2η Εργασία στο μάθημα “Τεχνικές εξόρυξης Δεδομένων”, εαρινό εξάμηνο 2018-19,

Ομαδική εργασία **2 ατόμων**

ΘΕΜΑ: Ανάλυση, περιγραφή, και αξιολόγηση σε μεγάλα δεδομένα.

Σκοπός της εργασίας είναι η κατανόηση και η εξερεύνηση (data exploration) των δεδομένων εισόδου. Η υλοποίηση της εργασίας θα γίνει στην γλώσσα προγραμματισμού Python (όπως και η 1η άσκηση) με την χρήση των εργαλείων/βιβλιοθηκών: jupyter notebook, pandas, κτλ.

Το θέμα (dataset) με το οποίο θα ασχοληθείτε είναι το εξής:

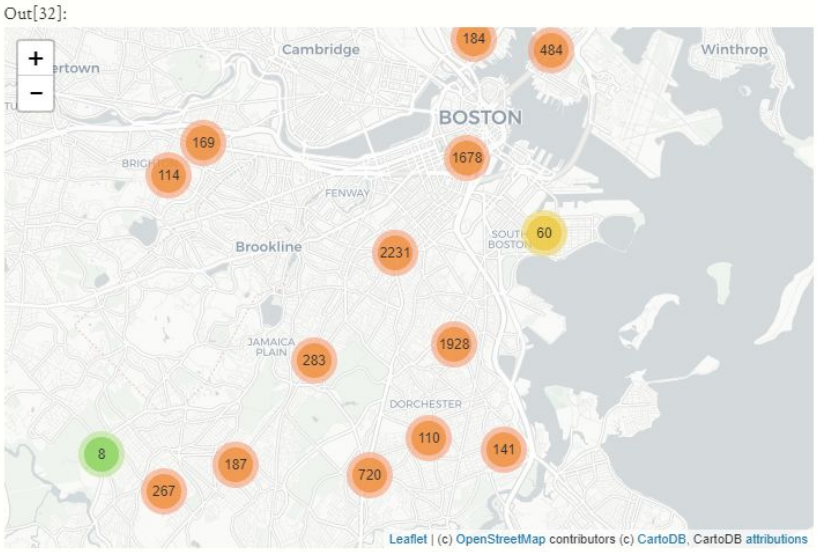
Crime Data : είναι ένα σύνολο δεδομένων που περιέχει αρχεία από το σύστημα αναφοράς εγκλημάτων στη Βοστώνη, και περιλαμβάνει ένα σύνολο πεδίων που αφορούν την καταγραφή του είδους του συμβάντος καθώς και πότε και πού συνέβη. Τα δεδομένα βρίσκονται σε σχετικό φάκελο στο eclass.

Τα ερωτήματα τα οποία πρέπει να απαντήσετε είναι τα παρακάτω:

1. Παρουσιάστε τα γραφήματα που δείχνουν το πλήθος των εγκλημάτων ανά χρόνο, ανά μήνα και ανά ημέρα. Επίσης το πλήθος των εγκλημάτων ανά περιοχή (`DISTRICT`)
2. Μελετήστε τα δεδομένα χρησιμοποιώντας την στήλη Shootings. Ποια χρονιά έχουμε τα περισσότερα shootings ; Σε ποια περιοχή (`DISTRICT`) εμφανίζονται τα περισσότερα περιστατικά shootings ;
3. Χρησιμοποιήστε την πληροφορία από τη στήλη ‘HOUR’ και φτιάξτε μία καινούρια στήλη που αντιπροσωπεύει την πληροφορία “Day or Night”. Αν η ώρα είναι μεταξύ 18:00 μμ - 06:00 πμ θεωρούμε ότι είναι νύχτα, αλλιώς θεωρούμε ότι είναι μέρα. Είναι περισσότερα τα εγκλήματα την ημέρα ή τη νύχτα ;
4. Συνδιάστε την νέα στήλη που προέκυψε (Day or Night) με την στήλη “OFFENSE_CODE_GROUP“ και απαντήστε στην ερώτηση : Ποιος είναι ο πιο συχνός τύπος εγκλήματος που συμβαίνει την ημέρα ;

5. Χρησιμοποιώντας τις στήλες Lat και Log και την βιβλιοθήκη KMeans θα εφαρμόσετε clustering με βάση την γεωγραφική τοποθεσία. Δοκιμάστε τον KMeans με 2,3,5,10 clusters. Στη συνέχεια συνδιάστε περισσότερες στήλες από τα δεδομένα σας και εφαρμόστε το clustering, δηλαδή με (location, OFFENSE_CODE) και (location, MONTH).

Bonus: Χρησιμοποιώντας τη βιβλιοθήκη folium (https://github.com/python-visualization/folium) φτιάξτε έναν interactive χάρτη όπου θα φαίνονται clusters για ένα συγκεκριμένο περιστατικό (για παράδειγμα επιλέγουμε τις γραμμές που έχουν ως περιστατικό “Drug Violation” από τη στήλη OFFENSE_CODE_GROUP)
(hint: χρησιμοποιήστε το MarkerCluster plugin και τη στήλη “location”)



Το παραδοτέο και σε αυτή την εργασία θα είναι ένα jupyter notebook στο οποίο θα έχετε τον κώδικα και τις οπτικοποιήσεις (plots) καθώς και σύντομη περιγραφή των συμπερασμάτων που θα βγουν από τα δεδομένα σας.