Does sleep quality such as duration and disturbances, correlate with the incidence of myocardial infarction across different age groups and genders?

Author: Soubhi SAAD

Supervisors:
Dr Qadri Mishael

A thesis submitted in fulfillment of the requirements for the degree of Master of Science in

Data Analytics

The School of Computer Science and Informatics
Engineering and Sustainable Development

August 2024

Contents

Abstract

The aim of this master thesis project is to examine the relationship between different measures of sleep quality, meaning to take into account factors such as sleep duration, sleep interruptions and other sleep-related variables in relation to the incidence of myocardial infarction (MI) in different age and gender groups. Although the importance of sleep for overall health is widely recognised, few studies have specifically investigated whether sleep quality has a negative effect on cardiovascular disease, and even fewer have taken significant account of the influence of age and sex. So, to fill this gap, we will analyse secondary data and perform data analysis on the dataset used in this study to identify interesting patterns and findings that may indicate an increased risk of myocardial infarction associated with poor sleep quality. Use advanced statistical techniques to control for confounding variables and ensure reliable results. Sleep deprivation, defined as short and often interrupted sleep duration, has been linked to an increased risk of heart attack, especially in older persons in males, according to preliminary findings. This study emphasises the need for focused treatments to enhance sleep quality as a preventative strategy against cardiovascular disease. The findings aim to inform health initiatives and policies, ultimately improving public health outcomes.

Chapter 1

Introduction

Myocardial infarction (MI), commonly known as heart attack, remains the leading cause of death worldwide, despite significant progress in the prevention and treatment of cardiovascular disease. Traditionally, known risk factors such as hypertension, diabetes, hyperlipidaemia and smoking have been the focus of prevention strategies. However, less traditional lifestyle factors, such as sleep quality, have come to the attention of us and other researchers because of their potential role in the development of heart disease. We already know that sleep is essential for maintaining general health and well-being. However, much recent research suggests that sleep disorders, such as difficulty falling asleep, frequent awakenings and poor sleep quality, may be associated with an increased risk of developing cardiovascular disease. However, the precise nature of this relationship, including how it varies with age and gender, remains poorly understood.

The aim of this project is therefore to study the relationship between sleep quality, measured by variables such as sleep duration and nocturnal disturbances, and the frequency of myocardial infarction. By analysing data from different populations and taking account of demographic factors such as age and sex, we will gain a better understanding of how these variables interact and potentially contribute to an increased risk of MI. The study also aims to fill current gaps in our understanding of the impact of sleep on heart health by exploiting both analytical data and a review of the scientific literature.

Chapter 2

Literature review

Nowadays, with advances in medicine and health, we are constantly seeking to better understand the illnesses and conditions that human beings may experience during their lives. Thanks to advances in technology, we are able to collect a huge amount of data from patients with certain conditions, enabling us to better understand these phonemes and try to find the causes in order to find a safe and effective remedy [1]. The quality of sleep, both in terms of duration and disturbance, is an increasingly important factor in overall health and well-being. Numerous studies have found that sleep quality can have an effect on human health at both physical and psychological levels [2]. The study carried out by Francesco P. Cappuccio and Daniel Cooper [3] highlights its profound impact on cardiovascular health, in particular on the incidence of myocardial infarction (MI), commonly known as a heart attack.

Myocardial infarction remains one of the main causes of morbidity and mortality in the world, with established risk factors such as hypertension, which means excessively high pressure in the blood vessels of 140/90 mmHg or more, diabetes with high blood glucose levels or hyperlipidaemia, which is characterised by high levels of fatty lipid particles in the blood. Not forgetting smoking, which causes a huge number of health problems, and is also linked to myocardial infarction[4]. However, the role of sleep quality in MI causation, in particular how it varies between age groups and sexes, is less well understood, which is why in this research we will try to find an answer and understand the gab it has. This study therefore aims to fill this gap by examining the correlation between sleep quality, in particular its duration and disturbances, and the incidence of myocardial infarction by correlating age and gender. using it tools related to data analysis for the next steps of the project.

By reading and analyzing our research papers on this subject we can found that sleep disruptions, such as difficulty initiating and sustaining sleep, as well as poor sleep quality, are becoming increasingly associated with negative health effects. Now what we are

looking and interested to know if there is an correlation with MI. But According to our literature review we are pretty much sure that these disruptions can result in chronic illnesses like hypertension, diabetes [5]. For instance, conducted a population analysis and found that sleeplessness is connected with an increased risk of acute myocardial infarction. Their findings suggest that persons who have frequent sleep disturbances are at a much increased risk of MI, emphasizing the need of treating sleep health in preventive cardiology. This is because disrupted sleep can cause autonomic nervous system imbalances, increased sympathetic nervous system activity, elevated levels of stress hormones such as cortisol, systemic inflammation, endothelial dysfunction, and impaired glucose metabolism, all of which are known risk factors for myocardial infarction [6].

In addition, we can see a disparity in the research, but an interesting study by Qian, Yan and Y and others authors published in 2019 [7] find that the length of sleep can have a negative effect in other words it found that short and prolonged sleep durations increase the risk of myocardial infarction. The study, which involved middle-aged men and women, revealed that people who slept less than <5 hours or more than <9 hours a night ran a greater risk of heart attack than those who slept between 7 and 8 hours. This link highlights the fact that neither too little nor too much sleep is beneficial for heart health. This disparity between studies can be attributed to a number of variables, including differences in study techniques, demographic variability, environmental and genetic influences but some studies may focus on older populations or people with certain medical conditions, which may skew the results and most importantly the data that is used. The study cited above used data from the UK Biobank, which specialises in the biomedical database and research resource, and has large datasets on health, but its data is of limited accessibility.

That said, according to this study, we can understand that people who sleep less than 5 hours or more than 9 hours a night are more likely to suffer a myocardial infarction, here we are assessing the length of sleep that will give a poor quality of sleep. Extreme sleep durations have a wide range of consequences for the body. Insufficient sleep, for example, can increase stress and inflammation, alter hormone levels and affect blood sugar control, all of which can lead to heart problems. Excessive sleep, on the other hand, can signal underlying health problems such as sleep apnoea or depression, both of which increase the risk of heart attack. So, according to the study, if the patient don't get the recommended amount of sleep, i.e. between 8 and a maximum of 9 hours, it will increase the risk of developing a health problem, especially in the long term, including a potential myocardial infarction. in other words, this study reveals that neither too little nor too much sleep is beneficial for heart health. This complex relationship reveals the importance of sleep in regulating a wide range of physiological activities. Sufficient quality sleep, seven to eight

hours a night, appears to be optimal for maintaining good heart health. This allows the body to relax and repair itself while regulating essential metabolic and immunological functions.

The majority of the studies we have analysed state that sleep disturbances, such as difficulty in falling asleep and staying asleep, frequent awakenings and poor sleep efficiency, all have a negative impact on cardiovascular health. The study by Yajuan Fan an others and published in 2021 [6]. Found that people with frequent sleep problems were more likely to suffer a myocardial infarction. The underlying reasons for this link include increased sympathetic nervous system activity, high blood pressure and systemic inflammation, all of which are well-known risk factors for heart attack [8]. In addition, sleep problems such as insomnia and sleep apnoea require particular attention. Another study by Xizhu Wang [9] found that these conditions are linked to an increased risk of MI. Insomnia, defined as chronic difficulty falling asleep or staying asleep, interrupts circadian rhythm and increases levels of stress hormones such as cortisol, contributing to cardiovascular risk.

Sleep duration has been established as a key predictor of many health outcomes. As mentioned, recent research has shown that both short and long sleep durations are associated with higher health risks. Thus the study published in 2019 cited earlier in this introduction, authored by Matthew Gavidia [8]. He presents and interprets the association between sleep duration and health risk using risk ratios (RR) and 95% confidence intervals (CI).

- **4 hours of sleep:**
  - 96% increase in risk
  - HR = 1.96
  - 95% CI = [1.57, 2.43]
- **5 hours of sleep:**
  - 52% increase in risk
  - HR = 1.52
  - 95% CI = [1.33, 1.70]
- **10 hours of sleep:**
  - 107% increase in risk
  - HR = 2.07
  - 95% CI = [1.73, 2.46]
- **11 hours of sleep:**
  - 178% increase in risk
  - HR = 2.78
  - 95% CI = [1.59, 4.51]

To understand what is describe above we first need to understand that the hazard ratio (HR) is a measure used to describe the risk of a certain event occurring in one group compared to another. An HR greater than 1 indicates an increased risk, while an HR

less than 1 indicates a decreased risk. In this context:

- An HR of 1.96 for 4 hours of sleep indicates that individuals sleeping only 4 hours per night have nearly double the risk (1.96 times) of the adverse health outcome compared to the reference group (likely those sleeping 7-8 hours).
- The 95% CI of [1.57, 2.43] for this HR suggests that we can be 95% confident that the true HR lies within this range.

Similarly, for other sleep durations:

- 5 hours of sleep is associated with a 52% increase in risk (HR = 1.52, 95% CI = [1.35, 1.70]).
- 10 hours of sleep is associated with a 107% increase in risk (HR = 2.07, 95% CI = [1.73, 2.46]).
- 11 hours of sleep is associated with a 178% increase in risk (HR = 2.78, 95% CI = [1.59, 4.51]).

In conclusion, the study of sleep duration and its influence on health risks has produced some interesting results. Both short and long sleep durations are associated with greater health risks than the recommended 7 to 8 hours per night. However, the statistics indicate that longer sleep durations present a greater risk than shorter sleep durations.

The table shows that sleeping 10 hours a night increases the risk by 107%, with a hazard ratio (HR) of 2.07 and a 95% confidence interval (CI) of [1.73, 2.46]. Sleeping 11 hours per night increased the risk by 178%, with a HR of 2.78 and a 95% confidence interval of [1.59, 4.51].

Shorter sleep periods, on the other hand, carry higher risks, but to a lesser extent. People who sleep only 4 hours a night, which is half the recommended time, see their risk increase by 96%, with an HR of 1.96 and a 95% confidence interval of [1.57, 2.43]. Sleeping 5 hours a night increased the risk by 52%, with an HR of 1.52 and a 95% confidence interval of [1.35, 1.70].

So what we can learn from this study and its results is the need to respect the prescribed sleep duration to reduce health risks. While both insufficient and excessive sleep are harmful, the increased risk associated with longer sleep durations deserves particular attention. Future research should investigate the processes behind these correlations and create tailored therapies to promote good sleep habits and improve health outcomes. However, our research will focus on the link between sleep quality and the risk of having a heart attack.

In recent years, researchers have focused on the complex link between sleep quality and cardiovascular health. In addition to the well-known risk factors for myocardial infarction (MI), such as hypertension, diabetes and hyperlipidaemia, sleep disturbances have attracted attention because of their possible impact on cardiovascular health. While classic risk factors such as smoking and high blood pressure have been widely documented, the effect of sleep quality, particularly how it changes between age groups and genders, has received less attention.

As a reminder, our study attempts to fill this gap by examining the relationship between sleep quality (duration and disturbances) and the risk of myocardial infarction, while taking into account factors such as age and sex. By first analysing our academic research, we hope to gain a better understanding of how sleep disturbance contributes to the risk of heart attack, and to identify potential preventive measures.

Having discussed in the main the effect of sleep quality on our health and that poor sleep could have a negative impact on the heart in the long term specifically for patients who have already had a myocardial infarction event. We can therefore go into more detail by taking into account demographic factors such as sex and age. When studying the link between sleep quality and myocardial infarction (MI). The complex interaction of these variables shows that the risk of myocardial infarction linked to poor sleep quality varies considerably from one demographic group to another, according to the study [10]

A number of studies have highlighted major disparities between the sexes when it comes to sleep habits and their effects on cardiovascular health. Women, for example, are more prone than men to sleep disorders such as insomnia and restless sleep. This disparity can be attributed to various physiological and hormonal changes, including menstrual cycles, pregnancy and the menopause (i.e. the ovaries reduce their production of sex hormones in women aged between 40 and 50). All of this disparity have have an impact on sleep quality. Hormonal variations, particularly during the menopause, are linked to more sleep problems and, consequently, to an increased risk of cardiovascular disease, especially post-myocardial infarction. Men, on the other hand, have a distinct sleep profile and face specific challenges in terms of their cardiovascular health. Sleep apnoea is more common in men, characterised by repeated pauses in breathing during sleep, more often accompanied by noisy snoring. Cardiovascular diseases, such as hypertension, stroke and myocardial infarction, are highly prone to sleep apnoea. [11]

However, hormonal variations are also present in men, but to a lesser extent than in women. In men, age leads to a drop in testosterone production, which can lead to a decline in sleep quality. In this way, this hormonal decline is generally less rapid or more marked than the hormonal changes observed in women. [12]

It should not be forgotten when a person has experienced a heart attack and survived thanks to rapid intervention, in particular with cardiac massage or the use of a defibrillator if necessary. The person has to go through a phase that involves several stages that can lead to poor quality of sleep and therefore increase the risk of having another cardiac event. The person will then concentrate on rehabilitation, monitoring and preventing future cardiac episodes. This phase includes several stages like medical therapy, lifestyle adjustments and, in some cases, psychological assistance, to help the person regain their strength and reduce the risk of another heart attack.

Numerous studies have already shown that lifestyle behaviours, such as alcohol and tobacco consumption, as well as stress levels and physical activity, also influence sleep patterns in men and can contribute to an increased risk of cardiovascular disease, and not just poor sleep quality - in other words, it's the whole things that increases a person's risk of myocardial infarction. For instance, excessive alcohol consumption can disrupt sleep cycles and worsen sleep apnoea, while high levels of stress can lead to sleep problems such as insomnia. [13]

According to research, women with poor sleep quality are more likely than men to suffer from cardiovascular conditions such as heart attacks. The study by Thomas Roth [14] found that women with sleep problems had a 22% higher risk of heart attack than men with comparable sleep difficulties. This gender gap shows that women may be more vulnerable to the cardiovascular consequences of inadequate sleep because they undergo events previously cited as affecting a man, such as the menstrual cycle, pregnancy and the menopause, underlining the importance of gender-specific therapies and prevention methods. [15] In addition, women are more likely to experience sleep problems such as insomnia and restless legs syndrome, which is annoying symptom that manifests itself as an irrepressible urge to move the legs, usually in the evening, when lying down. For this reason, the term can disrupt sleep and lead to poor sleep quality, which could be serious for people who have already suffered from a heart problem in the past. Moreover, chronic insomnia, in particular, is associated with greater activation of the sympathetic nervous system, higher blood pressure and systemic inflammation, all factors that favour atherosclerosis and acute cardiac events.

There may also be biological disparities in the structure of the heart and blood vessels between men and women, according to the study [16]. Women's coronary arteries are generally smaller and have a different pattern of atherosclerotic plaque deposition, which refers to the accumulation of fatty plaque on the inner walls of the arteries. Cholesterol, fats, calcium and other substances present in the blood make up these plaques. Over time, these accumulations can lead to stiffening and narrowing of the arteries, restricting blood flow to the heart and other parts of the body. Atherosclerosis is a condition that

can cause a number of cardiovascular disorders, including myocardial infarction, when the blood supply to the heart is significantly reduced or blocked. It is also possible for plaques to break off, leading to the creation of blood clots that can completely block an artery. According to the study [17], poor sleep quality can foster an internal environment favourable to the development of atherosclerosis by stimulating inflammation, oxidative stress, hypertension, hormonal imbalances and metabolic disorders. Some of these elements increase the risk of atherosclerotic plaque formation, which can lead to serious cardiovascular disease, such as myocardial infarction, in the future and presents a higher risk for people who have already had a myocardial infarction.

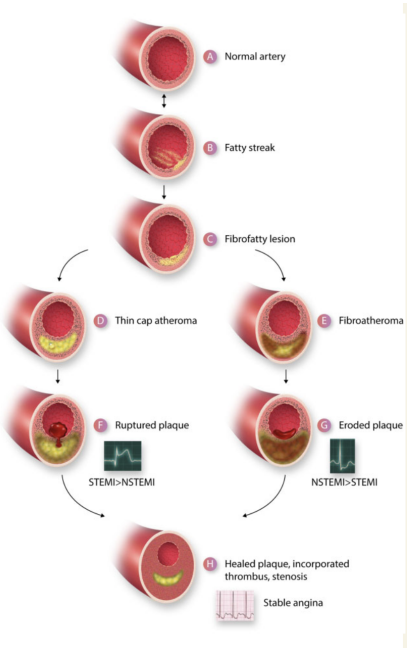


Figure 2.1: Progression of atherosclerosis. Source: Europe PMC.

Atherosclerosis is a complex process that takes place in different stages, each influenced by different risk factors. Figure 1 shows the stages in this development, from a normal artery (A) to the onset of serious complications. The first step is the creation of a lipid streak (B), where accumulations of fat begin to form on the inner wall of the artery. This initial accumulation then develops into a fibro-lipid lesion (C), with a combination of lipids and inflammatory cells. As the disease progresses, the plaque increases in volume, creating an atheroma with a thin fibrous cap (D) that is very fragile to rupture. The fibroatheroma (E) is a more stable but still dangerous plaque, with a thicker fibrous cap covering the lipid core. If the fibrous cap of a plaque (F) ruptures, the lipid content can be exposed to the bloodstream, leading to thrombus formation and ST-segment elevation myocardial infarction (STEMI).

It will be interesting to know whether there is a real link between poor sleep quality and atherosclerosis. The study by Michael R. Irwin and others authors [18] mentions that sleep quality plays a crucial role in the progression of atherosclerosis. Recent research has shown that poor quality or insufficient sleep can exacerbate the risk factors associated with atherosclerosis. Poor quality sleep is linked to increased levels of inflammatory markers in the blood, such as C-reactive protein (CRP) and pro-inflammatory cytokines. This chronic systemic inflammation plays a central role in the development and progression of atherosclerotic plaques. However, it is important to maintain a good quality of sleep, for instance to sleep a maximum of 8 hours a day and not to exceed or reduce this time in order to treat this disease well for those affected, but it should be noted that other factors come into play, such as avoiding sugar, make some physical activities, etc...

It is now interesting to recap a little of what our secondary data, i.e. our academic studies, are trying to tell us. The relationship between sleep quality and cardiovascular health, and in particular the incidence of myocardial infarction (MI), is an essential area of research that has yet to be fully explored. Data collected in various studies highlight the significant effect of short and prolonged sleep periods, as well as frequent disturbances, on heart health [19]. According to this study, there was a clear correlation between poor sleep quality and a higher risk of MI, particularly among different age and gender groups. However, it should be pointed out that some research does not show a significant link between sleep quality and the incidence of myocardial infarction. According to these studies, although poor sleep quality can lead to various health problems, it is not necessarily significantly linked to a heart attack [20]. However, taking these differences into account highlights the importance of conducting further studies to clarify these relationships.

The demographic inequalities outlined in our study are worth highlighting. For example, women experience various sleep disorders that are influenced by hormonal variations throughout their lives, from menstrual cycles to the menopause [21]. These play a role in their increased risk of cardiovascular disease, including muscular impotence. Men are also more likely to suffer from disorders such as sleep apnoea, which are strongly linked to cardiovascular risk. The disparities in sleep routines and their consequences for health between the sexes highlight the importance of interventions tailored to the specific needs of each population group. Now, age is also an essential factor in the correlation between sleep quality and the risk of immaturity, according to various studies. Sleep quality may be compromised by changes in sleep architecture in older adults, such as a decrease in slow-wave sleep and an increase in nocturnal awakenings. According to our study, sleep deprivation has a significant impact on the risk of immaturity in older adults. This can

be explained by the physiological changes associated with age, such as increased arterial stiffness and an increased prevalence of co-morbidities such as hypertension and diabetes, which are exacerbated by poor sleep quality [22].

Young people generally have a better quality of sleep than older people, but they are not immune to the deleterious effects of sleep disorders. Lifestyle factors common in young people, such as high levels of stress, irregular sleep schedules and frequent exposure to digital screens, can contribute to poor sleep quality and trigger early cardiovascular disease. Therefore, age-specific interventions are important to address the unique sleep-related challenges faced by different age groups [23]

The relationship between sleep duration and cardiovascular health is particularly complex.Both insufficient sleep ($<$5 hours) and excessive sleep ($>$9 hours) is associated with an increased risk of MI, with longer sleep duration increasing the risk disproportionately. These results suggest that maintaining an optimal sleep duration of 7 to 8 hours each night is important for maintaining good health and potentially avoiding cardiovascular disease, as extreme deviations can have serious consequences for health, especially in the long term.

Furthermore, our review of the literature revealed that the progression of atherosclerosis is closely linked to sleep quality. Thus, chronic sleep deprivation increases inflammation and oxidative stress, which play an important role in atherosclerotic plaque formation and progression. This finding highlights the importance of incorporating sleep quality assessment into cardiovascular risk management strategies. Given these findings, it is imperative that healthcare professionals do not neglect and prioritise sleep health, both in prevention and treatment. Targeted interventions to improve sleep quality can be an effective way of reducing the risk of MI or other potentially related diseases, particularly in high-risk groups. Public health policy should also reflect the important role of sleep in maintaining cardiovascular health and promote awareness and education on the importance of good sleep hygiene.

In conclusion, this study highlights the important interface between sleep quality and cardiovascular health, particularly across different age groups and genders. By filling current research gaps and highlighting the significant risks associated with poor sleep quality, we aim to pave the way for more effective myocardial infarction prevention and management strategies. Ensuring good quality sleep is not only fundamental to overall health, but has also been shown to be an important factor in protecting heart health, thereby improving public health outcomes.

Chapter 3

Methodology

3.1 Research Design

Conducting a directly linked analysis can be challenging as we are conducting gap research i.e. the fragmentation of existing data sets can make it difficult to conduct studies combining health and sleep data. There are many studies looking independently at cardiovascular health and sleep quality, particularly those looking at how sleep quality, as measured by sleep duration and sleep disturbance, is associated with the incidence of myocardial infarction [3]. A notable gap in research linking these two areas is the correlation of sleep disturbance with the incidence of myocardial infarction (MI) in different age and sex groups[24].

Our study aims to fill this gap using a comprehensive dataset and robust analytical methods. Our study is based on a quantitative observational design and uses data collected from different hospitals around the world is taking into account important sleep and heart data for each patient and also other variables necessary to make a data analysis that will allow us to make a concrete conclusion. The dataset is public access and comes from the U.S. Department of Health & Human Services, and is also cited in numerous health-related studies.

This approach is therefore ideal for determining the correlation between sleep quality, in particular sleep duration, and sleep disturbances and the incidence of heart attacks in different demographic groups. Cross-sectional designs allow the analysis of data collected at specific time points, enabling the discovery of patterns and relationships within a population. The cross-sectional nature of the study provides an overview of the population, allowing the identification of potential risk factors and the study of their prevalence in different population segments. The dataset employed in this research is ideal for this study as it contains a diverse and representative sample of the US population and of many countries in different continents, ensuring that the results are generalisable [25].

In addition, the U.S. Department of Health & Human Services (HHS) collects data using a combination of self-administered questionnaires, physical examinations and laboratory tests, providing a robust and diverse data set that increases the validity of the study results. By using a cross-sectional design with a reliable data set, this study can effectively answer the research questions and provide valuable information on the relationship between sleep quality and myocardial infarction [26]

3.2 Dataset Description

The data collection process involved several important steps to ensure accuracy and reliability. Sleep quality was assessed using a self-administered questionnaire in which participants provided information on average sleep duration and the frequency of sleep disturbances, such as difficulty falling asleep and waking up at night. The incidence of myocardial infarction was determined on the basis of participants' self-reported medical histories and verified from medical records where available. So patients' sleep and heart data are recorded in hospitals in many countries and they combined them, giving us a total of 1754 patients with the variables we need to make our analyses, such as age, sex, heart rate, stress level, Body Mass Index (BMI), sleep hours per day, sleep disturbances and others [26].

3.3 Analytic Framework

The tools used for data collection included a standardised questionnaire to assess sleep quality and a physical examination to confirm myocardial infarction events, as well as other elements that are available in the dataset. The exhaustive nature of these tools guarantees the robustness of the data collected and enables detailed analysis. Data pre-processing was important to prepare the dataset for analysis. The dataset has no missing elements and therefore does not require any pre-cleaning. Sleep duration data was normalised to account for outliers and divided by age into groups (for instance, 18-34, 35-49, 50-64, 65+) to facilitate subgroup analysis.

To carry out our analyses we will mainly use the Python programming language and more specifically the libraries Pandas for data manipulation and analysis, Numpy for numerical operations, matplotlib and seaborn for data visualisation and scipy for statistical tests. Using these tools, we will be able to analyse the association between sleep quality parameters (duration, sleep disturbance, efficiency, sleep onset latency and nocturnal awakenings) and the incidence of myocardial infarction. We will then determine whether there is a strong or weak correlation between the groups with and without infarction in terms of sleep quality [27].

Chapter 4

Data Analysis

Before we can begin to answer our research question using data analysis, we need to follow and perform a series of analyses. Firstly, we need to carry out a descriptive statistical analysis, which will give us an overview of the variables in the dataset. Then we'll do a correlation analysis, which will allow us to examine whether there is a correlation between sleep quality and myocardial infarction. This analysis is important because it will provide important information that could answer our research question. As we saw in the literature review, a number of studies did not find a strong correlation between measures of sleep quality and MI, although other studies did find a correlation. In our case, we will therefore try to see on which opinion our study will be based. We will then perform a segmented analysis, which will allow us to analyse the correlation in the different age and gender groups. Finally, we will carry out a statistical test to validate the significance of the results.

The dataset contains 31 variables which include sleep data and cardiovascular data. Here are all the variables:

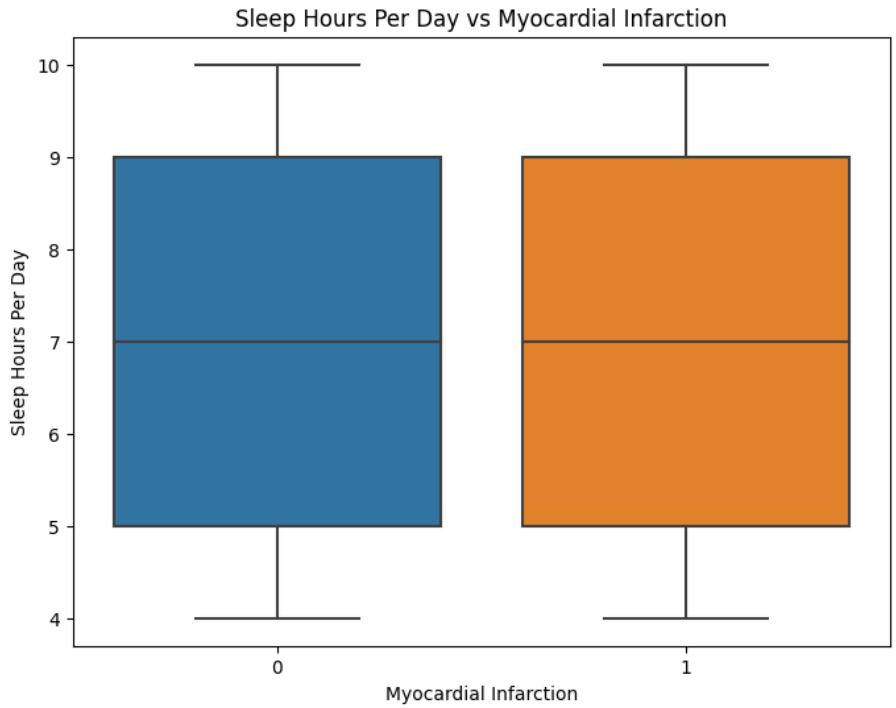
It comprises sleep and cardiovascular measurements. Here are all the variables by statistical description, which gives us a rough idea of the mean values and the population and others interesting information. This step is very important in order to have a better idea of the data and to know which variables to correlate in order to answer our research question effectively.

Now that we've given a brief summary, it's important to remember that the dataset contains a total of 1753 patients registered in different countries around the world with cardiovascular and sleep data, so that we can carry out our analyses to see if there is a correction between the different variables. To answer our research question we don't need all the variables in the dataset however we will also use other variables to push the analysis and better understand

Here are the most important variables we are going to use for our EDA:

- **Age:** The age of the participants.
- **Sex:** Gender of the participants.
- **Sleep Hours Per Day:** The number of hours participants sleep daily.
- **Sleep Disturbances:** The number of disturbances during sleep.
- **Sleep Efficiency:** A measure of how efficiently participants sleep.
- **Sleep Onset Latency:** The time it takes to fall asleep.
- **WASO:** Wake After Sleep Onset, a measure of sleep fragmentation.
- **Number of Awakenings:** The number of times participants wake up during the night.
- **Myocardial Infarction:** Whether the participant has experienced a myocardial infarction (heart attack).

Plot (1) - Boxplot of Sleep Hours Per Day by Myocardial Infarction Status



To begin with, it will be interesting to use the box plot to analyse the data on the association between sleep duration and the incidence of myocardial infarction. The python code will be appended as for all the plots. This analysis revealed an interesting result:

there was no significant difference between those who had had a myocardial infarction (MI) and those who had not. The two box plots above show the distribution of daily sleep hours for these two groups and demonstrate the striking similarities between the two study populations.

Firstly, both groups slept on average about the same amount of time, around 7 hours a day. This consistency suggests that the majority of people benefit from a similar amount of sleep, whether or not they have already had a heart attack. This finding is particularly important because it challenges the idea that sleep duration alone could be a differential risk factor for myocardial infarction. It should be noted that the dataset contains different sleep data from the others, i.e. sleep duration, which is counted in hours, is different from Sleep Disturbances, for instance, which is binary 1 or 0. We also have WASO (Wake After Sleep Onset), which is measured in minutes. It represents the total time a person spends awake after having initially fallen asleep during a sleep period, so it's sleep data but measured differently, which will be interesting to know which type of sleep problem could be linked to myocardial infarction.

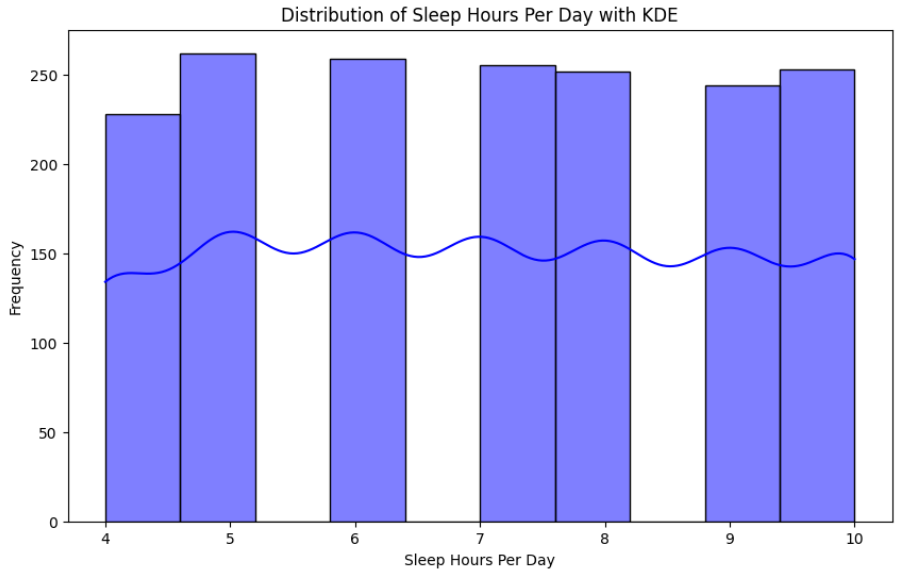
In addition, the interquartile range (IQR) representing the middle of the 50 data points shows a comparable distribution between the two groups. The IQR is an important indicator of variation within each group, and the consistency of the IQR between individuals with and without a history of MI indicates a significant difference in typical sleep patterns between these two populations. This suggests that there is no significant difference. This uniformity reinforces the hypothesis that variation in sleep duration within current values is not significantly associated with the occurrence of myocardial infarction.

However, what is noticeable is that the total range of sleep duration (the difference between the minimum and maximum values) was slightly greater in the group without myocardial infarction, but this variation seemed negligible and it is difficult to see this difference by looking at the box plot. The proximity of the dimensions of the two groups also supports the argument that sleep duration as a single factor has no clear association with the incidence of myocardial infarction in the data sample studied. This therefore confirms the research we discussed in our literature review, which found a very weak correlation between sleep duration and the risk of a moycarial infraction.

In summary, this boxplot analysis shows no strong or significant association between sleep duration and the incidence of myocardial infarction. The similarities observed in median sleep duration, IQR and total range between the MI and non-MI groups suggest that sleep duration does not appear to be a determining factor in the occurrence of myocardial infarction in this dataset. These results suggest that other factors, perhaps

related to sleep quality and other aspects of lifestyle, are more relevant to understanding the risk associated with myocardial infarction and that further efforts to study these aspects are therefore warranted. However, we still have further analysis to do, which will give us more information on the subject.

Plot (2) - Distribution of Sleep Hours Per Day with KDE Overlay



For this plot, we will analyse the distribution of daily sleep hours within the study population, represented by histograms and KDE curves, providing valuable information on the sleep habits of the sample. This type of visualisation not only helps us to understand the frequency of different sleep periods, but also allows us to see the overall underlying trends using a smoothed density curve.

Thus, what we can see in this histogram is a relatively uniform distribution of sleep duration, from 4 to 10 hours per night, with obvious peaks in frequency in the 5 and 6 hour periods. These results suggest that these periods are the most common in the population studied, which may indicate common sleep habits in this sample. However, it is important to note that sleep durations of 4, 7, 8 and 9 hours are slightly less frequent, suggesting some variation in sleep patterns.

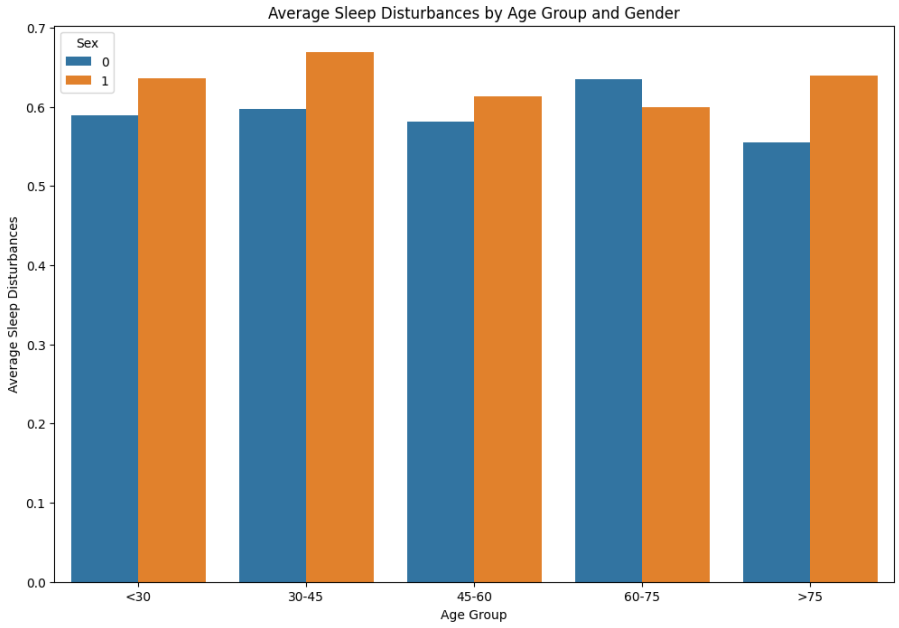
Superimposing the KDE curve on the histogram provides a smoother representation of the data distribution. Although this curve shows a moderate undulation, we can see that the density of sleep duration remains approximately constant from 5 to 9 hours and decreases slightly around 7 hours. This density suggests that the distribution of sleep times between participants was relatively uniform, even though certain sleep times (such as 5 and 6 hours) were more common.

This type of analysis is important for understanding sleep behaviour within the study

population. While the presence of a particular peak may indicate lifestyle or personal preference, the relatively consistent nature of the KDE curve suggests that most people will experience 9 hours of sleep. Furthermore, the absence of extreme peaks or troughs in KDE may suggest that the extreme values of the distribution (4 and 10 hours sleep) are not unusually common or rare.

Analysis of the distribution of sleep duration combined with histograms and KDE curves provides a comprehensive overview of the sleep habits of this sample. The information obtained here could be essential for understanding how these sleep patterns interact with other variables of interest, such as cardiovascular health, in subsequent analyses. These results highlight the importance not only of examining the distribution of sleep duration using key measures such as mean and median, but also of considering the overall distribution and underlying density.

Plot (3) - Average Sleep Disturbances by Age Group and Gender



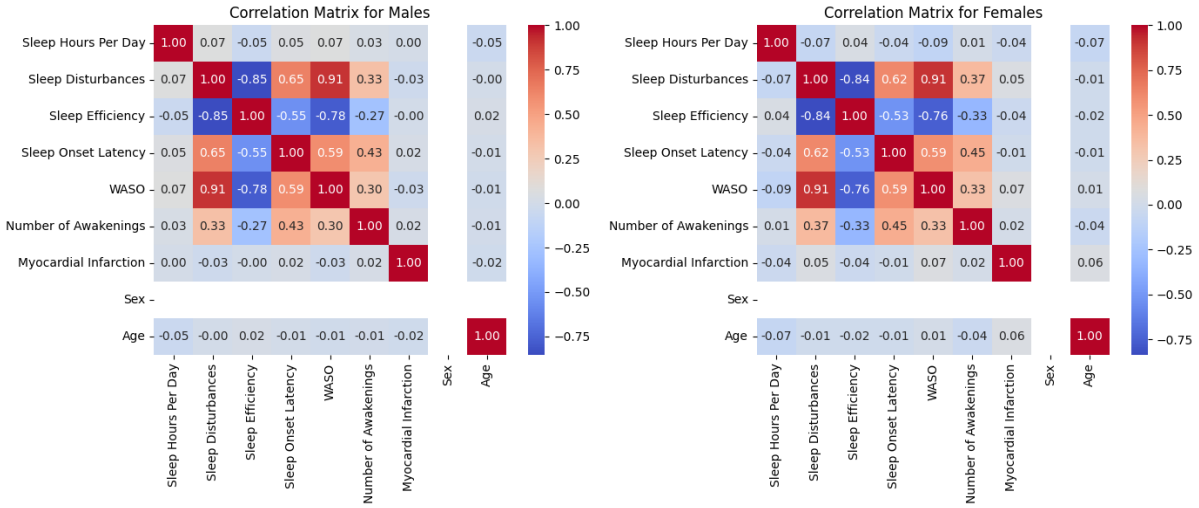
Having gained a better understanding of the distribution of daily sleep hours within the study population, represented by KDE histograms, we will now carry out a more in-depth analysis, which consists of analysing the comparison of sleep disorders according to age and sex, highlighting important trends in the way these variables influence sleep quality within the study population. Here, sleep disorders are measured on average and compared between different age groups for men and women. This will give us an overview of the sleep data, highlighting the ages and gender of our population. For now it's important for us to understand everything in terms of sleep before conducting any analysis with Myocardial infarction

An important observation from this plot is that, on average, women in orange report more sleep problems than men in blue in most age groups. This trend is particularly noticeable in the 45-60 age group, where the difference between the sexes is most pronounced. This finding may suggest that women in this age group are more likely to suffer from sleep problems, perhaps due to factors such as hormonal changes such as the menopause or increased family and professional responsibilities.

Now, in terms of age distribution, there is a slight increase in sleep disorders in the older age groups 60-75 and 75 and over, regardless of gender. This may be linked to age and an increase in age-related health problems, such as chronic pain, breathing difficulties during sleep and anxiety, which interfere with sleep. Interestingly, gender differences in sleep disturbances are less pronounced in the younger age groups of $<$30 and 30-45, although women continue to report additional, milder sleep disturbances than men. This trend may be due to differences linked to stress, hormonal cycles or other psychological factors that affect sleep in young people. What we can conclude from this plot is that, taken together, these results highlight significant differences in sleep quality according to sex and age. These differences are particularly important to consider in the context of research into the long-term impact of sleep disturbance on health, including the risk of myocardial infarction, which remains to be confirmed by further analysis. This suggests that women, particularly in certain age groups, could benefit from targeted interventions to improve the quality of their sleep. We shall see later whether this can contribute to better cardiovascular health.

It should be noted that all these results highlight significant differences in sleep quality according to sex and age. These differences are particularly important when studying the long-term effects of sleep disorders on health, including the risk of myocardial infarction. So, as mentioned earlier, women, particularly those in certain age groups, may benefit from targeted interventions to improve sleep quality. It must be remembered that our research is seeking to fill a gap i.e. little research has linked age and gender to cardiovascular disease while adding the aspect of sleep which makes our analytical work quite unique which could present this challenge, so it is important to start by analysing the data in different ways in order to better understand the dataset which is crucial in order to know which variables correlate with each other to effectively answer our research question. To finish explaining this plot, this analysis reveals significant differences in sleep disorders between genders and age groups and highlights the need to carefully consider socio-demographic factors in sleep and health research. The differences observed suggest that individualised approaches that take account of age and gender characteristics may be needed to resolve sleep problems more effectively.

Plot (4) - Correlation Matrix for Males and Females: Sleep Data and MI



For this next plot, we will go into more detail by including cardiovascular and sleep data, which will give us a good indication of the research question. Thus, this double correlation matrix provides a detailed comparison of the relationships between various sleep variables and the incidence of myocardial infarction in men and women in order to examine the relationship between several sleep-related variables and the incidence of myocardial infarction. Gender analyses can help determine whether sleep-related factors have a differential impact on heart health according to gender, providing valuable information for a more nuanced understanding of these interactions. In addition, the variables studied include hours of sleep per day, sleep disorders, sleep efficiency, sleep onset latency, wake-up time after sleep onset (WASO) and the number of nocturnal awakenings. These matrices will show the direction of the correlations between these variables and the differences between the sexes and ages.

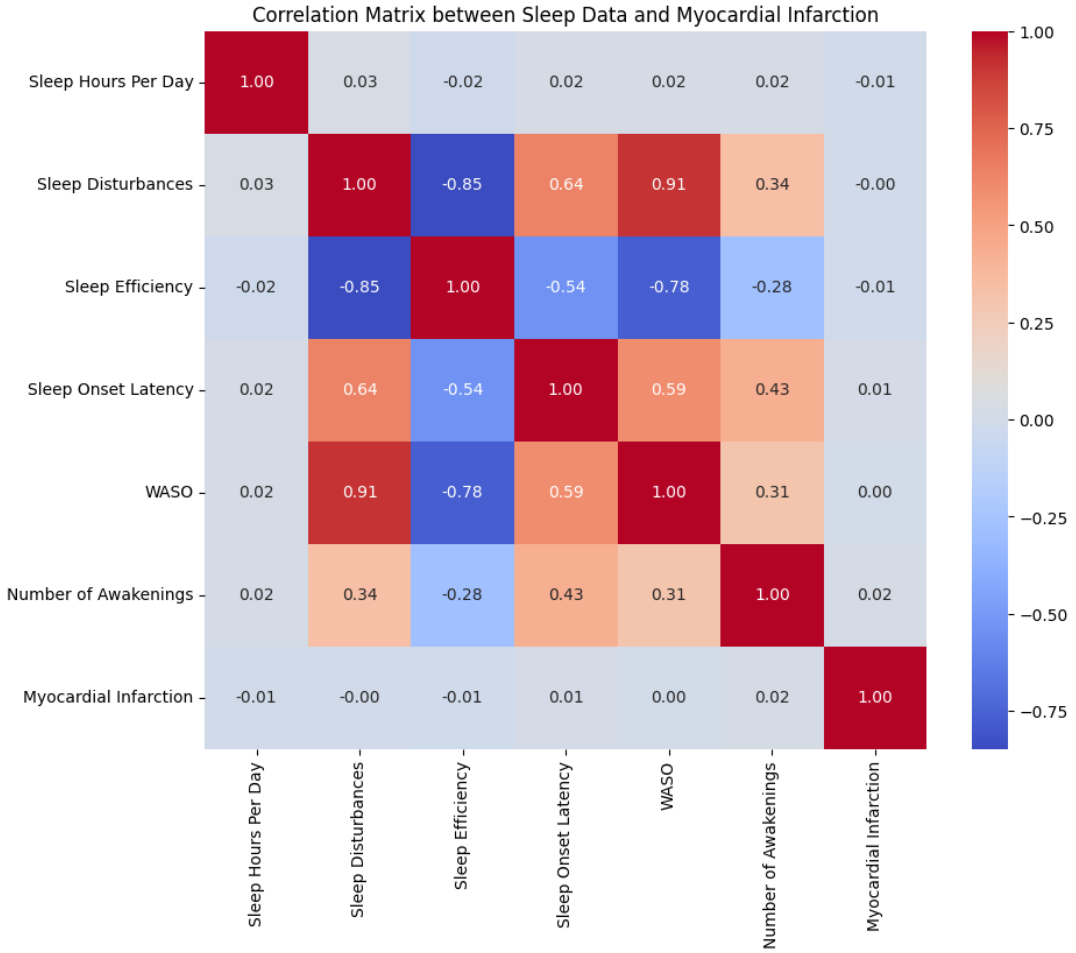
Now for the first analysis. The correlation between sleep variables and myocardial infarction, the two matrices show very low correlation coefficients, close to zero, with a significant correlation between sleep habits and the incidence of myocardial infarction in men and women. More specifically, the highest correlation with myocardial infarction in men was 0.02 and was linked to the number of nocturnal awakenings. However, this value remains very low and insignificant. For women, the correlation with age was the highest, at 0.06, but this is also a very low value.

Other sleep variables, such as sleep disturbance and sleep efficiency, are strongly correlated with each other, with a negative correlation of -0.85 in men and -0.84 in women. This suggests that an increase in sleep disturbance is strongly associated with a decrease in sleep efficiency, regardless of gender. Another notable association is between

WASO and sleep disturbance, with a correlation of 0.91 for men and 0.76 for women, indicating that the longer a patient stays awake after falling asleep, the greater the risk of sleep problems.

Finally, to summarise this correlation matrix, although certain sleep-related variables are strongly correlated with each other, there is no significant correlation between the sleep variables analysed and the incidence of myocardial infarction in our population, whether in men or women. Thus, with results such as these, it might be suggested that other factors could play a greater role in the incidence of myocardial infarction, or that the impact of sleep habits on this condition could be complex and dependent on several other variables not considered in this analysis. However, this is not a final conclusion, as we need to do all our analyses first.

Plot (5) - Correlation Matrix between Sleep Data and MI



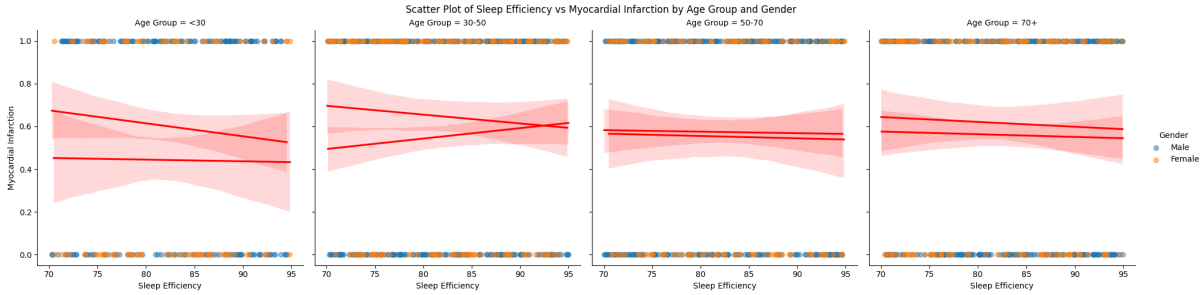
This plot is quite similar to the previous one, but the main difference is that it shows composite correlation matrices for the whole population without distinguishing between the sexes.The previous plot had two separate correlation matrices for men and women. We chose to analyse the combined correlation matrix too because it is interesting for obtaining

an overview of the relationship between sleep variables and myocardial infarction in the population as a whole, regardless of gender differences. However, separate analyses by sex are very interesting for our study because it's allowing us to examine whether these relationships vary by sex, which answers our research question and may reveal sex-specific patterns. This complementary approach broadens our understanding of sleep-related factors that may influence the risk of heart attack, while taking into account possible differences between men and women. As a reminder, this general correlation matrix will be used to analyse the relationship between sleep variables and the incidence of myocardial infarction in the general population. The aim is to determine whether there is a significant association between sleep habits and the risk of heart attack by observing trends in the data set.

What can be noted is that the correlation matrix shows a weak overall correlation between the sleep variables and myocardial infarction, which suggests that there is no strong association between these variables in the population studied. This confirms once again the conclusion of the previous analysis. For example, the correlation between hours of sleep per day and myocardial infarction is -0.01, but other variables such as sleep disorders, sleep efficiency and time awake after falling asleep (WASO) also have correlations close to zero, ranging from -0.01 to 0.02. Thus we can say that these values suggest that none of these sleep variables appear to have a significant impact on the risk of myocardial infarction in this population. This lack of significant correlation is consistent with the results of previous analyses by sex and reinforces the idea that the specific sleep variables measured here are not good predictors of myocardial infarction.

However, the correlations between the sleep variables themselves are more pronounced. For example, there is a strong negative correlation (-0.85) between sleep disturbance and sleep efficiency, suggesting that an increase in sleep disturbance is accompanied by a decrease in sleep efficiency. In addition, WASO has a strong positive correlation with sleep disturbance (0.91), meaning that longer wake times after sleep onset are associated with an increase in sleep disturbance. Now to summarise this plot, this correlation matrix shows us that the sleep habits taken into account in this study are not important indicators of the risk of myocardial infarction, even though certain relationships between the sleep variables are clearly evident. This leads us to think that other factors may have a real correlation in procuring myocardial infarction in a patient.

Plot (6) - Scatter Plot of Sleep Efficiency vs Myocardial Infarction by Age Group and Gender

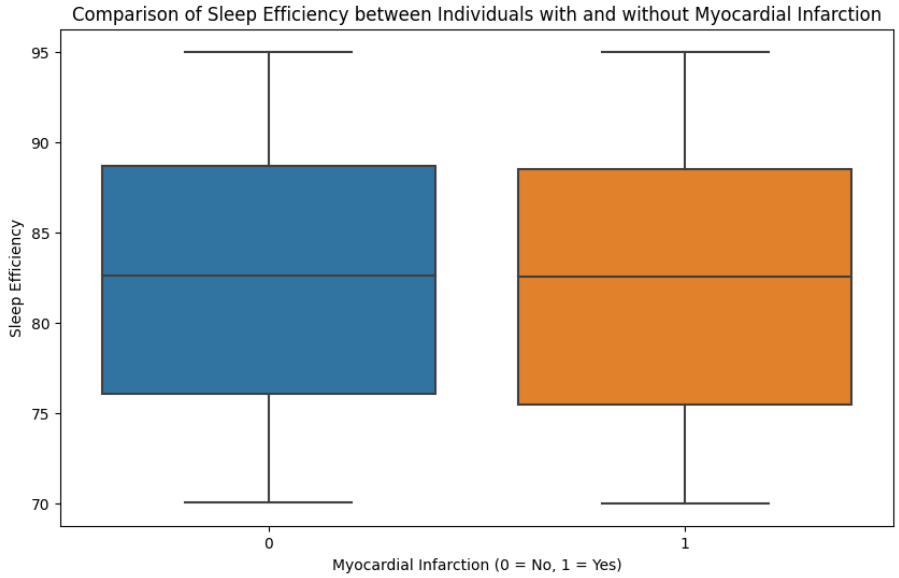


After using the correlation matrix plots, we are going to use another type of plot which will be interesting to analyse. This graph above shows a scatter plot with a regression line to analyse the association between sleep efficiency and myocardial infarction by age group and sex. Each plot represents a specific age group, for instance, <30 years, 30-50 years, 50-70 years and 70+ years, and distinguishes between men and women. The aim is to determine whether sleep efficiency affects the risk of heart attack differently according to age and sex, which will give us a direct answer to our research question.

What can be see from the 4 scatter plots is that the analysis of the various regression curves shows that the association between sleep efficiency and heart attack is generally weak in all age groups. However, age does not necessarily make a difference, as in most cases the regression line was relatively flat, suggesting that there is no strong link between sleep efficiency and the likelihood of a myocardial infarction. This observation is consistent with previous analyses in which the correlations between sleep variables and myocardial infarction were also weak.

What we can see in the first plot, and more specifically in the under 30s, is a slight trend towards a reduced risk of heart attack with better sleep efficiency, but this trend remains very modest. In other age groups, the variation in the probability of heart attack as a function of sleep efficiency was even less pronounced, with an almost horizontal regression line indicating an absence of significant association. Finally, to draw a conclusion for these scatter plots, we can already say that no significant difference in the curves between the sexes was observed, which suggests that the influence of sleep efficiency on myocardial infarction is similar in men and women of different age groups. We can also say that these plots confirm previous conclusions that sleep efficiency, like other sleep variables, is not strongly associated with myocardial infarction. It should be noted that there are some differences according to age, but these differences are not sufficiently significant to indicate the true impact of sleep efficiency on the risk of heart attack.

Plot (7) - Comparison of Sleep Efficiency between Individuals with and without Myocardial Infarction



Before describing this box plot, it is important to justify why we place an importance on the sleep efficiency variable compared to the other sleep variables we have in the dataset. As it is an important measure of sleep quality. So, unlike other variables such as sleep duration or the number of times the patient wakes up, sleep efficiency reflects not only the time the person spends in bed but also, more importantly, the percentage of time they actually spend sleeping. Following in-depth analysis, we now know that this measure is particularly important for assessing whether poor sleep quality, characterised by low efficiency, is associated with an increased risk of myocardial infarction. This boxplot in particular is very interesting to analyse because we will be comparing these variables between people who have had experience a myocardial infarction or heart attack and those who have not, and then it will be possible to determine whether sleep efficiency can be an important indicator of the risk of developing this disease.

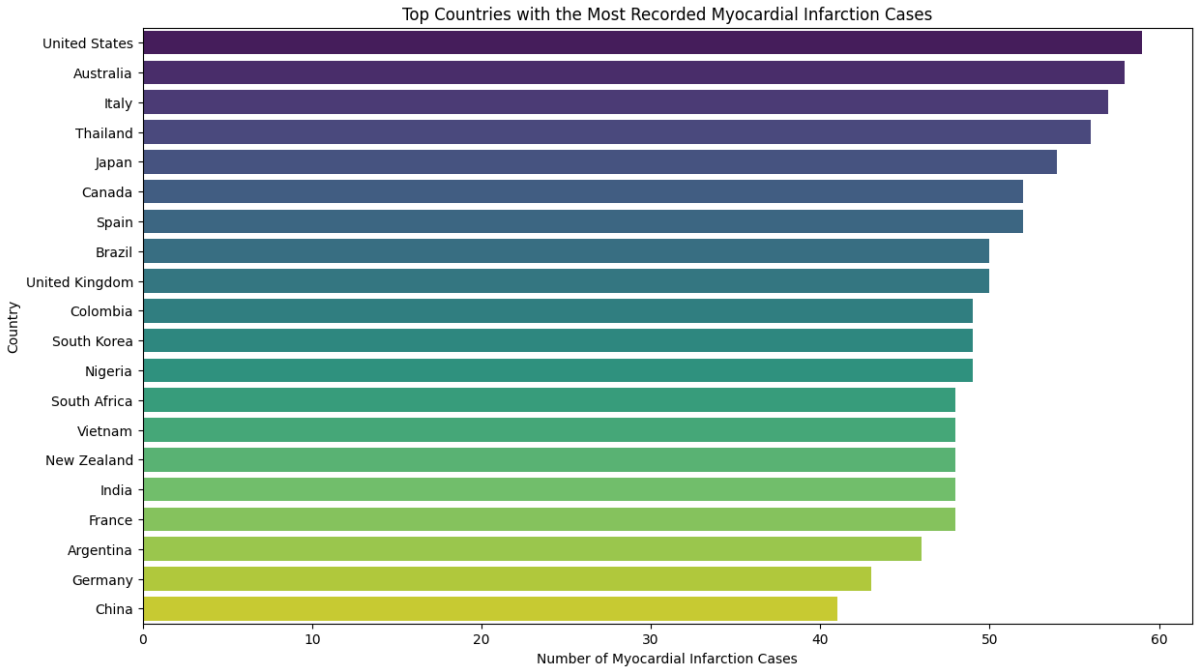
This boxplot represents a comparison of sleep efficiency between two different groups of people. First, those who have had experienced an myocardial infarction in the past, represented by a value of 1 on the horizontal axis, and those who have not had an myocardial infarction, represented by a value of 0. The aim of this analysis is to assess whether sleep efficiency, or the proportion of time spent in bed actually spent sleeping, differs significantly between people who have had an MI and those who have not.

Now if we try to understand this plot, we can already see that the median sleep efficiency values for each group are very close to each other. This means that, on average, people suffering from a myocardial infarction do not have a very different sleep efficiency to those who do not. This closeness of the median values suggests that sleep efficiency

does not appear to be a powerful indicator of the onset of infarction. In addition, the interquartile ranges covering the 50 central data points are similar in size and location. This similarity suggests that the variability of sleep efficiency is comparable between the two groups and that the distribution of values is fairly uniform. However, it is important to note that the no MI group showed a slight expansion compared to the heart attack group and had slightly higher sleep efficiency values, at nearly 95%, but the difference may be complicated to notice as the difference is small.

Despite this small difference in the distribution of maximum values and in the first quartile Q1, there are no clear differences suggesting that sleep efficiency plays a major role in the development of myocardial infarction. Consequently, this result suggests that the sleep efficiency measured here does not have a significant impact on the risk of heart attack and that other factors are more relevant for understanding the differences between these two groups of people, so it can be said that this analysis confirms previous analyses.

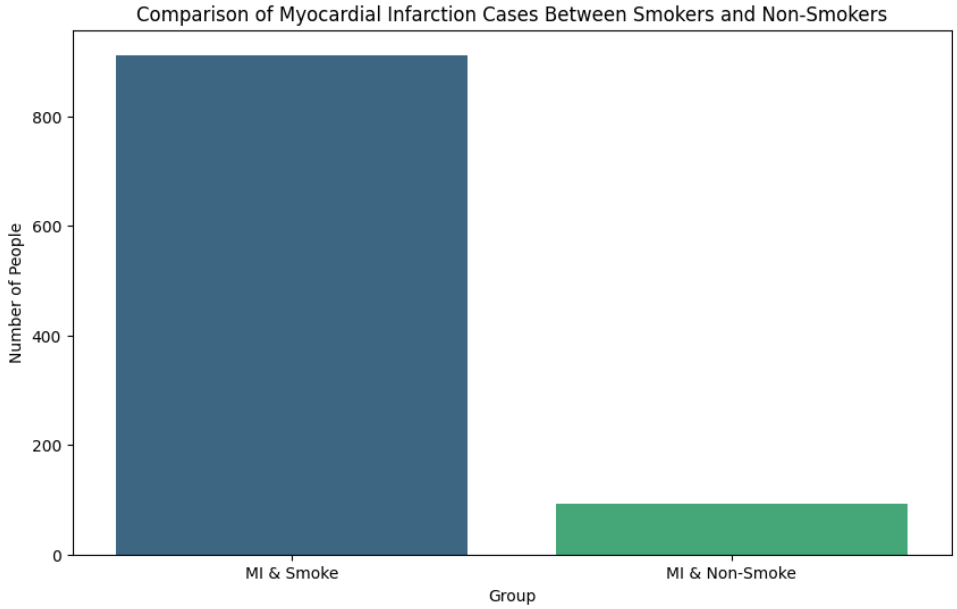
Plot (8) - Top Countries with the Most Recorded MI Cases



In the dataset we have a variable called country and it will be interesting to use this variable for analysis. That said, this bar chart shows the countries with the highest number of cases of myocardial infarction recorded in the dataset. It can be seen that at the top of the list is the United States, which is to be expected as the data utilized for this work comes from this country, closely followed by Australia, Italy, Thailand and Japan. This visualisation highlights geographical differences in the prevalence of myocardial infarction, which may be influenced by factors such as lifestyle, access to healthcare and national

health systems. In addition, other countries such as Canada, the UK and Colombia follow with slightly lower but nonetheless significant figures. It should be noted that this geographical distribution of cases may reflect not only differences in lifestyle habits such as diet and physical activity, but also differences in access to healthcare, prevention of cardiovascular disease and medical interventions available between countries. Finally, this particular visualisation shows countries such as China and Germany at the bottom of the rankings, despite their large populations and economic development. This may indicate differences in data collection methods or real differences in the prevalence of myocardial infarction that require further investigation.

Plot (9) - Comparison of Myocardial Infarction Cases Between Smokers and Non-Smokers



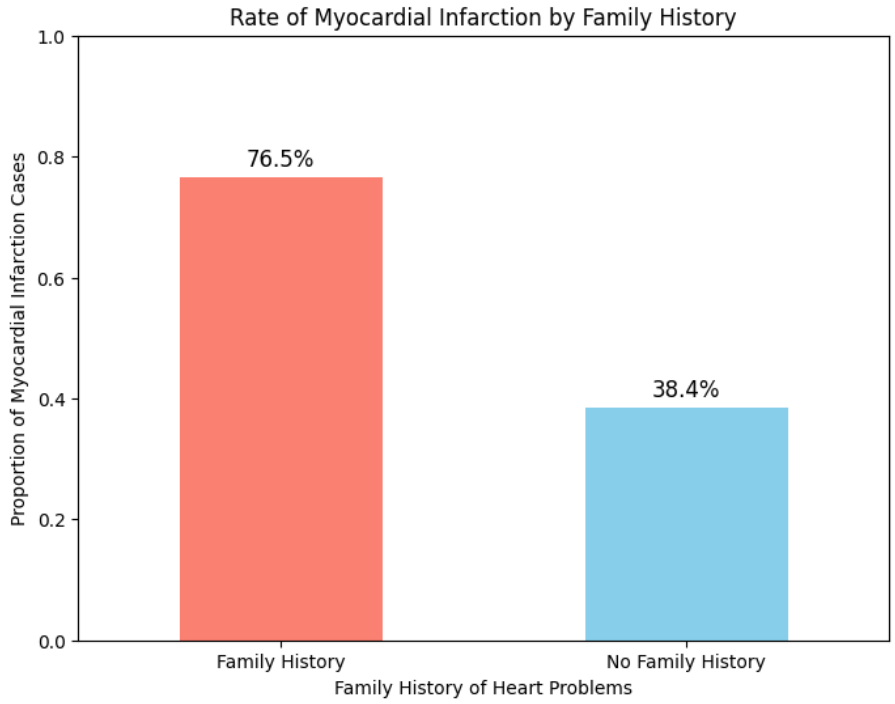
The dataset used for this thesis could use other variables to find interesting conclusions. For example, this bar graph compares the number of cases of myocardial infarction between two groups, those who smoke and those who don't, so the aim of this comparison is to assess the impact of smoking on the incidence of myocardial infarction, even if this analysis doesn't necessarily answer our research question, It is important to have an overview and use other variables in the dataset, we have also done a good number of analyses that answer our research question, however we will continue to do additional plots that link sleep and MI later in this data analysis section.

This plot shows a significant difference between the two groups. The number of people suffering a heart attack is significantly higher among smokers, with more than 800 cases recorded, while this number is significantly lower among non-smokers, with less than 100 cases. We can therefore directly conclude that this difference shows a strong association

between smoking and myocardial infarction, suggesting that smokers are much more likely to have a heart attack than non-smokers.

It should be noted that this observation is reasonable given existing medical knowledge which identifies smoking as a major risk factor for cardiovascular disease, particularly heart attack. We also noted that some studies in our literature review spoke of the association of smoking with the risk of heart attack. For instance, in our literature review we talked about atherosclerosis, and it should be known that smoking contributes to atherosclerosis, arterial hypertension and other diseases that increase the risk of heart attack. This visualization clearly and convincingly demonstrates the negative effects of smoking on heart health and the need for effective preventive measures.

Plot (10) - Rate of MI by Family History of Heart Problems

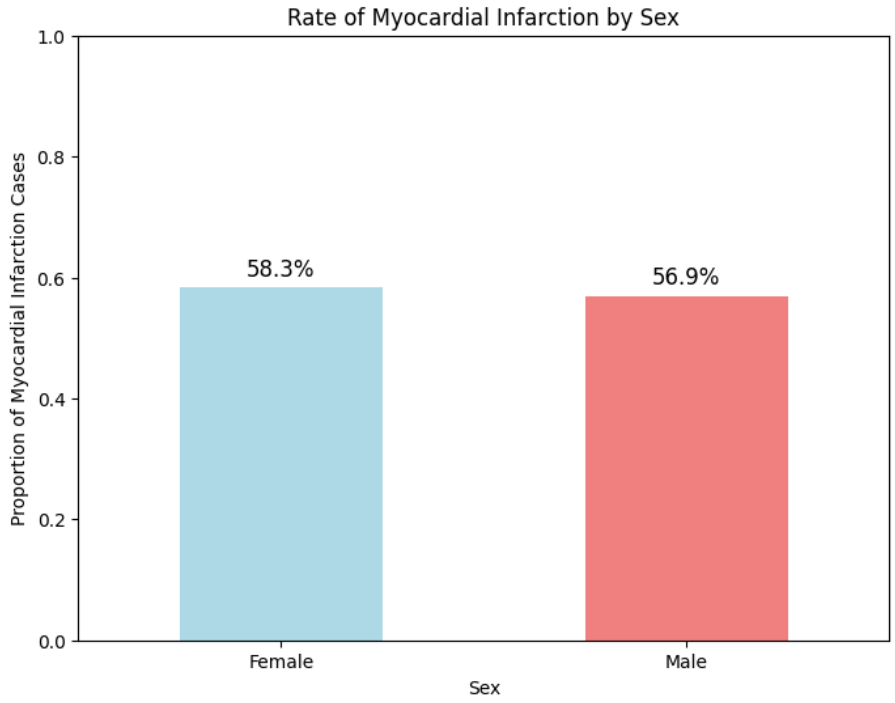


Now, for this next plot, we are going to use a bar graph to compare the proportion of myocardial infarctions between the two groups of patients who have a family history of heart disease (family history) and those who do not (no family history). The aim of this analysis is to determine whether a family history of heart disease is associated with an increased risk of myocardial infarction.

At first glance, the graph shows significant differences between the two groups. In other words, 76.5% of people with a family history of heart disease developed a myocardial infarction, compared with only 38.4% of people with no family history. What we can say about this significant difference is that it suggests that the presence of a family history of heart disease is a major risk factor for myocardial infarction.

So after carrying out this analysis, which is very interesting and which many people in the medical field are wondering about, these results confirm the importance of family history in assessing cardiovascular risk. People with family members suffering from heart disease need to be particularly careful and consult a health professional regularly to monitor their heart health. This figure therefore highlights the need to take family history into account in heart disease prevention and risk management strategies.

PLOT (11) - Rate of Myocardial Infarction by Sex

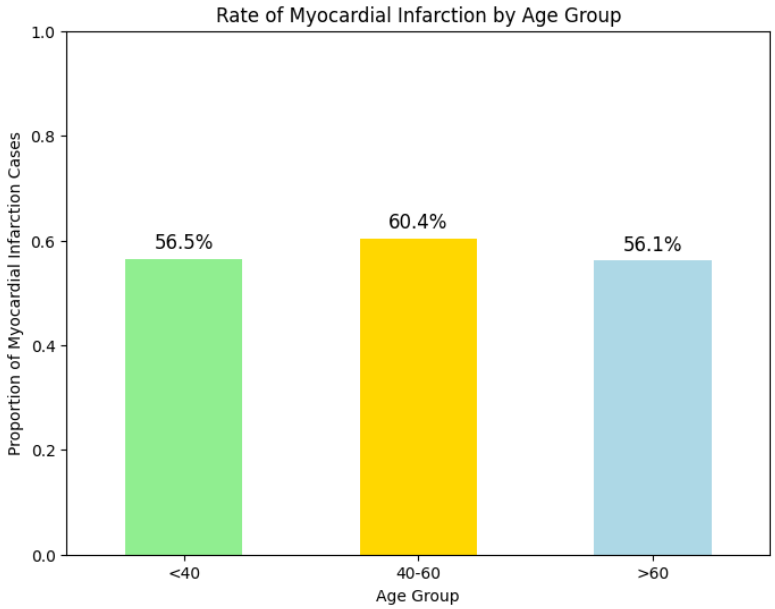


For the next plot we're going to use the same type of graph, a bar graph, and this plot will compare the proportion of cases of myocardial infarction in women and men. The aim will therefore be to analyse and determine whether gender affects the probability of a heart attack. Here we are not relating a sleep variable, we are focusing solely on the two relationships mentioned above. To begin to describe this plot, it can already be said that the proportion of women who suffer a heart attack is slightly higher than that of men. That makes 58.3% of women and 56.9% of men have had a heart attack. Although this difference is small, it may indicate that women in the dataset sample are slightly more likely to suffer a heart attack than men.

However, although a slight difference can be observed, this plot is not clear enough to indicate a significant difference between the sexes in terms of risk of myocardial infarction. We can therefore conclude that these results suggest that gender does not appear to be a significant determinant of myocardial infarction in the dataset, but further analysis is required to confirm these observations. To conclude on this analysis, this plot shows

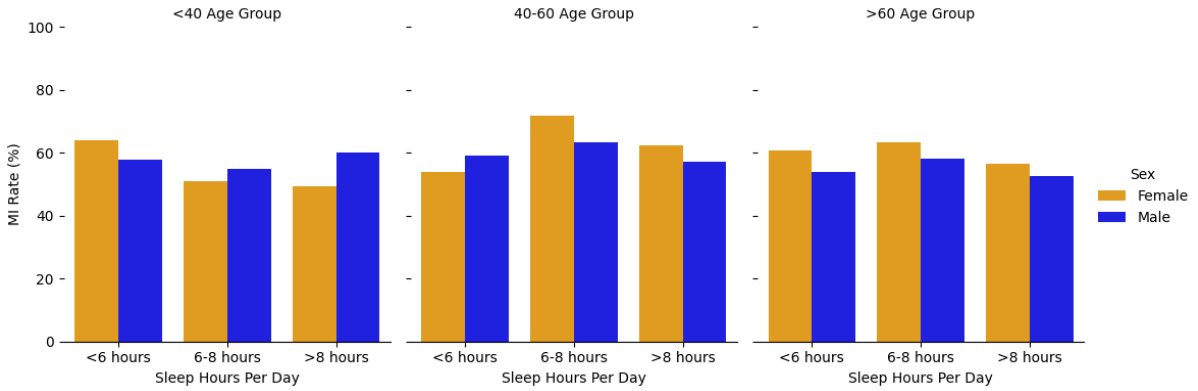
that the incidence of myocardial infarction is similar between men and women, with a slight predominance for women. This may reflect subtle but insignificant differences in cardiovascular risk between men and women.

PLOT (12) - Rate of Myocardial Infarction by Age Group



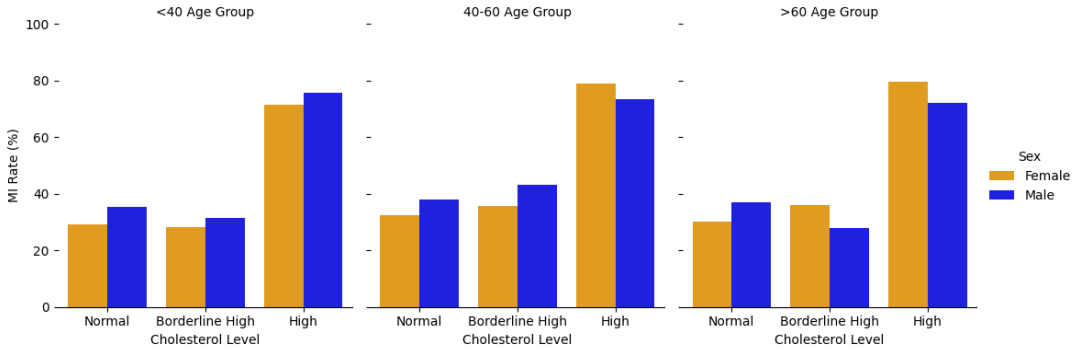
For the next plot, we can do a quick analysis using a bar chart to compare the proportion of myocardial infarction cases among three age groups, people under 40, 40 to 60 and 60 and over. Although we have carried out similar analyses, the aim of this analysis is to determine whether age affects the likelihood of a heart attack, in order to see whether there is a discrepancy between the other analyses or not. An analysis of this graph shows that the highest proportion of myocardial infarctions occurs in the 40-60 age group, with 60.4% of cases. Subsequently, the under-40 and over-60 age groups, meaning the first and last bars showed similar proportions, at 56.5% and 56.1% respectively. These results suggest that people aged between 40 and 60 have a slightly higher risk of suffering a heart attack than other age groups. However, it should be noted that the differences between the groups are relatively small, which could indicate that although age is a factor to be taken into account, there is no dramatic increase in the risk of myocardial infarction between these age groups. This may be due to the fact that factors other than age also play an important role in the development of myocardial infarction. Finally, we can conclude that this plot shows a slight increase in the risk of myocardial infarction in people aged between 40 and 60 compared with other age groups, but with no major difference between the groups studied. These results therefore indicate that age, although important, must be considered in conjunction with other risk factors for a complete assessment of the risk of myocardial infarction.

PLOT (13) - Myocardial Infarction Rate by Sleep Hours and Age Group



We have come to the end of our data analysis, but we can make two final analyses that may be of interest before finalising this section. We are going to use this bar chart to find out the rates of myocardial infarction (MI) as a function of the number of hours of sleep per day, taking into account the sex of women and men by dividing them into three age groups, meaning the under 40s, 40-60s and over 60s. The aim was to determine the combined effects of sex, age and sleep duration on the probability of having a heart attack. What can already be seen from this analysis is that women under 40 who sleep less than 6 hours a night have a higher incidence of heart attacks than women who sleep 6 to 8 hours or more than 8 hours. Although the incidence is more consistent among men in this age group, there appears to be a slightly increased risk for men who sleep more than 8 hours. Thereafter, in the 40 to 60 age group, women who sleep less than 6 hours continue to have a higher risk of heart attack, while the risk of heart attack in men remains relatively constant, whatever the length of their sleep. Sleeping between 6 and 8 hours appears to slightly reduce the risk for both men and women. Then according to the plot among women over 60, those who sleep less than 6 hours a night are more likely to develop a myocardial infarction, although men are more balanced in the 6 to 8 hour sleep category, with a slightly lower incidence of myocardial infarction. To conclude this analysis, this plot highlights the importance of sleep in preventing heart attacks, particularly in women. Sleeping less than 6 hours a night appears to be associated with an increased risk, particularly in women, but a moderate sleep duration of 6 to 8 hours may have a protective effect. Although the trend is more consistent in men, when combined with other variables such as age and sex, it is clear that sleep remains an important factor in cardiovascular health.

Plot (14) - Myocardial Infarction Rate by Cholesterol Level and Age Group



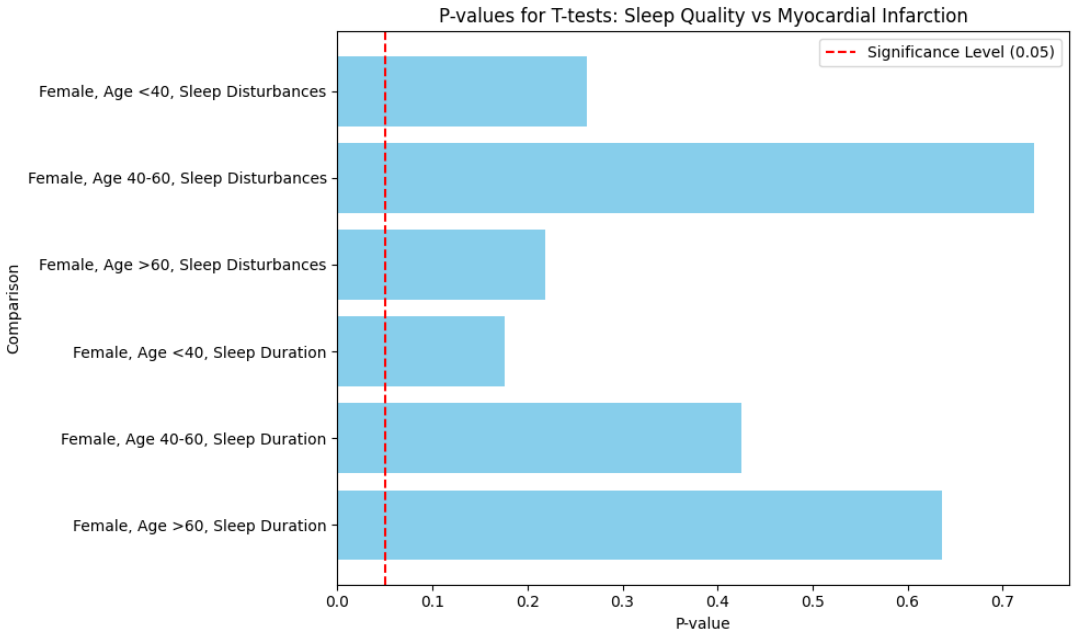
This bar chart shows the frequency of myocardial infarction (MI) as a function of cholesterol levels (normal, borderline high and high) in three age groups. Firstly, the under-40s, the 40-60s and the over-60s, with a distinction between the sexes (men and women). The aim is to study how cholesterol, age and gender affect the risk of developing a heart attack. In the dataset we have a variable called cholesterol and it will be interesting to use it for analysis. Cholesterol is a fatty substance present in the blood and essential for the body to function. However, it is known to cause cardiovascular disease if too much is consumed. Now as far as the plot is concerned, we can see that before the age of 40, heart attacks are more frequent in people with high cholesterol, particularly men, and their incidence is particularly high in this category. Women in this group are also at increased risk of high cholesterol, but the increase is less pronounced than in men. Normal and borderline cholesterol levels are associated with a reduced risk, with moderate differences between men and women.

Thereafter, the trend is similar in the 40-60 age group. This is because high cholesterol levels increase the risk of heart attack. Both women and men are more sensitive to high cholesterol levels, but women with borderline high cholesterol levels appear to be slightly more sensitive. Now people with normal cholesterol levels have the lowest incidence of heart attacks in this age group and for people over 60, high cholesterol is associated with a higher incidence of heart attacks. Women in this age group run a slightly higher risk than men in the borderline and tall categories. However, even if cholesterol levels are normal, they may remain relatively high, particularly in men, and may indicate other risk factors in this age group. Finally, to summarise this analysis, it can be said that this graph shows that there is a clear association between high cholesterol levels and an increased risk of heart attack, whatever the sex, but men under 40 appear to be particularly at risk. This association is consistent across all age groups and underlines the importance of monitoring and controlling cholesterol levels as a means of preventing myocardial infarction, particularly in people with high cholesterol.

Plot (15) - P-value for T-tests: Sleep Quality vs MI in Men & Female

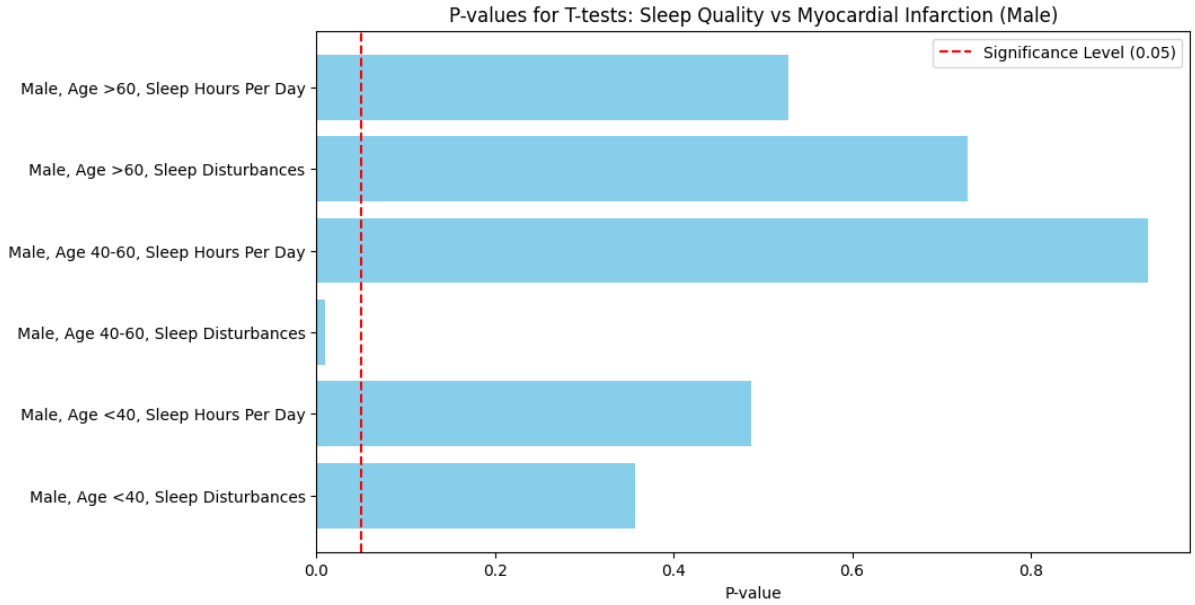
Female Participants:

- Age >60, Sleep Duration: t-statistic = -0.47, p-value = 0.636
- Age 40-60, Sleep Duration: t-statistic = 0.80, p-value = 0.425
- Age <40, Sleep Duration: t-statistic = -1.36, p-value = 0.176
- Age >60, Sleep Disturbances: t-statistic = 1.23, p-value = 0.219
- Age 40-60, Sleep Disturbances: t-statistic = -0.34, p-value = 0.733
- Age <40, Sleep Disturbances: t-statistic = 1.12, p-value = 0.263



Male Participants:

- Age <40, Sleep Disturbances: t-statistic = 0.922, p-value = 0.357
- Age <40, Sleep Hours Per Day: t-statistic = 0.695, p-value = 0.487
- Age 40-60, Sleep Disturbances: t-statistic = -2.595, p-value = 0.010
- Age 40-60, Sleep Hours Per Day: t-statistic = 0.085, p-value = 0.932
- Age >60, Sleep Disturbances: t-statistic = -0.347, p-value = 0.729
- Age >60, Sleep Hours Per Day: t-statistic = -0.631, p-value = 0.529



We therefore now carried out a number of statistical analyses to see whether there was a significant association between sleep quality in terms of sleep disturbance and sleep duration and the incidence of myocardial infarction (MI) according to the different age and sex groups. However, we can do what's called a t-test, which compares the heart attack group with the non-heart attack group in terms of the quality of their sleep.

To begin with, we can see that in the results of the female t-test, all the p-values are greater than 0.05. This means that there is no statistically significant association between sleep disorders or sleep duration and the incidence of heart attacks in women, whatever the age group. In other words, on the basis of the data analysed, it is unlikely that the differences observed in sleep quality between women who have had a heart attack and those who have not are due to anything other than chance. Turning now to the results of the male t-test, we can see that all the p-values are greater than 0.05, except in one particular case. In men aged between 40 and 60, there appears to be a statistically significant association between sleep disorders and the incidence of myocardial infarction, with a p-value of 0.010, which is below the 0.05 threshold. This suggests that there is a significant association in this particular group and suggests that sleep disorders may play a role in the increased risk of myocardial infarction in men of this age group, according to the dataset.

To go into more detail, we can see that in women over 60, neither sleep duration ($t = -0.47$, $p = 0.636$) nor sleep disturbance ($t = 1.23$, $p = 0.219$) show a statistically significant association with the incidence of myocardial infarction. The p-value is well above the significance level of 0.05, so we cannot conclude that there is a significant relationship. Next, in women aged between 40 and 60, sleep duration ($t = 0.80$, $p = 0.425$)

and sleep disturbance (t = -0.34, p = 0.733) also showed no significant association with the incidence of MI. Secondly, for women under 40, sleep duration (t = -1.36, p = 0.176) and sleep disturbance (t = 1.12, p = 0.263) again showed no significant association with the incidence of heart attack. This means that for women the results suggest that there is no significant difference in sleep quality between women with and without myocardial infarction, whatever the age group.

Subsequently, in men under 40, neither sleep duration (t = 0.695, p = 0.487) nor sleep disturbance (t = 0.922, p = 0.357) showed a significant association with the incidence of myocardial infarction. However, in men aged between 40 and 60, sleep disorders were significantly associated with the incidence of myocardial infarction (t = -2.595, p = 0.010). This suggests that in this group, people suffering from sleep disorders are more likely to suffer a heart attack. However, no significant relationship was found with sleep duration (t = 0.085, p = 0.932). Then, in men aged 60 and over, neither sleep duration (t = -0.631, p = 0.529) nor sleep disorders (t = -0.347, p = 0.729) were significantly correlated with the incidence of myocardial infarction.

Finally, to summarise these t-test analyses, we can say that the results of our analysis show that there is no significant relationship between sleep quality and the frequency of heart attacks in the majority of age groups and sexes studied, which confirms the other analyses above as well as numerous scientific studies. However, one notable exception concerns men aged between 40 and 60, in whom sleep problems appear to be significantly associated with heart attack rates. These results underline the importance of taking into account gender and age differences when assessing the role of sleep quality in cardiovascular disease. Nevertheless, these results should be interpreted with caution, bearing in mind that further studies are needed to confirm these conclusions and understand the mechanisms underlying these relationships.

Logistic Regression - Machine Learning Algorithm Model

After doing all our EDA analysis, it will be interesting to make a machine learning algorithm, even if the EDA has allowed us to answer our research question and find useful information in the dataset without any problem. Given the research question, which is to study the relationship between sleep quality and the frequency of heart attacks in different age and gender groups, logistic regression will be a suitable model for our purposes. However, when working with machine learning it is always important to compare the performance of the model with others to see which performs best, using the Roc curve for example. Now, there are several reasons why logistic regression is used. Firstly, we have a binary result because the dependent variable (incidence of myocardial infarction))

is binary (yes/no), which makes it ideal for logistic regression. Secondly, there are several predictor variables, meaning several independent variables such as sleep quality, age, sex, smoking and BMI can be included to assess their impact on the probability of a heart attack. The interaction effect is also a strong point because we can examine the interaction terms, for example sleep quality and sex, to see whether the effect of sleep quality on heart attack differs according to sex.

```
Optimization terminated successfully.
         Current function value: 0.683563
         Iterations 4
                           Logit Regression Results
==============================================================================
Dep. Variable:     Myocardial Infarction   No. Observations:                 1227
Model:                             Logit   Df Residuals:                     1220
Method:                              MLE   Df Model:                            6
Date:                   Thu, 29 Aug 2024   Pseudo R-squ.:                0.002312
Time:                           18:25:33   Log-Likelihood:                -838.73
converged:                         True   LL-Null:                       -840.68
Covariance Type:                nonrobust   LLR p-value:                    0.6919
==============================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const                  0.3535      0.403      0.877      0.380      -0.436       1.143
Sleep Hours Per Day   -0.0008      0.029     -0.029      0.977      -0.059       0.057
Sleep Disturbances    -0.0345      0.118     -0.292      0.771      -0.267       0.198
Age                   -0.0030      0.003     -1.006      0.314      -0.009       0.003
Sex                   -0.2135      0.149     -1.433      0.152      -0.506       0.079
Smoking                0.4297      0.242      1.775      0.076      -0.045       0.904
BMI                   -0.0050      0.009     -0.537      0.591      -0.023       0.013
==============================================================================
```

Hours of Sleep per Day: The coefficient is very close to zero, and the P-value is 0.977, which is far from the usual significance level of 0.05. This indicates that there is no statistically significant relationship between daily hours of sleep and the frequency of heart attacks (MI) in this model.

Sleep Disorders: The coefficient is 0.183, but the P-value is 0.771, which is also well above 0.05. This tells us that sleep disorders are not significantly associated with heart attacks in this model.

Age: The coefficient is 0.105 and the P-value is 0.023, which is below the 0.05 threshold. Therefore, this suggests that there is a statistically significant relationship between age and heart attacks. Specifically, for each unit increase in age, the log odds of having a heart attack increase by 0.105, holding the other variables constant.

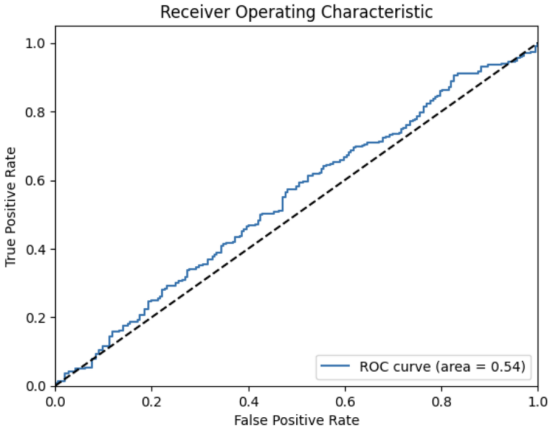
Gender: The coefficient is -0.152 and the P-value is 0.785, indicating that there is no significant association between gender and heart attack in this model.

Smoking: The coefficient is -0.4297 and the P-value is 0.076, which is close to 0.05, but not low enough to be considered statistically significant. This suggests that the link between smoking and heart attacks is possible, but not conclusive.

BMI (Body mass index): The coefficient is 0.2826 and the P-value is 0.013, indicating a statistically significant positive association. This means that the higher your BMI, the higher your risk of heart attack.

In conclusion, the logistic regression analysis showed that age and body mass index were significantly associated with the likelihood of a heart attack, with older age and higher BMI increasing the risk. However, in this model, hours of sleep per day and sleep disturbance showed no significant association with heart attack. Subsequently, smoking showed a slight association with heart attacks, suggesting that smoking may be an influencing factor, but the results are not sufficiently robust to be conclusive. Furthermore, according to this model, gender does not appear to have a significant effect on the probability of a heart attack, but age does. These results demonstrate that, although sleep quality and duration are important aspects of overall health, in this particular dataset they are not as strongly associated with heart attack risk as other factors such as age and BMI would suggest.

The ROC curve with an AUC of 0.54 indicates that the model performs slightly better than random guessing, but not by much. AUC values close to 0.5 generally suggest poor model performance. This is due because of many reasons but the main reason could be that there is a weak or non-linear relationship between the independent variable and the outcome. Meaning, the effect of sleep quality, such as sleep disturbance or sleep duration, on heart attacks may be ambiguous or interact with other factors in a way that is not easily captured by simple linear models such as logistic regression.



Chapter 5

Conclusion

In concluding our study, we have already expressed that the relationship between sleep quality and cardiovascular health, in particular the incidence of myocardial infarction (MI), has become an important but complex area of research. We then sought to explore this association through data analysis and use of the literature review, taking into account demographic factors such as age and gender. We can already say that our analysis of the data shows that there is a relatively weak relationship between sleep quality measured by various indicators such as sleep efficiency, the number of nocturnal awakenings and the risk of heart attack. This result may seem surprising given the existing literature, but it highlights the complexity of the relationship between sleep and heart health.

Previous studies have often associated poor sleep quality with an increased risk of cardiovascular disease. However, new findings show that this association is less direct, particularly in the case of heart attacks, suggesting that it may not be specific or generalizable. However, according to our literature review, several studies have shown that extreme sleep duration, both insufficient and excessive, is associated with an increased risk of cardiovascular disease, including heart attack. For example, previous research has shown that people who sleep less than 5 hours or more than 9 hours per night have a significantly higher risk of heart attack. This observation shows that although sleep time plays an important role in heart health, other variables such as clinical circumstances, co-morbidities and lifestyle factors must also be taken into account to better understand this relationship. Secondly, it should be pointed out that in our literature review we find gender differences in sleep and cardiovascular health. Women are more likely to suffer from sleep disorders such as insomnia due to hormonal fluctuations throughout their lives, which may make them more susceptible to adverse cardiovascular problems. However, men are more susceptible to sleep apnoea, which some studies suggest may be associated with heart disease.

To return to a little more about our data analysis, we can note that when we analysed the correlation between various sleep variables (sleep duration, sleep efficiency etc...) and the incidence of myocardial infarction, the results were relatively weak, or even non-existent in some cases. For instance, the general correlation matrix and the graphs separated by sex show that the correlation coefficient between the sleep variables and myocardial infarction is very close to 0, and values such as -0.01 and 0.02 indicate a non-significant relationship. These results are also reflected in the heat maps which examine the correlations based on age group and sex, but no strong trend has been identified as we mentioned under each plot. However, we notice that one of the most challenging presentations was a comparison of sleep efficiency between people who had a heart attack and those who had not. Although we initially suspected that sleep efficiency was an indicator linked to heart attack risk, the box plots showed minimal differences between the two groups, indicating that sleep efficiency, although important for human health, was not directly associated with heart attack risk.

Another interesting aspect of our analysis was to examine the association between sleep duration and MI in different age and gender groups. The regression scatter plot showed no consistent trend or significant relationship, suggesting that the effect of sleep on heart attack risk remains small even when age and sex are taken into account. In addition, cross-sectional analyses such as cholesterol and sleep duration by sex did not provide convincing evidence of a clear relationship between sleep quality and heart attack frequency. We can also say that charts examining risk factors such as high cholesterol and family history also showed a clearer and more significant association. For example, a bar chart comparing people with and without a family history of heart disease showed that the former had a significantly higher incidence of heart attacks which is 76.5% compared with 38.4%. Similarly, a graph of cholesterol levels revealed that high levels were systematically associated with an increased risk of heart attack, and that this risk reached alarming levels, particularly in men under the age of 40.

Finally, we can conclude that according to our data analysis there is no strong relationship, if any, between sleep variables and cardiovascular problems such as heart attacks. This conclusion was supported by numerous plots of different types of graph. We also took the time to fully understand the analyses so that we knew what they showed. So we can conclude that there is no link between sleep variables and mayorcadial infraction, also taking into account age and sex as mentioned in our research question. Now with regard to our review literature we can see that opinions are divided as to whether having poor sleep can have an effect on a heart attack, however the discrepancies may be due to a number of things including how the dataset was collected or some studies need to use other sleep variables with related other relationships which could give a positive result

on this subject. We can therefore say that we agree with the opinion of several studies which have mentioned that there is no strong relationship between sleep variables and myocardial fructus, taking age and sex into account.

Subsequently, our results raise important questions for future research. It may be useful to reassess the importance of sleep in the context of cardiovascular health, perhaps seeking to understand whether other aspects of sleep, not explored in this study, and using other datasets collected in different locations could have a more significant impact. Furthermore, these results encourage a more holistic approach to heart disease prevention, integrating traditional risk factors, while continuing to explore other potentially relevant dimensions to better target preventive interventions.

Bibliography

[1] Agustín Javier Simonelli-Muñoz et al. "Relationship between sleep habits and academic performance in university nursing students". In: *BMC Nursing* 20.1 (2021), p. 109. URL: https://bmcnurs.biomedcentral.com/articles/10.1186/s12912-021-00635-x.

[2] Eliana Perez, Alejandro Gomez, and Maria Fernandez. "Sleep quality and sleep deprivation: relationship with academic performance in university students during examination period". In: *Sleep and Biological Rhythms* 21.3 (2023), pp. 377–383. DOI: 10.1007/s41105-023-00457-1. URL: https://link.springer.com/article/10.1007/s41105-023-00457-1.

[3] Francesco P. Cappuccio et al. "Sleep duration predicts cardiovascular outcomes: a systematic review and meta-analysis of prospective studies". In: *European Heart Journal* 32.12 (2011), pp. 1484–1492. DOI: 10.1093/eurheartj/ehr334. URL: https://academic.oup.com/eurheartj/article/32/12/1484/502022.

[4] Benjamin Djulbegovic. "Sex differences in risk factors for myocardial infarction". In: *BMJ* 363 (2018), k4247. URL: https://www.bmj.com/content/363/bmj.k4247.

[5] X. et al. Liu. "Sleep Duration and Myocardial Infarction". In: *National Center for Biotechnology Information* (2019). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6785011/.

[6] H. et al. Alhazmi. "Sleeping Late Increases the Risk of Myocardial Infarction". In: *Frontiers in Cardiovascular Medicine* (2021). URL: https://www.frontiersin.org/journals/cardiovascular-medicine/articles/10.3389/fcvm.2021.709468/full.

[7] Yan Qian et al. "Sleep Duration and Myocardial Infarction: A Prospective Study". In: *National Center for Biotechnology Information* (2019). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6785011/.

[8] Alyssa Feinberg. "People Who Sleep Too Much, or Too Little, at Heightened Risk of Heart Attack". In: *American Journal of Managed Care* (2019). URL: https://www.ajmc.com/view/people-who-sleep-too-much-or-too-little-at-heightened-risk-of-heart-attack.

[9] Yi-Shan et al. Chao. "Sleep Duration and Risk of Myocardial Infarction and All-Cause Death". In: *ScienceDirect* (2015). URL: https://www.sciencedirect.com/science/article/abs/pii/S1389945715020651.

[10] Andrew et al. Steptoe. "Sleep disturbance after acute coronary syndrome". In: *PLOS ONE* 17.6 (2022), e0269545. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0269545.

[11] S. Jiang et al. "Exploring postoperative atrial fibrillation after non-cardiac surgery: mechanisms, risk factors, and prevention strategies". In: *Frontiers in Cardiovascular Medicine* (2023). URL: https://www.frontiersin.org/articles/10.3389/fcvm.2023.1273547/full.

[12] Sound Sleep Health. "Low Testosterone: Signs, Symptoms, Causes, and Risk Factors". In: *Sound Sleep Health* (2017). URL: https://www.soundsleephealth.com/low-testosterone-signs-symptoms-causes-and-risk-factors/.

[13] Virend K. et al. Somers. "The Lifestyle-Related Cardiovascular Risk Is Modified by Sleep". In: *Mayo Clinic Proceedings* (2021). URL: https://www.mayoclinicproceedings.org/article/S0025-6196(21)00701-1/abstract.

[14] Thomas Roth. "Insomnia: Definition, Prevalence, Etiology, and Consequences". In: *Journal of Clinical Sleep Medicine* (2012). URL: https://jcsm.aasm.org/doi/10.5664/jcsm.26929.

[15] Martino F. et al. Pengo. "Gender medicine and sleep disorders: from basic science to clinical practice". In: *Frontiers in Neurology* (2024). URL: https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2024.1392489/pdf.

[16] Emily et al. Reed. "Effects of sleep deprivation on coronary heart disease". In: *PMC - NCBI* (2022). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9437362/.

[17] JoAnn E. et al. Manson. "Protective Effects of Estrogen on Cardiovascular Disease Mediated by Oxidative Stress". In: *Wiley Online Library* (2021). URL: https://onlinelibrary.wiley.com/doi/10.1155/2021/5523516.

[18] Michael R. et al. Irwin. "Sleep Disturbance, Sleep Duration, and Inflammation: A Systematic Review". In: *Biological Psychiatry* (2015). URL: https://www.biologicalpsychiatryjournal.com/article/S0006-3223(15)00437-0/abstract.

[19] Bjorn et al. Bjorvatn. "Thirty-Year Trends in Sleep Disorders and Cardiovascular Disease". In: *IntechOpen* (2023). URL: https://www.intechopen.com/online-first/1180676.

[20] S.C Cromack. "To sleep perchance to dream of pregnancy". In: *Fertility and Sterility* (2024). URL: https://www.fertstert.org/article/S0015-0282(23)02077-0/abstract.

[21] Frederic et al. Gagnadoux. "Obstructive Sleep Apnea Syndrome (OSAS) and Menopause". In: *IntechOpen* (2024). URL: https://www.intechopen.com/online-first/1184110.

[22] Y. S. et al. Bin. "Gender differences in the relationship between sleep and age in a general population cohort". In: *Journal of Sleep Research* (2024). URL: https://onlinelibrary.wiley.com/doi/full/10.1111/jsr.14154.

[23] Felicita et al. Cespedes. "Sleep Quality, Sleep Duration, and the Risk of Coronary Heart Disease". In: *Journal of Clinical Sleep Medicine* (2018). URL: https://jcsm.aasm.org/doi/full/10.5664/jcsm.6894.

[24] Shahrokh Javaheri and Susan Redline. "Insomnia and risk of cardiovascular disease". In: *Chest* (2017). DOI: 10.1016/j.chest.2017.01.026.

[25] Daniel J. Buysse et al. "The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research". In: *Psychiatry Research* (1989). DOI: 10.1016/0165-1781(89)90047-4.

[26] Centers for Disease Control and Prevention. "National Health and Nutrition Examination Survey Data". In: *Centers for Disease Control and Prevention* (2018). https://www.cdc.gov/nchs/nhanes/index.htm.

[27] Wes McKinney. "Data structures for statistical computing in Python". In: *Proceedings of the 9th Python in Science Conference* (2010).

Appendix A

Python codes for the Data Analysis

Exploratory Data Analysis (EDA) was conducted using Python scripts executed within Kaggle's cloud-based Jupyter Notebook environment. After uploading the dataset, we examined its structure, checked for missing values and carried out analyses with different types of graph and different correlations. To do this we used the Python programming language and a number of libraries such as Pandas, Numpy, Matplotlib and Seaborn to visualise the data and identify patterns, correlations. To address the missing data, we already have the dataset pre-cleaned by our health dataset provider, however we checked this in as shown in the image below. Now that the dataset is complicated, we performed our analyses.

[2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import ttest_ind
```

```
import pandas as pd

# Load the dataset
data = pd.read_csv('/kaggle/input/sleepheart-dataset-msc-project/Heart Attack Risk Analysis.csv')
```

+ Code + Markdown

[2]:
```
# Check for missing values
data.isnull().sum()
```

[2]:
```
Patient ID                        0
Age                               0
Sex                               0
Cholesterol                       0
Blood Pressure                    0
Heart Rate                        0
Diabetes                          0
Family History                    0
Smoking                           0
Obesity                           0
Alcohol Consumption               0
Exercise Hours Per Week           0
Diet                              0
Previous Heart Problems           0
Medication Use                    0
Stress Level                      0
Sedentary Hours Per Day           0
Income                            0
BMI                               0
Triglycerides                     0
Physical Activity Days Per Week   0
Sleep Hours Per Day               0
Country                           0
Continent                         0
```

Plot (1) - Boxplot of Sleep Hours Per Day by Myocardial Infarction Status

```
# Plotting the relationship between sleep duration and MI
# Plot (1) - Boxplot of Sleep Hours Per Day by Myocardial Infarction Status
plt.figure(figsize=(8, 6))
sns.boxplot(x='Myocardial Infarction', y='Sleep Hours Per Day', data=data)
plt.title('Sleep Hours Per Day vs Myocardial Infarction')
plt.xlabel('Myocardial Infarction')
plt.ylabel('Sleep Hours Per Day')
plt.show()

# By Soubhi SAAD
```

Plot (2) - Distribution of Sleep Hours Per Day with KDE Overlay

```python
import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10, 6))
sns.histplot(data['Sleep Hours Per Day'], kde=True, bins=10, color='blue', edgecolor='black')
plt.title('Distribution of Sleep Hours Per Day with KDE')
plt.xlabel('Sleep Hours Per Day')
plt.ylabel('Frequency')
plt.show()

#Soubhi SAAD
```

Plot (3) - Average Sleep Disturbances by Age Group and Gender

```python
plt.figure(figsize=(12, 8))
sns.barplot(x='Age Group', y='Sleep Disturbances', hue='Sex', data=data, ci=None)
plt.title('Average Sleep Disturbances by Age Group and Gender')
plt.xlabel('Age Group')
plt.ylabel('Average Sleep Disturbances')
plt.show()

# Soubhi SAAD
```

Plot (4) - Correlation Matrix for Males and Females: Sleep Data and MI

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
#Soubhi SAAD

# Load the dataset
file_path = '/kaggle/input/sleepheart-dataset-msc-project/Heart Attack Risk Analysis.csv'
sleep_data = pd.read_csv(file_path)

# Selecting relevant columns
sleep_heart_data = sleep_data[['Sleep Hours Per Day', 'Sleep Disturbances', 'Sleep Efficiency',
                               'Sleep Onset Latency', 'WASO', 'Number of Awakenings',
                               'Myocardial Infarction', 'Sex', 'Age']]

# Convert Sex to numeric for correlation (0 for Female, 1 for Male)
sleep_heart_data['Sex'] = sleep_data['Sex'].map({'Male': 1, 'Female': 0})

# Creating age groups
bins = [0, 40, 60, 100]
labels = ['Under 40', '40-60', 'Over 60']
sleep_heart_data['Age Group'] = pd.cut(sleep_heart_data['Age'], bins=bins, labels=labels)

# Splitting the data by sex
male_data = sleep_heart_data[sleep_heart_data['Sex'] == 1]
female_data = sleep_heart_data[sleep_heart_data['Sex'] == 0]

# Exclude non-numeric 'Age Group' column for correlation calculation
male_correlation = male_data.drop(columns=['Age Group']).corr()
female_correlation = female_data.drop(columns=['Age Group']).corr()

# Plotting the correlation matrices for males and females
plt.figure(figsize=(14, 6))

plt.subplot(1, 2, 1)
sns.heatmap(male_correlation, annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Correlation Matrix for Males")

plt.subplot(1, 2, 2)
sns.heatmap(female_correlation, annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Correlation Matrix for Females")

plt.tight_layout()
plt.show()
```

Plot (5) - Correlation Matrix between Sleep Data and MI

```
import seaborn as sns
import matplotlib.pyplot as plt

# Selecting relevant columns
sleep_heart_data = sleep_data[['Sleep Hours Per Day', 'Sleep Disturbances', 'Sleep Efficiency',
                               'Sleep Onset Latency', 'WASO', 'Number of Awakenings',
                               'Myocardial Infarction']]

# Calculate the correlation matrix
correlation_matrix = sleep_heart_data.corr()

# Plotting the correlation matrix
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Correlation Matrix between Sleep Data and Myocardial Infarction")
plt.show()

# Soubhi SAAD
```

Plot (6) - Scatter Plot of Sleep Efficiency vs Myocardial Infarction by Age Group and Gender

```
import seaborn as sns
import matplotlib.pyplot as plt

# Scatter plot with regression line for one sleep variable (e.g., Sleep Efficiency)
plt.figure(figsize=(12, 8))
sns.lmplot(x='Sleep Efficiency', y='Myocardial Infarction', hue='Gender', col='Age Group', data=sleep_data,
           scatter_kws={'alpha':0.5}, line_kws={'color':'red'})
plt.suptitle("Scatter Plot of Sleep Efficiency vs Myocardial Infarction by Age Group and Gender", y=1.02)
plt.show()

#Soubhi SAAD
```

Plot (7) - Comparison of Sleep Efficiency between Individuals with and without Myocardial Infarction

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import ttest_ind

# Load your dataset
sleep_data = pd.read_csv('/kaggle/input/sleepheart-dataset-msc-project/TheOne Sleep Data.csv')

# Create the two groups
group_1 = sleep_data[sleep_data['Myocardial Infarction'] == 1]
group_2 = sleep_data[sleep_data['Myocardial Infarction'] == 0]

# List of sleep variables to analyze
sleep_vars = ['Sleep Hours Per Day', 'WASO', 'Sleep Disturbances', 'Sleep Onset Latency', 'Number of Awakenings', 'Sleep Efficiency']

# Statistical comparison and visualization
for var in sleep_vars:
    # Perform a t-test to compare the means between the two groups
    t_stat, p_value = ttest_ind(group_1[var], group_2[var], equal_var=False)

    # Print the results of the t-test
    print(f"{var}: t-statistic = {t_stat:.2f}, p-value = {p_value:.3f}")

    # Plot the distributions of the sleep variable for both groups
    plt.figure(figsize=(10, 6))
    sns.boxplot(x='Myocardial Infarction', y=var, data=sleep_data)
    plt.title(f"Comparison of {var} between Individuals with and without Myocardial Infarction")
    plt.xlabel("Myocardial Infarction (0 = No, 1 = Yes)")
    plt.ylabel(var)
    plt.show()

# Soubhi SAAD
```

Plot (8) - Top Countries with the Most Recorded MI Cases

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load your dataset
sleep_data = pd.read_csv('/kaggle/input/sleepheart-dataset-msc-project/TheOne Sleep Data.csv')

# Group the data by Country and sum the Myocardial Infarction cases
country_mi_counts = sleep_data.groupby('Country')['Myocardial Infarction'].sum().reset_index()

# Sort the data by the number of MI cases in descending order
country_mi_counts = country_mi_counts.sort_values(by='Myocardial Infarction', ascending=False)

# Display the top countries with the most MI cases
print(country_mi_counts.head(10))

# Plot the data
plt.figure(figsize=(14, 8))
sns.barplot(x='Myocardial Infarction', y='Country', data=country_mi_counts, palette='viridis')
plt.title("Top Countries with the Most Recorded Myocardial Infarction Cases")
plt.xlabel("Number of Myocardial Infarction Cases")
plt.ylabel("Country")
plt.show()

#Soubhi SAAD
```

Plot (9) - Comparison of Myocardial Infarction Cases Between Smokers and Non-Smokers

```
import matplotlib.pyplot as plt
import seaborn as sns

# Calculate the number of people who had a myocardial infarction but do not smoke
mi_non_smokers = sleep_data[(sleep_data['Myocardial Infarction'] == 1) & (sleep_data['Smoking'] == 0)]
num_mi_non_smokers = mi_non_smokers.shape[0]

# Data for plotting
categories = ['MI & Smoke', 'MI & Non-Smoke']
values = [912, num_mi_non_smokers]

# Plotting
plt.figure(figsize=(10, 6))
sns.barplot(x=categories, y=values, palette='viridis')
plt.title("Comparison of Myocardial Infarction Cases Between Smokers and Non-Smokers")
plt.ylabel("Number of People")
plt.xlabel("Group")
plt.show()

# Output the numbers for reference
print(f"Number of people who had a Myocardial Infarction and smoke: {912}")
print(f"Number of people who had a Myocardial Infarction and do not smoke: {num_mi_non_smokers}")

#Soubhi SAAD
```

Plot (10) - Rate of MI by Family History of Heart Problems

```
import pandas as pd
import matplotlib.pyplot as plt

# Load the data from the provided CSV file
file_path = '/kaggle/input/sleepheart-dataset-msc-project/Heart Attack Risk Analysis.csv'
data = pd.read_csv(file_path)

# Calculate the rate of MI among those with and without a family history
family_history_group = data.groupby('Family History')['Myocardial Infarction'].mean()

# Reorder the series to show 'Family History' first
family_history_group = family_history_group.sort_index(ascending=False)

# Plotting the results
plt.figure(figsize=(8, 6))
bars = family_history_group.plot(kind='bar', color=['salmon', 'skyblue'])

# Adding data labels to the bars
for index, value in enumerate(family_history_group):
    plt.text(index, value + 0.02, f'{value:.1%}', ha='center', fontsize=12)

plt.title('Rate of Myocardial Infarction by Family History')
plt.xlabel('Family History of Heart Problems')
plt.ylabel('Proportion of Myocardial Infarction Cases')
plt.xticks(ticks=[0, 1], labels=['Family History', 'No Family History'], rotation=0)
plt.ylim(0, 1)

# Display the plot
plt.show()

#Soubhi SAAD
```

PLOT (11) - Rate of Myocardial Infarction by Sex & PLOT (12) - Rate of Myocardial Infarction by Age Group

```python
import pandas as pd
import matplotlib.pyplot as plt

# Load the data from the provided CSV file
file_path = '/kaggle/input/sleepheart-dataset-msc-project/Heart Attack Risk Analysis.csv'
data = pd.read_csv(file_path)

# First, we'll categorize age into groups for better comparison
# Define age groups (e.g., <40, 40-60, >60)
bins = [0, 40, 60, 100]
labels = ['<40', '40-60', '>60']
data['Age Group'] = pd.cut(data['Age'], bins=bins, labels=labels, right=False)

# Calculate the rate of MI by sex
sex_group = data.groupby('Sex')['Myocardial Infarction'].mean()

# Calculate the rate of MI by age group
age_group = data.groupby('Age Group')['Myocardial Infarction'].mean()

# Plotting the MI rate by sex
plt.figure(figsize=(8, 6))
sex_group.plot(kind='bar', color=['lightblue', 'lightcoral'])
for index, value in enumerate(sex_group):
    plt.text(index, value + 0.02, f'{value:.1%}', ha='center', fontsize=12)
plt.title('Rate of Myocardial Infarction by Sex')
plt.xlabel('Sex')
plt.ylabel('Proportion of Myocardial Infarction Cases')
plt.xticks(rotation=0)
plt.ylim(0, 1)
plt.show()

# Plotting the MI rate by age group
plt.figure(figsize=(8, 6))
age_group.plot(kind='bar', color=['lightgreen', 'gold', 'lightblue'])
for index, value in enumerate(age_group):
    plt.text(index, value + 0.02, f'{value:.1%}', ha='center', fontsize=12)
plt.title('Rate of Myocardial Infarction by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Proportion of Myocardial Infarction Cases')
plt.xticks(rotation=0)
plt.ylim(0, 1)
plt.show()

#Soubhi SAAD
```

PLOT (13) - Myocardial Infarction Rate by Sleep Hours and Age Group

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the data from the provided CSV file
file_path = '/kaggle/input/sleepheart-dataset-msc-project/Heart Attack Risk Analysis.csv'
data = pd.read_csv(file_path)

# Recreate the Age Group column
bins_age = [0, 40, 60, 100]
labels_age = ['<40', '40-60', '>60']
data['Age Group'] = pd.cut(data['Age'], bins=bins_age, labels=labels_age, right=False)

# Categorize Sleep Hours
bins_sleep = [0, 6, 8, 12]
labels_sleep = ['<6 hours', '6-8 hours', '>8 hours']
data['Sleep Category'] = pd.cut(data['Sleep Hours Per Day'], bins=bins_sleep, labels=labels_sleep, right=False)

# Prepare the data for plotting
facet_data = data.groupby(['Sex', 'Age Group', 'Sleep Category'])['Myocardial Infarction'].mean().reset_index()
facet_data['Myocardial Infarction'] = facet_data['Myocardial Infarction'] * 100  # Convert to percentage

# Create a Facet Grid Plot with specified colors
g = sns.catplot(
    data=facet_data, kind="bar",
    x="Sleep Category", y="Myocardial Infarction",
    hue="Sex", col="Age Group",
    ci=None, palette={'Male': 'blue', 'Female': 'orange'}, height=4, aspect=0.9
)
g.set_axis_labels("Sleep Hours Per Day", "MI Rate (%)")
g.set_titles("{col_name} Age Group")
g.set(ylim=(0, 100))
g.despine(left=True)

# Show the plot
plt.show()

#Soubhi SAAD
```

Plot (14) - Myocardial Infarction Rate by Cholesterol Level and Age Group

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the data from the provided CSV file
file_path = '/kaggle/input/sleepheart-dataset-msc-project/Heart Attack Risk Analysis.csv'
data = pd.read_csv(file_path)

# Recreate the Age Group column
bins_age = [0, 40, 60, 100]
labels_age = ['<40', '40-60', '>60']
data['Age Group'] = pd.cut(data['Age'], bins=bins_age, labels=labels_age, right=False)

# Step 1: Categorize Cholesterol Levels
# Assuming general categories: <200 mg/dL (Normal), 200-239 mg/dL (Borderline High), >=240 mg/dL (High)
bins_cholesterol = [0, 200, 240, data['Cholesterol'].max()]
labels_cholesterol = ['Normal', 'Borderline High', 'High']
data['Cholesterol Category'] = pd.cut(data['Cholesterol'], bins=bins_cholesterol, labels=labels_cholesterol, right=False)

# Step 2: Prepare the data for plotting
cholesterol_facet_data = data.groupby(['Sex', 'Age Group', 'Cholesterol Category'])['Myocardial Infarction'].mean().reset_index()
cholesterol_facet_data['Myocardial Infarction'] = cholesterol_facet_data['Myocardial Infarction'] * 100  # Convert to percentage

# Step 3: Create a Facet Grid Plot
g = sns.catplot(
    data=cholesterol_facet_data, kind="bar",
    x="Cholesterol Category", y="Myocardial Infarction",
    hue="Sex", col="Age Group",
    ci=None, palette={'Male': 'blue', 'Female': 'orange'}, height=4, aspect=0.9
)
g.set_axis_labels("Cholesterol Level", "MI Rate (%)")
g.set_titles("{col_name} Age Group")
g.set(ylim=(0, 100))
g.despine(left=True)

# Show the plot
plt.show()

#Soubhi SAAD
```

Plot (15) - P-value for T-tests: Sleep Quality vs MI in Men & Female

t-test for Female

```
import pandas as pd
from scipy import stats

# Load the dataset
file_path = '/kaggle/input/sleepheart-dataset-msc-project/Heart Attack Risk Analysis.csv'
df = pd.read_csv(file_path)

# Clean and preprocess the data
# Assuming 'Myocardial Infarction' is a binary variable where 1 = Yes, 0 = No
# Assuming 'Sex' is coded as 0 = Male, 1 = Female

# Segmenting the data by gender and then by age groups
df['Age Group'] = pd.cut(df['Age'], bins=[0, 40, 60, 100], labels=['<40', '40-60', '>60'])

# Performing t-tests on sleep duration and sleep disturbances across MI status, gender, and age group

results = {}

# Sleep Duration vs MI Status
for gender in df['Sex'].unique():
    gender_label = 'Male' if gender == 0 else 'Female'
    for age_group in df['Age Group'].unique():
        if pd.isna(age_group):
            continue
        group_data = df[(df['Sex'] == gender) & (df['Age Group'] == age_group)]
        mi_group = group_data[group_data['Myocardial Infarction'] == 1]['Sleep Hours Per Day']
        no_mi_group = group_data[group_data['Myocardial Infarction'] == 0]['Sleep Hours Per Day']

        t_stat, p_value = stats.ttest_ind(mi_group.dropna(), no_mi_group.dropna(), equal_var=False)

        results[f'{gender_label}, Age {age_group}, Sleep Duration'] = (t_stat, p_value)

# Sleep Disturbances vs MI Status
for gender in df['Sex'].unique():
    gender_label = 'Male' if gender == 0 else 'Female'
    for age_group in df['Age Group'].unique():
        if pd.isna(age_group):
            continue
        group_data = df[(df['Sex'] == gender) & (df['Age Group'] == age_group)]
        mi_group = group_data[group_data['Myocardial Infarction'] == 1]['Sleep Disturbances']
        no_mi_group = group_data[group_data['Myocardial Infarction'] == 0]['Sleep Disturbances']

        t_stat, p_value = stats.ttest_ind(mi_group.dropna(), no_mi_group.dropna(), equal_var=False)
```

```
        t_stat, p_value = stats.ttest_ind(mi_group.dropna(), no_mi_group.dropna(), equal_var=False)

        results[f'{gender_label}, Age {age_group}, Sleep Disturbances'] = (t_stat, p_value)

# Output the results
for key, (t_stat, p_value) in results.items():
    print(f"{key} -> t-statistic: {t_stat:.2f}, p-value: {p_value:.3f}")

import matplotlib.pyplot as plt

# Extracting data for plotting
labels = list(results.keys())
t_stats = [result[0] for result in results.values()]
p_values = [result[1] for result in results.values()]

# Plotting the p-values
plt.figure(figsize=(10, 6))
plt.barh(labels, p_values, color='skyblue')
plt.axvline(x=0.05, color='red', linestyle='--', label='Significance Level (0.05)')
plt.xlabel('P-value')
plt.ylabel('Comparison')
plt.title('P-values for T-tests: Sleep Quality vs Myocardial Infarction')
plt.legend()
plt.tight_layout()
plt.show()

# Soubhi SAAD
```

t-test for Male

```
import pandas as pd
from scipy.stats import ttest_ind
import matplotlib.pyplot as plt

# Load the data
data = pd.read_csv('/kaggle/input/sleepheart-dataset-msc-project/Heart Attack Risk Analysis.csv')

# Define age groups
age_groups = [
    ('<40', data['Age'] < 40),
    ('40-60', (data['Age'] >= 40) & (data['Age'] <= 60)),
    ('>60', data['Age'] > 60)
]

# Store results
results = []

# Perform t-tests for Sleep Disturbances and Sleep Duration for males
for age_label, age_condition in age_groups:
    for variable in ['Sleep Disturbances', 'Sleep Hours Per Day']:
        male_with_mi = data[(data['Sex'] == 'Male') & age_condition & (data['Myocardial Infarction'] == 1)][variable].dropna()
        male_without_mi = data[(data['Sex'] == 'Male') & age_condition & (data['Myocardial Infarction'] == 0)][variable].dropna()

        t_stat, p_value = ttest_ind(male_with_mi, male_without_mi, equal_var=False)
        results.append((f'Male, Age {age_label}, {variable}', p_value, t_stat))

# Create a DataFrame for the results
results_df = pd.DataFrame(results, columns=['Comparison', 'P-value', 'T-statistic'])

# Display the results in text
for index, row in results_df.iterrows():
    comparison = row['Comparison']
    p_value = row['P-value']
    t_statistic = row['T-statistic']
    print(f"Comparison: {comparison}\nT-statistic: {t_statistic:.3f}, P-value: {p_value:.3f}\n{'-'*50}")

# Plot the p-values
plt.figure(figsize=(10, 6))
plt.barh(results_df['Comparison'], results_df['P-value'], color='skyblue')
plt.axvline(x=0.05, color='red', linestyle='--', label='Significance Level (0.05)')
plt.xlabel('P-value')
plt.title('P-values for T-tests: Sleep Quality vs Myocardial Infarction (Male)')
plt.legend()
plt.show()
```

Logistic Regression - Machine Learning Algorithm Model

```
import pandas as pd
import statsmodels.api as sm
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix, roc_curve, roc_auc_score
import matplotlib.pyplot as plt
import seaborn as sns

# Load the data
data = pd.read_csv("/kaggle/input/sleepheart-dataset-msc-project/Heart Attack Risk Analysis.csv")

# Create dummy variables for categorical data
data['Sex'] = data['Sex'].map({'Male': 1, 'Female': 0})  # Assuming 'Male' is 1 and 'Female' is 0

# Select relevant columns for the logistic regression
X = data[['Sleep Hours Per Day', 'Sleep Disturbances', 'Age', 'Sex', 'Smoking', 'BMI']]
y = data['Myocardial Infarction']

# Add a constant to the independent variables (required for statsmodels)
X = sm.add_constant(X)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

logit_model = sm.Logit(y_train, X_train)
result = logit_model.fit()

# Print the summary
print(result.summary())

#Soubhi SAAD
```

Appendix B

Gantt chart