Νίκος Χριστοδούλου – ΑΜ: p3190223

1η Εργασία

Μέθοδοι Στατιστικής και Μηχανικής Μάθησης

Χειμερινό Εξάμηνο 2022-2023

Προεπεξεργασία του dataset:

Αρχικά, παρατηρούμε ότι το dataset περιέχει NA στοιχεία μέσα. Συνεπώς, αντικαθιστούμε με τη βοηθεία του excel κάθε NA στοιχείο με 0 και μετατρέπουμε όλες τις μεταβλητές που ήταν character σε numeric, για να εφαρμόσουμε τις τεχνικές συσταδοποίησης που γνωρίζουμε.

i)

Για να απαντήσουμε στο εν λόγω ερώτημα, πρέπει να διερευνήσουμε τα δεδομένα με τα οποία δουλεύουμε. Κάνοντας μια σύντομη οπτικοποίηση των δεδομένων μας σε ενα dataframe, παρατηρούμε ότι δεν υπάρχει κάποια μεταβλητή – ετικέτα (label) που να μας προδιαθέτει για την κατάταξη των πόλεων σε κάποια πιθανή ομάδα. Συνεπώς, έχουμε να αντιμετώπισουμε ένα πρόβλημα **μη** επιβλεπόμενης μάθησης. Άρα, οι μέθοδοι μηχανικής μάθησης που θα προσπαθήσουμε να εφαρμόσουμε είναι ο αλγόριθμος των k Μέσων (**K-Means**), καθώς και ο **Hierarchical Clustering**, με τη βοήθεια των κατάλληλων τεχνικών για την εύρεση του εν δυνάμει βέλτιστου αριθμού k ομάδων.

Όσον αφορά, τώρα, στο ποια από τις δύο τεχνικές θα είναι καταλληλότερη, πρέπει να διερευνήσουμε τα δεδομένα ως προς τον όγκο τους, την κατανομή τους, τη μεταξύ τους συσχέτιση και τις τιμές τις οποίες λαμβάνουν. Ο αλγόριθμος K-Means γνωρίζουμε ότι είναι προτιμότερος σε ένα μεγάλο dataset λόγω του μικρού υπολογιστικού κόστους που απαιτεί. Βέβαια, ο όγκος των δεδομένων που έχουμε στην περίπτωση μας είναι μικρός, οπότε το υπολογιστικό κόστος είναι αμελητέο. Επίσης, πρέπει διαισθητικά να έχουμε κάποια ιδέα για τον αριθμό των ομάδων k. Με βάση το ζητούμενο της εκφώνησης, μια πρώτη σκέψη θα ήταν να διακρίνουμε τις πόλεις ως προς την ποιότητα ζωής που προσφέρουν στους πολίτες της σε:

- Καλή Ποιότητα
- Μέτρια Ποιότητα
- Κακή ποιότητα

Γενικά, θα περιμέναμε το k να είναι ανάμεσα σε 2-4, αλλά και πάλι δε μπορούμε να γνωρίζουμε ακριβώς το καλύτερο k. Επιπρόσθετα, θα πρέπει οι ομάδες που θα σχηματιστούν να είναι σφαιροειδείς ώστε να είναι ακριβείς τα αποτελέσματα της συσταδοποίησης. Ωστόσο, τρέχοντας τον αλγόριθμο για k = 2,3,4 βλέπουμε ότι τα περιεχόμενα της κάθε ομάδας δεν είναι διακριτώς διαχωρίσιμα και οι συστάδες που προκύπτουν δεν έχουν σφαιροειδές σχήμα. Ακόμα, βλέπουμε ότι κάθε ομάδα έχει εξαιρετικά ανόμοιο πλήθος παρατηρήσεων σε σχέση με τις άλλες, με ορισμένες ομάδες

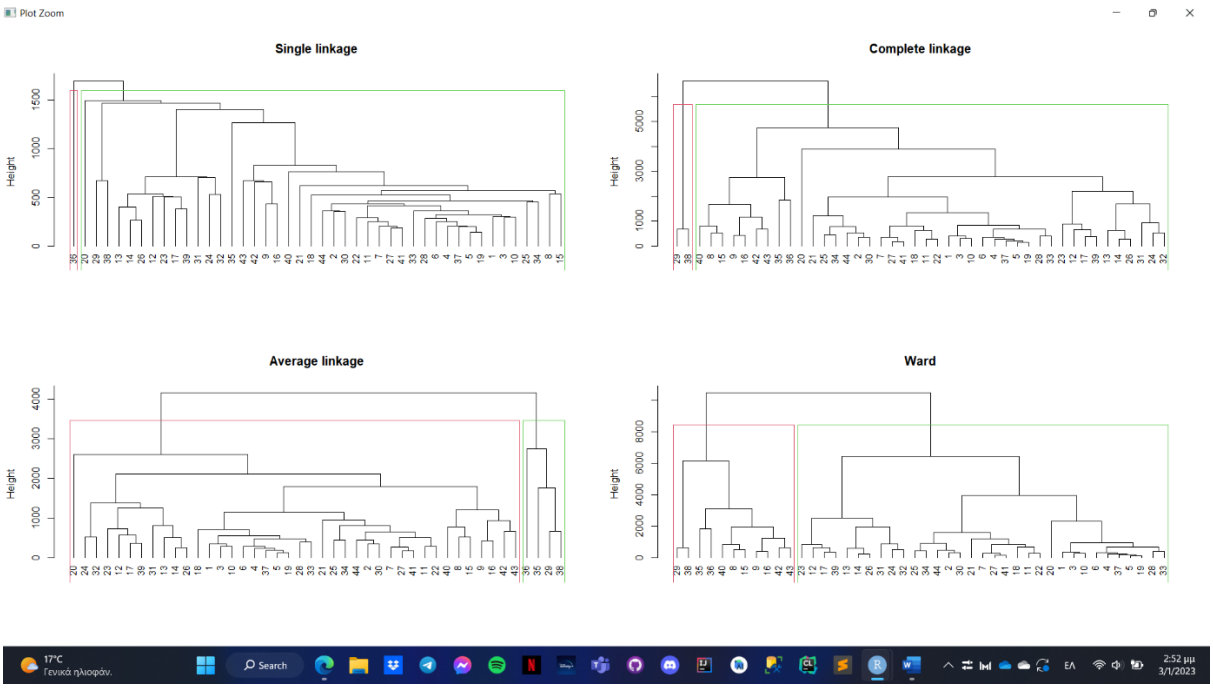
να περιέχουν ελάχιστες παρατηρήσεις. Για την οπτικοποίηση των παραπάνω, παραπεμφθείτε στο ερώτημα **iii)**. Συνεπώς, ο K-Means **δεν** είναι κατάλληλος για την περίπτωσή μας.

Περνώντας στον αλγόριθμο Hierarchical Clustering, θα χρησιμοποιήσουμε τις μεθόδους

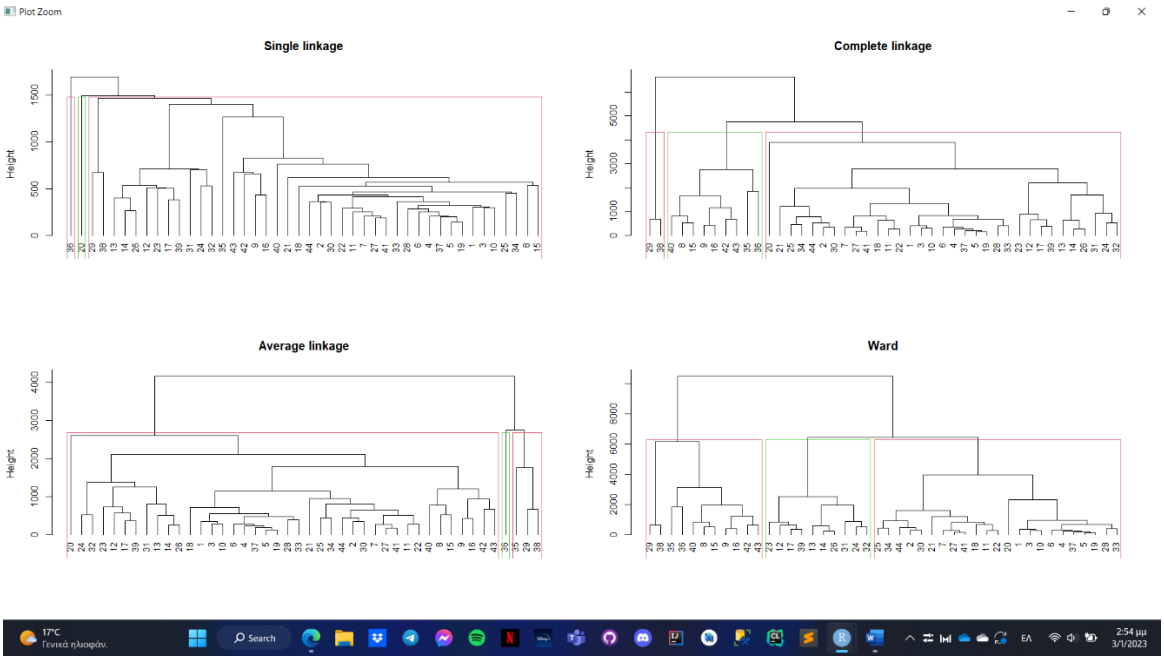
Single – Complete – Average Linkage, τη μέθοδο του **Ward**, καθώς και το **Silhouette coefficient** για να βρούμε τον κατάλληλο αριθμό των k συστάδων. Όπως γνωρίζουμε, θα σχηματίσουμε σε κάθε περίπτωση τα δεντροδιαγράμματα και θα χρησιμοποιήσουμε την συνάρτηση **cuttree** για να βρούμε τις τελικές συστάδες.

Παρατηρώντας τα δεντροδιαγράμματα από κάθε μέθοδο, βλέπουμε ότι οι τρεις πρώτες μέθοδοι (Singe – Complete – Average Linkage) συσταδοποιούν τις πόλεις σε ομάδες με αρκετά ανόμοιο αριθμό πόλεων. Αντίθετα, η μέθοδος του Ward, φαίνεται να κάνει μια πιο ομοιόμορφη ομαδοποίηση ως προς το πλήθος των πόλεων, για κάθε πιθανή τιμή του k = 2, 3, 4. Επίσης, παρατηρούμε ότι το Single και Average Linkage τείνει να δημιουργεί ομάδες με μικρότερο πλήθος πόλεων σε σχέση με τις άλλες μεθόδους για k=2,3 , με τις υπόλοιπες 2 μεθόδους να φέρνουν παρόμοια συσταδοποίηση των πόλεων για αντίστοιχες τιμές k, αλλά με την μέθοδο του Ward να έχει πιο διακριτή συσταδοποίηση και να είναι πιο consistent για κάθε τιμή k. Παρατίθενται τα δεντροδιαγράμματα κλαδεμένα κάθε φορά για k = 2, 3, 4.

Για k=2:



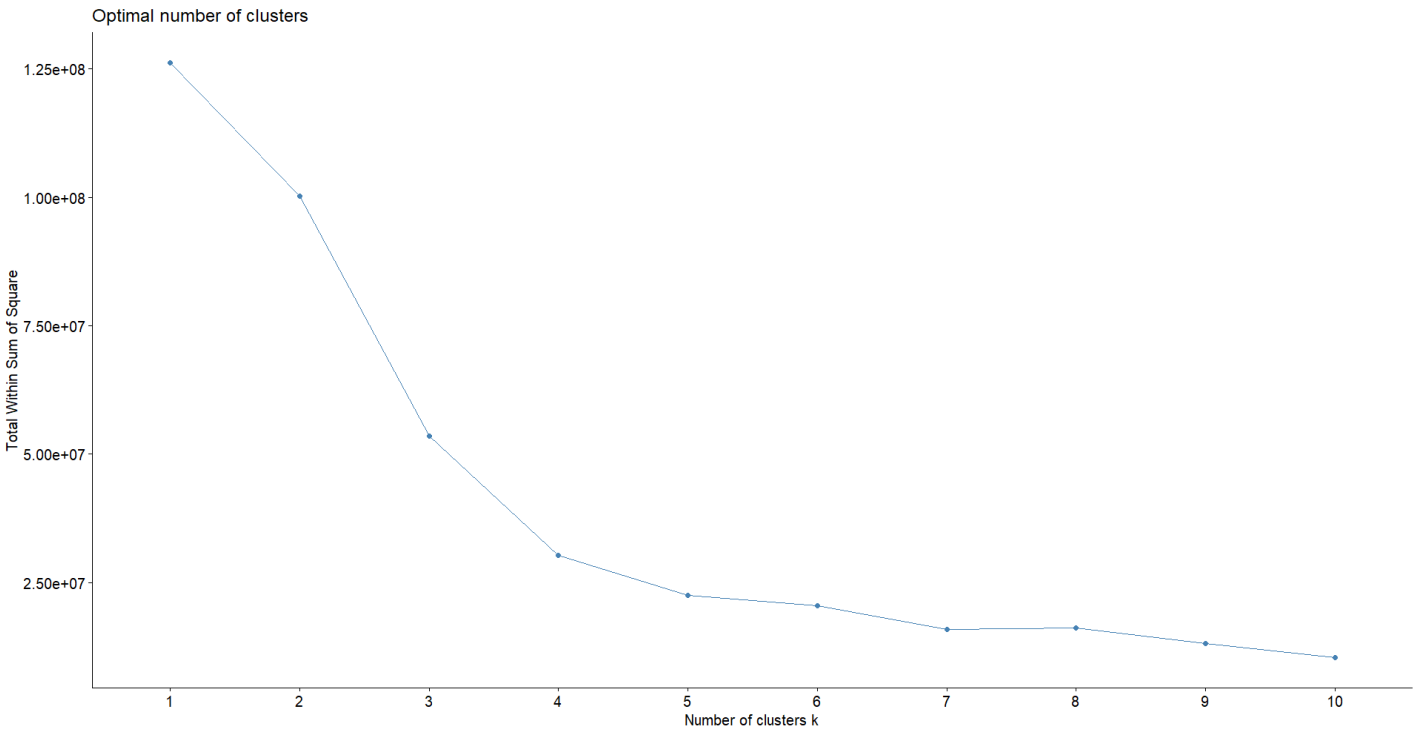
Για k=3:


Plot Zoom
Single linkage
Complete linkage
Average linkage
Ward
Height

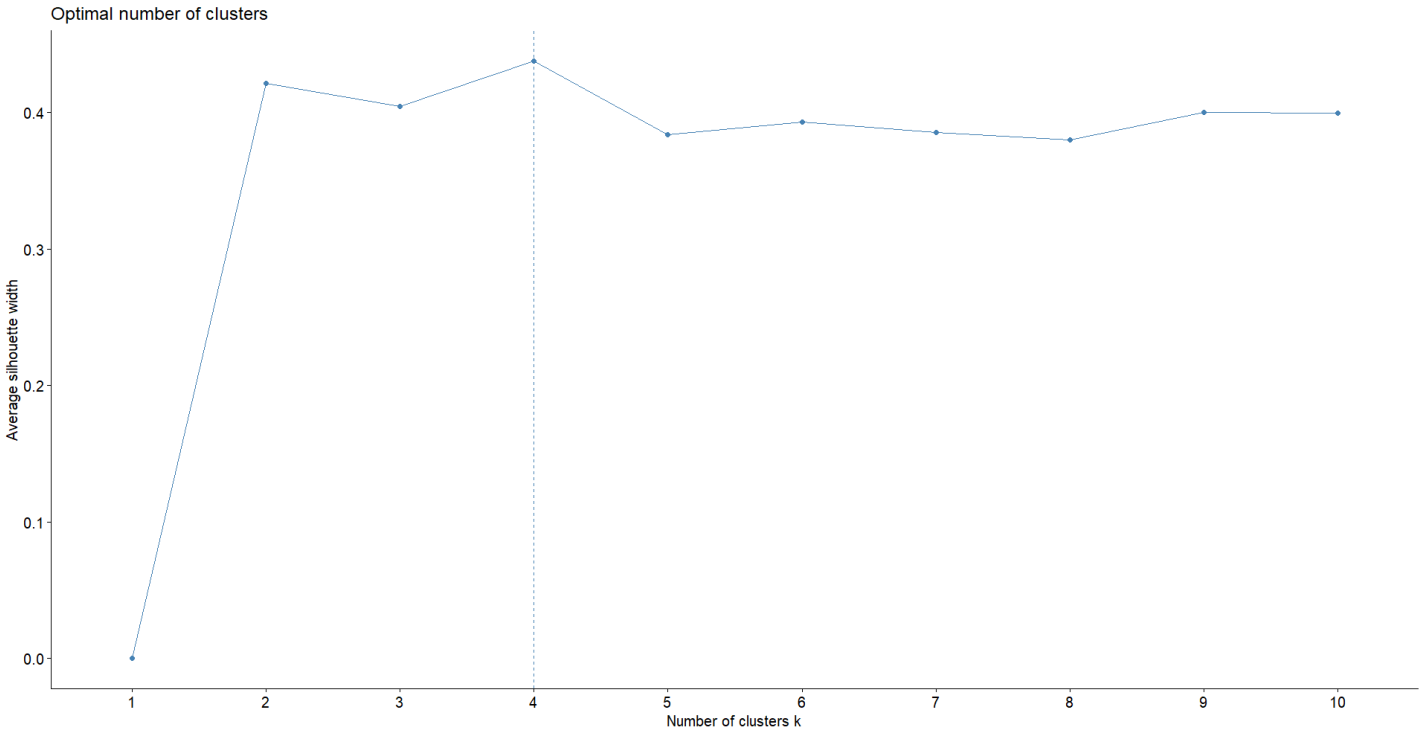
Για k=4:

Εν τέλει, αν και έχουμε μία καλή οπτικοποίηση των συστάδων με τη μέθοδο του Ward για k = 3 να φέρει πιο διακριτές και καλές συστάδες σε σχέση με τις άλλες μεθόδους, χρησιμοποιούμε τη συνάρτηση fviz_nbclust() τόσο για τον K-Means με τη χρήση της μεθόδου Elbow, όσο και για τον Hierarchical Clustering με τη βοήθεια του silhouette coefficient. Τα αποτελέσματα, όπως θα δούμε παρακάτω, μας παραπέμπουν στο εξής:

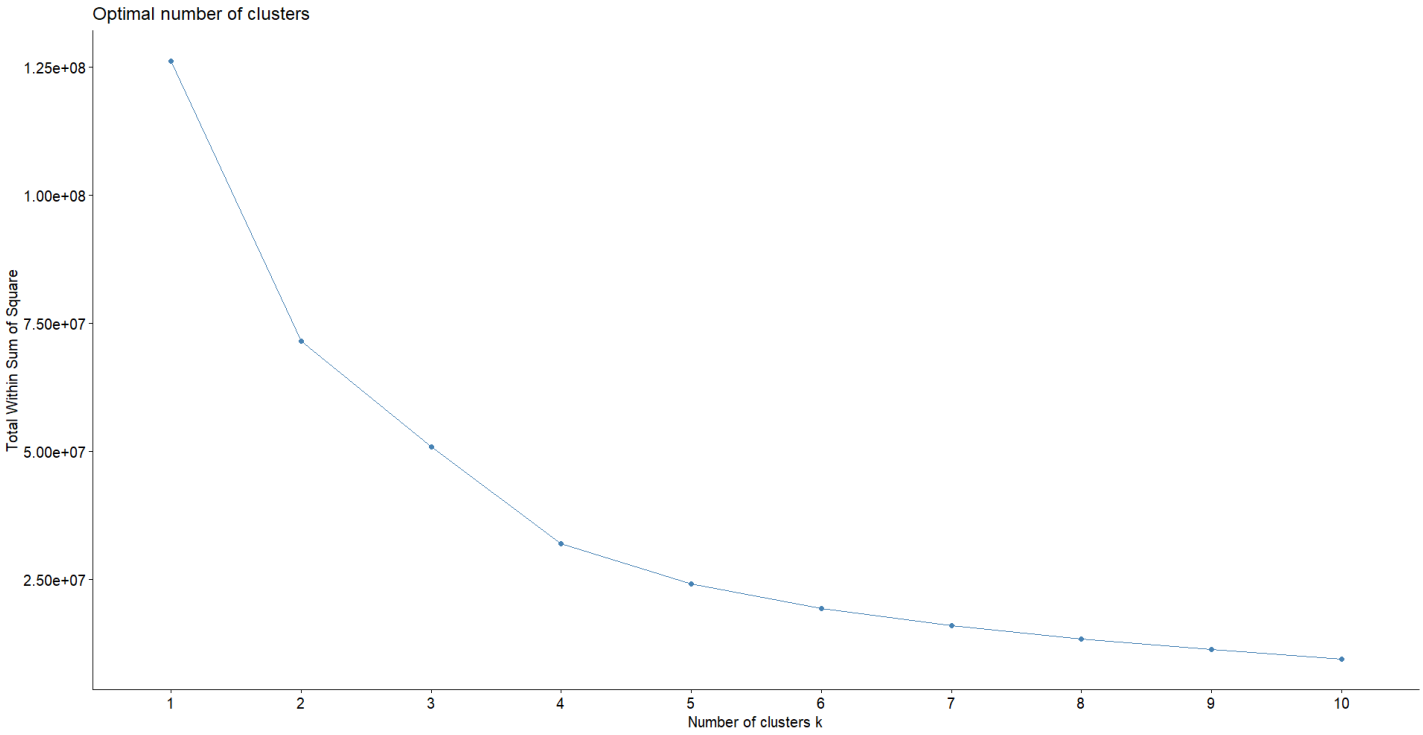
- Το σχήμα του K-Means με την χρήση του wss δημιουργεί κύρτωση στα k=3 και k=4. Ωστόσο, όπως αναφέραμε παραπάνω, η εν λόγω μέθοδος δεν είναι κατάλληλη για τα δεδομένα μας.



- Ο Hierarchical Clustering με βάση το Silhouette Coefficient εμφανίζει optimal k=4. Βέβαια, βλέποντας και το σχήμα τα αποτελέσματα φαίνονται αρκετά καλά και για k=2, με την αύξηση των ομάδων κατά 2 να φέρει μικρή βελτίωση.



- Ο Hierarchical Clustering με τη χρήση του wss εμφανίζει κύρτωση για k=2 και k=4, με k=4 να εμφανίζει καλύτερα αποτελέσματα συσταδοποίησης, αφού φτάνει το τετραγωνικό σφάλμα σε χαμηλότερη τιμή.



Έχοντας εφαρμόσει πλέον τις απαραίτες μεθόδους, έχουμε να διαλέξουμε την καλύτερη από τις παραπάνω. Βλέποντας τα αποτελέσματα που παίρνουμε κάθε φορά από την συνάρτηση fviz_nbclust(), ο βέλτιστος αριθμός k είναι ανάμεσα σε 2-3-4 όπως θα περιμενάμε αρχικά και διαισθητικά. Αν και τις περισσότερες φορές παίρνουμε βέλτιστο k=4, παρατηρούμε ότι με τη μέθοδο Ward για k=3 οι συστάδες που καταλήγουμε είναι καλύτερες σε σχέση με το k=4. Καταληγούμε, λοιπόν, στη μέθοδο **Hierarchical Clustering με τη χρήση του Ward.D2 για k=3.**

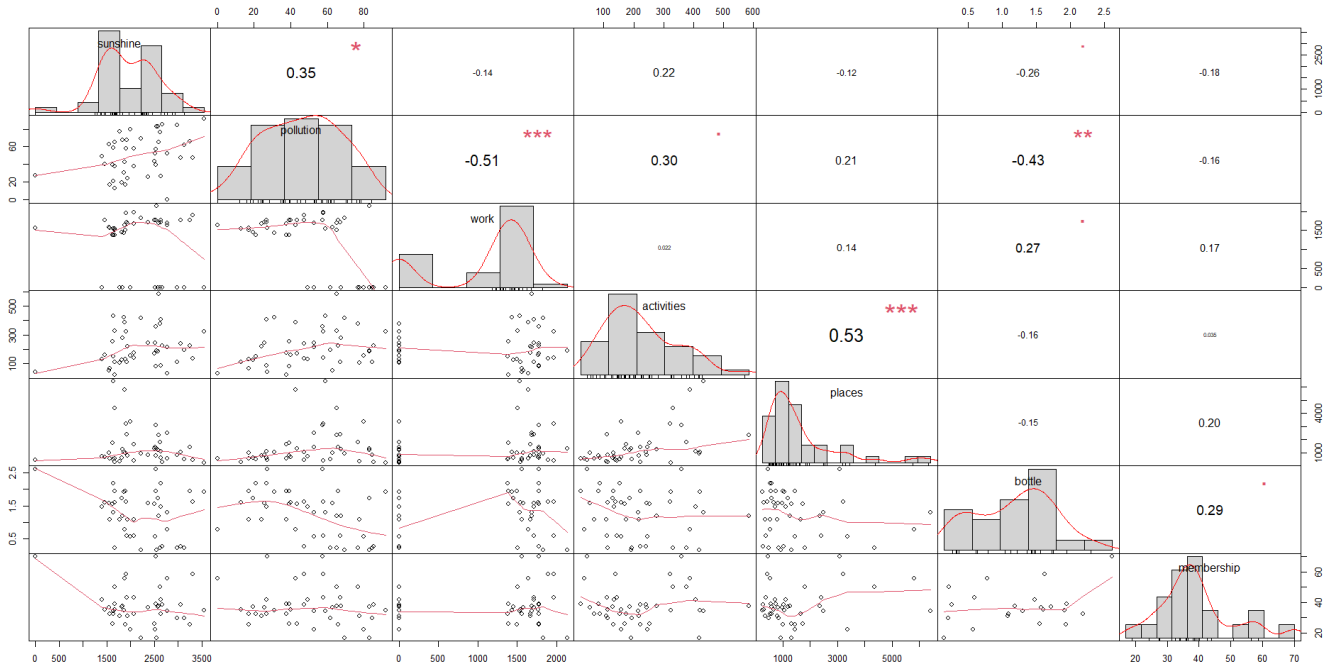
ii)

Η μέθοδος στην οποία καταλήξαμε πρέπει να κρίνουμε αν μπορεί να θεωρηθεί τελικά κατάλληλη ή όχι. Όπως, αναφέραμε και στο ερώτημα i) πρέπει να διερευνήσουμε τα δεδομένα μας.

Μετά την προεπεξεργασία που κάναμε στο dataset, προκύπτει πλέον ότι έχουμε κάποια outliers. Ωστόσο, λόγω του μικρού όγκου των δεδομένων μας, η αφαίρεση τους θα επηρεάσει το τελικό αποτέλεσμα, αφού θα πρέπει να βγάλουμε από τους υπολογισμούς την εκάστοτε πόλη που έχει το outlier και εν τέλει η συσταδοποίηση θα είναι ελλειπής.

Όσον αφορά στη συσχέτιση που έχουν τα δεδομένα μεταξύ τους, χρησιμοποιώντας το chart.Correlation, παρατηρούμε τόσο την ύπαρξη outliers που αναφέραμε, όσο και την

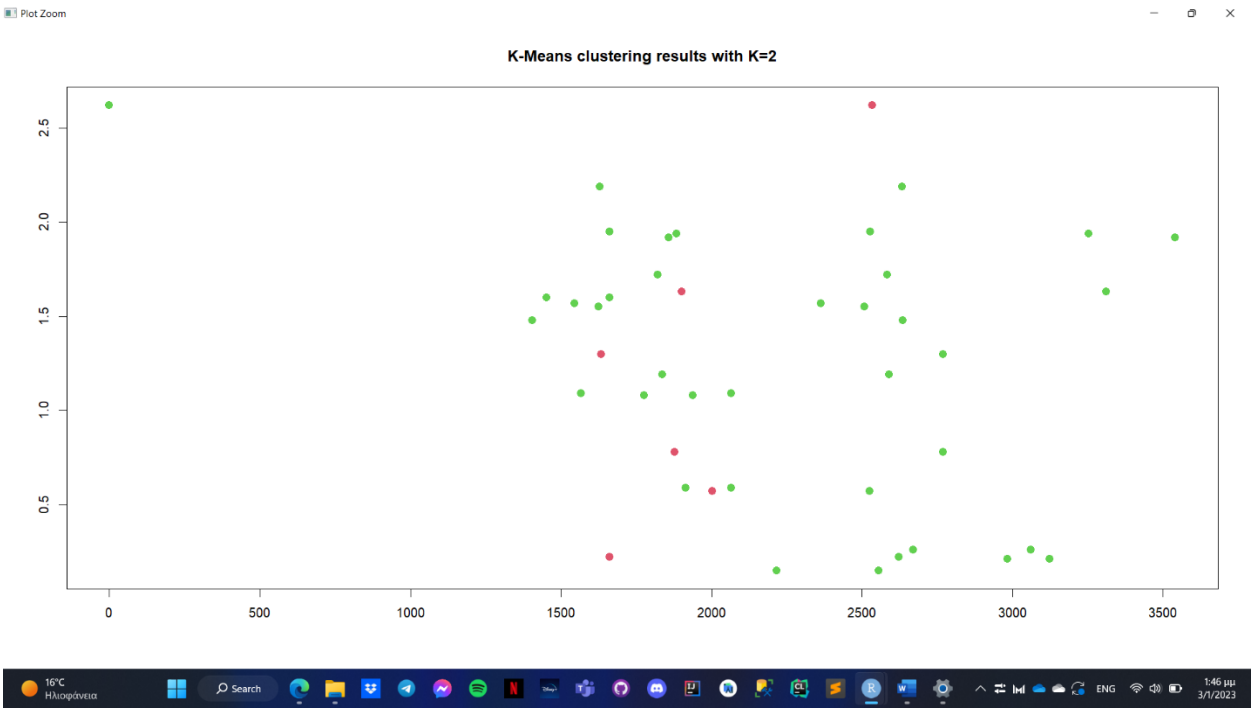
ασθενή συσχέτιση των μεταβλητών μεταξύ τους. Είναι, επίσης, εμφανές οτί η κατανομή των δεδομένων δεν είναι ευνοϊκή για την διεξαγωγή στατιστικά σημαντικών αποτελεσμάτων.

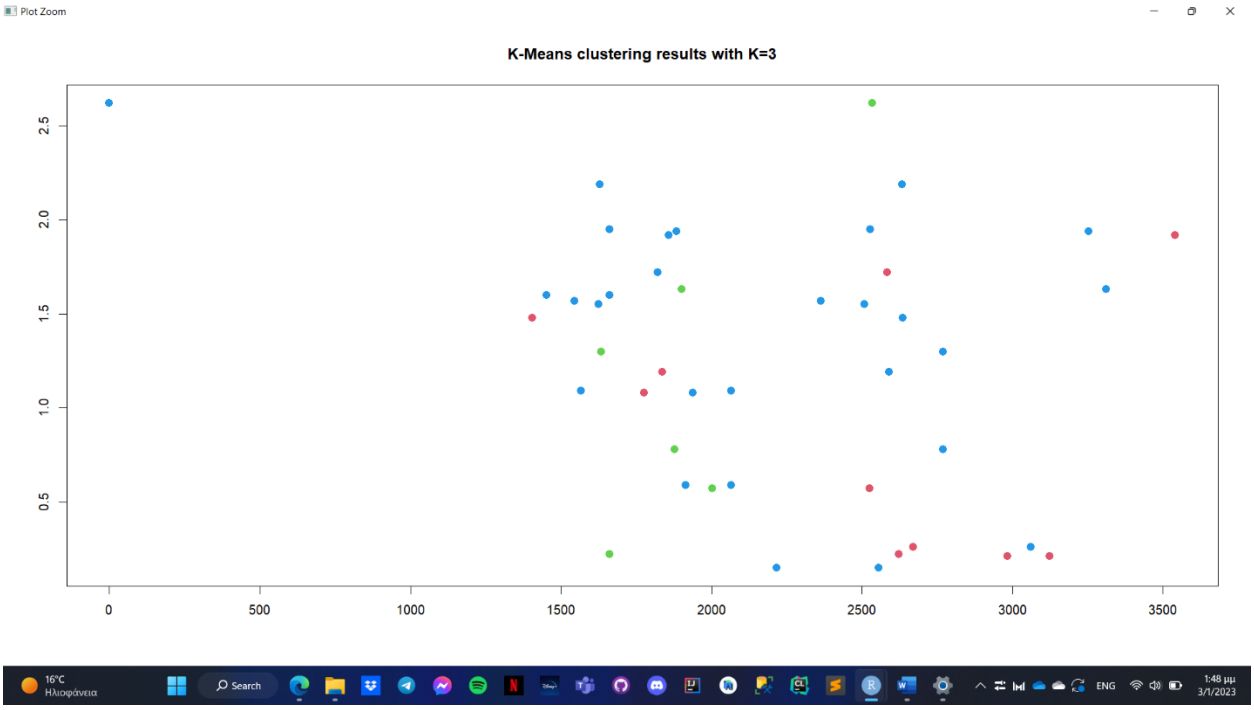


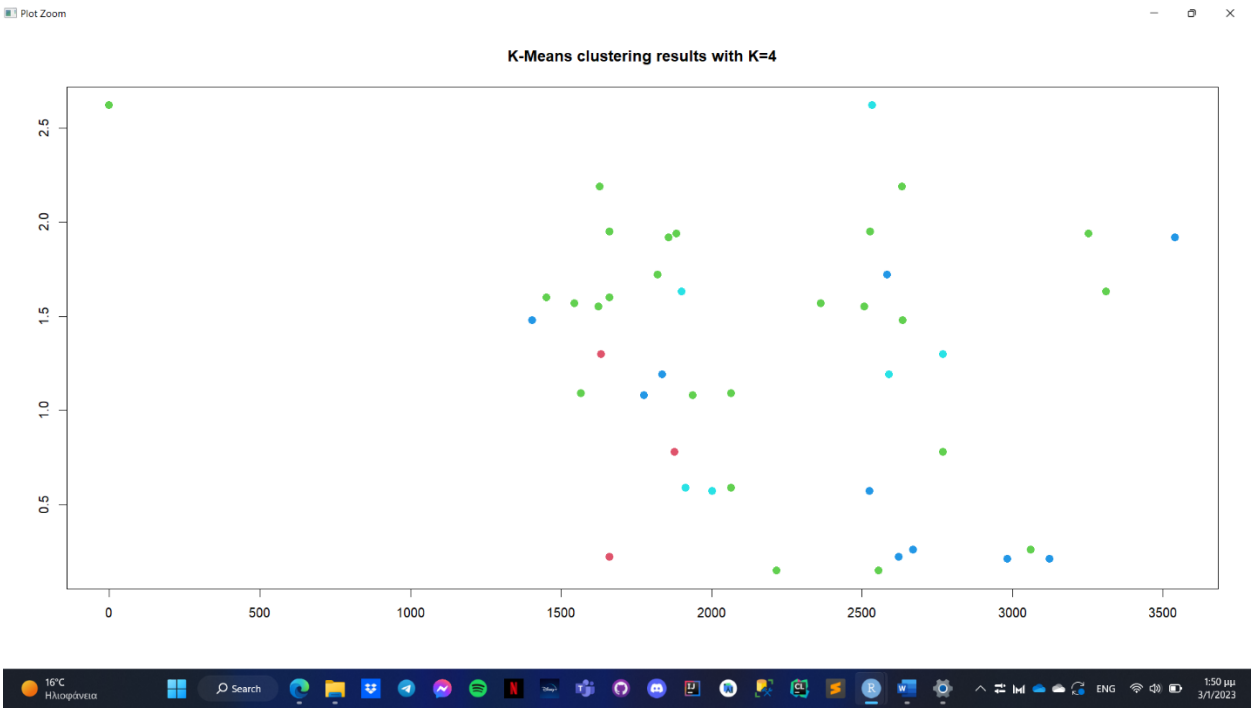
Τελικά, με βάση των παραπάνω, η τελική συσταδοποίηση ενδέχεται να **μην** είναι επιτυχημένη.

iii)

Όπως αναφέραμε στο ερώτημα i), σε πολλές από τις προσπάθειες συσταδοποίησης των δεδομένων, οι ομάδες δεν είναι διακριτώς διαχωρίσιμες. Αυτό μπορεί να φανεί ξεκάθαρα από την οπτικοποίηση των αποτελεσμάτων του K-Means. Προφανώς, οι ομάδες δεν έχουν σφαιροειδές σχήμα και τα δεδομένα της εκάστοτε ομάδας δεν διαχωρίζονται από τα δεδομένα της άλλης ομάδας. Αυτό δεν αλλάζει για οποιαδήποτε τιμή του k δοκιμάσουμε στον αλγόριθμο. Παρατίθονται τα σχετικά σχήματα:


Plot Zoom
K-Means clustering results with K=2
2.5
2.0
1.5
1.0
0.5
0
500
1000
1500
2000
2500
3000
3500
16°C
Ηλιοφάνεια
Search
ENG
1:46 μμ
3/1/2023


Plot Zoom
K-Means clustering results with K=3
2.5
2.0
1.5
1.0
0.5
0
500
1000
1500
2000
2500
3000
3500
16°C
Ηλιοφάνεια
Search
ENG
1:48 μμ
3/1/2023



Από την άλλη, όπως αναλύσαμε στο ερώτημα i), χρησιμοποιώντας Hierarchical Clustering και ιδίως με τη βοήθεια της μεθόδου Ward, οι ομάδες που προέκυψαν ήταν διακριτά διαχωρίσιμες είχαν συσταδοποιηθεί με σαφήνεια.

iv)

Για την ερμηνεία των αποτελεσμάτων μας έγινε αντιστοίχιση της κάθε πόλης με την εκάστοτε ομάδα, χρησιμοποιώντας την εντολή cbind. Συγκεκριμένα, μελετώντας τα δεδομένα:

- Η **ομάδα 1** συμπεριλαμβάνει τις πόλεις με την **καλύτερη** ποιότητα ζωής. Συνεπώς, στην ομάδα αυτή οι πόλεις έχουν πολλές επιλογές σε activities και μέρη ψυχαγωγιάς (places), τις χαμηλότερες τιμές εργατοωρών και τις μικρότερες τιμές στην ρύπανση.
- Η **ομάδα 2** περιλαμβάνει τις πόλεις με την **μεσαία** ποιότητα ζωής. Εδώ, οι πόλεις έχουν αρκετά ικανοποιητικό πλήθος σε ψυχαγωγία και δραστηριότητες, αυξημένες εργατοώρες και συνήθως μεγαλύτερο δείκτη ρύπανσης.
- Τέλος, **η ομάδα 3** περιέχει τις πόλεις με την **χειρότερη** ποιότητα ζωής. Συνήθως εδώ, ο δείκτης ρύπανσης της πόλης είναι πολύ μεγάλος, οι εργατοώρες αυξημένες και το πλήθος των δραστηριοτήτων μειωμένο.

Βέβαια, για τους λόγους που αναφέραμε στο ερώτημα ii), υπάρχουν ορισμένες πόλεις που δεν έχουν ενταχθεί στην κατάλληλη ομάδα (π.χ. το Mexico City θα έπρεπε να ενταχθεί στην ομάδα

2 ή 3 λόγω του υψηλού δείκτη ρύπανσης και των αυξημένων εργατοωρών, αλλά βρέθηκε στην ομάδα 1). Ωστόσο, συνολικά η συσταδοποίηση φαίνεται στην πλειοψηφία της να είναι έγκυρη.