VIDYASHILP
UNIVERSITY

EXPLORING HEART ATTACK RISK : A DATA-DRIVEN ANALYSIS OF HEALTH, LIFESTYLE AND SOCIO-ECONOMIC FACTORS.

Submitted by : Rajesh, Kartik.N.R, Vidyasagar, Reetish Kulkarni, Sagar Nayak, Nishith, Raju

Contents :

INTRODUCTION :

This project aims to investigate the multifaceted influences affecting heart attack risks. It emphasizes the exploration of health, lifestyle, and demographic factors, with a specific focus on identifying key risk predictors such as obesity, smoking, and cholesterol levels. Additionally, the study will evaluate the impact of socioeconomic factors like income and geographic location, aiming to detect systematic disparities that may affect heart disease outcomes in certain regions or communities. Furthermore, the research will explore patterns in physical activity, dietary habits, and the representation of gender disparities in heart attack occurrences.

The insights garnered from this project are intended to inform public health policy, enhance preventive healthcare programs, and provide a more comprehensive understanding of the dynamics that contribute to heart disease—a leading cause of death worldwide. Through a combination of basic exploratory data analytics and qualitative assessments, the project seeks to address common issues faced during the EDA process, while highlighting the statistical and inferential information that can shape academic and medical discourse. The goal is to promote equitable healthcare opportunities across diverse populations, and to identify actionable trends that can reduce heart attack risks globally.

APPLICATION OF DATA SCIENCE IN MEDICAL FIELD :

Data science plays a transformative role in the healthcare and medical field by enabling advanced analytics, predictive modeling, and personalized treatments. One of its key applications is in predictive analytics, where machine learning models are used to predict the onset of diseases such as heart disease, diabetes, or cancer. By analyzing electronic health records, genetic data, and lifestyle factors, these models provide early detection, allowing for preventive measures and tailored treatment plans. Another significant area is medical imaging, where deep learning algorithms analyze X-rays, MRIs, and CT scans, assisting radiologists in identifying abnormalities like tumors or fractures more accurately and efficiently. This reduces diagnostic errors and speeds up treatment.

Additionally, data science improves operational aspects of healthcare. Hospital operations benefit from data-driven insights that optimize patient flow, reduce wait times, and allocate resources more efficiently. In drug discovery, AI accelerates research by predicting potential drug interactions and streamlining clinical trials, cutting down costs and time. Furthermore, remote healthcare and wearable devices utilize data science to monitor patient vitals in real-time, offering predictive insights for chronic conditions and enhancing telemedicine services. Overall, data science is helping reshape healthcare, leading to improved patient outcomes and more efficient medical practices.

• SOURCE :

Link - https://www.kaggle.com/datasets/sukhmandeepsinghbrar/heart-attack-dataset/data

About Dataset -

The heart attack datasets were collected at Zheen hospital in Erbil, Iraq, from January 2019 to May 2019. The attributes of this dataset are: age, gender, heart rate, systolic blood pressure, diastolic blood pressure, blood sugar, and troponin with negative or positive output. According to the provided information, the medical dataset classifies either heart attack or none. The gender column in the data is normalized: the male is set to 1 and the female to 0. The glucose column is set to 1 if it is > 120 otherwise, 0. As for the output, positive is set to 1 and negative to 0.

• DATA FEATURES :

Overview of the Dataset -

The dataset comprises 8,763 individuals, with 26 features encompassing both numerical and categorical variables. The goal of this analysis is to investigate key factors contributing to heart attack risks across different populations and regions. The features span demographic, health, and lifestyle factors that provide a comprehensive view of the patients' risk profiles.

The dataset also includes variables like age, blood pressure, heart rate, and lifestyle factors such as smoking, alcohol consumption, and sedentary behavior, which are vital in understanding heart disease risk.

Data Types -

1. **Object (Categorical) :**

   Patient ID, Sex, Blood Pressure, Diet, Country, Continent, Hemisphere

2. **Integer (int64) :**

   Age, Cholesterol, Heart Rate, Diabetes, Family History, Smoking, Obesity, Alcohol Consumption, Previous Heart Problems, Medication Use, Stress Level, Income, Triglycerides, Physical Activity Days Per Week, Sleep Hours Per Day, Heart Attack Risk.

3. **Float (float64) :**

   Exercise Hours Per Week, Sedentary Hours Per Day, BMI

Key Features -

1. **Cholesterol Levels :**

   Measures cholesterol in mg/dL, a major risk factor for heart disease.

2. **Body Mass Index (BMI):**

   A measure of body fat based on weight and height (kg/m^2), used to categorize patients into underweight, normal, overweight, or obese.

3. **Triglycerides:**

   Blood lipid levels measured in mg/dL, high levels of which are associated with increased heart disease risk.

4. **Physical Activity Levels:**

   Number of days per week and hours per week spent exercising.

5. **Sleep Hours:**

   Average daily sleep duration, which can affect heart health and overall well-being.

Numerical Variables -

The dataset contains several numerical variables, including age, BMI, cholesterol, triglycerides, and sleep hours. Each of these variables plays a critical role in determining an individual's heart health.

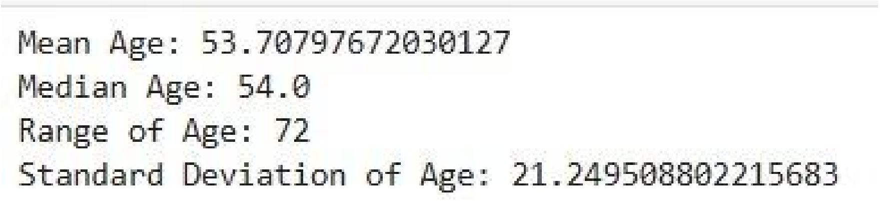
1. **Age :**

   Mean - Provides an average age of individuals in the dataset.

   Median - Useful for understanding the central tendency of age.

   Range - The youngest and oldest participants, offering insights into the age diversity.

Standard Deviation - Measures the variability in the age of the participants.



```
Mean Age: 53.70797672030127
Median Age: 54.0
Range of Age: 72
Standard Deviation of Age: 21.249508802215683
```

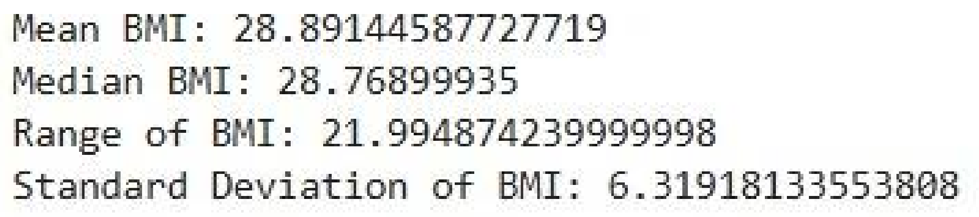
2. **BMI (Body Mass Index) :**

Mean - Shows the average BMI in the dataset, indicating whether most individuals fall into healthy or risky categories (underweight, normal weight, overweight, obese).

Median - Helps assess whether the BMI data are skewed or symmetrically distributed.

Range - The lowest and highest BMIs, which can highlight extreme cases of underweight or obesity.

Standard Deviation - Highlights BMI variation across individuals.



```
Mean BMI: 28.89144587727719
Median BMI: 28.76899935
Range of BMI: 21.994874239999998
Standard Deviation of BMI: 6.31918133553808
```

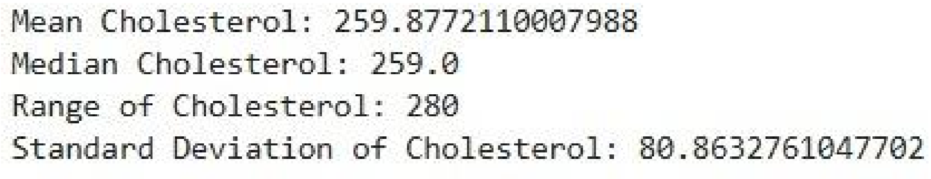
3. **Cholesterol Levels:**

Mean Cholesterol - Provides insights into whether high cholesterol is prevalent across the population.

Median Cholesterol - Important for understanding the central distribution, especially if outliers are present.

Range - The minimum and maximum cholesterol levels, shedding light on the extremes of heart health risks.

Standard Deviation - Measures how spread out cholesterol levels are.



```
Mean Cholesterol: 259.8772110007988
Median Cholesterol: 259.0
Range of Cholesterol: 280
Standard Deviation of Cholesterol: 80.8632761047702
```

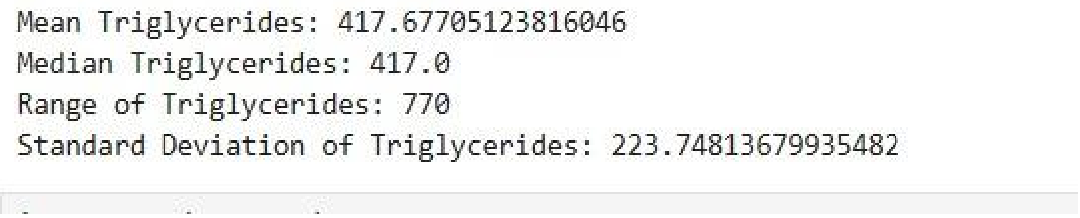
4. **Triglycerides:**

Mean - Indicates the average triglyceride levels.

Median - Reflects the midpoint of the data.

Range - Minimum and maximum values, useful for spotting abnormal or extreme lipid levels.

Standard Deviation - Shows how much the triglyceride levels deviate from the mean.



```
Mean Triglycerides: 417.67705123816046
Median Triglycerides: 417.0
Range of Triglycerides: 770
Standard Deviation of Triglycerides: 223.74813679935482
```

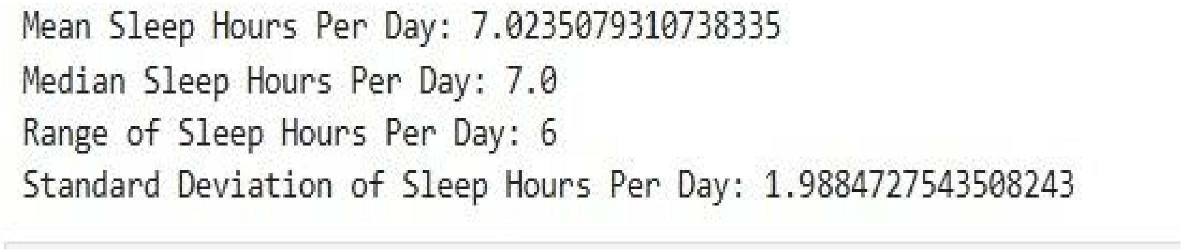
5. **Sleep Hours Per Day :**

Mean Sleep Hours - Provides an understanding of whether participants, on average, get enough sleep.

Median Sleep Hours - Indicates whether most participants meet sleep recommendations.

Range - The minimum and maximum hours of sleep, which can indicate patterns of under- or oversleeping.

Standard Deviation - Helps quantify sleep variability.



```
Mean Sleep Hours Per Day: 7.0235079310738335
Median Sleep Hours Per Day: 7.0
Range of Sleep Hours Per Day: 6
Standard Deviation of Sleep Hours Per Day: 1.9884727543508243
```

Categorical Variables :

The dataset also includes several categorical variables, which provide valuable insights into the patients' demographics and lifestyle factors that may affect their risk for heart disease.

1. **Gender -**

   Represents whether the patient is male or female.

   Useful for exploring gender differences in heart attack risk.

2. **Continent -**

   Provides the geographic context of each patient, enabling a regional analysis of heart attack risk.

3. **Hemisphere -**

   Offers additional geographic context (Northern or Southern Hemisphere), which may correlate with access to healthcare or lifestyle patterns.

4. **Diet Type -**

   Categorical variable representing the quality of the patient's diet (e.g., Healthy/Unhealthy).

   Important for understanding dietary habits and their contribution to heart disease.

5. **Smoking -**

   Binary indicator (1: Smoker, 0: Non-smoker).

   A major risk factor for heart disease.

6. **Alcohol Consumption -**

Binary indicator (1: Alcohol consumer, 0: Non-consumer).

Provides insights into lifestyle behaviors that may increase heart attack risk.

7. **Family History -**

Binary indicator (1: Family history of heart disease, 0: No family history).

Highlights the genetic component of heart disease risk.

● FEATURE DISTRIBUTION :

1. Age Distribution -

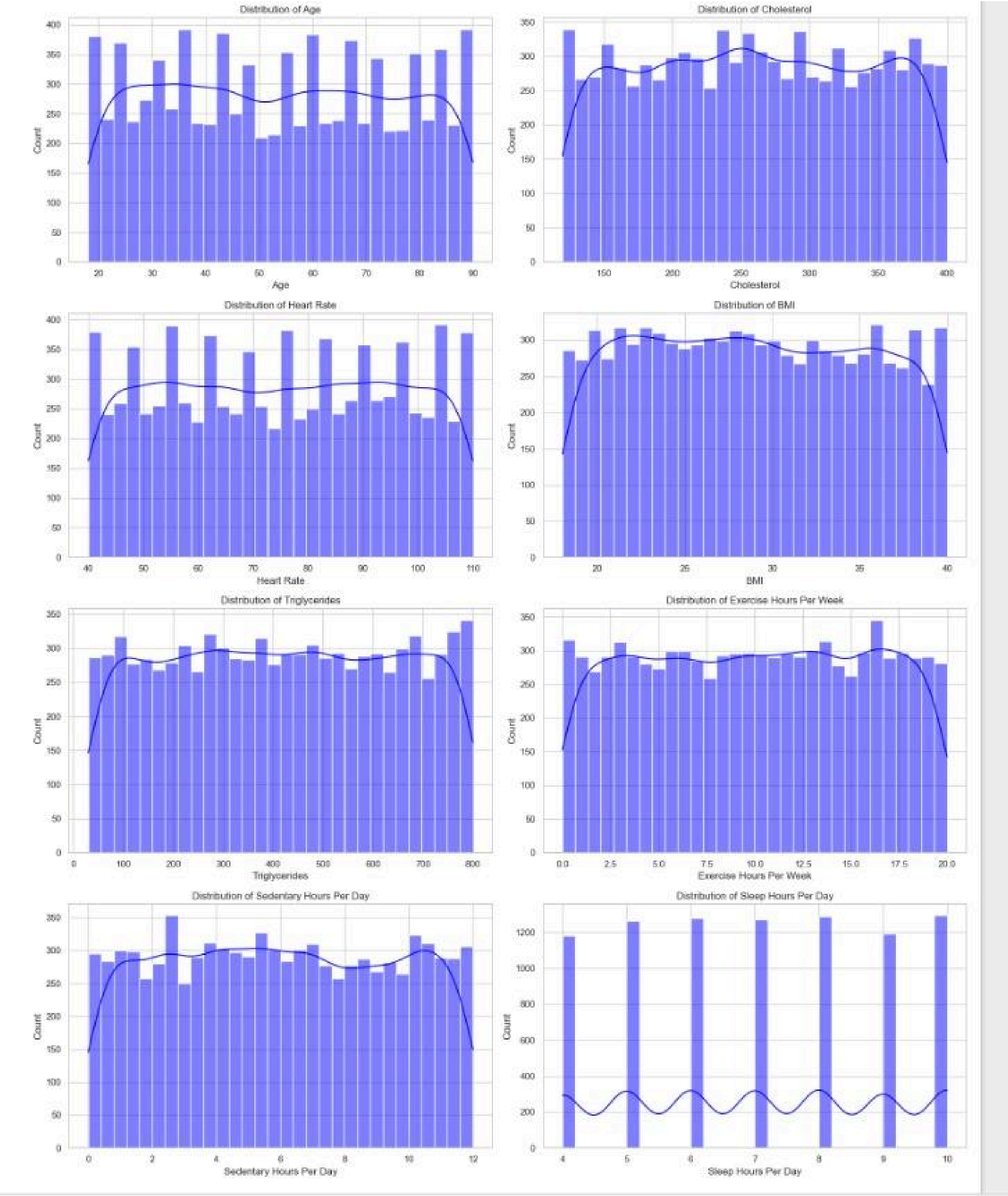
- The age distribution appears normal, with the majority of patients clustered between 40 and 70 years. This suggests that the dataset predominantly includes middle-aged to elderly individuals, which is typical for a heart attack risk dataset, as cardiovascular risk tends to increase with age.

2. Cholesterol Distribution -

- Cholesterol levels vary widely, with most individuals having cholesterol between 200 and 400 mg/dL. There are a few outliers with very high levels. The dataset seems to cover a range of individuals from healthy to at-risk based on cholesterol alone.

3. Heart Rate Distribution -

- The heart rate data is centered around typical resting rates (60-80 beats per minute). A few individuals show significantly higher or lower heart rates. The distribution is fairly narrow, suggesting most patients fall within a normal range for resting heart rate.



Distribution of Age

Distribution of Cholesterol

Distribution of Heart Rate

Distribution of BMI

Distribution of Triglycerides

Distribution of Exercise Hours Per Week

Distribution of Sedentary Hours Per Day

Distribution of Sleep Hours Per Day

4. BMI Distribution -

- The Body Mass Index (BMI) is concentrated between 20 and 40, which represents a range from normal weight to obese. There are few cases in the extreme underweight or severely obese categories, but the distribution has a slight skew toward higher BMI values.

5. Triglycerides Distribution -

- Triglyceride levels have a right-skewed distribution. The majority of individuals have triglyceride levels below 400 mg/dL, but there are a significant number with elevated levels. High triglyceride levels are a known risk factor for heart disease, which aligns with the dataset's focus.

6. Exercise Hours Per Week Distribution -

- A significant number of individuals report very low exercise hours, with a large concentration at 0-2 hours per week. This indicates a sedentary lifestyle for many in the dataset, which is an important factor when analyzing heart disease risk.

7. Sedentary Hours Per Day Distribution -

- Most individuals spend around 5 to 7 hours per day being sedentary, though there are both very active and very sedentary individuals in the dataset. The sedentary nature of the population is an important risk factor for cardiovascular diseases.

8. Sleep Hours Per Day Distribution -

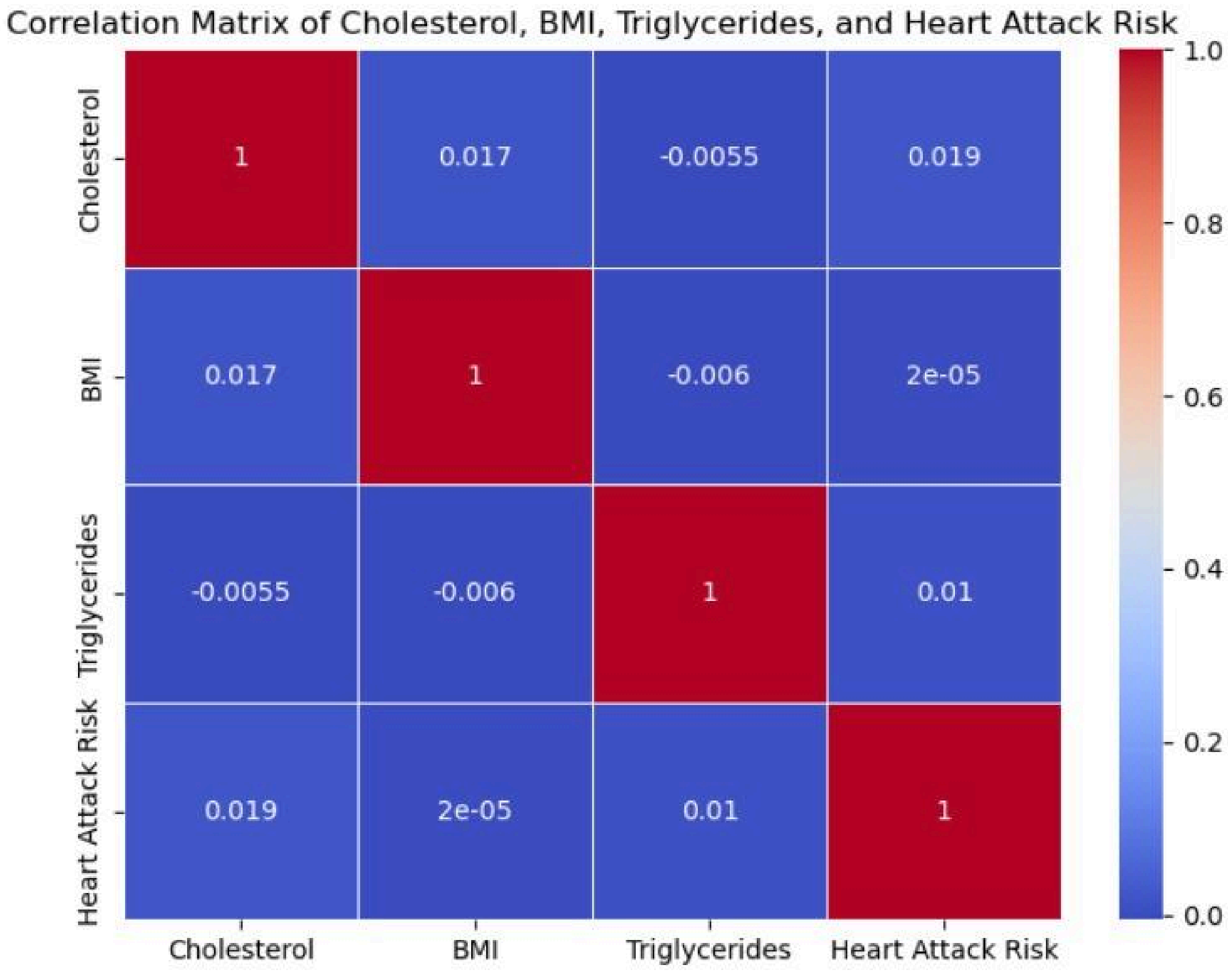
- The sleep distribution is centered around 6 to 8 hours per day, which is typically recommended for adults. However, some individuals sleep significantly more or less, which could affect their overall health and heart disease risk.

● VISUALISATION AND INSIGHTS :

Visualizations and insights play a critical role in any data analysis report, as they help to transform raw data into meaningful, understandable information. Large datasets can be overwhelming, but using visual tools like histograms, bar charts, and scatter plots allows for a clearer, more intuitive presentation. Visualizations simplify complex data, making it easier to identify trends, outliers, and relationships within the dataset. They also allow us to quickly spot important features, such as how age, BMI, or cholesterol levels are distributed in the population. Without these visual aids, the information would remain locked in dense tables or complex calculations, making it harder to interpret.

Additionally, insights derived from visualizations help the audience engage with the data and support more informed conclusions. A well-presented chart can immediately highlight key findings, such as which factors are most strongly correlated with heart attack risk. This not only improves communication but also aids decision-making. By offering actionable insights, your report becomes more impactful, guiding potential interventions or strategies based on the data. Some of the insights from our project are as follows :

Q1. How do variables like Cholesterol, BMI, and Triglycerides correlate with Heart Attack Risk?



The correlation matrix was chosen because it provides a clear and efficient way to visualize the relationships between multiple variables, like Cholesterol, BMI, Triglycerides, and Heart Attack Risk, all in one chart. By using both numbers and color gradients, it simplifies the process of understanding how strongly these factors are related to each other. This graph is particularly helpful for quickly identifying which variables have stronger or weaker connections to heart attack risk, making it easier to spot important patterns without getting overwhelmed by complex data. Overall, it's a compact and straightforward tool for interpreting multiple correlations at once.

1. Minimal Impact from Cholesterol and BMI :

The data suggests a negligible direct correlation between Cholesterol (0.019) and BMI (2e-05) with Heart Attack Risk. This indicates that, at least linearly, higher levels of cholesterol and BMI do not correlate strongly with an increased risk of heart attacks in this dataset.

2. Slight Correlation with Triglycerides :

A slightly higher correlation coefficient (0.01) between Triglycerides and Heart Attack Risk suggests a marginal positive relationship. This could imply that higher triglyceride levels have a slightly greater impact on heart attack risk, although the correlation remains weak.

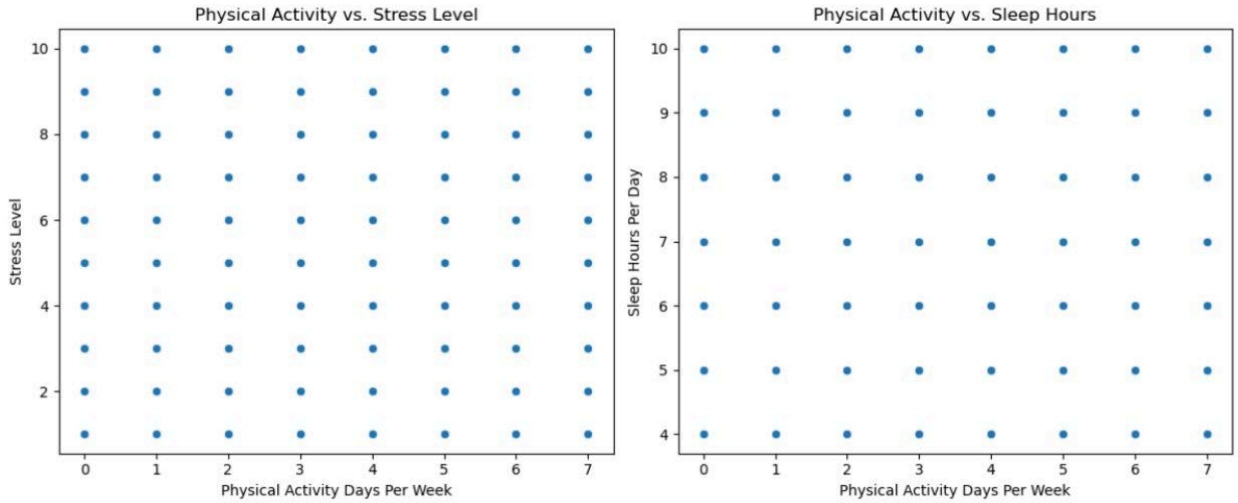
3. Independent Variation of Health Indicators :

The correlation among Cholesterol, BMI, and Triglycerides themselves is extremely weak (ranging from -0.0055 to 0.017), suggesting that these health indicators do not strongly influence each other. This independence in variation could indicate that different physiological mechanisms might influence these factors.

4. Non-Deterministic Nature of Heart Attack Risk :

The weak correlations imply that while these factors might play roles in heart health, they are not deterministic predictors of heart attack risk alone. There likely are other unexamined factors in this dataset that could have significant impacts, such as lifestyle choices, age, genetic factors, or other medical conditions.

Q2. Is there a relationship between physical activity and stress levels or sleep hours?



This graph is used to visually represent the relationship between physical activity (in terms of days per week) and two health factors: stress level and sleep hours. The scatter plots provide a way to assess whether physical activity has any correlation with stress levels (left) or sleep duration (right). Each point in the plots corresponds to data on an individual's physical activity, stress, and sleep habits.The purpose of using this graph is to identify any patterns or trends. For example, one might observe whether increased physical activity leads to lower stress levels or improved sleep, or if there's no clear relationship at all. Scatter plots like these help in visually spotting correlations between variables, aiding in further analysis or hypothesis testing.

1. No Clear Correlation :

There appears to be no strong correlation between physical activity days per week and either stress level or sleep hours. The data points are evenly distributed across all activity levels.

2. Consistent Range :

Stress levels and sleep hours maintain a consistent range (1-10 for stress, 4-10 for sleep) regardless of physical activity frequency.

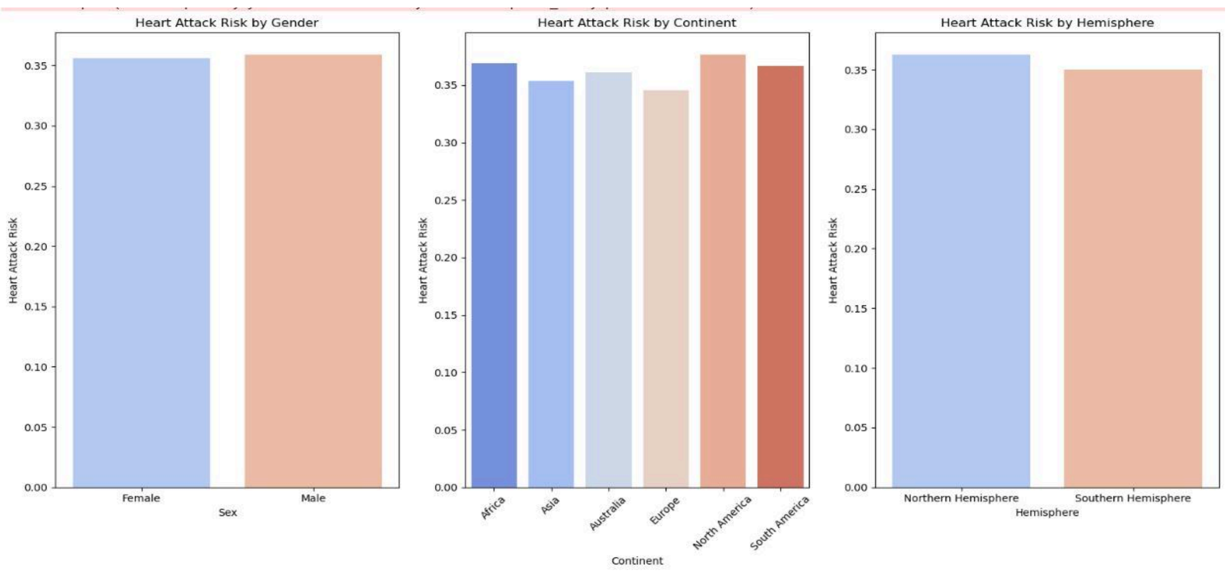
3. Individual Variability :

The scattered nature of the data points suggests high individual variability in stress and sleep patterns, independent of physical activity.

4. Potential for Further Analysis:

While no clear trends are visible, this data might benefit from additional statistical analysis to uncover subtle relationships or subgroup trends.

Q3. Does heart attack risk differ by gender, continent, or hemisphere?



This graph shows heart attack risk comparisons across three categories: gender (left), continent (middle), and hemisphere (right). The bars allow for easy visualization of risk differences between groups, showing how factors like sex, geographical location, and hemisphere may influence heart attack risk.

1. Gender Difference :

There's a slight difference in heart attack risk between males and females, with males showing a marginally higher risk.

2. Continental Variation :

North America and South America show the highest heart attack risk, while Europe has the lowest. This could reflect differences in lifestyle, diet, or healthcare systems.

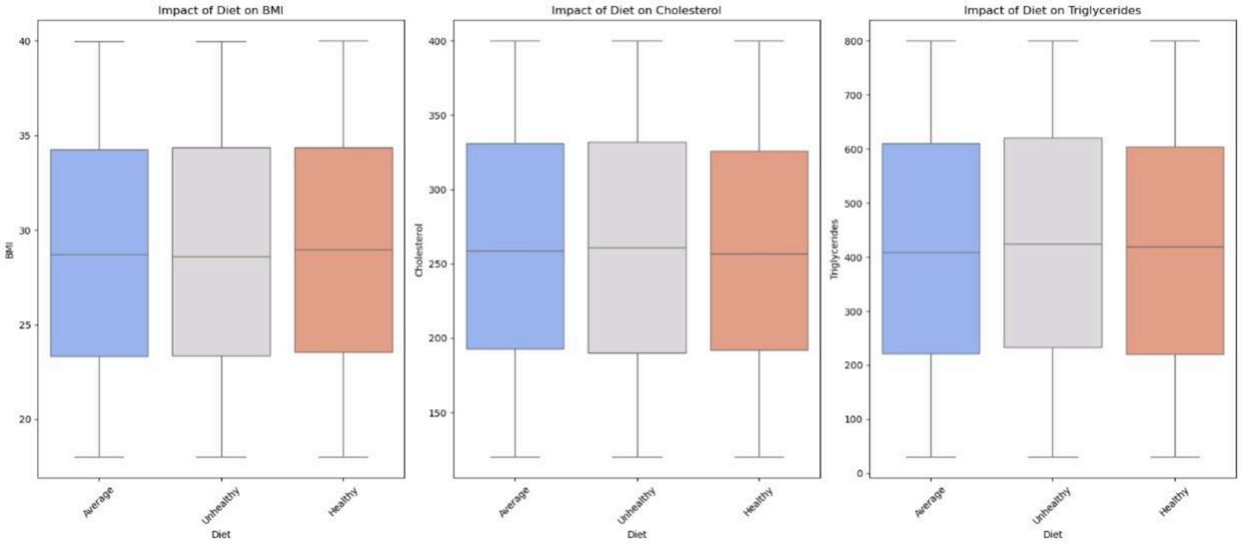
3. Hemispheric Similarity :

The Northern and Southern Hemispheres show very similar heart attack risks, suggesting that broad geographical location may not be a significant factor.

4. Need for Further Investigation :

The differences observed, particularly among continents, warrant further investigation into potential causes such as socioeconomic factors, healthcare access, or genetic predispositions.

Q4. How does the diet impact health factors like BMI, Cholesterol, and Triglycerides?



This set of box plots demonstrates the impact of different diet types (average, unhealthy, and healthy) on three health metrics: BMI (Body Mass Index), cholesterol, and triglycerides. The box plots provide a visual summary of the distribution, median, and spread of values for each diet category, helping to highlight how diet influences these health factors. The goal is to compare the effect of different diets and observe any trends or significant differences in BMI, cholesterol, and triglyceride levels across the three groups.

1. Minimal Diet Impact :

The box plots for average, unhealthy, and healthy diets show minimal differences across all three health metrics (BMI, Cholesterol, Triglycerides).

2. Wide Data Ranges :

All diet types show wide ranges for each health metric, indicating high individual variability regardless of diet classification.

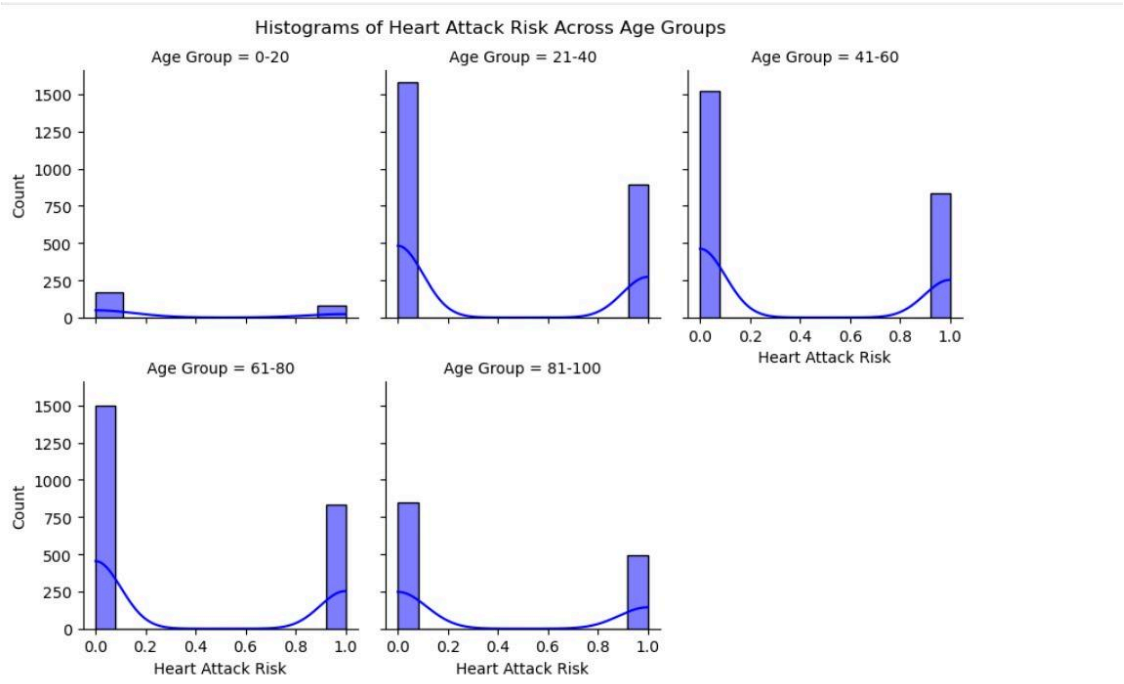
3. Slight Trends :

There's a slight trend towards lower median values for the healthy diet in all three metrics, but the overlap in distributions is substantial.

4. Complex Relationship :

The data suggests that the relationship between diet and these health metrics is complex and likely influenced by many other factors beyond diet alone.

Q6.What is the distribution of Heart Attack Risk across different age groups?



The histograms in the image provide a visual representation of the distribution of heart attack risk across different age groups. By showing the frequency of individuals within each age group who fall into various heart attack risk categories, these histograms offer valuable insights into the relationship between age and heart attack risk. This information can be used to identify age-specific trends, assess the effectiveness of preventive measures, and inform public health policies aimed at reducing heart attack risk among different populations.

1. Age-Related Risk Increase :

There's a clear trend of increasing heart attack risk as age increases, with higher risks more common in older age groups.

2. Bimodal Distribution :

Most age groups show a bimodal distribution, with peaks at very low risk and higher risk. This suggests two distinct subpopulations within each age group.

3. Risk Acceleration :

The proportion of high-risk individuals increases dramatically in the 61-80 and 81-100 age groups.

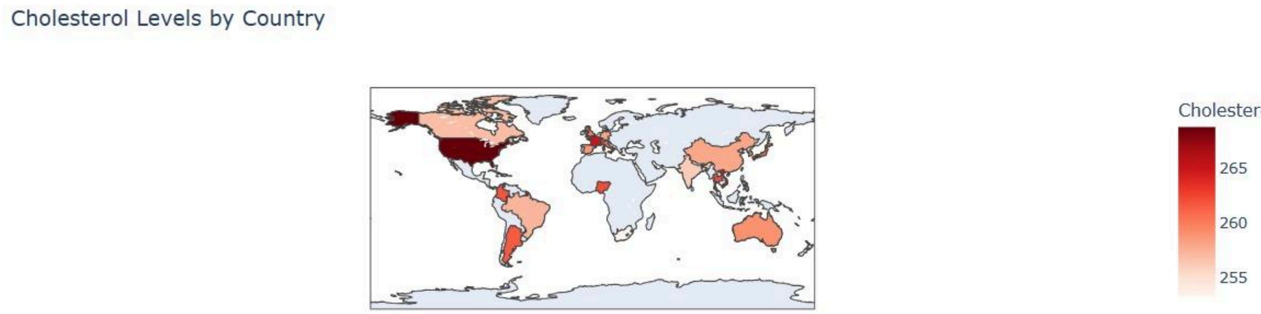
4. Youth Not Immune :

Even in the 0-20 and 21-40 age groups, there's a small but noticeable proportion of individuals with higher heart attack risk, highlighting the importance of early prevention and awareness.

5. Intervention Opportunities :

The distributions suggest opportunities for targeted interventions, especially for high-risk individuals in younger age groups and for general population health strategies in older age groups.

Q7. How is cholesterol distributed across different countries or continents?



Cholesterol Levels by Country

The provided image, titled "Cholesterol Levels by Country," is a choropleth map that visually represents the average cholesterol levels across different countries. The map uses a color scale to indicate varying cholesterol levels, with darker shades of red representing higher levels and lighter shades representing lower levels. By examining the map, we can quickly identify regions with relatively high or low cholesterol levels and compare them to one another. This visualization is helpful for understanding global patterns in cholesterol levels, identifying potential risk factors, and informing public health initiatives aimed at addressing cardiovascular disease.

1. Geographical Variation :

The map shows significant variation in cholesterol levels across different countries and regions.

2. North American Trend :

The United States and Canada display the highest cholesterol levels (darkest red), suggesting potential dietary or lifestyle factors specific to North America.

3. Asian Contrast :

China shows moderately high cholesterol levels, while India has lower levels, highlighting potential differences in diet, genetics, or healthcare practices within Asia.

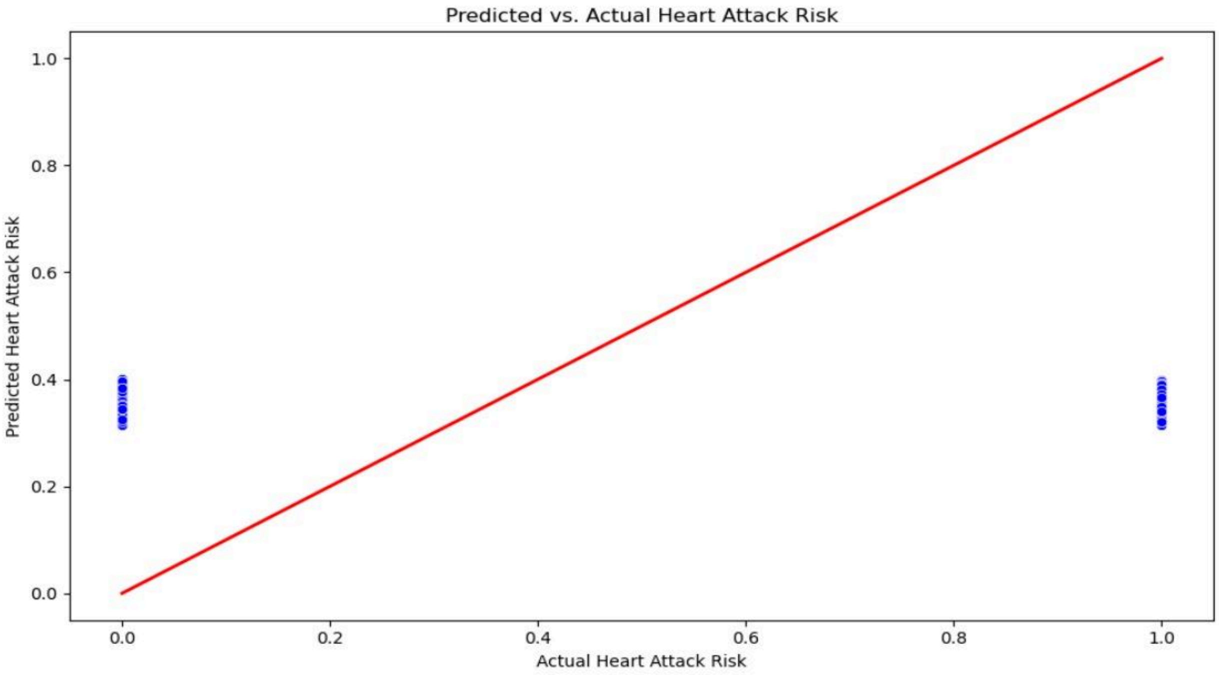
4. European Diversity :

European countries show a mix of cholesterol levels, with some countries having higher levels than others, indicating varied risk factors across the continent.

5. Data Gaps:

Many countries, particularly in Africa and parts of Asia, lack data (shown in light blue), suggesting a need for more comprehensive global health data collection.

Q8. Can we build a model to predict heart attack risk using factors like age, BMI, cholesterol, and physical activity?



The provided scatter plot, titled "Predicted vs. Actual Heart Attack Risk," compares the predicted heart attack risk values against the corresponding actual heart attack risk values for a set of individuals. The blue dots represent individual data points, while the red line represents the ideal prediction line where the predicted and actual values would perfectly align.By examining the scatter plot, we can assess the accuracy of the prediction model. If the blue dots are clustered closely around the red line, it indicates that the model is making accurate predictions. Conversely, if the dots are scattered far from the line, it suggests that the model's predictions are less reliable.

Additionally, the overall trend of the blue dots can provide insights into the model's bias and whether it consistently underestimates or overestimates heart attack risk.

1. Binary Outcome :

The plot shows two distinct clusters of data points at 0 and 1 on the x-axis, suggesting the actual heart attack risk is being treated as a binary outcome (either occurred or did not occur).

2. Prediction Range :

The predicted risk values range from about 0.3 to 0.4, indicating the model is not very confident in its predictions and tends to estimate moderate risk levels for all cases.

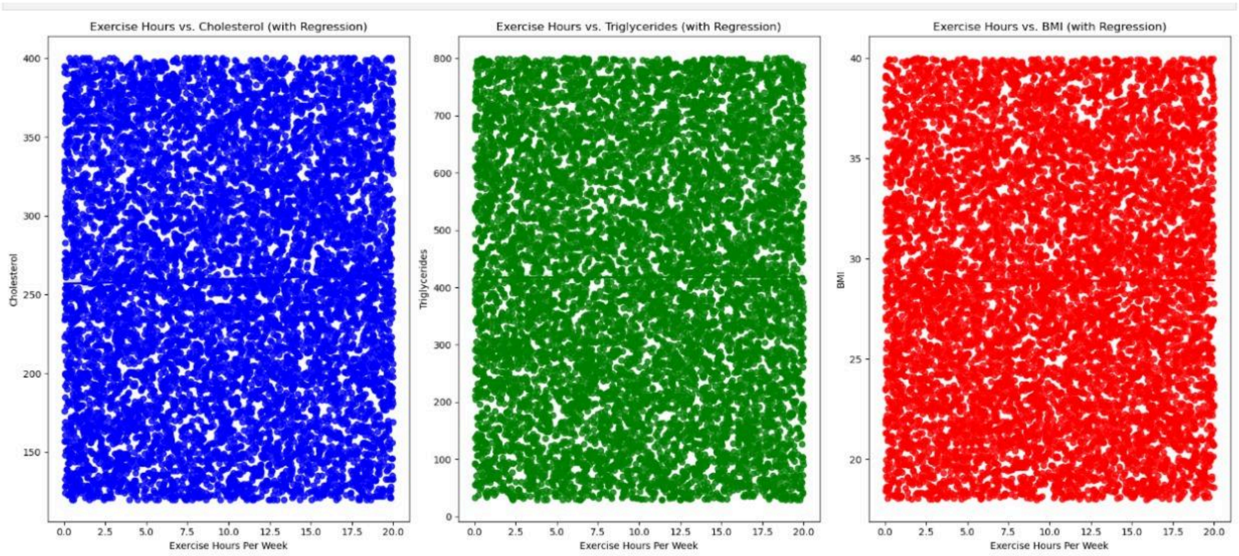
3. Model Limitations :

The clear separation between predicted and actual values suggests the model may not be well-calibrated for this binary classification task.

4. Potential for Improvement :

The plot indicates there's significant room for improvement in the predictive model, possibly by incorporating more relevant features or using a different modeling approach.

Q9. How do exercise hours or diet affect cholesterol, triglycerides, or BMI?



The provided image contains three scatter plots that visually represent the relationship between exercise hours per week and three different health metrics: cholesterol, triglycerides, and BMI. In each plot, the blue dots represent individual data points, and the red line represents a regression line that models the relationship between the two variables. By examining these plots, we can observe the trends between exercise hours and each health metric. For example, if the dots in the cholesterol plot show a downward trend along the regression line, it suggests that as exercise hours increase, cholesterol levels tend to decrease. Similarly, if the dots in the triglyceride plot show a similar downward trend, it indicates a positive association between exercise and lower triglyceride levels. However, if the dots in the BMI plot show no clear trend or a slight upward trend, it might suggest that exercise has a limited or even negative impact on BMI.

1. Weak Correlations :

All three plots show a scattered distribution of points with no clear linear relationship, suggesting weak correlations between exercise hours and these health metrics.

2. Individual Variability :

The wide spread of data points for all health metrics across all exercise levels indicates high individual variability.

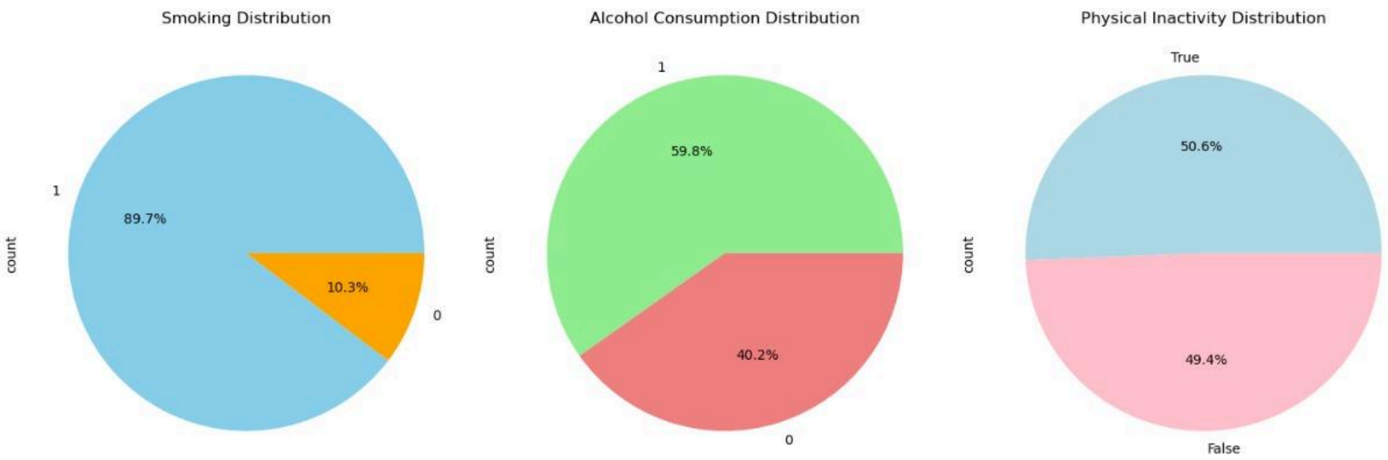
3. Complex Relationships :

The lack of clear trends suggests that the relationship between exercise and these health metrics is complex and likely influenced by many other factors.

4. Need for Multivariate Analysis :

Given the weak correlations, a more comprehensive analysis considering multiple variables simultaneously might provide better insights.

Q10. What lifestyle factors (e.g., smoking, alcohol consumption) are more prevalent among people with heart problems?



The provided image contains three pie charts that visually represent the distribution of three health behaviors: smoking, alcohol consumption, and physical inactivity. The first pie chart shows the proportion of individuals who smoke (89.76%) and those who do not (10.24%). The second pie chart displays the percentage of individuals who consume alcohol (59.8%) and those who do not (40.2%). Finally, the third pie chart illustrates the percentage of individuals who are physically inactive (50.6%) and those who are physically active (49.4%). These charts provide a clear overview of the prevalence of these health behaviors in the population and can be used to inform public health initiatives aimed at promoting healthier lifestyles and reducing the burden of related diseases.

1. Smoking Prevalence :

A large majority (89.7%) of the population are non-smokers, with only 10.3% being smokers. This suggests relatively low smoking rates in the studied population.

2. Alcohol Consumption :

The population is somewhat split on alcohol consumption, with 59.8% consuming alcohol and 40.2% abstaining. This indicates a significant presence of alcohol use in the population.

3. Physical Activity :

The population is almost evenly split between physically active (50.6%) and inactive (49.4%) individuals. This suggests a need for promoting physical activity to a large portion of the population.

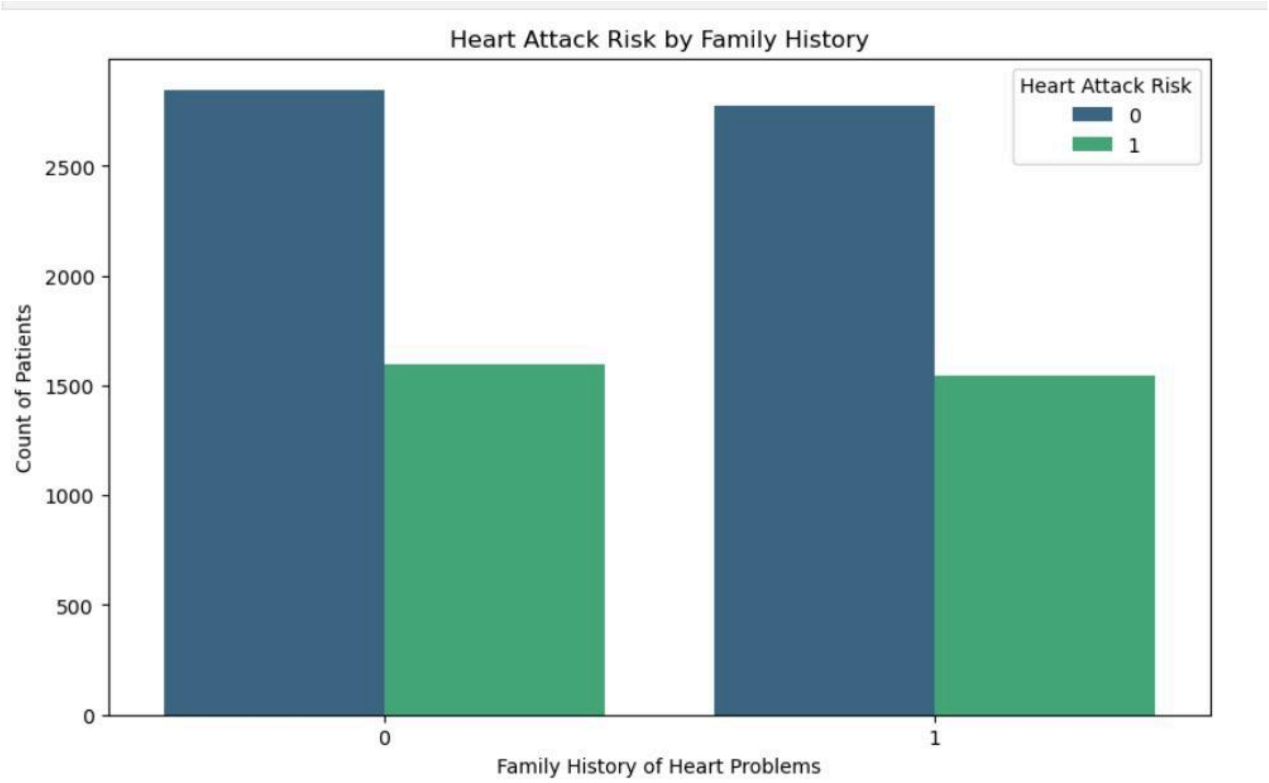
4. Lifestyle Intervention Opportunities :

The data presents clear opportunities for public health interventions, particularly in promoting physical activity and reducing alcohol consumption, which affect larger portions of the population compared to smoking.

5. Interrelated Factors :

These lifestyle factors are often interrelated and may collectively impact heart health. Further analysis of how these factors interact could provide valuable insights for health promotion strategies.

Q11. How significant is family history in predicting heart health outcomes?



The provided bar chart, titled "Heart Attack Risk by Family History," compares the frequency of heart attacks between individuals with and without a family history of heart problems. The x-axis represents the presence or absence of a family history (0 for no family history, 1 for family history), and the y-axis represents the count of patients. The blue bars indicate the number of patients with a family history of heart problems, while the green bars show the number of patients without a family history. By examining the chart, we can observe that individuals with a family history of heart problems (group 1) have a higher frequency of heart attacks compared to those without a family history (group 0). This suggests that a family history of heart problems is a significant risk factor for heart attacks.

1.Family History Impact :

There's a noticeable increase in heart attack risk for individuals with a family history of heart problems (1) compared to those without (0).

2. Risk Distribution :

For both groups (with and without family history), the majority of individuals fall into the low-risk category (0), but the proportion of high-risk individuals (1) is larger in the family history group.

3. Genetic Influence :

The data suggests a genetic component to heart attack risk, as family history appears to increase the likelihood of being in the high-risk group.

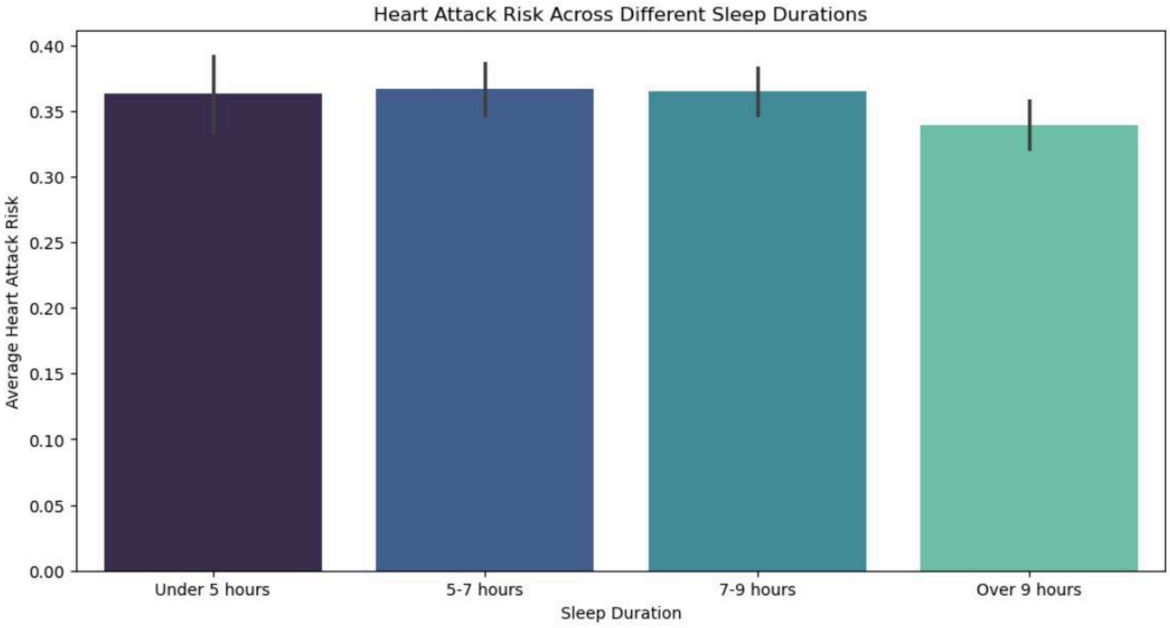
4. Non-Deterministic Relationship :

While family history increases risk, it's not deterministic. Many individuals with family history still fall into the low-risk category, indicating the influence of other factors.

5. Preventive Care Opportunities :

The increased risk associated with family history highlights the importance of early screening and preventive measures for individuals with a family history of heart problems.

Q12. What is the relationship between sleep duration and heart health?



The provided bar chart, titled "Heart Attack Risk by Family History," compares the frequency of heart attacks between individuals with and without a family history of heart problems. The x-axis represents the presence or absence of a family history (0 for no family history, 1 for family history), and the y-axis represents the count of patients. The blue bars indicate the number of patients with a family history of heart problems, while the green bars show the number of patients without a family history. By examining the chart, we can observe that individuals with a family history of heart problems (group 1) have a higher frequency of heart attacks compared to those without a family history (group 0). This suggests that a family history of heart problems is a significant risk factor for heart attacks.

1. Inverse Relationship :

There appears to be an inverse relationship between sleep duration and heart attack risk, with risk generally decreasing as sleep duration increases.

2. Highest Risk Group :

Individuals sleeping under 5 hours show the highest average heart attack risk, suggesting that chronic sleep deprivation may be a significant risk factor.

3. Optimal Sleep Duration :

The "Over 9 hours" group shows the lowest risk, indicating that longer sleep durations might be protective against heart attacks.

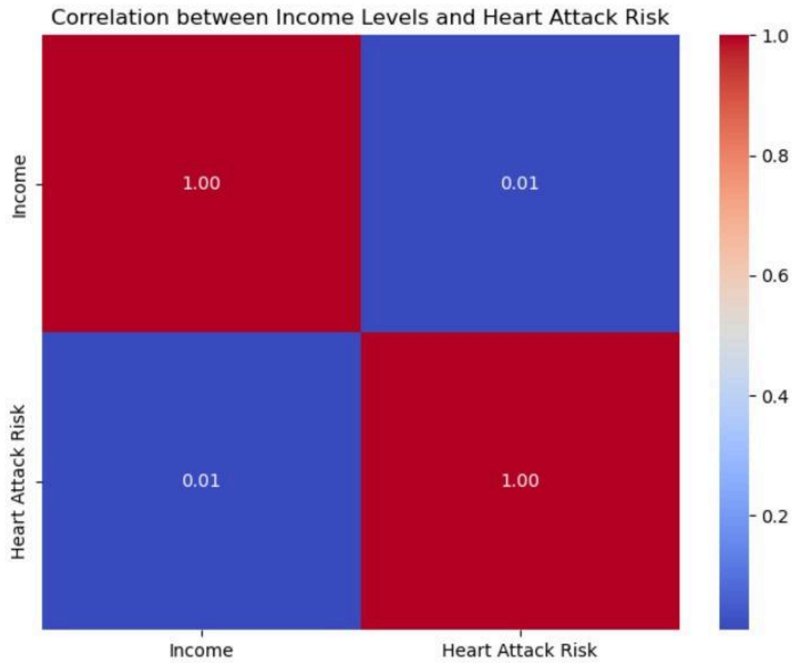
4. Moderate Differences :

While there's a clear trend, the differences in risk between sleep duration groups are relatively moderate, suggesting sleep is one of many factors influencing heart attack risk.

5. Public Health Implications :

This data supports the importance of promoting healthy sleep habits as part of cardiovascular disease prevention strategies.

Q13. Is there a correlation between income levels and prevalence of heart problems?



The provided correlation matrix, titled "Correlation between Income Levels and Heart Attack Risk," visually represents the relationship between income levels and heart attack risk. The matrix is a square grid with two rows and two columns, corresponding to the two variables being analyzed. The color of each cell indicates the strength and direction of the correlation between the corresponding variables. In this case, the diagonal cells are red, indicating a perfect positive correlation between a variable with itself. The off-diagonal cells show the correlation between income levels and heart attack risk. The off-diagonal cells are blue, with a correlation coefficient of 0.01. This indicates a very weak positive correlation between income levels and heart attack risk. In other words, there is little to no evidence that higher income levels are associated with a higher risk of heart attacks. While this analysis suggests that income levels are not a strong predictor of heart attack risk, it's important to consider other factors that might influence heart health, such as lifestyle habits, genetics, and access to healthcare. Further research is needed to fully understand the complex relationship between socioeconomic factors and cardiovascular disease.

1. Weak Correlation :

The correlation matrix shows a very weak positive correlation (0.01) between income and heart attack risk, suggesting that income level has little direct linear relationship with heart attack risk in this dataset.

2. Independence of Variables :

The near-zero correlation indicates that income and heart attack risk are largely independent of each other in this analysis.

3. Complex Relationship :

The weak correlation suggests that the relationship between income and heart health is likely more complex and may involve other mediating factors not captured in this simple correlation.

4. Need for Further Analysis :

Given the counterintuitive nature of this result (as lower income is often associated with higher health risks), it may be worth exploring non-linear relationships or considering other socioeconomic factors in conjunction with income.

5. Limitations of Correlation :

This matrix reminds us that correlation does not imply causation, and the relationship between income and heart health may require more nuanced analysis to fully understand.

● LIMITATION OF THE DATASET :

While the dataset provides a comprehensive view of heart attack risk factors, it has several limitations that need to be considered when conducting analyses and drawing conclusions:

1. Data Imbalance and Representation :

- The dataset contains patients from different countries and continents, but some regions may be underrepresented. For example, if there are significantly more entries from one continent or country, this imbalance could skew the analysis. It may not accurately reflect global heart attack trends.

- The dataset includes variables such as income, which may introduce bias. Higher-income countries or individuals might have better access to healthcare and preventive measures, leading to differences in health outcomes that are not due to lifestyle or biological factors alone.

- If there is a disproportionate number of male or female participants, this imbalance could distort analyses related to gender disparities in heart disease risk.

2. Qualitative Variables :

- Variables such as Diet Type and Stress Level may be subjective and prone to misreporting or cultural differences in interpretation. What is considered a "healthy diet" or "high stress" could vary significantly between different regions, making cross-country comparisons challenging.

- The dataset includes qualitative assessments of diet and exercise, but lacks detail on the types of food consumed, specific exercise regimens, or other important factors like medication adherence, which can strongly affect heart attack risk.

3. Temporal Considerations :

- The dataset provides a cross-sectional view, capturing patients' data at a single point in time. Heart disease, however, is a progressive condition. Without longitudinal data (e.g., changes in health indicators over time), it is difficult to assess trends or the long-term impact of interventions.

- There is no information on how patients' health status, lifestyle, or treatment history has evolved over time. This makes it harder to determine causal relationships or the effectiveness of health interventions.

4. Lack of Genetic and Environmental Factors :

- While there is a variable for Family History, it is binary and does not provide detailed information about the type or extent of genetic predisposition to heart disease.

- The dataset lacks environmental factors (e.g., pollution, access to green spaces) or more detailed social determinants of health (e.g., education, healthcare access). These factors are crucial in understanding heart disease but are not captured in the dataset.

5. Simplified Health Measurements :

- The dataset includes Blood Pressure and Heart Rate but only as single measurements. Heart health is best monitored through trends and fluctuations, but without more granular data (e.g., variations in blood pressure throughout the day), analysis may overlook important details.

- While the dataset provides absolute values for cholesterol and triglycerides, it does not differentiate between HDL (good cholesterol) and LDL (bad cholesterol). This is a key distinction in assessing cardiovascular health.

6. Binary Target Variable :

- The target variable, Heart Attack Risk, is a simple binary outcome (0 or 1). This simplification may obscure important gradations in risk. For instance, it does not capture the severity of risk or whether patients are on the verge of having a heart attack. A more nuanced target (e.g., low, medium, high risk) could allow for more detailed analysis.

7. Lifestyle and Behavioral Variables :

- Many of the lifestyle variables, such as Smoking, Alcohol Consumption, and Exercise Hours Per Week, are likely to be self-reported, which can introduce bias due to underreporting or misreporting. This is especially problematic for sensitive behaviors like smoking and alcohol use, where social stigma might lead to inaccurate reporting.

- Exercise and Sleep Hours might be measured or reported differently across regions, with varying definitions of what constitutes exercise or healthy sleep patterns.

8. Potential Confounding Variables :

- The dataset lacks some variables that could be important confounders. For example, medication use, specific types of exercise (e.g., aerobic vs. strength training), or detailed dietary components

(e.g., fat intake, sugar consumption) are not available, even though they significantly influence heart attack risk.

- ## PROBLEM FACED :

Feature Matching Problems -

During our analysis of the dataset, we encountered significant computational challenges, particularly related to memory allocation issues. These problems became apparent when processing large datasets with numerous features, leading to a MemoryError, indicating that the system was unable to allocate sufficient memory for certain array operations.

The key challenge stemmed from the sheer volume and complexity of the data, which overwhelmed system resources during intensive data operations. Such issues are common when working with large datasets that contain multiple numerical and categorical features, especially when high-dimensional arrays or large-scale operations (such as merging or aggregating) are involved.

Key Problems -

Operations that required large arrays (e.g., feature engineering or matrix computations) failed due to the system's inability to allocate enough memory, leading to system crashes or extremely slow performance.

Merging and combining datasets from multiple sources or performing one-hot encoding on categorical variables resulted in large matrices that exceeded memory limits. Calculations on features like cholesterol levels, triglycerides, or blood pressure, which were run across thousands of rows, resulted in memory errors.

Optimizations Implemented -

One of the first optimizations involved reviewing the data types used for storing the dataset. For instance:

- Numerical Variables : Instead of using standard floating-point or integer types (which can take up considerable memory), we downcasted numerical columns to smaller data types (e.g., from `float64` to `float32` or `int64` to `int32`).

- Categorical Variables : Categorical variables such as Gender, Continent, or Diet Type were converted into more memory-efficient data types (using `pandas.Categorical`), significantly reducing memory usage.

CONCLUSION :

In conclusion, the investigation into heart attack risks has provided valuable insights into the multifaceted influences on heart health. By analyzing both demographic and health-related factors, we have identified key patterns, such as the significant impact of age, family history, and physical activity levels on heart attack risk. The results highlight the complexity of the relationships between lifestyle choices, socioeconomic conditions, and geographic location.

Despite the limitations of the dataset, including its binary classification of heart attack risk and the absence of longitudinal data, our findings offer actionable recommendations for public health initiatives. By promoting healthier lifestyles, particularly among younger populations and those with a family history of heart disease, targeted interventions can be developed to mitigate risk factors.

Incorporating additional variables such as detailed dietary habits, genetic information, and environmental factors would enhance the comprehensiveness of future studies. This report underscores the importance of continued efforts in preventive healthcare, with the goal of reducing heart attack risks and promoting equitable healthcare access globally.

GITHUB :

Link - https://github.com/Rajesh-07-s/EDA