Your Whatsapp chats can tell a lot about you

Visualising WhatsApp chats using Python

Lately, I was looking for some small and exciting visualisation projects to explore the data visualisation field. Then I came across a feature in WhatsApp about exporting your chats in a text file, which is quite handy and easy to work with. Whatsapp claims that nearly 100 billion WhatsApp messages sent per day, or 69 million per minute in 2021.

I started using WhatsApp frequently after getting into college in 2017, so I thought of collecting and visualising my last four years of chats. I obtained around 62 text files from my WhatsApp, including personal conversations with friends and family members, and some group chats.

For people who aren't interested in the code, you can enjoy the images. For others, I have also uploaded the entire code in my **GitHub** repository**.** I used **Google Colaboratory** for this project, in case you are using this or some other platform then change the path of the files accordingly.

Loading the Messages

The messages in the text file are of the format - {Date}, {Time} — {Author}: {Message}

```
08/07/2018, 11:32 - Sourav @ NIT Dgp | Civil: Kal se class
```

The plain text files will have to be converted in a meaningful manner to store in a Pandas data frame. I saved all my conversations in a data folder so I could list, load, and merge them into one dataframe. Now our dataframe is ready and looks something like this.

	index	date	name	msg	msg_len	date1
29116	135	2021-10-07 16:14:00	Sagnik @ TCS	Sick leave could be an option	29	2021-10-07
29117	136	2021-10-07 16:14:00	Subhash Dixit	Do we have to provide documents for that?	41	2021-10-07

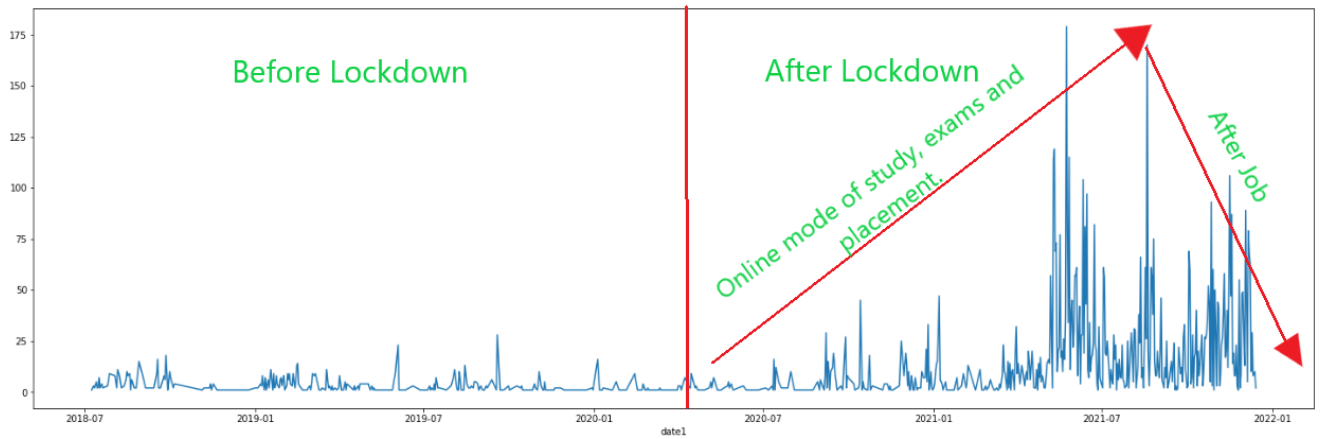
Some Statistics

How many messages have I sent in the last four years? How many different people have I talked to within the past four years?

In my case, I sent over 8188 messages, talked to over 412 different people.

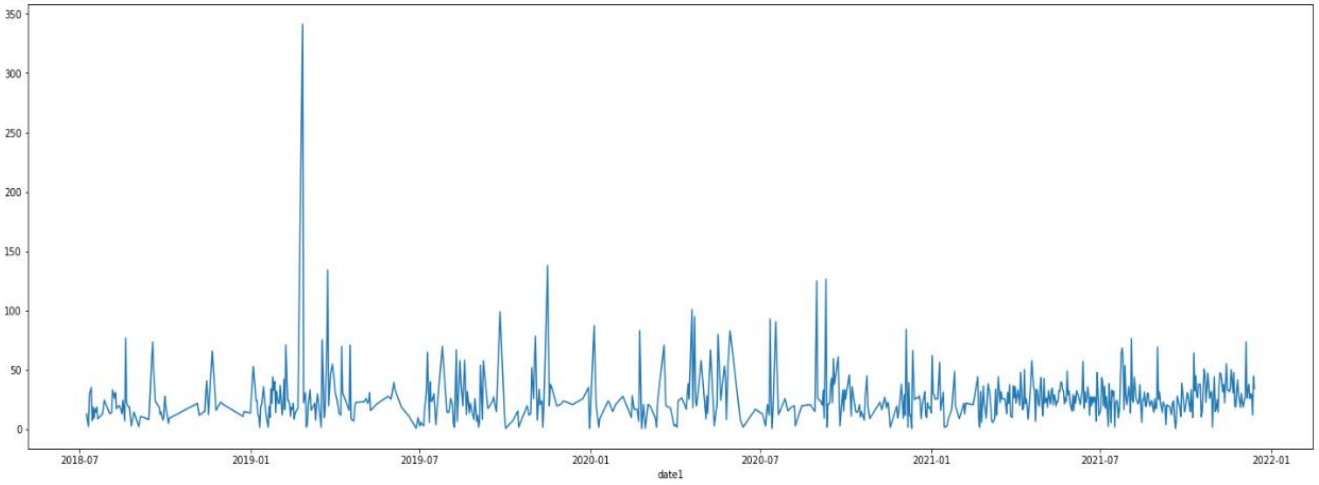
Data Exploration

This is the most exciting part of this article - Data Exploration. Let's dig out all the fascinating stories that these data are trying to tell us.



This plot is pretty impressive as it can very quickly identify the changes after and before lockdown in my texting pattern. The effect of coronavirus on my texting pattern can be determined very quickly (I guess everyone is going through the same situation). Apart from this, some peaks in the plots over a few months (May to August after lockdown) can be justified by Online examination and training period of job.

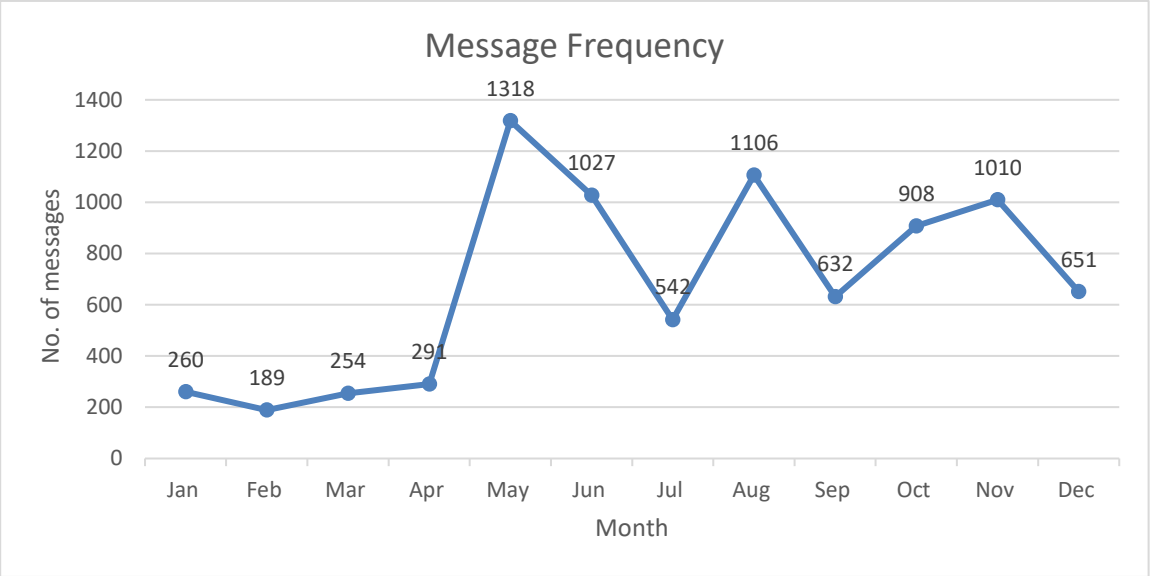
I also find it funny to see the length of my messages over the years.



So I tried to find that outlier, and I got this -

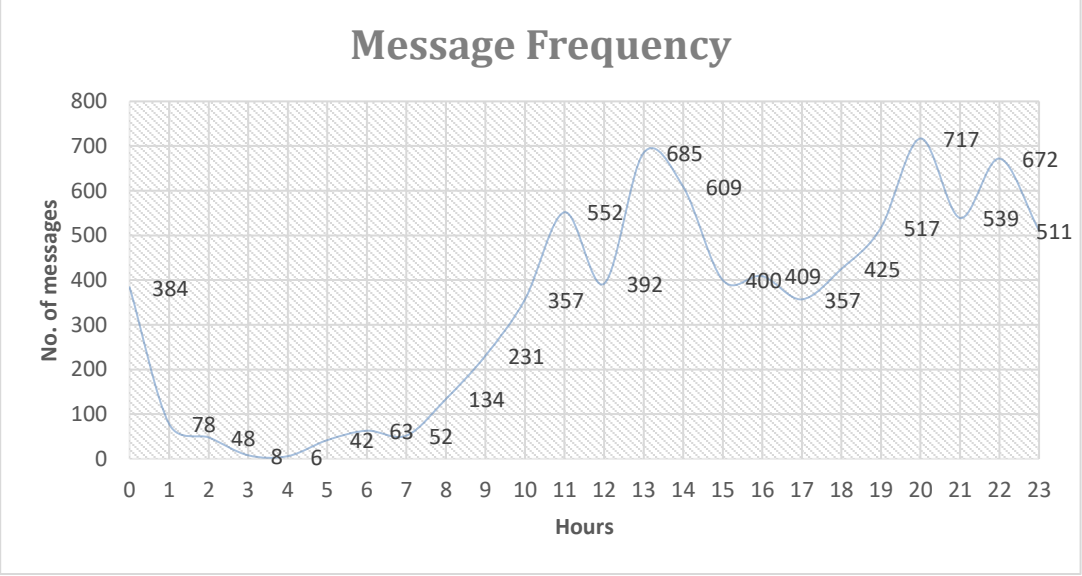
array(['Mar har CHiz Mai ready Hu...Aur Jo sab Karega wahi Mera decision hai....but remember one thing Humlog jyada SE jyada Director ke pass ja sakte hai....But yeh sab log sale ek Hei khet ke muli hai....Nahi sunega koi hamara bat.....chahe Humlog Kuch bhi Kar le....yehi reality hai Aur Jo accept Kar lega Uske liye Sahi Hoga.....SalA 3 Saal SE Kam try thodi Hua hai aajtak Kuch CHiz Nahi hua Humlog ke man SE....yeh log class Nahi chorte hai subject change Karna Toh Bahut Bari bat hai....Josh Josh Mai bas jyada Kuch na ho Jaye ....Nahi Toh consequences Bahut Bura ho Sakta hai.... atleast Sahi SE degree mil Jaye taki bad Mai kuch ho paye.... agar 30 log awaj uthayengee including me Toh Mai Sath Hu....majority Hona Chahiye....Nahi Toh Kuch log Ko yeh log ache se subject change Kara dega.....abhi SE count karte Hai....30 SE jyada hai ya Nahi majority Mai....'],
dtype=object)

After this, I plotted my Message frequency over the months.



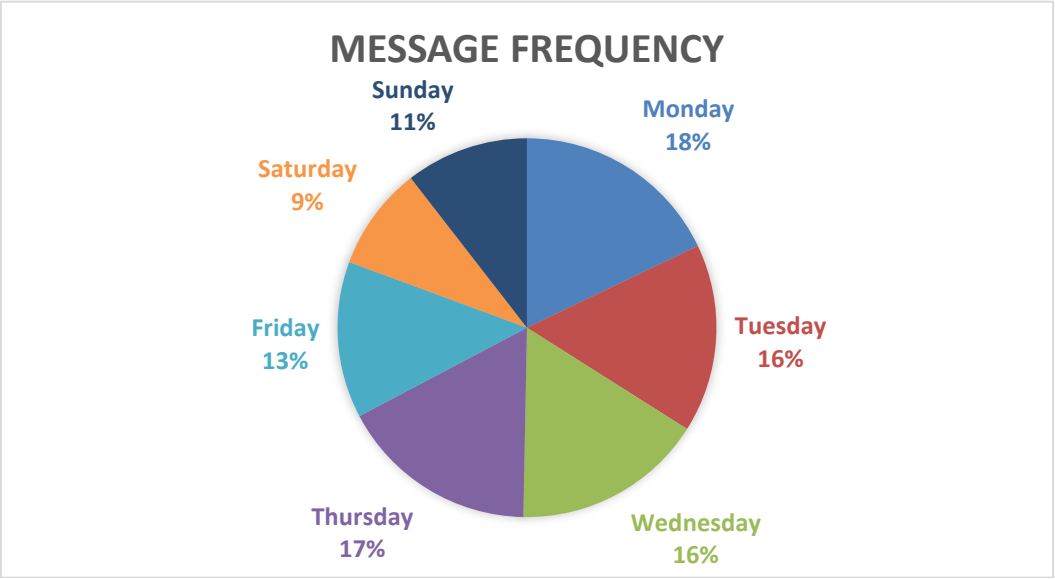
I was amazed to see this distribution. I guess this is also suggesting that I chat a lot in peak of summers .

Next plot is fascinating and tells about your sleeping time.



It suggests that I am not a Night-Owl and prefers to sleep after 11 or midnight. Also, the number of messages increases from morning to noon. It is contrary to the fact that during these hours one should work or study more .

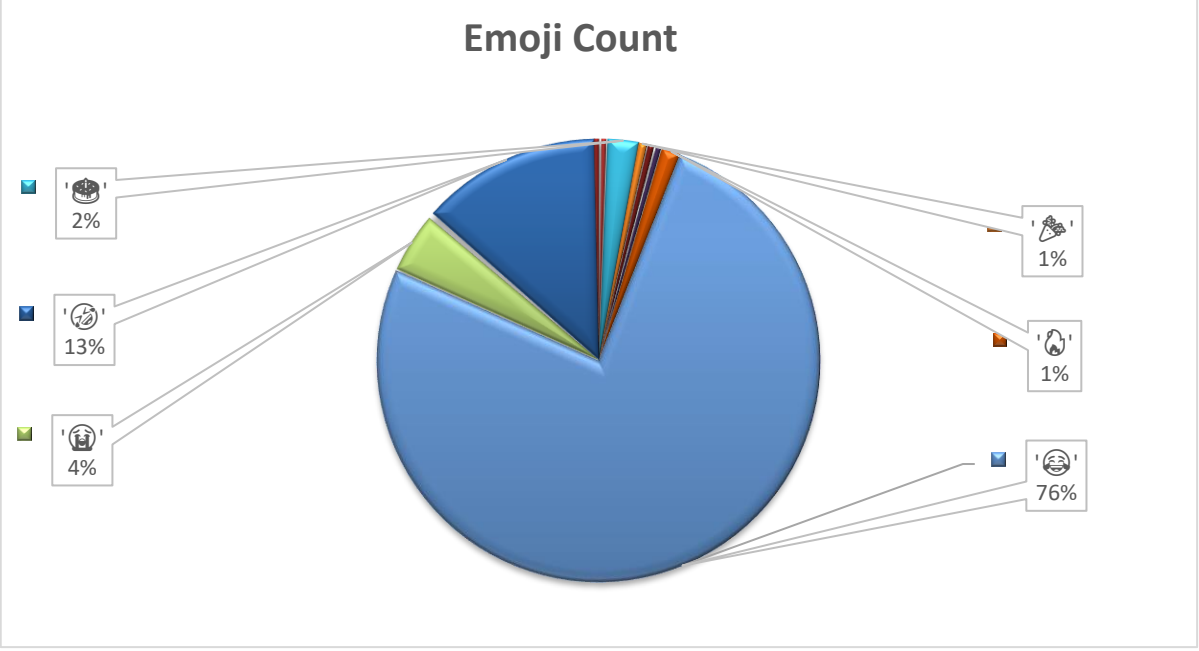
Next, I also plotted the frequency over the days in a week. Well, it looks like, I chat more in the weekdays as compared to the weekend. This also suggest that I play a lot in the weekend because of outdoor activities I don't get time to use whatsapp.



Now comes the most interesting plot of this article.

Let's Plot Emojis

The complete code for emoji is in my GitHub. I used Emoji library by Python to plot this. You can install this library using this line - *pip install emoji --upgrade*



Well, this plot suggest that I am not huge fan of emoji. I only use few emojis. Among that percentage of 🤣 and 😂 are 89 %. I was glad to see that evidently, my life is full of laughter. This also suggest that I use many jokes during the conversation.

Conclusion

It looks like this analysis while answering some questions has opened up a lot of new problems which can be further solved. Stay tuned for the next part of this article where I will try to use the concepts of NLP like Topic Modelling.

Did you find any of these insights useful? Or do you have suggestions about some valuable insights that I missed? Feel free to add your comments below.