**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](https://docs.google.com/a/stonybrook.edu/viewer?a=v&pid=sites&srcid=c3Rvbnlicm9vay5lZHV8Y3NlNTMyLXMyMHxneDo2MDYwZDU1M2IyNDEwNzll)) 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\_CITY,REPORTER\_STATE,REPORTER\_ZIP,REPORTER\_COUNTY,BUYER\_DEA\_NO,BUYER\_BUS\_ACT,BUYER\_NAME,BUYER\_ADDL\_CO\_INFO,BUYER\_ADDRESS1,BUYER\_ADDRESS2,BUYER\_CITY,BUYER\_STATE,BUYER\_ZIP,BUYER\_COUNTY,TRANSACTION\_CODE,DRUG\_CODE,NDC\_NO,DRUG\_NAME,QUANTITY,UNIT,ACTION\_INDICATOR,ORDER\_FORM\_NO,CORRECTION\_NO,STRENGTH,TRANSACTION\_DATE,CALC\_BASE\_WT\_IN\_GM,DOSAGE\_UNIT,TRANSACTION\_ID,Product\_Name,Ingredient\_Name,Measure,MME\_Conversion\_Factor,Combined\_Labeler\_Name,Reporter\_family,dos\_str,MME).

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**](https://docs.google.com/a/stonybrook.edu/viewer?a=v&pid=sites&srcid=c3Rvbnlicm9vay5lZHV8Y3NlNTMyLXMyMHxneDoyZjBlMjI3ZDZiNWIwNWJk)**. 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**](https://docs.google.com/a/stonybrook.edu/viewer?a=v&pid=sites&srcid=c3Rvbnlicm9vay5lZHV8Y3NlNTMyLXMyMHxneDoyMTk5ZTg2NjAwZGExZTY4)**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](https://sites.google.com/a/stonybrook.edu/cse532-s20/homework/homework1-2/examplecase))

 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.)