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 [8]:
```
import csv, sqlite3

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

In [9]:
```
!pip install -q pandas==1.1.5
```

In [10]:
```
%load_ext sql
```

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

In [11]:
```
%sql sqlite:///RealWorldData.db
```

Out[11]:
```
'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 [23]:
```
import pandas as pd
from sqlalchemy import create_engine

# Create a database connection
engine = create_engine('sqlite:///my_database.db')  # replace with your database co

# Load data from CSVs and write into the database
df = pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.clou
df.to_sql("CENSUS_DATA", engine, if_exists='replace', index=False, method="multi")

df = pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.clou
df.to_sql("CHICAGO_CRIME_DATA", engine, if_exists='replace', index=False, method="m

df = pd.read_csv("https://cf-courses-data.s3.us.cloud-object-storage.appdomain.clou
df.to_sql("CHICAGO_PUBLIC_SCHOOLS_DATA", engine, if_exists='replace', index=False,

print(df)
```

```
     School_ID                                    NAME_OF_SCHOOL  \
0       610038                  Abraham Lincoln Elementary School
1       610281  Adam Clayton Powell Paideia Community Academy ...
2       610185                Adlai E Stevenson Elementary School
3       609993                    Agustin Lara Elementary Academy
4       610513                      Air Force Academy High School
..         ...                                               ...
561     610172                William T Sherman Elementary School
562     609844                 William W Carter Elementary School
563     610088                Wolfgang A Mozart Elementary School
564     609977               Woodlawn Community Elementary School
565     610392                 World Language Academy High School

    Elementary, Middle, or High School         Street_Address     City State  \
0                                    ES        615 W Kemper Pl  Chicago    IL
1                                    ES  7511 S South Shore Dr  Chicago    IL
2                                    ES     8010 S Kostner Ave  Chicago    IL
3                                    ES     4619 S Wolcott Ave  Chicago    IL
4                                    HS        3630 S Wells St  Chicago    IL
..                                  ...                     ...      ...   ...
561                                  ES         1000 W 52nd St  Chicago    IL
562                                  ES    5740 S Michigan Ave  Chicago    IL
563                                  ES      2200 N Hamlin Ave  Chicago    IL
564                                  ES     6657 S Kimbark Ave  Chicago    IL
565                                  HS     3120 S Kostner Ave  Chicago    IL

     ZIP_Code    Phone_Number  \
0       60614  (773) 534-5720
1       60649  (773) 535-6650
2       60652  (773) 535-2280
3       60609  (773) 535-4389
4       60609  (773) 535-1590
..        ...             ...
561     60609  (773) 535-1757
562     60637  (773) 535-0860
563     60647  (773) 534-4160
564     60637  (773) 535-0801
565     60623  (773) 535-4334

                                              Link  \
0    http://schoolreports.cps.edu/SchoolProgressRep...
1    http://schoolreports.cps.edu/SchoolProgressRep...
2    http://schoolreports.cps.edu/SchoolProgressRep...
3    http://schoolreports.cps.edu/SchoolProgressRep...
4    http://schoolreports.cps.edu/SchoolProgressRep...
..                                                 ...
561  http://schoolreports.cps.edu/SchoolProgressRep...
562  http://schoolreports.cps.edu/SchoolProgressRep...
563  http://schoolreports.cps.edu/SchoolProgressRep...
564  http://schoolreports.cps.edu/SchoolProgressRep...
565  http://schoolreports.cps.edu/SchoolProgressRep...

                       Network_Manager  ... Freshman_on_Track_Rate__  \
0          Fullerton Elementary Network  ...                      NDA
1             Skyway Elementary Network  ...                      NDA
2             Midway Elementary Network  ...                      NDA
```

```
3           Pershing Elementary Network  ...                      NDA
4    Southwest Side High School Network  ...                     91.8
..                                  ...  ...                      ...
561                         AUSL Schools  ...                      NDA
562     Burnham Park Elementary Network  ...                      NDA
563        Fullerton Elementary Network  ...                      NDA
564     Burnham Park Elementary Network  ...                      NDA
565       West Side High School Network  ...                       76

    X_COORDINATE Y_COORDINATE   Latitude  Longitude COMMUNITY_AREA_NUMBER  \
0    1171699.458  1915829.428  41.924497 -87.644522                     7
1    1196129.985  1856209.466  41.760324 -87.556736                    43
2    1148427.165  1851012.215  41.747111 -87.731702                    70
3    1164504.290  1873959.199  41.809757 -87.672145                    61
4    1175177.622  1880745.126  41.828146 -87.632794                    34
..           ...          ...        ...        ...                   ...
561  1170500.817  1870373.159  41.799788 -87.650255                    61
562  1178101.365  1866810.123  41.789841 -87.622490                    40
563  1150644.396  1914368.955  41.920927 -87.721925                    22
564  1185825.188  1860883.579  41.773400 -87.594356                    42
565  1147521.302  1883405.128  41.836020 -87.734195                    30

    COMMUNITY_AREA_NAME  Ward Police_District                     Location
0          LINCOLN PARK    43              18  (41.92449696, -87.64452163)
1           SOUTH SHORE     7               4  (41.76032435, -87.55673627)
2               ASHBURN    13               8  (41.74711093, -87.73170248)
3              NEW CITY    20               9    (41.8097569, -87.6721446)
4         ARMOUR SQUARE    11               9  (41.82814609, -87.63279369)
..                  ...   ...             ...                          ...
561            NEW CITY    16               9  (41.79978772, -87.65025483)
562     WASHINGTON PARK    20               2  (41.78984129, -87.62248974)
563        LOGAN SQUARE    35              25  (41.92092734, -87.72192541)
564            WOODLAWN     5               3  (41.77339962, -87.59435584)
565      SOUTH LAWNDALE    22              10  (41.83601953, -87.73419465)

[566 rows x 78 columns]
```

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

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

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 [20]:
```
# type in your query to retrieve list of all tables in the database

%sql SELECT name FROM sqlite_master WHERE type='table'
```

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

Out[20]:

name
CENSUS_DATA
CHICAGO_CRIME_DATA
CHICAGO_PUBLIC_SCHOOLS_DATA

Double-click **here** for a hint

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 [21]:
```
# type in your query to retrieve the number of columns in the SCHOOLS table
%sql SELECT count(name) FROM PRAGMA_TABLE_INFO('CHICAGO_PUBLIC_SCHOOLS_DATA');
```

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

Out[21]:

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 [24]:
```
# type in your query to retrieve all column names in the SCHOOLS table along with t
%sql SELECT name,type,length(type) FROM PRAGMA_TABLE_INFO('CHICAGO_PUBLIC_SCHOOLS_D
```

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

Out[24]:

name	type	length(type)
School_ID	INTEGER	7
NAME_OF_SCHOOL	TEXT	4
Elementary, Middle, or High School	TEXT	4
Street_Address	TEXT	4
City	TEXT	4
State	TEXT	4
ZIP_Code	INTEGER	7
Phone_Number	TEXT	4
Link	TEXT	4
Network_Manager	TEXT	4
Collaborative_Name	TEXT	4
Adequate_Yearly_Progress_Made_	TEXT	4
Track_Schedule	TEXT	4
CPS_Performance_Policy_Status	TEXT	4
CPS_Performance_Policy_Level	TEXT	4
HEALTHY_SCHOOL_CERTIFIED	TEXT	4
Safety_Icon	TEXT	4
SAFETY_SCORE	REAL	4
Family_Involvement_Icon	TEXT	4
Family_Involvement_Score	TEXT	4
Environment_Icon	TEXT	4
Environment_Score	REAL	4
Instruction_Icon	TEXT	4
Instruction_Score	REAL	4
Leaders_Icon	TEXT	4
Leaders_Score	TEXT	4
Teachers_Icon	TEXT	4
Teachers_Score	TEXT	4
Parent_Engagement_Icon	TEXT	4
Parent_Engagement_Score	TEXT	4

name	type	length(type)
Parent_Environment_Icon	TEXT	4
Parent_Environment_Score	TEXT	4
AVERAGE_STUDENT_ATTENDANCE	TEXT	4
Rate_of_Misconducts__per_100_students_	REAL	4
Average_Teacher_Attendance	TEXT	4
Individualized_Education_Program_Compliance_Rate	TEXT	4
Pk_2_Literacy__	TEXT	4
Pk_2_Math__	TEXT	4
Gr3_5_Grade_Level_Math__	TEXT	4
Gr3_5_Grade_Level_Read__	TEXT	4
Gr3_5_Keep_Pace_Read__	TEXT	4
Gr3_5_Keep_Pace_Math__	TEXT	4
Gr6_8_Grade_Level_Math__	TEXT	4
Gr6_8_Grade_Level_Read__	TEXT	4
Gr6_8_Keep_Pace_Math_	TEXT	4
Gr6_8_Keep_Pace_Read__	TEXT	4
Gr_8_Explore_Math__	TEXT	4
Gr_8_Explore_Read__	TEXT	4
ISAT_Exceeding_Math__	REAL	4
ISAT_Exceeding_Reading__	REAL	4
ISAT_Value_Add_Math	REAL	4
ISAT_Value_Add_Read	REAL	4
ISAT_Value_Add_Color_Math	TEXT	4
ISAT_Value_Add_Color_Read	TEXT	4
Students_Taking__Algebra__	TEXT	4
Students_Passing__Algebra__	TEXT	4
9th Grade EXPLORE (2009)	TEXT	4
9th Grade EXPLORE (2010)	TEXT	4
10th Grade PLAN (2009)	TEXT	4
10th Grade PLAN (2010)	TEXT	4

name	type	length(type)
Net_Change_EXPLORE_and_PLAN	TEXT	4
11th Grade Average ACT (2011)	TEXT	4
Net_Change_PLAN_and_ACT	TEXT	4
College_Eligibility__	TEXT	4
Graduation_Rate__	TEXT	4
College_Enrollment_Rate__	TEXT	4
COLLEGE_ENROLLMENT	INTEGER	7
General_Services_Route	INTEGER	7
Freshman_on_Track_Rate__	TEXT	4
X_COORDINATE	REAL	4
Y_COORDINATE	REAL	4
Latitude	REAL	4
Longitude	REAL	4
COMMUNITY_AREA_NUMBER	INTEGER	7
COMMUNITY_AREA_NAME	TEXT	4
Ward	INTEGER	7
Police_District	INTEGER	7
Location	TEXT	4

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 [25]:
```
%sql select count(*) from CHICAGO_PUBLIC_SCHOOLS_DATA where "Elementary, Middle, or
```

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

Out[25]:

count(*)
462

Double-click **here** for a hint

Double-click **here** for another hint

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

Problem 2

What is the highest Safety Score?

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 a hint

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

Problem 3

Which schools have highest Safety Score?

In [27]:
```
%sql select Name_of_School, Safety_Score from CHICAGO_PUBLIC_SCHOOLS_DATA where \
  Safety_Score= (select MAX(Safety_Score) from CHICAGO_PUBLIC_SCHOOLS_DATA)
```

```
 * 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 [28]:
```
%sql select Name_of_School, Average_Student_Attendance from CHICAGO_PUBLIC_SCHOOLS_
    order by Average_Student_Attendance desc nulls last limit 10
```

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

Out[28]:

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%

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

Problem 5

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

In [29]:
```
%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[29]:

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%

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

Problem 6

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

In [30]:
```
%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[30]:

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

Double-click **here** for a hint

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

Problem 7

Which Schools have Average Student Attendance lower than 70%?

In [31]:
```
%sql SELECT Name_of_School, Average_Student_Attendance  \
     from CHICAGO_PUBLIC_SCHOOLS_DATA \
     where CAST ( REPLACE(Average_Student_Attendance, '%', '') AS DOUBLE ) < 70 \
     order by Average_Student_Attendance
```

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

Out[31]:

NAME_OF_SCHOOL	AVERAGE_STUDENT_ATTENDANCE
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%
Orr Academy High School	66.30%
Manley Career Academy High School	66.80%
Chicago Vocational Career Academy High School	68.80%
Roberto Clemente Community Academy High School	69.60%

Double-click **here** for a hint

Double-click **here** for another hint

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

Problem 8

Get the total College Enrollment for each Community Area

In [32]:
```
%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[32]:

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
EDISON PARK	910
ENGLEWOOD	6832
FOREST GLEN	1431
FULLER PARK	531
GAGE PARK	9915
GARFIELD RIDGE	4552
GRAND BOULEVARD	2809
GREATER GRAND CROSSING	4051

COMMUNITY_AREA_NAME	TOTAL_ENROLLMENT
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
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

COMMUNITY_AREA_NAME	TOTAL_ENROLLMENT
ROGERS PARK	4068
ROSELAND	7020
SOUTH CHICAGO	4043
SOUTH DEERING	1859
SOUTH LAWNDALE	14793
SOUTH SHORE	4543
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 a hint

Double-click **here** for another hint

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

Problem 9

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

In [33]:
```
%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[33]:

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 [34]:
```
%sql SELECT name_of_school, safety_score \
FROM CHICAGO_PUBLIC_SCHOOLS_DATA  where safety_score !='None' \
ORDER BY safety_score \
LIMIT 5
```

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

Out[34]:

NAME_OF_SCHOOL	SAFETY_SCORE
Edmond Burke Elementary School	1.0
Luke O'Toole Elementary School	5.0
George W Tilton Elementary School	6.0
Foster Park Elementary School	11.0
Emil G Hirsch Metropolitan High School	13.0

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

Problem 11

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

In [35]:
```
%%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[35]:

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 [36]:
```
%sql select community_area_number, community_area_name, hardship_index from CENSUS_
   where community_area_number in \
   ( select community_area_number from CHICAGO_PUBLIC_SCHOOLS_DATA order by college
```

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

Out[36]:

COMMUNITY_AREA_NUMBER	COMMUNITY_AREA_NAME	HARDSHIP_INDEX
5.0	North Center	6.0

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