Monday, 12 July, 2021

WHATSAPP GROUP CHAT ANALYSIS

Summary Report by mellchii

Project 1

Introduction

WhatsApp Messenger is a cross-platform messaging and voice-over-IP (VoIP) service that allows users to send text messages and voice messages, make voice and video calls, and share images, documents, user locations, and other content.

The WhatsApp application runs on mobile devices but is also accessible from desktop computers, as long as the user's mobile device remains connected to the Internet while they use the desktop app.

In this project, the usage of WhatsApp in a Group Chat made up of colleagues from the same workspace in analyzed. The intent is to carry out an exploratory data analysis on the interactions within the group in order to discover patterns, test hypothesis, and summarize its main characteristics using basic visualization methods.

A couple of questions are also asked of the data as well, and no doubt, it sure gave us some interesting answers. 😎😚

Exploratory Data Analysis

Background Information

Data Overview, Scope, Boundary, Approach

The data source of the analysis is a download of the Group's WhatsApp Chat record. WhatsApp has a functionality that enables a download of the conversation logs of individual and group chats. To do this, just select any conversation, click on the dots on the top right, click on '**More**' then '**Export chat**'.

For this project, we have exported the logs "**without media**" files. This generates a **.txt** text file which is a time-ordered list of events that occurs within the chat mostly made up of text messages. Each line is a single message and is in the following format:

date, time - sender: message

The exported data for this project covers a period that ranges from:

30/06/2020, 09:21 to 30/06/2021, 17:04

And does not include an analysis of the usage of multimedia files such as audio, images, voice notes, video, etc sent within the group.

Our approach first involves importing all the necessary python libraries into the Jupyter notebook, that will be used for the analysis.

We then define a python function that takes in the raw file (exported WhatsApp conversation) as it's input and this file is parsed through a python regex function in order to convert it into a pandas dataframe, carrying the right type and format for the data in each column.

Some Questions Asked:

- Most active users in the group?
- Average number of messages per day?
- Most active time of day in chat room?
- Most active months of conversations?
- The emoji y'all used the most or the least?
- Emoji usage by specific group members?
- A word cloud of the commonly used words?
- Distinct conversations within the group?
- Who starts the most conversations?
- Weekday versus Weekend usage pattern?
- …

Total messages

7,807

This represents the total number of individual messages (including messages comprised only of emoji's but excluding multimedia messages) sent by the group within the period analyzed

Total Words

72,922

The group has sent a total of 72,922 individual words (including emoji's) within the period. No thanks to Toba's and Lulu's updates from Gov. Wike's media broadcasts. 😑😏

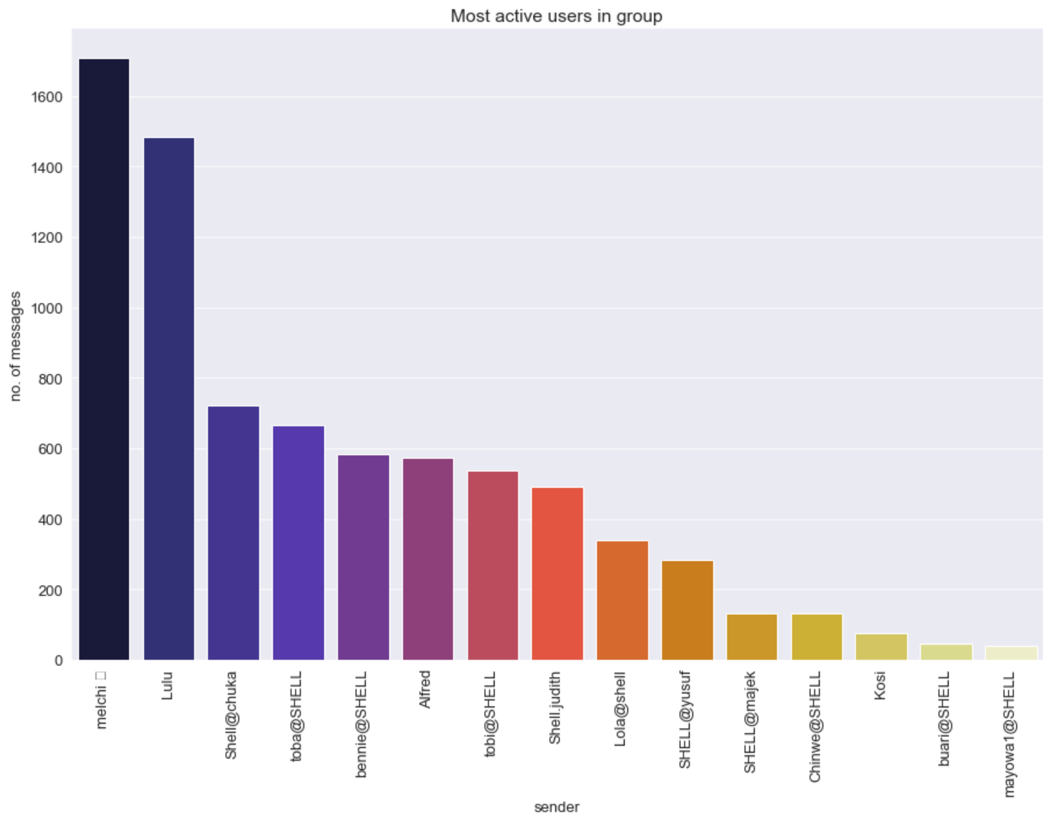
No. of Participants

15

Of course, the group is made up of 15 talkatives. We'll soon get to see the major culprits.

Insights and Visualizations

Let’s hear what the data is saying



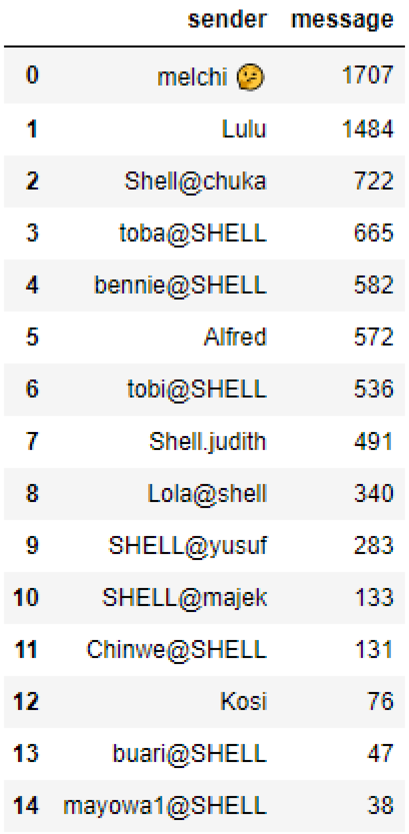
Viz 1: Most Active Users in Chat Group

For a period covering 2020-06-30 09:21:00 to 2021-06-30 17:04:00, which represents **365 days (exactly 1 year) of activity** in the WhatsApp group, a total of 7,807 individual text and emoji--based messages were sent. This turned out a total of 72,922 words across all 15 participants in the group.

Now see the culprits! Not surprised to see Lulu’s name topping charts, since we can all assume not to know who melchi🤔 is. Lulu sent 1,484 messages within the period; and that has not taken into account multimedia messages by the way.

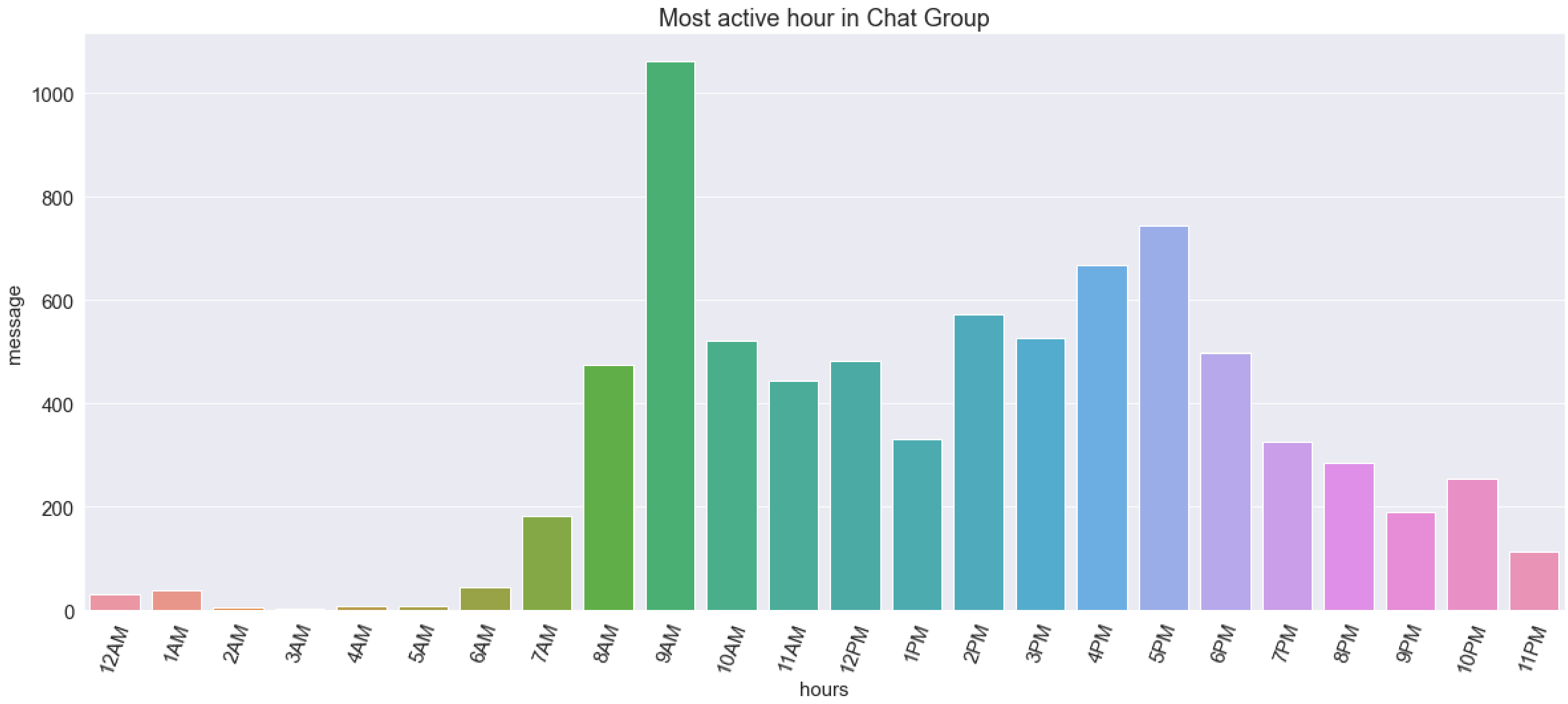
At the bottom of the chart is Madam Owoni, clocking at just 38 messages. Majority of other participants hover around 491 messages (Judith) to 722 messages (Major).

A simple division of total number of messages sent by 365 days means that the group is sending out **an average of 21 messages per day.** We take a further dive into this result in order to separate the weekdays from the weekends, as well as an analysis on time of day in which most of the messages are sent.



	sender	message
0	melchi 🤔	1707
1	Lulu	1484
2	Shell@chuka	722
3	toba@SHELL	665
4	bennie@SHELL	582
5	Alfred	572
6	tobi@SHELL	536
7	Shell.judith	491
8	Lola@shell	340
9	SHELL@yusuf	283
10	SHELL@majek	133
11	Chinwe@SHELL	131
12	Kosi	76
13	buari@SHELL	47
14	mayowa1@SHELL	38

Viz 2: Sender Vs Total Messages Sent



Viz 3: Most Active Hour in Chat Group

Now, I look at this and I wonder who are those people sending messages by 12am, 1am 🙄. Could be the vampires within the group? Or the diasporians in a different timezone? Unfortunately, our analysis did not unravel this mystery yet. The group activity seem to start at about 6am, peaking at 9am and 5am, and then winding down at about 10pm.

We could infer that the peak at 9am could be related to the times when the usual Morning Operations call is just finished and y’all come to the group to rant, ‘shook’ pin, and gossip about your Oga’s. I have not inference for the 5pm peak…for now. What you thinking?

Insights and Visualizations

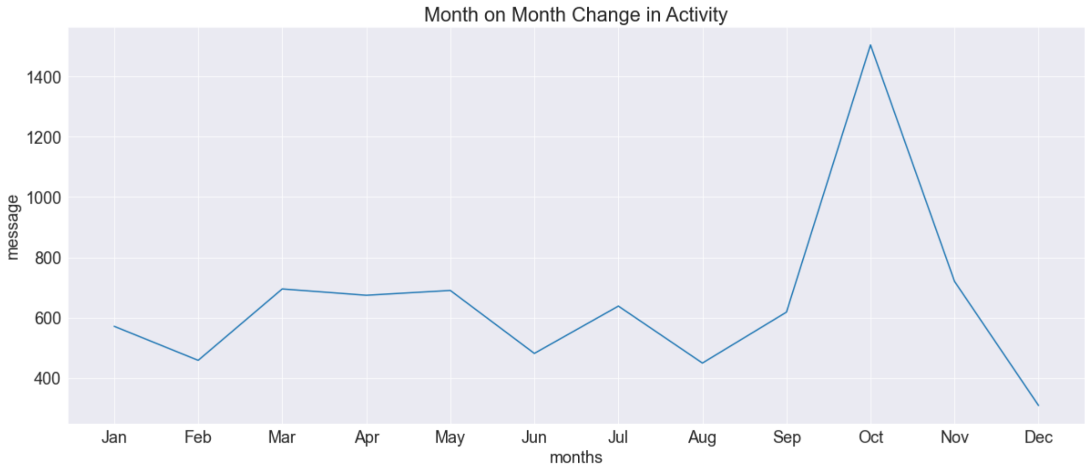
Let's hear what the data is saying

In the month of October 2020, the group registered a significant increase in conversations, more than doubling its average of about 700 messages per month to peak at 1500 messages in October 2020. Being the month of independence, was someone complaining too much about the politics of the nation? 😥 😪

A quick look at October 2020 shows that this spike in conversations within the group coincided with the **#EndSARS** protests that happened around the country. This definitely dominated conversations within the group and accounted for this spike.

December ranks lowest in terms of group activity, coinciding with the holidays and periods in which group members are mostly out of the office (on leave from work). We could deduce that this may mean most conversations in the group are office related, driven by experiences of members in day-to-day office interactions. Hence the fall in conversations in December could be due to members absence from the office work/environment.

On the flipside, it could just be the holidays and nothing else; a time when members are spending more time with family and loved one and away from their chatting away on phone. What's your take?



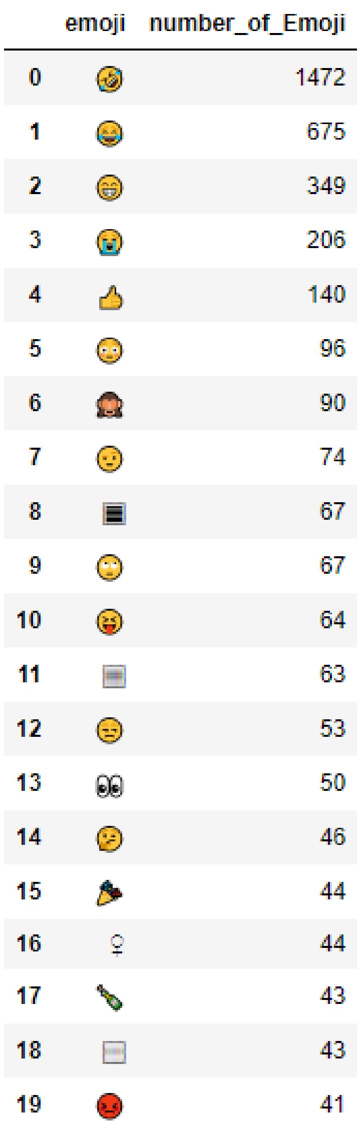
Viz 4: Month on Month change in activity

For the period in view, the breakdown of emoji usage in the group chat conversations is presented below. **Viz 5** shows a sorted order of the total number of emoji's used by count, while **Viz 6** plots this dataframe in a pie chart.

By inference, I'd say you guy smile and grin a lot with the **ROTFL** 🤣 emoji accounting for 39.5% of total emoji usage. That's 1,472 times one of you in the group is actually rolling on the floor laughing.

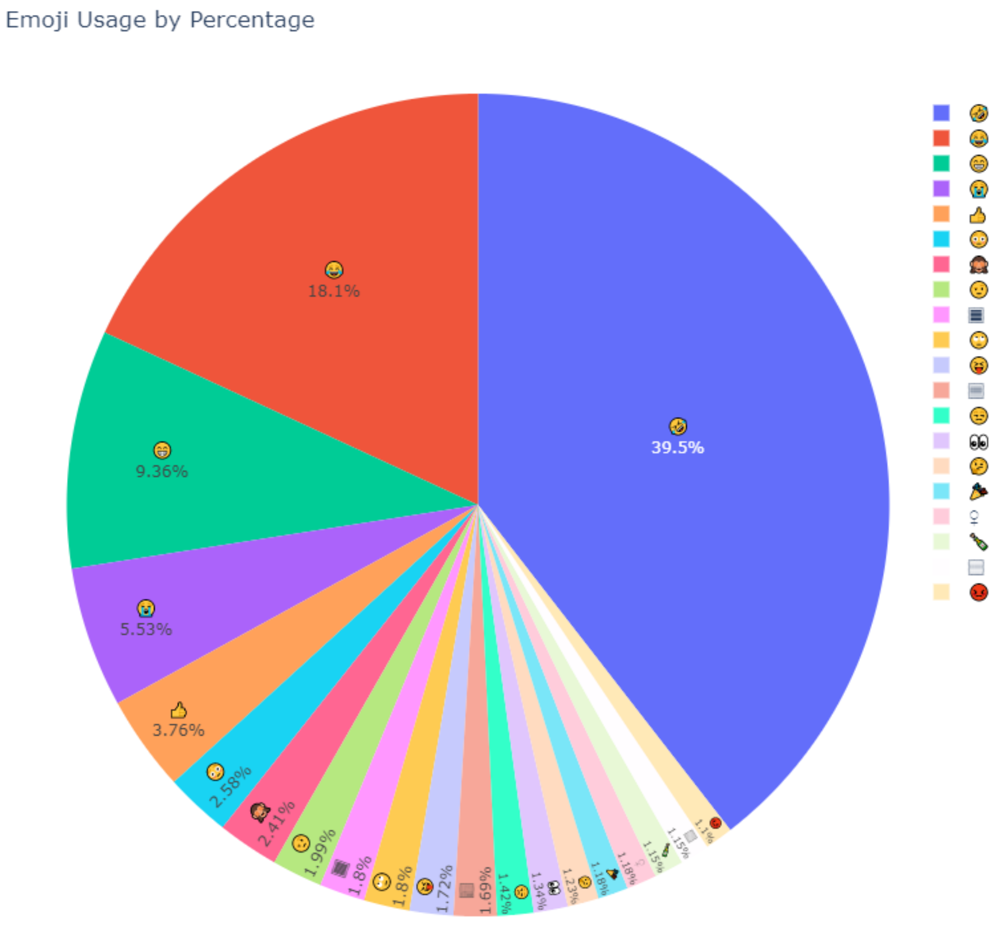
At the bottom of our spectrum is the angry face 😡, meaning there are fewer times than we think where Toba misbehaved and got a group member angry.

It's worth noting that the group also has a healthy amount of affirmation during conversations; supported by 140 thumbs up 👍 during the 365-day period analyzed.



	emoji	number_of_Emoji
0	🤣	1472
1	😂	675
2	😁	349
3	😭	206
4	👍	140
5	😳	96
6	🙈	90
7	🤨	74
8	▤	67
9	🙄	67
10	😝	64
11	▢	63
12	😒	53
13	👀	50
14	🤔	46
15	🎉	44
16	♀	44
17	🍾	43
18	▢	43
19	😡	41

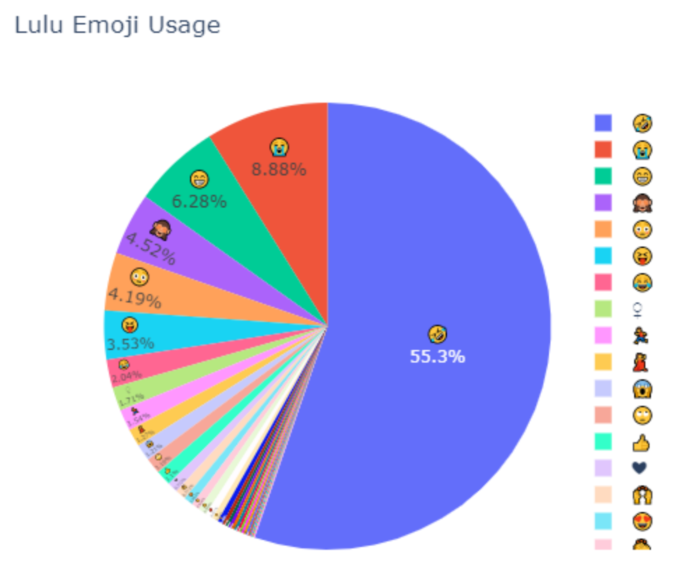
Viz 5: Total Emoji Usage by Count



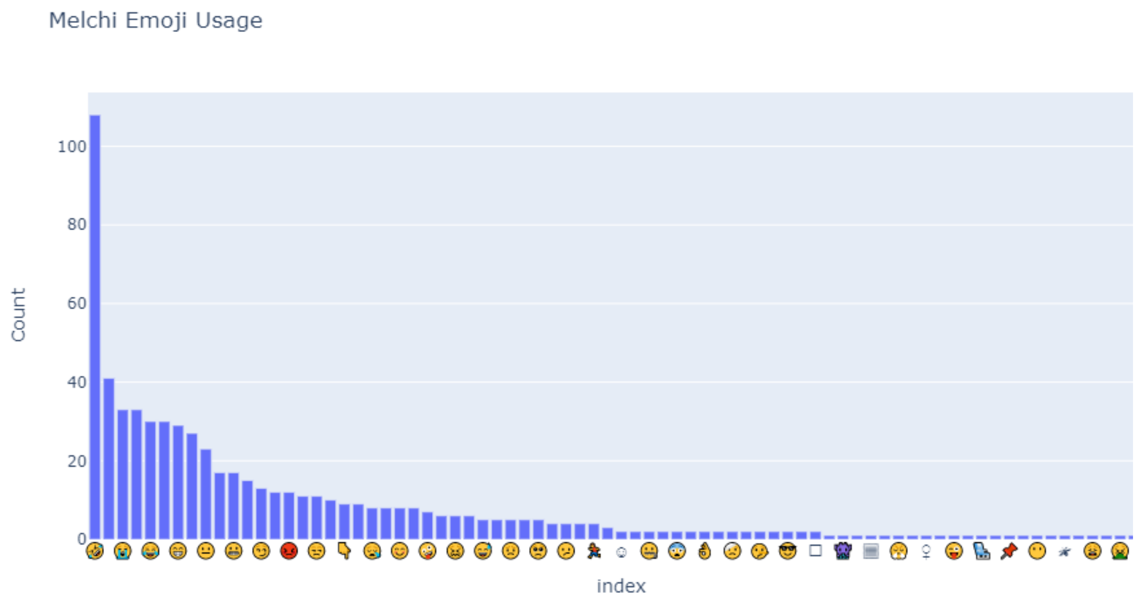
Viz 6: Emoji Usage by Percentage

Insights and Visualizations

Emoji Usage based on Group Member



Viz 7: Lulu Emoji Usage by Percentage



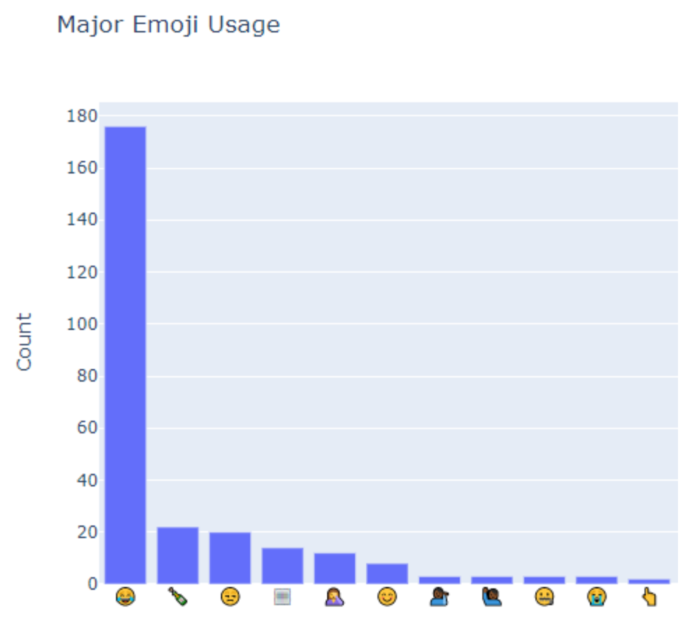
Viz 9: Melchi Emoji Usage by Count

This was definitely the most difficult insight to generate. I eventually got a breakthrough after watching **The Ice Road** movie to cool off and then taking a revision class on python dictionaries, regex and learning to parse through text to identify emoji.UNICODE_EMOJI in English language 😭.

Anyways, not to bore you; here we present the result of emoji usage for each individual group member. Some of the visualizations will be presented in pie charts by percentage, while others will be in bar charts by count.

A summary table of the results in shown on **Tab 1** to give a quick overview. The data shows that melchi is using the widest range of emoji in posted text compared to the rest of group members. Lulu outclasses the rest of the group with a whooping 1,003 🤣 posted within the period of examination.

I think it's safe to associate certain emojis with particular engineers based on **Tab 1**: ❤️ for molly, 🍾 for major, 🎉 for owoni, 🤨 for alhaji, 🙏 for buari, 🤣 for lulu, 😊 for judith, 👀 for melchi. With 🥳 standing as majek's most used emoji, I'm beginning to wonder if there's a glitch in the code written. Is majek our undercover party guy?



Viz 8: Major Emoji Usage by Count

Member	No. of unique emojis used	Your Top 3 Emojis by Count		
		1st	2nd	3rd
melchi	77	🤣 - 108	👀 - 41	😭 - 33
tobi	11	😁 - 33	😜 - 5	😘 - 4
major	22	😂 - 176	🍾 - 22	😒 - 20
kosi	26	🤣 - 9	😝 - 6	😘 - 5
lulu	39	🤣 - 1,003	😭 - 161	😁 - 114
bennie	17	🤣 - 105	□ - 67	👍 - 40
alhaji	18	🤨 - 11	😆 - 7	😏 - 5
majek	13	🥳 - 5	□ - 5	🤣 - 5
toba	22	🤣 - 92	🤨 - 55	😁 - 42
molly	29	🤣 - 116	😂 - 24	❤️ - 15
buari	10	😂 - 12	□ - 8	🙏 - 5
alfred	43	😂 -286	🤣 - 74	👍 - 32
madamPSI	33	🤣 - 23	😂 -11	👍 - 10
owoni	15	🎉 - 18	🥳 - 12	😂 -5
judith	16	😁 - 93	😊 - 22	👍 - 21

Tab 1: Emoji Usage by Percentage

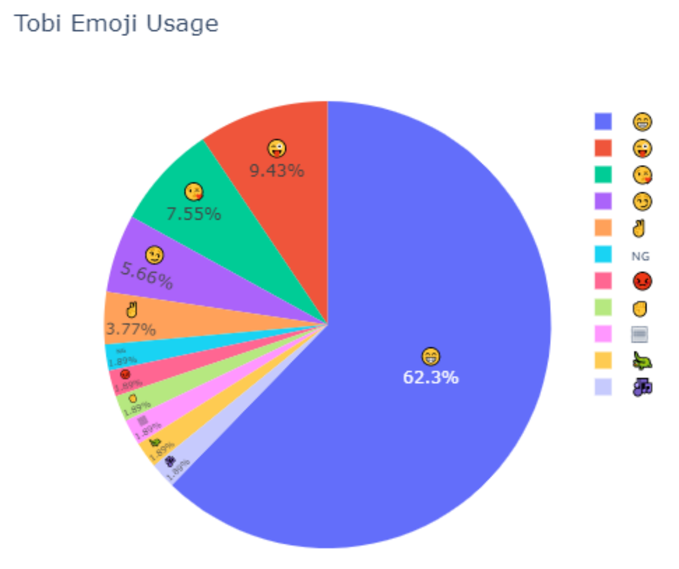
Note that the □ emoji present in Tab 1 isn't really an emoji, but rather an associated Unicode which usually appears just after some specific emojis, but was rendered by python as an emoji itself. I'm still figuring out how to work a way around this and present a cleaner representation.

madamPSI used a pretty wide range of emojis (33) when compared to total messages sent within the period - 131 messages in all. We could infer that she's the most inclined to using SPECIFIC emoji's in group conversations.

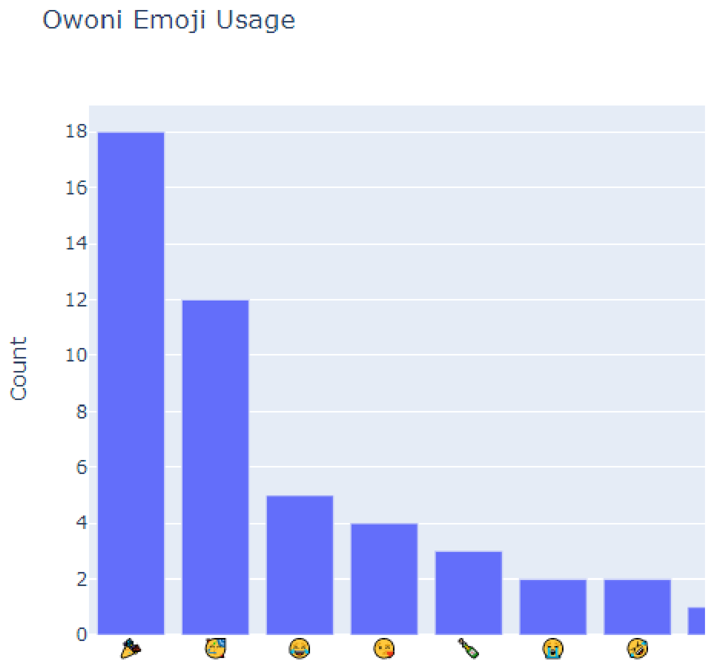
Insights and Visualizations

Emoji Usage based on Group Member

A sample of the charts for selected group members is presented in the following to enable us draw more insights or inferences. It can show the emoji's particular group members are more inclined to use, or it could show if a member displays a more balanced usage of a range of emojis.



Viz 10: Tobi Emoji Usage by Percentage

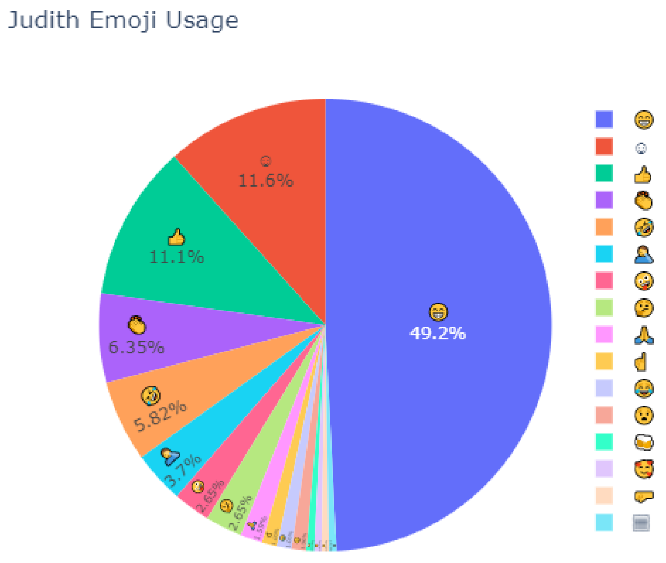


Viz 12: Owoni Emoji Usage by Count

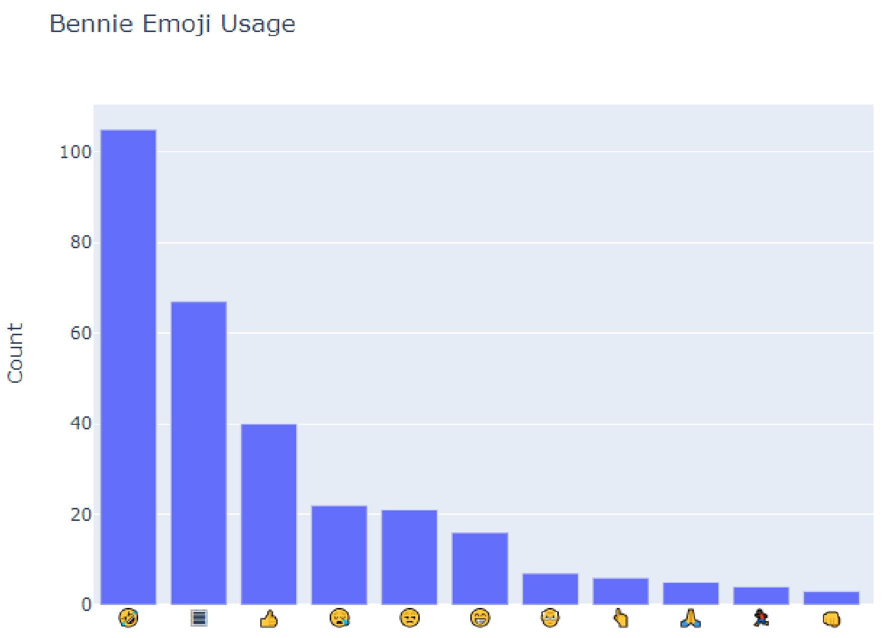
If we ignore the first 2 bars on **Viz 11**, we see that bennie is throwing in more affirmation in the group with the 👍 emoji. He also seems to be the only one with 👴 making the list. Lol. Osheee baba!

Judith and Alfred have stuck with a particular emoji 50% of the time compared to their total emoji usage.

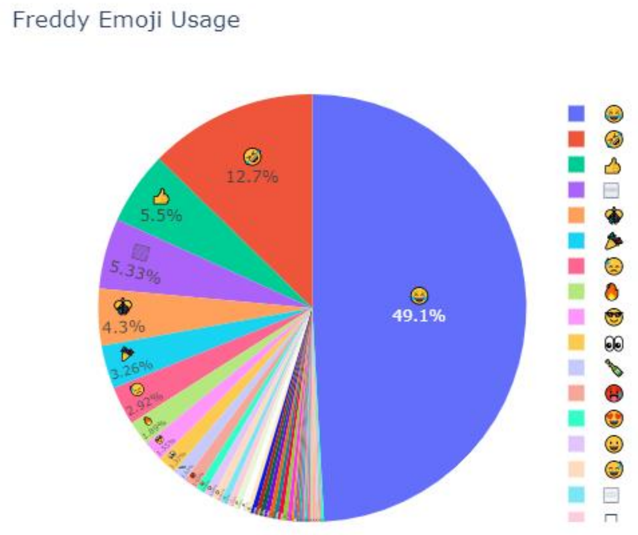
Owoni's selection of emoji usage (🎉🥳🍾) no lie at all. We needed no data analysis for this one. If you've missed owoni's house party, let me hear you say aye! The aye's have it.



Viz 13: Judith Emoji Usage by Count



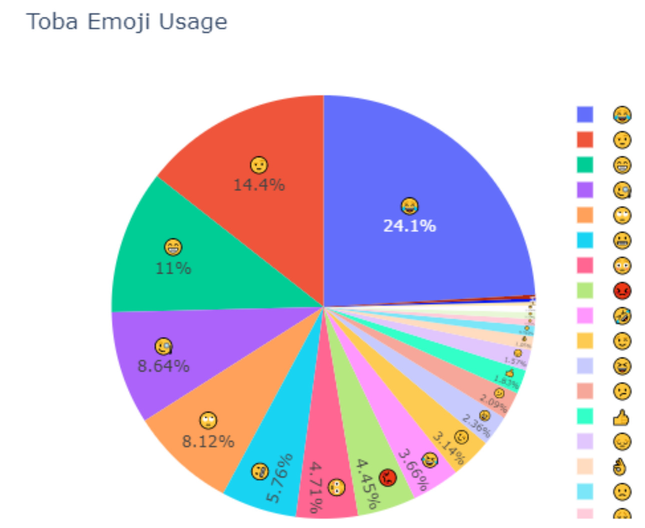
Viz 11: Bennie Emoji Usage by Count



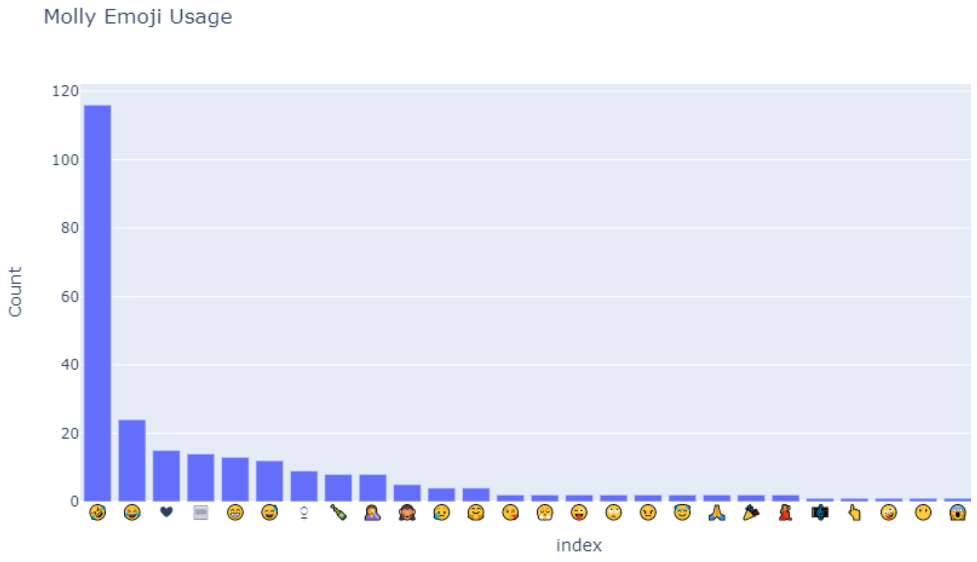
Viz 14: Freddy Emoji Usage by Count

Insights and Visualizations

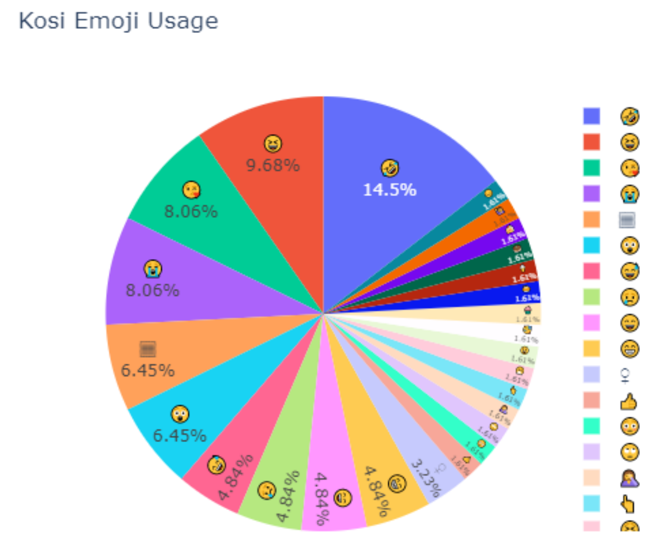
Emoji Usage based on Group Member



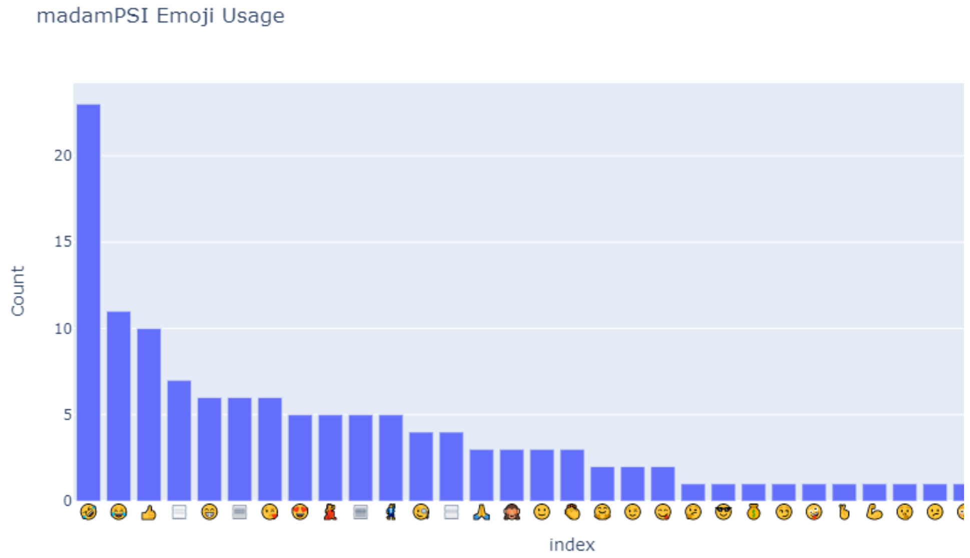
Viz 15: Toba Emoji Usage by Percentage



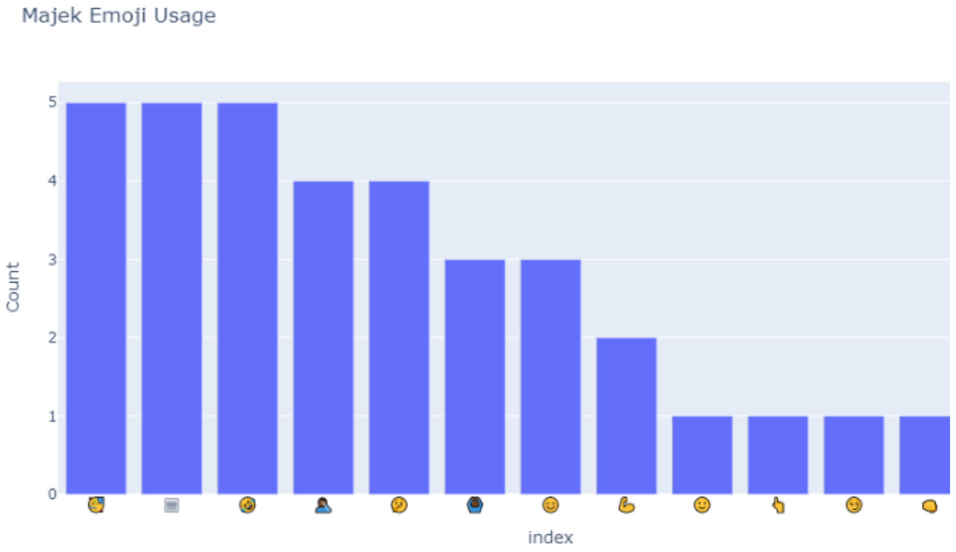
Viz 18: Molly Emoji Usage by Count



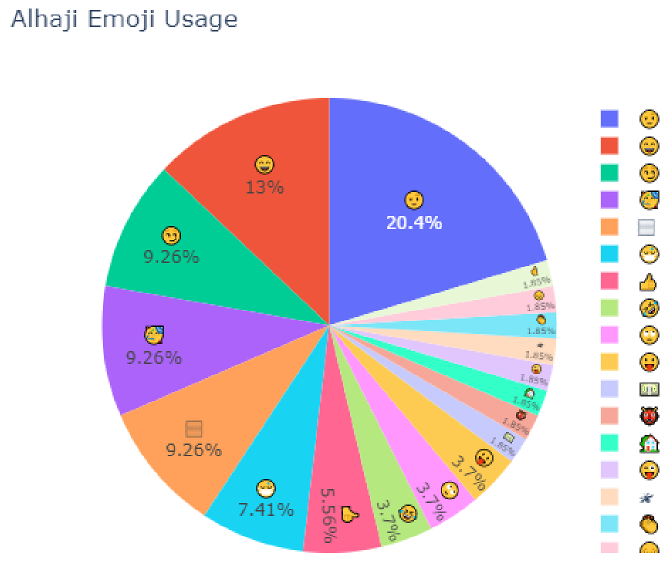
Viz 16: Kosi Emoji Usage by Percentage



Viz 19: madamPSI Emoji Usage by Count



Viz 20: Majek Emoji Usage by Percentage



Viz 17: Alhaji Emoji Usage by Percentage

Toba, kosi, alhaji seem to also display a more distributed use of emoji's, at least for their top 5 records. What other questions can we ask of the data on emoji usage?

Insights and Visualizations

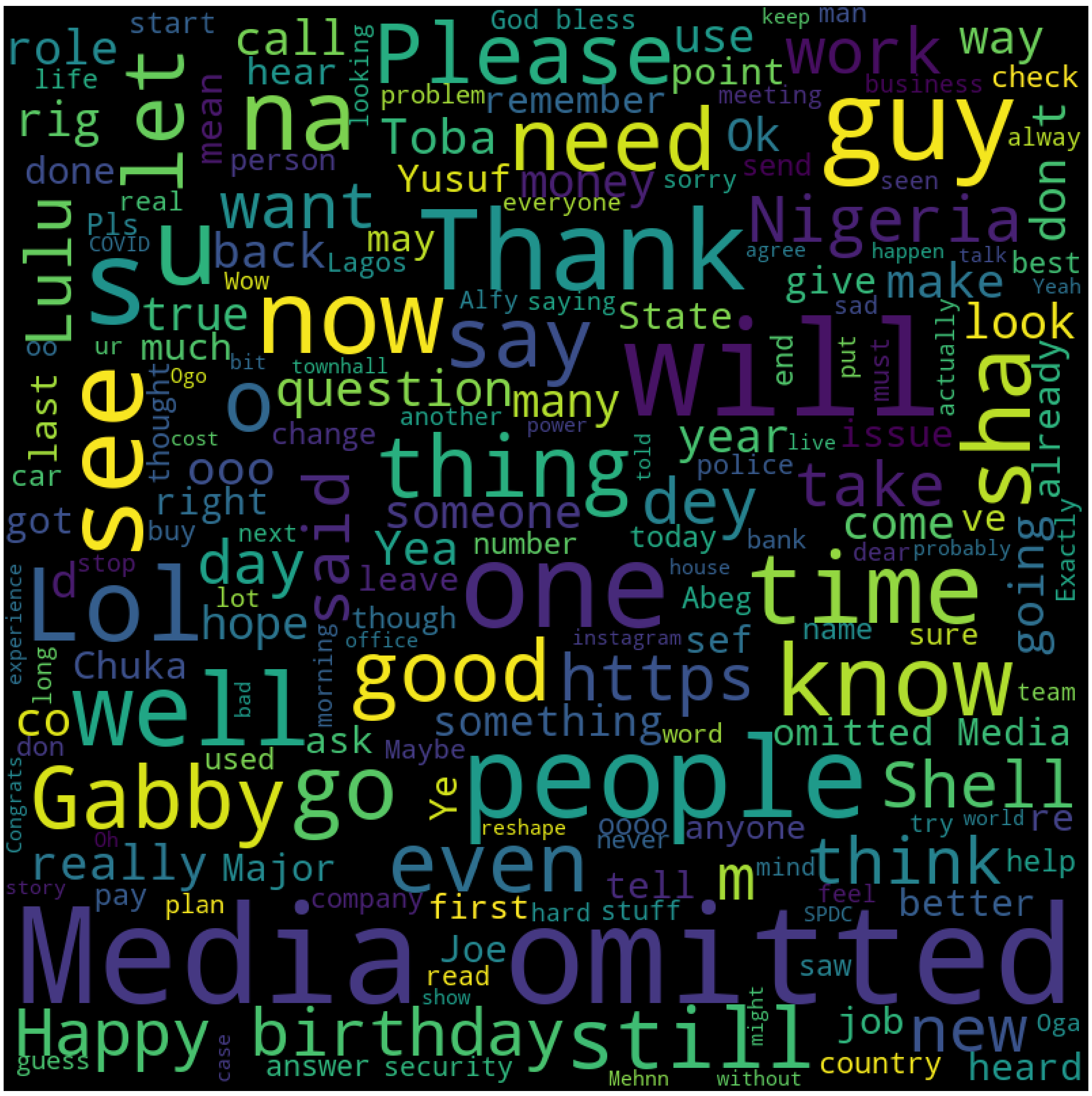
Words Usage and Frequency

Stepping away from our love for emoji's, I performed an analysis on the group's commonly used words, which is presented in a Word Cloud (Viz 21) using a wordcloud library.

I'll let y'all make out what you think of this result; nonetheless, permit me to draw your attention to a few pointers:

- At the bottom of the image, you see "**Media omitted**". This refers to all the times one form of multimedia messages were sent in the group. Remember, that for simplicity and size of WhatsApp exported data, I had chosen to download the files '**without media**'. Hence, for all the places where bennie sent his favorite 'aki' & 'pawpaw' gifs or memes 😏👀, or toba posts something he thinks is funny about Wike, 🤦‍♂️, WhatsApp rendered them in the .txt file as <Media omitted>. Our word cloud then captures this as a text. From the size of the word "Media omitted", it means y'all send out quite a lot of multimedia messages in the group.
- Is that 'Joe' I see in between the words Media and omitted? Nawa o, una harsh oo...😬, y'all spend this much time talking about a certain 'Joe'?
- And can someone explain to me why the word '**Gabby**' is this large in the cloud; probably an anomaly- let's skip that.
- On a serious note, pay attention to the group's highflyer words: Will, Thank, U, People, See, Now, Lol, Well, One, Need, Guy, Sha, Know, Time, Go, Still, Good, Please...
- And by extension the least used words: Oga, show, case, bad, story, house, team, dear, probably, told, Instagram, cost, bit, ur, talk keep, COVID...
- For the Nigerian identity, the word 'sha' made a decent representation.

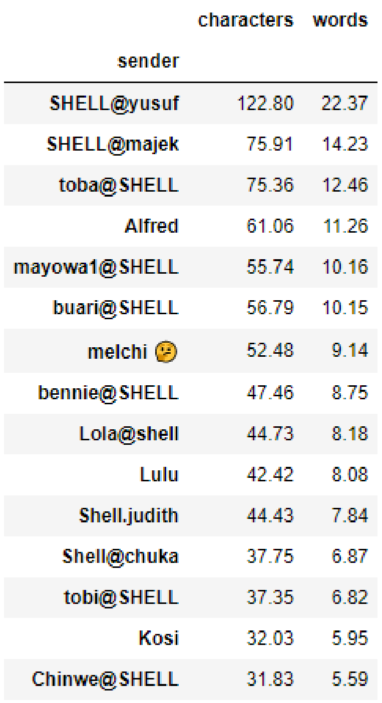
There's a whole world to explore from these results with opportunities to associate words here and there to group members personality or otherwise. These words can also be further aggregated to draw out correlations with the other aspects of this analysis that have been previously shown. Maybe someday when I have the time...for now, let's move on.



Viz 21: Commonly Used Words in the group Conversations

Deep dive

Message Length & Distinct Conversations



sender	characters	words
SHELL@yusuf	122.80	22.37
SHELL@majek	75.91	14.23
toba@SHELL	75.36	12.46
Alfred	61.06	11.26
mayowa1@SHELL	55.74	10.16
buari@SHELL	56.79	10.15
melchi 🤔	52.48	9.14
bennie@SHELL	47.46	8.75
Lola@shell	44.73	8.18
Lulu	42.42	8.08
Shell.judith	44.43	7.84
Shell@chuka	37.75	6.87
tobi@SHELL	37.35	6.82
Kosi	32.03	5.95
Chinwe@SHELL	31.83	5.59

Viz 22: Average message character/ words per member

Viz 22 shows a breakdown of the average number of words and characters used in a line of text message posted by a user. This has been sorted in descending order.

Now isn't this interesting, see the folks toping the charts 🤣🤣. Let me sound smart with my next sentence. At the top of the chart are 'serious' minded people, who take their time to sit down and construct their complete line of thought in a text before posting. Lulu, with all her group activity no reach top 5 🙊; she's typing far fewer words and characters per line of text sent, which then explains why she's considered highly active (due to number of texts posted). I'm not talking about melchi cos I can't be judge and jury, sorry. 😋.

Permit me to remove Toba from the previous list of 'serious' minded people. He's is an outlier; that high average number of words per text is probably his copy and paste from Wike's broadcasts. 😏

Top Word

"will"

In the space of 365 days of conversation on the group chat, this word stands out as the most used by group members. 💪

Unexpected Entry

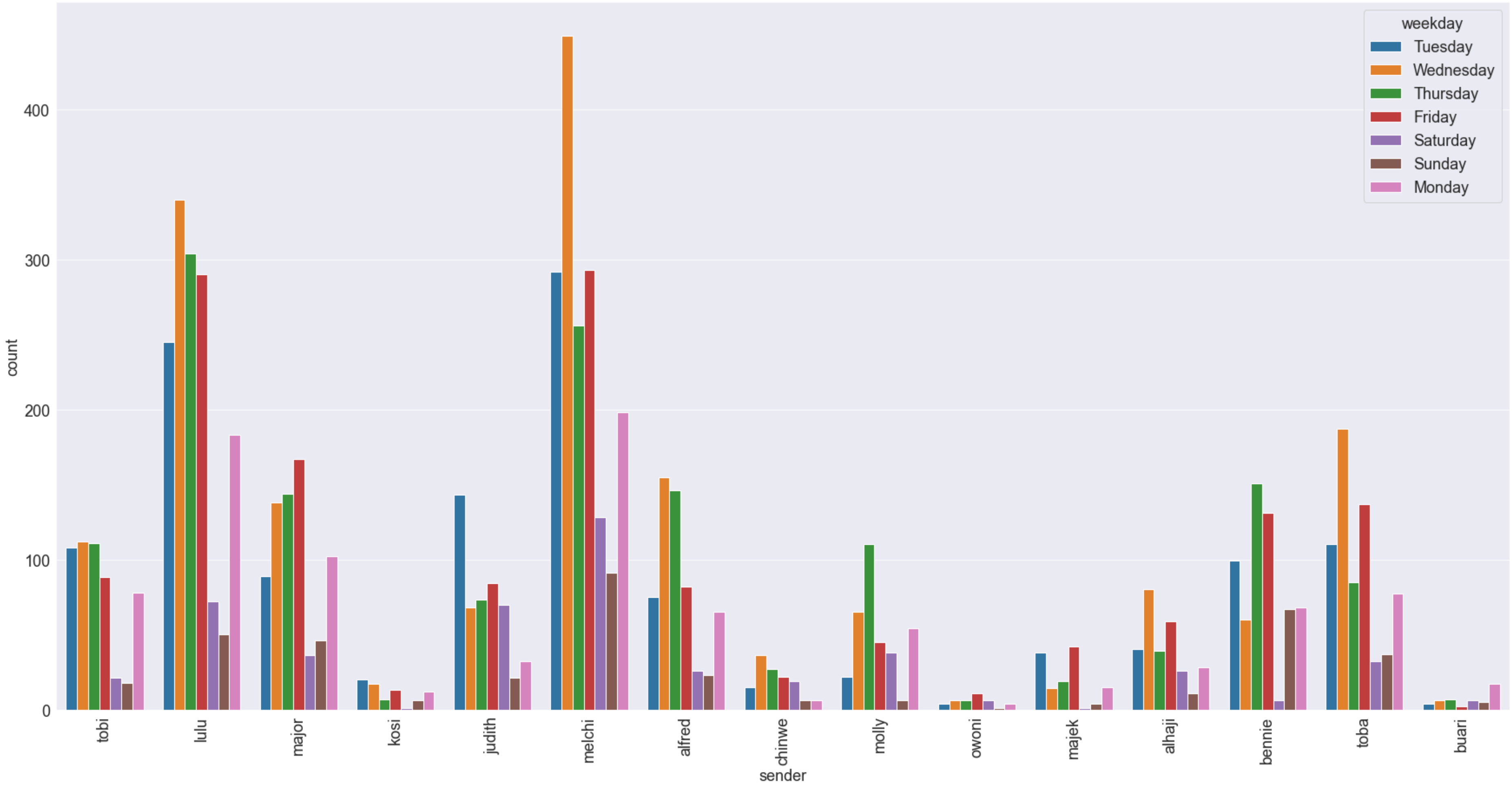
"Joe"

Let me pretend that I have no idea how this 'word' makes the list. If I say this could be related to some people's stress now, bennie will shout 'DSS is watching you" 😏

Group Identity

"rig"

This was once a common 'language' for the group. Gradually becoming history for some. To rub salt into that injury, some people came to show us DS10. 😡

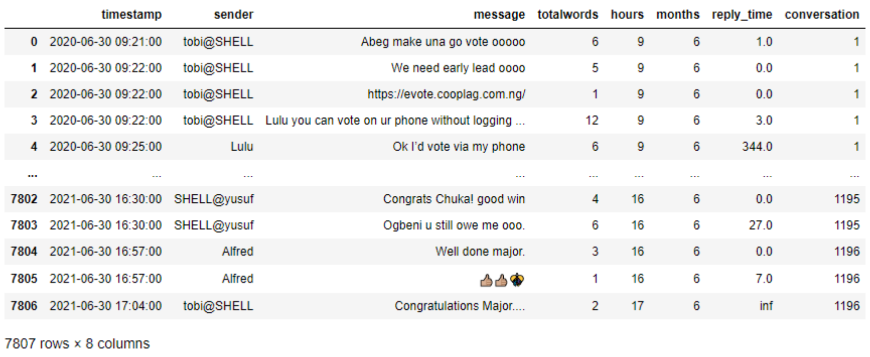


Viz 23: Message Count by Weekday

Still on **Viz 22**, e be like say Yusuf sef dey do copy and paste o 😕. When we put it all together, the group is doing an average of 54 characters and 9 words per text message in chat conversations

Distinct Conversations within the group

For my next analysis, I have made an assumption by defining a conversation to be a stream of messages where the gap between any two messages in not longer than 20 minutes. So, iterating through the list of text sent in chronological order, I find the difference between the current and previous message and if they are less than 20 minutes apart, they are grouped as part of the same conversation, and if more than 20 minutes apart, that message is considered part of the next group of conversation.



	timestamp	sender	message	totalwords	hours	months	reply_time	conversation
0	2020-06-30 09:21:00	tobi@SHELL	Abeg make una go vote ooooo	6	9	6	1.0	1
1	2020-06-30 09:22:00	tobi@SHELL	We need early lead oooo	5	9	6	0.0	1
2	2020-06-30 09:22:00	tobi@SHELL	https://evote.cooplag.com.ng/	1	9	6	0.0	1
3	2020-06-30 09:22:00	tobi@SHELL	Lulu you can vote on ur phone without logging ...	12	9	6	3.0	1
4	2020-06-30 09:25:00	Lulu	Ok I'd vote via my phone	6	9	6	344.0	1
...	...	...	...	...	...	...	...	...
7802	2021-06-30 16:30:00	SHELL@yusuf	Congrats Chuka! good win	4	16	6	0.0	1195
7803	2021-06-30 16:30:00	SHELL@yusuf	Ogbeni u still owe me ooo.	6	16	6	27.0	1195
7804	2021-06-30 16:57:00	Alfred	Well done major.	3	16	6	0.0	1196
7805	2021-06-30 16:57:00	Alfred	👍👍🎊	1	16	6	7.0	1196
7806	2021-06-30 17:04:00	tobi@SHELL	Congratulations Major....	2	17	6	inf	1196

7807 rows × 8 columns

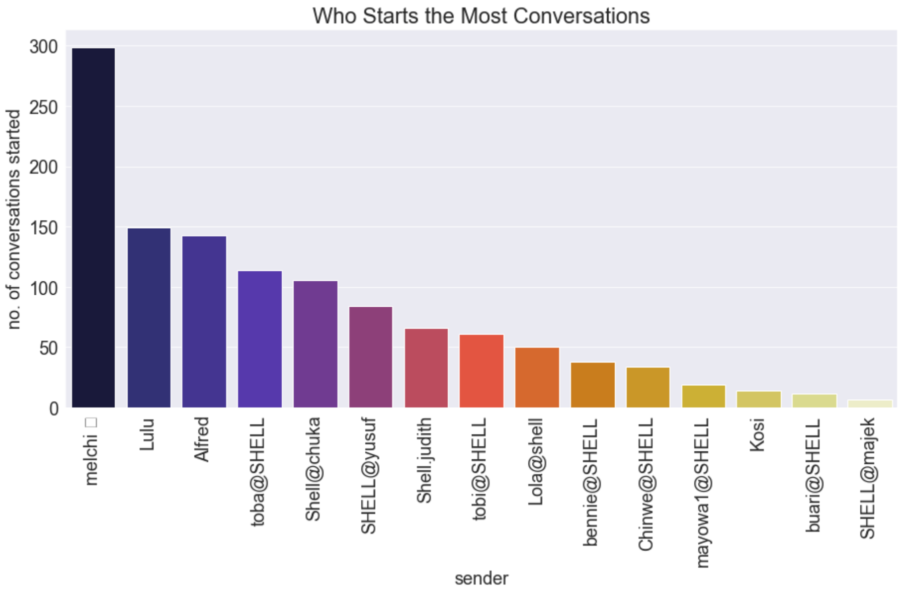
Viz 23: Conversation dataframe

And the results are in! For the period under analysis, and following my assumption stated above, there have been 1,196 separate conversations, lasting an average of 10.9 minutes and surprisingly made up of 7 messages from just 2 participants.

Hmm, knowing this group, this doesn't seem to be correct – I probably need to adjust my assumptions and test again. What do you guys think is average time to consider a conversation closed based on the length of the 'silent' period after a message is sent, before the next one.

I believe we also average more than 2 participants in group conversations. 😕

As a follow-up to the previous result though, we tried to find out which group member is responsible for starting the most conservations; whoever this turns out to be, permit me to say, he or she is responsible for starting off smart conversations, in short, very smart conversation. 😊



Viz 24: Conversation Starters

And the prize goes to melchi, yippee! 🎉🎊✨. Tseink you, tseink you 🙏.

Weekend versus Weekday Pattern

The visualization on **Viz 23** is a concatenated plot of group activity per group member, per day of the week. Any surprises why Wednesday is the most active day on the group? In one word – gossip. Gossip is sweet sha. 😊 Either major is asking lulu to run commentary from Townhall, or toba is throwing a tantrum about management. Y'all talk on Wednesdays the most.

Not for all though: major, judith, molly, majek, bennie seem to be more in sync with Fridays, Tuesdays, Thursdays, Fridays and Thursdays respectively.

Kosi and majek have taken oaths against sending group chat messages on Saturdays, while it's Sunday for owoni.

With a total word count of **62,920 sent on Weekdays** (Monday to Friday) compared to **10,002 words sent on Weekends** within the period analyzed, it puts the group on an average of **12,584 words per day on Weekdays** and **5,001 words per day on Weekends.**

Finally, I present a heat map on Viz 25 to give a visual indication of the interaction between day of the week, time of day and group activity by number of messages sent. The darker the cell, the more messages were sent within that intersection on '**hours**' and '**weekday**'.

Wednesday 9am, Friday 9am, Monday 3pm, Friday 4pm, Thursday 5pm are points of interest. You can make out your own inferences.

Midnight to 7am is pretty quiet, as should be the case, so we all cool there.

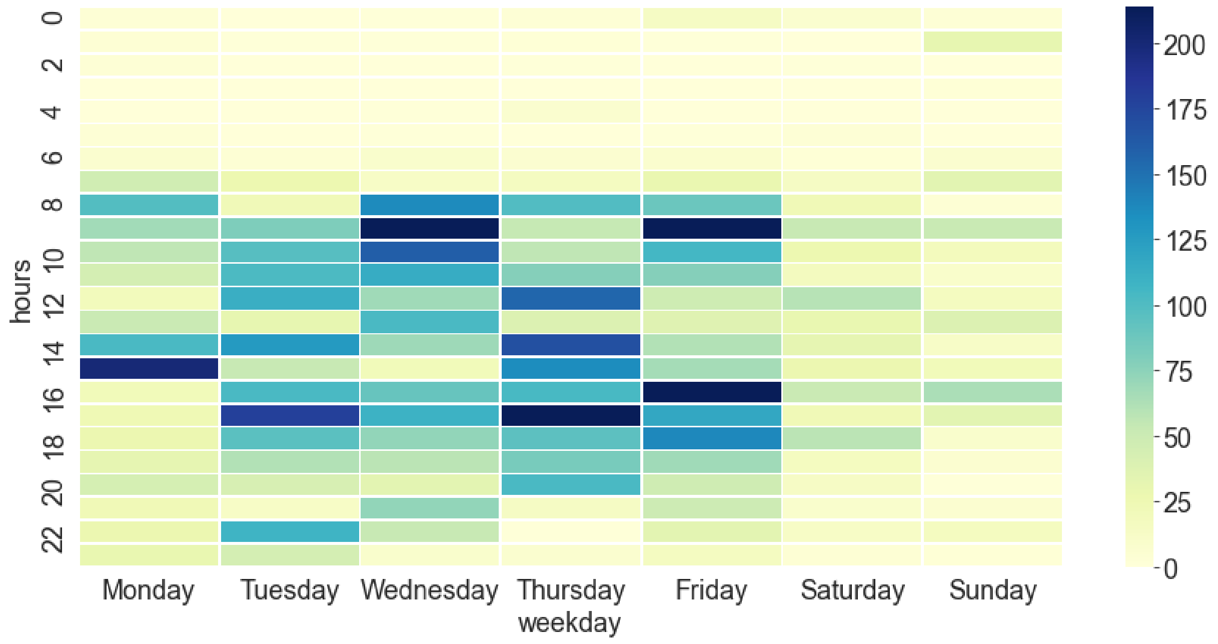
Final words

With every Data Analysis Project, there's always room to expand upon the results, there's always more questions that can be asked of the data, there's always a hypothesis to test, there's always more insights that can be gleaned.

But permit this brother to call it a stop at this junction, pending when some interesting dataset comes my way.

Looking forward to hearing your thoughts on this, especially challenging the inferences and conclusions that the data (which is fact by the way 😏) has presented to us.

Let me know of any questions you would love to ask of the data; until then I don't want any of you fighting over the results in this report. God bless y'all.



Viz 25: Heat Map of Message Count per Weekday per Time-of-day