Homework 3

Instructions

Due on Friday, Oct. 9^{th}, 2015, at class time (2:30PM)

You must submit this assignment electronically, via Blackboard, as a PDF.

Questions

1. Ch. 4, problem 4.9
2. Ch. 7, Problem 7.14
3. Ch. 7, Problem 7.20
4. Ch. 7, Problem 7.21
5. Ch. 8, Problem 8.29

Database Loading

Linked off the calendar page on the syllabus is a file 'ml-1m.zip' which is a collection of movies and rating information. Your assignment will involve loading this data into MySQL and performing some queries on it. The data is described in the README inside the archive, so read over that first to give you clues as to how you should create your tables.

You will not be able to load the data in directly with the "LOAD DATA INFILE" command in MySQL, though it may work on some of the tables. In particular, some of the queries that you'll be making will require you to split the genres and years into a proper relationship, and for that, you'll need to do a bit of programming. Create a script in your favorite language to read in the data from the zip files, parse it, and output a series of INSERT commands and then run this script directly into your database.

You can then run:

```
mysql -u rhoyle -p rhoyle < commands.sql
```

to load your data. You should make sure that you have all of the records loaded correctly, as your answers will be off if you have errors on the load.

We'll talk next week about how to do this with ODBC directly from a script, which is a more efficient way of doing this.

A hint... Since the data is in UTF8, if you don't want to deal with getting the syntax right in your script, you can do a

```
LOAD DATA INFILE LOCAL '/tmp/movies.dat'
        INTO TABLE movies
        CHARACTER SET 'utf8' FIELDS TERMINATED BY '::';
```

to load the movie names, then have your script manipulate the data based on movie ID to generate the proper genre information.

I was able to create a PERL script that just ran a sequence of INSERT statements to get all the data loaded.

Update:

Loading the ratings seems to be what slowed things down. I got it to work by creating a script to create the database structure for all the tables and loading the movie and user data into them via INSERT statements, as I wrote above. To get the ratings, I loaded the data using the MySQL command

```
LOAD DATA INFILE LOCAL '/tmp/ratings.dat'
        INTO TABLE ratings
        FIELDS TERMINATED BY '::';
```

This took around a minute. Note where the file is located. MySQL has issues with reading files from directories that are not globally readable. The easiest solution was to put the data in /tmp and do it from there. To prevent collisions, I suggest you rename the files with your username, such as `rhoyle-ratings.dat`, though if you are loading an unmodified one, you can use the one I put there.

Answer the following questions:

1. How many movies are in each genre?
2. What genre has the highest average star rating?
3. What year produced the highest average star rating?
4. What age group rated the most movies?
5. By genre, which gender enjoyed it the most?
6. By genre, which age group enjoyed it the most?

Please submit the script that you used to generate your data with your homework.

If you are having issues with timeouts due to network inactivity, you may want to consider using the UNIX command `screen`. You can find a tutorial at

`https://www.mattcutts.com/blog/a-quick-tutorial-on-screen/`

Honor Code

Please affix the required "I have adhered to the Honor Code in this assignment." at the bottom of your submission.