ORIE 5270 - Homework 7
Due Friday March 13th at 11:59pm

You are not allowed to collaborate with your classmates. For this assignment, you have to implement a python script, write SQL queries in txt files, and write a solution file explaining how to run everything (as usual). Submit this assignment by uploading all the requested files to Canvas. Please upload each file, no need to compress them in a .zip.

Problem 1 (Exploratory data analysis)

For this problem, we will use the Chinook database, which can be downloaded from

https://cdn.sqlitetutorial.net/wp-content/uploads/2018/03/chinook.zip

For this question, you will be asked to query this database using SQLite. You will be asked to submit .txt files with your queries, e.g. if you are asked to output all the rows and columns of table X, your solution for the problem will be a .txt file whose content should be

```
SELECT * FROM X;
```

Submit 5 .txt files one per each item in the following list, their names should follow be `p1-<i>.txt` where `<i>` is a number between 1 and 5 indicating the number of the question. Additionally, you need to submit 5 .txt files called `output1-<i>.txt` with the output of each one of your queries.

1. Create a table called `producers`, which will list record producers. The table should have two columns: `producerId` (integer, this should be the key of the table) and `name` (text). Additionally, you need to insert the following five producers into the table: Joe Meek, George Martin, Quincy Jones, Nile Rodgers, and Phil Spector. Finally, query the table to see all the columns and rows. The file `p1-1.txt` should include all the queries you executed for this question. The output `output1-1.txt` only needs to include the output of the last query.

2. Delete the entry for Nile Rodgers in your table (your query has to look for it by name) and then display all the columns and rows of `producers`.

3. Find all the cities that have at least two customers.

4. Get the maximum, minimum, and average length of songs (in milliseconds) for each composer, output the results sorted by average length (in non-ascending order).

5. Find all the tracks that do not have a composer and then do another query to find the average size (in bytes) of these songs.

Problem 2 (Python, SQL, and friends)

For this problem, we will use the same datasets we used in Problem 2 of Homework 2. Write a python script (called `p2.py`) that executes the following tasks:

1. It **directly** scribes the information from

   https://snap.stanford.edu/data/facebook_combined.txt.gz

https://people.cam.cornell.edu/md825/names.txt

That is, your code has to download the information directly from the links (you cannot assume you have the files locally). With these data, it creates an SQLite database called `p2.db` with two tables: `people` and `friends`. The first one will have two columns `personId` and `name` and contains everyone from `names.txt`. The first few columns of this table would be

PersonId	Name
0	Randy Lavergne
1	Jamie Jones
2	Jana Griggs

The `friends` table should have too columns `personId1` and `personId2` and contain all the entries in `facebook_combined.txt`. The first few rows of this table should be

PersonId1	PersonId2
0	1
0	2
0	3

2. It calls `sqlite3` to output a list of people (names) and the numbers of friends they have, your result should be sorted in non-ascending order. Print the results in a readable form (a table with column and row name is fine). Note that to achieve this you will have to do a JOIN operation between two tables, explain in your solution file which kind JOIN (INNER, LEFT, RIGHT) you used and why.