Inventado, Charles Fredric G.
Rodelas, John Vincent B.
Valles, James Vincent V.

4CSE - Tres Sigmas - LAB ACT - MVDA

Part 1: Explore the Dataset & Form Questions

Multivariate Data Analysis based questions the Tres Sigmas group wanted to figure out.

1. Determine how vehicle characteristics such as Horsepower, Weight, and Acceleration could influence fuel efficiency such as Real-World MPG and CO2 emissions such as Real-World CO2 g/mi, and identify underlying dimensions that represent these combined impacts.
2. Find ways to possibly reduce the complexity of vehicle performance and environmental metrics into a smaller set of key indicators that are most relevant to understanding progress towards clean energy (SDG7) and climate action goals (SDG13).
3. Figure a combination of vehicle specifications and environmental data, differentiate these vehicles that align better with sustainable transportation objectives such as lower emissions, and higher MPG, and those that do not.

Part 2: Choose Analysis Techniques & Justify

MVDA Q1
Analysis Technique # 1 used : Factor Analysis (FA)
Explanation on the use of the specific analysis technique :
Explanation on the use of the specific analysis technique : Factor Analysis is suitable for Question 1 because the approach can help identify underlying latent factors that can explain the correlations of these vehicle specifications and their environmental impact. The technique can group related variables together into fewer factors, better explainability while reducing the complexity of understanding these relationships between variables.

Analysis Technique # 2 used (if applicable): Scatter Plot Matrix
Explanation on the use of the specific analysis technique :
A Scatter Plot Matrix is useful for Question 1 to visualize the pairwise relationships between two numerical variables related to vehicle specifications and environmental metrics. The approach helps in identifying potential linear or non-linear correlations just by observation before applying more complex techniques like Factor Analysis.

MVDA Q2
Analysis Technique # 1 used : Principal Components Analysis (PCA)
Explanation on the use of the specific analysis technique :

Regarding Q2, PCA is a dimensionality reduction technique by reducing the original set of identified correlated features into a smaller set of uncorrelated principal components. The dimensionality reduction allows us to retain most of the important information (with minimum variance) in the dataset with fewer variables, simplifying the further analysis and visualization.

Analysis Technique # 2 used (if applicable): Parallel Coordinates Plot
Explanation on the use of the specific analysis technique :
A Parallel Coordinates Plot helps to visualize the relationships between multiple features simultaneously such as using 3 or more variables. Parallel coordinates plot allows for the identification of patterns and clusters in the data after dimensionality reduction with PCA, helping to understand how different vehicles perform across the reduced set of components.

MVDA Q3
Analysis Technique # 1 used : Discriminant Analysis (DA)
Explanation on the use of the specific analysis technique :
Using Discriminant Analysis is appropriate for Question 3 to help classify and find a linear combination of features that best separates different vehicle types or regulatory classes. The approach helps in building a classification model to predict the regulatory class of a vehicle based on the specifications and environmental metrics.

Analysis Technique # 2 used (if applicable): Color-based Visualizations
Explanation on the use of the specific analysis technique :
Color-based visualizations can be used in combination with techniques such as Discriminant Analysis for Question 3 to visually represent the separation of different vehicle types or regulatory classes in a scatter plot of the discriminant functions. Using different colors for each class helps visually distinguish between the groups or datapoints.

Part 3: Conduct the Analysis & Build Visualizations in Python &

Part 4: Analyze Each Result & Visualization

MVDA Q1 Related

P-value & KMO value Preliminaries test to justify using Factor Analysis :

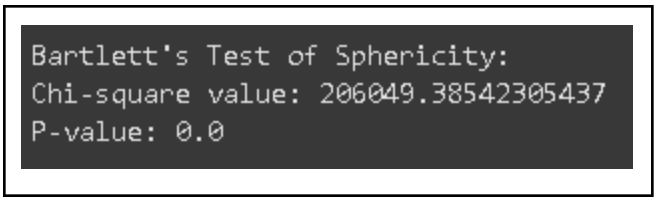

Figure 1.0 - Bartlett's Test Preliminary Test Result for Factor Analysis

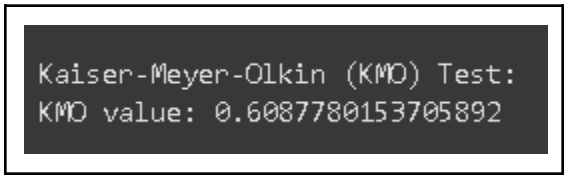

Figure 1.1 - Kaiser-Meyer-Olkin (KMO) Test Preliminary Test Result for Factor Analysis

Decision parameter for p-value :
If $p < 0.05$, variables are correlated enough to justify the use of Factor analysis. Can also mean the correlation matrix was significantly different from the identity matrix.

∴ Since p-value is $0 < 0.05$, or zero is less than 0.05, the variables are correlated enough to justify the use of Factor analysis.

Decision parameter for KMO value :
If KMO value > 0.6 is generally considered good for Factor analysis.

∴ Since the KMO value $0.61 > 0.6$, or KMO value is greater than 0.6, the data is considered suitable for Factor analysis.

Conclusion on preliminary for Factor analysis:
Both p-value and KMO values on the preliminary tests satisfy the conditions to justify the use of Factor analysis.

Visualizations :

Chart # 1 - Factor Analysis - To what extent do these original variables contribute to the defined latent factors?

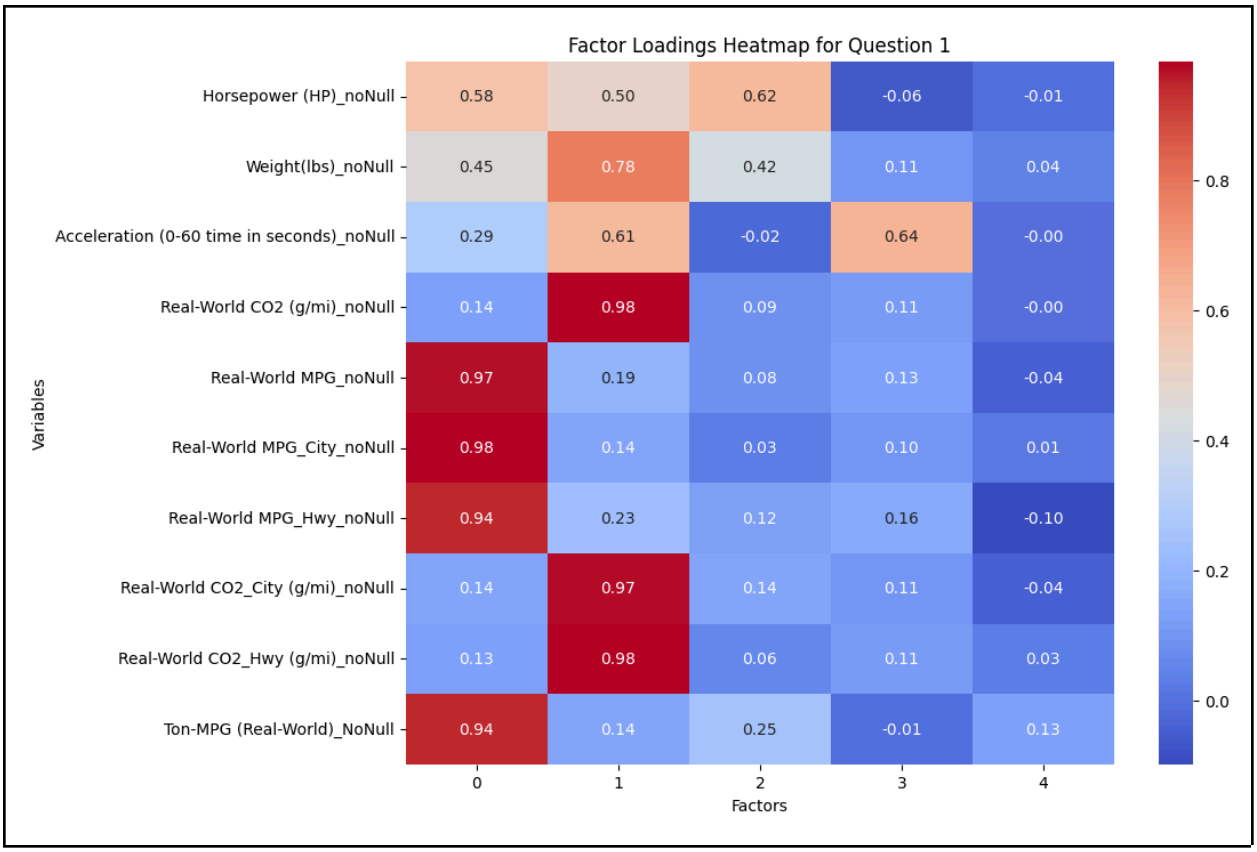

Figure 1.2 - Factor Loadings of Vehicle Performance and Environmental Metrics Heatmap

Analysis :

Figure 1.2 depicts the use of Bartlett's Test of Sphericity and the Kaiser-Meyer-Olkin (KMO) Test as preliminary checks for Factor analysis. Follows suit is the Factor Analysis itself and the heatmap visualization of the factor holdings. The Barlett's Test checks if the variables are uncorrelated, while the KMO Test measures the sampling adequacy. By using Factory Analysis we could identify the underlying factors that explain the correlation between variables, while the heatmap shows how each variable loads onto each factor strongly.

	0	1	2	3	4
Horsepower (HP)_noNull	0.579043	0.504455	0.615825	-0.060339	-0.010214
Weight(lbs)_noNull	0.454257	0.776072	0.416652	0.108550	0.037874
Acceleration (0-60 time in seconds)_noNull	0.294175	0.612073	-0.024565	0.636099	-0.002377
Real-World CO2 (g/mi)_noNull	0.135556	0.979660	0.087410	0.111536	-0.004434
Real-World MPG_noNull	0.968757	0.186499	0.084781	0.126945	-0.041155
Real-World MPG_City_noNull	0.982573	0.140964	0.029094	0.096949	0.014820
Real-World MPG_Hwy_noNull	0.944562	0.232610	0.122086	0.158857	-0.097424
Real-World CO2_City (g/mi)_noNull	0.141562	0.970814	0.143917	0.106357	-0.041038
Real-World CO2_Hwy (g/mi)_noNull	0.131961	0.980659	0.060128	0.113407	0.026634
Ton-MPG (Real-World)_NoNull	0.943938	0.142708	0.245375	-0.011547	0.131023

Variance Explained by Factors:

	Variance	Proportional Variance	Cumulative Variance
0	4.371166	0.437117	0.437117
1	4.224427	0.422443	0.859559
2	0.668559	0.066856	0.926415
3	0.507542	0.050754	0.977169
4	0.032529	0.003253	0.980422

Figure 1.0 - Factor Loadings and Variance Explained by Principal Components

1. **Main takeaway in one sentence**

   The preliminary tests depict that Factor Analysis can be performed based on the data at hand, and the factor loading heatmap shows how vehicle characteristics (HP, Weight, Acceleration) and environmental metrics (Real-World CO2, Real-World MPG, etc.) group together into underlying factors.

2. **One design or analysis decision and its benefit**

   We noticed that the use of Barlett's Test and the KMO test before Factor Analysis provides statistical evidence for the suitability of the data for this particular technique while ensuring that the identified factors are well-defined and meaningful.

Additional Insights :

The interpretation of meaning of each factor by showing which variables load highly on which factor can be depicted well using a factor loadings heatmap. An

example would be a factor with high loadings on MPG variables and low loadings on CO2 variables that might represent ‘Fuel Efficiency’, while a factor with high loadings on ‘Horsepower’ and ‘Weight’ might represent ‘Vehicle Power/Size’. The variance explained by each factor indicates the proportion of the total variance in the original variables accounted for by each factor that helps in deciding how many factors to retain.

The interpretation of factors based on loadings can be subjective and different rotation methods might lead to different factor structures.

Visualizations :

Chart # 2- Scatter Plot - Finding correlations on Vehicle Features and Environmental variables

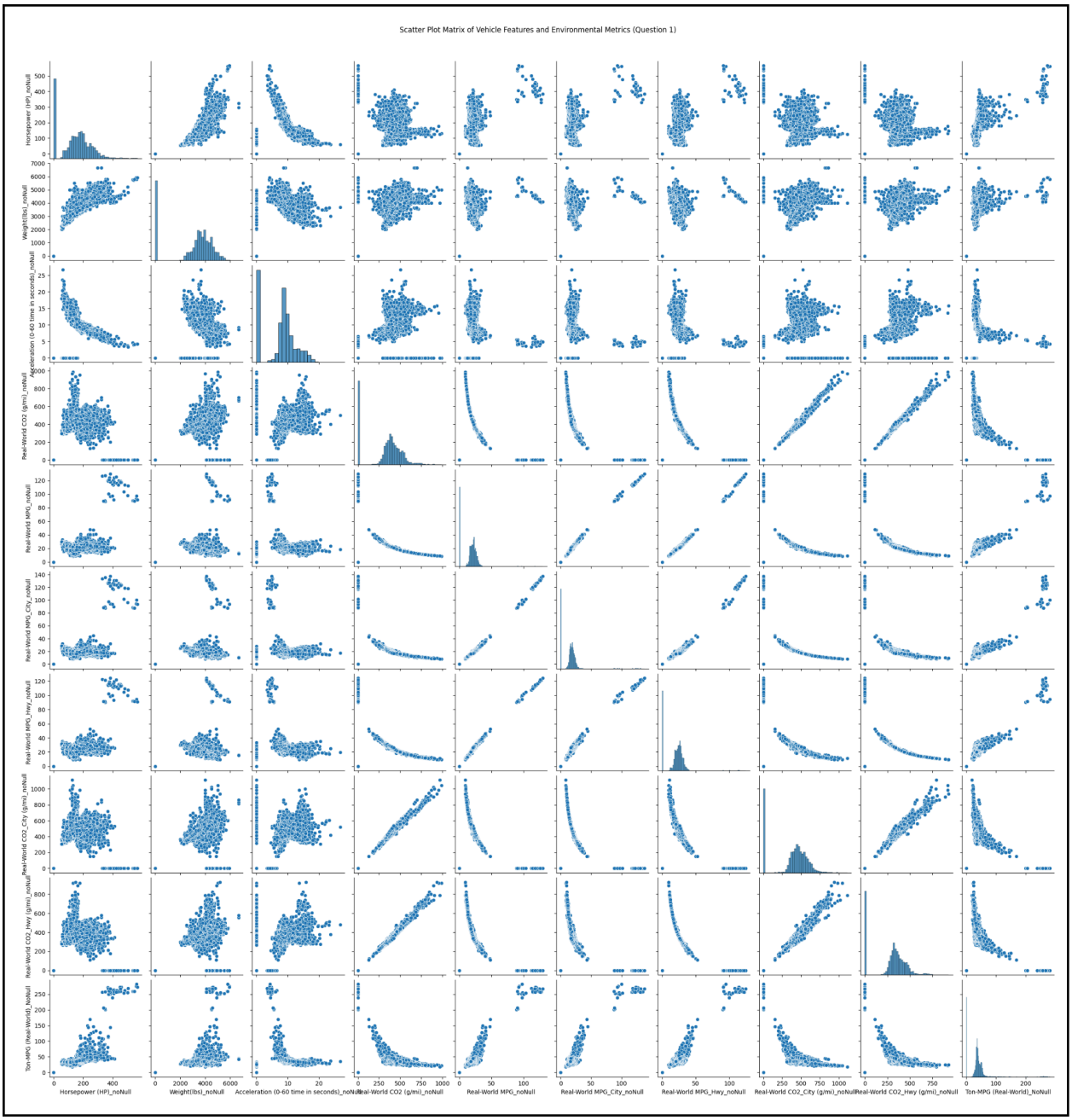

Figure 2.0 - Pairwise Scatter Plot and Univariate Histograms Matrix of Vehicle Features and Environmental Metrics

Analysis :

In Figure 2.0, we can see that it involves a scatter plot matrix that displays the pairwise relationships between the selected vehicle features and environmental metrics for

the first question. The figure depicts not only scatter plots for every combination of two variables but also histograms for each individual variable along the diagonal.

1. **Main takeaway in one sentence**

   The scatter plot matrix above highlights the visual overview of the pairwise relationships and distributions of the relevant vehicle features and environmental metrics while highlighting the potential correlations.

2. **One design or analysis decision and its benefit**

   Generating the Scatter Plot Matrix as a preliminary step allows us to quickly identify visually the potential linear or non-linear relationships between variables before applying more complex multivariate techniques like Factor Analysis.

Additional Insights :

We could only show the relationship between two variables on 1 graph , so from all of these graphs present in the matrix from both pairwise scatter plot and univariate histogram, we have a limitation that we could not see and observe the interaction occurring if we want three or more variables interact at the same time.

Visual-wise from the size of the scatter plot and number of variables is considered to be cluttered thus there are difficulties in briefly interpreting the scatter plot matrix results.

MVDA Q2 Related

P-value & KMO value Preliminaries test to justify using Principal Component Analysis (PCA) :

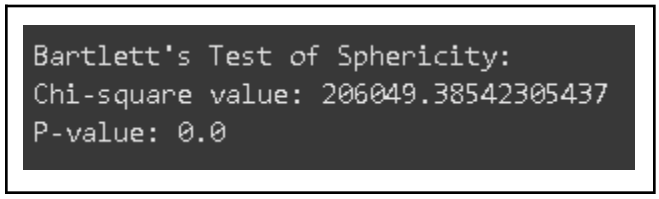

Figure 3.0 - - Bartlett's Test Preliminary Test Result for PCA

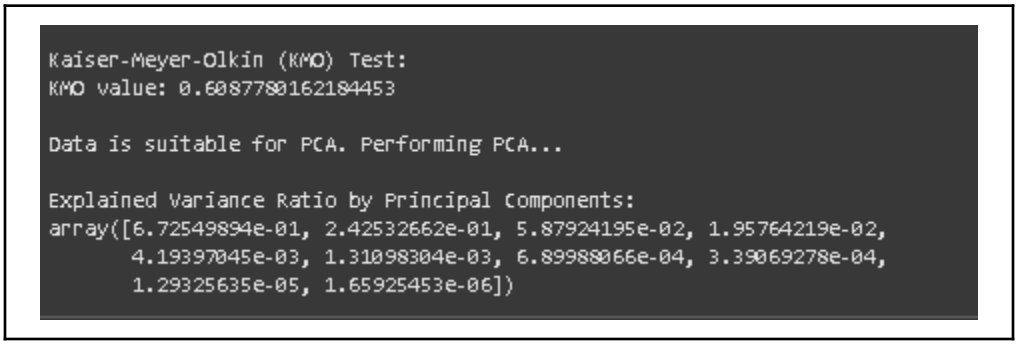

Figure 3.1 - Kaiser-Meyer-Olkin (KMO) Test Preliminary Test Result for PCA

Decision parameter for p-value :

If $p < 0.05$, variables are correlated enough to justify the use of PCA. Can also mean the correlation matrix was significantly different from the identity matrix.

∴ Since p-value is $0 < 0.05$, or zero is less than 0.05, the variables are correlated enough to justify the use of PCA.

Decision parameter for KMO value :

If KMO value > 0.6 is generally considered good for Principal Component Analysis (PCA).

∴ Since the KMO value $0.61 > 0.6$, or KMO value is greater than 0.6, the data is considered suitable for Principal Component Analysis (PCA).

Conclusion on preliminary for PCA :

Both p-value and KMO values on the preliminary tests satisfy the conditions to justify the use of Principal Component Analysis (PCA).

Visualizations :

Chart # 3 - Principal Component Analysis (PCA) - 4 principal components explain virtually all (nearly 100%) of the data's total variance.

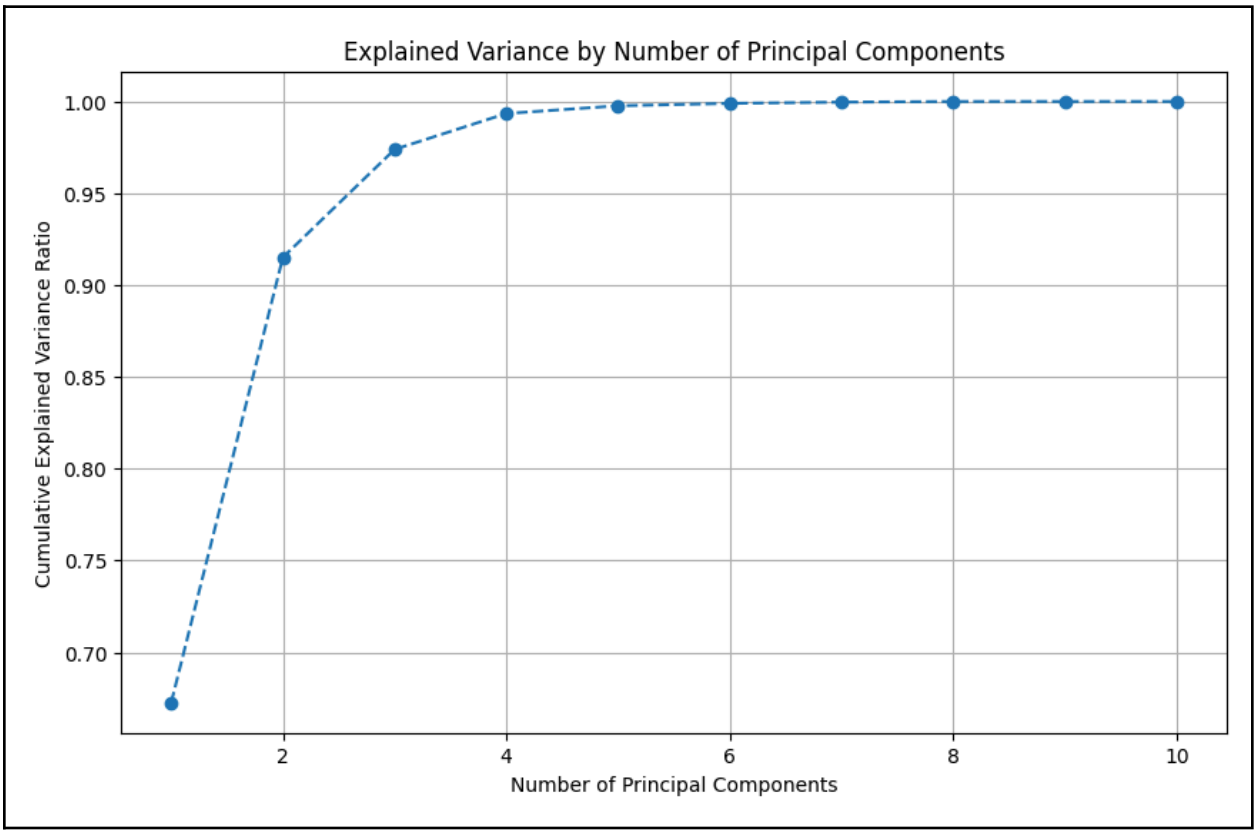

Figure 3.2 - Scree Plot Showing Cumulative Explained Variance by Number of Principal Components

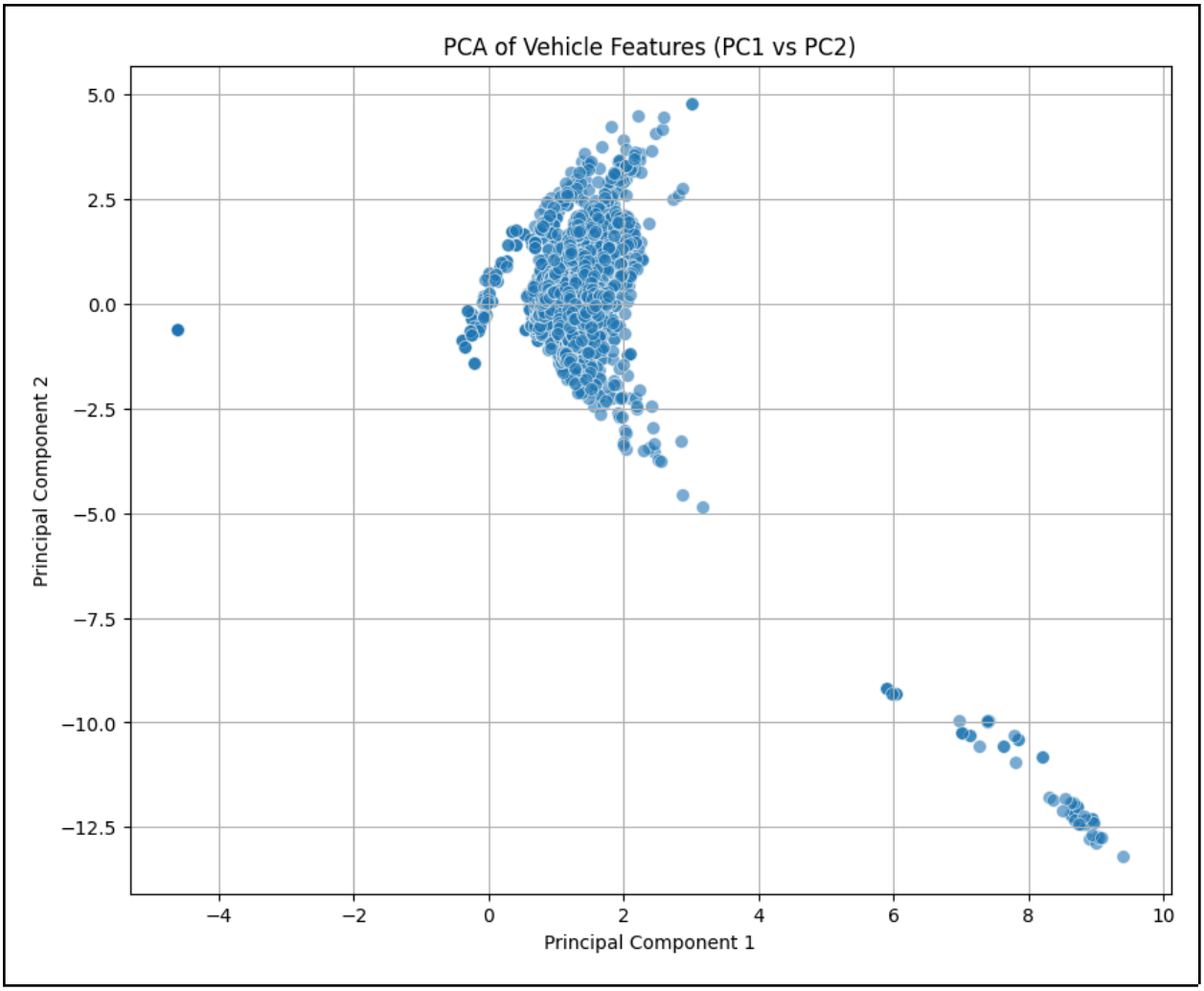

Figure 3.3 - Scatter Plot of Principal Component 1 (PC1) versus Principal Component 2 (PC2) from the PCA of Vehicle Features

Analysis :

The figures above depict Principal Component Analysis (PCA) on the selected vehicle features to reduce their dimensionality. We can see in the first chart the cumulative explained variance by the number of principal components that help determine the number of components needed to capture a sufficient amount of the total variance. As for the second chart, it involves a scatter plot of the first two principal components, which includes PC1 and PC2, that visualizes the data in a reduced 2D space.

1. **Main takeaway in one sentence**

   The first few principal components take into account a significant part of the variance in the features of vehicles, allowing for dimensionality reduction while keeping the important information.

2. **One design or analysis decision and its benefit**

We discovered that using the cumulative explained variance plot helped us in deciding on a reasonable number of principal components to retain, which then reduces the complexity of the dataset all the while preserving most of the variability at the same time.

Additional Insights :

The scatter plots on the figures prior (PC1 vs PC2) provides a visual representation of how vehicles are distributed in the reduced dimensional space. By observing the clusters or patterns in the plots can give us insights into the relationships between vehicles based on the combined influence of the original features. The original variables' loadings on the principal components, that of which can be obtained from the PCA object, would provide further insight into what each component represents in terms of the original features.

Chart # 4 - Parallel Coordinates Plot - Regular Classes Traversing Through Principal Components

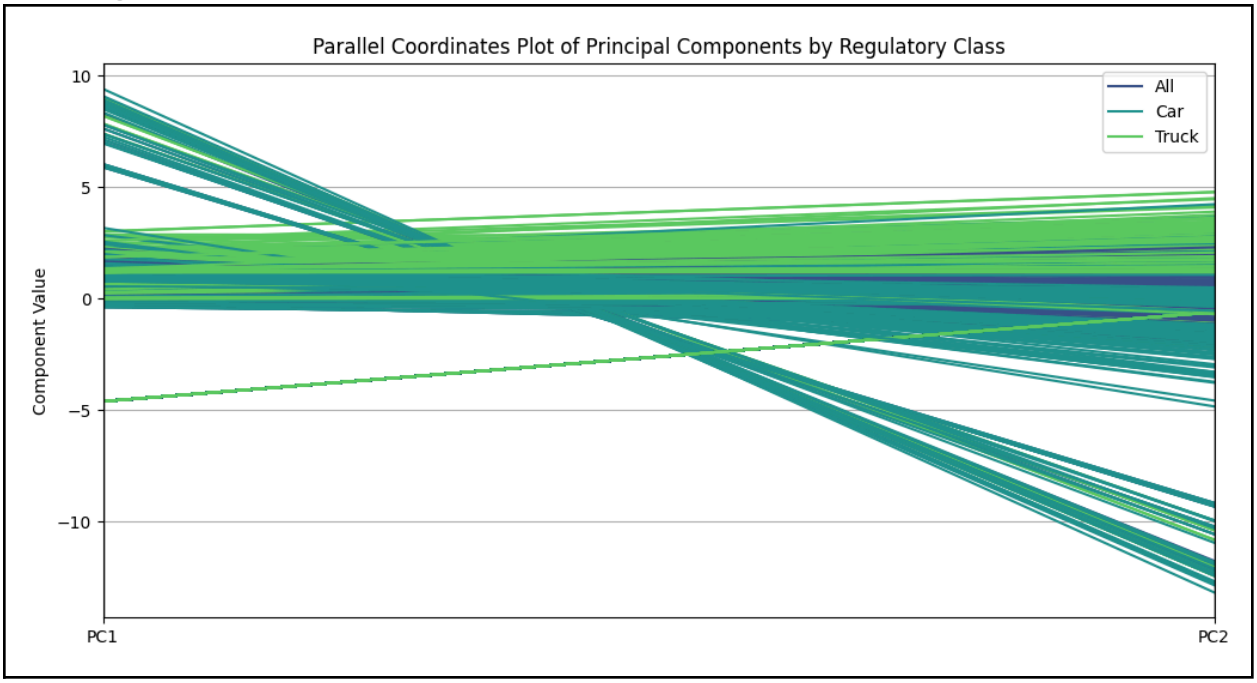

Figure 4.0 - Parallel Coordinates Plot of Principal Components by Regulatory Class

Chart # 4.5 - Parallel Coordinates Plot - Regular Classes Traversing Through Original Relevant Vehicle Features

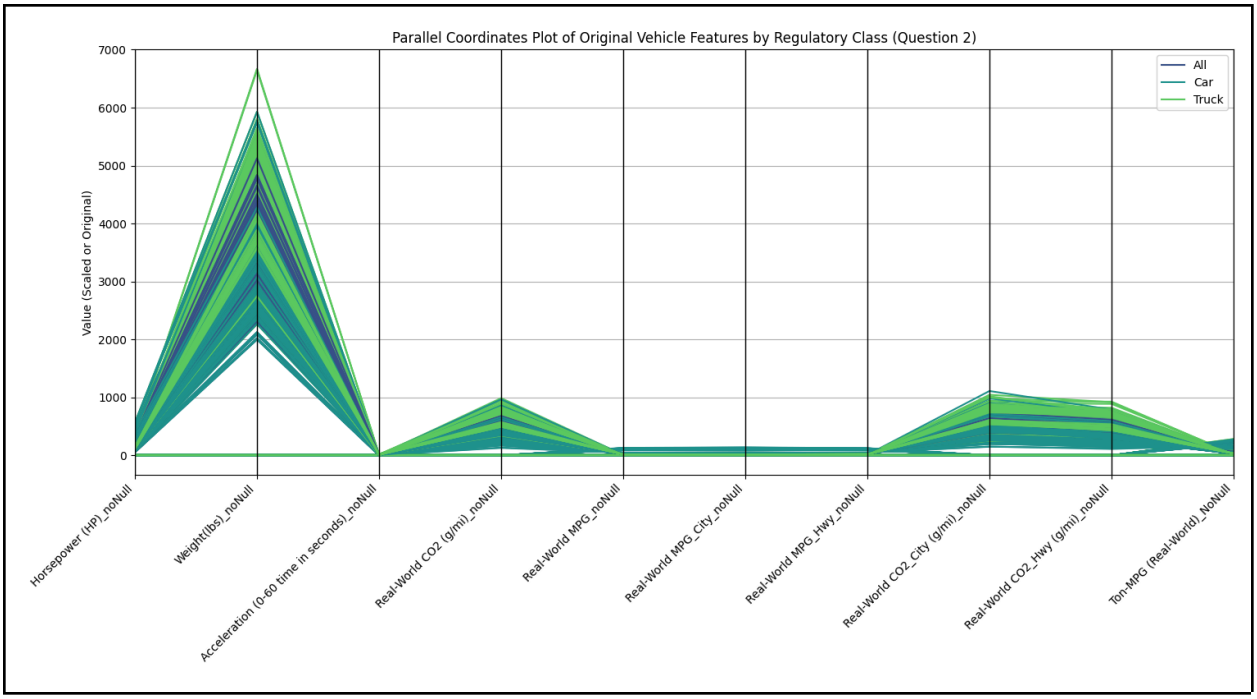

Figure 4.1 - Parallel Coordinates Plot of Original Relevant Vehicle Features by Regulatory Class

Analysis :

In Figure 4,1, we can see a Parallel Coordinates Plot that visualizes the principal components for different vehicles that are colored by 'Regulatory Class.' Each line that we see represents a vehicle, while the line traverses the vertical axes that represent the principal components, such as PC1, PC2, and potentially more if kept. Such plot helps to observe patterns and relationships across multiple dimensions simultaneously in the reduced PCA space.

3. **Main takeaway in one sentence**

   The Parallel Coordinates Plot allows for a visual inspection of how different vehicle regulatory classes are distributed across the principal components all the while providing insights into their multivariate characteristics.

4. **One design or analysis decision and its benefit**

   We found out that coloring the lines by 'Regulatory Class' helps to visually differentiate between the groups and observe if there are any distinct patterns or clusters for the variables 'All', 'Car', and 'Truck' vehicles in the reduced dimensional space

Additional Insights :

The visual distortion caused by the extreme difference in variable scales, leading to significant overplotting and a loss of detail. The axes at Figure 4.1 are somewhat visually distinct are X-axis Weight(lbs), Real World CO2(g/mi), Real World CO2_City,

and Real World CO2_Hwy while the rest suffers the aforementioned overplotting. It becomes nearly impossible to discern variation, patterns, or clusters within these critical efficiency metrics, severely limiting the plot's ability to compare efficiency nuancedly across the regulatory classes. The same goes for Figure 4.0 because there are many lines representing many observations.

MVDA Q3 Related :

Visualizations :

Chart # 5 & Chart # 6 - Linear Discriminant Analysis (LDA) - Regulatory Class Separated by Vehicle Performance

```
Explained Variance Ratio by Linear Discriminant Components:
array([0.98185535, 0.01814465])
```

Figure 5.0 - Variance Ratio for the Linear Discriminant Charts

What the sigma is the Variance

Figure 5.1 - Linear Discriminant Analysis Chart on Vehicle Performance by Regulatory Class

Chart # 6.5 - Linear Discriminant Analysis (LDA) - Regulatory Class Separated by Vehicle Performance with Approximated Decision Boundaries

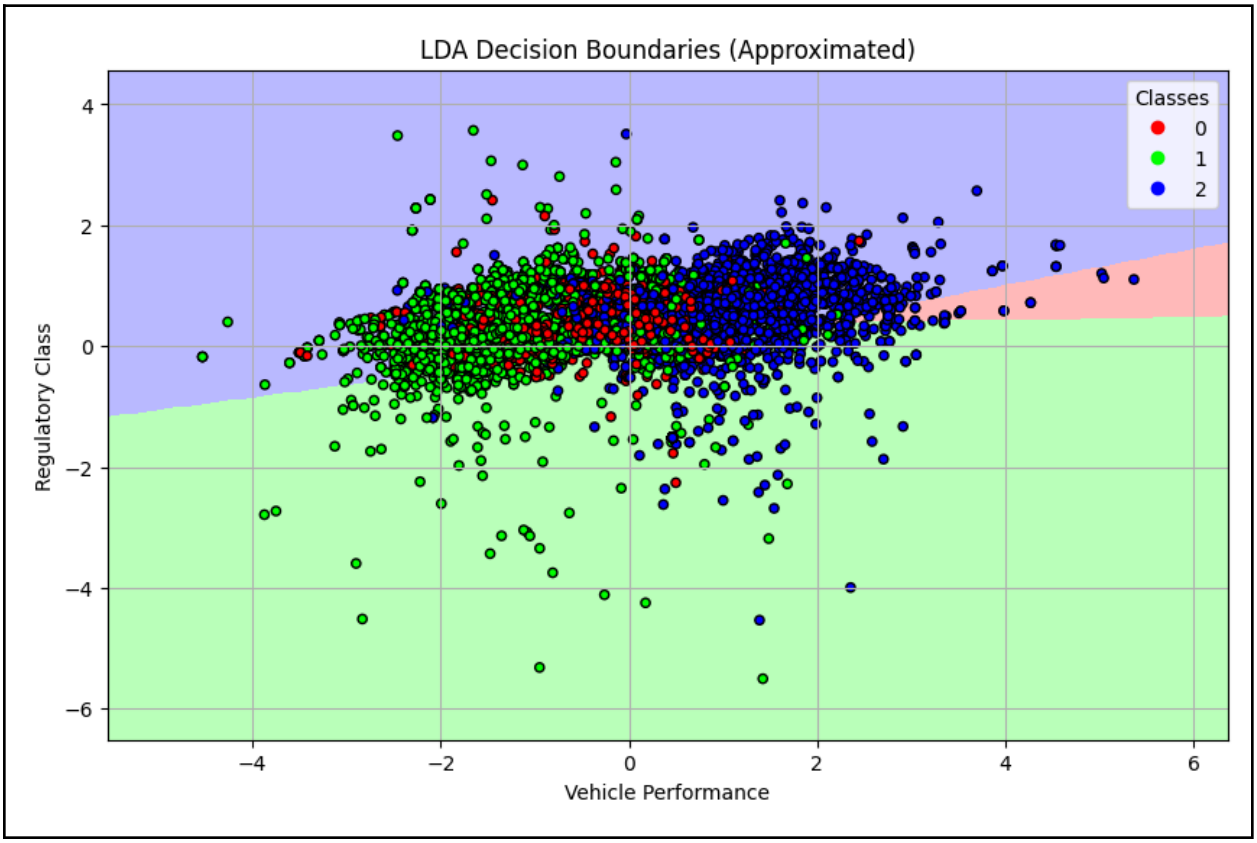

Figure 5.2 - Linear Discriminant Analysis Chart on Vehicle Performance by Regulatory Class with Approximated Boundaries

Analysis :

In Figure 5.1 and 5.2, we can see that it involves a linear discriminant analysis chart that displays a separation between variables in terms of regulatory class. The figure 5.2 shows a more clearer presentation of how the variables are separated by coloring the boundaries and the area they fill which shows “Trucks” are on the top, “All” are in the middle, and “Car” cover most of the chart at the bottom. Vehicle performance show that the data points of “Truck” sets at the right, “All” sets in the middle, and “Car” sets at the left

1. **Main takeaway in one sentence**

   Cars are lighter, more fuel-efficient, and lower power, while Trucks are heavier, less efficient, and higher power

2. **One design or analysis decision and its benefit**

   Data points are colored to their respective regulatory class to better distinguish the separation.

Additional Insights :

You really wouldn't be able to know what the vehicle performance would be able to tell since it combines 10 numerical features of a vehicle such as Horsepower, Weight, Acceleration, MPG, CO2 emissions, and etc. You would initially think that the graph shows that cars are worse than trucks which isn't what it is supposed to entail because it shows that car trade off some features in order to serve how cars work and that trucks are supposed to be better in gas and power since they usually go long treks and carry heavy cargo.