DA5020 – Assignment 6

This assignment provides you with an opportunity to practice SQL queries for retrieving data from relational databases. This is used in practice for two primary reasons. One, to collect data from database sources — relational being the most common database architecture in use today. Second, eventually we will store cleaned data in an analytics store and relational databases are a common choice.

The assignment uses the SQLite database system, which is not truly a database server as it does not manage data on disk and does not provide multi-user access to the data. However, it is a fully functioning database that allows us to practice SQL queries.

The data camp "Introduction to SQL" and "Intermediate SQL" may be helpful to complete before attempting this assignment and will count as bonus points if both are completed.

Download the SQLite database called **imdb.db**, from canvas, and save it in a local folder on your computer, then connect to that database in R. Once connected, formulate SQL queries for each of these questions. Show each result set to demonstrate that the correct data was retrieved.

The database contains one table: **movie_info**. The **movie_info** table stores the ID (primary key), Series_Title, Release_Year, Runtime, Genre, IMDB_Rating, Director_ID and Gross.

Useful Resources

- SQLite CSV import examples
- Installation Tutorial for SQLite on Windows 10
- Installation Tutorial for SQLite on MacOS
- SQLite Download & SQLite Studio GUI
- Wickham, H. RSQLite Quick Tutorial

Bonus — (+5 bonus points, *All instructions must be followed and both parts must be completed to earn bonus points)*

1. Complete the data camps "Introduction to SQL" and a separate data camp"Intermediate SQL". A link to Data Camp will be shared within the course, so that you can access the resource for free using your Northeastern email. Please submit the statement of accomplishment from both data camps to earn credit (must also have your name on it). (No partial credit)
2. What is the average runtime for the **thriller** movie genre.

Question 1 — (10 points)

This question should be done completely in the SQLite Console, not in R. Start by loading the **imdb.db** file using the console and download the **directors.csv** file (from Canvas). Perform the following tasks:

1. (5pts) Create a table named **director_info** using SQLite; the columns are: Director_ID, and Director_Name. The **Director_ID** should be the primary key.
2. (5pts) Import the entire data from the CSV file into the **director_info** table using the SQLite **.import** command (see helpful resources below). Verify that your data was imported correctly.

Copy the queries above into a comment chunk in your Rmd file.

Question 2 — (40 points)

This question should be done in RStudio. Connect to the database, using R, and write queries to answer the questions below (answer each question in a separate R chunk). Do not load the entire database or its tables in your R environment.

1. (5 pts) Count the number of rows in the **movie_info** and **director_info** tables.

2. (5 pts) How many movies were released between 2010 and 2020 (inclusive)? Visualize the results.

3. (5 pts) What is the minimum, average and maximum ratings for "Action" movies. Ensure that you query the genre using wild cards.

4. (5 pts) What are the 25 highest-grossing movies within the dataset? Display the title, genre and gross.

5. (10 pts) Which directors have the highest-grossing movies. Display the director name and the total gross. Ensure that you join the necessary tables. Visualize the results using a Bar chart.

6. (10 pts) Create a function called **verifyDirector()** that takes a director name as its argument, and queries the database to check if the director exists. Your function should display a message to notify the user if the director was found or not.

Submission Details

- Your submission must contain two files: the .Rmd file and a knitted PDF or HTML (from the .rmd). Name your .Rmd file, DA5020.A6.FirstName.LastName.Rmd and your PDF/HTML DA5020.A6.FirstName.LastName.{pdf,html}, where *FirstName.LastName* is your first and last name.

- The .Rmd file must be fully commented and properly "chunked" R code and detailed explanations. Make sure that it is easy to recognize which question you answer and that your code runs from beginning to end (because that is how we will test it). Code that doesn't execute, stops, throws errors will receive no points. If the TAs have to "debug" your code or spend any effort getting it to run, substantial points will be deducted.

- Not submitting a knitted PDF or HTML will result in reduction of 30 points.

- Not submitting the .Rmd file (or both) will result in a score of 0.