**Analyzing Retail Sales and Customer Behavior**

Deeksha Chawla

Bellevue University

DSC 540 Data Preparation

Prof. Catherine Williams

April 6th, 2024

**Project Subject Area:**

Analyzing Retail Sales and Customer Behavior

**Data Sources:**

1. **Flat File (CSV Dataset from Kaggle):**

Dataset: https://www.kaggle.com/datasets/tigboatnc/amazon-all-categories-best-sellers-reviews/data

Description: This dataset contains transactional data from an online retail store, including product name, category, category rank, cost.

Link: /kaggle/input/amazon-all-categories-best-sellers-reviews/abs080922 - clear.csv

1. **Website (Product Information):**

Website: Amazon Best Sellers

Description: Amazon Best Sellers lists top-selling products across various categories, providing valuable product information like rankings, categories, and customer reviews.

Link: <https://www.amazon.com/Best-Sellers/zgbs>

1. **API (Weather Data API):**

API: OpenWeatherMap API

Description: The OpenWeatherMap API provides weather data based on geographical coordinates, which can be used to analyze the impact of weather on retail sales.

Link: <https://openweathermap.org/api>

**Relationships and Integration:**

* **Matching Products (Dataset and Website):**
  + Will utilize product descriptions or product names from the Online Retail dataset to match with corresponding products listed on the Amazon Best Sellers website. This connection will allow you to enrich transactional data with detailed product information and popularity metrics.
* **Enhancing Customer Transactions with Weather Data (API):**
  + Will use customer locations (e.g., postal codes) from the transaction dataset to query the OpenWeatherMap API. This will provide weather conditions (e.g., temperature, precipitation) at the time and location of each transaction, enabling weather-based analysis of sales trends.

**Implementation Steps:**

1. Data Collection:
   1. Download the Online Retail dataset from Kaggle.
   2. Explore Amazon Best Sellers website to understand product categories and rankings.
2. Data Preprocessing:
   1. Clean and preprocess the transaction dataset (CSV) by handling missing values, standardizing formats, and ensuring data consistency.
3. Web Scraping (Product Information):
   1. Use web scraping techniques (e.g., Beautiful Soup in Python) to extract product details (e.g., categories, rankings, reviews) from relevant Amazon Best Sellers pages.
4. Matching Products and Integration:
   1. Match products from the transaction dataset with scraped product information based on common attributes (e.g., product descriptions, IDs).
   2. Integrate the matched product details (e.g., categories, rankings) into the transaction dataset for comprehensive analysis.
5. API Integration (Weather Data):
   1. Develop scripts to query the OpenWeatherMap API using customer locations (e.g., postal codes) from the transaction dataset.
   2. Retrieve weather data (e.g., temperature, precipitation) for each transaction location and timestamp.
6. Analysis and Visualization:
   1. Conduct exploratory data analysis (EDA) to uncover insights, such as the relationship between weather conditions and sales patterns, popular product categories, and customer preferences.
7. Modeling and Prediction:
   1. Build predictive models (e.g., regression, time series forecasting) to predict future sales based on factors including weather conditions, product attributes, and historical transaction data.

**Concerns/Challenges:**

1. Data Quality and Consistency:
   1. Incomplete or Missing Data: The datasets obtained from different sources may have missing values or incomplete records, which can affect the accuracy of your analysis.
   2. Inconsistent Data Formats: Different datasets may use varying formats for dates, product IDs, or other important fields, requiring careful standardization and preprocessing.
2. Data Integration:
   1. Matching and Merging Datasets: Integrating data from multiple sources (e.g., CSV file, website, API) can be challenging, especially when identifying common keys or attributes for merging datasets.
   2. Handling Large Volumes of Data: Depending on the size of the datasets and web scraping results, managing, and processing large volumes of data efficiently can be demanding.
3. Web Scraping and API Limitations:
   1. Rate Limiting: Some websites and APIs impose rate limits or restrictions on data retrieval, which may require implementing delays or batch processing to avoid being blocked.
   2. Dynamic Web Content: Websites may use dynamic content or JavaScript frameworks that complicate web scraping, requiring advanced techniques or tools.

**Ethical Considerations:**

* Data Privacy: Anonymize and handle customer data securely to protect individual privacy.
* Transparency: Clearly communicate the use of data collected from websites and APIs and ensure compliance with privacy regulations.
* Bias Mitigation: Address potential biases in data and modeling techniques to ensure fair and equitable analysis and predictions.