# Εξόρυξη δεδομένων με χρήση τεχνικών μηχανικής μάθησης

Τμ. Μηχανικών Πληροφορικής και Υπολογιστών

# Παράδειγμα- Αντιπροσωπία Αυτοκίνητων

# Πρόβλημα

Ο Νίκος είναι πωλητής σε μια αντιπροσωπία αυτοκίνητων μάρκας Audi. Η εταιρία παρατήρησε ότι ενώ ο αριθμός των επισκεπτών στην έκθεση αυξανόταν αντιθέτως οι πωλήσεις μειώνονταν. Έτσι ανάθεσε στον Νίκο να καταγράψει ότι πληροφορίες σχετίζονταν με τους πελάτες από τη στιγμή που επισκέπτονταν την αντιπροσωπεία μέχρι την στιγμή που φεύγουν. Σκοπός τους είναι σε πρώτη φάση να εντοπίσουν τα χαρακτηριστικά των ομάδων-πελατών που αγοράζουν και αυτών που δεν αγοράζουν και στη συνέχεια να δούν σε ποιες ομάδες πελάτων θα μπορούσαν ενδεχομένως να κάνουν κάποιες διορθωτικές κινήσεις προκειμένου να αυξήσουν τις πωλήσεις τους.

# Τα βήματα για την Κατηγοριοποίηση

Τα βήματα που απαιτούνται είναι:

1. Καθαρισμός των δεδομένων
2. Κανονικοποίηση σε περίπτωση που απαιτείται
3. Υπολογισμός της απόστασης κάθε αντικειμένου από το κέντρο του.
4. Δημιουργία των clusters με βάση την απόσταση μεταξύ των αντικειμένων.
5. Υπολογισμός του κέντρου του cluster (*centroid)*.
6. Υπολογισμός της απόστασης κάθε αντικειμένου από το κέντρο του cluster στο οποίο ανήκει. Αν δεν υπάρχει κάποιο νέο αντικείμενο η διαδικασία ολοκληρώνεται. Διαφορετικά θα πρέπει να πάμε ξανά στο βήμα 3.

# Τα βήματα για την Κατηγοριοποίηση

Αν θεωρήσουμε ότι έχουμε ένα σύνολο δεδομένων που αποτελείται από 100 γραμμές και 4 στήλες και τα δεδομένα αυτά θέλουμε να τα χωρίσουμε σε 4 clusters, ο χρόνος που θα απαιτηθεί για να γίνουν όλοι οι παραπάνω υπολογισμοί θα είναι τεράστιος.

Με τη βοήθεια του εργαλείου Rapid Miner, θα δούμε πως τέτοιες διαδικασίες εκτελούνται σε μόλις μερικά λεπτά.

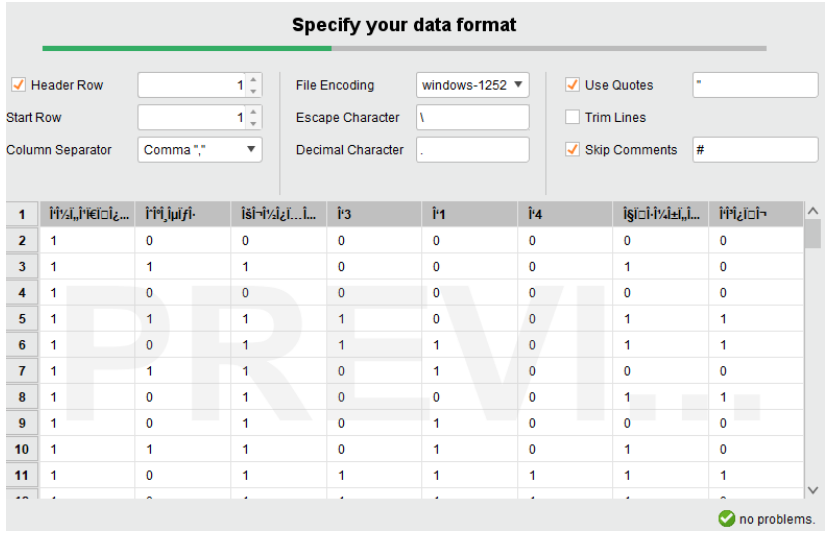
# Τα Δεδομένα

Τα δεδομένα τα οποία αποφασίστηκε να συλλέγονται είναι

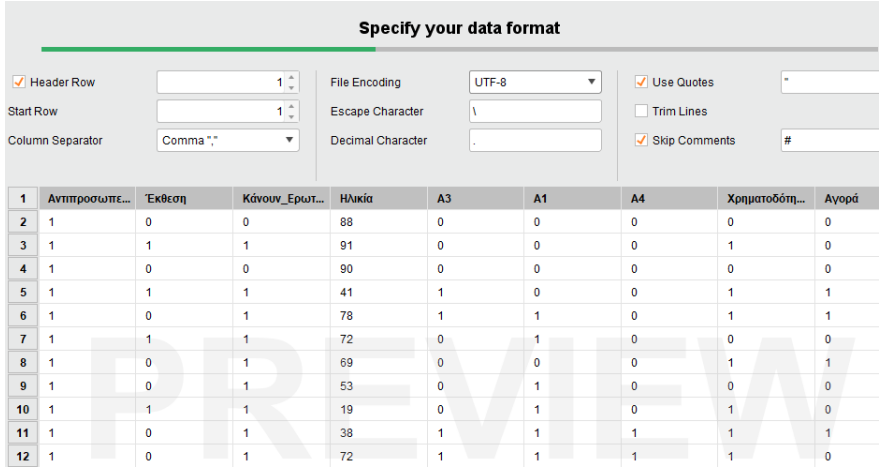
- Αντιπροσωπεία {0,1}, Αν επισκέπτονται την αντιπροσωπεία χωρίς να μπουν μέσα
- Έκθεση {0,1}, Αν μπαίνουν στην Έκθεση
- Κάνουν_Ερωτήσεις {0,1}, Αν κάνουν ερωτήσεις
- Ηλικία {19,88}
- A3 {0,1}, Αν ενδιαφέρονται για το A3
- A1 {0,1}, Αν ενδιαφέρονται για το A1
- A4 {0,1}, Αν ενδιαφέρονται για το A4
- Χρηματοδότηση {0,1}, Αν επιλέγουν κάποιο πακέτο χρηματοδότησης
- Αγορά {0,1}, Αν αγοράζουν.

# Προετοιμασία των Δεδομένων

- Το αρχείο με τα δεδομένα θα το βρείτε στον φάκελο dataset της ενότητας.
- Επειδή τα ονόματα των στηλών στο αρχείο είναι με Ελληνικούς χαρακτήρες, όταν θα εισάγετε τα δεδομένα ενδεχομένως να δείτε αυτή την οθόνη

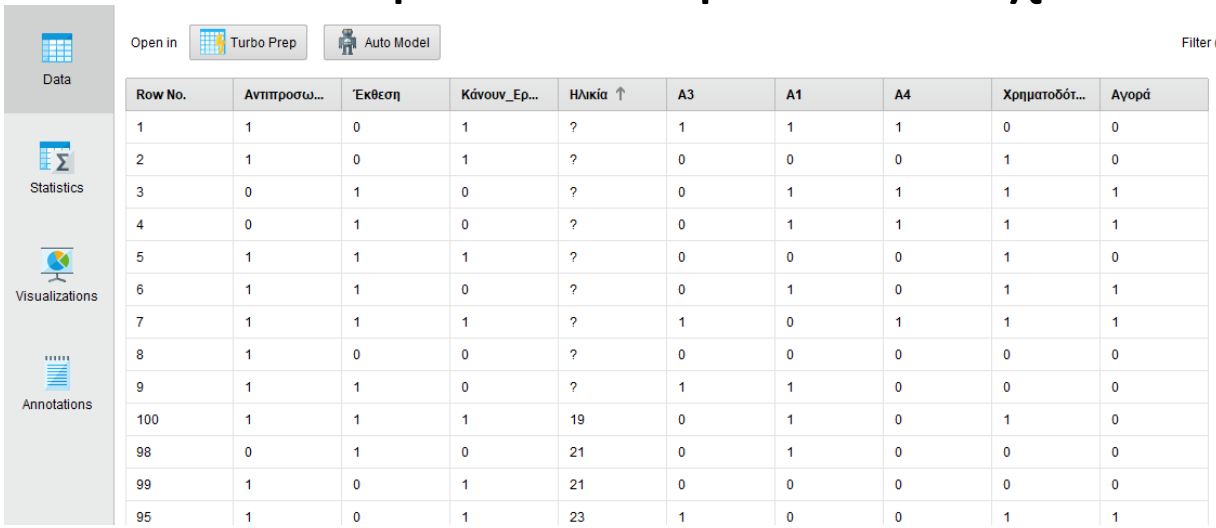



- Για να διορθωθεί το πρόβλημα με την κωδικοποίηση θα πρέπει να επιλέξετε File Encoding: UTF-8

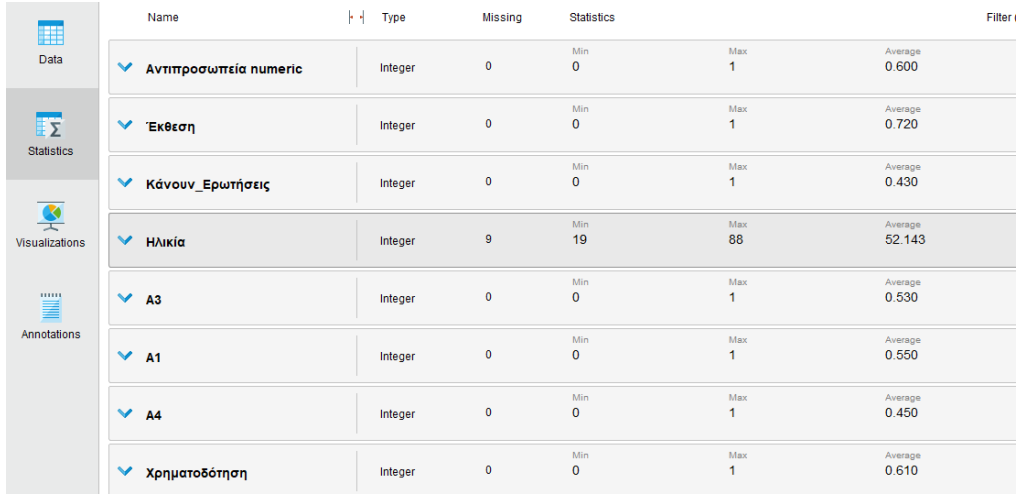


# Προετοιμασία των Δεδομένων

Τα δεδομένα θα πρέπει να έχουν αυτή την μορφή.



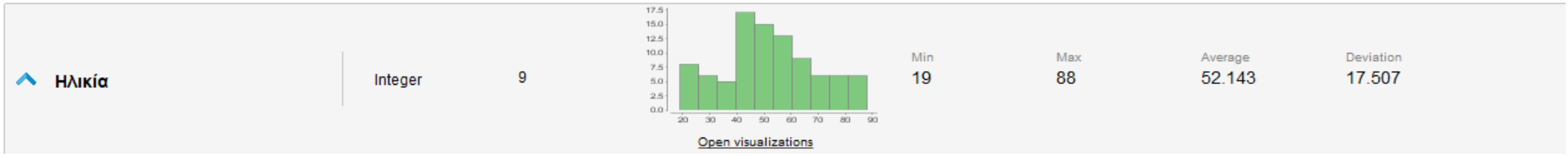

| Row No. | Αντιπροσω... | Έκθεση | Κάνουν_Ερ... | Ηλικία | A3 | A1 | A4 | Χρηματοδότ... | Αγορά |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 1 | ? | 1 | 1 | 1 | 0 | 0 |
| 2 | 1 | 0 | 1 | ? | 0 | 0 | 0 | 1 | 0 |
| 3 | 0 | 1 | 0 | ? | 0 | 1 | 1 | 1 | 1 |
| 4 | 0 | 1 | 0 | ? | 0 | 1 | 1 | 1 | 1 |
| 5 | 1 | 1 | 1 | ? | 0 | 0 | 0 | 1 | 0 |
| 6 | 1 | 1 | 0 | ? | 0 | 1 | 0 | 1 | 1 |
| 7 | 1 | 1 | 1 | ? | 1 | 0 | 1 | 1 | 1 |
| 8 | 1 | 0 | 0 | ? | 0 | 0 | 0 | 0 | 0 |
| 9 | 1 | 1 | 0 | ? | 1 | 1 | 0 | 0 | 0 |
| 100 | 1 | 1 | 1 | 19 | 0 | 1 | 0 | 1 | 0 |
| 98 | 0 | 1 | 0 | 21 | 0 | 1 | 0 | 0 | 0 |
| 99 | 1 | 0 | 1 | 21 | 0 | 0 | 0 | 0 | 0 |
| 95 | 1 | 0 | 1 | 23 | 1 | 0 | 0 | 1 | 1 |



# Προετοιμασία των Δεδομένων

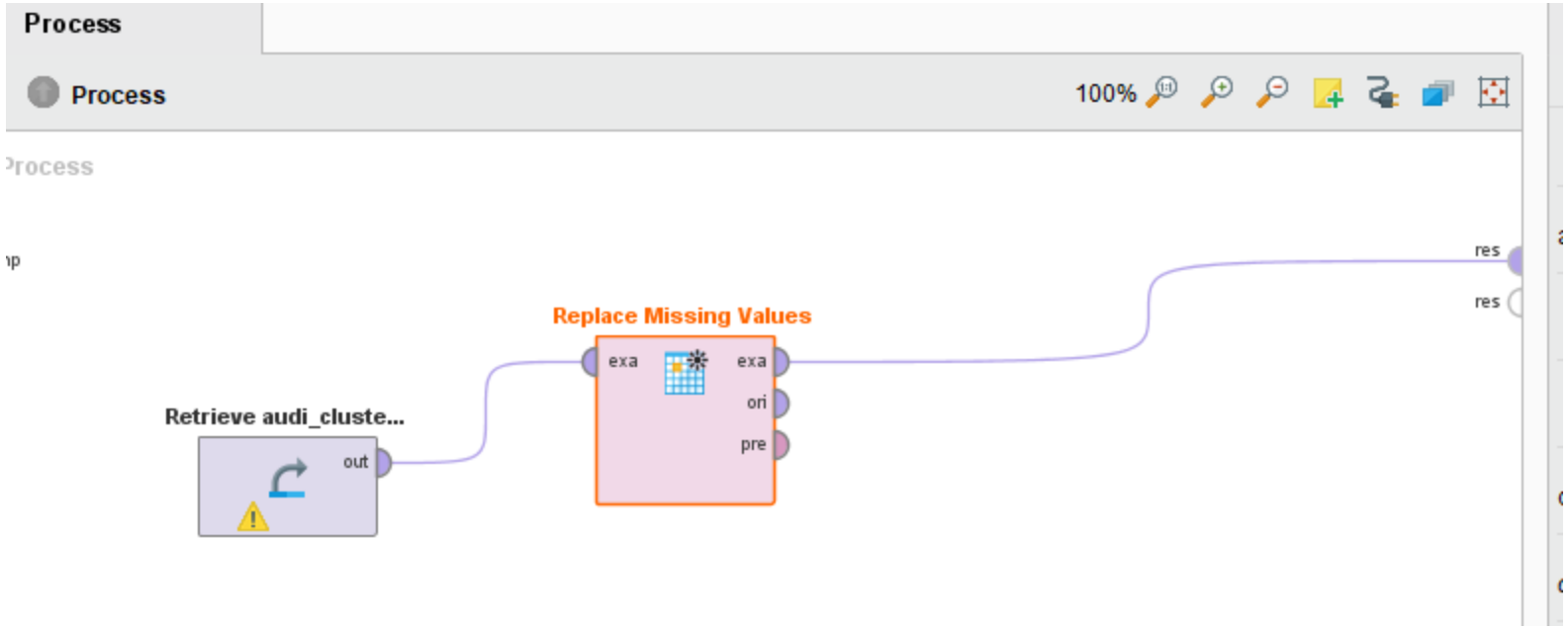
Όπως προκύπτει από τα δεδομένα

- Θα πρέπει να αντικαταστήσουμε τις μηδενικές τιμές (9 τιμές)
- Και θα χρειαστεί να κάνουμε κανονικοποίηση στην ηλικία καθώς έχει τιμές από 19 μέχρι 88, ενώ όλο το δείγμα για τις υπόλοιπες μεταβλητές παίρνει τιμές 0 ή 1.



# Προετοιμασία των Δεδομένων – Missing Values

Προκειμένου να αντικαταστήσουμε τις μεταβλητές επιλέγουμε τον operator “Replace Missing Values”

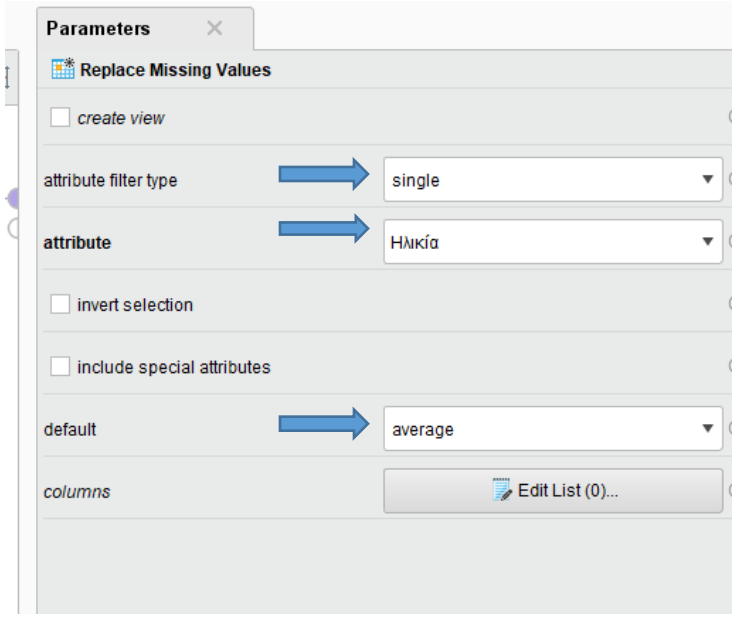


# Προετοιμασία των Δεδομένων – Missing Values

Επιλέγουμε το “Replace Missing Values” και μας ανοίγει από δεξιά το παράθυρο «Parameters» .

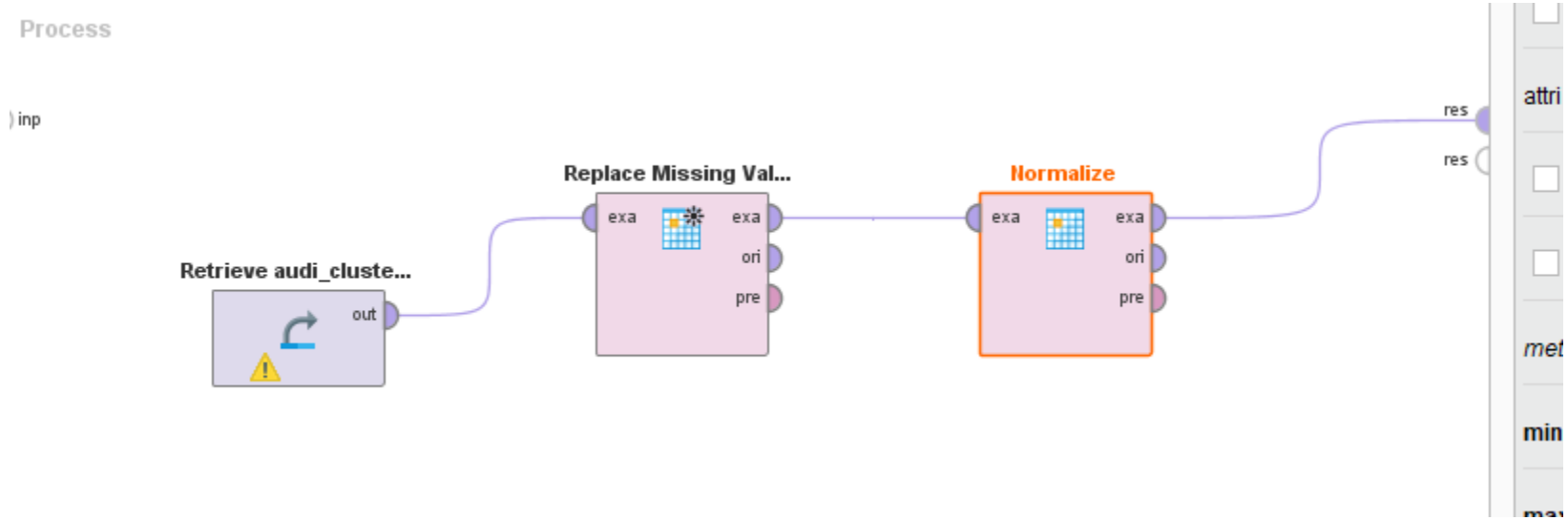
Και επιλέγουμε

- Attribute filter type: Single
- Attribute: Ηλικία
- Default : Average



# Προετοιμασία των Δεδομένων – Normalization

Στην συνέχεια προσθέτουμε στο μοντέλο το «Normalize»

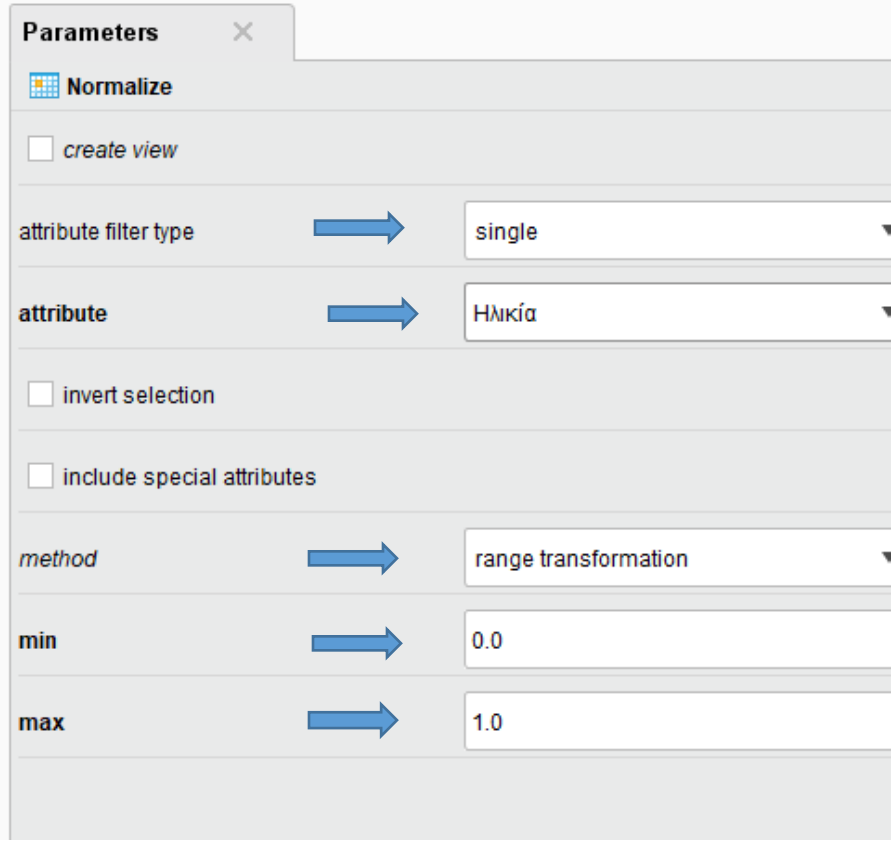


# Προετοιμασία των Δεδομένων – Normalization

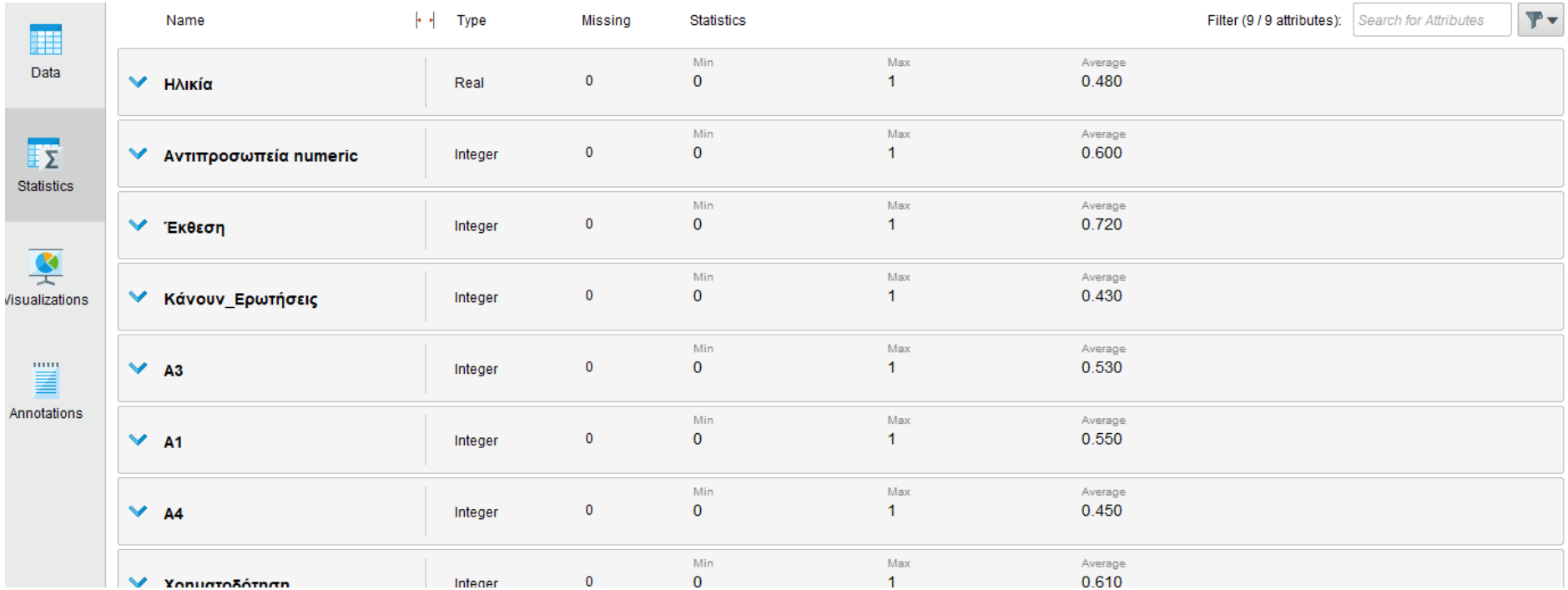
Επιλέγουμε το “Normalize” και μας ανοίγει από δεξιά το παράθυρο «Parameters» .

Και επιλέγουμε

- Attribute filter type: Single
- Attribute: Ηλικία
- Method : Range Transformation
- Min : 0.0
- Max : 1.0

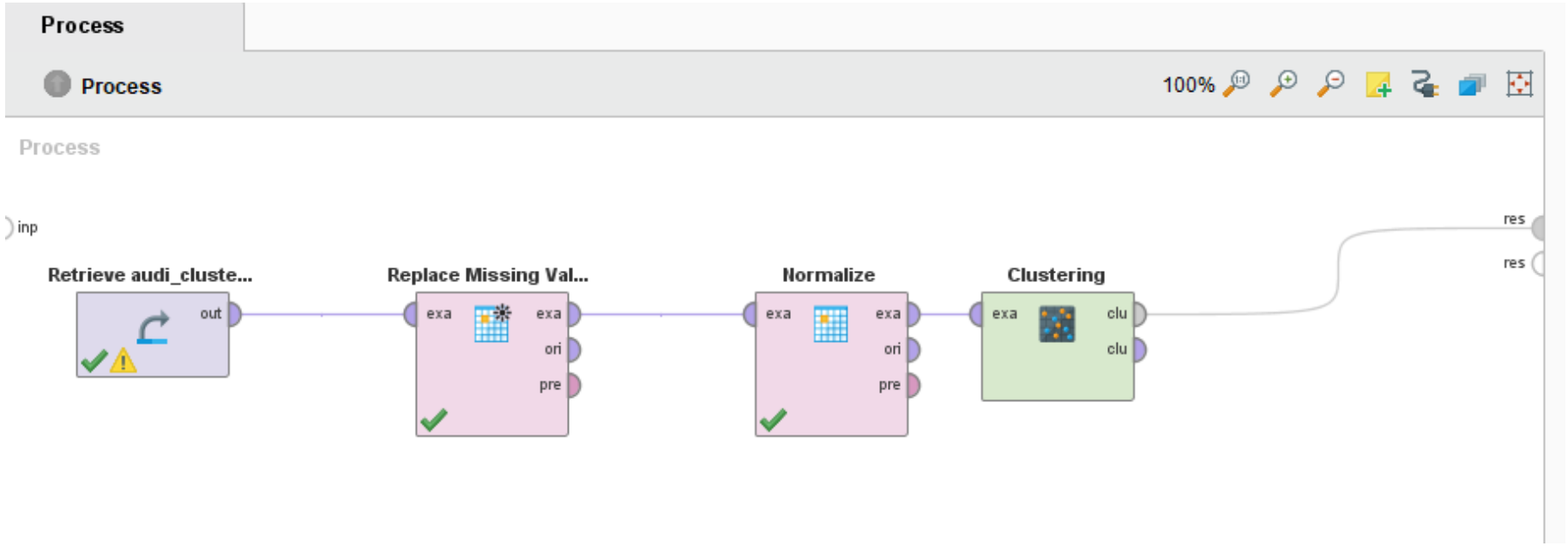


# DATA PREPARATION – Τελική Μορφή



# Μοντελοποίηση

- Προκειμένου να πραγματοποιήσουμε την κατηγοριοποίηση των δεδομένων θα χρησιμοποιήσουμε τον αλγόριθμο K-means.
- Οπότε το μοντέλο μας θα έχει την παρακάτω μορφή



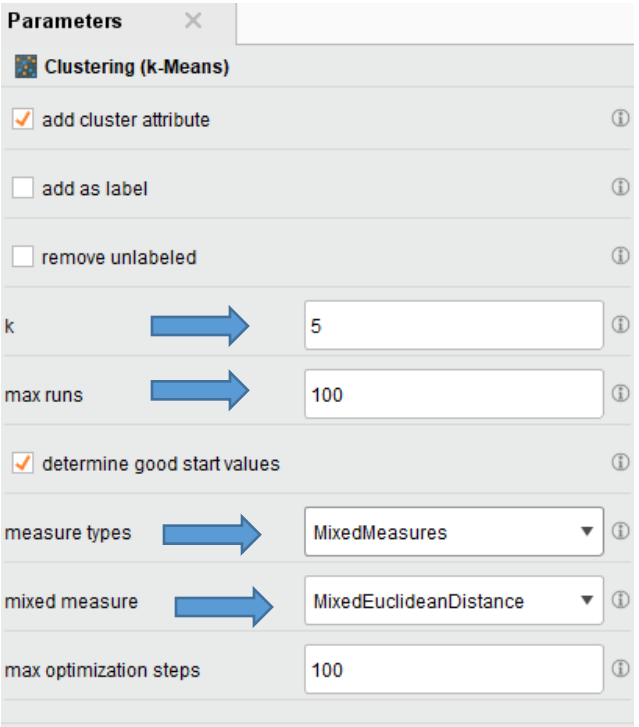
# Μοντελοποίηση

Επιλέγουμε το “Clustering” και μας ανοίγει από δεξιά το παράθυρο «Parameters».

Στο παράθυρο parameters επιλέγουμε

- K: Ο αριθμός των clusters
- Max_runs: 100 . Αυτή η παράμετρος καθορίζει τον μέγιστο αριθμό επαναλήψεων όπου θα επιλέγονται τυχαία αντικείμενα σαν κέντρα των clusters.

Measure types και mixed measure : Επιλέγουμε τον αλγόριθμο που θα χρησιμοποιηθεί για την μέτρηση της απόστασης μεταξύ των αντικειμένων.

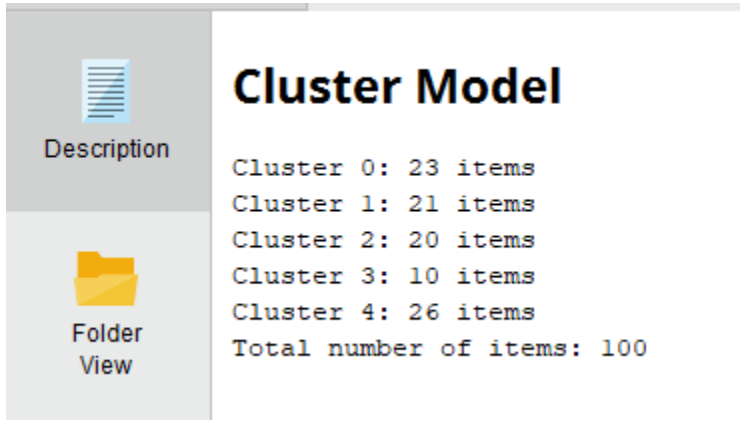


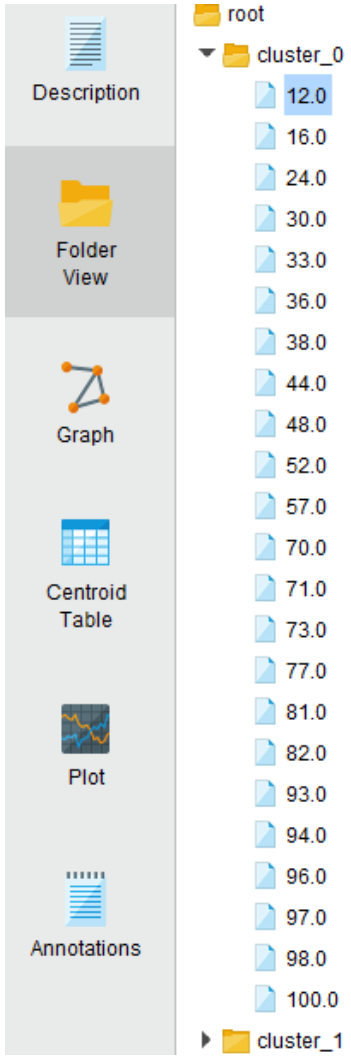

Useful Link

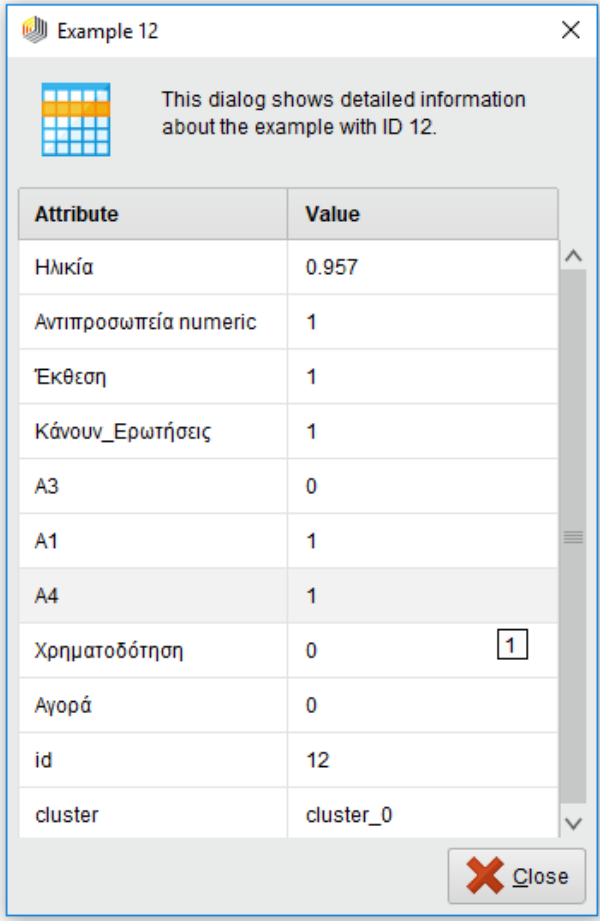
# Αποτελέσματα

Όπως βλέπουμε («Description») έχουν δημιουργηθεί 5 clusters με τα αντίστοιχα αντικείμενα.

Επίσης, αν επιλέξουμε «Folder View» βλέπουμε αναλυτικά ποια αντικείμενα βρίσκονται εντός των clusters.



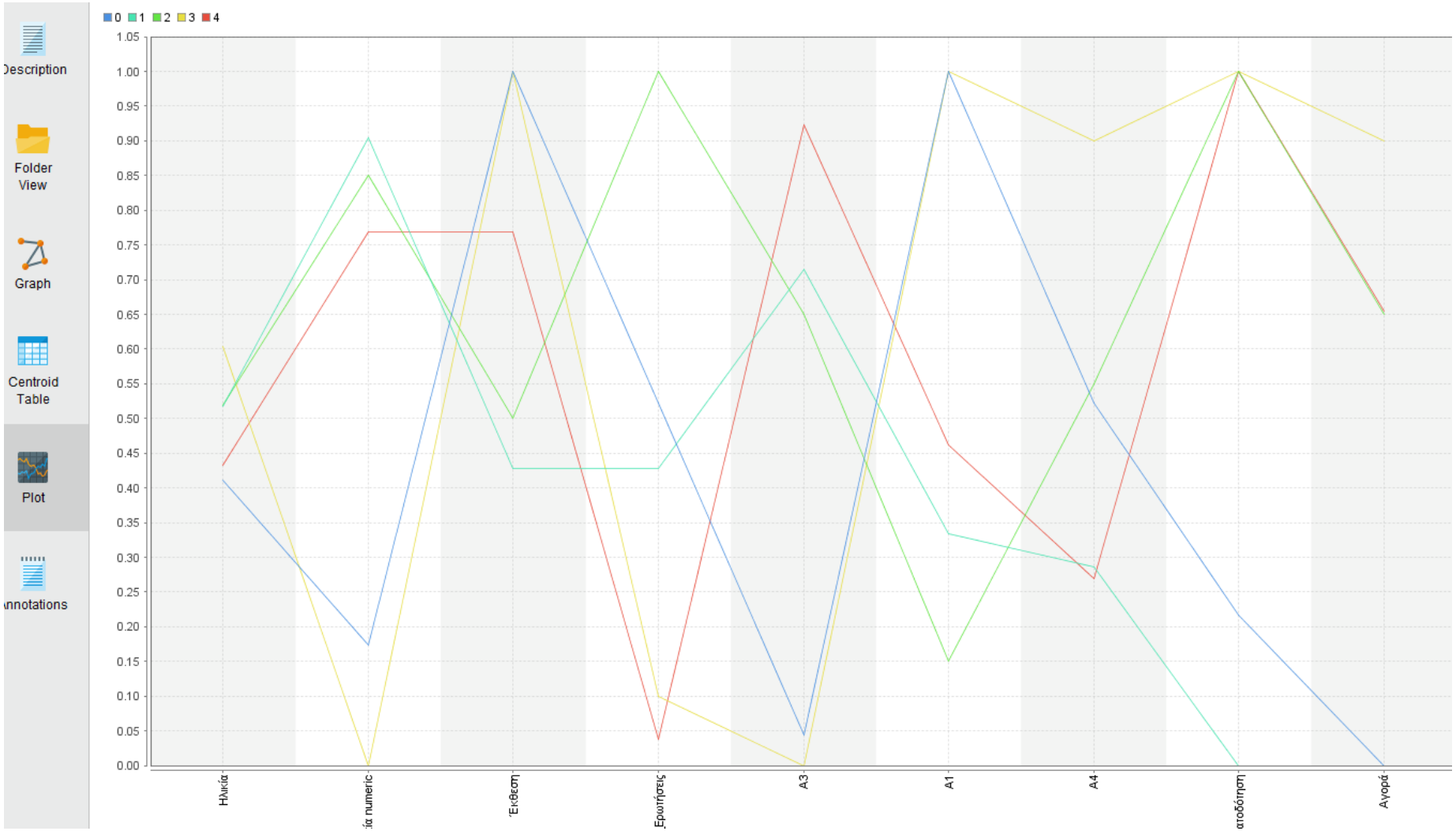



# Αποτελέσματα

| Attribute | cluster_0 | cluster_1 | cluster_2 | cluster_3 | cluster_4 |
|---|---|---|---|---|---|
| Ηλικία | 0.411 | 0.518 | 0.519 | 0.604 | 0.433 |
| Αντιπροσωπεία | 0.174 | 0.905 | 0.850 | 0 | 0.769 |
| Έκθεση | 1 | 0.429 | 0.500 | 1 | 0.769 |
| Κάνουν_Ερωτήσεις | 0.522 | 0.429 | 1 | 0.100 | 0.038 |
| A3 | 0.043 | 0.714 | 0.650 | 0 | 0.923 |
| A1 | 1 | 0.333 | 0.150 | 1 | 0.462 |
| A4 | 0.522 | 0.286 | 0.550 | 0.900 | 0.269 |
| Χρηματοδότηση | 0.217 | 0 | 1 | 1 | 1 |
| Αγορά | 0 | 0 | 0.650 | 0.900 | 0.654 |

# Αποτελέσματα



# Αποτελέσματα

- **Cluster 0**— Αυτή η ομάδα αποτελείται από ανθρώπους οι οποίοι δεν κάθονται πολλή ώρα στην αντιπροσωπεία να κοιτάνε αυτοκίνητα αλλά προτιμούν να μπαίνουν μέσα στην έκθεση και να κάνουν ερωτήσεις. Δυστυχώς, αυτή η κατηγορία είναι η χειρότερη καθώς δεν αγοράζει ποτέ τίποτα και απασχολεί το προσωπικό με ερωτήσεις.
- **Cluster 1**— Στην ομάδα 2 ανήκουν αυτοί που απασχολούν περισσότερο τους υπαλλήλους με ερωτήσεις. Συνήθως αυτοί οι πελάτες παρατηρούν για πολλή ώρα τα αυτοκίνητα χωρίς να μπουν στην έκθεση και δυστυχώς και αυτή η κατηγορία πραγματοποιεί αγορές πολύ σπανία.
- **Cluster 2**— Σε αυτή την ομάδα ανήκουν αυτοί που τους αρέσει το A3, A4. Αυτοί οι πελάτες απασχολούν για πολλή ώρα τους υπάλληλους καθώς κάνουν πολλές ερωτήσεις, ευτυχώς όμως οι περισσότεροι από αυτούς πραγματοποιούν την αγορά κάποιου προϊόντος.

# Αποτελέσματα

- **Cluster 3**— Η τελευταία ομάδα είναι η αγαπημένη μας ομάδα. Δεν απασχολεί καθόλου το προσωπικό με ερωτήσεις μπαίνει κατευθείαν στην έκθεση και πραγματοποιεί σχεδόν πάντα κάποια αγορά.
- **Cluster 4**—Η ομάδα 4 μοιάζει αρκετά με την ομάδα 2 με την διαφορά ότι αυτοί οι πελάτες ενδιαφέρονται για το A3 και επίσης δεν απασχολούν το προσωπικό με ερωτήσεις.

# Αποτελέσματα

Όπως παρατηρούμε, δημιουργήθηκαν 5 clusters. Αν θεωρούμε ότι θέλουμε πιο λίγες ομάδες τότε μπορούμε να μειώσουμε την τιμή του K στις παραμέτρους. Οπότε έστω ότι K=3. Τότε οι ομάδες θα είχαν τα παρακάτω χαρακτηριστικά

| Attribute | cluster_0 | cluster_1 | cluster_2 |
|---|---|---|---|
| Ηλικία | 0.475 | 0.516 | 0.438 |
| Αντιπροσωπεία | 0.065 | 0.897 | 0.767 |
| Έκθεση | 1 | 0.590 | 0.600 |
| Κάνουν_Ερωτήσεις | 0.355 | 0.462 | 0.467 |
| A3 | 0.032 | 0.692 | 0.833 |
| A1 | 1 | 0.359 | 0.333 |
| A4 | 0.677 | 0.385 | 0.300 |
| Χρηματοδότηση | 0.452 | 0.436 | 1 |
| Αγορά | 0.290 | 0 | 1 |

ΤΕΛΟΣ