Dataset Selection Document: Analysis of NYC Yellow Taxi Trip Data

1. Business Domain Chosen and Why

Business Domain: Transportation & Urban Logistics

Justification: This domain was selected for its direct relevance to Roado's operations in the logistics and transportation sector. Analyzing urban trip data from a large, mature market provides a strong parallel to the challenges and opportunities in route optimization, peak demand management, and service efficiency. The insights derived from this analysis can offer valuable, actionable intelligence for strategic planning.

2. Data Source(s)

- **Source:** NYC Taxi and Limousine Commission (TLC) Trip Record Data
- **Specific Dataset:** Yellow Taxi Trip Records for January 2025. Using a single recent month provides a dataset that is large enough for meaningful analysis (over 3 million records) while remaining manageable for a time-constrained project.
- **Link:** https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

3. Brief Description of Raw Data

The raw data is provided in a Parquet file format and contains detailed, anonymized records for each taxi trip. Key fields include:

- tpep_pickup_datetime / tpep_dropoff_datetime: Timestamps for the start and end of each trip.
- PULocationID / DOLocationID: Numeric IDs for the pickup and dropoff taxi zones.
- passenger_count: The number of passengers in the vehicle.
- trip_distance: The total distance of the trip in miles.
- payment_type: A categorical code for the payment method (e.g., 1=Credit Card, 2=Cash).
- fare_amount, tip_amount, tolls_amount, total_amount: Itemized financial details for the trip.

4. Business Questions to Answer

This analysis aims to answer the following core business questions:

1. **Demand & Pricing:** What are the temporal patterns of trip demand (hourly, daily), and how do they correlate with fare amounts?
2. **Geospatial Hotspots:** Which pickup zones are the *busiest* (highest volume) versus the most *profitable* (highest average fare)?
3. **Trip Efficiency:** What is the relationship between trip distance, duration (calculated from timestamps), and overall fare, particularly in high-congestion scenarios?
4. **Payment Behavior:** Is there a standard tipping behavior for customers who pay with credit cards?
5. **Airport Traffic:** How do airport trips (JFK, LaGuardia, Newark) compare to standard city trips in terms of fare and profitability?