DB0201EN-PeerAssign-v5_SQLite

May 22, 2022

Assignment: Notebook for Peer Assignment

1 Introduction

Using this Python notebook you will:

1. Understand three Chicago datasets
2. Load the three datasets into three tables in a Db2 database
3. Execute SQL queries to answer assignment questions

1.1 Understand the datasets

To complete the assignment problems in this notebook you will be using three datasets that are available on the city of Chicago's Data Portal:

1. Socioeconomic Indicators in Chicago
2. Chicago Public Schools
3. Chicago Crime Data

1.1.1 1. Socioeconomic Indicators in Chicago

This dataset contains a selection of six socioeconomic indicators of public health significance and a "hardship index," for each Chicago community area, for the years 2008 – 2012.

A detailed description of this dataset and the original dataset can be obtained from the Chicago Data Portal at: https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

1.1.2 2. Chicago Public Schools

This dataset shows all school level performance data used to create CPS School Report Cards for the 2011-2012 school year. This dataset is provided by the city of Chicago's Data Portal.

A detailed description of this dataset and the original dataset can be obtained from the Chicago Data Portal at: https://data.cityofchicago.org/Education/Chicago-Public-Schools-Progress-Report-Cards-2011-/9xs2-f89t

1.1.3 3. Chicago Crime Data

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days.

A detailed description of this dataset and the original dataset can be obtained from the Chicago Data Portal at: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

1.1.4 Download the datasets

This assignment requires you to have these three tables populated with a subset of the whole datasets.

In many cases the dataset to be analyzed is available as a .CSV (comma separated values) file, perhaps on the internet. Click on the links below to download and save the datasets (.CSV files):

- Chicago Census Data
- Chicago Public Schools
- Chicago Crime Data

NOTE: Ensure you have downloaded the datasets using the links above instead of directly from the Chicago Data Portal. The versions linked here are subsets of the original datasets and have some of the column names modified to be more database friendly which will make it easier to complete this assignment.

1.1.5 Store the datasets in database tables

To analyze the data using SQL, it first needs to be loaded into SQLite DB. We will create three tables in as under:

1. **CENSUS_DATA**
2. **CHICAGO_PUBLIC_SCHOOLS**
3. **CHICAGO_CRIME_DATA**

Let us now load the ipython-sql extension and establish a connection with the database

- Here you will be loading the csv files into the pandas Dataframe and then loading the data into the above mentioned sqlite tables.
- Next you will be connecting to the sqlite database **FinalDB**.

Refer to the previous lab for hints .

Hands-on Lab: Analyzing a real World Data Set

[2]:
```
import csv, sqlite3

con = sqlite3.connect("RealWorldData2.db")
cur = con.cursor()
```

[3]:
```
%load_ext sql
```

[4]:
```
!pip install -q pandas==1.1.5
```

[5]:
```
%sql sqlite:///RealWorldData2.db
```

[5]:
```
'Connected: @RealWorldData2.db'
```

[7]:
```
import pandas

df = pandas.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.
 ↪appdomain.cloud/IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork/labs/
 ↪FinalModule_Coursera_V5/data/ChicagoCensusData.csv")
df.to_sql("CENSUS_DATA", con, if_exists='replace', index=False,method="multi")

df = pandas.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.
 ↪appdomain.cloud/IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork/labs/
 ↪FinalModule_Coursera_V5/data/ChicagoCrimeData.csv")
df.to_sql("CHICAGO_CRIME_DATA", con, if_exists='replace', index=False,␣
 ↪method="multi")

df = pandas.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.
 ↪appdomain.cloud/IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork/labs/
 ↪FinalModule_Coursera_V5/data/ChicagoPublicSchools.csv")
df.to_sql("CHICAGO_PUBLIC_SCHOOLS", con, if_exists='replace', index=False,␣
 ↪method="multi")
```

[8]:
```
#Make sure the tables got loaded:
%sql SELECT name FROM sqlite_master WHERE type='table'
```

```
 * sqlite:///RealWorldData2.db
Done.
```

[8]:
```
[('CENSUS_DATA',), ('CHICAGO_CRIME_DATA',), ('CHICAGO_PUBLIC_SCHOOLS',)]
```

1.2 Problems

Now write and execute SQL queries to solve assignment problems

1.2.1 Problem 1

Find the total number of crimes recorded in the CRIME table.

[9]:
```
%sql select count(*) from CHICAGO_CRIME_DATA
```

```
 * sqlite:///RealWorldData2.db
Done.
```

[9]:
```
[(533,)]
```

1.2.2 Problem 2

List community areas with per capita income less than 11000.

```
[10]: %sql select COMMUNITY_AREA_NAME, COMMUNITY_AREA_NUMBER,  PER_CAPITA_INCOME from␣
       ↪CENSUS_DATA \
      where PER_CAPITA_INCOME < 11000;
```

```
 * sqlite:///RealWorldData2.db
Done.
```

```
[10]: [('West Garfield Park', 26.0, 10934),
       ('South Lawndale', 30.0, 10402),
       ('Fuller Park', 37.0, 10432),
       ('Riverdale', 54.0, 8201)]
```

1.2.3 Problem 3

List all case numbers for crimes involving minors?(children are not considered minors for the purposes of crime analysis)

```
[11]: %sql select CASE_NUMBER, DESCRIPTION from CHICAGO_CRIME_DATA \
      where UPPER(DESCRIPTION) like '%MINOR%';
```

```
 * sqlite:///RealWorldData2.db
Done.
```

```
[11]: [('HL266884', 'SELL/GIVE/DEL LIQUOR TO MINOR'),
       ('HK238408', 'ILLEGAL CONSUMPTION BY MINOR')]
```

1.2.4 Problem 4

List all kidnapping crimes involving a child?

```
[12]: %sql select * from CHICAGO_CRIME_DATA where PRIMARY_TYPE = 'KIDNAPPING' \
      AND UPPER(DESCRIPTION) like '%CHILD%';
```

```
 * sqlite:///RealWorldData2.db
Done.
```

```
[12]: [(5276766, 'HN144152', '2007-01-26', '050XX W VAN BUREN ST', '1792',
      'KIDNAPPING', 'CHILD ABDUCTION/STRANGER', 'STREET', 0, 0, 1533, 15, 29.0, 25.0,
      '20', 1143050.0, 1897546.0, 2007, 41.87490841, -87.75024931, '(41.874908413,
      -87.750249307)')]
```

1.2.5 Problem 5

What kinds of crimes were recorded at schools?

```
[13]: %sql select distinct(PRIMARY_TYPE) from CHICAGO_CRIME_DATA \
      where UPPER(LOCATION_DESCRIPTION) like '%SCHOOL%';
```

```
 * sqlite:///RealWorldData2.db
Done.
```

```
[13]: [('BATTERY',),
       ('CRIMINAL DAMAGE',),
       ('NARCOTICS',),
       ('ASSAULT',),
       ('CRIMINAL TRESPASS',),
       ('PUBLIC PEACE VIOLATION',)]
```

1.2.6 Problem 6

List the average safety score for each type of school.

```
[14]: #Use double quotes for the mixed case column

      %sql select "Elementary, Middle, or High School", AVG(Safety_Score) from␣
       ↪CHICAGO_PUBLIC_SCHOOLS \
      group by "Elementary, Middle, or High School"
```

```
 * sqlite:///RealWorldData2.db
Done.
```

```
[14]: [('ES', 49.52038369304557), ('HS', 49.62352941176471), ('MS', 48.0)]
```

1.2.7 Problem 7

List 5 community areas with highest % of households below poverty line

```
[15]: %sql select COMMUNITY_AREA_NAME, PERCENT_HOUSEHOLDS_BELOW_POVERTY from␣
       ↪CENSUS_DATA \
      order by PERCENT_HOUSEHOLDS_BELOW_POVERTY desc limit 5;
```

```
 * sqlite:///RealWorldData2.db
Done.
```

```
[15]: [('Riverdale', 56.5),
       ('Fuller Park', 51.2),
       ('Englewood', 46.6),
       ('North Lawndale', 43.1),
       ('East Garfield Park', 42.4)]
```

1.2.8 Problem 8

Which community area is most crime prone?

```
[23]: %sql select CR.COMMUNITY_AREA_NUMBER, CE.COMMUNITY_AREA_NAME, count(CR.ID) as␣
       ↪CRIME_COUNT \
      from CHICAGO_CRIME_DATA CR, CENSUS_DATA CE where CR.COMMUNITY_AREA_NUMBER = CE.
       ↪COMMUNITY_AREA_NUMBER \
      group by CR.COMMUNITY_AREA_NUMBER order by CRIME_COUNT desc limit 1;
```

```
 * sqlite:///RealWorldData2.db
Done.
```

```
[23]: [(25.0, 'Austin', 43)]
```

Double-click **here** for a hint

1.2.9 Problem 9

Use a sub-query to find the name of the community area with highest hardship index

```
[17]: %sql select COMMUNITY_AREA_NAME, HARDSHIP_INDEX \
      from CENSUS_DATA where HARDSHIP_INDEX = \
      (select MAX(HARDSHIP_INDEX) from CENSUS_DATA)
```

```
 * sqlite:///RealWorldData2.db
Done.
```

```
[17]: [('Riverdale', 98.0)]
```

1.2.10 Problem 10

Use a sub-query to determine the Community Area Name with most number of crimes?

```
[55]: ##Put the subquery in the from clause! Tried it in the where clause but had no␣
        ↪success.
      ##Also, I only got this to work by ONLY referencing the crime table within the␣
        ↪subsuery and not again separately

      %sql select CRI.COMMUNITY_AREA_NUMBER, CE.COMMUNITY_AREA_NAME, max(CRIME_COUNT)␣
        ↪as MAX_CRIME_COUNT \
      from CENSUS_DATA CE, \
      (select CR.COMMUNITY_AREA_NUMBER, count(CR.ID) as CRIME_COUNT \
       from CHICAGO_CRIME_DATA CR group by CR.COMMUNITY_AREA_NUMBER) CRI\
      where CRI.COMMUNITY_AREA_NUMBER = CE.COMMUNITY_AREA_NUMBER
```

```
 * sqlite:///RealWorldData2.db
Done.
```

```
[55]: [(25.0, 'Austin', 43)]
```

1.3 Author(s)

Hima Vasudevan

Rav Ahuja

Ramesh Sannreddy

1.4 Contribtuor(s)

Malika Singla

1.5 Change log

Date	Version	Changed by	Change Description
2022-03-04	2.5	Lakshmi Holla	Changed markdown.
2021-05-19	2.4	Lakshmi Holla	Updated the question
2021-04-30	2.3	Malika Singla	Updated the libraries
2021-01-15	2.2	Rav Ahuja	Removed problem 11 and fixed changelog
2020-11-25	2.1	Ramesh Sannareddy	Updated the problem statements, and datasets
2020-09-05	2.0	Malika Singla	Moved lab to course repo in GitLab
2018-07-18	1.0	Rav Ahuja	Several updates including loading instructions
2018-05-04	0.1	Hima Vasudevan	Created initial version

##