Mid-Journey Report

Phase 1: Problem Formulation

Problem Definition:

The dataset used in this analysis consists of **New York City Taxi Trip Records from 2019**. The data includes details about taxi rides such as pickup and drop-off timestamps, passenger count, trip distance, fare amount, and location IDs.

Objectives:

- **Understand ride patterns** over time and geography.
- **Identify correlations and anomalies** in fare, distance, and passenger counts.
- **Clean and prepare the data** for model building.
- **Engineer new features** that can improve predictive performance.
- **Apply dimensionality reduction** for pattern recognition.
- **Select an appropriate model** for predictive or descriptive tasks.

Phase 2: Data Analysis and Cleansing

Dataset Source:

- Dataset: NYC Taxi Trips 2019
- Source: Kaggle (Dataset)

Preprocessing Steps:

- Data was loaded using Pandas.
- Columns with too many missing or irrelevant values were dropped.
- Filtering was applied to remove invalid data (e.g., negative fares or zero distances).
- Datetime fields were parsed to extract **day, month, hour** features.
- The categorical payment_type and RatecodeID were encoded for modeling.
- Fare and distance features were scaled using MinMaxScaler and StandardScaler.

Data Cleansing & Standardization:

- Outliers in trip distances and fare amounts were removed based on thresholding.
- Passenger count limited to 1–6 for sanity checks.
- Missing values were either dropped or imputed.

Exploratory Data Analysis (EDA)

Descriptive Statistics & Visualizations:

- Histograms plotted for **trip distance**, **fare amount**, and **passenger count**.
- Box plots used to detect **outliers** in fare and distance.
- Heatmaps and scatter plots used for correlation analysis.

Key Patterns Identified:

- **Positive correlation** between fare amount and trip distance.
- Weekends and evenings showed **higher ride demand**.
- Most rides had **1–2 passengers**, indicating typical urban travel.

Dimension Reduction:

- **PCA (Principal Component Analysis)** was applied for linear reduction.
    - Helped in understanding feature variance.
- **UMAP (Uniform Manifold Approximation and Projection)** used for non-linear projection.
    - Showed better cluster separation for trip types.

Initial Insights:

- Fare calculation is **not only distance-dependent**—likely includes time, location, and payment type.
- Certain location pairs had **high frequency**, indicating commute hotspots.

Hypothesis Testing

Hypotheses Formulated:

- **H□ (Null Hypothesis):** There is no difference in mean fare between cash and card payments.

- **H□ (Alternative Hypothesis):** There is a statistically significant difference in mean fare based on payment type.

Statistical Test Used:

- **Two-sample t-test** was applied between fare amounts for cash and card payments.

Results:

- The test **rejected H□**, indicating a significant difference in fares by payment method.

Phase 3: Model Selection

Feature Engineering:

At least 10 new features were created:

1. **Hour of Day**
2. **Day of Week**
3. **Month**
4. **Is Weekend**
5. **Trip Duration (estimated)**
6. **Trip Speed (distance/time)**
7. **Is Rush Hour**
8. **Pickup Area Type (Urban/Suburban based on zone)**
9. **Fare per Mile**
10. **Distance Bucketed Feature (short, medium, long)**

Model Candidates:

- **Linear Regression**: For fare prediction baseline.
- **Random Forest**: For handling non-linear relations and mixed data.
- **Gradient Boosting (XGBoost)**: For high accuracy and robustness.
- **K-Means**: For unsupervised pattern analysis on locations.

Validation Strategy:

- **Train/Test Split (80/20)** was used for simple models.
- **K-Fold Cross-Validation** was considered for final model tuning.

Justification:

- Feature distributions were skewed; hence, tree-based models performed better than linear models.
- Dimensionality reduction helped reduce model complexity.
- Feature scaling and encoding ensured fair comparisons between algorithms.