DB0201EN-Week4-1-1-RealDataPractice-v5_sqlite_Learner

February 3, 2025

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

Estimated time needed: **30** minutes

1.1 Objectives

After completing 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

1.2 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.

1.2.1 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 pandas
!pip install ipython-sql prettytable

import prettytable
prettytable.DEFAULT = 'DEFAULT'
```

[ ]:
```
!pip install ipython-sql
%load_ext sql
```

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

1.2.2 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.

[ ]:
```

```

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

1.2.3 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
```

Double-click **here** for a hint

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

1.2.4 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
```

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

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
```

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

1.2.5 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 "_"?

1.3 Problems

1.3.1 Problem 1

How many Elementary Schools are in the dataset?

[ ]:
```

```

Double-click **here** for a hint

Double-click **here** for another hint

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

1.3.2 Problem 2

What is the highest Safety Score?

[ ]:
```

```

Double-click **here** for a hint

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

1.3.3 Problem 3

Which schools have highest Safety Score?

[ ]:
```

```

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

1.3.4 Problem 4

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

[ ]:
```

```

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

1.3.5 Problem 5

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

[ ]:
```

```

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

1.3.6 Problem 6

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

```
[ ]:
```

Double-click **here** for a hint

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

1.3.7 Problem 7

Which Schools have Average Student Attendance lower than 70%?

```
[ ]:
```

Double-click **here** for a hint

Double-click **here** for another hint

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

1.3.8 Problem 8

Get the total College Enrollment for each Community Area

```
[ ]:
```

Double-click **here** for a hint

Double-click **here** for another hint

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

1.3.9 Problem 9

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

```
[ ]:
```

Double-click **here** for a hint

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

1.3.10 Problem 10

List 5 schools with lowest safety score.

```
[ ]:
```

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

1.3.11 Problem 11

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

```
[ ]:
```

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

1.3.12 Problem 12

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

```
[ ]:
```

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

1.4 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.

1.5 Author

Rav Ahuja

##