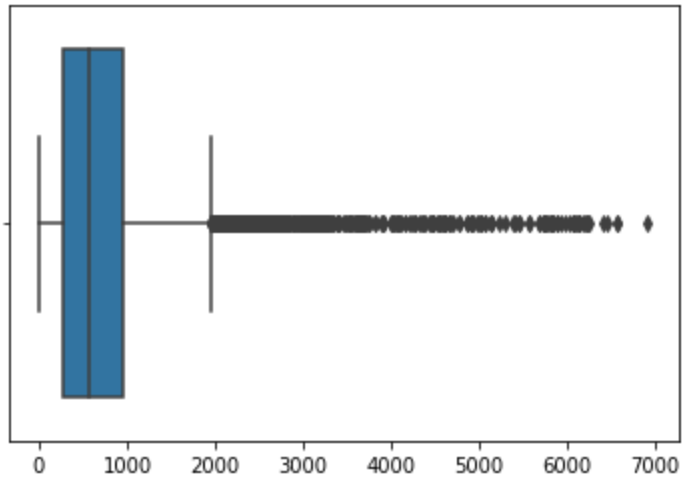
1) Extracted 20,327 news URLs from 2,326 emails
2) Downloaded and parsed urls into title, body text, etc. 17,014 articles were successfully downloaded and parsed.
3) Checked the word count of each article to remove potentially bad data. Limited articles to those between 500 and 2000 words, which is approximately the 50th percentile up to the outlier limit (75th percentile + (1.5 * IQR)). 8,875 articles were of this length.



4) Sample the articles into 4 corpora:
    a) Top Publishers: All articles from the top 5 publishers represented were extracted.
    b) Sentiment: All articles were quantified. The 500 most negative and 500 most positively ranked articles were extracted.
    c) Politics: All articles were searched for four term groups. 4,095 articles had at least one term. The 250 articles with the most mentions in each term group were extracted. There can be overlap between these groups.
    d) Chronology: The articles were split into three years (4-2017 to 3-2018, 4-2018 to 3-2019, 4-2019 to 3-2020). A random sample of 30 articles from each month of those yearly groupings was taken.