Module 3 Assignment

DAT-375

Brandon Hobbs

March 12, 2023

# Overview

I am being asked to analyze the 21,973 data records scraped from a social media site to get a sense of usage.

Specifically, I am being asked to answer the following questions:

* Find the average number of messages per username
* Find average number of reshares per username
* Find the time frame with the highest number of original messages

Note: there were errors on data importation and ~3000 records were stripped. Most of the errors look to be encoding issues. However, as the project is using UTF-8 these encoding issues have not been fixed as they looked to be characters from an improper character set, e.g., ϋ or ό.

# Average Messages per Username

There are multiple methods to find the average messages per username. This could be the global average, e.g., COUNT OF MESSAGES / COUNT OF USERS, or this could be the average over time for a single user. As we are not tracking individual user profiles, yet, the global average will be used.

The script chosen is as follows:

*SELECT Count(message) / Count(distinct username)*

*from hobbs.mod\_3;*

Which produces a global average of 945.2308 messages per username.

# Average Reshares per Username

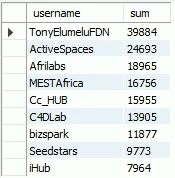
To find the average reshare per username a similar approach will be used for average messages.

The script chosen is as follows:

*SELECT Sum(reshares) / Count(distinct username)*

*from hobbs.mod\_3;*

Which produces a global average of 7385.5385 reshares per username. This of course makes it appear as though all users have an average reshare of greater than 7300. This is not true. Only 9 user account have this as a true condition – the other 17 have far fewer reshares.



Which is found using this query:

*Select username, Sum(reshares) as sum*

*from hobbs.mod\_3*

*Group by username*

*Having sum > (*

*Select Sum(reshares) / Count(distinct username)*

*from hobbs.mod\_3)*

*order by sum DESC;*

# Timeframe with the Greatest Number of Original Messages

Timeframe is open to interpretation. Sense every message is tagged with a time string when created we could do a simple query to see how many messages occur at that times string, below:

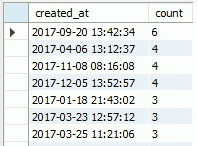
*Select created\_at, COUNT(message) as count*

*from hobbs.mod\_3*

*Group by Created\_at*

*order by count DESC;*

Which shows that 6 messages were created at 13:42:34 in September 20, 2017.



The month could also be extracted from the date field and the messages could be grouped per month, below:

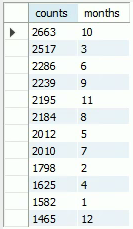
*Select COUNT(message) as counts, month(created\_at) as months*

*from hobbs.mod\_3*

*Group by months*

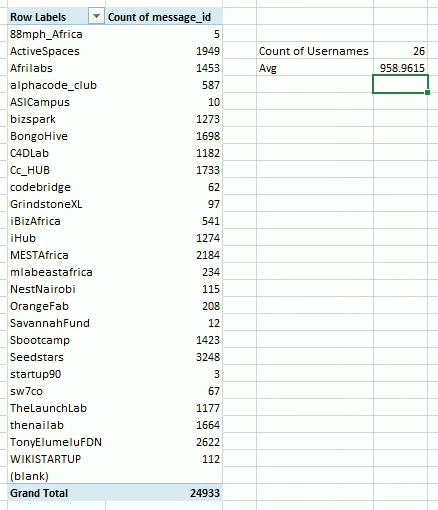
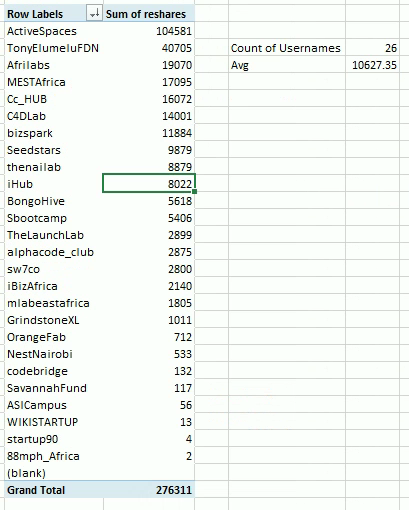
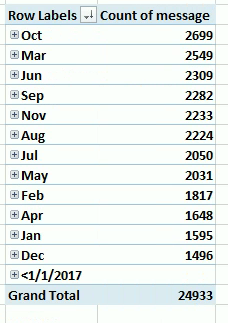
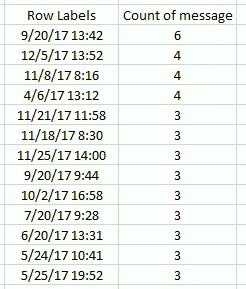
*order by counts DESC;*

Which shows the most messages were created during October (month 10).



# Validation

To validate the scripts Excel was used. A series of pivot tables was created as the dataset is small.

* Average messages per username  
    
  The average found within Excel is larger because of the larger number of messages. The ~3000 records that were rejected within SQL were not rejected in Excel, but the fact that the average is larger confirms that the script value is probably correct.  
  
* Average reshares per username  
    
  Again the Excel value are larger which is being attributed to the missing records.  
  
* Time period with the greatest frequency  
    
  Excel confirms that October has the greatest number of original messages:  
    
    
  And that 6 messages occurred on September 20, 2017:  
  

# Conclusions

For the above analysis each analysis and accompanying script is laid out. At this early phase where we have not defined user profiles and started tracking those profiles it is more interesting to use a global average to see how a generic user might behave. These generic profile averages were what were used above.   
  
Using this generic profile idea, the scripts used were simple and only focused on the count of distinct records. The validation techniques in Excel also used the distinct usernames to define the global averages.

Validation in Excel was not able to fully validate the scripts due to the encoding issues. Further attempts or a change in encoding should be evaluated.