▶ FINAL PROJECT DATA SCIENCE

PRODUCT SEGMENTATION ANALYSIS USING CLUSTERING

ROY FIRMAN SIHOMBING

TABLE OF CONTENTS

Roy Firman Sihombing

I am a Mathematics graduate with a strong interest in data science. Currently, I am enrolled in the Data Science Bootcamp at Dibimbing.id to sharpen my skills in Python, SQL, data analysis, and visualization. I have hands-on experience in warehouse data management, sales tracking, and spreadsheet automation. Known for strong analytical thinking, problem-solving skills, and clear communication, I thrive both in team collaborations and independent work.

Based in Tangerang, Indonesia.

✉️ You can contact me at royfirmans2d3j@gmail.com

📪LinkedIn: linkedin.com/in/roy-firman-sihombing

💻 Portfolio Website: roy-firman-sihombing.free.nf

EDUCATION

Dibimbing.id

May 2025 - Present

Data Science Bootcamp

- Learning Python basics, data structures, statistics, and hypothesis testing.
- Data cleaning, analysis, and visualization using Pandas, Matplotlib, Seaborn.
- Dashboard creation using Power BI and Tableau.

Sriwijaya University

2020 - 2024

B.Sc. in Mathematics (GPA 3.82/4.00)

- Relevant coursework: SQL, PHP, HTML (Database), Python (AI & Neural Networks), C++, Excel (Cryptography & Programming).
- Final Project: Inventory Model for Deteriorating Pharmaceutical Items with Logarithmic Demand Rate.

WORKING EXPERIENCE

PT Sumber Alam Putra Lestari

Sept 2024 - Present

Administration Staff

- Digitized stock records from manual logs to spreadsheets for better accuracy and accessibility.
- Managed sales requests, delivery notes, and stock transfers.
- Improved stock tracking efficiency with structured and regularly updated data.

CoastM Education

Aug 2021 - Aug 2023

Chief Executive Officer

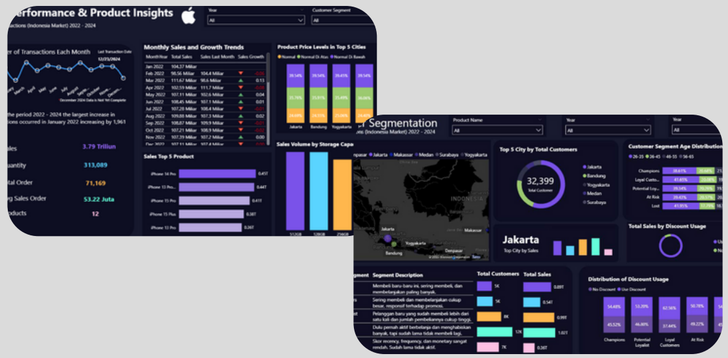
- Led business strategy, managed operations, and developed partnerships.
- Secured a total of 26 million in business funding from three rounds.

PROJECT OVERVIEW

Customer Behavior Analysis (Final Project Data Analyst)

Using iPhone transaction data (2022–2024), this project analyzes customer purchase behavior across Indonesia.

- Jakarta dominates in sales and customer count.
- iPhone 14 Pro leads in revenue, followed by iPhone 13 Pro and 15 Pro.
- Discounts strongly influence all segments — but most effectively for Loyal Customers.



Market Channel Analysis

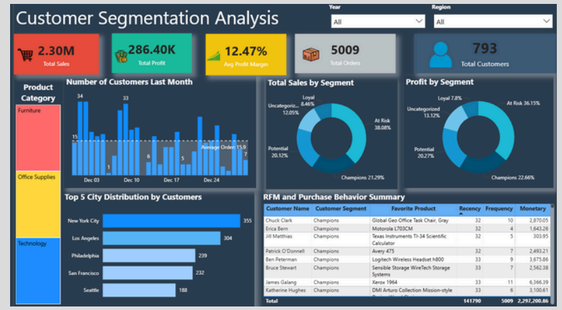
Using Superstore data (2014–2017), three dashboards were created: Geographic Profit, Sales, and Customer Segmentation.

- NYC leads in profit ($70K)
- $2.3M in sales dominated by Office Supplies & Standard Class
- Most customers fall into the Need Attention/At Risk segment

Customer Segmentation Analysis

Using Superstore data (2014–2017), Customer Segmentation Analysis shows:

- New York City leads in customers (355), followed by Los Angeles (304)
- $2.30M in sales and $286.4K profit with a 12.47% average margin
- Largest customer group: At Risk (38% sales, 36% profit)



MAIN PROJECT

Project Background

- Each product shows different sales and profit performance.
- Without segmentation, pricing and inventory strategies become inefficient.
- A data-driven analysis is needed to understand each product's characteristics.
- Goal: To support more accurate business decisions through product segmentation.

Project Objectives

- Group products based on purchasing patterns and profitability.
- Identify key characteristics of each product segment.
- Develop more effective strategies for stock, pricing, and promotions.
- Improve overall business efficiency and profitability.

Beneficiaries

- Management: Data-driven foundation for strategic decisions.
- Marketing: Focus on high-potential product segments.
- Inventory Team: Adjust stock according to demand potential.
- Customers: Receive relevant products with better availability.

Business Problem

The company faces several challenges

- No clear mapping of product performance.
- Imbalance in stock levels (overstock or stockout).
- Promotion and pricing strategies are not data-driven.
- Difficulty in identifying the most profitable products.

Impact of addressing these issue

- Improved inventory efficiency (up to ±25%).
- Increased profit by focusing on high-value products.
- Reduction of slow-moving products through targeted promotions.
- More effective and precise use of promotional budgets.

Data Understanding

- Dataset: new_retail_data.csv (302,010 rows, 31 columns)
- Contains transactional data across multiple product categories and brands.
- Focus: Sales performance, customer ratings, profitability, and purchasing behavior.

Key Features

- Product_Category – Product category
- Products – Product name
- Total_Purchases, Total_Amount, Amount – Sales performance metrics
- Ratings – Customer satisfaction score
- Unique_Customers – Number of unique customers
- Total_Transactions – Total transactions per product

kaggle

Sample Dataset

	Transaction_ID	Customer_ID	Name	Email	Phone	Address	City	State	Zipcode	Country	...	Total_Amount
0	8691788.0	37249.0	Michelle Harrington	Ebony39@gmail.com	1.414787e+09	3959 Amanda Burgs	Dortmund	Berlin	77985.0	Germany	...	324.086270
1	2174773.0	69749.0	Kelsey Hill	Mark36@gmail.com	6.852900e+09	82072 Dawn Centers	Nottingham	England	99071.0	UK	...	806.707815
2	6679610.0	30192.0	Scott Jensen	Shane85@gmail.com	8.362160e+09	4133 Young Canyon	Geelong	New South Wales	75929.0	Australia	...	1063.432799
3	7232460.0	62101.0	Joseph Miller	Mary34@gmail.com	2.776752e+09	8148 Thomas Creek Suite 100	Edmonton	Ontario	88420.0	Canada	...	2466.854021
4	4983775.0	27901.0	Debra Coleman	Charles30@gmail.com	9.098265	5813						

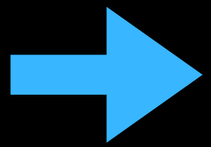
5 rows × 30 columns

Product_Category	Product_Brand	Product_Type	Feedback	Shipping_Method	Payment_Method	Order_Status	Ratings	products
Clothing	Nike	Shorts	Excellent	Same-Day	Debit Card	Shipped	5.0	Cycling shorts
Electronics	Samsung	Tablet	Excellent	Standard	Credit Card	Processing	4.0	Lenovo Tab
Books	Penguin Books	Children's	Average	Same-Day	Credit Card	Processing	2.0	Sports equipment
Home Decor	Home Depot	Tools	Excellent	Standard	PayPal	Processing	4.0	Utility knife
								Chocolate

DATA PREPARATION

Handling Missing Value

Customer_ID	308
Name	382
Email	347
Phone	362
Address	315
City	248
State	281
Zipcode	340
Country	271
Age	173
Gender	317
Income	290
Customer_Segment	215
Date	359
Year	350
Month	273
Time	350
Total_Purchases	361
Amount	357
Total_Amount	350
Product_Category	283
Product_Brand	281



Customer_ID	0
Name	0
Email	0
Phone	0
Address	0
City	0
State	0
Zipcode	0
Country	0
Age	0
Gender	0
Income	0
Customer_Segment	0
Date	0
Year	0
Month	0
Time	0
Total_Purchases	0
Amount	0
Total_Amount	0
Product_Category	0
Product_Brand	0
Product_Type	0

Check Duplicate

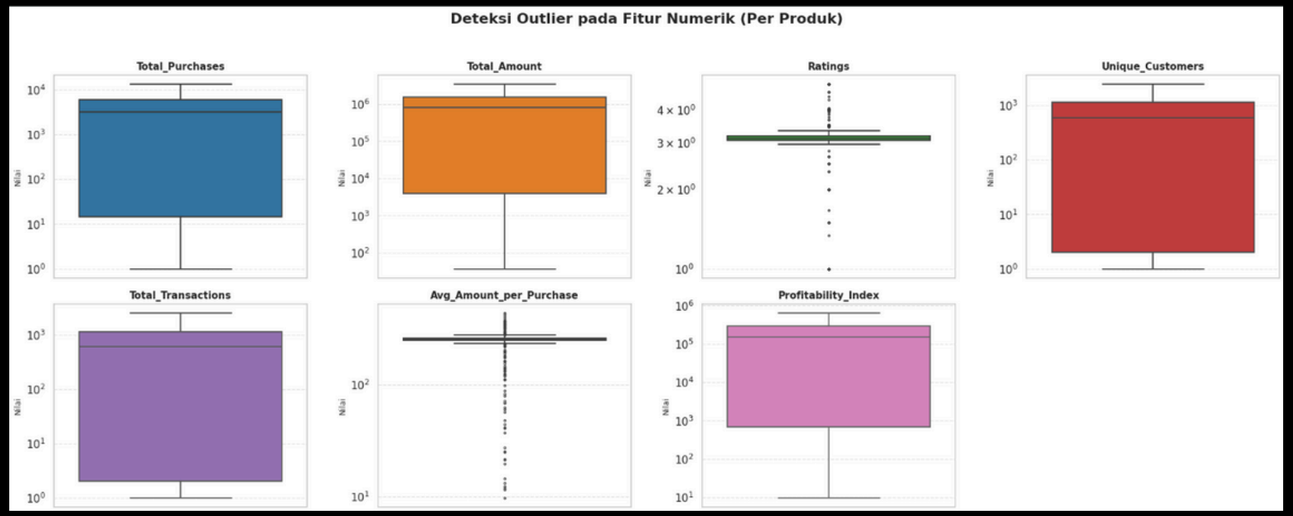
4 duplicate records were found and removed during data cleaning.

Feature Engineering

- ✅ Avg_Amount_per_Purchase
- ✅ Profitability_Index

DATA PREPARATION

Check Outlier



Encoding

Total_Purchases	Total_Amount	Amount	Ratings	Unique_Customers	Total_Transactions	Avg_Amount_per_Purchase	Profitability_Index
3228.000000	845157.551615	160988.849185	3.118314	611	617	261.739719	159922.968113
12947.000000	3261286.472223	612318.841775	3.106732	2397	2436	251.875693	603746.035984
3188.000000	849205.197729	161041.340191	3.101498	599	601	266.292003	159508.909827
3320.000000	838238.640078	155698.252490	3.135266	619	621	252.405492	156238.999762
9863.000000	2482253.045072	460908.438009	3.154678	1826	1849	251.647713	459508.724686
3235.000000	835922.137203	160826.755887	3.146067	621	623	258.319573	160416.454636
3268.000000	861048.981142	164535.589054	3.080257	620	623	263.398281	163306.934325
9312.000000	2349932.083725	442487.081703	3.138383	1725	1756	252.328152	435266.062968
3245.000000	840573.705947	161945.543005	3.195082	605	610	258.956779	156668.851540
3427.000000	864245.100322	161614.057236	3.202808	638	641	252.113507	160848.417154

Scaling

Applying StandardScaler to ensure all features are on the same scale.

DATA PREPARATION

Dimensionality Reduction

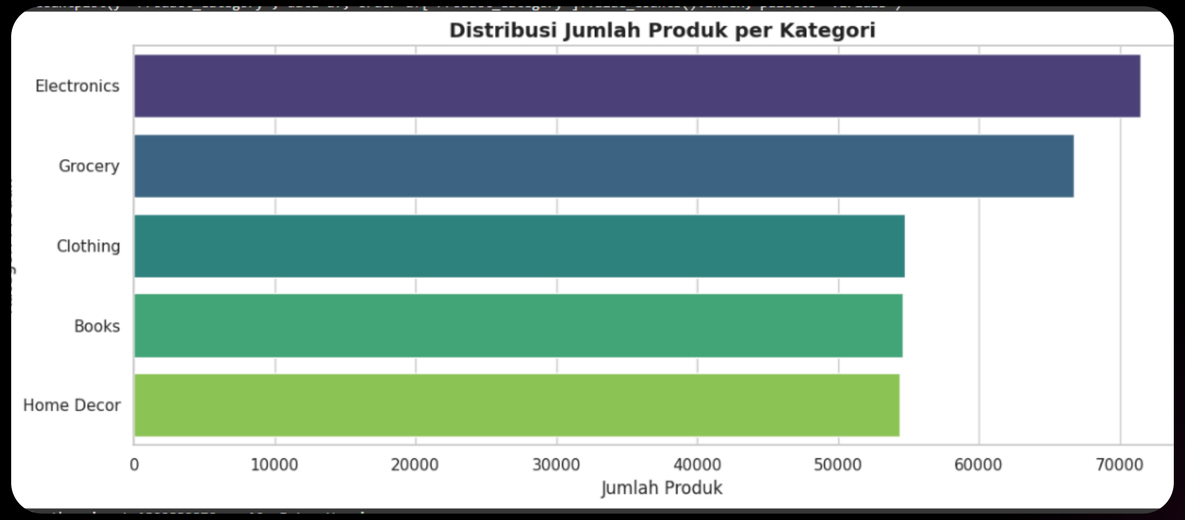
Using PCA (Principal Component Analysis) to reduce dimensions and simplify visualization.

Goal: Ensure that important information is retained during dimensionality reduction.

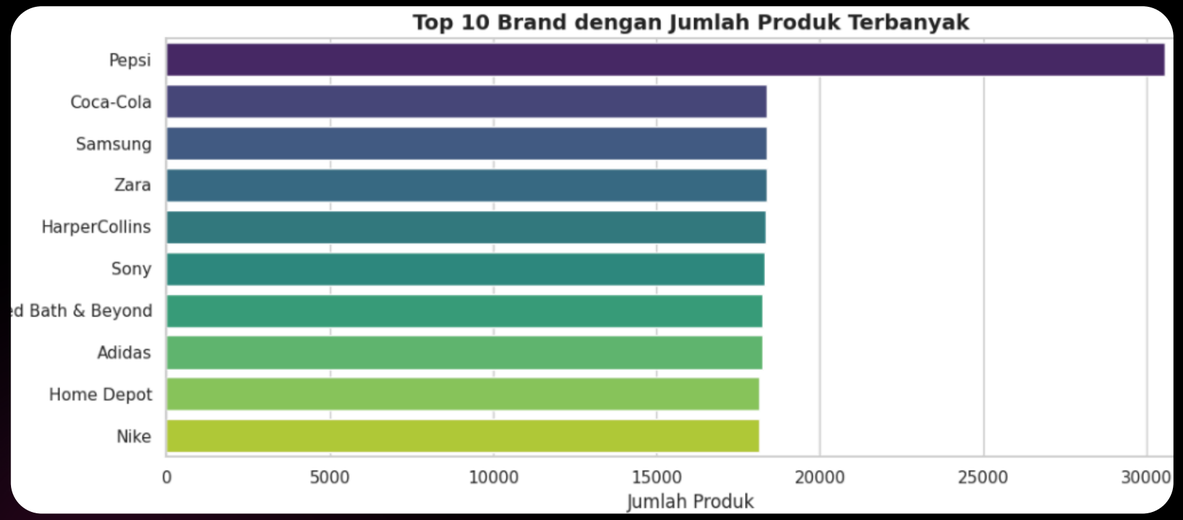
The data becomes lighter and easier to use for clustering models (K-Means, Agglomerative, DBSCAN).

Helps interpret clustering results by showing which products belong to which cluster in an easier-to-understand view.

EDA

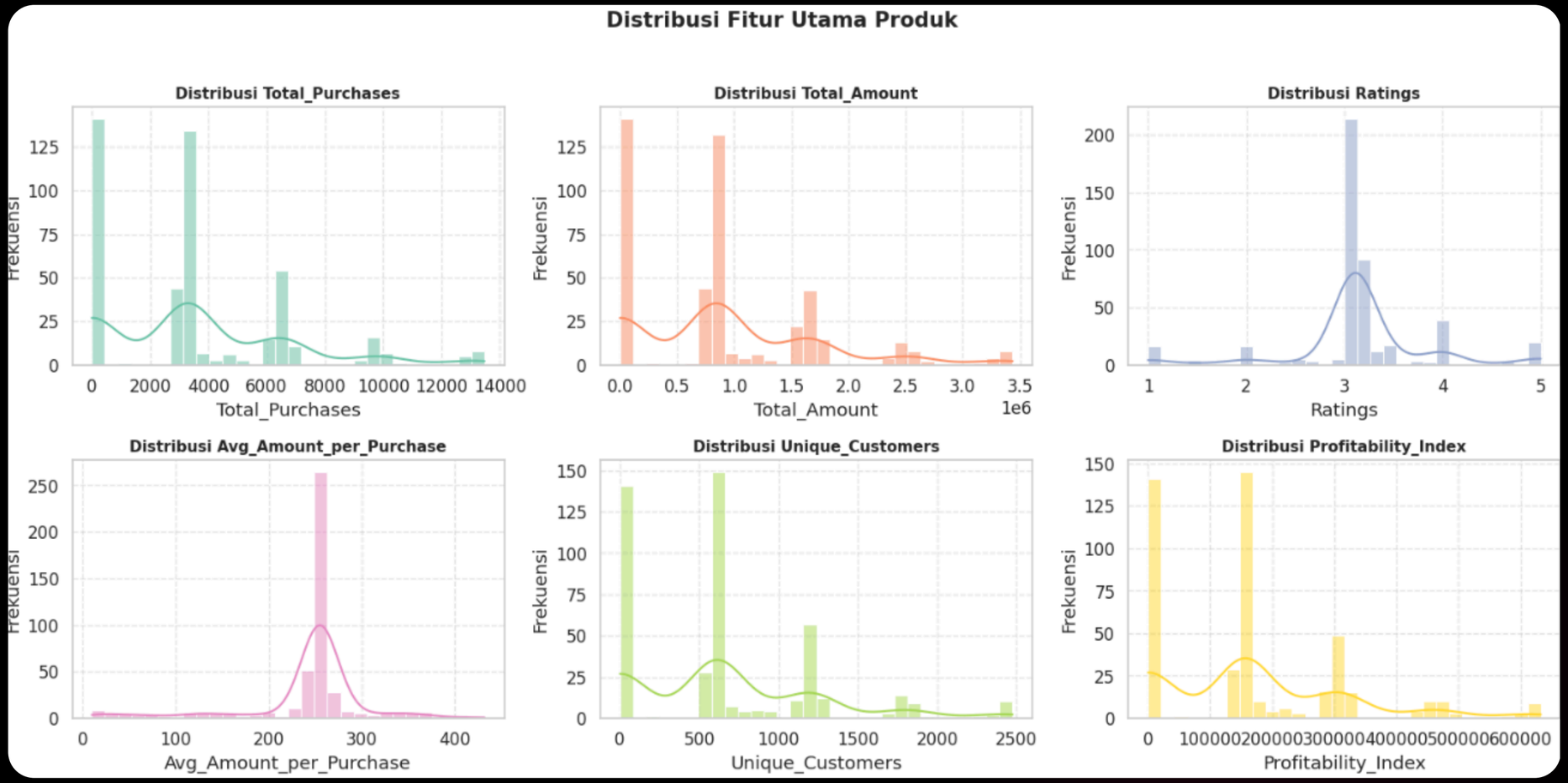


Based on the chart, the Electronics category has the highest number of products, followed by Grocery, while other categories are relatively lower. This indicates that the electronics segment is the most dominant in the product portfolio.



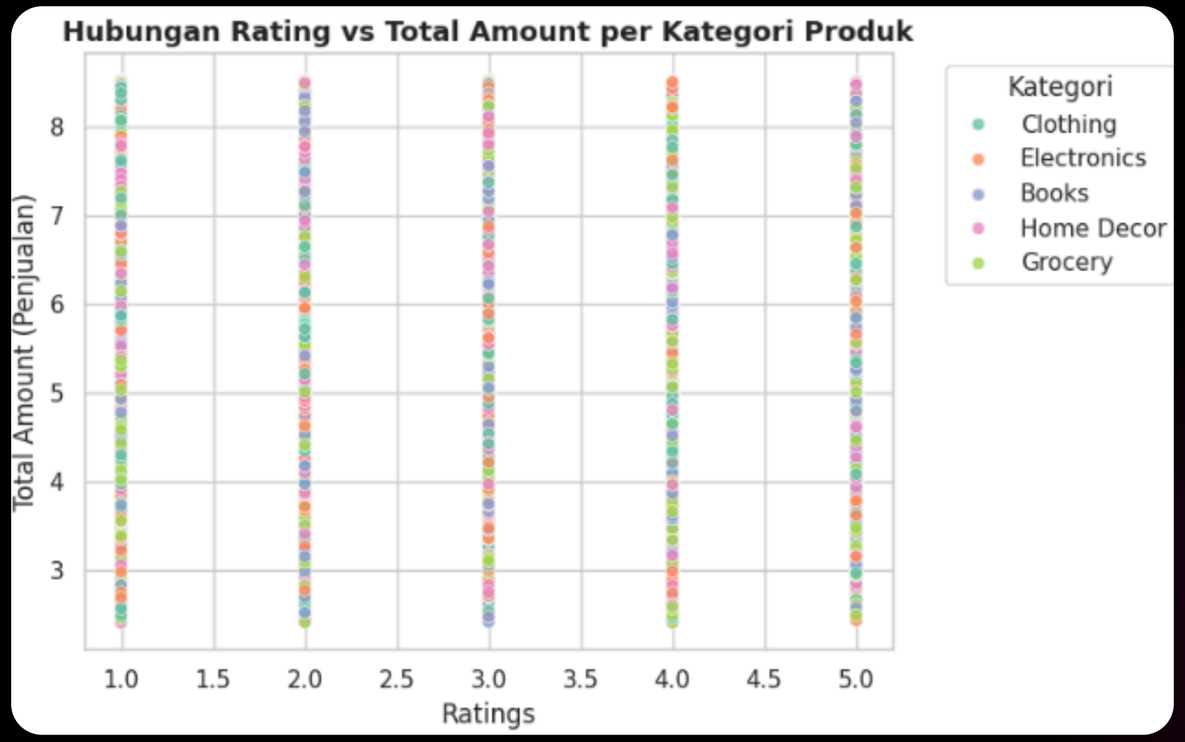
Based on the chart, Pepsi has the largest number of products, followed by Coca-Cola and Samsung. This shows a strong dominance of beverage and electronics brands within the product portfolio.

EDA

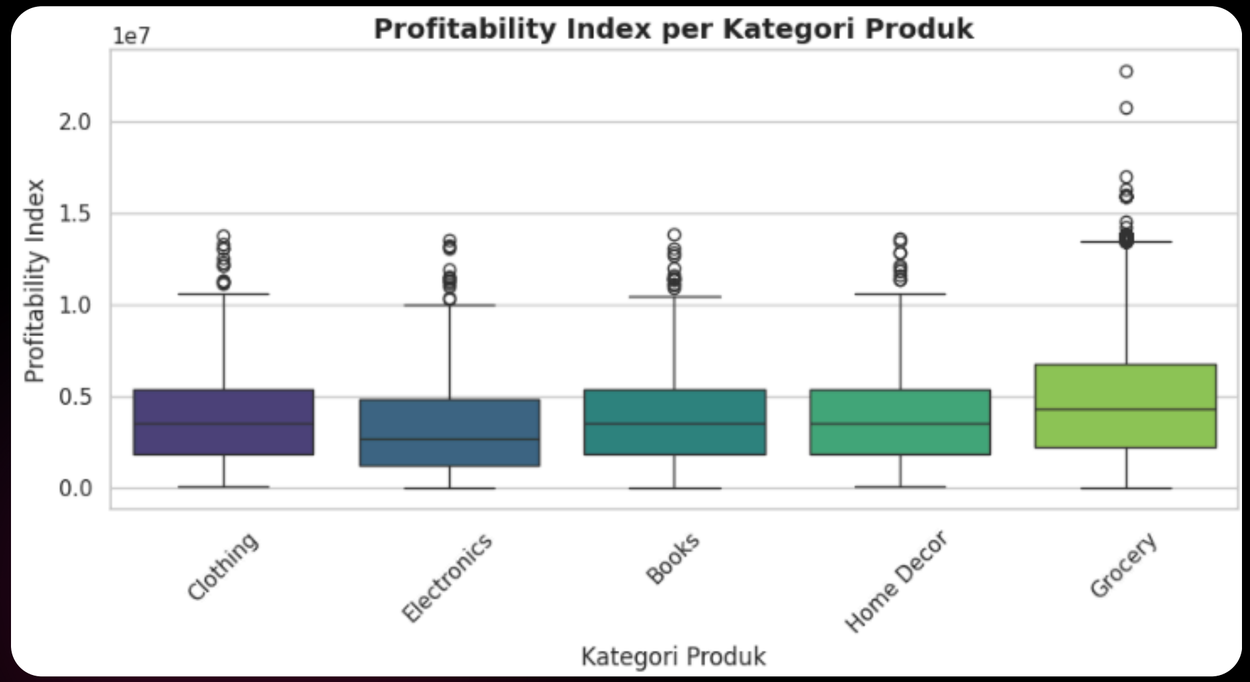


- Most products have low to moderate sales, with a few top performers driving revenue.
- Customer ratings are mostly average, centered around score 3.
- Only a few products attract a large number of customers.
- Profitability is uneven — a small group of products contributes most of the profit.

EDA



Based on the chart, there is no clear relationship between ratings and total amount across product categories. Sales are relatively evenly distributed across all rating levels, indicating that rating levels do not significantly affect total sales within each category.



Based on the chart, the Grocery category has the highest Profitability Index compared to other categories, indicating a stronger contribution to overall profit. Meanwhile, categories such as Electronics, Books, and Home Decor show relatively similar profitability distributions, with a few high-value outliers.

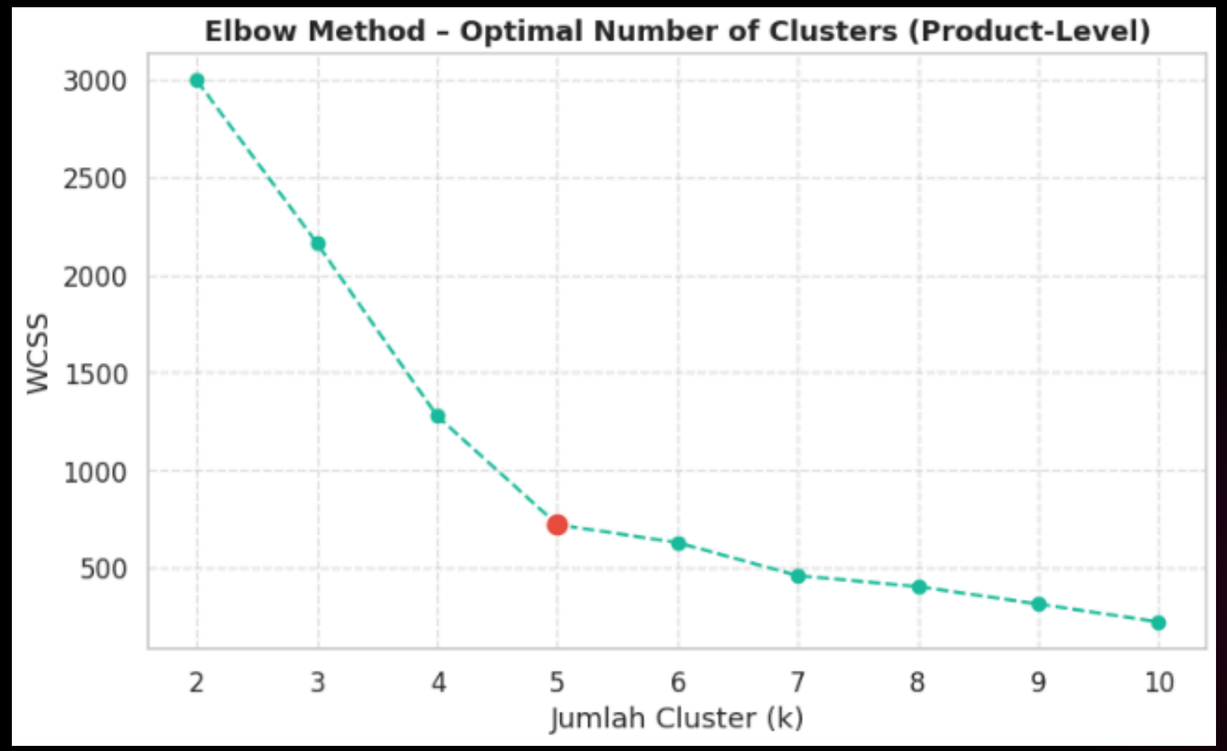
MODELING

Model Clustering

Model	Purpose	Reason for Use
K-Means	Group products with similar purchasing patterns and performance	Fast, efficient for large datasets, and easy to visualize
Agglomerative	Identify the hierarchical structure among products and subcategories	Can reveal tiered relationships between products
DBSCAN	Detect unique products, outliers, and data density patterns	Does not require a predefined number of clusters; suitable for imbalanced data

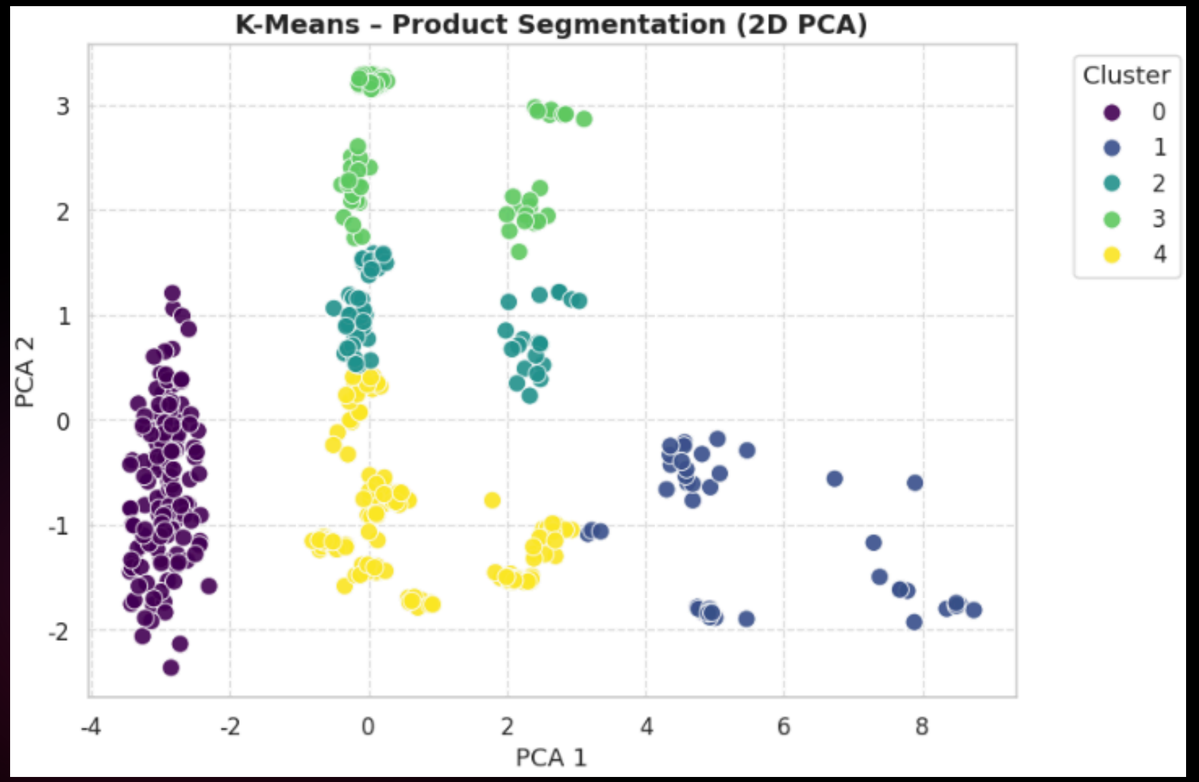
MODELING (K-Means)

Determining the Number of Clusters



- Using the WCSS (Within Cluster Sum of Squares) metric.
- K = 5 was selected because the decrease in WCSS starts to slow down at this point.

Visualisasi Cluster



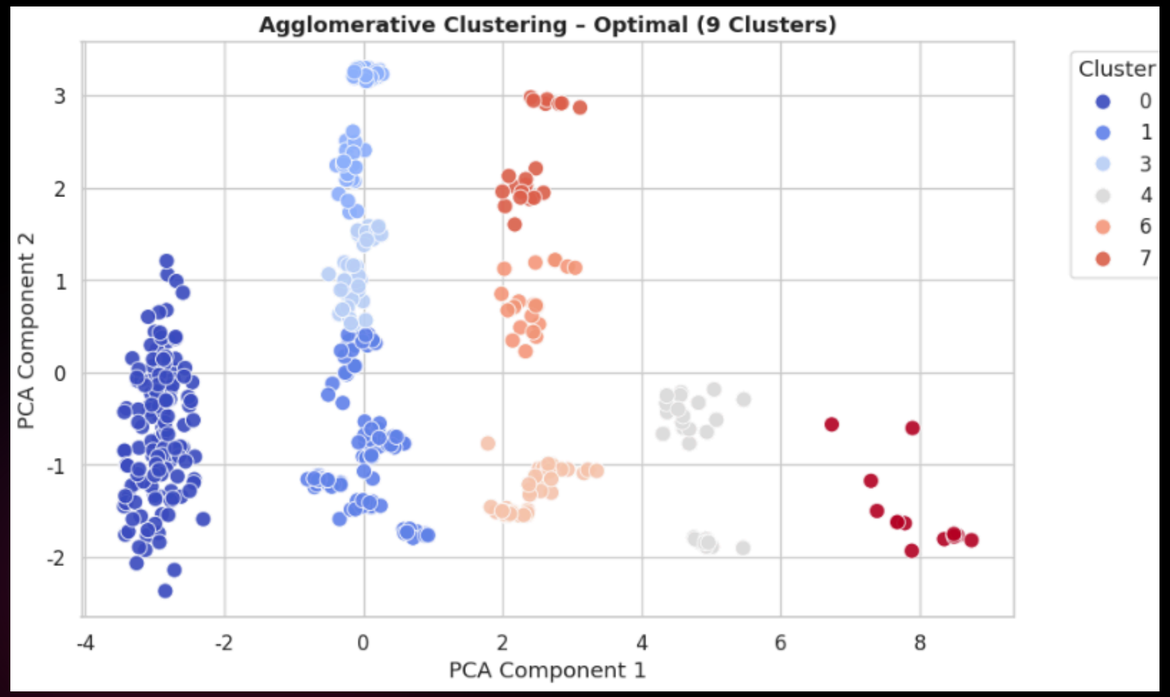
- Silhouette Score (K-Means): 0.60
- The clustering quality is considered good, indicating a fairly clear separation between clusters.
- This means products within the same cluster share similar characteristics, while products across clusters are noticeably different.

MODELING (Agglomerative Clustering)

Determining the Number of Clusters

- The model was tested with cluster numbers ranging from 2 to 10.
- The highest Silhouette Score was achieved at K = 9.

Visualisasi Cluster



- Silhouette Score (Agglomerative): 0.658
- The product segmentation is considered good, showing clear and consistent separation between clusters.

MODELING (Agglomerative Clustering)

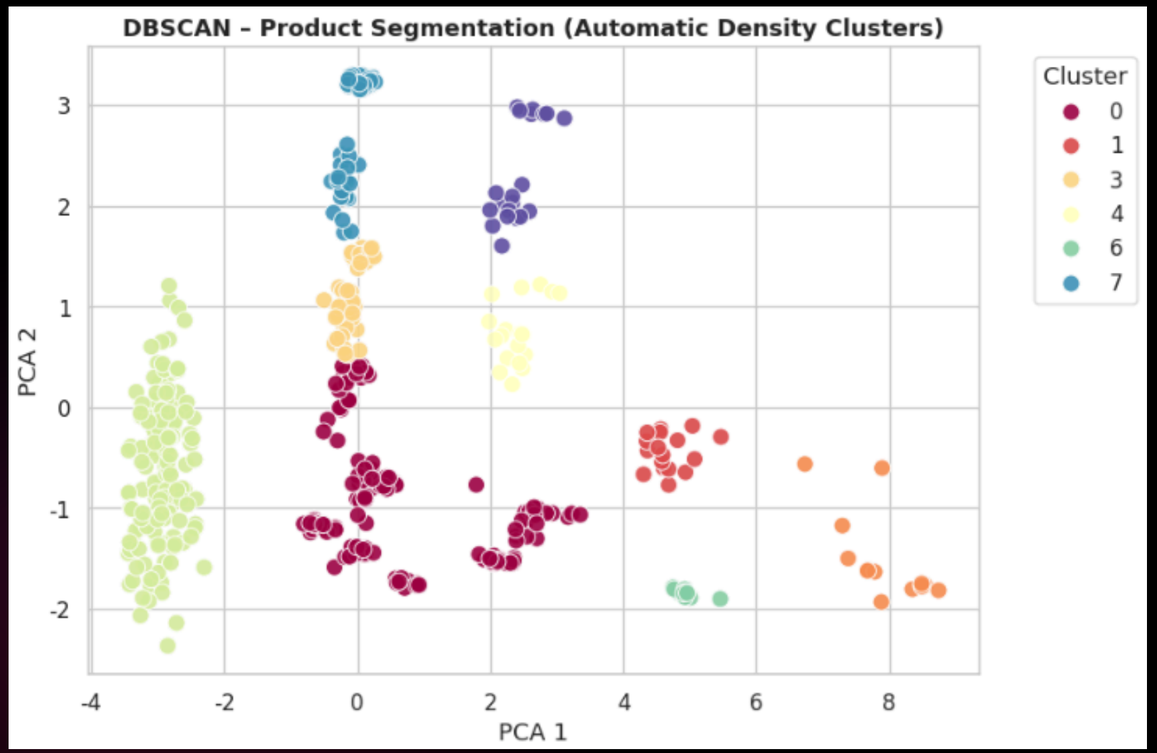
Determining the Number of Clusters

Hasil Uji Kombinasi Parameter DBSCAN (Top 10):

	eps	min_samples	n_clusters	silhouette_score
0	1.0	5	9	0.620
1	1.0	10	9	0.607
5	1.2	15	6	0.571
3	1.2	5	7	0.570
2	1.0	15	7	0.569
7	1.4	10	7	0.567
8	1.4	15	6	0.567
4	1.2	10	7	0.540
6	1.4	5	6	0.529
10	1.6	10	6	0.392

- The optimal parameters were found to be eps = 1.0 and min_samples = 5.
- With these parameters, a total of 9 clusters were formed.

Visualisasi Cluster



- Silhouette Score (DBSCAN): 0.620
- The product segmentation is fairly good, with 9 clusters consistent with other models.
- DBSCAN excels at detecting outliers and identifying product density patterns.

MODELING

Model Comparison

Algorithm	Silhouette Score	Number of Clusters	Brief Analysis
K-Means	0.6	5	Segmentation is fairly good, but some products still overlap between clusters.
Agglomerative	0.658	9	Best result with the clearest and most consistent cluster separation.
DBSCAN	0.62	9	Stable and able to detect outliers; results are consistent with Agglomerative.

- Agglomerative Clustering provided the best results with the highest Silhouette Score (0.658).
- DBSCAN produced the same number of clusters (9), indicating consistent data structure.
- The pattern of 9 product segments is likely a natural division within the data.
- K-Means remains useful due to its simplicity and speed, though its results are slightly less optimal compared to the other two models.

MODELING

Model Interpretation

Ringkasan Tiap Cluster – Agglomerative Clustering (Product Level)

	Cluster	Dominant_Category	Dominant_Product	Avg_Total_Purchases	Avg_Ratings	Avg_Total_Amount	Avg_Profitability_Index
0	0	Electronics	Action	17.218310	3.170227	4317.784019	763.880936
1	1	Electronics	Acer Swift	3475.082474	3.241343	886534.171615	164853.402912
2	2	Home Decor	Bathtub	3243.540000	3.109444	829082.857798	153858.610608
3	3	Clothing	A-line dress	3263.340000	3.101744	830536.939434	154102.322546
4	4	Books	Biography	9779.740741	3.115744	2505839.601502	461560.873377
5	5	Electronics	4K TV	6489.450000	3.120227	1652005.420458	306216.283589
6	6	Clothing	Boots	6537.850000	3.106445	1668121.831960	307273.681057
7	7	Home Decor	Bed	6460.850000	3.109950	1644557.106411	304625.828766
8	8	Grocery	Adventure	13081.153846	3.252404	3334116.591720	612850.087754

Products in the Grocery and Books categories show the highest performance, while some segments in Electronics and Clothing still require promotional strategies to boost sales and profitability.

MODELING

Model Interpretation

📊 Interpretasi Segmentasi Produk (Berdasarkan Rata-Rata Tiap Cluster – Agglomerative Clustering):

Cluster	Kategori & Produk Dominan	Karakteristik Produk	Implikasi Inventory	Nama Segmen
Cluster 0	Electronics – Action	Penjualan dan profit sangat rendah dengan pelanggan sedikit. Produk dengan performa pasar lemah dan rating standar.	Butuh promosi agresif atau reposisi produk agar lebih kompetitif.	🔴 Low-Performing Electronics
Cluster 1	Electronics – Acer Swift	Produk dengan profit dan total penjualan tinggi serta rating stabil. Mewakili segmen elektronik menengah ke atas dengan loyalitas pelanggan kuat.	Pertahankan ketersediaan stok dan dorong strategi promosi berbasis nilai tambah.	🔵 Mid-Tier Tech Segment
Cluster 2	Home Decor – Bathtub	Produk dekorasi rumah dengan performa stabil, profit cukup tinggi, dan pelanggan setia. Menunjukkan pasar yang konsisten dan loyal.	Fokus pada stok produk unggulan dan variasi desain untuk mempertahankan penjualan.	🟢 Stable Home Decor
Cluster 3	Clothing – A-line dress	Produk fashion dengan penjualan konsisten dan profit menengah. Rating stabil, cocok untuk segmen pakaian kasual atau harian.	Pertahankan desain populer dan dorong promosi musiman.	🔵 Consistent Fashion Segment
Cluster 4	Books – Biography	Produk dengan profit dan total penjualan sangat tinggi. Segmen dengan pelanggan luas dan stabilitas permintaan tinggi.	Pertahankan stok populer dan perluas distribusi untuk menjaga performa.	🟢 High-Value Literature
Cluster 5	Electronics – 4K TV	Produk elektronik premium dengan profit besar dan rating stabil. Target pasar menengah ke atas yang sensitif terhadap kualitas visual.	Fokus pada diferensiasi fitur dan kampanye teknologi canggih.	🟣 Premium Visual Tech
Cluster 6	Clothing – Boots	Produk fashion dengan profit tinggi dan permintaan stabil. Rating konsisten, cocok untuk segmen produk musiman atau tren tertentu.	Manfaatkan strategi limited edition dan stok adaptif berdasarkan musim.	🔵 Trend-Driven Apparel
Cluster 7	Home Decor – Bed	Produk furnitur bernilai tinggi, profit besar dan pelanggan cukup luas. Rating stabil dengan loyalitas kuat.	Pastikan kualitas bahan dan waktu pengiriman untuk menjaga kepuasan pelanggan.	🟢 Premium Furniture Segment
Cluster 8	Grocery – Adventure	Cluster dengan total penjualan, profit, dan jumlah pelanggan tertinggi. Menjadi pendorong utama penjualan secara keseluruhan.	Pastikan pasokan selalu terjaga dan sistem distribusi efisien.	🟢 Core Grocery Segment

Recommendation

01 Improve the performance of the low-performing segment (Low-Performing Electronics) through more active promotions, quality improvements, and better product positioning in the market.

02 Maintain the strong performance of the top segments (Core Grocery, High-Value Literature) by ensuring consistent stock availability and smooth distribution.

03 Develop flexible promotions for the mid-tier segments (Mid-Tier Tech, Trend-Driven Apparel) with innovative ideas and campaigns aligned with seasonal trends.

04 Enhance the image of premium segments (Premium Visual Tech, Premium Furniture) by improving product quality and providing exceptional customer service.

Streamlit

Let’s Connect

I'm always open to networking, collaboration, or just a quick chat.

linkedin.com/in/roy-firman-sihombing

github.com/Roysihombing

roy-firman-sihombing.free.nf/