The Sleep Cipher:

Deciphering the Hidden Language of Disrupted Sleep

Name: Pritam Naskar

Roll No: STUG/042/21

Registration No: A03-1122-0199-21

Ramakrishna Mission Residential College (Autonomous)

Narendrapur, Kolkata - 700103

Abstract:

In this study, we delve into the critical realm of sleep disorder prediction, harnessing a dataset from Kaggle comprising 374 observations spanning 13 variables.

Our endeavour aims to unravel the intricate interplay between lifestyle factors and sleep disorders. Central to our pursuit is the pressing need to underscore the importance of early detection, given the often overlooked yet profound ramifications of sleep disorders on overall well-being and productivity.

The genesis of this project stems from an intriguing query posed by an insurance company regarding the potential influence of sleep disorders on policy premiums. Amidst the quest for suitable datasets across various platforms, the discovery of this Kaggle dataset seized my attention, prompting a deep dive into this pertinent subject matter. Thus, our study not only sheds light on the nexus between sleep disorders and insurance policies but also underscores their broader societal implications.

We begin by pre-processing the data and conducting exploratory data analysis (EDA) to elucidate the relationships between various variables and sleep disorders. Based on our findings, we opt for logistic regression as it aligns with our goal of predicting the presence of sleep disorders. Employing stepwise AIC technique for model selection and addressing multicollinearity using VIF, we derive our final model. Notably, variable selection is rigorously addressed throughout our analysis to ensure the model's robustness and interpretability using Pearson's chi-square and Goodman-Kruskal gamma.

With a threshold probability of 0.5 to classify individuals as having a sleep disorder, our confusion matrix validation on testing data demonstrates an approximate 94% accuracy. Moreover, cross-validation on a randomly selected sample of 10 individuals achieves a perfect 100% accuracy. Following this, we've developed an HTML form for straightforward input of individual data, providing both probabilities of sleep disorder affliction and a diagnosis of sleep disorder.

Our conclusion emphasizes the criticality of early detection and highlights the practical value of our predictive model in enabling timely intervention. This model not only assists the medical industry in delivering personalized treatment but also offers various other benefits.

Table of Contents:

Introduction:

Sleep disorders are on the rise in our modern, tech-driven society, posing challenges to health and well-being. As technology and demanding lifestyles disrupt sleep patterns, there's a growing interest in predictive models to tackle these issues and promote better sleep habits.

- **Ancient Wisdom and Modern Challenges:**

  Ancient cultures, like in the Ayurvedic "Charka Samhita," have long emphasized the importance of sleep for health and longevity. Despite this wisdom, today's focus on productivity often leads to sleep deprivation due to electronic devices, irregular work, and stress, increasing sleep disorders.

- **Harnessing Modern Techniques for Sleep Health:**

  Modern techniques, like data mining and predictive modelling, offer ways to analyse sleep patterns and identify risk factors. These technologies enable early detection of sleep disorders, leading to smart solutions like wearable devices for real-time sleep monitoring.

- **Sanskrit Sloka:**

  Sleep has been considered vital for overall health since ancient times. Here are two examples:

1. "रात्रौ द्विजातिष्ठन्ति वै रोगाः"
   Translation: "Diseases arise from disregarding the night."

   This Sanskrit sloka from ancient texts underscores the timeless importance of respecting and maintaining regular sleep patterns for overall health.

2. "यत्र नास्ति निद्रा तत्र नास्ति श्रीः"
   Translation: "Where there is no sleep, there is no prosperity."

   This sloka highlights the essential connection between sleep and well-being, emphasizing that without proper sleep, one cannot achieve or maintain prosperity or well-being.

Purpose of Study: Enhancing Sleep Health in Today's World

In our bustling modern world, sleep disorders are often overlooked but require care just like any other health issue. This study passionately advocates for the importance of addressing sleep disorders with the same urgency and compassion as other diseases. By developing predictive models, we aim to identify sleep disorder risks early, emphasizing the need for timely and personalized care. Our heartfelt goal is to reduce the societal burden of sleep-related issues, nurturing overall well-being and emotional health. Furthermore, we strive to inform and inspire workplace wellness programs to boost productivity, safety, and employee well-being. Leveraging cutting-edge technology like wearable devices, we aspire to monitor sleep in real-time, contributing to advancements in compassionate and personalized healthcare.

Motivation:

Sleep disorders significantly impact quality of life, yet they often go undiagnosed, straining healthcare systems globally.

The inspiration for my project came from an insurance company's curiosity about whether sleep disorders should influence policy premiums. While searching for datasets on various platforms, I discovered this dataset for my project on Kaggle, which piqued my interest. This led me to choose this topic for my project, exploring the impact of sleep disorders on insurance policies and premiums.

Motivated by this, the project aims to revolutionize sleep disorder diagnosis and management.

➢ References:

- CDC reports 50-70 million affected by sleep disorders in the U.S., including conditions like insomnia and sleep apnea.
- WHO links sleep disorders to health issues like cardiovascular diseases and mental health disorders.
- Global studies show rising sleep disturbances due to stress, lifestyle changes, and screen time.
- Many remain undiagnosed or face treatment delays, increasing healthcare costs.

➢ Key Objectives:

- Develop a predictive model for early sleep disorder detection using wearable or sensor data.
- Tailor personalized treatment plans based on individual sleep patterns.
- Improve healthcare outcomes and reduce economic burden by enabling timely interventions.

The project aims to advance sleep medicine, enhancing diagnosis and treatment and ultimately improving global health and well-being through better sleep.

Definitions:

- **Sleep Duration:**

Sleep duration refers to the length of time an individual spends asleep during a specified period, typically measured in hours. It is a crucial aspect of sleep health and is influenced by various factors such as age, lifestyle, and individual differences. Adequate sleep duration is essential for overall well-being, as it allows the body and mind to rest, recover, and perform essential functions.

According to **American Academy of Sleep Medicine (AASM)**;

Age Group	Sleep Duration
Infants and New-borns (0-12 months)	
- New-borns (0-3 months)	14-17 hours/day
- Infants (4-11 months)	12-15 hours/day
Toddlers and Pre-schoolers (1-5 years)	
- Toddlers (1-2 years)	11-14 hours/day
- Pre-schoolers (3-5 years)	10-13 hours/day
School-age Children and Adolescents (6-18 years)	
- School-age children (6-12 years)	9-12 hours/day
- Adolescents (13-18 years)	8-10 hours/day
Adults (18-64 years)	7 or more hours/day
Older Adults (65+ years)	
- 61-64 years	7-9 hours/day
- 65+ years	7-8 hours/day

These guidelines serve as references for understanding the recommended sleep durations across different age groups.

- **Quality of Sleep:**

Quality of sleep refers to the subjective and objective characteristics of sleep that determine its restorative and rejuvenating effects on the body and mind. It encompasses various factors, including sleep depth, continuity, efficiency, and subjective experiences during sleep.

According to **American Academy of Sleep Medicine (AASM);**

Sleep Metric	Description
Sleep Latency	Time taken to fall asleep after initiating sleep onset, ideally within 15-20 minutes.
Sleep Duration	Total time spent asleep, meeting the recommended duration for age group and individual needs.
Sleep Architecture	Pattern and distribution of sleep stages throughout the night, including time in each stage (e.g., REM sleep, deep sleep).
Sleep Fragmentation	Presence of frequent awakenings or disruptions during sleep, impacting sleep continuity and quality.

By optimizing sleep quality, individuals can enhance their overall health, cognitive function, and quality of life.

- **Stress:**

According to **World Health Organization (WHO)**;

Stress can be defined as a state of worry or mental tension caused by a difficult situation. Stress is a natural human response that prompts us to address challenges and threats in our lives. Everyone experiences stress to some degree.

- **Body Mass Index (BMI):**

According to the **World Health Organization (WHO)**, Body Mass Index (BMI) is a weight-to-height ratio used to classify adults as underweight, overweight, or obese. BMI is calculated by dividing a person's weight in kilograms by the square of their height in meters.

Obesity:

Obesity is a chronic complex disease defined by excessive fat deposits that can impair health. Obesity can lead to increased risk of type 2 diabetes and heart disease, it can affect bone health and reproduction, it increases the risk of certain cancers. Obesity influences the quality of living, such as sleeping or moving.

A BMI of 18.5 to 24.9 is considered healthy, 25 to 29.9 is overweight, 30 or higher is obese, and 40 or higher is severely obese.

- **Sleep Disorder:**

A sleep disorder is a medical condition that disrupts a person's normal pattern of sleep. These disorders can affect the quality, timing, and duration of sleep, leading to difficulties in falling asleep, staying asleep, or waking up too early. Sleep disorders can have a significant impact on a person's overall health, well-being, and daily functioning.

Common types of sleep disorders include:

Sleep Disorder	Description
Insomnia	Difficulty falling or staying asleep, can be acute or chronic, often related to factors like stress, anxiety, medical conditions, or lifestyle habits.
Sleep Apnea	Disorder with interruptions in breathing during sleep, leading to fragmented sleep, snoring, and fatigue. Two main types: obstructive (airway blockage) and central (brain signalling issue).

About Dataset:

➢ Introduction to the Sleep Health and Lifestyle Dataset:

The Sleep Health and Lifestyle Dataset comprises 374 rows and 13 columns, capturing diverse variables from sleep patterns to daily routines. It includes demographic details, occupational information, sleep metrics, physical activity, stress levels, BMI, blood pressure, heart rate, and sleep disorder prevalence. This dataset is a valuable resource for exploring correlations and insights into sleep health and lifestyle behaviours, offering a rich foundation for in-depth research on factors influencing sleep quality and well-being.

➢ Overview of Dataset Columns:

Variable	Description
Person ID	Unique identifier for each individual in the dataset.
Gender	Categorizes individuals as Male or Female.
Age	Age of each individual in years.
Occupation	Profession or occupation pursued by each person.
Sleep Duration (hours)	Number of hours each person sleeps per day.
Quality of Sleep (scale: 1-10)	Subjective rating of sleep quality on a scale from 1 to 10.
Physical Activity Level (minutes/day)	Duration of daily physical activity engagement in minutes.
Stress Level (scale: 1-10)	Perceived stress level on a scale from 1 to 10.
BMI Category	Classification into BMI categories: Obese, Normal, or Overweight.
Blood Pressure (systolic/diastolic)	Blood pressure measurement indicated as systolic over diastolic.
Heart Rate (bpm)	Resting heart rate in beats per minute.
Daily Steps	Number of steps taken per day.
Sleep Disorder	Presence or absence of a sleep disorder categorized as None, Insomnia, or Sleep Apnea.

Summary:

- **Numeric Variables Overview:**

Variable	Range	Average	Notes
Age	27 to 59 years	~42 years	
Daily Steps	3,000 to 10,000 steps/day	~6,817 steps	
Heart Rate (bpm)	65 to 86 bpm	~70 bpm	15 outliers in data
Physical Activity Level (minutes/day)	30 to 90 minutes/day	~59 minutes	
Sleep Duration (hours)	5.8 to 8.5 hours	~7.13 hours	

- **Categorical Variables Insights related to Sleep Disorder:**

Variable	Details
Gender	2 categories: Male & Female
Occupation	10 different occupations: Doctor, Engineer, Accountant, Lawyer, Nurse, Salesperson, Sales Representative, Scientist, Manager, Teacher, Software Engineer
Blood Pressure	4 different categories: Normal, Elevated, Hypertension Stage 1, Hypertension Stage 2

Data Processing:

Manipulation in the column named "Occupation":

1. We have a total of 11 professions, with the database predominantly consisting of doctor, nurse, accountant, engineer, lawyer, salesperson and teacher. For a valid analysis involving this variable, at first, we merged sales representative (2) with salesperson (32), as they likely represent similar roles within the organization.

2. It is recommended to exclude manager (1), scientist (4), and software engineer (4) from the dataset because we have not sufficient data for those professionals.

3. So, I've opted to sacrifice a small portion of my actual database (2.41%) for the sake of project management convenience and to ensure the validity of my project.

Manipulation in the column named "BMI Category":

1. We have a total of four categories in the "BMI Category" column. Here, we observe two categories labelled as "Normal" (195) and "Normal Weight" (21) respectively.

2. We assume that these categories are identical and require modification. To address this, we merge the two categories into a single "Normal" category.

Manipulation in the column named "Blood Pressure":

1. We initially encounter data on blood pressure in the systolic/diastolic format, which impedes further analysis. Our first step involves splitting this data into two separate columns labelled "Systolic" and "Diastolic."

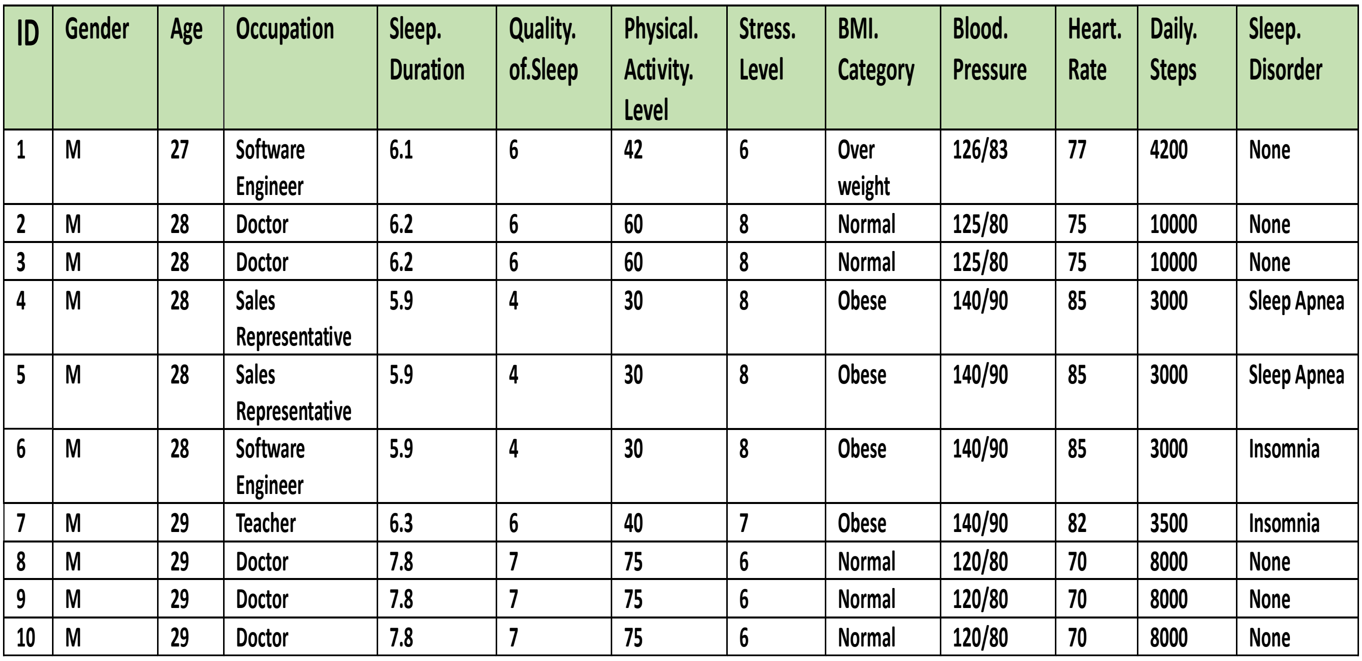
2. This variable presents certain challenges, particularly due to categories with low counts. However, it's worth noting that the American College of Cardiology, in collaboration with the American Heart Association, has provided guidelines for classification. Additional information can be found at CDC.gov.

3. Based on the aforementioned classification guidelines, I establish four categories and assign them to the column named "BP Category":

Blood Pressure Category	Systolic Blood Pressure	Diastolic Blood Pressure
Normal	<120 mmHg	<80 mmHg
Elevated	120-129 mmHg	<80 mmHg
Hypertension Stage 1	130-139 mmHg	80-89 mmHg
Hypertension Stage 2	≥140 mmHg	≥90 mmHg

Data Snapshot: First 10 Rows:

Here's a glimpse into the first 10 rows of our dataset, offering an overview of the variables and values present. This snapshot provides a foundational understanding of the dataset's structure and content, setting the stage for our subsequent analysis.



ID	Gender	Age	Occupation	Sleep. Duration	Quality. of.Sleep	Physical. Activity. Level	Stress. Level	BMI. Category	Blood. Pressure	Heart. Rate	Daily. Steps	Sleep. Disorder
1	M	27	Software Engineer	6.1	6	42	6	Over weight	126/83	77	4200	None
2	M	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
3	M	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
4	M	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
5	M	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
6	M	28	Software Engineer	5.9	4	30	8	Obese	140/90	85	3000	Insomnia
7	M	29	Teacher	6.3	6	40	7	Obese	140/90	82	3500	Insomnia
8	M	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
9	M	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None
10	M	29	Doctor	7.8	7	75	6	Normal	120/80	70	8000	None

Exploratory Analysis of Sleep Health and Lifestyle Variables:

In this basic search of the Sleep Health and Lifestyle Dataset, we aim to uncover insights into the relationships and patterns among various variables related to sleep health and lifestyle behaviours. Our focus will be on posing specific questions to guide our exploration, such as:

1. *How are the variables in my dataset correlated with each other?*
2. *Which professions are more prone to sleep disorders?*
3. *Is there a gender difference in susceptibility to sleep disorders?*
4. *After which age distribution do sleep disorders start to prevail notably?*
5. *What's the overall distribution of sleep disorders in our dataset?*
6. *How do sleep quality and duration impact the risk of sleep disorders?*
7. *Does physical activity and daily steps play a part in sleep disorders?*
8. *How does stress level influence the likelihood of sleep disorders?*
9. *Is there a trend between heart rate and sleep disorders?*
10. *Are there differences in sleep disorders across various BMI categories?*
11. *Between males and females, who typically exhibit better quality of sleep?*
12. *Which gender tends to have lower stress levels, males or females?*
13. *Among males and females, who generally shows lower heart rates?*
14. *How does physical activity level influence sleep duration?*
15. *How does stress level relate to sleep duration?*
16. *Is the level of physical activity related to the number of daily steps taken?*
17. *Is there a relationship between heart rate and BMI category?*

Through this exploratory analysis, we seek to gain a deeper understanding of the factors influencing sleep health and their implications for overall well-being.

Data Visualisation:

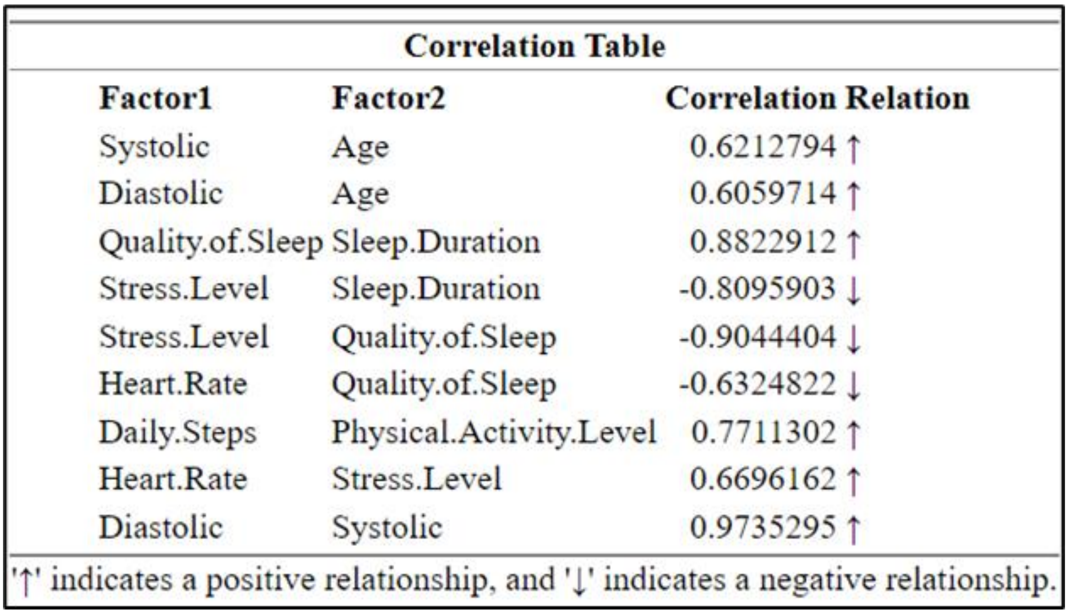
Data visualization simplifies complex datasets, highlighting patterns and relationships for clearer insights. In sleep disorders research, it reveals the factors affecting sleep disturbances effectively. Visualizations enhance understanding and aid in communicating findings to diverse audiences.

I will now proceed to present key visualizations based on our queries derived from the dataset.

1. How are the variables in my dataset correlated with each other?

Checking the correlation matrix helps identify relationships between variables and understand how they influence each other, offering insights into underlying patterns and dependencies in the dataset.

I've used the *corrplot* function from the *corrplot* library in *R* to visualize the correlations among variables. After generating the correlation matrix, I've used the *kable* function to present the insights efficiently. I've focused on correlations with an absolute value greater than 0.5. Key observations include:



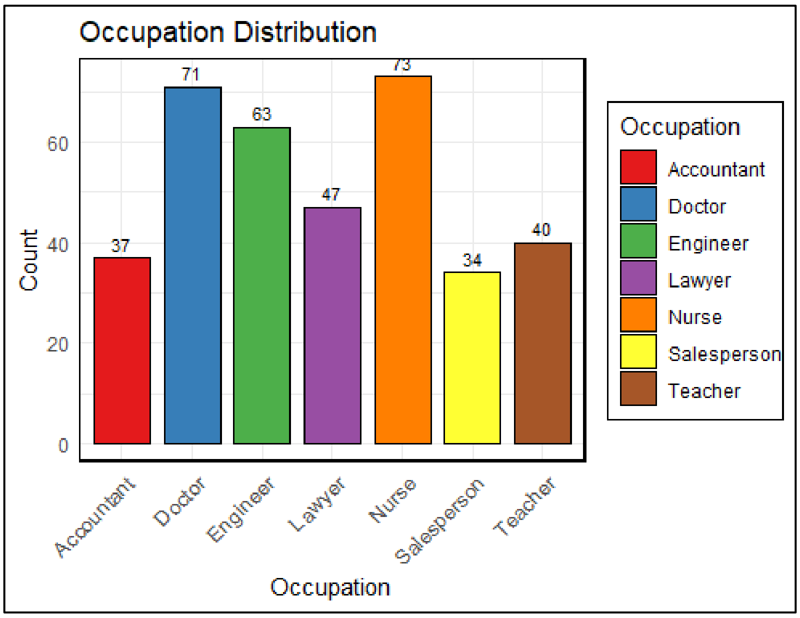
Correlation Table

Factor1	Factor2	Correlation Relation
Systolic	Age	0.6212794 ↑
Diastolic	Age	0.6059714 ↑
Quality.of.Sleep	Sleep.Duration	0.8822912 ↑
Stress.Level	Sleep.Duration	-0.8095903 ↓
Stress.Level	Quality.of.Sleep	-0.9044404 ↓
Heart.Rate	Quality.of.Sleep	-0.6324822 ↓
Daily.Steps	Physical.Activity.Level	0.7711302 ↑
Heart.Rate	Stress.Level	0.6696162 ↑
Diastolic	Systolic	0.9735295 ↑

'↑' indicates a positive relationship, and '↓' indicates a negative relationship.

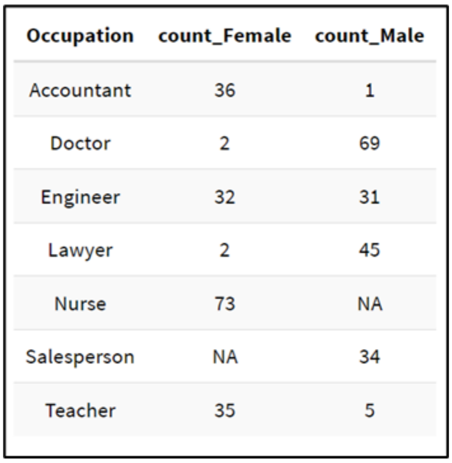
2. Which professions are more prone to sleep disorders?

To answer these questions, we'll explore the relationship between occupation and sleep disorders. First, let's look at the distribution of various occupations in the dataset with a visualization.

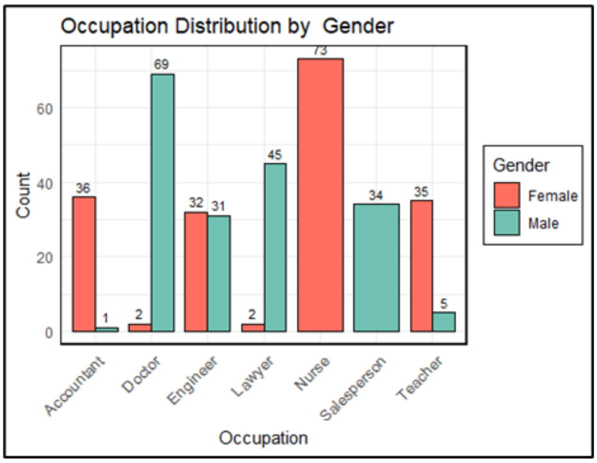


The graph illustrates the distribution of various occupations within the dataset, displaying the count for each occupation.

Now, I aim to present the distribution of occupations categorized by gender, as it provides the occupational composition across genders within the dataset, facilitating a comprehensive understanding of demographic variations in the context of sleep disorders research.



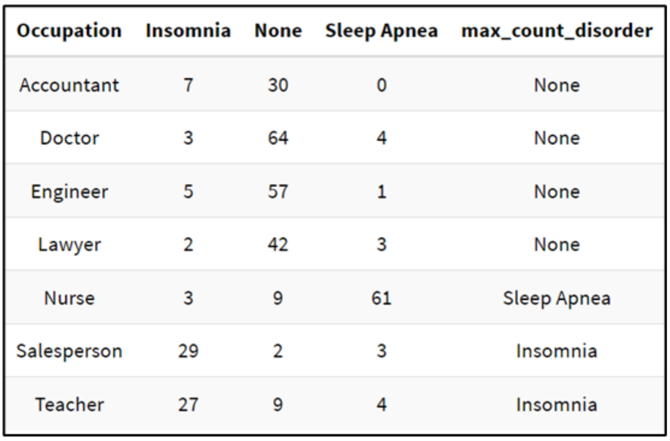
Occupation	count_Female	count_Male
Accountant	36	1
Doctor	2	69
Engineer	32	31
Lawyer	2	45
Nurse	73	NA
Salesperson	NA	34
Teacher	35	5



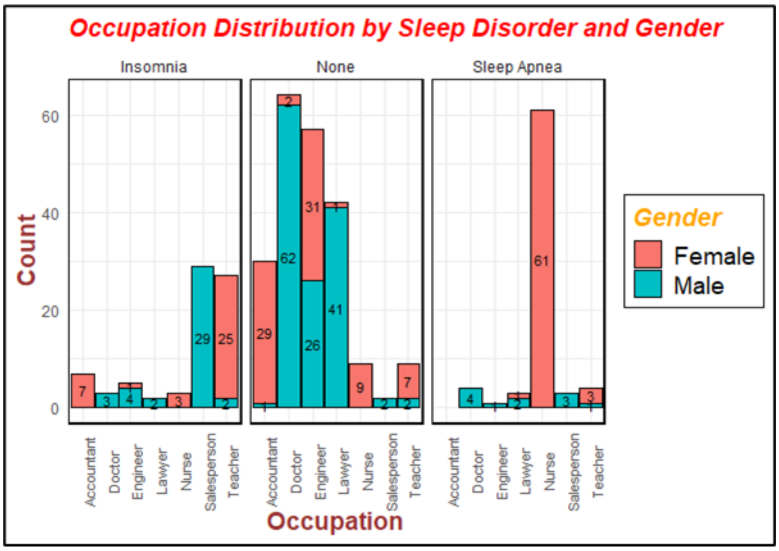
Upon closer examination of the data, several observations emerge:

- *Engineers exhibit a relatively balanced gender ratio.*
- *Nurses are predominantly female, aligning with common expectations.*
- *Salespersons predominantly comprise males.*
- *There is a greater representation of females among teachers, which may not fully reflect real-life demographics.*
- *Doctors and lawyers predominantly consist of males, while accountants are predominantly female.*

The graph illustrates the distribution of various occupations categorized by gender and sleep disorder within the dataset, displaying the count for each occupation.



Occupation	Insomnia	None	Sleep Apnea	max_count_disorder
Accountant	7	30	0	None
Doctor	3	64	4	None
Engineer	5	57	1	None
Lawyer	2	42	3	None
Nurse	3	9	61	Sleep Apnea
Salesperson	29	2	3	Insomnia
Teacher	27	9	4	Insomnia



After closely scrutinizing the dataset, the following insights emerge:

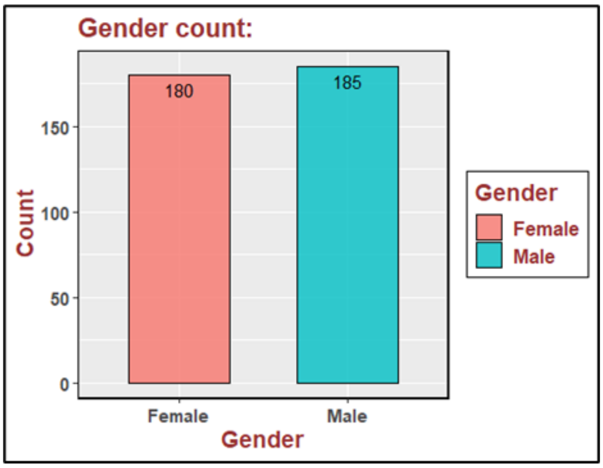
- *Insomnia is primarily observed among female teachers and male salespersons.*
- *No sleep disorders are reported among accountants, doctors, engineers, and lawyers.*
- *Sleep apnea is predominantly prevalent among nurses.*

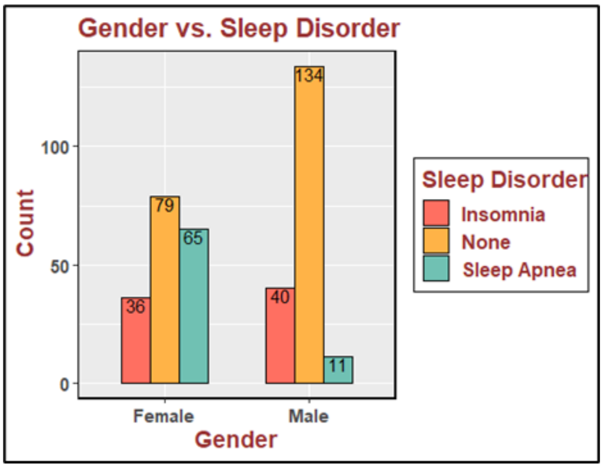
3. Is there a gender difference in susceptibility to sleep disorders?

We'll explore the link between gender and sleep disorders by visualizing gender distribution in the dataset to understand prevalence. Here's the visualization of our findings.

The graph shows nearly equal representation of males and females in the dataset, ensuring our gender-based analysis is balanced and representative.

Now, I will examine the occurrence of sleep disorders with respect to gender and provide the corresponding graph.

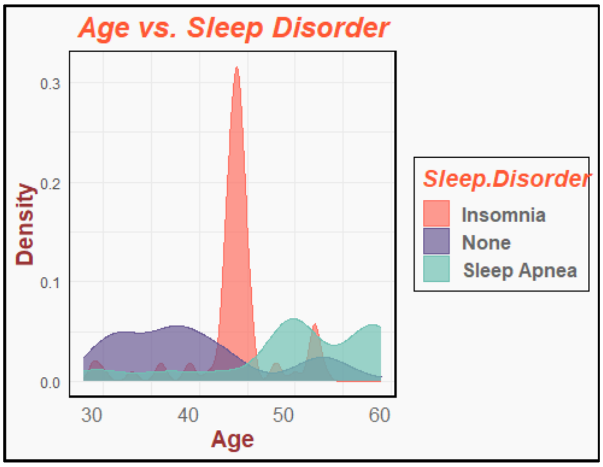




The graph indicates more males in the normal category and more females with sleep disorders, notably nurses with sleep apnea and female teachers with insomnia. This aligns with our previous findings, strengthening our analysis.

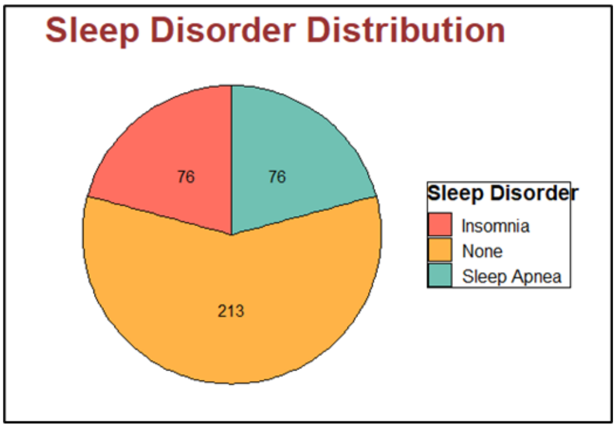
4. After which age distribution do sleep disorders start to prevail notably?

Age plays a crucial role in sleep disorders, with prevalence increasing with advancing age. The graph highlights a higher proportion of normal individuals under 45, while sleep apnea patients are predominantly aged 45 and older. Notably, insomnia patients peak around age 44. Overall, the graph reflects the expected positive correlation between age and sleep disorder prevalence, mirroring real-life trends.



5. What's the overall distribution of sleep disorders in our dataset?

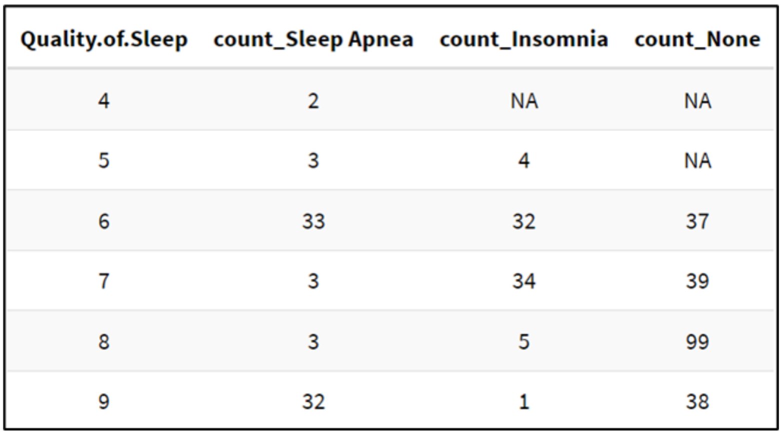
I'll create a pie chart to show the sleep disorder distribution.



Most data are in the normal category, with sleep apnea and insomnia categories equally represented.

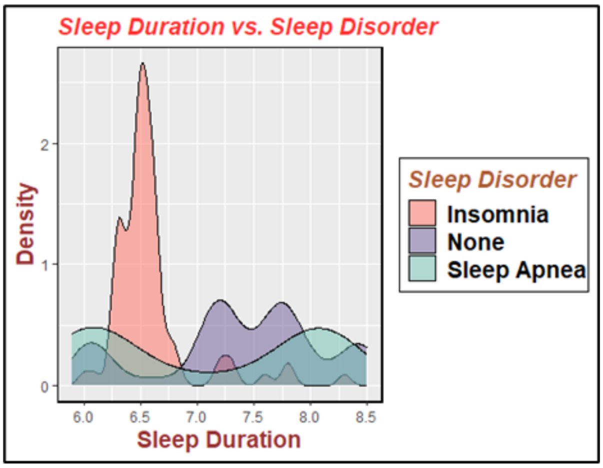
6. How do sleep quality and duration impact the risk of sleep disorders?

I'm exploring the link between sleep quality, rated on a 10-point scale, and sleep disorders. Here's the data to show this relationship.



Quality.of.Sleep	count_Sleep Apnea	count_Insomnia	count_None
4	2	NA	NA
5	3	4	NA
6	33	32	37
7	3	34	39
8	3	5	99
9	32	1	38

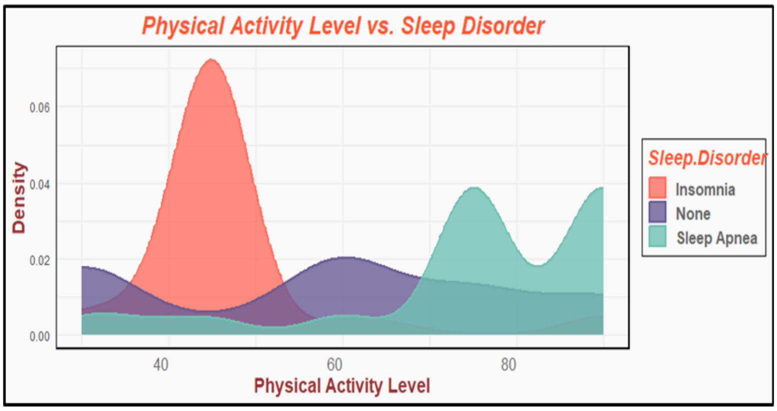
To explore the link between sleep duration and sleep disorders, we'll focus on the recommended 7 to 8 hours per night. The graph is expected to show: meeting this guideline reduces sleep disorder risk, while deviations may increase it. This highlights the importance of adequate sleep for healthy patterns and lower risk of sleep disorders.

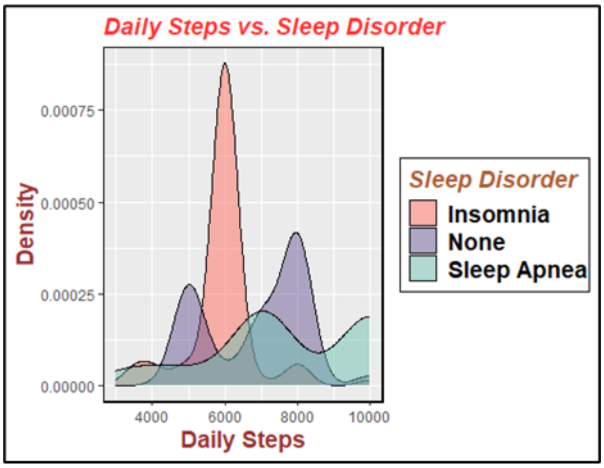


The graph confirms our earlier statement: those getting 7 to 8 hours of sleep have fewer sleep disorders, validating our analysis and its real-world relevance.

7. Does physical activity and daily steps play a part in sleep disorders?

This density plot shows different physical activity levels among sleep disorder patients. Insomnia patients have shorter workouts than normal individuals, while sleep apnea patients exceed the average duration.





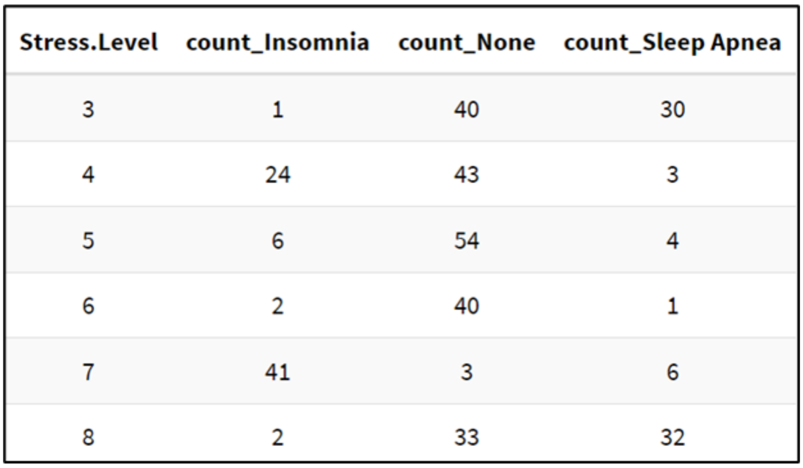
Using a density plot for daily workout minutes, we find:

- *Average activity is 59.50 minutes/day.*
- *Insomnia patients average around 45 minutes, below the mean.*

- *Normal individuals are close to the mean at 60 minutes.*
- *Sleep apnea patients often exceed the mean with longer workouts.*

Examining the correlation between daily steps and sleep disorders using a density plot, we find: as daily steps increase, sleep disorder likelihood decreases, supporting the link between physical activity and better sleep. Peaks at 6000 steps for insomnia patients and around 5000-8000 for others validate this. However, daily steps alone may not fully predict sleep disorders.

8. How does stress level influence the likelihood of sleep disorders?



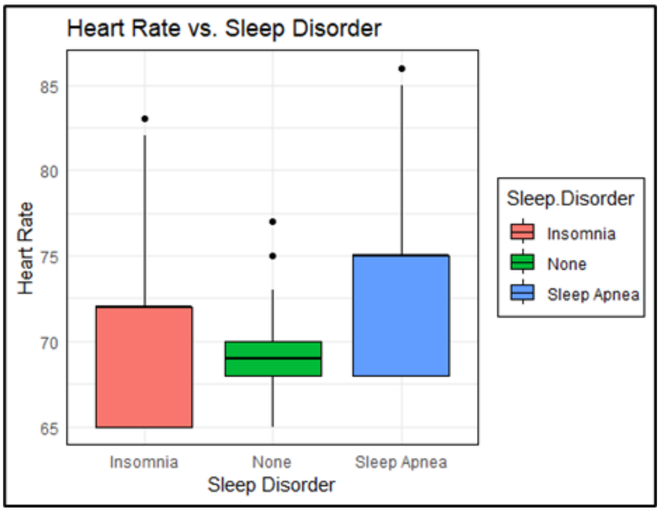
Stress.Level	count_Insomnia	count_None	count_Sleep Apnea
3	1	40	30
4	24	43	3
5	6	54	4
6	2	40	1
7	41	3	6
8	2	33	32

We're exploring the link between stress levels and sleep disorders. Higher stress often leads to poorer sleep quality and more sleep disorders. The data will show this correlation.

It is clear from the data that for people without sleep issues, stress levels usually range from 4 to 6 with a median of 5. Insomnia raises the median stress level to 7, and for sleep apnea, it's 6.5.

9. Is there a trend between heart rate and sleep disorders?

I want to compare heart rates between individuals without sleep disorders and those with sleep issues. The graph shows:

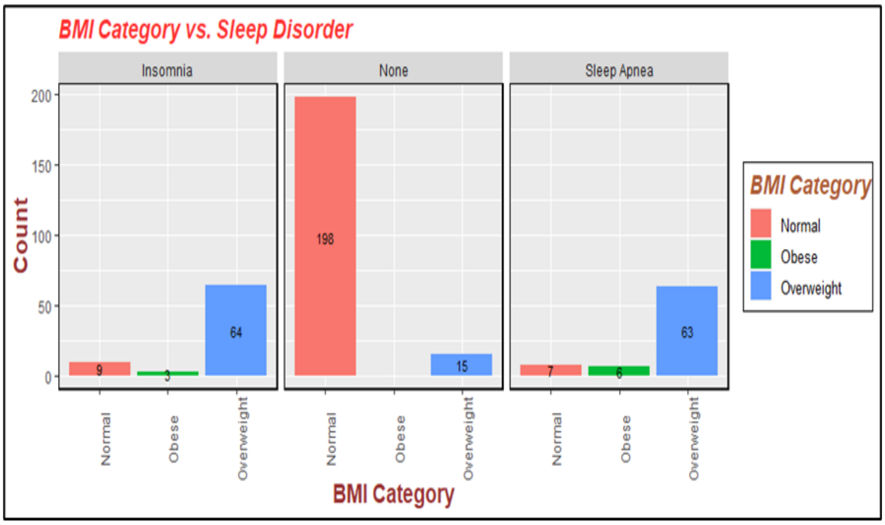


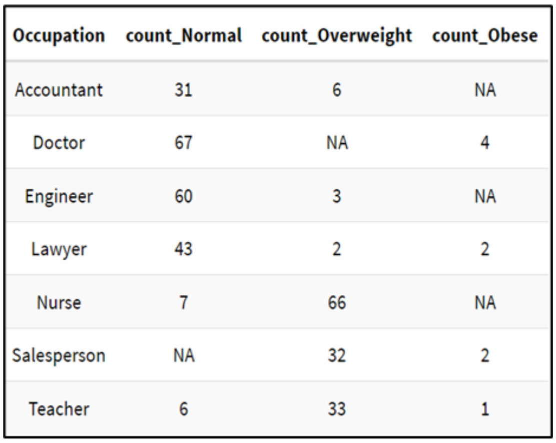
Sleep Category	Median Heart Rate (bpm)
Normal	68
Insomnia	72
Sleep Apnea	75

Individuals within the recommended heart rate range are less likely to have sleep disorders.

10. Are there differences in sleep disorders across various BMI categories?

The bar graph shows a link between BMI categories and sleep disorders. Overweight people are more prone to sleep disorders, while normal-weight people are less affected. There's ample data for normal and overweight groups, but less for obese individuals, reflecting real-world BMI prevalence.

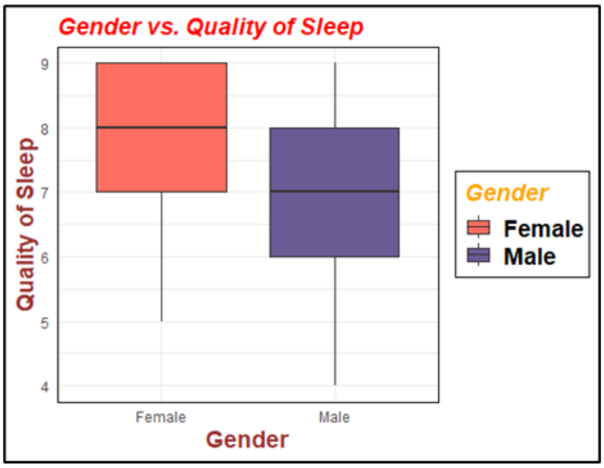




Occupation	count_Normal	count_Overweight	count_Obese
Accountant	31	6	NA
Doctor	67	NA	4
Engineer	60	3	NA
Lawyer	43	2	2
Nurse	7	66	NA
Salesperson	NA	32	2
Teacher	6	33	1

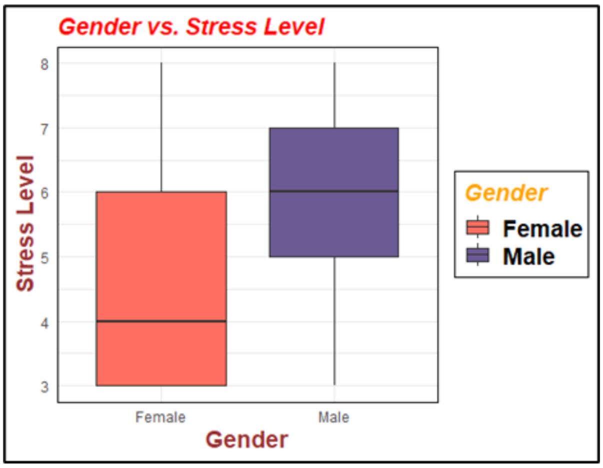
11. Between males and females, who typically exhibit better quality of sleep?

Using a box plot to compare sleep quality across genders, females more often rated their sleep at 8, while males rated theirs at 7, suggesting females generally have higher sleep quality than males.

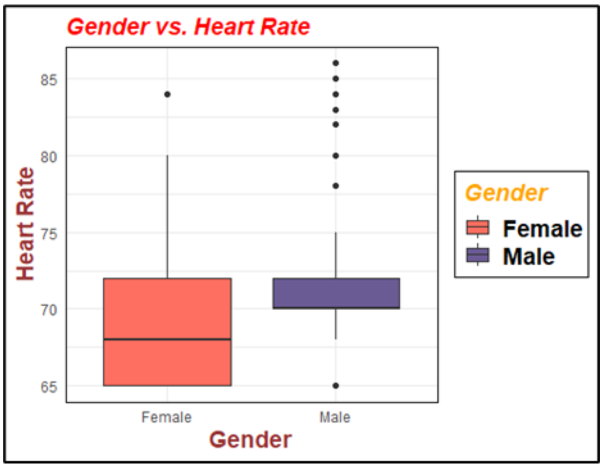


12. Which gender tends to have lower stress levels, males or females?

Switching to gender and stress levels, a box plot shows that higher stress often leads to poorer sleep. Consistent with females reporting better sleep quality, they also tend to have lower stress levels compared to males.

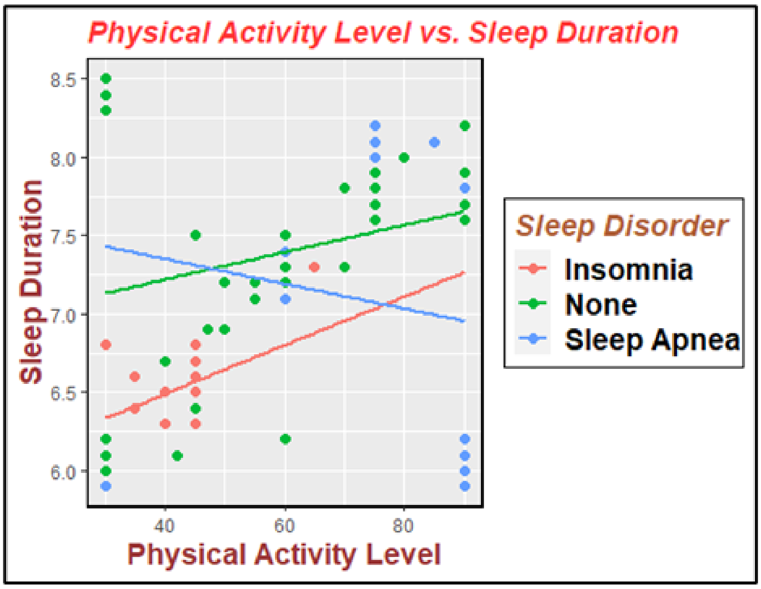


13. Among males and females, who generally shows lower heart rates?



Using a box plot to look at gender and heart rate, outliers below 85 are present but within the normal range of 60-100 bpm, so they're ignored. A positive correlation between heart rate and stress is observed. Females generally have a heart rate around 68 bpm, slightly lower than males at 70 bpm.

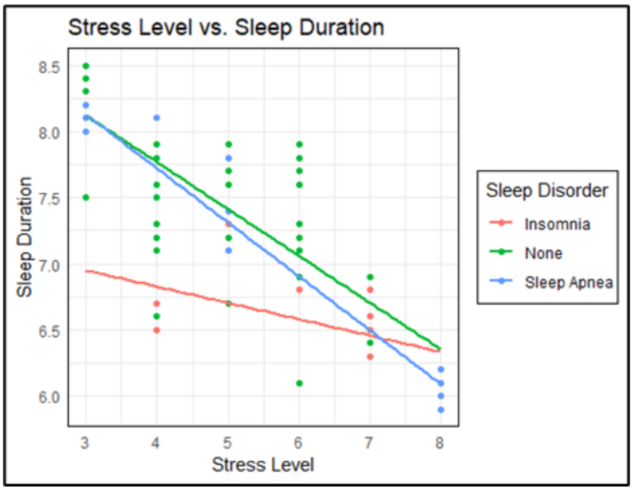
14. How does physical activity level influence sleep duration?



We're examining how physical activity level correlates with sleep duration to predict sleep disorders. Using a scatter plot color-coded by sleep disorder type, we see that insomnia patients and normal individuals generally sleep more with higher physical activity.

In contrast, sleep apnea patients tend to sleep less with increased activity. This aligns with real-world observations where sleep duration differs notably between those with insomnia, sleep apnea, and those without sleep disorders.

15. How does stress level relate to sleep duration?

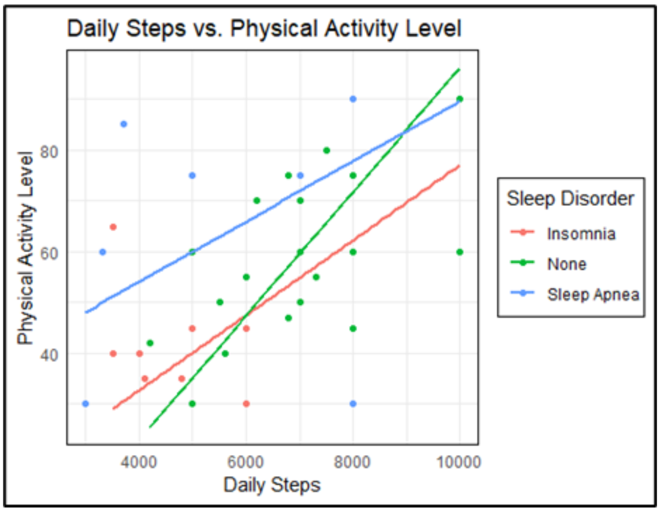


There's a strong negative correlation of 0.81 between stress level and sleep duration. A scatter plot with sleep disorder categories and a regression line show that as stress increases, sleep duration decreases, supporting the correlation.

The regression lines for normal and sleep apnea patients are steeper than for insomnia patients, indicating stress has a stronger impact on sleep duration in these groups.

16. Is the level of physical activity related to the number of daily steps taken?

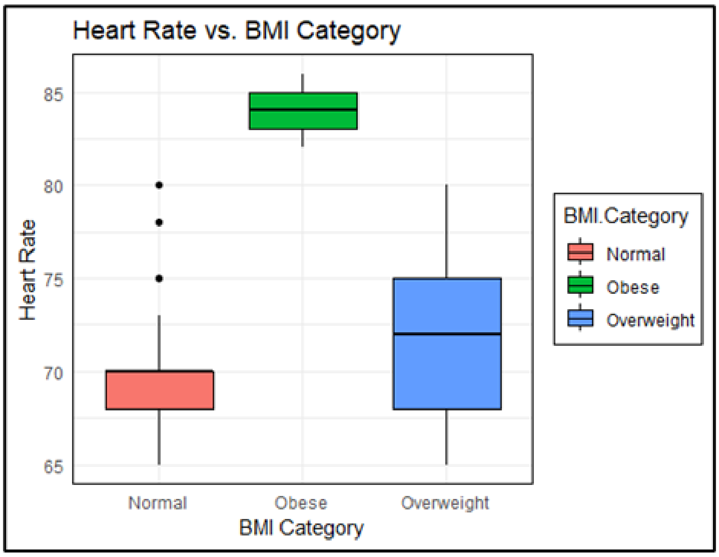
The level of physical activity and daily steps reflect an individual's overall activity and fitness. Using a scatter plot helps visualize this correlation, highlighting trends. A positive correlation ideally indicates that as physical activity increases, daily steps do too.



The graph supports these assumptions. For those without sleep disorders, the steeper slope suggests a stronger link between activity level and steps. In contrast, sleep disorder patients show a less pronounced slope. Among them, sleep apnea patients have a higher line than insomnia patients.

17. Is there a relationship between heart rate and BMI category?

Heart rate varies across BMI categories, showing a relationship between body weight and cardiovascular health. Individuals with a normal BMI typically have heart rates within the recommended range. In contrast, overweight individuals tend to have slightly higher heart rates, and those classified as obese exhibit the highest rates. The graph confirms these trends.



Exploring Occupational Risks for Sleep Disorders:

I've narrowed my focus to the occupation column, as I'm intrigued by understanding the prevalence of sleep disorders across different professions. Initially, I found that doctors, accountants, engineers, and lawyers are classified as 'normal,' while nurses, teachers, and salespersons are categorized under 'sleep disorder.' However, I suspect that solely examining counts may not provide sufficient insights. Therefore, I plan to analyse the relationship between occupation and sleep disorders in conjunction with other variables.

- *Upon incorporating age as a third variable, I've observed that professions like accountants, doctors, engineers, and lawyers are predominantly found in younger age groups, with fewer instances of sleep disorders. Conversely, nurse teachers and salespersons tend to show a higher prevalence, particularly among older age brackets.*

- *Occupations like doctors, accountants, lawyers, and engineers consistently demonstrate healthy sleep patterns within the recommended range. Nurses, associated with sleep apnea, exhibit variability in sleep duration, possibly due to symptoms of the disorder. Salespersons and teachers, experiencing insomnia, tend to have shorter sleep durations on average, reflecting challenges associated with the condition.*

- *In case of physical activity, lawyers and accountants cluster around the mean, with lawyers displaying greater variation. In contrast, doctors lead in activity, while engineers lag behind. Nurses, sharing activity levels with doctors, may face increased susceptibility to sleep apnea due to their demanding roles. Conversely, salespersons and teachers, with lower activity levels, could be at higher risk of insomnia.*

- *Accountants, engineers, and lawyers generally exhibit lower stress levels, aligning with the normal sleep category. Doctors, while still within the normal range, show slightly elevated stress levels due to their demanding responsibilities. Nurses display a wide range of stress levels reflective of their demanding workload. Salespersons show notably higher stress levels, while teachers exhibit lower stress levels.*

These validations solidify the link between occupation and sleep disorders with other predictor variables, reinforcing the credibility of our findings and offering essential insights for further research and intervention strategies.

Overall Conclusion:

- *Occupations such as doctors, accountants, engineers, and lawyers are primarily categorized as 'normal,' while nurses, teachers, and salespersons are more prevalent in the 'sleep disorder' category.*

- *Gender distribution shows more males in the 'normal' category and more females in the 'sleep disorder' category.*

- *Females tend to report higher quality of sleep compared to males, who often experience sleep deprivation.*

- *Females exhibit lower stress levels and heart rates compared to males.*

- *Individuals in the 'normal' category generally show no signs of sleep disorders based on BMI categories.*

- *Lower physical activity levels are associated with a higher likelihood of sleep disorders.*

- *Individuals aged 45 and above are more prone to sleep disorders, suggesting an age-related factor influencing sleep health.*

Our findings validate the overall pattern observed and underscore the project's validity. These insights hold relevance across various fields. The implications extend to the medical field, offering insights into broader discussions on sleep health.

Generalized Linear Model (GLM):

A Generalized Linear Model (GLM) is a statistical framework that extends the concepts of linear regression to a broader range of data distributions. It accommodates response variables that follow distributions beyond the normal distribution, such as binomial, Poisson, or gamma distributions.

GLMs consist of three essential components: the random component, the systematic component, and the link function.

- **Random Component:**

  In a GLM, the response variables $\boldsymbol{Y_i}$'s are considered the random component. These variables are assumed to be independent random variables, each following a distribution from the same family.

  For example, in logistic regression, the response variable follows a binomial distribution, representing the probability of an event occurring.

- **Systematic Component:**

  The systematic component is a function of the predictor variables $\boldsymbol{x_i}\ '\text{s}\ where\ \boldsymbol{i = 1(1)p}$, linearly related to the parameters that affect the mean of the response variables $\boldsymbol{Y_i}'s$. This component typically takes the form of $\boldsymbol{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}$, where **β**'s are the coefficients associated with each predictor variable.

- **Link Function:**

  Finally, the link function $\boldsymbol{g(\mu)}$ links the random and systematic components by asserting that $\boldsymbol{g(\mu_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}$ where $\boldsymbol{\mu_i} = \boldsymbol{E(Y_i)}$. For our case, link function is

  $$\boldsymbol{g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)}, \text{ where } \boldsymbol{\mu_i = P(Y_i = 1)}.$$

  This type of link function is known as canonical link. For our case, we use logit link for logistic regression.

Relationship to Logistic Regression:

Logistic regression is a statistical method used for modelling the relationship between a categorical dependent variable and one or more independent variables. It's particularly useful when the dependent variable is binary, having only two possible outcomes, such as "yes" or "no", "success" or "failure", or "1" or "0".

Thus, logistic regression can be seen as a special case of the generalized linear model framework, where the response variable follows a binomial distribution, and the link function is the logit function. The systematic component in logistic regression is a linear combination of the predictor variables, allowing for the prediction of the log-odds of the response variable being in a particular category based on the values of the predictors.

General Form of Logistic Regression:

The general form of logistic regression can be expressed using the logistic function, also known as the sigmoid function:

$$\boldsymbol{P(Y_i = 1) \ = \ \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p)}}, for\ i = 1(1)p \ldots\ldots (i)}$$

Where:

- $\mathbf{P(Y_i = 1)}$ *is the probability of the dependent variable* $\mathbf{Y_i}$ *being 1.*
- $\boldsymbol{\beta_0, \beta_1, \beta_2, \ldots, \beta_p}$ *are the coefficients of the model.*
- $\mathbf{X_1, X_2, \ldots, X_p}$ *are the independent variables.*
- **e** *is the base of natural logarithm.*

Mathematical Representation:

The logistic regression model assumes that the log-odds of the dependent variable is a linear combination of the independent variables. This can be represented mathematically as:

$$\boldsymbol{\log\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \ldots\ldots\ldots..(ii)}$$

Or equivalently:

$$\log\left(\frac{P(Y_i = 1)}{P(Y_i = 0)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \ldots\ldots\ldots..(iii)$$

Here, $P(Y_i = 1)$ and $P(Y_i = 0)$ are the probabilities of the dependent variables Y_i being 1 and 0, respectively.

Equation (ii) and (iii) can be rewritten as

$$\mu_i = \frac{e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p)}}{1 + e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p)}} \ldots.\ldots\ldots\ldots\ldots..(iv)$$

Or, more generally, we can write that

$$\mu(x) = \frac{e^{(\beta_0+\beta x)}}{1 + e^{(\beta_0+\beta x)}} \ldots\ldots\ldots.\ldots..(v)$$

From equation (v) it is clear that $0 < \mu(x) < 1$, which seems appropriate because $\mu(x)$ is a probability. But, if it is possible that $\mu(x) = 0$ **or** 1 for some x, then this model is not appropriate. If we examine $\mu(x)$ more closely, its derivative can be written as

$$\frac{d\mu(x)}{dx} = \frac{\beta e^{(\beta_0+\beta x)}}{(1 + e^{(\beta_0+\beta x)})^2}$$

$$= \beta\, \mu(x)(1 - \mu(x)) \ldots\ldots\ldots\ldots.(vi)$$

As the term $\mu(x)(1 - \mu(x))$ is always positive, the derivative of $\mu(x)$ is positive, 0 or negative according as **β** is positive, 0 or negative. If **β** is positive, $\mu(x)$ is strictly increasing function of x; if **β** is negative, $\mu(x)$ is strictly decreasing function of x; if **β** is 0, then

$$\mu(x) = \frac{e^{\beta_0}}{(1 + e^{\beta_0})} \ldots\ldots\ldots\ldots\ldots\ldots.(vii)$$

Setting $x = 0$ in (iv) yields that β_0 is the log-odds of success at $x = 0$. Evaluating (iv) at x and $(x + 1)$ yields, for any x,

$$\log\left(\frac{\mu(x+1)}{1-\mu(x+1)}\right) - \log\left(\frac{\mu(x)}{1-\mu(x)}\right) = \beta_0 + \beta(x+1) - \beta_0 - \beta x$$

$$= \beta \ldots\ldots\ldots\ldots\ldots \text{(vii)}$$

Thus, β is the change in the log-odds of success corresponding to a one-unit increase in x. Exponentiating both sides of (vii) yields

$$e^{\beta} = \left(\frac{\mu(x+1)}{1-\mu(x+1)}\right) \Big/ \left(\frac{\mu(x)}{1-\mu(x)}\right) \ldots\ldots\ldots..(viii)$$

The right-hand side is the odds ratio comparing the odds of success at (x+1) to the odds of success at x. In logistic regression model this ratio is constant as a function of x. Finally, from (viii),

$$\left(\frac{\mu(x+1)}{1-\mu(x+1)}\right) = e^{\beta}\left(\frac{\mu(x)}{1-\mu(x)}\right) \ldots\ldots\ldots\ldots...(ix)$$

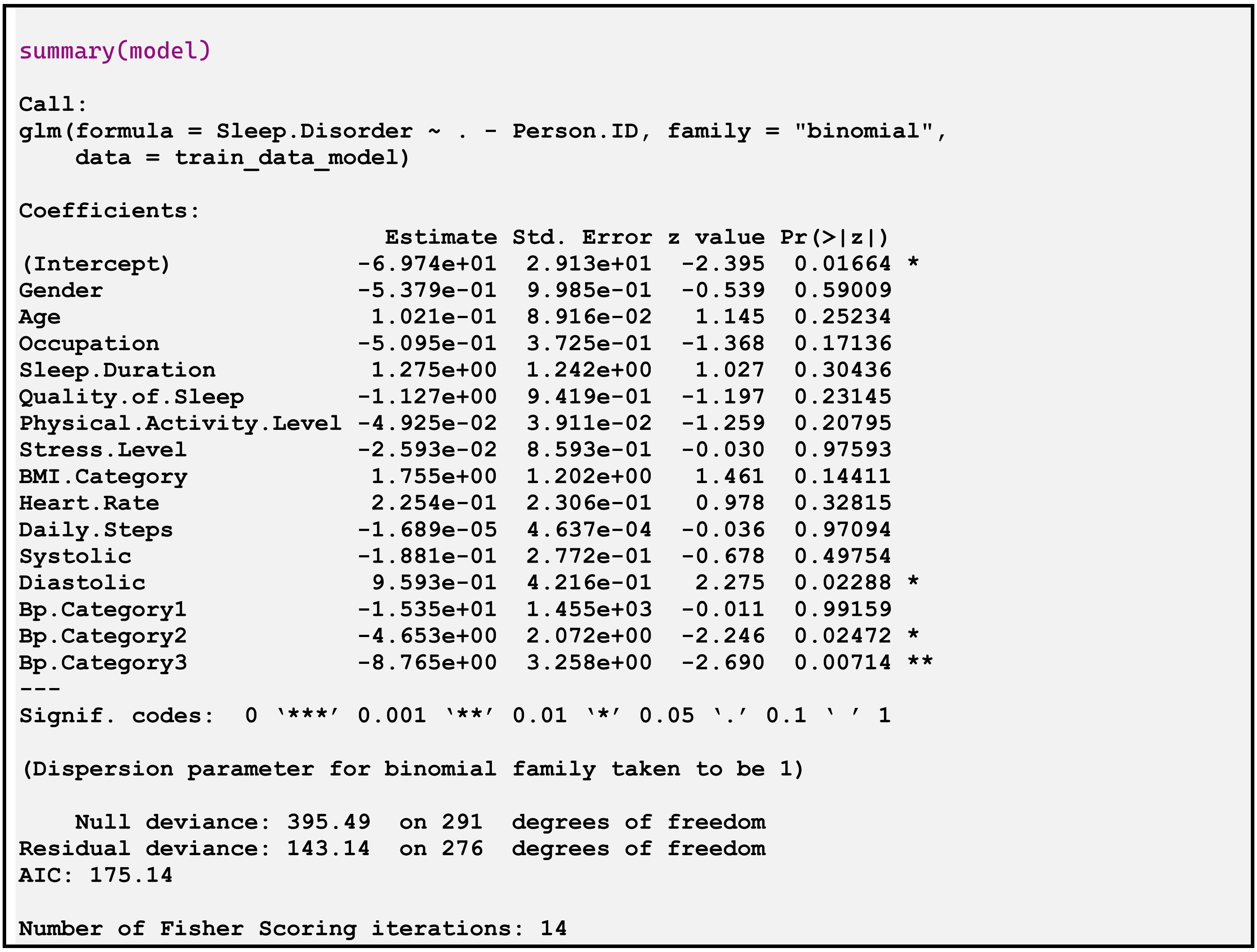
That is, e^{β} is the multiplicative change in the odds of success corresponding to a one-unit increase in x.

Data Pre-processing:

In our dataset, we have categorical columns such as $Gender$, $Occupation$, and $BMI.Category$, along with our response variable,$Sleep.Disorder$. Initially, I convert these categorical columns to factors, which is a basic requirement for fitting a GLM.
Next, I split the dataset into an 80:20 ratio, with 80% of the data used for training the model and 20% for testing the model.

Building a baseline model:

Building a baseline model is essential to understand how each predictor variable relates to the response variable and to identify which predictors significantly impact the response variable. I remove the $Person.ID$ column from my dataset since it merely represents a unique identifier for each patient and is not required for modelling purposes. I then fit a GLM, naming it $model$. The summary of this model is provided below:



```
summary(model)

Call:
glm(formula = Sleep.Disorder ~ . - Person.ID, family = "binomial",
    data = train_data_model)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -6.974e+01  2.913e+01  -2.395  0.01664 *
Gender                  -5.379e-01  9.985e-01  -0.539  0.59009
Age                      1.021e-01  8.916e-02   1.145  0.25234
Occupation              -5.095e-01  3.725e-01  -1.368  0.17136
Sleep.Duration           1.275e+00  1.242e+00   1.027  0.30436
Quality.of.Sleep        -1.127e+00  9.419e-01  -1.197  0.23145
Physical.Activity.Level -4.925e-02  3.911e-02  -1.259  0.20795
Stress.Level            -2.593e-02  8.593e-01  -0.030  0.97593
BMI.Category             1.755e+00  1.202e+00   1.461  0.14411
Heart.Rate               2.254e-01  2.306e-01   0.978  0.32815
Daily.Steps             -1.689e-05  4.637e-04  -0.036  0.97094
Systolic                -1.881e-01  2.772e-01  -0.678  0.49754
Diastolic                9.593e-01  4.216e-01   2.275  0.02288 *
Bp.Category1            -1.535e+01  1.455e+03  -0.011  0.99159
Bp.Category2            -4.653e+00  2.072e+00  -2.246  0.02472 *
Bp.Category3            -8.765e+00  3.258e+00  -2.690  0.00714 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 395.49  on 291  degrees of freedom
Residual deviance: 143.14  on 276  degrees of freedom
AIC: 175.14

Number of Fisher Scoring iterations: 14
```

Improvements using the *stepAIC* technique:

So far, I've been using all predictor variables, which might not be the most efficient approach. I aim to streamline the model by reducing the number of predictor variables using the *stepAIC* technique from the *MASS* library. I've updated the model and named it *model_step*. The summary of *model_step* is provided below:

Here, it's evident that the number of predictor variables has been reduced, accompanied by a decrease in AIC, indicating an improvement in the model. Let me clarify the concept of AIC:

Akaike Information Criterion (AIC):

The Akaike Information Criterion (AIC) is a measure used for model selection, particularly in the context of statistical modelling. It aims to balance the goodness of fit of the model and its complexity, penalizing overly complex models.

Formula:

$$AIC = 2K - 2ln(L)$$

Where:

- *k is the number of parameters in the model.*
- *L is the maximum value of the likelihood function for the model.*

Interpretation:

A model with a lower AIC is considered to be better. It suggests that the model fits the data well while using fewer parameters, thus being more parsimonious.

```
summary(model_step)

Call:
glm(formula = Sleep.Disorder ~ Occupation + Sleep.Duration +
    Quality.of.Sleep + Physical.Activity.Level + BMI.Category +
    Heart.Rate + Diastolic + Bp.Category, family = "binomial",
    data = train_data_model)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -77.48966   23.46201  -3.303 0.000957 ***
Occupation                -0.62400    0.32050  -1.947 0.051537 .
Sleep.Duration             1.77506    0.87259   2.034 0.041927 *
Quality.of.Sleep          -1.11309    0.63847  -1.743 0.081272 .
Physical.Activity.Level   -0.04835    0.02018  -2.396 0.016560 *
BMI.Category               2.71398    0.91275   2.973 0.002945 **
Heart.Rate                 0.21297    0.14538   1.465 0.142936
Diastolic                  0.78310    0.20258   3.866 0.000111 ***
Bp.Category1             -15.32027 1455.39821  -0.011 0.991601
Bp.Category2              -5.37633    1.85268  -2.902 0.003709 **
Bp.Category3              -9.03279    2.92206  -3.091 0.001993 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 395.49  on 291  degrees of freedom
Residual deviance: 144.55  on 281  degrees of freedom
AIC: 166.55

Number of Fisher Scoring iterations: 14
```

Improving the model by addressing multicollinearity:

While it's beneficial that the number of predictor variables has decreased, it's crucial to ensure that the model is free from multicollinearity. Multicollinearity can be problematic even if the model is simple and interpretable. Therefore, achieving a balance in the model is essential. To assess multicollinearity, I employ the *vif* function from the *car* library. Let's delve into what *VIF* (Variance Inflation Factor) represents:

Variance Inflation Factor (VIF):

The Variance Inflation Factor (VIF) is a statistical measure used to assess multicollinearity in regression models. It quantifies how much the variance of an estimated regression coefficient is increased due to collinearity with other predictor variables. Here are the key points about VIF:

Interpretation:

A VIF value greater than 1 indicates collinearity. Commonly, a VIF threshold of 5 or 10 is used to identify problematic multicollinearity. If a predictor's VIF is high (e.g., above 10), it suggests that the predictor is highly correlated with other predictors in the model.

Calculation:

The VIF for a predictor is computed as the ratio of the variance of the estimated coefficient when including all predictors to the variance of the estimated coefficient when excluding that specific predictor.

Mathematically, for a predictor i, the VIF is given by:

$$VIF_i = \frac{1}{1-{R_i}^2}$$

where $\boldsymbol{R_i}^{\mathbf{2}}$ is the coefficient of determination for the regression of predictor i on all other predictors.

Generalized VIF (GVIF):

The GVIF extends the concept of VIF to sets of predictor variables. It accounts for the fact that multiple columns in the model matrix and multiple coefficients may be associated with a single covariate (e.g., polynomial terms).

To make GVIFs comparable across dimensions, it's recommended to use $\boldsymbol{GVIF}^{(\mathbf{1}/\,(\mathbf{2}*\boldsymbol{Df}))}$, where Df is the number of coefficients in the subset.

For a single coefficient, GVIF simplifies to the usual VIF.

vif for $model\_step$ is given below:

Predictor Variable	GVIF	Df	GVIF^(1/(2*Df))
Occupation	10.144	1	3.185
Sleep Duration	9.439	1	3.072
Quality of Sleep	11.375	1	3.373
Physical Activity Level	3.751	1	1.937
BMI Category	4.130	1	2.032
Heart Rate	3.667	1	1.915
Diastolic	25.639	1	5.063
BP Category	12.902	3	1.531

It's evident from the output that the $model\_step$ is affected by multicollinearity issues, as many of the predictor variables have VIF values significantly large. This indicates high multicollinearity among the variables. To address this, I employ two methods for modification:

- *Firstly, I use the* Pearsonian chi square test *from the* stats *library, which tests for independence between categorical variables.*

- *Secondly, I utilize the* Goodman Kruskal Gamma *from the* DescTools *library, which measures the association between ordinal variables.*

Considerations for Variable Selection to Avoid Multicollinearity:

- **Occupation vs. Sleep Disorder:**

  The result of *Pearsonian chi square test* is given below:

```
chisq.test(table(data$Occupation,data$Sleep.Disorder))

        Pearson's Chi-squared test

data:  table(data$Occupation, data$Sleep.Disorder)
X-squared = 206.04, df = 6, p-value < 2.2e-16
```

  As null hypothesis is rejected, that indicates a significant association between *Occupation* and *Sleep.Disorder*.

- **Quality of sleep vs sleep disorder:**

  Since the quality of sleep is measured on a 10-point scale, it qualifies as an ordinal variable. Therefore, the *Pearsonian chi square test* is not suitable for this variable; instead, I use *Goodman Kruskal Gamma* to assess its association.

```
GoodmanKruskalGamma(data$Quality.of.Sleep,data$Sleep.Disorder)
[1] -0.4145457
```

  There is a moderate negative association between these two variables. To address multicollinearity, considering the high correlation between quality of sleep and sleep duration (0.88), as well as the negative correlation between sleep duration and stress level, removing both variables could enhance model stability and interpretability.

- **Stress level and sleep disorder:**

  Since the stress level is also measured on a 10-point scale, making it an ordinal variable, I employ the *Goodman Kruskal Gamma* to evaluate its association.

```
GoodmanKruskalGamma(data$Stress.Level,data$Sleep.Disorder)
[1] 0.1949237
```

  A slight positive association is evident. It would be beneficial to examine the association between stress level and quality of sleep. The result is presented below:

```
GoodmanKruskalGamma(data$Quality.of.Sleep,data$Stress.Level)
```

```
[1] -0.9144241
```

A strong negative association is evident. Therefore, for my case, I choose to include only the stress level to further improve the model, as it is easier to interpret.

- **BMI Category vs sleep disorder:**

  As BMI category is ordinal variable, *Goodman Kruskal Gamma* will be beneficial. The result is given below:

```
GoodmanKruskalGamma(table(data$BMI.Category,data$Sleep.Disorder))
[1] 0.9824195
```

A strong positive association is identified.

- **Bp category vs sleep disorder:**

  As bp category is ordinal variable, *Goodman kruskal Gamma* is used.

```
GoodmanKruskalGamma(table(data$Bp.Category,data$Sleep.Disorder))
[1] 0.9126208
```

A strong positive association is found. Now, it is valid to check association between bmi category and bp category. The result is given below:

```
GoodmanKruskalGamma(table(data$BMI.Category,data$Bp.Category))
[1] 0.9726704
```

As a strong association is observed, I take BMI category only for my future model.

- **Diastolic vs sleep disorder:**

  The result is given below:

```
GoodmanKruskalGamma(table(data$Diastolic,data$Sleep.Disorder))
[1] 0.9117596
```

It is obvious to find a positive association between diastolic and bp category. So, for my case, I remove *Diastolic* variable for future model.

- **Heart rate vs sleep disorder:**

I've observed a significant positive correlation between heart rate and stress level, as well as a notable negative correlation between heart rate and sleep quality. To mitigate multicollinearity in my final model, I'm considering removing the heart rate variable.

Summary:

After analysing the results, I have decided to exclude variables such as *Quality.of.Sleep*, *Sleep.Duration*, *Bp.Category*, *Heart.Rate* and *Diastolic* and adding *Stress.Level* from my future model.

My Final Model Selection:

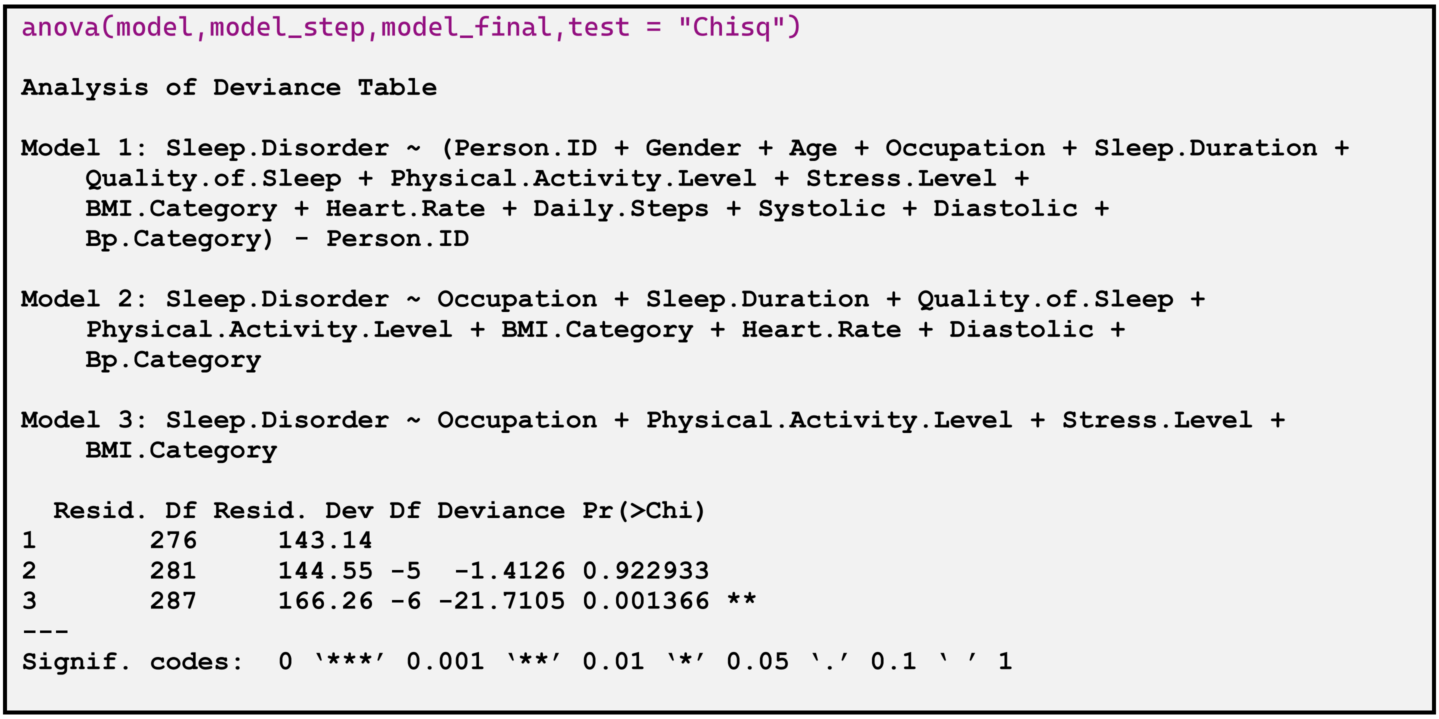
After addressing the reduced variable set and resolving multicollinearity issues, I fit my final model, naming it *model_final*. The summary of *model_final* is provided below:

```
summary(model_final)

Call:
glm(formula = Sleep.Disorder ~ Occupation + Physical.Activity.Level +
    Stress.Level + BMI.Category, family = "binomial", data = train_data_model)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -4.308e+00  1.165e+00  -3.699 0.000216 ***
Occupation               2.906e-01  1.405e-01   2.067 0.038698 *
Physical.Activity.Level  3.295e-06  1.151e-02   0.000 0.999772
Stress.Level             2.357e-01  1.313e-01   1.796 0.072478 .
BMI.Category             3.594e+00  5.682e-01   6.325 2.53e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 395.49  on 291  degrees of freedom
Residual deviance: 166.26  on 287  degrees of freedom
AIC: 176.26

Number of Fisher Scoring iterations: 5
```

Here, we can observe a slight increase in AIC; however, the model is now free from multicollinearity. The VIF results are presented below:

Variable	VIF
Occupation	2.136154
Physical.Activity.Level	1.129739
Stress.Level	1.191529
BMI.Category	1.898208

It's evident that *model_final* is the optimal model for my dataset. This can be theoretically demonstrated using the *anova* function to compare the baseline model, *model_step*, and *model_final*.



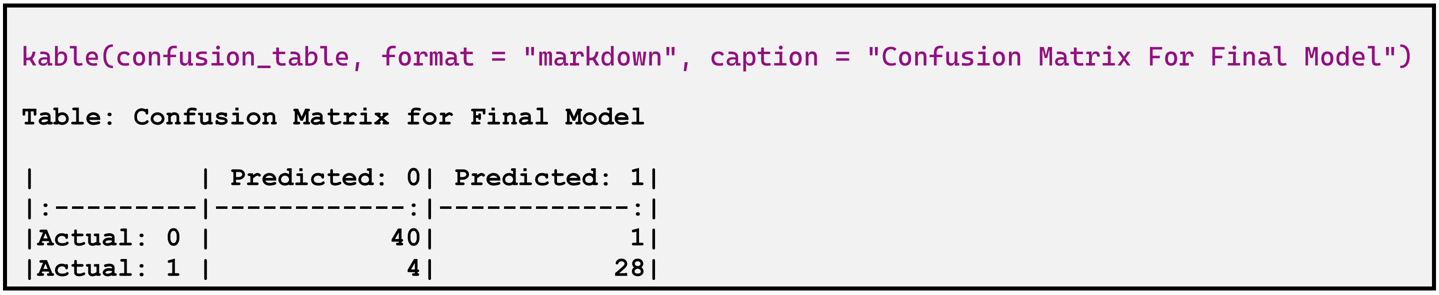
```
anova(model,model_step,model_final,test = "Chisq")

Analysis of Deviance Table

Model 1: Sleep.Disorder ~ (Person.ID + Gender + Age + Occupation + Sleep.Duration + 
    Quality.of.Sleep + Physical.Activity.Level + Stress.Level + 
    BMI.Category + Heart.Rate + Daily.Steps + Systolic + Diastolic + 
    Bp.Category) - Person.ID

Model 2: Sleep.Disorder ~ Occupation + Sleep.Duration + Quality.of.Sleep + 
    Physical.Activity.Level + BMI.Category + Heart.Rate + Diastolic + 
    Bp.Category

Model 3: Sleep.Disorder ~ Occupation + Physical.Activity.Level + Stress.Level + 
    BMI.Category

  Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
1       276     143.14                        
2       281     144.55 -5  -1.4126 0.922933   
3       287     166.26 -6 -21.7105 0.001366 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Predictions and Accuracy Calculation:

I have our final model named *model_final*. Now, I want to make predictions on my test data, which is named *test_data_model*, using this trained model. To achieve this, I'll use the *predict* function to obtain the probabilities and store them in a variable called *predictions_model_final*.

Now, I convert the predictions to binary values (0 or 1) based on a threshold of 0.5, and store them in a variable named *predicted_classes_model_final*.

Next, I create a confusion table named *confusion_table* using the *kable* function from the *knitr* package. This table compares the actual sleep disorder values (0 or 1) from *test_data_model* with the predicted sleep disorder values from *predicted_classes_model_final*. The table is presented below:



```
kable(confusion_table, format = "markdown", caption = "Confusion Matrix For Final Model")

Table: Confusion Matrix for Final Model

|          | Predicted: 0| Predicted: 1|
|:---------|------------:|------------:|
|Actual: 0 |           40|            1|
|Actual: 1 |            4|           28|
```

Now, I am interested in calculating the accuracy of my model. Let me explain what accuracy means beforehand:

Accuracy:

Accuracy of the model is the proportion of correctly predicted observations to the total observations. Formally, accuracy has the following definition:

$$accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

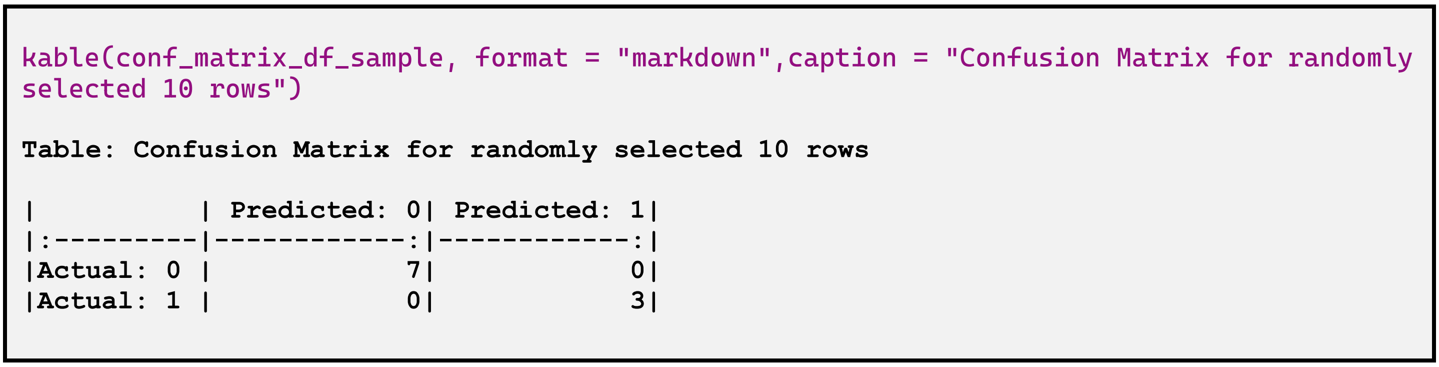
For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

$$accuracy = \frac{(true\ positive + true\ negative)}{(true\ positive + true\ negative + false\ positive + false\ negative)}$$

Our model has an accuracy rate of 93.15%.

Cross Validation:

Now, I want to select a random sample of size 10 from the test data (*test_data_model*). After calculating the predicted probabilities and assigning them to predicted classes as before, I'll assess the accuracy of my model based on this sample. To do this, I generate a confusion matrix for this random sample of size 10. Here are the results:



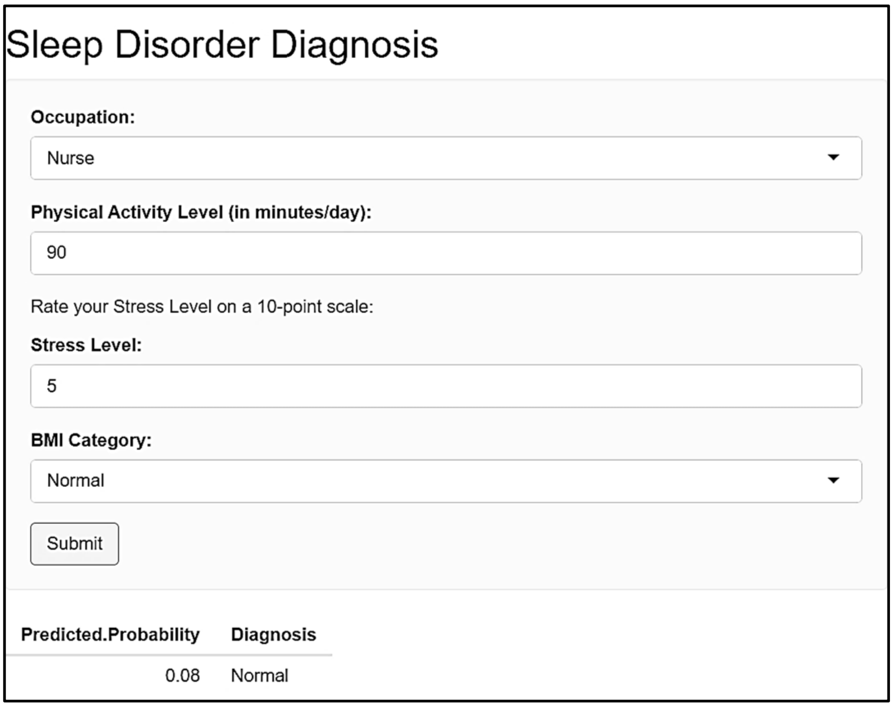
```
kable(conf_matrix_df_sample, format = "markdown",caption = "Confusion Matrix for randomly
selected 10 rows")

Table: Confusion Matrix for randomly selected 10 rows

|          | Predicted: 0| Predicted: 1|
|:---------|------------:|------------:|
|Actual: 0 |            7|            0|
|Actual: 1 |            0|            3|
```

In a promising development, my model accurately predicts sleep disorders for all 10 randomly selected rows, achieving 100% accuracy. This perfect accuracy rate signifies a significant breakthrough in sleep disorder prediction.

To further enhance the accessibility of this technology, I've constructed a user-friendly application. This HTML form, built using the *shiny* package in *R*, allows users to input their data and receive a calculated probability of having a sleep disorder. Based on the generated probability, the form provides a preliminary diagnosis, classifying them as either low-risk or potentially affected.

This user-interface serves as a valuable tool for raising public awareness and potentially encouraging individuals to seek professional medical evaluation.

Take a look at the HTML form:



Sleep Disorder Diagnosis

Occupation:

Nurse

Physical Activity Level (in minutes/day):

90

Rate your Stress Level on a 10-point scale:

Stress Level:

5

BMI Category:

Normal

Predicted.Probability	Diagnosis
0.08	Normal

Limitations:

- *Note that the form is designed specifically for the occupations listed in my dataset. If someone from these professions inputs their details, the form will provide accurate results with an overall accuracy of up to 94%. Inputs for occupations not included in my dataset will not be accepted.*

- *For accurate predictions, it's crucial that the input provider takes good care of themselves and provides honest and accurate information. They should know their daily activity level in minutes and rate their stress level based on their own perception, as stress can be subjective.*

- *Additionally, they should be aware of their BMI category. However, it's not uncommon for people to inaccurately report their BMI category due to societal pressures or personal concerns. When incorrect or misleading information is provided, it can lead the model to generate inaccurate predictions. Therefore, it's essential for users to be honest and provide genuine data for the model to produce reliable results.*

Project Evolution: Paving the Path for Future Enhancements:

My successful development of a sleep disorder prediction model, leveraging basic statistical tools from my BSc studies, marks an initial milestone in my project's journey. While adhering strictly to syllabus constraints, I've achieved a commendable accuracy of up to 94%. Looking ahead, my pursuit of advanced concepts in my master's program promises to refine and evolve the model, integrating sophisticated tools and methodologies for even greater predictive precision. This project represents not only the culmination of my undergraduate studies but also the foundation upon which future advancements in sleep disorder prediction will be built.

Here's a concise version of your future enhancement plan:

- *Utilized Kaggle dataset, partitioned for training and testing due to survey constraints.*
- *Prioritize gathering gender-specific data on sleep patterns and disorders to inform the project's extension effectively.*
- *Currently reliant on logistic regression but plan to explore other machine learning models post advanced AI study for greater accuracy.*
- *Considering additional professions and data collection methods to broaden occupational scope.*
- *Simplified sleep disorder categories for logistic regression, aiming to incorporate multinomial logistic regression in the future.*
- *Recognized data biases and aim to include new variables like smoking habits, medical conditions, etc., for real-world readiness*
- *Developed an HTML form, intending to expand it into a free website for sleep disorder diagnosis and plan to link form with Excel database for ongoing model enhancement.*
- *Future collaboration with medical professionals to refine model based on real-world feedback.*

Acknowledgement

In culmination of my Bachelor of Science (Honours in Statistics) journey, I humbly present this project work as an essential component of my evaluation for the DSE-04 paper. I extend my sincere gratitude to my esteemed college, Ramakrishna Mission Residential College (Autonomous) and the Department of Statistics for affording me the opportunity to embark on this academic venture.

Foremost, I express my deepest appreciation to my project supervisor, Sri Palas Pal, whose unwavering guidance, insightful suggestions, and steadfast support have been invaluable throughout this endeavour. Their expertise and constructive feedback have not only shaped the trajectory of this project but also enriched its quality.

I am profoundly thankful to the esteemed professors of our Department, Dr. Parthasarathi Chakrabarti, Subhadeep Banerjee, and Tulsidas Mukhopadhyay, whose mentorship has equipped me with the requisite tools to comprehend the philosophical underpinnings of the subject and translate theory into practice with adeptness.

Moreover, I wish to extend heartfelt gratitude to my friends, whose unwavering encouragement, motivation, and intellectual discourse have been instrumental in propelling the success of this endeavor. Their collaborative efforts and invaluable insights have undeniably enriched the outcome of this project.

Lastly, I express my deepest appreciation to my family for their unwavering support and understanding throughout this journey. Their unwavering encouragement, patience, and unwavering belief in my capabilities have served as the cornerstone of my perseverance and determination to see this project to fruition.

References

This project benefitted from a diverse range of data sources and analytical resources. The data collection phase drew upon various sources, while the analysis phase leveraged assistance from multiple resources. They are:

1. ***Sleep Health and Lifestyle Dataset***, ***Kaggle***

   https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

2. *For definitions and more field knowledges*
   - The American Academy of Sleep Medicine (AASM)
   - World Health Organisation (WHO)

3. *Dalpiaz, David. "**Applied Statistics with R**"*

4. *Bruce, Peter, Andrew Bruce, and Peter Gedeck. "**Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python**" O'Reilly Publication.*

5. *Braun, W. John, and Duncan J. Murdoch. "**A First Course in Statistical Programming with R**"*

6. *"**Akaike Information Criterion (AIC)**"* Akaike information criterion (AIC)

7. ***"Variance Inflation Factor (VIF)"*** Variance Inflation Factor (VIF)

8. *"**Accuracy**"* Accuracy