Assignment 4: SQL & Aggregation

This set of exercises will give you more practice using all the elements of SQL which have been covered in class, including: selections, projections, joins, SQL functions, aggregation, grouping, and nested sub-queries.

Remember to work on your queries by identifying smaller sub-queries which are testable, and working your way out. You may find it useful to sketch out a simple relational schema to understand the relationships between tables, and which attributes are being used as foreign keys.

Attribution Statement

The `movies` database in this assignment consists of data scraped from the TMDB ("The Movie Database" [https://www.themoviedb.org/]) in 2017 and hosted on Kaggle at the following location: https://www.kaggle.com/rounakbanik/the-movies-dataset.

Part 0: Loading the database

Included with this assignment is a MySQL database backup file "`data604_a4_movies.sql`". This will create a new database called `movies` which you will use for the exercises contained in this document.

Use mysqlsh to run the commands in this file. First, connect to your chosen database server.

`<username>@<hostname or IP>:33060`

> For example:
>
> `root@localhost:33060` (if MySQL is installed on your own machine)
>
> `lewu@162.246.156.87:33060` (if you are the user lewu and are accessing the remote instance available for DATA 604). **REMEMBER that you must edit your SQL file** so that you are using an appropriate database name containing your username as part of the name, on each of lines 22, 24, and 26 if you are using this remote instance).

Then copy the path to your script and run it by entering in your mysqlsh window in SQL mode.

```
\sql
\source <path to your script>
```

Loading this data may take some time (3-5 minutes) - **be patient**!

Part 1: Warmup

1. What are the titles of every movie that has lost money, but had a rating higher than 8/10? (i.e. Have a lower revenue than their budget) [1 mark]
2. What are the names of all the production companies have been involved in producing movies that were produced or filmed in `Liechtenstein`? [2 mark]
3. What are the names of all of the cast in the movie `Cast Away`? [2 mark]

Part 2: Aggregation

4. How many movies have a budget over $100M? [1 mark]
5. How many unique countries have had movies in the `movies` database produced in them? [1 mark]
6. What are the 5 largest movie collections (in order from largest to smallest), and how many movies are in each? [2 marks]
7. What is the average revenue of movies produced in each country? [2 marks]
8. What production company has produced the most movies? [2 marks]
9. How many movies of each genre in the database has `Tom Hanks` acted in, and what is the average revenue in each of these genres? [2 marks]

Part 3: More Advanced Queries

10. What are the 3 highest grossing movie collections (in order from greatest to smallest) and what is the total revenue of each of these collections? [2 marks]
11. What is the maximum and average budget of movies in each genre that are produced at least partially in `Canada`? [2 marks]

Part 4: Six Degrees of Kevin Bacon

We define the degree of a connection between 2 people as the smallest chain of mutual connections between them. Eg. If A knows B and B knows C and C knows D, but A does not know C or D and B does not know D, then the degree between A and D is 3. We will say 2 actors know each other directly if they have worked on the same movie together.

12. How many actors are 1st-degree connections with `Kevin Bacon`? [2 mark]
13. How many actors are 1st or 2nd-degree connections with `Kevin Bacon`? [2 marks]
14. How many actors are 6th-degree (or less) connections with `Kevin Bacon`? [2 marks]
15. How many actors are there total in the database? [1 mark]

Submission Guidelines

Submit only one file. You may choose one of the following formats:

Submit a single *.sql file containing the SQL queries (not the result sets) you have created in response to each of the above tasks. Order them by the task number they respond to. Use SQL comments on the lines immediately preceding each query to declare the task which the query is for, and to state any assumptions you need to make.

OR

Submit a Jupyter notebook which runs each query in Python using `mysql-connector`. Each query should be its own cell.