movie

February 24, 2024

```
[1]: import pandas as pd
```

```
[2]: import numpy as np
```

```
[3]: pip install --upgrade pandas
```

```
Requirement already satisfied: pandas in /usr/local/lib/python3.10/dist-packages
(2.2.1)
Requirement already satisfied: numpy<2,>=1.22.4 in
/usr/local/lib/python3.10/dist-packages (from pandas) (1.25.2)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.10/dist-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.10/dist-
packages (from pandas) (2023.4)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.10/dist-
packages (from pandas) (2024.1)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.10/dist-
packages (from python-dateutil>=2.8.2->pandas) (1.16.0)
```

```
[4]: pd.__version__
```

```
[4]: '2.2.1'
```

```
[5]: train_path = "train_data.txt"
```

```
[6]: train_data = pd.read_csv("train_data.txt", header=None, sep=":::", names=["ID",␣
      ↪"Title", "Genres","Description"], engine='python')
```

```
[7]: train_data
```

```
[7]:           ID                               Title        Genres  \
     0          1      Oscar et la dame rose (2009)         drama
     1          2                      Cupid (1997)      thriller
     2          3  Young, Wild and Wonderful (1980)         adult
     3          4             The Secret Sin (1915)         drama
     4          5            The Unrecovered (2007)         drama
     …        …                                 …          …
     11175  11176        Gokudô no onna-tachi (1986)         crime
```

```
11176  11177      Les rendez-vous d'Anna (1978)          drama
11177  11178            Chipurile deltei (2006)    documentary
11178  11179                 La pisseuse (1997)          short
11179  11180         Legion of the Black (2012)             NaN

                                             Description
0       Listening in to a conversation between his do…
1       A brother and sister with a past incestuous r…
2       As the bus empties the students for their fie…
3       To help their unemployed father make ends mee…
4       The film's title refers not only to the un-re…
…                                                  …
11175   While her husband is in prison doing time, Ta…
11176   Anna, a detached and diffident director, arri…
11177   The Danube Delta. The modern era slowly creep…
11178   This is the great day for her: she has an imp…
11179                                               None

[11180 rows x 4 columns]
```

[8]:
```
train_data.head(2)
```

[8]:
```
   ID                          Title      Genres  \
0   1   Oscar et la dame rose (2009)       drama
1   2                  Cupid (1997)    thriller

                                         Description
0   Listening in to a conversation between his do…
1   A brother and sister with a past incestuous r…
```

[9]:
```
train_data.duplicated().sum()
```

[9]:
```
0
```

[10]:
```
train_data.shape
```

[10]:
```
(11180, 4)
```

[11]:
```
train_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11180 entries, 0 to 11179
Data columns (total 4 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   ID           11180 non-null  int64
 1   Title        11180 non-null  object
 2   Genres       11179 non-null  object
 3   Description  11179 non-null  object
dtypes: int64(1), object(3)
memory usage: 349.5+ KB
```

[12]:

```
train_data.describe
```

[12]:

```
<bound method NDFrame.describe of           ID
Title         Genres  \
0          1       Oscar et la dame rose (2009)          drama
1          2                       Cupid (1997)       thriller
2          3   Young, Wild and Wonderful (1980)          adult
3          4              The Secret Sin (1915)          drama
4          5             The Unrecovered (2007)          drama
…      …                                 …            …
11175  11176        Gokudô no onna-tachi (1986)          crime
11176  11177      Les rendez-vous d'Anna (1978)          drama
11177  11178            Chipurile deltei (2006)    documentary
11178  11179                 La pisseuse (1997)          short
11179  11180         Legion of the Black (2012)             NaN

                                             Description
0       Listening in to a conversation between his do…
1       A brother and sister with a past incestuous r…
2       As the bus empties the students for their fie…
3       To help their unemployed father make ends mee…
4       The film's title refers not only to the un-re…
…                                                  …
11175   While her husband is in prison doing time, Ta…
11176   Anna, a detached and diffident director, arri…
11177   The Danube Delta. The modern era slowly creep…
11178   This is the great day for her: she has an imp…
11179                                               None

[11180 rows x 4 columns]>
```

[13]:

```
train_data.describe()
```

[13]:

```
                 ID
count  11180.000000
mean    5590.500000
std     3227.532339
min        1.000000
25%     2795.750000
50%     5590.500000
75%     8385.250000
max    11180.000000
```

[14]:

```python
test_path = "test_data.txt"
```

[15]:

```python
test_data = pd.read_csv("test_data.txt", header=None, sep=":::", names=["ID",␣
 ↪"Title", "Genres","Description"], engine='python')
```

[16]:

```python
test_data
```

[16]:

```
          ID                                        Title  \
0          1                        Edgar's Lunch (1998)
1          2                    La guerra de papá (1977)
2          3                 Off the Beaten Track (2010)
3          4                      Meu Amigo Hindu (2015)
4          5                           Er nu zhai (1955)
…      …                                          …
11516  11517                        La extranjera (2008)
11517  11518                          Molly Crows (2013)
11518  11519   How to Survive a Zombie Apocalypse (2015)
11519  11520                              Respect (1994)
11520  11521                                           "S

                                                  Genres  Description
0       L.R. Brane loves his life - his car, his apar…          NaN
1       Spain, March 1964: Quico is a very naughty ch…          NaN
2       One year in the life of Albin and his family …          NaN
3       His father has died, he hasn't spoken with hi…          NaN
4       Before he was known internationally as a mart…          NaN
…                                                  …          …
11516   It is Maria's history : a woman of character …          NaN
11517   When seven year old Jess and her alcoholic Mo…          NaN
11518   The latest technology 5G is infecting mankind…          NaN
11519   Young journalist has an interview with a succ…          NaN
11520                                               None          NaN

[11521 rows x 4 columns]
```

[17]:

```python
test_data.head(2)
```

[17]:

```
   ID                       Title  \
0   1       Edgar's Lunch (1998)
1   2   La guerra de papá (1977)

                                              Genres  Description
0   L.R. Brane loves his life - his car, his apar…          NaN
1   Spain, March 1964: Quico is a very naughty ch…          NaN
```

[18]:

```python
test_data.duplicated().sum()
```

[18]:
```
0
```

[19]:
```
test_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11521 entries, 0 to 11520
Data columns (total 4 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   ID           11521 non-null  int64
 1   Title        11521 non-null  object
 2   Genres       11520 non-null  object
 3   Description  0 non-null      float64
dtypes: float64(1), int64(1), object(2)
memory usage: 360.2+ KB
```

[20]:
```
test_data.describe()
```

[20]:
```
                ID  Description
count  11521.00000          0.0
mean    5761.00000          NaN
std     3325.97056          NaN
min        1.00000          NaN
25%     2881.00000          NaN
50%     5761.00000          NaN
75%     8641.00000          NaN
max    11521.00000          NaN
```

[21]:
```
test_data.shape
```

[21]:
```
(11521, 4)
```

[22]:
```
import matplotlib.pyplot as plt
import seaborn as sns
import re
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
```

[38]:
```
def clean_description(text):
    text = re.sub(r'[^\w\s]', '', text)
    text = re.sub(r'\s+', ' ', text)
    text = re.sub(r"\s+", " ", text).strip()

    return text
```

[34]:

```
train_data['Clean_Description'] = train_data['Description'].
  ↪apply(clean_description)
test_data['Clean_Description'] = test_data['Description'].astype(str).
  ↪apply(clean_description)
```

[33]:

```
palette = sns.color_palette("pastel")

plt.figure(figsize=(12, 15))
sns.countplot(data=train_data, y="Genres", order=train_data["Genres"].
  ↪value_counts().index, palette=palette)
plt.xlabel('Genre', fontsize=12)
plt.ylabel('Count', fontsize=12)
plt.xticks(fontsize=10)
```

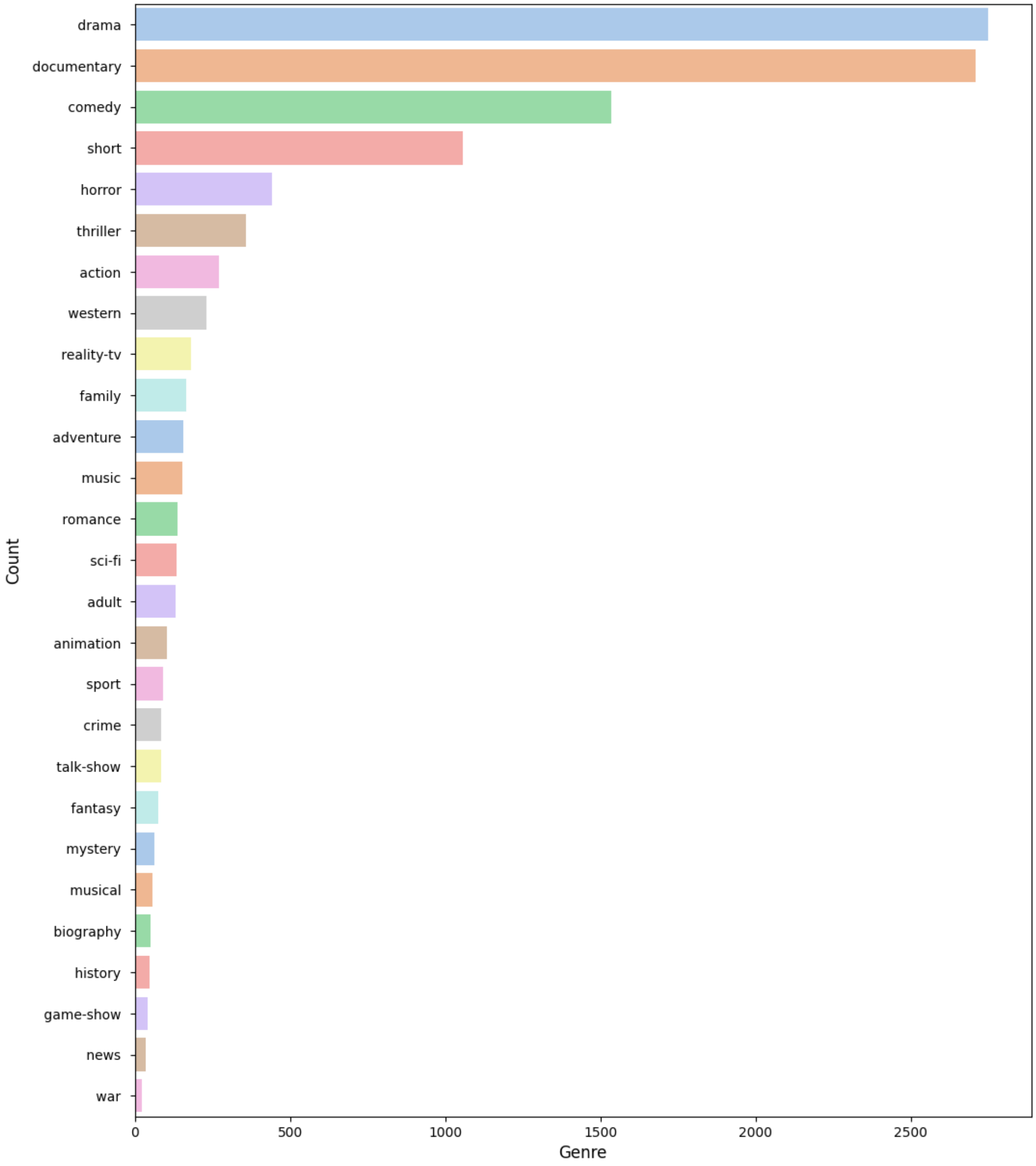
```
<ipython-input-33-827a2c795626>:4: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in
v0.14.0. Assign the `y` variable to `hue` and set `legend=False` for the same
effect.

  sns.countplot(data=train_data, y="Genres",
order=train_data["Genres"].value_counts().index, palette=palette)
<ipython-input-33-827a2c795626>:4: UserWarning:
The palette list has fewer values (10) than needed (27) and will cycle, which
may produce an uninterpretable plot.
  sns.countplot(data=train_data, y="Genres",
order=train_data["Genres"].value_counts().index, palette=palette)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
```

```
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
```

```
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
```

```
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
/usr/local/lib/python3.10/dist-packages/seaborn/_base.py:949: FutureWarning:
When grouping with a length-1 list-like, you will need to pass a length-1 tuple
to get_group in a future version of pandas. Pass `(name,)` instead of `name` to
silence this warning.
  data_subset = grouped_data.get_group(pd_key)
```

```
[33]: (array([   0.,  500., 1000., 1500., 2000., 2500., 3000.]),
       [Text(0.0, 0, '0'),
        Text(500.0, 0, '500'),
        Text(1000.0, 0, '1000'),
        Text(1500.0, 0, '1500'),
        Text(2000.0, 0, '2000'),
        Text(2500.0, 0, '2500'),
        Text(3000.0, 0, '3000')])
```



[41]:
```
train_data['Original_Length'] = train_data['Description'].apply(len)
train_data['Cleaned_Length'] = train_data['Clean_Description'].apply(len)
```

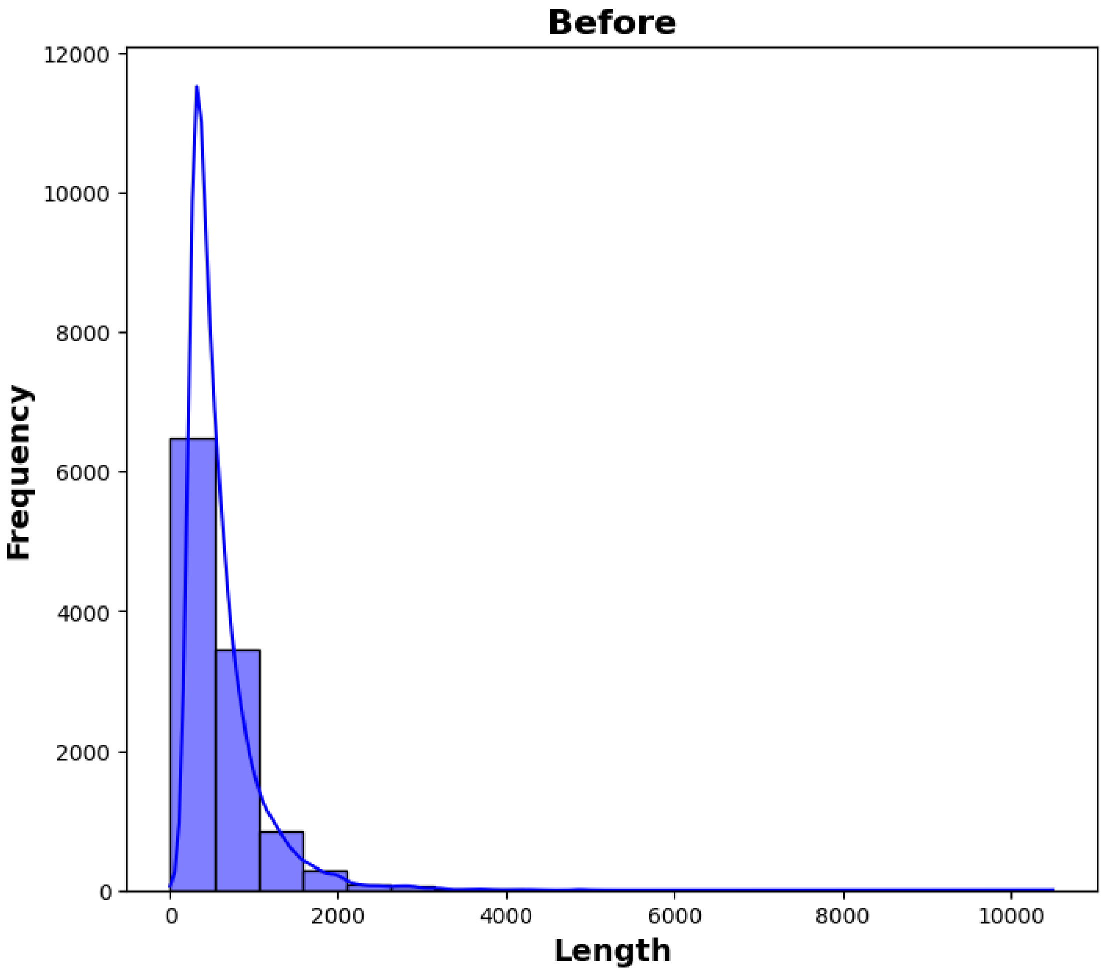
[42]:
```
train_data.sample(5)
```

[42]:
```
        ID                              Title         Genres  \
545    546        Will of the Warrior (2013)        history
3789  3790    King of the Wild Horses (1947)        western
426    427           His Will Be Done (2009)         horror
```

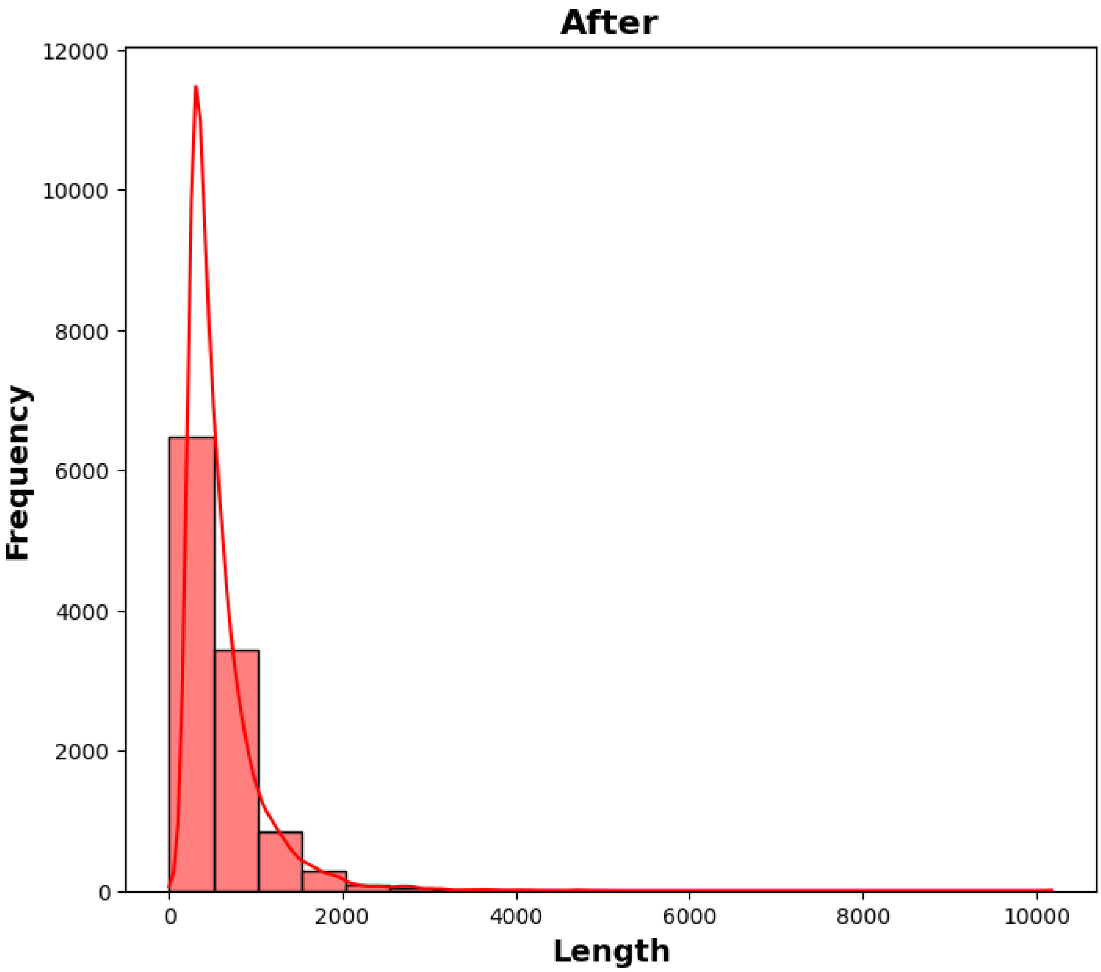
```
9279  9280                      Yugo (2009)          short
6813  6814                   Bully (2011/I)    documentary

                                            Description  \
545    This behind-the-scenes documentary focuses on…
3789   An orphan goes to live with his uncle and cou…
426    Powerful witch Morgan heads to a festival at …
9279   A Serb profiteer, driving supplies to the sol…
6813   This year, over 13 million American kids will…

                                      Clean_Description  Original_Length  \
545   This behindthescenes documentary focuses on Ma…              555
3789  An orphan goes to live with his uncle and cous…              268
426   Powerful witch Morgan heads to a festival at a…              419
9279  A Serb profiteer driving supplies to the soldi…              308
6813  This year over 13 million American kids will b…             2440

      Cleaned_Length
545              538
3789             260
426              411
9279             301
6813            2381
```

[45]:

```
plt.figure(figsize=(8, 7))
sns.histplot(data=train_data, x='Original_Length', bins=20, kde=True,␣
 ↪color='blue')
plt.xlabel('Length', fontsize=14, fontweight='bold')
plt.ylabel('Frequency', fontsize=14, fontweight='bold')
plt.title('Before', fontsize=16, fontweight='bold')
plt.show()
```

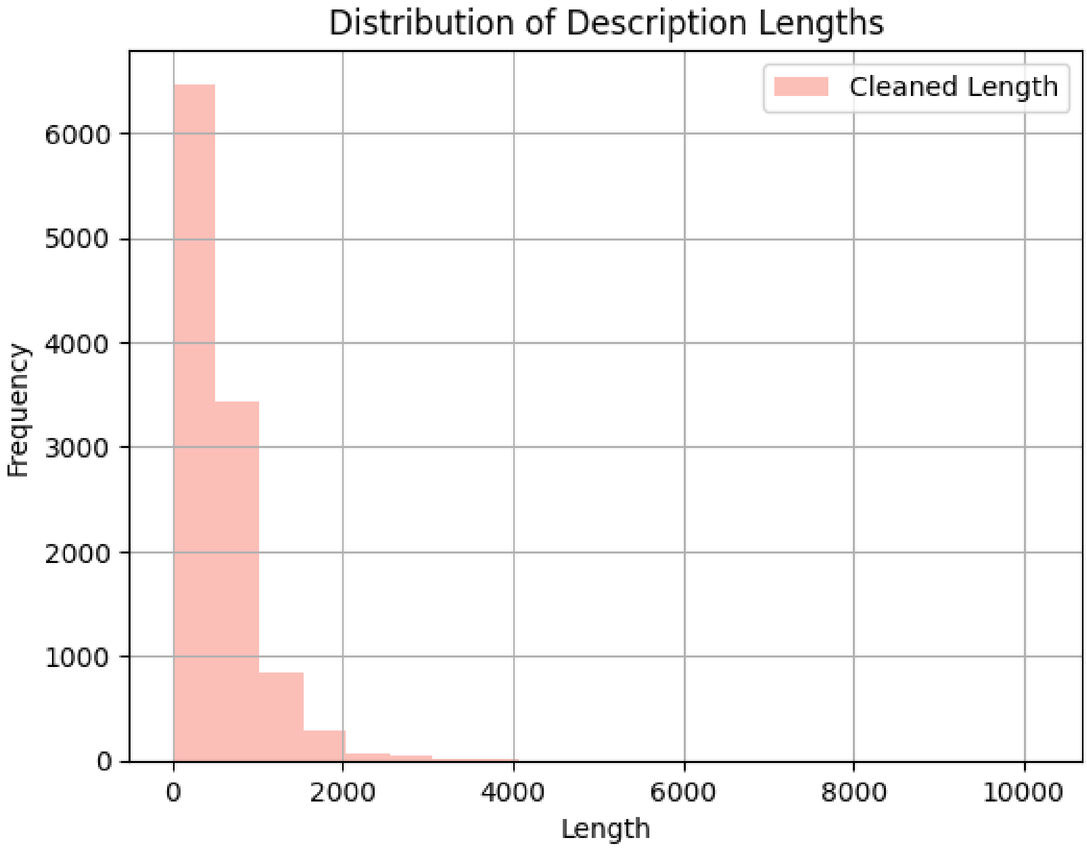


[46]:
```
plt.figure(figsize=(8, 7))
sns.histplot(data=train_data, x='Cleaned_Length', bins=20, kde=True,␣
 ↪color='red')
plt.xlabel('Length', fontsize=14, fontweight='bold')
plt.ylabel('Frequency', fontsize=14, fontweight='bold')
plt.title('After', fontsize=16, fontweight='bold')
plt.show()
```



[59]:

```
plt.hist(train_data['Cleaned_Length'], bins=20, color='salmon', alpha=0.5, label='Cleaned Length')

plt.title('Distribution of Description Lengths')
plt.xlabel('Length')
plt.ylabel('Frequency')
plt.legend()
plt.grid(True)
plt.show()


removed_characters = sum(train_data['Original_Length'] - train_data['Cleaned_Length'])
print("Total characters removed during cleaning:", removed_characters)
```



```
Total characters removed during cleaning: 192058
```

[60]:
```
X = train_data['Description']
y = train_data['Genres']
```

[61]:
```
X_train, X_val, y_train, y_val = train_test_split(X, y, test_size= 0.2,␣
↪random_state=123)
```

[62]:
```
vectorize = TfidfVectorizer()
```

[63]:
```
X_train_tfidf  = vectorize.fit_transform(X_train)
X_test_tfidf = vectorize.transform(test_data['Clean_Description'])
X_val_tfidf  = vectorize.transform(X_val)
```

[67]:
```
svm_classifier = SVC()
svm_classifier.fit(X_val_tfidf, y_val)
```

[67]:
```
SVC()
```

[68]:

```
y_pred_val = svm_classifier.predict(X_val_tfidf)
valAccuracy = accuracy_score(y_val, y_pred_val)
```

[69]:

```
print("Validation Accuracy:", valAccuracy)
```

```
Validation Accuracy: 0.8305008944543828
```

[ ]:

```
```