Final Project Notes:

- Data set is missing tons of “types”, the scraper that was used, replaced any type that was “unknown” with “bs”. It’s about 11,000 values.
    - After considering the way the data set was created, I found that it used a curated repository of unreliable websites that had a tagged type. Could try to fill in the bs values and get a more accurate list.
    - After replacement, the distribution is *much* more reasonable. I’m assuming that whoever created this dataset made some sort of mistake when getting the type, or perhaps because it’s a little older the list of websites has increased.
- Data set scraped a chosen 244 websites and seemed to stop after 100 articles on each site. This makes finding the top “fake news” websites a bit pointless, but still worth investigating as some sites didn’t reach 100.
- Initial Histogram of all articles over time shows that the articles decrease over time. Is this biased because of the way that the dataset was created? Or was this from some event?
- Clickbait is a major draw for this type of fake news, is there a common theme among these titles? Is it worth doing a sentiment analysis on the title as well as the data?
- The sentiment data is from personal experiences and are only about one sentence. How will this play into classifying the data set?
- After creating the initial histogram, there seems to be a downward trend. Is there a way to prove that this is biased?
    - Maybe an x-y plot that has the crawled date on the y and the published date on the x?