Extract and Load the Dataset

```
import zipfile
import os

# Define path and extract the zip file
zip_path = '/content/sample_data/archive (7).zip'
extract_dir = '/content/sample_data/genre_dataset'
with zipfile.ZipFile(zip_path, 'r') as zip_ref:
    zip_ref.extractall(extract_dir)

# View extracted files
os.listdir(extract_dir)
```

```
['Genre Classification Dataset']
```

Double-click (or enter) to edit

```
import pandas as pd

# Load and parse train_data.txt
train_path = os.path.join(extract_dir, 'Genre Classification Dataset/train_data.txt')
ids, titles, genres, descriptions = [], [], [], []
with open(train_path, 'r', encoding='utf-8') as f:
    for line in f:
        parts = line.strip().split(" ::: ")
        if len(parts) == 4:
            ids.append(parts[0])
            titles.append(parts[1])
            genres.append(parts[2].lower().split(', '))  # handles multiple genres
            descriptions.append(parts[3])

# Create DataFrame
df = pd.DataFrame({
    'id': ids,
    'title': titles,
    'genres': genres,
    'description': descriptions
})
df.head()
```

1 to 5 of 5 entries Filter

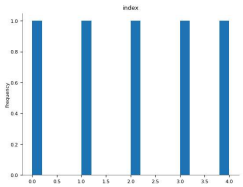
index	id	title	genres	description
0	1	Oscar et la dame rose (2009)	drama	Listening in to a conversation between his doctor and parents, 10-year-old Oscar learns what nobody has the courage to tell him. He only has a few weeks to live. Furious, he refuses to speak to anyone except straight-talking Rose, the lady in pink he meets on the hospital stairs. As Christmas approaches, Rose uses her fantastical experiences as a professional wrestler, her imagination, wit and charm to allow Oscar to live life and love to the full, in the company of his friends Pop Corn, Einstein, Bacon and childhood sweetheart Peggy Blue.
1	2	Cupid (1997)	thriller	A brother and sister with a past incestuous relationship have a current murderous relationship. He murders the women who reject him and she murders the women who get too close to him.
2	3	Young, Wild and Wonderful (1980)	adult	As the bus empties the students for their field trip to the Museum of Natural History, little does the tour guide suspect that the students are there for more than just another tour. First, during the lecture films, the coeds drift into dreams of the most erotic fantasies one can imagine. After the films, they release the emotion of the fantasies in the most erotic and uncommon ways. One slips off to the curator's office for a little "acquisition. " Another finds the anthropologist to see what bones can be identified. Even the head teacher isn't immune. Soon the tour is over, but as the bus departs, everyone admits it was quite an education.
3	4	The Secret Sin (1915)	drama	To help their unemployed father make ends meet, Edith and her twin sister Grace work as seamstresses . An invalid, Grace falls prey to the temptations of Chinatown opium and becomes an addict, a condition worsened by a misguided physician who prescribes morphine to ease her pain. When their father strikes oil, the family enjoys a new prosperity and the sisters meet the eligible Jack Herron, a fellow oil prospector. To Grace's shock, Jack falls in love with Edith and in her jealousy, Grace tells Jack that Edith, not she, has a drug problem. Hinting that her sister will soon need more morphine, Grace arranges for a dinner in Chinatown with the couple. While her sister and Jack dance, Grace slips away to an opium den. Edith follows her, but ends up in the wrong den and is arrested in an ensuing drug raid. After he bails her out of jail, Edith takes an angry Jack to search for Grace and stumbles across her half-conscious body lying in the street. The truth about the sisters is revealed, and after sending Grace to a sanitarium in the country, Jack and Edith are married.
4	5	The Unrecovered (2007)	drama	The film's title refers not only to the un-recovered bodies at ground zero, but also to the state of the nation at large. Set in the hallucinatory period of time between September 11 and Halloween of 2001, The Unrecovered examines the effect of terror on the average mind, the way a state of heightened anxiety and/or alertness can cause the average person to make the sort of imaginative connections that are normally made only by artists and conspiracy theorists-both of whom figure prominently in this film. The Unrecovered explores the way in which irony, empathy, and paranoia relate to one another in the wake of 9/11.

Show 100 per page

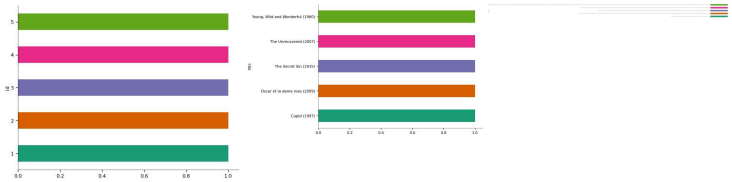
Like what you see? Visit the data table notebook to learn more about interactive tables.

```
WARNING:root:Quickchart encountered unexpected dtypes in columns: "(['genres'],)"
```

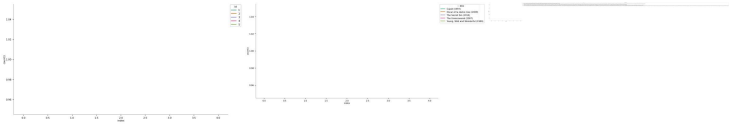
Distributions



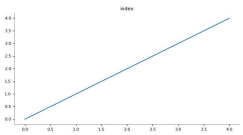
Categorical distributions



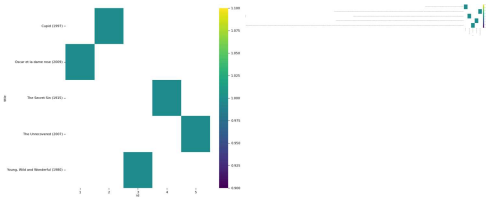
Time series



Values



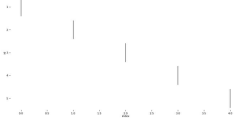
2-d categorical distributions



Faceted distributions

```
<string>:5: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `le
```



```
<string>:5: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `y` variable to `hue` and set `le
```

Next steps: Generate code with df | View recommended plots | New interactive sheet

Text Cleaning

```
import re
import nltk
from nltk.corpus import stopwords
from nltk.stem import SnowballStemmer
nltk.download('stopwords')
stop_words = set(stopwords.words('english'))
stemmer = SnowballStemmer("english")
def clean_text(text):
    text = text.lower()
    text = re.sub(r'\W+', ' ', text)  # remove punctuation
    text = re.sub(r'\d+', '', text)   # remove digits
    words = text.split()
    words = [stemmer.stem(word) for word in words if word not in stop_words]
    return ' '.join(words)
df['clean_description'] = df['description'].apply(clean_text)
df[['description', 'clean_description']].head()
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

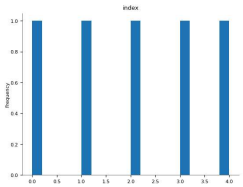
1 to 5 of 5 entries Filter

index	description	clean_description
0	Listening in to a conversation between his doctor and parents, 10-year-old Oscar learns what nobody has the courage to tell him. He only has a few weeks to live. Furious, he refuses to speak to anyone except straight-talking Rose, the lady in pink he meets on the hospital stairs. As Christmas approaches, Rose uses her fantastical experiences as a professional wrestler, her imagination, wit and charm to allow Oscar to live life and love to the full, in the company of his friends Pop Corn, Einstein, Bacon and childhood sweetheart Peggy Blue.	listen convers doctor parent year old oscar learn nobodi courag tell week live furious refus speak anyon except straight talk rose ladi pink meet hospit stair christma approach rose use fantast experi profession wrestler imagin wit charm allow oscar live life love full compani friend pop corn einstein bacon childhood sweetheart peggi blue
1	A brother and sister with a past incestuous relationship have a current murderous relationship. He murders the women who reject him and she murders the women who get too close to him.	brother sister past incestu relationship current murder relationship murder women reject murder women get close
2	As the bus empties the students for their field trip to the Museum of Natural History, little does the tour guide suspect that the students are there for more than just another tour. First, during the lecture films, the coeds drift into dreams of the most erotic fantasies one can imagine. After the films, they release the emotion of the fantasies in the most erotic and uncommon ways. One slips off to the curator's office for a little "acquisition. " Another finds the anthropologist to see what bones can be identified. Even the head teacher isn't immune. Soon the tour is over, but as the bus departs, everyone admits it was quite an education.	bus empti student field trip museum natur histori littl tour guid suspect student anoth tour first lectur film co drift dream erot fantasi one imagin film releas emot fantasi erot uncommon way one slip curat offic littl acquisit anoth find anthropologist see bone identifi even head teacher immun soon tour bus depart everyon admit quit educ
3	To help their unemployed father make ends meet, Edith and her twin sister Grace work as seamstresses . An invalid, Grace falls prey to the temptations of Chinatown opium and becomes an addict, a condition worsened by a misguided physician who prescribes morphine to ease her pain. When their father strikes oil, the family enjoys a new prosperity and the sisters meet the eligible Jack Herron, a fellow oil prospector. To Grace's shock, Jack falls in love with Edith and in her jealousy, Grace tells Jack that Edith, not she, has a drug problem. Hinting that her sister will soon need more morphine, Grace arranges for a dinner in Chinatown with the couple. While her sister and Jack dance, Grace slips away to an opium den. Edith follows her, but ends up in the wrong den and is arrested in an ensuing drug raid. After he bails her out of jail, Edith takes an angry Jack to search for Grace and stumbles across her half-conscious body lying in the street. The truth about the sisters is revealed, and after sending Grace to a sanitarium in the country, Jack and Edith are married.	help unemploy father make end meet edith twin sister grace work seamstress invalid grace fall prey temptat chinatown opium becom addict condit worsen misguid physician prescrib morphin eas pain father strike oil famili enjoy new prosper sister meet elig jack herron fellow oil prospector grace shock jack fall love edith jealousi grace tell jack edith drug problem hint sister soon need morphin grace arrang dinner chinatown coupl sister jack danc grace slip away opium den edith follow end wrong den arrest ensu drug raid bail jail edith take angri jack search grace stumbl across half conscious bodi lie street truth sister reveal send grace sanitarium countri jack edith marri
4	The film's title refers not only to the un-recovered bodies at ground zero, but also to the state of the nation at large. Set in the hallucinatory period of time between September 11 and Halloween of 2001, The Unrecovered examines the effect of terror on the average mind, the way a state of heightened anxiety and/or alertness can cause the average person to make the sort of imaginative connections that are normally made only by artists and conspiracy theorists-both of whom figure prominently in this film. The Unrecovered explores the way in which irony, empathy, and paranoia relate to one another in the wake of 9/11.	film titl refer un recov bodi ground zero also state nation larg set hallucinatori period time septemb halloween unrecov examin effect terror averag mind way state heighten anxieti alert caus averag person make sort imagin connect normal made artist conspiraci theorist figur promin film unrecov explor way ironi empathi paranoia relat one anoth wake

Show 100 per page

Like what you see? Visit the data table notebook to learn more about interactive tables.

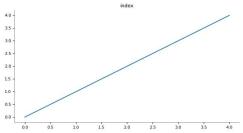
Distributions



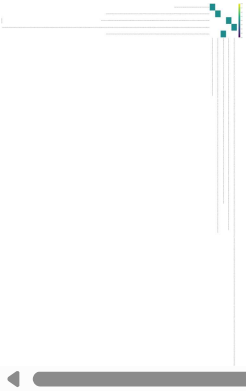
Categorical distributions

Time series

Values



2-d categorical distributions



TF-IDF Vectorization

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer(max_features=5000)
X = vectorizer.fit_transform(df['clean_description'])
```

Encode Genre Labels with MultiLabelBinarizer

```
from sklearn.preprocessing import MultiLabelBinarizer
mlb = MultiLabelBinarizer()
y = mlb.fit_transform(df['genres'])
print("Genres:", mlb.classes_)
print("Shape of X:", X.shape)
print("Shape of y:", y.shape)
```

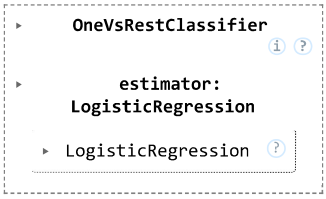
```
Genres: ['action' 'adult' 'adventure' 'animation' 'biography' 'comedy' 'crime'
 'documentary' 'drama' 'family' 'fantasy' 'game-show' 'history' 'horror'
 'music' 'musical' 'mystery' 'news' 'reality-tv' 'romance' 'sci-fi'
 'short' 'sport' 'talk-show' 'thriller' 'war' 'western']
Shape of X: (54214, 5000)
Shape of y: (54214, 27)
```

Train-Test Split

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Train Classifiers with OneVsRest

```
from sklearn.linear_model import LogisticRegression
from sklearn.multiclass import OneVsRestClassifier
logreg = OneVsRestClassifier(LogisticRegression(max_iter=1000))
logreg.fit(X_train, y_train)
```



Evaluation

```
from sklearn.metrics import classification_report, accuracy_score, hamming_loss
y_pred = logreg.predict(X_test)
print("Classification Report:\n", classification_report(y_test, y_pred, target_names=mlb.classes_))
print("Hamming Loss:", hamming_loss(y_test, y_pred))
```

```
Classification Report:
                precision    recall  f1-score   support

        action       0.83      0.08      0.14       263
         adult       0.73      0.07      0.13       112
     adventure       0.44      0.05      0.09       139
     animation       0.00      0.00      0.00       104
     biography       0.00      0.00      0.00        61
        comedy       0.76      0.30      0.43      1443
         crime       0.67      0.02      0.04       107
   documentary       0.81      0.68      0.74      2659
         drama       0.69      0.46      0.55      2697
        family       1.00      0.03      0.06       150
       fantasy       0.00      0.00      0.00        74
     game-show       1.00      0.23      0.37        40
       history       0.00      0.00      0.00        45
        horror       0.82      0.34      0.48       431
         music       0.64      0.24      0.35       144
       musical       0.00      0.00      0.00        50
       mystery       0.00      0.00      0.00        56
          news       0.00      0.00      0.00        34
    reality-tv       0.75      0.03      0.06       192
       romance       0.00      0.00      0.00       151
        sci-fi       0.67      0.06      0.10       143
         short       0.73      0.11      0.19      1045
         sport       0.69      0.10      0.17        93
     talk-show       0.75      0.04      0.07        81
      thriller       0.54      0.04      0.08       309
           war       0.00      0.00      0.00        20
       western       0.98      0.56      0.71       200
```

```
   micro avg       0.76      0.37      0.49     10843
   macro avg       0.50      0.13      0.18     10843
weighted avg       0.71      0.37      0.45     10843
 samples avg       0.36      0.37      0.36     10843

Hamming Loss: 0.027787171105440957
/usr/local/lib/python3.11/dist-packages/sklearn/metrics/_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined ar
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.11/dist-packages/sklearn/metrics/_classification.py:1565: UndefinedMetricWarning: Precision is ill-defined ar
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

```
from sklearn.ensemble import RandomForestClassifier
rf = OneVsRestClassifier(RandomForestClassifier(n_estimators=100))
rf.fit(X_train, y_train)
print("RF Accuracy:", rf.score(X_test, y_test))
```

```
RF Accuracy: 0.20308032832241998
```

Feature Importance (for Logistic Regression)

```
import numpy as np

# Feature importance for each genre
for i, genre in enumerate(mlb.classes_):
    top10 = np.argsort(logreg.estimators_[i].coef_[0])[-10:]
    print(f"\nTop features for genre '{genre}':")
    print([vectorizer.get_feature_names_out()[j] for j in top10])
```

```
Top features for genre 'action':
['skill', 'hero', 'kill', 'reveng', 'gangster', 'singh', 'fight', 'martial', 'assassin', 'action']

Top features for genre 'adult':
['sexi', 'fantasi', 'seduc', 'orgi', 'bound', 'scene', 'hot', 'sexual', 'gag', 'sex']

Top features for genre 'adventure':
['tie', 'ship', 'bound', 'kidnap', 'rescu', 'buxom', 'jungl', 'gag', 'bondag', 'adventur']
```