0.0.1 Question 1d

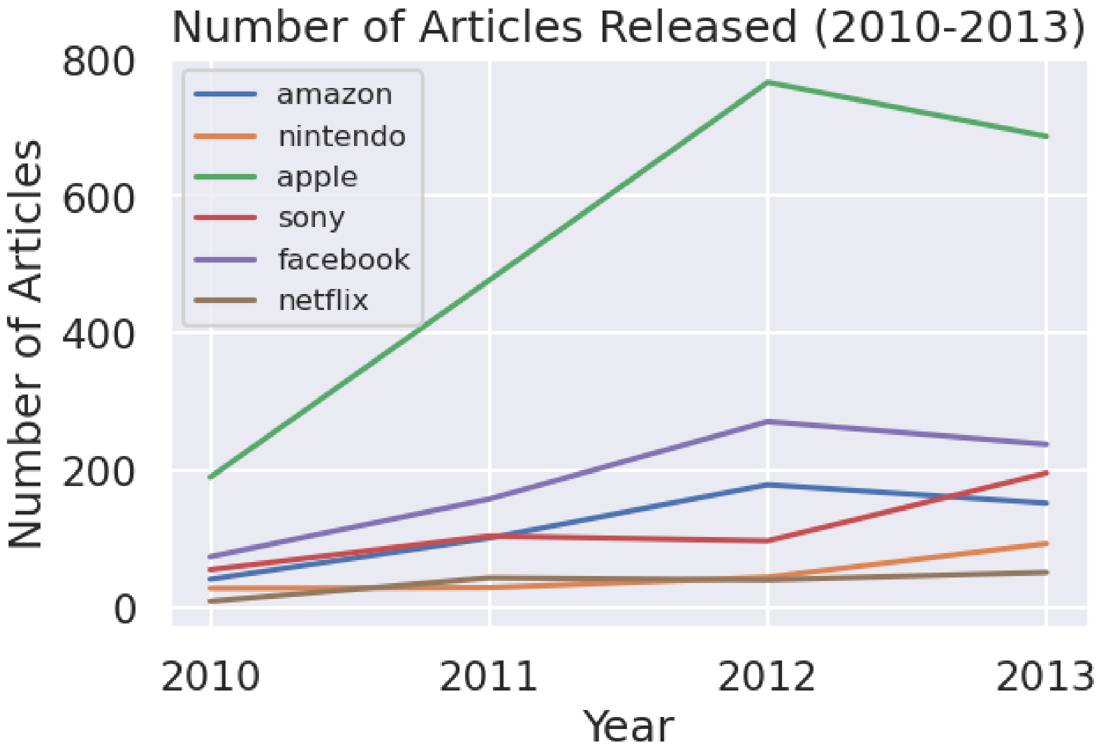
Suppose we are interested in using the news to predict future stock values. What additional data would we need to predict stock prices, and how could we connect that data to news articles? In addition, what attributes or characteristics of the news might help predict the stock value?

To assist in predicting future stock values it would be helpful to merge columns that join the corresponding company's stock ticker based on the time under 'released_at'. Additionally I would parse through the title and content to identify key words that may allude to an increase in stock valuation of a company such as "jumps" or "surges" (wording used in first entry columns).

Question 2d, Part ii Given your code in the previous part is correct, after running the cell below, you should be able to see the number of articles released mentioning `companies` for each year. The plot should look like this:

```
In [20]: plt.figure(figsize=(6, 4))

         for company in companies:
             sns.lineplot(data=year_news.reset_index(),
                          x="Year",
                          y=company,
                          label=company)
         plt.legend(fontsize="12")
         plt.xticks(np.arange(2010, 2014), np.arange(2010, 2014))
         plt.ylabel("Number of Articles")
         plt.xlabel("Year")
         plt.title("Number of Articles Released (2010-2013)");
```



What trends do you notice in the plot above? Feel free to reference or Google any events to explain the trends seen in the graph. What are some limitations of using data and the corresponding plot to analyze the performance of different companies or trends?

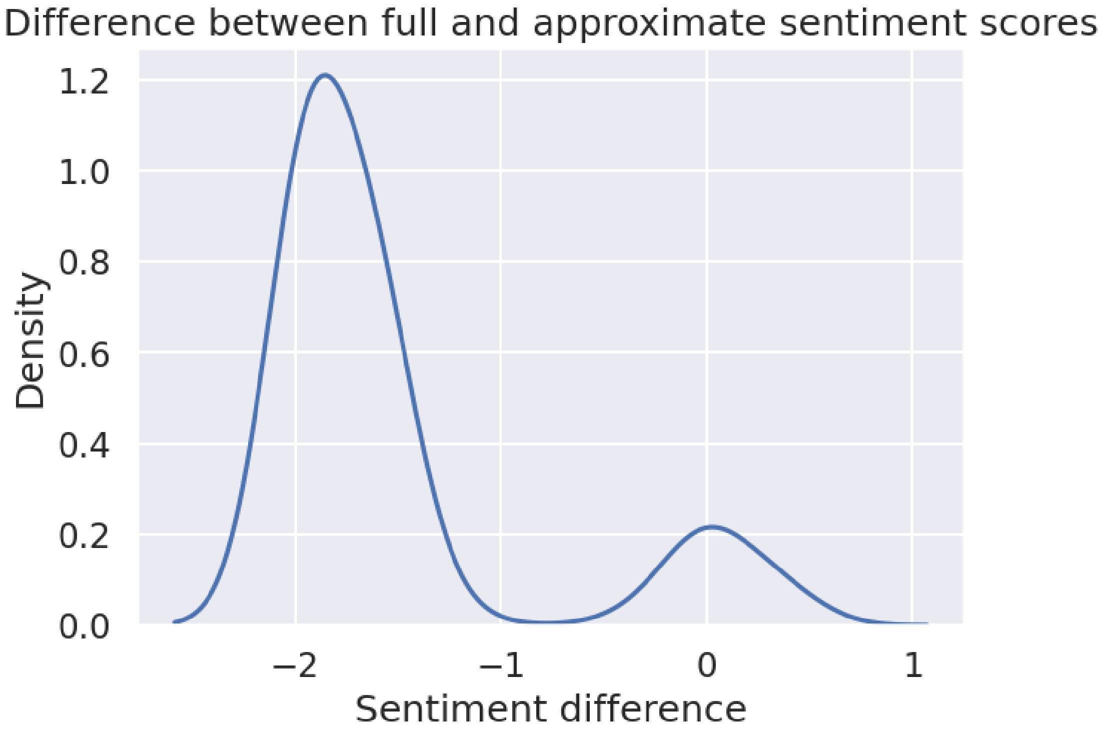
Hint: Remember the source of the articles and the subset of the articles we are analyzing in this assignment.

According to the plot, Apple received the most mentions within bloomberg articles from 2010 - 2013 relative to the other major companies by a significant margin. Pointing to the peak in 2012, there can be several factors and events in Apple's timeline that may have influenced this and caused media coverage. Such events includes the loss in life of Steve Jobs in 2011, launch of the iPhone 5, trial case versus the rival Samsung and morality issues surrounding the outsourcing to China. As with any tragic loss of a public figure representing a major company or a pivotal court case, you can expect headlines surrounding these topics from major journals.

Meanwhile, other companies tend to experience steady and consistent mentions throughout the plot. The data does not capture a company's relative performance as these mentions in the articles can have a binary positive or negative story. Thus, the data without further manipulation cannot entail the performance of different companies as it is restricted to capturing the bloomberg coverage which can be biased based on the business model of Bloomberg and who they choose to report on.

Question 3c, Part ii Below we have provided a plot looking at these differences. Comment on why we see differences when calculating the sentiment of an article as the sentiment of the first sentence mentioning "microsoft" or "msft" in the article versus the sentiment of the entire article itself. How does context play a role when evaluating the sentiment of a text?

```
In [165]: sns.kdeplot(msft_scores_2010['sentiment_difference'])
          plt.xlabel('Sentiment difference')
          plt.title('Difference between full and approximate sentiment scores');
```



Since we were strictly observing the first sentence of the article's content, the sentiment of the entire article and its stance was unclear. Oftentimes, articles present counterarguments that can steer the sentiment detection model towards the polar opposite or an author can simply be warming up into his main argument in revealing a negative sentiment despite the first sentence of the respective article expressing neutral sentiment. Therefore, it is extremeley important to gain context of the article and assess its sentiment upon full review to prevent misleading computational results. In the plot, we observe the density plot peaking at the difference of -2 with a density of >1, highlighting the differences amongst the full article's sentimenet against the first sentence's sentiment.