

A PROJECT REPORT

ON

“Human Age-Prediction Synthetic Dataset”

**Submitted in partial fulfillment of the requirement for the Course CSE303**

**(Course title:** Data Science**, Section:**02**)**

Of

BACHELOR OF SCIENCE

IN

COMPUTER SCIENCE AND ENGINEERING

**Submitted By: Group-09**

1. Shanta Islam (ID: 2022-1-60-288)

2. Md Sifatullah Sheikh (ID: 2022-1-60-029)

3. Arpita Saha (ID: 2022-1-60-093)

**Under the supervision of**

**Dr. Mohammad Manzurul Islam**

(Assistant Professor, Department of Computer Science and Engineering)

**TABLE OF CONTENTS**

1.Introduction

2.Dataset description

3.Data Preprocessing

4.Exploratory Data Analysis (EDA)

5.Machine learning model

6.conclusion

**Introduction**

This project focuses on predicting human age using a synthetic dataset sourced from Kaggle.

In this project, we are aiming to predict the age using different features of human such as height, weight, blood pressure, BMI etc as well as daily basic habits with different medical conditions.

Age prediction is a valuable tool in various applications such as healthcare, marketing and demographic analysis. Accurate age prediction can help in personalizing user experiences, targeted advertising and improving user engagement.

Through this project, we are trying to build a machine learning model which will predict age accurately based on given features and to analyze the relationships among them.

**Dataset Description**

This dataset contains synthetic data designed for predicting age based on various health and lifestyle factors. It includes 3,000 rows with 26 features, each representing different aspects of physical health and lifestyle.

**Features:**

* **Height (cm)**: The height of the individual in centimeters.
* **Weight (kg)**: The weight of the individual in kilograms.
* **Blood Pressure (s/d)**: Blood pressure (systolic/diastolic) in mmHg.
* **Cholesterol Level (mg/dL)**: Cholesterol level in milligrams per deciliter.
* **BMI**: Body Mass Index, calculated from height and weight.
* **Blood Glucose Level (mg/dL)**: Blood glucose level in milligrams per deciliter.
* **Bone Density (g/cm²)**: Bone density in grams per square centimeter.
* **Vision Sharpness**: Vision sharpness on a scale from 0 (blurry) to 100 (perfect).
* **Hearing Ability (dB)**: Hearing ability in decibels.
* **Physical Activity Level**: Categorized as ‘Low’, ‘Moderate’, or ‘High’.
* **Smoking Status**: Categorical values including ‘Never’, ‘Former’, and ‘Current’.
* **Alcohol Consumption**: Frequency of alcohol consumption, categorized as ‘No’, ‘Occasional’, ‘Frequent’.
* **Diet**: Type of diet, categorized as ‘Balanced’, ‘High-fat’, ‘Low-carb’, ‘Vegetarian’.
* **Chronic Diseases**: Presence of chronic diseases (e.g. ‘No’, ‘Hypertension’, ‘Diabetes’, ‘Heart Disease’)
* **Medication Use**: Usage of medication categorized as ‘No’, ‘Regular’ ‘Occasional’
* **Family History**: Presence of family history of age-related conditions categorized as ‘No’, ‘Hypertension’, ‘Diabetes’, ‘Heart Disease’.
* **Cognitive Function**: Self-reported cognitive function on a scale from 0 (poor) to 100 (excellent).
* **Mental Health Status**: Self-reported mental health status on a scale from 0 (poor) to 100 (excellent). (e.g. ‘Poor’, ‘Good’, ‘Fair’, ‘Excellent’)
* **Sleep Patterns**: Average number of sleep hours per night. (e.g. ‘Normal’, ‘Insomnia’, ‘Excessive’)
* **Stress Levels**: Self-reported stress levels on a scale from 0 (low) to 100 (high).
* **Pollution Exposure**: Exposure to pollution measured in arbitrary units.
* **Sun Exposure**: Average sun exposure in hours per week.
* **Education Level**: Highest level of education attained categorized as ‘No’, ‘High School’, ‘Undergraduate’, ‘Postgraduate’.
* **Income Level**: Annual income in USD categorized as ‘Low’, ‘Medium’ or ‘High’
* **Age (years)**: The target variable representing the age of the individual.

**Data Preprocessing**

Preprocessing is a crucial step to ensure the data is clean, consistent, and suitable for modeling. The steps undertaken include:

1. **Handling missing values**: in our dataset, there is 10 features which have null values.
   * **height**: there are 29 null values. By commanding some code, we find out that there is only one outlier so we use average height by calling Mean function.
   * **BMI**: there are 34 null values. From the description we can see the mean and the median are almost same but there are 16 outliers, that’s why we choose median to fill-up the null values.
   * **Blood Glucose Level (mg/dL):** there are 33 null values with 13outliers. There is also slightest difference between mean and median as well as Q3 and we fill the null values with median.
   * **Bone Density (g/cm²):** there are 41 null values with 0 outliers. We choose average/mean value to replace null values.
   * **Vision Sharpness:** there are 40 null values with 0 outliers. Here, for vision sharpness, selecting the average vision is the best choice for us.
   * **Hearing Ability (dB):** there are 59 null values with 9 outliers. Because of the outliers, median is the best choice to avoid it.
   * **Cognitive Function:** there are 82 null values with 11 outliers. We choose median for this as well.
   * **Stress Levels:** there are 54 null values with 0 outliers. Considering average stress level to fill up the null values.
   * **Pollution Exposure:** there are 69 null values with 0 outliers and we replace the null values with mean value.
   * **Sun Exposure:** there are 62 null values with 0 outliers, replaced the null values with mean.
2. **Encoding Categorical Variables**: there are 12 categorical features in our dataset. To build the model from this dataset, it’s important to have these features, that’s why we converted these features into numerical form using numbers like 0,1,2,3 as per needed. Categorical features:

Physical Activity Level

Smoking Status

Alcohol Consumption

Diet

Gender

Chronic Diseases

Medication Use

Family History

Mental Health Status

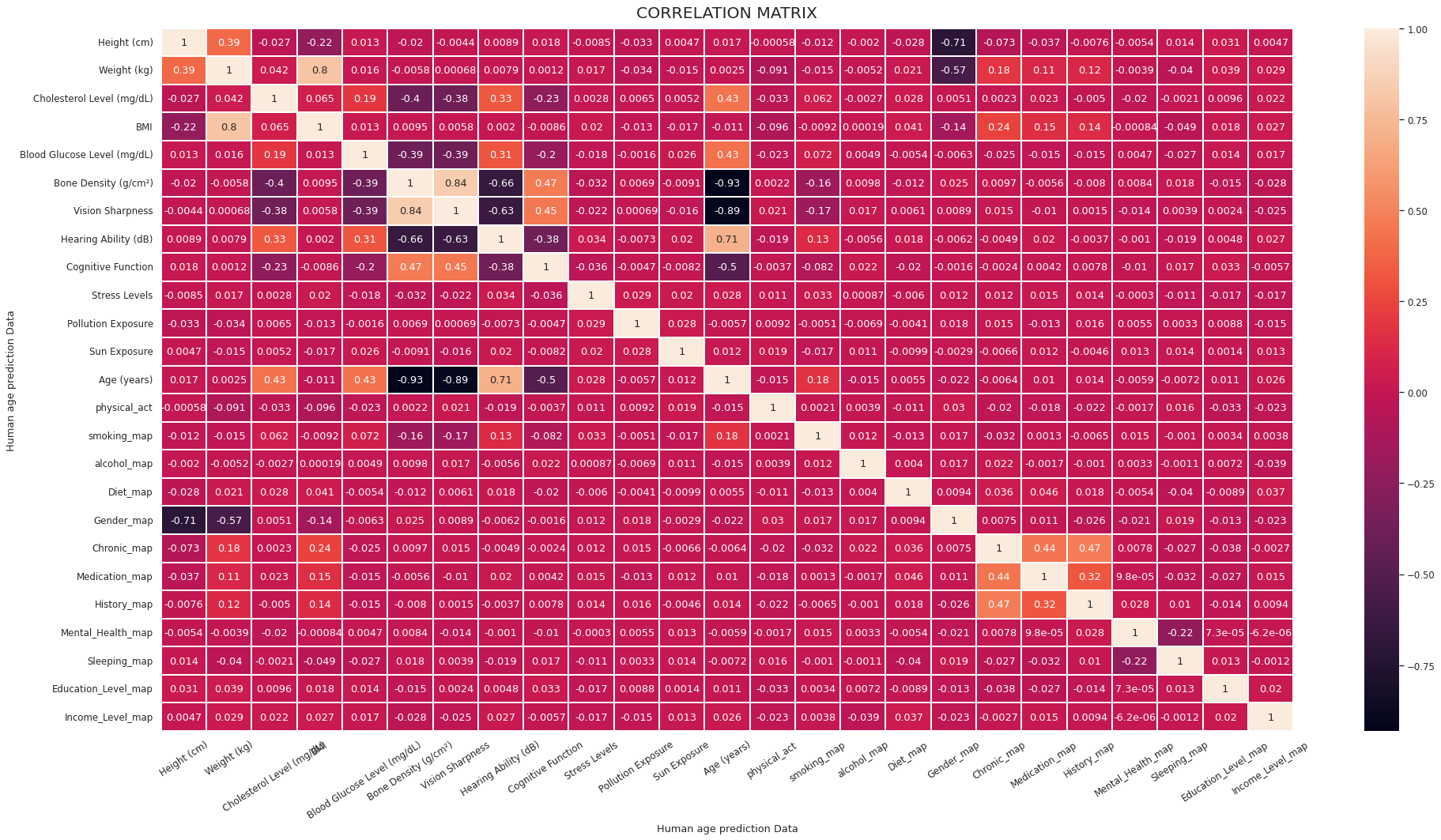
Sleep Patterns

Education Level

Income Level

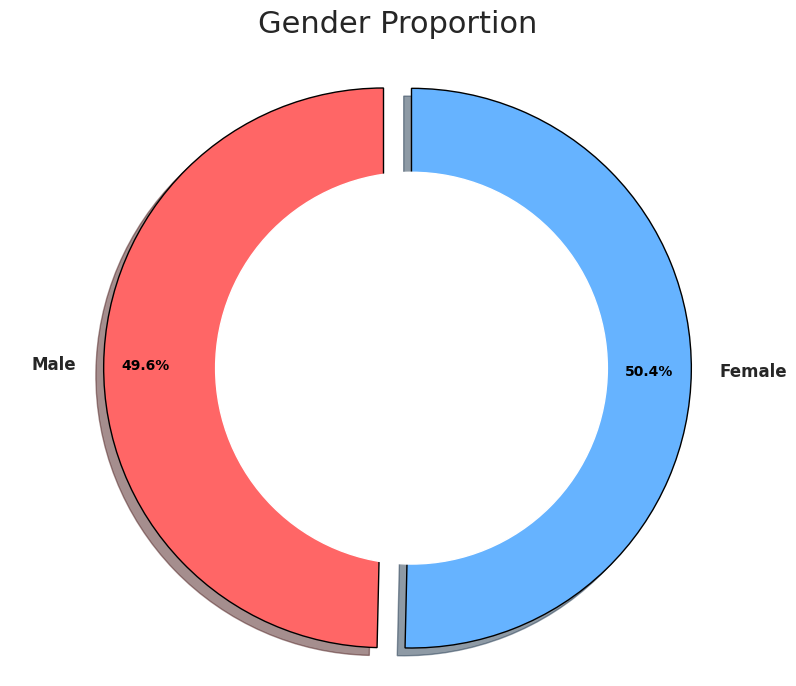
**Exploratory Data Analysis (EDA)**

EDA helps to understand the underlying patterns in the data. Key insights from EDA include:

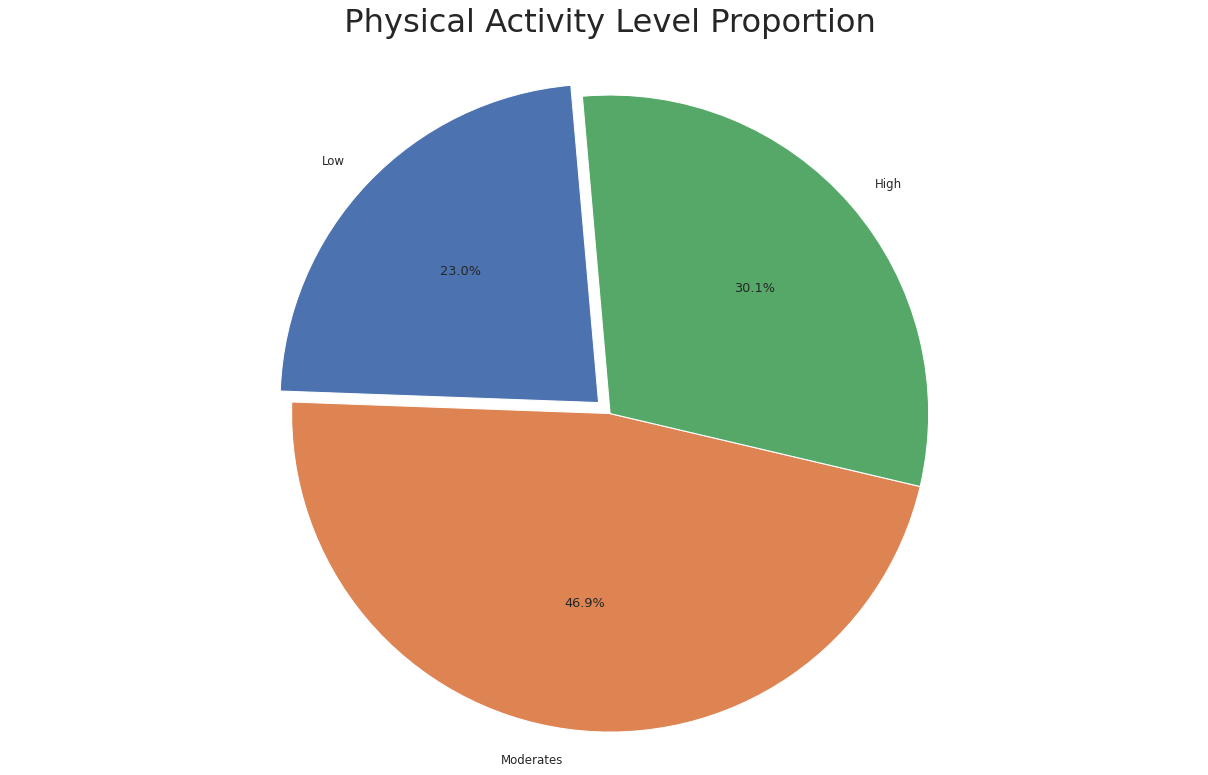
**1.Correlation Matrix**: A heatmap was generated to visualize the correlations between features. 

From this correlation matrix, it’s clear that every features are correlated to each other either strongly positive correlation or strongly negative correlation or weakly positive or negative. The main thing is the target variable which is age is highly correlated with hearing ability, blood glucose level and cholesterol level. It also has strongly negative correlation with bone density, vision sharpness. Smoking\_map also has much effect on age as well. Other features have a little effect on age features.

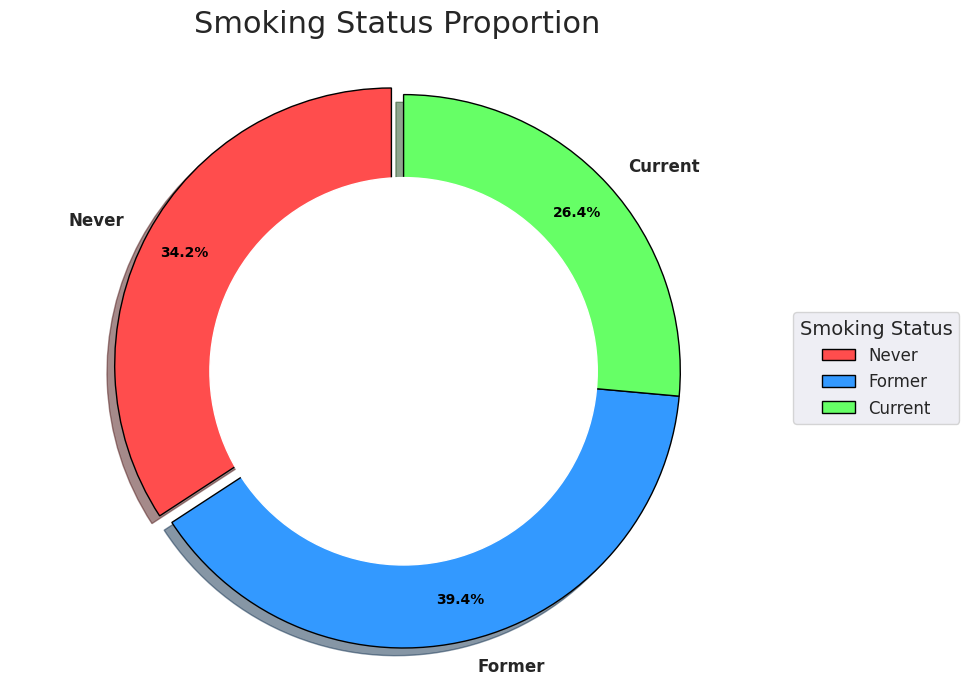
**2.Distribution Plots:**

****

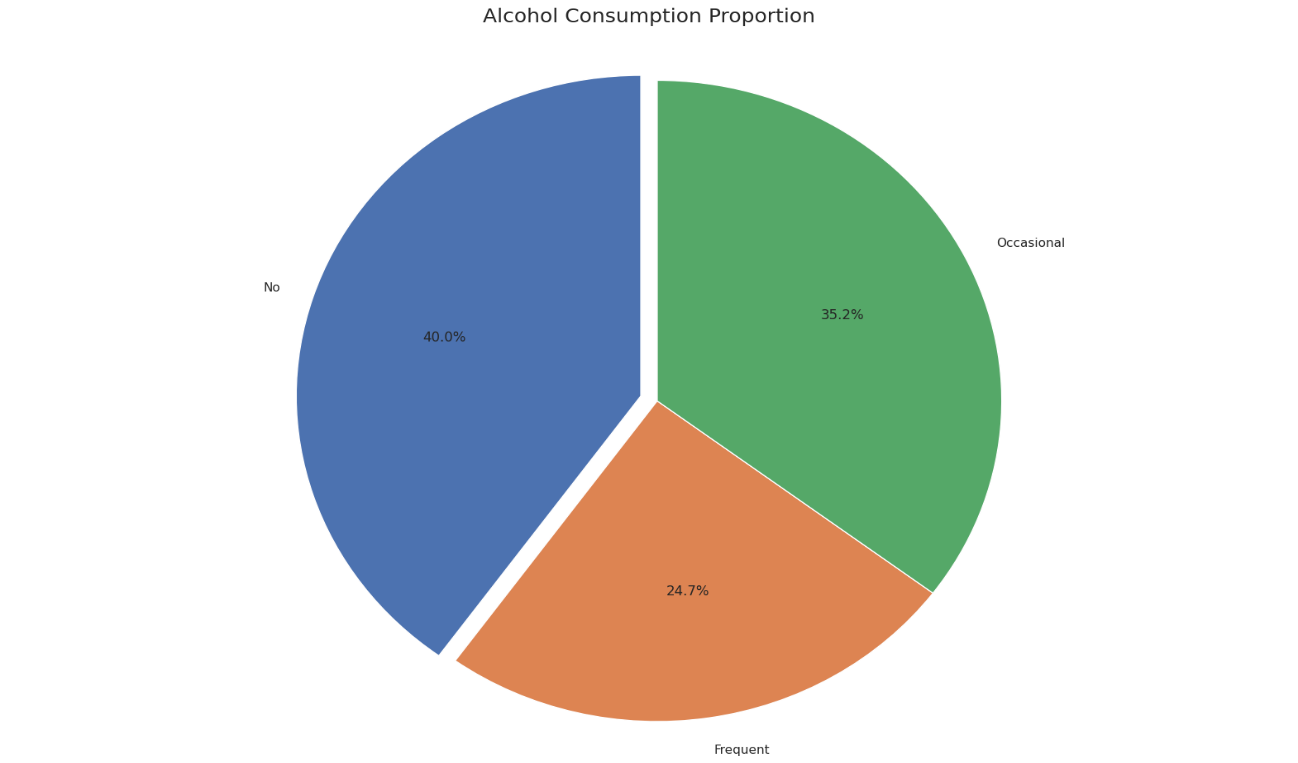
In this section, we created a donut chart for the Gender proportion of the human based on their gender. That is, how many of the data were male and how many of data are female. We created a donut chart for the based on their proportion. If we consider total percentage of the human as 100%, from the graph, we can see that 49.6% were male and 50.4% were female.



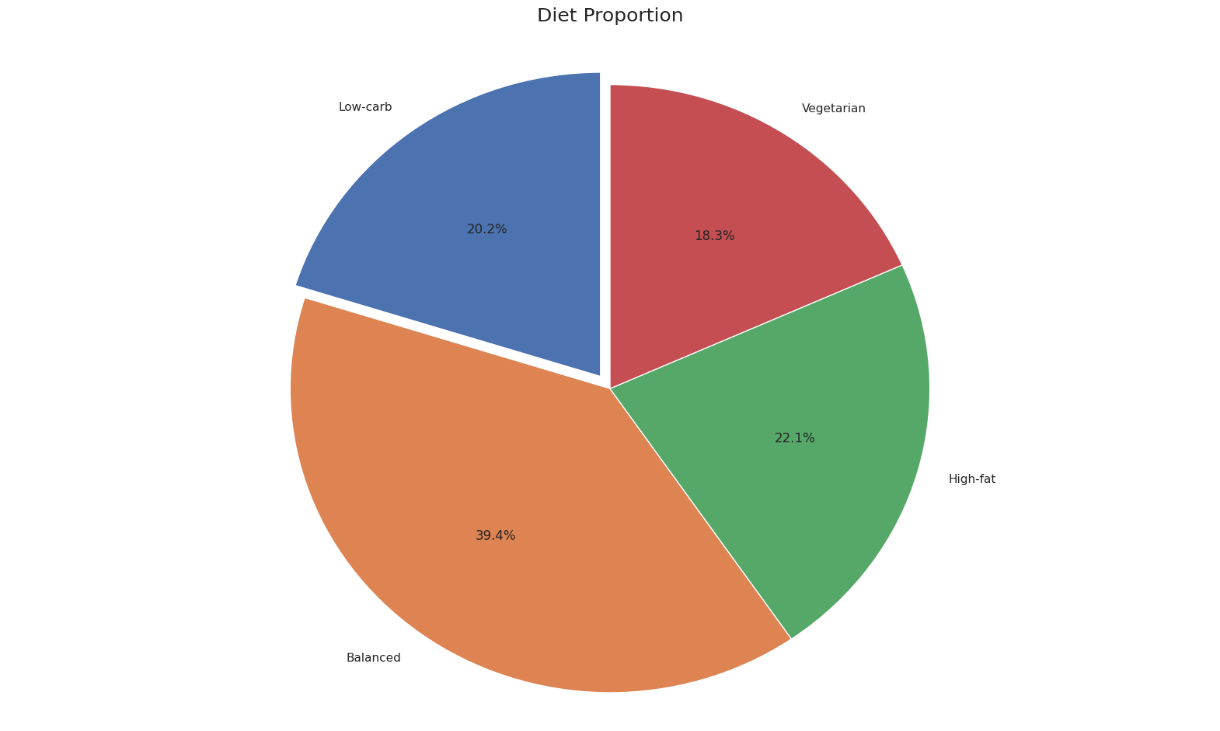
In this section, we created a pie chart for the Physical Activity Level Proportion based on our data. That is, how many of the humans have high activity level, how many have moderates activity level and how many have low activity level. We created a pie chart for the based on their proportion. If we consider total percentage of the data as 100%, from the graph, we can see that 30.1% of the data have high activity level, 46.9% have moderate activity level and 23.0% have low activity level.



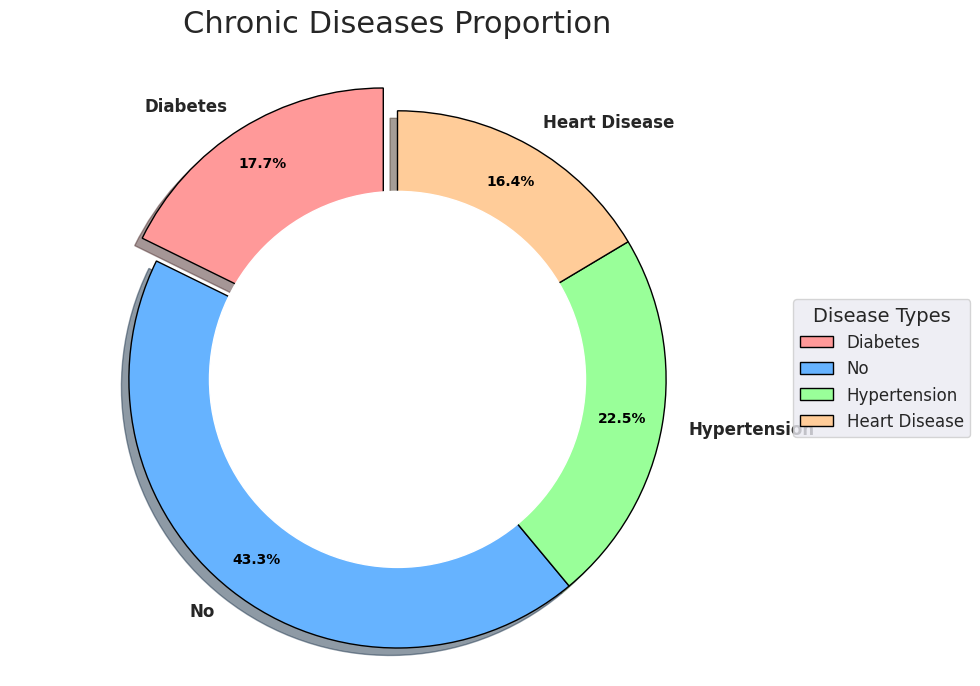
In this section, we created a donut chart for the Smoking Status proportion of individuals based on their smoking habits. That is, how many of the data points represent never, formar and current smokers. We created a donut chart based on their proportion. If we consider the total percentage of individuals as 100%, from the graph, we can see that 34.2% were Never, 26.4% were current and 39.4% were former smokers.



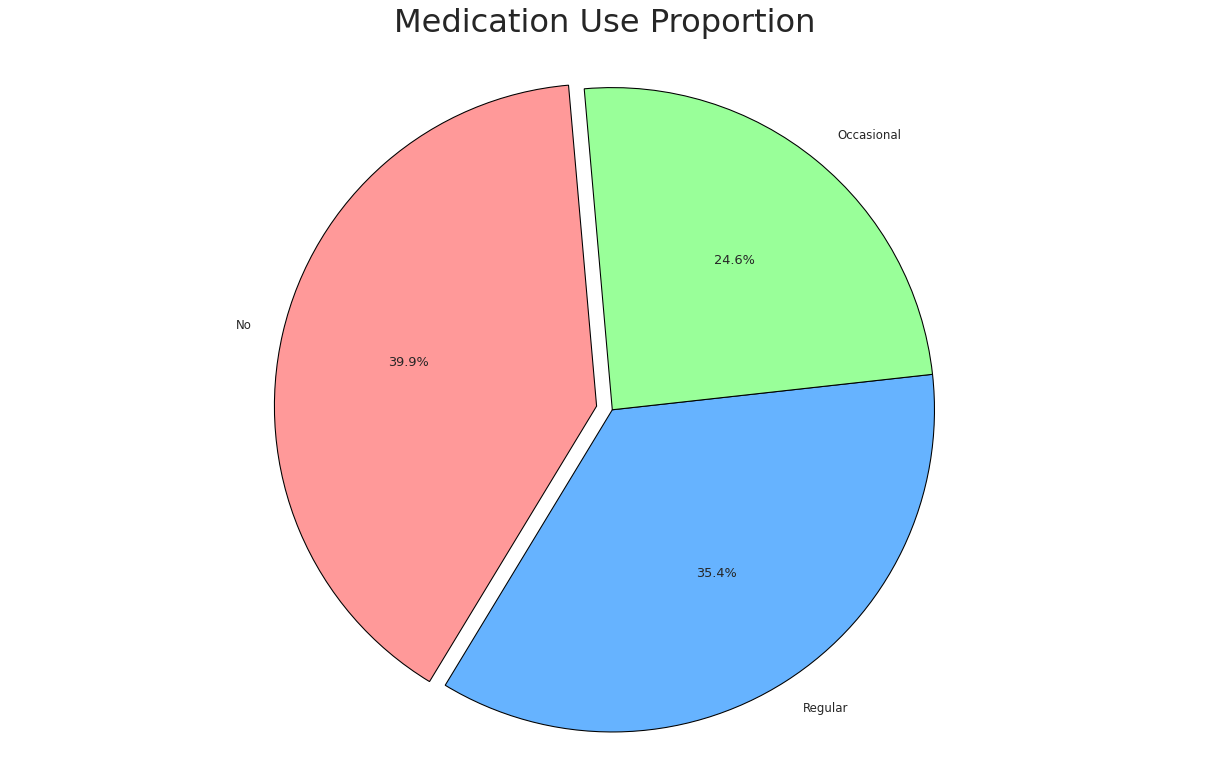
In this section, we created a pie chart for the Alcohol Consumption Proportion based on our data. That is, how many individuals do not consume alcohol, how many are frequent drinkers, and how many are occasional drinkers. We created a pie chart based on their proportion. If we consider the total percentage of the data as 100%, from the graph, we can see that 40.0% of the individuals do not consume alcohol, 24.7% are frequent drinkers and 35.2% are occasional drinkers.



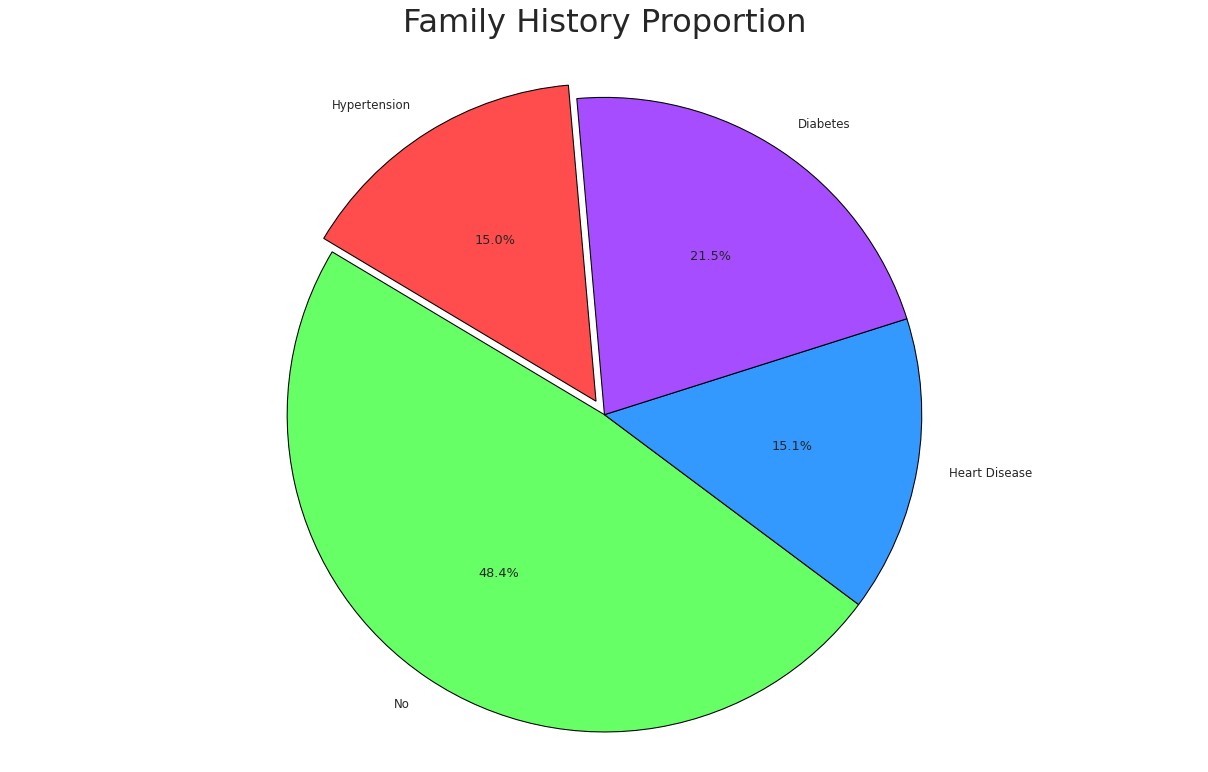
In this section, we created a pie chart for the Diet Proportion based on our data. That is, how many individuals follow a low-carb diet, a balanced diet, a high-fat diet, or a vegetarian diet. We created a pie chart based on their proportion. If we consider the total percentage of the data as 100%, from the graph, we can see that 20.2% of the individuals follow a low-carb diet, 39.4% follow a balanced diet, 22.1% follow a high-fat diet, and 18.3% follow a vegetarian diet.



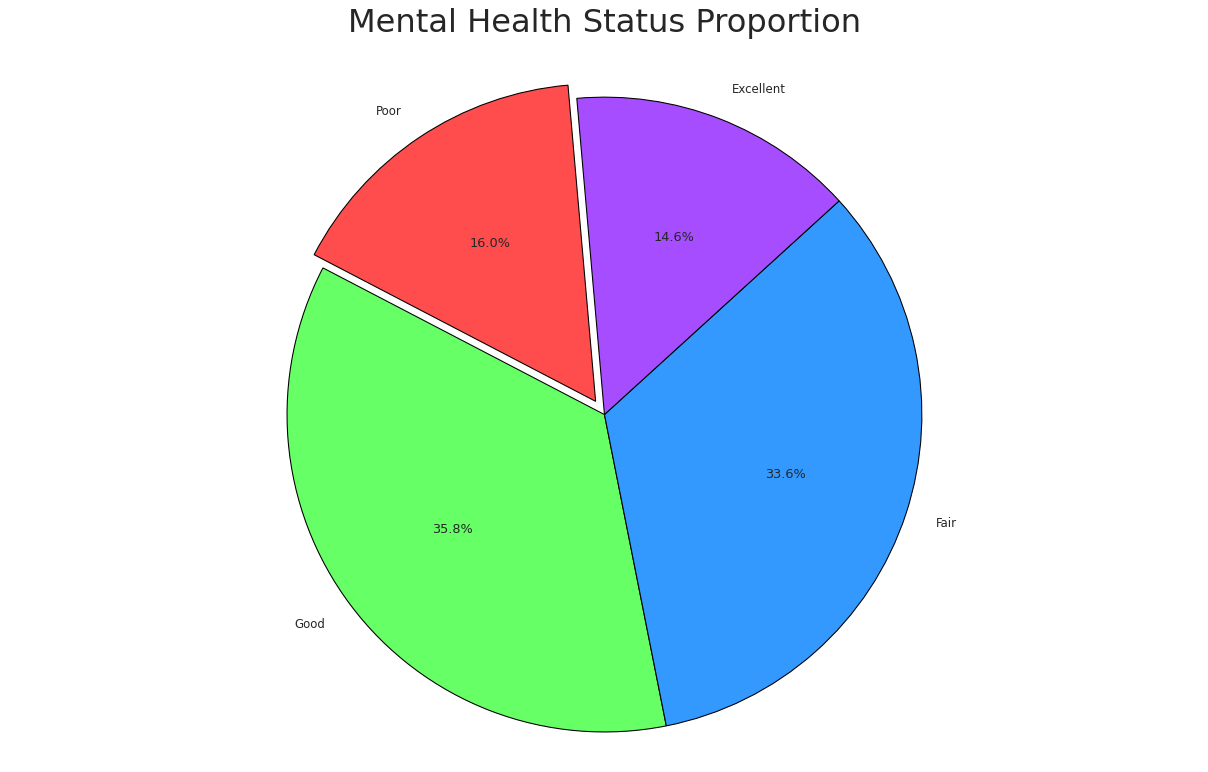
In this section, we created a donut chart for the Chronic Diseases proportion based on the health conditions present in the data. That is, how many of the individuals have diabetes, hypertension, heart disease or none of these conditions. We created a donut chart based on their proportion. If we consider the total percentage of individuals as 100%, from the graph, we can see that 17.7% had diabetes, 43.3% had no chronic disease, 22.5% had hypertension and 16.4% had heart disease.



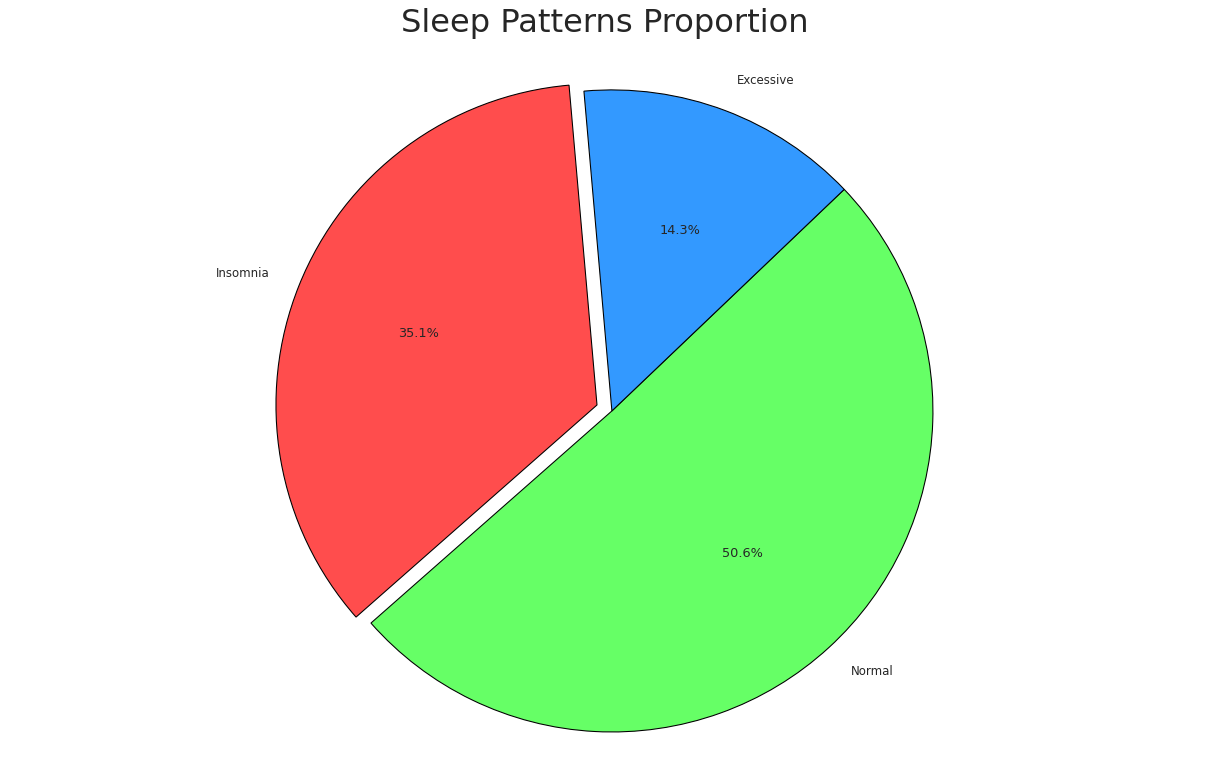
In this section, we created a pie chart for the Medication Use Proportion based on our data. That is, how many individuals do not use medication, how many use medication regularly, and how many use medication occasionally. We created a pie chart based on their proportion. If we consider the total percentage of the data as 100%, from the graph, we can see that 39.9% of the individuals do not use medication, 35.4% use medication regularly, and 24.6% use medication occasionally.



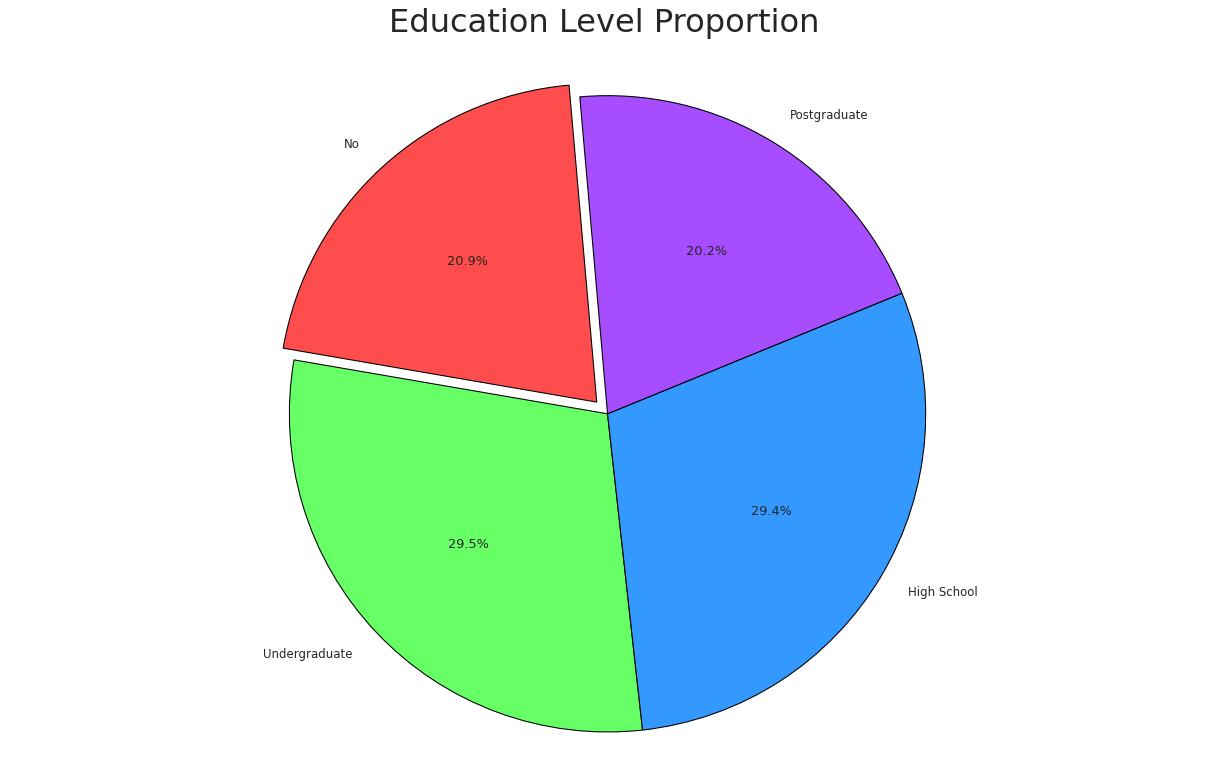
In this section, we created a pie chart for the Family History Proportion based on our data. That is, how many individuals have a family history of hypertension, no family history of chronic diseases, a history of heart disease or a history of diabetes. We created a pie chart based on their proportion. If we consider the total percentage of the data as 100%, from the graph, we can see that 15.0% of the individuals have a family history of hypertension, 48.4% have no family history of chronic diseases, 15.1% have a family history of heart disease and 21.5% have a family history of diabetes.



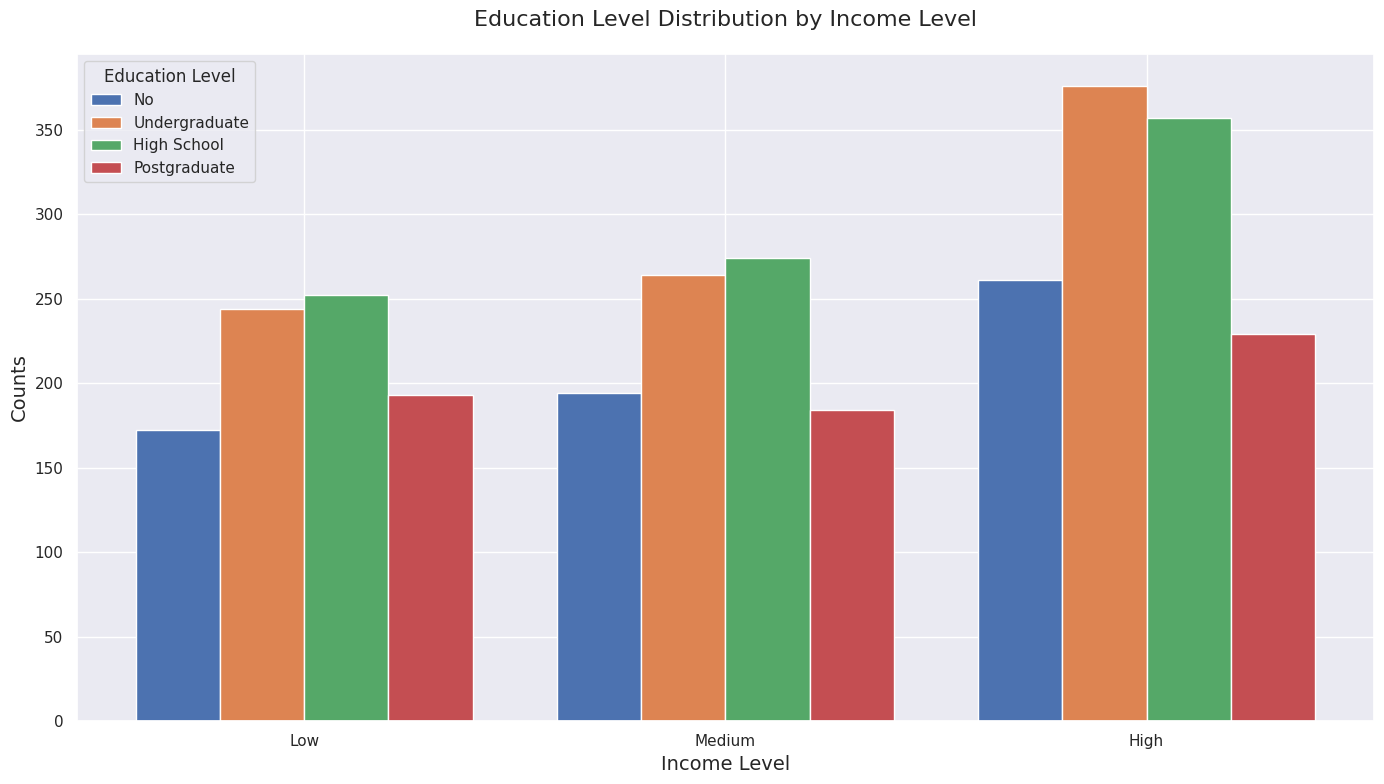
In this section, we created a pie chart for the Mental Health Status Proportion based on our data. That is, how many individuals reported poor, fair, good or excellent mental health. We created a pie chart based on their proportion. If we consider the total percentage of the data as 100%, from the graph, we can see that 16.0% of the individuals reported poor mental health, 33.6% reported fair mental health, 35.8% reported good mental health and 14.6% reported excellent mental health.



In this section, we created a pie chart for the Sleep Patterns Proportion based on our data. That is, how many individuals experience insomnia, have normal sleep patterns or have excessive sleep. We created a pie chart based on their proportion. If we consider the total percentage of the data as 100%, from the graph, we can see that 35.1% of the individuals experience insomnia, 50.6% have normal sleep patterns and 14.3% have excessive sleep.



In this section, we created a pie chart for the Education Level Proportion based on our data. That is, how many individuals have no formal education, completed high school, pursued undergraduate studies or obtained a postgraduate degree. We created a pie chart based on their proportion. If we consider the total percentage of the data as 100%, from the graph, we can see that 20.9% of the individuals have no formal education, 29.4% completed high school, 29.5% pursued undergraduate studies and 20.2% obtained a postgraduate degree.

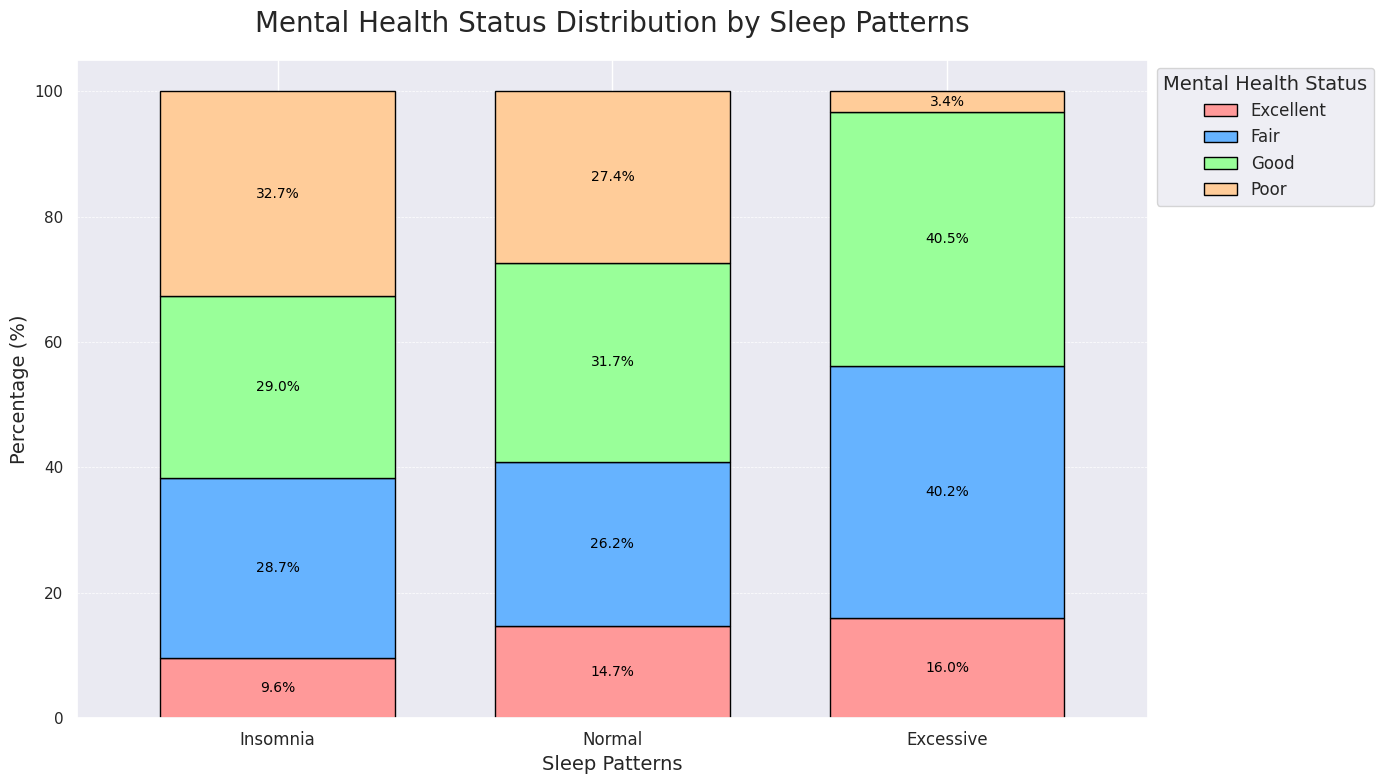


The conclusion deduced from the group bar chart analysis is, education and income are highly correlating. People who have at least finished high school or obtained a university degree are also more likely to retire with higher income levels than those without much education. On the other hand, there was data that showed while individuals with medium and low income levels have had a similar type of educational attainment which could mean education might aid in garnering better earnings.

Income can have large effects on other parts of life, and it also correlates with how rich someone truly is—how they live day-to-day, the healthcare that they access or not have enough of, and what insurances they may be covered by.

Having more money usually means you can afford to spend more of it on quality healthcare, nicer homes, and generally improved life decisions that contribute to better health and a longer lifespan. That relationship calls into question the prediction of age. Being able to afford food, housing and health care makes it easier for people to live a stable life full of physical and emotional well-being—and that may translate into a longer lifespan. People with more money generally have healthier lifestyles: better diets, regular check-ups and lower stress levels; they also have access to resources that can promote longevity. Meanwhile, those in lower income brackets may be exposed to multiple stressors and have fewer opportunities for receiving healthcare that could impact their health as well as life expectancy. In brief, this is a very complicated relationship between education, income and age prediction.

A better quality of life is associated with longevity and higher income, something which generally results from higher education levels.

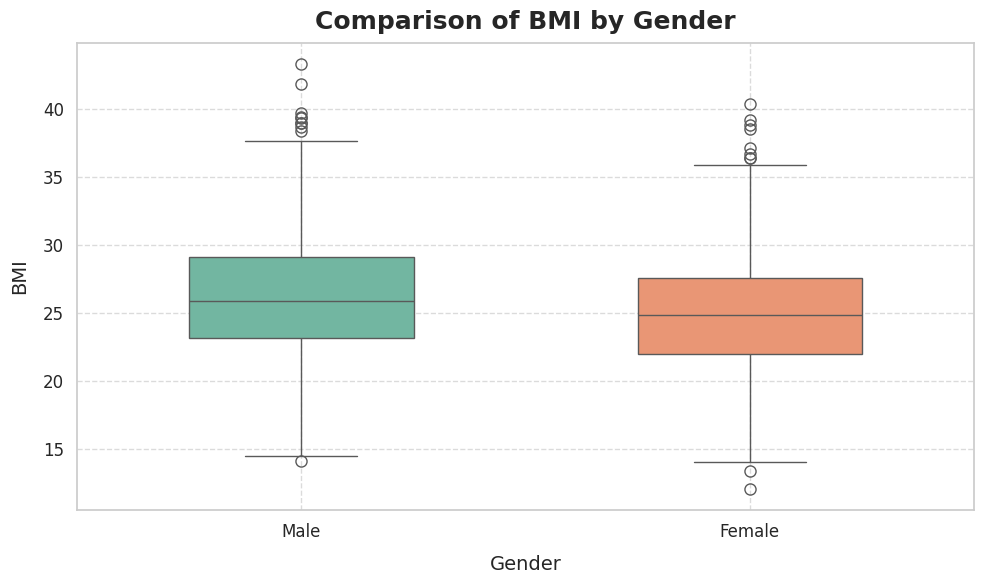


Inspection of the stacked bar chart makes it clear: those suffering from insomnia are vastly reporting worse mental health. Interestingly, merely 9.6% report excellent mental health and more than a third suffer poor demise with the highest in over a decade (32.7%). These data support a robust relationship between sleep disruptions and mental health phenotype.

On the opposite pole, those who report normal sleep distribution display a more positive – less polarized- distribution of mental health statuses. The numbers after adjusting for age are similar: 14.7% of this group have excellent mental health, while 26.2% call it fair; 31.7% call it good; and 27.4% say it is poor. This difference reflects the positive effects that normal sleep patterns may have on mental health more broadly.

In addition, a whopping 40 percent of those who sleep too much among folks with good mental health. Hereby showing that sleep enough and regularly is pivotal to our mental health — supporting the notion that sleep is needed for recovery, energy replenishment in the body.

On a more general level, sleep is critical for mental health and can dramatically impact biological age prediction datasets. Poor sleep quality has been known to contribute to mental well-being in a negative way, but can it have long-term health outcomes as well, suggesting the implementation effective sleep management strategies for healthier lives.

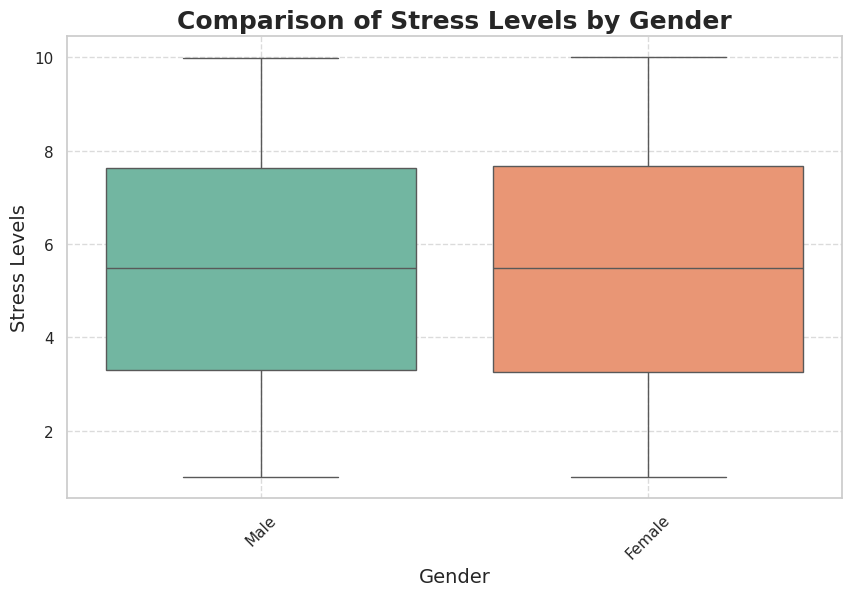


This boxplot is an example of comparing Body Mass Index (BMI) between males and females, showing some interesting outliers Cite this: Males have an average BMI of 25.35, whereas females have an average BMI recorded at 24.9. This means that, men have relatively higher BMI than women.

By the way, that famous 18–24.9 BMI healthy range places a substantial number of females in the pink fat patterns too. In men, the average BMI approaches the desired range although it is slightly above the upper limit, suggesting that some men may be overweight.

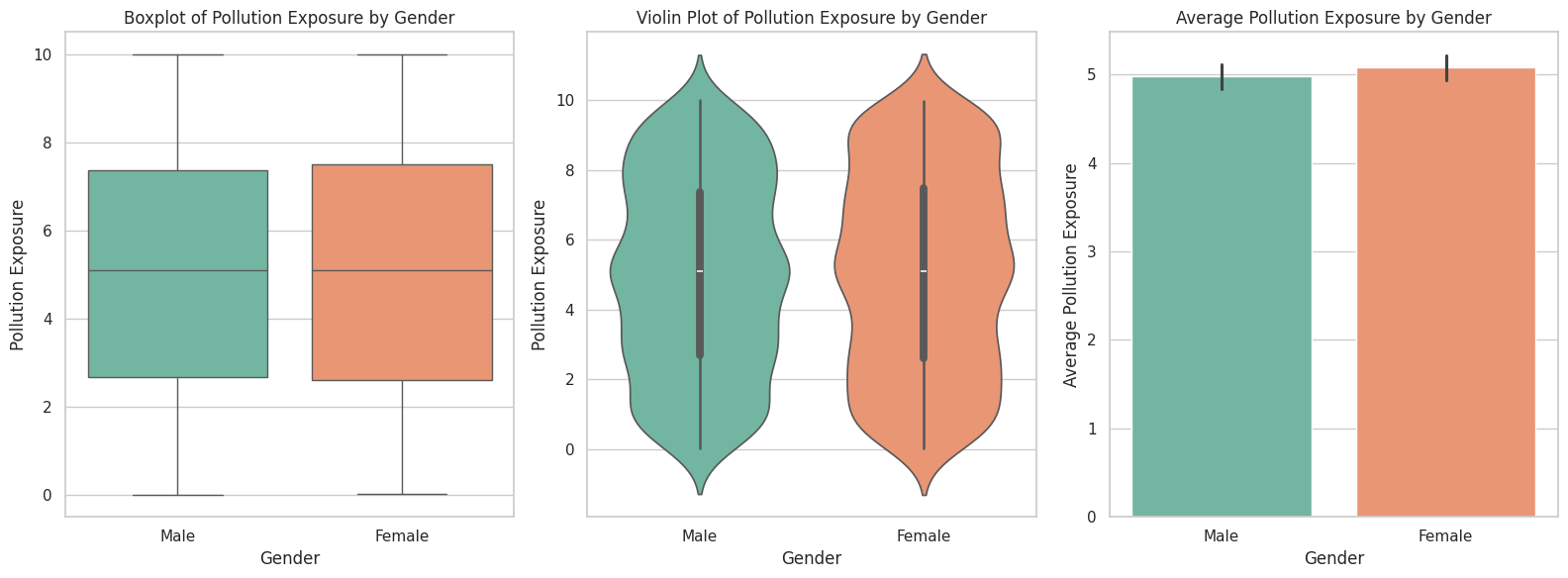
It is important to have Body Mass Index (BMI) in age prediction datasets since it shows the general health and physical condition. A low body mass index is linked to healthier life outcomes as well, including a lower risk of developing chronic conditions and greater physical health. So, awareness of this and tips for how to get (and stay) at an optimal BMI are important in a continuing effort towards healthy living and potentially longer lifespans.

Overall, this boxplot reveals that monitoring BMI is crucial not only for health assessments but also analyses of age prediction.



The boxplot visually describes the comparison of male and female stress levels. Interestingly, no outliers are present in the data, which means that stress levels appear to be more evenly distributed across the two groups.

Both men and women seem more or less equally spun up: their average stress levels reach 5.8 out of 10, the survey results suggested. This finding implies that men and women undergo the same types of stress, indicating environmental or social factors (familial structure, choice of hobbies, professional duties) that lead to similar levels of stress for sentinels have an impact on both species. All in all, the boxplot shows that stress perception is an average point and no outliers are present.



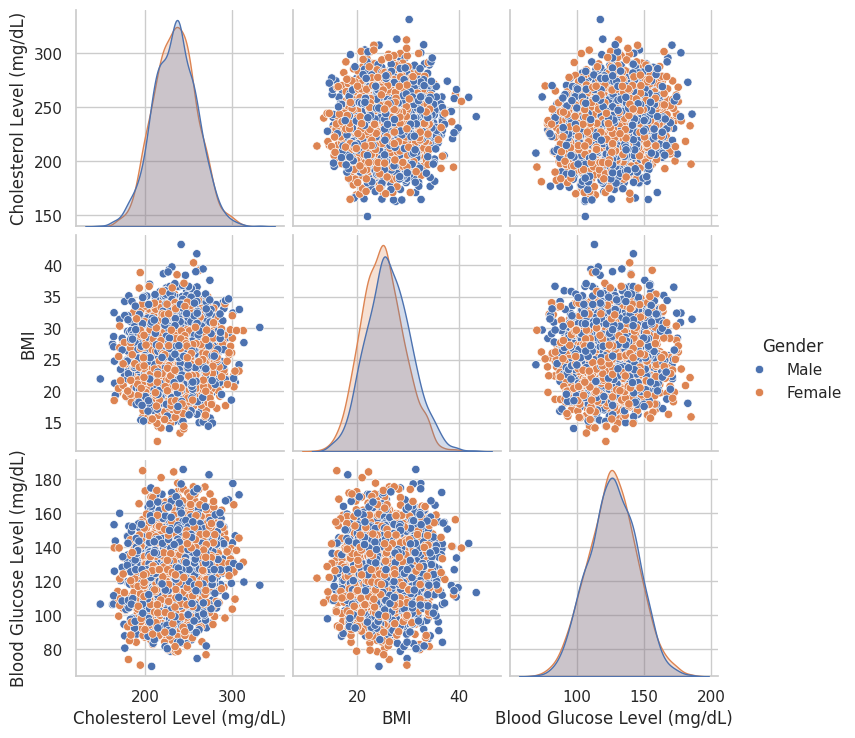
To do this comparison we will take a look at pollution exposure by gender through a boxplot, a violin plot and bar chart in the following analysis.

The boxplot shows that there is no outlier in the data, which in turn indicated a constant distribution of pollution exposure levels. The other is the revelation that females experience even higher levels of pollution than males. This trend is also confirmed by the violin plot that shows a more precise distribution of pollution exposure which and density that confirms female are more hit by pollution than males.

A bar chart for the average pollution exposure levels makes a quantitative case again for what we already observed — female pollution exposure averages 5.1 and male exposure at 6. That tiny variation is representative of the usual flux observed in other plots of this type.

In age prediction datasets this could be valuable information, since constant exposure to pollution over years is related to numerous health complications that can have a substantial impact on life span and overall healt. Understanding sex differences in exposure to pollutants can help us design public health interventions and policies, which target these risks.

**3.Scatter plots:**

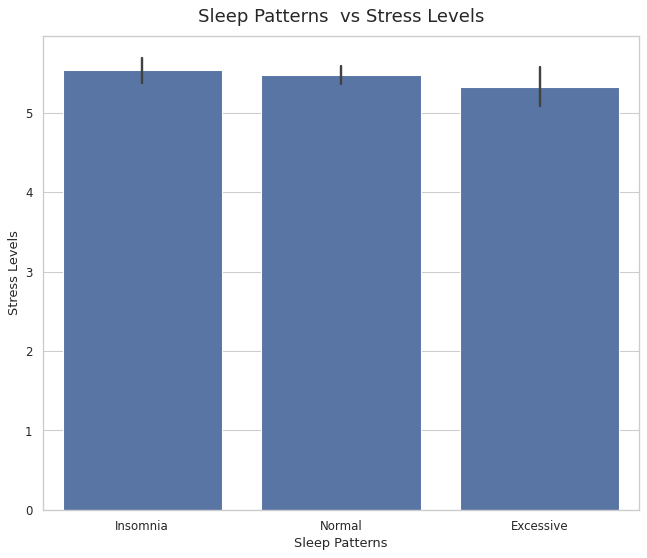


The pair-plot offers a comprehensive comparison of three critical health metrics: lipid profiles, anthropometrics, and biochemical parameters, including cholesterol levels, BMI, and blood glucose levels; the results were presented separately by the gender. The circle is labelled to show that the male participants are depicted by the blue markers while the female participants are depicted by the yellow markers.

From the analysis, several insights are obtained as explained below. First, it emerges that male subjects had a higher cholesterol mean than females, which could mean that male could be at a greater risk of cardiovascular diseases compared to female and therefore there might be a need to encapsulate this aspect to formulate health policies and intervention for the male subjects in particular. Further, if taking into account the difference in parameters between males and females, it can also be observed that males have a higher average BMI, or which might be linked to lifestyle or diet choices.

On the other hand, female’s outcome reveals high blood glucose level than male outcome. This can be either an indication that there are health problems that need medical attention including those involving high risks of diabetes and other diseases associated with high blood glucose levels.

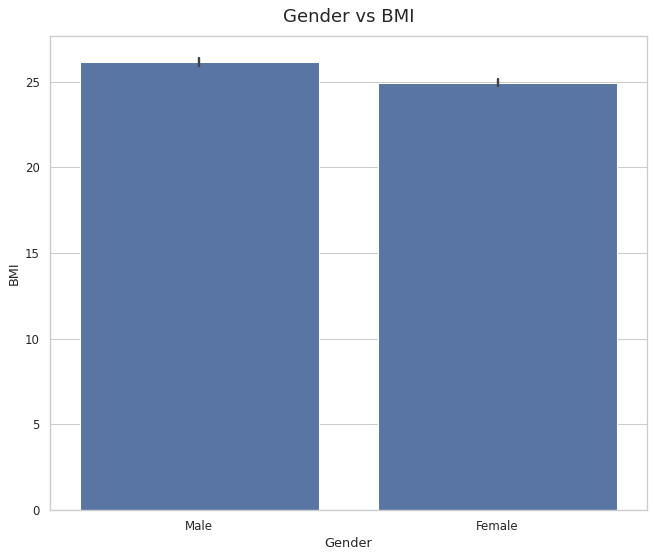
In general, the pair-plot is helpful in revealing the gender differences in these health indicators while always highlighting the relationships between those indicators.



The bar plot displayed in the analysis shows the correlation between the sleep patterns and stress level and it is clearly visible that people with isomnia stress levels are visibly higher than people with other types of sleeping habits. This association indicates that sleep disorders have a significant potential to impair mental health and cause various limitations in everyday life.

It is important not to define the effects of these findings in terms of stress alone; lack of sleep also has a ripple effect on other aspects of an individual’s lifestyle, efficiency, and psychological well-being. Constant stress and sleep deprivation are among the factors that cause different physical illnesses, for example, cardiovascular diseases, obesity, and a weakened immune system. Thus, disturbed sleep patterns may have negative effects on the health of an individual in the long run that can even affect the entire life cycle of the person and his potential lifespan.

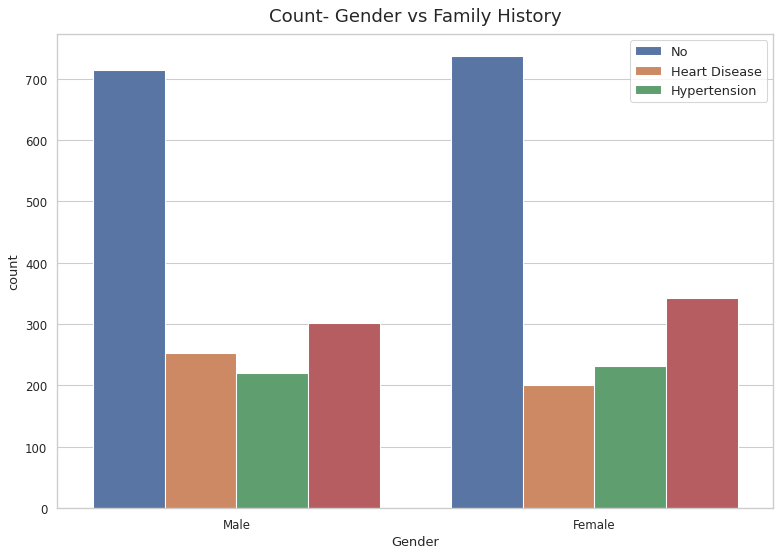
To sum up, stress and all the threats connected to it recall that healthy sleep is a crucial factor to prevent stress. Thus, acknowledging the connections between sleep stress it is possible to encourage rolling of certain changes in a daily routine promoting a better night sleep and, therefore, improving the physical health all through the human life cycle.



This barplot is an example of comparing Body Mass Index (BMI) between males and females, Males have an average BMI of 25.35, whereas females have an average BMI recorded at 24.9. This means that, men have relatively higher BMI than women.

By the way, that famous 18–24.9 BMI healthy range places a substantial number of females in the pink fat patterns too. In men, the average BMI approaches the desired range although it is slightly above the upper limit, suggesting that some men may be overweight.

Overall, this boxplot reveals that monitoring BMI is crucial not only for health assessments but also analyses of age prediction.

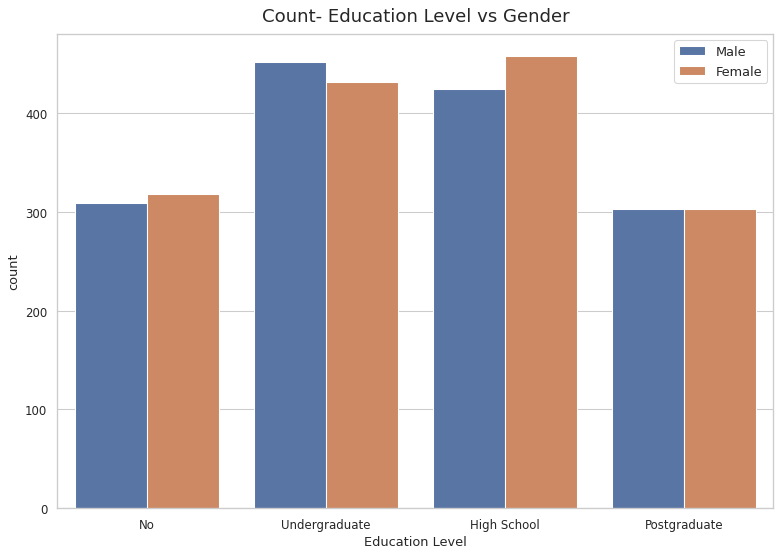


The count plot offers a simplistic view of the health status of families headed by either a male or a female. The information that has been analyzed concerns shows that female-headed families usually have fewer diseases in comparison with male-headed families.

For instance, male families indicate an increased prevalence of this ailment commonly referred to as heart disease which can be explained by numerous causes such as preliminary life choices, type of food intake, and heredity among others. Such findings call for analysing the social and environmental factors affecting family health to address the observed differences in health status.

These disparities may, for example, indicate that female-headed families are more likely to make health-supportive decisions, place a higher premium on prevention, or have superior infrastructures for coping with these diseases. On the other hand, the trend toward higher rates of heart disease among male families may suggest the need for enhanced health promotion efforts, such as raising awareness of heart disease risk factors, improving access to preventative care, and adopting healthy lifestyles.

Overall, the count plot raises worries regarding gender distinctions in family health status and underlines the need for gendered public health approaches. Thus, raising awareness of these discrepancies is a step toward promoting improved health and physical conditions in families, which in turn will benefit the health of society.



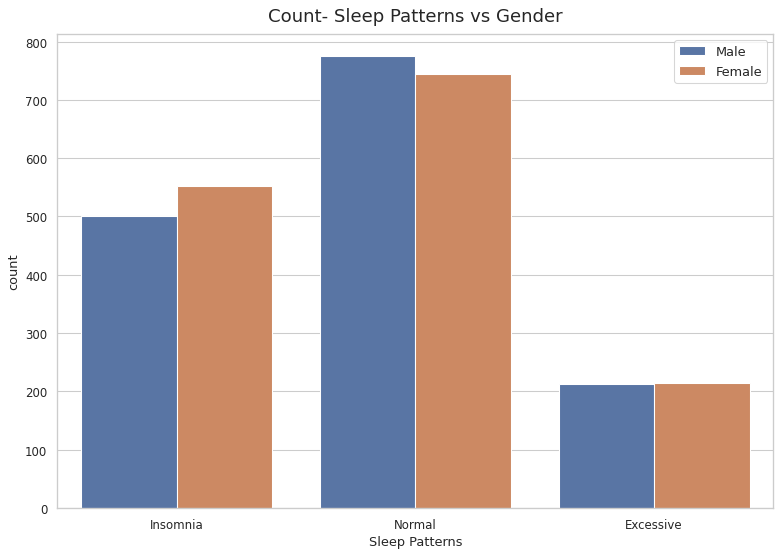
The count plot clearly signifies specially the educational attainment achieved by the male and female subjects. This proves that a larger number of male students complete their undergraduate degree than female students meaning that male students might have better chance or are more inclined in having higher education at this level.

On the other hand, female has a higher passing rate in high school than male and this indicates women could easily finish the secondary education. This has clearly shown that there is a change in the earlier disparities in the educational attainment since females are almost catching-up with males in earlier stages of education.

As for the last level which goes down to the doctoral, the data reveals that males and females have quite similar level of PhD holders, thus implying that once the male and female have entered the higher tier of education, their completion rates are quite similar.

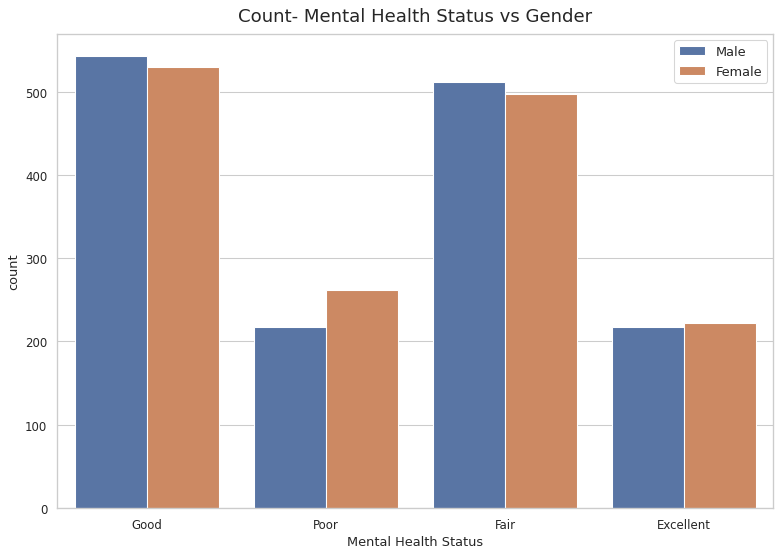
Furthermore, the count plot shows that females with no formal education are higher in number than male with no formal education. This discovery is important because it brings into focus a cause for concern mapping to limit Democratic space that of some females may experience to access education.

All in all, the count plot that has been presented in the paper provides quite sophisticated trends in educational attainment by gender, as well as successes and inequalities. These findings may be useful in designing the educational policies and kindergartens’ programs and practices that can support the equality of genders and access to education for everyone.



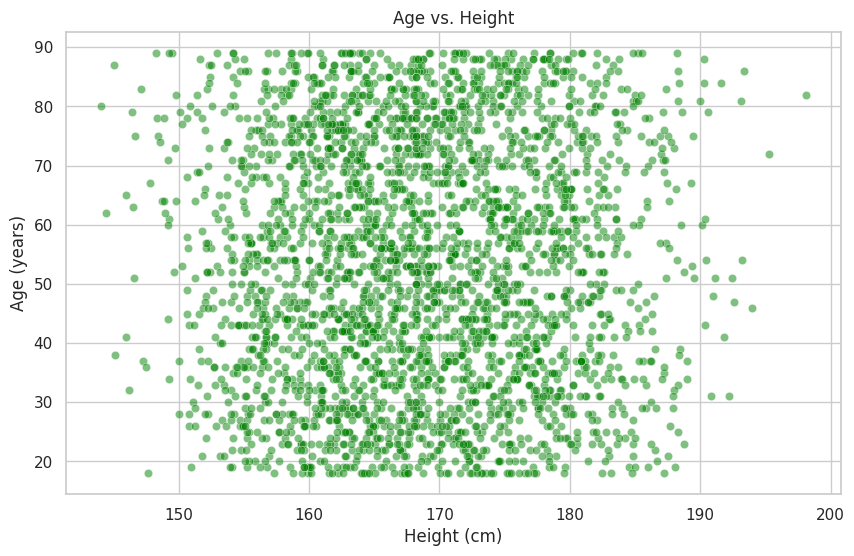
This is most easily visualized in the count plot which shows some very interesting differences in sleep habits between the sexes. This reveals that males prefer to have normal sleep pattern more frequently than females meaning that they might enjoying better quality and longer sleep.

On the other hand, females have a higher prevalence of isomnia which means that they have many difficulties to overcome in order to have a sound sleep. The reason for such a divergence might be found in fluctuations of sex hormones and stress levels, as well as in gender- specific workload in women.

Sleeping differences of the two groups bring light into essential aspects of mental as well as physical well-being. Although normal sleep is effective for male, greater prevalence of isomnia in females that leads to enhanced stress level and fatigue may impair their health and quality of life.

The count plot shows that a higher percentage of females report poorer mental health than males, thus highlighting the gender gap in mental health. This implies that women are more vulnerable to experience psychological disorders than men due to certain factors including, social, stress and some health conditions.

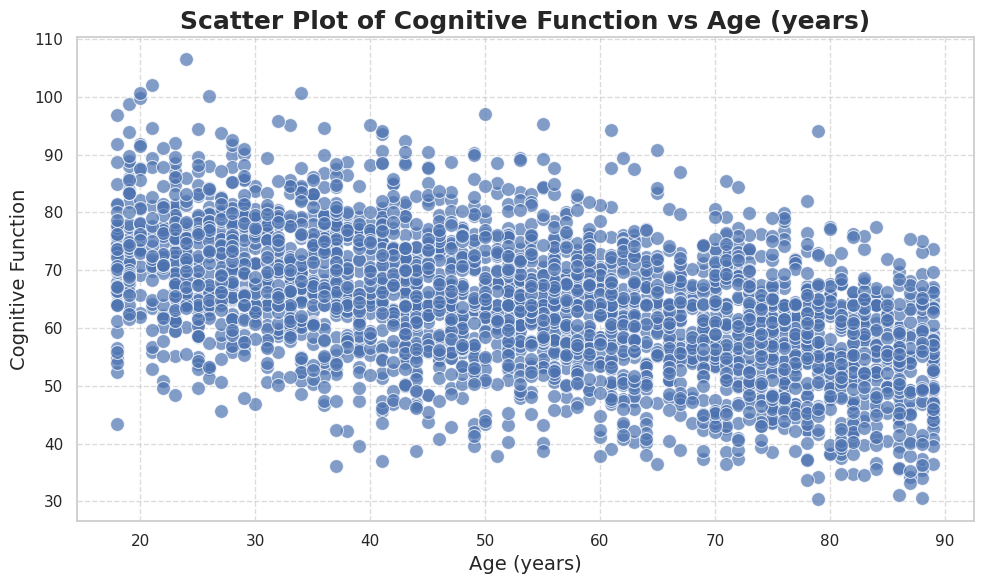
Thus, having discussed the quantitative distribution of males and females by the criteria of good, fair, and excellent mental health, it is possible to conclude that the observed differences are not significant. This means that whereas a greater percentage of females suffer with poor mental health, females in the better classifications: good, fair, and excellent; have comparable results as their counterparts in males.



From the scatter plot, we notice that majority of the people in the given age limit lie within the height range of 160cm – 175 cm. The heights associated with this cluster raise an argument that these heights are normal among the aged people in the given data set.

It will be important knowing this height distribution as a mean to analyze the relation between age and height besides enabling computation of general health status and demographic examinations. It also reflects possibly genetic, or an environmental or nutritional factor which determines heights in people of certain ages.

All in all, we can observe a reasonable degree of population density in regard to the specified height range and it is worthy of further research concerning physical well-being and longevity.



The scatter plot suggests that individuals in their early 20s and 30s exhibit higher levels of creativity and intelligence, as well as enhanced cognitive function, compared to other age groups. This trend may be attributed to various factors, including the neurological development that peaks during these years, as well as increased exposure to new ideas and experiences.

**Machine Learning Models**

In this project, we are using linear regression model to predict our target variable (Age).

Linear Regression is a fundamental statistical method used to model the relationship between a dependent variable and one or more independent variables. The model assumes a linear relationship, meaning it predicts the dependent variable as a weighted sum of the independent variables plus a constant (intercept). The goal is to find the best-fitting line that minimizes the sum of squared differences between the actual and predicted values. It is widely used for predicting and understanding relationships between variables and is easy to implement and interpret.

**Advantages:**

* Easy to interpret and implement.
* Provides clear insight into the relationship between variables.

**Disadvantages:**

* Assumes linearity, which may not always be true.
* Sensitive to outliers, which can distort the results.
* Can be affected by multicollinearity in multiple linear regression.

**Reason of using this model:**

The dataset that we are working on is to predict age. The target variable (Age) is a continuous numerical variable, making Linear Regression a suitable model since it predicts continuous outcomes. Linear Regression provides a straightforward, interpretable model that shows how features like height, weight, and gender contribute to age prediction. It serves as a good baseline to compare against more complex models. If it performs well, it may reduce the need for more sophisticated methods. Linear Regression is computationally efficient and scales well with large datasets.

Due to the continuous nature of target variable (Age), Linear Regression is selected to perform as model.

**Results:**

By running this model in our dataset, the prediction was well performed as we get the Mean Squared Error (MSA) about 30.71 which can be considered as accepted error to perform a model.

**Conclusion**

The synthetic dataset effectively demonstrates how features relate to predicting human age. Linear Regression was used because of its simplicity and interpretability. While the model performed well, adding more relevant features or using more complex models could improve accuracy. The findings from the model can be applied in various contexts, such as targeted marketing strategies or public health initiatives, where understanding age demographics is crucial. The significant predictors identified can inform future research and decision-making processes.

The synthetic nature limits real-world application, but it's great for practicing machine learning techniques. Future work could explore non-linear models and additional features like lifestyle factors to better capture the nuances in age prediction. This dataset serves as a valuable tool for learning and experimentation.