Part A - covid19 Data Analysis with R

Phuoc Vinh Dat Nguyen

2024-07-06

Part B: Shell Commands

In this section, we will perform some basic EDA to Olympics_tweets datset by

- view first 10 lines
- count the number of lines
- view the column names
- count number of columns
- check if there is any missing values in the dataset

```
iconv -f utf-8 -t ascii//translit Olympics_tweets.csv -o cleaned_olympics_tweets.csv

# save working dataset in a variable
dataset="cleaned_olympics_tweets.csv"

#first few lines of the dataset
echo "First 10 lines of the dataset:"
head -n 10 $dataset
echo ""

# Count the number of lines
echo "Number of lines in the dataset "
wc -l < $dataset
echo ""

# View the column names
echo "Column names:"
head -n 1 $dataset
echo ""

# count the number of columns
echo "Number of columns in the dataset:"
head -n 1 $dataset | tr ',' '\n' | wc -l
echo ""

# Check for missing values
echo "Checking for missing values in each column:"
awk -F, '
NR == 1 {
  for (i = 1; i <= NF; i++) {
    colnames[i] = $i;
  }
}
NR > 1 {
  for (i = 1; i <= NF; i++) {
    if ($i == "" || $i == "NA" || $i == " ") {
      count[i]++;
    }
  }
}
END {
  for (i = 1; i <= length(colnames); i++) {
    printf "column %s has %d missing values\n", colnames[i], count[i];
  }
}' $dataset
```

```
## First 10 lines of the dataset:
## id,text,user_screen_name,user_location,retweet_count,favorited,favorite_count,user_description,user_c
## 141896231718678e4,Mirabai Chanu's maiden Olympic silver helped India clinch joint-12th spot on the me
## 141896235133692e4,.@mirabai_chanu brings home first silver medal on Day 1. DSEU congratulates her on
## 1418962410673689900,Heartiest congratulations to Mirabai Chanu for starting  the medal tally for Indi
## 141896246682283e4,Hearty congratulations to ace Indian weightlifter @mirabai_chanu for winning a Silv
## 1418962529569670100,Congratulations to @mirabai_chanu for winning silver medal in weight lifting. Thi
## 141896254617485e4,@narendramodi @Tokyo2020 @mirabai_chanu Noble Group congratulates  India's Pride Mi
## 1418962573756550100,Mirabai Chanu's maiden Olympic silver helped India clinch the joint-12th spot on
## 141896264720307e4,Congratulations #india #olympics https://t.co/SApwOjumgR,YSR4Ever,United States,1,F
## 141896272332555e4,Congratulations to Mirabai Chanu getting first medal for INDIA on the first day of
##
## Number of lines in the dataset
## 114214
##
## Column names:
## id,text,user_screen_name,user_location,retweet_count,favorited,favorite_count,user_description,user_c
##
## Number of columns in the dataset:
## 13
##
## Checking for missing values in each column:
## column id has 0 missing values
## column text has 0 missing values
## column user_screen_name has 757 missing values
## column user_location has 31114 missing values
## column retweet_count has 1078 missing values
## column favorited has 445 missing values
## column favorite_count has 198 missing values
## column user_description has 12843 missing values
## column user_created_at has 1645 missing values
## column user_followers has 818 missing values
## column user_friends has 608 missing values
## column date has 286 missing values
## column language has 478 missing values
```

Task 1: Write commands to count and then remove lines with an id that is not a number of 19 digits long, i.e., id values that contain anything other than numbers OR are of a length more/less than 19. Store the filtered set in a file named ***filtered_tweets_1.csv.***

```
# save working dataset in a variable
dataset="cleaned_olympics_tweets.csv"

# Count any line with invalid id values
awk -F, 'NR>1 && !($1 ~ /^[0-9]{19}$/)' $dataset | wc -l

# Remove lines with invalid id values and save to a new file as filtered_tweets_1.csv
awk -F, 'NR==1 || ($1 ~ /^[0-9]{19}$/)' $dataset > filtered_tweets_1.csv


# taking a look at the result
head -n 10 filtered_tweets_1.csv
```

```
## 92823
## id,text,user_screen_name,user_location,retweet_count,favorited,favorite_count,user_description,user_
## 1418962410673689900,Heartiest congratulations to Mirabai Chanu for starting  the medal tally for Indi
## 1418962529569670100,Congratulations to @mirabai_chanu for winning silver medal in weight lifting. Thi
## 1418962573756550100,Mirabai Chanu's maiden Olympic silver helped India clinch the joint-12th spot on
## 1418963345806329900,Proud to take a Bow and Congratulate @mirabai_chanu on the maiden Olympic Medal i
## 1418964164048030200,SCIKEY congratulates MIRABAI CHANU for winning the first medal for India in Toky
## 1418964829944089900,Congratulations @mirabai_chanu #Olympics #Olympics2020 #OlympicGames #WINNER #Ol
## 1418964856213049900,Congratulations dY'? @mirabai_chanu! What an amazing start of India on first day
## 1418964969647969800,Fabulous start by @mirabai_chanu  in the Tokyo Olympics. Congratulations for you
## 1418965006578809900,Rachita Panda Mistry- A former Indian Olympian who was participated in 2000 Sydn
```

Task 2: Identify the date range of the tweets. Please note that the file is not guaranteed to be sorted and Nulls (NA and empty values) should not be considered. Hint: you can change the delimiter of the dataset to sort the date.

```
# Define the dataset
dataset="filtered_tweets_1.csv"

# Convert the delimiter to a tab and sort it out by date column and remove the invalid values
awk -F, 'NR > 1 && $9 ~ /^[0-9]{2}\/[0-9]{2}\/[0-9]{4}/ {print $9}' $dataset | sort | uniq > sorted_date

echo "Date range of the tweets:"
head -n 1 sorted_dates.txt
tail -n 1 sorted_dates.txt
```

```
## Date range of the tweets:
## 10/01/2007 15:39
## 31/12/2020 6:21
```

Task 3: Select a subset that includes the keyword 'Australia,' and then display the top 10 most frequent user_screen_names in that subset.

```
#set up some key variables
dataset="filtered_tweets_1.csv"
keyword="Australia"

# filter keyword 'Australia'from row 'user_location'
awk -F, -v keyword="$keyword" '$4 ~ keyword' $dataset > subset_australia.csv

# extract the user_screen_name columnfrom the subset
awk -F, 'NR > 1 {print $3}' subset_australia.csv > user_screen_names.txt
```

```
# count of each user_screen_name and sort out
sort user_screen_names.txt | uniq -c | sort -nr > user_screen_name_counts.txt

# print top 10 most frequent user_screen_names
echo "10 most frequent user_screen_names:"
head -n 10 user_screen_name_counts.txt
```

```
## 10 most frequent user_screen_names:
##       7 Avatar5991
##       4 LizannV
##       4 abcnews
##       3 westaustralian
##       3 theage
##       3 newscomauHQ
##       3 greysfan
##       3 AUSOlympicTeam
##       3 abcsport
##       2 tvtonightau
```

Task 4: Filter your data based on the following conditions:

- Keep only these columns: id, user_screen_name, user_created_at, user_followers, user_friends, and date (Note: keep column names as well)
- Keep the tweets with user_friends and user_followers each larger than 1000

Export the above-selected data to a new file named ***filtered_tweets_2.csv***.

In the file ***filtered_tweets_2.csv***, how many tweets have an NA value in the last column (i.e., the column 'date')? How many user accounts were created prior to 2020? Please note that the column 'user_created_at' specifies when a user account was created and one Twitter user might have produced multiple tweets and you are supposed to count the accounts.

```
# set dataset variable
dataset="filtered_tweets_1.csv"

# we have the following id: 1, user_screen_name: 3, user_created_at: 9, user_followers: 10, user_friend
awk -F, 'BEGIN {OFS=","}
    NR == 1 {print $1, $3, $9, $10, $11, $12}
    NR > 1 && $10 > 1000 && $11 > 1000 {print $1, $3, $9, $10, $11, $12}' $dataset > filtered_tweets_2.

# print new file
echo "First 10 lines of the filtered dataset for Task 4:"
head -n 10 filtered_tweets_2.csv
```

```
## First 10 lines of the filtered dataset for Task 4:
## id,user_screen_name,user_created_at,user_followers,user_friends,date
## 1418965464978310100,saumyadadoo,27/06/2009 10:01,2968,3661,24/07/2021 16:05
## 1418976894129689900,nileshtrivedi,1/05/2008 12:01,4021,4462,24/07/2021 16:50
## 1418980762146150100,aJPY? a?2aJPY? a?1aJPYEUR a?+a?^a?,0,A BRAVE SOLDIER WHO SERVE FOR NATION BY CRP
## 1418982990286310100,Ram__Rathod,a?oaJPY?a??a??a?? a?oa? 3/4 a??aJPY?a??a?oa??aJPY?a??a? 3/4  a?--aJP
## 1418983231257510100,vadabuoy, a?,a?|aJPY?a?ua??a?aaJPY?a??a? 3/4  a?!a?1aJPY?a??a? 3/4  a?ua?|a??aJP
## 1419331606913950200,stripeyspotty,1/01/2012 8:39,3109,1595,25/07/2021 16:19
## 1419331622290150100,SilentFancyFox,?aEUR?dYOE^dY?odY?,| Mostly Art Retweets | Fancy Fox dY|S | Video
## 1419331659262950100,StrangeKeith,18/11/2009 9:00,4720,1194,25/07/2021 16:20
## 1419331701004710100,naganosilver98, #LanguageLearning dY?-dY?.dY?adY?,dY?udY?1 #Photoshop dYZ? #Writ
```

```
# set up new dataset
dataset="filtered_tweets_2.csv"

# Count tweets with 'NA' values or NULL in the date
na_count=$(awk -F, 'NR > 1 && ($6 == "" || $6 == "NA" || $6 == " ") {count++} END {print count}' $datase

echo "Number of tweets with 'NA' or NULL in the 'date': $na_count"


# Count unique user accounts were created before 2020
unique_user_count=$(awk -F, 'NR > 1 && $3 != "" {split($3, a, "/"); if (a[3] < 2020) print $2}' $dataset
echo "Number of unique user accounts created before 2020: $unique_user_count"
```

```
## Number of tweets with 'NA' or NULL in the 'date': 17
## Number of unique user accounts created before 2020: 3758
```