Project Proposal Report

Project Title: Unsupervised Learning for H&M Customer Segmentation

---

1. Problem Statement:

H&M operates an extensive online store with millions of customers and over 100,000 articles. This vast scale makes it challenging to provide relevant and personalized shopping experiences. The current marketing and recommendation strategies may be too general, leading to inefficient resource allocation and sub-optimal engagement. We lack a clear, data-driven understanding of the diverse customer base.

---

2. Business Goal:

The primary goal is to **develop an interpretable machine learning model (clustering)** to **segment the H&M customer base** into 5-7 distinct groups based on their purchasing behavior and demographics. This model will provide the Marketing and E-commerce teams with **actionable insights** for **personalized communication and product targeting**, ultimately aiming to **increase customer engagement and overall sales conversion within specific segments.**

---

3. Data Source

We will use the "H&M Personalized Fashion Recommendations" dataset. This large-scale, real-world transactional dataset is ideal for deriving behavioral and preference features essential for effective customer segmentation.

- **Source Platform:** Kaggle
- Full Citation:
  H&M Group. (n.d.). H&M Personalized Fashion Data [Data set]. Kaggle. Retrieved October 8th, 2025, from https://www.kaggle.com/datasets/sohyunjun0401/h-and-m-personalized-fashion-data

---

4. Tools & Technologies

- **Programming Language:** Python
- **Core Libraries:**
  - Data Manipulation & Analysis: **Pandas, NumPy**
  - Machine Learning: **Scikit-learn** (for Clustering, PCA, and Scaling)
  - Data Visualization & Storytelling: **Matplotlib, Seaborn**
- **Development Environment:** [Our team will primarily use **Google Colab** which makes cross team collaboration easier]
- **BI Tools (Optional):** Tableau or Power BI to visualize the final segment profiles.

---

5. Project Workflow

The project will follow a structured data science lifecycle, with a heavy emphasis on **Feature Engineering** for effective clustering:

Data Acquisition→Data Cleaning & Preprocessing→Feature Engineering→Exploratory Data Analysis (EDA)→Clustering & Evaluation→Reporting & Visualization

Step	Focus Area	Key Activities

1. Data Acquisition	Raw Data Retrieval	Fetch `transactions.csv`, `customers.csv`, and `articles.csv` from Kaggle.
2. Preprocessing	Data Quality	Handle missing `Age` and `Club_Member_Status` values. Encode categorical variables for merging.
3. Feature Engineering	**Customer Profile Creation**	Calculate **RFM** (Recency, Frequency, Monetary) metrics. Derive customer preference features (e.g., favorite product group, color).
4. EDA	Feature Validation	Analyze the distribution and correlation of the new customer profile features. Perform **Feature Scaling** (Standardization).
5. Modeling	Unsupervised Learning	Apply **K-Means Clustering**. Determine the optimal number of clusters (K) using the **Elbow Method** and **Silhouette Score**.
6. Evaluation	Cluster Interpretation	Analyze the mean/mode values of features within each cluster to create clear, descriptive **Segment Profiles**.
7. Visualization	Reporting	Use **Principal Component Analysis (PCA)** to visualize the distinct segments in 2D/3D space. Present segment profiles and key findings.

---

6. Data Extraction

The large size of the H&M dataset necessitates a robust, reproducible extraction process:

- **Source Files:** The three main CSV files (`transactions.csv`, `customers.csv`, `articles.csv`) will be acquired directly from the Kaggle repository.
- **Automate the Process:** We will write a Python script utilizing the official **Kaggle API** to connect to the source, download the zipped files, and ensure a professional, reproducible workflow.
- **Loading:** The script will handle the unzipping of the downloaded files and load the data directly into **Pandas DataFrames**, making it immediately available for the Feature Engineering phase.
- **Notebook:** `data_extraction.ipynb`

---

7. Schema/Data Dictionary

An initial high-level schema from the source files will be used, with a detailed Data Dictionary of the **engineered features** created as part of the project documentation.

Original File	Key Columns to Use	Purpose
transactions.csv	customer_id, article_id, price, t_dat	**Primary source** for RFM () and behavioral feature calculation.
customers.csv	customer_id, age, club_member_status	**Demographic features** for segmentation.
articles.csv	article_id, product_group_name, colour_group_name	**Content features** used to derive customer product preference.