Automobile Dataset Analysis using NumPy

Presented by: Riya Manoj

Introduction

- NumPy is a powerful open-source Python library for numerical and scientific computing.
- Useful for handling large arrays and mathematical operations.
- Forms the foundation for data science and Machine Learning libraries like Pandas and TensorFlow.
- Project goal: Apply NumPy on the automobile dataset to clean data and analyze horsepower.

Problem Statement

- Horsepower values were stored as strings, not numbers.
- Needed to analyze cars with respect to horsepower and mpg.
- Challenge: connect numerical horsepower with car names for insights.

Proposed Solution

- Used pandas to load dataset and identify missing values.
- Replaced missing horsepower values with mean (data cleaning).
- Converted horsepower into NumPy array for operations.
- Applied slicing and sorting to extract insights.

Dataset Overview

- Dataset: Auto MPG , 398 rows, 9 columns.
- Key attributes: mpg, horsepower, car name.
- Issue: horsepower column had non-numeric values.

Tools Used

- Python (programming language)
- NumPy (numerical operations, slicing, sorting)
- pandas (data cleaning and dataset handling)
- Google Colab (implementation environment)

Implementation

- Step 1: Loaded dataset and checked info.
- Step 2: Converted non numeric horsepower column to numeric.
- Step 3: Handled missing horsepower values with mean.
- Step 4: Converted horsepower column into NumPy array.
- Step 5: Performed slicing to view specific ranges.
- Step 6: Sorted horsepower to find top 5 cars.

Insights

- Missing horsepower values were successfully replaced.
- Cars with higher horsepower generally had lower mpg.
- Top 5 cars by horsepower identified (e.g., Pontiac grand prix , Pontiac Catalina, Buick estate wagon).
- Data cleaning made analysis possible.
- *Note: The dataset is historical, so high-horsepower cars identified (Pontiac, Buick , etc.)were considered powerful in that era, but may not match today's high-performance standards*

Conclusion

- NumPy simplified numerical processing and array handling.
- Preprocessing improved dataset quality and accuracy.
- Analysis provided real-world insights into automobile performance.
- Project highlighted the importance of data cleaning + NumPy operations in data analytics.

References

- Auto MPG Dataset -Kaggle
- NumPy Official Documentation.
- pandas Official Documentation.

Thank You