

1 Bash & Data Management

Dataset: Hyphen-delineated with Format: [INDEX, Name, age, country, height, hair colour]
For this section, you have been provided with an incorrectly formatted dataset. The first half of this section will involve using Bash commands/scripts to correct the issues in the data. You will then write two Bash scripts to insert the cleaned data into both an SQL and MongoDB database, before performing a number of operations on the data there.

```
wget --no-check-certificate  
→ 'https://docs.google.com/uc?export=download&id=1FDqa2hJ9rPhLG1YEcvHlfUliLymiul4u'  
→ -O bashdm.csv
```

Dictionary file for Q4 (format: [INDEX&CODE&NAME]):

```
wget --no-check-certificate  
→ 'https://docs.google.com/uc?export=download&id=1h62yHjiRXMElbusQdCOM6W0_kgr-sQiv'  
→ -O dictionary.csv
```

Clean-up Tasks:

NB - for this section, there are multiple tools that may satisfy the required correction. You may use whichever tool you feel fits best - we will evaluate on the correctness of your output

1. There is an error in the name column - every name starts with an erroneous string "#)". Write a command/script to remove this string from the name column.
2. Currently the database is delineated by a "-" character. Write a command/script to convert this file into a comma delineated file. **NB - there are several triple-barelled names in the dataset, in the format "first-second-thirdname". These must not be split into separate columns. You may address these in a separate preparatory step if you wish**
3. There are two columns in this dataset with non-useful values. Write a script/command that identifies the columns that do not change, and removes them from the output.
4. Lastly, the country names in this dataset have been incorrectly added as country codes. Using the dictionary file provided (**Link above**) find and replace each country code with its correct country name.