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

```
import csv, sqlite3
```

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

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

```
%load_ext sql
```

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

```
'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.

```
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_DATA", con, if_exists='replace',
index=False, method="multi")
```

```
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages/
pandas/core/generic.py:2615: UserWarning: The spaces in these column
names will not be changed. In pandas versions < 0.14, spaces were
converted to underscores.
  method=method,
```

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

```
# 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.
```

```
[('CENSUS_DATA',), ('CHICAGO_CRIME_DATA',),
('CHICAGO_PUBLIC_SCHOOLS_DATA',)]
```

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?

```
# 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.
```

```
[(78,)]
```

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

```
# type in your query to retrieve all column names in the SCHOOLS table
along with their datatypes and length
%sql SELECT name,type,length(type) FROM
PRAGMA_TABLE_INFO('CHICAGO_PUBLIC_SCHOOLS_DATA')
```

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

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

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?

```
%sql select count(*) from CHICAGO_PUBLIC_SCHOOLS_DATA where
"Elementary, Middle, or High School"='ES'

 * sqlite:///RealWorldData.db
Done.

[(462,)]
```

Problem 2

What is the highest Safety Score?

```
%sql SELECT MAX(safety_score) FROM CHICAGO_PUBLIC_SCHOOLS_DATA

 * sqlite:///RealWorldData.db
Done.

[(99.0,)]
```

Problem 3

Which schools have highest Safety Score?

```
%sql select Name_of_School, Safety_Score from
CHICAGO_PUBLIC_SCHOOLS_DATA where Safety_Score = 99
```

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

```
[('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)]
```

Problem 4

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

```
%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.
```

```
[('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

```
%sql SELECT Name_of_School, Average_Student_Attendance  \
     from CHICAGO_PUBLIC_SCHOOLS_DATA \
     order by Average_Student_Attendance \
     LIMIT 5
```

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

```
[('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%')]
```

Problem 6

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

```
%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.
```

```
[('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')]
```

Problem 7

Which Schools have Average Student Attendance lower than 70%?

```
%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.
```

```
[('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%')]
```

Problem 8

Get the total College Enrollment for each Community Area

```
%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.
```

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

Problem 9

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

```
%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.

[('OAKLAND', 140),
 ('FULLER PARK', 531),
 ('BURNSIDE', 549),
 ('OHARE', 786),
 ('LOOP', 871)]
```

Problem 10

List 5 schools with lowest safety score.

```
%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.

[('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)]
```

Problem 11

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

```
%%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.

[(6.0,)]
```

Problem 12

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

```
%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)
```

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

```
[(5.0, 'North Center', 6.0)]
```

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