**UCI Automobile Dataset: Project Requirements Document**

**Dataset Overview**

The UCI Automobile dataset contains information on various car models, focusing on specifications, pricing, and other attributes that could affect a car's market price and performance. The goal of this project is to predict the car price based on these features by implementing data cleanup, exploratory data analysis (EDA), hypothesis testing, and machine learning modeling.

**Dataset Description**

**Features Overview:**

1. **Symboling** - Insurance risk rating (-3 to +3).
2. **Normalized Losses** - Average loss payment for each car in the dataset.
3. **Make** - Manufacturer of the car (e.g., Audi, BMW).
4. **Fuel Type** - Fuel type of the car (e.g., gas, diesel).
5. **Aspiration** - Engine aspiration type (e.g., standard, turbo).
6. **Num of Doors** - Number of doors in the car.
7. **Body Style** - Body style of the car (e.g., sedan, hatchback).
8. **Drive Wheels** - Drive type (e.g., front-wheel drive, rear-wheel drive).
9. **Engine Location** - Position of the engine (front, rear).
10. **Wheel Base** - Distance between the front and rear wheels.
11. **Length, Width, Height** - Car dimensions in inches.
12. **Curb Weight** - Weight of the car without occupants or cargo.
13. **Engine Type** - Type of engine (e.g., dohc, ohcv).
14. **Num of Cylinders** - Number of engine cylinders.
15. **Engine Size** - Engine size in cubic inches.
16. **Fuel System** - Fuel injection system type.
17. **Bore, Stroke** - Bore and stroke dimensions of the engine.
18. **Compression Ratio** - Engine compression ratio.
19. **Horsepower** - Engine horsepower.
20. **Peak RPM** - Peak revolutions per minute.
21. **City MPG, Highway MPG** - Fuel efficiency in miles per gallon for city and highway.
22. **Price** - Car price in dollars.

**Project Objectives**

1. **Data Preparation and Cleanup**
   * Handle missing values and replace them with appropriate imputations.
   * Convert data types where necessary (e.g., numeric encoding for categorical variables).
   * Identify and handle outliers that may skew analysis or model training.
2. **Exploratory Data Analysis (EDA)**
   * Generate descriptive statistics for each feature (mean, median, variance, etc.).
   * Visualize distributions of key features (histograms, box plots) to observe patterns.
   * Create correlation matrices to identify relationships between features and target variable (Price).
   * See Appendix
3. **Hypothesis Testing**
   * Perform statistical hypothesis tests to examine:
     + Whether there is a significant difference in Price based on Fuel Type (use a t-test or ANOVA).
     + Correlation between Engine Size and Horsepower using a correlation test.
     + Association between Body Style and Drive Wheels using a chi-square test.
4. **Machine Learning Modeling**
   * Split the dataset into training and testing sets (e.g., 80-20 split).
   * Build and evaluate multiple machine learning models for price prediction:
     + **Linear Regression** – Create a baseline model for price prediction.
     + **Random Forest Regressor** – Use an ensemble model and tune hyperparameters
   * Evaluate models based on Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared metrics.
   * Identify the model with the best predictive accuracy and interpret its results, do a cross validation if necessary
5. **Model Interpretation and Insights**
   * Analyze feature importance to understand which features have the most impact on car price.
   * Create partial dependence plots for top features to interpret relationships with price.
   * Summarize findings and insights into a report format.
6. **Python API using Flask**
   * Develop a simple Flask REST API server for predicting the values
   * Use POSTMAN or similar technique to test it

**Project Timeline**

**Total Duration**: 4 hours

| **Task** | **Description** | **Estimated Time** |
| --- | --- | --- |
| **Data Preparation and Cleanup** | Clean missing data, handle data types and outliers | 45 minutes |
| **Exploratory Data Analysis** | Visualize and analyze feature relationships | 45 minutes |
| **Hypothesis Testing** | Conduct statistical tests on selected features | 30 minutes |
| **Machine Learning Modeling** | Build, train, and evaluate ML models | 1 hour |
| **Model Interpretation and Insights** | Analyze model importance and create final report | 30 minutes |
| **API Development** | Python FLASK API | 30 mins |

**Submission Requirements**

1. **Python code** and **Jupyter Notebooks** with all code, visualizations, and comments explaining each step.
2. **Summary Report** detailing:
   * Key findings from the EDA and hypothesis tests.
   * Model comparison and evaluation metrics.
   * Interpretation of the best model and top influential features.
   * Presentation for 10 minutes

**Appendix:**

**🚗 Automobile Dataset**

**Goal: Predict car price or identify influential features.**

**📊 EDA Questions**

1. **What is the distribution of car prices? Are there outliers or skewness?**
2. **Which features are most correlated with price? (e.g., horsepower, engine size)**
3. **Compare price distributions across fuel type (fuel-type) and body style (body-style).**
4. **How does the number of doors or drive wheels affect price?**
5. **Do luxury brands (e.g., BMW, Mercedes) show distinctly higher price ranges?**
6. **Is there a trend between engine size and fuel efficiency (city-mpg, highway-mpg)?**
7. **Are there missing values in numeric features like normalized-losses?**

**🛠️ Feature Engineering Questions**

1. **Encode make, body-style, fuel-type using label or one-hot encoding.**
2. **Impute missing values in normalized-losses using mean or regression.**
3. **Create a performance index: horsepower / curb-weight. Is it predictive?**
4. **Create a luxury\_brand binary feature for expensive brands.**
5. **Engineer a fuel cost feature based on city-mpg and assumed gas price.**
6. **Normalize skewed numerical features (price, engine-size) using log transformation.**