ΠΑΝΕΠΙΣΤΗΜΙΟ ΔΥΤΙΚΗΣ ΑΤΤΙΚΗΣ
UNIVERSITY OF WEST ATTICA

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

ΕΡΓΑΣΙΑ

ΑΝΑΛΥΣΗ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΠΑΡΟΥΣΙΑΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ

ΣΤΟΙΧΕΙΑ ΟΜΑΔΑΣ

ΣΤΟΙΧΕΙΑ ΦΟΙΤΗΤΗ 1: ΑΘΑΝΑΣΙΟΥ ΒΑΣΙΛΕΙΟΣ ΕΥΑΓΓΕΛΟΣ (ΠΑΔΑ-19390005)
ΣΤΟΙΧΕΙΑ ΦΟΙΤΗΤΗ 2: ΠΕΤΡΟΠΟΥΛΟΣ ΠΑΝΑΓΙΩΤΗΣ (ΠΑΔΑ-20390188)
ΣΤΟΙΧΕΙΑ ΦΟΙΤΗΤΗ 3: ΤΑΤΣΗΣ ΠΑΝΤΕΛΗΣ (ΠΑΔΑ-20390226)

ΥΠΕΥΘΥΝΟΣ ΜΑΘΗΜΑΤΟΣ: ΑΝΔΡΙΤΣΟΣ ΠΕΡΙΚΛΗΣ
ΗΜΕΡΟΜΗΝΙΑ ΥΠΟΒΟΛΗΣ: 21/6/2024
ΗΜΕΡΟΜΗΝΙΑ ΠΡΟΘΕΣΜΙΑΣ: 21/06/2024

ΠΕΡΙΕΧΟΜΕΝΑ

Εισαγωγή

Εισαγωγή

Στην ακόλουθη εργασία επιχειρήσαμε τη σύγκριση 2 datasets για κοινά έτη (2015-2016), με στόχο την ανάλυση και την διερεύνηση πιθανής σχέσης μεταξύ της ανεργίας και των περιστατικών δολοφονίας από αστυνομικούς σε διάφορες πολιτείες των ΗΠΑ. Το πρώτο dataset αφορά τα ποσοστά ανεργίας , ενώ το δεύτερο dataset τις δολοφονίες από αστυνομικούς.

Ως κοινό παρονομαστή των 2 datasets χρησιμοποιήσαμε την κάθε πολιτεία, ενώ στη συνέχεια θεωρήσαμε απαραίτητη την χρήση ενός επιπλέον dataset που θα περιλαμβάνει τον πληθυσμό των πόλεων στις ΗΠΑ. Αυτό έγινε για να προσδώσουμε μία πιό ολοκληρωμένη αξιολόγηση των δεδομένων μας.

Η τεχνική που χρησιμοποιήσαμε για την ανάλυση των δεδομένων μας ήταν αυτή της συσταδοποίησης και πιό συγκεκριμένα αξιοποιήσαμε τον αλγόριθμο k-means. Με τη χρήση αυτής της τεχνικής μπορέσαμε να ομαδοποιήσουμε τα δεδομένα μας σε συστάδες και να παρατηρήσουμε μοτίβα και συσχετίσεις σε αυτές. Για την αξιολόγηση της ποιότητας που έχουν αυτές οι συστάδες, εμπιστευτήκαμε μη προβλεπόμενα μέσα όπως το SSE (sum of squarred error) και τον συντελεστή περιγράμματος (silhouette coefficient), εξασφαλίζοντας έτσι πιο ποιοτικές συστάδες.

Η εργασία συνολικά περιέχει μια ενδελεχή ανάλυση μεταξύ των 2 κύριων μεταβλητών μας (ανεργία, φόνοι σε ποσοστά και καθαρούς αριθμούς), συνεισφέροντας στη συζήτηση για το αν κοινωνικο-οικονομικοί παράγοντες όπως η ανεργία δύναται να σχετίζονται με την έξαρση της αστυνομικής βίας.

1. Ορισμός προβλήματος και κίνητρο

1.1 Περιγραφή του Προβλήματος

Υπάρχουν πολλά επαγγέλματα τα οποία θα μπορούσαν να ωφεληθούν από μία έρευνα που μπορεί να συσχετίσει ποσοστά ανεργίας ανά πολιτεία με αντίστοιχα περιστατικά φόνων τις ίδιες χρονιές. Επαγγέλματα που έχουν να κάνουν με την μελέτη της ανθρώπινης συμπεριφοράς (κοινωνιολόγοι, ψυχολόγοι, οικονομολόγοι κλπ), μπορούν να αξιοποιήσουν τα όποια πορίσματα σε άλλες έρευνες ή να εμβαθύνουν στα πραγματικά αίτια του προβλήματος.

1.2 Κίνητρα Αξιοποίησης της Έρευνας

Οι λόγοι που μπορεί να οδηγήσουν στην αξιοποίηση της έρευνας που κάνουμε (στην περίπτωση που αυτή δείξει κάποια συσχέτιση μεταξύ των δεδομένων που συγκρίνουμε), είναι πολλοί. Ωστόσο, δεν πρέπει να ξεχνάμε ότι η συσχέτισης μεταξύ 2 παραγόντων δεν συνεπάγεται σε καμία περίπτωση με αιτιότητα πίσω από κάποια ανθρώπινη συμπεριφορά. Μπορεί οι παράγοντες που οδηγούν σε φόνους να είναι άλλοι από αυτούς της ανεργίας. Η συσχέτιση είναι απλά η αρχή της κατανόησης πιθανών αιτιών που προκαλούν ένα πρόβλημα. Ωστόσο, ακόμα και αυτή μπορεί να οδηγήσει σε λήψη μέτρων που θα προσφέρουν μία βελτίωση στην δημόσια ασφάλεια.

1.2.1 Κατανόηση της εγκληματικότητας

Ένα από τα κυριότερα κίνητρα για να αναζητήσει κάποιος μία έρευνα που αφορά την ενδεχόμενη συσχέτιση μεταξύ υψηλής ανεργίας και υψηλών περιστατικών φόνων, είναι στην περίπτωση που γίνεται κάποια μελέτη της εγκληματικότητας και αναζητούνται ορισμένα αίτια της. Στην περίπτωση που ο συλλογισμός της έρευνας ευσταθεί και ταυτίζεται με άλλες παρόμοιες έρευνες, τότε μπορεί να χρησιμοποιηθεί ως αίτιο της εγκληματικότητας.

Υπενθυμίζουμε ωστόσο ότι αυτό δεν σημαίνει πως το πρόβλημα της εγκληματικότητας μπορεί να κατανοηθεί πλήρως, καθώς πάντα τα περιστατικά παραβατικότητας είναι απόρροια της ανθρώπινης συμπεριφοράς.

1.2.2 Δημιουργία ενεργειών κοινωνικής πολιτικής

Άλλο ένα κίνητρο για την επαλήθευση μιας έρευνας που αφορά τη συσχέτιση οικονομίας με μια ανθρώπινη παραβατική ενέργεια είναι η διαμόρφωση πολιτικών αποφάσεων. Αν αποδειχθεί μέσα από έρευνες ότι η ανεργία μπορεί να επηρεάσει σημαντικά τον αριθμό των φόνων που πραγματοποιούνται, υπάρχει ένας επιπλέον λόγος τα κράτη να μεριμνήσουν για το ζήτημα και να αναζητήσουν λύση στο πρόβλημα της ανεργίας.

2. Σύντομη περιγραφή του συνόλου δεδομένων

2.1 Τα Δεδομένα της Εργασίας

Τα δεδομένα της εργασίας αντλήθηκαν από το Kaggle, ένα μέσο που αξιοποιούν data scientists. Τα δεδομένα μας περιλαμβάνουν ποσοστά και αριθμούς ανθρώπων σε φτώχεια και τον αριθμό φόνων από αστυνομικούς σε διάφορες πολιτείες των ΗΠΑ για διάφορα έτη. Συγκεκριμένα, το σύνολο δεδομένων περιλαμβάνει τις μεταβλητές Ποσοστό Φτώχειας . Αριθμούς σε φτώχεια και Κάθε φόνος ξεχωριστά ως εγγραγή για κάθε πολιτεία. Τα δεδομένα χρησιμοποιήθηκαν για να διερευνηθεί η πιθανή συσχέτιση μεταξύ της ανεργίας και της εγκληματικότητας κατά την εν λόγω περίοδο. Ως κλειδί για την ένωση των 2 αυτών μεταβλητών χρησιμοποιήσαμε την πολιτεία.

Στη διάρκεια της εργασίας κρίναμε απαραίτητη την εισαγωγή μιας επιπλέον μεταβλητής, του Συνολικού Πληθυσμού Πόλεων στις Ηνωμένες Πολιτείες και την μετατροπή του σε πληθυσμό ανα πολιτεία. Η άντληση του συγκεκριμένου dataset έγινε από μια γνωστή πλατφόρμα διαχείρισης και διάθεσης δεδομένων, την opendatasoft.

2.1.1 Ανάλυση των Δεδομένων των Φόνων από Αστυνομική Βία

Τα δεδομένα που ανακτήσαμε έπρεπε να τα επεξεργαστούμε με τέτοιο τρόπο, ώστε να περιορίζονται μεταξύ των ετών 2015-2016. Η αποθήκευση τους για την απαραίτητη προεπεξεργασία, γίνεται σε .csv αρχεία. Από αυτά τα έτη είχαμε δεδομένα για όλη την χρονιά. Ορισμένα από τα στοιχεία του πίνακα των φόνων από αστυνομικούς κατά τα έτη που αναφέραμε είναι: η πολιτεία που τελέστηκε ο φόνος, η ημερομηνία, ο αριθμός των σφαιρών που χρησιμοποιήθηκε, η αιτία θανάτου, η ηλικία του θύματος, η πόλη στην οποία έγινε ο φόνος.

Υπάρχουν και άλλες ανησυχητικές πληροφορίες στα δεδομένα μας, οι οποίες και στα 2 έτη είναι κατά συντριπτική πλειοψηφία παρόμοιες. Όλα τα θύματα ήταν άνδρες, σχεδόν όλοι δεν ειχαν δείγματα ψυχικής ασθένειας, ενώ σχεδόν σε κανένα φόνο δεν υπήρχε εφαρμογή της αστυνομικής κάμερας.

Τα τελικά μας δεδομένα περιορίστηκαν στην Πολιτεία και τον αριθμό και ποσοστό των θανάτων σε αυτή.

2.1.2 Ανάλυση των Δεδομένων της Ανεργίας ανά Πολιτεία

Τα δεδομένα της ανεργίας ανά πολιτεία περιορίζονται και αυτά στα έτη 2015-2016, διότι θέλουμε να ασχοληθούμε με αυτές τις συγκεκριμένες χρονικές περιόδους. Τα δεδομένα αυτά αποθηκεύτηκαν και αυτά σε .csv αρχεία ώστε να καταφέρουμε να τα επεξεργαστούμε κατάλληλα. Ορισμένα από τα αρχικά στοιχεία των δεδομένων μας ήταν η πολιτεία, ο αριθμός των ανέργων της, το έτος και το ποσοστό των ανέργων σε σχέση με τον πληθυσμό της πολιτείας.

Τα τελικά μας δεδομένα περιορίστηκαν σε πολιτεία, αριθμό ανέργων και ποσοστό ανέργων.

2.1.3 Ανάλυση δεδομένων Πληθυσμού ανα Πολιτεία

Τα δεδομένα του πληθυσμού εισήχθησαν μεταγενέστερα στην εργασία ως μέτρο επιβεβαίωσης των αποτελεσμάτων. Το περιεχόμενό τους αφορούσαν το έτος 2015 . Ο πληθυσμός αφορούσε τις πόλεις της Αμερικής με πληθυσμό μεγαλύτερο της τάξης των 65.000 ανθρώπων.Η επεξεργασία αυτών των δεδομένων ήταν καθοριστική για το τελικό πόρισμα της εργασίας, καθώς μας έδωσε μια καλύτερη σύνδεση μεταξύ του συνολικού πληθυσμού και των αποτελεσμάτων που βρήκαμε από τη χρήση των 2 προηγούμενων datasets.

3. Περιγραφή της μεθόδου ανάλυσης των δεδομένων

Πριν αρχίσουμε την ανάλυση σε αυτό το σημείο να υπενθυμίσουμε οτι το dataset με τον πληθυσμό μπήκε στο τέλος περισσότερο πειραματικά να επιβεβαιώσει τα αποτελέσματα του πειραματικού μέρους.

3.1 Προεπεξεργασία Δεδομένων

Σε αυτή την υποενότητα, θα αναλύσουμε όλη την προετοιμασία που χρειάστηκε να γίνει πάνω στα δεδομένα πριν την συσταδοποίηση.

3.1.1 Κλάδεμα δεδομένων

Όπως αναλύσαμε και στο προηγούμενο κεφάλαιο, χρησιμοποιούμε δύο dataset όπου το καθένα έχει διαφορετικό χρονικό εύρος (συγκεκριμένα το dataset με τους φόνους έχει δεδομένα από το 2015 έως τα μέσα του 2017 και το dataset με τα ποσοστά φτώχειας για τα έτη 2011 έως και 2021). Τα δεδομένα που θα χρησιμοποιήσουμε πρέπει να συμπίπτουν χρονικά , αλλιώς δεν θα είχε νόημα να προσπαθήσουμε να τα συσχετίσουμε. Γι αυτό, επιλέξαμε ότι τα δεδομένα πρέπει να περιοριστούν για τα έτη 2015, 2016 και όχι το έτος 2017 διότι τα δεδομένα των φόνων δεν ήταν πλήρεις πάρα μόνο μέχρι τις 31/07/2017.

3.1.2 Καθορισμός στοιχείου ένωσης των dataset

Εφόσον έχουμε 3 διαφορετικά datasets έπρεπε να βρούμε τον «συνδετικό κρίκο τους», δηλαδή το στοιχείο στο οποίο τα δυο dataset θα γίνουν join. Ανακαλώντας τον πρωταρχικό μας στόχο, ο οποίος είναι να αποδείξουμε ότι οι πολιτείες με υψηλά ποσοστά ανεργίας, έχουν και υψηλά περιστατικά φόνων -αντίστοιχα για χαμηλά επίπεδα ανεργίας έχουμε λίγα περιστατικά φόνων- καταλήξαμε ότι το join πρέπει να γίνει με βάση την εκάστοτε πολιτεία.

3.1.3 Μετασχηματισμός δεδομένων σε κοινή κλίμακα

Ένα από τα σημαντικότερα προβλήματα που κληθήκαμε να λύσουμε είναι αυτό της εύρεσης και μετασχηματισμού των δεδομένων σε μία κοινή κλίμακα. Αρχικά τα δεδομένα της φτώχειας περιείχαν τα ποσοστά ανεργίας για την εκάστοτε πολιτεία (“Percentage in Poverty”) και αριθμούς ανέργων. Ωστόσο στο dataset με τους φόνους, δεν περιείχε το πλήθος τον φόνων ανα πολιτεία και προφανώς με βάση αυτό ούτε τα ποσοστά φόνων ανα πολιτεία . Επομένως δουλειά μας ήταν αρχικά να μετρήσουμε τους φόνους ανά έτος και τους φόνους σε κάθε πολιτεία σε “καθαρούς” αριθμούς και στην συνέχεια να τους μετατρέψουμε σε ποσοστά επί τοις εκατό για κάθε πολιτεία. Επίσης και στο dataset με τον πληθυσμό , πάλι δεν είχαμε τον πληθυσμό της εκάστοτε πολιτείας , αλλα τον πληθυσμό ανα πόλη της Αμερικής(για πληθυσμό > 65.000 κατοίκων). Επομένως έπρεπε και για αυτό να αθροίσουμε το πληθυσμό από κάθε πόλη κάθε πολιτείας.

3.1.4 Μετασχηματισμός δεδομένων σε κοινή μορφολογία

Σημαντικό πρόβλημα ήταν και η διαφορετική ονομασία με την οποία αναγράφονταν οι πολιτείες στα δεδομένα που είχαμε βρει. Στο dataset με τους φόνους οι πολιτείες ήταν γραμμένες σε συντομογραφία (π.χ. NW για Νέα Υόρκη), ενώ στο dataset με τα ποσοστά φτώχειας αναγράφονταν με το πλήρες όνομά τους. Έπρεπε λοιπόν να καθαρίσουμε μια κοινή μορφολογία. Αποφασίσαμε ότι θα πρέπει να μετατρέψουμε τα πλήρη ονόματα του dataset της φτώχειας σε συντομογραφία με την βοήθεια ενός πίνακα αντιστοίχησης.

Τέλος, για τυπικούς λόγους εκτός από τον υπολογισμό του ποσοστού των φόνων, στα υπόλοιπα πεδία (ηλικία, φύλο, φυλή κλπ) επιλέξαμε μια αντιπροσωπευτική τιμή σε κάθε πολιτεία με βάση την πιο συχνή εμφάνιση περιστατικών φόνων.

3.1.5 Αθροιση φόνων

Όπως είπαμε τα δεδομένα τον φόνων ήταν σε μορφή εγγραφής, δηλαδή κάθε εγγραφή ήταν ενας ξεχωριστός φόνος , επομένως μετρήσαμε κάθε φόνο ξεχωριστά για κάθε πολιτεία και υπολογίσαμε και το ποσοστό σε σχέση με τους συνολικούς φόνους.

3.1.6 Διόρθωση πολιτειών χωρίς φόνους

Κάποιες πολιτείες , δεν είχαν φόνους για κάποιο έτος (π.χ. η πολιτεία Rhode Island είχε 0 φόνους για το έτος 2015) επομένως αυτό κατα το join των πινάκων θα έπρεπε να ληφθεί υπόψιν, αλλιώς δεν θα μπορούσε να συμπεριληφθεί(εισαγωγή τιμής 0).

3.1.7 Διόρθωση κενών δεδομένων

Τα δεδομένα ήταν πολύ καλα και δεν υπήρχαν κενά

3.1.8 Κανονικοποίηση δεδομένων ως τυπικές αποκλίσεις

Τα δεδομένα πριν εισαχθούν στο αλγόριθμο μετατράπηκαν σε τυπικές αποκλίσεις γύρω από τον 0

3.1.9 Λογαριθμική κανονικοποίηση

Τα καθαρά δεδομένα , δηλαδή ο απόλυτος αριθμός των ανέργων σε κάθε πολιτεία και ο απόλυτος αριθμός των φόνων σε κάθε πολιτεία πριν την εφαρμογή του αλγορίθμου , κανονικοποιήθηκαν με χρήση λογαριθμικής κανονικοποίηση λόγω της διαφοράς σε volume (λόγω της διαφοράς του πληθυσμού).

3.2 Ανάλυση Δεδομένων

Για την ανάλυση των δεδομένων υπάρχουν πάρα πολλές διαφορετικές τεχνικές που μπορούν να επιτύχουν αυτόν τον σκοπό (π.χ. clustering analysis,regression analysis,factor analysis). Στην συγκεκριμένη εργασία χρησιμοποιήσαμε cluster analysis.

3.2.1 Γιατί Cluster Analysis

- Βοηθάει στην καλύτερη κατανόηση των διαφορών των ομάδων και στην εξαγωγή των συμπερασμάτων τους.
  - Στην δική μας περίπτωση , την κατανόηση του γιατί μπορεί κάποια πολιτεία της Αμερικής να έχει παραπάνω φόνους από κάποια άλλη.
- Εύκολη στην υλοποίηση.
- Μπορεί να προσφέρει γρήγορα συμπεράσματα χωρίς να παρέμβουμε πολύ πάνω στα δεδομένα.
- Εντοπισμός ασυνήθιστων δεδομένων , δηλαδή αν κάποια δεδομένα είναι outliers μπορεί να μπούν μόνα τους σε μία μικρή συστάδα το οποίο είναι σημάδι ότι αποτελούν εξαιρέσεις.
- Δεν απαιτεί ετικέτες δεδομένων.
-

3.2.2 Γιατί K-Means

- Ταχύτερος σε σχέση με τις άλλες κατηγορίες συσταδοποίηση όπως ιεραρχική και πυκνωτική
- Καλύτερος σε όχι τόσο μεγάλα δεδομένα (στην δική μας περίπτωση 51)
- Είναι απλός και εύκολος στην υλοποίηση
- Καλή λύση στην περίπτωση μας που δεν έχουμε δεδομένα που γνωρίζουμε τις ετικέτες τους , όπως για παράδειγμα έχουμε στα δεδομένα iris

- Είναι καλός στο να εξακριβώνει κρυφά μοτίβα ειδικά σε δεδομένα που δεν γνωρίζουμε από την αρχή την αποτύπωση τους στο χώρο .

3.4 Ποιότητα συστάδων

3.4.1 Sum of Squared Error

Γενικά το SSE είναι ένα μέτρο το οποίο υπολογίζει την απόκλιση των πραγματικών τιμών σε σχέση με των προβλεπόμενων σε ένα σύνολο δεδομένων

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Μπορεί στην συσταδοποίηση να χρησιμοποιηθεί ως μέτρο για να κατανοήσουμε το πόσο κοντά είναι τα δεδομένα συνολικά σε όλες τις συστάδες με τύπο:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2$$

3.4.2 Συντελεστής Περιγράμματος

Σε μια συσταδοποίηση είναι πολύ σημαντικό τα δεδομένα σε κάθε cluster να είναι πολύ κοντά μεταξύ τους(συνοχή) και καλά διαχωρισμένα σε σχέση με τις άλλες συστάδες(διαχωρισμός).Ο συντελεστής περιγράμματος είναι ένα μέτρο το οποίο λαμβάνει υπόψη και τα δυό .

$$s_i = (b_i - a_i) / \max(a_i, b_i)$$

Όπου a είναι η μέση απόσταση του i από τα σημεία της συστάδας του και β η μικρότερη απόσταση του i απο το κοντινότερο σημείο μιας άλλη συστάδας.Λαμβάνει τιμές από το -1 εως το 1. Θετική τιμή σημαινει οτι η συσταδοποίηση είναι αποδεκτή(όσο πιο κοντα στο 1 τόσο το καλύτερο) ενώ μια αρνητική τιμή δεν είναι αποδεκτή.

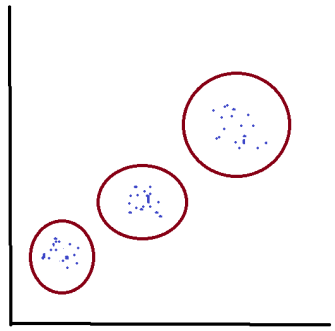
3.2.5 Στάδια Ανάλυσης

1. Στο πρώτο στάδιο χρησιμοποιήσαμε τα ποσοστά ανεργίας ανα πολιτεία και τα ποσοστά φόνων σε σχέση με τους συνολικούς φόνους ανα πολιτεία. Τα αποτελέσματα αυτής της φάσης και έπειτα από επανεξέταση του τρόπου ανάλυσης αποφασίσαμε ότι έπρεπε να αλλάξουμε την μεθοδολογία μας.
2. Στο δεύτερο στάδιο χρησιμοποιήσαμε τους καθαρούς αριθμούς. Δηλαδή τον απόλυτο αριθμό φόνων ανα πολιτεία και τον απόλυτο αριθμό ανθρώπων σε φτώχεια ανα πολιτεία.
3. Στο Τρίτο στάδιο χρησιμοποιήσαμε λογαριθμική κανονικοποίηση των απόλυτων αριθμών.
4. Τέλος χρησιμοποιήσαμε το τρίτο dataset με τον πληθυσμό για την κανονικοποίηση των δεδομένων σε καθαρούς αριθμούς.

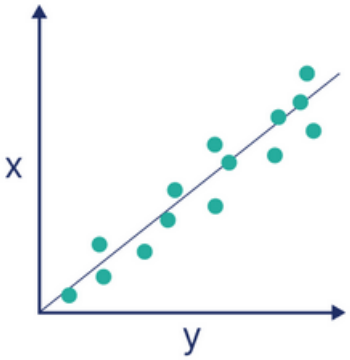
Τους λόγους που αλλάξαμε μεθοδολογία και πήγαμε από την πρώτη φάση στην τελευταία θα τους εξηγήσουμε στην κριτική συζήτηση των αποτελεσμάτων.

3.5 Αποτελέσματα που αναμένουμε

Σε αυτό το σημείο θα δείξουμε πως φανταζόμαστε τα αποτελέσματα για να εξάγουμε το συμπέρασμα οτι υπάρχει συσχέτιση



ή

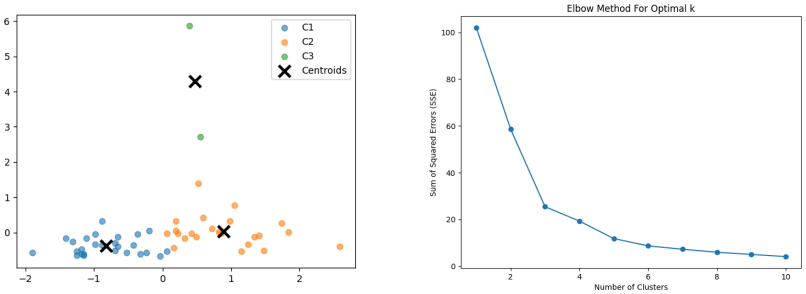


4. Πειραματικά αποτελέσματα

4.1 Αποτελεσμάτων ποσοστών για τα έτη 2015, 2016 και μέσος όρος 2015-2016

4.1.1 Αποτελέσματα ποσοστών για το έτος 2015

Ποσοστά φτώχειας και ποσοστά φόνων



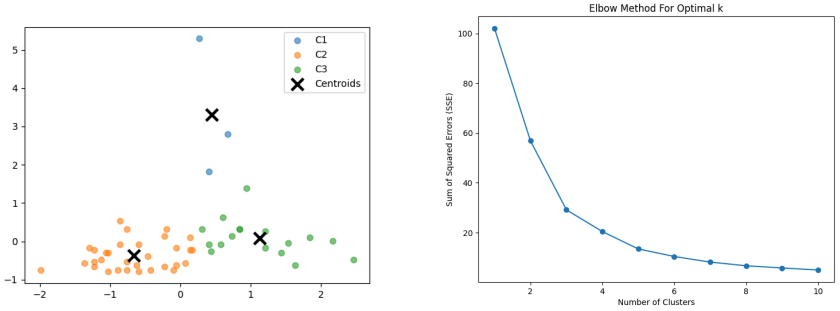
Εικόνα 1

SSE (Sum of Squared Errors) = 25.464

Silhouette Coefficient: 0.518

4.1.2 Αποτελέσματα ποσοστών για το έτους 2016

Ποσοστά φτώχειας και ποσοστά φόνων



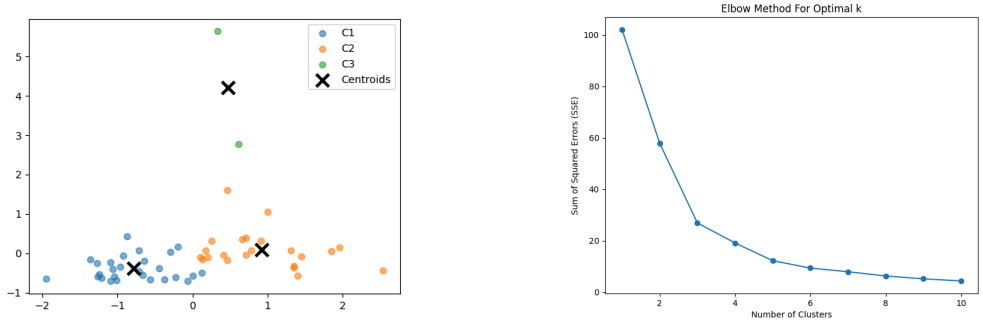
Εικόνα 2

SSE (Sum of Squared Errors) = 29.151

Silhouette Coefficient: 0.495

4.1.3 Αποτελέσματα μέσον όρων ποσοστών για τα ετοι 2015-2016

Ποσοστά φτώχειας και ποσοστά φόνων



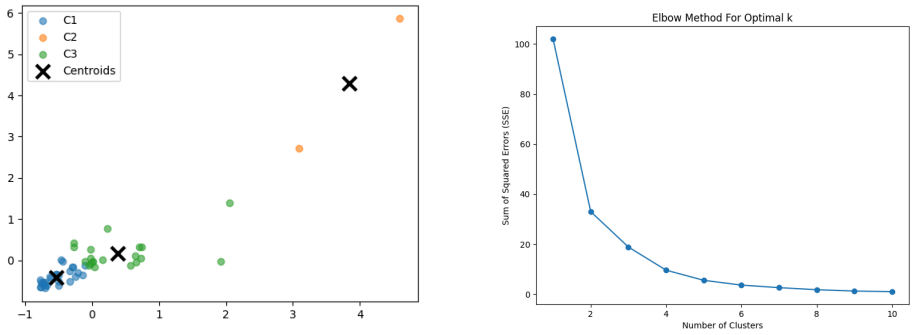
Εικόνα 3

SSE (Sum of Squared Errors) = 26.879

Silhouette Coefficient: 0.501

4.2.1 Αποτελέσματα καθαρών αριθμών για το έτος 2015

Αριθμός ανθρώπων σε φτώχεια και αριθμός φόνων



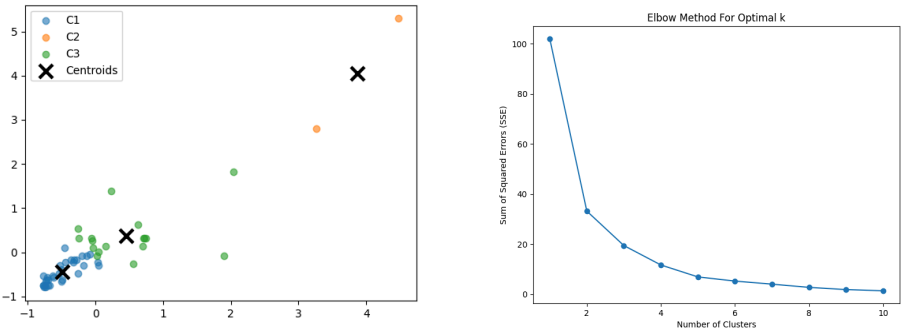
Εικόνα 4

SSE (Sum of Squared Errors) = 18.885

Silhouette Coefficient: 0.481

4.2.2 Αποτελέσματα καθαρών αριθμών για το έτος 2016

Αριθμός ανθρώπων σε φτώχεια και αριθμός φόνων



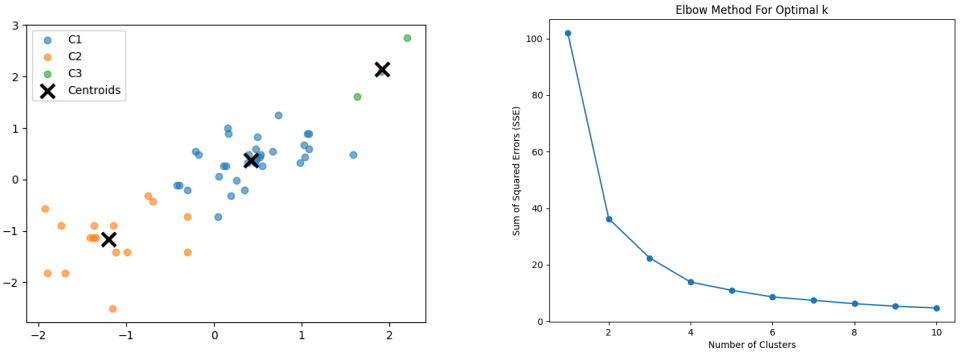
Εικόνα 5

SSE (Sum of Squared Errors) = 19.478

Silhouette Coefficient: 0.484

4.2.3 Αποτελέσματα καθαρών αριθμών λογαριθμική κανονικοποίηση για το έτος 2015

Αριθμός ανθρώπων σε φτώχεια και αριθμός φόνων



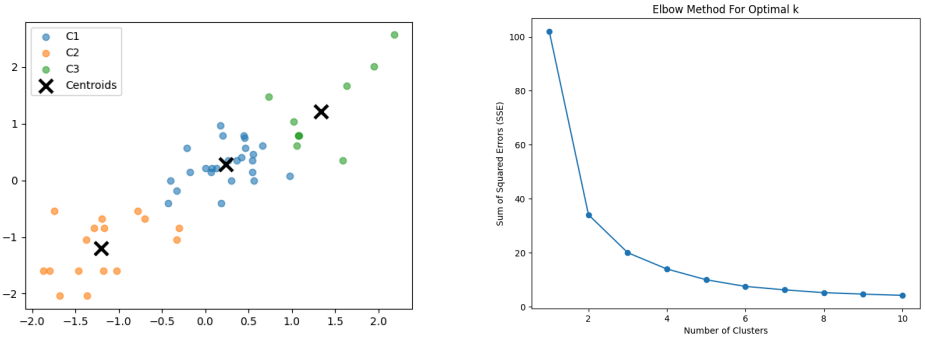
Εικόνα 6

SSE (Sum of Squared Errors) = 22.391

Silhouette Coefficient: 0.554

4.2.4 Αποτελέσματα καθαρών αριθμών λογαριθμική κανονικοποίηση για το έτος 2016

Αριθμός ανθρώπων σε φτώχεια και αριθμός φόνων



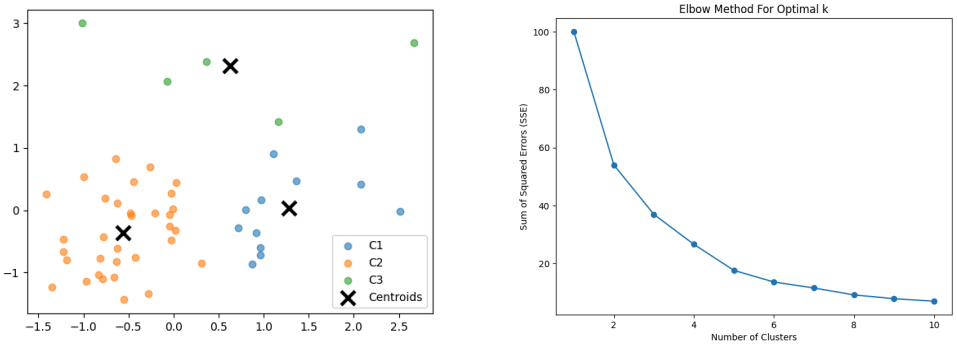
Εικόνα 7

SSE (Sum of Squared Errors) = 20.014

Silhouette Coefficient: 0.491

4.2.4 Αποτελέσματα καθαρών αριθμών κανονικοποίηση με πληθυσμο για το έτος 2015

Αριθμός ανθρώπων σε φτώχεια και αριθμός φόνων και αριθμός πληθυσμού.



Εικόνα 8

SSE (Sum of Squared Errors) = 36.972

Silhouette Coefficient: 0.453

5. Συζήτηση/Κριτική αποτίμηση αποτελεσμάτων

Στα δεδομένα που θα δούμε κάθε σημείο αποτελεί κάποια πολιτεία. Οι άξονες χ και y αλλάζουν με βάση το πιο πεδίο του dataset επιλέγουμε.Ο άξονας x ανήκει στην κατηγορία αστυνομικής βίας και ο y στην ανεργία.Ο αριθμός των k στις συσταδοποιήσεις προ καθορίζεται με βάση την elbow method.

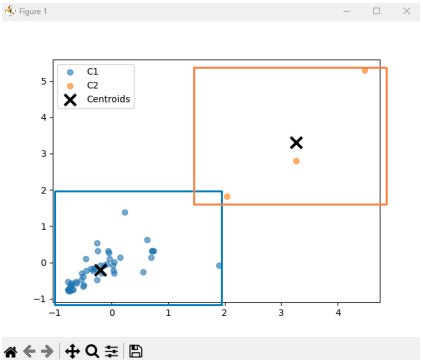
5.1 Κριτική Ανάλυση Αποτελεσμάτων ποσοστών για τα έτη 2015, 2016 και μέσος όρος 2015-2016

Για τα έτη 2015 , 2016 και συνδυασμός αυτών με χρήση μέσου όρου(εικόνες 1,2,3 αντίστοιχα) παρατηρούμε οτι συμπέρασμα που θέλουμε να αποδείξουμε , δηλαδή ότι στις πολιτείες με περισσότερη ανεργία υπάρχουν περισσότερη αστυνομική βια δεν αποτυπώνεται αφού δεν αυξάνεται το ποσοστό φόνων καθώς μεγαλώνουν τα ποσοστά ανεργίας. Παρατηρούμε ότι στο γράφημα γενικά υπάρχει μια νόρμα σημείων κάτω του y=1. Με λίγα λόγια καθώς αυξάνονται οι φόνοι δεν αυξάνεται και η ανεργία.

Η πρώτη φάση της ανάλυσης είχε απογοητευτικά αποτελέσματα. Σε αυτό το σημείο όμως καταλάβαμε ότι έχουμε κάνει σφάλμα στην ανάλυση . Δεν έχουμε συμπεριλάβει τον παράγοντα πληθυσμό που είναι πάρα πολύ σημαντικός. Είναι λογικό ότι πολιτείες με μεγάλα ποσοστά πληθυσμού(οπως η California) να έχουν μεγαλύτερους ποσοστά σε φόνους και μικρότερες πολιτείες (όπως η Rhode Island) να έχουν μικρότερα . Επίσης εσφαλμένη ήταν και η λογική με τα ποσοστά διότη , το ποσοστό της φτώχειας το οποίο μας δίνεται απο το dataset έχει υπολογιστεί με βάση τον πληθυσμό της κάθε πολιτείας (τα δεδομένα πληθυσμού δεν μας δίνονται και αυτός είναι ένας λόγος που χρησιμοποιούμε δικό μας dataset με πληθυσμό στην συνέχεια) ενώ το ποσοστό τον φόνων ανα πολιτεία(που το υπολογίζουμε εμείς) έχει υπολογιστεί με βάση τον συνολικό αριθμό φόνων και όχι με τον πληθυσμό. Άρα συγκρίναμε αποτελέσματα με διαφορετικό παρονομαστή. Το dataset είχε επίσης και καθαρούς αριθμούς ανθρώπων σε ανεργία και εμείς έχουμε υπολογίσει τους φόνους που αντιστοιχούν σε κάθε πολιτεία. Επομένως χρησιμοποιήσαμε αυτά τα δεδομένα για την περεταίρω ανάλυση.

5.2 Κριτική Ανάλυση Αποτελεσμάτων καθαρών αριθμών για τα έτη 2015, 2016

Μετά την εφαρμογή του K-Means στα δεδομένα (εικόνα 4,5) παρατηρούμε οτι μια πιο ελπιδοφόρα εικόνα. Τα περισσότερα δεδομένα και για τα δύο έτη είναι συγκεντρωμένα πριν το x = 1 και y = 1. Βλέπουμε λοιπόν μια νόρμα σε αυτό το σημείο, ενώ υπάρχουν κάποια δεδομένα που είναι πολύ μακριά από αυτή.



Αν εξαιρέσουμε την δεύτερη συστάδα που φαίνεται οτι τα δεδομένα δεν ακολουθούν την νόρμα , βλέπουμε στα δεδομένα της πρώτης οτι υπάρχει γενικά μια αυξητική τάση αποτέλεσμα που ικανοποιεί το συμπέρασμα που θέλουμε να εξάγουμε. Ωστόσο πάλι εδώ υπάρχει ένα σφάλμα στην ανάλυση. Δεν έχουμε πάλι κανονικοποίηση τα δεδομένα σε σχέση με τον πληθυσμό που είχαμε αναφέρει στην 5.1

5.3 Κριτική Ανάλυση Αποτελεσμάτων λογαριθμικά κανονικοποιημένων καθαρών αριθμών για τα έτη 2015, 2016

Τα δεδομένα των εικόνων 6 και 7 μας παρέχουν ένα πάρα πολύ καλό αποτέλεσμα . Βλέπουμε μια ξεκάθαρα αυξητική τάση και συστάδες είναι σχετικά καλές με βάση τα δυο μέτρα. Με αυτή την εικόνα μπορούμε να πούμε ότι όντως υπάρχει συσχέτιση των δύο αυτών στοιχείων . Επειδή όμως η λογαριθμική κανονικοποίηση είναι τυφλή. Δηλαδή δεν γνωρίζει τον πληθυσμό αλλα απλά εφαρμόζει έναν μαθηματικό τύπο που συμπιέζει

δεδομένα με ανόμοια διακύμανση , το ιδανικότερο θα ήταν να επιβεβαιωθεί μέσω του πραγματικού πληθυσμού ανα πολιτεία. Αυτό ακριβώς θα κάνουμε στο επόμενο βήμα (εδώ να σημειωθεί ότι τα δεδομένα πληθυσμού αφορουν μονο το έτος 2015 διότη δεν μπορούσαμε να βρούμε δεδομένα για το 2016).

5.4 Κριτική Ανάλυση Αποτελεσμάτων κανονικοποιημένων με βάση τον πληθυσμό καθαρών αριθμών για τα έτη 2015, 2016

Στην εικόνα 8 βλέπουμε μια όχι τόσο καλή εικόνα όπως σε αυτή της λογαριθμικής κανονικοποίησης . Παρατηρούμε οτι και πάλι υπάρχει μια αυξητική τάση ωστόσο τα δεδομένα είναι αρκετά αραιά (αποτυπώνεται και στο sse και στον συντελεστή περιγράμματος)με αποτέλεσμα να μην μπορούμε να εξάγουμε με ασφάλεια ότι υπάρχει συσχέτιση. Ωστόσο είναι η πιο σωστή μεθοδολογία από όλες διότη ανταποκρίνεται σε πραγματικά δεδομένα πληθυσμού και όχι σε τυφλές μεθόδους και υποθέσεις. Το σφάλμα της μεθόδου είναι ότι ο πληθυσμός αφορά πόλεις που έχουν πληθυσμό >= 65.000 κατοίκων μη προσμετρώντας τις μικρότερες πόλεις.

5.5 Για μελλοντική ανάλυση

Μερικά πράγματα που θα μπορούσαν να γίνουν για πιο ορθολογικά αποτελέσματα θα ήταν να εκφραστούν τα δεδομένα του πληθυσμού σε σχέση με τους φόνους και τους ανθρώπους σε φτώχια σε ποσοστά επί της 100. Αυτό μπορεί να φέρει τα δεδομένα πιο κοντα . Επίσης θα μπορούσε να χρησιμοποιηθεί ενα καλύτερα dataset με πληθυσμό που εμπεριέχει για όλες τις πόλεις των πολιτειών και όχι μόνο τις μεγαλύτερες.

Συμπεράσματα

Τα κοινωνικά δεδομένα είναι σύνθετα και πολυπαραγοντικά. Μια ανάλυση με δυο χαρακτηριστικά δεν είναι απαραίτητο να μπορεί να αποδείξει ότι υπάρχει σύνδεση ανάμεσα στην αστυνομική βία και την ανεργία. Για ασφαλέστερα αποτελέσματα θα έπρεπε να εισαχθούν και άλλες διαστάσεις στην ανάλυση. Όμως για τα συγκεκριμένα δεδομένα που έχουμε και την συλλογιστική που ακολουθήσαμε λόγο της αυξητικής φύσης των δεδομένων περισσότερο στην λογαριθμική κανονικοποίηση λιγότερο στην κανονικοποίηση με βάση τον πληθυσμό και στα καθαρά δεδομένα (δεν προσμετράμε την αρχική υπόθεση που κάναμε στα 5.1 επειδή βασίζεται σε λανθασμένες βάσεις.) πιστεύουμε ότι υπάρχει ναι μεν μια συσχέτιση αλλα όχι τόσο ισχυρή που μπορούμε να εξάγουμε με αυτοπεποίθηση το συμπέρασμα ότι ισχύει . Με λίγα λόγια θα λέγαμε ότι υπάρχει με βάση τα δεδομένα μια ένδειξη που απαιτεί επαλήθευση και από άλλα δεδομένα.

Σας ευχαριστούμε για την προσοχή σας.