Title: Customer Segmentation using Clustering(Instacart Market Basket Analysis)

AisleAffinity – Cartographer: Mapping Tastes to Carts

Business Goal:

- Group Instacart customers into meaningful behavioral segments to improve:
- **Targeted marketing**
- **Cross-sell & up-sell recommendations**
- **Expected AOV uplift: 5–8%**

Dataset Source:

Kaggle – Instacart Market Basket Analysis
Includes: orders, prior orders, train orders, products, aisles & departments
Link:
https://www.kaggle.com/datasets/psparks/instacart-market-basket-analysis

Citation: Instacart (2017), *Instacart Market Basket Analysis*, Kaggle. Retrieved from the link above.

Team Members:

- **Debjit Adak – 2025EM1100283**
- **Anisha Das – 2025EM1100192**
- **Akash Ghosh – 2025EM1100017**
- **Kritanhik Biswas – 2025EM1100104**

Supervisor:

- **Mr. Shubham Gupta**

Data & Feature Engineering (High-Level)

Dataset Summary (from proposal + typical Instacart stats):

- ~3.4M prior order lines
- ~200K customers
- ~134 aisles, 21 departments
- Products enriched using aisle & department metadata

Preprocessing Steps:

- Joined product → aisle → department
- Built baskets per **prior** order
- Cleaned missing values, duplicates, type issues
- Removed extremely rare products
- Created consolidated workspace for modeling (Phase 3)

Feature Engineering:

- **Customer Embeddings (128 dimensions)**
  Derived from product vectors (Product2Vec / aggregation of product embeddings).
- **Behavioural Features:**
  - Prior order count
  - Average basket size
  - Reorder ratio
  - Mean days between orders
  - Basic RFM-style proxies

Flow Diagram:

Raw Instacart Data → Cleaning/EDA → Feature Engineering → Customer Vectors

Clustering Models & Metrics

Clustering Models Used:

- **KMeans** – simple centroid-based baseline
- **Agglomerative / Hierarchical** – tree-based clustering using linkage merges
- **UMAP + HDBSCAN** – density-based approach; handles noise and identifies micro-segments

Evaluation Metrics (Test Set):

Model	Silhouette ↑	Davies-Bouldin ↓
KMeans	~ 0.32	~0.92
Hierarchical	~ 0.30	~ 1.71
HDBSCAN	~ 0.60	~ 0.45

(Silhouette: higher = better; DBI: lower = better)

Conclusion:

- **HDBSCAN provides the best overall cluster quality**
  - → Highest separation
  - → Smallest intra-cluster variance
- **Hierarchical** shows similar structure but weaker numeric scores
- **KMeans** kept as a baseline reference model

Final Segments & Behaviour Insights (HDBSCAN + RFM)

Key Visuals:

- **Bar Chart:** Customers per HDBSCAN cluster (noise removed)
- **RFM Behaviour Plots:**
  - Recency, Frequency, Avg. Basket Size (Monetary-proxy)
- **Cluster Behaviour Comparison:**
  - Reorder ratio
  - Days-since-last-order

Segment Insights (HDBSCAN):

Segment 0 – Loyal High-Value Buyers

- High frequency, large baskets
- Low recency (buy often)

Action: Loyalty rewards, premium cross-sell bundles

Segment 1 – Regular Moderate Buyers

- Medium basket size, stable frequency

Action: Personalized aisle/department recommendations

Segment 2 – At-Risk / Churn Segment

- High recency (long gaps), low engagement

Action: Reactivation offers, push notifications

Segment 3 – Bulk but Infrequent Buyers

- Large baskets but low frequency

Action: Bulk-purchase deals, stock-up reminders

Conclusions

1. Built complete end-to-end segmentation pipeline

   (Cleaning → Features → Clustering → Insights)

2. Tried 3 clustering techniques;

   HDBSCAN clearly performs best (Silhouette ≈ 0.60, DBI ≈ 0.45)

3. RFM-style analysis confirms meaningful customer groups:

- High-value loyal customers
- Regular moderate buyers
- Infrequent / at-risk customers
- Occasional bulk buyers

Business Actions:

- **High-value:** loyalty rewards, premium cross-sell bundles
- **At-risk:** retention messages, discount reactivation
- **Bulk buyers:** offer category-specific deals
- **General:** segment-based email campaigns