Executive Summary

Project Title: Data Mining Analysis of Brazilian E-commerce Platform (Olist Dataset) Course: Data Warehouse (DSA 2040A) Group: Group 8

---

1. Project Overview

This project analyzes the Olist E-commerce dataset, a Brazilian marketplace that connects small and medium-sized retailers to customers across the country. The objective is to explore customer behavior, delivery logistics, and product trends using a complete data mining pipeline that includes ETL, EDA, modeling, and visualization.

Key questions addressed:

- *What product categories are most profitable?*
- *How do discounts affect sales and profit?*
- *Are there regional or order-priority trends?*
- *Can we segment customers or transactions using clustering?*

---

2. ETL Summary

- *Loading & Inspection: Raw CSV files from Olist are ingested and previewed to understand structure and schema.*
- *Cleaning: Dates are parsed into datetime objects, missing values are treated, and irrelevant fields dropped.*
- *Feature Engineering: New columns like `profit`, `profit margin`, and `delivery_delay` are derived for better insights.*
- *Saving: Cleaned datasets are stored in structured formats (CSV/Parquet) for efficient querying and visualization.*

---

3. Exploratory Data Analysis (EDA)

- *Sales Trends: Time series plots show monthly revenue growth and peak sales months.*
- *Profitability: Electronics and furniture consistently rank as high-profit categories.*
- *Discounts: Analysis reveals non-linear relationships between discount levels and profitability.*
- *Geographical Patterns: Regional heatmaps reveal top-performing states by revenue and delivery speed.*
- *Order Priority: Boxplots show how express orders influence profit and delivery time.*

---

4. Data Mining Techniques

- *Clustering (K-Means): Customers and transactions were grouped into segments based on spending, frequency, and location.*
- *Classification: Attempted to predict late delivery based on product type, seller location, and shipping method.*
- *Association Rules: Identified frequent itemsets using Apriori algorithm to suggest complementary products.*

- *NLP Sentiment Analysis: Cleaned and tokenized review comments, applied VADER for polarity scoring, visualized positive/neutral/negative sentiment distributions.*
- *Text Classification: Trained a logistic regression model using TF-IDF features to classify sentiment labels.*

---

5. Key Findings

- *Top 3 product categories by profit: Computers, Office Supplies, and Furniture.*
- *Discounts beyond 20% often reduced profit margins significantly.*
- *Southeast region had highest number of orders but also highest delays.*
- *Three distinct customer clusters were discovered: high-spenders, regional buyers, and bargain-seekers.*
- *Review sentiment was predominantly positive, and correlated weakly with delivery performance.*

---

6. Recommendations

- *Focus marketing on high-margin categories with stable demand.*
- *Limit discounts to below 20% to preserve profit.*
- *Improve delivery logistics in high-traffic regions (e.g., São Paulo).*
- *Personalize offers based on customer cluster segmentation.*
- *Integrate NLP-driven sentiment monitoring into feedback analysis loops.*

---

7. Tools & Technologies

- *Programming: Python (Pandas, NumPy, Seaborn, Plotly, Scikit-learn, NLTK, VADER, WordCloud)*
- *Environment: Jupyter Notebook*
- *Visualization: Power BI*
- *Version Control: Git & GitHub*

---

8. Conclusion

This project effectively used the Olist dataset to explore key trends in online retail. From raw data cleaning to customer segmentation and sentiment classification, the project highlights how data mining techniques can inform strategy, optimize logistics, and drive business insights.

Appendices

- *Confusion Matrix from Classification*
- *Cluster Profiles from K-Means*
- *Profit Analysis Tables by Region and Category*