KU LEUVEN

Multivariate Statistics

Group #27

Assignment 3: Cluster Analysis & Discriminant Analysis

Natalia Eremeeva r0648321
Rana Cansu Kebabcı r0772826
Ömer Yiğit r0767950

Prof. Vandebroek Martina 2019-2020

TASK 1 - CLUSTER ANALYSIS:

Introduction and Problem Statement:

We are looking at the dataset that contains information about endurance, strength, power, speed, agility, flexibility, nerve, durability, hand-eye coordination and analytic aptitude for 60 different sports. We would like to perform *clustering analysis* on this dataset. The idea of clustering is to be able to identify groups of objects that are more similar to each other within the same group and different compared to any other group. In our case, we will try to identify the groups of sports that are more alike than others based on the abovementioned criteria (strength, power, speed, etc.).

Descriptive Statistics:

The data consists of 10 numerical variables with means between 4.61 and 5.17 and standard deviations between 1.58 and 2.44 which seems to be good enough to perform analysis without normalizing the data. Data does not show any outliers.

Testing the Assumptions:

The beauty of cluster analysis is that it does not require any assumptions about the data itself. We are first going to use hierarchical clustering methods which do not require a prespecified number of clusters and then we can look at algorithms with a prespecified number of clusters that we determined through hierarchical methods and compare results.

Conducting the Method and Interpretation:

We are going to perform cluster analysis that combines data points based on distance from points to the centroid (centroid method), based on the distance between two closest data points in any two separate clusters (single linkage) and based on the distance between the two furthest data points in any two separate clusters (complete linkage). After we determine the number of clusters with hierarchical methods we will re-run the analysis using a nonhierarchical cluster method called k-means with improved initial seeds and compare the results.
You can find results for centroid, single linkage and complete linkage in the tables below.

Method	Number of Clusters based on R-squared	Number of Clusters based on Pseudo t-squared	Number of Clusters based on Pseudo F
Centroid	9,10 (R2 = 67-70%)	3,8,15	3,9
Simple linkage	17, 18 (R2 = 71%)	4,17	4,17
Complete Linkage	8, 9 (R2 = 72%)	6,9	6,8

Number of clusters based on different methods and statistics

As we can see, the number of clusters depending on the criteria we are looking at vary, we can see that number of clusters suggested by looking at R-squared (proportion of variance accounted for by the clusters) is different from judgment based on pseudo t-squared or pseudo F statistic which seem to give more or less the same recommendations.
Judging by the suggested number of clusters we will not pursue with single linkage method as we only have 60 countries to group and 17 clusters (based on R-squared) is too much, therefore we are going to look at cluster solutions with 9 clusters for both centroid and complete linkage methods and compare results.

Cluster #	Sports included *Centroid method*	Sports included *Complete linkage method*
#1	Bowling, Curling, Billiards, Shooting, Archery, Fishing, Golf	Bowling, Curling, Billiards, Shooting, Archery, Fishing, Golf
#2	Track and Field: High Jump, Track and Field: Long, Triple jumps, Skiing: Nordic, Swimming (all strokes): Distance, Track and Field: Distance, Swimming (all strokes): Sprints, Track and Field: Sprints, Track and Field: Middle Distance, Canoe/Kayak, Rowing, Roller Skating, Cycling: Distance, Speed Skating	Track and Field: High Jump, Track and Field: Long, Triple jumps, Swimming (all strokes): Sprints, Track and Field: Sprints, Track and Field: Middle Distance, Cycling: Sprints, Speed Skating
#3	Field Hockey, Lacrosse, Badminton, Racquetball/Squash, Team Handball, Volleyball, Basketball, Ice Hockey, Soccer, Tennis, Fencing, Water Polo, Baseball/Softball, Table Tennis	Field Hockey, Lacrosse, Basketball, Ice Hockey, Soccer, Tennis, Boxing, Football, Baseball/Softball
#4	Skiing: Freestyle, Surfing, Skateboarding, Water Skiing, Martial Arts, Skiing: Alpine, Ski Jumping, Bobsledding/Luge, Cycling: Sprints, Rodeo: Calf Roping, Track and Field: Pole Vault, Rodeo: Steer Wrestling, Figure Skating, Gymnastics, Wrestling, Diving, Rodeo: Bull/Bareback/Bronc Riding	Skiing: Nordic, Swimming (all strokes): Distance, Track and Field: Distance, Canoe/Kayak, Rowing, Cycling: Distance
#5	Track and Field: Weights, Weight Lifting	Badminton, Racquetball/Squash, Team Handball, Volleyball, Fencing, Table Tennis
#6	Equestrian, Horse Racing	Skiing: Freestyle, Surfing, Skateboarding, Water Skiing, Bobsledding/Luge, Ski Jumping, Rodeo: Calf Roping, Cheerleading, Roller Skating, Diving, Rodeo: Bull/Bareback/Bronc Riding
#7	Football, rugby, boxing	Track and Field: Weights, Weight-Lifting
#8	Cheerleading	Martial Arts, Skilling: Alpine, Rugby, Water Polo, Wrestling, Rodeo: Steer Wrestling, Track and Field: Pole Vault, Figure Skating, Gymnastics
#9	Auto Racing	Equestrian, Horse Racing, Auto Racing

Clusters based on different methods

Judging from the table of classification of sports into 9 clusters we see a lot of similarities and some differences between the final groupings with 2 methods. Clustering is all about interpretability, so let us have a closer look at the results proposed by the complete linkage method. Please have a look at the results below:

Cluster #	Why got clustered together?
#1	Sports with the lowest levels of endurance and speed
#2	Sports with low levels of flexibility, stress and hand-eye coordination
#3	Sports with high levels of durability and power
#4	Sports with the highest levels of endurance and lowest levels of hand-eye coordination
#5	Sports with the highest level of hand-eye coordination
#6	Sports with low levels of endurance, strength, and power
#7	Sports with the highest levels of power and strength
#8	Sports with high levels of power, stress (nerves) and durability
#9	Sports with the lowest levels of power and speed and the highest levels of durability

Explanation of the clusters

Alternative Solution:

When performing clustering analysis, there is no such thing as right or wrong solution, as results are open to interpretation and if it is well-justified, then it can be counted as a success. Alternative solutions can be in this case the results of Centroid method from the previous section, the results of Simple linkage method, as well as performing non-hierarchical clustering with maximum number of clusters equal to 9 as determined in the previous section.
As discussed above, simple linkage method proposed too many clusters to retain in order to account for at least 70% of variability, therefore this method for our dataset is not suitable. As for centroid method, we prefer not to take that one as a final solution, because most of the clusters group very small (1-3) amount of sports inside, while others take the vast majority of sports under investigation. Thus, centroid method with 9 clusters grouping makes it hard to interpret the solution. The last alternative solution is to run k-means with improved initial seeds. After it has been performed we got the following results:

Cluster #	Sports included *K-means with improved seeds method*
#1	Canoe/Kayak, Cycling: Distance, Roller Skating, Rowing, Skiing: Nordic, Swimming (all strokes): Distance
#2	Auto Racing
#3	Badminton, Baseball/Softball, Fencing, Field Hockey, Lacrosse, Racquetball/Squash, Soccer, Table Tennis, Team Handball, Tennis, Volleyball
#4	Cycling: Sprints, Speed Skating, Swimming (all strokes): Sprints, Track and Field: High Jump. Track and Field: Long, Triple jumps, Track and Field: Middle Distance, Track and Field: Sprints
#5	Archery, Cheerleading, Equestrian, Golf
#6	Billiards, Bowling, Curling, Fishing, Shooting
#7	Track and Field: Wights, Weight-Lifting
#8	Basketball, Boxing, Figure Skating, Football, Gymnastics, Ice Hockey, Martial Arts, Rugby, Skiing: Alpine, Water Polo, Wrestling
#9	Bobsledding/Luge, Diving, Horse Racing, Rodeo: Bull/Bareback/Bronc Riding, Rodeo: Calf Roping, Rodeo: Steer Wrestling, Skateboarding, Ski Jumping, Skiing: Freestyle, Surfing, Track and Field: Pole Vault, Water Skiing

Clusters according to K-means method

We can already spot similarities between the clusters chosen by complete linkage method and k-means. Cluster #6 in k-means is the same as cluster #1 with complete linkage; except, complete linkage puts extra sport ‘archery’ in the same cluster. Track and Field: Wights and Weight-Lifting are classified together as a cluster by both methods. While there are similarities in classification, cluster 5 in k-means, on the other hand, is completely different from complete linkage grouping, as in complete linkage all 4 sports belong to 4 different clusters. Moreover, let us validate how good the clusters are:

Statistics for Variables				
Variable	**Total STD**	**Within STD**	**R-Square**	**RSQ/(1-RSQ)**
END	2.09422	1.19420	0.718919	2.557697
STR	1.71536	1.00908	0.700869	2.343021
PWR	1.98040	1.19331	0.686154	2.186276
SPD	2.26640	1.10648	0.793970	3.853664
AGI	1.92727	0.87924	0.820092	4.558390
FLX	1.70437	1.35883	0.450562	0.820043
NER	2.44265	1.34439	0.738151	2.818995
DUR	1.88210	1.16826	0.666951	2.002558
HAN	1.96633	1.14037	0.709267	2.439584
ANA	1.57683	0.98130	0.665227	1.987101
Overall	1.97149	1.14663	0.707598	2.419952

Table for validation of the clusters

Clusters suggested by k-means method explain (account for) the variability of variables in clusters pretty well, except for variable *Flexibility*.

Conclusion:

Clustering is the task of dividing the population into a number of groups such that data points in the same groups are more similar to data points in the same group than those in other groups. After performing various clustering methods (simple linkage, centroid, complete linkage and k-means with improved initial seeds) we take complete linkage method grouping with 9 clusters:

- Sports with the lowest levels of endurance and speed
- Sports with low levels of flexibility, stress and hand-eye coordination
- Sports with high levels of durability and power
- Sports with the highest levels of endurance and lowest levels of hand-eye coordination
- Sports with the highest level of hand-eye coordination
- Sports with low levels of endurance, strength, and power
- Sports with the highest levels of power and strength
- Sports with high levels of power, stress (nerves) and durability
- Sports with the lowest levels of power and speed and the highest levels of durability

This makes the most sense to us interpretability wise and as the method explains a decent part of the variability in clusters we keep complete linkage as a final method. Alternative methods classify sports slightly differently, but some patterns stay the same after all.

TASK 2 - DISCRIMINANT ANALYSIS:

Problem Statement:
The cars dataset consists of 12 variables measured on 74 observations. The variables are: Name - Name of the car, Price - Price of the car, Mileage - Miles per gallon, Headroom - Room above front seat, Rearseatcl - Space between front and rear seat, Trunk - Trunk space, Weight - Weight of the car, Length - Length of the car, Turning - Turning diameter, Volume - Total volume of the car, Gearratio - Gear ratio for high gear, Headq - Company headquarters / US(1), Japan(2), Europe(3).
The problem is that if we can describe the differences amongst cars by headquarters, using the measured variables. In order to achieve some conclusions, discriminant analysis is to be performed.

Descriptive Statistics:
First, to have a better understanding about the data, some descriptive statistics should be extracted. It is also helpful to detect outliers, incorrect or missing data etc. *proc means* procedure is used on SAS to obtain the basic statistics. According to the table below, no anomalies are detected. Plus, the standardized data is checked. There is only one observation with the price value being outside the range of 6σ. However, it is not considered as a significant outlier for the entire dataset and kept.

Variable	N	Mean	Std Dev	Minimum	Maximum
price	74	6192.28	2938.06	3291.00	15906.00
mileage	74	21.2972973	5.7855032	12.0000000	41.0000000
headroom	74	2.9864865	0.8398180	1.5000000	5.0000000
rearseatcl	74	26.8175676	3.1269644	18.5000000	37.5000000
trunk	74	13.7432432	4.2685881	5.0000000	23.0000000
weight	74	3010.81	783.8957101	1760.00	4840.00
length	74	188.0675676	22.4073930	142.0000000	233.0000000
turning	74	39.7972973	4.3193725	32.0000000	51.0000000
volume	74	197.2972973	91.8372190	79.0000000	425.0000000
gearratio	74	3.0181081	0.4532037	2.1900000	3.8900000

Output of the *proc means* procedure

Assumptions:
Before applying discriminant analysis, the main assumptions of the procedure should be stated.

- *The dataset is suitable for discriminant analysis.*

This assumption can be tested with Wilks' Lambda (Λ) value. SAS uses Rao-approximation for the distribution of Λ under the null hypothesis.

H0 : $\mu 1 = \mu 2 = \mu 3$

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.20533733	7.48	20	124	<.0001

Test for Wilks' Lambda

According to the p-value, the dataset is suitable for discriminant analysis.

- *Samples come from multivariate normal distribution.*

Histograms are checked for the distribution of the variables. Price and Volume variables seem to have non-normal distributions. However, it is important that the sample of the subset of discriminatory variables should come from multivariate normal distribution. In this analysis, the most important variables (turning & rearseatcl) are normally distributed.

- *Covariance matrices among groups are the same.*

The covariance matrices of the groups are calculated. Then, the equality of the matrices are tested pairwise.
H0: $\sum i = \sum j$; $i,j \in \{1,2,3\}$; $i \neq j$
H1: $\sum i \neq \sum j$; $i,j \in \{1,2,3\}$; $i \neq j$

According to the hypothesis testing, ∑1 is accepted to be different than ∑2 and ∑3. This is actually against the assumption of discriminant analysis, therefore this result may affect the power and the reliability of the analysis.

- *Misclassification costs are equal.*

This is for the simplicity of the analysis. We consider that the misclassifications of the cars will not have significantly different costs.

Conducting the Method and Interpretation:

We are going to perform canonical discriminant analysis in which indexes (discriminant functions) are composed to distinguish the groups in the dataset. These discriminant functions are used to find out which variables are significantly discriminating and to predict which classified group the new observation belongs. In the scope of the assignment, Fischer Approach is used. It divides the space mutually exclusive and collectively exhaustive regions by creating new axes (discriminant functions) which are linear combinations of chosen variables (discriminating variables). In that approach, groups are separated by maximizing the Mahalanobis distance between group means and minimizing the intra group variance. Relatedly, classification of the new observation is done by Mahalanobis Distance Approach. The new observation is assigned to a group which is closest in terms of euclidean distance between the discriminant score of the new observation and mean score of the group means (Mahalanobis distance).

In this analysis we are going to check if discriminant functions explain the variability across headquarter sufficiently and in which variable aspect, groups differ among each other. At the end we are going to check the accuracy of the classification rule as well.

As the first step of the analysis we check the Mahalanobis distance between the group means which gives the first measure of the applicability of the discriminant analysis.

Squared Distance to headq			
From headq	1	2	3
1	0	7.19277	17.36233
2	7.19277	0	8.38362
3	17.36233	8.38362	0

Squared mahalanobis distances among groups

Prob > Mahalanobis Distance for Squared Distance to headq			
From headq	1	2	3
1	1.0000	<.0001	<.0001
2	<.0001	1.0000	0.0003
3	<.0001	0.0003	1.0000

p-values of squared mahalanobis distances among groups

Multivariate Statistics and F Approximations					
S=2 M=3.5 N=30					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.20533733	7.48	20	124	<.0001
Pillai's Trace	1.00075145	6.31	20	126	<.0001
Hotelling-Lawley Trace	2.86637546	8.77	20	101.05	<.0001
Roy's Greatest Root	2.45806210	15.49	10	63	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

Test statistics for discriminatory power

From the results, we conclude that the distances between group means are significant, therefore discriminant analysis is applicable for the groups of headquarters. The number of discriminant functions is min{p, G-1}. In this sense, we have two canonical discriminant functions in the analysis — min{10, 3-1} = 2. In order to check their significance of discrimination, we checked the Wilk's Lambda. The null hypothesis is that the eigenvalues of the discriminant functions are zero. Eigenvalue is the measurement of the ratio of the sum of squares between groups and sum of squares within groups.

As seen above, at least one of the two discriminant functions has a significant discriminating power. In order to evaluate the discriminating power in detail, we check the eigenvalues and regarding explained percent of variance among groups.

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)			
					Eigenvalue	Difference	Proportion	Cumulative
1	0.843102	0.820343	0.033846	0.710821	2.4581	2.0497	0.8576	0.8576
2	0.538452	0.472411	0.083107	0.289931	0.4083		0.1424	1.0000

Canonical correlations and eigenvalues for the discriminant analysis

Test of H0: The canonical correlations in the current row and all that follow are zero				
Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
0.20533733	7.48	20	124	<.0001
0.71006924	2.86	9	63	0.0069

Hypothesis test results for canonical correlations

As seen above, we can conclude that the first discriminant function's ability to discriminate the groups is almost 6 times higher. We can also check the squared canonical correlation (R2) which is an equivalent result to the previous evaluation. There we can see that R2 value of the discriminant function is about 50%. It is a moderate value for the R square. However, since 0.0069<0.05 based on the F-test, we conclude that both discriminant functions' eigenvalues are significantly different than zero and they have discriminatory power. As a next step, we check the relative discriminating ability of the discriminating variables by standardized canonical coefficients.

Pooled Within-Class Standardized Canonical Coefficients		
Variable	Can1	Can2
price	-0.813185136	-0.023967204
mileage	0.461275685	-0.318769711
headroom	0.337000607	-0.504679011
rearseatcl	-0.507482832	0.715874436
trunk	-0.482634797	0.981565171
weight	0.560999273	0.473382901
length	0.235841677	-1.683542964
turning	0.530076401	0.626129788
volume	0.269852053	-0.022556851
gearratio	-0.636511921	0.047817295

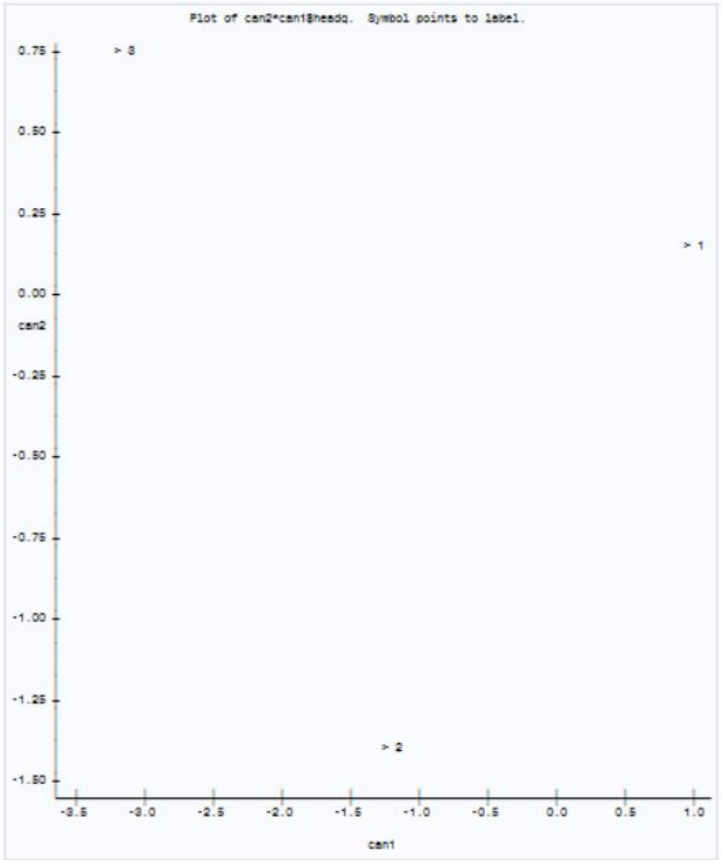
Pooled Within Canonical Structure		
Variable	Can1	Can2
price	-0.079026	0.306724
mileage	-0.236678	-0.441958
headroom	0.212622	0.243224
rearseatcl	0.052663	0.662529
trunk	0.177257	0.685427
weight	0.416624	0.468584
length	0.400871	0.424328
turning	0.531371	0.580240
volume	0.455685	0.428752
gearratio	-0.554005	-0.396118

The absolute values indicates the partial importance(contribution) of the discriminating variable to predict the discriminant function. We took standardized table to avoid the scale differences between variables. In the table, it is seen that price and gear ratio has higher discriminating power on the first canonical variable. But the relation is negative. Turning and weight follow them as positively related powerful discriminating variables. On the other hand; length, trunk and space between rear and front seat play more important role to predict the second canonical variable score. While the length has a negative relation, trunk space and rear-front seat space have positive relations.

The next step is labeling the discriminant functions with respect to variables. We benefit from pooled within canonical structure matrix which gives correlation between the canonical variable and the discriminating variables based on structural loading. So that, the misleading results caused by multicollinearity are avoided. When we check the differences between absolute values, we see that relatively higher differences are observed for price, mileage, rearseatcl and trunk.

As follows, raw coefficients of canonical variables and mean scores of the groups are provided. With the help of these outputs, canonical variable-group mean is plotted to see how the groups fall apart based on new variables (canonical variables).We can see that the groups are well classified with respect to new variables. They look quite distant to each other.
With the help of pooled within canonical structure matrix, we can label the cars of different headquarters as:

- US (1) — Cars with high turning diameter and size (weight & volume)
- Japan (2) — Cars with low price and high gear ratio
- Europe (3) — Cars with high price and low turning diameter



Plot of the group means on canonical variables

Class Means on Canonical Variables		
headq	Can1	Can2
1	0.938304103	0.139632826
2	-1.251798839	-1.408339317
3	-3.183820556	0.748256867

Raw Canonical Coefficients		
Variable	Can1	Can2
price	-0.000280200	-0.000008258
mileage	0.086758509	-0.059955436
headroom	0.421662050	-0.631464696
rearseatcl	-0.174307717	0.245885045
trunk	-0.125615612	0.255472483
weight	0.000869093	0.000733359
length	0.012578744	-0.089792682
turning	0.163791330	0.193471413
volume	0.003648834	-0.000305005
gearratio	-1.867831331	0.140318884

The last step is checking accuracy of the classification rule. Misclassified observations as a result of the new variable (discriminant function) are determined and the ratio of the misclassification is evaluated. For this procedure the prior probability to belong to a group in population should be indicated. In the scope of the assignment, prior probabilities are taken equal as 0.33 for each group. Besides, we assume that the misclassification costs are equal as well for simplicity.
We start with the misclassified observations posterior probability table.

Posterior Probability of Membership in headq						
Obs	From headq	Classified into headq		1	2	3
1	1	2	*	0.4151	0.5233	0.0616
3	1	2	*	0.4577	0.5423	0.0001
10	1	2	*	0.4073	0.5926	0.0001
27	1	2	*	0.0506	0.9367	0.0128
56	1	3	*	0.1120	0.4336	0.4543
57	1	2	*	0.0262	0.9724	0.0015
65	1	3	*	0.0142	0.3171	0.6686

Misclassified observations

In models where misclassification costs are equal, observations are assigned to the group where the posterior probabilities is higher. We see the observations are classified in wrong group based on their posterior probabilities. We can observe misclassifications in terms of percent ratios by the following table.

Number of Observations and Percent Classified into headq				
From headq	1	2	3	Total
1	45 86.54	5 9.62	2 3.85	52 100.00
2	0 0.00	11 100.00	0 0.00	11 100.00
3	0 0.00	0 0.00	11 100.00	11 100.00
Total	45 60.81	16 21.62	13 17.57	74 100.00
Priors	0.33333	0.33333	0.33333	

Confusion matrix of the discriminant analysis

Error Count Estimates for headq				
	1	2	3	Total
Rate	0.1346	0.0000	0.0000	0.0449
Priors	0.3333	0.3333	0.3333	

Error Count Estimates for headq				
	1	2	3	Total
Rate	0.1731	0.0000	0.0909	0.0880
Priors	0.3333	0.3333	0.3333	

ECM tables without (above) and with (below) cross-validation

Misclassified observations belong to group 1 is 13,46%. On the other hand group 2 and group 3 are classified 100% correct. Another measurement of the misclassification is error count estimates which is the overall expected cost of misclassification which is calculated by misclassification ratio(13,46%) times prior probability(0,3333) when the costs are equal.
We compare this value with the upper bound (G*p1*p1')which is ECM of proportional assignment according to prior probabilities. The aim is to see if the misclassification is significantly high.
As mentioned before, 0.0449=0.1346*0.333 which is less than (0.333*0.667*3)=0.666. Therefore, we can conclude that the misclassification is not significant. Same conclusion is made by the cross-validation method as well, despite a neglectable change in misclassified observations.

Alternative Solution(s):

Alternatively, stepwise discriminant analysis is applied to determine discriminatory variables. In 5 steps, the variables used for discriminant analysis as follows:

Turning > Rearseatcl > Gearratio > Price

Step	Number In	Entered	Removed	Label	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	turning		turning	0.4540	29.52	<.0001	0.54599579	<.0001	0.22700211	<.0001
2	2	rearseatcl		rearseatcl	0.1981	8.65	0.0004	0.43781876	<.0001	0.31489265	<.0001
3	3	gearratio		gearratio	0.1749	7.31	0.0013	0.36123555	<.0001	0.36081012	<.0001
4	4	price		price	0.1971	8.35	0.0006	0.29003359	<.0001	0.40372979	<.0001

Summary table for stepwise discriminant analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.812200	0.802049	0.039833	0.659668
2	0.384436	0.359297	0.099743	0.147791

The result shows us the reduction in discriminating variables didn't affect the significance of the discriminant variables. And practical significance (CR^2) of the first discriminant function hasn't decreased remarkably as well compared to the

unreduced first model. Below we observe the confusion matrix of reduced model. This result shows the misclassification increased and ECM value is still good as 0.2028< 0,666. Also, KNN method is applied for misclassifications which is non-parametric approach. The k value is selected to be 5. Smaller k value may cause underfitting and larger k value may cause overfitting. The result is very close to the Fischer Approach's.

Number of Observations and Percent Classified into headq				
From headq	1	2	3	Total
1	44 84.62	8 15.38	0 0.00	52 100.00
2	0 0.00	9 81.82	2 18.18	11 100.00
3	0 0.00	3 27.27	8 72.73	11 100.00
Total	44 59.46	20 27.03	10 13.51	74 100.00
Priors	0.33333	0.33333	0.33333	

Number of Observations and Percent Classified into headq				
From headq	1	2	3	Total
1	44 84.62	7 13.46	1 1.92	52 100.00
2	0 0.00	11 100.00	0 0.00	11 100.00
3	0 0.00	1 9.09	10 90.91	11 100.00
Total	44 59.46	19 25.68	11 14.86	74 100.00
Priors	0.33333	0.33333	0.33333	

Confusion matrix of the reduced discriminant model (left) and KNN method (right)

Error Count Estimates for headq				
	1	2	3	Total
Rate	0.1538	0.1818	0.2727	0.2028
Priors	0.3333	0.3333	0.3333	

Error Count Estimates for headq				
	1	2	3	Total
Rate	0.1538	0.0000	0.0909	0.0816
Priors	0.3333	0.3333	0.3333	

ECM tables of reduced discriminant model (left) and KNN method (right)

Conclusion:

In terms of misclassification the results between KNN method and Fischer's Approach did not differ remarkably. Although Fischer gave better results, KNN is helpful to have good classification when the distribution is not normal. On the other hand, the reduced model gave worse discriminant analysis in which ECM increased to 0.2028 from 0.0445 but still non-significant. As a summary, the misclassification tables for both methods are given in the table below. According to the table, Fisher's approach gives the best results.

Method	Observation	Variable	Misclassification	ECM
Fischer	74	10	7/74 = 9.46%	0.0449
Reduced Model	74	4	13/74 = 17.57%	0.2028
KNN	74	10	9/74 = 12.16%	0.0816

ECM results of tested methods

APPENDIX A - SAS CODES FOR TASK 1

```
FILENAME REFFILE '/folders/myfolders/sasuser.v94/sport.xlsx';
PROC IMPORT DATAFILE=REFFILE
        DBMS=XLSX
        OUT=WORK.IMPORT3;
        GETNAMES=YES;
RUN;
PROC CONTENTS DATA=WORK.IMPORT3; RUN;
proc print data = work.import3; run;
proc  cluster  data=work.import3  simple  noeigen  method=centroid  rsquare  rmsstd  pseudo  nonorm
out=tree;
        id SPORT;
        var END STR PWR SPD AGI FLX NER DUR HAN ANA;
proc tree data=tree out=clus9 nclusters=9;
id SPORT;
copy END STR PWR SPD AGI FLX NER DUR HAN ANA;
proc sort; by CLUSTER;
proc print data = clus9;
run;
proc cluster data=work.import3 simple noeigen method=single rmsstd rsquare
                pseudo nonorm out=tree;
        id SPORT;
        var END STR PWR SPD AGI FLX NER DUR HAN ANA;
proc cluster data=work.import3 simple noeigen method=complete rmsstd rsquare
                pseudo nonorm out=tree;
        id SPORT;
        var END STR PWR SPD AGI FLX NER DUR HAN ANA;
proc tree data=tree out=clus9 nclusters=9;
id SPORT;
copy END STR PWR SPD AGI FLX NER DUR HAN ANA;
proc sort; by CLUSTER;
proc print data = clus9;
run;
proc fastclus data = work.import3 radius =1 replace = full
maxclusters = 9 maxiter = 20 out = fastnew;
var END STR PWR SPD AGI FLX NER DUR HAN ANA;
proc sort data=fastnew; by cluster;
proc print data=fastnew; by cluster;
var sport cluster distance END STR PWR SPD AGI FLX NER DUR HAN ANA;
run;
```

APPENDIX B - SAS CODES FOR TASK 2

```
*Importing and sorting the data;
libname c xlsx ".../cars.xlsx";
data cars;
	set c.cars;
run;
proc sort data=cars;
	by headq;
run;
proc print data=cars;
run;
* Descriptive statistics;
proc standard data=cars mean=0 std=1 out=cars_st;
	var price mileage headroom rearseatcl trunk weight length turning volume gearratio;
proc means data=cars nolabel;
	var price mileage headroom rearseatcl trunk weight length turning volume gearratio;
run;
* Covariance matrices for each group;
data us;
	set cars;
	where headq=1;
data jp;
	set cars;
	where headq=2;
data eu;
	set cars;
	where headq=3;
run;
* Here is only one pairwise test;
proc calis covpattern=eqcovmat;
	var price mileage headroom rearseatcl trunk weight length turning volume gearratio;
	group 1 / data=jp nobs=11;
	group 2 / data=eu nobs=11;
	fitindex NoIndexType On(only)=[chisq df probchi];
run;
* Normality of variables;
proc univariate data=cars;
	var price mileage headroom rearseatcl trunk weight length turning volume gearratio;
	histogram price mileage headroom rearseatcl trunk weight length turning volume
gearratio;
run;
* Discriminant analysis;
proc candisc data=cars distance out=cars_discr;
	class headq;
	var price mileage headroom rearseatcl trunk weight length turning volume gearratio;
proc sort;
	by headq;
proc means noprint;
	var can1 can2;
	by headq;
proc plot;
	plot can2*can1 $ headq;
run;
proc discrim data=cars listerr crossvalidate crosslisterr;
	class headq;
	var  price mileage headroom rearseatcl trunk weight length turning volume gearratio;
run;
* Stepwise;
proc stepdisc data=cars;
	class headq;
	var price mileage headroom rearseatcl trunk weight length turning volume gearratio;
run;
* KNN for misclassification;
proc discrim data=cars method=npar k=5 listerr;
	class headq;
	var price mileage headroom rearseatcl trunk weight length turning volume gearratio;
run;
```