

I found a dataset which list the Top 500 news outlets top 10K stories from Nov 16 to May 17.

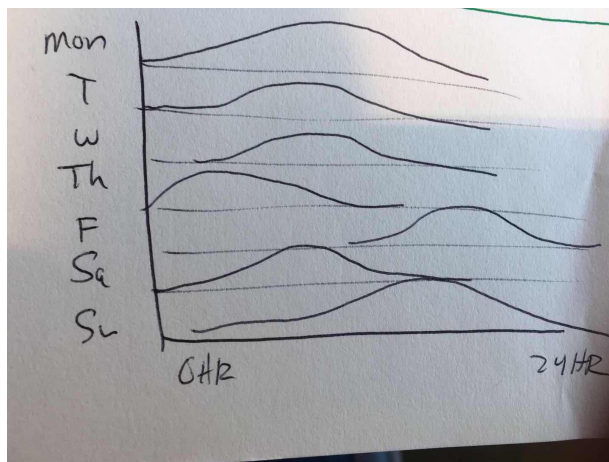
It contains 10001 rows and looks like this:

Publisher	Headline	Link	Date	Keywords	Excerpt	Facebook	FB_Overperform	Twitter	TW_Overperform	LinkedIn	hasVideo
worldnewsdailyreport.co	Woman arreste	https://w	3/31/17 15:11	Crime, Mental Illnes	45-year old Janic	1906460	13.99663102	2780	1.298582122	12	FALSE
buzzfeed.com	Only People W	https://w	3/6/17 11:14	Only People With P	Do not adjust yo	1250909	164.8301673	3630	3.178659062	245	FALSE
nationalreview.com	Trumps Execut	http://ww	1/28/17 17:32	David French, Politi	To read the online	1191089		0	28090	0	FALSE
huffingtonpost.com	Instead Of Trun	http://ww	12/19/16 11:39	What's Working,Clir	It would create jo	987710		0	4260	0	FALSE
burrardstreetjournal.com	Female Legisla	http://ww	1/24/17 2:42	nan	WASHINGTON (965871		0	5250	0	FALSE
worldnewsdailyreport.co	Babysitter trans	http://wor	5/3/17 18:24	nan	Cincinnati Ohio	893711	6.423431302	730	0.109616803	308	FALSE
viralnuashot.com	Anorv Woman	http://vira	2/15/17 0:26	null	A woman in Scott	835820		0	600	0	FALSE

There are some thoughts for my final projects:

1) The stories with Videos are more likely to do share? Or less?

2)what time of day is the most frequent time publishers send out the news? Which day during the week?



3)Not many people use LinkedIn to read the news, what kind of stories they are sharing on LinkedIn?

4)Go through all of the keywords to see the top 30 keywords are most popular. (Then present it as word cluster analysis. Like this:

