CSC343:
Term Project
Phase 2

November 19th, 2021

Prerak Chaudhari | 1005114760
Jaakulan Subeethakumar | 1005225757

Design Decisions

In phase 1, one of our proposed changes to the schema from what was provided in our dataset's data directory was to further split 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 wished to enforce attribute requirements for all columns in our tables.

We received feedback from the TA that this change was unnecessary because there is a 1-to-1 relationship between car ID and these parameters. Therefore, we will be keeping the csv file as is, removing the MilesPerGallon and Horsepower relations, and adding nullable columns for mpg and horsepower to our Parameters table. Given that both mpg and horsepower are never both null for any row in this csv file, and our queries never involve both mpg and horsepower at the same time, we can simply add a where clause to our queries that filters out tuples with a null in the interested column.

Aside from this, we made some further changes to the schema on our own accord. Firstly, for the Parameters table, we updated the column type of Displacement to be FLOAT because we found a row with a float value for this column. Secondly, to better represent the dataset and domain, we also updated the CountryName and ContinentName columns of the respective Countries and Continents relations to be keys.

Our schema in phase 1 was described using relational algebra. Some of our relations have multiple keys, which isn't allowed in SQL. Therefore, for each relation, we opted to use the integer ID key as the primary key of its respective SQL table, with any other keys being unique columns.

We are still 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. For example, 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 (integers are smaller than strings).

Now, we believe the shown schema in our demo.txt file is good because not only is it easy to understand, but it eliminates redundancies in each table. Furthermore, each table 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. Finally, the check constraints in our Parameter table prevent one from inserting nonsensical car metrics (i.e. negative values for miles per gallon, acceleration, weight, etc).

Cleaning Process

Detailed below are the steps taken to clean our dataset. Pictures are included with every step for added clarity. While this causes the report to go past 2 pages, we believe that these visual cues provide greatly beneficial added context that would help one reproduce these steps with ease.

1. The nonsensical manufacturer 'hi' whose country of origin is null was removed from the car-makers csv file to enforce attribute requirements:

	A	B	C	D
1	Id	Maker	FullName	Country
2	1	'amc'	'American Motor Company'	1
3	2	'volkswagen'	'Volkswagen'	2
4	3	'bmw'	'BMW'	2
5	4	'gm'	'General Motors'	1
6	5	'ford'	'Ford Motor Company'	1
7	6	'chrysler'	'Chrysler'	1
8	7	'citroen'	'Citroen'	3
9	8	'nissan'	'Nissan Motors'	4
10	9	'fiat'	'Fiat'	5
11	10	'hi'	'hi'	null
12	11	'honda'	'Honda'	4
13	12	'mazda'	'Mazda'	4
14	13	'daimler benz'	'Daimler Benz'	2
15	14	'opel'	'Opel'	2
16	15	'peugeaut'	'Peugeaut'	3
17	16	'renault'	'Renault'	3
18	17	'saab'	'Saab'	6
19	18	'subaru'	'Subaru'	4
20	19	'toyota'	'Toyota'	4
21	20	'triumph'	'Triumph'	7
22	21	'volvo'	'Volvo'	6
23	22	'kia'	'Kia Motors'	8
24	23	'hyundai'	'Hyundai'	8

2. The corresponding model was removed from the model-list csv file to avoid foreign key violations:

	A	B	C
1	ModelId	Maker	Model
2	1	1	'amc'
3	2	2	'audi'
4	3	3	'bmw'
5	4	4	'buick'
6	5	4	'cadillac'
7	6	5	'capri'
8	7	4	'chevrolet'
9	8	6	'chrysler'
10	9	7	'citroen'
11	10	8	'datsun'
12	11	6	'dodge'
13	12	9	'fiat'
14	13	5	'ford'
15	14	10	'hi'
16	15	11	'honda'
17	16	12	'mazda'
18	17	13	'mercedes'
19	18	13	'mercedes-benz'
20	19	5	'mercury'
21	20	8	'nissan'
22	21	4	'oldsmobile'
23	22	14	'opel'
24	23	15	'peugeot'
25	24	6	'plymouth'
26	25	4	'pontiac'
27	26	16	'renault'
28	27	17	'saab'
29	28	18	'subaru'
30	29	19	'toyota'
31	30	20	'triumph'
32	31	2	'volkswagen'
33	32	21	'volvo'
34	33	22	'kia'
35	34	23	'hyundai'
36	35	6	'jeep'
37	36	19	'scion'

3. There is 1 car of model ‘hi’ in the car-names csv file that had to be deleted to avoid foreign key violations:

	A	B	C
1	Id	Model	Make
2	1	'chevrolet'	'chevrolet chevelle malibu'
3	2	'buick'	'buick skylark 320'
4	3	'plymouth'	'plymouth satellite'
5	4	'amc'	'amc rebel sst'
6	5	'ford'	'ford torino'
7	6	'ford'	'ford galaxie 500'
8	7	'chevrolet'	'chevrolet impala'
9	8	'plymouth'	'plymouth fury iii'
10	9	'pontiac'	'pontiac catalina'
11	10	'amc'	'amc ambassador dpl'
12	11	'citroen'	'citroen ds-21 pallas'
13	12	'chevrolet'	'chevrolet chevelle concours (sw)'
14	13	'ford'	'ford torino (sw)'
15	14	'plymouth'	'plymouth satellite (sw)'
16	15	'amc'	'amc rebel sst (sw)'
17	16	'dodge'	'dodge challenger se'
18	17	'plymouth'	'plymouth cuda 340'
19	18	'ford'	'ford mustang boss 302'
20	19	'chevrolet'	'chevrolet monte carlo'
21	20	'buick'	'buick estate wagon (sw)'
22	21	'toyota'	'toyota corona mark ii'
23	22	'plymouth'	'plymouth duster'
24	23	'amc'	'amc hornet'
25	24	'ford'	'ford maverick'
26	25	'datsun'	'datsun pl510'
27	26	'volkswagen'	'volkswagen 1131 deluxe sedan'
28	27	'peugeot'	'peugeot 504'
29	28	'audi'	'audi 100 ls'
30	29	'saab'	'saab 99e'
31	30	'bmw'	'bmw 2002'
32	31	'amc'	'amc gremlin'
33	32	'ford'	'ford f250'
34	33	'chevrolet'	'chevy c20'
35	34	'dodge'	'dodge d200'
36	35	'hi'	'hi 1200d'
37	36	'datsun'	'datsun pl510'
38	37	'chevrolet'	'chevrolet vega 2300'
39	38	'toyota'	'toyota corona'
40	39	'ford'	'ford pinto'
41	40	'volkswagen'	'volkswagen super beetle 117'
42	41	'amc'	'amc gremlin'
43	42	'plymouth'	'plymouth satellite custom'
44	43	'chevrolet'	'chevrolet chevelle malibu'
45	44	'ford'	'ford torino 500'
46	45	'amc'	'amc matador'

4. This car's metrics also had to be deleted from the cars-data csv file to avoid foreign key violations:

	A	B	C	D	E	F	G	H
1	Id	MPG	Cylinders	Edispl	Horsepower	Weight	Accelerate	Year
2	1	18	8	307	130	3504	12	1970
3	2	15	8	350	165	3693	11.5	1970
4	3	18	8	318	150	3436	11	1970
5	4	16	8	304	150	3433	12	1970
6	5	17	8	302	140	3449	10.5	1970
7	6	15	8	429	198	4341	10	1970
8	7	14	8	454	220	4354	9	1970
9	8	14	8	440	215	4312	8.5	1970
10	9	14	8	455	225	4425	10	1970
11	10	15	8	390	190	3850	8.5	1970
12	11	null	4	133	115	3090	17.5	1970
13	12	null	8	350	165	4142	11.5	1970
14	13	null	8	351	153	4034	11	1970
15	14	null	8	383	175	4166	10.5	1970
16	15	null	8	360	175	3850	11	1970
17	16	15	8	383	170	3563	10	1970
18	17	14	8	340	160	3609	8	1970
19	18	null	8	302	140	3353	8	1970
20	19	15	8	400	150	3761	9.5	1970
21	20	14	8	455	225	3086	10	1970
22	21	24	4	113	95	2372	15	1970
23	22	22	6	198	95	2833	15.5	1970
24	23	18	6	199	97	2774	15.5	1970
25	24	21	6	200	85	2587	16	1970
26	25	27	4	97	88	2130	14.5	1970
27	26	26	4	97	46	1835	20.5	1970
28	27	25	4	110	87	2672	17.5	1970
29	28	24	4	107	90	2430	14.5	1970
30	29	25	4	104	95	2375	17.5	1970
31	30	26	4	121	113	2234	12.5	1970
32	31	21	6	199	90	2648	15	1970
33	32	10	8	360	215	4615	14	1970
34	33	10	8	307	200	4376	15	1970
35	34	11	8	318	210	4382	13.5	1970
36	35	9	8	304	193	4732	18.5	1970
37	36	27	4	97	88	2130	14.5	1971
38	37	28	4	140	90	2264	15.5	1971
39	38	25	4	113	95	2228	14	1971
40	39	25	4	98	null	2046	19	1971
41	40	null	4	97	48	1978	20	1971
42	41	19	6	232	100	2634	13	1971
43	42	16	6	225	105	3439	15.5	1971
44	43	17	6	250	100	3329	15.5	1971
45	44	19	6	250	88	3302	15.5	1971

5. In the car-names csv file, there are 2 instances of erroneous whitespace padding that were corrected by removing the whitespace before the opening quotation mark:

```
302  302,'mazda', 'mazda glc deluxe'
340  340, 'volkswagen','volkswagen rabbit'
```

6. The Model column in the cars-names csv file is replaced with its numerical identifier from the model-list csv file.

   This change was accomplished using the pandas library:

```
import pandas as pd

carNames = pd.read_csv('car-names.csv')
modelList = pd.read_csv('model-list.csv')

combined = pd.merge(carNames, modelList, on='Model')
df = combined[['Id', 'ModelId', 'Make']].sort_values(by=['Id'])

df.to_csv('car-names.csv', index=False)
```

7. All instances of the word "null" in the cars-data csv file are replaced by an unquoted empty string as per standard CSV format for handling null values: