhận xét và đưa ra kết luận)
- Cào dữ liệu tốn khá nhiều thời gian

MENU



TỰ ĐÁNH GIÁ

➡ Từng thành viên đã học được những gì?

Trần Đình Nhật Trí

- Cách xử lý conflict khi merge code.
- Rèn dũa kĩ năng thảo luận, đóng góp và hỗ trợ các thành viên trong nhóm.
- Quy trình làm một dự án dữ liệu hoàn chỉnh, rèn luyện được những kĩ năng cứng và mềm trong một dự án khoa học dữ liệu.
- Tự học và nghiên cứu, tìm cách đưa ra vấn đề và giải quyết.
- Học thêm được nhiều kiến thức mới lạ, thú vị trong chuyên ngành và biết cách triển khai một sản phẩm ra cho người dùng.

Nguyễn Thùy Uyên

- Kĩ năng sử dụng github, kĩ năng thảo luận và làm việc nhóm.
- Quản lí thời gian phù hợp và hiệu quả.
- Tự học hỏi các kiến thức mới, cách giải quyết vấn đề trong python (sử dụng các function...).
- Học được từ phần công việc của teammate (như Pipeline, deploy mô hình,...)
- Quy trình hoàn thiện một dự án Khoa Học Dữ Liệu.

Nguyễn Ngọc Gia Minh

- Các bước thực hiện một project dữ liệu hoàn chỉnh là như thế nào?
- Khả năng chịu đựng áp lực, bất đồng giữa thành viên trong nhóm
- Khả năng sử dụng Git, Github gia tăng
- Tự học được thêm nhiều thư viện, hàm hỗ trợ trong Python; các mô hình học máy, chọn lựa tham số tối ưu cho mô hình, triển khai mô hình
- Cách làm việc trong một đội nhóm thật sự

MENU



TỰ ĐÁNH GIÁ

➔ Nếu có nhiều thời gian hơn, nhóm sẽ:

Nhóm em sẽ thực hiện thêm một số thứ như sau:

- Tiến hành cào thêm các cột dữ liệu như đánh giá của từng người chơi game (số giờ chơi, bình luận về game, đề xuất game hay không, số người dùng thấy đánh giá hữu ích, vui vẻ,)
- Sentiment analysis trên tập dữ liệu đó, xây dựng NLP, Recommender System,
- Khai thác nhiều hơn thông tin của từng game cào được như System requirements (hệ điều hành tương thích, lưu trữ tối thiểu cần có, RAM,) để cụ thể hơn việc khai thác thị trường người chơi game
- Thủ cào dữ liệu bằng các cách mang lại hiệu suất cao hơn (ít tốn thời gian và tài nguyên máy tính hơn)
- Tìm hiểu thêm các mô hình của học máy dùng để xử lý text
- Tìm hiểu và cài đặt fine tuning cho mô hình được lựa chọn tốt hơn

CÔNG VIỆC CỤ THỂ

NGUYỄN THỦY UYÊN	NGUYỄN NGỌC GIA MINH	TRẦN ĐÌNH NHẬT TRÍ
Cào dữ liệu	Tiền xử lý	Tiền xử lý
Khai thác dữ liệu để đặt câu hỏi mang ý nghĩa	Khai thác dữ liệu để đặt câu hỏi mang ý nghĩa	Khai thác dữ liệu để đặt câu hỏi mang ý nghĩa
Fine tuning	Cài đặt code cho model selection, Cross validation	Feature engineering, tổ chức model dưới dạng Pipeline, deploy model
Chuẩn bị slide thuyết trình	Chuẩn bị slide thuyết trình	Chuẩn bị slide thuyết trình

[BACK TO AGENDA PAGE](#)

MENU



THANK YOU!