Theory Activity No. 1

Name: Kartik Somnath Uphade

Div : CS41 PRN : 20240104015
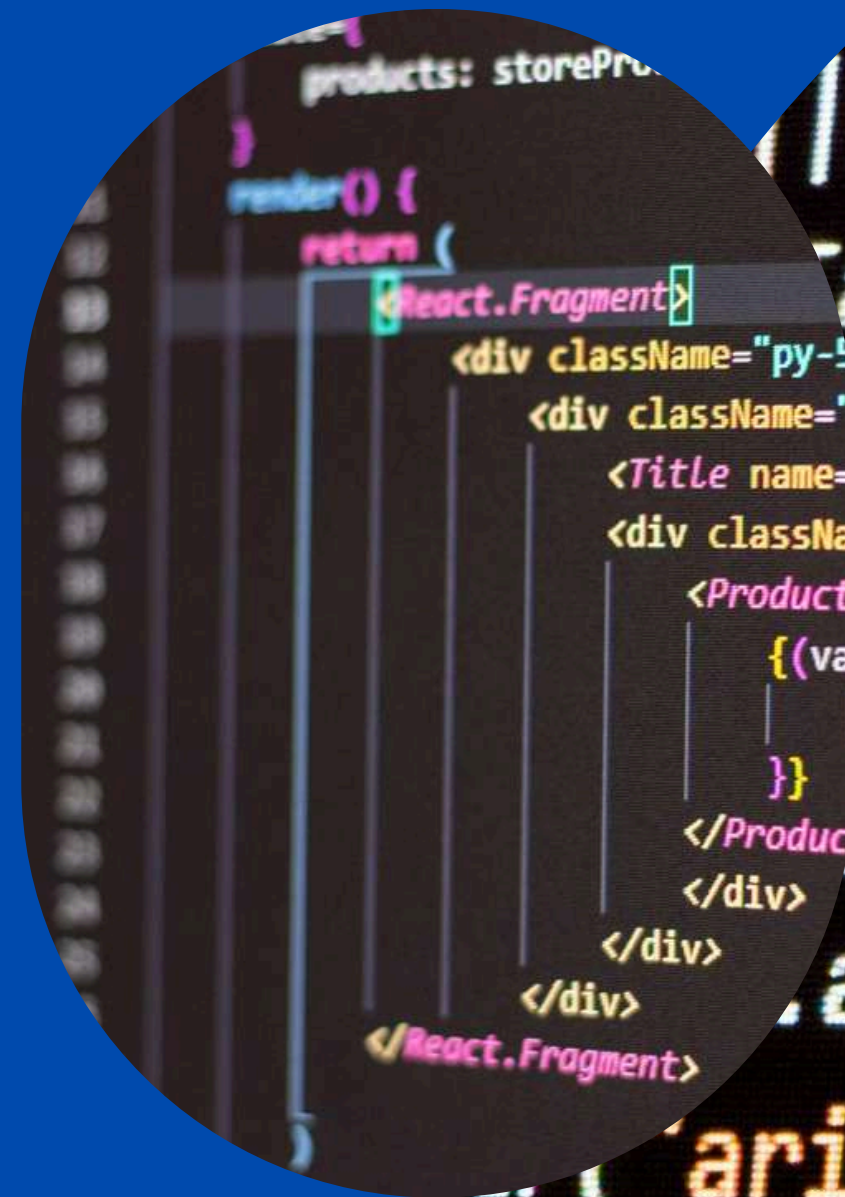
Roll no. : CS4-22

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.

Real-life dataset - **Twitter US Airline Sentiment**

Problem 1 : load the dataset

Solution :

```
import pd as pandas
import np as numpy
 df=pd.read_csv("Tweets.csv")
```

Problem 2 : Find the total number of tweets in dataset.

Solution :

```
total_tweets = len(df)
print(total_tweets)
```

Problem 3 : Find top 10 most common airline sentiment

Solution :

top_airline = df["airline_sentiment"].value_counts().head(10)
print(top_airline)

Problem 4: Find average sentiment confidence for each sentiment

```
avg_sent_confidence =
df.groupby("airline_sentiment")
["airline_sentiment_confidence"].mean()
print(avg_sent_confidence)
```

Problem 5 : How many tweets are missing the 'negativereason' column

Solution :

```
missing_tweets = df['negativereason'].isnull().sum()
print(missing_tweets)
```

Problem 6: Find average number of charaters in tweets for each sentiment

Solution:

```
df["text_length"]=df["text"].str.len()
avg_no_of_char=df.groupby("airline_sentiment")
["text_length"].mean()
print(avg_no_of_char)
```

Problem 7: Find multiple aggregation of airline_sentiment_confidence

Solution :

mul_agg=df.groupby('airline_sentiment_confidence').agg(['mean','max', 'min'])
print(mul_agg)

Problem 8 : List all the unique airlines present

Solution :

```
unique_airlines = df['airline'].unique()
print(unique_airlines)
```

Problem 9 : Find the airline with the maximum negative tweets

Solution :

```
most_negative_airline =
df[df['airline_sentiment']=='negative']
['airline'].value_counts().idxmax()
print(most_negative_airline)
```

Problem 10 : Find the percentage of negative tweets for each airline

```
negative_percentage =
(df[df['airline_sentiment']=='negative']
['airline'].value_counts() / df['airline'].value_counts()) * 100
print(negative_percentage)
```

Problem 11 : Find the tweet with the highest sentiment confidence

Solution :

```
highest_confidence_tweet =
df.loc[df['airline_sentiment_confidence'].idxmax()]
print(highest_confidence_tweet[['text',
'airline_sentiment_confidence']])
```

Problem 12 : Find out how many tweets mention 'late' in the text

Solution :

```
late_mentions = df['text'].str.contains('late', case=False,
na=False).sum()
print(late_mentions)
```

Problem 13 : Create a Pivot Table showing counts of Sentiments for each Airline

Solution :

```
pivot_sentiment_airline = pd.pivot_table(df, index='airline',
columns='airline_sentiment', values='tweet_id',
aggfunc='count', fill_value=0)
print(pivot_sentiment_airline)
```

Problem 14 : Find the airline with the least number of positive tweets

Solution :

```
least_positive_airline = df[df['airline_sentiment'] ==
'positive']['airline'].value_counts().idxmin()
print(least_positive_airline)
```

Problem 15: Find the day with the maximum tweets posted

Solution :

```
df['tweet_created'] = pd.to_datetime(df['tweet_created'])
most_active_day = df['tweet_created'].dt.date.value_counts().idxmax()
print(most_active_day)
```

Problem 16 : Check for duplicate tweets in the dataset

Solution :

```
duplicate_tweets = df.duplicated(subset='text').sum()
print(duplicate_tweets)
```

Problem 17 : Find the distribution of sentiments (positive, negative, neutral)

Solution :

```
sentiment_distribution =
df['airline_sentiment'].value_counts()
print(sentiment_distribution)
```

Problem 18 : Find the tweets where the sentiment confidence is less than 0.5.

Solution :

low_confidence_tweets = df[df['airline_sentiment_confidence'] < 0.5]
print(low_confidence_tweets[['text', 'airline_sentiment_confidence']])

Problem 19 : Find the shortest tweet in the dataset.

Solution :

```
shortest_tweet = df.loc[df['text_length'].idxmin()]
print(shortest_tweet[['text', 'text_length']])
```
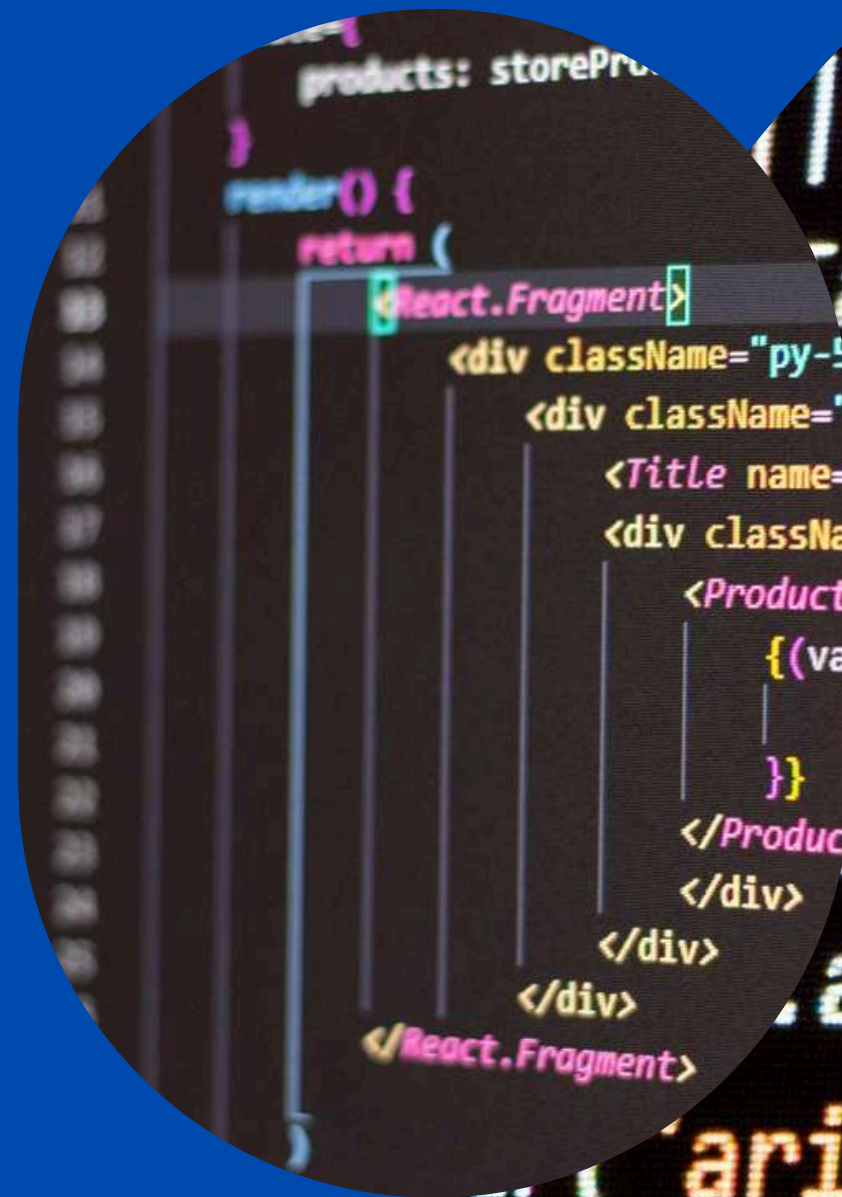
Problem 20 : Find the number of tweets per day of the week

Solution :

```
tweets_per_day =
df['tweet_created'].dt.day_name().value_counts()
print(tweets_per_day)
```

Thank you for your time!