karan.shah@stonybrook.edu ▾

Theory of Database Systems
CSE532
Spring 2020

[Stony Brook University]

Home

CSE532

Home
Syllabus
Schedule
Homework
Textbooks
Examples
Instructor
Resources

Homework >

Homework1: Opioid Pills Analysis

In this homework, we will use a dataset from Washington Post for opioid pills analysis.
Read the story: Drilling into the DEA's pain pill database
The New York State dataset can be downloaded here (login with SBU email). The zip file size is 323MB, The unzipped dataset is 4.72GB.
The database represents transaction of opioid pills sales by a buyer sold by a provider. We mainly analyze the transactions by the buyers.

Total points: 15 (1 extra credit). Due: Feb 24, midnight (11:59pm)

1. Create DEA table, load the data, and create indexes.

1a (2 points**). Create a DEA New York database table with following information** (column descriptions) **and put the** SQL into a file createdeatable.sql:
cse532.dea_ny(REPORTER_DEA_NO,REPORTER_BUS_ACT,REPORTER_NAME,REPORTER_ADDL_CO_INFO,REPORTER_ADDRESS1,REPORTER_ADDRESS2,REPORTER_CI
Note that TRANSACTION_DATE needs to be a DATE format. Please make sure your datatypes can support the queries below.
To save space and improve performance, you can use the COMPRESS option for your table. e.g., create my table(...) COMPRESS YES.

1b (1)**. Load the csv file into the database by modifying the loading script. Run it as:**
db2 -tf load.sql

Validate there were 4,281,954,931 prescription pain pills supplied to New York.
(DOSAGE_UNIT is the total number of pills in a transaction/record.)

1c (2). Based on queries in 3, create proper indexes to make the queries more efficient (createdeaindexes.sql). Attributes used in predicates and used in GROUP BY are candidates for indexes.

2. Create and load data for zip code population table (the population in each zip code).

2a (1 points)**. Create a table CSE532.ZIPPOP (ZIP, COUNTY, GEOID, ZPOP) on populations of zip codes in NY (**createzip.sql**).** "ZIP" is the zip code, and "ZPOP" is the population of the zip code. Note that there are duplicated zip codes in the table. Some zip codes have zero populations.

2b (1 point)**. Create a loading script to load the csv file to the database (**zipload.sql**)**.

3. Write SQL queries. (Note that we are only querying pills by buyers.)

3a (5 points). **Return monthly counts of pills and smooth counts of pills with a two-month window (preceding one month, following one month) (**q3a.sql**).** DOSAGE_UNIT is the total number of pills in a transaction.
Draw the two curves with Excel or other drawing tools to show the difference (q3a.jpg**).**
Note:
1) month() and year() functions extract month and year from a DATE datatype respectively.
2) You may use CONCAT or "||" to concatenate two values. To generate a uniform "yeardate" representation (e.g., 200801), you may also consider to use CASE function. (example case function)
 3) Consider to use common table expression.

3b (5 points)**. Find the top five zip codes with most pills sold in terms of MME when normalized by the population in the zip codes, i.e., zip codes with most sold total MME of pills per person (**q3b.sql**).**
(**Extra 1 point credit** if you write the query with the RANK function).
(MME is a normalized amount which better represents the total strength of the pills sold in the transaction.)

Homework Submission
Please zip your SQL scripts and results:
createdeatable.sql
load.sql
createdeaindexes.sql
createzip.sql
zipload.sql
q3a.sql (q3a.jpg and result as a text file or screenshot)
q3b.sql (and result as a text file or screenshot)

A readme.txt file explaining anything not included.

Please go to blackboard, and submit the zip file under homework 1.

Subpages (1):   examplecase

WPDEADatabase.pdf (2538k)	Fusheng Wang, Feb 13, 2020, 7:32 AM	v.1		⬇
load.sql (0k)	Fusheng Wang, Feb 13, 2020, 3:21 PM	v.1		⬇
nyzip.csv (912k)	Fusheng Wang, Feb 13, 2020, 3:25 PM	v.1		⬇
schemas.xlsx (14k)	Fusheng Wang, Feb 13, 2020, 3:16 PM	v.1		⬇

Report Abuse | Remove Access | Powered By **Google Sites**