ข้อสอบปฏิบัติการวิชา 01418323
Introduction to Data Science

ภาคปลาย ปีการศึกษา 2561

อ.ดร.เสฏฐวิทย์ เกิดผล

ข้อกำหนดในการสอบ

1. เข้าสอบให้ตรงเวลา ถ้าเข้าสายกว่า 15 นาทีหลังเริ่มสอบจะหมดสิทธิ์สอบ
2. ให้นำเอกสารชีทแล็บเข้าได้ สามารถบันทึกโน้ตในชีทแล็บได้
3. ไม่ให้รับส่งหรือนำเข้าไฟล์อิเล็คโทรนิคส์ใดๆ ยกเว้นไฟล์ที่จะให้ดาวน์โหลดก่อนเริ่มสอบเท่านั้น

การบันทึกไฟล์คำตอบ

1. ให้ตั้งชื่อไฟล์ด้วยรหัสนิสิต
2. นิสิตสามารถทำไฟล์คำตอบเป็นไฟล์ .py (ไพธอน) หรือ .ipynb (Jupyter Notebook) ก็ได้
3. ให้บันทึกไฟล์ไว้ที่ไดรฟ D: เพื่อป้องกันไม่ให้ไฟล์หาย

การตอบคำถาม

1. ให้ใส่ชื่อและรหัสนิสิตเป็นคำตอบข้อ 0
2. การตอบคำถามแต่ละข้อให้ใส่คอมเม้นต์เบอร์ข้อไว้ก่อนเริ่มตอบ
3. ต้องแสดงผลทุกครั้งที่จะตอบ
4. ให้แสดงผลตามที่ถาม เช่นถามจำนวนก็ตอบเป็นตัวเลข ถามว่ามีอะไรบ้างก็ต้องตอบเป็นชื่อ
5. การโหลดไฟล์ให้นำไฟล์มาใส่ไว้ใน directory เดียวกันกับไฟล์คำตอบ
6. นิสิตควรศึกษาข้อมูลในไฟล์ที่จะนำมาวิเคราะห์ก่อนจะเริ่มตอบคำถาม

คำถาม

1. โหลดไฟล์ kc_house_data.csv ขึ้นมาใส่ตารางชื่อ house_sales ซึ่งเป็นข้อมูลการขายบ้านใน King County มลรัฐ Washington จงแสดงจำนวนแถวและคอลัมน์ของตาราง
2. แสดงจำนวนข้อมูลที่เป็น NaN ในแต่ละคอลัมน์
3. แสดงค่าที่ไม่ซ้ำกันทั้งหมดในคอลัมน์ bathrooms
4. แสดงแถวที่ราคาบ้าน (price) มากกว่า $500,000
5. แสดงแถวที่ราคาบ้านน้อยกว่า $100,000 แต่พื้นที่บ้าน (sqft_living) ไม่น้อยกว่า 800 sq.ft.
6. เรียงลำดับข้อมูลตามพื้นที่บ้านจากน้อยไปมาก ถ้าขนาดเท่ากันให้เรียงราคาบ้านจากมากไปน้อย
7. [ยาก] เราต้องการจัดกลุ่มบ้านที่ขายแต่ละหลังตามรหัสไปรษณีย์ (zipcode) แต่จำนวนรหัสไปรษณีย์นั้นมากเกินไป เราจึงต้องการจัดกลุ่มตามรหัสไปรษณีย์ 3 ตัวแรก ให้สร้างคอลัมน์ใหม่ชื่อ Zip3 และใส่รหัสไปรษณีย์ 3 ตัวแรกที่คอลัมน์นี้ จากนั้นแสดงจำนวนบ้านที่ขายไปในแต่ละ Zip3
8. แสดงค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐานของราคาบ้าน
9. บ้านที่มีห้องน้ำเยอะที่สุดมีกี่ห้องน้ำ
10. แสดงจำนวนบ้านแยกตามจำนวนห้องนอน
11. สร้างตารางแสดงจำนวนบ้านแยกตาม grade และ zipcode จากนั้นแสดงตารางนี้
12. [ยาก] แสดงจำนวนของบ้านแต่ละเกรดเรียงลำดับจำนวนบ้านจากมากไปน้อย จัดกลุ่มตาม zipcode

13. [ยาก] หาเกรดของบ้านที่มีจำนวนเยอะที่สุดในแต่ละ zipcode แสดงเกรดและจำนวนบ้านที่มากที่สุดแยกตาม zipcode
14. ไฟล์ bld_cond_desc.csv เก็บคำอธิบาย condition ของบ้านไว้ตามคะแนนที่ได้ ให้นำคำอธิบายจากไฟล์มาใส่ในตารางแทนคะแนน (ห้ามใส่คำอธิบายเอง) จากนั้นแสดงคอลัมน์ 'price','sqft_living','condition' ของ 5 แถวแรก
15. เปลี่ยนคำอธิบาย condition ให้เป็น category และให้ประเภทจัดเรียงตามลำดับจากแย่ไปดี จากนั้นแสดง 5 แถวแรกของคอลัมน์นี้ (pandas จะแสดงรายละเอียด category ให้ด้วย)
16. มีบ้านกี่หลังที่ปีที่ขาย (date) เกิดขึ้นอย่างน้อย 60 ปีหลังจากการสร้าง (yr_built)
17. มีบ้านกี่หลังที่ขายได้ระหว่างวันที่ 1 มิถุนายน 2014 ถึง 28 กุมภาพันธ์ 2015
18. แสดงค่าเฉลี่ยของราคาบ้านที่ขายได้ในแต่ละไตรมาส
19. โหลดไฟล์ชื่อ King_County_Sheriff_s_Office.csv ซึ่งเก็บการปฏิบัติงานของเจ้าหน้าที่ตำรวจไว้ จงสร้างตารางแสดงจำนวนอาชญากรรมผิดกฎหมายแยกตามประเภทของการทำผิดกฎหมาย (ในคอลัมน์ incident_parent_type) และตามรหัสไปรษณีย์ (zip) จากนั้นแสดงตารางนี้
20. เราสนใจอาชญากรรมที่เกี่ยวข้องกับบ้าน ดังนั้นเราจะเน้นที่อาชญากรรมสามประเภทคือ 'Property Crime','Breaking & Entering' กับ 'Drugs' ให้คำนวณตารางใหม่ที่แสดงว่าแต่ละ zip นั้นมีอาชญากรรมทั้งสามเกิดขึ้นคิดเป็นกี่เปอร์เซ็นเทียบกับอาชญากรรมประเภทนั้นรวมในทุก zip และแทนตำแหน่งที่ไม่มีข้อมูลด้วย 0.0 ด้วย แสดงตารางผลลัพธ์
21. นำข้อมูลเปอร์เซ็นอาชญากรรมทั้งสามประเภทไปรวมกับตารางการขายบ้าน (ตารางแรก) ที่ตำแหน่งรหัสไปรษณีย์เดียวกัน แสดงตาราง
22. พล็อต boxplot ของราคาบ้าน (price) แยกตาม grade
23. พล็อต scatter plot ของราคาบ้าน (price) เทียบกับพื้นที่ใช้สอย (sqft_living) และความติดน้ำของบ้าน (waterfront)
24. สร้าง barplot เทียบราคาเฉลี่ยของบ้านที่ขายได้ในแต่ละเดือน แยกตามจำนวนชั้นของบ้าน
25. [Freestyle] ให้นิสิตคิดวิธีการสรุปข้อมูลที่แสดงประเด็นที่น่าสนใจจากข้อมูลในการสอบครั้งนี้ และแสดงผลการสรุปนั้นด้วยภาพ คะแนนจะคิดจากความน่าสนใจของการสรุปข้อมูลเป็นหลัก