[1]:
```
# Import libraries used in data preparation
import pandas as pd
```

[2]:
```
# Load the youtube_videos dataset .. in same folder
dataset = pd.read_csv("youtube_videos.csv") #load dataset into DataFrame
dataset # ensure read it successfully
```

[2]:

	index	title	description	publish_date	publish_time	likes	views	category	channel_id
0	1	Joe Biden congratulates Donald Trump as pair s...	NaN	2024-11-13	19:00:32	2167	73267	news	UC16niRr50-MSBwiO3YDb3RA
1	2	The Queen vs The President: "Boom" - BBC News	The Queen and Prince Harry have responded to B...	2016-04-29	17:59:12	79031	4319639	news	UC16niRr50-MSBwiO3YDb3RA
2	3	Science rejuvenates woman's skin cells to 30 y...	Researchers have rejuvenated a 53-year-old wom...	2022-04-08	19:00:11	32469	1099386	news	UC16niRr50-MSBwiO3YDb3RA
3	4	World's largest coral found in Pacific Ocean |...	The largest coral ever recorded has been found...	2024-11-16	18:00:00	1871	189218	news	UC16niRr50-MSBwiO3YDb3RA
4	5	Dinosaur fossil from asteroid strike that caus...	A dinosaur's leg, stunningly preserved, has be...	2022-04-07	13:00:31	54317	2281702	news	UC16niRr50-MSBwiO3YDb3RA
...	...	...	...	...	...	...	...	...	...
1195	1196	If You're Happy arabic no music (learn arabic)...	If You're Happy arabic no music (learn arabic)...	2018-01-07	08:25:45	1991	1204353	children	UCWIdqSQekeGmUWlSFeCiEnA
1196	1197	أنشودة توت توت بدون موسيقى	أنشودة توت توت بدون موسيقى اشتركوا معنا \n\n...	2019-08-29	18:03:35	128068	30724511	children	UCWIdqSQekeGmUWlSFeCiEnA
1197	1198	أنشودة الخضروات - السيدة بطاطا بدون موسيقى - ق...	أنشودة الخضروات - السيدة بطاطا - بدون موسيقى...	2019-11-08	16:09:45	11687	6161082	children	UCWIdqSQekeGmUWlSFeCiEnA
1198	1199	أغنية سباق الأرنب والسلحفاة - أغاني أطفال بال...	أغنية الأرنب والسلحفاة - أغاني أطفال باللغة ا...	2020-06-29	13:14:12	136618	86745467	children	UCWIdqSQekeGmUWlSFeCiEnA
1199	1200	Islamic cartoon for kids in english - The Secr...	Osratouna entertains & educates your child wit...	2018-09-12	13:22:59	1632	389924	children	UCWIdqSQekeGmUWlSFeCiEnA

1200 rows × 9 columns

[3]:
```
# get the dataset Size
size = dataset.shape
size
```

[3]:
```
(1200, 9)
```

[4]:
```
# get the dataset dimension
dimension = dataset.ndim
dimension
```

[4]:
```
2
```

[5]:
```
# get columns' titles
titles = dataset.columns
titles
```

[5]:
```
Index(['index', 'title', 'description', 'publish_date', 'publish_time',
       'likes', 'views', 'category', 'channel_id'],
      dtype='object')
```

[6]:
```
# Step 1 : Data Exploration
# Display the first 5 rows of the dataset using head method
dataset.head()
```

[6]:

	index	title	description	publish_date	publish_time	likes	views	category	channel_id
0	1	Joe Biden congratulates Donald Trump as pair s...	NaN	2024-11-13	19:00:32	2167	73267	news	UC16niRr50-MSBwiO3YDb3RA
1	2	The Queen vs The President: "Boom" - BBC News	The Queen and Prince Harry have responded to B...	2016-04-29	17:59:12	79031	4319639	news	UC16niRr50-MSBwiO3YDb3RA
2	3	Science rejuvenates woman's skin cells to 30 y...	Researchers have rejuvenated a 53-year-old wom...	2022-04-08	19:00:11	32469	1099386	news	UC16niRr50-MSBwiO3YDb3RA
3	4	World's largest coral found in Pacific Ocean |...	The largest coral ever recorded has been found...	2024-11-16	18:00:00	1871	189218	news	UC16niRr50-MSBwiO3YDb3RA
4	5	Dinosaur fossil from asteroid strike that caus...	A dinosaur's leg, stunningly preserved, has be...	2022-04-07	13:00:31	54317	2281702	news	UC16niRr50-MSBwiO3YDb3RA

[7]:
```
# Display information about the dataset
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1200 entries, 0 to 1199
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   index         1200 non-null   int64
 1   title         1200 non-null   object
 2   description   1167 non-null   object
 3   publish_date  1200 non-null   object
 4   publish_time  1200 non-null   object
 5   likes         1200 non-null   int64
 6   views         1200 non-null   int64
 7   category      1200 non-null   object
 8   channel_id    1200 non-null   object
dtypes: int64(3), object(6)
memory usage: 84.5+ KB
```

[8]:
```
# Display descriptive statistics of the dataset "for numeric columns"
dataset.describe()
```

[8]:

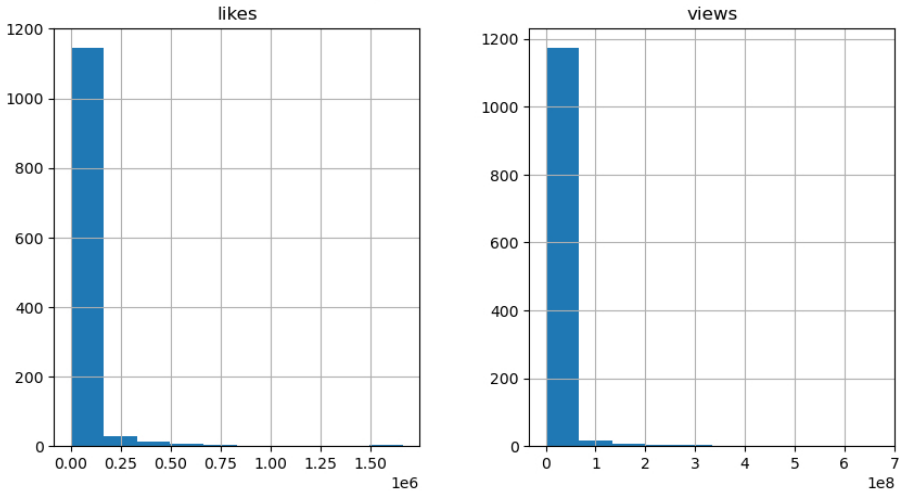
	index	likes	views
count	1200.000000	1.200000e+03	1.200000e+03
mean	600.500000	3.221288e+04	7.381537e+06
std	346.554469	1.145511e+05	3.574763e+07
min	1.000000	0.000000e+00	9.000000e+00
25%	300.750000	2.322500e+02	1.650100e+04
50%	600.500000	2.140000e+03	2.709995e+05
75%	900.250000	1.114850e+04	1.226443e+06
max	1200.000000	1.662526e+06	6.675060e+08

[9]:
```
dataset.dtypes
```

[9]:
```
index            int64
title           object
description     object
publish_date    object
publish_time    object
likes            int64
views            int64
category        object
channel_id      object
dtype: object
```

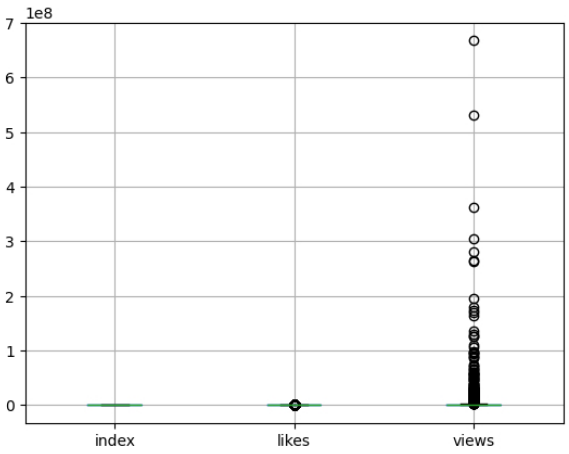
[10]:
```
dataset[['likes', 'views']].hist(figsize=(10, 5))
```

[10]:
```
array([[<Axes: title={'center': 'likes'}>,
        <Axes: title={'center': 'views'}>]], dtype=object)
```



[11]:
```
dataset.boxplot()
```

[11]:
```
<Axes: >
```



[12]:
```
# Step 2 : Data Cleaning

#count missing values in each attribute
dataset.isnull().sum() #or pd.DataFrame({'missing': dataset.isnull().sum()})
```

[12]:
```
index            0
title            0
description     33
publish_date     0
publish_time     0
likes            0
views            0
category         0
channel_id       0
dtype: int64
```

[13]:
```
# handel missing values in description
dataset['description'] = dataset['description'].fillna('No description available')

dataset.isnull().sum()
```

[13]:
```
index          0
title          0
description    0
```

```
publish_date    0
publish_time    0
likes           0
views           0
category        0
channel_id      0
dtype: int64
```

[14]:
```
dataset.duplicated()
```

[14]:
```
0       False
1       False
2       False
3       False
4       False
        ...
1195    False
1196    False
1197    False
1198    False
1199    False
Length: 1200, dtype: bool
```

[15]:
```
# Remove duplicates if exsist but we ensure no dupliaction
dataset = dataset.drop_duplicates()

dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1200 entries, 0 to 1199
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   index         1200 non-null   int64
 1   title         1200 non-null   object
 2   description   1200 non-null   object
 3   publish_date  1200 non-null   object
 4   publish_time  1200 non-null   object
 5   likes         1200 non-null   int64
 6   views         1200 non-null   int64
 7   category      1200 non-null   object
 8   channel_id    1200 non-null   object
dtypes: int64(3), object(6)
memory usage: 84.5+ KB
```

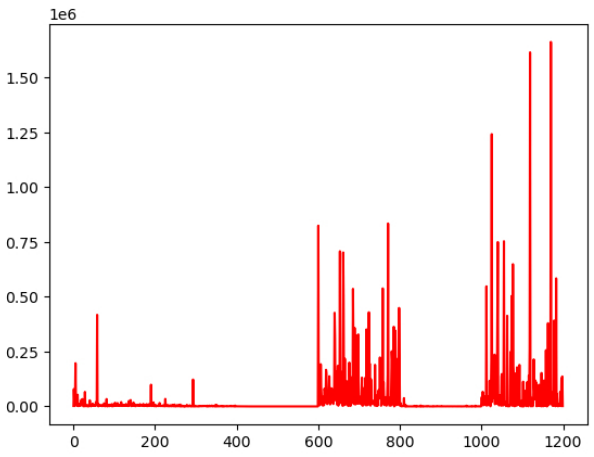
[16]:
```
# Display summary statistics for all columns
dataset.describe(include='all')
```

[16]:

	index	title	description	publish_date	publish_time	likes	views	category	channel_id
count	1200.000000	1200	1200	1200	1200	1.200000e+03	1.200000e+03	1200	1200
unique	NaN	1163	966	769	1020	NaN	NaN	5	6
top	NaN	فيلم 3D	To license ISU footage: https://bit.ly/3M7bcaF...	2024-11-22	17:00:02	NaN	NaN	news	UC16niRr50-MSBwiO3YDb3RA
freq	NaN	4	53	28	10	NaN	NaN	400	200
mean	600.500000	NaN	NaN	NaN	NaN	3.221288e+04	7.381537e+06	NaN	NaN
std	346.554469	NaN	NaN	NaN	NaN	1.145511e+05	3.574763e+07	NaN	NaN
min	1.000000	NaN	NaN	NaN	NaN	0.000000e+00	9.000000e+00	NaN	NaN
25%	300.750000	NaN	NaN	NaN	NaN	2.322500e+02	1.650100e+04	NaN	NaN
50%	600.500000	NaN	NaN	NaN	NaN	2.140000e+03	2.709995e+05	NaN	NaN
75%	900.250000	NaN	NaN	NaN	NaN	1.114850e+04	1.226443e+06	NaN	NaN
max	1200.000000	NaN	NaN	NaN	NaN	1.662526e+06	6.675060e+08	NaN	NaN

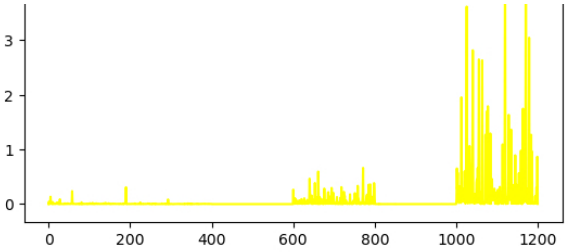
[17]:
```
dataset['likes'].plot.line(color = 'red')
```

[17]:
```
<Axes: >
```



[18]:
```
dataset['views'].plot.line(color = 'yellow')
```

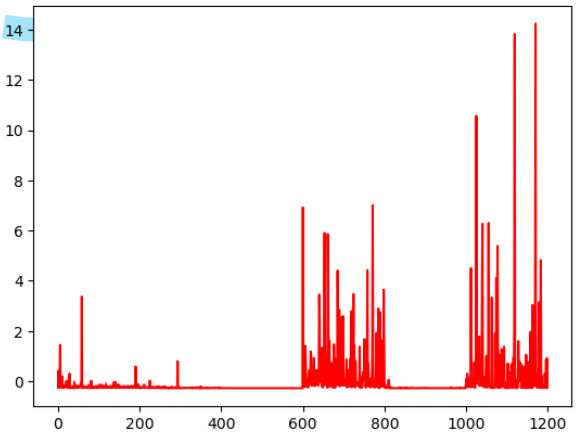
[18]:
```
<Axes: >
```



[19]:
```
# Step 3 : Data Transformation

from sklearn.preprocessing import StandardScaler

# numeric features
numeric_features = ['likes', 'views']

# Normlize numeric features using StandardScaler
numeric_transformer = StandardScaler()
scaler = StandardScaler()

# Apply the fit_transform method to the numeric columns
dataset[numeric_features] = scaler.fit_transform(dataset[numeric_features])
```

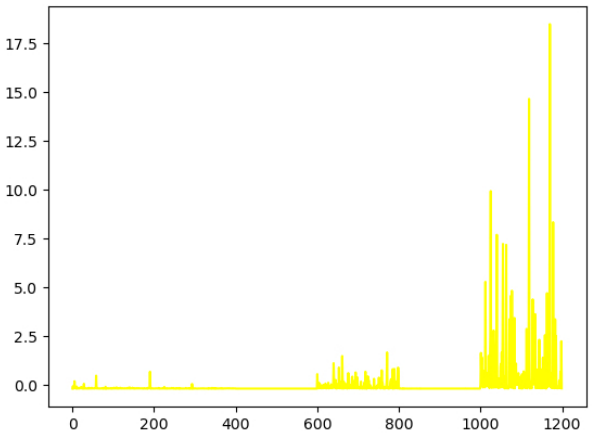
[20]:
```
dataset['likes'].plot.line(color = 'red')
```

[20]:
```
<Axes: >
```



[21]:
```
dataset['views'].plot.line(color = 'yellow')
```

[21]:
```
<Axes: >
```



[22]:
```
from sklearn.preprocessing import LabelEncoder


# LabelEncoder
label_encoder = LabelEncoder()
# Fit and transform the category column
dataset['category_encoded'] = label_encoder.fit_transform(dataset['category'])

#  mapping of categories to encoded integers ad dictionary
category_mapping = dict(zip(label_encoder.classes_, range(len(label_encoder.classes_))))
#Label_encoder.classes_ :  array of unique categories found in category column
#range(len(label_encoder.classes_)) : range of integers from 0 to the number of unique categories.
#zip(label_encoder.classes_, range(len(label_encoder.classes_))) Pairs each category with its corresponding integer.

category_mapping
```

[22]:
```
{'children': 0, 'cooking': 1, 'education': 2, 'news': 3, 'sport': 4}
```

[23]:
```
dataset.head()
```

[23]:

	index	title	description	publish_date	publish_time	likes	views	category	channel_id	category_encoded
0	1	Joe Biden congratulates Donald Trump as pair s...	No description available	2024-11-13	19:00:32	-0.262402	-0.204526	news	UC16niRr50-MSBwiO3YDb3RA	3
1	2	The Queen vs The President: "Boom" - BBC News	The Queen and Prince Harry have responded to B...	2016-04-29	17:59:12	0.408880	-0.085689	news	UC16niRr50-MSBwiO3YDb3RA	3

2	3	Science rejuvenates woman's skin cells to 30 y...	Researchers have rejuvenated a 53-year-old wom...	2022-04-08	19:00:11	0.002237	-0.175809	news	UC16niRr50-MSBwiO3YDb3RA	3
3	4	World's largest coral found in Pacific Ocean |...	The largest coral ever recorded has been found...	2024-11-16	18:00:00	-0.264987	-0.201281	news	UC16niRr50-MSBwiO3YDb3RA	3
4	5	Dinosaur fossil from asteroid strike that caus...	A dinosaur's leg, stunningly preserved, has be...	2022-04-07	13:00:31	0.193043	-0.142722	news	UC16niRr50-MSBwiO3YDb3RA	3

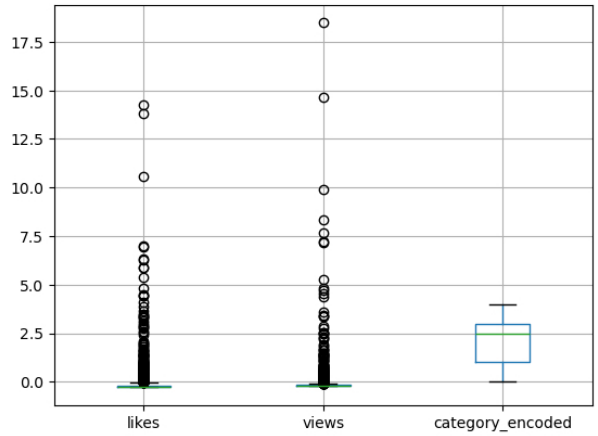
[24]:
```
dataset.drop(columns=['index'], inplace=True) #usless coulmn
dataset.describe()
```

[24]:

	likes	views	category_encoded
count	1.200000e+03	1.200000e+03	1200.000000
mean	2.368476e-17	1.184238e-17	2.166667
std	1.000417e+00	1.000417e+00	1.344270
min	-2.813268e-01	-2.065761e-01	0.000000
25%	-2.792985e-01	-2.061146e-01	1.000000
50%	-2.626374e-01	-1.989923e-01	2.500000
75%	-1.839629e-01	-1.722537e-01	3.000000
max	1.423812e+01	1.847394e+01	4.000000

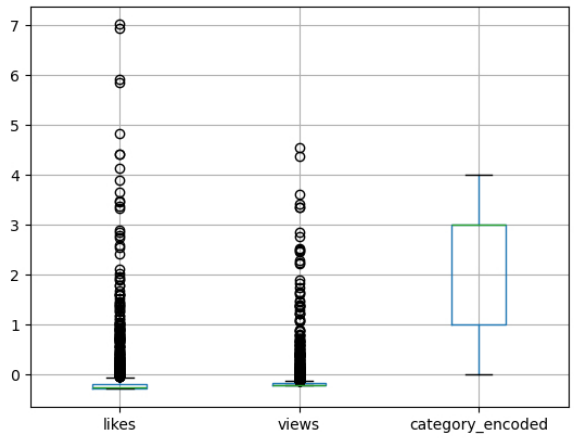
[25]:
```
dataset.boxplot()
```

[25]:
```
<Axes: >
```



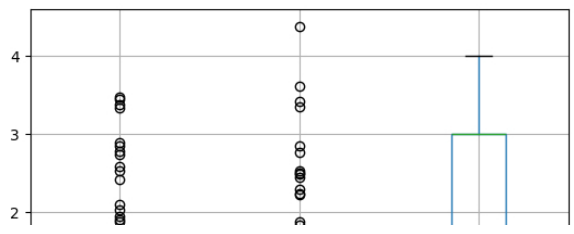
[26]:
```
# indices of the top largest views
indices = dataset['views'].nlargest(10).index

# drop this 10 rows from the dataset
dataset = dataset.drop(indices)

dataset.boxplot()
```

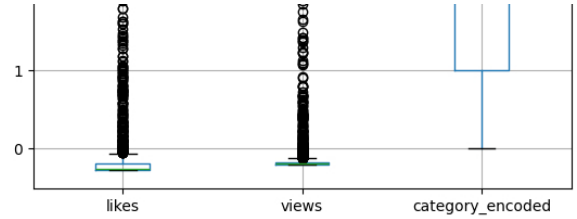
[26]:
```
<Axes: >
```



[27]:
```
# indices of the top largest likes
indices = dataset['likes'].nlargest(10).index

# drop this 10 rows from the dataset
dataset = dataset.drop(indices)

dataset.boxplot()
```

[27]:
```
<Axes: >
```





[28]:
```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1180 entries, 0 to 1199
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   title             1180 non-null   object
 1   description       1180 non-null   object
 2   publish_date      1180 non-null   object
 3   publish_time      1180 non-null   object
 4   likes             1180 non-null   float64
 5   views             1180 non-null   float64
 6   category          1180 non-null   object
 7   channel_id        1180 non-null   object
 8   category_encoded  1180 non-null   int32
dtypes: float64(2), int32(1), object(6)
memory usage: 87.6+ KB
```

[29]:
```
# step 4 : Feature Engineering
# new features
dataset['title_length'] = dataset['title'].apply(len)
dataset['description_length'] = dataset['description'].apply(len)

# show new features
dataset[['title', 'title_length', 'description', 'description_length']].head()
```

[29]:

	title	title_length	description	description_length
0	Joe Biden congratulates Donald Trump as pair s...	80	No description available	24
1	The Queen vs The President: "Boom" - BBC News	45	The Queen and Prince Harry have responded to B...	379
2	Science rejuvenates woman's skin cells to 30 y...	69	Researchers have rejuvenated a 53-year-old wom...	480
3	World's largest coral found in Pacific Ocean |...	55	The largest coral ever recorded has been found...	520
4	Dinosaur fossil from asteroid strike that caus...	94	A dinosaur's leg, stunningly preserved, has be...	680

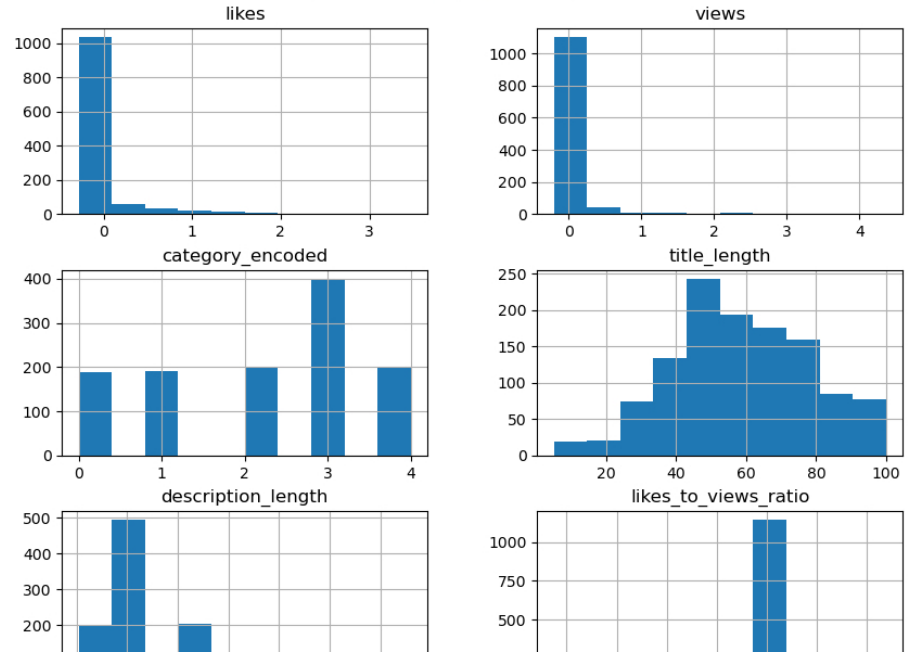
[30]:
```
# Calculate the likes to views ratio feature
dataset['likes_to_views_ratio'] = dataset['likes'] / dataset['views']

# Display the new feature and some other columns to verify
dataset[['title', 'likes', 'views', 'likes_to_views_ratio']].head()
```

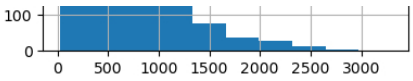
[30]:

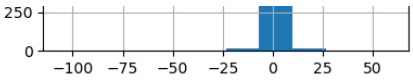
	title	likes	views	likes_to_views_ratio
0	Joe Biden congratulates Donald Trump as pair s...	-0.262402	-0.204526	1.282975
1	The Queen vs The President: "Boom" - BBC News	0.408880	-0.085689	-4.771676
2	Science rejuvenates woman's skin cells to 30 y...	0.002237	-0.175809	-0.012723
3	World's largest coral found in Pacific Ocean |...	-0.264987	-0.201281	1.316501
4	Dinosaur fossil from asteroid strike that caus...	0.193043	-0.142722	-1.352586

[31]:
```
# Generate histograms for each column
dataset.hist(figsize=(10, 8))
```

[31]:
```
array([[<Axes: title={'center': 'likes'}>,
        <Axes: title={'center': 'views'}>],
       [<Axes: title={'center': 'category_encoded'}>,
        <Axes: title={'center': 'title_length'}>],
       [<Axes: title={'center': 'description_length'}>,
        <Axes: title={'center': 'likes_to_views_ratio'}>]], dtype=object)
```




100
0
0
500
1000
1500
2000
2500
3000


250
0
−100
−75
−50
−25
0
25
50

[ ]: