HOME ALONe

behavioural analysis of Reddit communities during the COVID-19 pandemic

made by: anastasiia bakalyna and anna romanchuk
group: no.12

course: computational social science
professor: andrew kurochkin
presentation date: 12/09/25

agenda

1. data collection
2. dataset overview
3. EDA workflow
4. research questions
5. key insights
6. future work
7. GitHub repository

data collection

sources:
– Pushshift Reddit Archive (comments only)

targeted subreddits:
r/covid19, r/covid19_support, r/coronavirus,
r/depression, r/anxiety, r/offmychest,
r/mentalhealthsupport

tools:
– Python
– reddit-dump-extractor
– Google Colab

exploratory data analysis workflow

timeframe: feb–sep 2020

raw size: ~140GB ZST compressed files

final cleaned dataset: ~1.8M rows

challenges:
- huge data size
- inconsistent formatting
- mixed-timezone UTC
- corrupted rows requiring chunk-based cleaning

dataset description

final dataset fields:

- id
- author
- subreddit_clean
- created_utc
- month
- text
- text_len
- score

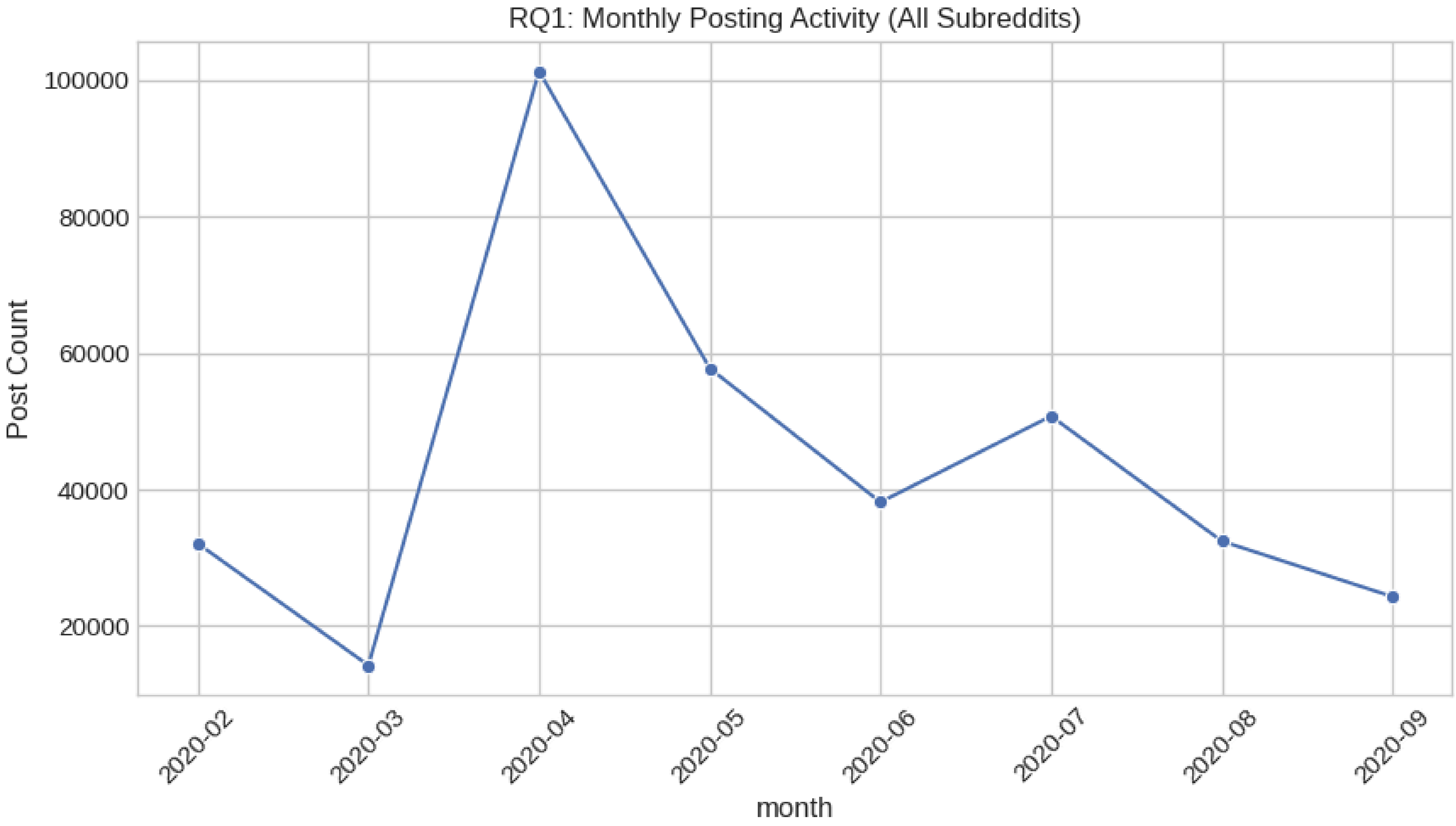
total posts: 1.8M
subreddits included: 7
months covered: 8

EDA workflow overview

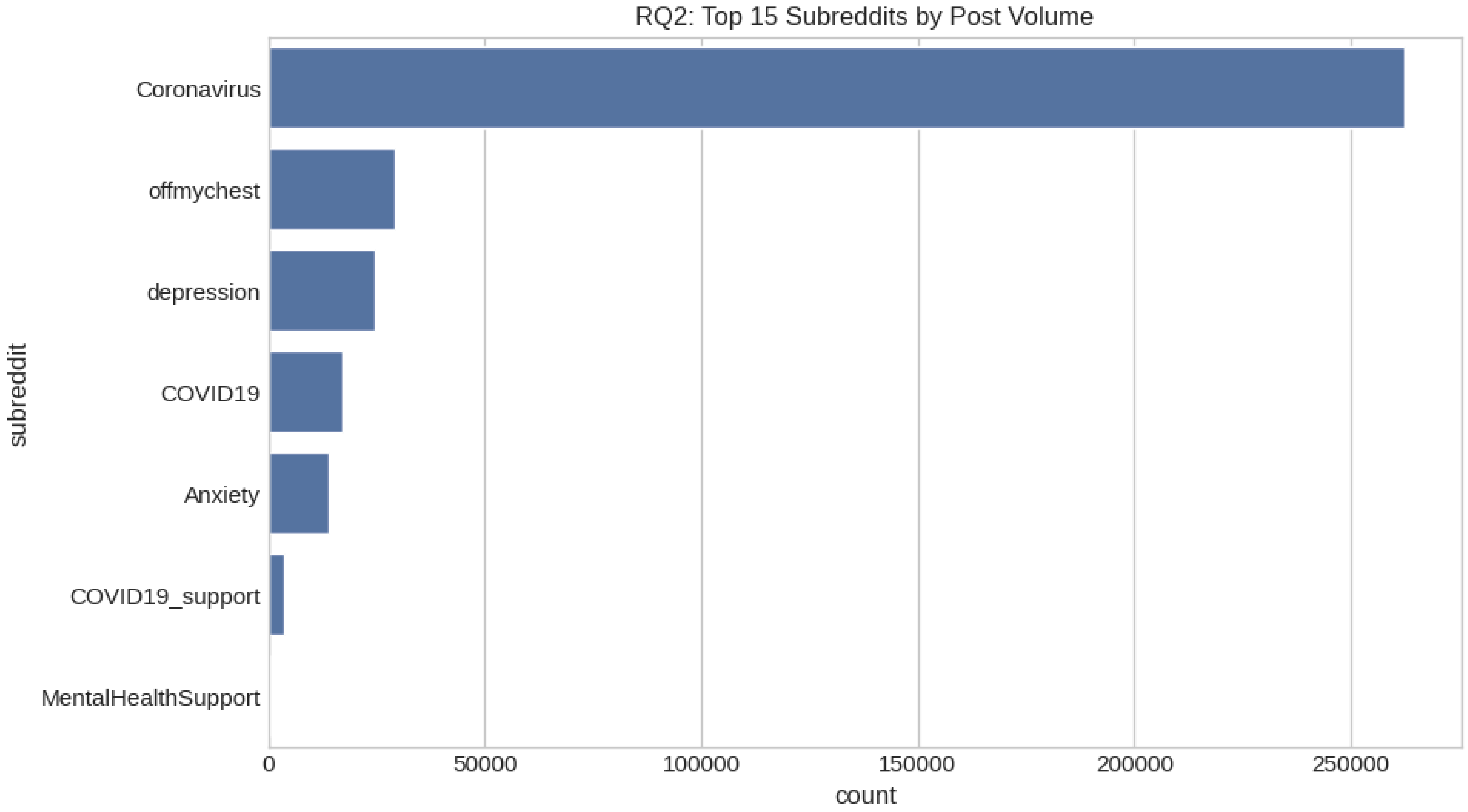
main steps:

- cleaning & normalization of timestamps
- removal of corrupted rows
- subreddit mapping and noise filtering
- text length engineering
- aggregation by month, author, subreddit
- behavioural pattern extraction

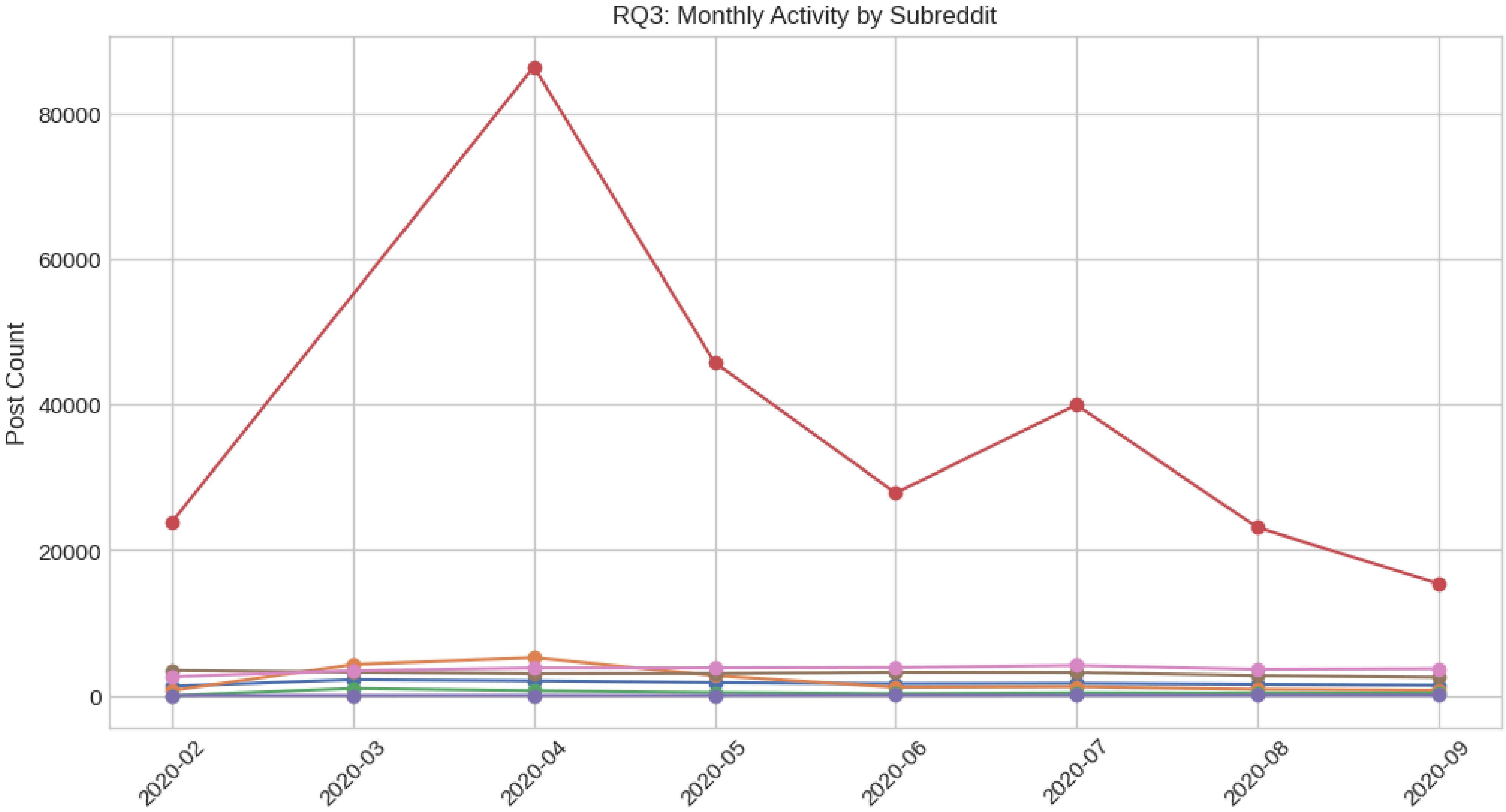
RQ1: monthly activity



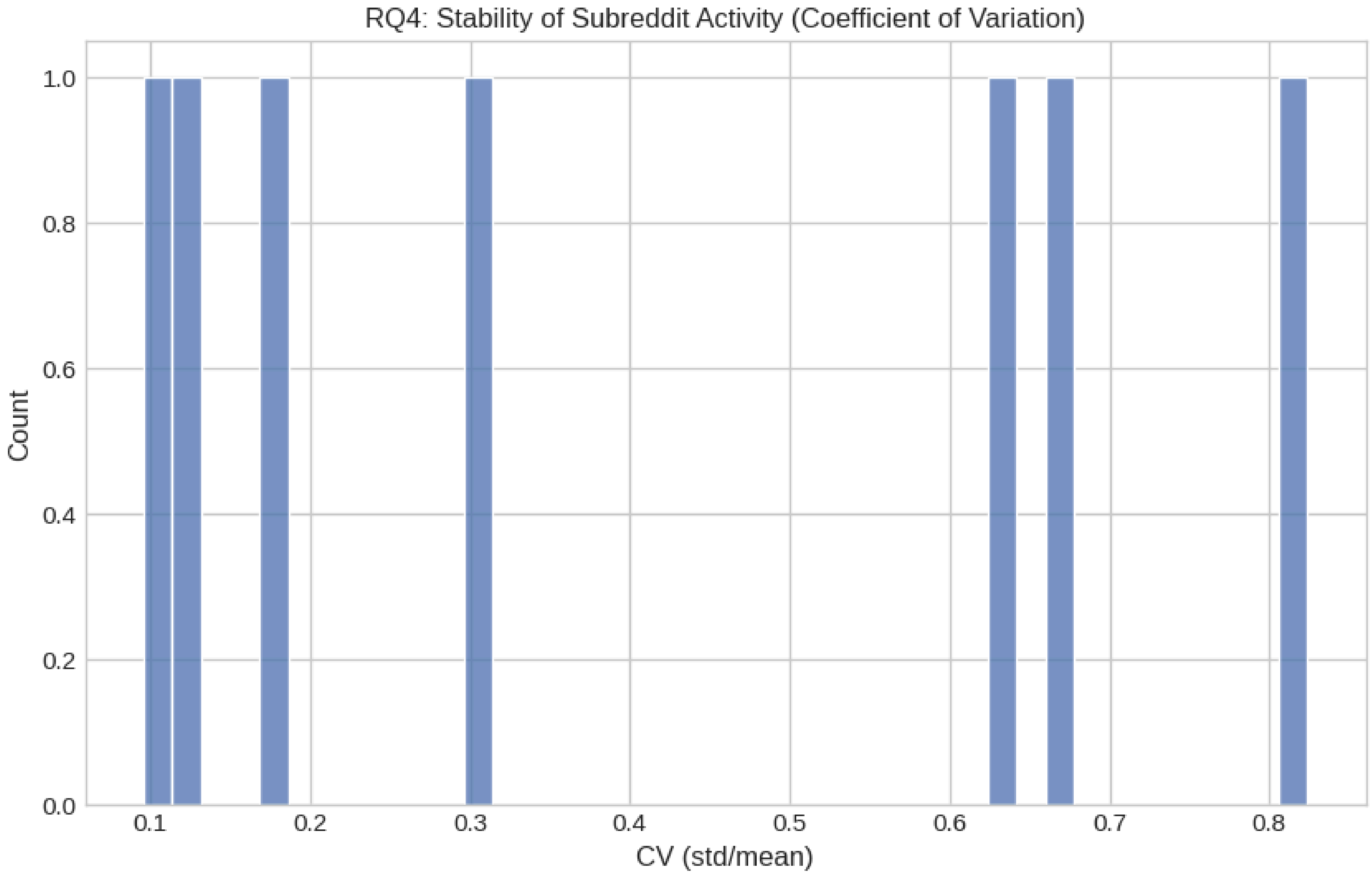
RQ2: subreddit size comparison



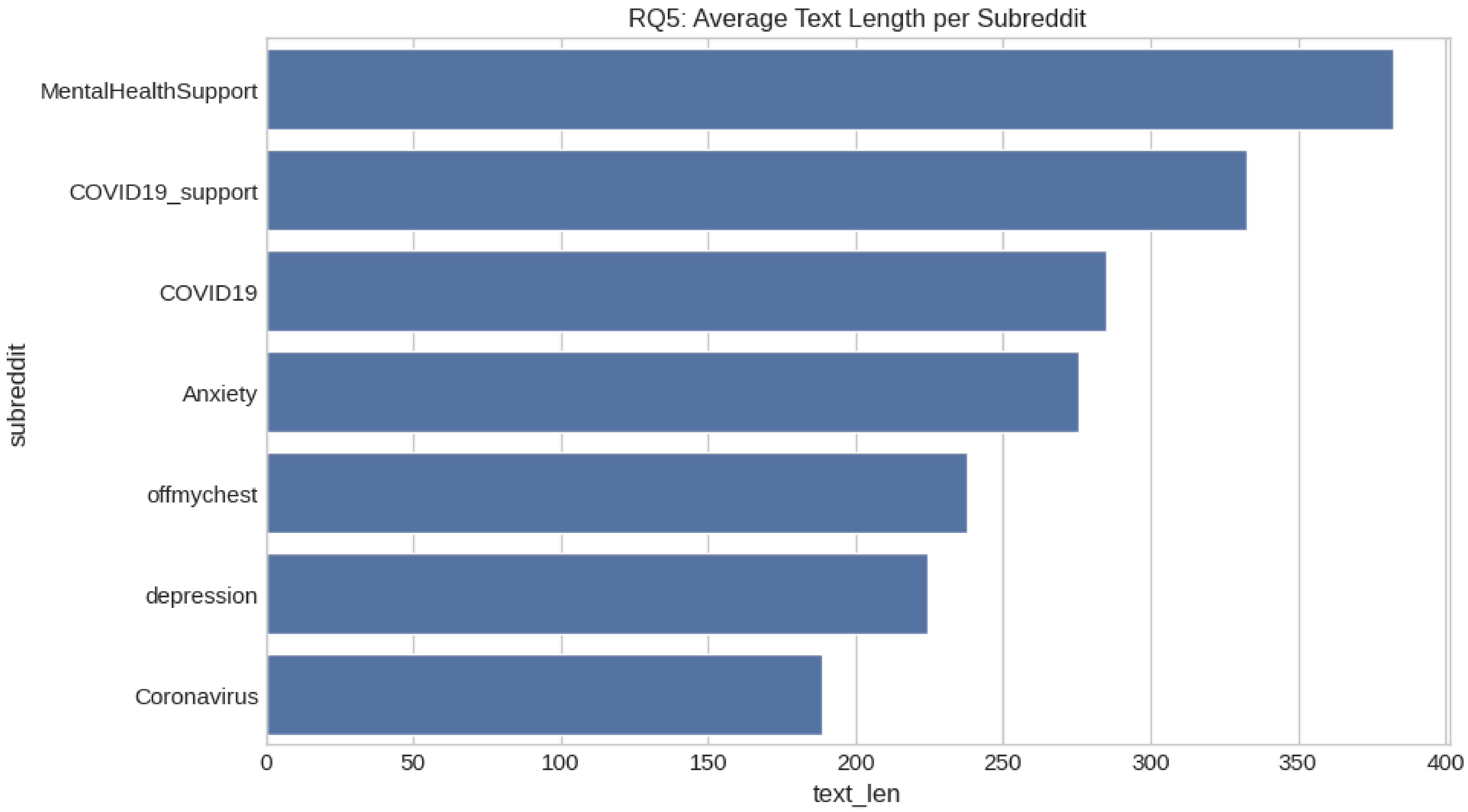
RQ3: monthly activity per subreddit



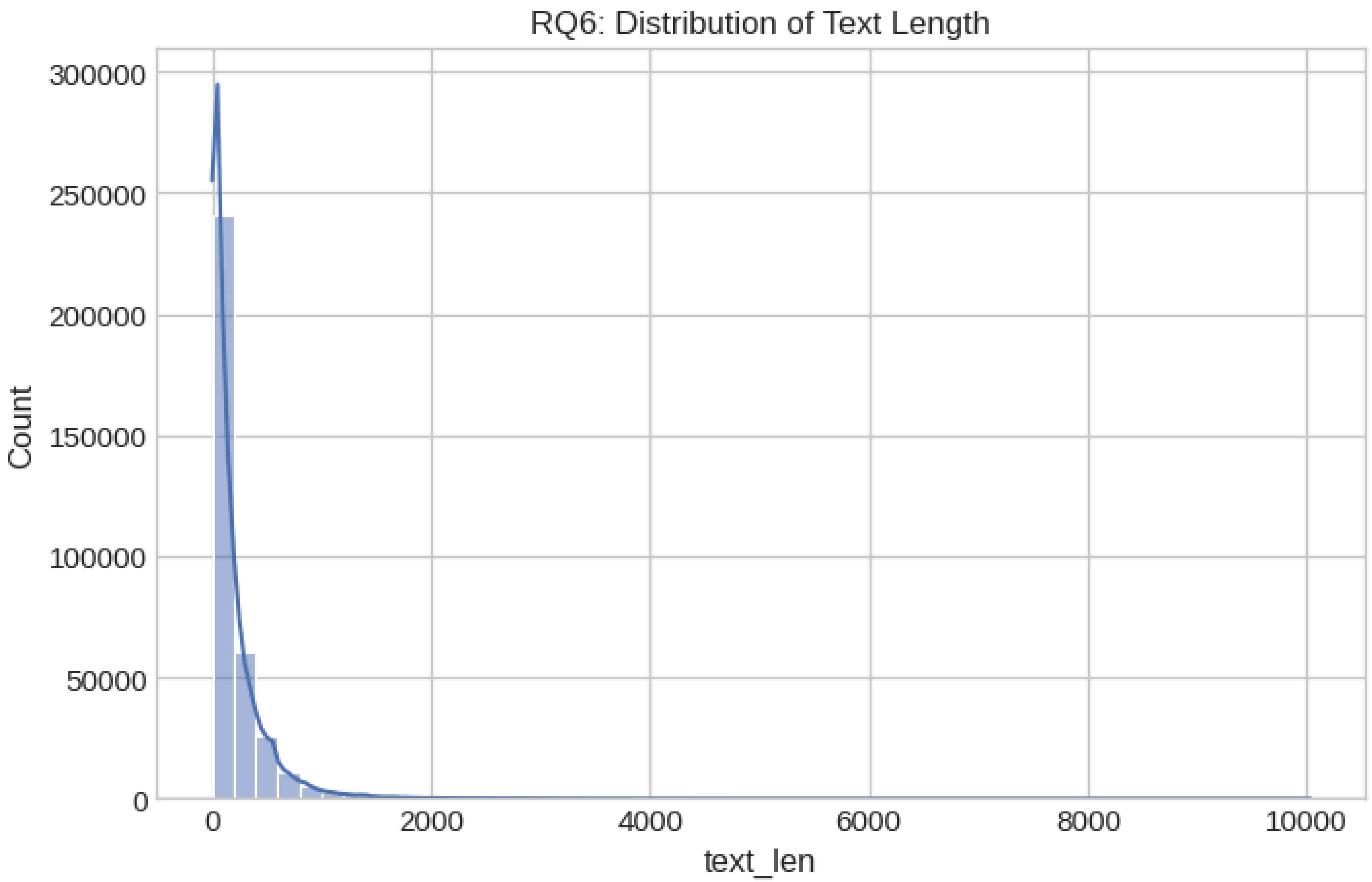
RQ4: stability & volatility



RQ5: text length behaviour



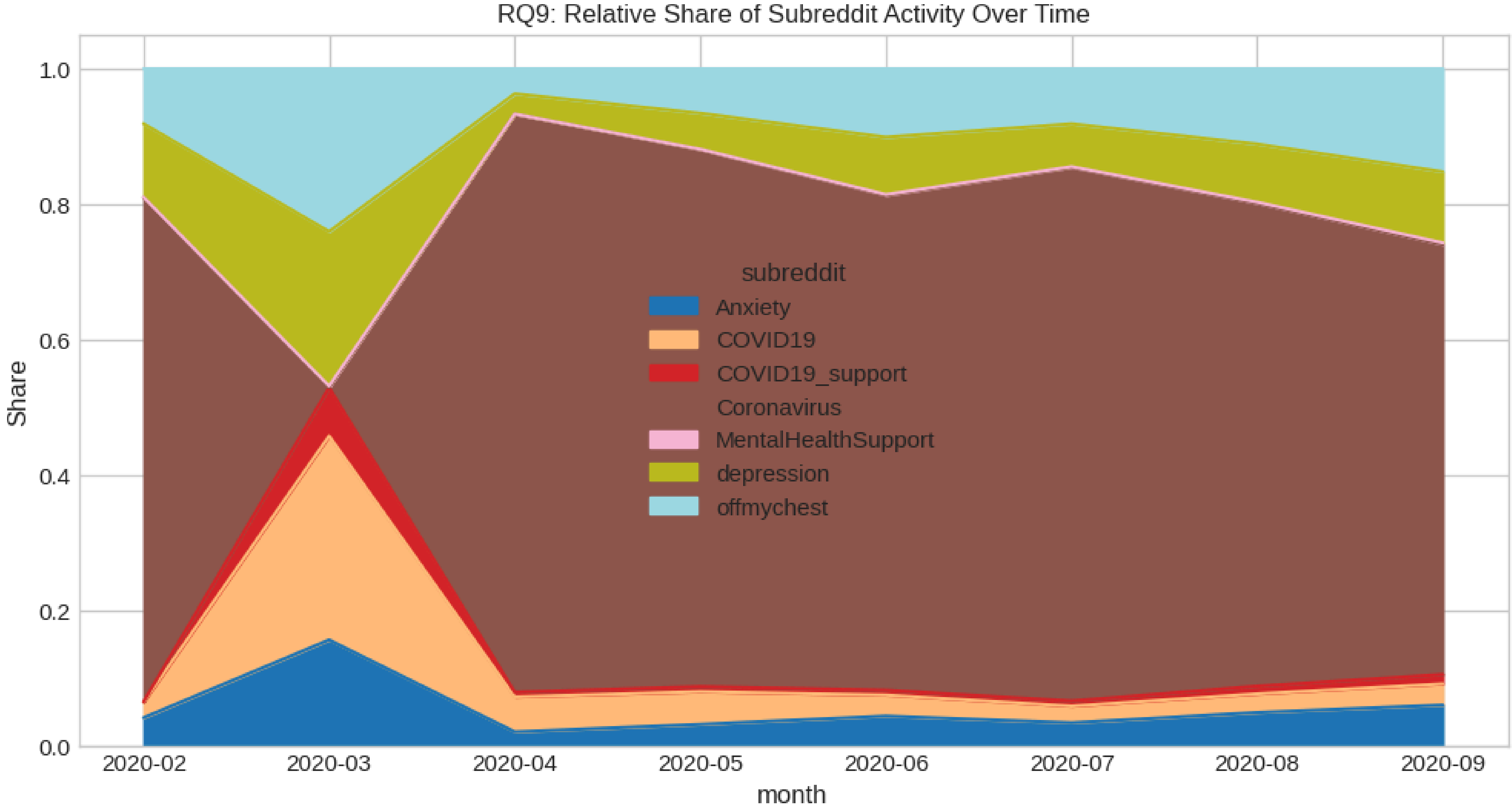
RQ6: stability & volatility



final insights

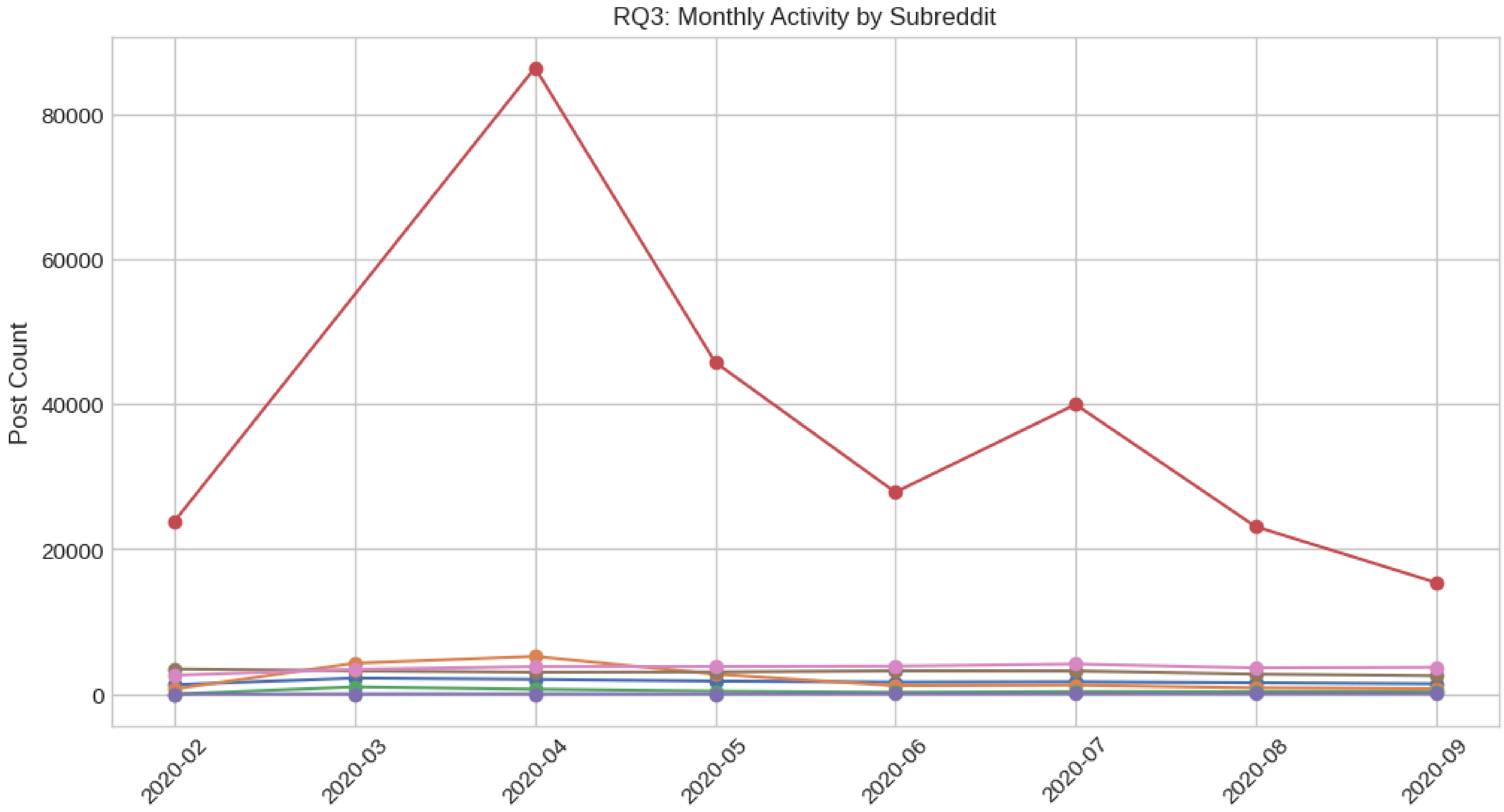
information shock → emotional spillover

in March–April 2020, COVID-related subreddits spike first, as users search for urgent information. however, starting from May, their relative share drops sharply.



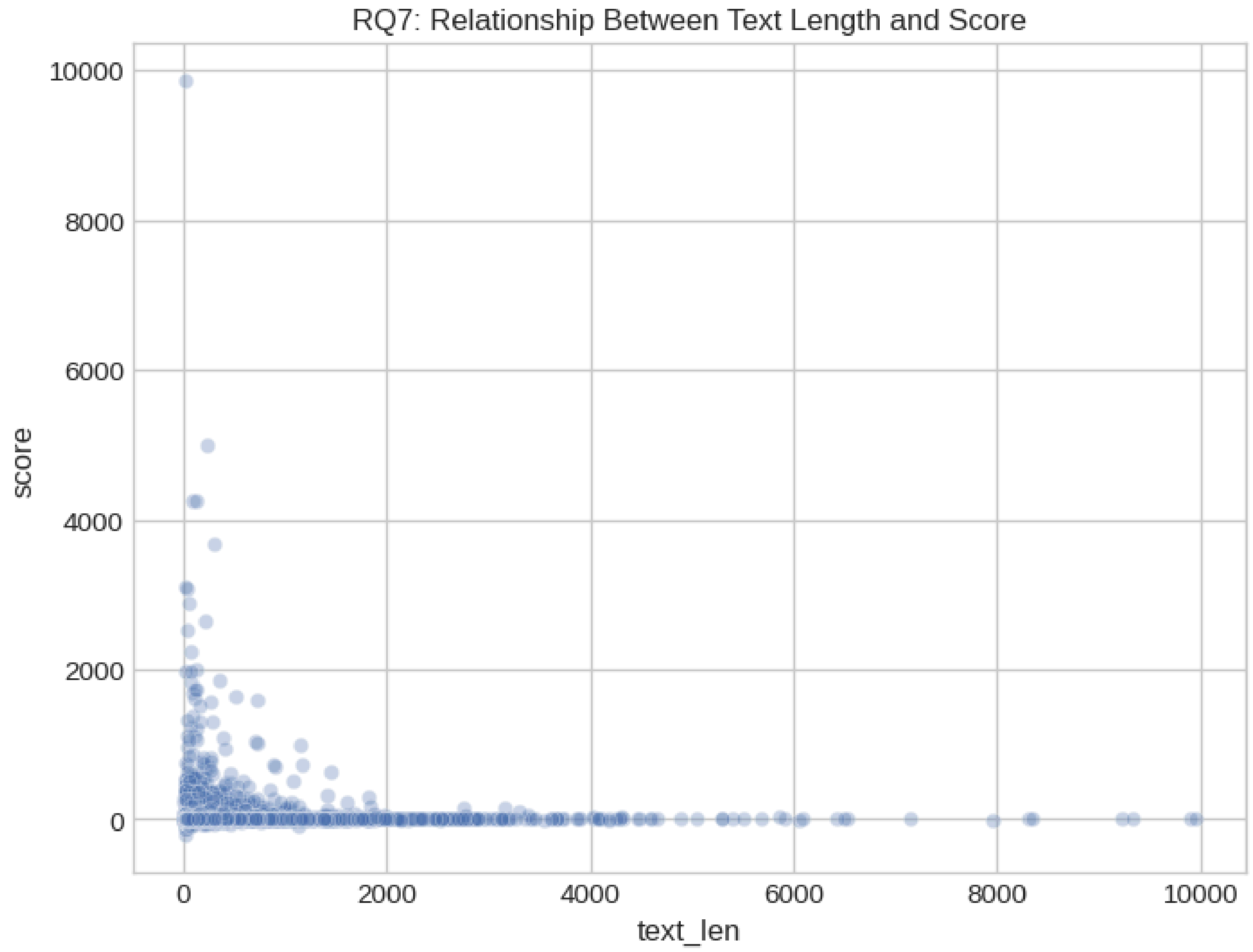
emotional communities stay active even after covid interest declines

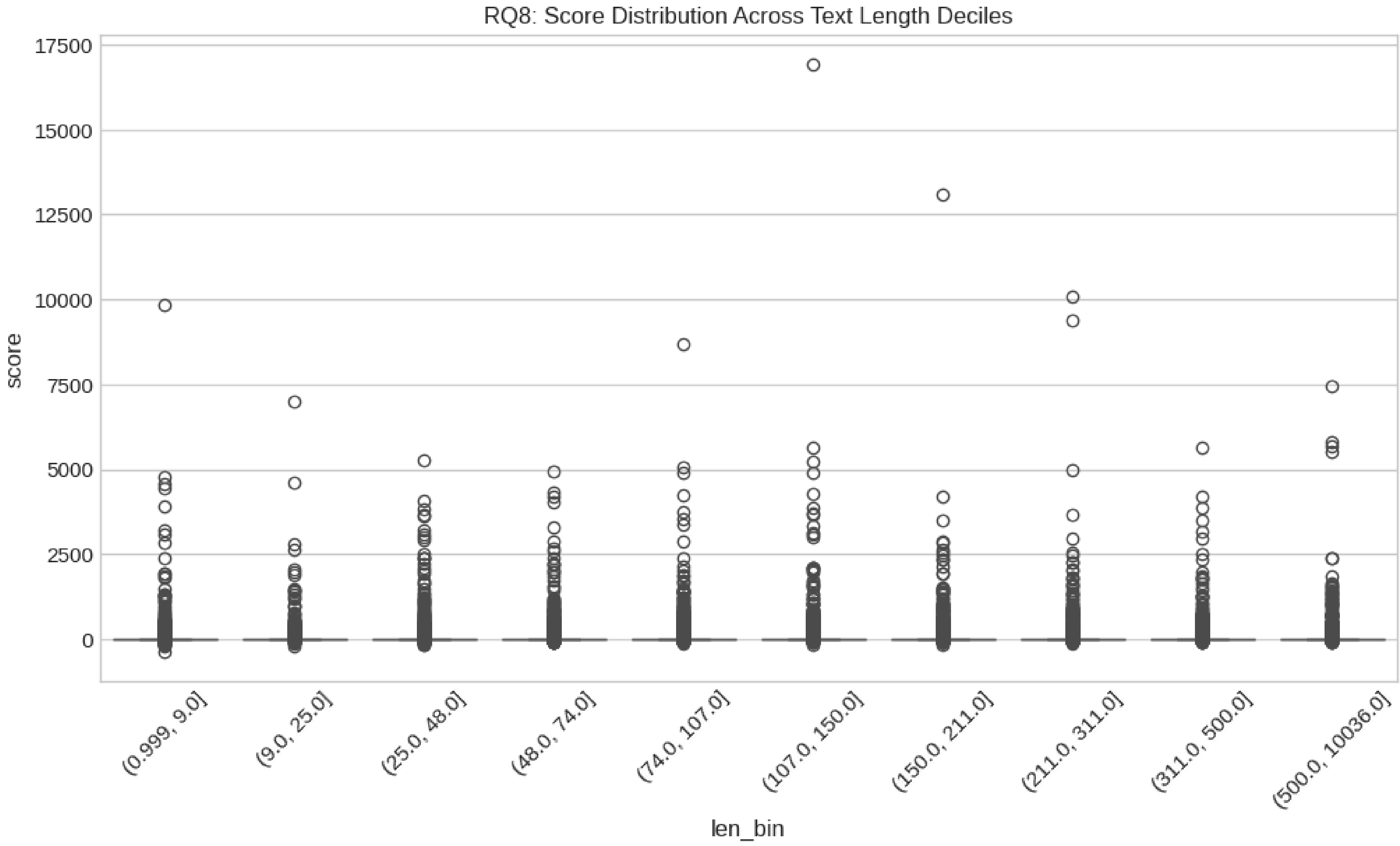
short explanation:
 covid subreddits show a classic crisis peak — sharp rise and sharp decline.
 but emotional support communities (offmychest, depression, anxiety) stay consistently high across all months.



engagement depends on emotion, not length

there is no meaningful correlation between post length and user score.
short posts can receive very high engagement; long posts may receive none.

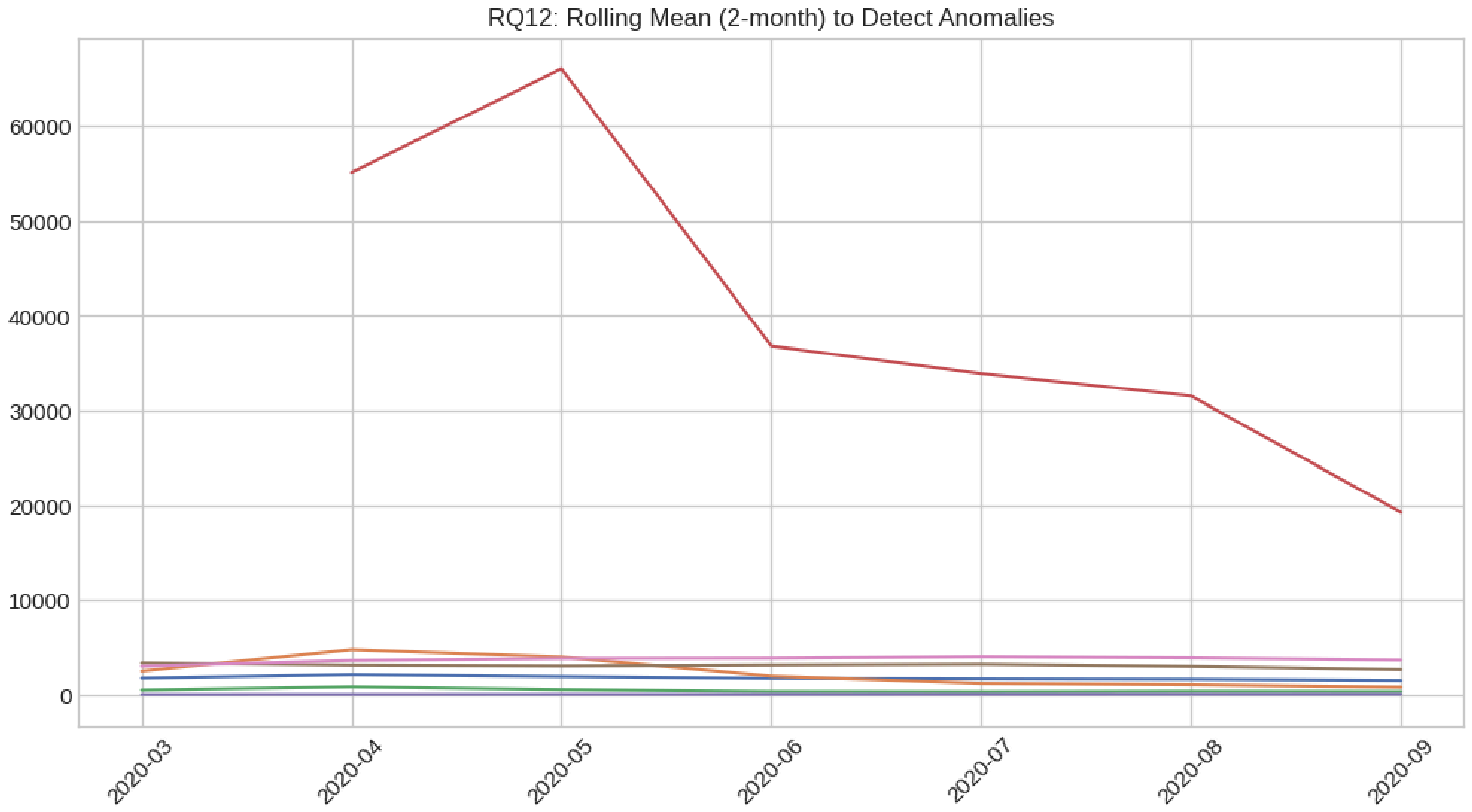




anomaly detection reveals stress peaks across subreddits

rolling averages reveal several behavioural anomalies:

- extreme surge in covid19 activity in early lockdown,
- unexpected rise in emotional posting toward late summer,
- irregular spikes in anxiety.



future work

– sentiment analysis
– TF-IDF / topic modeling
– user segmentation
– subreddit transition graphs
– day/night posting patterns
– merging comments + submissions
– advanced anomaly detection
– cross-language behavioural analysis

github repository link

https://github.com/AnnaRomanchuk/CSS-Reddit-MentalHealth