# **Milestone 3: Beyond Descriptive Stats**

## **Dive Deeper**

Look deeper into the features you are investigating, consider:

* Relationships / Correlation, Pearson Correlation
* Linear Regression for future prediction (if the relationship is linear)
* Textual Analysis for TF-IDF (Term Frequency-Inverse Document Frequency; Row-based and column-based, stop-word removal?

Specify 1-2 correlations you discovered. List the fields that you found to be correlated and describe what you learned from these correlations.

I performed a simple correlation matrix with the numerical data in the dataset and I found a poor correlation between retweet\_count and favorite\_count. This means that in general the count of each is random.

I also performed a TF-IDF analysis between 2008 and 2017 years. I found that he most used word in the tweets is attending. This suggest that the congressmen are always tweeted which events they are attending. Interestingly Arkansas is mentioned a lot throughout the years, which probably suggest there is a lot of importance of that place.

Finally, I did a sentiment analysis and found that generally the tweets were neutral with a slight skew to the positive side. There are quite a significant number of “very” positive tweets that is worth investigating.

## **Go Broader**

Expand the features you are investigating. Look for connections/relationships that you may have initially missed.

1. What jumps out at you now?
2. Use the descriptive stats to point you to features that you may now want to consider.

What key terms did you discover in any text analysis, for whom? Any themes? If you are not analyzing text, summarize what other things you are considering in your analysis?

One of the things that jumps to me is the small amount of negative tweets that congressmen tweet. A deeper look at it would be good to have.

#pork is mentioned a lot throughout all the years of the series what is the cause of this?

It would be good to see what the top 5 retweeted peoples sentiment tweet is and see if there are new insights coming out of it.

Also it would be good to find what are the users that are mentioned the most and show the top 10.

## **New Metric**

Create 1 or 2 new metrics to track relationships of data you discovered. Explain why you created them.

I have created one column to track when the top 5 words is used for easier filtering of the data.

Another metric I have computed is the time it took each user tweet the next tweet. This can give an insight into the number of days each person tweets.

One last metric is to generate 2 columns each with the list of “@” or “#” that the tweet contains. This can be useful to then find the most mentioned/popular people or topics and how this evolves over time.