Propaganda data investigation

Vladyslav Bezborodov CS-3
Id name 24
Computational Social Science
Teacher: Andrew Kurochkin

5.12.2023

Agenda

1. Introduction
2. Data acquisition
3. Exploratory data analysis
4. Further work
5. Sources

Introduction

- A few seconds about dataset
- What I tried investigating
- Presenting my results

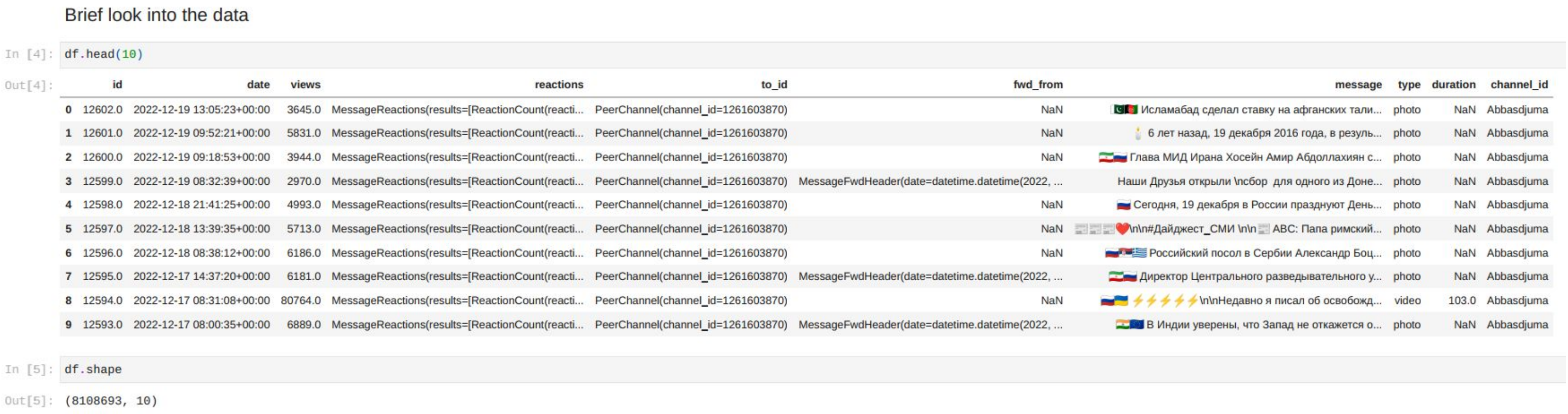
Data acquisition

- Used a propaganda dataset by Kate Burovova
- ~300 of russian telegram propaganda channels
  (just needed to join them in one file, used ipynb file for merging from previous homework)
- Obtaining time : 30-60 min
- Possible problems : had to adapt ipynb file for current task

- Data statistics :

  Posts quantity : 8108693

  Dataset size : 6670 Megabytes



Brief look into the data

```
In [4]: df.head(10)
```

Out[4]:

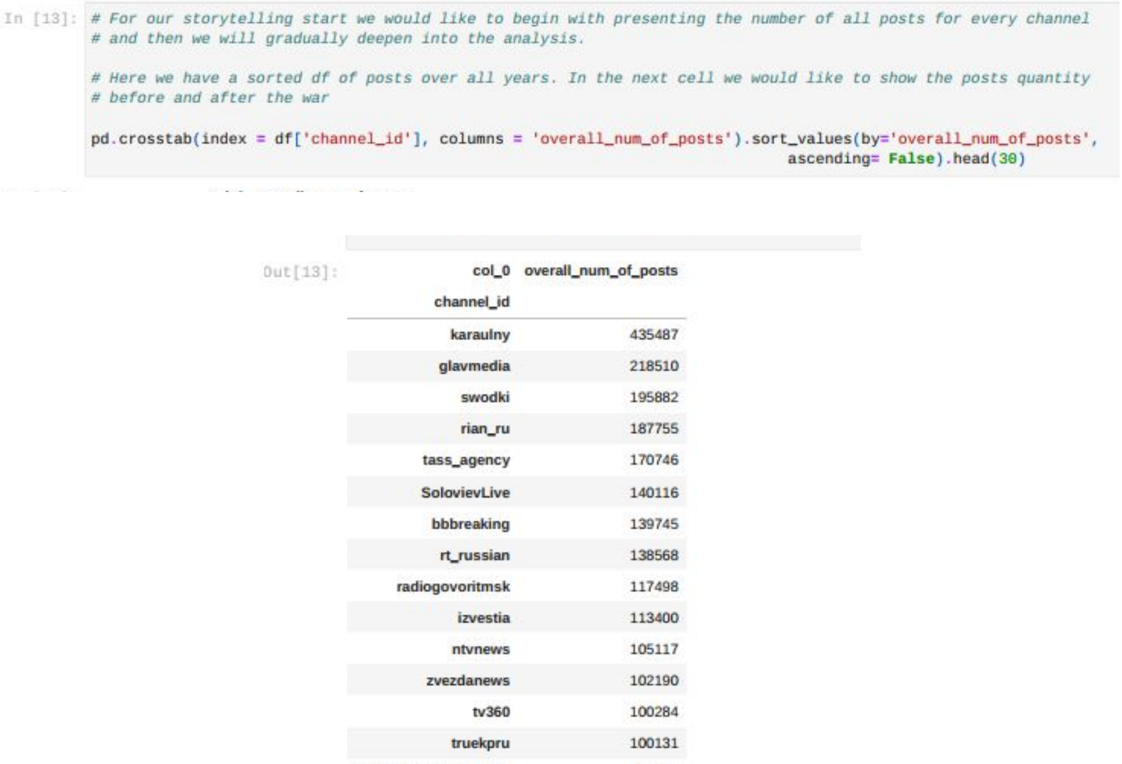
	id	date	views	reactions	to_id	fwd_from	message	type	duration	channel_id
0	12602.0	2022-12-19 13:05:23+00:00	3645.0	MessageReactions(results=[ReactionCount(reacti...	PeerChannel(channel_id=1261603870)	NaN	🇵🇰🇦🇫 Исламабад сделал ставку на афганских тали...	photo	NaN	Abbasdjuma
1	12601.0	2022-12-19 09:52:21+00:00	5831.0	MessageReactions(results=[ReactionCount(reacti...	PeerChannel(channel_id=1261603870)	NaN	🕯 6 лет назад, 19 декабря 2016 года, в резуль...	photo	NaN	Abbasdjuma
2	12600.0	2022-12-19 09:18:53+00:00	3944.0	MessageReactions(results=[ReactionCount(reacti...	PeerChannel(channel_id=1261603870)	NaN	🇮🇷🇷🇺 Глава МИД Ирана Хосейн Амир Абдоллахиян с...	photo	NaN	Abbasdjuma
3	12599.0	2022-12-19 08:32:39+00:00	2970.0	MessageReactions(results=[ReactionCount(reacti...	PeerChannel(channel_id=1261603870)	MessageFwdHeader(date=datetime.datetime(2022, ...	Наши Друзья открыли \nсбор для одного из Доне...	photo	NaN	Abbasdjuma
4	12598.0	2022-12-18 21:41:25+00:00	4993.0	MessageReactions(results=[ReactionCount(reacti...	PeerChannel(channel_id=1261603870)	NaN	🇷🇺 Сегодня, 19 декабря в России празднуют День...	photo	NaN	Abbasdjuma
5	12597.0	2022-12-18 13:39:35+00:00	5713.0	MessageReactions(results=[ReactionCount(reacti...	PeerChannel(channel_id=1261603870)	NaN	📰📰📰❤️\n\n#Дайджест_СМИ \n\n📰 ABC: Папа римский...	photo	NaN	Abbasdjuma
6	12596.0	2022-12-18 08:38:12+00:00	6186.0	MessageReactions(results=[ReactionCount(reacti...	PeerChannel(channel_id=1261603870)	NaN	🇷🇺🇷🇸🇬🇷 Российский посол в Сербии Александр Боц...	photo	NaN	Abbasdjuma
7	12595.0	2022-12-17 14:37:20+00:00	6181.0	MessageReactions(results=[ReactionCount(reacti...	PeerChannel(channel_id=1261603870)	MessageFwdHeader(date=datetime.datetime(2022, ...	🇮🇷🇷🇺 Директор Центрального разведывательного у...	photo	NaN	Abbasdjuma
8	12594.0	2022-12-17 08:31:08+00:00	80764.0	MessageReactions(results=[ReactionCount(reacti...	PeerChannel(channel_id=1261603870)	NaN	🇷🇺🇺🇦 ⚡⚡⚡⚡⚡\n\nНедавно я писал об освобожд...	video	103.0	Abbasdjuma
9	12593.0	2022-12-17 08:00:35+00:00	6889.0	MessageReactions(results=[ReactionCount(reacti...	PeerChannel(channel_id=1261603870)	MessageFwdHeader(date=datetime.datetime(2022, ...	🇮🇳🇪🇺 В Индии уверены, что Запад не откажется о...	photo	NaN	Abbasdjuma

```
In [5]: df.shape

Out[5]: (8108693, 10)
```

Data Analysis

Investigated questions

- How has the overall activity (number of posts) in these propaganda channels was evolving before and after the beginning of the war?

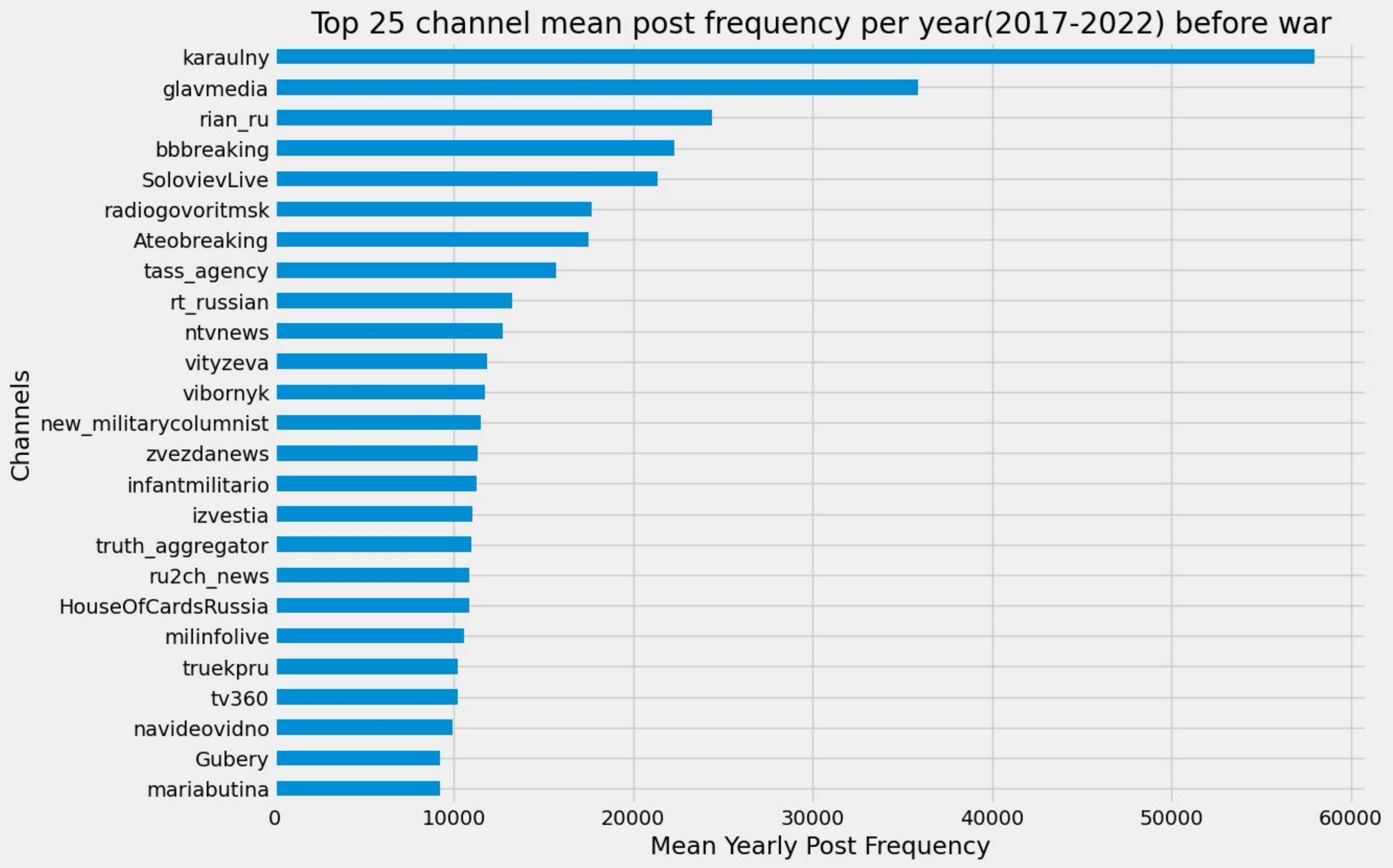


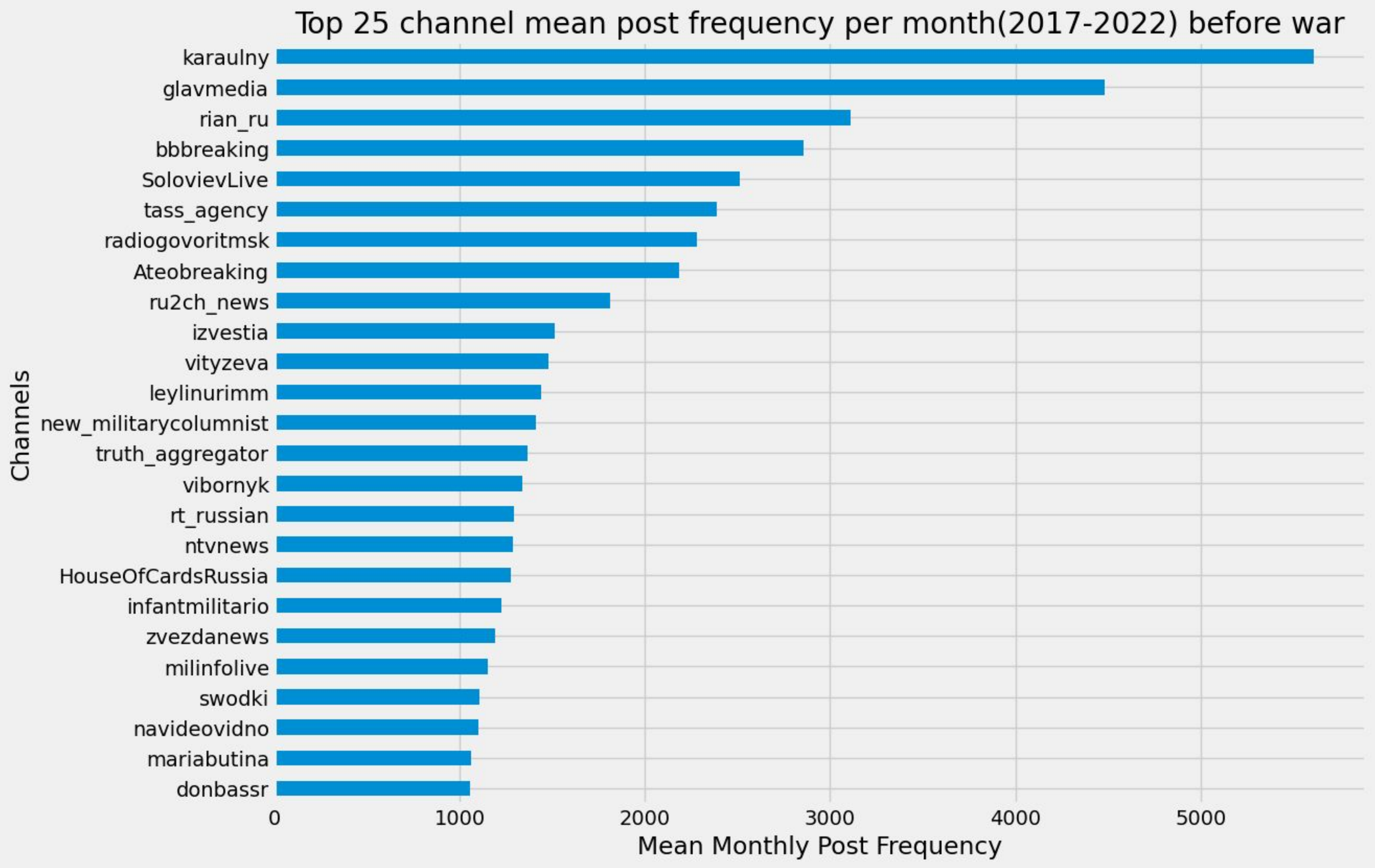
In [13]:

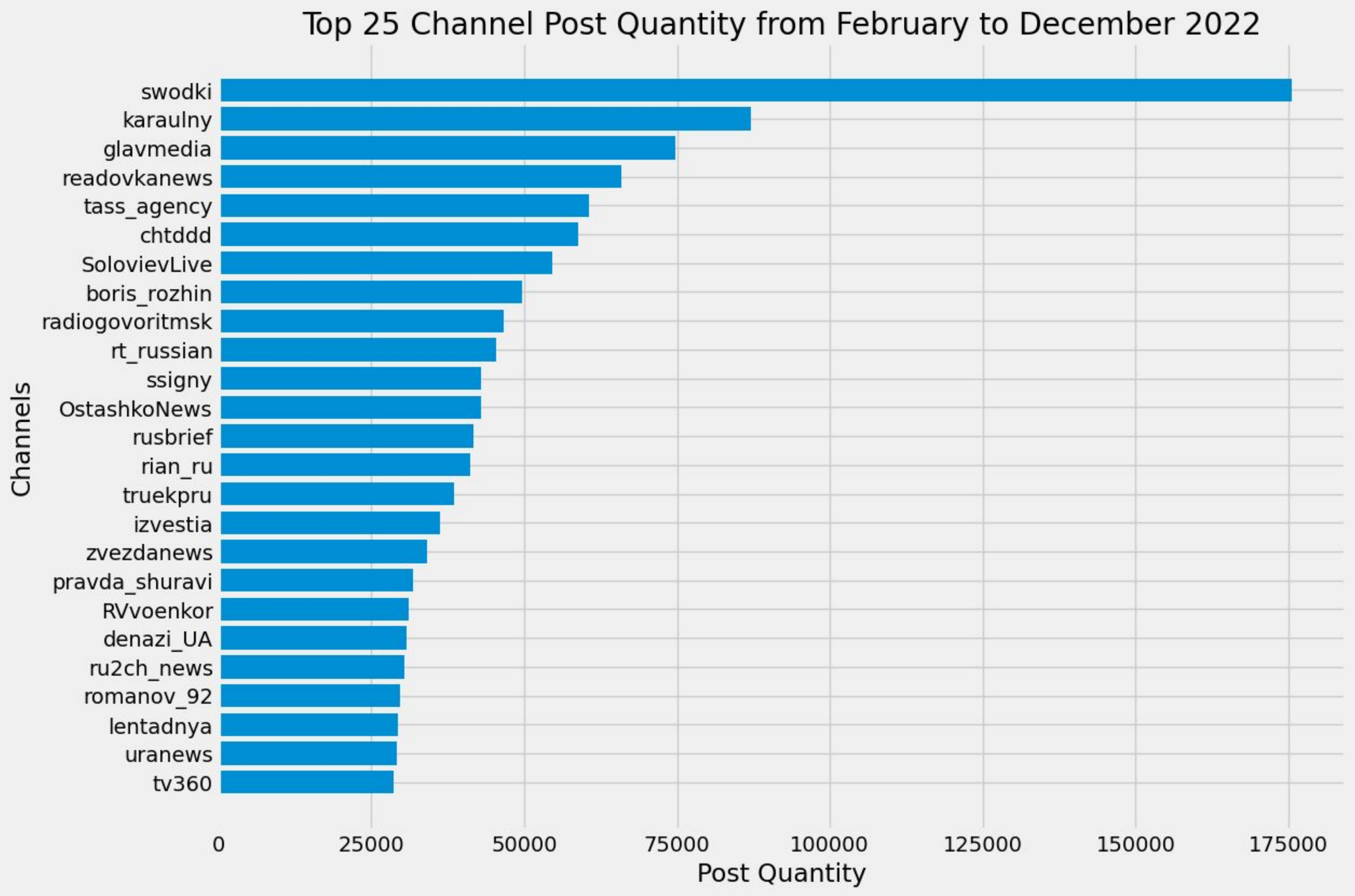
```
# For our storytelling start we would like to begin with presenting the number of all posts for every channel
# and then we will gradually deepen into the analysis.

# Here we have a sorted df of posts over all years. In the next cell we would like to show the posts quantity
# before and after the war

pd.crosstab(index = df['channel_id'], columns = 'overall_num_of_posts').sort_values(by='overall_num_of_posts',
                                                                                    ascending= False).head(30)
```

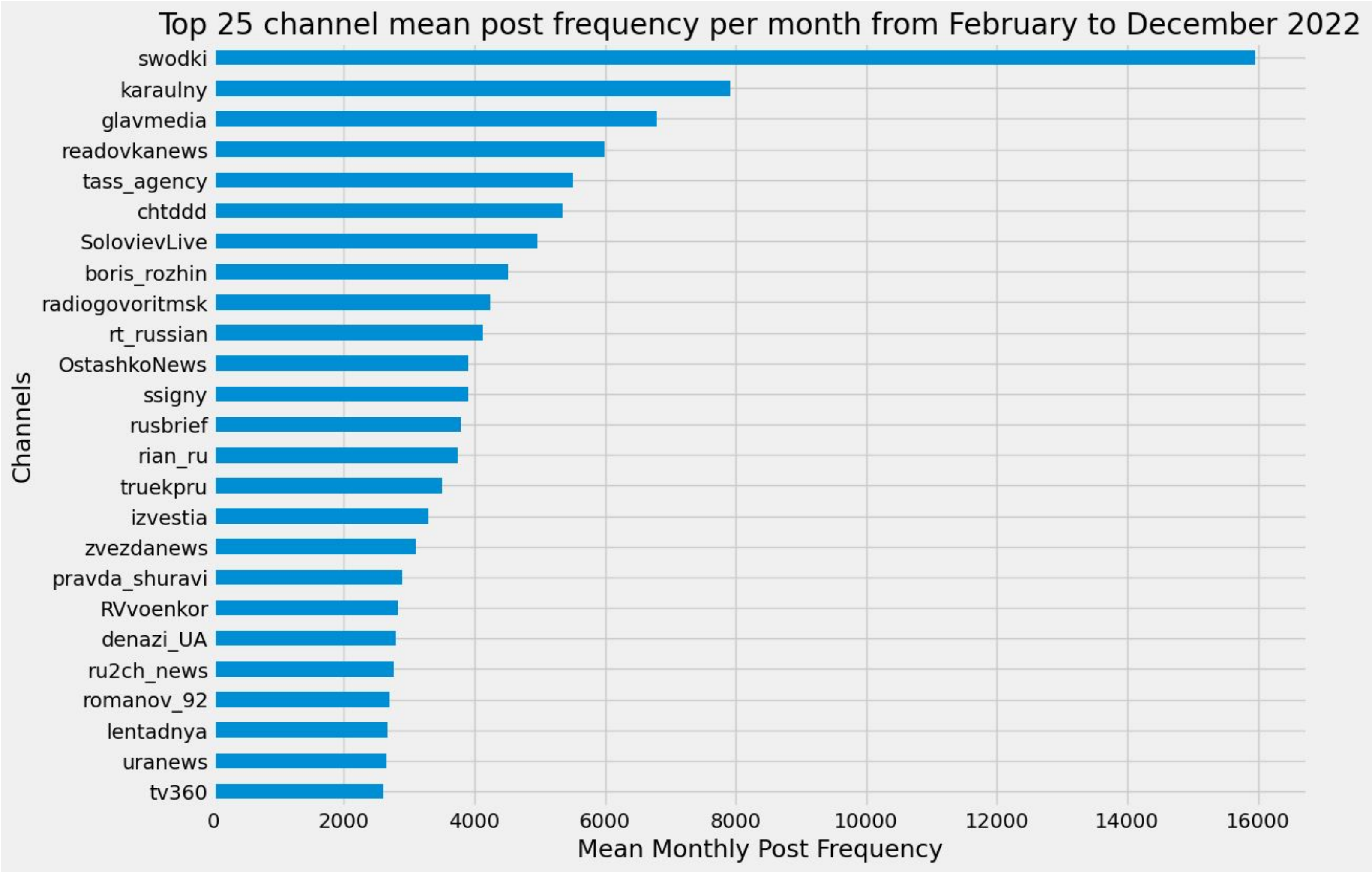
Out[13]:

col_0	overall_num_of_posts
channel_id	
karaulny	435487
glavmedia	218510
swodki	195882
rian_ru	187755
tass_agency	170746
SolovievLive	140116
bbbreaking	139745
rt_russian	138568
radiogovoritmsk	117498
izvestia	113400
ntvnews	105117
zvezdanews	102190
tv360	100284
truekpru	100131


Top 25 channel mean post frequency per year(2017-2022) before war
Channels
karaulny
glavmedia
rian_ru
bbbreaking
SolovievLive
radiogovoritmsk
Ateobreaking
tass_agency
rt_russian
ntvnews
vityzeva
vibornyk
new_militarycolumnist
zvezdanews
infantmilitario
izvestia
truth_aggregator
ru2ch_news
HouseOfCardsRussia
milinfolive
truekpru
tv360
navideovidno
Gubery
mariabutina
0
10000
20000
30000
40000
50000
60000
Mean Yearly Post Frequency

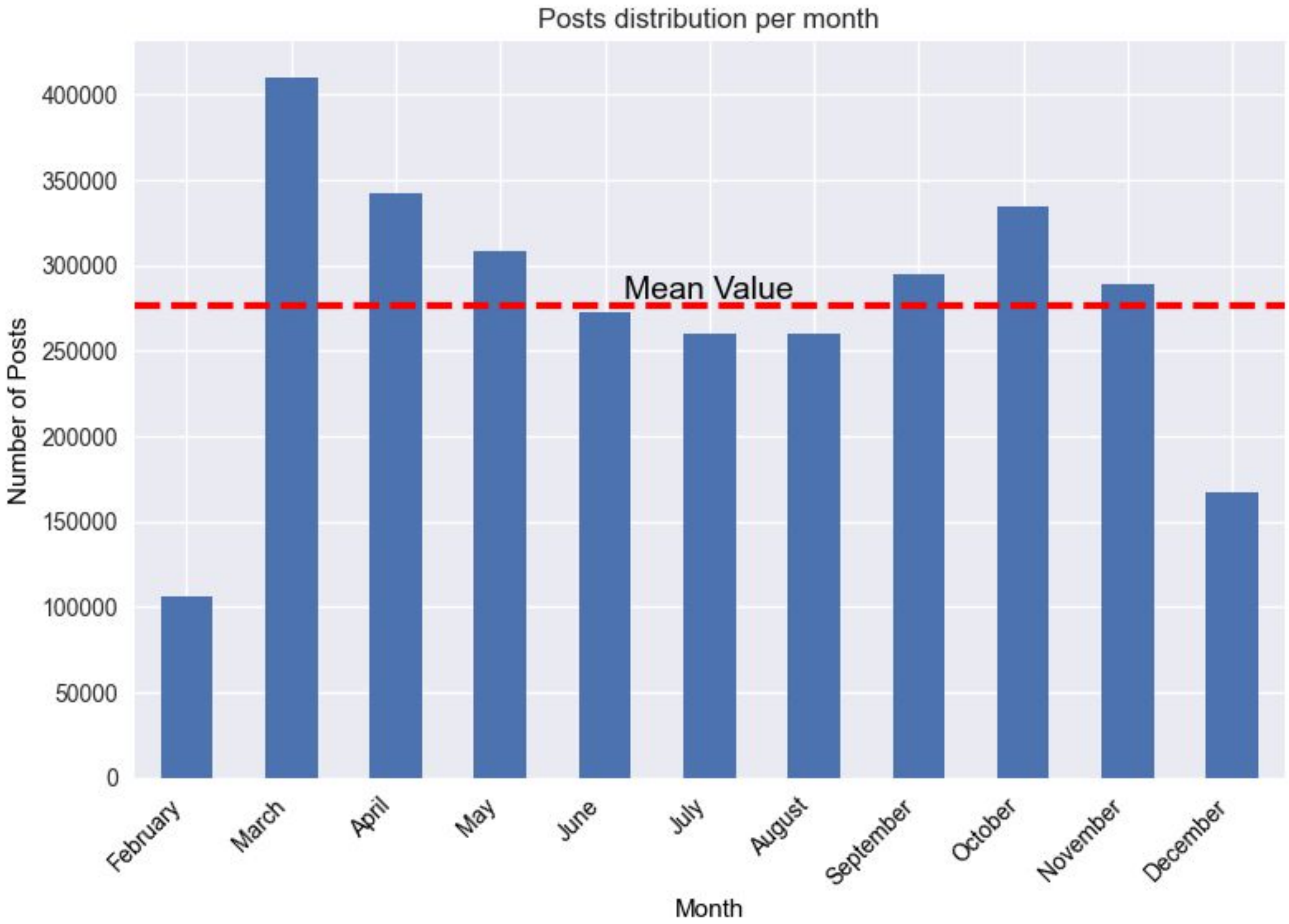

Top 25 channel mean post frequency per month(2017-2022) before war
Channels
karaulny
glavmedia
rian_ru
bbbreaking
SolovievLive
tass_agency
radiogovoritmsk
Ateobreaking
ru2ch_news
izvestia
vityzeva
leylinurimm
new_militarycolumnist
truth_aggregator
vibornyk
rt_russian
ntvnews
HouseOfCardsRussia
infantmilitario
zvezdanews
milinfolive
swodki
navideovidno
mariabutina
donbassr
0
1000
2000
3000
4000
5000
Mean Monthly Post Frequency


Top 25 Channel Post Quantity from February to December 2022
Channels
swodki
karaulny
glavmedia
readovkanews
tass_agency
chtddd
SolovievLive
boris_rozhin
radiogovoritmsk
rt_russian
ssigny
OstashkoNews
rusbrief
rian_ru
truekpru
izvestia
zvezdanews
pravda_shuravi
RVvoenkor
denazi_UA
ru2ch_news
romanov_92
lentadnya
uranews
tv360
0
25000
50000
75000
100000
125000
150000
175000
Post Quantity



Top 25 channel mean post frequency per month from February to December 2022

- What is the distribution of posts across months?


Posts distribution per month
Number of Posts
400000
350000
300000
250000
200000
150000
100000
50000
0
Mean Value
February
March
April
May
June
July
August
September
October
November
December
Month

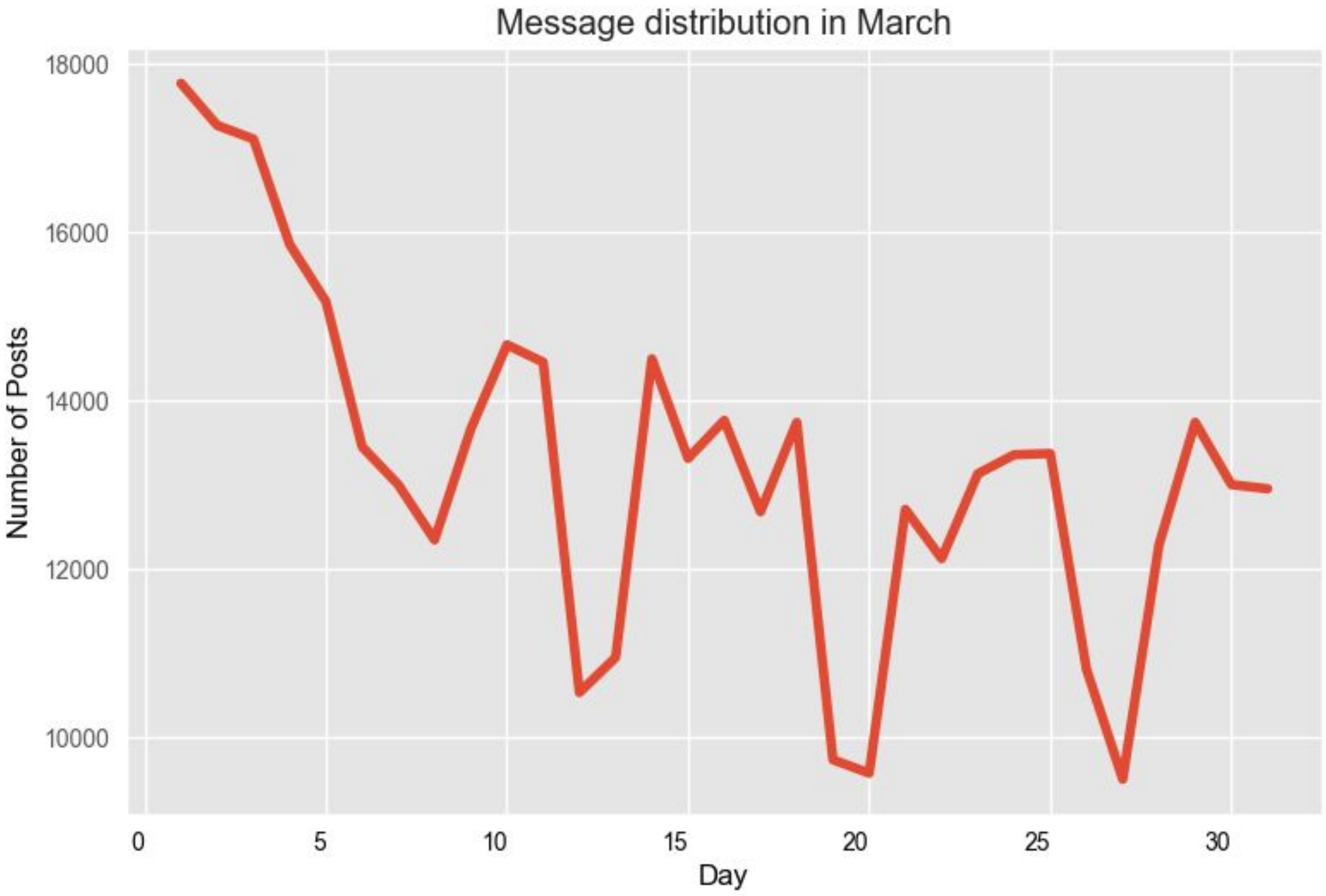
As observed, there are 6 months that surpass the mean threshold value for post activity, with the highest post amounts recorded in March, April, and October, respectively. Notably, these peaks coincide with significant events that likely influenced the surge in post activity.

In March, there was heightened activity due to the commencement of warfare, particularly in the Kyiv region, with notable fights in Bucha, Irpin, and Hostomel, which may explain the increased posts during this period.

Continuing into April and May, conflicts persisted in the East and South of Ukraine, culminating in the extended battle in Mariupol, which garnered substantial attention and continued through May. The conclusion of this conflict, particularly impactful for russia, likely contributed to a surge in posts during this timeframe.

Moreover, in September were counter-offensive operation in the Kharkiv region and russia announced mobilization, a significant event that could have provoked the high post flow observed during that month. In October was first explosion on the Kerch bridge and massive air missiles atacks from russia side. And in November Ukraine managed to get the right-bank occupied districts of the Kherson region back.

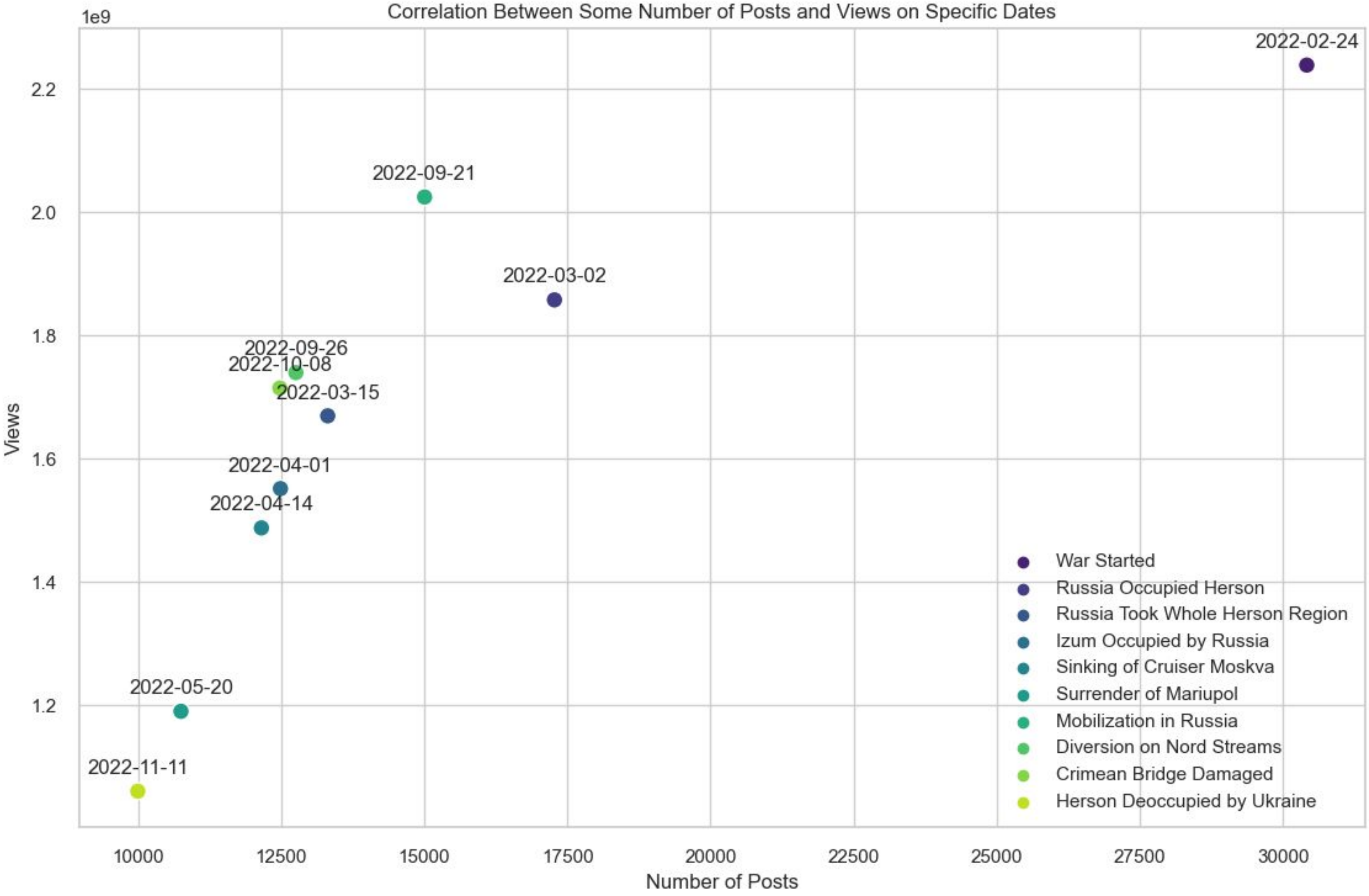
- Let's take a closer look at March as it stands out for high news flow.


Message distribution in March
18000
16000
14000
12000
10000
Number of Posts
0
5
10
15
20
25
30
Day

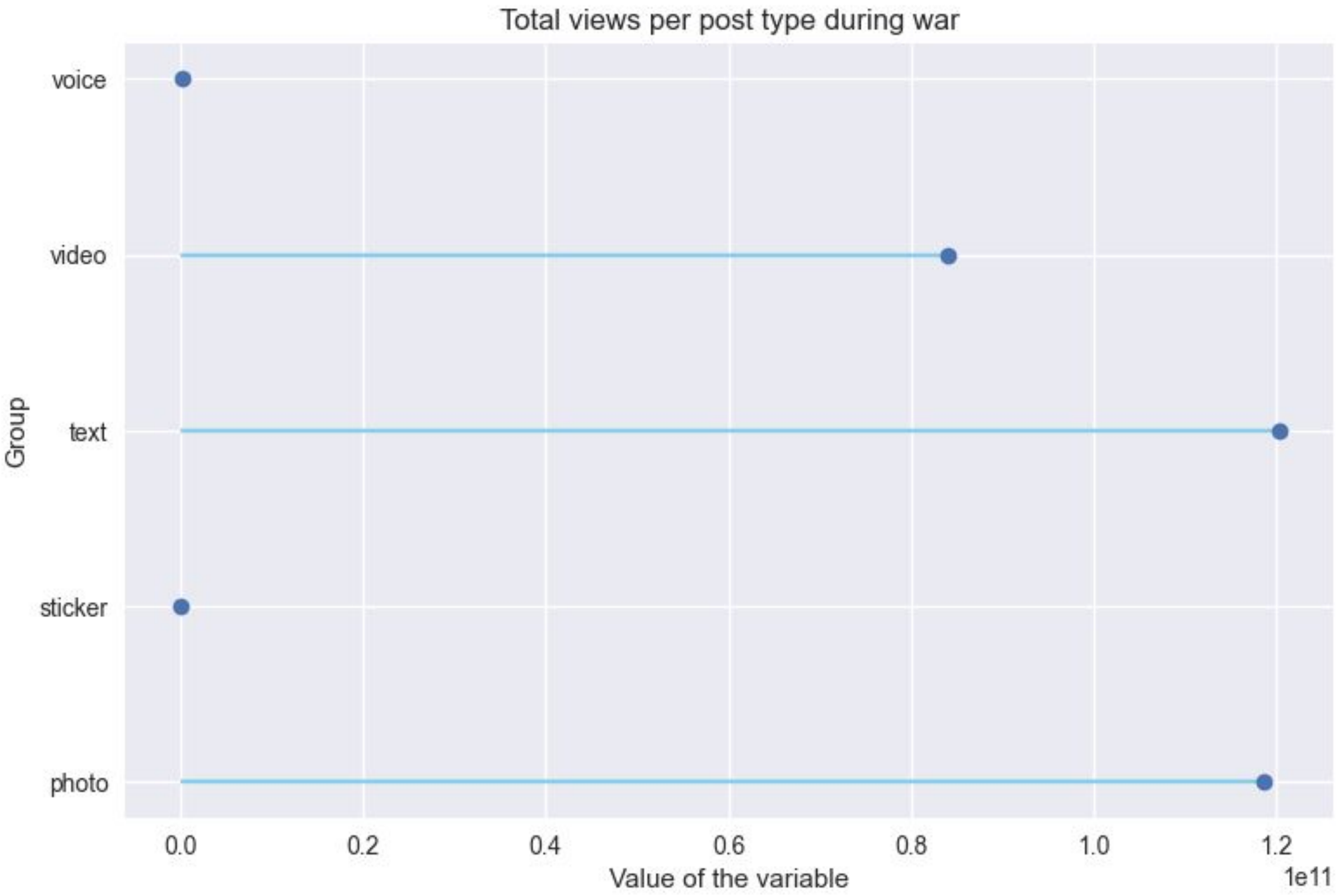
The high news flow was only at the beginning of the month as in this days might have happened most essential events for russian army.

According to the official sources, by the first days of March Russia has taken vast territories and was fighting in main directions toward Kyiv and south region.

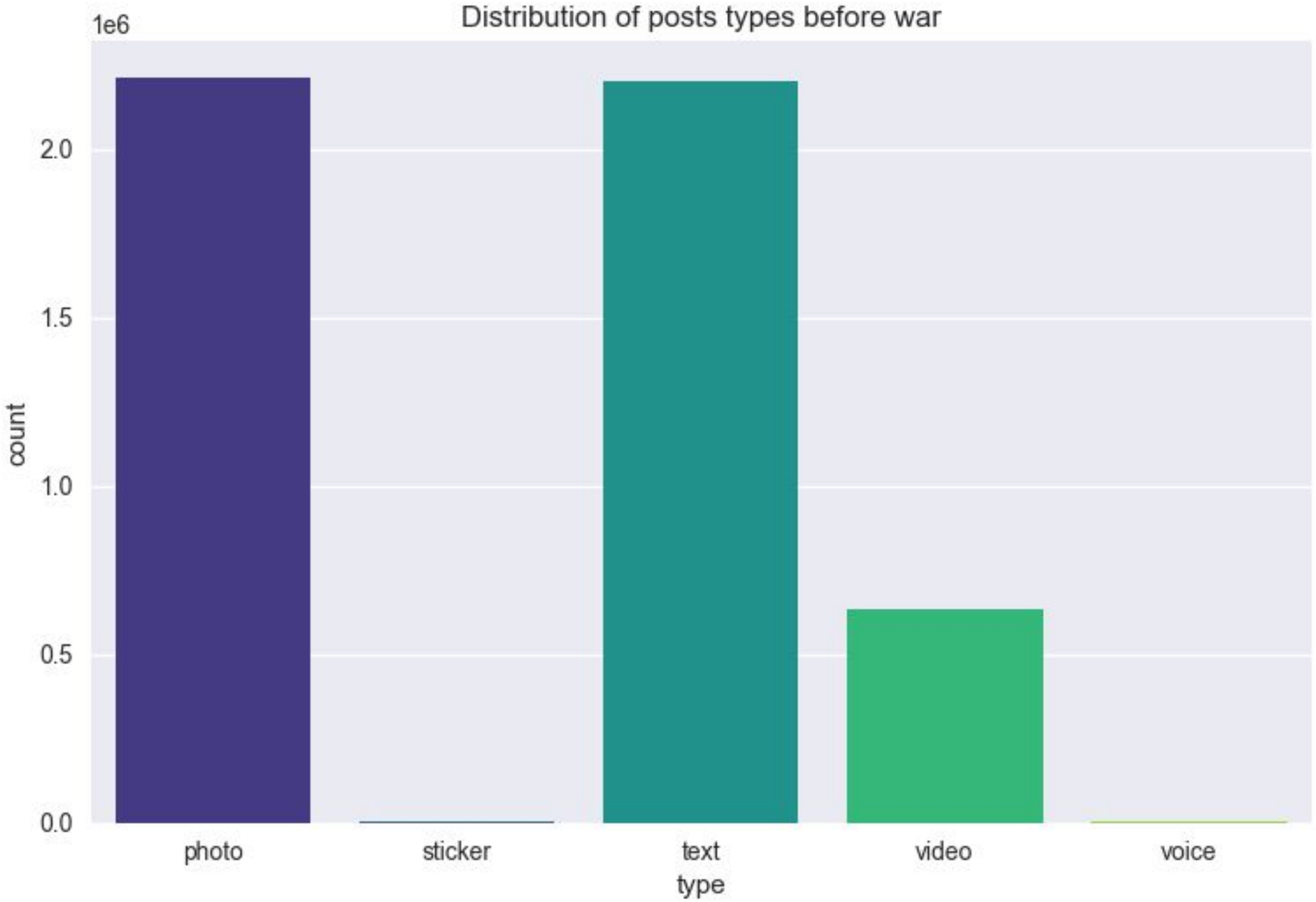
- Can we identify any significant events or spikes in post activity based on specific dates?

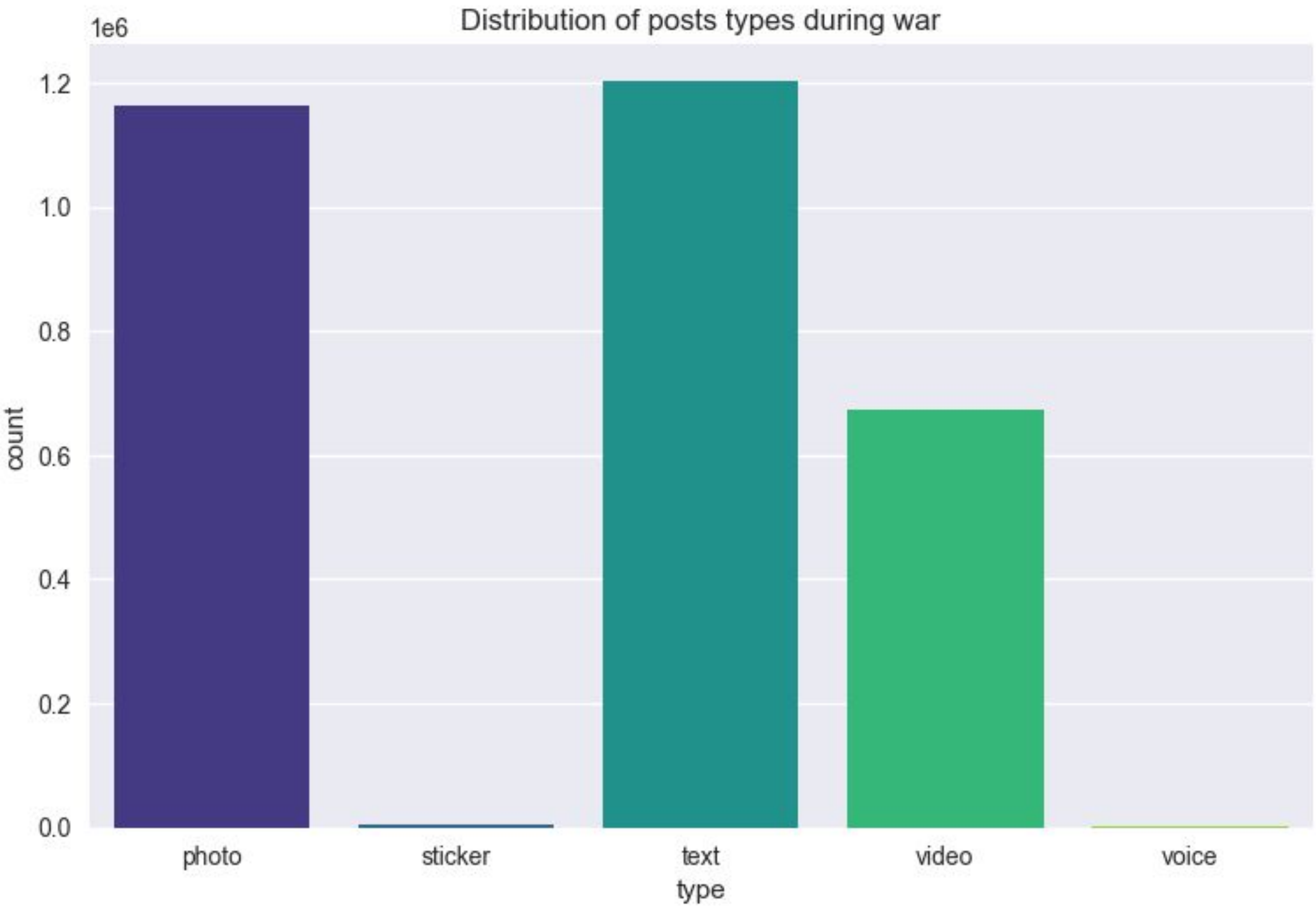

Correlation Between Some Number of Posts and Views on Specific Dates
1e9
2.2
2.0
1.8
1.6
1.4
1.2
Views
10000
12500
15000
17500
20000
22500
25000
27500
30000
Number of Posts
2022-02-24
2022-09-21
2022-03-02
2022-09-26
2022-10-08
2022-03-15
2022-04-01
2022-04-14
2022-05-20
2022-11-11
● War Started
● Russia Occupied Herson
● Russia Took Whole Herson Region
● Izum Occupied by Russia
● Sinking of Cruiser Moskva
● Surrender of Mariupol
● Mobilization in Russia
● Diversion on Nord Streams
● Crimean Bridge Damaged
● Herson Deoccupied by Ukraine

- What is the distribution of views per post types?


Total views per post type during war
Group
voice
video
text
sticker
photo
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1e11
Value of the variable

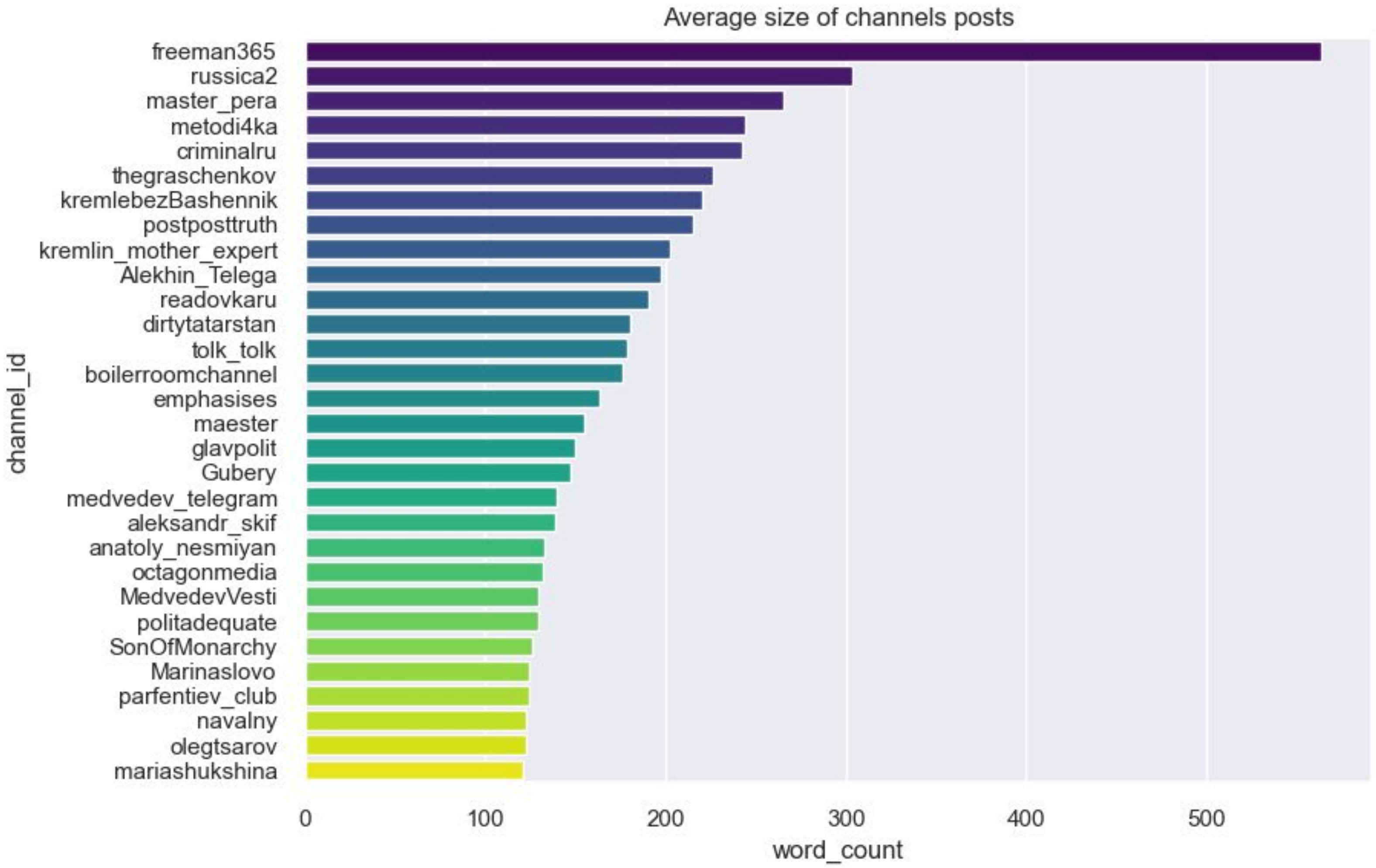
- What is the distribution of posts types?


Distribution of posts types before war
1e6
2.0
1.5
1.0
0.5
0.0
count
photo
sticker
text
video
voice
type

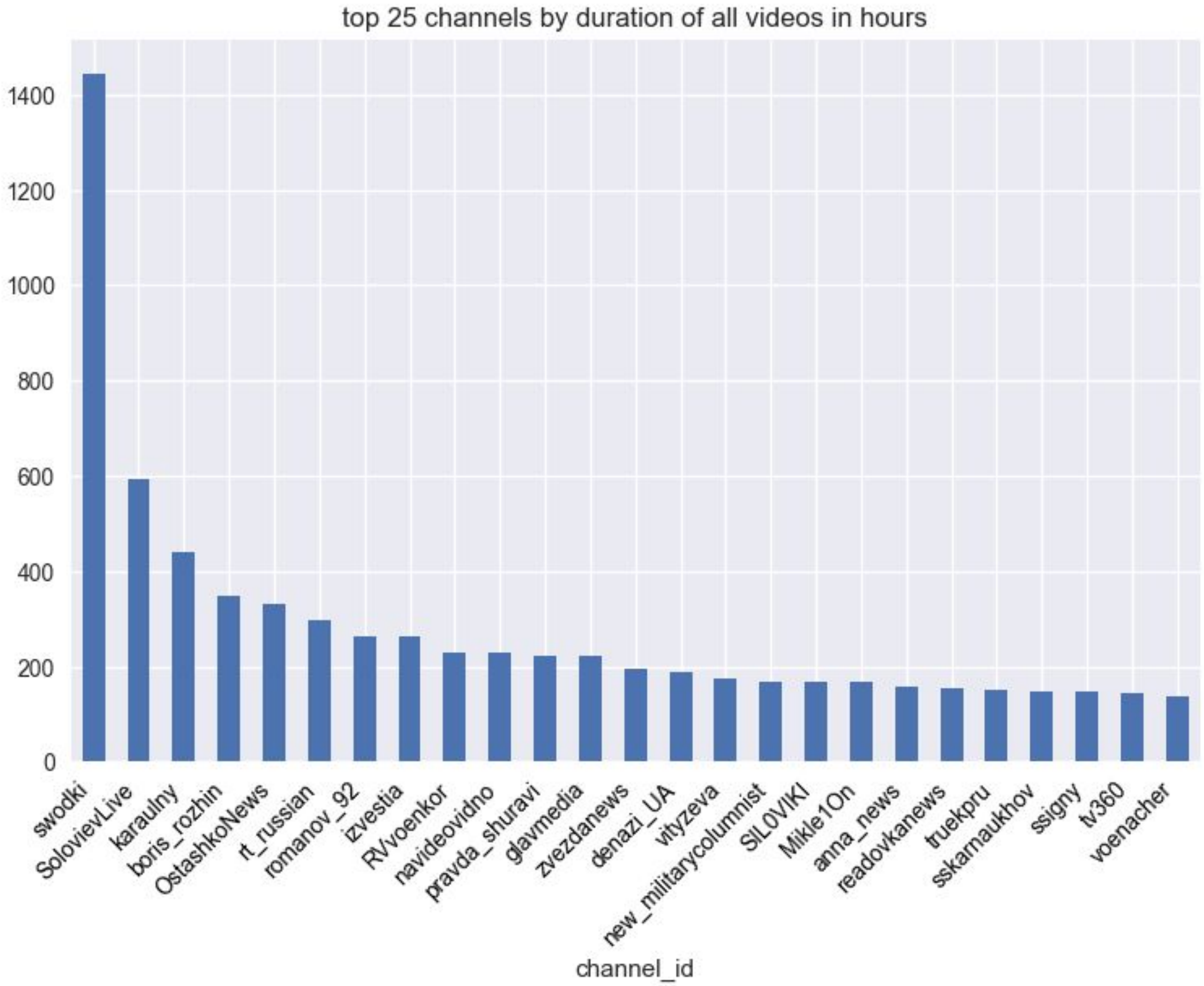


Distribution of posts types during war
1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
count
photo
sticker
text
video
voice
type

- What is the average post size of channels posts?

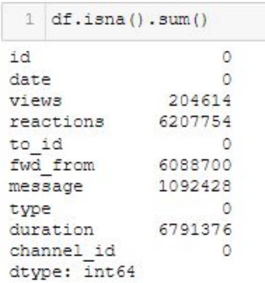

Average size of channels posts
channel_id
freeman365
russica2
master_pera
metodi4ka
criminalru
thegraschenkov
kremlebezBashennik
postposttruth
kremlin_mother_expert
Alekhin_Telega
readovkaru
dirtytatarstan
tolk_tolk
boilerroomchannel
emphasises
maester
glavpolit
Gubery
medvedev_telegram
aleksandr_skif
anatoly_nesmiyan
octagonmedia
MedvedevVesti
politadequate
SonOfMonarchy
Marinaslovo
parfentiev_club
navalny
olegtsarov
mariashukshina
0
100
200
300
400
500
word_count

- What is total duration of all videos per channel?


top 25 channels by duration of all videos in hours
1400
1200
1000
800
600
400
200
0
swodki
SolovievLive
karaulny
boris_rozhin
OstashkoNews
rt_russian
romanov_92
izvestia
RVvoenkor
navideovidno
pravda_shuravi
glavmedia
zvezdanews
denazi_UA
vityzeva
new_militarycolumnist
SIL0VIKI
Mikle1On
anna_news
readovkanews
truekpru
sskarnaukhov
ssigny
tv360
voenacher
channel_id

- How the file duration(audio video) might influence engagement(views)?

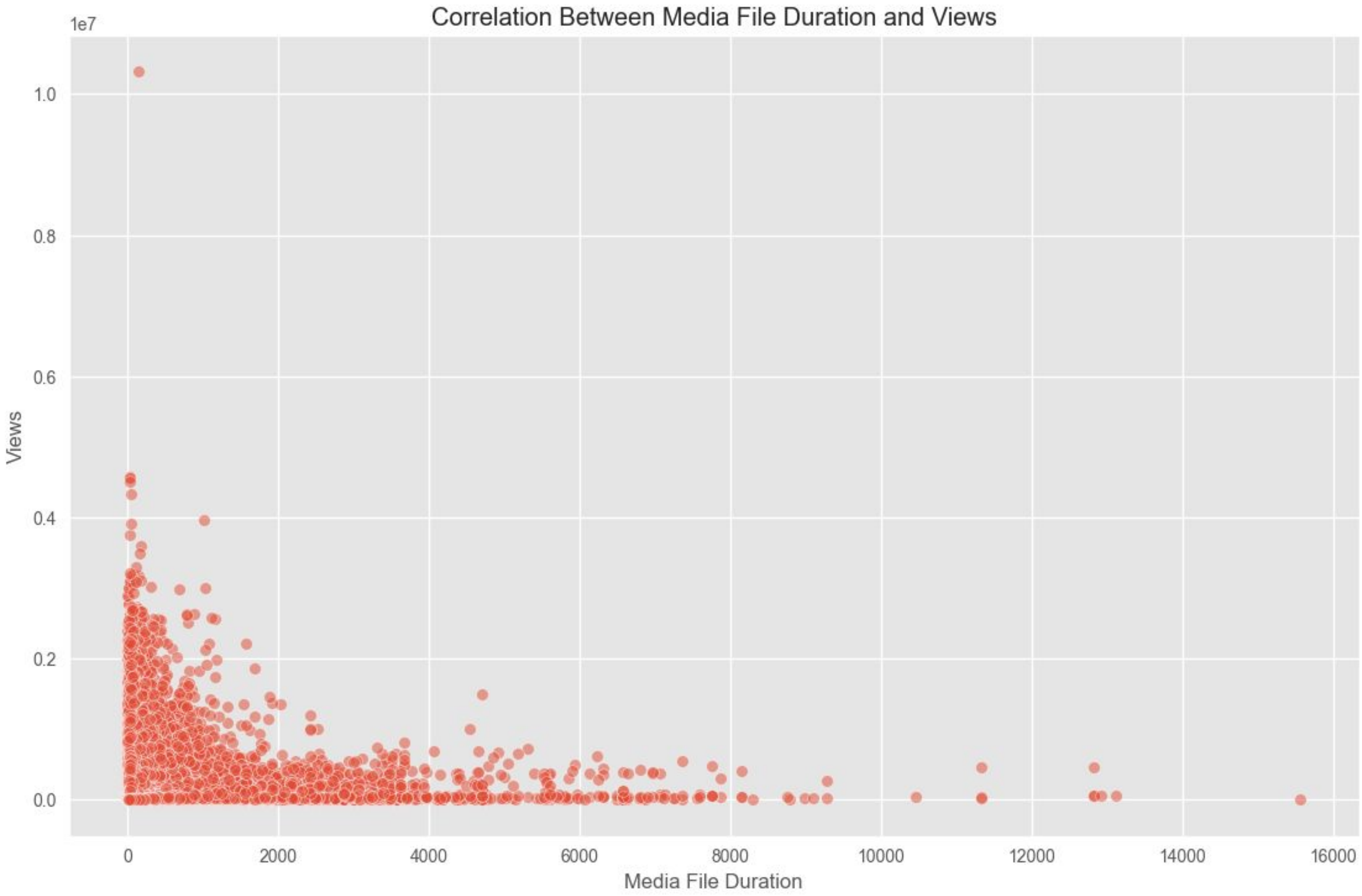
  reactions under posts are not included because 66% of all reactions are lost(NaN values)



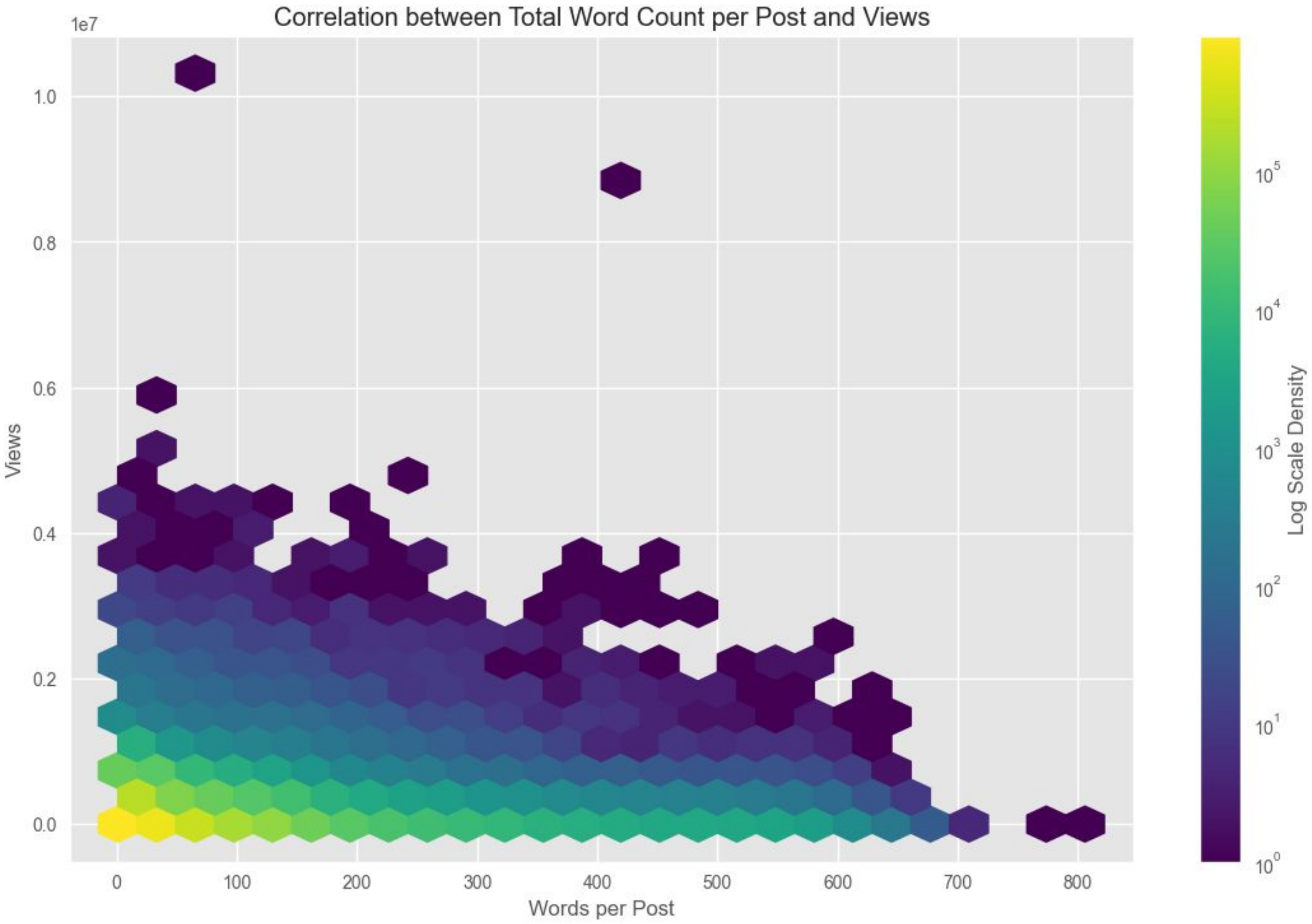
```
df.isna().sum()
```

```
id                  0
date                0
views          204614
reactions     6207754
to_id               0
fwd_from      6088700
message       1092428
type                0
duration      6791376
channel_id          0
dtype: int64
```

here is a screenshot that shows distribution of lost data
in columns of dataframe


Correlation Between Media File Duration and Views
1e7
1.0
0.8
0.6
0.4
0.2
0.0
Views
0
2000
4000
6000
8000
10000
12000
14000
16000
Media File Duration

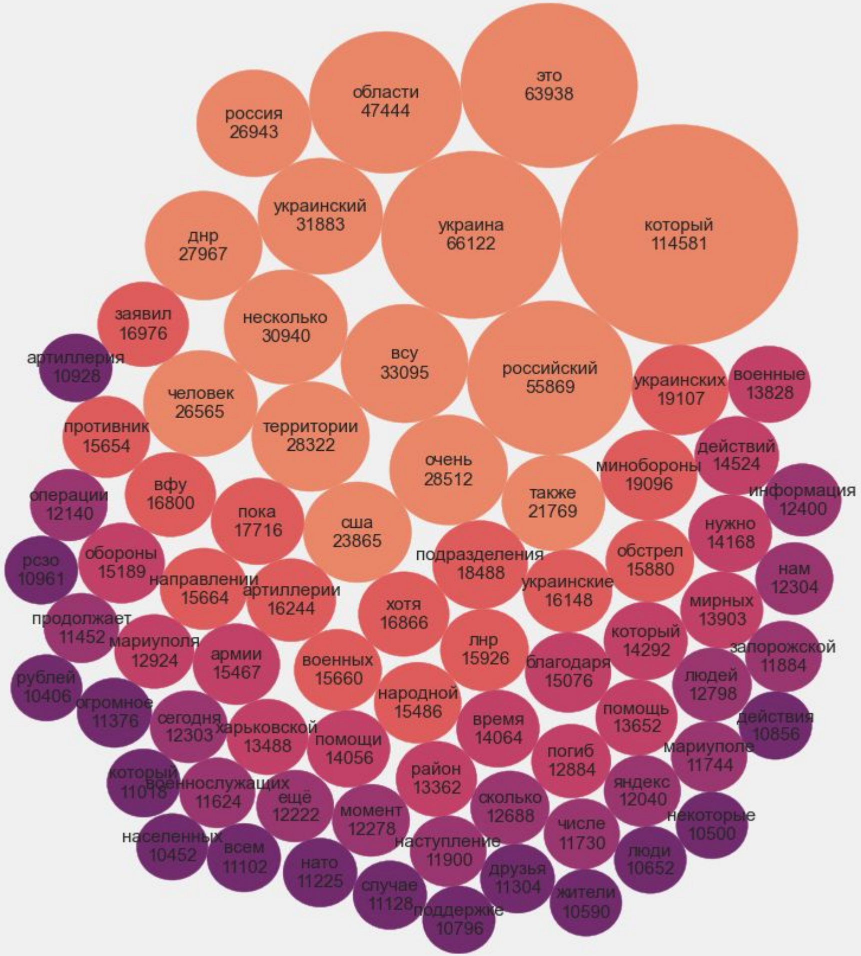
- How post size might influence engagement?


Correlation between Total Word Count per Post and Views
1e7
1.0
0.8
0.6
0.4
0.2
0.0
Views
0
100
200
300
400
500
600
700
800
Words per Post
10^0
10^1
10^2
10^3
10^4
10^5
Log Scale Density

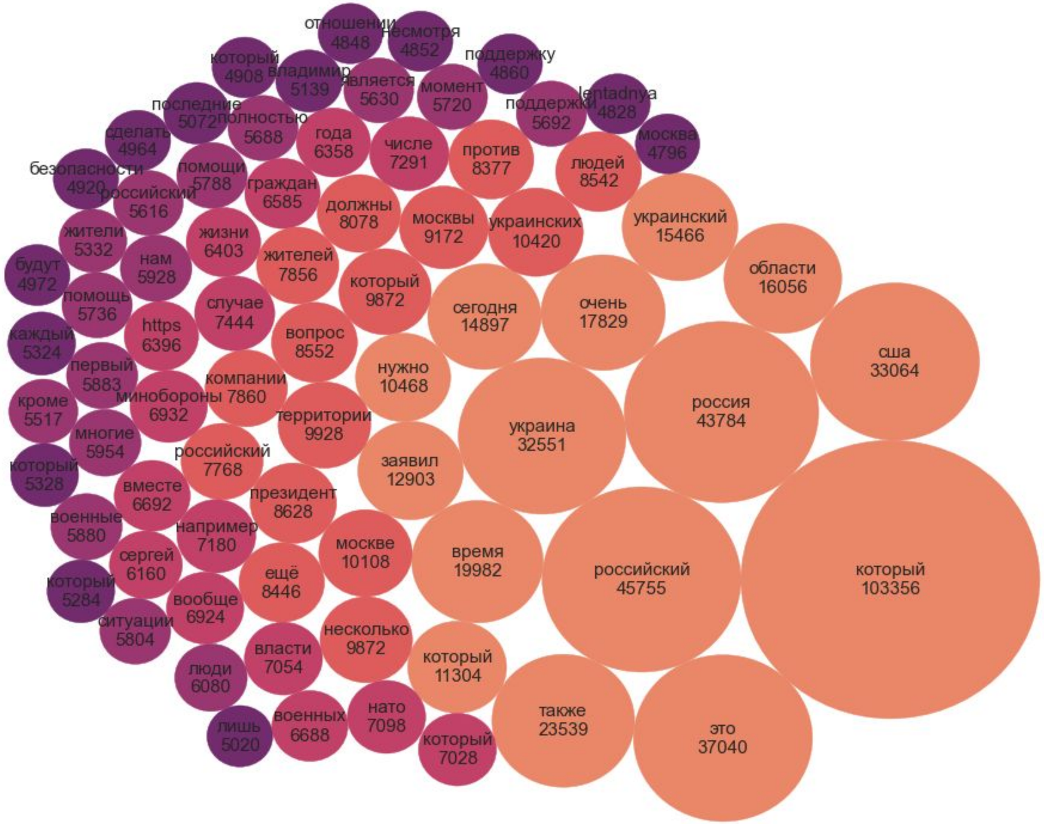
- What are the most frequently used words in the posts throughout the warfare?

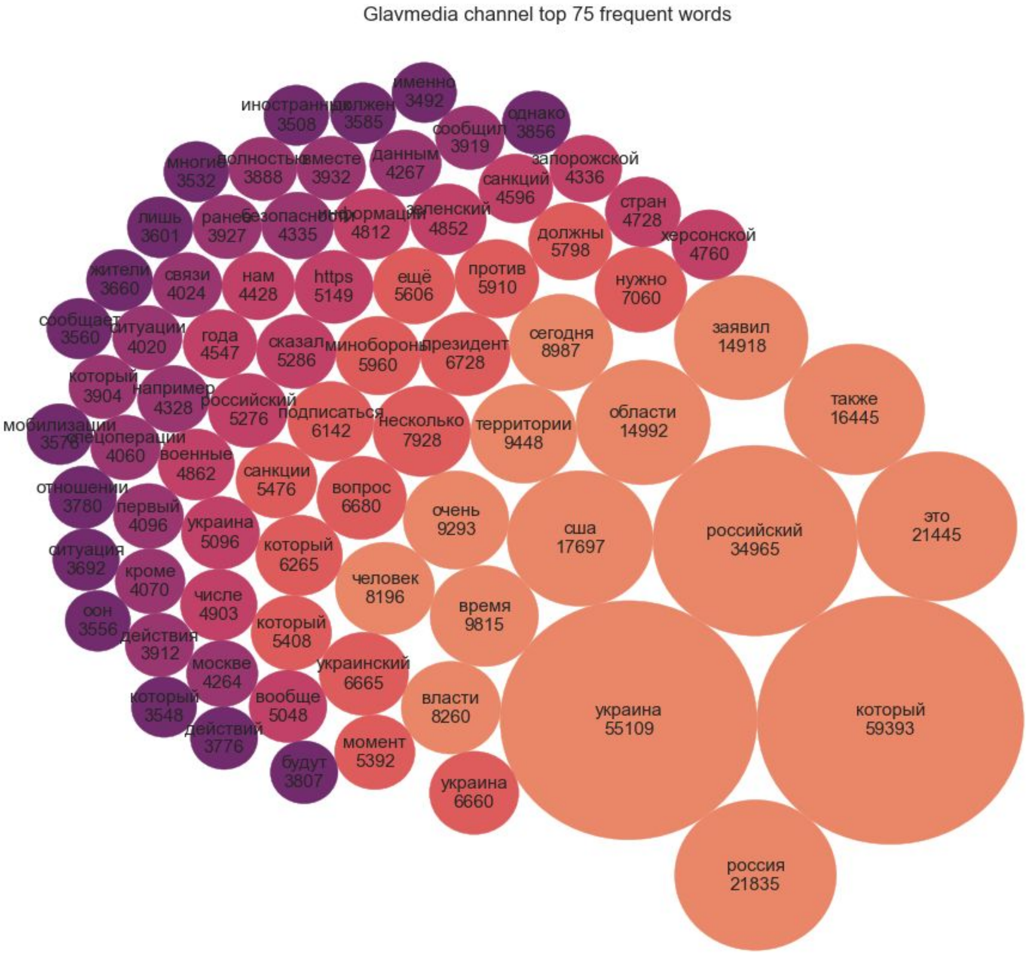
  Let's take the most active channels, swodki, karaulny and glavmedia, and based on their posts investigate the question.

Swodki channel top 75 frequent words

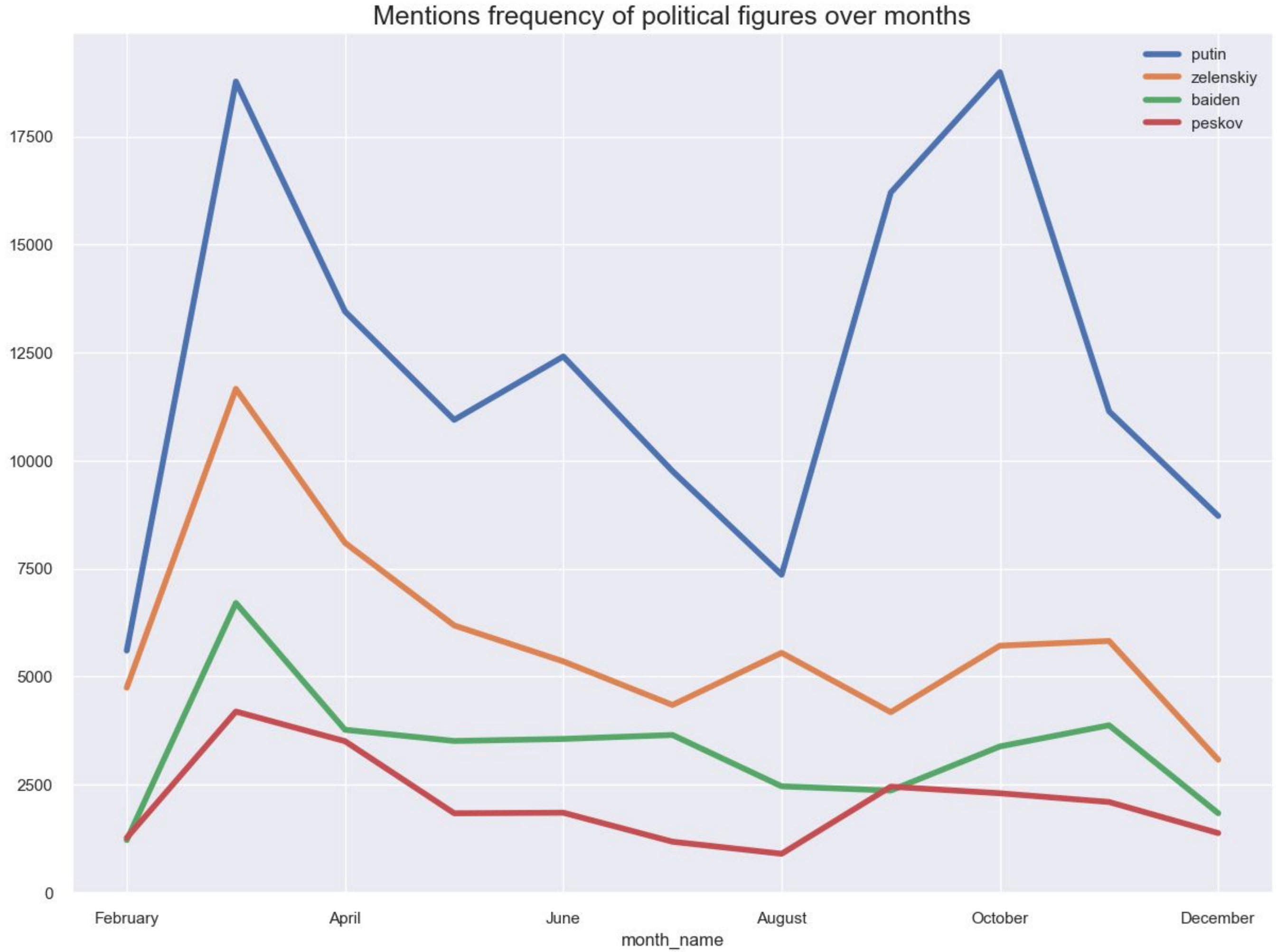


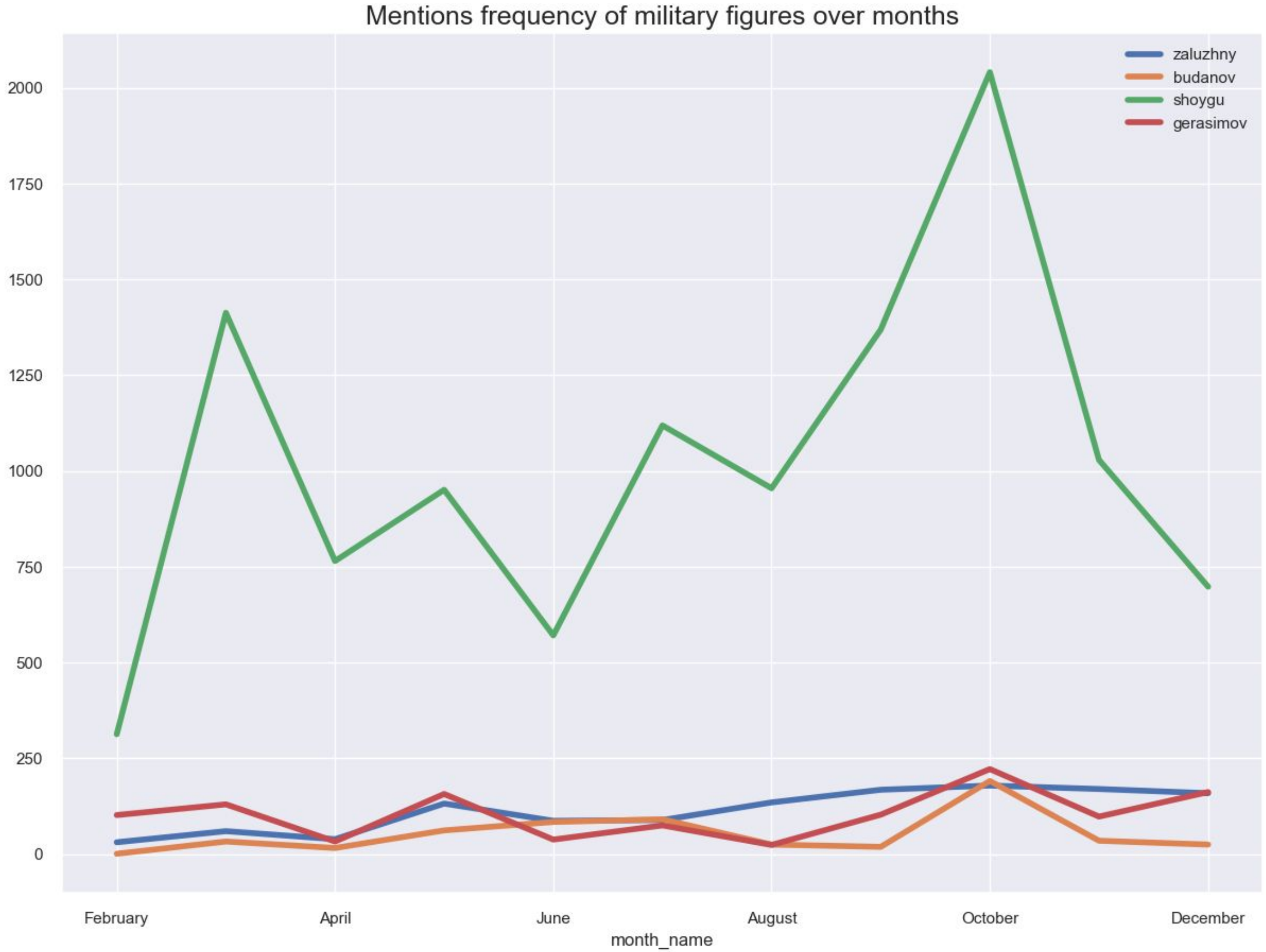
Karaulny channel top 75 frequent words

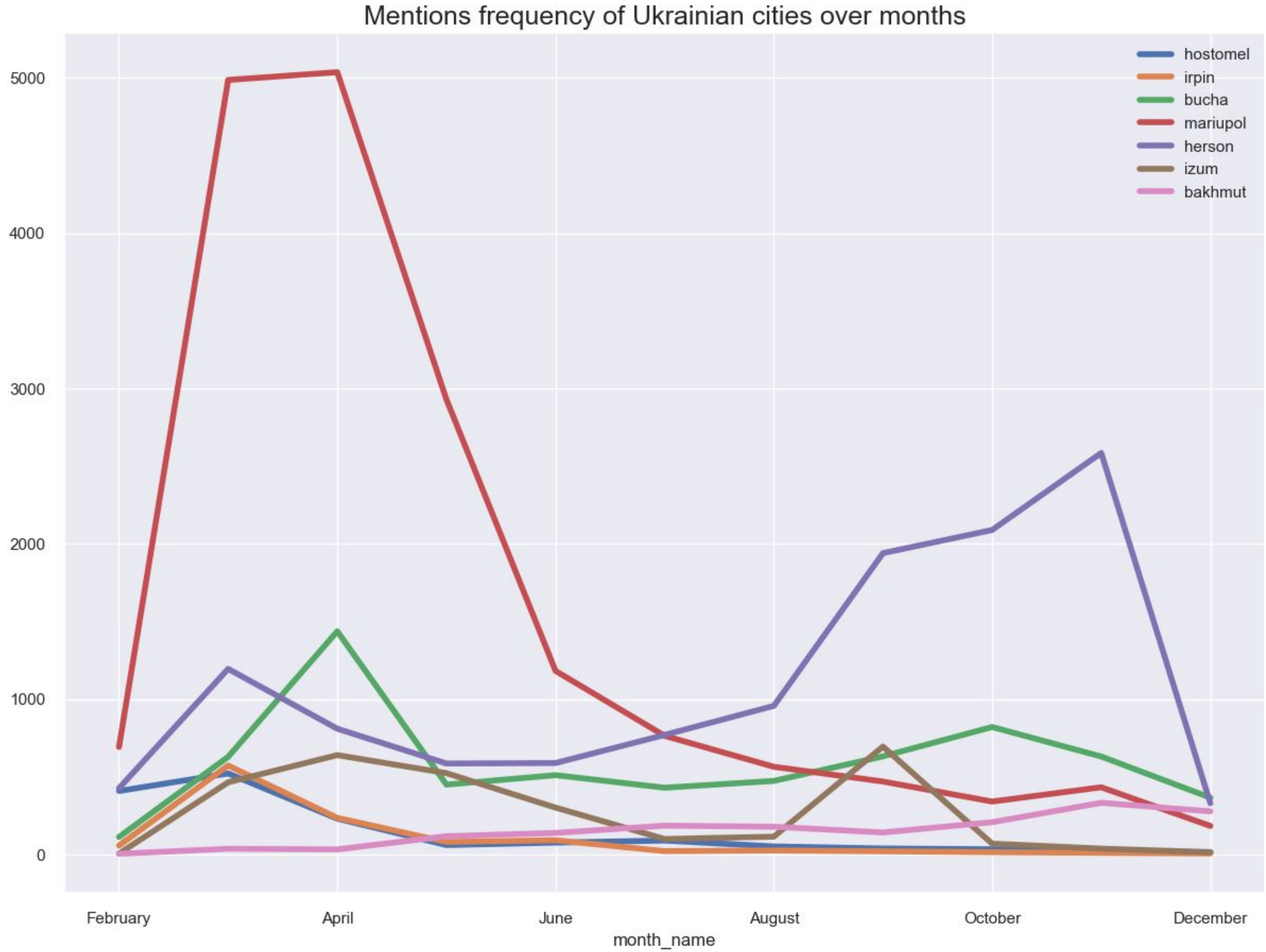



Glavmedia channel top 75 frequent words
именно 3492
иностранных 3508
должен 3585
однако 3856
сообщил 3919
многие 3532
полностью 3888
вместе 3932
данным 4267
санкций 4596
запорожской 4336
лишь 3601
ранее 3927
безопасности 4335
информации 4812
зеленский 4852
должны 5798
стран 4728
херсонской 4760
жители 3660
связи 4024
нам 4428
https 5149
ещё 5606
против 5910
нужно 7060
сообщает 3560
ситуации 4020
года 4547
сказал 5286
минобороны 5960
президент 6728
сегодня 8987
заявил 14918
который 3904
например 4328
российский 5276
подписаться 6142
несколько 7928
территории 9448
области 14992
также 16445
мобилизации 3576
спецоперации 4060
военные 4862
санкции 5476
вопрос 6680
отношении 3780
первый 4096
украина 5096
который 6265
очень 9293
сша 17697
российский 34965
это 21445
ситуация 3692
кроме 4070
числе 4903
человек 8196
время 9815
оон 3556
действия 3912
который 5408
москве 4264
украинский 6665
который 3548
вообще 5048
власти 8260
украина 55109
который 59393
действий 3776
момент 5392
будут 3807
украина 6660
россия 21835

- What is the dynamic change of some words, such as Ukrainian cities, political or military figures in whole dataset?


Mentions frequency of political figures over months
putin
zelenskiy
baiden
peskov
17500
15000
12500
10000
7500
5000
2500
0
February
April
June
August
October
December
month_name


Mentions frequency of military figures over months
zaluzhny
budanov
shoygu
gerasimov
2000
1750
1500
1250
1000
750
500
250
0
February
April
June
August
October
December
month_name


Mentions frequency of Ukrainian cities over months
hostomel
irpin
bucha
mariupol
herson
izum
bakhmut
5000
4000
3000
2000
1000
0
February
April
June
August
October
December
month_name

Further work

I have minimum three ideas what to explore in the dataset:

- Do successes or failures of russians provoke more reactions in posts?
- Co-Occurrence Plot Analysis
- Find a distribution of posts types per channel

Sources

- Link to github repository :

  https://github.com/vladyslavBrothervinn/russian-propaganda-data-investigation

- Other links :

  https://python-graph-gallery.com/

  https://www.data-to-viz.com/

  https://seaborn.pydata.org/index.html

Thank you for your attention!