MGMTMSA405 Final Project Proposal

Group Members:

Name	Uid	GitHub Username
Yuqi Gu	506539826	77yuki
Jiayi Chen	406538813	JiayiChen123456
Wenyan (Kyle) Zhang	006539800	kylezzz25
Man (Sophia) Mei	106539885	Sophiameiman
Steven Chen	806547491	Steven-Chen-53

Dataset and Total Size:

Link	Size	Description	Source
Google Drive Link	1.87 GB	This dataset aggregates e-commerce activity logs from 2020 Jan to Apr. Each record includes information about the event, product, and user to support comprehensive market behavior analysis. Subject to size down.	**Source**
https://www.kaggle.com/datasets/arashnic/covid19-case-surveillance-public-use-dataset	2.67GB	This dataset includes individual-level data reported to U.S. states and autonomous reporting entities, including New York City and the District of Columbia (D.C.), as well as U.S. territories and states, covering time range from Dec. 2019 to Dec. 2020.	Same as link

Part I: Analysis Overview

This project explores the relationship between eCommerce behavior and the COVID-19 pandemic by analyzing consumer transaction data and public health records from January to April 2020. The analysis tracks short-term fluctuations in online shopping activity alongside pandemic trends.

1.1 Consumer Behavior Analysis

The study examines variations in online shopping patterns, including event types (views, cart additions/removals, purchases), product categories, brands, and pricing. Additionally, it investigates cart abandonment trends and uncertainty-driven purchasing behavior.

1.2 COVID-19 Trends & Correlation Analysis

By overlaying pandemic trends (case surges, hospitalizations, mortality) with eCommerce activity, the study aims to assess consumer response during lockdowns and demographic-based variations in shopping behavior.

1.3 Predictive Insights

Machine learning and time-series forecasting will help predict shopping trends during crises, aiding businesses in inventory management, pricing, and marketing strategies.

Part II: Dataset Value and Interactive Dashboard

An interactive dashboard will visualize online shopping trends alongside pandemic data.

2.1 Real-Time Consumer Trends

Users can explore shopping activity trends, filtering by event type, product category, and brand.

2.2 Pandemic Impact on Commerce

COVID-19 case trends will be overlaid with eCommerce activity, allowing analysis of demographic-driven consumer responses.

2.3 Customizable Analysis & Forecasting

Users can adjust timeframes, demographic filters, and models to explore crisis-driven consumer behavior shifts.

Part III: Spark Processing Framework

To efficiently process large-scale eCommerce and COVID-19 datasets, we use Apache Spark because of its in-memory processing and parallel computation for fast data transformation, aggregation, and analysis.

3.1 Data Ingestion

- Load raw eCommerce data (Amazon S3) and COVID-19 data (EC2).
- Use Spark's DataFrame API for structured CSV data handling.
- Validate schema and partition data by timestamps for efficient time-series analysis.

3.2 Data Cleaning & Preprocessing

- Handle missing and duplicate values; normalize timestamps.
- Remove outliers and inconsistencies (e.g., negative purchase amounts).

3.3 Data Transformation & Feature Engineering

- **eCommerce Data:** Categorize transactions, aggregate trends over time, segment purchases by product and price, and compute cart abandonment rates.
- **COVID-19 Data:** Summarize daily cases, hospitalizations, and deaths; segment by demographics; apply smoothing techniques for trend analysis.

3.4 Data Integration & Correlation Analysis

- Join datasets on timestamps to align shopping activity with pandemic trends.
- Compute correlation metrics and assess lagged effects of COVID-19 events on consumer behavior.

3.5 Optimized Storage & Scalability

- Store processed data in Parquet format for efficient querying via DuckDB.
- Use distributed computing on EC2 nodes, Spark's lazy evaluation, and caching to optimize performance.

By implementing this Spark pipeline, we ensure efficient data processing, correlation analysis, and interactive insights in our project dashboard.

Part IV: ETL Pipeline Overview

A scalable ETL pipeline (EC2, Spark, DuckDB) will process, integrate, and store datasets for efficient analysis and visualization. This streamlined ETL and analysis approach ensures efficient processing and insightful correlations between consumer behavior and pandemic events.

4.1 Data Extraction

- Large eCommerce data stored in Amazon S3; COVID-19 data on EC2.
- Apache Spark ingests and processes high-volume datasets.

4.2 Data Transformation

- Cleaning, categorizing event types, and aggregating COVID-19 case records.
- Joining datasets on date fields for trend analysis and correlation studies.

4.3 Data Loading

- Processed data stored in Parquet format and loaded into DuckDB for fast querying.

4.4 Data Serving & Visualization

- Interactive Tableau dashboard visualizing consumer behavior trends alongside COVID-19 case patterns.
- Predictive modeling for businesses and policymakers to anticipate future trends.