ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ

ΔΗΜΗΤΡΗΣ ΖΕΡΚΕΛΙΔΗΣ 03400049
ΜΑΡΙΑ ΚΑΪΚΤΖΟΓΛΟΥ 03400052
ΜΑΡΙΑ-ΦΙΛΙΠΠΑ ΤΡΙΒΥΖΑ 03400080

Εισαγωγή

Αντικείμενο έρευνας

Η εργασία σχετίζεται με τα έντονα καιρικά φαινόμενα στις ΗΠΑ από το 1950 έως και το 2019. Πυρήνας της ανάλυσης είναι οι προκληθείσες καταστροφές σε σχέση με το χρόνο, την τοποθεσία και το είδος του φαινομένου. Μπορούμε να συνοψίσουμε την ανάλυσή μας στα παρακάτω τρία ερωτήματα:

- Η εξέλιξη στο χρόνο του μεγέθους καταστροφών από τα καιρικά φαινόμενα.
- Η γεωγραφική κατανομή των καταστροφών, κατά είδος και μέγεθος.
- Η σχέση του μεγέθους των καταστροφών με τον τύπο καταστροφών.

Χρησιμοποιήθηκε το Google Cloud για την εύρεση του dataset “Severe Storm Event Details”. Πηγή είναι το National Oceanic Atmospheric Administration (NOAA) και ανήκει στην Severe Weather Data Inventory (SWDI). Οι πληροφορίες αφορούν το είδος του φαινομένου, την τοποθεσία και τον αντίκτυπο σε υλικές ζημιές και στους ανθρώπους. Η διαδικασία με την οποία δομήσαμε την ανάλυσή μας αποτελείται από:

1. την προεπεξεργασία δεδομένων (data pre-processing)
2. την εξερέυνηση δεδομένων (exploratory data analysis)
3. την εξόρυξη δεδομένων (data mining, itemsets, clustering (DBSCAN))
4. την παρουσίαση των αποτελεσμάτων εξόρυξης (data presentation)

Προεπεξεργασία δεδομένων

Οι καταγραφές κατά τα πρώτα χρόνια περιέχουν πολλά ελλιπή στοιχεία και ότι υπάρχει σημαντική ασυνέπεια. Επιπλέον, υπήρχαν πολλές στήλες, οι οποίες δε χρησίμευαν για τους σκοπούς της ανάλυσής μας. Επομένως, πραγματοποιήσαμε τις εξής επεξεργασίες:

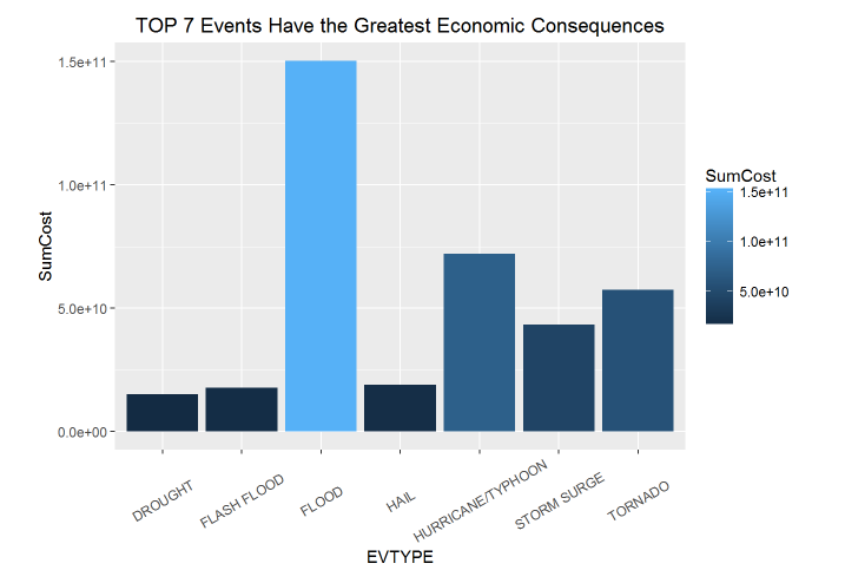
- Missing values: διαγραφή σειρών, συμπλήρωση σειρών με μέσους όρους ή 0
- Ενοποίηση χαρακτηριστικών (π.χ. class mapping, δηλαδή έχουν αντιστοιχηθεί event types σε άλλα παρόμοια για να μικρύνουμε τον αριθμό κατηγοριών)
- Διόρθωση ασυνεπών τιμών (π.χ. πληθωρισμό)
- Ομαδοποίηση τιμών χαρακτηριστικών
- Διαγραφή χαρακτηριστικών (π.χ. αφαίρεση event types τα οποία έχουν πολύ μικρή συχνότητα εμφάνισης)
- Δημιουργία νέων χαρακτηριστικών (π.χ. συνδυασμός άμεσων με έμμεσους θανάτους με ένα άθροισμα, το ίδιο και για τους τραυματισμούς)
- Μετασχηματισμός χαρακτηριστικών

ΠΡΟΣΕΓΓΙΣΕΙΣ ΣΕ ΣΧΕΤΙΚΑ ΠΡΟΒΛΗΜΑΤΑ

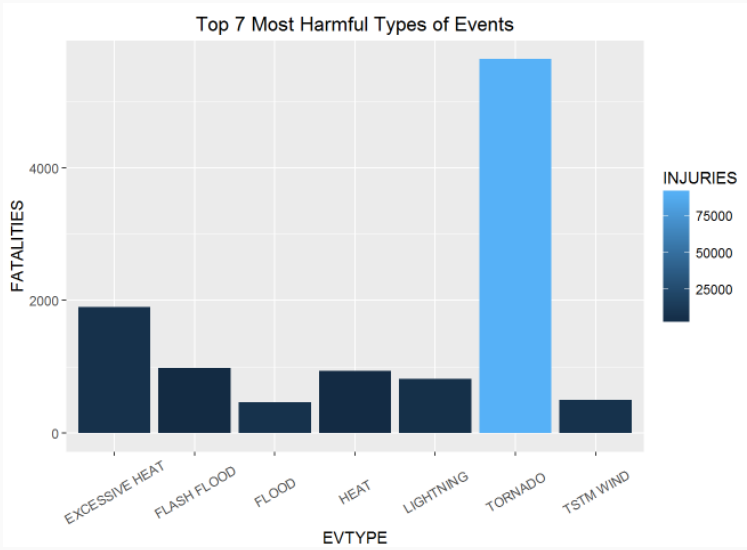
Exploratory Data analysis on the U.S. NOAA storm i

Σκοπός της ανάλυσης είναι η περιγραφή του αντίκτυπου έντονων καιρικών φαινομένων στην υγεία και την οικονομία, και ο έλεγχος της υπόθεσης ότι οι ανεμοστρόβιλοι είναι το καιρικό φαινόμενο που έχει πλήξει περισσότερο τις Η.Π.Α. και το πιο επιζήμιο οικονομικά. Τα αποτελέσματα της συγκεκριμένης ανάλυσης είναι τα εξής:

- Οι ανεμοστρόβιλοι είναι το καιρικό φαινόμενο, που προκάλεσε τους περισσότερους θανάτους και τραυματισμούς.
- Η πλημμύρα έχει προκαλέσει τις μεγαλύτερες οικονομικές καταστροφές.

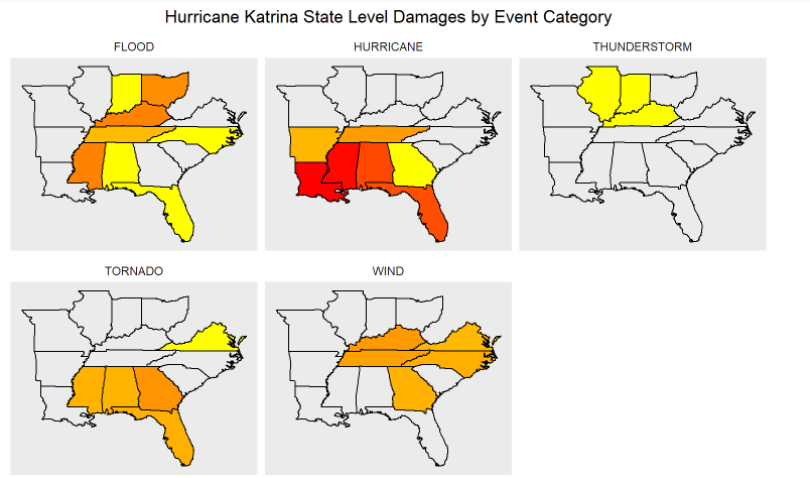


Σχήμα 1: Top 7 events that have the greatest economic consequences.

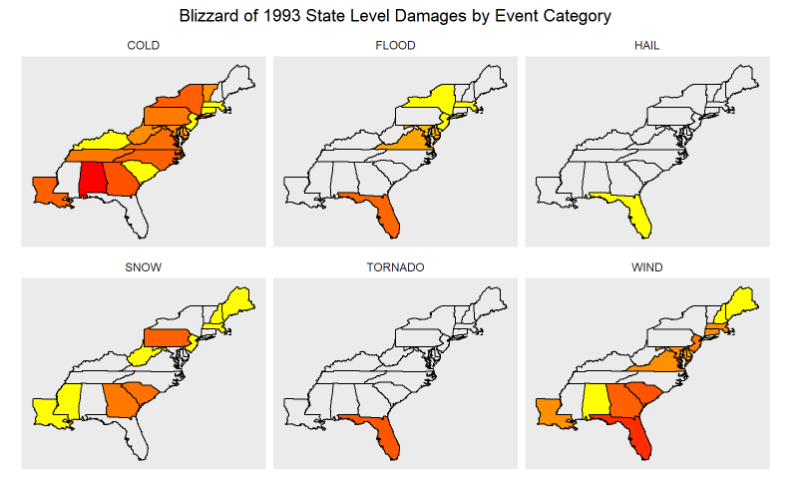


Σχήμα 2: Top 7 most harmful types of events.

Σκοπός και αυτής της ανάλυσης είναι η μελέτη του αριθμού θυμάτων και του μεγέθους της καταστροφής περιουσιών. Στην ανάλυση αυτή συγχωνεύθηκαν καιρικά φαινόμενα, τα οποία είναι όμοια μεταξύ τους, ενώ άλλα φαινόμενα αφαιρέθηκαν τελείως. Έτσι, έμειναν συνολικά έξι φαινόμενα. Ωστόσο, η ανάλυση λαμβάνει χώρα με βάση τα πιο 'καταστροφικά' φαινόμενα ανά κατηγορία.



Σχήμα 3: Hurricane Katrina state level damages by event category.



Σχήμα 4: Blizzard of 1993 state level damages by event category.

Google Cloud Sample queries i

Η Google προσφέρει κάποια queries, που απαντάνε στις παρακάτω ερωτήσεις.

- Ποιο είδος κακοκαιρίας ήταν πιο συχνό την περίοδο 1950-2000;
- Ποιοι ταχυδρομικοί κώδικες έχουν βιώσει τις περισσότερες καταιγίδες τα τελευταία 10 χρόνια;
- Ποιες καταιγίδες, που συνέβησαν τα τελευταία 15 χρόνια, προκάλεσαν την μεγαλύτερη ζημιά στις περιουσίες των ανθρώπων;

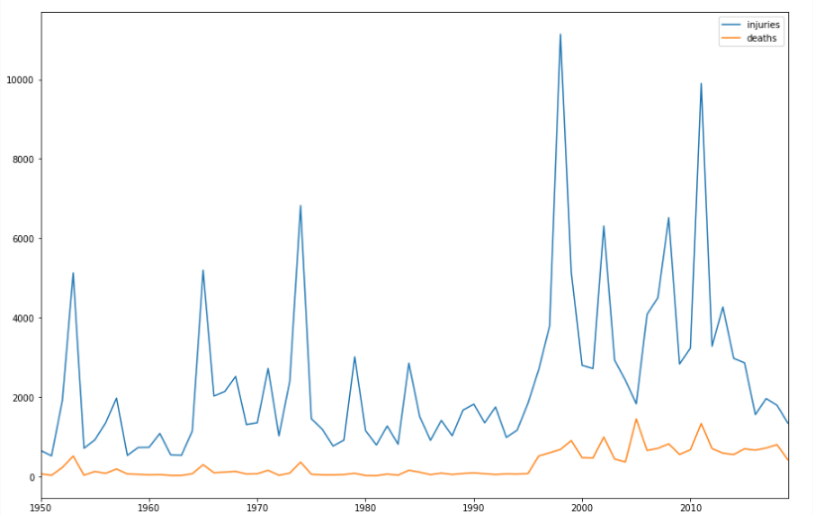
Παρακάτω βλέπουμε ορισμένες απαντήσεις.

event_type	count_storms
thunderstorm wind	169873
hail	134565
tornado	44202
heavy snow	16045
flash flood	15186

Πίνακας 1: Τα πιο συχνά είδη κακοκαιρίας την περίοδο 1950-2000.

ΕΞΕΡΕΥΝΗΣΗ ΔΕΔΟΜΕΝΩΝ

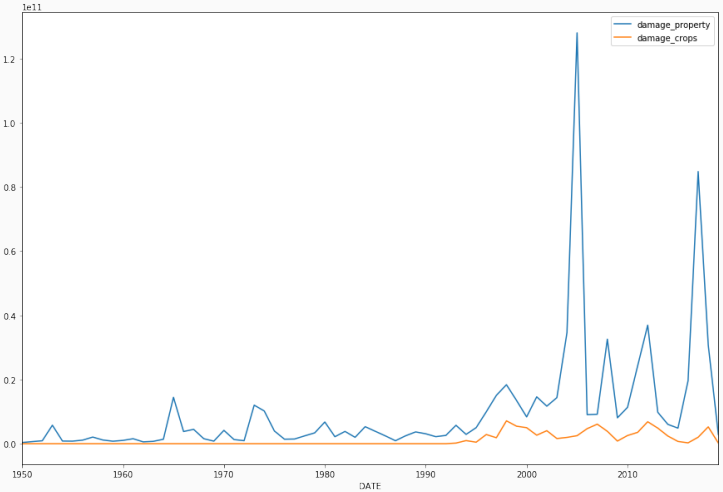
Η εξέλιξη του μεγέθους των καταστροφών στο χρόνο i



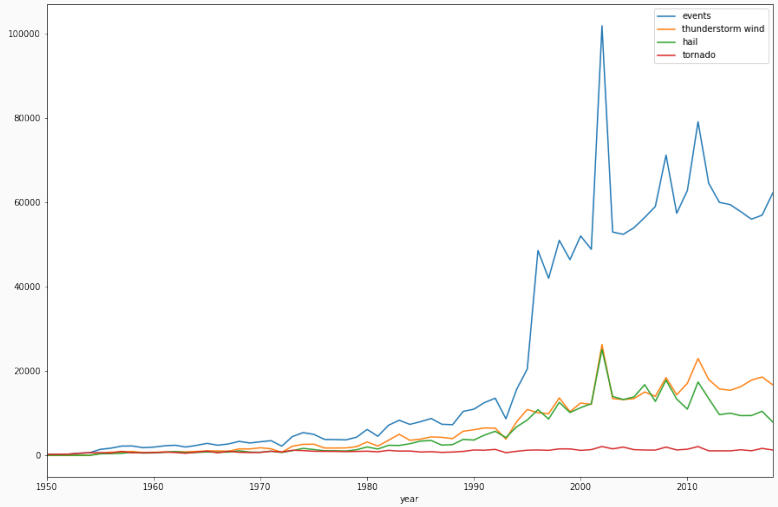
Σχήμα 5: Διάγραμμα χρόνου - τραυματισμών/θανάτων.

- Όσο μεγάλη και να είναι μια καταστροφή, οι απώλειες δεν ξεφεύγουν ιδιαίτερα.
- Οι τραυματισμοί έχουν διάφορα peaks στο χρόνο, γεγονός που μπορεί να αιτιολογηθεί από μεγάλες κακοκαιρίες εκείνη την περίοδο.
- Αν και παρατηρείται αυτή η κυματομορφή, τα ‘κάτω’ και ‘πάνω’ άκρα κινούνται αυξητικά με την πάροδο του χρόνου.

Η εξέλιξη του μεγέθους των καταστροφών στο χρόνο ii



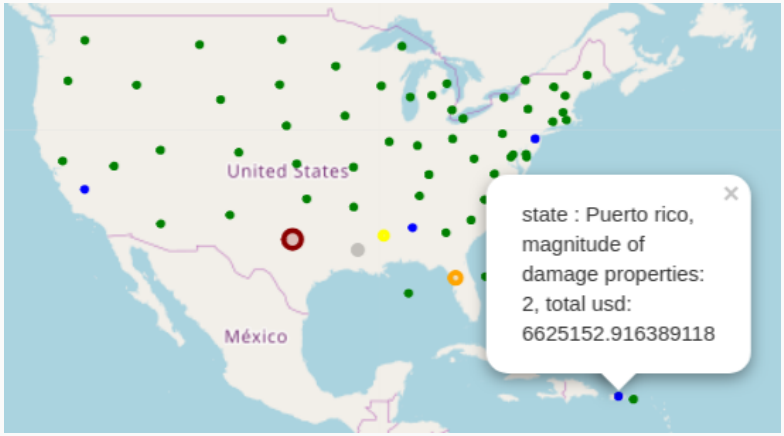
Σχήμα 6: Διάγραμμα ζημιών σε καλλιεργήσιμες εκτάσεις/ιδιωτικές περιουσίες - χρόνου.



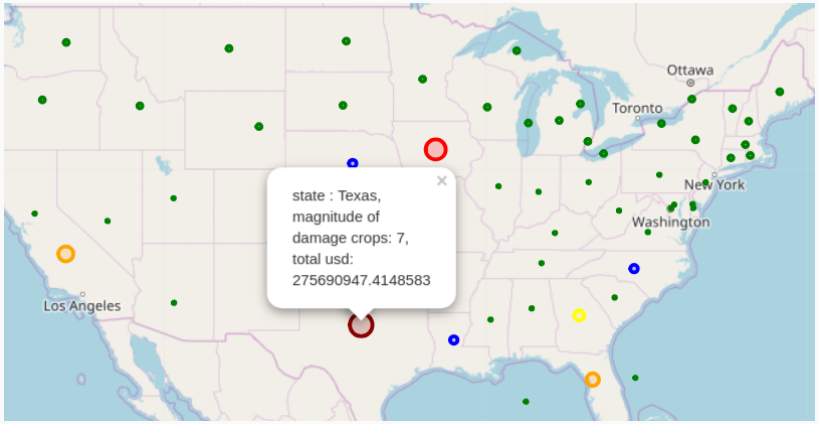
Σχήμα 7: Διάγραμμα αριθμού φαινομένων - χρόνου.

- Με την πάροδο του χρόνου, οι ζημιές σε καλλιεργήσιμες εκτάσεις είναι σχετικά σταθερές, ωστόσο δεν ισχύει το ίδιο για ιδιωτικές περιουσίες ανθρώπων. Από το 2000 και μετά, είχαμε κάποιες πολύ μεγάλες καταστροφές.
- Τα tornado δεν αυξάνονται με την πάροδο του χρόνου. Όμως, άλλα είδη κακοκαιρίας όπως hail ή thunderstorm wind ανεβαίνουν, ειδικά από το 1990 και μετά. Όλες οι κακοκαιρίες μαζί έχουν δραματική άνοδο από ένα σημείο και μετά.

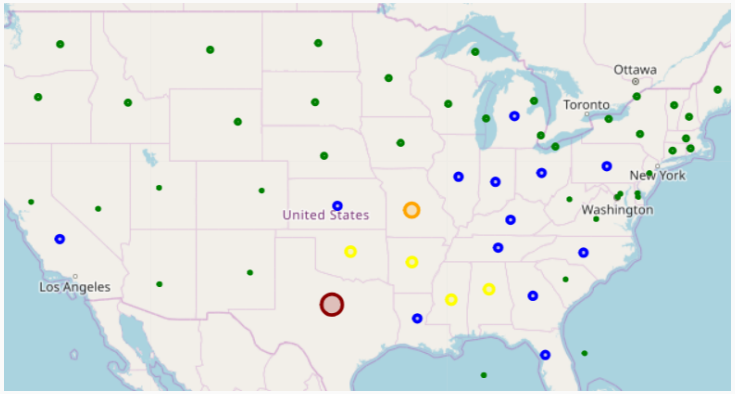
Γεωγραφική κατανομή των καταστροφών, κατά είδος και μέγεθος i



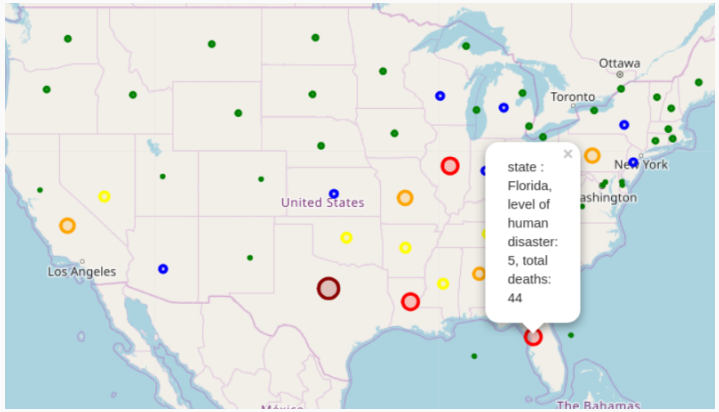
Σχήμα 8: Γεωγραφική κατανομή των καταστροφών, κατά είδος και μέγεθος.



Σχήμα 9: Γεωγραφικός χάρτης με τις ζημιές των καλλιεργήσιμων εκτάσεων.

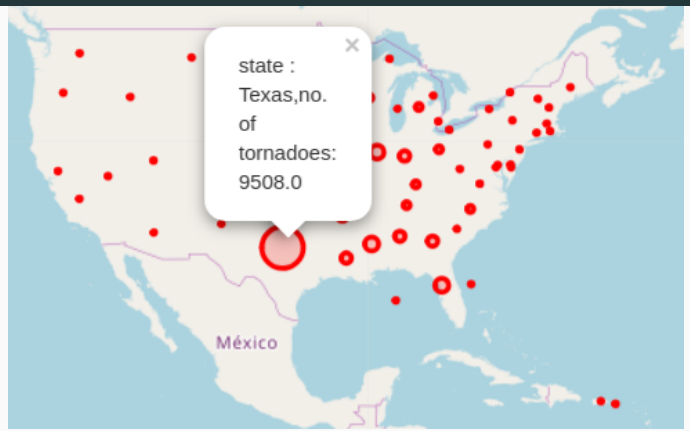


Σχήμα 10: Γεωγραφικός χάρτης με θανάτους ανά πολιτεία.

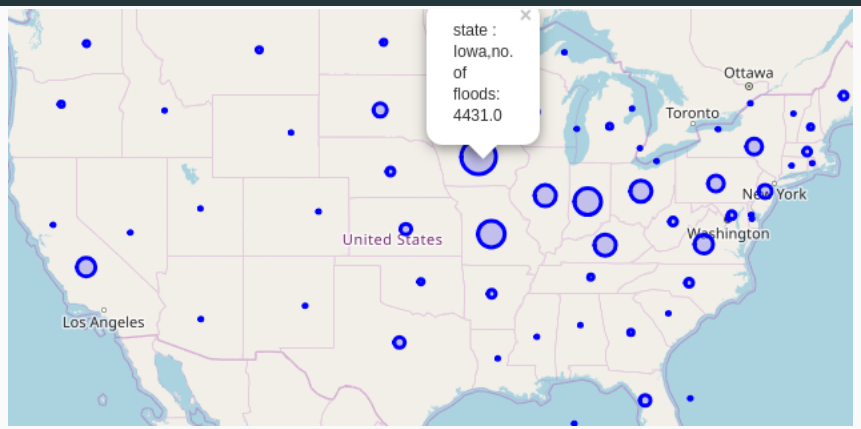


Σχήμα 11: Γεωγραφικός χάρτης με τραυματισμούς ανά πολιτεία.

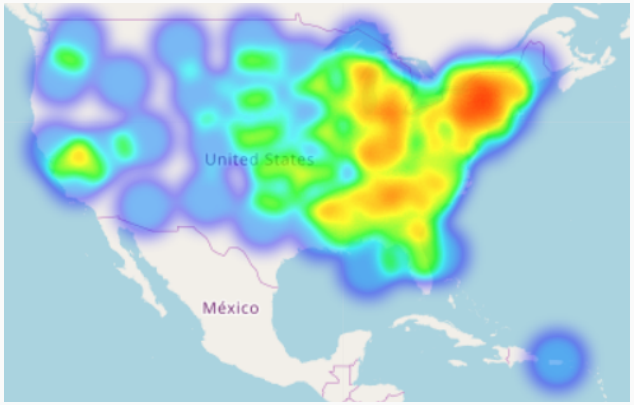
Γεωγραφική κατανομή των καταστροφών, κατά είδος και μέγεθος ii



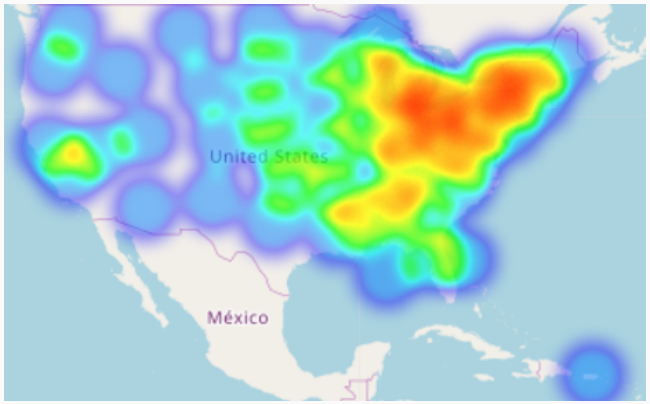
Σχήμα 12: Γεωγραφικός χάρτης με τον αριθμό των tornado κάθε πολιτείας.



Σχήμα 13: Γεωγραφικός χάρτης με τον αριθμό των flood events κάθε πολιτείας.

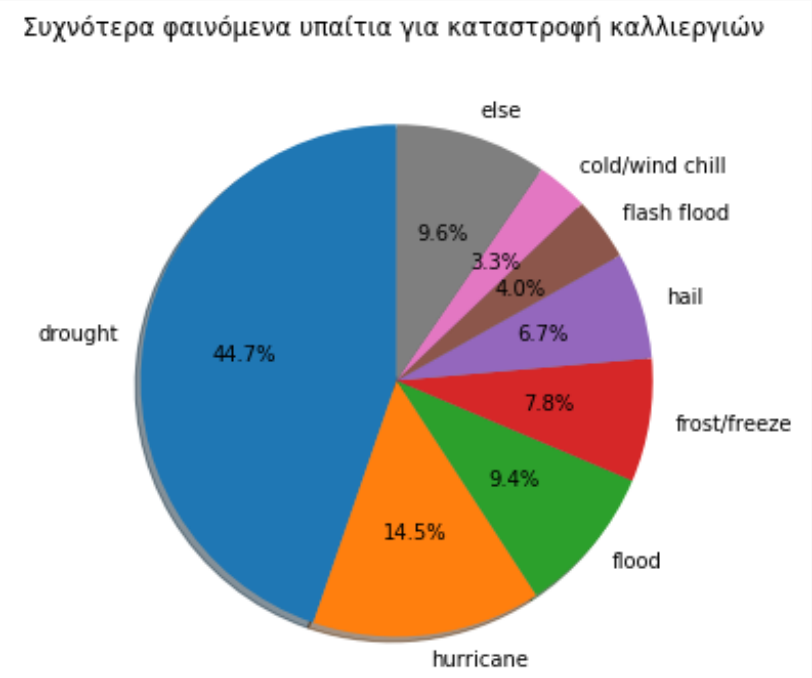


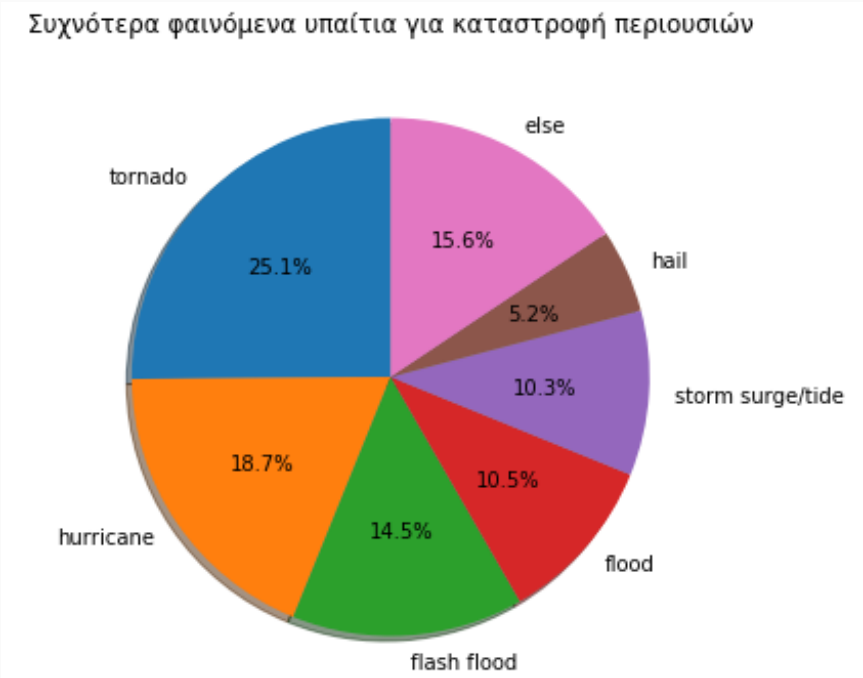
Σχήμα 14: Heatmap για κάθε πολιτεία και ζημιά περιουσιών.



Σχήμα 15: Heatmap για κάθε πολιτεία και ζημιά καλλιεργήσιμης έκτασης.

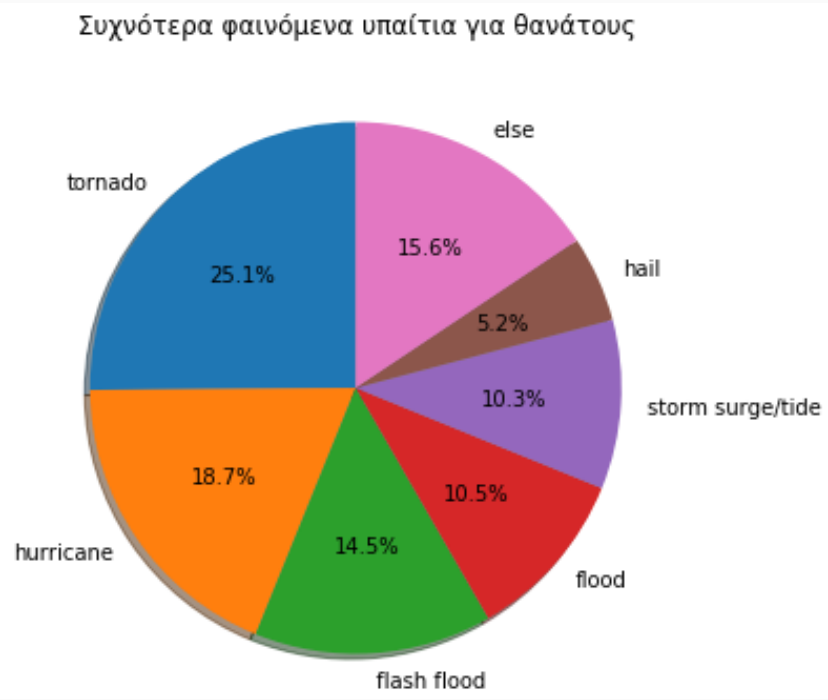
Σχέση μεγέθους–τύπου καταστροφών i

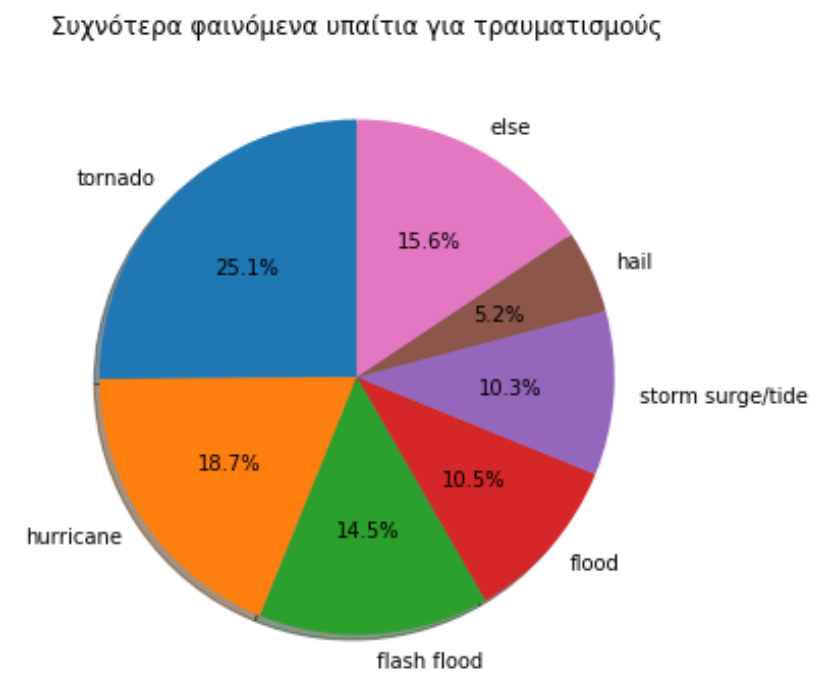




- Η ξηρασία είναι με μεγάλη διαφορά (44.7%) η πρώτη αιτία καταστροφής των καλλιεργιών. Ακολουθούν οι τυφώνες (14.5%).
- Η καταστροφή περιουσιών παρουσιάζει μικρότερες διαφορές. Η πρώτη αιτία είναι οι ανεμοστρόβιλοι (25.1%), στη συνέχεια έχουμε τους τυφώνες (18.7%).

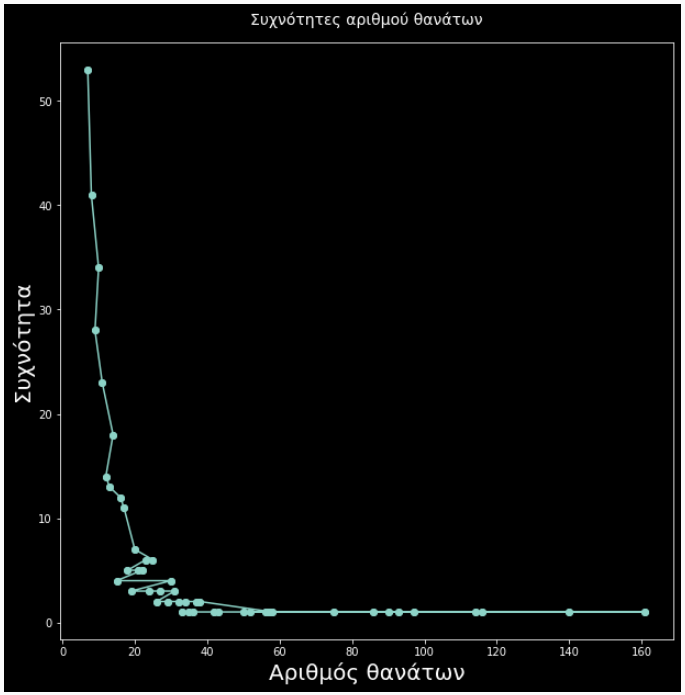
Σχέση μεγέθους–τύπου καταστροφών ii





- Οι ανεμοστρόβιλοι, οι τυφώνες, καθώς και οι ξαφνικές πλημμύρες σε παράκτιες περιοχές ευθύνονται για παραπάνω από το 50% του συνόλου των θανάτων και των τραυματισμών.

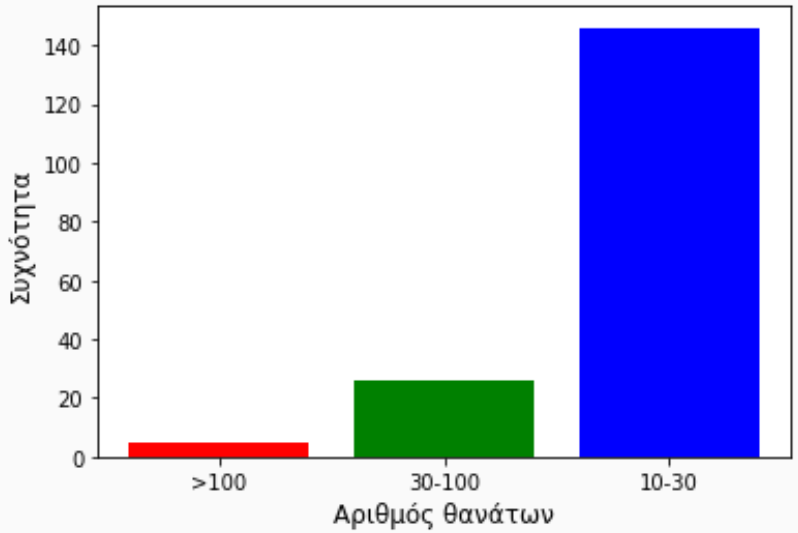
Σχέση μεγέθους–τύπου καταστροφών iii

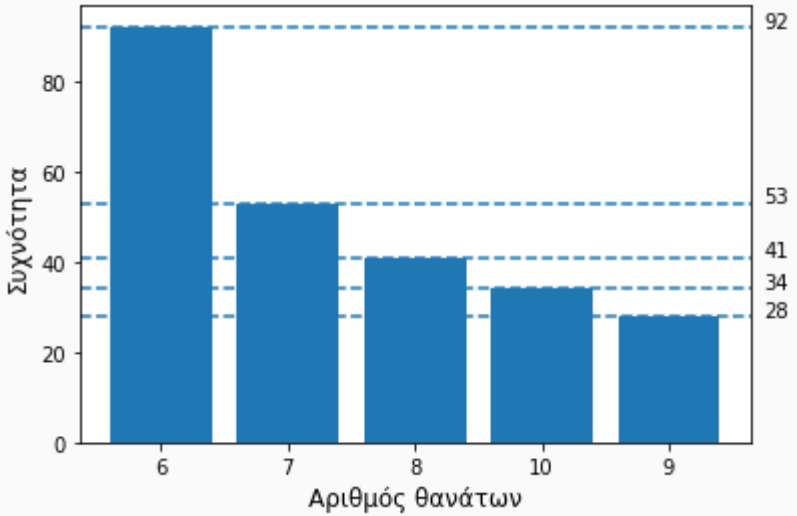


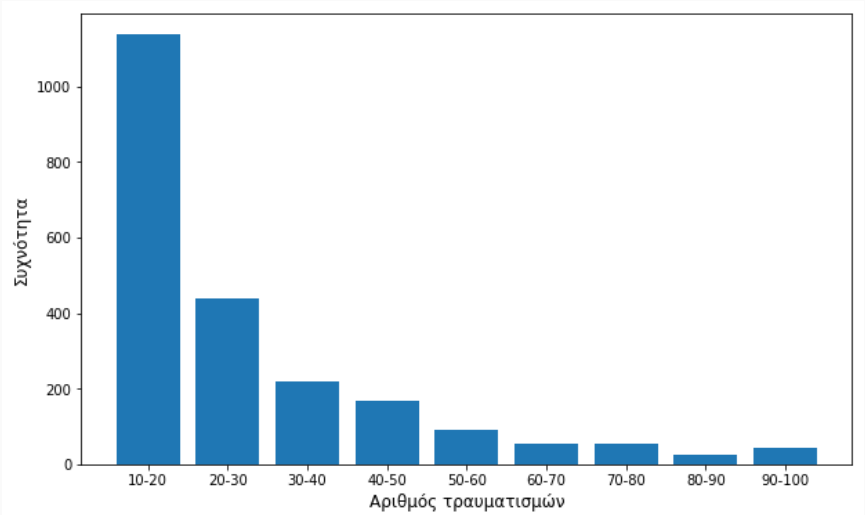
Σχήμα 16: Αριθμός θανάτων που προκαλλούνται από ένα φαινόμενο και οι αντίστοιχες συχνότητες.

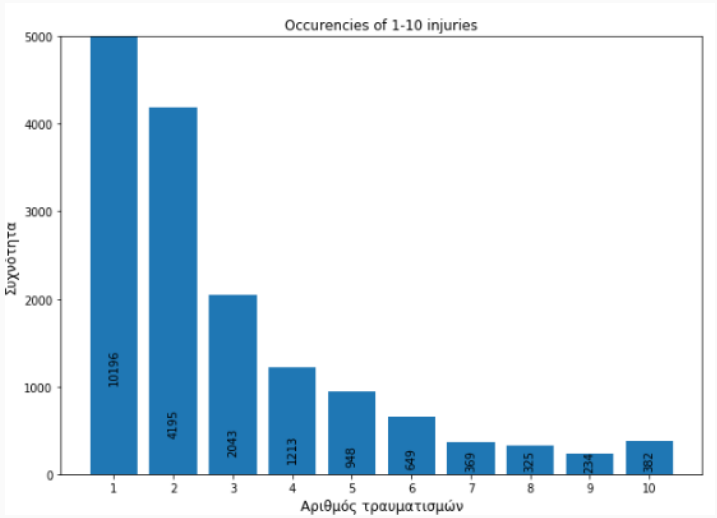
- Το παραπάτω διάγραμμα δημιουργεί μια υπερβολική καμπύλη που δείχνει πως μικρό πλήθος θανάτων (<10) έχει πολύ υψηλή συχνότητα, ενώ μεγάλο πλήθος θανάτων χαμηλή συχνότητα.

Σχέση μεγέθους–τύπου καταστροφών iv









ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΚΑΙ ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ

Association Rules - Itemsets i

Μελετάμε την ύπαρξη στοιχειοσυνόλων και κανόνων συσχέτισης. Τα στοιχειοσύνολα ορίζονται ως συλλογές k-πλήθους αντικειμένων.

event_type	drought	fire	flood	hail	hurricane	rain	snow	storm	thunderstorm	tide	tornado	tsunami	wind
episode_id													
2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
7.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0
8.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
10.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0

Χρησιμοποιούμε τον αλγόριθμο a-priori για minimum support=0.02 δημιουργούνται οι εξής κανόνες:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(tornado)	(thunderstorm wind)	0.049267	0.355667	0.020297	0.411977	1.15832	0.002774	1.095760
1	(thunderstorm wind)	(tornado)	0.355667	0.049267	0.020297	0.057067	1.15832	0.002774	1.008272

Association Rules - Itemsets ii

- Οι τιμές του support και του confidence είναι μικρές, και
- οι lift τιμές κοντά στο 1, που σημαίνει ότι δεν υπάρχουν ισχυρές συσχέτισεις.

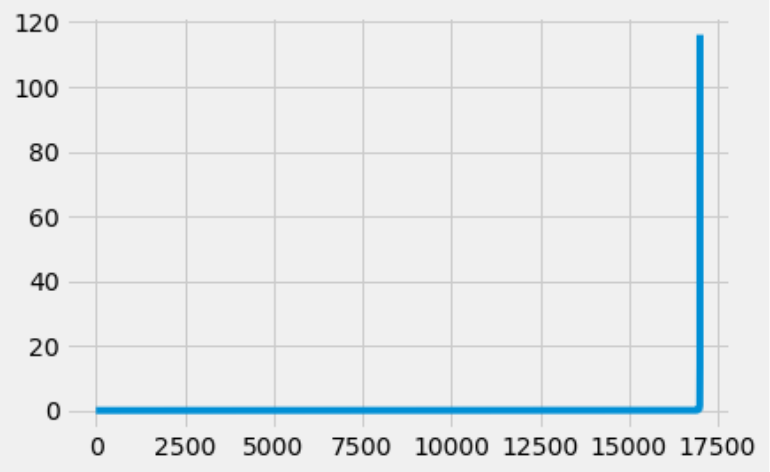
Με τον αλγόριθμο fp-growth, για min_support=0.05, λαμβάνουμε τα εξής στοιχειοσύνολα:

	support	itemsets
0	0.355667	(thunderstorm wind)
1	0.285685	(hail)
2	0.107548	(flash flood)

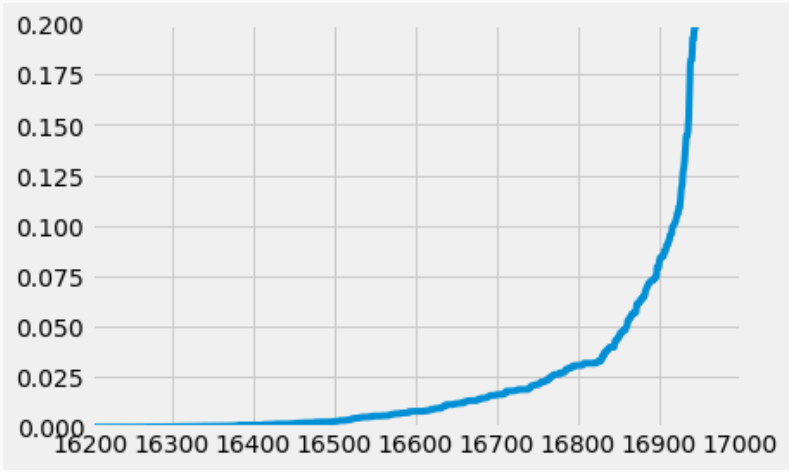
- Παρατηρούμε πως οι αλγόριθμοι a-priori και fp-growth βρίσκουν ότι τα πιο δημοφιλή καιρικά φαινόμενα είναι τα thunderstorm wind και hail.

DBSCAN Clustering i

Μελετάμε αν τα δεδομένα συσταδοποιούνται, όταν λαμβάνουμε υπόψιν μόνο τα χαρακτηριστικά που αφορούν τις καταστροφές.



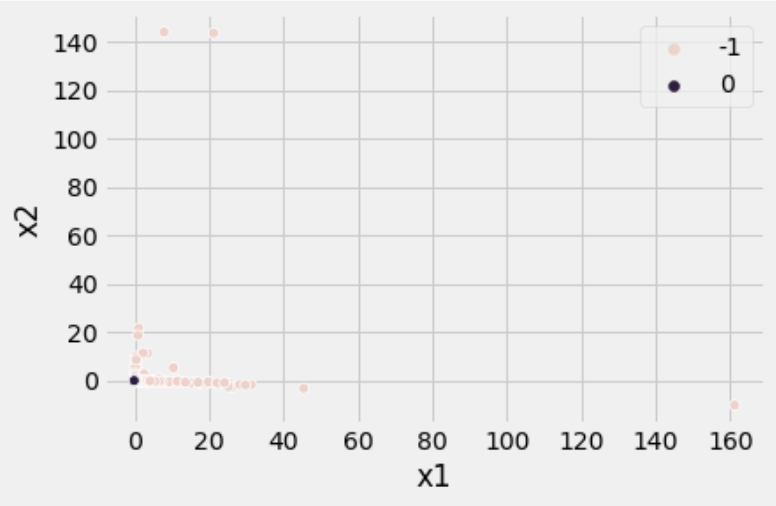
Σχήμα 17: Αποστάσεις σημείων από τα 4 πλησιέστερα.



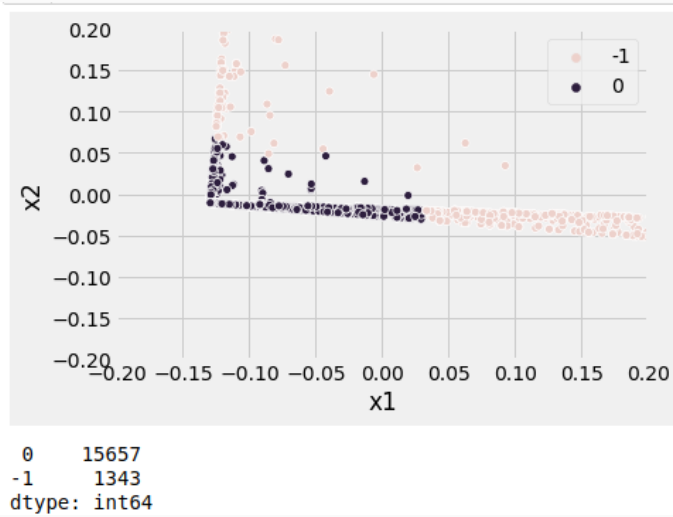
Σχήμα 18: Αποστάσεις σημείων από τα 4 πλησιέστερα – zoomed.

- Εκτιμούμε την τιμή epsilon του DBSCAN αλγόριθμου να είναι περίπου .04, παίρνοντας ένα τυχαίο δείγμα μεγέθους 17.000.
- Εφαρμόζουμε τώρα το DBSCAN επιλέγοντας για min points 500 δεδομένου του μεγέθους του δείγματος που πήραμε.

DBSCAN Clustering ii



Σχήμα 19: Clusters.



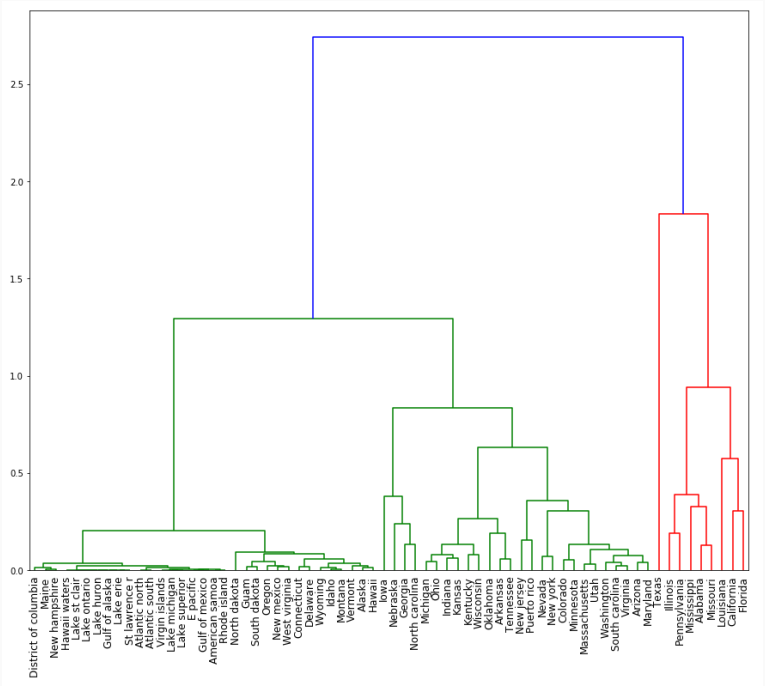

```
 0    15657
-1     1343
dtype: int64
```

Σχήμα 20: Clusters zoomed.

- Δημιουργούνται 2 clusters. Το ένα (μαύρες κουκίδες) σχηματίζεται στις χαμηλές τιμές των αξόνων και περιέχει την πλειοψηφία των σημείων (95%). Η διασπορά είναι μικρή και τα σημεία πολύ πυκνά στο χώρο. Το άλλο (ροζ κουκίδες) περιέχει μόλις το 5% των σημείων. Η διασπορά είναι μεγάλη και τα σημεία αραιά στο χώρο.
- Silhouette Score = 0.85.

Hierarchical Clustering i

Βρίσκουμε πόσες συστάδες θα δημιουργηθούν, αν έχουμε σα μέτρο ομοιότητας τη ζημιά (damage_property, damage_crops, injuries, deaths) για κάθε state της Αμερικής.

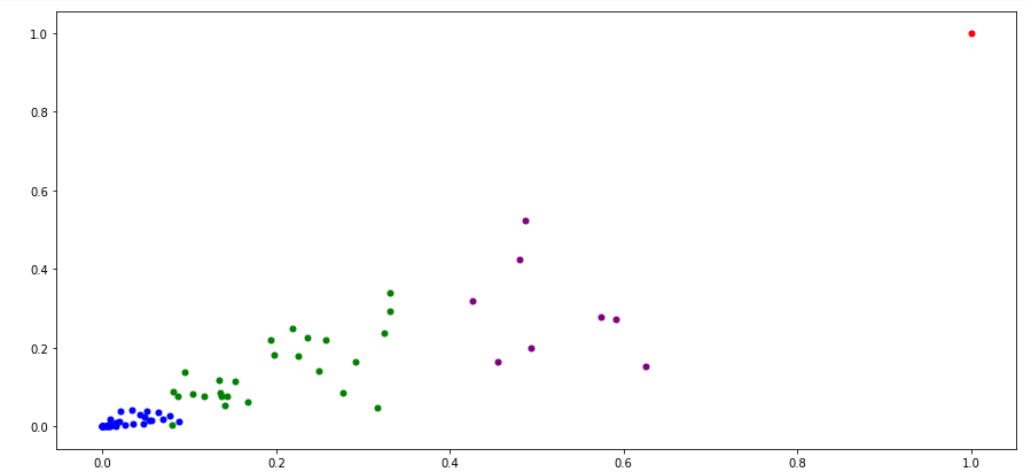


Σχήμα 21: Δενδρόγραμμα.

- Για τη δημιουργία του δενδρογράμματος, χρησιμοποιούμε τη μέθοδο Ward.
- Επιλέγουμε με βάση αυτό τον αριθμό των συστάδων: n_clusters = 4.

Hierarchical Clustering ii

Δημιουργούμε το παρακάτω διάγραμμα διασποράς για n_clusters = 4.

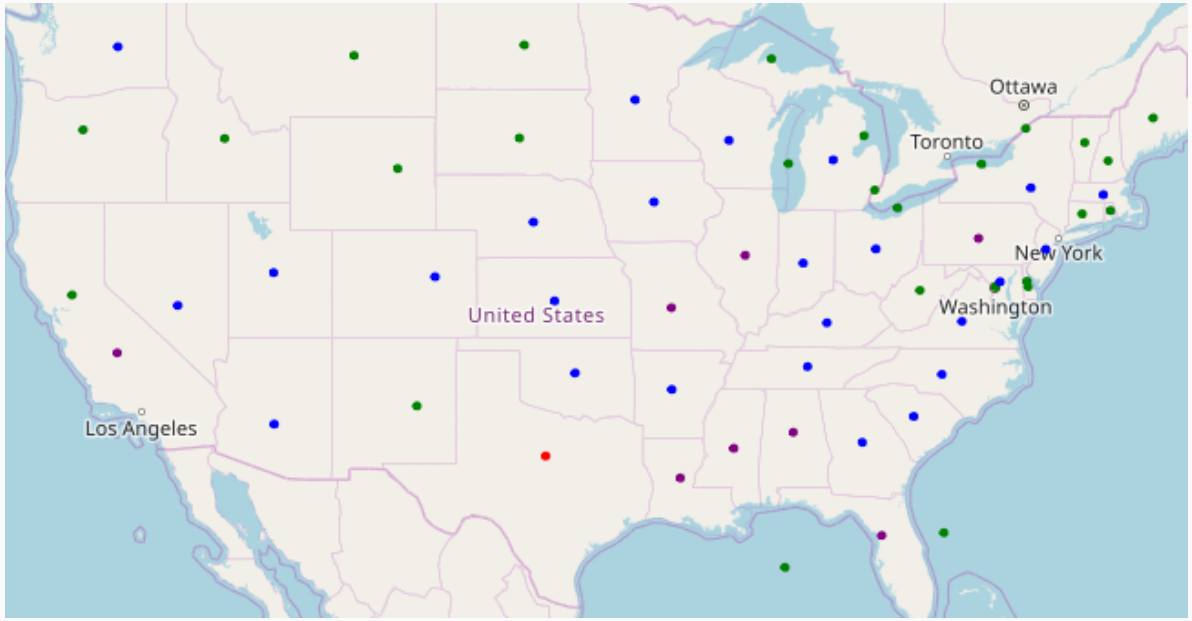



```
Out[32]: state                Texas
         event_longitude   -98.7248
         event_latitude     32.2986
         damage_property          1
         damage_crops             1
         deaths                   1
         injuries                 1
         Name: 59, dtype: object
```

Σχήμα 22: Διάγραμμα διασποράς για n_clusters = 4 και Τέξας (κόκκινη κουκίδα) - χαρακτηριστικά.

- Η κόκκινη κουκίδα του διαγράμματος διασποράς αντιστοιχίζεται στην πολιτεία του Τέξας.
- Το Τέξας παρουσιάζει τη μεγαλύτερη ζημιά και τους περισσότερους θανάτους και τραυματισμούς (τα αποτελέσματα είναι κανονικοποιημένα από 0 έως 1).

Hierarchical Clustering iii



Σχήμα 23: Χάρτης αντιστοίχισης πολιτειών - συστάδων.

- Ο παραπάνω χάρτης, ανάλογα με το χρώμα, δείχνει σε ποια συστάδα ανήκει η κάθε πολιτεία και τι μέγεθος καταστροφής είχε σε σχέση με τις υπόλοιπες.
- Οι πολιτείες στη μεριά του Ατλαντικού είναι πιο επιρρεπείς στις μεγάλες ζημιές και στις ανθρώπινες απώλειες.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Συμπεράσματα

- Από τα διαγράμματα καταστροφών-χρόνου είδαμε ότι υπάρχει μια αυξητική τάση στον αριθμό των καταστροφών.
- Όσον αφορά τη γεωγραφία, τα έντονα καιρικά φαινόμενα έχουν λάβει χώρα περισσότερο στο ανατολικό παρά στο δυτικό τμήμα των ΗΠΑ.
- Το Τέξας είναι η πολιτεία που έχει πληγεί περισσότερο από ανεμοστρόβιλους και η Καλιφόρνια από πλημμύρες.
- Αν δούμε συνολικά τις καταστροφές, το 60% έχει προκληθεί από ανεμοστρόβιλους και χαλάζι, ενώ οι τυφώνες, οι πλημμύρες, το χαλάζι και οι ξηρασίες σχετίζονται πιο άμεσα με κάποιο είδος καταστροφής ξεχωριστά.
- Η πλειοψηφία των τραυματισμών και των θανάτων κυμένεται στο εύρος 0-10, ενώ η πιο θανατηφόρα καταστροφή είχε 638 θύματα.
- Οι αλγόριθμοι για συσχέτιση χαρακτηριστικών και εξαγωγή κανόνων έδειξαν ότι δεν υπάρχει κάποια συσχέτιση μεταξύ των δεδομένων.
- Βρέθηκαν 2 clusters με τη μέθοδο DBSCAN, ένα με το 95% των δεδομένων και μικρή διασπορά, κι ένα με το υπόλοιπο 5% και πολύ μεγάλη διασπορά. Το DBSCAN αξιολογήθηκε με το Silhouette Score, με σκορ 0.85.
- Βρέθηκαν 4 clusters με τη μέθοδο της ιεραρχικής συσταδοποίησης, και συγκεκριμένα με τη μέθοδος Ward. Η πολιτεία του Τεξας παρουσιάζει τις μεγαλύτερες απώλειες, και οι πολιτείες που βρίσκονται στη μεριά του Ατλαντικού, είναι πιο επιρρεπείς στις ζημιές και στις ανθρώπινες απώλειες.

Τέλος ☺