**นายกฤตภาสณ์ พูลวงษ์ ID:6610754314**

Dataset ที่เลือกใช้จาก Kaggle Sleep Health and Lifestyle Dataset ผมตั้งชื่อใหม่เพื่อให้ง่ายต่อการเรียกใช้เป็น “sleep.csv”

Gender (เพศ)

Age (อายุ)

Occupation (อาชีพ)

Sleep Duration (ชั่วโมงการนอน)

Quality of Sleep (คุณภาพการนอน)

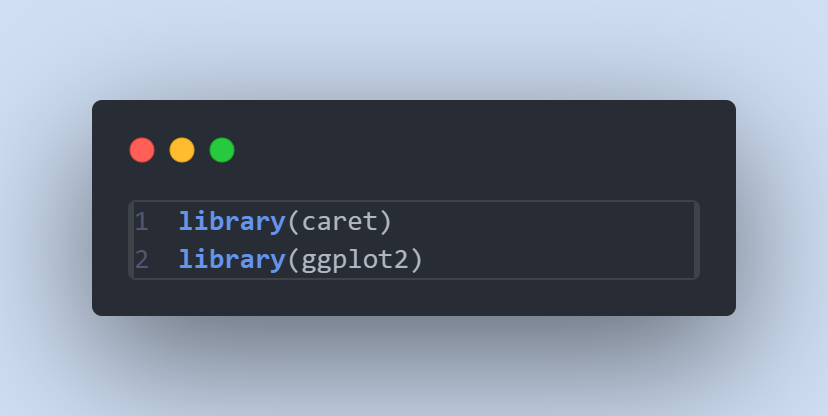
Physical Activity Level (ระดับการออกกำลังกาย)

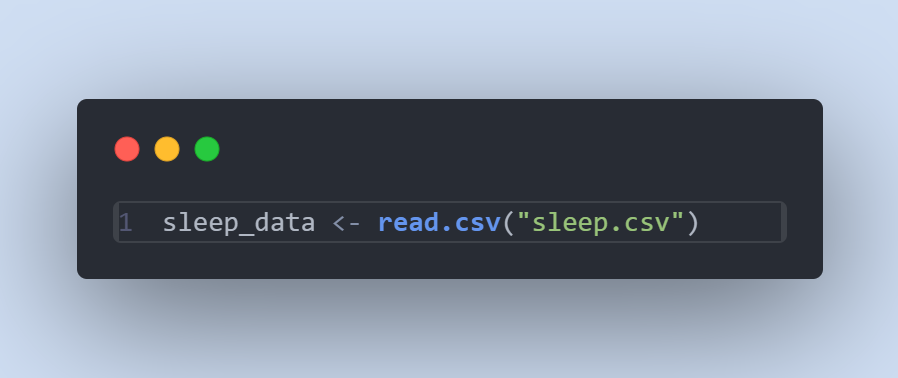
Stress Level (ระดับความเครียด)

BMI Category, Blood Pressure, Sleep Disorder (ข้อมูลสุขภาพ)

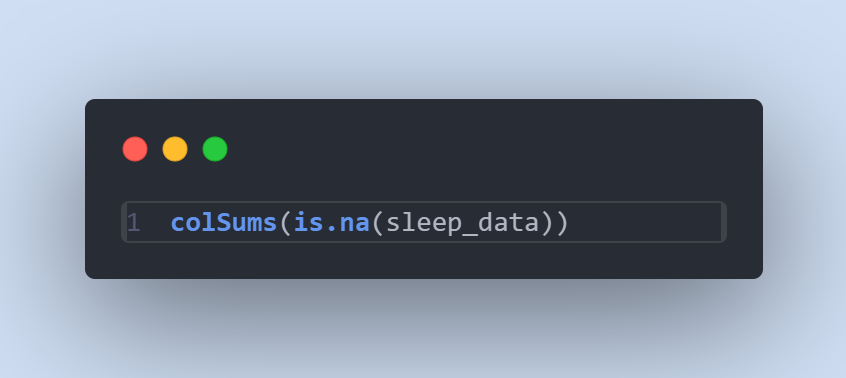
Daily Steps (จำนวนก้าวเดินต่อวัน)

Library ที่ใช้คือ caret ซึ่งใช้สำหรับแบ่งข้อมูลเป็น train กับ test เพื่อความแม่นยำในการ Run **Linear Regression** และเพื่อให้แบบจำลองมีประสิทธิภาพเมื่อเจอกันข้อมูลทีไม่เคยเจอมาก่อน และใช้ ggplot2 สำหรับการสร้างกราฟแสดงความสัมพันธ์ระหว่างตัวแปร เช่น การแสดงความสัมพันธ์ระหว่าง **Sleep Duration** กับ **Quality of Sleep**.

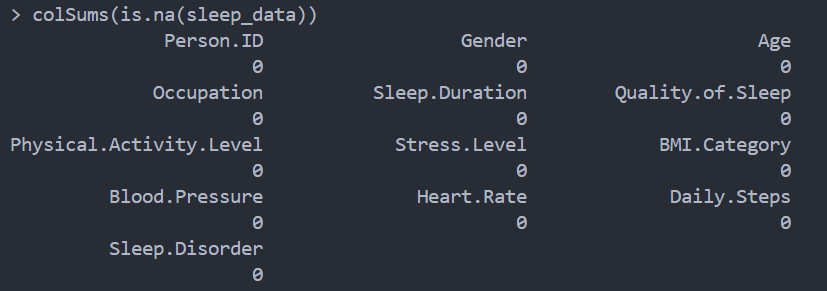
****

**Import dataset**

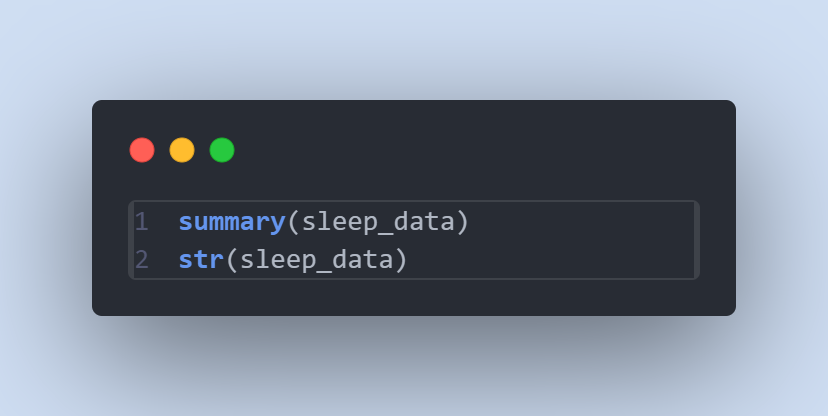
**Check missing value**



**Output: ไม่มี missing value ใน column ต่างๆ**



**ดูข้อมูลแบบภาพรวม**



A screenshot of a computer

AI-generated content may be incorrect.

Person.ID: เป็นตัวเลขตั้งแต่ 1 ถึง 374 (จำนวนแถวในข้อมูลคือ 374)

**Gender, Occupation, BMI Category, Blood Pressure, Sleep Disorder เป็น character variables**

**Age:**

Min: 27 ปี, Max: 59 ปี

ค่ามัธยฐาน (Median) = 43 ปี

**Sleep Duration:**

ค่ามัธยฐาน 7.2 ชั่วโมง, ค่าเฉลี่ย 7.132 ชั่วโมง

มีค่าต่ำสุดที่ 5.8 ชั่วโมง และสูงสุดที่ 8.5 ชั่วโมง

**Quality of Sleep:**

มัธยฐาน 7, ค่าเฉลี่ย 7.313 (ช่วงคะแนน 4-9)

**Physical Activity Level:**

มัธยฐาน 60, ค่าเฉลี่ย 59.17 (ช่วงคะแนน 30-90)

**Stress Level:**

มัธยฐาน 5, ค่าเฉลี่ย 5.385 (ช่วงคะแนน 3-8)

**Heart Rate:**

มัธยฐาน 70 BPM, ค่าเฉลี่ย 70.17 BPM (ช่วง 65-86 BPM)

**Daily Steps:**

มัธยฐาน 7,000 ก้าว, ค่าเฉลี่ย 6,817 ก้าว (ช่วง 3,000-10,000 ก้าว)

**สร้างตัวแปร Dummy Variable เพื่อใช้ในการวิเคราะห์ Regression กับตัวแปรที่สนใจในที่นี้สนใจตัวแปรเพศกับตัวแปรอาชีพ ว่ามีผลต่อคุณภาพการนอนหลับหรือไม่**



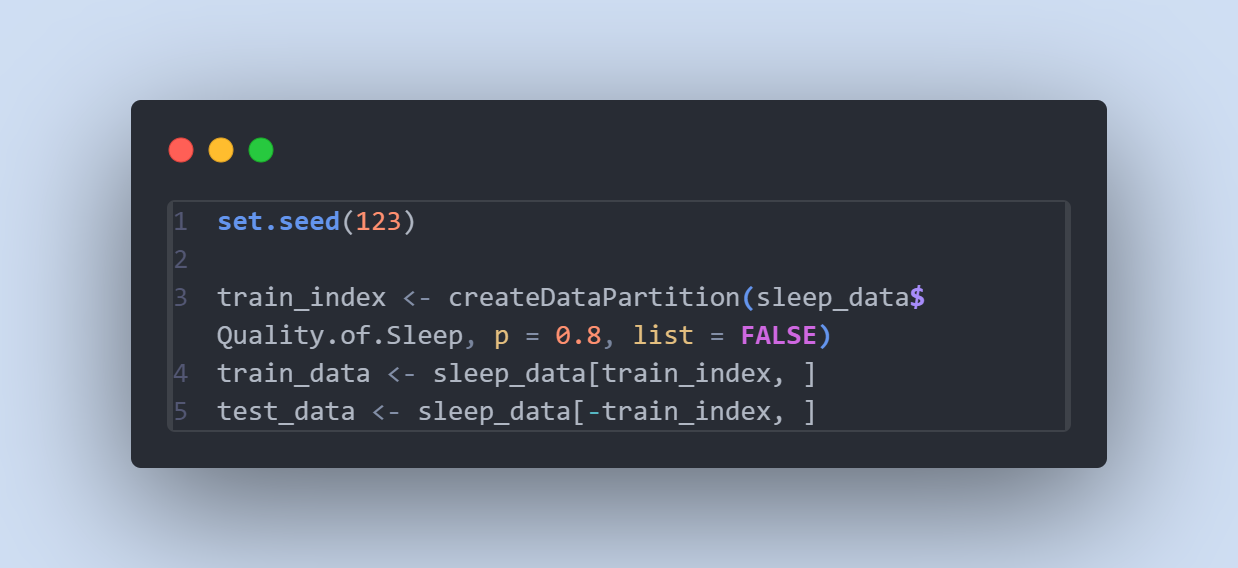
แปลง Gender เป็นตัวเลข (1 = Male, 0 = Female)

แปลง Occupation เป็น factor เพื่อใช้ในการวิเคราะห์

ใช้ model.matrix() สร้าง **Dummy Variables** สำหรับ Occupation

ใช้ cbind() เพิ่มคอลัมน์ใหม่เข้าไปใน sleep\_data

Set ข้อมูลสำหรับการ train model โดยจะใช้ชุด train 80% ของข้อมูลทั้งหมด ส่วนชุด test ใช้ข้อมูล 20% ที่เหลือ เพื่อดูความแม่นยำของ model เทียบกับใช้ข้อมูลแบบ 100%



set.seed(123) เพื่อทำให้ code ที่รันออกมาทุกครั้งได้การแบ่งชุดข้อมูลที่เหมือนเดิม

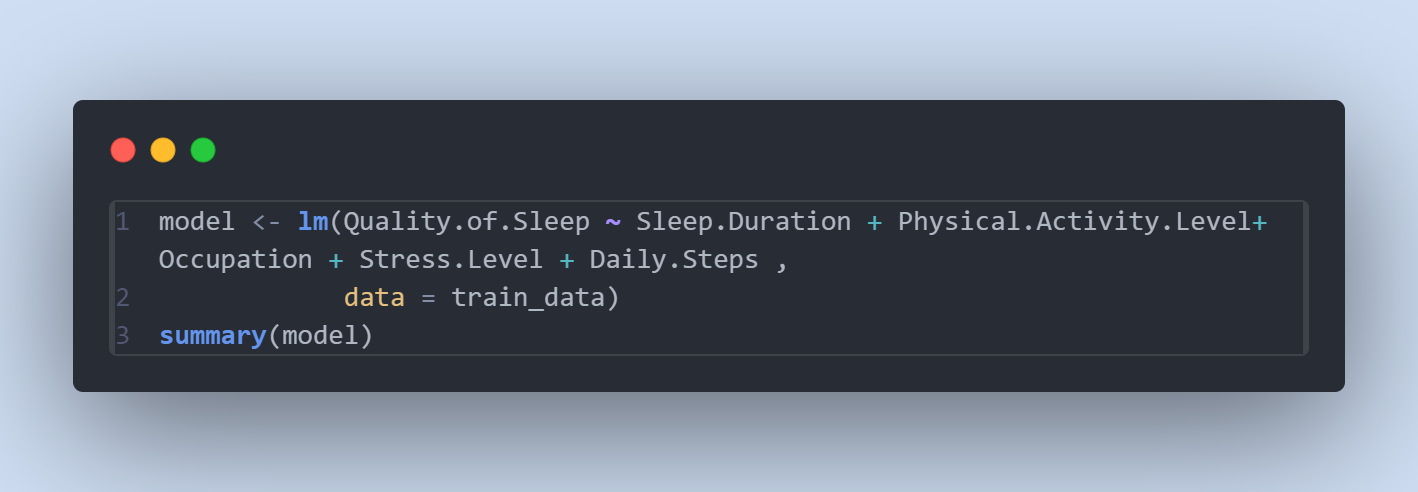
ใช้ฟังก์ชัน createDataPartition() จาก caret package

sleep\_data$Quality.of.Sleep เป็นตัวแปรที่ใช้แบ่งและเลือกกลุ่มตัวอย่าง (Stratified Sampling)

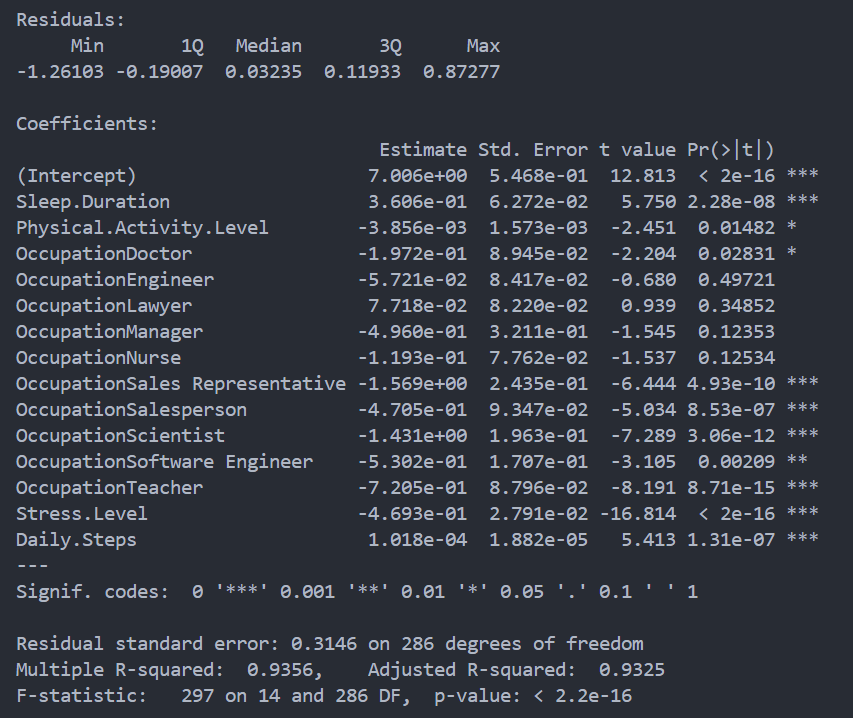
p = 0.8 หมายถึงแบ่งกลุ่มตัวอย่างออกเป็น 80% ของข้อมูลทั้งหมด ให้เป็น Training Set

train\_data ใช้ข้อมูลที่แบ่งออกมา 80% จากข้อมูลทั้งหมด ส่วน test\_data ใช้ข้อมูลอีก 20% ที่เหลือ

**ทำการ Run regression กับแบบจำลองแต่ใช้ข้อมูลชุด train ที่มีข้อมูลแค่ 80% ของข้อมูลทั้งหมด**



**Summary(model) เพื่อดูผลลัพธ์จากการ Run regression**



**แปรผลค่าสัมประสิทธ์**

**Sleep Duration** (ระยะเวลาในการนอนหลับ) ถ้านอนหลับเพิ่มขึ้น 1 ชั่วโมง  
ส่งผลให้คุณภาพการนอนเพิ่มขึ้น 0.3606 หน่วย

**Physical Activity Level** (ระดับการออกกำลังกาย) ถ้าระดับการออกกำลังกายเพิ่มขึ้น 1 หน่วยส่งผลให้คุณภาพการนอนลดลง 0.0003856 หน่วย

**Occupation Doctor** หากทำอาชีพหมอทำให้คุณภาพการนอนลดลง 0.1972 หน่วย

**Occupation Engineer** หากทำอาชีพวิศวกรทำให้คุณภาพการนอนลดลง 0.005721 หน่วย

**Occupation Lawyer** หากทำอาชีพทนายทำให้คุณภาพการนอนเพิ่มขึ้น 0.007718 หน่วย

**Occupation Manager** หากทำอาชีพผู้จัดการทำให้คุณภาพการนอนลดลง 0.4960 หน่วย

**Occupation Nurse** หากทำอาชีพพยาบาลทำให้คุณภาพการนอนลดลง 0.1193 หน่วย

**Occupation Sales Representative** หากทำอาชีพตัวแทนขายของบริษัท  
ทำให้คุณภาพการนอนลดลง 1.569 หน่วย

**Occupation Salesperson** หากทำอาชีพนักขายของบริษัททำให้คุณภาพการนอนลดลง 0.4705 หน่วย

**Occupation Scientist** หากทำอาชีพนักวิทยาศาสตร์ทำให้คุณภาพการนอนลดลง 1.431 หน่วย

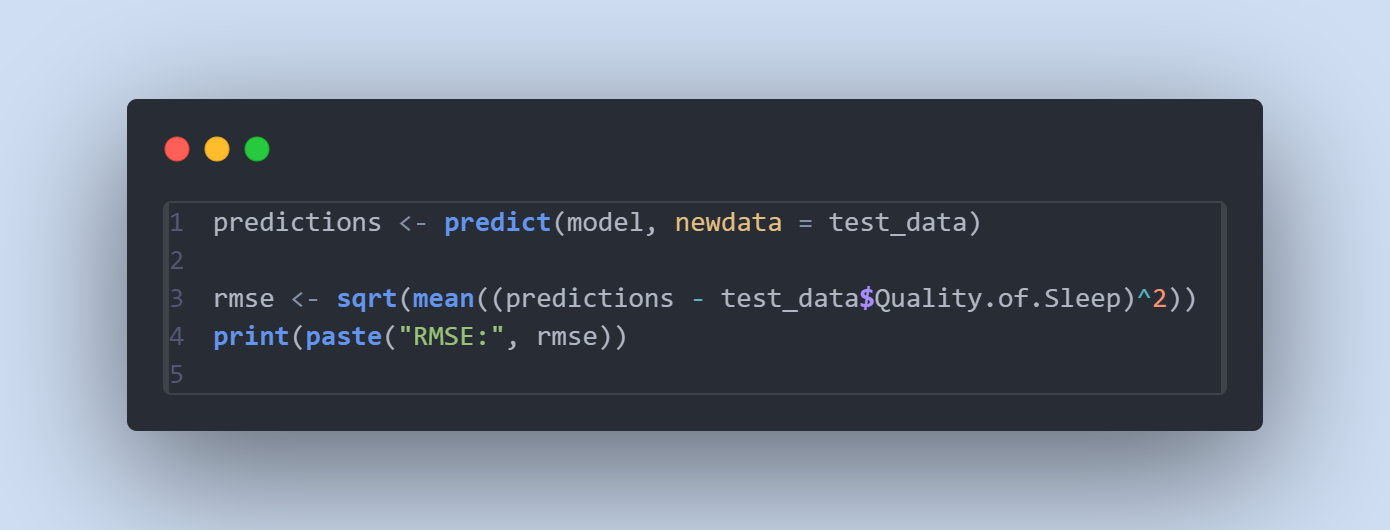
**Occupation Software Engineer** หากทำอาชีพ Software Engineer   
ทำให้คุณภาพการนอนลดลง 0.5302 หน่วย

**Occupation Teacher** หากทำอาชีพครูทำให้คุณภาพการนอนลดลง 0.7205 หน่วย

**Stress Level** หากระดับความเครียดเพิ่มขึ้น 1 หน่วยทำให้คุณภาพการนอนลดลง 0.4693 หน่วย

**Daily Steps** หากเดินในแต่ละวันเพิ่มขึ้น 1 ก้าวทำให้คุณภาพการนอนเพิ่มขึ้น 0.00001018 หน่วย

ทีนี้ผมจะมาลอง Run model ตัวเดิมแต่เปลี่ยนข้อมูลเป็นชุด test ที่มีข้อมูลแค่ 20% ของข้อมูลทั้งหมดและดูว่าสามารถพยากรณ์ได้ใกล้เคียงข้อมูลชุด train แค่ไหน



Output:

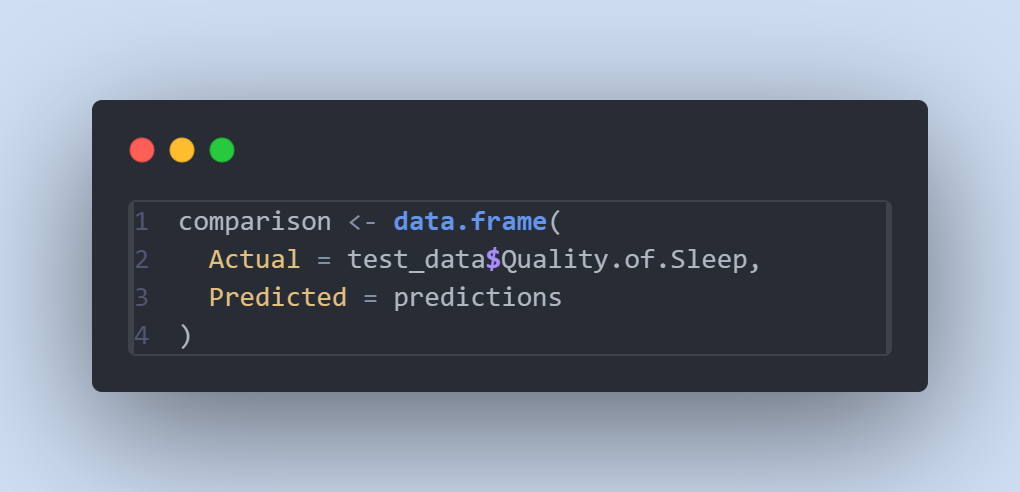
A screen shot of a computer

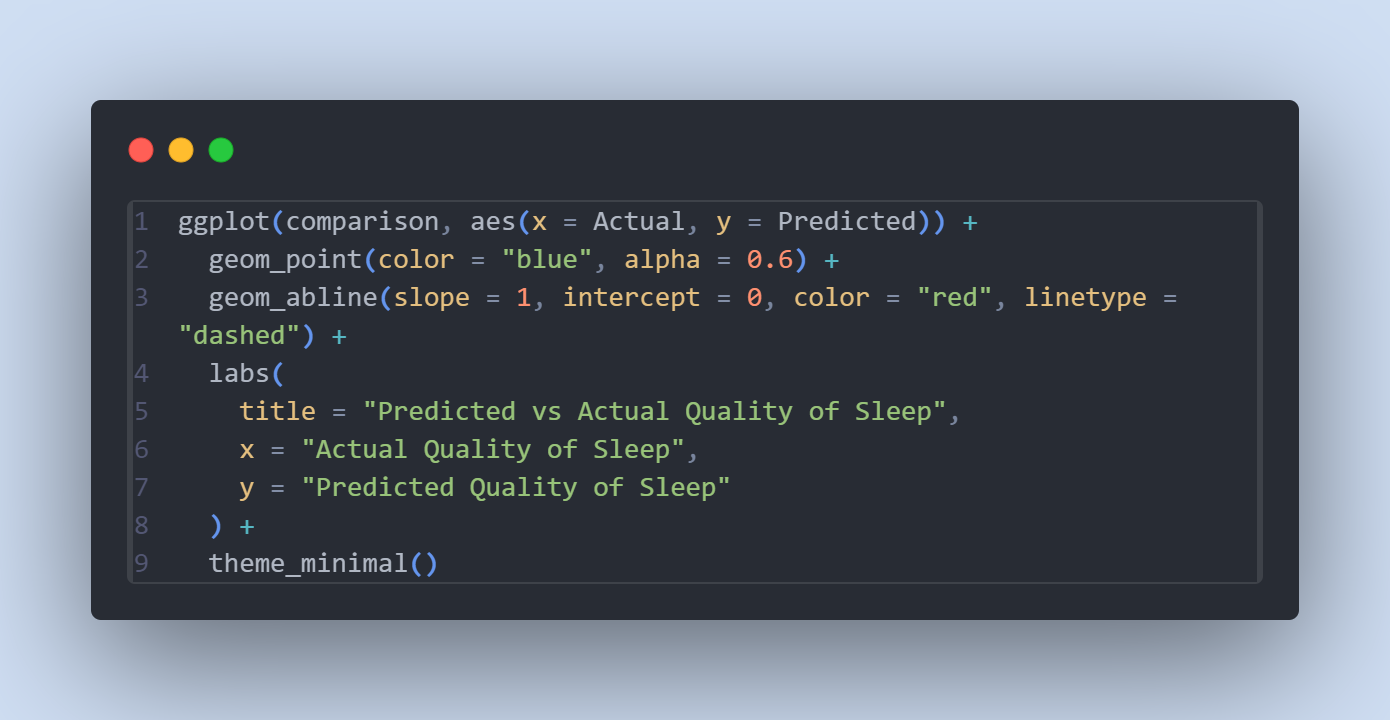
AI-generated content may be incorrect.

ค่า Root mean square error :RMSE เป็นค่าที่บอกว่าข้อมูลชุด test มีความแม่นยำมากแค่ไหน ยิ่งค่า RMSE มีค่าน้อยเท่าไร ยิ่งดี ในที่นี้ RMSE = 0.2696 หมายความว่าโดยเฉลี่ยแล้ว ค่าที่โมเดลที่ทำการพยากรณ์ (predictions) จะมีความคลาดเคลื่อนจากค่าจริง (Quality of Sleep) ประมาณ 0.27 หน่วย

ทำการ plot เปรียบเทียบข้อมูลชุด test กับค่าจริง Quality of Sleep

สร้าง data frame ในรูปตัวแปร comparison เพื่อจะนำข้อมูลไป plot เปรียบเทียบ



ใช้ ggplot2 เพื่อทำการ plot เปรียบเทียบ

**รูปกราฟที่ได้**

A graph with red and blue dots

AI-generated content may be incorrect.

จุดส่วนใหญ่ (สีน้ำเงิน) เรียงตัวใกล้กับ เส้นแดง (y = x) แสดงว่าโมเดลพยากรณ์ได้ดี

ค่าคาดการณ์ (Predicted Quality of Sleep) อยู่ในช่วงเดียวกับค่าจริง (Actual Quality of Sleep)

ดู Source Code ทั้งหมดได้ที่ <https://github.com/JINX23606/R-Project.git>