Module 3 Assignment

DAT-375

Brandon Hobbs

March 19, 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

To perform this type of analysis then quantitative counts and summations are warranted. After all, the count of distinct usernames, the summation of retweets, etc. can be calculated.

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 keyword distinct should be used in the query so that we can see the averages per unique user. For example, if a single user sends 5 messages the expectation is the script would return an average of 5 (5 messages / 1 user) and not an average of 1 (5 messages / 5 occurrences of username).

Incorrect script:

Graphical user interface, text

Description automatically generated

The script chosen is as follows:

Graphical user interface, text, application, email

Description automatically generated

*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.

Again, the distinct keyword is warranted here as if a single user were to reshare 10 messages, 100 times each the expectation is that 1000 would be returned (10 \* 100 / 1 user) and not 100 (10 \* 100 / 10 records).

Graphical user interface, text, application, email

Description automatically generated

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.

Graphical user interface, text, application

Description automatically generated

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 those 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 on September 20, 2017.

Graphical user interface, text, application

Description automatically generated

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).

Graphical user interface, text, application

Description automatically generated

# 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.  
    
  The pivot table created finds the unique usernames (26) and then divides the count of message id to find the average messages per distinct username, 958.96 messages per unique user.  
  Graphical user interface, application, table

  Description automatically generated
* Average reshares per username  
    
  The value of reshares was not in the proper format so a new field is added that converts the text value of reshares to an integer so that the summation of reshares may be determined. The number of unique users is still the same as before, 26, and the summation is 276,311. This is determined through this pivot table, which results in an average reshare per unique user of 10,627.35.  
    
    
  Graphical user interface, table

  Description automatically generated  
    
  Note: this value is significantly smaller than the SQL scripts because SQL is only finding reshares totaling 192024:  
    
    
  Graphical user interface, text, email

  Description automatically generated
* Time period with the greatest frequency  
    
  To determine the time period with the greatest frequency the created at timestamp had the month extracted using the Excel function MONTH() into a new attribute called month. Then this was counted via a pivot table. The individual, unique, time strings were also counted and pivoted. Both analyses are shown below.  
    
  Excel confirms that October has the greatest number of original messages:  
    
  Graphical user interface, application, table, Excel

  Description automatically generated  
    
  And that 6 messages occurred on September 20, 2017:  
  Graphical user interface

  Description automatically generated with low confidence

# 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.