มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี

วิชา CSS 341 Introduction to Data Science สำหรับ นศ. ภาควิชาคณิตศาสตร์

ข้อสอบครั้งที่ 1 ปีการศึกษา 2565 วันอังคารที่ 9 กันยายน 2565 เริ่มต้นเวลา 12:00 น.

คำแนะนำและคำสั่ง

ข้อสอบมี 2 ส่วน (Parts) รวม 12 ข้อ คะแนน 19 + 12 รวม 31 คะแนน ให้ทำทุกข้อ โดยที่....

1. เขียน Python 3.8 หรือ 3.9 ในรูปแบบ ipynb **ไฟล์เดียวต่อเนื่อง**ไปเลยทุกข้อ
2. Data files ทุกไฟล์อยู่ใน folder ที่อยู่ข้างๆ คือระดับเดียวกับ folder ที่บรรจุ ipynb ของท่าน ดังนั้น เวลาอ่าน data file ให้กำหนด path เป็น '`../data?/???.???`' เมื่อ `???.???` คือชื่อไฟล์ที่อ่านเข้ามา หากผิดกติกานี้จะถูกหัก 2 แต้มต่อการอ่าน data file 1 ครั้ง
3. เขียน Markdown ให้เหมาะสมชัดเจน นั่นคือ สำหรับแต่ละข้อให้ ควรมีโจทย์ (ย่อมาสั้นๆ) ตามด้วยโค้ด และตามด้วยการวิเคราะห์ผลลัพธ์เพื่อสรุปตอบ เรียงลำดับข้อไป
4. การส่ง ให้ส่งขึ้น LEB2 ด้วย file ipynb เท่านั้น. ไม่ต้องส่ง data file มา เพราะมีอยู่แล้ว กลุ่มที่ส่งสายเกินกำหนดจะถูกหัก 3 นาทีละ 1 คะแนน
5. ไฟล์ ipynb ที่ส่ง ให้ตั้งชื่อไฟล์ในรูปแบบ YY-aa-bb-cc-dd-ee เมื่อ YY บอกชั้นปี 63 หรือ 64 ส่วน aa, bb, ... เป็น student ID สองหลักท้าย (เรียงลำดับมาด้วย) ของสมาชิกทุกคน รวมทั้งการตั้งชื่อกลุ่ม (ถ้ามี) ตอนส่งงานเข้า LEB2 ก็ตั้งชื่อแบบเดียวกัน
6. **ส่วนต้นสุด**ของ ipynb file ที่ส่งจะต้องมี student ID และตามด้วยรายชื่อของสมาชิกทุกคนชัดเจน (เขียนเป็น Markdown ไว้) หากไม่มีจะถูกหัก 3 คะแนน

Part 1 (19 คะแนน)

Data description

- sales.csv – 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

1. ในข้อมูล sales.csv ให้ตัดรายการที่มีค่า item_cnt_day หรือ item_price น้อยกว่า 0 และแสดงสรุปข้อมูลให้เห็นในภาพรวมว่าตัดออกไปแล้ว (1 คะแนน)
2. แสดงค่าเฉลี่ย, Median และพิสัยระหว่างควอร์ไทล์ (Interquartile Range) ของข้อมูล item_price (1 คะแนน)
3. รวมข้อมูลในไฟล์ *items* dataset เข้าไปใน *sales* dataset (1 คะแนน)
4. เปลี่ยนชนิดข้อมูล *date* จาก *string* เป็นวันที่ *date* เพื่อการวิเคราะห์ต่อไป (1 คะแนน)
5. มีทั้งหมดกี่ items จากทุก shop รวมกัน (1 คะแนน)
6. item ใดที่ขายได้มากที่สุดในแต่ละ shop โดยแสดง shop_id, item_id และจำนวนที่ขายได้สูงสุด 10 shops พอ (4 คะแนน)
7. ให้แสดงค่าเฉลี่ย (รวมทั้ง s.d., max, min) ของราคาขายสินค้าของแต่ละร้านค้า โดยแสดงเพียงร้านที่มีค่าเฉลี่ยสูงสุด 5 อันดับและต่ำสุด 5 อันดับ (เรียงลำดับจากมากไปน้อย) (4 คะแนน)
8. ให้แสดงจำนวนสินค้าที่ขายได้รวมในแต่ละวันของสัปดาห์ และ ยอดขายรวมในแต่ละวันของสัปดาห์ โดยเอาเฉพาะปี 2013 และ 2014 และเทียบผลของสองปีนี้เคียงข้างกัน (ซ้ายเป็นของ 2013, ขวาเป็น 2014) (6 คะแนน)

Part 2 (12 คะแนน)

Data description

- customers.csv – Daily historical data from January 2013 to October 2015.
- tx2010.csv to tx2015 – Transaction file of each year
- zip_to_state_map.csv

Problems

9. ให้นำเข้า (load) ไฟล์ tx20xx ทั้ง 6 ไฟล์ จากนั้นรวมต่อเข้าด้วยกันเป็นข้อมูลเดียว และเตรียมข้อมูลให้เหมาะสม เพื่อใช้วิเคราะห์และนำเสนอข้อมูลรวมทั้ง 6 ปีนี้ ในแง่มุมต่างๆ ในข้อที่ **Error! Reference source not found.** ถึง 12 (1 คะแนน)
10. ยอดขายรวม (หลังหักส่วนลด) ของลูกค้าแต่ละคน (4 คะแนน)
    - โดยเรียงลำดับจากมากไปน้อย พร้อมระบุ zipcode และรัฐ (ตามหลัง)
    - และแสดงเฉพาะลูกค้ายอดขาย top 20 และ bottom 20 พอ
11. ยอดขายรวม (หลังหักส่วนลด) ของแต่ละรัฐ (6 คะแนน)
    - โดยที่สำหรับแต่ละรัฐให้แยกรายปี (เรียงปีจากน้อยไปมาก)
    - และเรียงรัฐจากยอดขายมากไปน้อย (ให้แสดงยอดรวมของแต่ละรัฐด้วย)
12. จากข้อที่แล้ว ให้แสดงเฉพาะรัฐที่เป็น top 15 และ bottom 15 พอ (1 คะแนน)