**Modification Log:**

**—-----------—-----------—-----------—-----------—-----------—-----------—-----**

**Date Description Author ATBT**

**—-----------—-----------—-----------—-----------—-----------—-----------—-----**

**09.06.25 Initial Document Syed Yes**

**11.06.25 Added User stories Syed Yes**

*Added: User Story 1.4,* User Story 2.4

**Optimizing E-Commerce Sales and Marketing Strategies**

## **Project Overview**

This capstone project focuses on leveraging the **Brazilian E-Commerce Public Dataset by Olist**, which encompasses over 100,000 anonymized orders from 2016 to 2018 across multiple Brazilian marketplaces. The dataset is comprehensive, including transactional records, customer demographics, product details, payment information, and customer reviews, making it an ideal resource for in-depth sales, retail, and marketing analysis. Students will work with this large, multi-table dataset to clean, process, and analyze data, applying advanced data analytics techniques to uncover actionable insights. The project emphasizes handling big data, performing exploratory and predictive analyses, and delivering data-driven recommendations to optimize e-commerce operations, enhance customer retention, and boost sales through targeted marketing strategies.

The project requires students to integrate multiple data tables, address real-world business questions, and present findings through visualizations and strategic recommendations. By working with this real-world e-commerce dataset, students will gain hands-on experience in managing large datasets, applying statistical and machine learning techniques, and translating insights into business value, preparing them for data analytics roles in the retail and e-commerce sectors.

## **Objective**

The primary objective is to analyze the Olist dataset to:

* Identify customer purchasing patterns and segment customers for targeted marketing.
* Optimize marketing campaigns by analyzing promotion effectiveness and customer behavior.
* Improve product offerings by identifying high-demand or underperforming categories.
* Enhance customer retention by analyzing factors like delivery performance and review sentiment.
* Provide actionable recommendations to increase sales, streamline operations, and improve customer satisfaction in an e-commerce setting.

**Phase Manner**

* Phase / Sprint 1: Project Setup and Data Acquisition
* Phase / Sprint 2: Data Processing and Cleaning
* Phase / Sprint 3: Exploratory Data Analysis (EDA)
* Phase / Sprint 4: Advanced Analytics
* Phase / Sprint 5: Marketing and Business Recommendations
* Phase / Sprint 6: Visualization and Reporting
* Phase / Sprint 7: Project Review and Iteration

**Phase 1/Sprint 1: Project Setup and Data Acquisition**

Objective: Set up the project environment, acquire the dataset, and understand its structure.

User Story 1.1: As a data analyst, I want to download the Olist dataset from Kaggle, so that I can access the raw data for analysis.

- Acceptance Criteria:

- Dataset is downloaded as a ZIP file containing all nine CSV files.

- Files are unzipped and stored in an accessible directory.

**Status: Complete**

User Story 1.2: As a data analyst, I want to review the dataset schema and documentation, so that I understand the relationships between tables and their contents.

- Acceptance Criteria:

- Schema for all tables (e.g., orders, customers, products) is documented.

- Relationships (e.g., `order\_id` links) are mapped in a diagram or notes.

**Status: Complete**

User Story 1.3: As a data analyst, I want to set up a Python environment with necessary libraries, so that I can process and analyze the dataset efficiently.

- Acceptance Criteria:

- Python environment includes Pandas, Dask, Scikit-learn, Matplotlib, Seaborn, and Plotly.

- PostgreSQL or SQLite is installed for database operations.

- Environment is tested with a sample data load.

**Status: NA  
*Note:*** *To revisit after the python class completion to CRUD Operations in SQL*

User Story 1.4: As a project supervisor, I want the team to create a project plan with milestones, so that the project stays on track and meets deadlines.

- Acceptance Criteria:

- Plan includes phases, tasks, and deadlines (e.g., data cleaning by week 2).

- Plan is shared with the team and approved by the supervisor.

**Status: Team to decide**

User Story 1.4: As a data analyst, I want to set up a version control system for code and data artifacts, so that the team can collaborate and track changes effectively.

* Acceptance Criteria:
  + A Git repository is created and shared with the team.
  + Initial dataset and code files are committed.
  + Branching strategy is documented (e.g., main branch, feature branches).

**Status: Complete**

**Phase 2: Data Preprocessing and Cleaning**

Objective: Load, clean, and integrate the dataset to prepare it for analysis.

User Story 2.1: As a data analyst, I want to load all CSV files into a relational database or Pandas DataFrames, so that I can query and manipulate the data easily.

- Acceptance Criteria:

- All nine CSV files are loaded successfully.

- Data types (e.g., datetime for timestamps, float for prices) are correctly assigned.

User Story 2.2: As a data analyst, I want to clean the dataset by handling missing values and inconsistencies, so that the data is reliable for analysis.

- Acceptance Criteria:

- Missing values (e.g., `review\_comment\_message`) are identified and handled (e.g., imputed or flagged).

- Inconsistencies (e.g., negative prices, invalid timestamps) are corrected or removed.

- Cleaning steps are documented in a notebook.

User Story 2.3: As a data analyst, I want to join relevant tables (e.g., orders, order\_items, customers), so that I can perform comprehensive analyses.

- Acceptance Criteria:

- Tables are joined using keys (e.g., `order\_id`, `customer\_id`).

- A merged dataset is created with relevant columns for analysis.

- Joins are validated (e.g., no missing orders due to incorrect joins).

User Story 2.4: As a data analyst, I want to optimize table join operations, so that data integration is performed efficiently.

* Acceptance Criteria:
* Join operations are tested for performance (e.g., execution time) on a subset of data.
* Indexing or partitioning is applied to large tables (e.g., order\_items) if using a database.
* Optimized joins are documented with performance metrics.

User Story 2.5: As a data analyst, I want to use Dask or Spark for processing the large geolocation dataset, so that I can handle its ~1 million rows efficiently.   
***Label:*** *No knowledge about Dask / Spark*

- Acceptance Criteria:

- Geolocation dataset is processed without memory issues.

- Aggregated geolocation data (e.g., by zip code) is integrated with customer data.

**Phase 3: Exploratory Data Analysis (EDA)**

Objective: Explore the dataset to identify trends, patterns, and initial insights.

User Story 3.1: As a data analyst, I want to analyze sales trends by time, category, and region, so that I can identify high-performing periods and products.

- Acceptance Criteria:

- Sales are aggregated by month, product category, and customer state.

- Visualizations (e.g., line charts for sales over time, bar charts for categories) are created.

User Story 3.2: As a data analyst, I want to explore customer behavior metrics like average order value and repeat purchases, so that I can understand purchasing patterns.

- Acceptance Criteria:

- Metrics (e.g., average order value, purchase frequency) are calculated using `order\_items` and `customers` data.

- Visualizations (e.g., histograms, box plots) show distributions.

User Story 3.3: As a data analyst, I want to investigate the relationship between review scores and delivery performance, so that I can identify factors affecting customer satisfaction.

- Acceptance Criteria:

- Delivery time is calculated (`order\_delivered\_customer\_date` - `order\_purchase\_timestamp`).

- Correlation between delivery time and `review\_score` is analyzed.

- Scatter plots or heatmaps visualize the relationship.

User Story 3.4: As an e-commerce business manager, I want a summary of EDA findings, so that I can understand key trends and areas for improvement.

- Acceptance Criteria:

- EDA report includes visualizations and written insights.

- Report is shared in a Jupyter Notebook or PDF format.

**Phase 4: Advanced Analytics**

Objective: Apply advanced techniques to derive deeper insights and predictive models.

User Story 4.1: As a data analyst, I want to segment customers using RFM analysis or K-means clustering, so that I can identify high-value customer groups for targeted marketing.

- Acceptance Criteria:

- RFM metrics (Recency, Frequency, Monetary) are calculated using `orders` and `order\_payments`.

- K-means clustering is applied to segment customers.

- Segments are visualized (e.g., scatter plot of Recency vs. Monetary).

User Story 4.2: As a data analyst, I want to perform market basket analysis, so that I can identify products frequently bought together for cross-selling opportunities.

- Acceptance Criteria:

- Apriori algorithm is applied to `order\_items` data.

- Association rules (e.g., {Product A} → {Product B}) are generated with support, confidence, and lift.

- Top rules are summarized in a table or visualization.

User Story 4.3: As a data analyst, I want to build a time-series model to forecast future sales, so that I can help the business plan inventory and promotions.

- Acceptance Criteria:

- Sales data is aggregated by month using `order\_purchase\_timestamp`.

- A model (e.g., ARIMA or Prophet) is trained and validated.

- Forecast for the next 6 months is visualized with confidence intervals.

User Story 4.4: As a data analyst, I want to analyze review sentiment using NLP, so that I can understand customer feedback and its impact on sales.

- Acceptance Criteria:

- `review\_comment\_message` is processed using NLP libraries (e.g., NLTK, TextBlob).

- Sentiment scores (positive, neutral, negative) are assigned.

- Correlation between sentiment and `review\_score` or sales is analyzed.

**Phase 5: Marketing and Business Recommendations**

Objective: Translate insights into actionable strategies for sales and marketing.

User Story 5.1: As an e-commerce business manager, I want recommendations for targeted marketing campaigns, so that I can increase sales among high-value customers.

- Acceptance Criteria:

- High-value customer segments are identified from RFM or clustering.

- Specific campaign strategies (e.g., discounts, email marketing) are proposed.

- Expected impact (e.g., sales increase) is estimated.

User Story 5.2: As an e-commerce business manager, I want suggestions for product bundling based on market basket analysis, so that I can improve cross-selling.

- Acceptance Criteria:

- Top product associations from market basket analysis are listed.

- Bundling recommendations (e.g., “Bundle Product A with Product B”) are provided.

User Story 5.3: As an e-commerce business manager, I want insights on improving customer retention, so that I can reduce churn and increase loyalty.

- Acceptance Criteria:

- Factors affecting retention (e.g., delivery time, review scores) are identified.

- Strategies (e.g., faster shipping, loyalty programs) are recommended.

**Phase 6: Visualization and Reporting**

Objective: Create visualizations and deliver a comprehensive report.

User Story 6.1: As a data analyst, I want to create an interactive dashboard, so that stakeholders can explore sales, customer, and marketing insights.

- Acceptance Criteria:

- Dashboard is built using Tableau, Power BI, or Plotly Dash.

- Includes visuals for sales trends, customer segments, and product performance.

- Dashboard is user-friendly and interactive (e.g., filters for time or category).

User Story 6.2: As a project supervisor, I want a final report summarizing all findings and recommendations, so that I can evaluate the team’s work.

- Acceptance Criteria:

- Report includes EDA, advanced analytics, and business recommendations.

- Visualizations are embedded with clear explanations.

- Report is submitted as a PDF or slide deck.

User Story 6.3: As an e-commerce business manager, I want a presentation of key insights and strategies, so that I can implement them in my business.

- Acceptance Criteria:

- Presentation is concise (10-15 slides).

- Includes visuals, insights, and actionable recommendations.

- Delivered in a professional format (e.g., PowerPoint, Google Slides).

**Phase 7: Project Review and Iteration**

Objective: Review the project, gather feedback, and refine deliverables.

User Story 7.1: As a project supervisor, I want to provide feedback on the team’s deliverables, so that they can improve their work before final submission.

- Acceptance Criteria:

- Feedback is provided on EDA, models, dashboard, and report.

- Team incorporates feedback into revised deliverables.

User Story 7.2: As a data analyst, I want to test the predictive models with new data splits, so that I can ensure their robustness.

- Acceptance Criteria:

- Models (e.g., sales forecast, clustering) are re-evaluated with a holdout dataset.

- Performance metrics (e.g., RMSE for forecasting) are reported.

User Story 7.3: As a data analyst, I want to document the project codebase and methodology, so that future teams can replicate or extend the work.

- Acceptance Criteria:

- Code is organized in a repository (e.g., GitHub) with README.

- Methodology (e.g., data cleaning steps, model parameters) is documented.

### **Dataset Schema** The dataset includes nine CSV files with the following schema (simplified for clarity):

1. **orders\_dataset.csv** (Order-level data)  
   Columns:

* order\_id: Unique order identifier (string)
* customer\_id: Unique customer identifier for the order (string)
* order\_status: Status (e.g., delivered, shipped) (string)
* order\_purchase\_timestamp: Purchase date/time (datetime)
* order\_delivered\_customer\_date: Delivery date/time (datetime)
* order\_estimated\_delivery\_date: Estimated delivery date (datetime)

1. **order\_items\_dataset.csv** (Order item details)  
   Columns:

* order\_id: Links to orders table (string)
* order\_item\_id: Item sequence in the order (integer)
* product\_id: Unique product identifier (string)
* seller\_id: Unique seller identifier (string)
* price: Item price (float)
* freight\_value: Shipping cost (float)

1. **customers\_dataset.csv** (Customer information)  
   Columns:

* customer\_id: Unique customer identifier (string)
* customer\_unique\_id: Unique customer identifier across orders (string)
* customer\_zip\_code\_prefix: First 5 digits of zip code (integer)
* customer\_city: City (string)
* customer\_state: State (string)

1. **products\_dataset.csv** (Product details)  
   Columns:

* product\_id: Unique product identifier (string)
* product\_category\_name: Product category (string)
* product\_weight\_g: Weight in grams (float)
* product\_length\_cm: Length in cm (float)

1. **order\_payments\_dataset.csv** (Payment details)  
   Columns:

* order\_id: Links to orders table (string)
* payment\_type: Payment method (e.g., credit\_card, boleto) (string)
* payment\_value: Payment amount (float)

1. **order\_reviews\_dataset.csv** (Customer reviews)  
   Columns:

* review\_id: Unique review identifier (string)
* order\_id: Links to orders table (string)
* review\_score: Rating (1-5) (integer)
* review\_comment\_message: Review text (string)

1. **sellers\_dataset.csv** (Seller information)  
   Columns:

* seller\_id: Unique seller identifier (string)
* seller\_zip\_code\_prefix: First 5 digits of zip code (integer)
* seller\_city: City (string)
* seller\_state: State (string)

1. **product\_category\_name\_translation.csv** (Category translations)  
   Columns:

* product\_category\_name: Category in Portuguese (string)
* product\_category\_name\_english: Category in English (string)

1. **geolocation\_dataset.csv** (Geolocation data)  
   Columns:

* geolocation\_zip\_code\_prefix: First 5 digits of zip code (integer)
* geolocation\_lat: Latitude (float)
* geolocation\_lng: Longitude (float)
* geolocation\_city: City (string)
* geolocation\_state: State (string)