DB0201EN-Week4-1-1-Analyzing-2-py

December 3, 2018

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

1 Introduction

This notebook shows how to work with a real world dataset using SQL and Python. In this lab you will: 1. Understand the dataset for Chicago Public School level performance 1. Store the dataset in an Db2 database on IBM Cloud instance 1. Retrieve metadata about tables and columns and query data from mixed case columns 1. Solve example problems to practice your SQL skills including using built-in database functions

1.1 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

Now download a static copy of this database and review some of its contents: https://ibm.box.com/shared/static/0g7kbanvn5l2gt2qu38ukooatnjqyuys.csv

1.1.1 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.

While it is easier to read the dataset into a Pandas dataframe and then PERSIST it into the database as we saw in the previous lab, it results in mapping to default datatypes which may not be optimal for SQL querying. For example a long textual field may map to a CLOB instead of a VARCHAR.

Therefore, **it is highly recommended to manually load the table using the database console LOAD tool, as indicated in Week 2 Lab 1 Part II**. The only difference with that lab is that in Step 5 of the instructions you will need to click on create "(+) New Table" and specify the name of the table you want to create and then click "Next".

Now open the Db2 console, open the LOAD tool, Select / Drag the .CSV file for the CHICAGO PUBLIC SCHOOLS dataset and load the dataset into a new table called SCHOOLS.

1.1.2 Connect to the database

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

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

```
In [ ]: # Enter the connection string for your Db2 on Cloud database instance below
        # %sql ibm_db_sa://my-username:my-password@my-hostname:my-port/my-db-name
        %sql ibm_db_sa://
```

1.1.3 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

```
In [ ]: # type in your query to retrieve list of all tables in the database for your db2 schema
```

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

1.1.4 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?

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

Double-click **here** for a hint
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
```

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

1.1.5 Questions

1. Is the column name for the "SCHOOL ID" attribute in upper or mixed case?
2. What is the "Community Area Name" field called in your table? Have the spaces " " between the words been replaced by some other character?
3. Have the paranthesis (round brackets) in the "College Enrollment (number of students)" attribute been replaced by some other character?
4. Are there any columns in whose names the spaces and paranthesis (round brackets) have not been replaced by the underscore character "_"?

1.2 Problems

1.2.1 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.

1.2.2 Problem 2

What is the highest Safety Score? Double-click **here** for a hint
Double-click **here** for the solution.

1.2.3 Problem 3

Which schools have highest Safety Score? Double-click **here** for the solution.

1.2.4 Problem 4

What are the top 10 schools with the highest "Average Student Attendance"? Double-click **here** for the solution.

1.2.5 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.

1.2.6 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.

1.2.7 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.

1.2.8 Problem 8

Get the total College Enrollment (number of students) for each Community Area Double-click **here** for a hint
Double-click **here** for another hint
Double-click **here** for the solution.

1.2.9 Problem 9

Get the 5 Community Areas with the least total College Enrollment (number of students) sorted in ascending order Double-click **here** for a hint
Double-click **here** for the solution.

1.3 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 names. You also used built in database functions.