# Trip Duration Prediction

## 1 – Data Understanding

The project began with an in-depth exploration of the dataset to understand its structure and characteristics. This included: - Reviewing all columns and their meanings. - Identifying data types (numeric, categorical, datetime). - Detecting missing or inconsistent values. - Conducting statistical distribution analysis and examining variable relationships to gain early insights into data patterns.

## 2 – Baseline Model

A simple model was built to: - Assess dataset readiness for prediction. - Establish a benchmark for future advanced models. - Evaluate initial performance before applying preprocessing or feature enhancements.

## 3 – Feature Engineering

A critical phase aimed at improving input quality and predictive power: - Feature Extraction – Generated new features from raw data, including: • Time-based variables (rush hours, day of week, night trips). • Distance measures (Haversine, Manhattan, Euclidean). • Estimated speeds and passenger-related metrics. • Proximity indicators to landmarks or airports. - Feature Integration – Combined new and original features into a richer, more informative dataset.

## 4 – Outlier Capping

Applied the IQR capping method to: - Limit extreme values in numeric features and the target variable. - Reduce the influence of outliers on training. - Improve model stability and avoid unrealistic predictions.

## 5 – Implementation

Developed a full, modular code pipeline covering: - Data loading. - Preprocessing and feature engineering. - Outlier handling. - Model training using a Pipeline combining MinMax scaling with Ridge regression. - Organized code into separate Python files for clarity and reusability.

## 6 – Model Saving

Saved the trained model using joblib, enabling: - Reloading without retraining. - Reduced computational cost in production.

## 7 – Model Testing

Tested the saved model on unseen data after identical preprocessing and feature engineering. Performance metrics achieved: - $R^2$ Score: 0.6853 - RMSE: 293.85 - MAE: 210.72

## 8 – Deployment Preparation

Prepared the model for production use, ensuring that the preprocessing pipeline can be consistently applied to new incoming data.