โจทย์ Black Friday Dataset

ให้พิจารณาชุดข้อมูล BlackFriday_train.csv ซึ่งสามารถศึกษาข้อมูลทั่วไปของชุดข้อมูลได้ที่ https://www.kaggle.com/sdolezel/black-friday#train.csv โดยชุดข้อมูลดังกล่าวเก็บข้อมูลบันทึกรายการซื้อสินค้าของผู้คนในเทศกาล Black Friday ซึ่งแต่ละ record (row) แทนหนึ่ง transaction ของการซื้อสินค้า ซึ่งประกอบด้วยตัวแปรดังต่อไปนี้

User_ID: รหัสผู้ใช้ที่ซื้อสินค้า transaction ดังกล่าว

Product_ID: รหัสสินค้าที่ซื้อ

Gender: เพศของผู้ใช้

Age: ช่วงอายุของผู้ใช้

Occupation: หมายเลขอาชีพของผู้ใช้ (ชุดข้อมูลไม่ระบุชื่ออาชีพ)

City_Category: กลุ่มประเภทของเมือง

Stay_In_Current_City_Years: จำนวนปีที่ผู้ใช้อาศัยอยู่ในเมืองปัจจุบัน

Marital_Status: สถานะการแต่งงาน

Product_Category_1: หมายเลขของสินค้าสำหรับระบุในสินค้ากลุ่มประเภท 1

Product_Category_2: หมายเลขของสินค้าสำหรับระบุในสินค้ากลุ่มประเภท 2

Product_Category_3: หมายเลขของสินค้าสำหรับระบุในสินค้ากลุ่มประเภท 3

Purchase: ปริมาณค่าใช้จ่ายสำหรับสินค้า

```
!ls "/content/drive/My Drive/BlackFriday_train.csv"

```

```
'/content/drive/My Drive/BlackFriday_train.csv'
```

```
import pandas as pd
import numpy as np
from matplotlib import pyplot as plt

df = pd.read_csv('/content/drive/My Drive/BlackFriday_train.csv')

```

ให้เขียน Python Sript ผ่าน Google Colab หรือ Ipython Notebook เพื่อวิเคราะห์และตอบคำถามต่อไปนี้

ข้อ 1 (1 คะแนน) ใน dataset ดังกล่าว เก็บรายการใช้จ่ายของ User เป็นจำนวนทั้งหมดกี่คน?

```
df1 = df.pivot_table(index=['User_ID'], aggfunc='size').count()
df1
print('Numbers of User is ', df1)
```

```
Numbers of User is  5891
```

ข้อ 2 (2 คะแนน)

2.1) ผู้ที่มาซื้อเป็นเพศชาย (M) ทั้งหมดกี่คน และเป็นเพศหญิง (F) ทั้งหมดกี่คน? (1 คะแนน)

```
df2 = df.pivot_table(index=['Gender','User_ID'], aggfunc='size')
print('Numbers of Famale is' ,len(df2['F']))
print('Numbers of Male is' ,len(df2['M']))

```

```
Numbers of Famale is 1666
Numbers of Male is 4225
```

2.2) ผู้ที่มาซื้อส่วนใหญ่เป็นเพศชายหรือหญิง? (1 คะแนน)

```
print('Numbers of Famale is' ,len(df2['F']))
print('Numbers of Male is' ,len(df2['M']))
print('Male is more than Famale.')
```

```
Numbers of Famale is 1666
Numbers of Male is 4225
Male is more than Famale.
```

ข้อ 3 (1 คะแนน)

เมื่อพิจารณาเฉพาะผู้ซื้อที่เป็นเพศหญิง ค่าเฉลี่ยของการใช้จ่ายรวม (Sum of Purchase) ต่อผู้ซื้อหนึ่งคนเป็นจำนวนเท่าไร?

```
df3 = df.loc[(df['Gender']=='F'),'Purchase'].sum()
arg_sum_of_purchase = df3/len(df2['F'])
print('Arg sum of purchase by Famale is ' , arg_sum_of_purchase)
```

```
Arg sum of purchase by Famale is  712024.3949579832
```

ข้อ 4 (2 คะแนน)

เราต้องการพิจารณาเป็นราย transaction ว่าสินค้าแต่ละหมายเลขในตัวแปร Product_Category_1 มีอัตราส่วนการซื้อเป็นอย่างไร

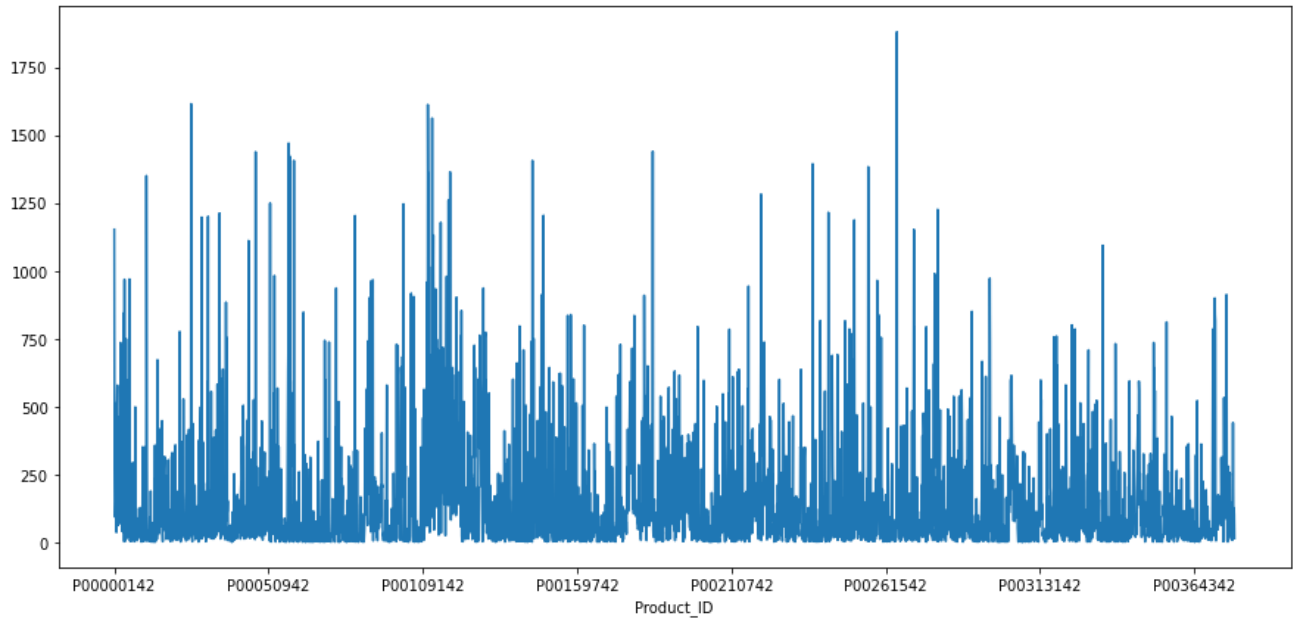
```
df4 = df.pivot_table(index=['Product_ID','Product_Category_1'], aggfunc='size')
df4
```

```
Product_ID  Product_Category_1
P00000142   3                    1152
P00000242   2                     376
P00000342   5                     244
P00000442   5                      92
P00000542   5                     149
```

4.1) พล็อตกราฟแท่ง (column chart) แสดงปริมาณการซื้อของทุกหมายเลขสินค้าในตัวแปร Product_Category_1 (ให้นับปริมาณตามจำนวน transaction โดยไม่ต้องดูจาก Purchase)

```
P00099842   3                     102
```

```
1 fig, ax = plt.subplots(figsize=(15,7))
2 df4 = df.groupby('Product_ID').count()['Product_Category_1'].plot(ax=ax)
3
```



4.2) ให้ระบุหมายเลขสินค้าที่มีอัตราส่วนการซื้อมากที่สุด 3 ลำดับแรก

```
1  # df.pivot_table(index=['Product_ID','Purchase'], aggfunc='size') #Tried to us
2 df4 = df.groupby('Product_ID').agg('sum')
3 df4.sort_values(by='Purchase',ascending=False).head(3)
```

Product_ID	User_ID	Occupation	Marital_Status	Product_Category_1	Produc
P00025442	1619931763	13355	652	1615	
P00110742	1616895727	13185	648	1612	
P00255842	1387479169	11396	535	22128	

ข้อ 5 (1 คะแนน)

จากข้อ 4.2) ปริมาณการซื้อสินค้าทั้ง 3 อันดับแรกรวมกัน คิดเป็นอัตราส่วนกี่เปอร์เซ็นต์ของปริมาณการซื้อสินค้าทั้งหมด

```
sum_top3 = df4.sort_values(by='Purchase',ascending=False).head(3).sum()
sumofPur = df['Purchase'].sum(axis = 0, skipna = True)
percentoftop3 = (sum_top3['Purchase']/sumofPur)*100
percentoftop3
```

```
1.5676878654031199
```

ข้อ 6 (3 คะแนน)

เราต้องการรู้ว่าตัวแปร Purchase มีความสัมพันธ์กับตัวแปร Age และ Stay_In_Current_City_Years มากน้อยแค่ไหน เช่น อายุเยอะขึ้น จะใช้จ่ายเยอะขึ้นหรือเปล่า หรือคนที่อาศัยอยู่ในเมืองปัจจุบันมานาน จะใช้จ่ายเยอะกว่า? โดยที่ให้พิจารณาเป็นราย transaction (ไม่ต้องดูเป็นรายบุคคล)

6.1) ให้ใช้ฟังก์ชัน pandas.DataFrame.corr() เพื่อพล็อตตาราง correlation matrix ระหว่าง 3 ตัวแปร ได้แก่ Age, Stay_In_Current_City_Years, และ Purchase

```
# df6 = df
# df6.corr()
# df6['Age'] = pd.to_numeric(df['Age'], errors='ignore')
# df6['Age'][0] = 17

df6['Age'][0]
```

```
17
```

6.2) พิจารณาค่าในตาราง แล้วตอบว่า Purchase มีความสัมพันธ์เชิงบวกกับตัวแปรใดมากกว่ากัน ระหว่าง Age และ Stay_In_Current_City_Years

```
1
```