
[Home](#)

[Home](#)
[Syllabus](#)
[Schedule](#)
[Homework](#)
[Textbooks](#)
[Examples](#)
[Instructor](#)
[Resources](#)

[Homework](#) >

Homework1: Queries of DEA Opioid Pill Database

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 database represents transaction of opioid pills sales by a buyer sold by a provider. We mainly analyze the transactions by the buyers.

In this homework, we will build a database for California and Kentucky, and compare the differences.

The California State dataset can be downloaded [here](#). The zip file size is 489MB, and the unzipped dataset is 7.66GB.

The Kentucky State dataset can be downloaded [here](#). The zip file size is 135MB, and the unzipped dataset is 2.15GB.

In addition, we have a [US zip code database](#), which includes population data for each zip code. The file is available to [download](#).

Total points: 20. Due: Sep 27, midnight (11:59pm)

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

1a (1 points). Create a DEA database table with following information ([column descriptions](#) in schemas.xlsx) and put the SQL into a file createdeatable.sql:

```
cse532.dea(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:

TRANSACTION_DATE needs to be a DATE format. Please make sure your datatypes can support the queries in Q3.

To save space and improve performance, you can use the COMPRESS option for your table. e.g., create mytable(...) COMPRESS YES.

1b (1 points). Load the csv file into the database by modifying the [loading script](#). Run it as:

```
db2 -tf load.sql
```

Validate there were 2,452,373,586 prescription pain pills supplied to Kentucky, and ~~10,573,780,785~~ 10,573,793,385 prescription pills supplied to California.

(DOSAGE_UNIT is the total number of pills in a transaction/record.)

1c (1 point). 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.ZIP (zip,state,county,pop) to represent the population of zip codes, where "pop" refers to "irs_estimated_population_2015" column in the [zip_code_database.csv](#) file ([createzip.sql](#)). The documentation of the zip code database can be found at [here](#).

Note that some zip codes there are duplicated zip codes in the table. Some zip codes have zero populations as they may represent a special administration or military region.

2b (1 point). Create a loading script to load the [csv file](#) to the database ([zipload.sql](#)). Note that you have to specify columns to be loaded. Refer to [load documentation](#) on how to specify columns to be loaded using "METHOD P".

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

3a (5 points). Return counts of pills for each county (of buyer), state, by month and by year. (q3a.sql). DOSAGE_UNIT is the total number of pills in a transaction. Save result in q3a.txt.

Note: month() and year() functions extract month and year from a DATE datatype respectively.

3b (5 points). **Return monthly smoothed counts of pills with a three-month history window (preceding three months) of Pike County in Kentucky** (q3b.sql). DOSAGE_UNIT is the total number of pills in a transaction. Save result in q3b.txt.

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.

3c (5 points). **Find the top 10 zip codes (together with their corresponding counties) 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.** Please use RANK function. (q3c.sql)

Note:

A zip code be be associated with multiple counties. You may want to check if that is true before writing the query. 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 (and result q3a.txt)
- q3b.sql (and result q3b.txt)
- q3c.sql (and result q3c.txt)

A readme.txt file (optional) 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.2	
 load.sql (0k)	Fusheng Wang, Sep 13, 2020, 8:50 PM	v.4	
 schemas.xlsx (14k)	Fusheng Wang, Feb 13, 2020, 3:16 PM	v.2	
 zip_code_database.csv (4358k)	Fusheng Wang, Sep 13, 2020, 4:38 PM	v.1	

Copyrights by Dr. Fusheng Wang