**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

ให้เขียน Python Sript ผ่าน Google Colab หรือ Ipython Notebook เพื่อวิเคราะห์และตอบคำถามต่อไปนี้

**ข้อ 1** (1 คะแนน)

ใน dataset ดังกล่าว เก็บรายการใช้จ่ายของ User เป็นจำนวนทั้งหมดกี่คน?

**ข้อ 2** (2 คะแนน)

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

2.2) ผู้ที่มาซื้อส่วนใหญ่เป็นเพศชายหรือหญิง? (1 คะแนน)

**ข้อ 3** (1 คะแนน)

เมื่อพิจารณาเฉพาะผู้ซื้อที่เป็นเพศหญิง ค่าเฉลี่ยของการใช้จ่ายรวม (Sum of Purchase) ต่อผู้ซื้อหนึ่งคนเป็นจำนวนเท่าไร?

**ข้อ 4** (2 คะแนน) เราต้องการพิจารณาเป็นราย transaction ว่าสินค้าแต่ละหมายเลขในตัวแปร Product\_Category\_1 มีอัตราส่วนการซื้อเป็นอย่างไร

4.1) พล็อตกราฟแท่ง (column chart) แสดงปริมาณการซื้อของทุกหมายเลขสินค้าในตัวแปร Product\_Category\_1 (ให้นับปริมาณตามจำนวน transaction โดยไม่ต้องดูจาก Purchase)

4.2) ให้ระบุหมายเลขสินค้าที่มีอัตราส่วนการซื้อมากที่สุด 3 ลำดับแรก

**ข้อ 5** (1 คะแนน) จากข้อ 4.2) ปริมาณการซื้อสินค้าทั้ง 3 อันดับแรกรวมกัน คิดเป็นอัตราส่วนกี่เปอร์เซ็นต์ของปริมาณการซื้อสินค้าทั้งหมด

**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

6.2) พิจารณาค่าในตาราง แล้วตอบว่า Purchase มีความสัมพันธ์เชิงบวกกับตัวแปรใดมากกว่ากัน ระหว่าง Age และ Stay\_In\_Current\_City\_Years