Executive Summary

DSA 2040A Group Project – Olist E-Commerce Data Mining

Project Objective

This project explores the **Olist e-commerce dataset** to uncover actionable insights across sales, customers, deliveries, and products. Techniques applied include ETL, EDA, feature engineering, and predictive modeling, aimed at improving business decision-making.

Team Structure & Roles

Member	Key Contributions
Hana Gashhaw	Led data cleaning, ETL, dashboards, and co-led modeling
Tizzah Nzioka	Co-led EDA and contributed to feature engineering
Ted Koiri	Co-led EDA and supported modeling phase
Selimah Tzindori	Built dashboards and co-led clustering & modeling
Levvin Ekxam	Managed documentation and visual storytelling
Angela Fungu	Assisted with documentation, presentation, and editing

---

Dataset Overview

- Public Olist dataset from Brazil
- ~1.7M records across 9 CSV files
- Focused on most recent 3 months of activity

ETL & Data Preparation

- Cleaned missing values and standardized column formats
- Merged tables and filtered recent data
- Handled outliers in pricing and freight
- Created structured data ready for analysis

Exploratory Data Analysis

- **Payment**: 73.5% used credit cards
- **Delivery**: 91% on-time delivery rate
- **Volume**: ~10K orders and revenue of 1.64M BRL

Feature Engineering

- Created new variables: delivery_delay, price_per_gram, and is_delayed
- Added time-based features like weekday, month, purchase_hour
- Applied transformations to normalize skewed data

Modeling Insights

- **Clustering**: Identified 3 distinct customer segments
- **Regression**: Modeled delivery times using logistics data
- **Classification**: Predicted review scores based on order details

---

Key Visual Insights

- Boxplots and scatterplots exposed price and freight patterns
- Heatmaps showed strong correlation between financial metrics
- Geo maps displayed regional product and sales distribution

Business Implications

- **Logistics**: Optimize shipping in delayed regions
- **Marketing**: Personalize campaigns based on customer segments
- **Inventory**: Focus stock on high-demand areas
- **Customer Retention**: Engage one-time buyers with offers
- **Product Review**: Monitor categories with lower satisfaction