# Exploratory Data Analysis (EDA) - Real-Time Pricing Predictor

## Introduction

This document provides a detailed step-by-step explanation of the Exploratory Data Analysis (EDA) conducted on the dataset for the Real-Time Pricing Predictor project. It includes insights, visualizations, and conclusions derived from the analysis.

## 1. Data Loading and Initial Inspection

The dataset was loaded, and the first few rows were inspected to understand its structure. The dataset contains columns such as distance, duration, traffic, weather, demand, service\_type, package\_weight, and price.

Initial Observations:

- Variables like distance, duration, and demand are likely to have a direct correlation with price.- Data is well-structured and ready for further analysis.

## 2. Data Cleaning and Preprocessing

The dataset was checked for missing values and data types. No missing values were found, and all columns had appropriate data types.

Insights: The dataset is complete and suitable for analysis without the need for extensive cleaning.

## 3. Descriptive Statistics

Descriptive statistics were generated to understand the central tendency and dispersion of numerical variables.

Insights:

- The distance ranges from short to long trips, duration varies accordingly, and price has a wide range.- High standard deviations in price and duration suggest the presence of outliers or significant variability.

## 4. Data Visualization

Various visualizations were created to explore the relationships between variables. Scatter plots, box plots, and line plots provided valuable insights.

Key Insights:

- Price vs. Distance: A positive correlation is observed; longer distances generally lead to higher prices.- Price vs. Duration: Longer durations often correspond to higher prices, but some short-duration trips are expensive, possibly due to high demand or premium services.- Service Type Impact: Premium services have a higher median price compared to economy services.- Demand Trends: Certain times of the day exhibit higher demand, indicating potential peak hours.

## 5. Correlation Analysis

A correlation matrix was computed and visualized using a heatmap. This helped identify relationships between variables.

Insights:

- Strong Correlations: Price shows strong positive correlations with distance and duration.- Moderate Correlations: Demand moderately correlates with price, suggesting higher prices during peak demand times.- Weak Correlations: Weather and traffic have weaker correlations with price, indicating less direct impact.

## 6. Feature Engineering

New features were created to enhance the dataset's predictive power. These include Price per Kilometer and a Travel Difficulty Score.

Insights:

- Price per Kilometer: Normalizes pricing over different distances, allowing for better comparison across trips.- Travel Difficulty Score: Combines weather and traffic conditions into a single metric to assess their impact on pricing.

## Conclusion

The EDA revealed valuable insights into the factors influencing real-time ride and delivery pricing. Distance, duration, and demand emerged as the primary drivers of price variability. Premium services command higher prices, and external factors like weather and traffic have a secondary impact. These findings set the stage for developing a robust predictive model for pricing.