ADS-500B
Module 2 Exercises
Filipp Krasovsky
11-6-2020

Foreword*: exercise files have been renamed ex1.csv, ex2.csv, and ex3.csv respectively.*
Due to formatting issues with the files downloaded from blackboard, native UNIX shell code, specifically the sort command, failed to properly operate on data without the data first being modified to coerce string values into numeric ones, or using AWK as an alternative.

1. **Dataset 1**
   - **a.** How many male and female record groups does the data have?

     here, we use a combination of the grep command to filter to records that contain the appropriate sex and the wc function to count the number of records.

     Command: **grep ‘Female’ Ex1.csv** | **wc**
     Output: 161 | 610 | 8632
     there are 161 records with female groups.

     Command: **grep ‘Male Ex1.csv** | **wc**
     Output: 163 | 636 | 8532
     there are 163 records with male groups.

   - **b.** How many white female groups are there? Copy the entire records of females to a new text file sorted by death count in descending order.
     Command: **grep ‘Female’ Ex1.csv** | **grep ‘White’ | wc**
     Output: 36 | 138 | 1912
     there are 36 records with white female groups.

     Command: **grep ‘Female’ Ex1.csv** | **grep ‘White’ > ex1_mod.csv**
     Command: **sort -t”,” -n -r -k5 ex1_mod.csv**

   - c. List all causes of death by frequency in descending order. What are the three most frequent causes of death for black males and five least frequent causes for Hispanic females?

     For black males:
     Command: **grep ‘Male’ Ex1.csv** | **grep ‘Black’ | sort -t”,” -k5 -n -r | head -3**
     Output:
     2010,Black,Male,Diseases of Heart,2015
     2010,Black,Male,Malignant Neoplasms (cancer),1540
     2010,Black,Male,Assault (Homicide),299

     the three deadliest causes of death for black males are Diseases of Heart, Malignant Neoplasms, and Assault.

For Hispanic females:

Command: **grep ‘Female’ Ex1.csv | grep ‘Hispanic’ | sort -t”,” -n -r -k5| tail -5**

Output:

```
2010,Hispanic,Female,Meningitis,4
2010,Hispanic,Female,Pneumonitis due to Solids and Liquids,2
2010,Hispanic,Female,Tuberculosis (TB),2
2010,Hispanic,Female,All Censored Causes,1
2010,Hispanic,Female,Anemias,1
```

the five least frequent causes of death for Hispanic females are anemias, censored causes, tuberculosis, pneumonitis due to solids and liquids, and meningitis.

2. **Dataset 2**
   a. Which country has the lowest percentage of urban population?

   To resolve this question, we put together an awk script that iterates over all records and modifies the variables *min* and *country* if the given row’s min value is smaller than the one declared at the top, which is set equal to NR=1.

   Command**:**

```
BEGIN{FS=”,”} NR==1 {min=$4;country=$1}
NR>1 && $4<min {min=$4; country=$1}
END{ print country,”:”,min}
```

   Output: Burundi: 13

   b. List all countries in which the urban population is more than 10 million but comprises less than half of the total population.

   Command:

```
BEGIN{FS=”,”} $3>10000000 && $4<50 {print $1,$3,$4}
```

   where $3 is the urbanized population and $4 is the % of total population.
   We print where $3 exceeds 10 million and where $4 is less than 50%.

   Output:

Country	Urban Population	% of Pop.
Bangladesh	164669750	36
Democratic Republic of the Congo	81339984	44
Egypt	97553148	43
Ethiopia	104957438	20
India	1339180125	34
Kenya	49699863	27
Myanmar	53370609	30
Pakistan	197015953	36
Philippines	104918094	47
Sudan	40533328	34
Thailand	69037516	49

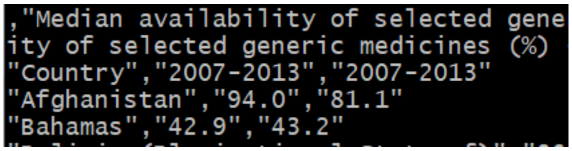
United Republic of Tanzania	57310020	33
Viet Nam	95540797	35
Yemen	28250420	36

3. **Dataset 3**

a. **Which country had the lowest percentage median availability generic medicines in private?**

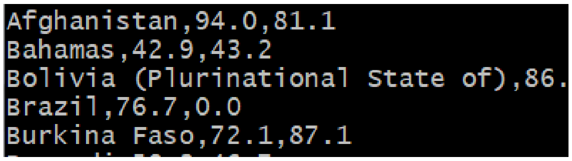
in order to make analysis on this dataset easier, cleanup is required. Since we already know that column 1 contains country data, column 2 contains private data, and column 3 contains public data, we can remove the first two header rows as well as the quotes around all values to make numerical evaluation possible in awk.

Initial State of Data when inspected through Cygwin terminal:



```
,"Median availability of selected gene
ity of selected generic medicines (%)
"Country","2007-2013","2007-2013"
"Afghanistan","94.0","81.1"
"Bahamas","42.9","43.2"
```

State of Data post-cleanup



```
Afghanistan,94.0,81.1
Bahamas,42.9,43.2
Bolivia (Plurinational State of),86.
Brazil,76.7,0.0
Burkina Faso,72.1,87.1
```

Commands:

sed ‘s/”//g’ ex3.csv > ex3_mod.csv

sed -e ‘1,2d’ ex3_mod.csv > ex3_mod2.csv

we will be working with **ex3_mod2.csv,** which has been attached as part of this assignment.

Commands: **sort -t”,” -k2 -n ex3_mod2.csv | cut -d’,’ -f1 | head -1**

Output: India

b. **Top 5 countries by private and public median access**

Private:

Command: **sort -t”,” -k2 -r -n ex3_mod2.csv | cut -d’,’ -f1 | head -5**

Output:

Russian Federation,
Syrian Arab Republic
Iran (Islamic Rep. of)
Afghanistan
Sudan

ADS-500B
Module 2 Exercises
Filipp Krasovsky
11-6-2020

Public:

Command: **sort -t”,” -k3 -n ex3_mod2.csv | cut -d’,’ -f1 | head -1**

Output:

```
Russian Federation
Cook Islands
Oman
Iran (Islamic Rep. of)
Syrian Arab Republic
```

c. **Top three countries where it is better to rely on private reliability of medicine than on public.**

Reasoning: the question asks us to consider where relying on the private sector is the optimal choice, not whether that country has a comparatively higher median access than other nations. Insofar as that’s the case, the calculation should depend on the highest difference between median private and public access.

Command: **awk -f ex_3.awk ex3_mod_2.csv | sort -t “,” -n -r -k2 | head -3**

where **ex_3.awk is:**

```
BEGIN{FS=”,” ; OFS=”,”}
{print $1, $2-$3}
```

Output: Brazil, Kyrgyzstan, Bolivia

As a note, the data downloaded does not contain any public data for Kyrgyzstan, so we have to default to the assumption that public data is not available. By discretion, we assume that the public median access is zero.

4. **Exercise 4: Python Script**

Code:

```
age = [25,18,9,13,34,15,22,17,12,37,15]
num_in_hs = 0

1    for age_val in age:
2            in_hs = age_val <= 18 and age_val >= 14
3            if not in_hs: not_in_hs+=1
4
5    ratio = not_in_hs/len(age)
7    print('% not in HS: ' + str(ratio) )
```

4.1 **:** the logic for determining if an individual is in high school is calculated on line 2 and is used to increment the number of people not in high school on line 3.

4.2 : the ratio of people not attending high school is calculated by dividing the variable we incremented by the total length of the array.