Comments on Hall group eda

Brendan Brown

SCORES

Total: 20/20

Summary: 4/4

Explore/clean: 8/8

Prelim results: 8/8

NOTES

\* good work. seems you have a plan for extending all of this, so keep it up.

\* the data collection effort is impressive. you might be the only ones doing true data collection here.

\* crazy huge difference in number of stories by month. is that really true, and not some data quirk or quality problem?

\* it's kind of weird and interesting that there's this hump shape to the word length plot over time

\* you could have done a little bit more on the analysis with number of words, word lengths. probably also worth looking at number of unique words, or number of words modulo common words. search around for a package with a list of common words --- or email me if you can't find one.

Other ideas here https://www.tidytextmining.com/tfidf.html

\* interested to know how the subjectivity metric is calculated in particular. doesn't need to be super explicit, but some indication of what they did would be good if possible. That would make it all the more interesting to see subjectivity scores basically flat across outlets

\* slight but noticable increases in subjectivity and polarity over time are interesting. seems too from the boxplots that there is a difference in the spread of these measures, rather than their centrality statistics.

\* for the regression: I forget whether turning datetime to numeric converts it automatically to days since a certain fixed date. worth checking to be sure it doesn't just give some nonesense like 121220 or whatever

\* interesting work on the subjectivity polarity correlation with word lengths and number of words. just be sure the model makes sense, because your outcome is in [0,1] --- mentioned quite often in class so perhaps you thought of this and didn;t mention it