ÔN THI GIỮA KỲ

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Cho sẵn dữ liệu TwitterUSAirlineSentiment.csv bao gồm các cột:

- airline_sentiment: Các mức độ phản hồi (positive, neutral, negative)
- airline_sentiment_confidence: đô tin cậy của phản hồi
- negativereason_confidence: độ tin cậy của lý do phủ định
- airline: Hãng máy bay (Virgin America, United, Southwest, Delta, US Airways, American)
- name: tên của hành khách phản hồi
- text: phản hồi dạng chữ
- user_timezone: múi giờ

1. Đọc dữ liệu và hiển thị 10 dòng đầu tiên:

```
df = pd.read_csv('Tweets.csv')
df.head(3)
```

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sent
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	
1	570301130888122368	positive	0.3486	NaN	0.0	Virgin America	
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	

2. (3 điểm) Tìm hiểu thông tin về dữ liệu:

a. Số lượng phản hồi của mỗi hãng máy bay và số lượng phản hồi mỗi mức độ của từng hãng máy bay.

```
# Số lượng phản hồi của mỗi hãng máy bay
airline_feedback_count = df['airline'].value_counts()

# Số lượng phản hồi mỗi mức độ của từng hãng máy bay
sentiment_by_airline = df.groupby(['airline', 'airline_sentiment']).size()

print(airline_feedback_count)
print(sentiment_by_airline)
```

```
airline
United            3822
US Airways        2913
American          2759
Southwest         2420
Delta             2222
Virgin America     504
Name: count, dtype: int64
airline         airline_sentiment
American        negative            1960
                neutral              463
                positive             336
Delta           negative             955
                neutral              723
                positive             544
Southwest       negative            1186
```

```
                  neutral              664
                  positive             570
US Airways        negative            2263
                  neutral              381
                  positive             269
United            negative            2633
                  neutral              697
                  positive             492
Virgin America    negative             181
                  neutral              171
                  positive             152
dtype: int64
```

b. Với mỗi mức độ phản hồi của từng hãng máy bay, hãy xác định các tham số thống kê: giá trị trung bình, độ lệch chuẩn của các thuộc tính đô tin cậy của phản hồi (airline_sentiment_confidence), độ tin cậy của lý do phủ định (negativereason_confidence); và độ đo xu hướng tập trung (central tendency) của thuộc tính múi giờ (user_timezone).

```
# Giá trị trung bình (mean) và độ lệch chuẩn (std) của thuộc tính 'airline_sentiment_confidence' và 'negativereason_confidence'
confidence_stats = df.groupby(['airline', 'airline_sentiment']).agg({'airline_sentiment_confidence':['mean','std'],
                                                                     'negativereason_confidence':['mean', 'std']})
# Xu hướng tập trung (central tendency) của thuộc tính múi giờ
timezone_mode = df.groupby(['airline', 'airline_sentiment'])['user_timezone'].agg(pd.Series.mode)

print(confidence_stats)
print(timezone_mode)
```

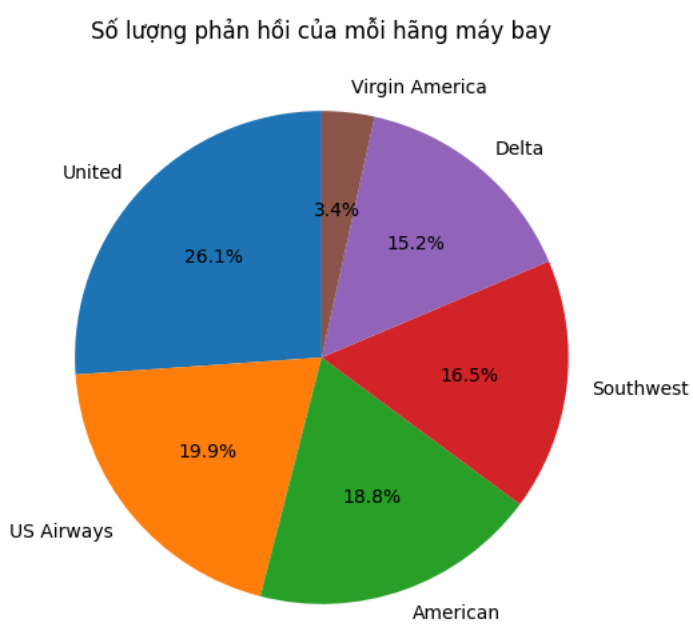
```
                                  airline_sentiment_confidence            \
                                                          mean       std
airline        airline_sentiment
American       negative                             0.944955  0.124992
               neutral                              0.825938  0.186387
               positive                             0.882302  0.169425
Delta          negative                             0.902202  0.162831
               neutral                              0.829264  0.185266
               positive                             0.867111  0.177139
Southwest      negative                             0.920533  0.147741
               neutral                              0.826109  0.186759
               positive                             0.886105  0.173846
US Airways     negative                             0.945714  0.126575
               neutral                              0.821922  0.187339
               positive                             0.859686  0.191301
United         negative                             0.933383  0.138631
               neutral                              0.809756  0.184240
               positive                             0.856012  0.188985
Virgin America negative                             0.901733  0.158833
               neutral                              0.838368  0.181370
               positive                             0.887978  0.173069

                                  negativereason_confidence
                                                       mean       std
airline        airline_sentiment
American       negative                            0.744644  0.235690
               neutral                             0.000000  0.000000
               positive                            0.000000  0.000000
Delta          negative                            0.710520  0.240584
               neutral                             0.000000  0.000000
               positive                            0.000000  0.000000
Southwest      negative                            0.732866  0.236287
               neutral                             0.000000  0.000000
               positive                            0.000000  0.000000
US Airways     negative                            0.750028  0.237247
               neutral                             0.000000  0.000000
               positive                            0.000000  0.000000
United         negative                            0.714719  0.240164
               neutral                             0.000000  0.000000
               positive                            0.000000  0.000000
Virgin America negative                            0.717003  0.228287
               neutral                             0.000000  0.000000
               positive                            0.000000  0.000000
airline         airline_sentiment
American        negative             Eastern Time (US & Canada)
                neutral              Eastern Time (US & Canada)
                positive             Central Time (US & Canada)
Delta           negative             Eastern Time (US & Canada)
                neutral              Eastern Time (US & Canada)
                positive             Eastern Time (US & Canada)
Southwest       negative             Central Time (US & Canada)
                neutral              Eastern Time (US & Canada)
                positive             Central Time (US & Canada)
US Airways      negative             Eastern Time (US & Canada)
```

```
                neutral             Eastern Time (US & Canada)
                positive            Eastern Time (US & Canada)
United          negative            Eastern Time (US & Canada)
```

3. (3 điểm) Đồ thị hóa dữ liệu:
   a. Số lượng phản hồi của mỗi hãng máy bay bằng đồ thị tròn.
   b. Bằng đồ thị hộp râu (box plot) biểu diễn trực quan về cách dữ liệu của thuộc tính đô tin cậy của phản hồi (airline_sentiment_confidence) theo hãng hàng không nằm rải rác trên mặt phẳng, so sánh giữa các mức độ phản hồi.
   c. Bằng đồ thị phù hợp biểu diễn sự tương quan bằng màu sắc giữa các thuộc tính dựa trên hệ số tương quan.

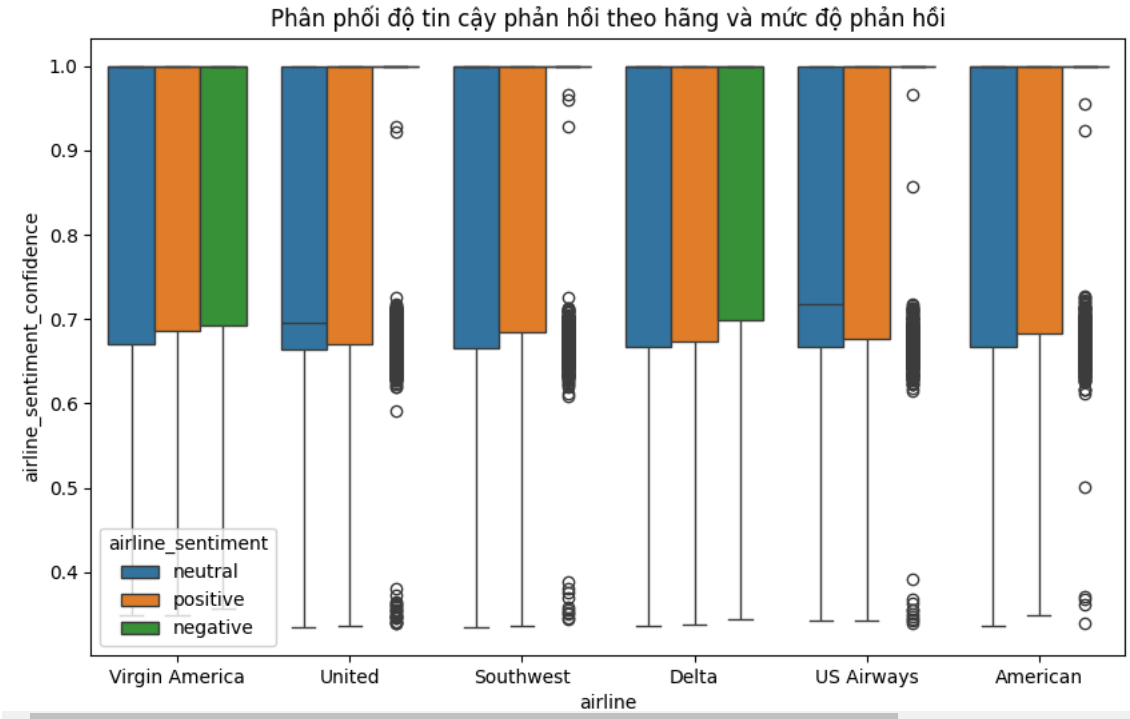
a. Số lượng phản hồi của mỗi hãng máy bay bằng đồ thị tròn.

```
plt.figure(figsize=(10,6))
airline_feedback_count.plot(kind='pie', autopct='%1.1f%%', startangle=90)
plt.title('Số lượng phản hồi của mỗi hãng máy bay')
plt.ylabel('')
plt.show()
```



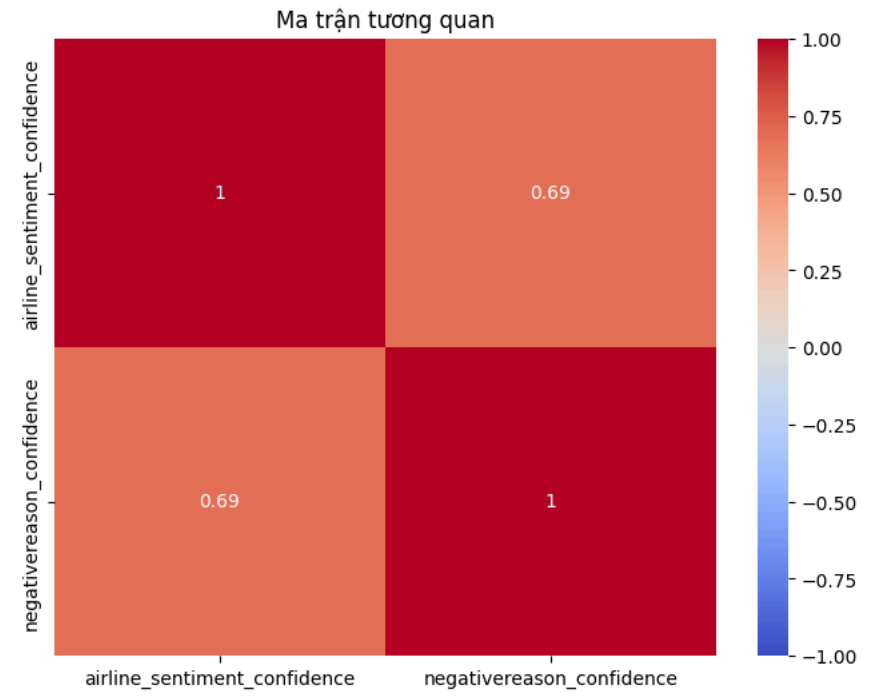
b. Bằng đồ thị hộp râu (box plot) biểu diễn trực quan về cách dữ liệu của thuộc tính đô tin cậy của phản hồi (airline_sentiment_confidence) theo hãng hàng không nằm rải rác trên mặt phẳng, so sánh giữa các mức độ phản hồi.

```
# Vẽ box plot cho 'airline_sentiment_confidence' theo từng hãng và mức độ phản hồi
plt.figure(figsize=(10, 6))
sns.boxplot(x='airline', y='airline_sentiment_confidence', hue='airline_sentiment', data=df)
plt.title('Phân phối độ tin cậy phản hồi theo hãng và mức độ phản hồi')
plt.show()
```



c. Bằng đồ thị phù hợp biểu diễn sự tương quan bằng màu sắc giữa các thuộc tính dựa trên hệ số tương quan.

```
# Tính ma trận tương quan
correlation_matrix = df[['airline_sentiment_confidence', 'negativereason_confidence']].corr()

# Vẽ heatmap thể hiện sự tương quan
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', vmin=-1, vmax=1)
plt.title('Ma trận tương quan')
plt.show()
```



+ Mã + Văn bản

4. (3 điểm) Đường kính của một loại chi tiết do một máy sản xuất có phân phối chuẩn, kỳ vọng 20mm, độ lệch chuẩn 0,2mm). Lấy ngẫu nhiên 1 chi tiết máy. Tính xác suất để

a. có đường kính nhỏ hơn 20,3mm

b. có đường kính trong khoảng 19,9mm đến 20,3mm,

c. có đường kính sai khác với kỳ vọng không quá 0,3mm\

```
from scipy.stats import norm

mu = 20  # Kỳ vọng
sigma = 0.2  # Độ lệch chuẩn

# a. P(X < 20.3)
p_a = norm.cdf(20.3, mu, sigma)

# b. P(19.9 < X < 20.3)
p_b = norm.cdf(20.3, mu, sigma) - norm.cdf(19.9, mu, sigma)

# c. P(|X - 20| <= 0.3)
p_c = norm.cdf(20.3, mu, sigma) - norm.cdf(19.7, mu, sigma)

print(f'P(X < 20.3): {p_a}')
print(f'P(19.9 < X < 20.3): {p_b}')
print(f'P(|X - 20| <= 0.3): {p_c}')
```

```
P(X < 20.3): 0.9331927987311424
P(19.9 < X < 20.3): 0.624655260005158
P(|X - 20| <= 0.3): 0.8663855974622847
```