ΕΡΓΑΣΙΑ 3

ΠΟΛΥΜΕΤΑΒΛΗΤΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

ΔΗΜΗΤΡΙΟΣ ΓΚΑΒΕΡΑΣ
ΑΛΕΞΑΝΔΡΟΣ ΣΚΟΝΔΡΑΣ

ΕΡΓΑΣΙΑ 3

ΔΗΜΗΤΡΙΟΣ ΓΚΑΒΕΡΑΣ (1112201500042)

ΑΛΕΞΑΝΔΡΟΣ ΣΚΟΝΔΡΑΣ (1112201500206)

A)

1) Κέντρα βάρους για τα γκρουπς 1 και 2 αντίστοιχα:

(50.2500 34.5000 14.5682 0.2477)

(59.3696 27.8696 42.5217 1.3239)

Πίνακες συσχέτισης αντίστοιχα:

R1 =

1.0000 0.7168 0.2419 0.2335

0.7168 1.0000 0.1556 0.1765

0.2419 0.1556 1.0000 0.3122

0.2335 0.1765 0.3122 1.0000

R2 =

1.0000 0.5944 0.7507 0.5483

0.5944 1.0000 0.6337 0.7142

0.7507 0.6337 1.0000 0.7863

0.5483 0.7142 0.7863 1.0000

2)

Οι συντελεστές της διακρίνουσας συνάρτησης δίνονται από το ιδιοδιάνυσμα που αντιστοιχεί στη μέγιστη ιδιοτιμή του πίνακα $W^{-1} * B$, δηλαδή το διάνυσμα:

u= (-0.0069 -0.0707 0.0778 0.9944)

Άρα η διακρίνουσα συνάρτηση είναι η u*(x*)

3)

Columns 1 through 25

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Columns 26 through 50

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

Columns 51 through 75

```
 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
```

Columns 76 through 90

```
 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
```

Το ποσοστό λανθασμένης ταξινόμησης είναι 0%.

4) Με βάση την ανάλυσή μας, τη νέα παρατήρηση την ταξινομούμε στο γκρουπ 2.

```
iris
X=iris_data(:,1:4); % observations
groups=iris_data(:,5); % which group each observation
belongs to
[n,p]=size(X); % n: sample size, p: number of
variables
uni=unique(groups); % groups that appear in the data
q=length(uni); % number of groups

xbar=sum(X)/n;
Xbar=repmat(xbar,[n,1]);
Xstar=X-Xbar; % centered observations
T=Xstar'*Xstar% matrix of total variation
Sp=T/(n-q) % pooled covariance matrix
cov(X) % sample covariance matrix
T/(n-1)

W=0; % W will be the matrix of variation within
groups
B=0; % B will be the matrix of variation between
groups
S_k=cell(1,q); % S_k will be a cell object containing
the group covariance matrices
m_k=cell(1,q); % m_k will be a cell object containing
the group means
nks=zeros(1,q);
for k=1:q
group_k=find(groups==k); % find observations that
belong to the k-th group
n_k=length(group_k); % number of observations in the
k-th group
nks(k)=n_k;
X_k=X(group_k,:);
xbar_k=sum(X_k)/n_k;
m_k{k}=xbar_k;
Xbar_k=repmat(xbar_k,[n_k,1]);
```

```
Xstar_k=X_k-Xbar_k; % centered observations in the k-
th group
    W=W+Xstar_k'*Xstar_k; % contribution to W from
the k-th group
    B=B+n_k*(xbar_k-xbar)'*(xbar_k-xbar);
S_k{k}=Xstar_k'*Xstar_k/(n_k-1);
end
W % matrix of variation within groups
B % B will be the matrix of variation between groups
W+B % W+B=T
T
m_k{1}
S_k{1}
D1=diag(diag(S_k{1}));
R1=D1^(-1/2)*S_k{1}*D1^(-1/2)
m_k{2}
S_k{2}
D2=diag(diag(S_k{2}));
R2=D2^(-1/2)*S_k{2}*D2^(-1/2)

A=inv(W)*B
[V,D]=eig(A) % eigenvalues and eigenvectors of the
matrix W^(-1)*B

r=min(q-1,p) % maximum number of discriminant
functions
discr_coef=V(:,2)' % one discriminant function since
there are two groups
% discr_coef=V(:,1:r)

% 1. Fisher's Rule
M_1=repmat(m_k{1},[n,1])-Xbar;
M_2=repmat(m_k{2},[n,1])-Xbar;

scores=discr_coef*Xstar' % the values that the
discriminant function takes
% at the observed sample

index=ones(1,n);
index(abs(scores-discr_coef*M_2')<abs(scores-
discr_coef*M_1'))=2

compare=(index==groups') % compare allocations with
true groups
prop=sum(compare==0)/n % proportion of incorrect
allocations
```

```
score_x=discr_coef*(x-xbar)' % the value that the
discriminant function takes
% at the new observation

m_1=m_k{1}-Xbar(1,:);
m_2=m_k{2}-Xbar(1,:);
index_x=1;
index_x(abs(score_x-discr_coef*m_2')<abs(score_x-
discr_coef*m_1'))=2
```

B)

AnalysisCaseProcessingSummary

UnweightedCases		N	Percent
Valid		391	100,0
Excluded	Missing or out-of-range group codes	0	,0
	At least one missing discriminating variable	0	,0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	,0
	Total	0	,0
Total		391	100,0

Από αυτόν τον πίνακα συμπεραίνουμε ότι χρησιμοποιούμε το 100% των παρατηρήσεών μας.

DescriptiveStatistics

	N	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
Lkm	391	-23,55	-5,05	-11,2241	,19471	3,85015	14,824
Displacement	391	1114,32	7456,13	3181,2558	86,71069	1714,59293	2939828,917
Horsepower	391	46	230	104,24	1,936	38,278	1465,220
Weight	391	725,85	2313,00	1337,8937	19,24888	380,62186	144873,004
Time to Accelerate from 0 to 60 mph (sec)	391	-25	-8	-15,53	,139	2,758	7,608
ModelYear (modulo 100)	391	70	82	75,99	,186	3,676	13,513
Valid N (listwise)	391						

Παρατηρούμε μεγάλες διαφορές στις τιμές των μέσων και των αποκλίσεων, πράγμα το οποίο δικαιολογείται από το γεγονός ότι τα χαρακτηριστικά, ως προς τα οποία μελετάμε τις παρατηρήσεις, έχουν διαφορετικές μονάδες μέτρησης.

GroupStatistics

Country of Origin		Mean	Std. Deviation	Valid N (listwise) Unweighted	Valid N (listwise) Weighted
American	lkm	-12,8642	3,77636	244	244,000
	Displacement	4052,2142	1614,32044	244	244,000
	Horsepower	118,7459	39,69618	244	244,000
	Weight	1515,1131	356,47906	244	244,000
	Time to Accelerate from 0 to 60 mph (sec)	-14,9758	2,73235	244	244,000
	ModelYear (modulo 100)	75,6148	3,65024	244	244,000
European	lkm	-8,9998	2,09517	68	68,000
	Displacement	1796,5563	371,87606	68	68,000
	Horsepower	80,5588	20,15787	68	68,000
	Weight	1095,0618	221,31624	68	68,000
	Time to Accelerate from 0 to 60 mph (sec)	-16,7941	3,08781	68	68,000
	ModelYear (modulo 100)	75,6765	3,42267	68	68,000
Japanese	lkm	-8,0729	1,77056	79	79,000
	Displacement	1683,1004	379,19956	79	79,000
	Horsepower	79,8354	17,81920	79	79,000
	Weight	999,5525	144,22376	79	79,000
	Time to Accelerate from 0 to 60 mph (sec)	-16,1722	1,95494	79	79,000
	ModelYear (modulo 100)	77,4430	3,65059	79	79,000
Total	lkm	-11,2241	3,85015	391	391,000
	Displacement	3181,2558	1714,59293	391	391,000
	Horsepower	104,2430	38,27820	391	391,000
	Weight	1337,8937	380,62186	391	391,000
	Time to Accelerate from 0 to 60 mph (sec)	-15,5338	2,75832	391	391,000
	ModelYear (modulo 100)	75,9949	3,67598	391	391,000

Παρατηρώντας τον πίνακα παραπάνω, διαπιστώνουμε ότι η Αμερική υπερταιρεί των άλλων δύο σε όλους τους τομείς εκτός του lkm, στον οποίο υστερεί, και στο modelyear, στο οποίο δε διαφέρει πολύ από την Ευρώπη και την Ιαπωνία. Να

σημειωθεί ότι το βάρος (weight) είναι παράγοντας δυσερμήνευτος, με την έννοια ότι δεν καθορίζει με κάποιο ξεκάθαρο τρόπο την απόδοση ενός αυτοκινήτου.
Η Ευρώπη κι η Ιαπωνία έχουν πολύ κοντά τις μέσες τιμές των χαρακτηριστικών τους. Αξίζει να σημειωθεί, επίσης, ότι η Αμερική έχει εμφανώς μεγαλύτερη απόκλιση από τις Ευρώπη κι Ιαπωνία στα περισσότερα χαρακτηριστικά (σε όλα εκτός του lkm).

CovarianceMatrices[a]

Country of Origin		lkm	Displacement	Horsepower	Weight	Time to Accelerate from 0 to 60 mph (sec)	ModelYear (modulo 100)
American	lkm	14,261	-5226,371	-122,880	-1170,443	-4,883	8,776
	Displacement	-5226,371	2606030,497	58019,971	529270,013	2758,856	-2926,926
	Horsepower	-122,880	58019,971	1575,787	11835,107	79,903	-70,880
	Weight	-1170,443	529270,013	11835,107	127077,324	441,737	-520,451
	Time to Accelerate from 0 to 60 mph (sec)	-4,883	2758,856	79,903	441,737	7,466	-3,911
	ModelYear (modulo 100)	8,776	-2926,926	-70,880	-520,451	-3,911	13,324
European	lkm	4,390	-448,518	-31,171	-283,710	-1,488	2,769
	Displacement	-448,518	138291,805	4662,979	73377,397	-43,671	263,395
	Horsepower	-31,171	4662,979	406,340	2728,593	33,885	-9,130
	Weight	-283,710	73377,397	2728,593	48980,878	-113,664	134,969
	Time to Accelerate from 0 to 60 mph (sec)	-1,488	-43,671	33,885	-113,664	9,535	-1,850
	ModelYear (modulo 100)	2,769	263,395	-9,130	134,969	-1,850	11,715
Japanese	lkm	3,135	-225,945	-21,527	-143,957	-1,462	3,575
	Displacement	-225,945	143792,306	4933,823	46017,458	397,039	167,272
	Horsepower	-21,527	4933,823	317,524	2229,663	25,087	-14,003
	Weight	-143,957	46017,458	2229,663	20800,493	160,057	23,885
	Time to Accelerate from 0 to 60 mph (sec)	-1,462	397,039	25,087	160,057	3,822	,007

	ModelYear (modulo 100)	3,575	167,272	-14,003	23,885	,007	13,327
Total	lkm	14,824	-5765,659	-125,870	-1298,793	-5,059	7,850
	Displacement	-5765,659	2939828,917	58966,689	609458,753	2592,939	-2314,730
	Horsepower	-125,870	58966,689	1465,220	12572,470	74,048	-57,827
	Weight	-1298,793	609458,753	12572,470	144873,004	446,699	-424,233
	Time to Accelerate from 0 to 60 mph (sec)	-5,059	2592,939	74,048	446,699	7,608	-3,004
	ModelYear (modulo 100)	7,850	-2314,730	-57,827	-424,233	-3,004	13,513

a. The total covariance matrix has 390 degrees of freedom.

PooledWithin-GroupsMatrices[a]

		lkm	Displacement	Horsepower	Weight	Time to Accelerate from 0 to 60 mph (sec)	ModelYear (modulo 100)
Covariance	lkm	10,320	-3396,089	-86,669	-810,966	-3,609	6,693
	Displacement	-3396,089	1684914,334	38134,306	353397,578	1800,116	-1753,990
	Horsepower	-86,669	38134,306	1120,897	8331,599	60,937	-48,783
	Weight	-810,966	353397,578	8331,599	92226,667	289,204	-297,845
	Time to Accelerate from 0 to 60 mph (sec)	-3,609	1800,116	60,937	289,204	7,090	-2,768
	ModelYear (modulo 100)	6,693	-1753,990	-48,783	-297,845	-2,768	13,047
Correlation	lkm	1,000	-,814	-,806	-,831	-,422	,577
	Displacement	-,814	1,000	,877	,896	,521	-,374
	Horsepower	-,806	,877	1,000	,819	,684	-,403
	Weight	-,831	,896	,819	1,000	,358	-,272
	Time to Accelerate from 0 to 60 mph (sec)	-,422	,521	,684	,358	1,000	-,288
	ModelYear (modulo 100)	,577	-,374	-,403	-,272	-,288	1,000

a. The covariance matrix has 388 degrees of freedom.

ΣτονPooledWithin-Groupsπίνακασυνδιακύμανσης, μπορούμε πάλι να παρατηρήσουμε μεγάλες διαφορές στις τιμές των στοιχείων του, το οποίο είναι λογικό, εφόσον τα χαρακτηριστικά διαφέρουν ως προς τις μονάδες μέτρησής τους.

Όσον αφορά τις συσχετίσεις, βλέπουμε πως όλα τα χαρακτηριστικά έχουν σημαντικές συσχετίσεις μεταξύ τους, εκτός του Modelyear (το οποίο έχει εμφανώς χαμηλές συσχετίσεις με τα υπόλοιπα) και του TimetoAcceleratefrom 0 to 60 mph (το οποίο έχει χαμηλές συσχετίσεις με όλα τα άλλα χαρακτηριστικά εκτός του Horsepower).

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
lkm	,693	86,107	2	388	,000
Displacement	,570	146,235	2	388	,000
Horsepower	,761	60,901	2	388	,000
Weight	,633	112,313	2	388	,000
Time to Accelerate from 0 to 60 mph (sec)	,927	15,244	2	388	,000
ModelYear (modulo 100)	,961	7,965	2	388	,000

Από τη στήλη sig του πίνακα, testsofEqualityofGroupMeans, βλέπουμε τα p-values από τον εξής έλεγχο υποθέσεων :

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6$ *vs* $H_1: \exists\, i, j \in \{1,2 \dots 6\} \mu_i: \neq \mu_j$

Τα p-values μας δείχνουν την πιθανότητα να παρατηρήσουμε κάτι ακραίο, υπό την υπόθεση H_0. Εδώ τα p-values είναι πολύ μικρά(0) , άρα απορρίπτεται η $H_{0,}$ που σημαίνει ότι έχουμε διαφορετικά μ_i δηλαδή ο πληθυσμός μας διαχωρίζεται από τις μεταβλητές μας.Τελικά τα δεδομένα μας είναι κατάλληλα για χρήση διακρίνουσας ανάλυσης.

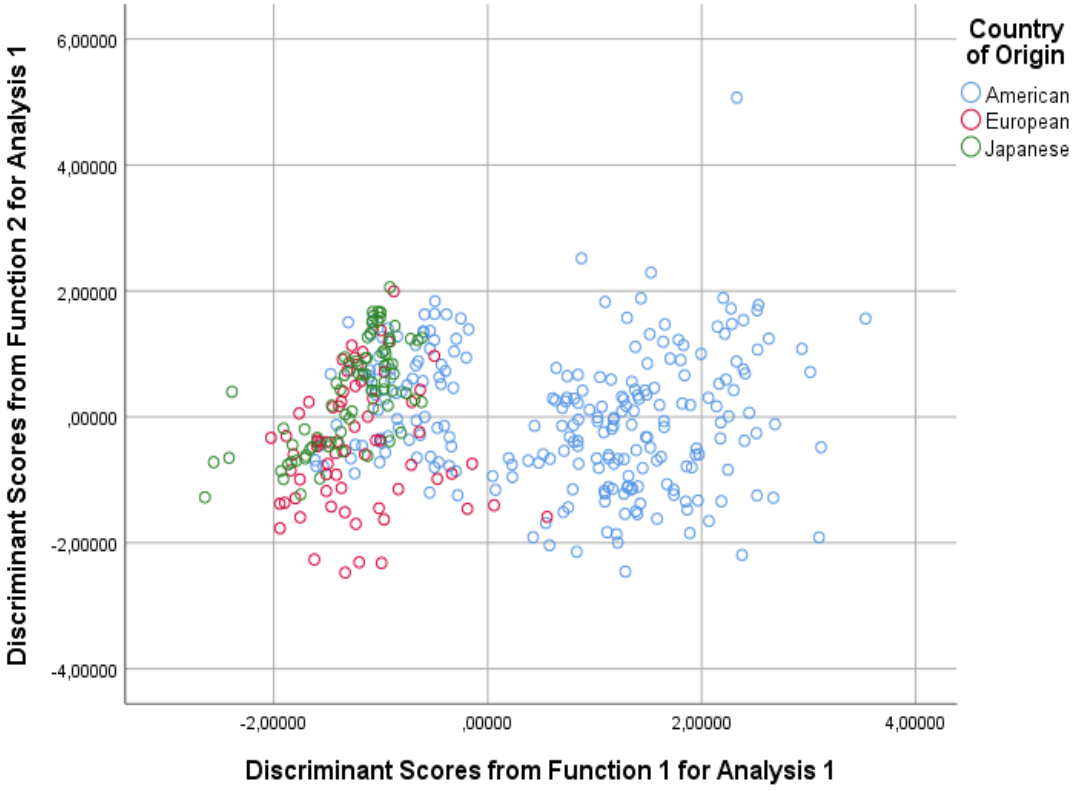
2)

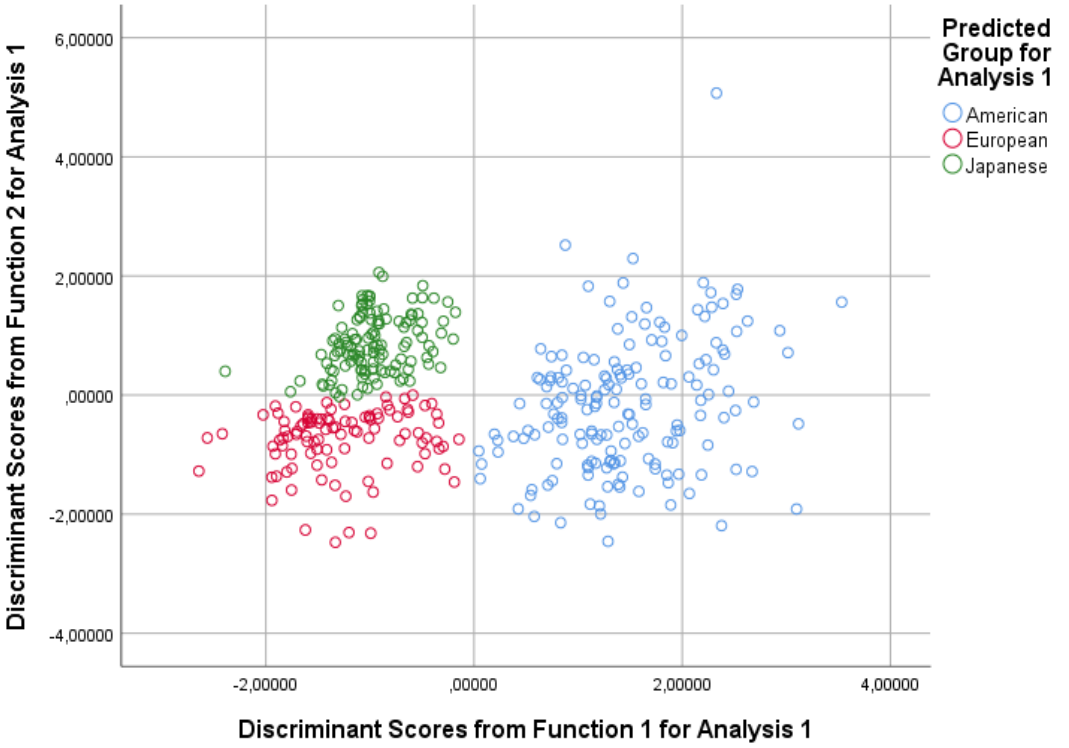
Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,948[a]	93,2	93,2	,698
2	,069[a]	6,8	100,0	,254

a. First 2 canonical discriminant functions were used in the analysis.

Στον παραπάνω πίνακα φαίνονται οι ιδιοτιμές που χρησιμοποιήθηκαν για τον υπολογισμό των δυο διακρινουσών συναρτήσεων.





Στο πρώτο γράφημα, βλέπουμε τα Scores χρωματισμένα σύμφωνα με το CountryofOriginκαι ύστερα στο δεύτερο γράφημα σύμφωνα με τις ομάδες που έχουν προβλεφθεί από την Διακρίνουσα. Αυτό σημαίνει ότι σε ενδεχόμενη προσπάθεια ταξινόμησης μιας νέας παρατήρησης, θα γίνει σύμφωνα με αυτό το διάγραμμα.
Συνεχίζοντας την ερμηνεία των διαγραμμάτων, άξιο μνείας θεωρείται το γεγονός ότι οι τρεις ομάδες που παρατηρούνται διαφέρουν στο πόσο απλωμένες είναι. Αυτό οφείλεται στις διαφορετικές τιμές των διασπορών που έχουν οι χώρες/ήπειροι και μια μεγαλύτερη διασπορά αυτόματα συνεπάγεται μεγαλύτερη ποικιλία αυτοκινήτων ως προς τα χαρακτηριστικά τους. Επίσης από την θέση των παρατηρήσεων στο γράφημα βλέπουμε ότι τα αμερικάνικα αυτοκίνητα είναι πιο ποιοτικά από τα υπόλοιπα, ενώ τα ιαπωνικά και τα ευρωπαικά είναι κοντά σε ποιότητα .Στο δεύτερο διάγραμμα, το οποίο είναι πιο «καθαρό», πρέπει να αναφέρουμε ότι υπάρχουν παρατηρήσεις, οι οποίες είναι ταξινομημένες σε διαφορετικό CountryofOrigin. Αυτό οφείλεται στη διακρίνουσα ανάλυση που κάναμε και συνήθως αυτές οι παρατηρήσεις είναι οι «οριακές», δηλαδή αυτές που παρατηρώντας το πρώτο γράφημα βλέπουμε ότι βρίσκονται κοντά στο κέντρο βάρους μιας άλλης ομάδας από αυτήν που είναι στην πραγματικότητα . Μάλιστα, ποσοστιαία και πιο αναλυτικά ανά την CountryofOriginφαίνεται από τον παρακάτω πίνακα.

ClassificationResults[a,c]

			PredictedGroupMembership			
		Country of Origin	American	European	Japanese	Total

Original	Count	American	166	32	46	244
		European	2	43	23	68
		Japanese	0	26	53	79
	%	American	68,0	13,1	18,9	100,0
		European	2,9	63,2	33,8	100,0
		Japanese	,0	32,9	67,1	100,0
Cross-validated[b]	Count	American	166	32	46	244
		European	3	42	23	68
		Japanese	0	26	53	79
	%	American	68,0	13,1	18,9	100,0
		European	4,4	61,8	33,8	100,0
		Japanese	,0	32,9	67,1	100,0

a. 67,0% of original grouped cases correctly classified.

b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

c. 66,8% of cross-validated grouped cases correctly classified.

Χαρακτηριστικά, από τα συνολικά 244 Αμερικάνικα αυτοκίνητα, τα 166 ταξινομήθηκαν ορθώς(67%), ενώ τα υπόλοιπα 32+46=78 ταξινομήθηκαν λανθασμένα σε Ευρώπη κι Ιαπωνία αντίστοιχα. Ακολουθώντας αυτόν τον τρόπο ερμηνεύουμε και τα υπόλοιπα στοιχεία του πίνακα.

FunctionsatGroupCentroids

	Function	
Country of Origin	1	2
American	,753	,003
European	-1,222	-,465
Japanese	-1,273	,390

Unstandardized canonical discriminant functions evaluated at group means

Ο παραπάνω πίνακας δίνει τα τρία κέντρα βάρους των τριών ομάδων, CountriesofOrigin, τα οποία μπορούμε να επιβεβαιώσουμε και διαισθητικά από το γράφημά μας πιο πάνω.

3)

Variables in the Analysis

Step		Tolerance	F toRemove	Wilks' Lambda
1	Displacement	1,000	146,235	
2	Displacement	,230	90,945	,761
	Horsepower	,230	19,604	,570
3	Displacement	,230	91,301	,736
	Horsepower	,223	17,188	,544
	ModelYear (modulo 100)	,836	6,860	,518
4	Displacement	,131	47,918	,600
	Horsepower	,216	17,514	,524
	ModelYear (modulo 100)	,807	9,078	,503
	Weight	,185	7,847	,500

Στον παραπάνω πίνακα, έχοντας χρησιμοποιήσει stepwisemethodγια την ανάλυσή μας, βλέπουμε ότι αρκούν οι 4 μεταβλητές Displacement, Horsepower, ModelYear, Weightγια την ανάλυσή μας, δηλαδή για την ταξινόμηση μιας νέας παρατήρησης σε κάποια ομάδα CountryofOrigin. Αυτό μπορούμε να το δούμε και καλύτερα στον επόμενο πίνακα, στον οποίο από 6 μεταβλητές που έχουμε στην αρχή, σε τεσσερα βήματα αφαιρούνται 2 μεταβλητές, ώστε να καταλήξουμε στο προαναφερθέν συμπέρασμα. Αυτές οι 2 μεταβλητές που αφαιρέθηκαν ή δε συμβάλλουν το ίδιο πολύ στην ανάλυση μας, και κατά συνέπεια στην πληροφόρησή μας, ή η πληροφορία που παρέχουν καλύπτεται μέχρι ένα βαθμό επαρκώς από τις υπόλοιπες 4 μεταβλητές.

Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	F toEnter	Wilks' Lambda
0	lkm	1,000	1,000	86,107	,693
	Displacement	1,000	1,000	146,235	,570
	Horsepower	1,000	1,000	60,901	,761
	Weight	1,000	1,000	112,313	,633
	Time to Accelerate from 0 to 60 mph (sec)	1,000	1,000	15,244	,927
	ModelYear (modulo 100)	1,000	1,000	7,965	,961
1	lkm	,337	,337	3,311	,561
	Horsepower	,230	,230	19,604	,518
	Weight	,196	,196	5,523	,554
	Time to Accelerate from 0 to 60 mph (sec)	,729	,729	6,799	,551
	ModelYear (modulo 100)	,860	,860	9,133	,544
2	lkm	,301	,197	3,764	,508
	Weight	,192	,134	5,639	,503

	Time to Accelerate from 0 to 60 mph (sec)	,506	,160	1,892	,513
	ModelYear (modulo 100)	,836	,223	6,860	,500
3	lkm	,230	,192	2,802	,493
	Weight	,185	,131	7,847	,480
	Time to Accelerate from 0 to 60 mph (sec)	,505	,157	2,027	,495
4	lkm	,165	,131	3,617	,472
	Time to Accelerate from 0 to 60 mph (sec)	,399	,128	,078	,480

Structure Matrix

	Function	
	1	2
Displacement	,892*	-,049
Weight	,777*	-,322
lkm[b]	-,606*	,382
Horsepower	,576*	,009
Model Year (modulo 100)	-,142	,564*
Time to Accelerate from 0 to 60 mph (sec)[b]	,261	,266*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions

Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

b. This variable not used in the analysis.

Στον παραπάνω πίνακα φαίνεται η συσχέτιση κάθε χαρακτηριστικού με τις δυο διακρίνουσες συναρτήσεις. Για παράδειγμα το χαρακτηριστικό displacement, σχετίζεται πολύ ισχυρά και με την πρώτη διακρίνουσα, ενώ τα χαρακτηριστικά accel και lkm έχουν συσχέτιση και με τις δύο διακρίνουσες.

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
Displacement	1,712	1,223
Horsepower	-,848	,758

Weight	-,023	-1,812
Model Year (modulo 100)	,150	,835

Στον παραπάνω πίνακα εμφανίζονται οι συντελεστές των τεσσάρων αυτών χαρακτηριστικών για τον υπολογισμό των σκορ.

ΑΛΕΞΑΝΔΡΟΣ ΣΚΟΝΔΡΑΣ (1112201500206)
ΔΗΜΗΤΡΙΟΣ ΓΚΑΒΕΡΑΣ (1112201500042)