Muhammad Shabir Khan

Module: R2: The Missing Semester
Section: Data Wrangling **Task:** 06

Task:

Download this file: https://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/forestfires.csv, The columns everyone can choose are up to them, we have 8 columns Fetch it using curl and extract out just two columns of numerical data. If you're fetching HTML data, **pup** might be helpful. For JSON data, try **jq**. Find the min and max of one column in a single command and the difference of the sum of each column in another.

Explanation:

1. I have extracting the data of two column from the link to perform the task and save this data to a csv file named as **extracted_data.csv**. The command to extract the data from the link is,

curl https://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/forestfires.csv | tail -n +2 | awk -F',' '{print $1, $2}' > extracted_data.csv

Here, **tail –n+2** is used to remove the first line of the both column. Since the first line has non numeric digit.

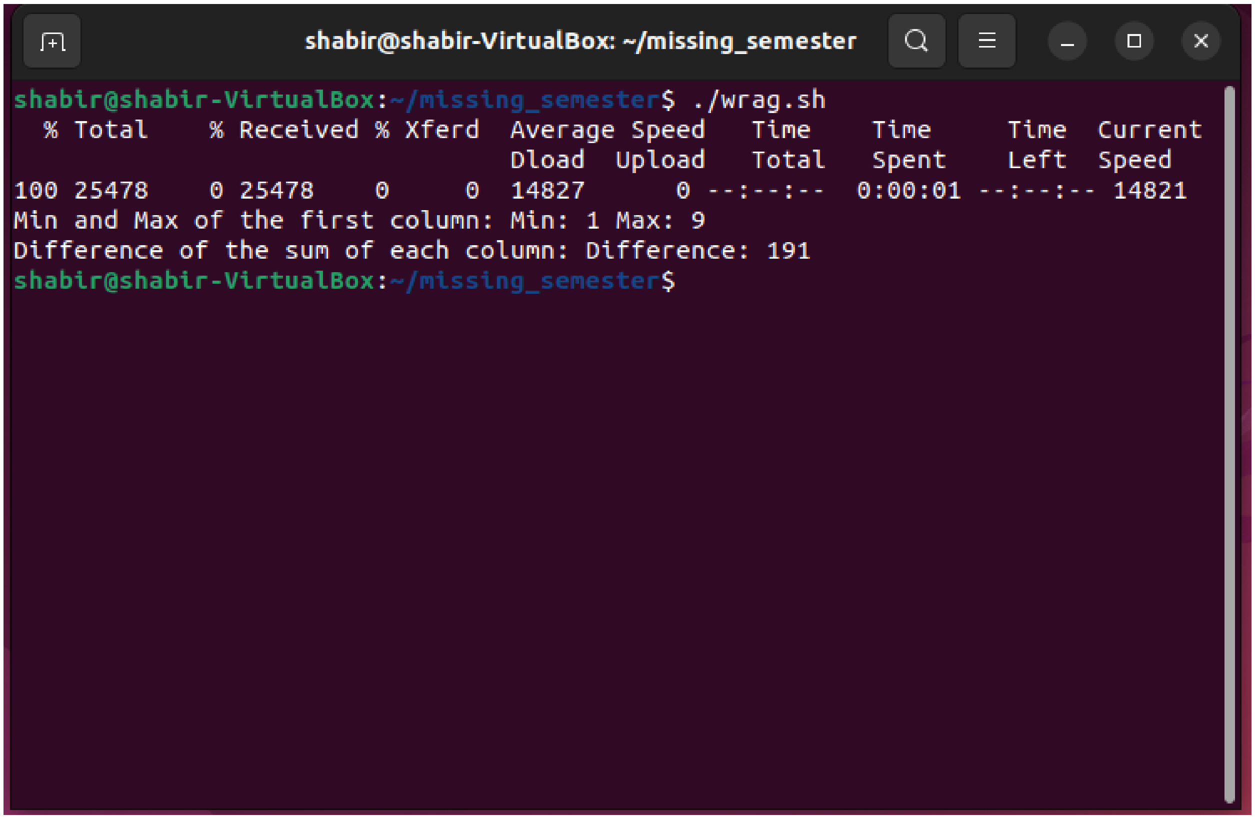
2. To find the Min and Max of the first and second columns I have used the **awk** command and wrote the script to find the min and max of each column. The command is shown below,

awk 'BEGIN {min=1000000; max=-1000000} {if ($1 < min) min=$1; if ($1 > max) max=$1} END {print "Min:", min, "Max:", max}' extracted_data.csv

3. To find the difference between both columns after addition of each column, I used the **awk** command and wrote a script. The command is shown below,

awk '{sum1+=$1; sum2+=$2} END {print "Difference:", sum1 - sum2}' extracted_data.csv

4. The Output of the script is shown below:



```
shabir@shabir-VirtualBox:~/missing_semester$ ./wrag.sh
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 25478    0 25478    0     0  14827      0 --:--:--  0:00:01 --:--:-- 14821
Min and Max of the first column: Min: 1 Max: 9
Difference of the sum of each column: Difference: 191
shabir@shabir-VirtualBox:~/missing_semester$
```

Appendix

The Script of the task is shown below:

```
#!/bin/bash

# Step 1: Download the CSV file and extract numerical columns, removing the first line
curl https://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/forestfires.csv | tail -n +2 | awk -F',' '{print $1, $2}' > extracted_data.csv

# Step 2: Find Min and Max of the first column
min_max_first_column=$(awk 'BEGIN {min=1000000; max=-1000000} {if ($1 < min) min=$1; if ($1 > max) max=$1} END {print "Min:", min, "Max:", max}' extracted_data.csv)
echo "Min and Max of the first column: $min_max_first_column"

# Step 3: Find the Difference of the Sum of Each Column
difference_sum_columns=$(awk '{sum1+=$1; sum2+=$2} END {print "Difference:", sum1 - sum2}' extracted_data.csv)
echo "Difference of the sum of each column: $difference_sum_columns"
```