Chicago Assignment

 Posted on 2019-11-27 |  Edited on 2019-12-05 |  In [Tech](http://bingcs.com/categories/Tech/) |  Comments: [1](http://bingcs.com/2019/11/27/chicago-assignment/#comments)| 4505 Readings

 Symbols count in article: 5.4k |  Reading time ≈ 5 mins.

Assignment of the Coursera IBM Data Science course.

**Datasets**

* **Chicago Socioeconomic Indicators**

This dataset contains a selection of six socioeconomic indicators of public health significance and a “hardship index,” for each Chicago community area, for the years 2008 – 2012.

For this assignment you will use a snapshot of this dataset which can be downloaded from [here](https://ibm.box.com/shared/static/05c3415cbfbtfnr2fx4atenb2sd361ze.csv).

A detailed description of this dataset and the original dataset can be obtained from the Chicago Data Portal at [this page](https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2).

* **Chicago Public Schools**

This dataset shows all school level performance data used to create CPS School Report Cards for the 2011-2012 school year. This dataset is provided by the city of Chicago’s Data Portal.

For this assignment you will use a snapshot of this dataset which can be downloaded from [here](https://ibm.box.com/shared/static/f9gjvj1gjmxxzycdhplzt01qtz0s7ew7.csv).

A detailed description of this dataset and the original dataset can be obtained from the Chicago Data Portal at [this page](https://data.cityofchicago.org/Education/Chicago-Public-Schools-Progress-Report-Cards-2011-/9xs2-f89t).

* **Chicago Crime Data**

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days.

This dataset is quite large - over 1.5GB in size with over 6.5 million rows. For the purposes of this assignment we will use a much smaller sample of this dataset which can be downloaded from [here](https://ibm.box.com/shared/static/svflyugsr9zbqy5bmowgswqemfpm1x7f.csv).

A detailed description of this dataset and the original dataset can be obtained from the Chicago Data Portal at [this page](https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2).

**Download the datasets**

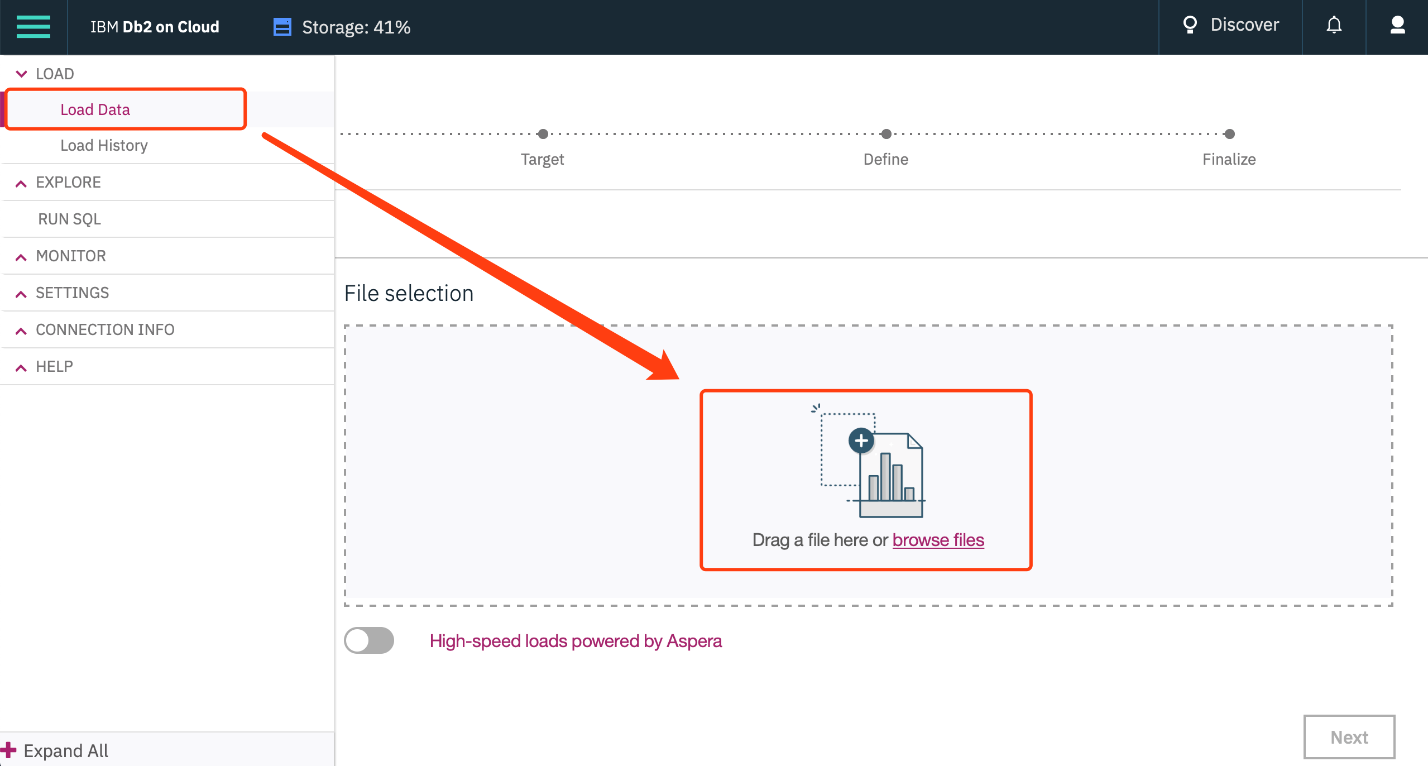
In many cases the dataset to be analyzed is available as a .CSV (comma separated values) file, perhaps on the internet. Click on the links below to download and save the datasets (.CSV files):

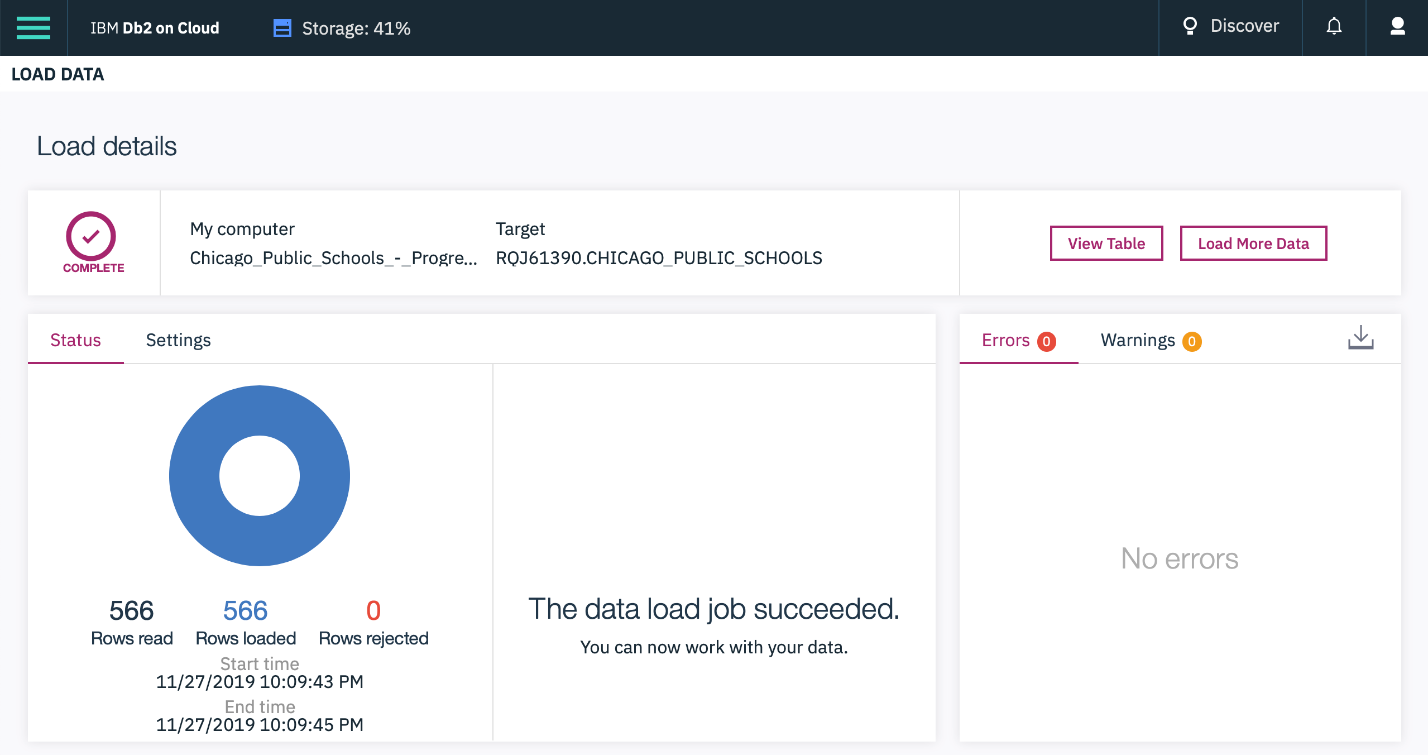
1. [CENSUS\_DATA](https://ibm.box.com/shared/static/05c3415cbfbtfnr2fx4atenb2sd361ze.csv)
2. [CHICAGO\_PUBLIC\_SCHOOLS](https://ibm.box.com/shared/static/f9gjvj1gjmxxzycdhplzt01qtz0s7ew7.csv)
3. [CHICAGO\_CRIME\_DATA](https://ibm.box.com/shared/static/svflyugsr9zbqy5bmowgswqemfpm1x7f.csv)

**NOTE:** Ensure you have downloaded the datasets using the links above instead of directly from the Chicago Data Portal. The versions linked here are subsets of the original datasets and have some of the column names modified to be more database friendly which will make it easier to complete this assignment.

**Load data**

To analyze the data using SQL, it first needs to be stored in the database. Through the IBM DB2 on Cloud, We will later connect our Python program to this database. Before you starting to load data, first prepare the 3 datasets mentioned above, register for IBM Cloud Service and run a resource of DB2.





*Notice: you may face the problem that the format of the date time in your CSV file does not fit the build-in format in the IBM DB2 Cloud platform when loading your tables with a .csv file, just refer to how I solve this problem in my another article:*

[Bing’s Blog: Date-time Transformation in CSV file](http://bingcs.com/2019/11/27/datetime-format-tran-py/)

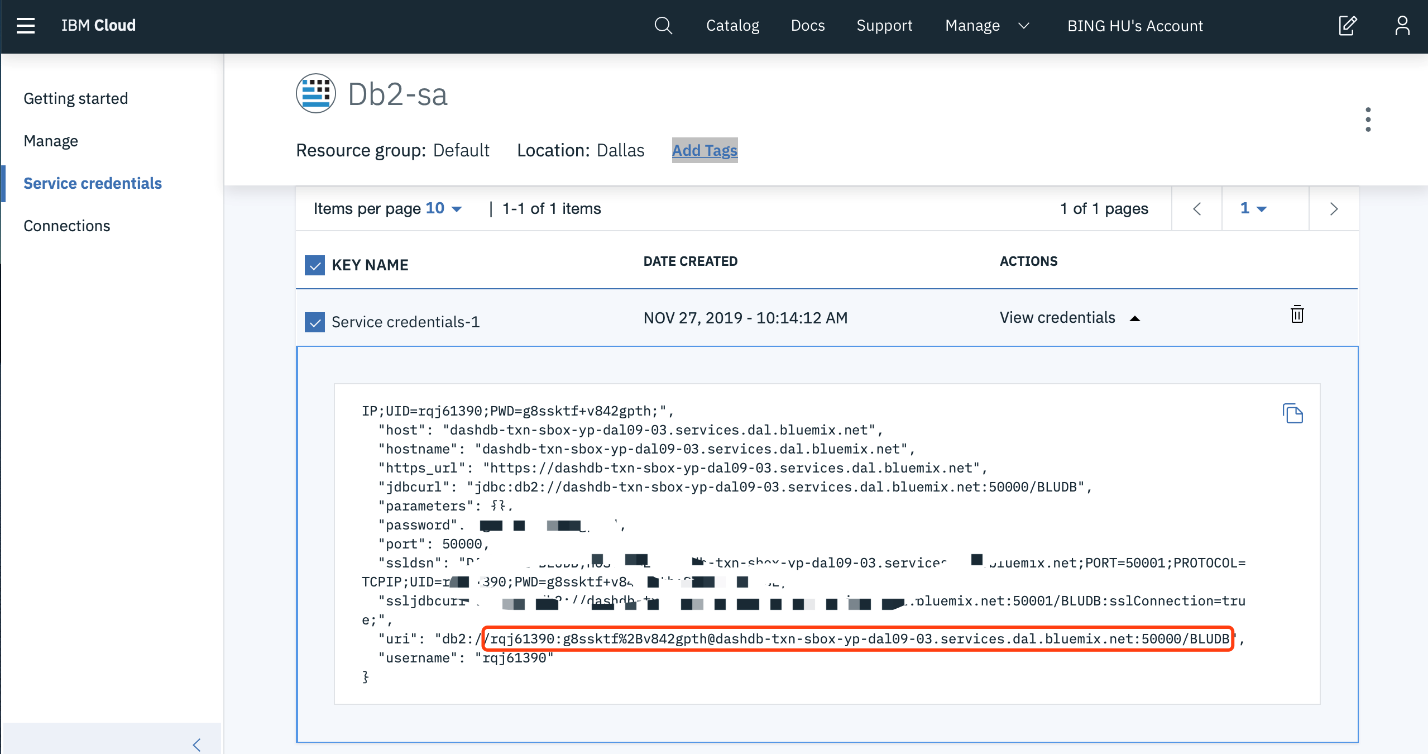
**Connect to the database**

First load the SQL extension and establish a connection with the database

|  |  |
| --- | --- |
| 1 | %load\_ext sql |

Enter db2 connection string. Recall the created Service Credentials for our Db2 instance. From the **uri** field of the Db2 service credentials copy everything after db2:// and paste it.

|  |  |
| --- | --- |
| 1 2 3 4 | # Remember the connection string is of the format: # %sql ibm\_db\_sa://my-username:my-password@my-hostname:my-port/my-db-name # Enter the connection string for your Db2 on Cloud database instance below %sql ibm\_db\_sa://rqj61390:g8ssktf%2Bv842gpth@dashdb-txn-sbox-yp-dal09-03.services.dal.bluemix.net:50000/BLUDB |



**Problems**

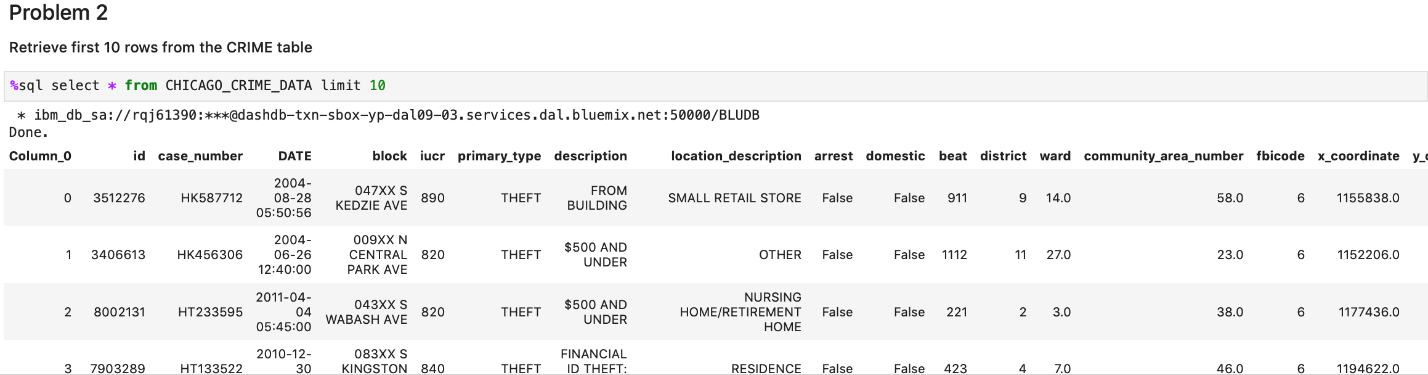
1. Find the total number of crimes recorded in the crime table.

|  |  |
| --- | --- |
| 1 | %sql select count(\*) as num\_crimes from CHICAGO\_CRIME\_DATA |



1. Retrieve first 10 rows from the CRIME table.

|  |  |
| --- | --- |
| 1 | %sql select \* from CHICAGO\_CRIME\_DATA limit 10 |



1. How many crimes involve an arrest?

|  |  |
| --- | --- |
| 1 | %sql select count(\*) from CHICAGO\_CRIME\_DATA where arrest='True' |



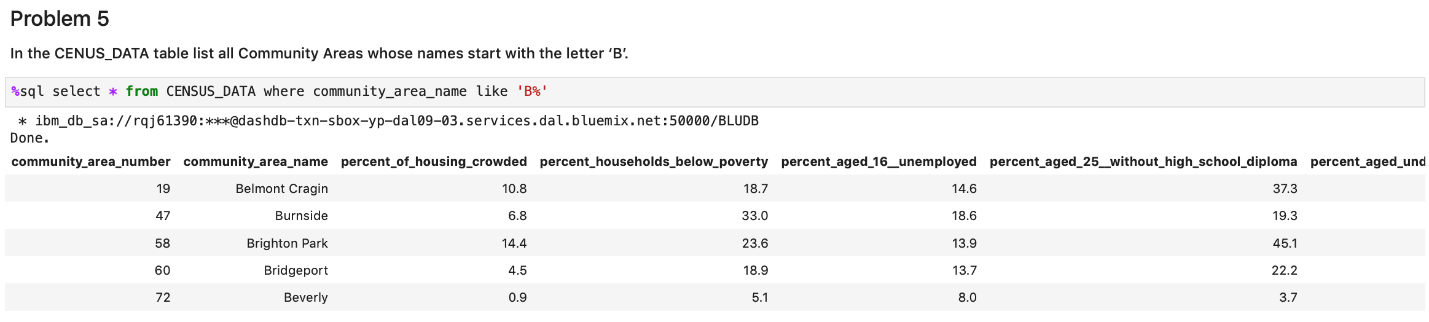
1. Which unique types of crimes (e.g. THEFT) have been recorded at a GAS STATION locations? (\*\**This answer maybe wrong. \**)

|  |  |
| --- | --- |
| 1 2 3 | %sql select distinct(primary\_type) from CHICAGO\_CRIME\_DATA where location\_description='GAS STATION' \ and primary\_type not in \ (select distinct(primary\_type) from CHICAGO\_CRIME\_DATA where location\_description != 'GAS STATION') |



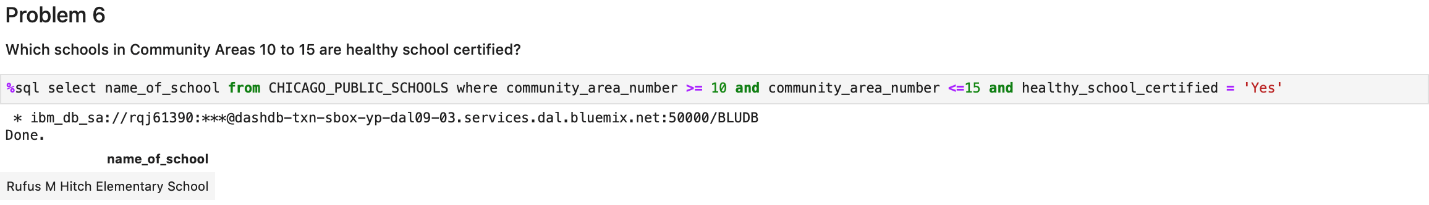
1. In the CENUS\_DATA table list all community areas whose names start with the letter ‘B’.

|  |  |
| --- | --- |
| 1 | %sql select \* from CENSUS\_DATA where community\_area\_name like 'B%' |



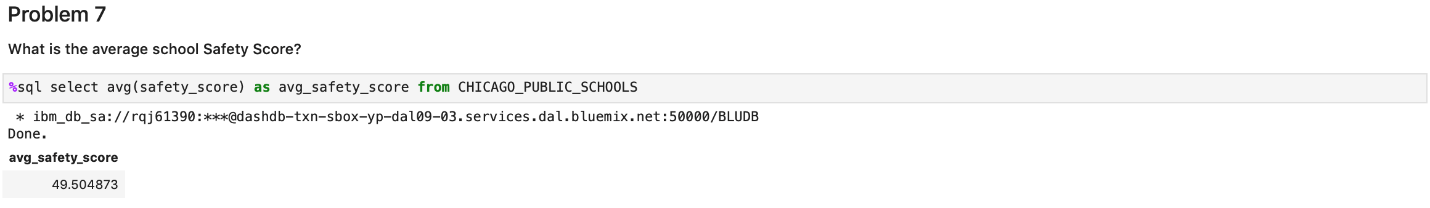
1. List the schools in community areas 10 to 15 that are healthy school certified.

|  |  |
| --- | --- |
| 1 | %sql select name\_of\_school from CHICAGO\_PUBLIC\_SCHOOLS where community\_area\_number >= 10 and community\_area\_number <=15 and healthy\_school\_certified = 'Yes' |



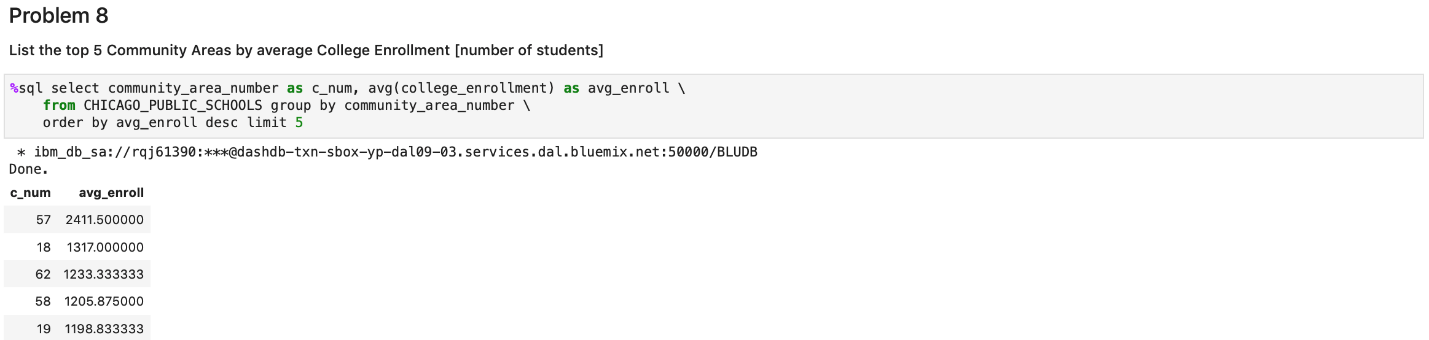
1. What is the average school Safety Score?

|  |  |
| --- | --- |
| 1 | %sql select avg(safety\_score) as avg\_safety\_score from CHICAGO\_PUBLIC\_SCHOOLS |



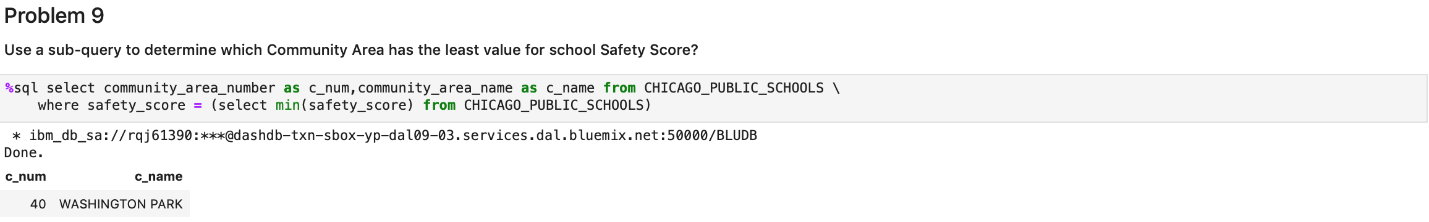
1. Find the top 5 Community Areas by average College Enrollment [number of students].

|  |  |
| --- | --- |
| 1 2 3 | %sql select community\_area\_number as c\_num, avg(college\_enrollment) as avg\_enroll \  from CHICAGO\_PUBLIC\_SCHOOLS group by community\_area\_number \  order by avg\_enroll desc limit 5 |



1. Use a sub-query todeterminewhich Community Area has the least value for school Safety Score?

|  |  |
| --- | --- |
| 1 2 | %sql select community\_area\_number as c\_num,community\_area\_name as c\_name from CHICAGO\_PUBLIC\_SCHOOLS \  where safety\_score = (select min(safety\_score) from CHICAGO\_PUBLIC\_SCHOOLS) |



1. [Without using an explicit JOIN operator] Find the Per Capita Income of the Community Area which has a school Safety Score of 1.

|  |  |
| --- | --- |
| 1 2 3 | # query in multi-table, we could use a sub-query instead of using an explicit JOIN operator %sql select per\_capita\_income from CENSUS\_DATA \  where community\_area\_number = (select community\_area\_number from CHICAGO\_PUBLIC\_SCHOOLS where safety\_score = 1) |

