Peer Review for Allison Barnes

by Matthew Keiser

This is a really interesting idea, I've only seen some surface level work with the Twitter API, but this is a pretty deep application. I don't have much experience with the twitter API, the function that pulls the tweet data into a dataframe is interesting, but it could probably be implemented with more efficient code - I would recommend looking at the .append function to create lists (hashtags, usernames, etc.) inside the for loop, then put them all together using the pd.DataFrame() function. Your code looks like it will work, this is just a different approach.

Have you considered which machine learning algorithm you will be using for your regression? It might make sense to start with Ridge or Lasso, which will let you find the most important features by looking at the model coefficients. This can also be done with RandomForestRegressor or AdaBoostRegressor by pulling the Feature Importance attribute. It would be interesting to compare these methods with KNN to see if there is a significant difference in the mean squared error between the simpler models and the more complex ones. There aren't a large number of features compared to the number of tweets, so you don't need to worry about using PCA or any other kind of feature reduction method.