Social Media Data Analysis

**Choosing Analysis**

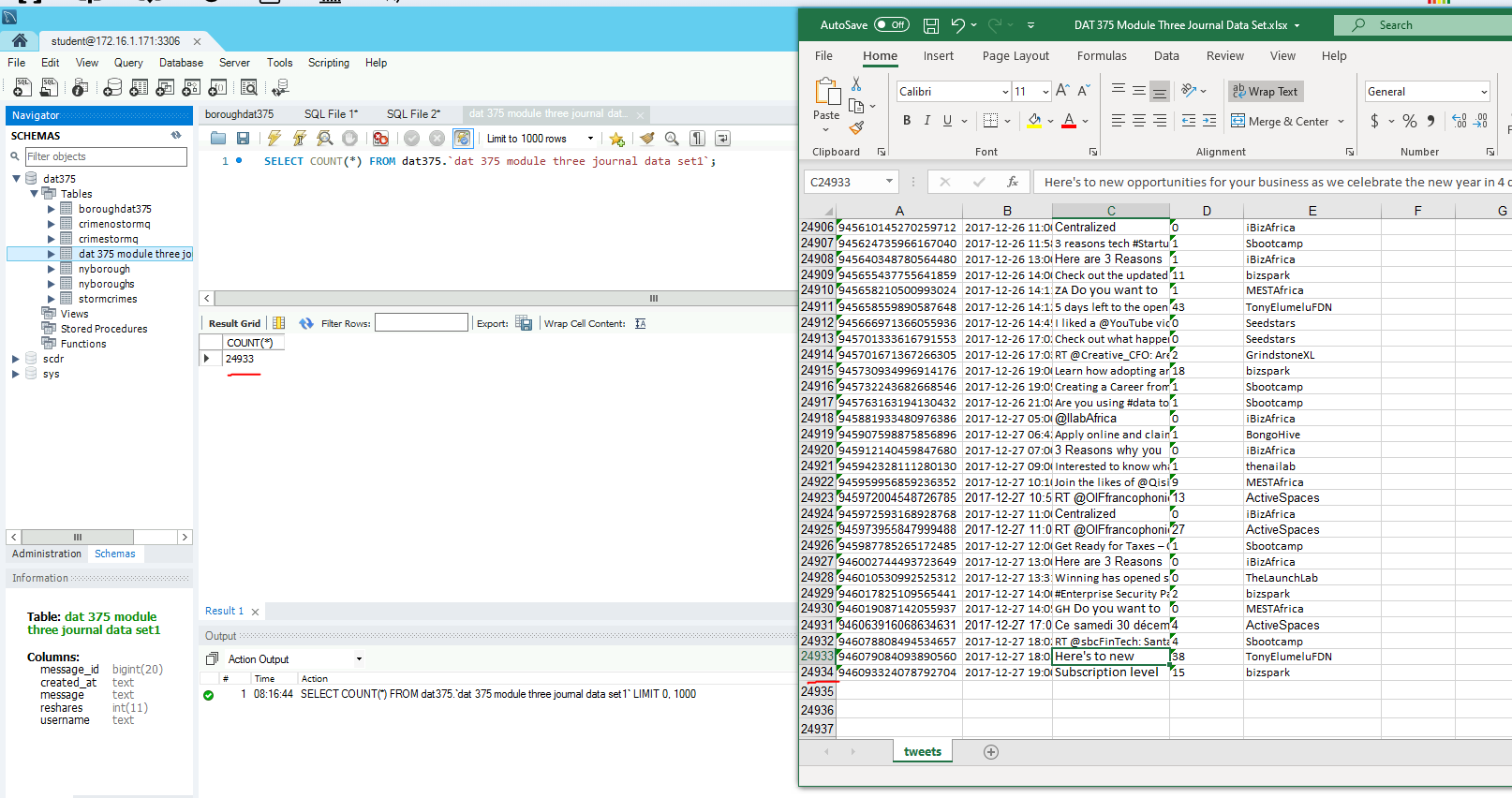
The problems stated are to find the average number of messages per username, the average number of reshares per username, and timeframe that has the highest number of original messages. The keyword here is average, also known as mean, and highest, also known as max, and the specific language of requirements used provides the technique to analyze which would be descriptive analysis. (Bhatia, 2019) In some essence the request and naming of analysis also provides reassurance that the appropriate method is used. Descriptive analysis simply describes the data set with numbers.

**Data Preparation**

The first objective to selecting scripts to query the dataset is to import into MySQL and verify data integrity. The first issue is the file was provided as an excel file, ASCII encoding, and with several carriage return and line feeds all over the message column. The formatting was generally easy to correct by saving as CSV, opening with Notepad ++ and converting to UTF-8, and doing a replace of all ‘LF’ to ‘’, and all ‘CR’ to ‘CRLF’. This would ensure MySQL would be able to ingest messages safely without breaking records. However, a new problem was identified and that is the use of quotation marks around several messages caused shifting of cells to fall under improper columns as MySQL import would assume “” as a cell value and the next characters would be part of the next column. For example the following record:

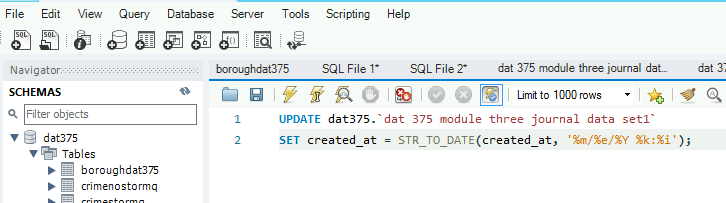
*815540465319690000,1/1/2017 12:49,”””If it were easy, somebody else would have done it already.”” Dare to be ambitious in 2017: Happy new year.”,8,Seedstars*

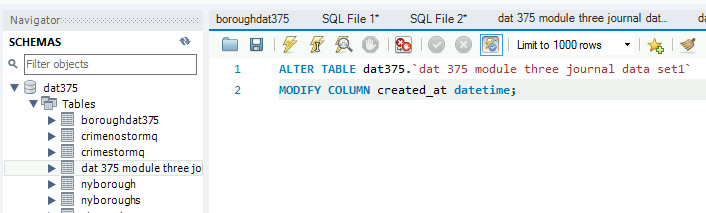
To correct this, a replacement of ‘“”’ (see that is two quotation marks) to ‘’ was done. Now all of the data can be safely imported into MySQL with every field being imported correctly. This was quality checked from MySQL results to the source excel file.



*Please note the excel file contains a header so it is total rows - 1.*

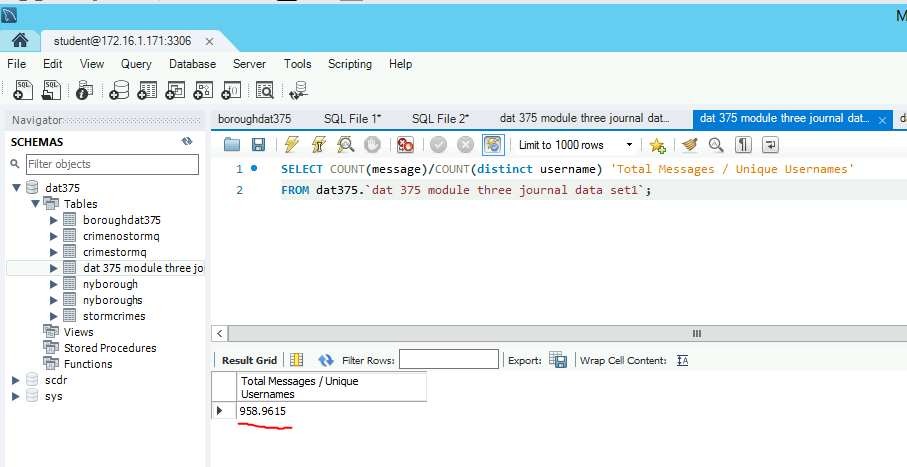
Finally, there was a reformat needed on the created\_at attribute into a datetime:



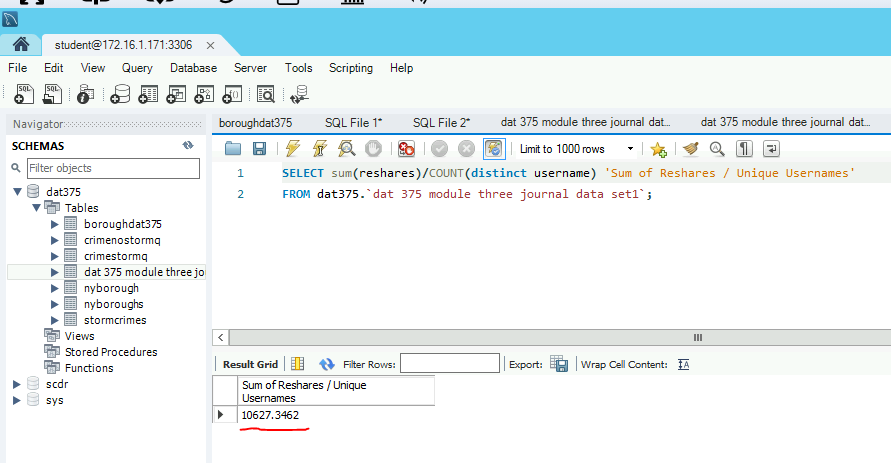


**Selecting Scripts**

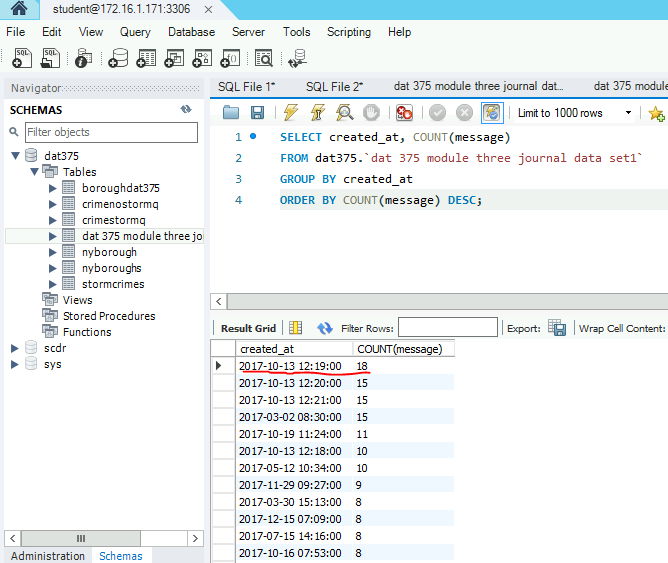
Beginning with the first question “Find the average number of messages by a single user name”, the following query and results should suffice:



The next question asks “Find the average number of reshares by a single username”, the following query and results should suffice:

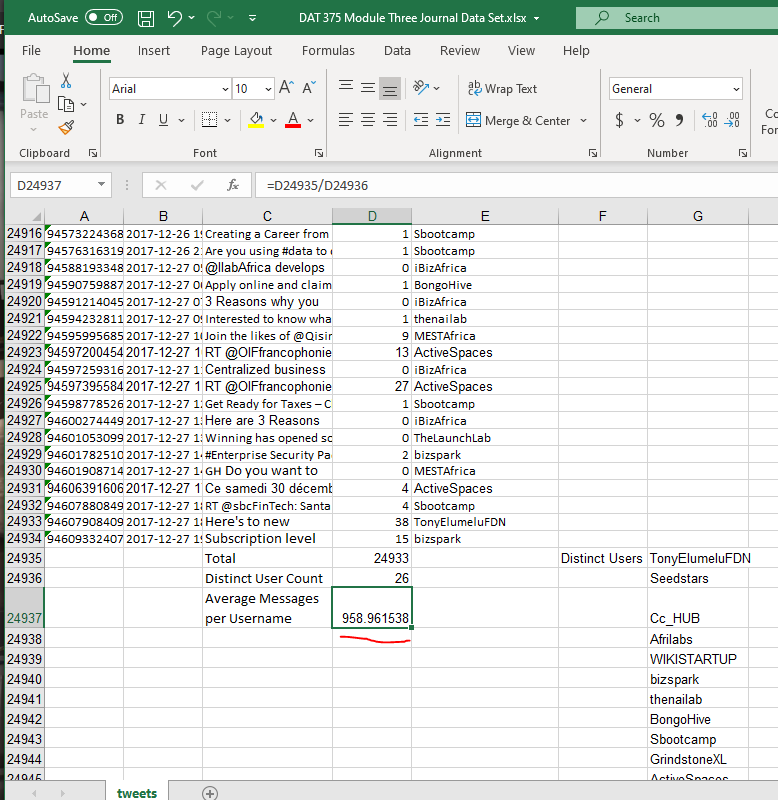


The final question asks “Find the timeframe that has the highest number of original messages”, the following query and results should suffice:

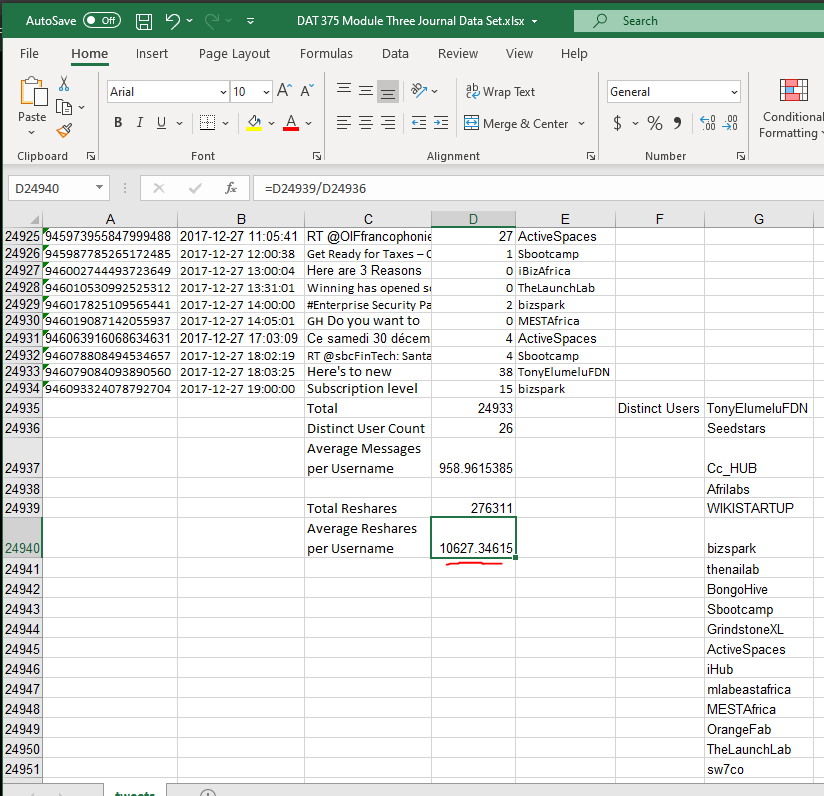


**Validating Scripts**

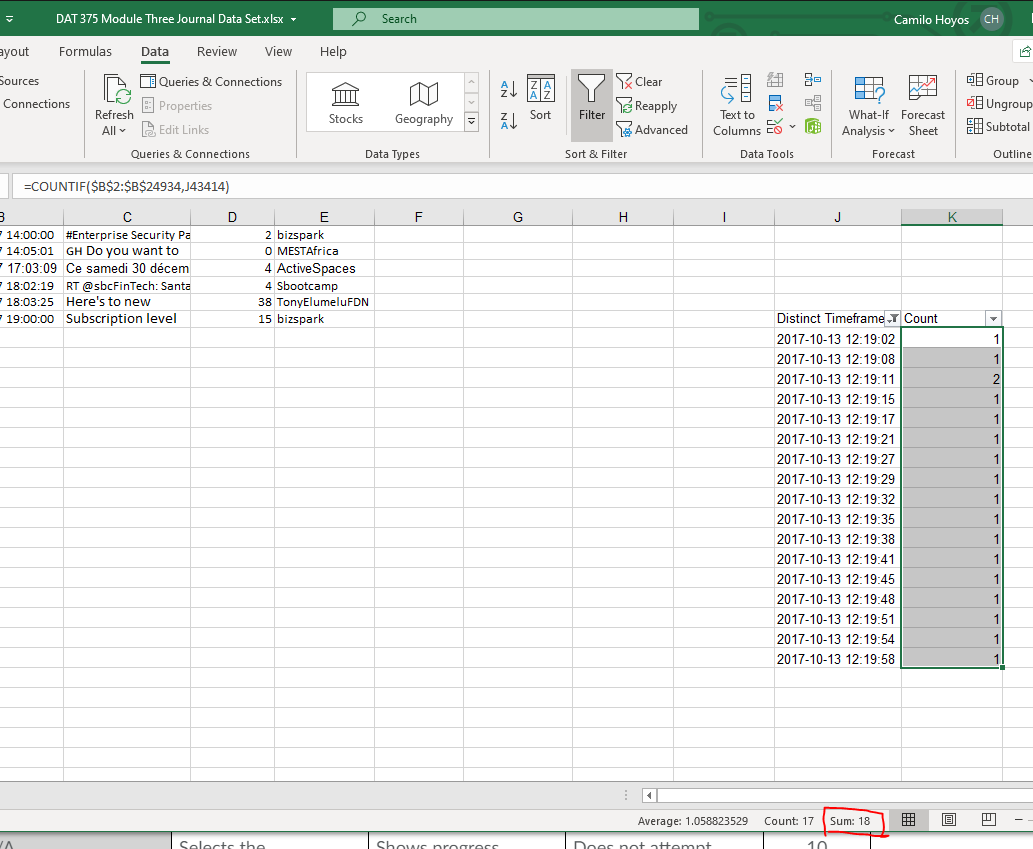
Since this is a small dataset, it can be manually verified. Regarding the first question, I can prove that the average number of messages per username is 958.96 with the calculation below:



I can prove the average number of reshares per username is 10627.34 with the calculation below:



I can prove that 2017-10-13 at 12:19 had the highest original message count at 18 with the calculation below:



References

Bhatia, M. (2019, July 12). Your Guide to Qualitative and Quantitative Data Analysis Methods - Atlan: Humans of Data. Retrieved January 19, 2021, from https://humansofdata.atlan.com/2018/09/qualitative-quantitative-data-analysis-methods/