TRƯỜNG ĐẠI HỌC ĐẠI NAM

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**HỌC PHẦN: DỮ LIỆU LỚN**

**ĐỀ TÀI: PHÂN TÍCH HÀNH VI MUA SẮM CỦA NGƯỜI TIÊU DÙNG**

**Giảng viên: Ths. Lê Thị Thuỳ Trang**

**Ts. Trần Quý Nam**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TT | Mã sv | Họ và Tên | Ngày Sinh | Lớp |
| 1 |  | Dương Xuân Thành |  | CNTT16-01 |
| 2 |  | Hồ Quang Huy |  | CNTT16-01 |
| 3 |  | Nguyễn Mạnh Cường |  | CNTT16-01 |
| 4 |  | Ngô Thành Đạt |  | CNTT16-01 |

**Hà Nội, năm 2024**

TRƯỜNG ĐẠI HỌC ĐẠI NAM

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo, company name

Description automatically generated

**BÀI TẬP LỚN**

**HỌC PHẦN: DỮ LIỆU LỚN**

**ĐỀ TÀI: PHÂN TÍCH HÀNH VI MUA SẮM CỦA NGƯỜI TIÊU DÙNG**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TT | Mã sv | Họ và Tên | Ngày Sinh | Điểm | |
| Bằng Số | Bằng Chữ |
| 1 |  | Dương Xuân Thành |  |  |  |
| 2 |  | Hồ Quang Huy |  |  |  |
| 3 |  | Nguyễn Mạnh Cường |  |  |  |
| 4 |  | Ngô Thành Đạt |  |  |  |

|  |  |
| --- | --- |
| **CÁN BỘ CHẤM THI 1**  **Trần Quý Nam** | **CÁN BỘ CHẤM THI 2** |

### 

**Hà Nội, năm 2024**

**LỜI NÓI ĐẦU**

Trong thời đại công nghệ số hiện nay, việc khai thác dữ liệu đóng vai trò quan trọng trong việc đưa ra các quyết định kinh doanh chiến lược. Với sự bùng nổ của dữ liệu, các doanh nghiệp cần những công cụ mạnh mẽ để xử lý, phân tích và trực quan hóa thông tin một cách hiệu quả.

Bài nghiên cứu này tập trung vào việc ứng dụng **Apache Spark** thông qua thư viện **sparklyr** trong ngôn ngữ lập trình **R** để xử lý dữ liệu bán hàng lớn. Thông qua quá trình phân tích, bài làm sẽ giúp nhận diện các **xu hướng doanh thu**, **sản phẩm bán chạy** và **hành vi mua hàng của khách hàng**. Đồng thời, ứng dụng **Shiny** được xây dựng nhằm trực quan hóa kết quả phân tích, giúp người dùng dễ dàng tương tác và khám phá dữ liệu.

Mục tiêu chính của bài nghiên cứu là phát triển một công cụ phân tích mạnh mẽ, cho phép doanh nghiệp có cái nhìn tổng quan về hiệu quả kinh doanh, từ đó đưa ra các chiến lược phù hợp để tối ưu hóa hoạt động bán hàng và tiếp cận khách hàng tiềm năng.

**MỤC LỤC**

[**CHƯƠNG 1. TỔNG QUAN VÀ CƠ SỞ LÝ THUYẾT** 5](#_Toc192959801)

[**1.** **Giới thiệu về hệ thống bán hàng và hành vi khách hàng.** 5](#_Toc192959802)

[**2.** **Cơ sở lý thuyết về phân tích dữ liệu bán hàng và mô hình hóa hành vi người tiêu dùng** 5](#_Toc192959803)

[**3.** **Các phương pháp phân tích dữ liệu và công nghệ hỗ trợ** 6](#_Toc192959804)

[**CHƯƠNG 2. PHÂN TÍCH DỮ LIỆU VÀ ĐÁNH GIÁ KẾT QUẢ** 8](#_Toc192959805)

[**1.** **Mô tả dữ liệu.** 8](#_Toc192959806)

[**2.** **Tiền xử lý dữ liệu.** 8](#_Toc192959807)

[*2.1.* *Cài đặt thư viện Spark.* 8](#_Toc192959808)

[*2.2.* *Đọc dữ liệu.* 9](#_Toc192959809)

[*2.3.* *Lựa chọn các cột cần thiết.* 9](#_Toc192959810)

[*2.4.* *Đưa dữ liệu lên Spark.* 9](#_Toc192959811)

[*2.5.* *Chuẩn hoá tên cột.* 9](#_Toc192959812)

[*2.6.* *Kiểm tra và xử lý dữ liệu.* 9](#_Toc192959813)

[*2.7.* *Thu thập dữ liệu từ Spark và R.* 10](#_Toc192959814)

[**3.** **Phân Tích Xu Hướng Doanh Thu và Hành Vi Khách Hàng.** 10](#_Toc192959815)

[**4.** **Đánh Giá Sự Khác Biệt Giữa Các Quốc Gia.** 10](#_Toc192959816)

[**CHƯƠNG 3. KẾT QUẢ XỬ LÝ, PHÂN TÍCH DỮ LIỆU VÀ ỨNG DỤNG** 11](#_Toc192959817)

[**1.** **Tổng quan về chương trình Phân tích hành vi mua sắm của người tiêu dụng.** 11](#_Toc192959818)

[***1.1.*** *Cài đặt và nạp thư viện.* 11](#_Toc192959819)

[***1.2.*** *Kết nối Spark.* 11](#_Toc192959820)

[***1.3.*** *Lọc dữ liệu từ Excel.* 12](#_Toc192959821)

[***1.4.*** *Lựa chọn các cột cần thiết.* 12](#_Toc192959822)

[***1.5.*** *Chuyển dữ liệu từ Spark.* 12](#_Toc192959823)

[***1.6.*** *Chuẩn hoá tên cột.* 13](#_Toc192959824)

[***1.7.*** *Xử lý dữ liệu.* 13](#_Toc192959825)

[***1.8.*** *Chuyển dữ liệu từ Spark về R.* 13](#_Toc192959826)

[***1.9.*** *Tạo giao diện người dùng (UI) với Shiny:* 13](#_Toc192959827)

[***1.10.*** *Xây dựng server.* 14](#_Toc192959828)

[***1.11.*** *Ngắt kết nối Spark.* 16](#_Toc192959829)

[***1.12.*** *Chạy ứng dụng.* 16](#_Toc192959830)

[***2.*** *Kết quả xử lý dữ liệu.* 17](#_Toc192959831)

[***3.*** ***Tổng quan về bài toán.*** 19](#_Toc192959832)

[**3.1.** Mục tiêu: 19](#_Toc192959833)

[**3.2.** Công cụ sử dụng: 19](#_Toc192959834)

[3.4. Quy trình thực hiện: 20](#_Toc192959835)

[3.5. Phân tích dữ liệu: 20](#_Toc192959836)

[3.6. Trực quan hóa: 20](#_Toc192959837)

[3.7. Ứng dụng: 20](#_Toc192959838)

[**KẾT LUẬN** 21](#_Toc192959839)

[**DANH MỤC TÀI LIỆU THAM KHẢO** 22](#_Toc192959840)

**CHƯƠNG 1. TỔNG QUAN VÀ CƠ SỞ LÝ THUYẾT**

1. **Giới thiệu về hệ thống bán hàng và hành vi khách hàng.**

Trong bối cảnh nền kinh tế số hiện nay, các hệ thống bán hàng ngày càng đóng vai trò quan trọng trong việc kết nối doanh nghiệp với khách hàng, đặc biệt là trong môi trường thương mại điện tử và bán lẻ hiện đại. Hệ thống bán hàng không chỉ bao gồm các quy trình từ tiếp nhận đơn hàng, xử lý thanh toán, quản lý tồn kho, đến giao hàng và hậu mãi, mà còn là một công cụ quan trọng để thu thập và phân tích dữ liệu khách hàng nhằm tối ưu hóa chiến lược kinh doanh.

Hành vi khách hàng là tập hợp những hành động, thói quen, và quyết định của người tiêu dùng trong quá trình mua sắm và sử dụng sản phẩm hoặc dịch vụ. Hiểu được hành vi khách hàng giúp doanh nghiệp dự đoán nhu cầu thị trường, cải thiện chất lượng dịch vụ, và cá nhân hóa trải nghiệm mua sắm. Thông qua việc phân tích hành vi này, doanh nghiệp có thể xác định các yếu tố ảnh hưởng đến quyết định mua hàng như giá cả, chất lượng, chương trình khuyến mãi, và các yếu tố tâm lý khác.

1. **Cơ sở lý thuyết về phân tích dữ liệu bán hàng và mô hình hóa hành vi người tiêu dùng**

Phân tích dữ liệu bán hàng là quá trình thu thập, xử lý và diễn giải dữ liệu nhằm khám phá các xu hướng và mô hình ẩn trong hành vi mua sắm của khách hàng. Đây là công cụ mạnh mẽ giúp doanh nghiệp đưa ra quyết định dựa trên dữ liệu thực tế, từ đó tối ưu hóa chiến lược kinh doanh và tăng cường sự hài lòng của khách hàng.

Các phương pháp phân tích dữ liệu bán hàng phổ biến bao gồm:

* **Phân tích mô tả**: Xem xét các thông kê cơ bản như doanh thu, số lượng bán, và tần suất mua hàng nhằm tạo ra cái nhìn tổng quan về hoạt động kinh doanh.
* **Phân tích khám phá**: Tìm kiếm các mối quan hệ giữa các biến số trong dữ liệu, như mối tương quan giữa giá bán và số lượng sản phẩm bán ra.
* **Phân tích dự báo**: Áp dụng các mô hình dự đoán như hồi quy tuyến tính, mô hình ARIMA, hay máy học để dự báo doanh thu, nhu cầu sản phẩm, hoặc hành vi khách hàng trong tương lai.
* **Phân khúc khách hàng**: Phân loại khách hàng dựa trên các đặc điểm chung như tần suất mua hàng, giá trị đơn hàng trung bình, hay sở thích tiêu dùng.

Mô hình hóa hành vi người tiêu dùng là quá trình xây dựng các mô hình nhằm mô phỏng và dự đoán các quyết định mua sắm của khách hàng dựa trên dữ liệu thu thập được. Một số mô hình phổ biến bao gồm:

* **Mô hình RFM (Recency, Frequency, Monetary)**: Phân tích khách hàng dựa trên thời gian mua hàng gần nhất, tần suất mua hàng, và tổng giá trị chi tiêu.
* **Mô hình phễu mua hàng (Purchase Funnel)**: Theo dõi hành trình mua hàng của khách từ nhận thức về sản phẩm đến quyết định mua sắm.
* **Mô hình dự đoán churn (Customer Churn Prediction)**: Dự đoán khả năng khách hàng ngừng sử dụng sản phẩm hoặc dịch vụ, giúp doanh nghiệp đưa ra các biện pháp giữ chân khách hàng kịp thời.

1. **Các phương pháp phân tích dữ liệu và công nghệ hỗ trợ**

Sự phát triển của công nghệ đã mở ra nhiều cơ hội mới cho việc phân tích dữ liệu bán hàng và hành vi khách hàng. Các công cụ và nền tảng hỗ trợ việc xử lý dữ liệu lớn và trực quan hóa thông tin đang trở thành yếu tố không thể thiếu trong việc nâng cao hiệu quả phân tích.

* **Big Data**: Việc thu thập dữ liệu từ nhiều nguồn khác nhau như hệ thống POS (Point of Sale), website, mạng xã hội, và các nền tảng thương mại điện tử tạo ra khối lượng dữ liệu khổng lồ. Phân tích Big Data cho phép doanh nghiệp nhận diện xu hướng thị trường, tối ưu hóa chuỗi cung ứng, và cá nhân hóa trải nghiệm khách hàng.
* **Apache Spark**: Đây là một nền tảng xử lý dữ liệu lớn mạnh mẽ, cho phép xử lý dữ liệu theo thời gian thực và hỗ trợ nhiều ngôn ngữ lập trình. Việc sử dụng Spark giúp tăng tốc độ xử lý dữ liệu thông qua tính năng tính toán phân tán, đặc biệt hữu ích khi làm việc với các tập dữ liệu lớn.
* **Shiny**: Là một framework cho phép tạo ra các ứng dụng web tương tác bằng ngôn ngữ R. Shiny giúp trực quan hóa dữ liệu một cách dễ dàng, cho phép người dùng không chuyên về lập trình vẫn có thể tương tác với dữ liệu và khám phá những thông tin quan trọng thông qua các biểu đồ trực quan.
* **Machine Learning (ML)**: Các thuật toán học máy như hồi quy, cây quyết định, mạng nơ-ron nhân tạo giúp phát hiện các mẫu hình phức tạp trong dữ liệu, từ đó dự đoán hành vi khách hàng và tối ưu hóa chiến lược bán hàng.

Việc kết hợp các công nghệ này trong quá trình phân tích dữ liệu bán hàng mang lại nhiều lợi ích cho doanh nghiệp như: giảm thời gian xử lý dữ liệu, phát hiện nhanh chóng những biến động trong hành vi khách hàng, và tạo ra các mô hình dự đoán chính xác. Qua đó, doanh nghiệp có thể xây dựng các chiến lược kinh doanh linh hoạt hơn, đáp ứng kịp thời nhu cầu thị trường, và nâng cao khả năng cạnh tranh.

**CHƯƠNG 2. PHÂN TÍCH DỮ LIỆU VÀ ĐÁNH GIÁ KẾT QUẢ**

1. **Mô tả dữ liệu.**

Dữ liệu bán hàng được thu thập từ hệ thống quản lý bán lẻ, bao gồm các thông tin về hóa đơn, sản phẩm, số lượng, ngày giao dịch, giá bán, ID khách hàng và quốc gia. Dữ liệu này cung cấp cái nhìn tổng quan về hành vi mua sắm của khách hàng cũng như xu hướng tiêu dùng trên phạm vi toàn cầu.

1. **Tiền xử lý dữ liệu.**

Quá trình tiền xử lý dữ liệu bao gồm việc lọc bỏ các giá trị thiếu, chuẩn hóa tên cột và chuyển đổi kiểu dữ liệu phù hợp để phục vụ cho việc phân tích. Ngoài ra, dữ liệu được đưa lên Spark để xử lý song song, giúp tăng tốc độ xử lý đối với tập dữ liệu lớn.

* 1. *Cài đặt thư viện Spark.*

install.packages(c("sparklyr", "readxl", "dplyr", "ggplot2", "shiny"))

library(sparklyr)

library(readxl)

library(dplyr)

library(ggplot2)

library(shiny)

# Kết nối Spark

sc <- spark\_connect(master = "local")

* 1. *Đọc dữ liệu.*

# Đường dẫn file

file\_path <- "C:/Users/TITBB/OneDrive/Máy tính/Big Data/Assignment-1\_Data.xlsx"

# Kiểm tra file tồn tại trước khi đọc

to\_load <- if (file.exists(file\_path)) read\_excel(file\_path) else stop("File không tồn tại!")

* 1. *Lựa chọn các cột cần thiết.*

# Chọn các cột quan trọng

to\_load <- to\_load %>% select(BillNo, Itemname, Quantity, Date, Price, CustomerID, Country)

* 1. *Đưa dữ liệu lên Spark.*

# Chuyển dữ liệu lên Spark

sales\_tbl <- copy\_to(sc, to\_load, "sales\_data", overwrite = TRUE)

* 1. *Chuẩn hoá tên cột.*

# Đổi tên cột về dạng chữ thường

sales\_tbl <- sales\_tbl %>% rename\_with(tolower)

* 1. *Kiểm tra và xử lý dữ liệu.*

# Lọc các giá trị thiếu và chuyển đổi kiểu dữ liệu ngày tháng

sales\_tbl <- sales\_tbl %>%

filter(!is.na(date), !is.na(country), !is.na(quantity)) %>%

mutate(date = as.Date(date))

* 1. *Thu thập dữ liệu từ Spark và R.*

# Chuyển dữ liệu từ Spark về R để vẽ biểu đồ

sales\_data\_r <- sales\_tbl %>% collect()

1. **Phân Tích Xu Hướng Doanh Thu và Hành Vi Khách Hàng.**

Phân tích được thực hiện thông qua các biểu đồ trực quan, giúp làm rõ các xu hướng doanh thu, sản phẩm bán chạy và hành vi khách hàng. Các kết quả từ phân tích hỗ trợ việc đưa ra các quyết định kinh doanh và chiến lược tiếp thị hiệu quả hơn.

1. **Đánh Giá Sự Khác Biệt Giữa Các Quốc Gia.**

Dữ liệu được phân tách theo từng quốc gia để so sánh hành vi mua hàng, từ đó nhận diện các thị trường tiềm năng và điều chỉnh chiến lược kinh doanh phù hợp với từng khu vực.

**CHƯƠNG 3. KẾT QUẢ XỬ LÝ, PHÂN TÍCH DỮ LIỆU VÀ ỨNG DỤNG**

1. **Tổng quan về chương trình Phân tích hành vi mua sắm của người tiêu dụng.**
   1. *Cài đặt và nạp thư viện.*

install.packages(c("sparklyr", "readxl", "dplyr", "ggplot2", "shiny"))

**install.packages()**: Cài đặt các gói cần thiết cho việc phân tích và trực quan hóa dữ liệu:

* **sparklyr**: Kết nối R với Apache Spark để xử lý dữ liệu lớn.
* **readxl**: Đọc dữ liệu từ các tệp Excel.
* **dplyr**: Hỗ trợ thao tác và xử lý dữ liệu dạng bảng.
* **ggplot2**: Vẽ biểu đồ trực quan.
* **shiny**: Tạo giao diện web tương tác với dữ liệu.

library(sparklyr)

library(readxl)

library(dplyr)

library(ggplot2)

library(shiny)

**library()**: Nạp các gói thư viện đã cài đặt vào môi trường làm việc.

* 1. *Kết nối Spark.*

# Kết nối Spark

sc <- spark\_connect(master = "local")

**spark\_connect()**: Tạo kết nối với Spark:

* **master = "local"**: Chạy Spark trên máy tính cục bộ, không yêu cầu cụm máy chủ phân tán.
  1. *Lọc dữ liệu từ Excel.*

# Đường dẫn file

file\_path <- "C:/Users/TITBB/OneDrive/Máy tính/Big Data/Assignment-1\_Data.xlsx"

# Kiểm tra file tồn tại trước khi đọc

to\_load <- if (file.exists(file\_path)) read\_excel(file\_path) else stop("File không tồn tại!")

**file.exists()**: Kiểm tra sự tồn tại của file tại đường dẫn được cung cấp.

**read\_excel()**: Đọc dữ liệu từ file Excel vào R.

**stop()**: Dừng chương trình nếu file không tồn tại và thông báo lỗi.

* 1. *Lựa chọn các cột cần thiết.*

# Chọn các cột cần thiết

to\_load <- to\_load %>% select(BillNo, Itemname, Quantity, Date, Price, CustomerID, Country)

**select()**: Chọn ra các cột quan trọng từ dữ liệu thô:

* **BillNo**: Mã hóa đơn.
* **Itemname**: Tên sản phẩm.
* **Quantity**: Số lượng sản phẩm.
* **Date**: Ngày giao dịch.
* **Price**: Giá bán sản phẩm.
* **CustomerID**: Mã khách hàng.
* **Country**: Quốc gia.
  1. *Chuyển dữ liệu từ Spark.*

# Chuyển dữ liệu lên Spark

sales\_tbl <- copy\_to(sc, to\_load, "sales\_data", overwrite = TRUE).

**copy\_to()**: Chuyển dữ liệu từ R lên Spark để thực hiện xử lý song song:

* **sc**: Kết nối Spark đã tạo trước đó.
* **to\_load**: Dữ liệu cần chuyển lên Spark.
* **"sales\_data"**: Tên bảng dữ liệu trong Spark.
* **overwrite = TRUE**: Ghi đè bảng dữ liệu nếu tồn tại từ trước.
  1. *Chuẩn hoá tên cột.*

# Chuẩn hóa tên cột

sales\_tbl <- sales\_tbl %>% rename\_with(tolower)

**rename\_with(tolower)**: Chuyển tất cả tên cột về dạng chữ thường để dễ xử lý sau này.

* 1. *Xử lý dữ liệu.*

# Kiểm tra và xử lý dữ liệu

sales\_tbl <- sales\_tbl %>%

filter(!is.na(date), !is.na(country), !is.na(quantity)) %>%

mutate(date = as.Date(date))

**filter()**: Lọc bỏ các dòng có giá trị thiếu trong cột date, country, và quantity.

**mutate()**: Chuyển cột date từ kiểu dữ liệu chuỗi sang kiểu ngày tháng (Date).

* 1. *Chuyển dữ liệu từ Spark về R.*

# Chuyển dữ liệu từ Spark về R để vẽ biểu đồ

sales\_data\_r <- sales\_tbl %>% collect()

**collect()**: Tải toàn bộ dữ liệu từ Spark về R để thực hiện trực quan hóa.

* 1. *Tạo giao diện người dùng (UI) với Shiny:*

# UI của Shiny

ui <- fluidPage(

titlePanel("Phân tích hành vi bán hàng"),

sidebarLayout(

sidebarPanel(

selectInput("country", "Chọn quốc gia", choices = unique(sales\_data\_r$country), selected = unique(sales\_data\_r$country)[1])

),

mainPanel(

plotOutput("salesTrend"),

plotOutput("productSales"),

plotOutput("customerTrends")

)

)

)

**fluidPage()**: Tạo giao diện web linh hoạt, tự động điều chỉnh theo kích thước màn hình.

**titlePanel()**: Đặt tiêu đề chính cho ứng dụng.

**sidebarLayout()**: Chia giao diện thành hai phần: thanh bên (sidebar) và khung chính (main panel).

**selectInput()**: Tạo menu thả xuống cho phép người dùng chọn quốc gia cần phân tích.

**plotOutput()**: Tạo các khung cho biểu đồ.

* 1. *Xây dựng server.*

# Server của Shiny

server <- function(input, output, session) {

filtered\_data <- reactive({

sales\_data\_r %>% filter(country == input$country)

})

**reactive()**: Tạo một phản ứng động để lọc dữ liệu theo quốc gia được chọn.

* + 1. Doanh thu theo thời gian.

output$salesTrend <- renderPlot({

ggplot(filtered\_data(), aes(x = date, y = quantity \* price)) +

geom\_line(color = "blue") +

ggtitle(paste("Doanh thu theo thời gian -", input$country)) +

xlab("Ngày") + ylab("Doanh thu")

})

**renderPlot()**: Tạo biểu đồ tuyến thể hiện doanh thu theo thời gian.

**ggplot()**: Tạo biểu đồ với trục x là ngày và trục y là tổng doanh thu (quantity \* price).

* + 1. Phân tích sản phẩm bán chạy.

output$productSales <- renderPlot({

filtered\_data() %>%

group\_by(itemname) %>%

summarise(total\_sales = sum(quantity), .groups = 'drop') %>%

ggplot(aes(x = reorder(itemname, -total\_sales), y = total\_sales, fill = itemname)) +

geom\_bar(stat = "identity") +

coord\_flip() +

ggtitle(paste("Số lượng sản phẩm bán ra -", input$country)) +

xlab("Sản phẩm") + ylab("Số lượng") +

theme(legend.position = "none")

})

**group\_by()** và **summarise()**: Tổng hợp dữ liệu theo itemname và tính tổng số lượng bán ra.

**geom\_bar()**: Tạo biểu đồ cột thể hiện số lượng sản phẩm bán ra.

* + 1. Phân tích khách hàng.

output$customerTrends <- renderPlot({

filtered\_data() %>%

group\_by(customerid) %>%

summarise(total\_spent = sum(quantity \* price), purchase\_count = n(), .groups = 'drop') %>%

ggplot(aes(x = purchase\_count, y = total\_spent)) +

geom\_point(color = "red") +

ggtitle(paste("Xu hướng khách hàng -", input$country)) +

xlab("Số lần mua hàng") + ylab("Tổng chi tiêu")

})

**geom\_point()**: Tạo biểu đồ phân tán thể hiện mối quan hệ giữa số lần mua hàng và tổng chi tiêu.

* 1. *Ngắt kết nối Spark.*

onStop(function() {

spark\_disconnect(sc)

})

**onStop()**: Đảm bảo ngắt kết nối với Spark khi ứng dụng kết thúc.

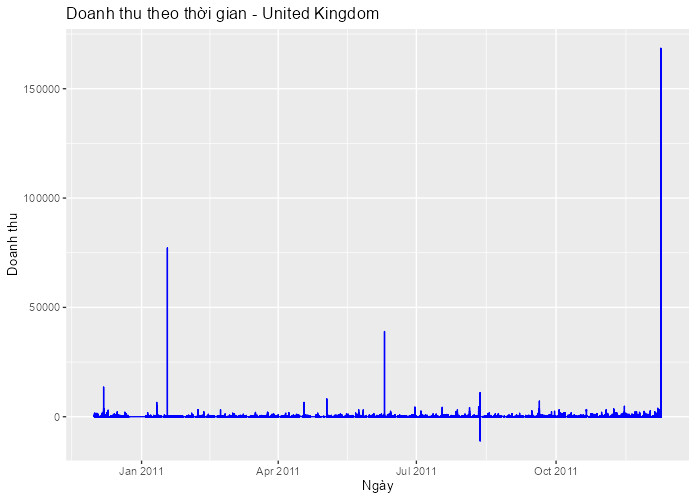
* 1. *Chạy ứng dụng.*

shinyApp(ui = ui, server = server)

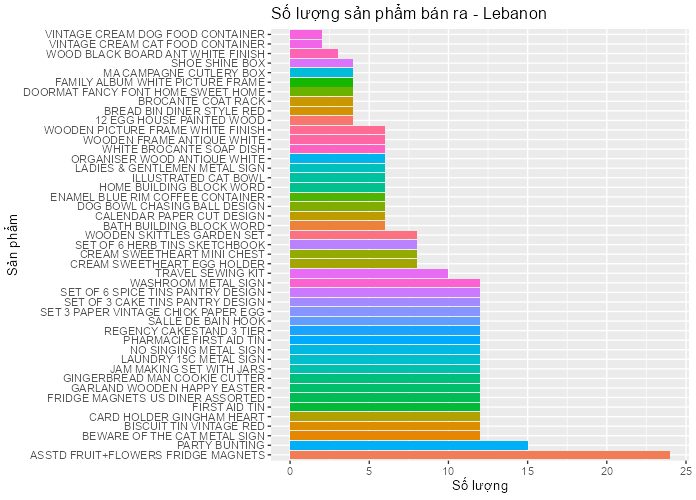
1. *Kết quả xử lý dữ liệu.*

Quá trình xử lý dữ liệu đã giúp loại bỏ các giá trị thiếu, chuẩn hóa các tên cột, và chuyển đổi kiểu dữ liệu phù hợp. Việc áp dụng Spark trong quá trình này đã tăng tốc độ xử lý, giúp xử lý một lượng dữ liệu lớn trong thời gian ngắn.

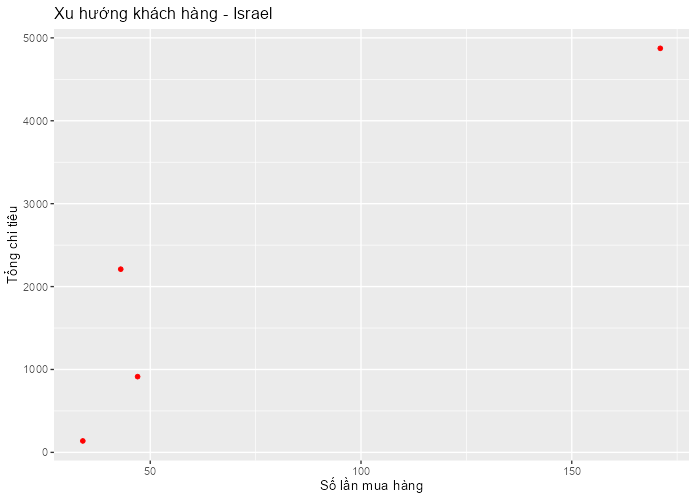
* Doanh thu theo thời gian:



* Số lượng sản phẩm bán ra:



* Xu hướng mua háng:



1. ***Tổng quan về bài toán.***

Bài này tập trung vào việc phân tích hành vi bán hàng bằng cách kết hợp các công cụ mạnh mẽ trong hệ sinh thái R và Apache Spark để xử lý dữ liệu lớn, sau đó trực quan hóa kết quả bằng ứng dụng Shiny. Các bước chính trong quá trình thực hiện bao gồm:

* 1. Mục tiêu:

Phân tích doanh thu theo thời gian.

Xác định các sản phẩm bán chạy.

Đánh giá hành vi mua sắm của khách hàng theo từng quốc gia.

* 1. Công cụ sử dụng:

Apache Spark: Xử lý dữ liệu lớn thông qua Sparklyr.

R: Xây dựng logic xử lý và trực quan hóa.

Shiny: Tạo giao diện web để hiển thị kết quả phân tích.

* 1. Quy trình thực hiện:

Tiền xử lý dữ liệu:

Đọc dữ liệu từ file Excel.

Loại bỏ các giá trị thiếu, chuẩn hóa tên cột.

* 1. Phân tích dữ liệu:

Tính toán tổng doanh thu theo ngày.

Xác định sản phẩm bán chạy nhất.

Đánh giá mức chi tiêu và tần suất mua sắm của khách hàng.

### Trực quan hóa:

Biểu đồ đường thể hiện doanh thu theo thời gian.

Biểu đồ cột cho thấy lượng sản phẩm bán ra.

Biểu đồ phân tán thể hiện mối quan hệ giữa tần suất mua hàng và tổng chi tiêu của khách hàng.

* 1. Ứng dụng:

Xây dựng giao diện web cho phép người dùng chọn quốc gia và xem kết quả phân tích theo thời gian thực.

Tạo môi trường tương tác trực quan, giúp dễ dàng nhận biết các xu hướng trong hành vi mua hàng.

**KẾT LUẬN**

*Qua quá trình xử lý và phân tích dữ liệu, bài nghiên cứu đã cho thấy tính hiệu quả của việc áp dụng Apache Spark trong việc xử lý các tập dữ liệu lớn, giúp giảm thiểu thời gian xử lý và tăng cường độ chính xác trong phân tích.*

*Các kết quả phân tích đã làm sáng tỏ những khía cạnh quan trọng trong hoạt động kinh doanh, bao gồm:*

* *Xác định xu hướng doanh thu theo thời gian, giúp nhận biết các giai đoạn cao điểm trong mua sắm.*
* *Phát hiện các sản phẩm bán chạy, từ đó hỗ trợ doanh nghiệp tối ưu hóa danh mục sản phẩm và quản lý tồn kho hiệu quả hơn.*
* *Phân tích hành vi khách hàng theo từng quốc gia, cho phép doanh nghiệp điều chỉnh chiến lược tiếp cận phù hợp với từng thị trường.*

*Bên cạnh đó, ứng dụng Shiny đã cung cấp một giao diện trực quan và thân thiện, giúp người dùng dễ dàng khám phá dữ liệu và theo dõi các chỉ số kinh doanh một cách trực quan.*

*Tóm lại, bài nghiên cứu không chỉ minh chứng khả năng xử lý dữ liệu lớn của Spark mà còn mang đến một công cụ hữu ích cho việc phân tích dữ liệu bán hàng, hỗ trợ doanh nghiệp đưa ra những quyết định kinh doanh chính xác và kịp thời. Đây là nền tảng quan trọng để phát triển các hệ thống phân tích dữ liệu lớn hơn trong tương lai.*

**DANH MỤC TÀI LIỆU THAM KHẢO**

Javier Luraschi, Kevin Kuo, Edgar Ruiz (2019). *Mastering Spark with R*. O'Reilly Media.

Hadley Wickham, Garrett Grolemund (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.

Apache Spark Documentation: Truy cập từ <https://spark.apache.org/>.

sparklyr Package Documentation: Truy cập từ https://spark.rstudio.com/.

Hadley Wickham (2021). *dplyr: A Grammar of Data Manipulation*. Truy cập từ https://dplyr.tidyverse.org/.

Winston Chang (2013). *R Graphics Cookbook: Practical Recipes for Visualizing Data*. O'Reilly Media.

Shiny Documentation: Truy cập từ https://shiny.rstudio.com/.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Truy cập từ <https://www.r-project.org/>.

Kaggle Datasets: Nguồn dữ liệu tham khảo từ cộng đồng Kaggle tại <https://www.kaggle.com/>.

****