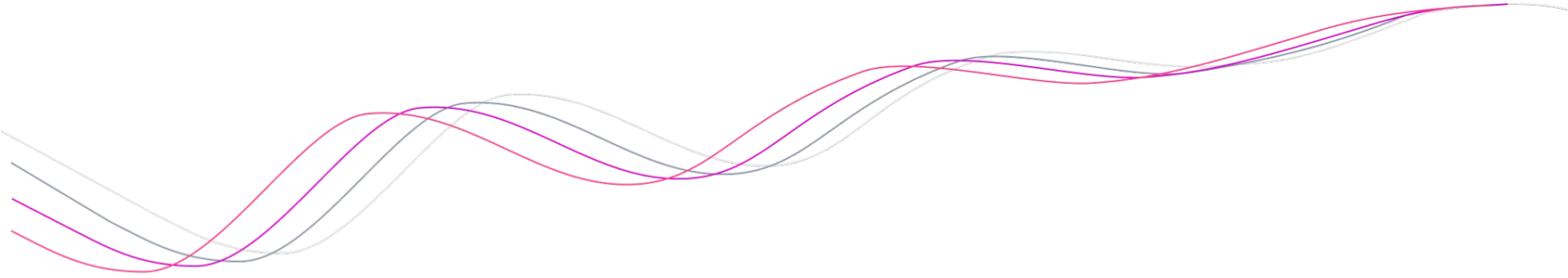


Discord Community Dynamics: Analyze Ocean Protocol Interactions



Anamaria Loznianu

Overview

1. [Preprocessing](#)
2. [General trends](#)
3. [Correlations](#)
4. [Community questions](#)
5. [Community activity](#)
6. [Scam & Spam](#)
7. [Technical issues](#)
8. [Prediction model](#)

Preprocessing

Target: Clean up the dataset, extract additional content, date and reactions features, remove irrelevant & empty messages, filter bot messages.


The provided dataset comprises 84,754 records across six columns:


- **Channel** - 84,754 entries
- **AuthorID** - 84,754 entries
- **Author** - 84,754 entries
- **Date** - 84,754 entries
- **Content** - 64,609 entries
- **Attachments** - 1959 entries
- **Reactions** - 6073 entries

The following steps were undertaken to enhance its utility and richness:

- Separated the Channel in two columns: **ChannelName** and **ChannelID**
- Renamed Author column into **AuthorName**
- Extracted additional Date features: **Time**, **Year**, **Hour**, **DayOfWeek**
- Filtered all entries that don't have **Content**
- Filtered bot messages: **GitHub**, **MEE6#4876**, **OceanDiffusion#4502**, **OceanGPT#0740**, **Ocean Protocol • TweetShift**
- Filtered **join server** messages
- Filtered **good morning** and **github notifications** channels: **911643560594505809**, **773926934601269248**
- Extracted additional Content features: **WordCount**, **CharCount**, **HasAttachment**, **ReactionCount**

The computed dataset comprises **30,883** entries across **17** columns: **ChannelName**, **ChannelID**, **AuthorName**, **AuthorID**, **DateTime**, **Content**, **Reactions**, **Attachments**, **Date**, **Time**, **Year**, **Hour**, **DayOfWeek**, **WordCount**, **CharCount**, **HasAttachment**, **ReactionCount**.

 Note: The computed file is available in the [github](#) repository and the code in the [colab](#).

 Note: This report doesn't cover the channels that aren't in the dataset we got, like **"tech-issues"** (forum-channel). With the access I have, I could only grab info from the last 10 active threads but no more.

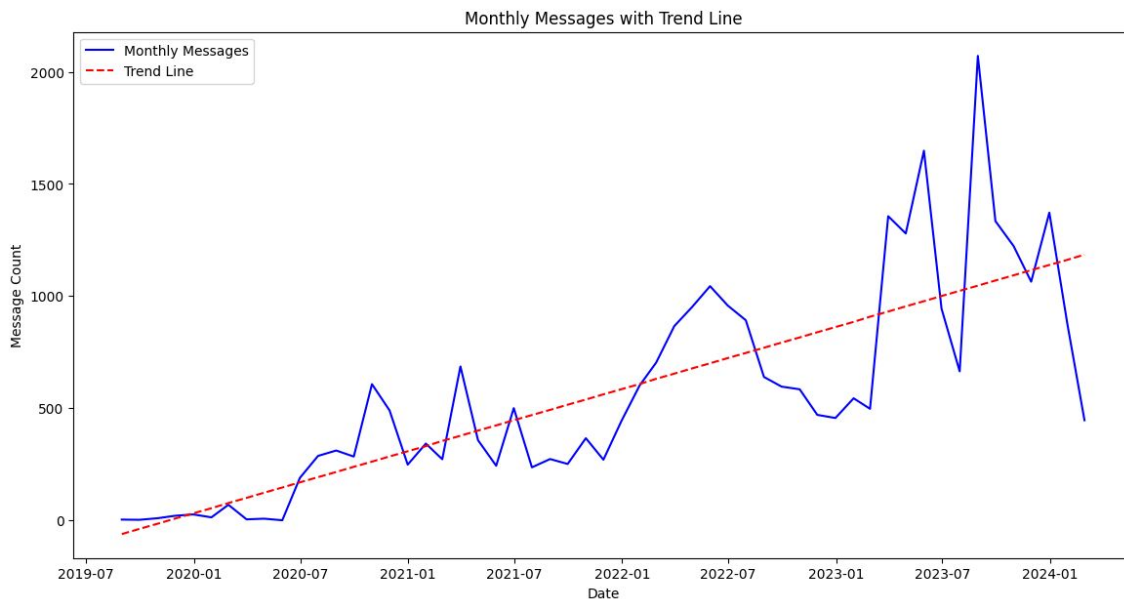
General trends

Target: Analyze the evolution of message numbers over time on a weekly, monthly, or quarterly basis. Determine if all channels follow the same trend and identify outlier periods with potential causes.

The [general trend](#) line plotted over the monthly message data shows an **upward** trajectory, indicating that the number of messages has been generally increasing over time. The positive slope of the trend line, approximately (23.08), suggests that, on average, the monthly message count increases by this amount each month.

The statistical significance of this trend is supported by a very small p-value, far below any conventional threshold, indicating that the observed increase in message volume is **statistically significant and not due to random chance**.

In summary, ***the general trend in the data is an increase in the number of messages over time, suggesting growing engagement or activity within the server community. This positive trend is a strong indication of an active, expanding community.***



General trends

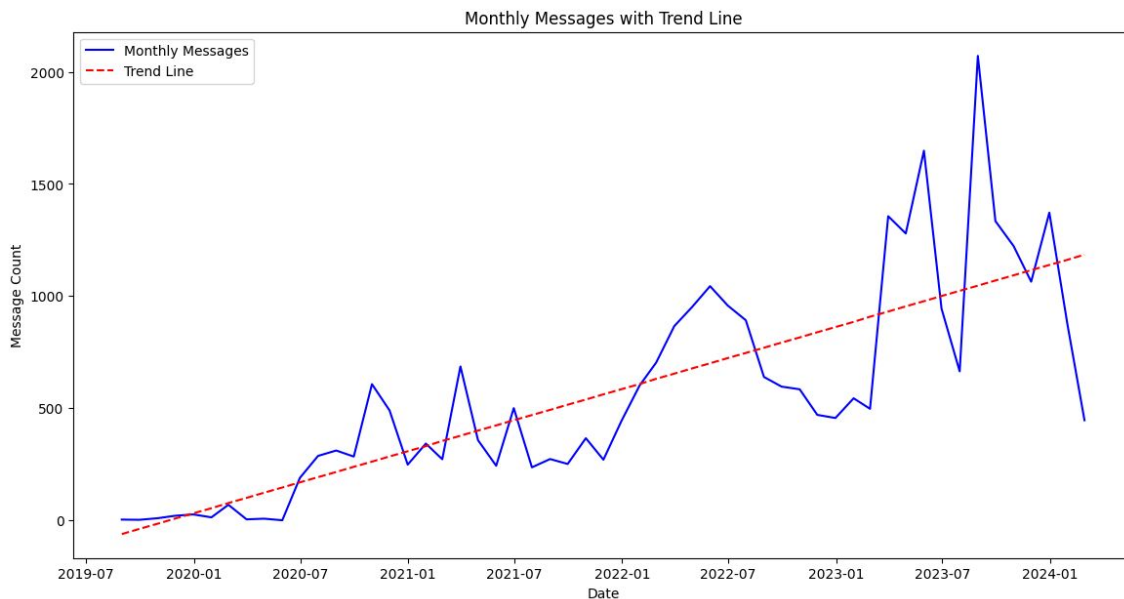
Target: Analyze the evolution of message numbers over time on a weekly, monthly, or quarterly basis. Determine if all channels follow the same trend and identify outlier periods with potential causes.

Growth in Message Volume: Starting with 60 messages in 2019, there has been a significant increase in total messages, peaking at 13,996 messages in 2023. This growth indicates an expanding active engagement within the community.

Word and Character Count: The average word count and character count per message have generally increased, peaking in 2020 with an average of 23.31 words and 136.61 characters. Thereafter, there's a slight fluctuation but the counts remain relatively stable, suggesting consistency in the depth of conversation over time.

Reaction Count: The average reaction count per message has gradually increased, from 0.017 in 2019 to 0.240 in 2023, before slightly decreasing to 0.190 in 2024. This trend signifies a growing interaction among community members over the years.

Attachment Rate: The rate of messages with attachments shows an interesting pattern, with a significant jump in 2022 (5.49%) compared to earlier years.



General trends

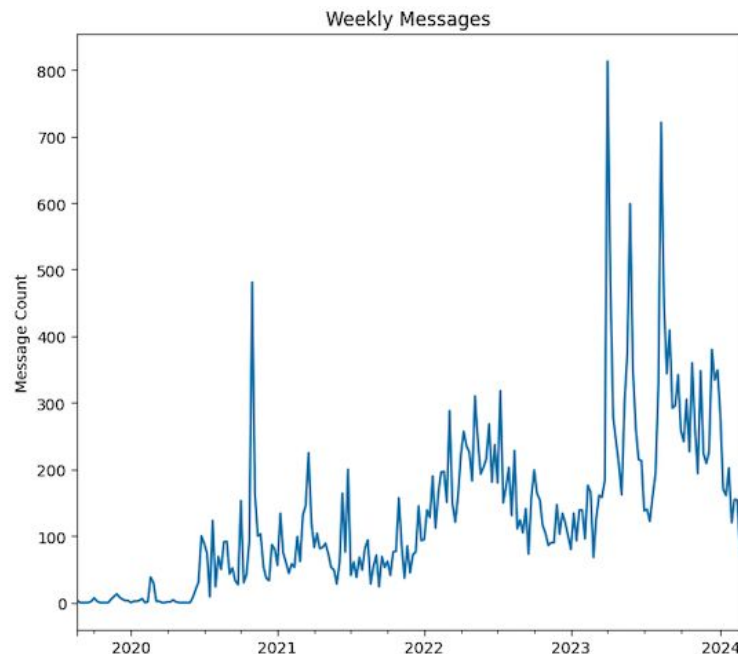
Target: Analyze the evolution of message numbers over time on a weekly, monthly, or quarterly basis. Determine if all channels follow the same trend and identify outlier periods with potential causes.

Weekly Insights

High Variability: The weekly data shows significant fluctuations, indicating that engagement can vary greatly from week to week. This suggests that specific events, discussions, or content releases can have a substantial immediate impact on community activity.

Engagement Opportunities: The variability in weekly activity highlights potential opportunities for increasing engagement through targeted weekly events or content. By analyzing weeks with high activity, you can identify topics or events that resonate with your community and plan similar activities to maintain or increase engagement.

Monitoring Trends: Regularly monitoring weekly trends can help quickly identify shifts in community engagement, allowing for timely interventions to address decreases in activity or capitalize on increasing interest in certain topics.



General trends

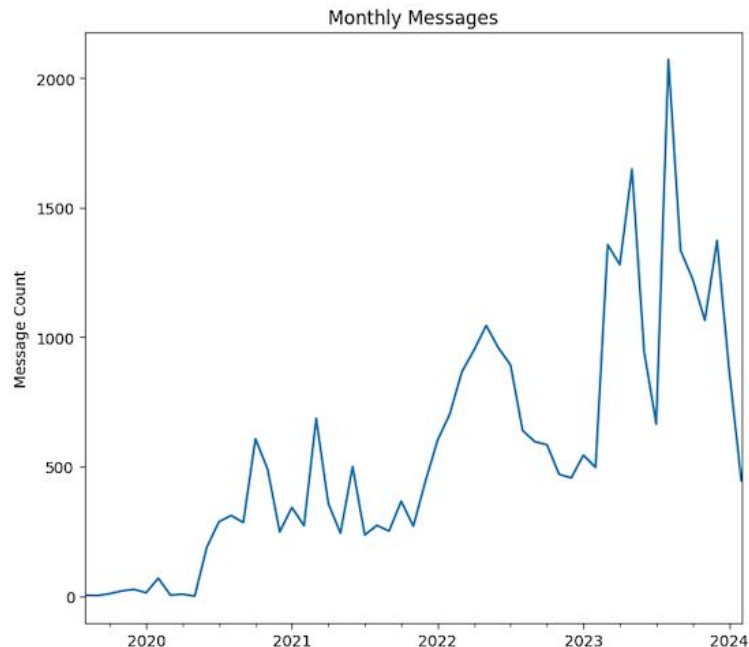
Target: Analyze the evolution of message numbers over time on a weekly, monthly, or quarterly basis. Determine if all channels follow the same trend and identify outlier periods with potential causes.

Monthly Insights

Smoothing of Fluctuations: Monthly aggregation smooths out the short-term fluctuations seen in weekly data, making it easier to identify more sustained trends in engagement over time.

Growth Trends: The linear regression analysis of the monthly message volumes shows a statistically significant upward trend, indicating that the community is **growing** in terms of engagement. **This long-term positive trend is a good sign of a healthy and active community.**

Seasonal and Event Impact: While the seasonal decomposition will specifically highlight cyclical patterns, monthly trends also reflect the impact of seasonal events, holidays, or major community milestones. The launch impact of Ocean V3, V4, Data Farming, Predictoor and the pool draining attacks are visible in this monthly view.



General trends

Target: Analyze the evolution of message numbers over time on a weekly, monthly, or quarterly basis. Determine if all channels follow the same trend and identify outlier periods with potential causes.

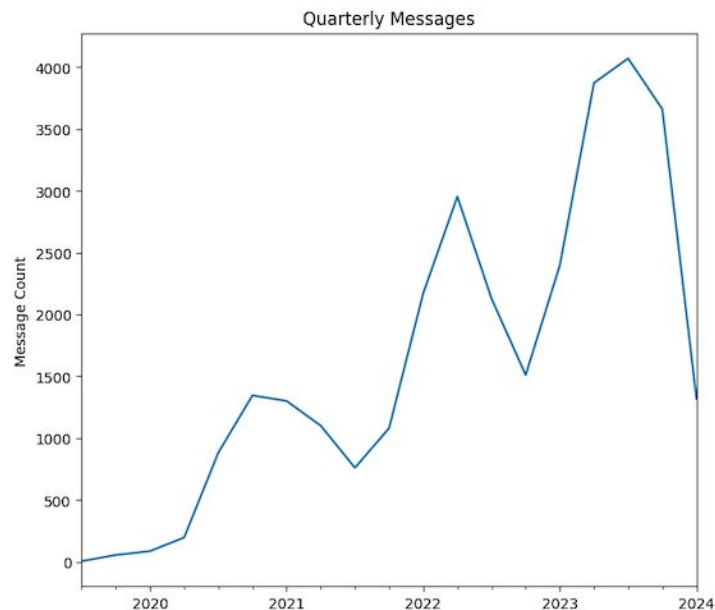
Quarterly Insights

Peak Engagement in Q2: The highest average message count in Q2 could be attributed to specific events, seasonal factors, or community initiatives that drive increased interaction during these months (**April, May, June**). This period might coincide with spring or early summer activities, which could naturally encourage more engagement due to seasonal variations in user availability or interest.

Lowest Engagement in Q1: Conversely, Quarter 1 (January, February, March) is identified as the period with the lowest average message volume, averaging 1454.4 messages. The reduced activity during this quarter may be influenced by post-holiday slowdowns, the start of the academic year in some regions, or other seasonal factors that affect member engagement.

Strategic Planning for High Activity Quarters: Knowing that Q2 has historically seen more messages, community managers and content creators can plan to align major events, releases, or engagement strategies with this period to maximize participation and interaction.

Understanding Seasonality: The fact that one quarter stands out in terms of message volume highlights the importance of understanding seasonal patterns and their impact on community engagement. This insight allows for more informed decision-making regarding content scheduling, community initiatives, and resource allocation to match these patterns.



General trends

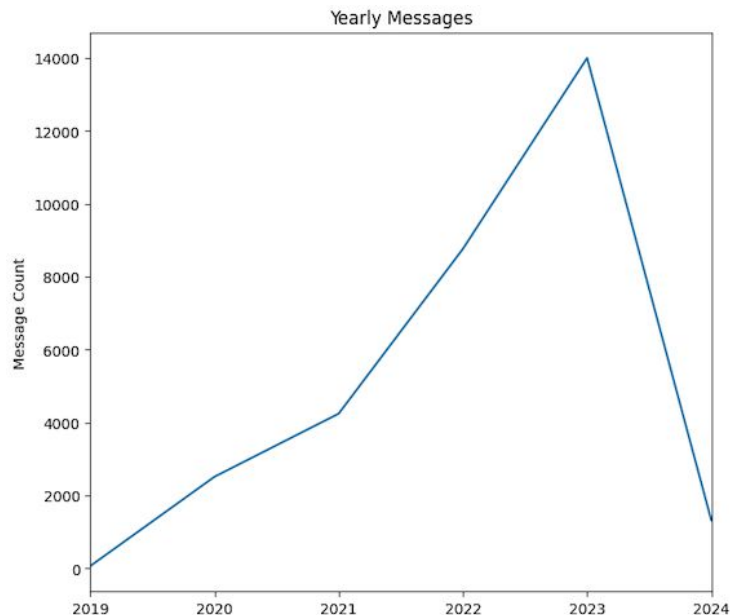
Target: Analyze the evolution of message numbers over time on a weekly, monthly, or quarterly basis. Determine if all channels follow the same trend and identify outlier periods with potential causes.

Yearly Insights

The analysis of yearly message volumes reveals that **2023** experienced the **highest level of activity**, with a total of 13,996 messages. This indicates a peak in engagement within the year, as visually represented in the bar chart.

Significant Growth in 2023: The spike in messages during 2023 suggests a notable increase in community activity and engagement.

The significant increase in community engagement observed in 2023, as highlighted by the yearly message volume analysis, could potentially be correlated with the **general trends in the cryptocurrency market**. Historically, cryptocurrency markets have experienced periods of high volatility and rapid growth, which often lead to increased public interest and activity within related communities. 2023 was a year of notable developments, major price movements, regulatory changes, and technological advancements in the crypto space, this could explain the heightened activity levels. The surge in messages and engagement within the community might reflect broader market enthusiasm, investor interest, and a desire for information and discussion about crypto-related topics.



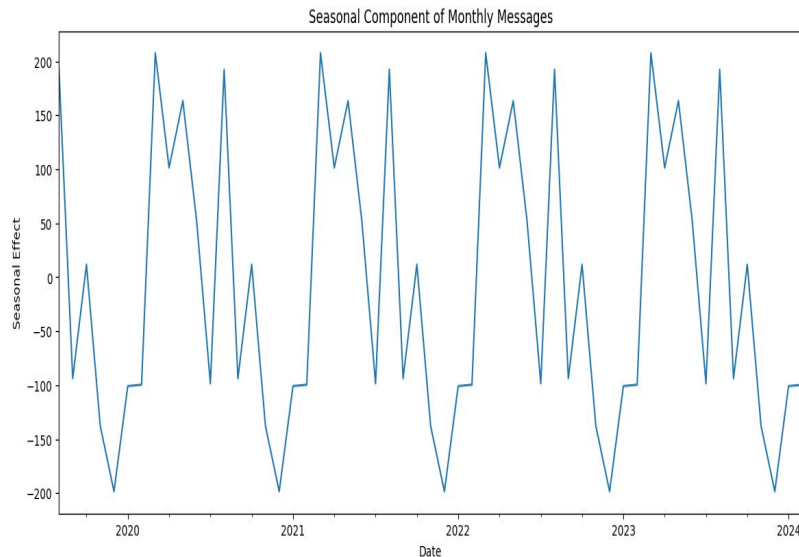
General trends

Target: Analyze the evolution of message numbers over time on a weekly, monthly, or quarterly basis. Determine if all channels follow the same trend and identify outlier periods with potential causes.

Seasonal Insights

Alignment with Market Cycles: The peaks in community engagement, especially in months like **May**, could coincide with periods of increased market activity or optimism within the cryptocurrency space. For instance, the crypto market generally experiences a surge in activity during the spring and early summer months due to increased investor interest or significant announcements, and this could drive more discussions and engagement within related communities.

Seasonal Downturns: Conversely, the periods identified as having lower engagement, such as **February**, might align with seasonal downturns or periods of consolidation in the cryptocurrency market. These times see reduced trading volume or a decrease in market-moving news, which could mirror the reduced activity within the community.



General trends

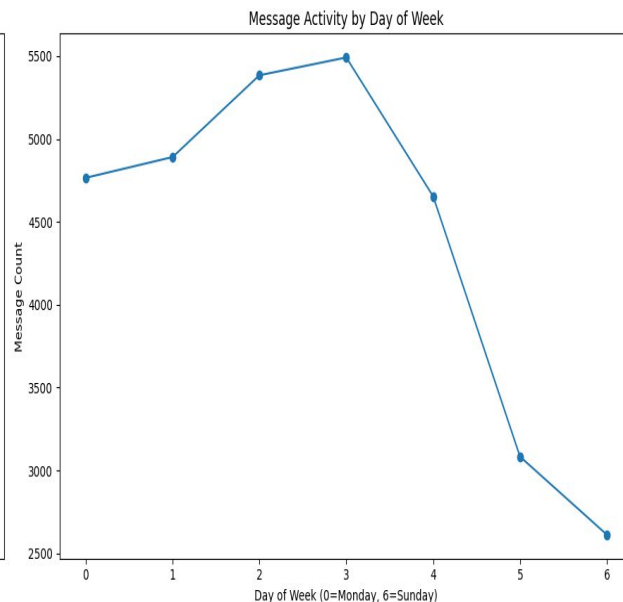
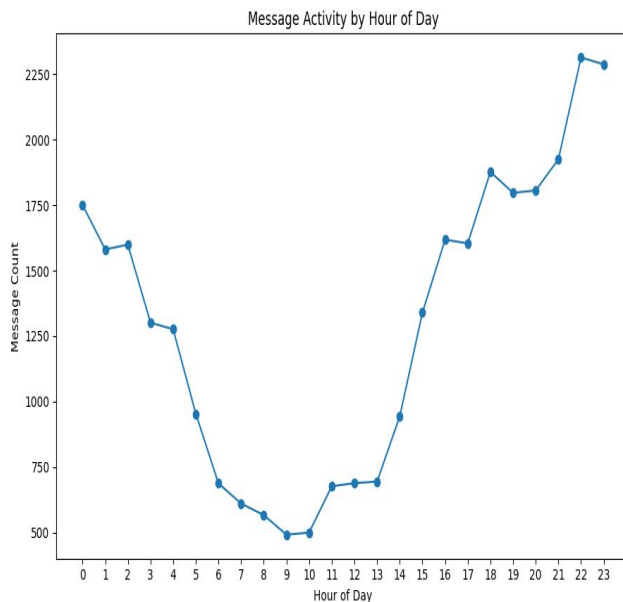
Target: Analyze the evolution of message numbers over time on a weekly, monthly, or quarterly basis. Determine if all channels follow the same trend and identify outlier periods with potential causes.

Peak Activity Insights

The analysis of message volumes by hour of the day and day of the week reveals that the community is most active at **22:00 (10 PM) and on Thursdays**.

Evening Engagement: The peak activity at 10 PM suggests that community members are most active during the evening hours. This could be due to users engaging with the community after work or school hours, indicating a preference for late-night discussions and interactions.

Midweek Activity: The fact that Thursday is the peak day for messages suggests a midweek surge in engagement. This could be due to the accumulation of topics and discussions throughout the week or specific weekly events or announcements that stimulate activity.



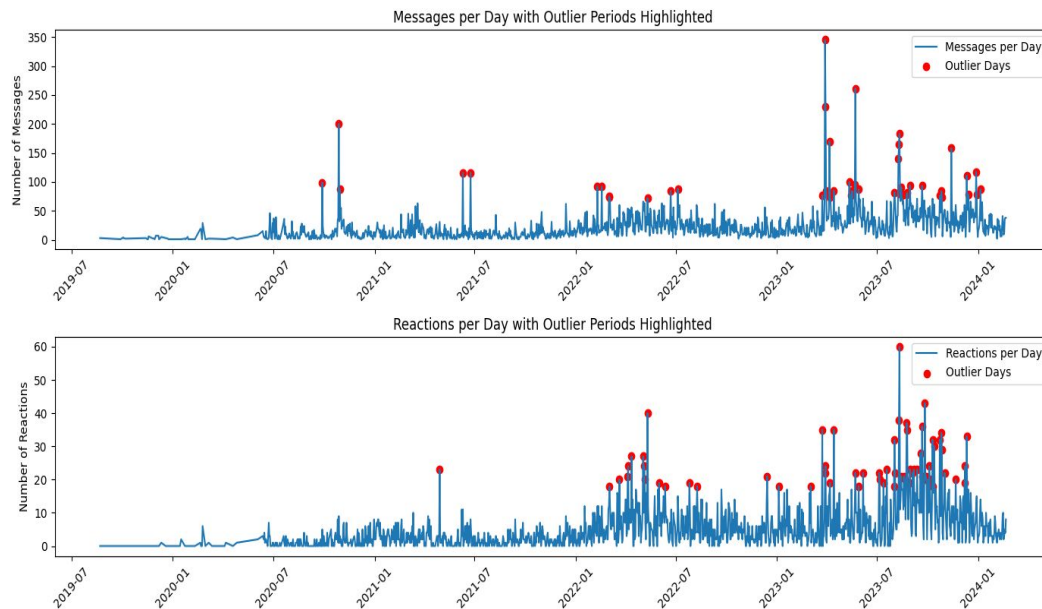
General trends

Target: Analyze the evolution of message numbers over time on a weekly, monthly, or quarterly basis. Determine if all channels follow the same trend and identify outlier periods with potential causes.

Outlier periods insights

Notable dates with unusually high numbers of messages:

1. April 17, 2021: 7 messages
2. May 4, 2021: 10 messages
3. July 5-6 and 8, 2022: 242 messages
4. August 6-7 and 11, 2022: 200 messages
5. March 29-30, 2023: 575 messages
6. May 23, 2023: 261 messages

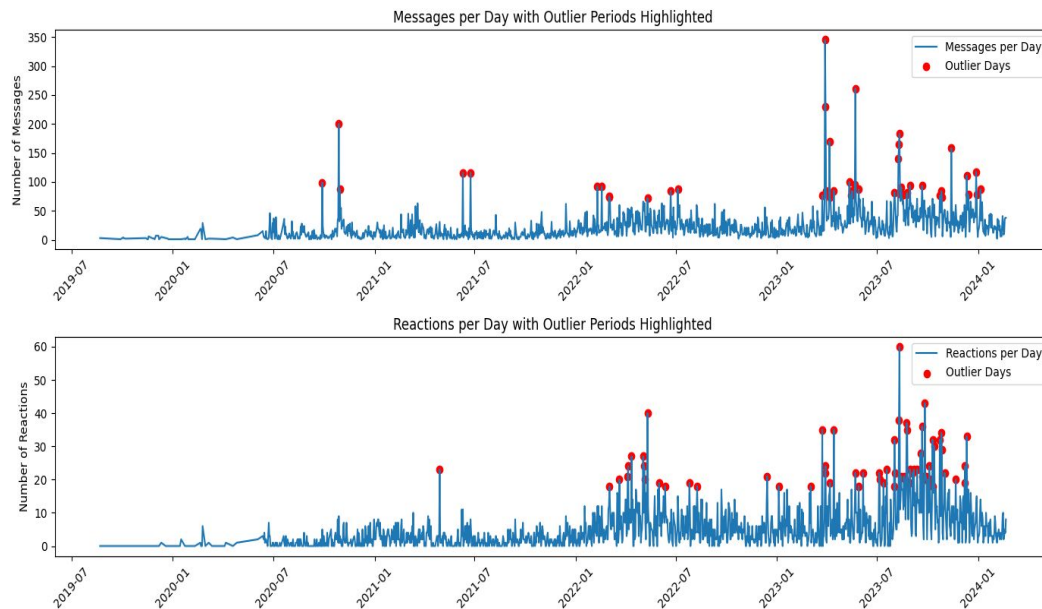


General trends

Target: Analyze the evolution of message numbers over time on a weekly, monthly, or quarterly basis. Determine if all channels follow the same trend and identify outlier periods with potential causes.

Outlier periods insights potential causes

1. **April - May 2021:** Frequent mentions of "warning", "scam", "account", "click", "links" and "validator" suggest discussions around security warnings and scam alerts. Also, a potential cause - This period is around new partnership announcements and the [Gaia-X](#) announcement.
2. **July - August 2022:** The terms "ocean", "data", "dataset" and "pool" point towards the Ocean Onda release and the draining pool [attack](#).
3. **March 2023:** Dominated by mentions of "oceandiffusion", "oceangpt", "ocean" and "protocol" reflecting engagement with the newly added tools.
4. **May 23 2023:** Common words include "ocean", "wallet", "discord" and "airdrop". These might be related to the Discord attack and also to the [minting of the ocean tokens](#).



General trends

Target: Analyze the evolution of message numbers over time on a weekly, monthly, or quarterly basis. Determine if all channels follow the same trend and identify outlier periods with potential causes.

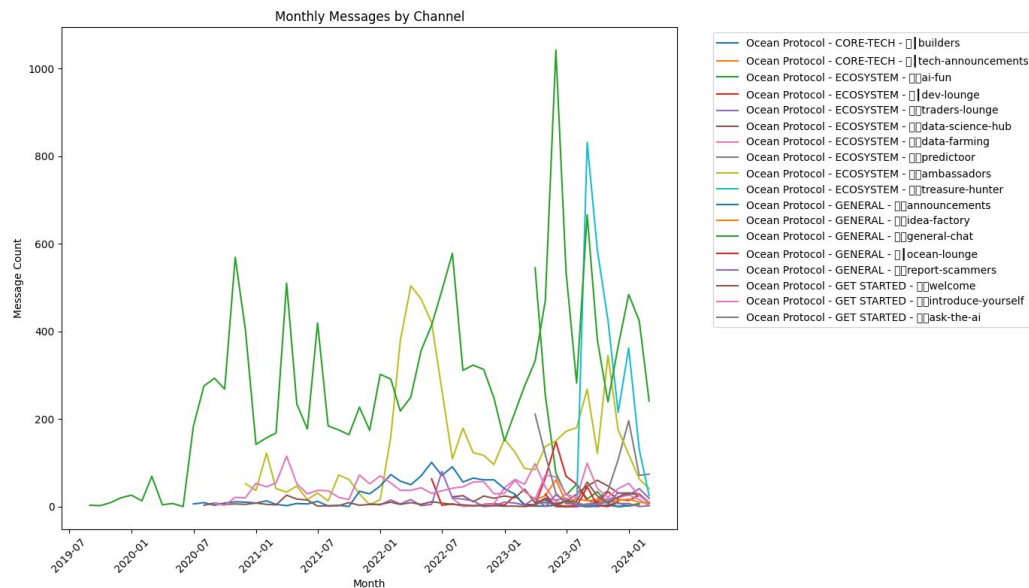
Channel trends

Channels with Most Messages (All-Time):

1. GENERAL - 🗨️ | general-chat: 15,074 messages
2. ECOSYSTEM - 🤖 | ambassadors: 5,622 messages
3. ECOSYSTEM - 🏴‍☠️ | treasure-hunter: 2,575 messages
4. GET STARTED - 🙋 | introduce-yourself: 1,788 messages
5. GENERAL - 🎉 | announcements: 1,056 messages

Channels with Least Messages (All-Time):

1. CORE-TECH - 🚀 | tech-announcements: 14 messages
2. CORE-TECH - 🏗️ | builders: 56 messages
3. ECOSYSTEM - 📊 | traders-lounge: 80 messages
4. GENERAL - 🧠 | ocean-lounge: 87 messages
5. GENERAL - 💡 | idea-factory: 214 messages



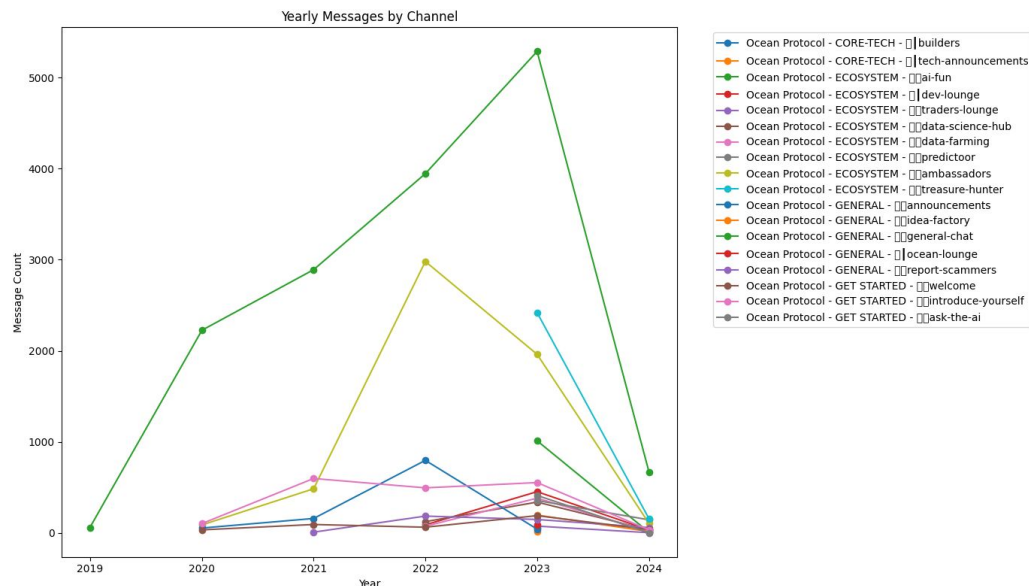
General trends

Target: Analyze the evolution of message numbers over time on a weekly, monthly, or quarterly basis. Determine if all channels follow the same trend and identify outlier periods with potential causes.

Channel trends

Most Active Channels in 2023: The "GENERAL - 🗨️ / general-chat" channel was the most active in 2023, with a significant increase to 5,287 messages. This suggests a very active general community discussion. The "ECOSYSTEM - 🏠 / treasure-hunter" and "ECOSYSTEM - 🏠 / ambassadors" channels followed, with 2,420 and 1,962 messages, respectively, indicating strong engagement in ecosystem-related discussions.

Engagement Metrics: The engagement levels in terms of average word and character count per message and reaction count varied widely across channels: The "ECOSYSTEM - 🏠 / treasure-hunter" channel exhibited **deep engagement** with an average of 42.79 words and 263.29 characters per message, along with a high average reaction count of 0.421 per message, suggesting intensive discussions or sharing of information. In contrast, the "GENERAL - 🗨️ / general-chat" had lower average word (14.53) and character (86.36) counts per message, but a very **high number of messages**, indicating a lot of quick, casual conversations.



General trends

Target: Analyze the evolution of message numbers over time on a weekly, monthly, or quarterly basis. Determine if all channels follow the same trend and identify outlier periods with potential causes.

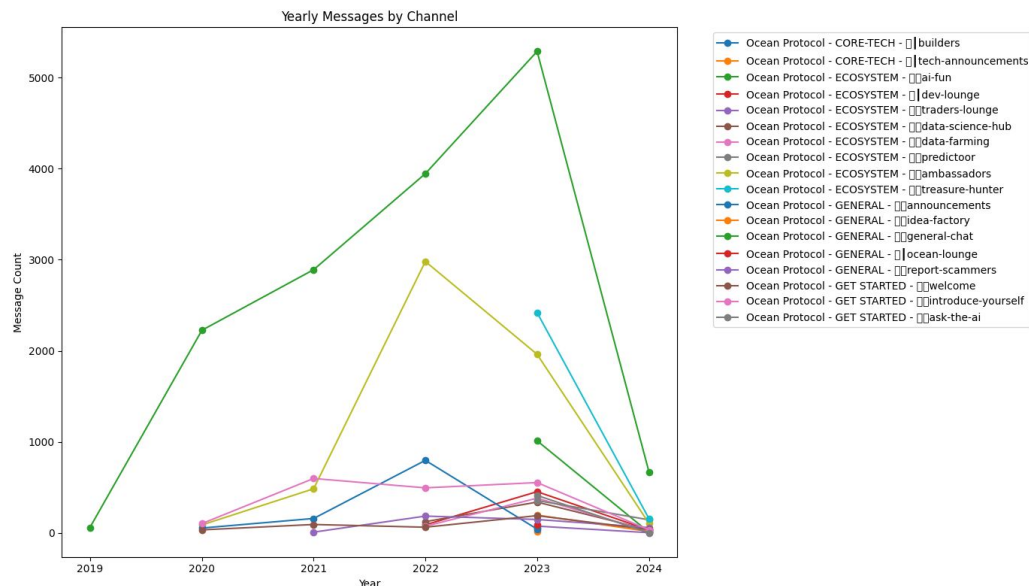
Channel trends

In 2023, the channels that saw the most significant growth in activity, based on the change in the number of messages from the beginning to the end of the year, include:

1. GET STARTED - 🗨️ | ask-the-ai: Infinite growth (from no activity to some activity).
2. ECOSYSTEM - 🏠 | treasure-hunter: 178.5% growth.
3. ECOSYSTEM - 🏠 | predictoor: 176.1% growth.
4. GENERAL - 💡 | idea-factory: 70.0% growth.
5. ECOSYSTEM - 📊 | data-science-hub: 19.6% growth.
6. GET STARTED - 🙋 | introduce-yourself: 11.96% growth.
7. GENERAL - 🗨️ | general-chat: 0.27% growth.
8. GET STARTED - 🙋 | welcome: No change.
9. ECOSYSTEM - 🚗 | data-farming: -10.26% decline.
10. GENERAL - 🗨️ | announcements: -14.04% decline.

On the other hand, channels that experienced the most significant drop in activity include:

1. ECOSYSTEM - 📊 | traders-lounge: -75.00% decline.
2. ECOSYSTEM - 🎮 | ai-fun: -71.43% decline.
3. ECOSYSTEM - 🖥️ | dev-lounge: -66.67% decline.
4. GENERAL - 📢 | report-scammers: -37.50% decline.
5. ECOSYSTEM - 🏠 | ambassadors: -30.34% decline.



General trends

Target: Analyze the evolution of message numbers over time on a weekly, monthly, or quarterly basis. Determine if all channels follow the same trend and identify outlier periods with potential causes.

Channel trends

While **some channels follow** the server's general trend of increasing discussions over time, such as "GENERAL - 🗨️ | general-chat", "ECOSYSTEM - 🏰 | data-farming," and "ECOSYSTEM - 🗨️ | predictoor," **it's evident that not all channels do**. For example, activity in the ambassadors channel dropped off after 2023, making way for the treasure-hunter channel.

Meanwhile, other channels like:

"CORE-TECH - 🚀 | tech-announcements",

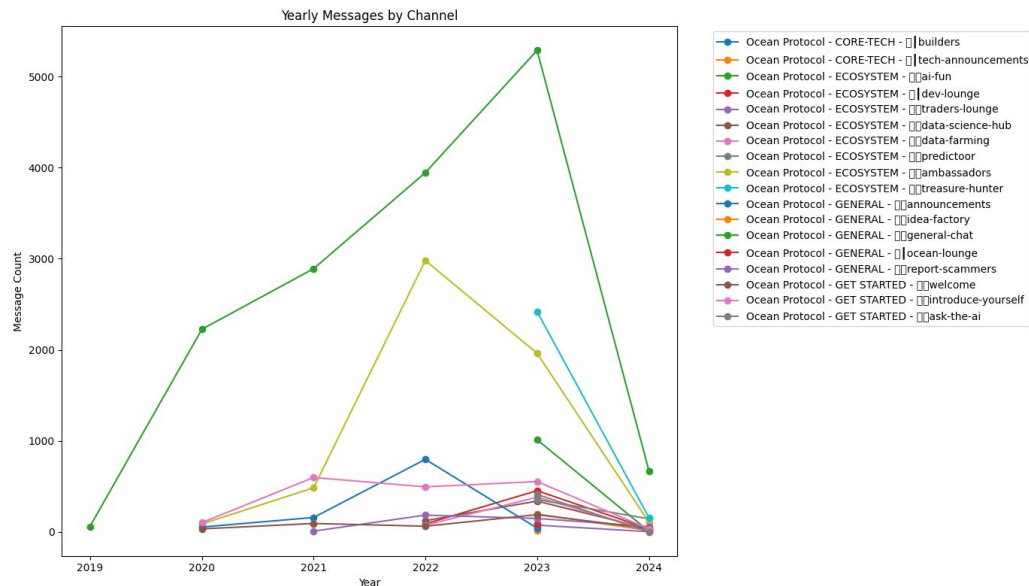
"CORE-TECH - 🌊 | builders",

"ECOSYSTEM - 📈 | traders-lounge",

"GENERAL - 🗨️ | ocean-lounge",

"GENERAL - 💡 | idea-factory",

have remained more or less consistent since they started with **not a lot of activity**.



Correlations

Target: Calculate the correlation between the price of \$OCEAN and the number of messages, new users, and active individuals on the server. Provide insights from these correlations and suggest statistical methods for analysis.

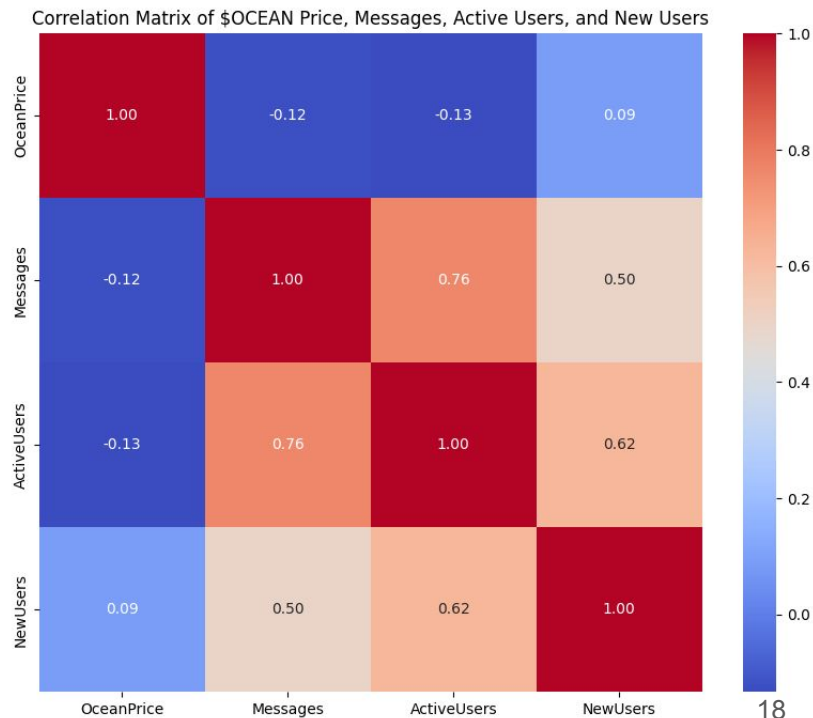
The Ocean token price is [downloaded](#) from yahoo finance and merged in the dataset.

\$OCEAN price and Messages: The correlation coefficient of **-0.121** suggests a very **weak** negative linear relationship between the price of **\$OCEAN** and the total number of messages. This implies that as the price of \$OCEAN changes, there's a slight tendency for the number of messages to move in the opposite direction, though the relationship is weak and might not be significant.

\$OCEAN price and Active Users: With a coefficient of **-0.134**, there's a similarly weak negative linear relationship between the price of **\$OCEAN** and the number of active users. This suggests a slight tendency for active user count to decrease as the price of \$OCEAN increases, or vice versa, but again, the relationship is very weak.

\$OCEAN price and New Users: The correlation coefficient of **0.089** indicates a very weak positive linear relationship between the price of \$OCEAN and the number of new users joining the server. This might suggest that higher \$OCEAN prices slightly correlate with more new users, but the relationship is very weak and may not be practically significant.

Messages, Active Users, and New Users: There are **stronger positive correlations** among the number of messages, active users, and new users. Notably, the correlation between Messages and Active Users is **0.764**, indicating a strong positive linear relationship. This suggests that days with more messages tend to also have more active users.



Correlations

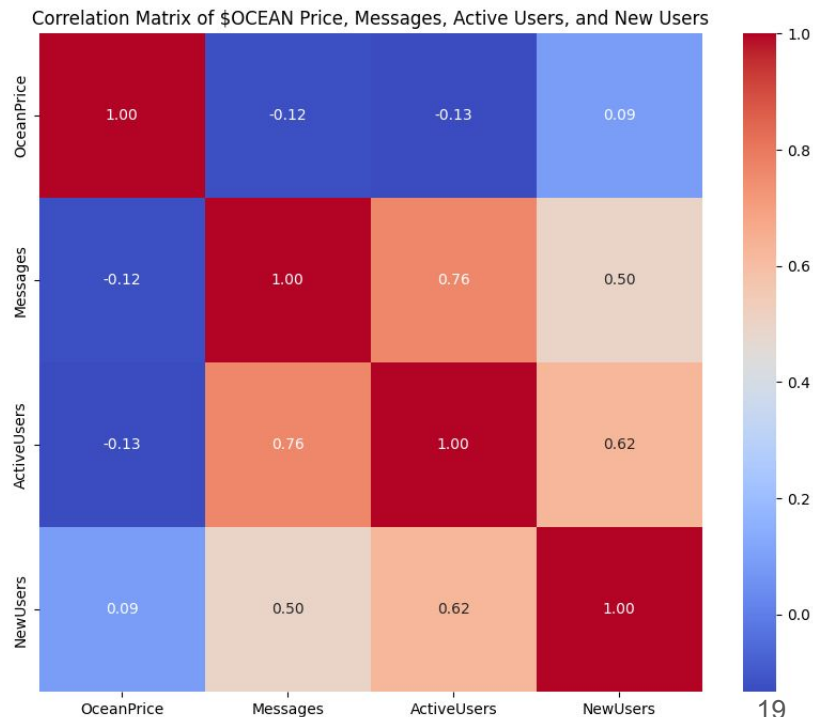
Target: Calculate the correlation between the price of \$OCEAN and the number of messages, new users, and active individuals on the server. Provide insights from these correlations and suggest statistical methods for analysis.

Insights:

The activity on the server (measured by the number of messages, active users, and new users) shows **some degree of interrelation**, especially between the number of messages and active users. This implies that more engagement (as seen through messages) is associated with higher numbers of active participants.

The **price** of \$OCEAN shows very **weak correlations** with **server activity** metrics, suggesting that daily price movements may not significantly impact daily server activity, at least not in a direct or linear manner.

These insights suggest that while day-to-day activity levels in terms of messaging and user engagement show some association with the price of \$OCEAN, **the relationships are not strongly positive or negative, indicating other factors might play a more significant role in price movements.**



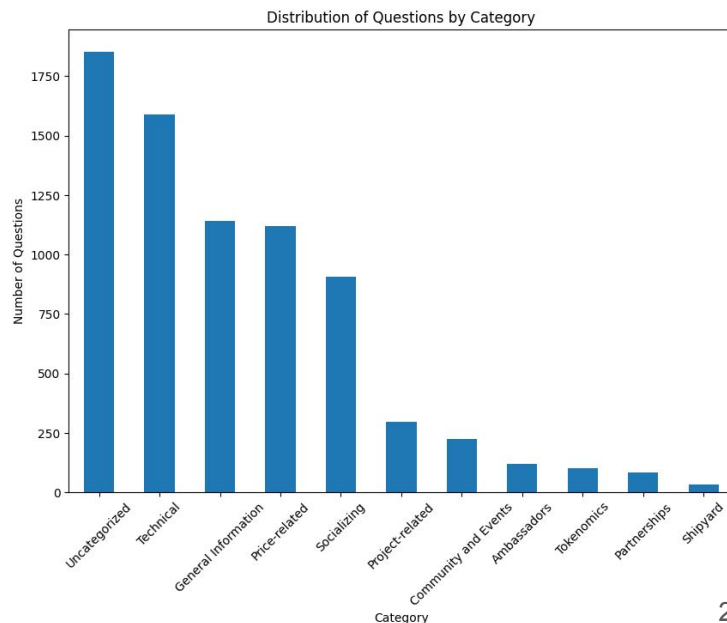
Community Questions

Target: Identify and categorize the most frequently asked questions into themes (technical, price-related, general information). Conclude on the subjects that generate the most questions.

Categories

1. **Technical:** Shows consistent engagement across years, highlighting a sustained interest in technical aspects of the project.
2. **General Information:** This category also show substantial activity, reflecting the community's need for information.
3. **Socialising:** This category also show substantial activity, reflecting the community's need for social interaction.
4. **Price-related:** Interest in price-related discussions appears stable, reflecting ongoing interest in the financial aspects of the project.
5. **Project-related, Community and Events, Ambassadors, and Tokenomics:** These categories show varying levels of engagement over the years, possibly correlating with specific project phases, events, or initiatives.
6. **Partnerships:** This category shows enough activity to require special attention on how to handle incoming proposals considering the spikes we see when new relevant partnerships are announced.
7. **AI Tools:** A new category in 2023, indicating a recent focus or development within the community related to AI technologies.

The majority of questions are primarily technical in nature, followed by inquiries about general information (such as token, project details, and releases), social interactions, and lastly, questions related to pricing.



Community Activity

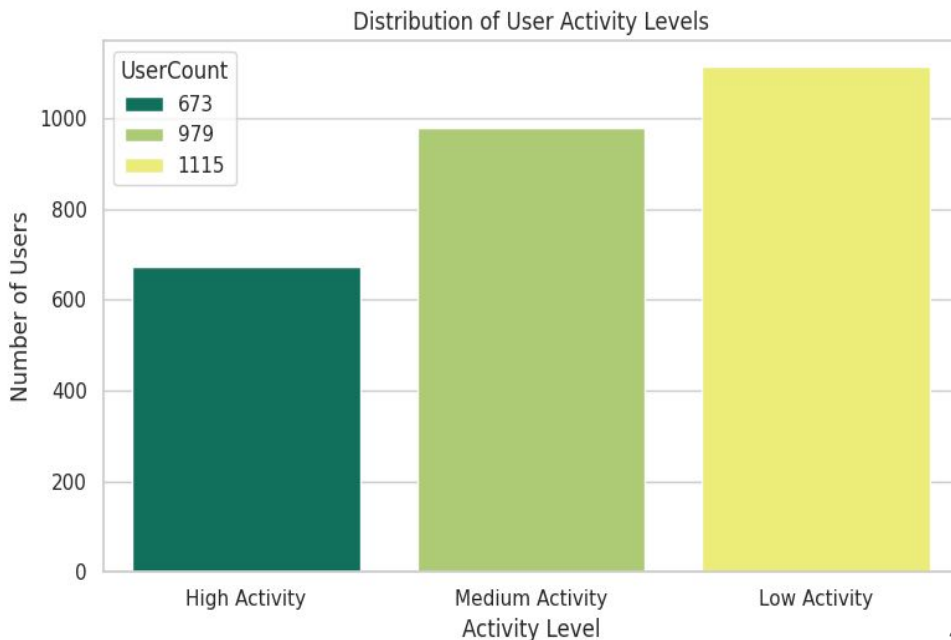
Target: Rank the most active non-bot users by various metrics (messages, words, characters, attachments sent, reactions received). Analyze the time of day/week for peak user activity and classify users into categories based on their activity.

The user [activity](#) has been aggregated by various metrics: total messages sent, total words, total characters, attachments sent, and reactions received for each user.

Activity thresholds:

- **High Activity:** Users in the top 25% of message counts.
- **Medium Activity:** Users between the top 25% and 75% of message counts.
- **Low Activity:** Users in the bottom 25% of message counts.

The distribution is relevant as it shows that more than 50% of the users are somehow active and approx 20% are highly active.



Community Activity

Target: Rank the most active non-bot users by various metrics (messages, words, characters, attachments sent, reactions received). Analyze the time of day/week for peak user activity and classify users into categories based on their activity.

All time user ranking(top 15)

Messages Sent: The user with the username **blockchainlugano** 🕶️ sent the most messages, followed by **kreigdk**, **dotunwilfred.eth**, **bhavingala**, and **zippy1979**.

Words Sent: Indicates the total word count of messages sent by each user. **blockchainlugano** 🕶️ also leads in this metric, suggesting not only a high number of messages but also substantial message lengths.

Characters Sent: Reflects the total character count, with **blockchainlugano** 🕶️ again showing significant activity.

Attachments Sent: **dotunwilfred.eth** 🕶️ has sent the most attachments, which is distinct from the other metrics and indicates a different form of engagement.

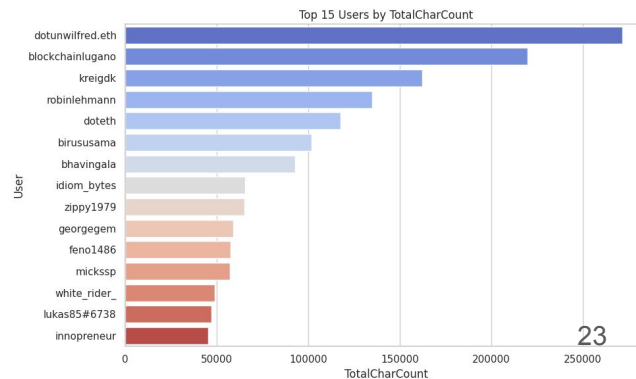
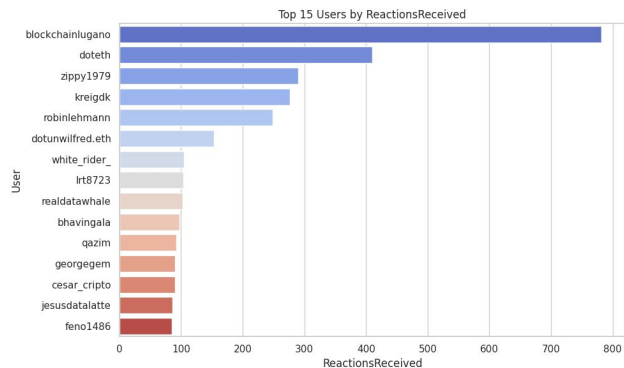
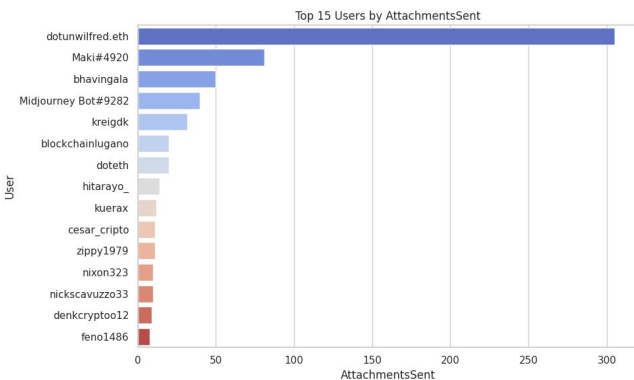
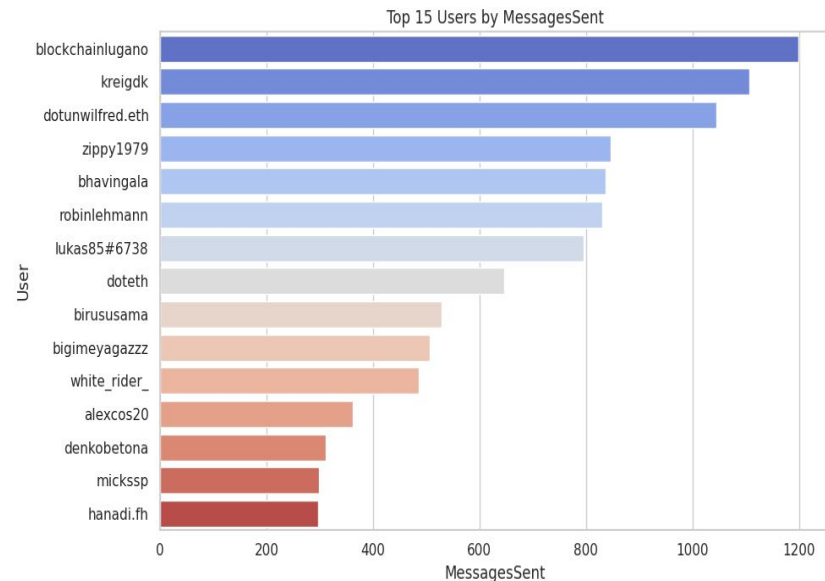
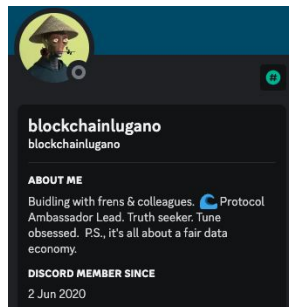
Reactions Received: **blockchainlugano** 🕶️ received the most reactions, reinforcing their high level of engagement within the community.

	MessagesSent	WordsSent	CharactersSent	AttachmentsSent	ReactionsReceived
AuthorName					
blockchainlugano	1199	37092	219677	20	782
kreigdk	1106	24580	162288	32	277
dotunwilfred.eth	1044	36381	271737	305	153
zippy1979	847	11151	65177	11	290
bhavingala	837	11037	92881	50	97
robinlehmann	830	22436	134804	8	249
lukas85#6738	795	7680	47038	5	31
doteth	647	18194	117544	20	410
birususama	529	15086	101738	4	50
biglmeiyagazzz	507	2095	10628	1	30
white_rider_	486	7919	49103	5	105
alexcos20	362	4567	26515	3	32
denkobetona	311	2133	13401	5	31
mickssp	298	8097	57315	4	47
hanadi.fh	297	4468	32268	0	49

Community Activity

Target: Rank the most active non-bot users by various metrics (messages, words, characters, attachments sent, reactions received). Analyze the time of day/week for peak user activity and classify users into categories based on their activity.

Most active user throughout Ocean discord history is: blockchainlugano 🏆



Community Activity

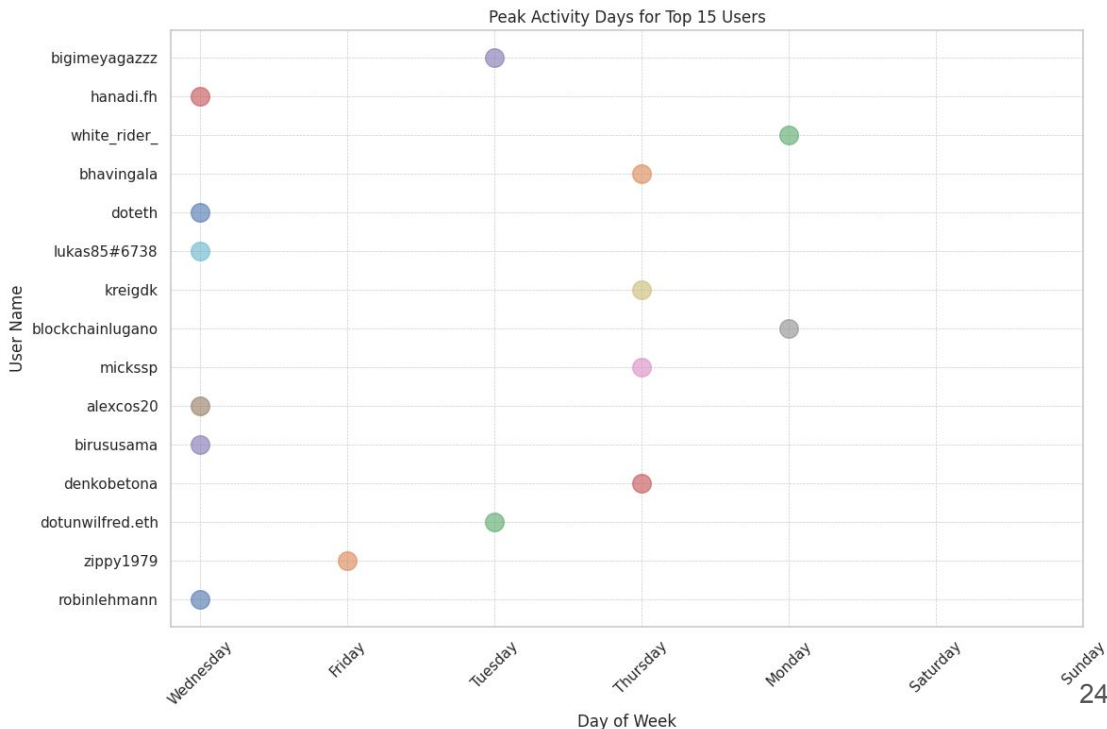
Target: Rank the most active non-bot users by various metrics (messages, words, characters, attachments sent, reactions received). Analyze the time of day/week for peak user activity and classify users into categories based on their activity.

Peak activity days

The day-of-week analysis for these users shows no singular trend, with peak activity days scattered across the week. Mondays through Sundays all serve as peak days for different users, reflecting the absence of a universal pattern in day-wise engagement.

This variability underscores the **dynamic** nature of the community, where each day brings its unique rhythm of activity.

Such findings imply that the community's vibrancy is not confined to traditional **workdays or weekends**, making it a robust platform for continuous interaction and engagement.



Community Activity

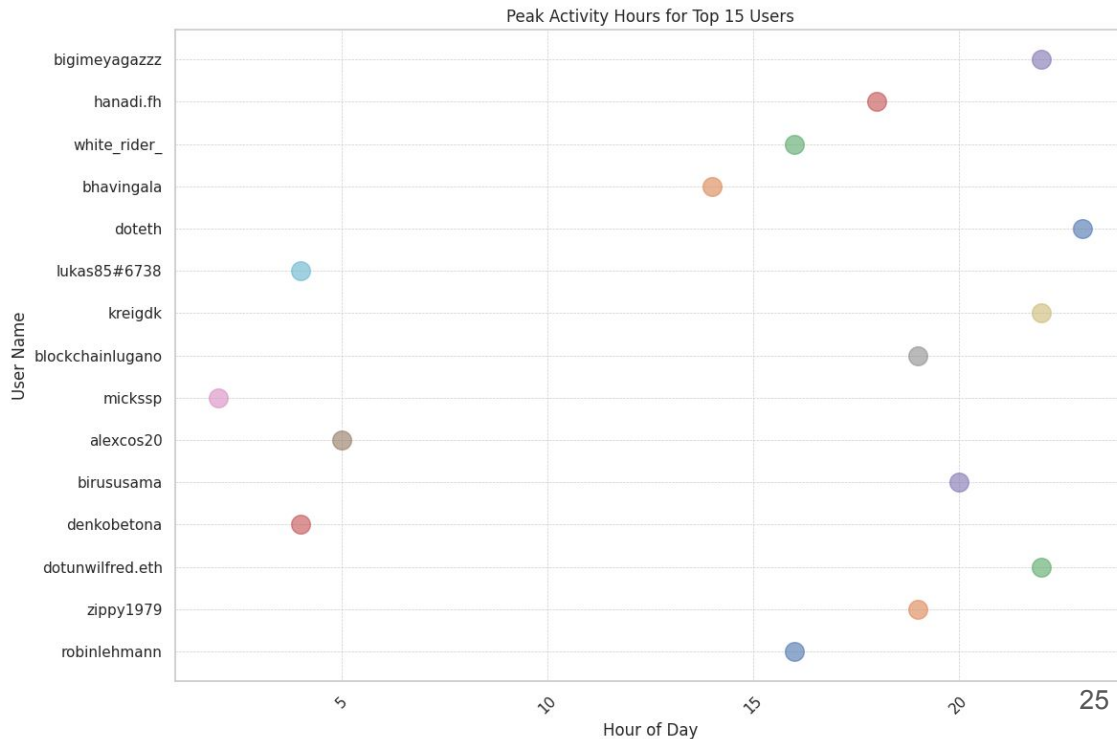
Target: Rank the most active non-bot users by various metrics (messages, words, characters, attachments sent, reactions received). Analyze the time of day/week for peak user activity and classify users into categories based on their activity.

Peak activity hours

The analysis of peak activity times for the top 15 most active users reveals a broad spectrum of engagement hours, highlighting the diverse nature of the community's participation.

The range of peak activity hours extends from the early morning, with some users most active around 2 AM and 4 AM, to late at night, with peak activity recorded at 19 PM, 22 PM, and extending to 23 PM.

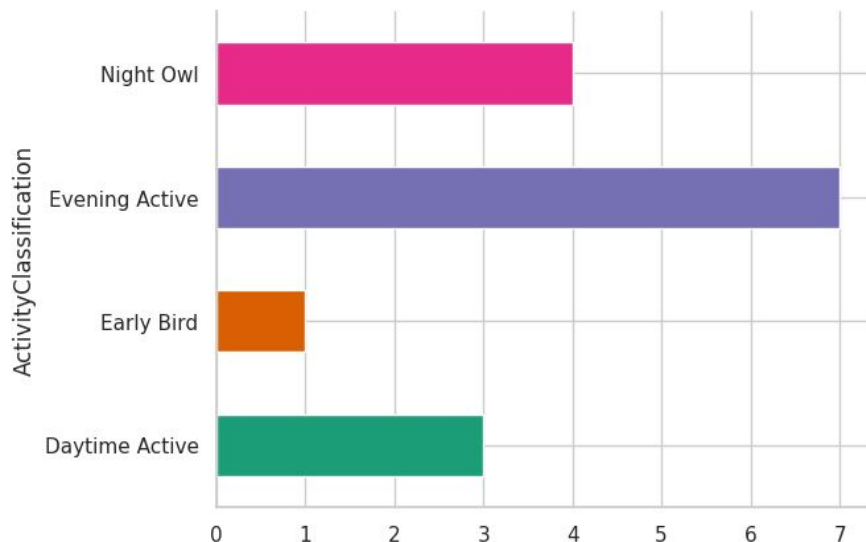
This wide distribution suggests that the community **accommodates a global audience**, with users participating from various time zones. The diversity in active hours underlines the non-uniformity in personal schedules and preferences for engagement, indicating that the community is alive and interactive across all hours, providing a continuous stream of communication and interaction.



Community Activity

Target: Rank the most active non-bot users by various metrics (messages, words, characters, attachments sent, reactions received). Analyze the time of day/week for peak user activity and classify users into categories based on their activity.

Peak activity classification



	AuthorID	AuthorName	ActivityClassification
0	194817764236460034	robinlehmann	Daytime Active
1	209963946432528384	zippy1979	Evening Active
2	344879785173843970	dotunwilfred.eth	Evening Active
3	368405653217345536	denkobetona	Night Owl
4	387401160656683034	birususama	Evening Active
5	625415196713943051	alexcos20	Early Bird
6	647661378198306821	mickssp	Night Owl
7	717363377269244015	blockchainlugano	Evening Active
8	739132787499597824	kreigdk	Evening Active
9	768163222095134732	lukas85#6738	Night Owl
10	804604311015129089	doteth	Night Owl
11	837901335470538772	bhavingala	Daytime Active
12	843831770062913568	white_rider_	Daytime Active
13	911229119235227718	hanadi.fh	Evening Active
14	985478606283223041	bigimeyagazzz	Evening Active

Community Activity

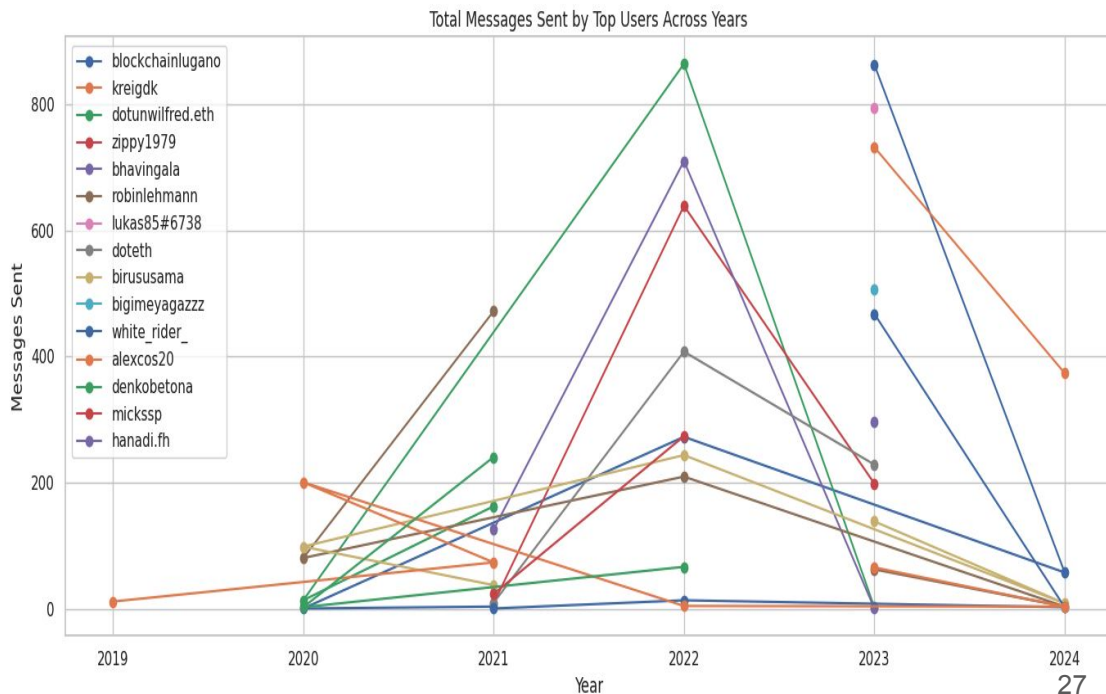
Target: Rank the most active non-bot users by various metrics (messages, words, characters, attachments sent, reactions received). Analyze the time of day/week for peak user activity and classify users into categories based on their activity.

Most active users - evolution

This perspective allows us to observe the engagement patterns of the most active users throughout the entire history of the Discord server. Notably, it's apparent that user engagement fluctuates over the years. For instance, **robinlehmman's** activity decreased after 2021, while **kreigdk** began participating actively in 2023 and has maintained a strong momentum since.

Such fluctuations in user activity are typical within communities, with members joining, participating, and sometimes leaving. This trend is clearly illustrated in our analysis.

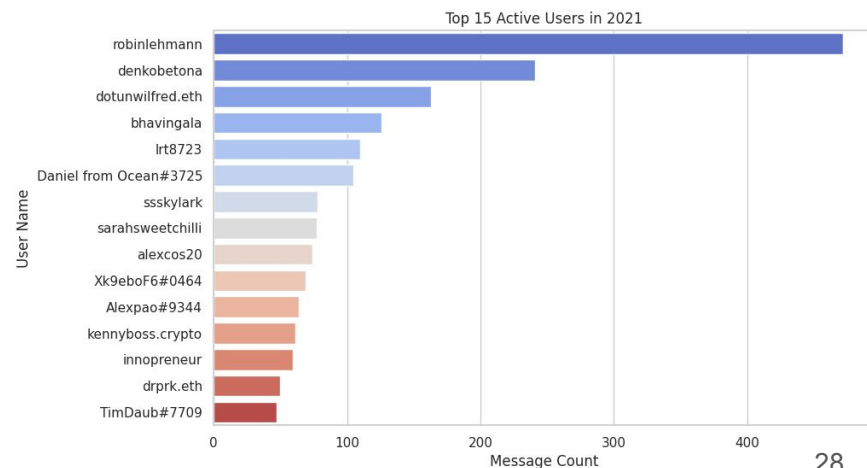
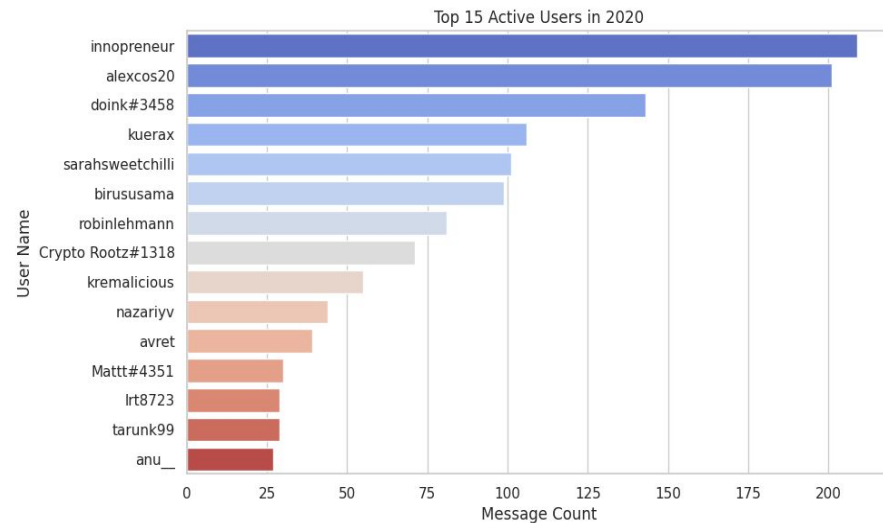
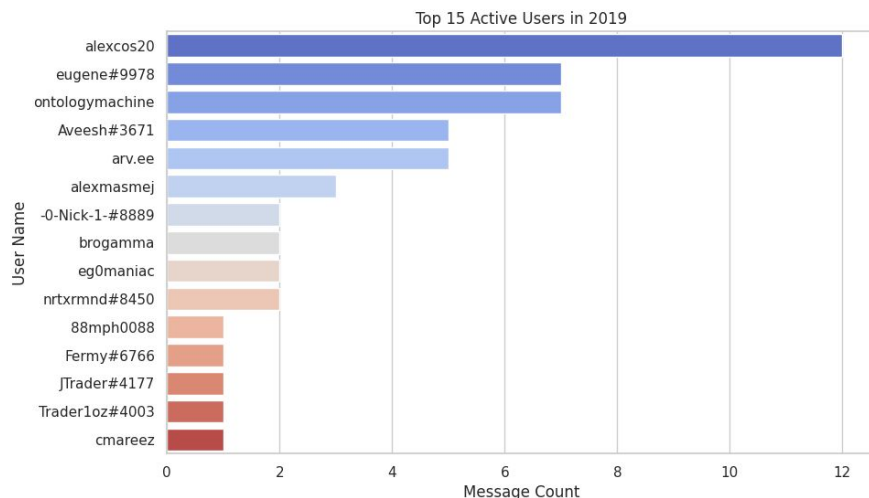
Aside from **blockchainlugano** and **alexcos20**, who have demonstrated steady engagement, the pattern of users coming and going is evident.



Community Activity

Target: Rank the most active non-bot users by various metrics (messages, words, characters, attachments sent, reactions received). Analyze the time of day/week for peak user activity and classify users into categories based on their activity.

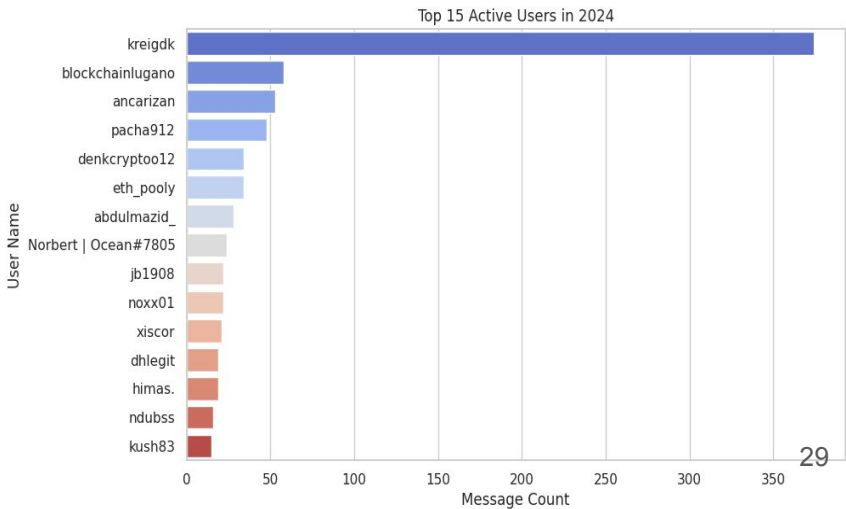
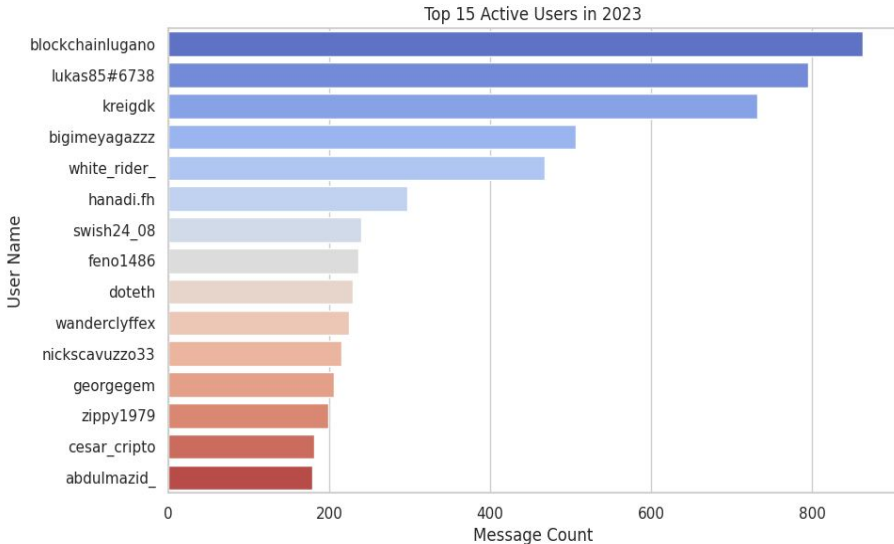
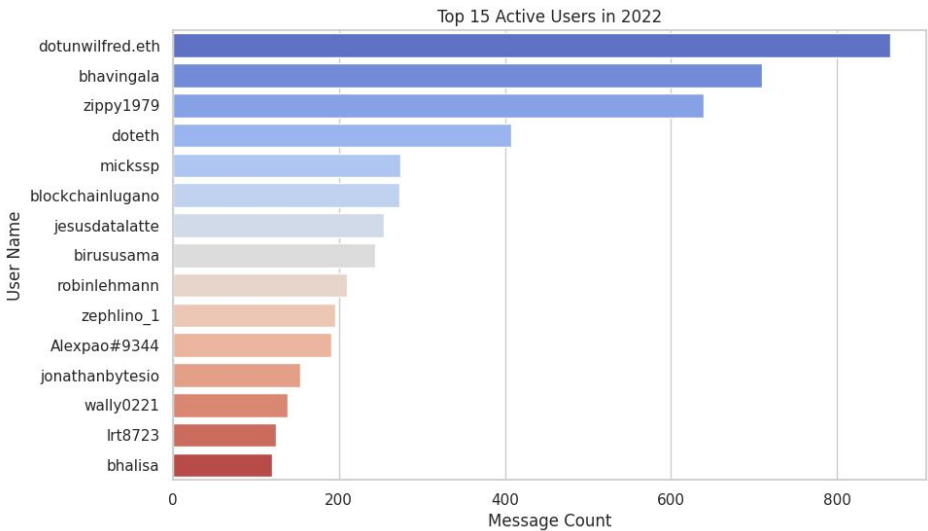
Most active users per year



Community Activity

Target: Rank the most active non-bot users by various metrics (messages, words, characters, attachments sent, reactions received). Analyze the time of day/week for peak user activity and classify users into categories based on their activity.

Most active users per year

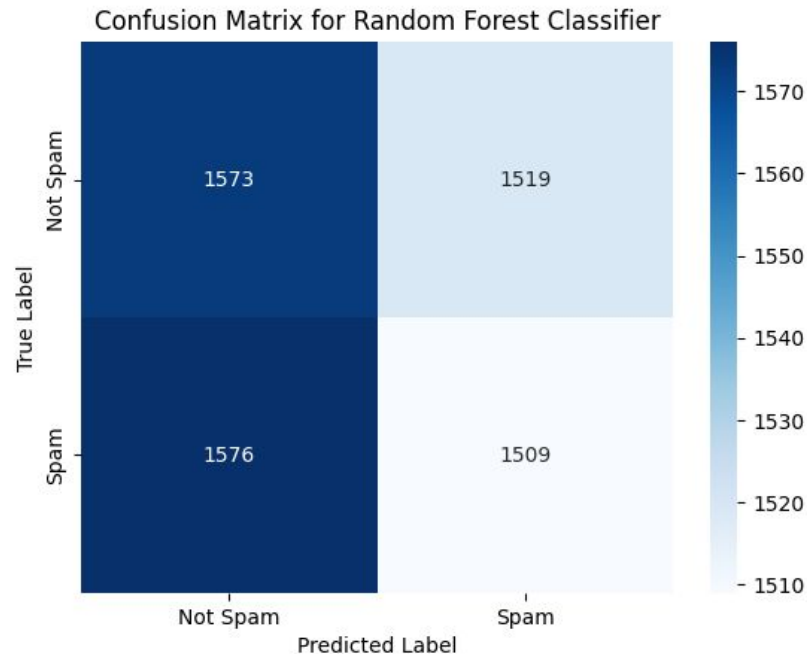


Scam & Spam

Target: Use an existing machine learning model to identify likely spam or scam messages, specifying chosen features for identification. Justify the model choice and describe the main characteristics of spam/scam messages.

Given the context of the dataset and the inherent challenges in detecting SCAM messages without explicit labels within the dataset, the task of accurately identifying SCAM content is notably complex. Initially, the application of **BERT (Bidirectional Encoder Representations from Transformers)** was considered for its advanced capability to understand the context and nuances of text data. Subsequently, a [Random Forest classifier](#) was trained as an alternative approach. However, the outcomes of the methods applied are not reliable indicators for SCAM message detection, **primarily because the dataset lacks actual SCAM messages**. Typically, upon the identification of SCAM messages or scammers, community managers or team members promptly intervene by banning the individuals involved and deleting their messages, **resulting in a dataset almost void of explicit SCAM content**.

Even though using a machine learning model to directly identify SCAM messages in the dataset didn't really work out—the outcomes were basically a toss-up, with each message having a **50/50** chance of being tagged as a scam—we can still take some active steps and come up with strategies to better filter out SCAM and SPAM messages on the server. You'll find a few suggestions in the following pages.



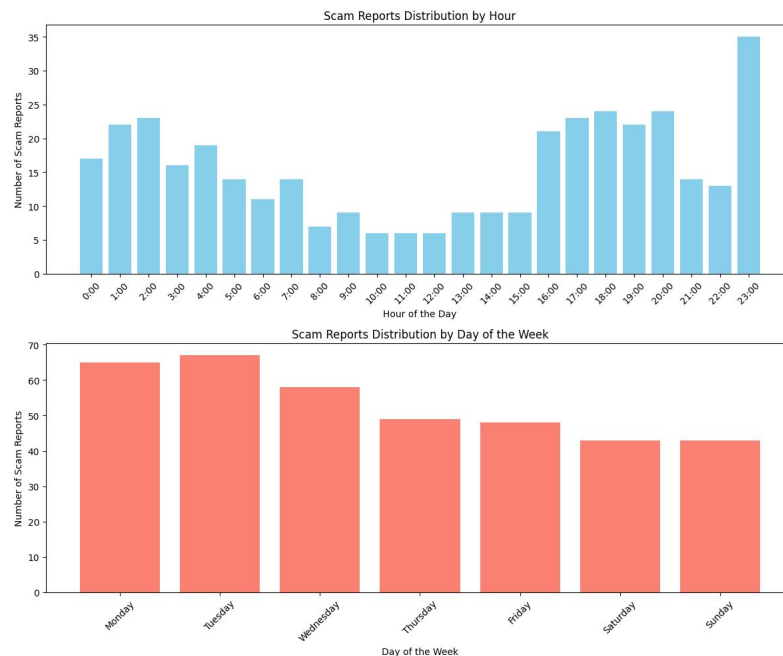
Scam & Spam

Target: Use an existing machine learning model to identify likely spam or scam messages, specifying chosen features for identification. Justify the model choice and describe the main characteristics of spam/scam messages.

Scam reports

Although the dataset doesn't directly allow us to pinpoint scam messages, it enables us to derive insights from reports submitted by users about potential scam activities.

- **Hourly Distribution:** The scam reports appear to be somewhat evenly distributed throughout the day, indicating that scam attempts don't have a strict timing pattern. However, there is a noticeable peak in scam reports around **23:00**. This spike might suggest that scammers prefer to target users during late-night hours, possibly betting on lower vigilance among community members.
- **Day of the Week Distribution:** Scam reports are observed to be spread across all days of the week, lacking a definitive pattern that points to a particular day. However, there's a **slight inclination towards the beginning of the week**, suggesting that scammers may choose to initiate their activities as a new week commences.

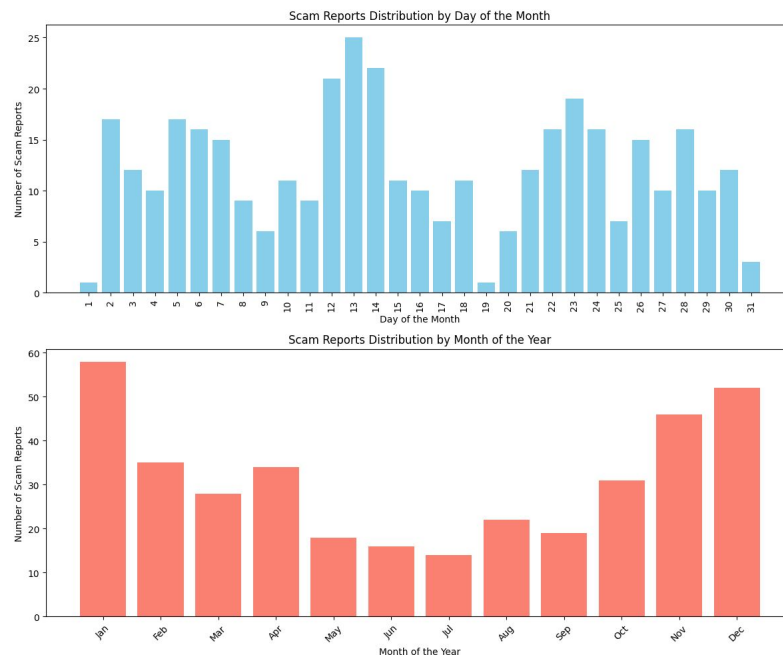


Scam & Spam

Target: Use an existing machine learning model to identify likely spam or scam messages, specifying chosen features for identification. Justify the model choice and describe the main characteristics of spam/scam messages.

Scam reports

- **Day of the Month Distribution:** The scam reports appear to be somewhat evenly distributed throughout the month, indicating that scam attempts don't adhere to a strict timing pattern. However, there is a **noticeable peak in scam reports around the middle (12th-13th) and towards the end (21st-23rd) of the month**. This could suggest that scammers might target specific times towards the middle & end of the month, possibly to exploit patterns related to financial cycles or pay periods when individuals might be more engaged in online transactions.
- **Month of the Year Distribution:** Scam reports are observed to be spread across the entire year, without significant peaks in most months, suggesting a consistent threat from scam activities year-round. However, there's a **slight inclination towards the beginning and the end of the year**. This trend may reflect scammers taking advantage of periods where there might be increased online activity, such as during new year resolutions or holiday shopping seasons, when users might be more susceptible to fraudulent schemes.



Scam & Spam

Target: Use an existing machine learning model to identify likely spam or scam messages, specifying chosen features for identification. Justify the model choice and describe the main characteristics of spam/scam messages.

Filter SCAM messages

Improving SCAM & SPAM message filtering can greatly enhance user experience and uphold community integrity. Here are some tips:

1. Keyword-Based Filters

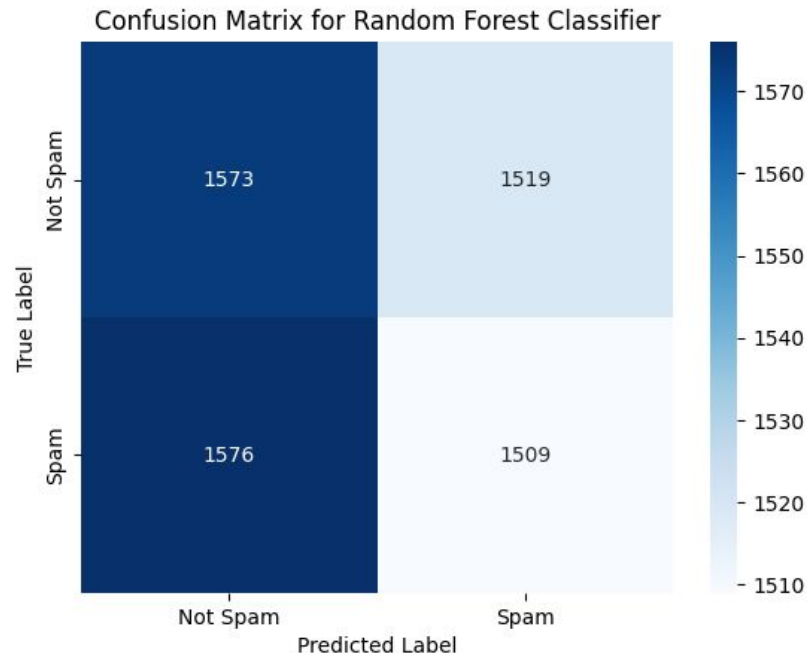
Implement filters that automatically flag or block messages/users containing specific high-risk keywords commonly used like:

- Phrases like **"Support ticket"**, **"airdrop"**, **"airdrops"**, **"NFT drops"**, **"giveaway"**, **"altcoins"**, **"helpdesk"**, **"support team"**, **"chosen few"**, **"admin-tech"** which are often used to lure users into participating in fraudulent schemes.
- Explicit calls to action, such as **"Claim now"**, **"Free access"**, **"Urgent"** or **"Urgent response required"**, **"You are chosen"** that create a false sense of urgency or offer too-good-to-be-true rewards.
- Usernames like: **"Ticket"**, **"Support"**, **"Ticket tool"**, **"Ticket-Support"**, **"Support-team"**, **"Helpdesk"** can indicate a potential malicious intent.

2. Link Analysis

Automatically screen messages for suspicious links, especially to phishing sites or those demanding immediate action. Focus on:

- **Links containing discord invites** (e.g., <https://discord.com/invite/...>, <https://discord.com/invite/...>) which can be used to divert users to scam servers.
- **Shortened URLs** that obfuscate the destination, often used to hide malicious sites.



Scam & Spam

Target: Use an existing machine learning model to identify likely spam or scam messages, specifying chosen features for identification. Justify the model choice and describe the main characteristics of spam/scam messages.

3. Restrictions Based on User Status

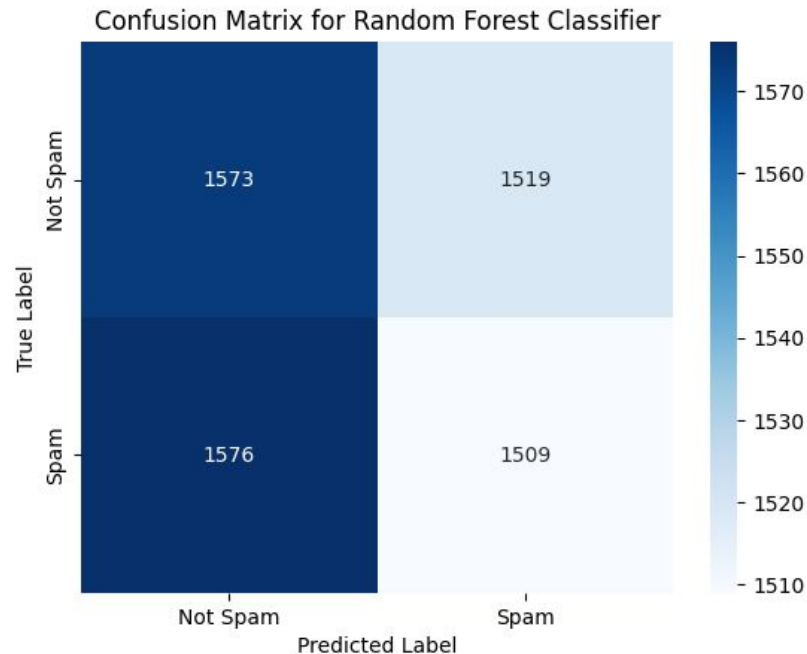
Limit the capabilities of new or low-ranking users to send certain types of messages, reducing the ability of scammers to exploit your platform:

- Prevent users with **no ranking** or a new account status from **sending messages containing any links**. This includes hyperlinks, file attachments, or embedded media, until they reach a certain trust level or account age.
- Implement a system where users must **earn** the privilege to **share links** or specific content types by contributing positively to the community or undergoing a verification process.

4. Authentication and Verification

Strengthen account security and authenticity through:

- Mandatory **phone number** verification during account creation or before enabling certain privileges. This adds a layer of difficulty for scammers looking to create multiple fake accounts.



Technical Issues

Target: Identify and categorize the most common technical issues, indicating their potential sources (user-related, system-related, external factors).

Activities:

- **Filter** the dataset with the time range: 01.01.2023 - Last day
- **Preprocess** the Content column to clean and prepare the text data.
- Use a **Natural Language Processing (NLP)** technique to extract features from the textual data.
- Apply a **clustering algorithm (K-means)** to group similar issues together.
- **Analyze** the clusters to categorize them and attempt to identify their potential sources (user-related, system-related, external factors).

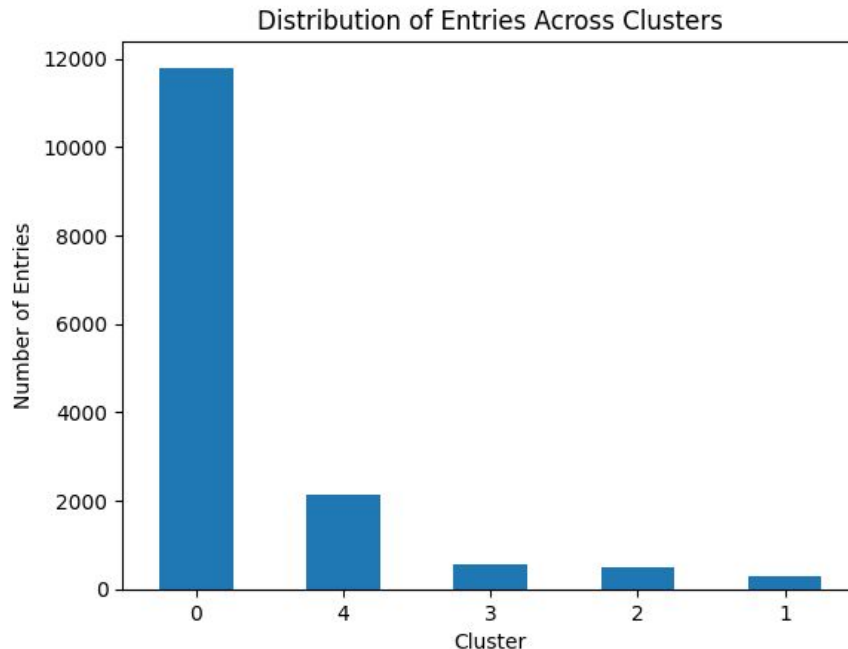
Extracted:

Cluster 0: Topics related to data, ambassador activities, and AI, indicating discussions might be around data use, community engagement, and AI technologies.

Cluster 4: Dominated by references to Ocean Protocol, data, and protocol use, suggesting technical discussions about Ocean Protocol's functionalities and data handling.

Cluster 3: Unique terms like "oceandiffusion", "image", "logo", which may indicate specific discussions about branding, imaging, or perhaps a project within the Ocean ecosystem.

Cluster 1 & 2: Both clusters contain more general terms like "hi", "hello", "guys", indicating social interactions or introductory messages, with Cluster 2 also mentioning "help", "welcome", possibly relating to newcomer assistance or community support.

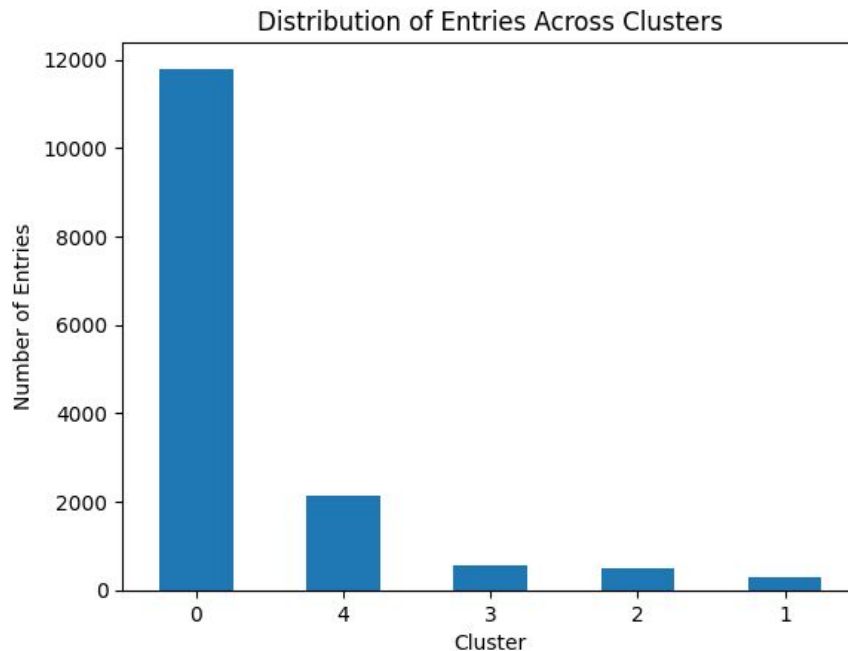


Technical Issues

Target: Identify and categorize the most common technical issues, indicating their potential sources (user-related, system-related, external factors).

Cluster 4 messages suggest a focus on Ocean Protocol's functionalities and applications, with discussions around:

- **Integrating Ocean Protocol with other technologies** ("how would using Ocean Protocol work with...").
- **Technical inquiries and project interest** ("liked the Ocean project but I didn't understand...").
- **Discussions about Ocean Protocol's features and benefits** ("allows people to sell and buy data").
- **Troubleshooting and error handling** ("invalid-token" error)

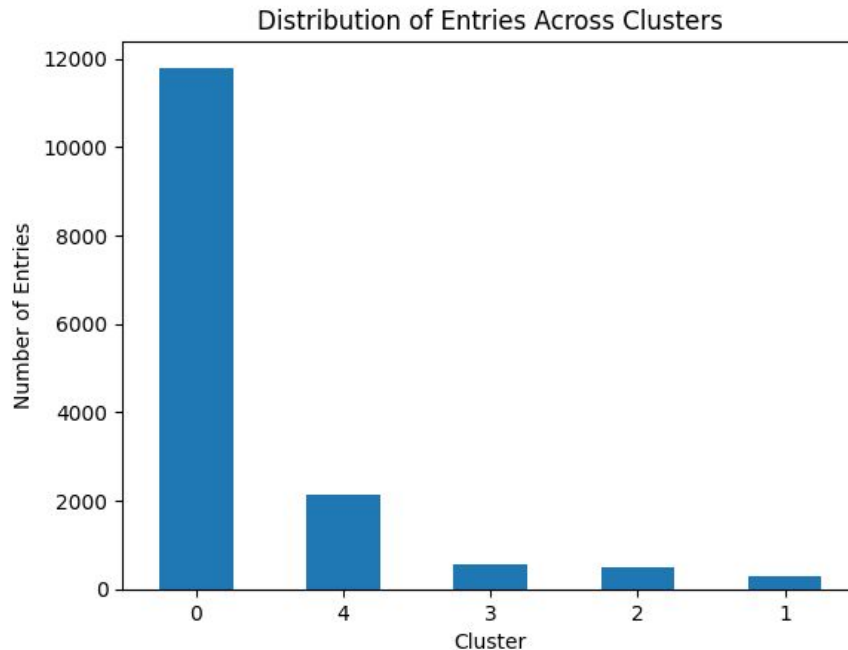


Technical Issues

Target: Identify and categorize the most common technical issues, indicating their potential sources (user-related, system-related, external factors).

Extracted key questions/categories

- 1. How to Use Ocean Protocol for Data Sharing:** "Can you provide resources or guides on how to utilize Ocean Protocol for sharing and monetizing data?"
- 2. Data Privacy and Integration:** "How does Ocean Protocol ensure data privacy when integrating various data sources in the data market?"
- 3. Improving Data Liquidity:** "What strategies does Ocean Protocol employ to enhance data liquidity and facilitate easier access to data?"
- 4. Monetizing Personal Data with GDPR Compliance:** "How can personal data be monetized using Ocean Protocol while ensuring compliance with GDPR and other privacy regulations?"
- 5. Integration Challenges:** "What are the common challenges faced when integrating Ocean Protocol with existing data management systems, and how can they be addressed?"
- 6. Troubleshooting 'Invalid Token' Errors:** "Why might someone encounter an 'invalid token' error when accessing the data market, and how can this issue be resolved?"
- 7. Development and Smart Contracts:** "Where can developers find comprehensive documentation on using Ocean Protocol's smart contracts for dApp development?"
- 8. Differences from Traditional Data Marketplaces:** "How does Ocean Protocol differentiate itself from traditional data marketplaces, particularly in terms of integrating with AI models and enhancing user privacy?"



Technical Issues

Target: Identify and categorize the most common technical issues, indicating their potential sources (user-related, system-related, external factors).

Questions categories:

1. User-related:

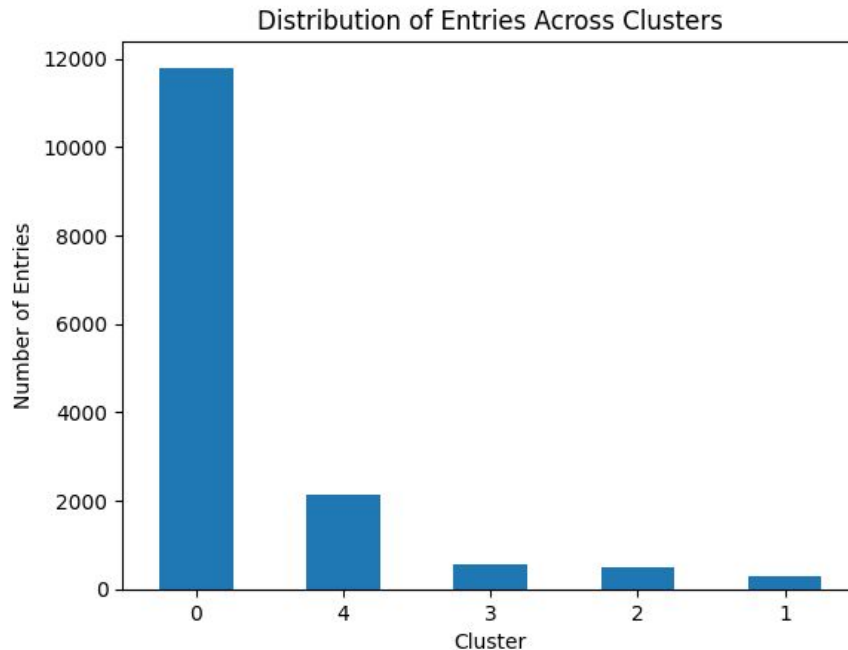
The dataset reveals a considerable volume of technical questions tied to user-related concerns, highlighting a community actively seeking assistance and information.

2. System-related:

A significant portion of system-related inquiries originates from core infrastructure, about ocean protocol components and its correlation with other projects/blockchains.

3. External factors:

The least amount of questions are about external factors that include other industries and projects.



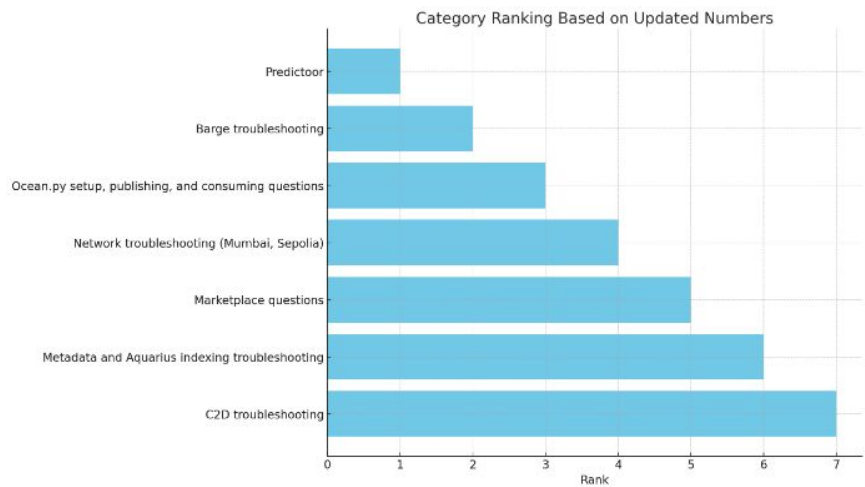
Technical Issues

Target: Identify and categorize the most common technical issues, indicating their potential sources (user-related, system-related, external factors).

The dataset doesn't cover messages from the "tech-issues" forum channel, however, a brief review of the tech-issues channel reveals the following question topics:

1. C2D troubleshooting
2. Metadata and Aquarius indexing troubleshooting
3. Marketplace questions
4. Network troubleshooting (Mumbai, Sepolia)
5. Ocean.py setup, publishing, and consuming questions
6. Barge troubleshooting
7. Predictoor

⚠ Recently, there's been a noticeable emphasis on C2D, suggesting that allocating more resources to this area could be beneficial. This approach is in line with the Ocean Protocol 2024 roadmap, indicating that the direction is already anticipated and addressed

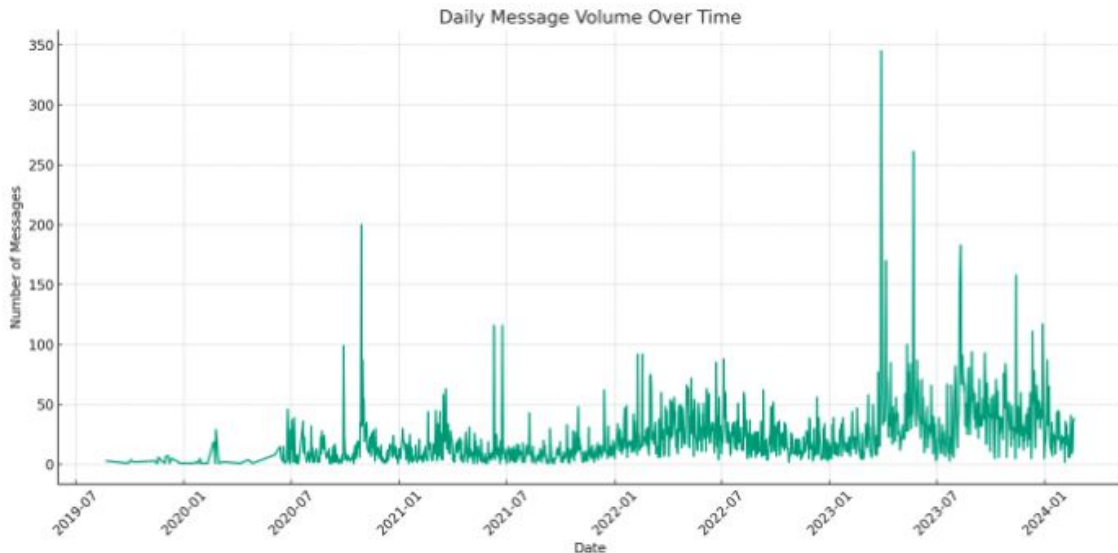


Prediction Model

Target: Develop a forecasting models to predict future server activity, providing the rationale for model selection. Define specific performance metrics for model evaluation. Think of how this model could be applied on the same data structure for other projects, not just Ocean Protocol.

Considering the temporal nature of the data and likely patterns such as daily or weekly cycles in chat activity, models that can capture seasonality, trend, and cyclic behavior would be appropriate. Choices include:


- **ARIMA/SARIMA:** Good for univariate time series data without external influences if the series is stationary or made stationary through differencing. SARIMA is particularly suited if the data exhibits seasonality.
- **Prophet:** Developed by Facebook, it's particularly user-friendly for forecasting with daily observations that display patterns on different time scales such as holidays, weekends, etc. It can handle outliers, missing data, and shifts in the trend well.
- **LSTM** (Long Short-Term Memory) Neural Networks: Suitable for capturing complex patterns in time series data, especially if there are long-term dependencies.



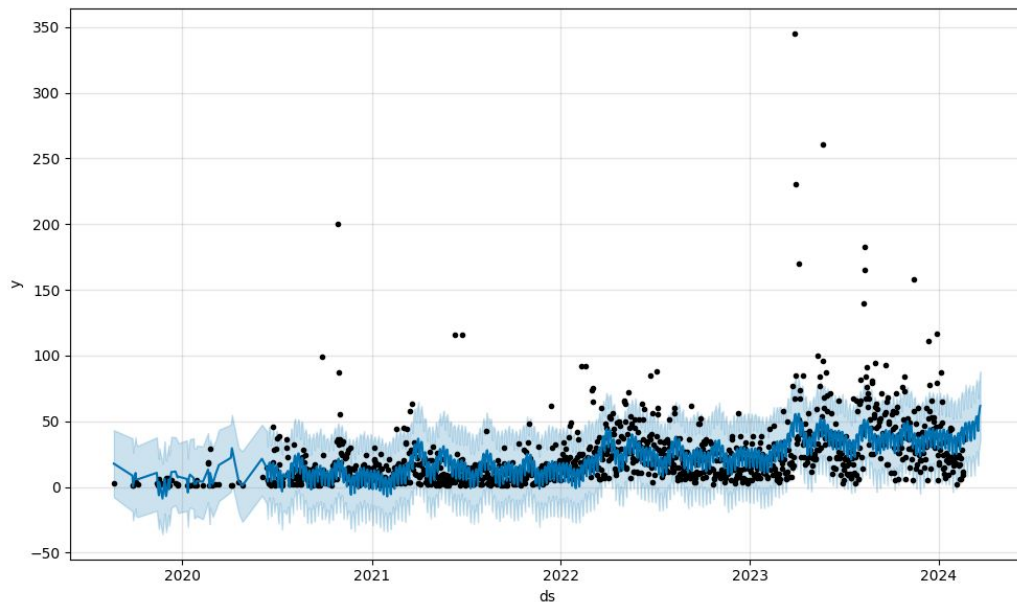
Prediction Model

Target: Develop a forecasting models to predict future server activity, providing the rationale for model selection. Define specific performance metrics for model evaluation. Think of how this model could be applied on the same data structure for other projects, not just Ocean Protocol.

Prediction using the Prophet

The forecast indicates an overall **upward trend**  in server activity. This trend aligns with Ocean Protocol's allocation of resources towards community development and the evolution of the protocol, suggesting continued growth in server activity.

Additionally, the **positive market** sentiment surrounding cryptocurrencies in 2024 is predicted to attract more individuals to crypto communities, further driving server activity. The correlation between artificial intelligence (AI) and blockchain technology represents a promising intersection that is expected to attract even more people over time, fueling continued growth in server activity.



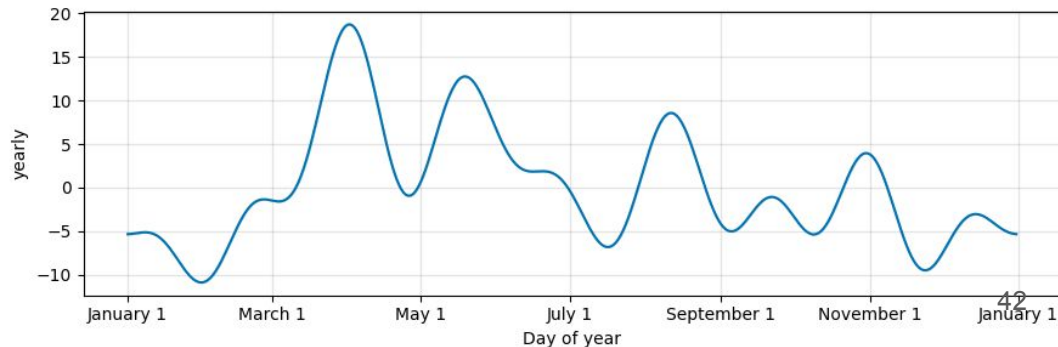
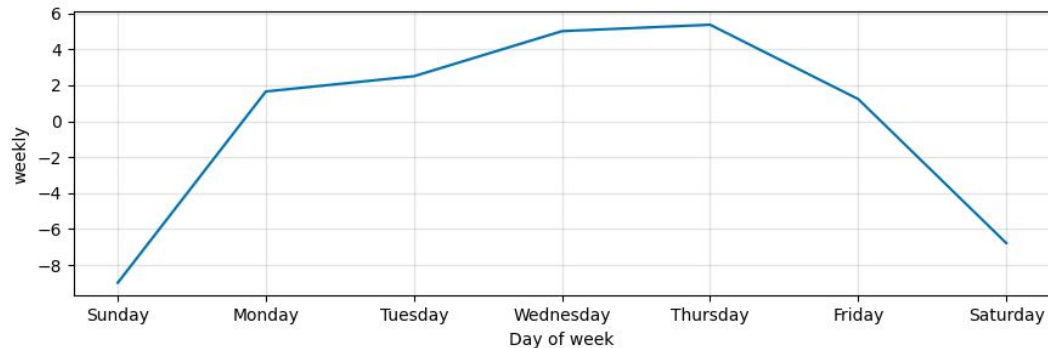
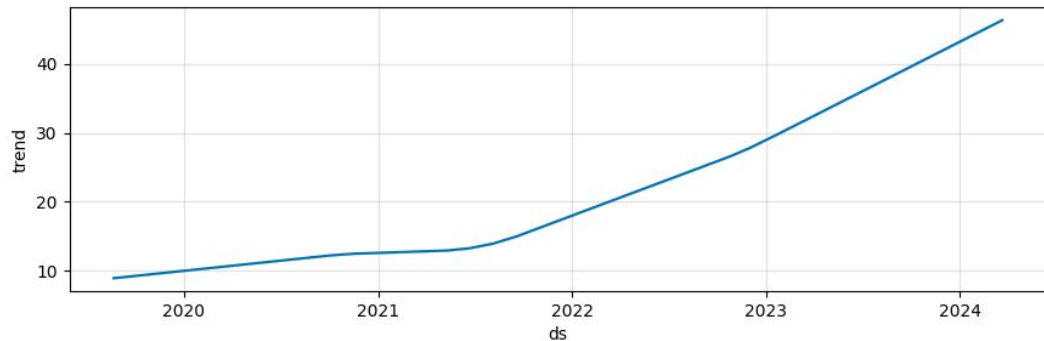
Prediction Model

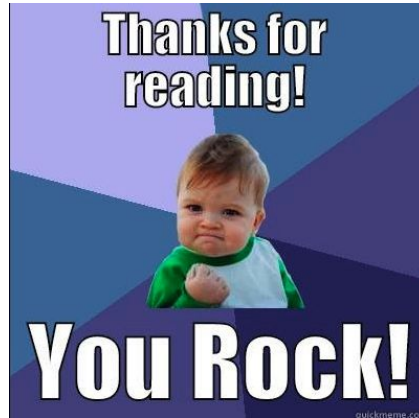
Target: Develop a forecasting models to predict future server activity, providing the rationale for model selection. Define specific performance metrics for model evaluation. Think of how this model could be applied on the same data structure for other projects, not just Ocean Protocol.

Prediction using the Prophet

Based on the forecasted series, it is advisable to allocate more resources and attention towards community engagement during specific periods:

- **Spring to early summer:** During this period, typically spanning from spring to early summer, heightened server activity is predicted.
- **Wednesday and Thursday:** Mid-week, particularly on Wednesdays and Thursdays, is forecasted to experience higher server activity. Focusing efforts on community engagement strategies during these days of the week can maximize outreach and interaction with users.
- **Evening times:** Evening hours are projected to see elevated server activity. Investing resources in community engagement efforts during these time periods can effectively target users who are more active during evenings, potentially increasing participation and engagement.





For the code and additional resources, head over to the [GitHub](#) repository. If you have any uncertainties or questions, please reach out via Discord(white_rider_) or [Twitter](#)!