UNSUPERVISED MACHINE LEARNING MODELS TO DETECT ROUTINE PATTERNS THAT INFLUENCE SLEEP QUALITY

Vania Cabral de Oliveira

(10607174)

Dissertation submitted in partial fulfilment of the requirements for the degree of
Master of Science in Data Analytics
at Dublin Business School

Supervisor: Ahmed Makki

Word Count: 10.937

20th May 2024

DECLARATION

I affirm that this dissertation, which I have presented to Dublin Business School to fulfil the requirements for the Master of Science in Data Analytics, is the culmination of my independent research efforts unless otherwise specified and duly acknowledged through references. Additionally, I confirm that this work has not been submitted for any other academic degree.

Signed: Vania Oliveira

Student Number: 10607174

Date: 20th May, 2024

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my partner, Kevin Doherty, for his unwavering support, encouragement, and patience throughout the year and a half of my studies. His constant presence and belief in my abilities have been invaluable to me.

I am also deeply grateful for the support of my family, who, despite living in another country, have always stood by me and cheered for my success. Their love and encouragement have been a source of strength and motivation during this journey.

I extend my sincere appreciation to my classmates, whose camaraderie and shared knowledge have enriched my learning experience. Their insights and collaboration have played a significant role in shaping my understanding of data analytics.

Special thanks are due to my supervisor, Ahmed Makki, whose expertise and guidance have been instrumental in the development of this project. His invaluable advice and constructive feedback have not only refined my research but also inspired me to strive for excellence.

Thank you all for being part of this incredible journey.

Table of Contents

ABSTRACT

Understanding patterns and routines that contribute to quality sleep remains a critical challenge in sleep research. Traditional methods often fail to uncover complex relationships in large datasets, highlighting the need for advanced analytical techniques. Quality sleep is vital for overall health and performance. This research employs unsupervised machine learning models to analyse sleep and routine data, offering a cost-effective and natural approach to improving sleep habits. By leveraging K-means and DBSCAN clustering algorithms, alongside PCA and UMAP for dimensionality reduction, this study reveals insightful patterns in sleep behaviours. The models successfully identified distinct clusters, providing valuable insights for personalized lifestyle recommendations. Integrating cutting-edge machine learning techniques with sleep research, this study aligns with and extends the current state of the art, contributing to machine learning applications in health optimization. The findings offer practical applications for developing targeted sleep hygiene programs, though future research should address data gaps on physical activities, meal times, and other factors.

CHAPTER ONE

Introduction

1.1 Background

In recent years, there has been a notable increase in the prevalence of individuals experiencing difficulties falling asleep. A survey conducted in 2020 in the United States reveals that 14.5% of adults face sleep-related challenges (Adjaye-Gbewonyo et al., 2022). This widespread issue is concerning for the general population, as poor sleep can lead to complications such as heart disease, diabetes, obesity, stroke, hypertension, and physical disorders (Shahin et al., 2018), as well as mental health issues (Fernandez-Mendoza & Vgontzas, 2013).

Research by Siegel (2003) highlights that neuronal activity reaches near-maximum levels during specific sleep phases, contradicting the misconception that sleep is merely a dormant state. Medical studies have observed variations in heart temperatures and protective patterns during non-REM sleep, and during REM sleep, brain activity similar to wakefulness has been noted. This has led to some uncertainty among scientists about the precise function of REM sleep in humans. Slyusarenko & Fedorin (2020) categorize sleep into four distinct stages: Rapid Eye Movement (REM), Light, Deep, and Wake. Each stage serves a specific physiological function, and optimal awakening is recommended during the Light or Wake stages to mitigate sleep inertia, which typically lasts for 15 to 60 minutes following awakening from the REM or Deep stages. Given the myriad factors influencing sleep quality, an inquiry arises concerning measures to ameliorate these circumstances.

These and countless other studies have been made possible due to technological advances aimed at detecting, collecting, and analysing sleep data. Technology has played a pivotal role in broadening the horizons of research, data collection, and monitoring of sleep-related metrics. The exploration of alternative techniques, such as ballistocardiography (BCG), dates back to the late 1930s, underscoring the enduring interest of researchers, scientists, and medical professionals in understanding sleep patterns. The evolution of methodologies over time attests to a commitment to advancing our understanding of sleep physiology (Majoe et al., 2010; Paalasmaa et al., 2012).

The growing interest of the general population in sleep-related information is evident in the escalating demand for wearable devices. Recently, there has been a surge in the adoption of devices such as BodyMedia FIT, SleepTracker, and Fitbit, which use wrist-based accelerometry to capture sleep-related data, providing comprehensive insights into the sleep cycle. The data collected is presented to users through dedicated applications on smartphones, computers, or websites (Paalasmaa et al., 2012). This trend reflects a societal inclination towards leveraging technological advancements for personalized sleep monitoring and analysis. The increasing public interest has spurred the development of various machine learning models aimed at exploring sleep patterns, including predicting sleep strategies (Ablao et al., 2021), forecasting sleep and wake times (Khademi et al., 2018), and assessing the likelihood of chronic insomnia (Islam et al., 2020).

Given the multiplicity of internal and external factors influencing sleep quality and the variety of machine learning models analysing sleep patterns, this research aims to address the gap in unsupervised machine learning models for identifying sleep patterns. It focuses on comparing and analysing clusters based on the similarities of sleep and routine data from the dataset.

1.2 Research Problem and Justification

Numerous medical and therapeutic interventions target sleep disorders.(Pattyn et al., n.d.) highlights that although natural light impacts sleep conditions, the significant influence of individuals' daily routines on sleep modulation must be recognized. Additionally, age is another relevant factor. Cudney (2022) conducted a systematic review elucidating the correlation between self-perceived sleep quality and physiologically recorded sleep, emphasizing its fundamental role in accurately identifying sleep disorders. Cudney’s findings stress age as a crucial variable in assessing sleep quality, noting that older adults tend to prioritize their perception of sleep quality over polysomnography (PSG) measurements, which significantly impacts their overall quality of life. Recognizing the importance of sleep-related data and recent technological advances in data collection and storage, there is an unexplored opportunity to use unsupervised machine learning algorithms to group sleep and routine data into clusters. This approach can identify patterns that could serve as recommendations for individuals willing to make small changes in their lifestyle to improve sleep quality.

1.3 Research Question

The guiding objective of this project is to evaluate unsupervised clustering models regarding their effectiveness in efficiently grouping individuals to identify sleep patterns among these individuals. To achieve this, we will evaluate several aspects:

1. Compare dimension reduction techniques to determine which one is most appropriate for this study.
2. Use K-means and DBSCAN to label clusters and analyse the results delivered by each model.
3. Analyse how the findings of this project can be effectively implemented in the context of sleep recommendations.

The study will also address the following questions:

1. Which unsupervised machine learning model performs best in this study?
2. Which features contribute most significantly to data clustering?

1.4 Significance of the Study

The significance of this study lies in utilizing techniques for analysing and interpreting data within a scientific and medical context. By combining machine learning models with the aim of promoting advances in the medical field, this research advocates for an intervention strategy that bypasses pharmacological approaches. Instead, it focuses on modifications to usual practices and routines, guided by the systematic analysis of population data. This paradigm shift not only presents potential cost advantages over pharmaceutical interventions but also aligns with a preventative healthcare approach, promoting holistic well-being. Consequently, this methodology offers a more economical, conscious, and convenient alternative to certain current treatment modalities.

1.5 Structure of this Report

This report is based on the previously prepared research proposal, which identified gaps in existing studies and highlighted the research questions explored in this study. The chapters contain crucial information for the current study and clearly present the ideas addressed and discussed, providing the reader with a comprehensive understanding of the problem, research, development, and results. The chapters are divided as follows:

- **Chapter 1: Introduction and Problem Identification**

  The opening chapter provides a brief summary of the physiological functions and phases of sleep, emphasizing the importance of the processes that occur during sleep. It also discusses the use of machine learning in sleep monitoring, identifies gaps in this area, raises relevant questions, and suggests potential solutions.

- **Chapter 2: Literature Review**

  This chapter reviews previous studies that evaluated various aspects influencing sleep quality to understand the importance of the features in the dataset. It also explores different types of treatments for sleep disorders available in the market, ensuring that this study is novel and has not been conducted under the same circumstances before.

- **Chapter 3: Methodology and Implementation**

  This chapter presents and discusses the methodology and implementation of the study. It includes the project plans, codes, methodological approaches, interpretations of graphs, details about data processing and the first models results.

- **Chapter 4: Results and Critical Analysis**

  This chapter presents the proposed solution of the study. It includes the final results, analysis and comparison of the chosen models, and answers the research questions previously raised.

- **Chapter 5: Discussion and Conclusion**

  The final chapter reflects on the results obtained, insights generated, and conclusions about the effectiveness of the models for the scientific and medical fields. It also offers suggestions for further studies and highlights the difficulties encountered throughout the research.

CHAPTER TWO

Literature Review

2.1 Preamble

Several studies conducted in recent years have shown that sleep disorders are directly related to various diseases, including depression, cardiovascular diseases, cancer, hypertension, gastroesophageal reflux, and stress, among others (Fernandez-Mendoza & Vgontzas, 2013; Sejbuk et al., 2022; Shahin et al., 2018). Most of these studies use the Pittsburgh Sleep Quality Index (PSQI) as a measure of the quality, quantity, and latency of sleep (Manzar et al., 2018). The increasing prevalence of conditions such as insomnia and sleep apnea has raised societal concerns. A systematic review analysing data from the UK revealed that about 16-21% of participants exhibited symptoms of insomnia (Ohayon & Reynolds, 2009). Similarly, the Korean National Health Insurance Service reported that the number of patients receiving medical treatment for sleep disorders increased from 358,000 in 2010 to 414,000 in 2014 (Cooper & Relton, 2010). This indicates a significant rise in the number of individuals experiencing difficulties in achieving quality sleep. There is much speculation about treatments that can aid in addressing sleep disorders, ranging from medical interventions, including medications, to more natural solutions such as homeopathy (Cooper & Relton, 2010) and acupuncture (Cristina Novak et al., 2019). To better understand how these treatments work and to elucidate the features considered in our dataset, we will explore the relationships between sleep and various factors that can influence sleep quality.

2.2 Sleep and Exercise

An intriguing association between sleep and exercise is explored in the document by Staples (2015). Contrary to common belief, the quantity of exercise does not necessarily equate to an improved night's sleep. However, a hypothesis posits that consistent engagement in physical activity may contribute to enhancing the overall quality of sleep. It is worth noting that the perception of a limited connection between sleep and exercise is still prevalent. Nevertheless, an interesting perspective emerges, suggesting that a restful night's sleep can potentially serve as a motivating factor for individuals to engage in exercise on the subsequent day. Furthermore, findings from another study, as presented by Park (2023) assert that individuals who engage in regular exercise tend to experience higher-quality sleep and dedicate more time to effective sleep, manifested by a longer duration spent in bed. Conversely, those who do not engage in exercise exhibit shorter sleep durations. This underscores the multifaceted interplay between physical activity and sleep patterns. The research by Dolezal (2017) highlights a significant observation: approximately one-third of the American adult population does not meet the minimum recommended hours of sleep (set at seven hours per night). This same survey also provides data indicating a persistent struggle within the American population to sustain daily physical activities. Are these two points related? The results of this research suggest a potential correlation, wherein one factor may exert influence over the other.

2.3 Sleep and Stimulants

The consumption of coffee as a means to enhance alertness and vigilance has become increasingly prevalent among individuals striving for improved performance, as noted by Slyusarenko & Fedorin (2020). This practice is underpinned by the stimulant effects of caffeine, which accelerates brain activity. However, it is noteworthy that individuals can develop a tolerance to the stimulating effects of caffeine over time, leading to a gradual attenuation of its efficacy. In contrast, research indicates that the ingestion of approximately 200mg of caffeine, particularly in the early evening, can impede the release of melatonin and adenosine, thereby potentially delaying the onset of sleep by up to 40 minutes (Reichert et al., 2022). Consequently, it is prudent to infer that avoiding the consumption of caffeinated substances within specific timeframes is advisable for optimal sleep hygiene. In light of research findings, it is evident that the prolonged use of sleeping pills, as opposed to providing beneficial effects, can be detrimental and potentially contribute to a reduction in lifespan, as discussed by Siegel (2003). The work presented in the study by Kripke (2013) emphasizes that managing the consumption of hypnotics has been associated with a reported decrease in mortality linked to insomnia disorder. This suggests that individuals who utilize sleeping medications do not experience a significant alteration in their lifespan attributable to these substances. Despite these findings, it is noteworthy that both sleeping pills and other stimulants are often resorted to as alternatives for promoting wakefulness or artificially inducing sleep. In the subsequent discussion, we will scrutinize the intricate relationship between sleep and routine.

2.4 Sleep and Routine

Sleep is an integral component of everyone's routine, often recognized as a fundamental activity for general well-being; paradoxically, it continues to be undervalued. Studies advocating the use of mobile applications to provide timely alerts and guidance aimed at improving sleep quality have been met with skepticism by the public. Notably, individuals, despite expressing concern regarding their sleep patterns, demonstrate a propensity to ignore recommendations that could potentially improve their sleep quality, as evidenced in a study by Sano (2015). Our daily routine is significantly influenced by a crucial factor that has received special attention in recent years: the Circadian Cycle. Researchers claim that our physiological functions are intrinsically synchronized with the natural progression of daylight. This synchronization is particularly evident in the regulation of hormonal release, such as melatonin, which plays a fundamental role in improving sleep quality and is predominantly reserved for the nighttime period (Dawson & Encel, 1993; Roenneberg et al., 2022; Walch et al., 2016). Tests conducted in Australia with the participation of 33 people divided into two groups, wherein one group received melatonin supplementation while the other group received a placebo, showed a specific drop in PSQI results (Chan & Lo, 2022), suggesting that melatonin could be a viable supplementation option for individuals with sleep problems.

Besides melatonin, other supplements such as magnesium, amino acids, and vitamin D were tested in different studies analysed by Chan & Lo (2022), all showing positive contributions to participants' sleep processes. According to Sletten (2018), the benefits of melatonin can be enhanced when aligned with sleep-wake behavioural programming. In their tests, the group that received melatonin treatment along with sleep-wake behavioural programming achieved a 52.8% improvement in sleep delay compared to the placebo (24.0%). Later, we will explore light treatments involving the production of melatonin to control the circadian cycle. Based on this understanding, our research will now delve deeper into the relationship between sleep and age since melatonin levels are produced in different quantities according to individuals' age, thus making the age factor relevant in our research.

2.5 Sleep and Age

Sleep patterns and duration exhibit variability contingent upon the age group and individual needs. Newborns necessitate a sleep duration ranging from 13 to 16 hours to substantiate optimal physical and cognitive developmental outcomes. An intriguing observation lies in the divergence of sleep duration between formula-fed and breast-fed infants, with formula-fed babies exhibiting an extended sleep duration of approximately 2 hours. This phenomenon is attributed to the necessity for increased resting intervals to facilitate the absorption of the comprehensive nutrient content inherent in formula-feeding (Jones & Brennan, 2010). In the case of adolescents, ensuring a sufficient duration of sleep is imperative. Nevertheless, empirical research indicates that a substantial 45% of American teenagers experience a nightly sleep duration of fewer than eight hours on school nights. This prevalence is attributed, in part, to the pervasive use of screens, particularly smartphones, which disrupts the initiation of the sleep onset process (Xu et al., 2019). In the context of adults, the prevailing recommendation is approximately 7 hours of nightly sleep (Adjaye-Gbewonyo et al., 2022). However, a considerable number of adults encounter challenges in adhering to this guideline due to the demands of a strenuous daily regimen and the accrual of multiple activities within a compressed timeframe. Notably, sleep disorders such as sleep apnea and insomnia are frequently cited by adults as principal contributors to compromised sleep quality. In advanced age, alterations in both the quantity and quality of sleep become apparent. It is crucial to underscore that the aging process does not inherently dictate an increased or decreased requirement for hours of sleep. Rather, the aging trajectory introduces shifts in sleep patterns, influencing both the manner and efficacy of sleep, while concurrently adapting to the individual needs of each elderly person (Kudale et al., 2023). Waldhauser (1998) suggests a correlation between age and melatonin production rates, which vary throughout our lives, starting with high levels of production and declining over the years. This decline is justified by bodily changes, wherein the size of the human body grows approximately 500 to 800% from childhood to adolescence, in addition to the increase in hormone production, concluding that due to the decrease in the natural production of melatonin, the quality of sleep in the elderly also decreases. However, research by Espiritu (2008) states the opposite. Although Espiritu comments on sleep latency, it is known that melatonin is the hormone responsible for inducing sleep; therefore, sleep latency and melatonin are directly linked. In this research, it was detected that sleep latency patterns are not linear as expected but rather triphasic, increasing until the age of 30, remaining stable between 30 and 50 years, and increasing again. In this sense, the elderly supposedly have an easier time falling asleep. However, in the analysis carried out by Waldhauser (1998), groups of elderly people produce less melatonin, and coincidentally, it is the group with the highest incidence of insomnia.

2.6 Sleep Disorder and Treatments

Based on the aspects above, treatments for sleep disorders such as insomnia and sleep apnea were researched. There is a lot of talk about light therapy, which consists of different light stimuli that are supposed to regulate the circadian cycle. In the research by Zhang (2023), which analysed studies involving people of advanced ages, it was found that more than 10 studies were carried out involving phototherapy and the circadian cycle. In this same analysis, the different ways of delivering this light therapy were also explored. For example, exposure to external natural light, internal artificial lights, light boxes positioned at a distance of 1 meter from patients, and simulation of the transition between dawn and dusk. Both in this research and in Najjar (2014) study, changes in the circadian cycle were observed. In Najjar's work, the use of lights, including 17,000k blue light, caused a delay in the circadian cycle, keeping the participants awake for longer even after early dark in the South Pole area, where the study was carried out. On the other hand, Riemann (2017), in his systematic review including papers from 1966 to 2016, reports that phototherapy, as well as exercises, can be useful in the treatment of sleep disorders, but they have low-quality evidence. Studies conducted by Dadgostar (2023), analysing the PSQI score of a group of 32 participants, concluded that resistance exercises combined with aerobic activities did improve sleep quality. Regarding the use of therapeutic pharmaceuticals, Zhang (2023) states that medicines can cause, especially in the elderly population, chemical dependency, dizziness, risk of fractures, cognitive degradation, among others. In a study with medical students between 18 and 25 years old conducted by Segundo (2017), it is also clear that the use of medication is not a good option, since the participants who used medication to induce sleep were the ones with low sleep quality according to calculations based on the PSQI. For Dadgostar (2023) and Ell (2023), the gateway to sleep disorder treatments is Cognitive Behavioural Therapy (CBT). However, this method involves sleep hygiene, relaxation exercises, cognitive therapy, and behavioural change, making it expensive and difficult to access for certain people (Riemann et al., 2017). For some more specific sleep disorders, such as Obstructive Sleep Apnea (OSA), which consists of breathing difficulty that impairs sleep quality, we have a study that evaluated respiratory devices (Knappe & Sonnesen, 2018). Although they are efficient, devices such as CPAP may be unfeasible for certain patients due to their cost and design. We also have complementary and alternative medicine that includes acupuncture treatments, aromatherapy, foot reflexology, music therapy, yoga, among others. However, none of these have sufficient evidence to prove their effectiveness against sleep disorders (Riemann et al., 2017).

CHAPTER 3

Methodology and Implementation

3.1 Research Methodology

This study aims to discern patterns within the routines of individuals who experience restful sleep and subsequently categorize them based on these discerned patterns. This methodology facilitates the identification of individuals characterized by high-quality sleep, with a nuanced analysis of their behavioural attributes serving as a basis for generating tailored recommendations for those facing challenges in achieving adequate sleep. In essence, our proposal advocates for an intervention strategy that circumvents pharmacological approaches, opting instead for modifications in habitual practices and routines, guided by the systematic analysis of population-wide data. This paradigm shift not only has potential cost advantages over pharmaceutical interventions but also aligns with a paradigm of preventive healthcare, fostering holistic well-being. Consequently, this methodology holds promise as a more economical, health-conscious, and convenient alternative to certain current treatment modalities.

The research was conducted based on the public dataset "Sleep Health and Lifestyle dataset." It contained 374 rows and 13 columns containing variables related to sleep and daily habits. As observed during the literature review, this dataset was chosen because it contains variables that, according to the reviewed studies, are highly valuable for the project's purpose. The following is a description of these variables.

Person ID: An identifier for each individual

Gender: Male/Female

Age: Age of the person in years

Occupation: Occupation or profession of each person

Sleep Duration (hours): The number of hours the person sleeps per day

Quality of Sleep (1 to 10): A subjective rating of the quality of sleep

Physical Activity Level (minutes/day): The number of minutes the person engages in physical activity daily.

Stress Level (1 to 10): A subjective rating of the stress level experienced by the person

BMI Category: The BMI category of the person (e.g., Underweight, Normal, Overweight).

Blood Pressure (systolic/diastolic): The blood pressure measurement of the person, indicated as systolic pressure over diastolic pressure.

Heart Rate (bpm): The resting heart rate of the person in beats per minute.

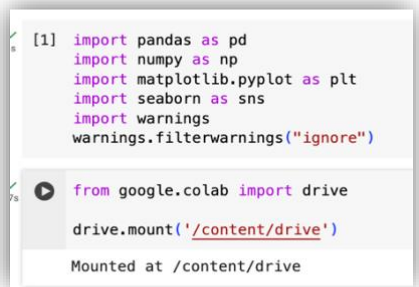
Daily Steps: Number of steps the person takes per day.

Sleep Disorder: The presence or absence of a sleep disorder in the person (None, Insomnia, Sleep Apnea).

This research entails a primary, quantitative, and experimental approach, focusing on the observation of data within the context of lifestyle to discern underlying patterns. The investigation will employ unsupervised clustering models, including K-Means and DBSCAN, alongside certain dimensionality reduction techniques such as Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP). The objective is to evaluate the effectiveness of these models by grouping the data into clusters with different people and their sleep patterns.

3.2 Research Implementation

Python is consistently considered one of the top 10 programming languages (Mukhiya & Ahmed, 2020). Moreover, it is compatible with the Google Colab platform. Hence, Python was chosen for generating the codes for data exploration and analysis. To initiate the analysis, fundamental libraries were imported to access and understand the data. An essential aspect of the project involved enabling access to Google Drive, facilitating the storage of label clusters in a new file at a later stage.



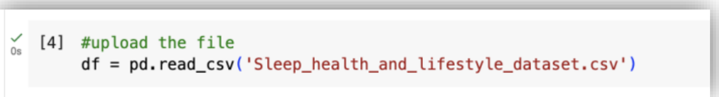
```
[1] import pandas as pd
    import numpy as np
    import matplotlib.pyplot as plt
    import seaborn as sns
    import warnings
    warnings.filterwarnings("ignore")
```

```
from google.colab import drive

drive.mount('/content/drive')
```

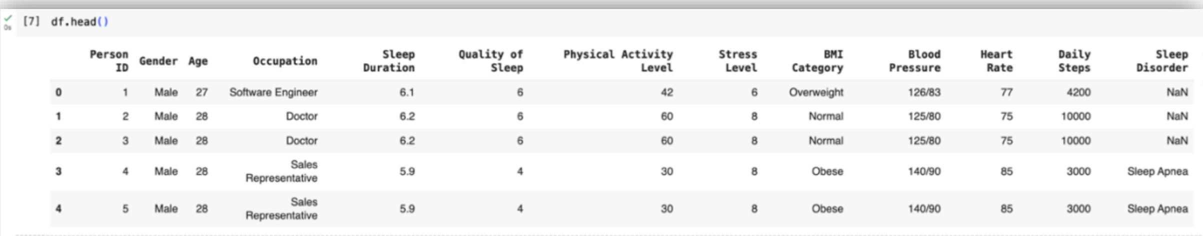
```
Mounted at /content/drive
```

3.2.1 Data Understanding

During this stage, the dataset is assigned to a DataFrame named 'df'. This marks the commencement of data exploration, where simple commands such as **.head()**, **.info()**, and **.shape()** are utilized. These commands provide insight into the type of data under analysis, aiding in determining the subsequent steps for data processing and cleaning (Da Poian et al., 2023).

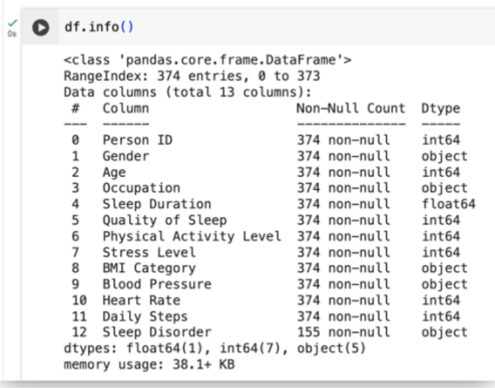


```
[4] #upload the file
    df = pd.read_csv('Sleep_health_and_lifestyle_dataset.csv')
```

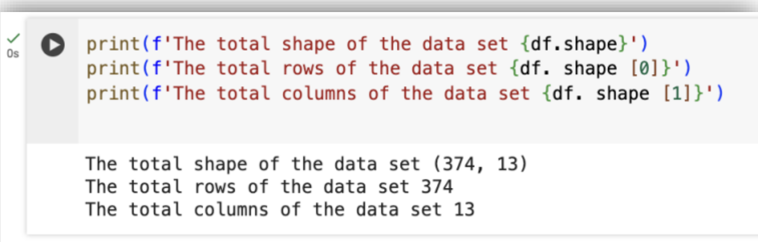


```
[7] df.head()
```

	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	NaN
1	2	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	NaN
2	3	Male	28	Doctor	6.2	6	60	8	Normal	125/80	75	10000	NaN
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



```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 374 entries, 0 to 373
Data columns (total 13 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Person ID                374 non-null    int64
 1   Gender                   374 non-null    object
 2   Age                      374 non-null    int64
 3   Occupation               374 non-null    object
 4   Sleep Duration           374 non-null    float64
 5   Quality of Sleep         374 non-null    int64
 6   Physical Activity Level  374 non-null    int64
 7   Stress Level             374 non-null    int64
 8   BMI Category             374 non-null    object
 9   Blood Pressure           374 non-null    object
 10  Heart Rate               374 non-null    int64
 11  Daily Steps              374 non-null    int64
 12  Sleep Disorder           155 non-null    object
dtypes: float64(1), int64(7), object(5)
memory usage: 38.1+ KB
```



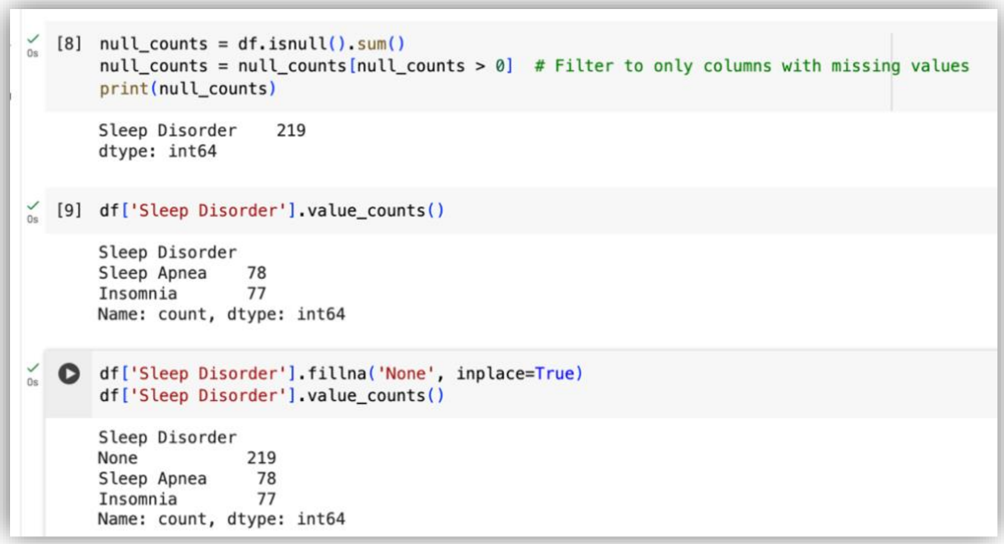
```
print(f'The total shape of the data set {df.shape}')
print(f'The total rows of the data set {df. shape [0]}')
print(f'The total columns of the data set {df. shape [1]}')

The total shape of the data set (374, 13)
The total rows of the data set 374
The total columns of the data set 13
```

3.2.2 Data Cleaning

Data cleaning is a critical step aimed at ensuring the accuracy of results. As emphasized by Chai (2020), this process involves addressing missing, erroneous, or duplicate values within the dataset. Proper cleaning of the data is essential for enhancing the performance of machine learning models, as it promotes data balance and consistency. Specifically, identifying and rectifying issues such as missing values, outliers, duplicate rows, and errors is imperative for generating more reliable results.

3.2.2.1 Missing Values

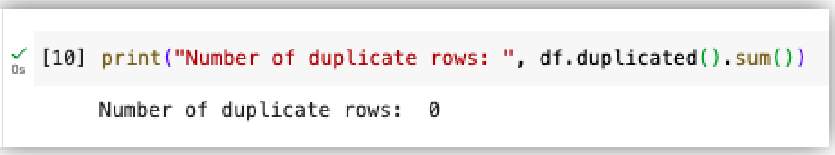
Upon inspecting for missing values, an issue was identified within the Sleep Disorder column of the dataset. Specifically, there are 219 missing values recorded as NaN, corresponding to participants who reported not having sleep disorders. While common approaches to handling missing data include imputation with averages or discarding records, neither option is suitable in this scenario. Instead, we aim to validate these entries as individuals without sleep disorders. To address this, the **.fillna()** command is utilized to transform NaN values into None. By doing so, we ensure that these entries are accurately counted as individuals without sleep disorders.



```
[8]  null_counts = df.isnull().sum()
     null_counts = null_counts[null_counts > 0]  # Filter to only columns with missing values
     print(null_counts)

     Sleep Disorder    219
     dtype: int64

[9]  df['Sleep Disorder'].value_counts()

     Sleep Disorder
     Sleep Apnea    78
     Insomnia       77
     Name: count, dtype: int64

     df['Sleep Disorder'].fillna('None', inplace=True)
     df['Sleep Disorder'].value_counts()

     Sleep Disorder
     None           219
     Sleep Apnea     78
     Insomnia        77
     Name: count, dtype: int64
```

3.2.2.2 Duplicate Rows

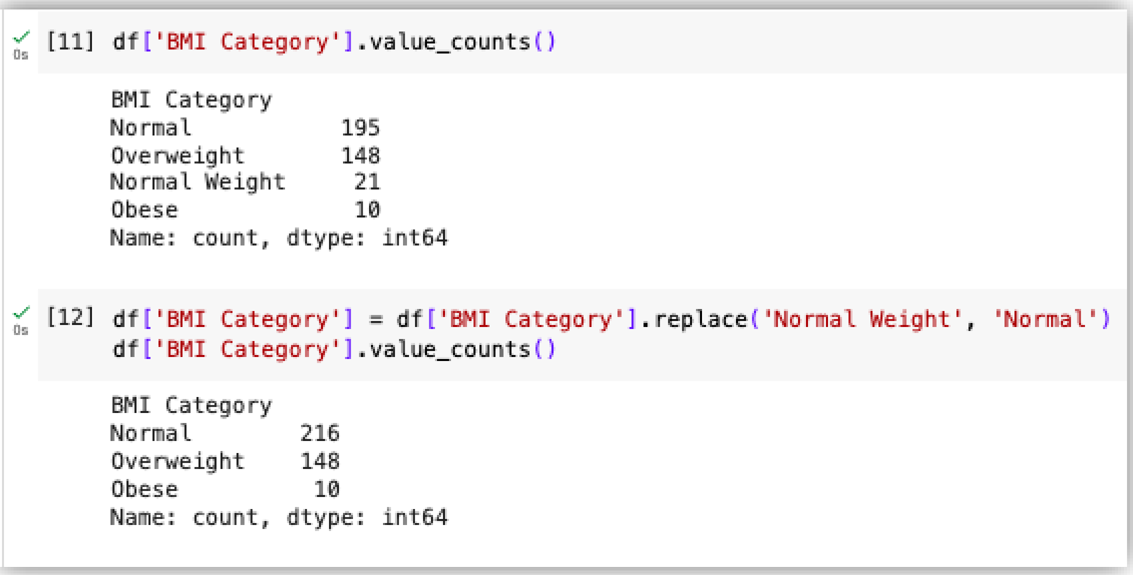
Duplicate rows refer to instances where all values within a specific record are replicated. Such occurrences can disrupt model results by introducing redundancy into the dataset, thereby inflating its dimensionality and potentially skewing analytical outcomes. The following code snippet verifies the absence of duplicate rows in the dataset:



```
[10] print("Number of duplicate rows: ", df.duplicated().sum())

     Number of duplicate rows:  0
```

3.2.2.3 Erroneous Data

Erroneous data encompasses instances where information is misconstrued, either due to discrepancies in the data recording process (e.g., capitalization, extra spaces, accents, or different characters) or when similar information is recorded differently (Mukhiya & Ahmed, 2020). In this dataset, one such issue was identified and addressed. The 'BMI Category' feature contains both "Normal" and "Normal Weight," which can be considered equivalent. To rectify this inconsistency, the **.replace()** function is employed to merge "Normal" and "Normal Weight" into a single category labelled "Normal."



```
[11] df['BMI Category'].value_counts()

     BMI Category
     Normal           195
     Overweight       148
     Normal Weight     21
     Obese             10
     Name: count, dtype: int64

[12] df['BMI Category'] = df['BMI Category'].replace('Normal Weight', 'Normal')
     df['BMI Category'].value_counts()

     BMI Category
     Normal        216
     Overweight    148
     Obese          10
     Name: count, dtype: int64
```

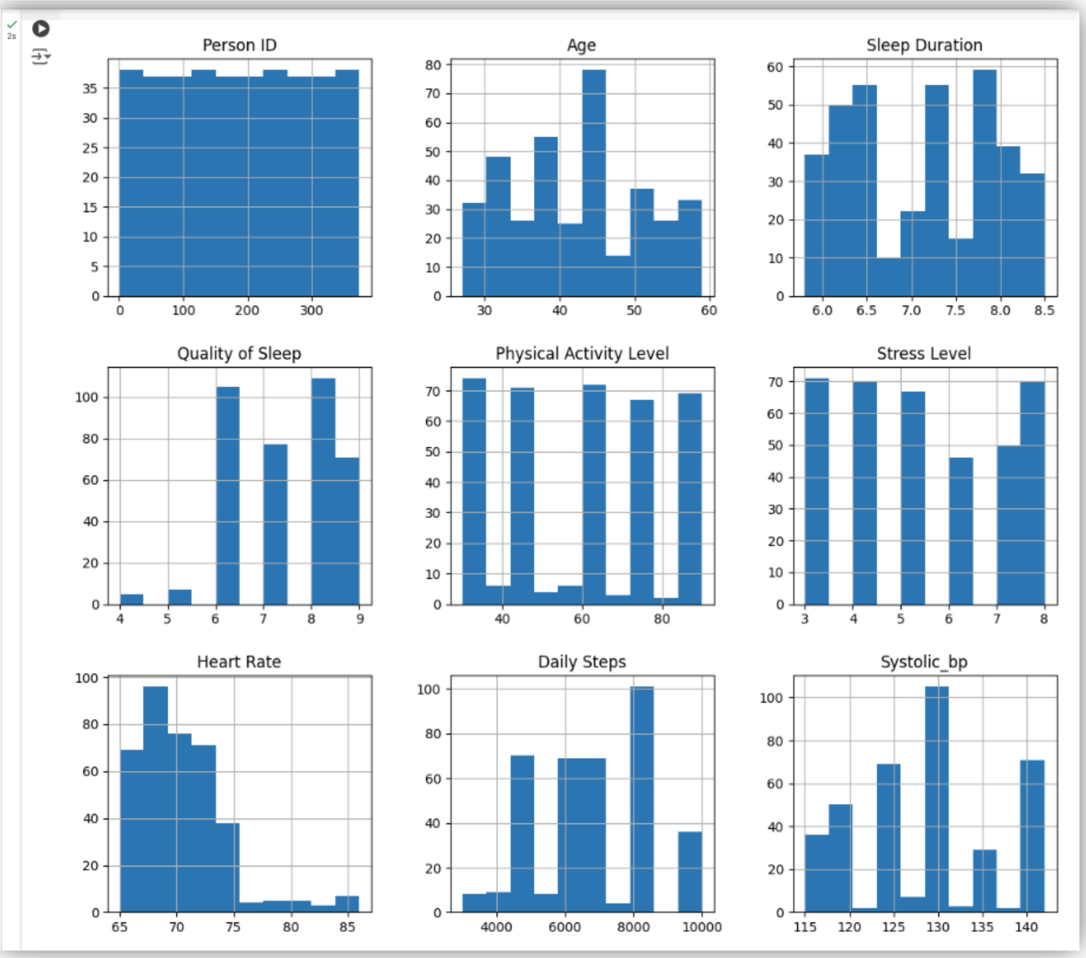
3.2.3 Exploratory Data Analysis (EDA)

3.2.3.1 Data Visualization

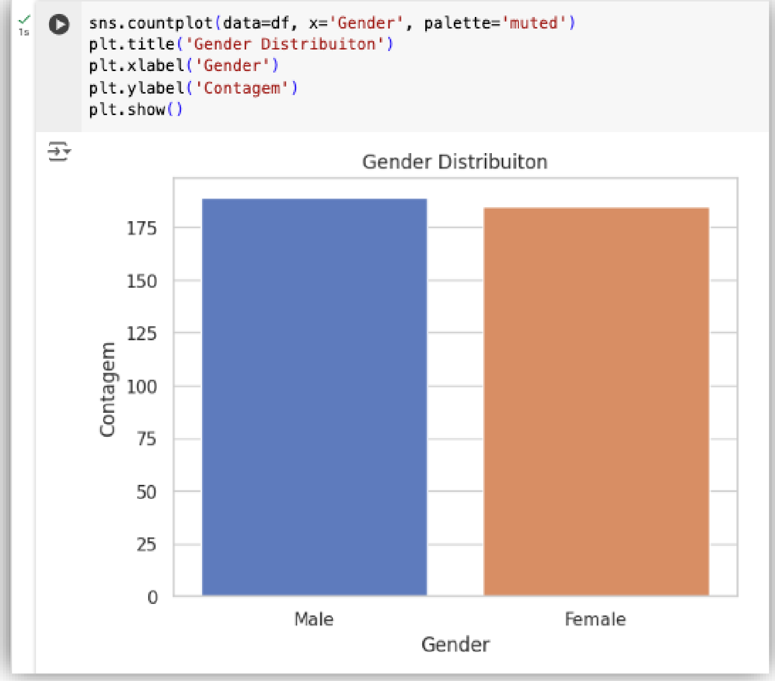
Data visualization is a foundational practice essential for understanding data and gaining a comprehensive overview of the information contained within the dataset. Despite technological advancements that facilitate improved data visualization, human intervention remains crucial for meticulously elaborating and structuring visualizations. This human involvement ensures clarity and establishes a strong connection between information, visualization, and the audience (Midway, 2020). In this section, graphical representations provide enhanced elucidation, revealing the interrelationships among various features within the dataset.

3.2.3.1.1 Univariate Analysis

Univariate analysis is employed to examine the relationship of a single variable in the dataset, aiming to comprehend the distribution of data, characteristics, and patterns for that specific variable. The subsequent visualization provides a comprehensive overview of the numerical variables in the dataset and their distributions. This visualization showcases examples of various distributions, including Uniform distribution (e.g., Person ID), Right-skewed distribution (e.g., Heart Rate), and Multimodal distribution (e.g., Systolic).

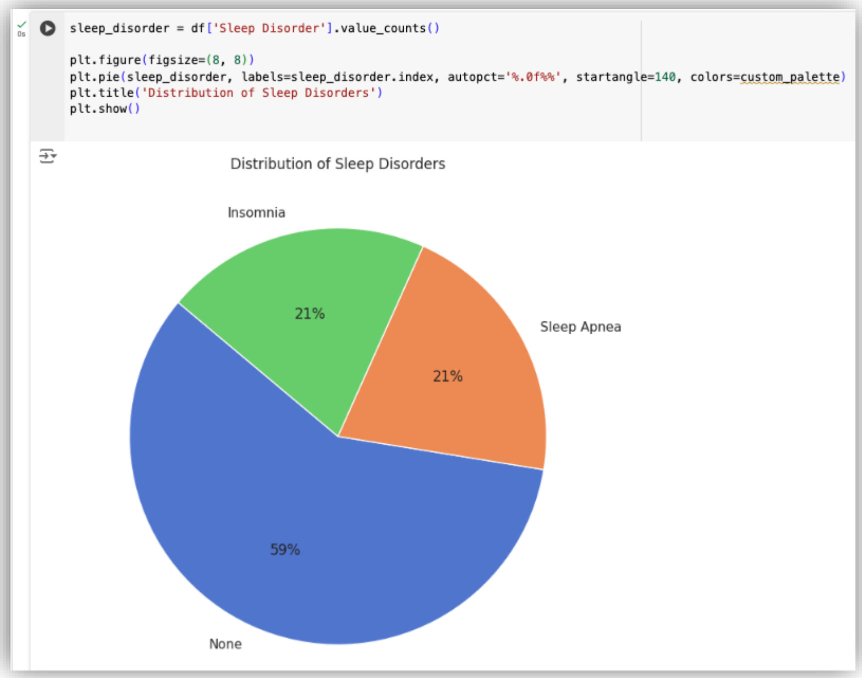


Below is a gender graph illustrating the distribution between females and males. It is evident that the genders are almost equally distributed in terms of quantity.



```
sns.countplot(data=df, x='Gender', palette='muted')
plt.title('Gender Distribuiton')
plt.xlabel('Gender')
plt.ylabel('Contagem')
plt.show()
```

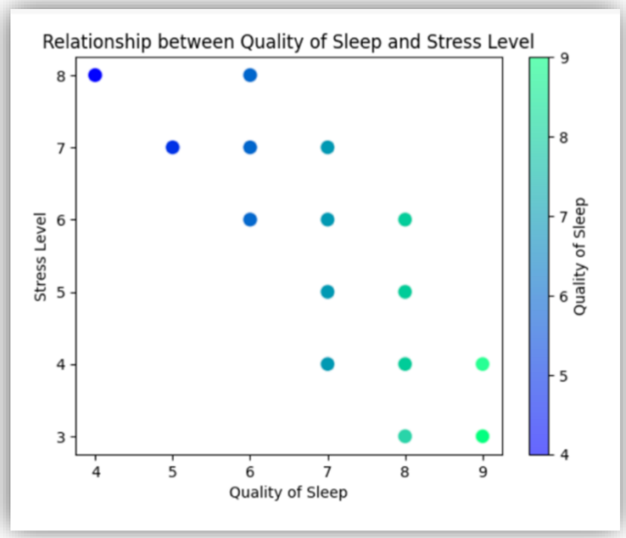
In the pie chart, we visualize the Sleep Disorder feature, and it can be concluded that the percentage of people who have sleep problems is very close to the percentage of people who claim they do not.



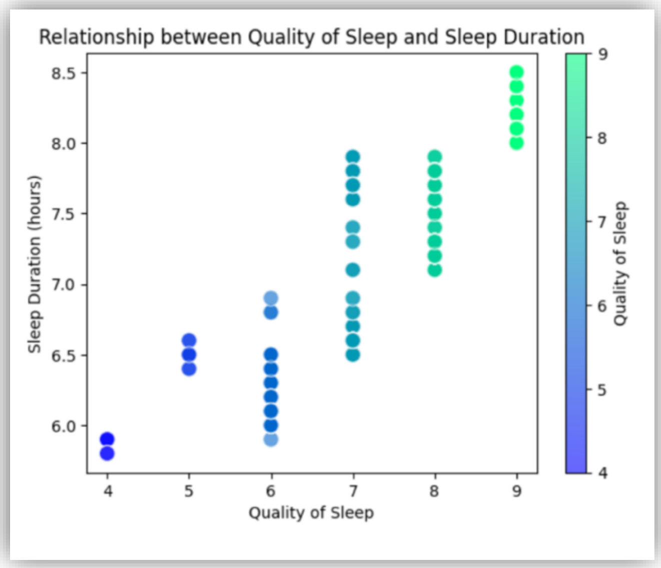
```
sleep_disorder = df['Sleep Disorder'].value_counts()

plt.figure(figsize=(8, 8))
plt.pie(sleep_disorder, labels=sleep_disorder.index, autopct='%.0f%%', startangle=140, colors=custom_palette)
plt.title('Distribution of Sleep Disorders')
plt.show()
```

3.2.3.1.2 Bivariate Analysis

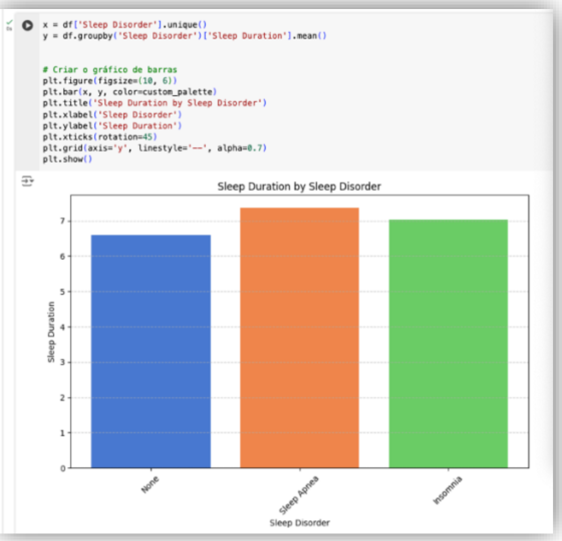
In bivariate analysis, we observe the relationship between two variables, considering correlations, patterns, associations, or trends between them. In the first image, we depict the relationship between Quality of Sleep and Stress Level. It is notable that individuals who rated their sleep quality highest also reported lower levels of stress. Conversely, those who rated their sleep quality lower tended to report higher levels of stress.



In the second image, we observe the relationship between Quality of Sleep and Sleep Duration. It is evident that individuals who rated their sleep quality lowest are those who sleep less than 7 hours per night. Conversely, those who rated their sleep quality highest predominantly sleep more than 7 hours per night.

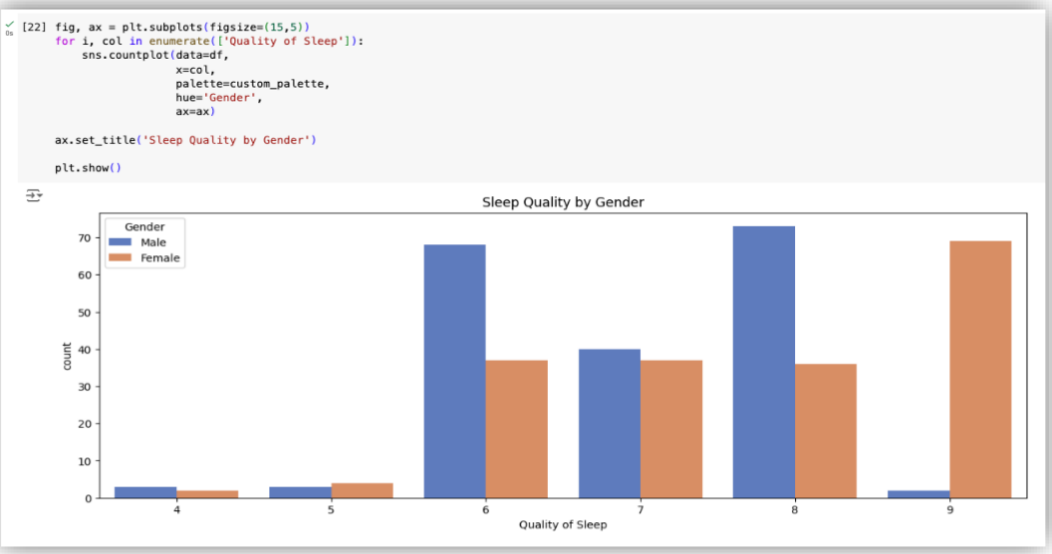


In the third image, we explore the relationship between Sleep Disorder and Sleep Duration. We observe that within the Sleep Apnea category, there are individuals who sleep for more than 7 hours, while within the Insomnia category, there are people who sleep for up to 7 hours. In the category without sleep disorders, individuals sleep a maximum of around 6 hours and 30 minutes. This suggests that even if one sleeps for an acceptable duration, they may still experience sleep disturbances.



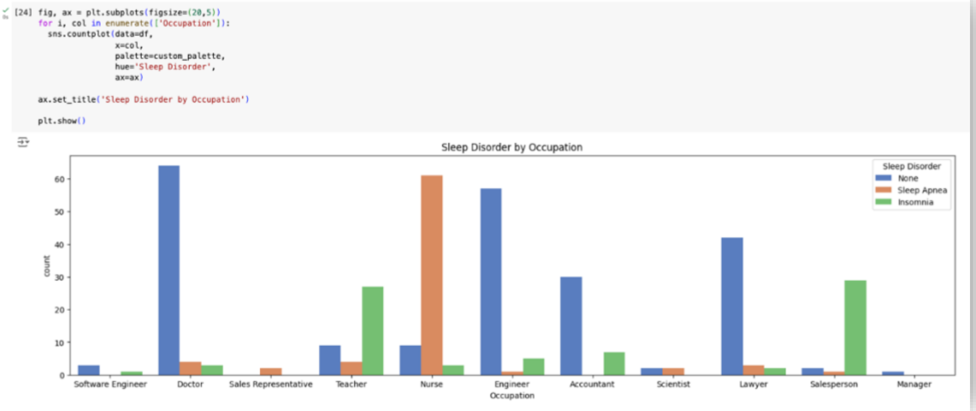
```
x = df['Sleep Disorder'].unique()
y = df.groupby('Sleep Disorder')['Sleep Duration'].mean()


# Criar o gráfico de barras
plt.figure(figsize=(10, 6))
plt.bar(x, y, color=custom_palette)
plt.title('Sleep Duration by Sleep Disorder')
plt.xlabel('Sleep Disorder')
plt.ylabel('Sleep Duration')
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()
```

In the fourth image, we examine Sleep Quality by Gender. It is evident that females report better sleep quality, with a score of 9, compared to males, who score an 8. Additionally, it is observed that both females and males exhibit balanced levels concerning low scores for sleep quality.



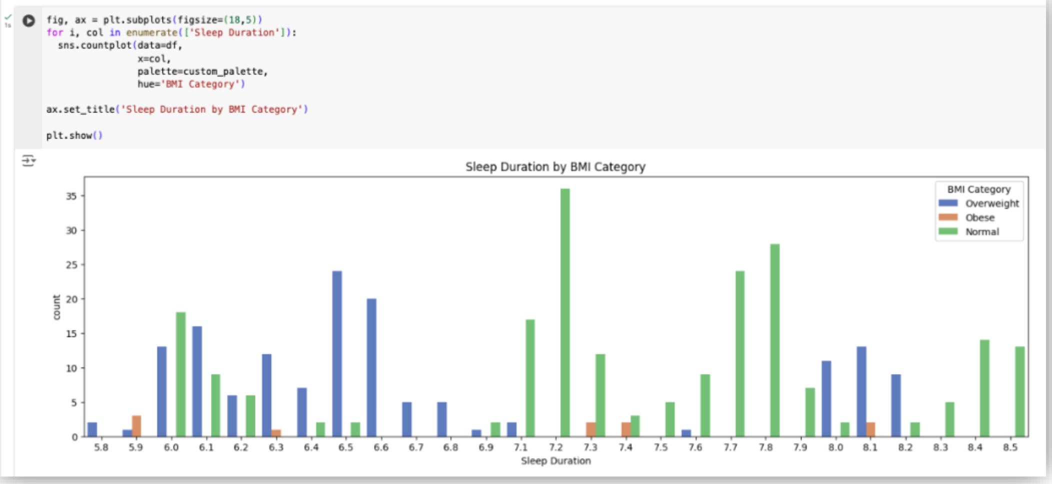
```
[22] fig, ax = plt.subplots(figsize=(15,5))
     for i, col in enumerate(['Quality of Sleep']):
         sns.countplot(data=df,
                       x=col,
                       palette=custom_palette,
                       hue='Gender',
                       ax=ax)

     ax.set_title('Sleep Quality by Gender')

     plt.show()
```

In the following visualization, we depict Sleep Disorder by Occupation. This graph offers valuable insights into the relationship between various professions and sleep disorders. Notably, it is evident that a majority of doctors report not having sleep disorders, whereas a significant portion of nurses experience sleep apnea. Interestingly, salespersons and teachers exhibit the highest rates of insomnia among the professions represented in the dataset. Moreover, additional insights and comparisons can be gleaned, such as the predominance of low rates of sleep apnea among most professions in the dataset, with the notable exception of nurses.



```
[24] fig, ax = plt.subplots(figsize=(20,5))
     for i, col in enumerate(['Occupation']):
       sns.countplot(data=df,
                     x=col,
                     palette=custom_palette,
                     hue='Sleep Disorder',
                     ax=ax)

     ax.set_title('Sleep Disorder by Occupation')

     plt.show()
```

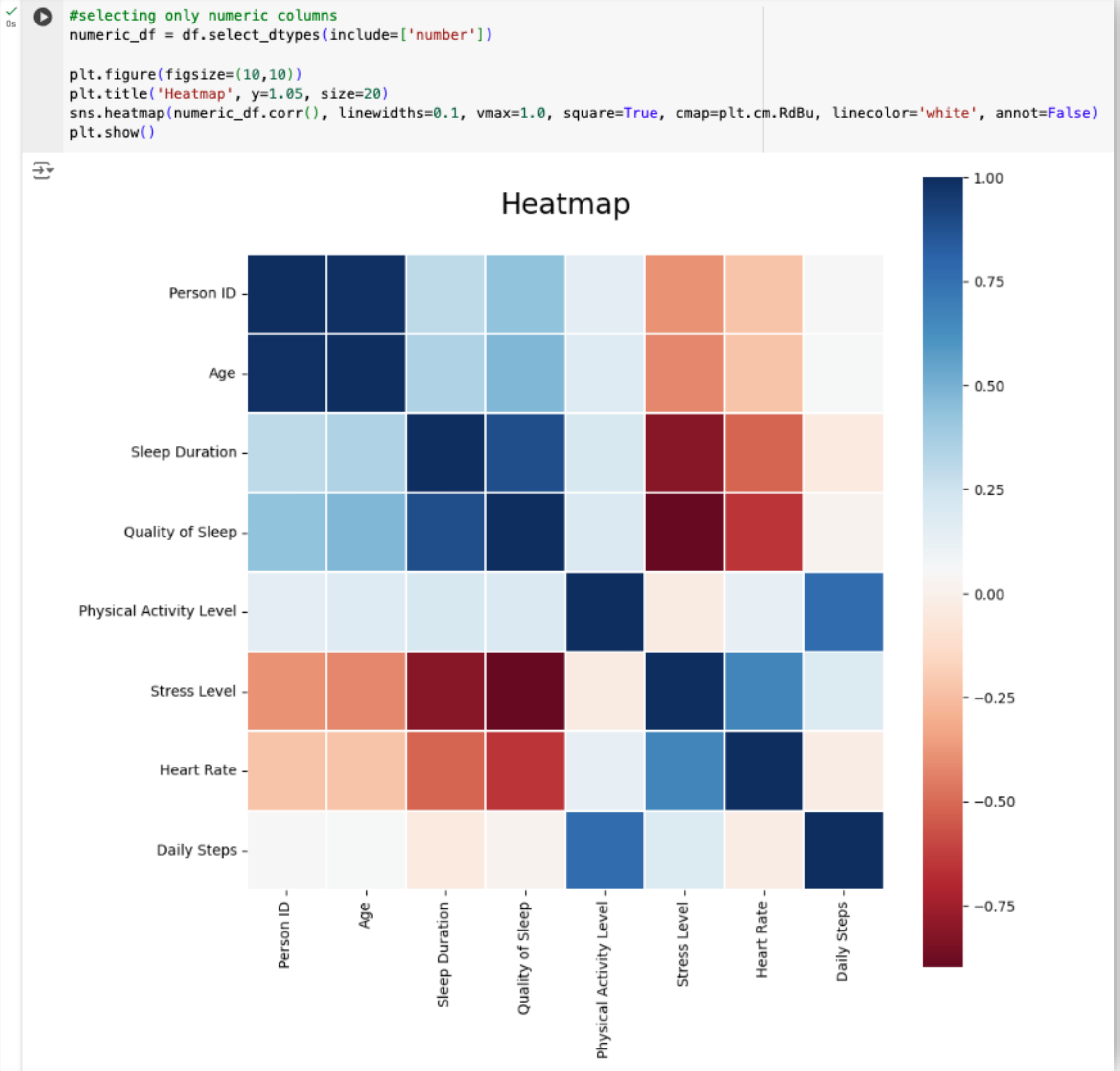
In this figure, we observe the relationship between Sleep Duration and BMI Category. It is evident that individuals with a normal BMI tend to sleep more, averaging between 7 to 8.5 hours per night. Conversely, those with an overweight BMI predominantly sleep fewer hours, typically ranging between 5.8 to 7.1 hours per night. These patterns suggest a potential correlation between being overweight and experiencing a reduction in the duration of sleep.



```
fig, ax = plt.subplots(figsize=(18,5))
for i, col in enumerate(['Sleep Duration']):
  sns.countplot(data=df,
                x=col,
                palette=custom_palette,
                hue='BMI Category')

ax.set_title('Sleep Duration by BMI Category')

plt.show()
```

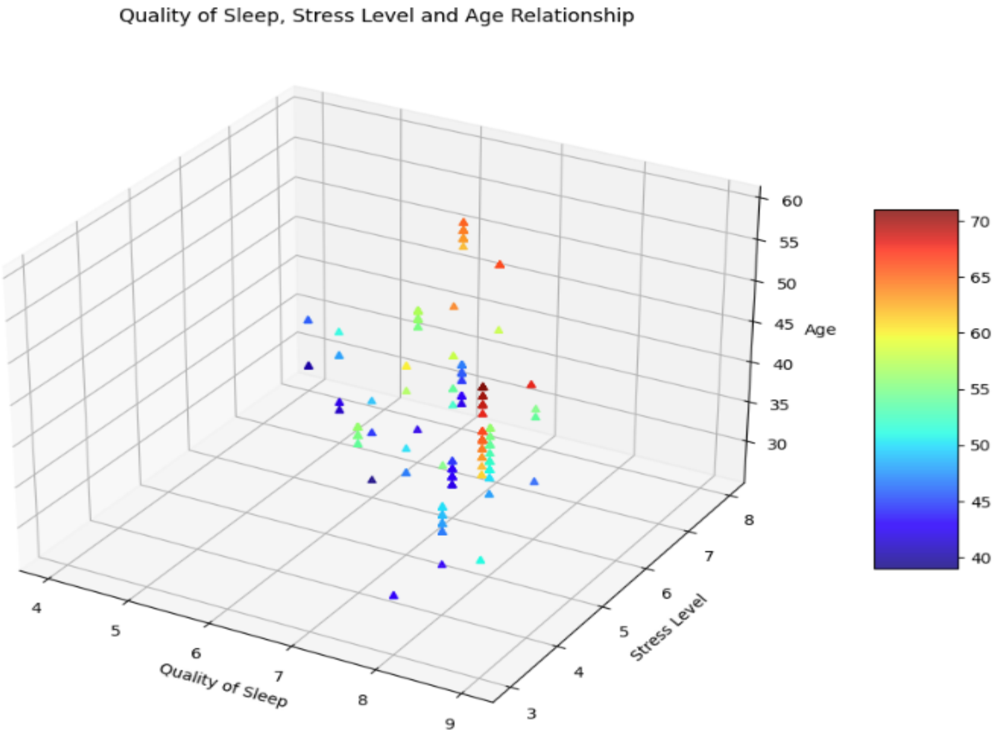
3.2.3.1.3 Multivariate analysis

In multivariate analysis, graphs incorporate three or more variables into a single plot. This visualization type offers a comprehensive insight into the relationships, patterns, and trends among these variables, aiding the reader's understanding of the data. In the initial multivariate visualization, a heatmap illustrates the strength of relationships between numerical variables in the dataset. Quality of Sleep demonstrates a strong correlation with variables like Personal ID, Age, and Sleep Duration. Conversely, Stress Level and Heart Rate exhibit weak correlations, while Physical Activity Level displays a reasonably correlated relationship. Upon closer examination of the Physical Activity Level variable, it reveals mostly reasonable correlations with other variables, with only one weak relationship observed with Stress Level.



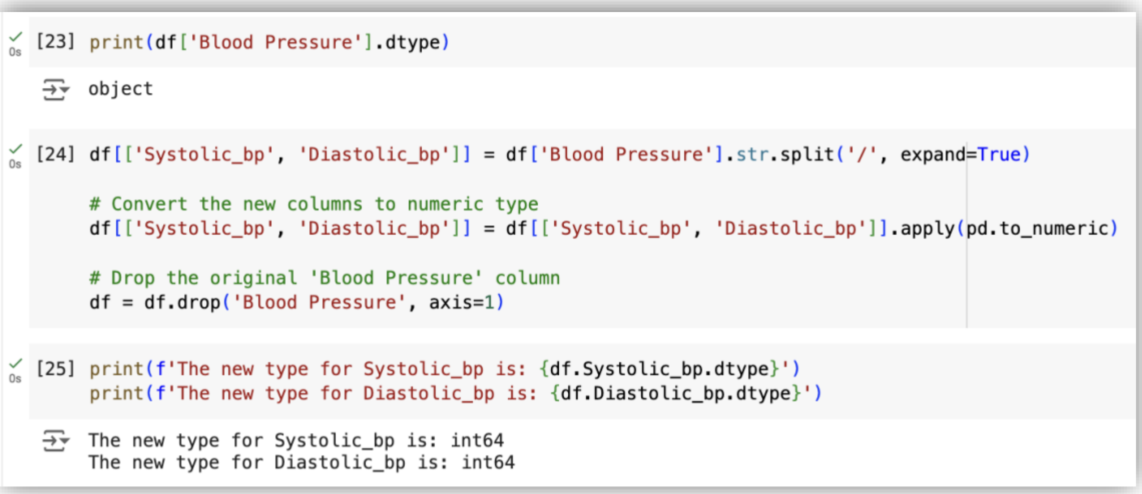
```
#selecting only numeric columns
numeric_df = df.select_dtypes(include=['number'])

plt.figure(figsize=(10,10))
plt.title('Heatmap', y=1.05, size=20)
sns.heatmap(numeric_df.corr(), linewidths=0.1, vmax=1.0, square=True, cmap=plt.cm.RdBu, linecolor='white', annot=False)
plt.show()
```

In the final visualization, we encounter a 3D graph featuring Quality of Sleep, Stress Level, and Age. The graph illustrates a concentration of data points at elevated stress levels (6, 7, and 8). Regarding Age and Quality of Sleep, the data exhibits a more dispersed distribution across the center.



3.2.3.2 Feature Engineering

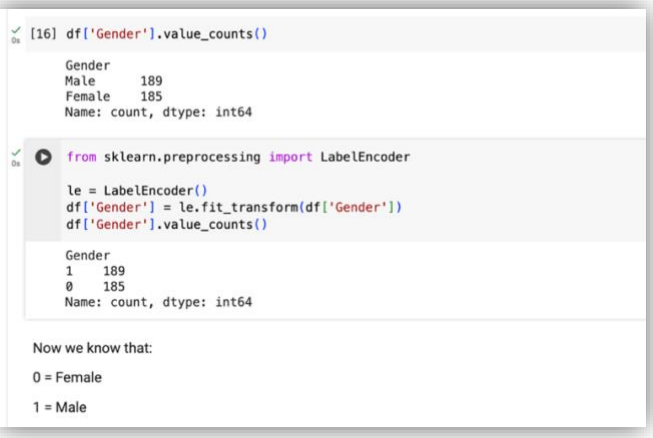
Feature Engineering is fundamental to learning the algorithm. It is the process of creating new features based on knowledge acquired from existing features. This way, it is possible to implement the inputs that the algorithm will receive, obtaining better results (Rawat & Khemchandani, 2017). The first step was to correct the Blood Pressure feature, which was previously identified as a problem due to it containing slash '/'. In this case, all data in this column are considered objects because they have a slash, making it difficult to use this information in generating visualizations and statistical analyses. The solution found was to remove the slash using **.str.split()** and create two new features that receive the Systolic and Diastolic values, excluding the Blood Pressure column.



```
[23] print(df['Blood Pressure'].dtype)

     object

[24] df[['Systolic_bp', 'Diastolic_bp']] = df['Blood Pressure'].str.split('/', expand=True)

     # Convert the new columns to numeric type
     df[['Systolic_bp', 'Diastolic_bp']] = df[['Systolic_bp', 'Diastolic_bp']].apply(pd.to_numeric)

     # Drop the original 'Blood Pressure' column
     df = df.drop('Blood Pressure', axis=1)

[25] print(f'The new type for Systolic_bp is: {df.Systolic_bp.dtype}')
     print(f'The new type for Diastolic_bp is: {df.Diastolic_bp.dtype}')

     The new type for Systolic_bp is: int64
     The new type for Diastolic_bp is: int64
```

3.2.3.3 Data Consistency

Data consistency refers to the reliability and quality of data within a dataset (Ilyas & Chu, 2015). In essence, high-quality data accurately reflect their intended meaning. For example, an age feature should consist solely of integers, not floating-point numbers. Data integrity is also vital, ensuring completeness, freedom from errors, and consistency in format. Maintaining data consistency is crucial for the effective operation of machine learning models. Since the models chosen for this study only accept numeric inputs, categorical data must be transformed into numeric data. To achieve this, we utilized a function from the Sklearn package known as LabelEncoder. This allowed us to assign the value 0 to females and the value 1 to males, thereby encoding categorical variables into a numeric format



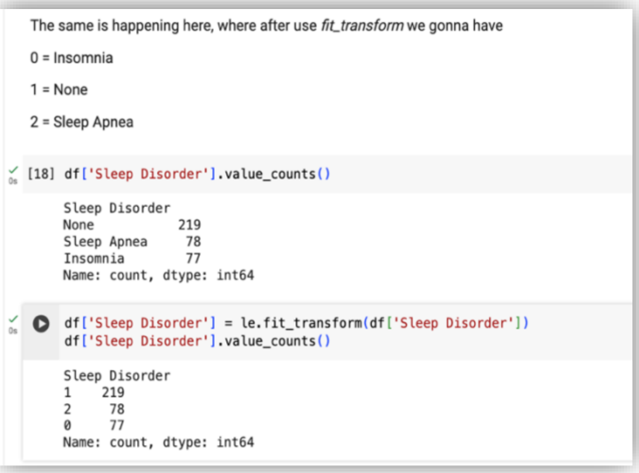
```
[16] df['Gender'].value_counts()

     Gender
     Male      189
     Female    185
     Name: count, dtype: int64

     from sklearn.preprocessing import LabelEncoder

     le = LabelEncoder()
     df['Gender'] = le.fit_transform(df['Gender'])
     df['Gender'].value_counts()

     Gender
     1    189
     0    185
     Name: count, dtype: int64
```

Now we know that:

0 = Female

1 = Male

In this scenario, we opted for a different technique known as **.fit_transform()** to convert the categorical data from the Sleep Disorder column into numeric values. Similarly, we applied the same function to handle the Occupation column, converting professions into numerical representations.



The same is happening here, where after use *fit_transform* we gonna have

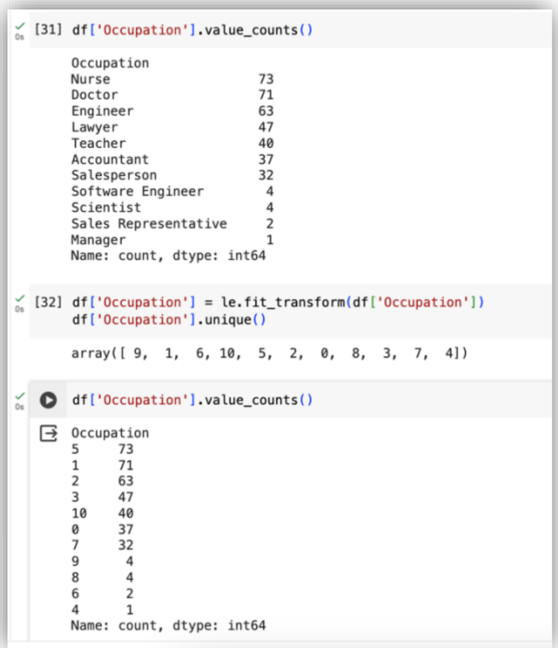
0 = Insomnia

1 = None

2 = Sleep Apnea

```
[18] df['Sleep Disorder'].value_counts()

Sleep Disorder
None           219
Sleep Apnea     78
Insomnia        77
Name: count, dtype: int64
```

```
df['Sleep Disorder'] = le.fit_transform(df['Sleep Disorder'])
df['Sleep Disorder'].value_counts()

Sleep Disorder
1    219
2     78
0     77
Name: count, dtype: int64
```

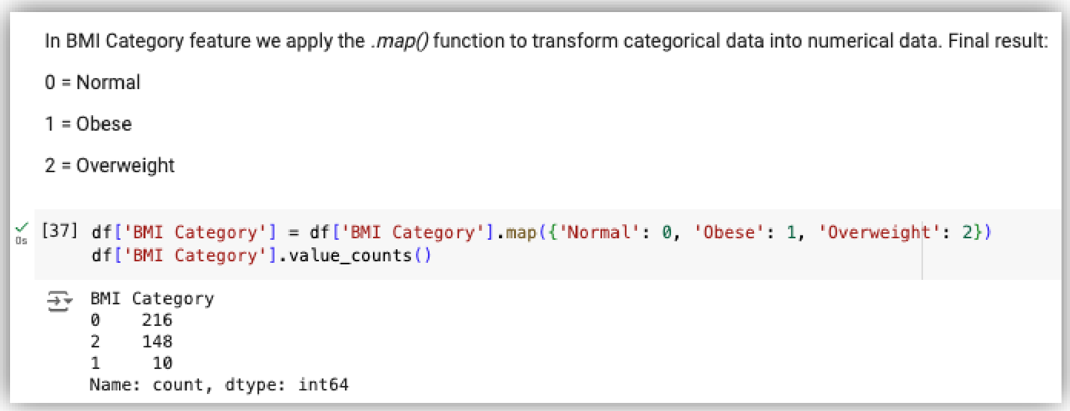


```
[31] df['Occupation'].value_counts()

Occupation
Nurse                   73
Doctor                  71
Engineer                63
Lawyer                  47
Teacher                 40
Accountant              37
Salesperson             32
Software Engineer        4
Scientist                4
Sales Representative     2
Manager                  1
Name: count, dtype: int64
```

```
[32] df['Occupation'] = le.fit_transform(df['Occupation'])
     df['Occupation'].unique()

array([ 9,  1,  6, 10,  5,  2,  0,  8,  3,  7,  4])
```

```
df['Occupation'].value_counts()

Occupation
5     73
1     71
2     63
3     47
10    40
0     37
7     32
9      4
8      4
6      2
4      1
Name: count, dtype: int64
```

Previously, we rectified the BMI variable, which contained duplicate entries representing the same category: Normal BMI. Now, the task at hand is to convert categorical data into numeric formats. To achieve this, we utilized the **.map()** function, assigning 0 to Normal, 1 to Obese, and 2 to Overweight.



In BMI Category feature we apply the *.map()* function to transform categorical data into numerical data. Final result:

0 = Normal

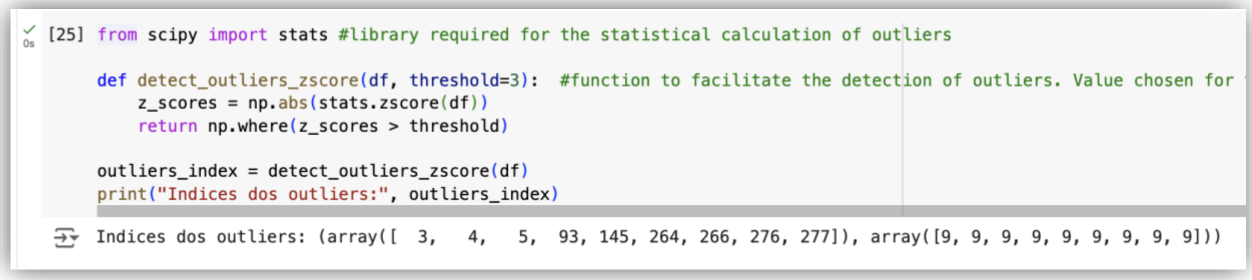
1 = Obese

2 = Overweight

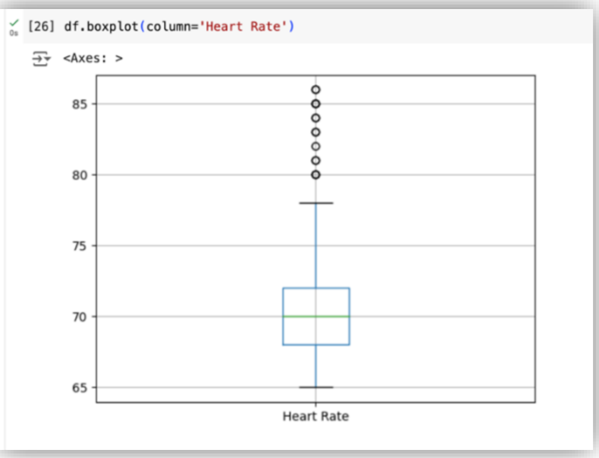
```
[37] df['BMI Category'] = df['BMI Category'].map({'Normal': 0, 'Obese': 1, 'Overweight': 2})
     df['BMI Category'].value_counts()

BMI Category
0    216
2    148
1     10
Name: count, dtype: int64
```

3.2.3.4 Outliers

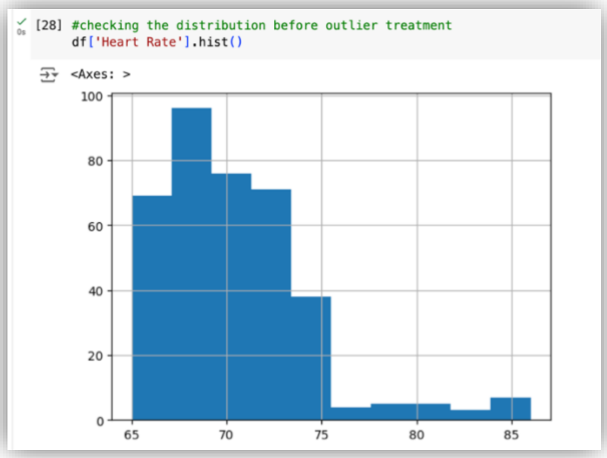
Outliers are data patterns that deviate significantly from the expected behaviour, standing out from the rest of the dataset. They often result from noise or errors in the data and can adversely affect dataset analysis, as well as model fitting (Rousseeuw & Hubert, 2011; Singh & Upadhyaya, 2012). To identify and address outliers within the dataset, we initiate by crafting a code containing a function (def) that computes and exhibits the outliers. Subsequently, we generate a boxplot to visually identify these outliers. Additionally, we generate a histogram to grasp the distribution of the data.



```
[25] from scipy import stats #library required for the statistical calculation of outliers

     def detect_outliers_zscore(df, threshold=3):  #function to facilitate the detection of outliers. Value chosen for
         z_scores = np.abs(stats.zscore(df))
         return np.where(z_scores > threshold)

     outliers_index = detect_outliers_zscore(df)
     print("Indices dos outliers:", outliers_index)

     Indices dos outliers: (array([  3,   4,   5,  93, 145, 264, 266, 276, 277]), array([9, 9, 9, 9, 9, 9, 9, 9, 9]))
```

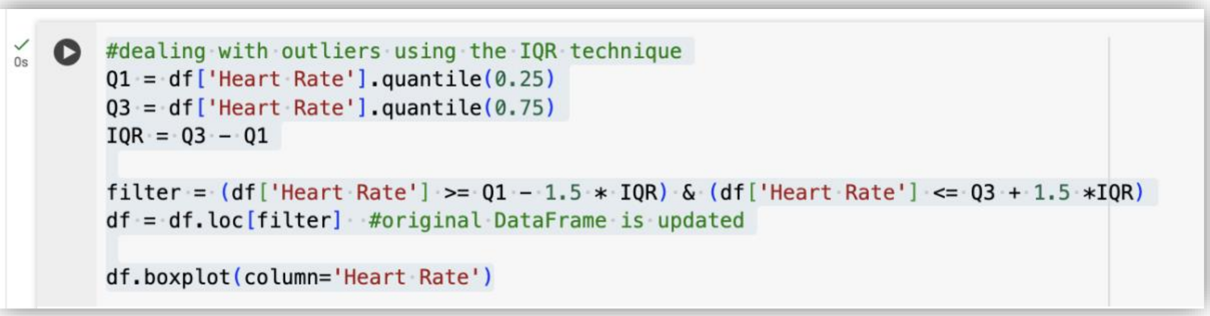


```
[26] df.boxplot(column='Heart Rate')

<Axes: >
```

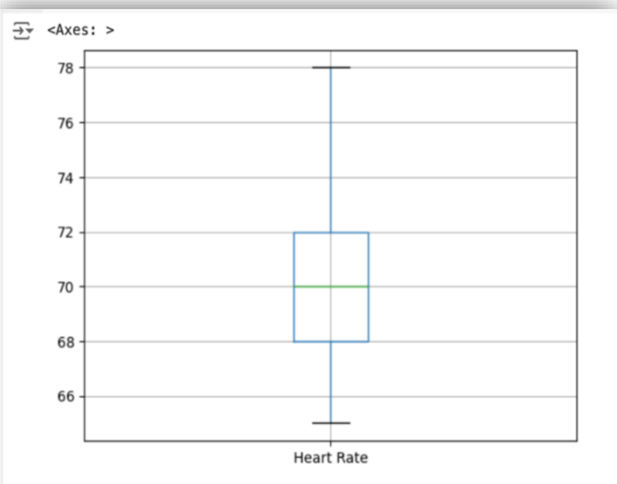


```
[28] #checking the distribution before outlier treatment
     df['Heart Rate'].hist()

<Axes: >
```

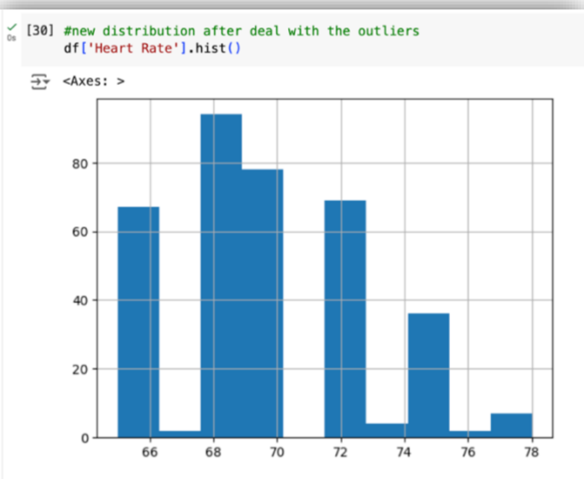
To address outliers, we employ the Interquartile Range (IQR) technique. IQR serves as a fundamental measure of data spread, renowned for its resilience in handling up to 25% of outliers. Calculating IQR involves sorting the data and computing the difference between the 25th and 75th percentile values. These percentiles delineate the data into quartiles, representing 25% and 75% of the dataset, respectively (Buch, 2015). Subsequently, we rectify the outliers and visualize the data distribution once more to assess the changes.



```
#dealing with outliers using the IQR technique
Q1 = df['Heart Rate'].quantile(0.25)
Q3 = df['Heart Rate'].quantile(0.75)
IQR = Q3 - Q1

filter = (df['Heart Rate'] >= Q1 - 1.5 * IQR) & (df['Heart Rate'] <= Q3 + 1.5 *IQR)
df = df.loc[filter]  #original DataFrame is updated

df.boxplot(column='Heart Rate')
```



```
<Axes: >
```

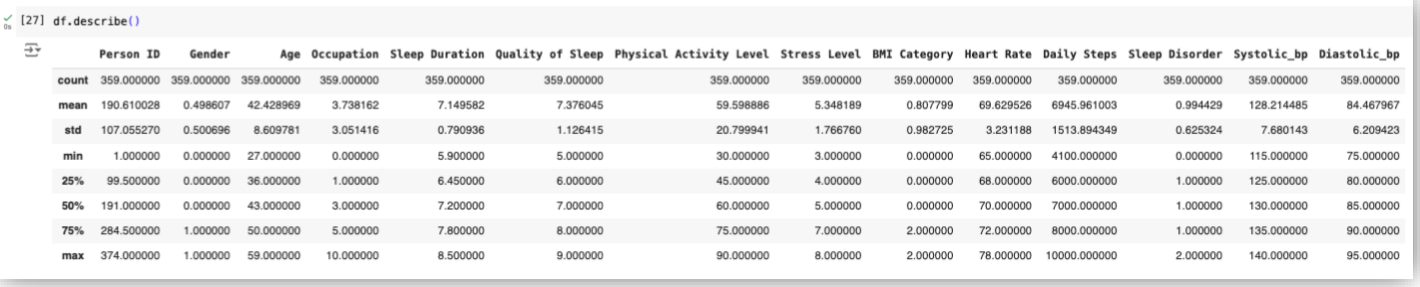


```
[30] #new distribution after deal with the outliers
     df['Heart Rate'].hist()
```

```
<Axes: >
```

3.2.3.5 Descriptive Statistics

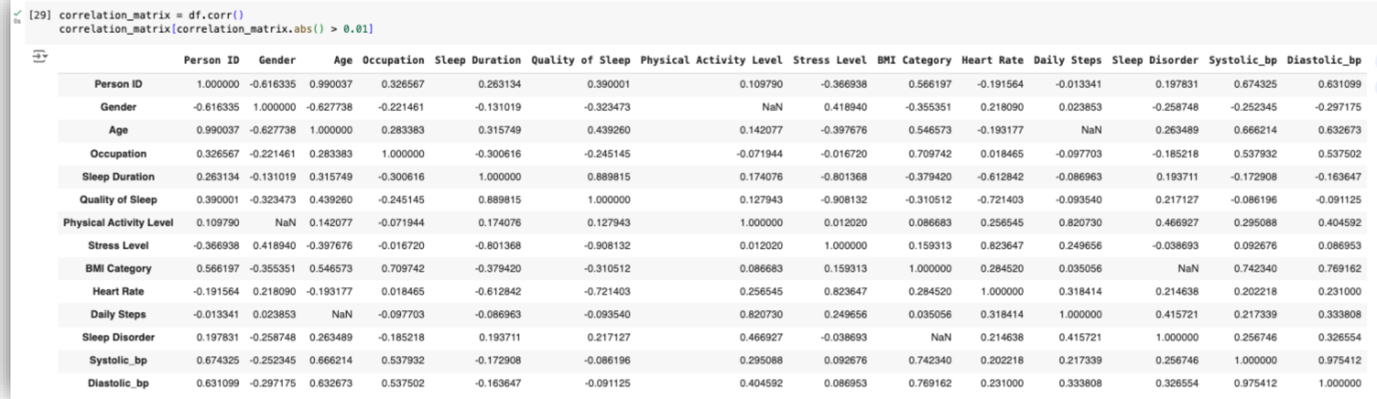
Statistics is the art of collecting, organizing and interpreting data (Case & Ambrosius, 2007). Descriptive statistics and graphs are used to summarize data in a way that is easy to understand. This helps us find patterns in the data and communicate our results clearly. The mean is a number that represents the middle of the data. It's like the "typical" value of the sample. However, it can be affected by extreme values, so attention is required when using it as a measurement. The median is the value in the middle of the ordered data. It's like the midpoint of the sample. It is less affected by extreme values, so it is useful when we have unusual data. Quartiles divide data into four equal parts. The interquartile range (IQR) is the difference between the first quartile (25%) and the third quartile (75%). It is a way of measuring the spread of data. Standard deviation is a common measure of how much data spreads around the mean. It is sensitive to extreme values, but is useful because it is related to the mean and is used in many statistical tests.



```
[27] df.describe()
```

	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Heart Rate	Daily Steps	Sleep Disorder	Systolic_bp	Diastolic_bp
count	359.000000	359.000000	359.000000	359.000000	359.000000	359.000000	359.000000	359.000000	359.000000	359.000000	359.000000	359.000000	359.000000	359.000000
mean	190.610028	0.498607	42.428969	3.738162	7.149582	7.376045	59.598886	5.348189	0.807799	69.629526	6945.961003	0.994429	128.214485	84.467967
std	107.055270	0.500696	8.609781	3.051416	0.790936	1.126415	20.799941	1.766760	0.982725	3.231188	1513.894349	0.625324	7.680143	6.209423
min	1.000000	0.000000	27.000000	0.000000	5.900000	5.000000	30.000000	3.000000	0.000000	65.000000	4100.000000	0.000000	115.000000	75.000000
25%	99.500000	0.000000	36.000000	1.000000	6.450000	6.000000	45.000000	4.000000	0.000000	68.000000	6000.000000	1.000000	125.000000	80.000000
50%	191.000000	0.000000	43.000000	3.000000	7.200000	7.000000	60.000000	5.000000	0.000000	70.000000	7000.000000	1.000000	130.000000	85.000000
75%	284.500000	1.000000	50.000000	5.000000	7.800000	8.000000	75.000000	7.000000	2.000000	72.000000	8000.000000	1.000000	135.000000	90.000000
max	374.000000	1.000000	59.000000	10.000000	8.500000	9.000000	90.000000	8.000000	2.000000	78.000000	10000.000000	2.000000	140.000000	95.000000

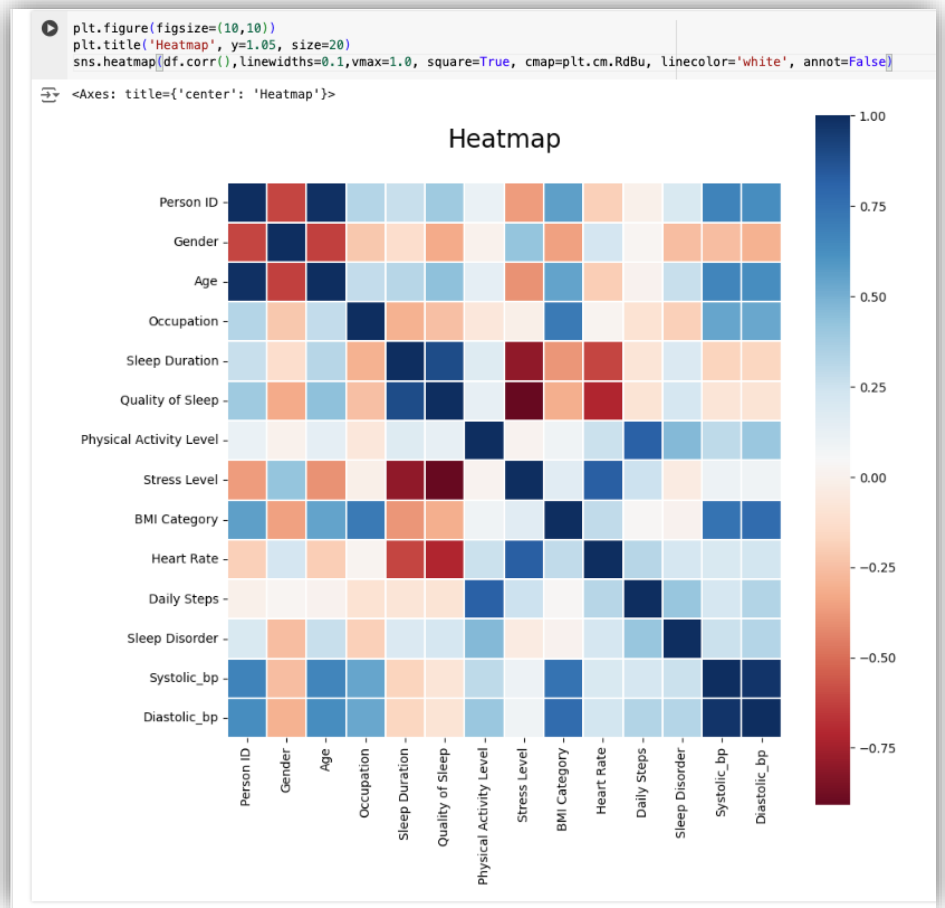
The correlation matrix is a table that shows the correlation between variables in a dataset. It calculates the correlation coefficient between all possible combinations of variables, showing how related they are to each other.



```
[29] correlation_matrix = df.corr()
     correlation_matrix[correlation_matrix.abs() > 0.01]
```

	Person ID	Gender	Age	Occupation	Sleep Duration	Quality of Sleep	Physical Activity Level	Stress Level	BMI Category	Heart Rate	Daily Steps	Sleep Disorder	Systolic_bp	Diastolic_bp
Person ID	1.000000	-0.616335	0.990037	0.326567	0.263134	0.390001	0.109790	-0.366938	0.566197	-0.191564	-0.013341	0.197831	0.674325	0.631099
Gender	-0.616335	1.000000	-0.627738	-0.221461	-0.131019	-0.323473	NaN	0.418940	-0.355351	0.218090	0.023853	-0.258748	-0.252345	-0.297175
Age	0.990037	-0.627738	1.000000	0.283383	0.315749	0.439260	0.142077	-0.397676	0.546573	-0.193177	NaN	0.263489	0.666214	0.632673
Occupation	0.326567	-0.221461	0.283383	1.000000	-0.300616	-0.245145	-0.071944	-0.016720	0.709742	0.018465	-0.097703	-0.185218	0.537932	0.537502
Sleep Duration	0.263134	-0.131019	0.315749	-0.300616	1.000000	0.889815	0.174076	-0.801368	-0.379420	-0.612842	-0.086963	0.193711	-0.172908	-0.163647
Quality of Sleep	0.390001	-0.323473	0.439260	-0.245145	0.889815	1.000000	0.127943	-0.908132	-0.310512	-0.721403	-0.093540	0.217127	-0.086196	-0.091125
Physical Activity Level	0.109790	NaN	0.142077	-0.071944	0.174076	0.127943	1.000000	0.012020	0.086683	0.256545	0.820730	0.466927	0.295088	0.404592
Stress Level	-0.366938	0.418940	-0.397676	-0.016720	-0.801368	-0.908132	0.012020	1.000000	0.159313	0.823647	0.249656	-0.038693	0.092676	0.086953
BMI Category	0.566197	-0.355351	0.546573	0.709742	-0.379420	-0.310512	0.086683	0.159313	1.000000	0.284520	0.035056	NaN	0.742340	0.769162
Heart Rate	-0.191564	0.218090	-0.193177	0.018465	-0.612842	-0.721403	0.256545	0.823647	0.284520	1.000000	0.318414	0.214638	0.202218	0.231000
Daily Steps	-0.013341	0.023853	NaN	-0.097703	-0.086963	-0.093540	0.820730	0.249656	0.035056	0.318414	1.000000	0.415721	0.217339	0.333808
Sleep Disorder	0.197831	-0.258748	0.263489	-0.185218	0.193711	0.217127	0.466927	-0.038693	NaN	0.214638	0.415721	1.000000	0.256746	0.326554
Systolic_bp	0.674325	-0.252345	0.666214	0.537932	-0.172908	-0.086196	0.295088	0.092676	0.742340	0.202218	0.217339	0.256746	1.000000	0.975412
Diastolic_bp	0.631099	-0.297175	0.632673	0.537502	-0.163647	-0.091125	0.404592	0.086953	0.769162	0.231000	0.333808	0.326554	0.975412	1.000000

Similar to the correlation matrix is the heatmap which is used to identify patterns and trends in data, analyse correlations between features, understand distributions and highlight outliers. Through colors, it is possible to assess the correlations between the features, where we have shades of blue for positive correlations, shades of red for negative correlations and shades of faded red and faded blue for no correlation. An example of some of these correlations can be seen in Table 1.



```
plt.figure(figsize=(10,10))
plt.title('Heatmap', y=1.05, size=20)
sns.heatmap(df.corr(),linewidths=0.1,vmax=1.0, square=True, cmap=plt.cm.RdBu, linecolor='white', annot=False)
```

```
<Axes: title={'center': 'Heatmap'}>
```

Type of correlation	Correlation matrix values
Positive correlation (1)	
BMI Category and Occupation	0.709742
Sleep Duration and Quality of Sleep	0.889815
Physical Activity Level and Daily Steps	0.820730
Heart Rate and Stress Level	0.823647
Negative correlation (2)	
Stress Level and Sleep Duration	-0.801368
Quality of Sleep and Stress Level	-0.908132
Sleep Duration and Heart Hate	-0.612842
Gender and Age	-0.627738
No correlation (0)	
Physical Activity Level and Gender	NaN
Daily Steps and Age	NaN
Sleep Disorder and BMI Category	NaN

Table 1: Examples of positive, negative and non-correlations

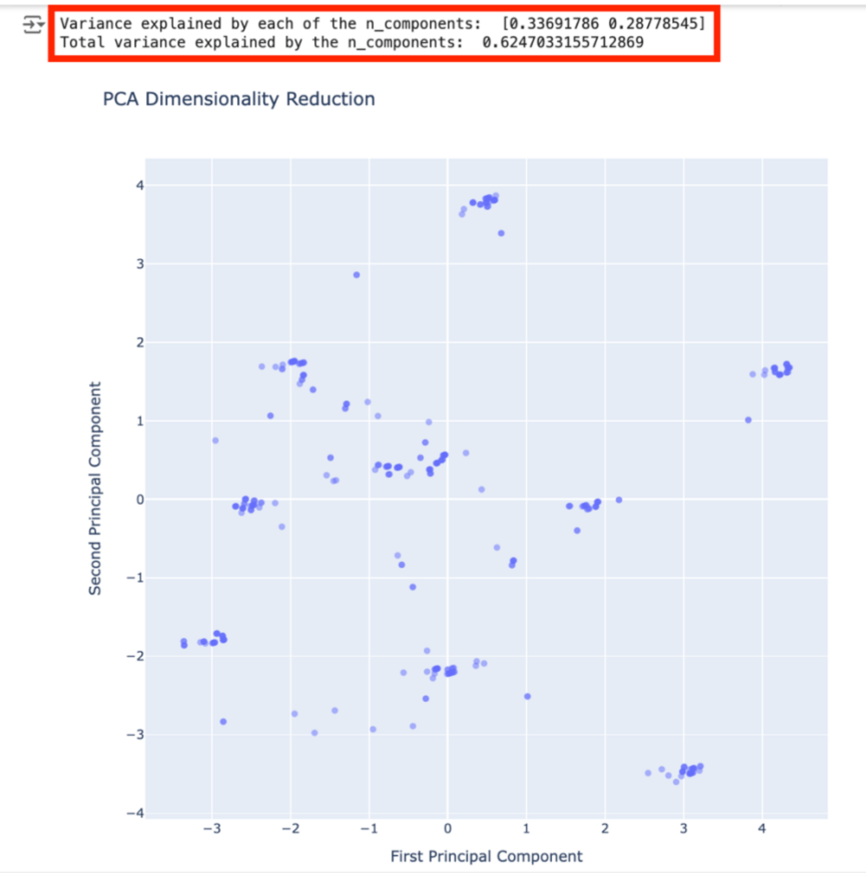
3.2.4 Applying the Models and Checking the First Results

3.2.4.1 Dimension Reduction Techniques

Contrary to common belief, dimension reduction is not about eliminating variables; that technique is known as feature selection, which can also reduce the dimensions of a dataset. Dimension reduction, however, involves robust and modern techniques that assess the importance of variables and the information they contain, creating new variables that encapsulate the most significant information for the model (Sorzano et al., n.d.). High-dimensional datasets, which contain numerous features, can make working with machine learning or data analysis models challenging and complex. Dimension reduction can help prevent overfitting and yield more accurate model results (Cunningham, n.d.). In this study, two dimension reduction techniques were explored: Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP). The following sections will detail these techniques and evaluate their performance in clustering.

3.2.4.1.1 Principal Component Analysis (PCA)

Principal Component Analysis, commonly known as PCA, is one of the most widely used linear dimension reduction techniques (van der Maaten et al., 2007). Originally developed by Karl Pearson in 1901 (Sorzano et al., n.d.), PCA reduces the dimensions of a dataset by creating new variables. These new variables are linear combinations of the original variables that capture the maximum variance (Al-Kandari & Jolliffe, 2005). Essentially, PCA calculates the variance of the variables and creates two or three new features (2D or 3D) that can be used in models. However, PCA has its limitations. It does not handle complex nonlinear data well, which can lead to what is known as the crowding problem due to the large amount of non-linear data in a dataset. Additionally, because PCA is based on variance calculation, it may not perform well on datasets with many discrete variables. To ensure that PCA provides meaningful results, it is important to check how much variance is captured. The sum of the variances of the principal components (n_components) should be at least 75% (0.75); otherwise, PCA may not be effective. In this research, we will use PCA to observe its performance with the selected dataset.



```
Variance explained by each of the n_components:  [0.33691786 0.28778545]
Total variance explained by the n_components:  0.6247033155712869
```

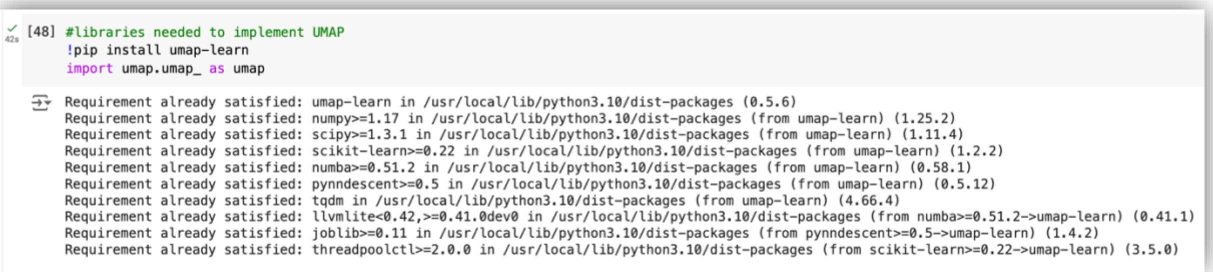
PCA Dimensionality Reduction
Second Principal Component
First Principal Component
4
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3
4

As we can see from the image above, PCA captured 0.62 (less than 0.75 of variance), justifying the lack of clarity in visualizing the clusters. We will proceed with UMAP which can deliver more satisfactory results.

3.2.4.1.2 Uniform Manifold Approximation and Projection (UMAP)

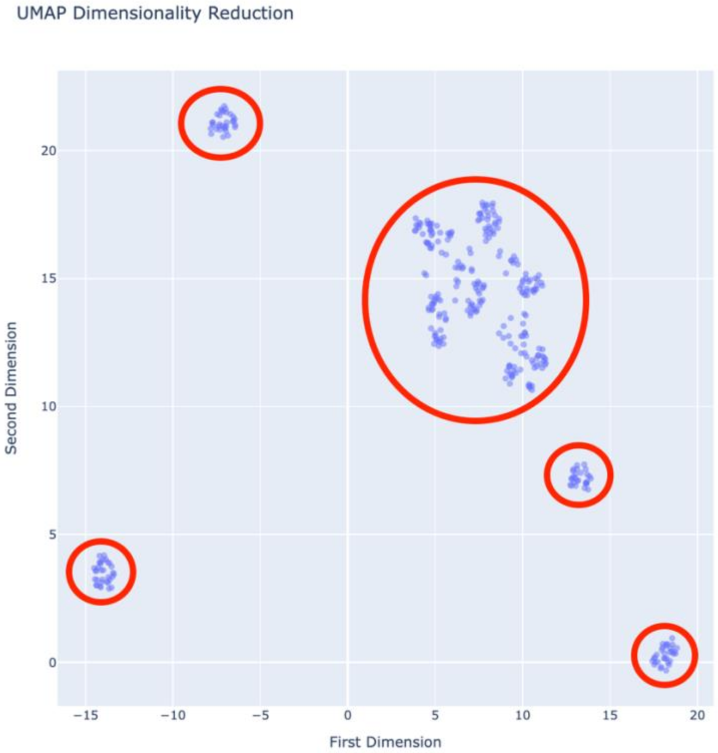
To deal with the limitations of PCA, other techniques were created, including Uniform Manifold Approximation and Projection (UMAP). Created in 2018 by McInnes, Healy, and Melville (McInnes et al., 2020), UMAP is a dimensionality reduction algorithm used to visualize and understand high-dimensional datasets. It works by building a neighborhood graph to capture local connectivity between data points. It then optimizes a projection of this data into a lower dimensional space, such as 2D or 3D, preserving the global structure of the data. This is done by minimizing a cost function that compares the characteristics of the diffusion function in high-dimensional space with the projection in low-dimensional space. Behind this, we have calculations based on algebraic topology (Ghojogh et al., 2021). Some of the advantages we have when using UMAP are: preservation of the global structure of the data, good performance in high-dimensional datasets, it handles non-linear data and complex structures well, and it can be used in different contexts, being very versatile. Furthermore, when compared to t-SNE, UMAP is faster in process time and better handles large datasets (Vermeulen et al., 2021). Some of the points that can be considered negative are sensitivity to the choice of parameters, and the organization of datapoints, which may seem messy, but make sense, since UMAP places datapoints and similar clusters close to each other.

Next, we evaluate the performance of this algorithm with our dataset.



```
[48] #libraries needed to implement UMAP
     !pip install umap-learn
     import umap.umap_ as umap
```

```
Requirement already satisfied: umap-learn in /usr/local/lib/python3.10/dist-packages (0.5.6)
Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-packages (from umap-learn) (1.25.2)
Requirement already satisfied: scipy>=1.3.1 in /usr/local/lib/python3.10/dist-packages (from umap-learn) (1.11.4)
Requirement already satisfied: scikit-learn>=0.22 in /usr/local/lib/python3.10/dist-packages (from umap-learn) (1.2.2)
Requirement already satisfied: numba>=0.51.2 in /usr/local/lib/python3.10/dist-packages (from umap-learn) (0.58.1)
Requirement already satisfied: pynndescent>=0.5 in /usr/local/lib/python3.10/dist-packages (from umap-learn) (0.5.12)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages (from umap-learn) (4.66.4)
Requirement already satisfied: llvmlite<0.42,>=0.41.0dev0 in /usr/local/lib/python3.10/dist-packages (from numba>=0.51.2->umap-learn) (0.41.1)
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.10/dist-packages (from pynndescent>=0.5->umap-learn) (1.4.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in /usr/local/lib/python3.10/dist-packages (from scikit-learn>=0.22->umap-learn) (3.5.0)
```



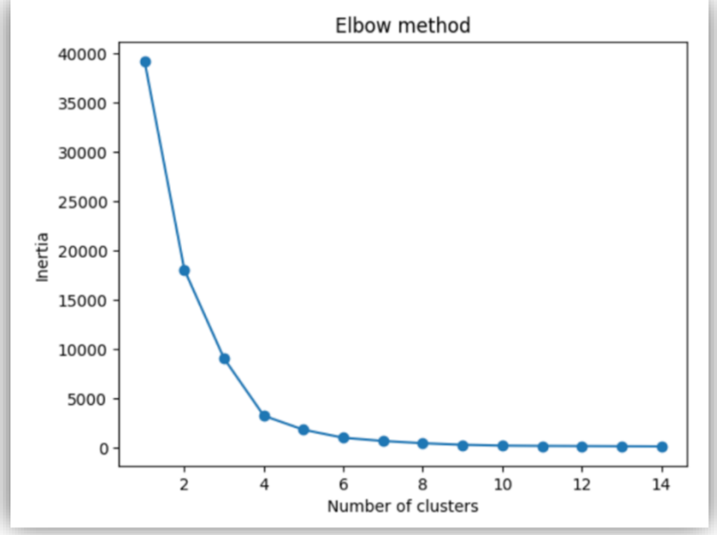
When monitoring the performance of UMAP we noticed a better division between the clusters formed. Clearly we have 4 visibly distinct clusters plus a large cluster that is more spread out. To create the labels we will use K-means and DBSCAN, observing the performance of these models in the next step.

3.2.4.2 Clusters

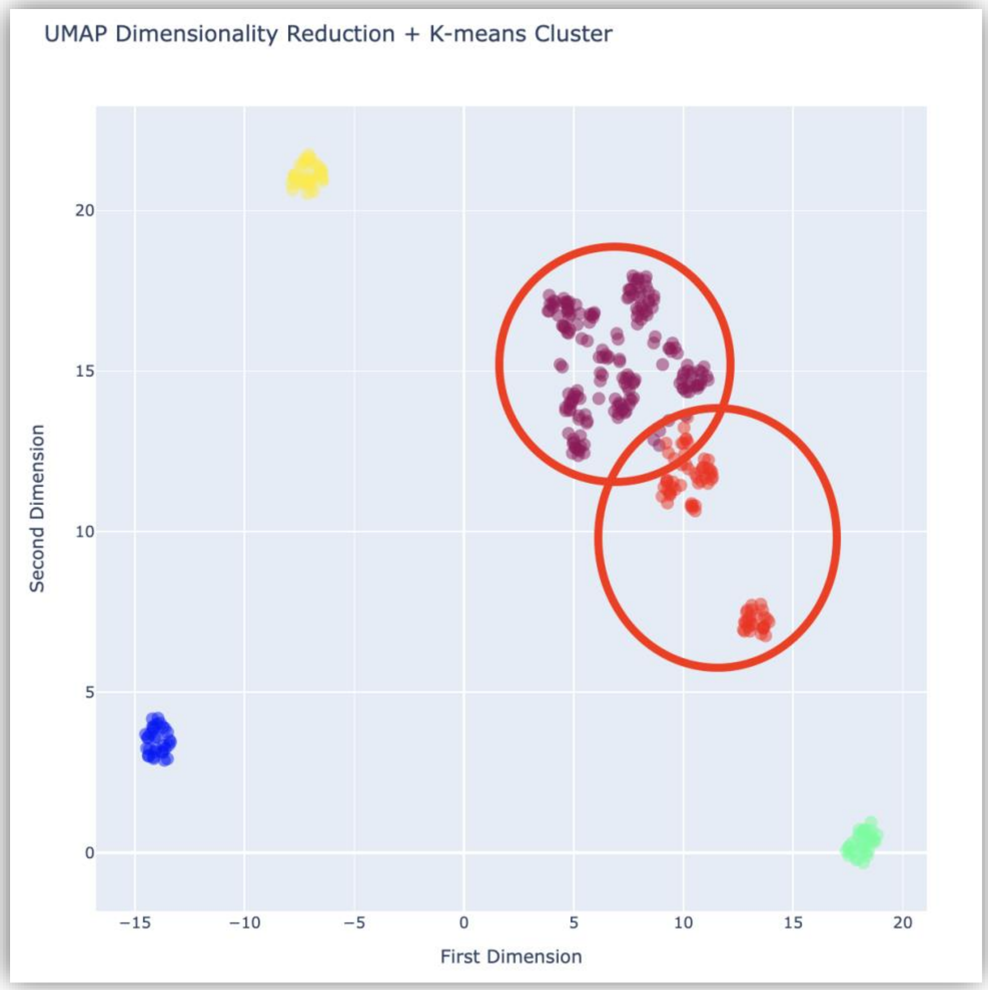
Clusters refer to the task of grouping a set of data objects into subgroups or clusters such that objects within the same group are more similar to each other than to objects in other groups. This technique is commonly used in data analysis to discover intrinsic patterns in data and identify relationships between different observations, in addition to being considered one of the most important unsupervised learning problem (Madhulatha, 2012). Clusters can be found in a variety of domains, from customer segmentation in marketing to document classification in text mining. In the case of our project, the intention is to form clusters that separate the data according to the lifestyle and sleeping patterns of each individual, so we can understand how these groups work and we can recommend changes to the routine, seeking to improve the quality of sleep of those who sleep poorly. Some of the best-known algorithms for clustering are: K-Means, DBSCAN, Hierarchical Clustering and Gaussian Mixture Models (GMM). Next, we will explore the first two in more detail.

3.2.4.2.1 K-Means

K-means is one of the most popular and most used algorithms in the area of unsupervised learning (Dubey & Choubey, n.d.). Initially created by Stuart Lloyd in 1957, and later implemented by James MacQueen in 1967 when it was called K-means (Shafeeq, 2012), this is an algorithm based on of calculating the Euclidean distance, making dimension reduction mandatory, since the Euclidean distance is only capable of correctly calculating within 2 or 3 dimensions (Faisal et al., 2020). It works as follows: first, K points are chosen as the initial centroids of the clusters. K is the number of clusters we want to find. We then assign each point to the cluster whose centroids it is closest to. Each set of points assigned to a centroids is a cluster. Then, the centroids of each cluster are updated based on the points assigned to them. These steps are repeated until no point changes cluster, that is, until the cluster centroids remain the same. This is done to find natural groups in the data, where points within each group are similar to each other and different from points in other groups (Komarasamy & Wahi, n.d.). However, some important details must be highlighted: we must avoid K-means if the clusters touch each other; it is essential to know the cluster number, otherwise we cannot run the algorithm without a pre-established value for K; in case of non-spherical data, such as circles or moons, K-means is going to fail, this is because it performs better on data globular/blobs. Furthermore, K-means probably fails in case of outliers or noise, since the algorithm needs to include all datapoints in some place/cluster (Morissette & Chartier, 2013). Among so many important details for the proper functioning of K-means, we also have the Elbow method, which can be used to get an idea of how many clusters we can attribute to the K value. This is a method based on the Sum Square error (SSE) calculation, where we observe a significant change in the slope of the curve, also known as "elbow", a curve plotted within a graph where the x-axis represents the number of clusters and the y-axis represents the value of the cost function. This way it is possible to have a good idea of a value for K (Bholowalia, n.d.; Humaira & Rasyidah, 2020). Below are the results of using the Elbow method in conjunction with K-means to create cluster labels.



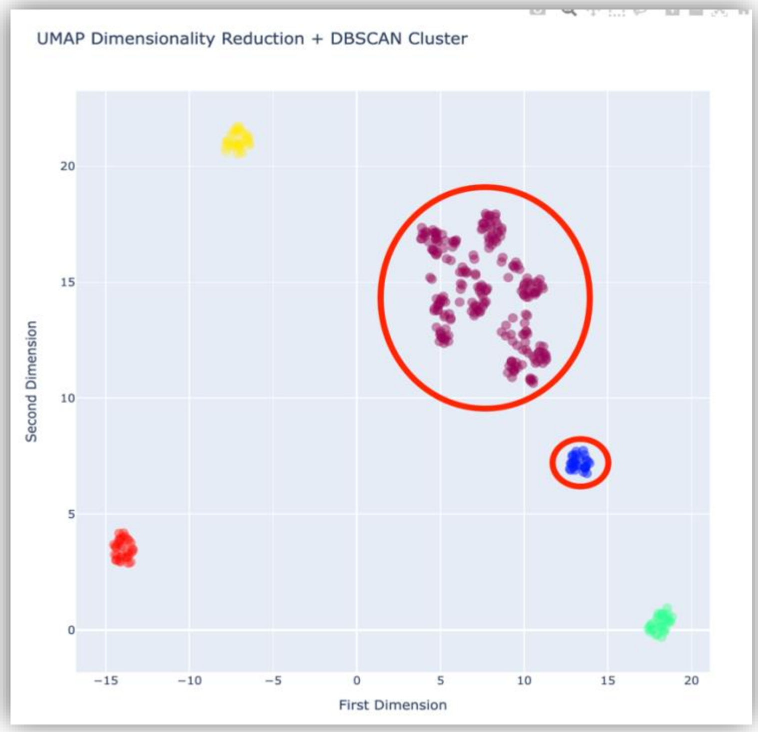
The elbow technique suggests some value between 4, 5 and 6. As we were able to previously visualize 5 clusters, we used K=5.



It can be seen that K-means was not able to color the clusters appropriately, mixing together clusters that should have different labels and colors. Next, we will use DBSCAN and observe how the algorithm will behave in this scenario.

3.2.4.2.2 Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

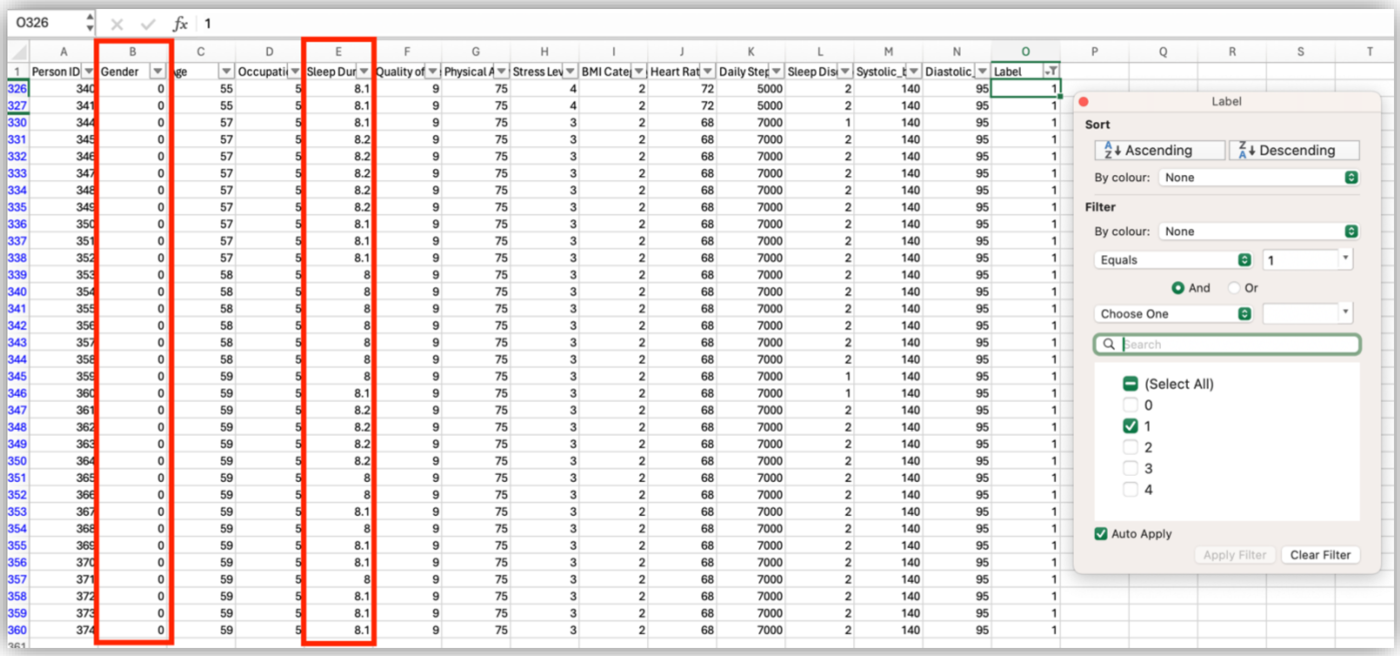
Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering algorithm developed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu, and introduced in 1996 (Ester et al., n.d.). This method is especially effective for identifying clusters of arbitrary shapes and detect noise points in datasets. DBSCAN operates with two main parameters: ε (epsilon), which defines the maximum neighborhood radius of a point, and MinPts, which establishes the minimum number of points necessary to form a dense cluster. The execution of the algorithm begins with marking all points as unvisited. Then, each point is visited and, if it is found with at least neighboring MinPts within radius ε, it is marked as a central point or core, and a new cluster is started. This cluster is then expanded, including all neighboring points that are within radius ε. If these neighboring points are also core points, the expansion process continues until all points within the cluster are identified. Points that do not belong to any cluster are classified as noise. One of the main advantages of DBSCAN is its ability to identify clusters of arbitrary shapes, unlike algorithms such as K-means, which assume spherical shapes for clusters. Furthermore, DBSCAN is robust in detecting noise, making it suitable for applications in image analysis, anomaly detection, and geospatial data segmentation. However, the choice of parameters ε and MinPts can be challenging and requires prior knowledge about the structure of the data (Wang et al., 2015). Furthermore, the computational complexity of DBSCAN can become a hindrance in large datasets, especially at high dimensionality. Next, following DBSCAN's performance in creating labels for our model.



DBSCAN adequately identified and colored the clusters previously separated by UMAP, showing more efficiency than K-means in this scenario.

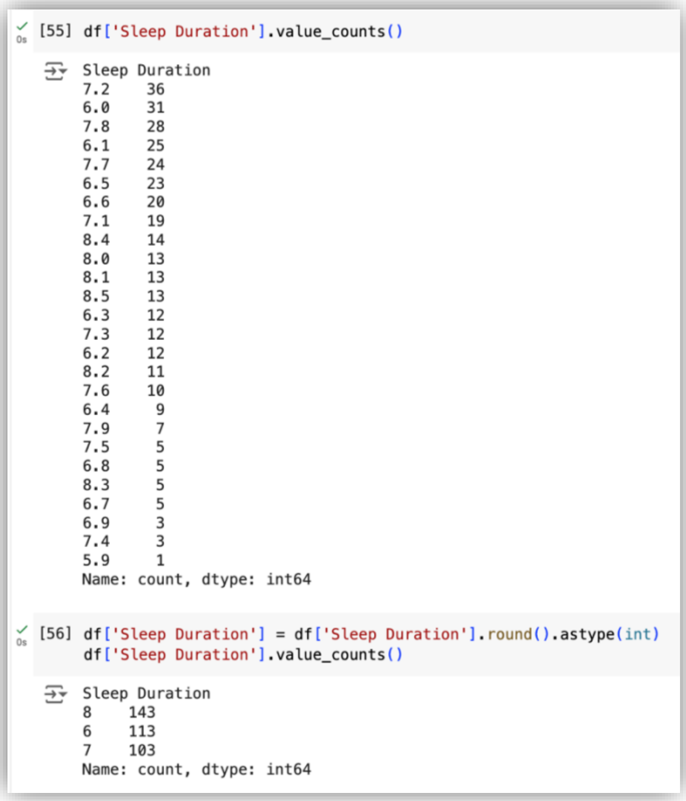
3.2.4.3 Modifications and Implementation to Improve Results

After implementing the models, we examined the new dataset with the generated labels. At the beginning of our code, we have a code cell that requests permission to access Google Drive. Following the application of K-means and DBSCAN, we included a part of the code that stores the new dataset in Google Drive, ensuring the results are saved for subsequent observations and analysis. Upon careful analysis of the results, we observed that most of the clusters formed were primarily using the gender feature as a grouping criterion. This outcome is not ideal for our project, as our goal is to use the clusters to recommend lifestyle and sleep changes. For these recommendations to be effective, it is preferable for the clusters to include individuals of both genders. To address this issue, we employed the **RandomForestClassifier()** method to generate a table containing feature selection recommendations. Subsequently, we removed the 'Gender' feature, as it does not align with our research objectives, 'Person ID,' which offers no contributions to the clustering process, 'Sleep Disorder,' as it falls outside the scope of our research objectives, and 'Labels,' given that we intend to rerun the models. Additionally, it was noted that the clusters were using Sleep Duration as a primary division criterion. To improve the accuracy and objectivity of these values, we considered applying the **.round()** and **.astype()** methods.



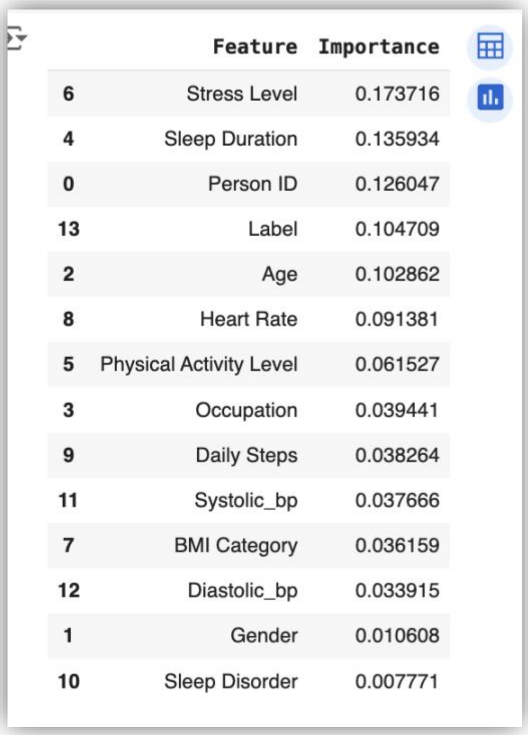
O326 fx 1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Person ID	Gender	ge	Occupati	Sleep Dur	Quality of	Physical A	Stress Lev	BMI Cate	Heart Rat	Daily Step	Sleep Dis	Systolic_b	Diastolic_	Label
326	340	0	55	5	8.1	9	75	4	2	72	5000	2	140	95	1
327	341	0	55	5	8.1	9	75	4	2	72	5000	2	140	95	1
330	344	0	57	5	8.1	9	75	3	2	68	7000	1	140	95	1
331	345	0	57	5	8.2	9	75	3	2	68	7000	2	140	95	1
332	346	0	57	5	8.2	9	75	3	2	68	7000	2	140	95	1
333	347	0	57	5	8.2	9	75	3	2	68	7000	2	140	95	1
334	348	0	57	5	8.2	9	75	3	2	68	7000	2	140	95	1
335	349	0	57	5	8.2	9	75	3	2	68	7000	2	140	95	1
336	350	0	57	5	8.1	9	75	3	2	68	7000	2	140	95	1
337	351	0	57	5	8.1	9	75	3	2	68	7000	2	140	95	1
338	352	0	57	5	8.1	9	75	3	2	68	7000	2	140	95	1
339	353	0	58	5	8	9	75	3	2	68	7000	2	140	95	1
340	354	0	58	5	8	9	75	3	2	68	7000	2	140	95	1
341	355	0	58	5	8	9	75	3	2	68	7000	2	140	95	1
342	356	0	58	5	8	9	75	3	2	68	7000	2	140	95	1
343	357	0	58	5	8	9	75	3	2	68	7000	2	140	95	1
344	358	0	58	5	8	9	75	3	2	68	7000	2	140	95	1
345	359	0	59	5	8	9	75	3	2	68	7000	1	140	95	1
346	360	0	59	5	8.1	9	75	3	2	68	7000	1	140	95	1
347	361	0	59	5	8.2	9	75	3	2	68	7000	2	140	95	1
348	362	0	59	5	8.2	9	75	3	2	68	7000	2	140	95	1
349	363	0	59	5	8.2	9	75	3	2	68	7000	2	140	95	1
350	364	0	59	5	8.2	9	75	3	2	68	7000	2	140	95	1
351	365	0	59	5	8	9	75	3	2	68	7000	2	140	95	1
352	366	0	59	5	8	9	75	3	2	68	7000	2	140	95	1
353	367	0	59	5	8.1	9	75	3	2	68	7000	2	140	95	1
354	368	0	59	5	8	9	75	3	2	68	7000	2	140	95	1
355	369	0	59	5	8.1	9	75	3	2	68	7000	2	140	95	1
356	370	0	59	5	8.1	9	75	3	2	68	7000	2	140	95	1
357	371	0	59	5	8	9	75	3	2	68	7000	2	140	95	1
358	372	0	59	5	8.1	9	75	3	2	68	7000	2	140	95	1
359	373	0	59	5	8.1	9	75	3	2	68	7000	2	140	95	1
360	374	0	59	5	8.1	9	75	3	2	68	7000	2	140	95	1

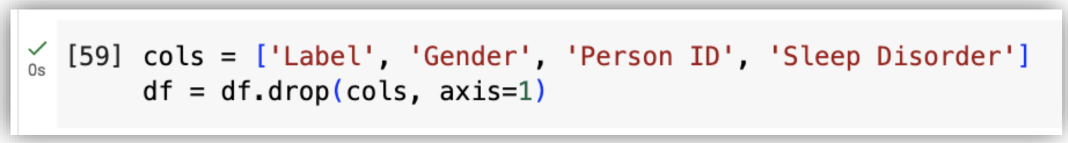


```
[55] df['Sleep Duration'].value_counts()

Sleep Duration
7.2    36
6.0    31
7.8    28
6.1    25
7.7    24
6.5    23
6.6    20
7.1    19
8.4    14
8.0    13
8.1    13
8.5    13
6.3    12
7.3    12
6.2    12
8.2    11
7.6    10
6.4     9
7.9     7
7.5     5
6.8     5
8.3     5
6.7     5
6.9     3
7.4     3
5.9     1
Name: count, dtype: int64
```

```
[56] df['Sleep Duration'] = df['Sleep Duration'].round().astype(int)
     df['Sleep Duration'].value_counts()

Sleep Duration
8    143
6    113
7    103
Name: count, dtype: int64
```



	Feature	Importance
6	Stress Level	0.173716
4	Sleep Duration	0.135934
0	Person ID	0.126047
13	Label	0.104709
2	Age	0.102862
8	Heart Rate	0.091381
5	Physical Activity Level	0.061527
3	Occupation	0.039441
9	Daily Steps	0.038264
11	Systolic_bp	0.037666
7	BMI Category	0.036159
12	Diastolic_bp	0.033915
1	Gender	0.010608
10	Sleep Disorder	0.007771



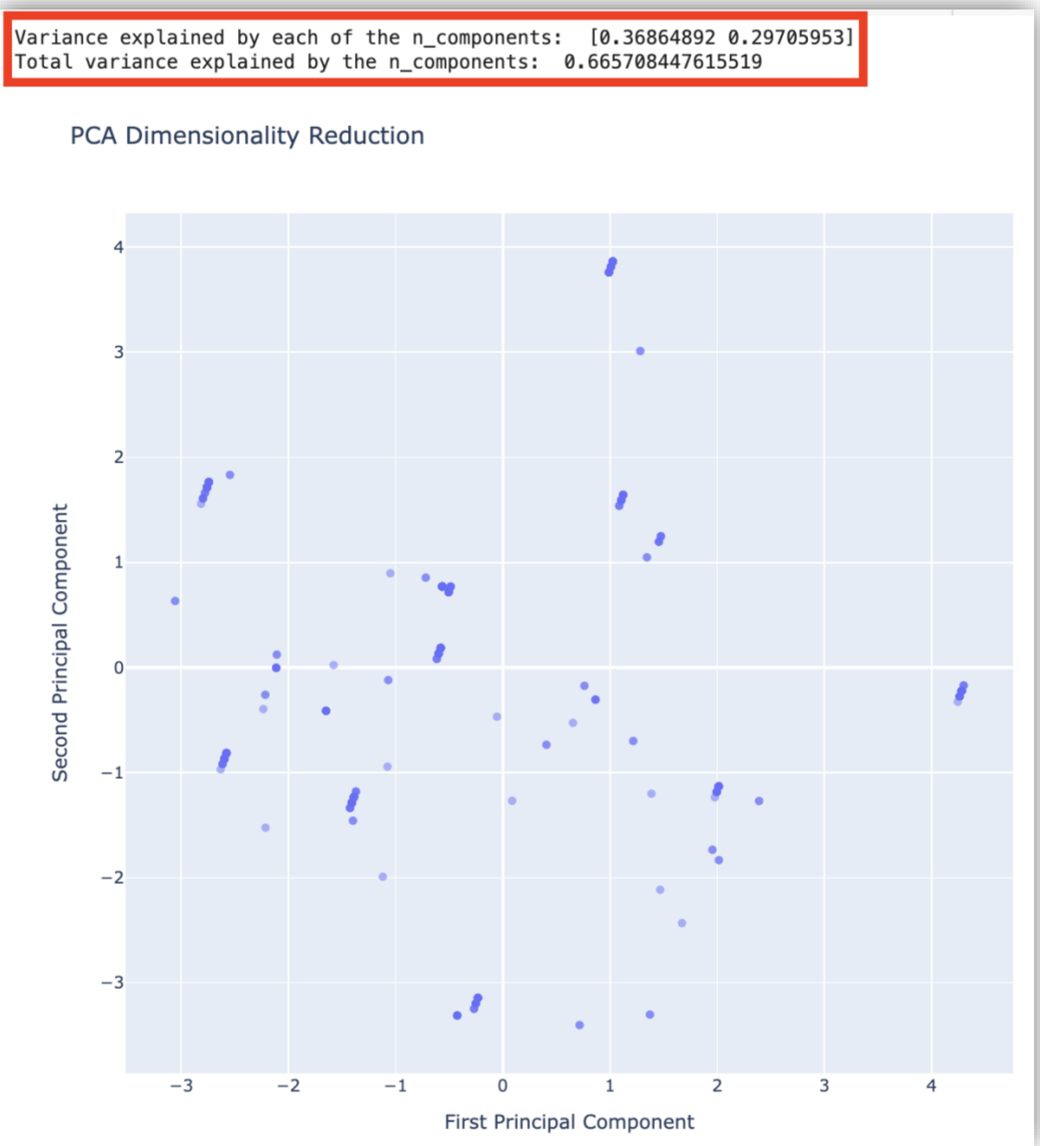
```
[59] cols = ['Label', 'Gender', 'Person ID', 'Sleep Disorder']
     df = df.drop(cols, axis=1)
```

CHAPTER 4

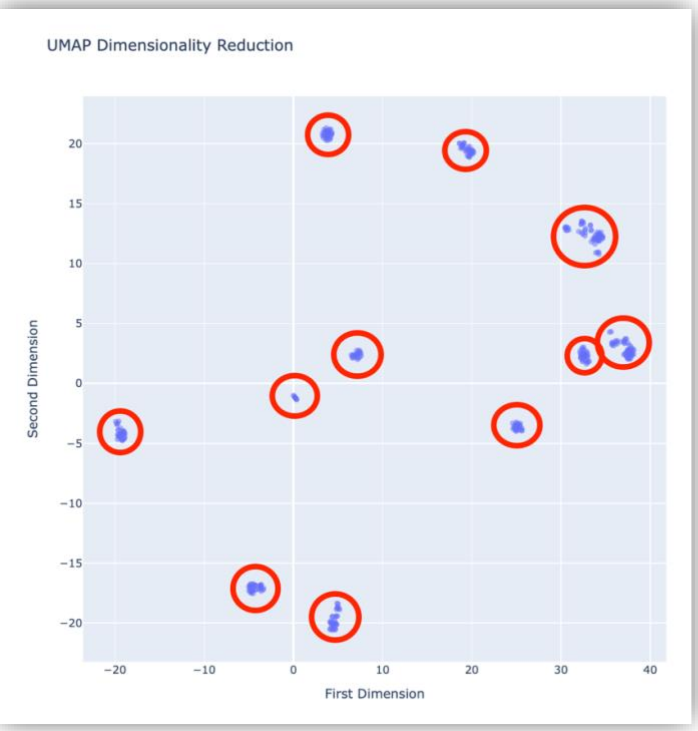
Final Results and Critical Analysis

4.1 PCA and UMAP Final Results and Critical Analysis

After implementing the modifications, we reapplied the PCA and UMAP dimension reduction techniques. Despite PCA continuing to capture less than 75% of the data variance, resulting in a different visualization, it still lacks the necessary clarity for identifying clusters. Conversely, UMAP performs significantly better, distinctly revealing N different clusters. Consequently, we have decided to retain UMAP for implementing the K-means and DBSCAN algorithms.

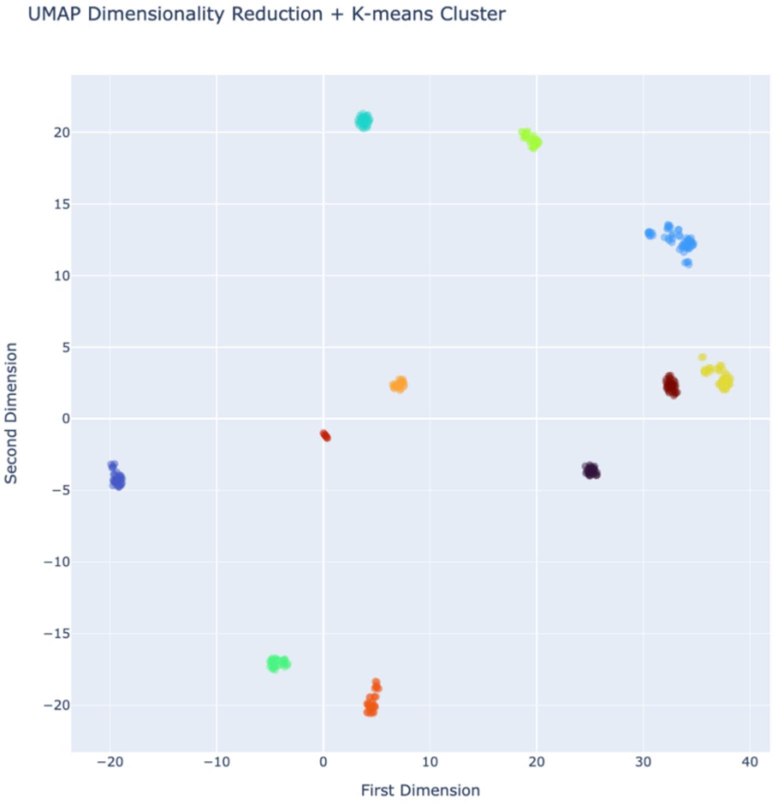


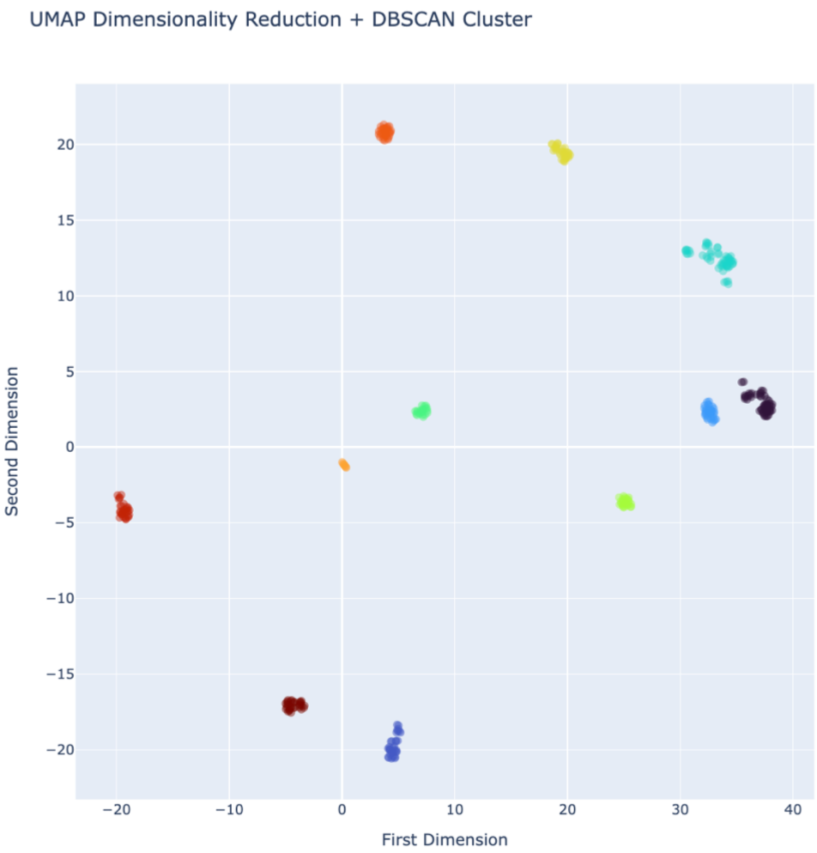
```
Variance explained by each of the n_components:  [0.36864892 0.29705953]
Total variance explained by the n_components:  0.665708447615519
```



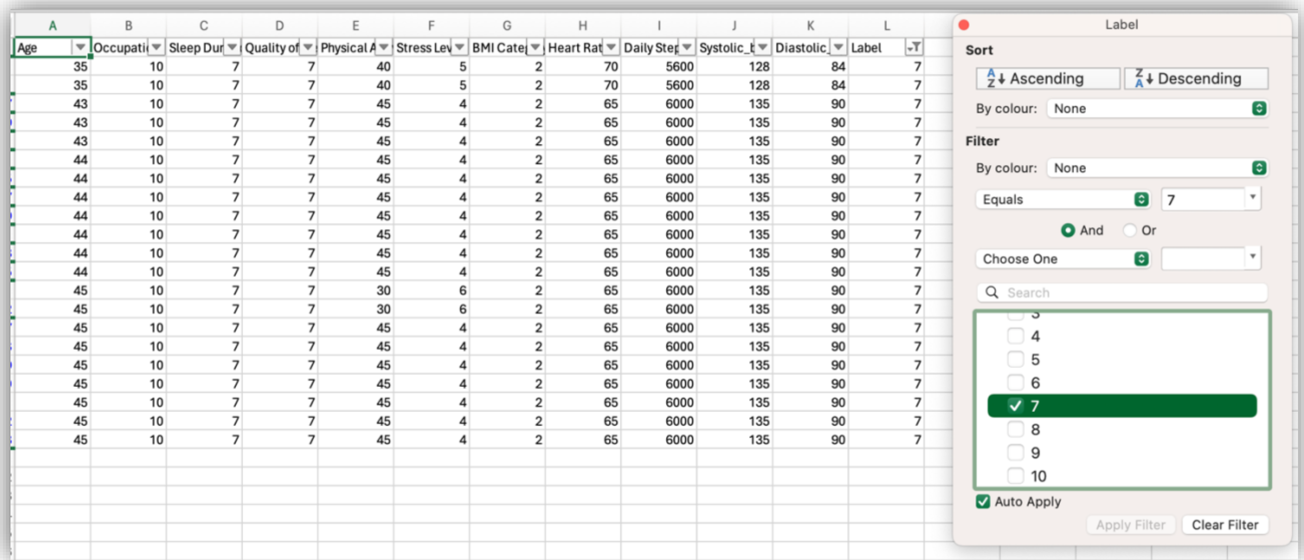
4.2 K-means and DBSCAN Final Results and Critical Analysis

Now that we have successfully identified the clusters using UMAP, we will proceed to rerun the K-means and DBSCAN algorithms to generate the corresponding labels.





After the modifications, K-means is able to properly create the labels for the clusters, coloring each cluster appropriately. DBSCAN continues to perform well, maintaining constant parameters since the initial attempt. When we check the labels created through the new files that are saved on Google Drive, we are able to understand the individuals' routine and sleep patterns. In the next chapter we will discuss more about this.



A	B	C	D	E	F	G	H	I	J	K	L
Age	Occupati	Sleep Dur	Quality of	Physical A	Stress Lev	BMI Cate	Heart Rat	Daily Step	Systolic_b	Diastolic_	Label
35	10	7	7	40	5	2	70	5600	128	84	7
35	10	7	7	40	5	2	70	5600	128	84	7
43	10	7	7	45	4	2	65	6000	135	90	7
43	10	7	7	45	4	2	65	6000	135	90	7
43	10	7	7	45	4	2	65	6000	135	90	7
44	10	7	7	45	4	2	65	6000	135	90	7
44	10	7	7	45	4	2	65	6000	135	90	7
44	10	7	7	45	4	2	65	6000	135	90	7
44	10	7	7	45	4	2	65	6000	135	90	7
44	10	7	7	45	4	2	65	6000	135	90	7
44	10	7	7	45	4	2	65	6000	135	90	7
44	10	7	7	45	4	2	65	6000	135	90	7
45	10	7	7	30	6	2	65	6000	135	90	7
45	10	7	7	30	6	2	65	6000	135	90	7
45	10	7	7	45	4	2	65	6000	135	90	7
45	10	7	7	45	4	2	65	6000	135	90	7
45	10	7	7	45	4	2	65	6000	135	90	7
45	10	7	7	45	4	2	65	6000	135	90	7
45	10	7	7	45	4	2	65	6000	135	90	7
45	10	7	7	45	4	2	65	6000	135	90	7
45	10	7	7	45	4	2	65	6000	135	90	7

Label

Sort

By colour: None

Filter

By colour: None

Equals 7

● And ○ Or

Choose One

□ 4
□ 5
□ 6
☑ 7
□ 8
□ 9
□ 10

☑ Auto Apply

CHAPTER 5

Discussions and Conclusions

5.1 Summary of Findings

An interesting observation is that the K-means algorithm initially underperformed, but subsequent attempts showed improvement, highlighting the algorithm's sensitivity and limitations. This underscores the importance of carefully selecting the best model for each dataset or situation. Notably, parameter adjustments were necessary for K-means, with the initial attempt using K=5 and the subsequent one using K=11. In contrast, DBSCAN maintained consistent values for the ε (epsilon) and MinPts parameters, requiring fewer adjustments than K-means. Regarding cluster quality, both K-means and DBSCAN produced meaningful and effective labels. Analysis of cluster patterns revealed that UMAP, the basis for group creation, followed a pattern influenced by age, sleep quality, and sleep duration. When using Sleep Quality as a reference, it is evident that some clusters exhibit uniform sleep quality scores, while others display mixed but proximate values (e.g., 6 and 7, 8 and 9). Throughout the process, multiple parameter adjustments were necessary to find values that met project needs and suited the data used. The model application was divided into two stages; even post-EDA, initial results were suboptimal. Thus, it is advisable to exclude certain features before dimensionality reduction, despite these models retaining the most significant variables. Excluding potentially problematic features can enhance the reduction process. As demonstrated, minor but significant changes improved model performance. Comparing clusters enabled identification of individual patterns, showing that the clusters are well-separated and logical, reflecting the techniques' effectiveness.

5.2 Conclusions

Summarizing the key points from the Research Question session, we conclude that UMAP was the most effective dimensionality reduction technique for this project, outperforming PCA by delivering superior results in both initial and subsequent attempts. PCA, on the other hand, failed in both scenarios, capturing less than 75% variance. Additionally, although both K-means and DBSCAN effectively served their purposes, DBSCAN demonstrated greater efficiency from the outset. Its dynamic nature, capability to handle large-scale data, and adaptability to various types of datasets made DBSCAN more suitable for this project.

Regarding the novelty of the project, it is evident that unsupervised machine learning can successfully identify sleep patterns and routines, offering a cost-effective and easy-to-implement alternative for treating sleep disorders, since we can compare the data and use it as a basis for making changes to our routine. Furthermore, in addressing one of the research questions, we found that 'Age', 'Sleep Quality', and 'Sleep Duration' are the most significant features, as these were fundamental in defining the clusters.

5.3 Limitation

Some limitations were identified in relation to the dataset. Although the research has been completed, the literature review revealed that additional information about individuals' routines and sleep patterns would significantly enhance the results. Including data such as meal times, physical activity schedules, caffeine and nicotine consumption, vitamin intake, screen time, and geographical location could provide more references, resulting in more diversified clusters and better reflecting the public's reality. Another limitation was the age range of the participants, which varied between 27 and 59 years. Incorporating a wider age range and diverse professions would also add significant value to the study.

5.4 Recommendations

As recommendations, we suggest enhancing the database by adding additional significant information for the analysis of sleep patterns and routines, as mentioned in the limitations section. We also recommend testing other unsupervised machine learning models and dimension reduction techniques to compare their efficiency. Additionally, exploring various EDA and Feature Engineering methods could improve clustering results. For instance, given the way clusters are organized by age, performing age binning—such as categorizing into young, young adult, adult, young elderly, and elderly—may enhance the clustering outcomes. It is also likely that other EDA techniques not explored in this project could contribute positively.

Reference List

Ablao, L. G., Tupaz, Z. Y. B., Dela Cruz, J. C., & Ibera, J. (2021). Machine Learning Sleep Phase Monitoring using ECG and EMG. *2021 IEEE 11th International Conference on System Engineering and Technology (ICSET)*, 320–325. https://doi.org/10.1109/ICSET53708.2021.9612546

Adjaye-Gbewonyo, Z., Ng, A., & Black, L. (2022). *Sleep Difficulties in Adults: United States, 2020*. National Center for Health Statistics (U.S.). https://doi.org/10.15620/cdc:117490

Al-Kandari, N. M., & Jolliffe, I. T. (2005). Variable selection and interpretation in correlation principal components. *Environmetrics*, *16*(6), 659–672. https://doi.org/10.1002/env.728

Bholowalia, P. (n.d.). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. *International Journal of Computer Applications*, *105*(9).

Buch, K. (2015). Decision based non-linear filtering using interquartile range estimator for Gaussian signals. *11th IEEE India Conference: Emerging Trends and Innovation in Technology, INDICON 2014*. https://doi.org/10.1109/INDICON.2014.7030658

Case, D., & Ambrosius, W. (2007). Power and Sample Size. *Methods in Molecular Biology (Clifton, N.J.)*, *404*, 377–408. https://doi.org/10.1007/978-1-59745-530-5_19

Chai, C. (2020). The Importance of Data Cleaning: Three Visualization Examples. *CHANCE*, *33*, 4–9. https://doi.org/10.1080/09332480.2020.1726112

Chan, V., & Lo, K. (2022). Efficacy of dietary supplements on improving sleep quality: A systematic review and meta-analysis. *Postgraduate Medical Journal*, *98*(1158), 285–293. https://doi.org/10.1136/postgradmedj-2020-139319

Cooper, K. L., & Relton, C. (2010). Homeopathy for insomnia: A systematic review of research evidence. *Sleep Medicine Reviews*, *14*(5), 329–337. https://doi.org/10.1016/j.smrv.2009.11.005

Cristina Novak, V., Mayara Tilpp, S., Raquel Bim, C., & Cristina Carrasco, A. (2019). Efeito da acupuntura na melhora da ansiedade, sono e qualidade de vida. *O Mundo Da Saúde*, *43*(3), 782–795. https://doi.org/10.15343/0104-7809.20194303782795

Cudney, L. E., Frey, B. N., McCabe, R. E., & Green, S. M. (2022). Investigating the relationship between objective measures of sleep and self-report sleep quality in healthy adults: A review. *Journal of Clinical Sleep Medicine*, *18*(3), 927–936. https://doi.org/10.5664/jcsm.9708

Cunningham, P. (n.d.). *University College Dublin*.

Da Poian, V., Theiling, B., Clough, L., McKinney, B., Major, J., Chen, J., & Hörst, S. (2023). Exploratory data analysis (EDA) machine learning approaches for ocean world analog mass spectrometry. *Frontiers in Astronomy and Space Sciences*, *10*. https://doi.org/10.3389/fspas.2023.1134141

Dadgostar, H., Basharkhah, A., Ghalehbandi, M. F., & Kashaninasab, F. (2023). An Investigation on the Effect of Exercise on Insomnia Symptoms. *International Journal of Preventive Medicine*, *14*, 16. https://doi.org/10.4103/ijpvm.ijpvm_204_21

Dawson, D., & Encel, N. (1993). Melatonin and sleep in humans. *Journal of Pineal Research*, *15*(1), 1–12. https://doi.org/10.1111/j.1600-079X.1993.tb00503.x

Dolezal, B. A., Neufeld, E. V., Boland, D. M., Martin, J. L., & Cooper, C. B. (2017). Interrelationship between Sleep and Exercise: A Systematic Review. *Advances in Preventive Medicine*, *2017*, 1–14. https://doi.org/10.1155/2017/1364387

Dubey, A., & Choubey, A. D. A. (n.d.). *A Systematic Review on K-Means Clustering Techniques*. *6*(6).

Ell, J., Schmid, S. R., Benz, F., & Spille, L. (2023). Complementary and alternative treatments for insomnia disorder: A systematic umbrella review. *Journal of Sleep Research*, *32*(6), e13979. https://doi.org/10.1111/jsr.13979

Espiritu, J. (2008). Aging-Related Sleep Changes. *Clinics in Geriatric Medicine*, *24*, 1–14, v. https://doi.org/10.1016/j.cger.2007.08.007

Ester, M., Kriegel, H.-P., & Xu, X. (n.d.). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*.

Faisal, M., Zamzami, E. M., & Sutarman. (2020). Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance. *Journal of Physics: Conference Series*, *1566*(1), 012112. https://doi.org/10.1088/1742-6596/1566/1/012112

Fernandez-Mendoza, J., & Vgontzas, A. N. (2013). Insomnia and Its Impact on Physical and Mental Health. *Current Psychiatry Reports*, *15*(12), 418. https://doi.org/10.1007/s11920-013-0418-8

Ghojogh, B., Ghodsi, A., Karray, F., & Crowley, M. (2021). *Uniform Manifold Approximation and Projection (UMAP) and its Variants: Tutorial and Survey* (arXiv:2109.02508). arXiv. http://arxiv.org/abs/2109.02508

Humaira, H., & Rasyidah, R. (2020). Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm. *Proceedings of the Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia*. Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA) 2018, 24-25 January 2018, Padang, Indonesia, Padang, Indonesia. https://doi.org/10.4108/eai.24-1-2018.2292388

Ilyas, I. F., & Chu, X. (2015). Trends in Cleaning Relational Data: Consistency and Deduplication. *Foundations and Trends® in Databases*, *5*(4), 281–393. https://doi.org/10.1561/1900000045

Islam, Md. M., Masum, A. K. M., Abujar, S., & Hossain, S. A. (2020). A systematic way of collecting data of insomniac patients: An analytical survey. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–7. https://doi.org/10.1109/ICCCNT49239.2020.9225485

Jones, S., & Brennan, M. J. (2010). *Great Expectations: Baby Sleep Guide*. Union Square & Co.

Khademi, A., El-Manzalawy, Y., Buxton, O. M., & Honavar, V. (2018). Toward personalized sleep-wake prediction from actigraphy. *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 414–417. https://doi.org/10.1109/BHI.2018.8333456

Knappe, S. W., & Sonnesen, L. (2018). Mandibular positioning techniques to improve sleep quality in patients with obstructive sleep apnea: Current perspectives. *Nature and Science of Sleep*, *Volume 10*, 65–72. https://doi.org/10.2147/NSS.S135760

Komarasamy, G., & Wahi, A. (n.d.). *An Optimized K-Means Clustering Technique using Bat Algorithm*.

Kripke, D. F. (2013). Surprising View of Insomnia and Sleeping Pills. *Sleep*, *36*(8), 1127–1128. https://doi.org/10.5665/sleep.2868

Kudale, K. M., Kannan, G., Navulla, D., & S, S. (2023). Social Media Influences in Sleeping Patterns of Human. *2023 International Conference on Disruptive Technologies (ICDT)*, 224–227. https://doi.org/10.1109/ICDT57929.2023.10151147

Madhulatha, T. S. (2012). AN OVERVIEW ON CLUSTERING METHODS. *IOSR Journal of Engineering*, *02*(04), 719–725. https://doi.org/10.9790/3021-0204719725

Majoe, D., Bonhof, P., Kaegi-Trachsel, T., Gutknecht, J., & Widmer, L. (2010). Stress and sleep quality estimation from a smart wearable sensor. *5th International Conference on Pervasive Computing and Applications*, 14–19. https://doi.org/10.1109/ICPCA.2010.5704068

Manzar, M. D., BaHammam, A. S., Hameed, U. A., Spence, D. W., Pandi-Perumal, S. R., Moscovitch, A., & Streiner, D. L. (2018). Dimensionality of the Pittsburgh Sleep Quality Index: A systematic review. *Health and Quality of Life Outcomes*, *16*(1), 89. https://doi.org/10.1186/s12955-018-0915-x

McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* (arXiv:1802.03426). arXiv. http://arxiv.org/abs/1802.03426

Midway, S. R. (2020). Principles of Effective Data Visualization. *Patterns*, *1*(9), 100141. https://doi.org/10.1016/j.patter.2020.100141

Morissette, L., & Chartier, S. (2013). The k-means clustering technique: General considerations and implementation in Mathematica. *Tutorials in Quantitative Methods for Psychology*, *9*, 15–24. https://doi.org/10.20982/tqmp.09.1.p015

Mukhiya, S. K., & Ahmed, U. (2020). *Hands-On Exploratory Data Analysis with Python: Perform EDA techniques to understand, summarize, and investigate your data*. Packt Publishing Ltd.

Najjar, R. P., Wolf, L., Taillard, J., Schlangen, L. J. M., Salam, A., Cajochen, C., & Gronfier, C. (2014). Chronic Artificial Blue-Enriched White Light Is an Effective Countermeasure to Delayed Circadian Phase and Neurobehavioral Decrements. *PLoS ONE*, *9*(7), e102827. https://doi.org/10.1371/journal.pone.0102827

Ohayon, M. M., & Reynolds, C. F. (2009). Epidemiological and clinical relevance of insomnia diagnosis algorithms according to the DSM-IV and the International Classification of Sleep Disorders (ICSD). *Sleep Medicine*, *10*(9), 952–960. https://doi.org/10.1016/j.sleep.2009.07.008

Paalasmaa, J., Waris, M., Toivonen, H., Leppakorpi, L., & Partinen, M. (2012). Unobtrusive online monitoring of sleep at home. *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 3784–3788. https://doi.org/10.1109/EMBC.2012.6346791

Park, S., Zhunis, A., Constantinides, M., Aiello, L. M., Quercia, D., & Cha, M. (2023). Social dimensions impact individual sleep quantity and quality. *Scientific Reports*, *13*(1), Article 1. https://doi.org/10.1038/s41598-023-36762-5

Pattyn, N., Puyvelde, M. V., Fernandez-Tellez, H., Roelands, B., & Mairesse, O. (n.d.). *From the midnight sun to the longest night: Sleep in Antarctica*. Retrieved March 26, 2024, from https://core.ac.uk/reader/82918463?utm_source=linkout

Rawat, T., & Khemchandani, V. (2017). Feature Engineering (FE) Tools and Techniques for Better Classification Performance. *International Journal of Innovations in Engineering and Technology*, *8*(2). https://doi.org/10.21172/ijiet.82.024

Reichert, C. F., Deboer, T., & Landolt, H.-P. (2022). Adenosine, caffeine, and sleep–wake regulation: State of the science and perspectives. *Journal of Sleep Research*, *31*(4), e13597. https://doi.org/10.1111/jsr.13597

Riemann, D., Baglioni, C., Bassetti, C., Bjorvatn, B., Dolenc Groselj, L., Ellis, J. G., Espie, C. A., Garcia-Borreguero, D., Gjerstad, M., Gonçalves, M., Hertenstein, E., Jansson-Fröjmark, M., Jennum, P. J., Leger, D., Nissen, C., Parrino, L., Paunio, T., Pevernagie, D., Verbraecken, J., … Spiegelhalder, K. (2017). European guideline for the diagnosis and treatment of insomnia. *Journal of Sleep Research*, *26*(6), 675–700. https://doi.org/10.1111/jsr.12594

Roenneberg, T., Foster, R. G., & Klerman, E. B. (2022). The circadian system, sleep, and the health/disease balance: A conceptual review. *Journal of Sleep Research*, *31*(4), e13621. https://doi.org/10.1111/jsr.13621

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *WIREs Data Mining and Knowledge Discovery*, *1*(1), 73–79. https://doi.org/10.1002/widm.2

Sano, A., Johns, P., & Czerwinski, M. (2015). HealthAware: An advice system for stress, sleep, diet and exercise. *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 546–552. https://doi.org/10.1109/ACII.2015.7344623

Segundo, L. V. G., Neto, B. F. C., & de Araujo, D. (2017). *ASPECTOS RELACIONADOS À QUALIDADE DO SONO EM ESTUDANTES DE MEDICINA / FEATURES RELATED TO QUALITY OF SLEEP IN MEDICAL STUDENTS*.

Sejbuk, M., Mirończuk-Chodakowska, I., & Witkowska, A. M. (2022). Sleep Quality: A Narrative Review on Nutrition, Stimulants, and Physical Activity as Important Factors. *Nutrients*, *14*(9), 1912. https://doi.org/10.3390/nu14091912

Shafeeq, A. (2012). *Dynamic Clustering of Data with Modified K-Means Algorithm*. https://doi.org/10.13140/2.1.4972.3840

Shahin, M., Mulaffer, L., Penzel, T., & Ahmed, B. (2018). A Two Stage Approach for the Automatic Detection of Insomnia. *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 466–469. https://doi.org/10.1109/EMBC.2018.8512360

Siegel, J. M. (2003). WHY WE SLEEP. *Scientific American*, *289*(5), 92–97. https://doi.org/10.1038/scientificamerican1103-92

Singh, K., & Upadhyaya, D. S. (2012). *Outlier Detection: Applications And Techniques*. *9*(1).

Sletten, T. L., Magee, M., Murray, J. M., Gordon, C. J., Lovato, N., Kennaway, D. J., Gwini, S. M., Bartlett, D. J., Lockley, S. W., Lack, L. C., Grunstein, R. R., Rajaratnam, S. M. W., & for the Delayed Sleep on Melatonin (DelSoM) Study Group. (2018). Efficacy of melatonin with behavioural sleep-wake scheduling for delayed sleep-wake phase disorder: A double-blind, randomised clinical trial. *PLOS Medicine*, *15*(6), e1002587. https://doi.org/10.1371/journal.pmed.1002587

Slyusarenko, K., & Fedorin, I. (2020). Smart alarm based on sleep stages prediction. *2020 42nd Annual International Conference of the IEEE Engineering in Medicine*

& Biology Society (EMBC), 4286–4289. https://doi.org/10.1109/EMBC44109.2020.9176320

Sorzano, C. O. S., Vargas, J., & Pascual, A. (n.d.). *A survey of dimensionality reduction techniques*.

Staples, S. (2015). *The relationship between exercise and self-esteem, sleeping patterns, anxiety and energy levels*. http://hdl.handle.net/10788/2503

van der Maaten, L., Postma, E., & Herik, H. (2007). Dimensionality Reduction: A Comparative Review. *Journal of Machine Learning Research - JMLR*, *10*.

Vermeulen, M., Smith, K., Eremin, K., Rayner, G., & Walton, M. (2021). Application of Uniform Manifold Approximation and Projection (UMAP) in spectral imaging of artworks. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, *252*, 119547. https://doi.org/10.1016/j.saa.2021.119547

Walch, O. J., Cochran, A., & Forger, D. B. (2016). A global quantification of “normal” sleep schedules using smartphone data. *Science Advances*, *2*(5), e1501705. https://doi.org/10.1126/sciadv.1501705

Waldhauser, F., Ková>cs, J., & Reiter, E. (1998). Age-related changes in melatonin levels in humans and its potential consequences for sleep disorders. *Experimental Gerontology*, *33*(7–8), 759–772. https://doi.org/10.1016/S0531-5565(98)00054-0

Wang, W.-T., Wu, Y.-L., Tang, C.-Y., & Hor, M.-K. (2015). Adaptive density-based spatial clustering of applications with noise (DBSCAN) according to data. *2015 International Conference on Machine Learning and Cybernetics (ICMLC)*, 445–451. https://doi.org/10.1109/ICMLC.2015.7340962

Xu, F., Adams, S. K., Cohen, S. A., Earp, J. E., & Greaney, M. L. (2019). Relationship between Physical Activity, Screen Time, and Sleep Quantity and Quality in US

Adolescents Aged 16–19. *International Journal of Environmental Research and Public Health*, *16*(9), 1524. https://doi.org/10.3390/ijerph16091524

Zhang, M., Wang, Q., Pu, L., Tang, H., Chen, M., Wang, X., Li, Z., Zhao, D., & Xiong, Z. (2023). Light Therapy to Improve Sleep Quality in Older Adults Living in Residential Long-Term Care: A Systematic Review. *Journal of the American Medical Directors Association*, *24*(1), 65-74.e1. https://doi.org/10.1016/j.jamda.2022.10.008