CSC343 Project Presentation

Jaakulan Subeethakumar
and
Prerak Chaudhari

Domain

Cars from 1970-1982

Investigative question 1

For each year, find the continent 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. In case of ties between continents for a given year, report all the tied continents and their assorted metrics.

Investigative question 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. In case of ties between models for a given year, report all the tied models and their assorted metrics.

Investigative question 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 vehicles (denoted type A) and least fuel-efficient low cylinder vehicles (denoted type B). There may be multiple vehicles of either type in the event of MPG ties. For every year, report every pairing of type A and type B cars. More explicitly, for a given pairing, with the type A car denoted as Car A, and the type B car as Car B, report the following columns in a row:

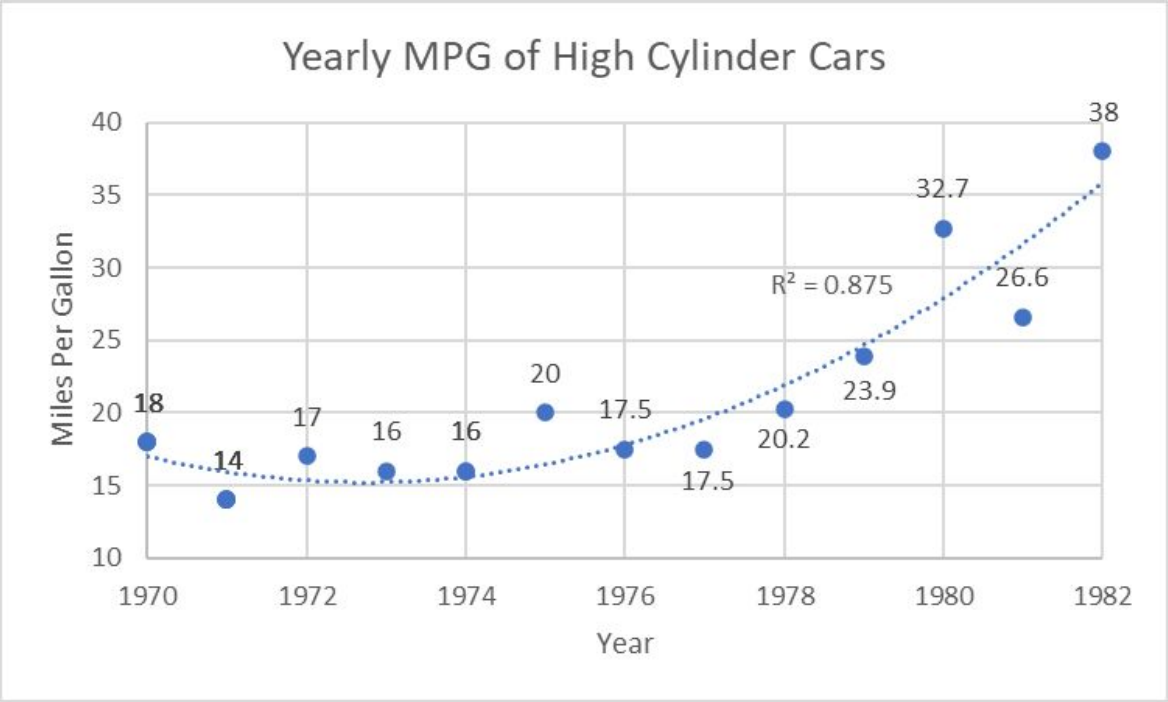
A. Year
B. Make description of Car A
C. Engine displacement of Car A
D. Fuel efficiency of Car A
E. Cylinder Count of Car A
F. Make description of Car B
G. Engine displacement of Car B
H. Fuel efficiency of Car B
I. Cylinder Count of Car B
J. Difference in MPG between Car B and Car A (i.e. column H - column D)
K. Difference in engine displacement between Car B and Car A (i.e. column G - column C)

Results - Better engine design

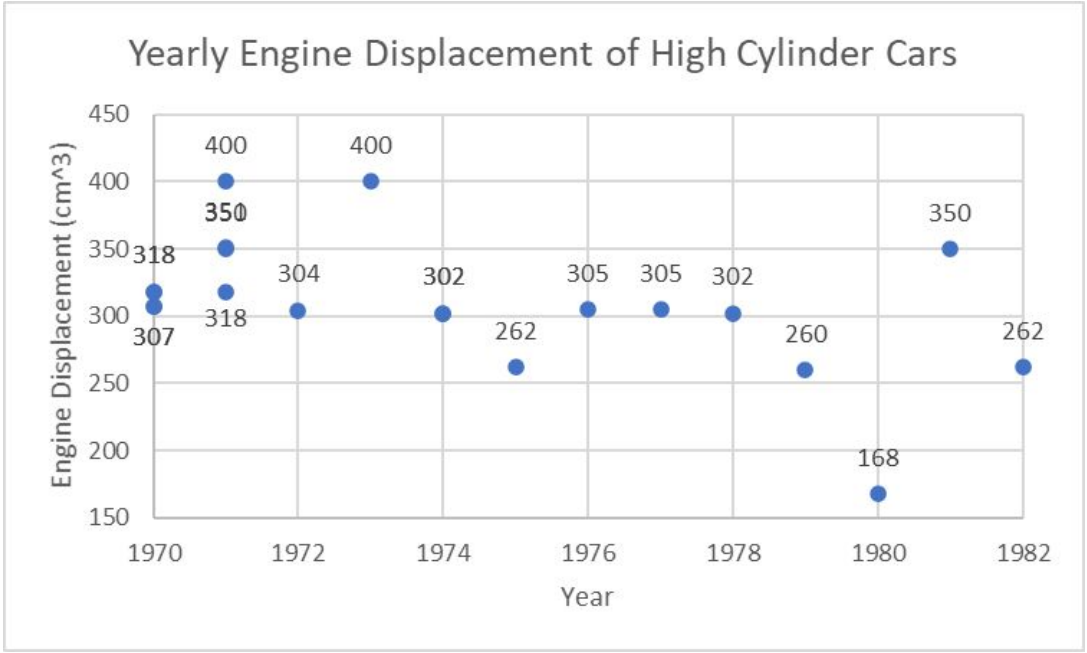
A trend for better engine design:

- Higher MPG year by year
- Massive decrease in engine throughout the later years
- Potentially massive changes in engine technology
- Fuel efficiency with difference in cylinders

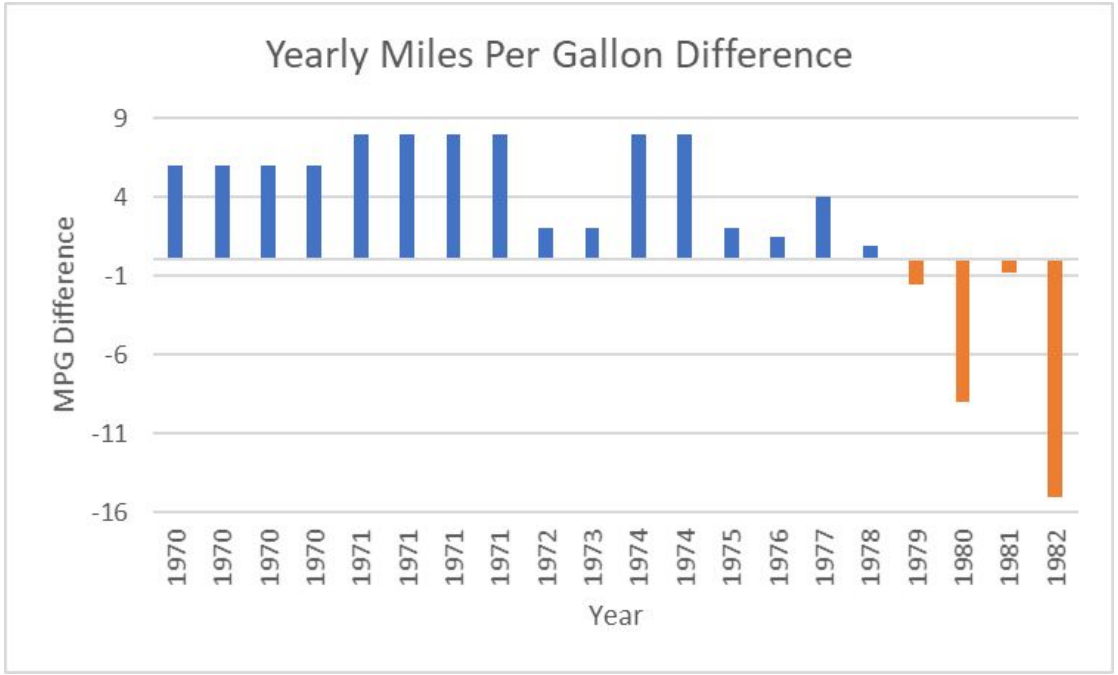
Higher MPG year by year



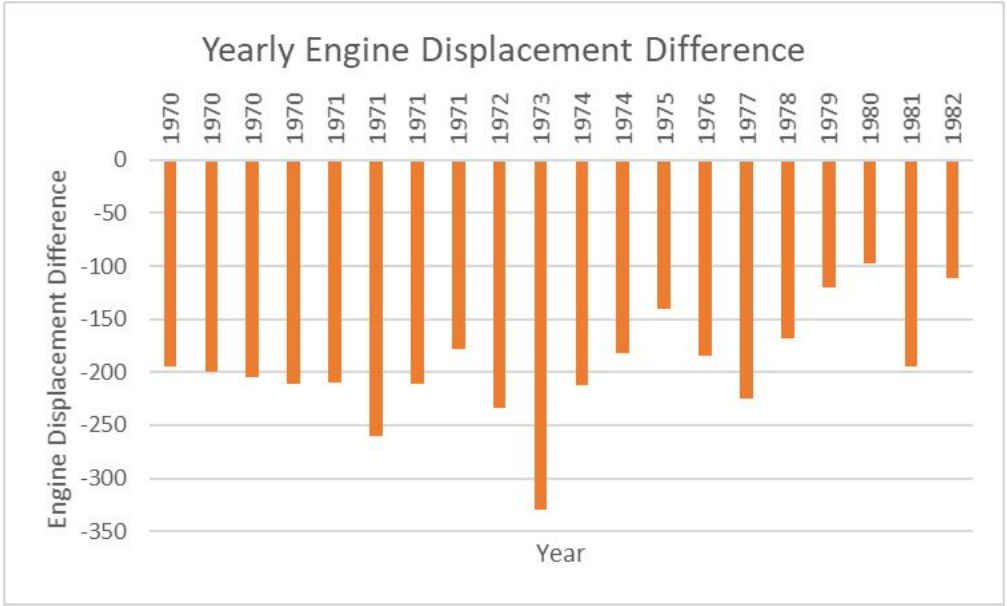
Massive decrease in engine displacement



Potentially massive changes in engine technology



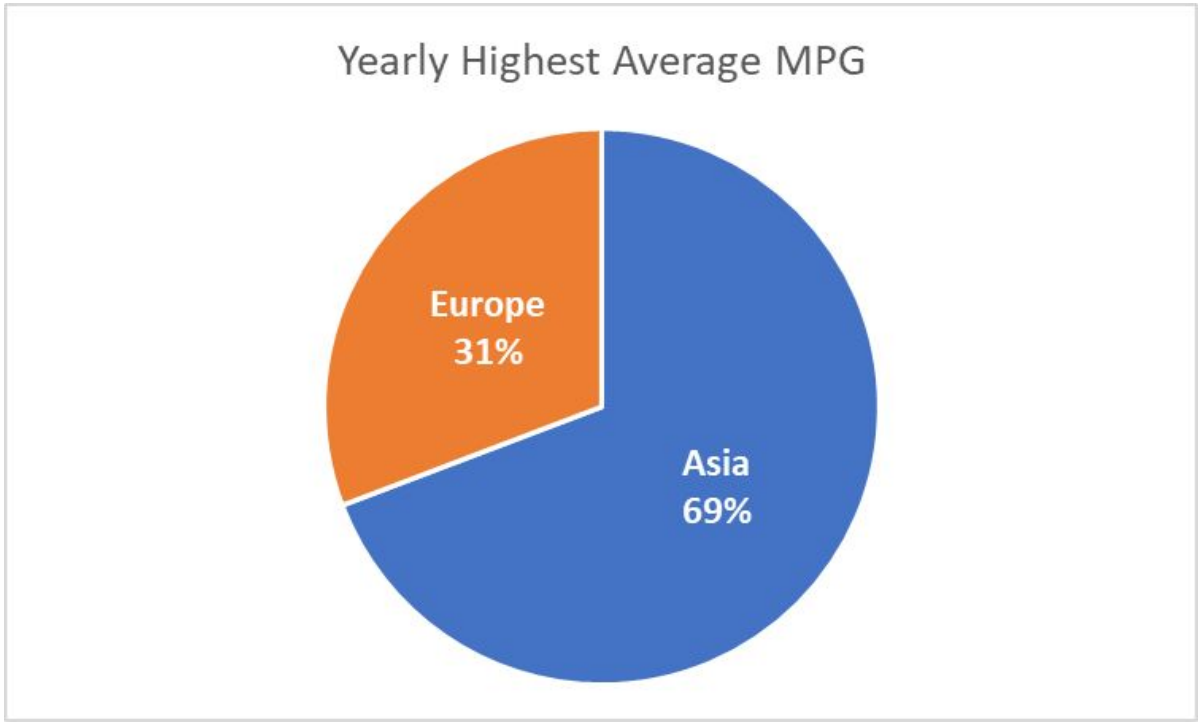
Engine displacement with difference in cylinders



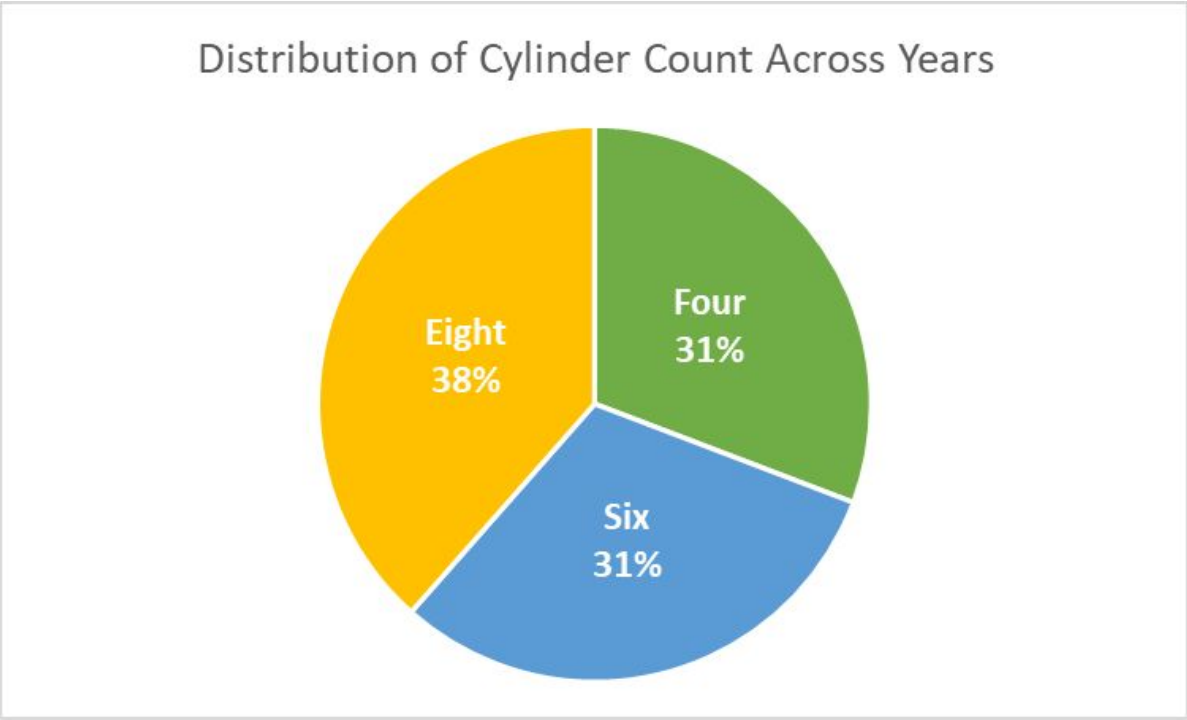
Results - Interesting tidbits

- American cars of this era were not fuel efficient
- More cylinders make it easier to achieve a higher power-to-weight ratio

The worst continent for fuel efficiency



More cylinders, more power



Challenges - Figuring out which questions?

To find the best questions we divided our car consumers into demographics:

- What would car enthusiasts want to know?
- What would a normal consumer want to know?
- How can we devise questions to answer all the intriguing questions?

What would car enthusiasts want to know?

What would a normal consumer want to know?

How can we devise questions to answer all the intriguing questions?

Challenges - Imperfect data

- CSV files had nullable columns (Country, MPG, Horsepower)
  => Implications of nullable columns on investigative questions
- Erroneous whitespace and null value formatting
  => Lost rows from data cleaning and unable to load dataset into database
- String foreign keys
  => Bad schema design to use non-integer foreign keys

Solution: query foresight and data cleaning

- The nullable columns did not impact our investigative questions
- Data for car manufacturer with null country of origin was removed
- Data cleaning process also fixed the erroneous CSV formatting and converted the string foreign key into an integer

Removing car manufacturers

	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

Remove the nonsensical car manufacturer “hi” whose country of origin is null

Removing car models

	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'

Remove the corresponding model to avoid foreign key violations

Removing car makes

	A	B	C
1	Id	Model	Make
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'

Remove the corresponding car to avoid foreign key violations

Removing car parameters

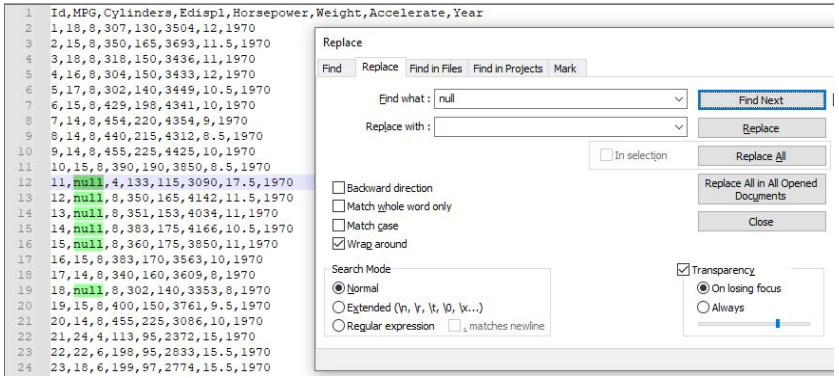
	A	B	C	D	E	F	G	H
1	Id	MPG	Cylinders	Edispl	Horsepower	Weight	Accelerate	Year
32	31	21	6	199	90	2648	15	1970
33	32	10	8	360	215	4615	14	1970
34	33	10	8	307	200	4376	15	1970
35	34	11	8	318	210	4382	13.5	1970
36	35	9	8	304	193	4732	18.5	1970
37	36	27	4	97	88	2130	14.5	1971
38	37	28	4	140	90	2264	15.5	1971
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

Remove the car's corresponding metrics to avoid foreign key violations

Fixing erroneous CSV formatting

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

There are 2 instances of erroneous whitespace padding that were corrected by removing the whitespace before the opening quotation mark



All instances of the word “null” were replaced with an unquoted empty string as per standard CSV format for handling null values

Converting string foreign key into an integer

Original:

Id	Model	Make
1	'chevrolet	'chevrolet chevelle malibu'
2	'buick'	'buick skylark 320'
3	'plymouth	'plymouth satellite'
4	'amc'	'amc rebel sst'
5	'ford'	'ford torino'
6	'ford'	'ford galaxie 500'
7	'chevrolet	'chevrolet impala'
8	'plymouth	'plymouth fury iii'
9	'pontiac'	'pontiac catalina'
10	'amc'	'amc ambassador dpl'
11	'citroen'	'citroen ds-21 pallas'

Desired:

Id	ModelId	Make
1	7	'chevrolet chevelle malibu'
2	4	'buick skylark 320'
3	24	'plymouth satellite'
4	1	'amc rebel sst'
5	13	'ford torino'
6	13	'ford galaxie 500'
7	7	'chevrolet impala'
8	24	'plymouth fury iii'
9	25	'pontiac catalina'
10	1	'amc ambassador dpl'
11	9	'citroen ds-21 pallas'

The string identifiers in the Model column
were replaced with their numerical counterparts

Performing CSV file manipulation with Python

```
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)
```

This change was accomplished using the Pandas library

Lessons - Jaakulan

Databases are great and easy way to store data from any dataset:

- You can reformat the data to answer multiple questions
- You can create different types of tables for output
- Very fast to work
- Great for organizing data in a specific way (i.e no duplicates)

Lessons - Prerak

- Real life data may not be perfect
- Assignments did not require us to worry about null values
- Project forced us to think about the implications of nullable columns
- Course work has equipped me with the tools to deal with non-idealities

Thanks for Listening

Any Questions?