Assignment on Dataset using Numpy and Panda

Name :- Soham Kand

PRN :- 202401070215

Roll No :- 63

Division :- ET2

Subject :- EDS Assignment On Dataset using Numpy and Panda

Dataset :- Twitter US Airline Sentiment.

Formulate 20 problem statements for a given dataset using Numpy and Pandas and Apply

Numpy and pandas methods to find the solution for the formulated problem statements. Datasheet is Twitter US Airline Sentiment.

First, a brief about the Twitter US Airline Sentiment dataset:

This dataset contains tweets about airlines in the United States labeled as *positive*, *negative*, and *neutral*. Main columns include:

- **`tweet_id`**
- **`airline_sentiment` (positive/negative/neutral)**
- **`airline`**
- **`airline_sentiment_confidence`**
- **`negativereason` (if sentiment is negative)**
- **`negativereason_confidence`**
- **`text`**
- **`tweet_created`**
- **`user_timezone` etc.**

✨ Here's the list:

No.		Solution using Pandas/Numpy
1	Find total number of tweets in the dataset.	`len(df)` or `df.shape[0]`
2	Count the number of positive, negative, and neutral tweets.	`df['airline_sentiment'].value_counts()`
3	Find the airline with the most tweets.	`df['airline'].value_counts().idxmax()`
4	Find the mean sentiment confidence score	`df['airline_sentiment_confidence'].mean()`

No.		Solution using Pandas/Numpy
	for all tweets.	
5	Find the airline with the highest average sentiment confidence.	`df.groupby('airline')['airline_sentiment_confidence'].mean().idxmax()`
6	Find the tweet with the highest sentiment confidence.	`df.loc[df['airline_sentiment_confidence'].idxmax()]`
7	List top 5 reasons for negative tweets.	`df['negativereason'].value_counts().head(5)`
8	Calculate the	`df['airline_sentiment'].value_counts(normalize=True) * 100`

No.		Solution using Pandas/Numpy
	percentage of tweets for each sentiment.	
9	Find how many tweets have missing `negativereason` values.	`df['negativereason'].isna().sum()`
10	Find the most common user timezone among all tweets.	`df['user_timezone'].value_counts().idxmax()`
11	Find standard deviation	`df['airline_sentiment_confidence'].std()`

No.		Solution using Pandas/Numpy
	of sentiment confidence.	
12	Identify airlines which have only positive tweets.	`df.groupby('airline')['airline_sentiment'].nunique()` (then filter where only 'positive')
13	For each airline, count how many negative tweets exist.	`df[df['airline_sentiment']=='negative'].groupby('airline').size()`
14	Calculate the total number of tweets created per day.	`df['tweet_created'].dt.date.value_counts()`

No.		Solution using Pandas/Numpy
15	Find the tweet(s) posted at the earliest time.	`df.loc[df['tweet_created'].idxmin()]`
16	Find average negativereason_confidence where sentiment is negative.	`df[df['airline_sentiment']=='negative']['negativereason_confidence'].mean()`
17	Find the number of unique users who tweeted.	`df['tweet_id'].nunique()`
18	Find which negative reason is	`df[(df['airline']=='United') & (df['airline_sentiment']=='negative')]['negativereason'].value_counts().idxmax()`

No.		Solution using Pandas/Numpy
	most common for 'United' airline.	
19	Check if there are any duplicate tweets based on `text`.	`df['text'].duplicated().sum()`
20	Find which airline has the highest proportion of positive tweets.	(group airline & calculate ratio positive/total)

☑ Here’s an example of code to perform this:

```
python
CopyEdit
import pandas as pd
import numpy as np

# Load the dataset
df = pd.read_csv('Tweets.csv')   #
Adjust if your file name differs

# 1. Total tweets
total_tweets = len(df)

# 2. Sentiment count
sentiment_counts =
df['airline_sentiment'].value_counts()

# 3. Airline with most tweets
most_tweeted_airline =
df['airline'].value_counts().idxmax()

# 4. Mean sentiment confidence
mean_confidence =
df['airline_sentiment_confidence'].mea
n()

# 5. Airline with highest avg
sentiment confidence
```

```
best_airline_confidence =
df.groupby('airline')['airline_sentime
nt_confidence'].mean().idxmax()

# 6. Tweet with highest sentiment
confidence
highest_confidence_tweet =
df.loc[df['airline_sentiment_confidenc
e'].idxmax()]

# 7. Top 5 negative reasons
top_negative_reasons =
df['negativereason'].value_counts().he
ad(5)

# 8. Sentiment percentage
sentiment_percentage =
df['airline_sentiment'].value_counts(n
ormalize=True) * 100

# 9. Missing negativereason
missing_negativereason =
df['negativereason'].isna().sum()

# 10. Most common timezone
common_timezone =
df['user_timezone'].value_counts().idx
max()
```

```
# 11. Std deviation of sentiment
confidence
std_confidence =
df['airline_sentiment_confidence'].std
()

# 12. Airlines with only positive
tweets
positive_airlines =
df[df['airline_sentiment']=='positive'
]['airline'].value_counts()

# 13. Negative tweets per airline
negative_tweets_airline =
df[df['airline_sentiment']=='negative'
].groupby('airline').size()

# 14. Tweets per day
df['tweet_created'] =
pd.to_datetime(df['tweet_created'])
tweets_per_day =
df['tweet_created'].dt.date.value_coun
ts()

# 15. Earliest tweet
earliest_tweet =
df.loc[df['tweet_created'].idxmin()]
```

```
# 16. Average
negativereason_confidence
avg_neg_reason_conf =
df[df['airline_sentiment']=='negative'
]['negativereason_confidence'].mean()

# 17. Unique users
unique_users =
df['tweet_id'].nunique()

# 18. Most common negative reason for
United
common_neg_reason_united =
df[(df['airline']=='United') &
(df['airline_sentiment']=='negative')]
['negativereason'].value_counts().idxm
ax()

# 19. Duplicate tweets based on text
duplicate_tweets =
df['text'].duplicated().sum()

# 20. Airline with highest positive
ratio
positive_ratio =
(df[df['airline_sentiment']=='positive
```

```
'].groupby('airline').size() /
df.groupby('airline').size()).idxmax()
```