CAB220 Portfolio Item 1 - Requirements

Semester 2, 2019

This portfolio item is designed to evaluate your capacity for applying data manipulation knowledge and skills on a real dataset.

This assignment requires you to answer 3 questions about some data contained in a file. Each question should be answered in two ways:

(1) creating bash scripts using **bash commands** (not SQL commands), and
(2) creating bash scripts using **SQL commands** (i.e. writing an SQL command within a bash script like
```
sqlite3 database.sqlite 'SELECT * FROM mytable;'
```

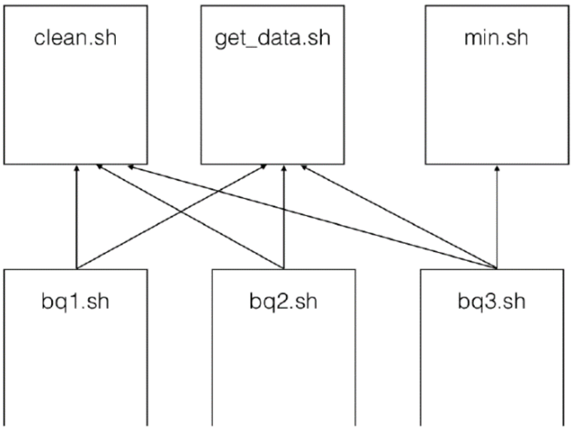
Templates for structuring the assignment:

The template is designed in “portfolio1.zip”, which contains a **.sh** file for each question. You must write your answer to each question into that file. Your answer must be an executable script that produces an answer to the question the script refers to. You may use auxiliary .sh scripts to break up your question scripts to make them easier to work with if you wish: in this way you can also use common scripts across questions, for example to clean the data and make sure you "link" your scripts in the answer script where relevant. We should be able to take your scripts for an answer by running them on the VM to obtain the correct output.

The template files where you have to store your solution scripts for the questions are:

- Bash commands scripts bq1.sh, bq2.sh, and bq3.sh (Q1, Q2, Q3 respectively).
- SQL commands based scripts sq1.sh, sq2.sh, and sq3.sh (Q1, Q2, Q3 respectively).

Example script files you may produce for the assignment (bash part only shown), and references between them (arrows) are portrait in the image below:



Dataset

This dataset contains details about biopics: a specific genre of movies. Unlike documentaries, which typically include raw footage and interviews, biopics are dramatizations, loosely based on the real-life events of actual people. Biopics offer an interpretation of lives deemed important (and profitable) by Hollywood, and they often try to make a statement about their subjects' historical or cultural significance. So which figures filmmakers spotlight matters, as does whom they ignore (or can't get the funding to feature).

Text and data have been taken from the article at https://fivethirtyeight.com/features/straight-outta-compton-is-the-rare-biopic-not-about-white-dudes/. Note, the original data has been manipulated to ease the tasks you are faced with. The data can only be used for CAB220 assessment items, please do not release it to other people.

Q1: How many movies has a director made?

Firstly, we would like you to spend some time exploring with the data, and cleaning it up (pay particular attention to ensure your output does not include "strings" that are not names, for example, – or "). Then, you should code up scripts that answer the question: How many movies has a director produced?

Download the dataset file **biopics.csv** from the blackboard.

Your **bq1.sh** script must do the following:

(1) Clean the data in an appropriate way to answer this question (and possibly the next ones)
(2) Produce a text file called **ba1.txt** that contains a list of directors along with how many movies they have produced. Note, each director should be named only once. The first column of your answer must be a list of directors sorted alphabetically and the second column must be the total number of movies they have made.

Your **sq1.sh** script must do the following:

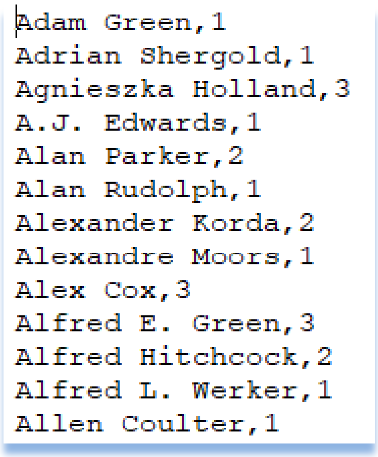
(1) Clean the data.
(2) Import the data into SQLite using (all in one line):

```
python3 csv2sqlite.py --table-name biopics --input biopics.csv
--output biopics.sqlite
```

Where `csv2sqlite.py` is a *Python script* that converts any delimited-type file (for instance csv). This file is included in the archive distributed with the assignment. Just like and `sqlite3` and bash scripts, we need to invoke it with the `python3` command. It takes several arguments, but the required ones have been filled out for you in the command above. If you would like an explanation of the arguments and the command, run `python3 csv2sqlite.py --help`.

(3) Produce a text file called **sa1.txt** that contains a list of directors along with how many movies they have produced. Note, each director should be named only once. The first column of your answer must be a list of directors sorted alphabetically and the second column must be must be the total number of movies they have made.

Example Output



```
Adam Green,1
Adrian Shergold,1
Agnieszka Holland,3
A.J. Edwards,1
Alan Parker,2
Alan Rudolph,1
Alexander Korda,2
Alexandre Moors,1
Alex Cox,3
Alfred E. Green,3
Alfred Hitchcock,2
Alfred L. Werker,1
Allen Coulter,1
```

...

Q2: Does gender influence how much a movie earns at the box office?

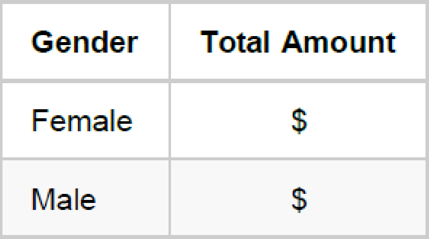
For the second question, we would like you to process the data to extract aggregate information.

Your bq2.sh script (**must** use bash commands) produces a HTML file called ba2.html that contains a HTML table that matches the example output (but with the actual correct amount of money). You do not need to style it, basic is better. Your table should summarise the total amount of money earned by each gender. Note that a movie may have more than one subject, and they may be of different genders. For example, a movie that earned $10M at the box office, may have a female and a male subject. In this case, count $10M as the contribution of the female subject, as well as the one of the male subject (i.e. no need to average or divide the contribution).

Your **sq2.sh** script (**must** use SQL queries) produces a HTML file called **sa2.html** that contains a HTML table that matches the example output (but with the actual correct amount of money). You do not need to style it, basic is better. Your table should summarise the total amount of money earned by each gender. Note that a movie may have more than one subject, and they may be of different genders. For example, a movie that earned $10M at the box office, may have a female and a male subject. In this case, count $10M as the contribution of the female subject, as well as the one of the male subject (i.e. no need to average or divide the contribution).

Please note, a movie for which box office income is not reported, it should not be calculated in the total amount.

Example Output



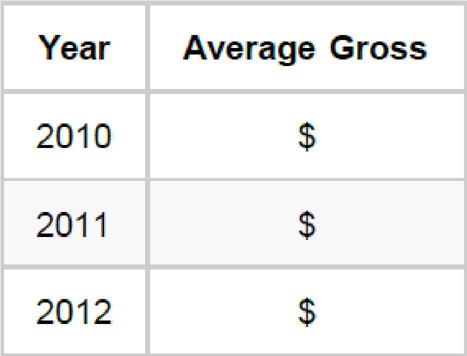
Gender	Total Amount
Female	$
Male	$

Q3: How much box office earnings do biopsies generate in each year?

For this question, your bq3.sh script (**must** use bash commands) produces a HTML file called ba3.html that contains a HTML table that matches the example output. You do not need to style it, basic is better. Your table should summarise the average of money earned each year by a biopic movie (each movie in the table is a biopic).

Your sq3.sh script (**must** use SQL queries) produces a HTML file called sa3.html that contains a HTML table that matches the example output. You do not need to style it, basic is better. Your table should summarise the total amount of money earned each year by a biopic movie (each movie in the table is a biopic).

Example Output



Year	Average Gross
2010	$
2011	$
2012	$

… …

Note, you need to include all the years in the final table.

Submission of the assessment

When working on this portfolio item, make sure you work in a new empty folder in your virtual machine. Once you have finished working on the assessment, you have to upload a zip file with all the content of your directory (and importantly the answer scripts) to blackboard (submission link in the BB folder for Portfolio Item 1, under the Assessment section in Blackboard). The zip file should be named as “yourStudentNumber_pi1.zip”, where yourStudentNumber is your student number without the leading letter “n”.

Remember, this assignment is to be performed **individually**.
Attempt to answer all questions.

The deadline for submitting your work for this portfolio item in CAB220 Blackboard is 11.59pm on Sunday **25 August 2019.**