AKRAM ALZAGHIR

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

Estimated time needed: **30** minutes

Objectives

After completing this lab you will be able to:

- Understand the dataset for Chicago Public School level performance
- Store the dataset in an Db2 database on IBM Cloud instance
- 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

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 (https://data.cityofchicago.org/Education/Chicago-Public-Schools-Progress-Report-Cards-2011-/9xs2-f89t? cm_mmc=Email_Newsletter-_-Developer_Ed%2BTech-_-WW_WW-_-SkillsNetwork-Courses-IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork-20127838&cm_mmca1=000026UJ&cm_mmca2=10006555&cm_mmca3=M12345678&cvosrc=email.Newslette _-Developer_Ed%2BTech-_-WW_WW-_-SkillsNetwork-Courses-IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork-20127838&cm_mmca1=000026UJ&cm_mmca2=10006555&cm_mmca3=M12345678&cvosrc=email.Newslette

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 (https://data.cityofchicago.org/api/assets/AAD41A13-BE8A-4E67-B1F5-86E711E09D5F? download=true&cm_mmc=Email_Newsletter-_-Developer_Ed%2BTech-_-WW_WW-_-SkillsNetwork-Courses-IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork-20127838&cm_mmca1=000026UJ&cm_mmca2=10006555&cm_mmca3=M12345678&cvosrc=email.Newslette

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 (https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DB0201EN-SkillsNetwork/labs/FinalModule_edX/data/Chicago_Public_Schools_-_Progress_Report_Cards__2011-2012-v3.csv).
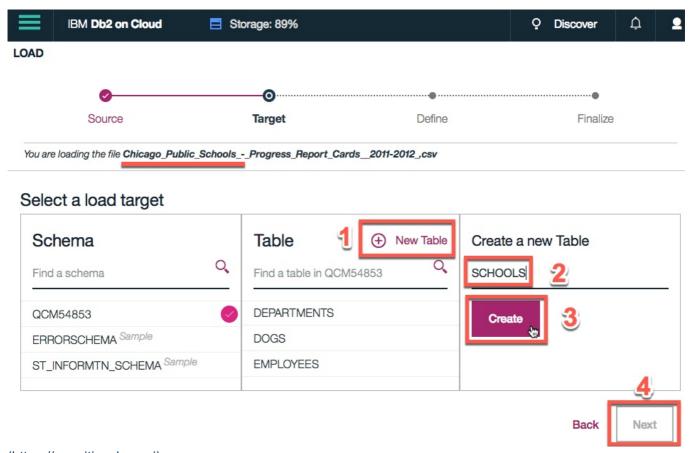
Now review some of its contents.

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.

(https://cognitiveclass.ai)

Connect to the database

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

In [1]:

```
%load_ext sql
```

In [2]:

```
# 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://tnm91075:krrsj3j7zjn%40nphp@dashdb-txn-sbox-yp-dal09-12.services.dal.b
luemix.net:50000/BLUDB
```

Out[2]:

'Connected: tnm91075@BLUDB'

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 [3]:

```
# type in your query to retrieve list of all tables in the database for your db2 schema
(username)
%sql select * from SCHOOLS limit 3
```

 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.

Out[3]:

School_ID	name_of_school	Elementary, Middle, or High School	Street_Address	City	State	ZIP_Code	Phone_N
610038	Abraham Lincoln Elementary School	ES	615 W Kemper Pl	Chicago	IL	60614	(773) 53
610281	Adam Clayton Powell Paideia Community Academy Elementary School	ES	7511 S South Shore Dr	Chicago	IL	60649	(773) 53
610185	Adlai E Stevenson Elementary School	ES	8010 S Kostner Ave	Chicago	IL	60652	(773) 53

◄   ▶

Click here for the solution

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 [4]:

```
# type in your query to retrieve the number of rows in the SCHOOLS table
%sql select count(*) as num_rows from SCHOOLS
```

 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.

Out[4]:

num_rows
566

In [5]:

```
#In Db2 the system catalog table called SYSCAT.COLUMNS contains the column metadata

%sql select count(*) from SYSCAT.COLUMNS where TABNAME = 'SCHOOLS'
```

 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.

Out[5]:

1
78

Click here for the solution

Now retrieve the the list of columns in SCHOOLS table and their column type (datatype) and length.

In [6]:

```
# type in your query to retrieve all column names in the SCHOOLS table along with their
datatypes and length
%sql select COLNAME, TYPENAME, LENGTH from SYSCAT.COLUMNS where TABNAME = 'SCHOOLS'
```

```
 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.
```

Out[6]:

colname	typename	length
School_ID	INTEGER	4
NAME_OF_SCHOOL	VARCHAR	65
Elementary, Middle, or High School	VARCHAR	2
Street_Address	VARCHAR	30
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	11
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

colname	typename	length
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

colname	typename	length
Police_District	SMALLINT	2
Location	VARCHAR	27

Click here for the solution

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 "_"?

Problems

Problem 1

How many Elementary Schools are in the dataset?

In [7]:

```
#Does the column name have mixed case, spaces or other special characters?
#If so, ensure you use double quotes around the "Name of the Column"
#so, we put the column name Elementary, Middle, or High School inside double quotes
%sql select count(*) as Total_ES from SCHOOLS where "Elementary, Middle, or High School" = 'ES'
```

 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.

Out[7]:

total_es
462

Click here for the hint

Click here for the solution

Problem 2

What is the highest Safety Score?

In [8]:

```
%sql select max(safety_score) as highest_score from SCHOOLS
```

 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.

Out[8]:

highest_score
99

Click here for the solution

Problem 3

Which schools have highest Safety Score?

In [9]:

```sql
%sql select name_of_school, safety_score from SCHOOLS \
where safety_score = (select max(safety_score) from SCHOOLS)
```

 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.

Out[9]:

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

Click here for the solution

Problem 4

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

In [10]:

```
# nulls last is to put null value at last, nulls first is to put the nulls value in the
first row
%sql select name_of_school, average_student_attendance from SCHOOLS \
order by average_student_attendance desc nulls last limit 10
```

* ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.

Out[10]:

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%

Click here for the solution

Problem 5

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

In [11]:

```
%sql select name_of_school, average_student_attendance from SCHOOLS \
order by average_student_attendance asc nulls last limit 5

#or
#%sql SELECT Name_of_School, Average_Student_Attendance  \
 #     from SCHOOLS \
  #    order by Average_Student_Attendance \
   #   fetch first 5 rows only
```

 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.

Out[11]:

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%

Click here for the solution

Problem 6

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

In [12]:

```
#REPLACE(string, 'old_string', 'new_string')
%sql SELECT Name_of_School, REPLACE(Average_Student_Attendance, '%', '') \
     from SCHOOLS \
     order by Average_Student_Attendance \
     fetch first 5 rows only
```

 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.

Out[12]:

name_of_school	2
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

Click here for the solution

Problem 7

Which Schools have Average Student Attendance lower than 70%?

In [13]:

```
%sql select Name_of_School, Average_Student_Attendance from SCHOOLS \
where Average_Student_Attendance < '70%'\
order by Average_Student_Attendance

#or
#%sql SELECT Name_of_School, Average_Student_Attendance  \
 #    from SCHOOLS \
  #   where CAST ( REPLACE(Average_Student_Attendance, '%', '') AS DOUBLE ) < 70 \
   #  order by Average_Student_Attendance

#or,

#%sql SELECT Name_of_School, Average_Student_Attendance  \
 #    from SCHOOLS \
  #   where DECIMAL ( REPLACE(Average_Student_Attendance, '%', '') ) < 70 \
   #  order by Average_Student_Attendance
```

 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.

Out[13]:

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%

Click here for the hint

Click here for the solution

Problem 8

Get the total College Enrollment for each Community Area

In [14]:

```
#%sql select Community_Area_Name, count(College_Enrollment) as total_College_Enrollment \
#from SCHOOLS group by Community_Area_Name
%sql select Community_Area_Name, sum(College_Enrollment) AS TOTAL_ENROLLMENT \
    from SCHOOLS \
    group by Community_Area_Name
```

```
  * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.
```

Out[14]:

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

community_area_name	total_enrollment
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

community_area_name	total_enrollment
WOODLAWN	4206

Click here for the hint

Click here for the solution

Problem 9

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

In [15]:

```
%sql select Community_Area_Name, sum(College_Enrollment) as TOTAL_ENROLLMENT \
from SCHOOLS group by Community_Area_Name \
order by TOTAL_ENROLLMENT limit 5
#or
#order by TOTAL_ENROLLMENT asc fetch first 5 rows only
```

 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.

Out[15]:

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

Click here for the solution

Problem 10

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

In [16]:

```
# College_Enrollment is inside SCHOOLS table,
# hardship_index is in chicago_socioeconomic_data table
%sql select hardship_index from chicago_socioeconomic_data CD, SCHOOLS SL \
where CD.ca = SL.community_area_number and College_Enrollment = 4368
# or use inner join as it match
# for inner join, we can use and clause also and it will return same result
# this is because the inner join will alawys return the values match
# both tables only, so if we put a clause, it will follow the clause result only
%sql select hardship_index from chicago_socioeconomic_data CD inner join SCHOOLS SL \
on CD.ca = SL.community_area_number where College_Enrollment = 4368
# or left join, return hardship_index value from left table (CD)
# when College_Enrollment = 4368 from right table (SL)
# if we use and clause here, it will return the value for hardship_index
#when College_Enrollment = 4368 first, then it will list the rest of values (all)
# for hardship_index for left table
%sql select hardship_index from chicago_socioeconomic_data CD left join SCHOOLS SL \
on CD.ca = SL.community_area_number where College_Enrollment = 4368
```

 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.
 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.
 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.

Out[16]:

hardship_index
6.0

Click here for the solution

Problem 11

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

In [17]:

```sql
%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);

# or
%sql select ca, community_area_name, hardship_index \
from chicago_socioeconomic_data \
where ca in (select community_area_number \
from SCHOOLS where College_Enrollment = (select max(College_Enrollment) from SCHOOLS));

# or
%sql select ca, hardship_index \
 from chicago_socioeconomic_data CD full outer join SCHOOLS SL\
 on CD.ca = SL.community_area_number \
 where College_Enrollment = (select max(College_Enrollment) from SCHOOLS);
```

 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.
 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.
 * ibm_db_sa://tnm91075:***@dashdb-txn-sbox-yp-dal09-12.services.dal.bluem
ix.net:50000/BLUDB
Done.

Out[17]:

ca	hardship_index
5.0	6.0

Click here for the solution

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.

Author

Rav Ahuja (https://www.linkedin.com/in/ravahuja/)

Change Log

Date (YYYY-MM-DD)	Version	Changed By	Change Description
2020-08-28	2.0	Lavanya	Moved lab to course repo in GitLab