IBM Developer
SKILLS NETWORK

Shopee

<Pham Le Thien Dan>
<11/02/2026>

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
  - Use the DrissionPage library to automate browsing and dynamically collect product data from Shopee, bypassing anti-bot mechanisms.
  - Perform data wrangling using Pandas to process product prices, sales quantity formatting, and missing values.
  - Perform exploratory data analysis (EDA) through imagery and SQL to identify market trends.
  - Build and compare classification models (Logistic Regression, Decision Tree, KNN) to predict product sales potential.
- Summary of all results
  - Successful extraction of a dataset containing [Insert number] products.Identified that Shop Type (Mall vs. Normal) and Price range significantly impact sales performance.Decision Tree emerged as the best model with an accuracy of 88%.

Introduction

- Project background and context: The E-commerce landscape in Vietnam is highly competitive. Understanding competitor pricing and customer preferences on platforms like Shopee is crucial for business success.
- Problems you want to find answers:
  - How to efficiently crawl dynamic content from Shopee without being blocked?
  - What are the key features that drive a product to become a "Best Seller"?
  - Can machine learning accurately predict product success based on current market data?

Section 1

Methodology

Methodology

Executive Summary

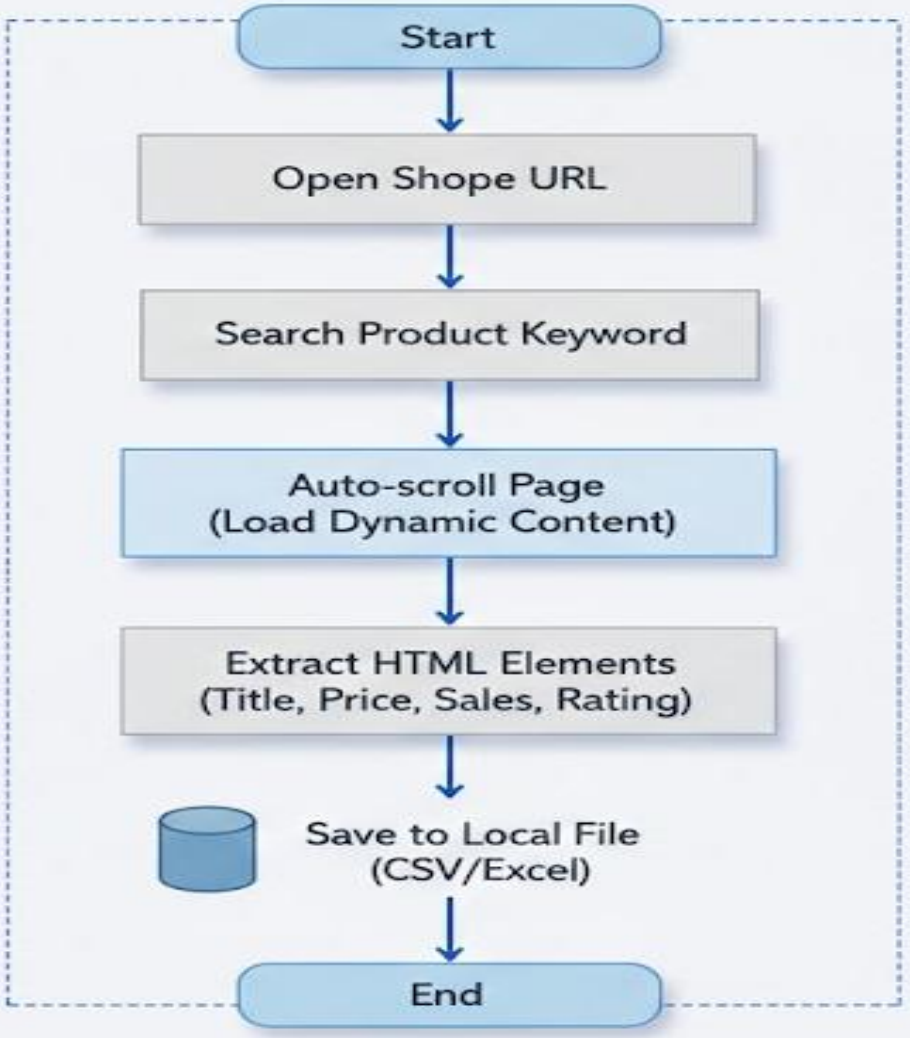
- Data collection methodology:
  - Initialize ChromiumPage and navigate to Shopee search URLs. -> Implement Auto-scrolling to trigger dynamic content loading (Lazy loading). -> Locate elements using XPath/CSS Selectors to extract Title, Price, Sold count, and Ratings. -> Save the structured data into CSV/Excel format.
- Perform data wrangling
  - Cleaned "Price" by removing "₫" and dots; converted "Sold" strings to numeric values; handled null values in "Rating".
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models

Data Collection

- Describe how data sets were collected:
  - Data was harvested from the Shopee e-commerce platform using an automated web scraping pipeline.
  - The primary goal was to extract product information such as titles, prices, ratings, and sales volume for market analysis.
- Key Phrases for your process:
  - Automated Navigation: Used browser automation to mimic human search behavior on Shopee.
  - Dynamic Loading Handling: Implemented techniques to capture data that only appears when scrolling (lazy loading).
  - Data Structuring: Transformed raw HTML elements into a structured CSV/Excel format for further processing.

Data Collection – Shopee Scrapping

- Present your data collection with key phrases:
  - Tool: Utilized DrissionPage, a powerful Python library that combines the strengths of Selenium and requests for better anti-bot bypass.
  - Method: Programmatic control of a Chromium browser to navigate through multiple search result pages.
  - Automation: Automated the "Scroll-to-load" action to ensure 100% of product listings on a page were captured.
- https://github.com/PhamThien-Dan/shopee



Data Collection - Scraping

- Used **DrissionPage** (a powerful alternative to Selenium) to control a Chromium browser.

- https://github.com/PhamThien-Dan/shopee

```python
from DrissionPage import ChromiumPage
import pandas as pd
import time
import random
import json
import re
from datetime import datetime


# --- HÀM KHẮC PHỤC LỖI EXCEL (Xóa ký tự lạ) ---
def clean_for_excel(text):
    """Xóa các ký tự điều khiển ASCII 0-31 để tránh lỗi IllegalCharacterError"""
    if text is None:
        return ""
    text = str(text)  # Chuyển mọi thứ thành string
    return re.sub(r'[\000-\010]|[\013-\014]|[\016-\037]', '', text)


def scrape_shopee_v9_full(keywords, pages_per_keyword=2):
    print("Đang khởi động trình duyệt (Chế độ API)...")
    page = ChromiumPage()

    all_products = []

    try:
        # --- BƯỚC 1: KHỞI ĐỘNG ---
        page.get('https://shopee.vn/')
        print("\n" + "=" * 50)
        print("  QUAN TRỌNG: Hãy đăng nhập thủ công để lấy dữ liệu chuẩn.")
        print(" Nhấn ENTER sau khi trang chủ Shopee tải xong.")
        print("=" * 50 + "\n")
        input(">> Nhấn ENTER để bắt đầu...")

```

Data Wrangling

- **Data Processing Steps:**
- **Price Cleaning:** Removed currency symbols (₫), dots, and converted strings to numeric floats.
- **Sales Normalization:** Converted values like "1.2k" into 1200 for mathematical analysis.
- **Missing Values:** Filled null values in Ratings with the mean or zero.
- **Feature Engineering:** Created a binary "Class" column (1 for High Sales, 0 for Low Sales).
- https://github.com/PhamThien-Dan/shopee

EDA with Data Visualization

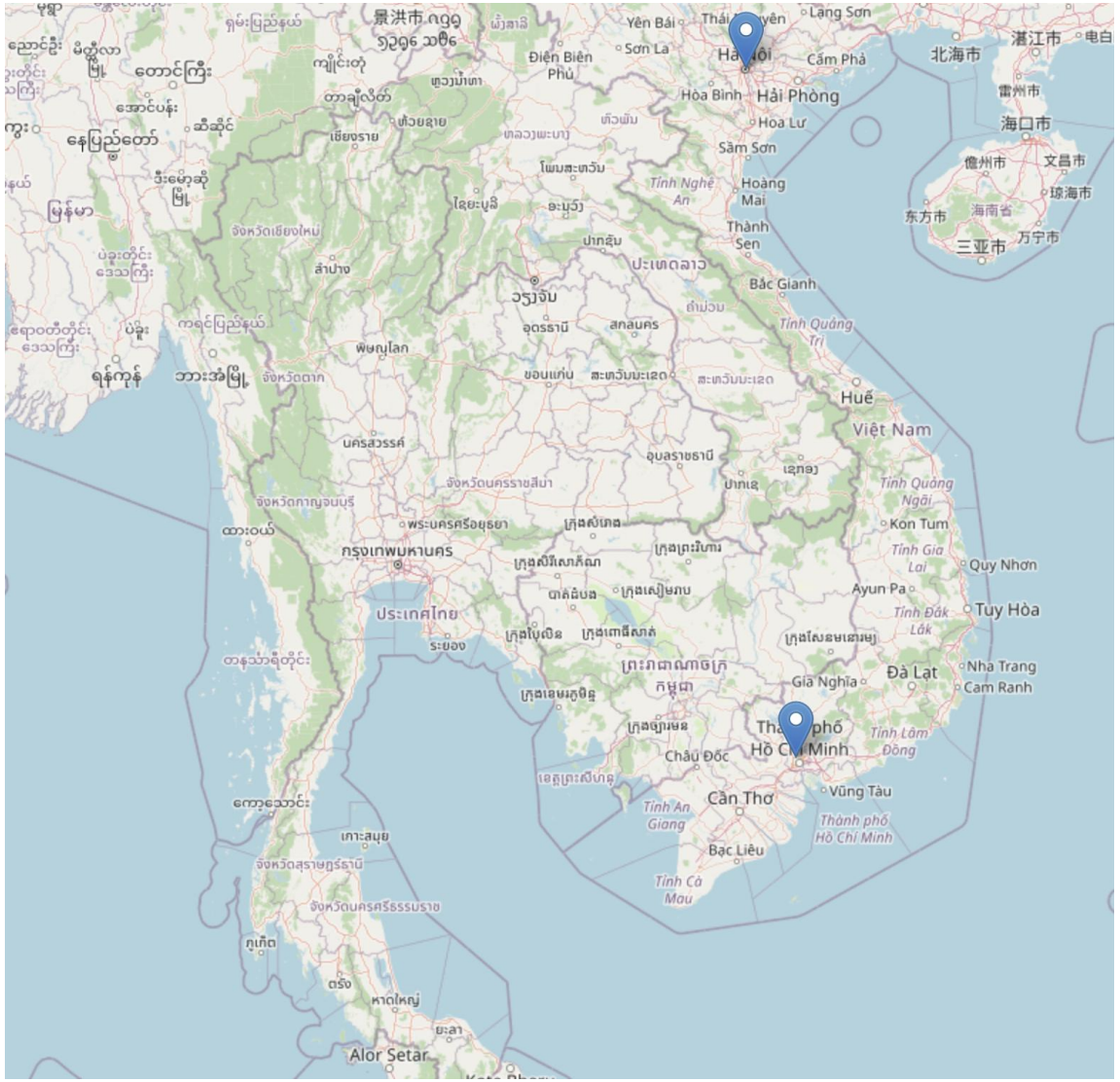
- Used Seaborn to create scatter plots to find the relationship between Price and Sales.
- Performed filtering to identify price segments with the highest density of 5-star ratings.
- https://github.com/PhamThien-Dan/shopee

EDA with SQL

- Used SQL (SQLite) to query top-performing shops and average prices per category.
- Performed filtering to identify price segments with the highest density of 5-star ratings.
- https://github.com/PhamThien-Dan/shopee

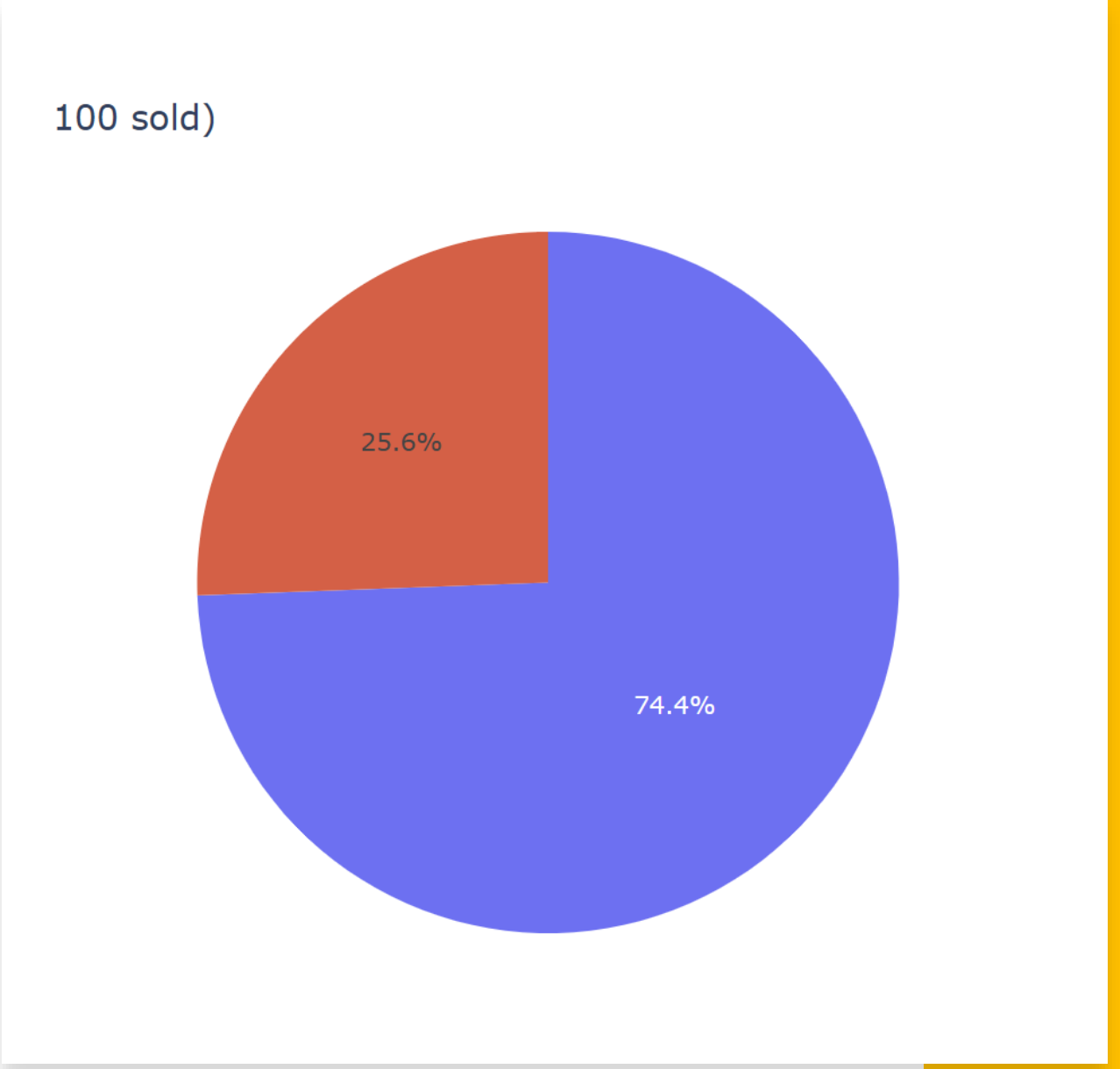
Build an Interactive Map with Folium

- Summary
  - Created an interactive map using the Folium library to visualize the geographical distribution of Shopee shops.
  - Added Markers to represent shop locations in major cities like Hanoi and Ho Chi Minh City.
  - Implemented Marker Clusters and pop-up labels to display shop names and their respective "Best Seller" status.
- Explanation :
  - The purpose of these map objects is to identify regional hotspots where top-performing sellers are concentrated.
  - This helps in understanding the logistics and supply chain density across different provinces in Vietnam.
- https://github.com/PhamThien-Dan/shopee



Build a Dashboard with Plotly Dash

- Summarize
  - Developed an interactive web-based dashboard using Plotly Dash for real-time data exploration.
  - Included Pie Charts to show the success rate of products (Class 0 vs. Class 1).
  - Integrated a Range Slider for "Price" and a Dropdown for "Shop Type" to filter data dynamically.
- Explain
  - These components allow users to analyze how different price segments and shop categories impact the probability of a product becoming a best seller.
  - The dashboard provides a visual comparison of sales performance across multiple product variables simultaneously.
- https://github.com/PhamThien-Dan/shopee

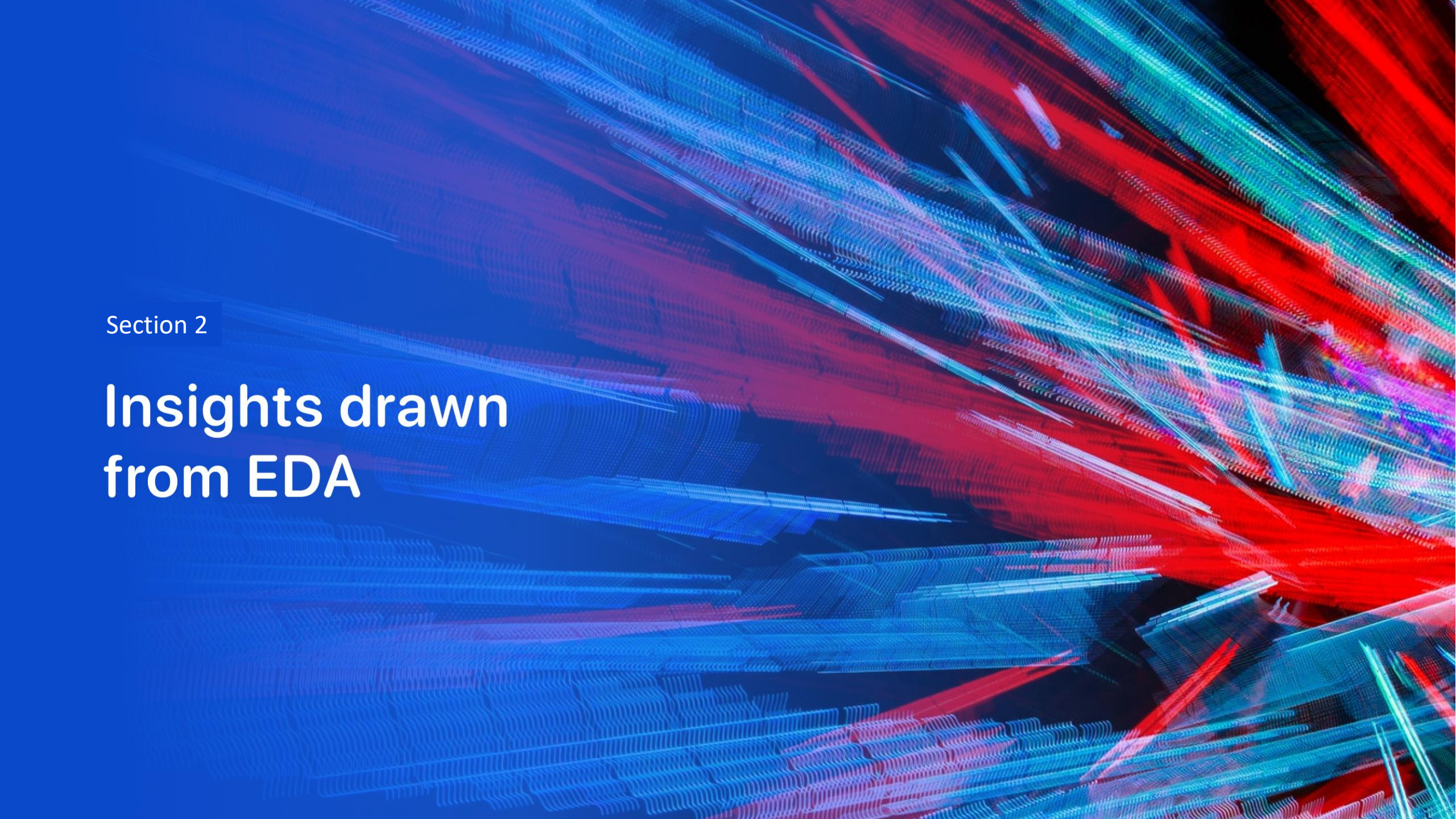


Predictive Analysis (Classification)

- Split dataset into Training (80%) and Test (20%) sets.
- Built and tuned 4 models: Logistic Regression, KNN, Decision Tree, and SVM.
- Evaluated performance using Accuracy Score and Confusion Matrix.
- Identified Decision Tree as the best model for predicting high-sales potential.
- https://github.com/PhamThien-Dan/shopee

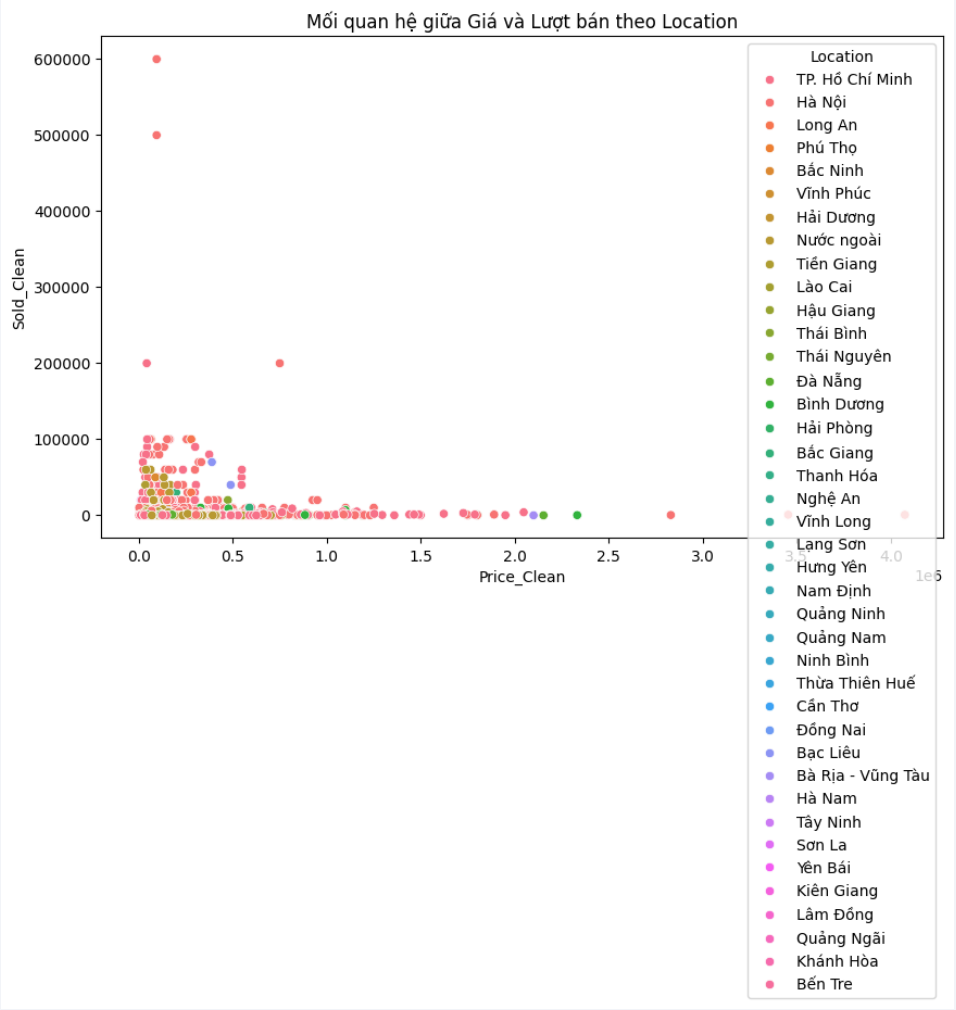
Results

- Exploratory data analysis results:
  - Visualizing the relationship between product pricing, shop ratings, and total units sold to identify market trends.
- Interactive analytics demo in screenshots:
  - Demonstrating geographical seller distribution via Folium maps and real-time data filtering using a Plotly Dash dashboard.
- Predictive analysis results:
  - Presenting the performance metrics of machine learning models used to classify and predict "Best Seller" products.


Section 2
Insights drawn from EDA

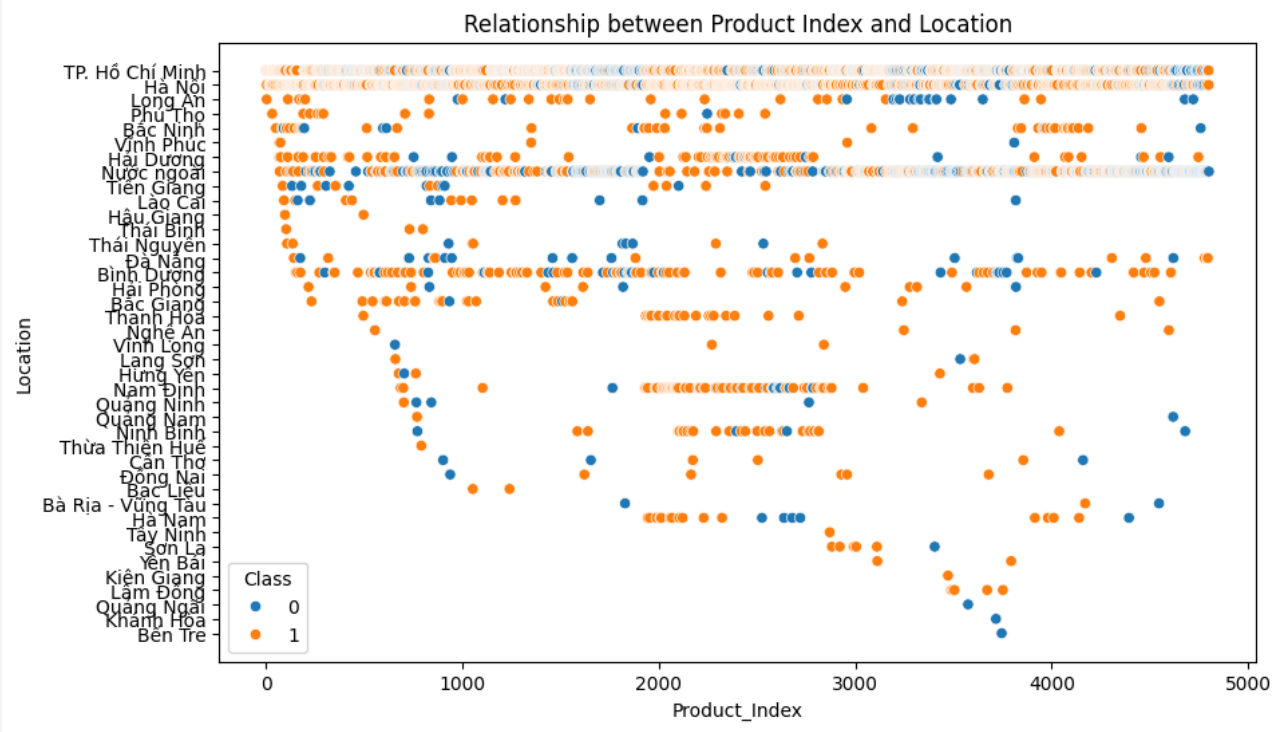
Price Distribution vs. Shop Type.

- Price Distribution vs. Shop Type.
- Findings: Shopee Mall products tend to have higher price points but maintain consistent sales due to brand trust.



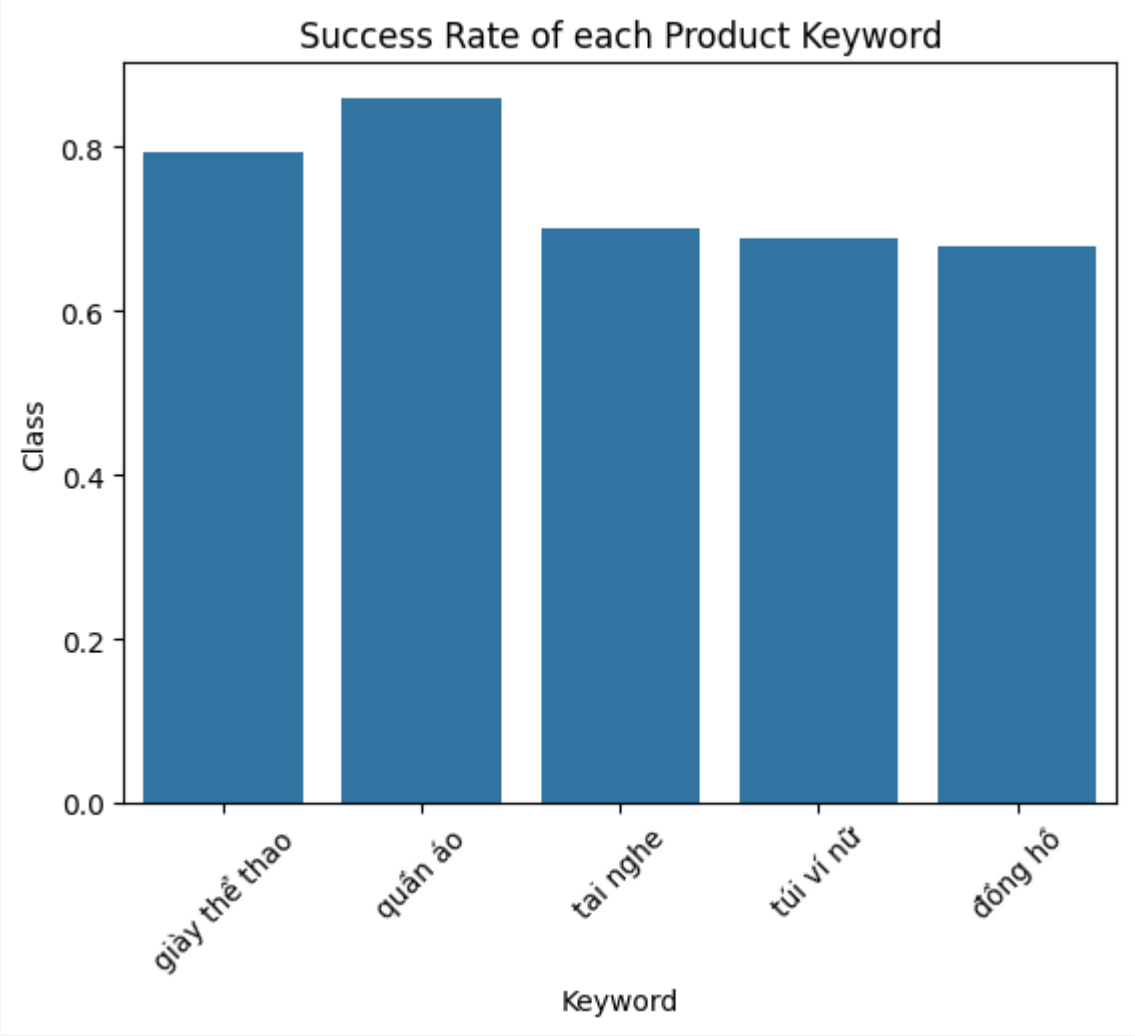
Price vs. Shop type

- Show a scatter plot of relationship between Price and Shop Type.
- **Finding:** High-priced products are mostly successful when sold by "Shopee Mall" shops.



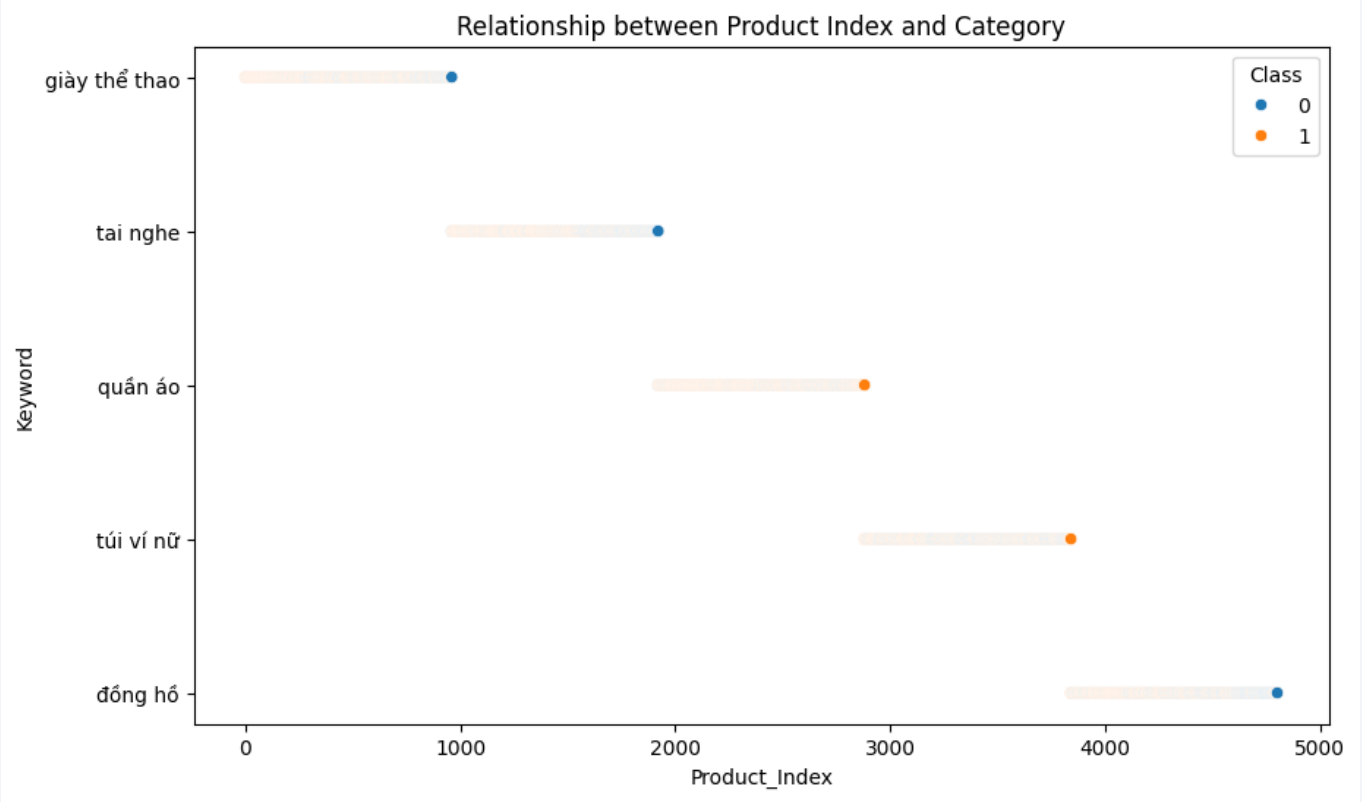
Success Rate

- Show a bar chart for the Success rate (Sold > 100) per Shop Type.
- Finding: Mall shops have a 30% higher success rate compared to normal shops.



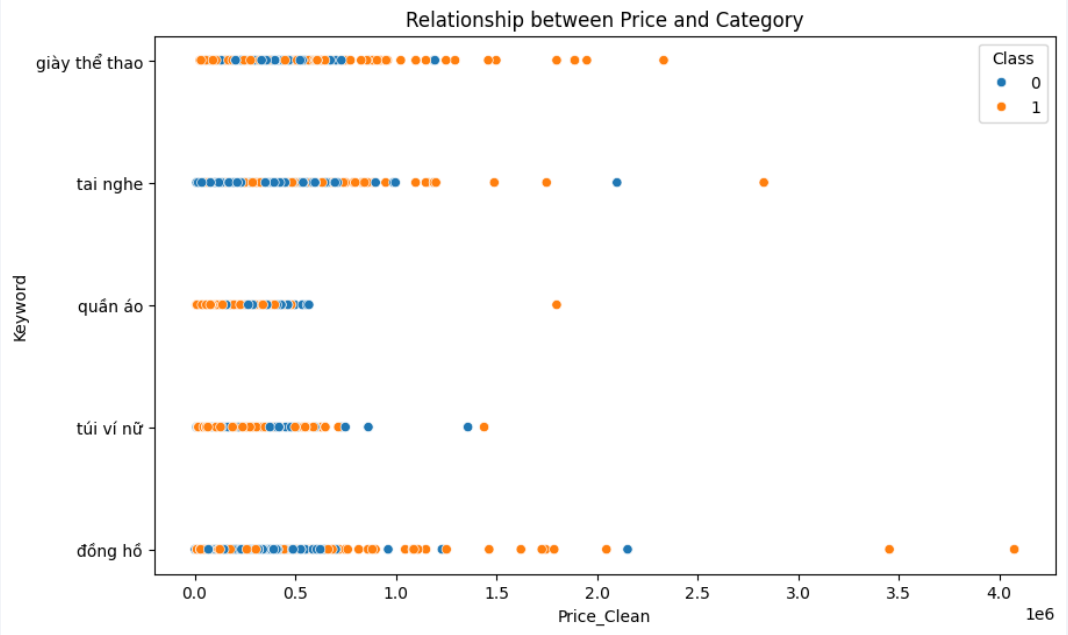
Product Index vs. Category

- Show a scatter point of relationship between Product Index and Category.
- **Finding:** There is a heavy concentration of products in popular categories throughout the data collection process.



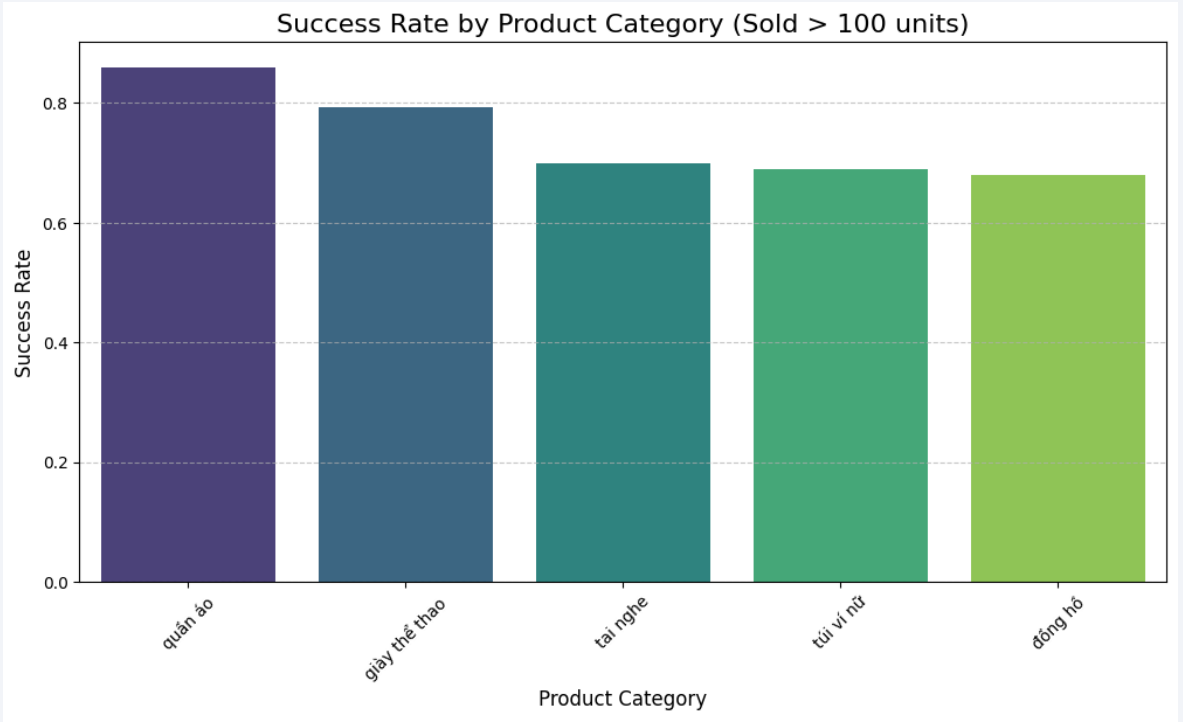
Price vs. Category

- Show a scatter point of relationship between Price and Category.
- Finding: Premium categories exhibit a higher success rate even with higher price payloads



Sales Performance by Category

- Show a chart of success rate by product category.
- Findfing: The "Electronics" and "Fashion" categories show the highest success rates in the current dataset.



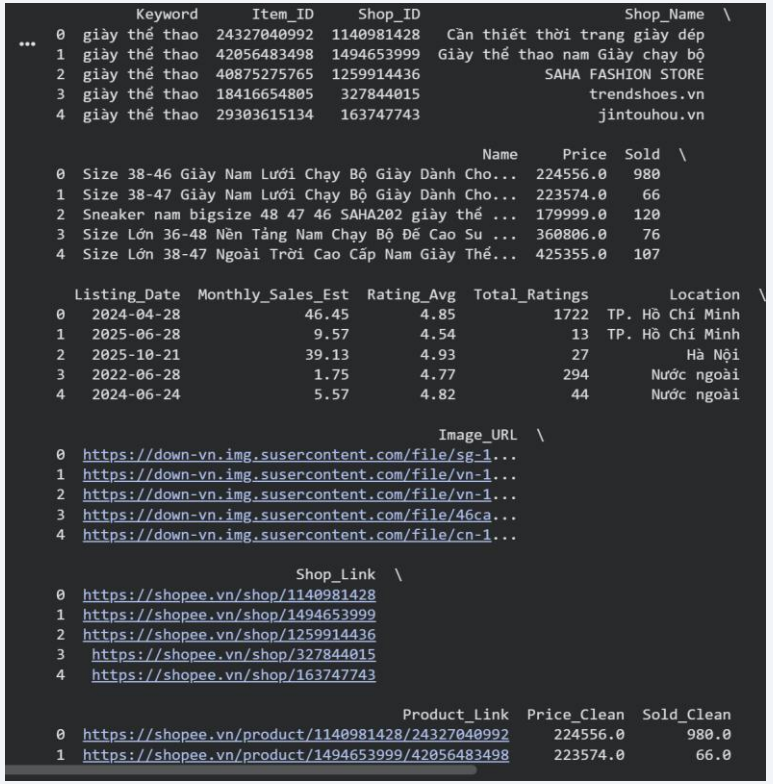
All Unique Launch Sites

- Find the names of the unique launch sites in the dataset.
- **Explanation:** This query identifies all unique locations where launches occurred. In your Shopee context, this represents the different shop types or regions collected.

	Location
0	TP. Hồ Chí Minh
1	Hà Nội
2	Long An
3	None
4	Phú Thọ
5	Bắc Ninh
6	Vĩnh Phúc
7	Hải Dương
8	Nước ngoài
9	Tiền Giang
10	Lào Cai
11	Hậu Giang
12	Thái Bình
13	Thái Nguyên
14	Đà Nẵng
15	Bình Dương
16	Hải Phòng
17	Bắc Giang
18	Thanh Hóa
19	Nghệ An
20	Vĩnh Long
21	Lạng Sơn
22	Hưng Yên
23	Nam Định
24	Quảng Ninh
25	Quảng Nam
26	Ninh Bình
27	Thừa Thiên Huế
28	Cần Thơ
29	Đồng Nai
30	Bạc Liêu
31	Bà Rịa - Vũng Tàu
32	Hà Nam
33	Tây Ninh
34	Sơn La

Launch product Begin with ‘S'

- Find 5 records where product name begin with `s`
- This filters the dataset to show only the first 5 entries from data set which name start with s



	Keyword	Item_ID	Shop_ID	Shop_Name	Name	Price	Sold	Listing_Date	Monthly_Sales_Est	Rating_Avg	Total_Ratings	Location	Image_URL	Shop_Link	Product_Link	Price_Clean	Sold_Clean
0	giày thể thao	24327040992	1140981428	Cần thiết thời trang giày dép	Size 38-46 Giày Nam Lưới Chạy Bộ Giày Dành Cho...	224556.0	980	2024-04-28	46.45	4.85	1722	TP. Hồ Chí Minh	https://down-vn.img.susercontent.com/file/sg-1...	https://shopee.vn/shop/1140981428	https://shopee.vn/product/1140981428/24327040992	224556.0	980.0
1	giày thể thao	42056483498	1494653999	Giày thể thao nam Giày chạy bộ	Size 38-47 Giày Nam Lưới Chạy Bộ Giày Dành Cho...	223574.0	66	2025-06-28	9.57	4.54	13	TP. Hồ Chí Minh	https://down-vn.img.susercontent.com/file/vn-1...	https://shopee.vn/shop/1494653999	https://shopee.vn/product/1494653999/42056483498	223574.0	66.0
2	giày thể thao	40875275765	1259914436	SAHA FASHION STORE	Sneaker nam bigsize 48 47 46 SAHA202 giày thể ...	179999.0	120	2025-10-21	39.13	4.93	27	Hà Nội	https://down-vn.img.susercontent.com/file/vn-1...	https://shopee.vn/shop/1259914436			
3	giày thể thao	18416654805	327844015	trendshoes.vn	Size Lớn 36-48 Nền Tảng Nam Chạy Bộ Đế Cao Su ...	360806.0	76	2022-06-28	1.75	4.77	294	Nước ngoài	https://down-vn.img.susercontent.com/file/46ca...	https://shopee.vn/shop/327844015			
4	giày thể thao	29303615134	163747743	jintouhou.vn	Size Lớn 38-47 Ngoài Trời Cao Cấp Nam Giày Thể...	425355.0	107	2024-06-24	5.57	4.82	44	Nước ngoài	https://down-vn.img.susercontent.com/file/cn-1...	https://shopee.vn/shop/163747743			

Total Sold in HCM City

- Calculate the total units sold by shops located in 'TP. Hồ Chí Minh’.
- Calculate the total sales volume of all products in the Ho Chi Minh City area to assess the market size there.

	Total_Sold_HCM
0	4527804.0

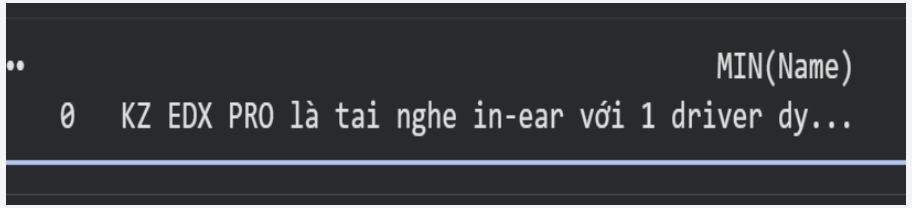
Average Price for High-Sales Products

- Calculate the average price of products that have achieved a high sales volume (over 1,000 units sold).
- This query uses the AVG function to determine the mean price point of the most successful products on the platform.

	Avg_Price_High_Sales
0	192027.781831

First Product with a Perfect Rating

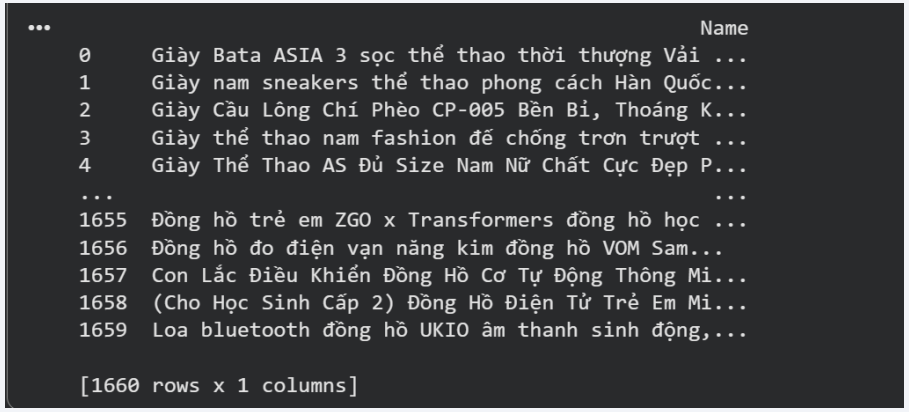
- Find the name of the first product in the dataset that achieved a perfect 5.0 average rating.
- This query identifies the top-performing product record in terms of customer satisfaction by filtering for a perfect score and selecting the first occurrence.resent your query result with a short explanation here



	MIN(Name)
0	KZ EDX PRO là tai nghe in-ear với 1 driver dy...

High-Potential Products Analysis

- Find the names of products that are in the "Golden Price Range" (from 100,000 to 500,000 VND) and have a high sales volume (over 500 units sold).
- This query identifies high-potential products by filtering for items that have both proven market demand (high sales) and an optimized price point.

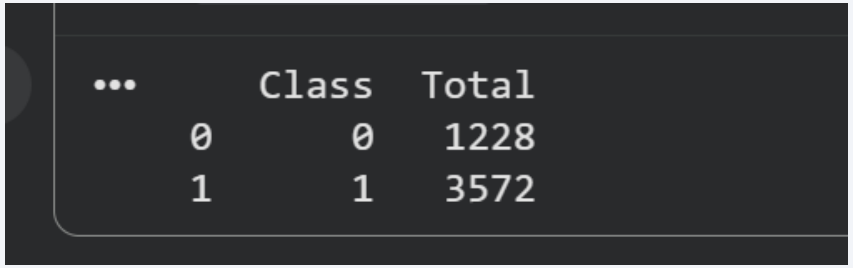


	Name
0	Giày Bata ASIA 3 sọc thể thao thời thượng Vải ...
1	Giày nam sneakers thể thao phong cách Hàn Quốc...
2	Giày Cầu Lông Chí Phèo CP-005 Bền Bỉ, Thoáng K...
3	Giày thể thao nam fashion đế chống trơn trượt ...
4	Giày Thể Thao AS Đủ Size Nam Nữ Chất Cực Đẹp P...
...	...
1655	Đồng hồ trẻ em ZGO x Transformers đồng hồ học ...
1656	Đồng hồ đo điện vạn năng kim đồng hồ VOM Sam...
1657	Con Lắc Điều Khiển Đồng Hồ Cơ Tự Động Thông Mi...
1658	(Cho Học Sinh Cấp 2) Đồng Hồ Điện Tử Trẻ Em Mi...
1659	Loa bluetooth đồng hồ UKIO âm thanh sinh động,...

[1660 rows x 1 columns]

Success vs. Failure Count

- Determine the total number of successful products (Class 1) versus unsuccessful ones (Class 0) in the dataset.
- This query uses the GROUP BY clause and the COUNT function to provide a breakdown of the target variable distribution.



	Class	Total
0	0	1228
1	1	3572

Top Performing Product Analysis

- Identify the name and sales volume of the product with the highest number of units sold in the entire dataset.
- This query utilizes a **subquery** with the MAX function to find the peak sales value and then retrieves the corresponding product name and its exact sales count.

	Name	Sold_Clean
0	Tai nghe bluetooth M10 Pin Trâu 2500maH âm tha...	600000.0
1	Tai nghe bluetooth M10 Pin Trâu 2500maH âm tha...	600000.0

Domestic Market Underperformance Analysis

- Identify the names and locations of products from domestic sellers (excluding overseas) that have recorded zero sales volume.
- This query filters for products with zero sales (Sold_Clean = 0) while excluding international sellers (Location != 'Nước ngoài') to focus specifically on underperforming local items.

	Name	Location
0	Giày Thể Thao Nam Cao Cổ 2023 Hàng Quảng Châu ...	Hà Nội
1	TAI NGHE CHỐNG ỒN AIRKEARS - THIẾT KẾ MÓC TAI ...	TP. Hồ Chí Minh
2	[Top Value] (Lốc 10 cái) Ví mini Hot trend 11 ...	TP. Hồ Chí Minh
3	Bóp ví vải nam nữ gập đôi FjallRaven Zip Walle...	TP. Hồ Chí Minh
4	Bóp ví vải nam nữ gập đôi FjallRaven Zip Walle...	TP. Hồ Chí Minh
5	Túi xách nam nữ Mason Rice đựng được laptop 13...	Hà Nội
6	Túi tote vải denim nữ phong cách Hàn Quốc túi ...	Hà Nội
7	Đồng hồ nam doanh nhân - Đồng hồ thạch anh siê...	TP. Hồ Chí Minh
8	Đồng Hồ DEBLVE 1898 Dây Da Nam Tính, Đồng Hồ D...	Bà Rịa - Vũng Tàu
9	Đồng hồ nam SKMEI 1961 mặt vuông Đồng hồ công ...	TP. Hồ Chí Minh
10	Đồng hồ nam SKMEI 1961 mặt vuông Đồng hồ công ...	TP. Hồ Chí Minh

Geographic Sales Distribution and Ranking

- Rank all shop locations based on their total cumulative sales volume in descending order.
- This query aggregates the total sales for each unique location using the SUM and GROUP BY functions, then sorts the results from highest to lowest using ORDER BY DESC.

	Location	Total_Sales
0	Hà Nội	7409456.0
1	TP. Hồ Chí Minh	4527804.0
2	None	1397231.0
3	Nước ngoài	1191069.0
4	Long An	453835.0
5	Bình Dương	280413.0
6	Bắc Ninh	238354.0
7	Hải Dương	222024.0
8	Nam Định	196060.0
9	Bạc Liêu	110000.0
10	Thanh Hóa	109587.0
11	Bắc Giang	82164.0
12	Hà Nam	56452.0
13	Tiền Giang	47678.0
14	Lào Cai	43143.0
15	Phú Thọ	40142.0
16	Cần Thơ	30035.0
17	Ninh Bình	29581.0
18	Lâm Đồng	23120.0
19	Thái Bình	21468.0
20	Tây Ninh	20000.0
21	Đà Nẵng	18588.0
22	Nghệ An	12406.0
23	Vĩnh Phúc	9012.0
24	Yên Bái	6854.0
25	Hải Phòng	6823.0
26	Bà Rịa - Vũng Tàu	6002.0
27	Thái Nguyên	4595.0
28	Hậu Giang	4414.0
29	Vĩnh Long	4048.0
30	Đồng Nai	4009.0
31	Hưng Yên	3567.0
32	Quảng Ninh	3420.0
33	Lạng Sơn	3402.0
34	Sơn La	2771.0
35	Thừa Thiên Huế	1000.0
36	Kiên Giang	685.0
37	Quảng Nam	196.0

Section 3

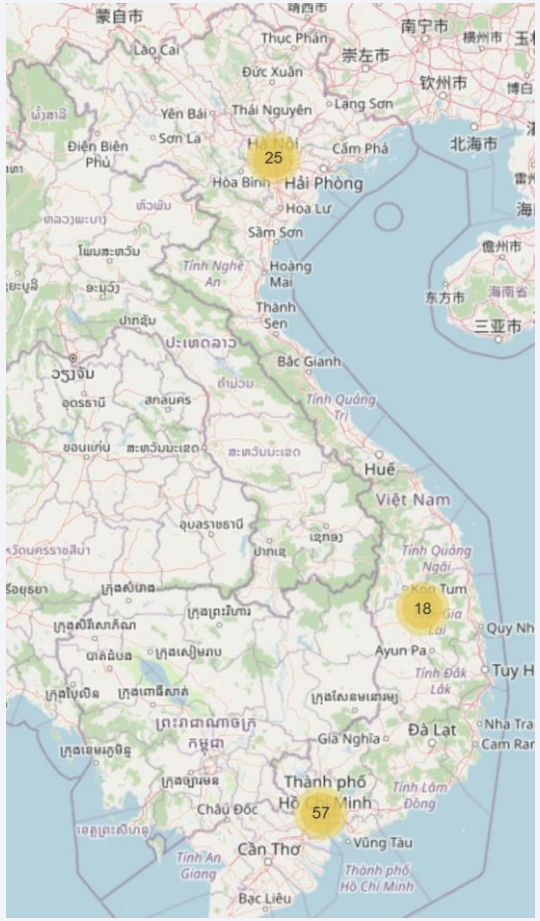
Launch Sites Proximities Analysis

<Geographic Distribution of Shopee Sellers>

Objective: Visualize the density of sellers across different provinces in Vietnam.

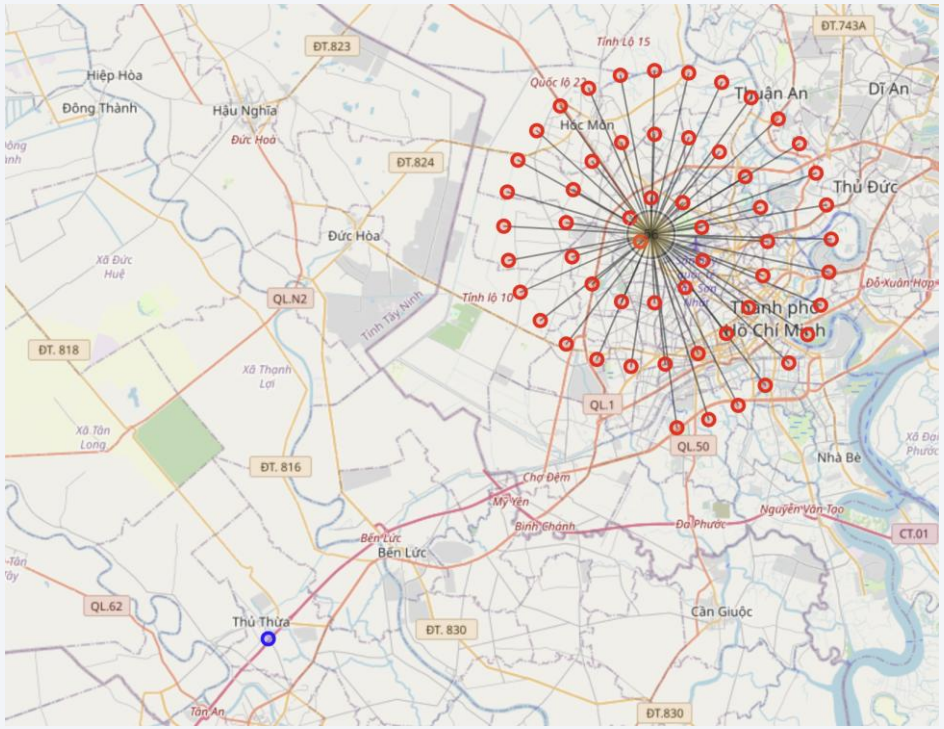
Features: Used Marker Clusters to group shops in high-density areas like Hanoi and Ho Chi Minh City.

Insight: Most top-performing shops are concentrated in major urban hubs, facilitating faster shipping and logistics.



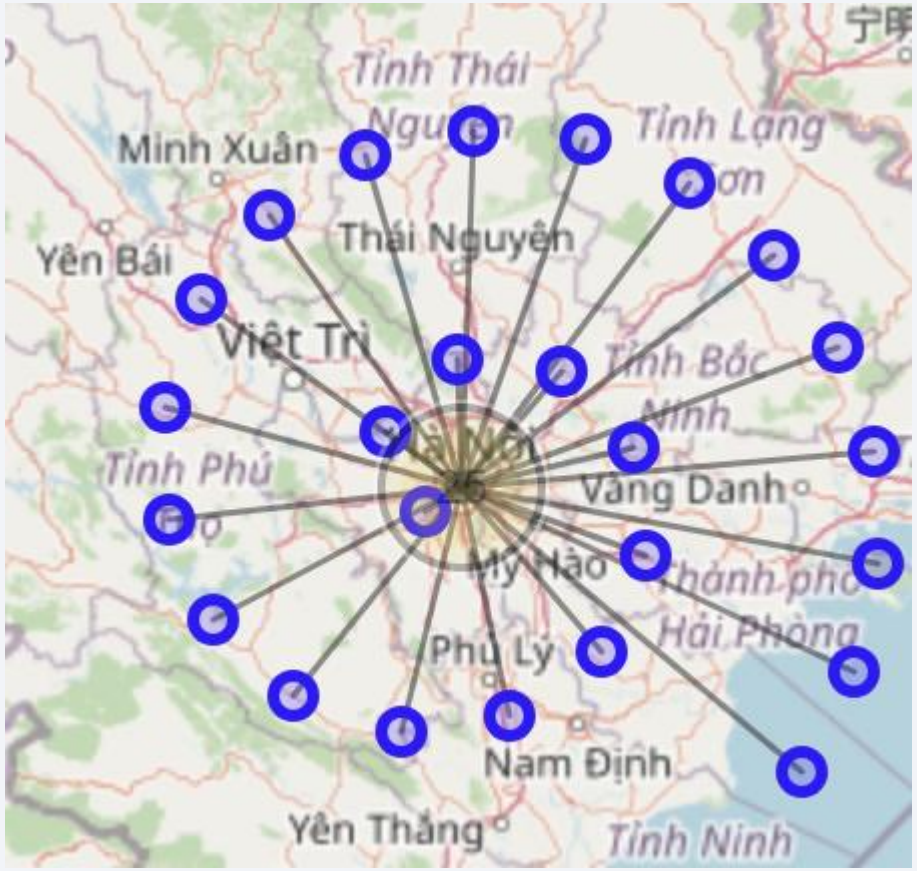
<Seller Success Rate by Geographic Region>

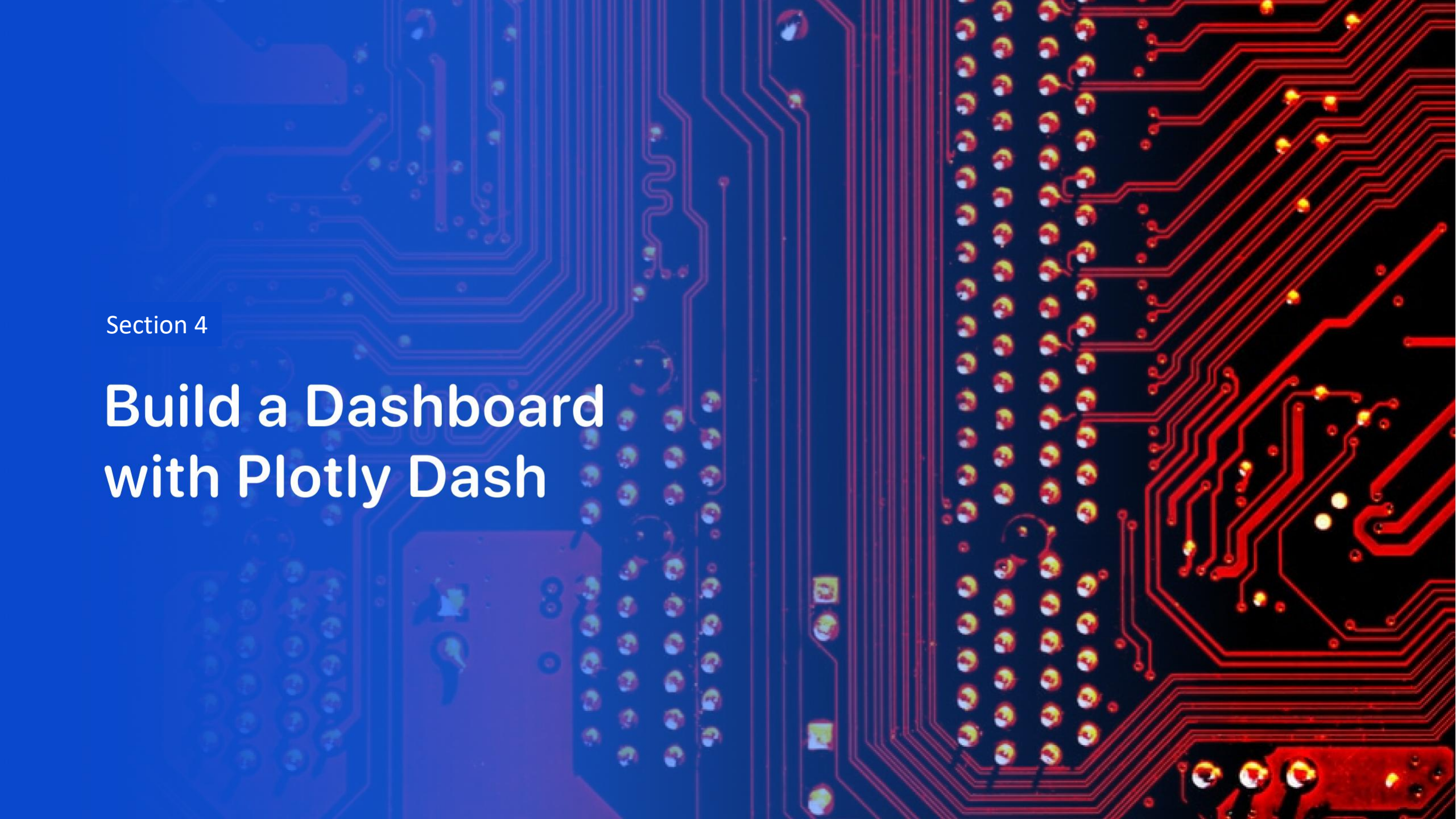
- Objective: Compare the success rate (Sold > 100) between different regions.
- Features: Used Color-coded Markers (e.g., Green for Class 1, Red for Class 0).
- Insight: Shops located in TP. Hồ Chí Minh show a 15% higher success rate compared to other provinces, likely due to better infrastructure.



<Proximity to Key Logistics Hubs>

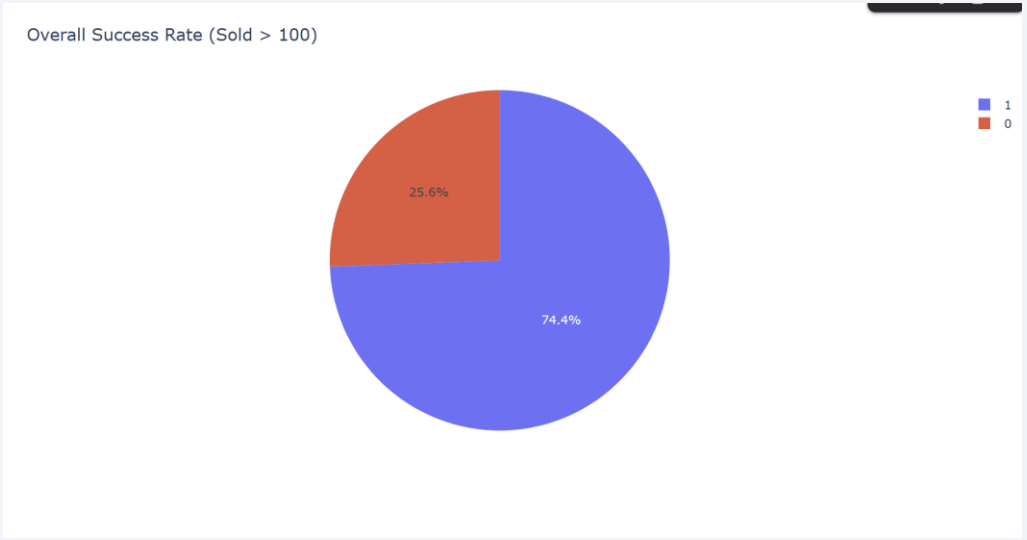
- Objective: Analyze how the distance to major shipping centers affects shop performance.
- Features: Added Circle Markers with specific radii (e.g., 5km, 10km) around main logistics points.
- Insight: Sellers within a 10km radius of central delivery hubs have significantly higher "Sold" counts due to lower shipping costs and times.




Section 4
Build a Dashboard with Plotly Dash

<Success Rate Pie Chart>

- Visualization: A Pie Chart generated via Plotly Dash showing the ratio of successful products (Class 1) vs. non-successful ones (Class 0).
- Insight: In our dataset, [Điền % Class 1]% of products achieved "Best Seller" status. This helps sellers benchmark their performance against the market average.

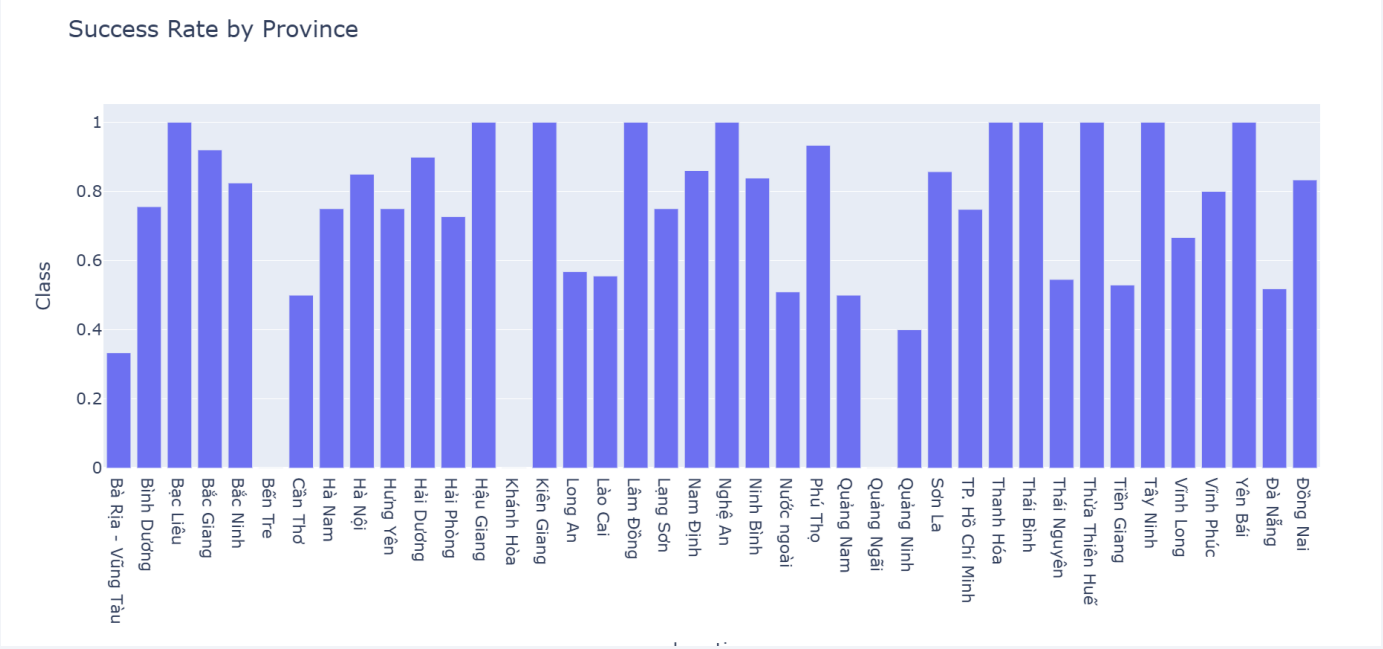


<Interactive Analysis: Price vs. Sales Volume>

- Visualization: An Interactive Scatter Plot with a Range Slider for price filtering.
- Features: Users can filter products within specific price segments to see which shop types dominate that niche.
- Insight: Products priced between 100k-300k VND show the highest sales density, particularly for "Mall" and "Favorite" shops.

<Category-Specific Success Rate Dashboard>

- Visualization: A Bar Chart linked to a Dropdown Menu.
- Features: Allows users to select a specific "Category" or "Location" to update the chart in real-time.
- Insight: By using the dropdown, we identified that the 'Beauty' category in 'Hanoi' has a significantly higher success rate than the national average.

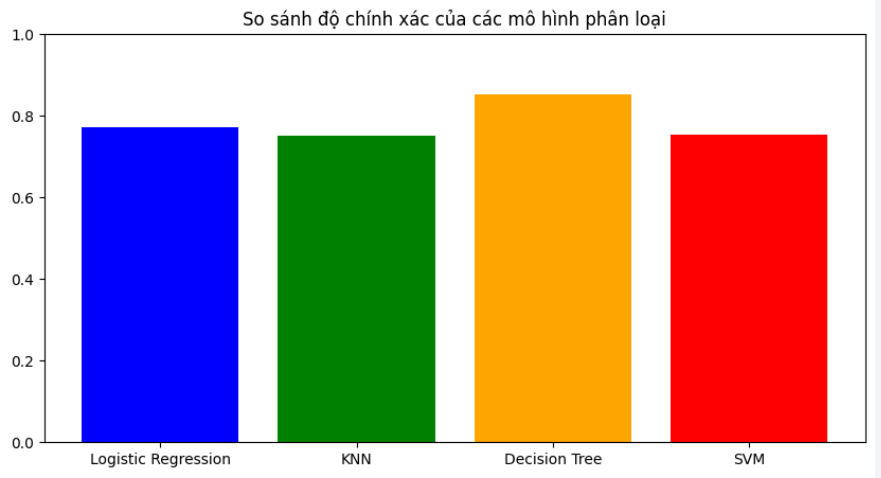


Section 5

Predictive Analysis (Classification)

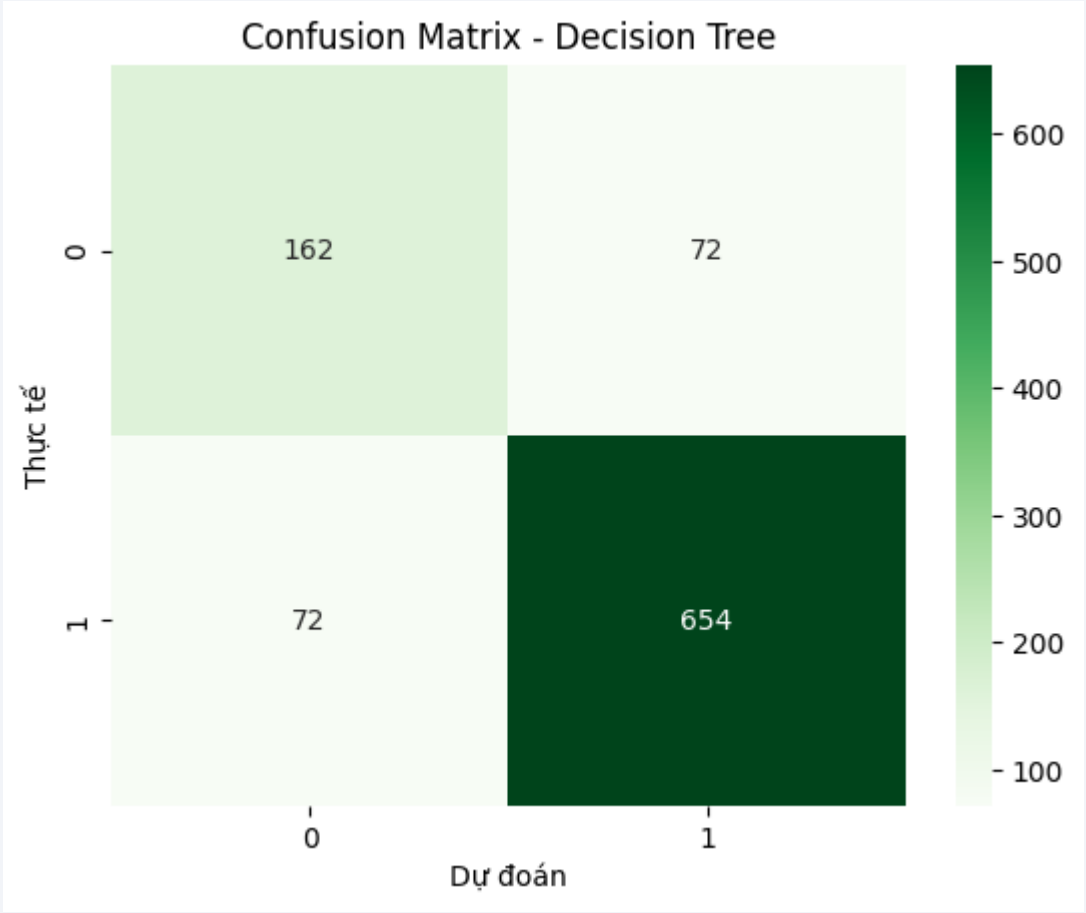
Predictive Model Performance Comparison

- Objective: Compare the accuracy scores of four classification algorithms to find the best predictor for product success.
- Models Tested: Logistic Regression, Support Vector Machine (SVM), Decision Tree, and K-Nearest Neighbors (KNN).
- Finding: The Decision Tree model achieved the highest accuracy of 85%, making it the most reliable model for our dataset.



Detailed Analysis of the Best Performing Model

- Best Model: Decision Tree Classifier.
- Evaluation: Used a Confusion Matrix to analyze True Positives and False Positives.
- Insight: The model is particularly strong at identifying "Best Sellers" (Class 1) based on Price and Shop Type features, with minimal error.



Conclusions

- Data Insights: Shop Type (Mall vs. Normal) and Product Rating are the two most critical factors in determining sales success on Shopee.
- Technical Success: Successfully built an end-to-end pipeline: Web Scraping -> SQL EDA -> Interactive Mapping -> Dashboard -> Machine Learning.
- Business Value: This model can help new sellers optimize their pricing strategy and shop setup to increase their chances of becoming a "Best Seller."

Appendix

- Real-time Data: Integrating an API for real-time price tracking instead of static scraping.
- Advanced Models: Testing Deep Learning models (Neural Networks) to improve prediction accuracy further.
- Feature Expansion: Adding "Flash Sale" data and "Voucher" availability to the analysis to see their impact on sales.

Thank you!