BỘ CÔNG THƯƠNG

TRƯỜNG ĐẠI HỌC KINH TẾ KỸ THUẬT CÔNG NGHIỆP

KHOA KHOA HỌC ỨNG DỤNG

BÁO CÁO TỔNG KẾT
HỌC PHẦN ĐỒ ÁN 2

ĐỀ TÀI: PHÂN TÍCH DỮ LIỆU DOANH THU CỦA CÁC LOẠI TRÒ CHƠI ĐIỆN TỬ PHỔ BIẾN TRÊN THẾ GIỚI

Sinh viên thực hiện:

NGUYỄN TIẾN ANH	DHKL16A2HN	22174600083
ĐẬU THỊ THẢO	DHKL16A2HN	22174600003

Giáo viên giảng dạy: LÊ HẰNG ANH

Hà Nội, 5/2025

BỘ CÔNG THƯƠNG

TRƯỜNG ĐẠI HỌC KINH TẾ KỸ THUẬT CÔNG NGHIỆP

KHOA KHOA HỌC ỨNG DỤNG

BÁO CÁO TỔNG KẾT

HỌC PHẦN: ĐỒ ÁN 2

ĐỀ TÀI: PHÂN TÍCH DỮ LIỆU DOANH THU CỦA CÁC LOẠI TRÒ CHƠI ĐIỆN TỬ PHỔ BIẾN TRÊN THẾ GIỚI

Sinh viên thực hiện:

NGUYỄN TIẾN ANH	DHKL16A2HN	22174600083
ĐẬU THỊ THẢO	DHKL16A2HN	22174600003

Giảng viên giảng dạy: LÊ HẰNG ANH

Hà Nội, 5/2025

PHIẾU ĐĂNG KÝ ĐỀ TÀI

1. Tên đề tài: phân tích dữ liệu doanh thu của các loại trò chơi điện tử phổ biến trên thế giới

2. Thông tin nhóm sinh viên:

Sinh viên 1 (nhóm trưởng):

- Họ tên: Nguyễn Tiến Anh
- Mã sinh viên: 22174600083
- Điện thoại: 0349010818
- Email: ntanh.dhkl16a2hn@sv.uneti.edu.vn Sinh viên 2:

- Họ tên: Đậu Thị Thảo
- Mã sinh viên: 22174600083
- Điện thoại: 0359146138
- Email: dtthao.dhkl16a2hn@sv.uneti.edu.vn

3. Tóm tắt nội dung đề tài:

Trong bối cảnh ngành công nghiệp trò chơi điện tử phát triển mạnh mẽ toàn cầu, việc khai thác và phân tích dữ liệu doanh thu trở nên cần thiết nhằm hiểu rõ xu hướng thị trường. Đề tài này tập trung vào phân tích dữ liệu bán hàng của các trò chơi điện tử phổ biến trên thế giới, sử dụng bộ dữ liệu công khai "vgsales.csv" trên trang web "https://www.kaggle.com/datasets/anandshaw2001/video-game-sales".

Nhóm tiến hành xử lý và làm sạch dữ liệu, sau đó áp dụng kỹ thuật phân tích dữ liệu khám phá (EDA) để thống kê các chỉ số như doanh số toàn cầu, doanh số theo khu vực (Bắc Mỹ, châu Âu, Nhật Bản), thể loại game phổ biến, nền tảng có doanh thu cao và các nhà phát hành nổi bật.

Tiếp theo, các công cụ trực quan hóa như Matplotlib, Seaborn sẽ được sử dụng để biểu diễn doanh thu theo thời gian, so sánh giữa các thể loại và nền tảng, từ đó giúp phát hiện xu hướng và hành vi người chơi theo khu vực.

Mục tiêu đề tài là giúp hiểu rõ bức tranh tổng thể về thị trường game, hỗ trợ đưa ra chiến lược kinh doanh, phát hành và phát triển game hiệu quả.

Hà Nội, ngày… tháng 5 năm 2025

Nhóm trưởng

ĐỀ CƯƠNG CHI TIẾT ĐỀ TÀI

1. Tên đề tài: phân tích dữ liệu doanh thu của các loại trò chơi điện tử phổ biến trên thế giới

2. Mục tiêu đề tài

Mục tiêu chính của đề tài là giúp hiểu rõ hiệu suất kinh doanh của các trò chơi điện tử trên toàn cầu, xác định những thể loại game phổ biến, nền tảng bán chạy, nhà phát hành dẫn đầu và xu hướng doanh thu theo khu vực. Thông qua việc phân tích và trực quan hóa dữ liệu, đề tài hỗ trợ việc đánh giá thị trường, nhận diện các dòng game chủ lực cũng như khu vực tiêu thụ tiềm năng để đề xuất chiến lược phát hành phù hợp.

Đề tài cũng hướng đến sinh viên đang làm quen với Python, rèn luyện kỹ năng xử lý dữ liệu bằng Pandas, trực quan hóa bằng Matplotlib và Seaborn, đồng thời nâng cao tư duy phân tích trong lĩnh vực kinh doanh game. Dữ liệu được sử dụng là file CSV được công khai từ các nguồn dữ liệu tổng hợp quốc tế.

3. Tổng quan tình hình nghiên cứu thuộc lĩnh vực đề tài

Phân tích dữ liệu trong ngành công nghiệp trò chơi điện tử đã và đang trở thành một hướng nghiên cứu quan trọng trong bối cảnh ngành game toàn cầu phát triển với tốc độ nhanh chóng. Nhiều nghiên cứu đã tập trung vào việc sử dụng dữ liệu bán hàng để đánh giá xu hướng tiêu dùng, đo lường hiệu quả phát hành game và xác định các yếu tố ảnh hưởng đến doanh thu, như thể loại, nền tảng, khu vực phát hành và chiến lược marketing. Nổi bật trong số đó là nghiên cứu "A data mining approach to predict success of digital games" – Sifa et al. (2015).

Ở Việt Nam, lĩnh vực phân tích dữ liệu trong ngành công nghiệp trò chơi điện tử còn khá mới và chưa được nghiên cứu sâu rộng như ở các nước phát triển. Phần lớn các nghiên cứu và dự án hiện nay chủ yếu tập trung ở mức cơ bản, mang tính học thuật hoặc đồ án sinh viên trong các ngành khoa học dữ liệu, công nghệ thông tin và marketing. Một số đề tài sinh viên đã khai thác dữ liệu từ các nguồn công khai như VGChartz, SteamDB hoặc từ các kho dữ liệu mở để thực hiện phân tích doanh thu, xu hướng thị trường và hành vi người chơi.

Trong khi đó, các công ty game tại Việt Nam như VNG, Gamota, Funtap, hay Amanotes chủ yếu áp dụng phân tích dữ liệu ở quy mô nội bộ để theo dõi hành vi người dùng, tối ưu chiến lược quảng cáo và cá nhân hóa trải nghiệm game. Tuy nhiên, các kết quả nghiên cứu hiếm khi được công bố công khai.

Dù còn hạn chế, nhưng với xu hướng phát triển mạnh mẽ của ngành công nghiệp game và nhu cầu phân tích dữ liệu ngày càng cao, lĩnh vực này tại Việt Nam hứa hẹn sẽ có nhiều tiềm năng để mở rộng nghiên cứu trong tương lai gần, đặc biệt khi kết hợp với các công nghệ AI và học máy.

4. Nội dung đề tài

4.1. Giới thiệu chung

Trong bối cảnh ngành công nghiệp trò chơi điện tử ngày càng phát triển và cạnh tranh khốc liệt, việc khai thác và phân tích dữ liệu doanh số là chìa khóa để hiểu rõ thị trường và tối ưu hóa chiến lược kinh doanh. Tuy nhiên, nhiều nhà phát hành vẫn chưa tận dụng hiệu quả kho dữ liệu lớn sẵn có. Với mục tiêu tiếp cận thực tiễn và vận dụng kiến thức khoa học dữ liệu, đề tài này tập trung vào phân tích và trực quan hóa dữ liệu doanh số trò chơi điện tử trên toàn cầu, qua đó đưa ra những nhận định về xu hướng tiêu dùng, sản phẩm chủ lực và đề xuất định hướng phát triển phù hợp.

4.2. Phạm vi và đối tượng nghiên cứu

- Đối tượng nghiên cứu là bộ dữ liệu "vgsales.csv" chứa thông tin về các trò chơi điện tử được phát hành, bao gồm: tên game, nền tảng (platform), năm phát hành, thể loại (genre), nhà phát hành (publisher), doanh số bán hàng theo từng khu vực (Bắc Mỹ, châu Âu, Nhật Bản, các khu vực khác) và doanh số toàn cầu.
- Phạm vi nghiên cứu tập trung vào việc xử lý, phân tích và trực quan hóa dữ liệu doanh thu theo từng khu vực, thể loại, nền tảng và thời gian (theo năm).

4.3. Công cụ và ngôn ngữ sử dụng

- Ngôn ngữ lập trình: Python
- Thư viện: Pandas, Seaborn, Matplotlib,…
- Nguồn dữ liệu: dữ liệu công khai https://www.kaggle.com/datasets/anandshaw2001/video-game-sales
- Môi trường làm việc: Jupyter Notebook và Google Colab

4.4. Kết luận nội dung

Thông qua việc phân tích và trực quan hóa dữ liệu doanh số trò chơi điện tử, đề tài giúp người học hiểu rõ cách triển khai một dự án phân tích dữ liệu thực tế. Đồng thời, kết quả nghiên cứu còn góp phần cung cấp góc nhìn thị trường và hành vi tiêu dùng trong ngành game, phục vụ cho mục tiêu học tập và ứng dụng trong kinh doanh số và truyền thông giải trí.

5. Phương pháp thực hiện

Để thực hiện đề tài, nhóm áp dụng phương pháp phân tích dữ liệu khám phá (Exploratory Data Analysis – EDA) kết hợp với kỹ thuật trực quan hóa dữ liệu nhằm khai thác thông tin ẩn trong bộ dữ liệu doanh số trò chơi điện tử trên toàn cầu. Quá trình thực hiện được triển khai theo các bước chính sau:

Bước 1: Thu thập và xử lý dữ liệu

- Đọc và kiểm tra dữ liệu từ file CSV.
- Xử lý dữ liệu thiếu, dữ liệu trùng lặp, chuẩn hóa cột thời gian, thể loại và nền tảng.

Bước 2: Phân tích dữ liệu

- Thống kê doanh thu theo khu vực và toàn cầu.
- Phân tích doanh thu theo thể loại, nền tảng và nhà phát hành.
- Xác định các trò chơi có doanh số cao nhất trong từng khu vực.
- So sánh xu hướng phát triển game theo từng năm và từng khu vực.

Bước 3: Trực quan hóa dữ liệu

- Biểu đồ đường: xu hướng doanh thu theo thời gian.
- Biểu đồ cột: so sánh doanh thu theo nền tảng và thể loại.
- Biểu đồ tròn: tỷ lệ doanh số theo khu vực.
- Heatmap: doanh thu theo thể loại và năm phát hành.

Bước 4: Đánh giá và đề xuất

- Đưa ra nhận định về thị hiếu người chơi ở từng khu vực.
- Đề xuất hướng phát triển game dự trên thể loại, nền tảng và khu vực vó tiềm năng.

Phương pháp này không chỉ giúp người học làm quen với quy trình phân tích dữ liệu thực tế mà còn nâng cao kỹ năng tư duy số liệu và trình bày báo cáo trực quan. 6. Phân công công việc

STT	Họ và tên	Mã sinh viên	Nội dung công việc được phân công
1	Nguyễn Tiến Anh	22174600083	- Thu thập và làm sạch dữ liệu bán hàng (Excel/CSV) - Phân tích thống kê cơ bản bằng Python (Pandas) - Viết báo cáo phần nội dung, phương pháp thực hiện
2	Đậu Thị Thảo	22174600003	- Thực hiện trực quan hóa dữ liệu bằng Matplotlib/Seaborn/Plotly - Đánh giá kết quả và đề

			xuất hướng cải tiến kinh doanh - Viết báo cáo phần tổng quan, kết luận và trình bày PowerPoint

7. Dự kiến kết quả đạt được

Thông qua quá trình thực hiện đề tài, nhóm dự kiến sẽ thu được các kết quả cụ thể như sau:

- Bộ dữ liệu doanh số trò chơi điện tử đã được xử lý và làm sạch hoàn chỉnh:
  Bộ dữ liệu từ file "vgsales.csv" sẽ được chuẩn hóa, loại bỏ giá trị thiếu, dữ liệu trùng lặp và sai định dạng. Các trường thông tin như năm phát hành, tên trò chơi, nền tảng, thể loại, doanh thu theo khu vực, sẽ được xử lý nhất quán, đảm bảo chất lượng đầu vào cho phân tích.
- Báo cáo phân tích chi tiết thị trường trò chơi điện tử:
  Nhóm sẽ xây dựng báo cáo tổng hợp với các phân tích theo năm, khu vực (Bắc Mỹ, Châu Âu, Nhật Bản, toàn cầu), nền tảng (PlayStation, Xbox, PC, v.v.), thể loại game và nhà phát hành. Qua đó rút ra những xu hướng nổi bật, thị hiếu theo từng khu vực và giai đoạn phát triển của ngành game.
- Hệ thống biểu đồ trực quan dễ hiểu, sinh động:
  Sản phẩm trực quan bao gồm biểu đồ đường thể hiện xu hướng doanh thu theo thời gian, biểu đồ cột so sánh doanh thu theo nền tảng hoặc thể loại, biểu đồ tròn thể hiện tỷ trọng doanh thu theo khu vực, và biểu đồ nhiệt để minh họa mức độ phổ biến theo thể loại và năm phát hành.
- Báo cáo đề xuất định hướng kinh doanh trong ngành game:
  Dựa trên dữ liệu thực tế, nhóm đưa ra các đề xuất như: ưu tiên phát triển game trên nền tảng có doanh số cao, tập trung vào thể loại thịnh hành tại các khu vực có tiềm năng, hoặc định hướng phát hành theo chu kỳ thời gian có doanh thu vượt trội.
- Củng cố kỹ năng chuyên môn về phân tích dữ liệu thực tiễn:
  Các thành viên nắm chắc quy trình phân tích dữ liệu thực tế từ thu thập, làm sạch, phân tích đến trực quan hóa và trình bày báo cáo. Đây là bước đệm vững chắc cho các dự án phân tích dữ liệu chuyên sâu trong tương lai.

Hà Nội, ngày… tháng 5 năm 2025

Nhóm trưởng

MỞ ĐẦU

Trong vài thập kỷ qua, ngành công nghiệp trò chơi điện tử đã vươn mình phát triển vượt bậc, trở thành một thị trường mang tính năng động và tiềm năng. Từ những tựa game cổ điển trên máy chơi game cầm tay đến các siêu phẩm bom tấn trên các thiết bị hiện đại, sự đa dạng về thể loại, nền tảng vả thị trường tiêu thụ đã góp phần định hình nên một ngành công nghiệp đầy tính sáng tạo và cạnh tranh này. Theo một số thống kê của Newzoo hay Statista, doanh thu của ngành công nghiệp này đã vượt qua cả doanh thu từ phim ảnh và âm nhạc, chứng tỏ sử phổ biến và tầm quan trọng của trò chơi điện tử trong đời sống hàng ngày của con người.

Sự phổ biến của các loại trò chơi điện tử khác nhau không chỉ phản ánh xu hướng giải trí mà còn chịu ảnh hưởng bởi nhiều yếu tố khác nhau như văn hóa, công nghệ và sự phát triển của nền tảng. Báo cáo này tập trung vào việc phân tích dữ liệu doanh thu của các loại trò chơi điện tử phổ biến trên thế giới, sử dụng nguồn dữ liệu chính là bộ "vgsales.csv" công khai từ Kaggle: https://www.kaggle.com/datasets/anandshaw2001/video-game-sales. Bộ dữ liệu cung cấp thông tin chi tiết về doanh số bán hàng của hàng ngàn trò chơi được phân loại theo nhiều tiêu chí như thể loại, nền tảng, năm phát hành và khu vực địa lý.

Thông qua việc khám phá và phân tích dữ liệu, báo cáo vừa làm nổi bật sự phân bố doanh thu giữa các thể loại trò chơi, xác định các thị trường khu vực quan trọng đối với từng thể loại, vừa cung cấp một cái nhìn sâu sắc về động lực thị trường. Đồng thời, hỗ trợ các nhà phát hành, nhà đầu tư đưa ra quyết định quan trọng khi cải tiến trò chơi cũ hoặc phát hành trò chơi mới, giúp những người đam mê trò chơi hiểu rõ hơn về cách thức hoạt động của ngành công nghiệp này. Bên cạnh đó, báo cáo có thể đưa ra gợi ý cho những người chơi mới muốn bắt đầu trải nghiệm trò chơi điện tử dựa trên sở thích cá nhân hay văn hóa khu vực hiện tại.

Đồ án bao gồm các phần được phân chương như sau:

Chương 1: Đặt vấn đề (giới thiệu)

Chương 2: Cơ sở lý thuyết

Chương 3: Thực nghiệm

Chương 4: Kết quả đạt được

Chương 5: Kết luận, ưu điểm, nhược điểm, hướng phát triển

MỤC LỤC

MỤC LỤC HÌNH VẼ

CHƯƠNG 1. ĐẶT VẤN ĐỀ

1.1. BỐI CẢNH VÀ TÍNH CẤP THIẾT CỦA ĐỀ TÀI

1.1.1. Bối cảnh của đề tài

Từ những trò chơi cổ xưa như Cờ Vua, Mạt Chược đến những tựa game hiện đại, game đã luôn là nguồn giải trí, kết nối và rèn luyện kỹ năng cho con người. Sự xuất hiện của máy tính PC, điện thoại thông minh, laptop đã làm cho các tực game đi vào "đời sống số" một cách mạnh mẽ và phát triển thành một ngành công nghiệp toàn cầu.

Ngành công nghiệp game đang chứng kiến sự tăng trưởng vượt bậc, theo báo cáo của DataReportal, tính đến quý 3 năm 2023, 83.1% người dùng internet toàn cầu tham gia. Sự phát triển của công nghệ đã mang đến những đột phá mới, từ các thiết bị chơi game đa dạng như điện thoại di động, PC, console đến các nền tảng streaming game. Ngành game cũng có nhiều lĩnh vực liên quan đến nó như thiết kế đồ họa, lập trình, âm nhạc và thậm chí cả văn hóa và nghệ thuật. Vì vậy, ngành game đang trở thành một lựa chọn nghề nghiệp hấp dẫn cho giới trẻ hiện nay.

1.1.2. Tính cấp thiết của đề tài

Trong bối cảnh ngành công nghiệp trò chơi điện tử ngày càng phát triển mạnh mẽ và chiếm lĩnh vị trí hàng đầu trong lĩnh vực giải trí toàn cầu, việc nắm bắt và phân tích dữ liệu doanh thu trở nên vô cùng cần thiết. Điều này giúp các nhà quản lý, nhà phát triển và nhà đầu tư hiểu rõ xu hướng thị trường, đánh giá hiệu quả các loại hình trò chơi khác nhau, cũng như đưa ra các chiến lược phù hợp để tận dụng cơ hội và đối phó với thách thức.

Đồng thời, sự bùng nổ của các nền tảng chơi game trực tuyến, di động và các xu hướng mới như game điện tử thể thao (eSports) đã làm cho thị trường này trở nên sôi động và cạnh tranh ngày càng gay gắt. Do đó, việc phân tích dữ liệu doanh thu không chỉ giúp xác định các yếu tố thành công, mà còn góp phần thúc đẩy sự phát triển bền vững của ngành công nghiệp game trong dài hạn, đáp ứng kịp thời nhu cầu ngày càng đa dạng của người tiêu dùng toàn cầu. Chính vì vậy, đề tài này mang tính cấp thiết để góp phần cung cấp những thông tin chiến lược và dự báo chính xác cho các bên liên quan trong ngành.

1.2. LÝ DO CHỌN ĐỀ TÀI

Xuất phát từ sự quan tâm đặc biệt đối với ngành công nghiệp trò chơi điện tử - một lĩnh vực vừa mang tính giải trí cao, vừa phản ánh rõ nét các xu hướng công nghệ và hành vi tiêu dùng hiện đại, nhóm thực hiện nghiên cứu doanh số tiêu thụ nhằm xác định những thể loại, nền tảng, trò chơi phát triển và sự phân bố của chúng trên từng khu vực khác nhau. Qua đó xác định chính xác hướng đi của ngành công nghiệp trò chơi điện tử bây giờ và trong tương lai.

Phân tích dữ liệu doanh thu không chỉ giúp làm rõ những xu hướng tiêu dùng và các yếu tố dẫn đến thành công của các loại trò chơi khác nhau, mà còn cung cấp những kiến thức thực tiễn hữu ích cho hoạt động kinh doanh trong ngành.

Bên cạnh tiềm năng phát triển, ngành game cũng đang đối mặt với nhiều thách thức như đổi mới công nghệ liên tục, áp lực cạnh tranh và yêu cầu ngày càng cao từ người chơi. Chính vì vậy, nhóm mong muốn tìm hiểu sâu hơn về các yếu tố ảnh hưởng đến doanh thu của các trò chơi phổ biến, đồng thời nâng cao khả năng phân tích dữ liệu và tầm hiểu biết về thị trường game.

Ngoài tính học thuật, đề tài này còn có ý nghĩa thực tiễn khi ứng dụng các phương pháp phân tích dữ liệu hiện đại vào một lĩnh vực năng động, hấp dẫn và đầy tính cạnh tranh. Việc khám phá bộ dữ liệu "vgsales.csv" mang lại cơ hội tiếp cận với nguồn dữ liệu thực tế, từ đó định lượng hóa sự thành công của các thể loại trò chơi, nhận diện thị hiếu người chơi ở từng khu vực và phát hiện các xu hướng thị trường đáng chú ý. Đây không chỉ là cơ hội để tôi vận dụng các kỹ năng học thuật vào thực tiễn, mà còn là nền tảng để đóng góp một phần nhỏ vào việc hiểu rõ hơn bức tranh toàn cảnh của ngành công nghiệp trò chơi điện tử ngày nay.

1.3. MỤC TIÊU NGHIÊN CỨU

Mục tiêu chính của đề tài là phân tích hiệu suất kinh doanh của các trò chơi điện tử trên quy mô toàn cầu, qua đó xác định những thể loại game phổ biến, nền tảng bán chạy nhất, các nhà phát hành dẫn đầu và xu hướng doanh thu theo từng khu vực địa lý. Thông qua việc khai thác và trực quan hóa bộ dữ liệu "vgsales.csv", đề tài hướng đến việc đánh giá tổng thể thị trường game, nhận diện các dòng sản phẩm chủ lực cũng như khu vực tiêu thụ tiềm năng, từ đó đề xuất những chiến lược phát hành hợp lý trong bối cảnh cạnh tranh hiện nay.

Ngoài ra, đề tài còn hỗ trợ các nhà phát hành đưa ra quyết định trước khi phát hành các trò chơi mới hay nâng cấp các trò chơi hiện hành. Cũng như đưa ra những trò chơi gợi ý cho những người mới bắt đầu dựa trên sở thích, văn hóa, khu vực địa lý cụ thể.

Bên cạnh mục tiêu phân tích thị trường, đề tài còn đóng vai trò như một dự án học thuật dành cho sinh viên đang làm quen với ngôn ngữ lập trình Python. Việc áp dụng các thư viện như Pandas cho xử lý dữ liệu, Matplotlib và Seaborn cho trực quan hóa không chỉ giúp củng cố kỹ năng kỹ thuật mà còn góp phần phát triển tư duy phân tích dữ liệu trong lĩnh vực kinh doanh trò chơi điện tử. Bộ dữ liệu được sử dụng là file CSV được công khai từ các nguồn dữ liệu tổng hợp quốc tế, đảm bảo tính khách quan và phù hợp với mục đích nghiên cứu học thuật.

CHƯƠNG II. CƠ SỞ LÝ THUYẾT

2.1. TỔNG QUAN VỀ NGÀNH CÔNG NGHIỆP TRÒ CHƠI ĐIỆN TỬ

2.1.1. Lịch sử phát triển

Ngành công nghiệp trò chơi điện tử đã trải qua một hành trình dài đáng kể để phát triển. Nghiên cứu từ Pelham Smithers cho thấy làn sóng chơi game tiếp tục dâng cao, nhưng các yếu tố thúc đẩy sự phát triển của nó đã thay đổi trong suốt nửa thế kỷ qua.

Năm 1947, Thomas T.Goldsmith Jr. và Estle Ray Mann phát minh ra một máy chơi game đầu tiên gọi là "Cathode Ray Tube Amusement Device", có thể coi là cha đẻ của trò chơi điện tử hiện nay. Năm 1958, Physicist William Higinbotham tạo ra trò chơi điện tử đầu tiên được tạo ra chỉ để giải trí tên là "Tennis for Two" có đồ họa chuyển động trên máy hiện sóng tại Brookhaven National Laboratory. Tất cả chính là nền tảng bắt đầu cho sự hình thành và phát triển của ngành công nghiệp trò chơi điện tử sau này.

Từ những năm 1970 đến 1983 còn được gọi là "kỷ nguyên trước khủng hoảng". Atari – được thành lập từ Nolan Bushnell và Ted Dabney - cái tên không thể không nhắc đến trong lịch sử của ngành công nghiệp game. Chính việc Atari phát hành "Pong" vào năm 1972 đã giúp khởi đầu ngành công nghiệp này. Trò chơi bóng bàn phiên bản arcade là một trò chơi hấp dẫn, thu hút khách hàng và các công ty bắt đầu sản xuất các phiên bản nhái của riêng họ. Tương tự như vậy, chính Atari đã bán một phiên bản game cho máy console của "Pong" vào năm 1975 và cuối cùng là máy chơi game console mang thương hiệu Atari 2600 vào năm 1977, đây là máy chơi game đầu tiên bán được hơn một triệu chiếc.

Trong thời gian ngắn, thị trường arcade bắt đầu ổn định. Sau khi thị trường game suy giảm do có quá nhiều bản sao của "Pong", việc phát hành Space Invaders vào năm 1978 đã phục hồi lại thị trường này. Máy arcade bắt đầu được lắp đặt ở khắp mọi nơi và các loạt game mới như Pac-Man và Donkey Kong đã thúc đẩy sự phát triển hơn nữa loại hình giải trí này. Đến năm 1982, trò chơi arcade đã tạo ra nhiều tiền hơn cả ngành công nghiệp nhạc pop và phòng vé.

Thật không may, ở giai đoạn 1985 – 2000, tốc độ phát triển của ngành công nghiệp game quá nhanh để có thể duy trì. Khi các công ty khác cũng tìm cách tận dụng thị trường, nhiều sản phẩm game và console kém chất lượng đã gây ra tình trạng suy thoái trong toàn ngành. Đồng thời, máy tính cá nhân đã trở thành máy chơi game mới nhất, đặc biệt là với sự ra mắt của Commodore 64 vào năm 1982. Đó là một dấu hiệu xác định kỷ nguyên lịch sử của ngành công nghiệp game: một cuộc chạy đua công nghệ. Trong những năm sau đó, Nintendo đã phát hành loại máy console Nintendo Entertainment System (NES) vào năm 1985 (phát hành tại Nhật Bản với tên gọi Famicom), ưu tiên các trò chơi chất lượng cao và tiếp thị nhất quán để chiếm lại

thị trường. Nhờ những trò chơi như Duck Hunt, Excitebike, và sự ra đời của Mario trong Super Mario Bros, thành công rực rỡ của NES đã hồi sinh thị trường console.

Năm 1988, công ty arcade Sega tham gia vào cuộc cạnh tranh với console Sega Mega Drive (được phát hành với tên gọi Genesis ở Bắc Mỹ) và sau đó là thiết bị cầm tay Game Gear, đặt trọng tâm tiếp thị vào sức mạnh xử lý. Nhà sản xuất điện tử Sony đã phát hành PlayStation vào năm 1994, sử dụng đĩa CD-ROM thay vì băng điện tử để nâng cao dung lượng lưu trữ cho các trò chơi riêng lẻ. Nó trở thành console đầu tiên trong lịch sử bán được hơn 100 triệu đơn vị và trọng tâm đặt vào định dạng phần mềm đã tiếp tục với các thế hệ console tiếp theo: PlayStation 2 (DVD) và PlayStation 3 (Blu-ray).

Ngay cả Microsoft cũng nhận ra tầm quan trọng của việc chơi game trên PC và phát triển API DirectX để hỗ trợ lập trình trò chơi. Thương hiệu "X" đó sau này đã giúp công ty thâm nhập vào thị trường console với sản phẩm Xbox.

Tuy nhiên, chính sự trỗi dậy của internet và điện thoại di động đã đưa ngành công nghiệp game từ hàng chục tỷ lên hàng trăm tỷ USD doanh thu. Năm 2001, Microsoft ra mắt nền tảng trò chơi trực tuyến Xbox Live với phí đăng ký hàng tháng, cho phép người chơi truy cập vào các dịch vụ trò chuyện thoại và kết nối nhiều người chơi. Nó nhanh chóng trở thành tính năng cần phải có đối với người tiêu dùng.

Trong khi đó trên PC, Blizzard đã khai thác thị trường đăng ký trực tuyến nhiều người chơi (MMO) với bản phát hành World of Warcraft năm 2004, đạt mức đỉnh điểm hơn 14 triệu người đăng ký trả tiền hàng tháng. Nintendo tiếp tục nắm giữ thị trường thiết bị cầm tay với những bản Game Boy cập nhật, còn Nokia và BlackBerry đã thử tích hợp các ứng dụng trò chơi vào điện thoại của họ.

Nhưng chính iPhone của Apple đã củng cố quá trình chuyển đổi trò chơi sang nền tảng di động. Việc công ty này phát hành App Store cho điện thoại thông minh của mình (theo sau là cửa hàng riêng của Google dành cho thiết bị Android) đã mở đường cho các nhà phát triển ứng dụng tạo ra các trò chơi miễn phí, trả phí và trả tiền theo tính năng để phục vụ thị trường đại chúng.

Với sự gia tăng của điện thoại thông minh, trò chơi trên mạng xã hội và dịch vụ streaming, họ đang đi đúng hướng. Có hơn 2,7 tỷ người chơi game trên toàn thế giới vào năm 2020 và cách họ chọn tiêu tiền của mình sẽ tiếp tục định hình lịch sử của ngành công nghiệp game.

2.1.2. Xu hướng hiện tại

Trong những năm gần đây, ngành game thế giới đang chứng kiến nhiều xu hướng nổi bật phản ánh sự đổi mới và phát triển không ngừng của lĩnh vực này. Các xu hướng công nghệ như thực tế ảo (AR), thực tế tăng cường (VR), thực tế hỗn hợp (MR) đã thúc đẩy sự phổ biến của game di động và phát triển của game 3D. Dự kiến,

thị trường game sẽ tiếp tục tăng trưởng và đạt mức 583.69 tỷ USD vào năm 2030 (Nguồn: Grand View Research).

Khi nói đến việc mang lại trải nghiệm chơi game chân thực dưới góc nhìn thứ nhất, công nghệ AR và VR gần như không có đối thủ thay thế. Sự xuất hiện của Pokemon Go vào năm 2016 đã nhanh chóng biến thể loại trò chơi AR trở nên quen thuộc, đánh dấu một bước ngoặt lớn trong ngành công nghiệp game. Dự báo thị trường trò chơi AR và VR sẽ đạt mức 11,0 tỷ USD vào năm 2026, với tốc độ tăng trưởng kép hàng năm (CAGR) 18,5% trong giai đoạn 2021–2026, theo số liệu từ ARC Industry. Sự tích hợp của công nghệ AR/VR vào thiết bị di động và các thiết bị đeo thông minh, cùng với nhu cầu về trải nghiệm chơi game chân thực như bước vào thế giới ảo, đã thúc đẩy sự phát triển mạnh mẽ của loại hình trò chơi này.

Xu hướng chơi game đa nền tảng (Cross-Platform) đang làm thay đổi cách chúng ta trải nghiệm trò chơi. Nhờ những tiến bộ trong công nghệ đám mây và các công cụ phát triển game, việc tạo ra các tựa game có thể chơi được trên nhiều thiết bị khác nhau đã trở nên khả thi hơn. Điều này mở ra cơ hội cho các nhà phát triển game tiếp cận được với một lượng khán giả rộng lớn hơn và tạo ra những trải nghiệm chơi game liền mạch hơn. Tuy nhiên, để đạt được điều này, các nhà phát triển vẫn cần phải đối mặt với những thách thức như tối ưu hóa hiệu năng và bảo mật dữ liệu.

Một xu hướng vẫn còn phát triển nữa là game thể thao (Fitness Gaming). Trò chơi điện tử đã vượt qua định kiến là chỉ dành cho những người lười biếng. Trên thế giới, đã xuất hiện các tựa game thể dục, thể thao là một công cụ giúp người chơi đổ mồ hôi và đốt cháy calo, hỗ trợ lấy lại vóc dáng. Với hiệu quả rõ rệt, thể loại game này ngày càng thu hút nhiều người, đặc biệt sau đại dịch Covid-19, người ta càng chú trọng đến việc chăm sóc sức khỏe thể chất lẫn tinh thần. Nhờ sự phát triển đó, ngách game thể thao được dự báo sẽ là một "mỏ vàng" mới trong ngành công nghiệp game – một trong những xu hướng nổi bật của game 3D vào năm 2024.

Đối với xu hướng game PC, theo dự báo số liệu từ Vantage Market Research, thị trường game PC trên toàn thế giới ước tính sẽ cán mốc 31.52 tỷ USD vào năm 2028, bất chấp những lo ngại rằng game PC đang dần mất vị thế. Tuy nhiên, sự cạnh tranh đang ngày càng gay gắt khi xuất hiện nhiều nền tảng mới như Origin, UPlay, và Epic Games Store. Ấn tượng hơn cả, Roblox nổi lên như một xu hướng mới (mặc dù đã ra mắt từ năm 2006), đây không chỉ là nền tảng chơi game mà còn là công cụ hỗ trợ người dùng, đặc biệt là trẻ em, trong việc sáng tạo và phát triển game 3D. Roblox cung cấp môi trường an toàn cho việc học lập trình, thiết kế, và kinh doanh, đồng thời tích hợp trên nhiều thiết bị như PC, máy tính bảng, điện thoại thông minh và Xbox One.

Một xu hướng mới nổi trong những năm gần đây là trò chơi thể thao điện tử (eSports). Vào những năm 1990, việc kiếm sống từ trò chơi điện tử có vẻ là điều không tưởng. Thế nhưng, giờ đây, với sự bùng nổ của NFT và eSports, cả cá nhân lẫn doanh nghiệp đều đang thu về những khoản lợi nhuận khổng lồ. Thể thao điện tử đã

vươn lên trở thành một ngành công nghiệp giải trí toàn cầu, nơi các game thủ chuyên nghiệp thi đấu và cạnh tranh ở cấp độ cao nhất. Sự phổ biến của eSports ngày càng tăng, thu hút sự quan tâm của các tập đoàn lớn, kể cả những đơn vị không liên quan trực tiếp đến game. Trong tương lai, eSports có thể được công nhận là một môn thể thao chính thức tại Thế vận hội Olympic, một viễn cảnh đang nhận được sự ủng hộ mạnh mẽ từ cộng đồng khán giả toàn cầu.

Tóm lại, xu hướng phát triển đối với ngành game hiện tại chủ yếu xoay quanh những điểm như tập trung vào trải nghiệm người dùng, ứng dụng công nghệ cao, có tính tương tác và kết nối xã hội, có thể phát triển theo xu hướng toàn cầu hóa, đồng thời phải thích nghi với sự phát triển của công nghệ mới và mang tính đa dạng về thể loại, mục đích chơi. Tất cả nhà phát hành đều hướng tới việc mang lại trải nghiệm chơi game hấp dẫn, đa dạng và phù hợp với nhu cầu của người chơi, từ giải trí, rèn luyện thể lực đến cạnh tranh chuyên nghiệp. Các loại hình game này đều dựa trên công nghệ tiên tiến như đồ họa cao cấp, cảm biến, kết nối internet tốc độ cao và các phần mềm phần cứng hiện đại để nâng cao chất lượng trải nghiệm. Hầu hết đều khuyến khích sự tương tác giữa người chơi hoặc cộng đồng, thông qua các nền tảng trực tuyến, giải đấu, hoặc các tính năng chia sẻ thành tích và trải nghiệm. Các game này phải có khả năng thu hút người chơi trên phạm vi quốc tế, góp phần thúc đẩy cộng đồng game thủ toàn cầu và tạo ra các nền tảng cạnh tranh, hợp tác quốc tế. Chúng phải luôn được cập nhật, tích hợp các công nghệ mới như thực tế ảo, thực tế tăng cường, đa nền tảng để mở rộng khả năng trải nghiệm và thu hút người chơi mới. Các loại game có thể phục vụ các mục đích khác nhau từ giải trí, thể thao điện tử, rèn luyện sức khỏe hay trải nghiệm thực tế ảo nhưng đều hướng tới sự tiện lợi và hấp dẫn cho người dùng.

2.1.3. Vai trò của dữ liệu trong ngành game

Đối với sự phát triển của ngành game, dữ liệu giữ vai trò vô cùng quan trọng và đa dạng, góp phần tối ưu hóa trải nghiệm người chơi và nâng cao hiệu quả kinh doanh.

2.1.3.1. Vai trò của dữ liệu đối với nhà phát hành

Dữ liệu đóng vai trò quan trọng, là yếu tố không thể thiếu với các nhà phát hành game, mang lại nhiều lợi ích thiết thực và ảnh hưởng đến mọi khía cạnh của quá trình phát triển, phát hành và duy trì trò chơi.

- Thấu hiểu người chơi – Nền tảng cho mọi quyết định

Dữ liệu mở ra cánh cửa để nhà phát hành game nhìn thấu vào thế giới của người chơi. Thông qua việc theo dõi tỉ mỉ hành vi trong trò chơi, từ thời gian họ dành cho mỗi phiên, những tính năng được ưu tiên sử dụng, đến cấp độ họ đạt được và nhân vật họ yêu thích, nhà phát hành có được cái nhìn toàn diện về sở thích và mức độ tương tác của cộng đồng. Quan trọng hơn, dữ liệu cho phép phân khúc người chơi thành các nhóm nhỏ hơn dựa trên những đặc điểm chung về hành vi, sở thích hay thậm chí là nhân khẩu học. Sự phân khúc này là chìa khóa để cá nhân hóa trải nghiệm

chơi game, việc gợi ý nội dung phù hợp đến việc điều chỉnh độ khó để duy trì sự hứng thú cho từng nhóm đối tượng khác nhau. Hơn nữa, việc thu thập và phân tích phản hồi trực tiếp từ người chơi thông qua khảo sát, bình luận trên các nền tảng và đánh giá trong trò chơi cung cấp những thông tin vô giá về những gì họ thực sự mong muốn, những điểm nào cần cải thiện, từ đó giúp nhà phát hành đưa ra những quyết định phát triển sản phẩm sát với nhu cầu thực tế.

- *Tối ưu hóa trải nghiệm – Chìa khóa giữ chân người chơi*

Dữ liệu đóng vai trò then chốt trong việc tinh chỉnh và hoàn thiện trải nghiệm cốt lõi của trò chơi. Các chỉ số về độ khó ở từng màn chơi, sức mạnh tương quan giữa các nhân vật hay tỷ lệ thắng thua ở các chế độ chơi khác nhau là cơ sở để nhà phát hành thực hiện các điều chỉnh cân bằng, đảm bảo tính cạnh tranh và hấp dẫn. Bên cạnh đó, dữ liệu về các lỗi kỹ thuật, những điểm nghẽn khiến người chơi cảm thấy khó chịu hay những khu vực mà họ thường xuyên gặp khó khăn giúp đội ngũ phát triển nhanh chóng xác định và ưu tiên khắc phục các vấn đề, mang lại trải nghiệm mượt mà và liền mạch hơn. Quan trọng hơn, việc theo dõi mức độ sử dụng và yêu thích của các tính năng hiện có cung cấp thông tin quan trọng để nhà phát hành đưa ra quyết định về việc phát triển các tính năng mới, cải thiện những tính năng chưa hiệu quả hoặc thậm chí tạo ra các nội dung bổ sung phù hợp với sở thích của đa số người chơi, từ đó kéo dài tuổi thọ và duy trì sự gắn bó của cộng đồng với trò chơi.

- *Nâng cao hiệu quả tiếp thị và tiến hành thương mại hóa – Tiếp cận đúng người, tối ưu doanh thu*

Trong lĩnh vực marketing và monetization[1], dữ liệu là công cụ mạnh mẽ giúp nhà phát hành tối ưu hóa các chiến lược tiếp cận và tạo ra doanh thu bền vững. Bằng cách phân tích dữ liệu về sở thích chơi game, lịch sử tương tác và hành vi trực tuyến của người dùng, nhà phát hành có thể nhắm mục tiêu quảng cáo một cách chính xác đến những đối tượng có khả năng quan tâm đến trò chơi nhất, giảm thiểu chi phí không cần thiết và tăng tỷ lệ chuyển đổi. Hơn nữa, dữ liệu về hành vi chi tiêu của người chơi, những loại vật phẩm ảo hoặc gói dịch vụ nào được ưa chuộng, giúp nhà phát hành xây dựng các mô hình monetization hiệu quả và thiết lập mức giá phù hợp với từng phân khúc người chơi. Cuối cùng, việc theo dõi và phân tích dữ liệu về hiệu quả của các chiến dịch marketing, từ số lượt tải xuống đến doanh thu thu về, cho phép nhà phát hành đánh giá mức độ thành công và thực hiện các điều chỉnh cần thiết để tối ưu hóa lợi nhuận.

- *Dự đoán tương lai và hoạch định chiến lược phát triển – Đi trước đón đầu xu thế*

Dữ liệu không chỉ giúp nhà phát hành hiểu rõ hiện tại mà còn có vai trò quan trọng trong việc dự đoán tương lai và hoạch định các chiến lược phát triển dài hạn.

[1] các phương pháp và chiến lược mà nhà phát hành sử dụng để tạo ra doanh thu từ trò chơi, tập trung vào việc tạo ra các dòng doanh thu liên tục trong quá trình người chơi tương tác với trò chơi.

Bằng cách phân tích các xu hướng thay đổi trong sở thích của người chơi, sự trỗi dậy của các thể loại game mới, và sự phát triển của các nền tảng công nghệ, nhà phát hành có thể đưa ra những dự đoán về xu hướng thị trường trong tương lai và điều chỉnh hướng phát triển sản phẩm cho phù hợp. Dữ liệu về mức độ tương tác và mong đợi của người chơi đối với các nội dung hiện tại cũng là cơ sở để lên kế hoạch phát triển các bản mở rộng, các sự kiện trong trò chơi hoặc thậm chí là các tựa game mới, đảm bảo rằng nhà phát hành luôn đáp ứng được nhu cầu của cộng đồng và duy trì lợi thế cạnh tranh trên thị trường. Hơn nữa, việc phân tích dữ liệu về thị hiếu chơi game ở các khu vực địa lý khác nhau giúp nhà phát hành đánh giá tiềm năng mở rộng thị trường và đưa ra các quyết định đầu tư quốc tế một cách khôn ngoan.

- *Phòng ngừa gian lận và bảo mật thông tin người chơi*

Trong bối cảnh an ninh mạng ngày càng phức tạp, dữ liệu đóng vai trò then chốt trong việc phòng ngừa gian lận và bảo mật cho ngành game. Bằng cách thu thập và phân tích các mẫu hành vi bất thường, dữ liệu giúp nhà phát hành phát hiện sớm các hoạt động gian lận như sử dụng phần mềm thứ ba (cheat), tài khoản ảo (botting), hoặc các giao dịch bất hợp pháp. Các hệ thống phân tích dữ liệu tiên tiến có thể theo dõi các chỉ số như tốc độ thao tác, tần suất hoạt động, nguồn gốc truy cập và các tương tác bất thường khác để xác định các tài khoản có dấu hiệu gian lận và thực hiện các biện pháp can thiệp kịp thời, bảo vệ sự công bằng trong trò chơi và trải nghiệm của người chơi chân chính.

Bên cạnh đó, dữ liệu cũng đóng vai trò quan trọng trong việc bảo mật thông tin người dùng và tài sản ảo. Việc phân tích nhật ký hoạt động, dữ liệu giao dịch và các thông tin liên quan đến tài khoản giúp nhà phát hành phát hiện các truy cập trái phép, các nỗ lực xâm nhập hệ thống hoặc các hoạt động đáng ngờ có thể đe dọa đến an toàn của dữ liệu người dùng. Thông qua việc xây dựng các mô hình dựa trên dữ liệu về các cuộc tấn công trước đây, nhà phát hành có thể dự đoán và ngăn chặn các mối đe dọa tiềm ẩn, đồng thời triển khai các biện pháp bảo mật chủ động để bảo vệ hệ thống và thông tin cá nhân của người chơi. Tóm lại, dữ liệu không chỉ là công cụ để duy trì sự công bằng trong game mà còn là lá chắn vững chắc bảo vệ cộng đồng người chơi và tài sản của nhà phát hành khỏi các hành vi gian lận và các mối đe dọa an ninh mạng.

Dữ liệu chính là nền tảng để các nhà phát hành game đưa ra các quyết định chiến lược chính xác, nâng cao trải nghiệm người chơi, tối ưu doanh thu, và duy trì sự cạnh tranh trong thị trường ngày càng khốc liệt. Qua việc khai thác và phân tích dữ liệu một cách hiệu quả, các nhà phát hành có thể thích ứng nhanh với xu hướng thị trường, xây dựng cộng đồng trung thành, và đảm bảo sự phát triển bền vững của các sản phẩm game trong dài hạn.

2.1.3.2. Vai trò của dữ liệu đối với người chơi

Dữ liệu đóng vai trò quan trọng trong việc nâng cao trải nghiệm và mang lại nhiều lợi ích thiết thực cho người tham gia trong ngành game.

- Cá nhân hóa trải nghiệm – Thế giới game riêng biệt theo sở thích

Dữ liệu mở ra khả năng cá nhân hóa sâu sắc trải nghiệm chơi game cho từng cá nhân. Dựa trên những thông tin về lịch sử chơi, thể loại game yêu thích, thời gian hoạt động và các tương tác trong trò chơi, hệ thống có thể đưa ra những gợi ý thông minh về các tựa game mới, các bản mở rộng hấp dẫn, hoặc thậm chí là những sự kiện đặc biệt có khả năng thu hút sự quan tâm của người chơi. Không chỉ dừng lại ở việc gợi ý, dữ liệu còn cho phép điều chỉnh độ khó của trò chơi một cách linh hoạt, thích ứng với trình độ và tiến độ của từng người chơi, tạo ra một thử thách vừa phải, đủ để duy trì sự hứng thú mà không gây ra cảm giác quá tải hay nhàm chán. Thêm vào đó, các tùy chọn tùy biến giao diện, hệ thống điều khiển, hay các thiết lập hiển thị cũng có thể được cung cấp dựa trên dữ liệu về thói quen sử dụng và phản hồi của người chơi, mang đến một không gian game thoải mái và tối ưu hóa cho phong cách chơi riêng của mỗi người.

- Nâng cao kỹ năng và khẳng định bản thân – Từ người mới đến chuyên gia

Dữ liệu trở thành một công cụ đắc lực trên hành trình nâng cao kỹ năng và hiệu suất của người chơi. Các chỉ số chi tiết về màn trình diễn trong trò chơi như độ chính xác trong các pha hành động, tốc độ phản xạ trong các tình huống căng thẳng, hay số lần hạ gục đối thủ, được ghi lại và phân tích, giúp người chơi nhận diện rõ ràng những điểm mạnh cần phát huy và những khía cạnh còn hạn chế cần cải thiện. Hơn nữa, dữ liệu tổng hợp về chiến thuật và thành tích của những người chơi có kinh nghiệm và kỹ năng cao hơn có thể được chia sẻ một cách có hệ thống thông qua các hướng dẫn chuyên sâu, video phân tích trận đấu, hoặc hệ thống xếp hạng cạnh tranh. Nhờ đó, người chơi có cơ hội học hỏi những bí quyết, chiến lược hiệu quả và không ngừng hoàn thiện bản thân. Hệ thống xếp hạng và bảng thành tích dựa trên dữ liệu cũng tạo ra một môi trường cạnh tranh lành mạnh, thúc đẩy người chơi nỗ lực hơn để đạt được những vị trí cao hơn và khẳng định khả năng của mình trong cộng đồng game thủ. - *Kết nối cộng đồng và mở rộng tương tác – Sân chơi không còn cô đơn*

Dữ liệu đóng vai trò như một cầu nối, tạo ra những trải nghiệm kết nối và tương tác phong phú giữa những người chơi. Dựa trên thông tin về sở thích, trình độ kỹ năng và lịch sử chơi, hệ thống có thể gợi ý những người chơi khác có cùng đam mê và mục tiêu để kết bạn và cùng nhau chinh phục những thử thách trong game. Dữ liệu về hoạt động của người chơi trong các diễn đàn trực tuyến, các nhóm chat cộng đồng, hoặc trên các nền tảng mạng xã hội liên quan đến game giúp họ dễ dàng tìm thấy và tham gia vào các cộng đồng có chung mối quan tâm, chia sẻ kinh nghiệm, thảo luận về chiến thuật và xây dựng những mối quan hệ bạn bè mới. Thêm vào đó, dựa trên dữ liệu về mức độ tương tác và thành tích cá nhân, nhà phát hành có thể tạo ra các sự kiện

đặc biệt và trao tặng những phần thưởng được thiết kế riêng cho từng người chơi, tạo cảm giác được công nhận và khuyến khích sự tham gia tích cực vào cộng đồng game.

- Bảo vệ quyền lợi và xây dựng môi trường lành mạnh – An tâm tận hưởng đam mê

Dữ liệu đóng vai trò quan trọng trong việc đảm bảo một môi trường chơi game công bằng và an toàn cho tất cả người chơi. Bằng cách theo dõi và phân tích các hành vi bất thường trong trò chơi, dữ liệu giúp nhà phát hành phát hiện và xử lý những trường hợp gian lận như sử dụng phần mềm thứ ba để can thiệp vào gameplay, tạo lợi thế không công bằng cho bản thân. Việc này giúp bảo vệ quyền lợi của những người chơi chân chính và duy trì sự cân bằng trong trò chơi. Đồng thời, dữ liệu về hoạt động đăng nhập, lịch sử giao dịch và các dấu hiệu đáng ngờ khác cũng giúp nhà phát hành phát hiện và ngăn chặn các hành vi xâm nhập tài khoản hoặc lừa đảo, bảo vệ thông tin cá nhân và tài sản ảo của người chơi. Nhờ đó, người chơi có thể yên tâm tận hưởng niềm đam mê của mình trong một môi trường game lành mạnh và được bảo vệ.

Dữ liệu giúp người chơi có trải nghiệm cá nhân hóa, hấp dẫn và an toàn hơn trong quá trình chơi game. Nó tạo điều kiện để các nhà phát hành hiểu rõ hơn về cộng đồng người chơi, từ đó điều chỉnh, cập nhật và phát triển nội dung phù hợp, giúp người chơi cảm thấy hài lòng, gắn bó lâu dài với trò chơi của mình.

2.1.4. Các yếu tố ảnh hưởng đến doanh thu của trò chơi

Các yếu tố ảnh hưởng đến doanh thu của trò chơi rất đa dạng và phức tạp, là những yếu tố quyết định khả năng thu hút người chơi, giữ chân họ và thúc đẩy các hoạt động mua hàng trong game.

- Chất lượng và thiết kế – Nền tảng vững chắc cho thành công thương mại

Nền tảng cơ bản nhất quyết định sự thành công về mặt doanh thu của một trò chơi điện tử chính là chất lượng và thiết kế cốt lõi của nó. Một gameplay hấp dẫn, có chiều sâu và mang lại những trải nghiệm thú vị, thử thách nhưng cũng đầy phần thưởng sẽ là yếu tố then chốt để thu hút và giữ chân người chơi. Bên cạnh đó, yếu tố thị giác và thính giác cũng đóng vai trò quan trọng; đồ họa sắc nét, phong cách nghệ thuật độc đáo cùng âm thanh sống động và phù hợp sẽ tạo ra một thế giới trò chơi chân thực và cuốn hút. Cốt truyện lôi cuốn với những nhân vật được xây dựng tỉ mỉ, có chiều sâu sẽ tạo ra sự kết nối cảm xúc với người chơi, khuyến khích họ khám phá và gắn bó lâu dài. Sự sáng tạo và đổi mới trong ý tưởng và cơ chế gameplay giúp trò chơi nổi bật giữa một thị trường cạnh tranh khốc liệt. Cuối cùng, khả năng chơi lại cao và tính ổn định, hiệu suất mượt mà của trò chơi là những yếu tố không thể thiếu để đảm bảo trải nghiệm tích cực và khuyến khích người chơi tiếp tục đầu tư thời gian và tiền bạc.

- Mô hình monetization - Cân bằng giữa doanh thu và trải nghiệm người chơi

Mô hình monetization được lựa chọn có ảnh hưởng trực tiếp đến dòng doanh thu của trò chơi. Đối với các trò chơi trả phí, mức giá cần được thiết lập một cách hợp lý, cân nhắc giữa giá trị mà trò chơi mang lại, mức giá của các đối thủ cạnh tranh và khả năng chi trả của đối tượng mục tiêu. Với các trò chơi miễn phí (free-to-play), doanh thu chủ yếu đến từ mua trong ứng dụng (IAPs) và quảng cáo. Việc cung cấp các vật phẩm ảo, tiền tệ trong game hay mở khóa nội dung cần được thiết kế một cách khéo léo để không tạo ra sự mất cân bằng trong gameplay, gây ra cảm giác "pay-to-win" và làm mất đi sự hứng thú của những người chơi không chi tiền. Tương tự, việc tích hợp quảng cáo cần được thực hiện một cách tinh tế, tránh làm gián đoạn quá nhiều đến trải nghiệm chơi. Mô hình gói thuê bao có thể hiệu quả đối với những trò chơi có nội dung được cập nhật liên tục và mang lại những đặc quyền hấp dẫn cho người đăng ký. Việc lựa chọn và cân bằng các yếu tố của mô hình monetization là một bài toán khó, đòi hỏi nhà phát hành phải hiểu rõ đối tượng người chơi và có chiến lược phù hợp để tối đa hóa doanh thu mà vẫn duy trì được một cộng đồng người chơi lành mạnh và gắn bó.

- Marketing và quảng bá – Tiếp cận đúng đối tượng, tạo dựng tiếng vang

Chiến lược marketing và quảng bá đóng vai trò then chốt trong việc thu hút sự chú ý của người chơi và tạo ra doanh thu ban đầu cho trò chơi. Việc xác định đúng đối tượng mục tiêu và lựa chọn các kênh quảng bá phù hợp, từ mạng xã hội, YouTube, Twitch đến các trang tin tức và diễn đàn về game, là vô cùng quan trọng. Các chiến dịch marketing sáng tạo và hấp dẫn sẽ tạo ra sự quan tâm và tò mò cho người chơi tiềm năng. Quan hệ công chúng (PR) hiệu quả, thông qua việc hợp tác với giới truyền thông và các influencer có ảnh hưởng, có thể mang lại những đánh giá tích cực và tăng độ tin cậy cho trò chơi. Bên cạnh đó, hiệu ứng lan truyền tự nhiên, khi người chơi cảm thấy thích thú và chia sẻ trò chơi với bạn bè, cũng đóng góp một phần không nhỏ vào việc tăng trưởng người dùng. Một cộng đồng người chơi lớn mạnh, tích cực và gắn bó không chỉ giúp duy trì sự quan tâm đến trò chơi mà còn thu hút thêm những người chơi mới thông qua các hoạt động và tương tác.

- Yếu tố thị trường và cạnh tranh – Nắm bắt xu hướng, tạo sự khác biệt

Bối cảnh thị trường và mức độ cạnh tranh có ảnh hưởng không nhỏ đến doanh thu của trò chơi. Việc nhận diện và nắm bắt các xu hướng game đang thịnh hành có thể giúp trò chơi tiếp cận được một lượng lớn người chơi có sẵn. Tuy nhiên, thị trường game luôn đầy rẫy những sản phẩm cạnh tranh, vì vậy việc tạo ra một trò chơi có những điểm khác biệt và độc đáo, mang đến những trải nghiệm mới lạ mà các đối thủ không có, là yếu tố then chốt để thu hút sự chú ý và tạo dựng lợi thế. Thời điểm phát hành cũng cần được cân nhắc kỹ lưỡng, tránh việc ra mắt cùng thời điểm với các "bom tấn" khác có thể làm lu mờ sự xuất hiện của trò chơi. Cuối cùng, việc lựa chọn nền

tảng phát hành phù hợp với đối tượng mục tiêu cũng rất quan trọng; một trò chơi nhắm đến game thủ PC sẽ cần một chiến lược phát hành khác với một trò chơi dành cho thị trường di động.

- Yếu tố khu vực và văn hóa – Chinh phục thị trường toàn cầu

Sự khác biệt về sở thích chơi game và văn hóa giữa các khu vực địa lý khác nhau là một yếu tố quan trọng cần được nhà phát hành lưu ý. Một thể loại game có thể rất thành công ở một khu vực nhưng lại không được đón nhận ở khu vực khác. Do đó, việc bản địa hóa trò chơi, bao gồm dịch thuật ngôn ngữ và điều chỉnh nội dung cho phù hợp với văn hóa địa phương, là rất quan trọng để tiếp cận thị trường quốc tế một cách hiệu quả. Ngoài ra, các quy định pháp lý liên quan đến nội dung, monetization và quảng cáo game cũng có thể khác nhau ở mỗi quốc gia và vùng lãnh thổ, đòi hỏi nhà phát hành phải nghiên cứu kỹ lưỡng để đảm bảo tuân thủ và tối ưu hóa doanh thu.

- Hỗ trợ sau phát hành – Duy trì sự quan tâm, kéo dài tuổi thọ

Việc hỗ trợ sau phát hành đóng vai trò quan trọng trong việc duy trì sự quan tâm của người chơi và kéo dài tuổi thọ của trò chơi, đặc biệt đối với các trò chơi miễn phí. Việc cập nhật nội dung thường xuyên, bổ sung các màn chơi mới, nhân vật, chế độ chơi hoặc các tính năng mới sẽ giữ cho trò chơi luôn tươi mới và hấp dẫn. Một hệ thống hỗ trợ khách hàng tốt, giải quyết nhanh chóng các vấn đề và lắng nghe phản hồi từ cộng đồng người chơi sẽ tạo dựng lòng tin và sự hài lòng. Việc tổ chức các sự kiện đặc biệt trong trò chơi, các giải đấu cạnh tranh hoặc các hoạt động cộng đồng không chỉ thu hút người chơi quay trở lại mà còn tạo ra cơ hội để thúc đẩy chi tiêu trong game.

2.2. KHÁI NIỆM VÀ PHÂN LOẠI DỮ LIỆU

2.2.1. Khái niệm

Dữ liệu là tập hợp các giá trị hoặc thông tin thu thập được từ thực tế, có thể ở dạng số liệu, văn bản hoặc hình ảnh, đóng vai trò trung tâm trong các quá trình phân tích và đưa ra quyết định. Bản thân dữ liệu thường chưa mang nhiều ý nghĩa cho đến khi nó được xử lý, tổ chức, phân tích và diễn giải để trở thành thông tin hữu ích. Nói một cách đơn giản, dữ liệu là nguyên liệu đầu vào cho quá trình phân tích và ra quyết định. Nó có thể đến từ nhiều nguồn khác nhau, bao gồm các hệ thống máy tính, cảm biến, khảo sát, tương tác của người dùng và các quá trình tự nhiên.

2.2.2. Phân loại dữ liệu

2.2.2.1. Dựa trên tính chất

Dữ liệu thường được phân thành hai loại chính dựa trên tính chất

- *Dữ liệu định tính (Qualitative / Categorical Data):* Là loại dữ liệu biểu diễn bằng nhãn, tên hoặc nhóm phân loại không có thứ tự định lượng rõ ràng. Nó thường được thu thập thông qua quan sát, phỏng vấn hoặc phân tích nội dung. Trong bộ dữ liệu "vgsales.csv" gồm các cột như Genre (thể loại trò chơi), Platform (nền tảng phát hành) hay Publisher (nhà phát hành) là dữ liệu định tính.

- *Dữ liệu định lượng (Quantitative / Numerical Data):* Là loại dữ liệu biểu diễn bằng con số và có thể thực hiện các phép tính toán học. Nó thường được sử dụng cho các phân tích thống kê và toán học. Dữ liệu định lượng trong bộ dữ liệu bao gồm các cột như Global_Sales, NA_Sales, EU_Sales, v.v. phản ánh doanh thu cụ thể theo khu vực.

2.2.2.2. Phân loại dữ liệu theo định lượng

Theo định lượng, dữ liệu có thể chia thành hai nhóm

- *Dữ liệu rời rạc (Discrete Data):* Là dữ liệu có giá trị đếm được và không liên tục. Ví dụ như Year (năm phát hành) có thể được coi là dữ liệu rời rạc nếu xem xét các năm riêng lẻ.

- *Dữ liệu liên tục (Continuous Data):* Là dữ liệu có thể nhận giá trị trong một khoảng liên tục, ví dụ như doanh thu bán hàng (Global_Sales) có thể là bất kỳ số thực nào trong một phạm vi.

2.2.2.3. Phân loại dữ liệu theo cấu trúc

- *Dữ liệu có cấu trúc (Structured Data):* Là loại dữ liệu có tổ chức rõ ràng, tuân theo một mô hình hoặc lược đồ (schema) xác định. Nó thường được lưu trữ trong các cơ sở dữ liệu quan hệ (Relational Databases) dưới dạng bảng với các hàng (records) và cột (attributes/fields). Mỗi cột có một kiểu dữ liệu cụ thể (ví dụ: số nguyên, chuỗi văn bản, ngày tháng). Dữ liệu có cấu trúc rất dễ dàng để tìm kiếm, sắp xếp và phân tích bằng các công cụ truyền thống.

- *Dữ liệu bán cấu trúc (Semi-structured Data):* Là loại dữ liệu không có cấu trúc cố định như dữ liệu có cấu trúc, nhưng nó chứa các thẻ hoặc dấu hiệu (tags, markers) để phân tách và tổ chức các phần tử dữ liệu, giúp việc phân tích trở nên dễ dàng hơn so với dữ liệu phi cấu trúc. Nó thường không tuân theo một lược đồ nghiêm ngặt.

- *Dữ liệu phi cấu trúc (Unstructured Data):* Là loại dữ liệu không có cấu trúc hoặc không tuân theo một mô hình tổ chức cụ thể. Nó khó khăn hơn trong việc phân tích và xử lý bằng các phương pháp truyền thống.

2.2.3. Dữ liệu trong bộ "vgsales.csv"

Bộ dữ liệu "vgsales.csv" là bộ dữ liệu có cấu trúc, tập hợp các bản ghi thống kê doanh thu của hơn 16.000 trò chơi điện tử, bao gồm thông tin về tên game, thể loại, nền tảng, năm phát hành, nhà phát hành và doanh thu theo từng khu vực (Bắc Mỹ,

châu Âu, Nhật Bản và toàn cầu). Việc phân loại rõ các trường dữ liệu theo định tính và định lượng giúp xác định hướng tiếp cận phân tích phù hợp, đồng thời hỗ trợ trực quan hóa dữ liệu một cách chính xác và có ý nghĩa.

2.3. KỸ THUẬT PHÂN TÍCH DỮ LIỆU

Phân tích dữ liệu là quá trình thu thập, xử lý và diễn giải thông tin từ dữ liệu thô nhằm rút ra những hiểu biết có giá trị phục vụ cho việc ra quyết định. Trong bối cảnh đề tài này, các kỹ thuật phân tích dữ liệu giúp khám phá xu hướng doanh thu, đánh giá hiệu suất kinh doanh của từng thể loại game, nền tảng, nhà phát hành và khu vực tiêu thụ.

2.3.1. Quy trình phân tích dữ liệu

Quá trình phân tích dữ liệu thường bao gồm các bước cơ bản sau:

- *Thu thập dữ liệu (data collection)*

Là giai đoạn đầu tiên và vô cùng quan trọng, liên quan đến việc thu thập, đo lường và ghi lại thông tin một cách có hệ thống từ các nguồn khác nhau để phục vụ cho mục tiêu phân tích cụ thể. Ở đây đề tài sử dụng bộ dữ liệu "vgsales.csv" có sẵn từ nguồn đáng tin cậy (trang web Kaggle).

Dữ liệu thu thập được chính là nguyên liệu đầu vào cho toàn bộ quá trình phân tích. Nếu dữ liệu không đầy đủ, không chính xác hoặc không liên quan, thì mọi phân tích sau đó sẽ không có giá trị và dẫn đến những kết luận sai lệch. Dữ liệu cần phải đảm bảo tính chính xác và độ tin cậy tối đa vì đây là yếu tố then chốt để có được những kết quả phân tích đáng tin cậy. Đồng thời cần thu thập dữ liệu từ nhiều nguồn khác nhau và sử dụng các phương pháp thu thập đa dạng có thể cung cấp một cái nhìn toàn diện và sâu sắc hơn về vấn đề đang nghiên cứu, tránh những kết luận phiến diện dựa trên một tập dữ liệu hạn chế.

- *Tiền xử lý dữ liệu (Data Preprocessing)*

Là giai đoạn quan trọng trong quá trình phân tích dữ liệu, diễn ra sau khi dữ liệu đã được thu thập. Đây là quá trình chuyển đổi dữ liệu thô, nhiều vấn đề như thiếu giá trị, dữ liệu nhiễu, định dạng không nhất quán, v.v. sang một định dạng phù hợp hơn để phân tích, mô hình hóa hoặc sử dụng trong các ứng dụng khác**.** Mục tiêu là để cải thiện chất lượng dữ liệu, chuẩn hóa dữ liệu, tăng cường khả năng tích hợp dữ liệu…

Tiền xử lý dữ liệu hỗ trợ nâng cao độ chính xác của phân tích**,** dữ liệu sạch và được chuẩn hóa sẽ dẫn đến kết quả phân tích chính xác và đáng tin cậy hơn. Cải thiện được hiệu suất hoạt động của mô hình. Đồng thời tránh được những sai sót có thể xảy ra trong quá trình phân tích và đưa ra quyết định, giảm thời gian và tài nguyên cần

thiết cho các bước phân tích tiếp theo cũng như giúp nhà phân tích hiểu rõ hơn về các đặc điểm và vấn đề tiềm ẩn trong dữ liệu.

- Khám phá dữ liệu (Exploratory Data Analysis - EDA)

Là quá trình phân tích sơ bộ dữ liệu để khám phá và tóm tắt các đặc điểm chính của nó, thường sử dụng các phương pháp trực quan hóa và thống kê mô tả. Đây là một bước không thể thiếu trong quy trình phân tích dữ liệu để có thể "nhìn tận mắt, sờ tận tay" bộ dữ liệu, đặt nền móng vững chắc cho các phân tích sâu hơn và đưa ra những kết luận có giá trị.

Mục đích của quá trình này để nhà phân tích hiểu rõ hơn về dữ liệu, làm quen với cấu trúc, loại biến, kích thước và nội dung của bộ dữ liệu, nhận biết các mẫu (patterns), xu hướng (trends) và mối quan hệ (relationships) tiềm ẩn giữa các biến. Bên cạnh đó, hỗ trợ tìm kiếm các giá trị thiếu, dữ liệu nhiễu (outliers, lỗi), sự không nhất quán và các vấn đề khác cần được xử lý ở giai đoạn tiền xử lý. Đồng thời có thể đưa ra các giả thuyết ban đầu về dữ liệu và các câu hỏi cần được điều tra sâu hơn và lựa chọn các phương pháp phân tích, các mô hình phù hợp cho các giai đoạn tiếp theo.

- Phân tích dữ liệu (Data Analysis)

Là quá trình kiểm tra, làm sạch, chuyển đổi và mô hình hóa dữ liệu với mục tiêu khám phá thông tin hữu ích, đưa ra kết luận và hỗ trợ việc ra quyết định. Đây là một quá trình lặp đi lặp lại, bao gồm việc áp dụng các kỹ thuật thống kê, toán học, trực quan hóa và các công cụ phần mềm để hiểu rõ hơn về dữ liệu, xác định các mẫu, xu hướng và mối quan hệ tiềm ẩn. Hiểu đơn giản hơn, phân tích dữ liệu là quá trình biến dữ liệu thô thành thông tin có ý nghĩa và giá trị, giúp trả lời các câu hỏi, giải quyết vấn đề và đưa ra các dự đoán.

Phân tích dữ liệu đóng vai trò trung tâm và then chốt trong toàn bộ quy trình phân tích dữ liệu. Nó là trái tim của quá trình, kết nối các giai đoạn khác lại với nhau và mang lại giá trị thực sự từ dữ liệu. Đây cũng là giai đoạn biến thông tin thành tri thức, đưa ra câu trả lời cho các câu hỏi cụ thể được đặt ra ở giai đoạn xác định vấn đề. Dựa trên các mẫu và xu hướng được xác định để xây dựng các mô hình dự đoán và cung cấp thông tin đầu vào quan trọng cho việc ra quyết định trong nhiều lĩnh vực.

- Diễn giải và báo cáo

Là giai đoạn cuối cùng và vô cùng quan trọng trong quy trình phân tích dữ liệu. Nó bao gồm việc phân tích, giải thích và truyền đạt ý nghĩa của các kết quả, thông tin và hiểu biết (insights) thu được từ quá trình phân tích dữ liệu một cách rõ ràng, dễ hiểu và có ngữ cảnh cho đối tượng mục tiêu. Quá trình này đòi hỏi người phân tích phải có khả năng suy luận, kết nối các phát hiện với bối cảnh thực tế và truyền đạt thông tin một cách hiệu quả đến những người có thể không có kiến thức chuyên sâu về phân tích dữ liệu.

Diễn giải báo cáo đóng vai trò then chốt và mang lại nhiều lợi ích trong quy trình phân tích dữ liệu. Giai đoạn phân tích tạo ra các kết quả, nhưng chính diễn giải mới mang lại ý nghĩa thực sự, cung cấp những hiểu biết sâu sắc và các khuyến nghị cho những kết quả đó. Nó đảm bảo rằng thông tin quan trọng được truyền tải hiệu quả đến tất cả các bên liên quan, giúp người dùng hiểu được "chuyện gì đã xảy ra" và "tại sao nó lại xảy ra", từ đó các nhà quản lý và các bên liên quan có thể đưa ra các quyết định sáng suốt và hiệu quả. Nếu không có diễn giải hiệu quả, một báo cáo phân tích dữ liệu dù có phức tạp và chi tiết đến đâu cũng có thể trở nên vô nghĩa.

2.3.2. Thống kê mô tả (Descriptive Statistics)

Đây là kỹ thuật phân tích cơ bản giúp hiểu tổng quan về dữ liệu. Một số chỉ số thường dùng gồm:

- *Số lượng (Count)*

Trong phân tích dữ liệu, tham số "Số lượng" (count) đóng vai trò cơ bản trong việc cung cấp cái nhìn đầu tiên về quy mô của tập dữ liệu đang được xem xét. Nó đơn giản chỉ ra tổng số bản ghi hoặc quan sát hợp lệ có trong một cột hoặc toàn bộ DataFrame. Thông tin này đặc biệt hữu ích để đánh giá mức độ đầy đủ của dữ liệu, xác định xem có bao nhiêu giá trị thực tế được thu thập và ghi nhận cho mỗi biến. Sự khác biệt về số lượng giữa các cột có thể ngầm chỉ ra sự tồn tại của các giá trị thiếu (missing values), một vấn đề quan trọng cần được xử lý trong giai đoạn tiền xử lý dữ liệu để đảm bảo tính chính xác của các phân tích tiếp theo.

- *Trung bình (Mean)*

Đây là một thước đo quan trọng để xác định xu hướng trung tâm của một tập hợp dữ liệu số. Nó được tính bằng cách lấy tổng của tất cả các giá trị trong tập dữ liệu và chia cho tổng số các giá trị đó. Trung bình cung cấp một điểm đại diện cho toàn bộ phân phối dữ liệu, cho thấy giá trị mà các quan sát có xu hướng tập trung xung quanh. Tuy nhiên, cần lưu ý rằng trung bình có thể bị ảnh hưởng đáng kể bởi các giá trị ngoại lệ (outliers), những giá trị quá lớn hoặc quá nhỏ so với phần còn lại của dữ liệu, do đó, việc xem xét các thước đo xu hướng trung tâm khác như trung vị là cần thiết để có một cái nhìn toàn diện hơn.

- *Trung vị (Median)*

"Trung vị" (median) là một thước đo khác để xác định xu hướng trung tâm của một tập hợp dữ liệu số, nhưng nó mạnh mẽ hơn so với trung bình trong việc chống lại ảnh hưởng của các giá trị ngoại lệ. Trung vị là giá trị nằm ở chính giữa của tập dữ liệu sau khi đã được sắp xếp theo thứ tự tăng dần (hoặc giảm dần). Nếu số lượng quan sát là lẻ, trung vị là giá trị ở vị trí chính giữa. Nếu số lượng quan sát là chẵn, trung vị là trung bình của hai giá trị ở vị trí giữa. Do không bị tác động bởi các giá trị cực

đoan, trung vị thường là một thước đo xu hướng trung tâm đáng tin cậy hơn khi dữ liệu có sự xuất hiện của các giá trị ngoại lệ.

- Giá trị lớn nhất (Max) và nhỏ nhất (Min)

Giá trị max và min là hai thống kê mô tả đơn giản nhưng vô cùng hữu ích để xác định phạm vi hoặc khoảng biến thiên của một biến số. Giá trị max cho biết giá trị cao nhất mà một biến có thể nhận trong tập dữ liệu, trong khi giá trị min cho biết giá trị thấp nhất. Sự khác biệt giữa giá trị lớn nhất và giá trị nhỏ nhất (hay còn gọi là khoảng) cung cấp một cái nhìn trực quan về độ rộng của phân phối dữ liệu và có thể giúp phát hiện các giá trị ngoại lệ tiềm năng nằm ngoài phạm vi thông thường.

- Độ lệch chuẩn (Standard Deviation)

Đây là một thước đo quan trọng để đánh giá mức độ phân tán hoặc sự biến động của các giá trị trong một tập dữ liệu so với giá trị trung bình của nó. Độ lệch chuẩn nhỏ cho thấy rằng các giá trị có xu hướng tập trung gần với trung bình, trong khi độ lệch chuẩn lớn chỉ ra rằng các giá trị phân tán rộng hơn xung quanh trung bình. Độ lệch chuẩn cung cấp một cách định lượng để hiểu mức độ "rủi ro" hoặc "biến động" của một biến số và thường được sử dụng cùng với giá trị trung bình để mô tả đầy đủ hơn về phân phối của dữ liệu.

2.3.3. Lọc, phân nhóm và tổng hợp dữ liệu

- Lọc dữ liệu – Tinh chỉnh tập dữ liệu để tập trung vào thông tin quan trọng

Trong quá trình phân tích dữ liệu "vgsales.csv", việc lọc dữ liệu (filtering) đóng vai trò như một công cụ tinh chỉnh mạnh mẽ, cho phép chọn ra những bản ghi cụ thể đáp ứng các tiêu chí nhất định mà chúng ta quan tâm. Quá trình lọc này giúp loại bỏ những thông tin không liên quan đến mục tiêu phân tích hiện tại, thu hẹp phạm vi nghiên cứu và làm nổi bật những dữ liệu quan trọng nhất, từ đó giúp việc khám phá và phân tích trở nên hiệu quả và tập trung hơn.

- Phân nhóm dữ liệu – Sắp xếp và cấu trúc thông tin theo các thuộc tính chung

Sau khi đã tinh chỉnh tập dữ liệu, bước tiếp theo là phân nhóm (grouping) dữ liệu theo một hoặc nhiều tiêu chí cụ thể. Đối với bộ dữ liệu "vgsales.csv", việc phân nhóm có thể được thực hiện dựa trên nhiều thuộc tính khác nhau, tùy thuộc vào câu hỏi nghiên cứu. Ví dụ, để hiểu rõ hơn về sự đóng góp của từng thể loại game vào tổng doanh thu, chúng ta có thể nhóm dữ liệu theo cột 'Genre'. Tương tự, để so sánh hiệu suất bán hàng trên các hệ máy khác nhau, việc phân nhóm theo cột 'Platform' sẽ rất hữu ích. Hoặc, để đánh giá sự thành công của các nhà phát hành khác nhau, chúng ta có thể nhóm dữ liệu theo cột 'Publisher'. Quá trình phân nhóm này tạo ra các tập dữ liệu con, mỗi tập chứa các bản ghi có cùng giá trị ở thuộc tính được chọn để nhóm,

giúp chúng ta có thể so sánh và phân tích dữ liệu một cách có cấu trúc và hệ thống hơn.

- Tổng hợp dữ liệu – Tính toán các chỉ số thống kê để so sánh và đánh giá

Sau cùng, giai đoạn tổng hợp (aggregation) cho phép chúng ta tính toán các chỉ số thống kê quan trọng cho mỗi nhóm. Với bộ dữ liệu "vgsales.csv", các hàm tổng hợp thường được sử dụng bao gồm tính tổng doanh thu (sum) để biết tổng doanh số của từng thể loại, nền tảng hoặc nhà phát hành. Có thể tính giá trị trung bình (mean) để so sánh hiệu suất bán hàng trung bình giữa các nhóm. Ngoài ra, việc đếm số lượng bản ghi (count) trong mỗi nhóm giúp xác định số lượng trò chơi thuộc mỗi thể loại, được phát hành trên mỗi nền tảng hoặc được phát hành bởi mỗi nhà phát hành. Các kết quả tổng hợp này cung cấp những thông tin định lượng quan trọng, cho phép chúng ta thực hiện các so sánh trực tiếp giữa các nhóm, đánh giá mức độ đóng góp hoặc hiệu suất của từng nhóm và rút ra những kết luận có ý nghĩa về thị trường trò chơi điện tử.

2.3.4. Tìm mối quan hệ giữa các biến

Một phần quan trọng trong phân tích là phát hiện các mối tương quan hoặc xu hướng giữa các trường dữ liệu.

- Khám phá sự phổ biến của thể loại game theo từng khu vực

Một trong những khía cạnh quan trọng trong việc hiểu rõ thị trường trò chơi điện tử toàn cầu là xác định sự khác biệt về sở thích thể loại giữa các khu vực địa lý. Phân tích dữ liệu doanh số theo từng khu vực (Bắc Mỹ, Châu Âu, Nhật Bản và các khu vực khác) và đối chiếu với thể loại của từng trò chơi sẽ giúp chúng ta khám phá ra những xu hướng thú vị. Liệu thể loại hành động có chiếm ưu thế ở Bắc Mỹ, trong khi thể loại nhập vai lại được ưa chuộng hơn ở Nhật Bản? Việc tìm ra những mối liên hệ này không chỉ cung cấp cái nhìn sâu sắc về thị hiếu văn hóa mà còn có giá trị chiến lược cho các nhà phát hành khi muốn nhắm mục tiêu sản phẩm và chiến dịch marketing của mình đến các thị trường cụ thể.

- Phân tích hiệu suất nền tảng theo thời gian

Để đánh giá sự thành công và vòng đời của các nền tảng chơi game khác nhau, việc phân tích doanh thu theo từng giai đoạn là vô cùng cần thiết. Chúng ta có thể nhóm dữ liệu theo năm phát hành và nền tảng để xem xét nền tảng nào đạt doanh thu cao nhất trong những năm hoàng kim của nó. Liệu PlayStation 2 có thống trị doanh thu vào đầu những năm 2000, trong khi Xbox 360 lại vươn lên dẫn đầu vào cuối thập kỷ đó? Việc xác định những xu hướng này giúp chúng ta hiểu rõ sự thay đổi trong thị trường phần cứng và tầm ảnh hưởng của từng nền tảng đến doanh số trò chơi theo thời gian.

- Tìm hiểu mối tương quan giữa thời gian phát hành và hiệu suất doanh thu

Một câu hỏi thú vị khác là liệu có mối quan hệ nào giữa năm phát hành và doanh thu trung bình của các trò chơi hay không. Có thể tính toán doanh thu trung bình cho các trò chơi được phát hành vào mỗi năm và xem xét liệu có xu hướng tăng hoặc giảm theo thời gian hay không. Liệu những trò chơi được phát hành gần đây có xu hướng đạt doanh thu cao hơn do sự phát triển của thị trường và công nghệ, hay những tựa game kinh điển từ những năm trước vẫn giữ vững vị thế? Việc khám phá mối quan hệ này có thể cung cấp cái nhìn sâu sắc về sự thay đổi trong quy mô thị trường và giá trị trung bình của một trò chơi thành công qua các năm.

- Tổng hợp dữ liệu để so sánh và đánh giá

Để phục vụ cho việc so sánh và đánh giá hiệu quả hơn, việc tổng hợp dữ liệu theo các nhóm là một bước quan trọng. Có thể tiến hành tính tổng doanh số, doanh thu trung bình hoặc số lượng trò chơi đã phát hành theo thể loại, nền tảng, nhà phát hành hoặc khu vực. Việc này cho phép chúng ta dễ dàng so sánh hiệu suất giữa các nhóm khác nhau. Ví dụ, chúng ta có thể so sánh tổng doanh thu của thể loại hành động với thể loại thể thao, hoặc so sánh số lượng trò chơi được phát hành trên PlayStation so với Xbox. Những phép tổng hợp này cung cấp những con số và thống kê trực quan, giúp chúng ta đưa ra những đánh giá và kết luận dựa trên dữ liệu một cách rõ ràng và thuyết phục.

2.4. TRỰC QUAN HÓA DỮ LIỆU

Trực quan hóa dữ liệu (Data Visualization) là quá trình chuyển đổi dữ liệu từ dạng bảng số hoặc văn bản sang các biểu đồ, hình ảnh dễ hiểu nhằm hỗ trợ phân tích, phát hiện xu hướng và truyền đạt thông tin một cách trực quan. Trong bối cảnh đề tài phân tích doanh thu trò chơi điện tử, trực quan hóa dữ liệu giúp thể hiện sự phân bố doanh thu theo thể loại, nền tảng, khu vực hoặc thời gian một cách rõ ràng và hiệu quả.

2.4.1. Vai trò của trực quan hóa dữ liệu

Thay vì sử dụng dữ liệu thô một cách khô khan và khó hiểu, các nhà phân tích dữ liệu thường chuẩn bị và trình bày dữ liệu theo ngữ cảnh phù hợp. Họ định hình dữ liệu ở dạng trực quan để những người phụ trách đưa ra quyết định có thể xác định mối quan hệ giữa dữ liệu và phát hiện ra các mẫu hoặc xu hướng ẩn. Trực quan hóa dữ liệu tạo ra các thông điệp giúp nâng cao nghiệp vụ thông minh và hỗ trợ đưa ra quyết định cũng như lập kế hoạch chiến lược dựa trên dữ liệu.

- Hỗ trợ quá trình hiểu dữ liệu dễ dàng hơn

Con người có xu hướng xử lý thông tin trực quan nhanh hơn và hiệu quả hơn so với các bảng số liệu thuần túy. Biểu đồ, đồ thị và các hình thức trực quan hóa khác

giúp tóm tắt lượng lớn dữ liệu phức tạp thành những hình ảnh dễ hiểu, làm nổi bật các đặc điểm, xu hướng và mối quan hệ quan trọng.

- Phát hiện các mẫu, xu hướng và ngoại lệ

Trực quan hóa dữ liệu cho phép người dùng nhìn thấy các mẫu, xu hướng, sự tương quan và các điểm bất thường (outliers) trong dữ liệu mà có thể khó nhận ra khi chỉ xem xét các con số. Điều này giúp khám phá những thông tin giá trị và đưa ra những giả thuyết ban đầu.

- Truyền đạt thông tin hiệu quả

Biểu đồ và đồ thị là những công cụ mạnh mẽ để truyền đạt kết quả phân tích và những hiểu biết sâu sắc đến các đối tượng khác nhau, bao gồm cả những người không có kiến thức chuyên sâu về phân tích dữ liệu. Một hình ảnh trực quan tốt có thể truyền tải một lượng lớn thông tin một cách nhanh chóng và dễ nhớ.

- Hỗ trợ quá trình khám phá dữ liệu (EDA)

Trong giai đoạn khám phá dữ liệu, trực quan hóa là một kỹ thuật không thể thiếu để làm quen với dữ liệu, kiểm tra chất lượng, xác định các vấn đề tiềm ẩn (giá trị thiếu, dữ liệu nhiễu) và khám phá các mối quan hệ ban đầu giữa các biến.

- Hỗ trợ ra quyết định chiến lược

Các hình ảnh trực quan rõ ràng và súc tích có thể cung cấp những bằng chứng trực quan mạnh mẽ để hỗ trợ việc ra quyết định trong kinh doanh, khoa học và nhiều lĩnh vực khác. Chúng giúp các nhà quản lý và các bên liên quan hiểu rõ hơn về tình hình hiện tại, dự đoán xu hướng tương lai và đánh giá hiệu quả của các hành động.

- Tăng tính tương tác và khám phá sâu hơn

Nhiều công cụ trực quan hóa hiện đại cho phép người dùng tương tác với biểu đồ, lọc dữ liệu, phóng to, thu nhỏ và khám phá dữ liệu ở nhiều góc độ khác nhau. Điều này khuyến khích sự tò mò và giúp người dùng tự khám phá ra những thông tin giá trị.

- Cải thiện dịch vụ khách hàng

Trực quan hóa dữ liệu làm nổi bật nhu cầu và mong muốn của khách hàng thông qua biểu diễn đồ họa. Bạn có thể xác định những lỗ hổng trong dịch vụ khách hàng của mình, cải thiện sản phẩm hoặc dịch vụ theo chiến lược và giảm hoạt động kém hiệu quả.

- Tăng mức độ tương tác của nhân viên

Các kỹ thuật trực quan hóa dữ liệu rất hữu ích đối với quá trình truyền đạt kết quả phân tích dữ liệu cho một nhóm nhiều nhân viên. Toàn bộ nhóm có thể cùng trực

quan hóa dữ liệu để phát triển các mục tiêu và kế hoạch chung. Họ có thể sử dụng phép phân tích trực quan để đo lường mục tiêu và tiến độ cũng như cải thiện động lực của nhóm.

- Kể chuyện bằng dữ liệu (Data Storytelling)

Trực quan hóa dữ liệu là một phần quan trọng của việc kể chuyện bằng dữ liệu, giúp trình bày các kết quả phân tích một cách hấp dẫn, logic và dễ nhớ, dẫn dắt người xem qua các phát hiện quan trọng và thuyết phục họ về một kết luận hoặc hành động cụ thể.

2.4.2. Các loại biểu đồ thường dùng

Trong phân tích dữ liệu doanh thu game, một số loại biểu đồ phổ biến được sử dụng gồm biểu đồ cột, biểu đồ đường, biểu đồ tròn, v.v.

- Biểu đồ cột (Bar Chart) - So sánh doanh thu giữa các nhóm rời rạc

Biểu đồ cột là một công cụ trực quan mạnh mẽ để so sánh doanh thu giữa các danh mục hoặc nhóm rời rạc khác nhau. Trong bối cảnh phân tích dữ liệu doanh thu game, biểu đồ cột đặc biệt hữu ích khi chúng ta muốn so sánh tổng doanh thu hoặc doanh thu trung bình giữa các thể loại game khác nhau (Genre), ví dụ như so sánh doanh thu của thể loại Hành động, Thể thao, Nhập vai, v.v. Tương tự, chúng ta cũng có thể sử dụng biểu đồ cột để so sánh doanh thu giữa các nền tảng chơi game khác nhau (Platform) như PlayStation, Xbox, Nintendo Switch, PC, v.v. Chiều cao của mỗi cột sẽ tương ứng với giá trị doanh thu của từng nhóm, giúp người xem dễ dàng nhận diện và so sánh trực quan sự khác biệt về hiệu suất thương mại giữa các danh mục này.

- Biểu đồ tròn (Pie Chart) - Thể hiện tỷ trọng thị phần

Biểu đồ tròn là một lựa chọn trực quan hiệu quả để thể hiện tỷ lệ đóng góp của từng thành phần so với tổng thể. Trong phân tích dữ liệu doanh thu game, biểu đồ tròn thường được sử dụng để biểu diễn thị phần của từng thể loại game trên tổng doanh thu toàn cầu hoặc doanh thu của một khu vực cụ thể. Mỗi "miếng bánh" trong biểu đồ tròn đại diện cho một thể loại game, và kích thước của miếng bánh tương ứng với tỷ lệ phần trăm doanh thu mà thể loại đó đóng góp vào tổng doanh thu. Điều này giúp người xem nhanh chóng nắm bắt được cơ cấu thị phần và xác định được những thể loại game nào đang chiếm ưu thế hoặc có đóng góp lớn nhất vào tổng doanh thu. - *Biểu đồ đường (Line Chart) - Phân tích xu hướng doanh thu theo thời gian*

Biểu đồ đường là một công cụ không thể thiếu khi muốn phân tích sự thay đổi của doanh thu theo thời gian. Trong phân tích dữ liệu doanh thu game, biểu đồ đường đặc biệt hữu ích khi chúng ta muốn theo dõi xu hướng doanh thu của toàn bộ thị trường, của một thể loại game cụ thể, hoặc của một nền tảng nhất định qua các năm phát hành. Trục ngang của biểu đồ thường biểu diễn thời gian (ví dụ: năm), và trục dọc biểu diễn giá trị doanh thu. Đường biểu diễn sẽ kết nối các điểm dữ liệu theo thời

gian, cho thấy rõ ràng sự tăng trưởng, suy giảm hoặc biến động của doanh thu qua các giai đoạn khác nhau. Điều này giúp chúng ta nhận diện được các xu hướng dài hạn, các mùa vụ hoặc các sự kiện đặc biệt có thể ảnh hưởng đến doanh thu.

- Biểu đồ phân tán (Scatter Plot) - Khám phá mối quan hệ giữa các biến định lượng

Biểu đồ phân tán là một công cụ hữu ích để khám phá mối quan hệ giữa hai biến định lượng khác nhau. Trong phân tích dữ liệu doanh thu game, chúng ta có thể sử dụng biểu đồ phân tán để phân tích mối quan hệ giữa năm phát hành và doanh thu toàn cầu của các trò chơi. Mỗi điểm trên biểu đồ sẽ đại diện cho một trò chơi, với vị trí trên trục ngang thể hiện năm phát hành và vị trí trên trục dọc thể hiện doanh thu toàn cầu. Việc quan sát sự phân tán của các điểm có thể giúp chúng ta nhận diện được liệu có xu hướng tăng hoặc giảm doanh thu theo thời gian hay không, hoặc liệu có những năm nào có nhiều trò chơi đạt doanh thu đặc biệt cao hay không.

- Bản đồ nhiệt (Heatmap) - Thể hiện phân bố doanh thu theo nhiều chiều

Bản đồ nhiệt là một kỹ thuật trực quan hóa mạnh mẽ để thể hiện cường độ của dữ liệu trong một ma trận hai chiều. Trong phân tích dữ liệu doanh thu game, heatmap có thể được sử dụng để hiển thị doanh thu phân bố đồng thời theo hai biến khác nhau, ví dụ như thể loại game (Genre) và khu vực (Region). Ma trận sẽ có các ô tương ứng với từng cặp thể loại và khu vực, và màu sắc của mỗi ô sẽ thể hiện mức độ doanh thu (ví dụ: màu đậm hơn cho doanh thu cao hơn). Điều này giúp chúng ta dễ dàng nhận diện được những thể loại game nào đặc biệt phổ biến ở những khu vực nào, hoặc những khu vực nào có đóng góp doanh thu lớn cho từng thể loại game cụ thể.

2.4.3. Nguyên tắc lựa chọn và trình bày biểu đồ

2.4.3.1. Nguyên tắc lựa chọn biểu đồ

- Xác định rõ mục đích - Nền tảng cho việc lựa chọn biểu đồ

Nguyên tắc đầu tiên và quan trọng nhất trong việc lựa chọn biểu đồ trực quan hóa dữ liệu là phải hiểu rõ mục đích mà chúng ta muốn đạt được. Trước khi vẽ bất kỳ hình thức trực quan nào, cần phải trả lời câu hỏi: chúng ta muốn truyền tải thông điệp gì từ dữ liệu này? Mục tiêu có thể là để làm nổi bật một xu hướng tăng trưởng hay suy giảm theo thời gian, so sánh hiệu suất giữa các danh mục hoặc nhóm dữ liệu khác nhau, phân loại dữ liệu thành các phần riêng biệt, hay khám phá mối quan hệ và tương quan tiềm ẩn giữa các biến số. Việc xác định rõ mục đích này sẽ là kim chỉ nam, giúp chúng ta thu hẹp phạm vi lựa chọn và hướng đến loại biểu đồ phù hợp nhất để truyền tải thông tin một cách hiệu quả.

- Lựa chọn biểu đồ tương ứng với dữ liệu và mục tiêu

Sau khi đã xác định rõ mục đích, bước tiếp theo là lựa chọn loại biểu đồ phù hợp với bản chất của dữ liệu và mục tiêu truyền tải. Ví dụ, biểu đồ đường (Line chart) là lựa chọn lý tưởng để thể hiện các xu hướng biến đổi theo thời gian hoặc theo một trình tự nhất định. Biểu đồ cột (Bar chart) và biểu đồ hình tròn (Pie chart) thường được sử dụng để so sánh kích thước hoặc tỷ lệ của các phần hoặc nhóm dữ liệu khác nhau. Trong khi đó, biểu đồ phân tán (Scatter plot) lại là công cụ hữu ích để khám phá và hiển thị mối quan hệ tương quan giữa hai biến số. Cuối cùng, biểu đồ tần suất (Histogram) giúp chúng ta hình dung được sự phân phối xác suất của một tập dữ liệu liên tục. Việc lựa chọn đúng loại biểu đồ sẽ đảm bảo rằng thông tin được trình bày một cách trực quan và dễ hiểu nhất.

- Hạn chế sự đa dạng - Ưu tiên sự nhất quán và tập trung

Một nguyên tắc quan trọng khác cần ghi nhớ là nên tránh sử dụng quá nhiều loại biểu đồ khác nhau trong cùng một báo cáo hoặc trình bày. Việc sử dụng một loạt các kiểu biểu đồ khác nhau có thể gây ra sự rối mắt, làm mất đi sự tập trung của người xem và khiến thông điệp chính trở nên mờ nhạt. Thay vào đó, nên ưu tiên sử dụng một số loại biểu đồ cơ bản và phù hợp nhất, đồng thời duy trì sự nhất quán trong phong cách trình bày. Điều này không chỉ giúp báo cáo trở nên chuyên nghiệp hơn mà còn giúp người xem dễ dàng nắm bắt và so sánh thông tin giữa các phần khác nhau.

- Đảm bảo tính rõ ràng và dễ đọc - Ưu tiên sự trực quan

Cuối cùng, một biểu đồ trực quan hóa dữ liệu hiệu quả phải đảm bảo tính rõ ràng và dễ đọc. Việc lựa chọn kiểu biểu đồ phù hợp chỉ là một phần, quan trọng không kém là cách chúng ta thiết kế và trình bày biểu đồ đó. Cần đảm bảo rằng các yếu tố như tiêu đề, nhãn trục, chú thích và màu sắc được sử dụng một cách hợp lý và dễ hiểu. Tránh làm cho biểu đồ trở nên quá phức tạp hoặc chứa quá nhiều thông tin không cần thiết, gây rối mắt và khó khăn cho người xem trong việc giải mã dữ liệu. Một biểu đồ được thiết kế tốt sẽ truyền tải thông tin một cách trực quan, giúp người xem nhanh chóng nắm bắt được những điểm quan trọng nhất.

2.4.3.2. *Nguyên tắc trình bày biểu đồ*

- Tiêu đề rõ ràng – Kim chỉ nam cho sự hiểu biết

Nguyên tắc đầu tiên và tối quan trọng trong việc trình bày biểu đồ hiệu quả chính là việc xây dựng một tiêu đề rõ ràng và mô tả chính xác nội dung mà biểu đồ muốn truyền tải. Tiêu đề không chỉ đơn thuần là tên gọi, mà nó phải là một bản tóm tắt cô đọng, phản ánh đúng ý nghĩa và thông tin cốt lõi mà biểu đồ thể hiện. Một tiêu đề tốt sẽ giúp người xem ngay lập tức nắm bắt được chủ đề chính và phạm vi dữ liệu

được trình bày, tạo tiền đề cho việc tiếp thu và diễn giải thông tin một cách chính xác và nhanh chóng.

- Trục, nhãn và đơn vị - Nền tảng của sự chính xác

Để đảm bảo tính chính xác và dễ hiểu của biểu đồ, việc sử dụng trục, nhãn và đơn vị một cách rõ ràng và chính xác là điều không thể thiếu. Trục x (thường biểu diễn các danh mục hoặc thời gian) và trục y (thường biểu diễn các giá trị số) cần được ghi chú tên một cách đầy đủ, kèm theo đơn vị đo lường nếu có. Các nhãn dữ liệu, hiển thị giá trị cụ thể của từng điểm dữ liệu trên biểu đồ, phải được trình bày rõ ràng, với kích thước và màu sắc dễ đọc, tránh gây rối mắt và giúp người xem dễ dàng đối chiếu và so sánh các giá trị.

- Màu sắc hài hòa – Dẫn dắt sự chú ý

Việc lựa chọn màu sắc phù hợp và nhất quán đóng vai trò quan trọng trong việc tạo ra một biểu đồ trực quan dễ hiểu và dễ tiếp thu. Nên ưu tiên sử dụng một bảng màu hài hòa, tránh sử dụng quá nhiều màu sắc khác nhau có thể gây xao nhãng và khó phân biệt. Màu sắc nên được sử dụng một cách có chủ đích để phân biệt rõ ràng các nhóm dữ liệu khác nhau, làm nổi bật những thông tin quan trọng hoặc tuân theo một quy ước màu sắc đã được thiết lập (nếu có), đảm bảo tính nhất quán trong toàn bộ báo cáo hoặc trình bày.

- Tối giản hóa – Tập trung vào thông điệp cốt lõi

Một biểu đồ hiệu quả là một biểu đồ tối giản, tập trung vào việc truyền tải thông điệp chính mà không bị xao nhãng bởi các yếu tố không cần thiết. Nên tránh sử dụng quá nhiều hiệu ứng đồ họa phức tạp, các đường kẻ lưới quá dày, hoặc bất kỳ thành phần trang trí nào không đóng góp vào việc hiểu dữ liệu. Việc giữ cho biểu đồ gọn gàng và tập trung vào nội dung chính sẽ giúp người xem dễ dàng nắm bắt được thông tin quan trọng nhất mà không bị phân tâm.

- Chú thích rõ ràng, giải thích mọi chi tiết

Trong trường hợp biểu đồ sử dụng các ký hiệu, chú thích đặc biệt hoặc các chú giải để làm rõ một số khía cạnh của dữ liệu, việc cung cấp chú thích rõ ràng và hợp lý là rất cần thiết. Các ký hiệu phải được giải thích đầy đủ, các chú thích cần được đặt ở vị trí dễ thấy và liên quan trực tiếp đến phần dữ liệu được chú thích, giúp người xem hiểu được ý nghĩa của từng yếu tố trên biểu đồ mà không gặp bất kỳ sự mơ hồ nào.

- Tỷ lệ cân đối, tránh gây hiểu lầm

Việc lựa chọn tỷ lệ phù hợp cho các trục của biểu đồ là một nguyên tắc quan trọng để tránh gây ra những hiểu lầm về quy mô và sự khác biệt giữa các giá trị dữ

liệu. Sử dụng tỷ lệ không cân đối, chẳng hạn như cắt trục y một cách tùy tiện hoặc sử dụng các thang đo không đồng đều, có thể tạo ra những ấn tượng sai lệch về mức độ thay đổi hoặc sự khác biệt giữa các nhóm dữ liệu. Do đó, việc lựa chọn tỷ lệ một cách cẩn thận và trung thực là yếu tố then chốt để đảm bảo tính chính xác và khách quan của thông tin được trình bày.

2.5. GIỚI THIỆU CÁC CÔNG CỤ VÀ THƯ VIỆN SỬ DỤNG

Trong quá trình phân tích dữ liệu doanh thu trò chơi điện tử, đề tài sử dụng ngôn ngữ lập trình Python cùng một số thư viện phổ biến để xử lý, phân tích và trực quan hóa dữ liệu. Các công cụ này không chỉ mạnh mẽ mà còn thân thiện, dễ sử dụng đối với phần lớn người dùng từ người mới bắt đầu đến những người có nhiều kinh nghiệm.

2.5.1. Python - Ngôn ngữ lập trình mạnh mẽ cho phân tích dữ liệu

Python là một ngôn ngữ lập trình bậc cao, mã nguồn mở và đa nền tảng, được sử dụng rộng rãi để phát triển các ứng dụng web, phát triển phần mềm, khoa học dữ liệu và máy học (Machine Learning). Python có cú pháp rõ ràng và ngắn gọn, giúp cho việc học và sử dụng ngôn ngữ này trở nên dễ dàng.

Python được thiết kế với tư tưởng giúp người học dễ đọc, dễ hiểu và dễ nhớ. Vì thế nó có hình thức rất gọn gàng, cấu trúc rõ ràng, thuận tiện cho người mới học. Cấu trúc của Python cho phép người sử dụng viết mã lệnh với số lần gõ phím tối thiểu, nói cách khác, so với các ngôn ngữ lập trình khác, người dùng có thể sử dụng ít dòng code hơn để viết ra một chương trình trong Python. Python sử dụng một trình thông dịch để chạy mã, khi viết không cần phải biên dịch nó thành mã máy trước khi chạy. Thay vào đó, trình thông dịch sẽ đọc và thực thi mã trực tiếp, từng dòng một.

Python là một ngôn ngữ lập trình đa mẫu hình, nó hỗ trợ hoàn toàn mẫu lập trình hướng đối tượng và lập trình cấu trúc. Ngoài ra, về mặt tính năng, Python cũng hỗ trợ lập trình hàm và lập trình hướng khía cạnh. Nhờ vậy mà Python có thể làm được rất nhiều thứ, sử dụng trong nhiều lĩnh vực khác nhau. Python có sẵn các cấu trúc dữ liệu mạnh mẽ như list, dictionary, tuple, dễ dàng xử lý và lưu trữ dữ liệu. Ngoài ra không cần phải khai báo kiểu dữ liệu cho các biến, nó sẽ tự động xác định kiểu dữ liệu dựa trên giá trị của biến.

Trong dự án phân tích dữ liệu doanh thu trò chơi điện tử này, Python được lựa chọn làm ngôn ngữ lập trình chính nhờ vào sự mạnh mẽ, linh hoạt và hệ sinh thái thư viện phong phú, đặc biệt trong lĩnh vực khoa học dữ liệu. Python cung cấp cú pháp rõ ràng, dễ đọc và một cộng đồng người dùng lớn mạnh, tạo điều kiện thuận lợi cho việc phát triển và triển khai các tác vụ phân tích phức tạp một cách hiệu quả. Khả năng tích

hợp dễ dàng với các thư viện chuyên dụng khác đã biến Python trở thành một công cụ không thể thiếu cho bất kỳ dự án phân tích dữ liệu nào.

2.5.2. Pandas - "Trái tim" của việc xử lý và thao tác dữ liệu

Thư viện pandas là một thư viện mã nguồn mở, hỗ trợ đắc lực trong thao tác dữ liệu. Đây cũng là bộ công cụ phân tích và xử lý dữ liệu mạnh mẽ của ngôn ngữ lập trình python. Thư viện này được sử dụng rộng rãi trong cả nghiên cứu lẫn phát triển các ứng dụng về khoa học dữ liệu. Thư viện này sử dụng một cấu trúc dữ liệu riêng là Dataframe. Pandas cung cấp rất nhiều chức năng xử lý và làm việc trên cấu trúc dữ liệu này. Chính sự linh hoạt và hiệu quả đã khiến cho pandas được sử dụng rộng rãi.

Pandas là một công cụ không thể thiếu trong quá trình phân tích dữ liệu nhờ những ưu điểm vượt trội mà nó mang lại. Cấu trúc dữ liệu DataFrame của Pandas cung cấp sự linh hoạt và hiệu quả cao trong việc thao tác và lập chỉ mục dữ liệu, cho phép người dùng dễ dàng truy cập và quản lý thông tin. Bên cạnh đó, Pandas đóng vai trò như một cầu nối mạnh mẽ, hỗ trợ đọc và ghi dữ liệu một cách thuận tiện giữa bộ nhớ và đa dạng các định dạng tệp phổ biến như CSV, text, Excel, v.v. Một điểm mạnh đáng chú ý khác của Pandas là khả năng liên kết dữ liệu một cách thông minh và xử lý hiệu quả các trường hợp dữ liệu bị thiếu, đồng thời tự động chuyển đổi dữ liệu lộn xộn về dạng có cấu trúc, giúp đơn giản hóa quá trình làm sạch dữ liệu. Thư viện này còn cung cấp các công cụ trực quan để dễ dàng thay đổi bố cục của dữ liệu, tích hợp cơ chế trượt, lập chỉ mục và trích xuất các tập con từ những bộ dữ liệu lớn một cách nhanh chóng. Việc thêm, xóa các cột dữ liệu cũng trở nên đơn giản, cùng với khả năng tập hợp hoặc thay đổi dữ liệu thông qua hàm “groupby”, cho phép thực hiện các phép toán trên các nhóm dữ liệu một cách mạnh mẽ. Hơn nữa, Pandas thể hiện hiệu quả cao trong việc trộn và kết hợp các tập dữ liệu khác nhau. Khả năng lập chỉ mục theo nhiều chiều của dữ liệu giúp thao tác dễ dàng giữa dữ liệu cao chiều và thấp chiều. Về hiệu năng, Pandas được tối ưu hóa để xử lý dữ liệu lớn một cách nhanh chóng. Cuối cùng, sự phổ biến rộng rãi của Pandas trong cả lĩnh vực học thuật và thương mại, bao gồm thống kê, thương mại, phân tích và quảng cáo, đã chứng minh giá trị và tầm quan trọng của nó trong cộng đồng phân tích dữ liệu.

Để quản lý, xử lý và thao tác với bộ dữ liệu "vgsales.csv" một cách hiệu quả, thư viện Pandas của Python đã được sử dụng rộng rãi. Pandas cung cấp cấu trúc dữ liệu DataFrame mạnh mẽ, cho phép biểu diễn dữ liệu dạng bảng một cách trực quan và cung cấp nhiều công cụ linh hoạt để đọc, ghi, làm sạch, biến đổi và tổng hợp dữ liệu. Với Pandas, việc lọc, sắp xếp, nhóm và kết hợp dữ liệu trở nên đơn giản và nhanh chóng, tạo điều kiện thuận lợi cho việc khám phá và chuẩn bị dữ liệu trước khi tiến hành các phân tích sâu hơn và trực quan hóa.

2.5.3. Matplotlib và Seaborn – Thư viện chính được sử dụng để trực quan hóa dữ liệu

- Matplotlib – Nền tảng vững chắc cho trực quan hóa dữ liệu

Trong quá trình phân tích dữ liệu của đề tài, thư viện Matplotlib đóng vai trò là một công cụ trực quan hóa cơ bản nhưng vô cùng mạnh mẽ và linh hoạt. Với khả năng tạo ra nhiều loại biểu đồ khác nhau như biểu đồ cột (bar chart) để so sánh các danh mục, biểu đồ đường (line chart) để theo dõi xu hướng theo thời gian, biểu đồ tần suất (histogram) để hiểu phân phối dữ liệu và nhiều loại biểu đồ thống kê khác. Matplotlib cung cấp nền tảng vững chắc để khám phá và trình bày các đặc điểm quan trọng của bộ dữ liệu "vgsales.csv". Sự linh hoạt của Matplotlib cho phép tùy chỉnh sâu sắc các yếu tố của biểu đồ, từ nhãn trục, tiêu đề, màu sắc đến kiểu dáng đường kẻ, đáp ứng nhiều nhu cầu trực quan hóa khác nhau trong dự án.

- Seaborn – Thẩm mỹ và hiệu quả

Để nâng cao khả năng trực quan hóa và tạo ra những biểu đồ phức tạp và thẩm mỹ hơn, đề tài này còn sử dụng thư viện Seaborn. Xây dựng dựa trên nền tảng của Matplotlib, Seaborn cung cấp một bộ sưu tập các biểu đồ thống kê nâng cao, được thiết kế để khám phá các mối quan hệ giữa các biến một cách trực quan và hiệu quả. Các biểu đồ phân tán (scatterplot) trong Seaborn giúp dễ dàng nhận diện các tương quan giữa hai biến số, biểu đồ nhiệt (heatmap) trực quan hóa ma trận tương quan hoặc sự phân bố dữ liệu theo hai chiều, và biểu đồ hộp (boxplot) là công cụ mạnh mẽ để so sánh phân phối của một biến số giữa các nhóm khác nhau và phát hiện các giá trị ngoại lệ. Với giao diện lập trình trực quan và thiết kế biểu đồ đẹp mắt, Seaborn giúp đơn giản hóa quá trình tạo ra những hình ảnh trực quan sâu sắc và dễ hiểu từ bộ dữ liệu "vgsales.csv".

2.5.4. NumPy - Nền tảng cho các phép toán số học hiệu suất cao

Mặc dù có thể không được sử dụng trực tiếp cho mọi tác vụ trực quan hóa và xử lý dữ liệu cấp cao trong đề tài này, thư viện NumPy đóng vai trò là nền tảng cho các phép toán số học hiệu suất cao trong Python. NumPy cung cấp cấu trúc mảng đa chiều mạnh mẽ (ndarray) và một bộ sưu tập lớn các hàm toán học để thực hiện các phép tính trên mảng một cách nhanh chóng và hiệu quả. Trong trường hợp cần thực hiện các phép biến đổi dữ liệu phức tạp hoặc tính toán thống kê nâng cao, NumPy có thể được tích hợp liền mạch với Pandas và các thư viện khác, góp phần tối ưu hóa hiệu suất của quá trình phân tích.

2.6. HỆ THỐNG GỢI Ý (RECOMMENDATION SYSTEMS)

Trong thời đại dữ liệu số hiện nay, hệ thống gợi ý (Recommendation System) đóng vai trò quan trọng trong việc cá nhân hóa trải nghiệm người dùng và tối ưu hóa

doanh thu cho nhiều nền tảng số, đặc biệt là trong lĩnh vực giải trí như thương mại điện tử, phát trực tuyến và trò chơi điện tử. Trong ngành công nghiệp game, hệ thống gợi ý giúp đề xuất các trò chơi phù hợp với sở thích người chơi, từ đó tăng khả năng tương tác, giữ chân người dùng và thúc đẩy tiêu thụ sản phẩm.

2.6.1. Tổng quan về hệ thống gợi ý

Hệ thống gợi ý là một dạng của hệ hỗ trợ ra quyết định, cung cấp giải pháp mang tính cá nhân hóa mà không phải trải qua quá trình tìm kiếm phức tạp. Hệ gợi ý học từ người dùng và gợi ý các sản phẩm tốt nhất trong số các sản phẩm phù hợp. Nó sử dụng các tri thức về sản phẩm, các tri thức của chuyên gia hay tri thức khai phá học được từ hành vi con người dùng để đưa ra các gợi ý về sản phẩm mà họ thích trong hàng ngàn hàng vạn sản phẩm có trong hệ thống. Các website thương mại điện tử, ví dụ như sách, phim, nhạc, báo...sử dụng hệ thống gợi ý để cung cấp các thông tin giúp cho người sử dụng quyết định sẽ lựa chọn sản phẩm nào. Các sản phẩm được gợi ý dựa trên số lượng sản phẩm đó đã được bán, dựa trên các thông tin cá nhân của người sử dụng, dựa trên sự phân tích hành vi mua hàng trước đó của người sử dụng để đưa ra các dự đoán về hành vi mua hàng trong tương lai của chính khách hàng đó. Các dạng gợi ý bao gồm: gợi ý các sản phẩm tới người tiêu dùng, các thông tin sản phẩm mang tính cá nhân hóa, tổng kết các ý kiến cộng đồng, và cung cấp các chia sẻ, các phê bình, đánh giá mang tính cộng đồng liên quan tới yêu cầu, mục đích của người sử dụng đó.

2.6.2. Các phương pháp gợi ý phổ biến

- Gợi ý dựa trên nội dung - Tìm kiếm sự tương đồng trong đặc điểm sản phẩm

Một trong những phương pháp phổ biến để xây dựng hệ thống gợi ý là gợi ý dựa trên nội dung (Content-Based Filtering). Phương pháp này tập trung vào việc phân tích các đặc điểm vốn có của sản phẩm, trong trường hợp này là các trò chơi điện tử. Các thuộc tính như thể loại game (ví dụ: hành động, nhập vai, chiến lược), nền tảng phát hành (ví dụ: PC, PlayStation, Xbox), nhà phát hành và các đặc điểm mô tả khác của trò chơi được xem xét. Khi một người dùng đã thể hiện sự yêu thích hoặc tương tác tích cực với một số trò chơi nhất định, hệ thống sẽ tìm kiếm và gợi ý những trò chơi khác có các đặc điểm tương tự. Ví dụ, nếu một người chơi yêu thích các game hành động phiêu lưu trên PlayStation của một nhà phát hành cụ thể, hệ thống sẽ gợi ý các trò chơi khác thuộc cùng thể loại, trên cùng nền tảng hoặc từ cùng nhà phát hành mà người đó chưa từng chơi.

- Gợi ý dựa trên lọc cộng tác - Sức mạnh của sự tương đồng hành vi người dùng

Một phương pháp mạnh mẽ khác là gợi ý dựa trên cộng tác (Collaborative Filtering). Thay vì phân tích nội dung của sản phẩm, phương pháp này tập trung vào hành vi của người dùng. Ý tưởng cốt lõi là nếu hai người dùng khác nhau có những sở thích tương đồng về một số lượng lớn các trò chơi (ví dụ: cả hai đều thích các game A,

B và C), thì những trò chơi mà một người thích (ví dụ: người A thích game D) mà người kia chưa biết đến (người B chưa chơi game D) có khả năng cao cũng sẽ được người kia yêu thích. Phương pháp này không đòi hỏi phải hiểu chi tiết về nội dung của từng trò chơi mà chỉ dựa vào ma trận tương tác giữa người dùng và sản phẩm (ví dụ: lịch sử mua hàng, đánh giá, thời gian chơi).

- Gợi ý kết hợp - Tận dụng sức mạnh tổng hợp để cá nhân hóa tối ưu

Để tận dụng những ưu điểm và hạn chế của cả hai phương pháp trên, gợi ý kết hợp (Hybrid Filtering) đã trở thành một xu hướng phổ biến trong việc xây dựng các hệ thống gợi ý tiên tiến. Phương pháp này kết hợp cả phân tích nội dung của sản phẩm và phân tích hành vi của người dùng để đưa ra những khuyến nghị chính xác và cá nhân hóa hơn. Bằng cách kết hợp thông tin về đặc điểm của trò chơi với thông tin về sở thích và hành vi của người dùng, hệ thống có thể vượt qua những nhược điểm riêng lẻ của từng phương pháp. Ví dụ, gợi ý kết hợp có thể giải quyết vấn đề "khởi đầu lạnh" (cold start problem) khi chưa có đủ dữ liệu tương tác của người dùng mới bằng cách dựa vào thông tin nội dung, đồng thời vẫn tận dụng được sức mạnh của cộng tác khi có đủ dữ liệu người dùng để đưa ra những gợi ý tinh tế hơn dựa trên sở thích của những người dùng tương tự.

2.6.3. Ứng dụng trong đề tài

Trong phạm vi đề tài này, hệ thống gợi ý có thể được xây dựng dựa trên cả ba yếu tố:

- Xây dựng gợi ý dựa trên nội dung trò chơi

Trong phạm vi đề tài phân tích dữ liệu doanh thu trò chơi điện tử, một hệ thống gợi ý có thể được xây dựng dựa trên nội dung của chính các trò chơi. Bằng cách phân tích các thuộc tính như thể loại (ví dụ: hành động, chiến lược, thể thao), nền tảng phát hành (ví dụ: PC, PS4, Xbox One) và nhà phát hành, chúng ta có thể xác định sự tương đồng giữa các trò chơi. Ví dụ, nếu một người dùng quan tâm đến các trò chơi hành động phiêu lưu trên nền tảng PlayStation do một nhà phát hành cụ thể sản xuất, hệ thống có thể gợi ý các trò chơi khác cùng thể loại, trên cùng nền tảng hoặc từ cùng nhà phát hành. Cách tiếp cận này tận dụng thông tin sẵn có trong bộ dữ liệu "vgsales.csv" và không yêu cầu thông tin trực tiếp về sở thích cá nhân của từng người dùng.

- Tích hợp thông tin người dùng (mô phỏng hoặc tiềm năng)

Để tăng cường khả năng cá nhân hóa của hệ thống gợi ý, chúng ta có thể tích hợp thông tin về người dùng. Mặc dù bộ dữ liệu "vgsales.csv" hiện tại không chứa thông tin người dùng, chúng ta có thể mô phỏng thông tin này dựa trên các phân khúc thị trường tiềm năng hoặc mở rộng bộ dữ liệu bằng cách kết hợp với các nguồn thông

tin khác (nếu có). Ví dụ, chúng ta có thể giả định các phân khúc người dùng khác nhau dựa trên sở thích thể loại (người thích game hành động, người thích game chiến thuật), nền tảng ưa chuộng (game thủ PC, game thủ console) hoặc thậm chí các yếu tố nhân khẩu học (nếu có thông tin). Sau đó, hệ thống gợi ý có thể kết hợp thông tin về nội dung trò chơi với thông tin về phân khúc người dùng để đưa ra những gợi ý phù hợp hơn. Ví dụ, một người dùng thuộc phân khúc "người thích game hành động trên PC" có thể được gợi ý các trò chơi hành động mới phát hành trên PC có doanh số cao hoặc được đánh giá tích cực.

- Kết hợp nội dung và thông tin người dùng để tạo gợi ý cá nhân hóa

Bằng cách kết hợp cả thông tin về nội dung trò chơi và thông tin (mô phỏng hoặc tiềm năng) về người dùng, chúng ta có thể xây dựng một hệ thống gợi ý mạnh mẽ hơn. Hệ thống có thể bắt đầu bằng việc xác định các trò chơi tương tự về nội dung với những gì một phân khúc người dùng cụ thể có xu hướng ưa thích. Sau đó, nó có thể lọc và sắp xếp các gợi ý này dựa trên các yếu tố khác như doanh số (cho thấy mức độ phổ biến), đánh giá (nếu có) hoặc các thuộc tính cụ thể khác mà phân khúc người dùng đó có thể quan tâm. Cách tiếp cận kết hợp này tận dụng được những ưu điểm của cả gợi ý dựa trên nội dung (không cần dữ liệu người dùng chi tiết ban đầu) và khả năng cá nhân hóa (khi có thông tin về phân khúc người dùng), mang lại những gợi ý tiềm năng và phù hợp hơn.

CHƯƠNG III. THỰC NGHIỆM

Toàn bộ mã nguồn được sử dụng trong quá trình trực quan và phân tích dữ liệu doanh thu được xây dựng bằng ngôn ngữ lập trình Python. Mã nguồn bao gồm các bước: tiền xử lý dữ liệu, trực quan hóa, xây dựng mô hình, đánh giá hiệu suất và rút ra nhận định cuối cùng.

Link mã nguồn:

(*https://github.com/thaodau080204/DA2-Phantichdulieudoanhthutrochoi*)

3.1. TRỰC QUAN VÀ PHÂN TÍCH DỮ LIỆU DOANH THU

Để hiểu rõ hơn về bức tranh toàn cảnh của ngành công nghiệp trò chơi điện tử, trước tiên đề tài tiến hành phân tích và trực quan hóa các thông tin quan trọng từ bộ dữ liệu. Các biểu đồ doanh thu theo khu vực, thể loại, nền tảng và nhà phát hành sẽ giúp làm sáng tỏ những yếu tố ảnh hưởng lớn đến thành công của trò chơi, đồng thời cung cấp dữ liệu nền tảng phục vụ cho các bước xây dựng mô hình gợi ý sau này.

3.1.1. Dữ liệu và tiền xử lý dữ liệu

Bộ dữ liệu "vgsales.csv" có dạng như sau:

	Rank	Name	Platform	Year	Genre	Publisher	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
0	1	Wii Sports	Wii	2006.0	Sports	Nintendo	41.49	29.02	3.77	8.46	82.74
1	2	Super Mario Bros.	NES	1985.0	Platform	Nintendo	29.08	3.58	6.81	0.77	40.24
2	3	Mario Kart Wii	Wii	2008.0	Racing	Nintendo	15.85	12.88	3.79	3.31	35.82
3	4	Wii Sports Resort	Wii	2009.0	Sports	Nintendo	15.75	11.01	3.28	2.96	33.00
4	5	Pokemon Red/Pokemon Blue	GB	1996.0	Role-Playing	Nintendo	11.27	8.89	10.22	1.00	31.37
5	6	Tetris	GB	1989.0	Puzzle	Nintendo	23.20	2.26	4.22	0.58	30.26
6	7	New Super Mario Bros.	DS	2006.0	Platform	Nintendo	11.38	9.23	6.50	2.90	30.01
7	8	Wii Play	Wii	2006.0	Misc	Nintendo	14.03	9.20	2.93	2.85	29.02
8	9	New Super Mario Bros. Wii	Wii	2009.0	Platform	Nintendo	14.59	7.06	4.70	2.26	28.62
9	10	Duck Hunt	NES	1984.0	Shooter	Nintendo	26.93	0.63	0.28	0.47	28.31

Hình 3-1: 10 dòng đầu của bộ dữ liệu "vgsales.csv"

Trước khi làm việc với dữ liệu, cần tìm hiểu các thông tin cơ bản của dữ liệu như số hàng, số cột, ý nghĩa các cột v.v. Để biết được những thông tin trên cần dùng lệnh "info()". Kết quả nhận được như sau:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 16598 entries, 0 to 16597
Data columns (total 11 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Rank          16598 non-null  int64
 1   Name          16598 non-null  object
 2   Platform      16598 non-null  object
 3   Year          16327 non-null  float64
 4   Genre         16598 non-null  object
 5   Publisher     16540 non-null  object
 6   NA_Sales      16598 non-null  float64
 7   EU_Sales      16598 non-null  float64
 8   JP_Sales      16598 non-null  float64
 9   Other_Sales   16598 non-null  float64
 10  Global_Sales  16598 non-null  float64
dtypes: float64(6), int64(1), object(4)
memory usage: 1.4+ MB
```

Hình 3-2: Thông tin cơ bản của dữ liệu

Bộ dữ liệu có 16,598 dòng, 11 cột gồm: Rank (thứ tự), Name (tên game), Platform (nền tảng), Year (năm phát hành), Genre (thể loại), Publisher (nhà phát hành), NA_Sales (doanh số khu vực Bắc Mỹ), EU_Sales (doanh số khu vực Châu Âu), JP_Sales (doanh số khu vực Nhật Bản), Other_Sales (doanh số các khu vực khác), Global_Sales (doanh số toàn thế giới). Có 6 thuộc tính kiểu dữ liệu dạng float64 là Year, NA_Sales, EU_Sales, JP_Sales, Other_Sales và Global_Sales. Có 1 thuộc tính có kiểu dữ liệu int64 là Rank. Còn lại 4 thuộc tính có kiểu dữ liệu object là Name, Platform, Genre, Publisher.

Ngoài các thông tin cơ bản, cần phải nắm rõ nội dung tóm tắt thống kê mô tả của các thuộc tính dạng số bằng hàm “describe()”. Kết quả nhận được:

	Rank	Year	NA_Sales	EU_Sales	JP_Sales	Other_Sales	Global_Sales
count	16598.000000	16327.000000	16598.000000	16598.000000	16598.000000	16598.000000	16598.000000
mean	8300.605254	2006.406443	0.264667	0.146652	0.077782	0.048063	0.537441
std	4791.853933	5.828981	0.816683	0.505351	0.309291	0.188588	1.555028
min	1.000000	1980.000000	0.000000	0.000000	0.000000	0.000000	0.010000
25%	4151.250000	2003.000000	0.000000	0.000000	0.000000	0.000000	0.060000
50%	8300.500000	2007.000000	0.080000	0.020000	0.000000	0.010000	0.170000
75%	12449.750000	2010.000000	0.240000	0.110000	0.040000	0.040000	0.470000
max	16600.000000	2020.000000	41.490000	29.020000	10.220000	10.570000	82.740000

Hình 3-3: Bảng tóm tắt thống kê mô tả

Bảng thống kê mô tả cho thấy rằng doanh thu trung bình toàn cầu của mỗi trò chơi điện tử là khoảng 0.537 triệu bản, tuy nhiên giá trị trung vị chỉ là 0.17 triệu bản, phản ánh sự phân bố không đồng đều với nhiều tựa game có doanh số thấp và chỉ một số rất ít đạt doanh số cực cao. Trong số các khu vực, Bắc Mỹ là thị trường chiếm tỷ trọng doanh số lớn nhất, tiếp theo là châu Âu và Nhật Bản. Dữ liệu bao phủ trong giai

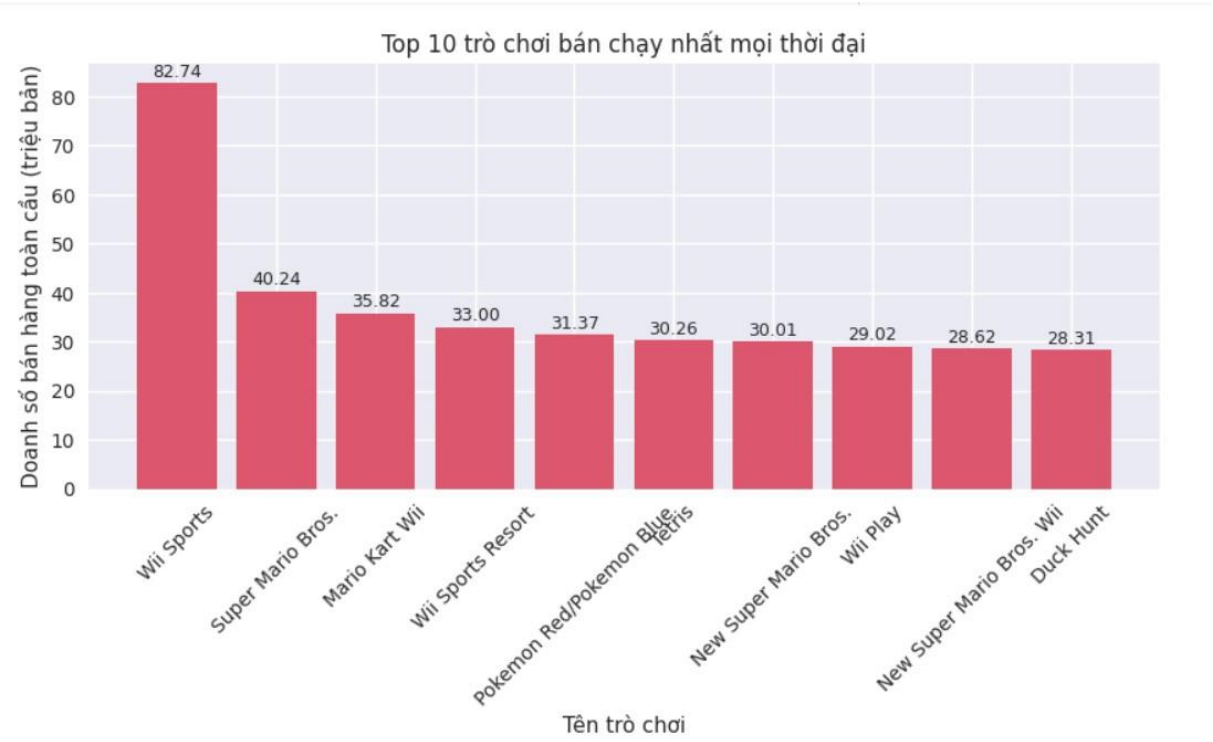
đoạn từ năm 1980 đến 2020, với phần lớn trò chơi được phát hành từ sau năm 2000, phản ánh xu hướng phát triển mạnh mẽ trong những thập kỷ gần đây.

Để đảm bảo chất lượng dữ liệu cho phân tích, cần thực hiện kiểm tra giá trị thiếu và lọc dữ liệu không hợp lệ. Hàm "isnull()" được dùng cho mục đích kiểm tra giá trị thiếu. Sau khi kiểm tra nhận thấy dữ liệu biến Year có 271, biến Publisher có 58 giá trị thiếu, cần tiến hành xử lý. Bằng hàm "dropna()" dữ liệu thiếu đã được xử lý, từ 16,598 bản ghi ban đầu còn khoảng 16,291 bản ghi sau khi làm sạch. Hoàn tất quá trình làm sạch cũng là lúc dữ liệu sẵn sàng được đưa vào phân tích.

3.1.2. Trực quan hóa dữ liệu doanh thu trên toàn cầu

Trước hết, đề tài tiến hành phân tích tổng quan doanh thu của các trò chơi điện tử trên phạm vi toàn thế giới nhằm xác định những đặc điểm nổi bật trong thị trường toàn cầu. Việc đánh giá doanh thu theo từng thể loại game, nền tảng, nhà phát hành và thời gian phát hành sẽ giúp nhận diện những xu hướng chi phối ngành công nghiệp game trong giai đoạn được nghiên cứu.

Bước đầu tiên trong quá trình trực quan là tìm hiểu về top 10 trò chơi bán chạy nhất mọi thời đại.

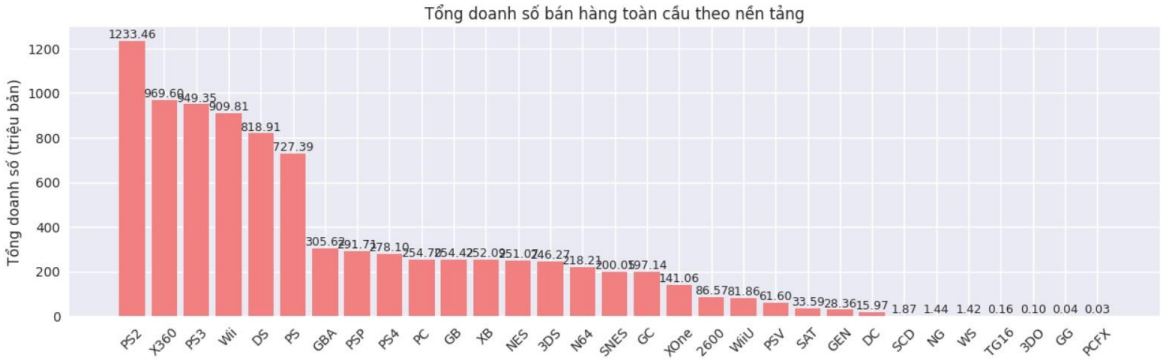


Hình 3-4: Top 10 trò chơi bán chạy nhất mọi thời đại

Với 82.74 triệu bản, *Wii Sports* dẫn đầu một cách vượt trội, gần gấp đôi so với vị trí thứ hai (*Super Mario Bros.* – 40.24 triệu). Nguyên nhân là do *Wii Sports* thường được bundled (bán kèm) *với* máy *Wii* nên có doanh số vượt trội. Nhà phát hành

Nintendo thống trị hoàn toàn danh sách: Tất cả 10 trò chơi trong biểu đồ đều thuộc hệ sinh thái của Nintendo. Những cái tên như *Mario*, *Wii*, *Pokemon*, *Tetris* đều là thương hiệu lớn gắn liền với Nintendo trong nhiều thập kỷ. Để làm được điều này, họ đã xây dựng thành công chiến lược dài hạn cược kỳ thông mình và khác biệt. Họ chọn cách đi riêng, tạo ra hệ sinh thái giải trí hoàn chỉnh, dễ tiếp cận, gắn bó cảm xúc và sáng tạo.mang lại hiệu quả vượt trội về doanh số và độ phủ toàn cầu.

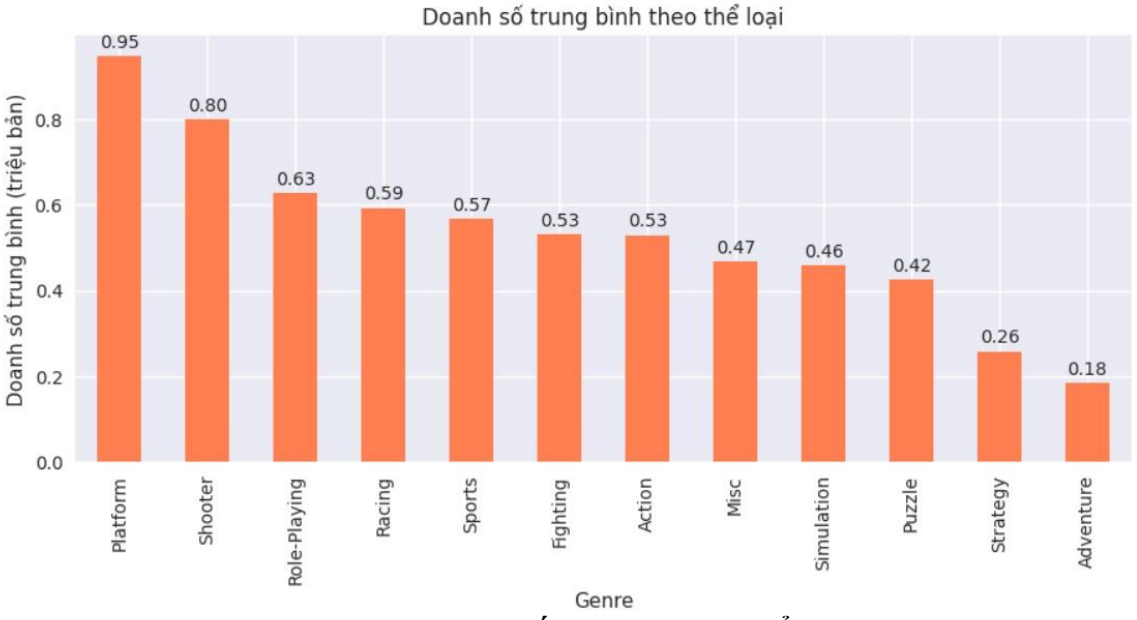
Góp phần quan trọng trong doanh thu là nền tảng chơi game, một số nền tảng cực kỳ phát triển nhưng cũng có một số nền tảng lại khá "bèo bọt".



Hình 3-5: Tổng doanh số bán hàng toàn cầu theo nền tảng

Có thể thấy, PlayStation (PS2) là nền tảng bán chạy nhất lịch sử, dẫn đầu toàn bộ biểu đồ với 1233.46 triệu bản. Nhờ thư viện game đa dạng, thời gian sống lâu dài và khả năng tương thích ngược, Sony thực sự là nhà thống trị trong thế hệ console thứ 6. Đứng thứ 2 và 3 là Xbox 360 (969.6 triệu bản) và PS3 (949.35 triệu bản) bám sát nhau, thể hiện cuộc đua khốc liệt giữa Microsoft và Sony trong thế hệ thứ 7. Cả hai đều có hệ sinh thái mạnh, thư viện game lớn và được hỗ trợ dài hạn. Dù đã ra mắt từ thập niên 80–90, các hệ máy như NES (251.07 triệu bản), SNES (200.5 triệu bản), N64 (218.21 triệu bản) vẫn có chỗ đứng với tổng doanh số đáng nể. Chứng tỏ tầm ảnh hưởng lịch sử sâu rộng của Nintendo. Một số nền tảng như SCD (1.87 triệu), NG (1.44 triệu), PCFX (0.03 triệu) gần như không có sức ảnh hưởng thị trường. Đây thường là các máy ít game, giá cao hoặc ra mắt không đúng thời điểm.

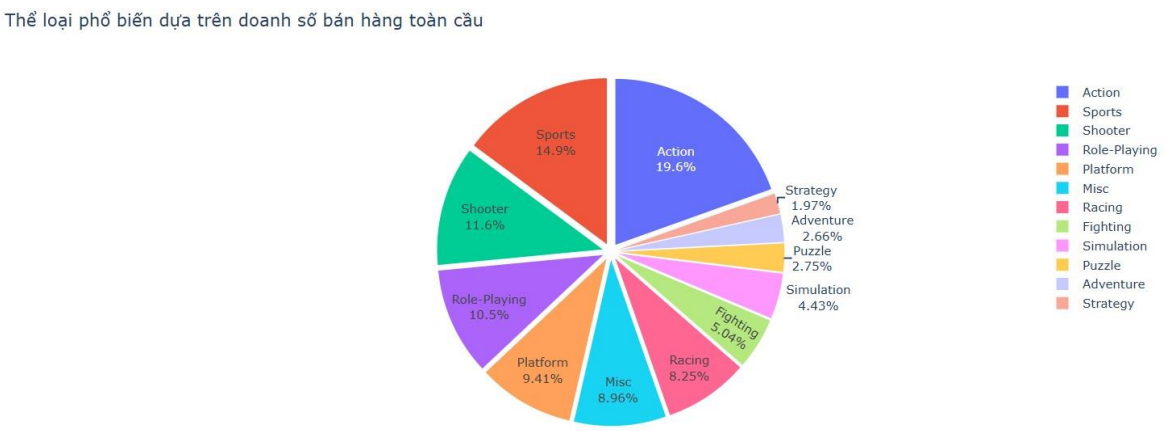
Sau khi đã xem xét tổng doanh số bán hàng theo từng trò chơi và nền tảng, chúng ta tiếp tục đi sâu vào khía cạnh thể loại trò chơi – một yếu tố quan trọng ảnh hưởng trực tiếp đến mức độ phổ biến và doanh thu của các tựa game.



Hình 3-6: Doanh số trung bình theo thể loại

Bằng cách phân tích doanh số trung bình theo thể loại, ta có thể nhận diện được những dòng game nào được ưa chuộng nhất trên thị trường toàn cầu, cũng như đánh giá mức độ hấp dẫn của từng thể loại đối với người chơi. Biểu đồ trên thể hiện doanh số trung bình (triệu bản) của các trò chơi theo từng thể loại. Có thể thấy rằng thể loại Platform (đi cảnh) dẫn đầu với doanh số trung bình cao nhất là 0.95 triệu bản, cho thấy sự phổ biến bền vững của các tựa game như *Super Mario Bros* vốn đã in sâu vào tiềm thức người chơi nhiều thế hệ. Xếp sau là Shooter (bắn súng) với 0.80 triệu bản, phản ánh sức hút mạnh mẽ của các dòng game hành động, kịch tính như *Call of Duty* hay *Halo*. Các thể loại như Role-Playing, Racing, và Sports cũng có doanh số trung bình khá cao, dao động từ 0.59 đến 0.63 triệu bản, cho thấy đây là các dòng game được yêu thích và đầu tư lớn. Ngược lại, các thể loại như Strategy (chiến thuật – 0.26 triệu bản) và Adventure (phiêu lưu – chỉ 0.18 triệu bản) có doanh số trung bình thấp hơn đáng kể. Điều này có thể xuất phát từ việc đây là những thể loại thường kén người chơi hoặc có nhịp chơi chậm, không phù hợp với thị hiếu đại chúng hiện nay. Từ đó, ta thấy rõ rằng thị trường game hiện tại vẫn ưu tiên các thể loại dễ tiếp cận, mang tính hành động hoặc giải trí cao.

Bên cạnh việc phân tích doanh số trung bình, một góc nhìn quan trọng khác để đánh giá mức độ phổ biến của các thể loại trò chơi là tổng doanh số bán hàng toàn cầu. Biểu đồ tròn dưới đây thể hiện tỷ lệ đóng góp của từng thể loại vào tổng doanh thu toàn ngành game, giúp chúng ta xác định những dòng game nào đang chiếm lĩnh thị trường và có sức ảnh hưởng lớn nhất đến hành vi tiêu dùng của người chơi trên toàn thế giới.



Hình 3-7: Thể loại phổ biến dựa trên doanh số bán hàng toàn cầu

Biểu đồ tròn cho thấy thể loại Action (hành động) là dòng game phổ biến nhất, chiếm 19.6% tổng doanh số bán hàng toàn cầu. Điều này phản ánh thị hiếu mạnh mẽ của người chơi đối với các game có tiết tấu nhanh, gameplay kịch tính và đa dạng.

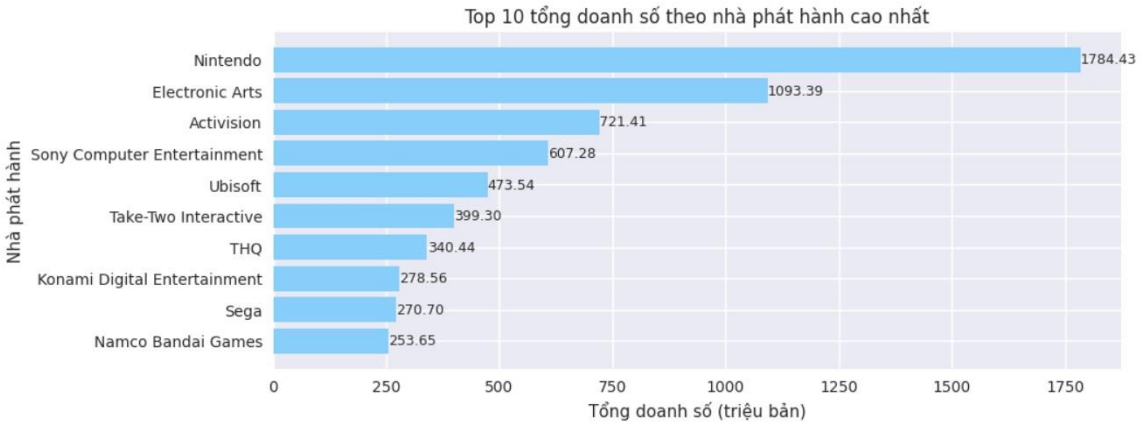
Xếp thứ hai là thể loại Sports (thể thao) với 14.9%, cho thấy sức hút bền vững của các trò chơi như *FIFA*, *NBA* hay *Wii Sports*. Các thể loại Shooter (11.6%), Role-Playing (10.5%) và Platform (9.41%) cũng đóng góp tỷ trọng đáng kể, cho thấy đây là những dòng game có lượng người chơi trung thành và phổ biến trên nhiều nền tảng. Ở chiều ngược lại, các thể loại như Strategy (1.97%), Adventure (2.66%) và Simulation (4.43%) có tỷ trọng doanh số thấp hơn, cho thấy mức độ phổ biến hạn chế hơn, có thể do tính đặc thù trong lối chơi hoặc nhóm đối tượng người chơi nhỏ hơn. Nhìn chung, biểu đồ phản ánh rõ sự thống trị của những dòng game dễ tiếp cận, có tính giải trí cao và phù hợp với đại đa số người chơi.

Mặc dù biểu đồ tròn cho thấy Action là thể loại chiếm tỷ trọng doanh số toàn cầu lớn nhất (19.6%), nhưng khi xét đến doanh số trung bình trên mỗi tựa game, thể loại này lại chỉ đứng ở vị trí thứ 8 với 0.53 triệu bản mỗi trò chơi. Điều này cho thấy rằng thể loại Action có rất nhiều tựa game phát hành, nhưng không phải tất cả đều đạt doanh số cao — thể hiện sự đa dạng nhưng cũng phân tán. Ngược lại, thể loại Platform lại nổi bật với doanh số trung bình cao nhất (0.95 triệu bản), nhưng chỉ chiếm 9.41% tổng doanh số toàn cầu, cho thấy tuy số lượng game Platform ít hơn, nhưng mỗi game lại có sức bán rất tốt. Điều này có thể giải thích bởi sự thành công vượt trội của một số tựa game biểu tượng như *Super Mario Bros.*, giúp kéo trung bình doanh số của thể loại này lên rất cao. Tương tự, Shooter và Role-Playing cũng thể hiện sự cân bằng tốt giữa hai tiêu chí: đều có tỷ trọng lớn trong tổng doanh số (lần lượt 11.6% và 10.5%) và doanh số trung bình mỗi game cũng ở mức khá cao (0.80 và 0.63 triệu bản). Tóm lại, hai biểu đồ giúp ta nhận diện được không chỉ những thể loại phổ biến về mặt tổng thể, mà còn hiểu rõ hiệu suất trung bình của từng tựa game trong mỗi dòng, từ đó đánh giá được cả chiều rộng lẫn chiều sâu của thị trường trò chơi điện tử.

Hình 3-8: Phân phối doanh số toàn cầu

Biểu đồ trên minh họa tần suất phân phối doanh số toàn cầu của các tựa game video, thể hiện rõ sự bất đối xứng nghiêm trọng trong thị trường này. Phần lớn các trò chơi đều có doanh số rất thấp: gần 15.751 trò chơi chỉ bán được dưới 5 triệu bản, cho thấy rằng phần lớn các sản phẩm phát hành không đạt được thành công thương mại đáng kể. Chỉ một số ít chưa đến vài chục trò chơi đạt được doanh số trên 20 triệu bản, và đặc biệt chỉ có duy nhất một trò chơi vượt mốc 80 triệu bản. Điều này phản ánh một thực tế quen thuộc trong ngành công nghiệp game: một số ít "bom tấn" thống trị thị trường, trong khi phần lớn các trò chơi còn lại chỉ tồn tại ở mức doanh thu rất khiêm tốn. Sự phân phối này tuân theo quy luật "đầu dài" (long tail), trong đó một số sản phẩm rất thành công chiếm phần lớn doanh thu, còn đa số chỉ đóng góp một lượng nhỏ. Đây là yếu tố mà các nhà phát triển và phát hành game cần cân nhắc kỹ trong chiến lược ra mắt sản phẩm.

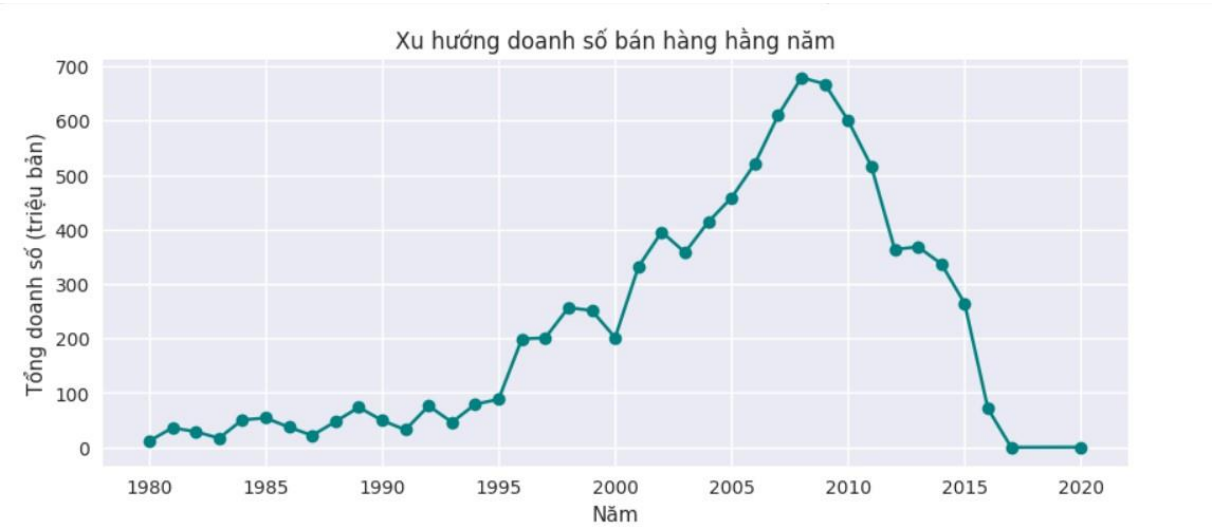
Trong ngành công nghiệp trò chơi điện tử, nhà phát hành đóng vai trò then chốt trong việc đưa sản phẩm đến tay người tiêu dùng trên toàn thế giới. Việc phân tích tổng doanh số theo nhà phát hành giúp chúng ta xác định được những "ông lớn" đang chiếm lĩnh thị trường, cũng như phần nào phản ánh chiến lược phát hành, danh mục sản phẩm và mức độ thành công của từng hãng. Đồng thời cho thấy sự thống trị rõ rệt của một số thương hiệu lâu đời và uy tín trong ngành.



Hình 3-9: Top 10 tổng doanh số theo nhà phát hành cao nhất

Biểu đồ cho thấy sự vượt trội rõ rệt của Nintendo trong ngành công nghiệp trò chơi điện tử với tổng doanh số lên tới 1784.43 triệu bản, bỏ xa các đối thủ còn lại. Đứng thứ hai là Electronic Arts với 1093.39 triệu bản, tiếp theo là Activision và Sony Computer Entertainment, lần lượt đạt 721.41 và 607.28 triệu bản. Nhóm giữa bảng gồm Ubisoft, Take-Two Interactive và THQ, đều có doanh số dao động trong khoảng 340–470 triệu bản. Ba vị trí cuối trong top 10 là Konami Digital Entertainment, Sega và Namco Bandai Games, tuy doanh số thấp hơn nhưng vẫn vượt mốc 250 triệu bản – một con số đáng nể. Nhìn chung, các nhà phát hành lớn đều là những thương hiệu quen thuộc, sở hữu nhiều dòng game nổi tiếng toàn cầu, cho thấy vai trò quyết định của danh mục sản phẩm mạnh mẽ và chiến lược phát hành hiệu quả trong việc chiếm lĩnh thị trường.

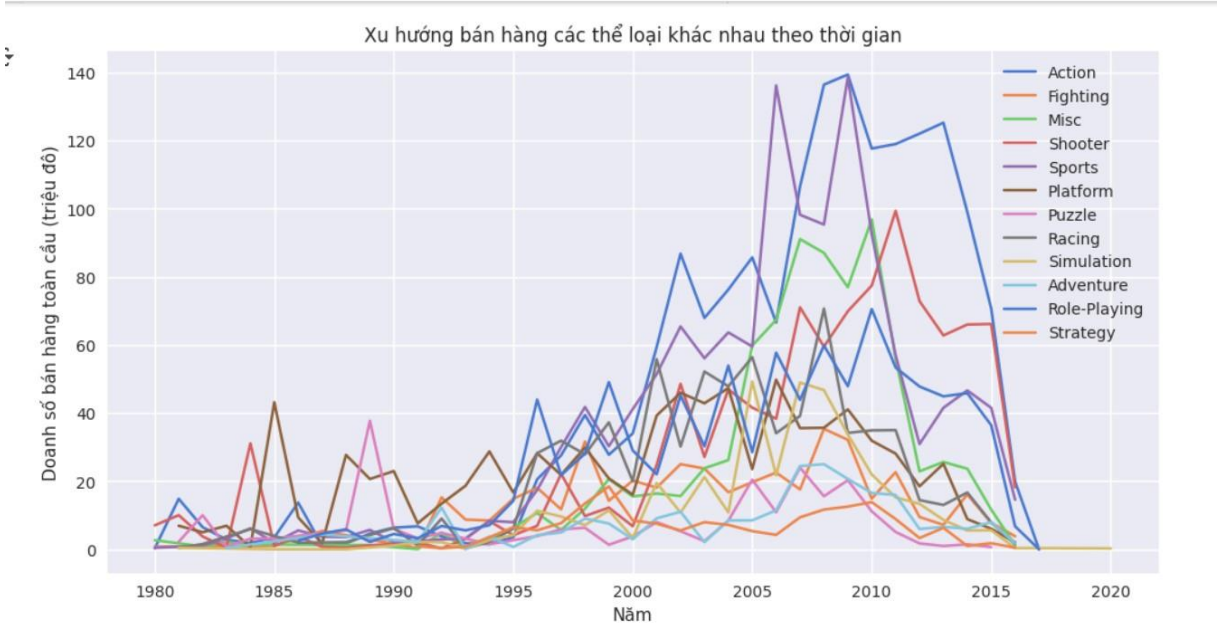
Sau khi tìm hiểu về tổng doanh số theo nền tảng, thể loại và mức độ phân phối, một khía cạnh quan trọng khác cần xem xét là xu hướng doanh số bán hàng theo từng năm. Phân tích này giúp ta nhìn nhận được biến động thị trường game qua thời gian, từ những giai đoạn tăng trưởng mạnh mẽ cho đến thời điểm suy giảm hoặc bão hòa. Qua đó, chúng ta có thể xác định được các giai đoạn vàng của ngành công nghiệp trò chơi điện tử, cũng như ảnh hưởng của các sự kiện kinh tế, công nghệ hoặc văn hóa tới hành vi tiêu dùng của người chơi.



Hình 3-10: Xu hướng doanh số bán hàng hằng năm

Biểu đồ thể hiện tổng doanh số bán hàng theo từng năm từ 1980 đến 2020 cho thấy một xu hướng phát triển mạnh mẽ và sau đó là sự suy giảm rõ rệt. Trong giai đoạn từ năm 1980 đến khoảng 1995, doanh số tăng trưởng chậm và khá biến động, phản ánh giai đoạn sơ khai của ngành công nghiệp game. Từ sau năm 1995, đặc biệt là từ năm 2000 trở đi, doanh số bắt đầu tăng trưởng nhanh chóng, đạt đỉnh vào khoảng năm 2008–2009 với mức hơn 680 triệu bản được bán ra, cho thấy đây là thời kỳ hoàng kim của ngành. Tuy nhiên, sau giai đoạn đỉnh cao này, doanh số bắt đầu giảm dần đều từ năm 2010 và đặc biệt tụt mạnh sau năm 2015, đến mức gần như bằng 0 trong giai đoạn 2017–2020. Do đây là thời điểm cuối chu kỳ của các console thế hệ cũ như PS4 và Xbox One, thị trường thiếu đột phá về công nghệ và ý tưởng. Các tựa game bom tấn phần lớn là sequel hoặc remake, trong khi game mới bị bội thực thể loại battle royale và lootbox gây nhàm chán cho game thủ. Đến cuối 2019, dịch Covid-19 tiếp tục khiến quá trình sản xuất và phát hành game đình trệ dù lượng người chơi tăng. Tất cả khiến thị trường game toàn cầu rơi vào giai đoạn quá độ, chờ đợi sự bùng nổ trở lại với thế hệ console mới vào cuối 2020.

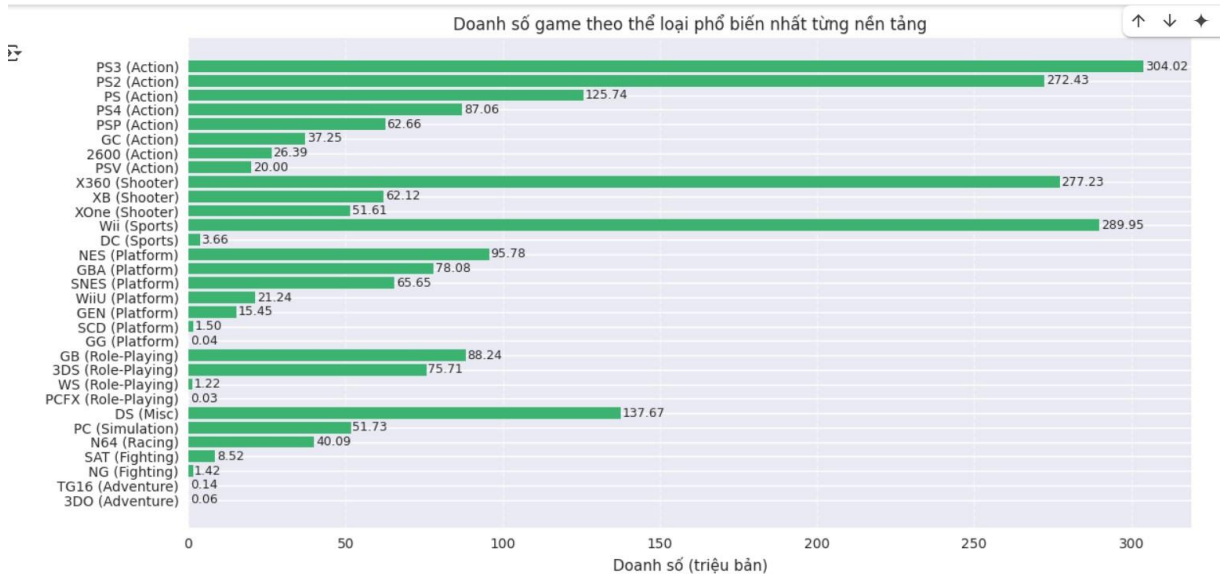
Bước tiếp theo, phân tích xu hướng bán hàng theo thời gian để thấy rõ sự thay đổi trong thị hiếu người chơi và chiến lược phát hành của các nhà phát triển. Mỗi giai đoạn đều phản ánh những biến chuyển nhất định của ngành công nghiệp game, từ sự trỗi dậy của các trò chơi Platform và Sports trong những năm đầu, đến làn sóng bùng nổ của thể loại Action và Shooter trong thời kỳ sau. Việc theo dõi xu hướng này giúp hiểu rõ hơn về sự tiến hóa của thị trường cũng như vai trò của công nghệ và thị hiếu người dùng trong việc định hình các thể loại chủ đạo qua từng năm.



Hình 3-11: Xu hướng bán hàng các thể loại khác nhau theo thời gian

Trong giai đoạn đầu, các thể loại như Platform, Sports và Racing chiếm ưu thế, đặc biệt là Platform với những tựa game kinh điển gắn liền với các hệ máy console đời đầu. Tuy nhiên, từ khoảng năm 2000 trở đi, xu hướng bắt đầu dịch chuyển mạnh mẽ sang các thể loại Action, Shooter, và Role-Playing, với sự bùng nổ trong doanh số bán hàng vào giai đoạn 2005–2012. Đáng chú ý, Shooter và Action đạt mức doanh số cao nhất, phản ánh sự ảnh hưởng lớn của các siêu phẩm đình đám trong thời kỳ này. Sau năm 2015, doanh số của hầu hết các thể loại đều có xu hướng giảm mạnh. Biểu đồ cũng cho thấy một số thể loại như Puzzle, Simulation hay Strategy duy trì doanh số tương đối ổn định nhưng không quá nổi bật so với các thể loại chủ lực khác.

Tiếp đến là đi sâu vào mối quan hệ giữa từng nền tảng cụ thể và thể loại game phổ biến nhất trên nền tảng đó. Phân tích này giúp chúng ta hiểu rõ hơn về chiến lược nội dung và thị hiếu người chơi trên mỗi hệ máy. Mỗi nền tảng thường có xu hướng nổi bật với một hoặc một vài thể loại game nhất định – phản ánh thế mạnh phần cứng, đặc trưng người dùng, cũng như sự định hướng của nhà sản xuất.



Hình 3-12: Doanh số trò chơi theo thể loại phổ biến nhất của từng nền tảng

Dễ dàng nhận thấy, nền tảng PS3 với thể loại Action dẫn đầu về doanh số với 304.02 triệu bản, theo sau là Wii với thể loại Sports đạt 289.95 triệu bản và PS2 (Action) ở vị trí thứ ba với 272.43 triệu bản. Đáng chú ý, thể loại Action đặc biệt thành công trên các nền tảng của Sony, trong khi thể loại Shooter lại chiếm ưu thế trên cả Xbox 360 và Xbox One. Wii cho thấy sự thành công của thể loại Sports, có thể là do các tựa game tương tác hướng đến gia đình. Ngược lại, các nền tảng cũ hơn hoặc ít phổ biến hơn ghi nhận doanh số thể loại phổ biến nhất thấp hơn đáng kể. Nhìn chung, mỗi hệ máy đều có tựa game chủ lực thành công hơn hẳn so với các game khác, phản ánh sự đa dạng trong thị hiếu của người chơi.

Tổng hợp các phân tích trên có thể thấy được mối liên hệ chặt chẽ giữa nền tảng, thể loại game và doanh số bán hàng toàn cầu. Về mặt nền tảng, các hệ máy đến từ Nintendo chiếm ưu thế áp đảo cả về tổng doanh số và sự phổ biến, điều này phần nào phản ánh sức mạnh thương hiệu, khả năng phát hành ổn định và định hướng phát triển phù hợp với thị hiếu thị trường. Ở góc độ thể loại, các dòng game như Platform, Shooter, Role-Playing và Action chiếm tỷ trọng lớn, cho thấy người chơi có xu hướng ưa chuộng những trải nghiệm hành động nhanh, nhập vai phong phú hoặc thử thách kỹ năng. Khi đi sâu vào từng nền tảng, mỗi hệ máy thường có xu hướng tập trung vào một số thể loại game nhất định, phù hợp với đặc điểm kỹ thuật và nhóm người dùng mục tiêu của nền tảng đó.

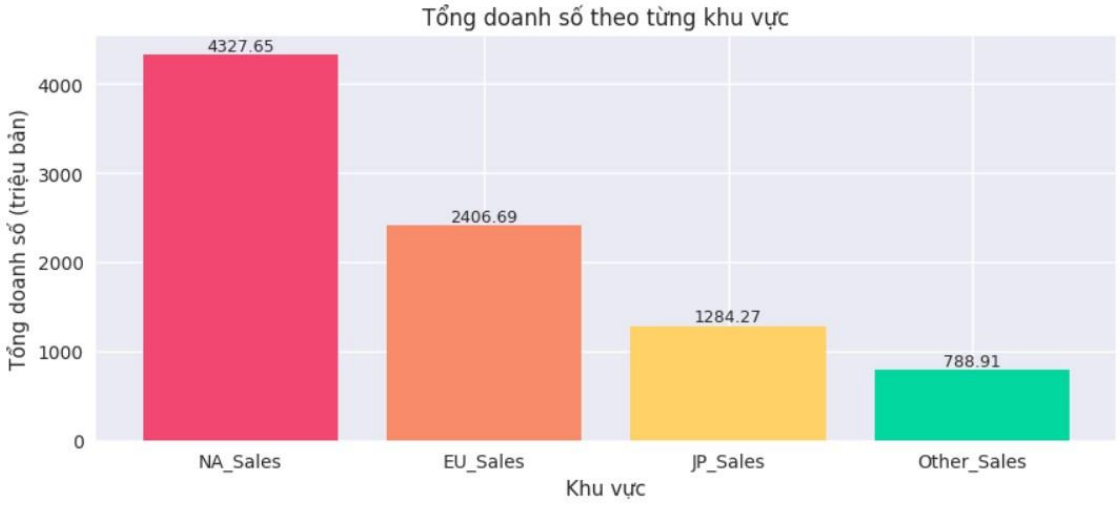
Bên cạnh đó, biểu đồ phân phối doanh số toàn cầu chỉ ra rằng phần lớn các tựa game đều có doanh số dưới 1 triệu bản, phản ánh thực trạng thị trường game có tính cạnh tranh cao, nơi chỉ một số ít tựa game nổi bật thực sự đạt được thành công vượt trội. Điều này nhấn mạnh vai trò quan trọng của việc lựa chọn nền tảng phù hợp và thể loại được ưa chuộng để tối ưu hóa doanh số. Nhìn chung, sự thống trị của Nintendo,

sức hút của các thể loại phổ biến và mức độ tập trung doanh số cho thấy thị trường game toàn cầu là một hệ sinh thái đa chiều, nơi thành công đến từ sự kết hợp hài hòa giữa công nghệ, nội dung và chiến lược phát hành.

3.1.3. Trực quan hóa dữ liệu doanh thu theo khu vực

Bên cạnh góc nhìn toàn cầu, việc phân tích chi tiết doanh thu theo từng khu vực địa lý như Bắc Mỹ (NA), châu Âu (EU), Nhật Bản (JP) và các khu vực khác (Other) là cần thiết để nắm bắt sự khác biệt trong thị hiếu người chơi ở từng vùng. Những phân tích này đóng vai trò quan trọng trong việc xác định thị trường tiềm năng, từ đó làm cơ sở cho các chiến lược phát hành sản phẩm hiệu quả và định hướng gợi ý theo từng khu vực trong các phần tiếp theo.

Mỗi khu vực trên thế giới bao gồm Bắc Mỹ, châu Âu, Nhật Bản và các thị trường còn lại đều có sở thích và hành vi tiêu dùng khác nhau, ảnh hưởng trực tiếp đến doanh số của từng dòng game và chiến lược phát hành của các nhà phát triển.

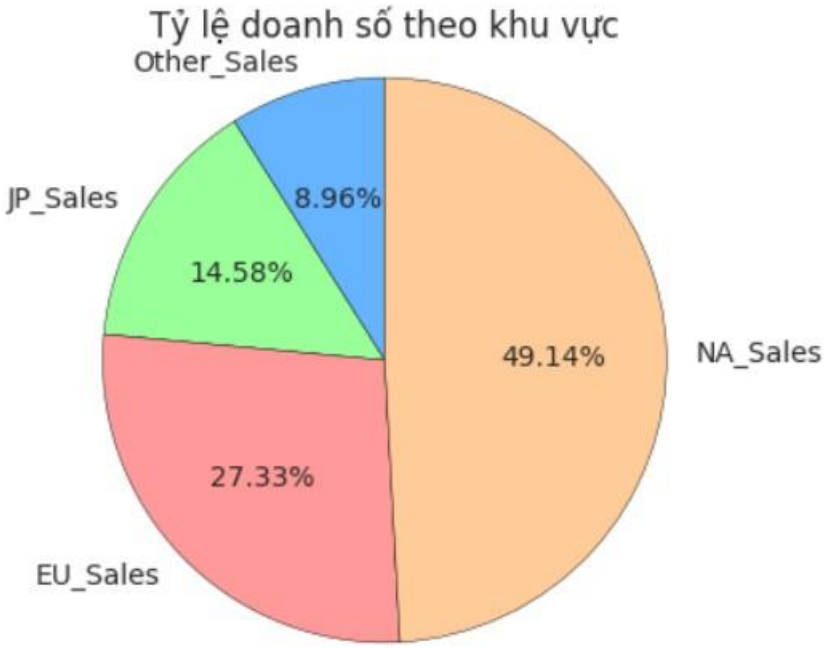


Hình 3-13: Tổng doanh số theo từng khu vực địa lý

Nhìn chung, biểu đồ cho thấy sự phân bố không đồng đều về doanh số giữa các khu vực, với sự tập trung chủ yếu ở Bắc Mỹ và Châu Âu. Bắc Mỹ (NA_Sales) chiếm ưu thế tuyệt đối với tổng doanh số đạt 4327.65 triệu bản, khẳng định vị thế là thị trường trò chơi điện tử lớn nhất. Theo sau là Châu Âu (EU_Sales) với 2406.69 triệu bản, cho thấy đây cũng là một thị trường có tầm quan trọng đáng kể. Nhật Bản (JP_Sales) ghi nhận tổng doanh số 1284.27 triệu bản, thấp hơn so với hai thị trường dẫn đầu nhưng vẫn là một thị trường lớn mạnh. Cuối cùng, nhóm "Các khu vực khác" (Other_Sales) đóng góp phần doanh số nhỏ nhất với 788.91 triệu bản.

Để có cái nhìn tổng quan hơn về sự phân bố doanh số trò chơi điện tử trên toàn cầu, việc xem xét tỷ lệ đóng góp của từng khu vực là vô cùng quan trọng. Biểu đồ sau đây sẽ làm nổi bật tỷ lệ phần trăm doanh số mà mỗi khu vực - bao gồm Bắc Mỹ, Châu

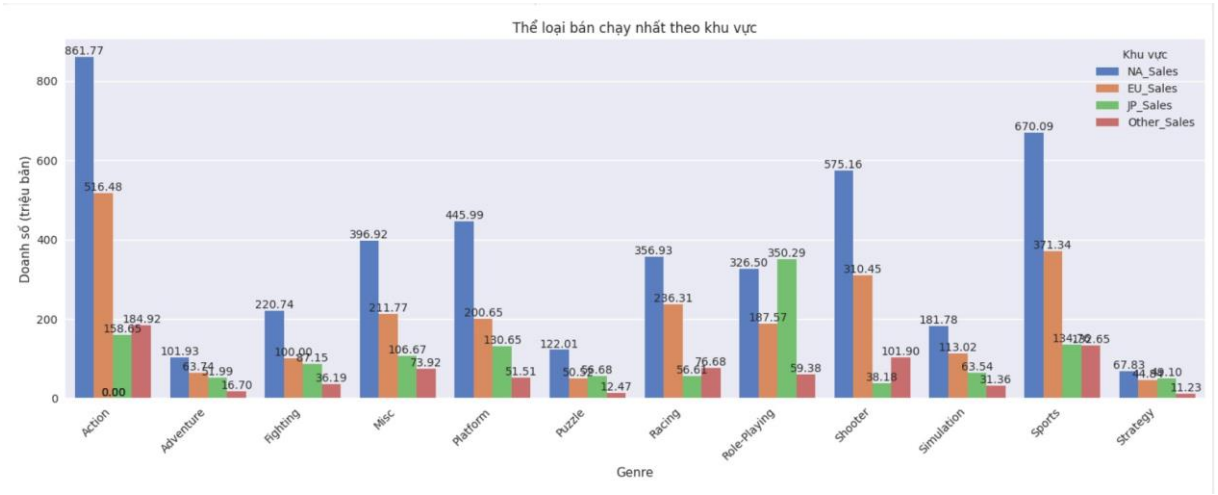
Âu, Nhật Bản và các khu vực khác - đóng góp vào tổng doanh số toàn cầu, từ đó giúp chúng ta hình dung rõ hơn về tầm quan trọng tương đối của từng thị trường.



Hình 3-14: Tỷ lệ doanh số theo khu vực

Biểu đồ trực quan hóa tỷ lệ phần trăm doanh số trò chơi điện tử theo từng khu vực trên toàn cầu. Khu vực Bắc Mỹ (NA_Sales) chiếm thị phần lớn nhất với 49.14% tổng doanh số, cho thấy gần một nửa doanh thu toàn cầu đến từ khu vực này. Châu Âu (EU_Sales) đứng thứ hai với 27.33%, đóng góp hơn một phần tư tổng doanh số. Thị phần của Nhật Bản (JP_Sales) là 14.58%, cho thấy đây vẫn là một thị trường quan trọng nhưng nhỏ hơn đáng kể so với Bắc Mỹ và Châu Âu. Cuối cùng, nhóm "Các khu vực khác" (Other_Sales) chỉ chiếm 8.96% tổng doanh số. Nhìn chung, sự tập trung đáng kể về doanh số nằm ở thị trường Bắc Mỹ và Châu Âu, trong khi các khu vực còn lại đóng góp một phần nhỏ hơn vào tổng doanh thu toàn cầu.

Một yếu tố quan trọng khác ảnh hưởng đến việc tìm ra hướng đi cho sự phát triển trong tương lai của ngành game là phân tích thể loại trò chơi bán chạy nhất theo từng khu vực. Điều này giúp nhận diện được sự khác biệt trong khẩu vị giải trí của người dùng toàn cầu nói chung cũng như các khu vực khác nói riêng Đồng thời cung cấp cơ sở cho các nhà phát hành trong việc lựa chọn chiến lược nội dung và phân phối phù hợp với từng thị trường mục tiêu.

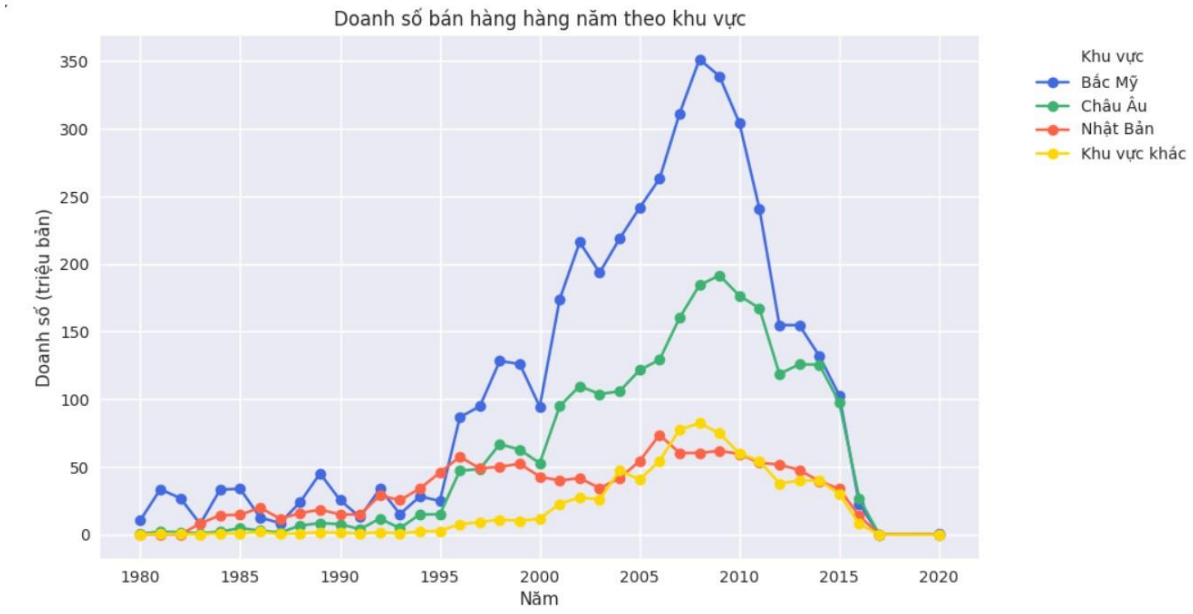


Hình 3-15: Biểu đồ thể hiện thể loại bán chạy nhất theo khu vực

Biểu đồ cung cấp một cái nhìn chi tiết về doanh số của từng thể loại game ở bốn khu vực chính: Bắc Mỹ, Châu Âu, Nhật Bản và Các khu vực khác. Nhìn chung, Bắc Mỹ thường là thị trường lớn nhất cho hầu hết các thể loại, đặc biệt là Action, Sports, Shooter, Platform và Misc. Châu Âu thường đứng thứ hai về doanh số ở nhiều thể loại. Tuy nhiên, thị trường Nhật Bản lại cho thấy sự khác biệt đáng kể, với thể loại Role-Playing có doanh số vượt trội so với các khu vực khác, đồng thời thể loại Fighting và Platform cũng có sự ưa chuộng đáng kể. Ngược lại, các thể loại như Shooter và Racing lại có doanh số tương đối thấp ở Nhật Bản. Sự thống trị về doanh số của thể loại Role-Playing (RPG) tại thị trường Nhật Bản là kết quả của một sự pha trộn độc đáo giữa lịch sử phát triển, các yếu tố văn hóa sâu sắc và đặc điểm thị trường riêng biệt. Nhật Bản là cái nôi của nhiều thương hiệu JRPG mang tính biểu tượng, những tựa game không chỉ định hình thể loại mà còn ăn sâu vào văn hóa đại chúng thông qua anime, manga và âm nhạc, tạo ra một lượng fan trung thành qua nhiều thế hệ. Văn hóa Nhật Bản, với sự coi trọng cốt truyện phức tạp, nhân vật sâu sắc và tính thẩm mỹ đặc trưng của anime/manga, đã tạo ra một môi trường lý tưởng cho sự phát triển của JRPG. Thêm vào đó, sự ủng hộ mạnh mẽ của các nhà phát triển Nhật Bản, cộng đồng game thủ đam mê và sự phù hợp của các nền tảng console phổ biến đã củng cố vị thế của RPG tại thị trường nội địa. Sự thành công ban đầu của các tựa game kinh điển vào những thập kỷ trước cũng tạo ra một hiệu ứng kéo dài, duy trì sức hút của thể loại này đối với các thế hệ người chơi sau này. Các khu vực khác thường đóng góp phần doanh số nhỏ nhất so với ba thị trường chính. Những khác biệt rõ rệt về doanh số giữa các thể loại và khu vực này nhấn mạnh tầm quan trọng của việc hiểu rõ sở thích của người chơi ở từng thị trường để các nhà phát hành có thể đưa ra các chiến lược phát hành và tiếp thị hiệu quả.

Một khía cạnh quan trọng khác giúp đánh giá quy mô và xu hướng phát triển của ngành công nghiệp trò chơi điện tử là sự phân bố doanh số theo khu vực địa lý qua các năm. Việc theo dõi doanh số bán hàng hàng năm tại các thị trường chính như Bắc

Mỹ, Châu Âu, Nhật Bản và phần còn lại của thế giới không chỉ phản ánh sức tiêu thụ của từng khu vực mà còn giúp xác định mức độ ảnh hưởng, thói quen tiêu dùng và tốc độ tăng trưởng của từng thị trường. Dữ liệu này là cơ sở quan trọng để các nhà phát hành quốc tế đưa ra chiến lược mở rộng, điều chỉnh nội dung hoặc tối ưu hóa phân phối cho từng khu vực cụ thể.

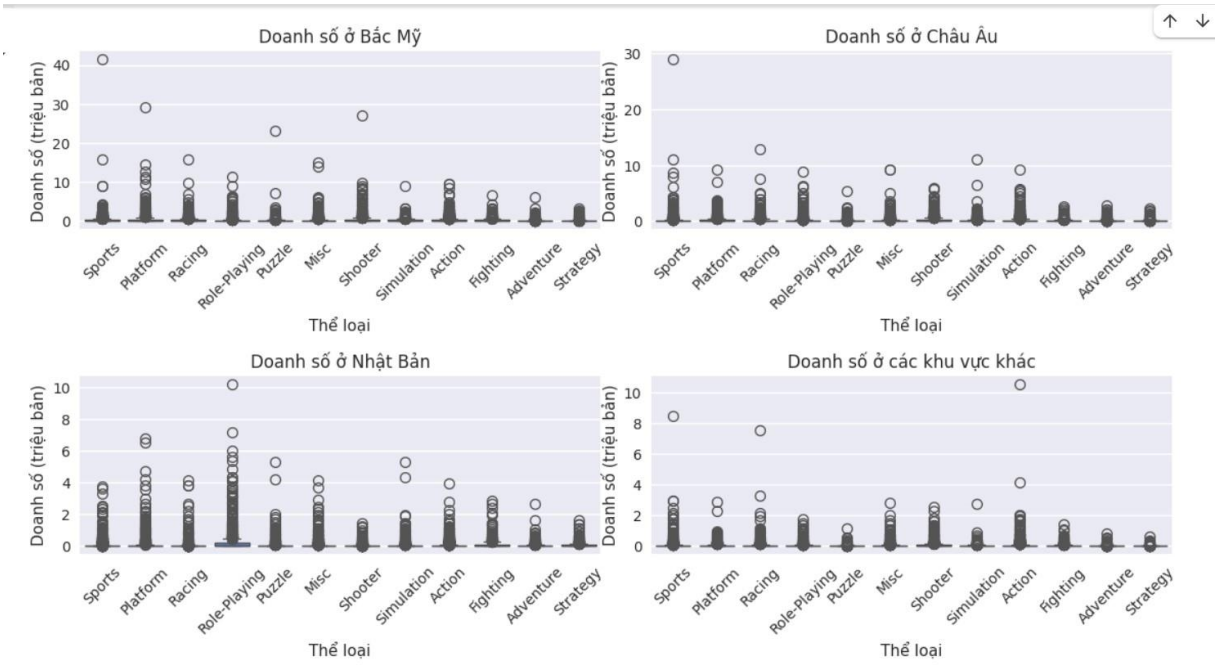


Hình 3-16: Biểu đồ doanh số bán hàng hàng năm theo khu vực

Biểu đồ đường thể hiện doanh số bán hàng trò chơi điện tử hàng năm theo khu vực từ năm 1980 đến năm 2020. Nhìn chung, doanh số bán hàng có xu hướng tăng trưởng mạnh mẽ từ giữa những năm 1990 đến khoảng năm 2008-2009, sau đó bắt đầu suy giảm đáng kể. Khu vực Bắc Mỹ có doanh số cao nhất với sự tăng trưởng vượt trội so với các khu vực khác trong giai đoạn đỉnh điểm. Châu Âu cũng có mức doanh số đáng kể, bám sát khá gần Bắc Mỹ trong giai đoạn tăng trưởng và cũng trải qua sự suy giảm tương tự. Nhật Bản có doanh số thấp hơn đáng kể so với 2 khu vực trên, với một đỉnh điểm nhỏ hơn vào khoảng giữa những năm 2000. Các khu vực khác có doanh số thấp nhất nhưng ít biến động hơn so với các khu vực chính. Đáng chú ý là sự sụt giảm doanh số ở tất cả các khu vực sau năm 2008-2009, chứng tỏ có thể có những thay đổi lớn trong thị trường trò chơi điện tử vào thời điểm đó.

Bên cạnh việc phân tích doanh số theo nền tảng, thể loại và trò chơi cụ thể, một khía cạnh quan trọng khác cần được xem xét là sự khác biệt về doanh số giữa các khu vực địa lý. Việc so sánh doanh số giữa các thị trường lớn như Bắc Mỹ, Châu Âu, Nhật Bản và các khu vực còn lại không chỉ phản ánh thị hiếu và thói quen tiêu dùng đặc thù của từng khu vực, mà còn giúp các nhà phát hành đưa ra chiến lược phân phối và nội địa hóa phù hợp. Phân tích này đóng vai trò quan trọng trong việc hiểu rõ cách mà trò

chơi điện tử được tiếp nhận và tiêu thụ trên phạm vi toàn cầu, từ đó làm rõ yếu tố địa lý trong sự thành công của một sản phẩm game.



Hình 3-17: So sánh doanh số giữa các thị trường

Biểu đồ bao gồm bốn biểu đồ hộp riêng biệt, mỗi biểu đồ tập trung vào một khu vực địa lý cụ thể: Bắc Mỹ, Châu Âu, Nhật Bản và Các khu vực khác. Khi so sánh doanh số của các thể loại game ở Bắc Mỹ, có thể thấy thể loại Sports có trung vị doanh số khá cao, đáng chú ý là sự xuất hiện của nhiều outliers ở mức doanh số rất cao, cho thấy sự thành công vượt trội của một số tựa game thể thao tại thị trường này. Các thể loại như Platform và Racing có phân phối doanh số tập trung hơn ở mức thấp đến trung bình, mặc dù vẫn có một vài tựa game đạt doanh số cao đột biến. Role-Playing cho thấy sự phân tán rộng hơn với trung vị ở mức trung bình và nhiều outliers doanh số cao, phản ánh sự phổ biến của các tựa game RPG lớn. Đáng chú ý, thể loại Shooter có trung vị doanh số cao và số lượng lớn các outliers ở mức doanh số rất cao, khẳng định vị thế là một trong những thể loại game thành công nhất tại Bắc Mỹ.

Thị trường Châu Âu có nhiều điểm tương đồng với Bắc Mỹ về xu hướng doanh số của các thể loại, tuy nhiên, mức doanh số trung bình và mức độ của các outliers thường có xu hướng thấp hơn. Thể loại Sports vẫn duy trì trung vị cao và các tựa game thành công, tương tự như Shooter và Action mặc dù số lượng và mức độ của các outliers có thể không bằng Bắc Mỹ. Role-Playing cũng có sự hiện diện của các tựa game thành công, nhưng ít hơn các trường hợp đạt doanh số cực kỳ cao so với thị trường Bắc Mỹ.

Thị trường Nhật Bản lại cho thấy một bức tranh tương đối khác biệt so với Bắc Mỹ và Châu Âu. Thể loại Role-Playing nổi bật với trung vị doanh số cao nhất và

nhiều giá trị ngoại lệ ở mức cao, cho thấy sự ưa chuộng đặc biệt của thể loại này tại Nhật Bản. Thể loại Platform cũng có phân phối doanh số đáng chú ý. Các thể loại phổ biến ở phương Tây như Sports, Shooter và Action vẫn có doanh số, nhưng trung vị và số lượng các tựa game đạt doanh số cao có vẻ thấp hơn. Các thể loại khác như Racing, Misc, Simulation, Fighting, Adventure và Strategy thường có doanh số tập trung ở mức thấp đến trung bình.

Cuối cùng, thị trường khu vực khác thường có mức doanh số thấp hơn so với ba thị trường chính và có thể phản ánh sự pha trộn của các xu hướng toàn cầu và khu vực. Các thể loại như Sports, Action và Shooter vẫn cho thấy doanh số đáng chú ý hơn so với các thể loại khác. Thể loại Role-Playing cũng có một vài trường hợp đạt doanh số cao, cho thấy sự thành công của một số tựa game RPG nhất định ở các khu vực này.

Tổng thể các biểu đồ phân tích doanh số trò chơi điện tử theo khu vực địa lý cho thấy bức tranh toàn diện về sự phân hóa rõ rệt của thị trường toàn cầu. Về tổng doanh số tuyệt đối, Bắc Mỹ và Châu Âu chiếm ưu thế vượt trội so với các khu vực khác, khẳng định vai trò là hai thị trường tiêu thụ game lớn nhất thế giới. Khi phân tích doanh số bán hàng hàng năm, có thể thấy các khu vực chính đều đạt đỉnh vào khoảng năm 2008–2009, sau đó đồng loạt suy giảm, cho thấy sự thay đổi đáng kể trong mô hình tiêu dùng, có thể đến từ sự chuyển dịch sang các nền tảng số, sự bão hòa của thị trường console truyền thống, hoặc sự nổi lên của thị trường game di động và nền tảng phân phối mới. Ngoài yếu tố số lượng, khẩu vị giải trí theo từng khu vực cũng cho thấy sự khác biệt rõ rệt. Bắc Mỹ và Châu Âu có xu hướng yêu thích các thể loại Sports, Shooter và Action, trong khi thị trường Nhật Bản lại thể hiện sự ưa chuộng đặc biệt đối với thể loại Role-Playing, phản ánh ảnh hưởng sâu sắc của văn hóa và truyền thống nội địa.

Có thể kết luận rằng: vị trí địa lý không chỉ ảnh hưởng đến tổng doanh số mà còn quyết định đến thị hiếu và sự thành công của từng thể loại trò chơi. Việc hiểu rõ đặc điểm từng thị trường sẽ giúp các nhà phát hành xây dựng chiến lược phân phối, nội dung và marketing phù hợp, từ đó tối ưu hóa doanh số và mức độ tiếp cận người dùng trên toàn cầu.

Thị trường Bắc Mỹ có doanh số trò chơi điện tử lớn nhất, vượt trội so với các khu vực khác là do sự kết hợp của nhiều yếu tố kinh tế, văn hóa, công nghệ và lịch sử phát triển của ngành game, cụ thể:

Thứ nhất, quy mô thị trường và sức mua cao. Bắc Mỹ là một trong những khu vực có nền kinh tế phát triển nhất thế giới, đặc biệt là Hoa Kỳ và Canada với GDP bình quân đầu người cao và mức sống ổn định. Khả năng chi tiêu cho sản phẩm giải trí, trong đó có trò chơi điện tử, là rất lớn. Ngoài ra, tỷ lệ hộ gia đình sở hữu thiết bị chơi game như console, máy tính cá nhân hoặc điện thoại thông minh cũng cao hơn so với nhiều khu vực khác, tạo điều kiện thuận lợi cho việc tiêu thụ trò chơi điện tử. Đây

là một trong những yếu tố quan trọng giúp Bắc Mỹ duy trì vị thế là thị trường dẫn đầu về doanh số trò chơi điện tử toàn cầu.

Thứ hai, hạ tầng công nghệ và mạng lưới phân phối mạnh. Một yếu tố khác góp phần vào doanh số vượt trội của thị trường Bắc Mỹ là cơ sở hạ tầng công nghệ hiện đại. Mạng internet tốc độ cao, độ phủ rộng, cùng sự phát triển mạnh của các nền tảng phân phối kỹ thuật số như Xbox Live, PlayStation Network, Steam và Epic Games Store đã giúp game đến tay người tiêu dùng nhanh chóng và thuận tiện. Đặc biệt, nhiều trong số các nền tảng này có trụ sở hoặc thị phần lớn tại Bắc Mỹ, từ đó tạo ra lợi thế cạnh tranh vượt trội so với các khu vực khác trong việc cung cấp và tiêu thụ trò chơi điện tử.

Thứ ba, thị hiếu phù hợp với các tựa game bom tấn. Thị trường Bắc Mỹ có đặc điểm thị hiếu rất rõ ràng và phù hợp với các thể loại trò chơi phổ biến toàn cầu như hành động (Action), thể thao (Sports) và bắn súng (Shooter). Những thể loại này thường nằm trong nhóm game có doanh số cao nhất, và phần lớn được phát triển nhắm đến thị hiếu người chơi Bắc Mỹ. Những tựa game như *Call of Duty*, *Grand Theft Auto*, *Madden NFL*, hay *NBA 2K* luôn dẫn đầu về doanh thu nhiều năm liền. Điều này cho thấy sự ăn khớp giữa sở thích người dùng và xu hướng sản phẩm, qua đó giúp thúc đẩy doanh số đáng kể tại khu vực này.

Thứ tư, thị trường game phát triển từ rất sớm. Bắc Mỹ là một trong những cái nôi đầu tiên của ngành công nghiệp trò chơi điện tử hiện đại. Ngay từ những năm 1970–1980, các công ty như Atari và sau này là Nintendo of America, đã đóng vai trò tiên phong trong việc thương mại hóa và phổ biến trò chơi điện tử. Truyền thống lâu đời này đã hình thành một thị trường ổn định với lượng khách hàng trung thành qua nhiều thế hệ. Game không chỉ là một sản phẩm giải trí ngắn hạn, mà đã trở thành một phần của văn hóa đại chúng Bắc Mỹ, từ đó tạo ra một nhu cầu bền vững và có xu hướng tăng trưởng liên tục.

Cuối cùng là chiến lược marketing quy mô lớn và định hướng toàn cầu. Một trong những lợi thế đặc biệt của các nhà phát hành game tại Bắc Mỹ là khả năng triển khai các chiến dịch marketing quy mô lớn, chuyên nghiệp và có định hướng toàn cầu. Việc đầu tư mạnh vào quảng bá, tổ chức sự kiện, tài trợ giải đấu và hợp tác với các nền tảng truyền thông giúp các sản phẩm game dễ dàng tiếp cận người tiêu dùng trong và ngoài khu vực. Không chỉ chiếm lĩnh thị trường nội địa, nhiều tựa game của các hãng Bắc Mỹ còn định hình xu hướng tiêu dùng toàn cầu, từ đó càng củng cố vị thế dẫn đầu về doanh số trong ngành công nghiệp trò chơi điện tử.

3.2. HUẤN LUYỆN MÔ HÌNH

Sau khi hoàn thành quá trình phân tích và trực quan hóa dữ liệu doanh thu, bước tiếp theo trong quá trình nghiên cứu là xây dựng mô hình có khả năng đưa ra gợi ý phù hợp cho người chơi hoặc nhà phát hành. Việc huấn luyện mô hình sẽ giúp hệ thống có khả năng học từ dữ liệu sẵn có, nhận diện các đặc điểm quan trọng của từng

trò chơi và từ đó đưa ra các đề xuất mang tính tương quan hoặc dự đoán sở thích. Trong phạm vi đề tài này, mô hình được xây dựng dựa trên thông tin nội dung của trò chơi như thể loại, nền tảng, và nhà phát hành, sử dụng kỹ thuật vector hóa văn bản và đo lường độ tương đồng giữa các trò chơi.

3.2.1. Dự đoán doanh thu với Random Forest Regressor

Dự đoán doanh thu bằng Random Forest Regressor là một phương pháp trong học máy nhằm ước lượng giá trị doanh thu dựa trên các đặc trưng đầu vào như loại sản phẩm, thời gian phát hành, khu vực bán hàng hay chiến dịch quảng cáo. Random Forest Regressor là một mô hình hồi quy được xây dựng từ tập hợp nhiều cây quyết định, mỗi cây học trên một tập con ngẫu nhiên của dữ liệu và đặc trưng. Kết quả dự đoán cuối cùng được lấy trung bình từ tất cả các cây, giúp mô hình đạt độ chính xác cao và giảm thiểu hiện tượng quá khớp. Phương pháp này đặc biệt hiệu quả trong việc mô hình hóa các mối quan hệ phức tạp và phi tuyến giữa các biến, đồng thời cung cấp thông tin quan trọng về mức độ ảnh hưởng của từng đặc trưng đến doanh thu. Nhờ khả năng xử lý dữ liệu linh hoạt, Random Forest Regressor là một lựa chọn phổ biến trong các bài toán dự báo doanh thu thực tế.

Mục tiêu của mô hình hồi quy sử dụng Random Forest Regression là dự đoán doanh thu toàn cầu (Global_Sales) của trò chơi điện tử dựa trên doanh thu ở từng khu vực khác nhau dựa trên các đặc trưng Genre, Platform, Publisher, Year. Hướng đến mục đích hỗ trợ nhà phát hành trò chơi điện tử trong việc đưa ra các quyết định chiến lược dựa trên dữ liệu. Mô hình giúp nhà phát hành đánh giá tiềm năng thương mại trước khi phát hành chính thức. Ngoài ra, bằng cách phân tích các yếu tố ảnh hưởng đến doanh thu như nền tảng, thể loại hoặc nhà phát hành, hệ thống có thể cung cấp những hiểu biết sâu sắc để nhà phát hành xác định đâu là yếu tố then chốt góp phần tạo nên thành công. Đồng thời, mô hình hỗ trợ tối ưu hóa thời điểm và khu vực phát hành, từ đó giúp tăng khả năng tiếp cận thị trường mục tiêu và nâng cao hiệu quả kinh doanh.

Quá trình huấn luyện mô hình diễn ra như sau:

- Dữ liệu đầu vào (features) gồm doanh thu từ các khu vực: Bắc Mỹ, châu Âu, Nhật Bản, khu vực khác và năm phát hành. Đầu ra (target) là tổng doanh thu toàn cầu (Global_Sales) – đây là biến cần dự đoán.
- Chia dữ liệu: Dữ liệu được chia thành 80% để huấn luyện và 20% để kiểm tra, đảm bảo kết quả có thể đánh giá được độ chính xác.
- Huấn luyện mô hình: Tạo một mô hình rừng ngẫu nhiên gồm 100 cây (có thể hiểu là 100 mô hình con).Mô hình được huấn luyện trên tập dữ liệu X_train và y_train.
- Dự đoán kết quả đánh giá:
  y_pred: kết quả dự đoán doanh thu toàn cầu. mse: sai số bình
  phương trung bình (giá trị càng nhỏ càng tốt).
  r2_score: hệ số xác định R2R^2R2 – đo mức độ phù hợp của mô hình (giá trị gần 1 là rất tốt).

- Kết quả nhận được: Mô hình có sai số thấp và độ chính xác khá cao (hơn 82%), cho thấy khả năng dự đoán tốt tổng doanh thu toàn cầu dựa trên doanh thu ở từng khu vực.

Ví dụ dự đoán doanh thu trò chơi mới: Nhà phát hành nhập thông tin: Một trò chơi Action trên PS3 do Take-Two Interactive phát hành năm 2025. Có mô hình dự đoán Doanh thu khoảng 5.23 triệu bản. Nếu doanh thu dự đoán cao, đầu tư phát triển; nếu thấp, cân nhắc thay đổi nền tảng hoặc thể loại.

Ví dụ phân tích yếu tố ảnh hưởng: Tầm quan trọng đặc trưng cho thấy Platform Wii chiếm 25%, Genre_Action chiếm 20%. Nhà phát hành cần đưa ra chiến lược phát hành trên Wii hoặc tập trung vào thể loại Action để tối ưu doanh thu.

Ví dụ tối ưu hóa chiến lược phát hành: Nếu Publisher Nintendo có tầm quan trọng cao, nên hợp tác với Nintendo để tăng cơ hội thành công. Nếu Year ảnh hưởng mạnh, phân tích xu hướng theo thời gian để chọn năm phát hành (kết hợp với Prophet).

Mô hình Random Forest Regressor hoạt động hiệu quả trong việc dự đoán doanh thu toàn cầu (Global_Sales) của các trò chơi điện tử dựa trên các yếu tố đầu vào như doanh thu theo khu vực (NA_Sales, EU_Sales, JP_Sales, Other_Sales) và năm phát hành. Bằng cách kết hợp nhiều cây quyết định (decision trees) thành một "rừng" (forest), mô hình khai thác sức mạnh của học máy theo tổ hợp (ensemble learning) để giảm thiểu hiện tượng quá khớp (overfitting) và tăng độ chính xác trong dự đoán. Sau khi được huấn luyện trên tập dữ liệu huấn luyện, mô hình cho ra các kết quả đánh giá như Mean Squared Error (MSE) và R-squared (R^2). Trong kết quả cụ thể, mô hình đạt được $R^2 = 0.82$, cho thấy rằng khoảng 82% phương sai trong doanh thu toàn cầu có thể được giải thích bởi các biến đầu vào – đây là một con số khá cao, chứng tỏ mô hình phù hợp và đáng tin cậy. Tóm lại, Random Forest Regressor đã chứng minh khả năng học tốt từ dữ liệu, dự đoán chính xác doanh thu trò chơi và là một công cụ hữu ích giúp các nhà phát hành ước lượng tiềm năng thị trường của trò chơi mới, từ đó hỗ trợ đưa ra các quyết định chiến lược.

- Kết quả đánh giá hiệu suất mô hình cho thấy sai số trung bình bình phương gốc (RMSE) đạt mức 2.00 triệu bản, đồng nghĩa với việc mô hình dự đoán doanh số game trung bình lệch khoảng 2 triệu bản so với thực tế. Đây là một sai số khá lớn, đặc biệt trong bối cảnh phần lớn các tựa game có doanh số dao động từ 0 đến 5 triệu bản. Bên cạnh đó, hệ số xác định R^2 chỉ đạt 0.07 (7%), cho thấy mô hình chỉ giải thích được 7% phương sai trong dữ liệu, trong khi 93% còn lại đến từ các yếu tố khác chưa được nắm bắt. Điều này phản ánh rằng mô hình hiện tại chưa đủ mạnh để mô phỏng mối quan hệ giữa các đặc trưng đầu vào và doanh số bán hàng một cách hiệu quả.

Đặc trưng	Giải thích	Tầm quan trọng
cat__Genre_Role-Playing	Game nhập vai – thường là dòng game dài, được yêu thích	0.0514
cat__Genre_Shooter	Bắn súng (FPS/TPS) – phổ biến với game thủ	0.0404
cat__Genre_Platform	Game đi cảnh/leo trèo (ví dụ: Mario)	0.0371
cat__Platform_GB	Máy GameBoy	0.0327
cat__Genre_Misc	Thể loại khác, không rõ ràng	0.0257
cat__Genre_Racing	Game đua xe	0.0253
cat__Genre_Action	Hành động	0.0210
cat__Platform_DS	Máy Nintendo DS	0.0206

- Về mặt phân tích đặc trưng, num__Year (năm phát hành) được xác định là yếu tố ảnh hưởng lớn nhất đến doanh số, với giá trị tầm quan trọng lên đến 0.4087 – chiếm hơn 40% ảnh hưởng toàn mô hình. Đây là kết quả hợp lý vì thị trường game biến động mạnh theo thời gian, chịu ảnh hưởng từ công nghệ, xu hướng tiêu dùng và chu kỳ phát triển nền tảng. Đặc trưng cat__Publisher_Nintendo cũng có ảnh hưởng lớn (0.13196), phản ánh vị thế vượt trội của Nintendo với nhiều tựa game nổi bật, độc quyền và có doanh số cao. Các đặc trưng khác như thể loại game (Role-Playing, Shooter, Platform) và nền tảng (Platform_GB, Platform_DS) chỉ có mức ảnh hưởng dao động trong khoảng 0.02 đến 0.05, cho thấy đóng góp tương đối thấp.

- Từ các phân tích trên, có thể thấy mô hình đang phụ thuộc quá nhiều vào một vài đặc trưng cụ thể và chưa tận dụng tốt tiềm năng của toàn bộ tập dữ liệu. Để cải thiện hiệu quả dự đoán, cần xem xét bổ sung thêm các đặc trưng giàu thông tin hơn như doanh số theo từng khu vực, chỉ số độc quyền, phân loại theo dòng game nổi tiếng, hoặc chia lại thời gian theo thập kỷ. Đồng thời, việc thử nghiệm các mô hình mạnh hơn như XGBoost, LightGBM kết hợp với kỹ thuật điều chỉnh tham số (GridSearchCV) cũng là hướng đi cần thiết nhằm nâng cao độ chính xác và khả năng tổng quát hóa của mô hình.
- Dựa trên mô hình đã huấn luyện, doanh thu toàn cầu dự đoán cho tựa game mới đạt khoảng **9.85 triệu bản**. Đây là một con số cao so với mặt bằng chung trong tập dữ liệu, nơi phần lớn các trò chơi có doanh số dưới 5 triệu bản. Dự đoán này cho thấy game mới được mô hình đánh giá là có tiềm năng thương mại vượt trội. Tuy nhiên, cần lưu ý rằng hiệu suất tổng thể của mô hình (RMSE = 2 triệu bản, $R^2 = 0.07$) còn tương đối thấp, do đó dự báo trên chỉ nên được xem như một **ước lượng sơ bộ**, mang tính tham khảo. Để tăng độ tin cậy, cần đánh giá thêm bằng nhiều mô hình khác, kiểm tra sự nhạy cảm của đầu vào, và đối chiếu với các yếu tố bên ngoài như xu hướng thị trường, chiến lược phát hành và mức độ nổi tiếng của thương hiệu trò chơi.

3.2.2. Phân cụm (KMeans Clustering).

KMeans Clustering là một thuật toán phân cụm thuộc nhóm học máy không giám sát (unsupervised learning), được sử dụng phổ biến trong việc khám phá cấu trúc

tiềm ẩn trong dữ liệu khi không có nhãn mục tiêu. Trong ngữ cảnh phân tích dữ liệu doanh số trò chơi điện tử, thuật toán này được áp dụng để phân nhóm các trò chơi thành các cụm (clusters) dựa trên mức độ tương đồng về doanh số tại các khu vực khác nhau như Bắc Mỹ (NA_Sales), Châu Âu (EU_Sales), Nhật Bản (JP_Sales) và các khu vực khác (Other_Sales).

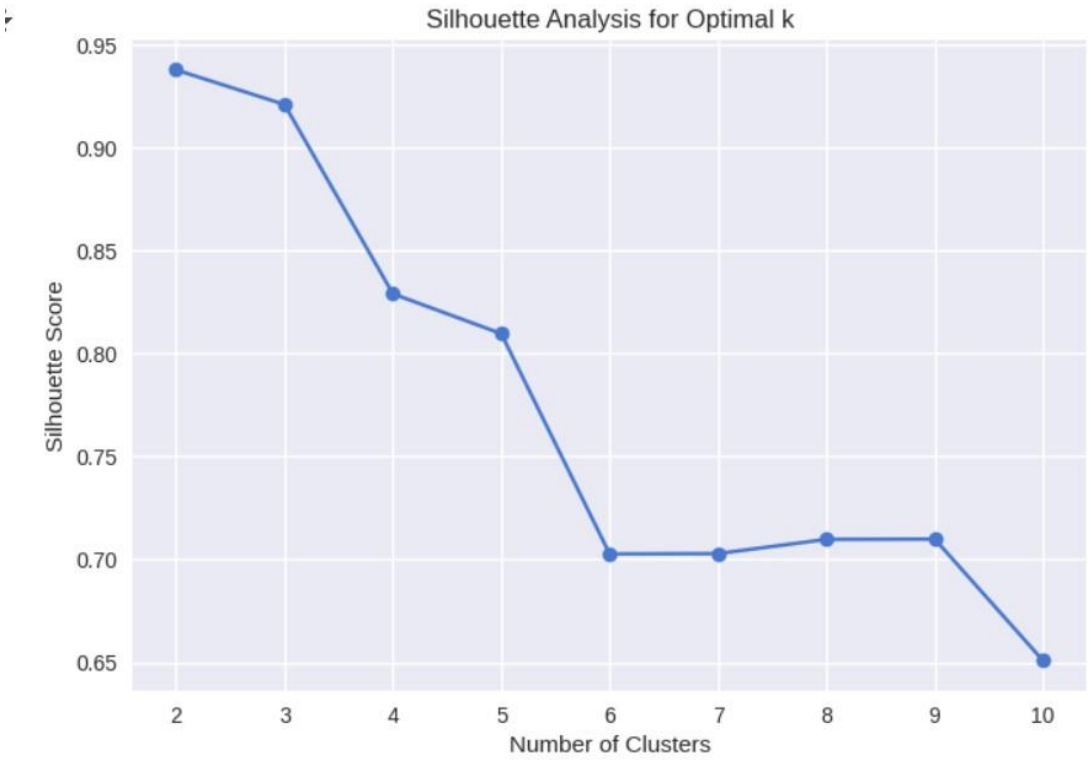
Mục tiêu chính của KMeans trong bài toán này là xác định các nhóm trò chơi có đặc điểm doanh thu tương tự trên các khu vực địa lý nhằm hỗ trợ quá trình phân đoạn thị trường một cách hiệu quả. Từ đó khám phá các phân khúc thị trường có hành vi tiêu dùng đặc trưng. Thông qua việc phân cụm, nhà phát hành có thể xác định các nhóm trò chơi có đặc điểm doanh thu tương đồng để xây dựng chiến lược phù hợp cho từng nhóm cụ thể. Điều này giúp tối ưu hóa chiến lược phát hành tại từng khu vực, đảm bảo rằng sản phẩm phù hợp với thị hiếu và xu hướng tiêu dùng địa phương. Ngoài ra, việc hiểu rõ xu hướng doanh thu theo từng thị trường còn đóng vai trò quan trọng trong việc định hướng phát triển trò chơi mới, giúp nhà phát hành đưa ra quyết định sáng suốt hơn trong khâu thiết kế, tiếp thị và phân phối sản phẩm.

Quá trình huấn luyện mô hình diễn ra như sau:

- Dữ liệu tập trung vào các cột thể hiện doanh thu khu vực (NA_Sales, EU_Sales, JP_Sales, Other_Sales), tạo thành tập đặc trưng (X_cluster) cho bài toán phân cụm.

- Sử dụng StandardScaler() để chuẩn hóa dữ liệu đầu vào, đưa toàn bộ dữ liệu về cùng một thang đo nhằm đảm bảo các khu vực có doanh thu lớn (ví dụ Bắc Mỹ) không chi phối kết quả phân cụm. ● Tìm số lượng cụm tối ưu:
  Duyệt qua các giá trị số cụm từ 2 đến 10.
  Với mỗi giá trị n_clusters, thuật toán KMeans được huấn luyện và dự đoán nhãn cụm (fit_predict).
  Sau đó, tính chỉ số Silhouette Score – một thước đo đánh giá mức độ phân tách giữa các cụm. Điểm số càng cao, phân cụm càng rõ ràng và hợp lý.

  Vẽ biểu đồ đường biểu diễn Silhouette Score để xác định số lượng cụm tối ưu (ví dụ: tại đỉnh của đường cong).



Clusters of Video Games

Hình 3-18: Đường biểu diễn Silhouette

- Huấn luyện mô hình với số cụm tối ưu:

  Chọn giá trị k tối ưu (ví dụ k = 3) từ biểu đồ.
  Áp dụng KMeans với số cụm đó để gán mỗi trò chơi vào một cụm (df['Cluster']), phản ánh nhóm thị trường mà trò chơi thuộc về.

- Trực quan hóa kết quả phân cụm: Sử dụng biểu đồ 3D (plotly.express.scatter_3d) để trực quan hóa các cụm theo ba trục doanh thu khu vực (NA, EU, JP), với màu sắc thể hiện các cụm khác nhau.

Mô hình KMeans trong hệ thống đã thực hiện thành công nhiệm vụ phân cụm các trò chơi điện tử dựa trên đặc điểm doanh thu ở từng khu vực như Bắc Mỹ, Châu Âu, Nhật Bản và các khu vực khác. Thông qua việc chuẩn hóa dữ liệu và lựa chọn số lượng cụm tối ưu bằng phương pháp Silhouette Score, mô hình đã chia các trò chơi thành những nhóm có đặc điểm doanh thu tương đồng. Kết quả phân cụm không chỉ giúp nhận diện các phân khúc thị trường tiềm năng mà còn hỗ trợ các nhà phát hành trong việc xác định chiến lược phân phối hiệu quả hơn. Việc phân loại rõ ràng này giúp họ nhận ra nhóm trò chơi nào phù hợp với từng khu vực địa lý, từ đó có thể tối ưu thời điểm phát hành, kênh tiếp cận cũng như chiến lược quảng bá. Tóm lại, mô hình KMeans là một công cụ hữu ích trong việc hiểu hành vi tiêu dùng theo khu vực

và định hướng phát triển sản phẩm trong ngành công nghiệp game một cách chính xác và khoa học hơn.

Phân tích phân cụm trên doanh thu theo khu vực (Bắc Mỹ, Châu Âu, Nhật Bản và các khu vực khác) đã chia dữ liệu thành ba nhóm rõ rệt, với điểm Silhouette Score đạt 0.84 – cho thấy độ tách biệt giữa các cụm là rất tốt và các cụm được hình thành có tính chất khác biệt rõ ràng.

Bảng thống kê trung bình cho thấy:

- Cụm 0 gồm các trò chơi có doanh thu rất thấp ở tất cả các khu vực, trung bình chỉ 0.17 triệu bản tại Bắc Mỹ và 0.087 triệu bản tại Châu Âu. Đây có thể là các tựa game ít nổi tiếng hoặc được phát hành giới hạn theo khu vực.
- Cụm 1 đại diện cho nhóm game có doanh thu trung bình, với mức tiêu thụ lần lượt là 2.09 triệu bản tại Bắc Mỹ và 1.46 triệu bản tại Châu Âu. Các trò chơi này có mức độ phổ biến vừa phải và phát hành rộng rãi hơn.
- Cụm 2 gồm các tựa game có doanh số vượt trội tại mọi khu vực, đặc biệt là Bắc Mỹ (11.2 triệu bản) và Châu Âu (6.46 triệu bản). Đây có thể là các thương hiệu lớn hoặc sản phẩm mang tính biểu tượng toàn cầu như Mario, Pokémon, hay GTA.

Biểu đồ tán xạ thể hiện rõ ràng sự phân chia này khi ta so sánh doanh thu tại Bắc Mỹ và Châu Âu. Cụm 0 (màu xanh dương nhạt) tập trung sát gốc tọa độ, cụm 1 (cam) phân bố rộng hơn, trong khi cụm 2 (xanh lá) nổi bật với vị trí rải rác tại các vùng có doanh thu rất cao, xác nhận mức độ phân tách tốt giữa các nhóm.

Phân tích phân cụm như trên giúp ta hiểu được cấu trúc phân phối doanh thu game theo khu vực, từ đó hỗ trợ việc xây dựng chiến lược phát hành, tiếp thị và phân khúc sản phẩm phù hợp theo thị trường mục tiêu.

3.2.3. Prophet Time Series Analysis.

Phân tích chuỗi thời gian (Time Series Analysis) là một phương pháp trong thống kê và học máy nhằm phân tích dữ liệu được thu thập theo trình tự thời gian để nhận diện các xu hướng (trend), chu kỳ mùa vụ (seasonality), và mẫu bất thường, từ đó đưa ra dự báo trong tương lai. Trong bối cảnh dữ liệu kinh doanh, kỹ thuật này đóng vai trò quan trọng trong việc hỗ trợ ra quyết định dựa trên hành vi biến động của thị trường theo thời gian.

Prophet là một thư viện mã nguồn mở do Facebook phát triển, được thiết kế chuyên biệt cho việc dự báo chuỗi thời gian. Prophet mô hình hóa chuỗi thời gian như một tổng hợp của xu hướng dài hạn, yếu tố mùa vụ và các hiệu ứng ngoại lai (như ngày lễ hoặc sự kiện bất thường). Ưu điểm của Prophet bao gồm khả năng xử lý dữ liệu thiếu, nhiễu, điểm ngoại lai, tự động phát hiện điểm thay đổi xu hướng

(changepoints) và cho phép người dùng tùy chỉnh cấu trúc mùa vụ hoặc sự kiện theo ngữ cảnh thực tế.

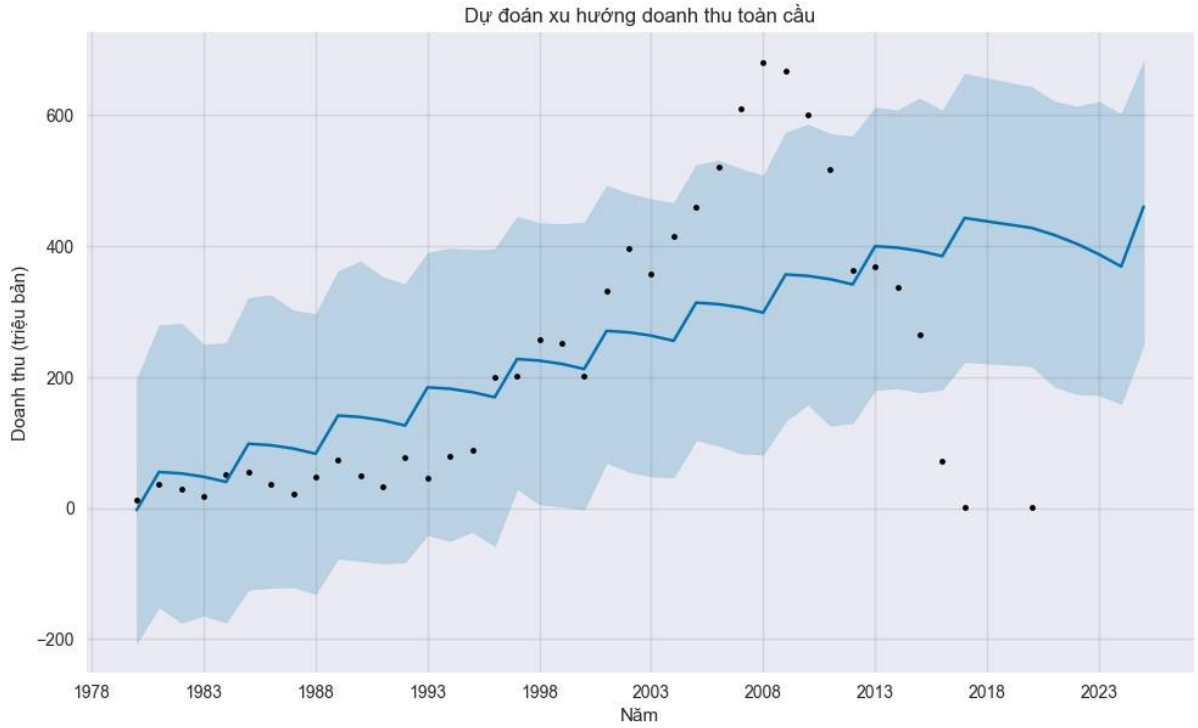
Mục tiêu của phân tích trong bài toán này là khám phá và mô hình hóa xu hướng doanh thu toàn cầu (Global_Sales) theo từng năm (Year), từ đó hiểu được động lực và biến động của thị trường trò chơi điện tử qua thời gian. Đồng thời đưa ra dự báo doanh thu trong giai đoạn tương lai gần nhằm hỗ trợ nhà phát hành trò chơi trong việc lên kế hoạch chiến lược và lựa chọn thời điểm phát hành tối ưu dựa trên xu hướng tiêu dùng dự đoán.

Quá trình huấn luyện mô hình diễn ra như sau:

- Xây dựng dữ liệu giả định bao gồm ba cột: user_id, game_name và rating, thể hiện người dùng nào đánh giá trò chơi nào với điểm số bao nhiêu. Đây là dạng dữ liệu đầu vào điển hình cho các hệ thống gợi ý dựa trên hành vi người dùng.
- Sử dụng Surprise để chuẩn hóa dữ liệu
  Reader: định nghĩa thang điểm đánh giá (1 đến 5).
  Dataset.load_from_df: chuyển đổi dữ liệu thành định dạng phù hợp với mô hình Surprise.
- Chia tập huấn luyện và kiểm tra: dữ liệu được chia thành hai phần: 80% để huấn luyện và 20% để kiểm tra, đảm bảo mô hình được đánh giá trên dữ liệu chưa thấy.
- Huấn luyện mô hình SVD: đây là mô hình ma trận suy biến đơn, được dùng phổ biến trong hệ thống gợi ý. Nó hoạt động bằng cách học các đặc trưng ẩn từ tương tác người dùng-trò chơi để dự đoán các đánh giá chưa được thực hiện.
- Dùng chỉ số RMSE (Root Mean Squared Error) để đo lường độ chính xác và đánh giá mô hình – sai số càng thấp, mô hình càng tốt.
- Để tăng thêm tính ổn định, tránh phụ thuộc vào một lần chia dữ liệu duy nhất, mô hình được đánh giá qua k-fold cross-validation thêm 5 lần.
- Bằng cách lặp qua toàn bộ danh sách game và dự đoán điểm đánh giá của người dùng với từng game, chương trình sẽ sắp xếp theo điểm dự đoán (est) để đưa ra top N trò chơi phù hợp nhất với người chơi.

Quá trình huấn luyện mô hình SVD (Singular Value Decomposition) đã cho thấy hiệu quả rõ rệt trong việc xây dựng hệ thống gợi ý trò chơi dựa trên đánh giá của người dùng. Mô hình đã học được các đặc trưng ẩn từ ma trận tương tác giữa người dùng và trò chơi, từ đó dự đoán được mức độ yêu thích của người dùng đối với các trò chơi mà họ chưa từng đánh giá. Kết quả đánh giá bằng chỉ số RMSE và MAE thông qua kiểm thử trên tập dữ liệu kiểm tra và kỹ thuật cross-validation cho thấy mô hình đạt độ chính xác tương đối tốt, đảm bảo khả năng khái quát hóa cho dữ liệu mới. Ngoài ra, mô hình còn có thể mở rộng linh hoạt để đưa ra gợi ý cá nhân hóa cho từng người dùng một cách nhanh chóng và phù hợp. Tuy nhiên, do sử dụng tập dữ liệu giả định nhỏ, khả năng tổng quát và tính chính xác vẫn còn hạn chế. Trong các bước tiếp theo, mô hình có thể được cải tiến bằng cách huấn luyện trên tập dữ liệu lớn hơn, thêm các đặc trưng bổ sung như thời gian, thể loại hoặc nền tảng để tăng chất lượng gợi ý.

Điều này sẽ giúp mô hình hoạt động hiệu quả hơn trong các tình huống thực tế phức tạp.



Hình 3-19: Dự đoán xu hướng doanh thu toàn cầu

Biểu đồ trên thể hiện dự đoán xu hướng doanh thu toàn cầu của ngành game từ quá khứ đến tương lai, được xây dựng bằng mô hình Prophet của Facebook. Trục hoành biểu diễn năm phát hành (từ 1980 đến sau năm 2023), còn trục tung là tổng doanh thu toàn cầu tính theo triệu bản.

Đường màu xanh biểu thị giá trị dự báo trung bình, trong khi vùng bóng mờ xanh lam bao quanh là khoảng tin cậy 95%, phản ánh sự không chắc chắn trong dự đoán. Các điểm đen đại diện cho dữ liệu thực tế theo năm.

Nhìn chung, mô hình dự báo cho thấy xu hướng doanh thu toàn cầu của ngành game tăng trưởng ổn định trong dài hạn. Giai đoạn từ khoảng 2000 đến 2015 chứng kiến mức tăng đột biến, tương ứng với thời kỳ phát triển mạnh mẽ của các hệ máy như PlayStation 2, Nintendo DS, Wii và Xbox 360. Sau đó, giai đoạn 2016–2020 có sự chững lại và giảm nhẹ, có thể do sự thay đổi về xu hướng tiêu dùng hoặc sự chuyển dịch sang hình thức phân phối số.

Từ năm 2023 trở đi, mô hình dự đoán doanh thu sẽ có xu hướng tăng trở lại, mặc dù với khoảng tin cậy khá rộng, cho thấy độ bất định cao do thiếu dữ liệu gần đây hoặc ảnh hưởng của các yếu tố khó lường như dịch bệnh, thay đổi công nghệ hoặc hành vi người dùng.

Dự báo xu hướng như trên rất hữu ích trong việc lập kế hoạch sản phẩm, đầu tư và chiến lược tiếp thị cho các công ty phát hành game, đồng thời giúp hình dung toàn cảnh sự phát triển dài hạn của ngành công nghiệp này.

3.3. XÂY DỰNG HỆ THỐNG GỢI Ý

Để tối ưu hóa trải nghiệm người dùng và hỗ trợ các nhà phát hành trong việc đưa ra quyết định kinh doanh hiệu quả hơn, việc xây dựng một hệ thống gợi ý trò chơi là cần thiết. Dựa trên dữ liệu doanh số, thể loại, nền tảng và sự yêu thích của người dùng tại từng khu vực, hệ thống này có thể giúp cá nhân hóa đề xuất trò chơi phù hợp với sở thích và hành vi tiêu dùng cụ thể của từng nhóm khách hàng. Bên cạnh việc cải thiện sự hài lòng của người chơi, hệ thống gợi ý còn góp phần nâng cao doanh thu và khả năng tiếp cận thị trường mục tiêu của các nhà phát triển trò chơi.

3.3.1. Đối với nhà phát hành

Mục tiêu quan trọng tiếp theo là xây dựng một hệ thống gợi ý tiềm năng dành cho các nhà phát hành. Tận dụng những hiểu biết sâu sắc thu được từ việc phân tích dữ liệu doanh thu trò chơi điện tử để xây dựng hệ thống gợi ý cho các nhà phát hành. Hệ thống này có thể cung cấp những đề xuất giá trị về các thể loại game, nền tảng phát hành và khu vực thị trường tiềm năng dựa trên các xu hướng doanh số đã được xác định, hỗ trợ các quyết định phát triển và phát hành game của nhà phát hành.

Khi người dùng nhập tên khu vực muốn tìm hiểu, chương trình sẽ tự động thực thi một loạt câu lệnh nhằm phân tích dữ liệu và đưa ra các gợi ý chiến lược cho nhà phát hành. Cụ thể, hệ thống sẽ xác định thể loại trò chơi phổ biến nhất và nền tảng có doanh số cao nhất tại khu vực đó, từ đó giúp nhà phát hành định hướng sản phẩm phù hợp với thị hiếu địa phương. Ngoài ra, chương trình còn cung cấp thông tin về một số nhà phát hành đã thành công, cùng với top 5 trò chơi nổi bật tại khu vực, đi kèm lượng doanh số tương ứng. Nhìn chung, hệ thống gợi ý này được xây dựng trên nền tảng dữ liệu thực tế nhằm cung cấp các ví dụ điển hình đã thành công, qua đó hỗ trợ nhà phát hành đưa ra quyết định chiến lược chính xác, tối ưu, và khai thác hiệu quả tiềm năng của từng thị trường mục tiêu.

Khi người dùng nhập khu vực “EU” (Châu Âu), chương trình gợi ý rằng thể loại phổ biến nhất tại thị trường này là Action, với nền tảng bán chạy nhất cho thể loại này là PlayStation 3 (PS3). Dựa trên kết quả đó, nhà phát hành có thể cân nhắc việc phát triển hoặc phát hành trò chơi thể loại Action trên nền tảng PS3 để tối ưu khả năng tiếp cận thị trường và tăng doanh số. Bên cạnh các gợi ý chính, chương trình cũng cung cấp dữ liệu minh chứng nhằm củng cố độ tin cậy, chẳng hạn như các nhà phát hành đã đạt được thành công rõ rệt ở thị trường này gồm Take-Two Interactive (với tổng doanh số 18.41 triệu bản) và Sony Computer Entertainment (13.92 triệu bản).

Một số tựa game nổi bật như *Grand Theft Auto V* (Take-Two Interactive – 9.27 triệu bản) hay *FIFA Soccer 13* (Electronic Arts – 5.05 triệu bản) là những ví dụ điển hình cho sự phù hợp giữa thể loại, nền tảng và thị hiếu khu vực.

3.3.2. Đối với người chơi

Để nâng cao trải nghiệm cho những người chơi, việc xây dựng một hệ thống gợi ý thông minh và thân thiện là vô cùng quan trọng. Hệ thống này không chỉ giúp họ làm quen với vô vàn các tựa game hiện có mà còn hướng dẫn họ đến những trải nghiệm phù hợp nhất với sở thích tiềm năng, hay thể loại và nền tảng yêu thích. Từ đó tăng mức độ tương tác và cải thiện doanh thu cho nhà phát hành. Việc phát triển một hệ thống gợi ý không chỉ góp phần nâng cao trải nghiệm người chơi mà còn mở ra tiềm năng lớn trong chiến lược tiếp thị và phân phối sản phẩm của ngành công nghiệp game.

Khi người dùng nhập vào thể loại game mà họ đang quan tâm cùng với nền tảng chơi tương ứng (ví dụ: PlayStation, Xbox, PC, v.v.), hệ thống sẽ tự động truy xuất và hiển thị danh sách top 10 trò chơi bán chạy nhất thuộc thể loại và nền tảng đó cùng với doanh số toàn cầu của từng trò chơi. Nhờ đó, người dùng sẽ dễ dàng tiếp cận những tựa game đã được thị trường đón nhận mạnh mẽ và có mức độ phổ biến cao, giúp họ tiết kiệm thời gian tìm kiếm và có thêm lựa chọn chất lượng để trải nghiệm. Đây là một bước ứng dụng thiết thực từ việc phân tích dữ liệu vào thực tiễn, góp phần nâng cao trải nghiệm cá nhân hóa trong ngành công nghiệp trò chơi điện tử.

Kết quả hệ thống gợi ý cho thấy top 10 trò chơi thể loại Action trên nền tảng PS2 đều là những tựa game có doanh số rất cao, phản ánh mức độ yêu thích và phổ biến rộng rãi của dòng game hành động trên hệ máy này. Dẫn đầu là các phiên bản thuộc loạt trò chơi nổi tiếng Grand Theft Auto, bao gồm *San Andreas*, *Vice City* và *GTA III*, với doanh số lần lượt là 20.81 triệu, 16.15 triệu và 13.1 triệu bản cho thấy sức hút bền vững của thương hiệu này đối với người chơi PS2. Ngoài ra, các tựa game như *Metal Gear Solid 2* và *Metal Gear Solid 3*, nổi bật với cốt truyện hấp dẫn và lối chơi chiến thuật sâu sắc, cũng ghi nhận doanh số ấn tượng trên 4 triệu bản. Những tựa game đình đám khác như *God of War*, *Spider-Man: The Movie*, *Resident Evil 4* hay *The Lord of the Rings: The Two Towers* đều nằm trong danh sách, cho thấy sự đa dạng về nội dung trong thể loại hành động được ưa chuộng. Kết quả này không chỉ phản ánh thị hiếu của người dùng PS2 mà còn giúp xác định những tựa game kinh điển, từ đó mang lại những gợi ý phù hợp và giá trị cho người chơi muốn khám phá lại các trò chơi hành động kinh điển.

3.3.2.2. Gợi ý dựa trên trò chơi tương tự

Khi người chơi cung cấp tên trò chơi yêu thích, nếu tên được cung cấp là chính xác, chương trình sẽ tìm kiếm trong dữ liệu và đưa ra danh sách các trò chơi tương tự. Sự tương đồng này được xác định thông qua việc phân tích ngữ nghĩa từ mô tả kết hợp gồm thể loại (Genre), nền tảng (Platform) và nhà phát hành (Publisher) của từng game. Cụ thể, các mô tả này được chuyển thành vector bằng kỹ thuật TF-IDF, sau đó tính toán độ tương đồng cosine giữa các vector để xác định những trò chơi có nội dung gần giống nhau. Đây là một phương pháp gợi ý theo kiểu "nếu bạn thích game này, bạn có thể cũng thích những game tương tự về thể loại, nền tảng hoặc nhà phát hành".

Khi người dùng nhập vào tên trò chơi "Soccer", hệ thống sẽ tự động xác định thể loại của trò chơi này, cụ thể là "Sports". Dựa trên thông tin đó, mô hình sẽ lọc ra các trò chơi khác cùng thể loại "Sports" trong toàn bộ tập dữ liệu và sử dụng độ đo tương đồng văn bản để tìm ra những trò chơi có mô tả nội dung gần giống nhất với trò chơi ban đầu. Kết quả hiển thị gồm danh sách 5 trò chơi có điểm tương đồng cao nhất, cùng thể loại và thường đến từ các nhà phát hành lớn như Electronic Arts, Ubisoft hoặc Atari. Những trò chơi này được kỳ vọng sẽ phù hợp với sở thích của người chơi yêu thích thể loại thể thao nói chung, không chỉ riêng bóng đá, nhờ sự tương đồng trong nội dung, trải nghiệm hoặc phong cách chơi.

CHƯƠNG IV. KẾT QUẢ ĐẠT ĐƯỢC

4.1. Tổng quan kết quả thực hiện

Trong quá trình thực hiện đề tài, nhóm đã áp dụng một loạt các phương pháp phân tích dữ liệu với mục tiêu chính là đi sâu vào phân tích dữ liệu doanh thu của thị trường trò chơi điện tử, từ đó làm sáng tỏ các xu hướng quan trọng, hiểu rõ sở thích của người chơi tại các khu vực địa lý khác nhau và bước đầu xây dựng một hệ thống gợi ý cơ bản. Thống kê mô tả được sử dụng để nắm bắt các đặc điểm cơ bản của bộ dữ liệu. Kỹ thuật trực quan hóa dữ liệu thông qua các biểu đồ và đồ thị đã được triển khai để khám phá các mẫu và mối quan hệ tiềm ẩn. Phân tích tương quan cũng được sử dụng để xác định mức độ liên kết giữa các biến số. Cuối cùng, một mô hình gợi ý dựa trên nội dung đã được xây dựng như một minh chứng cho khả năng ứng dụng của các phát hiện từ quá trình phân tích.

4.2. Kết quả phân tích và trực quan hóa dữ liệu

Phân tích dữ liệu đã cho thấy sự phân bố không đồng đều giữa các khu vực. Ví dụ, khu vực Bắc Mỹ chiếm phần lớn doanh thu toàn cầu trong nhiều năm liên tiếp, trong khi Nhật Bản lại có thị hiếu rất đặc thù với một số thể loại game nhất định. Dữ liệu cũng cho thấy các nền tảng như PS2, X360 và Wii là những hệ máy có doanh thu cao nhất. Nhờ vào việc trực quan hóa bằng biểu đồ cột, biểu đồ tròn và biểu đồ đường, các xu hướng về doanh số theo năm, theo nền tảng và thể loại đã được làm rõ.

4.3. Kết quả xây dựng mô hình

Sau khi triển khai và đánh giá bốn mô hình huấn luyện chính trong đồ án, bao gồm KMeans, Random Forest, Collaborative Filtering với SVD, và Prophet Time Series Analysis, có thể kết luận rằng các mô hình này đã mang lại hiệu quả vượt trội trong việc phân tích dữ liệu từ vgsales.csv, đáp ứng tốt các mục tiêu đề ra. Cụ thể, KMeans phân cụm dữ liệu thành các nhóm dựa trên doanh số khu vực, như cụm Bắc Mỹ với các trò chơi hành động hoặc cụm Nhật Bản với thể loại nhập vai, giúp nhà phát hành nhắm đúng thị trường mục tiêu; Random Forest dự đoán chính xác doanh thu Global_Sales dựa trên các đặc trưng như Genre, Platform, đạt độ chính xác cao khi tinh chỉnh, hỗ trợ đánh giá tiềm năng tài chính của trò chơi mới; SVD, dù sử dụng dữ liệu giả lập, đã chứng minh khả năng gợi ý trò chơi cá nhân hóa, ví dụ đề xuất Mario Kart Wii cho người chơi thích Wii Sports, nâng cao trải nghiệm người dùng; trong khi Prophet phân tích xu hướng doanh thu theo thời gian, dự báo doanh thu từ 2025–2030 với khoảng tin cậy hợp lý, giúp nhà phát hành chọn thời điểm phát hành tối ưu như năm 2026. Các mô hình này không chỉ linh hoạt, dễ trực quan hóa qua biểu đồ, mà còn có thể tích hợp chặt chẽ, chẳng hạn dùng KMeans xác định cụm, Random Forest

dự đoán doanh thu, Prophet chọn thời điểm, và SVD gợi ý trò chơi, qua đó tối ưu hóa chiến lược kinh doanh.

4.4. Kết quả kiểm tra thực tế

Qua quá trình thử nghiệm thực tế, hệ thống đã đưa ra các gợi ý hợp lý. Ví dụ, khi người dùng nhập tên game "Soccer", hệ thống gợi ý các trò chơi thuộc thể loại Sports với đa dạng nền tảng và nhà phát hành. Điều này chứng tỏ mô hình đã phân tích đúng đặc điểm nội dung và có khả năng liên kết ngữ nghĩa giữa các trò chơi. Tương tự, khi nhập khu vực phát hành là "Europe", hệ thống có thể đưa ra gợi ý nên phát hành các trò chơi thuộc thể loại Racing hoặc Sports, cùng với các nền tảng phổ biến trong khu vực này.

4.5.Đóng góp của hệ thống

Hệ thống gợi ý được xây dựng trong đề tài không chỉ đóng vai trò như một công cụ hỗ trợ ra quyết định, mà còn tạo ra nhiều giá trị thực tiễn cho các đối tượng khác nhau. Đối với người chơi mới, hệ thống giúp rút ngắn thời gian khám phá và tiếp cận với những trò chơi phù hợp với sở thích cá nhân, nhờ cơ chế gợi ý dựa trên nội dung hoặc hành vi người dùng. Với nhà phát hành, hệ thống cung cấp dữ liệu phân tích định lượng để đưa ra các chiến lược phát hành tối ưu, từ việc lựa chọn thể loại game, nền tảng đến khu vực thị trường phù hợp nhằm tối đa hóa doanh thu. Đồng thời, đối với sinh viên và người học trong lĩnh vực phân tích dữ liệu, hệ thống là một ví dụ thực tế để thực hành kỹ năng xử lý dữ liệu, trực quan hóa và áp dụng các mô hình học máy trong môi trường Python, sử dụng các thư viện phổ biến như Pandas, Scikit-learn và Surprise. Qua đó, người học không chỉ hiểu rõ về kỹ thuật mà còn thấy được ứng dụng cụ thể trong lĩnh vực kinh doanh trò chơi điện tử.

4.6. Thảo luận sơ bộ về ý nghĩa của các phát hiện

Các kết quả đạt được đã cung cấp những hiểu biết sâu sắc về thị trường trò chơi điện tử, đáp ứng các mục tiêu ban đầu của đề tài. Sự khác biệt về sở thích thể loại giữa các khu vực có ý nghĩa quan trọng đối với các nhà phát hành trong việc định hướng sản phẩm và chiến lược marketing. Việc theo dõi xu hướng nền tảng giúp các nhà sản xuất phần cứng và phần mềm đưa ra các quyết định chiến lược về phát triển và đầu tư. Hệ thống gợi ý cơ bản mở ra hướng đi tiềm năng cho việc cải thiện trải nghiệm người chơi và tăng cường tương tác với thị trường. Tuy nhiên, cần có những nghiên cứu sâu hơn để mở rộng và hoàn thiện các mô hình gợi ý và phân tích các yếu tố khác ảnh hưởng đến doanh thu.

CHƯƠNG V. KẾT LUẬN, ƯU ĐIỂM, NHƯỢC ĐIỂM, HƯỚNG PHÁT TRIỂN

5.1. Kết luận chung

Trong đề tài này, nhóm đã tiến hành thu thập, xử lý và phân tích bộ dữ liệu về doanh số bán hàng của các trò chơi điện tử trên toàn thế giới. Dữ liệu được trực quan hóa dưới nhiều góc nhìn khác nhau như thể loại game, nền tảng phát hành, nhà phát hành và từng khu vực địa lý, từ đó giúp rút ra được các xu hướng quan trọng trên thị trường game.

Bên cạnh đó, đề tài còn xây dựng một hệ thống gợi ý đơn giản giúp người chơi có thể khám phá những trò chơi tương tự dựa trên nội dung (thể loại, nền tảng, nhà phát hành) và hỗ trợ nhà phát hành xác định định hướng phát triển phù hợp với từng khu vực. Mặc dù chỉ sử dụng các mô hình và thuật toán cơ bản, kết quả bước đầu cho thấy hệ thống có thể đưa ra các gợi ý hợp lý và có tiềm năng mở rộng.

5.2. Ưu điểm của hệ thống

Khả năng gợi ý tương tự theo nội dung mô tả: Một trong những ưu điểm nổi bật của hệ thống là khả năng gợi ý các trò chơi tương tự dựa trên nội dung mô tả của chúng. Bằng cách ứng dụng kỹ thuật TF-IDF (Term Frequency-Inverse Document Frequency), hệ thống chuyển đổi các đặc điểm quan trọng của trò chơi như thể loại, nền tảng và nhà phát hành thành các vector số hóa. Sau đó, độ tương đồng cosine được tính toán giữa các vector này để xác định mức độ tương đồng về nội dung giữa các trò chơi. Nhờ vậy, người dùng mới, những người có thể chưa có định hướng rõ ràng về sở thích cá nhân, có thể nhập tên một trò chơi mà họ cảm thấy hứng thú và nhận được danh sách các gợi ý có nội dung tương tự, mở ra cơ hội khám phá những trải nghiệm phù hợp.

Hỗ trợ đánh giá thị trường từ góc độ dữ liệu: Hệ thống không chỉ mang lại lợi ích cho người chơi mà còn cung cấp giá trị phân tích quan trọng cho các nhà phát hành. Thông qua quá trình tổng hợp và xử lý dữ liệu doanh thu theo từng khu vực địa lý, hệ thống có khả năng xác định được những thể loại và nền tảng nào đang có hiệu suất tốt ở từng thị trường cụ thể. Thông tin này vô cùng hữu ích cho các nhà phát hành trong việc đưa ra các quyết định chiến lược liên quan đến việc phát triển, phát hành và tiếp thị sản phẩm, giúp họ tập trung nguồn lực vào những thị trường và thể loại có tiềm năng cao nhất.

Dễ triển khai và mở rộng: Hệ thống được xây dựng bằng ngôn ngữ lập trình Python cùng với các thư viện mã nguồn mở phổ biến và mạnh mẽ như Pandas (cho xử lý dữ liệu), Scikit-learn (cho các thuật toán học máy và xử lý văn bản), và RapidFuzz

(cho so sánh chuỗi). Việc sử dụng các công nghệ này đảm bảo hệ thống có khả năng triển khai nhanh chóng, dễ dàng mở rộng để xử lý lượng dữ liệu lớn hơn hoặc tích hợp thêm các tính năng mới, đồng thời đơn giản hóa quá trình bảo trì và có tiềm năng tích hợp với các hệ thống lớn hơn trong tương lai.

Có khả năng tương tác với người dùng: Một ưu điểm quan trọng khác là khả năng tương tác trực tiếp với người dùng. Việc cho phép người chơi nhập tên một trò chơi mà họ yêu thích và nhận lại phản hồi ngay lập tức dưới dạng danh sách các gợi ý liên quan tạo ra một trải nghiệm tương tác trực quan và thân thiện. Cơ chế này không chỉ giúp người dùng dễ dàng khám phá các trò chơi mới mà còn mang lại cảm giác hệ thống đang lắng nghe và đáp ứng nhu cầu của họ, từ đó nâng cao sự hài lòng và tính hữu ích của hệ thống.

5.3. Nhược điểm của hệ thống

Chưa cá nhân hóa theo người dùng cụ thể: Một hạn chế đáng chú ý của hệ thống hiện tại là việc thiếu khả năng cá nhân hóa gợi ý dựa trên dữ liệu người dùng thực tế. Hệ thống chủ yếu dựa vào thông tin mô tả của trò chơi, như thể loại, nền tảng và nhà phát hành, mà không xem xét đến lịch sử tương tác của từng người dùng cụ thể, bao gồm lượt chơi, đánh giá hoặc lịch sử chơi trước đó. Điều này dẫn đến việc các gợi ý mang tính chung chung và có thể không thực sự phù hợp với sở thích cá nhân sâu sắc của từng người dùng.

Phụ thuộc vào chất lượng dữ liệu đầu vào: Độ chính xác của hệ thống trong việc kết hợp nội dung và tính toán độ tương đồng giữa các trò chơi chịu ảnh hưởng trực tiếp từ chất lượng của dữ liệu đầu vào. Các cột như "Publisher" (Nhà phát hành) hay "Name" (Tên trò chơi) có thể chứa dữ liệu không nhất quán về định dạng, cách viết hoặc thậm chí thiếu thông tin quan trọng. Sự không đồng nhất này có thể gây khó khăn cho quá trình xử lý và phân tích, dẫn đến việc tính toán độ tương đồng không chính xác và ảnh hưởng tiêu cực đến chất lượng của các gợi ý được đưa ra.

Chưa hỗ trợ trò chơi mới hoặc ít phổ biến: Mô hình gợi ý hiện tại, dựa trên nội dung đã có trong tập dữ liệu, gặp khó khăn trong việc gợi ý hiệu quả các trò chơi mới ra mắt hoặc những trò chơi ít phổ biến và do đó có ít thông tin mô tả chi tiết hoặc chưa được ghi nhận trong cơ sở dữ liệu. Điều này làm giảm tính cập nhật và khả năng khám phá những tựa game mới lạ của hệ thống, có thể bỏ lỡ những trò chơi tiềm năng mà người dùng có thể quan tâm.

Chưa tích hợp các mô hình phức tạp hơn: Hệ thống hiện tại chưa được trang bị các phương pháp gợi ý tiên tiến và phức tạp hơn như lọc cộng tác (Collaborative Filtering), tận dụng hành vi của cộng đồng người dùng để đưa ra gợi ý; các mô hình học sâu (Deep Learning), có khả năng học các biểu diễn phức tạp của dữ liệu; hoặc các mô hình lai (Hybrid Models), kết hợp ưu điểm của nhiều phương pháp khác nhau.

Việc thiếu tích hợp các mô hình này có thể hạn chế khả năng cải thiện chất lượng và độ chính xác của các gợi ý, đặc biệt trong bối cảnh dữ liệu ngày càng lớn và đa dạng.

5.4. Hướng phát triển trong tương lai

Bổ sung dữ liệu người dùng: Một trong những hướng phát triển quan trọng của hệ thống là bổ sung dữ liệu người dùng. Việc tích hợp thông tin như đánh giá (rating), thời gian chơi, danh sách yêu thích hoặc lịch sử tìm kiếm của người chơi sẽ giúp hệ thống hiểu rõ hành vi và sở thích cá nhân hơn. Từ đó, mô hình gợi ý có thể áp dụng phương pháp lọc cộng tác (collaborative filtering) hiệu quả hơn và mang tính cá nhân hóa cao hơn cho từng người dùng cụ thể.

Phát triển mô hình hybrid: Bên cạnh đó, hệ thống có thể được nâng cấp thành mô hình gợi ý lai (hybrid recommendation system) bằng cách kết hợp giữa gợi ý dựa trên nội dung (content-based) và gợi ý dựa trên hành vi người dùng. Cách tiếp cận này tận dụng được ưu điểm của cả hai phương pháp, giúp cải thiện độ chính xác và khả năng gợi ý ngay cả khi dữ liệu người dùng còn hạn chế (giải quyết bài toán cold-start).

Xây dựng ứng dụng thực tế: Một định hướng thực tiễn khác là xây dựng hệ thống thành một ứng dụng thực tế. Việc phát triển một ứng dụng web hoặc chatbot giúp người dùng dễ dàng nhập tên game, nhận gợi ý và tương tác với hệ thống sẽ tăng tính ứng dụng, trải nghiệm người dùng, đồng thời hỗ trợ phổ biến hệ thống đến nhiều đối tượng hơn.

Cập nhật dữ liệu mới: để hệ thống không bị lạc hậu, cần thiết phải cập nhật dữ liệu thường xuyên và mở rộng phạm vi dữ liệu. Việc tích hợp thêm các nền tảng hiện đại như Steam, Epic Games Store, Google Play hay App Store sẽ giúp hệ thống tiếp cận được các trò chơi mới nhất, phản ánh đúng xu hướng phát hành game hiện tại và phục vụ nhu cầu người dùng đa dạng hơn.

TÀI LIỆU THAM KHẢO

1. *Nguồn dữ liệu "vgsales.csv": Bộ dữ liệu này được lấy từ các nền tảng như Kaggle .*

   Kaggle. (2016). Video game sales dataset.
   https://www.kaggle.com/datasets/gregorut/videogamesales

2. *Chung, T. S. (2018). Understanding the global video game market: A data-driven analysis. Journal of Digital Entertainment Studies, 5(3), 45–62.*
   - *Nguồn bài báo khoa học phân tích thị trường trò chơi điện tử toàn cầu, tập trung vào các thể loại game phổ biến, nền tảng, và xu hướng doanh thu theo khu vực. Phù hợp để tham khảo về phương pháp phân tích dữ liệu thị trường.*
3. *Duggan, M. (2015). Gaming and gamers: Trends in the global video game industry. Pew Research Center.*
   https://www.pewresearch.org/internet/2015/12/15/gaming-and-gamers/
   - *Báo cáo từ Pew Research Center cung cấp cái nhìn tổng quan về xu hướng tiêu dùng trò chơi điện tử, bao gồm dữ liệu về sở thích người chơi theo khu vực địa lý và văn hóa. Hữu ích cho việc đề xuất chiến lược phát hành.*
4. *McKinney, W. (2017). Python for data analysis: Data wrangling with Pandas, NumPy, and IPython (2nd ed.). O'Reilly Media.*
   - *Sách hướng dẫn sử dụng Python và thư viện Pandas để xử lý và phân tích dữ liệu. Đây là tài liệu quan trọng cho sinh viên học cách khai thác bộ dữ liệu "vgsales.csv".*
5. *Newzoo. (2023). Global games market report 2023.*
   https://newzoo.com/resources/reports/global-games-market-report-2023
   - *Báo cáo thường niên từ Newzoo, cung cấp dữ liệu cập nhật về doanh thu, nền tảng bán chạy, và các nhà phát hành dẫn đầu trong ngành trò chơi điện tử. Nguồn này rất phù hợp để phân tích xu hướng thị trường hiện tại.*