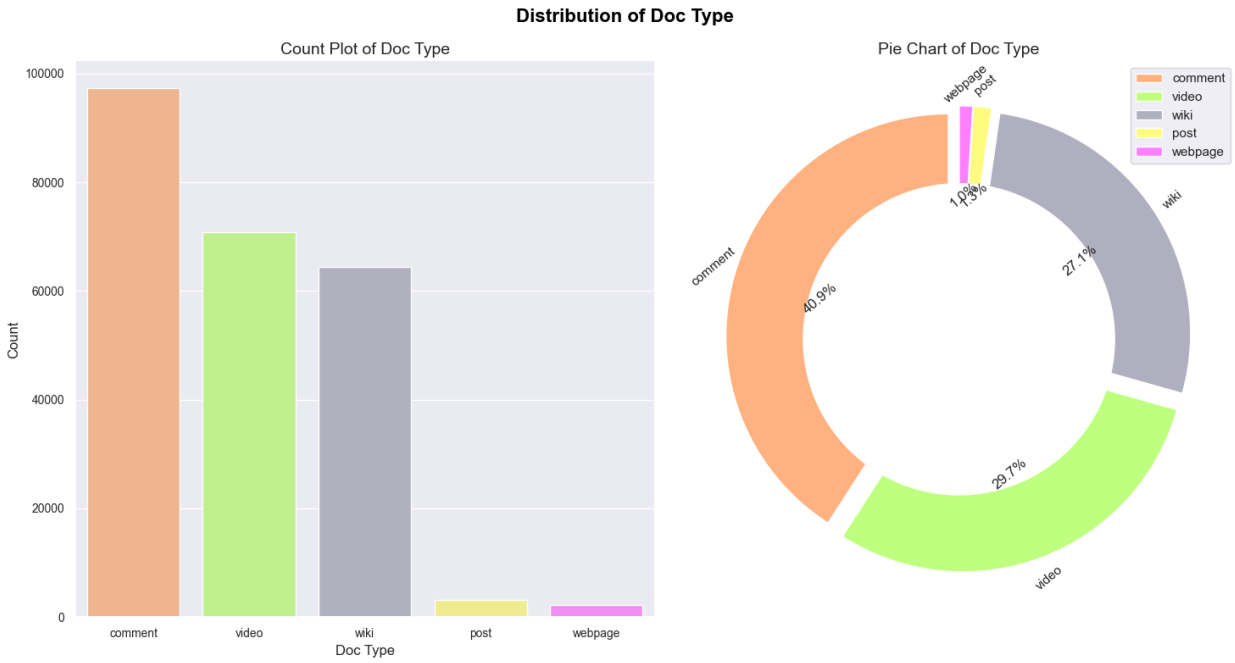
TASK – Data Cleaning Analysis

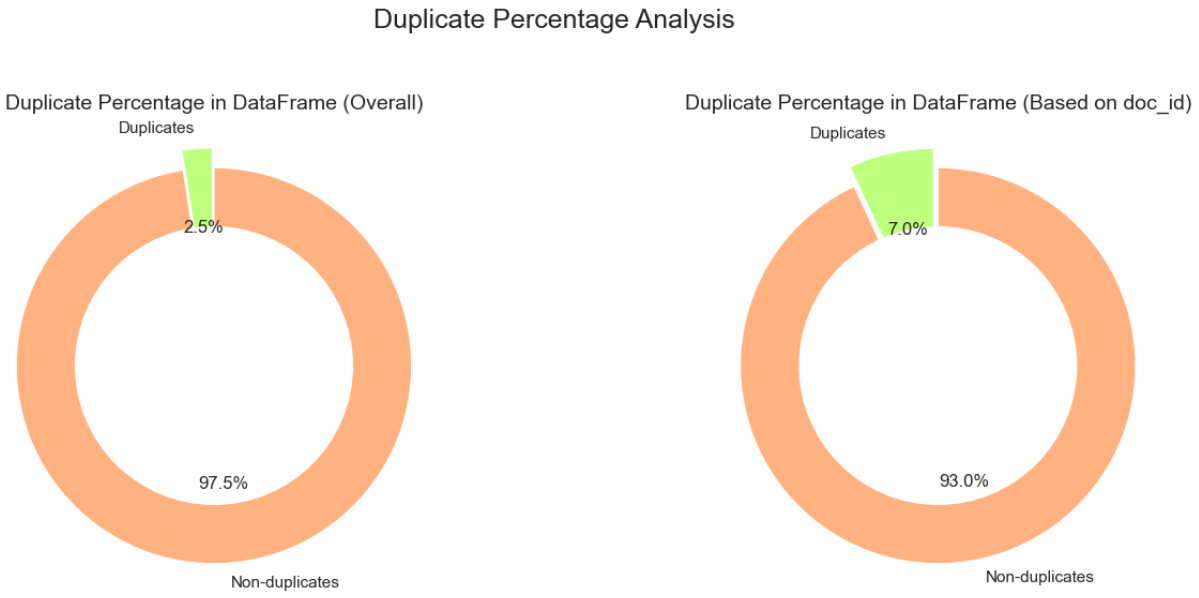
- There are a total of 193 JSONL files that I have found in the S3 bucket within the scrape folder.
- Inside those, there were a total of 2,38,111 documents stored.
- I have divided the entire data of 2,38,111 documents into 5 data frames as per their document types (wiki, webpage, post, comment, video).
- The distribution of documents as per doc_type is as per the below table.

Doc Type	Count
Comment	97380
Video	70820
Wiki	64497
Post	3118
Webpage	2296



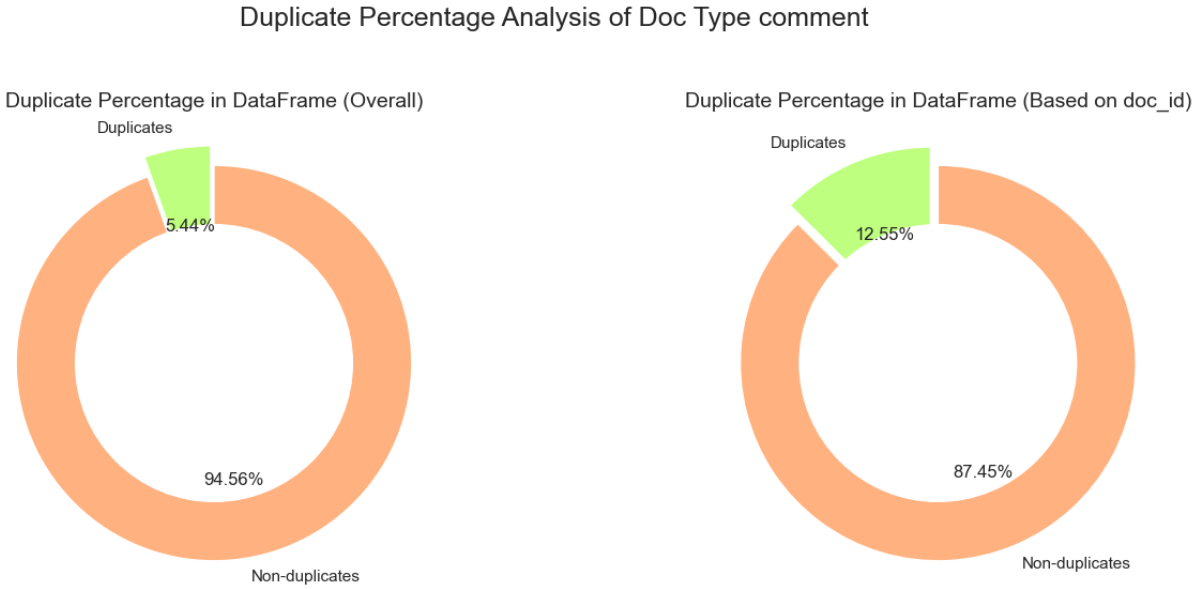
TASK – Data Cleaning Analysis

- Duplicates- Out of all documents 5915 records were duplicated.
  Duplicates as per document type-



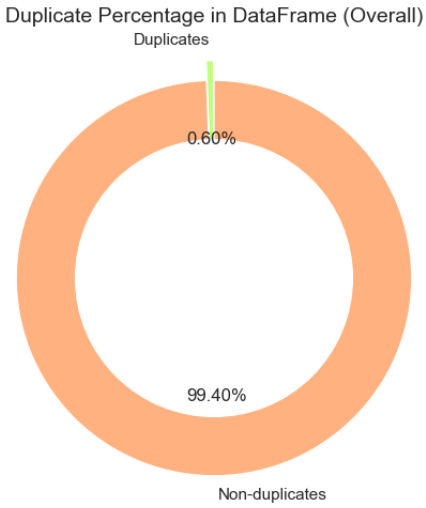
Duplicates based on Doc Type

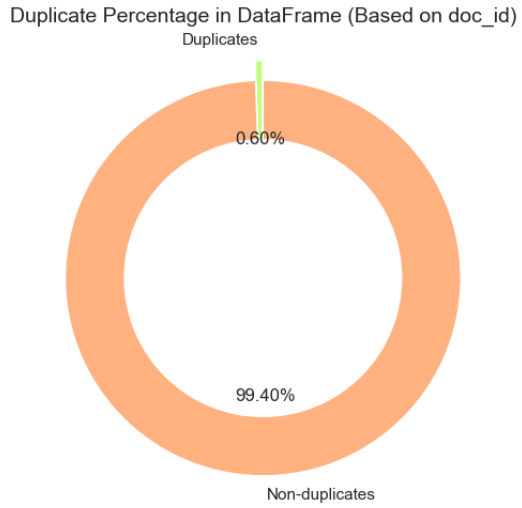
Doc Type	No. of duplicates	No. of duplicates (Doc ID)
Comment	5293	12217
Video	422	423
Post	165	165
Webpage	21	1313
Wiki	14	2479



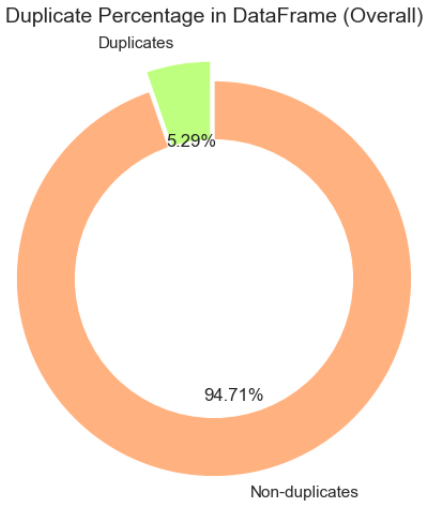
TASK – Data Cleaning Analysis

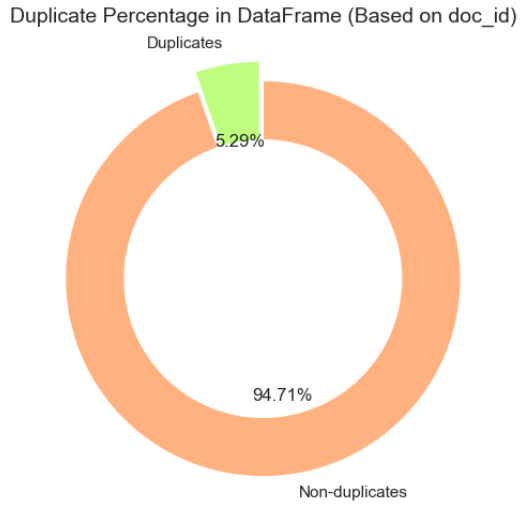
Duplicate Percentage Analysis of Doc Type video



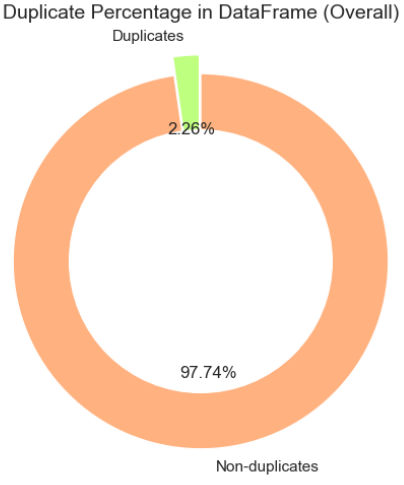


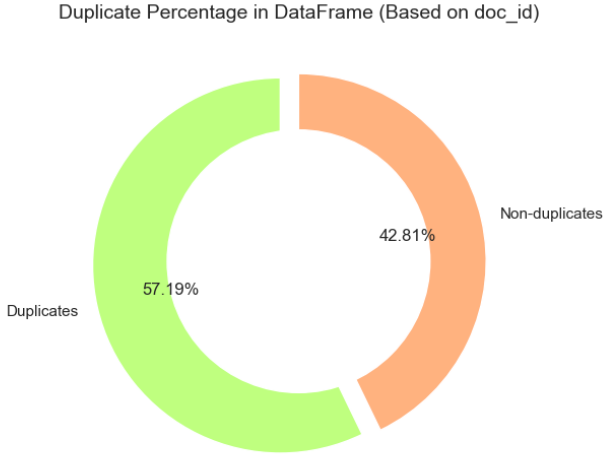
Duplicate Percentage Analysis of Doc Type post



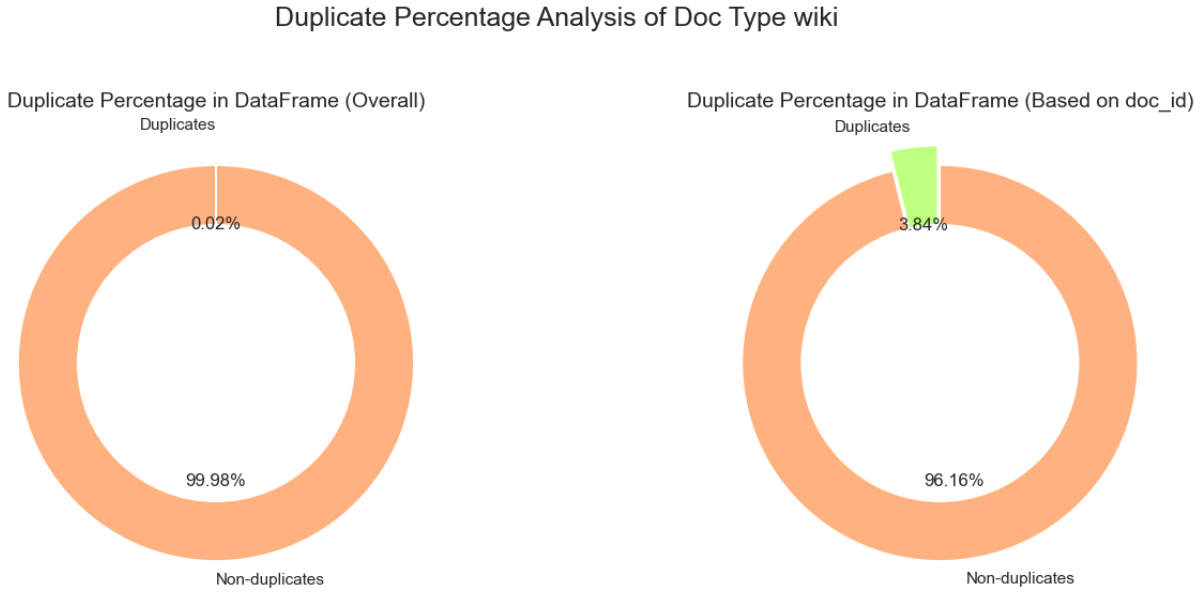


Duplicate Percentage Analysis of Doc Type webpage





TASK – Data Cleaning Analysis



Check that all fields exist in all documents and all doc types.

Doc Type	Columns Names
Comment	doc_id, source_platform, doc_type, user_id, user_name, channel_name, content, created_at, created_at_str, thread_id, parent_doc_id, title
Post	doc_id, source_platform, doc_type, user_id, user_name,channel_name, content, external_link, title, created_at, created_at_str
Video	doc_type, doc_id, source_platform, content, external_link, title, channel_name, created_at_str , 'created_at'

TASK – Data Cleaning Analysis

webpage	doc_id, doc_type, external_link, title, content, entities, attachments, source_platform, created_at, created_at_str
wiki	created_at, created_at_str, doc_id, doc_type, entities, external_link, title, content, source_platform

Common Fields vs Uncommon Fields

Common fields	title, doc_type, doc_id,created_at, created_at_str, content, source_platform
Uncommon fields	user_name, thread_id,external_link, entities, parent_doc_id, channel_name, attachments, user_id

For all fields: Check nonnull & non-empty strings.

- Table-wise fields with percentages of null values-
- Only those fields listed Below have some null values.

Table Name: - Comment

Column Name	**Null Value in %**
User Id	0.40 %

TASK – Data Cleaning Analysis

Username	0.40 %
Content	0.001 %
Title	59.56 %

Table Name: - Post

Column Names	**Null Value in %**
Content	19.11 %

Table Name: - Video

- The video Table has no null values.

Table Name: - Webpage

Column Names	**Null Value in %**
Attachments	99.04 %

Table Name: - Wiki

Column Names	**Null Value in %**
Source Platform	1.63 %

TASK – Data Cleaning Analysis

For specific fields

1). Document ID Check: Ensure that document IDs follow a specific format. It seems they are prefixed with "id:uds:uds::", followed by a long alphanumeric string. Count the docs with this regex for their ids, vs the ones that don't follow it.

- No documents of all the Doc Types follow such a format.

Example formats-

Table Name	**Doc ID**
Comment	t1_jk90ytb
Post	t3_13ib17s_0
Video	25d13ab737d5562273c3a639cad9289930d80df25f303fcb5f9976e76f3bebbe
Webpage	509ab47d97f4b27877d86176a1526aa2578f288160463bb7afac662a87a16706
Wiki	0db9ad670415179b013071139adce962e32d5b6bca469bb3e62e53c3b905a64a

TASK – Data Cleaning Analysis

2. Doc Type Validation: Get the set of all values of the Doc Type field. Get counts of the documents. Ensure the type is a valid string and from a predetermined set of valid document types, e.g., "webpage", "pdf", "article",etc.

- 5 Types- wiki, webpage, comment, post, video. Count already shown above with chart.
- There is no Null value and noise in the Doc Type, and it is satisfying all the conditions in the validation.

Doc Type	**Row count (Pass the validation)**
Comment	97380
Video	70820
Wiki	64497
Post	3118
Webpage	2296

TASK – Data Cleaning Analysis

3. External Link Validation: Ensure that "external_link" is not null and follows the proper URL format. It might be beneficial to add additional checks here to ensure that the link is accessible (returns a 200-status code, for example).

- All the links are in valid format.
- Checked the response only for Unique links.

Table Name	**200 responses**	**302 responses**	**404 responses**
Comment	No Links	No Links	No Links
Video	7397	0	0
Wiki	15994	4	3
Post	579	7	1210
Webpage	64	0	7

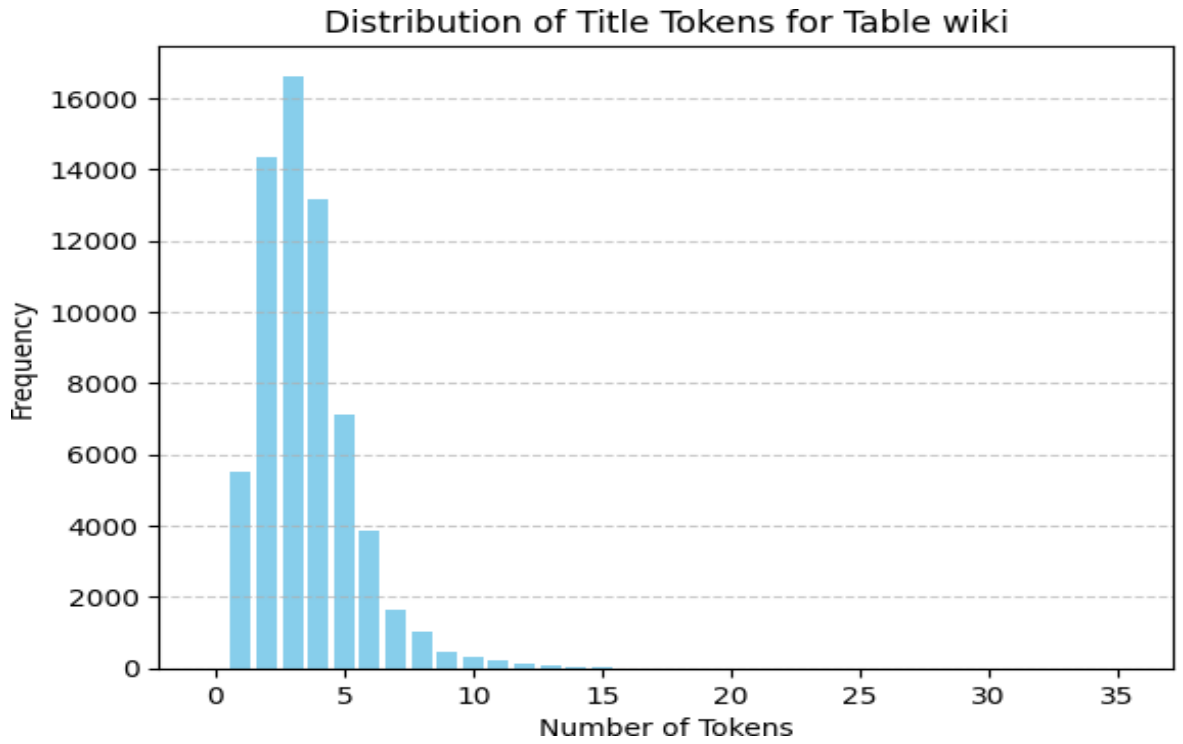
4. Title and Content Validation:

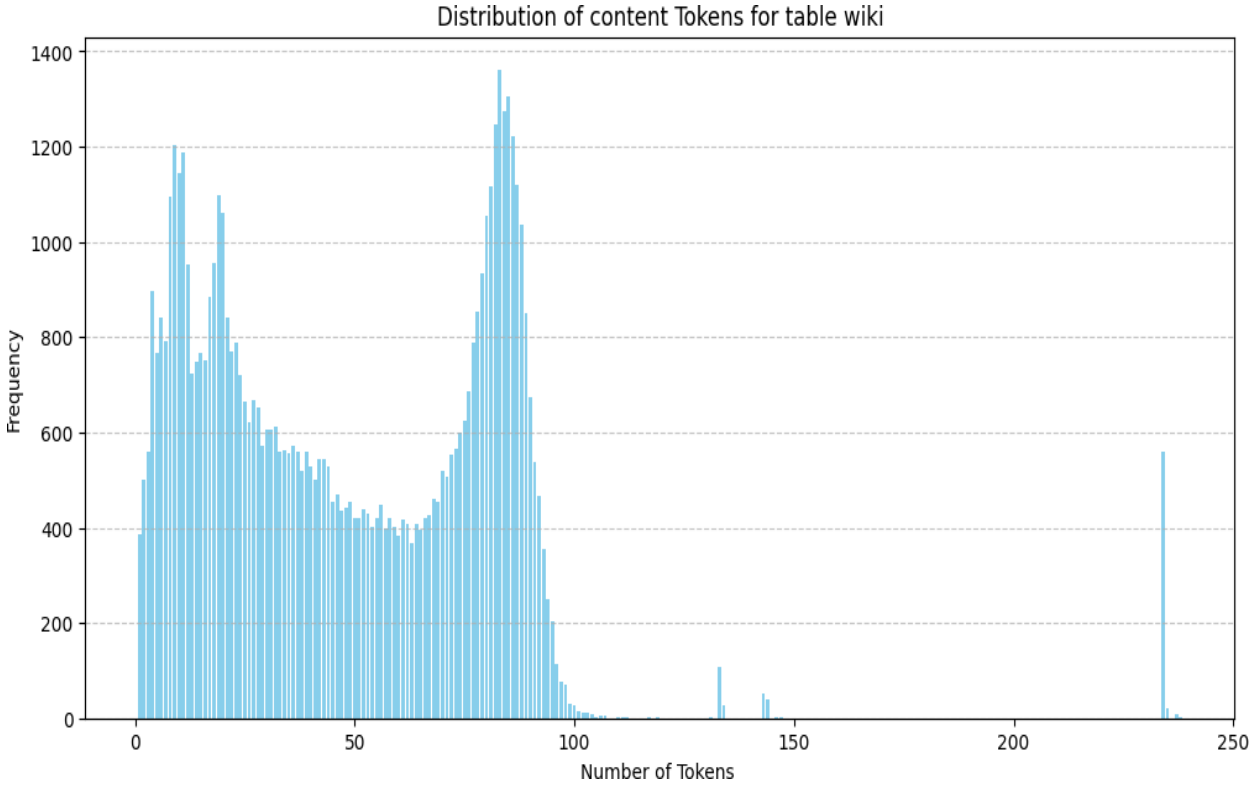
a. Ensure that the "title" and "content" fields are strings, not empty or null.

Table Names	**Null or empty string count**
Comment	57996
Video	0
Post	596 (content)
Webpage	0
Wiki	4 (content)

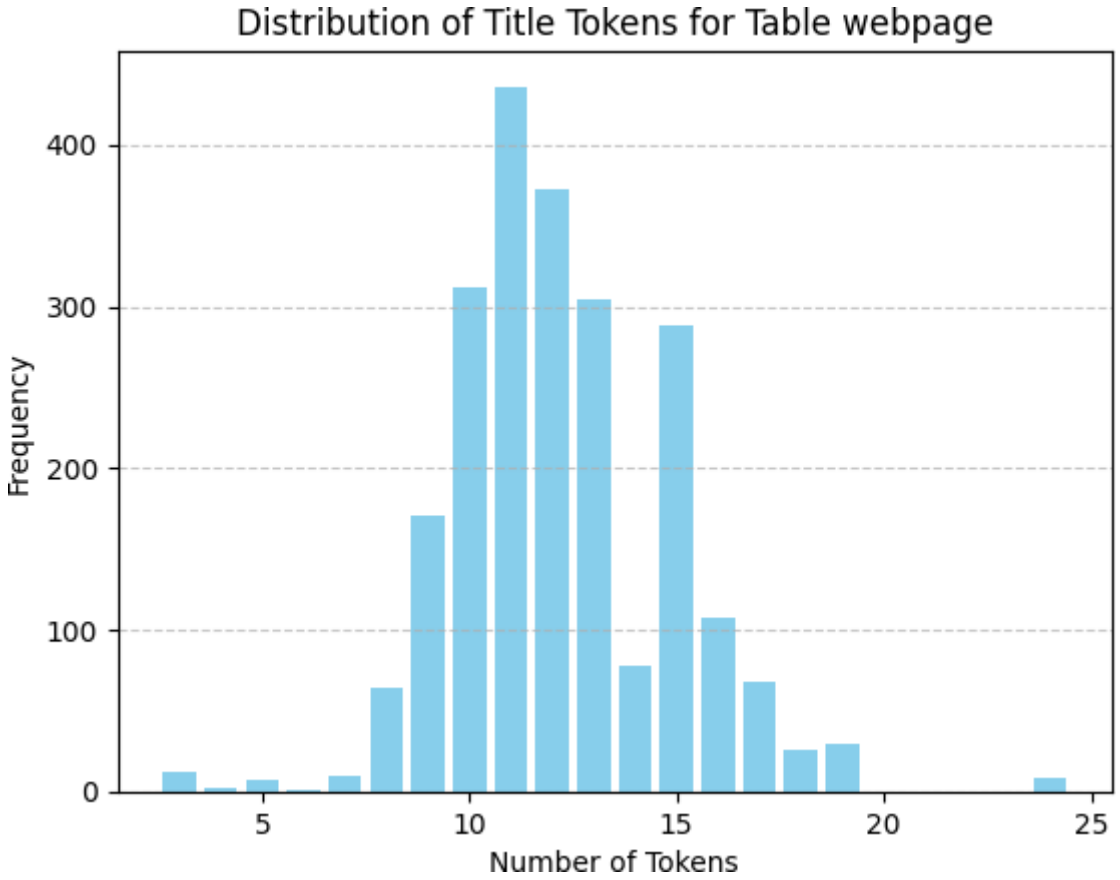
TASK – Data Cleaning Analysis

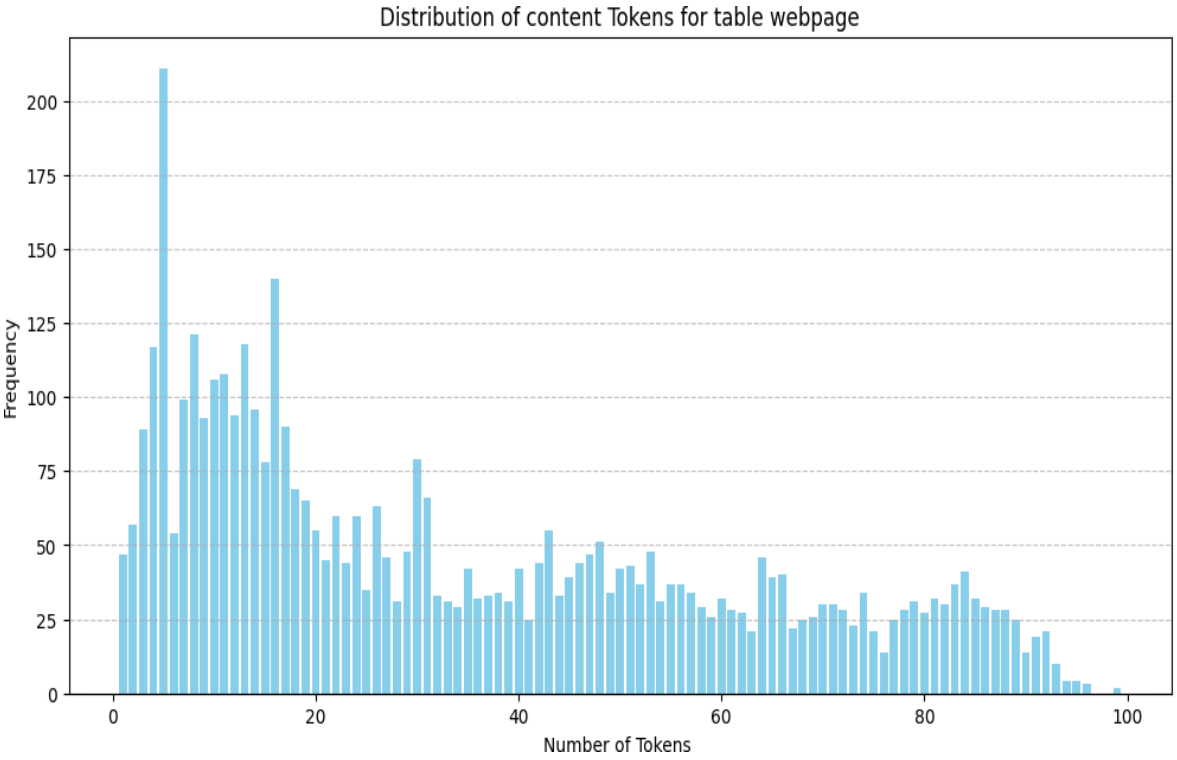
b. Check the distribution of the number of tokens for title and content.



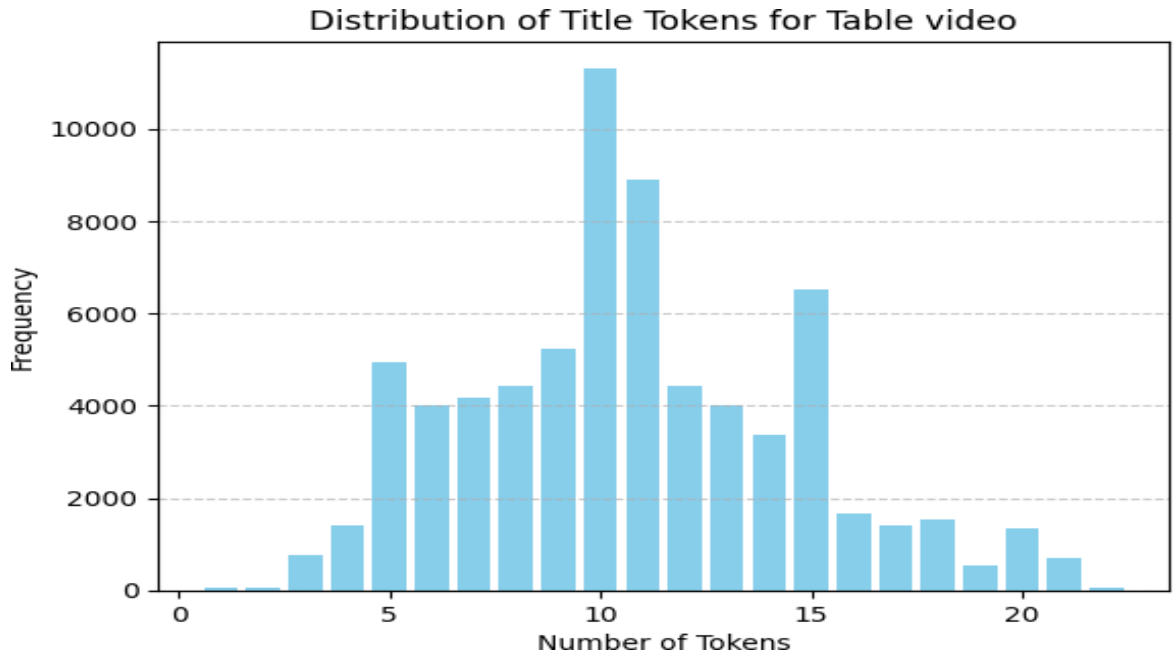


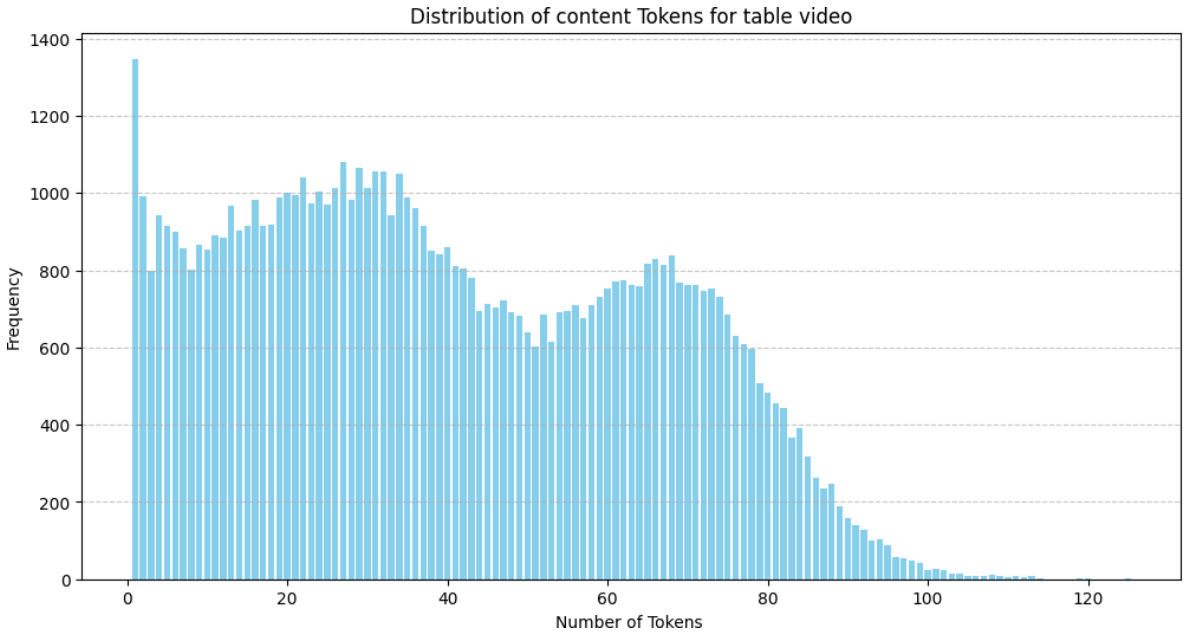
TASK – Data Cleaning Analysis



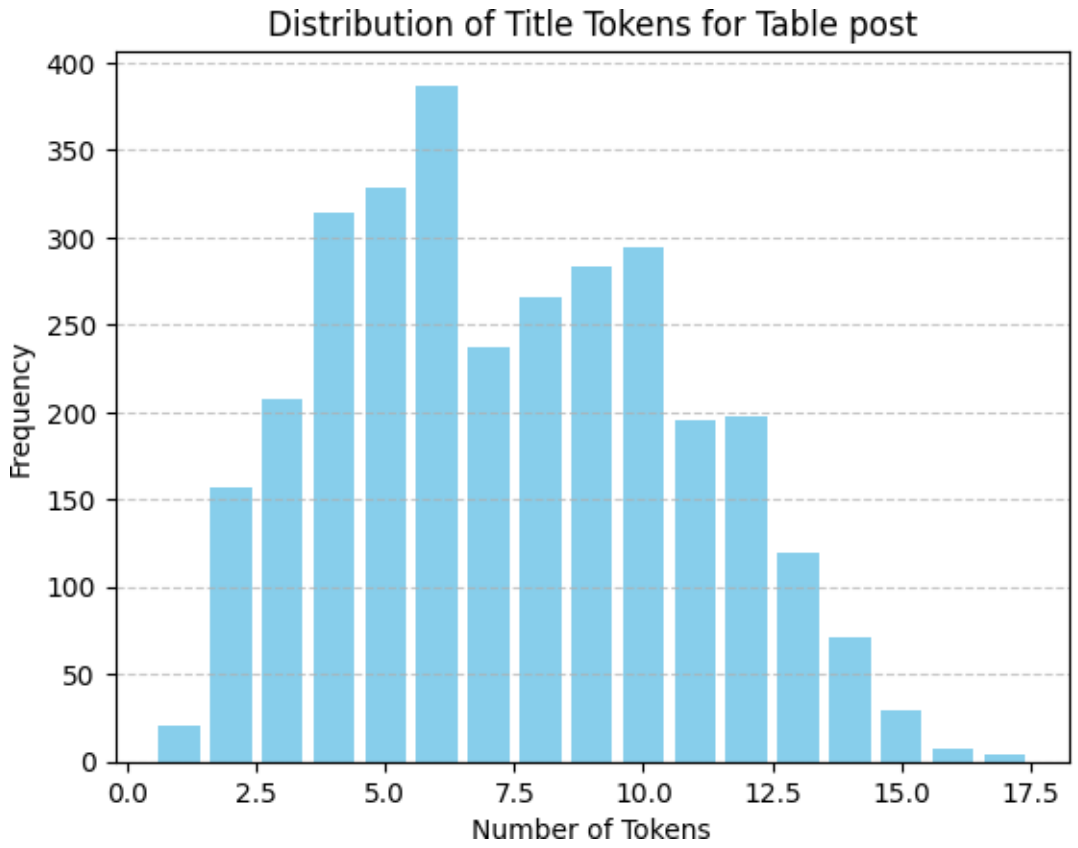


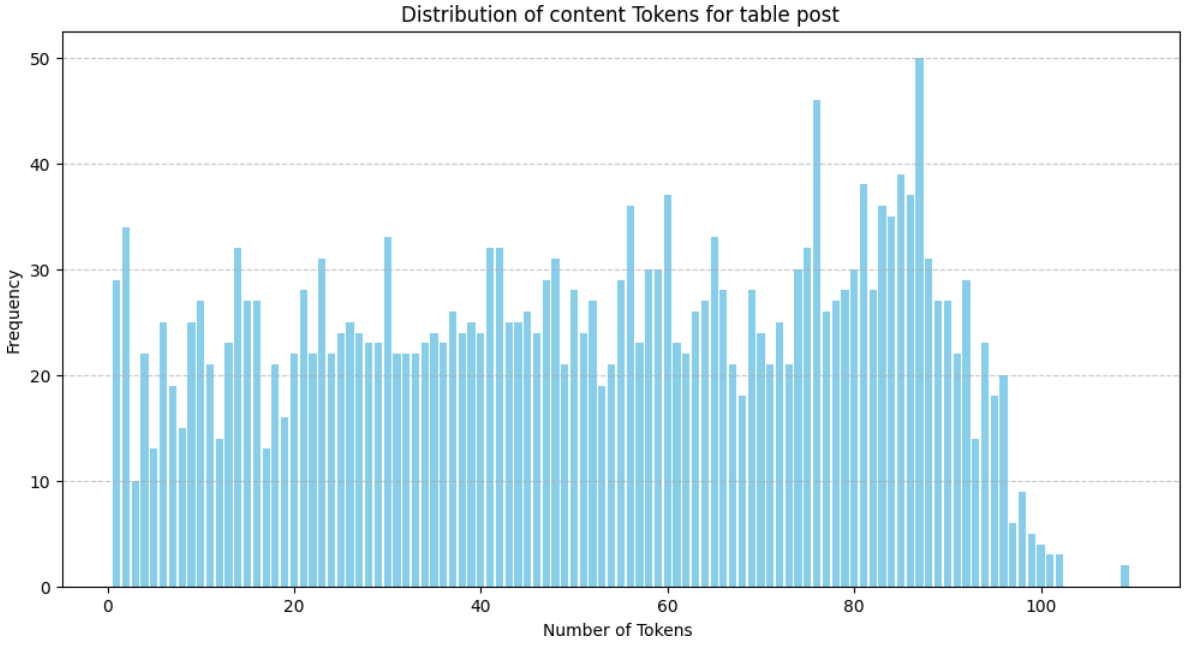
TASK – Data Cleaning Analysis



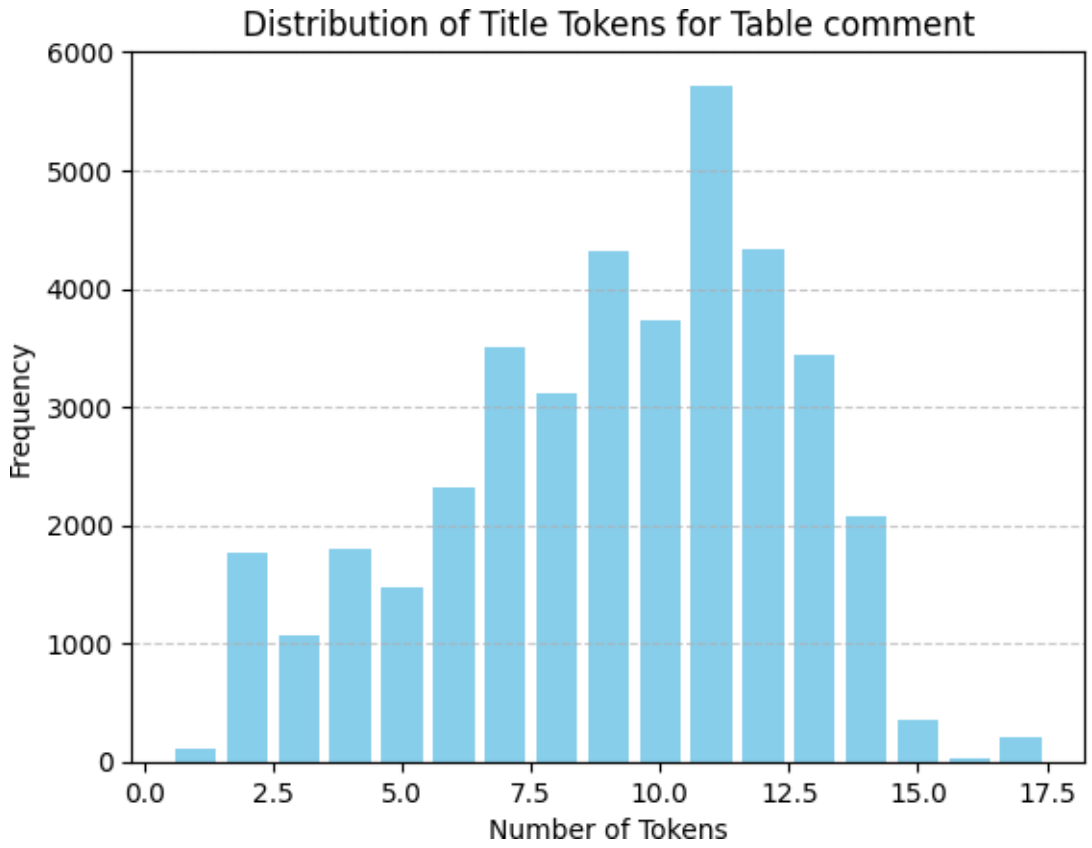


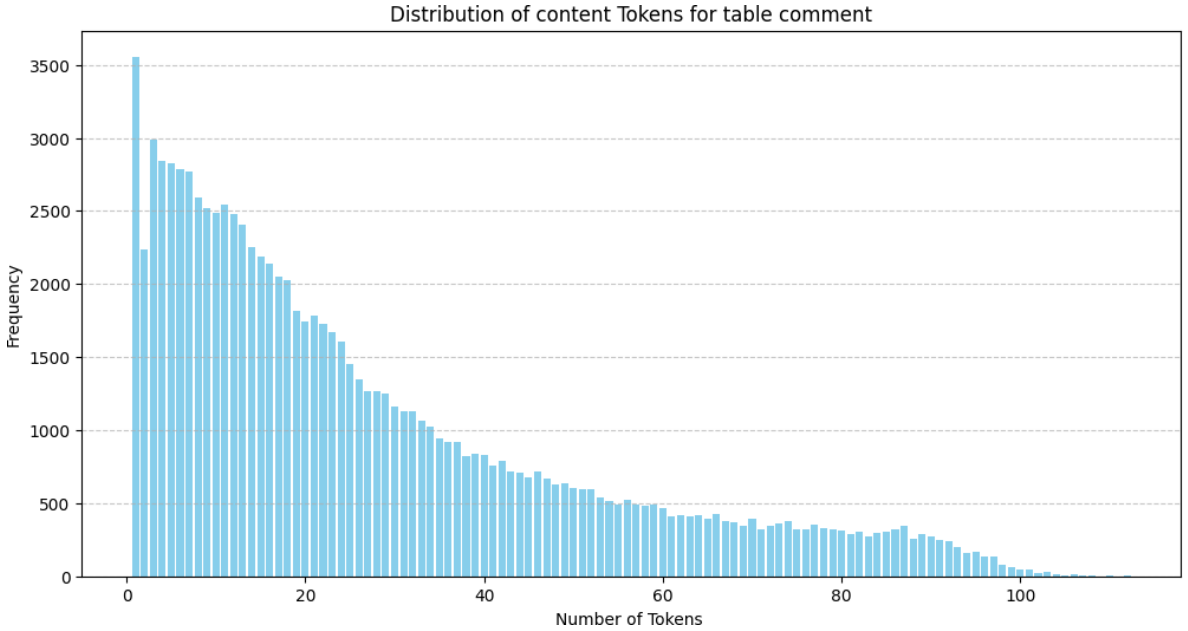
TASK – Data Cleaning Analysis





TASK – Data Cleaning Analysis





TASK – Data Cleaning Analysis

c. Check the number of documents with title and/or content less than X tokens. Start with X == 3.

Table Name	Less Than 3	Less than 2	Less than 1	Less than 0
Comment	3550	1238	0	0
Post	236	50	0	0
Video	2440	1391	0	0
Webpage	96	42	0	0
Wiki	20469	5782	1	1

d. Check documents that only have special characters in the content and/or title. (d & e both below).

Documents that have Special Characters

Table Name	Title	Content
Comment	0	163
Post	1	4
Video	0	129
Web Page	0	0
Wiki	1784	1703

e. Check documents that only have numbers in the content and/or title.

Documents that have only Numbers

Table Name	Title	Content
Comment	0	32
Post	0	0
Video	0	6
Web Page	0	0
Wiki	5	0

TASK – Data Cleaning Analysis

5. Entities Validation:

Ensure that "entities" is an array and contains valid strings. A number of documents that contain at least 1 entity vs those that contain any entities.

Entities field is only available in the webpage and wiki table, in others field is missing.

Table Name	Only 1	Any Entities (1 & 2)
Webpage	1100	2296
Wiki	0	64497

6. Source Platform Validation: Check that "source_platform" is a valid stringand from a pre-determined list of platforms.

pre-determined list of platforms –

Fandom	Official	Maxroll	Reddit	YouTube

Table Name	Available Platform
Comment	Reddit
Post	Reddit
Video	YouTube
Webpage	Official, Maxroll
Wiki	Fandom (nan is also there)

7. Timestamp Validation check:

Table name	Count of records where created_at > current date	Count of records where created_at and created_at_str are different
Wiki	0	0
Webpage	0	0
Video	0	10201
Post	0	845
Comment	0	35183

8. Is_Public Validation:

TASK – Data Cleaning Analysis

There is no such field in any of the documents.

9. Title case check:

Table name	Correct Records	Incorrect Records
Comment	33587	63793
Post	2682	436
Video	29405	41415
Webpage	2296	0
Wiki	61254	3243

10. Spam Document Detection: (Working on it)

I have found the top 10 repeated words in all the documents.

Still working on filtering the spam content out content column**.**

	content	count
0	foreign	532
1	[deleted]	258
2	thank you	251
3	Same	211
4	[	95
5	[removed]	91
6	Music]	90
7	Yes	67
8	Same here	46
9	Water	43

Conclusions

1. Handling Missing Values in the Wiki Table:

The "source_platform" column in the Wiki table has missing values.

Based on the analysis, all missing values in this column can be safely filled with 'fandom' as the unique value 'fandom' is associated with external links to the fandom website.

TASK – Data Cleaning Analysis

2. **Addressing Missing Values in the Webpage Table:**

The "attachments" column in the Webpage table contains missing values.

Observing that only 22 rows have a unique list of links in this column.

Suggestion: Consider filling in missing "attachments" values in other columns based on the content of these 22 rows, while carefully evaluating the relevance.

3**. No Null Values in Videos Table:**

The Videos table has no null values, which indicates data completeness and consistency in this table.

4. **Handling Null Values in the Post Table:**

The "content" column in the Post table has null values.

Upon closer inspection, it appears that rows with incorrect external link values lead to null values in the "content" column.

Recommendation: Investigate and rectify the issues causing these null values to ensure accurate content representation.

5. Addressing Null Values in the Comments Table:

Null values in the Comments table need further investigation to identify the root cause.

Consider exploring the relationship between null values and other columns, potentially external links or related data.

6. Duplicates and Data Quality:

Identified a certain percentage of duplicates within the dataset.

The overall duplicate rate is approximately 2.5% when considering all columns but increases to 7.0% when using the "Doc_id" column.

The Webpage table has the highest portion of duplicate IDs at 57.19%.

Recommendation: Implement data cleaning measures to remove or handle duplicates to improve data quality.

7. Quality of External Links:

The majority of external links are valid and result in a 200 response.

Noted that in the Post table, 579 links return a 200 response while 1210 links return a 404 error.

TASK – Data Cleaning Analysis

It's important to investigate the reasons behind the high number of 404 errors and address them for accurate and functional external references.

8. Future Steps:

Propose collaboration with the data engineer to implement appropriate data cleaning and preprocessing techniques to address missing values and duplicates.

Work on refining the external link validation process to minimize 404 errors and ensure accurate data.

Suggest regular data quality checks to maintain the accuracy and reliability of the dataset.

TASK – Data Cleaning Analysis