Telegram Investigation EDA

Author: Michael Fediuchenko

Teacher: Andrew Kurochkin

13.12.2022

Presentation outline

- Introduction
- Data mining
- Data exploration
- Final results
- Further work
- GIT references

Introduction

The main idea of the work was to research information downloaded from Telegram's feed and find regularities between data and real events. Define questions and try to answer them with understandable visualizations.

Data mining

Two codebases were used to get and merge information from Telegram with **Telegram API**:

- https://github.com/SanGreel/telegram-data-collection
- https://github.com/SanGreel/telegram-dialogs-analysis-v2

Let's take a look at the numbers:

- telegram data downloading: **4h 48m**
- size of downloaded data: **1 320 749 817 Б (1,34 ГБ)**

I had a problems with data downloading because used scrips require stable internet connection which was almost unreal in current situation

Data mining

Downloaded data contains information from 2015 about dialogs, groups, channels, and messages sent and received in my Telegram profile:

- 2 776 887 messages
  - 100 013 sent messages
  - 2 752 512 received messages
    - 2 268 067 text messages
    - 408 888 photos
    - 104 891 video
    - 47 071 stickers
    - 23 608 voice messages
- First message 2015-11-01 21:40:16
- Last downloaded message 2022-12-07 22:14:40
- The total duration of video and voice messages 4950399
  - Voice messages: 405584.0
  - Video : 4544815.0
- Groups and Chats:
  - Private dialogs: 326
  - Group: 120
  - Channel: 5
- Unique users: 18086

Data exploration

First of all I need to investigate the data which I get from Telegram API:

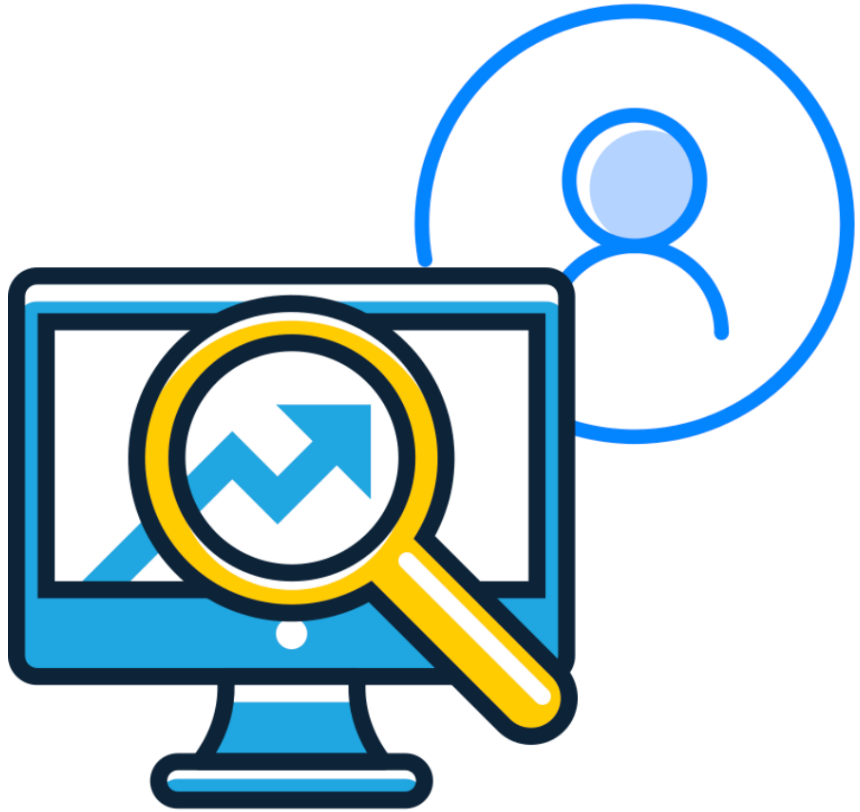
1. Took a look at every column and every table
2. Proceed with data preprocessing and clean up
3. Filled in missing data and make columns to have the same type

Only after this I came up with questions which could be answered with those amount and type of data that I had

Data exploration

Singled out the main areas of work:

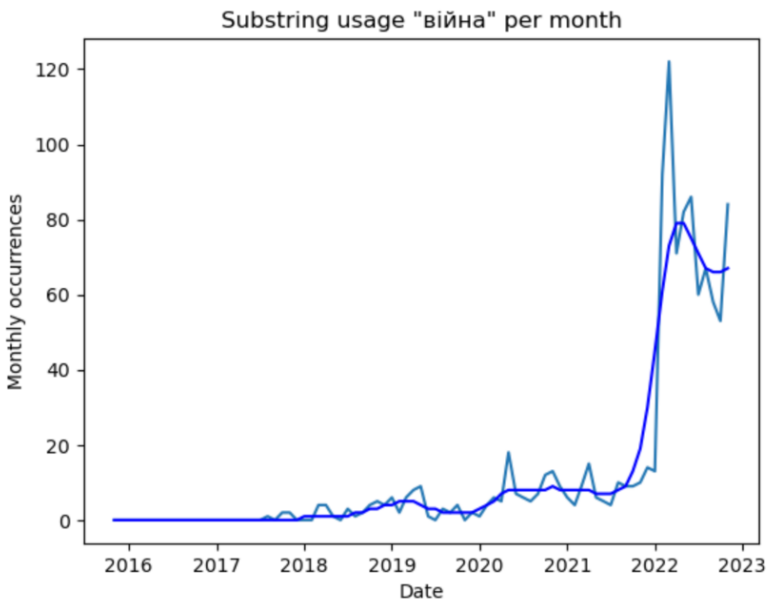
- Differences in user's behavior depending on user's sex
- Trends in sending different types of messages
- Correlation between historical events and user's interaction
- Trending substrings

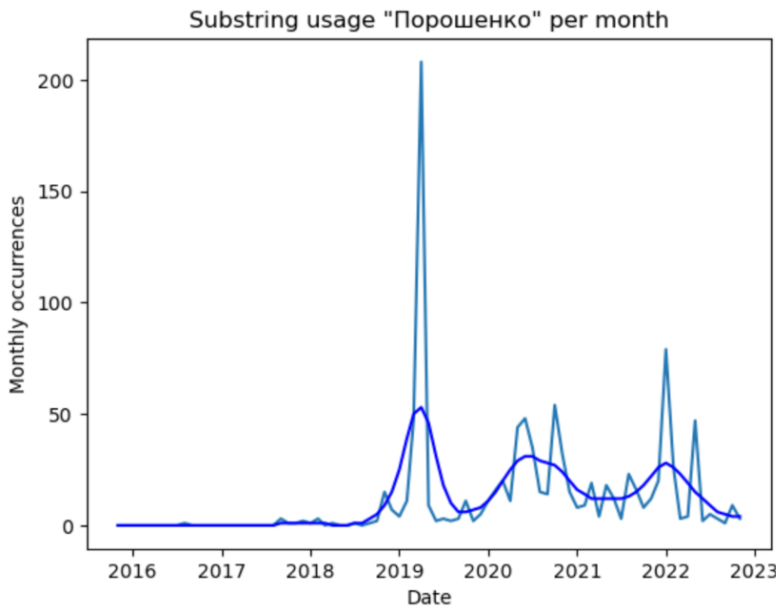


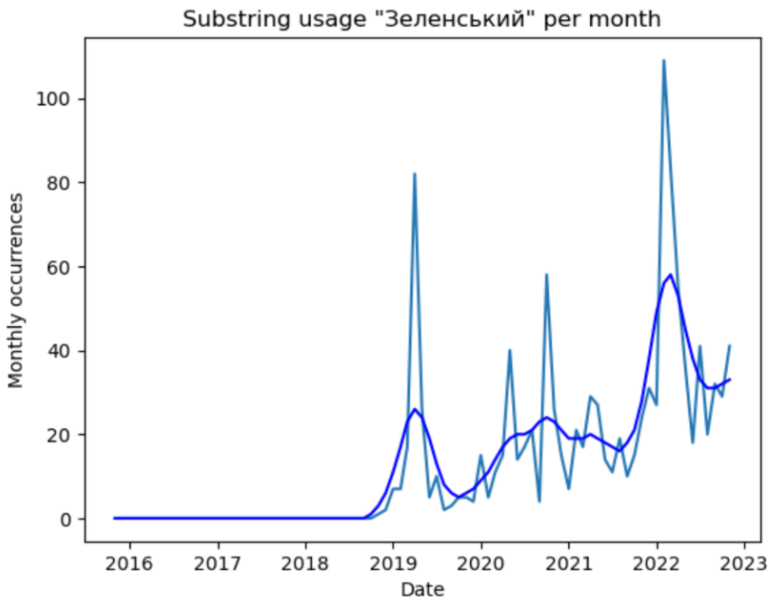
Data exploration

Question: The popularity of selected substrings during the time?

The **formatted date** column was used with the **count function** on the **message column**. To have smothed data I was using **1D Gaussian Smoothing**

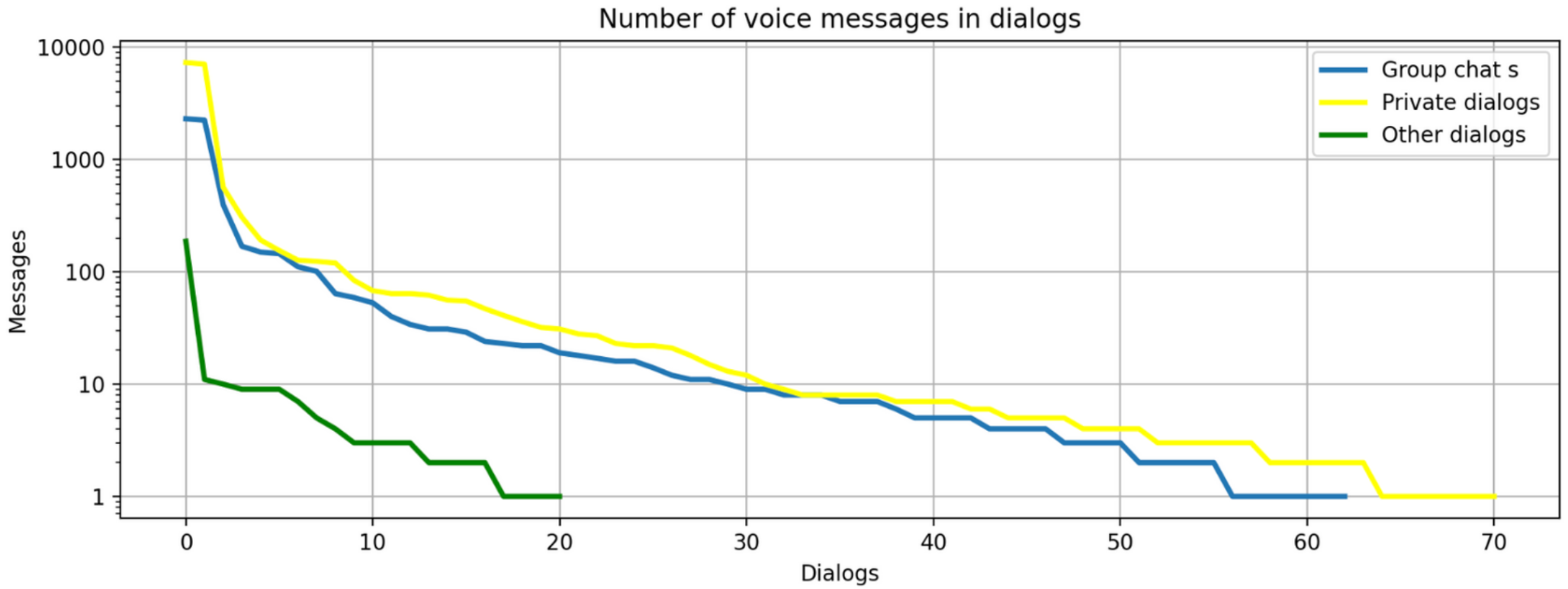






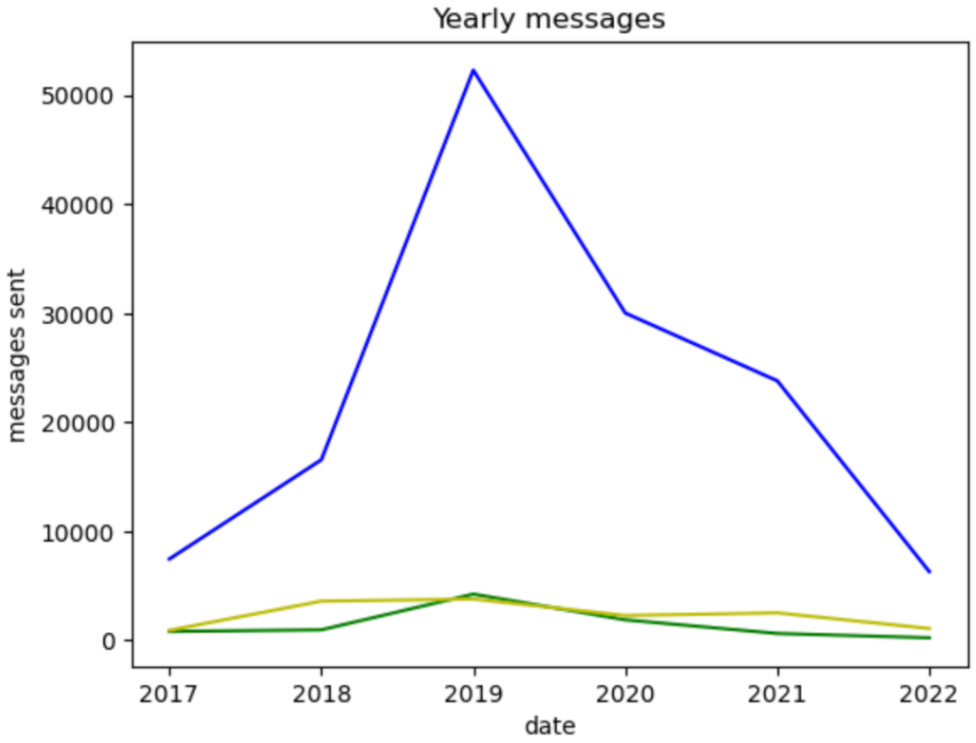
Data exploration

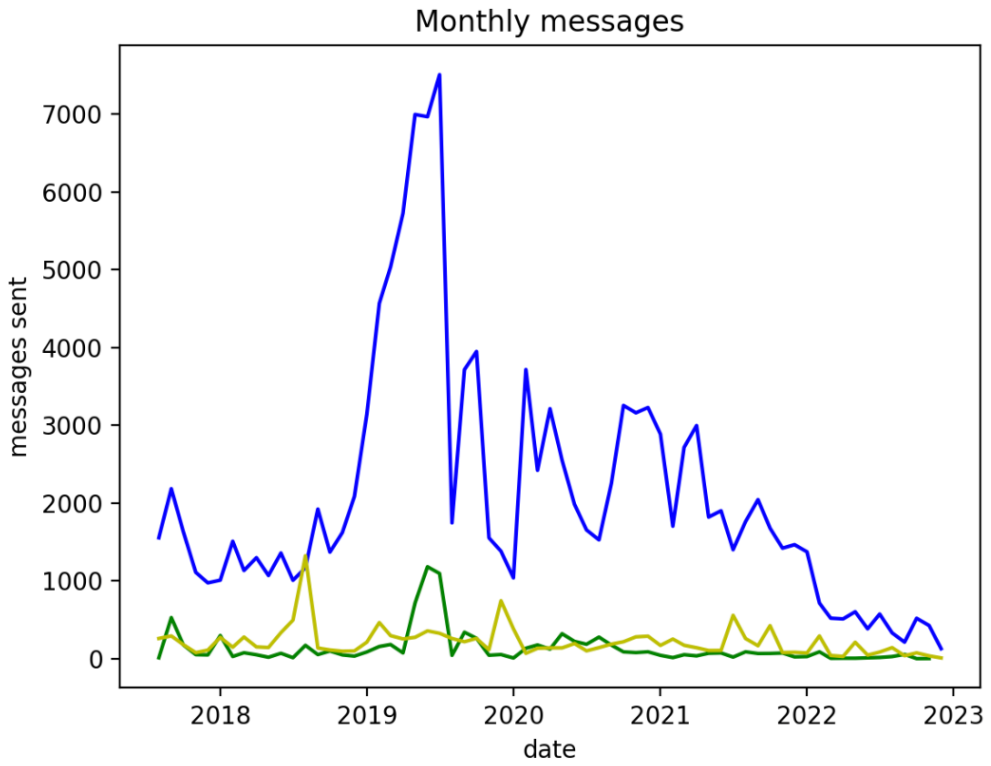
Question: How different is the amount of voice messages in dialogs, channels, and group channels?
It is noticeable how people are using voice messages in private chats more than in group.



Data exploration

Question: How my activity in TG looks like for the different type of messages?
It is noticeable that activity in Telegram decreased after 2019, which I think connected ti the fact that in 2019 I have found my first job





Data exploration

User data analysis:

- Users' distribution by sex
- Correlation of activity in telegram and COVID-19
- Correlation of vaccine names usage and COVID-19 pandemia
- Name popularity

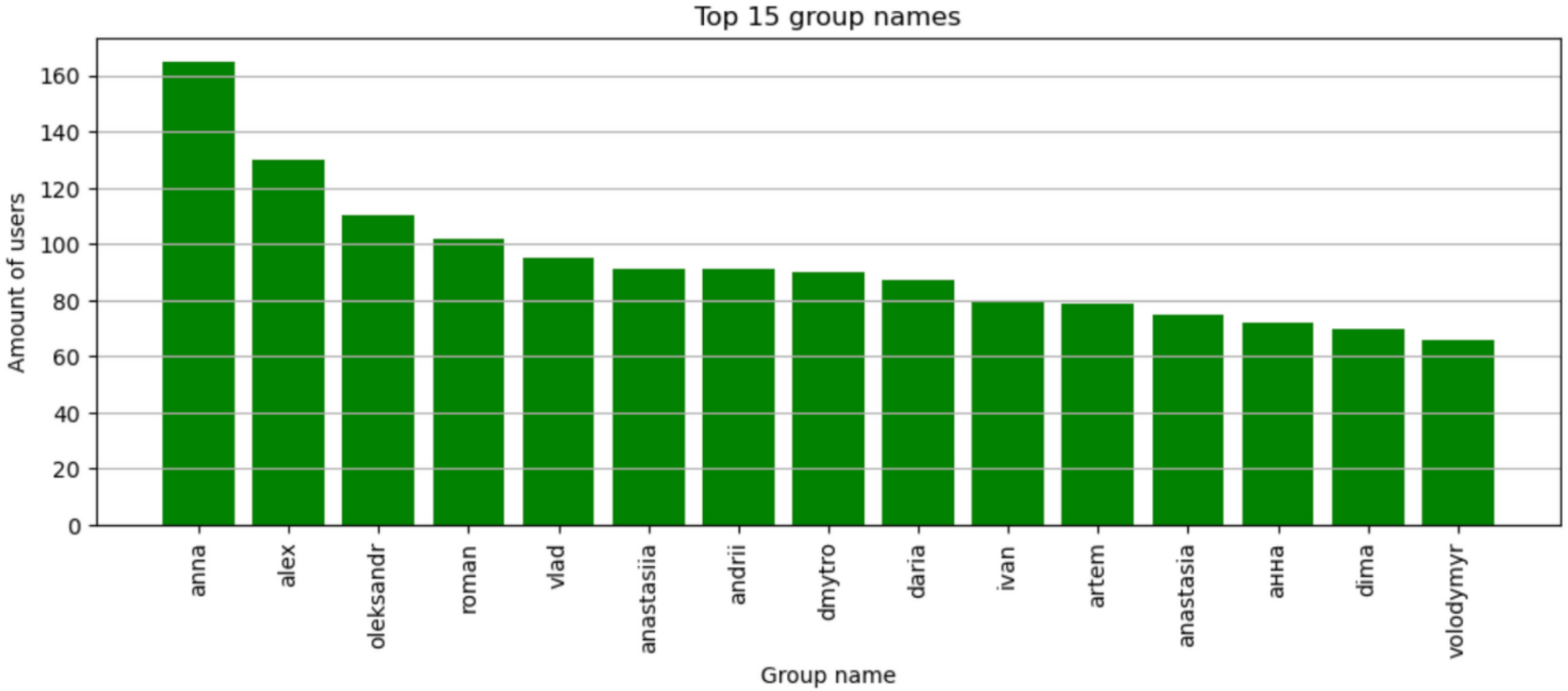
	name	t_name
10	валентина	valentine
11	людмила	ludmyla
12	галина	branch
13	ніна	Nina
14	катерина	Catherine
...	...	...
3382	эрнестина	ernestina
3383	эсмеральда	esmeralda
3384	юлиания	Juliania
3385	юния	Junia
3386	юстина	justina

	name	t_name
0	олександр	Alexander
1	Микола	Mykola
2	іван	Ivan
3	Сергій	Sergii
4	владимир	Vladimir
...	...	...
6227	хрисанф	chrysanthus
6228	эразм	Erasmus
6229	юстиниан	justinian
6230	януарий	januarium
6231	ярополк	Yaropolk

Data exploration

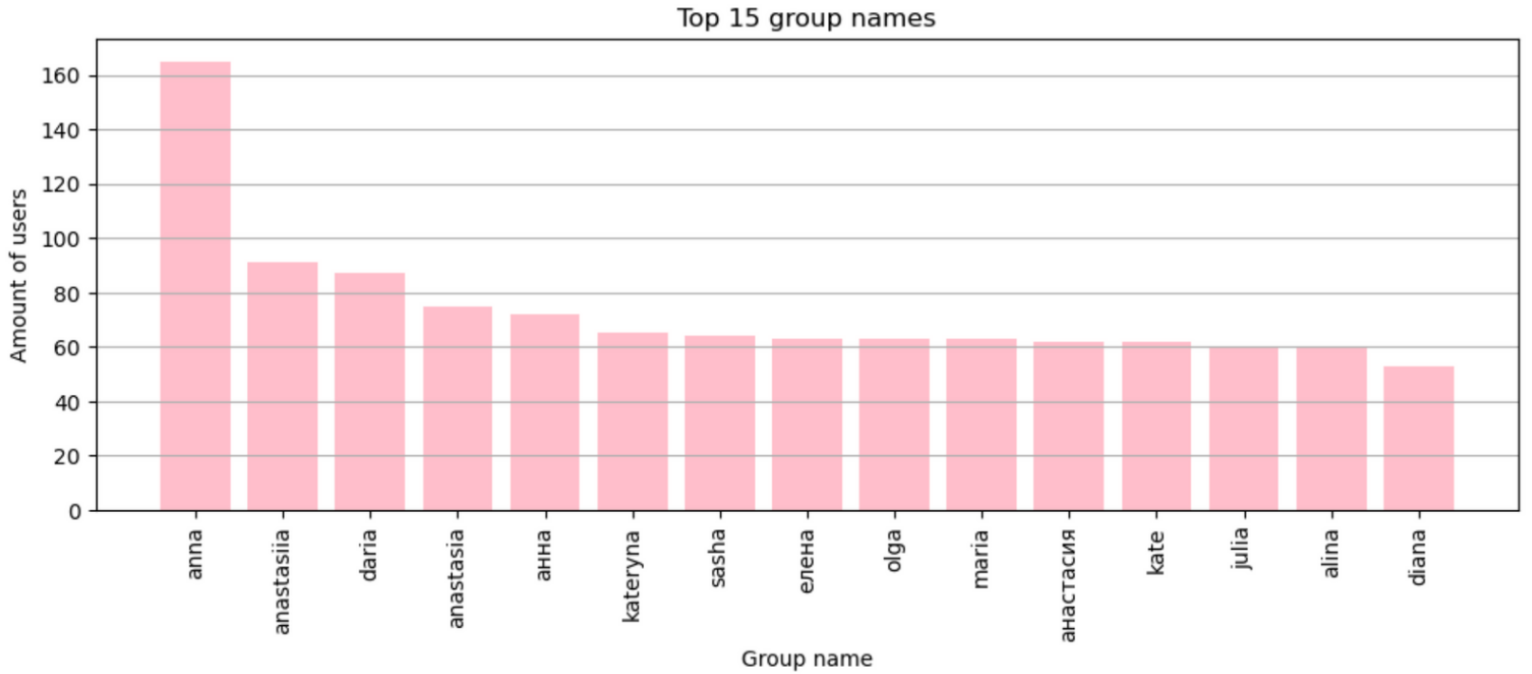
Defining the sex of user by first name mapping:

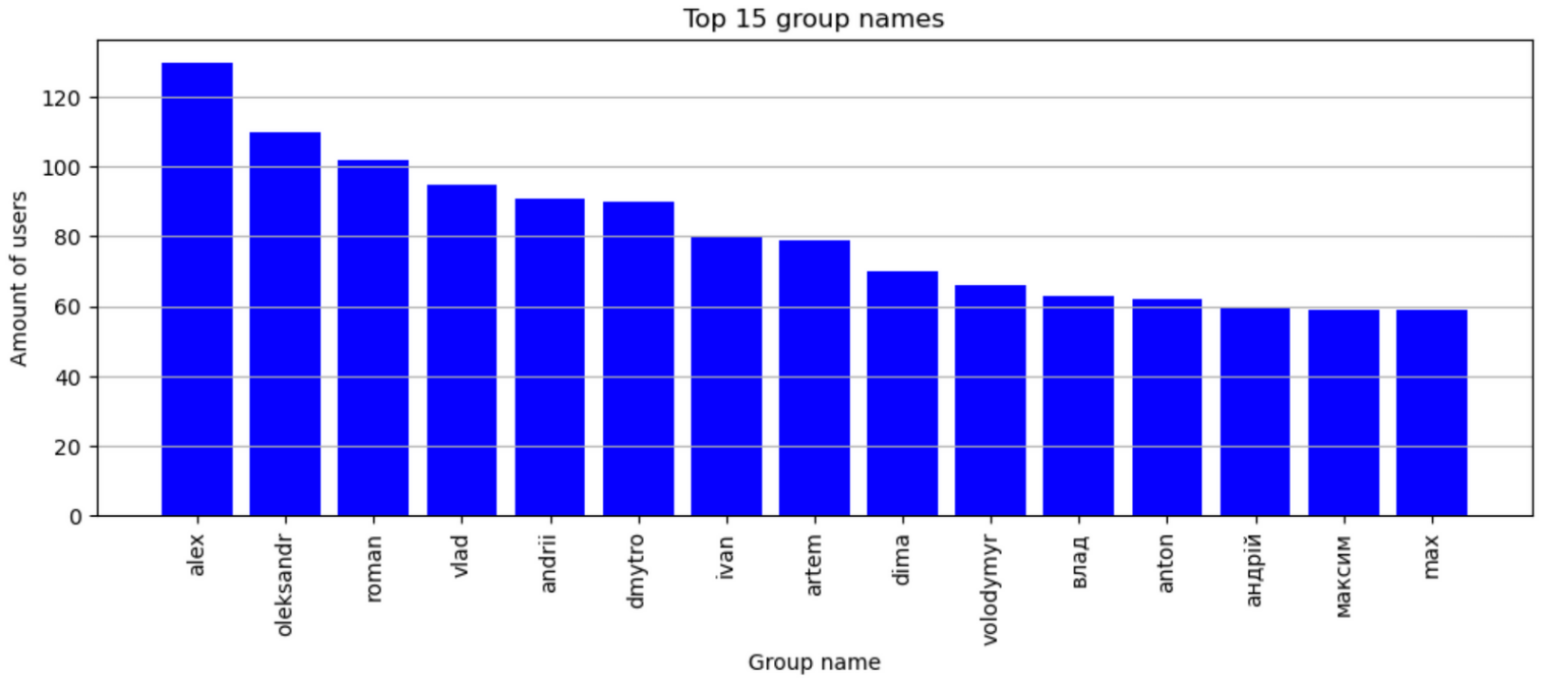
- Used dictionary for female and male names. from https://github.com/SanGreel/tone-dict-ukrainian
- Users' first name formatting and cleaning from non alpha characters
- Manual adding new names by defining it's sex
- Defined 1195 unique group of names



Data exploration

For different sex groups:

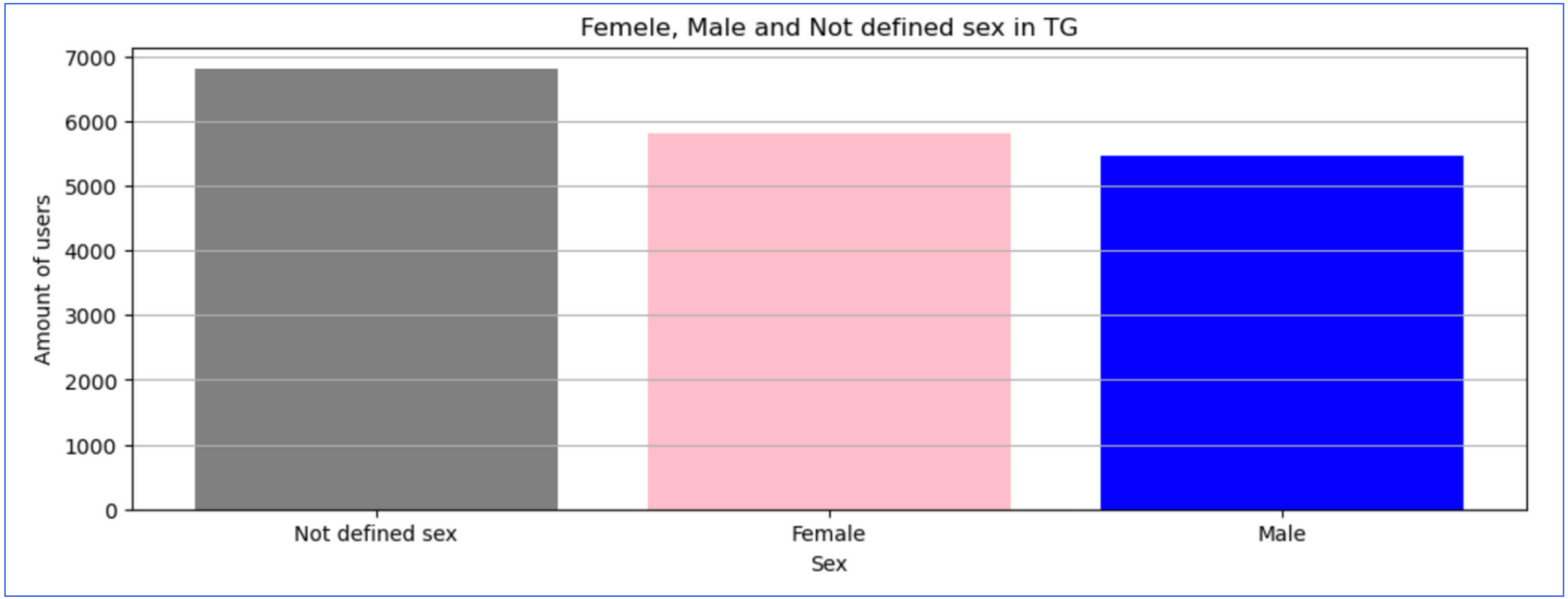




Data exploration

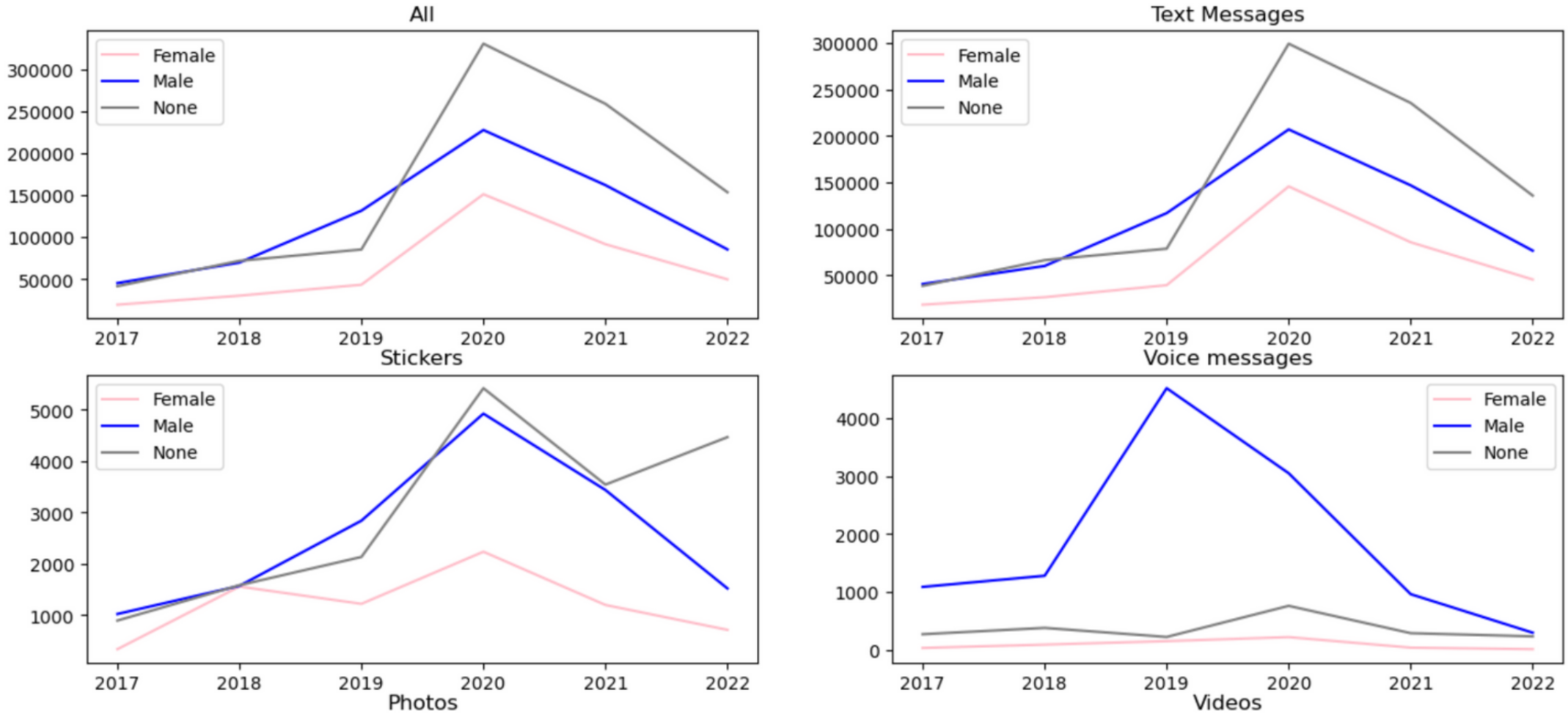
Total amount of users: **18 086**

Users with defined sex: **11 284**



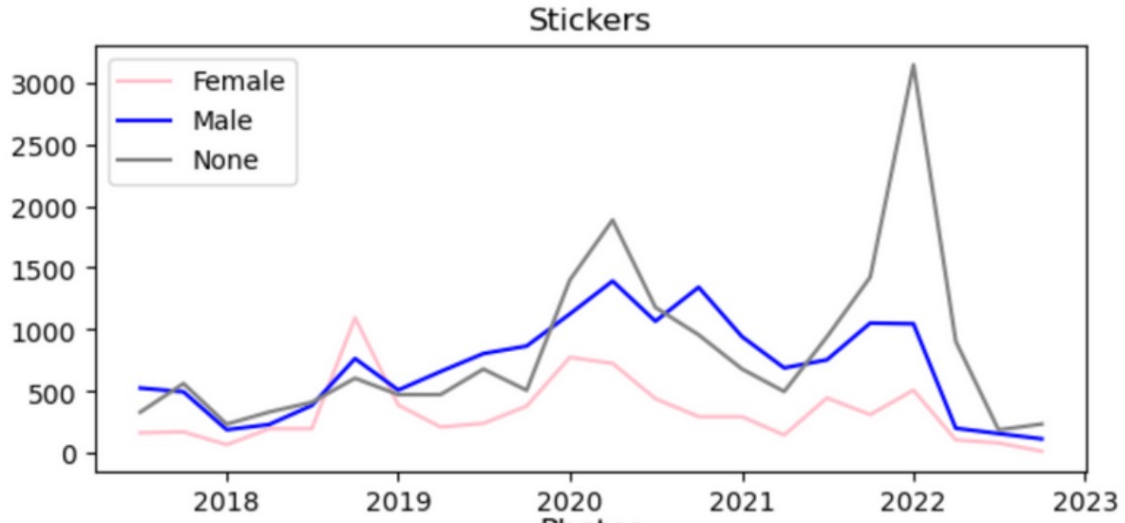
Final results

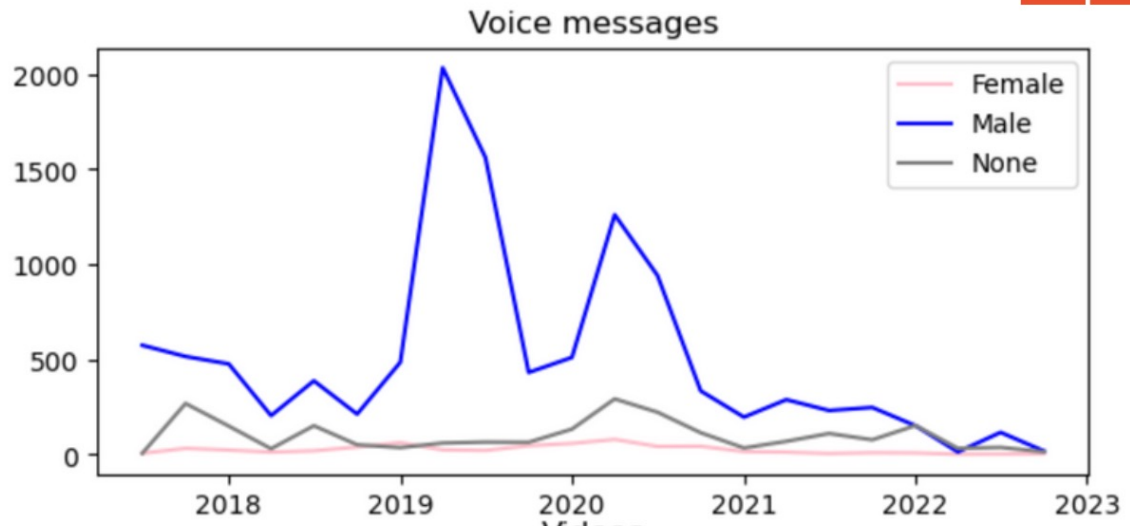
Amount of all messages sent Male vs Female vs None Defined per different timelines divided by message type

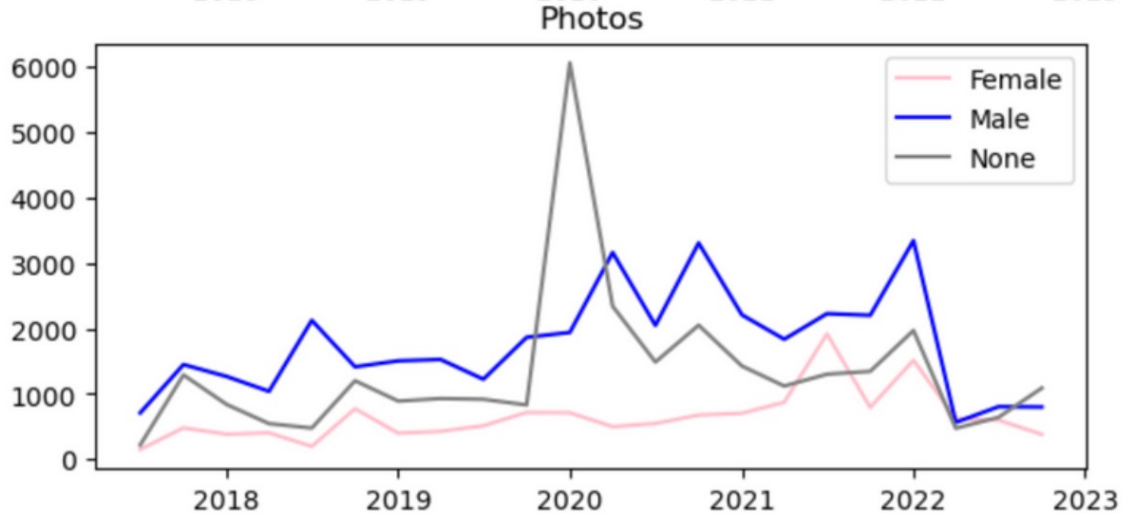


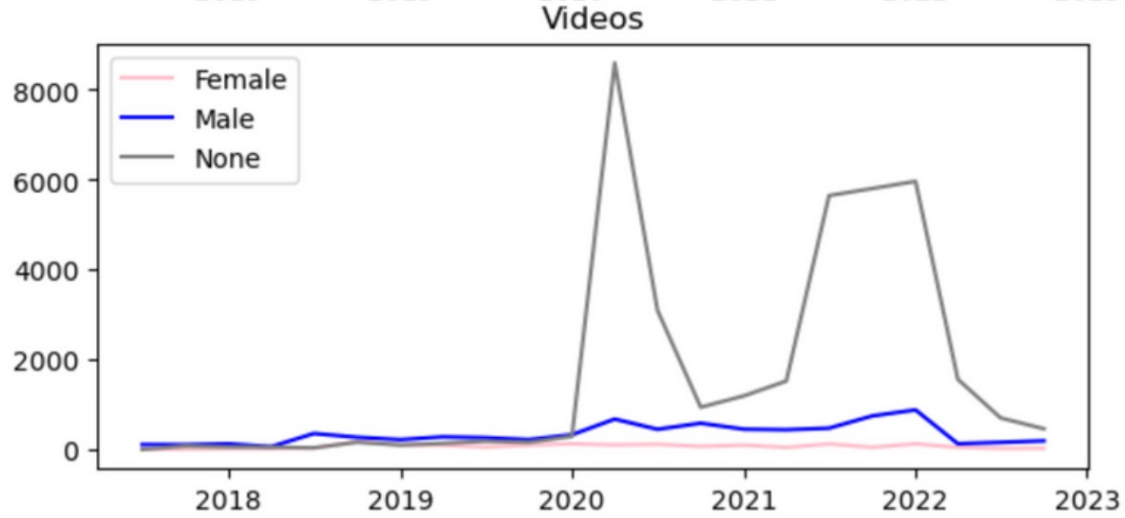
Per year

Final results



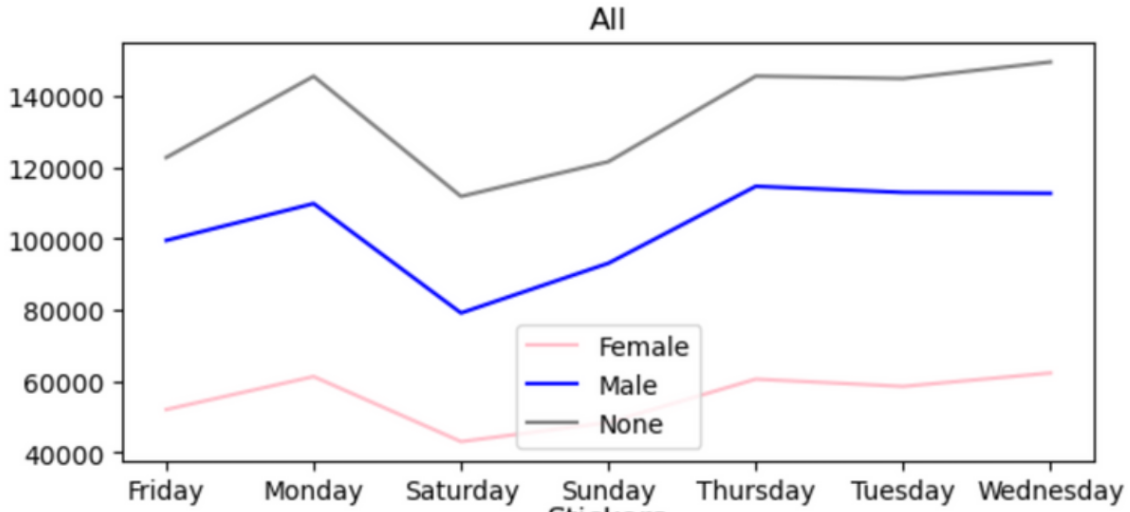


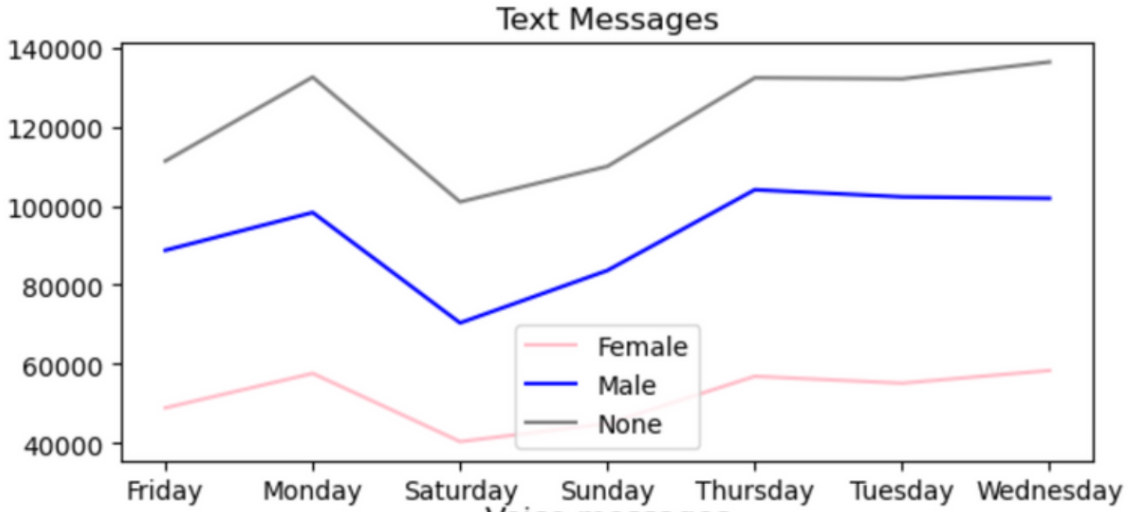


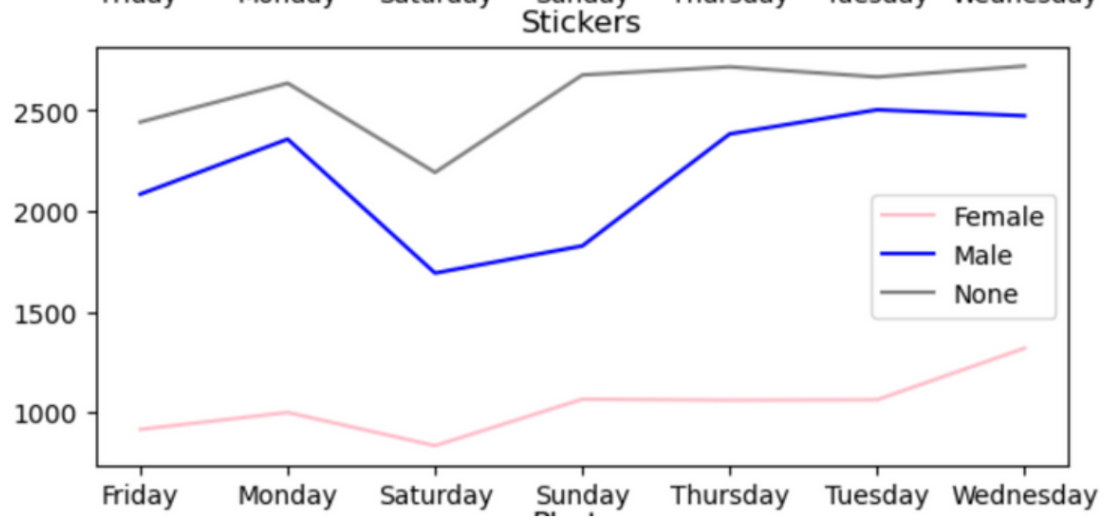


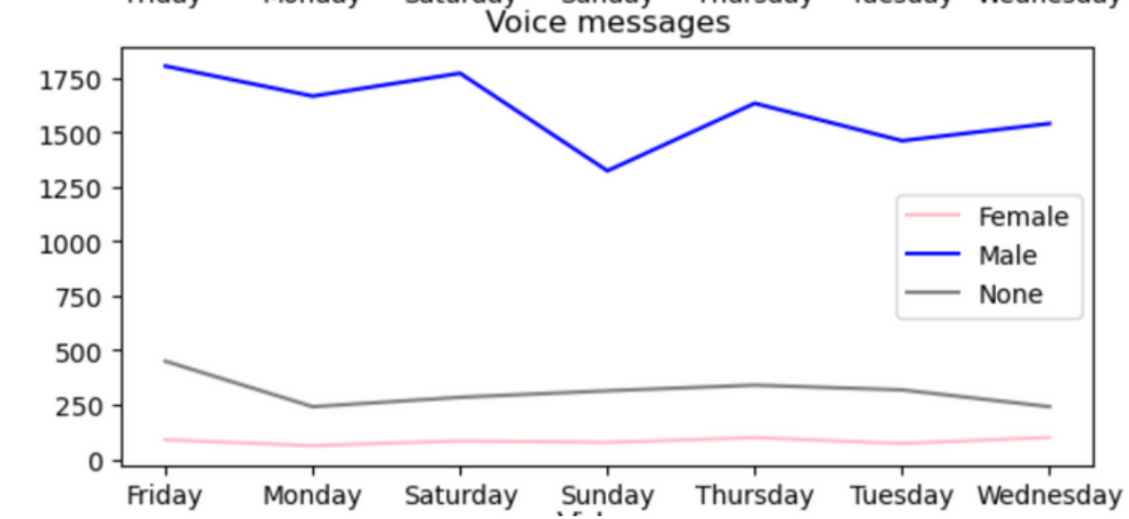
Per quarter

Final results



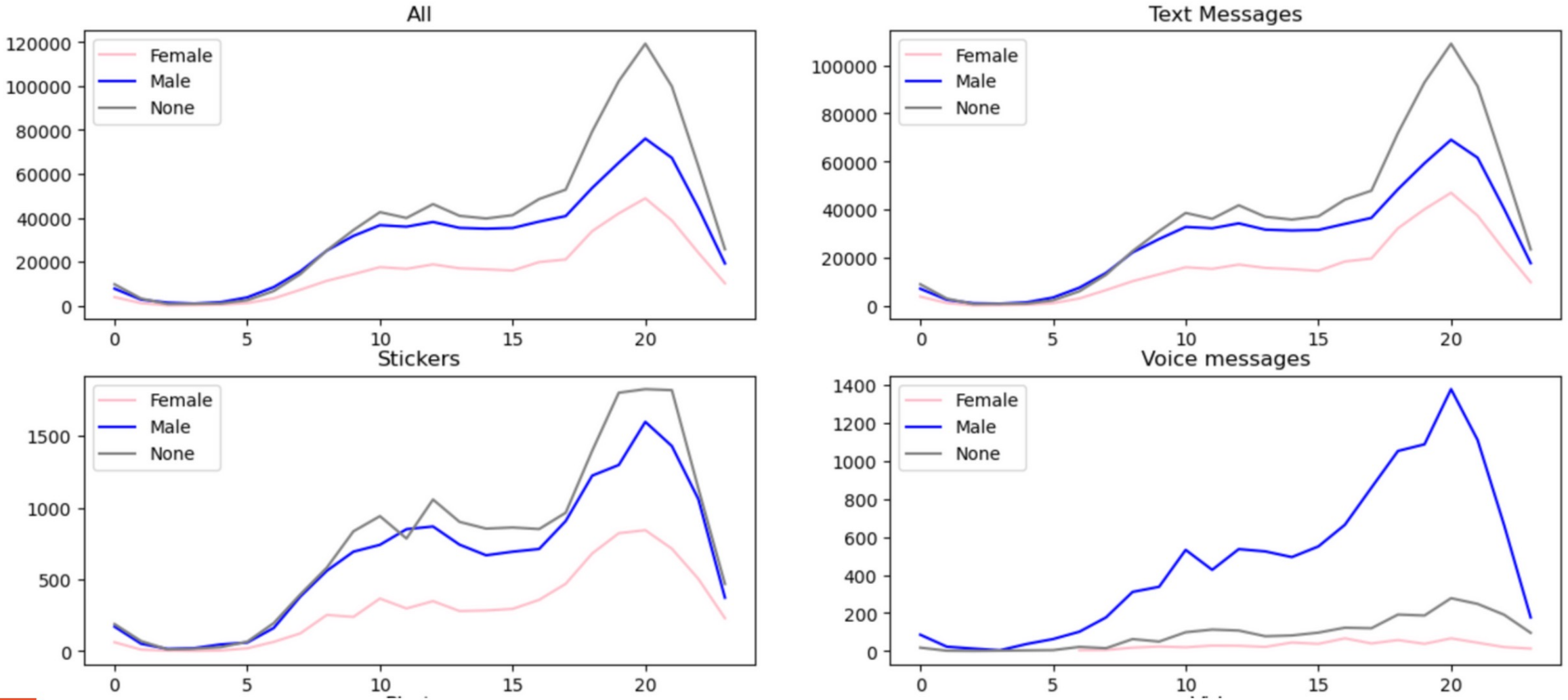






Per day of week

Final results



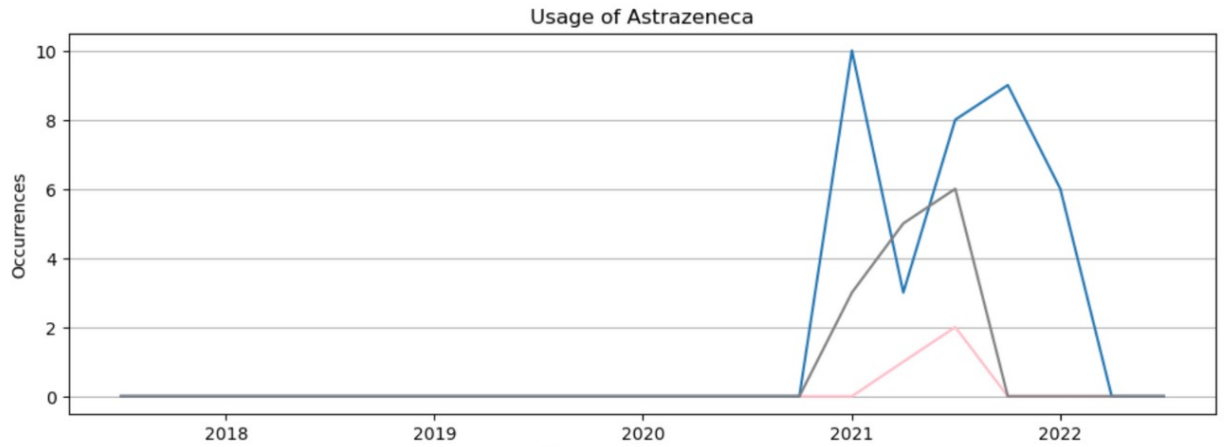
Per hour of the day

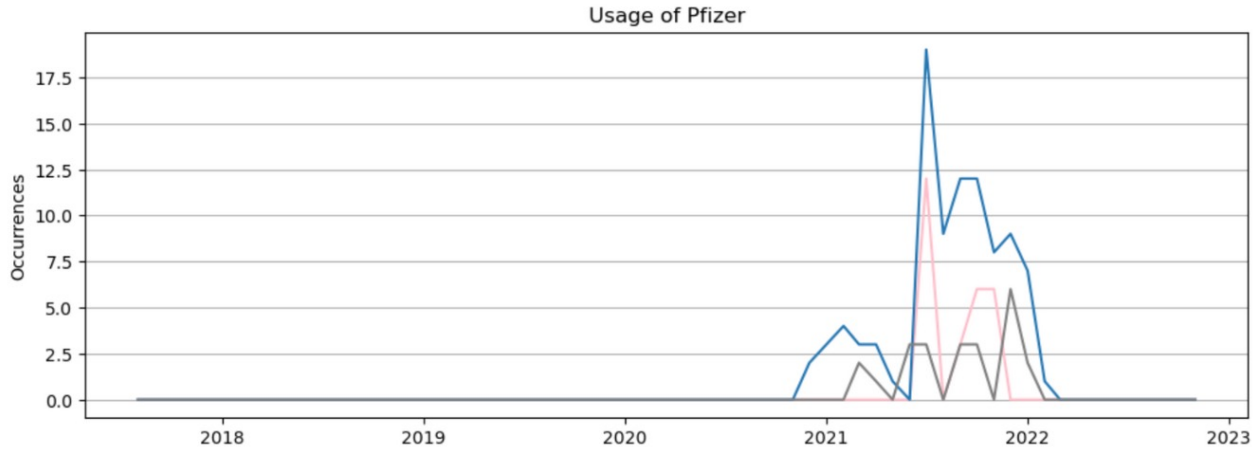
Final results

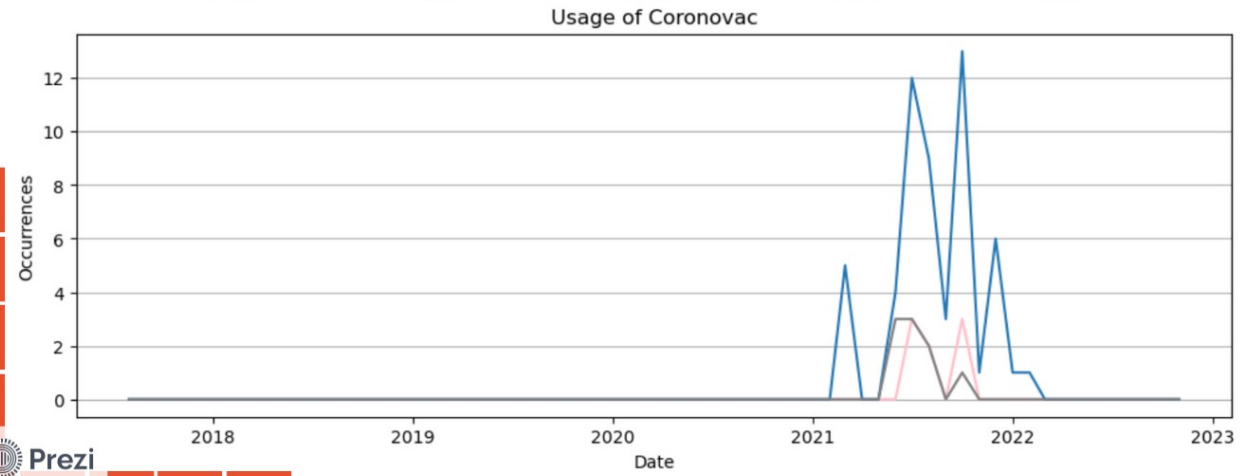
Question: How is the usage of vaccine names look over time for different sex groups?

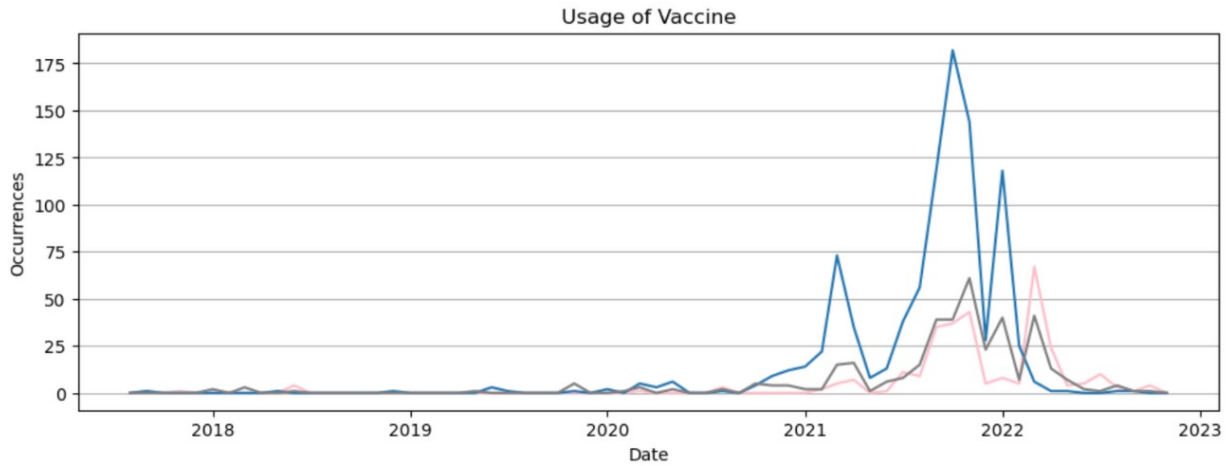
I created a dictionary of synonyms for a different types of vaccines

```
vaccines = {
    'astrazeneca': {'full_name':'Astrazeneca','synonyms_list':['астразенек', 'астра-зенек', 'zenec','astraz']},
    'pfizer': {'full_name':'Pfizer','synonyms_list':['pfizer','пфайзер','файзер','комирнати','комірнаті','comirnaty','комернат
    'coronovac': {'full_name':'Coronovac','synonyms_list':['coronovac','coronovak','короновак','коронавак','коронавок','coro
    'moderna': {'full_name':'Moderna','synonyms_list': ['moderna','maderna','мадерн','модерн']},
    'vaccine': {'full_name':'Vaccine', 'synonyms_list':['вакцин','прививка','vaccine','щепленн','шчеплення']}
}
```



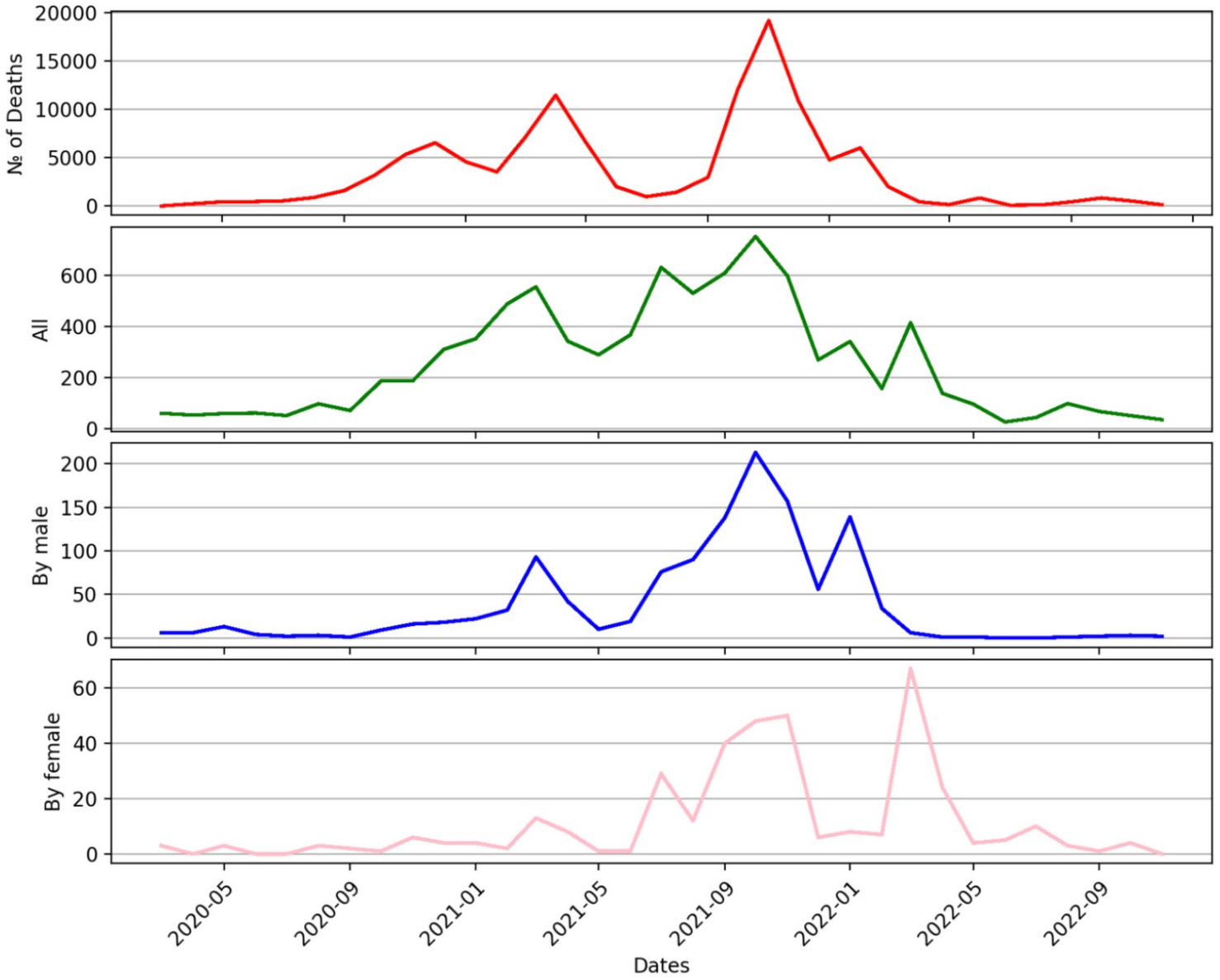






Final results

Question: Is there any correlation between the number of deaths in a time period and the frequency of usage vaccine names
I used open data from: https://ourworldindata.org/coronavirus/country/ukraine



Further work

- To get more accurate distribution data of female and male users can be also used last name and username columns.
- More interesting information can be get from analysing of sex groups behaviour
- Logic of substring counting can be improved because with usage of synonym list it can overlap and count extra times

References

- GitHub: https://github.com/MikLay/telegram-eda

Thank you for your attention