Part I: Article: In general, I particularly like how this article generally makes simple the terms they are using. Although this article may be more geared towards the perhaps slightly more educated reader, the metrics they use are explained and shown well enough that even without familiarity with the terminology, one can understand the overall message. All of the charts are consistent and very simple. Each chart is effectively some lines with time as the x-axis. They start with the simplest to understand, which is one about betting odds on economic downturn. They then increase the complexity slightly using consumer price index and then going to consumer sentiment. The last two graphs involve yield curve and Sahm's indicator, which are both terms more geared towards those familiar with economics. The thing that this article does well, however, is that using annotation, it makes it clear exactly what the plots mean, which makes truly understanding the metrics themselves. For example, I don't actually know the explicit way of measuring Sahm's rule (nor do they explain it), but I am still able to read the graph and understand that numbers near 0.5 are bad indicators and can see the marked economic recessions. I also like that the article uses quotations from experts to both lend credence to these charts and also to simplify the underlying meaning of the metrics themselves. Although I won't be able to do this in my own project, it is good to keep in mind my audience like they appear to do here.

Part 2: The topic for my final project has been simplified to simply be comparing the linguistic characteristics of presidents based on major speeches they've made. My research question is essentially: How do presidential speech patterns differ between individuals and also are there any discernible differences between Democrats and Republicans? One thing that will be interesting will be to see if certain clusters of speech tendencies make sense with any of the times that Democrats and Republicans have generally appeared to swap some ideological principles. I originally intended to make the target audience be the general public, but I think I actually would like to make it be perhaps the "technical public." What I mean by this is that I will introduce my project and detail what I am doing in a general sense, but I will be assuming that those reading my article are knowledgeable about NLP to some extent. I am planning on making the output of this project be a website generated from Python using Marimo, which will allow me to also have interactable elements. My goal is to have a streamlined, presentable project that also allows users to investigate the underlying code should they choose. My general structure will be one that is not necessarily strongly defined like a research paper might be. First, I will detail the data itself, a brief description of where it was sourced from, and possibly some descriptive metrics (i.e. number of speeches, political parties, time descriptions, etc.). Next, I will then present each graph that I make that compares linguistic characteristics for each president. As of right now, I am considering that to be sentiment and/or emotion analysis, readability score, lexical diversity, NER distribution, and finally topic modeling. I would also like to include metrics from the audio such as pitch, intonation, pauses, etc., but these would not be

plotted; they would be more for the part two of this. This is time-dependent since they are not easily accessible via their API like the transcripts are, so I need to scrape them. Additionally, audio analysis will add some additional complexity in terms of both researcher and computational time. Potentially, these could be distributed into a table and all organized that way, but I find tables for something like this difficult to compare, especially when some of these metrics will have distributions instead of hard numbers. As such, I do not plan to actually include any tables here. In general, most of the graphs in the project will focus on either the individual speakers or political party. For discerning political party, it will be easy; blue will always represent Democrats, and red will always represent Republicans. For individuals, I will have to differ depending on the plot type. For sentiment, I have two potential graph ideas. The first is simply a bar plot for the average sentiment by party, and then a second lollipop plot that contains individual average sentiment for each president by speech. This could also be replaced with a box and whisker violin plot hybrid, which would reduce space usage. The other option is to make a violin/bar chart that includes all of the presidents and simply colors them by political party; however, this would miss out on the across-party comparison. A similar pattern would be followed for readability score, since generally the same properties apply to it, so it would receive the same graphs. Next, for lexical diversity, I would do a similar plot to the one above; I may also do a plot relating to stop word presence or something else, but that would depend on the results of my exploration a little bit. For NER, I would like to have a similar plot to the lexical diversity where I visualize the types of named entities mentioned. Finally, there would be topic modeling. Here, I would probably do a single composite bar for each president and fill a bar for each speech with the percentage that the speech assigns to a topic. These are at least the ideas I have currently; however, I expect that when I get further into my project beyond the data cleaning and calculating some of the metrics that I have done so far, I will perhaps come up with more efficient or better-looking ideas. Although I believe that my plot selections are mostly good choices for visualizing the data, I would like to vary my chart choices within reason. It is particularly uninteresting to look at a bunch of bar charts, and although I do know that “interesting” is not exactly a good justification, if there are adequate alternatives, I would like to at least make the plots appealing. For each of these plots, I would also like to add a small amount of interpretation as well. I don’t plan on being as verbose or as thorough as I would were it a scientific publication, for example, but clear description and concise interpretation are important for making the website both interesting and useful.

For the second part of the project, I will then be taking all of the metrics calculated in the previous section and then using clustering and/or dimension reduction to try to find patterns. Initially, I would like to use PCA to visually place presidents and see visually if there are any interesting trends across them and across political lines. For this, really the only option is to use a biplot. Next, I would then use k-means to see if

speakers can be clustered easily by political party or if, in fact, that is not at all accurate. This is a perfect use case for k-means since I have a specific number of groups I will be expecting to see. If this proves to be consistent, then no further plot will be needed. However, if there are unexpected results with either result, I would like to then go back and try doing both techniques with individual speeches themselves and plotting that instead, as this may reveal further insight as to why the former would have failed to produce the expected results. I would like to make all of these plots interactable if possible. That will allow people to actually go in and see which presidents fall where and also see scores attached to them without being overwhelming with information. Another added bonus to making these plots interactable is that if I find multiple interesting findings, for example, maybe using a different number of clusters, then I can make this easily visualizable to the end user. Marimo allows you to attach things like sliders with set values to code so I can easily just allow the end user to quickly see my results in the same pane. In general, I know we are supposed to include tables for model results, but I don't think that including those for either the clustering or PCA will add any value and, in fact, will likely be confusing. These are both things that are much better visualized than seen numerically since that is generally the point of their function anyway.

Based on the focus of this project and the target audience, I do not believe that explaining methods is really necessary. It could certainly be interesting to do so, but I am primarily aiming to toe the line between scientific detail and general news article publication.