

HW6 – Like a final exam...!

Σας δίνεται το αρχείο albums.csv (comma delimited). Περιέχει 11461 εγγραφές με μουσικά album με τα εξής πεδία (όλα αλφαριθμητικά):

genre = μουσικό είδος στο οποίο ανήκει ο artist του album
artist = καλλιτέχνης/συγκρότημα
album = (χρονιά) τίτλος album

Σχόλια:

- Υπάρχουν 3692 διακριτοί artists.
- Υπάρχουν 21 διακριτά genres (Avant-Garde, Blues, Children's, Classical, Comedy-Spoken, Country, EasyListening, Electronic, Folk, Holiday, International, Jazz, Latin, NewAge, Pop-Rock, R&B, Rap, Reggae, Religious, Stage&Screen, Vocal). Είναι η ταξινομία που προτείνει το allmusic.com και σύμφωνα με την οποία κατηγοριοποιεί κάθε artist σε ένα ή περισσότερα genres.

Για παράδειγμα:

- "Pop-Rock", "Aerosmith" σημαίνει ότι οι Aerosmith ανήκουν στο είδος Pop-Rock.
- "Blues,Folk,Pop-Rock,R&B,Country", "Brown, Clarence Gatemouth" σημαίνει ότι ο Clarence Gatemouth Brown ανήκει σε πέντε είδη (πρωτίστως στο είδος Blues, δευτερευόντως στο είδος Folk, κ.κ.).

ΠΡΟΣΟΧΗ: Σε όλα τα album ενός artist εμφανίζονται τα ίδια ακριβώς genres. Συνεπώς, τα genre που χαρακτηρίζουν έναν artist, χαρακτηρίζουν και κάθε album του artist.

Προεπεξεργασία. Για να απαντήσετε στα παρακάτω ερωτήματα, αφού κάνετε τους σχετικούς ελέγχους για την ύπαρξη NA, θα πρέπει να μετασχηματίσετε τα αρχικά δεδομένα με όποιο τρόπο θέλετε (SQL, R, python, bash scripting, κλπ). Για παράδειγμα, στον πίνακα albums θα πρέπει να "σπάσετε" την τρίτη στήλη σε δυο διακριτές στήλες με ονόματα year και title. Επίσης, θα πρέπει να δημιουργήσετε έναν νέο πίνακα/data frame με όνομα artists που θα περιέχει τα ονόματα των 3692 artist καθώς και δυο έξτρα στήλες με το πρωτεύον και δευτερεύον genre του κάθε artist (για λεπτομέρειες δείτε παρακάτω στο [4]).

Απαντήστε στα παρακάτω ερωτήματα δίνοντας τα κατάλληλα γραφήματα.

[1] (μονάδες 2) Χρησιμοποιώντας μόνο τον πίνακα albums απαντήστε στα παρακάτω:

- (a) το πλήθος των album ανά genre (υπόθεση: κάθε album ανήκει μόνο στο πρωτεύον genre)
- (b) το πλήθος των album ανά genre (υπόθεση: κάθε album ανήκει σε όλα τα genres του)
- (c) το πλήθος των artist ανά genre (υπόθεση: κάθε artist ανήκει μόνο στο πρωτεύον genre)
- (d) το πλήθος των artist ανά genre (υπόθεση: κάθε artist ανήκει σε όλα τα genres του)

[2] (μονάδες 1) Δώστε την κατανομή των album ανά χρονιά.

[3] (μονάδες 2) Δώστε τις κατανομές των album των top 5 genre στο χρόνο (στο γράφημα να φαίνονται τα genre με διαφορετικά χρώματα). Στο συγκεκριμένο ερώτημα θεωρείστε ως genre του album το πρωτεύον genre του artist.

[4] (μονάδες 3) Χρησιμοποιώντας μόνο τον πίνακα artists, μελετήστε τις σχέσεις ανάμεσα στα genres, δηλαδή ποια ζεύγη έχουν υψηλή συσχέτιση. Ας εστιάσουμε μόνο στο πρωτεύον και δευτερεύον genre των artist.

Προσθέστε στον πίνακα artists τις στήλες primary_genre και secondary_genre και εισάγετε τη σωστή πληροφορία. Για παράδειγμα, για τον Clarence Gatemouth Brown θα βάλετε την τιμή Blues στη στήλη primary_genre και την τιμή Folk στη στήλη secondary_genre. Επίσης, για τους

artist με ένα ακριβώς genre, βάλτε το genre και στις δυο στήλες. Για παράδειγμα, για τους Aerosmith και οι δυο στήλες θα έχουν την τιμή Pop-Rock.

Τώρα είστε έτοιμοι/ες να ακολουθήσετε τις οδηγίες του tutorial περί [Two Way Tables](#) για να οπτικοποιήσετε τις σχέσεις όλων των ζευγών genre (με mosaicplot). Σε αντίθεση με το tutorial, εδώ οι γραμμές και οι στήλες θα περιέχουν τις ίδιες κατηγορίες (τα 21 genres).

[5] (μονάδες 4) Χρησιμοποιήστε το dataset world από τη βιβλιοθήκη spData. Πάρτε τα ονόματα των χωρών και για κάθε χώρα βρείτε ποιοι τίτλοι album την περιέχουν. Προσθέστε αυτή την πληροφορία σε μια νέα στήλη για το κάθε album. Αν στον τίτλο ενός album υπάρχουν πάνω από μια χώρες, κρατήστε μόνο μια (οποιαδήποτε). Προχωρήστε στην παραγωγή χωροπληθικού χάρτη για την οπτικοποίηση του πλήθους των album ανά χώρα (δείτε σχετικά κεφάλαια του βιβλίου "Geocomputation with R")

[6] (μονάδες 3) Δώστε τρεις (3) επιπλέον δικές σας ενδιαφέρουσες οπτικοποιήσεις των δεδομένων.

[7] (μονάδες 2) Ετοιμάστε ένα flexdashboard που σε τρεις καρτέλες να παρουσιάζει τις πιο σημαντικές κατά την άποψή σας διαδραστικές οπτικοποιήσεις των δεδομένων.

Παραδοτέο. Καταθέστε στο openeclass ένα αρχείο zip με τρία τουλάχιστον αρχεία: το **hw6.Rmd** με τις απαντήσεις σας, το αντίστοιχο knitted αρχείο **hw6.html** και το αρχείο **hw6-dashboard.Rmd**. Στην περίπτωση που δεν προεπεξεργαστήκατε τα δεδομένα με R, το αρχείο artists.csv και τα αρχεία με τον σχετικό κώδικα προεπεξεργασίας.