TP n°3 Unix [solution]

samuel.ortion@universite-paris-saclay.fr

Foreword

Wind turbines have an impact on biodiversity inducing an increase in bird and bat mortality. The main cause of this mortality is not the collision between the organism and the rotor blade but rather because of the high pressure induced by the rotation of the wind turbine.

The DREAL of the "Hauts de France" region deposits on data.gouv.fr a database that is updated regularly counting the dead corpse of bird and bats close to the wind turbine [1].

The objective of this practical exercise is to analyse some of these data using Unix commands.

1 Questions

1. How many corpse was discovered during the year 2020?
2. Count the number of corpse discovered per species. Which species was the most commonly found? Use the command `uniq -c`.
3. Use an `awk` script to answer the same question [2]
4. Two main taxonomic groups are affected by wind turbines: bats and birds. In the file "bird.txt" you have the list of bird species referenced in the database. Using this list, extract the list of other taxa (i.e., non bird taxa).
5. Based on the list of birds and bats, and using the commands `sort`, `join`, `cut` and `uniq` check that we found more bird corpses than bat corpses.
6. Looking at the content of the file `bird.txt`, maybe you discovered that the binomial names of the species were not always formatted with the genus name and species name as the standard *Genus species* format, with the first letter of the genus name in uppercase.

   Using an awk script, fix the formatting of all binomial names.

   Hint Don't forget the `toupper` and `substr` command.

2 Solution

1. Remember the `csv_fields` function? Let's use it here to identify the column number for the year.

```
csv_fields() {
    local filename="$1"
    local separator="${2:-\t}"
    awk -F "${separator}" 'NR == 1 { for (i=1; i < NF; i++) { print i, $i }}' "${filena
}
```

[1] Data source: `https://www.data.gouv.fr/datasets/cadavres-doiseaux-et-de-chiropteres-retrouves-au-pied-des-eoli` (visited on 7 novembre 2025).

[2] In the GNU Awk implementation, we have the function `asorti` which allows to sort a table. This might be useful.

```
<<csv_fields>>
csv_fields ./data/l_eolien_cadavres_mats_p_r32.csv ","
```

```
   1 code_icpe
   2 code_eolie
   3 nom_scient
   4 nom_vernac
   5 nombre
   6 oiseaux
   7 nom_du_par
   8 bureau_etu
   9 annee_deco
   10 date_deco
   11 sexe
   12 age
   13 distance_a
   14 id_map
   15 p_nominal
   16 departemen
   17 h_max
   18 h_mat
   19 h_nacelle
   20 diam_rotor
```

The number of records in the year 2020 can be obtained as follows with `awk`.

```
awk -F"," '$9 == 2020' ./data/l_eolien_cadavres_mats_p_r32.csv | wc -l
```

```
435
```

We can also use a Unix pipeline, for instance:

```
cut -d, -f 9 ./data/l_eolien_cadavres_mats_p_r32.csv | grep 2020 | wc -l
```

```
435
```

As rightfully pointed out by some students, the question asked the "number of corpse" discovered in 2020, not simply the number of data records. Indeed each row has a field "nombre" which counts how many corpse of this kind was discovered on a single visit in the wind turbine site.

To better answer this question, we can use `awk` and make the sum of every counts:

```
awk -F, '
BEGIN {
      # we initialize the total count
        total=0
}
# for each record of year 2020 (column 9 is 2020)
$9 == "2020" {
   # we add the count to the total number
```

```
    count = $5
    total += count
}
END {
    print total
}
' ./data/l_eolien_cadavres_mats_p_r32.csv
```

```
435
```

It happens by chance that we get the same results, because it seems that in 2020, every record line had a count of 1 corpse.

2. To count the number of records per species, we can use the following pipeline. Here I show only the 10 most reported species.

```
tail -n +2 ./data/l_eolien_cadavres_mats_p_r32.csv \
    | cut -d$',' -f 3 \
    | sort \
    | uniq -c \
    | sort -n --reverse \
    | head -10
```

```
    434 Pipistrellus pipistrellus
    127 Pipistrellus sp.
    124 Pipistrellus nathusii
    112 Apus apus
    111 Perdix perdix
    108 Columba palumbus
    105 Falco tinnunculus
     74 Buteo buteo
     70 Sturnus vulgaris
     67 Phasianus colchicus
```

Again, a way of improvement would be to take into account the number of corpse recorded for each line.

3.
```
awk -F',' '
NR > 1 {
  count[$3]++
}
function compare_count(i1, v1, i2, v2, l, r) {
  return v1 > v2
}
END {
  max = 0
  n = asorti(count, sorted_count, "compare_count")
  for (i = n; i >= 1; i--) {
    print sorted_count[i], count[sorted_count[i]]
  }
}
' ./data/l_eolien_cadavres_mats_p_r32.csv | head
```

```
Pipistrellus pipistrellus 434
Pipistrellus sp. 127
Pipistrellus nathusii 124
Apus apus 112
Perdix perdix 111
Columba palumbus 108
Falco tinnunculus 105
Buteo buteo 74
Sturnus vulgaris 70
Phasianus colchicus 67
```

4. `wc -l ./data/bird.txt`

```
93 ./data/bird.txt
```

```
mkdir -p tmp
```

```
 awk -F, '
# We first read the first file (bird.txt) and add all bird species to a table
NR == FNR {
     bird[$1] = 1; next;
 }
# Then, we read the second file, and report the third line (species latin name)
# when it is not present in the "bird" table
 {
     if (!($3 in bird)) {
         print $3
     }
 }
 ' ./data/bird.txt <(tail -n+2 ./data/l_eolien_cadavres_mats_p_r32.csv) | sort -u | tee
```

```
Chiroptera sp.
Myotis daubentonii
myotis emarginatus
Myotis nattereri
Nyctalus leisleri
Nyctalus noctula
Nyctula sp.
Vespertilio murinus
```

6. `awk -F' ' '`

```
{
   genus = $1;
   species = $2;
   first_letter = substr(genus, 1, 1);
   norm_first_letter = toupper(first_letter);
   genus_but_one = substr(genus, 2, length(genus) - 1);
   print norm_first_letter genus_but_one " " species }
   '
 ./data/bird.txt | tee ./tmp/oiseaux_norm_names.txt | head
```

Accipiter nisus
Acrocephalus scirpaceus
Aegithalos caudatus
Alauda arvensis
Alauda sp.
Alectoris rufa
Anthus pratensis
Anthus trivialis
Apus apus
Ardea cinerea