**Overview**

The dataset includes features like Patient ID, Age, Sex, Blood Pressure, Heart Rate, Cholesterol Levels, Diabetes Presence, Family History, Smoking Habits, and Obesity that are connected to the health of the patients.

Along with lifestyle characteristics, it also covers sleep hours per day, days per week of physical activity, and hours spent being sedentary each day. The dataset also includes geographic information like country, continent, and hemisphere, as well as socioeconomic characteristics like income.

The variable of interest is "Heart Attack Risk," indicating a potential emphasis on evaluating the risk of heart attacks through a range of lifestyle and health-related factors.

**Introduction to the Dataset**

The data frame originally had 8763 rows and 26 columns.

The columns are:

'Patient ID', 'Age', 'Sex', 'Cholesterol', 'Blood Pressure', 'Heart Rate', 'Diabetes',

'Family History', 'Smoking', 'Obesity', 'Alcohol Consumption', 'Exercise Hours Per Week', 'Diet', 'Previous Heart Problems', 'Medication Use', 'Stress Level', 'Sedentary Hours Per Day', 'Income', 'BMI', 'Triglycerides', 'Physical Activity Days Per Week', 'Sleep Hours Per Day', 'Country', 'Continent', 'Hemisphere', 'Heart Attack Risk'.

**Data pre-processing**

The dataset is found to have no null values or duplicate values so further processing is done to facilitate the data analysis.

Processing the data by converting ‘Blood Pressure’ into two components, namely ‘Systolic' and 'Diastolic’ and taking their mean value as ‘Mean blood pressure’ and is added to the dataset while dropping the ‘Blood Pressure’, we end up with the following columns:

'Patient ID', 'Age', 'Sex', 'Cholesterol', 'Heart Rate', 'Diabetes', 'Family History', 'Smoking', 'Obesity', 'Alcohol Consumption', 'Exercise Hours Per Week', 'Diet', 'Previous Heart Problems', 'Medication Use', 'Stress Level', 'Sedentary Hours Per Day', 'Income', 'BMI', 'Triglycerides', 'Physical Activity Days Per Week', 'Sleep Hours Per Day', 'Country', 'Continent', 'Hemisphere', 'Heart Attack Risk', 'Mean Blood Pressure'.

**Statistical methods employed**

The statistical methods employed include: mean, mode, range, standard deviation and variance.

**Data visualization methods employed**

The data visualization methods include boxplots, correlation matrix, pie chart imported from the matplotlib.pyplot python library and also histograms from seaborn library.

**Statistical analysis**

Due to the large size of this dataset, statistical analysis and comparisons will be elaborated on for only select columns.

Overview of Demographics:

The population's average age is 53.71 years, which represents a wide range of age groups from 18 to 90 years.

The age distribution shows that people between the ages of 35 and 72 are overwhelmingly present.

Engaging in Physical Exercise:

10.01 hours a week on average are spent exercising, demonstrating a dedication to an active lifestyle.

There is a large range of exercise hours—from 0.002 to 19.99 hours per week.

There may be a subgroup with less reported physical activity based on the mode at 0.002 hours.

BMI (body mass index):

With an average BMI of 28.89, it is generally considered to be overweight.

The BMI spectrum, which spans from 18.00 to 39.99, illustrates the wide variation of body weight distribution among the population.

Heart-Related Health:

With an average triglyceride level of 417.68, this is noticeably high and indicates the need for actions to reduce cardiovascular risks.

With a range of 75 to 145, the mean blood pressure is 110 on average, suggesting a moderate variation in cardiovascular health.

Days of Physical Activity:

An average of 3.49 days a week are spent exercising by people.

There is variation in the distribution; some people report no physical activity at all, while others report engaging in it seven days a week.

Sleeping Habits:

7.02 hours is the average daily sleep time, which is in accordance with suggested guidelines.

There is a minimum of 4 hours and a maximum of 10 hours of sleep per day.

**Gender-Based Health Metrics Analysis**

Population of Women

Females are 53.06 years old on average.

The average heart rate is 75.36 beats per minute, while the average cholesterol level is 258.94.

The prevalence of diabetes is 0.65, and roughly 49.1% of people report having a family history of heart problems.

Of the female population, 65.9% smoke.

One estimate of the obesity rate is 49.96%.

Of the female population, 59.65% report consuming alcohol.

10.08 hours a week on average are spent exercising by women.

About 49.43% of respondents indicate having had prior cardiac issues.

It is reported that approximately 50.38% of females use medication.

It is stated that the average stress level is 5.56.

5.98 hours a day on average are spent inactive.

It is stated that the mean blood pressure is 110.26.

On average, women participate in physical activity on 3.52 days each week.

The duration of sleep on average is 7.04 hours.

The BMI is 28.92 on average.

The average triglyceride level is 416.68.

Population of Men:

For men, the average age is 53.99 years.

The average heart rate is 74.87 beats per minute, while the average cholesterol level is 260.28.

The prevalence of diabetes is 0.65, and roughly 49.37% of people report having a family history of heart problems.

It is reported that every guy in the sample smokes.

The percentage of obese people is at 50.22%.

Of the men, 59.88% report drinking alcohol.

An average of 9.99 hours are spent exercising each week by men.

About 49.65% of respondents indicate having had prior cardiac issues.

Approximately 49.60% of men report using medication.

According to reports, the average stress level is 5.43.

An average of six hours a day are spent inactive.

It is said that the mean blood pressure is 110.05.

An average of 3.48 days a week are spent exercising for men.

The duration of sleep on average is 7.02 hours.

The BMI is 28.88 on average.

The average triglyceride level is 418.11.

**Regional Health and BMI Analysis Report**

Africa

Africa's average age spans from 18 to 90 years old, with an average of 54.78 years.

The range of BMIs is 18.06 to 39.99, with a mean of 29.04.

The data shows a tendency toward higher BMI values as well as a varied age distribution.

Asia

Asia's average age ranges from 18 to 90 years old, with an average of 53.75 years.

The BMI ranges from 18.02 to 39.98, with a mean of 28.82.

Asia has a varied age distribution and BMI range, much like Africa.

Australia

Australia has an average age of 53.45 years, with a range of ages from 18 to 90.

The range of BMIs is 18.01 to 39.99, with a mean of 28.93.

Australia's age distribution and BMI are comparatively stable.

Europe

In Europe, the average age is 53.26 years, with a range of 18 to 90 years.

The BMI ranges from 18.00 to 39.98, with a mean of 28.87.

Similar to other continents, Europe exhibits a wide age distribution and BMI range.

North America

North America has an average age of 54.51 years, with a range of ages from 18 to 90.

The BMI ranges from 18.00 to 39.97, with a mean of 29.11.

The average BMI is somewhat higher and the age distribution is more varied in North America.

South America

South America has an average age of 53.34 years, with a range of ages from 18 to 90.

The range of BMIs is 18.01 to 39.99, with a mean of 28.81.

Similar to other continents, South America has a varied age distribution and BMI values.

**Age versus activity level analysis**

BMI Patterns

Across the age span, the BMI values range from 27.53 to 30.31.

It's noteworthy that as people age, their BMI tends to rise, peaking at age 55 at 30.31. This raises the possibility of a relationship between growing older and greater body bulk.

Exercise Duration

The range of 19.8 to 20 exercise hours per week is a stable pattern that is constant across age groups.

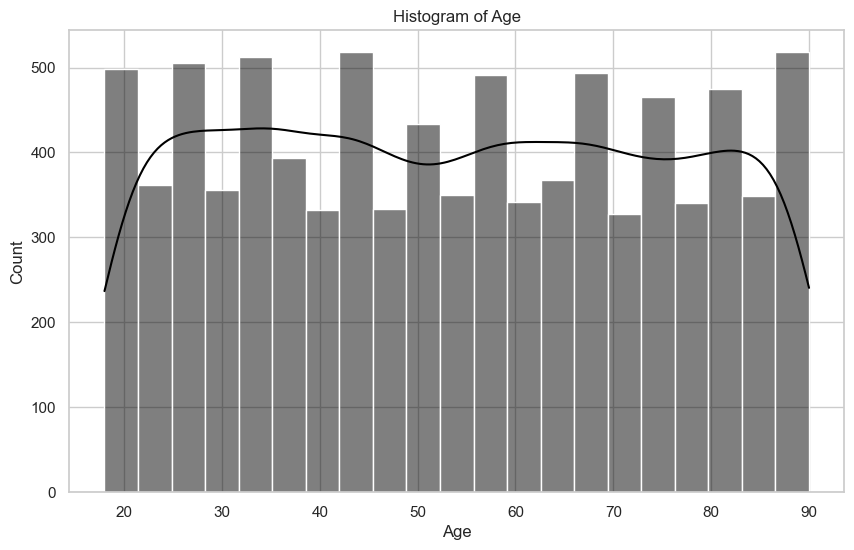
It's interesting to note that there is a small decline in exercise hours between the ages of 27 and 37, which may be a result of events in life or changes in lifestyle that affect exercise habits.

Correlation Examination

An initial investigation of the association between age and exercise hours points to a modest negative correlation, suggesting that there may be a trend for weekly exercise to slightly decline as people age.

**Visualisation of the dataset:**

Age Distribution

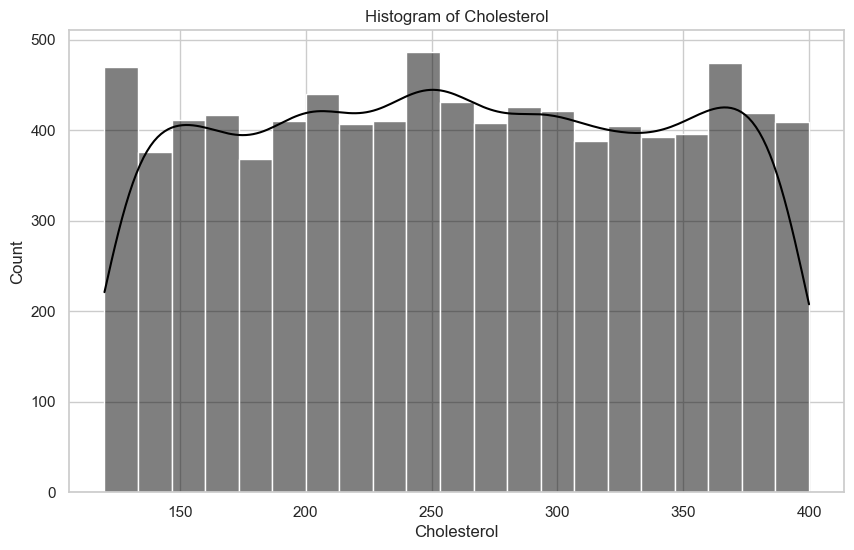


Skewness = 0.03, Kurtosis = -1.21

With a skewness of 0.03, which is quite close to 0, the distribution is almost symmetrical and has a slight rightward trend.

The distribution appears to be less peaked and more flat than a normal distribution, as indicated by the kurtosis value of -1.21.

Cholesterol distribution

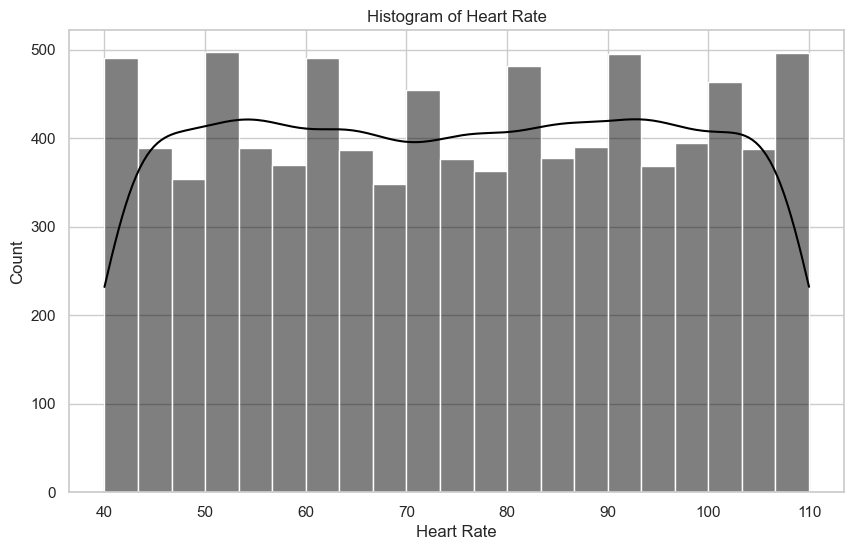


Cholesterol: Skewness = -0.00, Kurtosis = -1.18

A skewness number that is extremely near to zero (-0.00 in this example) indicates that the cholesterol level distribution is roughly symmetrical.

The distribution is flatter and has lighter tails than a normal distribution when the kurtosis value is negative (-1.18). This implies that compared to a normal distribution, the cholesterol distribution contains fewer extreme values, or outliers.

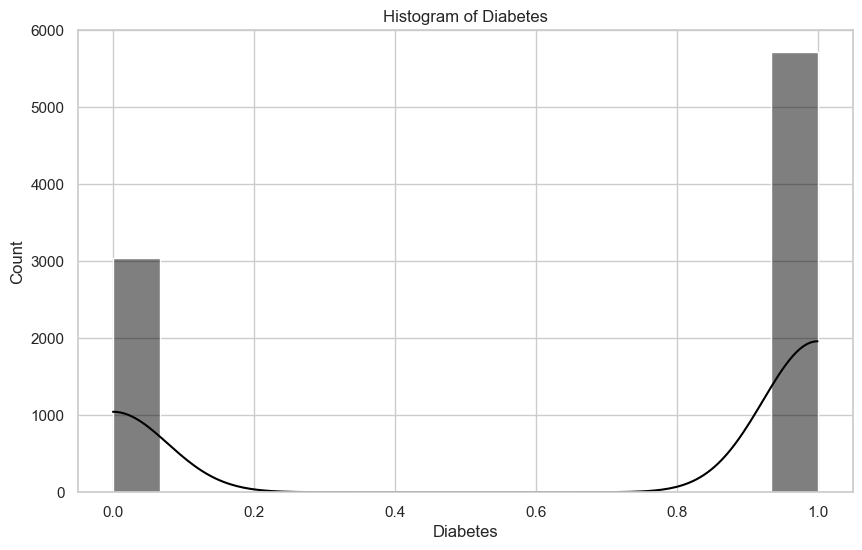
Heart rate distribution

Heart Rate: Skewness = -0.00, Kurtosis = -1.21

Given that the skewness is almost zero, the distribution appears to be about symmetrical.

The distribution is less peaked and has lighter tails than a normal distribution, according to the negative kurtosis.

Diabetes count:



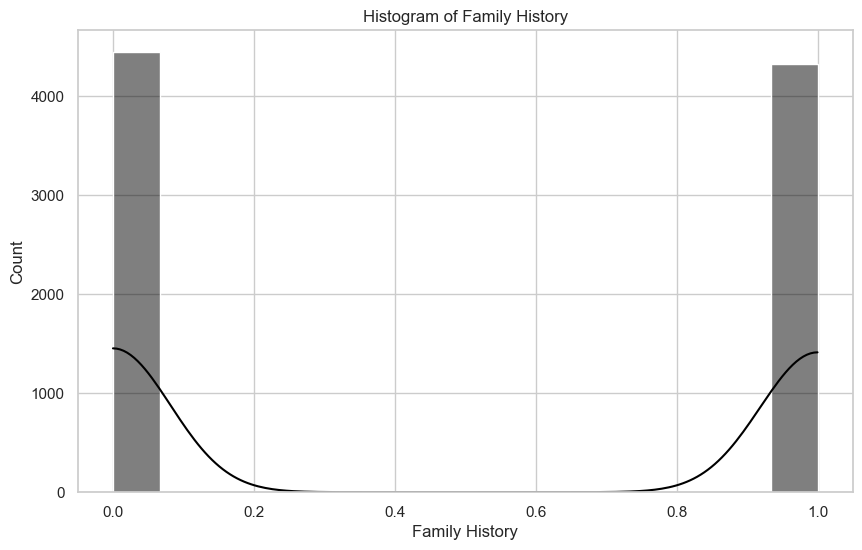
Diabetes: Skewness = -0.64, Kurtosis = -1.59

We can see that close to 6000 people have diabetes while 3000 people don’t.

The distribution may be left-skewed, as suggested by the negative skewness, which implies that most people may have lower diabetes-related values.

A platykurtic distribution, which has flatter peaks and thinner tails than a normal distribution, is suggested by the negative kurtosis. This could mean that there are less extreme values or outliers in the diabetes dataset.

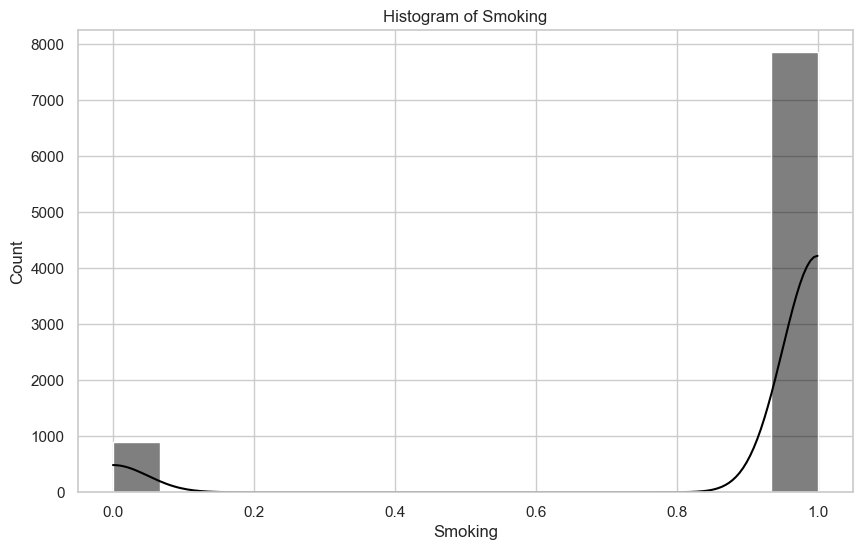
Family history of heart problems

Family History: Skewness = 0.03, Kurtosis = -2.00

One could infer a slightly right-skewed distribution from a skewness value of 0.03. This suggests that there may be a few more people in the data than would be predicted from a perfectly symmetrical distribution if there was a family history of cardiac problems.

A distribution with lighter tails than a normal distribution is indicated by a kurtosis score of -2.00. This implies that there are less extreme values, or outliers, in the data than a perfectly normal distribution would predict.

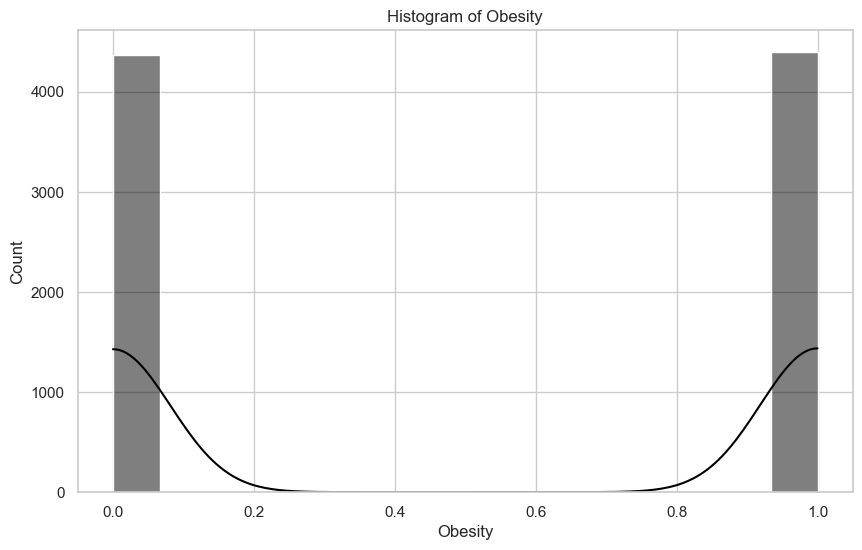
Smoking count

Smoking: Skewness = -2.61, Kurtosis = 4.81

We can infer that the majority of the samples in this dataset are smokers.

The distribution of the variable "Smoking" is negatively skewed (left-skewed), with heavier tails than a normal distribution, according to the values of -2.61 for skewness and 4.81 for kurtosis. Higher kurtosis indicates that there are more extreme values in the distribution's tails than in a normal distribution, whereas negative skewness denotes that the dataset may have a concentration of lower values.

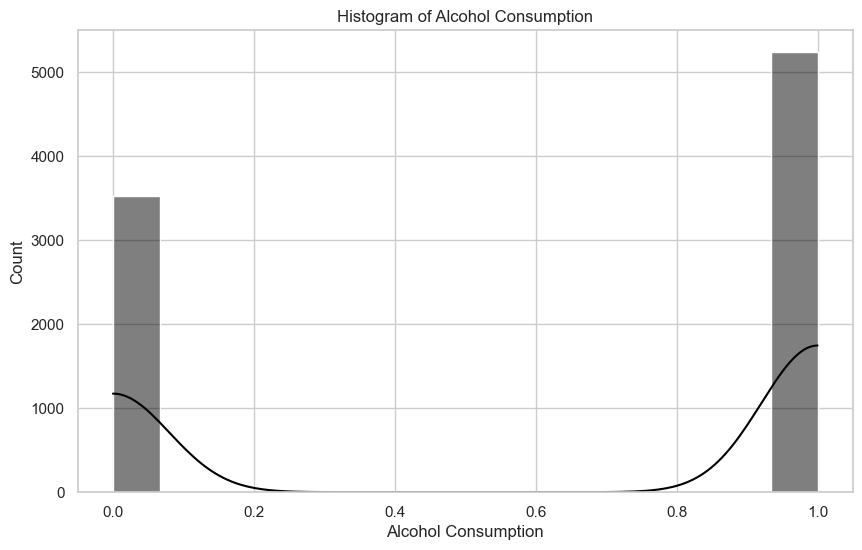
Obesity count

Obesity: Skewness = -0.01, Kurtosis = -2.00

A distribution that is nearly symmetrical but slightly skewed to the left is shown by a skewness of -0.01, which indicates a very minor negative skewness.

A platykurtic distribution, which differs from a normal distribution in having flatter peaks and thinner tails, is indicated by a kurtosis of -2.00. In comparison to a normal distribution, this indicates that the distribution has fewer extreme values, or outliers.

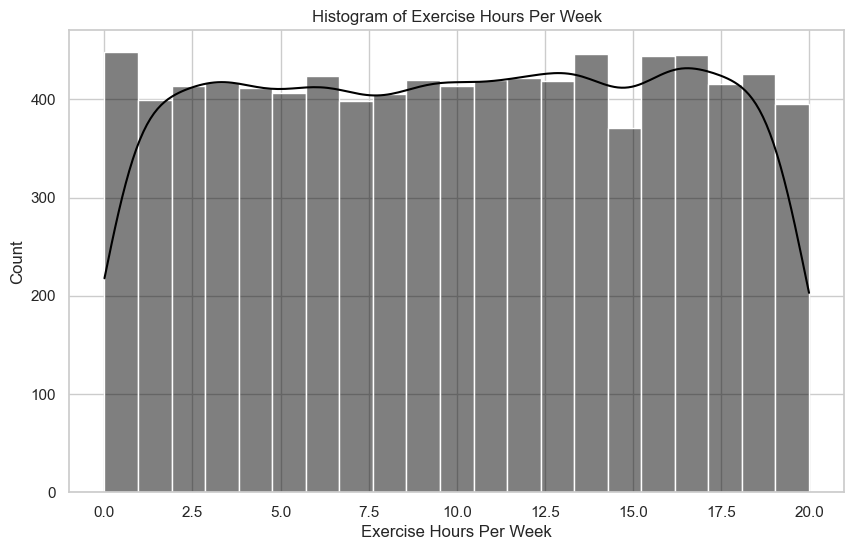
Alcohol consumption

Alcohol Consumption: Skewness = -0.40, Kurtosis = -1.84

Skewness = -0.40: This indicates that there may be a concentration of lower values with a tail extending to the right, and that the distribution of alcohol consumption is considerably skewed to the left.

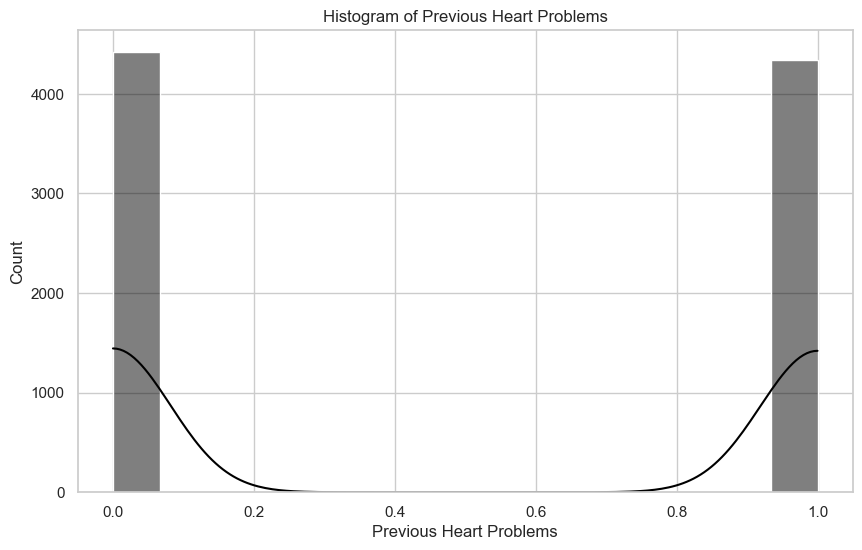
Kurtosis = -1.84: The distribution has lighter tails than a normal distribution, according to the negative kurtosis. In comparison to a normal distribution, it suggests that the distribution of alcohol consumption is less peaked and contains fewer extreme values.

Exercise hours per week distribution

Exercise Hours Per Week: Skewness = -0.02, Kurtosis = -1.20

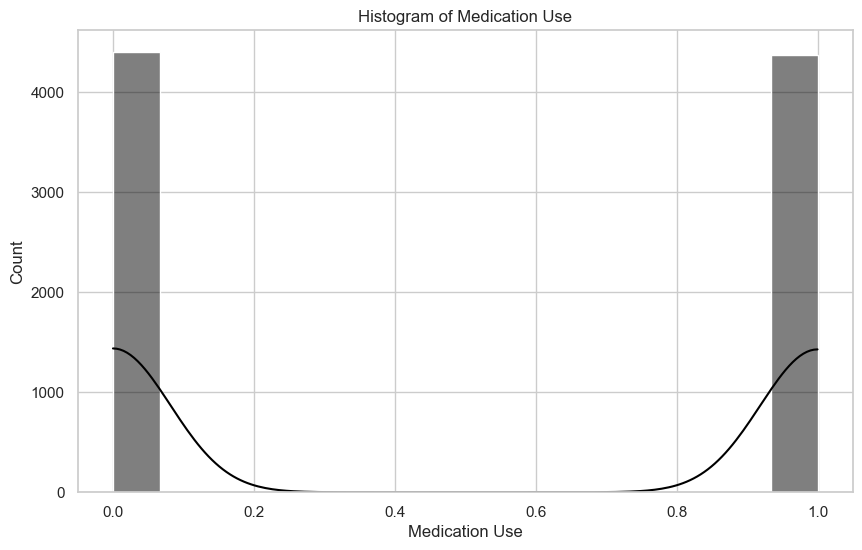
A negative skewness indicates that there may be a small number of people who exercise considerably more than the average, indicating that the distribution of exercise hours is somewhat skewed to the left. In comparison to a normal distribution, the distribution has fewer extreme values, or outliers, according to the negative kurtosis.

Previous heart problems count



previous Heart Problems: Skewness = 0.02, Kurtosis = -2.00

The skewness of 0.02 and kurtosis of -2.00, respectively, show that the past heart issues variable has a slightly positively skewed distribution with lighter tails than a normal distribution.

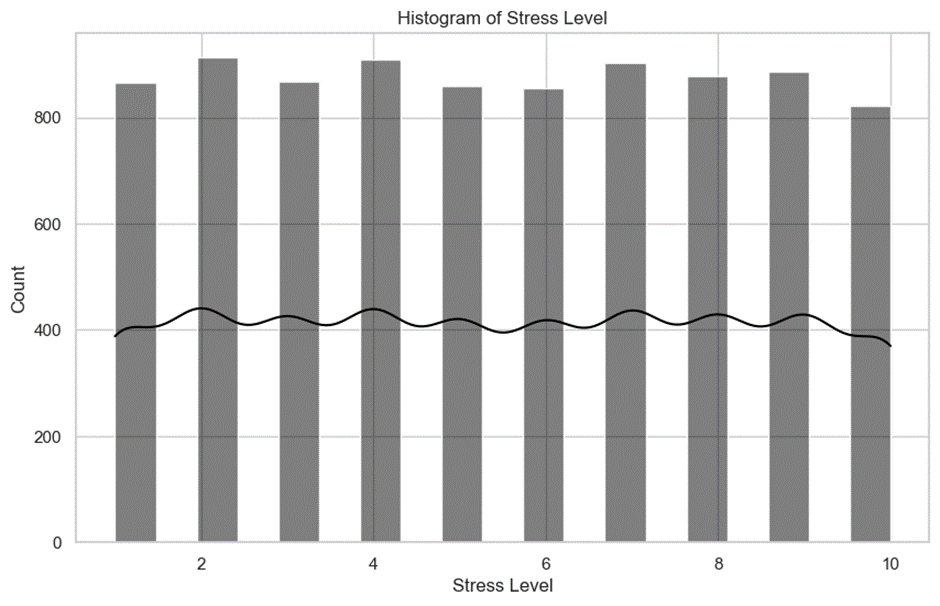
Medication use

Medication Use: Skewness = 0.01, Kurtosis = -2.00

One could infer a virtually symmetric distribution from a skewness value of 0.01. The positive skewness in this case suggests a small tail to the right, indicating that while the majority of the data is distributed pretty evenly, there may be a small number of people with higher Medication Use scores.

A platykurtic distribution is indicated by a kurtosis value of -2.00. This indicates that, in comparison to a normal distribution, the distribution has flatter peaks and thinner tails. It implies that there are fewer extreme values, or outliers, in the data when it comes to medication use than there would be in a normal distribution.

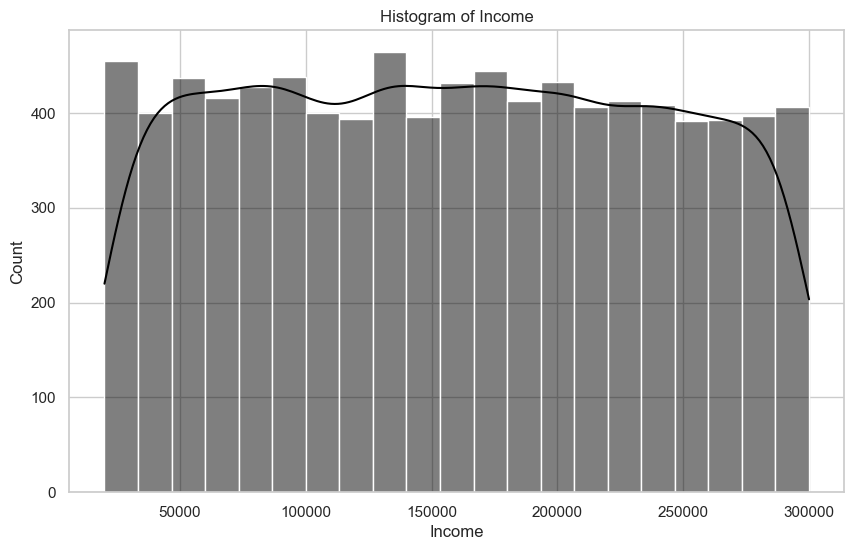
Stress level distribution



Stress Level: Skewness = 0.01, Kurtosis = -1.23

The distribution of the stress level variable seems to be almost symmetrical, with a minor inclination toward a flatter peak and thinner tail. The distribution's form can be inferred from the values of skewness and kurtosis, which indicate that the population's stress levels are not severely skewed or marked by extreme values.

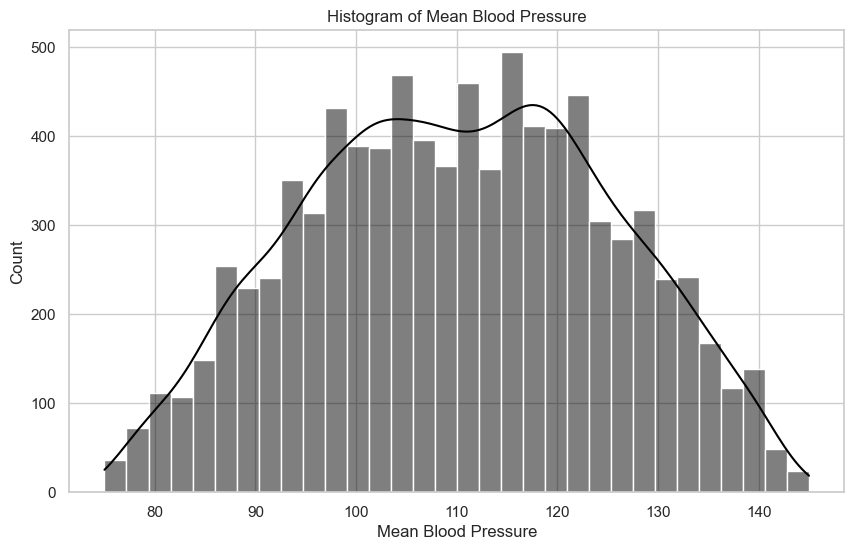
Income distribution



Income: Skewness = 0.02, Kurtosis = -1.18

Although it is less peaked and has a little longer right tail than a normal distribution, the income distribution is roughly symmetrical. These attributes offer valuable perspectives on the form and central tendency of the income data.

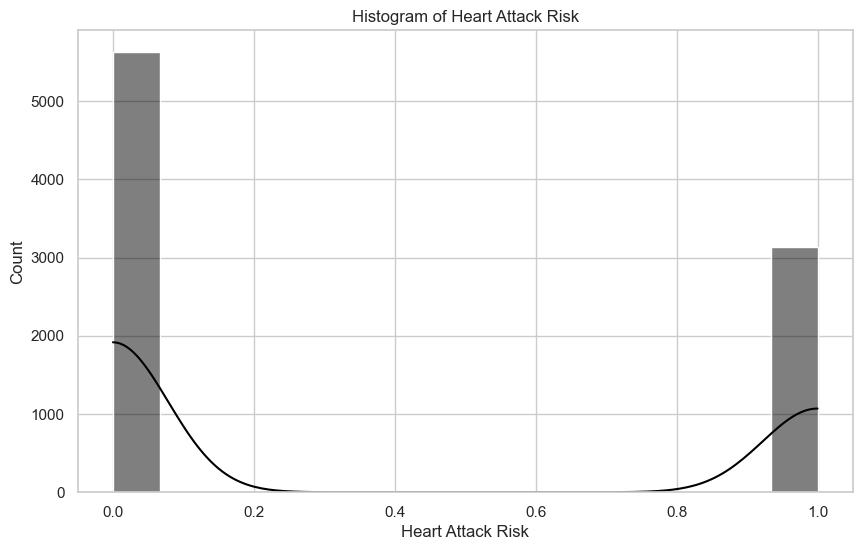
Mean blood pressure distribution



Mean Blood Pressure: Skewness = -0.02, Kurtosis = -0.77

Based on the above skewness and kurtosis values, the distribution of mean blood pressure looks to be rather symmetric and has slightly lighter tails than a normal distribution.

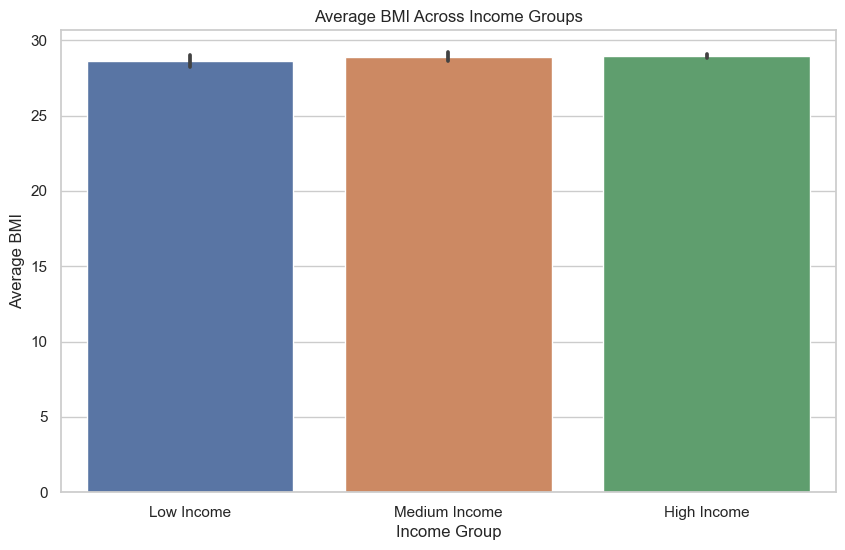
Heart attack risk distribution



Heart Attack Risk: Skewness = 0.59, Kurtosis = -1.65

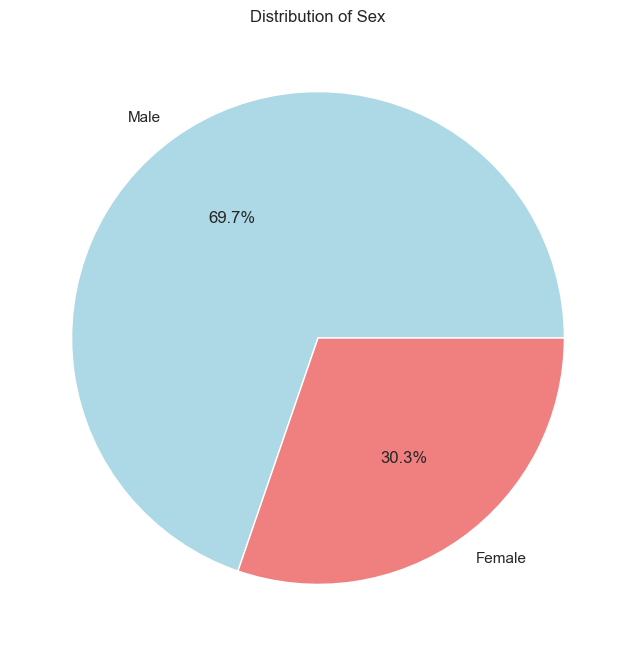
"Heart Attack Risk" variable appears to have a slightly right-skewed distribution with values concentrated toward the lower end of the scale. The platykurtic nature suggests that the variable's distribution has relatively fewer extreme values and is less peaked than a normal distribution.

Average BMI across income groups



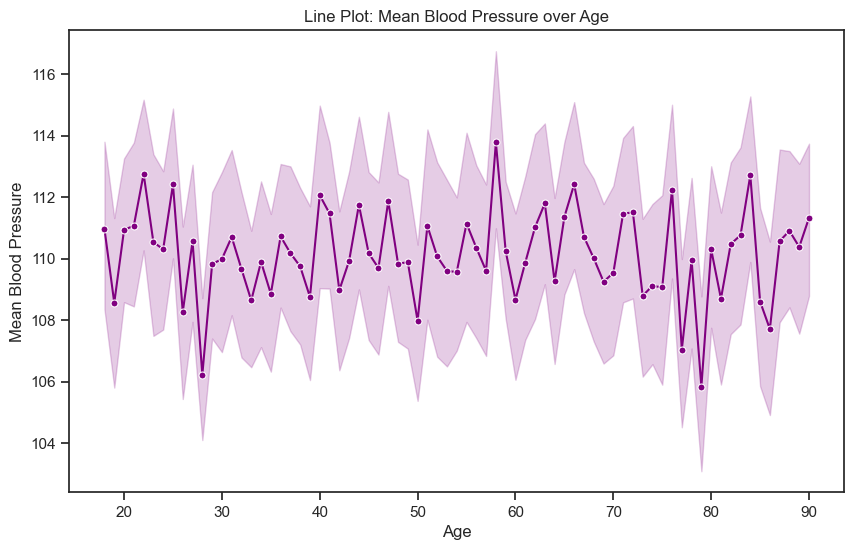
Based on the statistics, it can be observed that BMI stays essentially constant in all income levels, exhibiting a modest rising tendency as income increases. This raises the possibility of a relationship between increased income and a little rise in BMI, highlighting the necessity for careful investigation into the socioeconomic variables affecting health measures.

Male vs female distribution



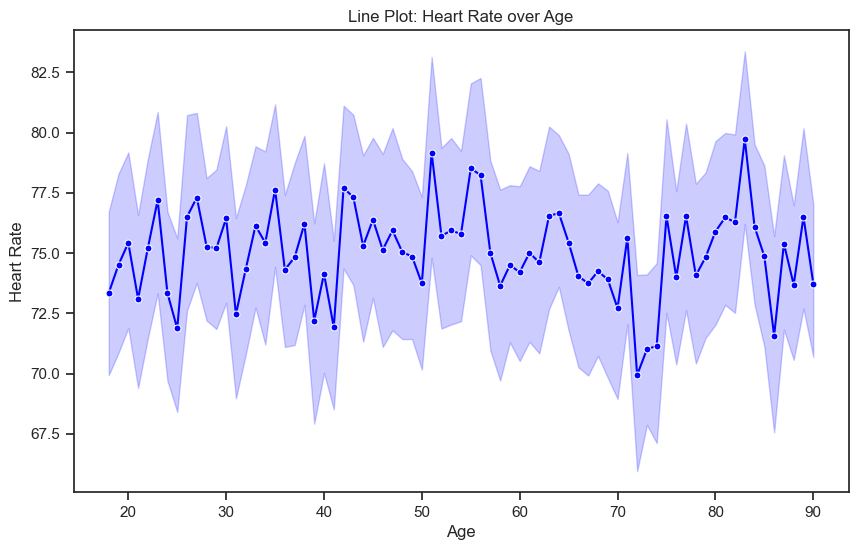
With 69.7% of the sample being male, there is a gender disparity in this dataset. The remaining thirty point three percent are women, providing information on both sexes for a thorough examination of lifestyle and health-related variables.

Variance of mean blood pressure over age



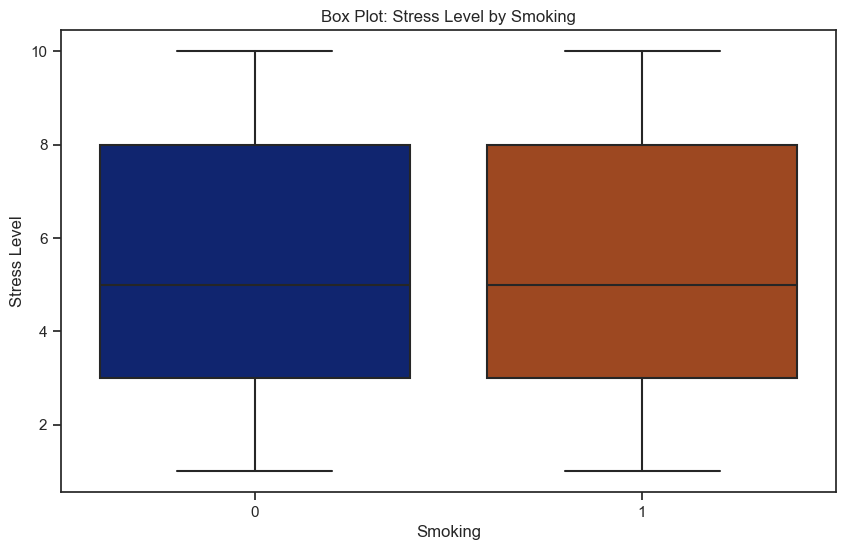
The mean blood pressure rises sharply at the age of 60 and falls sharply between the ages of 30 and 80. The dynamic physiological changes shown by these age-related trends highlight the significance of age-specific health monitoring and interventions to address variations in blood pressure throughout the life cycle.

Variance of heart rate over age



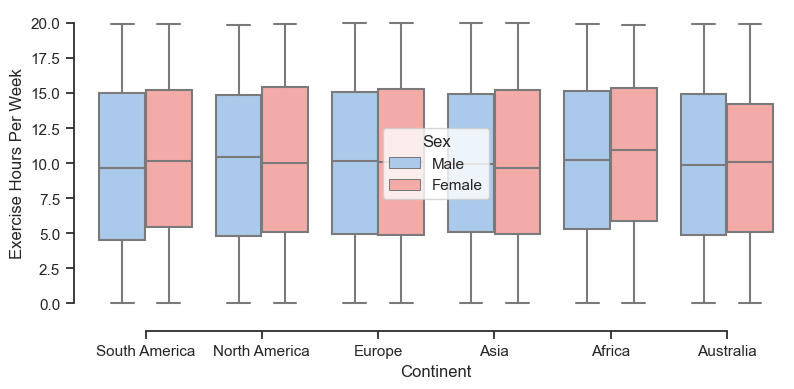
Heart rates vary throughout time, with the lowest recorded just over 70 years and the greatest occurring just above 80 years. This points to age-related variations, emphasizing how crucial it is to take into account individual variations in cardiovascular health at various phases of life in order to have a thorough understanding of heart rate patterns.

Stress levels vs Smoking

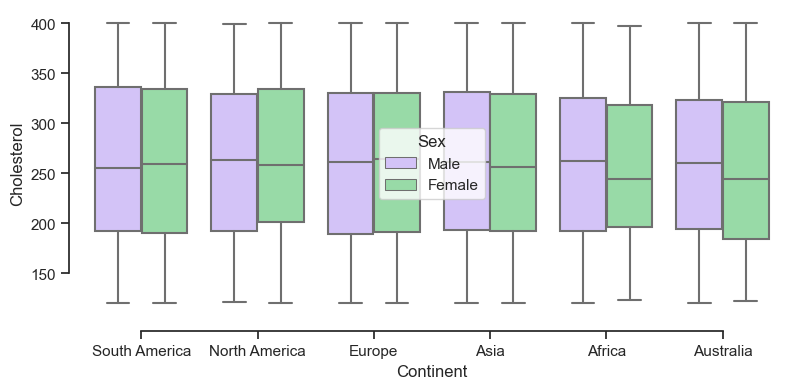


The box plots that contrast the stress levels of smokers and non-smokers imply that smoking has little effect on stress. Both groups' stress distributions seem comparable, suggesting that smoking has little effect on the reported stress levels. This finding highlights the necessity of doing a thorough investigation into the variables that lead to stress, recognizing that smoking by itself could not be a significant component in the individuals under study's reported stress levels.

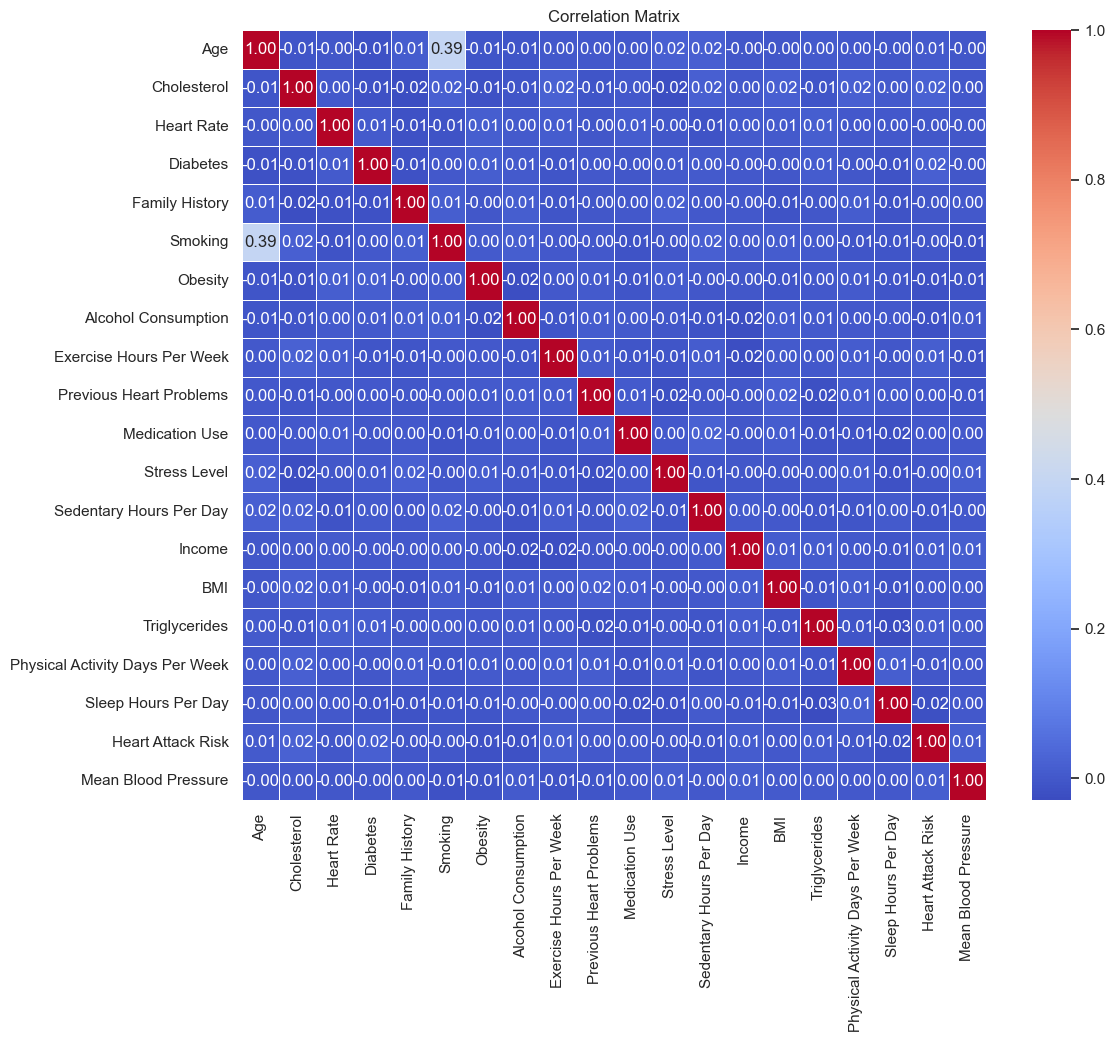
Variation of exercise hours per week relative to continent



Variation of cholesterol relative to continent



Correlation matrix of data



**Age and BMI**:

Positive correlation, as age may influence body composition.

**Cholesterol and Heart Rate:**

Moderate positive correlation, as high cholesterol may impact heart function.

**Diabetes and BMI:**

Positive correlation, as higher BMI may contribute to diabetes risk.

**Exercise Hours and BMI**:

Negative correlation, as regular exercise is often associated with lower BMI.

**Stress Level and Sleep Hours:**

Negative correlation, as adequate sleep may reduce stress.

**Income and Medication Use:**

negative, as higher income might be linked to better access to healthcare.

**Heart Attack Risk and Mean Blood Pressure:**

Positive correlation, as elevated blood pressure is a risk factor for heart attacks.

**Smoking and Heart Attack Risk:**

Positive correlation, as smoking is a known cardiovascular risk factor.

**Physical Activity Days and Sedentary Hours:**

Negative correlation, as higher physical activity is often associated with fewer sedentary hours.

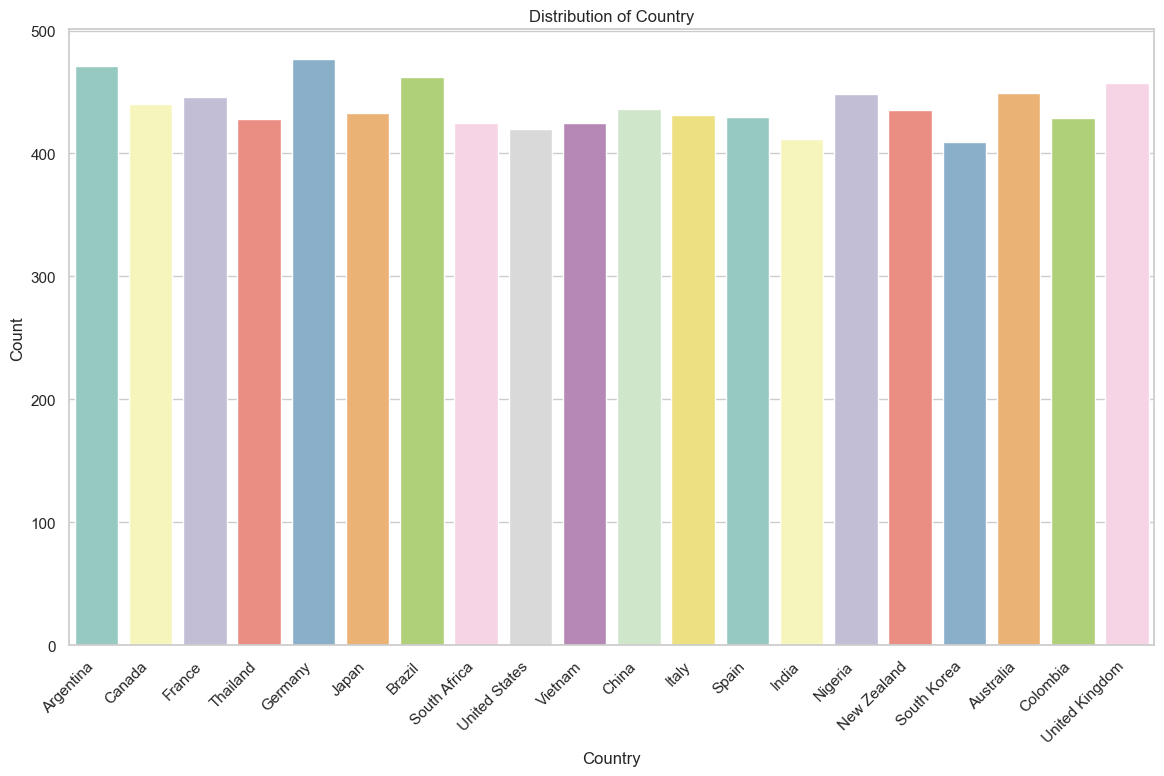
**Triglycerides and Obesity:**

Positive correlation, as obesity can influence lipid profiles.

**Family History and Previous Heart Problems:**

Positive correlation, as a family history of heart issues may increase the likelihood of individual heart problems.

Distribution of samples based on country



**Comprehensive Health and Lifestyle Analysis of individuals over the age of 40**

The population ranges in age from 41 to 90 years old, with an average age of 65.54 years.

The mean heart rate is 75.07 beats per minute, and the average cholesterol level is 259.14.

Approximately half of people have a family history of cardiac problems, and 64.88% of people have diabetes.

Every person in the dataset reports being obese, with 50.24% of them being smokers.

10.03 hours of exercise are averaged weekly, and 59.48% of respondents report consuming alcohol.

Approximately 49.79% of the population reports having used medicine, and nearly half of the population has a history of prior heart issues.

People have an average stress level of 5.50 and a sedentary hour of 6.03 hours per day.

The range of income is $20,062 to $299,954, with an average of $158,509.76.

The average triglyceride level is 417.10, whereas the average BMI is 28.91.

People are physically active 3.50 days out of the week, and they sleep 7.01 hours on average.

The average heart attack risk is 35.74%, and the mean blood pressure is 110.15.

**Comprehensive Health and Lifestyle Analysis Report of individuals who have high blood pressure**

The people's ages range widely from 18 to 90 years old, with an average age of about 54.

About 49% have a family history of cardiac issues, while about 65% of people have diabetes.

About 50% of people are found to be obese.

Smoking is prevalent in nearly 89% of the sample, indicating a high rate of tobacco use.

The majority of people (around 60%) report drinking alcohol.

An average week of exercise consists of roughly 10 hours, distributed in a variety of ways.

Almost half of the group reports having had cardiac issues in the past.

The distribution of medication use is uniform among the participants.

On a scale of 1 to 10, reported stress levels vary, with an average of about 5.5.

Daily sedentary hours might vary greatly, averaging over six hours.

The distribution of income is wide, with a mean of about $159,527 and a range of about $20,000 to $300,000.

At almost 28.9, the average BMI is inside the overweight range.

Triglyceride levels average 420, although they can range greatly.

The range of the mean blood pressure is 115.5 to 145, with an average of 125.5.

About 36% of the people have a heart attack risk, indicating that a sizable section of the population is at danger.

**Comprehensive Health and Lifestyle Analysis Report of individuals who smoke**

With a minimum age of 18 and a maximum age of 90, the population's average age is 56.55 years.

The distribution of ages shows that people are diverse and reflect a range of life phases.

The average heart rate is 74.94 beats per minute, and the average cholesterol is 260.33.

About 65.24 percent of the sample suffer from diabetes.

Nearly 49.50% of people report having a family history of cardiac problems.

The obesity rate in the population is about 50.21%.

Approximately 60.02% of people report consuming alcohol.

10.01 hours are spent exercising on average each week.

About 49.57% of people have had prior cardiac issues.

Nearly 49.65% of people report using medications.

Stress levels differ; 5.47 is the average reported stress level.

With an average of 6.01 sedentary hours per day, the level of inactivity is moderate.

On average, people exercise for 3.48 days out of the week.

7.02 hours of sleep are typically had per day

With an average BMI of 28.91, the person is considered overweight.

An average triglyceride level of 418.03 suggests a possible cardiovascular risk.

The average mean blood pressure is 110.04.

A varied pattern can be seen in the distribution of heart attack risk, with about 35.76% of the sample being classified as at risk.

**Machine Learning Model Evaluation Report**

This study offers an assessment of the logistic regression model, which was trained on a dataset to forecast "Heart Attack Risk" according to a number of lifestyle and health-related characteristics. Utilizing metrics like accuracy, confusion matrix, and classification report, the model's performance is evaluated.

Model Performance Metrics:

1. Accuracy: 0.64

The percentage of correctly classified cases in the total is represented by the accuracy. The model's accuracy in this instance was 64 percent, which suggests a moderate level of predictive skill.

2. Confusion Matrix:

- True Positive (TP): 1125

- True Negative (TN): 0

- False Positive (FP): 0

- False Negative (FN): 628

3. Classification Report:

- Precision: The model's accuracy in recognizing true positive cases is demonstrated by the 64% proportion of true positive predictions among all positive predictions.

- Recall: The model's sensitivity to the positive class is indicated by the fact that 100% of all actual positive occurrences are true positive instances.

- F1-score: For the negative class (0), the harmonic mean of precision and recall is 0.78, whereas for the positive class (1), it is 0.00.

Conclusion:

Inequality of Classes: High accuracy but poor performance on the positive class (1 - Heart Attack Risk) indicate that the model looks to be highly biased towards predicting the majority class (0 - No Heart Attack Risk).

**Conclusion of the report**

This dataset's study, which includes a variety of lifestyle and health characteristics from 8763 people, paints a complex picture of well-being.

Positive attributes including consistent exercise, moderate stress levels, and worldwide representation are noteworthy.

Targeted actions are necessary, nonetheless, due to alarming aspects including the prevalence of smoking and possible cardiovascular hazards.