ΤΕΙ ΠΕΙΡΑΙΑ

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ

ΕΠΙΣΤΗΜΗ ΤΩΝ ΑΠΟΦΑΣΕΩΝ ΜΕ ΠΛΗΡΟΦΟΡΙΑΚΑ ΣΥΣΤΗΜΑΤΑ

ΜΑΘΗΜΑ:

Ανάλυση Πολυδιάστατων (Πολυμεταβλητών) Δεδομένων και Συστήματα Εξόρυξης Δεδομένων (Multivariate Data Analysis and Data Mining Systems)

Θέμα εργασίας: **Ατομική Εργασία 2η**

Ονοματεπώνυμο φοιτητή: **ΧΟΥΡΔΑΚΗΣ ΕΥΣΤΡΑΤΙΟΣ** (Α.Μ.1317)

Επιβλέπουσα καθηγήτρια: Μοσχονά θ.

Ιούνιος 2014

Πίνακας περιεχομένων

Α. Παραγοντική Ανάλυση

Τα απαιτούμενα βήματα για την εφαρμογή της παραγοντικής ανάλυσης είναι:

1. Έλεγχος για το αν υπάρχουν συσχετίσεις ικανοποιητικές για να κάνουμε παραγοντική ανάλυση.
2. Εύρεση του αριθμού των παραγόντων και εκτίμηση των παραμέτρων του μοντέλου
3. Περιστροφή του μοντέλου με σκοπό να αυξήσουμε την ερμηνευτική του ικανότητα
4. Εκτίμηση των scores των παραγόντων για περαιτέρω στατιστική χρήση

Η μέθοδος της παραγοντικής ανάλυσης που θα εφαρμόσουμε είναι **η μέθοδος των κύριων συνιστωσών (principal components)**. Η μέθοδος αυτή λαμβάνει υπόψη τη συνολική διασπορά των μεταβλητών κατά φθίνουσα ακολουθία. Δηλαδή, η πρώτη κύρια συνιστώσα είναι ο γραμμικός συνδυασμός των αρχικών μεταβλητών με την μεγαλύτερη ποσότητα διασποράς στο δείγμα. Η δεύτερη κύρια συνιστώσα, η οποία είναι ασυσχέτιστη με την πρώτη, ερμηνεύει την υπόλοιπη μεγαλύτερη ποσότητα διασποράς, κ.τ.λ.

Για την ορθή εφαρμογή της εν λόγω μεθόδου απαιτείται οι διασπορές (ή ομοίως οι διακυμάνσεις) των μεταβλητών να μην διαφέρουν έντονα μεταξύ τους.

Τον έλεγχο των διασπορών των μεταβλητών τον κάνουμε μέσω της διαδικασίας: Analyze > Case Summaries, στο πεδίο Variables θέτουμε όλες της παρατηρούμενες μεταβλητές και στο Statistics επιλέγουμε Std Deviation και Variance.

Από τον **Πίνακα 1** διαπιστώνουμε ότι η τάξη των διασπορών (και συνακόλουθα των διακυμάνσεων) των μεταβλητών διαφέρει έντονα.

Πίνακας 1

Case Summaries

	N	Std. Deviation	Variance
Συνολικά Έσοδα	150	93412,39862	8725876215,786
Δαπάνες Διαφήμισης	150	3103,70587	9632990,103
Δαπάνες Ενοικίων	150	10248,12059	105023975,545
Δαπάνες για Προσωπικό	150	9981,14477	99623250,845
Αξιολόγηση από Περιοδικό	150	10,319	106,472
Αξιολόγηση από Πελάτες	150	1,169	1,368

Για το λόγο αυτό θα εφαρμόσουμε την παραγοντική ανάλυση στις **τυποποιημένες τιμές** των μεταβλητών ώστε η διασπορά (ή ομοίως, η διακύμανση) των τιμών της κάθε μεταβλητής να είναι μονάδα.

Σημειώνεται ότι η μέθοδος principal components μάς επιτρέπει να χρησιμοποιήσουμε τις αρχικές τιμές των μτβ (δηλ. όχι τις Z τιμές), αρκεί στο Extraction και στο πεδίο Analyze να επιλέξουμε Correlation Matrix. Τα αποτελέσματα της Factor Analysis στην περίπτωση αυτή είναι ακριβώς τα ίδια με εκείνα που θα λάβουμε χρησιμοποιώντας εξ' αρχής τις Z τιμές.

Από τον **Πίνακα 2** επιβεβαιώνεται ότι οι τιμές διασποράς των τυποποιημένων τιμών των μεταβλητών μας είναι μονάδα.

Πίνακας 2

Descriptive Statistics

Variables	N	Minimum	Maximum	Mean	Std. Deviation	Variance
ZΣυνολικά Έσοδα	150	-1,866	2,068	,0	**1,0**	**1,0**
ZΔαπάνες Διαφήμισης	150	-1,854	2,040	,0	**1,0**	**1,0**
ZΔαπάνες Ενοικίων	150	-1,686	2,000	,0	**1,0**	**1,0**
ZΔαπάνες για Προσωπικό	150	-2,013	2,203	,0	**1,0**	**1,0**
ZSCORE1 Αξιολόγηση από Περιοδικό	150	-2,366	2,770	,0	**1,0**	**1,0**
ZSCORE2 Αξιολόγηση από Πελάτες	150	-2,600	3,386	,0	**1,0**	**1,0**
Valid N (listwise)	150					

Ο έλεγχος των συσχετίσεων γίνεται με τη βοήθεια του SPSS ακολουθώντας τη διαδρομή: **Analyze> Dimension Reduction > Factor** τοποθετούμε τις τυποποιημένες τιμές των μεταβλητών στο variable και επιλέγουμε κατά περίπτωση:
Descriptives: Univariate discriptives, Initial solution, coefficients, reproduced, anti-image, KMO Batrlett's test
Extraction: Method principal components, correlation matrix, unrotated factor solution, screeplot, eigenvalues Over 1
Rotation: varimax, rotated solution, loading plots
Scores: Save as variables, method regression, display factor score coefficient matrix
Options: suppress absolute values less than 0,30

Ο **Πίνακας 3** (Correlation Matrix) εμφανίζει τους συντελεστές συσχέτισης μεταξύ των μεταβλητών.

Πίνακας 3

Correlation Matrix

		ZSALES	ZADV	ZRENT	ZPERS	ZSCORE1	ZSCORE2
Correlation	ZSALES	1,000	**,997**	**,915**	**,954**	-,006	-,015
	ZADV	**,997**	1,000	**,913**	**,951**	,003	-,010
	ZRENT	**,915**	**,913**	1,000	**,881**	-,035	-,044
	ZPERS	**,954**	**,951**	**,881**	1,000	-,033	-,029
	ZSCORE1	-,006	,003	-,035	-,033	1,000	**,854**
	ZSCORE2	-,015	-,010	-,044	-,029	**,854**	1,000

Από τον ανωτέρω πίνακα συνάγεται ότι υπάρχει έντονη συσχέτιση μεταξύ των μτβ «Συνολικά Έσοδα», «Συνολικές Δαπάνες για Διαφήμιση», «Συνολικές Δαπάνες για Ενοίκια» και «Συνολικές Δαπάνες για Προσωπικό», ενώ αντίστοιχα υψηλή συσχέτιση εμφανίζουν οι μτβ «Αξιολόγηση από Περιοδικό» και «Αξιολόγηση από δείγμα Πελατών». Αντίθετα, κάθε μία από τις μτβ αξιολόγησης παρουσιάζουν χαμηλή συσχέτιση με τις μεταβλητές που σχετίζονται με τα χρηματο-οικονομικά δεδομένα των επιχειρήσεων.
Από τον **Πίνακα 4** λαμβάνουμε τον δείκτη **Kaiser-Meyer-Olkin (KMO)**, που συγκρίνει τα

μεγέθη των παρατηρούμενων συντελεστών συσχέτισης προς τους συντελεστές μερικής συσχέτισης. Μικρές τιμές του δείκτη δηλώνουν ότι η παραγοντική ανάλυση δεν είναι κατάλληλη τεχνική για τα δεδομένα. **Στη συγκεκριμένη περίπτωση ο δείκτης ΚΜΟ λαμβάνει τιμή 0,773 και συνεπώς συμπεραίνουμε ότι το δείγμα μας είναι κατάλληλο για να εφαρμοστεί η παραγοντική ανάλυση**.

Πίνακας 4

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		**,773**
Bartlett's Test of Sphericity	Approx. Chi-Square	1543,607
	df	15
	Sig.	,000

Από τον ίδιο Πίνακα (Πιν. 4) διεξάγουμε τον έλεγχο σφαιρικότητας του **Bartlett**.
Διατυπώνουμε την μηδενική υπόθεση και ελέγχουμε την ισχύ της σε επίπεδο στατιστικής σημαντικότητας α=5%:
H_0: Τα δεδομένα είναι ένα δείγμα από ένα πολυμεταβλητό κανονικό πληθυσμό όπου όλοι οι συντελεστές συσχέτισης είναι ίσοι με το μηδέν,
H_1: Τα δεδομένα είναι ένα δείγμα από ένα πολυμεταβλητό κανονικό πληθυσμό όπου όλοι οι συντελεστές συσχέτισης είναι διάφοροι από το μηδέν.

Επειδή p-value = 0,0001 < 0,05 απορρίπτουμε τη μηδενική υπόθεση και δεχόμαστε την εναλλακτική ότι <u>οι συντελεστές συσχέτισης είναι διάφοροι του μηδενός</u>.

Πίνακας 5

Total Variance Explained

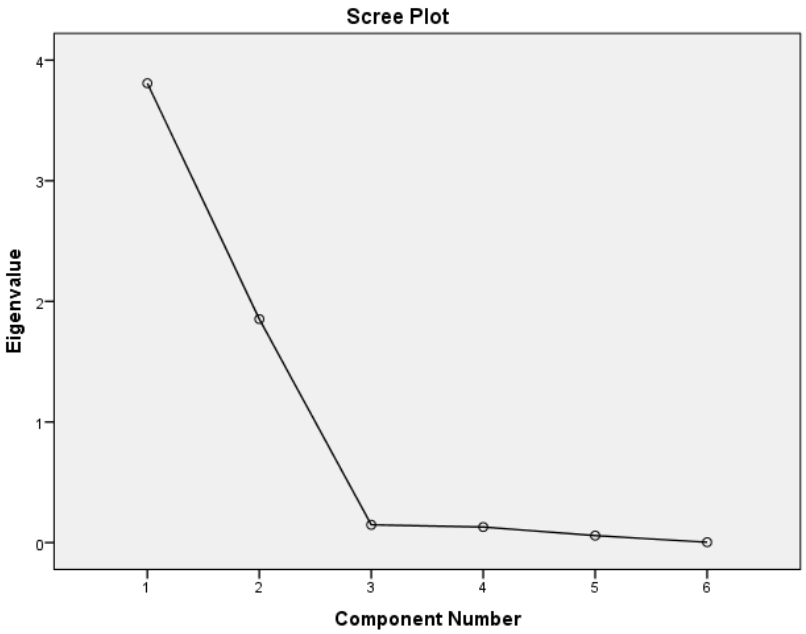
Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,808	63,471	63,471	3,808	63,471	63,471	3,806	63,437	63,437
2	1,853	30,888	94,359	1,853	30,888	94,359	1,855	30,921	94,359
3	,148	2,461	96,819						
4	,130	2,160	98,979						
5	,058	,966	99,945						
6	,003	,055	100,000						

Extraction **Method: Principal Component Analysis.**

Από τον **Πίνακα 5** (Total Variance Explained) διαπιστώνουμε ότι ο πρώτος παράγοντας ερμηνεύει το **63,47%** της συνολικής διασποράς των έξι μεταβλητών, ενώ ο δεύτερος παράγοντας ερμηνεύει το **30,89%.** Αθροιστικά, οι δύο πρώτοι παράγοντες ερμηνεύουν το **94,36%** της συνολικής διασποράς των έξι εξεταζόμενων μεταβλητών.

Ένας δεύτερος τρόπος προσδιορισμού του αριθμού των παραγόντων που θα χρησιμοποιήσουμε είναι μέσω του **Γραφήματος 1** (Scree Plot).

Γράφημα 1



Από το ανωτέρω γράφημα διαπιστώνεται ότι μόνο οι δύο πρώτοι παράγοντες έχουν ιδιοτιμές (δηλ. διασπορά των συνιστωσών) μεγαλύτερες της μονάδας ή αλλιώς μετά τον δεύτερο παράγοντα παρατηρείται τάση ευθυγράμμισης της γραμμής που ενώνει τις ιδιοτιμές με τους επόμενους υποψήφιους παράγοντες.
Συνοπτικά, από τα ανωτέρω (Πίνακας 5 και Γράφημα 1), επιλέγουμε να χρησιμοποιήσουμε στην παραγοντική μας ανάλυση δύο παράγοντες (συνιστώσες).

Από τον **Πίνακα 6** φαίνεται ότι το μοντέλο των δύο παραγόντων περιγράφει πολύ καλά τις αρχικές μεταβλητές, καθόσον η διασπορά της κάθε μεταβλητής που ερμηνεύεται από τις δύο συνιστώσες πλησιάζει την μονάδα.

Πίνακας 6

Communalities

	Initial	Extraction
ZSALES: Συνολικά Έσοδα	1,000	,983
ZADV: Δαπάνες Διαφήμισης	1,000	,980
ZRENT: Δαπάνες Ενοικίων	1,000	,903
ZPERS: Δαπάνες για Προσωπικό	1,000	,942
ZSCORE1: Αξιολόγηση από Περιοδικό	1,000	,927
ZSCORE2 : Αξιολόγηση από Πελάτες	1,000	,927

Extraction Method: Principal Component Analysis.

Στον **Πίνακα 7** φαίνονται οι παραγοντικές φορτίσεις μετά την ορθογώνια περιστροφή (varimax) των αρχικών παραγοντικών φορτίσεων.

Πίνακας 7

Rotated Component Matrix[a]

	Component	
	1	2
ZSALES: Συνολικά Έσοδα	,991	
ZADV: Δαπάνες Διαφήμισης	,990	
ZPERS: Δαπάνες για Προσωπικό	,970	
ZRENT: Δαπάνες Ενοικίων	,950	
ZSCORE1: Αξιολόγηση από Περιοδικό		,963
ZSCORE2: Αξιολόγηση από Πελάτες		,963

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 3 iterations.

Από τον ανωτέρω Πίνακα φαίνεται ότι τα *Συνολικά Έσοδα* της επιχείρησης, οι *Δαπάνες Διαφήμισης*, οι *Δαπάνες για το Προσωπικό* και οι *Δαπάνες Ενοικίων* είναι πολύ υψηλά συσχετισμένες με τον **1ο Παράγοντα**, ενώ η *Αξιολόγηση από περιοδικό* και η *Αξιολόγηση από δείγμα πελατών* εμφανίζουν υψηλή συσχέτιση με τον **2ο Παράγοντα**.
Τον 1ο Παράγοντα μπορούμε να τον ονομάσουμε **Χρηματοοικονομικά Δεδομένα** της Επιχείρησης και τον 2ο Παράγοντα **Εξωτερική Αξιολόγηση** της Επιχείρησης.

Πίνακας 8

Component Score Coefficient Matrix

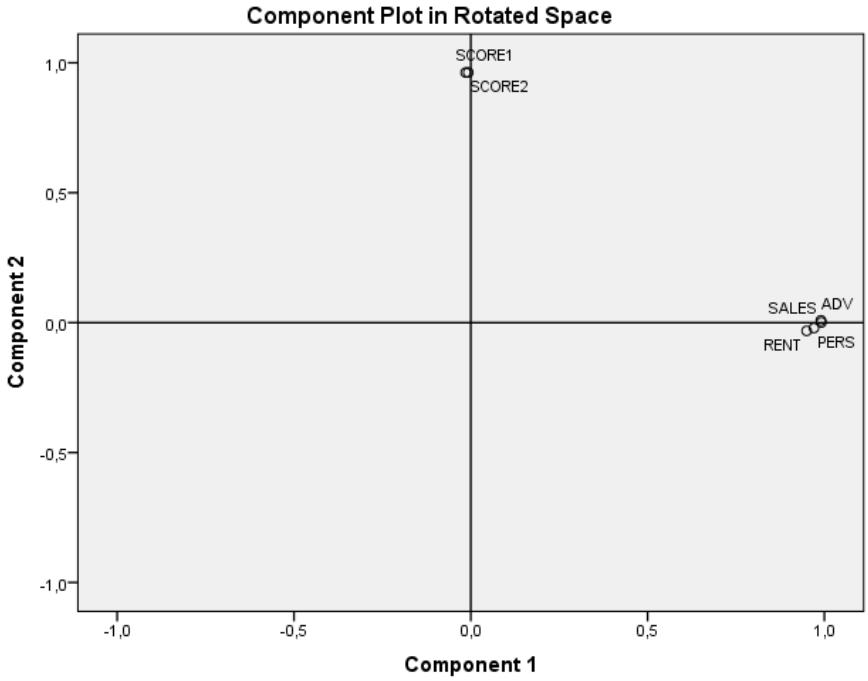
	Component	
	1	2
ZSALES: Συνολικά Έσοδα	,261	,009
ZADV: Δαπάνες Διαφήμισης	,260	,013
ZRENT: Δαπάνες Ενοικίων	,249	-,009
ZPERS: Δαπάνες για Προσωπικό	,255	-,003
ZSCORE1: Αξιολόγηση από Περιοδικό	,007	,519
ZSCORE2: Αξιολόγηση από Πελάτες	,005	,519

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Component Scores.

Οι συντελεστές των παραγοντικών score δίνονται από τον **Πίνακα 8**. Ειδικότερα, ο 1ο παράγοντας = 0,261*SALES + 0,260*ADV + 0,249*RENT + 0,255*PERS + 0,007*SCORE1 + 0,005*SCORE2 και ο 2ος παράγοντας= 0,009*SALES + 0,013*ADV - 0,009*RENT - 0,003*PERS + 0,519*SCORE1 + 0,519*SCORE2.

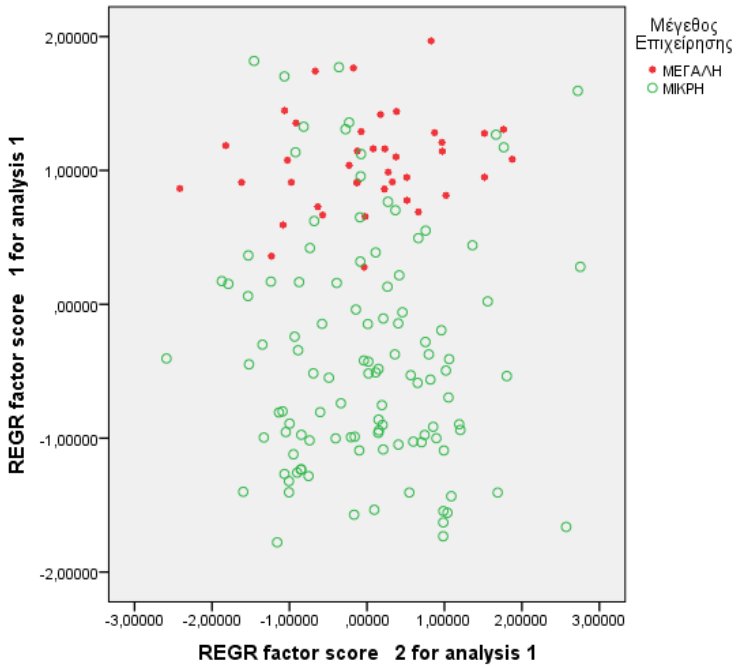
Στο **Γράφημα 2** δίνεται το διάγραμμα διάταξης των μεταβλητών χρησιμοποιώντας τις δύο κύριες συνιστώσες.

Γράφημα 2



Για να έχουμε μια εικόνα της κατανομής των παρατηρήσεων των παραγόντων σε σχέση με το μέγεθος των επιχειρήσεων κάνουμε ένα scatter plot διάγραμμα:

Γράφημα 3



Συμπεράσματα.

Μετά την ανάλυση των δεδομένων σε κύριες συνιστώσες καταλήξαμε στην αντιπροσώπευση των έξι μεταβλητών από δύο κύριες συνιστώσες.

Η πρώτη κύρια συνιστώσα, την οποία ονομάσαμε «**Χρηματοοικονομικά δεδομένα**», αντιπροσωπεύει τα συνολικά έσοδα των επιχειρήσεων και κάποιες από τις βασικότερες δαπάνες (Ενοικίων, Προσωπικού και Διαφήμισης) τους.

Τη δεύτερη κύρια συνιστώσα την ονομάσαμε «**Εξωτερική αξιολόγηση**» και αφορά στην

αξιολόγηση των επιχειρήσεων από ένα σχετικό περιοδικό και από ένα δείγμα τριάντα πελατών της κάθε επιχείρησης.

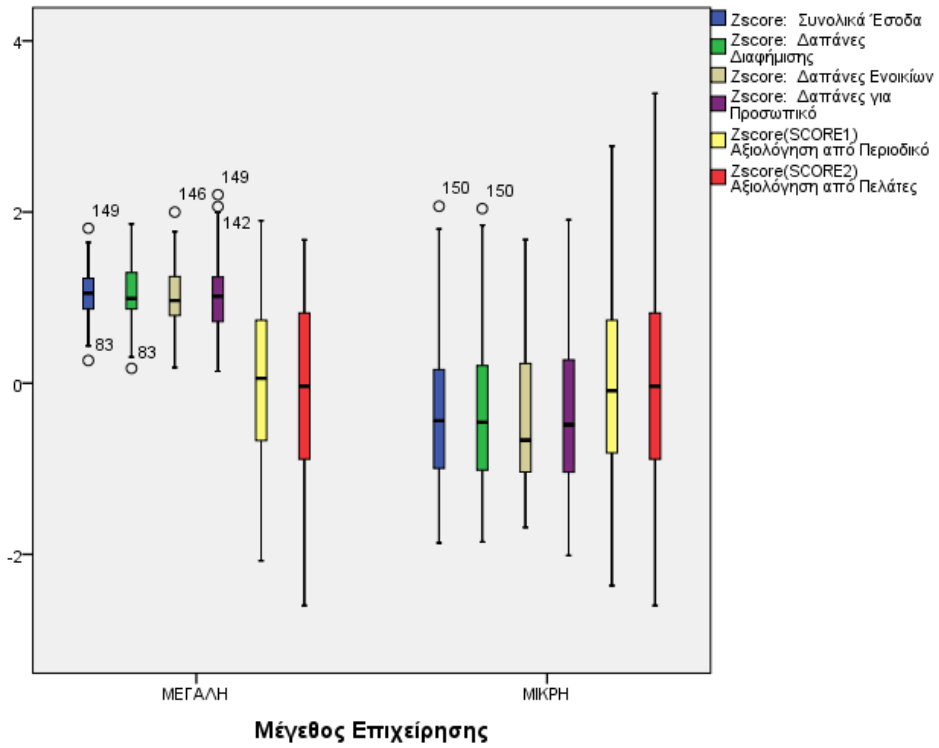
Β. Ανάλυση κατά Συστάδες

1. Μέθοδος K-Means

Η μέθοδος που αρχικά επιλέγουμε για να προβούμε σε ανάλυση κατά Συστάδες, είναι η K-Means. Επειδή οι μονάδες των μεταβλητών μας μετρούν διαφορετικά μεγέθη (χρηματικές μονάδες και βαθμούς αξιολογικής κατάταξης) και διαφέρουν αισθητά ως προς την τάξη μεγέθους τους, και επειδή η μέθοδος K-Means μετρά αποστάσεις μεταξύ των τιμών των μεταβλητών, θα χρησιμοποιήσουμε τις τυπικοποιημένες τιμές των μεταβλητών μας.

Επιπλέον, επειδή η μέθοδος K-Means Clustering είναι ευαίσθητη στις ακραίες τιμές των μτβ, γιατί συνήθως αυτές επιλέγονται ως initial Cluster centers, θα κάνουμε ένα Box Plot διάγραμμα για να ελέγξουμε αν υπάρχουν τέτοιες τιμές στο δείγμα μας.

Γράφημα 4



Από το **Γράφημα 4** δεν φαίνεται να υπάρχουν ακραίες τιμές ικανές να καθορίσουν αναπόδραστα τη μέση απόσταση της συστάδας που θα ενταχθούν ή να μας υποχρεώσουν να δημιουργήσουμε συστάδες με μικρό αριθμό cases που θα τις συμπεριλάβει.

Δοκιμή με Δύο Συστάδες

Αρχικά επιλέγουμε να ξεκινήσουμε την ανάλυσή μας χρησιμοποιώντας **δύο Συστάδες**.
Ακολουθώντας τη διαδρομή Analyze>Classify>K-Means Cluster, θέτουμε στα Variables: ZSALES, ZADV, ZRENT, ZPERS, ZSCORE1 και ZSCORE2 και επιλέγουμε Number of Clusters = 2, λαμβάνουμε τον **Πίνακα 9**.

Πίνακας 9

Initial Cluster Centers

	Cluster	
	1	2
ZSALES	-1,69007	1,80347
ZADV	-1,67608	1,84794
ZRENT	-1,38652	1,58657
ZPERS	-1,83040	1,90880
ZSCORE1	1,60746	-1,10610
ZSCORE2	3,38633	-1,74447

Πίνακας 10

Iteration History[a]

Iteration	Change in Cluster Centers	
	1	2
1	3,709	2,805
2	,186	,269
3	,166	,194
4	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. **The current iteration is 4**. The minimum distance between initial centers is 9,007.

Από τον **Πίνακα 9** λαμβάνουμε τις αρχικές τιμές των μέσων των Ζτιμών των μεταβλητών μας για κάθε μία από τις Συστάδες. Από τον Πίνακα αυτόν φαίνεται το αρχικό σημείο αναφοράς από το οποίο θα αρχίσουν οι επαναλήψεις με σκοπό την ομογενοποίηση των Συστάδων. Το κάθε case κατατάσσεται σε μια από τις δύο Συστάδες βάσει της απόστασης της κάθε μτβ από το αρχικό κέντρο της Συστάδας. Μετά την κατάταξη όλων των cases στις Συστάδες, επαν-υπολογίζονται τα νέα κέντρα των Συστάδων. Η διαδικασία επαναλαμβάνεται μέχρι να ελαχιστοποιηθεί (ή να μηδενιστεί) η διαφορά του επόμενου από το προηγούμενο κέντρο της κάθε Συστάδας. Στον **Πίνακα 10** φαίνεται ο αριθμός των επαναλήψεων που έγιναν. Έτσι, στο δείγμα μας, στην τέταρτη επανάληψη, βλέπουμε ότι η διαφορά του κέντρου της κάθε Συστάδας από την προηγούμενη επανάληψη ($3^{η}$ επανάληψη) είναι μηδέν (δηλ. στην τέταρτη επανάληψη δεν άλλαξαν τα κέντρα των Συστάδων σε σχέση με την τιμή που είχαν λάβει στην τρίτη επανάληψη). Συνεπώς, στην τέταρτη επανάληψη δεν μετακινήθηκαν cases μεταξύ των Συστάδων.

Τα τελικά κέντρα των Συστάδων για κάθε μεταβλητή φαίνονται στον **Πίνακα 11**.

Πίνακας 11

Final Cluster Centers

	Cluster	
	1	2
ZSALES	-,82193	,89042
ZADV	-,82021	,88856
ZRENT	-,84327	,91355
ZPERS	-,82289	,89147
ZSCORE1	,03449	-,03737
ZSCORE2	,05350	-,05796

Από τον **Πίνακα 11** συμπεραίνουμε ότι στην $1^{η}$ Συστάδα κατατάσσονται οι επιχειρήσεις που έχουν κατά μέσο όρο μικρότερες μέσες πωλήσεις (ZSALES) κατά 0,8219 τυποποιημένες

μονάδες από τον μέσο όρο των πωλήσεων όλων των επιχειρήσεων του δείγματος. Αντίστοιχα, οι επιχειρήσεις της **1ης Συστάδας** έχουν κατά περίπου 0,8 τυποποιημένες μονάδες λιγότερες δαπάνες για *διαφήμιση*, *ενοίκιο* και *έξοδα προσωπικού* από τις μέσες δαπάνες των επιχειρήσεων του δείγματος. Ενώ, οι επιχειρήσεις της ίδιας Συστάδας (1η Συστάδα) έχουν κατά μέσο όρο μεγαλύτερο score αξιολόγησης από το περιοδικό και τους πελάτες τους, σε σχέση με την μέση τιμή της αξιολόγησης που λαμβάνουν όλες οι επιχειρήσεις του δείγματος.

Τα Χρηματοοικονομικά δεδομένα (ZSALES, ZADV, ZRENT και ZPERS) των επιχειρήσεων της **2ης Συστάδας** είναι υψηλότερα κατά περίπου 0,9 τυποποιημένες μονάδες από τις αντίστοιχες μέσες τιμές όλων των επιχειρήσεων του δείγματος, ενώ η αξιολόγησή τους υστερεί σε σχέση με την μέση τιμή της αξιολόγησης όλων των επιχειρήσεων του δείγματος.

Στον **Πίνακα 12** φαίνεται ο αριθμός των επιχειρήσεων που κατατάσσεται σε κάθε Συστάδα. Έτσι, οι Συστάδες μας έχουν σχεδόν ίδιο πλήθος καθώς στην πρώτη κατατάσσονται 78 επιχειρήσεις και στην δεύτερη 72.

Πίνακας 12

Number of Cases in each Cluster

Cluster	1	78,000
	2	72,000
Valid		150,000
Missing		,000

Πίνακας 13

Distances between Final Cluster Centers

Cluster	1	2
1		3,449
2	3,449	

Στον **Πίνακα 13** εμφανίζεται η απόσταση μεταξύ των κέντρων των δύο Συστάδων.

Ακολουθώντας τη διαδρομή Analyze>Descriptive Statistics>Crosstabs και επιλέγοντας στα Row(s) *cluster number* και στα Column(s) *μέγεθος επιχείρησης*, λαμβάνουμε τον **Πίνακα 14**.

Πίνακας 14

Cluster Number of Case * Μέγεθος Επιχείρησης Crosstabulation

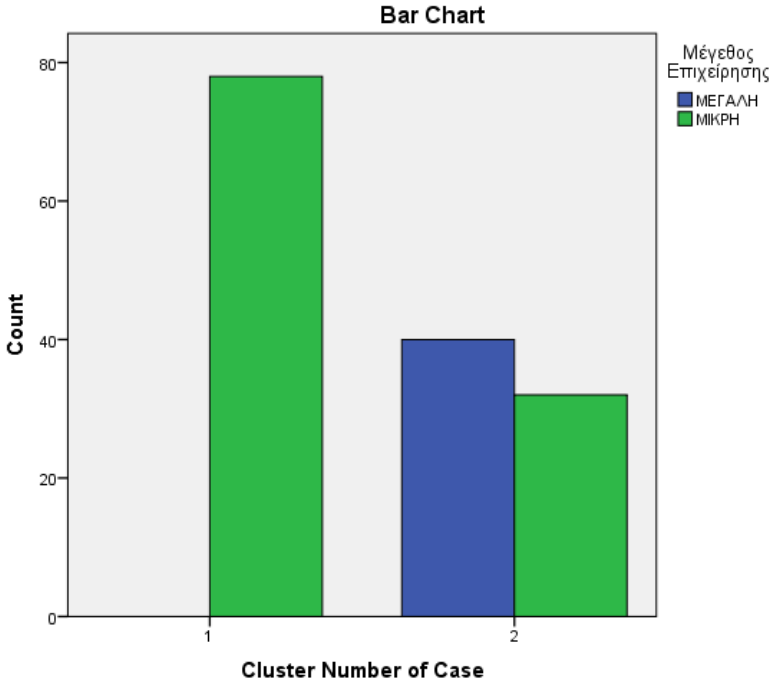
			Μέγεθος Επιχείρησης		Total
			ΜΕΓΑΛΗ	ΜΙΚΡΗ	
Cluster Number of Case	1	Count	0	78	78
		% within Cluster Number of Case	0,0%	100,0%	100,0%
		% within Μέγεθος Επιχείρησης	0,0%	70,9%	52,0%
	2	Count	40	32	72
		% within Cluster Number of Case	55,6%	44,4%	100,0%
		% within Μέγεθος Επιχείρησης	100,0%	29,1%	48,0%
Total		Count	40	110	150
		% within Cluster Number of Case	26,7%	73,3%	100,0%
		% within Μέγεθος Επιχείρησης	100,0%	100,0%	100,0%

Από τον ανωτέρω Πίνακα βλέπουμε ότι όλες οι **μεγάλες επιχειρήσεις** του δείγματός μας κατατάσσονται στην **2η Συστάδα** , η οποία συντίθεται κατά 55,6% από τις επιχειρήσεις αυτές και κατά 44,4% από μικρές επιχειρήσεις. Η 1η Συστάδα, όπου κατατάσσονται μόνο μικρές

επιχειρήσεις, περιλαμβάνει το 70,9% των μικρών επιχειρήσεων του δείγματος έναντι του υπολοίπου 29,1% που κατατάσσεται στη δεύτερη Συστάδα.

Στο **Γράφημα 5** απεικονίζονται οι Συστάδες, βάσει του μεγέθους των επιχειρήσεων που κατατάσσονται σε αυτές.

Γράφημα 5



Δοκιμή με Τρεις Συστάδες

Παρότι η κατάταξη του δείγματος σε δύο συστάδες εμφανίζεται να δίνει ευκρινή και ερμηνεύσιμα αποτελέσματα, θα δοκιμάσουμε να επαναλάβουμε την ανάλυση με τρεις Συστάδες.

Στον **Πίνακα 15** λαμβάνουμε τα αρχικά κέντρα των τριών Συστάδων και τον **Πίνακα 16** τον αριθμό των επαναλήψεων:

Πίνακας 15

Initial Cluster Centers

	Cluster		
	1	2	3
ZSALES	1,81083	-1,69007	-,48316
ZADV	1,86165	-1,67608	-,59016
ZRENT	1,77046	-1,38652	-,11375
ZPERS	2,20287	-1,83040	-,27104
ZSCORE1	,73525	1,60746	-2,36597
ZSCORE2	,82093	3,38633	-2,59960

Πίνακας 16

Iteration History[a]

Iteration	Change in Cluster Centers		
	1	2	3
1	2,028	2,991	2,490
2	,144	,196	,188
3	,078	,060	,117
4	,000	,040	,043
5	,000	,035	,038
6	,000	,031	,037
7	,000	,000	,000

a. Convergence achieved due to no or small change in cluster centers. The maximum absolute coordinate change for any center is ,000. The current iteration is 7. The minimum distance between initial centers is 6,501.

Στους **Πίνακες 17 και 18** βλέπουμε τα τελικά χαρακτηριστικά των Συστάδων μας.

Πίνακας 17

Final Cluster Centers

	Cluster		
	1	2	3
ZSALES	,99753	-,79326	-,64280
ZADV	,99237	-,77710	-,65299
ZRENT	1,01111	-,87198	-,57534
ZPERS	,97242	-,81330	-,58171
ZSCORE1	,07993	,72471	-,93591
ZSCORE2	,04724	,70939	-,86848

Πίνακας 18

Distances between Final Cluster Centers

Cluster	1	2	3
1		3,732	3,493
2	3,732		2,330
3	3,493	2,330	

Ειδικότερα, στην **1η Συστάδα** κατατάσσονται επιχειρήσεις με χρηματοοικονομικά δεδομένα (ZSALES, ZADV, ZRENT και ZPERS) που κατά μέσο όρο είναι υψηλότερα κατά περίπου μια τυποποιημένη μονάδα σε σχέση με τις μέσες τιμές των αντίστοιχων μεταβλητών όλων των επιχειρήσεων του δείγματος.

Οι επιχειρήσεις **της 2ης Συστάδας** έχουν τα μικρότερα χρηματοοικονομικά δεδομένα σε σχέση με τις επιχειρήσεις των άλλων Συστάδων, αλλά εμφανίζουν την καλύτερη (μεγαλύτερη) αξιολόγηση σε σχέση με τις επιχειρήσεις των άλλων Συστάδων.

Τέλος, στην **3η Συστάδα** κατατάσσονται επιχειρήσεις με μικρότερα χρηματοοικονομικά δεδομένα από τις μέσες τιμές των επιχειρήσεων του δείγματος, αλλά υψηλότερα σε σχέση με τα αντίστοιχα των επιχειρήσεων της 2ης Συστάδας. Επιπλέον, οι επιχειρήσεις της 3ης Συστάδας παρουσιάζουν τους χαμηλότερους δείκτες αξιολόγησης, σε σχέση με την αξιολόγηση που λαμβάνουν οι επιχειρήσεις των άλλων δύο Συστάδων, και ταυτόχρονα κατά μέσο όρο η αξιολόγηση που επιτυγχάνουν είναι περίπου κατά μία τυποποιημένη μονάδα μικρότερη της μέσης αξιολόγησης όλων των επιχειρήσεων του δείγματος.

Από τον **Πίνακα 18** βλέπουμε ότι η μεγαλύτερη απόσταση μεταξύ των κέντρων των Συστάδων είναι η απόσταση μεταξύ της 1ης και της 2ης Συστάδας (3,732) ενώ η μικρότερη είναι μεταξύ της 2ης και 3ης (2,330).

Από τον **Πίνακα 19** βλέπουμε ότι στην πρώτη Συστάδα κατατάσσονται 63 επιχειρήσεις, στη δεύτερη 46 και στην τρίτη 41 επιχειρήσεις του δείγματος.

Πίνακας 19

Number of Cases in each Cluster

Cluster	1	63,000
	2	46,000
	3	41,000
Valid		150,000
Missing		,000

Η επισκόπηση των Συστάδων βάσει του μεγέθους των επιχειρήσεων (**Πίνακας 20**) μάς δείχνει

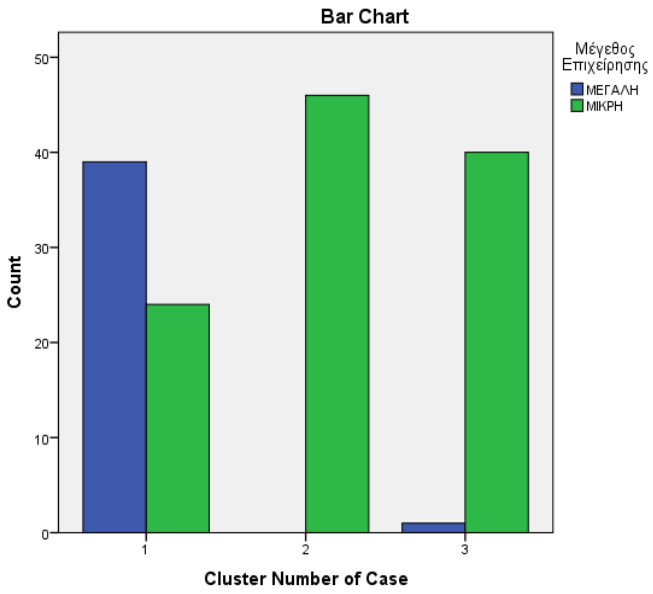
ότι το 97,5 % των μεγάλων επιχειρήσεων (39 από τις 40) κατατάσσονται στην **πρώτη Συστάδα**, δηλ. στη Συστάδα που εμφανίζει τα υψηλότερα Χρηματοοικονομικά Δεδομένα (Συνολικά έσοδα, Δαπάνες διαφήμισης, προσωπικού και ενοικίων), ενώ στη **δεύτερη συστάδα** κατατάσσονται μόνο μικρές επιχειρήσεις οι οποίες εμφανίζουν τα μικρότερα χρηματοοικονομικά δεδομένα αλλά ταυτόχρονα τους υψηλότερους βαθμούς αξιολόγησης. Στη συστάδα αυτή (2η Συστάδα) εντάσσεται το 41,8% των μικρών επιχειρήσεων του δείγματος.

Πίνακας 20

Cluster Number of Case * Μέγεθος Επιχείρησης Crosstabulation

			Μέγεθος Επιχείρησης		Total
			ΜΕΓΑΛΗ	ΜΙΚΡΗ	
Cluster Number of Case	1	Count	39	24	63
		% within Cluster Number of Case	61,9%	38,1%	100,0%
		% within Μέγεθος Επιχείρησης	97,5%	21,8%	42,0%
	2	Count	0	46	46
		% within Cluster Number of Case	0,0%	100,0%	100,0%
		% within Μέγεθος Επιχείρησης	0,0%	41,8%	30,7%
	3	Count	1	40	41
		% within Cluster Number of Case	2,4%	97,6%	100,0%
		% within Μέγεθος Επιχείρησης	2,5%	36,4%	27,3%
Total		Count	40	110	150
		% within Cluster Number of Case	26,7%	73,3%	100,0%
		% within Μέγεθος Επιχείρησης	100,0%	100,0%	100,0%

Γράφημα 6



Στο παραπλεύρως Γράφημα απεικονίζονται ο Συστάδες σε σχέση με το μέγεθος των επιχειρήσεων που κατατάσσονται σ' αυτές.

Συμπεράσματα

Συγκρίνοντας τις προηγηθείσες αναλύσεις κατά Συστάδες, θεωρούμε ότι η κατηγοριοποίηση των επιχειρήσεων του δείγματος σε τρεις συστάδες μας επιτρέπει επιπλέον να διακρίνουμε μεταξύ των μικρότερων επιχειρήσεων εκείνες που παρότι έχουν κατά μέσο όρο τα μικρότερα χρηματοοικονομικά δεδομένα (συνολικά έσοδα, δαπάνες για διαφήμιση…) λαμβάνουν υψηλή αξιολόγηση τόσο από το περιοδικό όσο και από τους πελάτες τους (επιχειρήσεις 2^{ης} Συστάδας). Επίσης, οι επιχειρήσεις της **3^{ης} Συστάδας** παρότι δεν παρουσιάζουν μεγάλες αποκλίσεις στα χρηματοοικονομικά τους δεδομένα σε σχέση με αυτές της 2ης Συστάδας, έχουν δυσμενέστερη αξιολόγηση κατά περίπου δύο τυπικές μονάδες (σε σχέση με αυτές της 2^{ης} Συστάδας) και ταυτόχρονα η αξιολόγησή τους υστερεί κατά περίπου μια τυπική μονάδα σε σχέση με τη μέση τιμή της αξιολόγησης όλων των επιχειρήσεων του δείγματος.

2. Μέθοδος Two-Step Cluster

Από τη διαδρομή Analyze>Classify>**TwoStep Cluster**, επιλέγω στο πεδίο Continuous Variables τις τυποποιημένες τιμές των μτβ: ZSALES, ZADV, ZRENT, ZPERS, ZSCORE1 και ZSCORE2, και στο πεδίο Categorical Variables το μέγεθος της επιχείρησης (SIZE). Στο Number of Clusters επιλέγω: Determine Automatically.

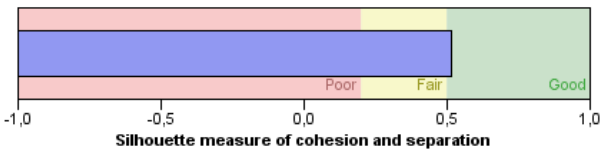
Τρέχοντας τη μέθοδο λαμβάνουμε το **Γράφημα 7** από το οποίο βλέπουμε ότι η κατάταξη των επιχειρήσεων του δείγματος γίνεται σε τρεις Συστάδες και ότι το μοντέλο μας αξιολογείται ως σχετικά καλό (Fair).

Γράφημα 7

Model Summary

Algorithm	TwoStep
Inputs	7
Clusters	3

Cluster Quality



Από τον **Πίνακα 21** φαίνεται ότι το 44,7% των επιχειρήσεων του δείγματος (ή 67 από τις 150) κατατάσσονται στην 1^{η} Συστάδα, το 28,7% στη 2^{η} Συστάδα και το υπόλοιπο 26,7% στην 3^{η} Συστάδα. Η γραφική απεικόνιση παρουσιάζεται στο **Γράφημα 8**.

Από τον **Πίνακα 22** βλέπουμε την κατάταξη των επιχειρήσεων στις τρεις Συστάδες βάσει του μεγέθους τους. Ειδικότερα, στην τρίτη Συστάδα κατατάσσονται μόνο οι μεγάλες επιχειρήσεις, ενώ στην πρώτη και τη δεύτερη μόνο μικρές.

Πίνακας 21

Cluster Distribution

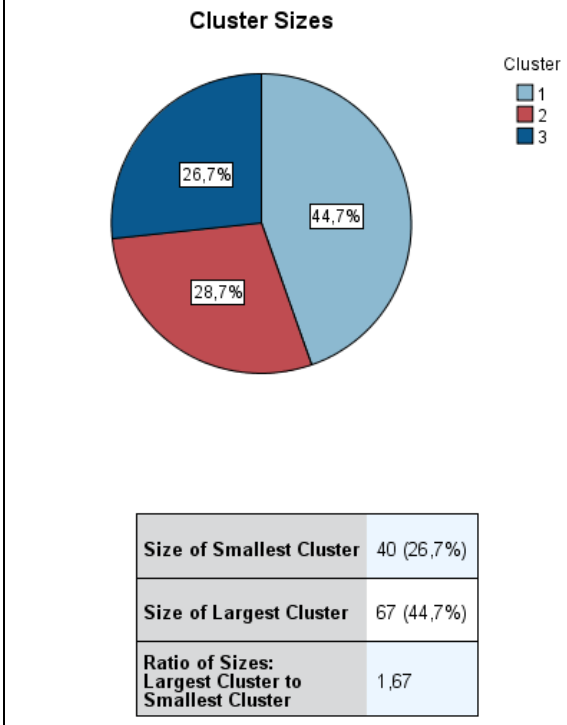
		N	% of Combined	% of Total
Cluster	1	67	44,7%	44,7%
	2	43	28,7%	28,7%
	3	40	26,7%	26,7%
	Combined	150	100,0%	100,0%
Total		150		100,0%

Πίνακας 22

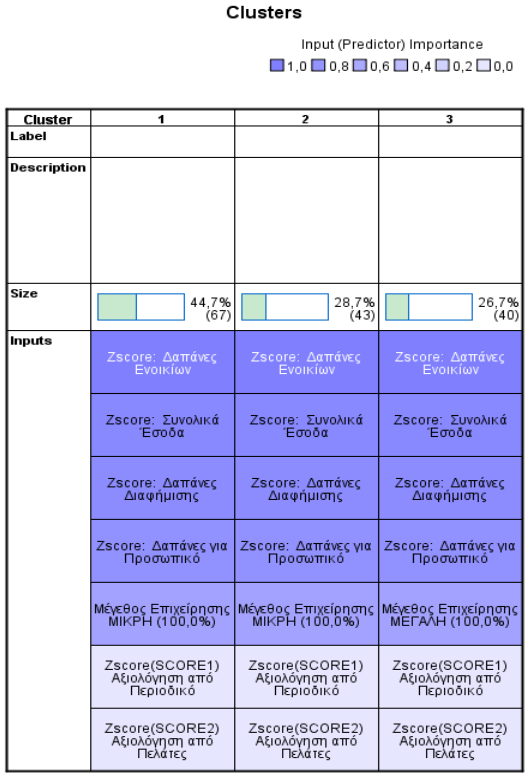
Μέγεθος Επιχείρησης

Cluster	ΜΕΓΑΛΗ		ΜΙΚΡΗ	
	Frequency	%	Frequency	%
1	0	0,0%	67	60,9%
2	0	0,0%	43	39,1%
3	40	100,0%	0	0,0%
Combined	40	100,0%	110	100,0%

Γράφημα 8



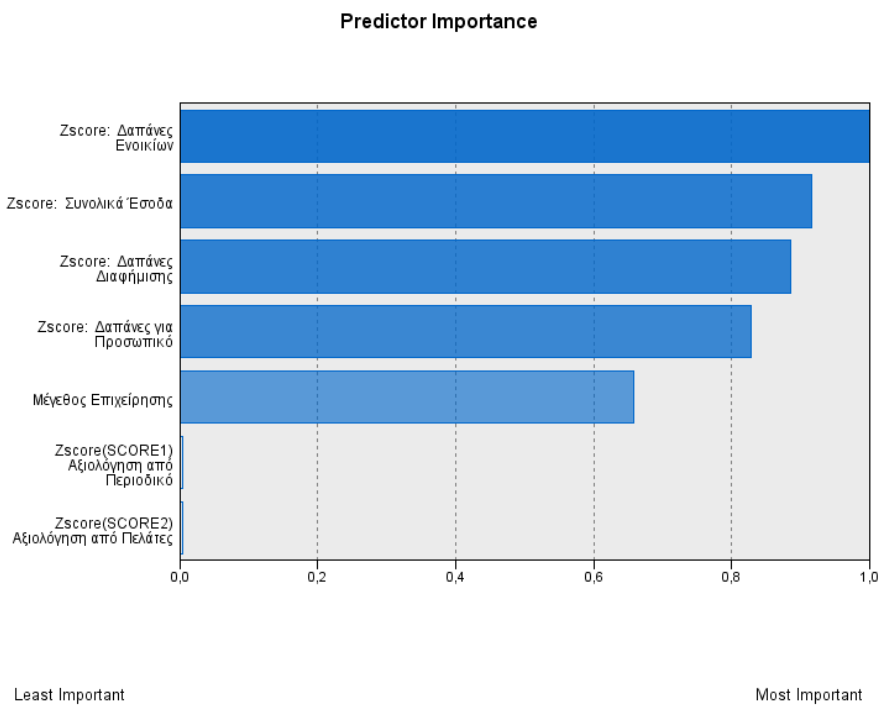
Γράφημα 9



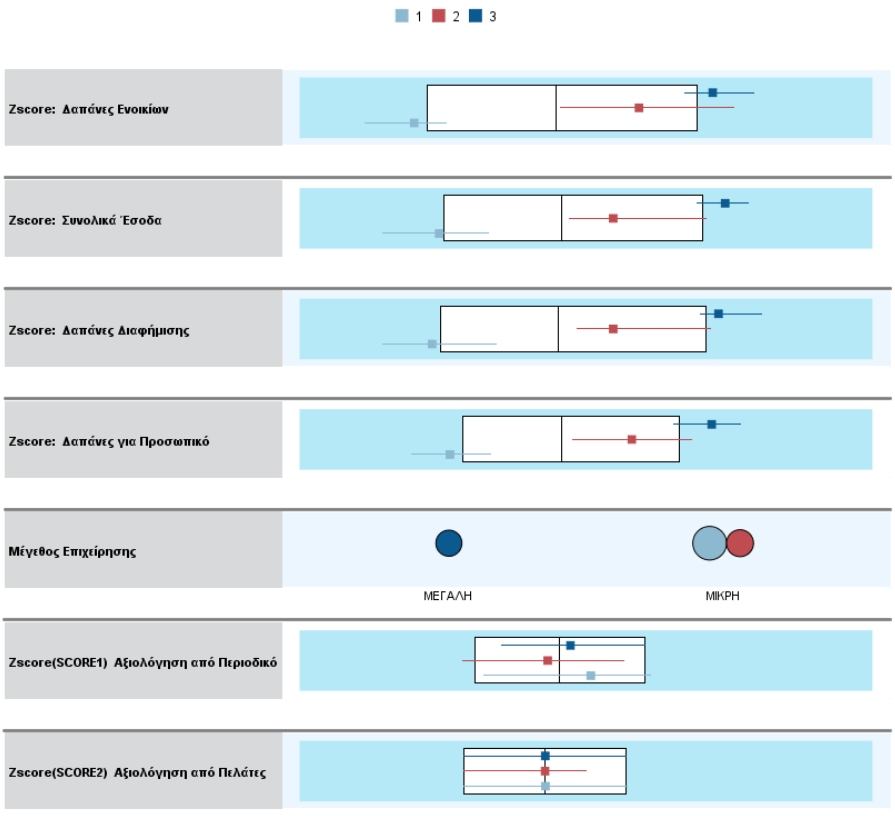
Στο **Γράφημα 9** φαίνονται κατά σειρά σπουδαιότητας τα κριτήρια που καθόρισαν την κατανομή των επιχειρήσεων στις τρεις συστάδες. Έτσι οι *Δαπάνες για Ενοίκιο* είναι το κυριότερο (σημαντικότερο) κριτήριο για την κατάταξη των επιχειρήσεων στις τρεις Συστάδες και ακολουθούν τα *Συνολικά Έσοδα* και οι *Δαπάνες για Διαφήμιση* κ.ο.κ. Το λιγότερο σημαντικό κριτήριο είναι η *Αξιολόγηση από τους πελάτες* και ακολουθεί η *Αξιολόγηση από το περιοδικό (*δηλ. οι βαθμοί αξιολόγησης δεν έχουν σχεδόν καμία επίδραση στη δημιουργία των Συστάδων*)*.

Διαγραμματικά η σπουδαιότητα των κριτηρίων απεικονίζεται στο **Γράφημα 10**.

Γράφημα 10



Γράφημα 11

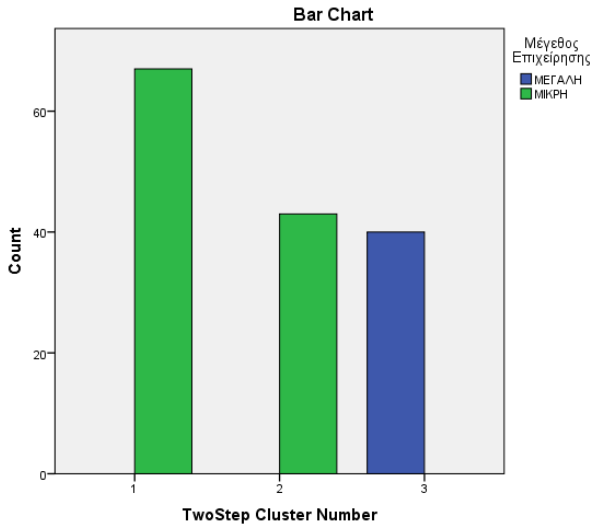


Στο **Γράφημα 11** εμφανίζεται για κάθε μεταβλητή η περιοχή τιμών μεταξύ του πρώτου και του τρίτου τεταρτημορίου, που σχηματίζεται βάσει των τιμών της συγκεκριμένης μεταβλητής στο σύνολο του δείγματος. Για κάθε Συστάδα, σε κάθε μτβ, εμφανίζεται το κέντρο της και το εύρος τιμών που λαμβάνουν τα cases που κατατάσσονται στη συγκεκριμένη συστάδα.

Έτσι, για την μτβ ZRENT (ΖΔαπάνες Ενοικίων) βλέπουμε ότι η 3^{η} Συστάδα έχει κέντρο θετικό και μεγαλύτερο του τρίτου τεταρτημορίου, ενώ η 1^{η} Συστάδα έχει κέντρο αρνητικό και μικρότερο του πρώτου τεταρτημορίου. Αυτή η παρατήρηση ισχύει για τις τέσσερις πρώτες μεταβλητές και έτσι μπορούμε να συμπεράνουμε ότι οι επιχειρήσεις της 3^{ης} Συστάδας έχουν υψηλά και θετικά (μεγαλύτερα της μέσης τιμής) Χρηματοοικονομικά δεδομένα (δαπάνες ενοικίων, Συνολικά έσοδα, δαπάνες διαφήμισης, δαπάνες προσωπικού), ενώ αντίθετα οι επιχειρήσεις της 1^{ης} Συστάδας έχουν αρκετά χαμηλότερα της μέσης τιμής Χρηματοοικονομικά δεδομένα. Οι επιχειρήσεις της 2^{ης} Συστάδας, έχουν θετικά αλλά όχι υψηλά Χρηματοοικονομικά δεδομένα.

Στο ίδιο γράφημα (Γράφημα 11), για την ποιοτική μτβ *Μέγεθος Επιχείρησης,* βλέπουμε ότι στην 3^{η} Συστάδα εντάσσονται μόνο μεγάλες επιχειρήσεις ενώ στην 1^{η} και 2^{η} εντάσσονται μόνο μικρές. Το πλήθος τους καθορίζει το μέγεθος του σημείου στο γράφημα.

Γράφημα 12



Στο παραπλεύρως Γράφημα[1] απεικονίζονται ο Συστάδες σε σχέση με το μέγεθος των επιχειρήσεων που κατατάσσονται σ' αυτές.

Συγκεντρωτικά Συμπεράσματα

Η μέθοδος Two-Step Cluster κατέταξε τις επιχειρήσεις του δείγματος σε τρεις συστάδες, δίνοντας βαρύτητα μόνο στα Χρηματοοικονομικά τους δεδομένα αγνοώντας τις μτβ που αφορούν στην Εξωτερική τους Αξιολόγηση. Αντίθετα με τη μέθοδο K-Means με τρεις Συστάδες, η κατάταξη των επιχειρήσεων λαμβάνει υπόψη της πέραν των Χρηματοοικονομικών δεδομένων και την Εξωτερική Αξιολόγηση. Με τον τρόπο αυτό γίνεται διάκριση μεταξύ εκείνων των επιχειρήσεων που παρότι έχουν παρόμοια χρηματοοικονομικά

[1] Το Γράφημα 12 δημιουργείται από τη διαδρομή Analyze>Descriptive Statistics>Crosstabs, στα Rows επιλέγω: TwoStep Cluster Numb και στα Columns: Μέγεθος Επιχείρησης

δεδομένα διαφέρουν ως προς την Εξωτερική Αξιολόγηση, που λαμβάνουν από το περιοδικό και από τους πελάτες του.

Συνοπτικά, η μέθοδος K-Means με τρεις Συστάδες κρίνεται καλύτερη για τα δεδομένα μας γιατί κατατάσσει τις επιχειρήσεις του δείγματος τόσο βάσει των Χρηματοοικονομικών δεδομένων όσο και βάσει των βαθμών Αξιολόγησης που λαμβάνουν.