Project Task: Primary analysis on text data

Dataisgood ¶

Sumit Sharma

Batch :-1204

The Company X is the world's largest positive and solution based content driven impact platform. They use their website to share the contents around the world in the form of articles and other forms of media. The problem we are facing is to segregate or automate the process of classification of these articles into separate categories using AI for better reach to the right audience. We are using deep learning for this purpose.

Dataset Description -

The columns/features in the given dataset are as follows: ● Article category- Type of news article (Target variable) ● News Headline- Headline of the news article ● Author of the news article ● Brief about what the headline is about ● Date of publishing of article

What is EDA?

Exploratory Data Analysis (EDA) for string or categorical data is essential for understanding the distribution, patterns, and characteristics of these types of variables. While EDA for numerical data often involves summary statistics and visualizations like histograms, EDA for string data focuses on different techniques to gain insights into categorical variables.

In [97]:

```
# importing some liberary for doing work
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib as plt
import plotly.express as px
```

Steps which all time remain same

In [99]:

```
df=pd.read_csv("D:/sql assientment make by sumit/news_syn.csv")
```

In [100]:

```
df.head()
```

Out[100]:

	Article category	News Headline	Author	news_link	short_description	date
0	WEDDINGS	Real Weddings: Couples Who Got Married This We...	NaN	_RARE_	If there's ever a time you need a little distr...	29/07/13
1	WELLNESS	The Moment I Knew	MeiMei Fox, Contributor\nNYTimes bestselling a...	_RARE_	NaN	02/05/13
2	POLITICS	Sunday Roundup	Arianna Huffington, Contributor	_RARE_	NaN	17/01/15
3	PARENTS	Funniest Parenting Tweets: What Moms And Dads ...	Hollis Miller	_RARE_	Kids may say the darndest things, but parents ...	15/07/16
4	BUSINESS	WATCH: 60 Seconds of Social Media	Shawn Amos, Contributor\nblues preacher | cont...	_RARE_	So, you think you're a real fashionista, hmm? ...	18/08/12

In [101]:

```
df.tail()
```

Out[101]:

	Article category	News Headline	Author	news_link	short_description	date
9995	ENTERTAINMENT	Real Weddings: Couples Who Got Married This We...	NaN	_RARE_	NaN	04/04/15
9996	FOOD & DRINK	Parenthesis: The Best Of The Mom And Dad Blogs...	Emma Mustich	_RARE_	To receive the eBay Roundup of Vintage Home Fi...	01/05/12
9997	THE WORLDPOST	Sunday Roundup	Arianna Huffington, Contributor	_RARE_	NaN	23/11/14
9998	TECH	Watch The Top 9 YouTube Videos Of The Week	NaN	_RARE_	If you're looking to see the most popular YouT...	19/11/12
9999	POLITICS	Sunday Roundup	Arianna Huffington, Contributor	_RARE_	NaN	14/06/15

In [13]:

```
df.describe()
```

Out[13]:

	Article category	News Headline	Author	news_link	short_description	date
count	10000	9087	7282	10000	3874	10000
unique	41	34	267	1	46	1813
top	POLITICS	Sunday Roundup	Arianna Huffington, Contributor	_RARE_	Kids may say the darndest things, but parents ...	28/06/13
freq	1611	2075	2120	10000	1018	51

In [9]:

```
df.shape
```

Out[9]:

```
(10000, 6)
```

In [15]:

```
df.columns
```

Out[15]:

```
Index(['Article category', 'News Headline', 'Author', 'news_link',
       'short_description', 'date'],
      dtype='object')
```

In [18]:

```
df.nunique()
```

Out[18]:

```
Article category       41
News Headline          34
Author                267
news_link               1
short_description      46
date                 1813
dtype: int64
```

In [20]:

```
df['Article category'].unique()
```

Out[20]:

```
array(['WEDDINGS', 'WELLNESS', 'POLITICS', 'PARENTS', 'BUSINESS',
       'FOOD & DRINK', 'TRAVEL', 'COMEDY', 'IMPACT', 'HOME & LIVING',
       'BLACK VOICES', 'PARENTING', 'QUEER VOICES', 'ENVIRONMENT',
       'WEIRD NEWS', 'WORLDPOST', 'HEALTHY LIVING', 'SPORTS',
       'ENTERTAINMENT', 'CRIME', 'COLLEGE', 'MEDIA', 'THE WORLDPOST',
       'WOMEN', 'ARTS & CULTURE', 'GOOD NEWS', 'STYLE & BEAUTY', 'STYLE',
       'CULTURE & ARTS', 'DIVORCE', 'RELIGION', 'MONEY', 'SCIENCE',
       'GREEN', 'TASTE', 'WORLD NEWS', 'TECH', 'ARTS', 'LATINO VOICES',
       'EDUCATION', 'FIFTY'], dtype=object)
```

In [21]:

```
df['News Headline'].unique()
```

Out[21]:

```
array(['Real Weddings: Couples Who Got Married This Weekend',
       'The Moment I Knew', 'Sunday Roundup',
       'Funniest Parenting Tweets: What Moms And Dads Said On Twitter This
Week',
       'WATCH: 60 Seconds of Social Media',
       'The Best Late Night Clips of the Week (VIDEO/PHOTOS)', nan,
       'Weekly Roundup of eBay Vintage Home Finds (PHOTOS)',
       'Real Weddings: Couples Who Got Married This Weekend (PHOTOS)',
       "This Week's Top 10 News Stories From Africa",
       'Best Parenting Tweets: What Moms And Dads Said On Twitter This Wee
k',
       'Parenthesis: The Best Of The Mom And Dad Blogs This Week',
       'Extreme Weather Photos Of The Week',
       'The Psychometer: Who Went Too Far Last Week?',
       'This Week In Pictures: Faith In Practice Around The World',
       'Animal Photos Of The Week',
       'Weekly Roundup of eBay Vintage Home Finds',
       'Days of Inspiration for the New Year!',
       'What To Watch On Hulu That’s New This Week',
       'The Funniest Tweets From Women This Week',
       'Weekly Roundup of eBay Vintage Clothing Finds',
       'Ikea Bag Dress Is Massive, In Need Of A Good Steam (PHOTOS)',
       'Extreme Weather Of The Week (PHOTOS)',
       "Fashionably Late Style Quiz: Test Your Knowledge Of This Week's Fa
shion News!",
       'The Funniest Someecards Of The Week (PICTURES)',
       "Wardrobe Malfunctions: See This Week's Almost-Dangerous Outfits (P
HOTOS)",
       'The Funniest Tweets From Parents This Week',
       'Watch The Top 9 YouTube Videos Of The Week',
       'What To Watch On Amazon Prime That’s New This Week',
       'The 20 Funniest Tweets From Women This Week',
       'Hot On Pinterest: 5 Pinners To Follow Now',
       'Weekly Roundup of eBay Vintage Clothing Finds (PHOTOS)',
       "Saturday's Morning Email: Funnies Edition",
       "Wardrobe Malfunctions Photos: See This Week's Almost-Dangerous Out
fits (PHOTOS)",
       'Fashionably Late: Style News You Might Have Missed This Week (PHOT
OS)'],
      dtype=object)
```

In [22]:

```
df['news_link'].unique()
```

Out[22]:

```
array(['_RARE_'], dtype=object)
```

In [23]:

```
df['short_description'].unique()
```

Out[23]:

```
array(["If there's ever a time you need a little distraction in your life,
it's during the divorce process. That's why we launched",
       nan,
       'Kids may say the darndest things, but parents tweet about them in
the funniest ways. So each week, we round up the most hilarious',
       "So, you think you're a real fashionista, hmm? Well, then step righ
t up and take HuffPost Style's Fashionably Late Style Quiz",
       'The stress and strain of constantly being connected can sometimes
take your life -- and your well-being -- off course. GPS',
       'Do you have a home story idea or tip? Email us at homesubmissions@
huffingtonpost.com. (PR pitches sent to this address will',
       "To receive the eBay Roundup of Vintage Home Finds via email, sign
up for Zuburbia's mailing list here. Your information will",
       "If you're looking to see the most popular YouTube videos of the we
ek, look no further. Once again, we're bringing you the",
       'PLEASE NOTE that Zuburbia does not endorse the use of fur, feather
s, leather or animal skins in home decor. Any of these',
       'Yikes!', '🔥🔥🔥',
       "Need a time out? You're in luck. In honor of Bliss' Triple Oxygen
Month, HuffPost's GPS for the Soul has teamed up with Bliss",
       'Yikes.',
       'No time to page through thousands of eBay listings? Then just snea
k a peek at my weekly eBay roundup of top vintage clothing',
       'The ladies of Twitter never fail to brighten our days with their b
rilliant — but succinct — wisdom. Each week, HuffPost Women',
       'Kids may say the darndest things, but parents tweet about them in
the funniest ways. So each week,\xa0we round up the most hilarious',
       'Which item is your favorite? Leave me a comment below to let me kn
ow and please take a minute to rate your favorite slides',
       'Who knew?',
       'No time to page through thousands of eBay listings? Then just snea
k a peek at my weekly eBay roundup of top vintage clothing finds.',
       'Ouch',
       "Welcome to Fashionably Late, where we round up the style scraps th
at didn't make it to our news page this week. Click through",
       'Do you have a home story idea or tip? Email us at homesubmissio
ns@huffingtonpost.com. (PR pitches sent to this address',
       'In our Beauty Street Style series, we find inspiring girls around
New York City and get the secrets to their look. This week',
       'The ladies of Twitter never fail to brighten our days with their b
rilliant -- but succinct -- wisdom. Each week, HuffPost',
       'Whoops.', 'Seriously?', 'Here we go again.',
       'This entry has expired',
       "We couldn't agree more. So HuffPost has joined with Laurie and eve
ry Friday afternoon, just in time for dinner, our editors",
       'Have something to say? Check out HuffPost Home on Twitter, Faceboo
k, Pinterest, Tumblr and Instagram. Do you have a home',
       'In her cookbook, The Family Dinner, Laurie David talks about the i
mportance of families making a ritual of sitting down to',
       'Wow.',
       'Want more HuffPost Style beauty content? Check us out on Twitter,
Facebook, Tumblr, Pinterest and Instagram at @HuffPostBeauty',
       'The 25 most profound “Shower Thoughts” on Reddit from the last wee
k.',
       'Want more? Be sure to check out HuffPost Style on Twitter, Faceboo
k, Tumblr, Pinterest and Instagram at @HuffPostStyle. The',
       'In our series FaceTime, we find inspiring girls around New York Ci
ty and get the secrets to their look. This week we met',
       'Want more? Be sure to check out HuffPost Style on Twitter, Faceboo
k, Tumblr, Pinterest and Instagram at @HuffPostStyle. PHOTOS',
       'Awww!',
```

In [29]:
```
df['date'].unique()
```

Out[29]:
```
array(['15/07/16', '18/08/12', '28/10/12', ..., '03/10/16', '16/09/15',
       '10/01/16'], dtype=object)
```

In [34]:
```
# cleaning the data
df.isnull().sum()
```

Out[34]:
```
Article category      0
News Headline       913
Author             2718
news_link             0
short_description   612
date                  0
dtype: int64
```

In [102]:
```
# drop row News Headline
df.dropna(inplace=True)
```

In [103]:
```
df.isnull().sum()
```

Out[103]:
```
Article category    0
News Headline       0
Author              0
news_link           0
short_description   0
date                0
dtype: int64
```

In [45]:
```
df.head()
```

Out[45]:

	Article category	News Headline	Author	news_link	short_description	date
3	PARENTS	Tweets: What Moms And Dads ...	Hollis Miller, Parenting Contributor	_RARE_	Kids may say the darndest things, but parents ...	15/07/16
4	BUSINESS	WATCH: 60 Seconds of Social Media	Shawn Amos, Contributor\nblues preacher | cont...	_RARE_	So, you think you're a real fashionista, hmm? ...	18/08/12
8	COMEDY	The Best Late Night Clips of the Week (VIDEO/P...	Matt Wilstein, Contributor\nEditor, Gotcha Med...	_RARE_	Do you have a home story idea or tip? Email us...	28/10/12
10	HOME & LIVING	Weekly Roundup of eBay Vintage Home Finds (PHO...	Mary Kincaid, Contributor\nFounder and Editor ...	_RARE_	To receive the eBay Roundup of Vintage Home Fi...	30/10/13
14	POLITICS	Sunday Roundup	Arianna Huffington, Contributor	_RARE_	To receive the eBay Roundup of Vintage Home Fi...	04/09/15

In [83]:
```
df.drop(columns=['news_link'],inplace=True)
```

In [84]:
```
df=df.groupby('News Headline').filter(lambda x:len(x)>20).reset_index(drop=True)
```

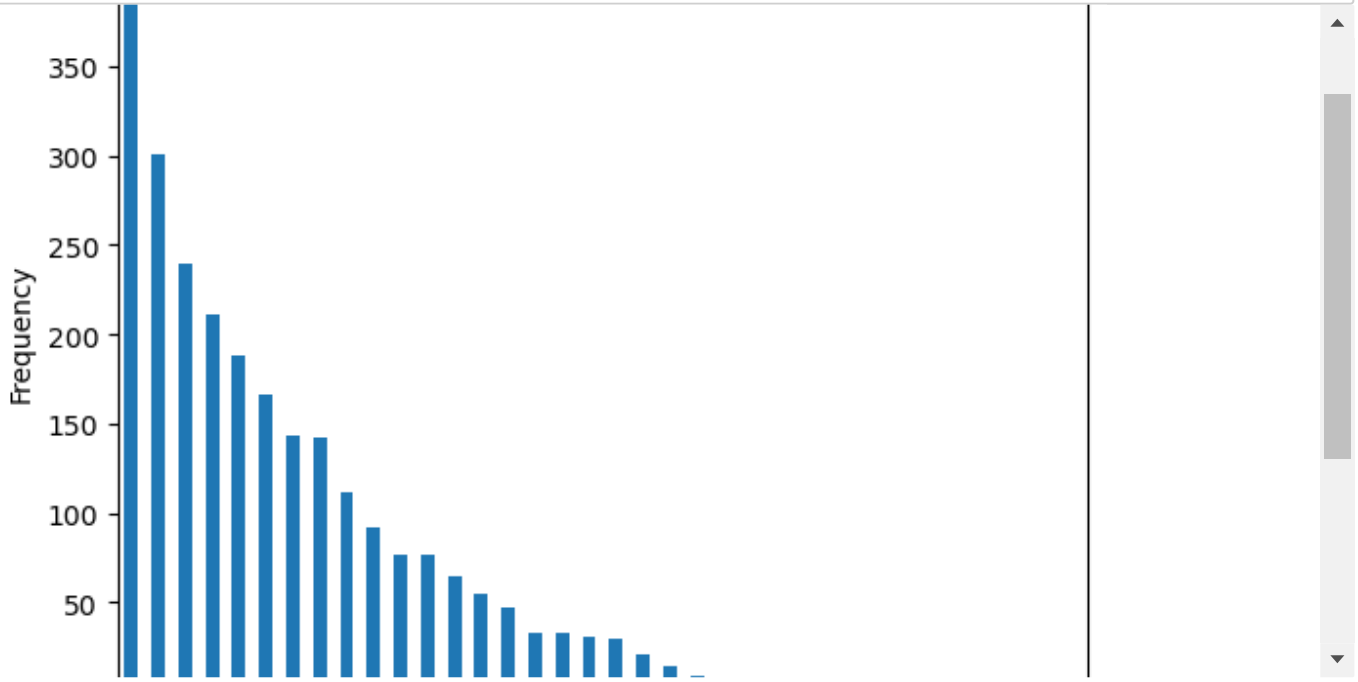
1.Frequency Distribution

Here i am going to visualize the distribuation of categories of data which do help me understand which categories are most commom

In [110]:

```
import pandas as pd
import matplotlib.pyplot as plt

# Plot the frequency distribution
df['Article category'].value_counts().plot(kind='bar')
plt.title('Frequency Distribution of Categories')
plt.xlabel('Article category')
plt.ylabel('Frequency')
plt.show()
```



In [111]:

```
df.head(1)
```

Out[111]:

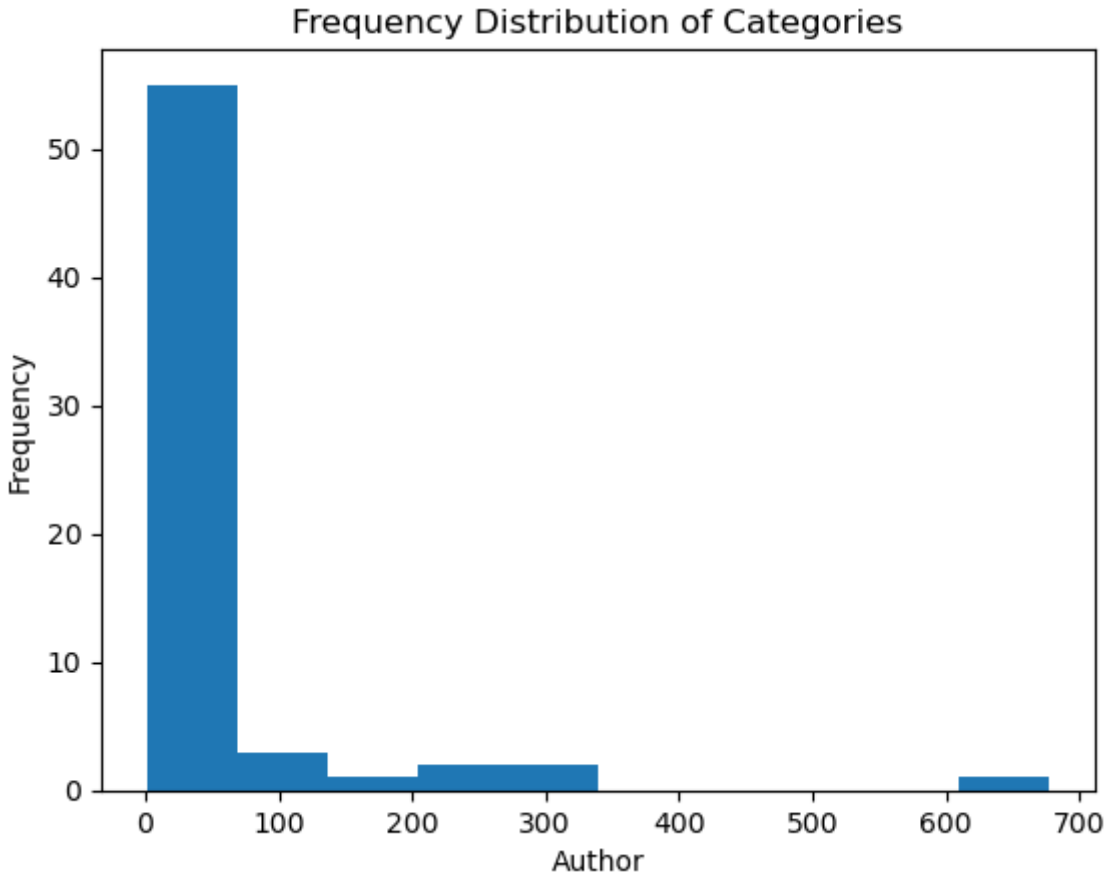
	Article category	News Headline	Author	news_link	short_description	date
3	PARENTS	Funniest Parenting Tweets: What Moms And Dads ...	Hollis Miller	_RARE_	Kids may say the darndest things, but parents ...	15/07/16

In [118]:

```
# Plot the frequency distribution
df['Author'].value_counts().plot(kind='hist')
plt.title('Frequency Distribution of Categories')
plt.xlabel('Author')
plt.ylabel('Frequency')
```

Out[118]:

```
Text(0, 0.5, 'Frequency')
```



2.Unique values:-

To see distinict categories present

In [88]:

```
unique_categories = df['Article category'].unique()
print("Unique Categories:", unique_categories)
```

```
Unique Categories: ['PARENTS' 'BUSINESS' 'COMEDY' 'HOME & LIVING' 'POLITIC
S' 'FOOD & DRINK'
 'WEIRD NEWS' 'ENTERTAINMENT' 'ARTS & CULTURE' 'TRAVEL' 'STYLE & BEAUTY'
 'STYLE' 'CULTURE & ARTS' 'SPORTS' 'QUEER VOICES' 'WOMEN' 'TASTE' 'GREEN'
 'HEALTHY LIVING' 'BLACK VOICES' 'TECH' 'PARENTING' 'CRIME' 'SCIENCE'
 'THE WORLDPOST' 'MEDIA' 'LATINO VOICES' 'DIVORCE' 'MONEY' 'GOOD NEWS'
 'WELLNESS' 'EDUCATION' 'FIFTY' 'IMPACT' 'RELIGION' 'ARTS']
```

3.Count of Unique Categories

In [89]:

```
num_unique_categories = len(unique_categories)
print("Number of Unique Categories:", num_unique_categories)
```

```
Number of Unique Categories: 36
```

4.Proportion of Categories

To check the % of each category to understand its relative importance in dataset.

In [90]:

```
category_proportions = df['Article category'].value_counts(normalize=True)
print("Category Proportions:\n", category_proportions)
```

```
Category Proportions:
 Article category
STYLE & BEAUTY    0.153630
ENTERTAINMENT     0.120738
POLITICS          0.095868
TRAVEL            0.084637
HOME & LIVING     0.068993
COMEDY            0.065383
PARENTS           0.057361
WOMEN             0.056959
FOOD & DRINK      0.043321
TASTE             0.036903
BUSINESS          0.030485
PARENTING         0.030485
WEIRD NEWS        0.023666
TECH              0.022062
QUEER VOICES      0.018853
SPORTS            0.013237
CRIME             0.013237
STYLE             0.012435
ARTS & CULTURE    0.011633
SCIENCE           0.008424
HEALTHY LIVING    0.005616
IMPACT            0.003610
BLACK VOICES      0.002808
GREEN             0.002407
LATINO VOICES     0.002407
DIVORCE           0.002407
CULTURE & ARTS    0.002006
MEDIA             0.002006
GOOD NEWS         0.002006
WELLNESS          0.002006
MONEY             0.001604
THE WORLDPOST     0.000802
FIFTY             0.000802
EDUCATION         0.000401
RELIGION          0.000401
ARTS              0.000401
Name: proportion, dtype: float64
```

5.Bar Plot:-

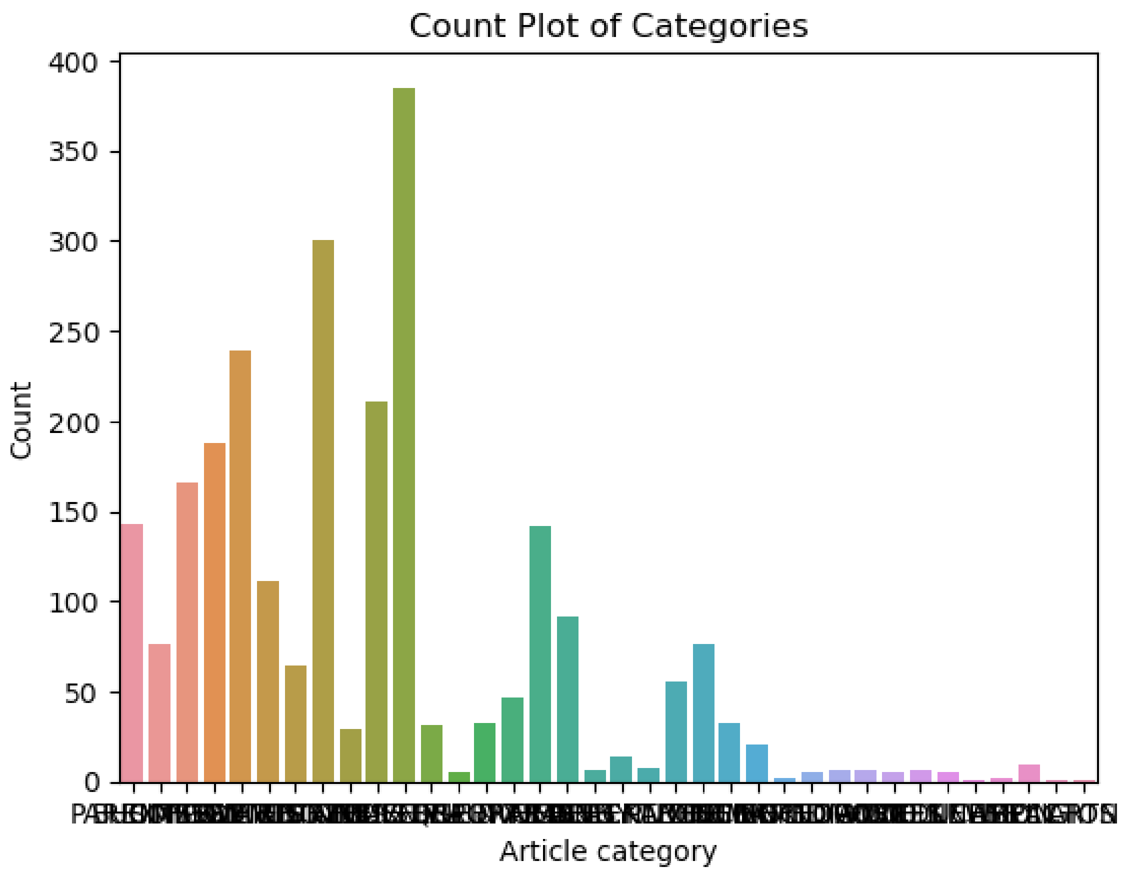
visualize the distribution of categories using bar plot or counnt plots

In [106]:

```
import seaborn as sns

sns.countplot(data=df, x='Article category')
plt.title('Count Plot of Categories')
plt.xlabel('Article category')
plt.ylabel('Count')
plt.show()
```

```
C:\Users\ASUS\anaconda3\lib\site-packages\seaborn\_core.py:1225: FutureWar
ning: is_categorical_dtype is deprecated and will be removed in a future v
ersion. Use isinstance(dtype, CategoricalDtype) instead
  if pd.api.types.is_categorical_dtype(vector):
C:\Users\ASUS\anaconda3\lib\site-packages\seaborn\_core.py:1225: FutureWar
ning: is_categorical_dtype is deprecated and will be removed in a future v
ersion. Use isinstance(dtype, CategoricalDtype) instead
  if pd.api.types.is_categorical_dtype(vector):
C:\Users\ASUS\anaconda3\lib\site-packages\seaborn\_core.py:1225: FutureWar
ning: is_categorical_dtype is deprecated and will be removed in a future v
ersion. Use isinstance(dtype, CategoricalDtype) instead
  if pd.api.types.is_categorical_dtype(vector):
```



6. Cross Tabulation:-

its use for explore relationship b/w two categorical variable.its help undestand how categories in one variable are distribuated across the categories of onother variable

In [107]:

```
df.head(1)
```

Out[107]:

	Article category	News Headline	Author	news_link	short_description	date
3	PARENTS	Funniest Parenting Tweets: What Moms And Dads ...	Hollis Miller	_RARE_	Kids may say the darndest things, but parents ...	15/07/16

In [109]:

```
cross_tab = pd.crosstab(df['Article category'],df['Author'])
print("Cross-Tabulation:\n", cross_tab)
```

```
Cross-Tabulation:
 Author            Alana Horowitz Satlin  Alanna Vagianos  Amanda McGowa
n  \
Article category
ARTS                                  0                0              0
ARTS & CULTURE                        0                0              0
BLACK VOICES                          0                0              0
BUSINESS                              0                0              0
COMEDY                                0                0              0
CRIME                                 0                0              0
CULTURE & ARTS                        0                0              0
DIVORCE                               0                0              0
EDUCATION                             0                0              0
ENTERTAINMENT                         0                5              0
FIFTY                                 0                0              0
FOOD & DRINK                          0                0              0
GOOD NEWS                             0                2              0
GREEN                                 0                1              0
HEALTHY LIVING                        0                0              0
```

In [ ]: