Chapter 4

Data Preparing With Google Sheet

Google Sheet คือ อะไร

- แอปสเปรดชีตออนไลน์สำหรับสร้างและจัดรูปแบบสเปรดชีตและทำงานร่วมกับคนอื่นๆ

ไฟล์ Lab

Part 1

- https://docs.google.com/spreadsheets/d/1UbDJpKVAT3b7oY0j-TwfxT5iHwPe4Op_aEoxQuSXH0I/edit?usp=sharing

ตรวจสอบข้อมูล เบื้องต้นด้วย Google Sheet

✓นับจำนวนคอลัมภ์ (No. Cols)

=columns(ช่วงข้อมูล)

✓นับจำนวนแถว (No. Rows)

=rows(ช่วงข้อมูล)

✓นับค่าซ้ำ (Duplicated)

=countif(คอลัมภ์ที่ต้องการนับ, ตำแหน่งข้อมูลที่ต้องการหาค่าซ้ำ)

ตรวจสอบข้อมูล เบื้องต้นด้วย Google Sheet

✓นับค่าที่หายไป (Missing Values)

=countblank(แถวที่ต้องการหา Missing Values)

✓หาค่าผิดปกติ (Outliers)

A. หา Percentile(ช่วงข้อมูล, ค่า 0.99 หรือ 0.01)

B. if(เงื่อนไข, ผลกรณีเงื่อนไขเป็นจริง, ผลกรณีเงื่อนไขเป็นเท็จ)

C. ล็อคเซลล์ ด้วยการกด F4

D. ArrayFormula ใช้ Ctrl+Shift+enter

ไฟล์ Lab

Part 2

- https://docs.google.com/spreadsheets/d/1jcSpBqyzfRn6Ig56h_LvIB5W5HB_R-4teuvEgHYe1jI/edit?usp=sharing

ทำความสะอาดข้อมูล เบื้องต้นด้วย Google Sheet

✓แทนที่ค่าว่าง Replace NULL

=unique(คอลัมภ์ที่ต้องการหาค่า)

=averageif(คอลัมภ์ที่เป็นเงื่อนไข, คอลัมภ์ที่เป็นเงื่อนไขที่ระบุ, คอลัมภ์ที่ต้องการหาค่าเฉลี่ย)

✓เปลี่ยนรูปแบบวันที่และดึงค่าปี

=year(คอลัมภ์ที่ต้องการดึงค่า)

✓ปรับข้อมูลให้ scale ใกล้เคียงกันด้วย Normalize

✓ปรับข้อมูลให้ scale ใกล้เคียงกันด้วย Standardize

ทำไมต้องทำ Normalization ข้อมูล

- ข้อมูลดิบที่เราได้รับมานั้นมีความหลากหลาย ทั้งชนิดข้อมูล รูปแบบข้อมูล และ Scale ช่วงของข้อมูล (ข้อมูลตัวเลข Cardinal) เช่น ข้อมูลเด็กมัธยม มี 3 Feature คือ อายุ [10, 20], น้ำหนัก [30, 200] ส่วนสูง [120, 180]
- สำหรับอัลกอริทึม Machine Learning หลาย ๆ ตัว ไม่สามารถรับข้อมูลหลากหลาย Scale แบบนี้ได้โดยตรง จำเป็นที่เราต้องทำ Normalization ก่อนที่เราจะป้อนข้อมูลให้กับโมเดลอัลกอริทึมถึงจะสามารถทำงานได้

ทำไมต้องทำ Standardization ข้อมูล

- อัลกอริทึม Machine Learning หลาย ๆ ตัว ต้องการให้เราปรับข้อมูลให้เป็นแบบนี้ก่อนที่จะป้อนให้โมเดลใช้เทรน