Heart Disease Analysis & Risk Factor Exploration

Understanding the Impact of Lifestyle and Clinical Factors on Heart Disease

Presented by : Jobina K v

Introduction

- **This project focuses on exploring and analyzing heart disease datasets to understand the key clinical and behavioral factors associated with heart health.**
- **Through Exploratory Data Analysis (EDA), patterns and relationships between various features such as age, sex, BMI, cholesterol, smoking habits, and physical activity are identified.**
- **The goal of this analysis is to understand the factors influencing heart disease occurrence and to compare how different lifestyle and medical attributes are related to heart disease status.**

Problem Statement

- **Heart disease is one of the leading causes of death worldwide.**
- **There is a need to understand how various clinical (like cholesterol, blood pressure, and age) and behavioral (like smoking, physical activity, and BMI) factors contribute to heart disease.**
- **This project aims to analyze and compare these factors using two datasets to identify :**

➔ **Key patterns and trends**

➔ **Relationships between lifestyle and medical attributes**

➔ **Factors commonly linked with heart disease occurrence**

Proposal Solution

- **We perform Exploratory Data Analysis (EDA) on two heart-related datasets one containing clinical data and another with behavioral data.**
- **Both datasets are combined to get a complete view of how medical and lifestyle factors work together.**
- **Through visualization and analysis, we aim to :**

➔ **Identify the major clinical and behavioral factors linked to heart disease.**

➔ **Compare how lifestyle habits and medical conditions influence heart health.**

➔ **Provide clear insights that help in understanding overall heart disease risk.**

Dataset Overview & Structure

- **Clinical dataset Contains medical records such as blood pressure, cholesterol, heart rate, etc.**
- **Behavioral dataset Contains lifestyle information such as smoking, alcohol consumption, physical activity, and diet.**
- **These datasets are analyzed to identify the key medical and lifestyle factors associated with heart disease.**

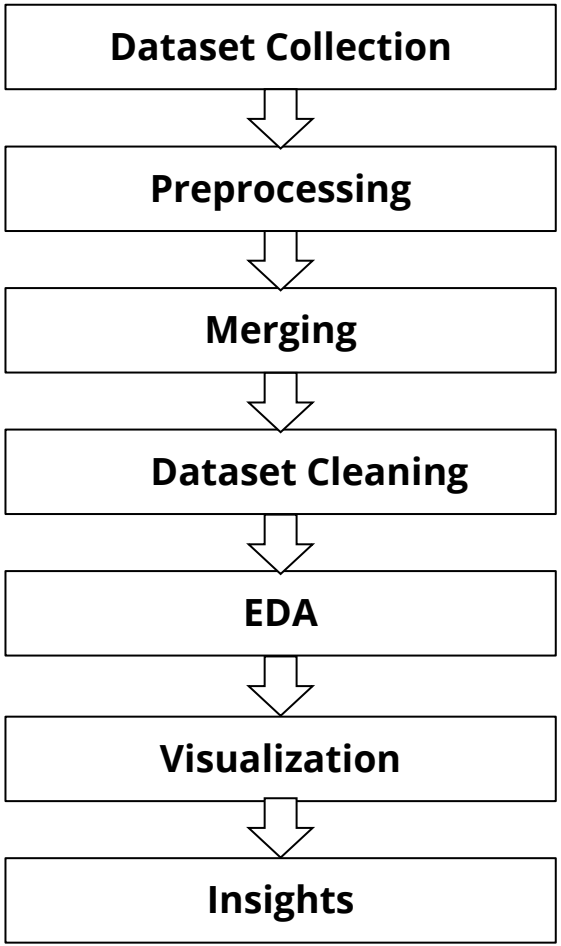
Behavioral Dataset

```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 19 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   HeartDisease      1000 non-null   object
 1   BMI               1000 non-null   float64
 2   Smoking           1000 non-null   object
 3   AlcoholDrinking   1000 non-null   object
 4   Stroke            1000 non-null   object
 5   PhysicalHealth    1000 non-null   float64
 6   MentalHealth      1000 non-null   float64
 7   DiffWalking       1000 non-null   object
 8   Sex               1000 non-null   object
 9   AgeCategory       1000 non-null   object
 10  Race              1000 non-null   object
 11  Diabetic          1000 non-null   object
 12  PhysicalActivity  1000 non-null   object
 13  GenHealth         1000 non-null   object
 14  SleepTime         1000 non-null   float64
 15  Asthma            1000 non-null   object
 16  KidneyDisease     1000 non-null   object
 17  SkinCancer        1000 non-null   object
 18  Person_ID         1000 non-null   int64
dtypes: float64(4), int64(1), object(14)
memory usage: 148.6+ KB
```

Clinical dataset

```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 15 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        1000 non-null   int64
 1   sex        1000 non-null   int64
 2   cp         1000 non-null   int64
 3   trestbps   1000 non-null   int64
 4   chol       1000 non-null   int64
 5   fbs        1000 non-null   int64
 6   restecg    1000 non-null   int64
 7   thalach    1000 non-null   int64
 8   exang      1000 non-null   int64
 9   oldpeak    1000 non-null   float64
 10  slope      1000 non-null   int64
 11  ca         1000 non-null   int64
 12  thal       1000 non-null   int64
 13  target     1000 non-null   int64
 14  Person_ID  1000 non-null   int64
dtypes: float64(1), int64(14)
memory usage: 117.3 KB
```

Workflow


Dataset Collection
Preprocessing
Merging
Dataset Cleaning
EDA
Visualization
Insights

Tools Used

- **Software: Python,Google Colab**
- **Libraries: pandas, matplotlib, Seaborn**

Implementation

- **Step 1 : Selected 50 individuals from both heart datasets for analysis.**
- **Step 2 : Checked data types, missing values, and duplicates to ensure data quality.**
- **Step 3 : Merged the clinical and behavioral datasets using the ID column.**
- **Step 4 : Plotted bar charts and other visualizations to study the distribution of :**

➔ **Age & Age Category**
➔ **Sex**
➔ **BMI**
➔ **Smoking, Alcohol Drinking, Stroke**
➔ **Physical & Mental Health**
➔ **Heart Disease status**

- **Step 5 : Performed correlation and comparative analysis to examine relationships between lifestyle/clinical factors and heart disease occurrence.**

Behavioral dataset info

```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 19 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   HeartDisease      1000 non-null   object
 1   BMI               1000 non-null   float64
 2   Smoking           1000 non-null   object
 3   AlcoholDrinking   1000 non-null   object
 4   Stroke            1000 non-null   object
 5   PhysicalHealth    1000 non-null   float64
 6   MentalHealth      1000 non-null   float64
 7   DiffWalking       1000 non-null   object
 8   Sex               1000 non-null   object
 9   AgeCategory       1000 non-null   object
 10  Race              1000 non-null   object
 11  Diabetic          1000 non-null   object
 12  PhysicalActivity  1000 non-null   object
 13  GenHealth         1000 non-null   object
 14  SleepTime         1000 non-null   float64
 15  Asthma            1000 non-null   object
 16  KidneyDisease     1000 non-null   object
 17  SkinCancer        1000 non-null   object
 18  Person_ID         1000 non-null   int64
dtypes: float64(4), int64(1), object(14)
memory usage: 148.6+ KB
```

Clinical dataset info

```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 15 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   age        1000 non-null   int64
 1   sex        1000 non-null   int64
 2   cp         1000 non-null   int64
 3   trestbps   1000 non-null   int64
 4   chol       1000 non-null   int64
 5   fbs        1000 non-null   int64
 6   restecg    1000 non-null   int64
 7   thalach    1000 non-null   int64
 8   exang      1000 non-null   int64
 9   oldpeak    1000 non-null   float64
 10  slope      1000 non-null   int64
 11  ca         1000 non-null   int64
 12  thal       1000 non-null   int64
 13  target     1000 non-null   int64
 14  Person_ID  1000 non-null   int64
dtypes: float64(1), int64(14)
memory usage: 117.3 KB
```

Merged dataset info

```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 32 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   HeartDisease      1000 non-null   int64
 1   BMI               1000 non-null   float64
 2   Smoking           1000 non-null   int64
 3   AlcoholDrinking   1000 non-null   int64
 4   Stroke            1000 non-null   int64
 5   PhysicalHealth    1000 non-null   float64
 6   MentalHealth      1000 non-null   float64
 7   DiffWalking       1000 non-null   int64
 8   Sex               1000 non-null   int64
 9   AgeCategory       1000 non-null   object
 10  Race              1000 non-null   object
 11  Diabetic          1000 non-null   int64
 12  PhysicalActivity  1000 non-null   int64
 13  GenHealth         1000 non-null   object
 14  SleepTime         1000 non-null   float64
 15  Asthma            1000 non-null   int64
 16  KidneyDisease     1000 non-null   int64
 17  SkinCancer        1000 non-null   int64
 18  Person_ID         1000 non-null   int64
 19  Age               1000 non-null   int64
 20  cp                1000 non-null   int64
 21  trestbps          1000 non-null   int64
 22  chol              1000 non-null   int64
 23  fbs               1000 non-null   int64
 24  restecg           1000 non-null   int64
 25  thalach           1000 non-null   int64
 26  exang             1000 non-null   int64
 27  oldpeak           1000 non-null   float64
 28  slope             1000 non-null   int64
 29  ca                1000 non-null   int64
 30  thal              1000 non-null   int64
 31  target            1000 non-null   int64
dtypes: float64(5), int64(24), object(3)
memory usage: 250.1+ KB
```

Data Cleaning

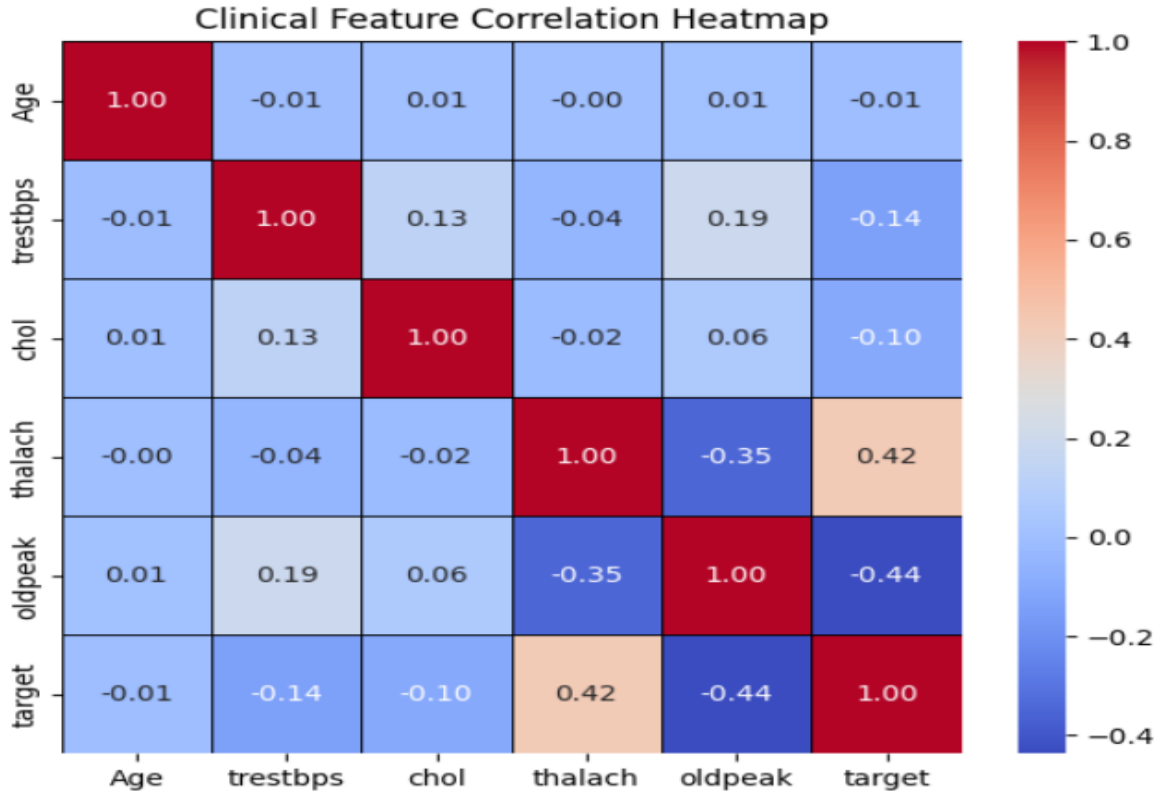
	0
HeartDisease	0
BMI	0
Smoking	0
AlcoholDrinking	0
Stroke	0
PhysicalHealth	0
MentalHealth	0
DiffWalking	0
Sex	0
AgeCategory	0
Race	0
Diabetic	0
PhysicalActivity	0
GenHealth	0
SleepTime	0
Asthma	0
KidneyDisease	0
SkinCancer	0
Person_ID	0

Age	0
cp	0
trestbps	0
chol	0
fbs	0
restecg	0
thalach	0
exang	0
oldpeak	0
slope	0
ca	0
thal	0
target	0

```
merged_df.duplicated().sum()
```

```
np.int64(0)
```

Clinical Feature Correlation Heatmap

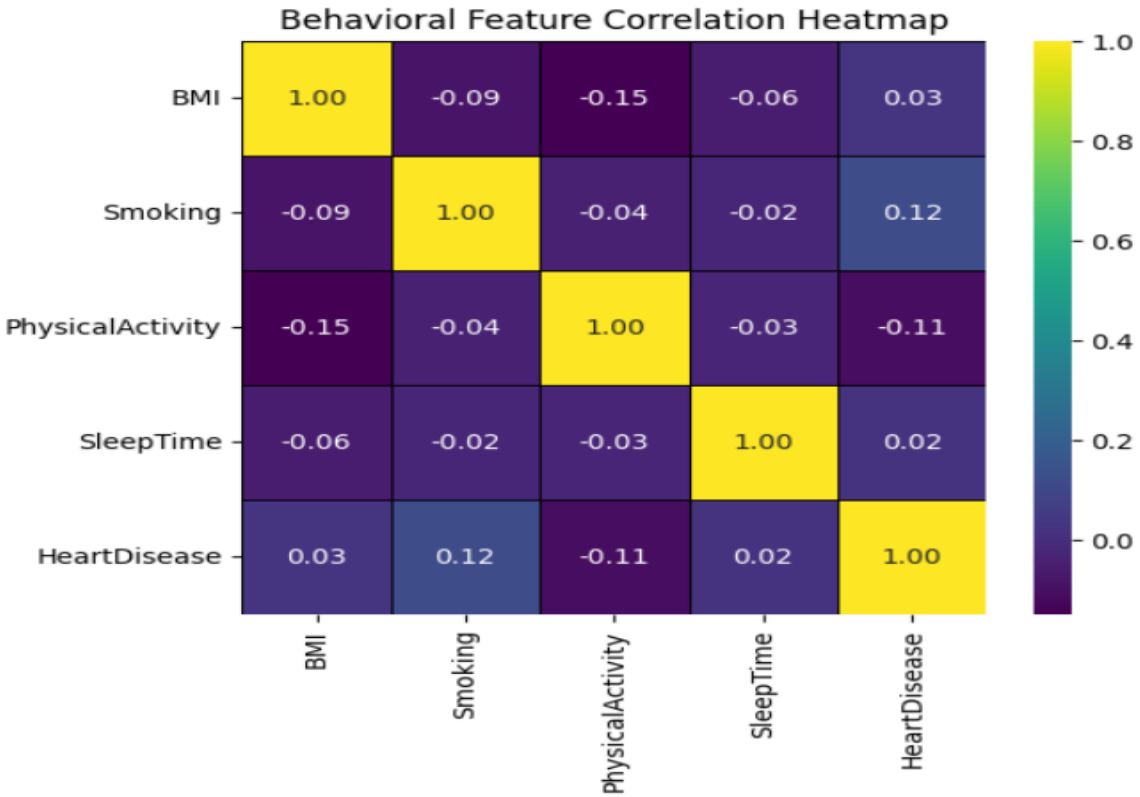


- **The Purpose of the graph is To study how clinical health indicators (like blood pressure, cholesterol, heart rate, etc.) are related to each other and to heart disease.**

Key Insights:

- **Higher thalach (maximum heart rate achieved) correlates positively with heart disease absence (r=0.42).**
- **Higher oldpeak (ST depression) has a negative correlation with heart disease (r=-0.43), meaning higher ST depression increases risk.**
- **Blood pressure (trestbps) and cholesterol show weaker correlations.**

Behavioral Feature Correlation Heatmap

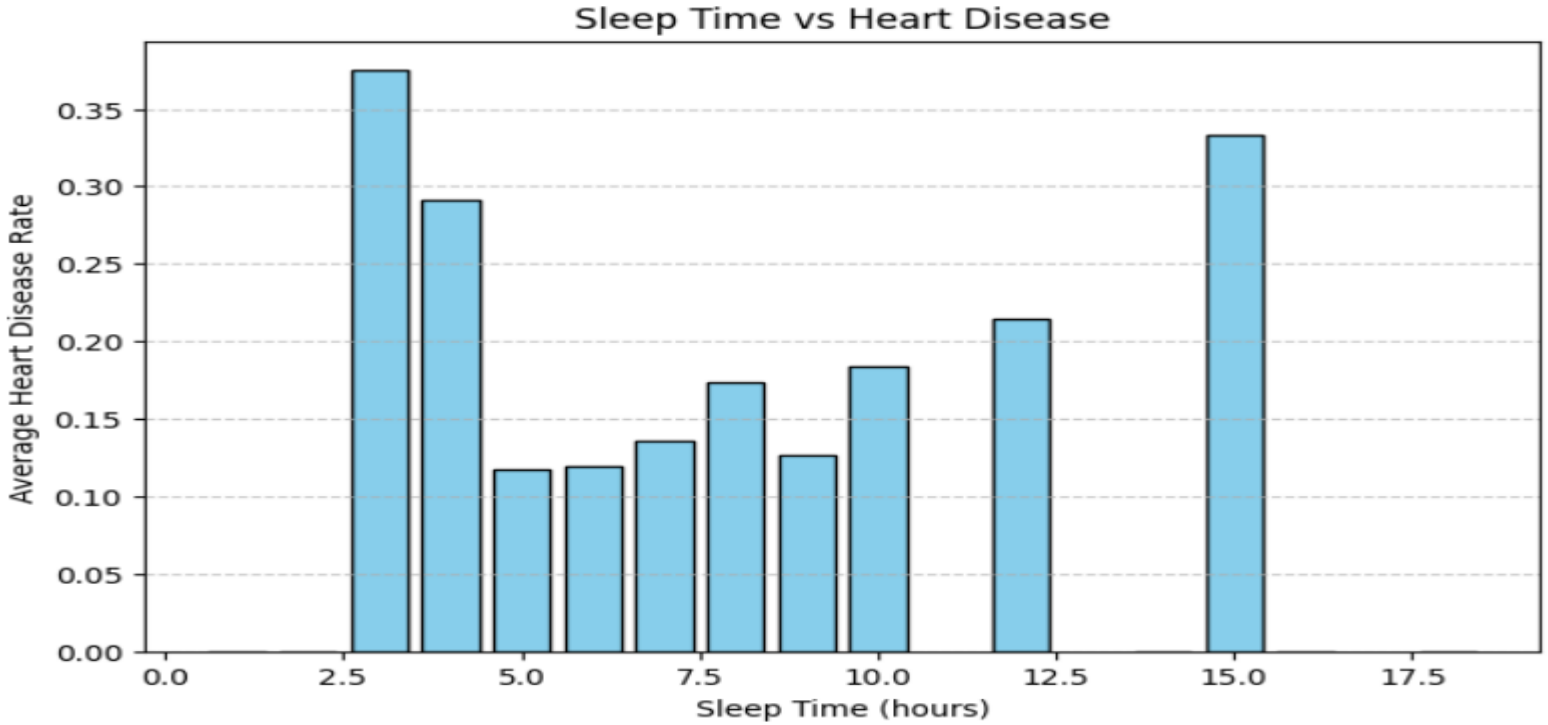


- **The Purpose of Graph is To understand how lifestyle behaviors (like BMI, smoking, sleep, and physical activity) relate to each other and to heart disease.**

Key Insights :

- **Smoking shows a mild positive correlation (r=0.11) with heart disease.**
- **Physical activity is negatively correlated (r=-0.11), suggesting an active lifestyle reduces heart disease risk.**
- **BMI and sleep time have weak correlations but still contribute cumulatively.**

Sleep Time vs Heart Disease

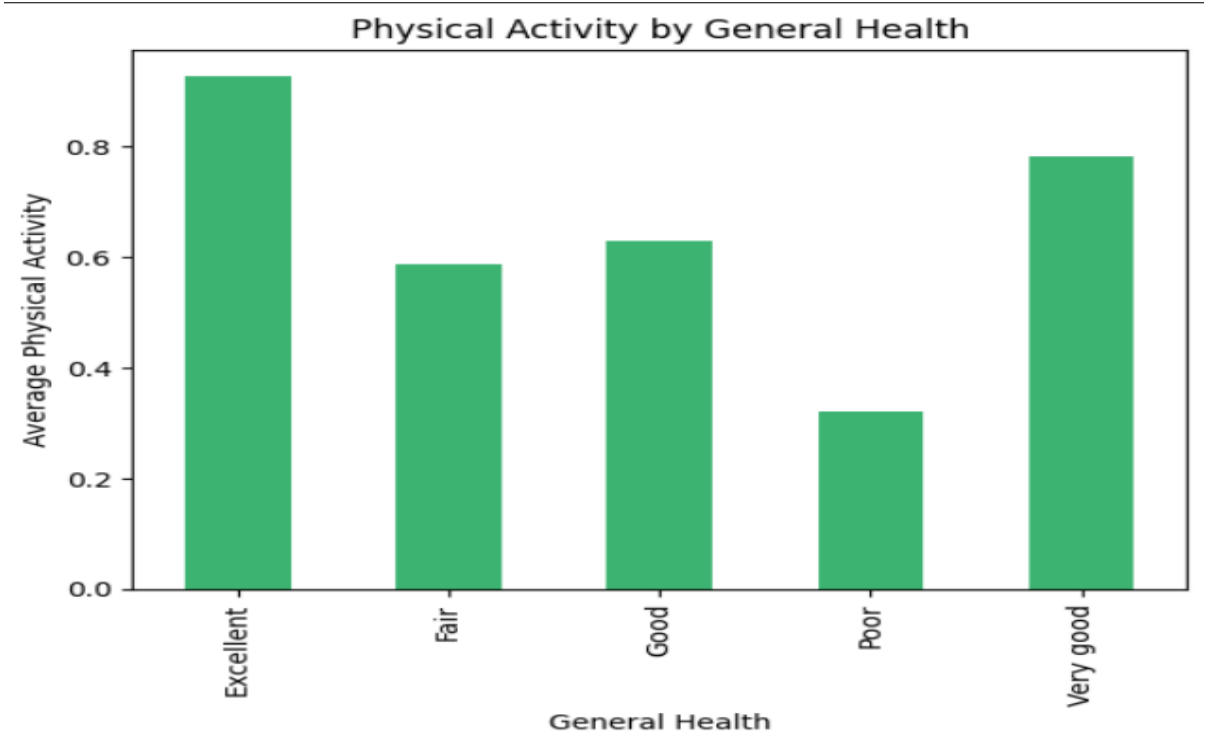


- **The Purpose ot the Graph is To understand how the number of hours a person sleeps per day affects the chances of developing heart disease.**

Key Insights :

- **People who sleep less than 6 hours or more than 10 hours show a higher likelihood of heart disease.**
- **The lowest risk appears between 7–8 hours of sleep per night.**
- **Both very short and very long sleep durations can negatively affect heart health.**
- **Balanced sleep (6–8 hours) lowers heart disease risk.**
- **sleeping less than 6 hours or more than 9 hours increases heart disease risk.**

Physical Activity by General Health

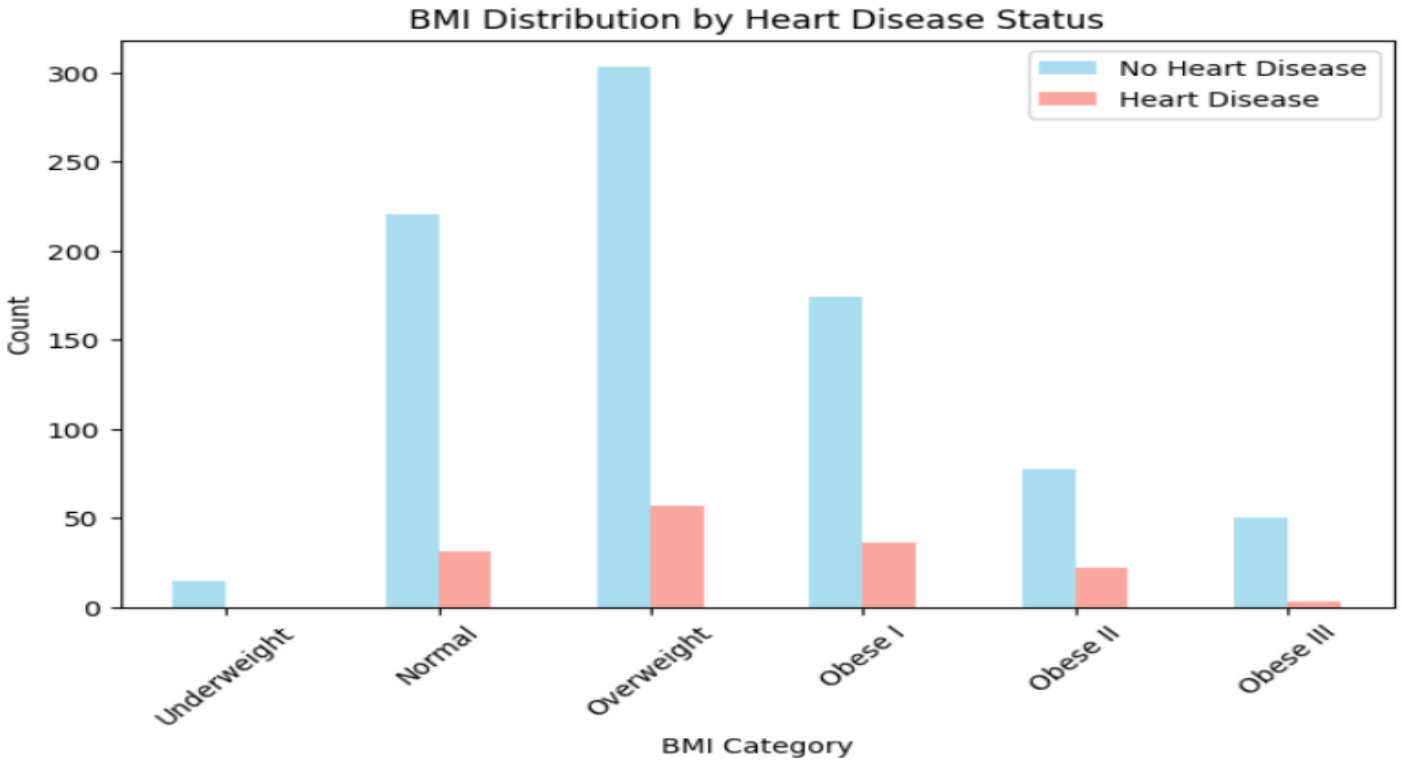


- **The Purpose of Graph is To examine how Physical Activity levels differ among people with different self-reported General Health conditions.**

Key Insights :

- **People with Excellent or Very Good health have the highest physical activity levels.**
- **Poor general health corresponds with the lowest activity rates.**
- **Suggests exercise contributes to better health perception.**
- **regular physical activity improves overall health and significantly reduces heart disease risk.**

BMI Distribution by Heart Disease Status

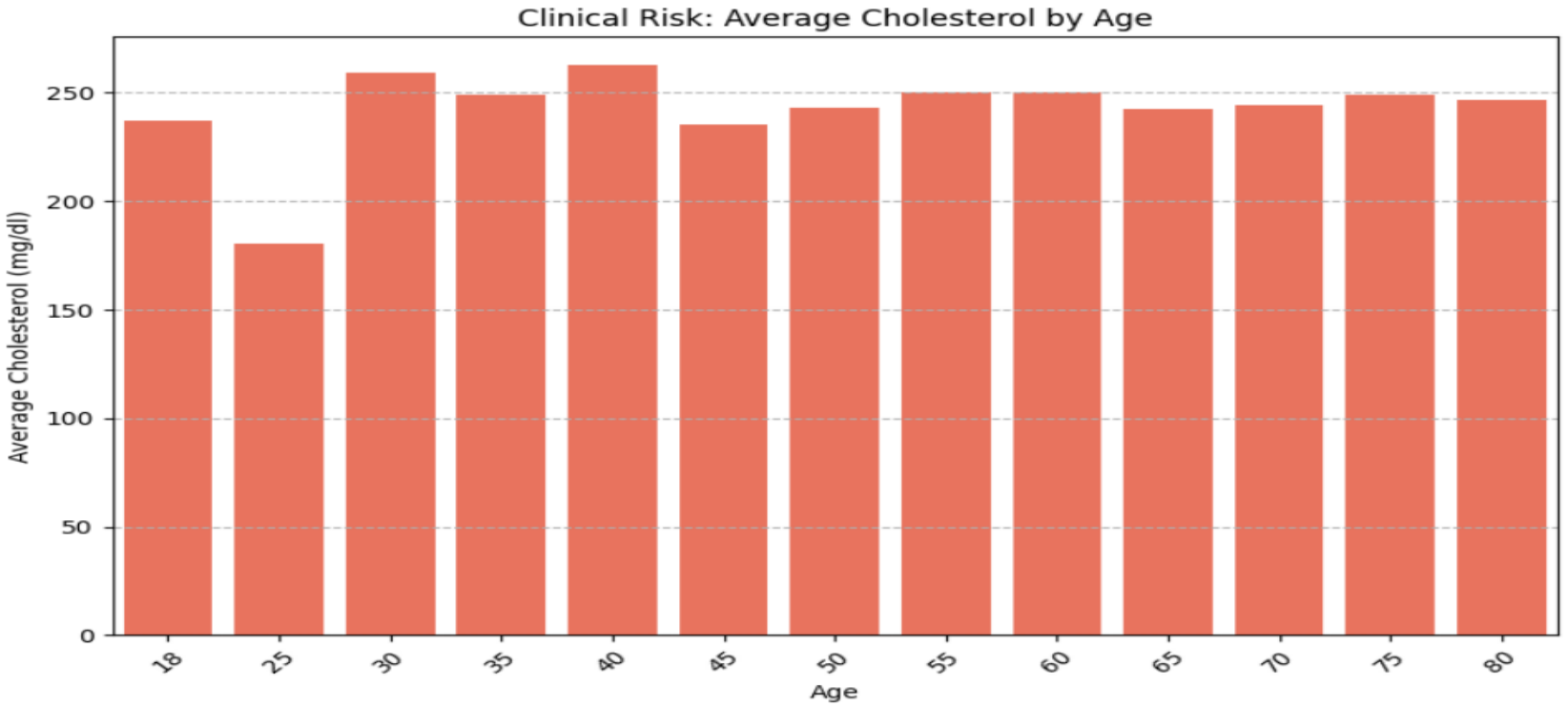


- **The Purpose of Graph is To study how Body Mass Index (BMI) categories are distributed among people with and without heart disease.**

Key Insights :

- **Overweight and Obese (Class I & II) individuals have higher heart disease counts.**
- **Underweight and Normal categories show lower risk.**
- **Suggests a positive link between higher BMI and risk of heart disease.**
- **overweight and obesity as major risk factors for heart diseases due to increased blood pressure.**

Average Cholesterol by Age

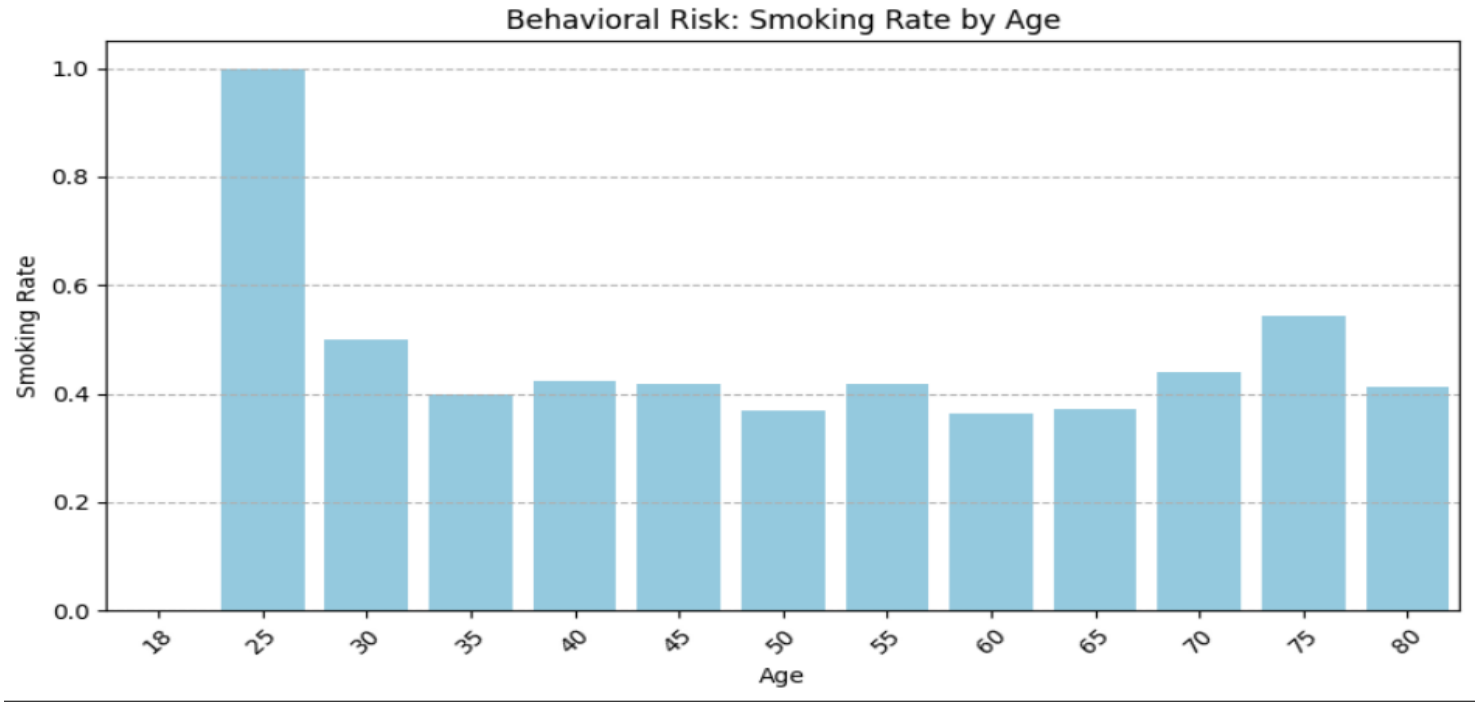


- **The Purpose of the Graph is To analyze how average cholesterol levels change with age.**

Key Insights :

- **Cholesterol levels generally increase with age until mid-60s, then stabilize.**
- **Age 40–60 shows particularly higher cholesterol averages (>250 mg/dL).**
- **cholesterol increases with age, and elevated cholesterol is a leading cause of coronary heart disease.**

Behavioral Risk – Smoking Rate by Age

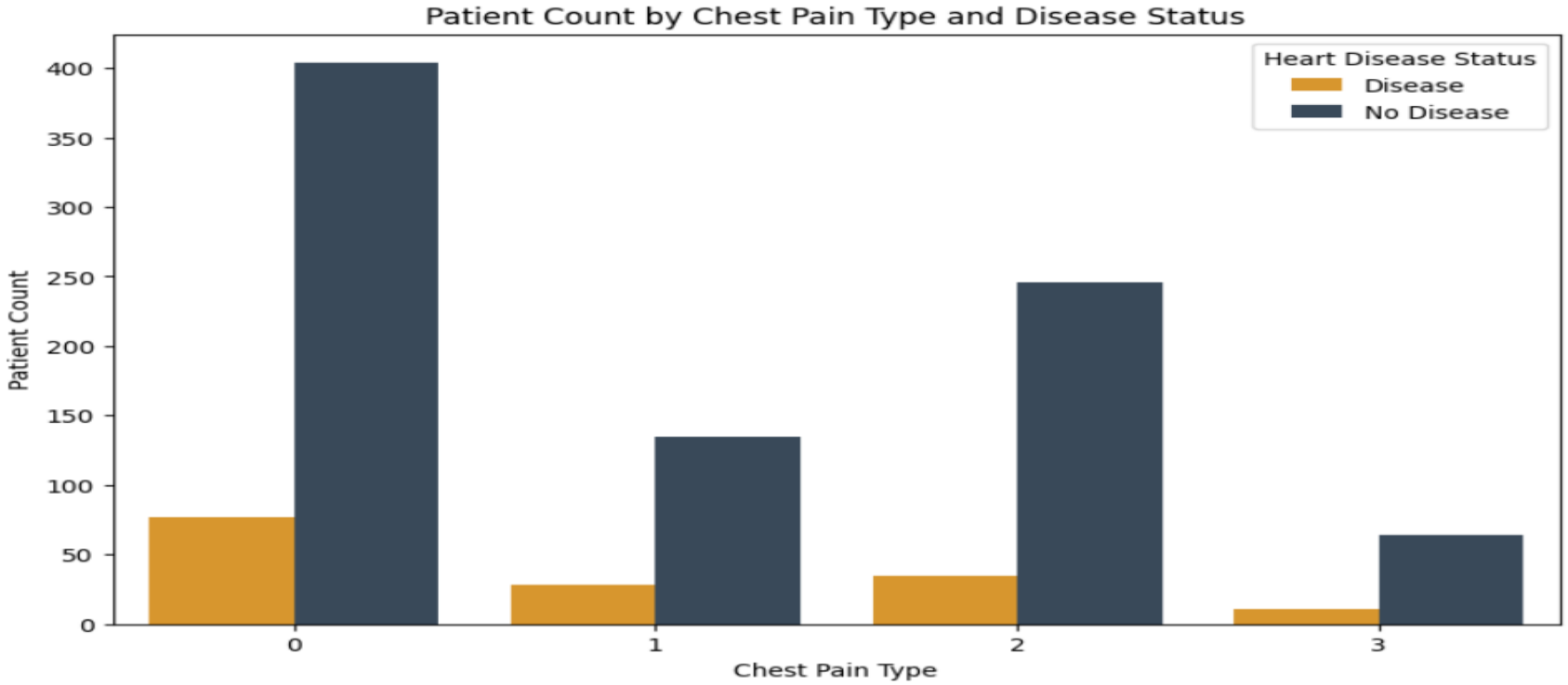


- **The Purpose of the Graph is To analyze smoking habits across different age groups and their relationship to heart disease risk.**

Key Insights :

- **Smoking rates rise from early adulthood (25–35) and remain high till 70+.**
- **Middle-aged adults (40–70) show the highest smoking prevalence.**
- **long-term smoking damages blood vessels and significantly increases heart attack risk, especially in adults over 35.**

Chest Pain Type and Disease Status

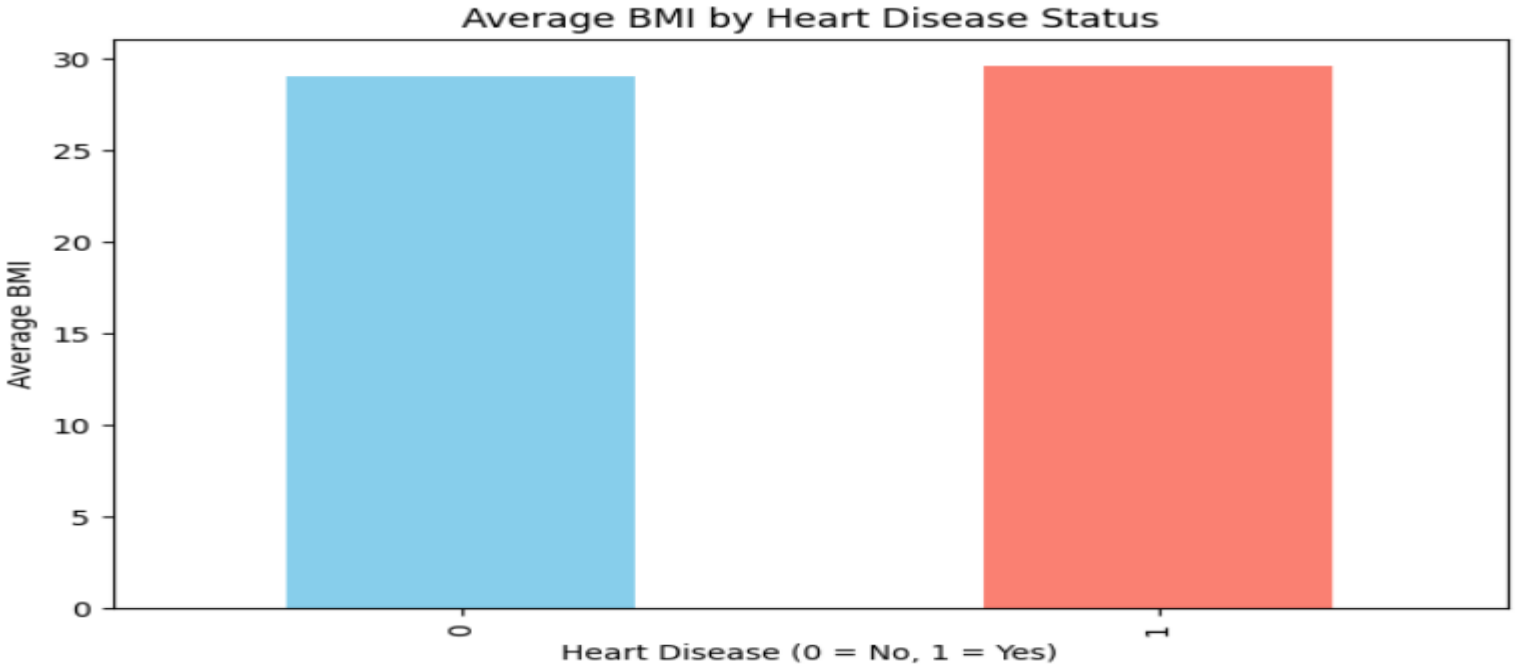


- **The Purpose of the Graph is To explore how different types of chest pain relate to the presence of heart disease.**

Key Insights :

- **People with chest pain type 0 (asymptomatic) mostly do not have heart disease.**
- **Chest Pain Type 0 (Typical Angina) shows the highest number of heart disease cases.**
- **Type 1 and 2 (Atypical/Non-anginal) have fewer cases.**
- **Highlights the diagnostic importance of chest pain type.**

Average BMI by Heart Disease Status

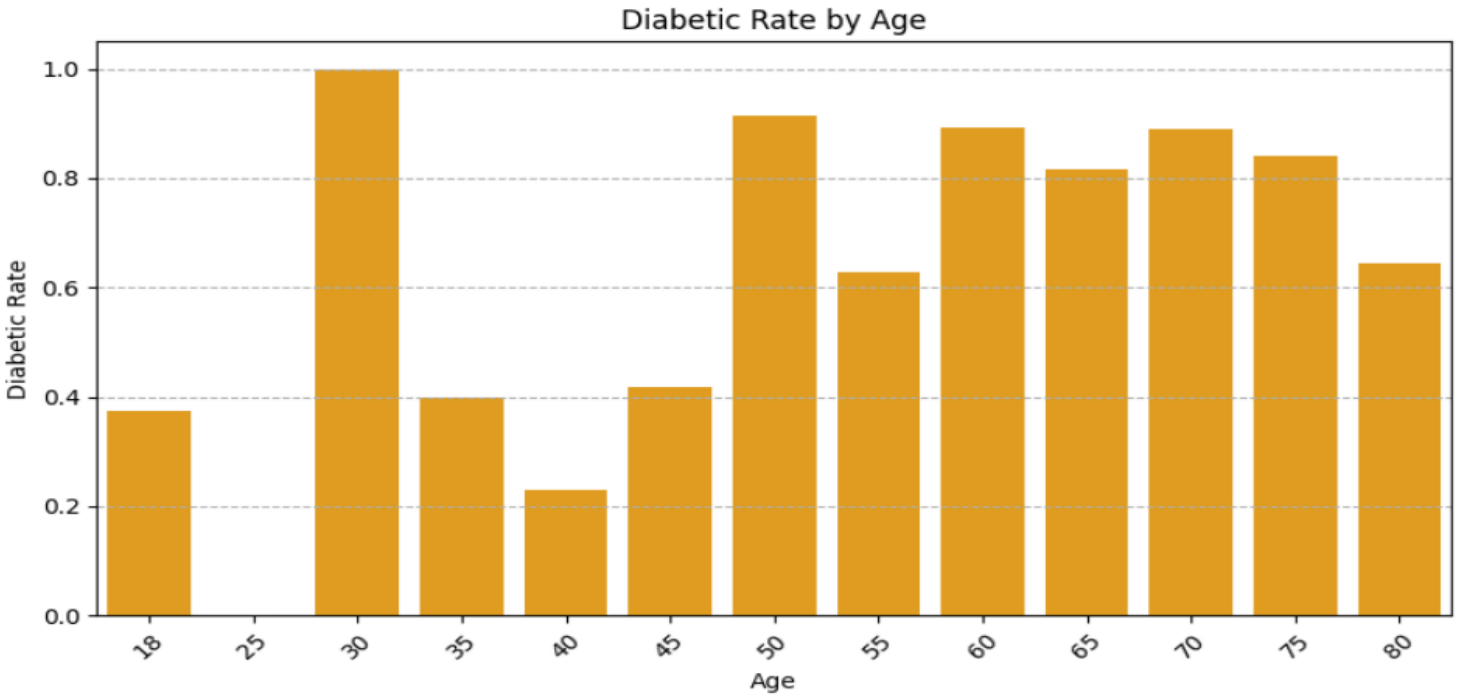


- **The Purpose of the Graph is To compare the average Body Mass Index (BMI) between people with and without heart disease.**

Key Insights :

- **People with heart disease have slightly higher average BMI (29.5) than those without (29.0).**
- **Indicates a possible link between higher BMI and heart disease risk.**
- **Useful for identifying BMI as a behavioral risk factor.**

Diabetic Rate by Age

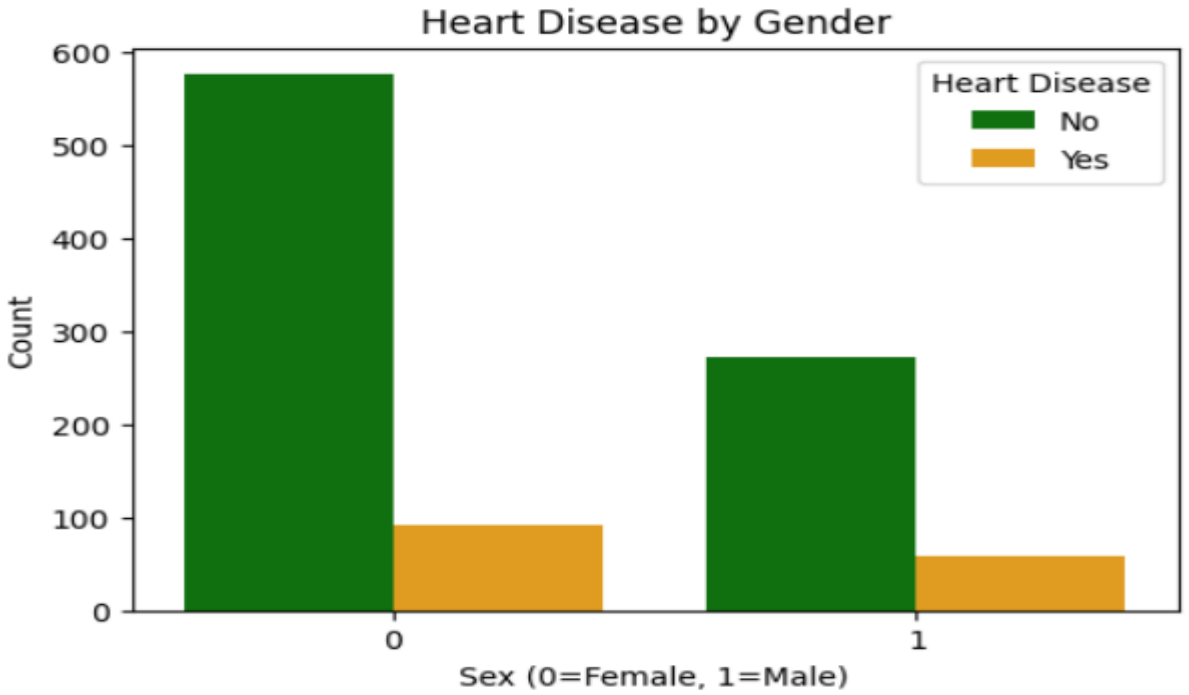


- **The Purpose of Graph is To explore how diabetes prevalence changes with age and its potential link to heart disease.**

Key Insights :

- **Diabetes prevalence increases sharply after age 45 and peaks around 60–70.**
- **Early detection and lifestyle management are key after midlife.**

Heart Disease by Gender

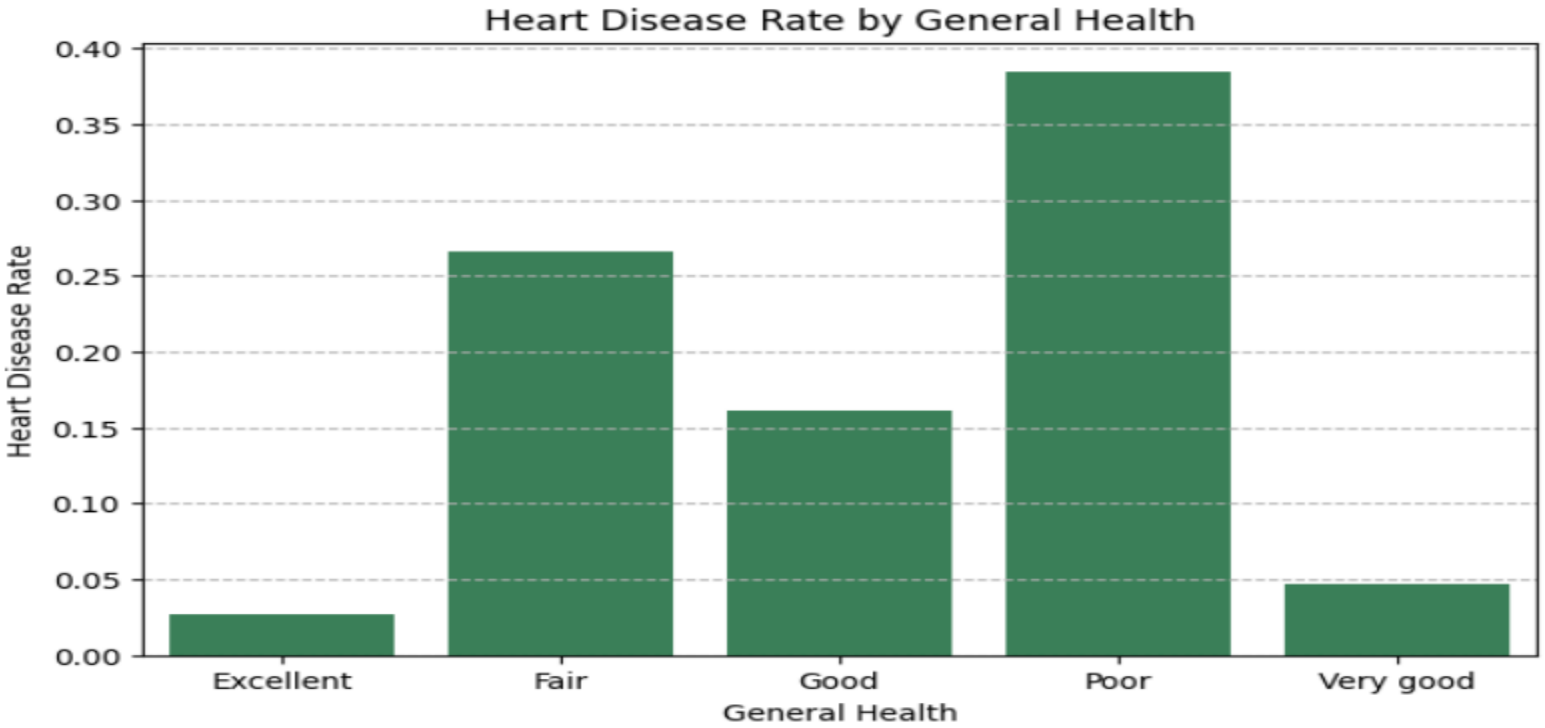


- **The Purpose of the Graph is To show how heart disease is distributed between males and females.**

Key Insights :

- **Males show higher counts of heart disease than females.**
- **Females have lower overall risk but rising rates after age 55.**

Heart Disease Rate by General Health



- **The Purpose of the Graph To analyze how individuals' self-assessed general health relates to heart disease risk.**

Key Insights :

- **Those reporting Poor or Fair health have significantly higher heart disease rates.**
- **People with Excellent or Very Good health show the lowest risk.**

Results

- Age, cholesterol, and blood pressure are key clinical risk factors.
- Age and gender also play major roles in determining heart disease likelihood.
- BMI, smoking, and physical activity are important behavioral risk factors.
- Heart disease is more common in older age groups and higher BMI categories.
- Males show slightly higher occurrence than females.
- Combined lifestyle and medical factors help understand heart disease risk.

Conclusion

- **Both clinical and behavioral factors contribute to heart disease.**
- **Age, BMI, physical activity, and smoking are key indicators.**
- **Regular health checks, a balanced diet, physical activity, and controlled BMI help reduce risks.**
- **EDA helped visualize clear patterns for understanding heart health better.**
- **Insights can guide preventive measures and lifestyle decisions.**

References

- **Heart Disease Datasets – Kaggle**
- **Python & Libraries – Pandas, NumPy, Matplotlib, Seaborn**

THANK YOU