Project Report: Mileage Prediction Using ETL + SQLite

Abstract This report outlines the development of an end-to-end ETL (Extract, Transform, Load) pipeline designed to analyze car data for insights into mileage prediction based on factors such as brand, condition, and price. The project demonstrates the integration of Python and SQLite for effective data manipulation, storage, and visualization. The aim is to streamline data preparation and analysis without relying on machine learning techniques, providing a robust framework for similar analytical tasks.

Introduction The automotive industry generates extensive data on various aspects of vehicles, including mileage, price, and condition. This project focuses on leveraging an ETL pipeline to extract insights from such data, emphasizing the relationship between mileage ranges and car pricing. The primary goal is to design a systematic pipeline that extracts data from a source, transforms it into a structured format, and loads it into a database for analysis and visualization.

Tools and Technologies

- **Python**: For data manipulation and scripting.
- **Pandas**: Used to clean, process, and transform the dataset.
- **SQLite**: Employed for structured data storage and querying.
- **Matplotlib**: For generating data visualizations.

Methodology

1. Extract Phase:

- The dataset containing car attributes such as brand, model, mileage, and price was loaded from a CSV file.
- Pandas was used to structure the data into a DataFrame for further processing.

2. Transform Phase:

- Removed irrelevant columns and handled missing data to ensure consistency.
- Standardized mileage data by creating a new column for "Mileage Range," categorizing values into predefined buckets (e.g., 0–50K, 50K–100K).
- Stored the cleaned data in a transformed CSV file for subsequent steps.

3. Load Phase:

- Imported the transformed data into an SQLite database.

- Created schemas and tables to organize data, including a table dedicated to mileage range analytics.

4. Data Analysis:

- SQL queries were executed to uncover trends such as:
    - Average car price by mileage range.
    - The distribution of mileage ranges across brands and conditions.
- Results were stored in a SQLite database for easy reference.

5. Visualization:

- A column chart was generated using Matplotlib to illustrate the average car price across mileage ranges.
- Visualizations facilitated the interpretation of trends and relationships within the data.

Results and Insights

- The analysis revealed that cars with lower mileage ranges (e.g., 0–50K) are associated with higher average prices, confirming the significant influence of mileage on car value.
- Mileage distribution varied notably across different brands, reflecting manufacturing and usage patterns.
- Data visualization made these trends accessible and actionable.

How to Run the Project

1. Install Python along with Pandas, SQLite3, and Matplotlib libraries.
2. Download the dataset (cleaned_cars_data.csv) and place it in the project directory.
3. Run the following scripts sequentially:
    - extract.py: Extracts and organizes raw data.
    - transform.py: Cleans and processes the data, saving it to a transformed file.
    - load.py: Loads the transformed data into an SQLite database.
    - query.py: Executes SQL queries to derive insights.
    - visualize.py: Creates visual representations of the analysis.
4. Review the processed data in the SQLite database and inspect the visual output.

Conclusion This project demonstrates the utility of ETL pipelines for data-driven decision-making in non-machine learning contexts. By integrating Python's data processing capabilities with SQLite's querying efficiency, this pipeline offers a scalable framework for analyzing complex datasets. The structured approach not only simplifies data analysis but also provides actionable insights, making it an invaluable tool for industries reliant on data analytics.