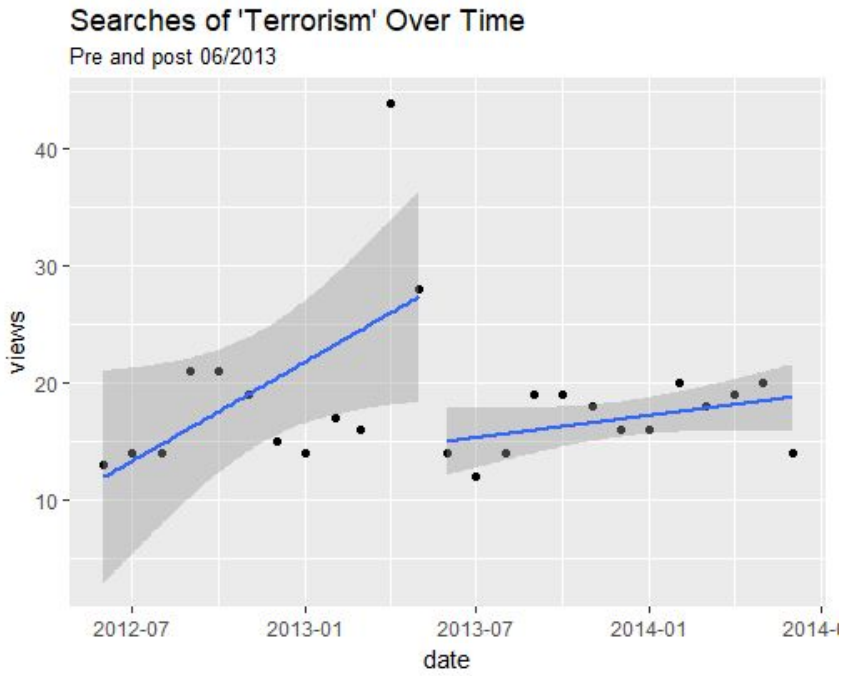
Exercise 2

Social Data Analytics 501

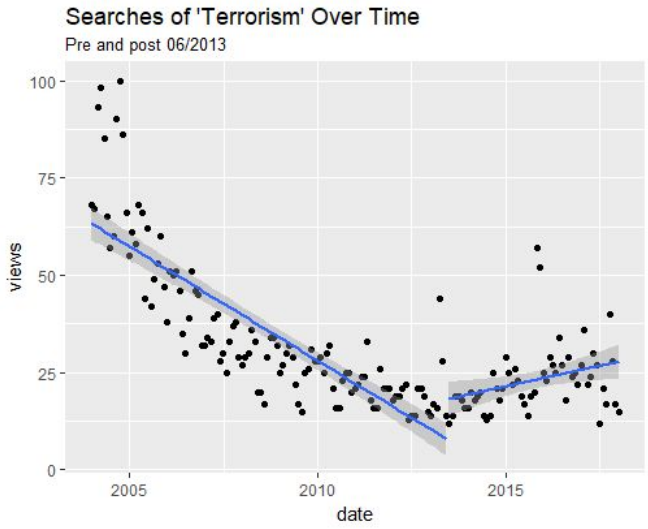
Claire Kelling

1) Try to replicate the Penney result in Google Trends data. Is it there, too? Speculate: why or why not? (https://trends.google.com — you can download data in spreadsheet form)



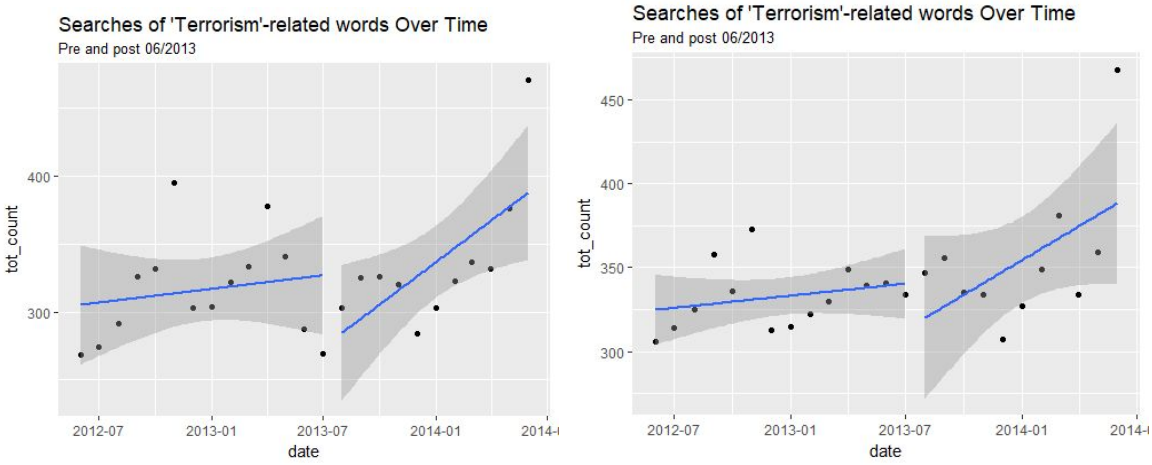
We were able to somewhat replicate this finding of a decrease in searching of “terrorism” over time, though it is not as convincing as the Penney article. There are a couple of high values for views right around the time of June 2013, which drastically impacted the line fit. I think otherwise, the points seem to pretty constant throughout the time. Also, there is a slight increase in the points after June 2013 according to this figure.

If you look at the full figure with all the data available, you will see a general decrease before June 2013 and a slight increase after June 2013. Therefore, I would say that this trend is even less convincing, given this data. This figure is included below.



The reason that Google trends may not have picked up on this trend is perhaps after July 2013 people were searching for the term, they just weren't clicking on the page. Also, perhaps because the Snowden controversy was specifically related to Wikipedia, people were staying away from Wikipedia articles.

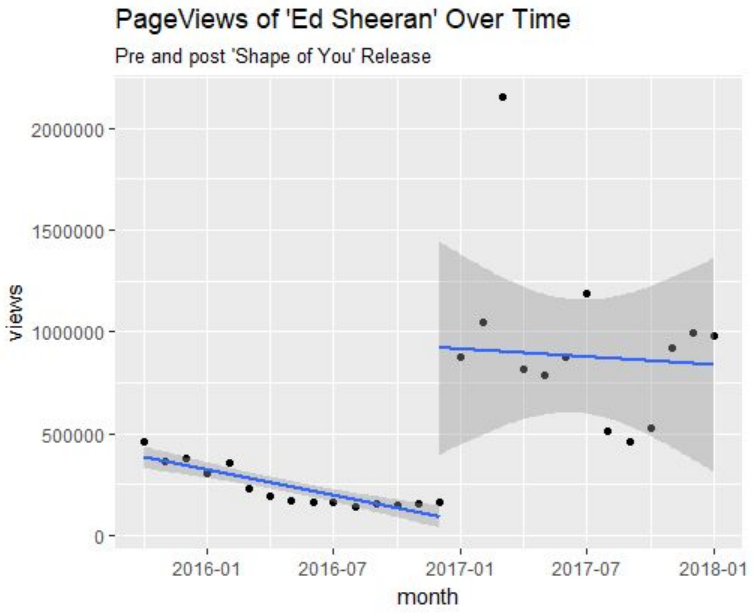
The above analysis includes only the word "terrorism." Using the Python package pytrends, we replicated this analysis for the full 48 words that Penney used in the paper. We find that the trend is even less convincing here, seen in the figure below. There is a small drop around June 2013 but it definitely does not stay low. We find a similar trend when we search the entire globe, suggesting that there was no chilling effect in the United States.



Just US World

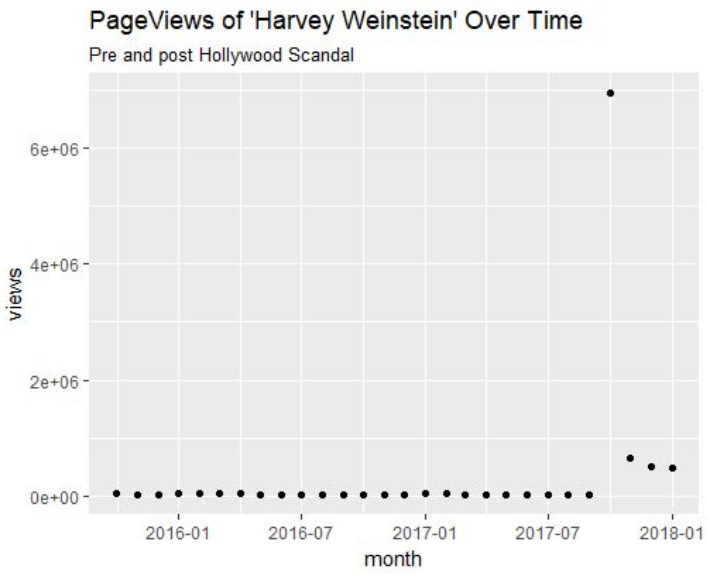
2) Use Penney's interrupted time series design, or something like it, to demonstrate a convincing shift up in Wikipedia page views in some topic as a result of an event (this should be easy). This will have to be based on data after July 2015. (I have confirmed that the R package "pageviews" works, as does the API at

wikimedia.org/api/rest_v1)
For our purposes, we used Ed Sheeran, because he is newly popular. After his songs “Shape of You” and “Castle on the Hill” were released in January 2017, there were significantly more views of his page on Wikipedia.



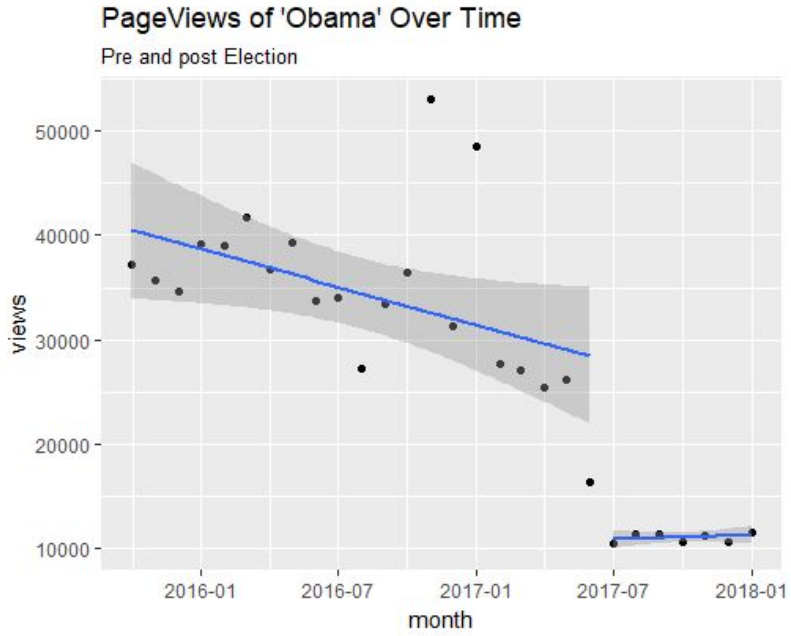
However, we found this quite difficult to do. In most instances, there would be a peak for a month, and then it would decrease after that. So, it did not follow the form of regression discontinuity, but rather something along the lines of an outlier.

Below, we have an example of such a page, the page on “Harvey Weinstein”, where there was a peak after his scandal, but in general, there is a decrease down to the original level of interest after the news broke. This intuitively makes sense, because after most current events, people will forget relatively quickly. Ed Sheeran’s page has consistently high views because he is a new trend. We thought there might be a similar trend with the something like “fidget spinners” but this page didn’t even exist before a time. That is why it is important for pages like Ed Sheeran’s to have a history before they become a trending page.



3) Can you find and demonstrate a similar shift down as the result of an event as Penney does? Why is this harder?

We actually did not have to search that long to find a shift down as the result of an end. This is mainly because we were looking for a regression discontinuity above, rather than an outlier. The main page that we were able that demonstrated the “shift down” was the page on Obama. We found that after the election and inauguration of President Trump, people did not search for Obama anymore. It appears that folks did not care about Obama as much after Trump was in office for a couple months.



This might be harder to find a shift down, in concept, because there aren’t usually events that result in a shift down, it is usually a more gradual decrease.