Assignment on Dataset using Numpy and Pandas

Name: Atul Gaikwad
PRN: 202401070070
Roll No: ET1-07
Subject: EDS Assignment on Dataset Using Numpy and Pandas
Dataset: Kaggle Text Classification Dataset

Sample Data Table

Message	Category	Message Length
"New exam schedule announced"	Education	29
"India wins world cup!"	Sports	21
"Top 10 technology trends"	Technology	28
"New book released on AI"	Books	25
"How to manage study time"	Study Tips	26

Problem Statements and Solutions

Que.1 Find the total number of records in the dataset.

```
total_records = df.shape[0]
print(total_records)
```

✅ Output: Total number of records

Que.2 Find the number of unique categories in the dataset.

```
unique_categories = df['category'].nunique()
print(unique_categories)
```

✅ Output: Unique category count

Que.3 Find the category that has the maximum number of messages.

```
max_category = df['category'].value_counts().idxmax()
print(max_category)
```

✅ Output: Category with most messages

Que.4 Find the message with the maximum message length.

```
max_length_message = df.loc[df['message_length'].idxmax(), 'message']
print(max_length_message)
```

✅ Output: Message with maximum length

Que.5 Find the average message length across all messages.

```
avg_length = df['message_length'].mean()
print(avg_length)
```

✅ Output: Average message length

Que.6 Find the total number of messages classified under 'Education'.

```
education_count = (df['category'] == 'Education').sum()
print(education_count)
```

✅ Output: Count of Education messages

Que.7 Find the shortest message in the dataset.

```
min_length_message = df.loc[df['message_length'].idxmin(), 'message']
print(min_length_message)
```

✅ Output: Shortest message

Que.8 Find all messages with length greater than 40 characters.

```
long_messages = df[df['message_length'] > 40]['message']
print(long_messages.tolist())
```

✅ Output: List of long messages

Que.9 Find the number of messages per category.

```
messages_per_category = df['category'].value_counts()
print(messages_per_category)
```

✅ Output: Messages per category

Que.10 Find the category with the least number of messages.

```
min_category = df['category'].value_counts().idxmin()
print(min_category)
```

✅ Output: Category with fewest messages

Que.11 Convert categories into numerical codes and find correlation with message length.

```
df['category_code'] = df['category'].astype('category').cat.codes
correlation = df['message_length'].corr(df['category_code'])
print(correlation)
```

✅ Output: Correlation value

Que.12 Find the top 3 longest messages.

```
top3_longest = df.sort_values('message_length', ascending=False).head(3)
print(top3_longest[['message', 'message_length']])
```

✅ Output: Top 3 longest messages

Que.13 Find messages from the 'Technology' category with message length above 40.

```
tech_messages = df[(df['category'] == 'Technology') & (df['message_length'] >
40)]['message']
print(tech_messages.tolist())
```

✅ Output: Technology messages > 40 length

Que.14 Find the percentage of 'Sports' category messages.

```
sports_percent = (df['category'].value_counts(normalize=True) *
100)['Sports']
print(sports_percent)
```

✅ Output: Sports message percentage

Que.15 Add a new column ‘word_count’ representing the number of words in each message.

```
df['word_count'] = df['message'].apply(lambda x: len(x.split()))
print(df[['message', 'word_count']])
```

✅ Output: Message and word count

Que.16 Find the message with the maximum number of words.

```
max_words_message = df.loc[df['word_count'].idxmax(), 'message']
print(max_words_message)
```

✅ Output: Message with most words

Que.17 Find messages that contain the word "exam".

```
exam_messages = df[df['message'].str.contains('exam', case=False,
na=False)]['message']
print(exam_messages.tolist())
```

✅ Output: Messages containing "exam"

Que.18 Find the mean and median word count across all messages.

```
mean_words = df['word_count'].mean()
median_words = df['word_count'].median()
print(mean_words, median_words)
```

✅ Output: Mean and median word count

Que.19 Find the category with the highest average message length.

```
category_avg_length =
df.groupby('category')['message_length'].mean().idxmax()
print(category_avg_length)
```

✅ Output: Category with highest average message length

Que.20 Find how many messages have a length between 30 and 50 characters.

```
messages_30_50 = df[(df['message_length'] >= 30) & (df['message_length'] <=
50)].shape[0]
print(messages_30_50)
```

✅ Output: Messages between 30 and 50 characters

End of Assignment