

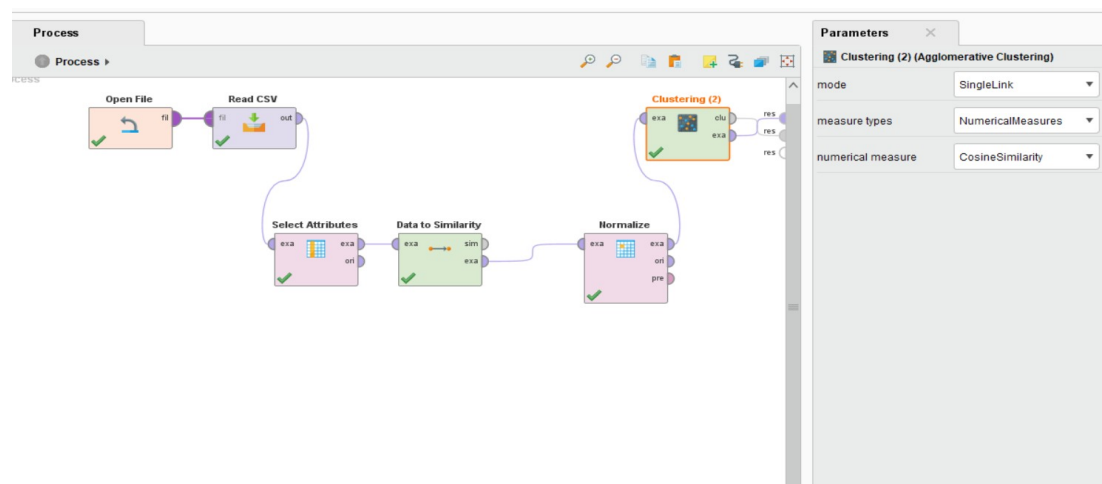


ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
ΣΧΟΛΗ ΤΕΧΝΟΛΟΓΙΚΩΝ ΕΦΑΡΜΟΓΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ
ΥΠΟΛΟΓΙΣΤΩΝ

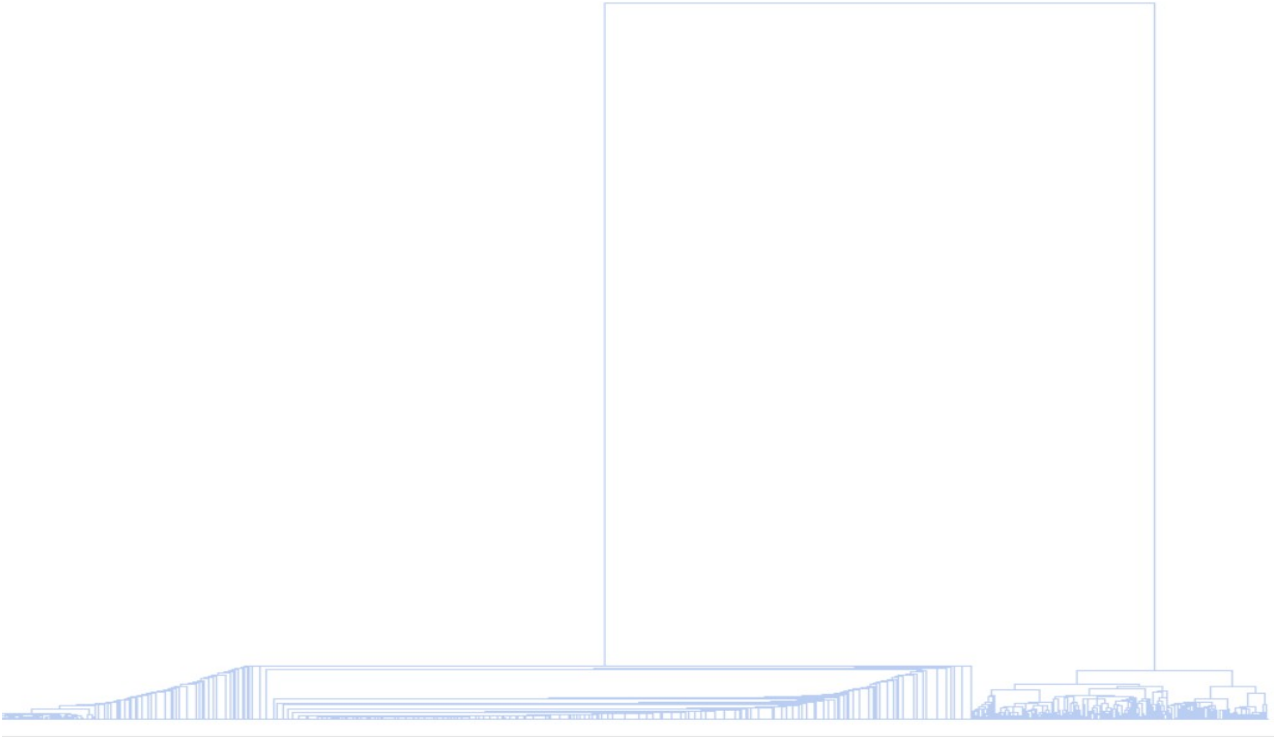
ΒΑΣΙΛΗΣ ΧΡΙΣΤΟΔΟΥΛΟΥ
ΑΜ:161028
ΕΞΟΡΥΞΗ ΔΕΔΟΜΕΝΩΝ
ΕΡΓΑΣΙΑ 1
ΣΥΣΤΑΔΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ

ΜΕΡΟΣ 1ο

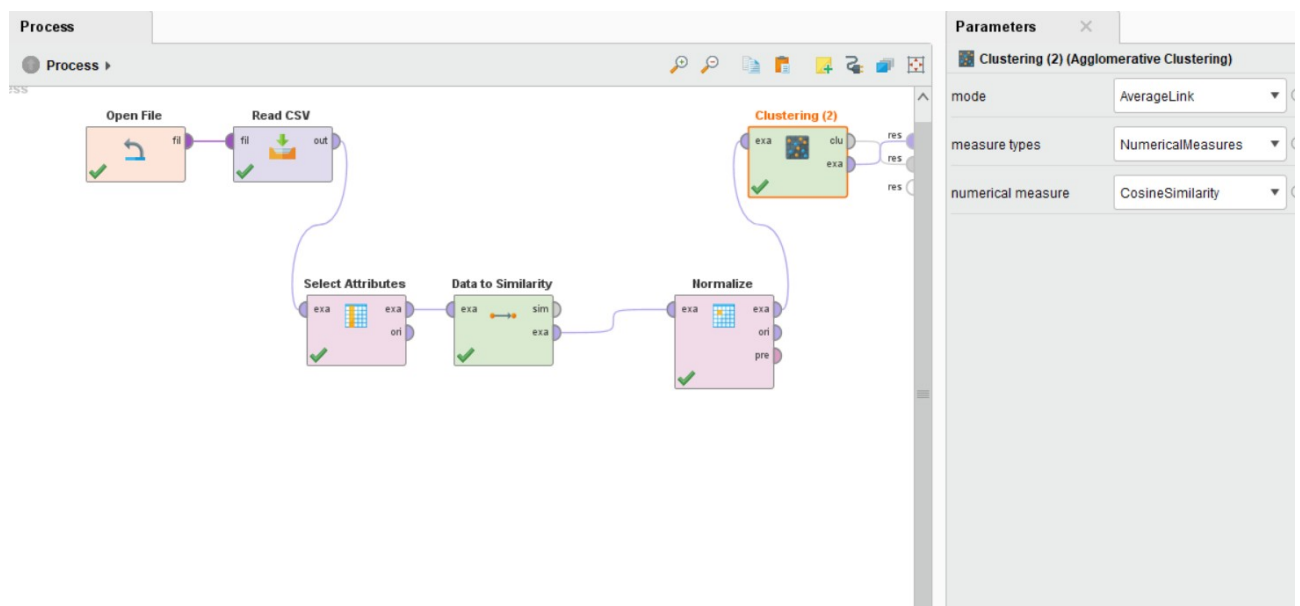
1.1



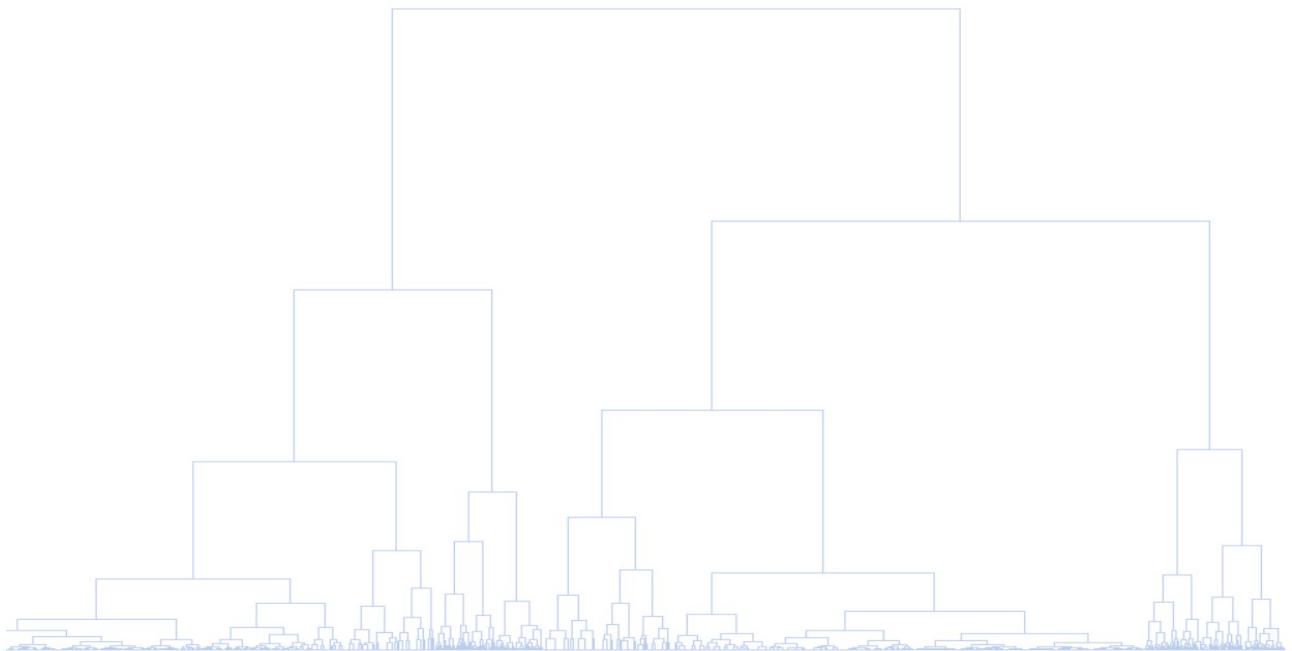
Δενδρόγραμμα:



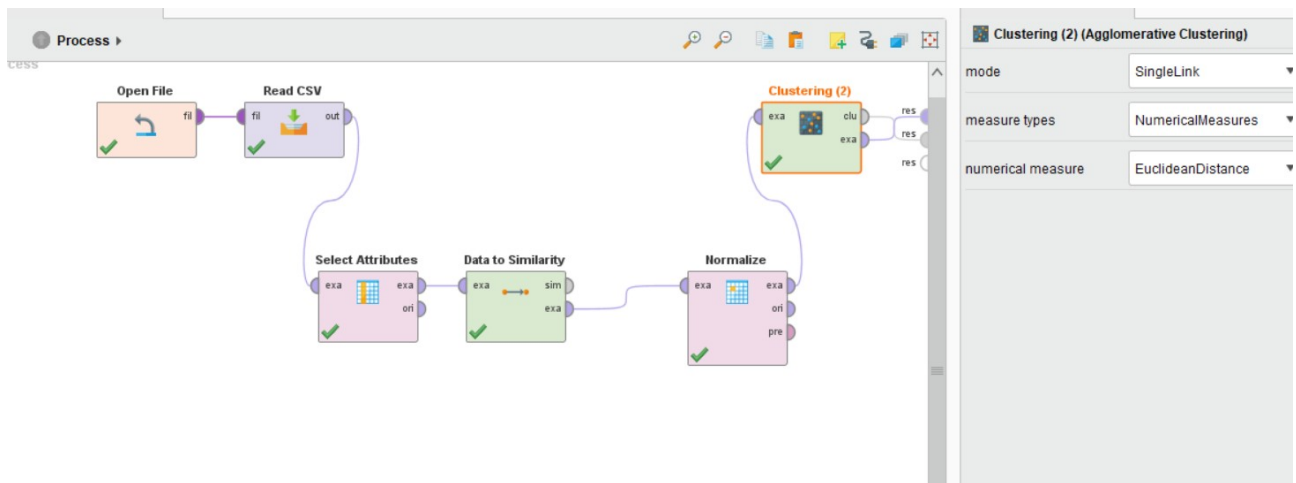
1.2



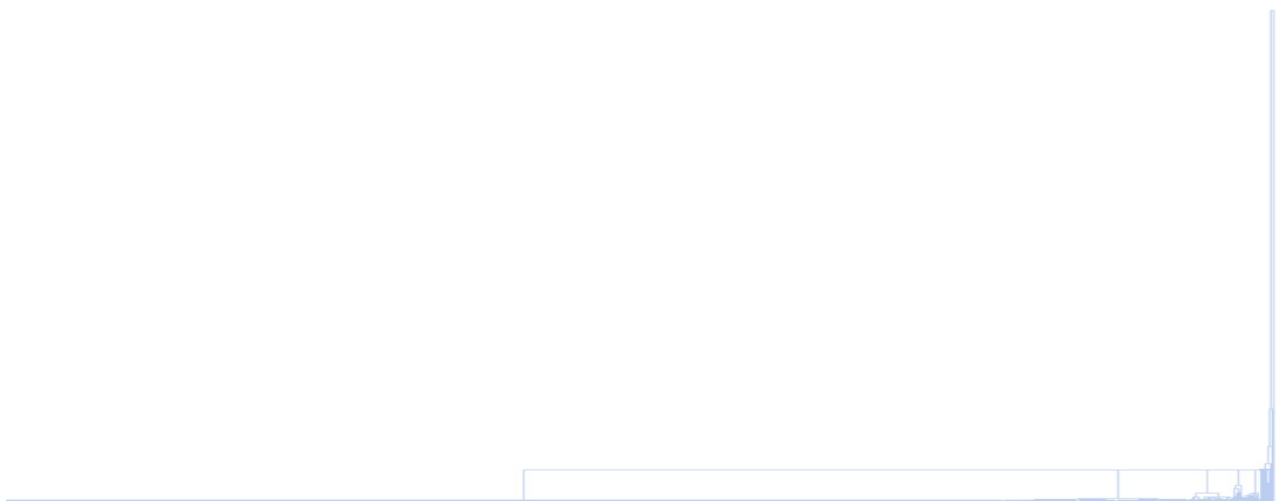
Δενδρόγραμμα:



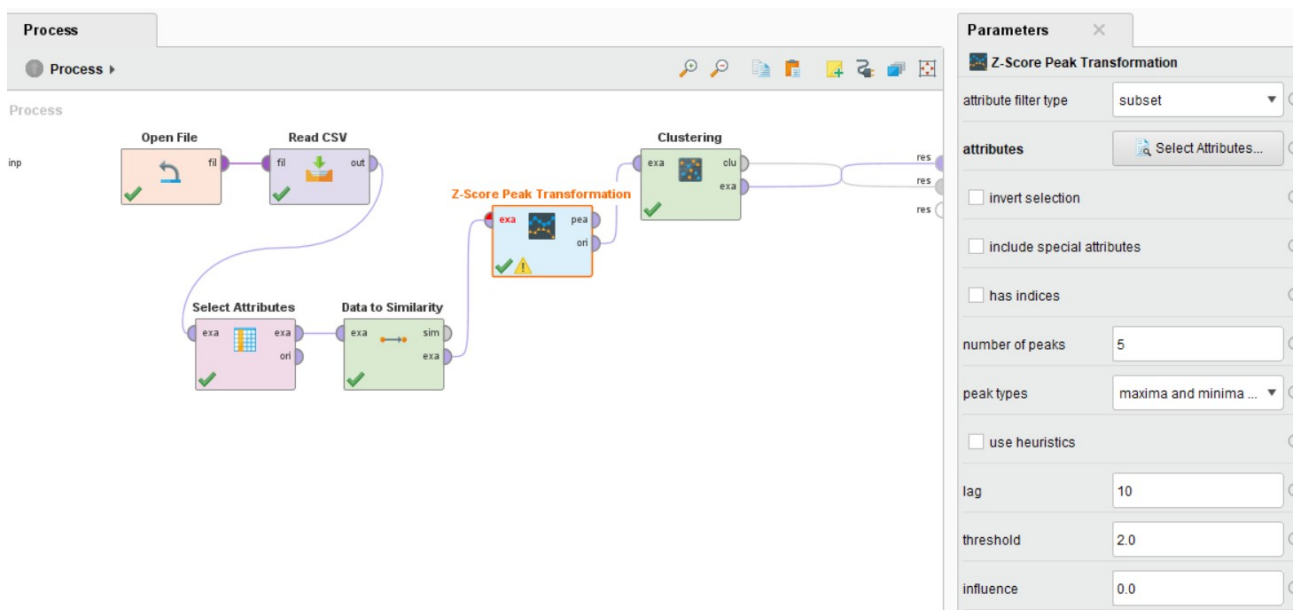
1.3

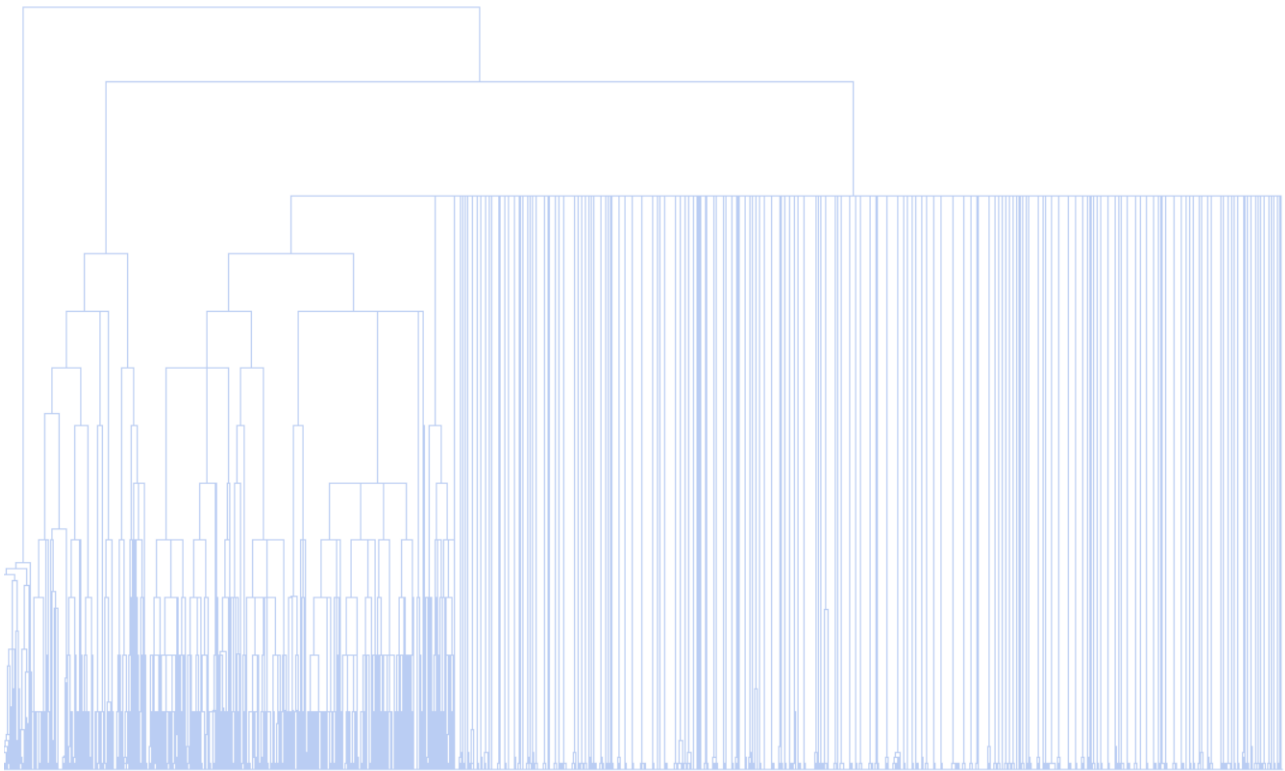


Δενδρόγραμμα:



Z-score:







1.4

Με βάση αυτά που προκύπτουν από τα παραπάνω ερωτήματα 1.1 ,1.2 και 1.3 βλέπουμε μέσα από τα δένδρογράμματα που παρήχθησαν ότι στην περίπτωση του 1.2 με δείκτη ομοιότητας Cosine και με τη μέθοδο του μέσου δεσμού στη δεύτερη και τρίτη στήλη του πίνακα `enron100` έχουμε καλύτερα αποτελέσματα στο δένδρογράμμα. Συγκεκριμένα αποτυπώνεται πολλή μεγαλύτερη πληροφορία και η τελική εικόνα είναι πιο ευκρινής για μελέτη και ανάλυση. Σε αντίθεση με τις άλλες δύο μεθόδους απλού δεσμού και δείκτη ομοιότητας Cosine και Ευκλείδεια απόσταση όπου τα αποτελέσματα δεν είναι εύκολο να διαβαστούν. Επίσης πολλή μεγάλη πληροφορία αποτυπώνεται στο μετασχηματισμό z-score.

Ακολουθούν εικόνες από τους operators που χρησιμοποιήθηκαν με τις παραμέτρους τους

Parameters ✕

 **Read CSV**

 Import Configuration Wizard... ⓘ

csv file ⓘ

column separators ⓘ

☐ trim lines ⓘ

☒ use quotes ⓘ


quotes character ⓘ

escape character ⓘ

☐ skip comments ⓘ

starting row ⓘ


Parameters ✕

 **Open File**

resource type ⓘ


filename ⓘ

Parameters ✕

 **Iris (Retrieve)**

repository entry ⓘ


Parameters ✕

 **Data to Similarity**

measure types ⓘ

numerical measure ⓘ

Parameters ✕


 **Clustering (Agglomerative Clustering)**

mode ⓘ

measure types ⓘ

numerical measure ⓘ

Parameters ✕

 **Clustering (Agglomerative Clustering)**

mode ⓘ

measure types ⓘ

numerical measure ⓘ

Parameters X

Z-Score Peak Transformation

attribute filter type: subset

attributes: Select Attributes...

☐ invert selection

☐ include special attributes

☐ has indices

number of peaks: 5

peak types: maxima and minima co...

☐ use heuristics

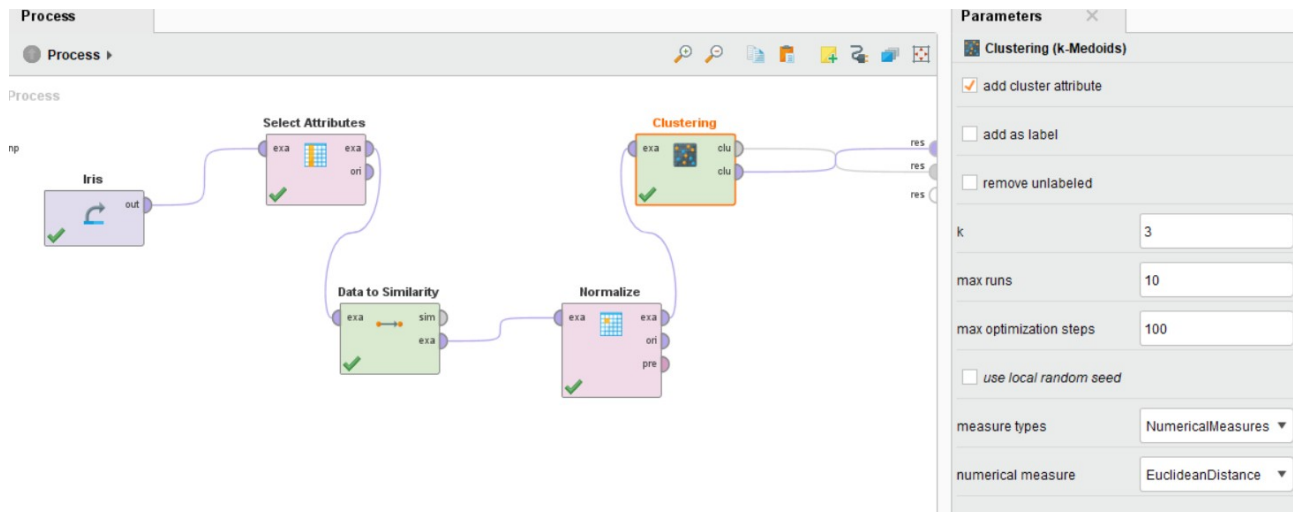
lag: 10

threshold: 2.0

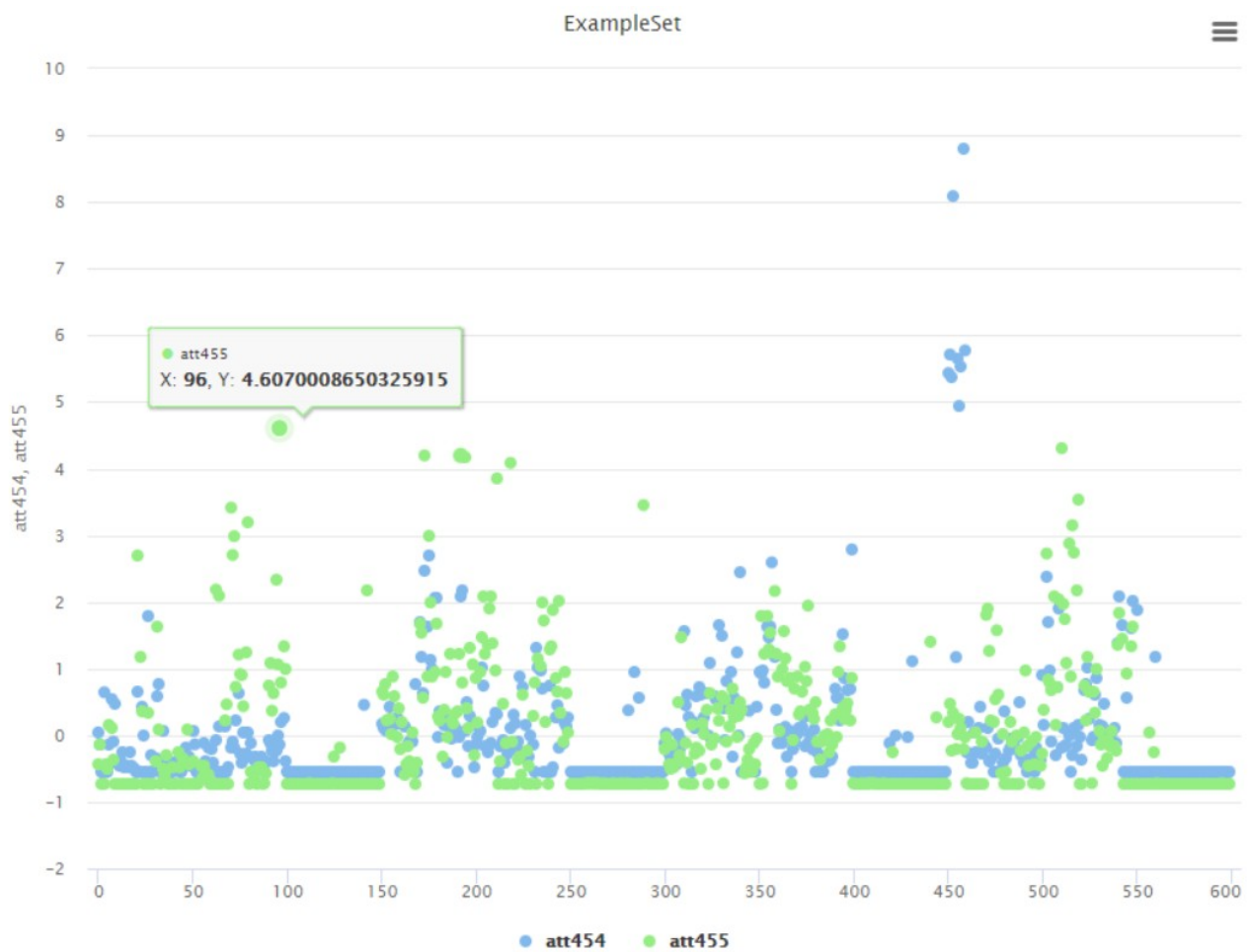
influence: 0.0

ΜΕΡΟΣ 2ο

2.1



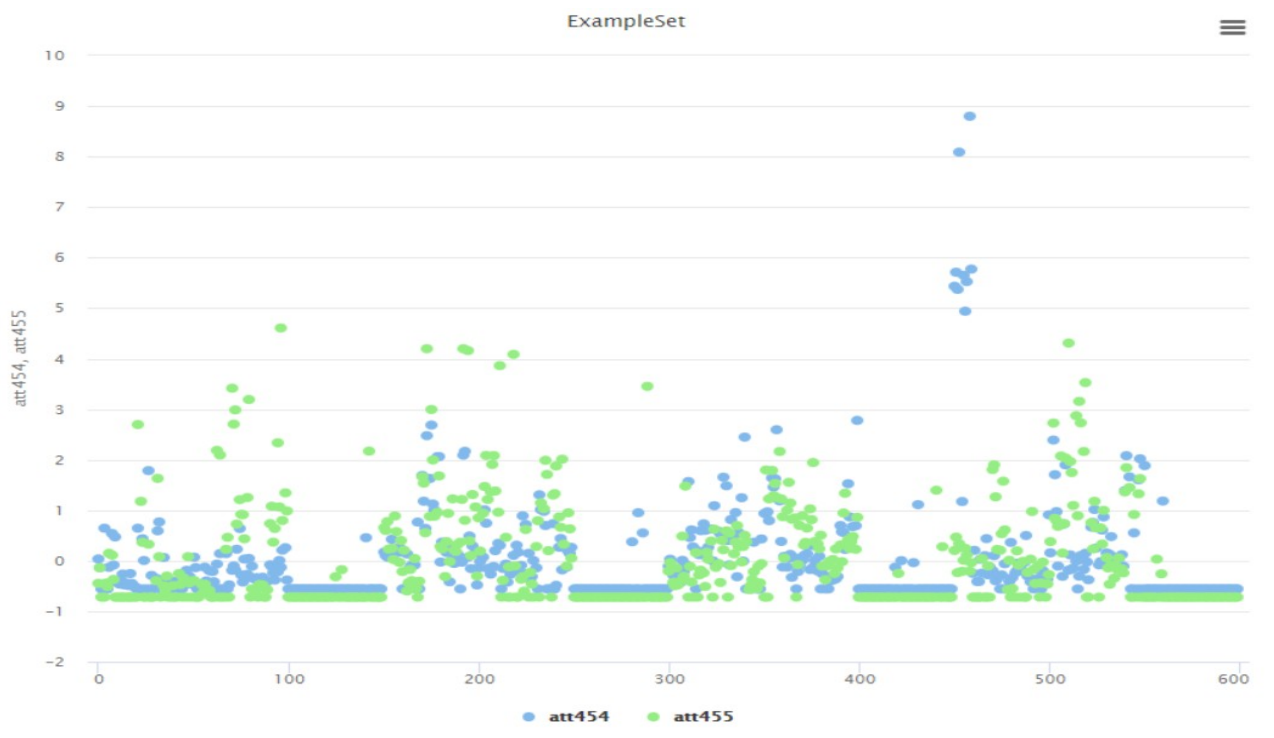
Scatter:



Αριθμητικά στατιστικά στοιχεία

Id	Integer	0	Min	Max	Average
id	Integer	0	1	600	300.500
Cluster	Nominal	0	Least	Most	Values
cluster	Nominal	0	cluster_2 (9)	cluster_0 (433)	cluster_0 (433), cluster_1 (158), ...[1 m
att454	Numeric	0	Min	Max	Average
att454	Numeric	0	-0.556	8.798	0
att455	Numeric	0	Min	Max	Average
att455	Numeric	0	-0.719	4.607	0

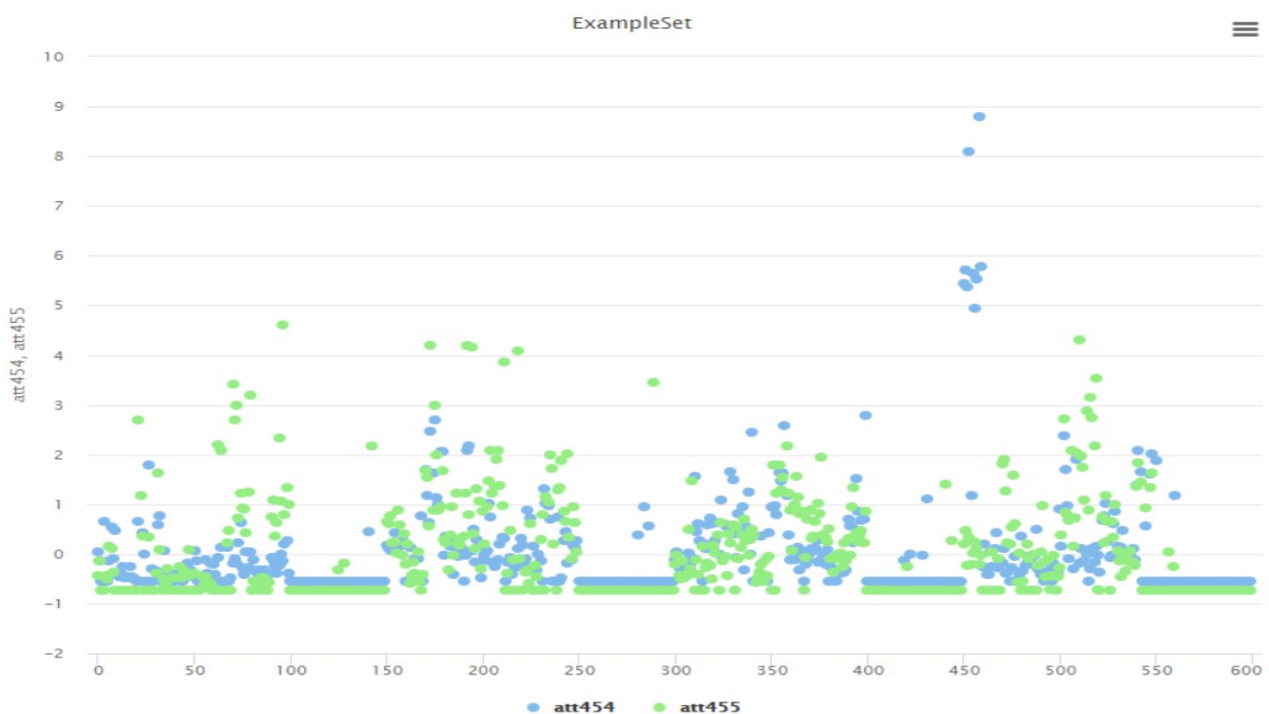
2.2 Manhattan scatter



Manhattan Αριθμητικά στατιστικά στοιχεία

Id		Integer	0	Min	1	Max	600	Average	300.500
Cluster		Nominal	0	Least	cluster_2 (9)	Most	cluster_0 (432)	Values	cluster_0 (432), cluster_1 (159), .
att454		Numeric	0	Min	-0.556	Max	8.798	Average	0
att455		Numeric	0	Min	-0.719	Max	4.607	Average	0

Cosine scatter



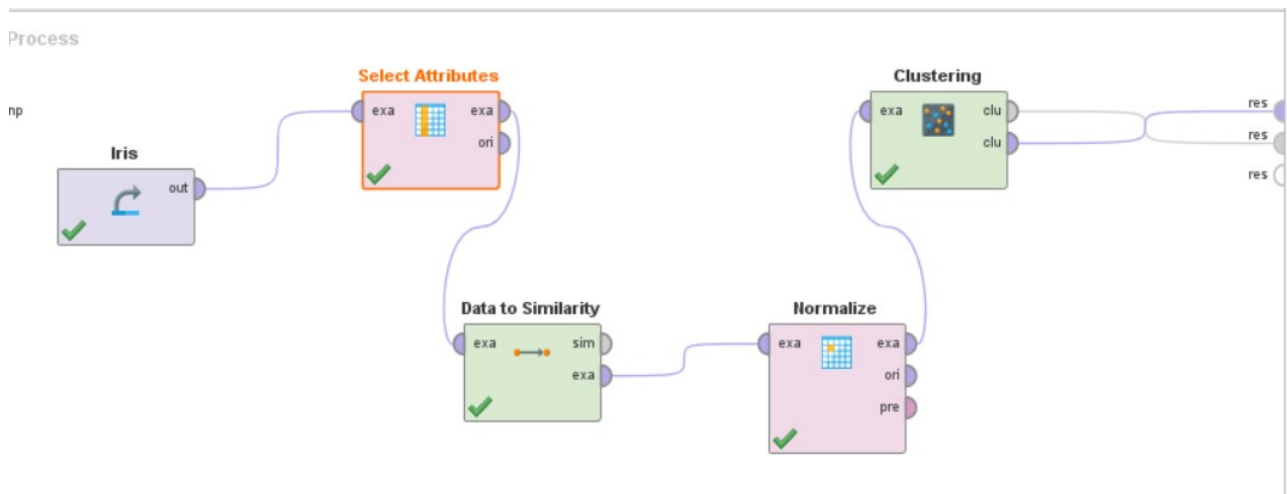
Cosine Αριθμητικά στατιστικά στοιχεία

Id id	Integer	0	Min 1	Max 600	Average 300.500
Cluster cluster	Nominal	0	Least cluster_1 (101)	Most cluster_0 (355)	Values cluster_0 (355), cluster_2 (144), .
att454	Numeric	0	Min -0.556	Max 8.798	Average 0
att455	Numeric	0	Min -0.719	Max 4.607	Average 0

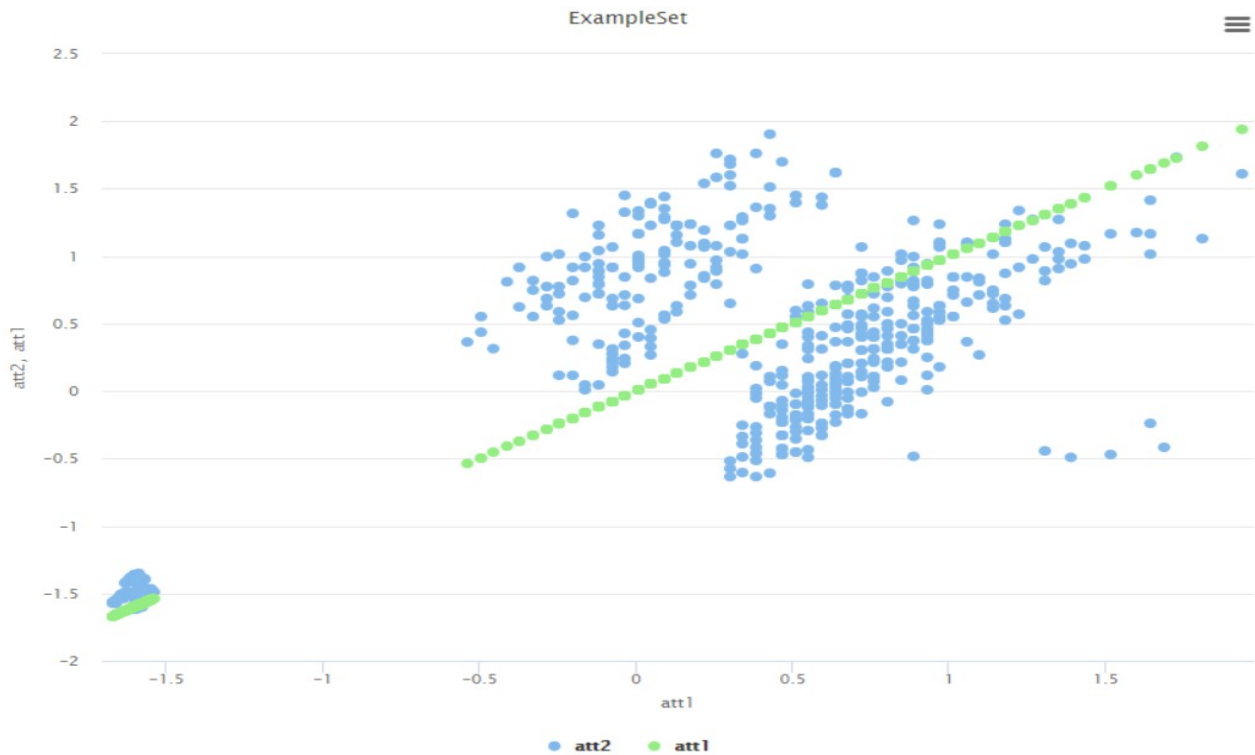
2.3

Παρατηρώντας τα παραπάνω αποτελέσματα που προκύπτουν καταλήγουμε στο τελικό συμπέρασμα ότι και με τα τρία μέτρα ομοιότητας που χρησιμοποιήθηκαν Ευκλείδεια απόσταση, Manhattan και Cosine τα αποτελέσματα είναι πανομοιότυπα και δεν διακρίνονται κάποιες εμφανείς διαφορές καθώς χρησιμοποιούμε την ίδια διαμεριστική μέθοδο συσταδοποίησης k-μέσων.

2.4



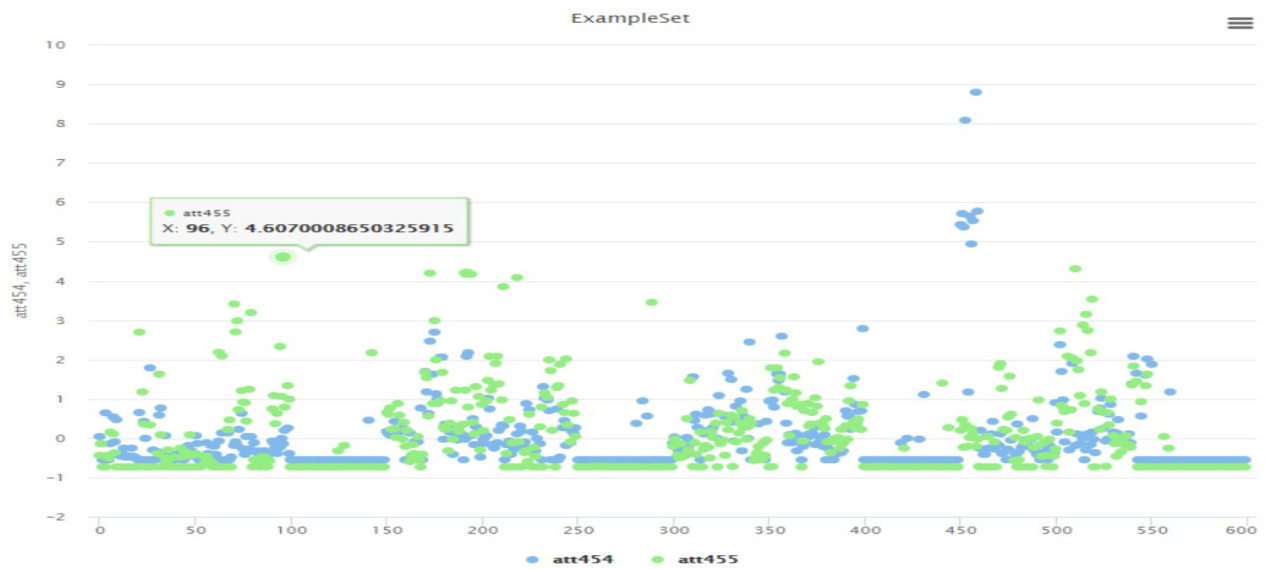
Scatter:



Αριθμητικά στατιστικά στοιχεία

Name	Type	Missing	Statistics			Filter (4 / 4 attributes): <input type="text" value="Search for Attributes"/>
id	Integer	0	Min	Max	Average	
id	Integer	0	1	600	300.500	
cluster	Nominal	0	Least cluster_2 (150)	Most cluster_1 (228)	Values cluster_1 (228), cluster_0 (222), ...[1	
att1	Numeric	0	Min -1.670	Max 1.938	Average -0.000	
att2	Numeric	0	Min -1.618	Max 1.906	Average 0.000	

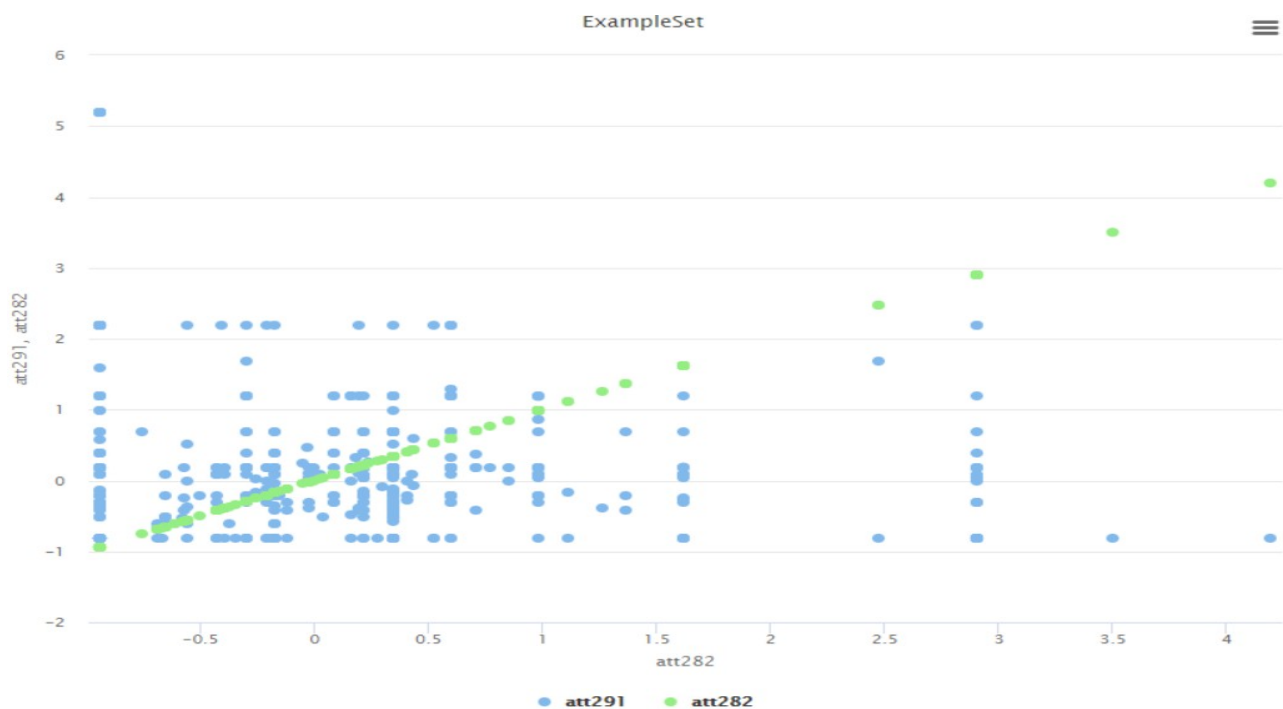
2.5 Scatter



id			Min	Max	Average
id	Integer	0	1	600	300.500
Cluster	Nominal	0	Least cluster_2 (9)	Most cluster_0 (433)	Values cluster_0 (433), cluster_1 (158), ...[1 m
att454	Numeric	0	Min -0.556	Max 8.798	Average 0
att455	Numeric	0	Min -0.719	Max 4.607	Average 0

2.6

Scatter



Αριθμητικά στατιστικά στοιχεία

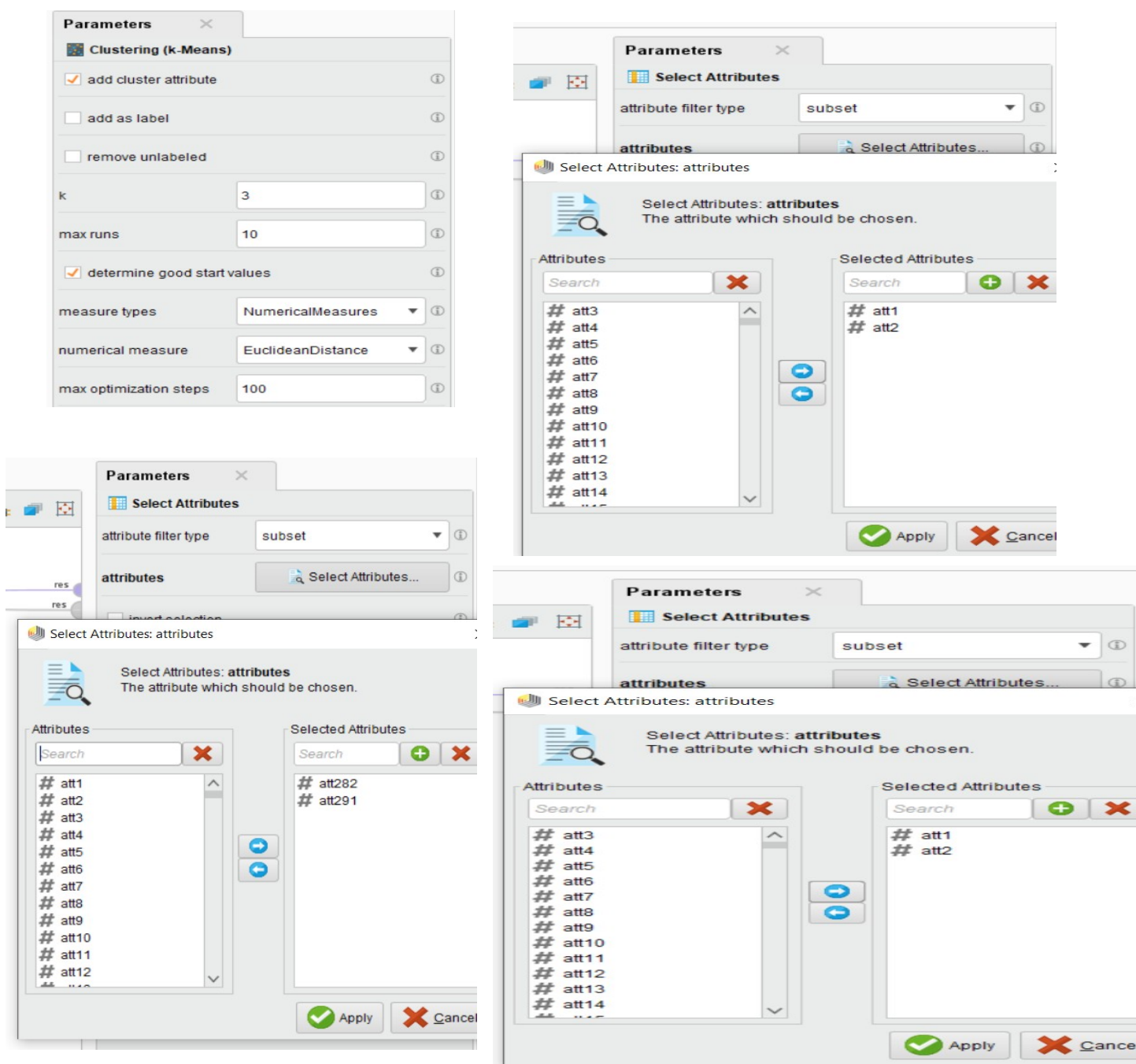
Id	Integer	0	Min 1	Max 600	Average 300.500
cluster	Nominal	0	Least cluster_1 (58)	Most cluster_0 (470)	Values cluster_0 (470), cluster_2 (72), ...[1 m
att282	Numeric	0	Min -0.939	Max 4.193	Average -0.000
att291	Numeric	0	Min -0.815	Max 5.195	Average -0.000

2.7

Εκτελώντας τα παραπάνω βήματα 2.4, 2.5 και 2.6 φτάνουμε στο συμπέρασμα ότι στην περίπτωση που χρησιμοποιήθηκαν όλα τα χαρακτηριστικά του πίνακα xV το αποτέλεσμα στο γράφημα είναι πιο πλήρες και αποτυπώνει μεγαλύτερη πληροφορία σε σχέση με τα υπόλοιπα δύο. Ενώ στις άλλες δύο περιπτώσεις το

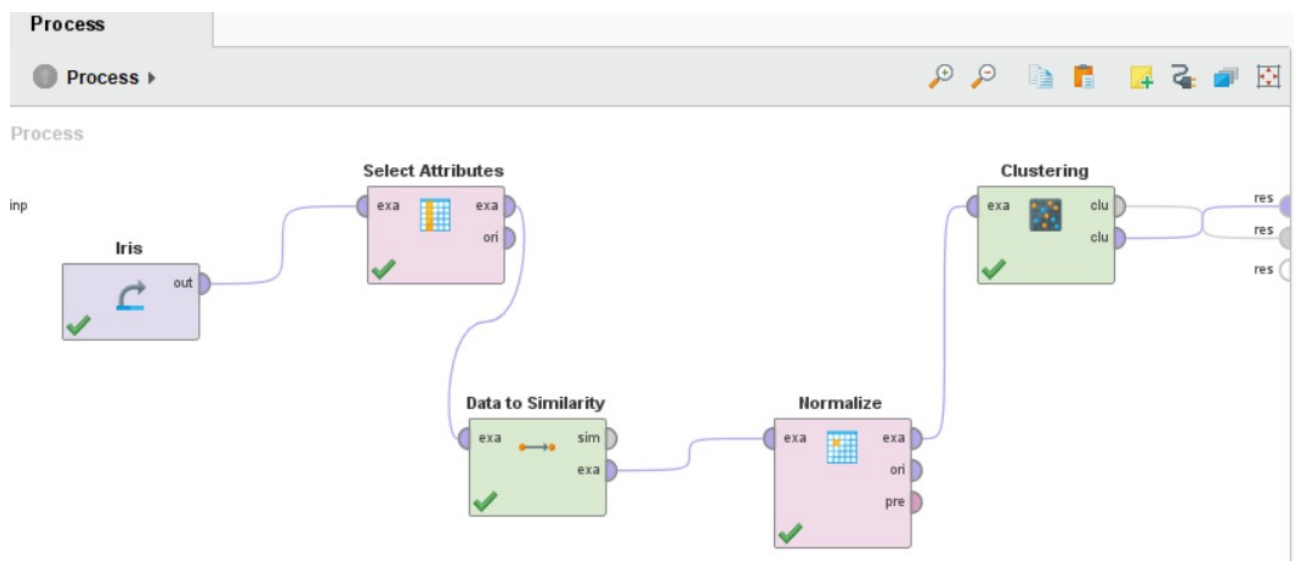
γραφήματα είναι πιο αραιά και δεν έχουν μεγάλη πυκνότητα των στοιχείων όπως συμβαίνει στο 2.6, επειδή γίνεται χρήση μόνο δύο χαρακτηριστικών την φορά. Όσον αφορά το μέτρο παρατηρούμε ότι το μέτρο ομοιότητας Cosine είναι αυτό που αποδίδει καλύτερα όπως φαίνεται και στα παραπάνω ερωτήματα δίνοντας πιο λεπτομερή γραφήματα σε αντίθεση με τα μέτρα απόστασης Ευκλείδεια και Manhattan.

Ακολουθούν εικόνες από τους operators που χρησιμοποιήθηκαν με τις παραμέτρους τους



ΜΕΡΟΣ 3ο

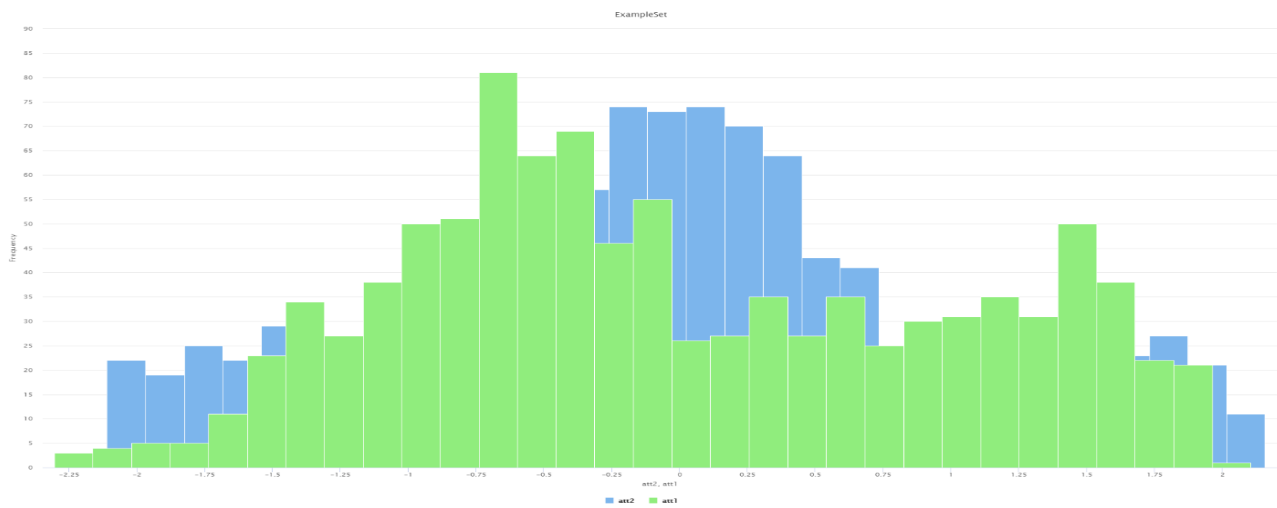
3.1



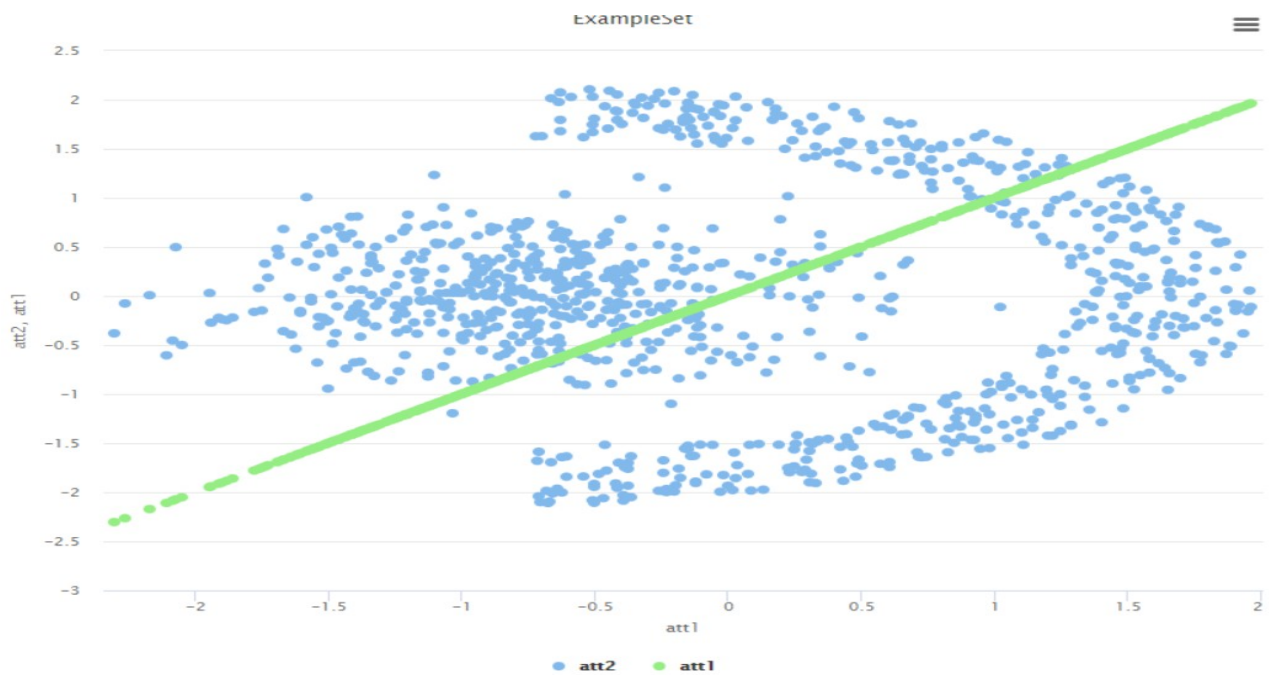
1ο γράφημα



2ο γράφημα

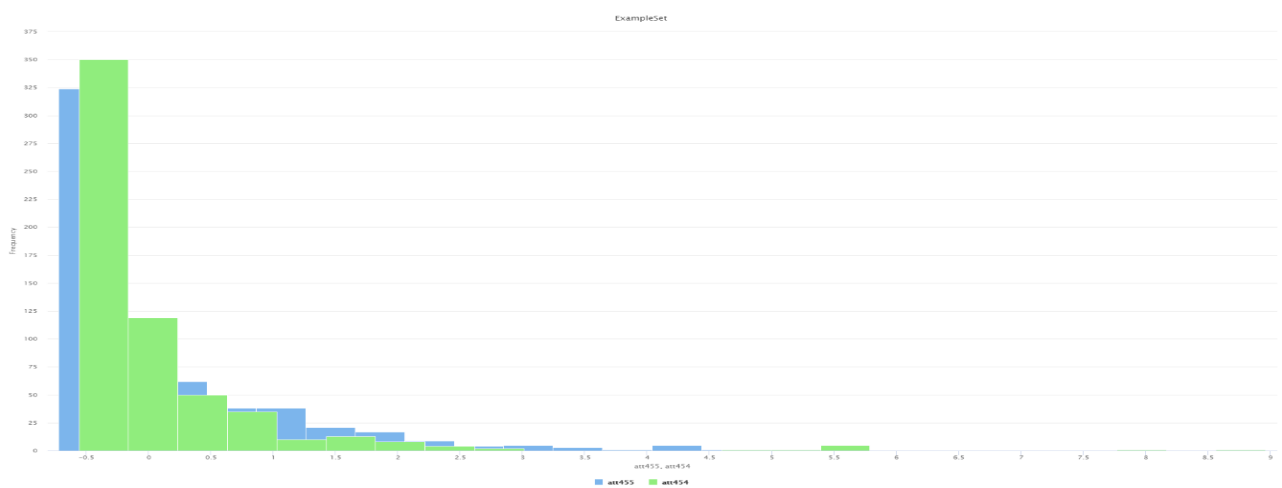
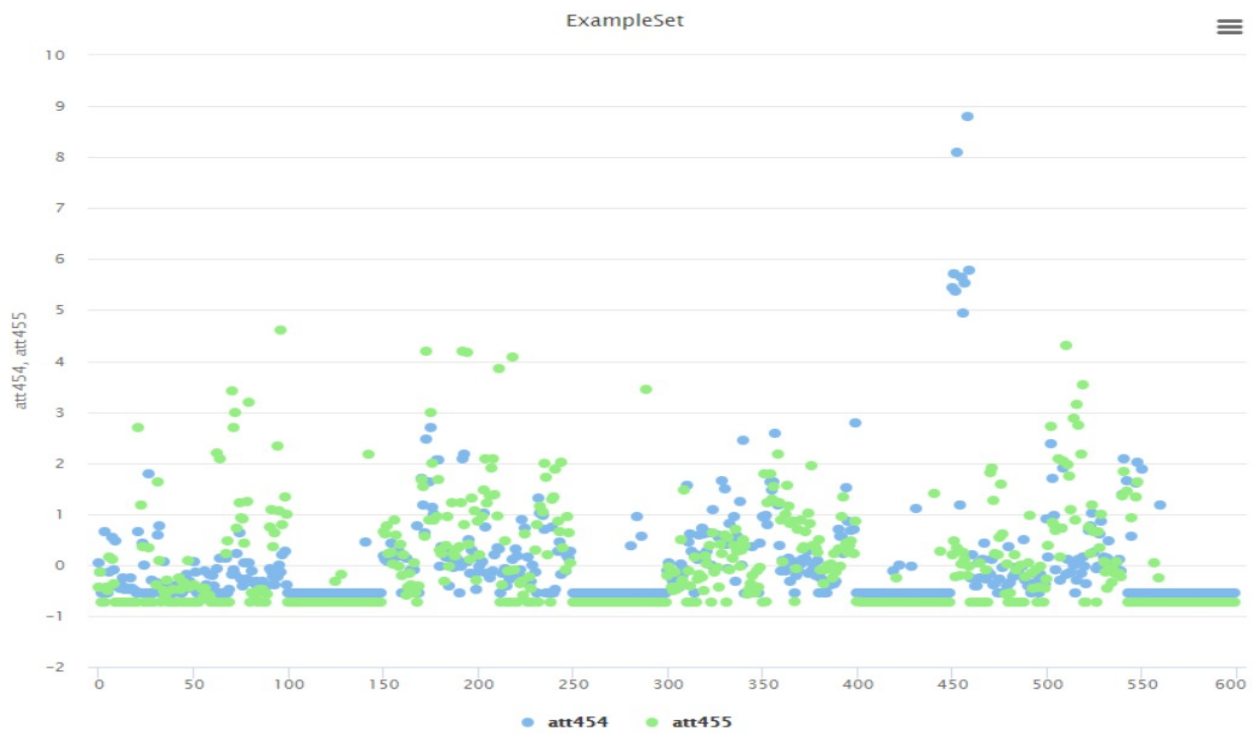


3.2

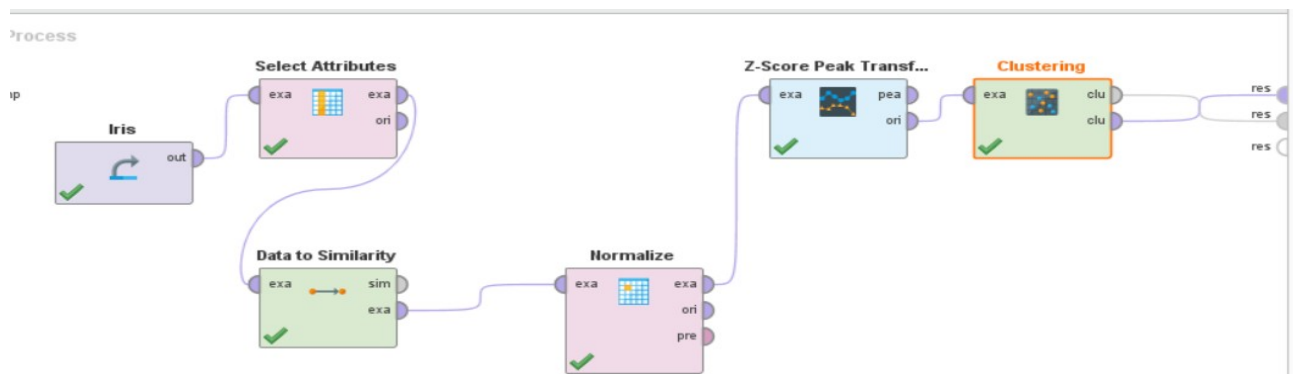


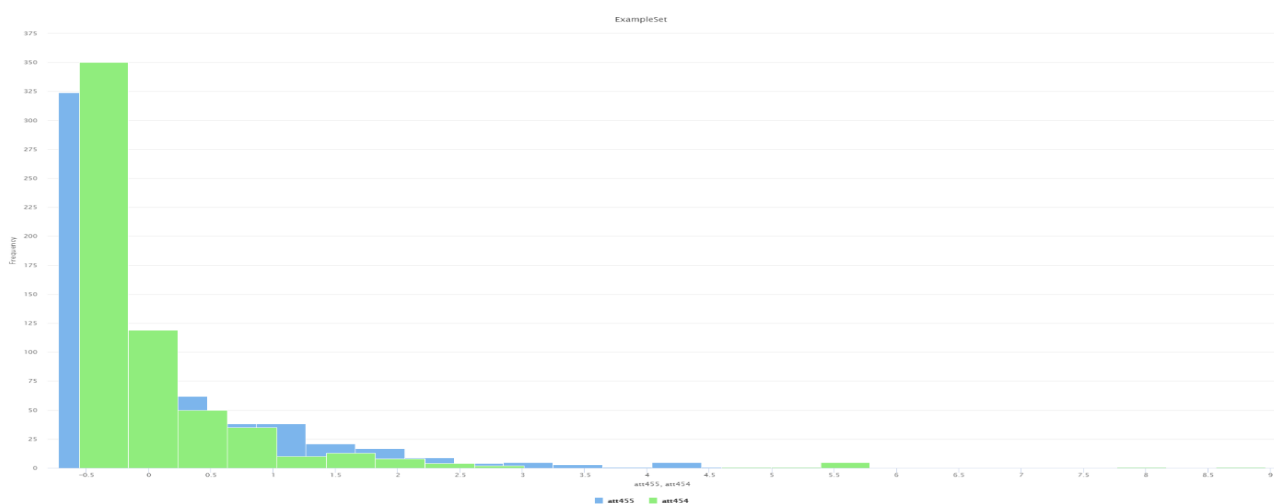
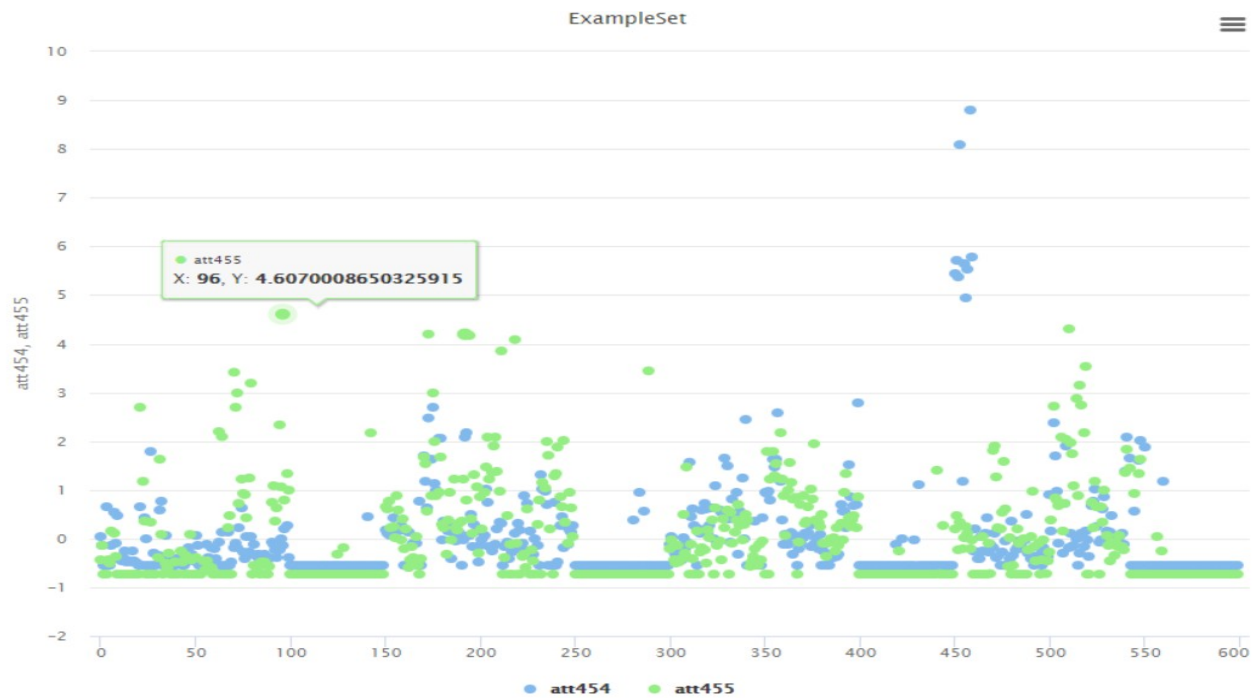
Στα δύο ερωτήματα συσταδοποίησης 3.1 και 3.2 παρατηρούμε στα γραφήματα διασποράς ότι με τις δύο μεθόδους DBSCAN και k-μέσων παράγονται αρκετά παρόμοια γραφήματα με την Ευκλείδεια απόσταση, όμως παρατηρούμε ότι στο δεύτερο γράφημα υπάρχει μία οπτική διαφορά καθώς φαίνεται να απεικονίζεται μεγαλύτερη πληροφορία από ότι στο πρώτο με αποτέλεσμα να είναι πιο πλήρες.

3.3



3.4



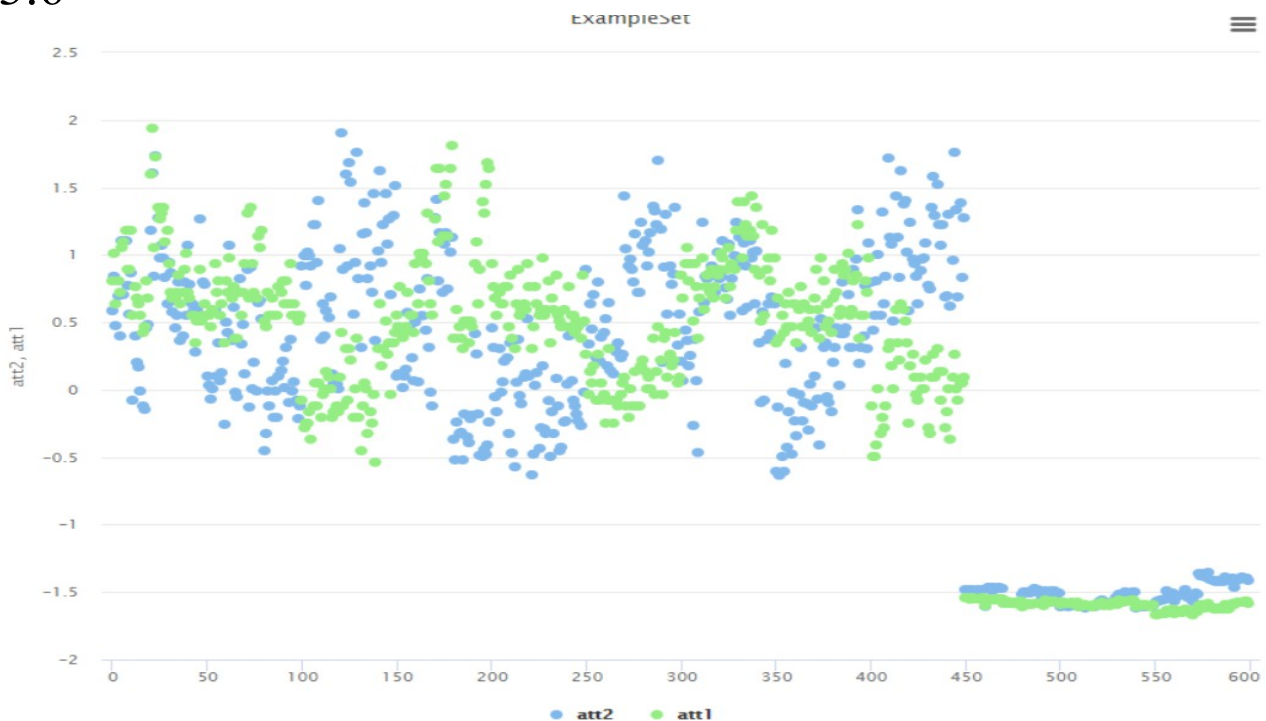


3.5

Στα προηγούμενα βήματα 3.3 και 3.4 χρησιμοποιήθηκαν μέθοδοι συσταδοποίησης DBSCAN στα δεδομένα iris και στο 3.4 έγινε κανονικοποίηση των δεδομένων με μέθοδο z-score. Έτσι παρατηρώντας τα αποτελέσματα προκύπτει η άποψη ότι και στις δύο μεθόδους έχουμε μεγάλη ομοιότητα στο τέλος και δεν είναι εύκολο να φανούν κάποιες διαφορές

είτε στο γράφημα διασποράς είτε στο γράφημα με τις συστάδες.

3.6



3.7

Επέλεξα για τιμές των παραμέτρων το 0.9 και το 40 αντίστοιχα για ϵ και MinPts



3.8

Στα δύο παραπάνω παραδείγματα 3.6 και 3.7 βλέπουμε ότι στα διαγράμματα διασποράς δεν υπάρχουν μεγάλες διαφορές σε σημείο που θα μπορούσαμε να πούμε ότι είναι ίδια. Αυτό συμβαίνει γιατί δεν υπήρξε μεγάλη αλλαγή στις τιμές των παραμέτρων (από 0.5 σε 0.9 και από 50 σε 40). Σε άλλη περίπτωση με πιο μεγάλες παραμέτρους θα παρατηρούσαμε μία πιο αισθητή διαφορά στα δύο γραφήματα

Ακολουθούν εικόνες από τους operators που χρησιμοποιήθηκαν με τις παραμέτρους τους

Parameters X

Clustering (DBSCAN)

epsilon 0.5 ⓘ

min points 15 ⓘ

☒ add cluster attribute ⓘ


☐ add as label ⓘ

☐ remove unlabeled ⓘ

Parameters X

Select Attributes

attribute filter type subset ⓘ


attributes  Select Attributes... ⓘ

☐ invert selection ⓘ


Select Attributes: attributes

Select Attributes: **attributes**
The attribute which should be chosen.



Attributes



Search 

Selected Attributes

Search 

att1
att2

 Apply 

Parameters X

Clustering (k-Means)

☒ add cluster attribute ⓘ

☐ add as label ⓘ

☐ remove unlabeled ⓘ

k 3 ⓘ

max runs 10 ⓘ

☒ determine good start values ⓘ

measure types NumericalMeasures ⓘ

numerical measure EuclideanDistance ⓘ

max optimization steps 100 ⓘ

Parameters X

Clustering (DBSCAN)

epsilon 0.1 ⓘ

min points 5 ⓘ

☒ add cluster attribute ⓘ

☐ add as label ⓘ

☐ remove unlabeled ⓘ

measure types NumericalMeasures ⓘ

numerical measure EuclideanDistance ⓘ

Parameters X

Clustering (DBSCAN)

epsilon 0.5 ⓘ

min points 50 ⓘ


☒ add cluster attribute ⓘ

☐ add as label ⓘ

Parameters X

Select Attributes

attribute filter type subset ⓘ


attributes  Select Attributes... ⓘ

☐ invert selection ⓘ

Select Attributes: attributes


Select Attributes: **attributes**
The attribute which should be chosen.

Attributes



Search 



att1
att2
att3
att4
att5
att6
att7
att8
att9
att10
att11

Selected Attributes


Search 

att454
att455

 Apply 

Parameters ✕


 **Clustering (DBSCAN)**

epsilon ⓘ

min points ⓘ


☒ add cluster attribute ⓘ


Parameters ✕

 **Select Attributes**

attribute filter type ⓘ

attributes ⓘ

 **Select Attributes: attributes**

 **Select Attributes: **attributes****
The attribute which should be chosen.

Attributes

✕

- # att3
- # att4
- # att5
- # att6
- # att7
- # att8
- # att9
- # att10
- # att11

Selected Attributes

- # att1
- # att2

▶