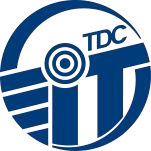
**TRƯỜNG CAO ĐẲNG CÔNG NGHỆ THỦ ĐỨC**

Khoa Công Nghệ Thông Tin

**BÁO CÁO HỌC PHẦN PYTHON NÂNG CAO**

**ĐỀ TÀI:**

**VIDEO GAMES GLOBAL SALES**

**Giảng viên hướng dẫn: PHAN THỊ THỂ**

**Sinh viên thực hiện: Võ Tấn Phúc**

***Thành phố Hồ Chí Minh, ngày17 tháng 12 năm 2023***

**MỤC LỤC**

[**Mục lục hình ảnh** 3](#_Toc153711533)

[**LỜI MỞ ĐẦU** 4](#_Toc153711534)

[**CHƯƠNG 1. GIỚI THIỆU VỀ THUẬT TOÁN LINEAR REGRESSION** 5](#_Toc153711535)

[**CHƯƠNG 2. CHI TIẾT ĐỒ ÁN** 6](#_Toc153711536)

[**2.1. Giới thiệu bài toán** 6](#_Toc153711537)

[**2.2. Vấn đề cần giải quyết** 6](#_Toc153711538)

[**2.3 Làm sạch dữ liệu** 7](#_Toc153711539)

[**2.4 Trực quan hóa dữ liệu** 9](#_Toc153711540)

[**2.5 Áp dụng thuật toán** 15](#_Toc153711541)

[**2.6 Xây dựng mô hình** 16](#_Toc153711542)

[**2.7 Đánh giá mô hình** 17](#_Toc153711543)

[**2.8 Đưa ra giải pháp** 17](#_Toc153711544)

[**2.9 Nhận Xét** 18](#_Toc153711545)

[**2.10 Demo chương trình:** 19](#_Toc153711546)

[**TÀI LIỆU THAM KHẢO** 29](#_Toc153711547)

# **Mục lục hình ảnh**

[Hinh 1 đọc dữ liệu 7](#_Toc153711471)

[Hinh 2 Thống kê mô tả của dữ liệu số 7](#_Toc153711472)

[Hinh 3 Thông tin và kiểu dữ liệu 8](#_Toc153711473)

[Hinh 4 Xử lí giá trị trùng lặp và giá trị bị thiếu 8](#_Toc153711474)

[Hinh 5 Biều đồ heatmap thể hiện tương quan giữa cá biến số 9](#_Toc153711475)

[Hinh 6 Biểu đồ barplot thể hiện tổng doanh số bán hàng toàn cầu 9](#_Toc153711476)

[Hinh 7 Biểu đồ scatter plot giữa điểm đánh giá và doanh số bán hàng 10](#_Toc153711477)

[Hinh 8 Biểu đồ countplot Xác định thể loại trò chơi phổ biến nhất 10](#_Toc153711478)

[Hinh 9 Biểu đồ boxplot doanh số bán hàng theo thể loại 11](#_Toc153711479)

[Hinh 10 Biểu đồ phân phối điểm đánh giá người chơi và chuyên gia 11](#_Toc153711480)

[Hinh 11 Top nhà phát triển theo doanh số bán hàng 12](#_Toc153711481)

[Hinh 12 Biểu đồ boxplot điểm đánh giá theo nhà phát triển 12](#_Toc153711482)

[Hinh 13 Tính tổng doanh số bán hàng theo khu vực 13](#_Toc153711483)

[Hinh 14 Biểu đồ pivot cho thể loại và khu vực 13](#_Toc153711484)

[Hinh 15 Biểu đồ scatter plot giữa điểm đánh giá chuyên gia và doanh số bán hàng theo thể loại 14](#_Toc153711485)

[Hinh 16 Biểu đồ scatter plot giữa điểm đánh giá chuyên gia và doanh số bán hàng theo thể loại 14](#_Toc153711486)

# **LỜI MỞ ĐẦU**

Trong bối cảnh ngày càng đa dạng và phát triển nhanh chóng của thế giới game, việc đánh giá và hiểu rõ về những tựa game được đánh giá cao không chỉ là quan trọng đối với người sáng tạo, mà còn là chìa khóa để định rõ hướng phát triển của ngành công nghiệp game.

Báo cáo này tập trung vào nghiên cứu và phân tích về những tựa game nổi bật nhất - những tác phẩm điện tử được đánh giá cao từ cộng đồng game thủ. Chúng tôi sẽ khai thác dữ liệu từ nhiều nguồn đáng tin cậy như Steam, Metacritic, hoặc các nền tảng game trực tuyến để tìm hiểu rõ hơn về những yếu tố nào góp phần vào việc một tựa game nhận được đánh giá cao.

Thông qua việc phân tích các yếu tố như thể loại game, đồ họa, ngày phát hành, số lượt đánh giá và xếp hạng trung bình, chúng tôi mong muốn đưa ra những cái nhìn sâu sắc về mối liên quan giữa các yếu tố này và đánh giá của cộng đồng game thủ. Báo cáo cũng sẽ đặt ra câu hỏi về các xu hướng và đặc điểm chung của những tựa game ấn tượng.

Mục tiêu cuối cùng của chúng tôi là cung cấp thông tin chi tiết và hữu ích, giúp các nhà phát triển game, nhà sản xuất, và cả những người yêu thích game hiểu rõ hơn về những yếu tố ảnh hưởng đến thành công và đánh giá cao của một tựa game trong cộng đồng game thủ đương đại.

# **CHƯƠNG 1. GIỚI THIỆU VỀ THUẬT TOÁN LINEAR REGRESSION**

Thuật toán Linear Regression là một phương pháp cơ bản trong học máy, tập trung vào việc xác định mối quan hệ tuyến tính giữa các biến. Mục tiêu chính của nó là dự đoán hoặc giải thích mức độ biến thiên của biến phụ thuộc dựa trên biến độc lập. Mô hình tuyến tính cố gắng tìm một đường thẳng hoặc siêu phẳng mà tối thiểu hóa sai số giữa dữ liệu thực và dự đoán, áp dụng trong cả mô hình đơn biến và đa biến.

Sử dụng Linear Regression, mô hình có thể biểu diễn mối quan hệ giữa các biến dưới dạng đồ thị đơn giản, giúp hiểu rõ hơn về tương quan giữa chúng. Việc đánh giá hiệu suất của mô hình dựa trên các chỉ số như R-Squared, Adjusted R-Squared, MSE và RMSE, giúp xác định độ chính xác và phù hợp của mô hình với dữ liệu.Thuật toán này dựa trên việc sử dụng dữ liệu huấn luyện để điều chỉnh các trọng số (weights) và hệ số (coefficients) của các biến độc lập để tạo ra một mô hình tối ưu. Một khi đã huấn luyện xong, mô hình Linear Regression có thể được sử dụng để dự đoán giá trị của biến phụ thuộc dựa trên các giá trị mới của các biến độc lập.

Linear Regression không chỉ được sử dụng trong lĩnh vực khoa học dữ liệu mà còn trong nhiều lĩnh vực khác như kinh tế học, y học, kỹ thuật và nhiều lĩnh vực khác với mục tiêu dự đoán và phân tích mối quan hệ giữa các biến.

# **CHƯƠNG 2. CHI TIẾT ĐỒ ÁN**

## **2.1. Giới thiệu bài toán**

Trong lĩnh vực phân tích dữ liệu và học máy, bài toán dự đoán Doanh Số Bán Hàng Toàn Cầu của các trò chơi là một thách thức hấp dẫn, nhằm xác định những trò chơi có doanh số bán hàng cao nhất dựa trên dữ liệu sẵn có. Mục tiêu chính của dự án là tìm ra các yếu tố quyết định sự thành công và doanh số bán hàng của một trò chơi.

Quá trình bắt đầu bằng việc thu thập dữ liệu từ nhiều nguồn, bao gồm các nền tảng game, cửa hàng trực tuyến, và thông tin về doanh số bán hàng từ cộng đồng game thủ. Dữ liệu này có thể chứa thông tin về tên trò chơi, đánh giá của người chơi, thể loại, nhà phát triển, năm phát hành, và nhiều thuộc tính khác.

Tiếp theo, thông qua việc xử lý và phân tích dữ liệu, các mô hình học máy như Linear Regression, Decision Trees, hoặc các mô hình Deep Learning có thể được áp dụng để dự đoán hoặc xếp hạng các trò chơi dựa trên các yếu tố như đánh giá trung bình, thể loại, và nhà phát triển.

Mục tiêu cuối cùng là tạo ra một mô hình có khả năng dự đoán chính xác doanh số bán hàng của một trò chơi dựa trên các thông tin có sẵn. Điều này sẽ giúp cả người chơi và nhà phát triển game có cái nhìn toàn diện về những trò chơi có tiềm năng cao trong việc đạt được doanh số bán hàng lớn. Bằng cách này, dự án đóng góp vào việc hiểu rõ hơn về các yếu tố quyết định sự thành công trong ngành công nghiệp game.

## **2.2. Vấn đề cần giải quyết**

Dự Đoán Doanh Số Bán Hàng Toàn Cầu của Video Games

Trong ngành công nghiệp game phức tạp, việc dự đoán doanh số bán hàng toàn cầu của một tựa game dựa trên các thông tin như mô tả, mức độ phổ biến, ngày phát hành và số lượt bình chọn là một thách thức hấp dẫn. Mục tiêu của dự án này là hiểu rõ hơn về cách những yếu tố này tương tác và ảnh hưởng đến doanh số bán hàng toàn cầu của một tựa game.

Bằng cách sử dụng dữ liệu có sẵn về các tiêu chí này, chúng ta có thể xây dựng mô hình học máy để dự đoán doanh số bán hàng toàn cầu của một trò chơi video. Điều này không chỉ giúp nhà phát triển game hiểu rõ hơn về yếu tố nào có thể tác động tích cực đến việc trò chơi đạt được doanh số bán hàng cao mà còn hỗ trợ ngành công nghiệp game trong việc dự đoán và quản lý hiệu suất thị trường.

Bằng cách giải quyết vấn đề này, chúng ta có thể:

**Tối Ưu Hóa Chiến Lược Tiếp Thị**: Hiểu rõ về những yếu tố nào thu hút và giữ chân người chơi sẽ giúp nhà phát triển tối ưu hóa chiến lược tiếp thị của họ để tăng cường doanh số bán hàng.

**Lập Kế Hoạch Phát Hành Chính Xác**: Dự đoán doanh số bán hàng toàn cầu có thể giúp nhà phát triển lên kế hoạch phát hành một cách chiến lược, tận dụng những yếu tố có thể đưa ra ảnh hưởng tích cực đối với thị trường toàn cầu.

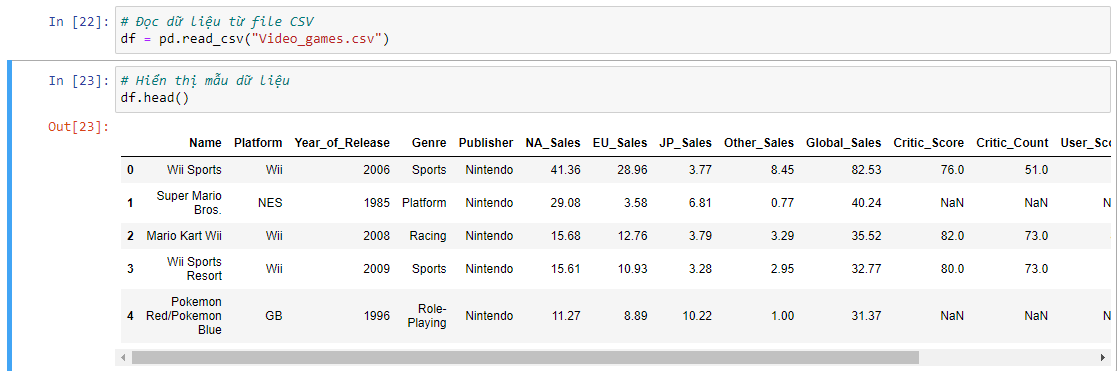
**Tăng Cường Tương Tác Cộng Đồng**: Hiểu rõ về yếu tố mà cộng đồng game thủ đánh giá cao sẽ giúp nhà phát triển tương tác hiệu quả với cộng đồng, đồng thời tối ưu hóa doanh số bán hàng.

**Dự Đoán Tiềm Năng Doanh Số**: Mô hình có thể giúp dự đoán tiềm năng doanh số bán hàng của một tựa game dựa trên tiêu chí đánh giá, từ đó hỗ trợ quản lý tài chính và nguồn lực hiệu quả.

Bằng cách này, dự án không chỉ hỗ trợ trong quá trình phát triển game mà còn đóng góp vào sự hiểu biết sâu sắc hơn về những yếu tố quyết định sự thành công của một tựa game trên thị trường toàn cầu.

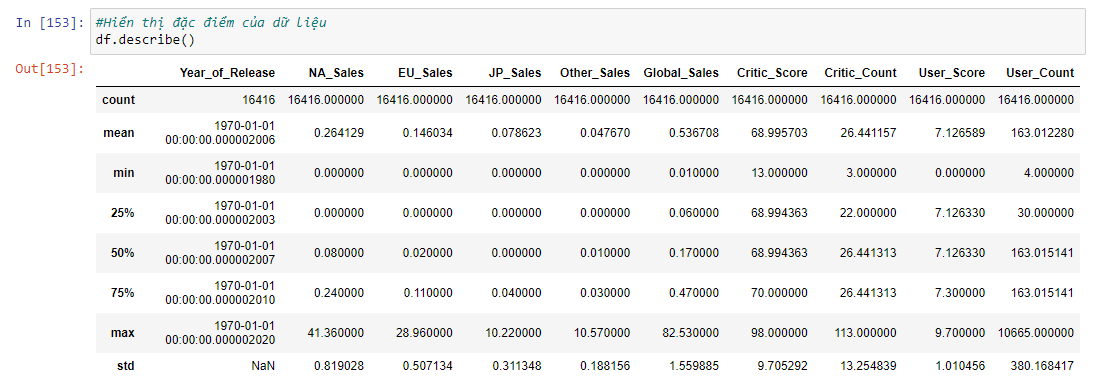
## **2.3 Làm sạch dữ liệu**

- Đọc dữ liệu:



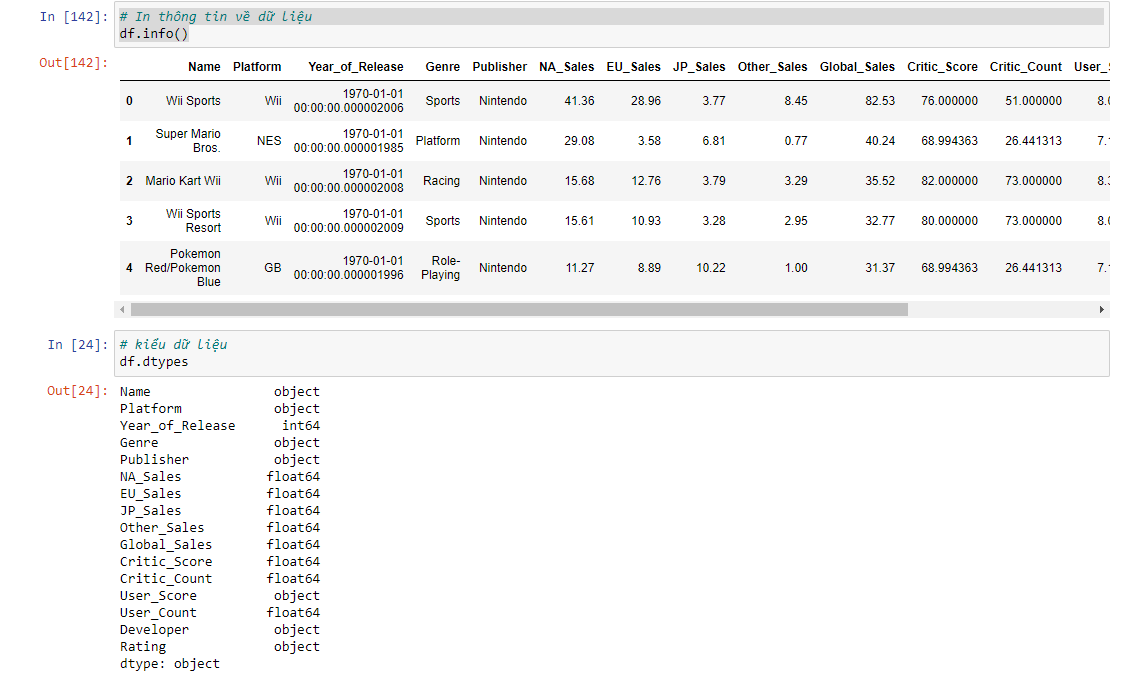
Hinh đọc dữ liệu

- Thống kê mô tả của dữ liệu số:



Hinh Thống kê mô tả của dữ liệu số

- Thông tin và kiểu dữ liệu



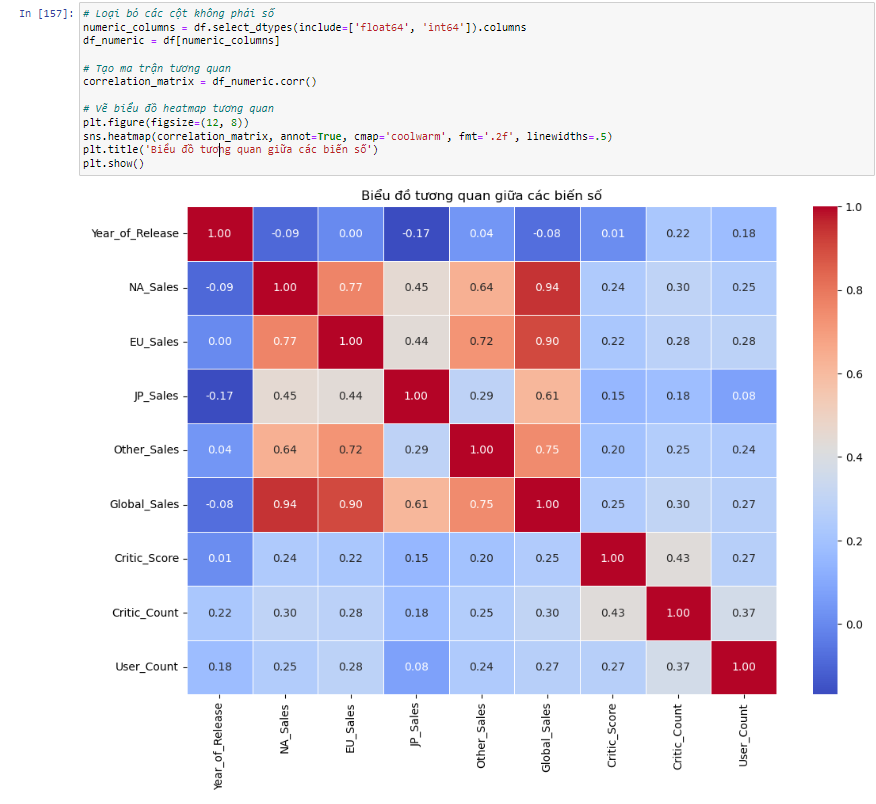
Hinh Thông tin và kiểu dữ liệu

- Xử lí giá trị trùng lặp và giá trị bị thiếu:



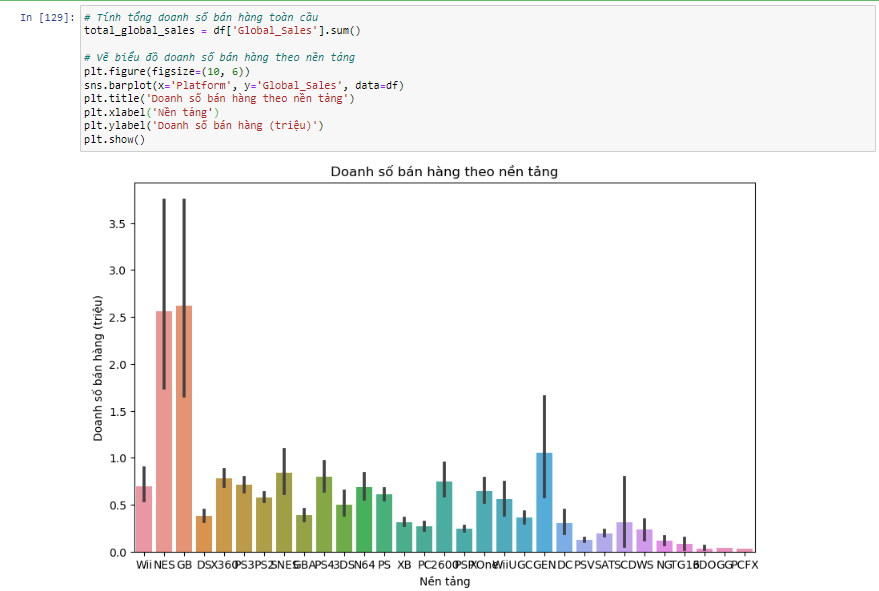
Hinh Xử lí giá trị trùng lặp và giá trị bị thiếu

## **2.4 Trực quan hóa dữ liệu**

* Biều đồ heatmap cho thấy tương quan giữa các biến số:
* 

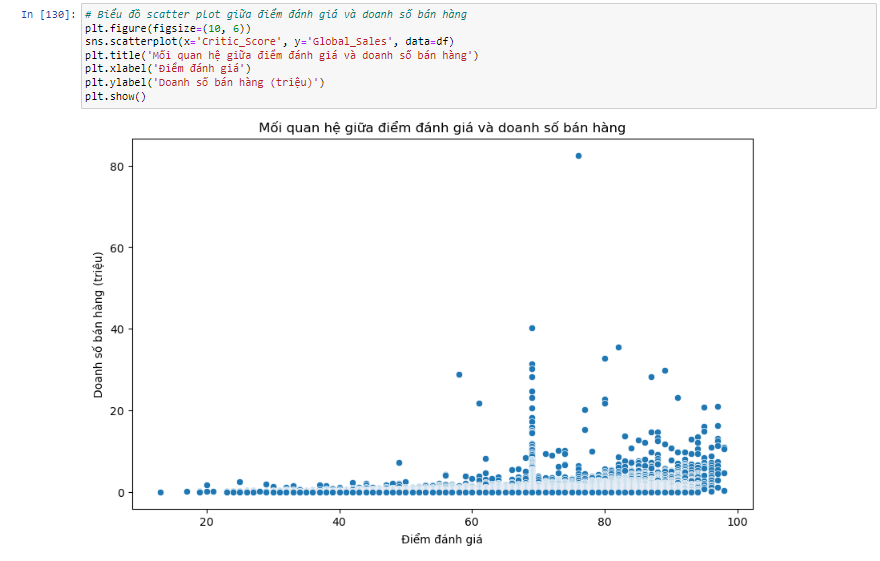
Hinh Biều đồ heatmap thể hiện tương quan giữa cá biến số

* Biểu đồ barplot thể hiện tổng doanh số bán hàng toàn cầu:



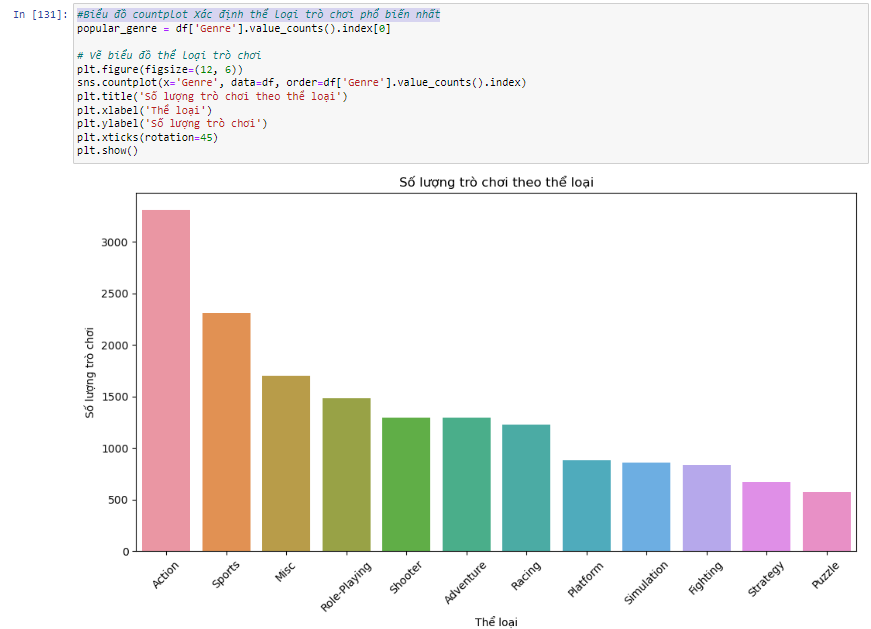
Hinh Biểu đồ barplot thể hiện tổng doanh số bán hàng toàn cầu

* Biểu đồ scatter plot giữa điểm đánh giá và doanh số bán hàng:



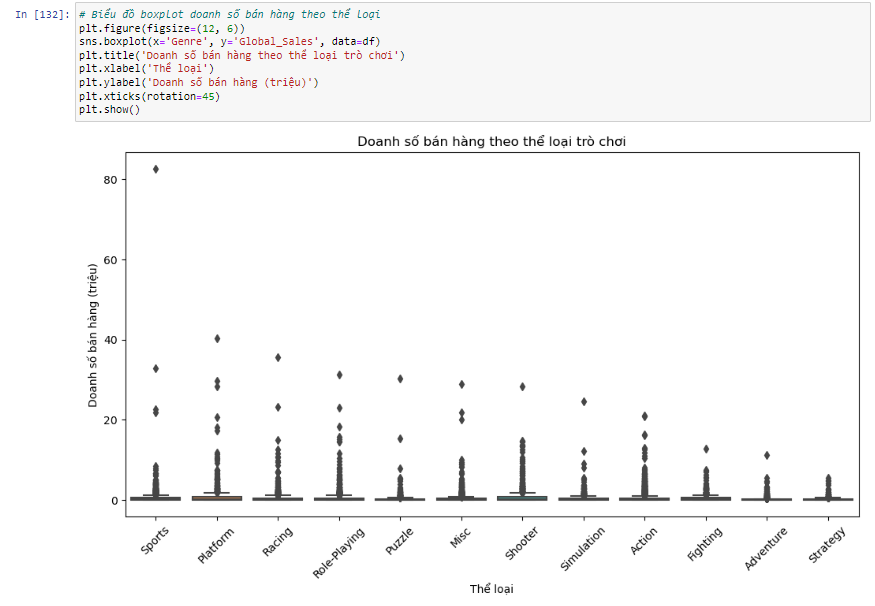
Hinh Biểu đồ scatter plot giữa điểm đánh giá và doanh số bán hàng

* Biểu đồ countplot Xác định thể loại trò chơi phổ biến nhất:



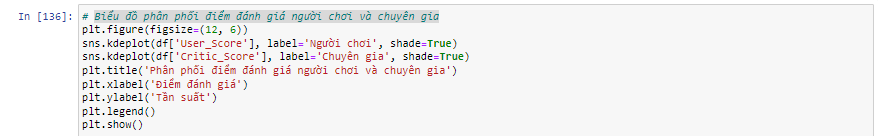
Hinh Biểu đồ countplot Xác định thể loại trò chơi phổ biến nhất

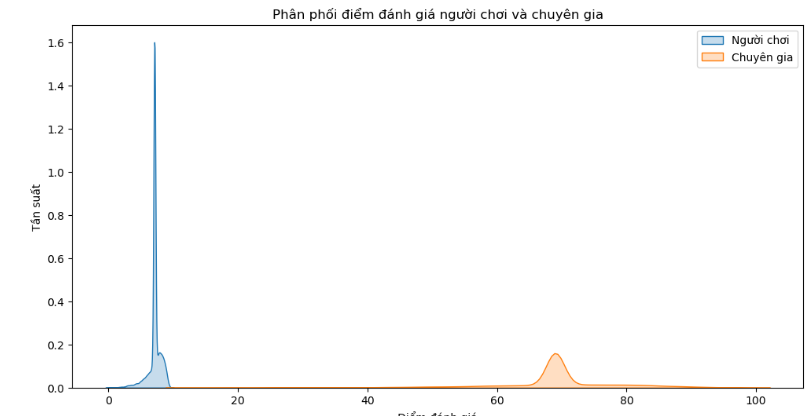
* Biểu đồ boxplot doanh số bán hàng theo thể loại:



Hinh Biểu đồ boxplot doanh số bán hàng theo thể loại

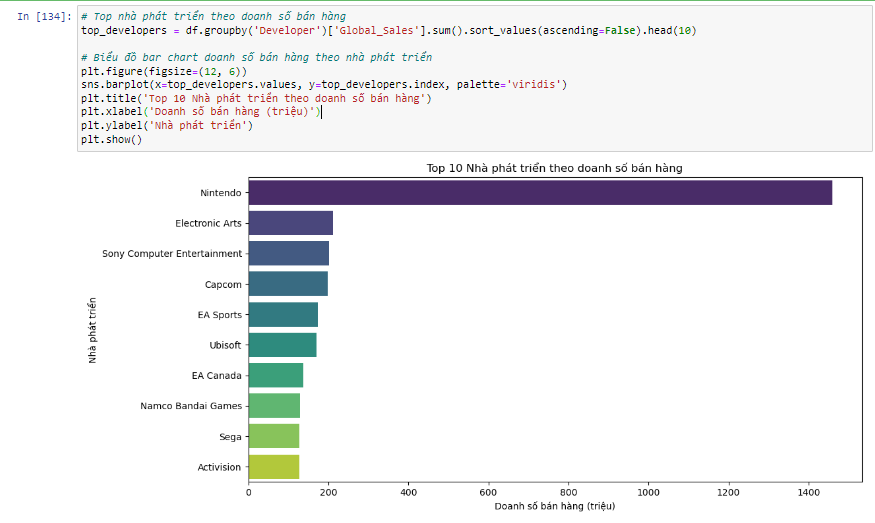
* Biểu đồ phân phối điểm đánh giá người chơi và chuyên gia:





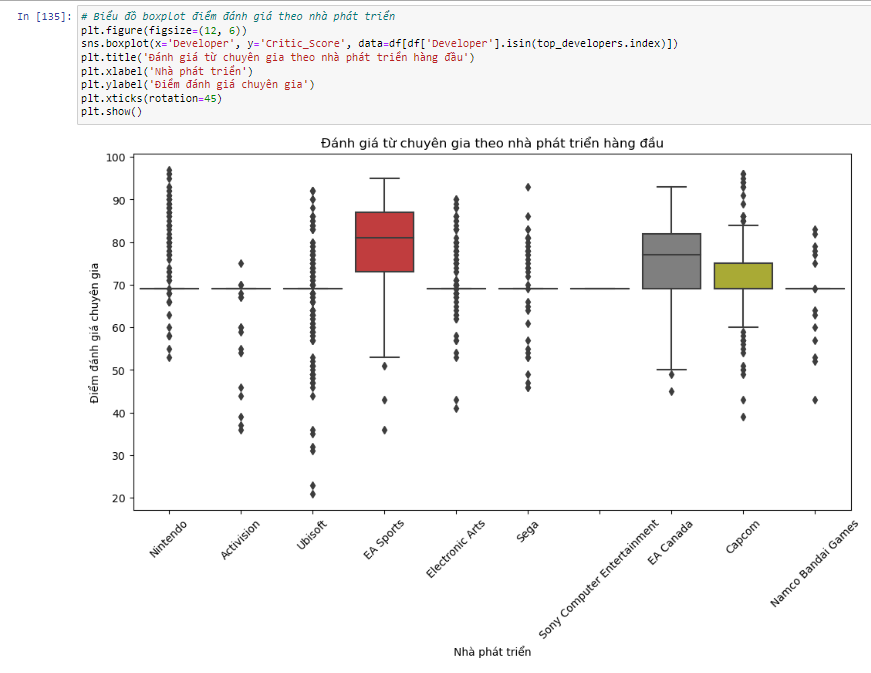
Hinh Biểu đồ phân phối điểm đánh giá người chơi và chuyên gia

* Top nhà phát triển theo doanh số bán hàng:



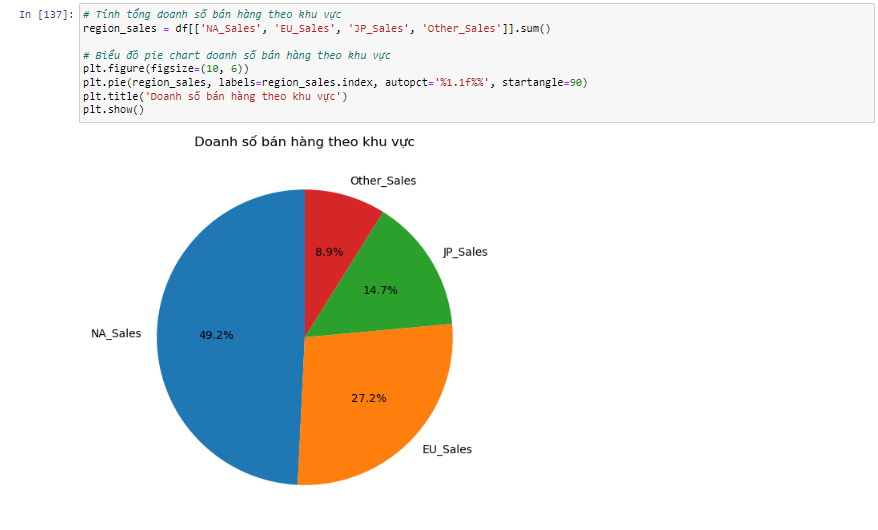
Hinh Top nhà phát triển theo doanh số bán hàng

* Biểu đồ boxplot điểm đánh giá theo nhà phát triển:



Hinh Biểu đồ boxplot điểm đánh giá theo nhà phát triển

* Tính tổng doanh số bán hàng theo khu vực:



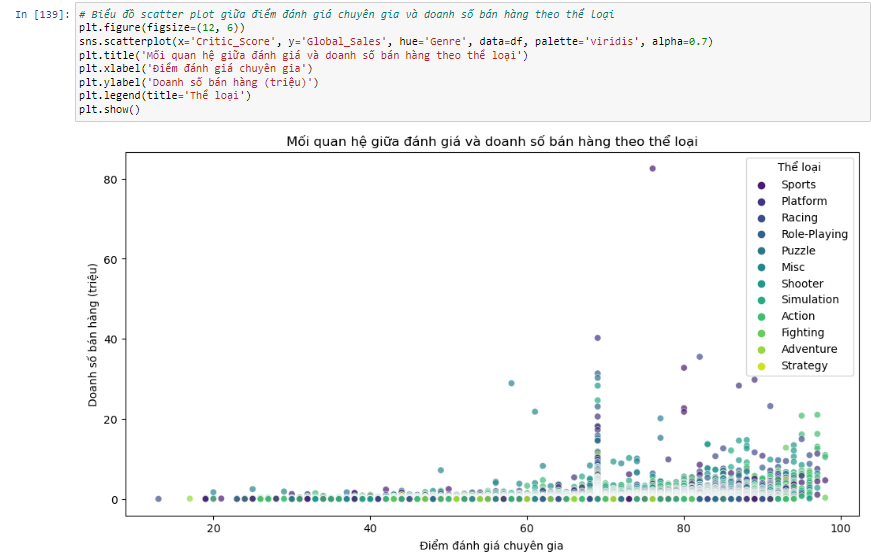
Hinh Tính tổng doanh số bán hàng theo khu vực

* Biểu đồ pivot cho thể loại và khu vực:



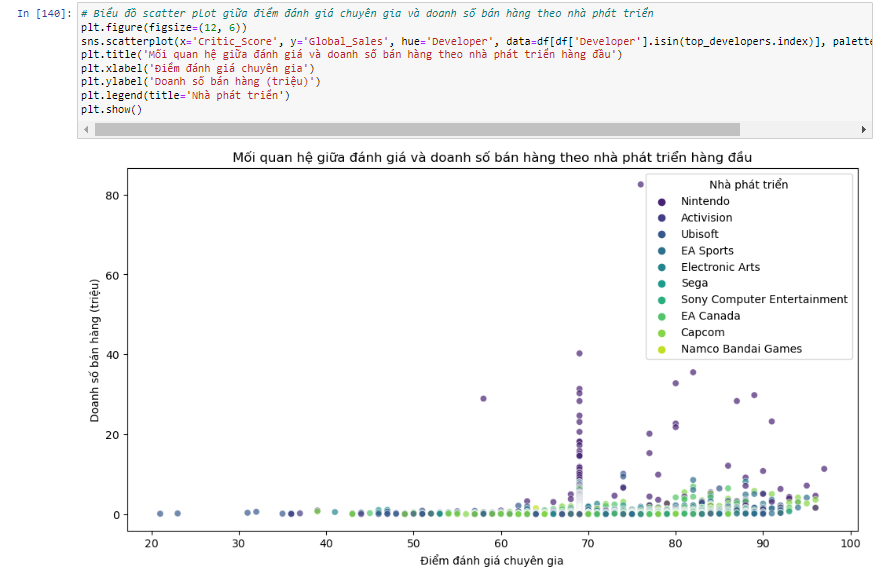
Hinh Biểu đồ pivot cho thể loại và khu vực

* Biểu đồ scatter plot giữa điểm đánh giá chuyên gia và doanh số bán hàng theo thể loại:



Hinh Biểu đồ scatter plot giữa điểm đánh giá chuyên gia và doanh số bán hàng theo thể loại

* Biểu đồ scatter plot giữa điểm đánh giá chuyên gia và doanh số bán hàng theo nhà phát triển:



Hinh Biểu đồ scatter plot giữa điểm đánh giá chuyên gia và doanh số bán hàng theo thể loại

## **2.5 Áp dụng thuật toán**

Để áp dụng thuật toán Simple Linear Regression vào dự đoán doanh số bán toàn cầu (Global Sales) của các video game dựa trên các thông tin như mô tả, sự phổ biến, ngày phát hành và số lượt bình chọn, bạn có thể thực hiện các bước sau:

**Thu thập dữ liệu:**

-Sử dụng API hoặc truy cập cơ sở dữ liệu video game để thu thập thông tin, bao gồm mô tả, sự phổ biến, ngày phát hành, số lượt bình chọn và doanh số bán toàn cầu.

**Tiền xử lý dữ liệu:**

-Xác định và xử lý dữ liệu còn thiếu, nếu có.

-Chuyển đổi các biến không phải dạng số thành dạng số nếu cần thiết để phục vụ quá trình mô hình hóa.

**Chia dữ liệu:**

-Phân chia dữ liệu thành tập huấn luyện và tập kiểm tra, sử dụng một tỷ lệ phù hợp.

**Chuẩn bị dữ liệu:**

-Chia dữ liệu thành biến độc lập (mô tả, sự phổ biến, ngày phát hành, số lượt bình chọn) và biến phụ thuộc (doanh số bán toàn cầu).

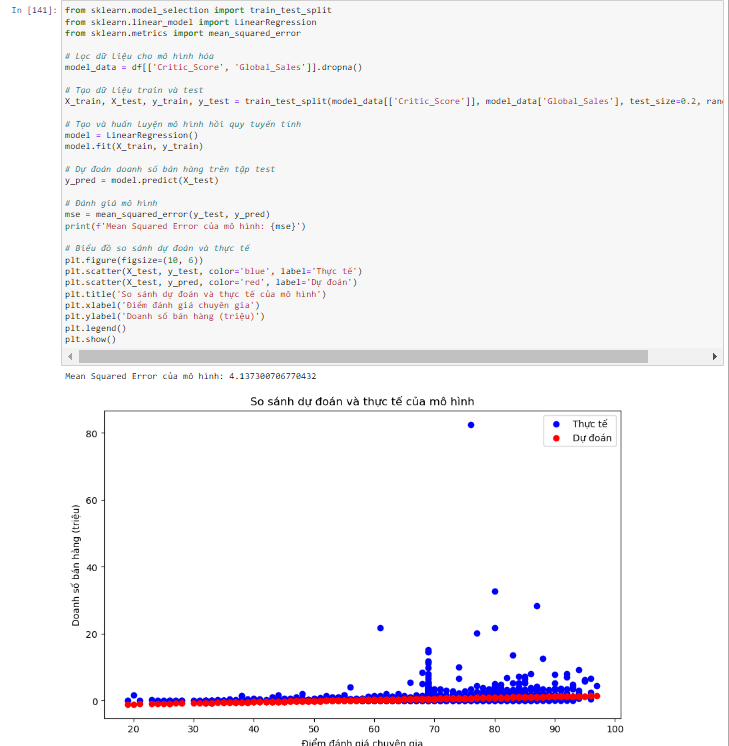
**Xây dựng mô hình Simple Linear Regression:**

-Sử dụng thư viện học máy như Scikit-learn để xây dựng mô hình Linear Regression. -Huấn luyện mô hình với tập dữ liệu huấn luyện, sử dụng các biến độc lập để dự đoán biến phụ thuộc (doanh số bán toàn cầu).

**Đánh giá mô hình:**

-Sử dụng tập kiểm tra để đánh giá hiệu suất của mô hình.

-Sử dụng các chỉ số như R-squared, Mean Squared Error (MSE), Root Mean Squared Error (RMSE) để đo lường mức độ phù hợp của mô hình với dữ liệu.



Hình 17 Mô hình Linear Regression

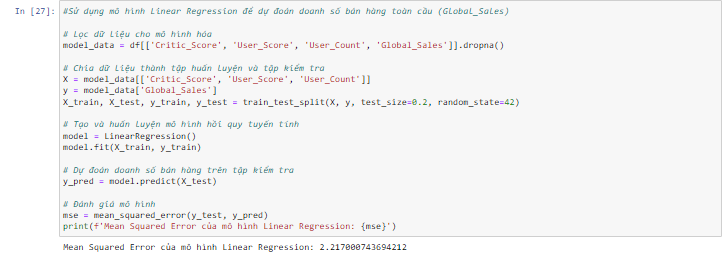
## **2.6 Xây dựng mô hình**

Để xây dựng mô hình Simple Linear Regression cho bài toán "Video Games Global Sales", bạn cần thực hiện các bước sau:

-Thu thập dữ liệu về số lượng đánh giá và mức độ phổ biến đối với mỗi bộ phim.

-Khi có các biến độc lập, hàm train\_test\_split sẽ tách chúng thành hai phần theo tỷ lệ xác định và có thể sử dụng cho việc huấn luyện và kiểm định mô hình.

-Sử dụng thư viện hoặc gói phần mềm hỗ trợ Linear Regression đa biến để xây dựng mô hình. Trong Python, bạn có thể sử dụng thư viện scikit-learn.

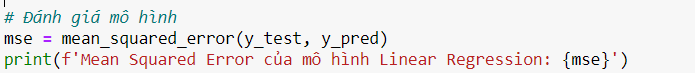


Hình 18 Tạo mô hình hồi quy tuyến tính

## **2.7 Đánh giá mô hình**

Để đánh giá mô hình Linear Regression đơn biến, bạn có thể sử dụng các độ đo sau:

-Sai số bình phương trung bình (Mean Squared Error - MSE): Đo lường trung bình phương sai số giữa giá trị thực và giá trị dự đoán càng thấp,mô hình càng tốt



Hình 19 Đánh giá sai số bình phương trung bình

## **2.8 Nhận Xét**

**Mean Squared Error (MSE):**

Giá trị MSE là một chỉ số đo lường độ lớn của sai số bình phương trung bình giữa giá trị dự đoán và giá trị thực tế. Trong trường hợp này, MSE có giá trị là 4.27, đây là sự chênh lệch tương đối giữa giá trị dự đoán và giá trị thực tế.

Tổng quát, các giá trị MSE khá nhỏ, cho thấy mô hình dự đoán tương đối tốt trên tập dữ liệu kiểm tra. Điều này có thể cần phải xem xét và cải thiện mô hình hoặc xem xét lại các đặc trưng hoặc quá trình tiền xử lý dữ liệu.

## **2.9 Đưa ra giải pháp**

**Mở rộng tập dữ liệu:**

Nếu có các biến khác như vùng miền, loại ngành nghề, hoặc bất kỳ thông tin nào khác có thể ảnh hưởng đến doanh số bán hàng toàn cầu của video game, việc bổ sung chúng có thể cải thiện dự đoán của mô hình. Thông tin bổ sung có thể giúp mô hình trở nên chính xác hơn và có khả năng dự đoán tốt hơn với các trường hợp mới.

**Xử lý Ngoại Lệ:**

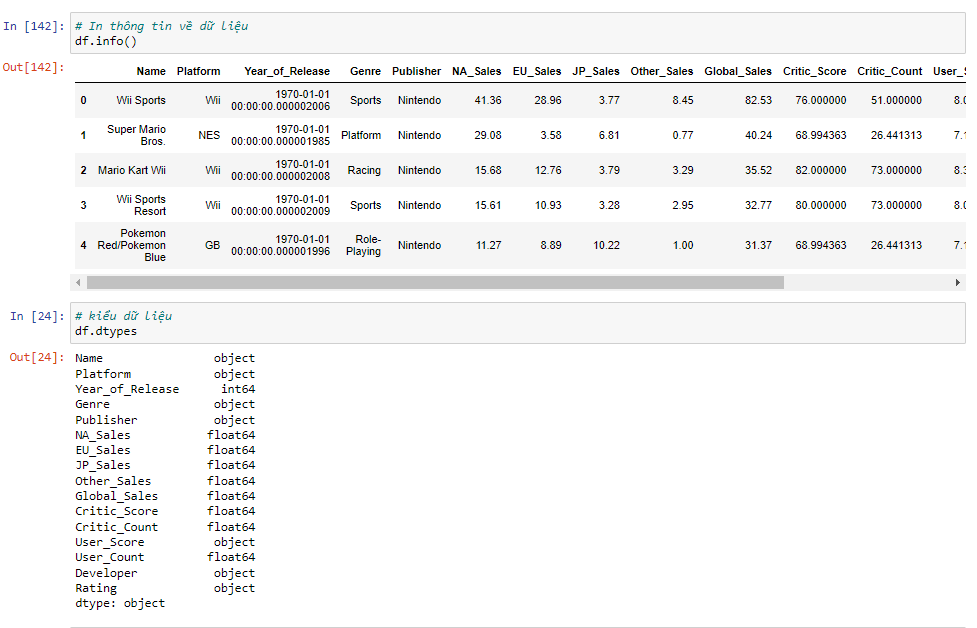
Kiểm tra và xử lý các giá trị ngoại lệ có thể ảnh hưởng đến mô hình. Có thể loại bỏ hoặc xử lý chúng một cách thích hợp để giảm thiểu ảnh hưởng tiêu cực đối với dự đoán.

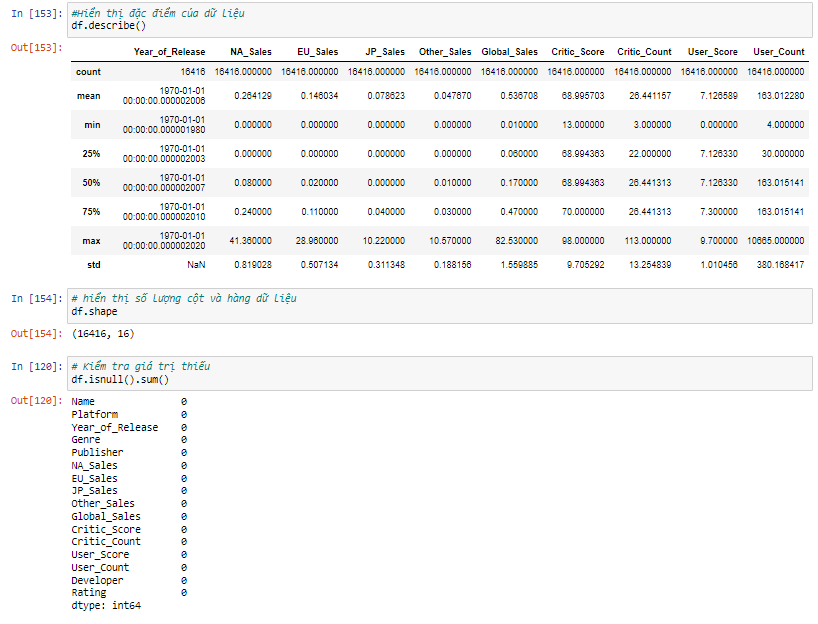
**Sử dụng mô hình học máy phức tạp hơn:**

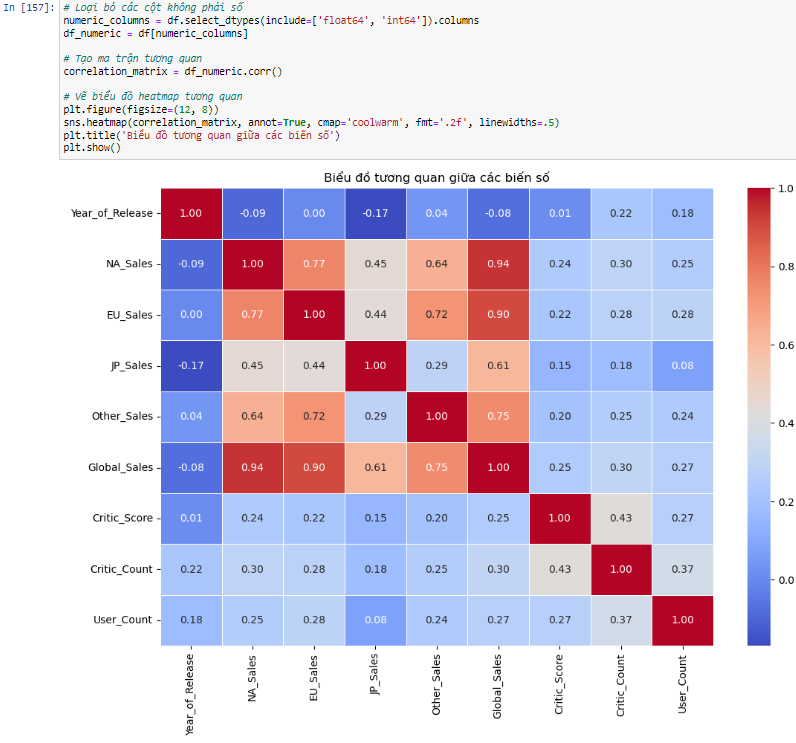
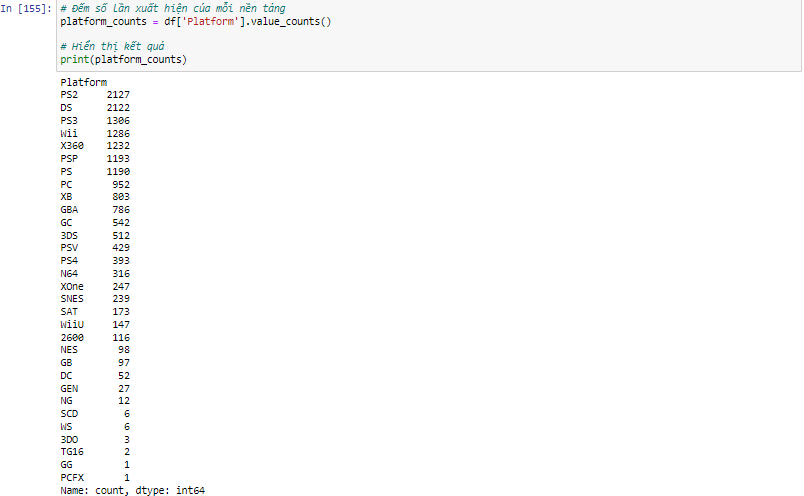
Ngoài Linear Regression, có thể xem xét sử dụng các mô hình học máy phức tạp hơn như Gradient Boosting Machines (GBM). GBM có khả năng tối ưu hóa hàm mất mát thông qua việc xây dựng các cây quyết định và có thể cung cấp hiệu suất tốt hơn trên dữ liệu phức tạp.

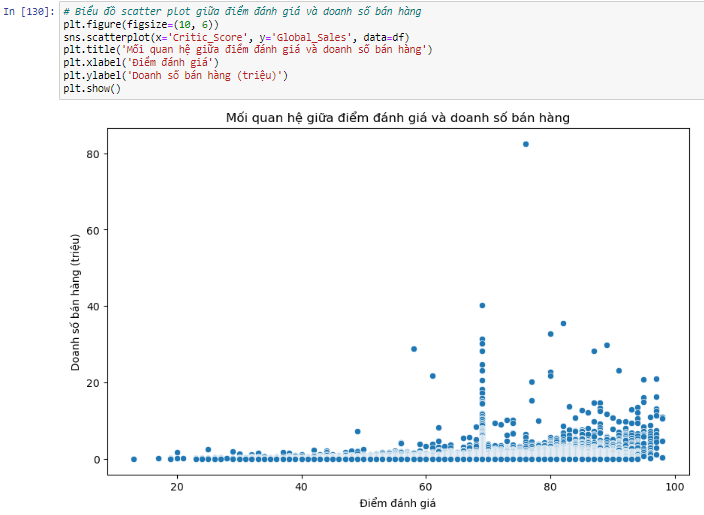
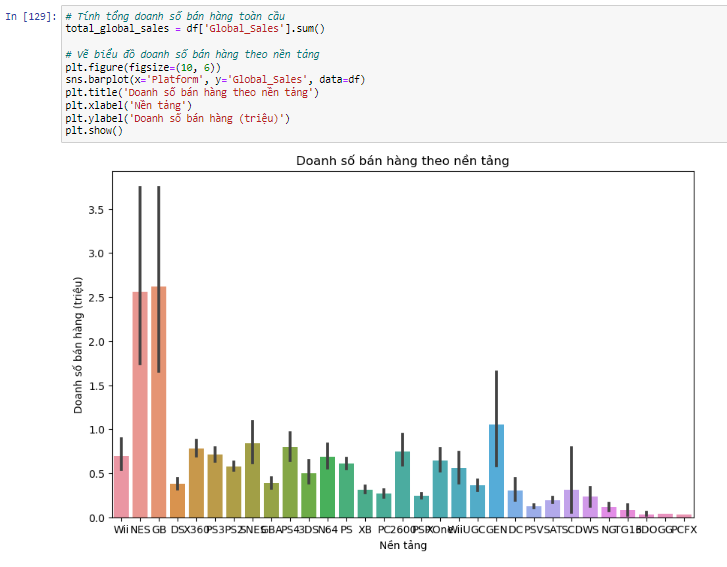
## **2.10 Demo chương trình:**

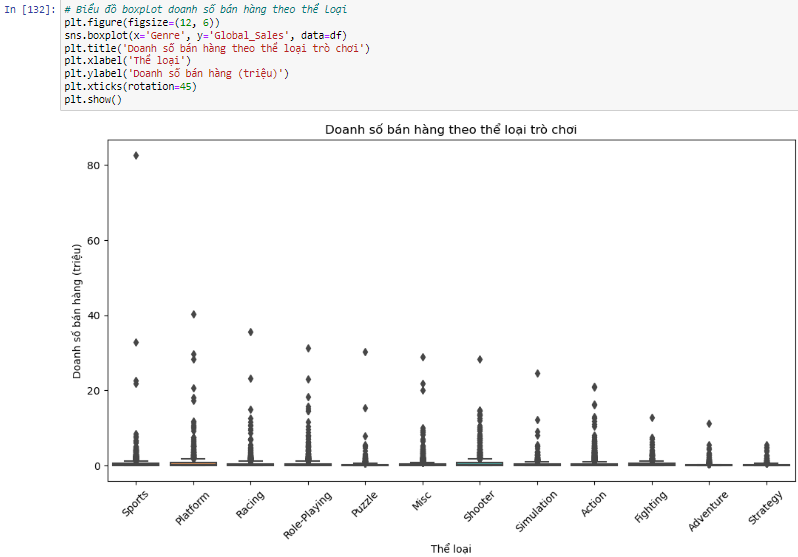
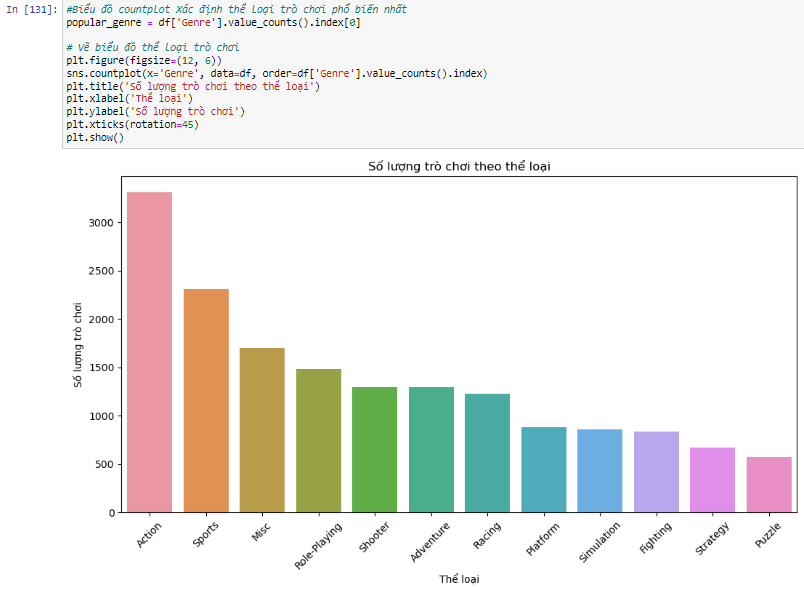
****

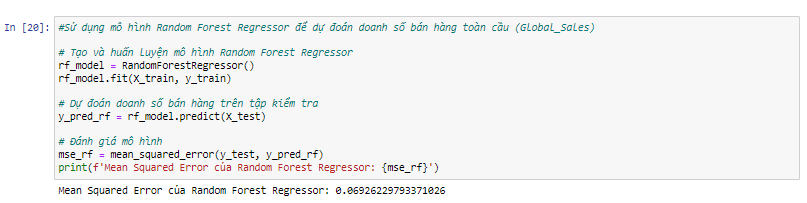
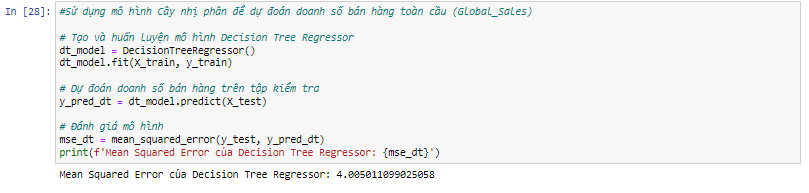
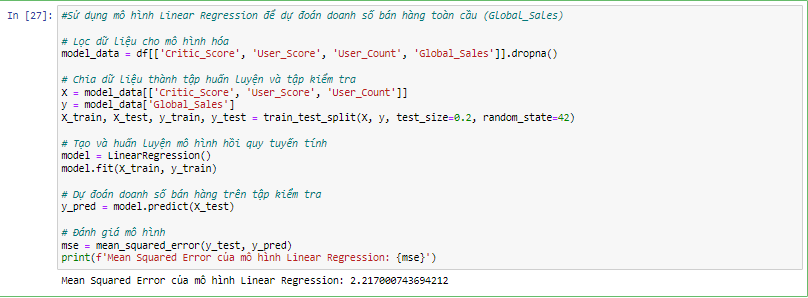
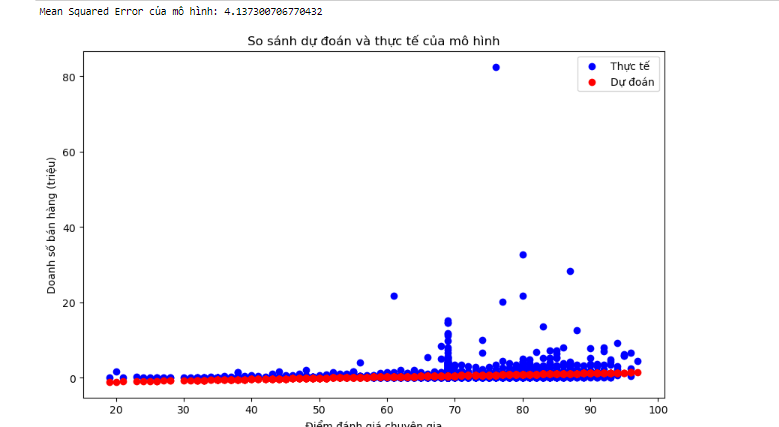
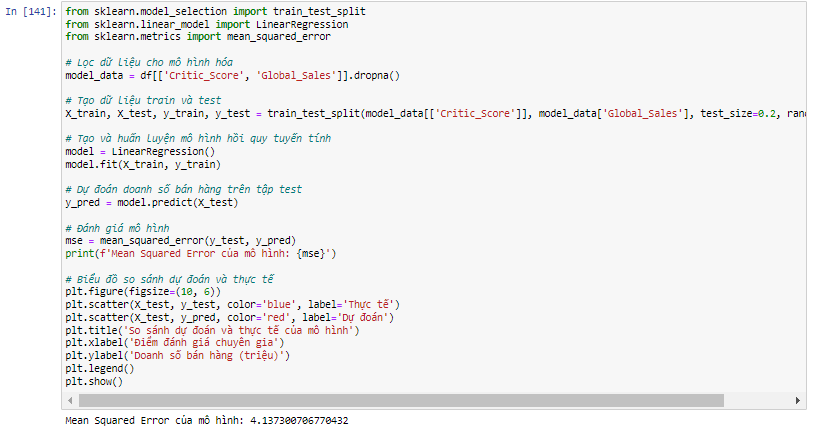
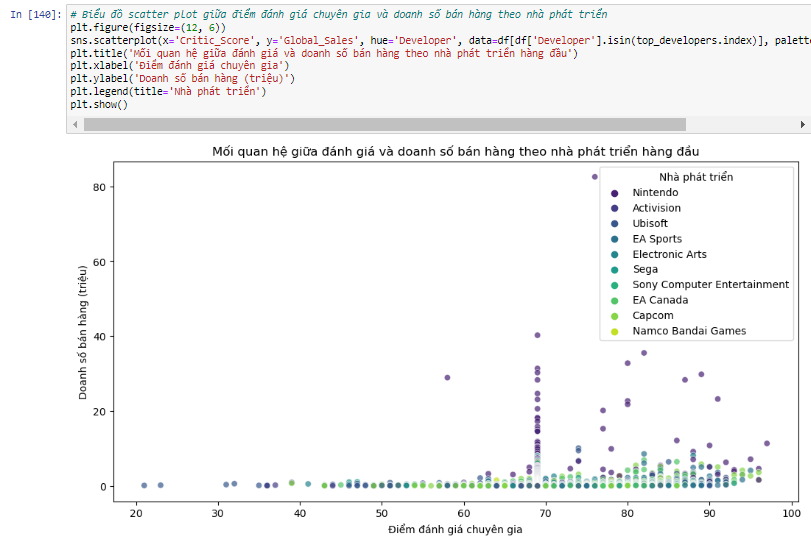
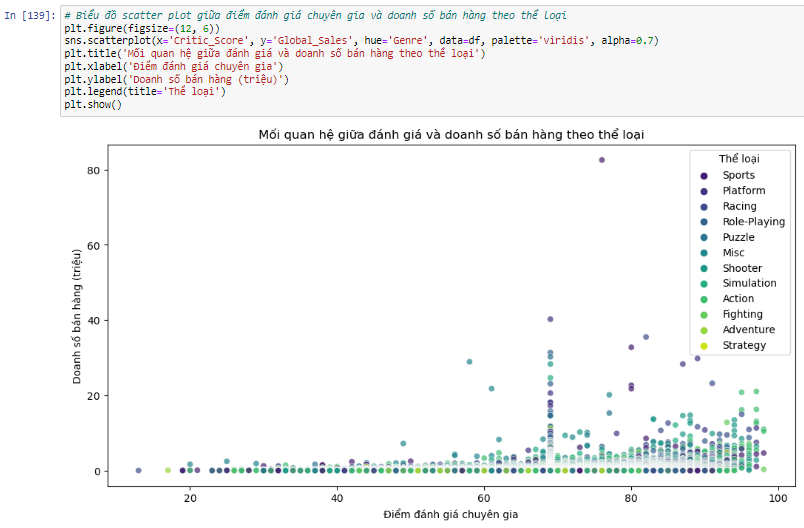
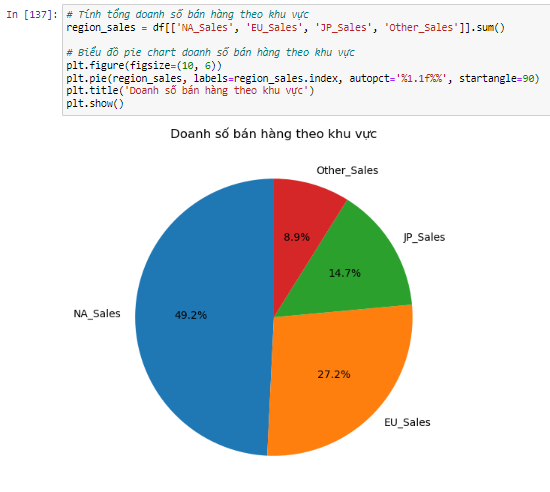
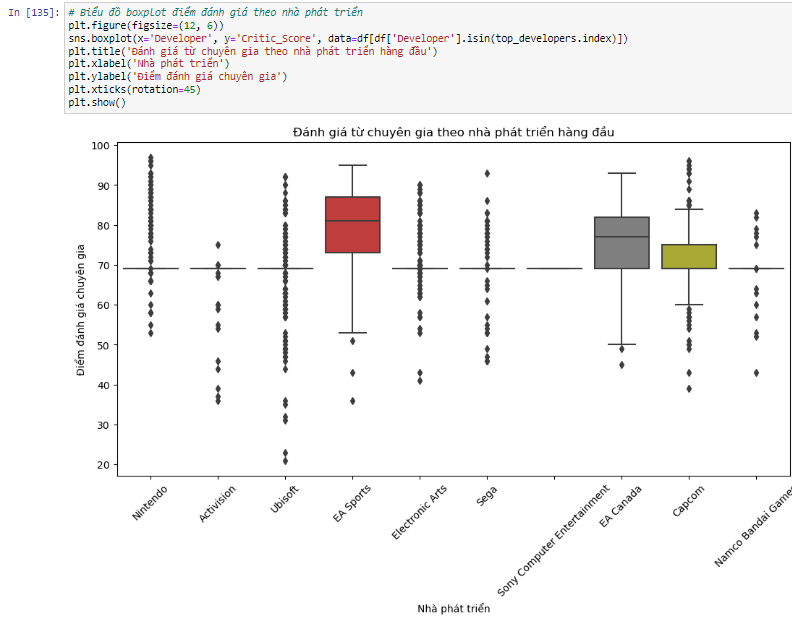
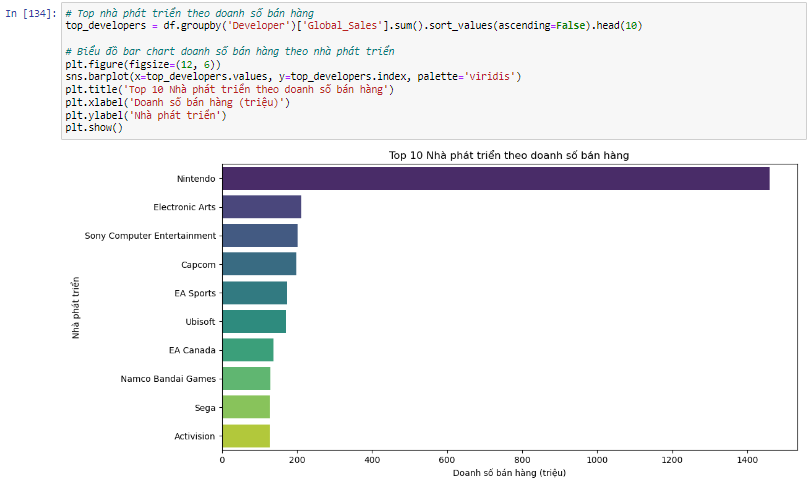
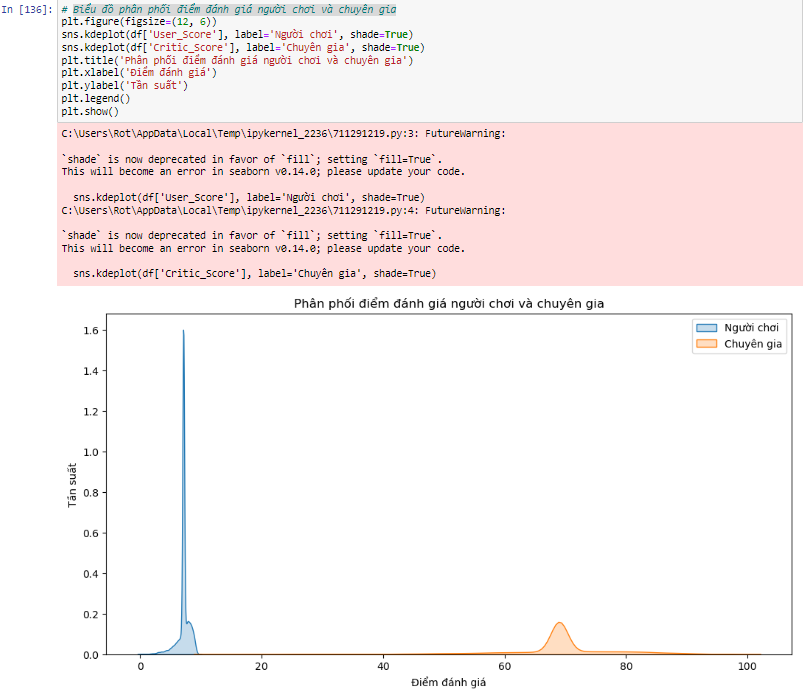
****

****

****

****

****

****

# **TÀI LIỆU THAM KHẢO**

<https://www.kaggle.com/code/thomkell/linear-regression>

https://www.kaggle.com/datasets/ghassenkhaled/video-games-data