# Delhivery Data Analysis and Feature Engineering


## Data Exploration & Quality

Comprehensive analysis of logistics data quality, identifying patterns and inconsistencies to ensure reliable insights.

## Feature Engineering

Transformation of raw logistics data into meaningful features, enabling better predictive modeling and decision-making.


## Operational Insights

Data-driven recommendations for improving forecasting accuracy and streamlining logistics operations.

By Mohammad Ahmed Khan

# Data Loading and Initial Exploration

The Delhivery dataset was loaded using pandas, and an initial exploration was conducted to understand its structure. The dataset consists of 144,867 rows and 24 columns. The **info()** method revealed the data types of each column, and the **isnull().sum()** method identified missing values in *source_name* and *destination_name* columns. Irrelevant columns such as *is_cutoff*, *cutoff_factor*, *cutoff_timestamp*, *factor*, and *segment_factor* were dropped to streamline the dataset.

# Missing Value Handling and Outlier Detection

Missing values in *source_name* and *destination_name* columns were removed, accounting for approximately 0.4% of the data. Outliers in numerical columns were detected using box plots. The numerical columns include *start_scan_to_end_scan*, *actual_distance_to_destination*, *actual_time*, *osrm_time*, *osrm_distance*, *segment_actual_time*, *segment_osrm_time*, and *segment_osrm_distance*. All numerical columns exhibited outliers, indicating extreme values that may affect model performance.

# Data Type Conversion and Descriptive Statistics

Time-related columns such as*trip_creation_time**od_start_time*, and *od_end_time* were converted to datetime objects for time series analysis. Descriptive statistics were computed using the **describe(include='all').T** method. The analysis revealed that FTL is the dominant route type and identified the most frequent source and destination centers.

- FTL is the primary mode of transportation.
- Most frequent source and destination center is "IND000000ACB"
- The trips took place approximately 3 weeks .


# Segment-wise Analysis and Feature Engineering

A new feature, *segmented_trip*, was created by concatenating *trip_uuid*, *source_name*, and *destination_name* for segment-wise analysis. Cumulative sums of *segment_actual_time*, *segment_osrm_time*, and *segment_osrm_distance* were calculated. This process aggregates the data to provide insights into the overall trip characteristics.

# Hypothesis Testing and Route Analysis

Paired t-tests were conducted to compare *actual_time* vs. *segment_actual_time_sum* and *actual_distance_to_destination* vs. *osrm_distance*. The statistical analysis was performed with a significance level of 0.05 (95% confidence interval). The t-tests
revealed significant differences ($p < 0.001$) in both comparisons, suggesting systematic deviations between predicted and actual route metrics.

For the time comparison, actual delivery times were consistently higher than the sum of segment times (mean difference = 45.3 minutes), indicating substantial cumulative delays. This difference was particularly pronounced in routes with multiple stops, suggesting that transition times between segments contribute significantly to overall delivery duration. The distance

analysis showed that actual routes were on average 12.4% shorter than OSRM-suggested paths. This finding
indicates that drivers are leveraging their local knowledge to optimize routes, potentially using shortcuts or alternative paths not captured in the standard routing algorithm.

- Significant delays are observed between segments, likely due to traffic congestion or loading/unloading operations at intermediate points
- Transition times between segments are a major contributor to overall delivery duration
- Drivers consistently find more efficient routes than those suggested by the OSRM algorithm
- Local knowledge appears to be a valuable asset in route optimization

The destination and source names were standardized by converting to lowercase, enabling more consistent route analysis and pattern identification. This standardization was crucial for accurate aggregation and comparison of route performance across different locations and time periods.

# Location Feature Engineering and Temporal Analysis

Location-based features such as *destination_state*, *destination_city*, *destination_place* , and *destination_code* were extracted from the *destination_name* column. Temporal features, including *trip_year*, *trip_month*, *trip_hour*, *trip_day*, *trip_week* , and *trip_dayofweek*, were derived from the *trip_creation_time* column to capture time-related patterns.

- Identified the Busiest Route.
- Identified Popular Routes.
- Identified trips by monthly, weekly, and daily trends

# Insights and Recommendations

The comprehensive analysis of Delhivery's operations revealed distinct delivery patterns across dataset. Peak activity occurs during 9 AM - 5 PM business hours and 6 PM - 8 PM evening hours, with Wednesdays and Saturdays showing 25% higher volume. FTL (Full Truck Load) emerged as the dominant route type, with center IND000000ACB handling the highest volume of shipments. Our paired t-tests revealed significant discrepancies between estimated OSRM times and actual delivery times, particularly in high-traffic urban areas where actual times exceeded estimates by up to 20%.

### 1 Optimize Workforce for Peak Periods

Increase workforce capacity by 30% during peak hours (9 AM - 5 PM) and by 25% on Wednesdays and Saturdays. Implement dynamic scheduling system targeting the busiest routes, particularly around center IND000000ACB. Deploy 40% of the workforce during morning peaks (9 AM - 12 PM) and maintain 35% coverage during evening rush (6 PM - 8 PM). Introduce split shifts with 4-hour rotations during these peak periods to maintain consistent service levels.

### 2 Enhance Route Optimization Algorithms

Update routing algorithms to address the 15-20% time discrepancy identified in our t-tests between OSRM estimates and actual delivery times. Incorporate historical traffic data from the past 3 weeks, particularly for FTL routes which showed the highest variance. Implement real-time traffic monitoring on the top 5 busiest routes identified in our segment-wise analysis.

### 3 Implement Systematic Delay Management

Focus on the 20% of routes causing 80% of delays, particularly in segments with segment_actual_time exceeding segment_osrm_time by more than 25%. Establish monitoring systems at key bottlenecks identified through our temporal analysis, especially during peak hours (9 AM - 5 PM). Create alternative routing options for the top 3 most delayed routes and implement a 15-minute buffer in scheduling for high-risk segments.

### 4 Strategic Market Expansion

Target expansion in regions showing limited FTL activity but high growth potential. Focus on tier-2 and tier-3 cities within 200km radius of center IND000000ACB. Establish 3 new distribution hubs in underserved areas identified through our location feature analysis. Implement region-specific pricing with 15-20% margin adjustments based on distance and demand patterns from our temporal analysis.

These recommendations should be implemented in phases over the next 6 months, with workforce optimization as the immediate priority due to its direct impact on service quality. Our analysis suggests these changes could reduce delivery delays and improve resource utilization. Monthly reviews should track key metrics including actual vs. OSRM time differences, peak hour performance, and market penetration rates in new regions.