1: Analyzing ChicagoCensusData Set

These libraries are pre-installed in the Skills Network Lab environment I used. If running in another environment, please uncomment the lines below to install them:

```
# !pip install --force-reinstall ibm_db==3.1.0 ibm_db_sa==0.3.3
# Ensure we don't load_ext with sqlalchemy>=1.4 (incompadible)
# !pip uninstall sqlalchemy==1.4 -y && pip install sqlalchemy==1.3.24
# !pip install ipython-sql

# Load the SQL extension and establish a connection with the database
%load_ext sql

# Connection string for the Db2 on Cloud database instance
%sql
ibm_db_sa://bcy01016:pPV11zoSkjmBw2pU@19af6446-6171-4641-8aba-9dcff8e1b6ff.c1ogj3s
d0tgtu0lqde00.databases.appdomain.cloud:30699/bludb?security=SSL

# Output:
'Connected: bcy01016@bludb'
```

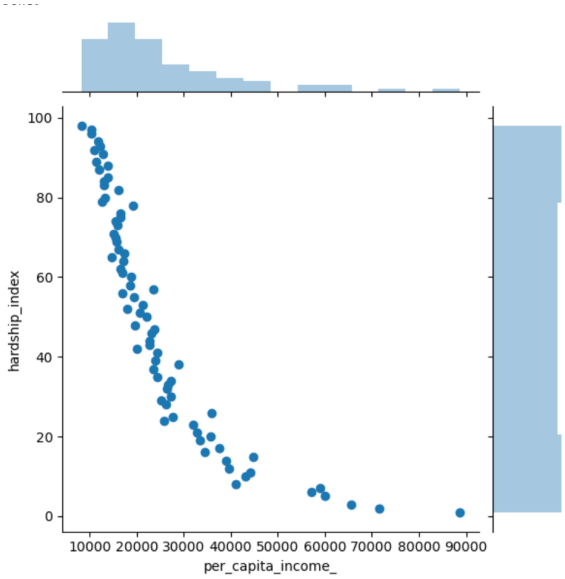
Store the data set in a table

```
# First, read the dataset source .CSV from the internet into a pandas dataframe
import pandas
chicago_socioeconomic_data =
pandas.read_csv('https://data.cityofchicago.org/resource/jcxq-k9xf.csv')

# Then, create a table in the database to store the dataset. The PERSIST command in SQL
"magic" simplifies the process of table creation and writing the data from a pandas dataframe
into the table
%sql PERSIST chicago_socioeconomic_data
```

Verify the table was created successfully by making a few basic queries

```
# Number of rows in the dataset
%sql SELECT COUNT(*) FROM chicago_socioeconomic_data;

# Output:
78

# Number of community areas in Chicago with a hardship index greater than 50.0
%sql SELECT COUNT(*) FROM chicago_socioeconomic_data WHERE hardship_index > 50.0;

# Output:
38
```

```
# Maximum value of hardship index in this dataset
%sql SELECT MAX(hardship_index) FROM chicago_socioeconomic_data;

# Output:
98.0

# Community area with the highest hardship index
%sql SELECT community_area_name FROM chicago_socioeconomic_data WHERE
hardship_index = (SELECT MAX(hardship_index) FROM chicago_socioeconomic_data);

# Output:
Community_area_name
Riverdale

# Chicago community areas with per-capita incomes greater than $60,000
%sql SELECT community_area_name FROM chicago_socioeconomic_data WHERE
per_capita_income_ > 60000;

# Output:
Community_area_name
Lake View
Lincoln Park
Near North Side
Loop

# Scatter plot using the variables per_capita_income_ and hardship_index.
# if the import command gives ModuleNotFoundError: No module named 'seaborn'
# then uncomment the following line i.e. delete the # to install the seaborn package
# !pip install seaborn==0.9.0

import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

income_vs_hardship = %sql SELECT per_capita_income_, hardship_index FROM
chicago_socioeconomic_data;
plot = sns.jointplot(x='per_capita_income_',y='hardship_index',
data=income_vs_hardship.DataFrame())

# Output:
```



Analysis: You can see that as Per Capita Income rises, the Hardship Index decreases. We see that the points on the scatter plot are somewhat closer to a straight line in the negative direction. Therefore, we have a negative correlation between the two variables.

2: Analyzing ChicagoPublicSchoolsData Set

To store this data set in the database, I manually created the table using the database console LOAD tool, rather than reading the dataset into a Pandas dataframe and then using the (Magic SQL) PERSIST command to write the data from the dataframe into the table. I did this manually in order to avoid any possibility of mapping to default data types, which may not be optimal for SQL querying.

Verify that the table creation was successful by retrieving the metadata of the SCHOOLS table
%sql select * from SYSCAT.TABLES where TABNAME = 'SCHOOLS'

Output:

tabschema	tabname	owner	ownertype	TYPE	status	base_tabschema	base_tabname	rowtypeschema	rowtypename	create_time	alter_time	invalidate_time	stats_time
BCY01016	SCHOOLS	BCY01016	U	T	N	None	None	None	None	2023-04-17 02:20:59.836922	2023-04-17 02:20:59.836922	2023-04-17 02:20:59.836922	2023-04-17 02:23:14.866899

Query the database system catalog to retrieve column metadata
Number of columns in this table
%sql select COUNT(*) from SYSCAT.COLUMNS where TABNAME = 'SCHOOLS'

Output:
78

Columns in the SCHOOLS table and their column datatype and length
%sql select COLNAME, TYPENAME, LENGTH from SYSCAT.COLUMNS where TABNAME = 'SCHOOLS'

Output:

[7]:

colname	typename	length
SCHOOL_ID	INTEGER	4
NAME_OF_SCHOOL	VARCHAR	64
Elementary, Middle, or High School	VARCHAR	2
STREET_ADDRESS	VARCHAR	29
CITY	VARCHAR	7
STATE	VARCHAR	2
ZIP_CODE	INTEGER	4
PHONE_NUMBER	VARCHAR	14
LINK	VARCHAR	78
NETWORK_MANAGER	VARCHAR	40
COLLABORATIVE_NAME	VARCHAR	34
ADEQUATE_YEARLY_PROGRESS_MADE_	VARCHAR	3
TRACK_SCHEDULE	VARCHAR	12
CPS_PERFORMANCE_POLICY_STATUS	VARCHAR	16
CPS_PERFORMANCE_POLICY_LEVEL	VARCHAR	15
HEALTHY_SCHOOL_CERTIFIED	VARCHAR	3
SAFETY_ICON	VARCHAR	11
SAFETY_SCORE	SMALLINT	2
FAMILY_INVOLVEMENT_ICON	VARCHAR	11
FAMILY_INVOLVEMENT_SCORE	VARCHAR	3
ENVIRONMENT_ICON	VARCHAR	11
ENVIRONMENT_SCORE	SMALLINT	2
INSTRUCTION_ICON	VARCHAR	11
INSTRUCTION_SCORE	SMALLINT	2
LEADERS_ICON	VARCHAR	4
LEADERS_SCORE	VARCHAR	3
TEACHERS_ICON	VARCHAR	11

TEACHERS_SCORE	VARCHAR	3
PARENT_ENGAGEMENT_ICON	VARCHAR	7
PARENT_ENGAGEMENT_SCORE	VARCHAR	3
PARENT_ENVIRONMENT_ICON	VARCHAR	7
PARENT_ENVIRONMENT_SCORE	VARCHAR	3
AVERAGE_STUDENT_ATTENDANCE	VARCHAR	6
RATE_OF_MISCONDUCTS__PER_100_STUDENTS_	DECIMAL	5
AVERAGE_TEACHER_ATTENDANCE	VARCHAR	6
INDIVIDUALIZED_EDUCATION_PROGRAM_COMPLIANCE_RATE	VARCHAR	7
PK_2_LITERACY__	VARCHAR	4
PK_2_MATH__	VARCHAR	4
GR3_5_GRADE_LEVEL_MATH__	VARCHAR	4
GR3_5_GRADE_LEVEL_READ__	VARCHAR	4
GR3_5_KEEP_PACE_READ__	VARCHAR	4
GR3_5_KEEP_PACE_MATH__	VARCHAR	4
GR6_8_GRADE_LEVEL_MATH__	VARCHAR	4
GR6_8_GRADE_LEVEL_READ__	VARCHAR	4
GR6_8_KEEP_PACE_MATH_	VARCHAR	4
GR6_8_KEEP_PACE_READ__	VARCHAR	4
GR_8_EXPLORE_MATH__	VARCHAR	4
GR_8_EXPLORE_READ__	VARCHAR	4
ISAT_EXCEEDING_MATH__	DECIMAL	4
ISAT_EXCEEDING_READING__	DECIMAL	4
ISAT_VALUE_ADD_MATH	DECIMAL	3
ISAT_VALUE_ADD_READ	DECIMAL	3
ISAT_VALUE_ADD_COLOR_MATH	VARCHAR	6
ISAT_VALUE_ADD_COLOR_READ	VARCHAR	6

STUDENTS_TAKING__ALGEBRA__	VARCHAR	4
STUDENTS_PASSING__ALGEBRA__	VARCHAR	4
9th Grade EXPLORE (2009)	VARCHAR	4
9th Grade EXPLORE (2010)	VARCHAR	4
10th Grade PLAN (2009)	VARCHAR	4
10th Grade PLAN (2010)	VARCHAR	4
NET_CHANGE_EXPLORE_AND_PLAN	VARCHAR	3
11th Grade Average ACT (2011)	VARCHAR	4
NET_CHANGE_PLAN_AND_ACT	VARCHAR	3
COLLEGE_ELIGIBILITY__	VARCHAR	4
GRADUATION_RATE__	VARCHAR	4
COLLEGE_ENROLLMENT_RATE__	VARCHAR	4
COLLEGE_ENROLLMENT	SMALLINT	2
GENERAL_SERVICES_ROUTE	SMALLINT	2
FRESHMAN_ON_TRACK_RATE__	VARCHAR	4
X_COORDINATE	DECIMAL	13
Y_COORDINATE	DECIMAL	13
LATITUDE	DECIMAL	18
LONGITUDE	DECIMAL	18
COMMUNITY_AREA_NUMBER	SMALLINT	2
COMMUNITY_AREA_NAME	VARCHAR	22
WARD	SMALLINT	2
POLICE_DISTRICT	SMALLINT	2
LOCATION	VARCHAR	27

Number of elementary schools in the dataset
%sql select count(*) from SCHOOLS where "Elementary, Middle, or High School" = 'ES'

Output:
462

Highest safety score
%sql select MAX(SAFETY_SCORE) AS MAX_SAFETY_SCORE from SCHOOLS

Output:
Max_safety_score
99

Schools with the highest safety score
```
%sql select Name_of_School, Safety_Score from SCHOOLS where \
  Safety_Score= (select MAX(Safety_Score) from SCHOOLS)
```

Output:

name_of_school	safety_score
Abraham Lincoln Elementary School	99
Alexander Graham Bell Elementary School	99
Annie Keller Elementary Gifted Magnet School	99
Augustus H Burley Elementary School	99
Edgar Allan Poe Elementary Classical School	99
Edgebrook Elementary School	99
Ellen Mitchell Elementary School	99
James E McDade Elementary Classical School	99
James G Blaine Elementary School	99
LaSalle Elementary Language Academy	99
Mary E Courtenay Elementary Language Arts Center	99
Northside College Preparatory High School	99
Northside Learning Center High School	99
Norwood Park Elementary School	99
Oriole Park Elementary School	99
Sauganash Elementary School	99
Stephen Decatur Classical Elementary School	99
Talman Elementary School	99
Wildwood Elementary School	99

Top 10 schools with the highest average student attendance
```
%sql select Name_of_School, Average_Student_Attendance from SCHOOLS \
   order by Average_Student_Attendance desc nulls last limit 10
```

Output:

name_of_school	average_student_attendance
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%

The 5 Schools with the lowest average student attendance sorted in ascending order based on attendance¶
%sql SELECT Name_of_School, Average_Student_Attendance from SCHOOLS order by Average_Student_Attendance nulls last LIMIT 5

Output:

name_of_school	average_student_attendance
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%

Schools with an average student attendance lower than 70%
%sql SELECT Name_of_School, Average_Student_Attendance from SCHOOLS where Average_Student_Attendance < '70%' order by Average_Student_Attendance

Output:

name_of_school	average_student_attendance
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%

Total college enrollment for each community area
%sql SELECT Community_Area_Name, SUM(College_Enrollment) AS TOTAL_ENROLLMENT from SCHOOLS group by Community_Area_Name

Output:

community_area_name	total_enrollment
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

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 SCHOOLS group by Community_Area_Name order by Total_Enrollment asc Limit 5

Output:

community_area_name	total_enrollment
OAKLAND	140
FULLER PARK	531
BURNSIDE	549
OHARE	786
LOOP	871

The 5 schools with the lowest safety score
%sql SELECT Name_of_School, Safety_Score from SCHOOLS order by Safety_Score asc Limit 5

Output:

name_of_school	safety_score
Edmond Burke Elementary School	1
Luke O'Toole Elementary School	5
George W Tilton Elementary School	6
Foster Park Elementary School	11
Emil G Hirsch Metropolitan High School	13

Hardship index for the community area which has college enrollment of 4368
%%sql
select hardship_index
  from chicago_socioeconomic_data CD, schools CPS
  where CD.ca = CPS.community_area_number /* ca represents Community_Area_number in the ChicagoCensusData table */
    and college_enrollment = 4368 /* this metric is in the SCHOOLS table */

Output:
Hardship_index:
6

Hardship index for the community area which has the school with the highest enrollment
%sql select ca, community_area_name, hardship_index from chicago_socioeconomic_data \
  where ca in \
  (select community_area_number from schools order by college_enrollment desc limit 1)

Output:

ca	community_area_name	hardship_index
5.0	North Center	6.0

3: Analyzing all 3 Data Sets (Census, School, and Crime)

Total number of crimes recorded in the CRIME table
%sql select count(*) from CHICAGO_CRIME_DATA

Output:
533

Community areas with per capita income less than 11,000
%%sql
select unique(COMMUNITY_AREA_NAME), Per_Capita_Income from CENSUS_DATA, CHICAGO_CRIME_DATA
where CENSUS_DATA.Community_Area_Number = CHICAGO_CRIME_DATA.Community_Area_Number and Per_Capita_Income < 11000

Output:

community_area_name	per_capita_income
Fuller Park	10432
Riverdale	8201
South Lawndale	10402
West Garfield Park	10934

All case numbers for crimes involving minors (children are not considered in this case)
%%sql
select Case_number,DESCRIPTION from CHICAGO_CRIME_DATA where LCASE(DESCRIPTION) like '%minor%'

Output:

case_number	description
HL266884	SELL/GIVE/DEL LIQUOR TO MINOR
HK238408	ILLEGAL CONSUMPTION BY MINOR

List all kidnapping crimes involving a child
%%sql
select Case_number, Primary_Type, DESCRIPTION from CHICAGO_CRIME_DATA where LCASE(PRIMARY_TYPE) like 'kidnapping' and LCASE(DESCRIPTION) like '%child%'

Output:

case_number	primary_type	description
HN144152	KIDNAPPING	CHILD ABDUCTION/STRANGER

Types of crimes that were recorded at schools
%%sql
select UNIQUE(Primary_Type) from CHICAGO_CRIME_DATA

Output:

primary_type
ARSON
ASSAULT
BATTERY
BURGLARY
CONCEALED CARRY
CRIM SEXUAL ASS
CRIMINAL DAMAGE
CRIMINAL TRESPA
DECEPTIVE PRACT
DOMESTIC VIOLEN
GAMBLING
HOMICIDE
HUMAN TRAFFICKI
INTERFERENCE WI
INTIMIDATION
KIDNAPPING
LIQUOR LAW VIOL
MOTOR VEHICLE T
NARCOTICS
NON - CRIMINAL
NON-CRIMINAL
NON-CRIMINAL (S
OBSCENITY
OFFENSE INVOLVI
OTHER NARCOTIC
OTHER OFFENSE
PROSTITUTION
PUBLIC INDECENC
PUBLIC PEACE VI
RITUALISM
ROBBERY
SEX OFFENSE
STALKING
THEFT
WEAPONS VIOLATI

Average safety score for each type of school
%%sql
select "Elementary, Middle, or High School",AVG(Safety_Score) AS "Average Safety Score" from CHICAGO_PUBLIC_SCHOOLS group by "Elementary, Middle, or High School"

Output:

Elementary, Middle, or High School	Average Safety Score
ES	49
HS	49
MS	48

The 5 community areas with highest % of households below poverty line
%%sql
select Community_Area_Name, Percent_Households_Below_Poverty from CENSUS_DATA order by Percent_Households_Below_Poverty desc limit 5

Output:

community_area_name	percent_households_below_poverty
Riverdale	56.5
Fuller Park	51.2
Englewood	46.6
North Lawndale	43.1
East Garfield Park	42.4

Most crime prone community area
%%sql
select Community_Area_Number, count(*) as Number_of_crimes from CHICAGO_CRIME_DATA group by Community_Area_Number order by Number_of_crimes desc Limit 1

Output:

community_area_number	number_of_crimes
25	43

Community area with highest hardship index
%%sql
select Community_Area_name,Hardship_index from CENSUS_DATA where Hardship_Index = (select MAX(Hardship_Index) from CENSUS_DATA)

Output:

community_area_name	hardship_index
Riverdale	98