มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

วิชา CSS 341 Introduction to Data Science สำหรับ นศ. ภาควิชาคณิตศาสตร์

ข้อสอบกลางภาคเรียนที่ 1 ปีการศึกษา 2564 วันพฤหัสบดีที่ 30 กันยายน 2564 เริ่มต้นเวลา 12:00 น.

คำแนะนำและคำสั่ง

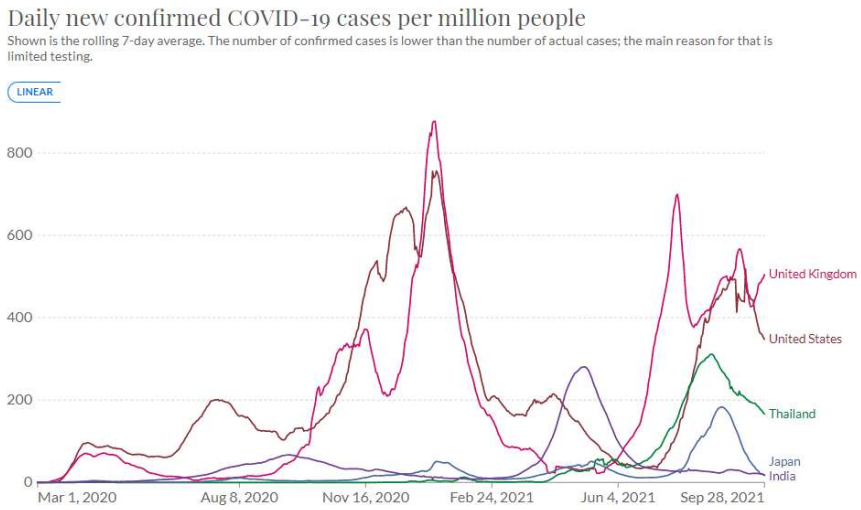
ข้อสอบมี 3 ส่วน (Parts) คิดเป็น 9, 25, และ 6 คะแนน รวม 40 คะแนน ให้ทำทุกข้อ โดยที่....

1. เขียน Python 3.8++ ใน ipynb file เดียว**ต่อเนื่อง**ไปเลย ทั้ง 3 Parts
2. Data files ทุกไฟล์อยู่ใน folder ที่อยู่ข้างๆ คือระดับเดียวกับ folder ที่บรรจุ ipynb ของท่าน ดังนั้น เวลาอ่าน data file ให้กำหนด path เป็น '`../input/dddd.eee`' เมื่อ `dddd.eee` คือชื่อไฟล์ที่อ่านเข้ามา หากผิดกติกานี้จะถูกหัก 2 แต้มต่อการอ่าน data file 1 ครั้ง
3. เขียน Markdown ให้เหมาะสมชัดเจน นั่นคือ สำหรับแต่ละข้อให้ มีโจทย์ (สั้นๆ) ตามด้วยโค้ด และตามด้วยการวิเคราะห์ผลลัพธ์เพื่อสรุปตอบ เรียงลำดับข้อไป
4. การส่ง ให้ส่งขึ้น LEB2 ด้วย file ipynb เท่านั้น. ไม่ต้องส่ง data file มาเลย เพราะมีอยู่แล้ว กลุ่มที่ส่งสายเกินกำหนดจะถูกหัก 5 นาทีละ 1 คะแนน
5. ไฟล์ ipynb ที่ส่ง ให้ตั้งชื่อไฟล์เป็นชื่อต้นของสมาชิกคนแรก (ภาษาอังกฤษ) และส่วนต้นของ ipynb file ที่ส่งจะต้องมี รายชื่อและ student ID ของสมาชิกทุกคน ชัดเจน (เขียนเป็น Markdown ไว้)

Part 1

จากข้อมูล Covid-19.csv ที่ให้ไป ให้วิเคราะห์เปรียบเทียบ

1. การเปลี่ยนแปลง (คิดเป็นร้อยละ) ของจำนวนผู้ป่วย**รายสองสัปดาห์**ล่าสุด ว่าขึ้นหรือลงกี่ % เทียบกับสองสัปดาห์ก่อนหน้า โดยแสดงค่าและชื่อประเทศนั้น ๆ ที่เพิ่มสูงสุด 10 ประเทศ และลดลงมากสุด 10 ประเทศ (5 คะแนน)
2. จำนวนผู้ป่วยใหม่รายวันของแต่ละประเทศโดยเริ่มตั้งแต่วันที่ Mar 1, 2020. ให้เลือกรายชื่อประเทศที่ต้องการ (ใน List) และวันเริ่มต้นได้โดยใส่ไว้ต้นของโค้ด Cell นั้น ผลลัพธ์ดังรูปตัวอย่างต่อไปนี้ (3 คะแนน)



Part 2

File Descriptions

- Asales.csv – the training set. Daily historical data from January 2013 to October 2015.
- items.csv – supplemental information about the items/products.
- item_categories.csv – supplemental information about the items categories.
- shops.csv – supplemental information about the shops.

Data Fields

- ID – an ID that represents a (Shop, Item) tuple within the test set
- shop_id – unique identifier of a shop
- item_id – unique identifier of a product
- item_category_id – unique identifier of item category
- item_cnt_day – number of products sold.
- item_price – current price of an item
- date – date in format dd/mm/yyyy
- date_block_num – a consecutive month number, used for convenience. January 2013 is 0, February 2013 is 1,..., October 2015 is 33
- item_name – name of item
- shop_name – name of shop
- item_category_name – name of item category

Problems

3. ในข้อมูล Asales.csv ให้ตัดรายการที่มีค่า item_cnt_day หรือ item_price น้อยกว่า 0 และแสดงสรุปข้อมูลให้เห็นในภาพรวมว่าตัดออกไปแล้ว (1 คะแนน)
4. แสดงค่าเฉลี่ย, Median และพิสัยระหว่างควอร์ไทล์ (Interquartile Range) ของข้อมูล item_price (1 คะแนน)
5. รวมข้อมูลในไฟล์ *items* dataset เข้าไปใน *sales* dataset (1 คะแนน)
6. เปลี่ยนชนิดข้อมูล *date* จาก *string* เป็นวันที่ *date* เพื่อการวิเคราะห์ต่อไป (1 คะแนน)

7. มีทั้งหมดกี่ items จากทุก shop รวมกัน (1 คะแนน)
8. item ใดที่ขายได้มากที่สุดในแต่ละ shop โดยแสดง shop_id, item_id และจำนวนที่ขายได้สูงสุด 10 shops พอ (4 คะแนน)
9. ให้แสดงค่าเฉลี่ย (รวมทั้ง s.d., max, min) ของราคาขายสินค้าของแต่ละร้านค้า โดยแสดงเพียงร้านที่มีค่าเฉลี่ยสูงสุด 5 อันดับและต่ำสุด 5 อันดับ (เรียงลำดับจากมากไปน้อย) (4 คะแนน)
10. ให้นำยอดขายของทุกร้านมาเรียงจากมากไปน้อย แล้วแบ่งจัดกลุ่มร้านทั้งหมดออกเป็น 6 กลุ่ม (ตามจำนวนร้านโดยปัดเศษตามสมควร) จากนั้นให้แสดง*ยอดขายรวมของกลุ่ม*ทั้ง 6 กลุ่มนั้นเปรียบเทียบกัน (เป็นรูปสัดส่วนอย่างสวยงามเหมาะสม) (6 คะแนน)
11. ให้แสดงจำนวนสินค้าที่ขายได้รวมในแต่ละวันของสัปดาห์ และ ยอดขายรวมในแต่ละวันของสัปดาห์ โดยนำเสนอข้อมูลทั้งสองนี้ข้าง ๆ กัน โดยเอาเฉพาะปี 2015 ปีเดียว (6 คะแนน)

Part 3

File Description of Part 3

ไฟล์ Bsales.csv เป็นข้อมูลยอดขายของบริษัทแห่งหนึ่ง (สมมติชื่อ B) โดยลักษณะข้อมูลดังต่อไปนี้

- Year ปี (ค.ศ.)
- Product ชื่อย่อสินค้า
- Rep ชื่อย่อของพนักงานขาย (Sales representative)
- Type ประเภทสินค้า
- North, South, East, West ยอดขายแยกรายภาคของสินค้าและพนักงานขายในปีที่ระบุ (หน่วยเป็นบาท)

Problems

12. ยอดขายรวมในภาคเหนือและภาคใต้ของพนักงานขายแต่ละคน โดยพิจารณาเฉพาะสินค้าสองตัวคือ 360 และ PS3 และเฉพาะปี 2018 และ 2019 และให้แสดงผลคอลัมน์จัดกลุ่มเป็น 2 ชั้นตามคือปีและสินค้า ดังตัวอย่างผลลัพธ์ทำนองนี้ (6 คะแนน)

Rep	2018				2019				Total Sum of North	Total Sum of South
	360		PS3		360		PS3			
	Sum of North	Sum of South	Sum of North	Sum of South	Sum of North	Sum of South	Sum of North	Sum of South		
5p			0.00	0.00					0.00	0.00
Ac	1.89	0.92	0.96	1.52	0.02	0.03	0.01	0.04	2.88	2.51
AlSo					0.00	0.02	0.00	0.04	0.00	0.06
AqPl			0.00	0.00			0.00	0.00	0.00	0.00
	.	.	.							
TrBlEn	0.00	0.03	0.00	0.03					0.00	0.06
Ub	0.28	0.15	0.10	0.14					0.38	0.29
Un							0.00	0.00	0.00	0.00
WaBrInEn	0.71	0.68	0.49	0.74	0.21	0.21	0.13	0.24	1.54	1.87
Total	7.11	4.77	4.76	7.43	0.36	0.40	0.40	0.80	12.63	13.40