Chapter 3

Google Sheet For Data Science

Google Sheet คือ อะไร

- แอปสเปรดชีตออนไลน์สำหรับสร้างและจัดรูปแบบสเปรดชีตและทำงานร่วมกับคนอื่น ๆ

ไฟล์ Lab

Part 1 คลิกขวาเปิดลิงก์ด้านล่างใน browser tab ใหม่ >>

https://docs.google.com/spreadsheets/d/1MHHvY2wv_tD68UaUfEg8Ma4ePW9n6Ms15FPPAzpFnSE/edit?usp=sharing

Step 1 ตรวจสอบข้อมูลเบื้องต้น (Assess)

ด้วย Google Sheet

ตรวจสอบข้อมูล เบื้องต้นด้วย Google Sheet

✓นับจำนวนคอลัมภ์ (No. Cols)

=columns(ช่วงข้อมูล)

✓นับจำนวนแถว (No. Rows)

=rows(ช่วงข้อมูล)

ตรวจสอบข้อมูล เบื้องต้นด้วย Google Sheet

✓นับค่าซ้ำ (Duplicated)

=countif(คอลัมภ์ที่ต้องการนับ, ตำแหน่งข้อมูลที่ต้องการหาค่าซ้ำ)

✓นับค่าที่หายไป (Missing Values)

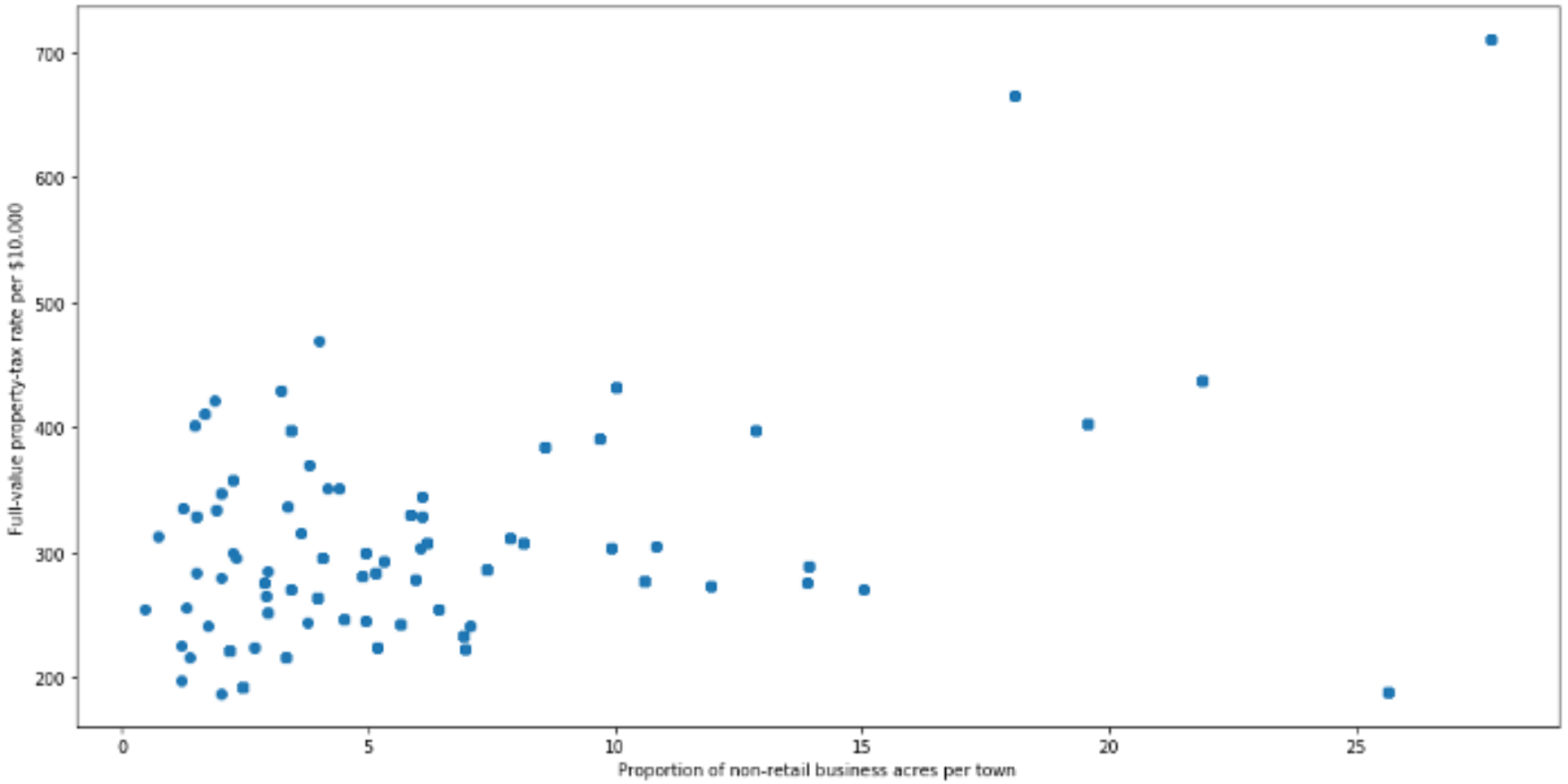
=countblank(แถวที่ต้องการหา Missing Values)

Outliers คือ ค่าผิดปกติ

ค่าผิดปกติ (Outliers) เป็นข้อมูลที่มีค่าแยกออกจากกลุ่มหรือผิดแผกแตกต่างไปจากข้อมูลค่าอื่น ๆ

เช่น IQ ของเด็กได้ 195 น้ำหนักของคน 220 กิโลกรัม ความสูงของคน 210 ซ.ม.

ตัวอย่างค่าผิดปกติ



สาเหตุส่วนใหญ่ของค่าผิดปกติในชุดข้อมูล

- ✓ข้อผิดพลาดในการป้อนข้อมูล (ข้อผิดพลาดจากมนุษย์)
- ✓ข้อผิดพลาดในการวัด (ข้อผิดพลาดของเครื่องมือ)
- ✓ข้อผิดพลาดในการทดลอง (การดึงข้อมูล หรือการวางแผนการทดลอง / การดำเนินการผิดพลาด)
- ✓เจตนา (ค่าผิดปกติหลอกที่ทำขึ้น เพื่อทดสอบวิธีการตรวจจับ)
- ✓ข้อผิดพลาดในการประมวลผลข้อมูล (การจัดการข้อมูล หรือการกลายพันธุ์โดยไม่ได้ตั้งใจของชุดข้อมูล)
- ✓ข้อผิดพลาดในการสุ่มตัวอย่าง (การแยกหรือผสมข้อมูลจากแหล่งที่ไม่ถูกต้อง หรือแหล่งต่างๆ)
- ✓ธรรมชาติ (ไม่ใช่ข้อผิดพลาด เป็นความแปลกใหม่ในข้อมูล)

Percentile คืออะไร

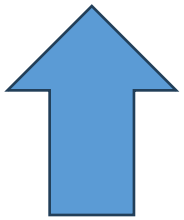
- เปอร์เซ็นต์ไทล์ หมายถึง ตำแหน่งที่แสดงให้ทราบว่า มีจำนวนร้อยละเท่าไหร่จากจำนวนทั้งหมดที่มี *ค่าต่ำกว่า* ค่าที่ตำแหน่งนั้น
- *ตัวอย่างเช่น* นักศึกษาคนหนึ่งสอบวิชาสถิติได้คะแนน 54 คะแนน และคะแนน 54 นี้อยู่ตำแหน่งเปอร์เซ็นต์ไทล์ ที่ 60 หมายความว่า ร้อยละ 60 ของนักศึกษากลุ่มนั้นได้คะแนนวิชาสถิติต่ำกว่า 54

Percentile ใน Lab นี้

คนที่เงินเดือนน้อยกว่า Percentile ที่ 1% คือ เงินเดือนต่ำกว่าปกติ

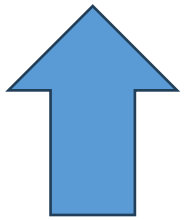
คนที่เงินเดือนปกติ

คนที่เงินเดือนมากกว่า Percentile ที่ 99% คือ เงินเดือนสูงมากกว่าปกติ



Percentile ที่ 1%

เราจะใช้สูตร Percentile เพื่อหาเงินเดือนจุดนี้ คือเงินเดือนเท่าไหร่เพื่อใช้เป็นเกณฑ์ในการหาค่าต่ำกว่าปกติ



Percentile ที่ 99%

เราจะใช้สูตร Percentile เพื่อหาเงินเดือนจุดนี้ คือเงินเดือนเท่าไหร่เพื่อใช้เป็นเกณฑ์ในการหาค่าสูงกว่าปกติ

ตรวจสอบข้อมูล เบื้องต้นด้วย Google Sheet

✓หาค่าผิดปกติ (Outliers) มีขั้นตอนดังนี้

1. หาค่าที่เป็นเงื่อนไข โดยใช้ฟังก์ชัน =percentile(ช่วงข้อมูลที่ต้องการหาค่า, ค่า percentile ที่ต้องการหา)
2. ใช้ฟังก์ชัน =if(เงื่อนไข, ผลกรณีเงื่อนไขเป็นจริง, ผลกรณีเงื่อนไขเป็นเท็จ)
3. ล็อคเซลล์ ด้วยการกด F4
4. แก้ไขช่วงข้อมูล
5. ใช้ ArrayFormula โดยกด Ctrl+Shift+enter

Step 2 ทำความสะอาดข้อมูลเบื้องต้น

ด้วย Google Sheet

ไฟล์ Lab

Part 2 คลิกขวาเปิดลิงก์ด้านล่างใน browser tab ใหม่ >>

https://docs.google.com/spreadsheets/d/1zAK-Zt59SOBu94aMq0z4CJCuHnaVWojqDqpLuYQ9stQ/edit?usp=sharing

ทำความสะอาดข้อมูล เบื้องต้นด้วย Google Sheet

✓Remove Duplicated

ไปที่เมนูข้อมูล > การล้างข้อมูล > นำรายการที่ซ้ำออก

✓แทนที่ค่าว่าง Replace NULL

=unique(คอลัมภ์ที่ต้องการหาค่า)

=averageif(คอลัมภ์ที่เป็นเงื่อนไข, ตำแหน่งที่เป็นเงื่อนไขที่ระบุ, คอลัมภ์ที่ต้องการหาค่าเฉลี่ย)

ทำความสะอาดข้อมูล เบื้องต้นด้วย Google Sheet

✓เปลี่ยนรูปแบบวันที่และดึงค่าปี

=year(คอลัมภ์ที่ต้องการดึงค่า)

=month(คอลัมภ์ที่ต้องการดึงค่า)

=day(คอลัมภ์ที่ต้องการดึงค่า)

*เพิ่มเติม**

ดึงข้อมูลจากเว็บไซต์ (Data Scraping) เบื้องต้น ด้วย Google Sheet

ดึงข้อมูลจากเว็บไซต์ (Data Scraping) เบื้องต้นด้วย Google Sheet

✓ดึงข้อมูลไฟล์ csv จากเว็บ

= IMPORTDATA(ลิงก์ไฟล์ csv)

ดึงข้อมูลจากเว็บไซต์ (Data Scraping) เบื้องต้นด้วย Google Sheet

✓ดึงข้อมูลไฟล์ html จากเว็บ

=IMPORTHTML(url, query, index)

Url: URL เว็บ

Query: ระบุว่าเป็น table หรือ list

Index: ลำดับตารางในไฟล์

ใบงานบทที่ 3

1. ดึงข้อมูลไฟล์ csv จากเว็บนี้ >> https://www.datablist.com/learn/csv/download-sample-csv-files (เลือกไฟล์ใดไฟล์หนึ่ง)

2. นำข้อมูลที่ได้มาทำต่อไปนี้ ด้วย Google Sheet

- นับจำนวนคอลัมภ์ (No. Cols)
- นับจำนวนแถว (No. Rows)
- ตรวจสอบค่าซ้ำ (Duplicated)
- ตรวจสอบค่าที่หายไป (Missing Values)

3. บันทึกไฟล์ด้วยชื่อ นามสกุล รหัสนักศึกษา แปะลิงก์แชร์มาที่ >>

ม.1 https://docs.google.com/spreadsheets/d/1WLcL4Zi3oBEZEtPsv7gwM1p12k8RPe6h8xNeh8KU9x8/edit?usp=sharing

ม.2 https://docs.google.com/spreadsheets/d/1_OFH42ZCePU-iKj-hG3A_VkGEAzgtdE45axRoYvjUMc/edit?usp=sharing

อ้างอิง

- Google Sheet For Data Science >>

  https://www.youtube.com/watch?v=DlRwRX-mtnA

- IMPORTHTML Google Sheets - The Ultimate Guide for 2023 >>https://www.lido.app/tutorials/importhtml-google-sheets

- เปอร์เซ็นต์ไทล์ >> https://tuemaster.com/blog/%E0%B9%80%E0%B8%9B%E0%B8%AD%E0%B8%A3%E0%B9%8C%E0%B9%80%E0%B8%8B%E0%B9%87%E0%B8%99%E0%B8%95%E0%B9%8C%E0%B9%84%E0%B8%97%E0%B8%A5%E0%B9%8C-%E0%B9%80%E0%B8%94%E0%B9%84%E0%B8%8B%E0%B8%A5%E0%B9%8C-%E0%B8%84/