**Data Preprocessing Report**

**Problem statement:**

Porter, India’s largest intra-city logistics marketplace, aims to accurately estimate delivery time for customer orders. Delivery time is influenced by factors such as order details, restaurant characteristics, and delivery partner availability. Through Exploratory Data Analysis (EDA), the goal is to understand these factors, handle data quality issues, engineer useful features, and prepare a clean dataset for building predictive models that improve customer satisfaction and optimize operations.

**Data cleaning:**

The timestamp columns created\_at and actual\_delivery\_time were first converted into proper datetime format to ensure consistency in calculations. Using these columns, the target variable delivery\_mins was derived as the difference between actual\_delivery\_time and created\_at, expressed in minutes. Records with invalid or negative delivery times were filtered out to maintain data quality and reliability for further analysis.

**Handling Missing Values:**

For handling missing values efficiently, numeric columns such as total\_onshift\_partners, total\_busy\_partners, and total\_outstanding\_orders were imputed using the median, as it is less sensitive to outliers and provides a more stable estimate of central tendency. For categorical columns like store\_primary\_category and order\_protocol, missing values were filled with the label ‘Unknown’, ensuring that no data was discarded while still preserving the overall category distribution.

**Outlier Treatment:**

Outliers in delivery\_mins were addressed by removing extreme records, specifically those with delivery times less than zero or greater than three days, as such values are unrealistic and could distort the analysis.

**Feature Engineering:**

Additional features were engineered from the created\_at timestamp to capture temporal patterns in delivery times. These included the hour of order placement, the day of the week, and an is\_weekend indicator to distinguish between weekdays and weekends. Furthermore, derived ratio-based features were created to represent operational dynamics more effectively: the busy\_ratio, calculated as total\_busy\_partners / total\_onshift\_partners, reflects partner availability, while items\_per\_distinct\_item, calculated as total\_items / num\_distinct\_items, captures the complexity of each order.

**Conclusion:**

The preprocessing steps produced a clean and consistent dataset. Missing values were handled systematically, categorical variables were encoded appropriately, outliers were capped, and meaningful features were engineered.