**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH**

**TẠ THỊ THIÊN THANH**

**<TÊN ĐỒ ÁN NGÀNH>**

**ĐỒ ÁN NGÀNH**

**NGÀNH KHOA HỌC MÁY TÍNH**

**TP. HỒ CHÍ MINH, 2024BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH**

**TẠ THỊ THIÊN THANH**

**< TÊN ĐỒ ÁN NGÀNH >**

**Mã số sinh viên: 2151013088**

**ĐỒ ÁN NGÀNH**

**NGÀNH KHOA HỌC MÁY TÍNH**

**Giảng viên hướng dẫn: ThS. PHẠM CHÍ CÔNG**

**TP. HỒ CHÍ MINH, 2024**

LỜI CẢM ƠN

NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

TÓM TẮT ĐỐ ÁN NGÀNH

ABSTRACT

**MỤC LỤC**

[LỜI CẢM ƠN 1](#_Toc176972785)

[NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN 2](#_Toc176972786)

[TÓM TẮT ĐỐ ÁN NGÀNH 3](#_Toc176972787)

[ABSTRACT 4](#_Toc176972788)

[DANH MỤC TỪ VIẾT TẮT 7](#_Toc176972789)

[DANH MỤC HÌNH ẢNH 8](#_Toc176972790)

[DANH MỤC BẢNG 9](#_Toc176972791)

[MỞ ĐẦU 10](#_Toc176972792)

[Chương 1. TỒNG QUAN VỀ ĐỀ TÀI 11](#_Toc176972793)

[1.1. Mục tiêu của đề tài 11](#_Toc176972794)

[1.1.1. Chủ đề cấp độ 3 11](#_Toc176972795)

[1.1.2. Chủ đề cấp độ 3 11](#_Toc176972796)

[1.2. Phạm vi đề tài 11](#_Toc176972797)

[1.3. Tổng quan về hệ thống gợi ý sản phẩm 11](#_Toc176972798)

[1.4. Các công nghệ và lý do chọn công nghệ 12](#_Toc176972799)

[1.4.1. Thư viện Selenium 12](#_Toc176972800)

[1.4.2. Thư viện NLTK 12](#_Toc176972801)

[1.4.3. Thư viện Pandas 13](#_Toc176972802)

[1.4.4. Thư viện Matplotlib 13](#_Toc176972803)

[1.4.5. Thư viện Seaborn 14](#_Toc176972804)

[1.5. Các bước thực hiện đề tài 14](#_Toc176972805)

[1.6. Thông tin về nguồn dữ liệu 14](#_Toc176972806)

[1.6.1. Lazada 14](#_Toc176972807)

[1.7. Mô tả dữ liệu 15](#_Toc176972808)

[1.7.1. Các sản phẩm của Lazada 15](#_Toc176972809)

[1.8. Trực quan hóa dữ liệu 16](#_Toc176972810)

[1.8.1. Các sản phẩm của Lazada 16](#_Toc176972811)

[1.8.2. Các đánh giá sản phẩm của Lazada 16](#_Toc176972812)

[1.9. Lưu trữ dữ liệu 16](#_Toc176972813)

[Chương 2. XÂY DỰNG HỆ THỐNG GỢI Ý SẢN PHẨM 17](#_Toc176972814)

[2.1. Chủ đề cấp độ 2 17](#_Toc176972815)

[2.1.1. Chủ đề cấp độ 3 17](#_Toc176972816)

[2.2. Chủ đề cấp độ 2 17](#_Toc176972817)

[2.2.1. Chủ đề cấp độ 3 17](#_Toc176972818)

[Chương 3. TÊN CHƯƠNG 3 18](#_Toc176972819)

[3.1. Chủ đề cấp độ 2 18](#_Toc176972820)

[3.1.1. Chủ đề cấp độ 3 18](#_Toc176972821)

[3.2. Chủ đề cấp độ 2 18](#_Toc176972822)

[TÀI LIỆU THAM KHẢO 19](#_Toc176972823)

[PHỤ LỤC 20](#_Toc176972824)

DANH MỤC TỪ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| TỪ VIẾT TẮT | CỤM TỪ ĐẦY ĐỦ | Ý NGHĨA |
| NLP | Natural Language Processing | Xử lý ngôn ngữ tự nhiên |
| CSV |  |  |
| DBSCAN |  |  |

DANH MỤC HÌNH ẢNH

[Hình 1.1 Logo Selenium 12](#_Toc177393549)

[Hình 1.2 Logo Pandas 13](#_Toc177393550)

[Hình 1.3 Logo Matplotlib 13](#_Toc177393551)

[Hình 1.4 Logo Seaborn 14](#_Toc177393552)

[Hình 1.5 Logo Lazada 14](#_Toc177393553)

[Hình 1.6 Các giá trị trong ‘categories\_keywords’ 16](#_Toc177393554)

[Hình 1.7 Biểu đồ cột các thông số thống kê giá tiền sản phẩm trên Lazada 16](#_Toc177393555)

[Hình 1.8 Biểu đồ box plot biểu thị sự phân phối giá tiền sản phẩm trên Lazada 17](#_Toc177393556)

[Hình 1.9 Biểu đồ phân cụm bằng DBSCAN theo giá tiền sản phẩm trên Lazada 18](#_Toc177393557)

[Hình 1.10 Số lượng sản phẩm ở mỗi loại sản phẩm quần áo nữ trên Lazada 18](#_Toc177393558)

[Hình 1.11 Giá tiền trung bình mỗi loại sản phẩm quần áo nữ trên Lazada 19](#_Toc177393559)

DANH MỤC BẢNG

[Bảng 1.1: Tên bảng 1 10](#_Toc367742567)

[Bảng 2.1: Tên bảng 1 11](#_Toc367742568)

MỞ ĐẦU

# TỒNG QUAN VỀ ĐỀ TÀI

## Mục tiêu của đề tài

Nội dung …………………

Nội dung………………….

### Chủ đề cấp độ 3

Nội dung

### Chủ đề cấp độ 3

#### Chủ đề cấp độ 4

Nội dung

Hình 1.1: Tên hình 1

Bảng 1.1: Tên bảng 1

## Phạm vi đề tài

Đề tài tập trung vào việc phát triển một hệ thống hỗ trợ người dùng trong tìm kiếm và lựa chọn sản phẩm phù hợp. Đề tài bao gồm các bước thu thập và xử lý dữ liệu từ các nền tảng thương mại điện tử phổ biến ở Việt Nam có bán các mặt hàng thời trang nữ đồng thời phân tích hành vi mua sắm và đặc điểm các sản phẩm có ở những nền tảng này mà cụ thể ở đây là Lazada, Shein và Tiki. Các nguồn dữ liệu sẽ được chọn lọc và kiểm tra kỹ lưỡng để đảm bảo độ tin cậy và đa dạng. Dựa trên những gì phân tích được, hệ thống sẽ phát triển các mô hình máy học và thuật toán gợi ý thích hợp.

## Tổng quan về hệ thống gợi ý sản phẩm

Nội dung ...............................

## Các công nghệ và lý do chọn công nghệ

### Thư viện Selenium



Hình 1.1 Logo Selenium

Selenium là một thư viện mã nguồn mở, là một công cụ tự động hóa tương tác với trình duyệt web với hai chức năng chính là tương tác với trình duyệt và kiểm thử ứng dụng web. Selenium hỗ trợ nhiều ngôn ngữ như Python, C#, Java, ...

Trong đề tài này, ta sử dụng trình duyệt Chrome và ngôn ngữ Python để tương tác và thu thập dữ liệu từ các nền tảng thương mại điện tử. Với Python, ta có thể cài đặt Selenium bằng pip `pip install selenium`. Để thu thập dữ liệu từ trình duyệt, ta cần tải xuống driver tương ứng với trình duyệt (ở đề tài này là [ChromeDriver](https://github.com/tathithienthanh/MajorProject_/tree/main/chromedriver_win32)) và cài đặt thư viện webdriver-manager (có thể cài đặt bằng pip `pip install webdriver-manager`).

Lợi thế của Selenium so với các thư viện với chức năng tương tự là Selenium hỗ trợ đa ngôn ngữ và nhiều trình duyệt; là công cụ mã nguồn mở và miễn phí phù hợp với nhiều đối tượng mà ở đây cụ thể là sinh viên đồng thời cho phép cộng đồng cải tiến liên tục; vì có một cộng đồng người dùng lớn, việc tìm hiểu cách sử dụng thông qua các hướng dẫn, mã code ví dụ và những hỗ trợ từ cộng đồng sẽ dễ dàng hơn.

### Thư viện NLTK

NLTK (Natural Language Toolkit) là một thư viện mã nguồn mở, là công cụ phổ biến trong lĩnh vực NLP bằng ngôn ngữ Python. Là một nền tảng làm việc với dữ liệu là ngôn ngữ con người, NLTK có các chức năng chính: xử lý văn bản, phân tích câu trúc từ vựng và ngữ pháp, phân tích cảm xúc, … Ta có thể cài đặt thư viện NLTK thông qua pip `pip install nltk`, tải các tài nguyên và dữ liệu cần thiết có sẵn của thư viện bằng câu lệnh `nltk.download()`. Vì NLTK chủ yếu hỗ trợ tiếng Anh nên trong quá trình xử lý tiếng Việt cần cài đặt thêm thư viện Unicodedata để đảm bảo các thao tác NLP được nhất quán và chính xác. Có thể cài đặt thư viện Unicodedata bằng pip `pip install unicodedata`.

Điểm mạnh của NLTK là lượng tài nguyên phong phú và đầy đủ các công cụ phục vụ các nhiệm vụ NLP từ cơ bản đến nâng cao; khả năng tích hợp với những thư viện khác như Pandas trong xử lý dữ liệu, Matplotlib trong trực quan hóa dữ liệu, …; đồng thời vì là thư viện mã nguồn mở, được sử dụng và hỗ trợ bởi số lượng lớn cộng đồng người dùng nên giúp dễ dàng tham khảo các giáo trình hướng dẫn, các mã code có sẵn miễn phí và các diễn đàn lớn từ cộng đồng.

### Thư viện Pandas

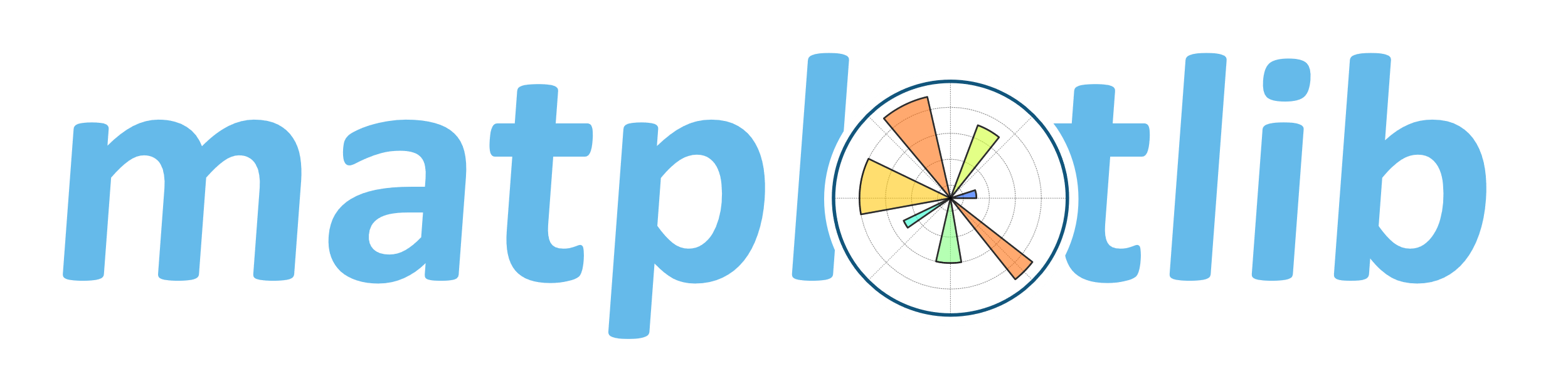


Hình 1.2 Logo Pandas

Pandas là một thư viện mã nguồn mở chuyên dùng để xử lý và phân tích dữ liệu. Với khả năng nổi bật là làm việc với dữ liệu dạng bảng, Pandas là một công cụ xử lý dữ liệu tốt với đầy đủ các tính năng lọc, tổng hợp, xử lý thiếu, nhiễu, lọc, … dữ liệu. Có thể cài đặt Pandas bằng pip `pip install pandas`.

Ưu điểm của Pandas là khả năng linh hoạt và tích hợp tốt với các thư viện khác giúp hỗ trợ tốt hơn trong thao tác với dữ liệu. Thư viện này cũng là mã nguồn mở và được hỗ trợ bởi cộng đồng lớn, cung cấp nhiều tài liệu hướng dẫn và ví dụ mã, giúp dễ dàng tìm kiếm và giải quyết các vấn đề liên quan đến dữ liệu.

### Thư viện Matplotlib



Hình 1.3 Logo Matplotlib

Matplotlib là một thư viện mã nguồn mở được sử dụng để tạo các biểu đồ và hình ảnh hóa, trực quan hóa dữ liệu. Matplotlib phổ biển nhờ việc hỗ trợ nhiều loại biểu đồ, giúp người dùng dễ dàng tạo ra các loại biểu đồ một cách trực quan và chuyên nghiệp. Có thể cài đặt Matplotlib bằng pip `pip install matplotlib`.

Hỗ trợ nhiều kiểu dáng, màu sắc và khả năng tùy chỉnh các chú thích, tiêu đề, trục, … Matplotlib giúp người dùng dễ dàng cá nhân hóa cũng như điều chỉnh các biểu đồ theo nhu cầu cụ thể. Hơn nữa khả năng tích hợp tốt với các thư viện khác như Numpy, [Pandas](#_Thư_viện_Pandas) giúp quá trình xử lý và hình ảnh hóa dữ liệu trở nên dễ dàng hơn.

### Thư viện Seaborn



Hình 1.4 Logo Seaborn

Seaborn là một thư viện mã nguồn mở được xây dựng dựa trên [Matplotlib](#_Thư_viện_Matplotlib), chuyên dùng để tạo các biểu đồ và trực quan hóa dữ liệu. Seaborn phổ biến nhờ việc cung cấp các biểu đồ thống kê nâng cao và trực quan hơn, giúp người dùng dễ dàng tạo ra các biểu đồ đẹp mắt và giàu thông tin. Có thể cài đặt Seaborn bằng pip `pip install seaborn`.

Seaborn hỗ trợ nhiều loại biểu đồ trực quan như biểu đồ phân tán, biểu đồ tương quan, biểu đồ hộp (box plot), và biểu đồ mật độ, với khả năng tùy chỉnh cao về màu sắc, kích thước và chú thích. Thư viện này giúp người dùng dễ dàng tạo ra các biểu đồ phức tạp chỉ với vài dòng mã. Seaborn tích hợp tốt với [Pandas](#_Thư_viện_Pandas) và Numpy, giúp việc xử lý và trực quan hóa dữ liệu từ các tập dữ liệu trở nên mượt mà và hiệu quả hơn.

### Thư viện Scikit-learn

## Các bước thực hiện đề tài

Nội dung ...............................

## Thông tin về nguồn dữ liệu

### Lazada



Hình 1.5 Logo Lazada

Lazada là một trong những công ty điều hành thương mại điện tử lớn nhất Đông Nam Á. Thành lập từ năm 2012 bởi Maximilian Bittner, Lazada cho phép các nhà bán lẻ và bên thứ ba bán sản phẩm trên nền tảng của mình. Lazada hoạt động tại nhiều quốc gia trong khu vực, trong đó có Việt Nam.

Shein

Tiki

## Mô tả dữ liệu

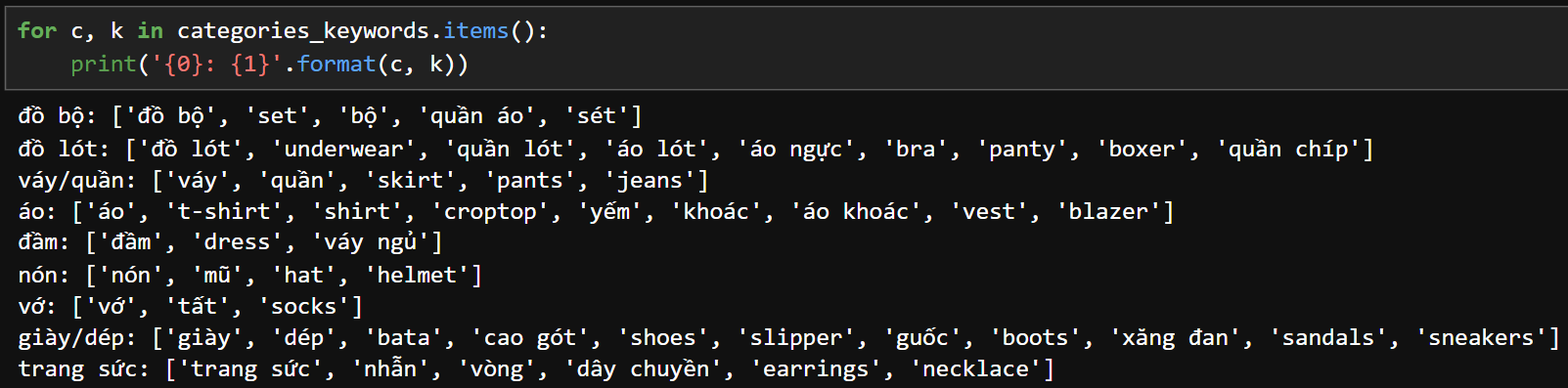
Dữ liệu thu thập được qua quá trình trích xuất dữ liệu tự động từ trang web bằng thư viện [Selenium](#_Thư_viện_Selenium) sẽ được lưu trữ dưới dạng file text (\*.txt) để backup dữ liệu và file CSV (\*.csv) để dễ dàng chuyển đổi thành dạng DataFrame trong quá trình xử lý và phân tích. Bao gồm:

* [Laz\_product\_urls.txt](https://github.com/tathithienthanh/MajorProject_/blob/main/Laz_shop_urls.txt)
* [lazada\_products.csv](https://github.com/tathithienthanh/MajorProject_/blob/main/lazada_products.csv)
* [lazada\_feedbacks.csv](https://github.com/tathithienthanh/MajorProject_/blob/main/lazada_feedbacks.csv)

### Các sản phẩm của Lazada

Bộ dữ liệu gồm 120 sản phẩm với từ khóa ‘Quần áo nữ’. Gồm các 7 trường:

* Url: đường dẫn của từng sản phẩm. (string)
* Name: tên sản phẩm. (string)
* Price: đơn giá mỗi sản phẩm được rao bán trên sàn Lazada tại thời điểm thu thập dữ liệu. (float)
* Voucher: phiếu giảm giá trên mỗi sản phẩm tại thời điểm thu thập dữ liệu. (string)
* Sold: Số lượng sản phẩm đã bán tại thời điểm thu thập dữ liệu. (int)
* NumRate: Số lượng những đánh giá của khách hàng tại thời điểm thu thập dữ liệu. (int)
* Category: Loại sản phẩm được phân loại bằng phương pháp n-grams dựa trên trường ‘Name’ và quy ước được quy định sẵn được lưu trong biến ‘*categories\_keywords*’ bằng kiểu dữ liệu ‘dict’. (string)



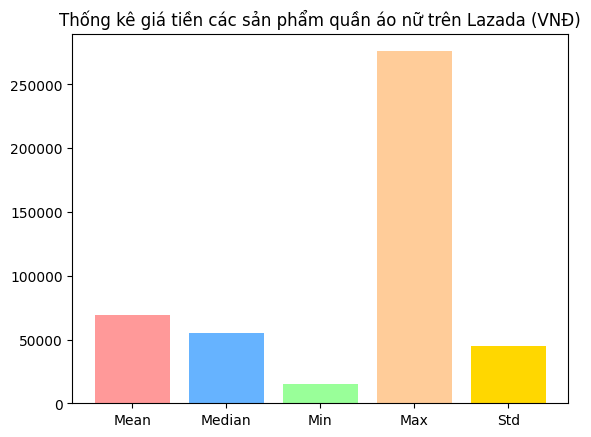
Hình 1.6 Các giá trị trong ‘categories\_keywords’

### Các đánh giá sản phẩm của Lazada

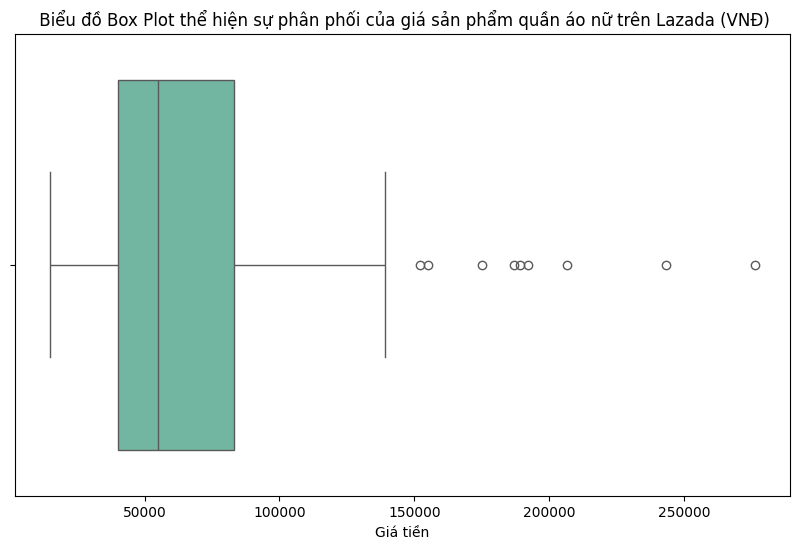
Nội dung ...............................

## Trực quan hóa dữ liệu

### Các sản phẩm của Lazada



Hình 1.7 Biểu đồ cột các thông số thống kê giá tiền sản phẩm trên Lazada

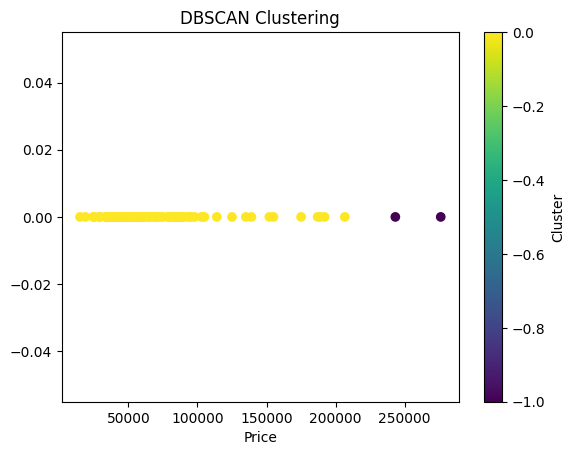


Hình 1.8 Biểu đồ box plot biểu thị sự phân phối giá tiền sản phẩm trên Lazada

Giá tiền cao nhất cho sản phẩm quần áo nữ là 276.051 VNĐ và thấp nhất là 15.000 VNĐ. Điều này cho thấy sự chênh lệch giá của các sản phẩm khá lớn.

Giá trung bình của sản phẩm là khoảng 68.981 VNĐ, trong khi độ lệch chuẩn là 44.870 VNĐ. Tỷ lệ giữa độ lệch chuẩn và giá trị trung bình đạt khoảng 0.65, cho thấy mức độ phân tán của dữ liệu là tương đối cao. Điều này chỉ ra rằng dữ liệu có sự biến động đáng kể. Hơn nữa, như đã trình bày trong Hình 1.8, trường dữ liệu ‘Price’ có khá nhiều giá trị ngoại lai.

Sự xuất hiện nhiều giá trị ngoại lai trong giá tiền của các sản phẩm quần áo nữ cho thấy sự đa dạng của sản phẩm được bán trong nền tảng này. Điều này có thể được lý giải bởi sự khác biệt về thương hiệu, chất liệu, kiểu dáng, và phân khúc khách hàng mà mỗi sản phẩm hướng đến. Sự đa dạng này tạo điều kiện cho người tiêu dùng có nhiều lựa chọn, từ các sản phẩm bình dân đến hàng cao cấp, phục vụ nhu cầu và sở thích đa dạng của thị trường.

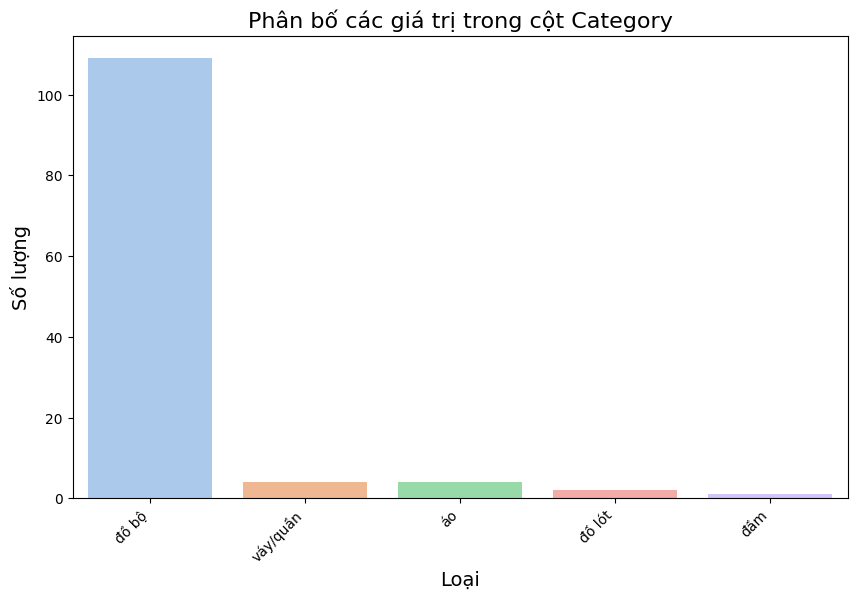


Hình 1.9 Biểu đồ phân cụm bằng DBSCAN theo giá tiền sản phẩm trên Lazada

Do dữ liệu giá tiền chứa nhiều giá trị ngoại lai và mục tiêu là phân tích và phân khúc khách hàng dựa trên giá tiền (với số lượng cụm chưa được xác định trước), thuật toán DBSCAN là sự lựa chọn phù hợp để thực hiện phân cụm.

Sau khi phân cụm bằng DBSCAN, ta có 2 cụm: cụm 1 (màu vàng) với giá tiền nằm dưới 207.000 VNĐ chiếm đa số và cụm 2 (màu tím) với những giá trị còn lại.

Tuy nhiên, dữ liệu tập trung chủ yếu ở mức giá dưới 150.000 VNĐ nên nếu chỉ xét theo giá tiền ta có 2 phân khúc khách hàng chính trên nền tảng Lazada là phân khúc giá thấp (dưới 150.000 VNĐ) và phân khúc giá cao (trên 150.000 VNĐ).

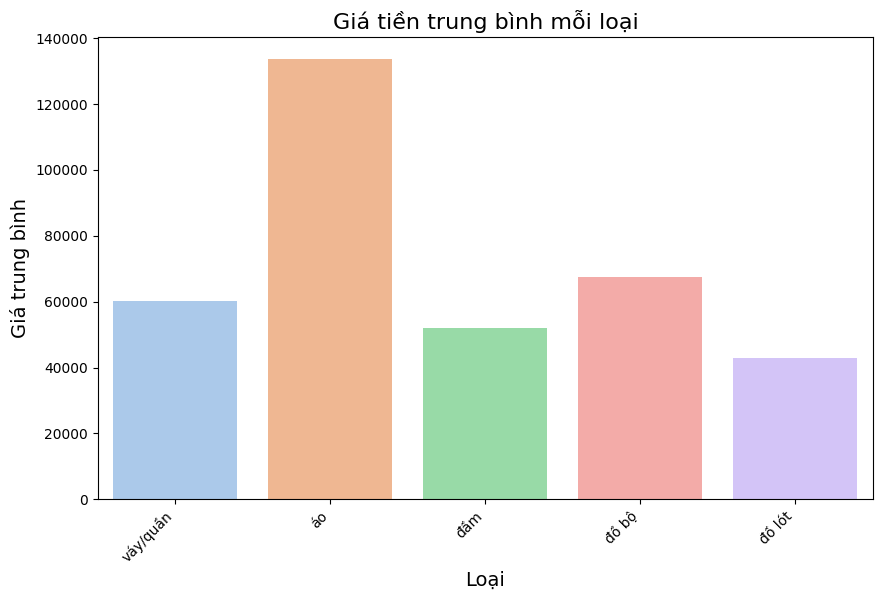


Hình 1.10 Số lượng sản phẩm ở mỗi loại sản phẩm quần áo nữ trên Lazada

Giá trị trong cột ‘Category’ được phân loại từ các giá trị trong cột ‘Name’ bằng phương pháp n-grams với chuẩn được quy định sẵn (Hình 1.6). Do việc xử lý NLP Tiếng Việt chưa được thực hiện một cách toàn diện, các giá trị không thể phân loại bằng phương pháp n-grams đã được xử lý thủ công.

Sau khi phân loại, trường ‘Category’ trong bộ dữ liệu các sản phẩm của Lazada có 5 loại: đồ bộ (109 sản phẩm), váy/quần (4 sản phẩm), áo (4 sản phẩm), đồ lót (2 sản phẩm) và đầm (1 sản phẩm).

Từ Hình 1.10, có thể thấy loại ‘đồ bộ’ chiếm đa số và chiếm số lượng quá lớn so với các loại quần áo nữ còn lại. Cho thấy xu hướng đăng bán sản phẩm trên nền tảng Lazada tập trung vào các mặt hàng đồ bộ như đồ ngủ và các set đồ đã phối sẵn nhằm thuận tiện hơn cho khách hàng và tăng độ đa dạng trong chọn lựa sản phẩm.



Hình 1.11 Giá tiền trung bình mỗi loại sản phẩm quần áo nữ trên Lazada



Hình 1.12 Các sản phẩm loại áo trên Lazada

Đối với áo, vì số lượng sản phẩm trên tổng số sản phẩm là quá thấp (4/120 sản phẩm) và giá tiền các sản phẩm trải đều ở cả hai phân khúc giá cao và giá thấp theo giá tiền như đã phân tích ở trên nên giá tiền trung bình của sản phẩm không phản ánh được ý nghĩa thực tế của sản phẩm áo trên nền tảng này.

Tương tự với các loại sản phẩm váy/quần, đầm và đồ lót, vì số lượng sản phẩm trên tổng số là quá ít nên ở đây ta chỉ tập trung phân tích loại sản phẩm đồ bộ.

### Các đánh giá sản phẩm của Lazada

## Lưu trữ dữ liệu

Nội dung ...............................

# XÂY DỰNG HỆ THỐNG GỢI Ý SẢN PHẨM

## Chủ đề cấp độ 2

### Chủ đề cấp độ 3

#### Chủ đề cấp độ 4

Bảng 2.1: Tên bảng 1

## Chủ đề cấp độ 2

### Chủ đề cấp độ 3

# TÊN CHƯƠNG 3

## Chủ đề cấp độ 2

Nội dung …………………

Nội dung………………….

### Chủ đề cấp độ 3

#### Chủ đề cấp độ 4

## Chủ đề cấp độ 2

TÀI LIỆU THAM KHẢO

NLTK Team, https://www.nltk.org/

PHỤ LỤC