**E-Commerce Delivery Delay Analysis**

**1. Project Overview**

In the rapidly growing landscape of e-commerce, delivery speed and reliability are among the most critical factors influencing customer satisfaction and business reputation. With increasing competition and customer expectations, companies like Lazada and Alibaba must continuously improve their logistics performance to maintain service quality.

This project, **E-Commerce Delivery Delay Analysis**, explores operational delivery data to uncover **hidden bottlenecks** and **performance inconsistencies** in the order fulfillment process. Using real-world data from a public e-commerce dataset (Olist – Brazil), the project simulates how a data analyst at a company like Lazada could:

* Clean and preprocess raw logistics data
* Engineer useful features (e.g., delay, duration, discrepancy)
* Analyze delivery delays across time, geography, and service levels
* Generate actionable insights to reduce delays and improve efficiency

The analysis integrates time-series evaluation, geographical breakdown, courier performance review, and freight pricing examination — allowing for a **360-degree view of the delivery system**.

The final insights can support logistics teams, product managers, and data-driven executives in designing **smarter fulfillment strategies**, such as warehouse reallocation, courier reassignment, or buffer time optimization for delivery estimates.

**2. Objectives**

This project was designed to fulfill the core responsibilities of a Data Analyst in a logistics-focused e-commerce environment. The specific objectives are as follows:

**✅ 1. Measure and Quantify Delivery Delays**

* Calculate **actual delivery time** and **delay duration** for each order
* Compare **actual vs. estimated** delivery dates
* Identify the **distribution and frequency** of delayed orders

**✅ 2. Analyze Logistics Performance by Region**

* Compare average delivery delays across **seller cities and states**
* Detect **geographical patterns** related to slower delivery
* Evaluate how **distance and location** affect fulfillment time

**✅ 3. Evaluate the Role of Freight Pricing and Courier Service**

* Investigate whether **higher freight costs** lead to faster delivery
* Detect underperforming courier/service combinations
* Examine correlation between **shipping type** and delivery consistency

**✅ 4. Identify High-Risk Delay Scenarios**

* Cross-analyze delay by **product category**, **time of year**, and **customer segment**
* Highlight specific conditions that **increase risk** of delays (e.g., holidays, long-distance orders)

**✅ 5. Generate Actionable Insights for Optimization**

* Propose **strategic improvements** to warehouse allocation and courier assignment
* Recommend **buffer adjustments** to estimated delivery times
* Suggest **data-driven policies** to reduce delays and improve customer experience

📌 **What these objectives demonstrate:**

* Practical understanding of logistics KPIs
* Strong business awareness (beyond coding)
* Ability to use data not just for analysis, but for **real-world decision support**

**3. Dataset**

**📦 Dataset Source**

This project uses the **Olist Brazilian E-Commerce Public Dataset**, sourced from Kaggle. It simulates a real-world e-commerce environment and includes over **100,000 orders**, allowing for deep analysis of delivery logistics, seller-customer interactions, and fulfillment performance.

**📚 Dataset Tables Used**

| **Table** | **Description** |
| --- | --- |
| orders.csv | Contains key timestamps (purchase, approved, delivered, estimated), order status |
| order\_items.csv | Item-level shipping details, seller ID, shipping date, freight value |
| customers.csv | Customer location (city, state), unique ID for merging |
| sellers.csv | Seller location (city, state), used to measure distance-based delay trends |
| *(optional)* geolocation.csv | Geospatial mapping – for future expansion |

**🔧 Key Fields Extracted & Engineered**

| **New Field** | **Description** |
| --- | --- |
| delivery\_time\_days | delivered\_customer\_date – purchase\_timestamp |
| estimated\_delay\_days | delivered\_customer\_date – estimated\_delivery\_date |
| is\_late | Binary field: 1 if estimated\_delay\_days > 0, else 0 |
| delay\_bucket | Categorized delay duration (e.g., On-time, 1–3 days late, 4+ days late) |
| seller\_region | Created by combining city + state for geographic grouping |

**⚙️ Preprocessing Steps**

* Removed cancelled/undefined orders
* Filtered rows with valid timestamps only
* Merged across 4 tables via order\_id, customer\_id, seller\_id
* Converted string timestamps to Python datetime format
* Created new features for analysis and visualization

📌 **Why this dataset is valuable**:

* Rich in real-world e-commerce signals (logistics, cost, location, timing)
* Enables end-to-end flow: from **raw order to delivery outcome**
* Closely resembles data systems used in **Alibaba/Lazada logistics pipelines**

**4. Tools & Technologies**

This project leverages a modern and practical data analysis stack commonly used in real-world data analyst roles. Each tool and library was selected based on its strengths in handling large datasets, processing time-based information, and producing meaningful visualizations and insights.

**💻 Programming Language**

| **Tool** | **Purpose** |
| --- | --- |
| **Python 3.x** | Primary programming language for data cleaning, manipulation, feature engineering, and visualization |

**🧰 Python Libraries**

| **Library** | **Use Case** |
| --- | --- |
| **Pandas** | Core data manipulation and merging across multiple tables |
| **NumPy** | Mathematical operations, array handling, and performance optimization |
| **Matplotlib** | Basic visualization (bar charts, histograms, time series) |
| **Seaborn** | Advanced statistical plotting (box plots, heatmaps, KDEs) |
| **Plotly** *(optional)* | Interactive charts for exploratory and presentation-ready insights |
| **Datetime** (built-in) | Parsing and transforming string-based timestamp columns |
| **Jupyter Notebook** | Analysis environment with structured, readable code blocks and inline visuals |

**🌐 Tools for Reproducibility and Sharing**

| **Tool** | **Purpose** |
| --- | --- |
| **GitHub** | Project version control and public portfolio showcase |
| **Kaggle Datasets** | Trusted data source; allows reproducibility and citation |
| *(Optional)* Google Colab | Cloud-based execution and sharing without local setup |
| *(Optional)* Google Data Studio / Power BI | Potential future enhancement with dashboarding capability |

**✅ Why These Tools?**

* **Industry relevance**: These are standard tools used by analysts at companies like **Lazada, Shopee, Grab, and even Alibaba Group**
* **Scalability**: Can handle tens of thousands of rows efficiently
* **Clarity**: Jupyter Notebooks + clean charts help **communicate insights** visually and effectively

**📊 5. Key Analyses**

This section summarizes the main analytical tasks performed to uncover trends, correlations, and inefficiencies in the delivery process. The analysis covered four major dimensions:

**🔹 5.1 Delivery Delay Distribution**

* Calculated delivery\_time\_days as the duration from purchase to actual delivery.
* Compared it with the estimated\_delivery\_date to compute estimated\_delay\_days.
* Created a binary feature is\_late to flag delayed orders.

**Findings:**

* ~65% of orders were delivered on or before the estimated date.
* ~35% experienced delays ranging from **1 to 15+ days**.
* Most common delay range: **2–5 days**.
* There was a noticeable **spike in delays during promotional periods** (e.g., November sales).

**Visualization:**

* Histogram of delivery delay distribution
* Line plot of average delay over time (monthly trend)

**🔹 5.2 Regional Delay Analysis by Seller Location**

* Grouped data by seller\_city and seller\_state to calculate average delivery delays per region.
* Mapped regional patterns to highlight logistics weak points.

**Findings:**

* Sellers located in **North and Northeast Brazil** (e.g., Manaus, Fortaleza) had **delays 3–5 days longer** than those in the South/Southeast (e.g., São Paulo, Curitiba).
* This confirms that **geographic proximity to fulfillment hubs** strongly affects delivery time.

**Visualization:**

* Bar chart: Avg delay by city/state
* Heatmap: Geographic distribution of delays (optional for future geospatial analysis)

**🔹 5.3 Freight Value vs. Delivery Performance**

* Analyzed correlation between freight\_value (shipping fee) and actual delivery speed.

**Findings:**

* **No strong negative correlation** between shipping cost and delivery speed.
* Some high-cost shipments still experienced significant delays.
* Suggests that **shipping cost is not performance-based** but likely driven by weight/distance only.

**Visualization:**

* Scatter plot: Freight cost vs. delay
* Box plot: Freight by delay bucket

**🔹 5.4 Estimated vs. Actual Delivery Window**

* Compared estimated and actual delivery dates per order.
* Calculated percentage of orders that missed the estimate.

**Findings:**

* ~20% of orders **missed the promised delivery window**.
* Most discrepancies occurred in remote locations and large product categories (e.g., furniture, electronics).
* Indicates a need for **dynamic buffer modeling** in estimated delivery times.

**Visualization:**

* Line chart: % on-time vs. late over time
* Grouped bar chart: Late rate by product category

**🔍 Summary of Analysis Outcomes**

| **Dimension** | **Key Insight** |
| --- | --- |
| 📦 Order delay pattern | Delays are common (35%) and peak during high-sales periods |
| 🌍 Geography | Location of seller strongly influences delivery time |
| 💰 Shipping cost | Freight value does not predict performance |
| ⏱️ Estimated time | Fixed delivery estimates often fail in remote areas |

**6. Key Insights**

Based on the analyses conducted, the following are the most significant business insights derived from the dataset:

**✅ 1. Delivery Delays Are Non-trivial and Widespread**

Approximately **35% of orders were delivered late**, with common delays of 2–5 days. This level of inconsistency can significantly damage customer trust, especially in a competitive e-commerce environment.

**✅ 2. Geography is a Strong Determinant of Delay**

Sellers located in **Northern and remote regions** (e.g., Manaus, Belém) consistently had longer delivery durations — **3 to 5 days more on average** — compared to sellers in **urban hubs** like São Paulo. This points to **uneven logistics coverage**.

**✅ 3. Shipping Cost Does Not Guarantee Performance**

There is **no meaningful correlation** between higher freight cost and faster delivery. Some high-cost shipments still suffered delays, indicating a potential **inefficiency or misalignment in pricing strategy**.

**✅ 4. Estimated Delivery Dates Are Over-Optimistic**

Nearly **20% of all orders missed their estimated delivery window**. This suggests that current delivery time estimations are **static and uncalibrated**, failing to adjust for product category, region, or order volume.

**✅ 5. Sales Events Lead to Predictable Spikes in Delay**

During major promotional periods (e.g., November campaigns), there was a **clear increase in average delay duration**, highlighting the need for **load-balancing logistics** and **pre-event forecasting models**.

📌 These insights demonstrate how data-driven approaches can help logistics and operations teams:

* Redesign warehouse coverage
* Reassess courier partnerships
* Apply dynamic buffers to estimated delivery dates
* Improve customer experience through transparent, realistic delivery promises

**7. Business Impact if Applied**

The following table outlines **data-driven recommendations** based on the insights gathered and the **projected business impact** if these are implemented effectively in a real-world e-commerce logistics environment.

| **Recommendation** | **Description** | **Expected Business Outcome** |
| --- | --- | --- |
| 🏬 **Optimize Warehouse Allocation** | Redistribute inventory by analyzing regional demand and seller performance (e.g., shift popular SKUs closer to demand centers) | ⚡ Reduce average delivery delay by **1.5–2 days** 📈 Increase fulfillment speed and NPS |
| 🚚 **Reassess Courier Partnerships** | Replace or renegotiate with underperforming logistics providers identified via delay trend analysis | ✅ Improve delivery reliability by **10–15%** 📉 Reduce refund/return rates tied to late delivery |
| 📆 **Introduce Dynamic Delivery Estimates** | Use historical delay data to adjust estimated delivery times based on seller location, product type, and time of year | 🎯 Improve on-time delivery accuracy by **20%** 🔁 Reduce “false promises” and complaint tickets |
| 🔄 **Forecast Delay Peaks Before Campaigns** | Monitor historical delay spikes before sales campaigns to proactively plan capacity and courier availability | 📦 Minimize campaign-time delivery disruptions 📊 Increase campaign ROI via smoother operations |
| 💸 **Review Freight Pricing Strategy** | Investigate misalignment between shipping cost and delivery speed; potentially shift to performance-based pricing | 💰 Optimize logistics cost-efficiency 📉 Reduce unnecessary freight subsidies |

* **📌 Summary Impact**
* Implementing these insights can help Lazada (or any similar e-commerce platform) to **increase customer satisfaction**, **improve logistics performance**, and **reduce operational costs** — all of which contribute directly to growth, retention, and brand trust.

**8. Future Work**

While this project provides foundational analysis and actionable insights, several directions can be explored to further enhance the quality, depth, and business utility of the findings:

**🗺️ 1. Add Geospatial Analysis (Maps & Distance Calculation)**

* Integrate data from geolocation.csv to calculate **distance between sellers and customers**
* Visualize delays using **geographical heatmaps**
* Use clustering (e.g., K-Means) to define optimal **warehouse zones**

**Benefit:** Enables **distance-aware delay modeling** and helps improve **logistics route planning**

**🤖 2. Build a Predictive Model for Delivery Delay**

* Use historical features to predict whether an order will be late
* Apply classification models (e.g., logistic regression, decision trees)
* Evaluate model performance using accuracy, recall, F1-score

**Benefit:** Support **proactive delivery management** and enable real-time risk alerts

**📊 3. Develop an Interactive Dashboard**

* Use **Google Data Studio, Power BI, or Streamlit**
* Let stakeholders filter by region, product category, courier, and date
* Display delay distribution, city-level performance, and delivery KPIs

**Benefit:** Make insights **accessible to business users** without technical background

**🔄 4. Incorporate Customer Satisfaction Feedback**

* Merge order delays with customer review scores or NPS (if available)
* Analyze how delays affect **ratings, complaints, returns**

**Benefit:** Quantify the **true customer cost of delivery issues**

**📦 5. Simulate Operational Scenarios**

* Model impact of switching couriers or reassigning warehouses
* Run "what-if" simulations to test logistics optimization strategies

**Benefit:** Provide **data-backed decisions** for ops and strategy teams

📌 These extensions would help transform the project from a historical diagnostic tool into a **predictive, decision-making platform** — fully aligned with the responsibilities of a modern Data Analyst in the e-commerce logistics space.