Final Assignment - Databases and SQL for Data Science

Introduction

Using this Python notebook you will:

1. Understand 3 Chicago datasets
2. Load the 3 datasets into 3 tables in a Db2 database
3. Execute SQL queries to answer assignment questions

Understand the datasets

To complete the assignment problems in this notebook you will be using three datasets that are available on the city of Chicago's Data Portal:

1. Socioeconomic Indicators in Chicago
2. Chicago Public Schools
3. Chicago Crime Data

1. Socioeconomic Indicators in Chicago

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: 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: https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

2. 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: 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: https://data.cityofchicago.org/Education/Chicago-Public-Schools-Progress-Report-Cards-2011-/9xs2-f89t

3. 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: 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: 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.

Store the datasets in database tables

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 Week 3 Lab 3, 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 first dataset, Next create a New Table, and then follow the steps on-screen instructions to load the data. Name the new tables as folows:

1. **CENSUS_DATA**
2. **CHICAGO_PUBLIC_SCHOOLS**
3. **CHICAGO_CRIME_DATA**

Connect to the database

Let us first load the SQL extension and establish a connection with the database

```
In [2]:  %load_ext sql
```

```
In [14]:  # Remember the connection string is of the format:
          # 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
```

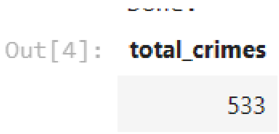
Problems

Now write and execute SQL queries to solve assignment problems

Problem 1

Find the total number of crimes recorded in the CRIME table

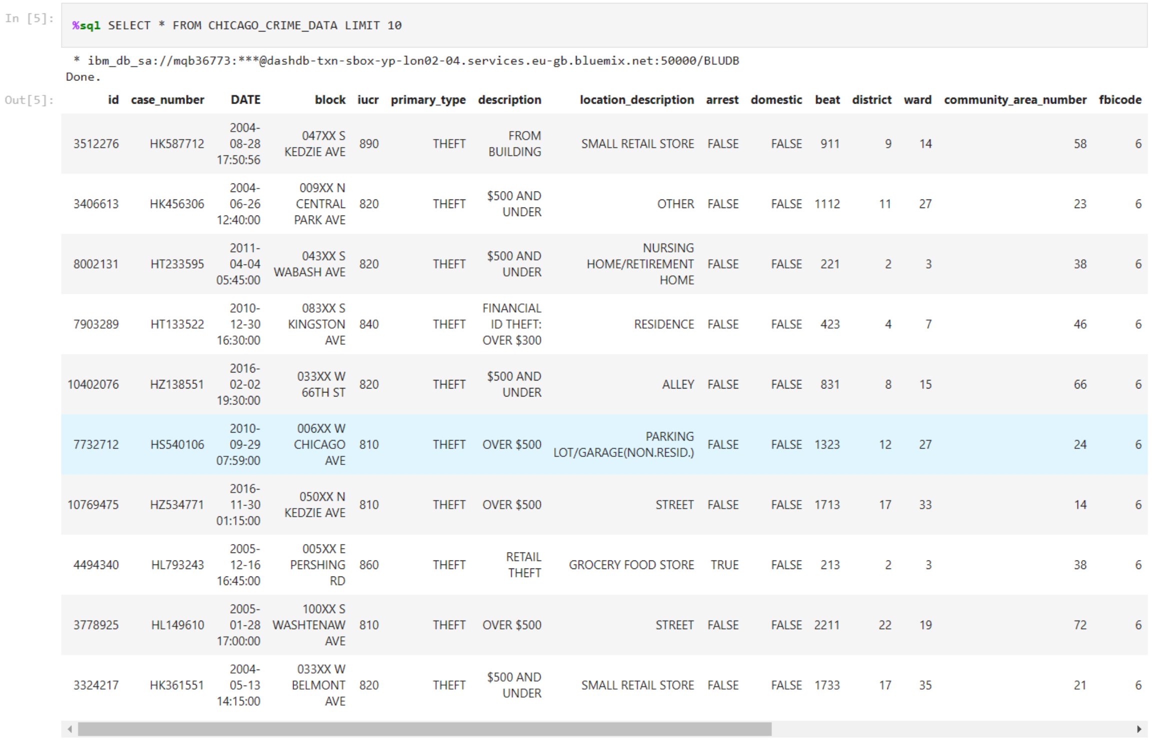
```
In [4]:  # Rows in Crime table

         %sql SELECT COUNT(*) AS TOTAL_CRIMES \
              FROM CHICAGO_CRIME_DATA
```



```
Done.
Out[4]:  total_crimes
                  533
```

Problem 2

Retrieve first 10 rows from the CRIME table

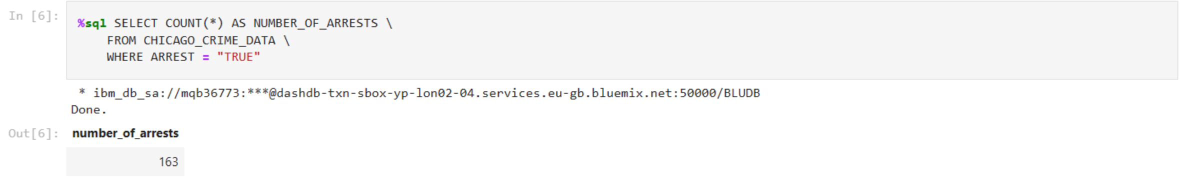


```
In [5]:  %sql SELECT * FROM CHICAGO_CRIME_DATA LIMIT 10

 * ibm_db_sa://mqb36773:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
Done.
Out[5]:
```

id	case_number	DATE	block	iucr	primary_type	description	location_description	arrest	domestic	beat	district	ward	community_area_number	fbicode
3512276	HK587712	2004-08-28 17:50:56	047XX S KEDZIE AVE	890	THEFT	FROM BUILDING	SMALL RETAIL STORE	FALSE	FALSE	911	9	14	58	6
3406613	HK456306	2004-06-26 12:40:00	009XX N CENTRAL PARK AVE	820	THEFT	$500 AND UNDER	OTHER	FALSE	FALSE	1112	11	27	23	6
8002131	HT233595	2011-04-04 05:45:00	043XX S WABASH AVE	820	THEFT	$500 AND UNDER	NURSING HOME/RETIREMENT HOME	FALSE	FALSE	221	2	3	38	6
7903289	HT133522	2010-12-30 16:30:00	083XX S KINGSTON AVE	840	THEFT	FINANCIAL ID THEFT: OVER $300	RESIDENCE	FALSE	FALSE	423	4	7	46	6
10402076	HZ138551	2016-02-02 19:30:00	033XX W 66TH ST	820	THEFT	$500 AND UNDER	ALLEY	FALSE	FALSE	831	8	15	66	6
7732712	HS540106	2010-09-29 07:59:00	006XX W CHICAGO AVE	810	THEFT	OVER $500	PARKING LOT/GARAGE(NON.RESID.)	FALSE	FALSE	1323	12	27	24	6
10769475	HZ534771	2016-11-30 01:15:00	050XX N KEDZIE AVE	810	THEFT	OVER $500	STREET	FALSE	FALSE	1713	17	33	14	6
4494340	HL793243	2005-12-16 16:45:00	005XX E PERSHING RD	860	THEFT	RETAIL THEFT	GROCERY FOOD STORE	TRUE	FALSE	213	2	3	38	6
3778925	HL149610	2005-01-28 17:00:00	100XX S WASHTENAW AVE	810	THEFT	OVER $500	STREET	FALSE	FALSE	2211	22	19	72	6
3324217	HK361551	2004-05-13 14:15:00	033XX W BELMONT AVE	820	THEFT	$500 AND UNDER	SMALL RETAIL STORE	FALSE	FALSE	1733	17	35	21	6

Problem 3

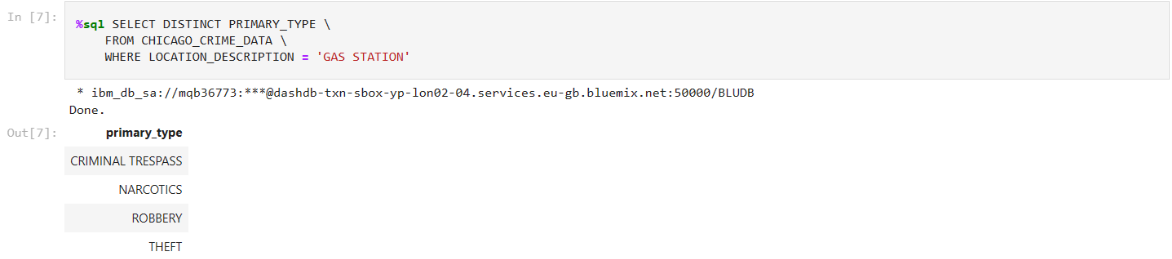
How many crimes involve an arrest?



```
In [6]: %sql SELECT COUNT(*) AS NUMBER_OF_ARRESTS \
            FROM CHICAGO_CRIME_DATA \
            WHERE ARREST = "TRUE"

         * ibm_db_sa://mqb36773:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
        Done.
Out[6]: number_of_arrests
                      163
```

Problem 4

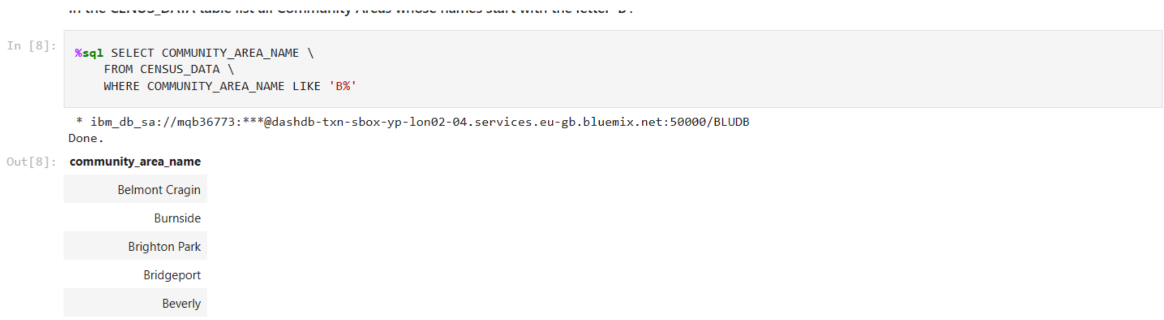
Which unique types of crimes have been recorded at GAS STATION locations?



```
In [7]: %sql SELECT DISTINCT PRIMARY_TYPE \
            FROM CHICAGO_CRIME_DATA \
            WHERE LOCATION_DESCRIPTION = 'GAS STATION'

         * ibm_db_sa://mqb36773:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
        Done.
Out[7]:         primary_type
        CRIMINAL TRESPASS
                NARCOTICS
                  ROBBERY
                    THEFT
```

Problem 5

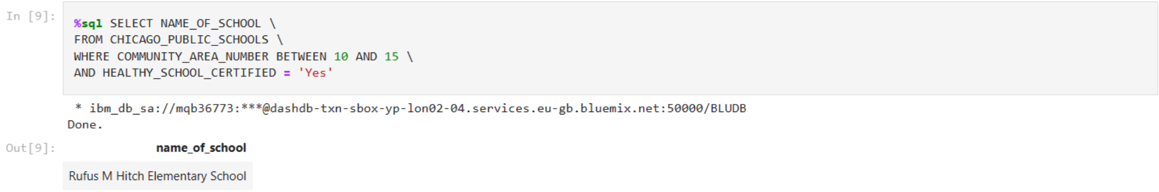
In the CENUS_DATA table list all Community Areas whose names start with the letter ‘B’.



```
In [8]: %sql SELECT COMMUNITY_AREA_NAME \
            FROM CENSUS_DATA \
            WHERE COMMUNITY_AREA_NAME LIKE 'B%'

         * ibm_db_sa://mqb36773:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
        Done.
Out[8]: community_area_name
             Belmont Cragin
                  Burnside
              Brighton Park
                Bridgeport
                   Beverly
```

Problem 6

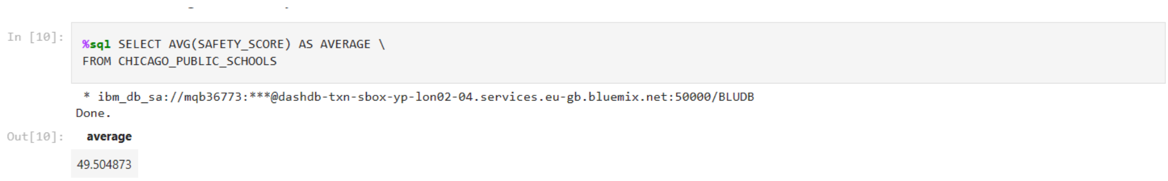
Which schools in Community Areas 10 to 15 are healthy school certified?



```
In [9]: %sql SELECT NAME_OF_SCHOOL \
        FROM CHICAGO_PUBLIC_SCHOOLS \
        WHERE COMMUNITY_AREA_NUMBER BETWEEN 10 AND 15 \
        AND HEALTHY_SCHOOL_CERTIFIED = 'Yes'

         * ibm_db_sa://mqb36773:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
        Done.
Out[9]:             name_of_school
        Rufus M Hitch Elementary School
```

Problem 7

What is the average school Safety Score?



In [10]:

```
%sql SELECT AVG(SAFETY_SCORE) AS AVERAGE \
FROM CHICAGO_PUBLIC_SCHOOLS
```

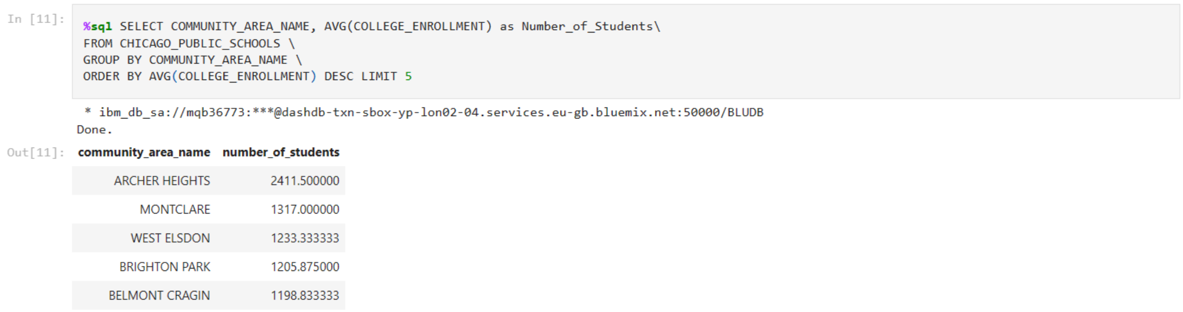
```
 * ibm_db_sa://mqb36773:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[10]:

average
49.504873

Problem 8

List the top 5 Community Areas by average College Enrollment [number of students]



In [11]:

```
%sql SELECT COMMUNITY_AREA_NAME, AVG(COLLEGE_ENROLLMENT) as Number_of_Students\
FROM CHICAGO_PUBLIC_SCHOOLS \
GROUP BY COMMUNITY_AREA_NAME \
ORDER BY AVG(COLLEGE_ENROLLMENT) DESC LIMIT 5
```

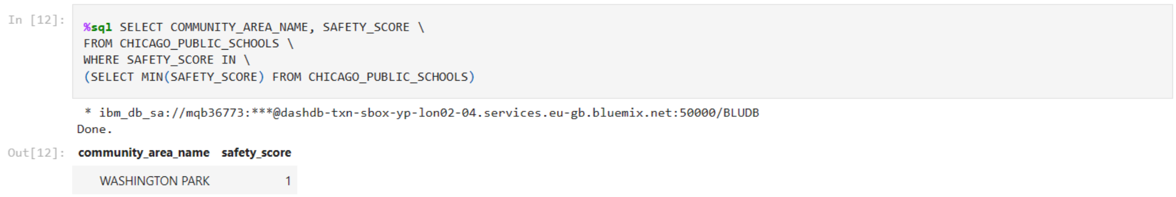
```
 * ibm_db_sa://mqb36773:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[11]:

community_area_name	number_of_students
ARCHER HEIGHTS	2411.500000
MONTCLARE	1317.000000
WEST ELSDON	1233.333333
BRIGHTON PARK	1205.875000
BELMONT CRAGIN	1198.833333

Problem 9

Use a sub-query to determine which Community Area has the least value for school Safety Score?



In [12]:

```
%sql SELECT COMMUNITY_AREA_NAME, SAFETY_SCORE \
FROM CHICAGO_PUBLIC_SCHOOLS \
WHERE SAFETY_SCORE IN \
(SELECT MIN(SAFETY_SCORE) FROM CHICAGO_PUBLIC_SCHOOLS)
```

```
 * ibm_db_sa://mqb36773:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[12]:

community_area_name	safety_score
WASHINGTON PARK	1

Problem 10

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

In [13]:

```
%sql SELECT CD.COMMUNITY_AREA_NAME, CD.PER_CAPITA_INCOME \
FROM CENSUS_DATA CD, CHICAGO_PUBLIC_SCHOOLS CPS \
WHERE CD.COMMUNITY_AREA_NUMBER = CPS.COMMUNITY_AREA_NUMBER \
AND CPS.SAFETY_SCORE = 1
```

```
 * ibm_db_sa://mqb36773:***@dashdb-txn-sbox-yp-lon02-04.services.eu-gb.bluemix.net:50000/BLUDB
Done.
```

Out[13]:

community_area_name	per_capita_income
Washington Park	13785