ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΜΥΕ041 - ΠΛΕ081: Διαχείριση Σύνθετων Δεδομένων

(ΕΑΡΙΝΟ ΕΞΑΜΗΝΟ 2022-23)

ΕΡΓΑΣΙΑ 1 - Ιστογράμματα

Προθεσμία: 31 Μαρτίου 2023

Τα ιστογράμματα χρησιμοποιούνται από συστήματα βάσεων δεδομένων για την προσεγγιστική αναπαράσταση της κατανομής τιμών σε πεδία πινάκων. Μπορούν να βοηθήσουν στην προσεγγιστική εκτίμιση της εξόδου κοινών τελεστών (όπως selection). Στην εργασία αυτή θα γράψετε ένα σε γλώσσα της επιλογής σας (C, C++, Java, Python, κλπ.) πρόγραμμα που θα δημιουργεί ιστογράμματα και θα τα χρησιμοποιεί για την εκτίμηση του πλήθους των αποτελεσμάτων τελεστών επιλογής.

Κατεβάστε το αρχείο acs2015_census_tract_data.csv από τη διεύθυνση https://www.kaggle.com/datasets/muonneutrino/us-census-demographic-data

Το αρχείο περιέχει ένα πίνακα σε μορφή csv με δημογραφικά στοιχεία σε διαφορετικές περιοχές των Ηνωμένων Πολιτειών.

Μέρος 1: Ζητείται να δημιουργήσετε ένα equi-width histogram και ένα equi-depth histogram για το **γνώρισμα Income** του πίνακα. Στο (κλασσικό) equi-width histogram όλα τα bins έχουν το ίδιο εύρος τιμών, π.χ. [0,1000), [1000,2000), [2000,3000), κλπ. και κάθε bin έχει τον αριθμό των δεδομένων που πέφτουν μέσα στο αντίστοιχο εύρος τιμών. Στο equi-depth histogram το άθροισμα των δεδομένων που πέφτουν σε κάθε bin είναι το ίδιο. Εν προκειμένω, τα όρια ορίζονται έτσι ώστε ο αριθμός των incomes που πέφτουν σε κάθε διάστημα είναι ο ίδιος. Οπότε τα όρια των bins μπορεί να διαφέρουν. Η αναπαράσταση ενός equi-width histogram είναι ένα διάνυσμα από αριθμούς δεδομένων (ένας αριθμός για κάθε διάστημα τιμών), ενώ η αναπαράσταση ενός equi-depth histogram είναι ένα διάνυσμα ορίων.

Διαβάστε το αρχείο και εξάγετε όλα τα δεδομένα του **γνωρίσματος Income** σε μια εσωτερική αναπαράσταση (π.χ. array, vector, list) του προγράμματός σας. Κατά την ανάγνωση, αγνοήστε γραμμές του CSV αρχείου, όπου το Income δεν καταγράφεται. Τυπώστε τον αριθμό των γραμμών για τις οποίες το Income έχει έγκυρη τιμή (αριθμός). Κατόπιν δημιουργήστε ένα equi-width ιστόγραμμα και ένα equi-depth histogram, ώστε το καθένα να έχει 100 bins. Για το equi-width ιστόγραμμα, θα χρειαστεί να γνωρίζετε την ελάχιστη και τη μέγιστη τιμή του Income, ώστε να χωρίσετε το εύρος τιμών σε 100 ίσα διαστήματα. Τυπώστε τα ιστογράμματα στην οθόνη. Για το equi-width ιστόγραμμα, τυπώστε το εύρος του κάθε bin και κατόπιν τα bins, ενώ για το equi-depth histogram τυπώστε τον αριθμό των στοιχείων σε κάθε bin και κατόπιν τα bins.

Παράδειγμα εξόδου:

```
72901 valid income values
minimum income = 2611.0 maximum income = 248750.0
equiwidth:
range: [2611.00,5072.39), numtuples: 13
range: [5072.39,7533.78), numtuples: 23
range: [7533.78,9995.17), numtuples: 85
...
equidepth:
range: [2611.00,14814.00), numtuples: 729
range: [14814.00,17456.00), numtuples: 729
range: [17456.00,19731.00), numtuples: 729
...
```

Μέρος 2: Στο δεύτερο μέρος θα δοκιμάσετε την ακρίβεια των ιστογραμμάτων στο να εκτιμούν τον αριθμό των δεδομένων που πέφτουν σε ένα διάστημα τιμών [α, β). Με άλλα λόγια, χρησιμοποιούμε ένα ιστόγραμμα για να εκτιμήσουμε πόσες πλειάδες έχει το αποτέλεσμα μιας ερώτησης επιλογής στο πεδίο Income, η οποία έχει σαν συνθήκη το α <= Income < β. Αν το α (ή το β) δεν είναι ακριβώς τα όρια των bins τότε γίνεται εκτίμηση του αποτελέσματος με βάση το ποσοστό των ορίων των bins που καλύπτει το [α, β). Π.χ., αν το bin έχει όρια [1000, 2000) στο οποίο πέφτουν 300 πλειάδες, τότε το διάστημα [1700, 6400) καλύπτει το διάστημα [1700, 2000) του bin δηλ. το 30% του bin, οπότε θεωρούμε ότι 90 πλειάδες (=300*30%) από αυτό το bin είναι μέρος του αποτελέσματος.

Αλλάξτε το πρόγραμμά σας από το πρώτο μέρος, ώστε να παίρνει σαν ορίσματα τα α, β, και να τυπώνει: (1) το εκτιμώμενο αποτέλεσμα με χρήση του equi-width histogram (2) το εκτιμώμενο αποτέλεσμα με χρήση του equi-depth histogram (3) το πραγματικό αποτέλεσμα χρησιμοποιώντας τα ίδια τα δεδομένα. Μέσω πειραματισμού αποφανθείτε για το αν το equi-width histogram υπερτερεί ή υστερεί του equi-depth histogram αφού δοκιμάσετε έναν μεγάλο αριθμό ερωτήσεων με διάφορα εύρη.

Παράδειγμα εξόδου για α=19000 και β=55000:

```
equiwidth estimated results: 39354.366524606026
equidepth estimated results: 39333.939948818304
actual results: 39361
```

Παραδοτέα:

Βάλτε σέ ένα zip αρχείο τα προγράμματά σας και ένα PDF αρχείο, το οποίο θα περιέχει πληροφορίες για τα προγράμματά σας και τη μελέτη του μέρους 2. Υποβάλετε το zip αρχείο σας μέσω turnin στο assignment1@mye041

Οδηγίες για τις υποβολές:

1) Αν χρησιμοποιήσετε Java, το πρόγραμμά σας θα πρέπει να γίνεται compile και να τρέχει και εκτός Eclipse στους υπολογιστές του εργαστηρίου. **Μην χρησιμοποιείτε packages**.
2) Αν χρησιμοποιήσετε Python, μην χρησιμοποιήστε τη βιβλιοθήκη pandas και μην υποβάλετε κώδικα για interactive programming (π.χ. ipython)

3) Υποβάλετε τις εργασίες σας σε ένα **zip** αρχείο (**όχι rar**) το οποίο πρέπει να περιλαμβάνει όλους τους κώδικες καθώς και ένα αρχείο τεκμηρίωσης το οποίο να περιγράφει τη μεθοδολογία σας και να περιλαμβάνει το PDF αρχείο. **Μην υποβάλετε αρχεία δεδομένων**.

4) Μην ξεχνάτε να βάζετε το όνομά σας (σε greeklish) και το ΑΜ σε κάθε αρχείο που υποβάλετε.

5) Ο έλεγχος των προγραμμάτων σας μπορεί να γίνει σε άλλα αρχεία εισόδου από αυτά που σας δίνονται, άρα θα πρέπει ο κώδικάς σας να μην εξαρτάται από τα συγκεκριμένα αρχεία εισόδου που σας δίνονται.