ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΕΞΟΡΥΞΗ ΓΝΩΣΗΣ ΑΠΟ ΔΕΔΟΜΕΝΑ

ΔΗΜΗΤΡΗΣ ΖΕΡΚΕΛΙΔΗΣ 03400049
ΜΑΡΙΑ ΚΑΪΚΤΖΟΓΛΟΥ 03400052
ΜΑΡΙΑ-ΦΙΛΙΠΠΑ ΤΡΙΒΥΖΑ 03400080

Μάρτιος 2020

ΠΕΡΙΕΧΟΜΕΝΑ

1. ΕΙΣΑΓΩΓΗ

1.1 Αντικείμενο έρευνας

Η εργασία σχετίζεται με τα έντονα καιρικά φαινόμενα στις ΗΠΑ από το 1950 ως 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, για τα οποία προσπαθούμε να εξάγουμε πληροφορίες:

- 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.

1.2 Δεδομένα

Χρησιμοποιήσαμε το Google Cloud για την εύρεση dataset, και εκεί βρήκαμε το "Severe Storm Event Details" dataset, με τελευταία ενημέρωση στις 8/27/2019. Η πηγή του dataset είναι το National Oceanic Atmospheric Administration (NOAA) και ανήκει στην ενσωματωμένη βάση δεδομένων Severe Weather Data Inventory (SWDI). Πρόκειται για μια βάση δεδομένων για καιρικά φαινόμενα μεγάλης δριμύτητας. Εκεί καταγράφονται πληροφορίες από το 1950 μέχρι και το 2019. Οι πληροφορίες, σε γενικές γραμμές, αφορούν το είδος του φαινομένου (π.χ. τυφώνας, πλημμύρα), την τοποθεσία και τον αντίκτυπο σε υλικές ζημιές και στους ανθρώπους. Τα δεδομένα καταγράφουν:

- Την εμφάνιση φαινομένων έντασης ικανής να προκαλέσει απώλειες ζωών, τραυματισμούς, καταστροφή περιουσιακών στοιχείων και διακοπή εμπορίου.
- Σπάνια, ασυνήθιστα καιρικά φαινόμενα, όπως, για παράδειγμα, χιονοθύελλες στη Florida ή στο παραθαλάσσιο τμήμα του San Diego.
- Άλλα σημαντικά μετεωρολογικά γεγονότα, όπως μέγιστες και ελάχιστες θερμοκρασίες, ή βροχοπτώσεις που λαμβάνουν χώρα σε συνδυασμό με άλλο καιρικό φαινόμενο.

Το dataset, το οποίο το αποθηκεύσαμε σε csv, περιείχε αρχικά 34 στήλες και 1639004 γραμμές.

1.3 Διαδικασία και μέθοδοι

Χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python και τα δεδομένα διαβάστηκαν σε Data Frame. Η διαδικασία με την οποία δομήσαμε την ανάλυσή μας αποτελείται από:

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

Βάσει των ερωτημάτων που θέσαμε, επιλέξαμε αντίστοιχες μεθόδους και τεχνικές, οι οποίες παρουσιάζονται στη συνέχεια.

1.3.1 Προεπεξεργασία δεδομένων

Ένα πολύ σημαντικό σκέλος στην εξόρυξη δεδομένων είναι η προεπεξεργασία των δεδομένων. Παρατηρήσαμε ότι 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. Το κυριότερο μέρος της προεπεξεργασίας, έγινε σε ένα ξεχωριστό notebook, το οποίο παράγει ένα csv αρχείο με τα επεξεργασμένα δεδομένα.

Στη συνέχεια, θα περιγράψουμε κάποιες από τις εξής επεξεργασίες:

- Missing values: διαγραφή σειρών, συμπλήρωση σειρών με μέσους όρους ή 0
- Ενοποίηση χαρακτηριστικών
- Διόρθωση ασυνεπών τιμών (π.χ. πληθωρισμός)
- Ομαδοποίηση τιμών χαρακτηριστικών
- Διαγραφή χαρακτηριστικών
- Δημιουργία νέων χαρακτηριστικών
- Μετασχηματισμός χαρακτηριστικών

1.3.2 Εξερεύνηση δεδομένων

Οι γραφικές τεχνικές είναι ένα από τα πιο βασικά εργαλεία της εξερεύνησης δεδομένων. Οι εικονικές αναπαραστάσεις βοηθούν πολύ στον εντοπισμό συχσετίσεων μεταξύ των χαρακτηριστικών και στην αναγνώριση μοτίβων. Η χρήση τους περιορίζεται αυστηρά στην παρατήρηση των συσχετισμών και δεν επεκτείνεται στην εύρεση των πιθανών αιτιατών μεταξύ των (το οποίο δεν είναι και σκοπός της εξερεύνησης δεδομένων).

Τα περιγραφικά στατιστικά, όπως τα μέτρα κεντρικής τάσης, προσφέρουν, επίσης, μια εικόνα για την κατανομή και τις τάξεις μεγέθους των χαρακτηριστικών. Ο συνδυασμός της εξαγωγής τέτοιων μέτρων με την οπτικοποίησή τους σε διαγράμματα είναι ακόμα πιο αποτελεσματικός ως προς το σκοπό της (διαισθητικής) κατανόησης των δεδομένων.

1.3.3 Εξόρυξη δεδομένων

Βασικοί στόχοι της εξόρυξης δεδομένων, τους οποίους και επιδιώξαμε, είναι η ανακάλυψη μοτίβων και η ταξινόμηση των δεδομένων. Επεκτείνεται και στην ανάλυση άτυπων/απόμακρων σημείων, στην πρόβλεψη κλπ, ωστόσο δεν ασχοληθήκαμε με κάτι τέτοιο στην εργασία αυτή. Επιλέξαμε τρεις μεθόδους για την ανακάλυψη μοτρίβων ενδιαφέροντος. Συγκεκριμένα,

- Clustering
- Itemsets

Ακολούθησε μετεπεξεργασία (post-processing) με απεικόνιση, ερμηνεία και αξιολόγιση των αποτελεσμάτων που εξήχθησαν. Στη συνέχεια, θα σχολιάσουμε πιο αναλύτικα πως εφαρμόστηκαν, ερμηνεύθηκαν και αξιολογήθηκαν αυτές οι μέθοδοι.

1.3.4 Παρουσίαση αποτελεσμάτων εξόρυξης

Χρησιμοποιήσαμε κατά βάση γραφικές τεχνικές για την παρουσίαση των αποτελεσμάτων, συνοδευόμενες από σχολιασμούς.

2. ΠΡΟΣΕΓΓΙΣΕΙΣ ΣΕ ΣΧΕΤΙΚΑ ΠΡΟΒΛΗΜΑΤΑ

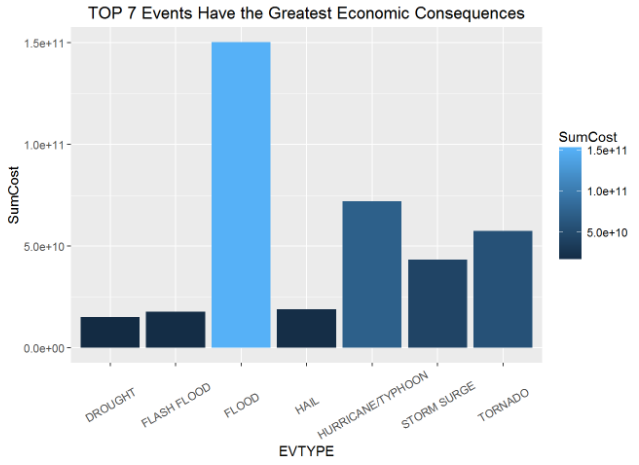
Σε αυτήν την ενότητα παρουσιάζουμε πώς προσέγγιστηκαν δύο παρόμοια ζητήματα από άλλες ομάδες/άτομα, καθώς και κάποια queries της Google πάνω στο dataset που χρησιμοποιούμε.

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

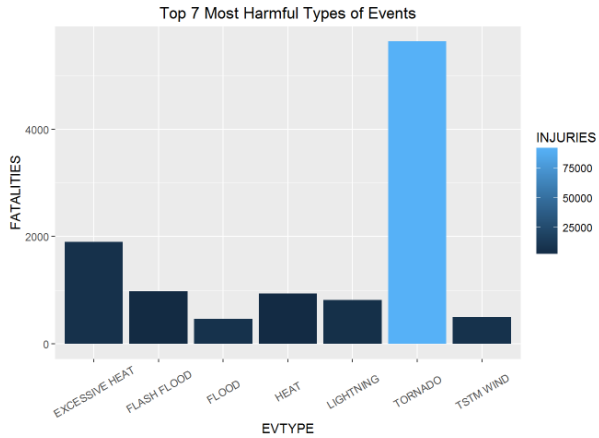
Η ανάλυση αυτή είναι δημοσιευμένη στην ιστοσελίδα rpubs.com. Σκοπός της είναι η περιγραφή του αντίκτυπου έντονων καιρικών φαινομένων στην υγεία και την οικονομία, και ο έλεγχος της υπόθεσης ότι οι ανεμοστρόβιλοι (tornadoes) είναι το καιρικό φαινόμενο που έχει πλήξει περισσότερο τις Ηνωμένες Πολιτείες Αμερικής και το πιο επιζήμιο οικονομικά. Ο λόγος για τον οποίο ερευνάται η συγκεκριμένη κακοκαιρία, είναι ότι οι Η.Π.Α. είναι περικυκλωμένες από τον Ειρηνικό και Ατλαντικό ωκεανό και αυτό ευνοεί το σχηματισμό ανεμοστρόβιλων. Τα αποτελέσματα της συγκεκριμένης ανάλυσης είναι τα εξής:

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

Για την εξαγωγή των παραπάνω αποτελεσμάτων, η δημοσίευση αναφέρει πως αγνοήθηκαν τα nan values, καθώς δεν επηρέαζαν τα αποτελέσματα. Επίσης, αθροίστηκαν οι μεταβλητές Damage Property και Damage Crops, καθώς αυτές συνοψίζουν την οικονομική καταστροφή που έφερε το κάθε καιρικό φαινόμενο. Μας δίνεται το πρώτο γράφημα. Σημειώνεται ότι, το sumCost δηλώνει το άθροισμα των μεταβλητών που αναφέρθηκαν για την οικονική ύφεση στον άξονα y, ενώ στον οριζόντιο άξονα έχουμε το EVTYPE, που είναι το είδος της κακοκαιρίας. Έπειτα, η ανάλυση αυτή μας προσφέρει το δεύτερο γράφημα, που δείχνει το μέγεθος θανάτων και τραυματισμών του κάθε καιρικού φαινομένου. Στον οριζόντιο άξονα έχουμε πάλι το είδος της κακοκαιρίας, ενώ στον κάθετο άξονα τον αριθμό θανάτων. Παράλληλα, το χρώμα δείχνει το μέγεθος των τραυματισμών.



Εικόνα 2.1: Top 7 events that have the greatest economic consequences.

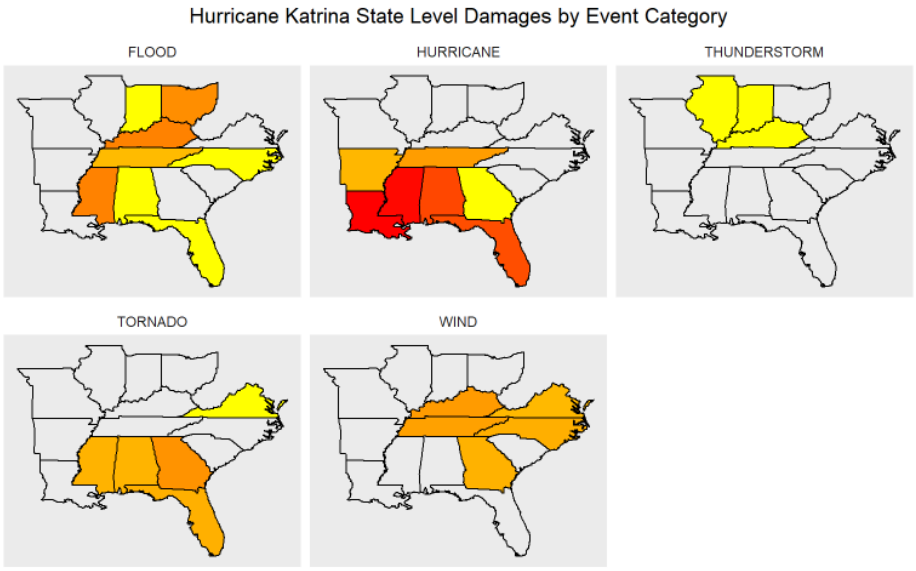


Εικόνα 2.2: Top 7 most harmful types of events.

2.2 Health and Economic Outcomes of Severe Weather Events

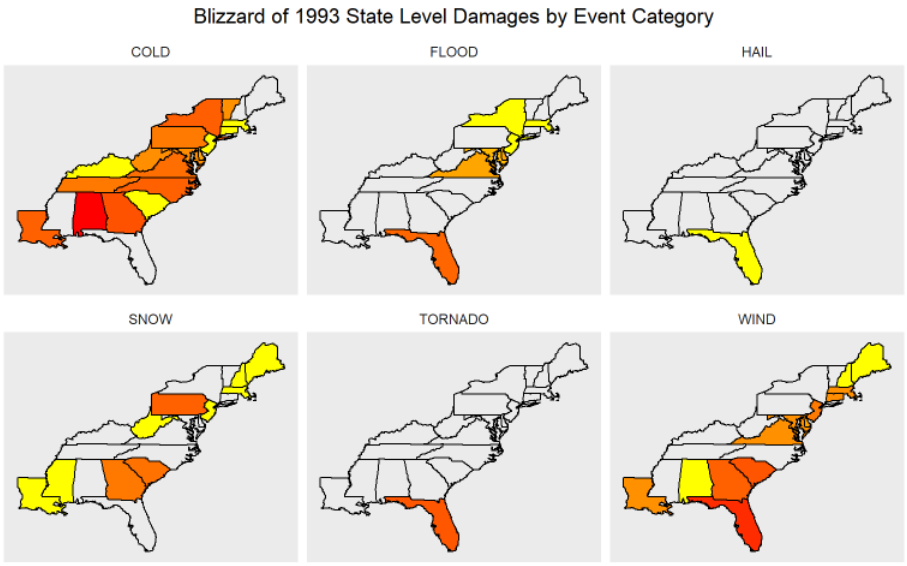
Σκοπός και αυτής της ανάλυσης είναι η μελέτη του αριθμού θυμάτων και του μεγέθους της καταστροφής περιουσιών. Στην ανάλυση αυτή έχουμε μια διαφορετική προσέγγιση του ζητήματος, καθώς συγχωνεύθηκαν καιρικά φαινόμενα, τα οποία είναι όμοια μεταξύ τους, όπως για παράδειγμα heavy rain flooding και flooding from surge, τα οποία μπορούν να ενωθούν στην κατηγορία flood, ενώ άλλα φαινόμενα αφαιρέθηκαν τελείως. Έτσι, έμειναν συνολικά έξι φαινόμενα. Ωστόσο, η ανάλυση λαμβάνει χώρα με βάση τα πιο 'καταστροφικά' φαινόμενα ανά κατηγορία.

Έγινε γεωγραφική ανάλυση για τις ζημιές που προκάλεσε ο τυφώνας κατρίνα, που είναι το άθροισμα των μεταβλητών Damage Crops και Damage Property, για κάθε κατηγορία κακοκαιρίας.



Εικόνα 2.3: Hurricane Katrina state level damages by event category.

Ένα παρόμοιο γράφημα έχουμε και για τη χιονοθύελα του 1993 (κόλπος Μεξικού).



Εικόνα 2.4: Blizzard of 1993 state level damages by event category.

2.3 Google Cloud Sample queries

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 την απάντηση στο 1ο ερώτημα, για τα 10 πιο συχνά φαινόμενα στην περίοδο 1950-2000. Παρατηρούμε πως, οι μη περιστρεφόμενοι άνεμοι (thunderstorm wind) έχουν τη μεγαλύτερη συχνότητα.

event_type	count_storms
thunderstorm wind	169873
hail	134565
tornado	44202
heavy snow	16045
flash flood	15186
winter storm	15166
high wind	14537
flood	9218
drought	7932
cold/wind chill	5497

Πίνακας 2.1: Τα πιο συχνά είδη κακοκαιρίας την περίοδο 1950-2000.

Στη συνέχεια, βλέπουμε τους ταχυδρομικούς κώδικες που βίωσαν τις περισσότερες καταστροφές από χιονοθύελλες (hail storms).

city	zip_code
Kelley city, Iowa	50134
Villa Park city, North Tustin CDP, Orange city, California	92869
Linn CDP, Texas	78563
Bushnell village, Nebraska	69128
Omaha city, Nebraska	68105

Πίνακας 2.2: Ταχυδρομικοί κώδικες που βίωσαν τις περισσότερες καταστροφές από χιονοθύελλες.

Τέλος, μας δίνονται οι καταστροφές με τις μεγαλύτερες ζημιές σε ανθρώπινες περιουσίες τα τελευταία 15 χρόνια. Παρουσιάζουμε τις 5 πρώτες.

episode_id	episode_month	counties	states	event_types	damage_property _in_billions
119753	8-2017	FORT BEND, MONTGOMERY, GALVESTON, SAN JACINTO, HARRIS	Texas	flash flood, tornado, tropical storm, funnel cloud	41.1238915
1198567	8-2005	ORLEANS, LOWER ST. BERNARD, LOWER JEFFERSON,UPPER PLAQUEMINES, UPPER LAFOURCHE	Louisiana	storm surge/tide	31.302999523
68471	10-2012	OCEAN, BURLINGTON, ATLANTIC, MERCER, CAMDEN	New jersey	flood, high surf, high wind, coastal flood	24.95909
120357	9-2017	EASTERN INTERIOR, COAMO, COROZAL, CENTRAL INTERIOR, AGUAS BUENAS	Puerto rico	hurricane, flash flood	18.26375
131864	11-2018	NORTHEAST FOOTHILLS/ SACRAMENTO VALLEY	California	wildfire	17.0

Πίνακας 2.3: 5 καταστροφές με τις μεγαλύτερες ζημιές σε ανθρώπινες περιουσίες τα τελευταία 15 χρόνια.

3. ΠΡΟΕΠΕΞΕΡΓΑΣΙΑ ΔΕΔΟΜΕΝΩΝ

3.1 Ενοποίηση και δημιουργία νέων χαρακτηριστικών

Συνδυάσαμε τους άμεσους με τους έμμεσους θανάτους με ένα άθροισμα. Το ίδιο και για τους τραυματισμούς.

- Deaths = indirect deaths + direct deaths
- Injuries = indirect injuries + direct injuries

3.2 CPI Multiplier - Ενσωμάτωση πληθωρισμού στα δεδομένα

Κατεβάσαμε το dataset CPI (Consumer Price Index). Ο δείκτης CPI είναι ένας τρόπος να συγκρίνουμε την αύξηση της τιμής ορισμένων αγαθών στο χρόνο. Αυτό το σύνολο δεδομένων μάς παρέχει το μέσο όρο του CPI για κάθε χρόνια ή μήνα, ανάλογως του τι θα διαλέξουμε. Δημιουργήσαμε τον πολλαπλασιαστή CPI για κάθε μήνα με τον παρακάτω τρόπο:

- CPI_BASE(2019) / CPI_OF_RECORD

Δηλαδή, μετατρέψαμε τις τιμές στη βάση του 2019. Το CPI OF RECORD αντικατοπτρίζει το CPI της συγκεκριμένης εγγραφής, ενώ το CPI BASE 2019 είναι πάντα σταθερό και είναι η βάση μετατροπής μας. Στη συνέχεια, βρήκαμε το μέσο όρο του CPI πολλαπλασιαστή για κάθε χρονιά, και έπειτα πολλαπλασιάσαμε τις στήλες damage_crops και damage_property με αυτόν τον πολλαπλασιαστή, ώστε να γίνει η μετατροπή της βάσης στο 2019.

3.3 Ημερομηνία

Δημιουργήσαμε τη στήλη 'DATE', που περιέχει μόνο το έτος για κάθε εγγραφή. Αυτό το νέο χαρακτηριστικό θα το χρησιμοποιήσουμε για να μελετήσουμε τις ζημιές και τη συχνότητα των καιρικών φαινομένων με την πάροδο των χρόνων.

3.4 Ασυνεπείς τιμές

Στη στήλη 'event types' υπήρχαν καταγραφές του ίδου φαινομένου και με μικρά και με κεφαλαία γράμματα, όπως για παράδειγμα το hail, που ήταν γραμμένο ως 'hail', αλλά και ως 'Hail'. Αυτή η ασυνέπεια είναι πρόβλημα για την case-sensitive Python, οπότε διορθώσαμε αυτό το πρόβλημα κάνοντας όλα τα stirngs, lowercase.

Το χαρακτηριστικό tor_f_scale παρουσίασε και αυτό ασυνέπεια στα δεδομένα του, καθώς οι τιμές EF0,EF1...EF5 εμφανίζονταν και ως F0,F1..,F5. Μετατρέψαμε όλες τις δεύτερες καταχωρίσεις όπως τις πρώτες.

3.4.1 Missing values

Στις εγγραφές που περιέχουν NaN/na τιμές σε στήλες, όπως damage_property , damage_crops, deaths και injuries, θέσαμε την τιμή 0, επειδή δεν μας επηρέαζε το αποτέλεσμα. Για τα missing values των χαρακτηριστικων event_latitude , event_longitude, τα οποία αναφέρονται σε συντεταγμένες, επιλέξαμε άλλη επεξεργασία. Εφόσον οι πολιτείες βρίσκονται πάντα στο ίδιο εύρος συντεταγμένων, για κάθε πολιτεία πήραμε το μέσο όρο τους και αντικαταστήσαμε τις κενές τιμές με αυτό το μέσο όρο.

3.4.2 Class – Mapping

Αφαίρεση event_type τα οποία είναι πολύ σπάνια και συγχώνευση κάποιων τα οποία μοιάζουν πολύ, όπως για παράδειγμα:

- thunderstorm_wind heavy rain και thunderstorm_wind / heavy rain -> thunderstorm wind

Το mapping βρίσκεται στο Παράρτημα Β του Appendix.

4. ΕΞΕΡΕΥΝΗΣΗ ΔΕΔΟΜΕΝΩΝ

4.1 Η εξέλιξη του μεγέθους των καταστροφών στο χρόνο

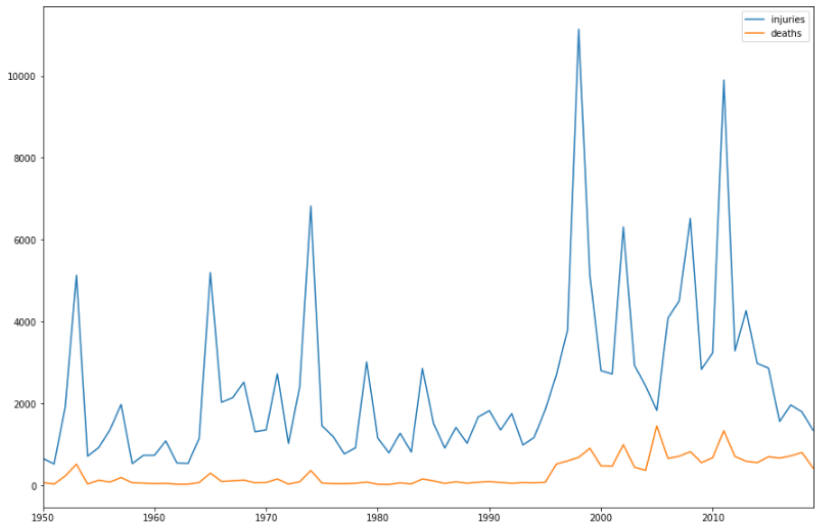
4.1.1 Χρόνος και αριθμός θανάτων/τραυματισμών

Αρχικά, δημιουργούμε ένα Data Frame που περιέχει την κάθε χρονολογία από το 1950 έως και το 2019, και αθροίζουμε για τη κάθε χρονολογία τους αντίστοιχους τραυματισμούς και θανάτους. Το Data Frame που προκύπτει είναι της μορφής:

	DATE	injuries	deaths
0	1950	659	70
1	1951	524	34
2	1952	1916	231
3	1953	5131	519

Εικόνα 4.1

Με το παραπάνω Data Frame δημιουργούμε το παρακάτω διάγραμμα χρόνου - τραυματισμών/θανάτων.

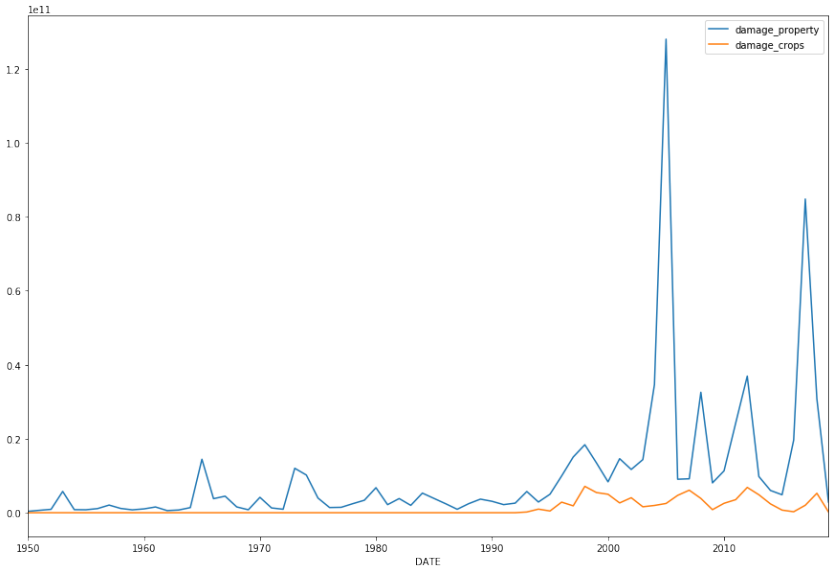


Εικόνα 4.2: Διάγραμμα τραυματισμών/θανάτων - χρόνου.

Η καμπύλη των θανάτων είναι πιο 'χαμηλά' και πιο ομαλή, κάτι που σημαίνει ότι, όσο μεγάλη και να είναι μια καταστροφή, οι απώλειες δεν ξεφεύγουν ιδιαίτερα. Οι τραυματισμοί έχουν διάφορα peaks στο χρόνο, γεγονός που μπορει να αιτιολογηθεί από μεγάλες κακοκαιρίες εκείνη την περίοδο, με αποτέλεσμα να τραυματιστεί ένας μεγάλος αριθμός ανθρώπων. Αν και παρατηρείται αυτή η κυματομορφή, τα 'κάτω' και 'πάνω' άκρα κινούνται αυξητικά με την πάροδο του χρόνου.

4.1.2 Χρόνος και ζημιά σε δολλάρια

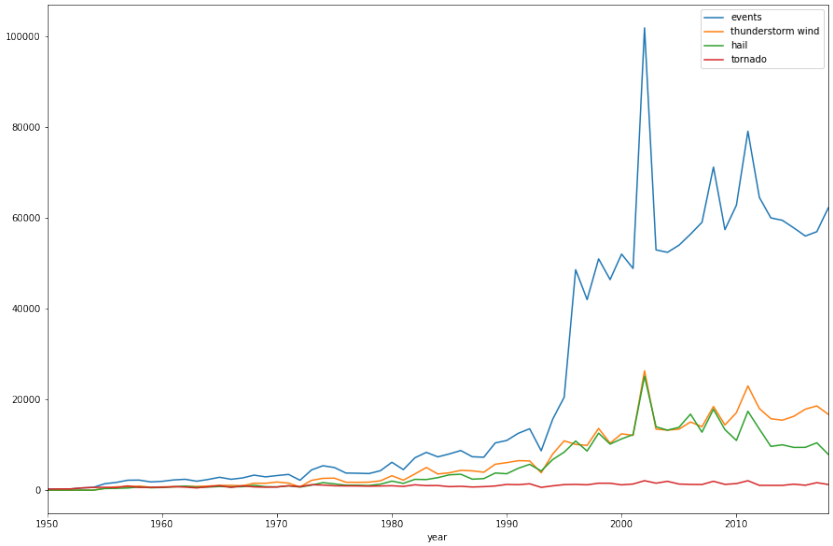
Παρατηρούμε, από το παρακάτω διάγραμμα, πως με την πάροδο του χρόνου, οι ζημιές σε καλλιεργήσιμες εκτάσεις είναι σχετικά σταθερές, ωστόσο δεν ισχύει το ίδιο για ιδιωτικές περιουσίες ανθρώπων. Είναι εμφανές πως, από το 2000 και μετά, είχαμε κάποιες πολύ μεγάλες καταστροφές, και για αυτό ανεβαίνει πάρα πολύ η ζημιά στο διάγραμμα.



Εικόνα 4.3: Διάγραμμα ζημιών σε καλλιεργήσιμες εκτάσεις/ιδιωτικές περιουσίες - χρόνου.

4.1.3 Χρόνος και και συχνότητα για κάποια μεγάλα καιρικά φαινόμενα

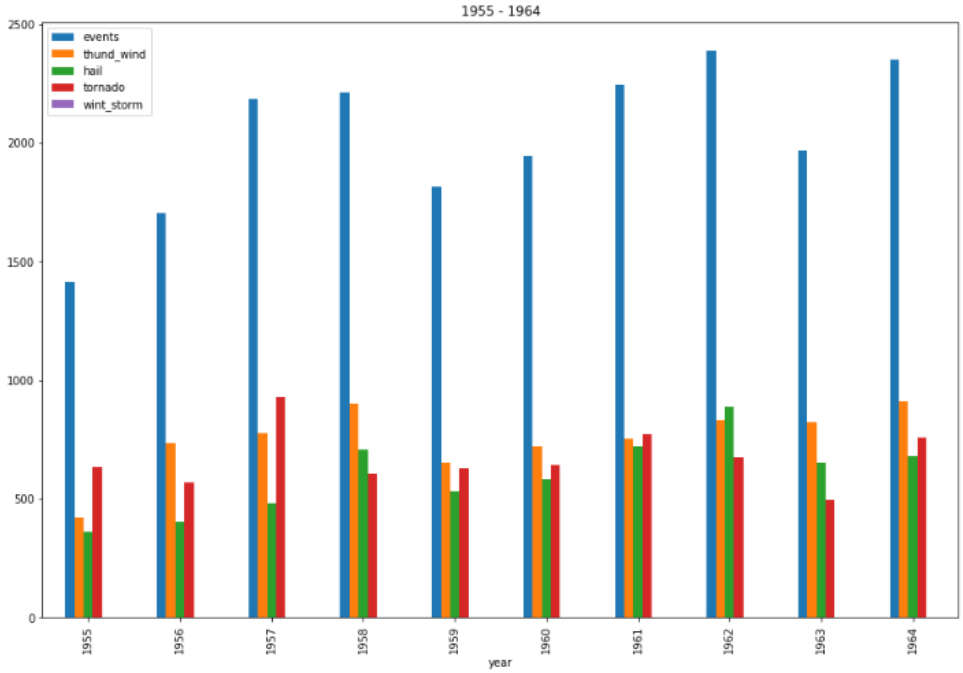
Το παρακάτω διάγραμμα δείχνει πως τα tornado δεν αυξάνονται με την πάροδο του χρόνου. Όμως, άλλα είδη κακοκαιρίας όπως hail ή thunderstorm wind ανεβαίνουν ειδικά από το 1990 και μετά. Γενικά, η μπλε γραμμή που δείχνει όλες τις κακοκαιρίες μαζί έχει δραματική άνοδο από ένα σημείο και μετά.



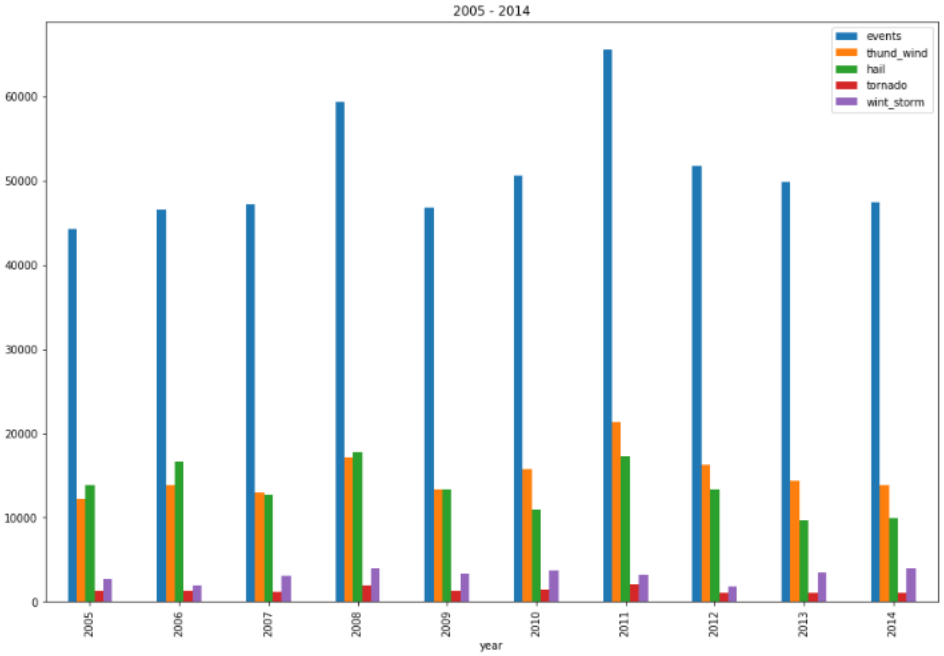
Εικόνα 4.4: Διάγραμμα αριθμού φαινομένων - χρόνου.

Στο επόμενο στάδιο θα μελετήσουμε τη συχνότητα για 4 είδη κακοκαιρίας ανά ορισμένες χρονικές περιόδους 10. Τα είδη κακοκαιρίας που θα μελετήσουμε είναι τα εξής:

- thunderstorm wind
- hail
- tornado
- wind_storm



Εικόνα 4.5: Συχνότητα για 4 είδη κακοκαιρίας τη χρονική περίοδο 1955-1964.



Εικόνα 4.6: Συχνότητα για 4 είδη κακοκαιρίας τη χρονική περίοδο 2005-2014.

Παρατηρούμε, παρά τις διακυμάνσεις, αύξηση εμφάνισης των φαινομένων με την πάροδο των χρόνων. Αυτό ίσως συμβαίνει, γιατί τις πρώτες δεκαετίες δεν υπήρχε η κατάλληλη τεχνολογία για να καταγραφούν τόσα πολλά καιρικά φαινόμενα όσο υπάρχουν τώρα.

4.2 Γεωγραφική κατανομή των καταστροφών, κατά είδος και μέγεθος

Σε αυτό το κομμάτι της εξερεύνησης δεδομένων, έχουμε δημιουργήσει ένα Data Frame με τις καταστροφές που υπέστη κάθε πολιτεία της Αμερικής. Οι καταστροφές χωρίζονται στις εξής κατηγορίες/χαρακτηριστικά:

- Τραυματισμοί (injuries)
- Απώλειες (deaths)
- Ζημιά σε ιδιωτικές περιουσίες
- Ζημιά σε καλλιέργειες

Επιπρόσθετα, για κάθε πολιτεία έχουμε συμπεριλάβει και τις συντεταγμένες της στο χάρτη.

	state	event_longitude	event_latitude	damage_property	damage_crops	deaths	injuries
0	Alabama	-86.794329	33.273120	1.862084e+10	2.756909e+08	931	9784
1	Alaska	-148.691622	62.656039	5.260037e+08	3.128348e+05	92	127
2	American samoa	-170.606827	-9.855073	3.232503e+08	2.283424e+07	44	184
3	Arizona	-111.818135	33.593247	5.202067e+09	3.504594e+08	325	1417
4	Arkansas	-92.645165	34.977785	9.368088e+09	3.176725e+08	643	6763
...	...	...	...	...	...	...	...

Εικόνα 4.7: DataFrame Name: Damage_each_state (πραγματικές τιμές στα χαρακτηριστικά των καταστροφών).

Έπειτα, αντιγράφουμε αυτό το Data Frame σε άλλη μεταβλητή και κανονικοποιούμε με τη χρήση Min-Max scaling τις στήλες damage_property , damage_crops, deaths και injuries, ώστε να αποφύγουμε την μεγάλη επιρροή των outliers. Επομένως, ο νέος πίνακας θα πάρει την παρακάτω μορφή με τιμές από 0 έως 1.

	state	event_longitude	event_latitude	damage_property	damage_crops	deaths	injuries
0	Texas	-98.724766	32.298647	1.000000e+00	1.000000	1.000000	1.000000
1	Louisiana	-92.241042	31.411497	7.476300e-01	0.147276	0.624742	0.152647
2	Florida	-82.522836	29.016986	5.468225e-01	0.474822	0.573711	0.279016
3	Mississippi	-89.661581	32.618340	3.584783e-01	0.121256	0.425773	0.319754
4	New jersey	-74.613006	40.319222	2.628344e-01	0.008935	0.152577	0.113035
...	...	...	...	...	...	...	...

Εικόνα 4.8: DataFrame Name: for_heatmap_df (τιμές για τις καταστροφές στο (0,1)).

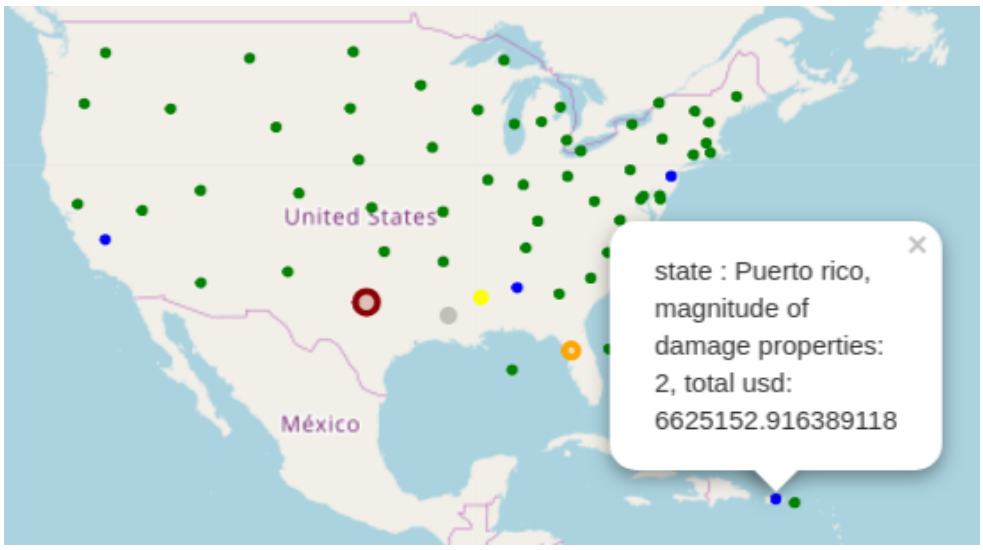
Στη συνέχεια, θα χωρίσουμε τις καταστροφές σε κατηγορίες με τη χρήση bins, ώστε να δούμε σε ποια κατηγορία ανήκει η κάθε μια. Για παράδειγμα, θα έχουμε 7 κατηγορίες. Η 7η θα σημαίνει μεγάλη καταστροφή ενώ η 1η μικρή. Επίσης το μέγεθος του κάθε κύκλου αναλογεί στο μέγεθος της καταστροφής.

	state	event_longitude	event_latitude	damage_property	damage_crops	deaths	injuries
0	Texas	-98.724766	32.298647	7	7	7	7
1	Louisiana	-92.241042	31.411497	6	2	5	2
2	Florida	-82.522836	29.016986	4	4	5	2
3	Mississippi	-89.661581	32.618340	3	1	3	3
4	New jersey	-74.613006	40.319222	2	1	2	1
...	...	...	...	...	...	...	...

Εικόνα 4.9: DataFrame Name: categorical_damage (κατηγορικές τιμές για τις μεταβλητές των ζημιών).

7:	darkred
6:	lightred
5:	red
4:	orange
3:	yellow
2:	blue
1:	green

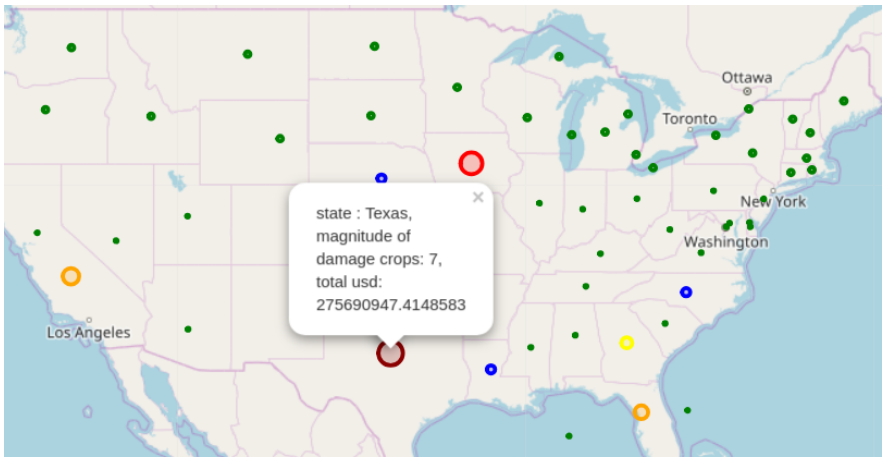
Πίνακας 4.1: Αντιστοιχία χρώματος-καταστροφών.



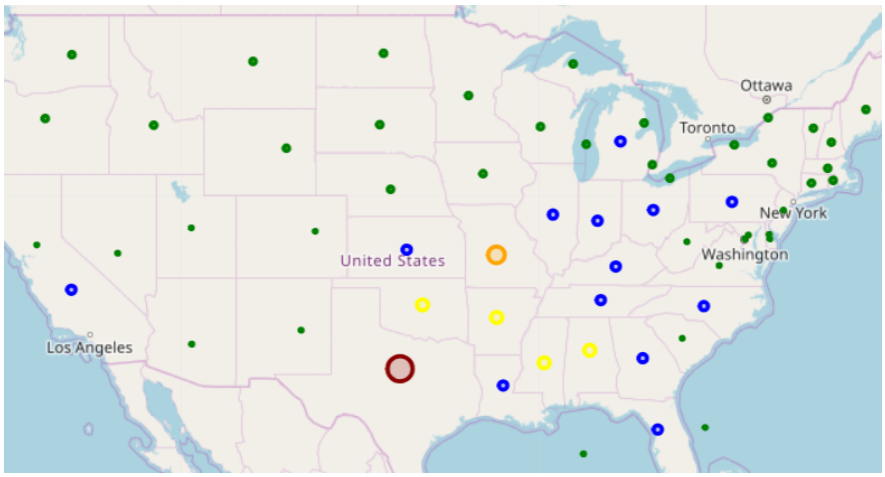
Εικόνα 4.10: Γεωγραφική κατανομή των καταστροφών, κατά είδος και μέγεθος.

Στον παραπάνω χάρτη (μέσω του maps.ipynb notebook), μπορούμε να κάνουμε zoom σε περιοχές και να κάνουμε κλικ πάνω σε μια απο τις κουκίδες για να πάρουμε διάφορες πληροφορίες. Στο παράδειγμα αυτό, βλέπουμε πως επιλέξαμε το Puerto Rico, με μέγεθος καταστροφής σε περιουσιακά στοιχεία 6625152.916389118 usd, το οποίο ανήκει στην κατηγορία 2, δηλαδή δεν έχει υποστεί ιδιαίτερα καταστροφικά καιρικά φαινόμενα σε σχέση με άλλες πολιτείες.

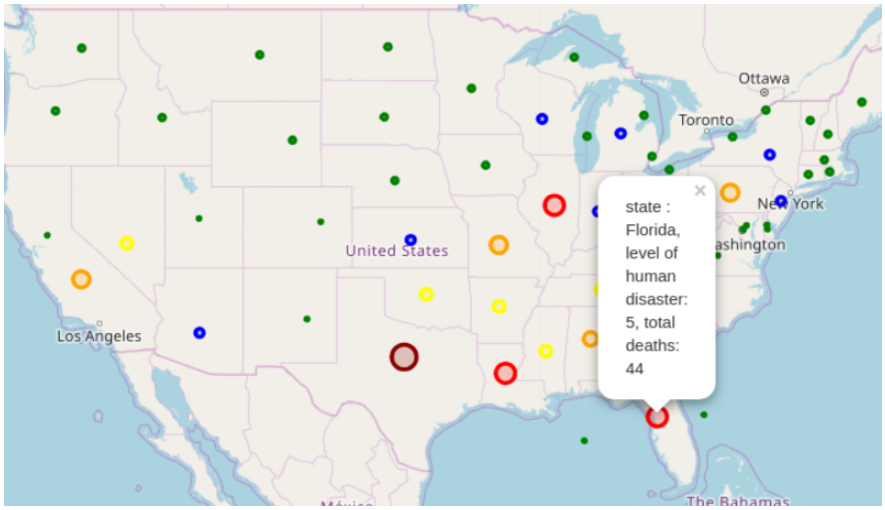
Παρακάτω, παρουσιάζουμε ορισμένους γεωγραφικούς χάρτες, οι οποίοι δείχουν τις ζημιές των καλλιεργήσιμων εκτάσεων (damage crops), τους θανάτους ανά πολιτεία και τους τραυματισμούς ανά πολιτεία.



Εικόνα 4.11: Γεωγραφικός χάρτης με τις ζημιές των καλλιεργήσιμων εκτάσεων.

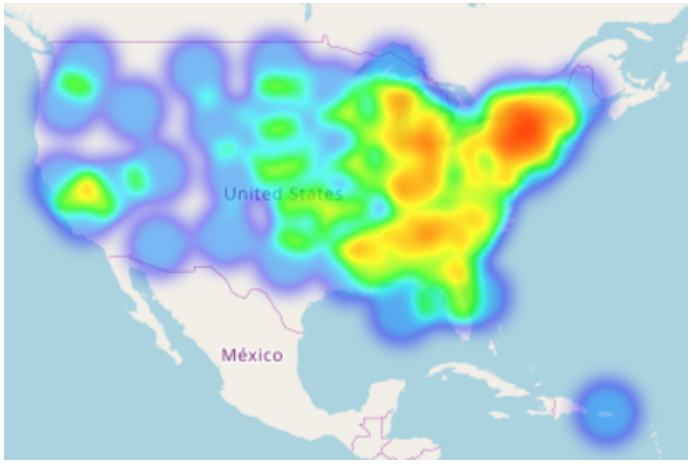


Εικόνα 4.12: Γεωγραφικός χάρτης με θανάτους ανά πολιτεία.

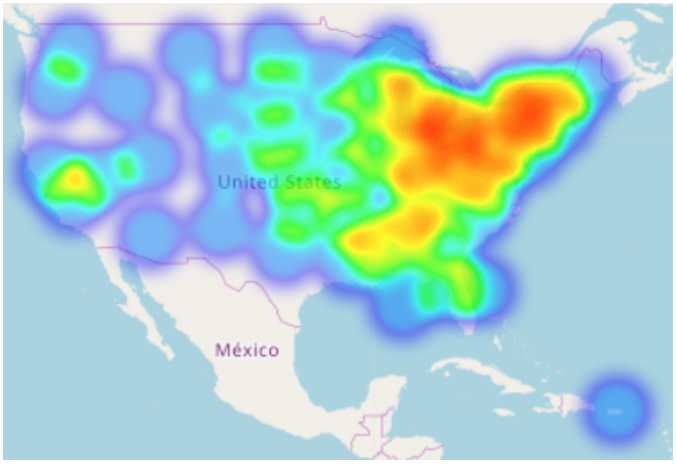


Εικόνα 4.13: Γεωγραφικός χάρτης με τραυματισμούς ανά πολιτεία.

Παρακάτω, παρουσιάζουμε τα heatmap για κάθε πολιτεία και ζημιά περιουσιων (damage_property), καθώς και για κάθε πολιτεία και ζημιά καλλιεργήσιμης έκτασης (damage_crops).

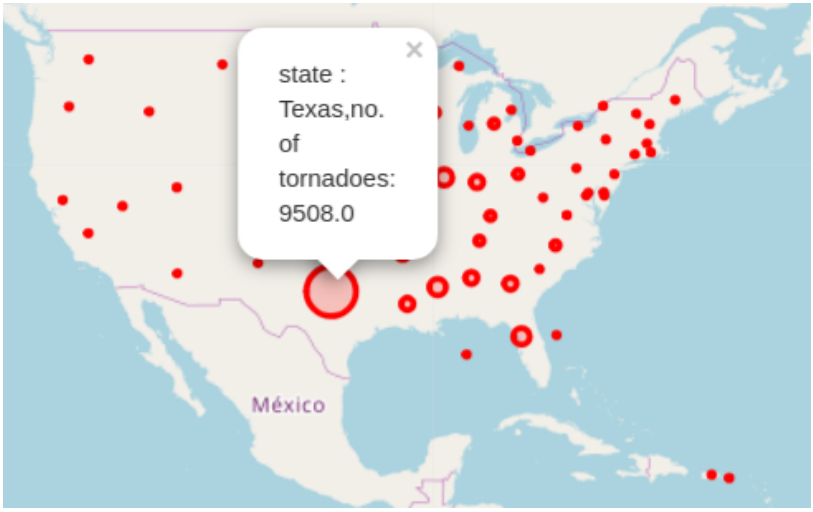


Εικόνα 4.14: Heatmap για κάθε πολιτεία και ζημιά περιουσιών.

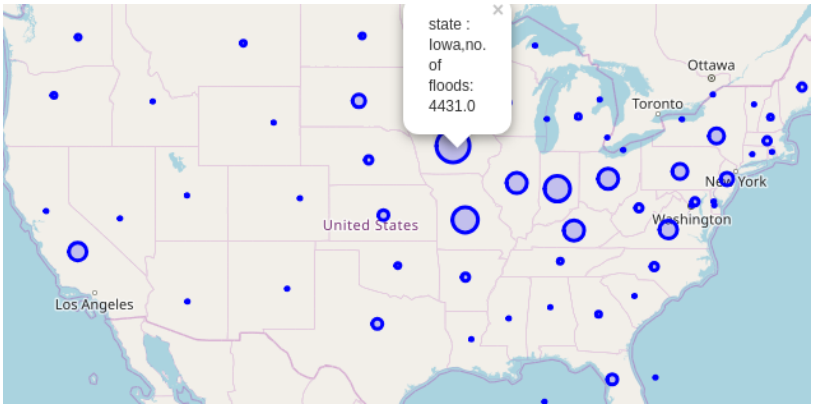


Εικόνα 4.15: Heatmap για κάθε πολιτεία και ζημιά καλλιεργήσιμης έκτασης.

Τέλος, παρουσιάζουμε άλλους δύο γεωγραφικούς χάρτες, όπου θα δούμε τη συχνότητα συγκεκριμένων φαινομένων ανά πολιτεία. Ο πρώτος χάρτης δείχνει πόσα tornado είχε κάθε πολιτεία, καθώς είναι το φαινόμενο με τις μεγαλύτερες ανθρώπινες απώλειες. Ο δεύτερος χάρτης δείχνει πόσα flood events είχε κάθε πολιτεία, καθώς είναι το φαινόμενο με τις μεγαλύτερες οικονομικές ζημιές.



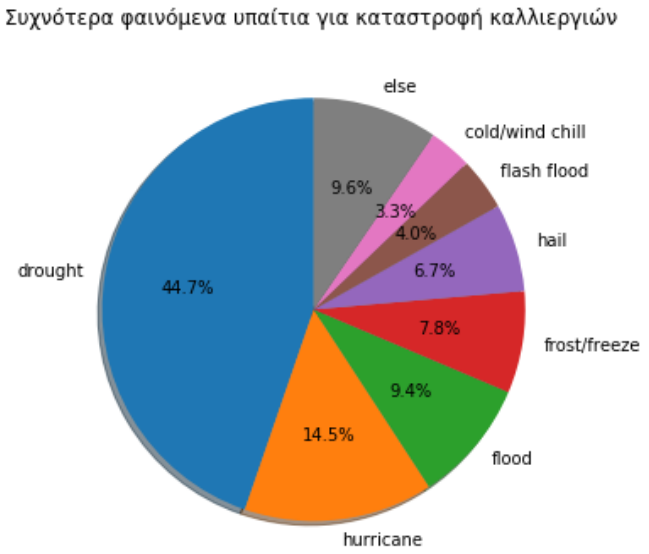
Εικόνα 4.16: Γεωγραφικός χάρτης με τον αριθμό των tornado κάθε πολιτείας.



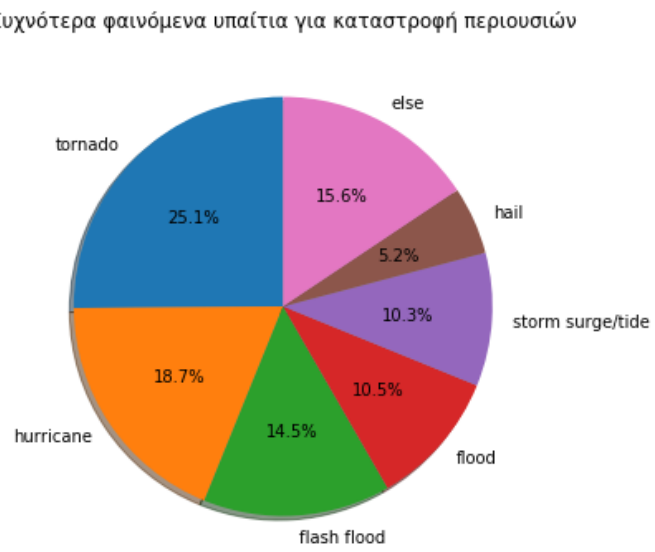
Εικόνα 4.17: Γεωγραφικός χάρτης με τον αριθμό των flood events κάθε πολιτείας.

4.3 Σχέση μεγέθους – τύπου καταστροφών

Αρχικά, διαχωρίσαμε τα τέσσερα είδη καταστροφών (deaths, injuries, damage in crops, damage in property) και εξετάσαμε το μέγεθος της κάθε μίας ξεχωριστά. Επειδή τα είδη των φαινομένων, που είναι καταγεγραμμένα στο dataset, είναι πολλά, κρατήσαμε αυτά με τη μεγαλύτερη συχνότητα εμφανίσεων ανά είδος καταστροφής. Συγκεκριμένα κρατήσαμε αυτά με ποσοστό εμφάνισης >3% επί του συνόλου (καθ' όλη τη διάρκεια των ετών), και τα υπόλοιπα τα κρατήσαμε σε μια κατηγορία 'else'. Παρακάτω παρουσιάζουμε τα pie charts με τα αντίστοιχα αποτελέσματα ποσοστών.

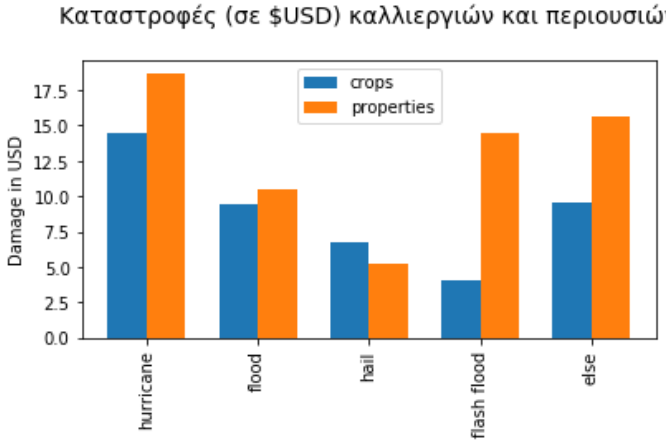


Εικόνα 4.18: Piechart συχνότερων φαινομένων υπαίτιων για καταστροφή καλλιεργιών.

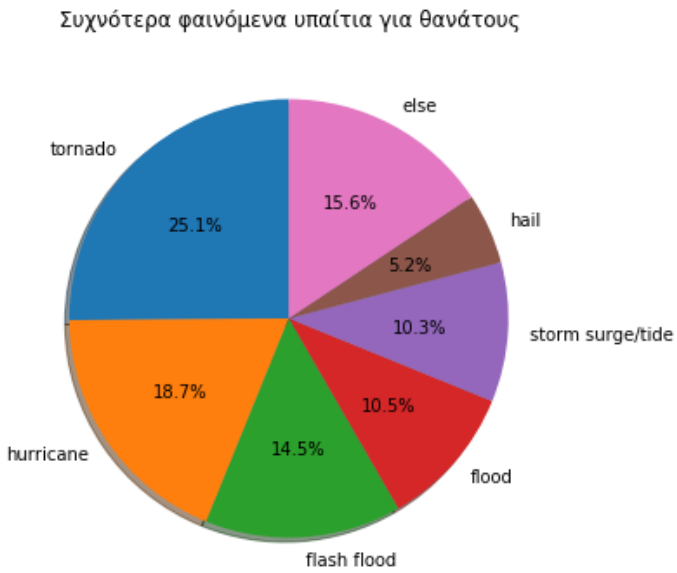


Εικόνα 4.19: Piechart συχνότερων φαινομένων υπαίτιων για καταστροφή περιουσιών.

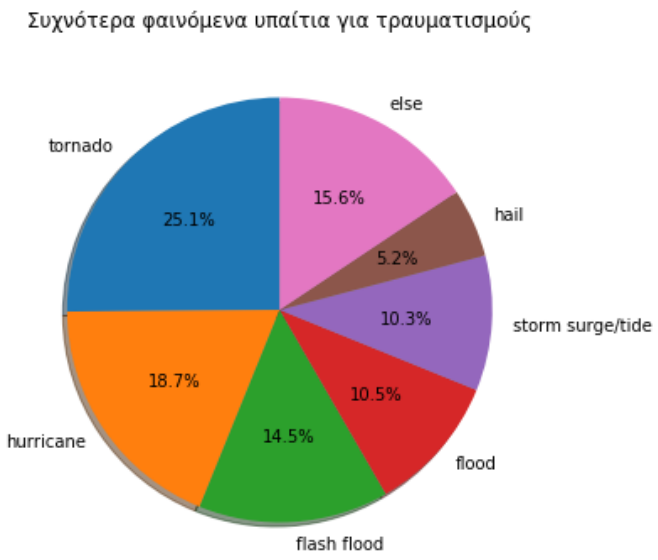
Παρατηρούμε ότι η ξηρασία είναι με μεγάλη διαφορά (44.7%) η πρώτη αιτία καταστροφής των καλλιεργιών. Ακολουθούν οι τυφώνες (14.5%). Η καταστροφή περιουσιών παρουσιάζει μικρότερες διαφορές. Η πρώτη αιτία είναι οι ανεμοστρόβιλοι (25.1%), στη συνέχεια έχουμε τους τυφώνες(18.7%).



Εικόνα 4.20: Barplot μεγέθους των καταστροφών που προξενήθηκαν από τα πιο κοινώς εμφανιζόμενα είδη φαινομένων.



Εικόνα 4.21: Piechart συχνότερων φαινομένων υπαίτιων για θανάτους.

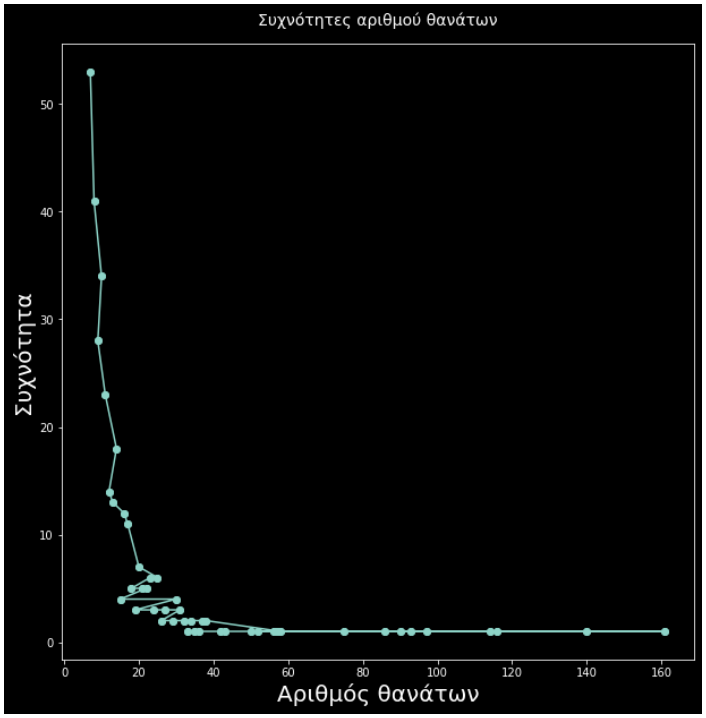


Εικόνα 4.22: Piechart συχνότερων φαινομένων υπαίτιων για τραυματισμούς.

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.

Το παρακάτω διάγραμμα αφορά τον αριθμό θανάτων που προκαλλούνται από ένα φαινόμενο και των αντίστοιχων συχνοτήτων τους. Δημιουργεί μια υπερβολική καμπύλη που δείχνει πως μικρό πλήθος θανάτων (<10) έχει πολύ υψηλή συχνότητα, ενώ μεγάλο πλήθος θανάτων χαμηλή συχνότητα.

Σημαντική σημείωση: Από τα δεδομένα για το πλήθος των θανάτων δε συμπεριλάβαμε τις τιμές 1-6 (όταν τόσοι άνθρωποι πέθαναν δηλαδή), καθώς και την τελευταία παρατήρηση γιατί ήταν μεγάλο outlier (648) και, κατά συνέπεια, παρουσιαζόταν δυσανάγνωστο διάγραμμα.



Εικόνα 4.23: Αριθμός θανάτων που προκαλλούνται από ένα φαινόμενο και οι αντίστοιχες συχνότητες.

Παραθέτουμε και τα μέτρα κεντρικής τάσης που υπολογίστηκαν:

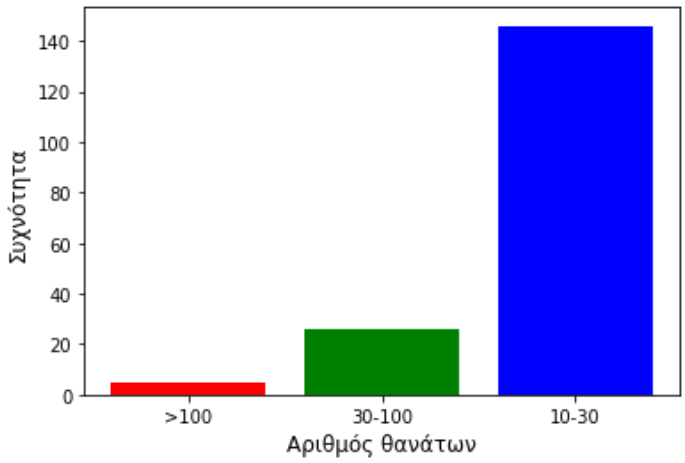
mean without the outlier value 638:	1.814
deaths average:	1.87
std:	7.238
median:	1.0
total number of deaths caused by storms:	11424

Πίνακας 4.2: Μέτρα κεντρικής τάσης.

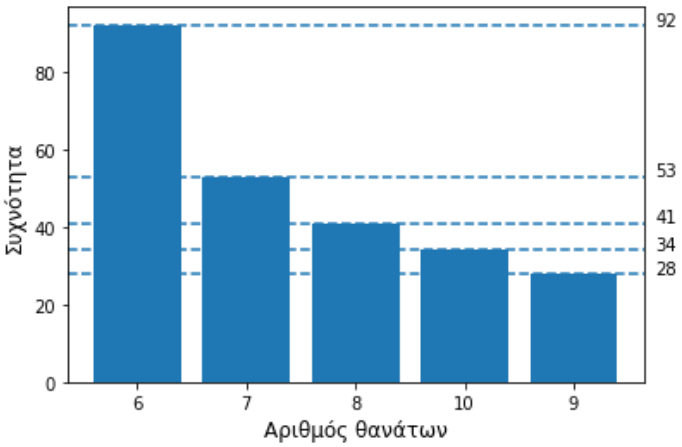
Χωρίσαμε επίσης σε 4 ομάδες το πλήθος των καταγεγραμμένων θανάτων.

- $>$ 100
- 30 - 100
- 10 - 30
- 5 - 10

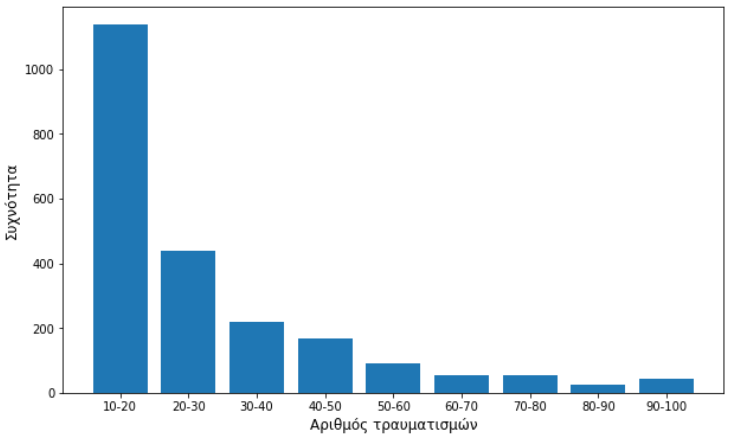
Για < 6, οι εμφανίσεις είναι πολύ μεγάλες και θα αλλοιωθεί η γραφική αναπαράσταση σε bar plot, οπότε επιλέξαμε να κάνουμε ξεχωριστό διάγραμμα. Παρακάτω παρουσιάζουμε μερικά περιγραφικά bar plots.



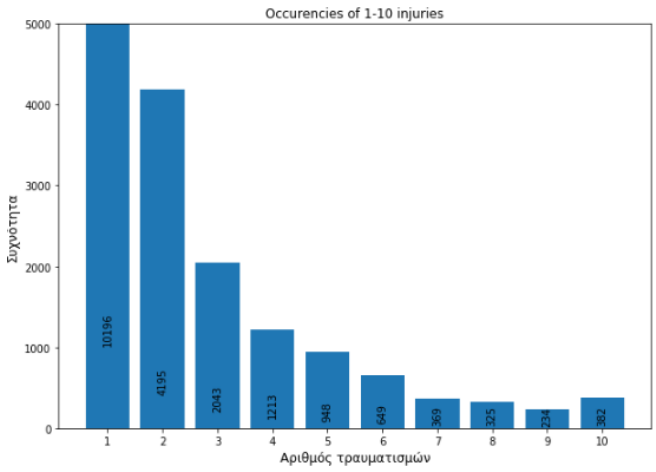
Εικόνα 4.24



Εικόνα 4.25



Εικόνα 4.26



Εικόνα 4.27

5. ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΚΑΙ ΚΑΝΟΝΕΣ ΣΥΣΧΕΤΙΣΗΣ

5.1 Association Rules - Itemsets (στοιχειοσύνολα)

Σε αυτό το μέρος της εργασίας, θα μελετήσουμε την ύπαρξη στοιχειοσυνόλων και κανόνων συσχέτισης. Τα στοιχειοσύνολα ορίζονται ως συλλογές k-πλήθους αντικειμένων. Χαρακτηρίζονται από την 'υποστήριξη', δηλαδή το πόσο συχνά εμφανίζεται το στοιχειοσύνολο στον πίνακά μας.

Η ισχύς κάθε κανόνα συσχέτισης μετριέται από την υποστήριξη (support) και την εμπιστοσύνη (confidence). Η εμπιστοσύνη δείχνει την πιθανοφάνεια να υπάρχει στο στοιχειοσύνολο ένα στοιχείο y, όταν υπάρχει ταυτόχρονα ένα άλλο στοιχείο x.

Επίσης, υπάρχει και μια άλλη μετρική που ονομάζεται ανύψωση (lift) και είναι το ίδιο με την εμπιστοσύνη, όμως συνυπολογίζει και την υποστήριξη που έχει και το στοιχείο y εκτός του x. Τιμές για το lift που είναι ίσες με 1 υποδεικνύουν ότι τα δύο στοιχεία είναι ανεξάρτητα. Τιμές μεγαλύτερες του 1 δείχνουν ότι είναι πιθανό, όταν υπάρχει το y, να υπάρχει και το x, ενώ τιμές μικρότερες του 1 δείχνουν ότι η ύπαρξη του x μειώνει την πιθανότητα ταυτόχρονης ύπαρξης του y.

Τέλος, έχουμε και την μετρική πεποίθηση (conviction), η οποία δηλώνει το ρυθμό με τον οποίο η αναμενόμενη συχνότητα του y προκύπτει χωρίς το στοιχείο x. Ουσιαστικά, μπορεί να διατυπωθεί ως 'το πόσο συχνά ο κανόνας που λέει ότι x -> y μπορεί να κάνει λάθος πρόβλεψη'.

Στο δικό μας dataset, λαμβάνουμε τον παρακάτω πίνακα συχνοτήτων για κάθε event type και episode_id. Σημειώνεται πως έχει γίνει class mapping, δηλαδή έχουν αντιστοιχηθεί event types σε άλλα παρόμοια για να μικρύνουμε τον αριθμό κατηγοριών, όπως αναφέρθηκε στο pre-processing part.

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

Εικόνα 5.1

Έπειτα, τρέχουμε τον αλγόριθμο a-priori για minimum support = 0.02. Λάβαμε υπόψιν το μεγάλο όγκο δεδομένων (1,600,000) και το ότι υπάρχουν 9 κατηγορίες, και δοκιμάσαμε διάφορες τιμές για minimum support. Τιμές μεγαλύτερες του 0.02 δεν έδωσαν κανόνες συσχέτισης. Το 0.02 έδωσε τα παρακάτω αποτελέσματα:

	support	itemsets
7	0.355667	(thunderstorm wind)
3	0.285685	(hail)
1	0.107548	(flash flood)
14	0.090780	(thunderstorm wind, hail)
9	0.070513	(wind)
2	0.069444	(flood)
6	0.055069	(snow)
8	0.049267	(tornado)
5	0.040317	(lightning)
10	0.032899	(winter storm)

Εικόνα 5.2: Αποτελέσματα a-priori για minimum support = 0.02.

Οι κανόνες που δημιουργήθηκαν είναι οι εξής:

	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

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 παίρνουμε τα παρακάτω στοιχειοσύνολα για min_support = 0.05

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

Εικόνα 5.3: Αποτελέσματα fp-growth για minimum support = 0.05.

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

5.2 DBSCAN Συσταδοποίηση (DBSCAN Clustering)

Θελήσαμε να δούμε αν τα δεδομένα συσταδοποιούνται, όταν λαμβάνουμε υπόψιν μόνο τα χαρακτηριστικά που αφορούν τις καταστροφές. Χρησιμοποιήσαμε τις στήλες 'deaths', 'damage', 'damage_crops', 'damage_property' εφαρμόζοντας τον αλγόριθμο DBSCAN.
Αρχικά δημιουργήσαμε δυο διδιάστατα διανύσματα, το ένα με τις τιμές deaths-injuries και το άλλο με τις τιμές crops-property damage. Αφότου κανονικοποιήσαμε τις τιμές των διανυσμάτων (standard scaling), εφαρμόσαμε PCA για να μειώσουμε τις διαστάσεις από δυο σε μία στο κάθε διάνυσμα. Έτσι, καταλήξαμε με δυο διανύσματα, ένα που περιέχει την πληροφορία για deaths-injuries και το άλλο για crops-property damage. Ενώνοντας αυτά τα δυο διανύσματα λαμβάνουμε ένα σύνολο samples δυο χαρακτηριστικών (των προαναφερθέντων). Σε αυτά τα samples εφαρμόσαμε το DBSCAN αλγόριθμο.

5.2.1 Αλγόριθμος DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Ο DBSCAN είναι αλγόριθμος συσταδοποίησης, που βασίζεται στην πυκνότητα των δεδομένων. Η πυκνότητα των δεδομένων σχετίζεται πολύ με τη διαστατικότητα, με την έννοια ότι, όσο μεγαλώνουν οι διαστάσεις, τα δεδομένα αραιώνουν στο χώρο και η εγγύτητα τους τείνει να γίνεται πιο ομοιόμορφη. Οι αλγόριθμοι που βασίζονται στην πυκνότητα λοιπόν εντοπίζουν περιοχές δεδομένων υψηλής πυκνότητας διαχωριζόμενες μεταξύ τους από περιοχές χαμηλής πυκνότητας. Λειτουργεί ως εξής:

1. Ορίζεται μια ακτίνα εγγύτητας (epsilon) μεταξύ των σημείων (samples), και το ελάχιστον πλήθος γειτόνων n (min points). Με κέντρο ένα σημείο του dataset (αρχικά ένα τυχαίο) κάθε άλλο σημείο κατατάσσεται συγκριτικά με αυτό σε μια από τις εξής 3 κατηγορίες:
   (a) core point, μέσα στην περιοχή της ακτίνας epsilon
   (b) border point, στην περιφέρεια της ακτίνας epsilon
   (c) noise point, έξω από αυτήν
2. Ο αλγόριθμος ξεκινάει ψάχνοντας να βρει τουλάχιστον n core ή border points. Αν δε βρεθούν τουλάχιστον n points, το αρχικό σημείο χαρακτηρίζεται ως θόρυβος. Αυτό βέβαια αργότερα μπορεί να αλλάξει, καθώς μπορεί να βρεθεί στη γειτονιά (γείτονας = core/border point) ενός άλλου σημείου που υπάρχει σε cluster. Ο DBSCAN δρα επαναληπτικά, συνεχίζοντας κάθε φορά με ένα σημείο το οποίο δεν έχει επισκεφθεί (ορίσει ως κέντρο).

3. Όσο εκτελείται ο αλγόριθμος, αυξάνονται οι διαστάσεις των clusters, καθώς κάθε σημείο του cluster έχει τη δυνατότητα να το επεκτείνει βρίσκοντας νέους γείτονες. τις διαστάσεις τους, καθώς ο αλγόριθμος επαναλαμβάνει αυτή τη διαδικασία μέσα σε κάθε cluster, δηλαδή πηγαίνει σε κάθε core και border point και βρίσκει καινούριους γείτονες.
4. Η διαδικασία τερματίζει όταν κανένα σημείο από τα υπολοιπόμενα δεν μπορεί να μπει σε κάποιο cluster. Να σημειωθεί ότι δεν υπάρχει με αυτόν τον τρόπο αλληλεπικάλυψη των clusters, δηλαδή κάθε σημείο ανήκει σε ένα και μόνο ένα cluster.

5.2.2 Προεπεξεργασία και μετασχηματισμός δεδομένων

Αφαιρέσαμε τις γραμμές που εμφάνιζαν NaN/na values στα χαρακτηριστικά 'damage_property', 'deaths', 'damage_crops' και 'injuries'.

Για να οπτικοποιήσουμε φυσικά το clustering, θα πρέπει να μειώσουμε τη διάσταση των δεδομένων μας σε 2 ή 3. Όπως αναφέραμε, θα εφαρμόσουμε PCA και θα μειώσουμε τις διαστάσεις σε 2. Πριν από το PCA, ωστόσο, θα κάνουμε scaling των δεδομένων. Θα χρησιμοποιήσουμε το standard scaling του sklearn. Το standard scaling αφαιρεί το μέσο όρο και διαιρεί με την τυπική απόκλιση, καταλήγοντας έτσι τα δεδομένα να έχουν μέση τιμή 0 και τυπική απόκλιση 1.

5.2.3 Επιλογή παραμέτρων

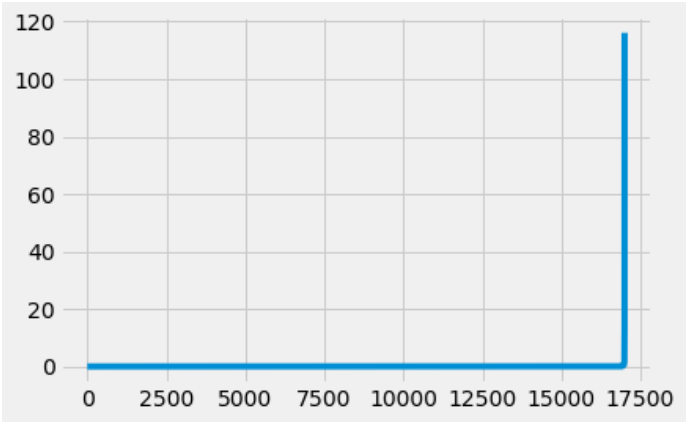
Δύο είναι οι σημαντικές παράμετροι που πρέπει να οριστούν με πολλή προσοχή στο DBSCAN, καθώς μικρές αλλαγές σε αυτούς οδηγούν σε έντονες διαφοροποιήσεις στα αποτελέσματα. Συγκεκριμένα, το epsilon και το min samples.

Epsilon: η ακτίνα μέσα στην οποία ψάχνουμε σημεία γύρω από ένα σημείο P.

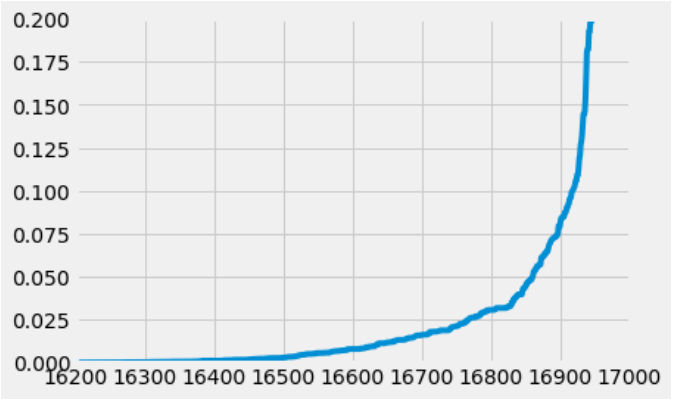
Min points: το ελάχιστο πλήθος σημείων που πρέπει να βρεθούν στην epsilon-γειτονιά ενός σημείου P.

Η εκτίμηση του epsilon δεν είναι εύκολη, ειδικά αν τα δεδομένα μας παρουσιάζουν διαφορετική πυκνότητα σε πεδία του χώρου. Ένας αλγόριθμος που βρίσκει καλή εκτίμιση του epsilon ακολουθεί την εξής διαδικασία:

Από κάθε δείγμα υπολογίζονται οι αποστάσεις από τα n πλησιέστερα σημεία (n = 4 εδώ), οι οποίες και ταξινομούνται. Έτσι, έχουμε ένα array με d αποστάσεις (d το μέγεθος του δείγματος) σε αύξουσα σειρά, και αυτές οι αποστάσεις γίνονται plotted (στον άξονα των X είναι απλά τα indices του array). Στο γράφημα της καμπύλης αναζητούμε πού εμφανίζεται η μέγιστη καμπυλότητα. Η y-συντεταγμένη στη μέγιστη καμπυλότητα είναι και το βέλτιστο epsilon. Για την εύρεση των αποστάσεων χρησιμοποιούμε knn. Πήραμε, αντί για όλο το Data Frame, ένα τυχαίο δείγμα μεγέθους 17000 (μέγεθος Data Frame περίπου 47.500), καθώς υπήρξε δυσκολία στην επεξεργασία περισσότερων δεδομένων. Παρακάτω είναι τα γραφήματα που προέκυψαν. Εκτιμούμε την τιμή αυτή να είναι περίπου .04.



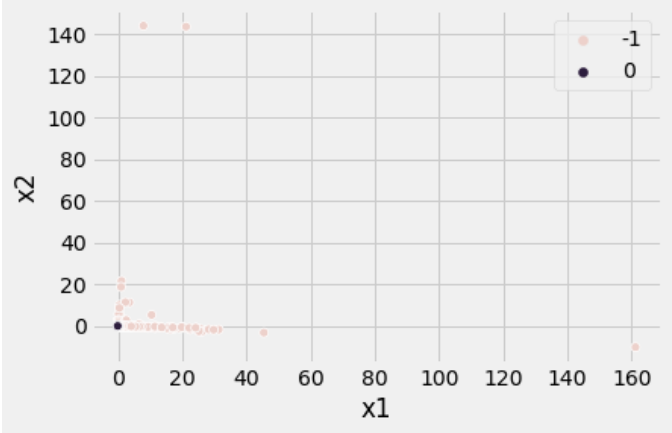
Εικόνα 5.4: Αποστάσεις σημείων από τα 4 πλησιέστερα.



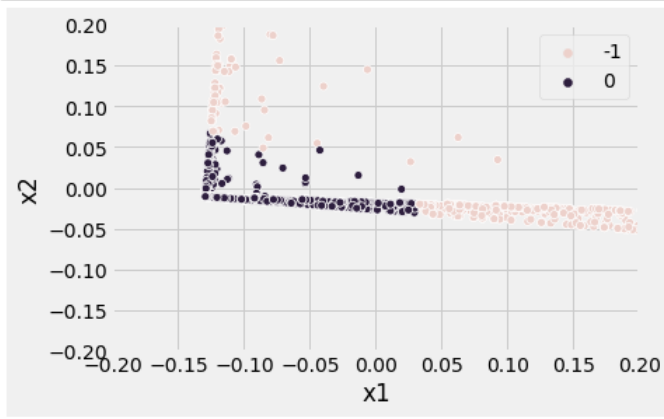
Εικόνα 5.5: Αποστάσεις σημείων από τα 4 πλησιέστερα – zoomed.

Εφαρμόζουμε τώρα το DBSCAN επιλέγοντας για min_samples (min points) 500 δεδομένου του μεγέθους του δείγματος που πήραμε.

Label	Counts
0	16237
-1	763



Εικόνα 5.6: Clusters.




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

Εικόνα 5.7: Clusters zoomed.

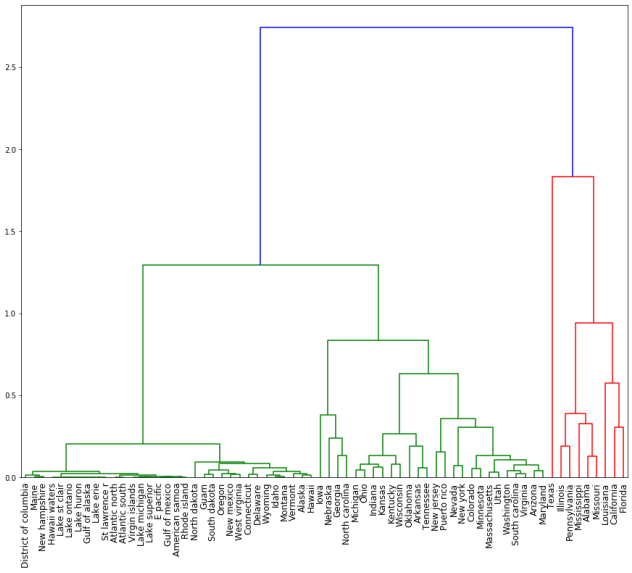
Δημιουργούνται 2 clusters. Το ένα (μαύρες κουκίδες) σχηματίζεται στις χαμηλές τιμές των αξόνων και περιέχει την πλειοψηφία των σημείων (95%). Η διασπορά είναι μικρή και τα σημεία πολύ πυκνά στο χώρο. Το άλλο (ροζ κουκίδες) περιέχει μόλις το 5% των σημείων. Η διασπορά είναι μεγάλη και τα σημεία αραιά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

5.2.4 Αξιολόγηση της απόδοσης του DBSCAN – Silhouette Score

Το Silhouette Score είναι μέτρο αξιολόγησης του clustering και υπολογίζεται χρησιμοποιώντας τη μέση τιμή των αποστάσεων των σημείων μέσα στα clusters και τη μέση τιμή των αποστάσεων μεταξύ των clusters (το κάθε ένα με το κοντινότερό του). Έτσι, ένα cluster με μεγάλη πυκνότητα, δηλαδή μικρές αποστάσεις των μεταξύ των σημείων, και μεγάλη απόσταστη από το πλησιέστερό του cluster, παρουσιάζει 'μοναδικότητα' κι έχει μεγάλο Silhouette Score. Γενικά, το σκορ κινείται στο διάστημα (-1,1) με -1 να είναι το χειρότερο σκορ και +1 το καλύτερο. Χρησιμοποιώντας την αντίστοιχη συνάρτηση του sklearn λαμβάνουμε αποτέλεσμα 0.85.

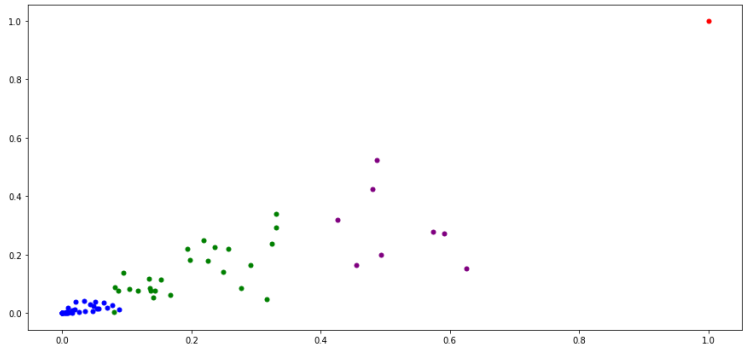
5.3 Ιεραρχική Συσταδοποίηση (Hierarchical Clustering)

Σκοπός είναι να βρούμε πόσες συστάδες θα δημιουργηθούν, αν έχουμε σα μέτρο ομοιότητας τη ζημιά (damage_property, damage_crops, injuries, deaths) για κάθε state της Αμερικής, επομένως χρησιμοποιούμε ένα από τα Data Frames που δημιουργήσαμε στην ενότητα 4.2 (βλέπε εικόνα 4.8). Για τη δημιουργία του δενδρογράμματος, χρησιμοποιούμε τη μέθοδο Ward. Λαμβάνουμε το παρακάτω δενδρόγραμμα, το οποίο και χρησιμοποιήσαμε ώστε να επιλέξουμε τον αριθμό των συστάδων (n_clusters = 4).



Εικόνα 5.8: Δενδρόγραμμα.

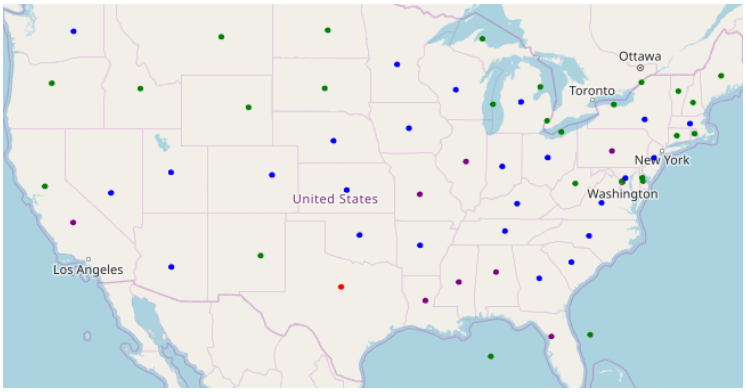
Η κόκκινη κουκίδα του παρακάτω διαγράμματος διασποράς για n_clusters = 4, αντιστοιχίζεται στην πολιτεία του Τέξας, η οποία παρουσιάζει τη μεγαλύτερη ζημιά και τους περισσότερους θανάτους και τραυματισμούς. Σημειώνουμε πως τα αποτελέσματα είναι κανονικοποιημένα από 0 έως 1.



```
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
```

Εικόνα 5.9: Διάγραμμα διασποράς για n_clusters = 4 και Τέξας (κόκκινη κουκίδα) - χαρακτηριστικά.

Τέλος, παρουσιάζουμε τον παρακάτω χάρτη, ο οποίος ανάλογα με το χρώμα δείχνει σε ποια συστάδα ανήκει η κάθε πολιτεία και τι μέγεθος καταστροφής είχε σε σχέση με τις άλλες πολιτείες. Παρατηρούμε πως, οι πολιτείες στη μεριά του Ατλαντικού είναι πιο επιρρεπείς στις μεγάλες ζημιές και στις ανθρώπινες απώλειες.



Εικόνα 5.10: Χάρτης αντιστοίχισης πολιτειών - συστάδων.

6. ΣΥΝΟΨΗ - ΣΥΜΠΕΡΑΣΜΑΤΑ

Το πρότζεκτ αυτό αποτελεί μια ανάλυση πάνω στις καταστροφικές συνέπειες καιρικών φαινομένων έντονης δριμύτητας στις ΗΠΑ την περίοδο 1950-2019. Η ανάλυση επικεντρώθηκε στη συσχέτιση του μεγέθους καταστροφών με (α) το χρόνο, (β) τη γεωγραφική τοποθεσία και (γ) το είδος του καιρικού φαινομένου. Το dataset που χρησιμοποιηθηκε είχε καταγεγραμμένες αντίστοιχες πληροφορίες (Appendix, Παράρτημα Α), το οποίο διαβάστηκε και επεξεργάστηκε ως Data Frame με την Python. Αρχικά έγινε μια πρώτη προεπεξεργασία στα δεδομένα ώστε να 'καθαρίσουν' από ελλιπείς και ασυνεπείς τιμές, διαγράφηκαν δεδομένα που δεν ήταν χρήσιμα και δημιουργήθηκαν νέα.

Ακολούθησε εξερεύνηση των δεδομένων με διάφορα διαγράμματα. Από τα διαγράμματα καταστροφών-χρόνου είδαμε ότι υπάρχει μια αυξητική τάση στον αριθμό των καταστροφών. Δεν είναι σκοπός μας να το συσχετίσουμε αιτιολογικά, ωστόσο, με την πάροδο των ετών αυξήθηκε και ο πληθυσμός στις ΗΠΑ. Όσον αφορά τη γεωγραφία, τα έντονα καιρικά φαινόμενα έχουν λάβει χώρα περισσότερο στο ανατολικό παρά στο δυτικό τμήμα των ΗΠΑ. Επίσης, το Τέξας είναι η πολιτεία που έχει πληγεί περισσότερο από ανεμοστρόβιλους και η Καλιφόρνια από πλημμύρες. Τέλος, διαπιστώσαμε ότι, 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.

Δοκιμάστηκαν αλγόριθμοι για συσχέτιση χαρακτηριστικών και εξαγωγή κανόνων, οι οποίοι, όμως, έδωσαν κανόνες μόνο με πολύ μικρή υποστήριξη και έδειξαν ότι δεν υπάρχει κάποια συσχέτιση μεταξύ των δεδομένων.

Εφαρμόστηκαν μέθοδοι clustering σε δεδομένα που είχαν ως χαρακτηριστικά μόνο τις τέσσερις καταστροφές. Έγινε αρχικά χρήση του DBSCAN, το οποίο βασίζεται στην πυκνότητα των δεδομένων. Η επιλογή κάποιων από τις παραμέτρους έγινε με αλγοριθμική μέθοδο. Βρέθηκαν δύο clusters, ένα με το 95% των δεδομένων και μικρή διασπορά, κι ένα με το υπόλοιπο 5% και πολύ μεγάλη διασπορά. Ο αλγόριθμος αξιολογήθηκε με το Silhouette Score, το οποίο αξιολόγησε το DBSCAN με σκορ 0.85. Στη συνέχεια, χρησιμοποιήθηκε η μέθοδος της ιεραρχικής συσταδοποίησης, και συγκεκριμένα η μέθοδος Ward, και βρέθηκαν 4 clusters, με την πολιτεία του Τεξας να παρουσιάζει τις μεγαλύτερες απώλειες. Τέλος, παρατηρήθηκε ότι, οι πολιτείες που βρίσκονται στη μεριά του Ατλαντικού, είναι πιο επιρρεπείς στις ζημιές και στις ανθρώπινες απώλειες.

APPENDIX

Παράρτημα Α: Επίσημη περιγραφή χαρακτηριστικών του dataset από τη Google

episode_id	STRING	NULLABLE	ID assigned by NWS to denote the storm episode; links the event details file with the information within location file
event_id	STRING	NULLABLE	ID assigned by NWS to note a single, small part that goes into a specific storm episode; links the storm episode between the three files downloaded from SPC's website
state	STRING	NULLABLE	The full text state name where the event occurred
state_fips_code	STRING	NULLABLE	Unique FIPS code identifier assigned to each state. State names and their corresponding FIPS codes are available as a BigQuery Public Dataset: bigquery-public-data.census_fips_codes.states_2016. The geographic polygons that define the perimeter of each state are available as a BigQuery Public Dataset: bigquery-public-data.geo_us_boundaries.u_states
event_type	STRING	NULLABLE	The only events permitted in Storm Data are listed in Table 1 of Section 2.1.1 of NWS Directive 10-1605 at http://www.nws.noaa.gov/directives/sym/pd01016005curr.pdf. The chosen event type is the one that most accurately describes the meteorological event leading to fatalities, injuries, damage, etc. However, significant events, such as tornadoes, having no impact or causing no damage, are also included in Storm Data.
cz_type	STRING	NULLABLE	Indicates whether the event happened in - C: County/Parish - Z: NWS zone - M: Marine
cz_fips_code	STRING	NULLABLE	Unique FIPS code identifier assigned to each county. State names and their corresponding FIPS codes are available as a BigQuery Public Dataset: bigquery-public-data.census_fips_codes.counties_2016. The geographic polygons that define the perimeter of each state are available as a BigQuery Public Dataset: bigquery-public-data.geo_us_boundaries.us_counties
cz_name	STRING	NULLABLE	(County/Parish, Zone or Marine Name assigned to the county FIPS number or NWS Forecast Zone NWS Forecast Zones are available as a BigQuery Public Dataset: bigquery-public-data.noaa_historic_severe_storms.nws_forecast_zones
wfo	STRING	NULLABLE	National Weather Service Forecast Office's area of responsibility (County Warning Area) in which the event occurred
event_begin_time	DATETIME	NULLABLE	The date and time that the event began. Note that episodes and events may have different start and end times if multiple events occured in the same episode
event_timezone	STRING	NULLABLE	The time zone in which the event_begin_time and the event_end_time is recorded.
event_end_time	DATETIME	NULLABLE	The date and time that the event ended. Note that episodes and events may have different start and end times if multiple events occured in the same episode
injuries_direct	INTEGER	NULLABLE	The number of injuries directly related to the weather event
injuries_indirect	INTEGER	NULLABLE	The number of injuries indirectly related to the weather event
deaths_direct	INTEGER	NULLABLE	The number of deaths directly related to the weather event
deaths_indirect	INTEGER	NULLABLE	The number of deaths indirectly related to the weather event
damage_property	INTEGER	NULLABLE	The estimated amount of damage to property incurred by the weather event, in USD at the time of the event. Values are not adjusted for inflation. Note: Values listed as 0 do not necessarily mean that no property damage occurred as a result of the event

damage_crops	INTEGER	NULLABLE	The estimated amount of damage to crops incurred by the weather event, in USD at the time of the storm. Values are not adjusted for inflation. Note: Values listed as 0 do not necessarily mean that no property damage occurred as a result of the event
source	STRING	NULLABLE	Source reporting the weather event. Note: This can be any entry. Values are not restricted to specific categories
magnitude	FLOAT	NULLABLE	Measured extent of the magnitude type. This is only used for wind speeds and hail size. Wind speeds are in MPH; Hail sizes are in inches
magnitude_type	STRING	NULLABLE	Differentiates between the type of mangitude measured. - EG = Wind Estimated Gust - ES = Estimated Sustained Wind - MS = Measured Sustained Wind - MG = Measured Wind Gust. No magnitude type is included for hail
flood_cause	STRING	NULLABLE	Reported or estimated cause of the flood
tor_f_scale	STRING	NULLABLE	Enhanced Fujita Scale describes the strength of the tornado based on the amount and type of damage caused by the tornado. The F-scale of damage will vary in the destruction area; therefore, the highest value of the F-scale is recorded for each event. - EF0 – Light Damage (40 – 72 mph) - EF1 – Moderate Damage (73 – 112 mph) - EF2 – Significant damage (113 – 157 mph) - EF3 – Severe Damage (158 – 206 mph) - EF4 – Devastating Damage (207 – 260 mph) - EF5 – Incredible Damage (261 – 318 mph)
tor_length	STRING	NULLABLE	Length of the tornado or tornado segment while on the ground (minimal of tenths of miles)
tor_width	STRING	NULLABLE	Width of the tornado or tornado segment while on the ground (in feet)
tor_other_wfo	STRING	NULLABLE	Indicates the continuation of a tornado segment as it crossed from one National Weather Service Forecast Office to another. The subsequent WFO identifier is provided within this field.
location_index	STRING	NULLABLE	Number assigned by NWS to specific locations within the same Storm event. Each event's sequentially increasing location index number will have a corresponding lat/lon point
event_range	FLOAT	NULLABLE	A hydro-meteorological event will be referenced, minimally, to the nearest tenth of a mile, to the geographical center (not from the village/city boundaries or limits) of a particular village/city, airport, or inland lake, providing that the reference point is documented in the Storm Data software location database.
event_azimuth	STRING	NULLABLE	16-point compass direction from a particular village/city, airport, or inland lake, providing that the reference point is documented in the Storm Data software location database of > 130,000 locations.
reference_location	STRING	NULLABLE	Reference location of the center from which the range is calculated and the azimuth is determined
event_latitude	FLOAT	NULLABLE	The latitude where the event occurred (rounded to the hundredths in decimal degrees; includes an '-' if it's S of the Equator)
event_longitude	FLOAT	NULLABLE	The longitude where the event occurred (rounded to the hundredths in decimal degrees; includes an '-' if it's W of the Prime Meridian)
event_point	GEOGRAPHY	NULLABLE	Geographic representation of the event_longitude and latitude

Παράρτημα Β: Αντιστοίχιση αγγλικών-ελληνικών όρων

drought	ξηρασία
hurricane	τυφώνας
flood	πλημμύρα
frost/freeze	παγετώνας
hail	χαλάζι
flash flood	ξαφνική πλημμύρα
cold/wind chill	κρύο/κρύος άνεμος
thunderstorm wind	Ευθυτενείς, μη περιστρεφόμενοι άνεμοι (!=ανεμοστρόβιλος)
tropical storm	Τυφώνας μέτριας έντασης (μεταξύ 39-74mph)
high wind	πολύ δυνατός άνεμος
tornado	ανεμοστρόβιλος
wildfire	πυρκαγιά σε περιβάλλον άγριας φύσης
coastal flood	πλημμύρα παρακτιων περιοχών (αύξηση του επιπέδου υδάτων)
heat	καύσωνας
rip current	υδάτινο ρεύμα
lightning	αστραπές
winter weather	χειμερινός καιρός
excessive heat	υπερβολικός καύσωνας

Παράρτημα Γ: Επεξηγήσεις για τις τιμές των στηλών 'magnitude', 'magnitude_type' και 'tor_f_scale'

'Magnitude' : Χρήση μόνο για ταχύτητα ανέμων και μέγεθος χαλαζιού. Η ταχύτητα του ανέμου μετριέται σε MPH(Miles per hour) ενώ το μέγεθος του χαλαζιού σε ίντσες.

'Magnitude type' : Διαφοροποιείται μεταξύ του είδους του μεγέθους(magnitude) που μετράται. Για το χαλάζι δεν συμπεριλαμβάνεται είδος magnitude.

- EG = Wind Estimated Gust (Αναμενόμενη τιμή ρίπων στον άνεμο)
- ES = Estimated Sustained Wind (Αναμενόμενη τιμή διάρκειας ανέμου)
- MS = Measured Sustained Wind (μέτρηση διάρκειας ανέμου)
- MG = Measured Wind Gust (Μέτρηση ρίπων στον άνεμο)

'tor_f_scale': "Enhanced Fujita Scale" Περιγράφει την ισχύ του ανεμοστρόβιλου(tornado) με βάση την ποσότητα και τον τύπο της ζημίας που προκλήθηκε από τον ανεμοστρόβιλο. Η κλίμακα F της βλάβης θα ποικίλει στην περιοχή καταστροφής. Επομένως, καταγράφεται για κάθε συμβάν η υψηλότερη τιμή της κλίμακας F.

- EF0 – Light Damage (40 – 72 mph)
- EF1 – Moderate Damage (73 – 112 mph)
- EF2 – Significant damage (113 – 157 mph)
- EF3 – Severe Damage (158 – 206 mph)
- EF4 – Devastating Damage (207 – 260 mph)
- EF5 – Incredible Damage (261 – 318 mph)

Επίσης, η τιμή 'EFU' που θα δούμε ότι υπάρχει είναι αντίστοιχο του 'unknown'.

Παράρτημα Δ: Class Mapping - λεξικό

'hail/icy roads', 'hail flooding', 'marine hail', 'hail': 'hail',

'thunderstorm winds funnel clou', 'thunderstorm winds/flash flood', 'thunderstorm winds heavy rain', 'thunderstorm winds/heavy rain', 'thunderstorm winds lightning', 'thunderstorm winds/flooding', 'thunderstorm winds/ flood', 'marine thunderstorm wind', 'thunderstorm wind/ trees', 'thunderstorm wind/ tree', 'thunderstorm wind': 'thunderstorm wind',

'hurricane (typhoon)', 'marine hurricane/typhoon', 'waterspout', 'hurricane': 'hurricane',

'tornadoes, tstm wind, hail', 'tornado/waterspout', 'tornado': 'tornado',
'flash flood': 'flash flood',
'lakeshore flood', 'coastal flood', 'high surf', 'flood': 'flood',
'drought', 'heat': 'drought',
'cold/wind chill': 'cold/wind chill',
'extreme cold/wind chill', 'marine strong wind', 'strong wind', 'heavy wind', 'high wind', : 'wind': 'wind',
'lake-effect snow', 'heavy snow', 'avalanche', 'high snow', 'blizzard': 'snow',
'wildfire': 'wildfire',
'heavy rain': 'heavy rain',
'storm surge/tide': 'storm surge/tide',
'dust storm': 'storm',
'tsunami': 'tsunami',
'winter weather': 'winter weather',
'winter storm': 'winter storm',
'frost/freeze': 'frost/freeze',
'marine tropical depression': 'marine tropical depression',
'marine tropical storm': 'marine tropical storm',
'astronomical low tide': 'astronomical low tide',
'tropical depression': 'tropical depression',
'marine high wind': 'marine high wind',
'volcanic ashfall': 'volcanic ashfall',
'marine dense fog': 'marine dense fog',
'marine lightning': 'marine lightning',
'northern lights': 'northern lights',
'excessive heat': 'excessive heat',
'tropical storm': 'tropical storm',
'debris flow': 'debris flow',
'dense smoke': 'dense smoke',
'volcanic ash': 'volcanic ash',
'freezing fog': 'freezing fog',
'sneakerwave': 'sneakerwave',
'rip current': 'rip current',
'funnel cloud': 'funnel cloud',
'landslide': 'landslide',
'dust devil': 'dust devil',
'ice storm': 'ice storm',
'dense fog': 'dense fog',
'lightning': 'lightning',
'seiche': 'seiche',
'sleet': 'sleet',
'other': 'other'