НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
«КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»

29/11/2024

Exploratory Telegram Data Analysis

Data collection & exploration of behavioural patterns

Project by Larysa Vasylenko
for Computational Social Science course;
Applied Mathematics, Bachelor's 3rd year

Teacher: Andrew Kurochkin

Table of Contents

1. Introduction
2. Data collection
3. Exploration process
4. Results (shortly)
5. Further work
6. Link on the GitLab repo with the code

Introduction

about the project

This data analysis looks at information taken from personal Telegram activity to find interesting **patterns in communication** and gain insights related to social behavior (for example Social Loafing, Groupthink, Synesthesia, homosociality). By studying messages, conversations, and interactions in private and group chats, the goal is to understand how people communicate, what they prefer, and explore any social trends that might appear.

Data collection

To collect the data for this project, I used the Python library **Telethon** along with a pre-existing script available on GitHub. Initially, I faced challenges during the data extraction process. My first attempt yielded a dataset of **649 583 messages**, which was significantly limited due to the interruption of the code execution. (The code ran for almost a day and in the middle of the night its execution was interrupted due to network problems.)

At that stage, I decided to proceed with the data I had for the task titled **“Homework #3: Download and Understand a Dataset.”**

However, it soon became evident that the dataset was insufficient for a comprehensive analysis. Determined to resolve the issue, I decided to solve the problem: I found the dialogues that were not loaded and uploaded them.

In the end, I was able to collect a much larger and more complete dataset, **comprising 1 262 176 messages**!

Data collection

final data statistics

number of messages

1 262 176

dataset size in MB (.csv file)

376.9 MB

Exploration process

My analysis began with reviewing the columns available in the dataframe and cleaning the data. During this process, I developed an understanding of the information and statistics I could visualize.

As a result, I had the following features at my disposal:

1. Date
2. Sender ID
3. Receiver ID
4. Message Content
5. Message Type
6. Duration
7. Reactions
8. Dialog ID
9. Forwarded From ID
10. Forwarded From Name
11. Dialog Type
12. Phone Number
13. Participant Information:
    Username
    User ID
    First Name
    Last Name

Exploration process

27 different aspects investigated!

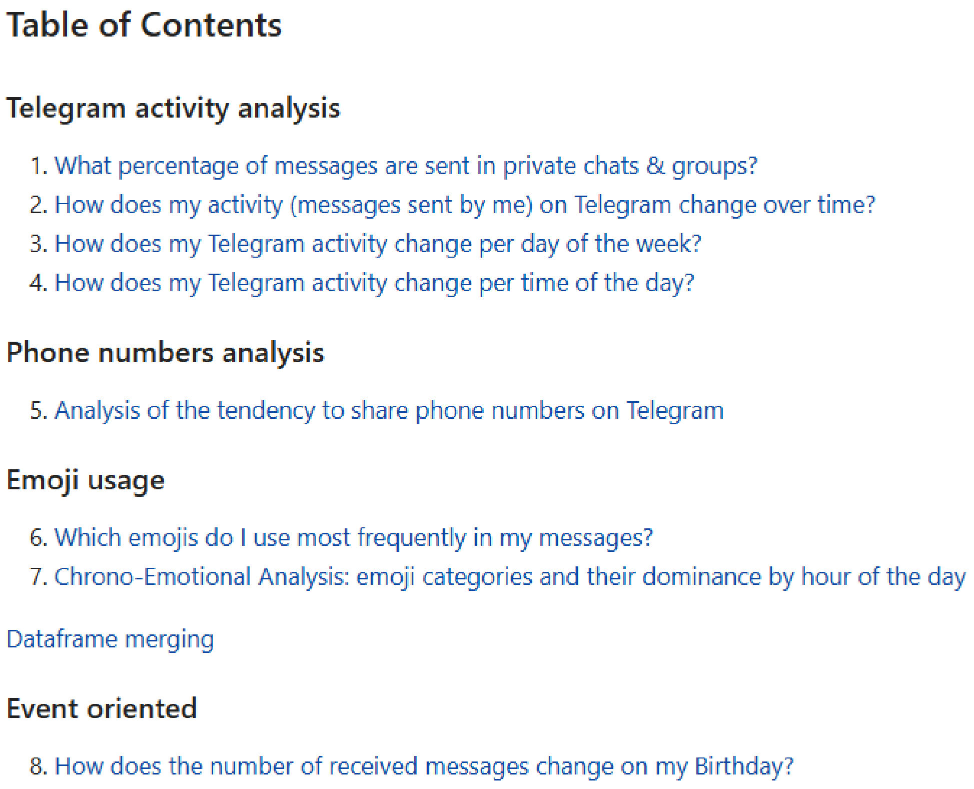


Table of Contents

Telegram activity analysis

1. What percentage of messages are sent in private chats & groups?
2. How does my activity (messages sent by me) on Telegram change over time?
3. How does my Telegram activity change per day of the week?
4. How does my Telegram activity change per time of the day?

Phone numbers analysis

5. Analysis of the tendency to share phone numbers on Telegram

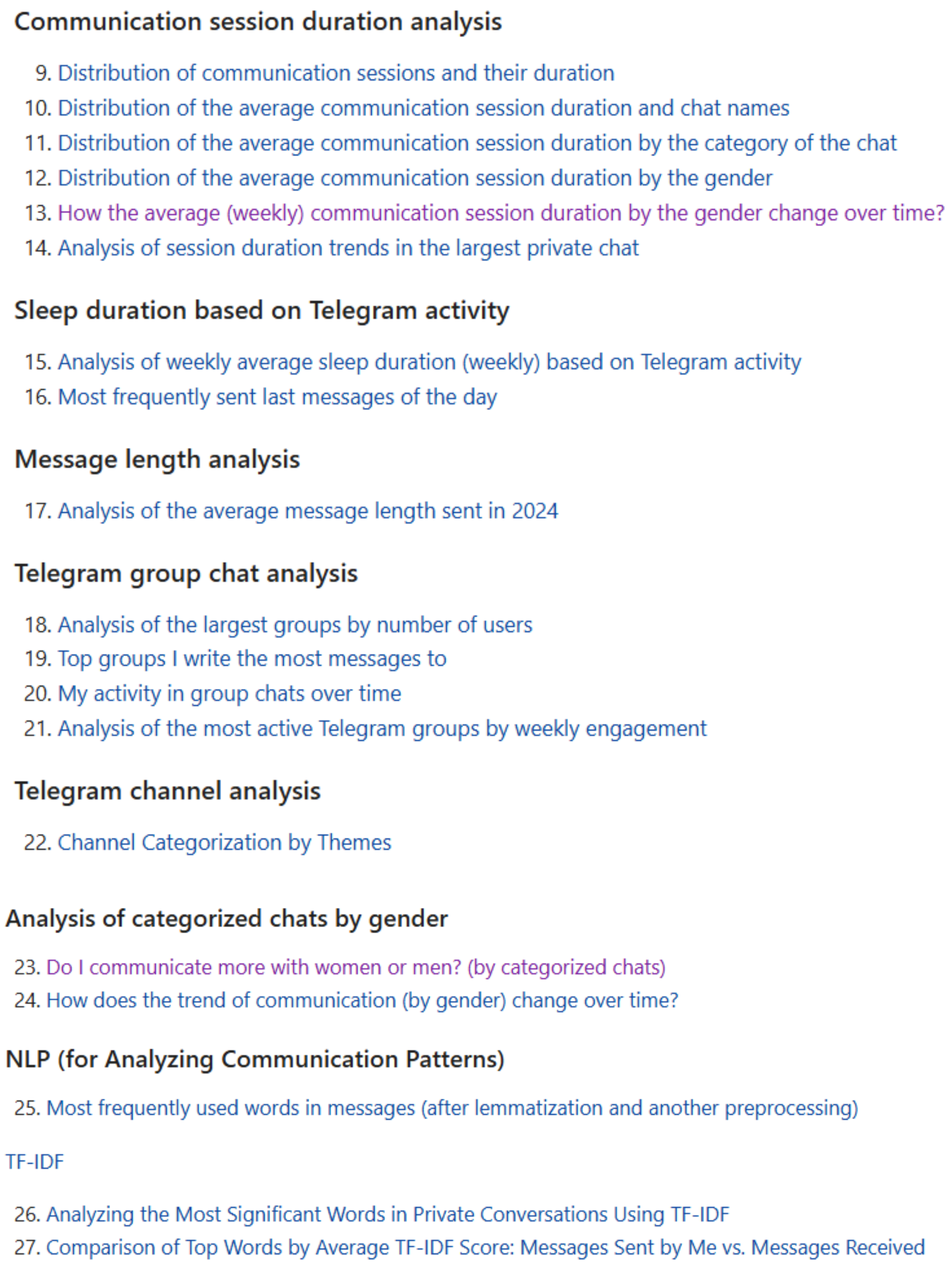
Emoji usage

6. Which emojis do I use most frequently in my messages?
7. Chrono-Emotional Analysis: emoji categories and their dominance by hour of the day

Dataframe merging

Event oriented

8. How does the number of received messages change on my Birthday?



Communication session duration analysis

9. Distribution of communication sessions and their duration
10. Distribution of the average communication session duration and chat names
11. Distribution of the average communication session duration by the category of the chat
12. Distribution of the average communication session duration by the gender
13. How the average (weekly) communication session duration by the gender change over time?
14. Analysis of session duration trends in the largest private chat

Sleep duration based on Telegram activity

15. Analysis of weekly average sleep duration (weekly) based on Telegram activity
16. Most frequently sent last messages of the day

Message length analysis

17. Analysis of the average message length sent in 2024

Telegram group chat analysis

18. Analysis of the largest groups by number of users
19. Top groups I write the most messages to
20. My activity in group chats over time
21. Analysis of the most active Telegram groups by weekly engagement

Telegram channel analysis

22. Channel Categorization by Themes

Analysis of categorized chats by gender

23. Do I communicate more with women or men? (by categorized chats)
24. How does the trend of communication (by gender) change over time?

NLP (for Analyzing Communication Patterns)

25. Most frequently used words in messages (after lemmatization and another preprocessing)

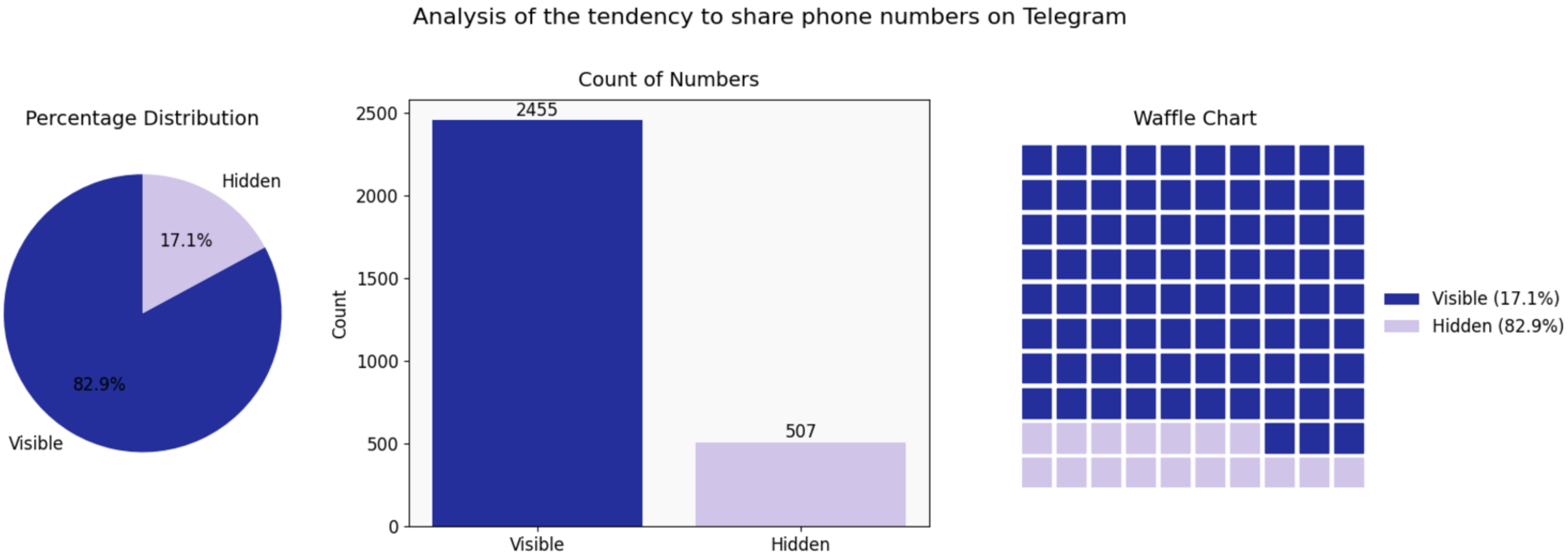
TF-IDF

26. Analyzing the Most Significant Words in Private Conversations Using TF-IDF
27. Comparison of Top Words by Average TF-IDF Score: Messages Sent by Me vs. Messages Received

Exploration process
& Results (shortly)

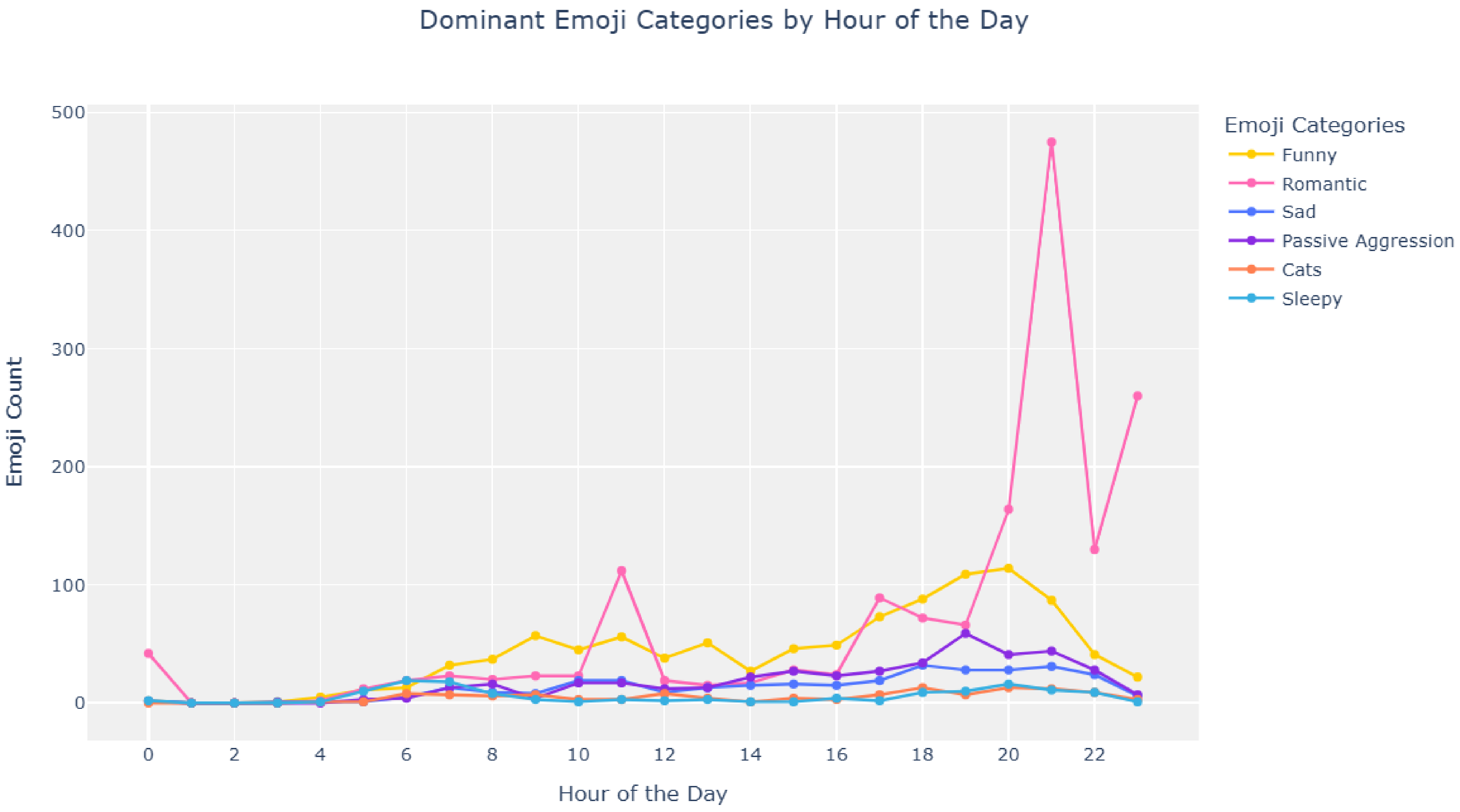
Phone Numbers

"Every sixth person makes their phone number visible on Telegram."



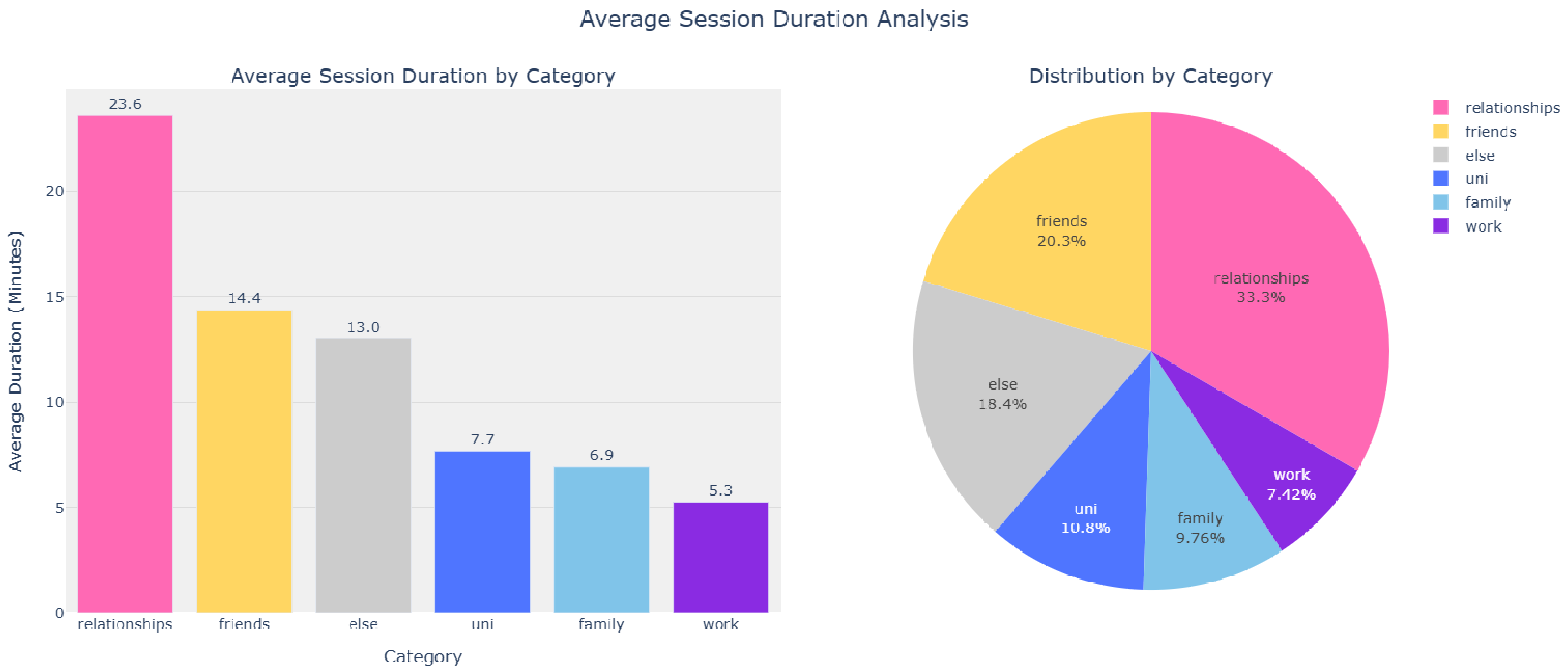
🤞 Emoji

How does my mood change throughout the day?



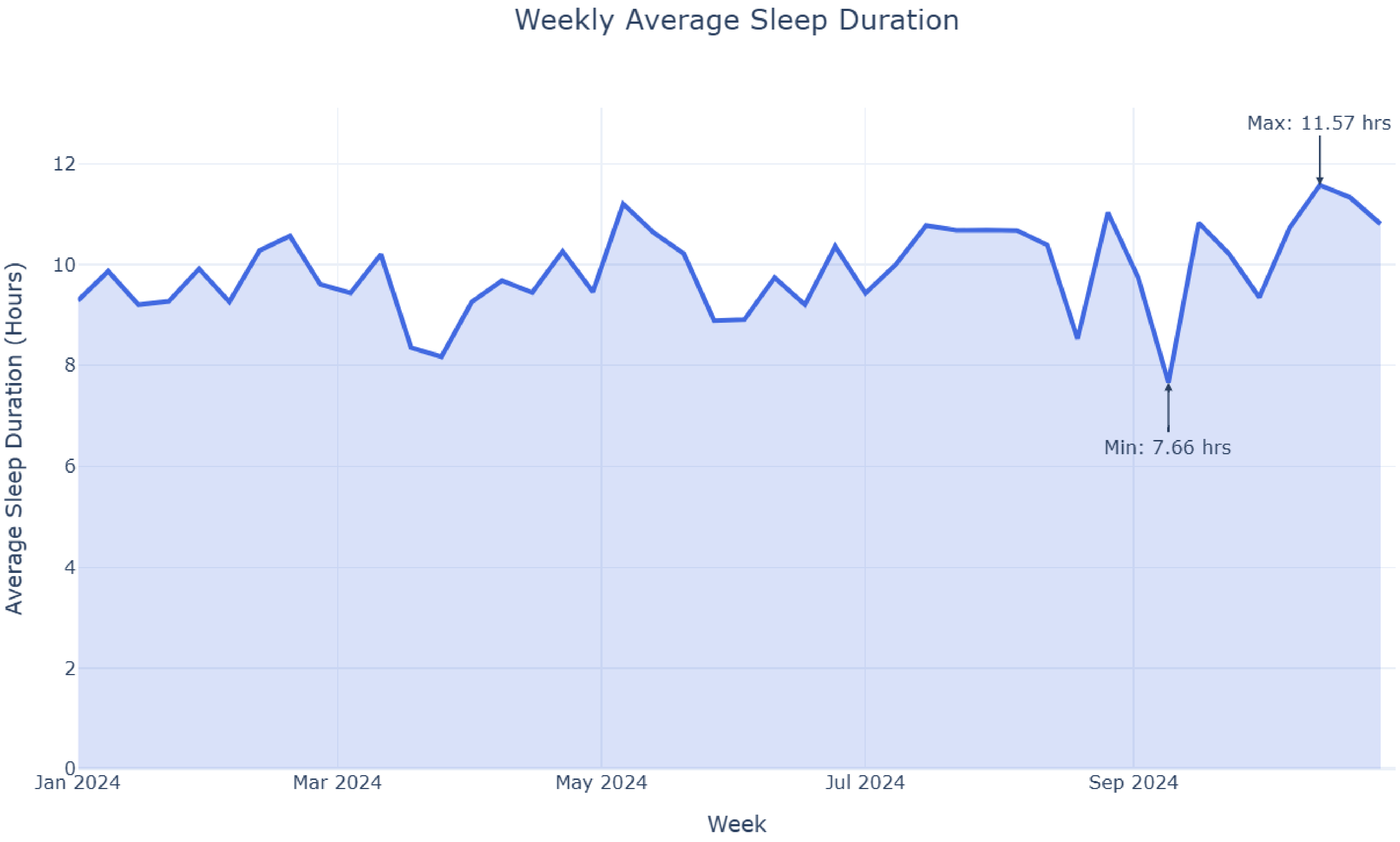
Session Duration

With which category of people do I have the longest communication session?

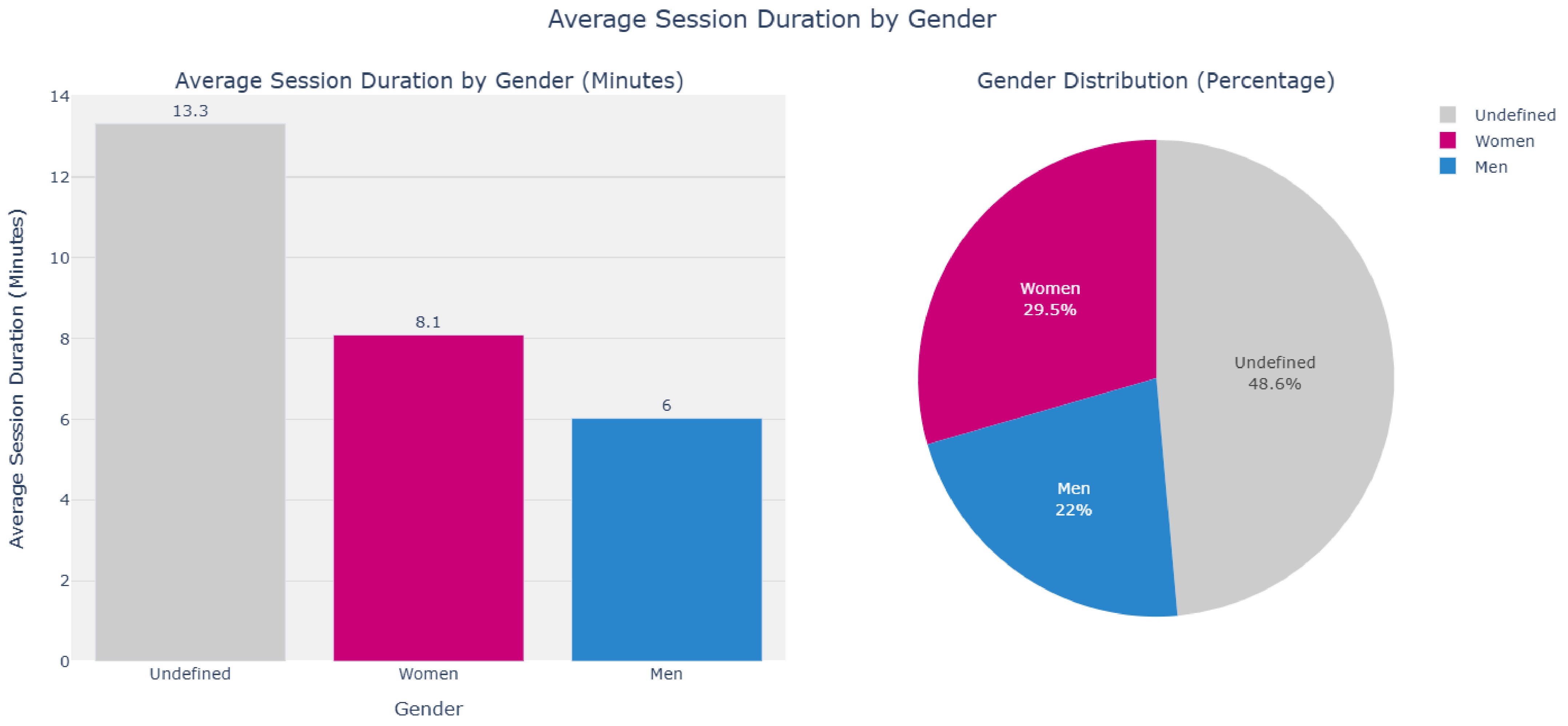


How long do I usually sleep?

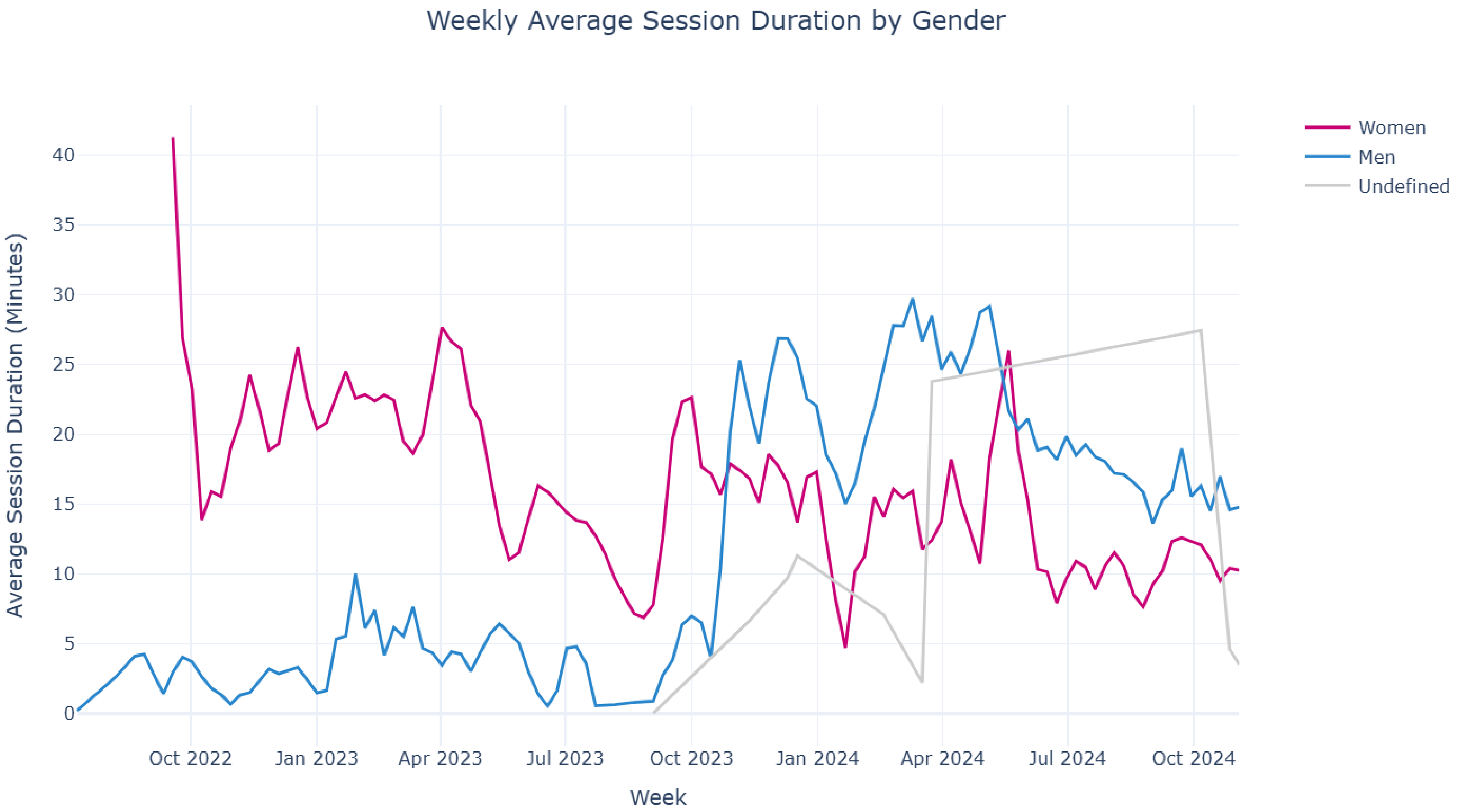
Average: 9.88 hours



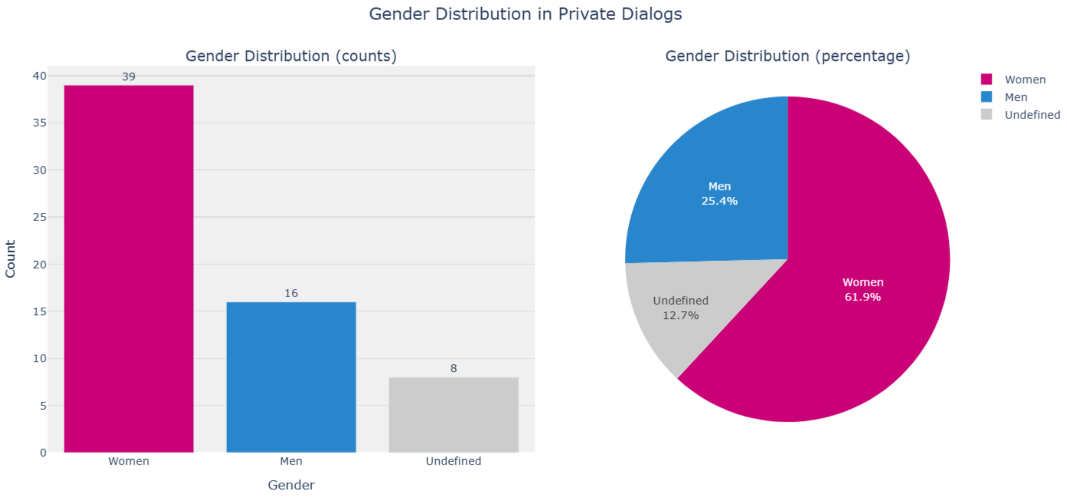
Which gender do I have the longest conversations with?



How has this changed over time?



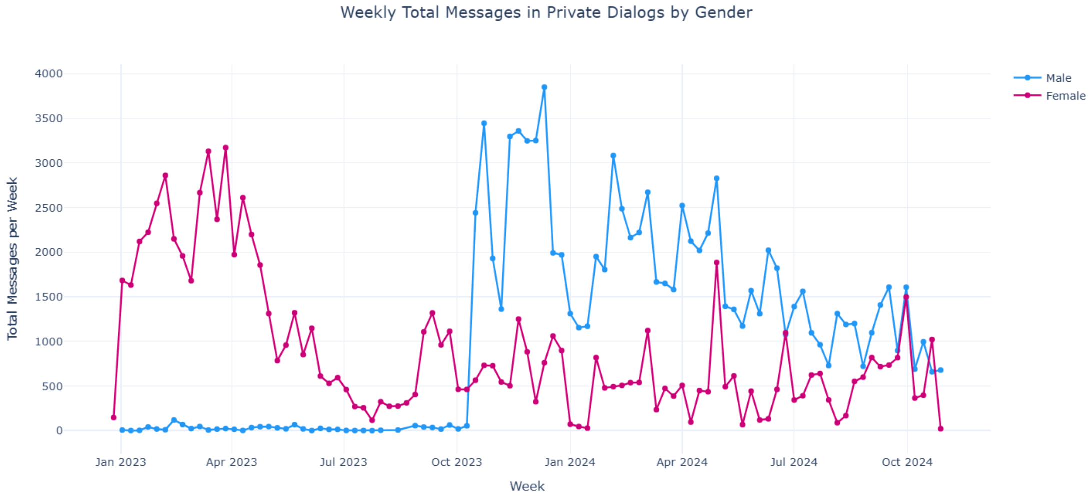
⚥ Genders



Gender Distribution in Private Dialogs

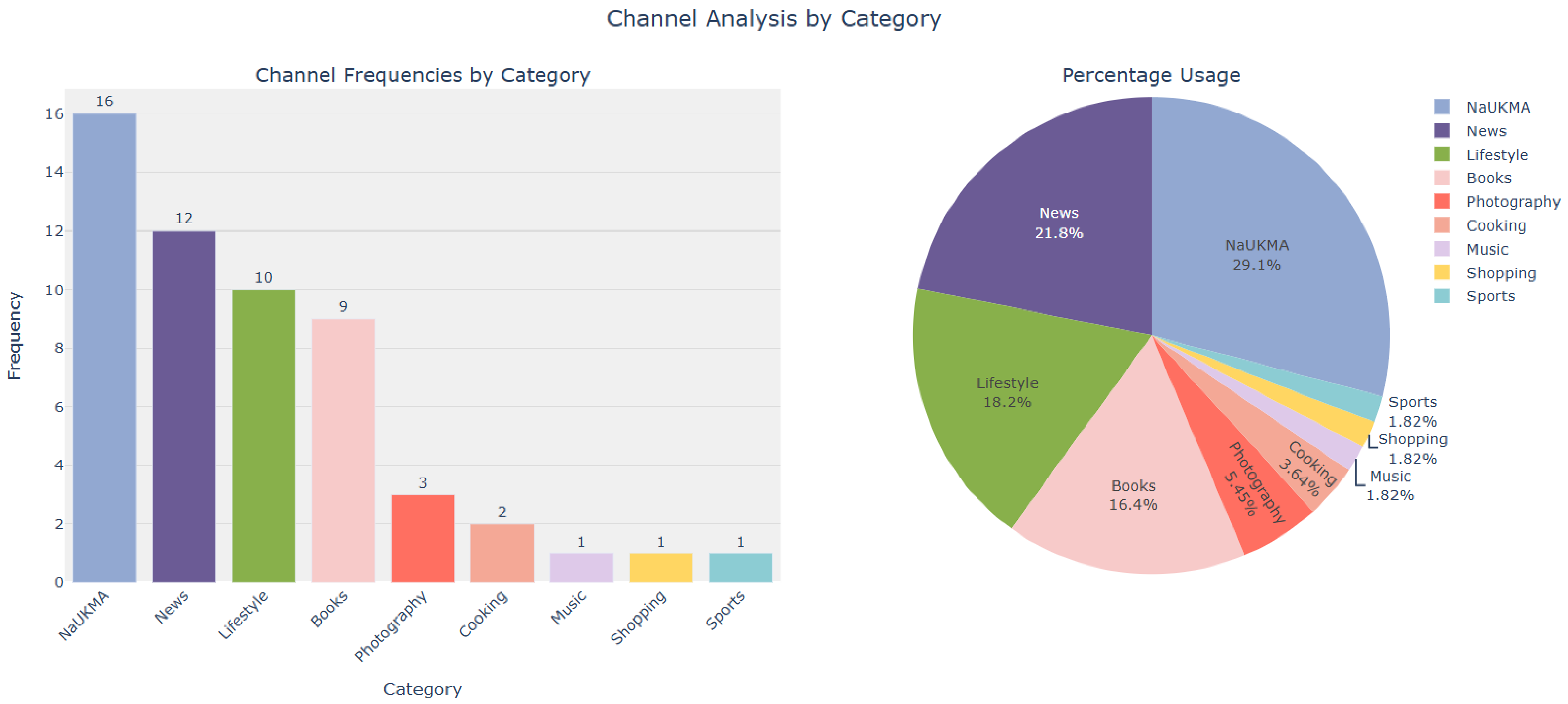
Gender Distribution (counts)

Gender Distribution (percentage)



Weekly Total Messages in Private Dialogs by Gender

Which Telegram channels do I mostly read?



Further work

1. Sentiment analysis
2. Deeper research into personal channels to increase activity there, for example, the channel of the NaUKMA student organization.
3. Using other available data, such as on reactions to messages, and building a model that would analyze the next message I want to send and predict the reaction of society. Or what message I will receive in response to mine!
4. Explore more with NLP

GitHub

repo with the code

To learn more about the insights I was able to find from my personal Telegram data, I invite you to the project repository!

Link: https://github.com/issaravas/Exploratory-Telegram-Data-Analysis/tree/main

Thanks for attention!
I will be happy to answer all questions about the project!