ΠΑΝΕΠΙΣΤΗΜΙΟ ΜΑΚΕΔΟΝΙΑΣ
ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ
ΤΜΗΜΑΤΟΣ ΕΦΑΡΜΟΣΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΚΗΣ

Ο ΑΛΓΟΡΙΘΜΟΣ K-MEANS ΣΕ PYTHON

Κωνσταντίνου Τσολάκη

Σκοπός

- Η παρούσα εργασία εντάσσεται στο πλαίσιο της τεχνικής Εξόρυξης Γνώσης από Δεδομένα που ονομάζεται ομαδοποίηση ή συσταδοποίηση (clustering).

- Η εν λόγω μέθοδος αφορά στη στατιστική ανάλυση και ταξινόμηση όμοιων ή συσχετιζόμενων αντικειμένων σε υποσύνολα/ ομάδες (clusters) έτσι ώστε αυτά να μοιράζονται κοινά χαρακτηριστικά (Δημητρακοπούλου, 2007).

- Η εργασία εστιάζει σε έναν από τους αλγόριθμους συσταδοποίησης, τον αλγόριθμο k-means, ο οποίος παρότι παρουσιάστηκε πριν περίπου 50 χρόνια, παραμένει ο πιο διαδεδομένος λόγω της απλότητας, της ευκολίας στην εφαρμογή και της αποτελεσματικότητας που τον διακρίνουν (Jain, 2010).

- Ο εν λόγω αλγόριθμος υλοποιείται με χρήση της γλώσσας προγραμματισμού Python (βλ. www.python.org).

Γενικά

Η χειροκίνητη εξαγωγή προτύπων από δεδομένα συμβαίνει εδώ και αιώνες. Οι πρώτες μέθοδοι για τον προσδιορσό προτύοων ήταν αυτές της θεωρίας του Bayes και της ανάλυσης παλινδρόμησης.

Η ολοένα και μεγαλύτερη αύξηση του όγκου δεδομένων προς επεξεργασία, οδήγησε στην γέννηση του όρου “Εξόρυξη Γνώσης από Δεδομένα”.

Μερικά παραδείγματα μεθόδων που χρησιμοποίουνται για αυτόν τον το σκόπο είναι:

- Οι **Κανόνες Συσχέτισης** (Association Rules)
- Η **Κατηγοριοποίηση** (Classification)
- Η **Δομημένη Πρόβλεψη** (Structured Prediction)
- Η **Συσταδοποίηση** (Clustering) κ.α.

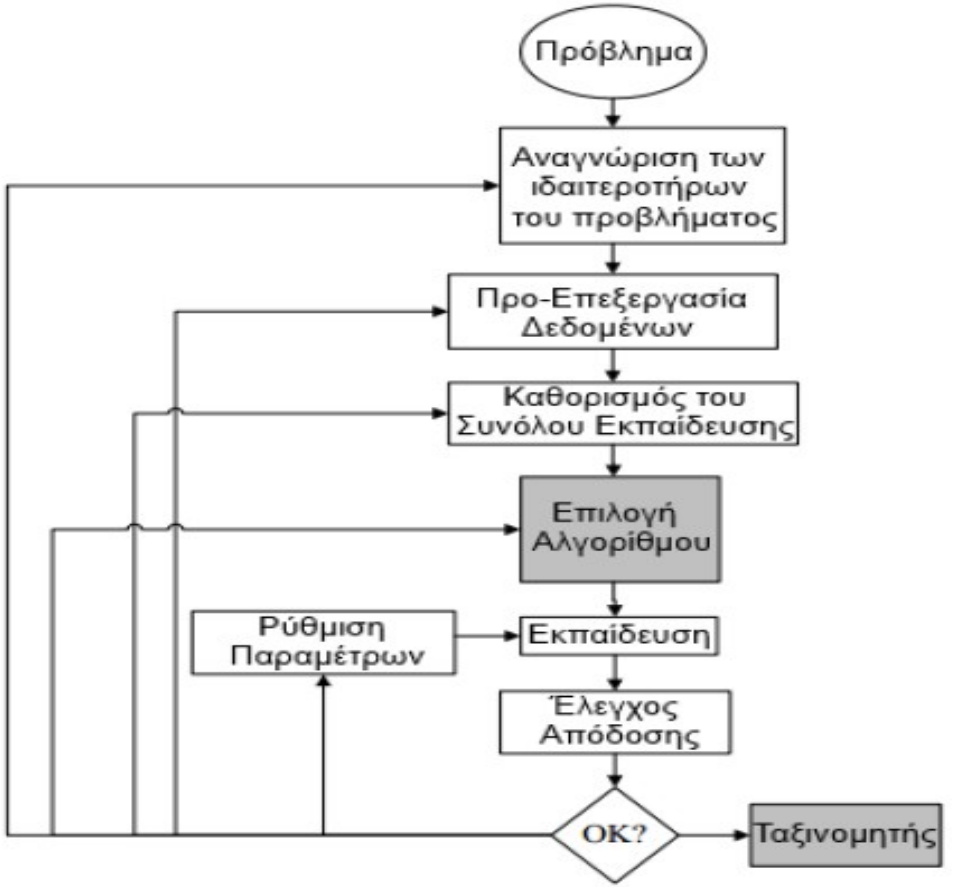
Με την τελευταία θα ασχοληθεί η συγκεκριμένη εργασία.

Μηχανική Μάθηση

Ο όρος ‘Μηχανική μάθηση’ αναφέρεται στον τρόπο με τον οποίο, μέσω προγραμμάτων ηλεκτρονικών υπολογιστών, αναγνωρίζονται περίπλοκα πρότυπα μέσα στα δεδομένα και να λάβουν έξυπνες αποφάσεις.

Το πεδίο περιλαμβάνει διάφορες εκδοχές, όπως (Han, 2012):

- Η καθοδηγούμενη μάθηση
- Η μη καθοδηγούμενη μάθηση
- Η ημι-καθοδηγούμενη μάθηση
- Η ενεργός μάθηση



Εικόνα 1: Διάγραμμα ροής αλγόριθμου μηχανικής μάθησης (Γούλας, 2015)

Συσταδοποίηση

Συνηθέστερα χρησιμοποιούμενη μέθοδος μη επιβλεπόμενης μάθησης αποτελεί αυτή της συσταδοποίησης, η οποία αφορά στο διαχωρισμό ενός συνόλου αντικειμένων σε επιμέρους υποσύνολα (συστάδες) με βάση κάποιο κοινό χαρακτηριστικό ή σχέση. (Αφεντουλίδης, 2015). Η συσταδοποίηση είναι στην ουσία μια διερευνητική τεχνική με σκοπό την ανεύρεση δομής σε ένα πλήθος δεδομένων.

Η μέθοδοι συσταδοποίησης κατηγοριοποιούνται με διάφορους τρόπους, ανάλογα με τα χαρακτηριστικά των αντικειμένων ή/ και το σκοπό εφαρμογής της εκάστοτε μεθόδου. Έτσι, αυτές διαχωρίζονται ανάλογα με (Αφεντουλίδης, 2015):

- την εξάρτηση των ομάδων (επίπεδη έναντι ιεραρχικής)
- την καθολικότητα της μεθόδου (ολική έναντι μερικής)
- το είδος των ομάδων

Ο Αλγόριθμος k- means

Συσταδοποιεί δεδομένα στο ν- διάστατο ευκλείδειο χώρο, συνήθως με χρήση του Ευκλείδειας απόστασης. Ο αντιπρόσωπος κάθε ομάδας ονομάζεται κεντροειδές (centroid) και υπολογίζεται συνήθως ως ο μέσος όρος των αντικειμένων κάθε ομάδας.

Η μέθοδος υλοποιείται σε δύο φάσεις (Nidheesh, 2017):

1η Φάση: ορίζουμε, συνήθως τυχαία, κάποια αντικείμενα ως κεντροειδή. Στην συνέχεια ο αλγόριθμος υπολογίζει την απόσταση κάθε αντικειμένου από κάθε κεντροειδές (με χρήση της Ευκλείδειας απόστασης) και το αντιστοιχεί στην ομάδα του κεντροειδούς με το οποίο η απόσταση αυτή είναι η μικρότερη.

2η Φάση: επαναϋπολογίζουμε τα κεντροειδή και επαναλαμβάνουμε τη διαδικασία. Τα νέα κεντροειδή υπολογίζονται ως μέσοι όροι των αντικειμένων κάθε ομάδας που έχει δημιουργηθεί από τη 1η φάση.

Υλοποίηση του αλγόριθμου k- means στην Python

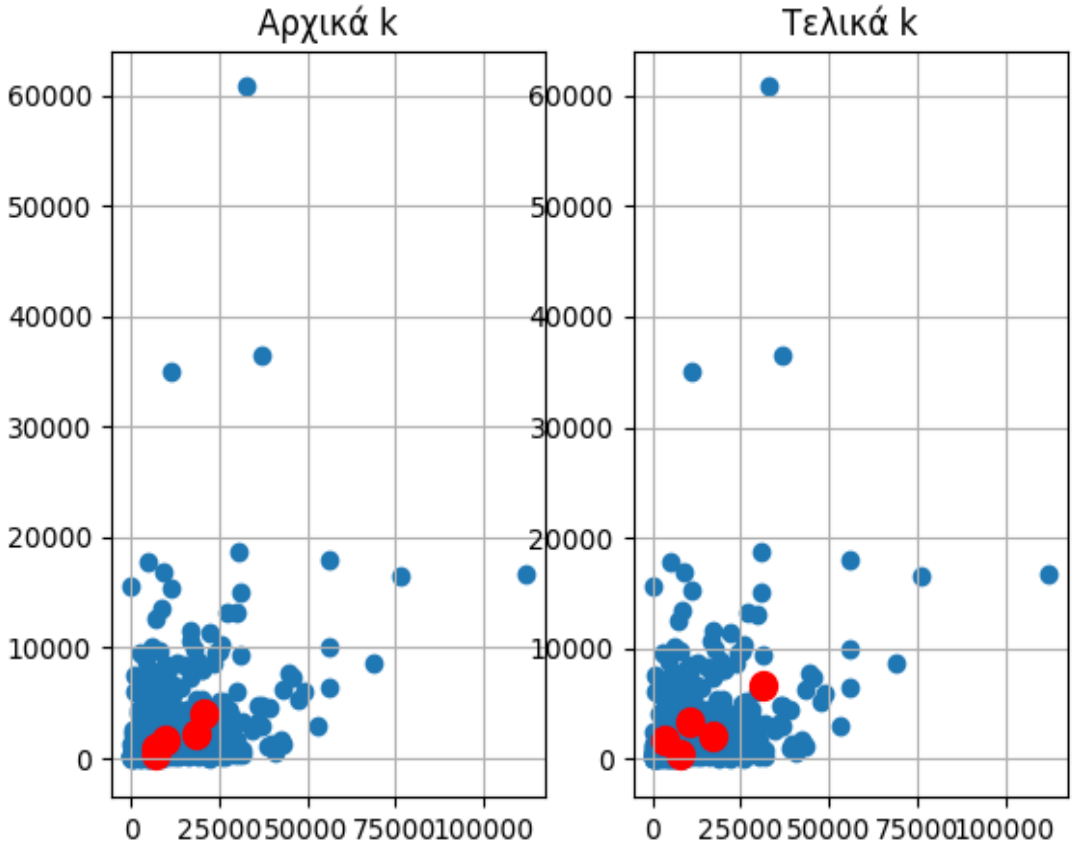
Για την υλοποίηση του αλγόριθμου χρησιμοποιήθηκε η διανομή της Python από την Anaconda (https://www.anaconda.com/) και πιο συγκεκριμένα το Spyder IDE.

Ο αλγόριθμος εφαρμόστηκε σε τέσσερα αρχεία δεδομένων. Το πρόγραμμα διαβάζει το εκάστοτε αρχείο (συγκεκριμένες στήλες που ορίζει ο χρήστης).

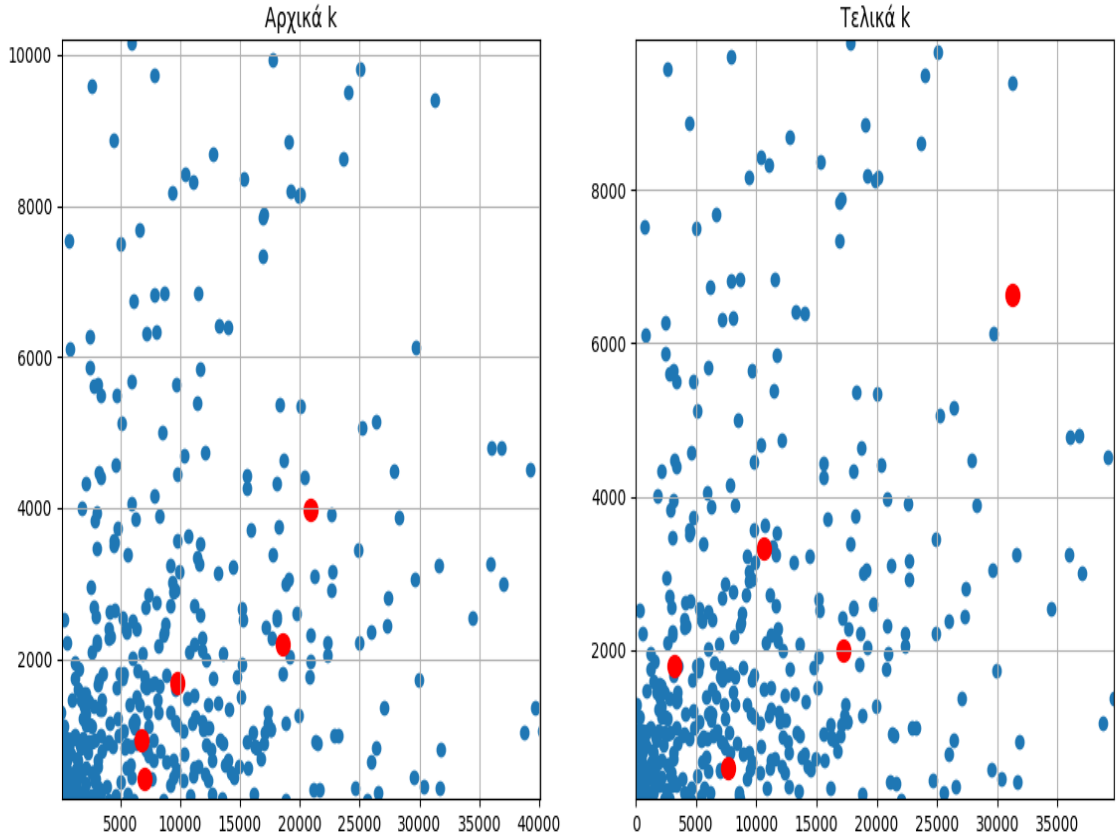
Ο χρήστης ορίζει τον αριθμό των κεντροειδών και το πρόγραμμα τα ορίζει. Στη συνέχεια ο αλγόριθμος μετράει τις αποστάσεις των δεδομένων από τα κεντροειδή με χρήση της Ευκλείδειας απόστασης και επιστρέφει τα νέα κεντροειδή. Η διαδικασία οπτικοποιείται.

Αποτελέσματα- Δεδομένα Wholesale

Το αρχείο περιέχει 440 καταγραφές (μέρος τους παρουσιάζεται στο παράρτημα 2) με δεδομένα πωλήσεων. Επιλέγονται μόνο οι στήλες Φρέσκα (Fresh) και Κατεψυγμένα (Frozen).

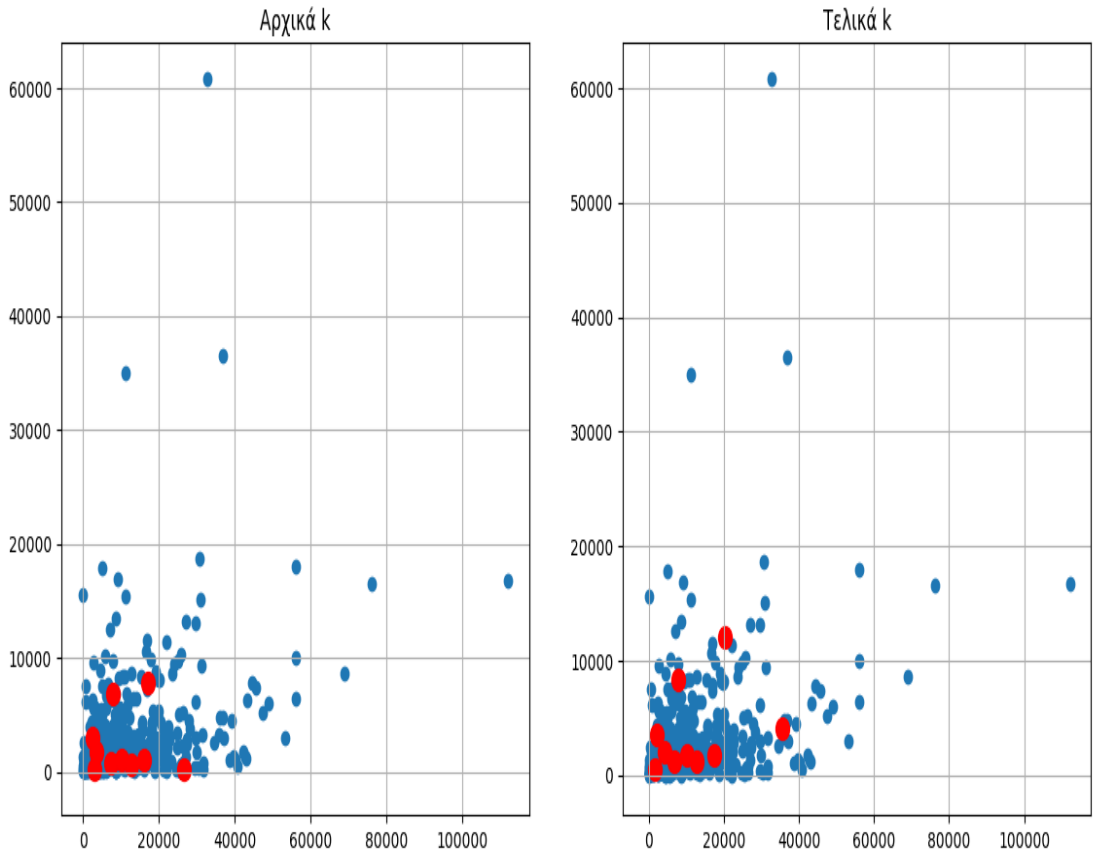


Εικόνα 6: Δεδομένα Wholesale, k=5

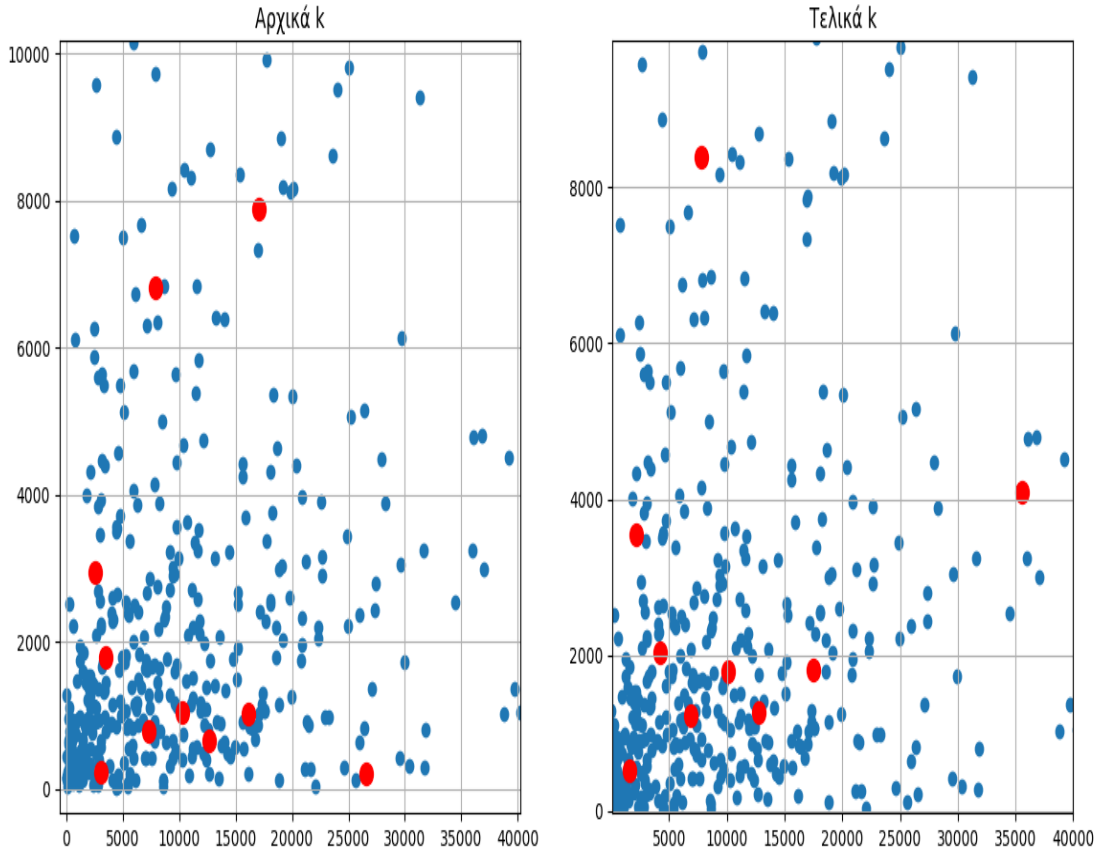


Εικόνα 7: Δεδομένα Wholesale, k=5 (Μεγέθυνση)

Αποτελέσματα- Δεδομένα Wholesale

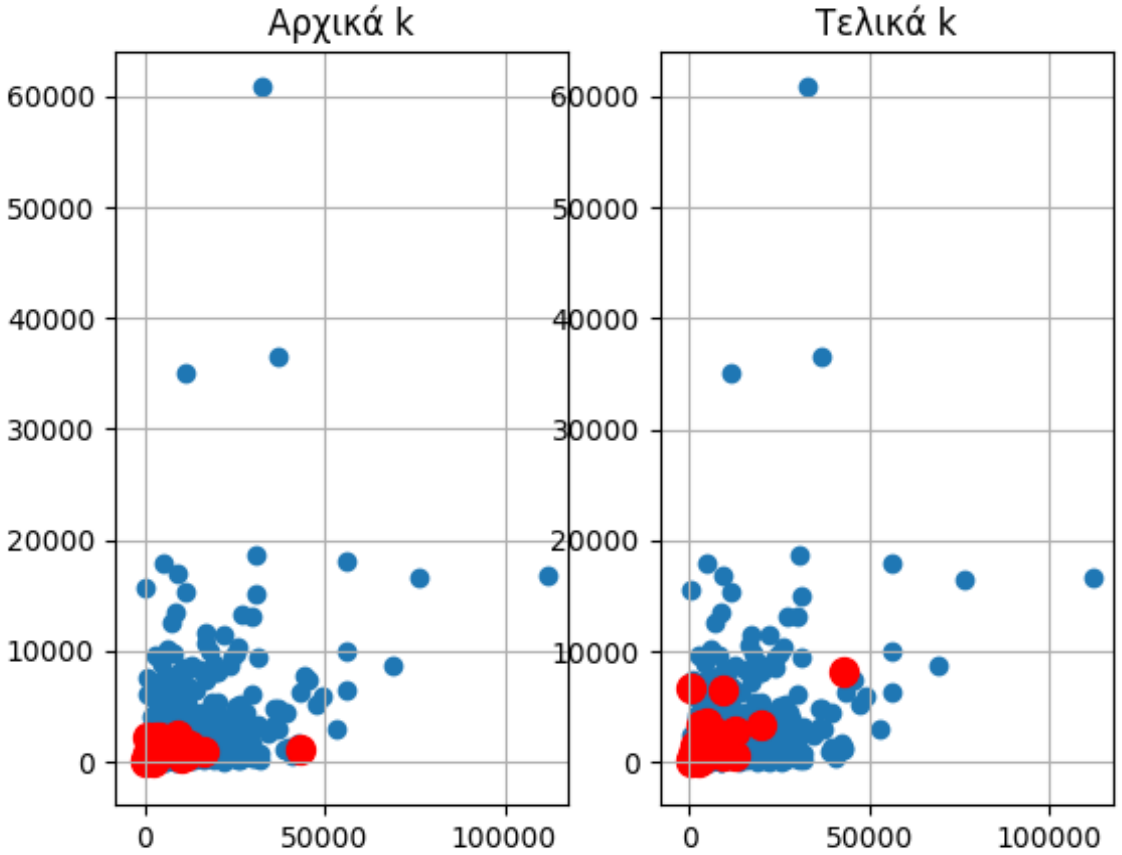


Εικόνα 8: Δεδομένα Wholesale, k=10 (Μεγέθυνση)

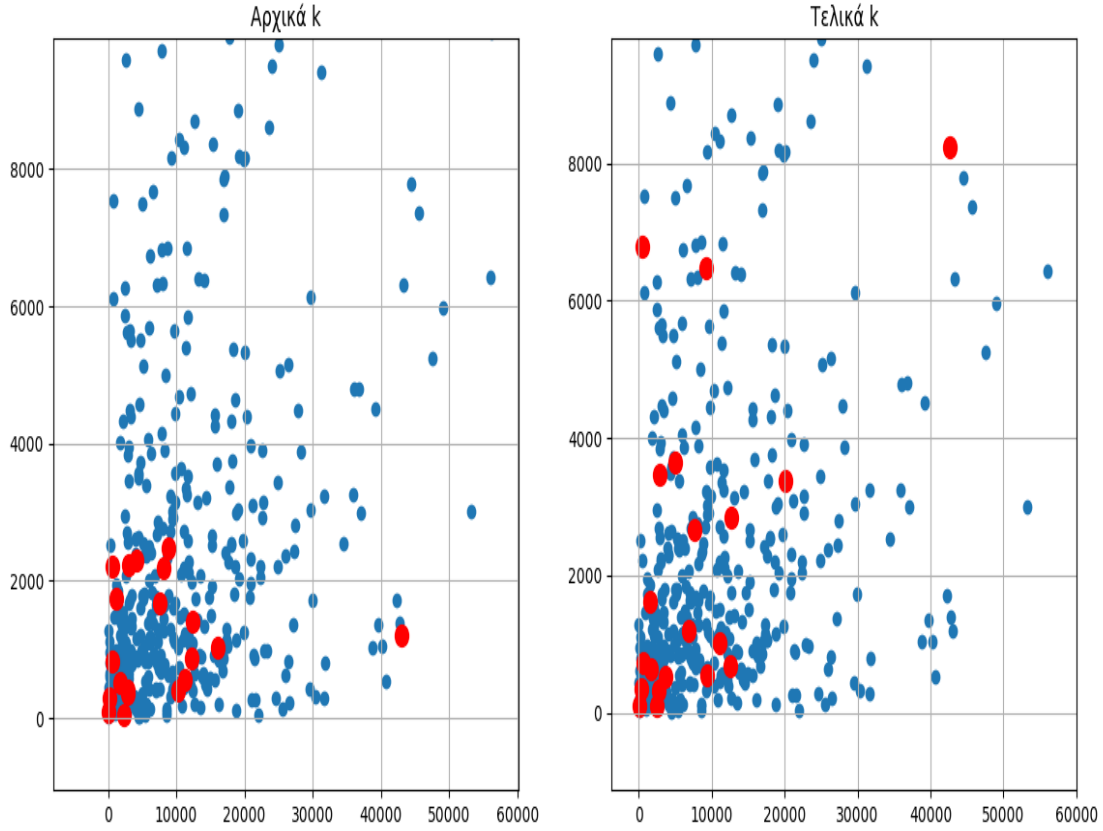


Εικόνα 9: Δεδομένα Wholesale, k=10 (Μεγέθυνση)

Αποτελέσματα- Δεδομένα Wholesale

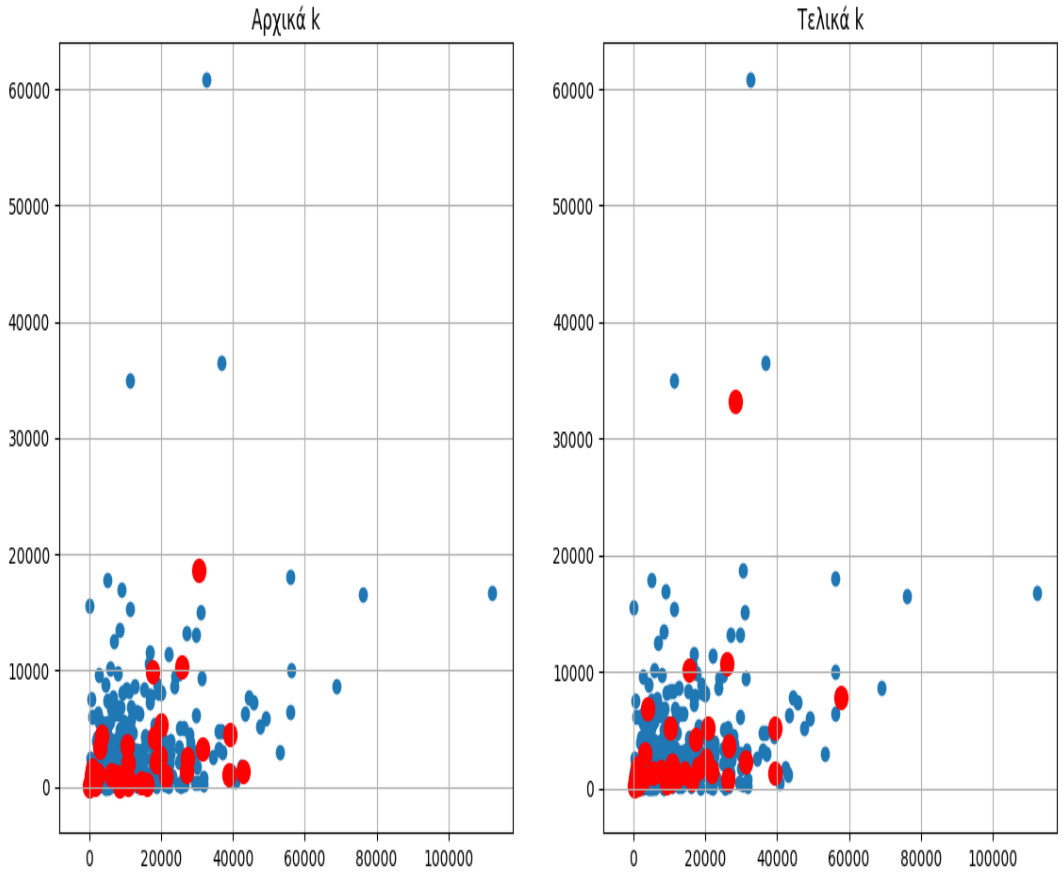


Εικόνα 10: Δεδομένα Wholesale, k=20

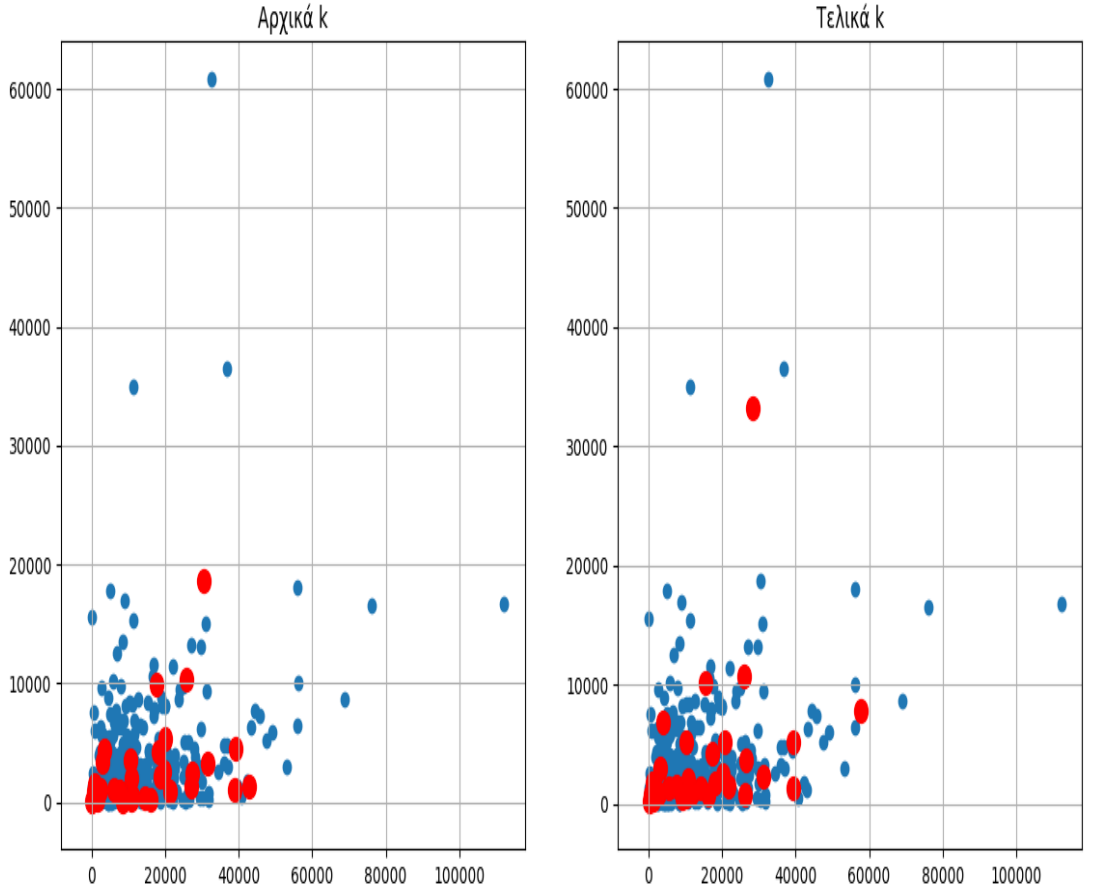


Εικόνα 11: Δεδομένα Wholesale, k=20 (Μεγέθυνση)

Αποτελέσματα- Δεδομένα Wholesale

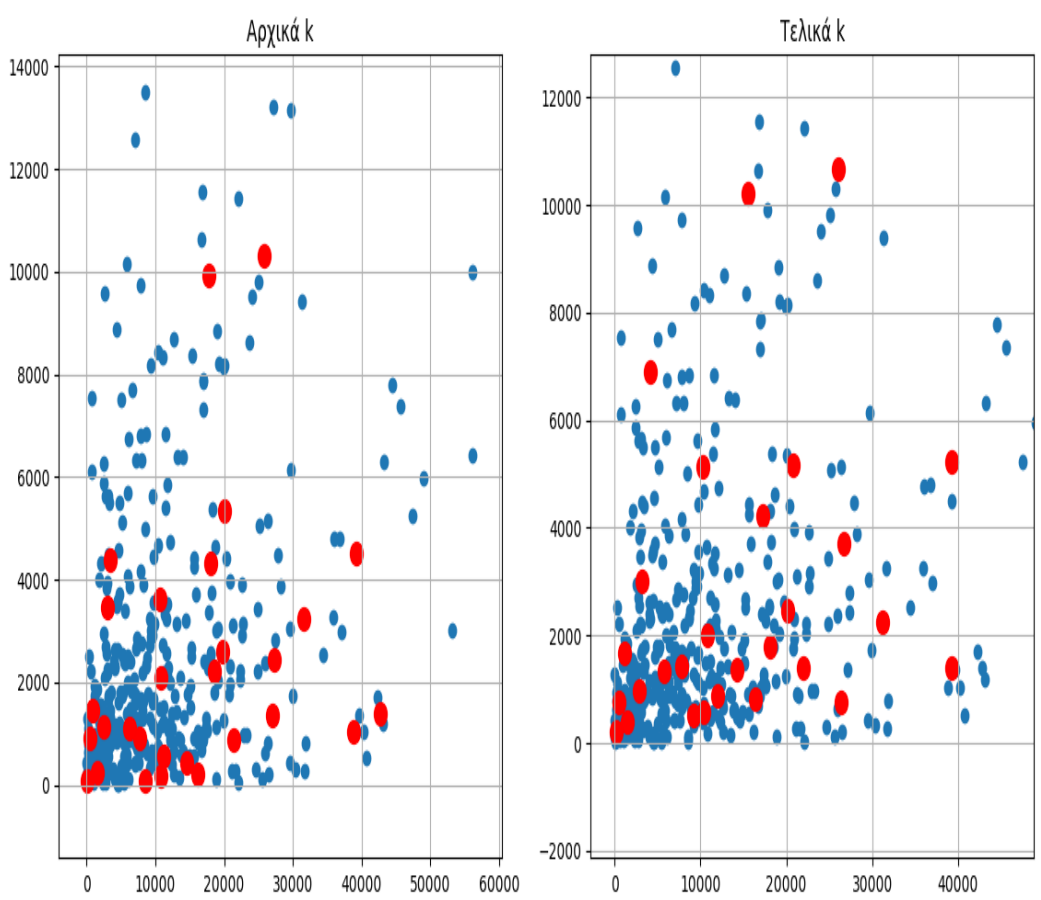


Εικόνα 12: Δεδομένα Wholesale, k=20 (Μεγέθυνση)

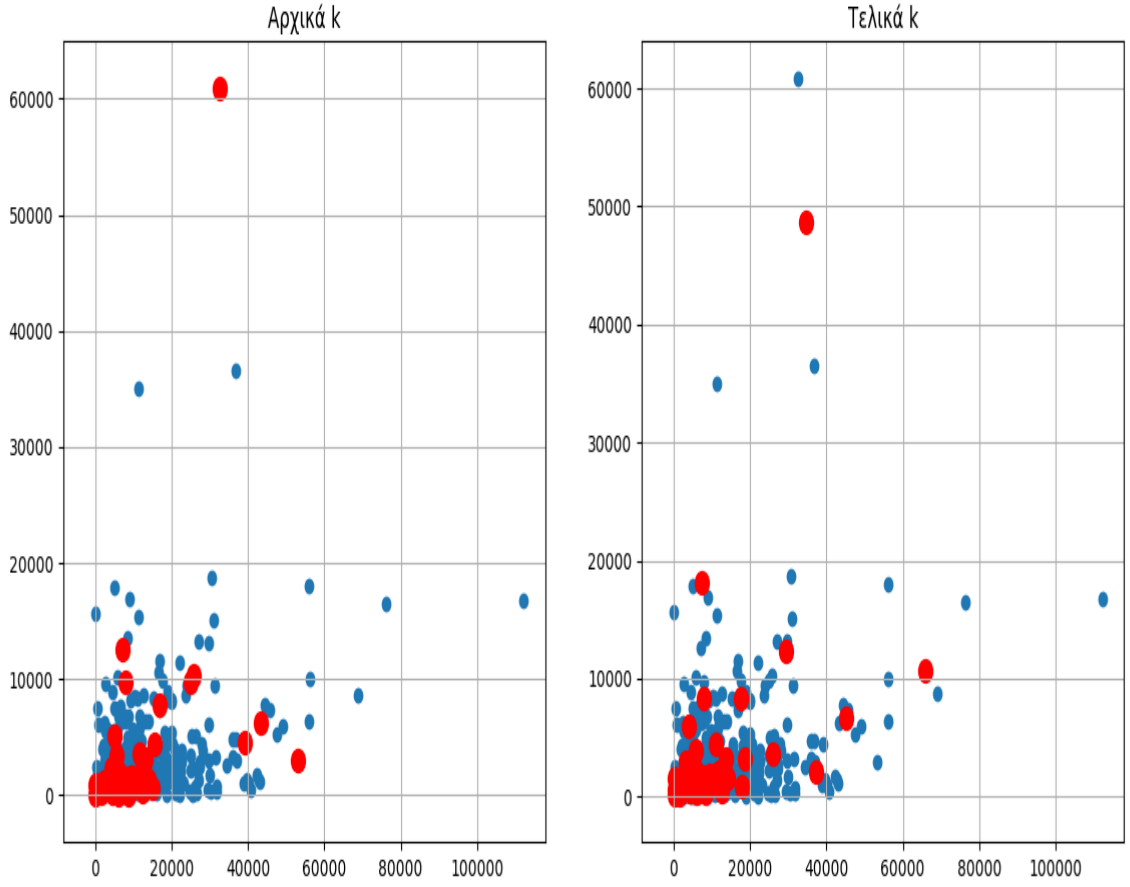


Εικόνα 13: Δεδομένα Wholesale, k=30

Αποτελέσματα- Δεδομένα Wholesale

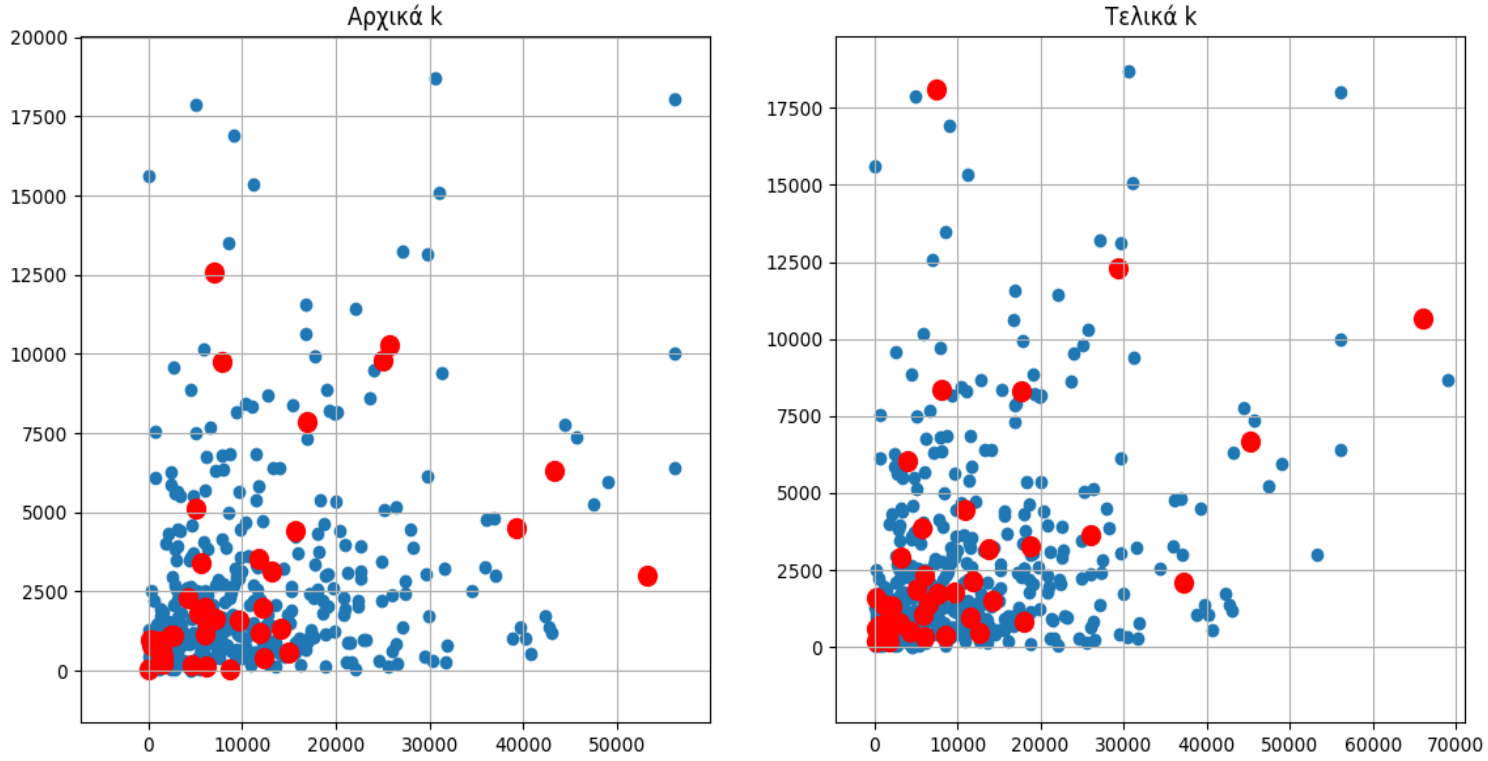


Εικόνα 14: Δεδομένα Wholesale, k=30 (Μεγέθυνση)



Εικόνα 15: Δεδομένα Wholesale, k=50

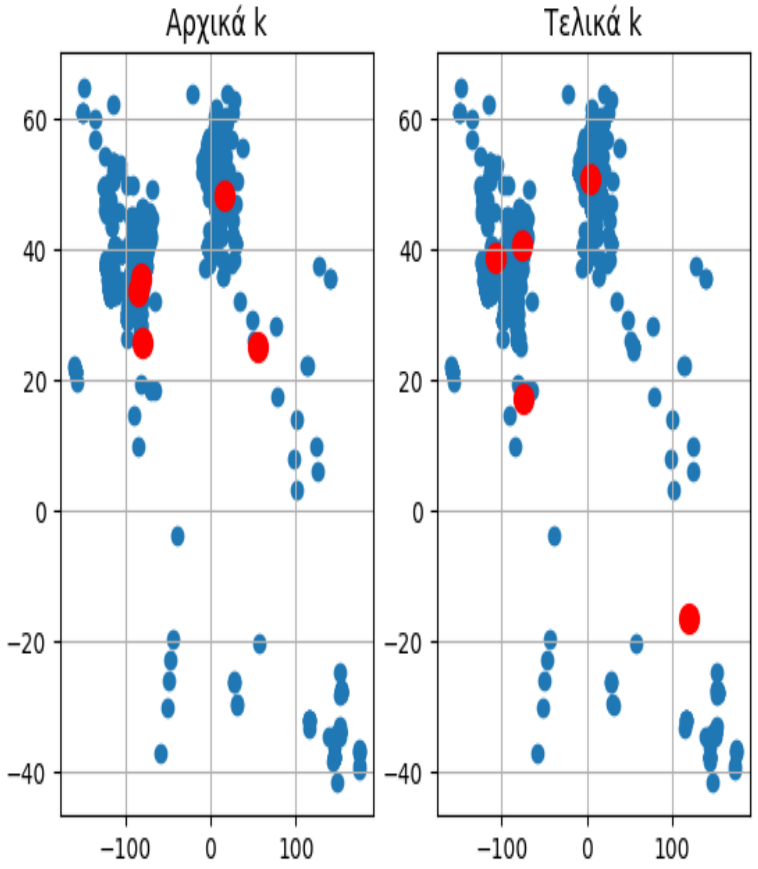
Αποτελέσματα- Δεδομένα Wholesale



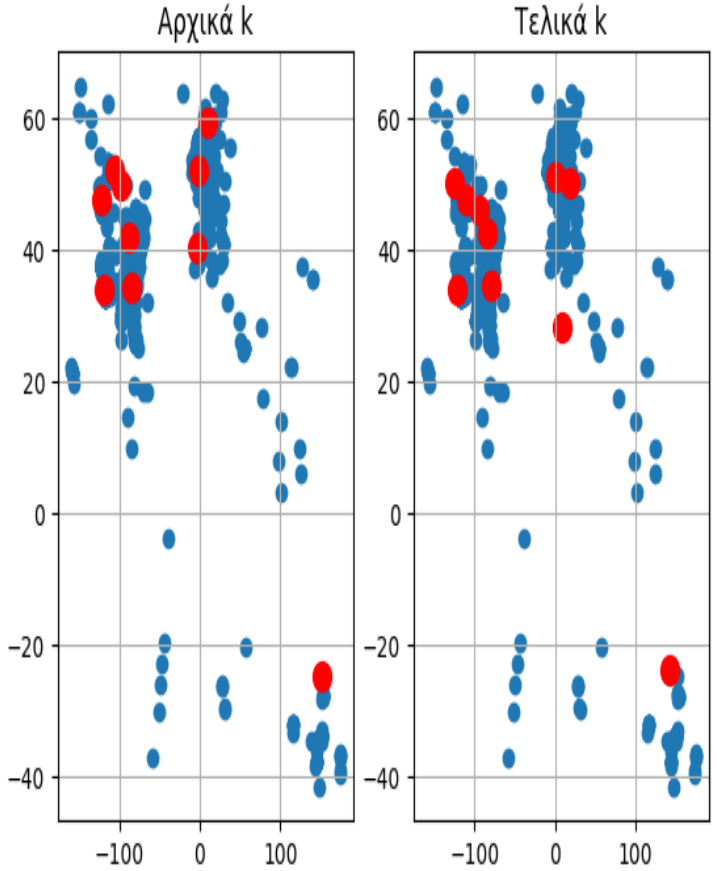
Εικόνα 16: Δεδομένα Wholesale, k=50 (Μεγέθυνση)

Αποτελέσματα- Δεδομένα Sales

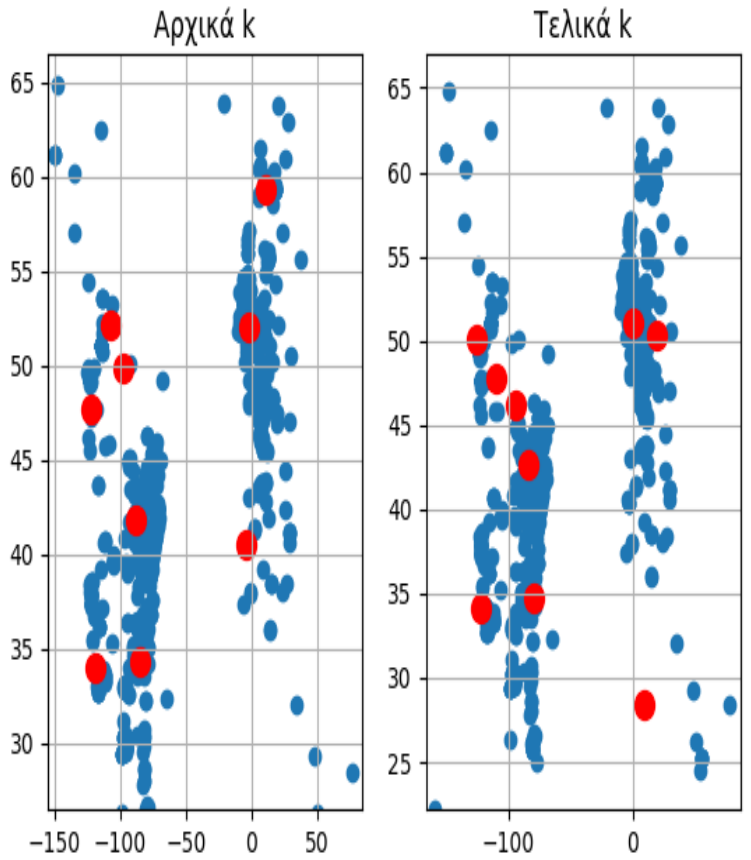
Το αρχείο περιέχει 998 καταγραφές (μέρος τους παρουσιάζεται στο παράρτημα 3) με δεδομένα πωλήσεων. Επιλέγονται μόνο οι στήλες Γεωγραφικό πλάτος (Latitude) και Γεωγραφικό μήκος (Longtitude).



Εικόνα 17: Δεδομένα Sales , k=5

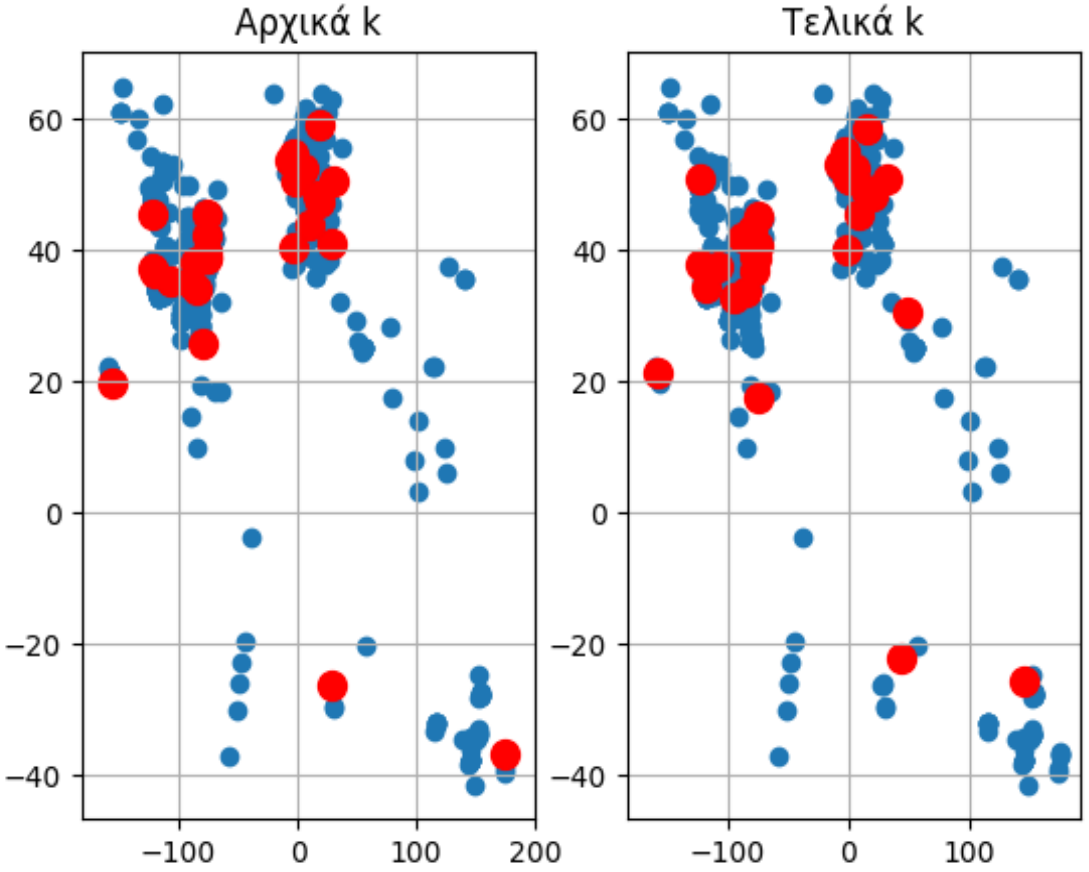


Εικόνα 18: Δεδομένα Sales , k= 10

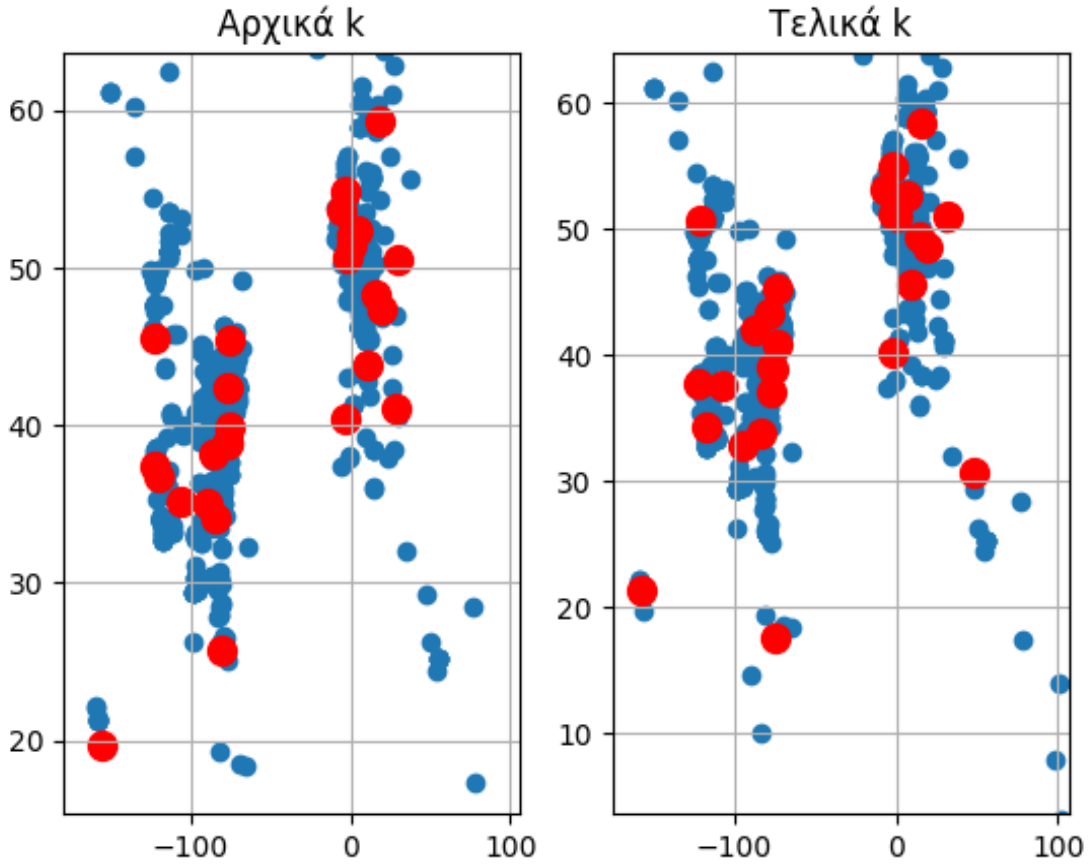


Εικόνα 19: Δεδομένα Sales , k=10 (Μεγέθυνση)

Αποτελέσματα- Δεδομένα Sales

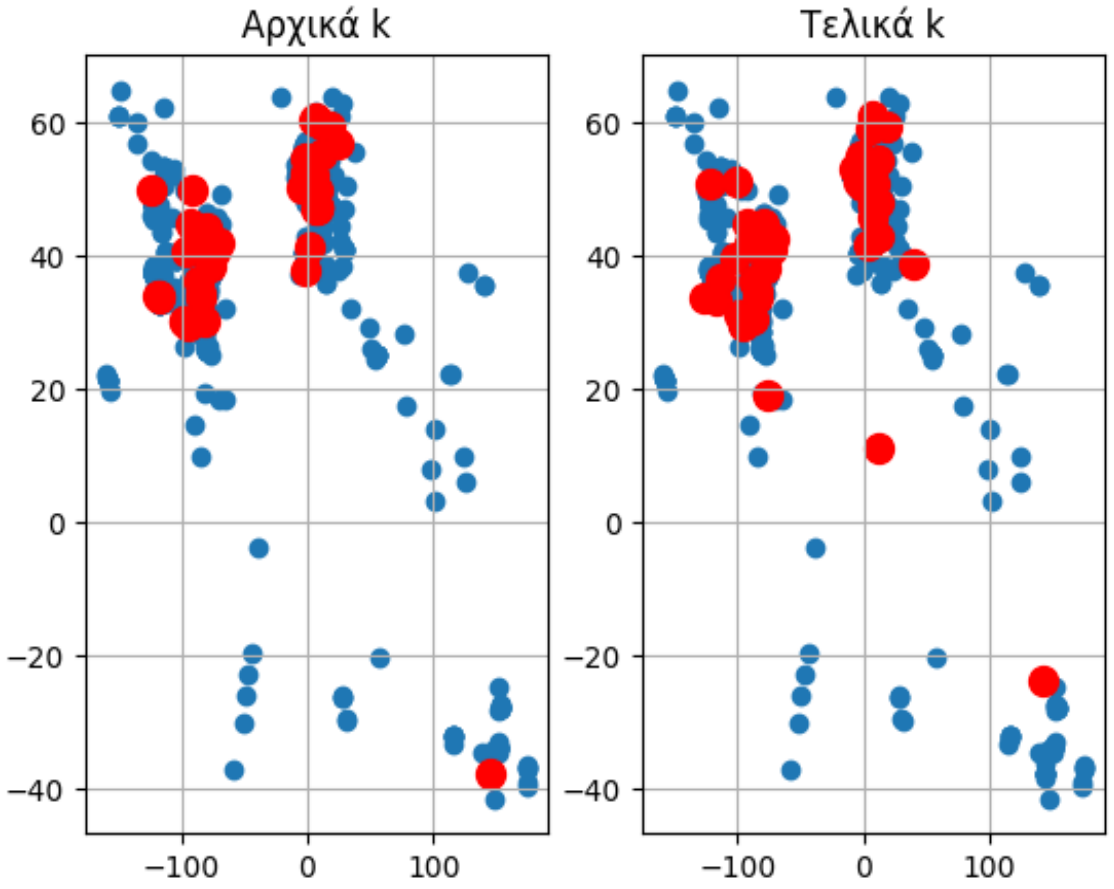


Εικόνα 20: Δεδομένα Sales , k=25

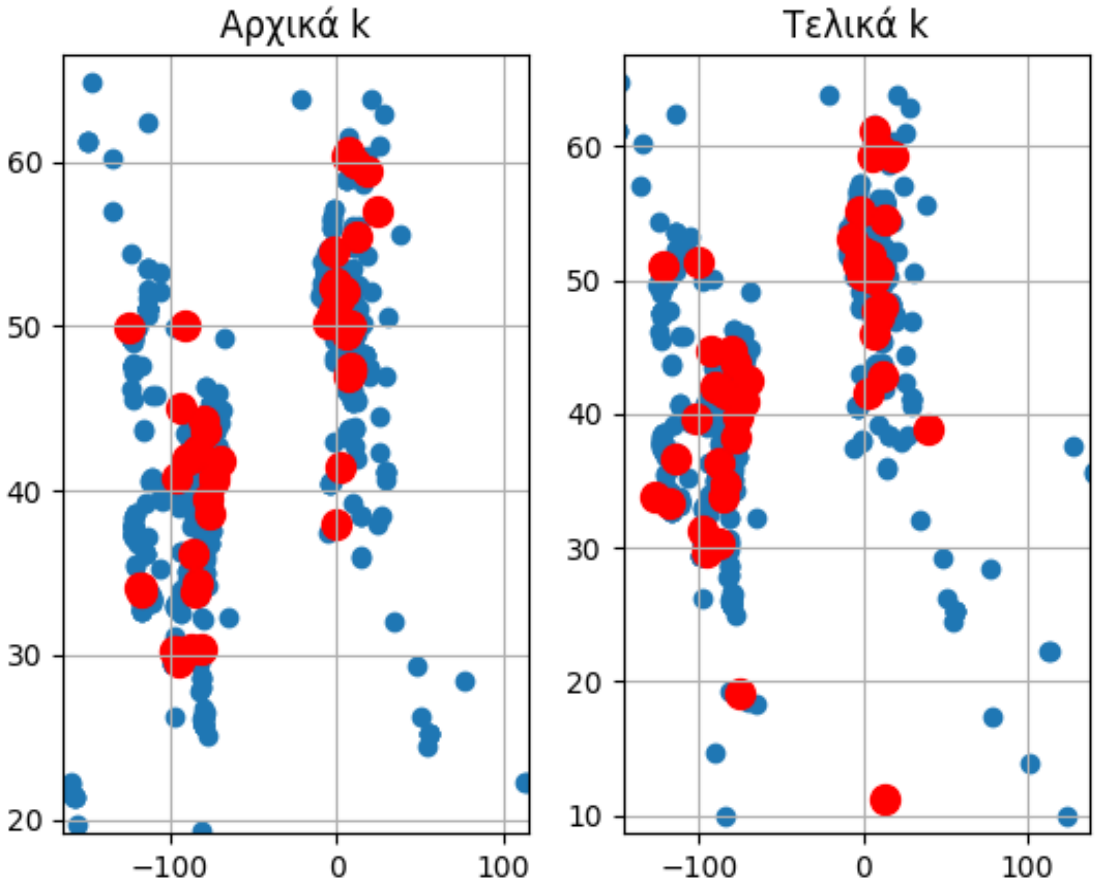


Εικόνα 21: Δεδομένα Sales , k=25 (Μεγέθυνση)

Αποτελέσματα- Δεδομένα Sales



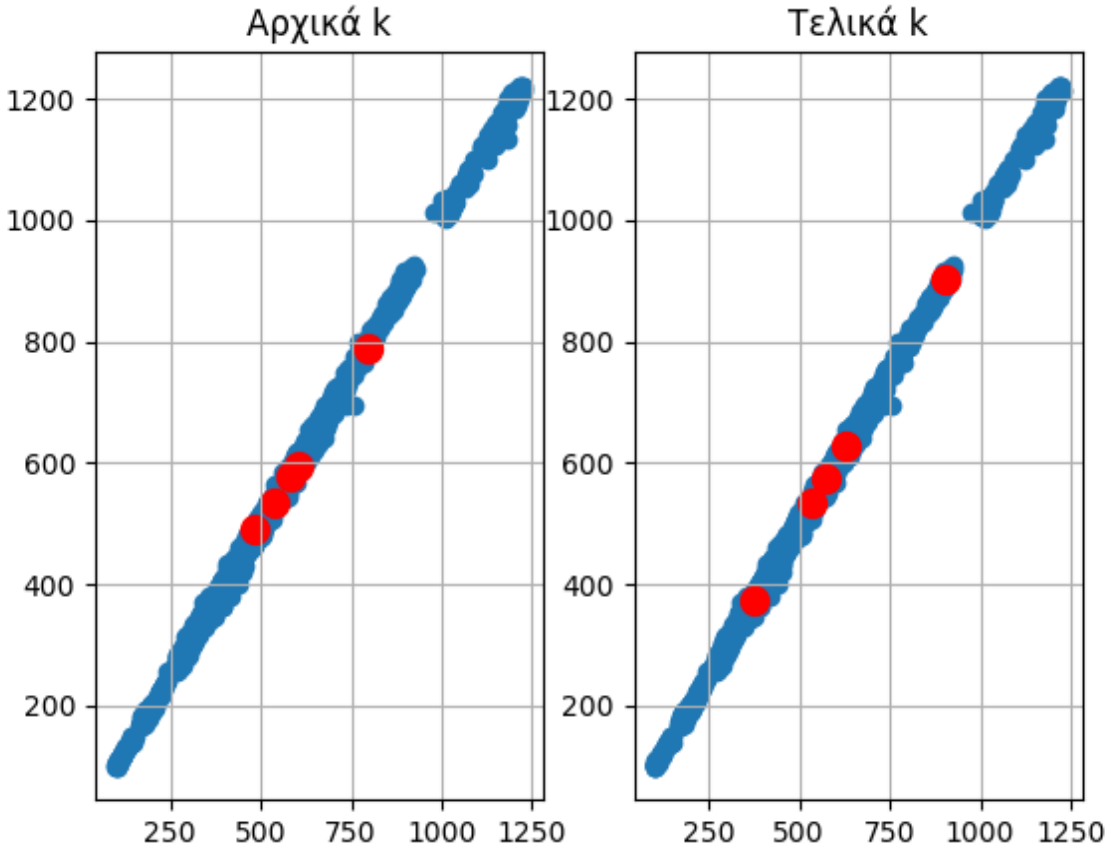
Εικόνα 22: Δεδομένα Sales , k=50



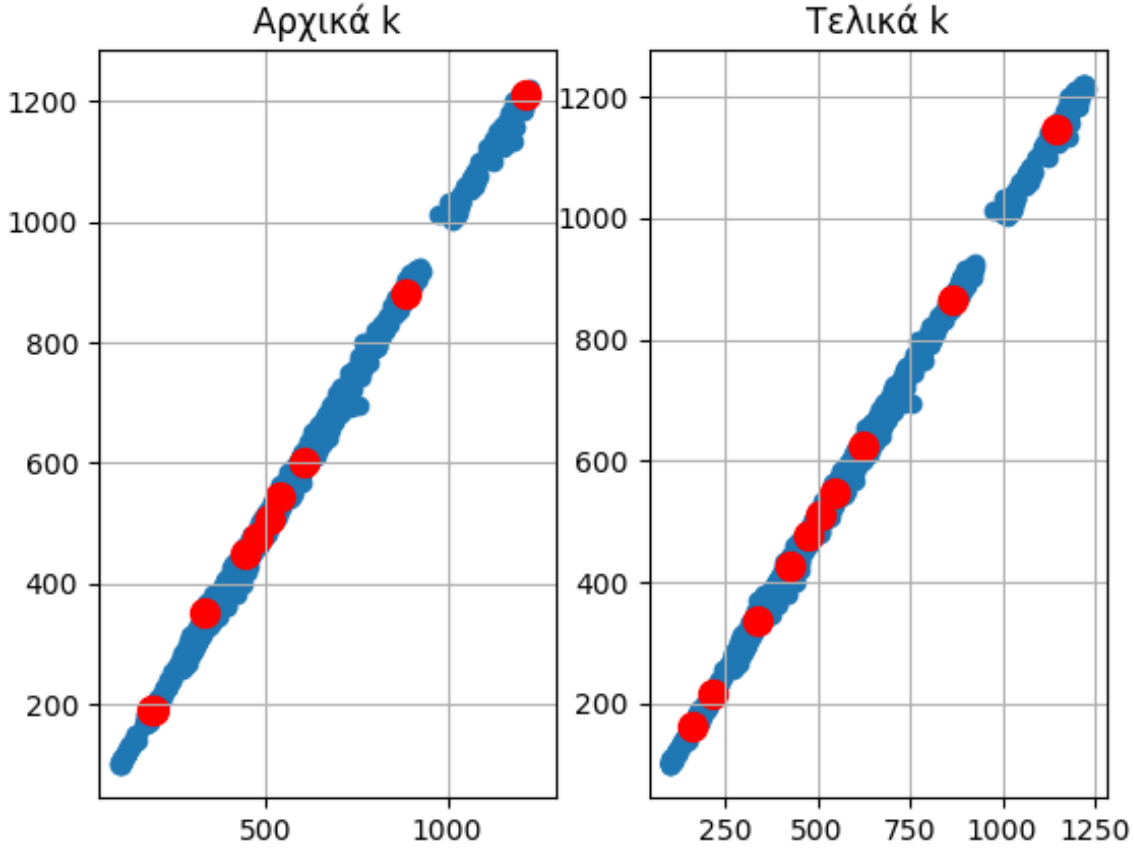
Εικόνα 23: Δεδομένα Sales , k=50 (Μεγέθυνση)

Αποτελέσματα-Δεδομένα Google

Το αρχείο περιέχει 2.518 καταγραφές (μέρος τους παρουσιάζεται στο παράρτημα 4) με δεδομένα τιμών της μετοχής της Google. Επιλέγονται μόνο η τιμή ανοίγματος (Open) και η τιμή κλεισίματος (Close).

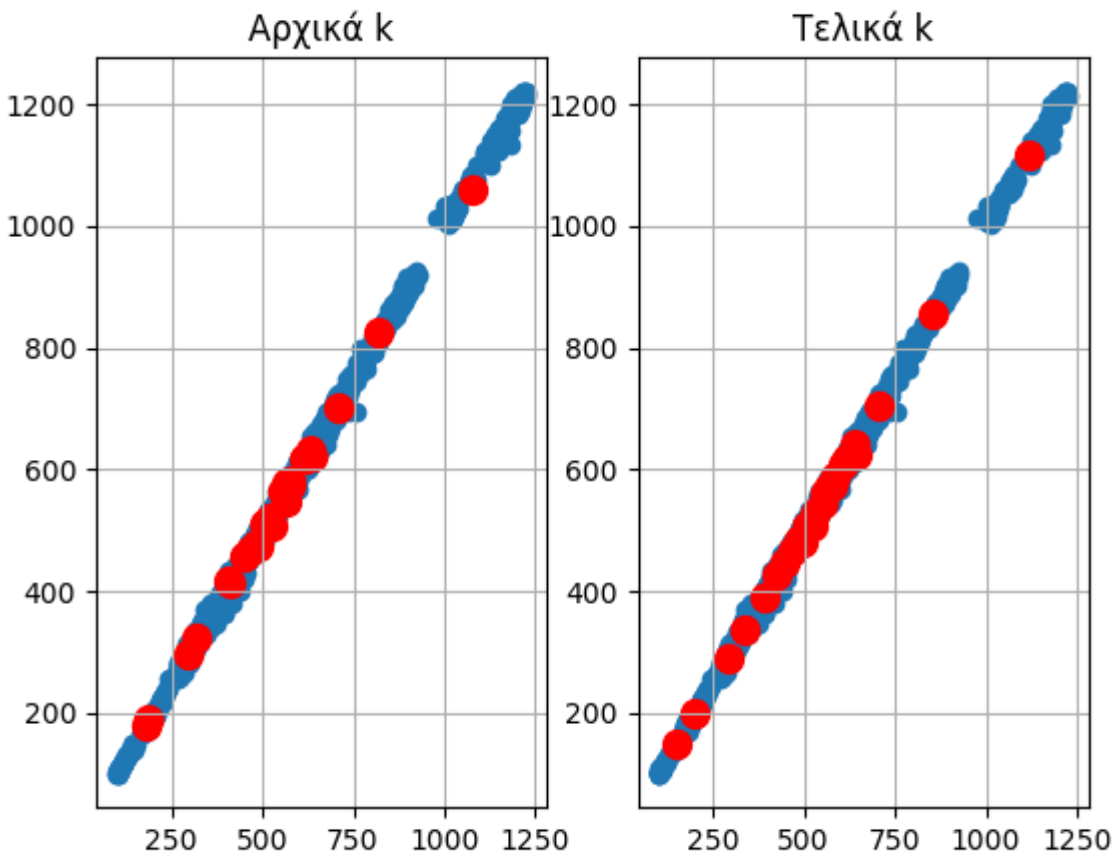


Εικόνα 24: Δεδομένα Google , k=5

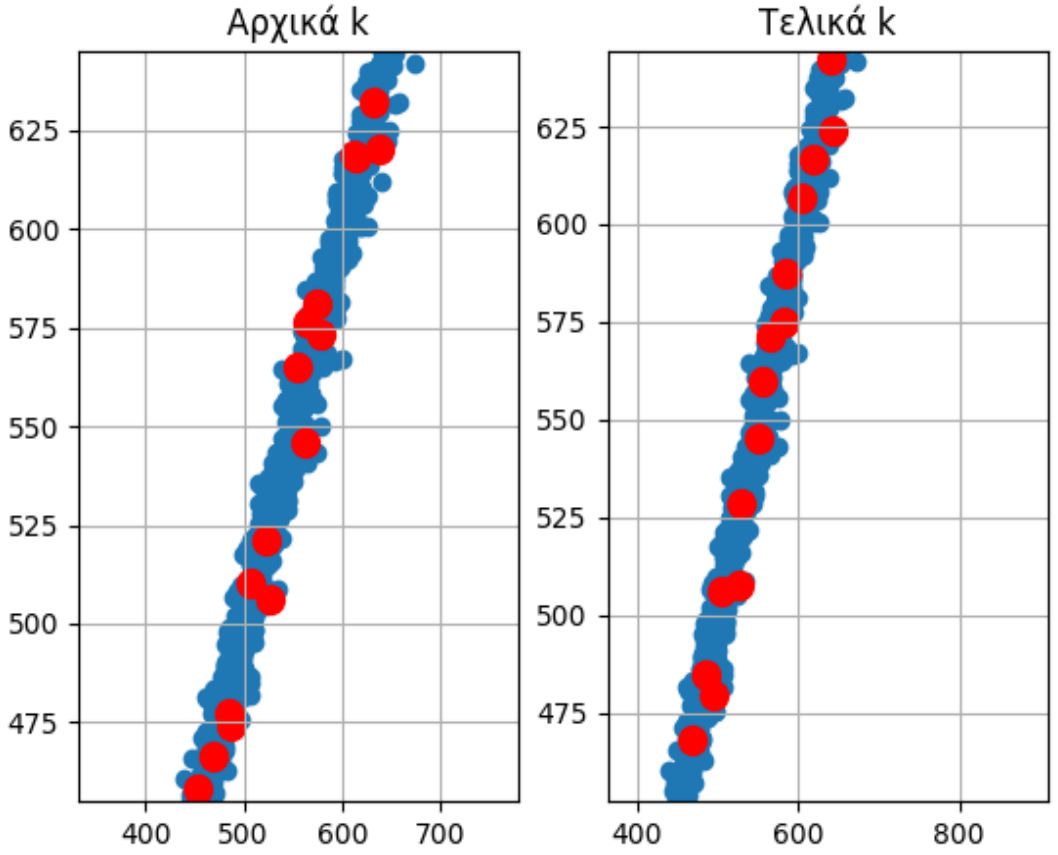


Εικόνα 25: Δεδομένα Google , k=10

Αποτελέσματα-Δεδομένα Google



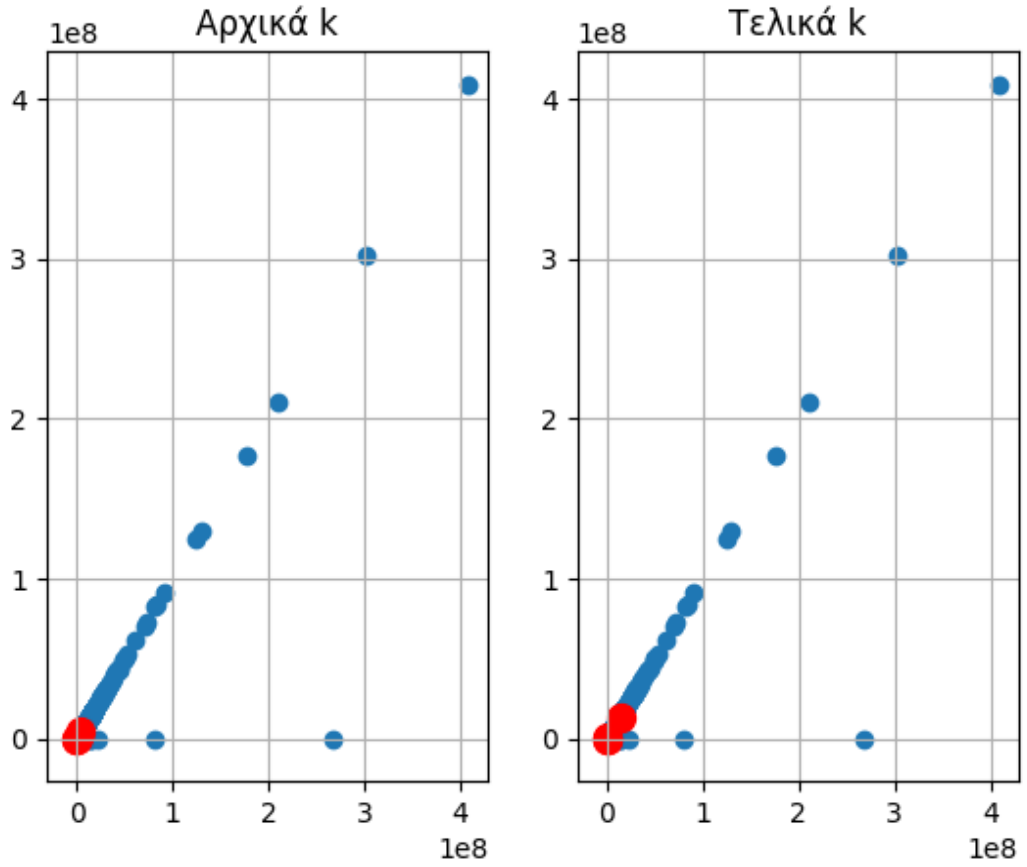
Εικόνα 26: Δεδομένα Google , k=15



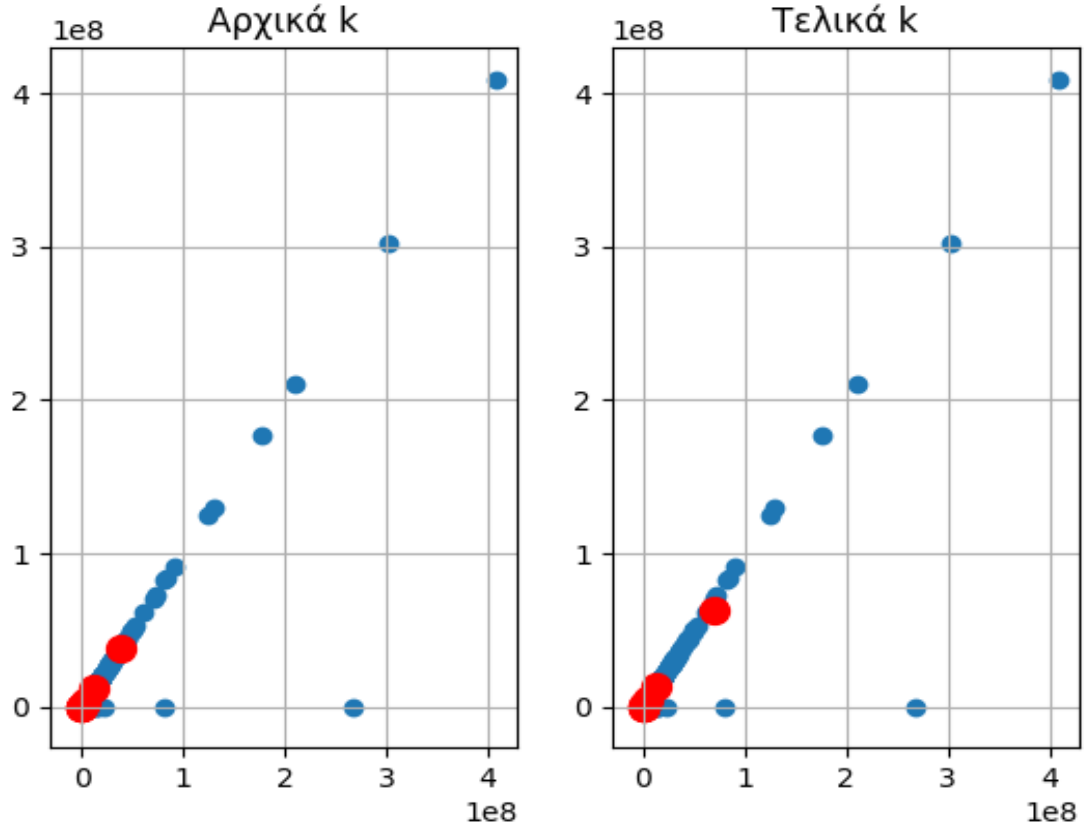
Εικόνα 27: Δεδομένα Google , k=15 (Μεγέθυνση)

Αποτελέσματα- Δεδομένα Insurance

Το αρχείο περιέχει 36.634 καταγραφές (μέρος τους παρουσιάζεται στο παράρτημα 5) με δεδομένα ασφαλιστικών συμβολαίων. Επιλέγονται μόνο οι στήλες E (hu_site_limit) και F(fl_site_limit).

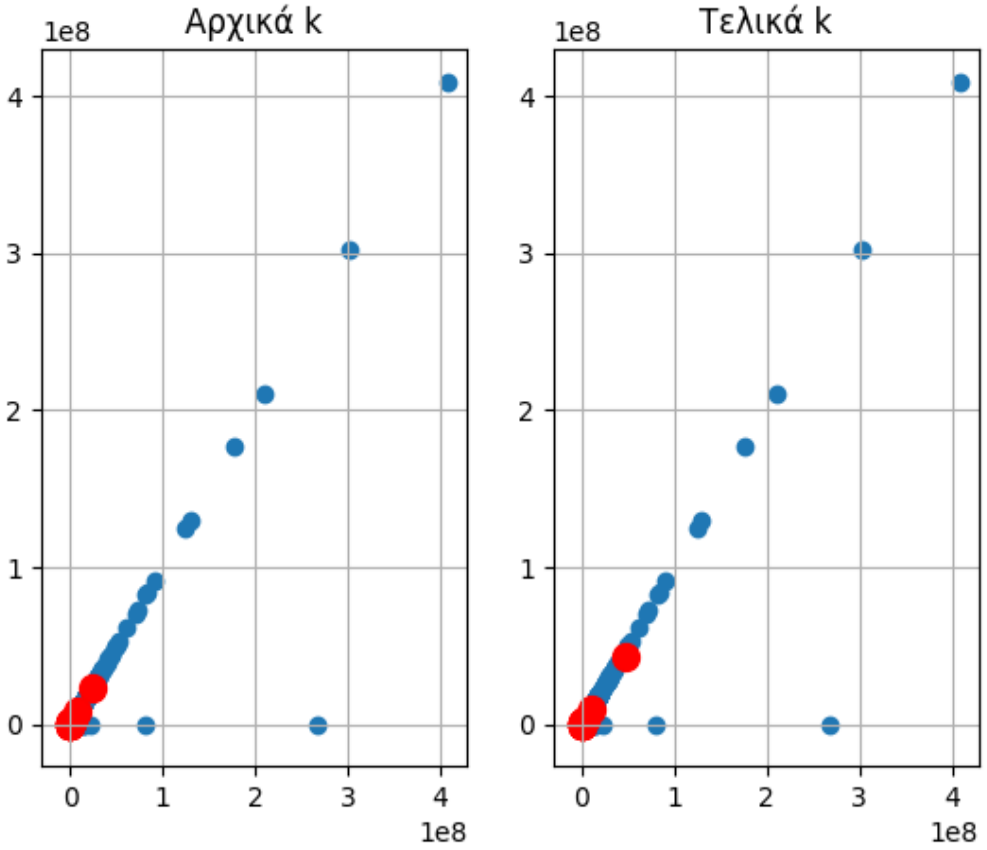


Εικόνα 28: **Δεδομένα Insurance , k=5**

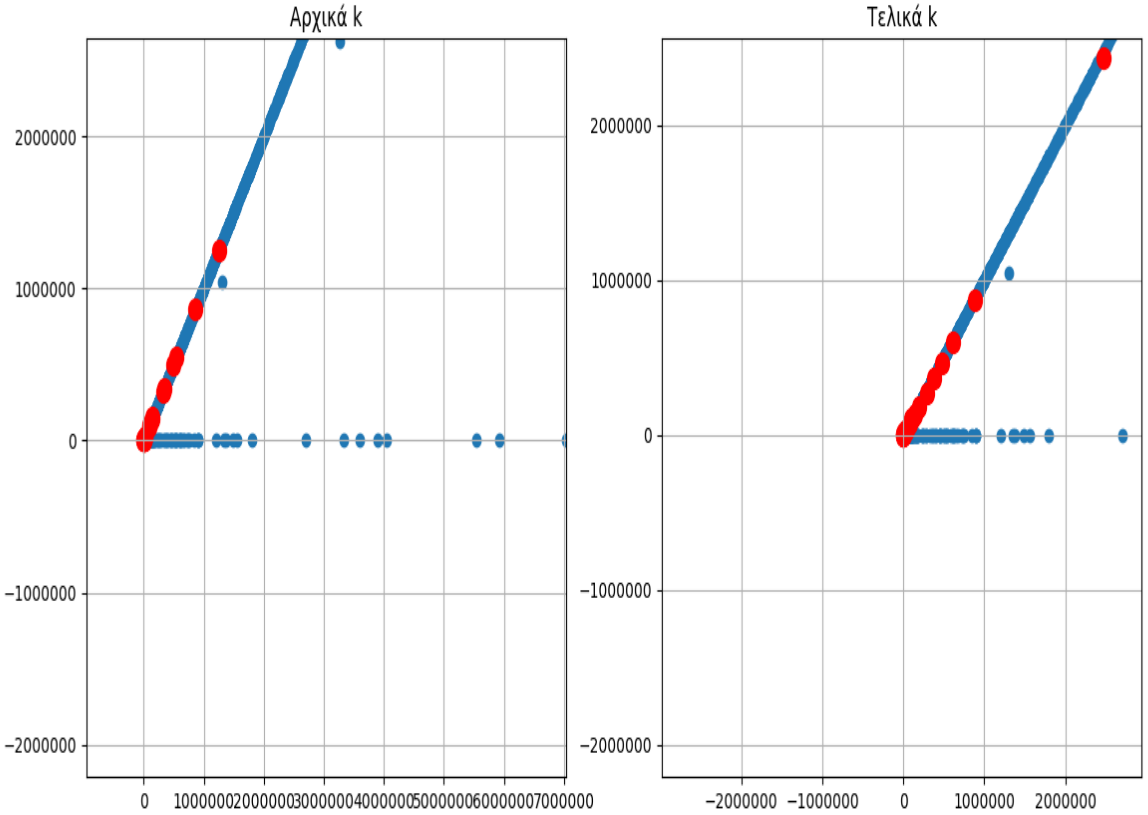


Εικόνα 29: Δεδομένα Insurance , k=10

Αποτελέσματα- Δεδομένα Insurance

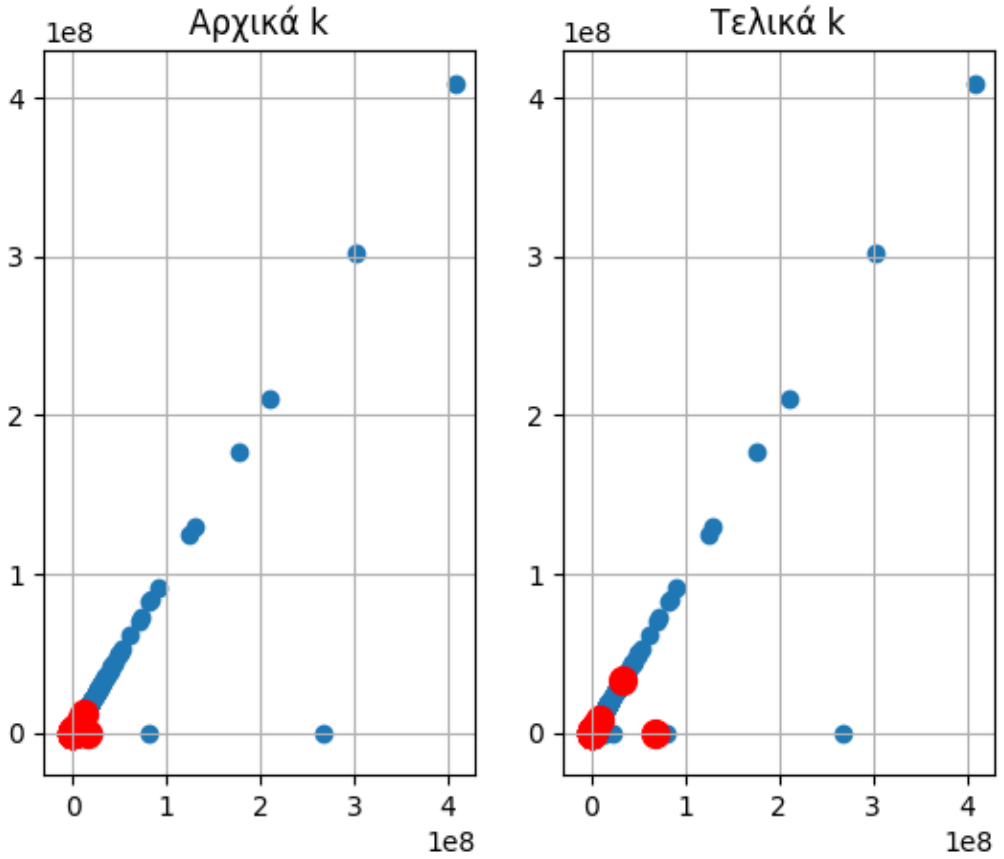


Εικόνα 30: Δεδομένα Insurance , k=25

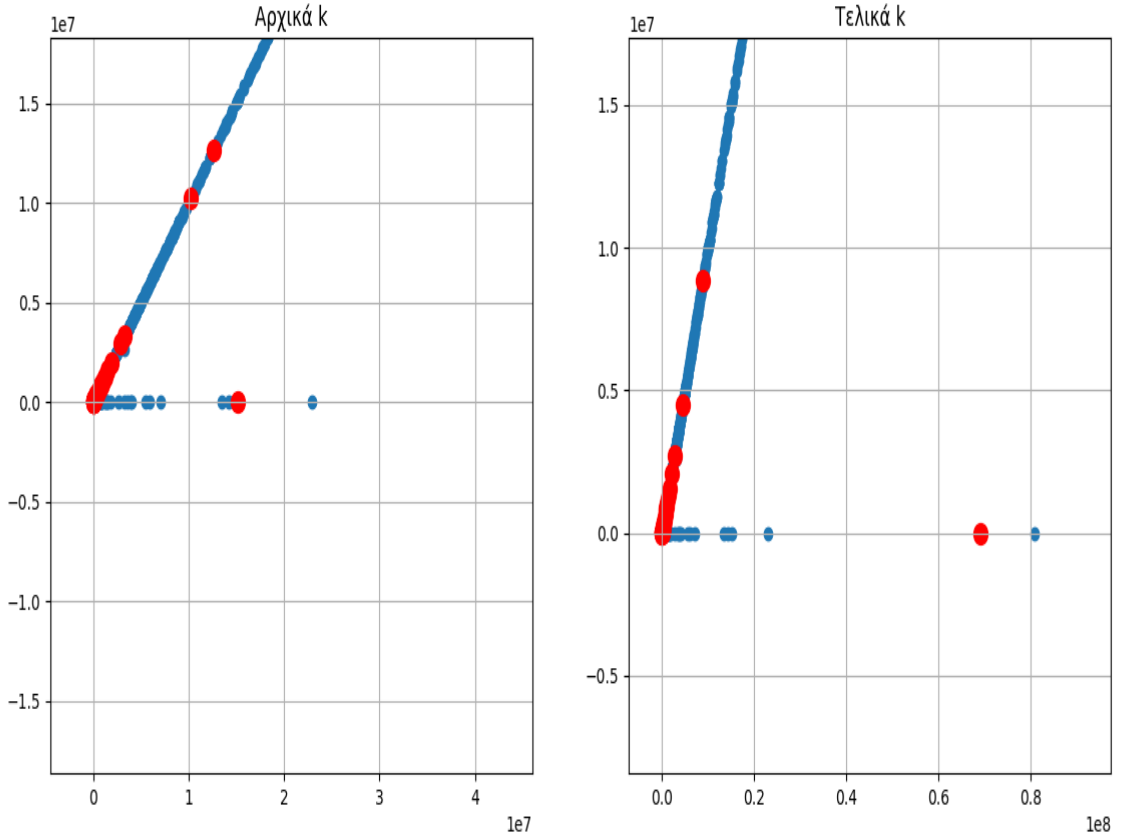


Εικόνα 31: Δεδομένα Insurance , k=25 (Μεγέθυνση)

Αποτελέσματα- Δεδομένα Insurance



Εικόνα 32: Δεδομένα Insurance , k=50



Εικόνα 33: Δεδομένα Insurance , k=50 (Μεγέθυνση)

Αποτελέσματα- Αξιολόγηση αλγόριθμου

Γενικά, ο αλγόριθμος ολοκληρώνεται σε μικρό χρονικό διάστημα (μερικά δευτερόλεπτα). Πρόβλημα παρουσιάστηκε με τα δεδομένα INSURANCE, καθώς ο όγκος τους είναι μεγάλος και το πρόγραμμα απαιτεί μεγάλο χρόνο να τα διαβάσει.

Τα δεδομένα WHOLESALES δεν παρουσιάζουν γραμμικότητα. Η μεγαλύτερη μάζα των δεδομένων συγκεντρώνεται στην περιοχή μηδέν, ενώ υπάρχουν και πολλά απομονωμένα στοιχεία. Ο αλγόριθμος διασπείρει τα κεντροειδή, συμπεριλαμβάνοντας και τις απομονωμένες περιοχές για k= 30 και πάνω.

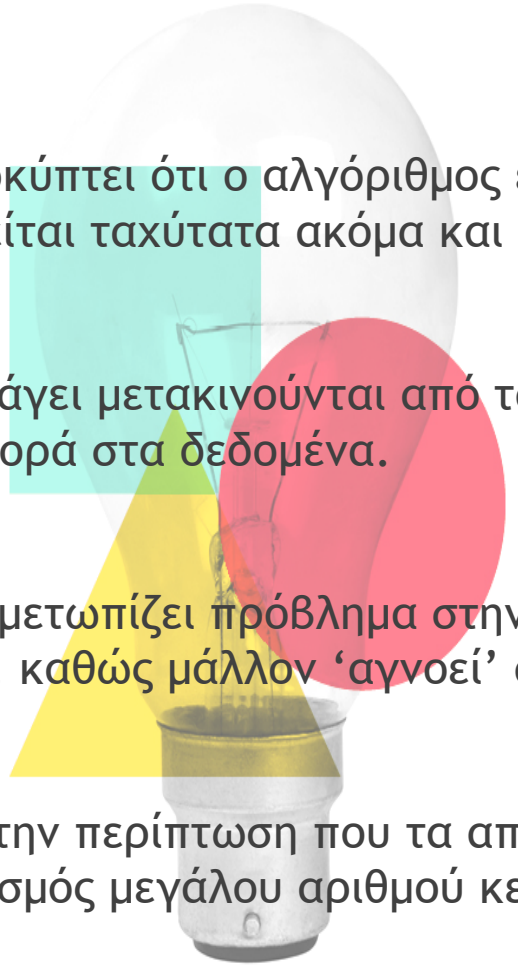
Τα δεδομένα SALES παρουσιάζουν διασπορά με δύο μεγάλες μάζες και αρκετά απομονωμένα στοιχεία. Οι απομονωμένες περιοχές συμπεριλαμβάνονται στον προσδιορισμός των κεντροειδών για k= 25 και πάνω.

Αποτελέσματα- Αξιολόγηση αλγόριθμου

Τα δεδομένα GOOGLE παρουσιάζουν γραμμική συσχέτιση με μία ασυνέχεια, ώστε να δημιουργούνται δύο διακριτά τμήματα. Για k= 5 ο αλγόριθμος δεν δημιουργεί κεντροειδές στο ένα τμήμα, ούτε υπολογίζει νέο κεντροειδές σε εκείνη την περιοχή. Για μεγαλύτερες τιμές k, ο αλγόριθμος συμπεριλαμβάνει και τη δεύτερη περιοχή. Ακόμα και για k= 15, όμως, ορίζει μόνο ένα κεντροειδές, το οποίο φαίνεται να προσαρμόζεται στο κέντρο των δεδομένων της περιοχής. Ο αλγόριθμος 'απλώνει' τα κεντροειδή μέχρι και στις μικρότερες τιμές των δεδομένων.

Τα δεδομένα INSURANCE παρουσιάζουν γραμμικότητα, αλλά και ασυνέχεια, με ένα κύριο σώμα και μερικά απομονωμένα στοιχεία τόσο στο διαγώνιο άξονα όσο και στον οριζόντιο. Το κύριο μέρος των δεδομένων συγκεντρώνονται στην περιοχή κοντά στο μηδέν. Ο αλγόριθμος δεν υπολογίζει κεντροειδή για τα απομονωμένα στοιχεία στον οριζόντιο άξονα ούτε για k = 25.

Συμπεράσματα



- Από την υλοποίηση προκύπτει ότι ο αλγόριθμος είναι εύκολος στην υλοποίηση και εύχρηστος. Υλοποιείται ταχύτατα ακόμα και σε μεγάλους όγκους δεδομένων.
- Τα κεντροειδή που παράγει μετακινούνται από τα αρχικώς ορισμένα και εμφανίζουν καλή διασπορά στα δεδομένα.
- Όμως, φαίνεται να αντιμετωπίζει πρόβλημα στην περίπτωση που τα δεδομένα εμφανίζουν ασυνέχειες, καθώς μάλλον ‘αγνοεί’ αυτά που είναι απομονωμένα.
- Λύση στο φαινόμενο, στην περίπτωση που τα απομονωμένα δεδομένα έχουν ενδιαφέρον, δίνει ο ορισμός μεγάλου αριθμού κεντροειδών.

Μελλοντικές Επεκτάσεις

- Στην παρούσα εργασία εξετάστηκαν δεδομένα δύο διαστάσεων. Θα ήταν χρήσιμο να επεκταθεί η εφαρμογή της μεθόδους και στην περίπτωση περισσότερων διαστάσεων, αλλά και να συγκριθεί με άλλου αλγόριθμους ομαδοποίησης.

- Επιπλέον, ο αλγόριθμος ίσως παρουσιάσει βελτίωση με χρήση άλλων αποστάσεων, πέραν της Ευκλείδειας.

- Προτείνεται, η εκτίμηση του αριθμού των κεντροειδών, μέσω μιας από τις μεθόδους που προτείνονται στη βιβλιογραφία και όχι ο τυχαίος ορισμός τους, κάτι που αναμένεται να βελτιώσει τα αποτελέσματα.

- Ενδιάφερον, θα είχε η παραλληλοποίηση του αλγόριθμου και η χρήση του στην ανάλυση δεδομένων μεγάλου όγκου.

ΤΕΛΟΣ

Σας ευχαριστώ για την προσοχή σας