The questions to be answered

This is the format

Preprogrammed questions to be asked to the user one by one

User can click and choose any one of the options

**Make**  - Tells the brand / company name of the car manufacturer

**Model** - tells about the specific car model

Body Type – SUV or sedan or hatchback or mpv or coupe or etc.

Engine Type – petrol or diesel or electric

Ex-Showroom Price ($) – price of new car bought from showroom

Year of Manufacture – tells the year of manufacture

Transmission Type - manual or automatic – for electric its always automatic its not a choice

Ownership status - 1st hand or 2nd hand

Resale Value ($) - resale value of car , depends on brand value and reliability of a given car

Maintenance Cost ($/yr) – higher the brand value and lesser the reliability more the maintenance.

Mileage (km) - applies only for 2nd hand cars.

Fuel Economy (km/l) - more the power of the car , lesser the fuel economy

User Rating - rating and feedbacks given by previous user who have bought the car – 0 to 5 rating

Safety Rating - depends on the build quality , air bag system , braking and other safety systems of the car - 0 to 5 rating

Comfort Level - 0 to 5 -

0-100 km/h Acceleration (s) – more power , lesser the acceleration time

**Car Recommendation System: Dataset Preparation and Machine Learning Model Requirements**

**Objective:**

The goal is to create a detailed, clean, and comprehensive car dataset to train a Deep Feedforward Neural Network (MLP) model for a car recommendation system. The dataset must be structured to reflect realistic scenarios and fulfill various user preferences for car purchase recommendations.

**Key Requirements for the Dataset:**

1. **Unique Records:**
   * Each record should represent a unique car model, with no repetitions or redundant data. This ensures a diverse dataset for effective model training.
2. **Diverse Attribute Range:**
   * The dataset should contain a broad range of values for each attribute to capture different real-world scenarios. Each attribute must cover the spectrum from the "worst" to the "best" conditions (e.g., low to high fuel efficiency, low to high maintenance costs).
3. **Real-Life Correlations Between Attributes:**
   * The dataset should reflect realistic correlations between car attributes to mirror real-world dynamics. For instance:
     + **High-performance cars** (e.g., high horsepower) typically have **low fuel efficiency**, **high maintenance costs**, and **high prices**.
     + **Fuel-efficient cars** often have **lower brand value**, **lower prices**, and **lower maintenance costs**.
     + **Higher safety ratings** are generally associated with **higher prices** and **larger body types** (e.g., SUVs).
     + **Luxury cars** tend to have **higher comfort levels** and **higher maintenance costs**.
     + **Manual transmissions** are usually found in **lower-priced** and **higher fuel-efficient** cars.
     + **All-Wheel Drive (AWD)** vehicles provide better performance in various terrains but may have lower fuel efficiency and higher prices.
     + **Electric cars** typically have **automatic transmissions**, **high battery capacity**, and **higher prices** due to advanced technology.
   * These correlations ensure the model provides logically correct recommendations that align with user preferences and real-world expectations.
4. **Balance and Diversity:**
   * The dataset should be balanced across different car types, brands, engine types, body types, and other attributes. This prevents the model from being biased toward any specific category and ensures it can handle various user preferences.

**User Preference Questions:**

The dataset needs to support dynamic filtering and combination logic based on the following user preference questions:

1. **Year of Manufacture**: User's preferred car manufacture year.
2. **Ex-Showroom Price**: User's budget for the car's ex-showroom price.
3. **Maintenance Cost**: Maximum annual maintenance cost the user is willing to incur.
4. **Engine Type**: User's preferred engine type (Petrol, Diesel, Electric, or Hybrid).
5. **Fuel Economy**: Minimum fuel economy (in km/l) the user is looking for.
6. **Body Type**: User's preferred car body type (e.g., Sedan, Hatchback, SUV).
7. **Transmission**: User's preference for automatic or manual transmission.
8. **User Rating**: Minimum user rating the user is willing to consider.
9. **Safety Rating**: Minimum safety rating required by the user.
10. **Comfort Level**: Desired comfort level in the car.
11. **0-100 km/h Acceleration**: Maximum acceptable time for a car to accelerate from 0 to 100 km/h.
12. **Warranty Period**: Preferred warranty period for the car.
13. **Performance Metrics**: Specific performance metrics (e.g., horsepower, torque) important to the user.
14. **Seating Capacity**: Minimum seating capacity required.
15. **Battery Capacity**: Minimum battery capacity for electric/hybrid cars.
16. **Drive Type**: User's preferred drive type (Front-Wheel Drive, Rear-Wheel Drive, All-Wheel Drive).

**Machine Learning Model: Deep Feedforward Neural Network (MLP)**

1. **Model Type**: The chosen model is a Deep Feedforward Neural Network (MLP).
   * Suitable for capturing complex patterns and relationships between different car attributes.
   * Capable of handling a wide range of inputs and providing accurate recommendations based on user preferences.
2. **Training Strategy**:
   * **Supervised Learning**: The model will be trained using labeled data to learn the mapping from input features (car attributes) to outputs (car recommendations).
   * **Backpropagation and Gradient Descent**: These techniques will be used to minimize the loss function and adjust the model's weights for better accuracy.
3. **Data Preparation for MLP**:
   * **One-Hot Encoding**: Convert categorical data (e.g., Engine Type, Body Type) into a binary matrix format suitable for the neural network.
   * **Normalization**: Normalize numerical attributes to ensure consistent scales and prevent any single feature from dominating the model training.
   * **Feature Engineering**: Create additional features or interaction terms that might enhance the model's ability to learn and provide accurate recommendations.

**Steps to Prepare the Dataset:**

1. **Data Cleaning and Deduplication**:
   * Remove duplicate records and ensure every record represents a unique car model.
2. **Attribute Diversity and Range Coverage**:
   * Ensure the dataset covers a wide range of values for each attribute, representing different real-world scenarios.
3. **Implement Real-Life Correlations**:
   * Ensure the dataset reflects realistic correlations between car attributes, such as high performance with high cost and low fuel efficiency, or high safety with higher prices.
4. **Data Encoding and Normalization**:
   * Apply one-hot encoding for categorical variables and normalize numerical data.
5. **Feature Engineering**:
   * Create derived features and interaction terms to capture complex relationships between attributes.
6. **Validate and Test Dataset**:
   * Ensure the dataset aligns with user questions and supports dynamic and logical recommendations.
7. **Final Dataset Creation**:
   * Prepare the final dataset tailored for the MLP model to ensure it is best suited for effective learning and recommendation accuracy.

**Challenges and Adjustments:**

* **File Handling Issues**: Encountered issues with saving and accessing large datasets, indicating possible environment limitations.
* **Alternative Approaches**: Considered smaller datasets, different file formats (e.g., JSON), and compression (ZIP) to handle dataset saving issues.

**Dataset Specifications:**

1. **Attributes Included**:
   * Car ID, Image (placeholder names), Make, Model, Body Type, Engine Type, Price ($), Year of Manufacture, Transmission Type, Ownership Status, Resale Value ($), Distance Driven (km), Fuel Economy (km/l), Performance Rating, User Rating, Safety Rating, Comfort Level, Maintenance Cost ($/yr), 0-100 km/h Acceleration (s), Warranty Period (years), Seating Capacity, Battery Capacity (kWh), Drive Type.
2. **Correlations and Constraints**:
   * Logical relationships between attributes have been defined to ensure realistic data representation. For example, a high-performance car would have low fuel efficiency, high maintenance costs, and a high price, while a budget-friendly car would have lower maintenance costs, higher fuel efficiency, and potentially lower safety and comfort ratings.

**Next Steps:**

1. **Final Dataset Preparation**: Create a dataset that meets all the outlined requirements and supports the machine learning model effectively.
2. **Testing and Validation**: Simulate various user scenarios to ensure the dataset can dynamically filter and provide accurate recommendations based on user inputs.
3. **Model Training**: Use the prepared dataset to train the MLP model and evaluate its performance in recommending cars based on user preferences.

**Conclusion:**

The dataset preparation is crucial for training a robust machine learning model that provides accurate car recommendations. By incorporating realistic data attributes, ensuring diversity and balance, and aligning with user preferences, the dataset will be a strong foundation for developing a recommendation system that meets user needs effectively.