Computational social science

created by Anton Marynych

teacher: Andrew Kurochkin
group №1

16 december 2022

plan

- introduction
- dataset properties
- exploratory data analysis
- results & conclusions

section 1

introduction

introduction

In this presentation I am going to show you the results of my analysis of my Telegram chats. I did research on different areas of my behaviour in Telegram. However, the main subject of this research was to analyse the chat with my friend Zakhar.

section 2

dataset properties

how did I get the data?

I got this data set using the program provided by my teacher. I followed the instructions on the GitHub page of this program and managed to download all of my data from Telegram.

how long did it take?

It took about 12 hours to download every chat. It also took about 2 hours to download each big group chat. The process of downloading this data didn’t stop even when my laptop was in the sleeping mode. Luckily enough, I didn’t face any problems while downloading the dataset.

dataset properties

1 624 113	messages in total
1 412 434	text messages
146 393	voice messages
21 324	video messages
36 986	stickers
340 mb	size

section 3

exploratory data analysis

parts of my research

I decided to divide my research into few parts. This parts describe different sides of my Telegram behaviour. The parts I chose were:

- chat with a friend
- tiktok
- chats with my friends
- me

chat with a friend

Here I wanted to analyze our communications in different topics. I used TF-IDF analysis to find out what do we talk about the most. I got clusters that represent topics. I also faced a problem here. Due to the small size of the dataset this algorithm couldn’t do clusterization properly and 75% of clusters were meaningless. So I had to choose the best clusters and divided them into groups that represent the most popular topics in our chat.

arranging a meeting

cluster 1

набери
опоздаю набери
опоздаю
попали
пенальті
позже
напишешь
контратака
контрактову
контрактовую

cluster 2

демеевской
пожалуйста
ближе
дивишся
курсах
метро
гулят
хмхмхм
олимпийской
скільки

cluster 3

тільки прийшов
напишешь будешь
готов
напишешь будешь
будешь готов
помиюсь
прийшов
готов

cluster 4

давай
впринципі
контрактовой
демеевской
напишу

arranging a meeting

As we can see from the results of my clusterization, when we talk about offline meetings we mention some familiar locations in Kyiv and the names of the metro stations that are close to the locations were we usually meet. Cluster 3 covers the topic of arranging online meeting, because that’s what we do a lot these days.

football

cluster 1

барселоне
баварии
давай
рона
трансферы
думаю баварии
забавная
давай давай
решить вопрос
придется

cluster 2

буває
барса
конференции
футболісти
після
дивись
работает
челсі
матчі
взагалі

cluster 3

ювентус
дивлюсь
действительно
шахтар
аталанта
порту
пирло

cluster 4

продадут
аталанте
некст сезоні
головы
малиновского
летом

football

We do talk about football a lot. As we can see from the clusters on the previous slide, we talk a lot about transfers in football in particular (clusters 1 and 4). Those clusters show the big variety of the clubs we talk about.

studies

cluster 1

алгоритмы

уровня

отличное

темы

горіли

книге

описание

дививмся

упражнений

интересно смотрел

cluster 2

ааааа

задача

задачу

завдання

більше

будешь

славка

насправді

контест

studies

This clusters show that we also talk about studies and programming in particular. In cluster 1 we mention some books, algorithms, tasks. In cluster 2 it is more specific, because here we mention a contest (which is a codeforces contest). We also mention some problems and the name of our mutual friend who often participates in those contests.

anime/manga

cluster 1

глави

розумію розумію

можна підозрюю

виходять

сторони

наверняка

моменту

выходят

спіши

спойлеры

cluster 2

назаре

серия

новая серия

назаре назаре

хочешь смотреть

анимесериал

дивився серію

anime/manga

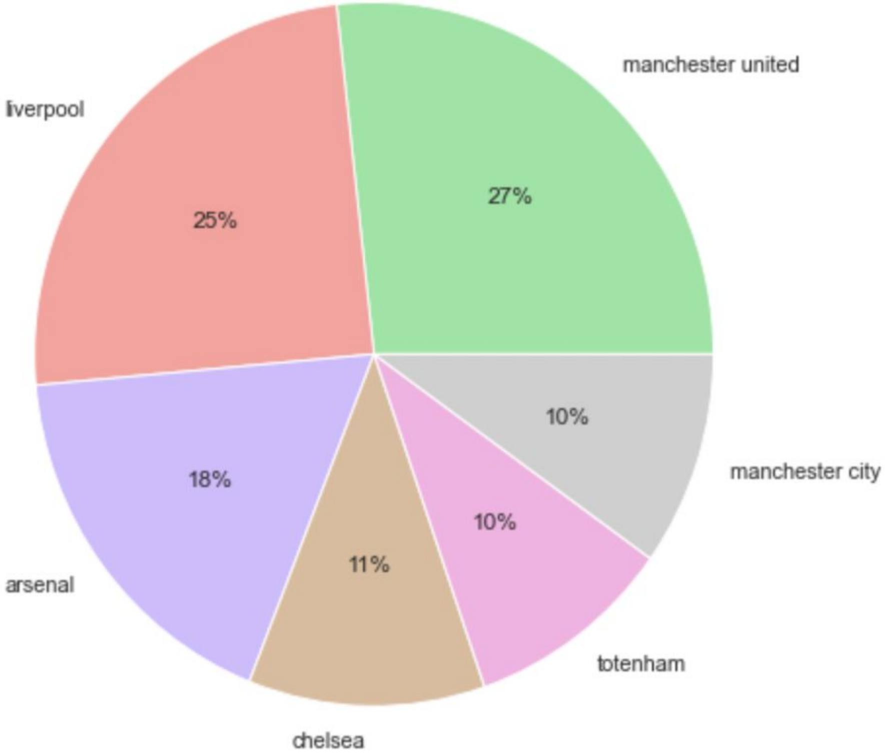
In cluster 1 we can see that we talk a lot about the chapters of manga, which are coming out and about spoilers and other stuff connected to chapters. In cluster 2 we can see that the topic here are new episodes of some anime series.

what did I decide to do next?

Basing on the results I’ve got from my TF-IDF analysis and clustering I decided to plot different information that covers one of the topics that I’ve mentioned before. I also included few questions that don’t cover any of these topics just to get some general information about our chat.

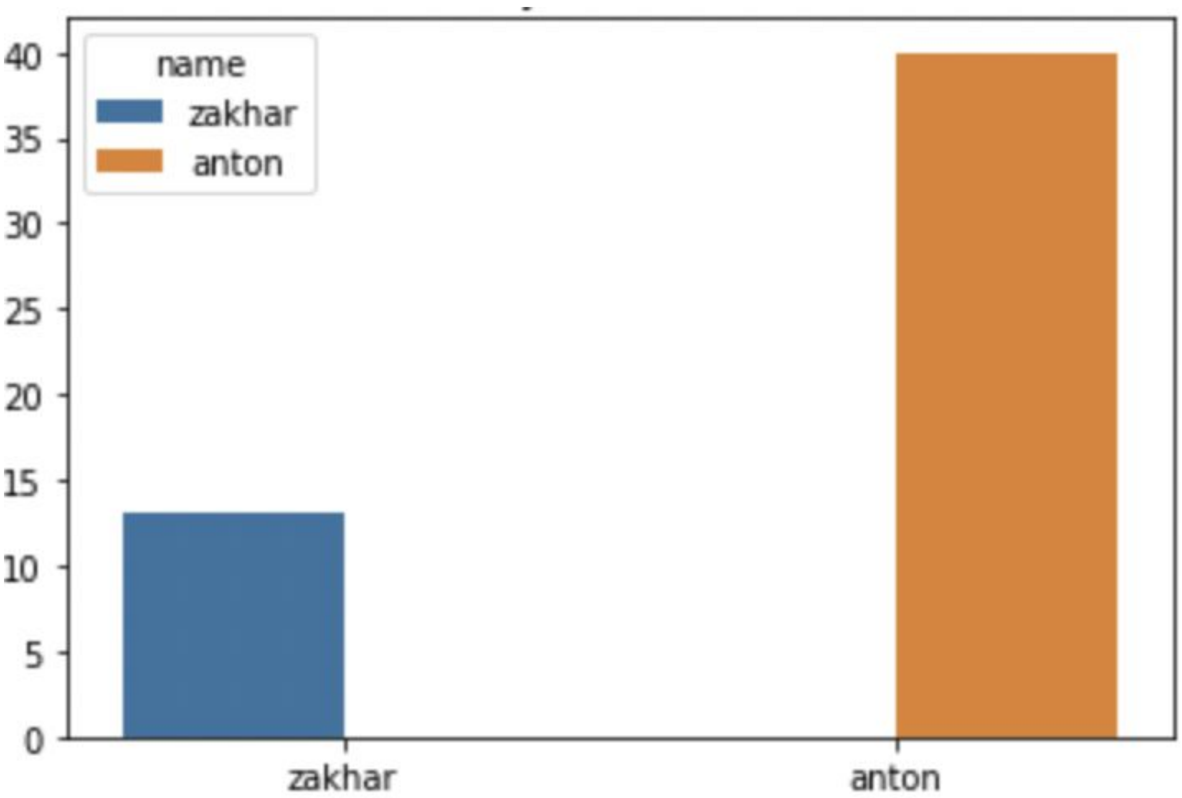
football

what english top football club is mentioned the most?



arranging meetings

who mostly invites for discord?



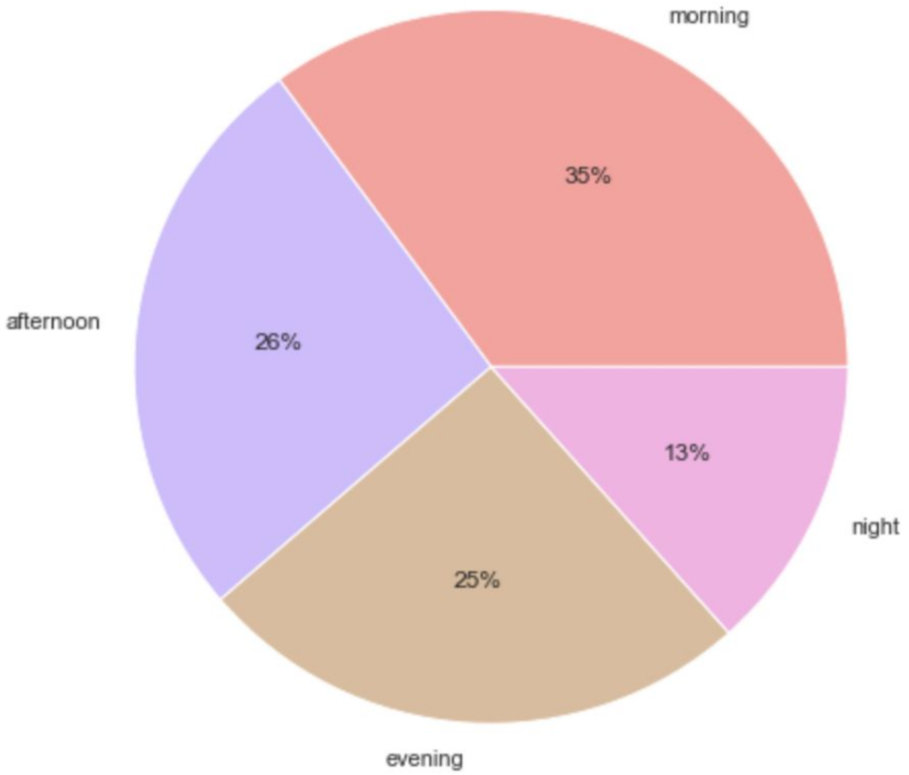
tiktok

In this part I am going to analyze people’s behaviour on tiktok basing on the tiktoks they send me during the day in Telegram.

Firstly, I decided to choose the question related to Zakhar to smoothly move from the previous part.

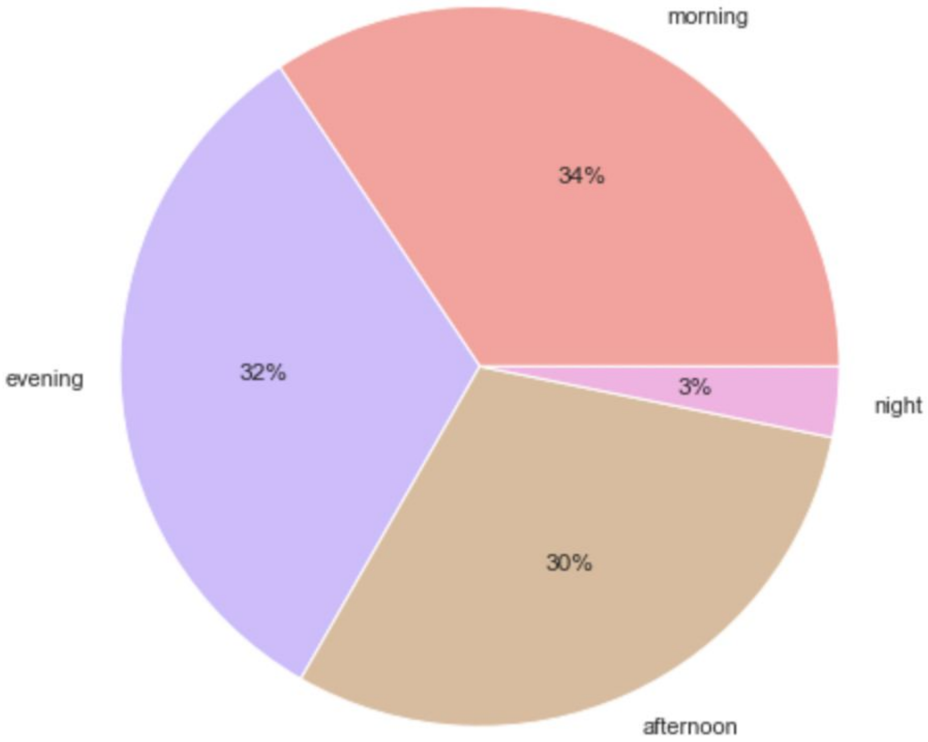
tiktok

When does Zakhar send me tiktoks during the day?



tiktok

The same exact question but here I used the data from the other chats with my friends to compare.

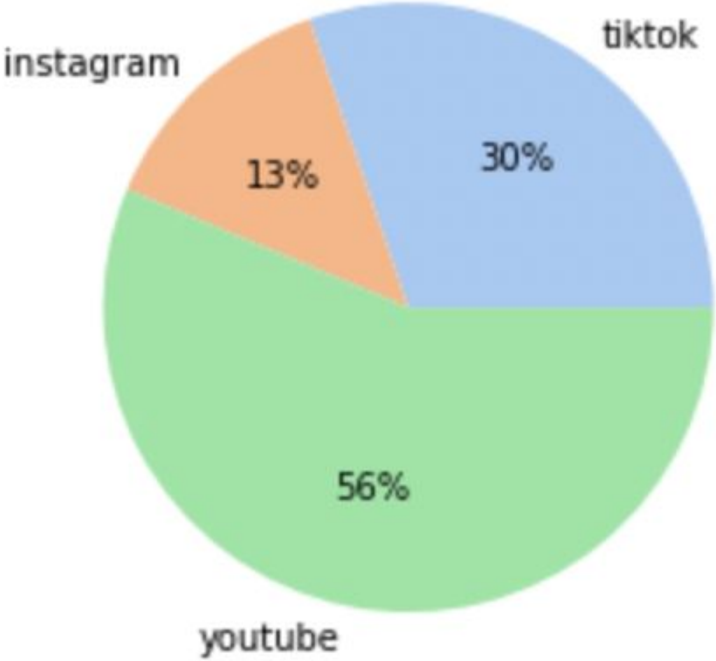


tiktok

Interesting fact: Zakhar sends me 0.64 tiktoks per day.

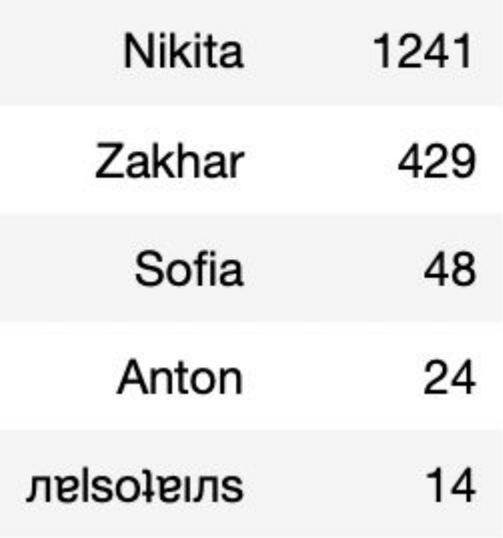
tiktok

Popularity of different types of links in my private dialogs.



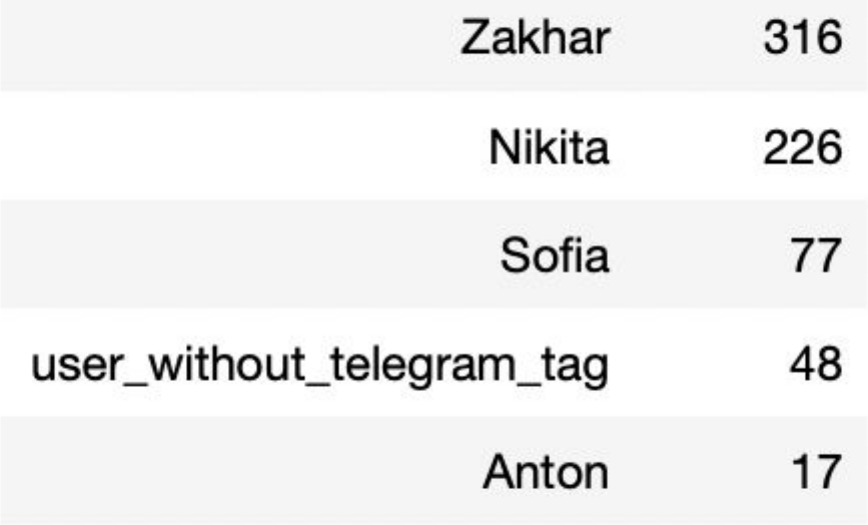
tiktok

Top 5 who send me



Nikita	1241
Zakhar	429
Sofia	48
Anton	24
ɹɐlsoʇɐıɹs	14

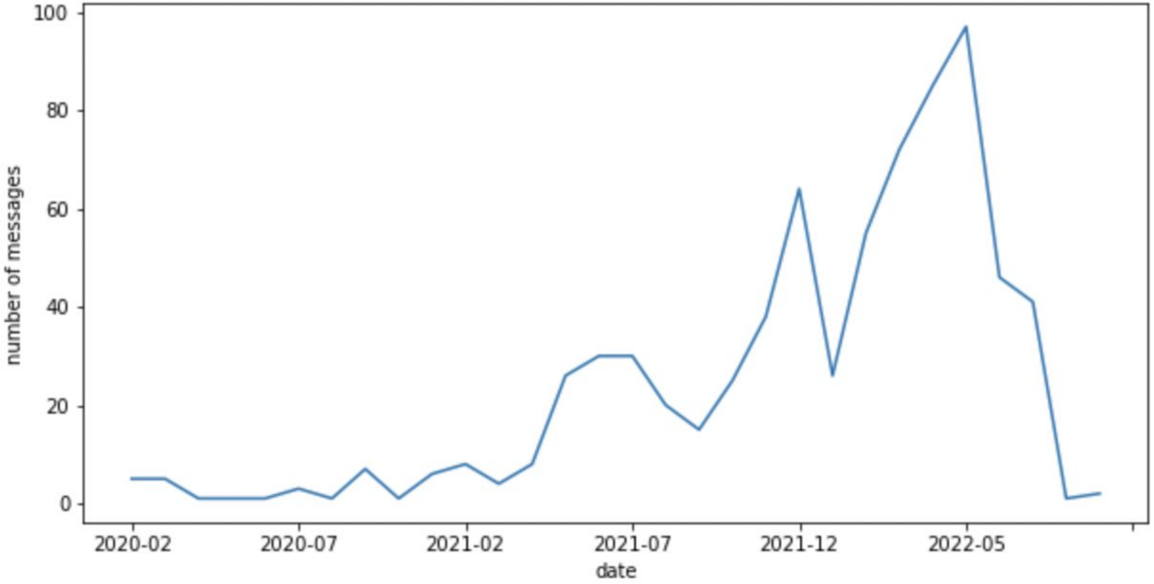
Top 5 who I send



Zakhar	316
Nikita	226
Sofia	77
user_without_telegram_tag	48
Anton	17

tiktok

My activity on sending tiktoks.

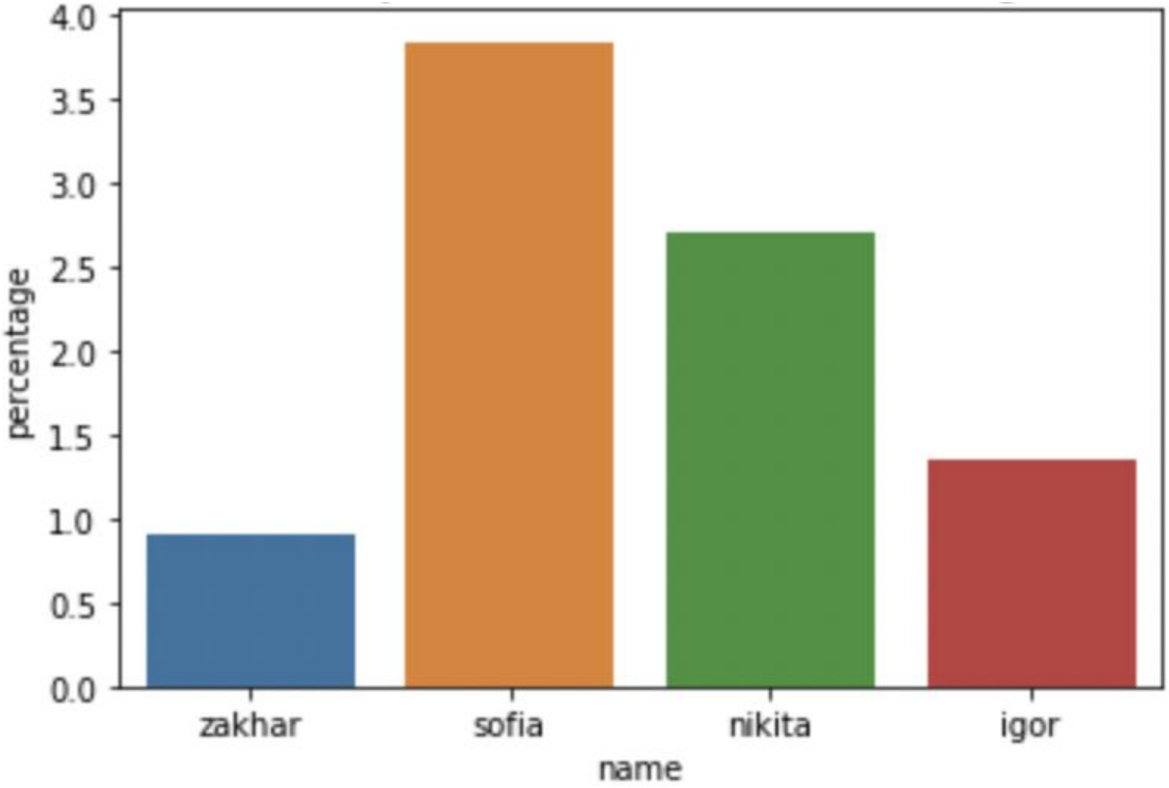


friends

In this part I am going to analyze the differences between my four close friends.

friends

Who texts me the most during the night?

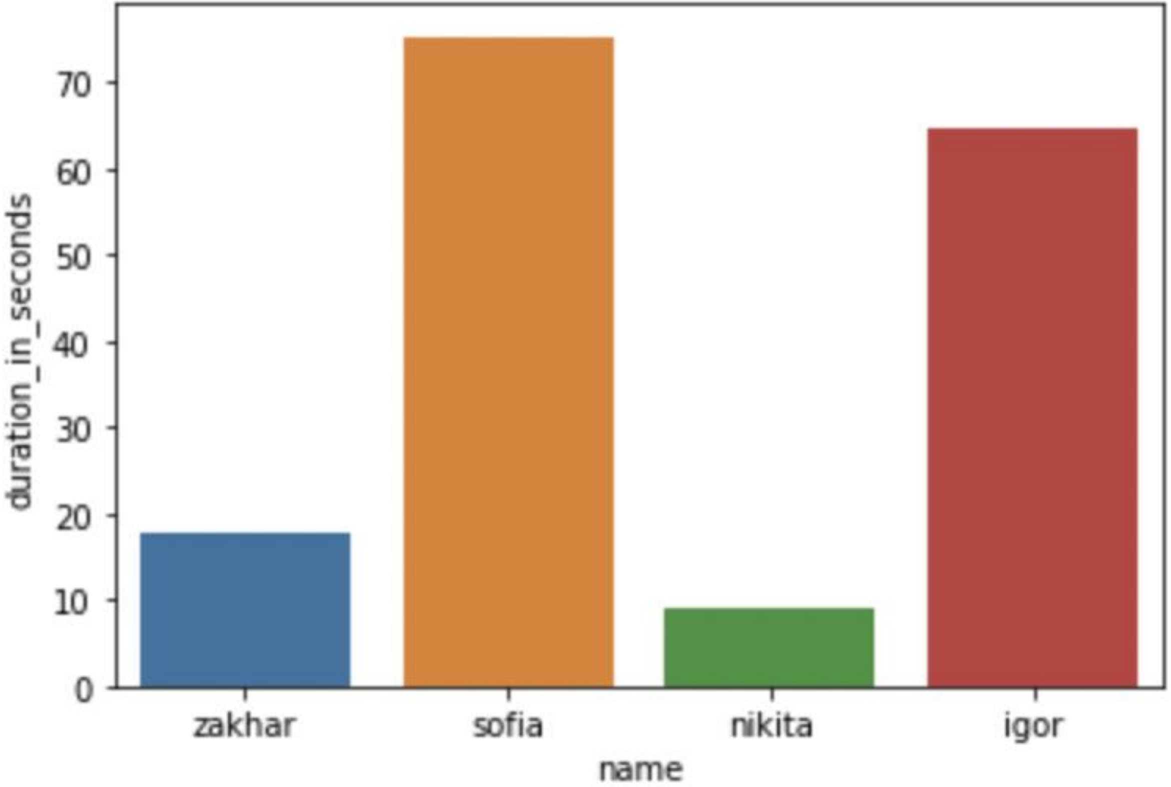


friends

I am really not surprised with the outcome here. I was expecting Zakhar to be the lowest here, because he has the best sleeping schedule out of all of my friends.

friends

Average duration of the voice message

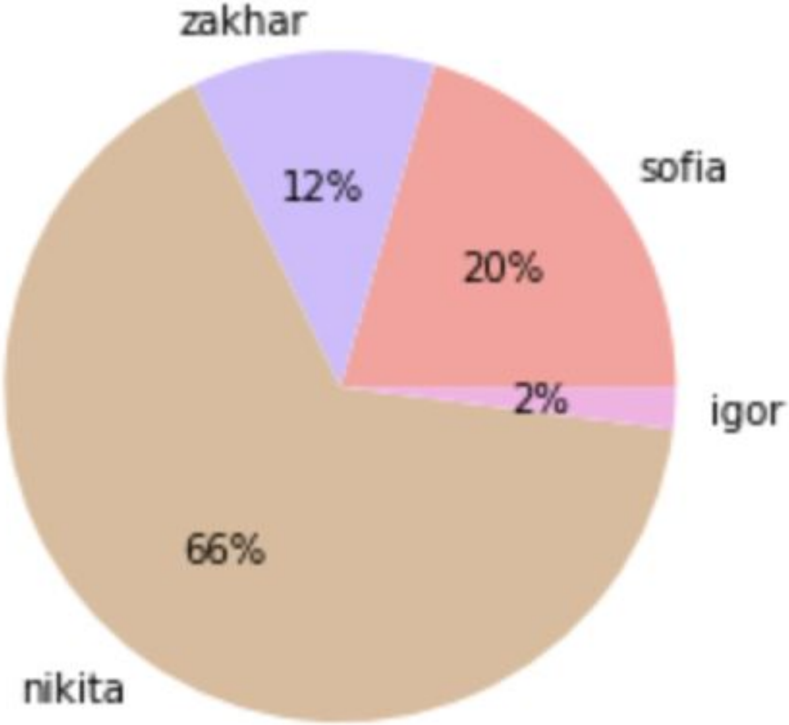


friends

Here I’ve got some expected and unexpected results. As I thought, Sofia is the first here. But I was honestly surprised for Igor to be the second. And it is also pretty wild that the average duration of the voice message that Sofia sends me is more than one minute.

friends

Who sends the most voice messages?

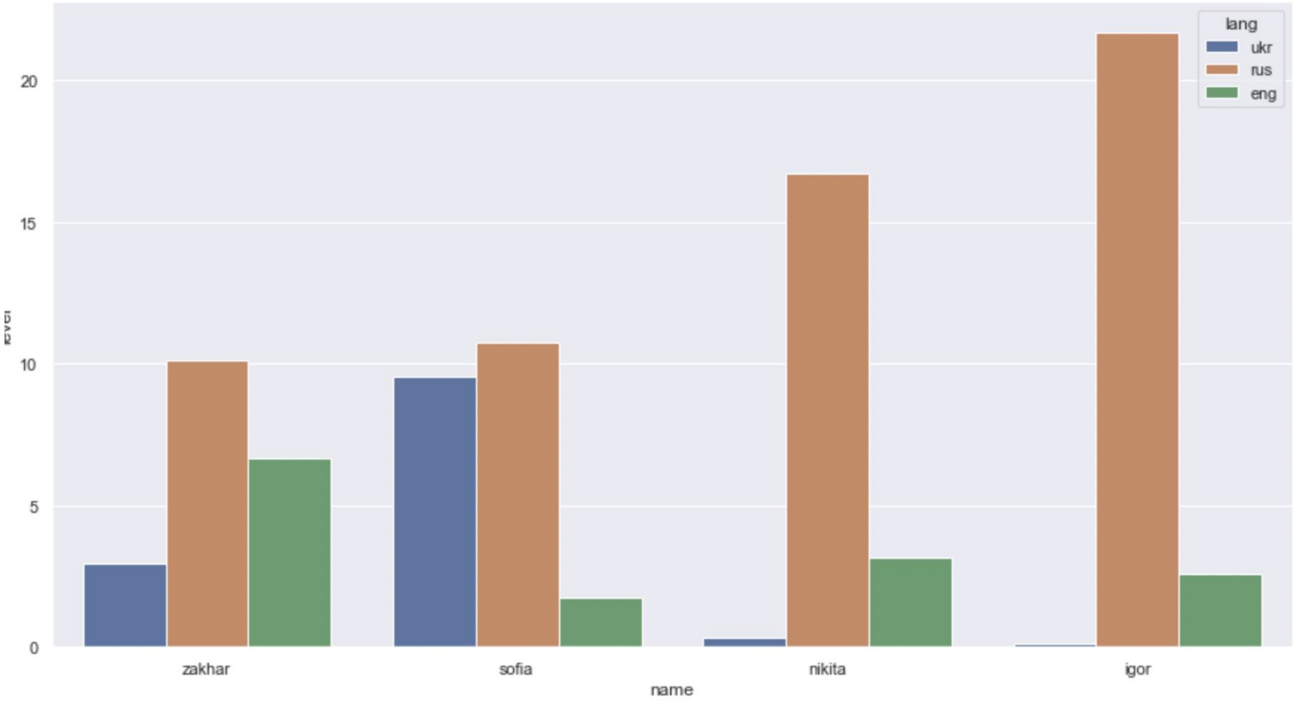


friends

From two previous charts we can assume that Nikita and Sofia like to send voice messages the most among my close friends. But Nikita sends many short voice messages, when Sofia doesn’t send many while her average duration is very high.

friends

Level of different languages in our chats

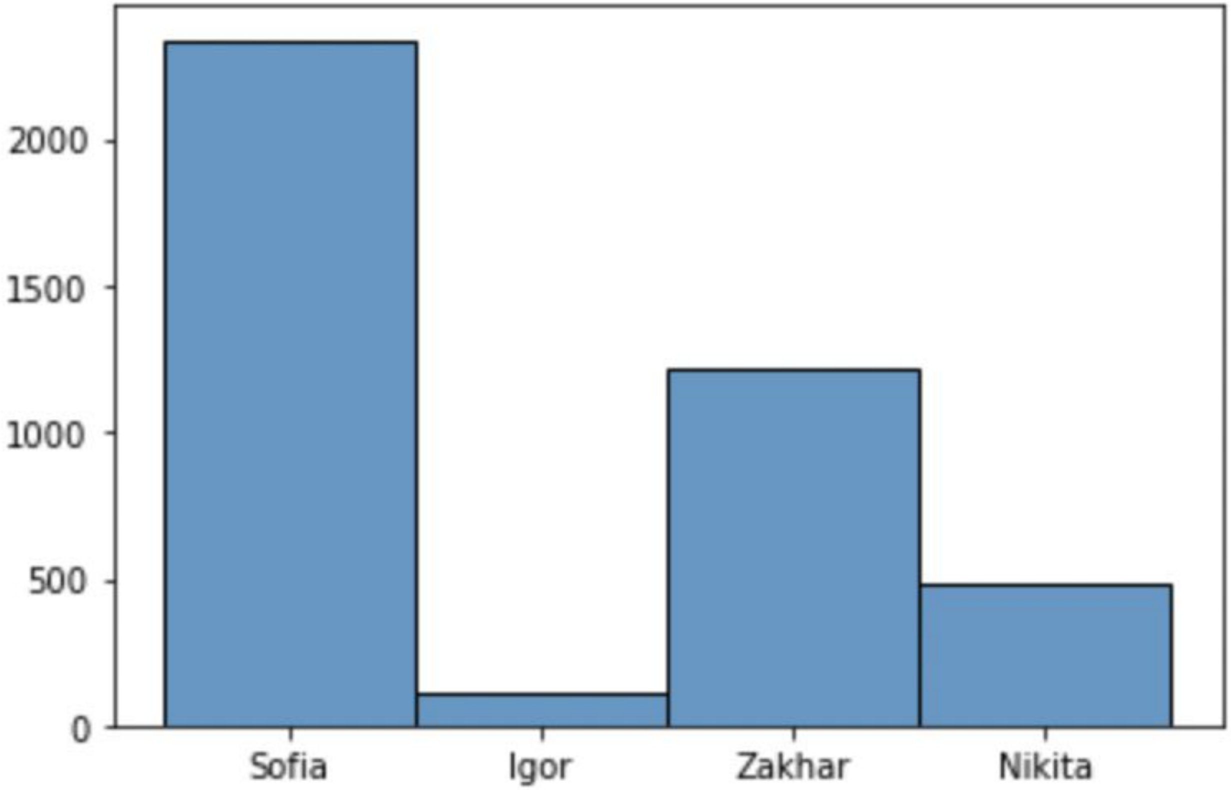


friends

The chart shows that Zakhar and Sofia mostly speak ukrainian, when my other friends speak russian. There is also a high level of english in my chat with Zakhar.

friends

Who laughs the most?

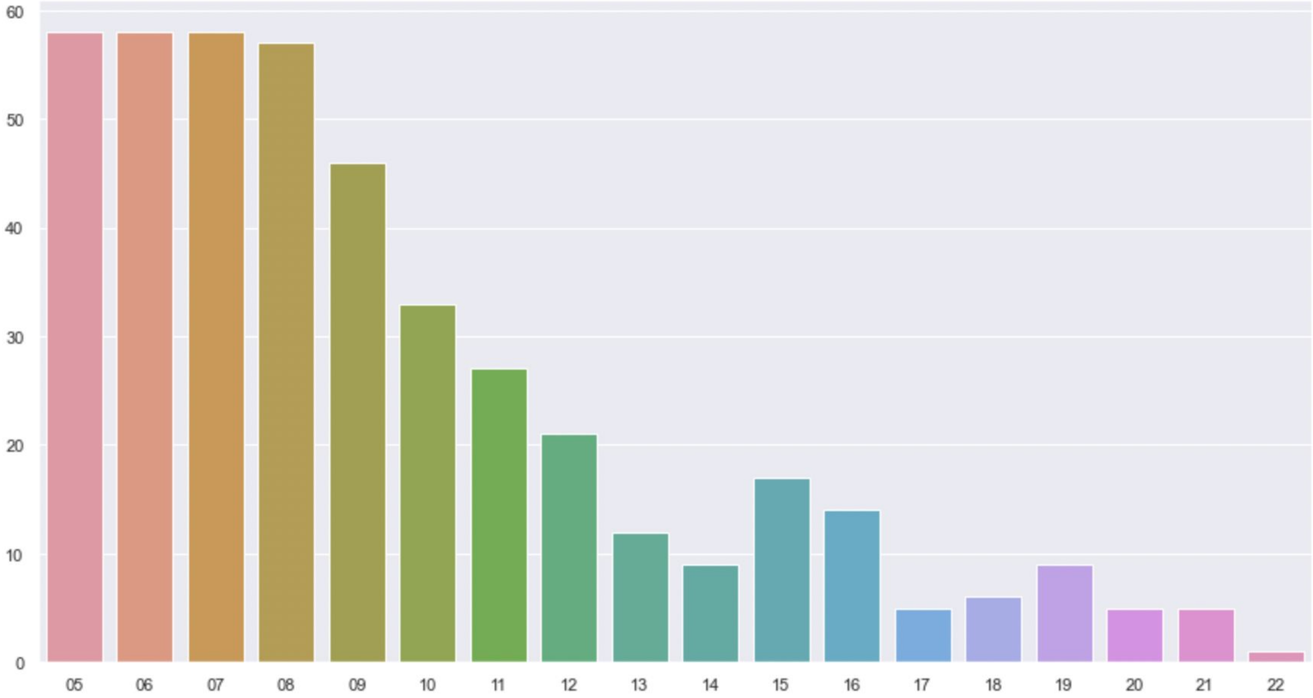


me

In this part I am going to analyze different sides of my behaviour in Telegram.

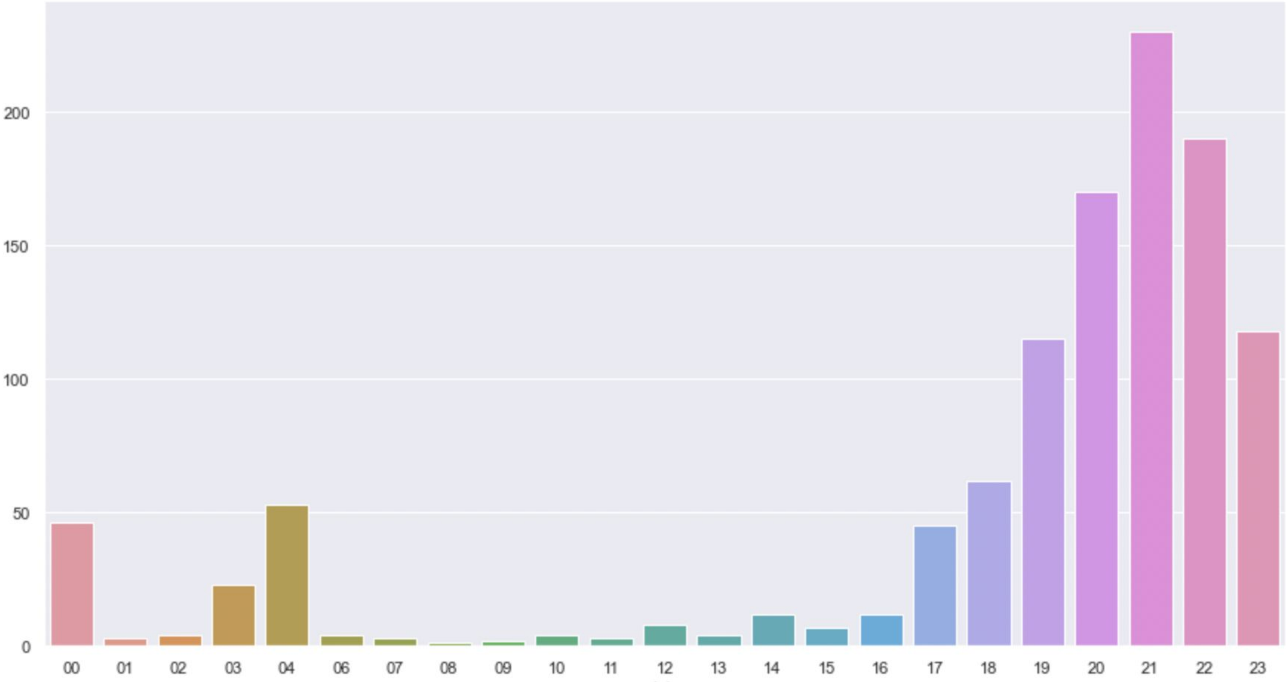
me

When does my Telegram day start?



me

When does my Telegram day end?

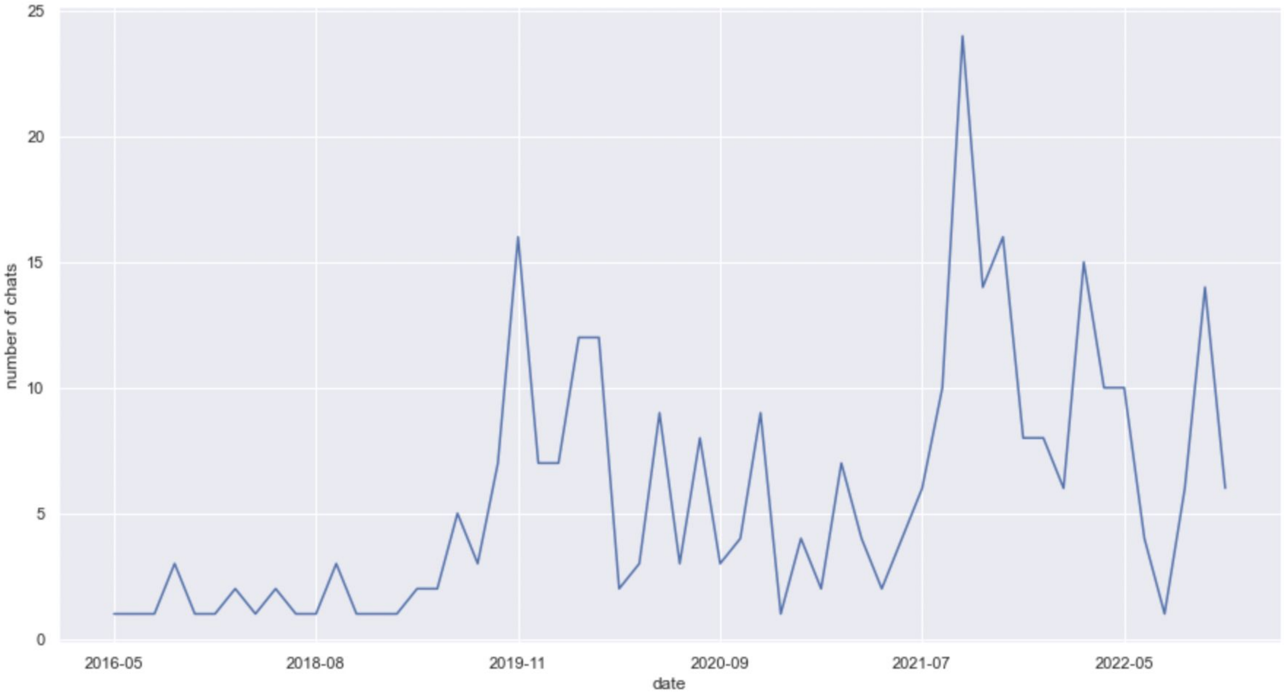


me

I am really surprised with the outcome of those questions. I would never thought that my Telegram day starts and ends so early.

me

When did I start my chats?



me

The main reason why I chose this question is to check if I started many chats when I started my studies in the university. The results totally match the expectations. You can clearly tell the month when I entered my university just by looking at this chart.

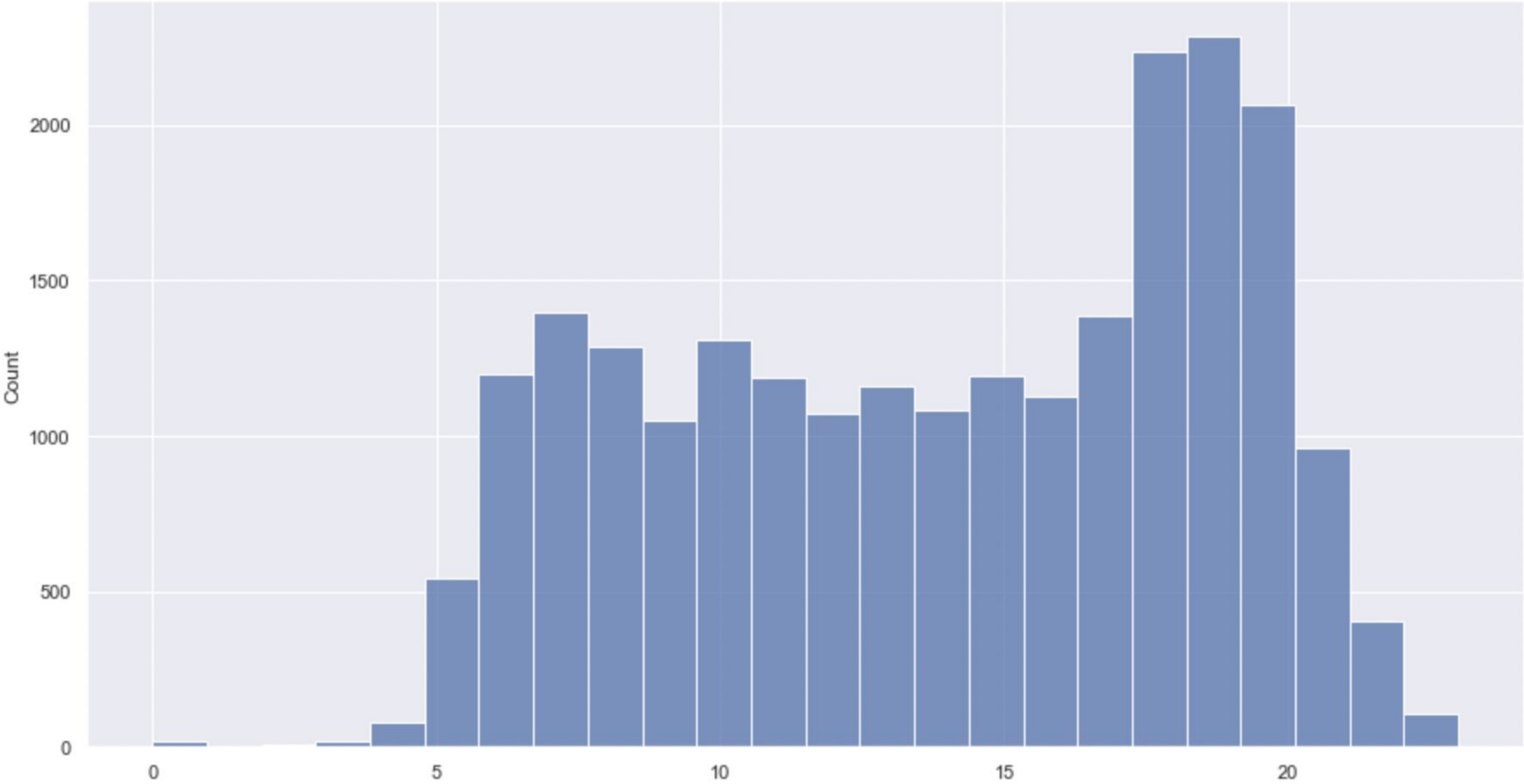
section 4

results & conclusions

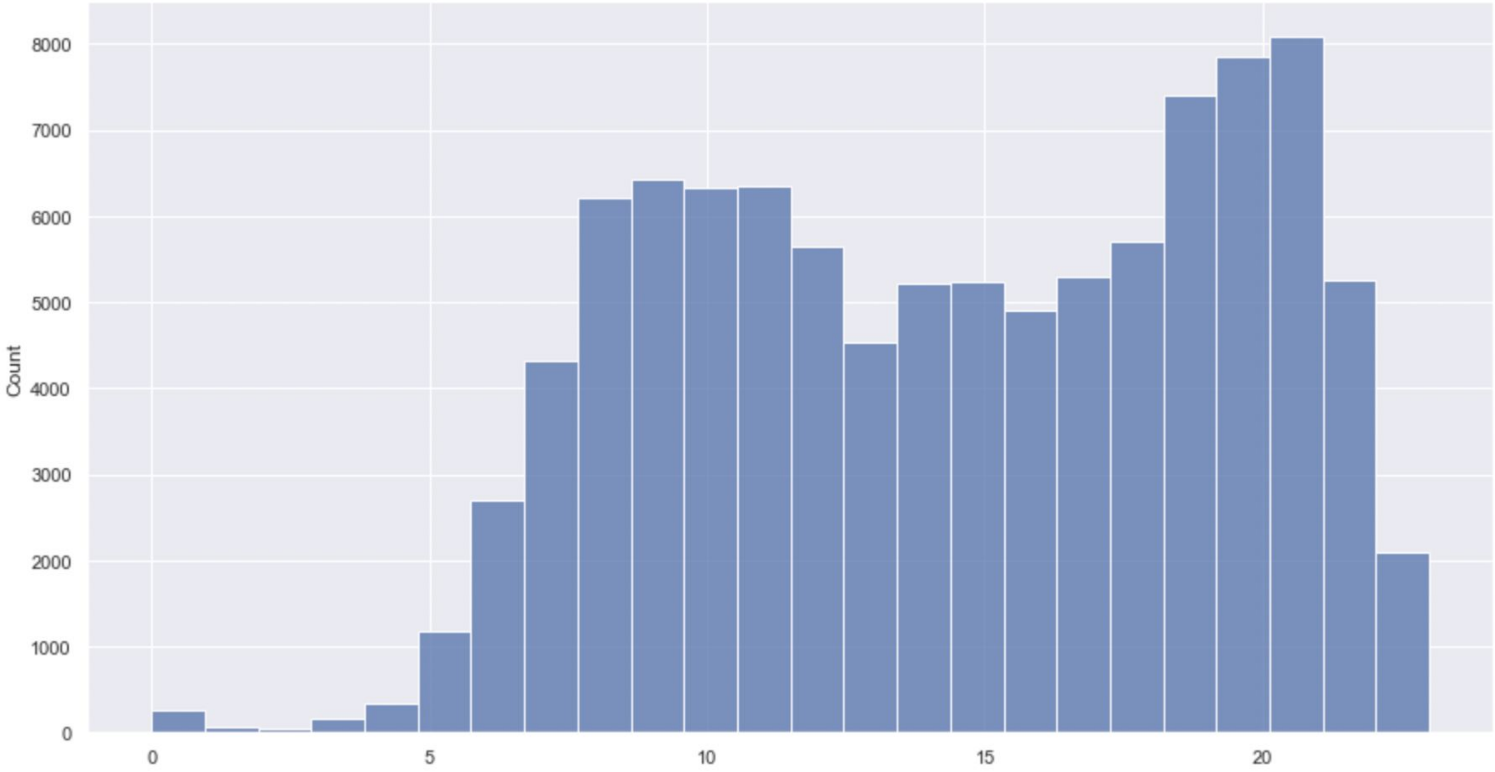
results

I am going tell you my personal top 4 the most interesting outcomes that I’ve got during this research. I didn’t mention them in the main part.

#4 when do we chat with zakhar?



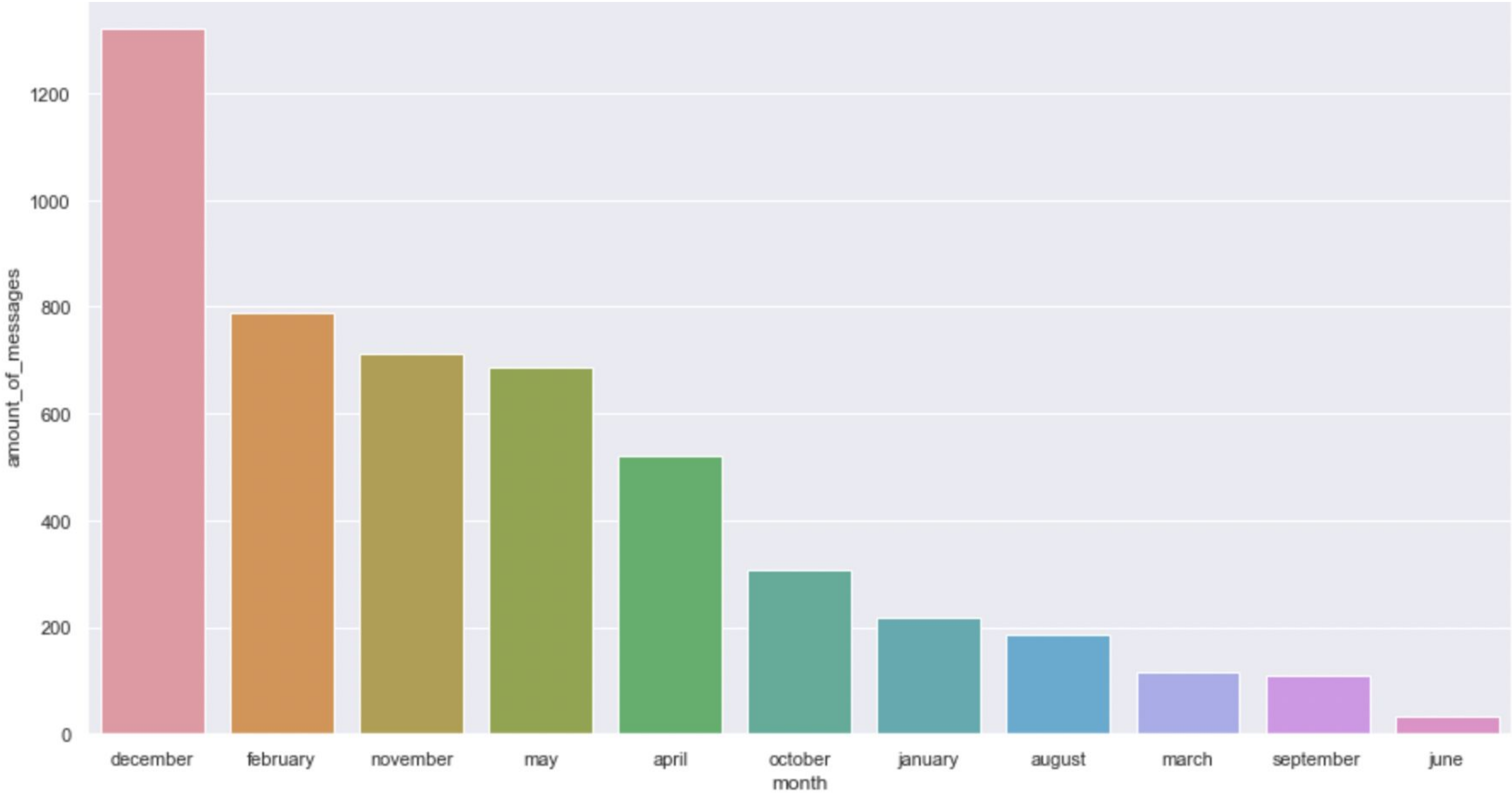
#4 when do we chat with zakhar?



#4 when do we chat with zakhar?

On the first chart you could see the activity in our chat during the day. On the second one you could see the same chart but about my other friends to compare it with the first one. They look similar, but with Zakhar there are few distinct hours when we chat more actively. Those hours are 6-8 p.m. The chart about my other friends is more steady.

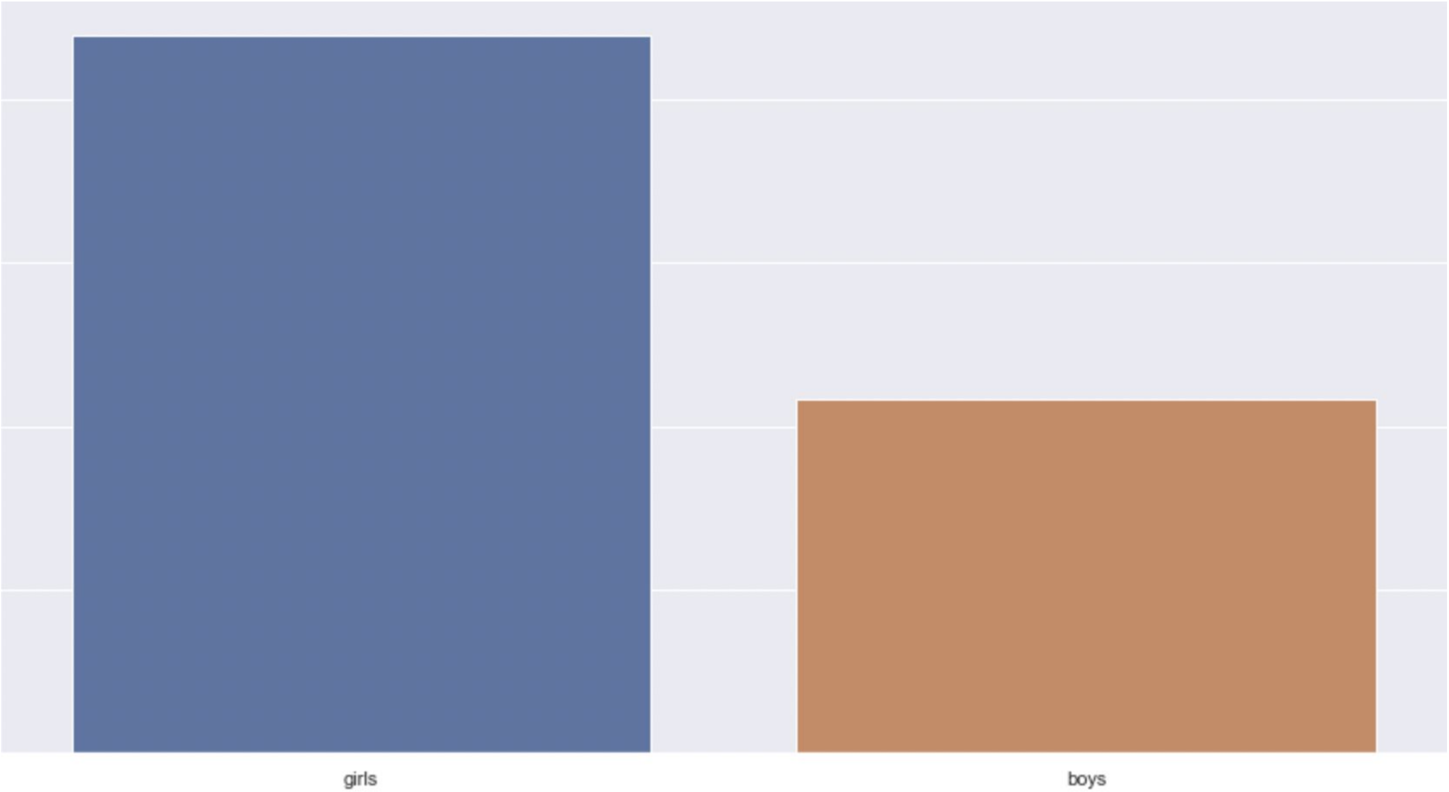
#3 my studying activity



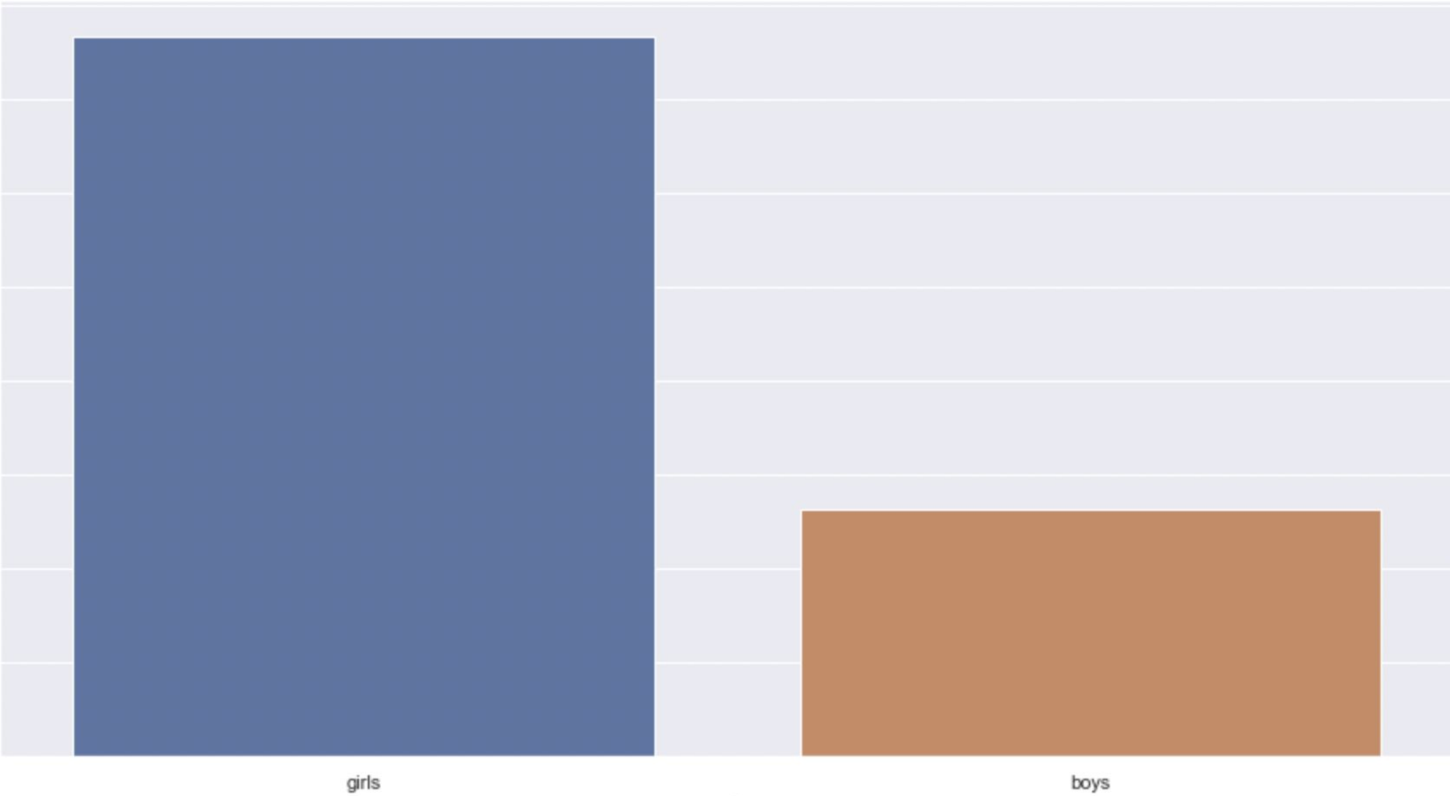
#3 my studying activity

The most interesting part of this topic is how I got the data. I just chose l person how I talk to a lot but only because of studies in the university. He is in my top 10 received & send messages, so I had enough data. The results of this approach are pretty accurate as you could see.

#2 boys & girls laugh



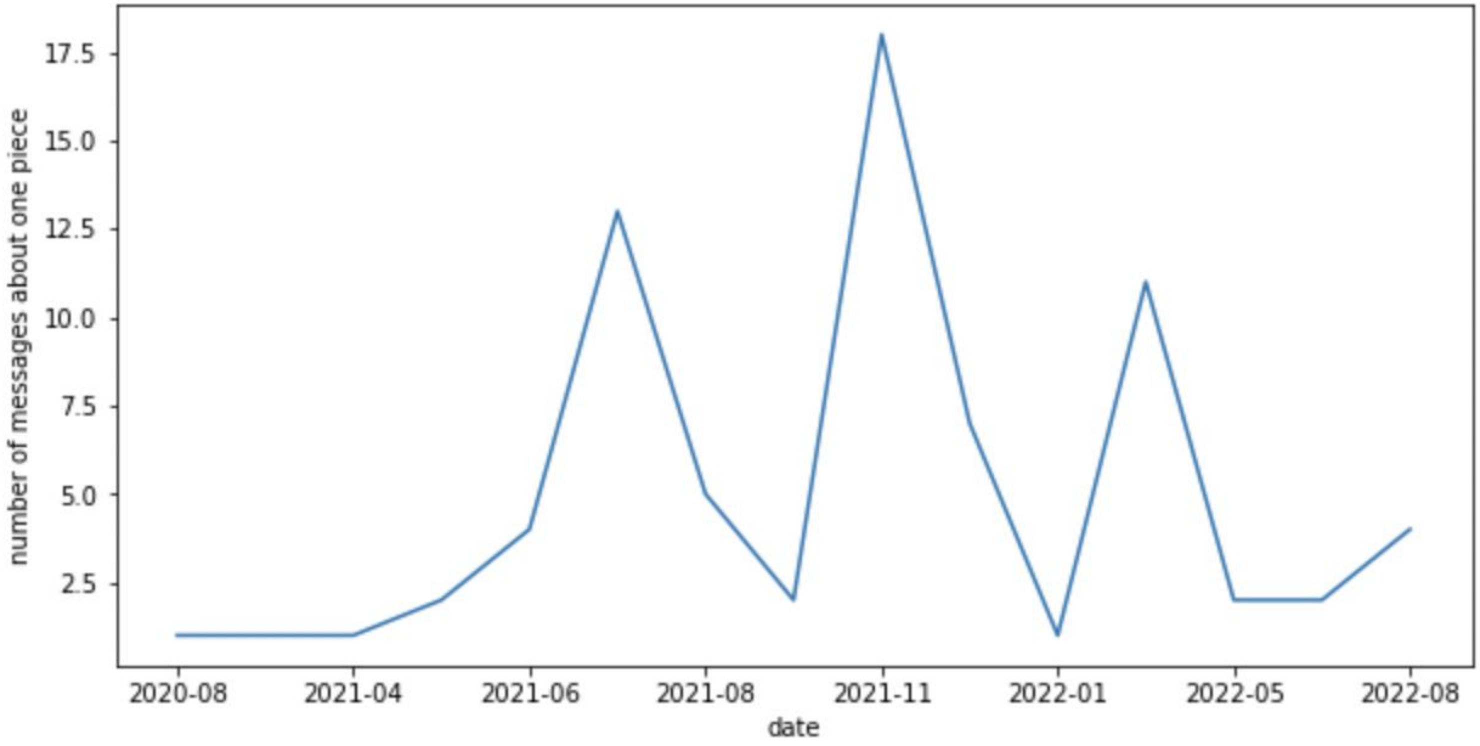
#2 boys & girls laugh



#2 Boys & girls laugh

I took 3 chats with boys and 3 chats with girls and compared the level of their laugh in chats with me (the first chart) and the level of my laugh with them (the second chart). The results only prove the popular theory that boys can be fully honest only with boys when it comes to jokes. That means that in boy&girl chat there might be a lot of fake laughing.

#1 one piece



number of messages about one piece
17.5
15.0
12.5
10.0
7.5
5.0
2.5
2020-08
2021-04
2021-06
2021-08
2021-11
2022-01
2022-05
2022-08
date

#1 one piece

The chart that you’ve just seen is #1 because it has three peaks that perfectly represent one piece related events that happened in my life.

#1 one piece

- the first peak is when I finished watching anime
- the second one is when I started reading manga
- the third one is around the time when everyone was hyped up because of one major event in the manga

conclusion

I’ve learned a lot about using python and different libraries to do analyze a dataset and I also got to know a lot about my friends behaviour in Telegram and I have plenty of fun graphs to share with my friends. I also might download the same dataset in a year or two to find out, what changed in during this time.

further work

The first thing that comes to my head is to do the same TF-IDF analysis, but not only on one chat. I would use all of my messages. The problem in my previous TF-IDF analysis was lack of data. Using all of my messages it should be more than enough.

further work

It is also possible for me to analyze my other friends as deep as I did it with Zakhar. That will help me understand unique features of each and every of my friends.

a very useful link to my github

https://github.com/antoshsha/telegram_analysis