

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC ĐẠI NAM



BÀI TẬP LỚN

TÊN HỌC PHẦN: DỮ LIỆU LỚN

ĐỀ TÀI: DỰ ĐOÁN DOANH SỐ BÁN TRÒ CHƠI ĐIỆN TỬ

Giáo viên hướng dẫn: LÊ THỊ THÙY TRANG

Sinh viên thực hiện:

STT	Mã sv	Họ và Tên	Ngày Sinh	Lớp
1	1671020092	Phùng Xuân Đức	22/02/2004	CNTT-1602
2	1671020207	Vũ Đức Minh	19/12/2004	CNTT-1602
3	1671020350	Đỗ Quốc Việt	14/01/2004	CNTT-1602
4	1671020044	Hà Minh Chiến	12/08/2004	CNTT-1602

Hà Nội, 2025

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC ĐẠI NAM**



BÀI TẬP LỚN

TÊN HỌC PHẦN: CÔNG NGHỆ PHẦN MỀM

ĐỀ TÀI: DỰ ĐOÁN DOANH SỐ BÁN TRÒ CHƠI ĐIỆN TỬ

STT	Mã Sinh Viên	Họ và Tên	Ngày Sinh	Điểm	
				Bảng Số	Bảng Chữ
1	1671020092	Phùng Xuân Đức	22/02/2004		
2	1671020207	Vũ Đức Minh	19/12/2004		
3	1671020350	Đỗ Quốc Việt	14/01/2004		
4	1671020044	Hà Minh Chiến	12/08/2004		

CÁN BỘ CHẤM THI 1

CÁN BỘ CHẤM THI 2

Trần Quý Nam

Hà Nội, 2025

LỜI NÓI ĐẦU

Trong thời đại công nghệ số phát triển mạnh mẽ, ngành công nghiệp trò chơi điện tử đang chứng kiến sự bùng nổ với hàng loạt trò chơi mới được phát hành mỗi năm. Điều này tạo ra sự cạnh tranh khốc liệt giữa các nhà phát triển và nhà phát hành game, đồng thời làm cho việc dự đoán doanh số bán hàng trở thành một yếu tố quan trọng giúp đưa ra quyết định chiến lược.

Bài tập lớn này nhằm nghiên cứu và áp dụng các phương pháp phân tích dữ liệu lớn cùng với mô hình học máy để dự đoán doanh số bán trò chơi điện tử. Bằng cách xử lý và phân tích tập dữ liệu doanh số bán game, nhóm em mong muốn xác định các yếu tố quan trọng ảnh hưởng đến doanh số, từ đó cung cấp những gợi ý hữu ích cho ngành công nghiệp game.

Trong quá trình thực hiện bài tập lớn, nhóm em đã cố gắng tìm hiểu và ứng dụng những kỹ thuật tiên tiến nhất trong lĩnh vực dữ liệu lớn và học máy. Tuy nhiên, do hạn chế về thời gian và kiến thức, bài báo cáo không thể tránh khỏi thiếu sót. Nhóm em mong nhận được sự góp ý của thầy cô và các bạn để bài nghiên cứu này được hoàn thiện hơn. Nhóm em xin chân thành cảm ơn các giảng viên đã hướng dẫn tận tình chỉ bảo trong suốt quá trình học tập thực hiện bài tập lớn này.

MỤC LỤC

CHƯƠNG 1: TỔNG QUAN CƠ SỞ LÝ THUYẾT -----	7
1.1 Mục đích nghiên cứu-----	7
1.2 Mục tiêu nghiên cứu -----	7
1.3 Phạm vi nghiên cứu -----	8
1.4 Phương pháp nghiên cứu -----	8
CHƯƠNG 2: MÔ TẢ TẬP DỮ LIỆU VÀ CÔNG NGHỆ SỬ DỤNG-----	10
2.1 Giới thiệu về Kaggle -----	10
2.2 Tập dữ liệu sử dụng -----	11
2.2.1 Các thuộc tính của tập dữ liệu -----	11
2.2.2 Mục Tiêu Phân Tích Dữ Liệu -----	12
2.2.3 Cách sử dụng tập dữ liệu -----	13
2.3 Giới thiệu Apache Spark -----	13
2.3.1 Thành phần của Spark -----	14
2.3.2 Những điểm nổi bật của Spark -----	15
2.3.3 So sánh spark với hadoop -----	16
2.3.4 Ứng dụng của spark-----	17
CHƯƠNG 3: KẾT QUẢ XỬ LÝ VÀ PHÂN TÍCH DỮ LIỆU -----	18
3.1 Kết quả xử lý-----	18
3.2 Phân tích dữ liệu -----	26
3.2.1 Xử lý dữ liệu -----	26
3.2.2 Chia dữ liệu thành tập huấn luyện và kiểm tra -----	27

3.2.3 Huấn luyện mô hình -----	27
3.2.4 Đánh giá mô hình -----	28
3.2.5 Dự báo doanh thu game -----	28
KẾT LUẬN-----	30
DANH MỤC TÀI LIỆU THAM KHẢO -----	31

MỤC LỤC HÌNH ẢNH

Hình 1: Trang chủ kaggle-----	11
Hình 2: Tập dữ liệu Explore Video Games Sales-----	12
Hình 3: Quá trình phát triển của spark -----	13
Hình 4: Các thành phần của spark -----	14
Hình 5: Biểu đồ số lượng game theo năm -----	20
Hình 6: Biểu đồ doanh thu theo năm -----	21
Hình 7: Biểu đồ thể loại có doanh thu cao nhất-----	23
Hình 8: Biểu đồ game có doanh thu cao nhất -----	24
Hình 9: Biểu đồ những nhà phát hành có số lượng game lớn-----	26
Hình 10: Biểu đồ dự đoán doanh thu game-----	29

CHƯƠNG 1: TỔNG QUAN CƠ SỞ LÝ THUYẾT

1.1 Mục đích nghiên cứu

Trò chơi điện tử là một trong những ngành công nghiệp giải trí phát triển nhanh nhất trên thế giới, thu hút hàng triệu người chơi và tạo ra doanh thu hàng tỷ đô la mỗi năm. Sự phát triển của công nghệ, đặc biệt là trí tuệ nhân tạo, đồ họa máy tính, và nền tảng chơi game trực tuyến, đã góp phần tạo nên sự bùng nổ của ngành công nghiệp này.

Cùng với sự mở rộng của thị trường, các nhà phát hành và phát triển game ngày càng quan tâm đến việc dự đoán xu hướng tiêu thụ để tối ưu hóa chiến lược kinh doanh. Việc phân tích dữ liệu doanh số bán hàng trong quá khứ có thể giúp xác định các yếu tố quan trọng ảnh hưởng đến mức độ thành công của một trò chơi. Những yếu tố này bao gồm thể loại game, nền tảng phát hành, khu vực tiêu thụ, năm phát hành, chiến lược tiếp thị và chất lượng nội dung. Nhờ đó, các công ty có thể đưa ra quyết định sáng suốt hơn trong việc phát triển và quảng bá trò chơi.

Ngoài ra, dự đoán doanh số trò chơi còn hỗ trợ nhà đầu tư, cửa hàng bán lẻ và các đối tác kinh doanh trong việc tối ưu hóa nguồn lực, quản lý hàng tồn kho, và lập kế hoạch chiến lược dài hạn. Do đó, nghiên cứu về dự đoán doanh số trò chơi điện tử không chỉ có giá trị đối với các nhà phát triển mà còn mang lại lợi ích cho toàn bộ hệ sinh thái ngành game.

1.2 Mục tiêu nghiên cứu

Nghiên cứu này nhằm:

- Xây dựng mô hình dự đoán doanh số bán trò chơi điện tử dựa trên dữ liệu lịch sử.
- Phân tích các yếu tố tác động đến doanh số bán hàng như thể loại, nền tảng, năm phát hành, nhà phát triển, nhà xuất bản và khu vực địa lý.
- So sánh hiệu suất của các mô hình dự đoán khác nhau, từ Hồi quy tuyến tính đến các

mô hình học máy tiên tiến như Random Forest, Gradient Boosting, và Mạng nơ-ron nhân tạo (ANN).

1.3 Phạm vi nghiên cứu

Nghiên cứu này tập trung vào bộ dữ liệu Explore Video Games Sales, bao gồm thông tin về các trò chơi điện tử phát hành từ trước đến nay trên nhiều nền tảng khác nhau. Phạm vi phân tích bao gồm:

- **Dữ liệu doanh số:** Bao gồm doanh số toàn cầu và theo khu vực (Bắc Mỹ, Châu Âu, Nhật Bản, phần còn lại của thế giới).
- **Các đặc điểm của trò chơi:** Gồm nền tảng phát hành, thể loại game, nhà phát triển, nhà xuất bản và đánh giá của người chơi.
- **Khoảng thời gian phân tích:** Tập trung vào dữ liệu từ năm 1980 đến thời điểm hiện tại, để đảm bảo mô hình có thể học được xu hướng theo thời gian.
- **Phương pháp dự đoán:** Sử dụng các mô hình hồi quy và học máy để tìm ra mối quan hệ giữa các biến và doanh số bán hàng.

1.4 Phương pháp nghiên cứu

Để đạt được mục tiêu trên, nghiên cứu này sử dụng các phương pháp sau:

Thu thập và tiền xử lý dữ liệu:

- **Làm sạch dữ liệu:** Loại bỏ các giá trị trùng lặp, xử lý dữ liệu bị thiếu.
- **Mã hóa dữ liệu:** Biến đổi dữ liệu danh mục thành dạng số để phù hợp với mô hình học máy.
- **Chuẩn hóa dữ liệu:** Sử dụng phương pháp Min-Max Scaling hoặc Standard Scaling để đảm bảo dữ liệu có phân phối hợp lý.

Phân tích dữ liệu khám phá (EDA - Exploratory Data Analysis):

- Trực quan hóa dữ liệu để hiểu rõ phân phối của doanh số bán hàng theo thời gian, theo khu vực và theo nền tảng.
- Phân tích tương quan giữa các biến để xác định yếu tố quan trọng ảnh hưởng đến doanh số.

Xây dựng mô hình dự đoán:

- Sử dụng các mô hình hồi quy tuyến tính để dự đoán doanh số trò chơi.
- Thử nghiệm các thuật toán học máy như Decision Trees, Random Forest, XGBoost để cải thiện độ chính xác.
- Sử dụng kỹ thuật kiểm tra chéo (cross-validation) để đánh giá hiệu suất mô hình.

Đánh giá mô hình:

- Sử dụng các chỉ số đánh giá như R-squared, Mean Absolute Error (MAE), Root Mean Square Error (RMSE) để đo lường độ chính xác của mô hình.
- So sánh hiệu suất giữa các mô hình để chọn ra phương pháp tối ưu nhất.

CHƯƠNG 2: MÔ TẢ TẬP DỮ LIỆU VÀ CÔNG NGHỆ SỬ DỤNG

2.1 Giới thiệu về Kaggle

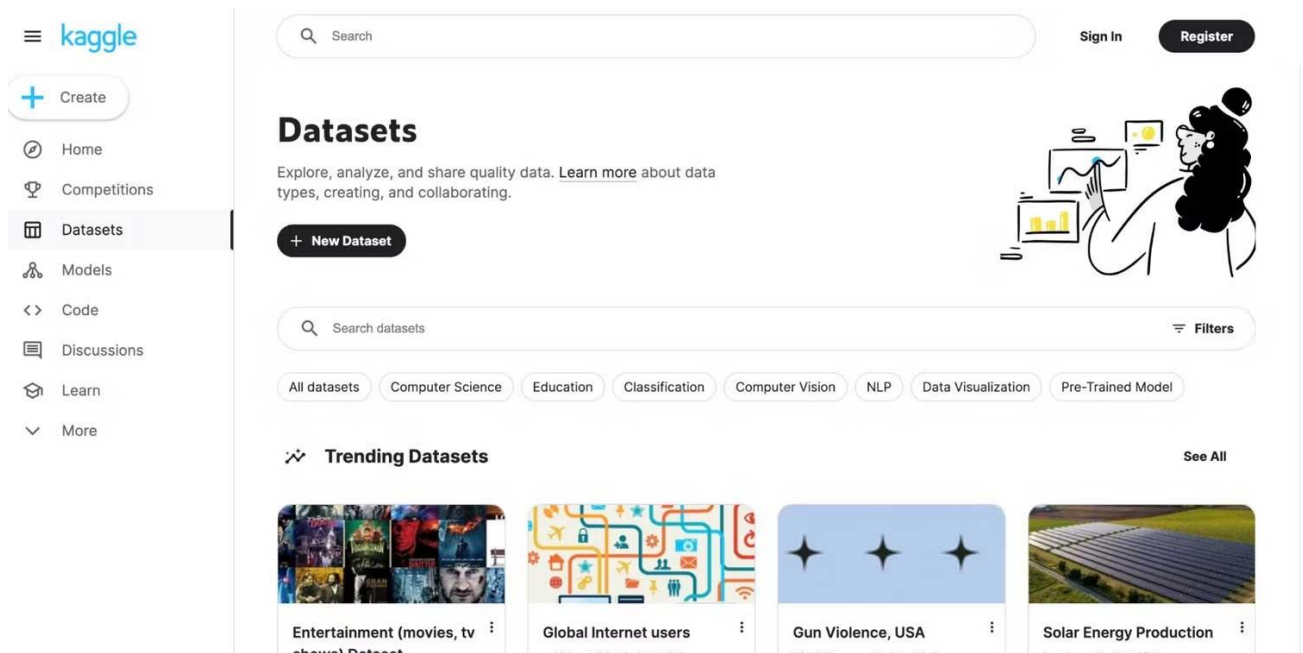
Kaggle là một cộng đồng online dành cho những ai đam mê khoa học dữ liệu và học máy (ML). Nó là công cụ học tập hàng đầu cho người mới và dân chuyên nghiệp với những vấn đề thực tế để mài giũa kỹ năng khoa học dữ liệu của bạn.

Được sở hữu bởi Google, nó hiện là nền tảng web có nguồn lực cộng đồng lớn nhất cho các nhà khoa học dữ liệu và học viên học máy. Kaggle cho bạn tiếp cận một số chuyên gia trong lĩnh vực mà có thể giúp bạn xây dựng ý tưởng, cạnh tranh và giải quyết những vấn đề đang gặp phải.

Dataset của Kaggle là tính năng hữu ích nhất vì việc tìm nguồn dữ liệu tại thời gian thực là vấn đề quan trọng với hầu hết các nhà khoa học dữ liệu. Hãy tưởng tượng bạn dành thời gian và tiền bạc để học lý thuyết nhưng lại không thể thực hành trong khi học thì sẽ khó đạt được hiệu quả cao.

Kaggle giải quyết vấn đề này bằng cách cung cấp hơn 50.000 dataset mà bạn có thể dùng trong khi huấn luyện các mô hình. Có một dataset cho bạn trên Kaggle. Tất nhiên, làm việc trên các dataset “hotter” có thể có lợi hơn cho người mới bắt đầu. Dù bạn có thể dùng kiến thức của bản thân để giải quyết vấn đề bất kỳ nhưng mọi việc sẽ dễ dàng hơn với sự trợ giúp của các dataset thông dụng. Ngoài ra, lưu ý rằng những dataset này xuất hiện ở nhiều định dạng file khác nhau, bao gồm CSV, JSON, SQLite...

Kaggle là cộng đồng trực tuyến một điểm dừng cho nhà khoa học dữ liệu bởi nó cho bạn cơ hội học hỏi từ người khác, qua network và trình bày công việc của bản thân. Bạn có thể đặt câu hỏi, kết nối với đồng nghiệp và xây dựng trên kiến thức hiện có qua cộng đồng. Trình bày tác phẩm cũng giúp bạn gây dựng danh tiếng như một chuyên gia trong lĩnh vực. Điều này rất hữu ích trong quá trình tìm việc.



Hình 1: Trang chủ kaggle

2.2 Tập dữ liệu sử dụng

Tập dữ liệu được sử dụng trong bài toán này là Explore Video Games Sales, một tập dữ liệu phổ biến về doanh số bán trò chơi điện tử, thường được tìm thấy trên các nền tảng như Kaggle. Tập dữ liệu này bao gồm thông tin về doanh số bán hàng của hàng nghìn trò chơi trên nhiều nền tảng khác nhau.

2.2.1 Các thuộc tính của tập dữ liệu

Tập dữ liệu chứa các cột quan trọng như:

- **Rank:** Xếp hạng của trò chơi theo doanh số bán toàn cầu.
- **Name:** Tên trò chơi.
- **Platform:** Hệ máy chơi game (PS4, Xbox One, PC, v.v.).
- **Year:** Năm phát hành trò chơi.
- **Genre:** Thể loại của trò chơi (Action, Sports, RPG, v.v.).

- **Publisher:** Nhà phát hành trò chơi.
- **NA_Sales:** Doanh số bán tại khu vực Bắc Mỹ (triệu bản).
- **EU_Sales:** Doanh số bán tại khu vực Châu Âu (triệu bản).
- **JP_Sales:** Doanh số bán tại khu vực Nhật Bản (triệu bản).
- **Other_Sales:** Doanh số bán tại các khu vực khác (triệu bản).
- **Global_Sales:** Tổng doanh số bán trên toàn cầu (triệu bản).

Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
1	Wii Sports	Wii	2006	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
2	Super Mario Bros.	NES	1985	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
3	Mario Kart Wii	Wii	2008	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
4	Wii Sports Resort	Wii	2009	Sports	Nintendo	15.75	11.01	3.28	2.96	33
5	Pokemon Red/Pokemon Blue	GB	1996	Role-Playing	Nintendo	11.27	8.89	10.22	1.31	37
6	Tetris	GB	1989	Puzzle	Nintendo	23.2	2.26	4.22	0.58	30.26
7	New Super Mario Bros.	DS	2006	Platform	Nintendo	11.38	9.23	6.5	2.9	30.01
8	Wii Play	Wii	2006	Misc	Nintendo	14.03	9.2	2.93	2.85	29.02
9	New Super Mario Bros.	Wii	2009	Platform	Nintendo	14.59	7.06	4.7	2.26	28.62
10	Duck Hunt	NES	1984	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31
11	Nintendogs	DS	2005	Simulation	Nintendo	9.07	11.1	1.93	2.75	24.76
12	Mario Kart DS	DS	2005	Racing	Nintendo	9.81	7.57	4.13	1.92	23.42
13	Pokemon Gold/Pokemon Silver	GB	1999	Role-Playing	Nintendo	9.6	18.7	2.0	71.23	1
14	Wii Fit	Wii	2007	Sports	Nintendo	8.94	8.03	3.6	2.15	22.72
15	Wii Fit Plus	Wii	2009	Sports	Nintendo	9.09	8.59	2.53	1.79	22
16	Kinect Adventures!	X360	2010	Misc	Microsoft Game Studios	14.97	4.94	0.24	1.67	21.82
17	Grand Theft Auto V	PS3	2013	Action	Take-Two Interactive	7.01	9.27	0.97	4.14	21.4
18	Grand Theft Auto: San Andreas	PS2	2004	Action	Take-Two Interactive	9.43	0.4	0.41	10.57	20.81
19	Super Mario World	SNES	1990	Platform	Nintendo	12.78	3.75	3.54	0.55	20.61
20	Brain Age: Train Your Brain in Minutes a Day	DS	2005	Misc	Nintendo	4.75	9.26	4.16	2.05	20.22
21	Pokemon Diamond/Pokemon Pearl	DS	2006	Role-Playing	Nintendo	6.42	4.52	6.04	1.37	18.36
22	Super Mario Land	GB	1989	Platform	Nintendo	10.83	2.71	4.18	0.42	18.14
23	Super Mario Bros.	NES	1988	Platform	Nintendo	9.54	3.44	3.84	0.46	17.28

Hình 2: Tập dữ liệu Explore Video Games Sales

2.2.2 Mục Tiêu Phân Tích Dữ Liệu

Tập dữ liệu Explore Video Games Sales sẽ được sử dụng để:

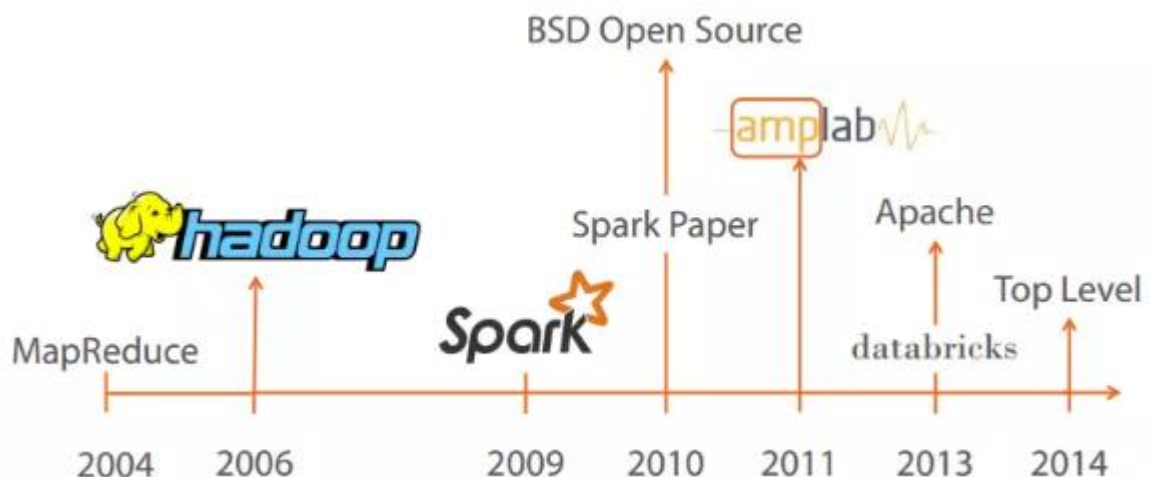
- **Phân tích xu hướng doanh số bán hàng** theo các yếu tố như thể loại, nền tảng, nhà phát hành và năm phát hành.
- **Trích xuất các đặc trưng quan trọng** ảnh hưởng đến doanh số bán hàng.
- **Huấn luyện mô hình dự đoán doanh số bán hàng** bằng các thuật toán học máy như Hồi quy tuyến tính (Linear Regression) hoặc các mô hình nâng cao hơn.

- **Đánh giá hiệu suất mô hình** dựa trên các chỉ số thống kê và thuật toán phù hợp.

2.2.3 Cách sử dụng tập dữ liệu

Tập dữ liệu sẽ được sử dụng để phân tích xu hướng doanh số theo các yếu tố khác nhau như thể loại, hệ máy và khu vực. Các đặc trưng quan trọng sẽ được trích xuất và xử lý để huấn luyện mô hình dự đoán doanh số. Việc đánh giá mô hình sẽ dựa trên dữ liệu này để đảm bảo tính chính xác của dự đoán. Việc sử dụng tập dữ liệu này sẽ giúp chúng tôi có một cái nhìn tổng quan về doanh số bán trò chơi trên thị trường, đồng thời áp dụng các phương pháp phân tích dữ liệu lớn để tạo ra những dự đoán có ý nghĩa.

2.3 Giới thiệu Apache Spark



Hình 3: Quá trình phát triển của spark

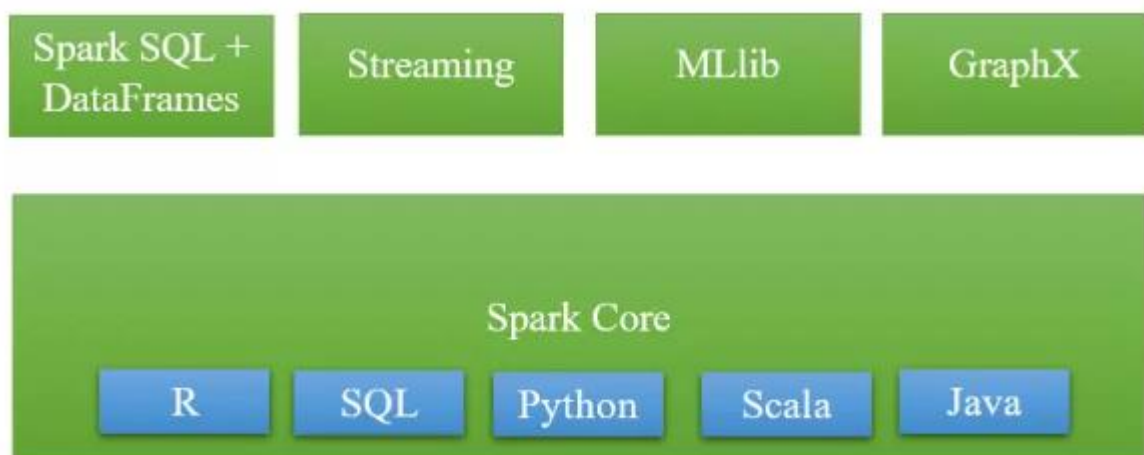
Apache Spark là một framework mã nguồn mở tính toán cụm, được phát triển sơ khởi vào năm 2009 bởi AMPLab. Sau này, Spark đã được trao cho Apache Software Foundation vào năm 2013 và được phát triển cho đến nay.

Tốc độ xử lý của Spark có được do việc tính toán được thực hiện cùng lúc trên nhiều máy khác nhau. Đồng thời việc tính toán được thực hiện ở bộ nhớ trong (in-memories) hay thực hiện hoàn toàn trên RAM.

Spark cho phép xử lý dữ liệu theo thời gian thực, vừa nhận dữ liệu từ các nguồn khác nhau đồng thời thực hiện ngay việc xử lý trên dữ liệu vừa nhận được (Spark Streaming).

Spark không có hệ thống file của riêng mình, nó sử dụng hệ thống file khác như: HDFS, Cassandra, S3. Spark hỗ trợ nhiều kiểu định dạng file khác nhau (text, csv, json...) đồng thời nó hoàn toàn không phụ thuộc vào bất cứ một hệ thống file nào.

2.3.1 Thành phần của Spark



Hình 4: Các thành phần của spark

Apache Spark gồm có 5 thành phần chính: Spark Core, Spark Streaming, Spark SQL, MLlib và GraphX, trong đó:

- **Spark Core** là nền tảng cho các thành phần còn lại và các thành phần này muốn khởi chạy được thì đều phải thông qua Spark Core do Spark Core đảm nhận vai trò thực hiện công việc tính toán và xử lý trong bộ nhớ (In-memory computing) đồng thời nó cũng tham chiếu các dữ liệu được lưu trữ tại các hệ thống lưu trữ bên ngoài.

- **Spark SQL** cung cấp một kiểu data abstraction mới (SchemaRDD) nhằm hỗ trợ cho cả kiểu dữ liệu có cấu trúc (structured data) và dữ liệu nửa cấu trúc (semi-structured data – thường là dữ liệu dữ liệu có cấu trúc nhưng không đồng nhất và cấu trúc của dữ liệu phụ thuộc vào chính nội dung của dữ liệu ấy). **Spark SQL** hỗ trợ DSL (Domain-specific language) để thực hiện các thao tác trên DataFrames bằng ngôn ngữ Scala, Java hoặc Python và nó cũng hỗ trợ cả ngôn ngữ SQL với giao diện command-line và ODBC/JDBC server.
- **Spark Streaming** được sử dụng để thực hiện việc phân tích stream bằng việc coi stream là các mini-batches và thực hiện kỹ thuật RDD transformation đối với các dữ liệu mini-batches này. Qua đó cho phép các đoạn code được viết cho xử lý batch có thể được tận dụng lại vào trong việc xử lý stream, làm cho việc phát triển lambda architecture được dễ dàng hơn. Tuy nhiên điều này lại tạo ra độ trễ trong xử lý dữ liệu (độ trễ chính bằng mini-batch duration) và do đó nhiều chuyên gia cho rằng Spark Streaming không thực sự là công cụ xử lý streaming giống như Storm hoặc Flink.
- **MLlib** (Machine Learning Library): MLlib là một nền tảng học máy phân tán bên trên Spark do kiến trúc phân tán dựa trên bộ nhớ. Theo các so sánh benchmark Spark MLlib nhanh hơn 9 lần so với phiên bản chạy trên Hadoop (Apache Mahout).
- **GrapX**: Grapx là nền tảng xử lý đồ thị dựa trên Spark. Nó cung cấp các Api để diễn tả các tính toán trong đồ thị bằng cách sử dụng Pregel Api.

2.3.2 Những điểm nổi bật của Spark

Xử lý dữ liệu: Spark xử lý dữ liệu theo lô và thời gian thực

Tính tương thích: Có thể tích hợp với tất cả các nguồn dữ liệu và định dạng tệp được hỗ trợ bởi cụm Hadoop.

Hỗ trợ ngôn ngữ: hỗ trợ Java, Scala, Python và R.

Phân tích thời gian thực:

- Apache Spark có thể xử lý dữ liệu thời gian thực tức là dữ liệu đến từ các luồng sự kiện thời gian thực với tốc độ hàng triệu sự kiện mỗi giây. Ví dụ: Data Twitter chẳng hạn hoặc lượt chia sẻ, đăng bài trên Facebook. Sức mạnh Spark là khả năng xử lý luồng trực tiếp hiệu quả.
- Apache Spark có thể được sử dụng để xử lý phát hiện gian lận trong khi thực hiện các giao dịch ngân hàng. Đó là bởi vì, tất cả các khoản thanh toán trực tuyến được thực hiện trong thời gian thực và chúng ta cần ngừng giao dịch gian lận trong khi quá trình thanh toán đang diễn ra.

2.3.3 So sánh spark với hadoop

Hadoop là một khung nguồn mở có Hệ thống tệp phân tán Hadoop (HDFS) làm kho lưu trữ, có YARN là phương pháp quản lý tài nguyên điện toán được sử dụng bởi các ứng dụng khác nhau và triển khai mô hình lập trình MapReduce như một công cụ thực thi. Trong triển khai Hadoop điển hình, các công cụ thực thi khác nhau cũng được triển khai như Spark, Tez và Presto.

Spark là một khung nguồn mở tập trung vào truy vấn tương tác, máy học và khối lượng công việc theo thời gian thực. Spark không có hệ thống lưu trữ riêng, nhưng chạy phân tích trên các hệ thống lưu trữ khác như HDFS hoặc các kho dữ liệu phổ biến khác như Amazon Redshift, Amazon S3, Couchbase, Cassandra và các kho dữ liệu khác. Spark trên Hadoop tận dụng YARN để chia sẻ cụm và tập dữ liệu chung như các công cụ Hadoop khác, việc này đảm bảo mức độ dịch vụ và phản hồi nhất quán.

2.3.4 Ứng dụng của spark

Spark là một hệ thống xử lý phân tán đa mục đích được sử dụng cho khối lượng công việc có dữ liệu lớn. Hệ thống này đã được triển khai trong mọi loại trường hợp sử dụng dữ liệu lớn để phát hiện các mẫu và cung cấp thông tin chuyên sâu theo thời gian thực. Các ví dụ về trường hợp sử dụng bao gồm:

Dịch vụ tài chính

Spark được sử dụng trong ngân hàng để dự đoán tỷ lệ khách hàng rời bỏ và đề xuất các sản phẩm tài chính mới. Trong ngân hàng đầu tư, Spark được sử dụng để phân tích giá cổ phiếu nhằm dự đoán xu hướng trong tương lai.

Chăm sóc sức khỏe

Spark được sử dụng để xây dựng dịch vụ chăm sóc bệnh nhân toàn diện bằng cách cung cấp dữ liệu cho nhân viên y tế tuyến đầu để phục vụ cho mọi tương tác với bệnh nhân. Spark cũng có thể được sử dụng để dự đoán/đề xuất phương pháp điều trị cho bệnh nhân.

Sản xuất

Spark được sử dụng để loại bỏ thời gian ngừng hoạt động của thiết bị kết nối internet bằng cách đề xuất thời điểm thực hiện bảo trì phòng ngừa.

Bán lẻ

Spark được sử dụng để thu hút và giữ chân khách hàng thông qua các dịch vụ và ưu đãi được cá nhân hóa.

CHƯƠNG 3: KẾT QUẢ XỬ LÝ VÀ PHÂN TÍCH DỮ LIỆU

3.1 Kết quả xử lý

```
library(ggplot2)
library(dplyr)
library(DT)
library(tidyr)
library(caret)
```

Khai báo và sử dụng thư viện cần thiết:

ggplot2: Thư viện hỗ trợ vẽ đồ thị và trực quan hóa dữ liệu.

dplyr: Cung cấp các hàm để xử lý và thao tác dữ liệu một cách dễ dàng, như lọc (filter), nhóm (group_by), sắp xếp (arrange)

DT: Hỗ trợ hiển thị dữ liệu dạng bảng HTML tương tác trong R.

tidyr: Cung cấp các hàm giúp làm sạch và tổ chức lại dữ liệu theo định dạng "tidy data".

caret: Một thư viện mạnh mẽ dùng để xây dựng và đánh giá mô hình Machine Learning, đặc biệt là các mô hình hồi quy và phân loại.

```
# Đọc file
videogamesales <- read.csv("E:/R/vgsales.csv")
# Xử lý dữ liệu
videogamesales <- videogamesales[!(videogamesales$Year %in% c("N/A", "2017", "2020")),]
videogamesales <- videogamesales %>% gather(Region, Revenue, 7:10)
videogamesales$Region <- factor(videogamesales$Region)
colnames(videogamesales) <- gsub("\\.", "", colnames(videogamesales))
videogamesales$Global_sales <- as.numeric(gsub(";", "", videogamesales$Global_sales))
```

Đọc dữ liệu từ file "E:/R/vgsales.csv" vào biến videogamesales.

Loại bỏ các dòng có giá trị Year là "N/A", "2017" hoặc "2020".

Sử dụng `gather()` từ thư viện **tidyr** để biến đổi dữ liệu từ dạng rộng (nhiều cột) sang dạng dài. Ở đây, cột từ vị trí **7 đến 10** (có thể là doanh số theo khu vực như `NA_Sales`, `EU_Sales`) được gộp thành hai cột mới:

- **Region:** Chứa tên khu vực.
- **Revenue:** Chứa giá trị doanh thu tương ứng.

Chuyển đổi cột **Region** thành kiểu **factor** để xử lý dữ liệu phân loại

Loại bỏ dấu “.” trong tên các cột để tránh lỗi khi truy xuất dữ liệu

Loại bỏ ký tự ";" trong cột `Global_Sales`, sau đó chuyển đổi sang kiểu số (numeric).

Dữ liệu sau khi được xử lý:

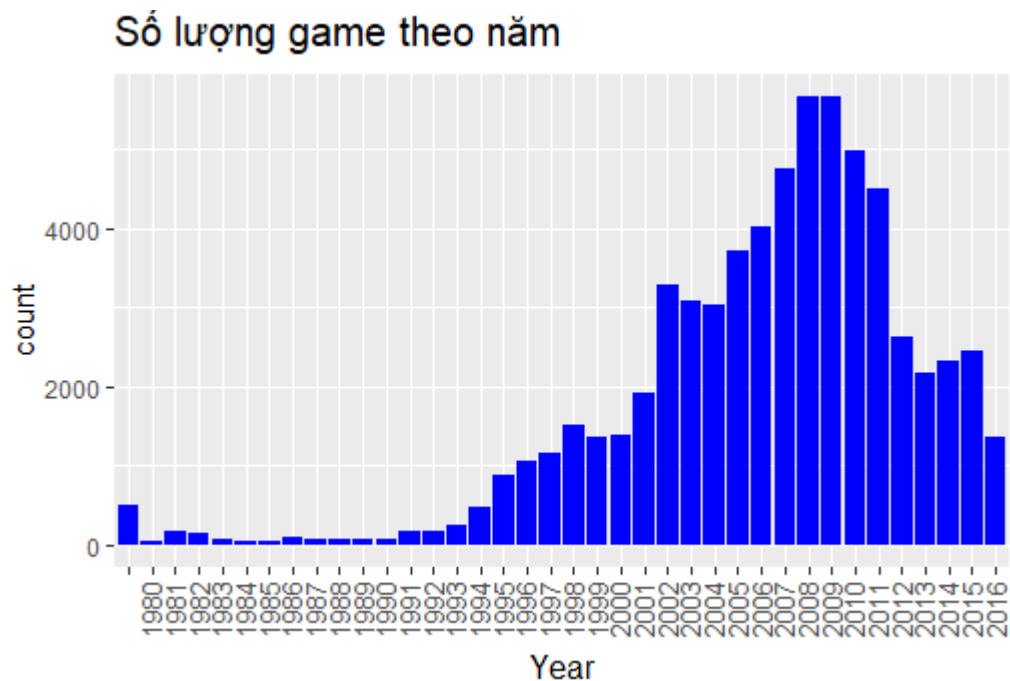
	Year	Genre	Publisher	Platform	Global_Sales
1	2006	Sports	Nintendo	Wii	82.74
2	1985	Platform	Nintendo	NES	40.24
3	2008	Racing	Nintendo	Wii	35.82
4	2009	Sports	Nintendo	Wii	33.00
5	1996	Role-Playing	Nintendo	GB	31.37
6	1989	Puzzle	Nintendo	GB	30.26
7	2006	Platform	Nintendo	DS	30.01
8	2006	Misc	Nintendo	Wii	29.02
9	2009	Platform	Nintendo	Wii	28.62
10	1984	Shooter	Nintendo	NES	28.31
11	2005	Simulation	Nintendo	DS	24.76
12	2005	Racing	Nintendo	DS	23.42
13	1999	Role-Playing	Nintendo	GB	23.10
14	2007	Sports	Nintendo	Wii	22.72
15	2009	Sports	Nintendo	Wii	22.00
16	2010	Misc	Microsoft Game Studios	X360	21.82
17	2013	Action	Take-Two Interactive	PS3	21.40
18	2004	Action	Take-Two Interactive	PS2	20.81

```
mytheme_1 <- function() {
  return(theme(axis.text.x = element_text(angle = 90, size = 10, vjust = 0.4), plot.title = element_text(size = 15, vjust = 2), axis.title.x = element_text(size = 12, vjust = 0.4))
}
mytheme_2 <- function() {
  return(theme(axis.text.x = element_text(size = 10, vjust = 0.4), plot.title = element_text(size = 15, vjust = 2), axis.title.x = element_text(size = 12, vjust = 0.4))
}
```

Định nghĩa hai hàm mytheme_1 và mytheme_2 để tùy chỉnh giao diện biểu đồ trong ggplot2.

```
# Biểu đồ số lượng game phát hành theo năm
ggplot(videogamesales, aes(Year)) + |
  geom_bar(fill = "blue") +
  mytheme_1() +
  ggtitle("Số lượng game theo năm")
```

Sử dụng thư viện ggplot2 để vẽ biểu đồ số lượng game phát hành theo năm



Hình 5: Biểu đồ số lượng game theo năm

```
revenue_by_year <- videogamesales %>%
  group_by(Year) %>%
  summarize(Revenue = sum(Global_Sales))

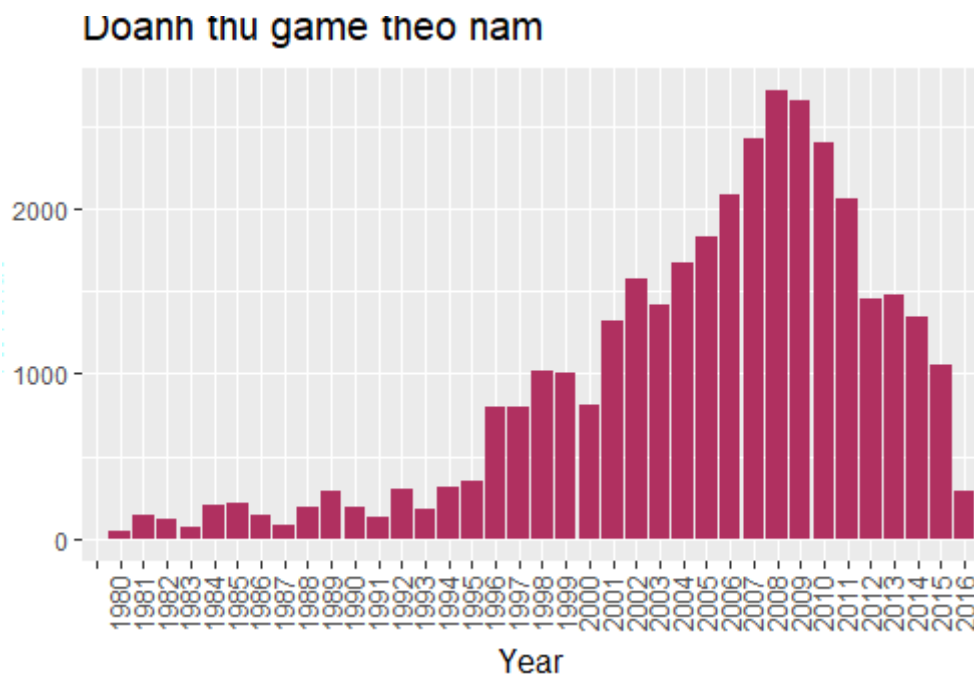
# Biểu đồ doanh thu game theo năm
ggplot(revenue_by_year, aes(Year, Revenue)) +
  geom_bar(fill = "maroon", stat = "identity") +
  mytheme_1() +
  ggtitle("Doanh thu game theo năm")
|
```

group_by(Year): Nhóm dữ liệu theo năm phát hành game.

summarize(Revenue = sum(Global_Sales)): Tính tổng doanh thu (Global_Sales) của tất cả các game trong từng năm.

Kết quả là một dataframe revenue_by_year chứa 2 cột: **Year** (năm) và **Revenue** (tổng doanh thu).

Sử dụng thư viện ggplot2 để vẽ biểu đồ doanh thu theo năm



Hình 6: Biểu đồ doanh thu theo năm

```
# Tính doanh thu cho từng loại
top_1 <- videogamesales %>%
  group_by(Year, Genre) %>%
  summarize(Revenue = sum(Global_Sales)) %>%
  top_n(1)

datatable(top_1)

# Biểu đồ thể loại có doanh thu cao nhất theo từng năm
ggplot(top_1, aes(Year, Revenue, fill = Genre)) +
  geom_bar(stat = "identity") +
  ggtitle("Thể loại có doanh thu cao nhất") +
  mytheme_1() +
  theme(legend.position = "top")
```

group_by(Year, Genre): Nhóm dữ liệu theo năm phát hành (Year) và thể loại (Genre).

summarize(Revenue = sum(Global_Sales)):

Tính tổng doanh thu (Revenue) của từng thể loại game trong từng năm.

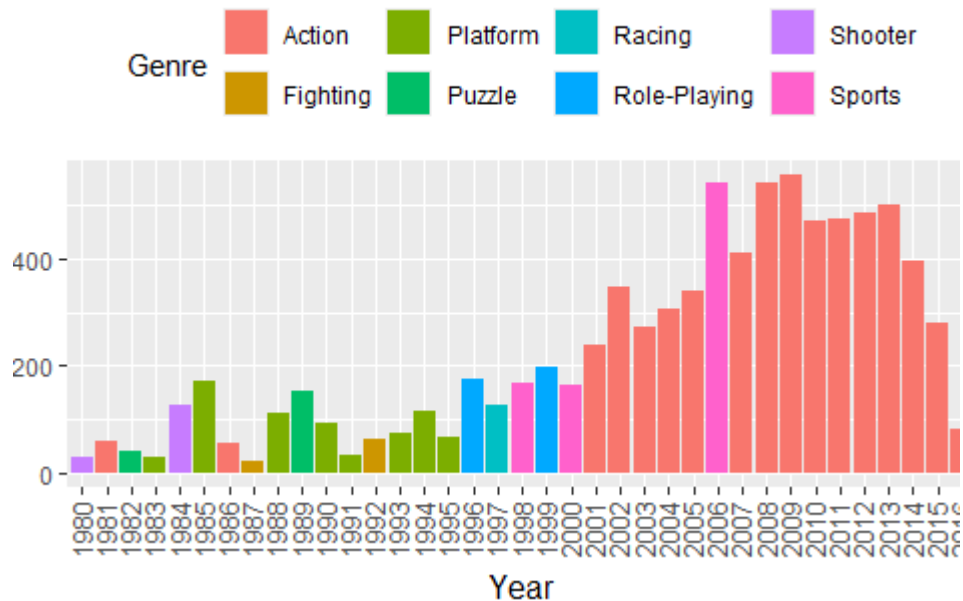
top_n(1):

Lấy thể loại game có doanh thu cao nhất trong mỗi năm.

datatable() từ thư viện DT giúp hiển thị dữ liệu top_1 dưới dạng bảng có thể tương tác.

Sử dụng thư viện ggplot2 để vẽ biểu đồ doanh thu theo năm

Thể loại có doanh thu cao nhất



Hình 7: Biểu đồ thể loại có doanh thu cao nhất

```
# Tính tổng doanh thu của từng game theo từng năm
top_games <- videogamesales %>%
  group_by(Year, Name) %>%
  summarize(Revenue = sum(Global_Sales)) %>%
  arrange(desc(Revenue)) %>%
  top_n(1)

datatable(top_games)

# Biểu đồ game có doanh thu cao nhất từng năm
ggplot(top_games, aes(Year, Revenue, fill = Name)) +
  geom_bar(stat = "identity") +
  mytheme_1() +
  ggtitle("Total Games by Revenue each year") +
  theme(legend.position = "top")
```

`group_by(Year, Name)`: Nhóm dữ liệu theo năm phát hành (Year) và tên game (Name).

`summarize(Revenue = sum(Global_Sales))`:

Tính tổng doanh thu (Revenue) của từng game trong từng năm.

`arrange(desc(Revenue))`:

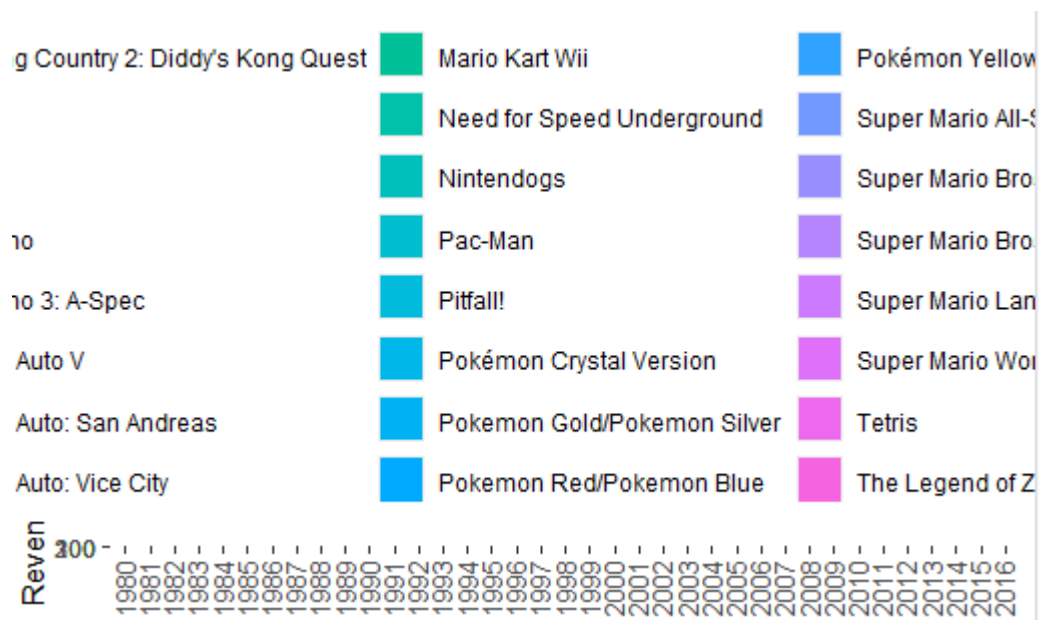
Sắp xếp dữ liệu theo doanh thu giảm dần.

top_n(1):

Lấy game có doanh thu cao nhất trong mỗi năm.

datatable() từ thư viện DT giúp hiển thị dữ liệu top_game dưới dạng bảng có thể tương tác.

Sử dụng thư viện ggplot2 để vẽ biểu đồ game có doanh thu cao nhất từng năm



Hình 8: Biểu đồ game có doanh thu cao nhất


```
# số lượng nhà phát triển
length(unique(videogamesales$Publisher))
# Lấy 10 nhà phát hành có số lượng lớn nhất
by_publishers <- videogamesales %>% group_by(Publisher) %>% summarize(Total = n()) %>% arrange(
  by_publishers$Percentage <- by_publishers$Total/dim(videogamesales)[1] * 100
  by_publishers$Publisher <- factor(by_publishers$Publisher)

# Cột đầu số lượng game, cột 2 chiếm bao nhiêu %
datatable(by_publishers, filter = "none")

# Biểu đồ những nhà phát hành có lượng game lớn
ggplot(by_publishers, aes(reorder(Publisher, Total), Total, fill = Publisher)) +
  geom_bar(stat = "identity") +
  ggtitle("Top những nhà phát hành có số lượng lớn nhất") +
  theme(legend.position = "none") +
  xlab("Publisher") +
  mytheme_2() +
  coord_flip()
```

unique(videogamesales\$Publisher): Lấy danh sách các nhà phát hành không trùng lặp.

length(): Đếm số lượng nhà phát hành.

group_by(Publisher): Nhóm dữ liệu theo nhà phát hành.

summarize(Total = n()): Đếm số lượng game của từng nhà phát hành.

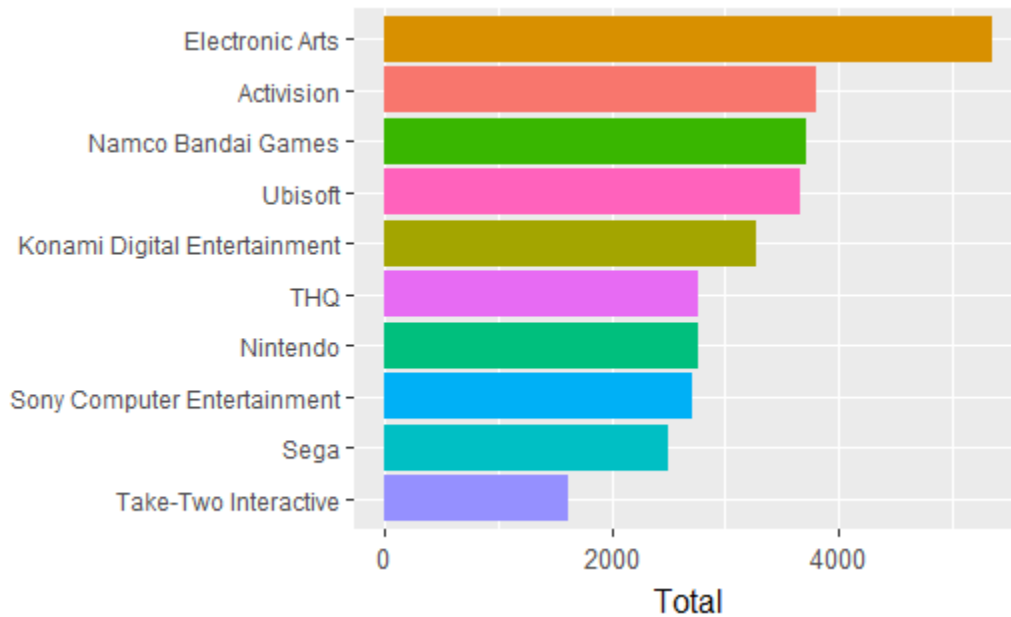
arrange(desc(Total)): Sắp xếp theo số lượng game giảm dần

dim(videogamesales)[1]: Lấy tổng số dòng (tổng số game).

Chia số lượng game của mỗi nhà phát hành cho tổng số game rồi nhân 100 để ra tỷ lệ phần trăm.

Dùng datatable() từ thư viện **DT** để hiển thị bảng dữ liệu.

filter = "none": Không có ô tìm kiếm trong bảng



Hình 9: Biểu đồ những nhà phát hành có số lượng game lớn

3.2 Phân tích dữ liệu

Bạn em sử dụng mô hình hồi quy tuyến tính để dự đoán doanh thu của game trong những năm tới

3.2.1 Xử lý dữ liệu

```
videogamesales <- videogamesales %>%
  select(Year, Genre, Publisher, Platform, Global_Sales) %>%
  na.omit()

videogamesales <- videogamesales %>%
  filter(Year != "N/A") %>%
  mutate(Year = as.numeric(Year),
         Global_Sales = as.numeric(Global_Sales),
         Genre = factor(Genre),
         Publisher = factor(Publisher),
         Platform = factor(Platform))
```

select(Year, Genre, Publisher, Platform, Global_Sales): Chỉ giữ lại các cột quan trọng.

na.omit(): Loại bỏ các dòng có giá trị NA trong bất kỳ cột nào.

filter(Year != "N/A"): Loại bỏ các dòng có Year là "N/A" (giá trị không hợp lệ).

mutate(Year = as.numeric(Year)): Chuyển Year thành kiểu số (numeric).

mutate(Global_Sales = as.numeric(Global_Sales)): Chuyển Global_Sales thành kiểu số.

mutate(Genre = factor(Genre)): Chuyển Genre thành kiểu phân loại (factor).

mutate(Publisher = factor(Publisher)): Chuyển Publisher thành kiểu phân loại.

mutate(Platform = factor(Platform)): Chuyển Platform thành kiểu phân loại.

3.2.2 Chia dữ liệu thành tập huấn luyện và kiểm tra

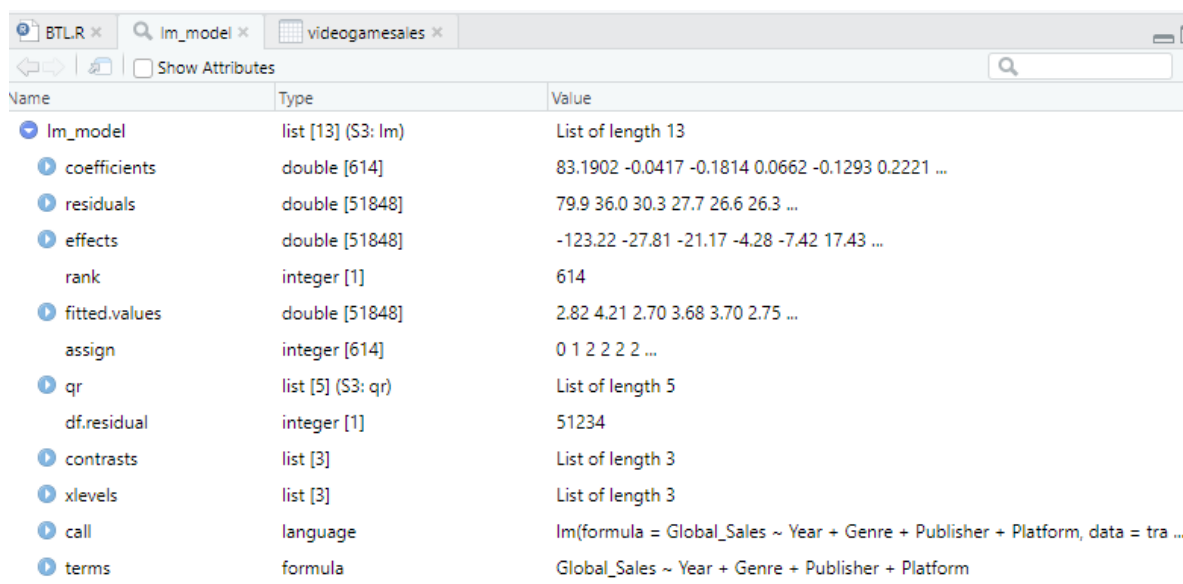
```
set.seed(123)
# Chia dữ liệu thành 80% train và 20% test
train_index <- createDataPartition(videogamesales$Global_Sales, p = 0.8, list = FALSE)

train_data <- videogamesales[train_index, ]
test_data <- videogamesales[-train_index, ]
```

3.2.3 Huấn luyện mô hình

```
# Xây dựng mô hình hồi quy tuyến tính
lm_model <- lm(Global_Sales ~ Year + Genre + Publisher + Platform, data = train_data)

# Xem tóm tắt mô hình
summary(lm_model)
```



The screenshot shows the RStudio interface with the 'lm_model' object selected in the Environment pane. The object is a linear model (lm) with the following attributes:

Name	Type	Value
lm_model	list [13] (S3: lm)	List of length 13
coefficients	double [614]	83.1902 -0.0417 -0.1814 0.0662 -0.1293 0.2221 ...
residuals	double [51848]	79.9 36.0 30.3 27.7 26.6 26.3 ...
effects	double [51848]	-123.22 -27.81 -21.17 -4.28 -7.42 17.43 ...
rank	integer [1]	614
fitted.values	double [51848]	2.82 4.21 2.70 3.68 3.70 2.75 ...
assign	integer [614]	0 1 2 2 2 2 ...
qr	list [5] (S3: qr)	List of length 5
df.residual	integer [1]	51234
contrasts	list [3]	List of length 3
xlevels	list [3]	List of length 3
call	language	lm(formula = Global_Sales ~ Year + Genre + Publisher + Platform, data = tra ...
terms	formula	Global_Sales ~ Year + Genre + Publisher + Platform

3.2.4 Đánh giá mô hình

```
# Dự đoán giá trị trên tập kiểm tra
predicted_revenue <- predict(lm_model, newdata = test_data)

# Đánh giá mô hình bằng cách so sánh với giá trị thực tế
actual_vs_predicted <- data.frame(Actual = test_data$Global_Sales, Predicted = predicted_r
print(head(actual_vs_predicted))

# Tính sai số RMSE (Sai số chung bình)
rmse <- sqrt(mean((actual_vs_predicted$Actual - actual_vs_predicted$Predicted)^2))
cat("Root Mean Square Error (RMSE):", rmse, "\n")

mean_sales <- mean(test_data$Global_Sales)
cat("Giá trị trung bình của doanh thu game:", mean_sales, "\n")
```

Root Mean Square Error (RMSE): 1.318755

Giá trị trung bình của doanh thu game: 0.5484776

3.2.5 Dự báo doanh thu game

```
future_years <- data.frame(
  Year = 2025:2030,
  Genre = factor(rep("Action", 6), levels = levels(videogamesales$Genre)),
  Publisher = factor(rep("Nintendo", 6), levels = levels(videogamesales$Publisher)),
  Platform = factor(rep("PS4", 6), levels = levels(videogamesales$Platform))
)

# Dự đoán doanh thu game cho các năm 2025 - 2030 (Triệu bản)
future_years$Predicted_Sales <- predict(lm_model, newdata = future_years)

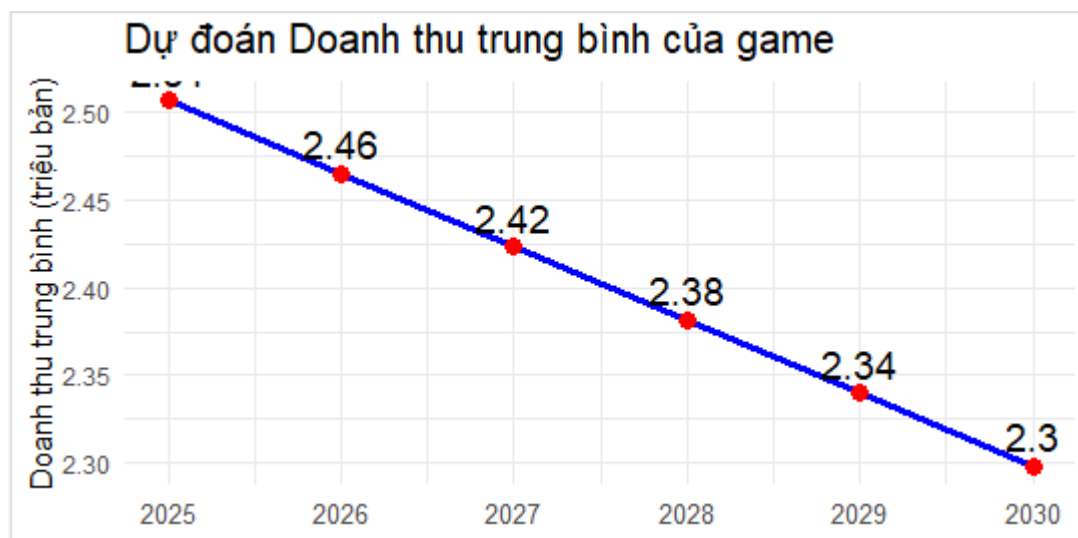
# Xem kết quả dự báo
print(future_years)

# vẽ biểu đồ xu hướng doanh thu game từ 2025 - 2030
ggplot(future_years, aes(x = Year, y = Predicted_Sales)) +
  geom_line(color = "blue", size = 1.2) +
  geom_point(color = "red", size = 3) +
  geom_text(aes(label = round(Predicted_Sales, 2)), vjust = -0.5, size = 5) +
  ggtitle("Dự đoán Doanh thu trung bình của game (2025 - 2030)") +
  xlab("Năm") +
  ylab("Doanh thu trung bình (triệu bản)") +
  scale_x_continuous(breaks = seq(2025, 2030, 1)) +
  theme_minimal() + mytheme_2()
```

Lựa chọn dự đoán từ năm 2025 đến 2030, Thể loại của trò chơi là Action và 6 lần lặp lại vector action của hãng Nintendo và trên PS4.

	Year	Genre	Publisher	Platform
1	2025	Action	Nintendo	PS4
2	2026	Action	Nintendo	PS4
3	2027	Action	Nintendo	PS4
4	2028	Action	Nintendo	PS4
5	2029	Action	Nintendo	PS4
6	2030	Action	Nintendo	PS4

	Year	Genre	Publisher	Platform	Predicted_Sales
1	2025	Action	Nintendo	PS4	2.506628
2	2026	Action	Nintendo	PS4	2.464951
3	2027	Action	Nintendo	PS4	2.423274
4	2028	Action	Nintendo	PS4	2.381597
5	2029	Action	Nintendo	PS4	2.339919
6	2030	Action	Nintendo	PS4	2.298242



Hình 10: Biểu đồ dự đoán doanh thu game

KẾT LUẬN

Trong bài tập lớn này, nhóm em đã nghiên cứu và triển khai các phương pháp phân tích dữ liệu lớn và học máy để dự đoán doanh số bán trò chơi điện tử. Quá trình thực hiện bao gồm thu thập, tiền xử lý, phân tích dữ liệu, xây dựng và đánh giá mô hình. Kết quả thu được giúp nhóm hiểu rõ hơn về các yếu tố quan trọng ảnh hưởng đến doanh số bán hàng và cách ứng dụng các thuật toán học máy để đưa ra dự đoán.

Mặc dù đã đạt được một số kết quả đáng kể, nhưng bài toán này vẫn còn nhiều thách thức như dữ liệu không đồng nhất, biến động của thị trường và sự ảnh hưởng của yếu tố con người trong quyết định mua hàng. Trong tương lai, có thể mở rộng nghiên cứu bằng cách sử dụng dữ liệu thời gian thực, tích hợp các yếu tố đánh giá từ người dùng và cải thiện mô hình bằng các thuật toán tiên tiến hơn.

Nhóm em hy vọng rằng những kết quả trong báo cáo này có thể đóng góp một phần nhỏ vào việc hiểu và dự đoán xu hướng bán hàng trong ngành công nghiệp trò chơi điện tử, đồng thời là nền tảng cho các nghiên cứu và ứng dụng tiếp theo trong lĩnh vực dữ liệu lớn.

DANH MỤC TÀI LIỆU THAM KHẢO