ΕΡΓΑΣΙΑ
ΑΝΑΛΥΣΗΣ ΔΕΔΟΜΕΝΩΝ 2022

Σύνολο δεδομένων

Το σύνολο δεδομένων σας θα είναι μια βάση SQLite με δεδομένα ποδοσφαιρικών αγώνων από ευρωπαϊκά πρωταθλήματα. Μερικές αρχικές πληροφορίες για τα το πως δημιουργήθηκε η βάση αυτή μπορείτε να βρείτε στον ακόλουθο σύνδεσμο:

https://www.kaggle.com/hugomathien/soccer

Άσκηση 1 (2 Μονάδες)

Κατεβάστε τοπικά το αρχείο **football-dataset.zip** και αποσυμπιέστε το. Χρησιμοποιείστε το πρόγραμμα ***DB Browser for SQLITE*** για να ανοίξετε το αρχείο **football-dataset.sqlite** και στη συνέχεια μελετήστε το. Ακολούθως κάντε τα ακόλουθα SQL ερωτήματα που να βρίσκουν:

a. (0,5M) Όλους τους ποδοσφαιριστές με ύψος από 160cm έως 180cm. Ο πίνακας που θα προκύψει από το ερώτημα να έχει 3 στήλες με το όνομα του παίκτη, το ύψος και το βάρος του. Τέλος οι τιμές να είναι σε φθίνουσα σειρά βάρους.
b. (0,5M) Για έναν αγώνα της επιλογής σας (επιλέξτε match_api_id), την ημερομηνία διεξαγωγής του, την γηπεδούχο ομάδα καθώς και τον αριθμό των goal που σκόραρε η γηπεδούχος ομάδα.
c. (0,5M) Σε ποια χώρα ανήκει η κάθε ομάδα. Θα χρειαστεί να συνδυάσετε 3 πίνακες για να το κάνετε αυτό. Ο τελικός πίνακας θα περιέχει μόνο 2 στήλες (όνομα ομάδας και χώρα) και θα επιστρέφει 296 γραμμές όσες είναι και όλες οι ομάδες. Αντί του SELECT να χρησιμοποιηθεί η εντολή SELECT DISTINCT η οποία επιστρέφει μόνο τις διαφορετικές γραμμές. Παράδειγμα χρήσης του SELECT DISTINCT: https://www.tutorialspoint.com/sqlite/sqlite_distinct_keyword.htm
d. (0,5M) Σε συνέχεια του ερωτήματος b κάντε το ερώτημα έτσι ώστε να επιστρέφει έναν πίνακα ο οποίος πέρα από την ημερομηνία, την γηπεδούχο ομάδα και τα goal που σημείωσε, θα περιέχει την αντίπαλο ομάδα καθώς και τα goal που σημείωσε αυτή.

Τα παραπάνω SQL ερωτήματα πρέπει να τα καταγράψετε σε αρχείο word με τη διαδικασία **αντιγραφή/επικόλληση** από το πρόγραμμα DB Browser for SQLITE. Απαντήσεις που θα περιέχουν μόνο εικόνες των SQL ερωτημάτων ή/και των αποτελεσμάτων τους **δε θα λαμβάνονται υπόψιν**.

Άσκηση 2 (8 Μονάδες)

Η παρακάτω άσκηση θα γραφεί και θα παραδοθεί σε ένα αρχείο .R και θα χρησιμοποιεί τις βιβλιοθήκες tidyverse και DBI.

1. (0,5Μ) Συνδεθείτε στην βάση δεδομένων *football-database.sqlite*
2. (0,5Μ) Διαβάστε όλους τους πίνακες της βάσης και ελέγξτε τη δομή τους
3. (1Μ) Χρησιμοποιήστε την συνάρτηση *tbl()* για να δημιουργήσετε pointers για τους 7 πίνακες (εξαιρέστε τον πίνακα sqlite_sequence) και ακολούθως ελέγξτε τις διαστάσεις τους. Τα ονόματα που θα δώσετε στα pointers να είναι ίδια με τα ονόματα που έχουνε οι πίνακες στη βάση αλλά με μικρά γράμματα (lowecase)
4. (2Μ) Ξαναγράψτε τα ερωτήματα SQL της Άσκησης 1 αλλά αυτή τη φορά με τον τρόπο της *dplyr*
5. (2Μ)
    a. Βρείτε πόσοι παίκτες είναι αριστεροπόδαροι και πόσοι δεξιοπόδαροι (Θα χρησιμοποιήσετε την στήλη *preferred_foot* του πίνακα *player_attributes* καθώς και τις συναρτήσεις *group_by* και *summarize*). Να αφαιρεθούν τα NA εφόσον υπάρχουν.
    b. Χρησιμοποιήστε τη βιβλιοθήκη *ggplot2* και κατασκευάστε ένα απλό ραβδόγραμμα για τη μεταβλητή *preferred_foot* του πίνακα *player_attributes*
6. (2Μ)
    a. Με χρήση της *ggplot2* κάντε ένα διάγραμμα διασποράς για τις μεταβλητές ύψος (*height*) και βάρος (*weight*) των παικτών (πίνακας *player*)
    b. Βελτιώσετε το παραπάνω γράφημα έτσι ώστε να αποφύγετε το overplotting