

▼ โจทย์ Black Friday Dataset

ให้พิจารณาชุดข้อมูล BlackFriday_train.csv ซึ่งสามารถศึกษาข้อมูลทั่วไปของชุดข้อมูลได้ที่ <https://www.kaggle.com/sdoлезel/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

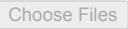
Product_Category_3: หมายเลขของสินค้าสำหรับระบุในสินค้ากลุ่มประเภท 3

Purchase: ปริมาณค่าใช้จ่ายสำหรับสินค้า

ให้เขียน Python Sript ผ่าน Google Colab หรือ Ipython Notebook เพื่อวิเคราะห์และตอบคำถามต่อไปนี้


▼ 1 ใน dataset ดังกล่าว เก็บรายการใช้จ่ายของ User เป็นจำนวนทั้งหมดกี่คน?

```
1 #solutions
2 from google.colab import files
3
4 files.upload()
```

 No file chosen Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.
Saving BlackFriday_train.csv to BlackFriday_train.csv
{'BlackFriday_train.csv': b'User_ID,Product_ID,Gender,Age,Occupation,City_Category,Stay_In_Current_City_Years,Marital_Status,Pr

```
1 import pandas as pd
2
3 df = pd.read_csv('BlackFriday_train.csv', sep=',')
```

```
1 print('ใน dataset เก็บรายการใช้จ่ายของ User เป็นจำนวน :', len(df.User_ID.unique()), 'คน')
```

 ใน dataset เก็บรายการใช้จ่ายของ User เป็นจำนวน : 5891 คน

▼ 2.1 ผู้ที่มาซื้อเป็นเพศชาย (M) ทั้งหมดกี่คน และเป็นเพศหญิง (F) ทั้งหมดกี่คน?

```
1 #solutions
2 check = df[['User_ID', 'Gender']]
3 check = check.drop_duplicates('User_ID', keep='first', ignore_index=True)
4 check['count'] = 1
5 gender = check.groupby('Gender')[['count']].sum()
6
7 print('ผู้ซื้อเพศชาย และเพศหญิงมีจำนวน')
8 gender
```

 ผู้ซื้อเพศชาย และเพศหญิงมีจำนวน

count	
Gender	
F	1666
M	4225

2.2 ผู้ที่มาซื้อส่วนใหญ่เป็นเพศชายหรือหญิง?

```
1 #solutions
2
3 print('ผู้ซื้อส่วนใหญ่ เป็นเพศ :', gender.index.max())
```

☞ ผู้ซื้อส่วนใหญ่ เป็นเพศ : M

3 เมื่อพิจารณาเฉพาะผู้ซื้อที่เป็นเพศหญิง ค่าเฉลี่ยของการใช้จ่ายรวม (Sum of Purchase) ต่อผู้ซื้อหนึ่งคนเป็นจำนวนเท่าไร?

```
1 gender.reset_index(inplace=True)
```

```
1 #solutions
2
3 sales = df[['User_ID', 'Gender', 'Purchase']]
4 sales = sales.groupby('Gender')[['Purchase']].agg('sum').reset_index()
5
6 sales = sales.merge(gender, how='inner', left_on='Gender', right_on='Gender')
7 sales['avg_purchase'] = sales.apply(lambda x : x['Purchase'] / x['count'], axis=1)
8
9
10 for row in sales.iterrows():
11     if row[1]['Gender'] == 'F':
12         print('ค่าเฉลี่ยของการใช้จ่ายรวมของผู้ซื้อเพศหญิง : %.2f' %row[1]['avg_purchase'])
```

☞ ค่าเฉลี่ยของการใช้จ่ายรวมของผู้ซื้อเพศหญิง : 712024.39

4 เราต้องการพิจารณาเป็นราย transaction ว่าสินค้าแต่ละหมายเลขในตัวแปร Product_Category_1

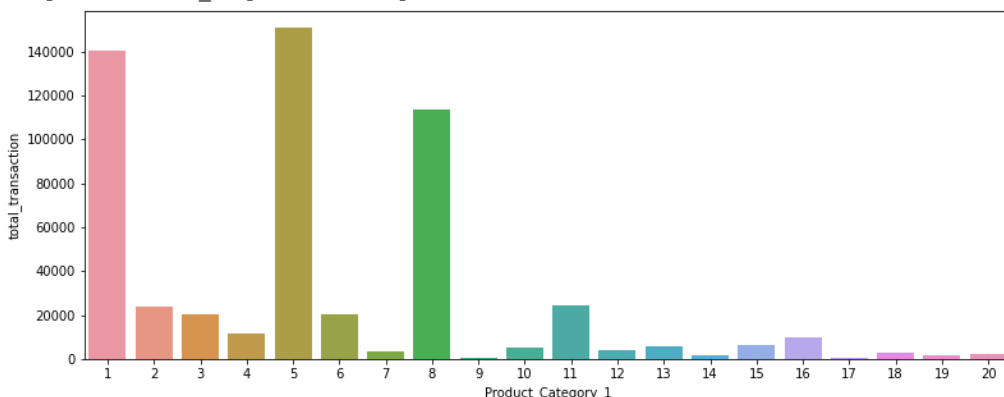
มีอัตราส่วนการซื้อเป็นอย่างไร

4.1 พล็อตกราฟแท่ง (column chart) แสดงปริมาณการซื้อของทุกหมายเลขสินค้าในตัวแปร Product_Category_1

(ให้เน้นปริมาณตามจำนวน transaction โดยไม่ต้องดูจาก Purchase)

```
1 #solutions
2 from collections import Counter
3 import matplotlib.pyplot as plt
4 import seaborn as sns
5
6 p1 = []
7
8 for row in df['Product_Category_1']:
9     p1.append(row)
10
11 p1_plot = pd.DataFrame.from_dict(dict(Counter(p1)), orient='index').reset_index()
12 p1_plot.rename(columns={'index' : 'Product_Category_1', 0 : 'total_transaction'}, inplace=True)
13
14 plt.figure(figsize=(13,5))
15 sns.barplot(x=p1_plot['Product_Category_1'], y=p1_plot['total_transaction'])
```

☞ /usr/local/lib/python3.6/dist-packages/statsmodels/tools/_testing.py:19: FutureWarning: pandas.util.testing is deprecated. Use import pandas.util.testing as tm
<matplotlib.axes._subplots.AxesSubplot at 0x7fcf64857f98>



▼ 4.2 ให้ระบุหมายเลขสินค้าที่มีอัตราส่วนการซื้อมากที่สุด 3 ลำดับแรก

```
1 #solutions
2 p1_plot['avg_transaction'] = p1_plot.apply(lambda x : x['total_transaction'] / (p1_plot['total_transaction'].sum()) * 100, axis=1)
3
4 p3 = p1_plot.nlargest(3, 'avg_transaction')
5
6 print('หมายเลขสินค้าที่มีอัตราส่วนการซื้อมากที่สุด 3 อันดับ')
7 for row in p3.iterrows():
8     print('>> %.0f' % row[1]['Product_Category_1'])
```

```
➤ หมายเลขสินค้าที่มีอัตราส่วนการซื้อมากที่สุด 3 อันดับ
>> 5
>> 1
>> 8
```

▼ 5 จากข้อ 4.2 ปริมาณการซื้อสินค้าทั้ง 3 อันดับแรกรวมกัน

คิดเป็นอัตราส่วนที่เปอร์เซ็นต์ของปริมาณการซื้อสินค้าทั้งหมด

```
1 #solutions
2
3 print('ปริมาณการซื้อสินค้าทั้ง 3 อันดับแรกรวมกันคิดเป็น %.2f' % p3['avg_transaction'].sum(), '% ของปริมาณการซื้อสินค้าทั้งหมด')
```

```
➤ ปริมาณการซื้อสินค้าทั้ง 3 อันดับแรกรวมกันคิดเป็น 73.67 % ของปริมาณการซื้อสินค้าทั้งหมด
```

▼ 6 เราต้องการรู้ว่าตัวแปร Purchase มีความสัมพันธ์กับตัวแปร Age และ

Stay_In_Current_City_Years มากน้อยแค่ไหน เช่น อายุเยอะขึ้น จะใช้จ่ายเยอะขึ้นหรือเปล่า หรือคนที่อาศัยอยู่ในเมืองปัจจุบันมานาน จะใช้จ่ายเยอะกว่า? โดยที่
ให้พิจารณาเป็นราย transaction (ไม่ต้องดูเป็นรายบุคคล)

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

Stay_In_Current_City_Years, และ Purchase

```
1 #solutions
2
3 cor = df[['Age', 'Stay_In_Current_City_Years', 'Purchase']]
4 cor[['c_Age']] = cor[['Age']].apply(lambda x:pd.Categorical(x).codes)
5 cor[['c_Years']] = cor[['Stay_In_Current_City_Years']].apply(lambda x:pd.Categorical(x).codes)
6
7
8 cor.corr()
```

```
➤ /usr/local/lib/python3.6/dist-packages/pandas/core/frame.py:2963: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-vs-returning-a-copy
self[k1] = value[k2]

	Purchase	c_Age	c_Years
Purchase	1.000000	0.015839	0.005422
c_Age	0.015839	1.000000	-0.004712
c_Years	0.005422	-0.004712	1.000000

▼ 6.2 พิจารณาค่าในตาราง แล้วตอบว่า Purchase มีความสัมพันธ์เชิงบวกกับตัวแปรใดมากกว่ากัน ระหว่าง Age และ

Stay_In_Current_City_Years

```
1 #solutions
2 print('Purchase มีความสัมพันธ์เชิงบวกกับตัวแปร Age มากกว่า')
```

```
➤ Purchase มีความสัมพันธ์เชิงบวกกับตัวแปร Age มากกว่า
```