Assignment 1 Report – Marketing Analytics

Customer Segmentation and Profiling

- (RetailX Case)

Submitted in partial fulfilment of the requirements for the award of the degree of

Ms. Data Science

IN

University of Europe for Applied Science, Dubai.

BY

Khaled Walid (70713127)
Mohammad Gousuddin (35197708)

Under the guidance of

Professor Eman Abukhousa

2025-2026

1. Objective

To identify natural customer segments based on behavioral data and develop targeted marketing strategies that enhance engagement, retention, and revenue for Retailx.

2. Methodology

- Clustering Technique: K-Means clustering using standardized behavioral features only.
- Feature Set for Clustering:
  - avg_order_size, avg_order_freq, crossbuy, multichannel, per_sale, tenure, return_rate, loyalty_card, avg_mktg_cnt
- Evaluation Metrics:
  - Elbow Method:
    - Formula:

$$J = \sum_{i=i}^{k} \sum_{x \epsilon Ci} ||x - \mu i||2$$

k = Number of clusters

Ci = Cluster i

μi = Centroid of cluster i

  - Silhouette Score: Highest at k = 3
    - Formula:

S= $\frac{b-a}{max(a,b)}$

a = average intra-cluster distance.

b = average nearest-cluster distance.

  - PCA Scatter: Clear separation of clusters
  - Calinsk – Harabasz:
    - Formula:

CH(K) = $\frac{Between-clust\quad dispresion}{Within-cluster\ dispersion} x\ \frac{n-k}{k-1}$

n = total number of data points

k= number of clusters

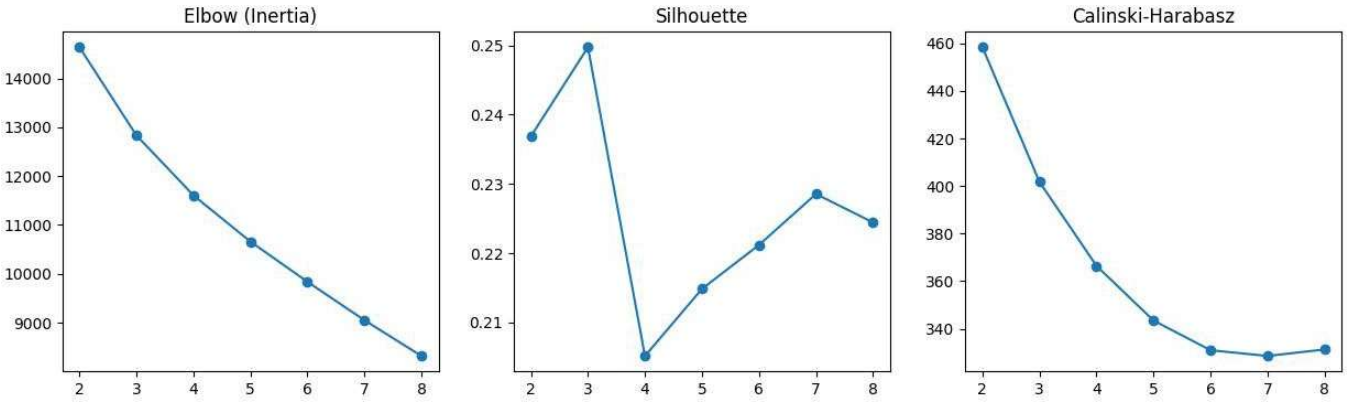
Between-cluster dispersion: Sum of squared distances between cluster centroids and the overall mean.

Within-cluster dispersion: Sum of squared distances between points and their cluster centroid.

  - Classification Accuracy:
    - Formula:

P(y=1| x) = $\frac{1}{1+e^{-(\beta_0+\beta_1+\beta_2+\ldots\ldots\ldots\beta_n)}}$

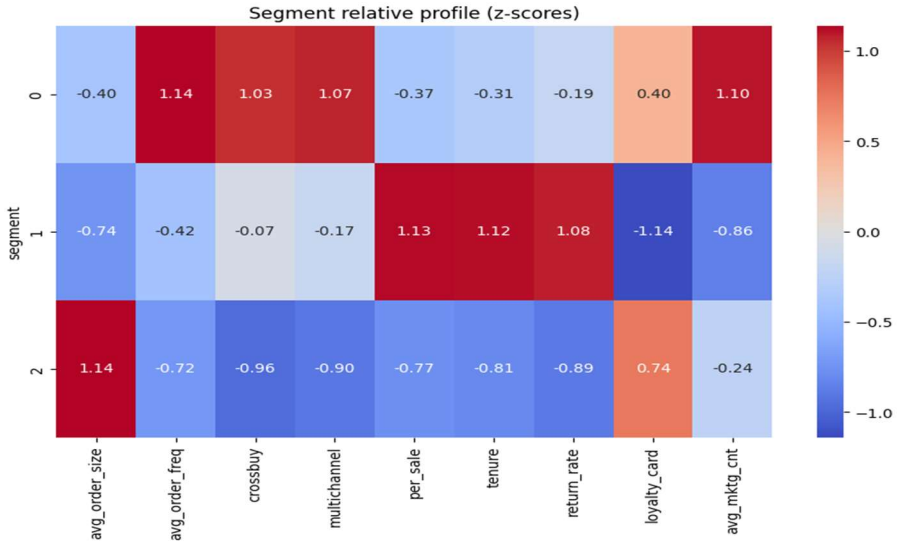
    - Logistic Regression: 98.7%
    - Random Forest: 95.0%



We tested k = 2, 3, and 4. While k = 2 had stronger statistical separation and k = 4 offered richer complexity, k = 3 provided the best balance of interpretability, behavioral diversity, and strategic clarity.

3. Segment Profiles (k = 3)

Segment	Behavioral Traits	Strategic Label	Marketing Strategy
0	High crossbuy, loyalty_card, moderate order_freq, low order_size	Loyalty-Driven Explorers	Reward loyalty, gamify engagement, and offer personalized bundles.
1	High multichannel, per_sale, tenure, low crossbuy, low return_rate	Multichannel Deal Seekers	Promote cross-channel offers, flash sales, and exclusive bundles.
2	High order_size, return_rate, low tenure, order_freq, multichannel	High-Spend Occasionalists	Focus on satisfaction guarantees, post-purchase follow-up, and retention campaigns.



4. Classification Performance

Model	Accuracy	Macro F1
Logistic Regression	**98.7%**	**98.5%**
Decision Tree	**93.5%**	**91.8%**
Random Forest	**95.0%**	**93.2%**

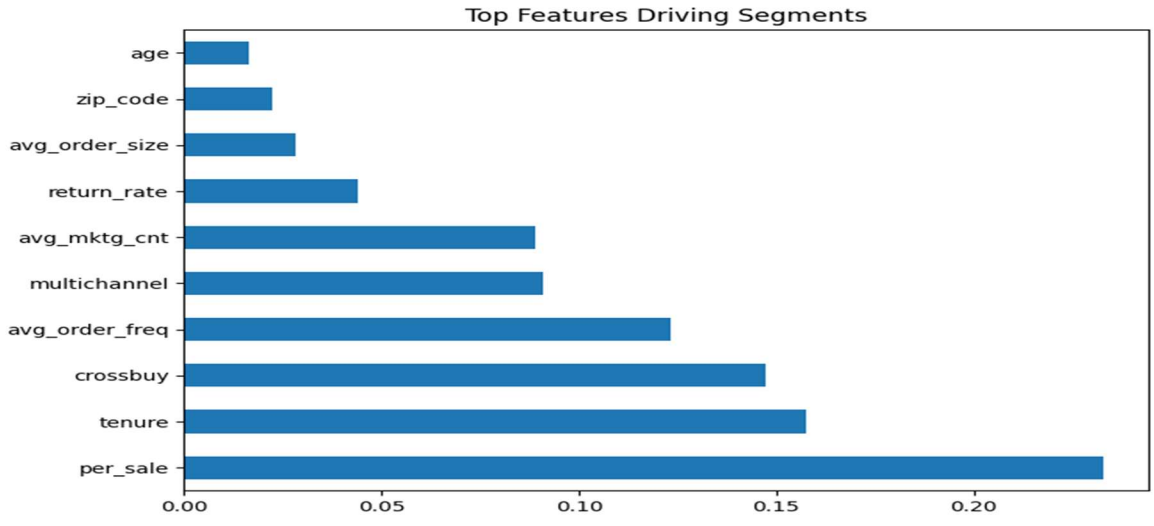
- **Confusion Matrix:** All models show strong predictive power with minimal misclassification.
- **Conclusion:** Segment membership is highly predictable using behavioral + descriptive features.

Feature Importance

```
/usr/local/lib/python3.12/dist-packages/sklearn/linear_model/_logistic.py:1247: FutureWarning: 'multi_class' was deprecated in version
  warnings.warn(

LogReg report:
              precision    recall  f1-score   support

           0       0.97      0.97      0.97       144
           1       0.99      0.99      0.99       106
           2       0.99      0.99      0.99       350

    accuracy                           0.99       600
   macro avg       0.98      0.98      0.98       600
weighted avg       0.99      0.99      0.99       600

Confusion matrix:
 [[140   1   3]
 [  1 105   0]
 [  3   0 347]]

Tree report:
              precision    recall  f1-score   support

           0       0.90      0.90      0.90       144
           1       0.92      0.87      0.89       106
           2       0.95      0.97      0.96       350

    accuracy                           0.94       600
...
avg_order_size    0.028128
zip_code          0.022257
age               0.016269
dtype: float64
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

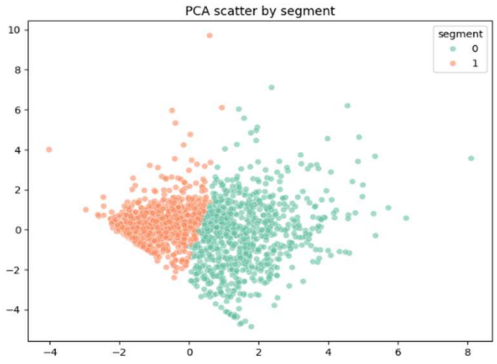
Based on a Random Forest classifier trained to predict segment membership:

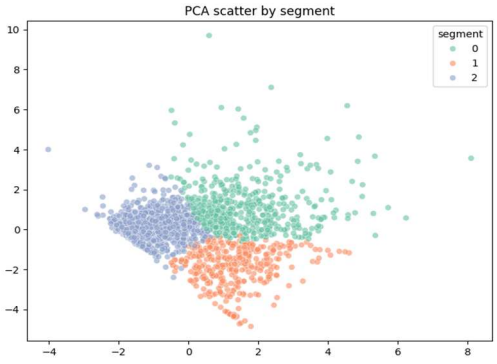
Rank	Feature	Importance
1	per_sale	0.23
2	tenure	0.16
3	crossbuy	0.15
4	avg_order_freq	0.12
5	multichannel	0.09

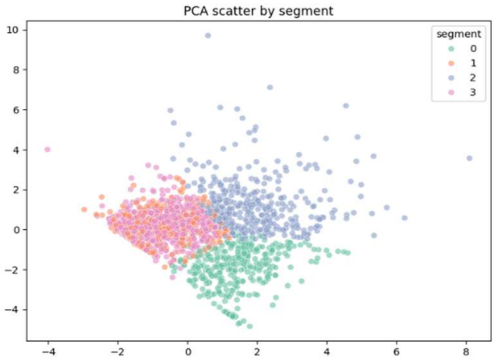


5. Strategic Implications

- **Segment 0:** Loyalty-focused campaigns, gamified rewards, and cross-category bundles.
- **Segment 1:** Multichannel promotions, flash deals, and exclusive offers for long-tenured shoppers.
- **Segment 2:** Retention-focused strategies, satisfaction guarantees, and post-purchase engagement.







6. Final Recommendation

Retailx should adopt the **k = 3 segmentation** for its marketing strategy. It reveals three distinct behavioral personas with clear engagement patterns and actionable insights. This structure balances statistical rigor with business relevance and supports scalable, targeted campaigns.

Code: Attached in .ipynb format in .zip.