

Sundus Yawar

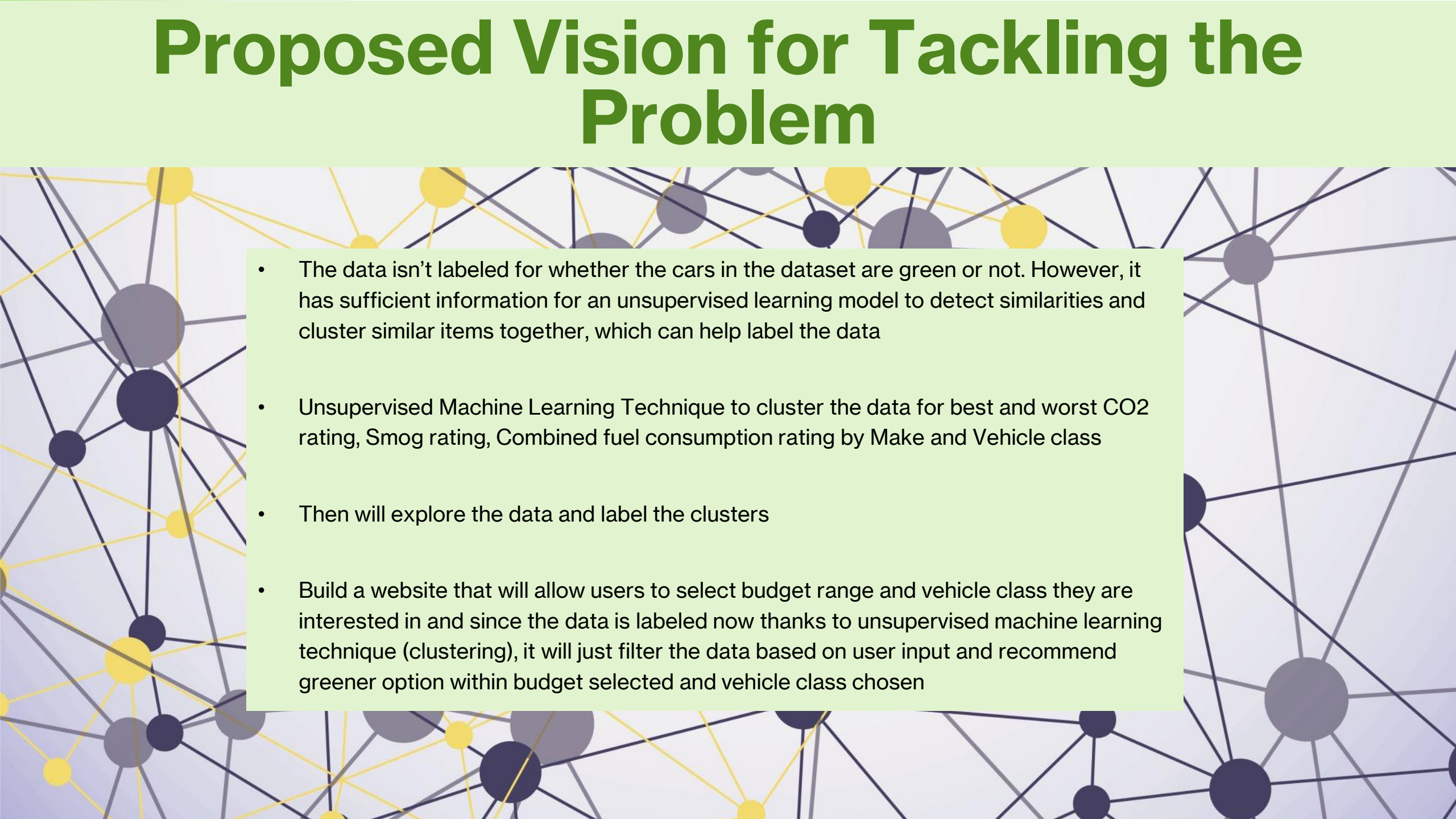
Budget and Environmentally Friendly Car Recommendation System

Non-Technical Overview & Potential Impact

- Not everyone can afford to buy and maintain a hybrid and electric vehicle despite wanting to make environmentally responsible choices
- Cost of living has increased substantially but salaries remain the same, last summer gas prices went up till 217 cents/Litre and this summer they went from 145 cents/Litre to 172 cents/Litre
- Helping users find a car that is both within their budget, is environmentally friendly and cost efficient in terms of fuel can help everyone make an environmentally responsible choice while staying within their budget



Proposed Vision for Tackling the Problem

- 
- The background of the slide features a complex network diagram. It consists of numerous circular nodes of varying sizes, colored in shades of yellow, grey, and dark blue. These nodes are interconnected by a web of thin, dark grey lines, creating a dense, interconnected pattern that fills the entire background.
- The data isn't labeled for whether the cars in the dataset are green or not. However, it has sufficient information for an unsupervised learning model to detect similarities and cluster similar items together, which can help label the data
 - Unsupervised Machine Learning Technique to cluster the data for best and worst CO2 rating, Smog rating, Combined fuel consumption rating by Make and Vehicle class
 - Then will explore the data and label the clusters
 - Build a website that will allow users to select budget range and vehicle class they are interested in and since the data is labeled now thanks to unsupervised machine learning technique (clustering), it will just filter the data based on user input and recommend greener option within budget selected and vehicle class chosen

Introduction to the Dataset

6,951 rows and 15 columns

Model_Year	Make	Model	Vehicle_Class	Engine_Size(L)	Cylinders	Transmi ssion	Fuel_Type	Fuel_Consumption-City(L/100 km)	Fuel_Consumption-Hwy(L/100 km)	Fuel_Consumption-Comb(L/100 km)	Fuel_Consumption-Comb(mpg)	CO2_Emissions(g/km)	CO2_Rating	Smog_Rating
Model_Year	Make	Model	Vehicle_Class	Engine_Size(L)	Cylinders	Transmi ssion	Fuel_Type	Fuel_Consumption-City(L/100 km)	Fuel_Consumption-Hwy(L/100 km)	Fuel_Consumption-Comb(L/100 km)	CO2_Emissions(g/km)	CO2_Rating	Smog_Rating	Price

- Prices data was brought in from other datasets, some of them included Price by Make, Model, Year and Mileage so I grouped by the 3 columns and took average then joined the grouped data with the main dataset.

- Found **data quality issues in the prices dataset where prices for cars were in 100s and the mileage wasn't even > 10,000. The cars in the dataset were from 2017-2020.**

		count	mean	std	min	25%	50%	75%	max	IQR	rank
Vehicle_Class	Make										
COMPACT	MITSUBISHI	25.0	7.680000	1.029563	5.0	7.0	8.0	8.0	9.0	1.0	29.0
FULL-SIZE	HONDA	61.0	7.032787	1.277592	5.0	6.0	7.0	8.0	10.0	2.0	24.0
MID-SIZE	HONDA	43.0	7.139535	1.319838	5.0	7.0	7.0	8.0	10.0	1.0	27.0
MINICOMPACT	FIAT	14.0	6.714286	0.611250	6.0	6.0	7.0	7.0	8.0	1.0	25.0

For each Vehicle Class: Top Make for CO2_Rating

		count	mean	std	min	25%	50%	75%	max	IQR	rank
Vehicle_Class	Make										
COMPACT	VOLKSWAGEN	64.0	6.687500	1.206793	3.0	7.00	7.0	7.00	8.0	0.00	28.00
FULL-SIZE	ACURA	3.0	6.666667	0.577350	6.0	6.50	7.0	7.00	7.0	0.50	26.50
MID-SIZE	JAGUAR	18.0	6.277778	1.994273	1.0	7.00	7.0	7.00	8.0	0.00	28.00
MINICOMPACT	MINI	36.0	6.611111	1.358103	3.0	7.00	7.0	7.00	8.0	0.00	28.00

For each Vehicle Class: Top Make for Smog_Rating

Introduction to the Dataset

	count	mean	std	min	25%	50%	75%	max	IQR	rank
Vehicle_Class										
MID-SIZE	939.0	9.619489	2.572342	4.4	7.8	9.2	11.0	20.0	3.2	24.4
COMPACT	713.0	9.190884	2.218288	4.5	7.6	8.7	10.3	17.4	2.7	24.5
STATION WAGON - SMALL	199.0	8.240201	1.411494	4.4	7.8	8.4	9.2	11.9	1.4	28.2
FULL-SIZE	548.0	11.156934	2.883345	4.0	9.1	11.5	12.9	17.1	3.8	29.8
SUBCOMPACT	619.0	10.635057	2.055811	6.8	9.1	10.4	12.1	16.7	3.0	30.0
STATION WAGON - MID-SIZE	61.0	11.226230	2.645304	5.8	9.6	10.2	12.7	16.6	3.1	30.3

Top 5 Vehicle Classes for Fuel Efficiency

Vehicle_Class	Make	count	mean	std	min	25%	50%	75%	max	IQR	rank
COMPACT	mitsubishi	25.0	6.852000	1.015759	6.0	6.20	6.50	6.600	9.3	0.400	24.200
FULL-SIZE	hyundai	47.0	7.068085	2.017220	4.0	4.70	7.90	8.700	9.6	4.000	13.200
MID-SIZE	toyota	88.0	7.190909	1.661784	4.4	5.10	7.50	8.400	9.7	3.300	15.300
MINICOMPACT	mercedes-benz	5.0	11.180000	3.425931	8.2	8.70	9.20	14.500	15.3	5.800	18.400
MINIVAN	toyota	14.0	9.300000	2.417882	6.5	6.70	10.85	11.525	11.7	4.825	20.625

For each Vehicle Class: Top Make for Fuel Efficiency

	count	mean	std	min	25%	50%	75%	max	IQR	rank
Make										
toyota	335.0	8.999104	2.930712	4.4	6.900	8.20	10.900	16.5	4.000	18.200
mitsubishi	66.0	8.325758	1.373968	6.0	6.525	8.90	9.300	10.6	2.775	22.525
NISSAN	227.0	9.712775	2.255311	6.7	8.000	8.70	11.400	15.5	3.400	23.200
KIA	189.0	8.640212	2.071558	4.4	7.500	8.60	10.200	13.3	2.700	24.100
HONDA	212.0	8.099528	1.667205	4.9	7.175	7.70	9.025	11.8	1.850	24.200
HYUNDAI	200.0	8.326000	1.883204	4.0	7.500	8.35	9.525	12.6	2.025	25.625

Top 5 Makes for Fuel Efficiency

Next Steps

Model_Year	Make	Model	Vehicle_Class	Engine_Size(L)	Cylinders	Transmi ssion	Fuel_Type	Fuel_Consumption-City(L/100 km)	Fuel_Consumption-Hwy(L/100 km)	Fuel_Consumption-Comb(L/100 km)	CO2_Emissions(g/km)	CO2_Rating	Smog_Rating	Price
------------	------	-------	---------------	----------------	-----------	------------------	-----------	---------------------------------	--------------------------------	---------------------------------	---------------------	------------	-------------	-------

Data Processing And Feature Engineering:

- One hot encoding for Make, Transmission, and Fuel Type
 - Determine any relationship between Transmission, Fuel with CO2 Emissions, Smog Ration and Combine Fuel Consumption
 - If there is no relevance for transmission then remove that column
 - One hot encoding can lead to dimensionality issues, so will need to handle that if it appears
 - # of Columns including Transmission: $40 + 26 + 5 + (15-5-3) = 78$
 - # of Columns excluding Transmission: $40 + 5 + (15-5-3) = 52$
- Remove the columns in yellow due to correlations heatmap results

Baseline Modeling:

- K-Means clustering algorithm
 - Simple to implement
 - Computationally efficient for moderately sized dataset
 - Easy interpretations as it assigns each data point to the nearest cluster center

