Guilherme Osorio

Kathryn Zwick

Jacob Karadsheh

Evan Yedinak

**TWEETS ON THE COVID 19 VACCINE**

**EXECUTIVE SUMMARY**

The data used in our presentation contains all tweets referencing the Covid-19 vaccine from December of 2020 until present day. This dataset has a total of 16 variables of interest which include, but are not limited to; the location of the user, number of retweets, the follower count of each user, date, and the actual text of the tweet itself. Three fields that contain numeric values in our dataset are user\_created, retweets, and user\_followers. The user\_created field contains the date of when the user who tweeted a tweet about the COVID-19 vaccines created their account and is represented as a continuous variable. Retweets and user\_followers are the number of retweets and user\_followers a tweet/user\_name has and is presented as a discrete variable. The original dataset contains 228,207 observations. We selected this dataset not only because it pertains to a currently relevant topic, but also because the content was easy to understand, the dataset has a good structure (cross-sectional), and there is a large amount of data to TIDY up. Given the inconsistency in the content provided throughout the dataset, a significant amount of cleaning and preprocessing was required prior to our exploratory analysis.

**THE PREPROCESS: CLEANING THE DATASET**

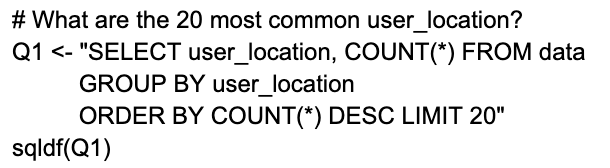
After examining the dataset, we realized that the user\_location variable had many locations that did not make sense such as “My bed” in addition to many NULL values. We knew we had to clean up this column of our data. We began by running a query that lists the 20 most common user\_locations in the dataset. Based on these results, we concluded that the most common user\_locations come from five different countries: India, Canada, US, China, and the UK. We decided to divide our user\_location column into these 5 countries.

After this, we created five new data frames based off of our five user\_locations and included only data from the respective five user\_locations. By only showing data from these five user\_locations, we consequently removed all NULL values in our tidied up dataset. In the process of creating the new data frames, we organized the user\_locations so that different cities in each country would classify the user\_location as either India, Canada, US, China, or UK. For example, for India, when user\_location == “Mumbai” or user\_location == “New Delhi”, these user\_locations would all be converted to India in the new data frame. We did this to tidy up the user\_location column and make it easier to run queries and visualizations with it. After creating our five new data frames, we merged all five data frames into one named data\_5.

Our initial dataset consisted of only one table, instead of the two required tables. Therefore, we had to create a new table (data frame). For our new table, we decided to use it to store user\_location information and refer to it as locations. The new table consisted of three columns: country\_ID, user\_location and GDP. We added GDP to this table to include more information about each of the user\_locations so people who read the dataset can find out more information about each user\_location. On the data\_5 data frame, we added a column to link the country\_id from the new table to the name of the user\_location on data\_5 and then removed the user\_location column from data\_5 so user\_location would only be represented by the country\_ID.

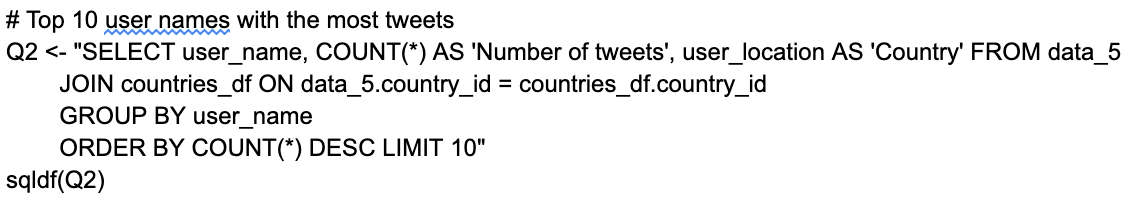
**EXPLORATORY ANALYSIS**

**QUERY #1:**



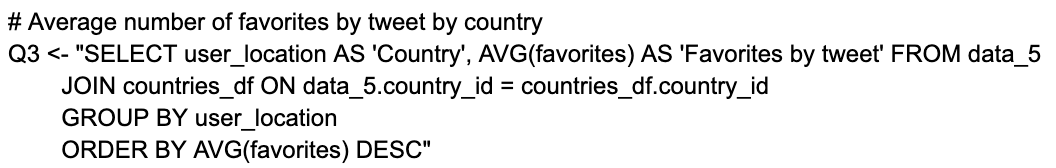
Our first query was a query to select the top 20 most frequent locations in the user\_location column. We used this query to determine which locations would be best to use for our tidied up data frame. From this query, we found out that the most common locations were: India, Canada, China, USA, and the UK. From this query, we were able to use these five locations in our tidied up data frame. In the new data frame, any city from the top 20 mentioned user\_locations was converted to one of the five most common country locations to make it easier to run future queries and visualizations. We also found out that out of the five most common locations, India was by far the most common user\_location.

**QUERY #2:**



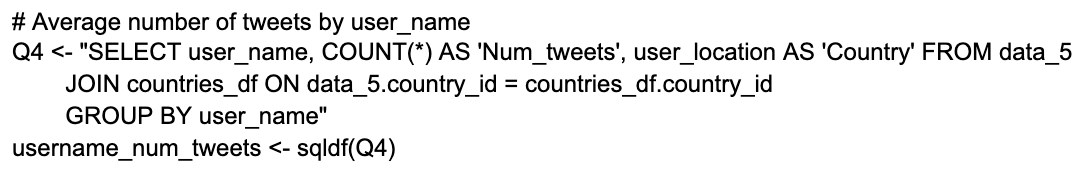
Our second query selected the top ten usernames and their respective country of the usernames that had the most tweets about the COVID-19 vaccine. Out of the top ten usernames that tweeted the most about the COVID-19 tweets, five of those were from India.

**QUERY #3:**



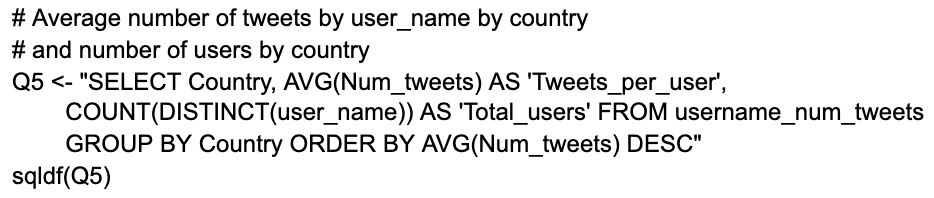
The third query selected the average number of favorites per tweets per country. The country with the highest average number of favorites per tweet was China. This shows that even though India is the country with the most tweets about COVID-19, tweets with user\_locations from China had the highest number of average favorites per tweet while India had the second highest average number of favorites per tweet. India was second with a difference of approximately eight less favorites per tweet.

**QUERY #4:**



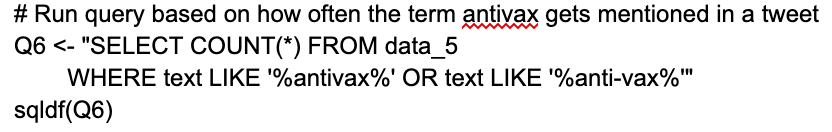
The fourth query shows the average number of tweets by user\_name and their associated country. From this query, we were able to observe how frequent users from each country tweet about COVID-19. From the results of this query, we observed that although usernames from India had the most cumulative number of tweets, typically usernames from India had less tweets per username about the COVID-19 vaccine. Therefore, more people from India tweeted about the COVID-19 vaccine as opposed to a few people from India tweeting many times about the COVID-19 vaccines.

**QUERY #5:**

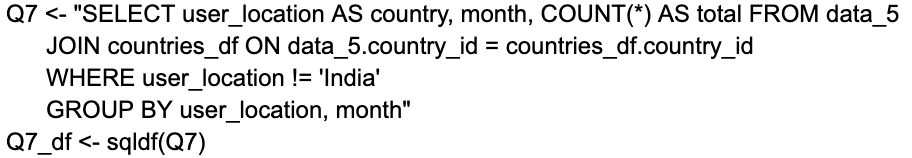


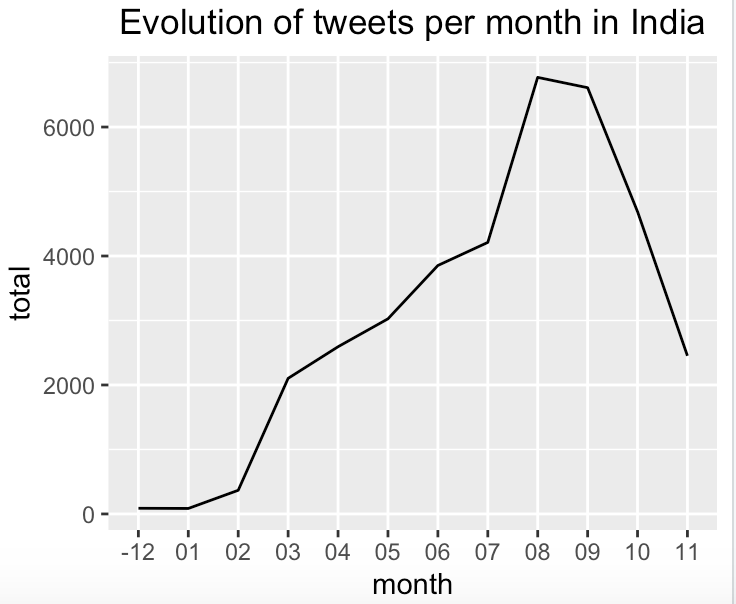
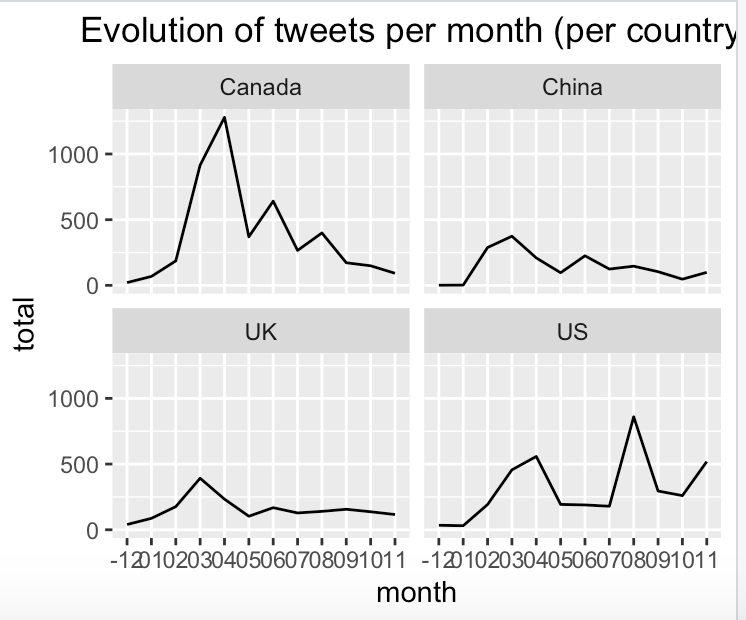
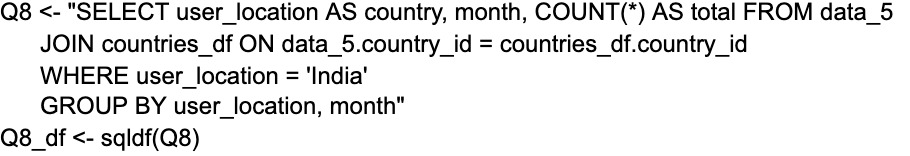
The fifth query shows the average number of tweets per country per user and the total number of usernames from each country. This query confirms the results of our fourth query. China had the highest number of tweets per username with an average of 28.08 tweets per username out of 61 users who tweeted about the COVID-19 vaccine. The number of average tweets per user from each country significantly dropped from here. The country with the second highest number of tweets per user was India with approximately 6.05 tweets per user with a total of 6094 users from India tweeting about the COVID-19 vaccine. Canada, the UK, and the US had less than six tweets per user.

**QUERY #6:**



The sixth query shows the count of how often the term “antivax” is mentioned in a tweet and returned with a result of 15. This returned lower than we initially thought it would since there were so many tweets being analyzed in the data frame.

**QUERIES #7, #8 AND VISUALIZATION #1:**

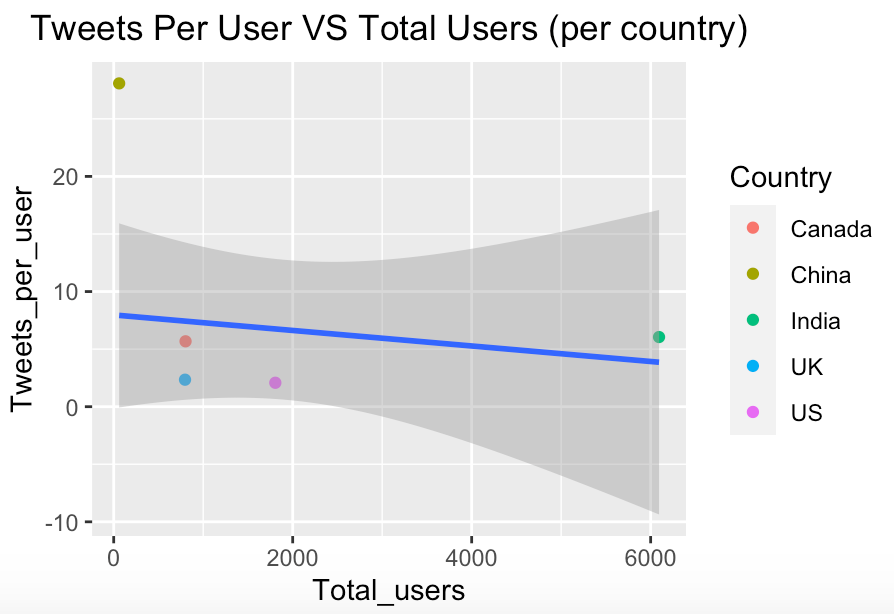


To create the first visualization, we added new columns to the data\_5 to extract the year and month of each tweet and added a new column to add the count of tweets for each month. In preparation for visualization one, we created a new data\_frame known as Q7\_df which included the country, month of tweet, and the total number of tweets from each month by joining the countries\_df and data\_5 country\_ID. We ran two queries (number 7 & 8) to select count of user\_location as country, month, and count of total tweets from data\_5 where user location is equal to anything except India and where user\_location is equal to India. We then created our first visualization using the results from this new dataframe.

Our first visualization is a collection of plot graphs with geom lines which show the evolution of the number of tweets through the year (so 12 months) for Canada, China, the UK, and the US. The country with the most variation in the number of tweets about the covid vaccine throughout the year was Canada. They had the most tweets around March-June and after that it declined. The other countries had a lot of tweets at the same time, but their peak was not as high. After the beginning of the year, Canada, China, and the UK’s number of tweets about the COVID-19 vaccine got lower. The US was different because their number of tweets about the vaccine reached an all time high in roughly August. Further analysis is needed to determine why the United States has an all time high number of tweets about the covid vaccine in August opposed to in April.

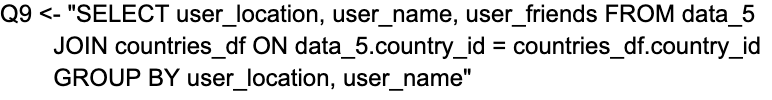
Our second part of visualization one shows the evolution of tweets per month in India. Unlike trends we saw in the previous visualization, we saw that India reached an all time high number of tweets around the month of August (just like the US). However, unlike the US, they did not have a spike around the beginning of the year and instead increased during the year until the month of August and then began to decline.

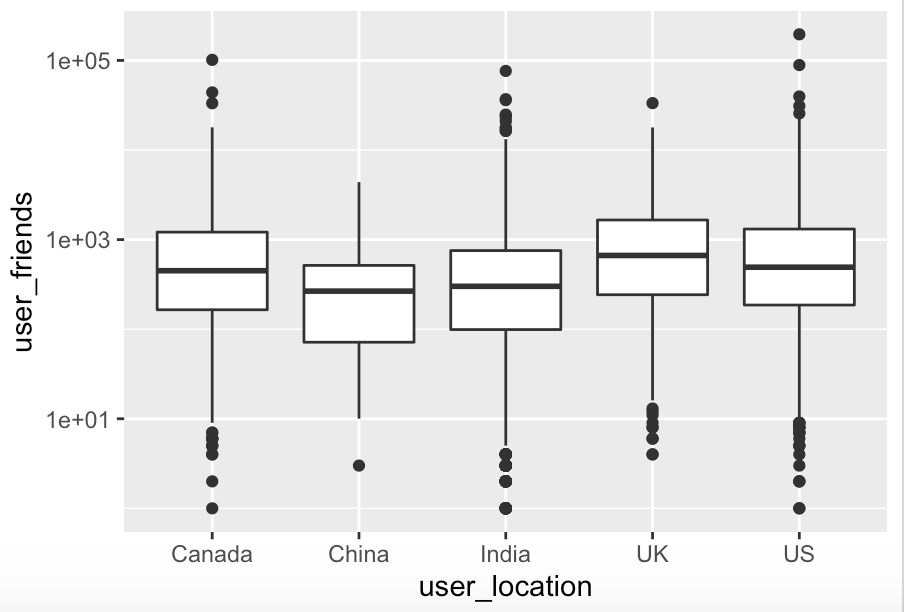
**VISUALIZATION #2:**

****

The second visualization is a visualization of favorites per tweet combined with the average number of tweets by country and total users (per country) named “Tweets per User VS Total Users (per country). This visualization looked at the results of queries 3 and 5, combined these results into a single variable, and used this to create the visualization. Our goal of this visualization was to see if there is a correlation between the number of tweets and favorites per tweet and total users. We used a plot graph with a smooth rlm line. The correlation for the graph is -0.3394. Based on these results, we can conclude that there is a negative correlation between these variables. However, the relationship is not very significant.

**QUERY #9 AND VISUALIZATION #3:**

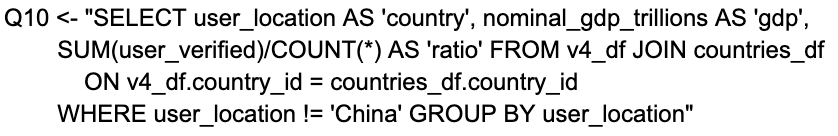
****

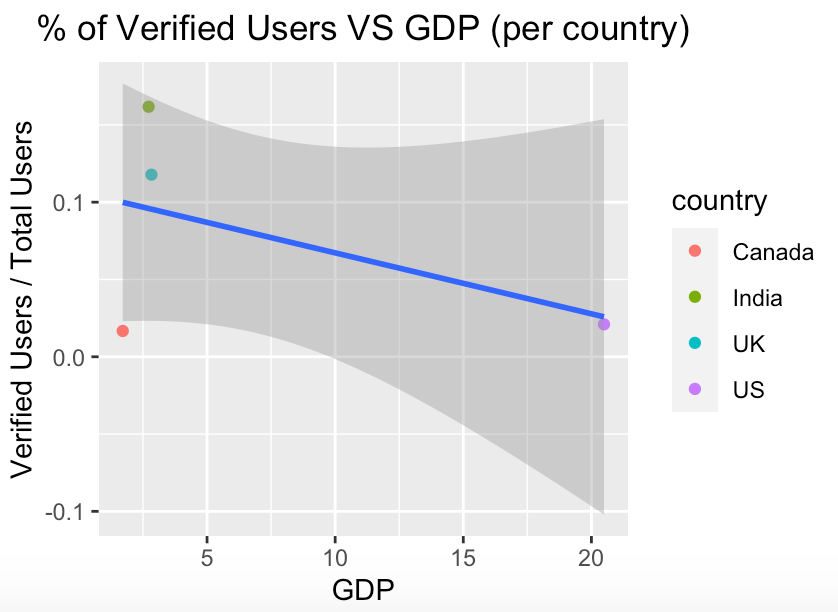
****

The third visualization is a box plot visualization that shows the number of user\_friends by country. From this visualization, we determined that the amount of user\_friends per location does not have a lot of variation. The median for each of the country’s user\_friends is fairly similar. However, the range between each country's user\_friends differs. China, the US, and India have large ranges for the amount of user\_friends while Canada and the UK have few outliers.

We used query #9 to create this visualization. We needed to join the country.id on data\_5 and the countries\_df table to select user\_location, user\_name, and user\_friends from data\_5.

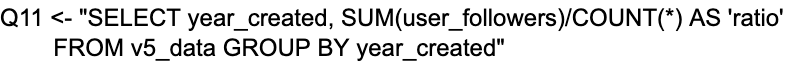
**QUERY #10 AND VISUALIZATION #4:**

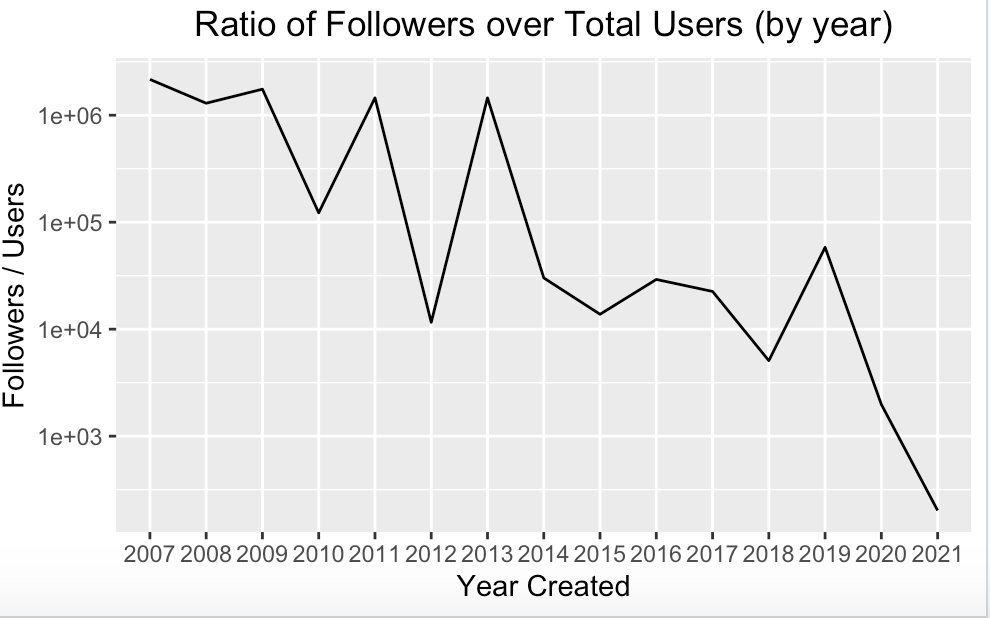
****

****

The fourth visualization shows the relationship between the ratio of count of user\_verified to total users and GDP. We removed China from this visualization because it was an outlier. We removed China from this visualization through query 10. Query 10 selects the user\_location, GDP, and ratio of user verified to number of users where user location is not equal to China. This query also joins country\_id from v4\_df and countries\_df to create visualization #4. The result of this query shows a negative correlation between the % of verified users to total users and GDP. The lower the GDP, the seemingly higher the verified users / total users ratio. However, the results of this could be skewed since three of the countries (Canada, UK, and India) had a GDP of less than 5 million. The US was the only country in our dataset that had a GDP of above 5 million. Therefore, since all the other countries and essentially being compared to the US, which has a lower verified users / total users ratio, the results could be skewed. Adding additional countries with higher GDP would make for a better comparison.

**QUERY #11 AND VISUALIZATION #5:**

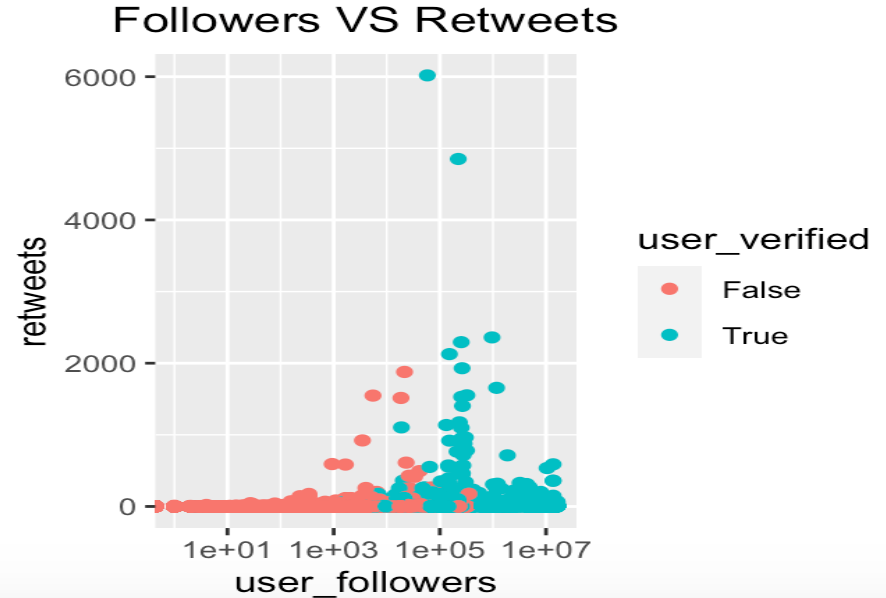
****

****

The fifth visualization shows the relationship between the year of when the user was created and the amount of user\_followers. We created a dataset to use for visualization 5 which includes a column with a distinct column of the year of when the user’s account was created. Query 11 selects the year the account was created and the sum of user followers/count from the dataset for visualization 5. With this dataset, referred to as v5\_df, we created a plot graph with a line which represents the ratio of user followers over total users and the year their account was created.

This graph shows fluctuating numbers for the amount of followers depending on when your account was created. However, for accounts created in 2014, their typical number of followers is much lower than accounts created before that. Another drop off occurs during the year 2019. Accounts that were created in 2019 have a very small amount of followers (most likely because they are newer). This is significant because it appears as if accounts created during the covid pandemic and tweeted about the vaccine do not have many followers.

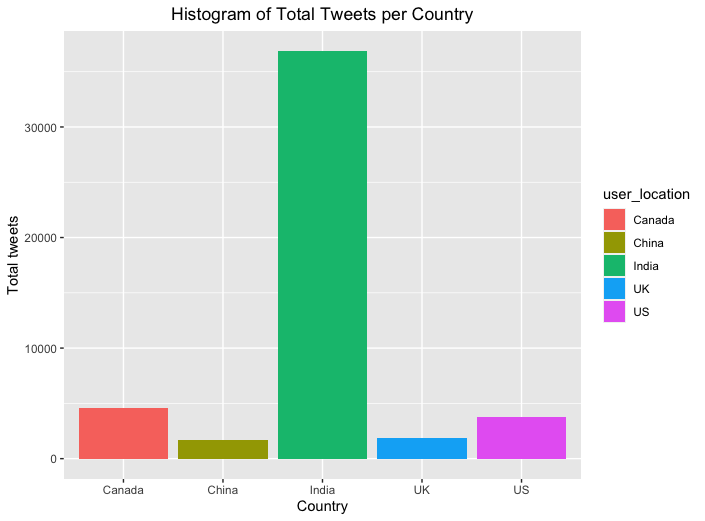
**VISUALIZATION #6:**

****

This visualization is a scatter plot of user followers and retweets. We also colored the dots so that we could determine between users that were verified and users that were not. It appears as if there is a relationship between being verified and a high amount of user followers. Nearly all of the users with a large amount of followers are verified. The relationship between user followers and retweets is unique because although retweets appear to increase as user followers grow larger, the tweets with the highest retweets are not the users with the most followers. It appears as if the highest retweeted tweets are towards the 1e+05 follower range and consist of both verified and unverified user followers but contain primarily more users that are verified than users that are not verified. There does appear to be a relationship between users that are verified and retweets. Out of the users that have tweets above 1000 retweets, there is a lot more users that are verified than users that are not verified implying that there is a positive correlation between those that are verified and users that have more retweets.

**QUERY #12 AND VISUALIZATION #7:**

****

****

This visualization is a histogram of the total number of tweets across the US, UK, China, India, and Canada. From this histogram we can clearly see that India has by far the most tweets of any of the countries with more than thirty thousand tweets followed by Canada, US, UK, and China with the least. The bars are also color coded to show which bar belongs to which country.

**CONCLUSION**

Through this project, we expanded on our knowledge of coding and TIDYing up our own dataset and realized how much work actually goes into these two steps when starting from scratch with a dataset.

​​ To conclude, we used a database of all the tweets that reference the COVID-19 vaccine from December of 2020 until present day for our Final Group Project. Since the original dataset contained 228,207 observations and a lot of NULL or nonsensical values in the user value field, we had to perform a lot of tidying up the dataset before we could perform our analysis.

In our analysis, we ran a total of 12 queries and 7 visualizations. In our queries, we determined variables such as the top 10 usernames with the most tweets, the average number of favorites by tweet by country and by user\_name, the average number of tweets by user\_name by country, and performed a lot of joining tables to complete our visualizations.

We performed a total of 7 visualizations. Our first visualization showed the evolution of tweets per month by country. Through this visualization, we observed that tweets about the vaccine varied per country but typically peaked in either April or August for each country. More analysis would be needed to determine why these specific months had the higest number of tweets about COVID-19. Our group hypothesizes these numbers may align with vaccine distribution.

In our second visualization, we used a smooth rlm line to determine that the variables of tweets per user and total users are slightly negatively correlated.

Our third visualization shows that user\_friends and user\_locations do not appear to have a very significant relationship. The median of each of the friends were around the same value. However, in countries (i.e. China and the UK) that had lower ranges of friends, our group believes further analysis is required to determine why these have lower range values.

Our fourth visualization shows the relationship between the % of verified users and GDP. From this visualization, we determined that the lower the GDP, the higher the ratio of verified users to total users. However, we believe that more countries with higher GDPs should be included in this analysis since only one of the countries has a GDP higher than $5 million (the US).

Our fifth visualization displays the ratio of followers over total users (by year) to year their account was created. The earlier the account was created, on average, the more followers the user has. Accounts created in the past 2 years have a significantly less follower count than accounts created before then.

Our sixth visualization is a scatter plot of user followers versus amount of retweets with color coded data points that show whether user is verified or not. With this visualization, we concluded that user followers and retweets are positively correlated to an extent, user verified = TRUE and user followers are positively correlated, and user verified = TRUE and retweets are positively correlated.

Our seventh visualization is a histogram of total tweets per country. India by far has the most number of tweets with over 30,000; while the rest of the countries have less than 5,000.

Through these visualizations, we figured out a lot of trends about our dataset. Through the process of coding, we also deepened our understanding of coding using the sqldf() and ggplot2() library. There is a lot of trial and error in coding, but we are very happy with our final project.