## Executive Summary: Olist E-Commerce Data Mining Project

### Objective

This project showcases a data analytics pipeline for Brazil's Olist e-commerce platform, covering ETL, EDA, feature engineering, and predictive modeling to derive actionable insights from three months of sales data.

### Team Roles

The project team was structured with Hana Gashhaw leading ETL, Tizzah Nzioka and Ted Koiri heading exploratory data analysis, Selimah Tzindori and Hana Gashhaw managing data mining, Hana Gashhaw and Selimah Tzindori developing the dashboard, and Levvin Ekxam and Angela Fungu handling documentation.

### Dataset

The Olist dataset comprises nine interlinked CSV files covering orders, payments, revenue, products, customers, and sellers, linked by 'customer_id' and 'product_id'. It includes 1.7 million transactions across 44 columns over three months.

### Data Preprocessing

The ETL phase involved extracting and cleaning data, handling null values, standardizing formats, capping outliers, and imputing missing values using median for numerical data and timestamps for dates. All data was filtered to include only the last three months.

### Exploratory Data Analysis

Exploratory data analysis revealed that credit cards were the dominant payment method, most orders were delivered on time with delays negatively impacting reviews, and São Paulo served as the central hub for both buyers and sellers.

### Feature Engineering

Feature engineering created delivery delay metrics, log-transformed price, freight, and payment values, price-per-gram ratios, and time-based fields such as month, hour, weekday, and an is_delayed flag to enhance modeling.

## Data Mining & Modeling

K-Means clustering segmented customers into high-frequency buyers, big spenders, and low-price frequent buyers. Regression and classification models, using standardized and one-hot encoded data, predicted delivery delays and customer review scores.

### Key Visual Insights

Visualizations showed that boxplots highlighted price and shipping cost variance by product category, scatter plots indicated a strong correlation between price and freight value, heatmaps revealed multicollinearity in financial features, and regional plots uncovered pricing differences across Brazilian states.

### Business Implications

To improve operations, late deliveries should be monitored, especially in high-delay regions, and shipping should be optimized to reduce freight costs. Products with low review scores need investigation, premium categories should be prioritized for high-margin marketing, and inventory and pricing strategies should be tailored based on regional trends.