Name: Syed Noor Razi Ali
StudentID:2070326

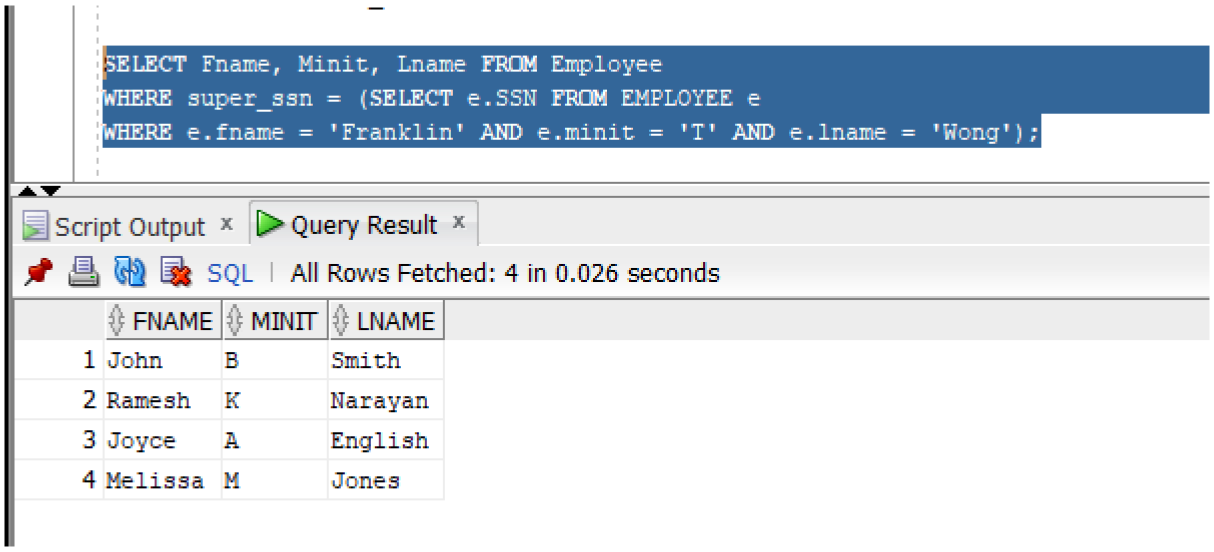
DSC 450: Database Processing for Large-Scale Analytics

Assignment Module 5

Part 1

Using the company.sql database (posted in with this assignment), write the following SQL queries.

1. Find the names of all employees who are directly supervised by 'Franklin T Wong' (you cannot use Franklin's SSN value in the query).

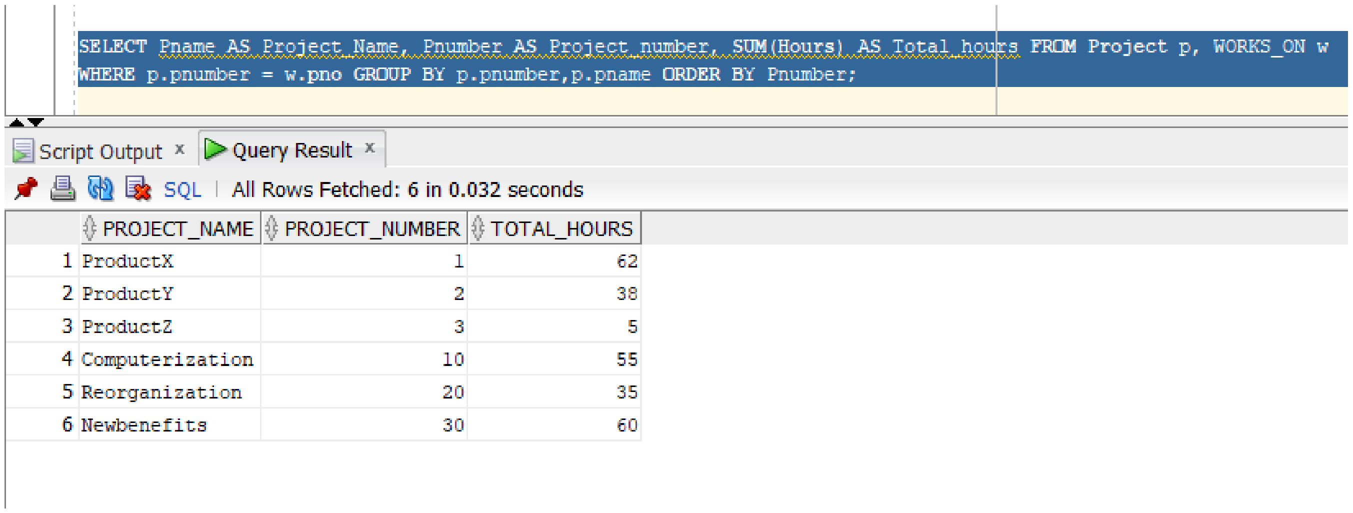


```
SELECT Fname, Minit, Lname FROM Employee
WHERE super_ssn = (SELECT e.SSN FROM EMPLOYEE e
WHERE e.fname = 'Franklin' AND e.minit = 'T' AND e.lname = 'Wong');
```

All Rows Fetched: 4 in 0.026 seconds

	FNAME	MINIT	LNAME
1	John	B	Smith
2	Ramesh	K	Narayan
3	Joyce	A	English
4	Melissa	M	Jones

2. For each project, list the project name, project number, and the total hours per week (by all employees) spent on that project.

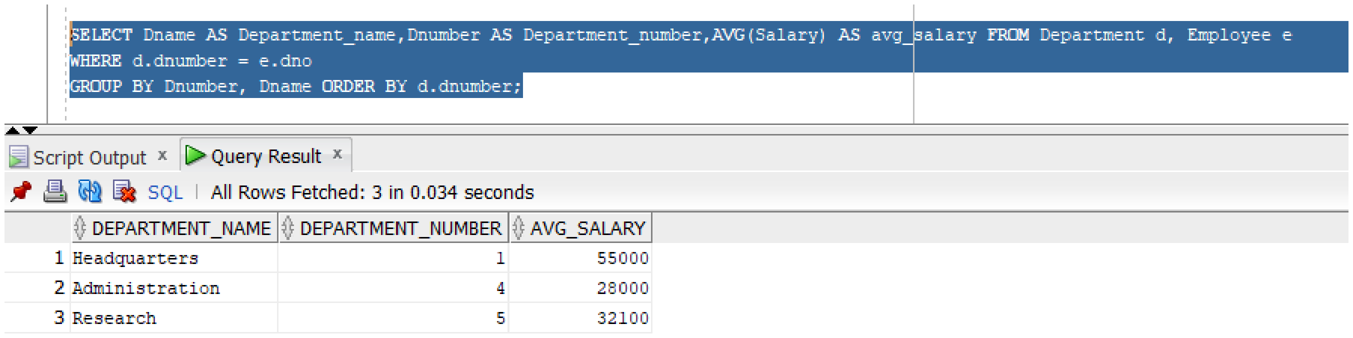


```
SELECT Pname AS Project_Name, Pnumber AS Project_number, SUM(Hours) AS Total_hours FROM Project p, WORKS_ON w
WHERE p.pnumber = w.pno GROUP BY p.pnumber,p.pname ORDER BY Pnumber;
```

All Rows Fetched: 6 in 0.032 seconds

	PROJECT_NAME	PROJECT_NUMBER	TOTAL_HOURS
1	ProductX	1	62
2	ProductY	2	38
3	ProductZ	3	5
4	Computerization	10	55
5	Reorganization	20	35
6	Newbenefits	30	60

3. For each department, retrieve the department name and the average salary of all employees working in that department. Order the output by department number in ascending order.



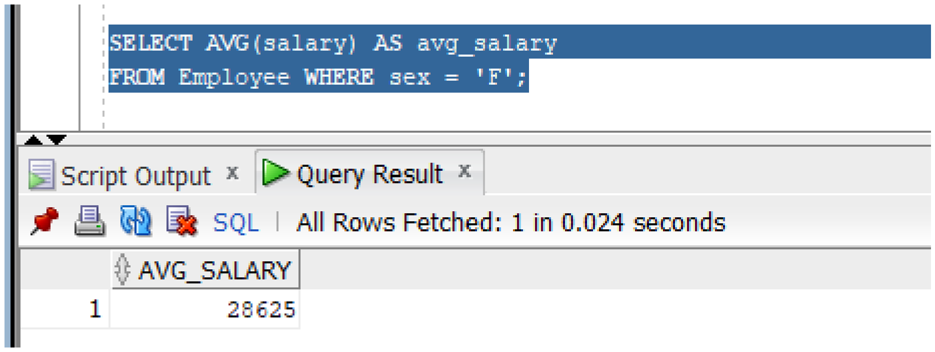
```
SELECT Dname AS Department_name,Dnumber AS Department_number,AVG(Salary) AS avg_salary FROM Department d, Employee e
WHERE d.dnumber = e.dno
GROUP BY Dnumber, Dname ORDER BY d.dnumber;
```

Script Output × Query Result ×

SQL | All Rows Fetched: 3 in 0.034 seconds

	DEPARTMENT_NAME	DEPARTMENT_NUMBER	AVG_SALARY
1	Headquarters	1	55000
2	Administration	4	28000
3	Research	5	32100

4. Retrieve the average salary of all female employees.



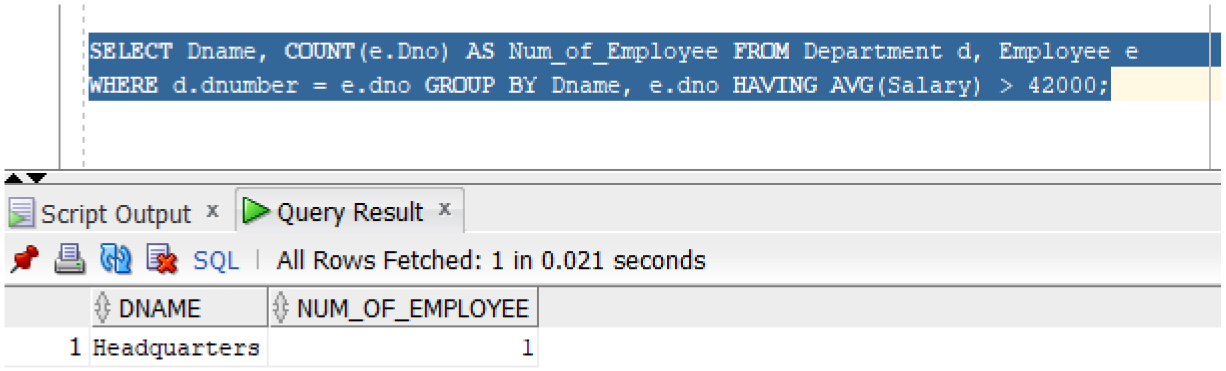
```
SELECT AVG(salary) AS avg_salary
FROM Employee WHERE sex = 'F';
```

Script Output × Query Result ×

SQL | All Rows Fetched: 1 in 0.024 seconds

	AVG_SALARY
1	28625

5. For each department whose average salary is greater than $42,000, retrieve the department name and the number of employees in that department.



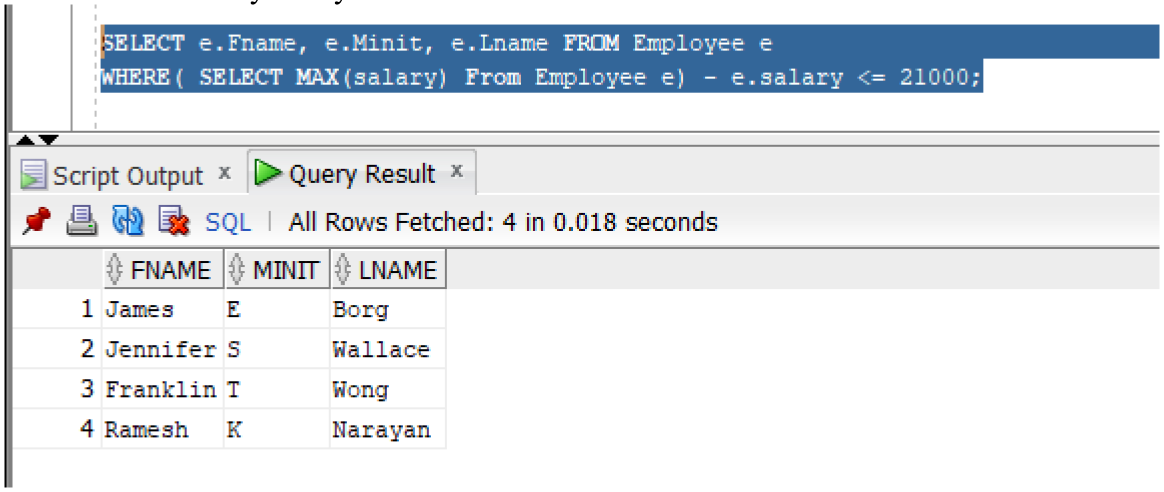
```
SELECT Dname, COUNT(e.Dno) AS Num_of_Employee FROM Department d, Employee e
WHERE d.dnumber = e.dno GROUP BY Dname, e.dno HAVING AVG(Salary) > 42000;
```

Script Output × Query Result ×

SQL | All Rows Fetched: 1 in 0.021 seconds

	DNAME	NUM_OF_EMPLOYEE
1	Headquarters	1

6. Retrieve the names of employees whose salary is within $21,000 of the salary of the employee who is paid the most in the company (e.g., if the highest salary in the company is $83,000, retrieve the names of all employees that make at least $62,000.). Naturally, your query should work for any salary.



```
SELECT e.Fname, e.Minit, e.Lname FROM Employee e
WHERE( SELECT MAX(salary) From Employee e) - e.salary <= 21000;
```

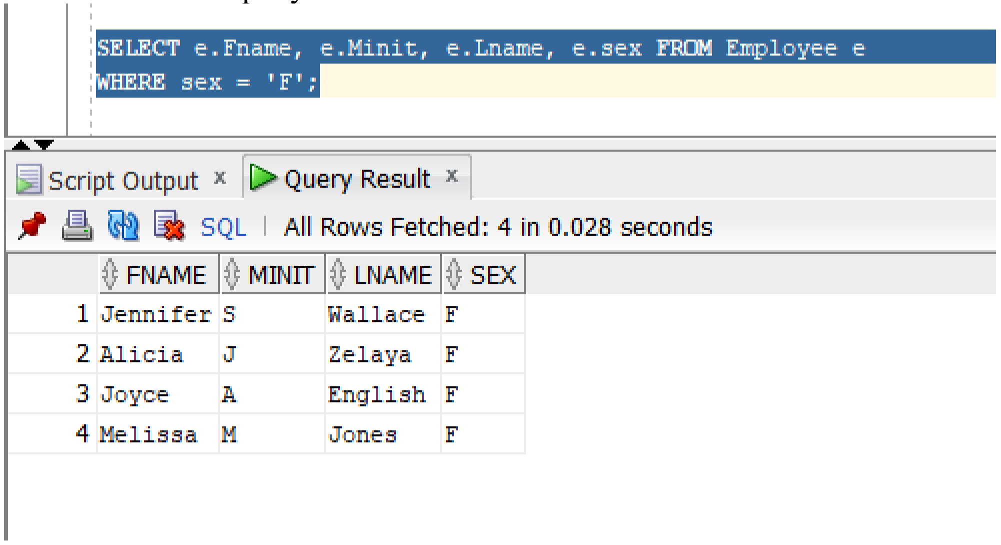
Script Output × | Query Result ×

SQL | All Rows Fetched: 4 in 0.018 seconds

	FNAME	MINIT	LNAME
1	James	E	Borg
2	Jennifer	S	Wallace
3	Franklin	T	Wong
4	Ramesh	K	Narayan

7. Find all female employees using:

a. Plain SELECT query



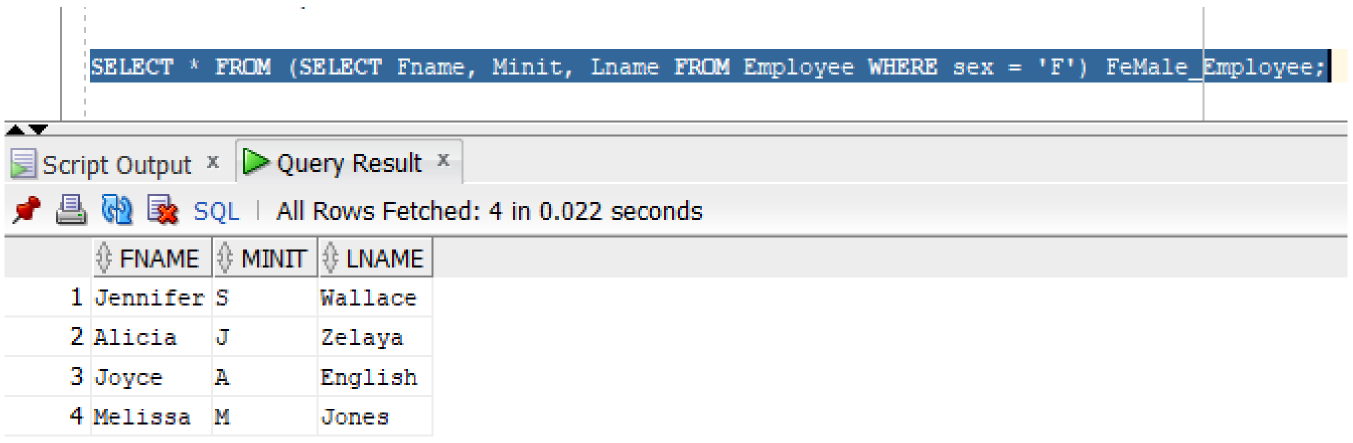
```
SELECT e.Fname, e.Minit, e.Lname, e.sex FROM Employee e
WHERE sex = 'F';
```

Script Output × | Query Result ×

SQL | All Rows Fetched: 4 in 0.028 seconds

	FNAME	MINIT	LNAME	SEX
1	Jennifer	S	Wallace	F
2	Alicia	J	Zelaya	F
3	Joyce	A	English	F
4	Melissa	M	Jones	F

b. Sub-query



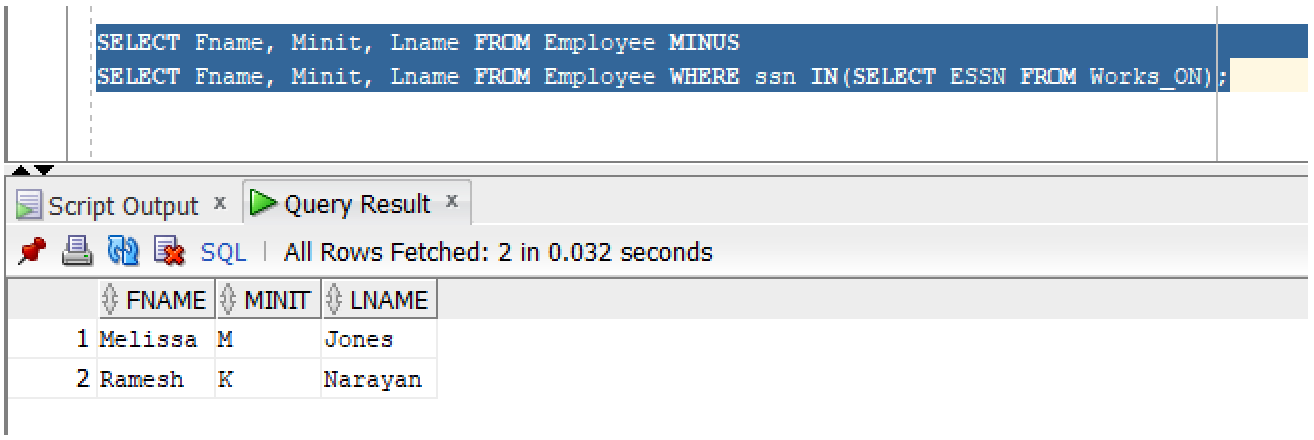
```
SELECT * FROM (SELECT Fname, Minit, Lname FROM Employee WHERE sex = 'F') FeMale_Employee;
```

Script Output × | Query Result ×

SQL | All Rows Fetched: 4 in 0.022 seconds

	FNAME	MINIT	LNAME
1	Jennifer	S	Wallace
2	Alicia	J	Zelaya
3	Joyce	A	English
4	Melissa	M	Jones

8. Find all employees who are not assigned to any project using SET operation in SQL



```
SELECT Fname, Minit, Lname FROM Employee MINUS
SELECT Fname, Minit, Lname FROM Employee WHERE ssn IN(SELECT ESSN FROM Works_ON);
```

Script Output × | Query Result ×

SQL | All Rows Fetched: 2 in 0.032 seconds

	FNAME	MINIT	LNAME
1	Melissa	M	Jones
2	Ramesh	K	Narayan

Part 2

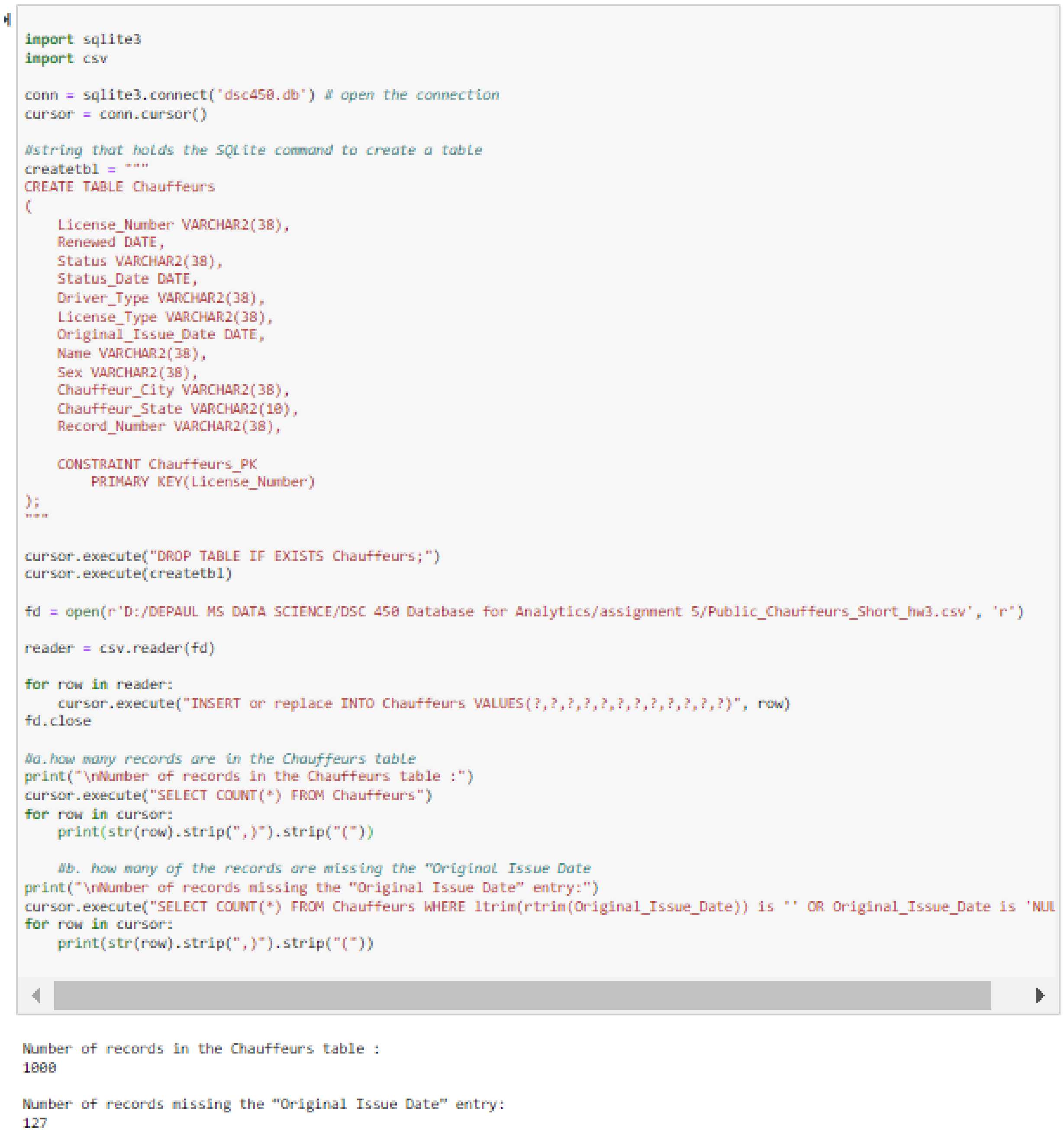
Create the table and use python to automate loading of the following file into SQLite:
http://dbgroup.cdm.depaul.edu/DSC450/Public_Chauffeurs_Short_hw3.csv

Find (using SQL)
a) how many records are in the Chauffeurs table and
b) how many of the records are missing the "Original Issue Date" entry.

It contains comma-separated data, with two changes: NULL may now be represented by NULL string **or** an empty string (e.g., either ,NULL, or ,,) and some of the names have the following form "Last, First" instead of "First Last", which is problematic because when you split the string on a comma, you end up with too many values to insert.

You can use csvreader to automatically load the data for you:

```
import csv
fd = open('Public_Chauffeurs_Short_hw3.csv', 'r')
reader = csv.reader(fd)
for row in reader:
    print(row)
fd.close()
```



```
import sqlite3
import csv

conn = sqlite3.connect('dsc450.db') # open the connection
cursor = conn.cursor()

#string that holds the SQLite command to create a table
createtbl = """
CREATE TABLE Chauffeurs
(
    License_Number VARCHAR2(30),
    Renewed DATE,
    Status VARCHAR2(30),
    Status_Date DATE,
    Driver_Type VARCHAR2(30),
    License_Type VARCHAR2(30),
    Original_Issue_Date DATE,
    Name VARCHAR2(30),
    Sex VARCHAR2(30),
    Chauffeur_City VARCHAR2(30),
    Chauffeur_State VARCHAR2(10),
    Record_Number VARCHAR2(30),

    CONSTRAINT Chauffeurs_PK
        PRIMARY KEY(License_Number)
);
"""

cursor.execute("DROP TABLE IF EXISTS Chauffeurs;")
cursor.execute(createtbl)

fd = open(r'D:/DEPAUL MS DATA SCIENCE/DSC 450 Database for Analytics/assignment 5/Public_Chauffeurs_Short_hw3.csv', 'r')

reader = csv.reader(fd)

for row in reader:
    cursor.execute("INSERT or replace INTO Chauffeurs VALUES(?,?,?,?,?,?,?,?,?,?,?,?)", row)
fd.close

#a.how many records are in the Chauffeurs table
print("\nNumber of records in the Chauffeurs table :")
cursor.execute("SELECT COUNT(*) FROM Chauffeurs")
for row in cursor:
    print(str(row).strip(",)").strip("("))

    #b. how many of the records are missing the "Original Issue Date
print("\nNumber of records missing the "Original Issue Date" entry:")
cursor.execute("SELECT COUNT(*) FROM Chauffeurs WHERE ltrim(rtrim(Original_Issue_Date)) is '' OR Original_Issue_Date is 'NUL
for row in cursor:
    print(str(row).strip(",)").strip("("))
```

```
Number of records in the Chauffeurs table :
1000

Number of records missing the "Original Issue Date" entry:
127
```

Part 3

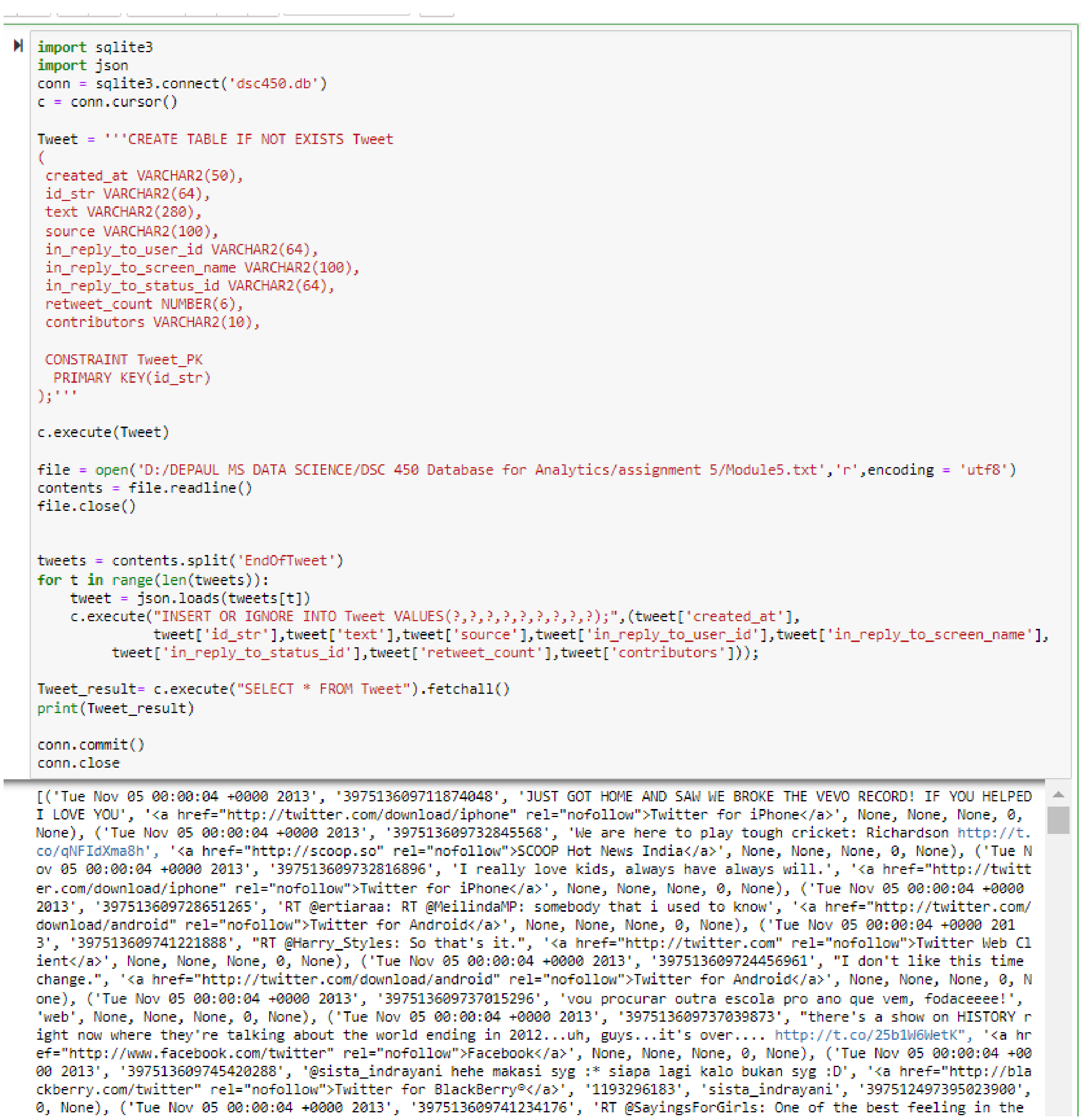
We are going to work with a small extract of tweets (about 200 of them), available here:
http://dbgroup.cdm.depaul.edu/DSC450/Module5.txt

NOTE 1: I do not recommend trying to copy-paste this text, because there is absolutely no knowing what might come out from paste on your system. You should be able to use "Save as…" function in your browser.

NOTE 2: The input data is separated by a string "EndOfTweet" which serves as a delimiter. The text itself consists of a single line, so using readline() or even readlines() will still only give you one row which needs to be split by the custom delimiter (i.e. .split('EndOfTweet')).

a. Create a SQL table to contain the following attributes of a tweet: "created_at", "id_str", "text", "source", "in_reply_to_user_id", "in_reply_to_screen_name", "in_reply_to_status_id", "retweet_count", "contributors". Please assign reasonable data types to each attribute and use SQLite for this assignment.

b. Write python code to read through the Module5.txt file and populate your table from part a. Make sure your python code reads through the file and loads the data properly (including NULLs).



```python
import sqlite3
import json
conn = sqlite3.connect('dsc450.db')
c = conn.cursor()

Tweet = '''CREATE TABLE IF NOT EXISTS Tweet
(
 created_at VARCHAR2(50),
 id_str VARCHAR2(64),
 text VARCHAR2(280),
 source VARCHAR2(100),
 in_reply_to_user_id VARCHAR2(64),
 in_reply_to_screen_name VARCHAR2(100),
 in_reply_to_status_id VARCHAR2(64),
 retweet_count NUMBER(6),
 contributors VARCHAR2(10),

 CONSTRAINT Tweet_PK
  PRIMARY KEY(id_str)
);'''

c.execute(Tweet)

file = open('D:/DEPAUL MS DATA SCIENCE/DSC 450 Database for Analytics/assignment 5/Module5.txt','r',encoding = 'utf8')
contents = file.readline()
file.close()


tweets = contents.split('EndOfTweet')
for t in range(len(tweets)):
    tweet = json.loads(tweets[t])
    c.execute("INSERT OR IGNORE INTO Tweet VALUES(?,?,?,?,?,?,?,?,?);",(tweet['created_at'],
              tweet['id_str'],tweet['text'],tweet['source'],tweet['in_reply_to_user_id'],tweet['in_reply_to_screen_name'],
         tweet['in_reply_to_status_id'],tweet['retweet_count'],tweet['contributors']));

Tweet_result= c.execute("SELECT * FROM Tweet").fetchall()
print(Tweet_result)

conn.commit()
conn.close
```

```
[('Tue Nov 05 00:00:04 +0000 2013', '397513609711874048', 'JUST GOT HOME AND SAW WE BROKE THE VEVO RECORD! IF YOU HELPED
I LOVE YOU', '<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>', None, None, None, 0,
None), ('Tue Nov 05 00:00:04 +0000 2013', '397513609732845568', 'We are here to play tough cricket: Richardson http://t.
co/qNFIdXma8h', '<a href="http://scoop.so" rel="nofollow">SCOOP Hot News India</a>', None, None, None, 0, None), ('Tue N
ov 05 00:00:04 +0000 2013', '397513609732816896', 'I really love kids, always have always will.', '<a href="http://twitt
er.com/download/iphone" rel="nofollow">Twitter for iPhone</a>', None, None, None, 0, None), ('Tue Nov 05 00:00:04 +0000
2013', '397513609728651265', 'RT @ertiaraa: RT @MeilindaMP: somebody that i used to know', '<a href="http://twitter.com/
download/android" rel="nofollow">Twitter for Android</a>', None, None, None, 0, None), ('Tue Nov 05 00:00:04 +0000 201
3', '397513609741221888', "RT @Harry_Styles: So that's it.", '<a href="http://twitter.com" rel="nofollow">Twitter Web Cl
ient</a>', None, None, None, 0, None), ('Tue Nov 05 00:00:04 +0000 2013', '397513609724456961', "I don't like this time
change.", '<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>', None, None, None, 0, N
one), ('Tue Nov 05 00:00:04 +0000 2013', '397513609737015296', 'vou procurar outra escola pro ano que vem, fodaceeee!',
'web', None, None, None, 0, None), ('Tue Nov 05 00:00:04 +0000 2013', '397513609737039873', "there's a show on HISTORY r
ight now where they're talking about the world ending in 2012...uh, guys...it's over.... http://t.co/25b1W6WetK", '<a hr
ef="http://www.facebook.com/twitter" rel="nofollow">Facebook</a>', None, None, None, 0, None), ('Tue Nov 05 00:00:04 +00
00 2013', '397513609745420288', '@sista_indrayani hehe makasi syg :* siapa lagi kalo bukan syg :D', '<a href="http://bla
ckberry.com/twitter" rel="nofollow">Twitter for BlackBerry®</a>', '1193296183', 'sista_indrayani', '397512497395023900',
0, None), ('Tue Nov 05 00:00:04 +0000 2013', '397513609741234176', 'RT @SayingsForGirls: One of the best feeling in the
```