CSC343:
Term Project
Phase 1

September 30th, 2021

Prerak Chaudhari | 1005114760
Jaakulan Subeethakumar | 1005225757

Domain

Cars from 1970-1982.

Dataset

Description:
This is a modified version of the 1983 ASA Data Exposition dataset about automobiles. The original dataset contained information about miles per gallon (mpg), cylinder count, engine displacement, horsepower, vehicle weight, 0-60 mph acceleration time, model year, and origin for all cars between 1970-1982. The modified version has already split the data into various tables and provided data directories. Furthermore, some data cleaning is done and auxiliary information is added. Both datasets are publicly available on the internet (thus, open-source).

Link to the dataset:
https://users.csc.calpoly.edu/~dekhtyar/365-Winter2015/data/CARS/

What information in it is relevant:
All information inside the dataset is useful except the nickname and full name of the car manufacturers; however, we are keeping these attributes present in our schema so other users of our database can perform queries in the future which may involve said information.

Any learning you will have to do in order to interpret the data:
Yes we will have to learn about engine displacement and how it affects certain cars as well as its relation to the amount of cylinders there are. We also need to know how the amount of cylinders can affect a car's weight and horsepower.

Any cleaning up you think you will have to do in order to use the data:
In the csv of car manufacturers, there is a nonsensical row detailing a car manufacturer named 'hi' whose home country is null. The row will need to be removed from this table as well as another one which details car models. Then, the ID column in some of our tables will need to be manually adjusted.

Also, the csv files will need to be reformatted as per our schema.

Questions

1. For each year find the continents whose models for that year had the highest average gas mileage. Report the year, the continent, the number of models produced that year, the highest average gas mileage and the avg acceleration of the models. Present the output in chronological order of year.
2. For each year, find the model that had the highest power-to-weight ratio. Report the year, make description, number of cylinders, horsepower, weight, and power-to-weight ratio.
3. Every year has some combination of cars with 3, 4, 5, 6 and 8 cylinders. For each year, determine which cars have the highest and lowest number of cylinders. Of these two groups, find the most fuel-efficient high cylinder vehicle (denoted Car A) and least fuel-efficient low cylinder vehicle (denoted Car B) (Fuel efficiency will be a measure of the highest miles per gallon). For Car A, report the year, make description, engine displacement, fuel efficiency, and cylinder count. Next to it, report the same for Car B, as well as the absolute difference in engine displacement and fuel efficiency between the two.

Schema

Our schema is based on the given data directories (see README.CARS.TXT inside the dataset's zip file), but with some changes.

Firstly, schema names and attributes have been changed for more clarity.

Secondly, we are further splitting the cars-data csv file into 3 separate files; one for miles per gallon, one for horsepower and one for the remaining metrics. This is because some rows are missing a value for either miles per gallon or horsepower and we wish to enforce attribute requirements for all columns in our tables.

Thirdly, we are replacing the Model column in the cars-names csv file with its numerical identifier from the model-list csv file. This is done to reduce redundancy as the model name should not be present in multiple relations. While this may come at a minor performance cost due to additional natural joins, we believe it is good practice to design the schema this way. Note that our first query deals with car manufacturers as we need to determine their country (thus, continent) of origin. This schema design means we don't need to re-lookup the model name; furthermore, our comparison and lookup times are faster because we are using an integer key versus a string key.

The resultant schema, as shown on the upcoming pages, is good because not only is it easy to understand, but it eliminates redundancies in each relation. Furthermore, each relation has an integer primary key and its foreign keys are also integers. This speeds up comparisons and lookups, compared to using a string key which is present in some relations, because integers are smaller than strings.

Integrity Constraints:

- Manufacturers[CountryID] ⊆ Countries[CountryID]
- Cars[ModelID] ⊆ Models[ModelID]
- MilesPerGallon[CarID] ⊆ Cars[CarID]
- Horsepower[CarID] ⊆ Cars[CarID]
- Parameters[CarID] ⊆ Cars[CarID]
- Countries[ContinentID] ⊆ Continents[ContinentID]
- Models[ManufacturerID] ⊆ Manufacturers[ManufacturerID]
- miles per gallon, cylinder count, engine displacement, horsepower, vehicle weight, 0-60 mph acceleration time and model year must be positive numbers

Manufacturers(<u>ManufacturerID</u>, <u>NickName</u>, <u>FullName</u>, CountryID)
A tuple in this relation represents a car manufacturer.

Attribute	Description	Type	Required
ManufacturerID	row-indexed identifier for car manufacturers in this dataset	INT	YES
NickName	car manufacturer's nickname	STRING	YES
FullName	car manufacturer's full name	STRING	YES
CountryID	numerical identifier of the car manufacturer's home country	INT	YES

Models(<u>ModelID</u>, ManufacturerID, <u>ModelName</u>)
A tuple in this relation represents a car model.

Attribute	Description	Type	Required
ModelID	row-indexed identifier for car models in this dataset	INT	YES
ManufacturerID	numerical identifier for car manufacturer of this model	INT	YES
ModelName	name of car model	STRING	YES

Cars(CarID, ModelID, Make)
A tuple in this relation represents the model and make description of a car.

Attribute	Description	Type	Required
CarID	row-indexed identifier for cars in this dataset	INT	YES
ModelID	numerical identifier of the car's model	INT	YES
Make	car's make description	STRING	YES

MilesPerGallon(CarID, MPG)
A tuple in this relation represents the fuel efficiency of a car in miles per gallon.

Attribute	Description	Type	Required
CarID	unique identifier for each car	INT	YES
MPG	mileage per gallon	FLOAT	YES

Horsepower(CarID, HP)
A tuple in this relation represents the horsepower of a car.

Attribute	Description	Type	Required
CarID	unique identifier for each car	INT	YES
Horsepower	engine power in horsepower	INT	YES

Parameters
(CarID, Cylinders, Displacement, Weight, Acceleration, Year)
A tuple in this relation represents the other operational parameters of a car.

Attribute	Description	Type	Required
CarID	unique identifier for each car	INT	YES
Cylinders	number of engine cylinders	INT	YES
Displacement	engine displacement volume in cubic inches	INT	YES
Weight	car's weight in pounds	INT	YES
Acceleration	0-60mph acceleration time in seconds	FLOAT	YES
Year	car's year of production	INT	YES

Continents(ContinentID, ContinentName)
A tuple in this relation represents a continent.

Attribute	Description	Type	Required
ContinentID	row-indexed identifier for continents in the dataset	INT	YES
ContinentName	continent name	STRING	YES

Countries(<u>CountryID</u>, <u>CountryName</u>, ContinentID)
A tuple in this relation represents a country.

Attribute	Description	Type	Required
CountryID	row-indexed identifier for countries in this dataset	INT	YES
CountryName	country name	STRING	YES
ContinentID	numerical identifier of the country's continent	INT	YES