**DADS 5001 Mini Project: Sleep health and lifestyle Analytic**

**1. DATA**

ข้อมูลที่นำมาใช้ในการวิเคราะห์ได้จาก: [**https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset**](https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset)

ข้อมูลประกอบด้วย

1. Person ID: รหัสเพื่อใช้ระบุตัวบุคคล
2. Gender: เพศของบุคคล (ชาย/หญิง)
3. Age: อายุของบุคคลเป็นปี
4. Occupation: อาชีพของบุคคล
5. Sleep Duration (hours): จำนวนชั่วโมงที่คนนอนหลับต่อวัน
6. Quality of Sleep (scale: 1-10): การให้คะแนนคุณภาพการนอนหลับตามอัตนัย ตั้งแต่ 1 ถึง 10
7. Physical Activity Level (minutes/day): จำนวนนาทีที่บุคคลนั้นทำกิจกรรมทางกายในแต่ละวัน
8. Stress Level (scale: 1-10): คะแนนส่วนตัวของระดับความเครียดที่บุคคลประสบ โดยมีค่าตั้งแต่ 1 ถึง 10
9. BMI Category: หมวดหมู่ BMI ของบุคคล (เช่น น้ำหนักน้อย, ปกติ, น้ำหนักเกิน,โรคอ้วน)
10. Blood Pressure (systolic/diastolic):การวัดความดันโลหิตของบุคคล ซึ่งระบุว่าเป็นความดัน systolic มากกว่าความดัน diastolic
11. Heart Rate (bpm): อัตราการเต้นของหัวใจขณะพักของบุคคลเป็นครั้งต่อนาที
12. Daily Steps: จำนวนก้าวเดินที่บุคคลนั้นใช้เดินต่อวัน
13. Sleep Disorder: การมีหรือไม่มีความผิดปกติของการนอนหลับในบุคคล (None, Insomnia, Sleep Apnea)

* None: บุคคลนี้ไม่แสดงอาการผิดปกติของการนอน
* Insomnia: ปัญหาในการนอนหลับหรือหลับไม่สนิท นำไปสู่การนอนหลับที่ไม่เพียงพอหรือมีคุณภาพต่ำของแต่ละบุคคล
* Sleep Apnea: การหยุดหายใจระหว่างการนอนหลับ ส่งผลให้รูปแบบการนอนหยุดชะงักและความเสี่ยงต่อสุขภาพที่อาจเกิดขึ้น

**2. Import Library**

ในโปรเจ็คนี้ ใช้ไลบรารี สำคัญสามตัวคือ pandas, seaborn, และ matplotlib.pyplot

1. pandas: ไลบรารีที่ใช้สำหรับการวิเคราะห์และจัดการข้อมูลที่มีโครงสร้างหรือเรียกอีกชื่อหนึ่งว่า DataFrame โดย pandas มีฟังก์ชันที่สามารถอ่านและเขียนไฟล์ข้อมูลต่าง ๆ เช่น CSV, Excel, SQL database และอื่น ๆ นอกจากนี้ยังมีเครื่องมือสำหรับการจัดการข้อมูล เช่น การกรองข้อมูล การจัดเรียง การเลือกคอลัมน์ การคำนวณสถิติพื้นฐาน เป็นต้น

2. seaborn: ไลบรารีที่ใช้สำหรับการพล็อตกราฟและแสดงข้อมูลทางสถิติ โดย seaborn ให้ความสามารถในการสร้างกราฟและการแสดงผลข้อมูลที่สวยงามและมีความสวยงามเพิ่มขึ้น โดยมีฟังก์ชันที่ช่วยในการสร้างกราฟหลากหลายรูปแบบ เช่น scatter plot, line plot, bar plot, histogram, box plot เป็นต้น

3. matplotlib: ไลบรารีที่ใช้สำหรับการพล็อตกราฟและการแสดงผลข้อมูล ซึ่ง pyplot เป็นโมดูลใน matplotlib ที่ให้เครื่องมือในการสร้างและแสดงผลกราฟ โดยมีฟังก์ชันที่ช่วยในการกำหนดแกน x และแกน y, การเพิ่มข้อมูลลงบนกราฟ เปลี่ยนแปลงสี ขนาด และรูปแบบของสัญลักษณ์ และฟังก์ชันอื่น ๆ ที่ช่วยในการปรับแต่งกราฟ

**3.Cleansing Data**

1. ตรวจสอบข้อมูลโดยใช้คำสั่ง .info()

A screenshot of a computer

Description automatically generated

จากข้อมูลที่ได้ 374 แถว, 13 คอลัมน์ ไม่พบค่าว่าง

2. เช็คข้อมูลที่เป็น object ใช้คำสั่ง .value\_counts()

2.1ข้อมูลของคอลัมน์ Gender : ผู้ชาย 189 คนและเป็นผู้หญิง 185 คน

A picture containing text, font, receipt, white

Description automatically generated

2.2 ข้อมูลของคอลัมน์ Occupation มีอาชีพทั้งหมด 11 อาชีพ

A screenshot of a computer

Description automatically generated with low confidence

2.3ข้อมูลของคอลัมน์ BMI มีข้อมูลเป็น 4 ส่วน, ในส่วนของ Normal, Normal Weight ซึ่งมีความหมายเดียวกัน จึงจะดำเนินการจัดเรียงข้อมูลใหม่ในลำดับถัดไป

A picture containing text, receipt, font, white

Description automatically generated

2.4ข้อมูลของคอลัมน์ Sleep Disorder แบ่งข้อมูลเป็น 3 ส่วน ได้แก่ None, Sleep Apnea และ Insomia

A close-up of numbers

Description automatically generated with low confidence

3. แยกตัวเลขความดัน Systolic และ Diastolic:

โค้ดนี้สร้างคอลัมน์ใหม่ใน DataFrame ชื่อ "Systolic" และ "Diastolic" โดยกำหนดค่าเริ่มต้นเป็น 0 สำหรับทั้งสองคอลัมน์นี้ จากนั้นใช้ลูป for และ while เพื่อแยกตัวเลขความดัน Systolic และ Diastolic จากคอลัมน์ "Blood Pressure" โดยใช้เครื่องหมาย "/" เป็นตัวแบ่ง และเก็บค่าตัวเลขทั้งสองลงในคอลัมน์ที่สร้างขึ้นมาใหม่ ดังตัวอย่างต่อไปนี้

|  |  |  |
| --- | --- | --- |
| **Blood Pressure** | **Systolic** | **Diastolic** |
| 120/80 | 120 | 80 |

4. เปลี่ยน "Normal Weight" เป็น "Normal":

โค้ดนี้ใช้คำสั่ง loc เพื่อเลือกแถวที่คอลัมน์ "BMI Category" เป็น "Normal Weight" และทำการเปลี่ยนค่าเป็น "Normal" แทน ซึ่งจะทำให้คอลัมน์ "BMI Category" เนื่องจากมีการใช้คำที่ซ้ำซ้อนซึ่งมีความหมายเดียวกัน

5. จัดกลุ่มข้อมูล outlier เป็น "others":

โค้ดนี้กำหนดค่าสำหรับการจัดกลุ่มข้อมูล outlier ของคอลัมน์ "Occupation" โดยกำหนดค่า cut\_off เป็น 10 ซึ่งหมายถึงถ้ามีจำนวนข้อมูลในกลุ่มนั้นต่ำกว่า 10 จะถือว่าเป็น outlier จากนั้นใช้คำสั่ง value\_counts เพื่อนับจำนวนข้อมูลในแต่ละกลุ่มของคอลัมน์ "Occupation" และเลือกกลุ่มที่จำนวนน้อยกว่า cut\_off และทำการเปลี่ยนค่าในคอลัมน์ "Occupation" ของกลุ่มนั้นเป็นค่าใหม่ที่กำหนดไว้ในตัวแปร new\_label ซึ่งคือ "others"

หลังจากนั้นโค้ดทำการ

พิมพ์ข้อความ "Successfully changed all the 'Normal Weight' to 'Normal'" เพื่อแสดงผลว่าเปลี่ยนค่าสำเร็จ

พิมพ์ข้อความ "Successfully separate SYS and DIA" เพื่อแสดงผลว่าแยกตัวเลขความดันสำเร็จ

พิมพ์ข้อความ "Successfully changed all occupation below {cut\_off} to {new\_label}" เพื่อแสดงผลว่าเปลี่ยนค่าข้อมูลอาชีพที่น้อยกว่า cut\_off สำเร็จ

**4.Categorical and Numerical Analysis**

**รูปภาพประกอบด้วย แผนภาพ, ภาพหน้าจอ, ไลน์, ออกแบบ

คำอธิบายที่สร้างโดยอัตโนมัติ**

**จาก Figure1 พบว่า**

**Age:** ข้อมูลเพศแสดงให้เห็นว่ามีการกระจายที่สม่ำเสมอและอาจมีผลต่อการนอนหลับของบุคคล อาจมีปัจจัยที่เกี่ยวข้องกับเพศที่ส่งผลต่อรูปแบบการนอนหลับ อย่างไรก็ตาม เพศไม่ได้ระบุว่ามีความสัมพันธ์อย่างชัดเจนกับพารามิเตอร์อื่น ๆ ดังนั้นการวิเคราะห์เพิ่มเติมอาจจำเป็นในการหาความสัมพันธ์ที่เป็นไปได้

**Occupational:** ข้อมูลอาชีพแสดงให้เห็นว่ามีหลายชนชั้นและอาจมีบางอาชีพที่ไม่มีตัวแทนเพียงพอที่จะสรุปได้ การวิเคราะห์เพิ่มเติมอาจจำเป็นเพื่อศึกษาและเข้าใจความสัมพันธ์ระหว่างอาชีพกับคุณภาพการนอน

**BMI Category:** ข้อมูลหมวดหมู่ค่าดัชนีมวลกายแสดงให้เห็นว่ามีการกระจายที่สมดุลระหว่างหมวดหมู่ที่มีน้ำหนักเกินและปกติ การวิเคราะห์เพิ่มเติมอาจช่วยในการตรวจสอบความสัมพันธ์ระหว่างค่าดัชนีมวลกายกับคุณภาพการนอน

**Sleep Disorder:** ข้อมูลความผิดปกติของการนอนหลับแสดงให้เห็นว่าการกระจายจะแบ่งเท่า ๆ กันระหว่างบุคคลที่มีและไม่มีความผิดปกติของการนอนหลับ อาจมีการแบ่งแยกระหว่างภาวะหยุดหายใจขณะหลับและโรคนอนไม่หลับ น่าสนใจว่าพารามิเตอร์นี้อาจเป็นตัวแปรตอบสนองที่สามารถเกี่ยวข้องกับคุณภาพการนอนได้

การวิเคราะห์เพิ่มเติมเกี่ยวกับพารามิเตอร์เหล่านี้จะช่วยให้เรามีภาพรวมที่ครอบคลุมและเข้าใจในความสัมพันธ์ที่เป็นไปได้ระหว่างพารามิเตอร์ต่าง ๆ

รูปภาพประกอบด้วย แผนภาพ, สี่เหลี่ยมผืนผ้า, สี่เหลี่ยม, วางแผน

คำอธิบายที่สร้างโดยอัตโนมัติ

จาก Figure2 เราใช้ข้อมูลจากชุดข้อมูลที่เกี่ยวข้องกับคุณภาพการนอน ปัจจัยที่สำคัญที่ถูกพิจารณาได้แก่ อายุ, ระยะเวลาการนอน,คุณภาพการนอน, ระดับกิจกรรมทางกาย, ระดับความเครียด, อัตราการเต้นของหัวใจ, จำนวนก้าวต่อวัน, ความดันโลหิตSystolic และDiastolic และพบว่า

**Age:** ข้อมูลอายุแสดงให้เห็นว่ามีการกระจายตัวคล้ายกับการแจกแจงปกติและไม่มีค่าที่ผิดปกติ

**Sleep Duration:** ข้อมูลระยะเวลาการนอนแสดงให้เห็นว่ามีการกระจายตัวแบบเส้นตรง

**Quality of Sleep:** ข้อมูลคุณภาพการนอนแสดงให้เห็นว่ามีการกระจายตัวค่อนข้างเร็วตามแบบเอ็กซ์โพเนนเชียลและเอียงไปทางค่าสูง นั่นหมายความว่ามีแนวโน้มที่การนอนที่ดีมากๆ

**Physical Activity Level:** ข้อมูลระดับกิจกรรมทางกายแสดงให้เห็นว่ามีการกระจายตัวแบบเส้นตรง

**Stress Level:** ข้อมูลระดับความเครียดแสดงให้เห็นว่าการกระจายตัวมีแนวโน้มที่สูงกว่าที่มีระดับความเครียดสูง

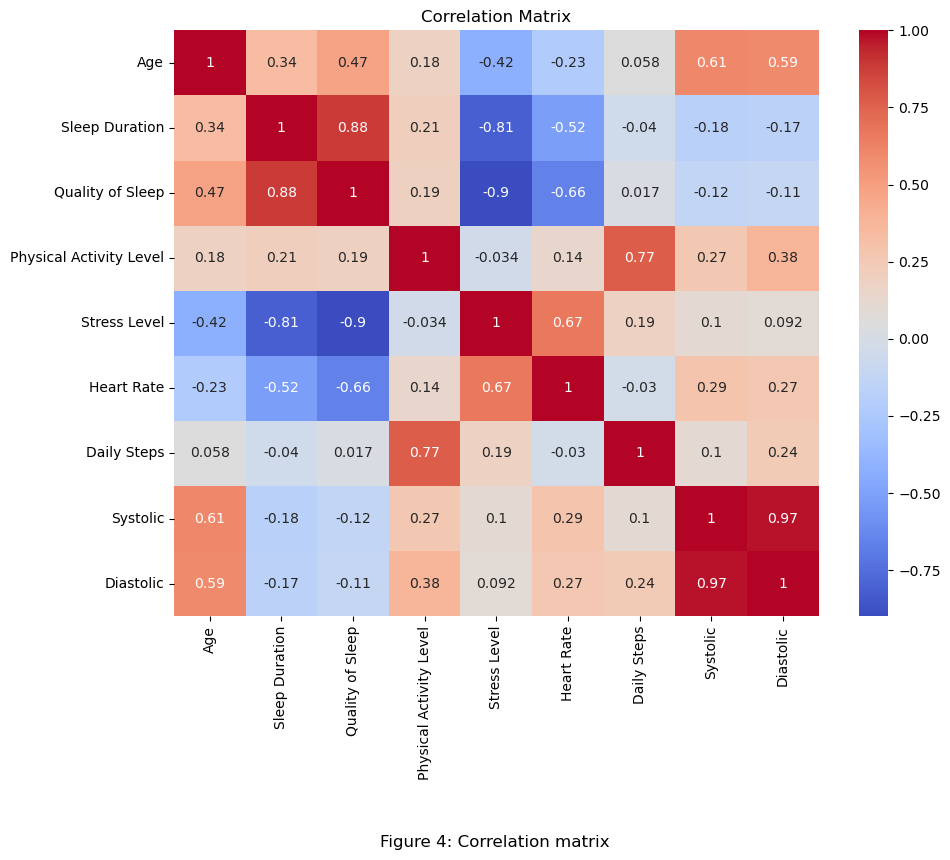
**Heart rate:** ข้อมูลอัตราการเต้นของหัวใจแสดงให้เห็นว่ามีการกระจายตัวค่อนข้างเร็วตามแบบเอ็กซ์โพเนนเชียลและเอียงไปทางค่าต่ำ

**Daily Steps:** ข้อมูลจำนวนก้าวต่อวันแสดงให้เห็นว่ามีการกระจายตัวคล้ายกับการแจกแจงปกติในชุดข้อมูล

**Systolic and Diastolic Blood pressure:** ข้อมูลความดันโลหิตสูงและต่ำแสดงให้เห็นว่ามีการ

กระจายตัวคล้ายกับการแจกแจงปกติ แต่อาจมีการแจกแจงตามแบบอื่นอีกแบบหนึ่ง น่าสนใจว่าการกระจายตัวระหว่างความดันโลหิตสูงและต่ำไม่เหมือนกัน

**5.Correlation Matrix**

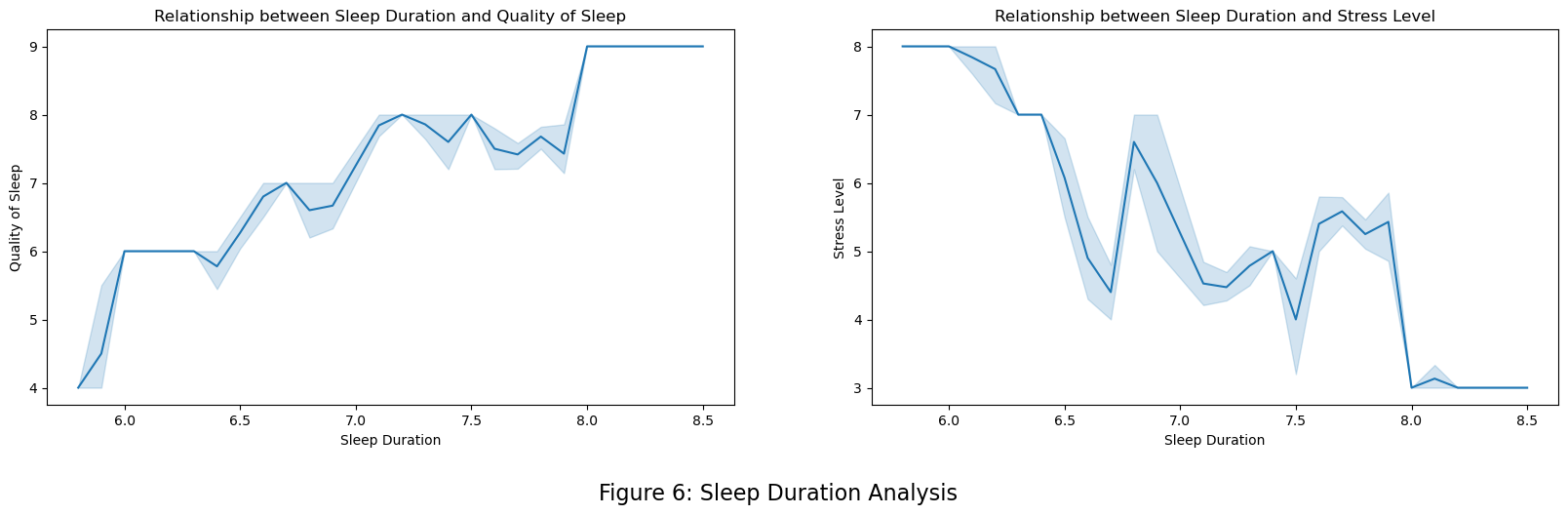
****

Correlation matrix ใน Figure 4 ให้ข้อมูลเชิงลึกที่น่าสนใจ คือ Stress Level และ Sleep Duration มีความสัมพันธ์กันอย่างมาก เช่นเดียวกับ Stress LevelและQuality of Sleep รวมถึงความสัมพันธ์ระหว่าง Sleep Duration และ Quality of Sleep

**6.Sleep Duration**

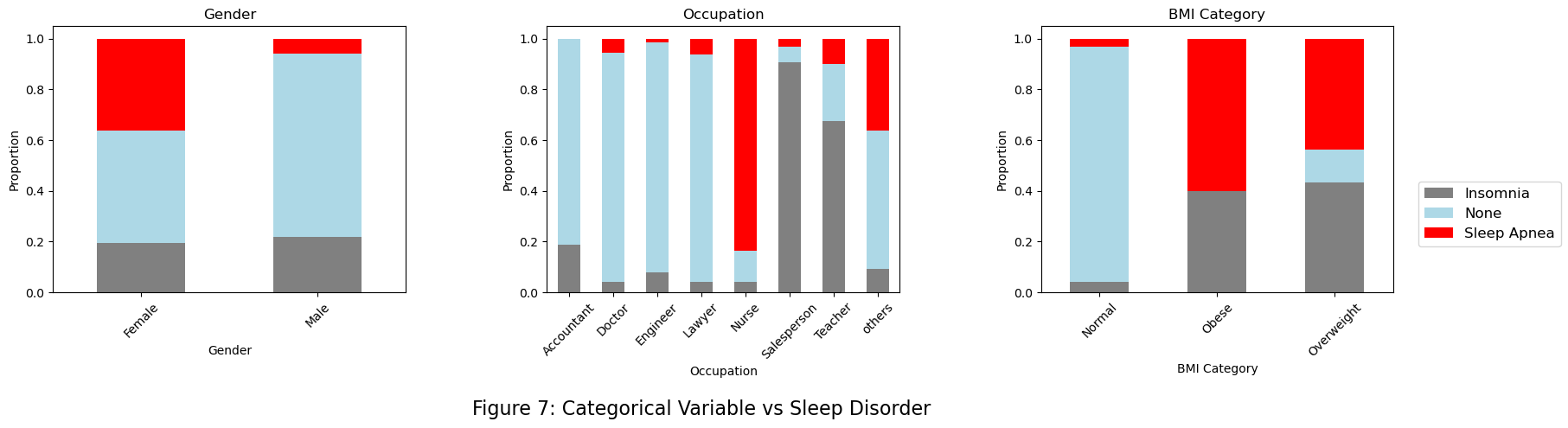
รูปภาพประกอบด้วย แผนภาพ, ไลน์, วางแผน, พล็อต

คำอธิบายที่สร้างโดยอัตโนมัติจาก Figure 5 พบว่า ความสัมพันธ์ระหว่างตัวแปรส่วนใหญ่กับระยะเวลาการนอนหลับจะค่อนข้างเป็นเส้นตรง อย่างไรก็ตาม มีปัญหากับกราฟที่จุดทับซ้อนกัน ส่งผลให้จุดเหล่านี้มีน้ำหนักมากขึ้น โดยตัวแปนที่น่าสนใจคือ Quality of Sleep และ Stress Level



และเมื่อดูกราฟจาก Figure 6: จะเห็นได้ชัดเจนว่า เมื่อ Sleep Duration สูงขึ้น จะทำให้ Quality of Sleep เพิ่มขึ้น และ Stress Level ลดลง

**7.Sleep Disorder**

****

ความสัมพันธ์ระหว่างตัวแปรของ Gender กับ Sleep Disorder จาก figure7 พบว่า ผู้หญิงมีอาการหยุดหายใจขณะนอนหลับมากกว่าผู้ชาย แต่มีปัญหาการนอนไม่หลับที่น้อยกว่าผู้ชาย

ความสัมพันธ์ระหว่างตัวแปรของ Occupation กับ Sleep Disorder ข้อมูลชุดนี้พบว่า พยาบาลมีอาการหยุดหายใจขณะนอนหลับมากที่สุด ในส่วนของอาการนอนไม่หลับอันดับ 1 เป็นอาชีพพนักงานขาย และครูเป็นอันดับต่อมา

ความสัมพันธ์ระหว่างตัวแปรของ BMI กับ Sleep Disorder ข้อมูลชุดนี้พบว่า คนที่เป็นโรคอ้วนมีอาการหยุดหายใจขณะนอนหลับและนอนไม่หลับสูง ในส่วนของคนที่น้ำหนักเกินจะอาหารนอนไม่หลับและหยุดหายใจขณะนอนหลับเป็นส่วนใหญ่