Final Project Report Part 6

Principal Component Analysis

On Car Attributes Data

By: Arsema Demeke
EMSE 6765
December 14, 2023

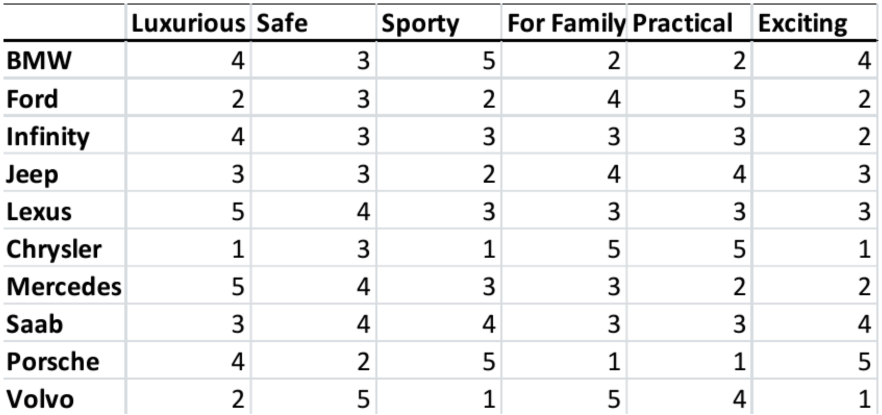
Table of Contents **Page**

Introduction:

A survey data was collected to study the relationship between different types of cars and specific attributes such as luxury and safey. Several variables and attributes were used in conducting the survey and results were collected. This report conducts data analysis on the information collected from the survey and performs a Principal Compoenent Anaysis for dimension reduction and to better understand the relationships observed in the survey.

The results form the collected survey is provided in Table 1 below:

Table 1: Survey Data



	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

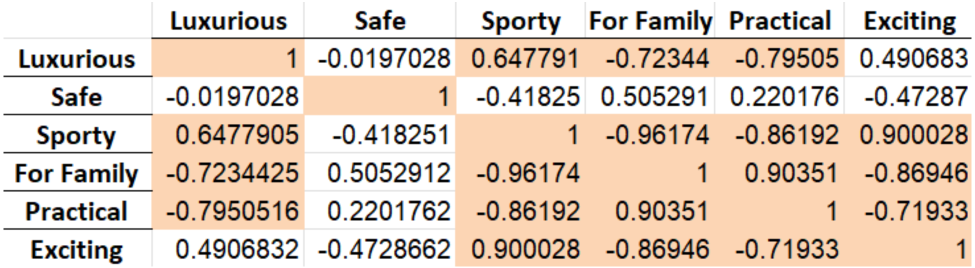
In the survey, the results shows range from numbers 1 to 5 with a higher score indicating that the car is better in terms of having that attribute.

Correlation Analysis:

Although there's plenty of information provided in Table 1, its data is not raw and thus requires the use of a correlation matrix to continue conducting a PCA.

Table 2 below shows the correlation matrix conducted on the attributes data:

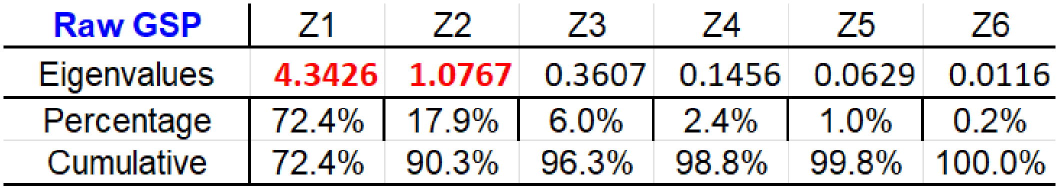
Table 2: Correlation Matrix on Attributes Data



	Luxurious	Safe	Sporty	For Family	Practical	Exciting
Luxurious	1	-0.0197028	0.647791	-0.72344	-0.79505	0.490683
Safe	-0.0197028	1	-0.41825	0.505291	0.220176	-0.47287
Sporty	0.6477905	-0.418251	1	-0.96174	-0.86192	0.900028
For Family	-0.7234425	0.5052912	-0.96174	1	0.90351	-0.86946
Practical	-0.7950516	0.2201762	-0.86192	0.90351	1	-0.71933
Exciting	0.4906832	-0.4728662	0.900028	-0.86946	-0.71933	1

In the correlation matrix displayed in Table 2, a threshold of 0.6 was used. Using the results from the correlation matrix, an Eigen Analysis was conducted using the software Minitab to retain the eigenvector and eigenvalues for the matrix data. Table 3 below depicts the results obtained through the performed calculation:

Table 3: Eigenvalue Analysis



Raw GSP	Z1	Z2	Z3	Z4	Z5	Z6
Eigenvalues	4.3426	1.0767	0.3607	0.1456	0.0629	0.0116
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%

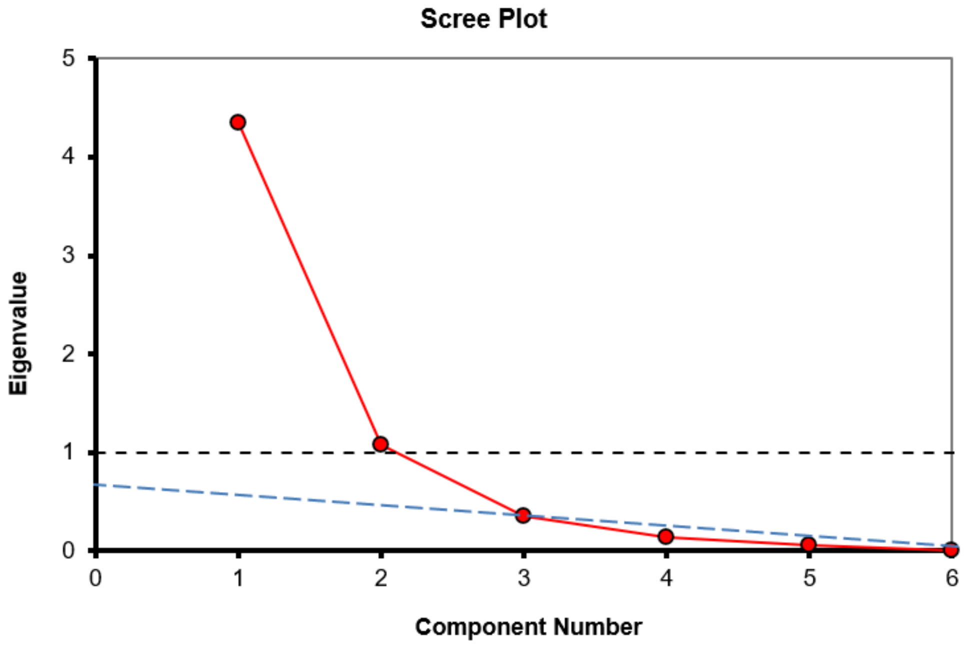
Table 3 above shows that the eigenvalues, which describe the variances of the Principal Components, are observed to be the highest in component 1 and 2 with values 4.3426 and 1.0767. Furthermore, these two components contribute the highest percent explanation of the variance of the overall data since they explain 90.3% of the total variance in the data.

Principal Component Analysis:

Next, a Scree plot can be plotted using the produced eigenvalues. This allows for understanding the principal components that generate the highest variances.

The resulting Scree plot is represented in Figure 1 below:

Figure 1: Scree Plot



Through the scatter plot in Figure 1, the following observations can be made:

- There are 6 Principal Components
- Principal Component 1 has the highest eigenvalue
- A straight line is drawn from right to left following the plotted points starting from the last Principal Component

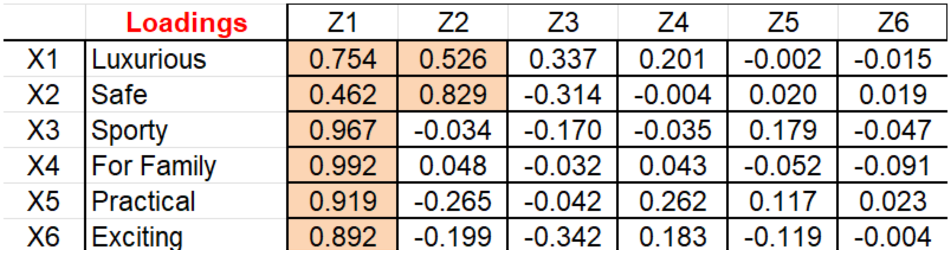
Principal Component Decision:
Using the information found above, a decision on the number of Principal Componenets to be used in the reduced model can be made.

The straight line drawn on Figure 1 allows for the use of the “Elbow Test”. Since the scatter plot points diverge from the linear line at Component Number 3, the number of Principal Components that should be used is 1 fewer than that. Thus, only two components, Principal Component 1 and 2, will be used for the PCA. This decision is further supported by the Kaiser Test which suggest that only components with eigenvalues greater than 1 should be used. The remaining Principal Components are thus general errors that are hard to intepret for this data and won’t be used. For these reasons, the model was reduced from six dimensions to two.

Analysis on Variables:
Next, several analysis can be conducted using the Principal Components.

Below, Table 4 provides calculated data on the correlation between the original variables and all the Principal Components:

Table 4: Correlation of Original Variables and Principal Components

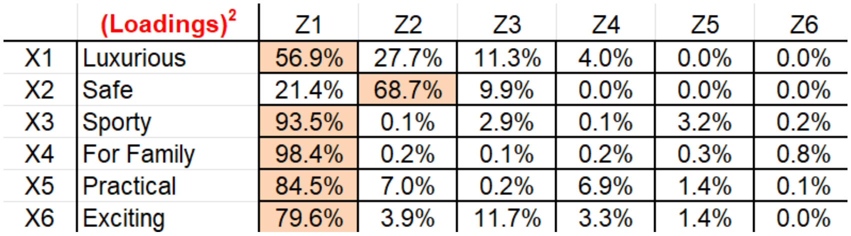


	Loadings	Z1	Z2	Z3	Z4	Z5	Z6
X1	Luxurious	0.754	0.526	0.337	0.201	-0.002	-0.015
X2	Safe	0.462	0.829	-0.314	-0.004	0.020	0.019
X3	Sporty	0.967	-0.034	-0.170	-0.035	0.179	-0.047
X4	For Family	0.992	0.048	-0.032	0.043	-0.052	-0.091
X5	Practical	0.919	-0.265	-0.042	0.262	0.117	0.023
X6	Exciting	0.892	-0.199	-0.342	0.183	-0.119	-0.004

A threshold of 0.6 was used in Table 4 above. Through this table, it is shown that Principal Component 1 has the highest correlation with the variables when compared to the other components and demonstrates this high correlation behavior with all of the variables. Thus, it can be concluded that the first component is a common source of variation amongst all of the variables and is a general activity component; likewise, Component 2 can thus be identified as a classification component.

To explore the relationships in this PCA further, Table 5 below is used to display the calculated percent explanation of the variation of each principal components with the original variables:

Table 5: 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%

Table 5 above demonstrates that Principal Compoent 1 and 2 have the highest individual percent variance explanation for the variables with the first component demonstrating this behavior for almost all the original variables.

Data Assessment:

Since the dimension of the original model has been reduced and the principal components have been chosen, a correlation of the two components can be conducted using a loading plot.

Figure 2 below illustrates the loading plot:

Figure 2: Loading Plot

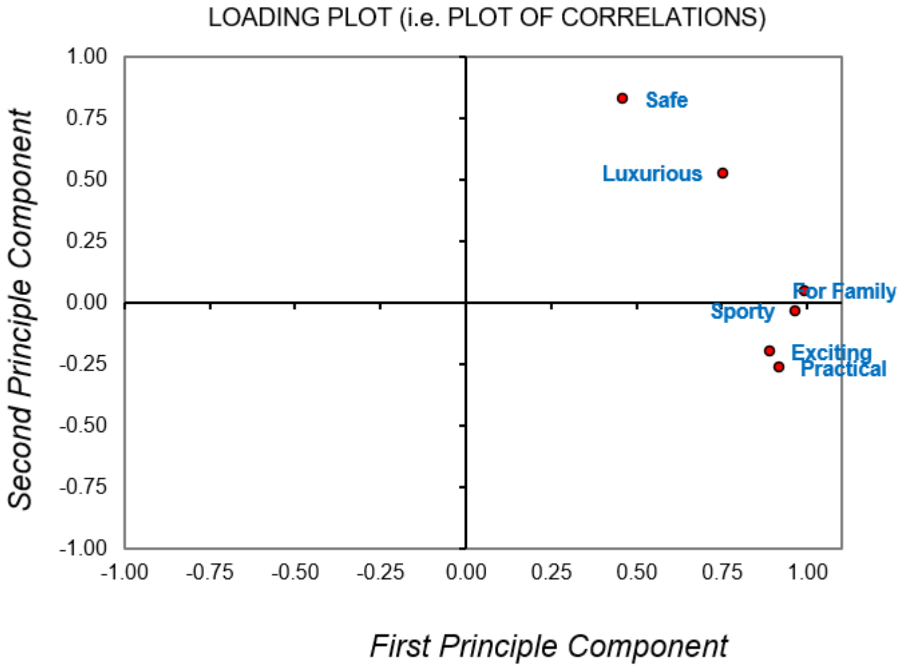


Figure 2 above illustrates the relationship between the two Principal Components that were chosen. The following observations can be made from the plot:

- All the variables are positively correlated with the First Component as they all lie high on the positive correlation scale
- The variables Exciting, Practical, Sporty, and For Family are negatively correlated with the Second Componenet as they lie in the negative correlation region for the second component
- It can be observed that the variables For Family, Sporty, Exciting, and Practical are the attributes that are clustered and thus indicate their strong correlation with each other
- Safe is the attribute that is farthest from the other variables
- Safe and the attributes Sporty and Exciting have a weak correlation
- Attributes like Luxurious and For Family have a weak correlation

However, more can be analyzed from this information through a creation of a score plot that plots the scores of the original data set. Figure 3 below demonstrates this:

Figure 3: Score Plot

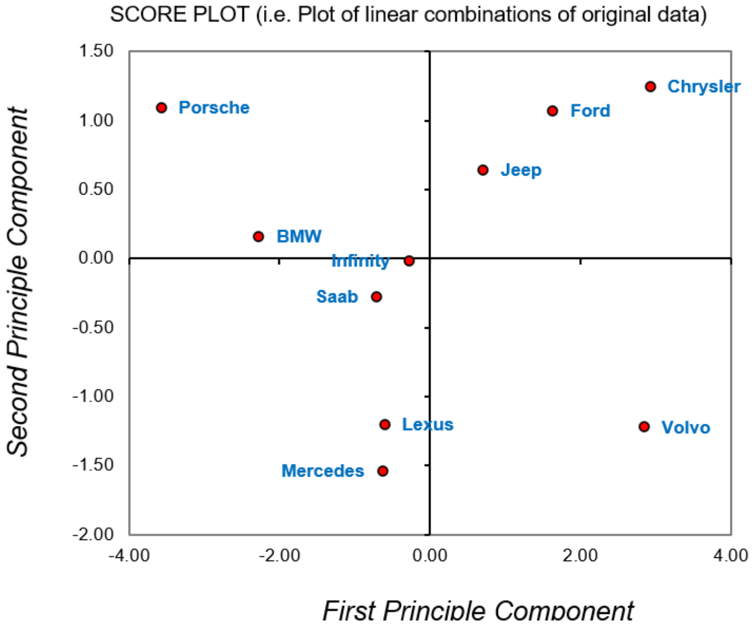


Figure 3 above was created using the Z scores of the original data which are the linear combinations that exist within the data. Its findings suggest that vehicle types like BMW, Infinity, and Saab share the same pattern in the observed data and are strongly correlated. A similar behavior can be observed between Mercedes and Lexus. However, the Porsche and Volve brands seem to be the least correlated with the other variables as they lie the farthest away.

Although not immediately clear, the correlations and trends observed in Figure 2 and 3 are supported by each other. For instance, Figure 5 displays the attribute For Family and Sporty to the right most section of the correlation plot. Similarly, the Volvo car brand shares a similar location in Figure 3's plot. Interestingly, the Volvo is known for being the best family car as it offers sportiness and is thus supported by Figure 2 and 3 above. Additionally, the Chrysler has been a top safety pick for car brands and likewise shares a similar location with the safety attribute from Figure 2 as they're both located in the top right. When it comes to the practical attribute, Mercedes is known for this feature and its reliability and is also supported by Figure 2 and 3.

Conclusion:

Through this assessment conducted using Principal Component Analysis, it can be validated that the original data should be reduced to the dimensions illustrated in the PCA as the remaining variables used in the reduced model are strongly associated and correlated.