Which party utilized Twitter more than the other and did it aid them in their elections?

<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/NMT4HP>

# Extracting, transforming, and loading

The dataset for this project came in a zip file that contained a .json file for each senator which I believe had most if not all the tweets for the 115th congress up between 2015 and 2019. The parties were separated by folder, so it was easy to identify their party. I extracted the data using a custom application I made that I will provide with this submission including the source code.

After giving the application the path to the dataset it goes through each folder/file and splits the data up into objects such as politicians and tweets. Afterwards the data is inserted into a MySQL database with a new structure: **politician**, **tweet**, and **politician\_tweets**. The **politician\_tweets** table is used to link politicians to tweets. Once all the files and folders have been looped through the database connection will close and the process is complete.

# Cleaning the data

One of the most notable actions I had to take to clean the data was correctly saving the politician’s name. The file name is the same as their twitter handle and using that the full name of their Twitter profile is fetched via searching the json for that handle. I had to do it this way because some records in the json had tweets from different users because of retweets.

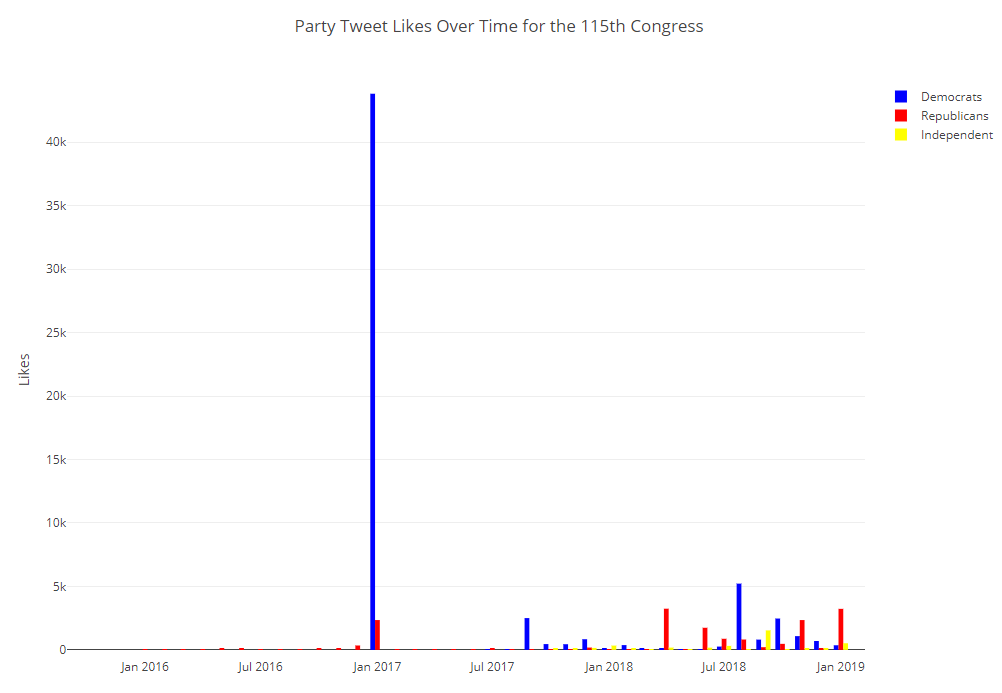
Even doing it this way there were different naming conventions for their twitter names such as prefixing it with “Senator” or “Sen.”. To make all the names uniform the application removes these words along with some others. However, after extracting all the data I discovered two congressmen whose Twitter names did not have spaces like Chuck Grassley’s and Dan Sullivan’s. I had to fix these manually via database queries because I didn’t bother adding a special case to the program to detect these specific cases. There was also the case of Mitch McConnel’s name being Leader McConnel.

Other cleaning that occurred was storing the tweets in the **tweet** table with the ID being the primary key, this way duplicate tweets would not be inserted into the database. If multiple politicians retweet the same tweet, it will be reflected in the **politician\_tweets** table.

# Analyzing data

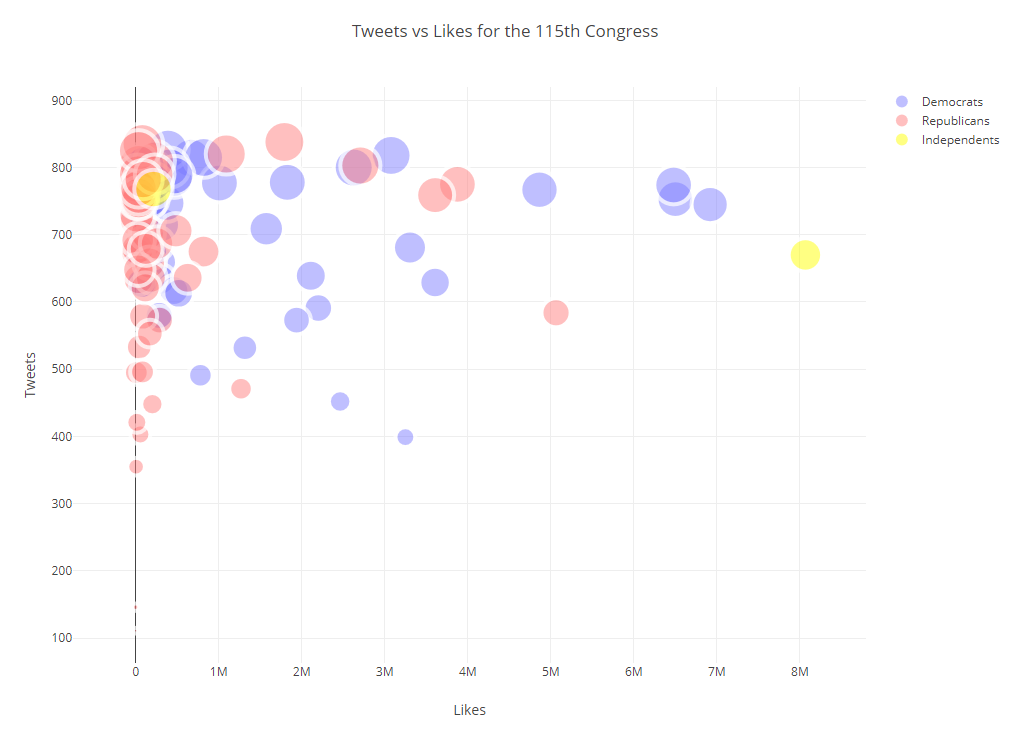
After the data has been extracted, transformed, loaded, and cleaned, the analysis can begin. At first, I planned to just use SQL queries but found a neat plugin for python called plotly that allows me to generate graphs. The code for generating these graphs will also be provided with the submission.

However, I should note to make the graphs more readable I had to limit the date of the tweets selected to only be from 2016 and onwards. For some reason the dataset included a few tweets dating back to 2009 which seemed a bit odd. While I could’ve removed these, I decided not to since I can easily just filter them out with a query.



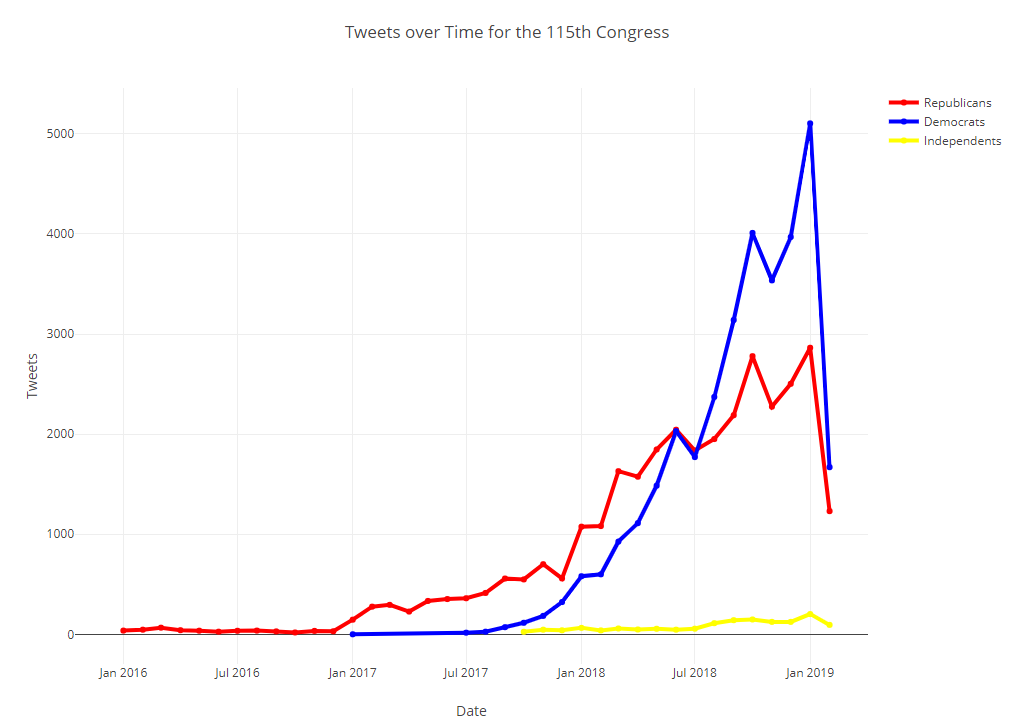
<https://plot.ly/~jon.antonop/2/party-tweet-likes-over-time-for-the-115th-congress/#/>

This first graph displays the amount of likes each party received over time. One thing that is extremely noticeable is that in January 2017 the Democrats received more likes than ever in this dataset. Although it causes the graph to be much larger, I couldn’t really leave it out. It’s difficult to speculate the reasoning for this, considering Trump had won the 2016 election. If I were to guess, there were possibly a lot of tweets attacking him as he was inaugurated and maybe it attracted unprecedented likes. Another small spike is seen in August 2018, which could be related to the then upcoming midterm election. What is clear from this graph is that Republicans do not receive many likes on Twitter compared to Democrats.



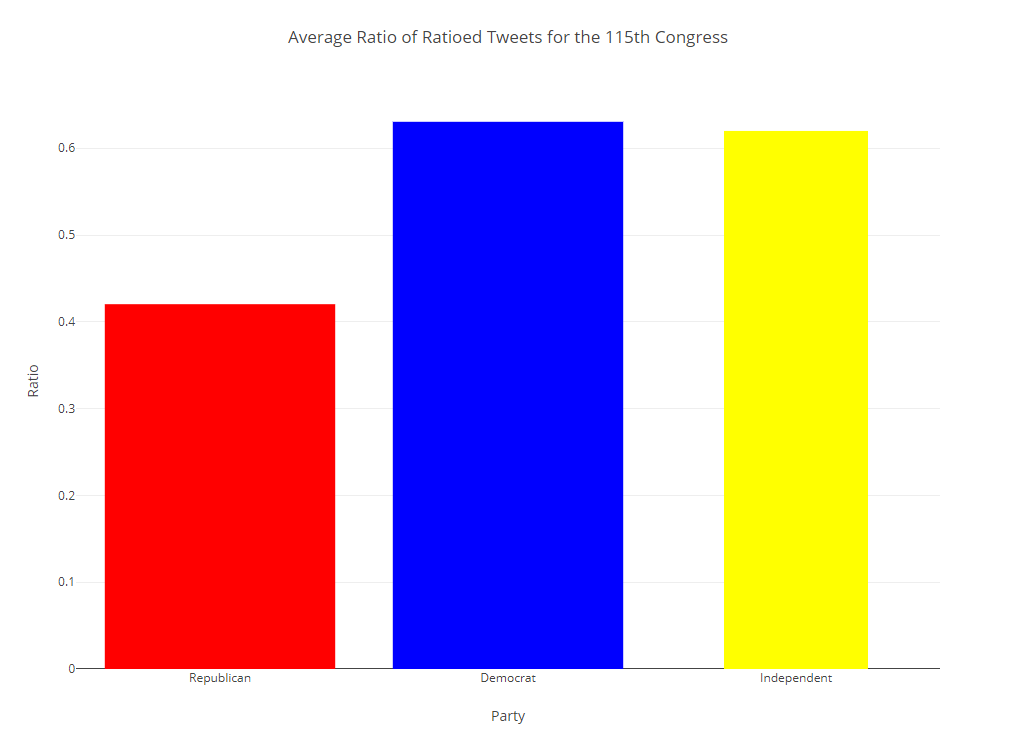
<https://plot.ly/~jon.antonop/0/tweets-vs-likes-for-the-115th-congress/#plot>

This next graph depicts the number of tweets a politician made and the total number of likes. If you’re viewing the graph on plot.ly you can see their names by hovering over the markers. The size of the marker is based on their tweet count. A major takeaway from this graph is that the tweets for Democrats receive a lot more likes than Republicans with most of them having under one million likes. Bernie Sanders comes in first place. This graph is interesting because it seems to demonstrate that Republicans are not received that well on Twitter. However, this could be due to multiple factors such as messaging/rhetoric, and/or the partisanship of Twitter. Another thing I noticed is that total tweet counts for each senator are weird, as most practically have about 800 tweets in this dataset. It makes me wonder if the dataset has been clamped in some way other than by date.



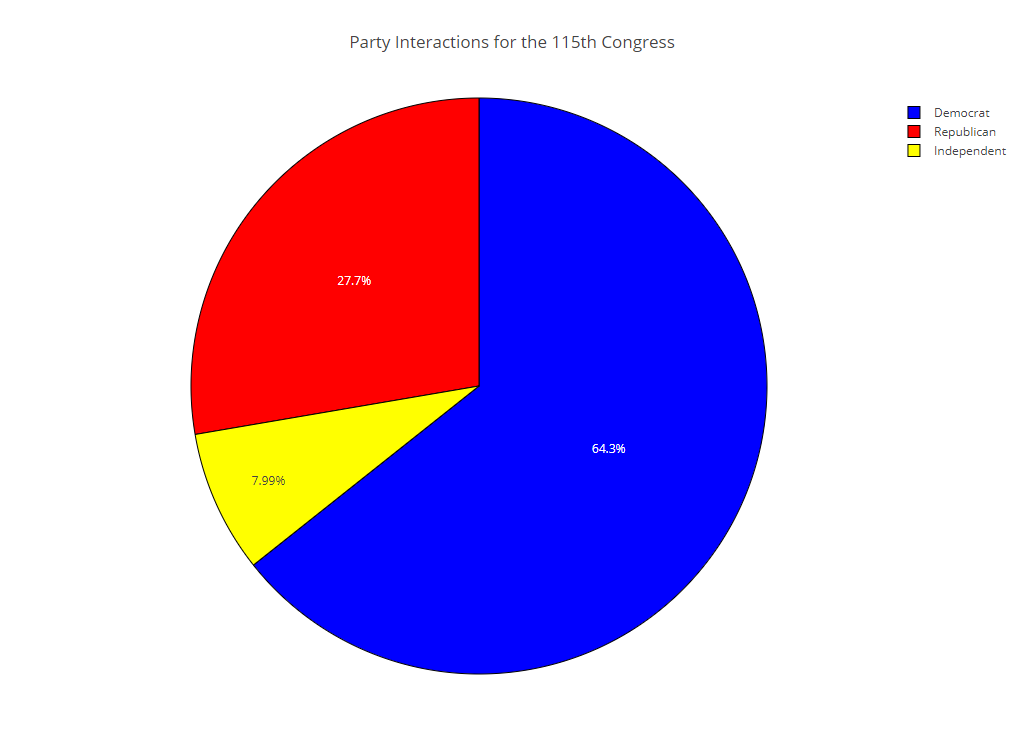
<https://plot.ly/~jon.antonop/8/tweets-over-time-for-the-115th-congress/>

I found this graph to be one of the most intriguing ones. There are multiple things that I’m able to speculate based on its illustration. For one, during the time of the 2018 midterms, the number of tweets for both parties practically doubled. At the beginning of the year the Republicans held a slight lead, but overtime Democrats doubled their tweet count. Another observation is if one assumed correlation implied causation, with this graph alone one could attribute the number of Democrat tweets being a factor in taking the House of Representatives during the midterms. However, that would be quite a stretch, but it is interesting seeing that a high tweet count accompanied their win. One last observation is that the number of tweets for both party’s plummet after January 2019. It’s difficult to think of a reason for this, especially considering the upcoming primaries and general election. Perhaps the importance of Twitter among politicians is fading, or maybe they were on holiday.



<https://plot.ly/~jon.antonop/6/average-ratio-of-ratioed-tweets-for-the-115th-congress/#/>

This graph may be a bit difficult to understand at first glance. Being “ratioed” on Twitter means either receiving more comments than retweets or receiving more comments than retweets and likes. Only the retweet/comment ratio is display in this graph. The higher the bar is, the better the average ratio was for tweets that were ratioed. This does not include non-ratioed tweets. One thing that this graph demonstrates that the others do as well is that Republicans again are less well received. Another aspect that should be pointed out is that Independent data mainly consists of just Bernie Sanders, so one can assume his tweets are generally very well received.



<https://plot.ly/~jon.antonop/14/party-interactions-for-the-115th-congress/#/>

The pie graph depicted above shows the percentage of interactions each party received, which includes likes, replies, and retweets. As the other data has indicated, Democrats hold a much larger percentage. They have received over double the number of interactions that Republicans did. However, not much of a conclusion can be drawn from this graph alone as interactions can be positive, neutral, or negative. It may indicate which party’s tweets had more viewers though. Considering the ratio graph above, combined with this one, you may be able to assume the Democrats received more pleasant interactions.

# Conclusion

Based on the data I’ve reviewed and analyzed, only the first part of my question can be definitively answered. The Democrats by far utilized Twitter more than the Republicans. Not only did they use it more, but they were often very well received compared to their colleagues. As I mentioned before this could be due to multiple factors, such as bad messaging and rhetoric, poor tweet composition, and the overall partisanship of the Twitter userbase. It is not exactly a mystery that Twitter seems to be mostly used by Democrats, Liberals, and others on the left side of the spectrum. So, my answer isn’t really surprising although it does confirm my initial suspicions that Republicans underutilized Twitter. While past speculation by media may have stated the opposite, my data does not seem to reflect that.

Attempting to gauge whether the Twitter activity played a role in the election is going to be speculative at best as it’s not something that can easily be proven if at all. However, based on the data during crucial election times Democrat Twitter activity shot far above the Republicans. During November 2018 the Democrat tweet count was nearly double the Republican’s. While this can’t really be directly attributed to the results of the election it could very well have played a factor. One thing is for certain, Republicans should probably learn to better utilize Twitter. It’s not going anywhere, and they may need to if they want to remain competitive.

The files submitted along with this document will include the python scripts used to generate the graphs, and the visual studio project for the ETL application. However, I scrubbed the credentials for security purposes, so to be used they would have to be replaced.