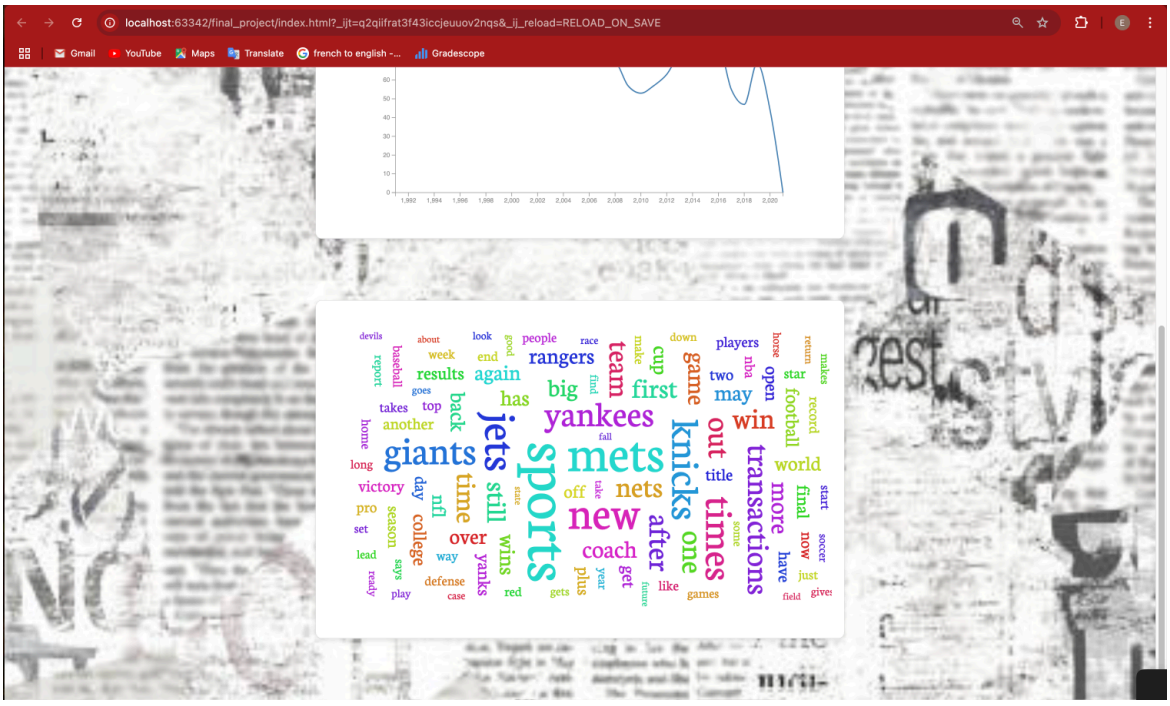
Week 11 - Prototype V1

- Contributors for version 1: Joshua Zhang Eric Vasquez Reyes
- Our design so far:

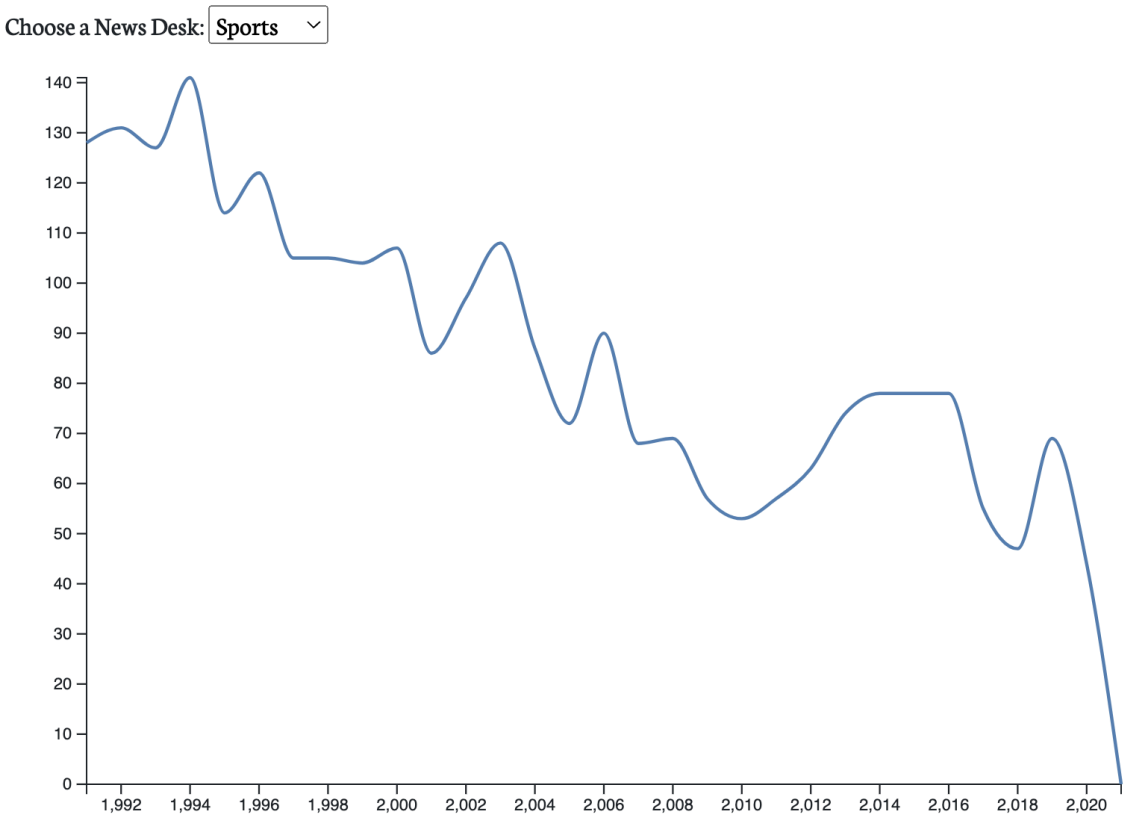


For our first prototype we decided to implement a line chart and a word cloud. The line chart measures the number of articles published by a selected news desk over the years. The word cloud shows the most used words in the article headlines also by the news desk

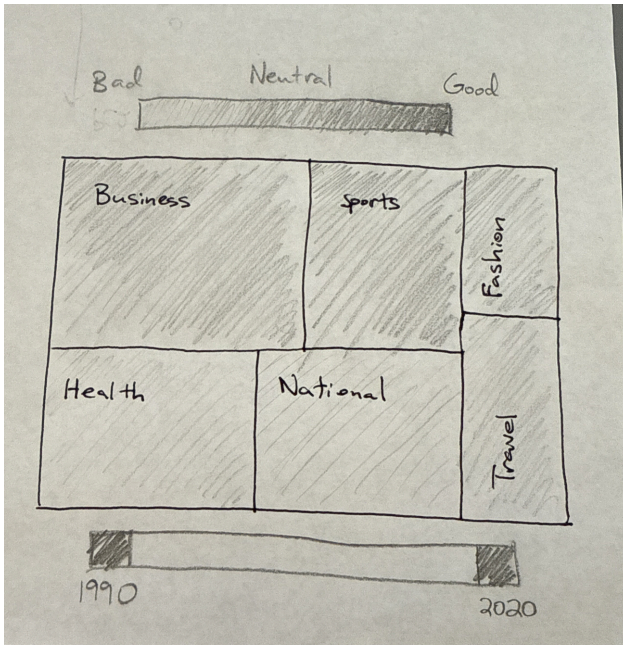
selected. As of now, both of these charts are filtered by a single drop down menu that the user interacts with at the top of the page. We may leave it like this or fine tune the personalization further later on if we deem it beneficial as we add more visualizations and improve our overall website design. For these two visuals, we followed the general pipeline and layout of using 3 functions: initVis(), wrangleData(), and updateVis(), which really helped us stay organized and successfully implement these two visuals.

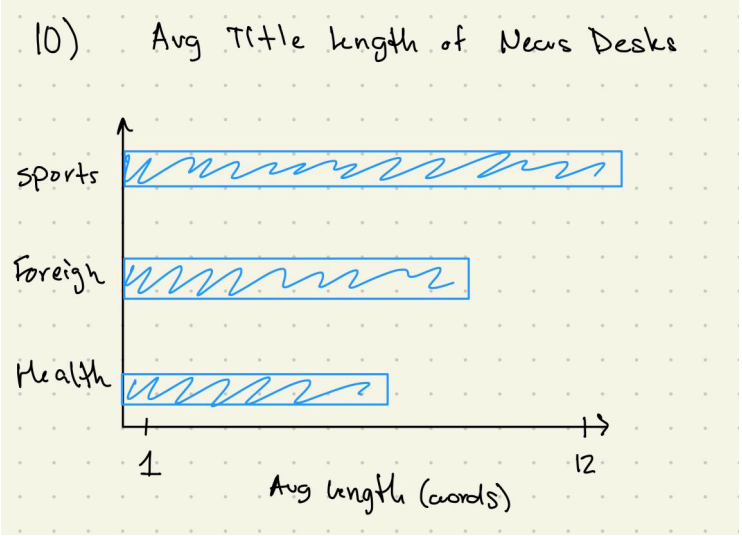
Data Clean Up

- The primary dataset of the project, as provided by the 109a project proposals, does not require significant clean up. We just plan to organize it into dataframes using the pandas library in Python and convert the data types of a few columns for consistency. Even though the data will be cleaned in Python, we plan to write it into a csv file so that we can access it with d3 as we have been.
- Much of our project for CS109 revolves around sentiment analysis and natural language processing. This may require us to deal with large strings and scraped data from the web, which would be done with the Beautiful Soup library in Python.
- However, the point of our work in CS109 is to output quantitative or categorical values for representing our emotional and interpretive findings, which is what we would be visualizing. Thus, as a result of our diligence in data cleaning for CS109, the goal is to keep the data for CS171 simple, as numbers and strings.
- Our dataset was actually too large to be processed reliably live while hosted in the browser. I tried to restrict it to 300,000 data points, but that was slowing the browser down. When I restricted it to 30,000 data points then it was finally small enough to be processed in the browser.
- We cleaned up the data and then randomly selected 1000 data points from each of the 30 years. The end product is shown below.



Next we plan to implement the following 2 visuals:

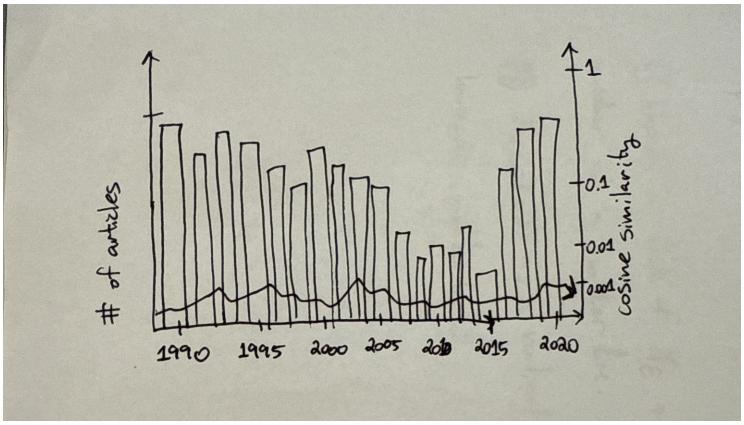
5 	5. How have sentiments in headlines changed over time?
10	10. How do the

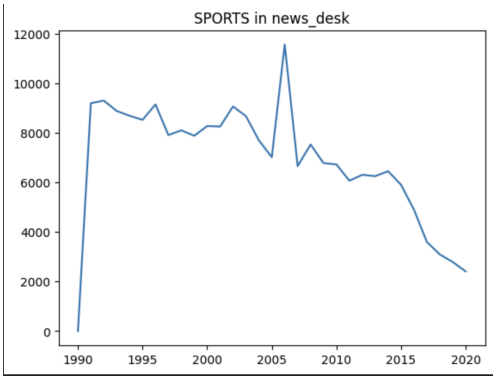
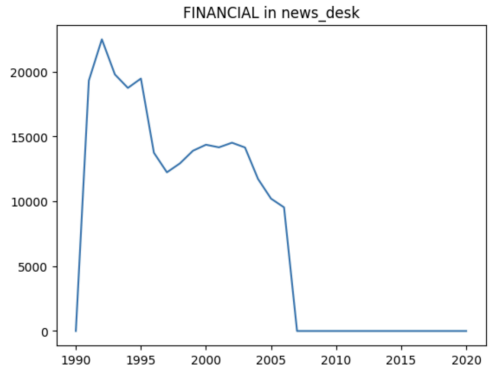
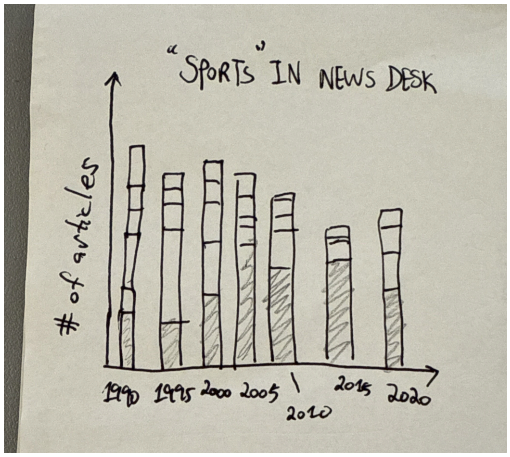
	lengths of article titles vary across different news categories?

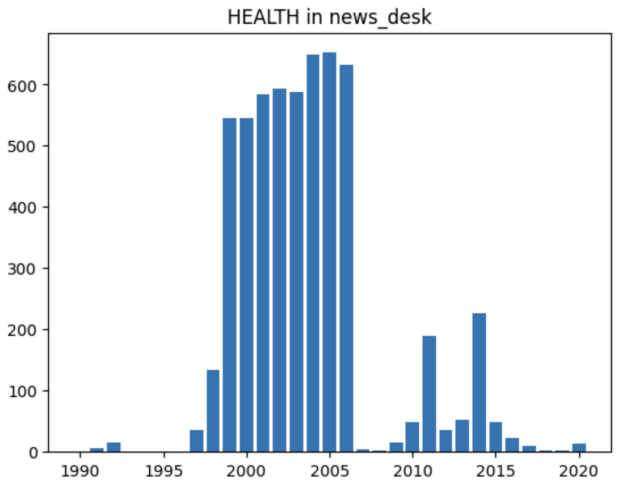
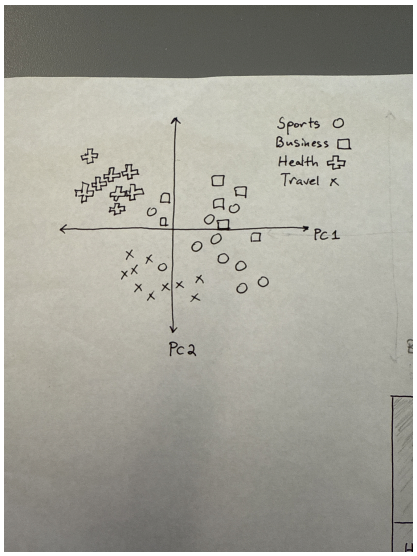
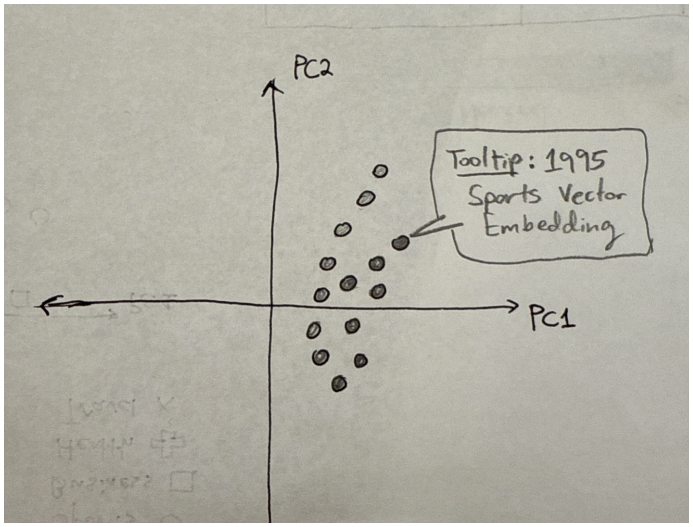
Week 10 - Data, Sketches, Decide & Storyboard

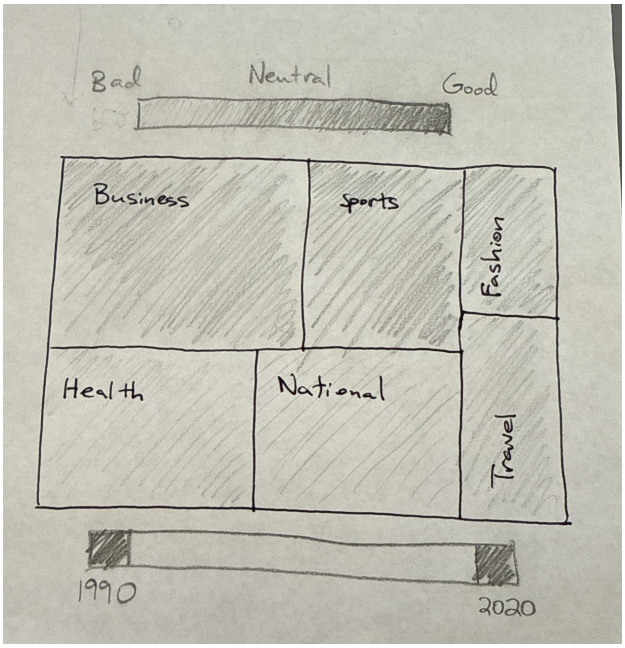
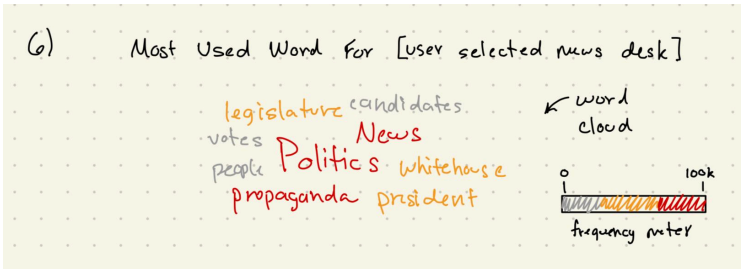
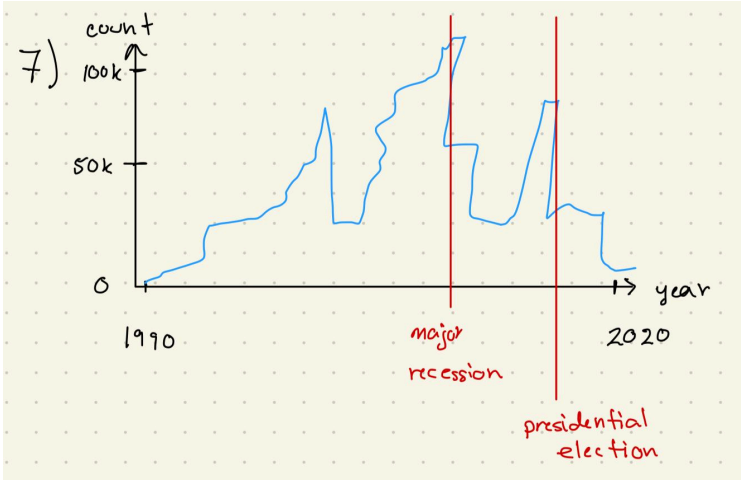
Our data:
https://drive.google.com/file/d/1U9Lf9K8MquYVf1QpgPTzvO4l5eSxVHR4/view?usp=share_link

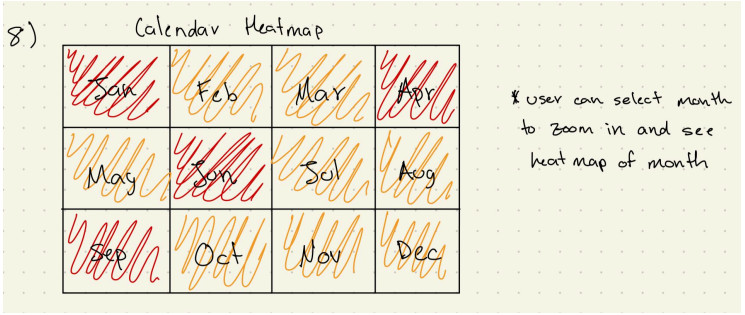
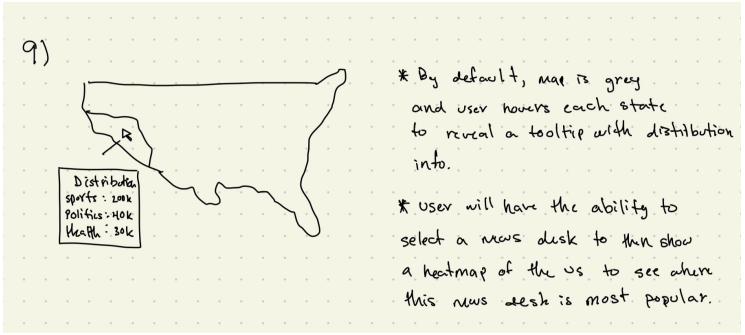
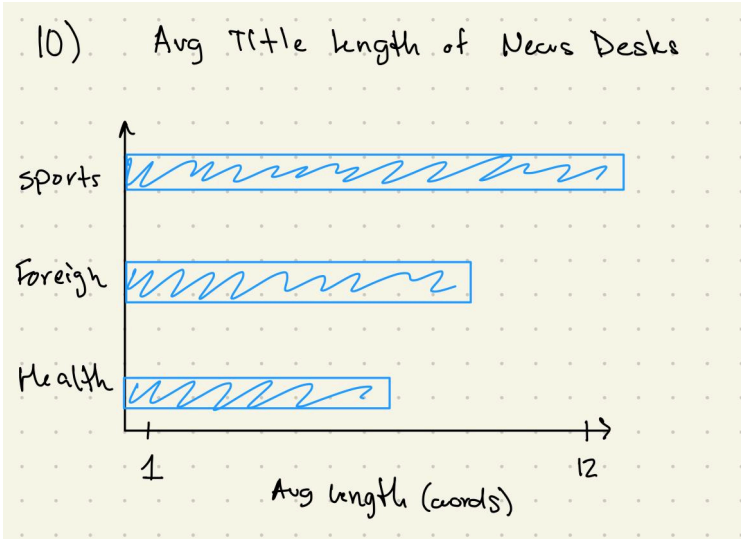
VOTING PROCESS

Sketch ID	Question ID	Author	Votes
1 	1: How have the number of articles in each category evolved over time? And, have headlines in the same new desk gotten more or less similar to each other? Already, we have some primitive line graphs implemented with Python.	JZ	II

 			
2 	2: How does the New York Times organize its different news desk categories? Are there many different categories with the same word?	JZ	

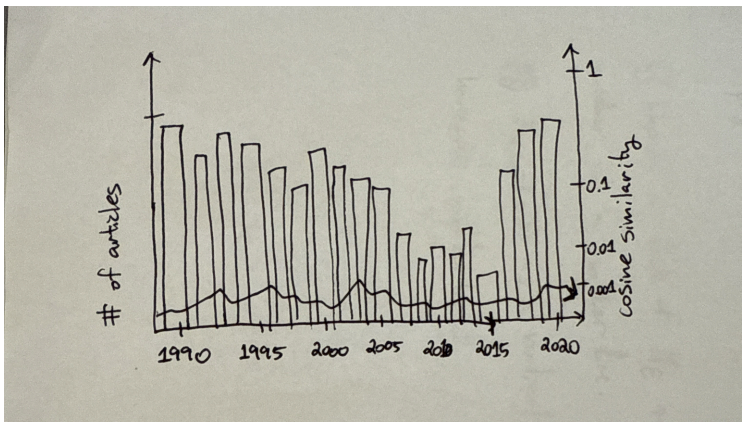
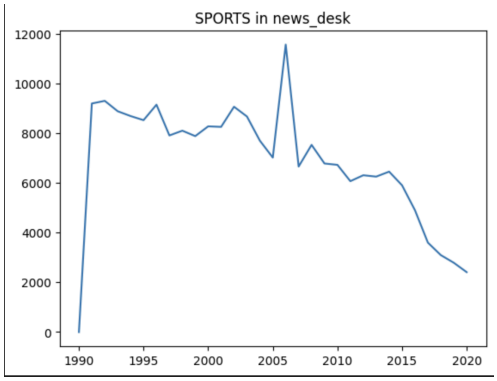
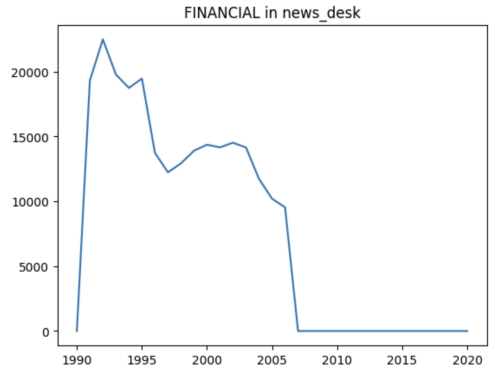
			
3 	3. What do the vector embeddings look like?	JZ	I
4 	4. How have these vector embeddings changed over time?	JZ	I

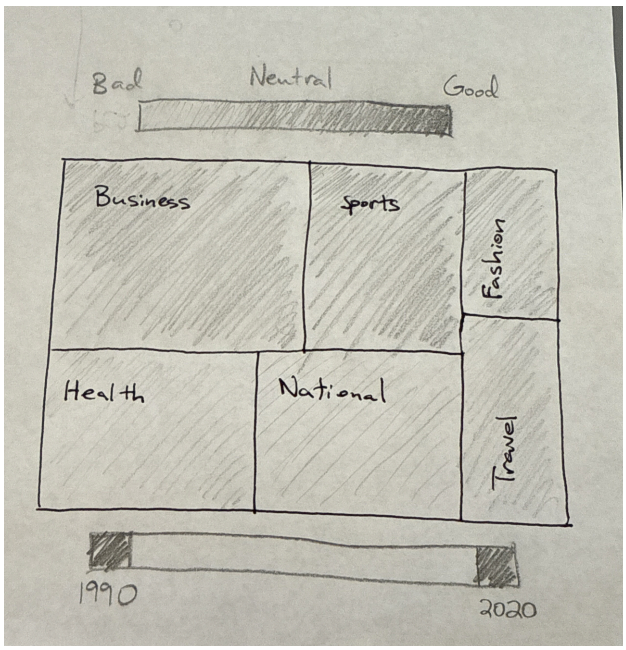
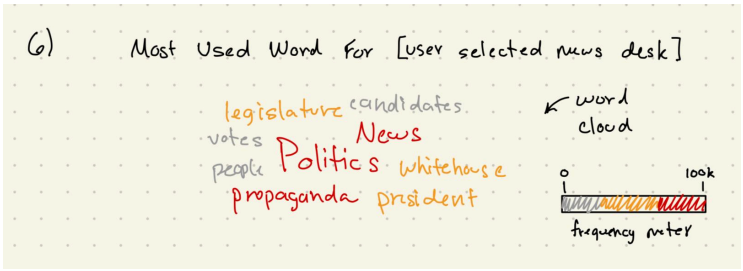
5 	5. How have sentiments in headlines changed over time?	JZ	II
6 	6. What are the most frequently used words in article titles within each category?	EV	I
7 	7. How are specific news desks titles correlated with real world events? (were there more articles from the politics news desks published during major political events?)	EV	I
8	8. Are there seasonal or monthly patterns in news coverage by	EV	

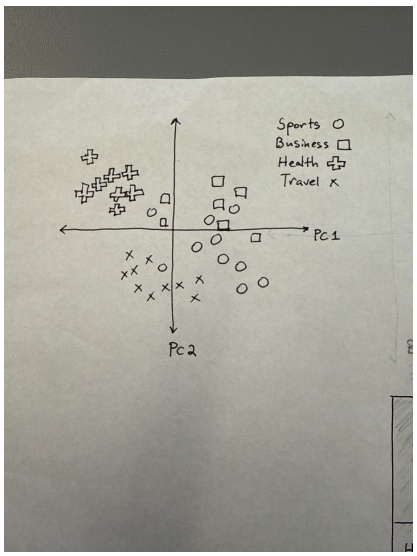
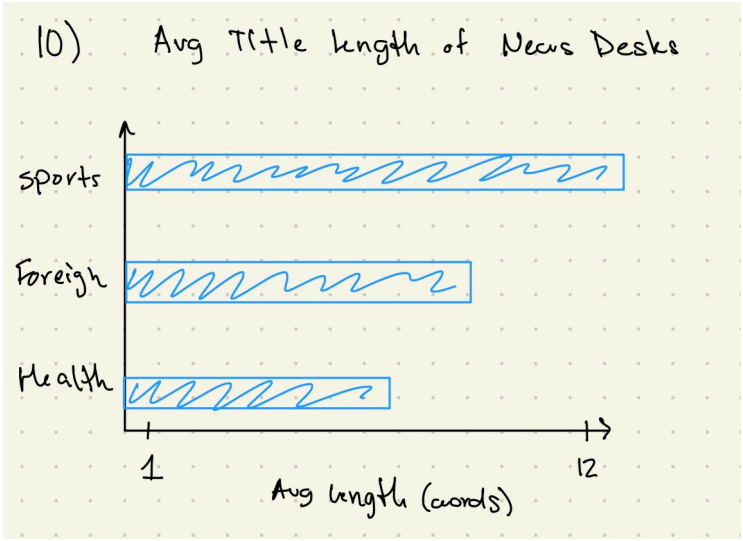
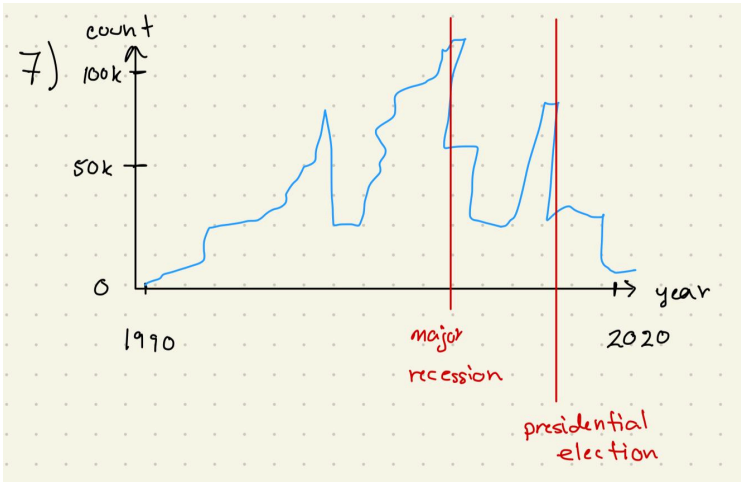
	category?		
9 	9. How does the distribution of categories differ by geographical region?	EV	
10 	10. How do the lengths of article titles vary across different news categories?	EV	I

SELECTED SKETCHES:

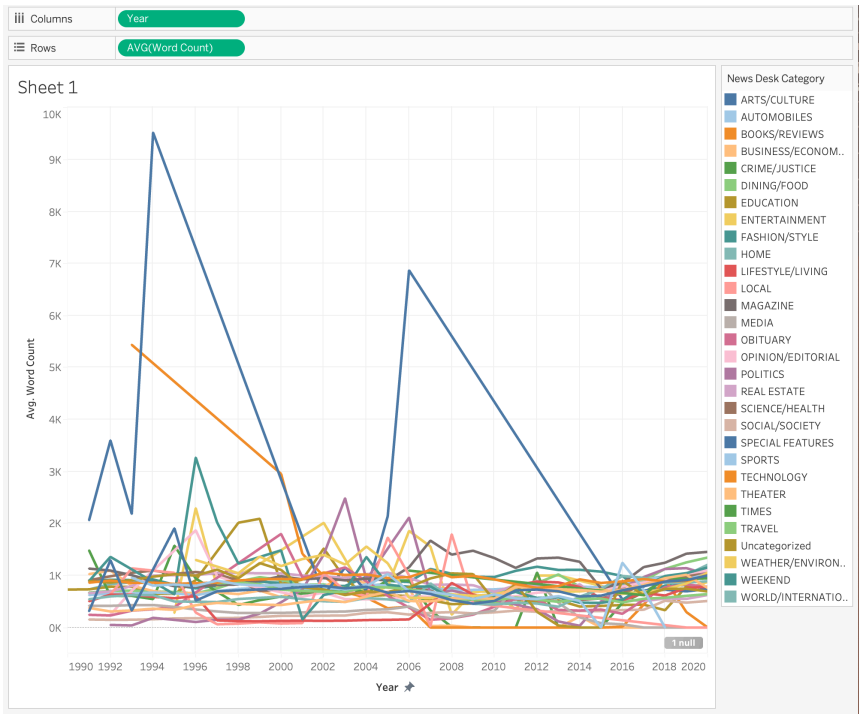
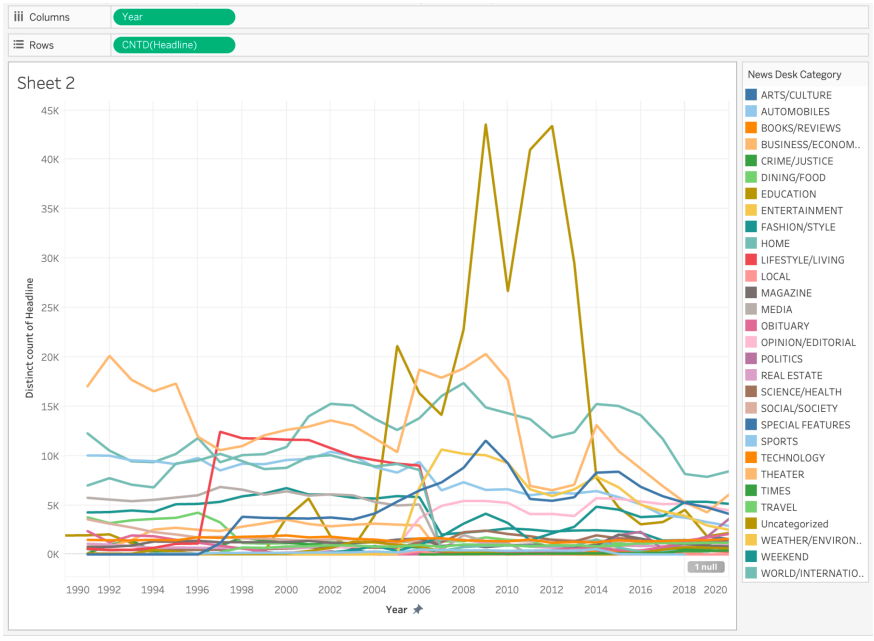
Sketch ID	Question ID	Reason for Selection
1	1: How has the	Visualizing the

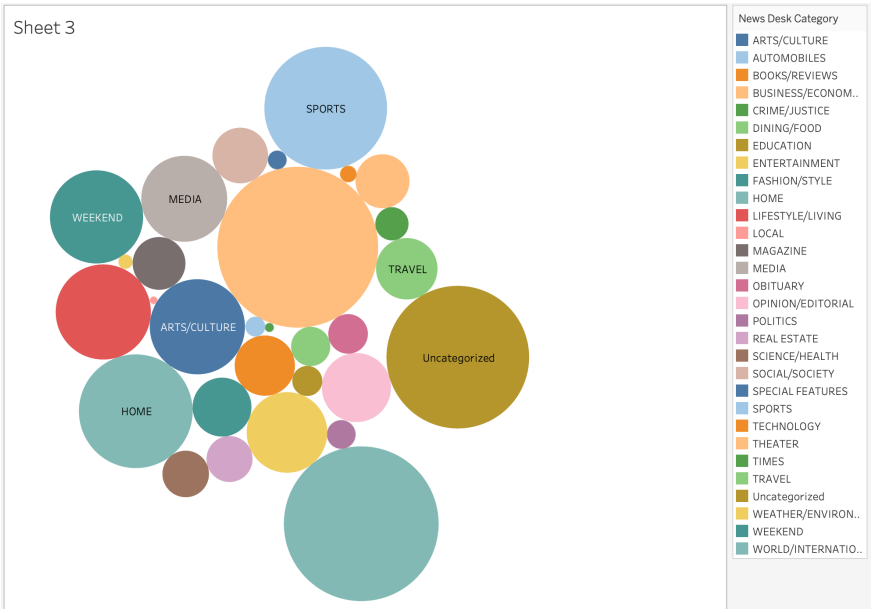
  	number of articles in each category evolved over time? And, have headlines in the same new desk gotten more or less similar to each other? Already, we have some primitive line graphs implemented with Python.	number of articles from each news desk over time is one of the main questions we ask and are working to answer for this project through our data analysis, so it makes sense to have this visualization as a priority.

5 	5. How have sentiments in headlines changed over time?	If time permits, we will create a model that will categorize all titles into different sentiments (bad, neutral, good). It will be very interesting to visualize this as this is not something that can easily be seen from simply looking at a large dataset.
	6. What are the most frequently used words in article titles within each category?	A word cloud for popular words used by news desks will help users easily visualize what words tend to be used the most across different categories. This will make it easy to spot trends across news desks and can lead to very interesting findings.

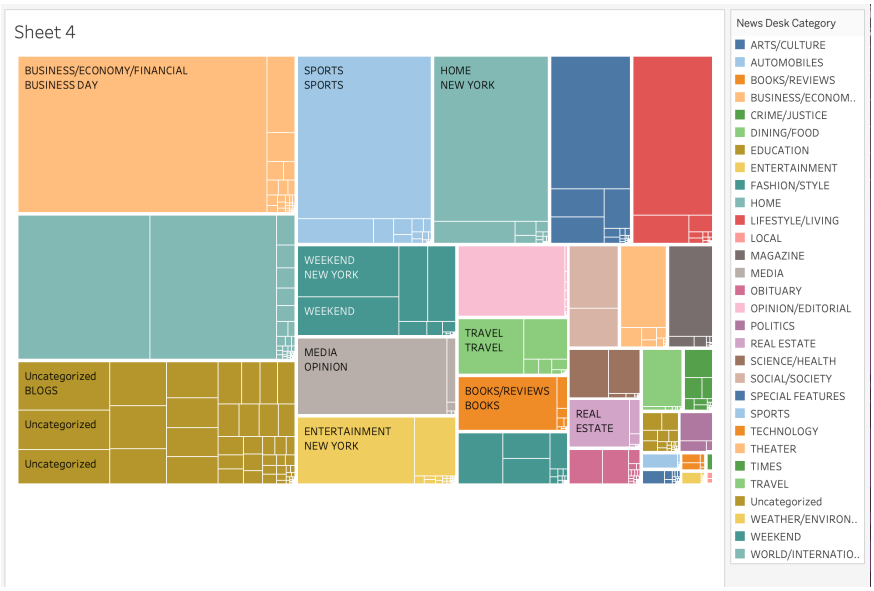
3 	3. What do the vector embeddings look like?	For the more advanced user, this will be an interesting visual to engage with. This visual will provide an easy way to understand how we approached our sentiment analysis.
	10. How do the lengths of article titles vary across different news categories?	NYT Title lengths are very relevant to the engagement of their corresponding articles. Longer titles could lead to less users engaging with an article, so it would be interesting to see how length has been used by different news desks and how it has changed over time.
	7. How are specific news desks titles correlated with real world events? (were there more articles from the politics news desks published during major political events?)	This visual is an important one because it will help explain any peaks or dips in our visualizations. It will be a very interesting visual tool for users who are very curious about why certain patterns were seen in the data.

STORYBOARDING

Insights	Description
	This visual shows the average word count of articles for each news desk category over time from 1990 to 2020. As we can see, Special Features tended to have the highest avg word count than all other categories for many years. If we look at what is happening in the area below, we see that most categories tend to have articles in the 500-1k word range.
	This visual shows the count of headlines over the years for each news desk category.

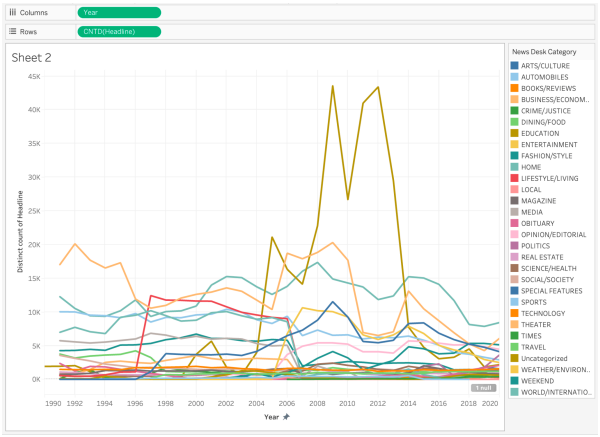


This bubble chart helps visualize the most prominent news desk categories across all years.



This tree map helps us see which sections are more prominent in different news desks categories. The first title for each group is the news desk category and the second title is the section.

Picked Insight: Headline Count Across News Desks Over the Years



Why did we pick it? -
We picked the Headline Count Across News Desk Over the Years because this is ultimately the question we aim to answer for our overall project in conjunction with CS109a. We seek to find trends in news desks across time in terms of the number of articles released from these categories.

So what? -
The analysis reveals shifting editorial priorities, with certain news desk categories showing significant growth in article count over time, indicating an evolving focus on topics that align with societal trends and reader interests.

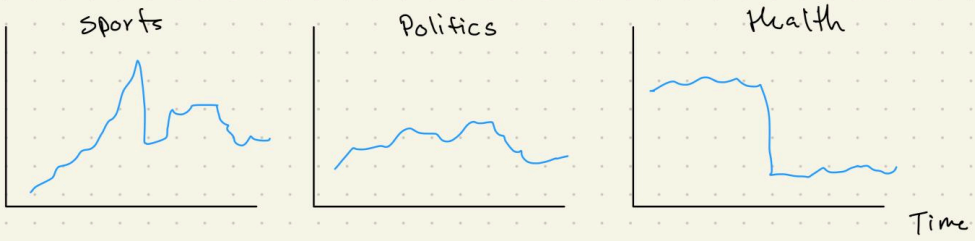
STORYBOARD

1) Hook: Contrast between 2 years with Bubble Chart

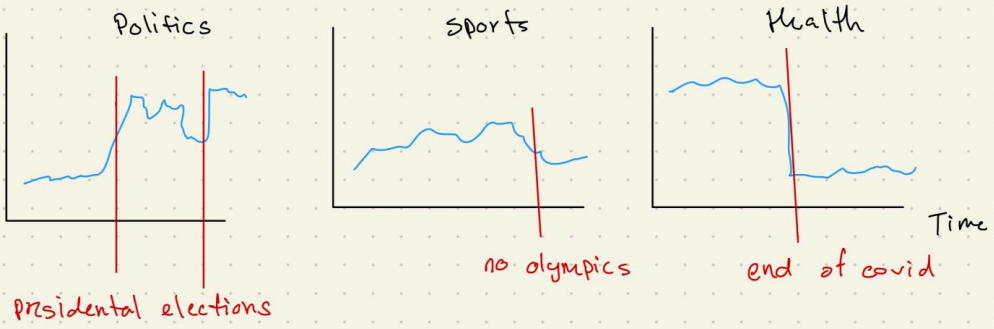
select year: 1990

Most popular desk: sports

2) Rising Insights: line charts showing desk prominance over years



3) Main Mussege: Connecting Desk Frequency and societal change



4) Solution:

section highlighting what newsrooms can do to improve their engagement with their audiences.

Week 9 - Project Map

1. Potential target audiences:

    a. The target audience could be the general American public. This could be narrowed down to those who are more familiar with the Internet and web applications.
    b. The target audience is other students in the general population of Harvard students. These are people who have written their own research projects in their own coursework, peer-read other people's projects, and perhaps are more accustomed to consuming and interpreting academic projects. As a result, they may be more willing to interact and play around with our visualization. If we were to target this audience, perhaps we could write with more formal vocabulary and a more academic tone.
    c. We could also try and specifically target students in COMPSCI 1710 or COMPSCI 1090A. These are students who have also been working on similar projects. If we were to target these students, we could assume a solid background in visualizations and data science. We could use the terms and vocabulary of the two courses and not have to explain what the concepts mean very in depth.
2. It makes the most sense to select the general population of Harvard students as our target audience. They may be the most accustomed to and interested in projects like ours, and we could assume a more academic background, allowing us to dive more rigorously into our findings.
3. Here are some interesting questions that we have come up with. Just to refresh, our project looks at the evolution of New York Times articles over time.
    a. Have the headlines of articles gotten more attention grabbing over the years? In internet terms, have they gotten more "clickbait-y"?
    b. Have the lengths of headlines gotten longer or shorter over time?
    c. How much does the headline actually match the content of the article? Does the headline say one thing, only for the article to read differently?
    d. How has coverage over topics like sports, business, economics, and politics changed over time?
    e. We looked into the data, and one complication is that articles are categorized into news desks with very strange names. For example, there are several different categories that articles about sports may fall into. We could ask: How have these categories changed over time?
    f. If we could access how many copies of an article were sold, or how many times an article was viewed online, we could ask if certain types of articles get more viewer attention or interaction?

g. Since we are processing text data, we have discussed vector embedding of the data with our project advisor. This means encoding seemingly qualitative data like text, pictures and videos, as vectors. We can then do work with dimensionality reduction, projecting these high dimensional vectors into a lower dimensional space. This is something that we as project members have just begun to look into and only understand a lower level. One question we could ask is: How should the audience interpret a scatter plot that has had its vector embeddings reduced to the second dimension?
h. Following the spirit of the previous question, are there any interesting observations we can make about the vector subspace that is all of our vector embeddings?
i. Have headlines of the same news_desk gotten more similar to each other or less similar? In other words, are they getting more or less unique?
j. How does public sentiment change after these headlines and articles are released? Does the change in public sentiment depend on how polarizing or how strongly worded the headline is?

4. We have data that includes the headlines, which news desk it is included in, which section it is in, the link to the article, what year the article was published, and other data. We have cleaned the data. Also, originally we had a different dataset for each year, but we combined all of the datasets into one using Python. Attached is the combined dataset. Most of the categories mentioned, all the ones that look like strings, are actually objects. So technically they are some abstract data type, which makes them categorical variables. This makes sense for columns like news_desk and section. When we were manipulating these points, we had no problem converting them into strings.

Week 8 - Team Agreement & Detailed Project Plan

Basic Info:

- Project title: Evolution of NYT Headlines over the Last Century
- Team Names: Eric Vasquez, Joshua Zhang
- Emails: evasquezreyes@college.harvard.edu, joshua_zhang@college.harvard.edu

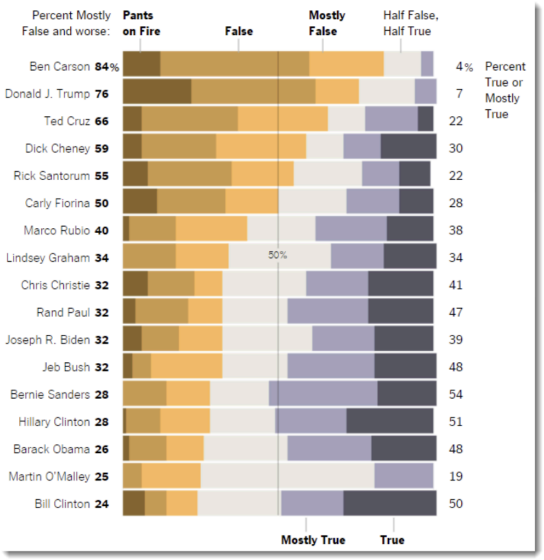
Background and Motivation

- We are both taking CS 1090a concurrently so we decided to work on a joint final project between 1090a and 171. Because of this, our 171 project is based on the 1090a project we chose, which looks into the evolution of NYT headlines over the

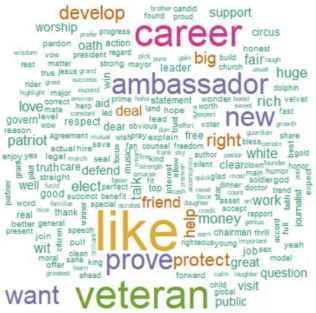
last century. We chose this project because we believed it was one of the few 1090a projects that will allow us to create a great visual after processing all the data.

Related Work

- Most of our inspiration came from our familiarity with the New York Times as a media company and how much work they put into their titles to capture audience attention. Before this project we didn't think much about the change of the titles over time, but given how often we come across these articles in our email boxes and websites we felt inspired to dedicate our final project to researching these changes.
- As for the technical side of the project, both the programming aspect of sentiment analysis and natural language processing as well as the visualization of our results were fascinating.



The tableau blog website itself (https://www.tableau.com/blog/how-visualize-sentiment-and-inclination-48534) gives us ideas on how we could visualize results of sentiment analysis, particularly categories that follow a set order, like the example above that goes from Pants on Fire to False to Half True to Mostly True. In practice, this could be implemented with regression for such a classification problem.



Bag of Positive Sentiments

Since our focus on titles of news pages has a strong emphasis on keywords, Medium.com (https://medium.com/analytics-vidhya/different-ways-of-visualizing-twitter-sentiments-analysis-in-r-270d5d459603) suggests ideas like a word cloud for certain words that show up in headlines and their frequencies. Such visualizations are easy to interpret and playful.

Data

- We were provided with two large datasets for our project within the 1090a project proposal. Those datasets can be found here:
  1. “Three Decades of New York Times Headlines”
     - https://www.kaggle.com/datasets/johnbandy/new-york-times-headlines
  2. ”New York Times Articles 1920-2020”
     - https://www.kaggle.com/datasets/tumanovalexander/nyt-articles-data
- The entire project proposal from 1090a can be found here:
  - https://canvas.harvard.edu/files/20807443/download?download_frd=1&verifier=QBRDza3Fe51L0AGCl9BDts8AFltmvML5mL0dBw36