Feature Engineering and Data Preparation - Conclusion & Insights

# 1. Overview of Feature Engineering and Data Preparation

The Feature Engineering (FE) and Data Preparation (DP) steps are crucial for transforming raw data into meaningful features that can be used to train predictive models. The following steps were performed:  
  
1. Handling Missing Data: Missing values were addressed using imputation.  
2. Encoding Categorical Features: Non-numeric variables were encoded to be processed by machine learning models.  
3. Feature Scaling: Numerical features were standardized to bring them to the same scale.  
4. Dimensionality Reduction with PCA: Reduced the feature space while preserving variance in the data.  
5. Mutual Information: Determined the relevance of features for predicting each target variable.

# 2. Data Preprocessing Steps Explained

## Step 1: Handling Missing Data (Imputation)

Missing values in both features and target columns were handled using the SimpleImputer from Scikit-learn. For features, the missing values were imputed with the most frequent value in each column (strategy='most\_frequent'), which works well for both categorical and numerical columns. For target columns, the missing values for continuous numeric variables were imputed with the mean value (strategy='mean').

## Step 2: Encoding Categorical Variables

The categorical columns were encoded using LabelEncoder, which converts non-numeric data into numerical format. This makes the data compatible with machine learning algorithms that require numerical inputs.

## Step 3: Scaling Numerical Features

The numerical features were scaled using StandardScaler. This step normalizes the features so that they have a mean of 0 and a standard deviation of 1. It is crucial because many machine learning algorithms, such as PCA, perform better when features are on the same scale.

## Step 4: Dimensionality Reduction with PCA

PCA was applied to reduce the dimensionality of the dataset while retaining 95% of the variance. This technique transforms the data into a set of orthogonal components that are linear combinations of the original features. This reduces the complexity of the model and helps in preventing overfitting.

## Step 5: Feature Importance (Mutual Information)

Mutual information was calculated using the mutual\_info\_regression method to determine the relevance of each feature with respect to the target variables. This step identifies the features that share the most information with the target variables and guides the selection of important features for model training.

# 3. Key Insights from the Outputs

## a. Mutual Information Results

After computing mutual information, features were ranked by their relevance to each target variable. Features with higher mutual information scores are prioritized for model training. For example, if 'Delivery\_Distance\_km' has a high mutual information value with 'trip\_duration\_min', it is an important feature for predicting the trip duration.

## b. Processed Data with PCA Features

The transformed dataset, which now includes principal components, has fewer dimensions but retains 95% of the original variance. This allows for a simpler model while maintaining most of the information.

## c. Ready-to-Use Dataset

The final processed dataset, which contains both the original and PCA features, is now ready for model training. This dataset includes engineered features and reduced PCA components, offering a hybrid approach for building predictive models.

# 4. Conclusion and Next Steps

The feature engineering and data preparation process has ensured that the data is clean, transformed, and ready for machine learning modeling. The following conclusions can be made from the steps performed:  
  
1. Missing Data Handling: Imputation ensures that no data points are lost, and all columns can be used effectively.  
2. Encoding Categorical Variables: Label encoding ensures that non-numeric data is made usable for machine learning.  
3. Scaling Numerical Features: Standardization improves model performance by ensuring no feature dominates due to scale differences.  
4. Dimensionality Reduction: PCA reduces the feature space, making the data more manageable and preventing overfitting.  
5. Feature Importance: Mutual information helps identify the most informative features for model training.  
  
The next steps include selecting machine learning models, training them on the processed dataset, and fine-tuning the models for optimal performance. Additionally, the insights from feature importance can guide the selection of the most relevant features for training the models.