Sundus Yawar

Budget and Environmentally Friendly Car Recommendation System

Non-Technical Overview & Potential Impact

- Opens doors to making environmentally responsible decision within a tight budget given the increased cost of living

    - environmentally friendly i.e. fuel efficient, low CO2 emissions, best smog rating

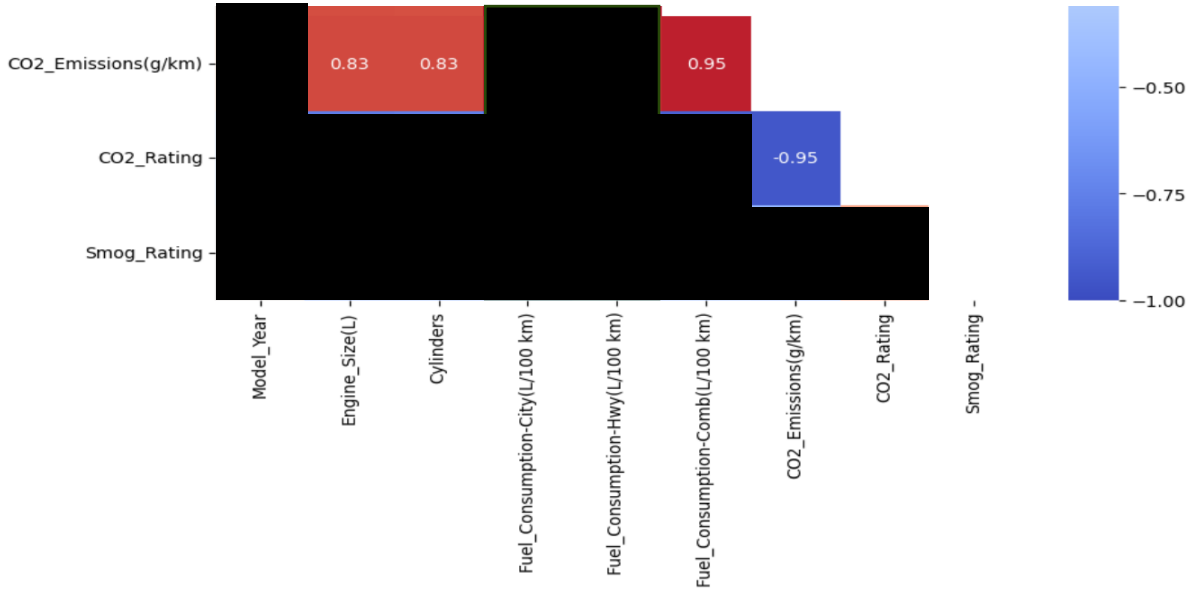
    - cost efficient i.e. within budget for purchase & fuel efficiency

Overview of the Dataset & Preprocessing Procedures

6,951 rows and 8 columns; 19 columns after 1 hot encoding

Model_Year	Make	Model	Vehicle_Class	Fuel_Type	Fuel_Consumption-Comb(L/100 km)	CO2_Emissions(g/km)	Smog_Rating

Vehicle_Class	Fuel_Consumption-Comb(L/100 km)	CO2_Emissions(g/km)	Smog_Rating



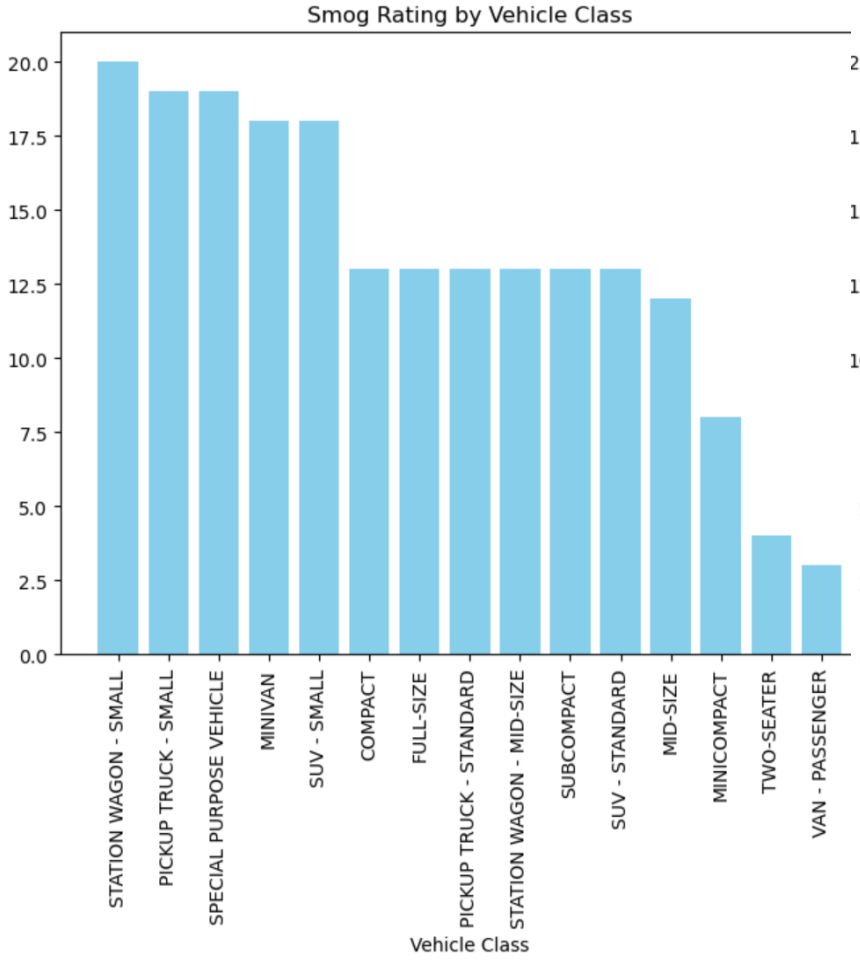
- Prices data was brought in from other datasets, some of them included Price by Make, Model, Year and Mileage so I grouped by the 3 columns and took average then joined the grouped data with the main dataset. 1 of them required currency conversion.

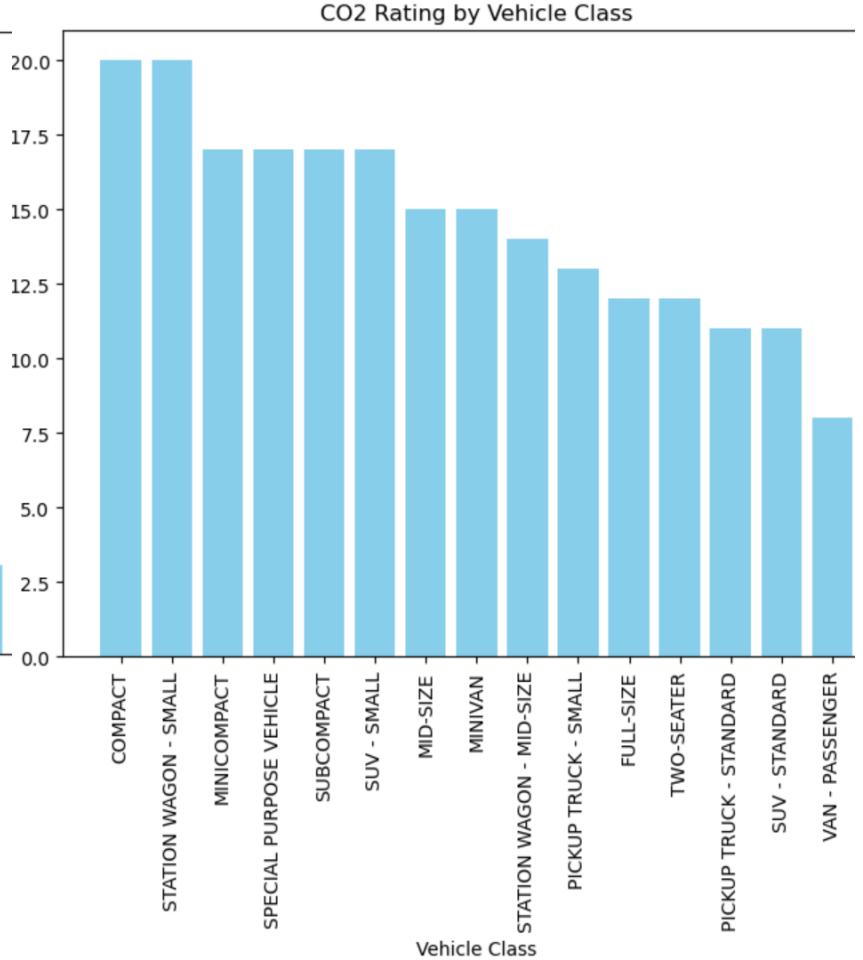
- Found data quality issues in the prices dataset where prices for cars were in 100s and the mileage wasn't even > 10,000. The cars in the dataset were from 2017-2020.

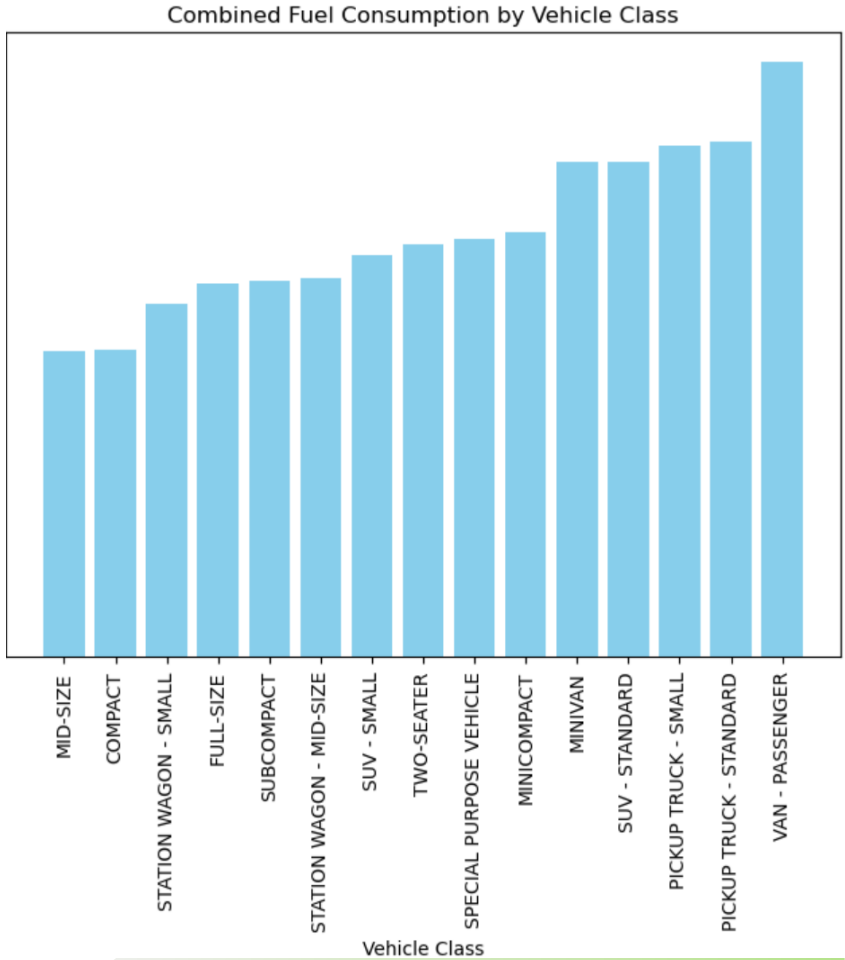
- Chi2 test (Model & Vehicle Class)

- Domain Knowledge & Correlation Matrix

- VIF Analysis for multicollinear columns



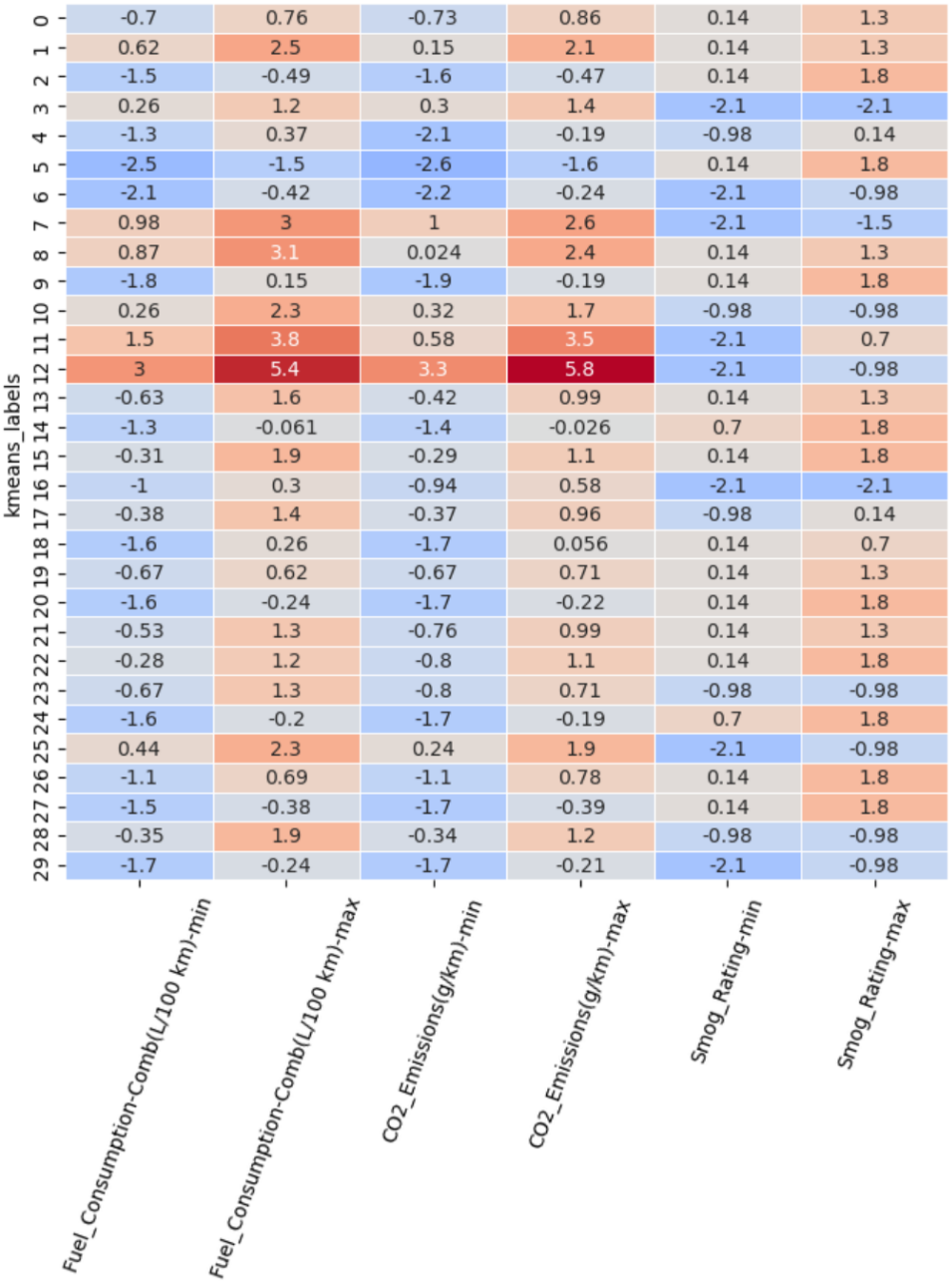


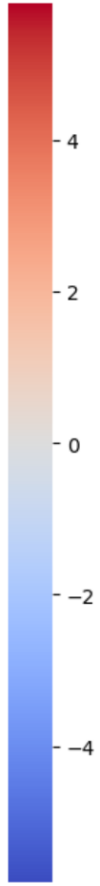


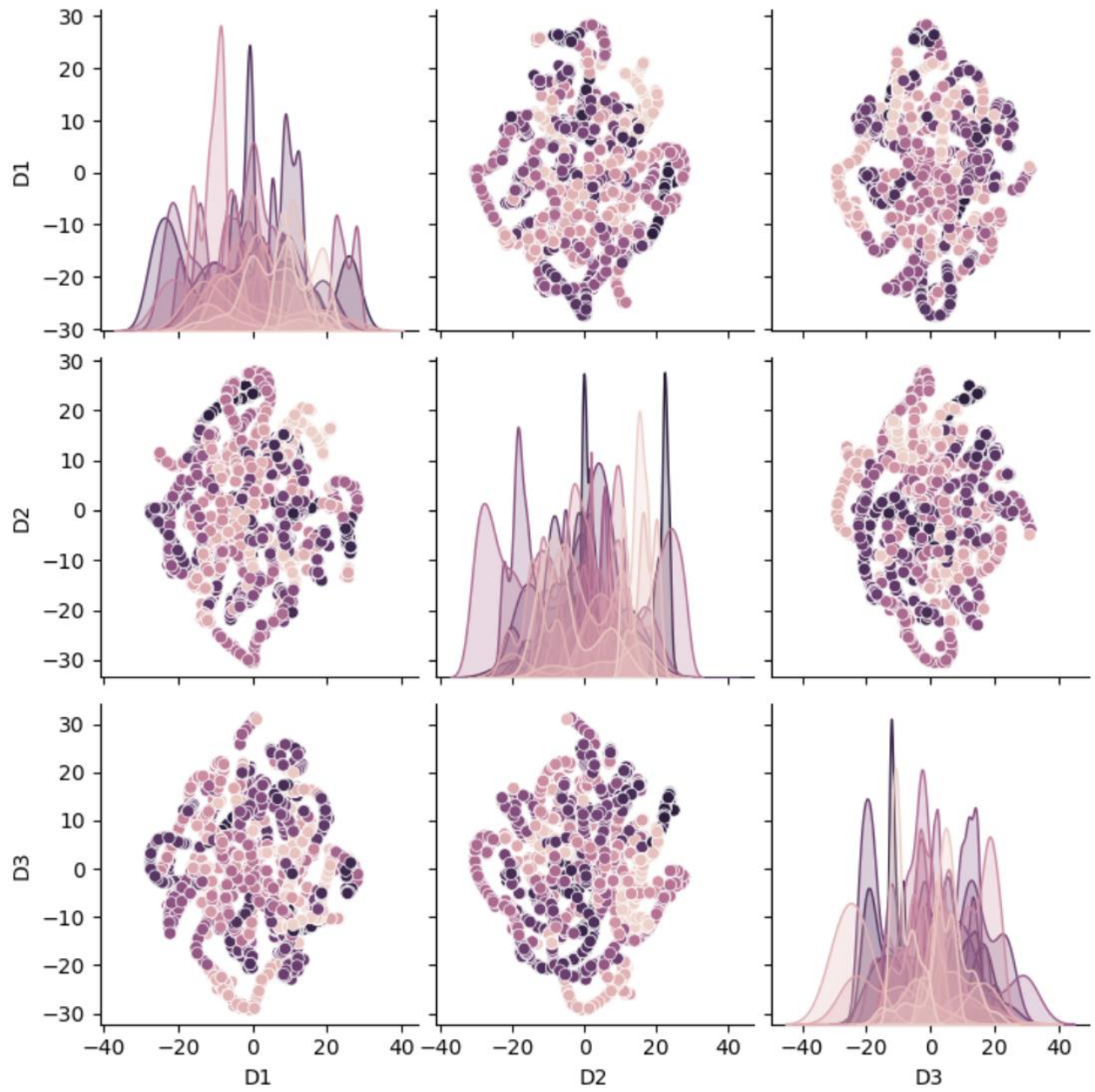
Few Important Findings from EDA

Baseline Model & Evaluation Metrics

kmeans_labels	Fuel_Consumption-Comb(L/100 km)-min	Fuel_Consumption-Comb(L/100 km)-max	CO2_Emissions(g/km)-min	CO2_Emissions(g/km)-max	Smog_Rating-min	Smog_Rating-max
0	-0.7	0.76	-0.73	0.86	0.14	1.3
1	0.62	2.5	0.15	2.1	0.14	1.3
2	-1.5	-0.49	-1.6	-0.47	0.14	1.8
3	0.26	1.2	0.3	1.4	-2.1	-2.1
4	-1.3	0.37	-2.1	-0.19	-0.98	0.14
5	-2.5	-1.5	-2.6	-1.6	0.14	1.8
6	-2.1	-0.42	-2.2	-0.24	-2.1	-0.98
7	0.98	3	1	2.6	-2.1	-1.5
8	0.87	3.1	0.024	2.4	0.14	1.3
9	-1.8	0.15	-1.9	-0.19	0.14	1.8
10	0.26	2.3	0.32	1.7	-0.98	-0.98
11	1.5	3.8	0.58	3.5	-2.1	0.7
12	3	5.4	3.3	5.8	-2.1	-0.98
13	-0.63	1.6	-0.42	0.99	0.14	1.3
14	-1.3	-0.061	-1.4	-0.026	0.7	1.8
15	-0.31	1.9	-0.29	1.1	0.14	1.8
16	-1	0.3	-0.94	0.58	-2.1	-2.1
17	-0.38	1.4	-0.37	0.96	-0.98	0.14
18	-1.6	0.26	-1.7	0.056	0.14	0.7
19	-0.67	0.62	-0.67	0.71	0.14	1.3
20	-1.6	-0.24	-1.7	-0.22	0.14	1.8
21	-0.53	1.3	-0.76	0.99	0.14	1.3
22	-0.28	1.2	-0.8	1.1	0.14	1.8
23	-0.67	1.3	-0.8	0.71	-0.98	-0.98
24	-1.6	-0.2	-1.7	-0.19	0.7	1.8
25	0.44	2.3	0.24	1.9	-2.1	-0.98
26	-1.1	0.69	-1.1	0.78	0.14	1.8
27	-1.5	-0.38	-1.7	-0.39	0.14	1.8
28	-0.35	1.9	-0.34	1.2	-0.98	-0.98
29	-1.7	-0.24	-1.7	-0.21	-2.1	-0.98







Next Steps

Model_Year	Make	Model	Vehicle_Class	Fuel_Type	Fuel_Consumption-Comb(L/100 km)	CO2_Emissions(g/km)	Smog_Rating

Data Processing And Feature Engineering:

- Try backward feature selection & see what # of clusters work best
- One hot encoding for Model_Year, Make and Fuel Type
  - One hot encoding can lead to dimensionality issues, so will need to handle that if it appears
- Handle the missing prices values

Modelling

- **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
- **Advanced Modelling:**
  - DBScan
- **Find better way for cluster visualization**
- **Look into the data and label clusters as green or not green**

Build the website and host it on heroku