Project Milestone

Project Title：Customer Segmentation Analysis for DataCo Global
Team Members: Yufei Shang, Shrinidhi Bhide, Ohm Srikiatkhachorn, Wenlin Zhao
Date: 03/2/2025

I. Proposal

1.1 Problem Statement

In today's highly digitalized retail landscape, businesses must go beyond traditional customer segmentation methods, which primarily rely on demographic and transactional data. These approaches fail to capture pre-purchase intent, browsing behaviors, and engagement trends, limiting their effectiveness in personalized marketing and product recommendations.

This project aims to develop a multi-faceted customer segmentation framework that integrates transactional data (purchases, order characteristics, discount usage) and aggregated behavioral data to enhance marketing strategies and customer insights. Rather than merging individual browsing sessions with purchase history, we aggregate product-level behavioral insights to analyze how browsing patterns influence conversions.

Key objectives:

1. **Behavioral Segmentation with Aggregated Clickstream Data**
   - Aggregate product-level browsing patterns to identify high-interest but low-conversion products.
   - Compare total product views vs. unique visitor count to understand engagement levels.
2. **Temporal and Sequential Pattern Analysis**
   - Analyze peak browsing hours vs. purchase times to optimize marketing strategies.
   - Identify common browsing paths leading to purchases through time-series clustering.
3. **Hybrid Clustering for Customer and Product Segmentation**
   - Use K-means and hierarchical clustering for purchase-based customer grouping.
   - Apply the modeling to classify products based on browsing interest.

By leveraging these techniques, our approach provides actionable insights to optimize targeted marketing, improve conversion rates, and refine customer retention strategies.

II. Exploratory Data Analysis

2.1 Data Source: DataCo SMART SUPPLY CHAIN FOR BIG DATA ANALYSIS

2.2 Summary Statistics

A. For DataCosupply ChainDatasets (Transactions: df1) ,

	Days for shipping (days)	Sales per customer ($)	Order item discount rate (%)
Mean	3.5	183	10.17
Median	3	163.99	-
Std	-	120	-

These key statistics will be used to analyze customer spending behavior, shipping performance, and discount impact on sales.

B. For Tokenized_access_log datasets (Clickstream: df2),

	Mean (hr)	Median (hr)	Std (hr)	25-75th percentile

Hourly activity	14.56	15	5.57	10 - 20

These statistics help identify peak browsing hours for customer engagement analysis The following steps outline our approach and key findings in detail.

1. Correlation Analysis

To understand relationships among numeric variables, we generated a correlation heatmap using the Pearson correlation coefficient. The heatmap revealed several variables with very high correlations (close to 1 or 0.999), suggesting potential multicollinearity, which could impair model performance. For instance, features like ‘Order_profit_per order’, and’ Sales per customer…’ displayed exceptionally high correlations with other variables. To address this, we dropped these highly correlated features to simplify the model and reduce redundancy. Additionally, a significant correlation of **0.89** was observed between 'Department ID' and 'Category ID', implying a hierarchical or structured relationship. This insight can guide feature selection or engineering in future analyses.

2. Distribution and Skewness Analysis

We examined the distribution of each numeric variable using histograms and Kernel Density Estimation (KDE) plots. The histograms indicated that several variables were skewed:

- Left-Skewed: 'Benefit per order' with a skewness of -4.74.
- Right-Skewed: 'Sales' (2.88) and 'Product Price' (3.19).

To mitigate skewness, we applied: 1. Log Transformation: On right-skewed variables using np.log1p to compress high values and enhance normality.2. Exponential Transformation: On the left-skewed variable using np.exp to spread low values more evenly.

3. Outlier Detection and Removal

We used the Interquartile Range (IQR) method on Sine-Cos Scaling to identify and handle outliers. Outliers were defined as values falling outside [Q1 - 1.5 * IQR, Q3 + 1.5 * IQR] for each variable. The results were:

- Data shape before outlier removal: 180,519 rows.
- Data shape after outlier removal: 180,504 rows.
- Only 15 rows were removed, indicating that the dataset was relatively clean initially.

Key Insights

- Multicollinearity Reduction: Successfully dropped highly correlated features to simplify the dataset and enhance model interpretability.
- Skewness Correction: Log and exponential transformations effectively reduced skewness, preparing the data for algorithms sensitive to normal distributions.
- Outlier Management: Outliers were efficiently handled with minimal data loss, improving model stability.
- Data Quality: No missing values detected, indicating a high-quality dataset ready for model development.

Categorical Variables: In the analysis of categorical variables for df1, we examined value counts and identified states like 'UT', 'NM', 'LA', and 'WA' with high order volumes. The 'Customer Country' column was simplified by removing 'Puerto Rico', reducing the dataset to 15 columns. Count plots revealed category popularity and potential imbalances, while bar charts highlighted preferences between 'Shipping Mode' and 'Customer Segment', offering insights for targeted marketing.

A new 'Loyalty' feature was created to capture repeat purchase behavior, defined as customers buying 20 or more items per year. Its correlation with the 'Order Item Discount Rate' suggested discounts might influence loyalty. Additionally, time-based features ('Order Hour', 'Order Day', 'Order Month', and 'Order Year') were added to uncover ordering patterns, identifying peak times for sales. These preprocessing steps refined the dataset for modeling and provided valuable insights into customer behavior and category performance.

III. Analysis & Experiments

The goal of Analysis: The goal was to reduce the dimensionality of the dataset using PCA and identify meaningful clusters with K-means clustering, thereby simplifying the data while retaining essential information for further analysis.

Method 1: Clustering with PCA And k_means

1. Preparation for df1 (transactions):
    - Applied one-hot encoding to categorical variables and sine-cosine transformations to cyclical features like months and days.
    - Standardized numerical features to eliminate bias and ensure equal contribution.
2. PCA for All Features:
    - Applied PCA to the entire dataset, revealing that approximately 30 principal components capture 95% of the variance, while 20 components capture 90%.
    - This suggests that a significant dimensionality reduction is possible without losing much information.
3. PCA for Numerical Features:
    - Performed PCA on standardized numerical features alone.
    - Three principal components captured over 90% of the variance, with the first component contributing the most.
    - A PCA loadings heatmap highlighted that 'Category Id', 'Department Id', 'Sales', and 'Product Price' were significant contributors.
    - Clustering: We use the Elbow Method (optimal k = 4) and Silhouette Score (optimal k = 3) for K-means clustering.
        - Visualizations showed reasonable cluster separation, indicating effective dimensionality reduction and meaningful clustering.
4. PCA for All Except ‘State’:
    - Excluded the ‘State’ feature, combined encoded categorical variables with standardized numerical features.
    - About 10 principal components captured 90% of the variance, and 14 captured 95%.
    - PCA loadings heatmap identified 'Category Id', 'Department Id', and 'Sales' as significant contributors.
    - Clustering: Elbow Method suggested k = 4; the Silhouette Score suggested k = 2.
    - Visualizations with 2D and 3D scatter plots indicated meaningful data separations.
5. Parameter Tuning and Adjustments: K-means Clustering:
    - Conducted parameter tuning using distortion scores and silhouette scores.
    - Selected k = 4 for deeper insights despite the silhouette score suggesting k = 2 or 3.

Conclusion: PCA effectively reduced dimensionality while preserving critical information. The K-means clustering, guided by PCA, identified meaningful patterns in the data, validating the approach’s effectiveness.

Df2 clickstream: PCA was applied to df2 (clickstream data) to reduce dimensionality after one-hot encoding categorical variables and standardizing numerical features. The cumulative explained variance showed that about 10 components could capture over 95% of the variance, indicating effective dimensionality reduction. Key features like unique_products, total_visits, and temporal variables significantly contributed to the principal components. The Elbow Method suggested k=4 as the optimal cluster count, while the Silhouette Score indicated k=2. Overall, PCA helped simplify the data while retaining most information, enabling meaningful clustering results.

PCA with Hierarchical Clustering

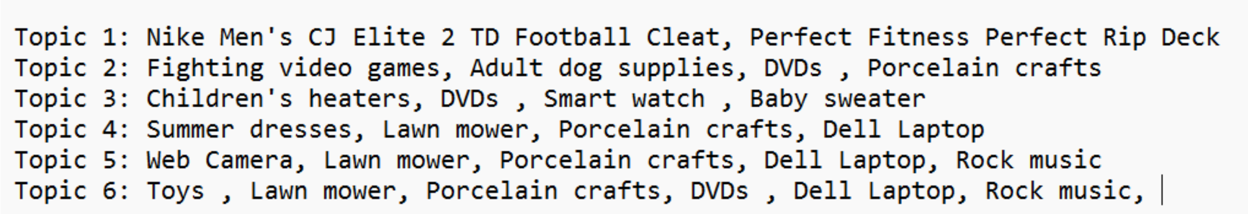
After the data preparation step, we filtered the data with Department ID and looked at the user and product data to identify patterns with hierarchical clustering. We created a pivot table that had unique User IDs as indexes and Products within that Department as columns. The values of the table were the total number of items the user bought in the whole dataset duration. There were 11 departments, from 2 to 12, and we applied PCA on all of them. The explained variance was never effectively captured by the first two PCs for any department, so visualization in 2D wasn't effective.

Method 2: UMAP:

UMAP Application: We applied UMAP after the data preparation step with n_neighbors=30, min_dist=0.1, n_components=2, and random_state=42 for dimensionality reduction. The UMAP plot showed distinct clusters, indicating effective structure capture compared to PCA.
Clustering with K-Means: K-Means clustering was performed on UMAP-reduced data with k=2. The silhouette score was 0.691, suggesting well-defined clusters with clear separation.
Conclusion: UMAP effectively captured complex structures and improved clustering quality, but interpretability and parameter tuning remain challenges. The visualization looked the same despite hyperparameter tuning.

Method 3: Topic Modeling

Similar to the pivot table in the Hierarchical Clustering model mentioned above, we generalized the pivot table to include all the departments and all the products. So the new pivot table had the unique User IDs as indices and Product Names as columns, with the total number of items they've ever bought as the values in the table. For this, we applied Topic Modeling and created the Heights and Weights matrices. The number of topics were adjusted for their reconstruction errors. The six topics and their top three corresponding products were:



```
Topic 1: Nike Men's CJ Elite 2 TD Football Cleat, Perfect Fitness Perfect Rip Deck
Topic 2: Fighting video games, Adult dog supplies, DVDs , Porcelain crafts
Topic 3: Children's heaters, DVDs , Smart watch , Baby sweater
Topic 4: Summer dresses, Lawn mower, Porcelain crafts, Dell Laptop
Topic 5: Web Camera, Lawn mower, Porcelain crafts, Dell Laptop, Rock music
Topic 6: Toys , Lawn mower, Porcelain crafts, DVDs , Dell Laptop, Rock music,
```

IV.Challenges, Dead Ends & Adjustments

1. **Preprocessing Limitations:**

- Since PCA can only be applied to numerical features, we used one-hot encoding for categorical variables. This transformation significantly increased the number of columns, making the dataset sparse and potentially causing a loss of interpretability for categorical information.

- The use of sine and cosine transformations for cyclical features like months and days also expanded the dataset dimensions, complicating the analysis.

2. **PCA Dimensionality Reduction Limitations:**
   Despite applying PCA, the dimensionality was not reduced as much as expected for the entire dataset; however, in the case of topic modeling we apply PCA after topic modeling we see the reduction of some dimensions

3. Impact of Excluding 'State':

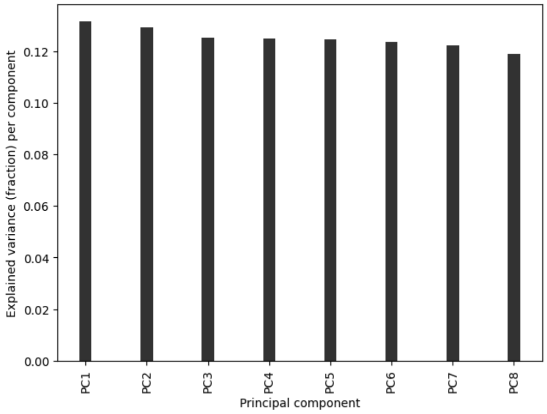
To address the complexity, we tried removing the 'State' feature and focused on numerical analysis. However, even after excluding 'State', a significant number of principal components were still required to retain 90-95% of the variance, suggesting that other features also contribute heavily to the dataset's complexity.

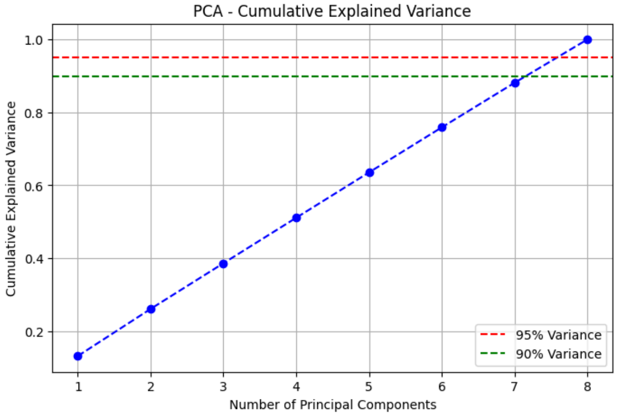
4. UMAP part challenges:

- Interpretability: UMAP lacks direct interpretability of original features, unlike PCA.
- Parameter Sensitivity: Results depend heavily on hyperparameters such as n_neighbors and min_dist.

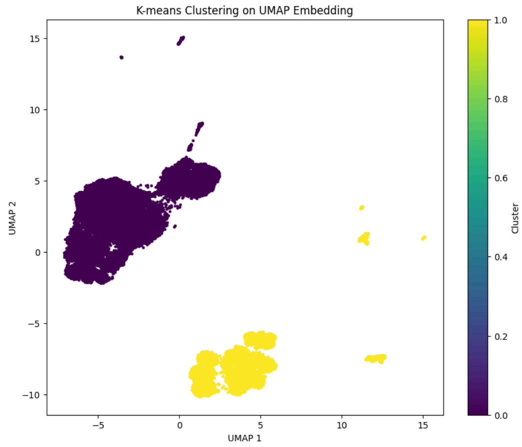
V.Findings and Interpretations

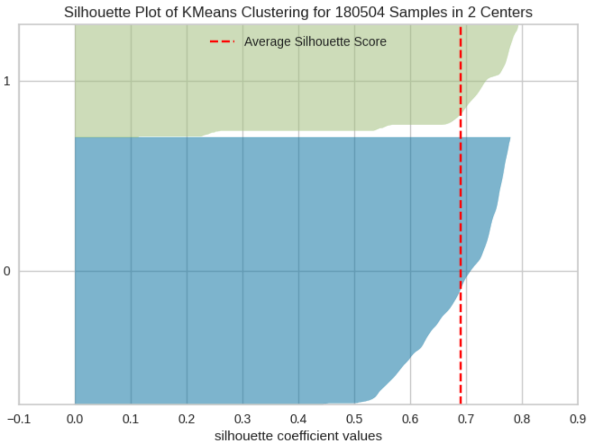
PCA with Hierarchical Clustering on df1 (transactions): PCs could not capture the variance in the first two PCs for most of the departments. For instance, this is what it looked like in Department 5, which contained aggregated transactions per user ID for sports apparel:



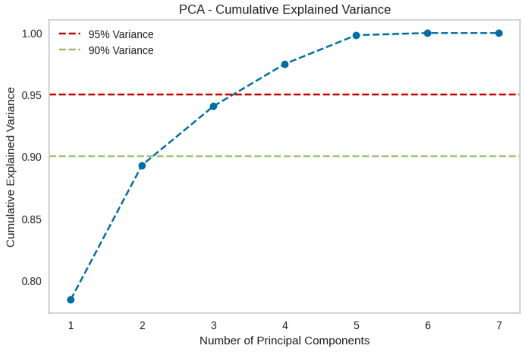


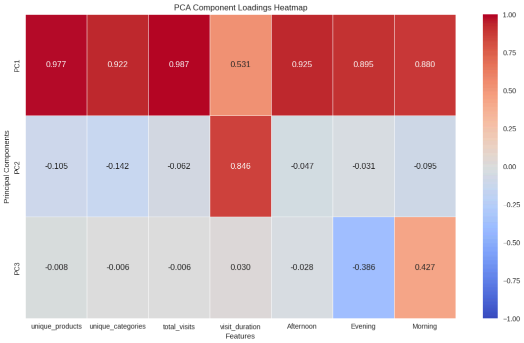
UMAP: UMAP clustering had the best silhouette score of 0.7, but the scope for qualitative interpretation was non-existent.





PCA and clustering on clickstream data: PC1 explains customer activity diversity (variety of products/categories and visit frequency). PC2 captures older users vs new users. PC3 differentiates between morning and evening browsing habits.





Application to the real world:

Our approach can be directly applied in retail analytics to improve targeted marketing and personalized recommendations. By using PCA to reduce dimensionality, we simplify complex datasets without losing essential information, enabling efficient analysis of customer behaviors. The clustering results help identify distinct customer segments based on browsing patterns, purchase history, and product interest. Retailers can leverage these insights to optimize marketing strategies, enhance product recommendations, and improve conversion rates. For instance, identifying high-interest but low-conversion products allows for focused promotions, while time-based analysis can inform targeted campaigns during peak browsing and purchase hours.

VI. Appendix

Contribution Table

Name	Done	Failures	Challenges/Dead ends
Yufei Shang	PCA, K_means cluster for df1, K_means and hierarchical cluster for df2, combining final colab code	Unclear classification boundary. The PCA process including states (categorical variables with lots of values) does not work.	Categorical encoding, High dimensions, Insufficient variance capture, the skew distribution of the dataset
Shrinidhi Bhide	UMAP, PCA & hierarchical for Df1, topic modeling	Df2 apriori to determine important months, PCA after Topic modeling	PCA hierarchy had low variance explained, UMAP with KMeans lacked interpretability
Ohm Srikiatkhachorn	Topic Modeling, Clustering, PCA for df1	PCA after Topic modeling	PCA of the pivot segment of df1 without adding 'sale'' and 'discount rate columns'
Wenlin Zhao	Data cleaning, data preprocessing, Feature selection and PCA, M2 reporting	Data loss, Overfitting Limited reduction, Information loss	High dimensions, Categorical handling, Insufficient reduction.

Distribution of Work(%)	Grades(/30)
[25,25,25,25]	[30,30,30,30]

Colab Link:
https://drive.google.com/file/d/1SwuREC0R9m6CH6TltL2ZSxva40SgG0vx/view?usp=drive_link

References/Gen Ai: We used generative AI for our PCA and method parts.
UMAP Generative AI link
Feactures Selection Gen AI
General Questions for Coding

Timeline:

Step 1 till Feb 5	Step 2 Feb 12	Step 3 Feb 19	Step 4 Feb 26	Step 5 Feb 28	Step 6 Mar 1	Step 7 Mar 3
Dataset finalization	EDA and preprocessing	Preprocessing finalization	Df2 apriori and analysis (dropped)	Df1 PCA clustering, topic modeling	UMAP, hierarchical, df2 hourly analysis	Finalization