0POSTS EDA:

A screenshot of a computer

Description automatically generated

1- around 33.33% of the posts published on the website was deleted.

A screenshot of a computer

Description automatically generated

A computer screen shot of a computer code

Description automatically generated

2- post types

A screenshot of a computer

Description automatically generated

The posts in stack exchange are divided into 7 types of posts

**1: Questions**

**2: Answer**

**3: Wiki:** This is a community-maintained post that provides general information about a topic.

**4: TagWikiExcerpt**: A short summary or introduction to a tag, typically displayed when hovering over a tag.

**5: TagWiki:** A detailed explanation about a tag, including usage guidance and relevant resources.

**6: ModeratorNomination:** Posts where users nominate themselves (or others) to become community moderators

**7: WikiPlaceholder:** A system-generated placeholder for a wiki that is pending creation.

Most of the posts are of type 1 or 2

A screenshot of a computer

Description automatically generated A graph of a number of numbers

Description automatically generated with medium confidence

It is obvious that the focus will be on the type 1: Questions & type 2: Answers

3- scores:

The Score column represents the **net votes** a post has received from the community:

Score=Upvotes−Downvotes

A white paper with numbers and letters

Description automatically generated

4- date columns

CreationDate, LastActivityDate, CommunityOwnedDate, and LastEditDate as object dtype means they are being treated as generic text. Later on we are gonna need change them

import pandas as pd

# Assuming your DataFrame is called df

date\_columns = ['CreationDate', 'LastActivityDate', 'CommunityOwnedDate', 'LastEditDate']

# Convert specified columns to datetime

df[date\_columns] = df[date\_columns].apply(pd.to\_datetime, errors='coerce')

A screenshot of a computer code

Description automatically generated

4- Parent ID

represents the ID of a "parent" post in a hierarchical structure.

* **Questions** are the top-level posts (they **do not** have a parent, so their ParentId is NULL).
* **Answers** are linked to a specific question (they **do** have a ParentId that refers to the question's
* **Other types of posts do not have parent ID**

5- AnswerCount:

Number of answers submitted in response to a question post.

Applies To: Only question posts (PostTypeId = 1).

Why is it NaN? For non-question posts (e.g., answers, tag wikis), AnswerCount does not apply, hence the missing values.

6- CommentCount:

Number of comments made on any post (questions, answers, wikis, etc.).

Applies To: All post types (both questions and answers).

Why is it 0? If no comments exist on a post, it will show 0

A screenshot of a black screen

Description automatically generated

7- Content License:

The ContentLicense values refer to the Creative Commons (CC) licenses under which the content on the platform is shared. These licenses determine how the content can be reused, shared, and modified. Here’s what each one means:

8- Columns to be dropped because of high null count:

AcceptedAnswerId 359639

LastEditorDisplayName 428238

CommunityOwnedDate 426583

OwnerDisplayName 421536

ClosedDate 415673

9- Correlation between the numeric features:

A screenshot of a computer screen

Description automatically generated

10- some relationships

A graph showing a blue dot

Description automatically generated with medium confidence

A screenshot of a computer program

Description automatically generated

11- time series analysis

A graph with a line

Description automatically generated

The reduction in Stack Exchange posts in 2020 can be attributed to several factors:

 **COVID-19 Pandemic Impact**

* The global pandemic disrupted daily life and work patterns. Many people faced stress, health concerns, and job insecurity, which could have reduced their participation in online communities.
* Remote work increased, but professionals may have had less time or energy to engage on Stack Exchange due to balancing personal and professional responsibilities.

 **Shifts in Online Behavior**

* During the pandemic, people turned to other communication platforms like **Slack, Discord, and GitHub Discussions** for real-time collaboration, reducing their reliance on traditional Q&A forums.
* Social media and specialized communities (e.g., Reddit, Twitter) also gained traction for quick answers and knowledge sharing.

 **Policy Changes and Community Discontent**

* In **2019-2020**, Stack Exchange faced **controversies** over moderation policies and the treatment of volunteer moderators, leading to community dissatisfaction. Some experienced users reduced their activity or left the platform entirely.
* Changes in the **Code of Conduct** and disputes over how moderators were treated caused tension, affecting engagement.

 **Maturity of Existing Knowledge**

* Many common technical questions were already answered, reducing the need for new posts.
* Increased use of **search engines** to find existing solutions rather than posting new questions.

The decline in Stack Exchange posts in **2022** can be attributed to several factors:​

1. **Emergence of AI Tools**  
   The release of AI language models like **ChatGPT** in **November 2022** provided users with alternative platforms to seek answers, reducing r\\\\eliance on traditional Q&A forums.
2. **Measurement Changes**  
   In **May 2022**, Stack Overflow recategorized its Google Analytics cookie to enhance user privacy, leading to a reported decrease in traffic data.
3. **Community Dynamics**  
   Perceptions of strict moderation and negative experiences may have discouraged participation, contributing to reduced engagement.
4. **Market Competition**  
   The rise of alternative platforms offering real-time collaboration and specialized communities diverted user activity from Stack Exchange.

12- Tags

A graph of a bar graph

Description automatically generated

A graph of a bar graph

Description automatically generated