Artificial Intelligence in planning and decision-making

Customer Segmentation for Marketing Decisions

Nigerian Retail & E-commerce Customer Segmentation Dataset

Presentation By: Joshua Daramola. Neptune

Code: BQ7I39

Problem Statement

Retail and e-commerce businesses in Nigeria face challenges in:

Understanding customer behavior

Identifying high-value vs. low-value customers

Targeting marketing campaigns efficiently

Preventing churn among profitable customers

Goal: Use AI (PCA + K-Means) to segment customers into meaningful groups and support data-driven marketing decisions.

Dataset Overview

The dataset contains 150,000 customers with 10 features:

Numerical Features

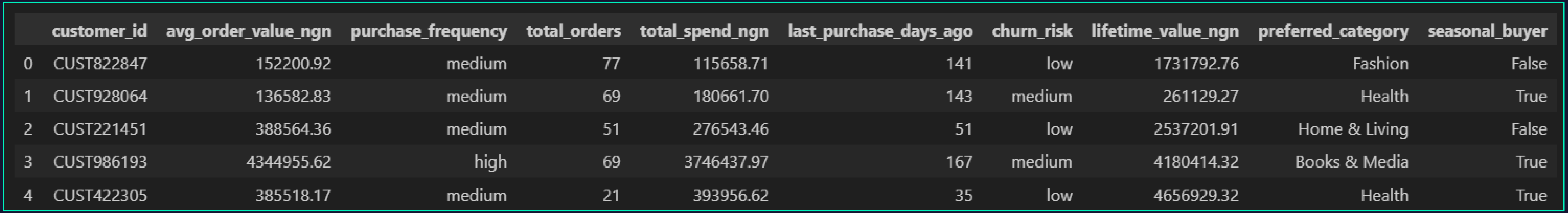
- avg_order_value_ngn
- total_orders
- total_spend_ngn
- last_purchase_days_ago
- lifetime_value_ngn

Categorical Features

- purchase_frequency (4 levels)
- churn_risk (3 levels)
- preferred_category (10 levels)
- seasonal_buyer (True/False)

Identifier

- customer_id

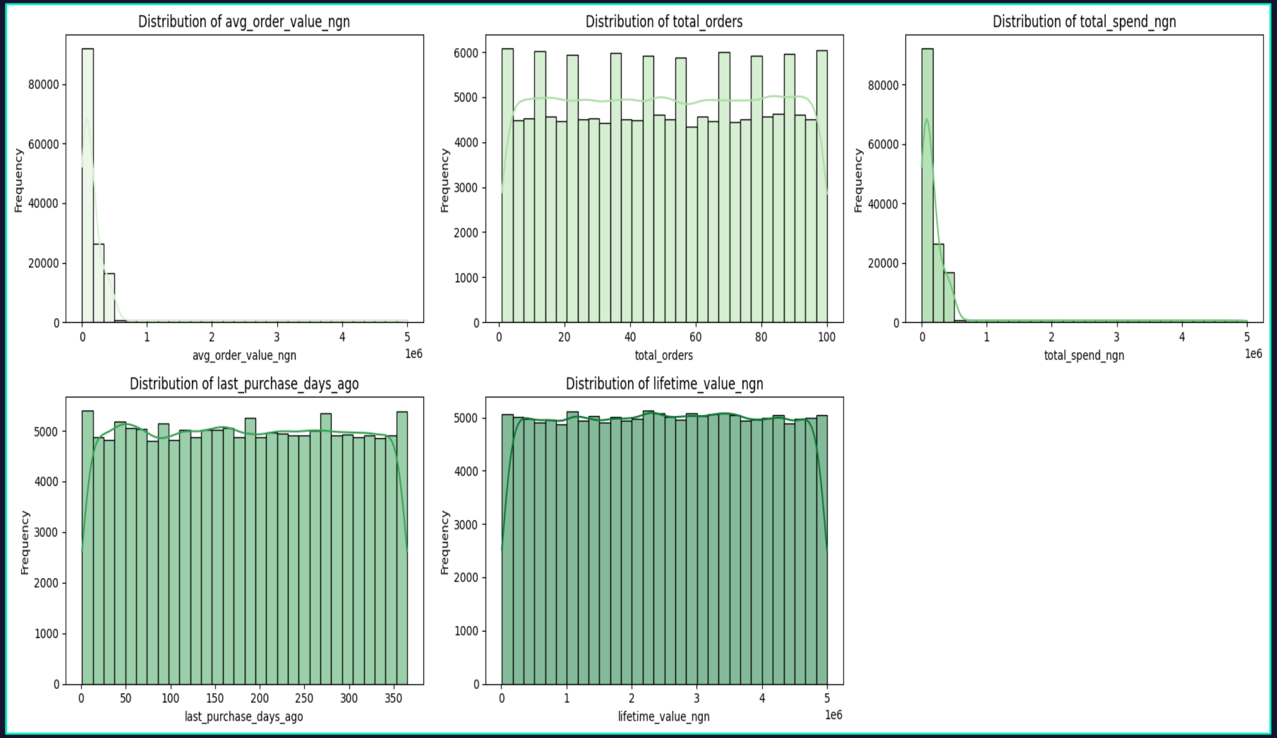


	customer_id	avg_order_value_ngn	purchase_frequency	total_orders	total_spend_ngn	last_purchase_days_ago	churn_risk	lifetime_value_ngn	preferred_category	seasonal_buyer
0	CUST822847	152200.92	medium	77	115658.71	141	low	1731792.76	Fashion	False
1	CUST928064	136582.83	medium	69	180661.70	143	medium	261129.27	Health	True
2	CUST221451	388564.36	medium	51	276543.46	51	low	2537201.91	Home & Living	False
3	CUST986193	4344955.62	high	69	3746437.97	167	medium	4180414.32	Books & Media	True
4	CUST422305	385518.17	medium	21	393956.62	35	low	4656929.32	Health	True

Exploratory Data Analytics

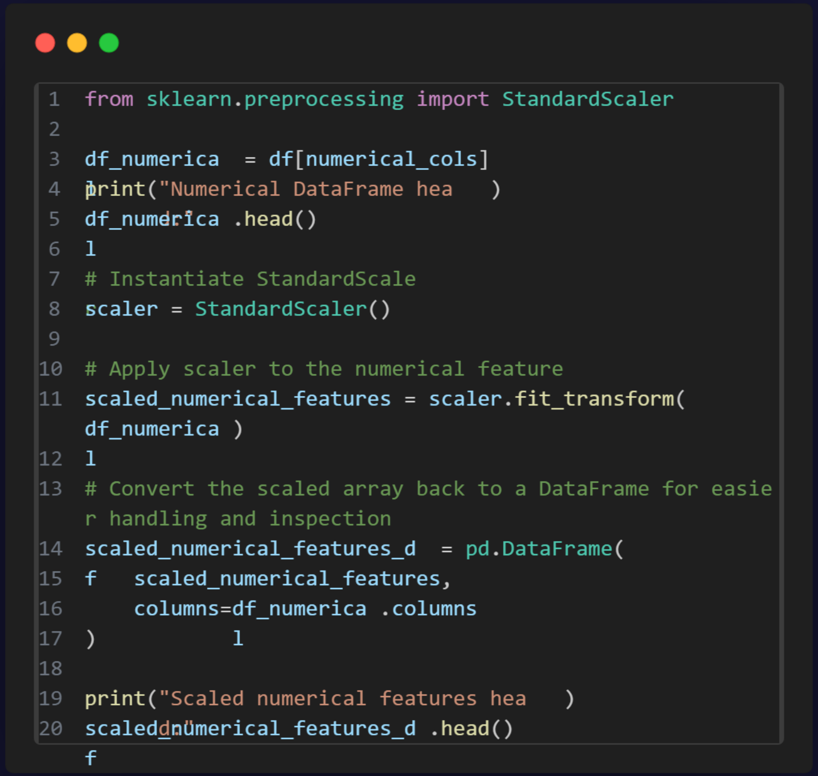
Key insights from EDA:

- Numerical features have wide ranges (e.g., total_spend up to hundreds of thousands of NGN).
- Categorical features show diversity:
  - 10 product categories
  - 3 churn risk levels
  - Balanced seasonal vs non-seasonal buyers
- High-cardinality ID column removed.
- Numerical distributions show skewness → scaling required.



PCA Preprocessing

- Selected only numerical columns for PCA.
- Scaled via StandardScaler for equal variance contribution.
- Prepared data for dimensionality reduction.



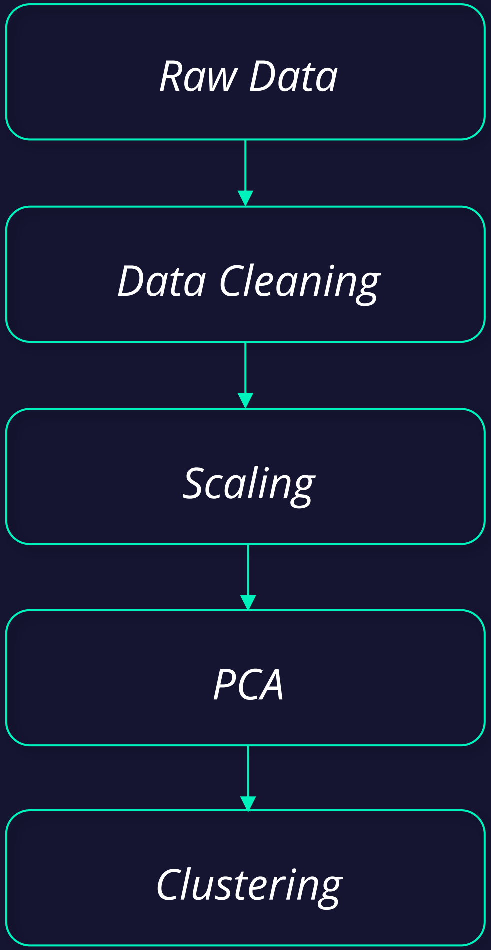
```
from sklearn.preprocessing import StandardScaler

df_numerica  = df[numerical_cols]
print("Numerical DataFrame hea   )
df_numerica .head()
l
# Instantiate StandardScale
scaler = StandardScaler()

# Apply scaler to the numerical feature
scaled_numerical_features = scaler.fit_transform(
df_numerica )
l
# Convert the scaled array back to a DataFrame for easie
r handling and inspection
scaled_numerical_features_d  = pd.DataFrame(
f   scaled_numerical_features,
    columns=df_numerica .columns
)           l

print("Scaled numerical features hea   )
scaled_numerical_features_d .head()
f
```

Flow Chart



PCA Result

PCA reduced the data into two components

Component contributions:

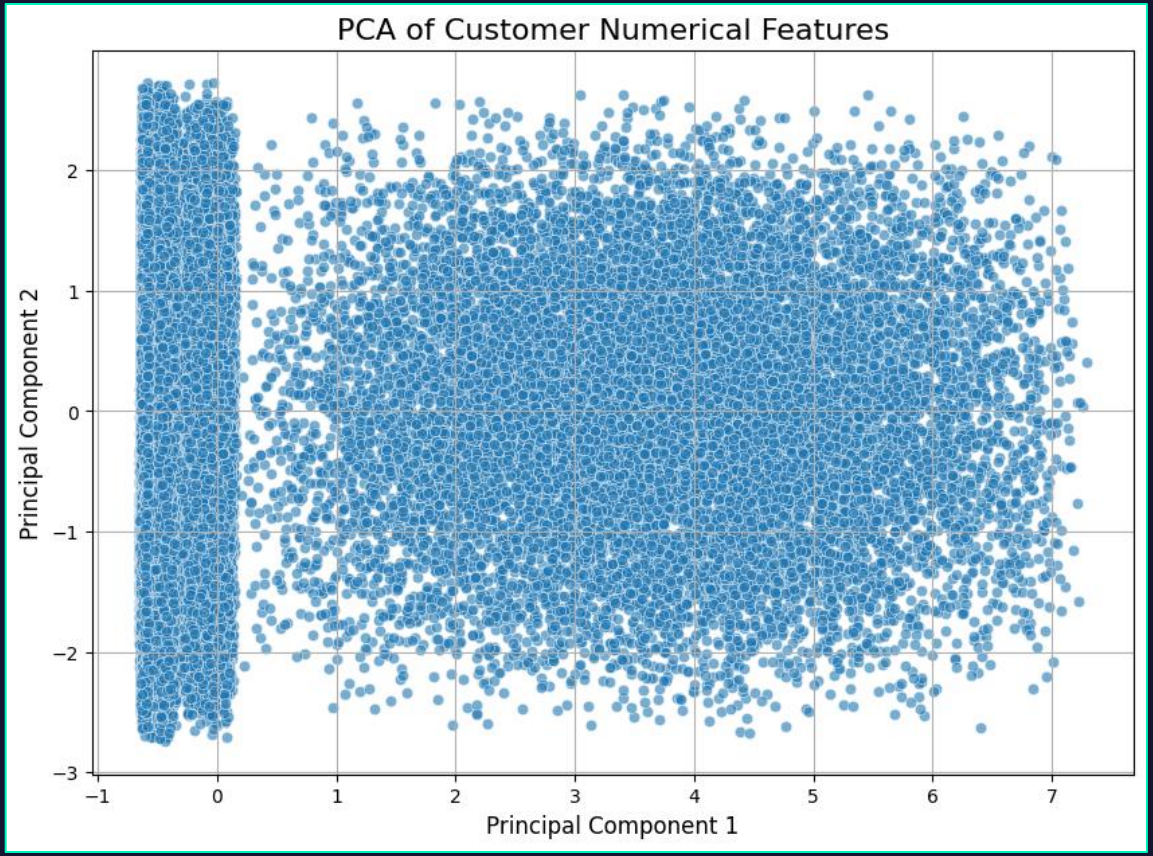
35.74%
PC1

20.11%
PC2

55.85%
Commulative

Interpretation:

- Adequate 2D representation.
- Customers cluster densely near the origin, with outliers representing unique profiles (High spenders, High Frequency Purchasers, and Long-lapsed customers)



K-means Clustering

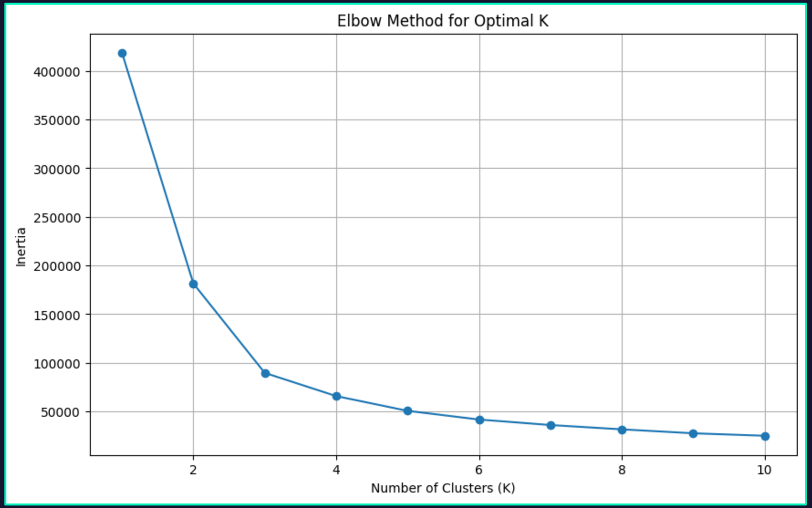
Tested K from 1 to 10.

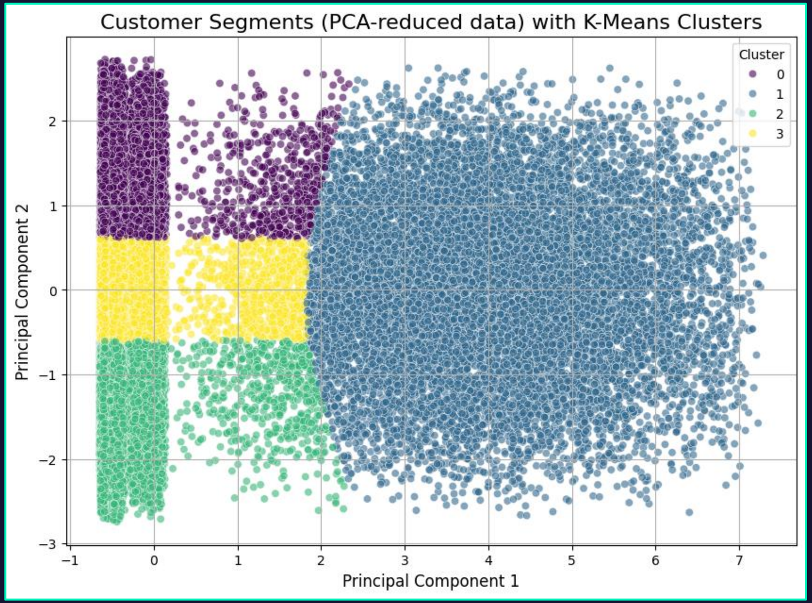
Result:

- Clear “elbow” at K = 4 indicates diminishing returns after 4 clusters.

Applied K-Means with 4 clusters.

- Assigns each customer to one segment in PCA space.
- Provides separable customer groups.





Cluster Profiles

Cluster 0 - High Lifetime Value, Less Recent

- Highest lifetime_value
- Not purchased recently
- Low churn risk

Cluster 1 – High Value, Engaged Spenders

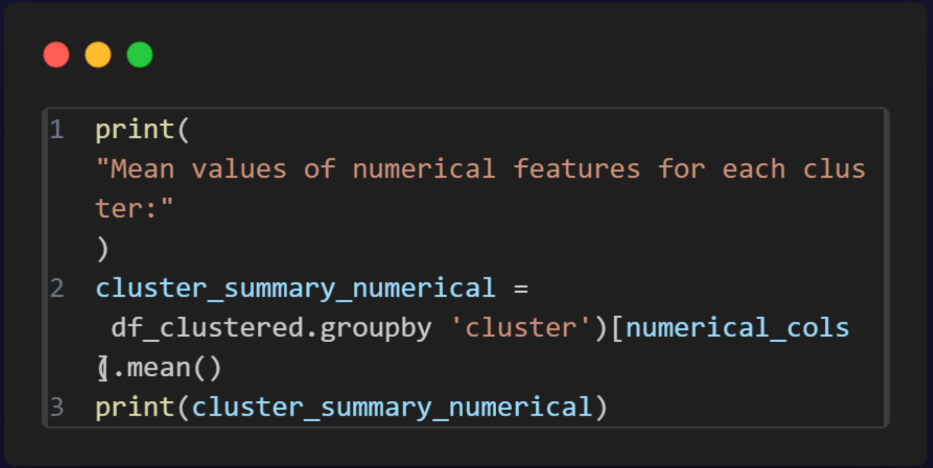
- Highest total spend & AOV
- High purchase frequency
- Prefers Books & Media, Electronics

Cluster 2 – Low Value, Recent Purchasers

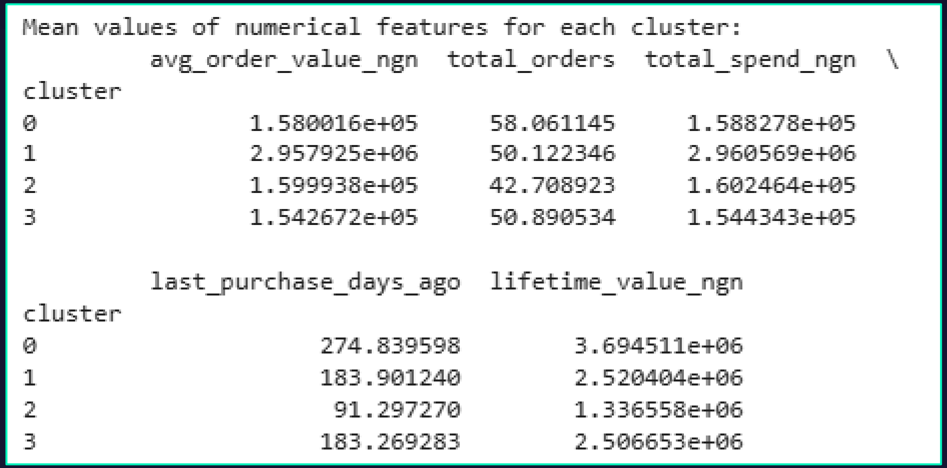
- Lowest monetary values
- Most recent purchases
- Likely new customers

Cluster 3 – Moderate Value, Average Engagement

- Average spend and recency
- Popular segment
- Prefers Fashion



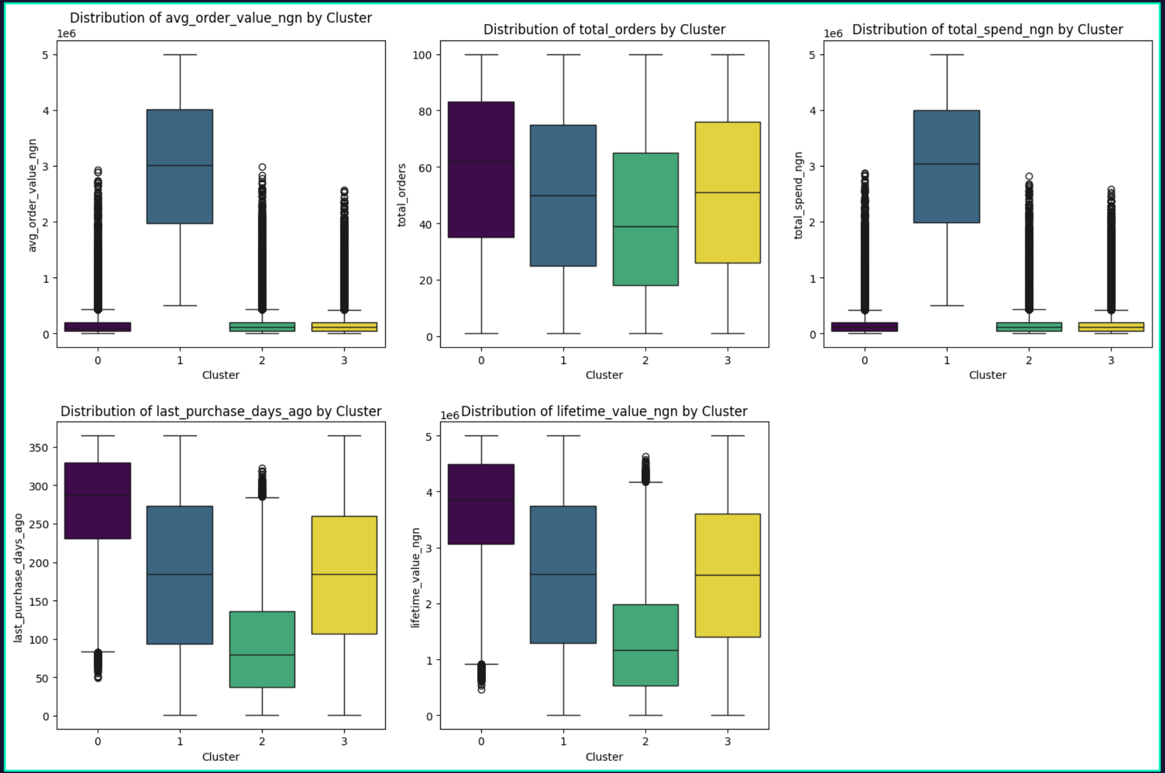
```
print(
"Mean values of numerical features for each clus
ter:"
)
cluster_summary_numerical =
 df_clustered.groupby 'cluster')[numerical_cols
].mean()
print(cluster_summary_numerical)
```

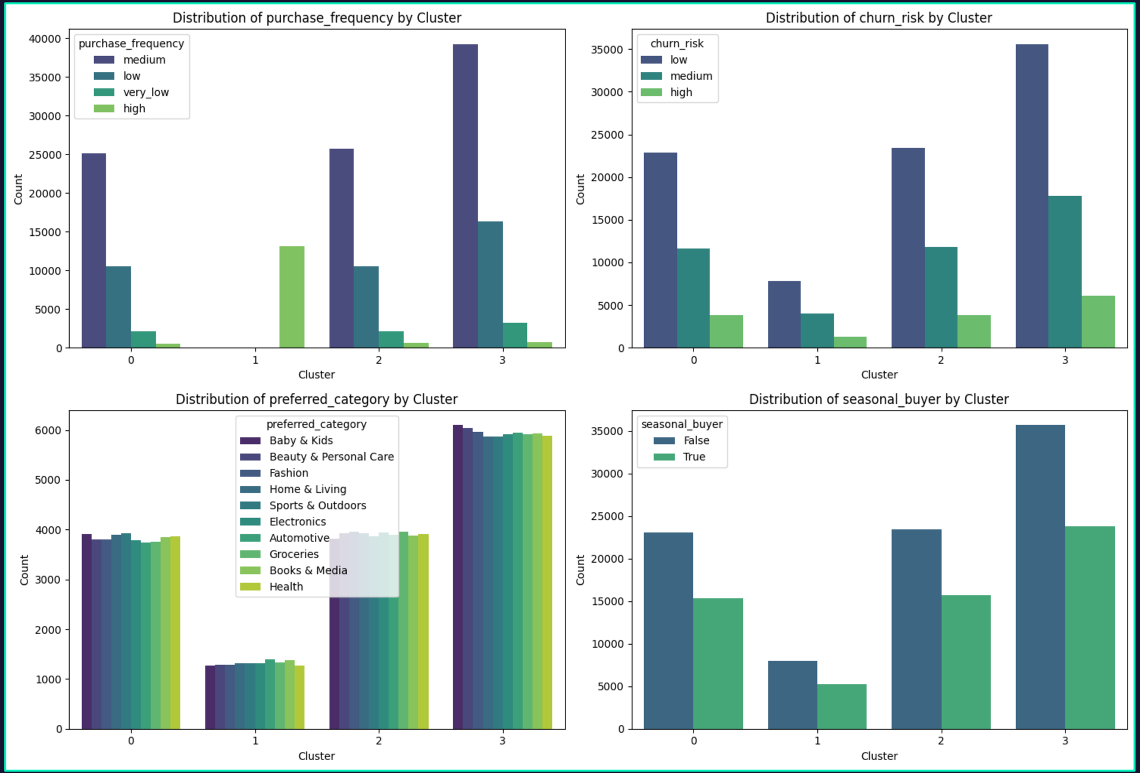


```
Mean values of numerical features for each cluster:
         avg_order_value_ngn  total_orders  total_spend_ngn  \
cluster
0               1.580016e+05     58.061145     1.588278e+05
1               2.957925e+06     50.122346     2.960569e+06
2               1.599938e+05     42.708923     1.602464e+05
3               1.542672e+05     50.890534     1.544343e+05

         last_purchase_days_ago  lifetime_value_ngn
cluster
0                    274.839598        3.694511e+06
1                    183.901240        2.520404e+06
2                     91.297270        1.336558e+06
3                    183.269283        2.506653e+06
```

Feature Difference Across Clusters





Key behavioral differences:
Cluster 1 = highest spenders
Cluster 0 = long-term high-value but inactive
Cluster 2 = low spend but active
Cluster 3 = stable mid-tier segment

Marketing Prioritization

Highest Priority:
Cluster 0
Why?

- VERY high lifetime value
- At risk due to inactivity

High Priority:
Cluster 1
Why?

- Highest spending customers

Medium Priority:
Cluster 3

- Large segment
- Moderate value

Lower Priority:
Cluster 2

- New or low-value customers

Overall Insight and Conclusion

Key Outcomes

- PCA captured 55.85% of variance.
- K-Means identified 4 actionable customer groups
- Each cluster has distinct values and behavioral signatures
- Clear evidence-based marketing strategy formed

Conclusion

AI-driven segmentation enables Nigerian e-commerce businesses to allocate budgets smartly, target profitable and at-risk customers, and optimize customer lifetime value.

Colab Link: Nigeria E-commerce customer segmentation

THANK YOU FOR YOUR ATTENTION

Appendix

The following Prompt was used in Google colab for the project;

- *Write a python code for clustering for customer segmentation so as to make a on which customer segment should the company prioritize for marketing investment and promotional targeting? Perform Exploratory data analysis first to understand the data sets and know the relationship between each variable / column. Use visualisations to show the weights and justify your decision.*

- *Prepare the numerical features of the df DataFrame by handling missing values and scaling them with StandardScaler. Apply PCA to reduce the data to 2 components and visualize the results using a Matplotlib scatter plot*

Resources:

- Images: Grok
- Environment: Google Colab
- Code Screenshots: Codesnap extension on Vscode
- Background image: Canva