

Trường Đại Học Sài Gòn - SGU

Khoa Toán - Ứng Dụng



**BÁO CÁO ĐỒ ÁN CUỐI KÌ KẾT THÚC HỌC
PHẦN**

Bộ Môn: Nhập Môn Khoa Học Dữ Liệu (858008)

**CHỦ ĐỀ: “Đánh Giá Hành Vi Người Chơi và Đề Xuất Trò
Chơi”**

Giảng viên: Vũ Ngọc Thanh Sang

Sinh Viên: Nguyễn Trương Cao Sơn

Lớp: DDU1231

TP Hồ Chí Minh, ngày 10, tháng 12, năm 2023

MỤC LỤC

MỤC LỤC.....	2
LỜI MỞ ĐẦU.....	4
Chương I: GIỚI THIỆU VỀ KHOA HỌC DỮ LIỆU.....	5
1.1. Giới thiệu về dữ liệu.....	5
1.1.1. Sơ lược về dữ liệu.....	5
1.2. Giới Thiệu Khoa học dữ liệu.....	6
1.2.1. Khái quát về khoa học dữ liệu.....	6
1.2.2. Sự phát triển của khoa học dữ liệu.....	7
1.2.3. Ứng dụng của khoa học dữ liệu.....	9
Chương II: TỔNG QUAN VỀ ĐỀ TÀI.....	10
2.1. Giới thiệu về ngành công nghiệp game.....	10
2.1.1. Khởi đầu của ngành game.....	10
2.1.2. Ngành game hiện tại.....	11
2.2. Giới thiệu về đề tài.....	13
2.2.1. Giới thiệu về đề tài.....	13
2.2.2. Lý do chọn đề tài.....	14
Chương III: Ứng dụng phương pháp vào bài phân tích thực tế..	16
3.1. Phương pháp nghiên cứu.....	16
3.1.1. Tổng quan về phương pháp.....	16
3.1.2. Lựa chọn phương pháp cho đề tài.....	17
3.2. Thu thập dữ liệu.....	17
3.2.1. Phương pháp thu thập dữ liệu.....	17
3.2.2. Xử lý và làm sạch dữ liệu.....	19

3.3. Phân tích dữ liệu.....	22
3.3.1. Biểu đồ và đồ thị thể hiện kết quả phân tích.....	22
3.3.2. Đưa ra đề xuất và nhận xét.....	27
 LỜI CẢM ƠN.....	33
TÀI LIỆU THAM KHẢO.....	34

LỜI MỞ ĐẦU

Trong những năm gần đây, việc nắm bắt được thông tin được coi là cơ sở của mọi hoạt động sản xuất, kinh doanh. Các nhân hoặc tổ chức nào thu thập và hiểu được thông tin, và hành động dựa trên các thông tin được kết xuất từ các thông tin đã có sẽ đạt được thành công trong mọi hoạt động.

Sự tăng trưởng vượt bậc của các cơ sở dữ liệu (CSDL) trong cuộc sống như: thương mại, quản lý đã làm nảy sinh và thúc đẩy sự phát triển của kỹ thuật thu thập, lưu trữ, phân tích và khai phá dữ liệu... không chỉ bằng các phép toán đơn giản thông thường như: phép đếm, thống kê... mà đòi hỏi một cách xử lý thông minh hơn, hiệu quả hơn. Các kỹ thuật cho phép ta khai thác được tri thức hữu dụng từ CSDL (lớn) được gọi là các kỹ thuật Khai phá dữ liệu (datamining). Đồ án nghiên cứu về những khái niệm cơ bản về khai phá dữ liệu, luật kết hợp và ứng dụng thuật toán khai phá luật kết hợp trong CSDL

Chương I: GIỚI THIỆU VỀ KHOA HỌC DỮ LIỆU

1.1. Giới thiệu về dữ liệu:

1.1.1. Sơ lược về dữ liệu:

Dữ liệu: "Data"

Dữ liệu là các giá trị thông tin định lượng hoặc định tính của các sự vật, hiện tượng trong cuộc sống. Trong khoa học dữ liệu, dữ liệu được dùng như một cách biểu diễn hình thức hoá của thông tin về các sự kiện, hiện tượng thích ứng với các yêu cầu truyền nhận, thể hiện và xử lý bằng máy tính.

Dữ liệu có hai loại chính: dữ liệu có cấu trúc và dữ liệu không có cấu trúc.

Dữ liệu có cấu trúc, còn được gọi là dữ liệu định lượng, là dạng dữ liệu và số liệu khách quan. Thông thường, nó được biểu diễn dưới dạng số hoặc chữ và được lưu trữ trong các hệ thống như **Excel**, **SQL** hoặc **Google Sheet**. Dữ liệu này dễ dàng thu thập, truy xuất, lưu trữ và sắp xếp, đồng thời cho phép trích xuất thông tin một cách dễ dàng.

Dữ liệu không có cấu trúc, còn được gọi là dữ liệu định tính, thường là các ý kiến chủ quan và đánh giá thương hiệu được biểu diễn dưới dạng văn bản. Nó chỉ tồn tại dưới dạng văn bản và có thể được lưu trữ trong các tài liệu **Word**, **Elasticsearch** hoặc **Solr**, nơi có thể thực hiện các truy vấn tìm kiếm từ và cụm từ. Dữ liệu không có cấu trúc khó thu thập và gây khó khăn cho việc xuất, lưu trữ và sắp xếp trong các cơ sở dữ liệu thông thường. Ngoài ra, không thể áp dụng các phương pháp và công cụ phân tích dữ liệu trực tiếp lên dữ liệu này.

Dữ liệu là nguồn thông tin quan trọng trong lĩnh vực khoa học dữ liệu. Nó thể hiện các thông tin, sự kiện và thuộc tính của các đối tượng trong thế giới thực hoặc trong một hệ thống.

Trong tất cả các lĩnh vực, dữ liệu đóng vai trò quan trọng trong việc cung cấp thông tin và kiến thức. Dữ liệu cung cấp cơ sở cho việc phân tích, nghiên cứu và đưa ra quyết định. Dữ liệu cũng là nguồn cung cấp thông tin quan trọng để phát hiện xu hướng, tìm ra mối quan hệ giữa các biến và xây dựng mô hình dự đoán.

1.2. Giới Thiệu Khoa học dữ liệu:

1.2.1. Khái quát về khoa học dữ liệu:

Khoa học dữ liệu là một lĩnh vực nghiên cứu và ứng dụng sử dụng các phương pháp, công cụ và kỹ thuật để trích xuất, xử lý, phân tích và hiểu dữ liệu. Nó kết hợp các lĩnh vực như toán học, thống kê, khoa học máy tính và tri thức kinh doanh để tìm hiểu thông tin ẩn chứa trong dữ liệu và tạo ra kiến thức có ích để đưa ra quyết định và dự đoán.

Trong thời đại số hóa hiện nay, dữ liệu được tạo ra và tích lũy với tốc độ chóng mặt từ nhiều nguồn khác nhau như mạng xã hội, cảm biến, máy móc và giao dịch điện tử. Khoa học dữ liệu giúp chúng ta khám phá, phân tích và tìm hiểu dữ liệu này để tạo ra giá trị. Các bước chính trong quá trình khoa học dữ liệu bao gồm:

(1) Thu thập dữ liệu: Đây là quá trình thu thập dữ liệu từ các nguồn khác nhau như cơ sở dữ liệu, tệp tin, trang web hoặc API.

(2) Tiền xử lý dữ liệu: Dữ liệu thường không hoàn hảo và có thể chứa lỗi, thiếu sót hoặc nhiễu. Bước này liên quan đến làm sạch, chuyển đổi và chuẩn hóa dữ liệu để nó có thể được sử dụng cho phân tích.

(3) Phân tích dữ liệu: Đây là quá trình tìm hiểu và khám phá dữ liệu bằng cách áp dụng các phương pháp và công cụ thống kê, khai phá dữ liệu và học máy. Mục tiêu là tìm ra mẫu, xu hướng và thông tin hữu ích từ dữ liệu.

(4) Xây dựng mô hình: Dựa trên các kết quả phân tích dữ liệu, các mô hình dự đoán có thể được xây dựng để đưa ra dự đoán và đưa ra quyết định. Điều này có thể bao gồm việc sử dụng các thuật toán máy học như học có giám sát, học không giám sát và học tăng cường.

(5) Triển khai và tối ưu hóa: Sau khi xây dựng mô hình, nó có thể được triển khai vào sản phẩm hoặc quá trình thực tế. Tuy nhiên, quá trình này còn liên quan đến việc áp dụng mô hình và công nghệ đã xây dựng vào môi trường thực tế. Điều này bao gồm kiểm tra và

đánh giá hiệu suất của mô hình, tối ưu hóa các tham số và quy trình để đạt được kết quả tốt nhất.

Khoa học dữ liệu là một lĩnh vực quan trọng trong thế giới kỹ nguyên số hiện đại. Nó liên quan đến việc khám phá, phân tích và hiểu dữ liệu để tạo ra thông tin và kiến thức có giá trị. Khoa học dữ liệu kết hợp các phương pháp, công cụ và kỹ thuật từ nhiều lĩnh vực như toán học, thống kê, khoa học máy tính và tri thức kinh doanh.

Qua các bước thu thập, tiền xử lý, phân tích dữ liệu và xây dựng mô hình, khoa học dữ liệu giúp chúng ta tìm hiểu thông tin tiềm ẩn trong dữ liệu và áp dụng nó vào quyết định và dự đoán. Quá trình này cũng liên quan đến việc làm sạch, chuẩn hóa và tối ưu hóa dữ liệu để đảm bảo tính chính xác và tin cậy của kết quả.

Trong lĩnh vực khoa học dữ liệu, dữ liệu là trung tâm của quá trình nghiên cứu. Khoa học dữ liệu liên quan đến việc thu thập, lưu trữ, xử lý và phân tích dữ liệu để tìm ra liệu để đảm bảo tính chính xác và tin cậy của kết quả.

1.2.2. Sự phát triển của khoa học dữ liệu:

Sự phát triển của khoa học dữ liệu đã là một xu hướng quan trọng trong thập kỷ gần đây và tiếp tục mở ra nhiều cơ hội và thách thức trong tương lai. Cùng với sự phát triển của khoa học dữ liệu đó thì cuộc sống ngày càng thuận tiện và dễ dàng hơn như:

- Tăng cường khả năng tính toán: Sự phát triển nhanh chóng của công nghệ tính toán, đặc biệt là việc sử dụng đám mây và công nghệ xử lý song song, đã cung cấp khả năng tính toán mạnh mẽ hơn cho việc xử lý và phân tích dữ liệu lớn. Điều này cho phép xử lý nhanh chóng và hiệu quả các tập dữ liệu phức tạp và lớn.
- Mở nguồn dữ liệu: Sự phát triển của Internet, truyền thông xã hội và các thiết bị di động

đã tạo ra một lượng dữ liệu khổng lồ. Sự gia tăng này trong nguồn dữ liệu đã tạo ra một cơ sở cho việc phân tích và khai thác thông tin từ các nguồn này. Ngoài ra, việc mở nguồn dữ liệu công cộng và khởi xướng các dự án dữ liệu mở đã tạo ra cơ hội rất lớn cho sự phát triển của khoa học dữ liệu.

- Tiến bộ trong kỹ thuật và công nghệ 🖥️: Các phương pháp và công nghệ trong khoa học dữ liệu đã trải qua sự phát triển đáng kể. Các thuật toán học máy và học sâu ngày càng được cải thiện, các mô hình và kiến trúc mới được phát triển, và các công cụ và framework phân tích dữ liệu để sử dụng đã xuất hiện. Điều này giúp giảm độ phức tạp và thời gian triển khai của các dự án khoa học dữ liệu.

- Học máy và trí tuệ nhân tạo 🤖: Khoa học dữ liệu đã đóng vai trò quan trọng trong sự phát triển của học máy và trí tuệ nhân tạo. Việc áp dụng các phương pháp và thuật toán học máy vào việc phân tích dữ liệu đã mang lại khả năng tự học và tự động hóa cho các hệ thống. Trí tuệ nhân tạo cũng đã được phát triển với việc sử dụng dữ liệu lớn để huấn luyện mô hình và tạo ra các ứng dụng thông minh trong nhiều lĩnh vực, như xe tự lái, robot hội thoại, chẩn đoán y tế và hơn thế nữa.

- Xử lý dữ liệu thời gian thực ⌚: Sự phát triển của khoa học dữ liệu đã đưa đến khả năng xử

lý dữ liệu thời gian thực. Với việc sử dụng các công nghệ như xử lý dữ liệu đám mây, cơ sở dữ liệu phân tán và hệ thống xử lý phân tán, chúng ta có thể phân tích và trích xuất thông tin từ dữ liệu được tạo ra và cập nhật liên tục. Điều này cho phép các ứng dụng theo thời gian thực như giám sát và dự đoán dựa trên dữ liệu liên tục.

- Bảo mật và quyền riêng tư 🛡️: Với việc sử dụng dữ liệu cá nhân và nhạy cảm, bảo mật và quyền riêng tư trở thành một vấn đề quan trọng trong khoa học dữ liệu. Sự phát triển của các phương pháp và kỹ thuật bảo mật đã đảm bảo rằng dữ liệu được bảo vệ và xử lý một cách an toàn và tuân thủ các quy định về quyền riêng tư.

- Điều chỉnh và luật pháp 📜: Sự phát triển của khoa học dữ liệu đã đặt ra nhiều thách thức về điều chỉnh và luật pháp. Vấn đề về quyền riêng tư, bảo vệ dữ liệu và trách nhiệm trong việc sử dụng dữ liệu đã trở thành một vấn đề quan trọng. Các quy định và quyền riêng tư dữ liệu đã được đưa ra để đảm bảo việc sử dụng dữ liệu đúng cách và có trách nhiệm....

- Ứng dụng đa lĩnh vực 🌀: Khoa học dữ liệu không chỉ được áp dụng trong lĩnh vực công nghệ thông tin và kỹ thuật, mà còn trong

nhiều lĩnh vực khác như y tế, tài chính, bán lẻ, marketing, sản xuất, vận tải và nhiều ngành công nghiệp khác.

1.2.3. Ứng dụng của khoa học dữ liệu:

Khoa học dữ liệu dùng trong phân tích dữ liệu; dự báo dự đoán; xây dựng hệ thống thông minh; tối ưu hóa và quyết định; khai thác dữ liệu xã hội; phân tích hình ảnh và video; kỹ thuật tăng cường thực tế (AR) và thực tế ảo (VR); kỹ thuật tìm kiếm và đề xuất;....

Một trong những ứng dụng cụ thể chúng ta thường thấy là: dự đoán số lượng hành khách; chiến lược quản lý quan hệ khách hàng, dự đoán trễ chuyến bay, tỷ lệ hủy; đảm bảo an toàn và bảo mật cho hành khách; tỷ lệ người tiêm vaccine covid-19 bị nhiễm covid-19; dự báo biến động thị trường tài chính; dự báo khách hàng trả nợ ngân hàng trước hạn hàng quý ...



Hình 1.1 Netflix sử dụng hệ thống Recommendations từ kho dữ liệu để phân tích lịch sử xem của người dùng và đề xuất các nội dung mới dựa trên sở thích cá nhân.

Chương II: TỔNG QUAN VỀ ĐỀ TÀI

2.1. Giới thiệu về ngành công nghiệp game:

2.1.1. Khởi đầu của ngành game:



Trước khi ngành công nghiệp game trở thành đế chế vô song, nó bắt đầu như một cuộc thách thức sáng tạo trong thế giới máy tính ngày càng phức tạp. Đỉnh điểm của sự khởi

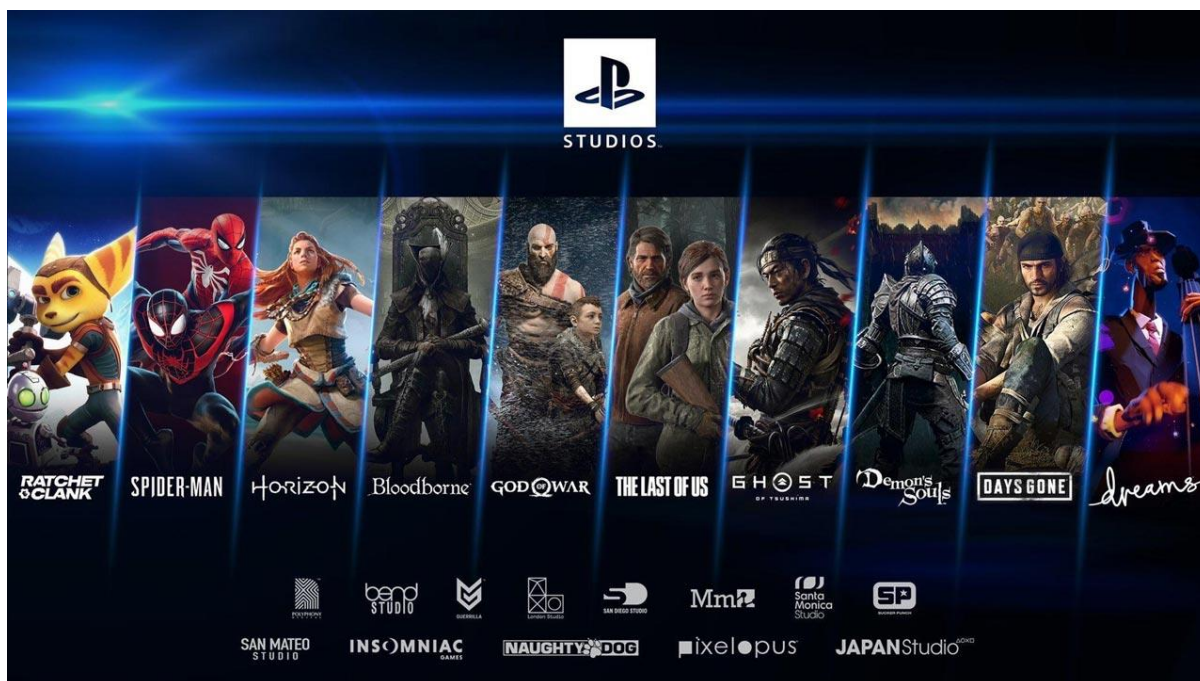
đầu này có thể được liên kết với "Spacewar!" - một trong những trò chơi đầu tiên xuất hiện vào những năm 1960. Được tạo ra bởi nhóm sinh viên tại Massachusetts Institute of Technology (MIT), "Spacewar!" không chỉ là một trò chơi đơn thuần, mà còn là một bước tiến quan trọng trong sự phát triển của ngành công nghiệp game.



Sự sáng tạo tiếp tục với sự xuất hiện của các máy arcade vào những năm 1970, như Pong của Atari, mở ra cánh cửa cho việc trải nghiệm trò chơi ngoại tuyến. Tính năng đầu tiên của ngành công nghiệp game - việc kết hợp giữa phần cứng và phần mềm để tạo ra những trải nghiệm tương tác - đã bắt đầu hình thành, tạo ra một nền tảng cho sự phát triển rộng lớn trong tương lai mặc dù vào thời điểm đó game vẫn bị xem như một thứ không lành mạnh cần tránh xa.

2.1.2. Ngành game hiện tại:

Ngày nay, ngành công nghiệp game đã trở thành một động lực mạnh mẽ của nền kinh tế toàn cầu. Công nghệ ngày càng tiến bộ, mang lại trải nghiệm chơi game chân thực và sống động hơn. Các hãng phát triển game hàng đầu như [Electronic Arts](#), [Ubisoft](#), và [Activision Blizzard](#) không chỉ tạo ra những tựa game giải trí mà còn xây dựng cộng đồng lớn mạnh quanh những tựa game của họ.





Người chơi ngày nay không chỉ thỏa mãn với việc chơi trên máy tính hoặc console mà còn khám phá thế giới ảo qua thực tế ảo và thực tế tăng cường. **Esports**, hay thể thao điện tử, trở thành một hiện thực với hàng triệu người hâm mộ và giải đấu quốc tế có giải thưởng lớn. Đồng thời, dịch vụ trực tuyến như **Xbox Game Pass** và **PlayStation Now** mang đến trải nghiệm chơi game theo mô hình dịch vụ, thay vì mua đĩa hoặc tải game riêng lẻ.



Từ những trò chơi đơn giản trên máy tính cỡ lớn đến cả thế giới ảo hiện đại, ngành công nghiệp game đã trải qua một hành trình đáng kinh ngạc, định hình văn hóa và giải trí của chúng ta.

2.2. Giới thiệu về đề tài:

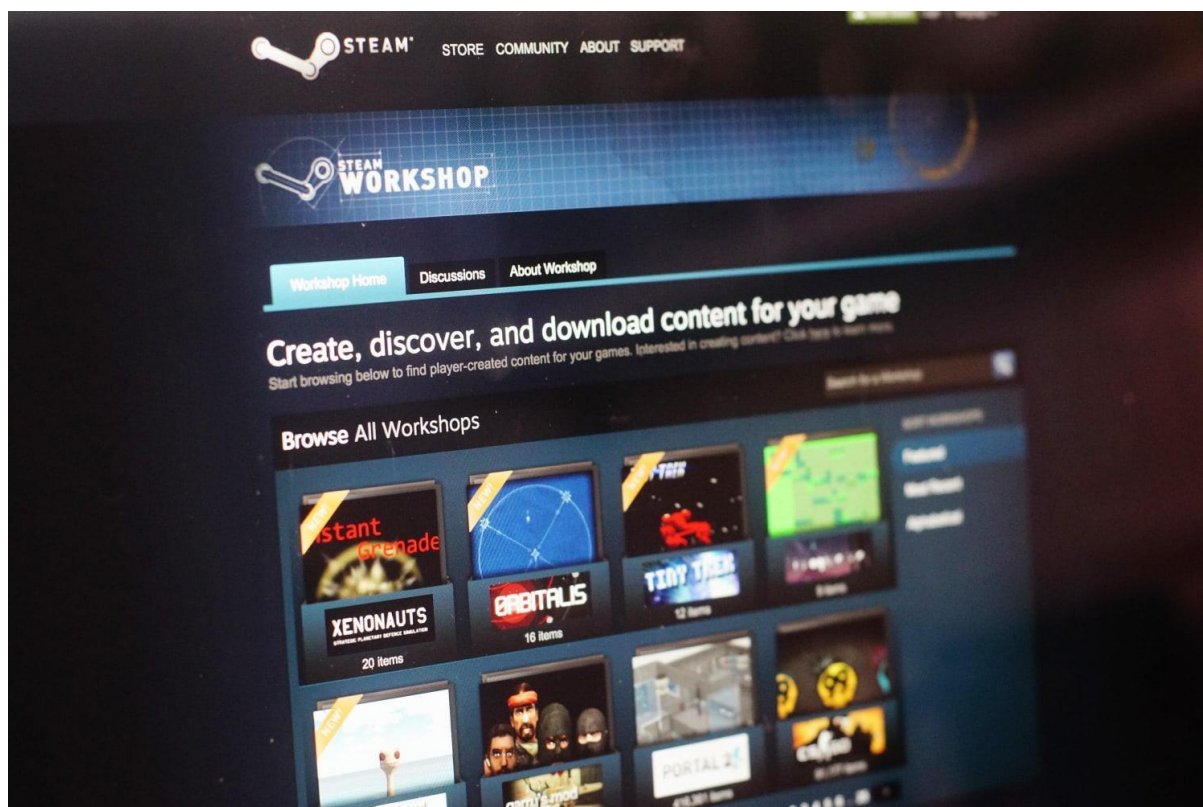
2.2.1. Giới thiệu đề tài:



Steam, nền tảng game trực tuyến do Valve Corporation phát triển, không chỉ là một cửa hàng điện tử cung cấp hàng nghìn tựa game mà

còn là một cộng đồng đa dạng của người chơi trên khắp thế giới. Với hơn 120 triệu người dùng tích cực hàng tháng, **Steam** đã trở thành trung tâm không thể thiếu của ngành công nghiệp game.

Nền tảng này không chỉ thu hút người chơi bằng việc cung cấp nhiều game đa dạng, từ đồ họa cao cấp đến indie độc đáo, mà còn tạo ra một hệ sinh thái mạnh mẽ với các tính năng như **Steam Workshop** cho việc chia sẻ nội dung sáng tạo và **Steam Community** cho giao lưu giữa cộng đồng.



Đề tài nghiên cứu về xử lý dữ liệu phân tích từ dataset [Steam](#) hành vi người chơi hứa hẹn là một cuộc đàm phán sâu rộng vào thế giới phức tạp của người chơi trực tuyến. Từ việc theo dõi thói quen chơi game đến việc phân tích tương tác trong cộng đồng, dataset này có thể cung cấp thông tin quý báu về hành vi người chơi, giúp hiểu rõ hơn về xu hướng, sở thích và đưa ra đề xuất hợp lý.

Đối diện với sự đa dạng ngày càng tăng về tựa game, đề án này mục tiêu tìm hiểu sâu rộng về hành vi người chơi và cung cấp đề xuất hợp lý cho việc chọn lựa tựa game.

2.2.2. Lý do chọn đề tài:

Chọn đề tài về xử lý dữ liệu phân tích từ dataset Steam hành vi người chơi là một quyết định tự nhiên dựa trên sự quan tâm đối với ngành công nghiệp game và khả năng cung cấp thông tin giá trị từ dữ liệu lớn. Dưới đây là một số lý do chính:

1. Sự Phổ Biến của Steam: Steam là một trong những nền tảng game lớn nhất và phổ biến nhất trên thế giới, có sức ảnh hưởng lớn đến cả ngành công nghiệp và cộng đồng người chơi. Nắm bắt thông tin từ người chơi trên Steam có thể mang lại cái nhìn tổng thể về xu hướng và sở thích trong cộng đồng này.

2. Đa Dạng trong Hành Vi Người Chơi: Steam thu hút người chơi với đa dạng loại game, từ AAA titles đến các tựa game độc lập. Phân tích dữ liệu từ nền tảng này có thể cung cấp cái nhìn sâu sắc về sự đa dạng trong hành vi người chơi và sự thay đổi theo thời gian.

3. Thách Thức Xử Lý Dữ Liệu Lớn: Với số lượng người chơi lớn và dữ liệu ngày càng tăng lên, việc xử lý dữ liệu lớn từ Steam đặt ra những thách thức thú vị trong việc phân tích và hiểu rõ hành vi người chơi.

4. Ứng Dụng Thực Tế: Kết quả từ nghiên cứu có thể được ứng dụng trong việc cải thiện trải nghiệm người chơi, phát triển game thông minh hơn, và thậm chí cung cấp thông tin hữu ích cho các nhà phát triển và nhà phát hành game.

Chọn đề tài này không chỉ mang lại niềm vui trong việc khám phá ngành công nghiệp game mà còn đặt ra những thách thức đáng giá trong việc nghiên cứu và xử lý dữ liệu phức tạp.



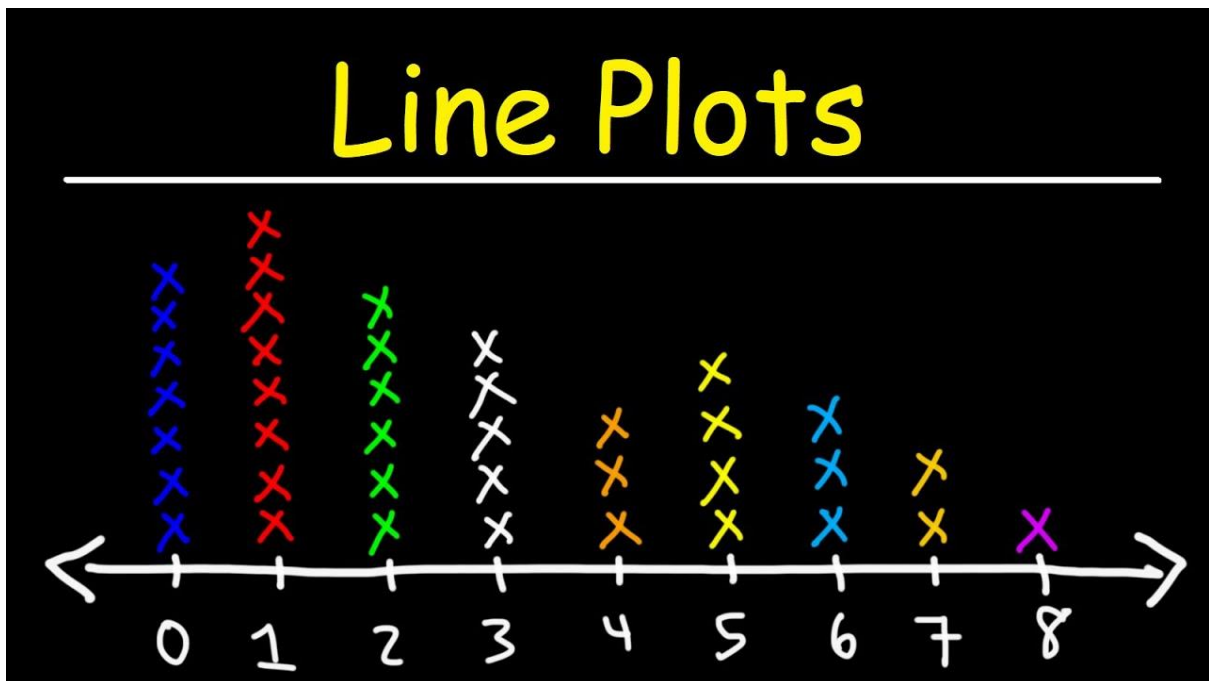
Chương III: Ứng dụng phương pháp vào bài phân tích thực tế

3.1. Phương pháp nghiên cứu:

3.1.1. Tổng quan về phương pháp:

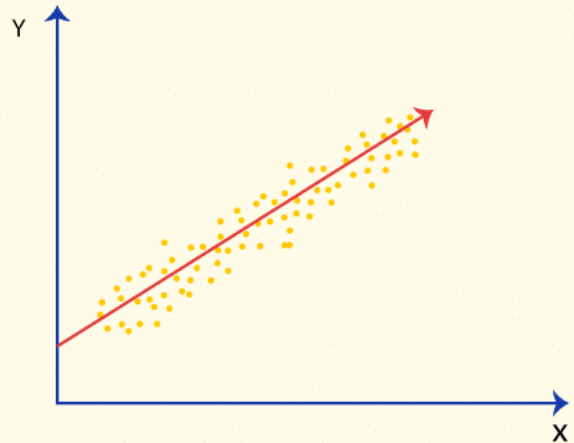
Để thực hiện bài phân tích, chúng tôi áp dụng một sự kết hợp linh hoạt giữa các phương pháp khoa học dữ liệu. Dưới đây là một số phương pháp chính:

- Phân tích thống kê: Áp dụng các kiểm định thống kê để đánh giá sự khác biệt giữa các nhóm dữ liệu và xác nhận tính độc lập giữa các biến.



- Biểu đồ và đồ thị: Sử dụng biểu đồ dạng line plot, scatter plot, và bar plot để trực quan hóa xu hướng và mối liên quan giữa các biến.

Linear Regression



- Phương pháp học máy đơn giản: Áp dụng mô hình học máy đơn giản như Linear Regression để dự đoán mối liên quan giữa mua và thời gian chơi.

3.1.2. Lựa chọn phương pháp cho đề tài:

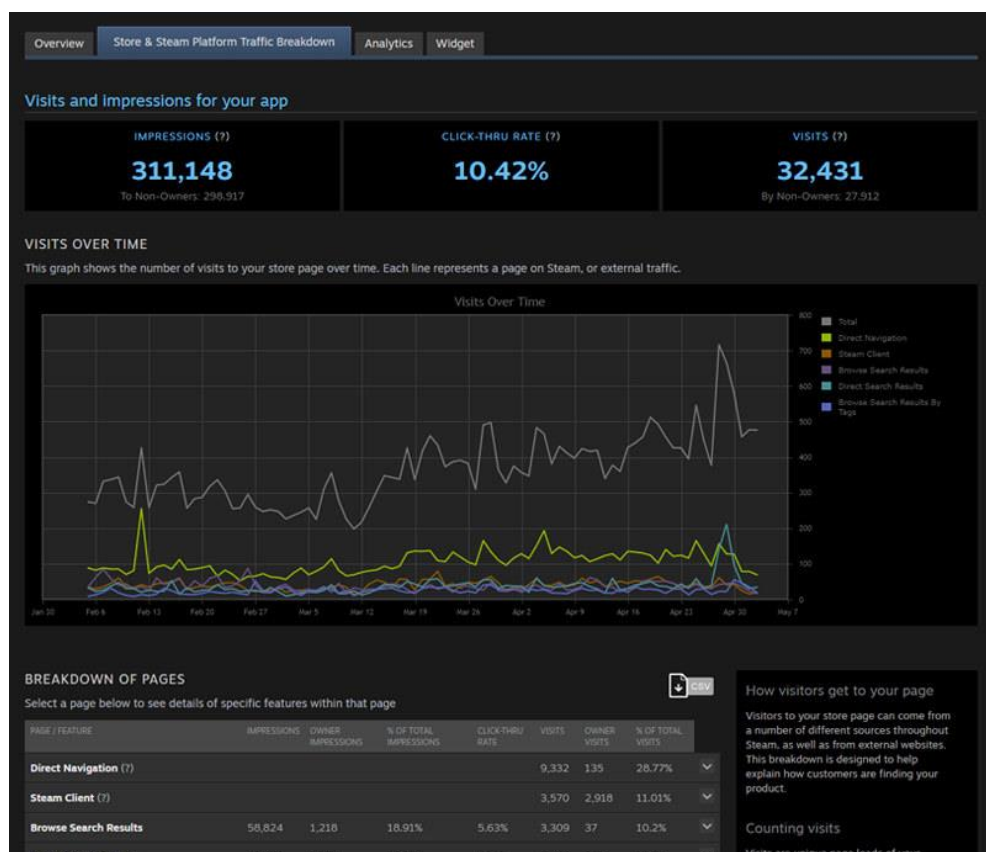
Với đối tượng nghiên cứu là ngành công nghiệp game và hành vi người chơi, chúng tôi chọn phương pháp nghiên cứu đa dạng để có cái nhìn toàn diện về dữ liệu. Sự kết hợp giữa phân tích thống kê và học máy giúp chúng tôi hiểu sâu hơn về mối liên quan và dự đoán trong ngữ cảnh ngành công nghiệp động đang phát triển.

3.2. Thu thập dữ liệu:

3.2.1. Phương pháp thu thập dữ liệu:



Dữ liệu được thu thập từ nguồn đáng tin cậy trong ngành công nghiệp game, bao gồm thông tin về User ID, tựa game, hành vi người chơi, và thời gian chơi. Quá trình thu thập được thực hiện với sự đảm bảo về quyền riêng tư và tuân thủ các chuẩn mực ngành.



* Hình ảnh chỉ mang tính chất minh họa.

3.2.2. Xử lý và làm sạch dữ liệu:

	User ID	Game	Behavior	Hoursplayed	Unnamed: 4
0	151603712	The Elder Scrolls V Skyrim	purchase	1.0	0
1	151603712	The Elder Scrolls V Skyrim	play	273.0	0
2	151603712	Fallout 4	purchase	1.0	0
3	151603712	Fallout 4	play	87.0	0
4	151603712	Spore	purchase	1.0	0
...
199995	128470551	Titan Souls	play	1.5	0
199996	128470551	Grand Theft Auto Vice City	purchase	1.0	0
199997	128470551	Grand Theft Auto Vice City	play	1.5	0
199998	128470551	RUSH	purchase	1.0	0
199999	128470551	RUSH	play	1.4	0

200000 rows × 5 columns

* *Tổng quan giá trị có trong dữ liệu*

*Dữ liệu bao gồm:

User ID (*int64*): Số nguyên đại diện cho ID của người chơi.

Game (*object*): Chuỗi ký tự chứa tên của trò chơi mà người chơi đã tham gia.

Behavior (*object*): Chuỗi ký tự chứa thông tin về hành vi của người chơi. (Purchase/Play)

Hoursplayed (*float64*): Số thực biểu thị số giờ mà người chơi đã chơi trò chơi.

Unnamed: 4 (*int64*): Một cột số nguyên bất định.

Trước khi phân tích, ta tiến hành xử lý và làm sạch dữ liệu. Bao gồm loại bỏ giá trị thiếu, kiểm tra và xử lý ngoại lệ, và chuyển đổi định dạng dữ liệu để phù hợp với mô hình phân tích.

-Xóa cột không mong muốn:

#INPUT:

```
# Remove unnecessary or erroneous columns, and print
the updated data table.
data = data.drop('Unnamed: 4', axis=1)
print(data)
```

#OUTPUT:

```

      User ID      Game  Behavior  Hoursplayed
0    151603712  The Elder Scrolls V Skyrim  purchase      1.0
1    151603712  The Elder Scrolls V Skyrim    play    273.0
2    151603712      Fallout 4  purchase      1.0
3    151603712      Fallout 4    play    87.0
4    151603712      Spore  purchase      1.0
...         ...         ...         ...
199995  128470551      Titan Souls    play      1.5
199996  128470551  Grand Theft Auto Vice City  purchase      1.0
199997  128470551  Grand Theft Auto Vice City    play      1.5
199998  128470551      RUSH  purchase      1.0
199999  128470551      RUSH    play      1.4
[200000 rows x 4 columns]
```

-Kiểm tra dữ liệu bị thiếu:

#INPUT:

```
# Checking
missing_values = data.isnull().sum()
missing_values
```

#OUTPUT:

```
User ID      0
Game        0
Behavior     0
Hoursplayed  0
dtype: int64
```

➤ Không có giá trị thiếu trong bộ dữ liệu, vì vậy chúng ta có thể tiếp tục phân tích mà không cần phải điền giá trị thay thế.

-Đếm số giá trị có trong DataSet:

#INPUT:

```
# Count the occurrences of each game (Sort alphabetically)
total_game =
data.groupby('Game')['Game'].agg('count')
total_game
```

#OUTPUT:

```
Game
007 Legends                2
ORBITALIS                  6
1... 2... 3... KICK IT! (Drop That Beat Like an Ugly Baby)  12
10 Second Ninja            8
10,000,000                 2
...
sZone-Online              160
samurai_jazz              1
the static speaks my name  21
theHunter                 372
theHunter Primal          8
Name: Game, Length: 5155, dtype: int64
```

#INPUT:

```
# Count the number of Play and Purchase.
total_purchase =
data.groupby('Behavior')['Behavior'].agg('count')
total_purchase
```

#OUTPUT:

```
Behavior
play      70489
purchase  129511
Name: Behavior, dtype: int64
```

- Các kết quả trên cho thấy 5155 tựa game trong dữ liệu có bao nhiêu người chơi, và hành vi giữa việc Chơi và Mua.

3.3. Phân tích dữ liệu:

3.3.1. Biểu đồ và đồ thị thể hiện kết quả phân tích:

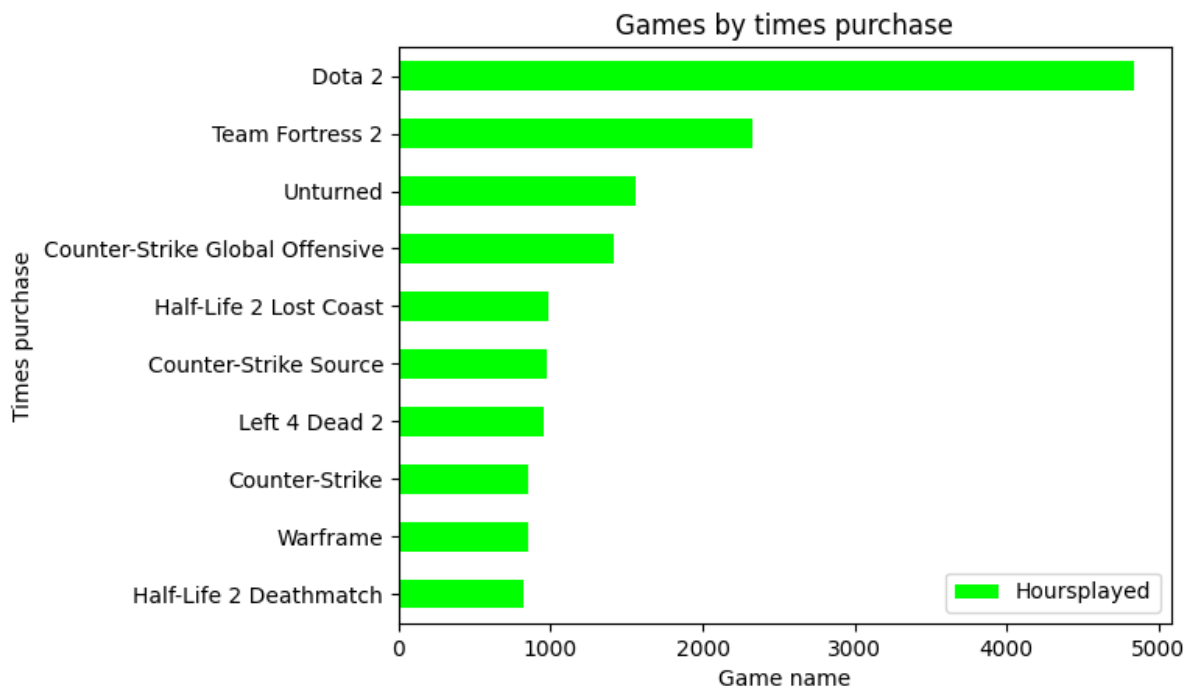
- Top 10 trò chơi được mua / cài đặt nhiều nhất là gì?

#INPUT:

```
data[data['Behavior']] ==
'purchase'][['Game', 'Hoursplayed']].groupby('Ga
me').sum().sort_values(by = 'Hoursplayed',
ascending = True).tail(10).plot(kind = 'barh',
cmap = 'brg_r', grid = False)

plt.title('Games by times purchase')
plt.xlabel('Game name')
plt.ylabel('Times purchase');
```

#OUTPUT:



✚ Như chúng ta có thể thấy trong mẫu này, các trò chơi như **Dota 2**, **TF2**, **Unturned** và **CS:GO** đang dẫn đầu. Nhưng những trò chơi này chủ yếu là miễn phí, điều này là do mẫu đếm số lần một trò chơi được tải về.

-Top 10 trò chơi được chơi nhiều nhất trong mẫu này là gì?
(Theo số giờ chơi)

#INPUT:

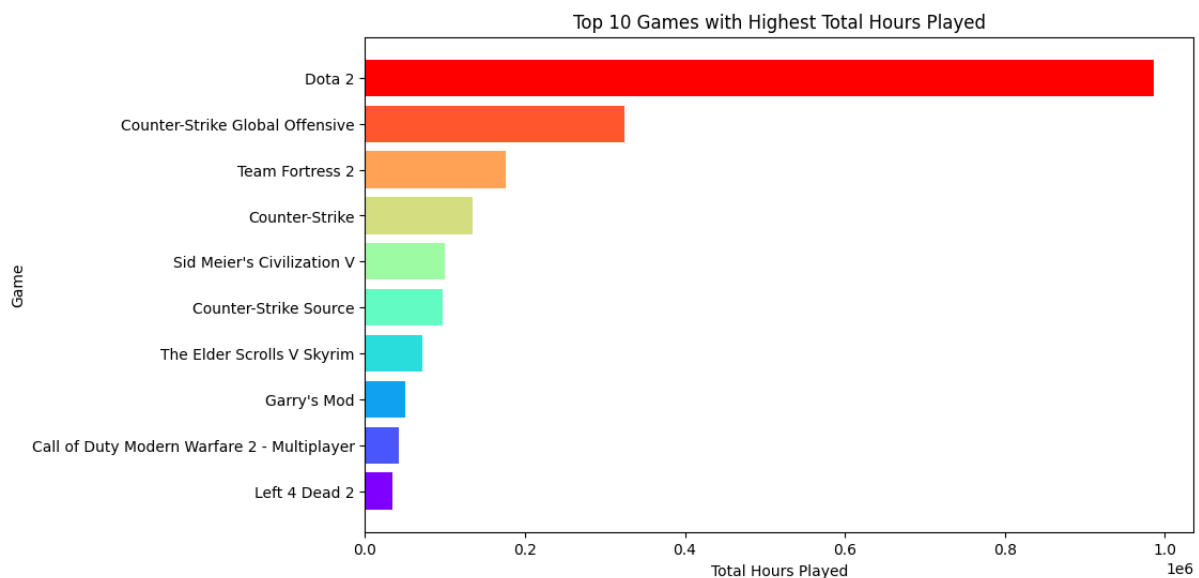
```
sorted_data = data.groupby('Game')['Hoursplayed'].sum().sort_values(ascending=True)

# Chose top 10 games.
top_games = sorted_data.tail(10)

colors = plt.cm.rainbow(np.linspace(0, 1, len(top_games)))
```

```
# Draw a horizontal bar chart.
plt.figure(figsize=(10, 6))
plt.barh(top_games.index, top_games.values,
color=colors)
plt.title('Top 10 Games with Highest Total Hours
Played')
plt.xlabel('Total Hours Played')
plt.ylabel('Game')
plt.show()
```

#OUTPUT:



✚ Như chúng ta có thể thấy, trò chơi **Dota 2** chiếm một lượng thời gian chơi đáng kể. Là một trò chơi MOBA, việc chơi nhiều trận đấu liên tục góp phần tạo ra sự chênh lệch đáng kể trong thời gian chơi so với các trò chơi khác và nó gấp ba lần so với trò chơi được chơi nhiều thứ hai (**CS:GO**).

-Top 10 trò chơi có số lượng người chơi cao nhất là gì? (Theo Số người chơi)

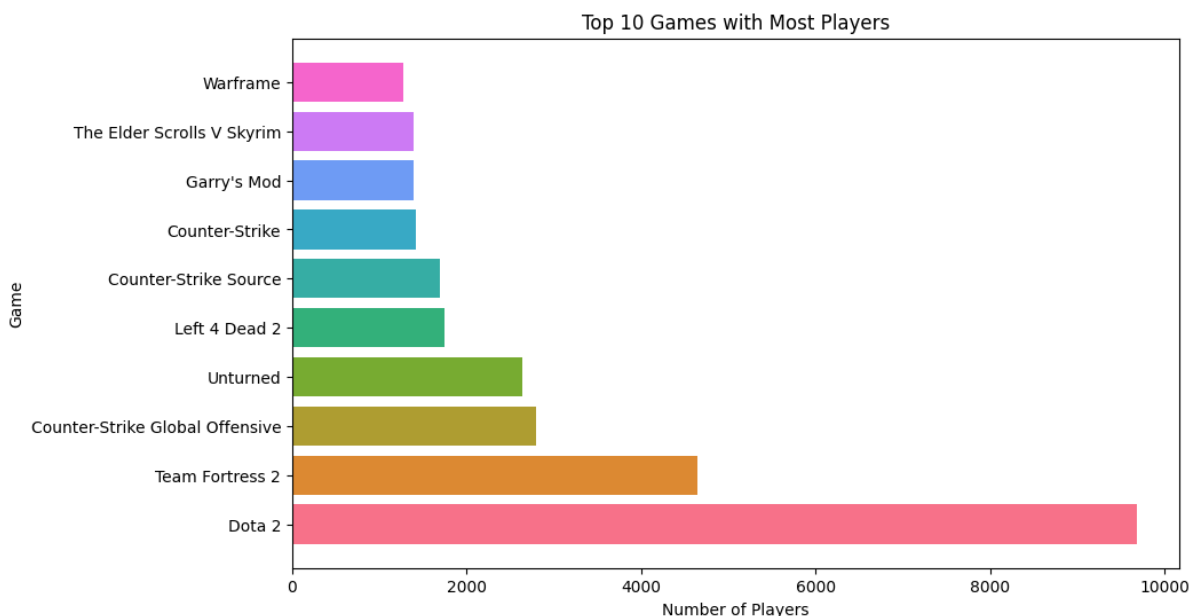
#INPUT:

```
top_games = data['Game'].value_counts().nlargest(10)

rainbow_palette = sns.color_palette("husl", 10)
plt.figure(figsize=(10, 6))
bars = plt.barh(top_games.index, top_games.values, color=rainbow_palette)

plt.title('Top 10 Games with Most Players')
plt.xlabel('Number of Players')
plt.ylabel('Game')
plt.show()
```

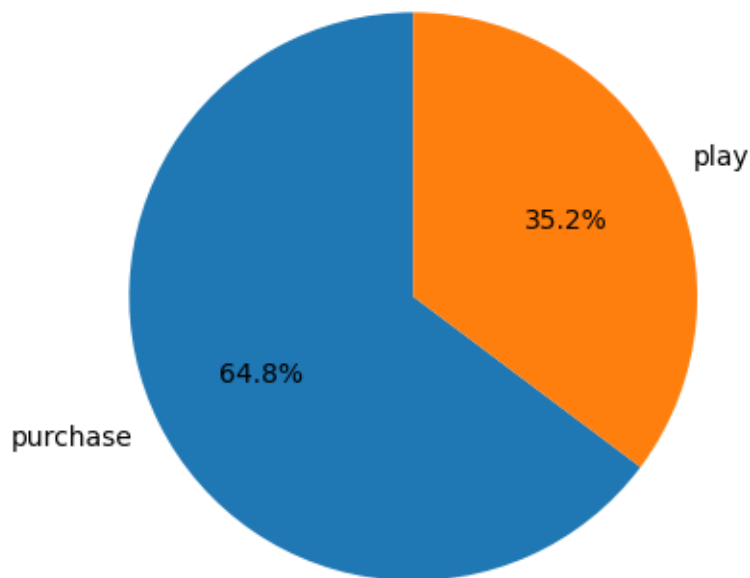
#OUTPUT:



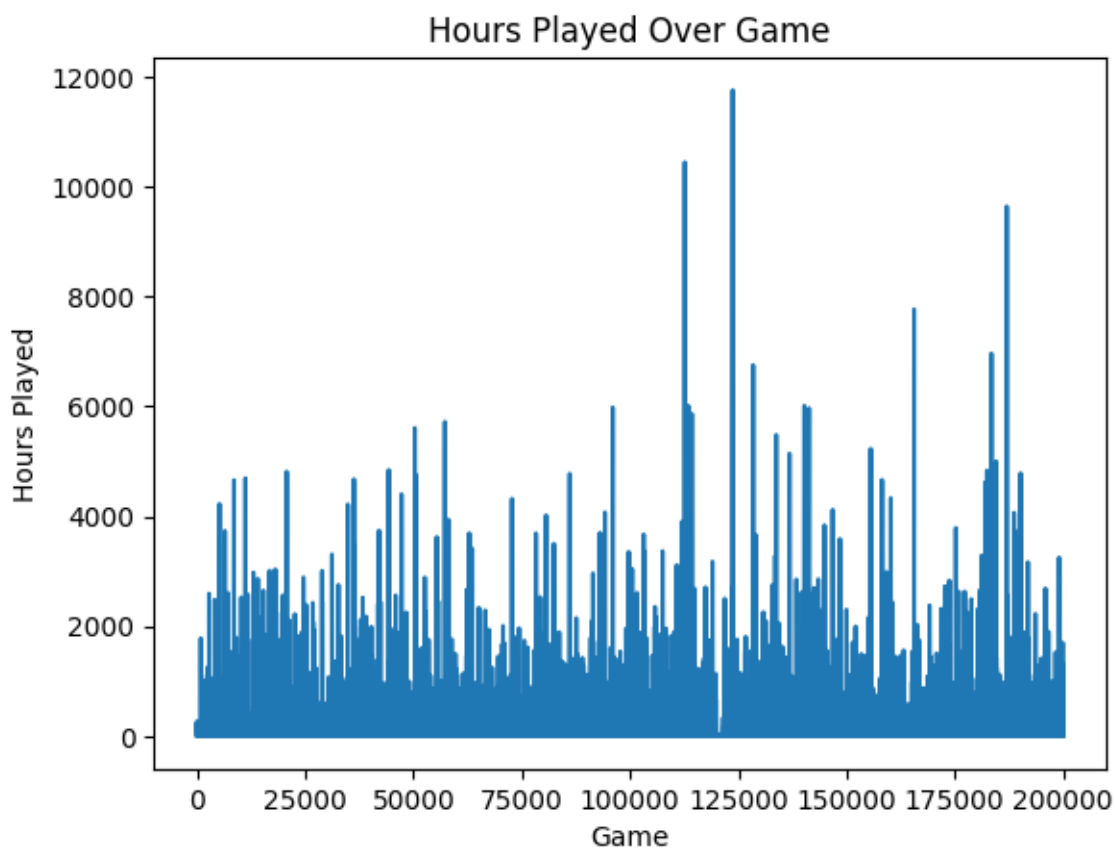
Ở đây, chúng ta có thể thấy rằng trò chơi hàng đầu là DOTA 2 với tổng cộng hơn 9000 người chơi, tiếp theo là Team Fortress 2 và CSGO. Top 3 trò chơi đều là những trò chơi **Esport** trực tuyến. Do độ cạnh tranh cao trong những trò chơi này, luôn có người chơi liên tục chơi vào bất kỳ thời điểm nào.

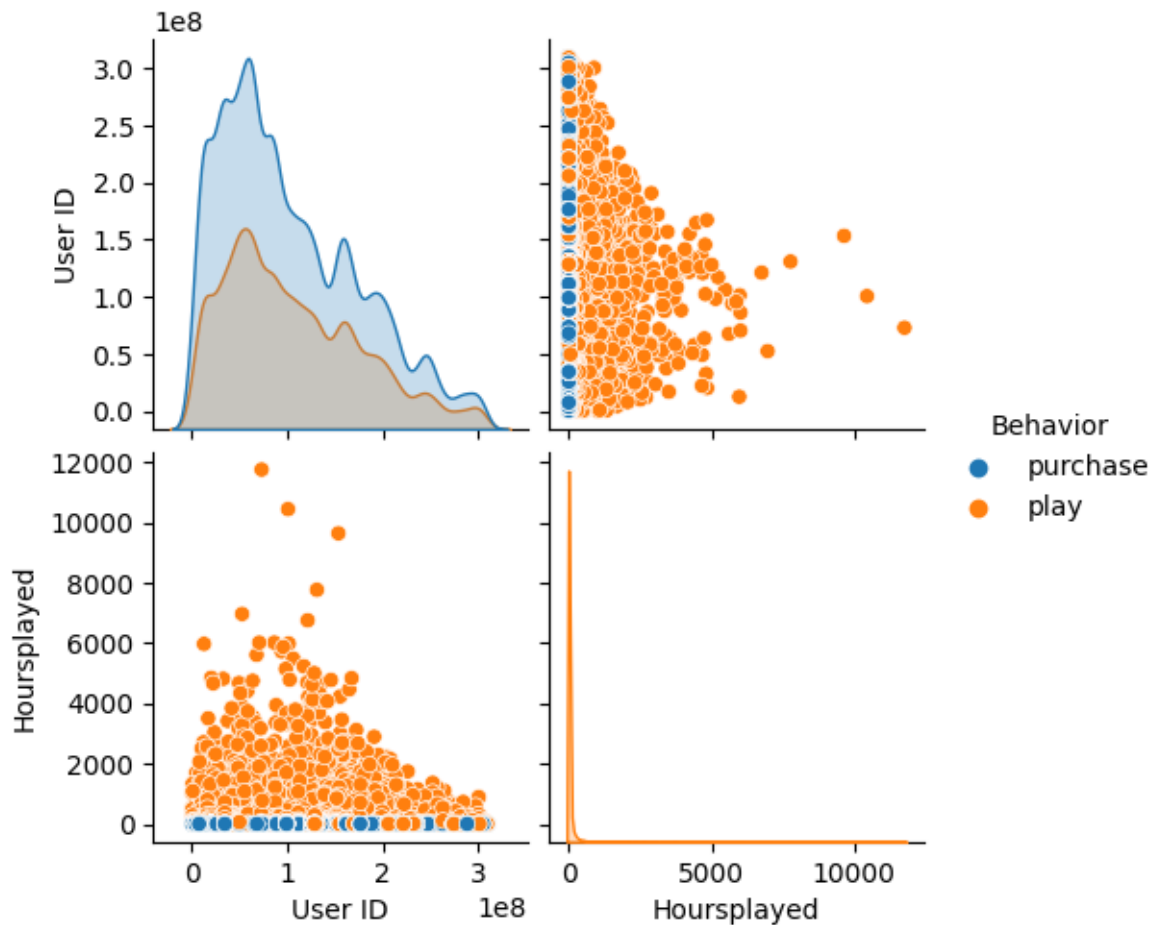
-Một số biểu đồ khác:

Distribution of Players in Each Behavior Category



Phần trăm đó chỉ ra rằng hành vi **Purchase** của người dùng cao hơn so với phần trăm **Play**, ngụ ý rằng nhiều người mua trò chơi nhưng chơi thì hiếm khi (*Gần như bằng không*).





3.3.2. Đưa ra đề xuất và nhận xét:

- ✚ Dựa trên những phân tích chi tiết về hành vi người chơi, ta đưa ra những đề xuất cụ thể để hỗ trợ quá trình chọn lựa tựa game.

3.3.2.1. Phân Khúc Đối Tượng Người Chơi:

***Người Chơi Nhiều Giờ:**

- **Đề Xuất:** Tựa game mới với thể giới mở rộng, nhiều nhiệm vụ phức tạp và tính năng độc đáo để kích thích sự hiếu kỳ và giữ chân người chơi lâu dài.



*Người Chơi Mua Nhiều:

- **Đề Xuất:** Gói mở rộng đặc biệt cho người chơi thường xuyên, cung cấp nhiều tính năng mới và quà tặng giá trị.



*Người Chơi Xã Hội:

- **Đề Xuất:** Tựa game đa người chơi (Coop, Mutiplayer...), có tính năng giao lưu, sự kiện cộng đồng, và khuyến khích chia sẻ trải nghiệm trên các nền tảng xã hội.



3.3.2.2. Yếu Tố Quyết Định Mua và Thời Gian Chơi:

***Mối Liên Quan giữa Mua và Thời Gian Chơi:**

- **Đề Xuất:** Tăng cường chiến lược tiếp thị để kích thích việc mua, đồng thời cập nhật thường xuyên nội dung mới để duy trì sự hứng thú.



***Ưu Đãi Đặc Biệt cho Người Chơi Thường Xuyên:**

- **Đề Xuất:** Tạo các ưu đãi đặc biệt, giảm giá, và quà tặng cho người chơi thường xuyên để khuyến khích họ duy trì hành vi mua.



3.3.2.3. Tổng Hợp Ý Kiến Đánh Giá:

***Tích Hợp Ý Kiến Đánh Giá:**

- **Đề Xuất:** Nâng cao trải nghiệm người chơi bằng cách tích hợp ý kiến đánh giá tích cực vào cửa hàng trực tuyến. Đặc biệt chú ý đến các tựa game được đánh giá cao.



***Xem Xét Phản Hồi Người Chơi:**

- **Đề Xuất:** Thường xuyên theo dõi và phản hồi với cộng đồng người chơi, điều này giúp cập nhật nhanh chóng các vấn đề và cải thiện sản phẩm.

3.3.2.4. Kết Hợp Các Yếu Tố:

***Tổng Hợp Các Yếu Tố:**

- **Đề Xuất:** Phát triển tựa game đa dạng, kết hợp nhiều yếu tố để phục vụ đa dạng nhu cầu của cộng đồng người chơi. Ví dụ: tựa game thú vị, nhiều tính năng mở rộng, và tương tác xã hội.



➤ Kết Luận những Đề Xuất:

Dựa trên các phân tích trên, đề xuất việc phát triển và quảng bá các tựa game có các yếu tố phù hợp với từng phân khúc người chơi. Chiến lược này giúp tối ưu hóa cơ hội thu hút và giữ chân người chơi trong môi trường cạnh tranh của ngành công nghiệp game.

LỜI CẢM ƠN

Mình xin bày tỏ lòng biết ơn sâu sắc đến những người đã hỗ trợ và đóng góp vào việc hoàn thành bài luận này. Công việc này không thể thực hiện được mà không có sự giúp đỡ, hướng dẫn và động viên từ gia đình, bạn bè và những người có kinh nghiệm.

Đặc biệt, xin gửi lời cảm ơn chân thành đến *Thầy Vũ Ngọc Thanh Sang* đã dành thời gian và tâm huyết để hướng dẫn mình qua từng bước của dự án này.

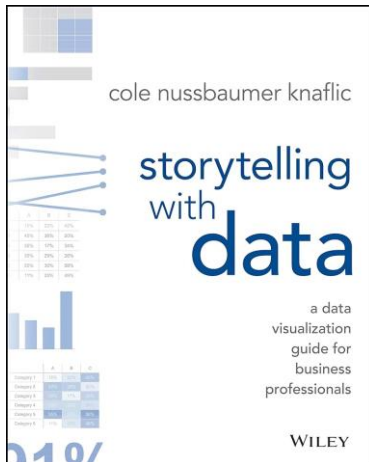
Cuối cùng, tôi gửi lời cảm ơn đến cộng đồng người làm khoa học dữ liệu và cộng đồng người chơi, nơi mình đã học hỏi được rất nhiều kiến thức và kinh nghiệm quý báu.

Trân trọng,

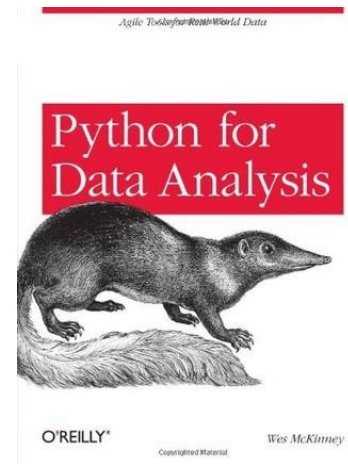
Nguyễn Trương Cao Sơn

10/12/2023

TÀI LIỆU THAM KHẢO

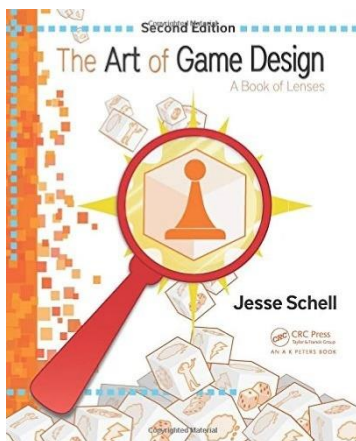


"Storytelling with Data" by Cole Nussbaumer Knaflc

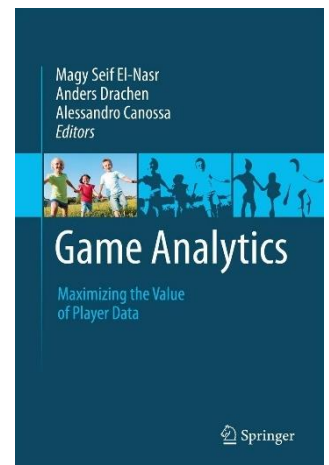


"Python for Data Analysis" by Wes McKinney

"Game Analytics: Maximizing the Value of Player Data" by Magy Seif El-Nasr, Anders Drachen, Alessandro Canossa



"The Art of Game Design: A Book of Lenses" by Jesse Schell



<https://dl.acm.org/doi/book/10.5555/559522>

<https://dl.acm.org/doi/10.1145/3340631.3398677>

<https://gameanalytics.com/blog/datasuite-player-warehouse-video/>