BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH

KHOA CÔNG NGHỆ THÔNG TIN

TIỂU LUẬN MÔN HỌC

CÔNG NGHỆ KHOA HỌC DỮ LIỆU

PHÂN TÍCH VÀ TRỰC QUAN HÓA DỮ LIỆU BÁN HÀNG SIÊU THỊ

Giảng viên hướng dẫn	**:**	**Th.S SỬ NHẬT HẠ**
Sinh viên thực hiện	**:**	**LÊ VÕ QUỐC HUY**
MSSV	**:**	**2000003954**
Lớp	**:**	**20DTH2A**
Ngành	**:**	**CÔNG NGHỆ THÔNG TIN**

Tp HCM, tháng 8 năm 2023

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH

KHOA CÔNG NGHỆ THÔNG TIN

TIỂU LUẬN MÔN HỌC

CÔNG NGHỆ KHOA HỌC DỮ LIỆU

PHÂN TÍCH VÀ TRỰC QUAN HÓA DỮ LIỆU BÁN HÀNG SIÊU THỊ

Giảng viên hướng dẫn	**:**	**Th.S SỬ NHẬT HẠ**
Sinh viên thực hiện	**:**	**LÊ VÕ QUỐC HUY**
MSSV	**:**	**2000003954**
Lớp	**:**	**20DTH2A**
Ngành	**:**	**CÔNG NGHỆ THÔNG TIN**

Tp HCM, tháng 8 năm 2023

LỜI CẢM ƠN

Trước hết, xin bày tỏ lòng biết ơn và sự tận tâm với Thầy Cô và người hướng dẫn trong suốt quá trình thực hiện Đồ Án Công Nghệ Khoa Học Dữ Liệu. Đặc biệt, chúng em muốn gửi lời cảm ơn chân thành tới Thầy Th.S Sử Nhật Hạ (ngành Công Nghệ Thông Tin, chuyên ngành Khoa Học Dữ Liệu, Trường Đại Học Nguyễn Tất Thành) vì sự hỗ trợ, chỉ dẫn, và định hướng nghiên cứu rất tận tâm từ Thầy. Thầy đã giúp chúng em chọn đề tài phù hợp, hướng dẫn chúng trong việc thực hiện và trình bày Đồ Án một cách chi tiết và cẩn thận.

Không chỉ riêng Thầy, mà chúng em còn có lòng biết ơn đối với tất cả Thầy Cô Giáo trong trường Đại học Nguyễn Tất Thành. Những kiến thức mà Thầy Cô đã chia sẻ và truyền đạt giúp chúng em có được nền tảng kiến thức vững chắc, chuẩn bị tốt cho cuộc hành trình trong tương lai.

Cuối cùng, bày tỏ lòng biết ơn sâu sắc tới các bạn cùng khóa K20 đã luôn ủng hộ và khuyến khích cùng nhau trong suốt thời gian học tập. Những người bạn này đã góp phần tạo nên môi trường học tập tích cực.

TP.HCM , ngày 22 tháng 08 năm 2023

Sinh viên

BM-ChT-11

TRƯỜNG ĐẠI HỌC NGUYỄN TẤT THÀNH
TRUNG TÂM KHẢO THÍ

KỲ THI KẾT THÚC HỌC PHẦN
HỌC KỲ …..… NĂM HỌC …….. - ……

PHIẾU CHẤM THI TIỂU LUẬN/BÁO CÁO

Môn thi: Công Nghệ Khoa Học Dữ Liệu.....Lớp học phần: 20DTH2A.....

Nhóm sinh viên thực hiện : .....

1.Nguyễn Minh Hoàng.....Tham gia đóng góp: 25%.....

2.Lê Võ Quốc Huy .....Tham gia đóng góp: 25%.....

3. Bùi Thị Thùy Trang.....Tham gia đóng góp: 25%.....

4. Võ Tuấn Kiệt.....Tham gia đóng góp: 25%.....

5.....Tham gia đóng góp:.....

6.....Tham gia đóng góp:.....

7.....Tham gia đóng góp:.....

8.....Tham gia đóng góp:.....

Ngày thi: 30/08/2023.....Phòng thi: .....

Đề tài tiểu luận/báo cáo của sinh viên : .....

.....

Phần đánh giá của giảng viên (căn cứ trên thang rubrics của môn học):

Tiêu chí (theo CĐR HP)	**Đánh giá của GV**	**Điểm tối đa**	**Điểm đạt được**
Cấu trúc của báo cáo	....		
Nội dung			
- Các nội dung thành phần	....		
- Lập luận	....		
- Kết luận	.....		
Trình bày	.....		
TỔNG ĐIỂM			

Giảng viên chấm thi
(ký, ghi rõ họ tên)

LỜI MỞ ĐẦU

Trong bối cảnh mạnh mẽ của cuộc cách mạng công nghiệp 4.0, sự phát triển của khoa học dữ liệu đã mở ra những cánh cửa mới về việc hiểu và tận dụng dữ liệu trong mọi khía cạnh cuộc sống. Trong lĩnh vực kinh doanh, đặc biệt là trong ngành bán lẻ và siêu thị, dữ liệu bán hàng đang trở thành một nguồn tài nguyên quý báu, định hình cách thức doanh nghiệp tương tác với khách hàng và phản ánh xu hướng tiêu dùng.

Đề tài 'Phân tích và Trực quan hóa Dữ liệu Bán hàng Siêu thị' đặt ra mục tiêu quan trọng là tìm hiểu cách mà khoa học dữ liệu có thể được áp dụng để nắm bắt, chuyển đổi và tận dụng dữ liệu bán hàng trong ngữ cảnh siêu thị. Phân tích dữ liệu đang trở thành công cụ quyết định quan trọng, giúp doanh nghiệp đào sâu vào những thông tin quan trọng như mức độ ưa thích của khách hàng, sự tương tác với các sản phẩm cụ thể, và thậm chí là dự đoán xu hướng tương lai. Từ những thông tin này, doanh nghiệp có thể điều chỉnh chiến lược kinh doanh, cải thiện trải nghiệm khách hàng và tối ưu hóa hoạt động để đáp ứng nhu cầu thị trường một cách hiệu quả.

Ngoài việc phân tích, việc trực quan hóa dữ liệu cũng đóng một vai trò quan trọng trong việc tạo nên sự thông tin một cách trực quan và dễ hiểu. Thông qua việc biểu đồ hóa, biểu đạt thông tin dưới dạng hình ảnh, doanh nghiệp có thể nhanh chóng nhận biết các mô hình và xu hướng tiềm ẩn trong dữ liệu. Điều này không chỉ giúp cho việc ra quyết định dựa trên dữ liệu trở nên hiệu quả hơn mà còn giúp doanh nghiệp truyền đạt thông tin đến đội ngũ quản lý và nhân viên một cách rõ ràng và nhanh chóng. Trong bối cảnh cạnh tranh khốc liệt của ngành bán lẻ, việc áp dụng khoa học dữ liệu để phân tích và trực quan hóa dữ liệu bán hàng siêu thị không chỉ là một xu hướng mà còn là một yếu tố cần thiết để tồn tại và phát triển. Khả năng hiểu rõ hơn về khách hàng, đáp ứng nhanh chóng các biến đổi thị trường và tối ưu hóa hoạt động kinh doanh là những lợi thế mà công nghệ này mang lại. Qua đề tài này, chúng ta sẽ đi sâu vào thế giới phức tạp của dữ liệu bán hàng siêu thị và khám phá cách mà khoa học dữ liệu có thể là chìa khóa đưa doanh nghiệp vượt qua những thách thức và đạt được sự thành công trong tương lai.

MỤC LỤC

DANH MỤC HÌNH

CHƯƠNG I: GIỚI THIỆU

1. Giới thiệu đề tài

Trong thời đại hiện nay, dữ liệu đóng vai trò quan trọng trong nhiều lĩnh vực và ngành công nghiệp. Đặc biệt, trong lĩnh vực kinh doanh, dữ liệu bán hàng từ các siêu thị trở thành một nguồn thông tin quý báu, giúp tiết lộ nhiều thông tin về thị trường, khách hàng và xu hướng tiêu dùng. Hiểu rõ hơn về cách khoa học dữ liệu có thể được áp dụng để phân tích và biểu đồ hóa dữ liệu bán hàng từ siêu thị là mục tiêu của đề tài "Phân Tích và Trực Quan Hóa Dữ Liệu Bán Hàng Siêu Thị".

Hình 1: Bán hàng siêu thị

2. Lý do chọn đề tài:

Lựa chọn đề tài xuất phát từ nhận thức về tầm quan trọng của dữ liệu trong quá trình ra quyết định kinh doanh. Trong bối cảnh cạnh tranh khốc liệt, việc hiểu sâu hơn về dữ liệu bán hàng có thể là yếu tố quyết định đến sự thành bại của một doanh nghiệp.

3. Mục tiêu của đề tài:

Mục tiêu của đề tài là nghiên cứu cách sử dụng khoa học dữ liệu để phân tích và biểu đồ hóa dữ liệu bán hàng từ siêu thị. Ta sẽ tìm hiểu cách xử lý dữ liệu bán

hàng, áp dụng các phương pháp phân tích để phát hiện xu hướng và biểu đồ hóa thông tin để hỗ trợ quá trình ra quyết định kinh doanh.

4. Phương pháp của đề tài

Để đạt được mục tiêu, đề tài sẽ sử dụng phương pháp nghiên cứu kết hợp cả khía cạnh định tính và định lượng. Dữ liệu bán hàng từ khách hàng thực tế sẽ được thu thập và các kỹ thuật phân tích dữ liệu sẽ được áp dụng để tìm hiểu hành vi mua sắm và tương tác của khách hàng.

5. Đối tượng và phạm vi nghiên cứu:

Đối tượng:

Dữ liệu bán hàng từ một siêu thị cụ thể tại các chi nhánh ở nước ngoài, mà chúng em sẽ cung cấp trong tiểu luận này.

Phạm vi nghiên cứu:

Nghiên cứu sẽ tập trung vào việc sử dụng công nghệ khoa học dữ liệu để phân tích và biểu đồ hóa dữ liệu bán hàng từ siêu thị, không đi quá sâu vào các khía cạnh kinh doanh khác của siêu thị.

6. Công nghệ áp dụng:

Các công nghệ chính sẽ được áp dụng bao gồm các phương pháp phân tích dữ liệu như phân tích thống kê, học máy và khai phá dữ liệu. Ngoài ra, việc sử dụng các công cụ biểu đồ một số biểu đồ phổ biến dễ quan sát sẽ giúp biểu đồ hóa thông tin một cách trực quan và dễ hiểu.

CHƯƠNG II: CƠ SỞ LÝ THUYẾT

1. Công Nghệ Khoa Học Dữ Liệu:

1.1 Tổng quan về Công Nghệ Khoa Học Dữ Liệu

Hình 2: Công Nghệ Khoa Học Dữ Liệu

Công nghệ Khoa học Dữ liệu là một lĩnh vực quan trọng trong ngành công nghệ hiện đại, tập trung vào việc thu thập, xử lý, phân tích và tạo ra thông tin từ dữ liệu. Nó kết hợp các kiến thức từ lĩnh vực khoa học dữ liệu, thống kê, máy học và các lĩnh vực liên quan khác nhau để khám phá thông tin từ dữ liệu và hỗ trợ quyết định.

Công nghệ Khoa học Dữ liệu có ứng dụng đa dạng và lan rộng trong nhiều lĩnh vực khác nhau. Dưới đây là một số ví dụ tiêu biểu:

Kinh doanh và Tài chính: Công nghệ Khoa học Dữ liệu được sử dụng để phân tích thị trường, dự đoán biến động giá cả, quản lý rủi ro tài chính và tối ưu hóa chuỗi cung ứng.

Y tế: Trong lĩnh vực y tế, dữ liệu từ bệnh nhân và thử nghiệm y tế giúp phát triển mô hình dự đoán bệnh, quản lý bệnh viện và hiểu rõ hơn về yếu tố di truyền.

Khoa học xã hội: Công nghệ Khoa học Dữ liệu hỗ trợ phân tích dữ liệu xã hội, dự đoán xu hướng xã hội và tạo mô hình mô phỏng tác động của chính sách xã hội.

Tiếp thị và Quảng cáo: Dữ liệu từ chiến dịch tiếp thị trực tuyến hỗ trợ tối ưu hóa quảng cáo, phân tích phản hồi khách hàng và phát triển chiến lược tiếp thị.

Giao thông và Đô thị thông minh: Dữ liệu từ hệ thống giao thông và cảm biến hỗ trợ tối ưu hóa luồng giao thông, dự đoán tình trạng giao thông và phát triển đô thị thông minh.

Nông nghiệp: Công nghệ Khoa học Dữ liệu được áp dụng trong quản lý nông nghiệp, từ dự đoán mùa vụ đến tối ưu hóa sử dụng tài nguyên như nước và phân bón.

Khám phá dược phẩm: Trong lĩnh vực nghiên cứu dược phẩm, dữ liệu về cấu trúc phân tử và thử nghiệm dược phẩm giúp phát triển các loại thuốc mới và dự đoán tác dụng phụ..

2. Định nghĩa về Công nghệ khoa học dữ liệu

2.1 Công nghệ khoa học dữ liệu là gì?

Công nghệ Khoa học Dữ liệu là phương pháp và hệ thống sử dụng các công cụ và kỹ thuật để hiểu và phân tích dữ liệu, nhằm tạo ra thông tin hữu ích và hỗ trợ trong quyết định. Quá trình này bao gồm thu thập, lưu trữ, xử lý và phân tích dữ liệu từ nhiều nguồn khác nhau, nhằm cung cấp cái nhìn tổng quan về một vấn đề hoặc tình huống cụ thể.

Trong thực tế, Công nghệ Khoa học Dữ liệu bắt đầu bằng việc thu thập dữ liệu từ nhiều nguồn khác nhau như cơ sở dữ liệu, tệp tin, cảm biến và trang web. Dữ liệu này thường không được cấu trúc sẵn, và nhiều lần cần được làm sạch, biến đổi và chuyển đổi thành định dạng có thể xử lý. Sau đó, các kỹ thuật và phương pháp của khoa học dữ liệu được áp dụng để phân tích dữ liệu, tìm ra mẫu, xây dựng mô hình dự đoán và trích xuất thông tin quan trọng.

Phân tích dữ liệu có thể sử dụng nhiều công cụ và kỹ thuật khác nhau như:

Khai phá dữ liệu (Data Mining): Khám phá thông tin ẩn trong dữ liệu bằng cách áp dụng các thuật toán máy học và kỹ thuật khai phá mẫu.

Học máy (Machine Learning): Xây dựng các mô hình dự đoán và phân loại dữ liệu dựa trên các thuật toán máy học.

Xử lý ngôn ngữ tự nhiên (NLP): Phân tích và xử lý dữ liệu dựa trên ngôn ngữ tự nhiên để trích xuất thông tin từ văn bản.

Thống kê: Áp dụng các phương pháp thống kê để kiểm tra giả thuyết và đưa ra kết luận dựa trên dữ liệu.

2.2 Ưu điểm & nhược điểm của công nghệ khoa học dữ liệu

Ưu điểm:

Công nghệ Khoa học Dữ liệu mang lại nhiều ưu điểm quan trọng, làm cho nó trở thành một công cụ mạnh mẽ trong nhiều lĩnh vực:

Tạo ra thông tin giá trị: Công nghệ Khoa học Dữ liệu giúp tạo ra thông tin có giá trị từ dữ liệu không có hướng dẫn cụ thể. Nhờ vào việc áp dụng các kỹ thuật phân tích và mô hình hóa, nó giúp khám phá mẫu và hiểu rõ hơn về dữ liệu.

Dự đoán và dự báo: Công nghệ này cho phép tạo ra các mô hình dự đoán dựa trên dữ liệu lịch sử, giúp dự đoán xu hướng tương lai và đưa ra quyết định một cách thông minh.

Tối ưu hóa quyết định: Công nghệ Khoa học Dữ liệu cung cấp thông tin hỗ trợ quyết định dựa trên dữ liệu thay vì dựa trên cảm tính. Điều này giúp tối ưu hóa quyết định và đảm bảo hiệu suất tốt hơn.

Tìm hiểu khách hàng: Trong lĩnh vực tiếp thị, Công nghệ Khoa học Dữ liệu cho phép phân tích dữ liệu về khách hàng và hiểu rõ hơn về hành vi mua sắm, sở thích và nhu cầu của họ.

Hiểu rõ môi trường kinh doanh: Công nghệ này giúp doanh nghiệp hiểu rõ hơn về môi trường kinh doanh thông qua việc phân tích dữ liệu thị trường, cạnh tranh và xu hướng ngành.

Nhược điểm:

Bên cạnh những ưu điểm, Công nghệ Khoa học Dữ liệu cũng đối mặt với một số nhược điểm:

Dữ liệu không đầy đủ: Quá trình phân tích và trực quan hóa dữ liệu yêu cầu dữ liệu đầy đủ và chất lượng. Nếu dữ liệu không đủ hoặc không chính xác, kết quả có thể bị sai lệch.

Khả năng hiểu lầm: Khi sử dụng các mô hình phức tạp và thuật toán phân tích, có thể xảy ra hiểu lầm về cách hoạt động của chúng. Điều này có thể dẫn đến việc áp dụng sai hoặc đưa ra quyết định sai lầm.

Phụ thuộc vào nguồn dữ liệu: Kết quả của quá trình phân tích và trực quan hóa dữ liệu phụ thuộc mạnh vào chất lượng và tính đúng đắn của nguồn dữ liệu. Nếu nguồn dữ liệu không đảm bảo, kết quả sẽ không chính xác.

Phức tạp trong triển khai: Triển khai các quá trình phân tích và trực quan hóa dữ liệu có thể phức tạp, đòi hỏi kiến thức sâu rộng về khoa học dữ liệu và công nghệ thông tin.

Bảo mật và quyền riêng tư: Xử lý và phân tích dữ liệu có thể đặt ra vấn đề về bảo mật và quyền riêng tư, đặc biệt là khi dữ liệu liên quan đến thông tin cá nhân.

Tóm lại, Công nghệ Khoa học Dữ liệu là một công cụ mạnh mẽ với khả năng mang lại nhiều lợi ích cho các tổ chức và doanh nghiệp. Tuy nhiên, để đảm bảo sự thành công, cần phải đối mặt và giải quyết những thách thức và nhược điểm của nó một cách cẩn thận và thông minh.

2.3 Áp dụng Công nghệ Khoa học Dữ liệu trong Bán hàng Siêu thị

Công nghệ Khoa học Dữ liệu có thể được áp dụng một cách hiệu quả trong việc phân tích và tối ưu hóa quá trình bán hàng trong siêu thị. Dưới đây là một số cách mà công nghệ này có thể được sử dụng:

Phân tích hành vi mua sắm: Công nghệ Khoa học Dữ liệu có thể phân tích dữ liệu về hành vi mua sắm của khách hàng, từ việc lựa chọn sản phẩm cho đến thời gian mua sắm và tần suất mua hàng. Điều này giúp siêu thị hiểu rõ hơn về nhu cầu của khách hàng và tạo ra chiến lược kinh doanh phù hợp.

Dự đoán xu hướng sản phẩm: Dựa trên dữ liệu lịch sử về doanh số bán hàng và hành vi mua sắm, Công nghệ Khoa học Dữ liệu có thể dự đoán xu hướng sản phẩm. Điều này giúp siêu thị lập kế hoạch tồn kho và quản lý cung ứng một cách hiệu quả hơn.

Phân tích hiệu suất kệ hàng: Công nghệ này có thể phân tích hiệu suất của từng kệ hàng trong siêu thị bằng cách theo dõi doanh số bán hàng và thời gian mua sắm. Điều này giúp quyết định về việc tái bố trí kệ hàng và quản lý không gian trưng bày sản phẩm.

Tối ưu hóa giá cả: Công nghệ Khoa học Dữ liệu có thể phân tích dữ liệu về giá cả, khuyến mãi và chiến lược giá của siêu thị. Điều này giúp đưa ra quyết định về việc định giá sản phẩm và xác định các chương trình khuyến mãi hợp lý.

Phân tích khách hàng tiềm năng: Dựa trên dữ liệu từ các khách hàng hiện tại, Công nghệ Khoa học Dữ liệu có thể xác định các đặc điểm và hành vi chung của khách hàng tiềm năng. Điều này giúp siêu thị tạo ra chiến lược tiếp thị và quảng cáo nhắm đến nhóm khách hàng có tiềm năng cao.

Tối ưu hóa chuỗi cung ứng: Công nghệ này có thể phân tích dữ liệu về chuỗi cung ứng từ nhà cung cấp đến siêu thị và từ siêu thị đến khách hàng. Điều này giúp tối ưu hóa quá trình vận chuyển, lưu trữ và phân phối sản phẩm.

2.4 Tương lai của Công nghệ khoa học dữ liệu

Tích hợp với Trí tuệ nhân tạo (AI): Công nghệ Khoa học Dữ liệu dự kiến sẽ tích hợp mạnh mẽ với Trí tuệ nhân tạo để tạo ra các hệ thống thông minh có khả năng học tập và ra quyết định tự động dựa trên dữ liệu.

Mở rộng vào lĩnh vực mới: Công nghệ này dự kiến sẽ mở rộng vào nhiều lĩnh vực mới như ngành y dược, năng lượng tái tạo, du lịch và giải trí, giúp tối ưu hóa hoạt động và cải thiện trải nghiệm khách hàng.

Dữ liệu thời gian thực: Xu hướng sử dụng dữ liệu thời gian thực trong phân tích dự kiến sẽ tăng cao, giúp doanh nghiệp có khả năng phản ứng nhanh chóng với các sự kiện và biến đổi trong thời gian thực.

Quản lý dữ liệu lớn và đa dạng: Với sự gia tăng về lượng dữ liệu và đa dạng dạng dữ liệu, Công nghệ Khoa học Dữ liệu sẽ phải đối mặt với thách thức quản lý và xử lý dữ liệu lớn, không chỉ dữ liệu cấu trúc mà còn dữ liệu phi cấu trúc như hình ảnh, âm thanh và văn bản.

An ninh dữ liệu: Vấn đề về bảo mật và quyền riêng tư dữ liệu sẽ tiếp tục là một thách thức quan trọng, đòi hỏi sự phát triển của các giải pháp bảo mật và chính sách quản lý dữ liệu.

Phát triển dịch vụ dựa trên dữ liệu: Các doanh nghiệp dự kiến sẽ phát triển nhiều dịch vụ mới dựa trên dữ liệu, từ dự đoán nhu cầu của khách hàng đến tư vấn sản phẩm và dịch vụ cá nhân hóa.

Giải quyết vấn đề xã hội: Công nghệ Khoa học Dữ liệu có thể được áp dụng để giải quyết các vấn đề xã hội như dự đoán đợt dịch bệnh, tối ưu hóa việc sử dụng tài nguyên và hỗ trợ quyết định chính trị.

Trong tương lai, Công nghệ Khoa học Dữ liệu sẽ tiếp tục phát triển và ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau, đem lại những lợi ích to lớn trong việc quản lý dữ liệu, tối ưu hóa quyết định và tạo ra giá trị cho xã hội. Tuy nhiên, việc thực hiện và quản lý công nghệ này cũng đặt ra nhiều thách thức cần được giải quyết một cách thận trọng và hiệu quả.

2.5 Kết luận:

Công nghệ Khoa học Dữ liệu là một lĩnh vực quan trọng trong ngành công nghệ hiện đại, có khả năng thu thập, xử lý, phân tích và tạo ra thông tin từ dữ liệu. Nó có ứng dụng đa dạng và lan rộng trong nhiều lĩnh vực như kinh doanh, y tế, khoa học xã hội, tiếp thị, giao thông, và nhiều lĩnh vực khác. Công nghệ này giúp tạo ra thông tin giá trị, dự đoán xu hướng, tối ưu hóa quyết định và hiểu rõ hơn về môi trường kinh doanh.

Tuy nhiên, để áp dụng Công nghệ Khoa học Dữ liệu một cách hiệu quả, cần đảm bảo tính chính xác và đầy đủ của dữ liệu, hiểu rõ về cách hoạt động của các mô hình phân tích, và giải quyết các vấn đề liên quan đến bảo mật và quyền riêng tư. Với khả năng tạo ra những thông tin hữu ích và hỗ trợ quyết định, Công nghệ Khoa học Dữ liệu chắc chắn sẽ tiếp tục đóng vai trò quan trọng trong tương lai.

3. Định nghĩa phân tích dữ liệu và trực quan hóa dữ liệu

3.1 Phân tích dữ liệu là gì?

Phân tích Dữ liệu là quá trình khám phá, kiểm tra và hiểu thông tin từ dữ liệu để tạo ra những hiểu biết có ý nghĩa. Trong ngữ cảnh của công nghệ Khoa học Dữ liệu, phân tích dữ liệu bao gồm việc áp dụng các phương pháp thống kê và thuật toán máy học để trích xuất thông tin, nhận biết mẫu, và hiểu rõ hơn về mối quan hệ trong dữ liệu.

3.2 Trực quan hóa dữ liệu là gì?

Trực quan hóa Dữ liệu là việc sử dụng biểu đồ, đồ thị và hình ảnh để hiển thị thông tin từ dữ liệu một cách trực quan và dễ hiểu. Trực quan hóa giúp hình dung thông tin phức tạp một cách rõ ràng và giúp người sử dụng dễ dàng nhận thức về mẫu, xu hướng và thông tin quan trọng.

3.3 Các bước trong phân tích dữ liệu & trực quan hóa dữ liệu:

3.3.1 Tiền xử lý dữ liệu:

Tiền xử lý dữ liệu là bước quan trọng để làm sạch và chuẩn bị dữ liệu trước khi thực hiện phân tích và trực quan hóa. Các bước tiền xử lý thường bao gồm:

Đọc dữ liệu từ nguồn.

Loại bỏ dữ liệu không hợp lệ hoặc thiếu bằng cách sử dụng hàm dropna().

Loại bỏ các dòng trùng lặp bằng hàm drop_duplicates().

Chuyển đổi kiểu dữ liệu cho các cột phù hợp.

Loại bỏ dữ liệu dư thừa, chuẩn hóa, hoặc biến đổi dữ liệu theo yêu cầu.

3.3.2 Phân tích dữ liệu:

Hình 3: Phân tích dữ liệu

Phân tích dữ liệu bao gồm việc áp dụng các kỹ thuật thống kê và máy học để khám phá thông tin ẩn sau dữ liệu. Các bước phân tích thường bao gồm:

Thống kê mô tả: Tổng hợp thông tin cơ bản về dữ liệu bằng các thống kê như trung bình, phương sai, tần số,...

Khai thác dữ liệu: Sử dụng các phương pháp khai phá dữ liệu để nhận biết mẫu, quy luật hoặc mối quan hệ trong dữ liệu.

Xây dựng mô hình: Áp dụng các thuật toán máy học để xây dựng các mô hình dự đoán, phân loại, gom cụm, hoặc học từ dữ liệu.

3.3.3 Trích xuất thông tin thời gian và thống kê:

Ta đã sử dụng thư viện pandas để trích xuất thông tin thời gian từ dữ liệu bán hàng. Chẳng hạn, Ta đã tạo biểu đồ cột thể hiện doanh thu hàng ngày và biểu đồ đường thể hiện xu hướng doanh thu theo từng ngày.

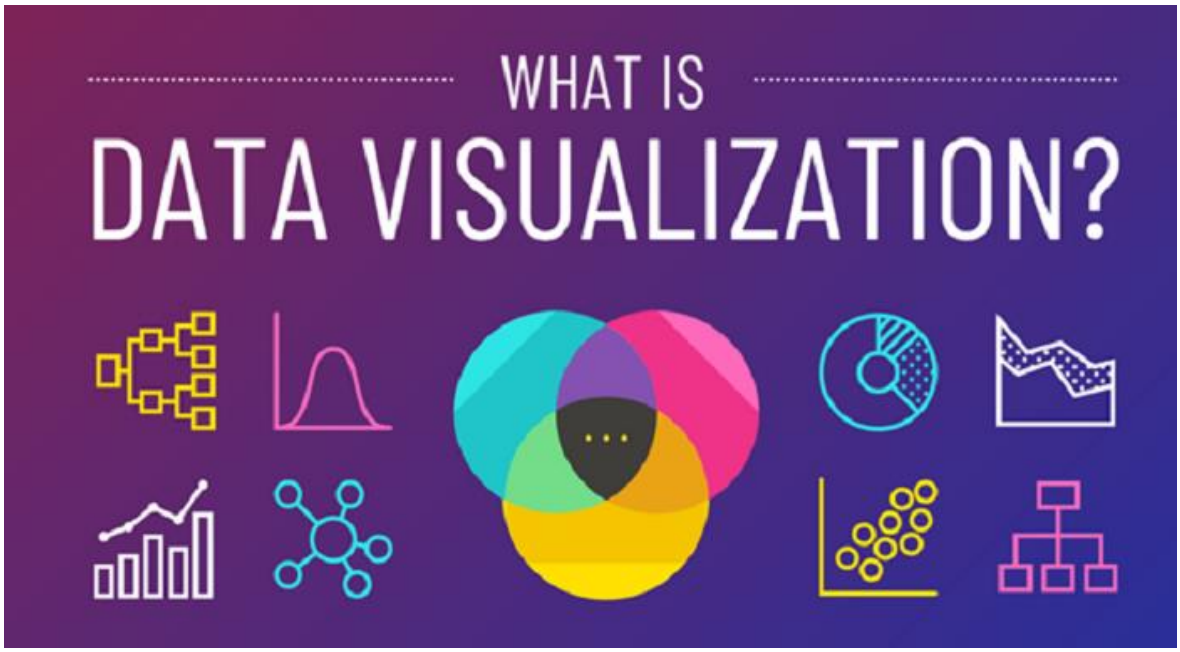
3.3.4 Trực quan hóa dữ liệu thời gian và thống kê hàng tuần, hàng tháng

Ta đã sử dụng phương pháp resample để tổng hợp dữ liệu hàng tuần và hàng tháng, sau đó tạo biểu đồ thể hiện doanh thu theo tuần và tháng.

3.3.5 Tính tổng doanh thu đạt được sau 90 ngày

Ta đã tính tổng doanh thu bằng cách tổng hợp cột "TotalSales" và sử dụng thư viện locale để định dạng tiền tệ.

3.3.6 Trực quan hóa dữ liệu:



Hình 4: Data Visualization

Trực quan hóa dữ liệu là việc biểu diễn thông tin từ dữ liệu bằng các biểu đồ và đồ thị. Mục tiêu của trực quan hóa là làm cho thông tin trở nên dễ hiểu và hấp dẫn. Các loại biểu đồ thường được sử dụng bao gồm:

Biểu đồ cột: Hiển thị dữ liệu dưới dạng các cột dọc hoặc ngang.

Biểu đồ đường: Biểu diễn mối quan hệ giữa các điểm dữ liệu trên một trục thời gian.

Biểu đồ hình tròn: Chia tỷ lệ phần trăm của các phần thành phần so với tổng thể.

Biểu đồ phân tán: Hiển thị mối quan hệ giữa hai biến số.

3.3.6.1 Lợi ích của việc trực quan hóa dữ liệu:

Trực quan hóa dữ liệu giúp tạo ra cái nhìn tổng quan về dữ liệu và tập trung vào các thông tin quan trọng. Điều này giúp hiểu rõ hơn về mẫu, xu hướng và biến đổi trong dữ liệu. Trực quan hóa còn giúp tạo ra những báo cáo dễ hiểu và hấp dẫn hơn.

3.3.6.2 Các loại biểu đồ trong trực quan hóa:

Có nhiều loại biểu đồ khác nhau để trực quan hóa dữ liệu như biểu đồ cột, biểu đồ đường, biểu đồ hình tròn và biểu đồ scatter. Mỗi loại biểu đồ có mục đích sử dụng khác nhau và thể hiện thông tin một cách rõ ràng.

4. Phân tích và trực quan hóa dữ liệu bán hàng siêu thị

Trong đề tài "Phân Tích và Trực Quan Hóa Dữ Liệu Bán Hàng Siêu Thị," ta sử dụng Công nghệ Khoa học Dữ liệu để khám phá thông tin từ dữ liệu bán hàng của một siêu thị. Quá trình này bao gồm các bước cơ bản từ việc tiền xử lý dữ liệu cho đến việc trực quan hóa thông tin, giúp ta hiểu rõ hơn về doanh số bán hàng, xu hướng mua sắm và sự phân bố của các sản phẩm.

4.1 Mô tả bài toán:

Bài toán này tập trung vào việc phân tích và trực quan hóa dữ liệu liên quan đến doanh số bán hàng tại một siêu thị. Mục tiêu là hiểu rõ hơn về các yếu tố ảnh hưởng đến doanh số bán hàng và tạo ra các biểu đồ thể hiện mô hình kinh doanh của siêu thị trong khoảng thời gian cụ thể.

4.2 Mô tả dữ liệu

Dữ liệu bán hàng được trích xuất từ tệp CSV "supermarket_sales.csv". Bộ dữ liệu này ghi chép thông tin về mỗi giao dịch bao gồm ngày bán hàng, chi nhánh, loại sản phẩm, số lượng bán, giá đơn vị và tổng doanh thu từ mỗi giao dịch.

4.3 Tiền xử lý dữ liệu

Trước khi bắt đầu phân tích, ta cần tiền xử lý dữ liệu để đảm bảo tính chính xác và độ tin cậy của kết quả. Điều này bao gồm việc đọc và kiểm tra dữ liệu, loại bỏ dữ liệu không hợp lệ hoặc trùng lặp, và chuyển đổi kiểu dữ liệu cho phù hợp với mục đích phân tích. Ta sử dụng các thư viện như Pandas và NumPy để thực hiện các bước tiền xử lý này qua các bước:

Bước 1: Đọc dữ liệu từ tệp CSV sử dụng thư viện Pandas.

Bước 2: Loại bỏ các dòng chứa dữ liệu thiếu (NaN) bằng hàm dropna().

Bước 3: Loại bỏ các dòng trùng lặp bằng hàm drop_duplicates().

Bước 4: Chuyển đổi kiểu dữ liệu của cột 'Date' thành dạng datetime.

Bước 5: Chuyển đổi kiểu dữ liệu của cột 'Quantity' thành kiểu số nguyên (downcast='integer').

Bước 6: Chuyển đổi kiểu dữ liệu của cột 'Unit price' thành kiểu số thực (downcast='float').

Bước 7: Loại bỏ khoảng trắng dư thừa trong cột 'Product line'.

Bước 8: Loại bỏ các giá trị trùng lặp trong cột 'Unit price'.

Bước 9: Tính toán doanh thu của mỗi giao dịch bằng cách nhân 'Quantity' và 'Unit price', và thêm cột 'TotalSales'.

4.4 Trực quan hóa thông tin

Mục tiêu chính của công việc phân tích là trực quan hóa thông tin để dễ dàng hiểu và tạo ra insights từ dữ liệu. Ta sử dụng các thư viện như Matplotlib và Seaborn để tạo ra các biểu đồ và biểu đồ thống kê để mô phỏng và thể hiện các thông tin quan trọng.

Đầu tiên, ta trực quan hóa doanh số bán hàng của từng chi nhánh bằng biểu đồ cột, giúp ta so sánh hiệu suất bán hàng giữa các chi nhánh.

Tiếp theo, ta xác định các sản phẩm bán chạy nhất dựa trên số lượng bán ra và doanh thu tương ứng. Để thể hiện điều này, ta sử dụng biểu đồ cột kết hợp với biểu đồ đường để thể hiện tình hình mua sắm của các sản phẩm này.

Ta cũng thực hiện trực quan hóa doanh thu hàng ngày, hàng tuần và hàng tháng bằng các biểu đồ cột và biểu đồ đường. Điều này giúp ta nhận biết các mô hình và xu hướng trong việc tiêu thụ hàng ngày của khách hàng.

Cuối cùng, ta tính tổng doanh thu đạt được sau 90 ngày để đánh giá hiệu suất kinh doanh trong một khoảng thời gian cụ thể.

4.5 Phương pháp và mô hình

Làm sạch dữ liệu: Đầu tiên, dữ liệu được kiểm tra để loại bỏ bất kỳ giá trị trống hoặc bản sao lặp lại nào bằng cách sử dụng phương thức dropna() và drop_duplicates(). Điều này giúp đảm bảo dữ liệu được đồng nhất và không chứa thông tin rác.

Chuyển đổi kiểu dữ liệu: Các cột 'Date', 'Quantity' và 'Unit price' được chuyển đổi thành các kiểu dữ liệu phù hợp. 'Date' được chuyển thành dạng ngày tháng để dễ dàng phân tích theo thời gian, còn 'Quantity' và 'Unit price' được chuyển thành dạng số để thực hiện tính toán.

Loại bỏ dữ liệu dư thừa: Cột 'Product line' và 'Unit price' được kiểm tra để loại bỏ giá trị dư thừa. Điều này đảm bảo rằng thông tin về sản phẩm và giá đơn vị là chính xác và không chứa sai sót.

Tính toán tổng doanh thu: Dựa trên số lượng và giá đơn vị, cột mới 'TotalSales' được tính toán để biểu thị tổng doanh thu từ mỗi giao dịch.

Tạo dữ liệu làm sạch: Sau khi làm sạch và xử lý, dữ liệu được lưu trữ trong tệp "cleaned_data.csv" để sử dụng cho các phân tích tiếp theo.

4.6 Kết quả thử nghiệm

Biểu đồ doanh số bán hàng theo chi nhánh: Sử dụng biểu đồ cột để thể hiện doanh số bán hàng của mỗi chi nhánh. Điều này giúp nhận biết sự phân phối của doanh số trong toàn bộ hệ thống.

Top sản phẩm bán chạy: Tạo biểu đồ kết hợp với cột và đường để thể hiện những sản phẩm bán chạy nhất dựa trên số lượng bán và doanh thu. Kết hợp giữa cả hai thông số giúp hiểu rõ hơn về mối liên hệ giữa chất lượng sản phẩm và doanh thu.

Biểu đồ doanh thu hàng ngày: Sử dụng biểu đồ cột và đường để mô tả doanh thu hàng ngày. Thông qua biểu đồ, ta có thể nhận thấy các biến đổi trong doanh thu theo thời gian.

Biểu đồ doanh thu hàng tuần và hàng tháng: Tạo biểu đồ đường và độ mờ để thể hiện mô hình doanh thu theo tuần và tháng. Điều này giúp xác định được các xu hướng dài hạn.

Tổng doanh thu sau 90 ngày: Bằng cách tính tổng toàn bộ doanh thu từ mọi giao dịch, ta có cái nhìn tổng quan về hiệu suất kinh doanh.

4.7 Điểm mạnh và hạn chế của phân tích

Trong quá trình phân tích dữ liệu và trực quan hóa, ta cần xem xét cả điểm mạnh và hạn chế của phương pháp đã sử dụng. Điểm mạnh có thể bao gồm khả năng hiển thị xu hướng, mô hình dữ liệu và hỗ trợ quyết định dựa trên thông tin. Tuy nhiên, cũng cần lưu ý rằng phân tích dữ liệu có thể bị ảnh hưởng bởi dữ liệu nhiễu hoặc thiếu sót, và việc hiểu sai hoặc sai lệch trong quá trình phân tích có thể dẫn đến quyết định không chính xác.

4.8 Tương lai của phân tích dữ liệu trong bán lẻ

Công nghệ khoa học dữ liệu và phân tích dữ liệu đang ngày càng trở nên quan trọng trong lĩnh vực bán lẻ. Với sự phát triển của dữ liệu số và khả năng tích hợp các công cụ thông minh, việc phân tích dữ liệu có thể giúp các doanh nghiệp hiểu rõ hơn về hành vi của khách hàng, dự đoán xu hướng thị trường và tối ưu hóa hoạt động kinh doanh.

4.9 Ưu và nhược điểm:

Ưu điểm:

Hiểu rõ hơn về doanh nghiệp: Phân tích và trực quan hóa dữ liệu giúp doanh nghiệp hiểu rõ hơn về các xu hướng, mô hình và hoạt động kinh doanh của mình.

Hỗ trợ quyết định: Dựa trên các thông tin phân tích, doanh nghiệp có thể đưa ra quyết định thông minh hơn, dự đoán xu hướng tương lai và tối ưu hóa chiến lược.

Phát hiện thông tin tiềm ẩn: Phân tích dữ liệu giúp phát hiện thông tin ẩn chưa được biết đến trước đây, giúp tạo ra cơ hội mới và mở rộng doanh nghiệp.

Hiệu suất tốt hơn: Trực quan hóa dữ liệu giúp hiển thị thông tin một cách trực quan và dễ hiểu, từ đó giúp các đội ngũ trong doanh nghiệp dễ dàng nắm bắt và đánh giá tình hình.

Tối ưu hóa quy trình kinh doanh: Phân tích dữ liệu có thể chỉ ra các vấn đề trong quy trình kinh doanh và đề xuất cách cải thiện, giúp tối ưu hóa hoạt động.

Nhược điểm:

Phụ thuộc vào chất lượng dữ liệu: Phân tích và trực quan hóa dữ liệu yêu cầu dữ liệu chất lượng và chính xác. Dữ liệu không đúng hoặc thiếu sót có thể dẫn đến kết quả không chính xác.

Phức tạp và tốn thời gian: Quá trình phân tích và trực quan hóa dữ liệu có thể phức tạp và tốn thời gian, đặc biệt khi xử lý dữ liệu lớn và áp dụng các phương pháp phức tạp.

Yêu cầu kiến thức chuyên sâu: Để thực hiện phân tích dữ liệu hiệu quả, người thực hiện cần phải có kiến thức về khoa học dữ liệu, thống kê, và các công cụ phân tích.

Khó khăn trong việc hiểu và giải thích: Các phương pháp phân tích phức tạp có thể khó hiểu và khó giải thích cho những người không có kiến thức về lĩnh vực này.

Rủi ro sai tưởng: Trong một số trường hợp, việc phân tích và trực quan hóa dữ liệu có thể dẫn đến những sai tưởng hoặc hiểu lầm về dữ liệu và xu hướng.

4.10 Nhận xét:

Phân tích và trực quan hóa dữ liệu bán hàng siêu thị đã giúp ta có cái nhìn tổng quan về hoạt động kinh doanh của siêu thị. Các biểu đồ và kết quả thử nghiệm cho thấy sự biến đổi và xu hướng của doanh thu theo thời gian, loại sản phẩm và chi nhánh. Điều này có thể hỗ trợ quyết định kinh doanh, kế hoạch tài chính và phân tích hiệu suất.

4.11 Kết luận:

Chương này đã tập trung vào mô tả và áp dụng các bước phân tích và trực quan hóa dữ liệu bán hàng siêu thị. Từ việc tiền xử lý dữ liệu cho đến việc tạo các biểu đồ thống kê, ta đã hiểu rõ hơn về bài toán ứng dụng và cách áp dụng các phương pháp cơ bản của công nghệ Khoa học Dữ liệu để nắm bắt thông tin quan trọng từ dữ liệu thô và hỗ trợ quyết định kinh doanh.

CHƯƠNG III: THỰC NGHIỆM

1. Tổng quan về bộ dữ liệu:

1.1 Mô tả bài toán

Bài toán trong thực nghiệm là phân tích dữ liệu bán hàng siêu thị để hiểu và trực quan hóa các khía cạnh về doanh số bán hàng, sản phẩm bán chạy, doanh thu hàng ngày, hàng tuần và hàng tháng.

1.2 Xây dựng bộ dữ liệu:

Dataset sử dụng là file **supermarket_sales - Sheet1.csv** được lấy từ trang Github.com (một trang web dataset nổi tiếng).

	Invoice ID	Branch	City	Customer	Gender	Product lir	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross mar	gross inco	Rating
2	750-67-84	A	Yangon	Member	Female	Health an	74.69	7	26.1415	548.9715	01-05-19	13:08	Ewallet	522.83	4.761905	26.1415	9.1
3	226-31-30	C	Naypyitaw	Normal	Female	Electronic	15.28	5	3.82	80.22	03-08-19	10:29	Cash	76.4	4.761905	3.82	9.6
4	631-41-31	A	Yangon	Normal	Male	Home and	46.33	7	16.2155	340.5255	03-03-19	13:23	Credit car	324.31	4.761905	16.2155	7.4
5	123-19-11	A	Yangon	Member	Male	Health an	58.22	8	23.288	489.048	1/27/2019	20:33	Ewallet	465.76	4.761905	23.288	8.4
6	373-73-79	A	Yangon	Normal	Male	Sports an	86.31	7	30.2085	634.3785	02-08-19	10:37	Ewallet	604.17	4.761905	30.2085	5.3
7	699-14-30	C	Naypyitaw	Normal	Male	Electronic	85.39	7	29.8865	627.6165	3/25/2019	18:30	Ewallet	597.73	4.761905	29.8865	4.1
8	355-53-59	A	Yangon	Member	Female	Electronic	68.84	6	20.652	433.692	2/25/2019	14:36	Ewallet	413.04	4.761905	20.652	5.8
9	315-22-56	C	Naypyitaw	Normal	Female	Home and	73.56	10	36.78	772.38	2/24/2019	11:38	Ewallet	735.6	4.761905	36.78	8
10	665-32-91	A	Yangon	Member	Female	Health an	36.26	2	3.626	76.146	01-10-19	17:15	Credit car	72.52	4.761905	3.626	7.2
11	692-92-55	B	Mandalay	Member	Female	Food and	54.84	3	8.226	172.746	2/20/2019	13:27	Credit car	164.52	4.761905	8.226	5.9
12	351-62-08	B	Mandalay	Member	Female	Fashion ac	14.48	4	2.896	60.816	02-06-19	18:07	Ewallet	57.92	4.761905	2.896	4.5
13	529-56-39	B	Mandalay	Member	Male	Electronic	25.51	4	5.102	107.142	03-09-19	17:03	Cash	102.04	4.761905	5.102	6.8
14	365-64-05	A	Yangon	Normal	Female	Electronic	46.95	5	11.7375	246.4875	02-12-19	10:25	Ewallet	234.75	4.761905	11.7375	7.1
15	252-56-26	A	Yangon	Normal	Male	Food and	43.19	10	21.595	453.495	02-07-19	16:48	Ewallet	431.9	4.761905	21.595	8.2
16	829-34-39	A	Yangon	Normal	Female	Health an	71.38	10	35.69	749.49	3/29/2019	19:21	Cash	713.8	4.761905	35.69	5.7
17	299-46-18	B	Mandalay	Member	Female	Sports an	93.72	6	28.116	590.436	1/15/2019	16:19	Cash	562.32	4.761905	28.116	4.5
18	656-95-93	A	Yangon	Member	Female	Health an	68.93	7	24.1255	506.6355	03-11-19	11:03	Credit car	482.51	4.761905	24.1255	4.6
19	765-26-69	A	Yangon	Normal	Male	Sports an	72.61	6	21.783	457.443	01-01-19	10:39	Credit car	435.66	4.761905	21.783	6.9
20	329-62-15	A	Yangon	Normal	Male	Food and	54.67	3	8.2005	172.2105	1/21/2019	18:00	Credit car	164.01	4.761905	8.2005	8.6
21	319-50-33	B	Mandalay	Normal	Female	Home and	40.3	2	4.03	84.63	03-11-19	15:30	Ewallet	80.6	4.761905	4.03	4.4
22	300-71-46	C	Naypyitaw	Member	Male	Electronic	86.04	5	21.51	451.71	2/25/2019	11:24	Ewallet	430.2	4.761905	21.51	4.8
23	371-85-57	B	Mandalay	Normal	Male	Health an	87.98	3	13.197	277.137	03-05-19	10:40	Ewallet	263.94	4.761905	13.197	5.1
24	273-16-66	B	Mandalay	Normal	Male	Home and	33.2	2	3.32	69.72	3/15/2019	12:20	Credit car	66.4	4.761905	3.32	4.4

supermarket_sales - Sheet1

*Hình 5: Dataset **supermarket_sales***

Bộ dữ liệu chứa 1001 dòng được trích xuất bằng cách sử dụng giao diện lập trình ứng dụng (API) của Github. Dữ liệu này chứa thông tin về ngày, chi nhánh, loại sản phẩm, số lượng, giá đơn vị và doanh thu.

2. Xây dựng mô hình:

2.1 Tiền xử lý dữ liệu

Trước khi bắt đầu phân tích và trực quan hóa, bước tiền xử lý dữ liệu cực kỳ quan trọng. Các bước chính trong tiền xử lý dữ liệu bao gồm:

2.1.1 Import các thư viện cần thiết

Import các thư viện cần thiết

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from google.colab import drive
drive.mount('/content/drive/')
import seaborn as sns
import locale
```

```
Drive already mounted at /content/drive/; to attempt to forcibly remount, call drive.mount("/content/drive/", force_remount=True).
```

Hình 6: Import các thư viện cần thiết.

Import các thư viện cần thiết: Trước hết, các thư viện cần thiết như pandas, numpy, matplotlib, seaborn và locale được import để sử dụng trong quá trình xử lý và trực quan hóa dữ liệu.

Bước tiếp theo là:

2.1.2 Đọc và xử lý dữ liệu:

Đọc dữ liệu: Dữ liệu từ tập tin CSV 'supermarket_sales.csv' được đọc vào DataFrame df bằng thư viện pandas.

Đọc dữ liệu và tiến hành loại bỏ các dữ liệu không hợp lệ:

```
[ ] df = pd.read_csv('/content/drive/MyDrive/ThuyetTrinhTechKHDL/supermarket_sales.csv')
    df.dropna(inplace=True)
    df.drop_duplicates(inplace=True)
    df.head()
```

	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
0	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.69	7	26.1415	548.9715	1/5/2019	13:08	Ewallet	522.83	4.761905	26.1415	9.1
1	226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.28	5	3.8200	80.2200	3/8/2019	10:29	Cash	76.40	4.761905	3.8200	9.6
2	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.33	7	16.2155	340.5255	3/3/2019	13:23	Credit card	324.31	4.761905	16.2155	7.4
3	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.22	8	23.2880	489.0480	1/27/2019	20:33	Ewallet	465.76	4.761905	23.2880	8.4
4	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.31	7	30.2085	634.3785	2/8/2019	10:37	Ewallet	604.17	4.761905	30.2085	5.3

Hình 7: Loại bỏ dữ liệu không hợp lệ

Loại bỏ dữ liệu không hợp lệ: Các hàng chứa giá trị thiếu và các hàng trùng lặp được loại bỏ để đảm bảo tính chính xác của phân tích.

dropna() được sử dụng để loại bỏ các hàng chứa giá trị thiếu (NaN).

drop_duplicates() được sử dụng để loại bỏ các hàng trùng lặp trong df.

Bước kế tiếp:

Kiểm tra và chuyển đổi kiểu dữ liệu

```
[ ] df['Date'] = pd.to_datetime(df['Date'])
    df['Quantity'] = pd.to_numeric(df['Quantity'], downcast='integer')
    df['Unit price'] = pd.to_numeric(df['Unit price'], downcast='float')
    df.head()
```

	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
0	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.690002	7	26.1415	548.9715	2019-01-05	13:08	Ewallet	522.83	4.761905	26.1415	9.1
1	226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.280000	5	3.8200	80.2200	2019-03-08	10:29	Cash	76.40	4.761905	3.8200	9.6
2	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.330002	7	16.2155	340.5255	2019-03-03	13:23	Credit card	324.31	4.761905	16.2155	7.4
3	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.220001	8	23.2880	489.0480	2019-01-27	20:33	Ewallet	465.76	4.761905	23.2880	8.4
4	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.309998	7	30.2085	634.3785	2019-02-08	10:37	Ewallet	604.17	4.761905	30.2085	5.3

Hình 8: Chuyển đổi kiểu dữ liệu

Chuyển đổi kiểu dữ liệu: Dữ liệu trong cột 'Date' được chuyển đổi thành định dạng ngày tháng, còn 'Quantity' và 'Unit price' được chuyển đổi thành kiểu dữ liệu phù hợp.

Cột 'Date' được chuyển đổi thành dạng ngày tháng bằng pd.to_datetime().

Bước kế tiếp:

Các cột 'Quantity' và 'Unit price' được chuyển đổi sang kiểu dữ liệu số bằng pd.to_numeric().

Loại bỏ dữ liệu dư thừa ở cột Product line

```
[ ] df["Product line"] = df["Product line"].str.strip()
    df.head()
```

	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
0	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.690002	7	26.1415	548.9715	2019-01-05	13:08	Ewallet	522.83	4.761905	26.1415	9.1
1	226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.280000	5	3.8200	80.2200	2019-03-08	10:29	Cash	76.40	4.761905	3.8200	9.6
2	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.330002	7	16.2155	340.5255	2019-03-03	13:23	Credit card	324.31	4.761905	16.2155	7.4
3	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.220001	8	23.2880	489.0480	2019-01-27	20:33	Ewallet	465.76	4.761905	23.2880	8.4
4	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.309998	7	30.2085	634.3785	2019-02-08	10:37	Ewallet	604.17	4.761905	30.2085	5.3

Hình 9: Loại bỏ dữ liệu dư thừa ở cột Product line

Cột 'Product line' được loại bỏ các khoảng trắng thừa bằng cách sử dụng .str.strip().

Loại bỏ dữ liệu dư thừa ở cột Unit price

```
[ ] # Loại bỏ các giá trị thừa trong cột 'Unit price'
    df['Unit price'] = df['Unit price'].drop_duplicates()
    df
```

	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating
0	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.690002	7	26.1415	548.9715	2019-01-05	13:08	Ewallet	522.83	4.761905	26.1415	9.1
1	226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.280000	5	3.8200	80.2200	2019-03-08	10:29	Cash	76.40	4.761905	3.8200	9.6
2	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.330002	7	16.2155	340.5255	2019-03-03	13:23	Credit card	324.31	4.761905	16.2155	7.4
3	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.220001	8	23.2880	489.0480	2019-01-27	20:33	Ewallet	465.76	4.761905	23.2880	8.4
4	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.309998	7	30.2085	634.3785	2019-02-08	10:37	Ewallet	604.17	4.761905	30.2085	5.3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
995	233-67-5758	C	Naypyitaw	Normal	Male	Health and beauty	40.349998	1	2.0175	42.3675	2019-01-29	13:46	Ewallet	40.35	4.761905	2.0175	6.2
996	303-96-2227	B	Mandalay	Normal	Female	Home and lifestyle	97.379997	10	48.6900	1022.4900	2019-03-02	17:16	Ewallet	973.80	4.761905	48.6900	4.4
997	727-02-1313	A	Yangon	Member	Male	Food and beverages	31.840000	1	1.5920	33.4320	2019-02-09	13:22	Cash	31.84	4.761905	1.5920	7.7

Hình 10: Loại bỏ dữ liệu dư thừa ở cột Unit price

Loại bỏ dữ liệu dư thừa: Dữ liệu trong cột 'Product line' và 'Unit price' được kiểm tra và loại bỏ giá trị thừa để tối ưu hóa dữ liệu.

Bước kế tiếp:

2.1.3 Kiểm tra dữ liệu:

Kiểm tra dữ liệu đã được loại bỏ giá trị thừa ở cột Product line và Unit price

```
[ ]  # Kiểm tra dữ liệu đã được loại bỏ giá trị thừa
     unique_values = df["Product line"].unique()
     print(unique_values)
```

```
['Health and beauty' 'Electronic accessories' 'Home and lifestyle'
 'Sports and travel' 'Food and beverages' 'Fashion accessories']
```

```
[ ]  # Kiểm tra dữ liệu sau khi loại bỏ giá trị thừa
     print(df['Unit price'])
```

```
0      74.690002
1      15.280000
2      46.330002
3      58.220001
4      86.309998
         ...
995    40.349998
996    97.379997
997    31.840000
998    65.820000
999    88.339996
Name: Unit price, Length: 1000, dtype: float32
```

Hình 11: Kiểm tra dữ liệu đã được loại bỏ giá trị thừa ở cột Product line và Unit price

Dữ liệu trong cột 'Product line' được kiểm tra sau khi loại bỏ giá trị thừa bằng .unique().

Dữ liệu trong cột 'Unit price' được kiểm tra để đảm bảo không có giá trị thừa bằng cách in ra nó.

2.2 Tính toán doanh thu và xây dựng DataFrame mới:

Tính toán và thêm cột TotalSales

```
[ ] df['TotalSales'] = df['Quantity'] * df['Unit price']
    df.head()
```

	Invoice ID	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Time	Payment	cogs	gross margin percentage	gross income	Rating	TotalSales
0	750-67-8428	A	Yangon	Member	Female	Health and beauty	74.690002	7	26.1415	548.9715	2019-01-05	13:08	Ewallet	522.83	4.761905	26.1415	9.1	522.830017
1	226-31-3081	C	Naypyitaw	Normal	Female	Electronic accessories	15.280000	5	3.8200	80.2200	2019-03-08	10:29	Cash	76.40	4.761905	3.8200	9.6	76.400002
2	631-41-3108	A	Yangon	Normal	Male	Home and lifestyle	46.330002	7	16.2155	340.5255	2019-03-03	13:23	Credit card	324.31	4.761905	16.2155	7.4	324.309998
3	123-19-1176	A	Yangon	Member	Male	Health and beauty	58.220001	8	23.2880	489.0480	2019-01-27	20:33	Ewallet	465.76	4.761905	23.2880	8.4	465.760010
4	373-73-7910	A	Yangon	Normal	Male	Sports and travel	86.309998	7	30.2085	634.3785	2019-02-08	10:37	Ewallet	604.17	4.761905	30.2085	5.3	604.169983

Tạo DataFrame mới chỉ chứa cột 'Date' và 'TotalSales':

```
[ ] df_cleaned = df[['Date', 'TotalSales']]
```

Hình 12: Tính toán và thêm cột TotalSales và tạo dataframe mới chỉ chứa cột date và TotalSales

Dựa trên dữ liệu đã được tiền xử lý, cột 'TotalSales' được tính toán bằng cách nhân số lượng ('Quantity') với giá tiền đơn vị ('Unit price').

DataFrame mới 'df_cleaned' chỉ chứa cột 'Date' và 'TotalSales' được tạo để tiếp tục thực nghiệm.

2.3 Lưu trữ dữ liệu đã được làm sạch:

Lưu trữ dữ liệu đã được làm sạch

```
[ ] df_cleaned.to_csv('/content/drive/MyDrive/ThuyetTrinhTechKHDL/cleaned_data.csv', index=False)
```

Hình 13: Lưu trữ dữ liệu đã được làm sạch

Dữ liệu trong DataFrame 'df_cleaned' sau khi được tiền xử lý được lưu trữ trong tập tin 'cleaned_data.csv' để dễ dàng sử dụng cho các phân tích và trực quan hóa sau này.

2.4 Hiển thị 5 dòng dữ liệu đã được làm sạch và chỉ chứa 2 cột "Date" & "TotalSales

Hiển thị 5 dòng dữ liệu đã được làm sạch và chỉ chứa 2 cột "Date" & "TotalSales

```
[ ]  print(df_cleaned.head())
```

```
        Date  TotalSales
0 2019-01-05  522.830017
1 2019-03-08   76.400002
2 2019-03-03  324.309998
3 2019-01-27  465.760010
4 2019-02-08  604.169983
```

Hình 14: Hiển thị 5 dòng dữ liệu sau khi làm sạch

Trong bước này, ta sẽ hiển thị ra màn hình 5 dòng dữ liệu từ DataFrame df_cleaned mà đã qua quá trình làm sạch, và chỉ bao gồm hai cột là "Date" và "TotalSales". Điều này giúp ta xem trước một phần nhỏ của dữ liệu đã được tiền xử lý và tạo sự kiểm tra nhanh về tính chính xác của các bước tiền xử lý trước khi thực hiện các phân tích và trực quan hóa dữ liệu.

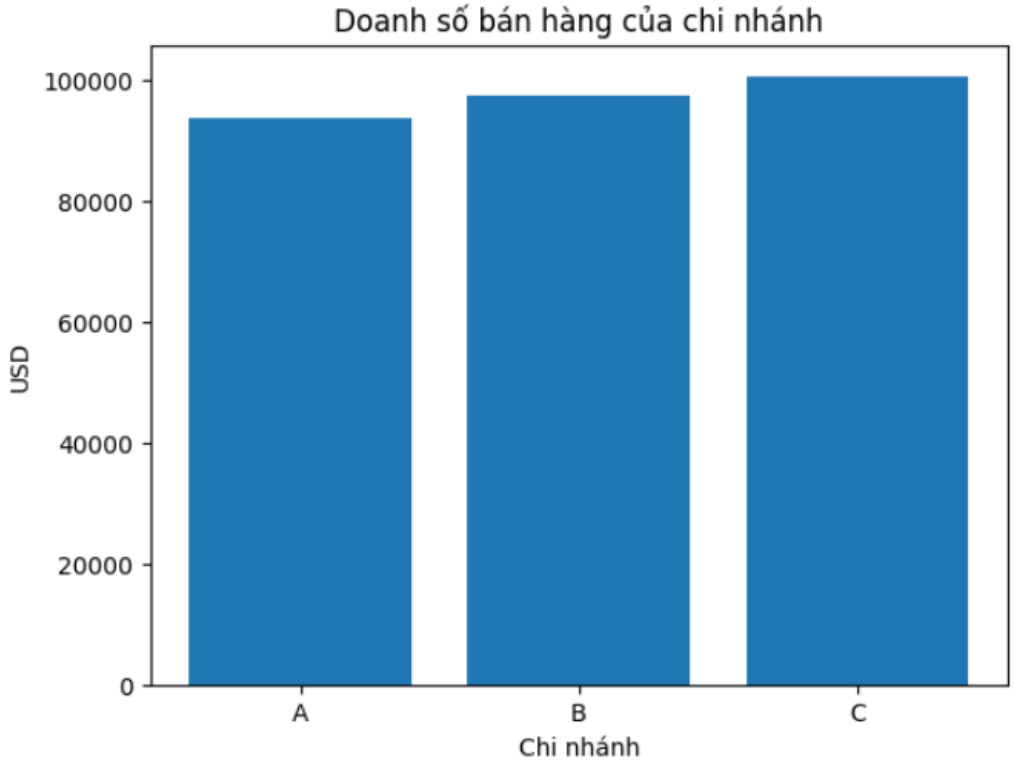
2.5 Trực quan hóa dữ liệu:

Sử dụng thư viện matplotlib và seaborn, các biểu đồ được tạo ra để trực quan hóa thông tin từ dữ liệu:

2.5.1 Biểu đồ cột doanh số bán hàng của từng chi nhánh.

Biểu đồ doanh số bán hàng của các chi nhánh

```
[ ]  branch_sales = df.groupby("Branch")["TotalSales"].sum()
     plt.bar(branch_sales.index, branch_sales.values)
     plt.xlabel("Chi nhánh")
     plt.ylabel("USD")
     plt.title("Doanh số bán hàng của chi nhánh")
     plt.show()
```



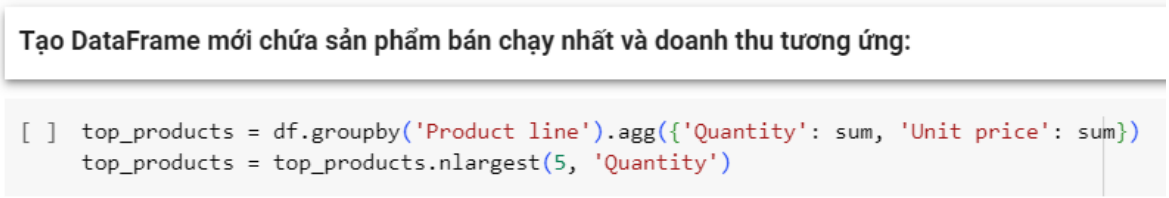
Hình 15: Doanh số bán hàng của 3 chi nhánh

Ta thực hiện phân tích doanh số bán hàng của từng chi nhánh bằng cách sử dụng phương thức groupby để tổng hợp dữ liệu theo cột "Branch". Ta tính tổng doanh thu của mỗi chi nhánh thông qua cột "TotalSales".

Sau khi tính toán, ta sử dụng hàm plt.bar() để tạo biểu đồ cột. Trục x của biểu đồ sẽ chứa tên các chi nhánh (lấy từ branch_sales.index), còn trục y sẽ biểu diễn doanh số bán hàng tương ứng của từng chi nhánh (lấy từ branch_sales.values). Các nhãn trục và tiêu đề của biểu đồ được thiết lập thông qua các hàm plt.xlabel(), plt.ylabel() và plt.title().

Cuối cùng, ta sử dụng plt.show() để hiển thị biểu đồ lên màn hình. Điều này giúp ta có cái nhìn trực quan về tình hình doanh số bán hàng của các chi nhánh khác nhau.

2.5.2 Tạo DataFrame mới chứa sản phẩm bán chạy nhất và doanh thu tương ứng:



Hình 16: Tạo DataFrame mới chứa sản phẩm bán chạy nhất và doanh thu tương ứng

Ta tiến hành xây dựng DataFrame mới để lưu trữ thông tin về sản phẩm bán chạy nhất và doanh thu tương ứng của chúng. Để làm điều này, ta sử dụng phương thức groupby trên cột "Product line" để tổng hợp dữ liệu dựa trên từng loại sản phẩm.

Trong hàm agg(), ta áp dụng hai hàm tổng hợp là sum lên cột "Quantity" và "Unit price". Điều này cho phép ta tính tổng số lượng sản phẩm đã bán và tổng doanh thu tương ứng.

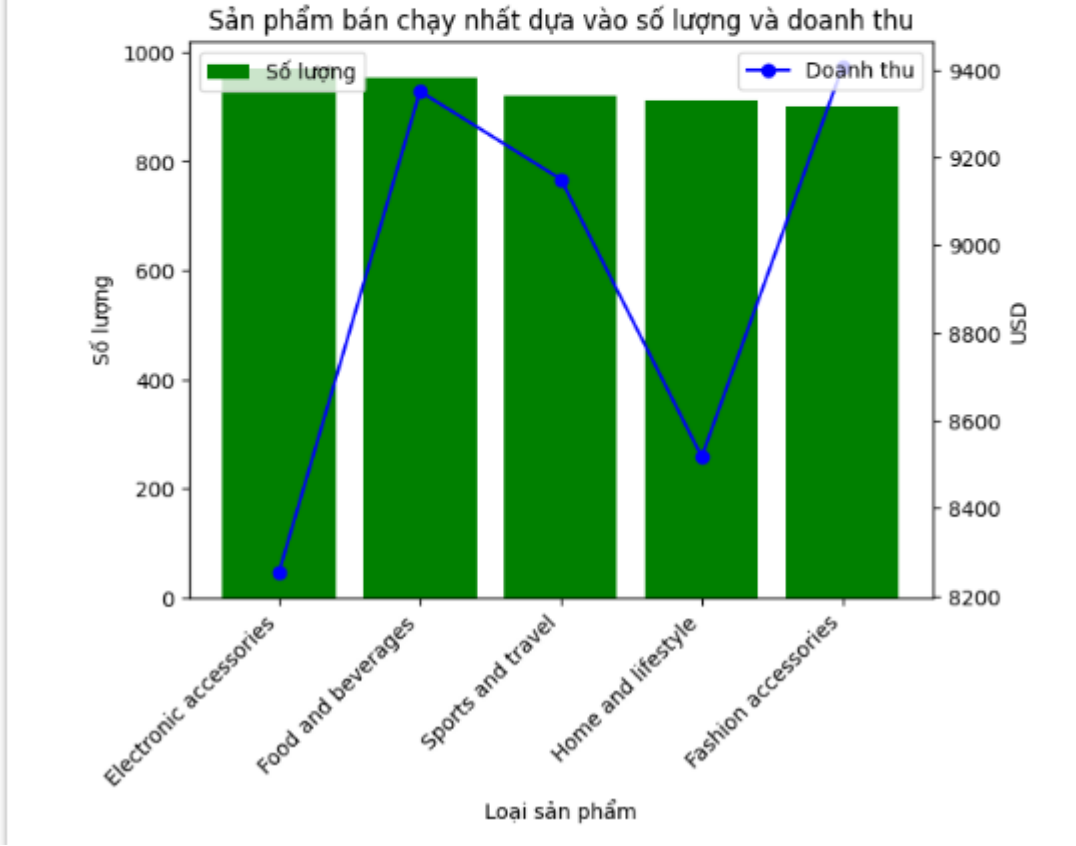
Sau khi tổng hợp dữ liệu, ta sử dụng phương thức nlargest(5, 'Quantity') để lựa chọn ra 5 loại sản phẩm có số lượng bán chạy nhất. Điều này giúp ta xác định được những sản phẩm quan trọng và đóng góp nhiều vào doanh thu.

Kết quả của quá trình này là DataFrame "top_products" chứa thông tin về 5 sản phẩm bán chạy nhất cùng với số lượng đã bán và tổng doanh thu tương ứng của chúng.

2.5.3 Biểu đồ sản phẩm bán chạy nhất dựa vào số lượng và doanh thu.

Biểu đồ sản phẩm bán chạy nhất dựa vào số lượng và doanh thu:

```
fig, ax1 = plt.subplots()
ax1.bar(top_products.index, top_products['Quantity'], color='green')
ax1.set_ylabel('Số lượng')
ax1.set_xlabel('Loại sản phẩm')
ax1.set_title('Sản phẩm bán chạy nhất dựa vào số lượng và doanh thu')
ax1.tick_params(axis='y')
ax2 = ax1.twinx()
ax2.plot(top_products.index, top_products['Unit price'], color='blue', marker='o')
ax2.set_ylabel('USD')
ax2.tick_params(axis='y')
ax1.legend(['Số lượng'], loc='upper left')
ax2.legend(['Doanh thu'], loc='upper right')
plt.setp(ax1.get_xticklabels(), rotation=45, ha='right')
plt.show()
```



Hình 17: sản phẩm bán chạy nhất dựa vào số lượng và doanh thu

ta sử dụng plt.subplots() để tạo một hình vẽ với hai trục xác định: ax1 cho biểu đồ cột và ax2 cho biểu đồ đường.

Trên ax1, ta tạo một biểu đồ cột với trục x là loại sản phẩm từ cột "Product line" và trục y là số lượng từ cột "Quantity". Màu xanh lá cây được sử dụng để thể hiện số lượng.

Trên ax2, ta tạo một biểu đồ đường với trục y là doanh thu từ cột "Unit price". Màu xanh dương với đánh dấu điểm bằng dấu "o" được sử dụng để thể hiện doanh thu.

Các trục y trên hai biểu đồ là khác nhau để đảm bảo rằng thông tin số lượng và doanh thu có thể được so sánh dễ dàng.

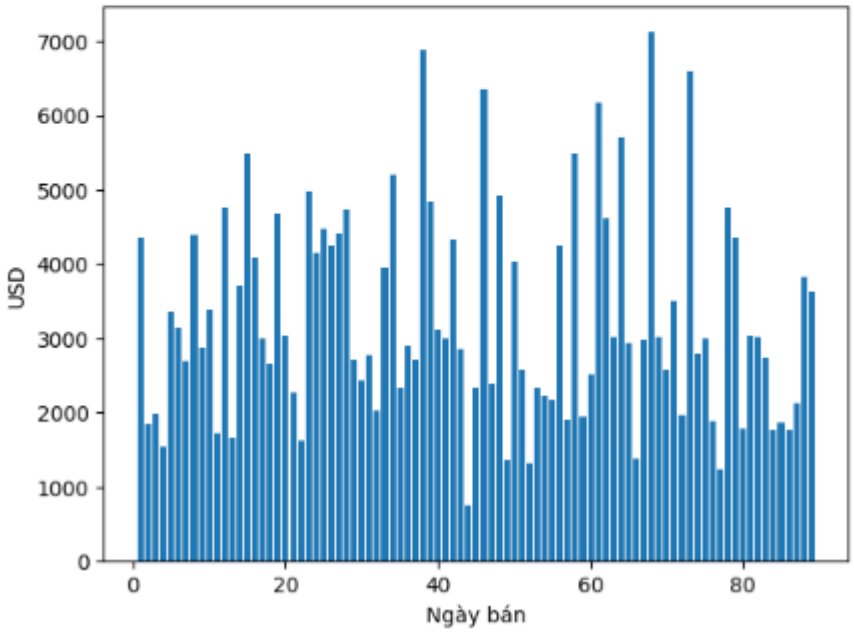
Ta cũng thiết lập các nhãn và tiêu đề cho biểu đồ, bao gồm tiêu đề "Sản phẩm bán chạy nhất dựa vào số lượng và doanh thu".

Cuối cùng, ta sử dụng plt.show() để hiển thị biểu đồ.

2.5.4 Biểu đồ doanh thu hàng ngày

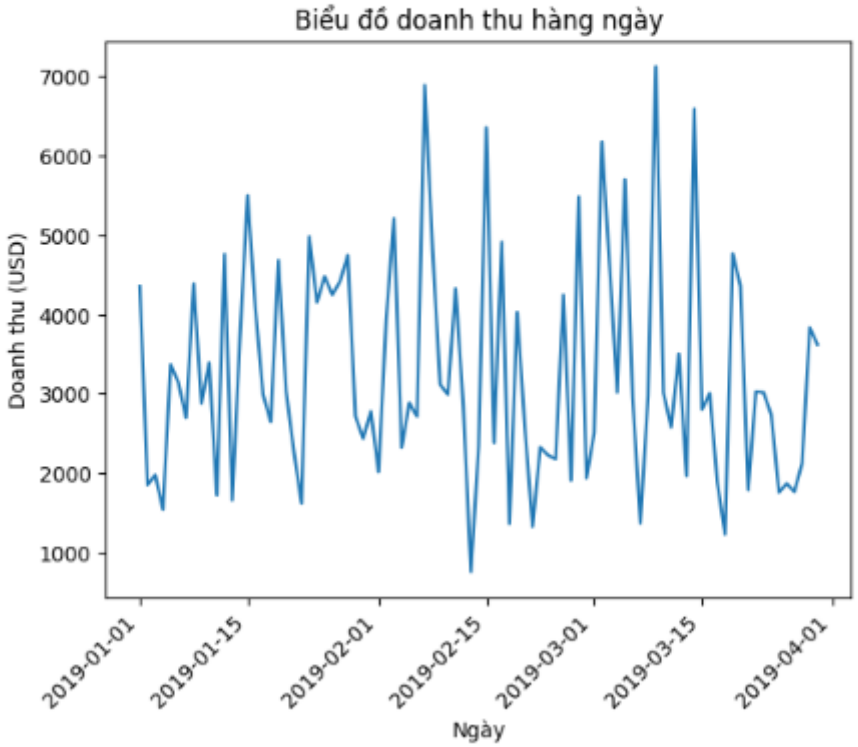
Biểu đồ doanh thu hàng ngày

```
[ ]  # Thêm cột 'TotalSales' là tổng doanh thu bán được (Quantity * Unit price)
     df['TotalSales'] = df['Quantity'] * df['Unit price']

     # Tạo một DataFrame mới chỉ chứa cột 'TotalSales'
     df_TotalSales = df.groupby('Date')['TotalSales'].sum()

     # Hiển thị biểu đồ cột
     days = range(1, len(df_TotalSales) + 1)
     plt.bar(days, df_TotalSales)
     plt.xlabel('Ngày bán')
     plt.ylabel('USD')
     plt.show()
```



Hình 18: doanh thu hàng ngày dạng cột

```
[ ]  #Line charts
     df_daily_sales = df.groupby('Date')['TotalSales'].sum()
     plt.plot(df_daily_sales.index, df_daily_sales.values)
     plt.xlabel('Ngày')
     plt.ylabel('Doanh thu (USD)')
     plt.title('Biểu đồ doanh thu hàng ngày')
     plt.xticks(rotation=45, ha='right')
     plt.show()
```



Hình 19: doanh thu hàng ngày dạng đường

Đầu tiên, ta thêm một cột mới 'TotalSales' vào DataFrame df bằng cách nhân số lượng ('Quantity') của mỗi sản phẩm với giá tiền đơn vị ('Unit price'). Điều này tạo ra tổng doanh thu từ mỗi giao dịch.

Tiếp theo, ta tạo một DataFrame mới 'df_TotalSales' chỉ chứa cột 'TotalSales'. Ta nhóm dữ liệu trong 'df_TotalSales' theo ngày và tính tổng doanh thu hàng ngày.

Ta sử dụng plt.bar() để tạo biểu đồ cột. Trục x biểu diễn các ngày bán hàng (số ngày từ ngày đầu tiên đến ngày cuối cùng trong tập dữ liệu), còn trục y biểu diễn tổng doanh thu hàng ngày.

Ta thiết lập các nhãn cho trục x và trục y bằng plt.xlabel() và plt.ylabel(), và đặt tiêu đề cho biểu đồ bằng plt.title().

Cuối cùng, ta sử dụng plt.show() để hiển thị biểu đồ cột thể hiện doanh thu hàng ngày.

Bước kế tiếp:

```
#Tổng các giá trị cột TotalSales theo từng ngày của cột Date
#max
TotalSales_value = df.groupby ('Date').sum(['total sales'])
TotalSales_value.max()
```

```
Unit price                1121.930054
Quantity                   128.000000
Tax 5%                     355.907000
Total                     7474.047000
cogs                      7118.140000
gross margin percentage     95.238095
gross income               355.907000
Rating                     151.400000
TotalSales                7118.140137
dtype: float64
```

```
#min
TotalSales_value.min()
```

```
Unit price                270.610016
Quantity                   18.000000
Tax 5%                     44.487500
Total                     934.237500
cogs                      889.750000
gross margin percentage     28.571429
gross income               44.487500
Rating                     36.700000
TotalSales                760.750000
dtype: float64
```

Hình 20: Tổng giá trị Min-Max

Ta sử dụng groupby('Date') để nhóm dữ liệu theo ngày. Sau đó, ta sử dụng hàm sum() để tính tổng các giá trị trong cột 'TotalSales' cho mỗi ngày.

Để tính giá trị lớn nhất (tổng doanh thu cao nhất) và giá trị nhỏ nhất (tổng doanh thu thấp nhất), ta sử dụng phương thức max() và min() trên đối tượng

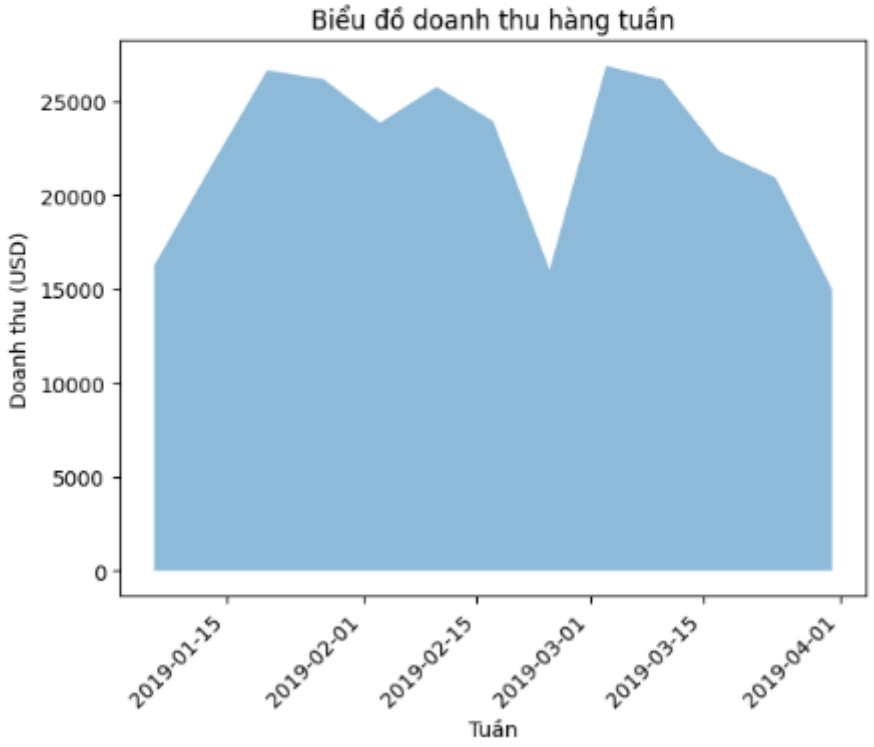
TotalSales_value mà ta đã tính toán trước đó. Điều này sẽ trả về giá trị lớn nhất và nhỏ nhất của tổng doanh thu theo ngày.

Đây là cách ta thực hiện tính toán và thống kê giá trị lớn nhất và nhỏ nhất của tổng doanh thu hàng ngày từ dữ liệu.

2.5.5 Biểu đồ doanh thu hàng tuần

Biểu đồ biểu diễn doanh thu hàng tuần

```
[ ] df_weekly_sales = df.resample('W', on='Date')['TotalSales'].sum()
    plt.fill_between(df_weekly_sales.index, df_weekly_sales.values, alpha=0.5)
    plt.xlabel('Tuần')
    plt.ylabel('Doanh thu (USD)')
    plt.title('Biểu đồ doanh thu hàng tuần')
    plt.xticks(rotation=45, ha='right')
    plt.show()
```



Hình 21: Doanh thu hàng tuần

Đầu tiên, ta sử dụng phương thức resample('W', on='Date') để tổng hợp dữ liệu hàng tuần từ dữ liệu hàng ngày trong cột 'TotalSales'. Kết quả của phép tổng hợp này được lưu vào DataFrame df_weekly_sales.

Sau đó, ta sử dụng hàm fill_between để tạo biểu đồ dạng diện tích, trong đó ta điền màu vào phía dưới đường biểu diễn doanh thu hàng tuần. Tham số alpha=0.5 điều chỉnh độ trong suốt của màu điền.

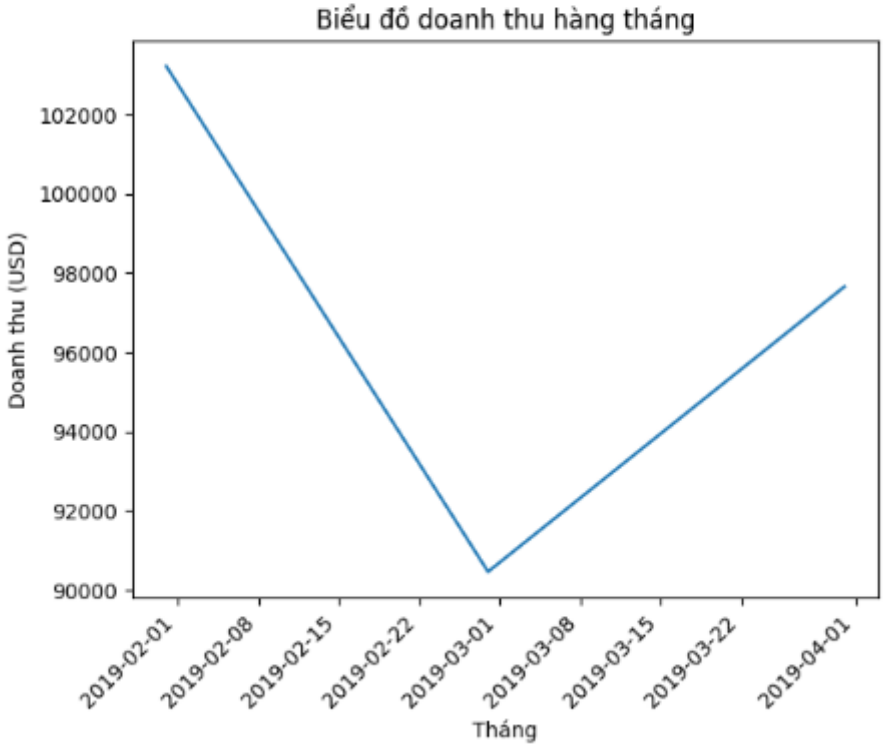
Các hàm xlabel, ylabel và title được sử dụng để đặt tên cho trục x, trục y và tiêu đề của biểu đồ tương ứng.

Cuối cùng, xticks(rotation=45, ha='right') được sử dụng để hiển thị các nhãn trục x ở góc nghiêng để tránh việc chồng chất khi có quá nhiều tuần.

2.5.6 Biểu đồ doanh thu hàng tháng

Biểu đồ doanh thu hàng tháng

```
[ ]  df_monthly_sales = df.resample('M', on='Date')['TotalSales'].sum()
     plt.plot(df_monthly_sales.index, df_monthly_sales.values)
     plt.xlabel('Tháng')
     plt.ylabel('Doanh thu (USD)')
     plt.title('Biểu đồ doanh thu hàng tháng')
     plt.xticks(rotation=45, ha='right')
     plt.show()
```



Hình 22: Doanh thu hàng tháng

Đầu tiên, ta sử dụng phương thức resample('M', on='Date') để tổng hợp dữ liệu hàng tháng từ dữ liệu hàng ngày trong cột 'TotalSales'. Kết quả của phép tổng hợp này được lưu vào Series df_monthly_sales.

Sau đó, ta sử dụng hàm plot để tạo biểu đồ đường, trong đó trục x là các ngày đầu tháng và trục y là doanh thu tương ứng với mỗi tháng.

Các hàm xlabel, ylabel và title được sử dụng để đặt tên cho trục x, trục y và tiêu đề của biểu đồ tương ứng.

Cuối cùng, xticks(rotation=45, ha='right') được sử dụng để hiển thị các nhãn trục x ở góc nghiêng để tránh việc chồng chất khi có quá nhiều tháng.

2.5.7 Tính toán tổng doanh thu sau 90 ngày:

Tổng doanh thu đạt được sau 90 ngày

```
[ ]  #Tính tổng doanh thu đạt được
     #USD
     locale.setlocale(locale.LC_ALL, 'en_US.UTF-8')
     total_revenue = df['TotalSales'].sum()
     formatted_total_revenue = locale.currency(total_revenue, grouping=True)

     print("Tổng doanh thu: ", formatted_total_revenue)

     Tổng doanh thu:  $291,320.38
```

Hình 23: Doanh thu đạt được sau 90 ngày

Đầu tiên, ta sử dụng hàm sum() để tính tổng doanh thu của tất cả các giao dịch.

Sau đó, ta sử dụng thư viện locale để định dạng tổng doanh thu thành một chuỗi tiền tệ theo định dạng của Hoa Kỳ. Điều này được thực hiện bằng cách đặt vùng địa lý (locale) thành 'en_US.UTF-8'.

Cuối cùng, ta in ra kết quả tổng doanh thu đã được định dạng để hiển thị.

2.6 Đánh giá hiệu suất mô hình:

Trong tài liệu về phân tích và trực quan hóa dữ liệu bán hàng siêu thị, sau khi thực hiện các bước xử lý dữ liệu và trực quan hóa, ta có thể đánh giá mô hình dựa trên các yếu tố sau:

Hiểu biết về doanh số bán hàng và sản phẩm: Qua việc trực quan hóa, ta có cái nhìn rõ ràng về doanh số bán hàng của từng chi nhánh, cũng như các sản phẩm bán chạy nhất và doanh thu tương ứng. Điều này giúp ta hiểu rõ hơn về hiệu suất kinh doanh và xu hướng của cửa hàng.

Phân tích các biểu đồ: Các biểu đồ thể hiện sự biến đổi của doanh thu theo ngày, tuần và tháng giúp ta nhận ra những ngày, tuần hoặc tháng có doanh thu cao hoặc thấp hơn. Điều này có thể giúp ta điều chỉnh chiến lược kinh doanh và quảng cáo theo thời gian.

Phát hiện xu hướng và biến đổi: Trực quan hóa dữ liệu giúp ta dễ dàng nhận thấy những xu hướng, biến đổi và sự tương quan giữa các yếu tố trong dữ liệu. Ví dụ, ta có thể thấy liệu có sự tương quan giữa giá bán và số lượng bán hàng không.

Phân tích sự thay đổi theo thời gian: Các biểu đồ thể hiện doanh thu hàng ngày, hàng tuần và hàng tháng giúp ta nhận ra mẫu thay đổi trong doanh thu theo thời gian. Điều này có thể giúp ta xác định những thời kỳ có nhiều cơ hội để tăng cường doanh thu.

Tổng quan về doanh thu: Kết quả tổng doanh thu sau 90 ngày cho thấy mức doanh thu đạt được trong thời gian đó. Điều này có thể giúp ta đánh giá hiệu suất kinh doanh trong giai đoạn thử nghiệm.

Tóm lại, việc phân tích và trực quan hóa dữ liệu giúp ta có cái nhìn tổng quan về hoạt động kinh doanh, từ đó hỗ trợ quyết định kinh doanh thông minh và điều chỉnh chiến lược trong tương lai.

CHƯƠNG IV: KẾT LUẬN

1. Kết luận:

1.1 Tóm tắt kết quả:

Ta đã thực hiện một quá trình phân tích dữ liệu bán hàng siêu thị một cách chi tiết và cặn kẽ. Từ việc tiền xử lý dữ liệu đến trực quan hóa, ta đã thực hiện các bước quan trọng để hiểu rõ hơn về doanh số bán hàng, sản phẩm phổ biến, chi nhánh hiệu quả, và xu hướng doanh thu theo thời gian.

1.2 Kết quả đạt được:

Ta đã thực hiện thành công việc tiền xử lý dữ liệu bằng cách loại bỏ dữ liệu không hợp lệ và chuyển đổi các kiểu dữ liệu phù hợp.

Xây dựng mô hình phân tích dựa trên các biểu đồ và trực quan hóa để tìm ra các xu hướng và thông tin quan trọng.

Đã thực hiện trực quan hóa dữ liệu bằng các biểu đồ, giúp hình dung rõ ràng về sự phân phối của doanh số bán hàng, sản phẩm bán chạy và doanh thu hàng ngày, tuần, tháng.

Áp dụng công nghệ khoa học dữ liệu vào thực tế, ta đã tạo ra các kết quả và thông tin hữu ích cho quản lý và ra quyết định.

Kết luận:

Tóm lại, ta đã thể hiện tiềm năng của công nghệ khoa học dữ liệu trong việc phân tích và trực quan hóa dữ liệu bán hàng siêu thị, đồng thời cũng gợi mở cho nhiều hướng phát triển tương lai.

1.3 Hạn chế

Mặc dù dự án đã đạt được một số kết quả tích cực, còn một số hạn chế cần lưu ý:

Dữ liệu có thể không hoàn toàn đại diện cho mọi tình huống thực tế trong siêu thị.

Các mô hình dự đoán có thể bị ảnh hưởng bởi các yếu tố không xác định, và cần được cải thiện và kiểm tra định kỳ.

Phân tích chỉ dựa trên dữ liệu lịch sử, không thể dự đoán một cách chính xác những biến đổi không mong đợi trong tương lai..

Kết luận:

Những hạn chế này cần được xem xét và giải quyết một cách cẩn thận trong quá trình nghiên cứu và thực hiện để đảm bảo tính khả thi và hiệu quả của mô hình.

1.4 Hướng phát triển

Mở rộng dự án sang các lĩnh vực khác, như dự đoán xu hướng thị trường, dự báo nguồn cung cấp và nhu cầu.

Kết hợp thêm các yếu tố bên ngoài như dữ liệu thời tiết, ngày lễ, sự kiện đặc biệt để cải thiện tính chính xác của mô hình.

Nâng cao mô hình dự đoán bằng cách sử dụng các thuật toán phức tạp hơn và tối ưu hóa tham số.

TÀI LIỆU THAM KHẢO

1. Smith, J. M. (2018). "Data Science for Business." Springer.
2. Zhang, Y., & Wu, Q. (2019). "Exploratory Data Analysis: Concepts and Methods." CRC Press.
3. McKinney, W. (2017). "Python for Data Analysis." O'Reilly Media.
4. Brownlee, J. (2021). "Better Machine Learning: Crash Course." Machine Learning Mastery.
5. Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). "Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies." MIT Press.
6. Raschka, S., & Mirjalili, V. (2019). "Python Machine Learning." Packt Publishing.

Mã nguồn chương trình hoàn chỉnh:

https://colab.research.google.com/drive/18mC8chQQvxOsoNR-COSIVggymMfKtMor?usp=sharing

QR code (source code):