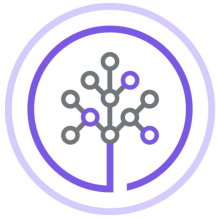

Working with a real world data-set using SQL and Python

Estaimted time needed: **30** minutes

Objectives

After complting this lab you will be able to:

- Understand the dataset for Chicago Public School level performance
- Store the dataset in SQLite database.
- Retrieve metadata about tables and columns and query data from mixed case columns
- Solve example problems to practice your SQL skills including using built-in database functions

Chicago Public Schools - Progress Report Cards (2011-2012)

The city of Chicago released a dataset showing all school level performance data used to create School Report Cards for the 2011-2012 school year. The dataset is available from the Chicago Data Portal: https://data.cityofchicago.org/Education/Chicago-Public-Schools-Progress-Report-Cards-2011-/9xs2-f89t

This dataset includes a large number of metrics. Start by familiarizing yourself with the types of metrics in the database: https://data.cityofchicago.org/api/assets/AAD41A13-BE8A-4E67-B1F5-86E711E09D5F?download=true

NOTE:

Do not download the dataset directly from City of Chicago portal. Instead download a static copy which is a more database friendly version from this link.

Now review some of its contents.

Connect to the database

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

The syntax for connecting to magic sql using sqllite is

%sql sqlite://DatabaseName

where DatabaseName will be your **.db** file

In [6]:
```
import csv, sqlite3

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

In [7]:
```
!pip install -q pandas==1.1.5
```

In [8]:
```
%load_ext sql
```

In [9]:
```
%sql sqlite:///RealWorldData.db
```

Out[9]:
```
'Connected: @RealWorldData.db'
```

Store the dataset in a Table

In many cases the dataset to be analyzed is available as a .CSV (comma separated values) file, perhaps on the internet. To analyze the data

using SQL, it first needs to be stored in the database.

We will first read the csv files from the given url into pandas dataframes

Next we will be using the df.to_sql() function to convert each csv file to a table in sqlite with the csv data loaded in it.

In [11]:
```
import pandas
df = pandas.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DE
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-DE
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-DE
df.to_sql("CHICAGO_PUBLIC_SCHOOLS_DATA", con, if_exists='replace', index=False, method="multi")
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
/tmp/ipykernel_68/185077871.py in <module>
      1 import pandas
      2 df = pandas.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNet
work-DB0201EN-SkillsNetwork/labs/FinalModule_Coursera_V5/data/ChicagoCensusData.csv")
----> 3 df.to_sql("CENSUS_DATA", con, if_exists='replace', index=False,method="multi")
      4
      5 df = pandas.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNet
work-DB0201EN-SkillsNetwork/labs/FinalModule_Coursera_V5/data/ChicagoCrimeData.csv")

NameError: name 'con' is not defined
```

In [12]:
```
import pandas
df = pandas.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DE
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-DE
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-DE
df.to_sql("CHICAGO_PUBLIC_SCHOOLS_DATA", con, if_exists='replace', index=False, method="multi")
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
/tmp/ipykernel_68/185077871.py in <module>
      1 import pandas
      2 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")
----> 3 df.to_sql("CENSUS_DATA", con, if_exists='replace', index=False,method="multi")
      4
      5 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")

NameError: name 'con' is not defined
```

Double-click **here** for the solution.

Query the database system catalog to retrieve table metadata

You can verify that the table creation was successful by retrieving the list of all tables in your schema and checking whether the SCHOOLS table was created

In [13]:
```
# type in your query to retrieve list of all tables in the database
```

```
  File "/tmp/ipykernel_68/4075914494.py", line 1
    Select * from syscat.tables# type in your query to retrieve list of all tables in the database
               ^
SyntaxError: invalid syntax
```

In [36]:
```
select * from sqlite_master where type= 'table'
```

```
  File "/tmp/ipykernel_68/739233235.py", line 1
    select * from sqlite_master where type= 'table'
               ^
SyntaxError: invalid syntax
```

Double-click **here** for a hint

In [22]:
```
%sql SELECT name FROM sqlite_master WHERE TYPE= 'table'
```

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

Out[22]:

name
CENSUS_DATA
CHICAGO_PUBLIC_SCHOOLS_DATA

Double-click **here** for the solution.

Query the database system catalog to retrieve column metadata

The SCHOOLS table contains a large number of columns. How many columns does this table have?

In [ ]:
```
# type in your query to retrieve the number of columns in the SCHOOLS table
```

In [34]:
```
select count(*) from chicago_public_schools_data
```

```
  File "/tmp/ipykernel_68/1780027626.py", line 1
    select count(*) from chicago_public_schools_data
               ^
SyntaxError: invalid syntax
```

In [23]:
```
%sql SELECT count(name) FROM PRAGMA_TABLE_INFO('CHICAGO_PUBLIC_SCHOOLS_DATA');
```

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

Out[23]:

count(name)
78

Double-click **here** for the solution.

Now retrieve the the list of columns in SCHOOLS table and their column type (datatype) and length.

In [ ]:
```
# type in your query to retrieve all column names in the SCHOOLS table along with their datatypes and length
```

In [40]:
```
%sql select distinct name, type, length(type) from pragma_table_info('chicago_public_schools_data')
```

```
UsageError: Line magic function `%sql` not found.
```

Double-click **here** for the solution.

Questions

1. Is the column name for the "SCHOOL ID" attribute in upper or mixed case?
2. What is the name of "Community Area Name" column in your table? Does it have spaces?
3. Are there any columns in whose names the spaces and paranthesis (round brackets) have been replaced by the underscore character "_"?

Problems

Problem 1

How many Elementary Schools are in the dataset?

In [48]:
```
SELECT count(*) FROM 'CHICAGO_PUBLIC_SCHOOLS_DATA' where "Elementary, Middle, or High School"='ES';
```

```
  File "/tmp/ipykernel_68/1038670752.py", line 1
    SELECT count(*) FROM 'CHICAGO_PUBLIC_SCHOOLS_DATA' where "Elementary, Middle, or High School"='ES';
               ^
SyntaxError: invalid syntax
```

Double-click **here** for a hint

Double-click **here** for another hint

In [24]:
```
%sql SELECT count(*) FROM 'CHICAGO_PUBLIC_SCHOOLS_DATA' where "Elementary, Middle, or High School"='ES';
```

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

Out[24]:

count(*)
462

Double-click **here** for the solution.

Problem 2

What is the highest Safety Score?

In [53]:
```
Select MAX(Safety_Score) AS MAX_SAFETY_SCORE FROM 'CHICAGOPUBLICSCHOOLSDATA';
```

```
  File "/tmp/ipykernel_68/582858078.py", line 1
    Select MAX(Safety_Score) AS MAX_SAFETY_SCORE FROM 'CHICAGOPUBLICSCHOOLSDATA';
           ^
SyntaxError: invalid syntax
```

Double-click **here** for a hint

In [26]:
```
%sql SELECT MAX(Safety_Score) AS MAX_SAFETY_SCORE FROM 'CHICAGO_PUBLIC_SCHOOLS_DATA';
```

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

Out[26]:

MAX_SAFETY_SCORE
99.0

Double-click **here** for the solution.

Problem 3

Which schools have highest Safety Score?

In [56]:
```
SELECT NAME_OF_SCHOOL, SAFETY_SCORE from ChicagoPublicSchools where Safety_score = 99;
```

```
  File "/tmp/ipykernel_68/1960369955.py", line 1
    SELECT NAME_OF_SCHOOL, SAFETY_SCORE from ChicagoPublicSchools where Safety_score = 99;
                           ^
SyntaxError: invalid syntax
```

In [27]:
```
%sql SELECT Name_of_School, Safety_Score from CHICAGO_PUBLIC_SCHOOLS_DATA where Safety_Score=99
```

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

Out[27]:

NAME_OF_SCHOOL	SAFETY_SCORE
Abraham Lincoln Elementary School	99.0
Alexander Graham Bell Elementary School	99.0
Annie Keller Elementary Gifted Magnet School	99.0
Augustus H Burley Elementary School	99.0
Edgar Allan Poe Elementary Classical School	99.0
Edgebrook Elementary School	99.0
Ellen Mitchell Elementary School	99.0
James E McDade Elementary Classical School	99.0
James G Blaine Elementary School	99.0
LaSalle Elementary Language Academy	99.0
Mary E Courtenay Elementary Language Arts Center	99.0
Northside College Preparatory High School	99.0
Northside Learning Center High School	99.0
Norwood Park Elementary School	99.0
Oriole Park Elementary School	99.0
Sauganash Elementary School	99.0
Stephen Decatur Classical Elementary School	99.0
Talman Elementary School	99.0
Wildwood Elementary School	99.0

Double-click **here** for the solution.

Problem 4

What are the top 10 schools with the highest "Average Student Attendance"?

In [ ]:
```
select name_of_school, average_student_attendance from CHIAGO_PUBLIC_SCHOOLS_DATA \ order by Average_Student_Aten
```

Double-click **here** for the solution.

In [30]:
```
%sql select Name_of_School, Average_Student_Attendance from CHICAGO_PUBLIC_SCHOOLS_DATA \
    order by Average_Student_Attendance desc nulls last limit 10
```

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

Out[30]:

NAME_OF_SCHOOL	AVERAGE_STUDENT_ATTENDANCE
John Charles Haines Elementary School	98.40%
James Ward Elementary School	97.80%
Edgar Allan Poe Elementary Classical School	97.60%
Orozco Fine Arts & Sciences Elementary School	97.60%
Rachel Carson Elementary School	97.60%
Annie Keller Elementary Gifted Magnet School	97.50%
Andrew Jackson Elementary Language Academy	97.40%
Lenart Elementary Regional Gifted Center	97.40%
Disney II Magnet School	97.30%
John H Vanderpoel Elementary Magnet School	97.20%

Problem 5

Retrieve the list of 5 Schools with the lowest Average Student Attendance sorted in ascending order based on attendance

In [35]:
```
%sql SELECT Name_of_School, Average_Student_Attendance  \
     from CHICAGO_PUBLIC_SCHOOLS_DATA \
     order by Average_Student_Attendance \
     LIMIT 5
```

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

Out[35]:

NAME_OF_SCHOOL	AVERAGE_STUDENT_ATTENDANCE
Velma F Thomas Early Childhood Center	None
Richard T Crane Technical Preparatory High School	57.90%
Barbara Vick Early Childhood & Family Center	60.90%
Dyett High School	62.50%
Wendell Phillips Academy High School	63.00%

In [ ]:

```
```

Double-click **here** for the solution.

Problem 6

Now remove the '%' sign from the above result set for Average Student Attendance column

In [ ]:

```
```

Double-click **here** for a hint

Double-click **here** for the solution.

In [38]:

```
%sql SELECT Name_of_School, REPLACE(Average_Student_Attendance, '%', '') \
     from CHICAGO_PUBLIC_SCHOOLS_DATA \
     order by Average_Student_Attendance \
     LIMIT 5
```

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

Out[38]:

NAME_OF_SCHOOL	REPLACE(Average_Student_Attendance, '%', '')
Velma F Thomas Early Childhood Center	None
Richard T Crane Technical Preparatory High School	57.90
Barbara Vick Early Childhood & Family Center	60.90
Dyett High School	62.50
Wendell Phillips Academy High School	63.00

In [ ]:

```
```

Problem 7

Which Schools have Average Student Attendance lower than 70%?

In [ ]:

```
select name_of_school, average_student_attendance \ from Chicago_public_schools_data \ where casta(replace Average_St
```

Double-click **here** for a hint

In [ ]:

```
```

Double-click **here** for another hint

Double-click **here** for the solution.

Problem 8

Get the total College Enrollment for each Community Area

In [57]:

```
Select Community_Area_Name, SUM(College_Enrollment) \ from Chicago_Public_Schools_Data \ group by Community_Area_Name
```

```
  File "/tmp/ipykernel_68/1167542610.py", line 1
    Select Community_Area_Name, SUM(College_Enrollment) \ from Chicago_Public_Schools_Data \ group by Community_Area_N
ame
                                ^
SyntaxError: invalid syntax
```

Double-click **here** for a hint

Double-click **here** for another hint

In [14]:
```
%sql select Community_Area_Name, sum(College_Enrollment) AS TOTAL_ENROLLMENT \
   from CHICAGO_PUBLIC_SCHOOLS_DATA \
   group by Community_Area_Name
```

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

Out[14]:

COMMUNITY_AREA_NAME	TOTAL_ENROLLMENT
ALBANY PARK	6864
ARCHER HEIGHTS	4823
ARMOUR SQUARE	1458
ASHBURN	6483
AUBURN GRESHAM	4175
AUSTIN	10933
AVALON PARK	1522
AVONDALE	3640
BELMONT CRAGIN	14386
BEVERLY	1636
BRIDGEPORT	3167
BRIGHTON PARK	9647
BURNSIDE	549
CALUMET HEIGHTS	1568
CHATHAM	5042
CHICAGO LAWN	7086
CLEARING	2085
DOUGLAS	4670
DUNNING	4568
EAST GARFIELD PARK	5337
EAST SIDE	5305
EDGEWATER	4600

COMMUNITY_AREA_NAME	TOTAL_ENROLLMENT
EDISON PARK	910
ENGLEWOOD	6832
FOREST GLEN	1431
FULLER PARK	531
GAGE PARK	9915
GARFIELD RIDGE	4552
GRAND BOULEVARD	2809
GREATER GRAND CROSSING	4051
HEGEWISCH	963
HERMOSA	3975
HUMBOLDT PARK	8620
HYDE PARK	1930
IRVING PARK	7764
JEFFERSON PARK	1755
KENWOOD	4287
LAKE VIEW	7055
LINCOLN PARK	5615
LINCOLN SQUARE	4132
LOGAN SQUARE	7351
LOOP	871
LOWER WEST SIDE	7257
MCKINLEY PARK	1552

COMMUNITY_AREA_NAME	TOTAL_ENROLLMENT
MONTCLARE	1317
MORGAN PARK	3271
MOUNT GREENWOOD	2091
NEAR NORTH SIDE	3362
NEAR SOUTH SIDE	1378
NEAR WEST SIDE	7975
NEW CITY	7922
NORTH CENTER	7541
NORTH LAWNDALE	5146
NORTH PARK	4210
NORWOOD PARK	6469
OAKLAND	140
OHARE	786
PORTAGE PARK	6954
PULLMAN	1620
RIVERDALE	1547
ROGERS PARK	4068
ROSELAND	7020
SOUTH CHICAGO	4043
SOUTH DEERING	1859
SOUTH LAWNDALE	14793
SOUTH SHORE	4543

COMMUNITY_AREA_NAME	TOTAL_ENROLLMENT
UPTOWN	4388
WASHINGTON HEIGHTS	4006
WASHINGTON PARK	2648
WEST ELSDON	3700
WEST ENGLEWOOD	5946
WEST GARFIELD PARK	2622
WEST LAWN	4207
WEST PULLMAN	3240
WEST RIDGE	8197
WEST TOWN	9429
WOODLAWN	4206

Double-click **here** for the solution.

Problem 9

Get the 5 Community Areas with the least total College Enrollment sorted in ascending order

In [60]:
```
Select Community_Area_Name, sum(College_Enrollment) AS TOTAL_ENROLLMENT \
From Chicago_Public_Schools_Data \
Group by Community_Area_Name \
Ordered by Total_Enrollment asc \
Limit 5;
```

```
  File "/tmp/ipykernel_68/1582964485.py", line 1
    Select Community_Area_Name, sum(College_Enrollment) AS TOTAL_ENROLLMENT \
                                ^
SyntaxError: invalid syntax
```

In [15]:
```
%sql select Community_Area_Name, sum(College_Enrollment) AS TOTAL_ENROLLMENT \
   from CHICAGO_PUBLIC_SCHOOLS_DATA \
   group by Community_Area_Name \
   order by TOTAL_ENROLLMENT asc \
   LIMIT 5
```

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

Out[15]:

COMMUNITY_AREA_NAME	TOTAL_ENROLLMENT
OAKLAND	140
FULLER PARK	531
BURNSIDE	549
OHARE	786
LOOP	871

Double-click **here** for a hint

Double-click **here** for the solution.

Problem 10

List 5 schools with lowest safety score.

In [63]:
```
%%sql Select Name_Of_School, Safety_Score \ from CHICAGO_PUBLIC_SCHOOLS_DATA where safety_Score !0='None')\ LImit 5;
```

```
UsageError: Cell magic `%%sql` not found.
```

In [17]:
```
%sql SELECT name_of_school, safety_score \
from CHICAGO_PUBLIC_SCHOOLS_DATA where safety_score !='None' \
LIMIT 5
```

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

Out[17]:

NAME_OF_SCHOOL	SAFETY_SCORE
Abraham Lincoln Elementary School	99.0
Adam Clayton Powell Paideia Community Academy Elementary School	54.0
Adlai E Stevenson Elementary School	61.0
Agustin Lara Elementary Academy	56.0
Air Force Academy High School	49.0

Double-click **here** for the solution.

Problem 11

Get the hardship index for the community area which has College Enrollment of 4368

In [64]:
```
select community_area_name, hardship_index from CHICAGO_PUBLIC_SCHOOLS_DATA where college_enrollment= 4368
```

```
  File "/tmp/ipykernel_68/3825082963.py", line 1
    select community_area_name, hardship_index from CHICAGO_PUBLIC_SCHOOLS_DATA where college_enrollment= 4368
                                ^
SyntaxError: invalid syntax
```

In [19]:
```
%%sql
select hardship_index from CENSUS_DATA CD, CHICAGO_PUBLIC_SCHOOLS_DATA CPS
where CD.community_area_number = CPS.community_area_number
and college_enrollment = 4368
```

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

Out[19]:

HARDSHIP_INDEX
6.0

Double-click **here** for the solution.

Problem 12

Get the hardship index for the community area which has the highest value for College Enrollment

In [ ]:
```
select community_area_name, hardship_index \ from CHICAGO_PUBLIC_SCHOOLS_DATA ordered where community_area_name in\ b
```

In [ ]:
```
%sql select community_area_number, community_area_name, hardship_index from CENSUS_DATA \
   where community_area_number in \
   ( select community_area_number from CHICAGO_PUBLIC_SCHOOLS_DATA order by college_enrollment desc limit 1 )
```

Double-click **here** for the solution.

Summary

In this lab you learned how to work with a real word dataset using SQL and Python. You learned how to query columns with spaces or special characters in their names and with mixed case names. You also used built in database functions and practiced how to sort, limit, and order result sets, as well as used sub-queries and worked with multiple tables.

Author

Rav Ahuja

Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2022-03-04	2.2	Lakshmi Holla	Made changes in markdown cells
2020-11-27	2.1	Sannareddy Ramesh	Modified data sets and added new problems
2020-08-28	2.0	Lavanya	Moved lab to course repo in GitLab
