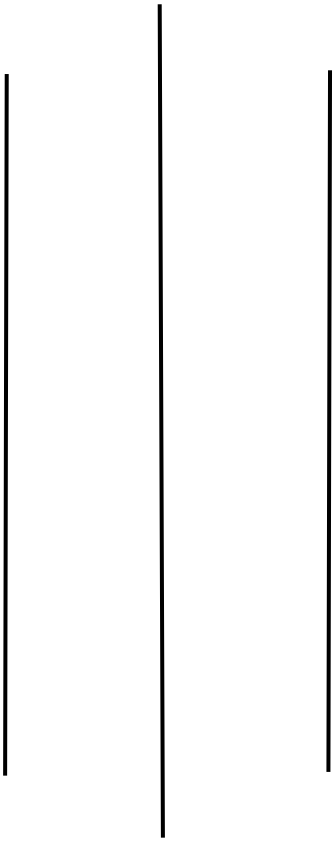

PCA

PRINCIPAL COMPONENT ANALYSIS PROJECT

FALL 2023

By:

ROBIN SAH

G48297336

EMSE 6765

DECEMBER 14, 2023

Table of Contents:

1. Introduction

Principal component analysis, commonly known as PCA, is a statistical technique used to simplify the complexity in high-dimensional data. It does this by condensing the data's dimensions, meaning it takes a dataset with many variables and distils it into a few key components.

The goal of PCA is to cut down the clutter. While it's true that in reducing dimensions some intricate details might be lost, the advantage gained is a more manageable set of key features. This trimmed set is far more practical to work with, particularly when it comes to visualizing data patterns or speeding up computational tasks like machine learning, all while retaining the essence of the original dataset.

In essence, PCA is about finding a balance - minimizing the number of variables while keeping the dataset's core information intact.

2. Dataset

The data is provided a survey data of 10 car brands in terms of 6 attributes. The higher score in a particular attribute, the better the car in terms of that attribute.

Figure 1. provides the dataset of the car survey in terms of 6 attributes. The ten car brands in the dataset are:

- BMW
- Chrysler
- Ford
- Mercedes
- Infinity
- Saab
- Jeep
- Porsche
- Lexus
- Volvo

<u>The 6 attributes in the dataset are:</u>

- Luxurious
- For Family
- Safe
- Practical
- Sporty
- Exciting

	Luxurious	Safe	Sporty	For Family	Practical	Exciting
BMW	4	3	5	2	2	4
Ford	2	3	2	4	5	2
Infinity	4	3	3	3	3	2
Jeep	3	3	2	4	4	3
Lexus	5	4	3	3	3	3
Chrysler	1	3	1	5	5	1
Mercedes	5	4	3	3	2	2
Saab	3	4	4	3	3	4
Porsche	4	2	5	1	1	5
Volvo	2	5	1	5	4	1

Fig 1. Survey data on 10 cars in terms of 6 attributes

3. Principal Component Analysis (PCA)

3.1 Correlation Analysis

Figure 2 presents a correlation matrix. The highlighted cells indicate particularly strong correlations, surpassing the chosen threshold of 0.6, suggesting that these attributes can influence each other.

The goal of PCA in this context is to synthesize this information, reducing the six attributes to a smaller number of principal components. These components will be new, underlying factors that explain the most variance in the data. Essentially, PCA will find the best way to summarize the dataset with fewer variables, losing as little of the original information as possible.

	Luxurious	*Safe*	*Sporty*	*For Family*	*Practical*	*Exciting*
Luxurious	1					
Safe	-0.0197028	1				
Sporty	0.6477905	-0.41825	1			
For Family	-0.7234425	0.505291	-0.96174	1		
Practical	-0.7950516	0.220176	-0.86192	0.90351	1	
Exciting	0.4906832	-0.47287	0.900028	-0.86946	-0.71933	1
	Threshold		0.6			

Fig 2. Correlation Matrix

3.2 Minitab Outputs for Eigenvalues, Coefficients and Scores

Figure 3 contains a table of eigenvalues and the corresponding eigenvectors (labelled U1 through U6). The eigenvalues represent the amount of variance captured by each principal component. Typically, a higher eigenvalue indicates that the component accounts for a larger amount of the total variance in the dataset.

Eigenvalue	U1	U2	U3	U4	U5	U6
4.3425536	-0.3620262	-0.5068949	-0.5609609	-0.5272533	0.00697324	0.13882677
1.07672686	0.22177344	-0.7987124	0.52320582	0.01040516	-0.0807493	-0.1802833
0.36065703	-0.4639205	0.03232421	0.28341337	0.09051488	-0.7125566	0.43298954
0.14560967	0.47598941	-0.0464761	0.05245977	-0.1129296	0.20750822	0.8442211
0.06288833	0.44115775	0.25547845	0.07019055	-0.6863015	-0.4662847	-0.2162239
0.0115645	-0.4281884	0.1914211	0.56883797	-0.4795206	0.47456608	0.03582014

Fig 3. Eigenvalues and Coefficients (MINITAB Output)

Figure 4 shows the scores (Z1 through Z6) for each observation (in this case, car brands) on the principal components. The scores essentially represent the projection of the original data onto the new axes formed by the principal components.

		1	2	3	4	5	6
	State	Z1	Z2	Z3	Z4	Z5	Z6
1	BMW	-2.2796	0.1619	0.3034	0.127	-0.3199	0.1407
2	Ford	1.6283	1.0746	-0.0514	-0.2992	-0.2958	-0.1606
3	Infinity	-0.2837	-0.0121	-0.8432	0.1077	-0.2345	-0.0002
4	Jeep	0.7024	0.6447	-0.0988	-0.5306	0.4184	0.1342
5	Lexus	-0.6115	-1.1951	-0.2168	-0.6327	0.0298	-0.0834
6	Chrysler	2.9195	1.251	-0.2109	0.3009	0.0017	0.0845
7	Mercedes	-0.6265	-1.5322	-0.6954	0.2471	0.0291	0.054
8	Saab	-0.7104	-0.2717	1.2429	-0.1404	-0.1175	0.0346
9	Porsche	-3.5781	1.0952	0.013	0.3677	0.3191	-0.129
10	Volvo	2.8397	-1.2163	0.5572	0.4526	0.1696	-0.075

Fig 4. Scores (MINITAB Output)

Figure 5 shows that the first principal component (Z1) accounts for the majority of the variance (72.4%), followed by the second principal component (Z2) with an additional 17.9%, and so on. Similarly, the first two principal components in the Minitab output have eigenvalues of 4.3 and 1.1, which are above the Kaiser criterion, indicating that they are significant. These two components represent the attributes 'Luxurious' and 'Safe'. They cumulatively account for 90.3% of the total variance.

	Raw GSP	Z1	Z2	Z3	Z4	Z5	Z6	Total
	Eigenvalues	4.343	1.077	0.361	0.146	0.063	0.012	6.0
	Percentage	72.4%	17.9%	6.0%	2.4%	1.0%	0.2%	
	Cumulative	72.4%	90.3%	96.3%	98.8%	99.8%	100.0%	
	Eigen Vectors	U1	U2	U3	U4	U5	U6	
X1	Luxurious	-0.362	-0.507	-0.561	-0.527	0.007	0.1388	
X2	Safe	0.2218	-0.799	0.5232	0.0104	-0.081	-0.18	
X3	Sporty	-0.464	0.0323	0.2834	0.0905	-0.713	0.433	
X4	For Family	0.476	-0.046	0.0525	-0.113	0.2075	0.8442	
X5	Practical	0.4412	0.2555	0.0702	-0.686	-0.466	-0.216	
X6	Exciting	-0.428	0.1914	0.5688	-0.48	0.4746	0.0358	

Fig 5. PCA Eigenvalues and their contribution

In the scree plot shown in figure 6, the first two components have significantly higher eigenvalues than the subsequent components, with a notable drop after the second component. The 'elbow' suggests that the first two components are sufficient in capturing the majority of the variance in the data. The number of components to be retained would be two. The eigenvalues for components beyond the second one fall close to zero, implying they contribute less to the explanation of the data's variance.

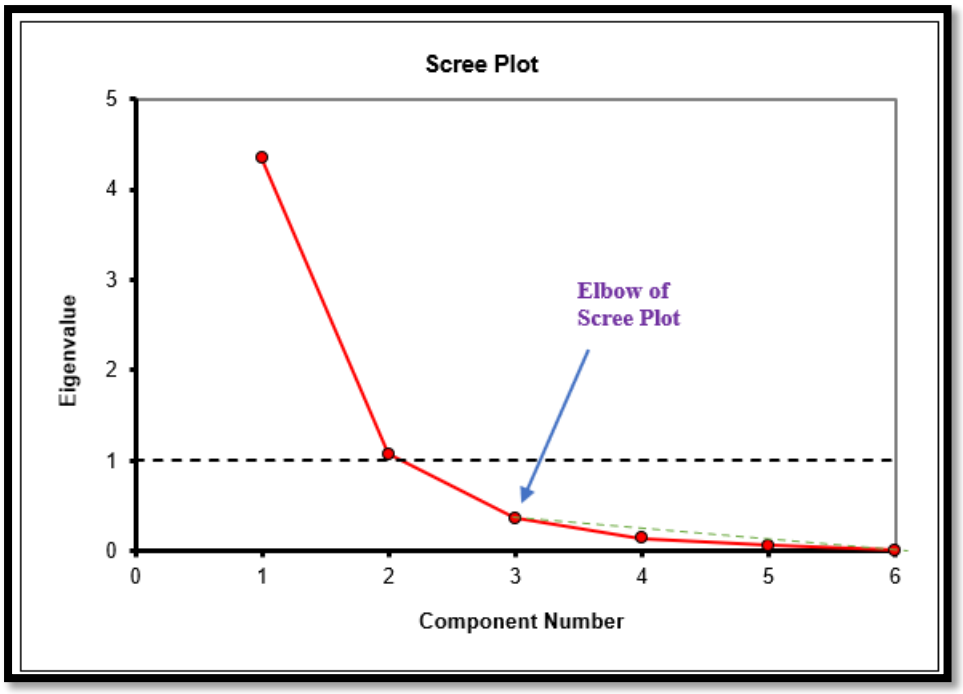

Fig 6. Scree plot

The (Loadings)^2 shown in the figure 7 represent the proportion of the variance in each original variable that is explained by each principal component. For instance, the first principal component (Z1) explains 56.9% of the variance in the variable 'Luxurious' (X1). The second principal component (Z2) explains an additional 27.7% of its variance, and so on.

	Correlations between original variables Xi and Principal Components Zi						
		1	2	3	4	5	6
	Loadings	Z1	Z2	Z3	Z4	Z5	Z6
X1	Luxurious	-0.754	-0.526	-0.337	-0.201	0.002	0.015
X2	Safe	0.462	-0.829	0.314	0.004	-0.020	-0.019
X3	Sporty	-0.967	0.034	0.170	0.035	-0.179	0.047
X4	For Family	0.992	-0.048	0.032	-0.043	0.052	0.091
X5	Practical	0.919	0.265	0.042	-0.262	-0.117	-0.023
X6	Exciting	-0.892	0.199	0.342	-0.183	0.119	0.004
	Explained Variance in Individual variables						
	(Loadings)^2	Z1	Z2	Z3	Z4	Z5	Z6
X1	Luxurious	56.9%	27.7%	11.3%	4.0%	0.0%	0.0%
X2	Safe	21.4%	68.7%	9.9%	0.0%	0.0%	0.0%
X3	Sporty	93.5%	0.1%	2.9%	0.1%	3.2%	0.2%
X4	For Family	98.4%	0.2%	0.1%	0.2%	0.3%	0.8%
X5	Practical	84.5%	7.0%	0.2%	6.9%	1.4%	0.1%
X6	Exciting	79.6%	3.9%	11.7%	3.3%	1.4%	0.0%

Fig 7. Explained Variance in each variable

	Cumulative Explained Variance in Individual variables						
		Z1	Z2	Z3	Z4	Z5	Z6
X1	Luxurious	56.9%	84.6%	95.9%	100.0%	100.0%	100.0%
X2	Safe	21.4%	90.0%	99.9%	99.9%	100.0%	100.0%
X3	Sporty	93.5%	93.6%	96.5%	96.6%	99.8%	100.0%
X4	For Family	98.4%	98.6%	98.7%	98.9%	99.2%	100.0%
X5	Practical	84.5%	91.5%	91.7%	98.6%	99.9%	100.0%
X6	Exciting	79.6%	83.6%	95.2%	98.6%	100.0%	100.0%

Fig 8. Cumulative Explained Variance in individual variable

3.3 Analysis from Loading Plot

- The first principal component correlates strongly with 'For Family', 'Practical', and 'Safe', indicating that this component may represent utility-focused features of a car.
- Conversely, it correlates negatively with 'Luxurious', 'Exciting', and 'Sporty', suggesting that these features are inversely related to the utility-focused nature of the first component.
- The second principal component shows a strong positive correlation with 'Practical', 'Exciting', and 'Sporty', which may reflect the car's performance or aesthetic appeal.
- It correlates negatively with 'For Family', 'Luxurious', and 'Safe', positioning these attributes on the opposite end of this spectrum.

We can observe that the first principal component could be interpreted as a 'Utility vs. Luxury' axis, where high scores indicate a focus on practicality and safety, suitable for regular use or family-oriented customers, while low scores suggest luxury and exclusivity. The second component might be viewed as a 'Performance vs. Compact' axis, differentiating between sports and muscular-looking cars and those that are more compact and possibly more efficient.

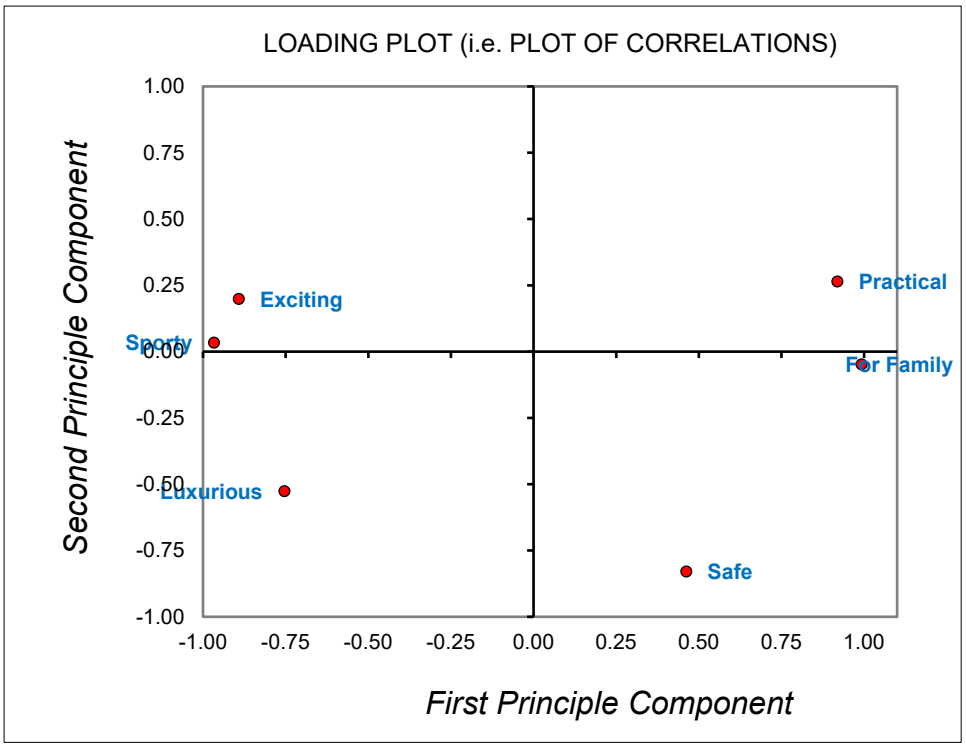

Fig 9. Loading Plot of Car Survey

3.4 Analysis from Score Plot:

• Porsche appears in the upper-left quadrant, suggesting a significant correlation between 'Exciting' and 'Sporty' features, exhibiting traits akin to those of a high-end sports vehicle.

• The brands Jeep, Ford, and Chrysler, which are positioned in the upper-right quadrant, indicate practical features. This means that people view them as practical and appropriate for daily commuting.

• Volvo is recognized for its practical family automobiles, as seen by its placement in the lower-right quadrant and its association with safety and family values.

The way brands like Mercedes, Saab, Lexus, and Infinity are arranged around the centre vertical axis implies that people regard them as having similar levels of luxury.

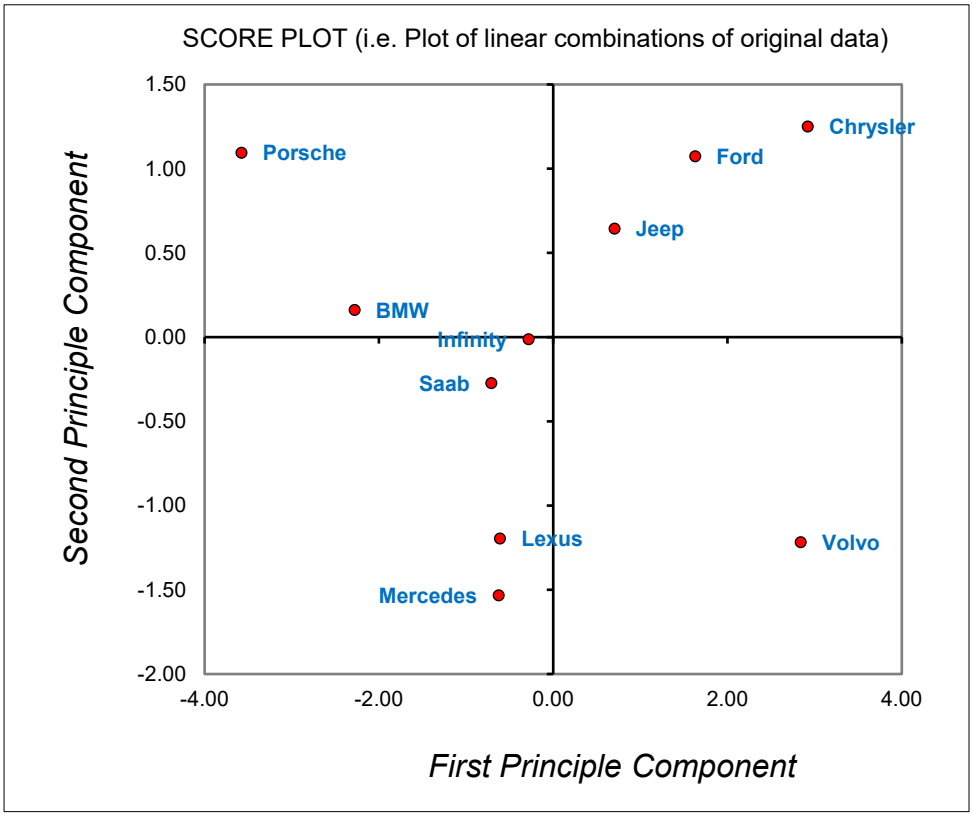

Fig 10. Score Plot of Car Survey

4. Conclusion

This project employed Principal Component Analysis (PCA) on car brand data which is a potent tool for reducing the complexity of multi-dimensional data and uncovering underlying patterns. The analysis revealed that two principal components were sufficient to capture the majority of the variability in the data. The first component differentiated cars based on utility and luxury, while the second contrasted performance-oriented attributes with compactness. The loading plot provided a clear visualization of how each original attribute contributed to these components, and the score plot displayed the positioning of car brands within this new framework, highlighting similarities and differences among them. Clusters identified in the score plot offered insights into brand positioning and consumer perception. Luxury sports brands, practical commuter cars, and family-safe vehicles occupied distinct areas in the PCA-derived space. This segmentation can serve as a strategic tool for stakeholders in the automotive industry to understand competitive positioning, inform marketing strategies, and identify potential gaps in the market.