20 Pandas and Numpy Problems from Tweets Dataset

- Name : Omkumar Rajput
- Division : CS31
- Roll No : CS3-22
- PRN : 202401040135

1. Find the number of tweets for each airline.

```
tweets_df['airline'].value_counts()
```

2. Find the percentage distribution of sentiments (positive, negative, neutral).

```
tweets_df['airline_sentiment'].value_counts(normalize=True) * 100
```

3. Calculate the average sentiment confidence.

```
tweets_df['airline_sentiment_confidence'].astype(float).mean()
```

4. Identify the top 5 users who posted the most tweets.

```
tweets_df['name'].value_counts().head(5)
```

5. Find the most common negative reason.

```
tweets_df['negativereason'].value_counts().idxmax()
```

6. Find the average number of retweets.

```
tweets_df['retweet_count'].mean()
```

7. Check how many tweets have no specified negative reason.

```
tweets_df['negativereason'].isnull().sum()
```

8. Find out which airline received the highest number of negative tweets.

```
tweets_df[tweets_df['airline_sentiment'] == 'negative']['airline'].value_counts().idxmax()
```

20 Pandas and Numpy Problems from Tweets Dataset

- Name : Omkumar Rajput
- Division : CS31

9. Find the airline with the highest average sentiment

confidence.tweets_df.groupby('airline')['airline_sentiment_confidence'].mean().idxmax()

- Roll No : ET1-21
- PRN : 202401070107

10. Extract the top 10 tweets with the highest retweet count.

```
tweets_df.nlargest(10, 'retweet_count')[['text', 'retweet_count']]
```

11. Find the time period (morning/afternoon/evening/night) when most tweets were made.

```
tweets_df['tweet_created'] = pd.to_datetime(tweets_df['tweet_created'])
def categorize_time(x):     hour = x.hour     if 5 <= hour < 12:         return
'Morning'     elif 12 <= hour < 17:         return 'Afternoon'     elif 17 <=
hour < 21:

return 'Evening'     else:         return 'Night'
tweets_df['time_of_day'] =
tweets_df['tweet_created'].apply(categorize_time)
tweets_df['time_of_day'].value_counts()
```

12. Find the number of tweets per timezone.

```
tweets_df['user_timezone'].value_counts()
```

13. Find which timezone tweets the most negatively.

```
tweets_df[tweets_df['airline_sentiment'] == 'negative']['user_timezone'].value_counts().idxmax()
```

- Roll No : ET1-21
- PRN : 202401070107

14. How many tweets were posted without a user timezone?

```
tweets_df['user_timezone'].isnull().sum()
```

20 Pandas and Numpy Problems from Tweets Dataset

- Name : Omkumar Rajput
- Division : CS31

15. Find tweets that mention 'cancelled' flights.

```
tweets_df[tweets_df['text'].str.contains('cancelled', case=False, na=False)]
```

16. Find out the correlation between retweet count and sentiment confidence.

```
tweets_df[['retweet_count', 'airline_sentiment_confidence']].corr()
```

17. Replace missing negative reasons with 'No Reason Provided'.

```
tweets_df['negativereason'].fillna('No Reason Provided', inplace=True)
```

18. Check how many unique users tweeted.

```
tweets_df['name'].nunique()
```

19. Find the standard deviation of sentiment confidence for each airline.

```
tweets_df.groupby('airline')['airline_sentiment_confidence'].std()
```

20. Find the number of tweets with coordinates provided.

```
tweets_df['tweet_coord'].notnull().sum()
```