Analysis on Factors that Contributes to Sleep Disorders

Sekolah Data Pacmann
Statistics for Business Project

Prepared by:
Student: Dewi Astuti (dewi-I4Zs)
Class: Continuation (BI+DS)
Email: dewiastutisuhendro@gmail.com

TABLE OF CONTENTS

Analysis on Factors that Contributes to Sleep Disorders
Statistics for Business Project

Dewi Astuti: dewiastutisuhendro@gmail.com

Abstract

As a healthcare provider, we analyze sleep as a fundamental physiological process that plays a crucial role in maintaining our overall health and well-being. Based on the recorded data of employees from different industries, the data is explored to identify underlying patterns in sleep data and the major factors to sleep disorders, enabling businesses to make informed decisions to enhance employee wellness programs and optimize productivity. The statistics are evaluated using the Regression Model.

Keywords: Sleep Disorders, OLS Regression, Hypothesis, Modelling

A. Introduction

A.1. Background

In today's fast-paced world, the importance of a good night's sleep cannot be overstated. Sleep is a fundamental physiological process that plays a crucial role in maintaining our overall health and well-being. However, sleep disorders have become a prevalent and concerning issue affecting millions of individuals worldwide.

This project aims to delve into the realm of sleep disorders from a statistical perspective and shed light on the various aspects that influence sleep quality and quantity among employees, enabling businesses to make informed decisions to enhance employee wellness programs and optimize productivity.

A.2. Problem Statement

Sleep disorders can significantly affect an individual's quality of life, leading to various health, psychological, and social issues. With the increasing prevalence of sleep-related problems, there is a growing need to understand the factors that contribute to sleep disorders. This analysis aims to examine the relationships between different demographic, lifestyle, and health factors (e.g., age, gender, physical activity, and mental health status) and the occurrence of sleep disorders.[1]

Using a sleep disorder dataset, this study seeks to identify key patterns and potential predictors of sleep disorders through various statistical techniques, including regression models and ANOVA. The findings from this analysis can provide insights into targeted interventions for improving sleep health.

Key research questions include:

1. What demographic and lifestyle factors are associated with the likelihood of developing sleep disorders?
2. How do mental and physical health conditions affect sleep quality and the prevalence of sleep disorders?
3. Can statistical models effectively predict the occurrence of sleep disorders based on available data?

A.3. Goals

The primary goal of this Statistics for Business project is to apply linear regression models to sleep-related data to conduct a comprehensive analysis of sleep disorders.[2] Through this analysis, we aim to achieve the following objectives:

1. Identify the prevalence and types of sleep disorders among the employee population.
2. Determine the factors that significantly influence sleep quality and quantity.
3. Explore the correlation between sleep patterns and work-related performance metrics.
4. Propose evidence-based strategies to improve sleep health and well-being among employees.
5. Provide actionable insights for organizations to implement targeted wellness programs that address sleep-related issues and enhance overall productivity.

B. Dataset

B.1. Description

The Sleep Health and Lifestyle Dataset is obtained from Kaggle and comprises a wide range of variables related to sleep and daily habits.

	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Blood Pressure	Heart Rate	Daily Steps	Sleep Disorder
0	1	Male	27	Software Engineer	6.1	6	42	6	Overweight	126/83	77	4200	None
1	2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
2	3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	None
3	4	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
4	5	Male	28	Sales Representative	5.9	4	30	8	Obese	140/90	85	3000	Sleep Apnea
...	...	...	...	...	...	...	...	...	...	...	...	...	...
369	370	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	Sleep Apnea
370	371	Female	59	Nurse	8.0	9	75	3	Overweight	140/95	68	7000	Sleep Apnea
371	372	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	Sleep Apnea
372	373	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	Sleep Apnea
373	374	Female	59	Nurse	8.1	9	75	3	Overweight	140/95	68	7000	Sleep Apnea

374 rows × 13 columns

Figure 1. Sleep Disorder Dataset

It includes details such as gender, age, occupation, sleep duration, quality of sleep, physical activity level, stress levels, BMI category, blood pressure, heart rate, daily steps, and the presence or absence of sleep disorders.

Data Variables:

Key Features	No.	Name	About
Personal Profile	1	Person ID	A unique identifier for each person.
	2	Gender	The person's gender (Male/Female).
	3	Age	The person's age in years.
	4	Occupation	The person's profession or job.
Comprehensive Sleep Metrics	5	Sleep Duration	The number of hours the person sleeps per day.
	6	Quality of Sleep	A subjective rating of the person's sleep quality, ranging from 1 to 10.
Lifestyle Factors	7	Physical Activity Level	The number of minutes the person engages in physical activity daily.
	8	Stress Level	A subjective rating of the person's stress level, ranging from 1 to 10.
	9	BMI Category	The person's BMI category, such as Underweight, Normal, or Overweight.
	10	Daily Steps	The number of steps the person takes per day.
Cardiovascular Health	11	Blood Pressure	The blood pressure measurement of the person, represented as systolic pressure over diastolic pressure.
	12	Heart Rate	In bpm. The person's resting heart rate in beats per minute.
Sleep Disorder Analysis	13	Sleep Disorder	Indicates whether the person has a sleep disorder (None, Insomnia, Sleep Apnea).

Figure 2. Data Variables

Details about Sleep Disorder Column:

- None: The person does not show any specific sleep disorder.
- Insomnia: The person faces challenges in initiating or maintaining sleep, which leads to insufficient or substandard sleep quality.
- Sleep Apnea: The person experiences interruptions in breathing while asleep, causing disruptions in sleep patterns and possible health hazards.

B.2. Data Cleaning

The observations go through 5 processes of data cleaning, namely:

1. Drop Unused Variable which is the ‘Person ID’
2. Replace the column names with simpler names:
   : ‘gender', 'age', 'occupation', 'sleep_duration', ‘quality_of_sleep', 'physical_activity_level', 'stress_level', 'bmi_category', 'blood_pressure', 'heart_rate', 'daily_steps', and 'sleep_disorder'
3. Check Number of Unique Values in Each Column
   - Person ID 374
   - Gender 2
   - Age 31
   - Occupation 11
   - Sleep Duration 27
   - Quality of Sleep 6
   - Physical Activity Level 16
   - Stress Level 6
   - BMI Category 4
   - Blood Pressure 25
   - Heart Rate 19
   - Daily Steps 20
   - Sleep Disorder 3

4. Check Missing Values and Duplicated Data
   Result:
   - Data is clean, no need to handle missing values.
   - No duplicated data, we don't need to handle duplicate data.
   - The nan value in sleep disorder stands for no sleep disorder, so it is not a missing value, thus has been encoded as 'None'.
5. Transform Categorical Parameters
   - Occupation (multi-class): Some occupations are underrepresented in the data, making it challenging to draw meaningful conclusions. So all occupations with fewer than 20 instances have been designated to the "unknown" class. The 'Sales Representative' occupation has also been referred to as 'Salesperson'.
   - BMI Category (multi-class): The distribution between overweight and normal categories is well-balanced. However, there appears to be a typo in the category label "Normal Weight," which will be corrected to "Normal."
   - Sleep Patterns and Blood Pressure. In another study, sleep patterns significantly predicted SBP and DBP, thus it is important to look in detail into Systolic Blood Pressure (SBP) and Diastolic Blood Pressure (DBP), while 'blood pressure' has been dropped.

C. Exploratory Data Analysis

The several steps of EDA namely:
C.1. Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
age	374.0	42.184492	8.673133	27.0	35.25	43.0	50.0	59.0
sleep_duration	374.0	7.132086	0.795657	5.8	6.40	7.2	7.8	8.5
quality_of_sleep	374.0	7.312834	1.196956	4.0	6.00	7.0	8.0	9.0
physical_activity_level	374.0	59.171123	20.830804	30.0	45.00	60.0	75.0	90.0
stress_level	374.0	5.385027	1.774526	3.0	4.00	5.0	7.0	8.0
heart_rate	374.0	70.165775	4.135676	65.0	68.00	70.0	72.0	86.0
daily_steps	374.0	6816.844920	1617.915679	3000.0	5600.00	7000.0	8000.0	10000.0
systolic	374.0	128.553476	7.748118	115.0	125.00	130.0	135.0	142.0
diastolic	374.0	84.649733	6.161611	75.0	80.00	85.0	90.0	95.0

```
{'gender': array(['Male', 'Female'], dtype=object),
 'occupation': array(['Unknown', 'Doctor', 'Salesperson', 'Teacher', 'Nurse', 'Engineer',
        'Accountant', 'Lawyer'], dtype=object),
 'bmi_category': array(['Overweight', 'Normal', 'Obese'], dtype=object),
 'sleep_disorder': array(['None', 'Sleep Apnea', 'Insomnia'], dtype=object)}
```

Figure 3. Descriptive Statistics

Based on the provided dataset of 374 individuals, here are a few observations and potential insights:

1. The average age is 42 years.
2. Sleep duration averages 7.13 hours, with a decent quality rating of 7.31/10.
3. Individuals exercise for about 59 minutes daily, but their step count (6816 steps) is below the recommended 10,000.
4. The average stress level is moderate at 5.38/10.
5. Heart rate averages 70 bpm, which is normal, while blood pressure (128/84) is nearing the upper limit of normal. Overall, the group appears relatively healthy, but some may face risks due to stress, insufficient activity, or elevated blood pressure.

General Insights:

- Health: While most metrics suggest that individuals in this dataset are relatively healthy, there are signs of potential health risks such as stress levels, insufficient sleep for some, and borderline high blood pressure.
- Activity & Lifestyle: Daily physical activity and steps are slightly below optimal levels, and there may be a relationship between lower activity levels and higher stress or sleep quality ratings.
- Further analysis could explore correlations between these variables (e.g., stress and sleep, physical activity and heart rate).

C. 2. Distribution of Key Numerical and Categorical Variables
Visualization of each feature is carried out to see the distribution of key numerical and categorical variables and their relationship with the presence of a sleep disorder.

Numerical - Boxplot

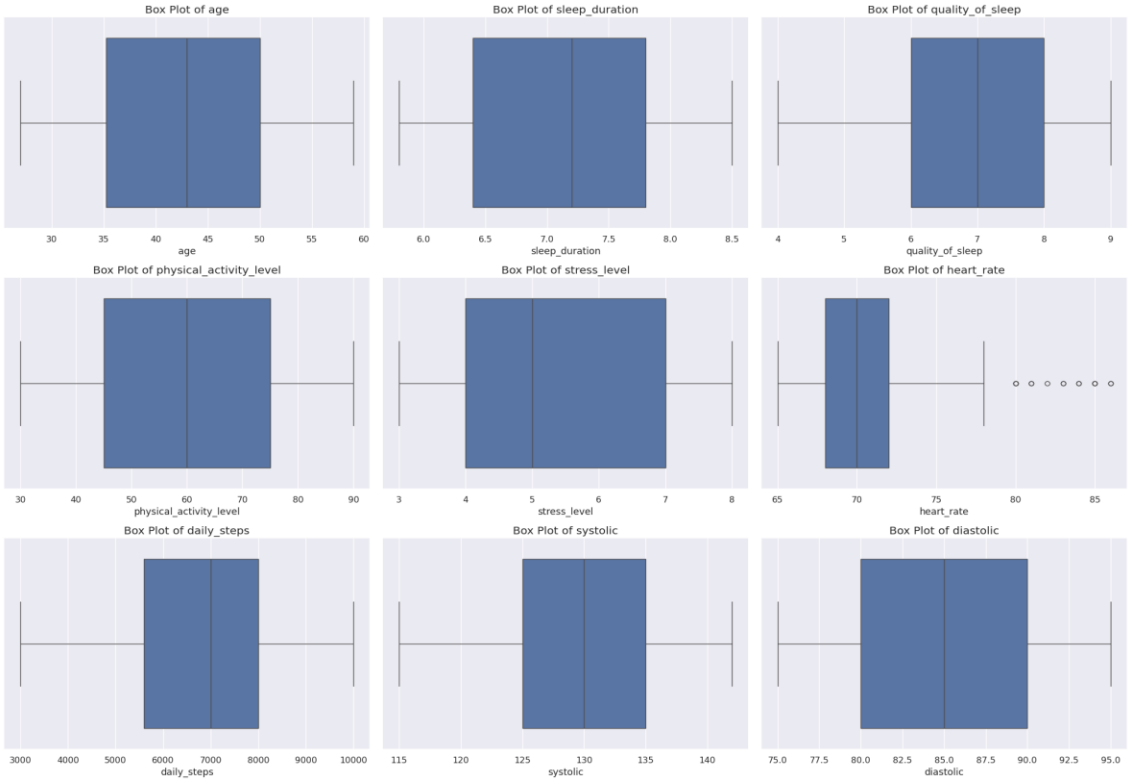

Figure 4. Boxplot of all numerical variables

Observations:

- Distribution of all numerical variables.
- Most variables are balanced/well-distributed with no major outliers, except for heart rate, which has a few values above 80 bpm.
- Sleep duration (6.5-7.5 hours) and quality (5-9) are consistent.
- Physical activity (50-70 minutes), stress (4-7), and daily steps (5000-8000) are moderate.
- Blood pressure is within normal ranges, with systolic (120-140) and diastolic (80-90) values.
- Overall, the dataset appears balanced, with only heart rate showing notable outliers.

Numerical - Correlation Matrix

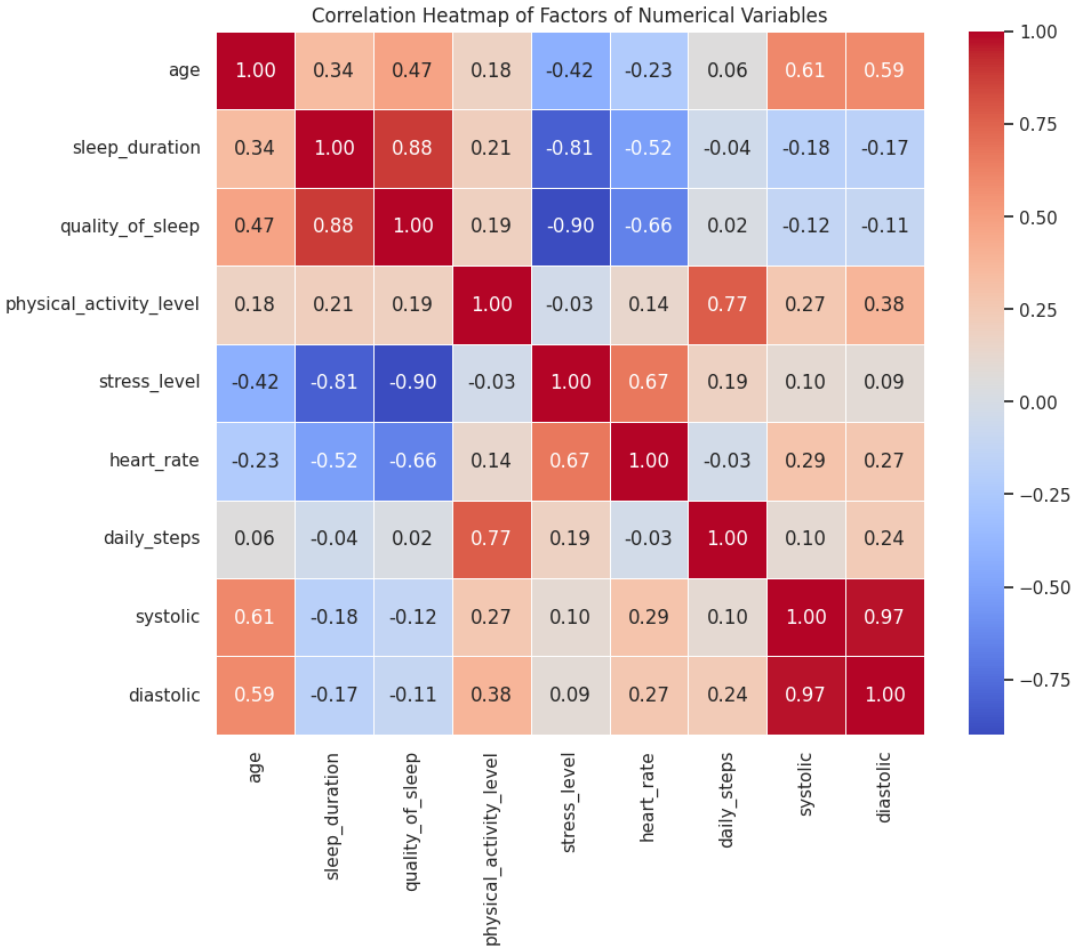

Figure 5. Correlation Heatmap of Factors of Numerical Variables

Observations:

- Analysis on the correlation between all numerical variables.
- Sleep quality is positively correlated with sleep duration (0.88) and negatively with stress (-0.90).
- Physical activity reduces stress (-0.66) and improves sleep quality (0.49).
- Age is strongly linked to higher blood pressure (systolic 0.61, diastolic 0.59).
- Heart rate is negatively related to sleep quality (-0.66) and positively to stress (0.60).
- Daily steps are strongly related to overall physical activity (0.77).
- These insights suggest stress, activity, and sleep duration are key factors influencing sleep quality.

Numerical - Pairplot

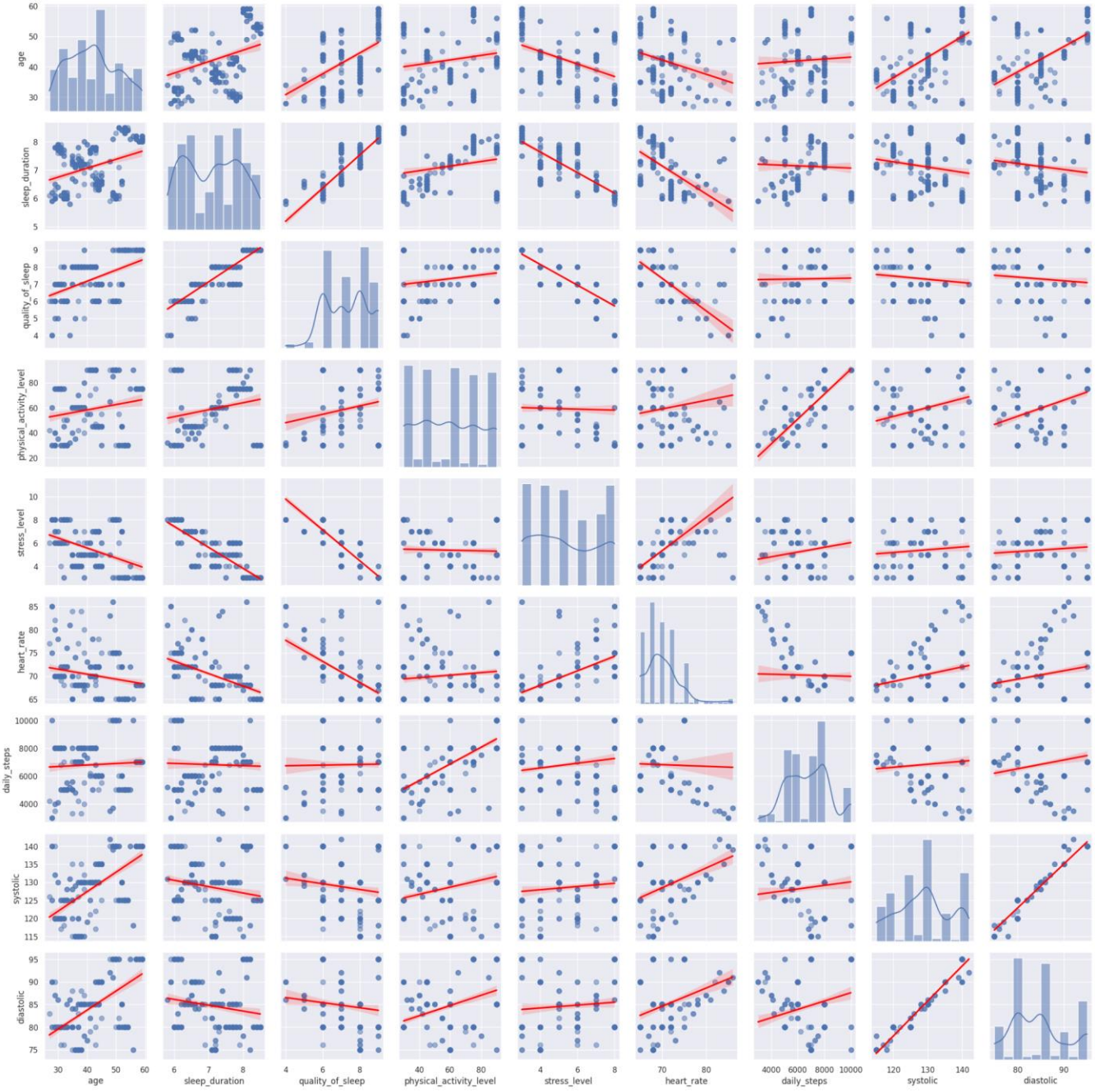

Figure 6. Pair Grid on sleep quality among all numerical variables

Observations:

- Analysis on sleep quality among all numerical variables.
- Sleep quality improves with longer sleep duration and more physical activity, but decreases with higher stress and heart rate.
- Age correlates positively with higher blood pressure.
- Stress reduces both sleep duration and sleep quality, but is lowered by more physical activity.
- Daily steps are strongly linked to overall physical activity.
- Key factors like stress, activity, and sleep duration heavily influence sleep quality.

Categorical - Barplot

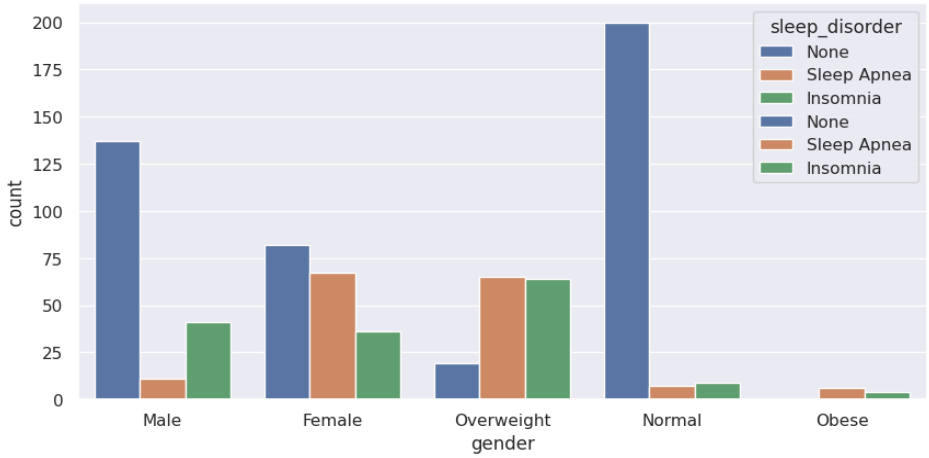

Figure 7. Barplot of sleep disorder among BMI categories and genders

Observations:

- Analysis on sleep disorders among BMI categories and genders.
- Sleep Apnea is more common in females and overweight/obese individuals.
- Insomnia is more prevalent in males and those overweight.
- Most normal-weight individuals are free from sleep disorders.
- Overall, sleep disorders are more frequent in females and those with higher weight.

Categorical - Count

Observations:

- Analysis on sleep disorders among occupations.
- The top 3 occupations with Insomnia are Salesperson, Teacher, and Accountant.
- The top 3 occupations with No Sleep Disorder are Doctor, Engineer, and Lawyer.
- Nurses have the highest incidence of Sleep Apnea, while Teachers and Salespersons have the highest rates of Insomnia.

sleep_disorder	occupation	count
Insomnia	Salesperson	29
	Teacher	27
	Accountant	7
	Engineer	5
	Doctor	3
	Nurse	3
	Lawyer	2
	Unknown	1
None	Doctor	64
	Engineer	57
	Lawyer	42
	Accountant	30
	Teacher	9
	Nurse	9
	Unknown	6
	Salesperson	2
Sleep Apnea	Nurse	61
	Doctor	4
	Teacher	4
	Lawyer	3
	Salesperson	3
	Unknown	2
	Engineer	1

dtype: int64

Figure 8. Table of sleep disorders among occupations

Categorical - Boxplot

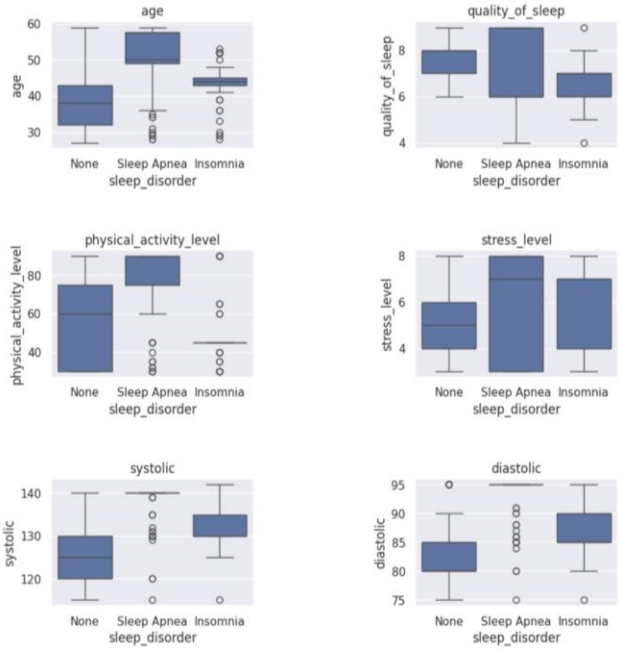

Figure 9. Boxplot on sleep disorder among several numerical variables

Observations:

- Analysis on sleep disorders among several numerical variables.
- Age: Individuals with Sleep Apnea tend to be older.
- Sleep Quality: Those with Insomnia and Sleep Apnea report lower sleep quality compared to those with no disorders.
- Physical Activity: Sleep Apnea patients are generally more active, while Insomnia sufferers have lower activity levels.
- Stress Levels: Insomnia is associated with higher stress levels, followed by Sleep Apnea.
- Blood Pressure: Both systolic and diastolic blood pressure are higher in individuals with Sleep Apnea.

Categorical - Boxplot

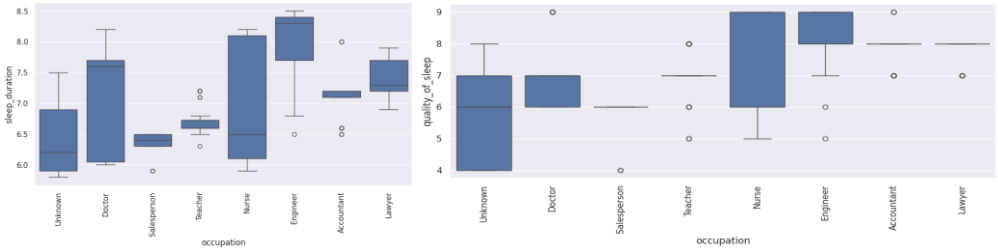

Figure 10. Boxplot on sleep duration and quality of sleep among occupations

Observations:

- Analysis on sleep duration and quality of sleep among occupations.[3]
- Sleep Duration: Nurses and Engineers have longer sleep durations, while Salespersons and Teachers report the shortest sleep.
- Quality of Sleep: Nurses and Engineers experience better sleep quality, while Teachers and Salespersons report lower sleep quality.
- Overall, nurses and engineers seem to fare better in terms of sleep, while salespersons and teachers struggle with both sleep duration and quality.

Categorical - Lineplot

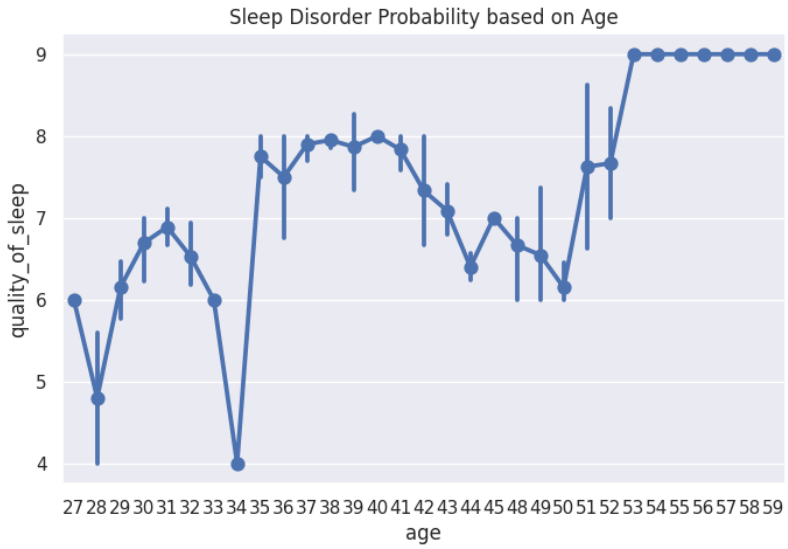

Figure 11. Lineplot with error bars on sleep disorder probability based on quality of sleep in people of various age

This chart appears to be a line plot with error bars, showing the relationship between age (on the x-axis) and the quality of sleep (on the y-axis). The title suggests it's related to the probability of sleep disorders across different age groups.

Observations:

- Analysis on sleep disorder probability based on quality of sleep in people of various ages.
- Fluctuations in younger ages (27-34): The sleep quality varies considerably in this range, with notable dips (such as at age 34) indicating possible higher probabilities of sleep disorders or disruptions.
- Stability in mid-30s to early 40s: Between ages 35 and 43, there's a relatively stable and higher sleep quality, suggesting a lower chance of sleep disorders.
- Decline from early 40s to late 40s: From 44 to 48, there is a noticeable decline in sleep quality, possibly indicating increasing sleep disturbances in this age range.
- Improvement and stabilization in early 50s: Sleep quality improves significantly around age 50 and remains high and stable up to age 59, suggesting a reduced probability of sleep disorders later in life.
- The error bars indicate the variability or uncertainty in the measurements at each age.

D. Statistical Test

D.1. Statistical Storyline

This is the structured flow that takes into account the research context, the types of variables involved, the hypotheses to be tested, and the insights we aim to derive from these analyses.

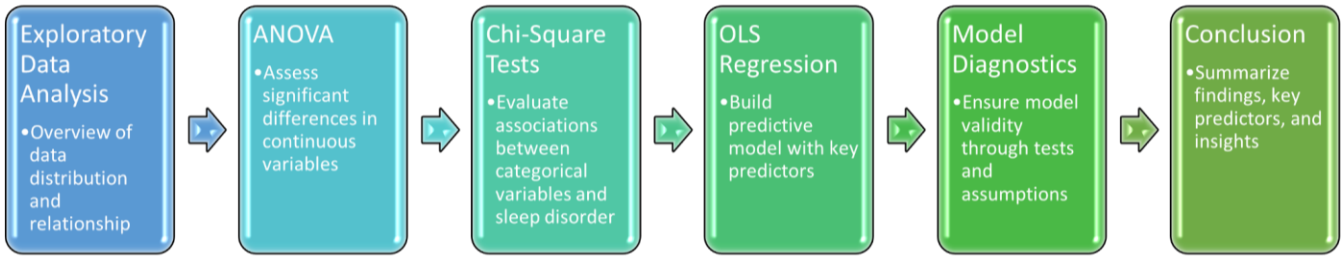

Figure 12. Statistical storyline

This storyline allows for a comprehensive statistical analysis of the dataset, exploring both categorical and continuous factors that could influence sleep disorders. By combining ANOVA, Chi-Square tests, and OLS regression, we aim to derive meaningful conclusions on how various factors affect the likelihood of experiencing sleep disorders.

D.2. Statistical Testing

We did a label encoding for categorical variables in X to convert categorical data into numerical format, which machine learning algorithms can more easily process. Now we have all the 12 categories in numerical format and we can do non-tree based models such as linear regression or logistic regression.

```
(    gender  age  occupation  sleep_duration  quality_of_sleep  \
0        1   27           7             6.1                 6
1        1   28           1             6.2                 6
2        1   28           1             6.2                 6
3        1   28           5             5.9                 4
4        1   28           5             5.9                 4

   physical_activity_level  stress_level  bmi_category  heart_rate  \
0                       42             6             2          77
1                       60             8             0          75
2                       60             8             0          75
3                       30             8             1          85
4                       30             8             1          85

   daily_steps  systolic  diastolic
0         4200       126         83
1        10000       125         80
2        10000       125         80
3         3000       140         90
4         3000       140         90  ,
array([1, 1, 1, 2, 2]))
```

Figure 13. Label Encoding

The label-encoded categorical variables allow the model to identify patterns and relationships between these features and the target variable (e.g., the presence or severity of sleep disorders), enabling better model training and predictions for sleep disorders.

Feature Importance

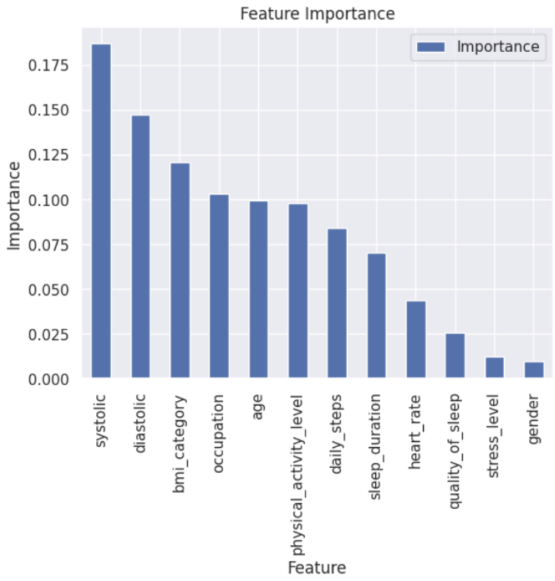

Figure 14. Feature Importance

We then trained Random Forest Classifiers to extract feature importance.

The feature importance analysis from the Random Forest model reveals that the most significant factor in predicting sleep disorders is Systolic blood pressure, with an importance of approximately 19%. This is followed by Diastolic blood pressure, BMI category, Occupation, Age, Physical Activity Level, and other features.[5]

We then excluded variables that don't have a significant effect on the sleep disorder condition, namely: stress level, quality of sleep, heart rate, sleep duration, and daily steps.

	gender	age	occupation	physical_activity_level	bmi_category	sleep_disorder	systolic	diastolic
0	Male	27	Unknown	42	Overweight	None	126	83
1	Male	28	Doctor	60	Normal	None	125	80
2	Male	28	Doctor	60	Normal	None	125	80
3	Male	28	Salesperson	30	Obese	Sleep Apnea	140	90
4	Male	28	Salesperson	30	Obese	Sleep Apnea	140	90
...	...	...	...	...	...	...	...	...
369	Female	59	Nurse	75	Overweight	Sleep Apnea	140	95
370	Female	59	Nurse	75	Overweight	Sleep Apnea	140	95
371	Female	59	Nurse	75	Overweight	Sleep Apnea	140	95
372	Female	59	Nurse	75	Overweight	Sleep Apnea	140	95
373	Female	59	Nurse	75	Overweight	Sleep Apnea	140	95

374 rows × 8 columns

Figure 15. After less significant variables been excluded

D.3. Hypothesis

ANOVA

We performed ANOVA (F-test) to check the hypothesis and significance level, so we came up with the conclusion based on the p-value.

```
ANOVA result for column:  Gender
                    sum_sq     df         F    PR(>F)
C(Q("Gender"))    3.490392    1.0  5.581156  0.018669
Residual        232.644582  372.0       NaN       NaN


ANOVA result for column:  Occupation
                       sum_sq     df          F        PR(>F)
C(Q("Occupation"))   83.728042    7.0  28.724287  1.703909e-31
Residual            152.406932  366.0        NaN           NaN


ANOVA result for column:  BMI Category
                          sum_sq     df          F        PR(>F)
C(Q("BMI Category"))   33.475890    1.0  61.448176  4.836096e-14
Residual              202.659083  372.0        NaN           NaN


ANOVA result for column:  Sleep Duration
                              sum_sq     df             F  PR(>F)
C(Q("Sleep Duration"))  2.361350e+02   26.0  2.279076e+29     0.0
Residual                1.382794e-26  347.0           NaN     NaN
```

Figure 16. ANOVA

This one-way ANOVA (Analysis of Variance) is done on multiple independent variables, testing how Gender, Occupation, BMI Category, and Sleep Duration impact one dependent variable with p-values below a chosen significance level (0.05) indicating statistically significant group differences.

The ANOVA tests the null hypothesis that the means of different groups (based on these categorical variables) are the same. The results show the sum of squares (sum_sq), degrees of freedom (df), F-statistic (F), and p-values (PR(>F)) for each variable, helping determine if there are statistically significant differences between group means.

In summary, all four variables have a statistically significant impact on sleep disorders, with occupation, BMI category, and sleep duration showing particularly strong effects.

Chi-square Test for Independence

Perform the Chi-square test for independence on BMI category and occupation to sleep disorder. Based on the results of the Chi-square tests for independence, here are key observations:

```
Chi-square Test for Independence
================================
Chi2 Statistic: 245.66534355746683
p-value: 5.5883512097923584e-52
Degrees of Freedom: 4

Chi-square Test for Independence
================================
Chi2 Statistic: 401.6055551972881
p-value: 5.8190777743484385e-77
Degrees of Freedom: 14
```

Figure 17. Chi-square test for independence

1. Test for BMI Category and Sleep Disorder:

- The p-value is much smaller than the typical significance level (e.g., 0.05), so we reject the null hypothesis (H0), which states there is no association between BMI category and sleep disorder.
- This suggests a significant association between BMI category and sleep disorder, indicating that individuals' BMI classifications (e.g., normal, overweight, obese) are related to whether they experience sleep disorders.

2. Test for Occupation and Sleep Disorder:
- Again, the p-value is far below the common significance threshold (e.g., 0.05), leading to the rejection of the null hypothesis (H0) that there is no association between occupation and sleep disorder.
- This means there is a significant association between occupation and sleep disorder, implying that the type of occupation a person holds is related to their likelihood of experiencing sleep disorders.

E. Regression Model

E.1. Model Design

Based on the results of previous statistical tests, there are several variables that significantly influence the condition of a bad sleeping habit. Those variables were prioritized in building the regression model.

The decision to use Ordinary Least Squares (OLS) regression model is because OLS Regression Model is a widely used method for estimating the coefficients in linear regression models, which illustrate the relationship between one or more independent quantitative variables and a dependent variable (whether it's simple or multiple linear regression).[4]

The performance of OLS is often measured using the R-squared statistic. The "least squares" refers to minimizing the sum of squared errors (SSE). Alternative methods to OLS include Maximum Likelihood Estimation and the Generalized Method of Moments.[6]

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         sleep_disorder   R-squared:                       0.562
Model:                            OLS   Adj. R-squared:                  0.546
Method:                 Least Squares   F-statistic:                     35.49
Date:                Fri, 20 Sep 2024   Prob (F-statistic):           1.39e-56
Time:                        09:29:58   Log-Likelihood:                -211.73
No. Observations:                 374   AIC:                             451.5
Df Residuals:                     360   BIC:                             506.4
Df Model:                          13
Covariance Type:            nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
Intercept                         3.5735      0.907      3.941      0.000       1.790       5.357
C(occupation)[T.Doctor]           0.3423      0.108      3.161      0.002       0.129       0.555
C(occupation)[T.Engineer]         0.3267      0.119      2.752      0.006       0.093       0.560
C(occupation)[T.Lawyer]           0.4673      0.140      3.326      0.001       0.191       0.744
C(occupation)[T.Nurse]            1.4166      0.160      8.832      0.000       1.101       1.732
C(occupation)[T.Salesperson]     -0.2400      0.135     -1.781      0.076      -0.505       0.025
C(occupation)[T.Teacher]         -0.0134      0.132     -0.101      0.919      -0.273       0.247
C(occupation)[T.Unknown]          0.5537      0.182      3.042      0.003       0.196       0.912
C(bmi_category)[T.Obese]          0.7409      0.195      3.801      0.000       0.358       1.124
C(bmi_category)[T.Overweight]    -0.0148      0.137     -0.108      0.914      -0.284       0.254
age                               0.0051      0.005      1.100      0.272      -0.004       0.014
physical_activity_level           0.0029      0.002      1.747      0.081      -0.000       0.006
diastolic                         0.0029      0.031      0.096      0.924      -0.057       0.063
systolic                         -0.0285      0.021     -1.360      0.175      -0.070       0.013
==============================================================================
Omnibus:                       53.681   Durbin-Watson:                   1.464
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              440.131
Skew:                          -0.201   Prob(JB):                     2.67e-96
Kurtosis:                       8.299   Cond. No.                     7.11e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.11e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
                                Coef.  Std.Err.
Intercept                    3.573481  0.906668
C(occupation)[T.Doctor]      0.342291  0.108294
C(occupation)[T.Engineer]    0.326703  0.118726
C(occupation)[T.Lawyer]      0.467268  0.140486
C(occupation)[T.Nurse]       1.416617  0.160391
C(occupation)[T.Salesperson] -0.240030  0.134809
C(occupation)[T.Teacher]     -0.013368  0.132148
C(occupation)[T.Unknown]      0.553684  0.182015
C(bmi_category)[T.Obese]      0.740931  0.194948
C(bmi_category)[T.Overweight] -0.014818  0.136656
age                           0.005096  0.004633
physical_activity_level       0.002910  0.001666
diastolic                     0.002925  0.030619
systolic                     -0.028508  0.020969
```

Figure 18. OLS Regression Result

E.2. Model Result

The provided OLS regression results summarize a model where the dependent variable is "sleep disorder." Here are some key observations:

1. Model Performance:
   - R-squared: 0.562 (= 56.2%) of the variance in the sleep disorder variable is explained by the independent variables in the model.
   - Adj. R-squared: 0.556 (= 55.6%), which is a slight decrease after accounting for the number of predictors, suggesting a moderately good fit.
   - F-statistic: 35.49 with a p-value of 1.39e-56, meaning the model as a whole is statistically significant.

2. Significant Predictors ($p < 0.05$):
  - Occupation:
    - Doctor ($p = 0.002$)
    - Engineer ($p = 0.006$)
    - Lawyer ($p = 0.001$)
    - Nurse ($p = 0.000$)
    - Unknown occupation ($p = 0.003$)
    - Salesperson and Teacher are not significant at $p < 0.05$.
  - BMI Category:
    - Obese ($p = 0.000$)
    - Overweight ($p = 0.914$) is not significant.

3. Other Predictors:
  - Age, Physical Activity Level, Diastolic, and Systolic are not statistically significant, as their p-values are greater than 0.05.

4. Multicollinearity Concerns:
  - The model's condition number is 7.11e+03, which suggests potential multicollinearity issues among the predictors, though it's not necessarily severe. You might want to explore variance inflation factors (VIF) to better assess this.

5. Error Terms:
  - The Omnibus test ($p = 0.000$) and Jarque-Bera test ($p = 0.000$) suggest that the residuals are not normally distributed, which may indicate some issues with model assumptions.
  - The Durbin-Watson statistic (1.464) suggests slight positive autocorrelation in the residuals, but it’s not alarming.

Overall, the OLS Regression model does a reasonably good job of explaining the variation in sleep disorder, with key predictors being occupation type and BMI category. It manages to show a moderate fit, with several independent variables explaining sleep disorders, though not all variables are statistically significant. However, the presence of multicollinearity and potential issues with residual normality might warrant further investigation.

F. Conclusion

F.1. Summary

Overall, the analysis reveals that occupation, BMI category, and sleep duration all have a statistically significant impact on sleep disorders. Chi-square tests confirm significant associations between both BMI category and occupation with the prevalence of sleep disorders in the studied population. The variables and model fitting can be understood as the following:

1. Occupational Influence:
   Certain occupations, such as Doctor, Engineer, Lawyer, Nurse, and Unknown, significantly predict sleep disorders. Other occupations, like Teacher and Salesperson, do not show a meaningful association with sleep disorders.
2. BMI and Sleep Disorders:
   Obesity is strongly correlated with sleep disorders, while being overweight shows a weaker, non-significant relationship ($p = 0.090$). This suggests that obesity has a more direct impact on sleep health than being moderately overweight.
3. Other Factors:
   Variables like age, physical activity, and blood pressure (diastolic and systolic) do not significantly predict sleep disorders, indicating they may play a lesser role or be overshadowed by stronger factors like occupation and BMI.
4. Model Fit and Multicollinearity:
   The model explains 56.2% of the variance, indicating a moderate fit. However, a high condition number suggests potential multicollinearity, meaning some predictors may be highly correlated, which could affect the reliability of the coefficient estimates.

F.2. Recommendations

1. Recommendations for Business:
   - Prioritize employee health programs, particularly those addressing obesity and sleep health, as these have a strong correlation with sleep disorders.
   - Consider offering specialized wellness initiatives for professions most affected, such as Doctors, Engineers, and Nurses.
   - Implement flexible working hours or stress management programs.

2. Recommendations for Experiment:
   - Further investigate the relationship between BMI and sleep disorders, focusing on the differences between overweight and obese categories.
   - Explore additional variables that may influence sleep disorders but weren't significant in this study, such as stress levels or dietary habits.
   - Address multicollinearity in future models by either refining or reducing correlated predictors to improve accuracy in estimating effects.

G. References

1. Aggarwal B., Makarem N., Shah R., Emin M., Wei Y., St-Onge M. P., Jelic S., (2018). Effects of Inadequate Sleep on Blood Pressure and Endothelial Inflammation in Women: Findings From the American Heart Association Go Red for Women Strategically Focused Research Network. Journal of the American Heart Association, Volume 7, Number 12. https://doi.org/10.1161/JAHA.118.008590

2. Alexander, H., Illowsky, B., & Dean, S. (2017). Introductory business statistics. Openstax.

3. Amelia VL, Jen HJ, Lee TY, Chang LF, Chung MH. Comparison of the Associations between Self-Reported Sleep Quality and Sleep Duration Concerning the Risk of Depression: A Nationwide Population-Based Study in Indonesia. Int J Environ Res Public Health. 2022 Nov 1;19(21):14273. doi: 10.3390/ijerph192114273. PMID: 36361153; PMCID: PMC9657645.

4. Faster Capital (2024). Building a Logistic Regression Model. Retrieved August 20, 2024, from https://fastercapital.com/topics/building-a-logistic-regression-model.html/6

5. Geeks For Geeks (Jul 18, 2024). Understanding Feature Importance in Logistic Regression Models. Retrieved August 20, 2024, from https://www.geeksforgeeks.org/understanding-feature-importance-in-logistic-regression-models/

6. Ghozali, I. (2016) Aplikasi Analisis Multivariate Dengan Program IBM SPSS 23. Edisi 8. Semarang: Badan Penerbit Universitas Diponegoro.

- END OF REPORT -