Introduction

Using this Python notebook you will:

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

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. 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

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

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

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.

Use the links below to read the data files using the Pandas library.

- Chicago Census Data

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork/labs/FinalModule_Coursera_V5/data/ChicagoCensusData.csv?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDeveloperSkillsNetworkDB0201ENSkillsNetwork20127838-2021-01-01

- Chicago Public Schools

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork/labs/FinalModule_Coursera_V5/data/ChicagoPublicSchools.csv?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDeveloperSkillsNetworkDB0201ENSkillsNetwork20127838-2021-01-01

- Chicago Crime Data

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork/labs/FinalModule_Coursera_V5/data/ChicagoCrimeData.csv?utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&utm_term=10006555&utm_id=NA-SkillsNetwork-Channel-SkillsNetworkCoursesIBMDeveloperSkillsNetworkDB0201ENSkillsNetwork20127838-2021-01-01

NOTE: Ensure you use the datasets available on 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.

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**

Load the `pandas` and `sqlite3` libraries and establish a connection to `FinalDB.db`

```
import csv, sqlite3

con = sqlite3.connect("RealWorldData.db")
cur = con.cursor()

!pip install -q pandas==1.1.5
```

Load the SQL magic module

```
%load_ext sql
```

```
The sql extension is already loaded. To reload it, use:
  %reload_ext sql
```

Use `Pandas` to load the data available in the links above to dataframes. Use these dataframes to load data on to the database `FinalDB.db` as required tables.

```
%sql sqlite:///RealWorldData.db
```

```
'Connected: @RealWorldData.db'
```

Establish a connection between SQL magic module and the database `FinalDB.db`

```
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?
utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&u
tm_term=10006555&utm_id=NA-SkillsNetwork-Channel-
SkillsNetworkCoursesIBMDeveloperSkillsNetworkDB0201ENSkillsNetwork2012
7838-2021-01-01.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?
utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&u
tm_term=10006555&utm_id=NA-SkillsNetwork-Channel-
SkillsNetworkCoursesIBMDeveloperSkillsNetworkDB0201ENSkillsNetwork2012
7838-2021-01-01.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?
utm_medium=Exinfluencer&utm_source=Exinfluencer&utm_content=000026UJ&u
tm_term=10006555&utm_id=NA-SkillsNetwork-Channel-
SkillsNetworkCoursesIBMDeveloperSkillsNetworkDB0201ENSkillsNetwork2012
7838-2021-01-01.csv")
df.to_sql("CHICAGO_PUBLIC_SCHOOLS_DATA", con, if_exists='replace',
index=False, method="multi")
```

You can now proceed to the the following questions. Please note that a graded assignment will follow this lab and there will be a question on each of the problems stated below. It can be from the answer you received or the code you write for this problem. Therefore, please keep a note of both your codes as well as the response you generate.

Problems

Now write and execute SQL queries to solve assignment problems

Problem 1

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

```
%sql select count(*) from CHICAGO_CRIME_DATA
```

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

```
[(533,)]
```

Problem 2

List community area names and numbers with per capita income less than 11000.

```
%sql SELECT COMMUNITY_AREA_NUMBER, COMMUNITY_AREA_NAME FROM
CENSUS_DATA WHERE PER_CAPITA_INCOME<11000
```

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

```
[(26.0, 'West Garfield Park'),
 (30.0, 'South Lawndale'),
 (37.0, 'Fuller Park'),
 (54.0, 'Riverdale')]
```

Problem 3

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

```
%sql SELECT DISTINCT CASE_NUMBER FROM CHICAGO_CRIME_DATA WHERE
DESCRIPTION LIKE '%minor%'
```

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

[('HL266884',), ('HK238408',)]
```

Problem 4

List all kidnapping crimes involving a child?

```
%sql SELECT * FROM CHICAGO_CRIME_DATA WHERE PRIMARY_TYPE =
"KIDNAPPING" AND DESCRIPTION LIKE '%child%'

 * sqlite:///RealWorldData.db
Done.

[(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)')]
```

Problem 5

List the kind of crimes that were recorded at schools. (No repetitions)

```
%sql select DISTINCT (PRIMARY_TYPE) from CHICAGO_CRIME_DATA where
LOCATION_DESCRIPTION LIKE '%school%'

 * sqlite:///RealWorldData.db
Done.

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

Problem 6

List the type of schools along with the average safety score for each type.

```
%sql SELECT "Elementary, Middle, or High School" AS SCHOOL_TYPE,
AVG(SAFETY_SCORE) AS AVERAGE_SAFETY_SCORE \
     FROM CHICAGO_PUBLIC_SCHOOLS_DATA GROUP BY "Elementary, Middle, or
High School"

 * sqlite:///RealWorldData.db
Done.

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

Problem 7

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

```
%sql SELECT COMMUNITY_AREA_NAME, PERCENT_HOUSEHOLDS_BELOW_POVERTY \
     from CENSUS_DATA \
     order by PERCENT_HOUSEHOLDS_BELOW_POVERTY desc \
     Limit 5

 * sqlite:///RealWorldData.db
Done.

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

Problem 8

Which community area is most crime prone? Display the coumminty area number only.

```
%sql SELECT PRIMARY_TYPE, COMMUNITY_AREA_NUMBER from
CHICAGO_CRIME_DATA GROUP BY COMMUNITY_AREA_NUMBER \
     ORDER BY COMMUNITY_AREA_NUMBER desc LIMIT 1

 * sqlite:///RealWorldData.db
Done.

[('THEFT', 77.0)]
```

Double-click **here** for a hint

Problem 9

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

```
%sql SELECT COMMUNITY_AREA_NAME FROM CENSUS_DATA WHERE HARDSHIP_INDEX
IN (SELECT MAX(HARDSHIP_INDEX) FROM CENSUS_DATA)

 * sqlite:///RealWorldData.db
Done.

[('Riverdale',)]
```

Problem 10

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

```
%%sql
SELECT community_area_name FROM CENSUS_DATA
WHERE COMMUNITY_AREA_NUMBER = (SELECT COMMUNITY_AREA_NUMBER FROM
```

```
CHICAGO_CRIME_DATA
    GROUP BY COMMUNITY_AREA_NUMBER
    ORDER BY COUNT(COMMUNITY_AREA_NUMBER) DESC
    LIMIT 1)
LIMIT 1
```

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

```
[('Austin',)]
```

Author(s)

Contribtuor(s)

Change log

Date	Version	Changed by	Change Description
2023-10-18	2.6	Abhishek Gagneja	Modified instruction set
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