מכללת הדסה, החוג למדעי המחשב

אנליזה של ביג דאטה / חננאל פרל

סמסטר ב', תשפ"ד

תרגיל 4 / 3.05.2024

תאריך הגשה:

יום שלישי 14.05.2024 בשעה 23:00

מטרות התרגיל:

תרגול שפת SQL מתקדם (advance)

הוראות (כמו בתרגיל 3):

הורד את הדאטה בייס:
https://drive.google.com/file/d/1C4MNtoLAAwGWAjBHOUGL1fTn0zo-rkHL/view?usp=sharing
יש להשתמש בטבלאות (tables) בלבד: City, Country, CountryLanguage. אין להשתמש ב views.

לצורך משחק ובדיקות אפשר לפתוח אותו בקלינט הזה (כמובן שיש גם אחרים):
https://sqlitebrowser.org/
פתור התרגילים.
הערה: כשכתוב בשאלות "או" / "וגם" הכוונה במובן הבוליאני OR / AND .

ההגשה צריכה להכיל:
א. קובץ טקסט אחד (.py) שמכיל תוכנית פייטון שניתן להריצה.
ב. קובץ טקסט אחד (.txt) של ההדפסות של הרצה שאתם עשיתם אצלכם.

התוכנית פייטון צריכה לפתוח את הדאטהבייס מקובץ לוקלי בסיפרייה נוכחית (ללא תתי תיקיות וכד), ולהריץ מולו את השאילתות שעונות על השאלות.
עבור כל שאלה יש להדפיס את:
1) מספר התרגיל.
2) את השאילתה עצמה בעימוד לפי הקונבנציה שלמדנו.
3) יש להדפיס גם את התוצאה באופן הבא:
   a. יש להדפיס את כמות השורות שיצאו סה"כ (לכתוב את מספר השורות, לא כולל כותרות כמובן)
   b. ויש להדפיס 5 שורות ראשונות וגם 5 שורות אחרונות של התוצאה (לא להדפיס את כל השורות, רק ראשונות ואחרונות). אם יש בתוצאה פחות מ 10 שורות, אז להדפיס את כל השורות של התוצאה פעם אחת.

<u>הדגשה</u>: השאילתה עצמה צריכה להחזיר את כל השורות לפי התרגיל. רק בהדפסה למסך בפייטון צריך להדפיס כאמור את השורות הראשונות והאחרונות.
השאילתה צריכה להיות כתובה בקונבנציה שלמדנו (בכמה שורות, אותיות גדולות איפה שצריך וכו)
בפייטון אפשר בנוחות לכתוב מחרוזת פרוסה על פני כמה שורות עם שלוש גרשיים """ (ואפשר באופן כזה במחרוזת לכתוב גם " ' ועוד). ככה:

```
query = """
SELECT *
FROM City
"""
```

כדי להדפיס בפייטון ללא הזחות מיותרות, אפשר להשתמש בפקודה textwrap.dedent
https://docs.python.org/3/library/textwrap.html#textwrap.dedent
יש לשים שורה ריקה וקו מפריד באורך 55 '=' (התו =) בין כל שאלה ושאלה.

הערה: אין לכתוב עברית בשאילתות ובקוד.
שאלה שאתם לא יודעים את התשובה עליה, עדיין לכתוב את הקו המפריד ומספר השאלה, אך להשאיר את השאילתה והתוצאה ריקים.

דוגמא להדפסה:

```
=========================================================
Question: 1
The query:

SELECT *
FROM City

Num of rows: 4079
The results:
   ID           Name CountryCode       District  Population
0   1          Kabul         AFG          Kabol     1780000
1   2       Qandahar         AFG       Qandahar      237500
2   3          Herat         AFG          Herat      186800
3   4  Mazar-e-Sharif         AFG          Balkh      127800
4   5      Amsterdam         NLD  Noord-Holland      731200
        ID        Name CountryCode    District  Population
4074  4075  Khan Yunis         PSE  Khan Yunis      123175
4075  4076      Hebron         PSE      Hebron      119401
4076  4077    Jabaliya         PSE  North Gaza      113901
4077  4078      Nablus         PSE      Nablus      100231
4078  4079       Rafah         PSE       Rafah       92020

=========================================================
Question: 2
The query:


=========================================================
Question: 3
..
..
```

השאלות לא בהכרח ממוינות לפי סדר קושי..

בהצלחה!

השאלות SQL 4:

1) עבור כל מדינה להראות את אחוז השטח שלה מתוך כלל שטח היבשת בה היא נמצאת (באמצעות פונקציית חלון). ממוין לפי שם יבשת סדר עולה ואז אחוז השטח בסדר יורד.
2) עבור כל הערים להראות את שם העיר, קוד מדינה, אוכלוסיה, ועמודה נוספת: האם האכלוסיה מעל הממוצע של כל הערים באותו מדינה. ממוין לפי קוד מדינה ואז אוכלוסיה.
3) להציג את כל טבלת המדינות, להוסיף עמודה בהתחלה שמראה מספור לפי שטח המדינה בסדר יורד (ז"א המדינה עם השטח הכי גדול מספר 1 וכו) למיין לפי עמודה זו.
4) להציג את כל טבלת המדינות, להוסיף עמודה בהתחלה שמראה מספור לפי שנת העצמאות של המדינה בסדר עולה, אם שנת העצמאות זהה יקבלו מספור דומה (למשל כזה 1,1,1,2,2,3,4,4). למיין לפי עמודה זו ואז לפי קוד מדינה.
5) להציג את רשימת המדינות יחד עם מספר דירוג (הכי גדול מס 1) של כל מדינה לפי מספר הערים שיש לה. למיין בסוף לפי הדירוג בסדר יורד.
6) להציג את הסכום המתגלגל של האוכלוסייה עבור כל מדינה, ממוין לפי אוכלוסיה בסדר יורד, להוסיף עמודה של כמה הסכום המתגלל הזה באחוזים מסך כל אוכלסיית העולם.
7) מה כמות המדינות המינימלית שיש להם לפחות 50 אחוז מאוכלסיית העולם?
8) הראה את המדינות הללו מהשאלה הקודמת.
9) עבור כל קוד מדינה להראות את שני השפות עם האחוז הכי גבוה.
10) להראות את טבלת הערים ממוינת לפי ID, להוסיף שתי עמודות: אחת עם הדירוג (הכי גדול מס 1) של כל עיר ביחס למדינתה על סמך האוכלוסייה, והשנייה עם הדירוג של כל עיר ביחס לכולן על סמך האוכלוסייה.
11) אם נמיין את כל המדינות לפי תוחלת חיים. כמה פעמים יהיה פער בין התוחלות חיים שבשני שורות סמוכות גדול ממש מ: 1 ?
12) להראות את כל המדינות שמבחינה קלנדרית יש מדינות אחרות עם ימי עצמאות גם בשנה אחת לפניהם וגם שנתיים לפניהם. למשל מדינת ישראל שנת עצמאות 1948 – אז אנו נראה אותה רק אם מתמלאים שני התנאים: קיימת לפחות מדינה אחת שיש לה עצמאות בשנת 1947 וגם קיימת לפחות מדינה אחת שיש לה עצמאות בשנת 1946.
13) נקח את כל המדינות ששנת עצמאותם גדולה ממש מ 1800. נסתכל על רשימת המדינות ממוינת לפי שנות עצמאות (כדי שהמיון יהיה יחודי נמיין לפי שנת עצמאות וקוד מדינה). אם יחסית ברשימה הממוינת נראה בין שתי מדינות סמוכות פער שנים גדול ממש מ: 5, אז נגדיר את זה כ"פער גדול". המשימה היא להציג את כל המדינות החל מהפער הגדול השביעי.

14) שאלה כללית:

- יש לכתוב **שלוש** (**3**) שאלות אינטלגנטיות ורלוונטיות על הדאטה.
- יש לכתוב את השאילתות שעונות על השאלות שיצרתם.
- יש להתייחס לכל שלושת הטבלאות: City, Country, CountryLanguage.
- יש להגיש את השאלות כתובות בעברית בקובץ וורד. ליד כל שאלה לרשום את השאילתה שעונה עליה.

- בנוסף בקובץ פייטון שאתם מגישים יש להוסיף את השאילתות שלכם עם ההדפסות כמו שמוסבר למעלה (כולל מספור, כמות שורות, 5 שורות בהתחלה ובסוף וכו בדיוק כמו קודם).
- יש לוודא שימוש נכון ונקי בפקודות (לא לאלץ שימוש בפקודה סתם).
- יש לוודא שהשאילתות מעומדות יפה.
- יש להשתמש בשאילתות במגוון הפעולות שלמדנו. אין שצורך שכל שאילתה תכיל את הכל. צריך לוודא שבסה"כ בכל שלושת השאילתות מופיעים הסעיפים הבאים:
  - כל אלו: WITH | SELECT | FROM
  - LEFT JOIN או JOIN
  - לפחות 2 מתוך אלו: AND | OR | LIKE | IN | IS
  - CASE WHEN
  - לפחות 1 פונקציה לטיפול ב strings (למשל: trim upper lower substr replace וכו)
  - GROUP BY | HAVING
  - לפחות 2 פונקציות אגרגציה (למשל: MIN MAX AVG COUNT וכו)
  - OVER | PARTITION | BETWEEN
  - לפחות 1 מאלו: ROWS | GROUPS | RANGE
  - לפחות 1 מאלו: CURRENT ROW | FOLLOWING | UNBOUNDED | PRECEDING
  - לפחות 2 פונקציות חלון:
    - לפחות אחת מסוג אגרגציה (למשל: MIN MAX AVG COUNT וכו)
    - לפחות אחת מסוג דירוג (למשל: LAG LEAD RANK ROW_NUMBER וכו)
  - ORDER BY
  - -- להוסיף הערות בשאילתה להסביר אותה
  - אופציונלי: UNION | INTERSECT | EXCEPT
  - אופציונלי: OFFSET | LIMIT | FILTER