Matplotlib, Pandas, Feature Engineering, Data Manipulation, & Data Visualization

Topic: Analyzing and visualizing insights of student test scores from study hours and physical relaxation. (30 นาที)

Dataset details

Column	**Description**
hours	Hours spent studying per week (ชั่วโมงศึกษาต่อสัปดาห์)
sleep	Hours of sleep per night (ชั่วโมงนอนหลับ)
score	Final exam score (0–100) (คะแนนสอบ)

Data constraints (ข้อมูลที่ให้มา) อ้างอิงตามไฟล์ CSV**:**

- **hours** → random integers from **1 to 15.**
- **sleep** → random integers from **4 to 9.**
- **score** → synthetic formula: score = **5 * hours + 3 * sleep + noise(±10)**

1. **Create data sampling and tabulate data:**

1.1. Download the dataset into your local computer. (ดาวน์โหลด Dataset ที่แนบมา)

1.2. Read the dataset (.csv file) and then transform it into pandas Dataframe type. (นำเข้า csv file แปลงเป็น dataframe)

1.3 นักศึกษาที่เป็นรหัสนักศึกษาลงท้ายด้วย **<u>เลขคี่</u>** ให้ใช้ข้อมูลตั้งแต่**แถว 2-42** อ้างอิงตามใน CSV

และ

นักศึกษาที่เป็นรหัสนักศึกษาลงท้ายด้วย **<u>เลขคู่</u>** ให้ใช้ข้อมูลตั้งแต่**แถว 11-51** อ้างอิงตามใน CSV

2. Show data details

2.1. Show the first 7 rows and tail 7 rows. (แสดงผล Dataframe 7 แถวแรกและ 7 แถวหลัง)

2.2. See data description. (แสดงผลคำบรรยายข้อมูล)

2.3. Check for missing values.

Hint: Use command "df.isnull().sum()" to summarize all null values.

(เช็คข้อมูลหรือค่าที่หายไป)

2.4. Plot histograms or boxplots or pair plots for all columns. (แสดงผลเป็น Histogram หรือ boxplot หรือ pair plot)

2.5. Compute the correlation matrix (Using "Pearson" method). (คำนวณ correlation ของข้อมูลโดยใช้วิธี Pearson)

3. Manipulate data

3.1. Impute missing values with Median. (เติมข้อมูลที่หายไปด้วยการตรวจสอบมัธยฐาน)

3.2. Create a new column "hours_squared" = **hours ** 2** (สร้างคอลัมน์ใหม่แบบ numerical)

3.3. Create a new column "sleep_quality" = **low/normal/high** categories (สร้างคอลัมน์ใหม่แบบ categorical/binning 3 ช่วง/ประเภท)

"low": less than 5 hours

"normal": from 5 to less than 7 hours

"high": more than 7 hours

3.4. Create a new column "efficiency" = **hours / sleep.** (สร้างคอลัมน์ใหม่จากการปฏิสัมพันธ์ 2 ฟีเจอร์)

4. Create simple plot

4.1. Scatter plot: sleep vs score (i.e., x-axis is sleep, y-axis is score). (แสดงผลเป็น Scatter plot 2 มิติ)

5. **Final Output**

5.1. Display the data frame first 10 rows and tail 10 rows with new features. (แสดงผล Dataframe 10 แถวแรกและ 10 แถวหลังที่มีฟีเจอร์ใหม่)

5.2. Print correlation with score with new and existing features.

Hint: Set the parameter "numeric_only" = True to avoid calculating string features.

(คำนวณ correlation ของข้อมูลที่มีฟีเจอร์ใหม่นี้)

วิธีการส่ง (ส่งใน assignment ของ google classroom)

- Capture วิธีการและผลลัพธ์ทั้งหมด จัดทำเป็น PDF ส่ง

  และ/หรือ
- ส่งด้วยไฟล์ ipython notebook (.ipynb) ที่มีใส่คำกำกับ markdown ควบคู่กับการ run แต่ละ cell