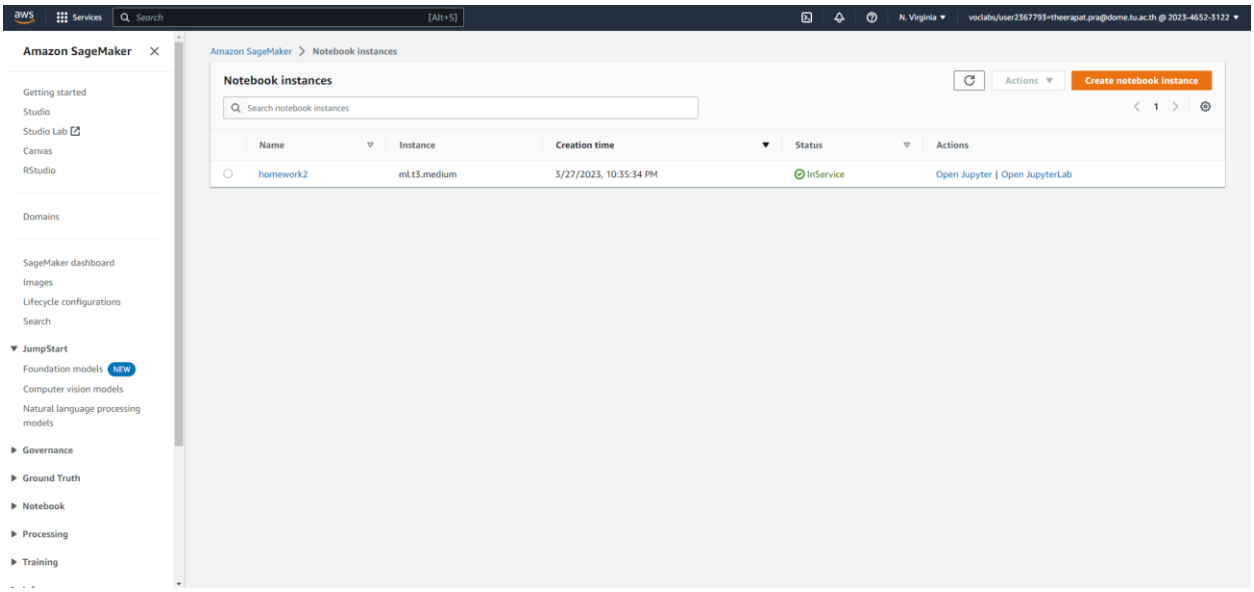
ชื่อ นายธีรภัทร์ ประจำทอง รหัสนักศึกษา 6509035256

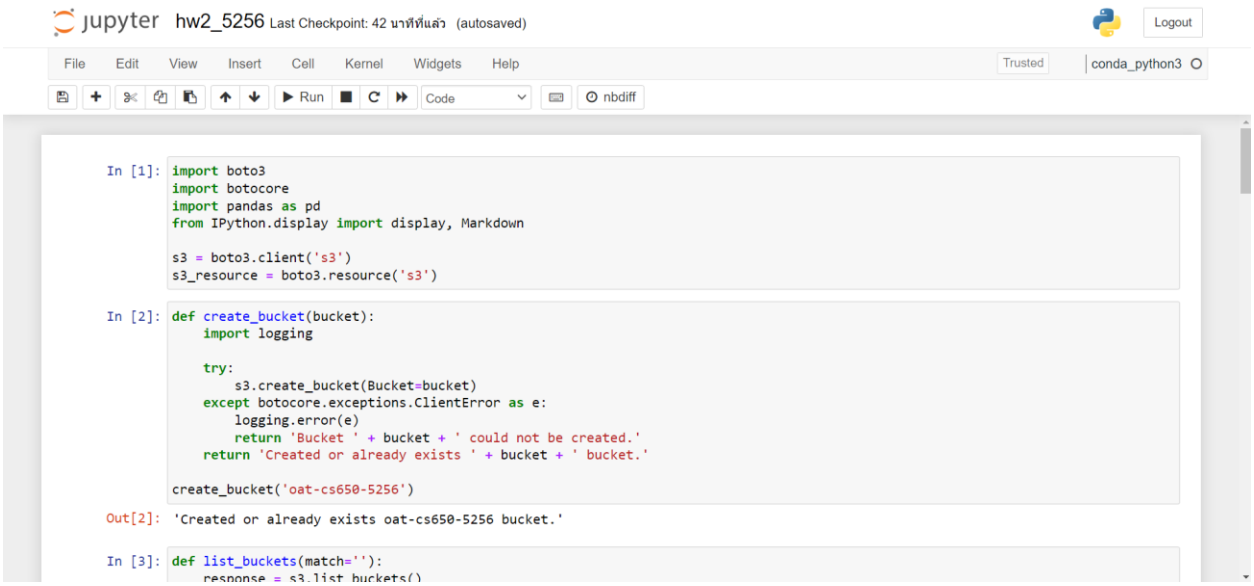
CS653 2/2565

HomeWork 2



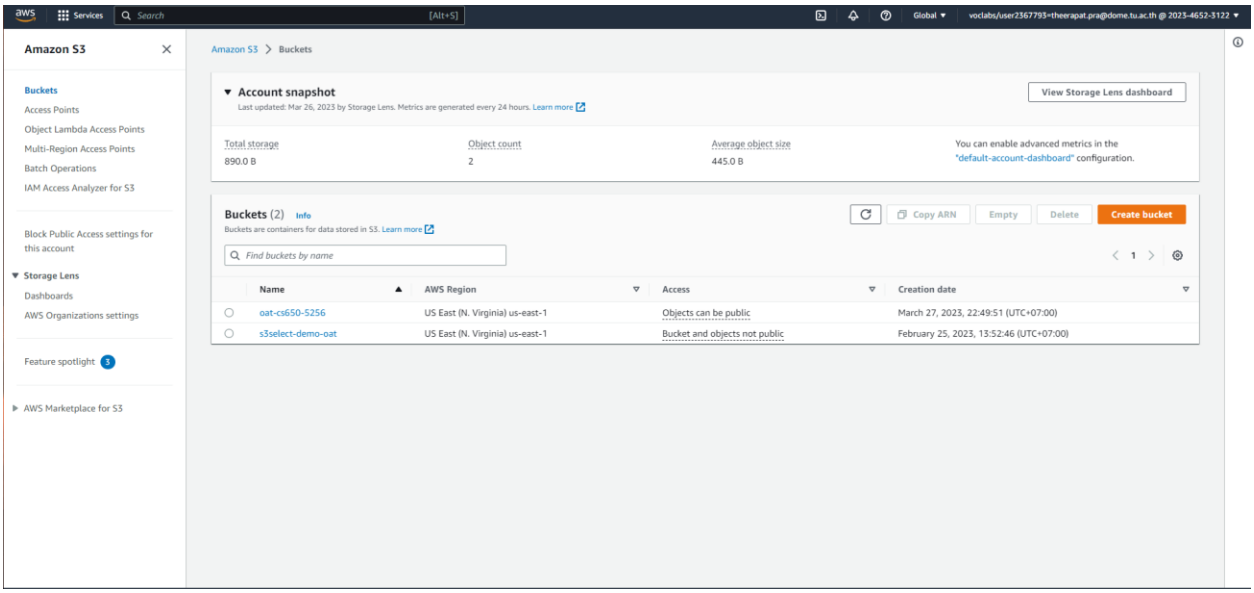
สร้าง GitHub Repo ชื่อ hw2_5256 ขึ้นมา และทำการ Create Notebook Instances ขึ้นมา เมื่อสามารถใช้งานได้แล้วให้ทำการ Open Jupyter ขึ้นมา พร้อมเปลี่ยนชื่อไฟล์เป็น hw2_5256.py

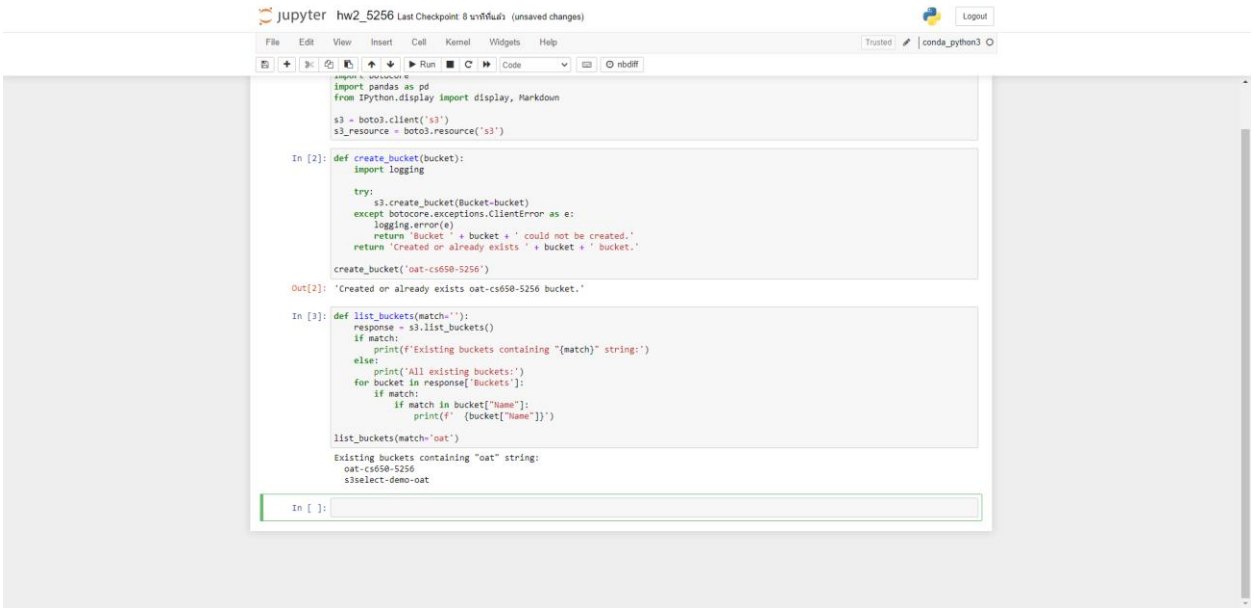
ขั้นตอนการทำงานของไฟล์ hw2_5256.py

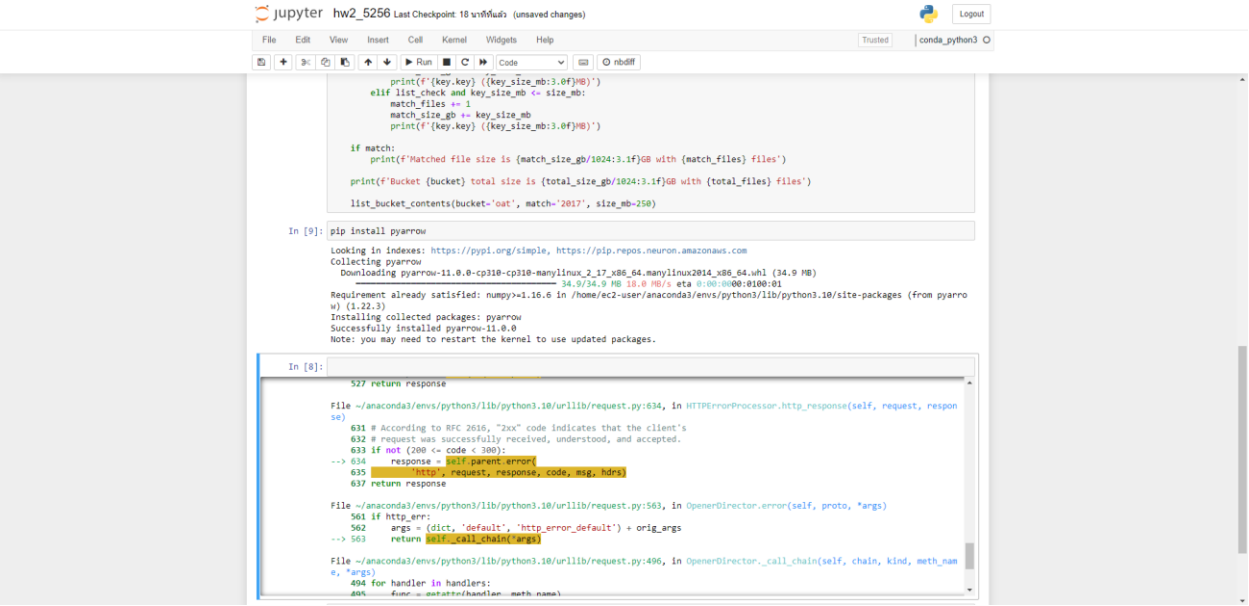


ชื่อ นายธีรภัทร์ ประจำทอง รหัสนักศึกษา 6509035256

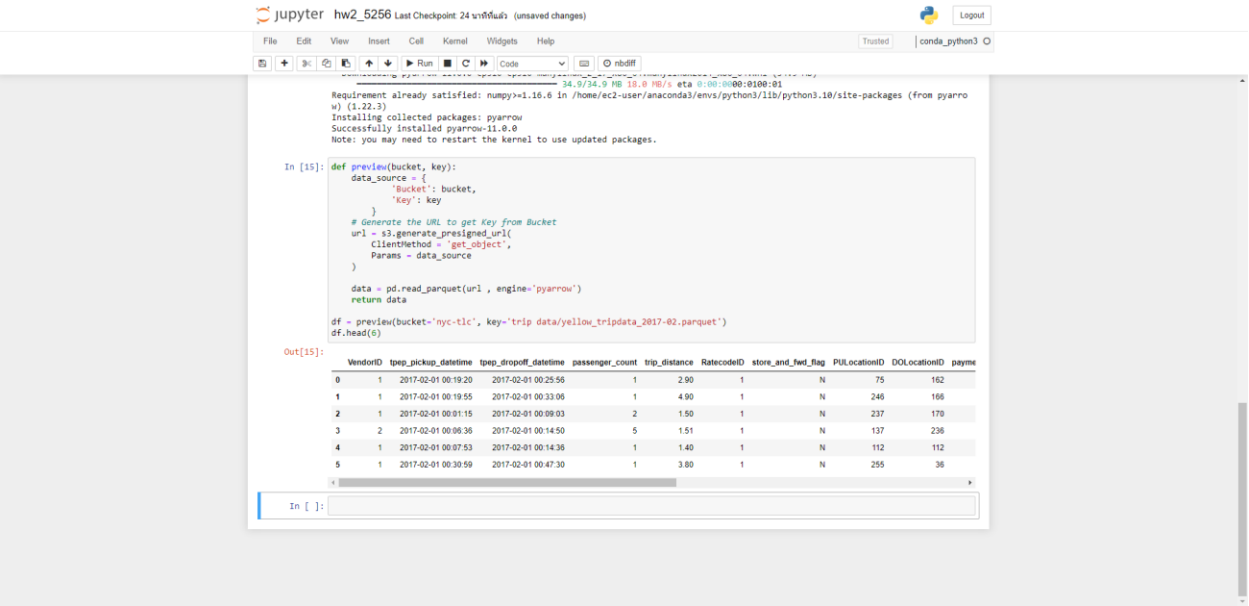
ทำการ import boto3 library และ botocore เป็นไลบรารี่ที่จำเป็นในการติดต่อกับ AWS S3 จากนั้นสร้าง Bucket ชื่อว่า oat-cs653-5033 ขึ้นมาเมื่อรันแล้วจะมี output ออกมา ‘''Created or already exists oat-cs650-5256 bucket.'’ ซึ่งจะขึ้นอยู่ที่ในส่วนของ Bucket แล้ว



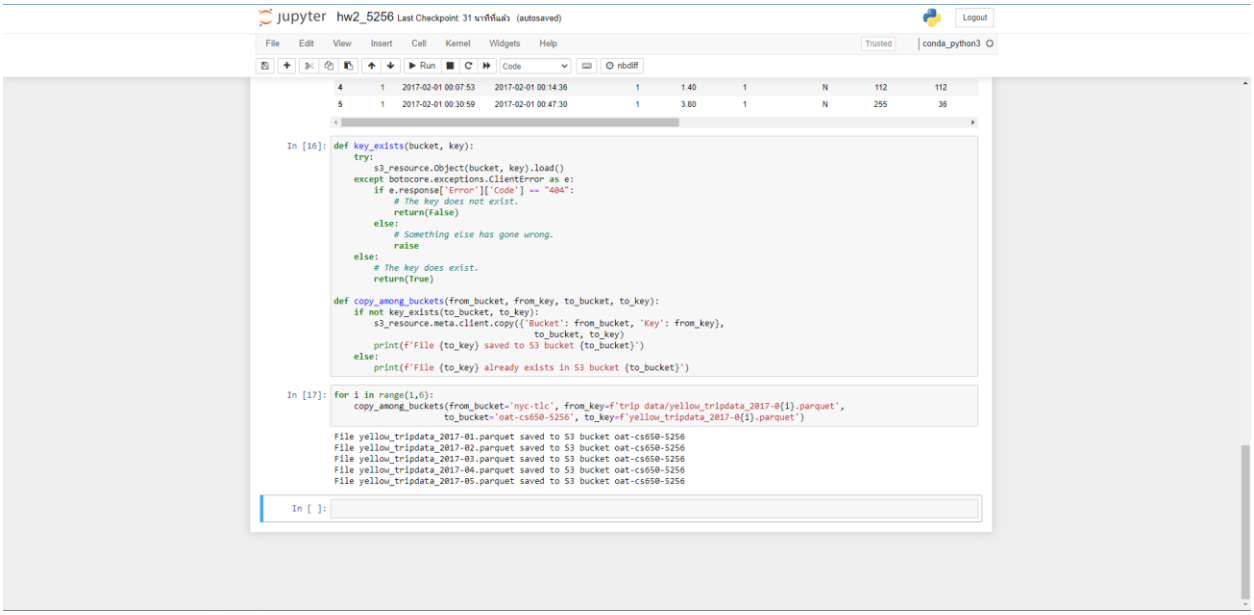




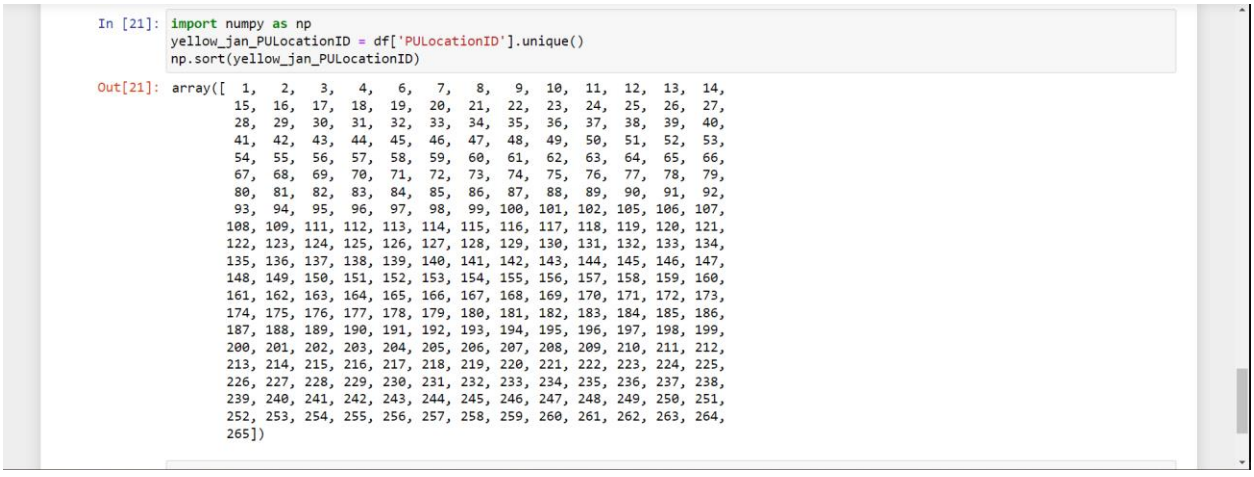
เนื่องจากไฟล์ข้อมูลแท็กซี่ประเภท yellow มีชนิดเป็น parquet จึงต้อง import ไลบรารี่ pyarrow ในไลบรารี่ดังกล่าวมีคำสั่งที่ช่วยในการอ่านไฟล์สกุล parquet เข้ามาใน dataFrame เราไม่จำเป็นต้องแปลงชนิดไฟล์จาก parquet เป็น csv เพราะว่า S3 Select รองรับการอ่านไฟล์ parquet อยู่แล้ว



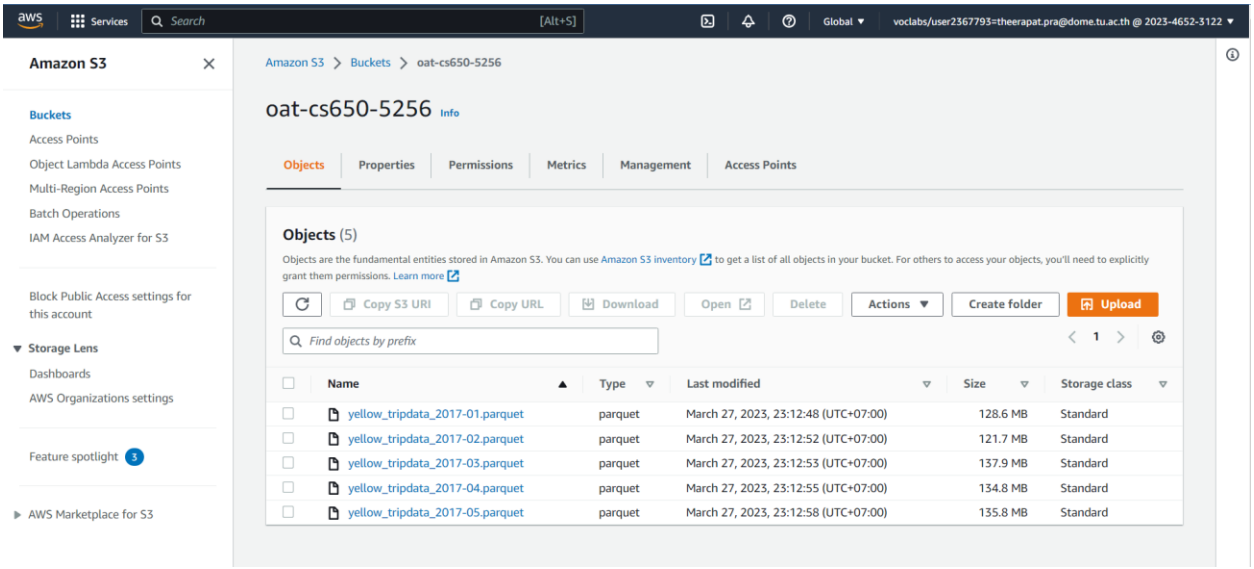
เมื่อลอง preview ข้อมูลแท็กซี่ yellow ในเดือน 01 ปี 2017 เพื่อสำรวจค่า attribute ต่าง ๆ ในที่นี้เราสนใจ payment_type



ทำสำเนา Bucket “nyc-tlc” จาก open dataset ของ AWS มายัง S3 Bucket ส่วนตัวของเราฟังก์ชัน key_exits เช็คว่าชื่อและคีย์ของ Bucket ส่วนตัวของเราถูกต้อง หรือมีอยู่จริงหรือไม่ถ้ามีอยู่จริงแล้ว ฟังก์ชัน copy_among_buckets จะทำสำเนาไฟล์มายัง Bucket ของเราไม่จำเป็นต้องเลือกทำสำเนาทุกไฟล์ เลือกเฉพาะไฟล์ที่จำเป็นในการทำการบ้านเท่านั้น มี 5 ไฟล์ ได้แก่ ไฟล์ yellow_tripdata_2017 ของตั้งแต่เดือน 01 ถึง 05



ชื่อ นายธีรภัทร์ ประจำทอง รหัสนักศึกษา 6509035256



ตอบคำถามข้อ a) ถึง c)

A. ในเดือน Jan 2017 มีจำนวน yellow taxi rides ทั้งหมดเท่าไร แยกจำนวน rides ตามประเภทการจ่ายเงิน (payment)

ใช้คำสั่งของ pandas เพื่อหาค่าทั้งหมดที่เป็นไปได้ของข้อมูล payment_type นั่นคือคำสั่ง dataFrame.['payment_type'].unique() ซึ่งคืนค่ามาว่า array([2, 1, 3, 4, 5]) แปลว่าค่าของ payment_type มีค่าที่แตกต่างกัน 5 ค่า

ผลลัพธ์

```
import boto3
s3 = boto3.client('s3')
sum = 0
for i in range(1,6):
    resp = s3.select_object_content(
        Bucket='oat-cs650-5256',
        Key='yellow_tripdata_2017-01.parquet',

        ExpressionType='SQL',

        Expression=f"SELECT COUNT(payment_type) FROM s3object s WHERE payment_type= {i}",
        InputSerialization={'Parquet': {}},
        OutputSerialization = {'CSV': {}},
    )

    for event in resp['Payload']:
        if 'Records' in event:
            records = event['Records']['Payload'].decode('utf-8')
            sum = sum + int(records)
            print(f"จำนวน yellow taxi ที่มี payment_type={i} เท่ากับ {records}")
            print(f"มี yellow taxi รวมทั้งสิ้น {sum} คัน")
```

```
จำนวน yellow taxi ที่มี payment_type=1 เท่ากับ 6506189

มี yellow taxi รวมทั้งสิ้น 6506189 คัน
จำนวน yellow taxi ที่มี payment_type=2 เท่ากับ 3144926

มี yellow taxi รวมทั้งสิ้น 9651115 คัน
จำนวน yellow taxi ที่มี payment_type=3 เท่ากับ 46257

มี yellow taxi รวมทั้งสิ้น 9697372 คัน
จำนวน yellow taxi ที่มี payment_type=4 เท่ากับ 13447

มี yellow taxi รวมทั้งสิ้น 9710819 คัน
จำนวน yellow taxi ที่มี payment_type=5 เท่ากับ 1

มี yellow taxi รวมทั้งสิ้น 9710820 คัน
```

B. ในเดือน Jan 2017 Yellow taxi rides ในแต่ละจุดรับผู้โดยสาร (Pickup location) เป็นจำนวน rides มากน้อยเท่าไร และมีค่าโดยสารรวมของ rides และจำนวนผู้โดยสารเฉลี่ยต่อ rides ในแต่ละจุดเท่าไร

เนื่องจากคำสั่ง DISTINCT ไม่สามารถใช้กับ S3 Select จึงใช้คำสั่งของ pandas เพื่อหาค่าทั้งหมดที่เป็นไปได้ของข้อมูล payment_type นั่นคือคำสั่ง dataFrame.['payment_type'].unique() แล้วจัดเรียงค่าจากน้อยไปหามาก มีการคืนค่ามา 265 ค่าดังภาพ แปลว่ามีจุดรับผู้โดยสารรวม 265 แห่ง

ผลลัพธ์

```
In [21]: import numpy as np
         yellow_jan_PULocationID = df['PULocationID'].unique()
         np.sort(yellow_jan_PULocationID)

Out[21]: array([  1,   2,   3,   4,   6,   7,   8,   9,  10,  11,  12,  13,  14,
                 15,  16,  17,  18,  19,  20,  21,  22,  23,  24,  25,  26,  27,
                 28,  29,  30,  31,  32,  33,  34,  35,  36,  37,  38,  39,  40,
                 41,  42,  43,  44,  45,  46,  47,  48,  49,  50,  51,  52,  53,
                 54,  55,  56,  57,  58,  59,  60,  61,  62,  63,  64,  65,  66,
                 67,  68,  69,  70,  71,  72,  73,  74,  75,  76,  77,  78,  79,
                 80,  81,  82,  83,  84,  85,  86,  87,  88,  89,  90,  91,  92,
                 93,  94,  95,  96,  97,  98,  99, 100, 101, 102, 105, 106, 107,
                108, 109, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121,
                122, 123, 124, 125, 126, 127, 128, 129, 130, 131, 132, 133, 134,
                135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147,
                148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160,
                161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173,
                174, 175, 176, 177, 178, 179, 180, 181, 182, 183, 184, 185, 186,
                187, 188, 189, 190, 191, 192, 193, 194, 195, 196, 197, 198, 199,
                200, 201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211, 212,
                213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225,
                226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238,
                239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251,
                252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264,
                265])
```

C. ในเดือน Jan - Mar 2017 มีจำนวน yellow taxi rides ทั้งหมดเท่าไร แยกจำนวน rides ตามประเภทการจ่ายเงิน (payment)

ทำการวนลูป 5 ครั้งซึ่งเป็นจำนวนเดือนตั้งแต่เดือนมกราคมถึงมีนาคม ในแต่ละเดือนมีการวนลูป 5 ครั้ง ซึ่งเป็นจำนวนค่าที่เป็นไปได้ของ payment_type และนำ list ทั้ง 3 ตัวมาสร้าง dataFrame พร้อมแสดงจำนวนเที่ยวทั้งหมดแบ่งตามประเภทการจ่ายค่าโดยหารในเดือนแต่ละเดือน

ผลลัพธ์

	month	payment type 1	payment type 2	payment type 3	payment type 4	payment type 5	sum
0	Jan	6506189	3144926	46257	13447	1	9710820
1	Feb	6261976	2849713	44719	13367	0	9169775
2	Mar	6994699	3231928	53815	14999	0	10295441
3	April	6695495	3281576	54383	15680	1	10047135
4	May	6780947	3250362	55027	15791	0	10102127

ชื่อ นายธีรภัทร์ ประจำทอง รหัสนักศึกษา 6509035256

การสะท้อนการเรียนรู้ของน.ศ.จากการบ้านครั้งนี้

1. เราได้ความรู้และทักษะอะไรจากการทำการบ้านครั้งนี้บ้าง และคิดว่านำไปใช้ประโยชน์อย่างไรได้บ้าง
   ได้ความรู้การใช้งาน Jupyter และการใช้งานของ AWS และได้ฝึกการ queue ข้อมูล
2. สิ่งที่เราชอบและไม่ชอบในการทำการบ้านครั้งนี้
   สิ่งที่ชอบคือได้ลองใช้ tool ใหม่ซึ่งไม่เคยลองใช้เลย เป็นประสบการณ์ที่ดี
3. คิดว่าตัวเองควรปรับปรุงอย่างไร หรือ มีอะไรอย่างอื่นที่ควรได้รับการปรับปรุงสำหรับการบ้านครั้งต่อไป
   ควรฝึกและลองทำเรื่อยๆ เพื่อให้เข้ามากขึ้น

URL ของไฟล์ hw2 _5256.py ที่ commit ไว้ใน github repo ส่วนตัวของน.ศ. ซึ่งเปิดให้อาจารย์ (rattanat@gmail.com) และ TA (ta1tonkit@gmail.com) สามารถเข้าถึงได้

***ระมัดระวังอย่าใส่ AWS credential ไว้บนไฟล์ใดๆ ของ repo เด็ดขาด ***

https://github.com/Oat123/hw2_5256