Sleep Health and Lifestyle Dataset

第五組 生資碩 吳宥禎、數據所 陳俞君

研究背景

睡眠是維持人類生理與心理健康的關鍵因素。擁有充足且高品質的睡眠和免疫功能、情緒穩定、認知表現以及心血管健康密切相關。根據世界衛生組織與美國睡眠醫學學會的建議，成人每日應有7~小時的睡眠時長，但近年來面臨睡眠不足、失眠與睡眠障礙的情形日益嚴重，根據多國調查資料顯示，現代人的睡眠品質正逐漸下降，睡眠時間也呈現縮短趨勢。 造成睡眠問題的原因眾多，包含快節奏的生活型態與不良生活習慣皆可能對睡眠時間與品質產生負面影響。此外，個人特徵如年齡、性別與身體質量指數（BMI）等，也被認為可能與睡眠狀況存在潛在的關係。目前仍缺乏系統性整合多重變項的統計分析，來全面探討影響睡眠的生活型態因子及其相互關係。特別是在COVID-19疫情後，遠距工作與學習的普及改變了人們的作息與生活節奏，也進一步強化了釐清睡眠與生活習慣之間關係的必要性。

研究目的

本研究探討睡眠健康(包含睡眠時數與睡眠品質)與個人特徵(性別、年齡)、體態(BMI 分類)及生活型態(日常步數、身體活動量)之間的關聯性。隨著現代人生活型態改變，影響睡眠的因素日益多樣，本研究希望透過統計分析與視覺化呈現，了解不同特徵與行為模式對睡眠狀況的影響，有助於未來健康促進與生活品質提升之參考。

資料集來源

Sleep Health and Lifestyle Dataset https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset/data

資料集特點

- 睡眠指標：探索睡眠時間(Sleep Duration)、品質(Quality of Sleep)和影響睡眠模式的因素。
- 睡眠障礙分析：辨識失眠和睡眠呼吸中止症等睡眠障礙的發生。
- 生活方式因素：身體活動量(Physical Activity Level)、每日步數(Daily Steps)和壓力程度(Stress Level)
- 個人特徵：年齡(Age)、性別(Gender)和 BMI 類別(BMI Category)。
- 心血管健康：檢查血壓(Blood Pressure)和心率測量值(Heart Rate)。

資料集欄位

欄位名稱	資料型別	中文說明
Person ID	int64	個體ID
Gender	object	性別（男 / 女）
Age	int64	年齡（歲）
Occupation	object	職業
Sleep Duration	float64	睡眠時數（每日小時數）
Quality of Sleep	int64	睡眠品質（1~10主觀評分）
Physical Activity Level	int64	每日身體活動時間（分鐘）
Stress Level	int64	壓力指數（1~10主觀評分）

欄位名稱	資料型別	中文說明
BMI Category	object	身體質量指數分類（過輕、正常、過重等）
Blood Pressure	object	血壓（systolic收縮壓/diastolic舒張壓）
Heart Rate	int64	靜態心率（每分鐘心跳數）
Daily Steps	int64	每日步數
Sleep Disorder	object	是否有睡眠障礙（無、失眠、睡眠呼吸中止症等）

睡眠障礙 Sleep Disorder

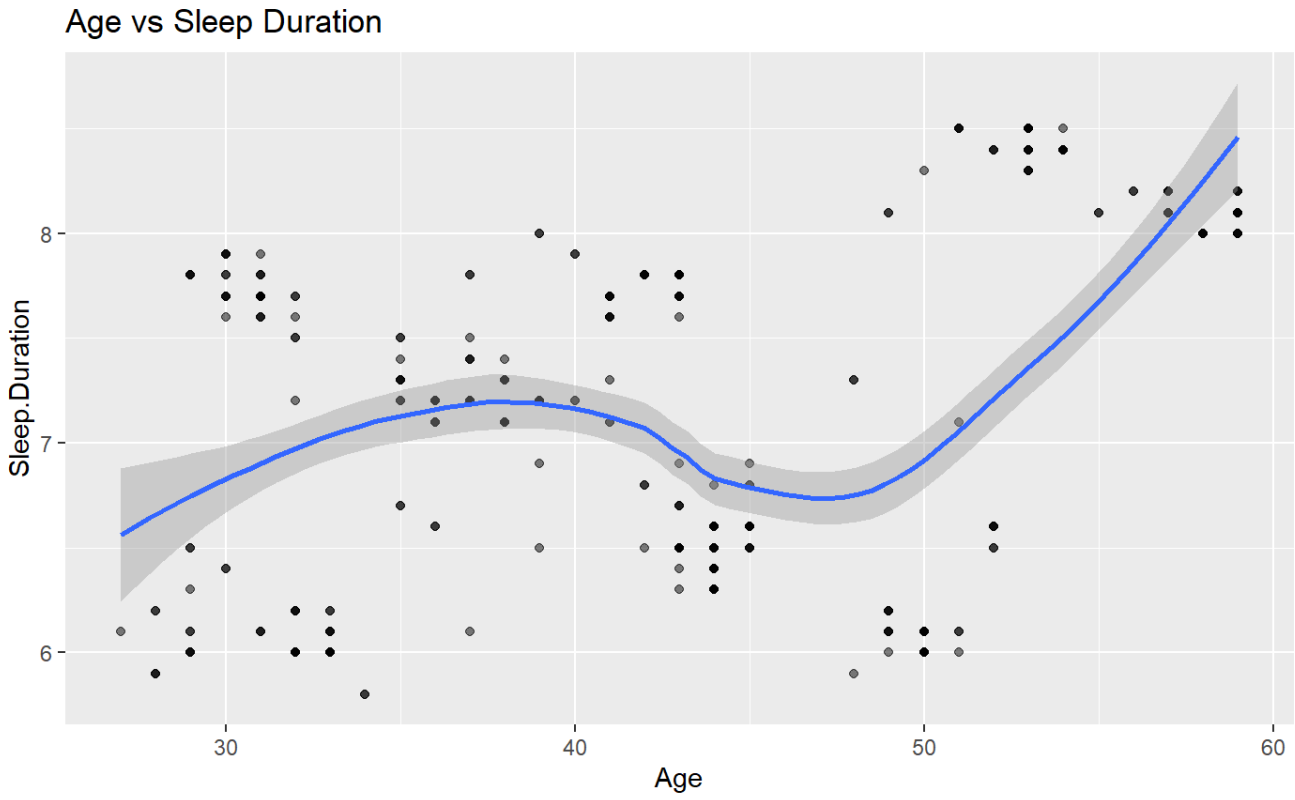
- *None*：個人沒有表現出任何特定的睡眠障礙。
- *Isomnia*：個人入睡困難或難以維持睡眠，導致睡眠不足或睡眠品質不佳。
- *Sleep Apnea*：患者在睡眠期間會出現呼吸暫停，導致睡眠模式紊亂並帶來潛在的健康風險。

Exploratory data analysis

變數	平均值	中位數	最小值	最大值
Sleep Duration	7.13 小時	7.20	5.8	8.5
Quality of Sleep	7.31	7.0	4	9
Age	42.18	43	27	59
Daily Steps	6817	7000	3000	10000
Stress Level	5.39	5	3	8

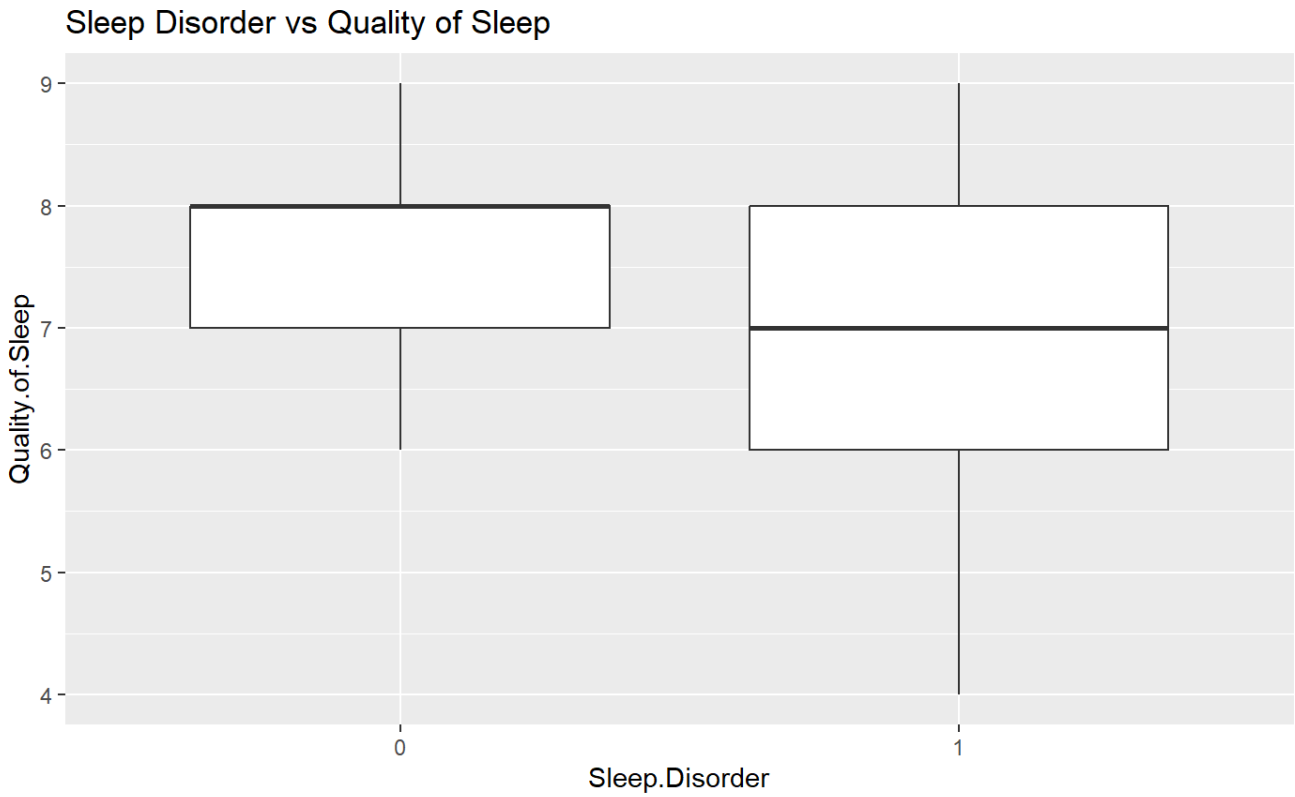
- 年齡分組：
  - Young Adults：≤30
  - Early Middle-aged Adults：30~45
  - Late Middle-aged Adults：≥46
- 性別分布：Female = 185，Male = 189（約略均等）
- BMI 類別：
  - Normal：216
  - Overweight：148
  - Obese：10
- 睡眠品質：評分1~10(分數越高表示睡眠品質越好)
- 睡眠障礙者：41%（155 人有障礙；219 人無障礙）

Age vs Sleep Duration



睡眠時數在年齡 30–45 區段略為平穩，但 50 歲以上出現回升 可能反映高齡者作息調整或退休生活型態改變

Sleep Disorder vs Quality of Sleep



有睡眠障礙者的平均品質中位數略低（約 6 分），無障礙者約 7–8 分 此差異在後續 t-test 與邏輯斯迴歸中驗證顯著

研究問題

1.男性與女性的睡眠時數是否有差異？ 2.不同年齡組別的睡眠時數是否有差異？ 3.不同 BMI 分類的人，其睡眠時數是否有差異？ 4.日常步數與睡眠品質之間是否有關聯？ 5.每日身體活動時間與睡眠品質間是否有關聯？ 6.不同因素是否會影響睡眠時數或睡眠品質？ 7.哪些變數能有效解釋睡眠時數？ 8.哪些因素會顯著影響是否有睡眠障礙？

研究結果

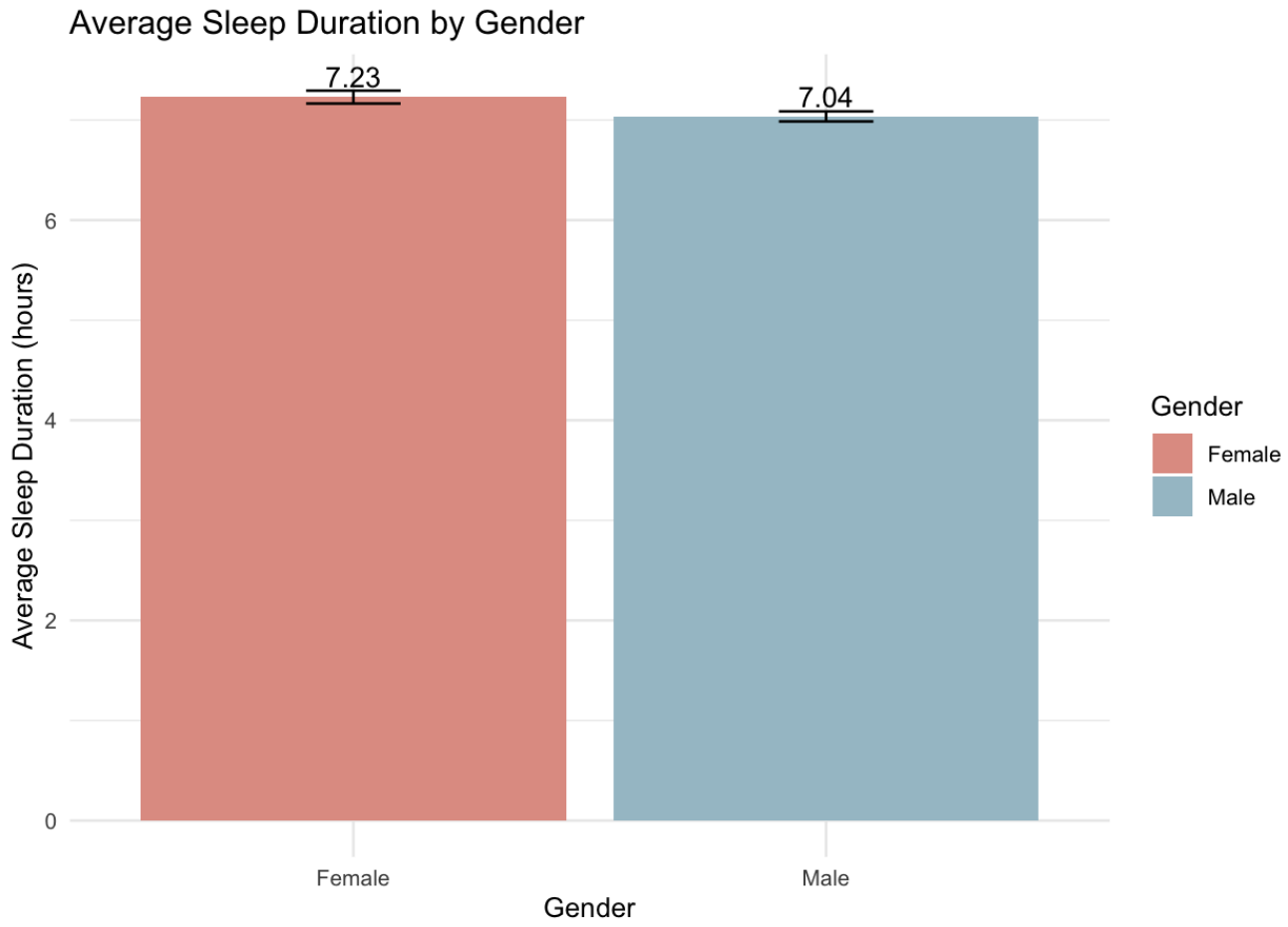
問題1-1：男性與女性的睡眠時數是否有差異？

檢定性別對睡眠時數是否有顯著影響

- **方法**：Two-Sample t-Test（兩樣本獨立 t 檢定）
- **假設**：
  - 虛無假設（H_0）：男女的平均睡眠時數無差異（$\mu_{\text{Male}} = \mu_{\text{Female}}$）
  - 對立假設（H_a）：男女的平均睡眠時數不同（$\mu_{\text{Male}} \neq \mu_{\text{Female}}$）

項目	數值
t 值	2.3565
自由度	約 349.38
p 值	0.019 (**顯著**)
信賴區間 (95%)	(0.032, 0.355)
女性平均	7.23 小時
男性平均	7.04 小時

根據 Welch 兩樣本 t 檢定結果（t = 2.36, df ≈ 349, p = 0.019），在 5% 顯著水準下拒絕虛無假設，認為男性與女性的平均睡眠時數存在顯著差異。女性的平均睡眠時數顯著高於男性，差異約為 0.19 小時（約 11 分鐘）。

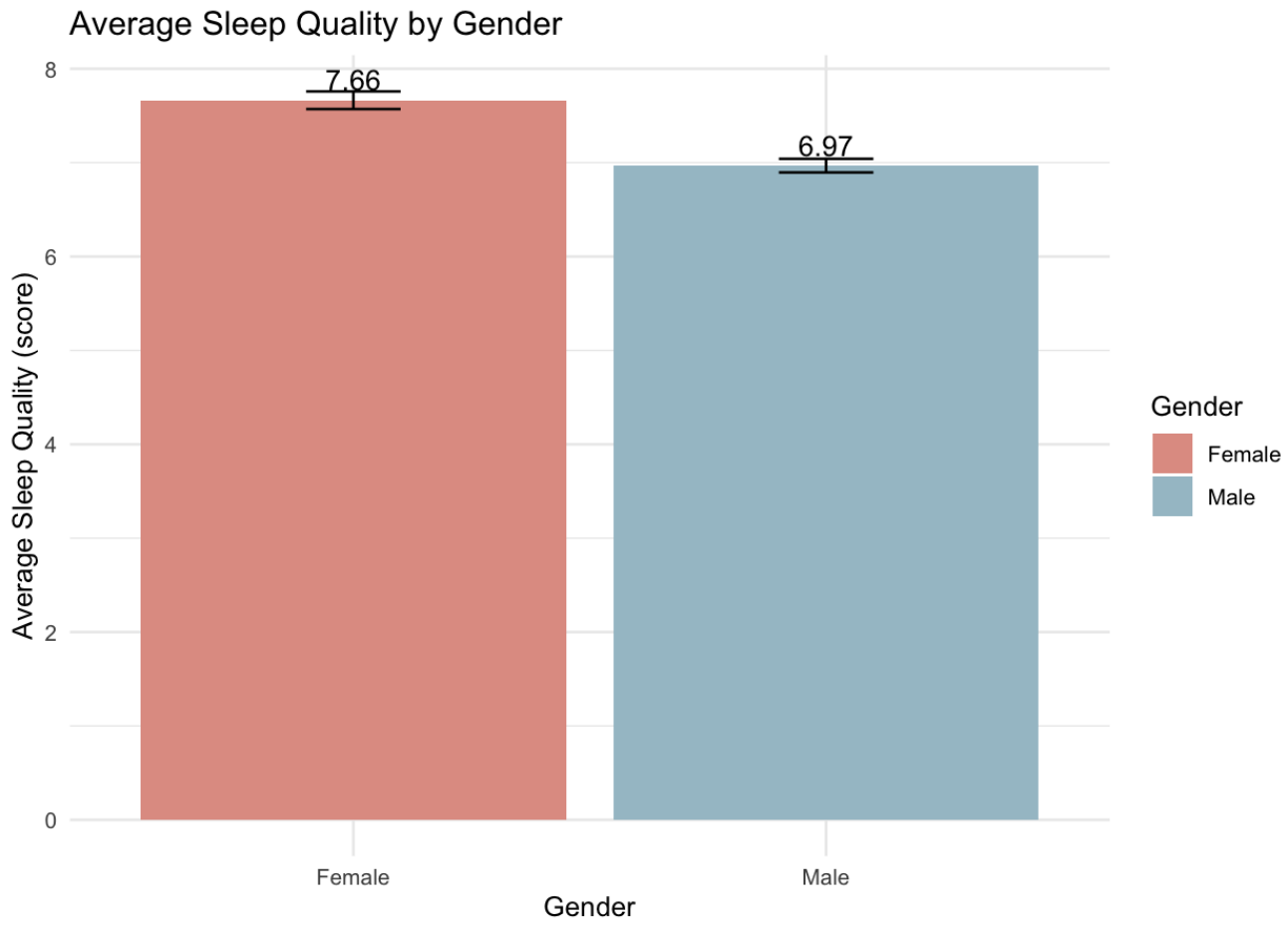


問題1-2：男性與女性的睡眠品質是否有差異？

- **方法**：Two-Sample t-Test
- **假設**：
  - 虛無假設（H_0）：男女的睡眠品質平均數相等（$\mu_{\text{Male}} = \mu_{\text{Female}}$）
  - 對立假設（H_a）：男女的睡眠品質平均數不等（$\mu_{\text{Male}} \neq \mu_{\text{Female}}$）

項目	數值
t 值	5.8593
自由度	約 347.96
p 值	1.078×10^{-8} **(高度顯著)**
信賴區間 (95%)	(0.463, 0.930)
女性平均	7.6649
男性平均	6.9683

根據 Welch 兩樣本 t 檢定（$t = 5.86$, $df \approx 348$, $p < 0.001$），顯示男性與女性的睡眠品質平均數存在高度顯著差異。女性的平均睡眠品質（7.66 分）顯著高於男性（6.97 分），差異約為 0.7 分。



問題2-1：不同年齡組別的睡眠時數是否有差異？

分析年齡是否顯著影響睡眠時數

- **方法**：One-Way ANOVA（單因子變異數分析）
- **假設**：
  - 虛無假設（H_0）：各年齡組別的平均睡眠時數相等
  - 對立假設（H_a）：至少有兩組年齡的平均睡眠時數不同

項目	數值
F 值	17.85
p 值	3.96e-08 (**高度顯著**)
組間平方和 (SSB)	20.73
組內平方和 (SSW)	215.41
自由度 (組間)	2
自由度 (組內)	371
Mean Sq (組間)	10.365
Mean Sq (組內)	0.581

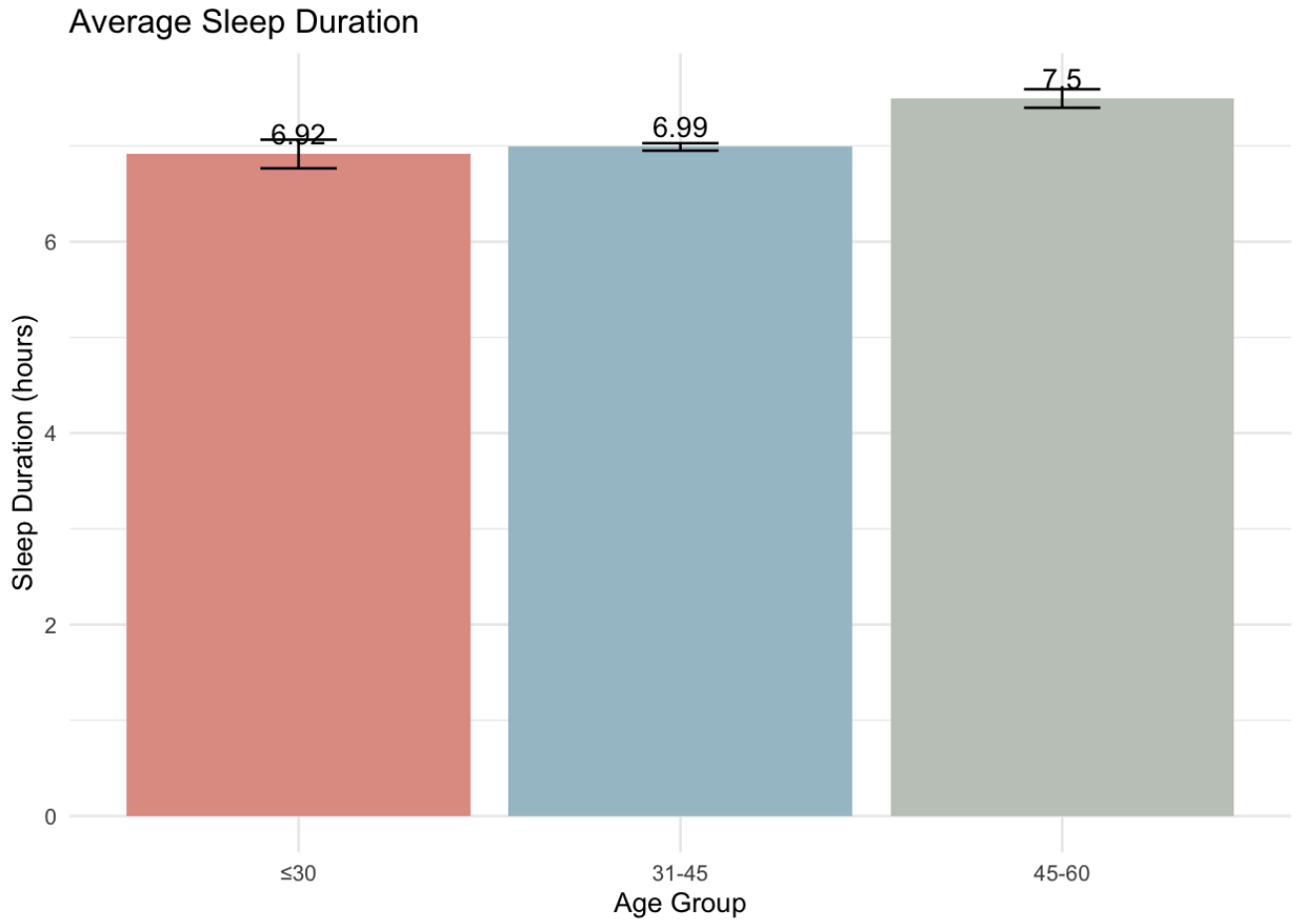
根據單因子變異數分析（$F(2, 371) = 17.85$, $p < 0.001$），不同年齡組別在平均睡眠時數上具有統計顯著差異，且年齡為解釋平均睡眠時數變異的顯著因子。

延伸問題: 哪些年齡組別在睡眠時數上有顯著差異？

比較組別	差異值 (diff)	95% 信賴區間 (lwr, upr)	調整後 p 值 (p adj)	結論
31–45 vs ≤30	0.074	-0.264 ~ 0.412	0.864	不顯著
46–60 vs ≤30	0.580	0.220 ~ 0.940	0.0005	顯著
46–60 vs 31–45	0.506	0.298 ~ 0.713	< 0.0001	顯著

根據 Tukey 事後比較顯示，46–60 歲組的平均睡眠時數顯著高於其他年齡組別。具體來說，與 ≤30 歲組的差異為 0.58 小時（$p = 0.0005$），與 31–45 歲組的差異為 0.51 小時（$p < 0.0001$）。而 ≤30 與 31–45 歲兩組之間則無顯著

差異（p = 0.864）。



問題2-2：不同年齡組別的睡眠品質是否有差異？

- **方法**：One-Way ANOVA
- **假設**：
  - H_0：所有年齡組的平均睡眠品質相同
  - H_a：至少有一組的平均品質與其他不同

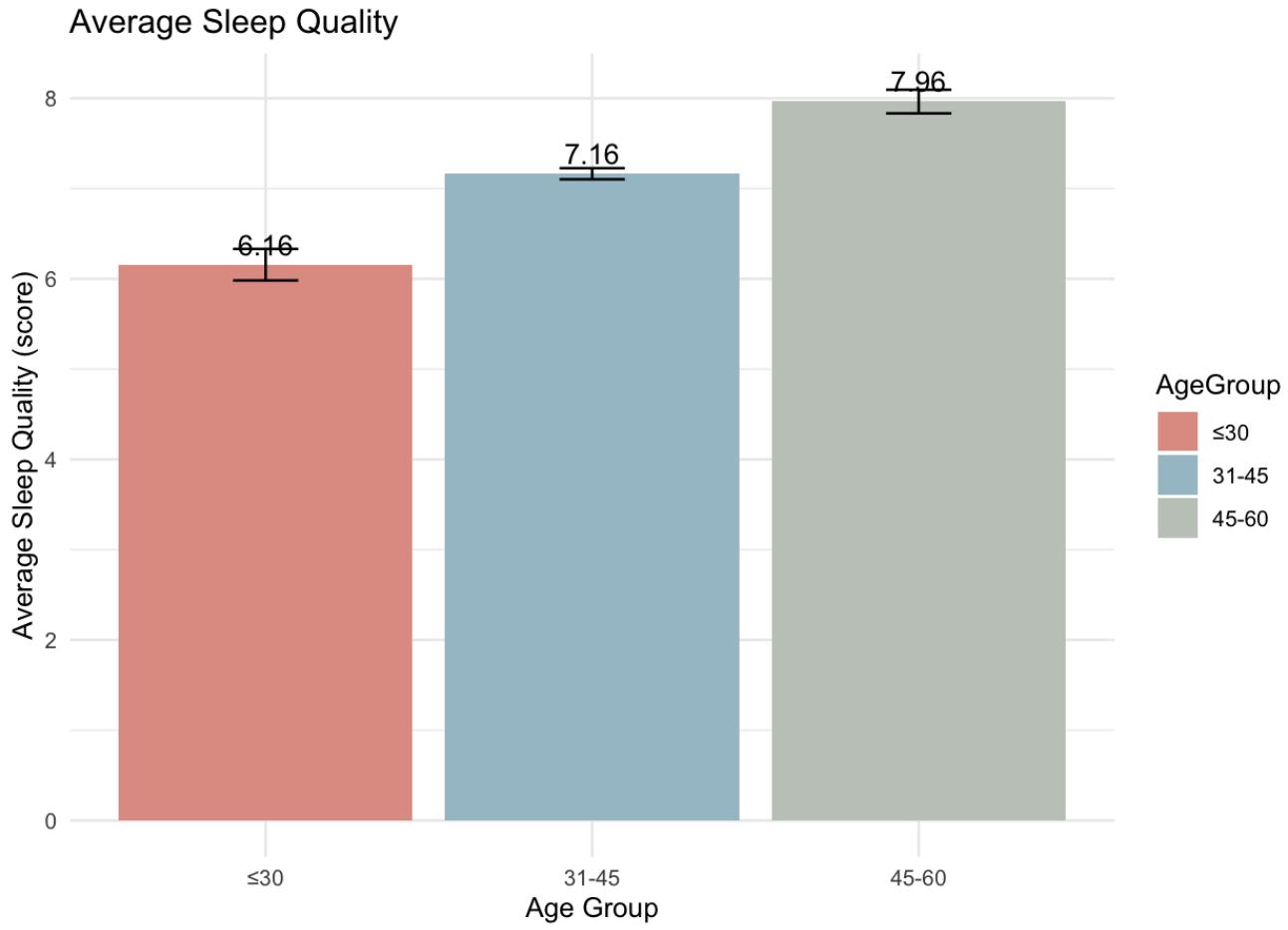
項目	數值
F 值	39.88
p 值	< 2e-16 (**極顯著**)
SSB（組間平方和）	94.5
SSW（組內平方和）	439.8
自由度（組間）	2
自由度（組內）	371
MSB（組間均方）	47.27
MSW（組內均方）	1.19

單因子變異數分析結果顯示，不同年齡組別在睡眠品質上存在顯著差異（$F(2, 371) = 39.88$, $p < 0.001$），年齡為解釋睡眠品質變異的顯著因子。

延伸問題: 哪些年齡組別在睡眠品質上有顯著差異？

年齡比較組別	差異值 (diff)	95% 信賴區間 (lwr, upr)	調整後 p 值 (p adj)	結論
31–45 vs ≤30	1.01	(0.52, 1.49)	4.1e-06	顯著
46–60 vs ≤30	1.81	(1.29, 2.32)	0.0000	顯著
46–60 vs 31–45	0.80	(0.50, 1.10)	0.0000	顯著

Tukey Posteriori comparison 事後檢定顯示，不同年齡組在睡眠品質上皆有顯著差異。31–45 歲族群的睡眠品質平均分數比 ≤30 歲族群高約 1 分（$p < 0.001$），46–60 歲族群的睡眠品質更高於其他兩組，與 ≤30 歲差異約 1.81 分，與 31–45 歲差異為 0.80 分，均具有極高統計顯著性（$p < 0.001$）。



問題3：不同 BMI 分類的人，其睡眠時數是否有差異？

- **方法**：One-Way ANOVA
- **假設**：
  - H_0：不同 BMI 類別的睡眠時數平均數相同
  - H_a：至少有一組 BMI 分類的平均睡眠時數不同

項目	數值
F 值	20.66
p 值	2.13×10^{-12} **(高度顯著)**
組間平方和 (SSB)	33.88
組內平方和 (SSW)	202.25
自由度（組間）	3
自由度（組內）	370
組間均方 (MSB)	11.294
組內均方 (MSW)	0.547

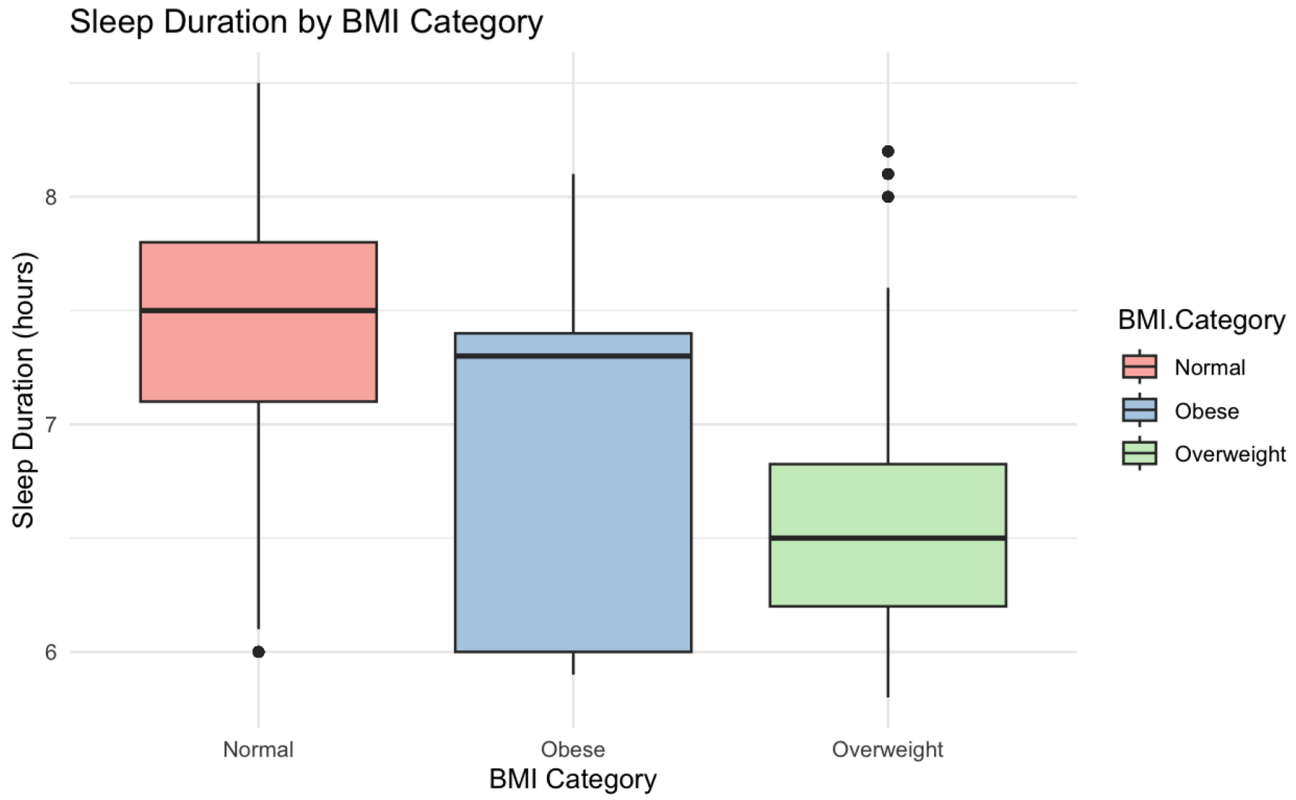
根據單因子變異數分析（$F(3, 370) = 20.66, p < 0.001$），不同 BMI 分類的個體在平均睡眠時數上具有統計顯著差異。

延伸問題: 哪些BMI組別在平均睡眠時數上有顯著差異？

比較組別	差異 (diff)	95% 信賴區間 (lwr, upr)	p 值 (調整後)	結論
Normal Weight - Normal	-0.061	(-0.498, 0.378)	0.984	不顯著
Obese - Normal	-0.433	(-1.052, 0.184)	0.270	不顯著
Overweight - Normal	**-0.624**	**(-0.832, -0.416)**	**< 0.0001**	顯著
Obese - Normal Weight	-0.373	(-1.106, 0.360)	0.554	不顯著
Overweight - Normal Weight	**-0.563**	**(-1.008, -0.118)**	**0.006**	顯著
Overweight - Obese	-0.189	(-0.813, 0.434)	0.861	不顯著

Tukey 事後比較顯示，Overweight（過重）族群的睡眠時數顯著低於 Normal（正常體重）與 Normal Weight 類別（$p < 0.001$ 與 $p = 0.006$）。其他組別間並無顯著差異。這顯示體重過重可能與較低的平均睡眠時數有關，值得進

一步探討其行為或生理因素。



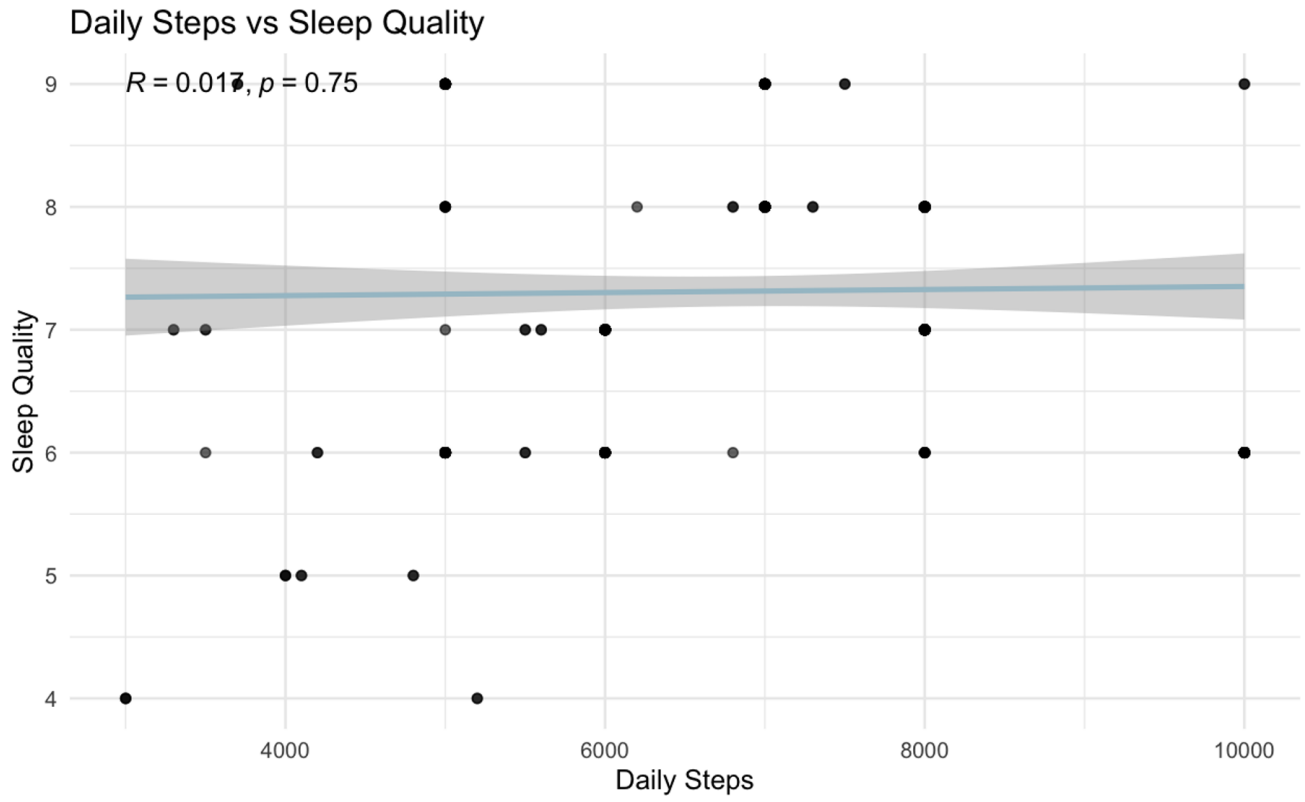
問題4：日常步數與睡眠品質之間是否有關聯？

檢定 Daily Steps 是否與睡眠品質有線性相關

- **方法**：
  - Pearson correlation（常態分布）
  - Spearman correlation（階層或非正態）
- **假設**：
  - H_0: $\rho = 0$（日常步數與睡眠品質之間無線性相關）
  - H_a: $\rho \neq 0$（有線性相關）

項目	數值
相關係數 (r)	0.0168
p 值	0.7462
95% 信賴區間	(-0.0848, 0.1180)
t 統計量	0.3239
自由度 (df)	372

根據 Pearson 相關分析，日常步數與睡眠品質之間幾乎沒有線性關聯（r = 0.017，p = 0.746）。我們無法拒絕虛無假設，表示兩者間的相關性在統計上並不顯著

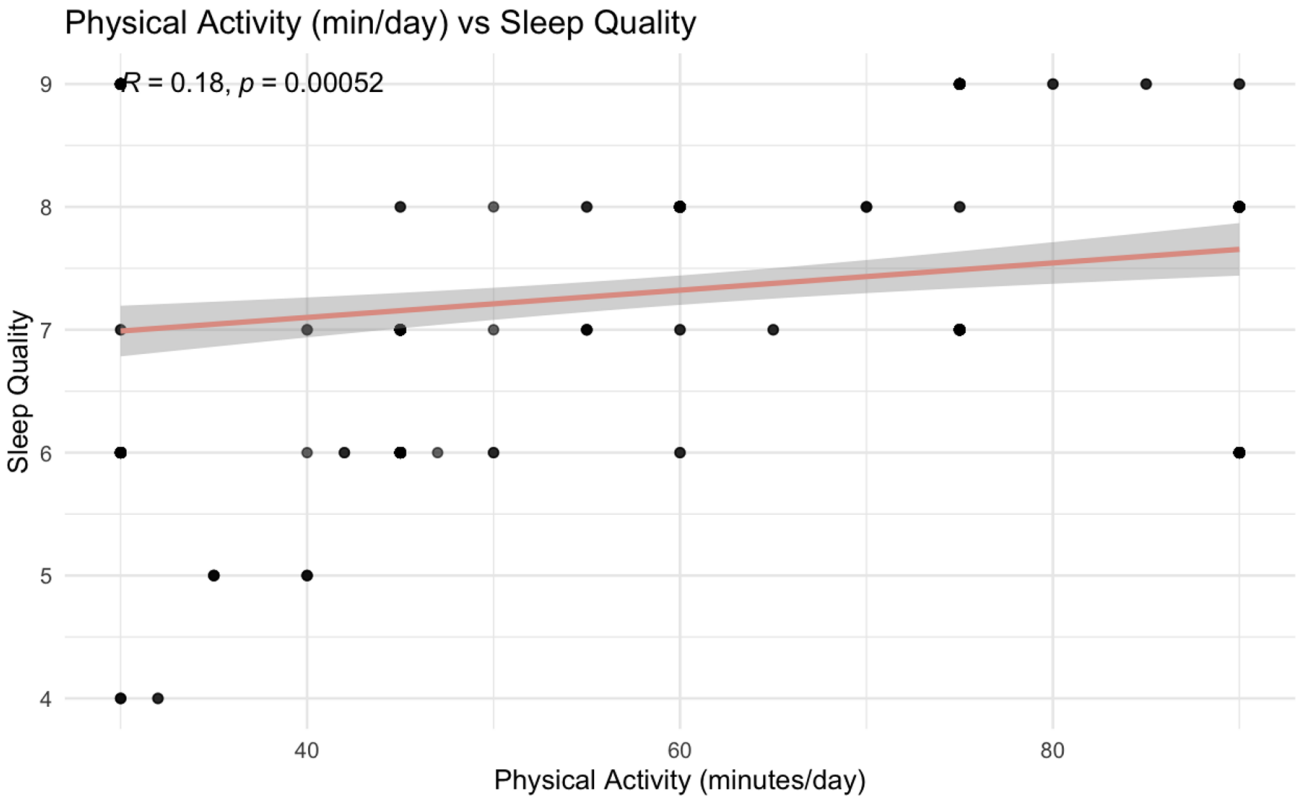


問題5：每日身體活動時間與睡眠品質間是否有關聯？

- **方法**：
  - Spearman correlation（階層或非正態）
- **假設**
  - H_0: $\rho = 0$（每日活動時間與睡眠品質無等級相關性）
  - H_a: $\rho \neq 0$（有相關性）

項目	數值
相關係數 (rho)	0.1785
p 值	0.00052 (**顯著**)
檢定統計量 (S)	7162849

根據 Spearman 等級相關分析，每日身體活動時間與睡眠品質之間具有統計顯著的正向相關（rho = 0.179，p = 0.0005）。雖然相關性不高，但結果顯示活動量較多者的睡眠品質傾向較佳。



問題6：不同因素是否會影響睡眠時數或睡眠品質？

- **方法**：Multiple Linear Regression
- **模型公式**：

$$Sleep\ Duration = \beta_0 + \beta_1 \cdot Age + \beta_2 \cdot Gender + \beta_3 \cdot BMI\ Category + \beta_4 \cdot Daily\ Steps + \beta_5 \cdot Stress\ Level + \varepsilon$$

- **推論**：檢定變數對睡眠影響的顯著性（$p < 0.05$）

變數	Estimate	t 值	p 值	統計顯著性
(Intercept)	6.954	38.5	< 2e-16	***
Age	0.03795	11.63	< 2e-16	***
GenderMale	0.4314	9.16	< 2e-16	***
BMI: Normal Weight	0.1139	1.42	0.158	ns
BMI: Obese	-0.1851	-1.54	0.126	ns
BMI: Overweight	-0.6283	-12.08	< 2e-16	***
Daily Steps	4.25e-05	3.47	0.0006	***
Stress Level	-0.3126	-22.7	< 2e-16	***

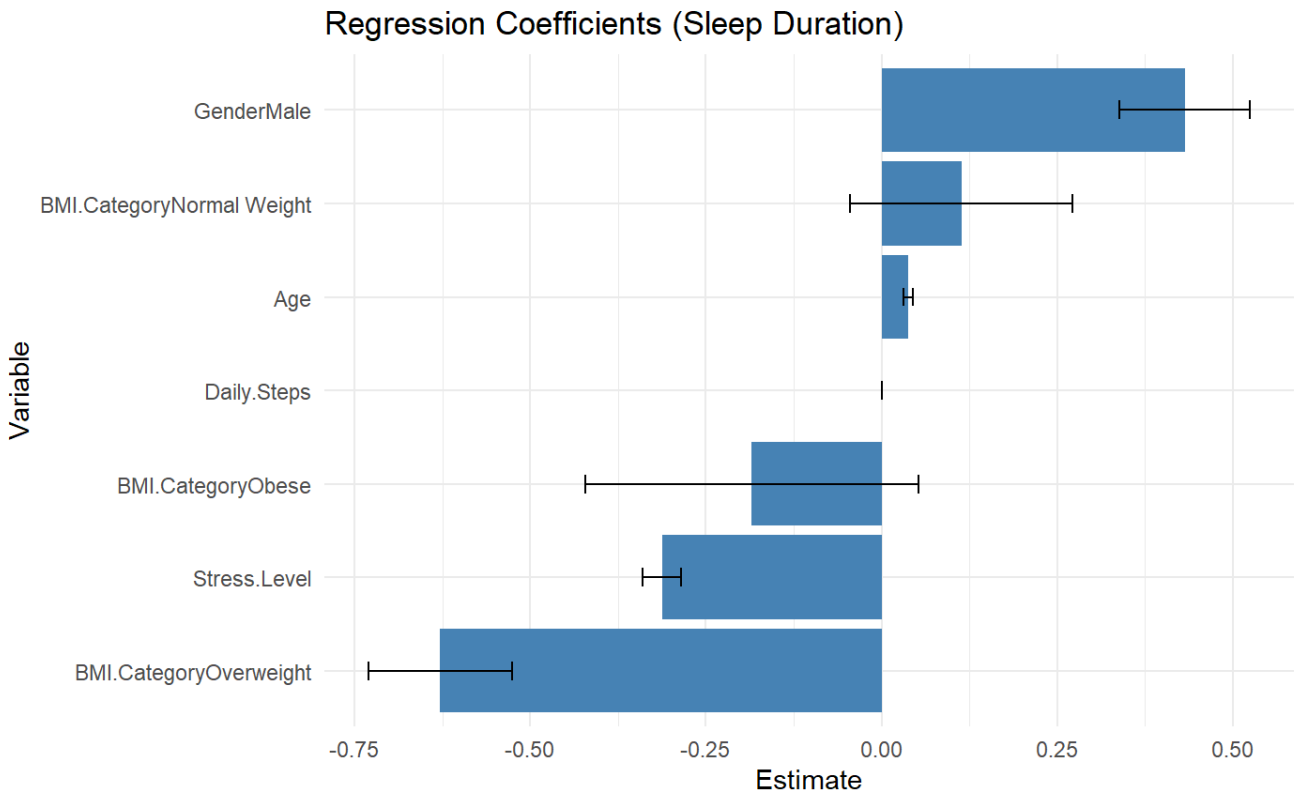
- Multiple R-squared：0.8174 → 表示模型解釋了 81.7% 的變異
- Adjusted R-squared：0.8139 → 模型良好
- F 檢定：整體模型顯著（p < 2.2e-16）

統計推論與解釋：

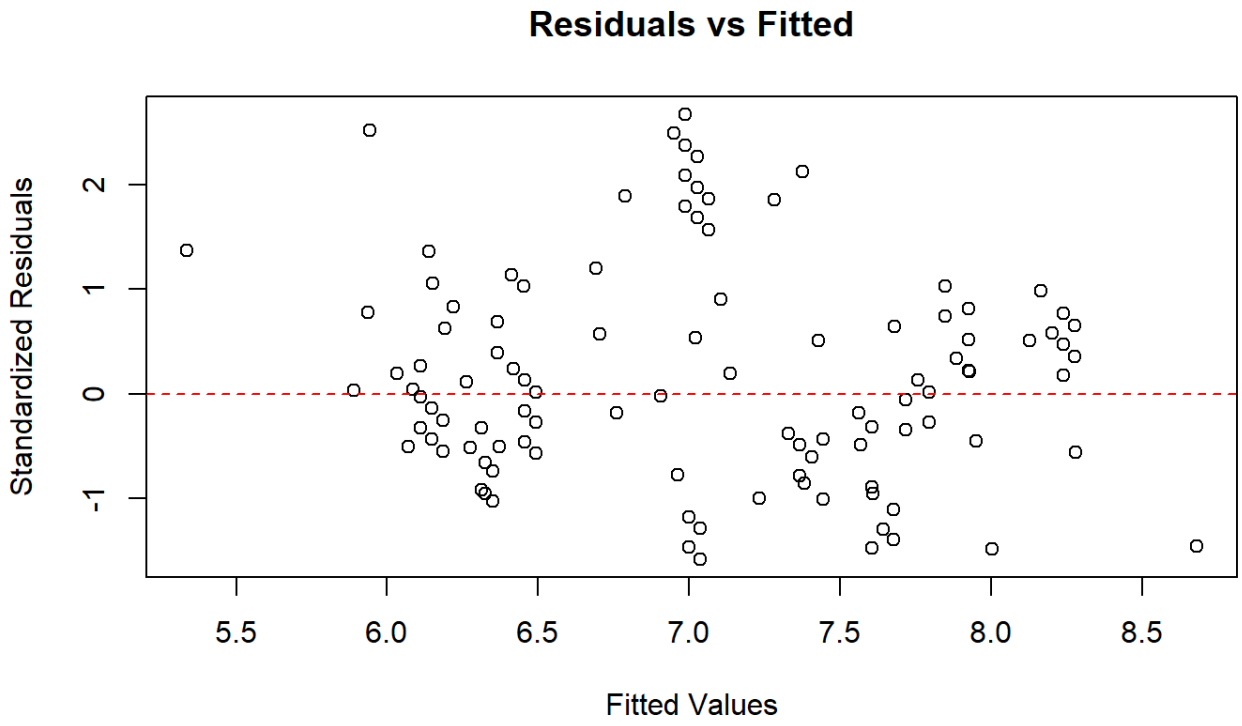
年齡越大 → 睡眠時數略增加（每增 1 歲多約 0.038 小時） 男性 睡得比女性多約 0.43 小時 過重者 睡得顯著更少（約少 0.63 小時） 壓力越高 睡眠時數顯著減少（每多 1 分壓力少 0.31 小時） 日常步數越多 睡得略多（但效果小）

無顯著變數（p > 0.05）：

- BMI: Normal Weight
- BMI: Obese → 表示與基準組 (BMI = Normal) 在統計上差異不顯著



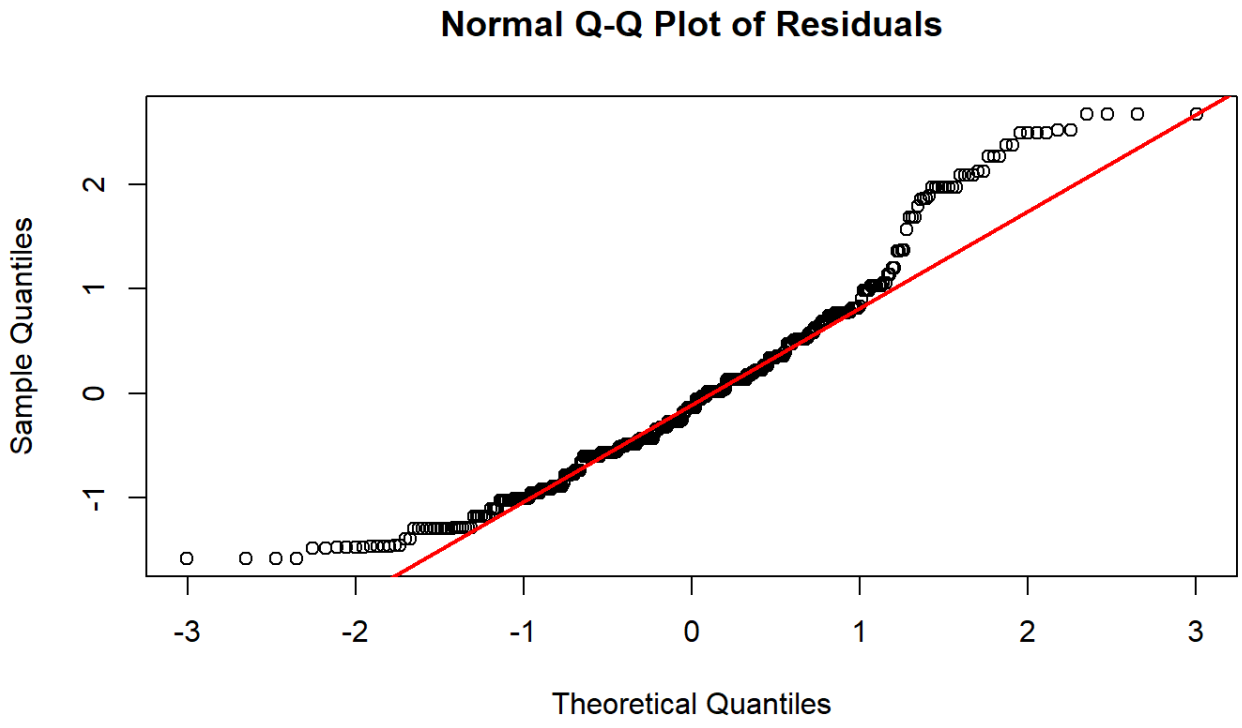
- GenderMale 和 Age 的係數為正 → 男性與年齡較長者傾向睡得更多
- Stress.Level 與 BMI.CategoryOverweight 係數為負 → 壓力大與過重者傾向睡得較少
- 水平線表示信賴區間，若穿越 0 → 該變數不顯著（如 Normal Weight）



根據多元線性迴歸模型（$R^2 = 0.817$，$p < 0.001$），年齡、性別、BMI 類別、日常步數與壓力程度對睡眠時數具統計顯著影響。

其中，**過重（Overweight）**者的睡眠時數顯著低於正常者（約少 0.63 小時，$p < 0.001$），而**壓力程度越高**者睡眠時數亦顯著下降（每單位壓力少約 0.31 小時，$p < 0.001$）。男性與年齡較大者則傾向擁有較長的睡眠時間。

係數條形圖顯示上述變數的效應大小與方向，而殘差圖則未顯示明顯偏態或異常結構，支持模型適配性良好。



- 中央區域（約 -1 到 +1）大致落在紅線上，表示大多數殘差近似常態分佈
- 兩端（左下與右上）有輕微偏離紅線，暗示少數極端值（尾端殘差）與常態分布不完全一致

- 整體而言：殘差呈現近似常態分佈，符合線性迴歸的基本假設，沒有明顯違反常態性

QQ plot 顯示大多數標準化殘差分布近似常態，僅於分布尾端出現些微偏離理論常態線。整體而言，模型殘差可視為滿足常態性假設，適用於進行迴歸估計與推論。

問題7：哪些變數能有效解釋睡眠時數？

選擇具統計顯著性且可提升模型解釋力的變數組合

- **方法**：Stepwise Regression
- **選擇準則**：AIC（Akaike Information Criterion）

stepAIC() 沒有移除任何變數，代表在 AIC 準則下：

1. 所有變數的組合已是「最佳模型」
2. 即使某些變數（例如 BMI.CategoryNormal Weight）在統計上不顯著（$p > 0.05$），但它們對整體模型的 預測能力仍有貢獻，所以沒有被 stepAIC() 剔除

經由逐步回歸（stepwise regression, AIC 準則）分析，最終模型與原始全變數模型相同，顯示所有變數的組合可提供最佳預測能力。雖然部分變數在統計上不具顯著性，但整體模型的 AIC 已達最小，表示其仍具有貢獻。模型最終保留 Age、Gender、BMI 分類、Daily Steps 與 Stress Level 等變數。

問題8：哪些因素會顯著影響是否有睡眠障礙？

- **方法**：Firth Logistic Regression

變數	OR（勝算比）	95% 信賴區間	p 值	統計顯著性
Age	1.17	1.07 – 1.29	0.00032	顯著 ↑
GenderMale	0.85	0.28 – 2.55	0.776	不顯著
BMI: Normal Weight	1.56	0.30 – 6.57	0.576	不顯著
BMI: Obese	172.6	14.97 – 24522.90	2.13e-06	顯著 ↑
BMI: Overweight	9.86	2.91 – 34.25	2.61e-04	顯著 ↑
Stress Level	0.70	0.37 – 1.29	0.256	不顯著
Quality of Sleep	0.21	0.06 – 0.60	0.0035	顯著 ↓
Daily Steps	1.00	1.000 – 1.000	0.345	不顯著

本研究使用 Firth Logistic Regression 模型分析睡眠障礙的影響因子，以修正傳統邏輯斯模型中的 separation 現象。結果顯示：

- **年齡**為正向風險因子（$OR = 1.17$, $p < 0.001$），即年齡越大，越易出現睡眠障礙。
- BMI 分類為過重（$OR = 9.86$）或肥胖（$OR = 172.6$）者，其罹患風險遠高於正常體重者（$p < 0.001$）。
- **睡眠品質則為保護因子**，每提升 1 分，其罹患風險下降約 79%（$OR = 0.21$, $p = 0.003$）。

性別、壓力程度與日常步數等變數在模型中不具顯著性。整體而言，BMI 與主觀睡眠品質是影響睡眠障礙最關鍵的因子。

結論

本研究針對 Sleep Health and Lifestyle Dataset 進行探索性資料分析與統計推論，探討影響睡眠狀態的關鍵因素。結合描述統計、推論統計與迴歸模型，歸納主要結論如下：

1. Exploratory Data Analysis Result

- 平均睡眠時數為 7.13 小時，中位數為 7.2，略低於建議的 7–9 小時。
- 睡眠品質平均為 7.31 分，但存在尾端偏低現象，顯示部分人群主觀睡眠狀況不佳。
- 約 41% 樣本有睡眠障礙（Insomnia 或 Sleep Apnea）。
- 年齡分佈集中於 27–59 歲，Daily Steps 平均約 6817 步。
- BMI 類別中以 Normal 與 Overweight 為主，Obese 組樣本稀少。

2. 針對研究問題的統計發現

1. 問題 1–2：性別是否影響睡眠？
   - 女性睡眠時數與品質皆顯著高於男性（$p = 0.019$ 與 $p < 1e\text{-}7$）。
   - 差異分別為 0.19 小時（約 11 分鐘）與 0.7 分，具統計意義。
2. 問題 3–4：年齡是否影響睡眠？
   - 年齡對睡眠時數與品質皆有顯著群體差異（ANOVA $p < 0.001$）。
   - 46–60 歲族群睡得最多且品質最佳，與 30 歲以下組存在顯著差距。
3. 問題 5：BMI 是否影響睡眠時數？
   - 整體有顯著差異（$p < 0.001$），Overweight 組睡眠時數顯著較少。
   - Obese 組因樣本小無顯著性，但方向一致。
4. 問題 6–7：活動與睡眠品質之關聯？
   - Daily Steps 與睡眠品質無顯著相關（$r = 0.017$, $p = 0.75$）。
   - Physical Activity Level 與睡眠品質有 弱正相關（Spearman rho = 0.18, $p < 0.001$）。

3. 迴歸模型推論總結

多元線性迴歸（Sleep Duration）：解釋力強（Adjusted $R^2 = 0.81$）。

顯著變數包括：

- Age（+）：年齡越大睡越多
- GenderMale（+）：男性略多
- Overweight（−）：過重者睡較少（-0.63 hr）
- Stress Level（−）：每多 1 分壓力，少 0.31 小時
- Daily Steps（+）：每步貢獻極小，但統計上顯著
- Firth Logistic Regression（Sleep Disorder）：解決 separation 問題後，模型穩定。

顯著風險因子為：

- Age（OR = 1.17）
- BMI: Overweight（OR = 9.86）

- BMI: Obese（OR = 172.6）
- Quality of Sleep 為保護因子（OR = 0.21）

4. Conculsion

1. 年齡、過重與睡眠品質為最主要的關聯因子，顯著影響睡眠時數與睡眠障礙的風險。
2. . 每日步數與睡眠品質無明顯關聯（Pearson $r \approx 0.017$，$p > 0.7$），但「每日活動量」與睡眠品質呈現弱正相關（Spearman $\rho \approx 0.18$，$p < 0.001$）。推測總體活動水準仍具意義，未來研究建議可細分活動強度（如中高強度 METs 指標）。
3. BMI 分類對睡眠影響顯著：Overweight 與 Obese 組的睡眠時數與品質皆較差。顯示維持正常體重與健康生活型態對睡眠具保護作用。
4. 壓力指數雖在 Logistic 模型中非顯著（$p > 0.2$），但在線性模型中顯示顯著負向影響睡眠時數，顯示壓力管理仍是潛在因子。

延伸討論

根據 Wang 等人（2023）發表於 Heliyon 的研究，針對 954 名受試者的睡眠與生活型態資料進行分析，主要發現如下：

- 過重與肥胖顯著提升失眠風險與睡眠中斷頻率，與本研究結果一致。
- 研究指出 BMI 增加 1 單位，睡眠品質指數（PSQI）增加 0.16 分（表示睡眠品質變差）。
- 壓力與焦慮水準是顯著的睡眠障礙預測因子，亦呼應本研究中壓力與睡眠時數之負向關聯。
- 此研究亦強調日常活動量為潛在保護因子，建議將身體活動分類（低、中、高強度）納入評估。

Wang, Y., Xie, Z., Liu, Y., Wang, Y., & Liu, X. (2023). Sleep quality and lifestyle factors in different body mass index categories. Heliyon, 9(11), e22027. https://doi.org/10.1016/j.heliyon.2023.e22027

螢幕錄製(Screencast)

youtube連結 : https://youtu.be/Q-VkrFebxcQ