Curating and Analyzing Customer Personality Data for Market Segmentation

Overview

Customer segmentation is an important component of marketing, enabling organizations to understand the unique needs, preferences, and behaviors of their clients. By identifying distinct groups within a customer base, businesses can design targeted marketing strategies, improve customer satisfaction, and optimize resource allocation. The Customer Personality Analysis (CPA) dataset, available on Kaggle, provides an opportunity to explore customer demographics, purchasing behavior, and campaign responsiveness for approximately 2,200 customers.

This project focuses on curating, cleaning, and enhancing the dataset to make it suitable for robust analysis. The ultimate goal is to explore customer profiles, identify patterns in purchasing behavior, and apply segmentation techniques to uncover actionable insights. In addition, to address the limitations of the original dataset and reduce potential biases, a larger synthetic dataset will be created to simulate a population of 10,000 customers. This will allow for more robust statistical analysis, better representation of potential market segments, and the ability to model more realistic spending patterns and demographic diversity.

Plan

The project will follow a structured approach, emphasizing reproducibility, transparency, and high-quality data curation. The workflow includes several key phases:

- **Create and Synthesis:** The initial step involves acquiring the dataset from Kaggle, recording provenance information such as source, version, and licensing, and creating a

metadata template to document the structure and content of the dataset. To enhance dataset diversity and address limitations of the original data, a synthetic dataset will be generated using Python libraries such as NumPy and Pandas. Distributions of demographic and purchasing variables will be designed to reflect real-world market trends in the United States, including income levels, age ranges, marital status, and spending behavior. The synthetic dataset will also introduce controlled missing values, outliers, and variability to simulate common real-world data challenges and provide an opportunity to test data cleaning and preprocessing methods.

- **Appraisal and Selection:** The original dataset will be evaluated for quality issues, including missing values, outliers, inconsistent entries, and irrelevant fields. Variables that are unlikely to contribute to meaningful segmentation or analysis will be documented and excluded if necessary. Data quality assessment will involve summaries, visualizations of distributions, and identification of anomalies. Selected fields will then be cleaned, standardized, and transformed to ensure they are analysis-ready.
- **Ingest and Store:** All curated datasets, scripts, and metadata will be organized into a structured repository. This repository will include a GitHub folder containing Jupyter notebooks, Python scripts for data cleaning and synthetic data generation, cleaned datasets, and a comprehensive README file detailing workflow, methodology, and metadata information ensures reproducibility, transparency, and facilitates future reuse or adaptation of the dataset for related projects. Metadata documentation will follow standard conventions, including variable definitions, data types, sources, and known limitations.
- **Access, Use, Reuse:** By publishing the repository, the curated dataset, along with workflow scripts and metadata, will be made accessible to the wider community. Detailed documentation will enhance discoverability and encourage reuse, while ethical considerations around anonymization, demographic bias, and fairness will be explicitly

addressed. The repository will also include steps to reproduce, generation of synthetic data.

- **Transform:** Once the dataset is clean and curated, exploratory data analysis will be performed to uncover patterns in demographics, purchasing behavior, and campaign responses. Visualization techniques such as graphs, histograms, and heatmaps etc will help identify trends, correlations, and outliers. Derived fields will be created to support segmentation, such as aggregated spending categories, customer response rates to marketing campaigns. Might try some clustering options to identify distinct customer groups. These insights can inform marketing strategy, product recommendations, and customer engagement initiatives.

Data Acquisition

The initial dataset will be sourced from Kaggle, specifically the “Customer Personality Analysis” dataset. To enhance analytical value and reduce the risk of biases due to limited sample size, a synthetic dataset will be generated using Python, expanding the original 2,240 records to approximately 10,000 customers. The synthetic dataset will include demographic variables, purchasing data, campaign response indicators, and other behavioral attributes. Missing values and outliers will be introduced to simulate real-world data characteristics, providing a realistic testbed for cleaning, preprocessing, and analysis.

Team and Timeline

This is an individual project. The project will be structured in three phases with defined milestones and deliverables:

- **Phase 1 – Project Plan:** A written plan outlining objectives, methodology, expected outcomes, and tools to be used.

- **Phase 2 – Progress Report:** Submission of synthetic dataset, data cleaning scripts, interim datasets, workflow scripts, and metadata documentation.
- **Phase 3 – Final Submission:** Complete workflow, fully curated dataset, exploratory analysis results, customer segmentation outputs, metadata, and final report.

Constraints and Mitigations

- **Sample Size:** The original dataset (~2,200 customers) may limit generalizability. Mitigation involves generating a synthetic dataset of 10,000 customers to enable more robust statistical analysis.

  Data Quality: Some variables, particularly income, have missing values, and spending variables exhibit heavy skew. Synthetic data creation and cleaning strategies will address these issues while retaining realistic variability.
- **Ethics:** Although the data is anonymized, demographic-based segmentation may introduce potential bias. Ethical considerations and careful documentation will ensure responsible use of the data.
- **Scope:** The project will focus on data curation, metadata documentation, reproducibility, and exploratory analysis rather than building production-grade predictive models.

References

- Imakash3011. (2020). Customer Personality Analysis. Kaggle https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis.
- ChatGPT for guidance on generating realistic value ranges for synthetic dataset creation.