# FINAL REPORT: HEALTHCARE ANALYTICS

By

**CHELSA MARIAM JOHN**

**12214129**

As part of

Analyse Health and Demographic Data to identify common traits leading to Heart Disease, by Practo


**School of Computer Science and Engineering**

Lovely Professional University

Phagwara, Punjab (India)

July, 2024

# Healthcare Analytics: Identifying Common Traits Leading to Heart Disease

## 1. Introduction to Healthcare Analytics

Healthcare analytics is the systematic use of data and analytical methods to improve decision-making, patient outcomes, and operational efficiency in healthcare settings. It involves collecting, analyzing, and interpreting vast amounts of healthcare data to derive meaningful insights that can inform clinical practice, policy-making, and healthcare management.

### Role in Heart Disease Prevention and Management

In the context of heart disease, healthcare analytics plays a crucial role in:

**1. Risk Prediction:** By analyzing large datasets, healthcare analytics can identify patterns and risk factors associated with heart disease, allowing for early intervention and prevention strategies.

**2. Personalized Treatment:** Analytics helps in tailoring treatment plans based on individual patient characteristics, improving outcomes and reducing adverse effects.

**3. Population Health Management:** It enables healthcare providers to identify high-risk populations and implement targeted interventions at a community level.

**4. Resource Allocation:** Analytics can optimize the distribution of healthcare resources by predicting demand and identifying areas of greatest need.

## 2. Methodology

This study utilized a comprehensive dataset (heart_disease_health_indicators.csv) containing information on various health indicators and heart disease status. We employed the following analytical methods:

**1. Univariate Analysis:** To understand the distribution and characteristics of individual variables.
**2. Bivariate Analysis**: To explore relationships between variables, particularly their association with heart disease.
**3**. **Visualization Techniques**: Including histograms, box plots, bar charts, pie charts, scatter plots, line plots, violin plots, heatmaps, and pair plots.

### 2.1 Dataset Overview [heart_disease_health_indicators.csv] :

```
print(df.info())
print("\nSample of the data:")
print(df.head())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 229781 entries, 0 to 229780
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype   
---  ------                --------------   -----   
 0   HeartDiseaseorAttack  229781 non-null  category
 1   HighBP                229781 non-null  category
 2   HighChol              229781 non-null  category
 3   CholCheck             229781 non-null  category
 4   BMI                   229781 non-null  int64   
 5   Smoker                229781 non-null  category
 6   Stroke                229781 non-null  category
 7   Diabetes              194684 non-null  float64 
 8   PhysActivity          229781 non-null  category
 9   Fruits                229781 non-null  category
 10  Veggies               229781 non-null  category
 11  HvyAlcoholConsump     229781 non-null  category
 12  AnyHealthcare         229781 non-null  category
 13  NoDocbcCost           229781 non-null  category
 14  GenHlth               229781 non-null  category
 15  MentHlth              229781 non-null  int64   
 16  PhysHlth              229781 non-null  int64   
 17  DiffWalk              229781 non-null  category
 18  Sex                   229781 non-null  category
 19  Age                   229781 non-null  int64   
 20  Education             229781 non-null  category
 21  Income                229781 non-null  category
dtypes: category(17), float64(1), int64(4)
memory usage: 12.5 MB
None

Sample of the data:
  HeartDiseaseorAttack HighBP HighChol CholCheck  BMI Smoker Stroke  Diabetes  \
0                    0      1        1         1   40      1      0       0.0   
1                    0      0        0         0   25      1      0       0.0   
2                    0      1        1         1   28      0      0       0.0   
3                    0      1        0         1   27      0      0       0.0   
4                    0      1        1         1   24      0      0       0.0   

  PhysActivity Fruits  ... AnyHealthcare NoDocbcCost GenHlth MentHlth  \
0            0      0  ...             1           0       5       18   
1            1      0  ...             0           1       3        0   
2            0      1  ...             1           1       5       30   
3            1      1  ...             1           0       2        0   
4            1      1  ...             1           0       2        3   

  PhysHlth  DiffWalk  Sex Age Education  Income  
0       15         1    0   9         4       3  
1        0         0    0   7         6       1  
2       30         1    0   9         4       8  
3        0         0    0  11         3       6  
4        0         0    0  11         5       4  

[5 rows x 22 columns]
```

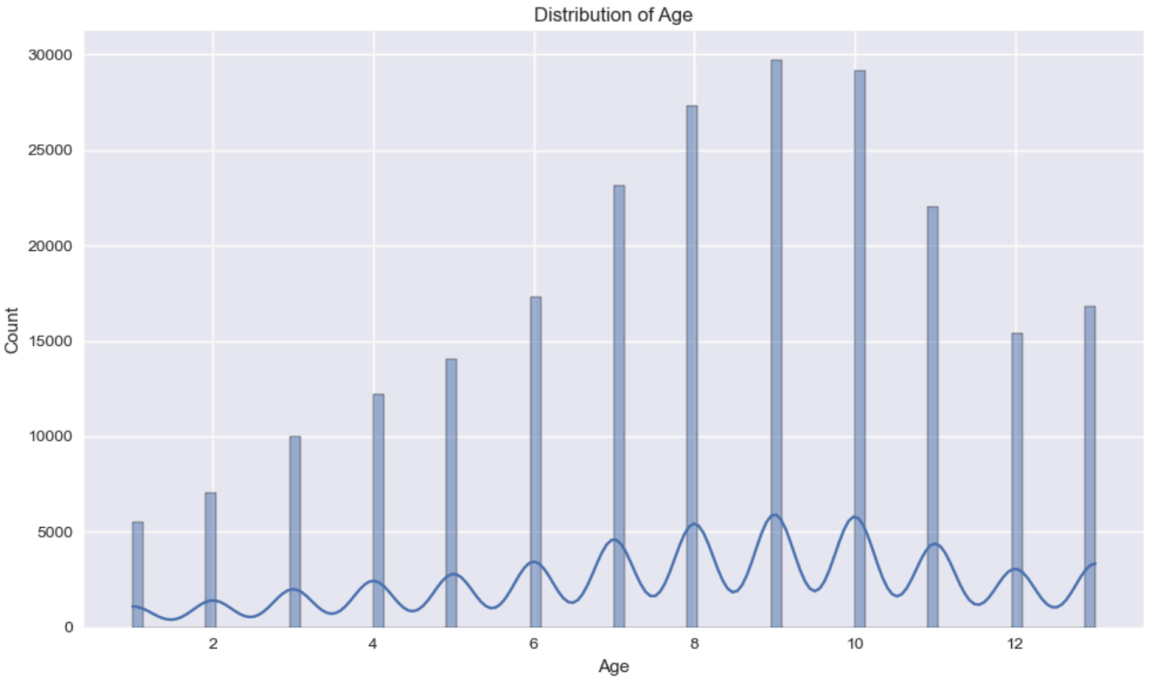
## 3. Common Traits Leading to Heart Disease

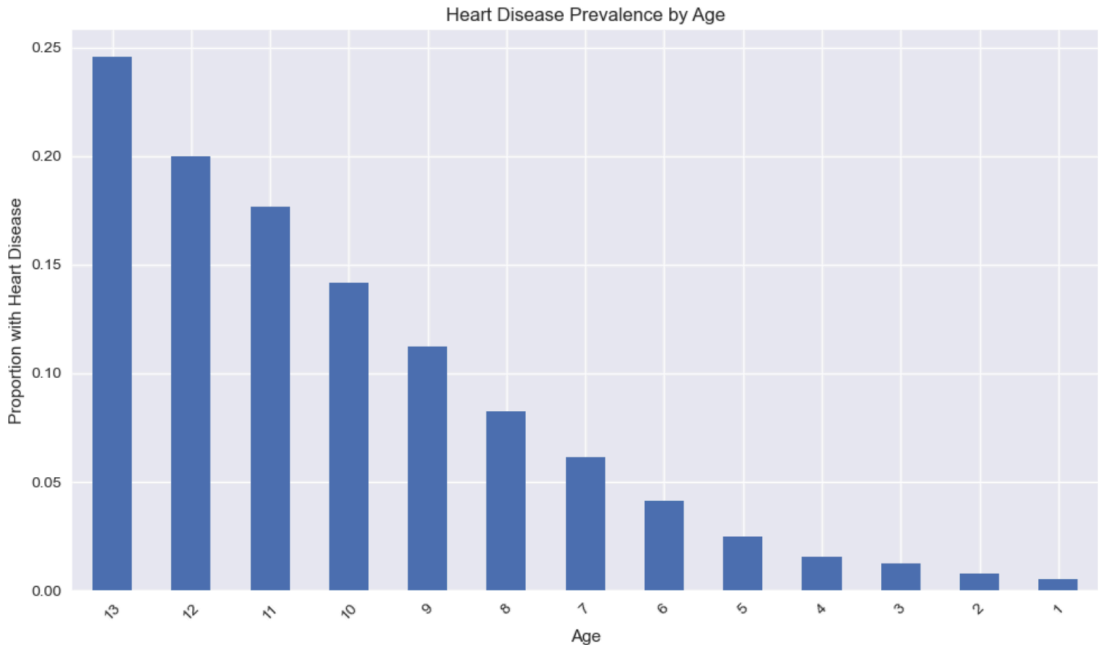
Our analysis identified several common traits associated with increased risk of heart disease:

### 3.1 Demographic Factors:

a) Age:

- Age categories range from 1 to 13
- Mean age category: 8.09 - Median age category: 8
- Standard deviation: 3.09
- Heart disease prevalence increases significantly with age, ranging from 0.5% in the youngest group to 24.55% in the oldest.

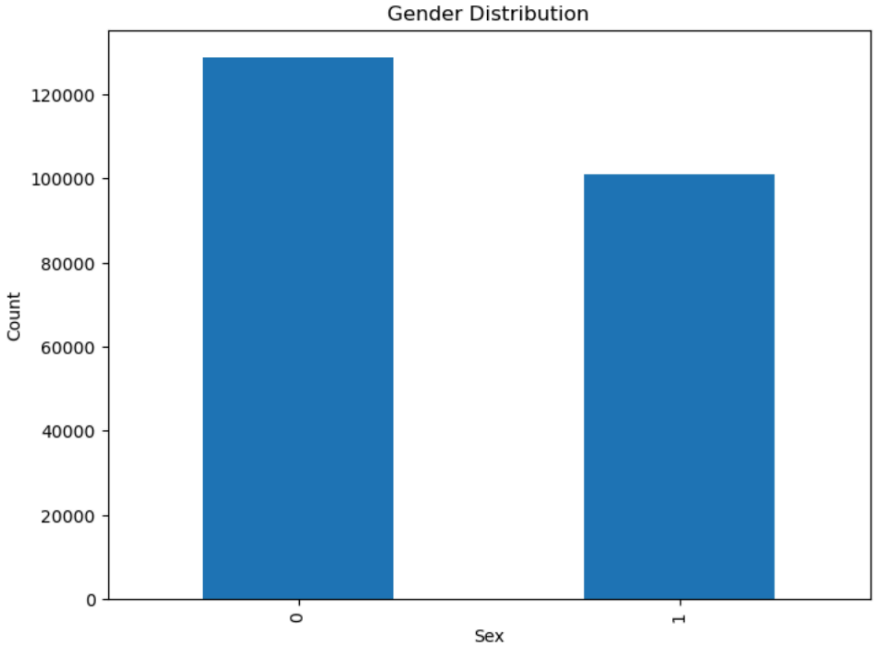


```
Heart Disease Prevalence by Age:
Age
13    0.245524
12    0.199818
11    0.176761
10    0.141765
9     0.112524
8     0.082341
7     0.061538
6     0.041101
5     0.024982
4     0.015776
3     0.012569
2     0.007640
1     0.005261
Name: HeartDiseaseorAttack, dtype: float64
```

b) Gender:

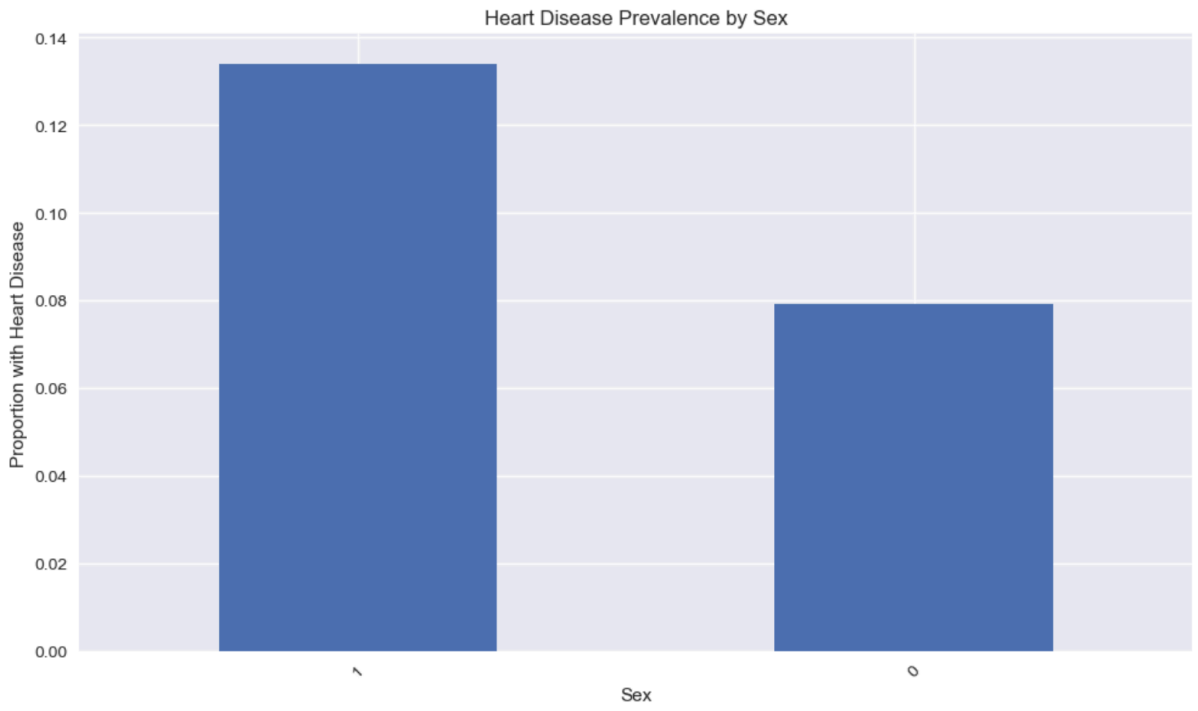
- Sex 1 (male): 13.41% heart disease prevalence
- Sex 0 (female): 7.91% heart disease prevalence



```
Heart Disease Rates by Gender:
Sex
0    0.079058
1    0.134057
Name: HeartDiseaseorAttack, dtype: float64

Heart Disease Rates by Gender (Percentage):
Sex
0     7.905847
1    13.405729
Name: HeartDiseaseorAttack, dtype: float64

Detailed Heart Disease Counts by Gender:
HeartDiseaseorAttack          0          1
Sex
0                     92.094153   7.905847
1                     86.594271  13.405729
```




```
Heart Disease Prevalence by Sex:
Sex
1    0.134057
0    0.079058
Name: HeartDiseaseorAttack, dtype: float64
```

**3.2 Lifestyle Factors:**

a) Smoking:

- 46.6% of the population are smokers
- Smokers have a 13.7% heart disease rate, compared to 7.3% for non-smokers

b) Physical Activity:

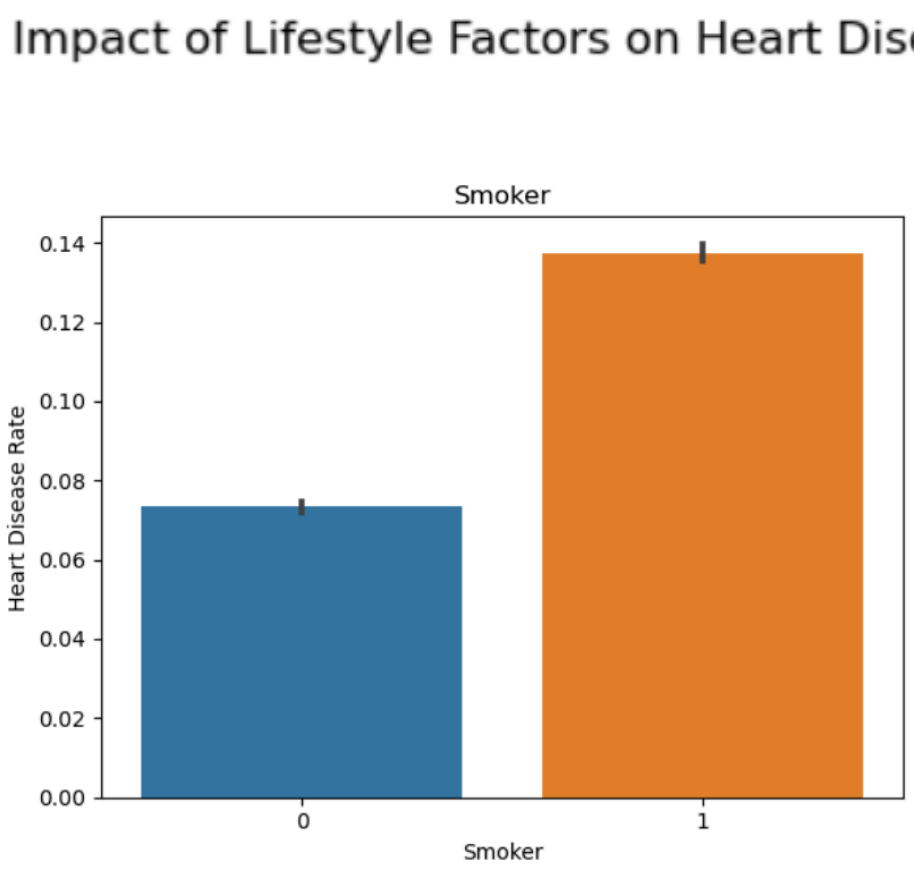
- 73.3% engage in regular physical activity
- Those who are physically active have a 9.0% heart disease rate, compared to 14.0% for inactive individuals

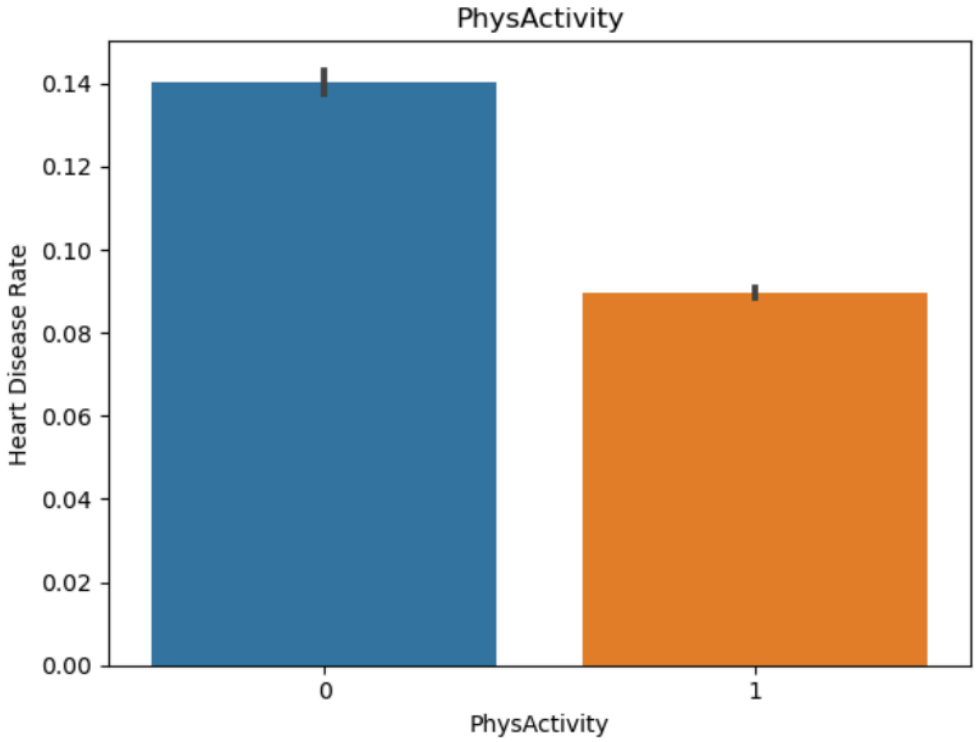
c) Diet:

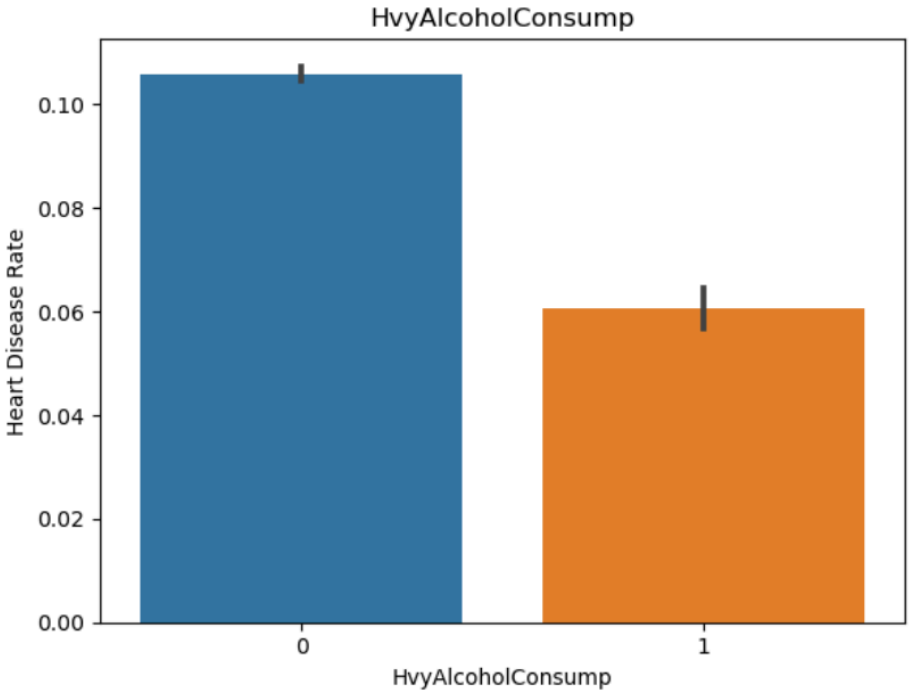
- Fruit Consumption: 61.3% consume fruits regularly (10.1% heart disease rate vs 10.6% for non-consumers)
- Vegetable Consumption: 79.5% consume vegetables regularly (9.9% heart disease rate vs 11.9% for non-consumers)

d) Alcohol Consumption:

- 6.1% are heavy alcohol consumers
- Heavy consumers have a 6.1% heart disease rate, compared to 10.6% for non-heavy drinkers





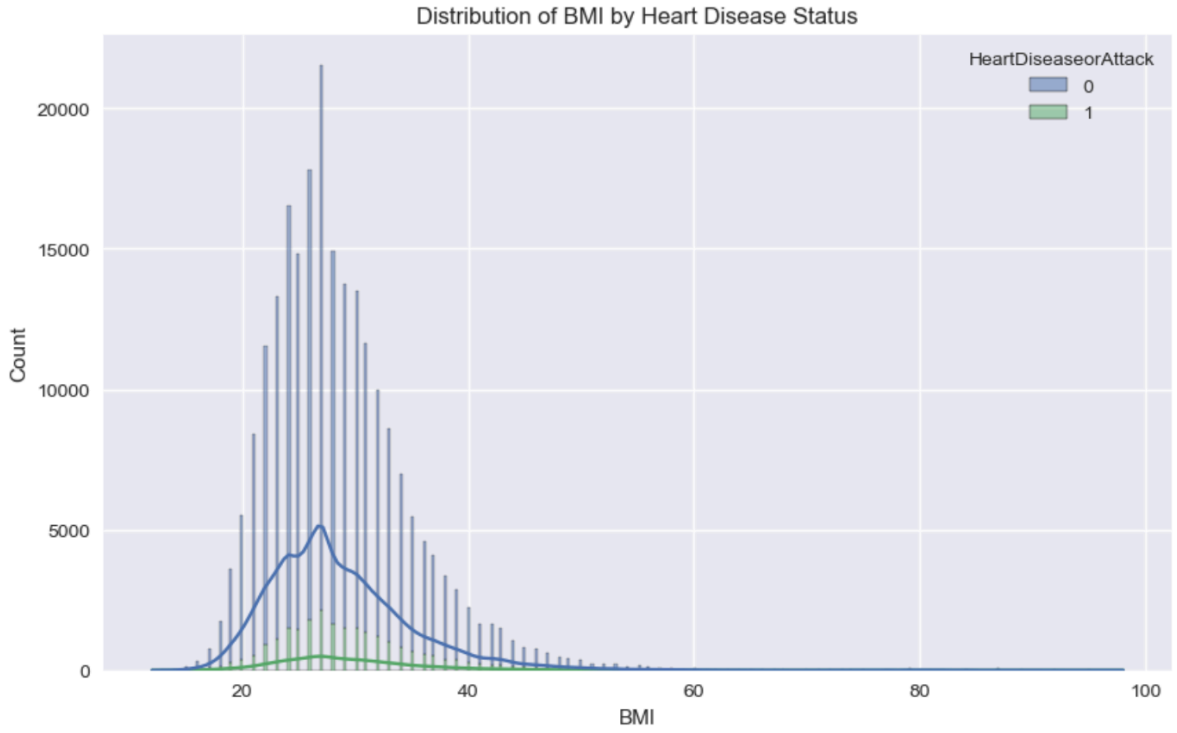



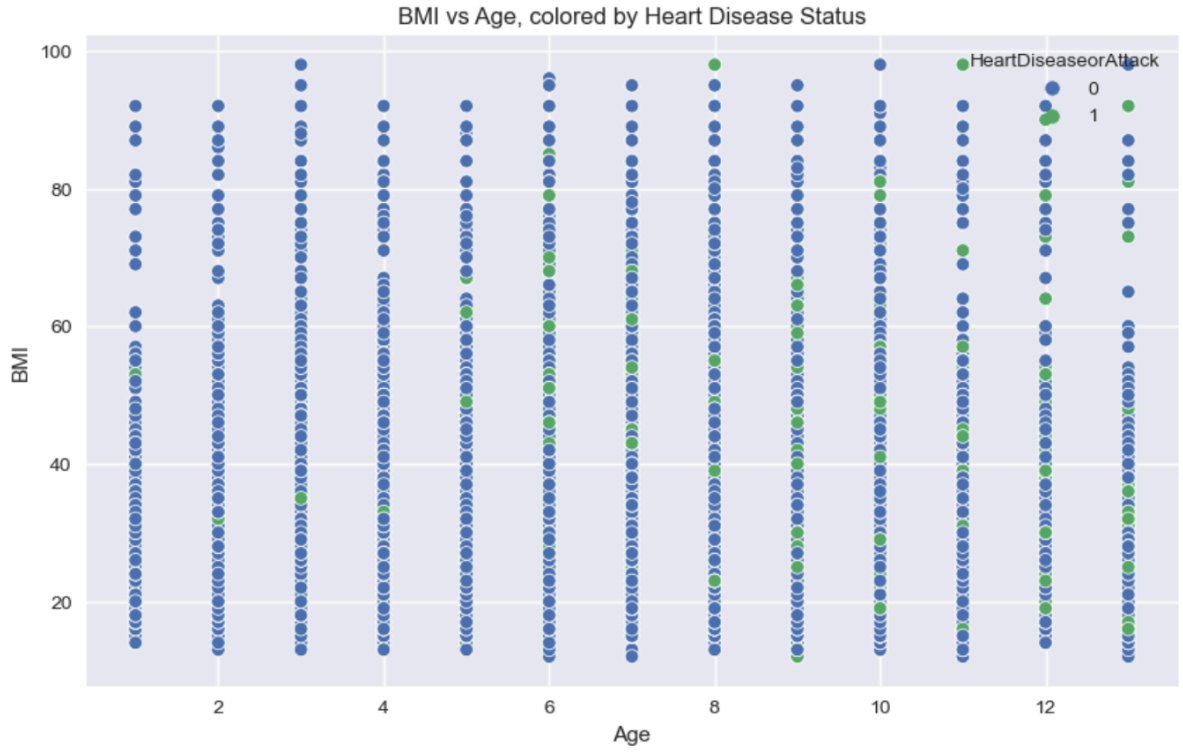

**3.3 Health Conditions:**

a) Body Mass Index (BMI):

- Mean BMI: 28.69 (overweight category)
- Median BMI: 27
- Range: 12 to 98
- Standard deviation: 6.79

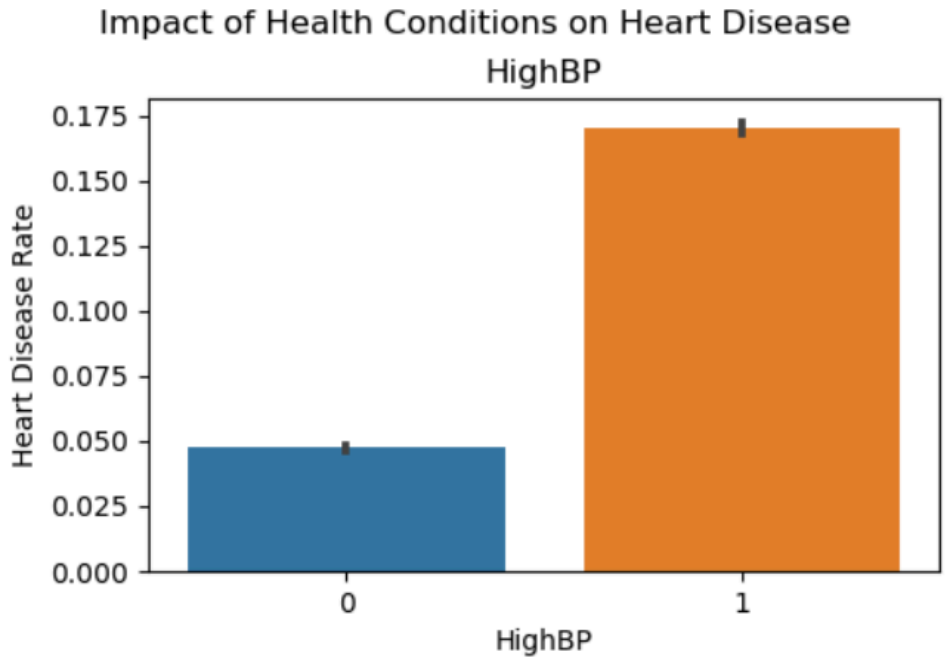
- Mean BMI for those with heart disease: 29.48
- Mean BMI for those without heart disease: 28.59

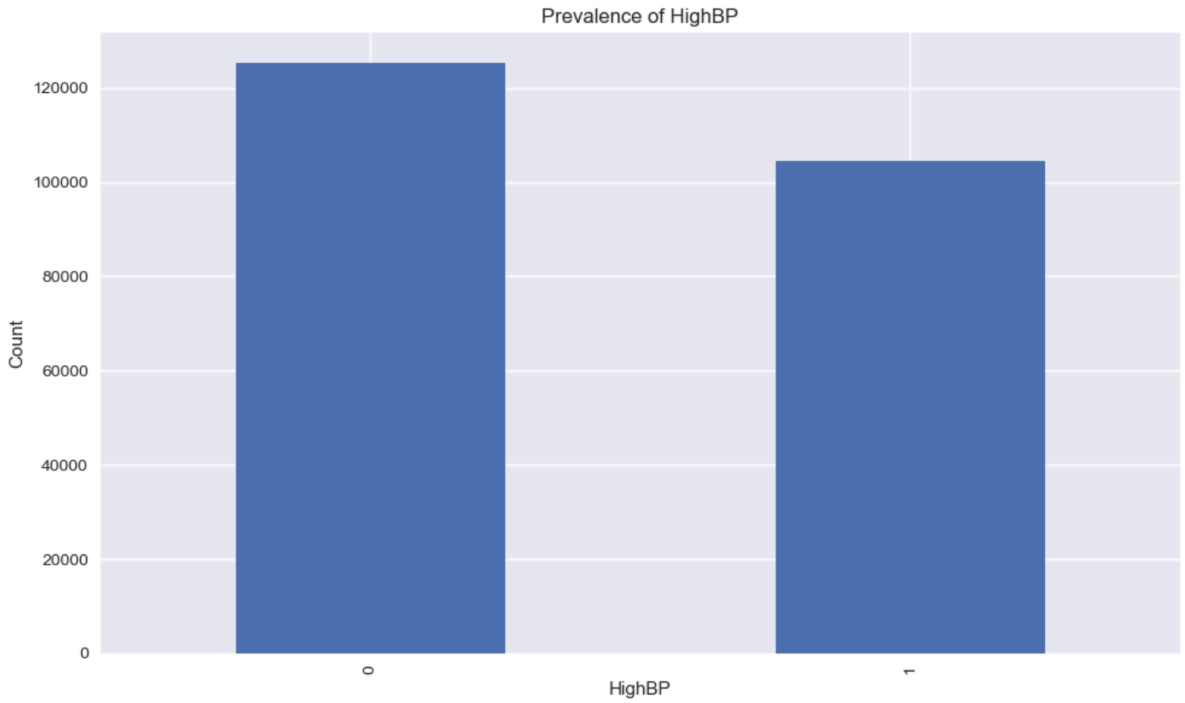


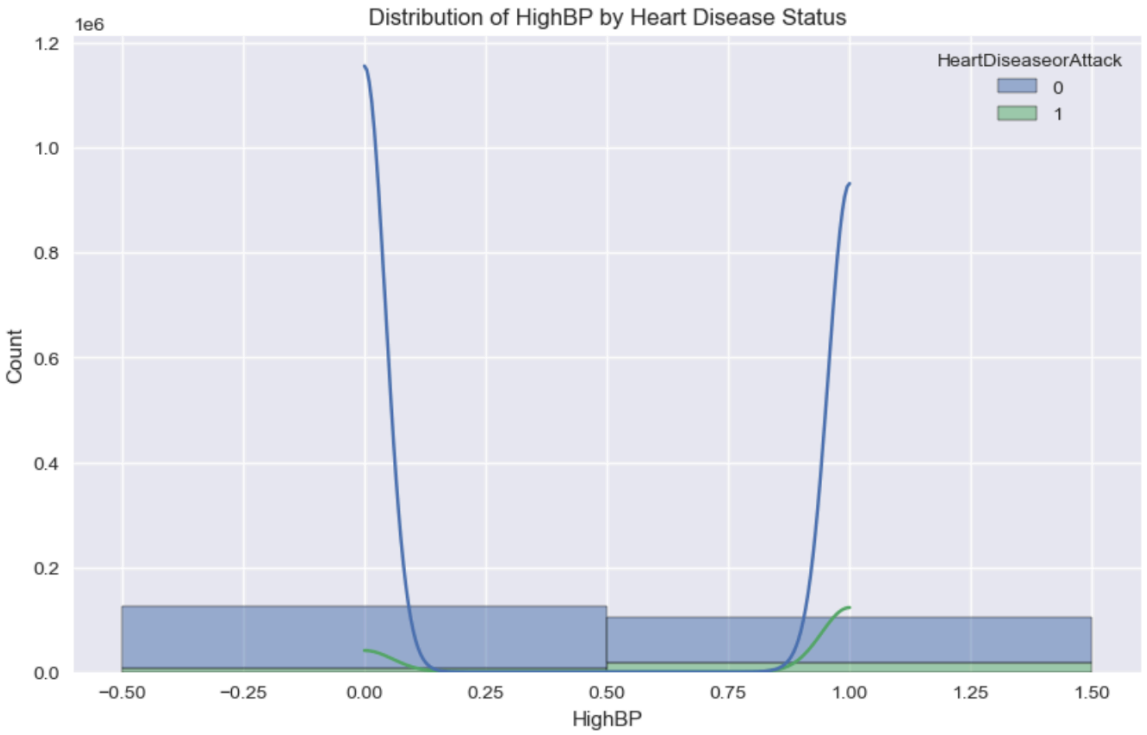
b) Blood Pressure:

- 45.44% of the population has high blood pressure
- 75% of individuals with heart disease have high blood pressure, compared to 42% without heart disease



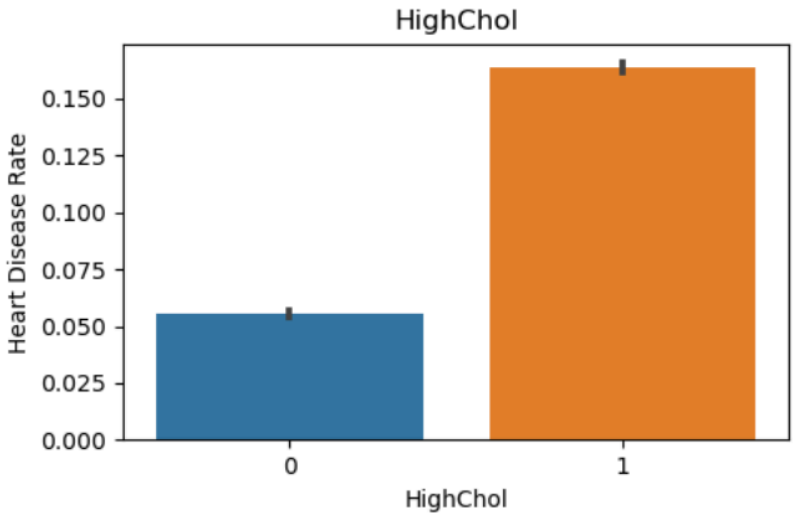
```
HighBP prevalence:
HighBP
0    0.545559
1    0.454441
Name: proportion, dtype: float64
```

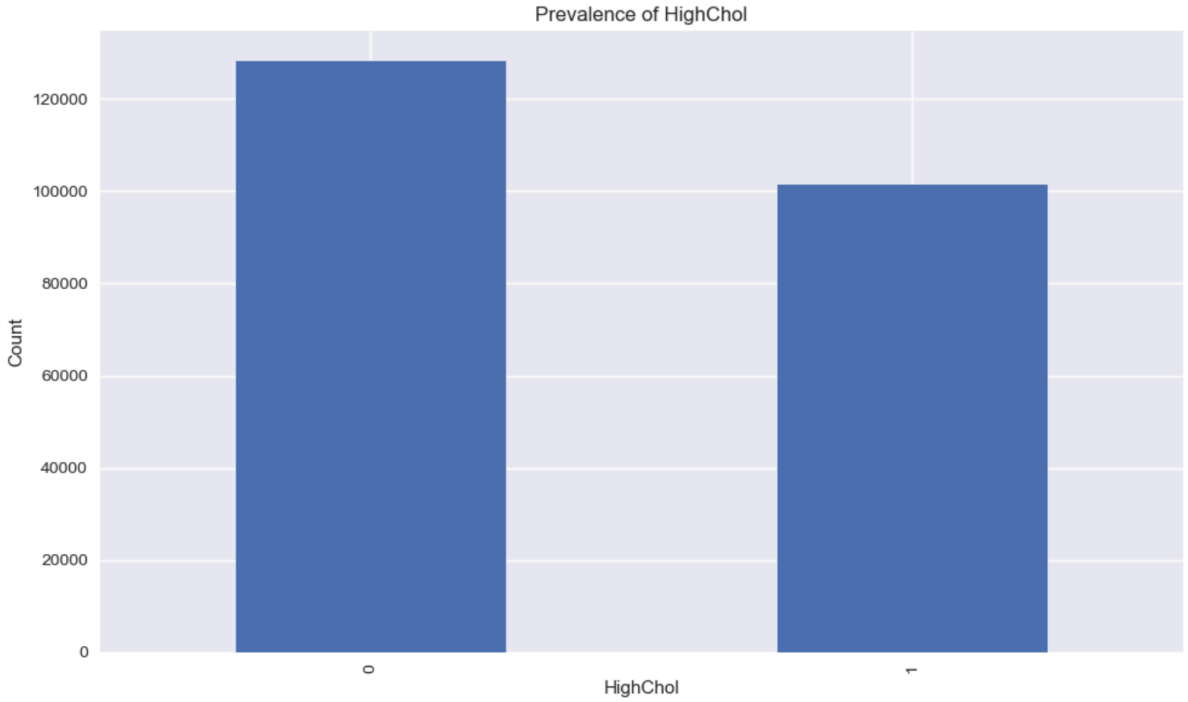



```
Summary Statistics for HighBP:
                        count  unique  top    freq
HeartDiseaseorAttack
0                      206064       2    0  119426
1                       23717       2    1   17784
```

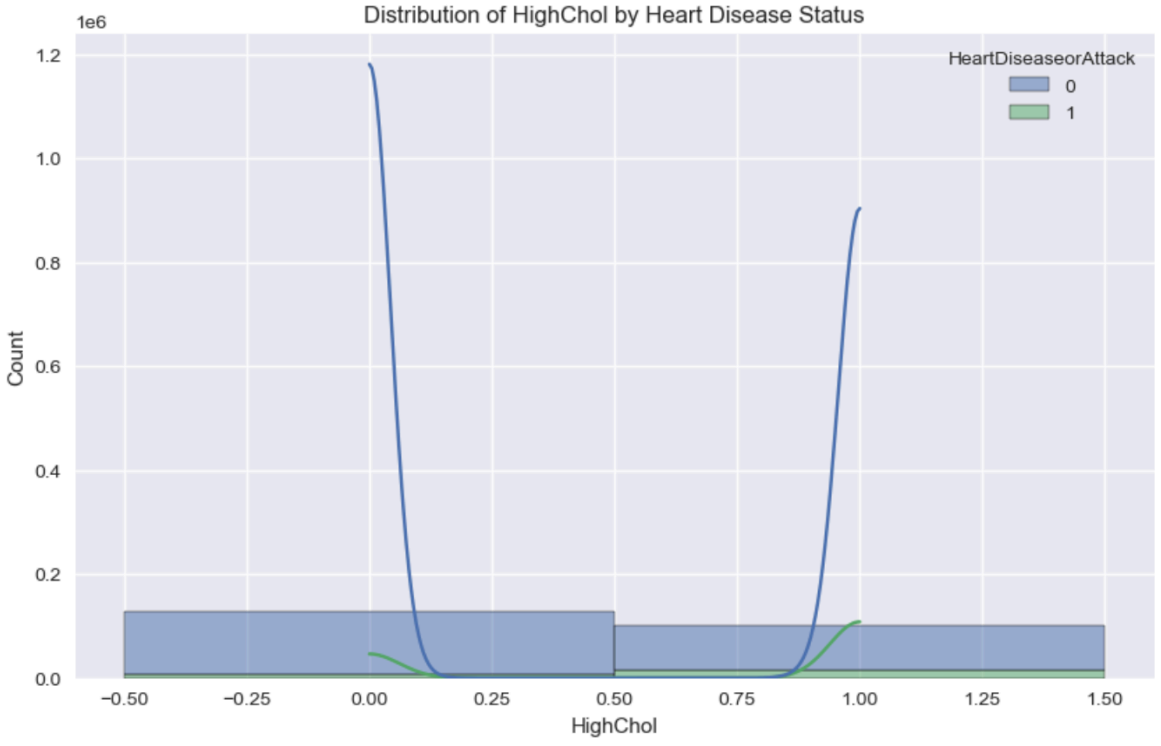
c) Cholesterol:

- 44.18% of the population has high cholesterol
- 70% of those with heart disease have high cholesterol, versus 41% without heart disease.





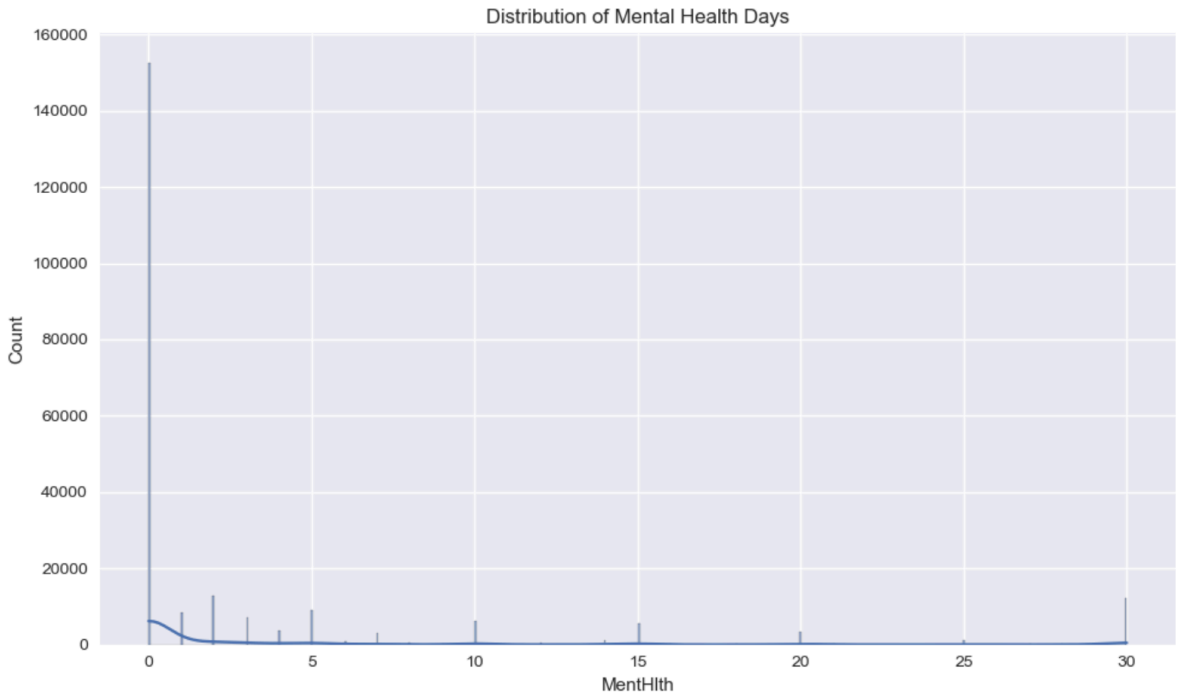

```
HighChol prevalence:
HighChol
0    0.55824
1    0.44176
Name: proportion, dtype: float64
```



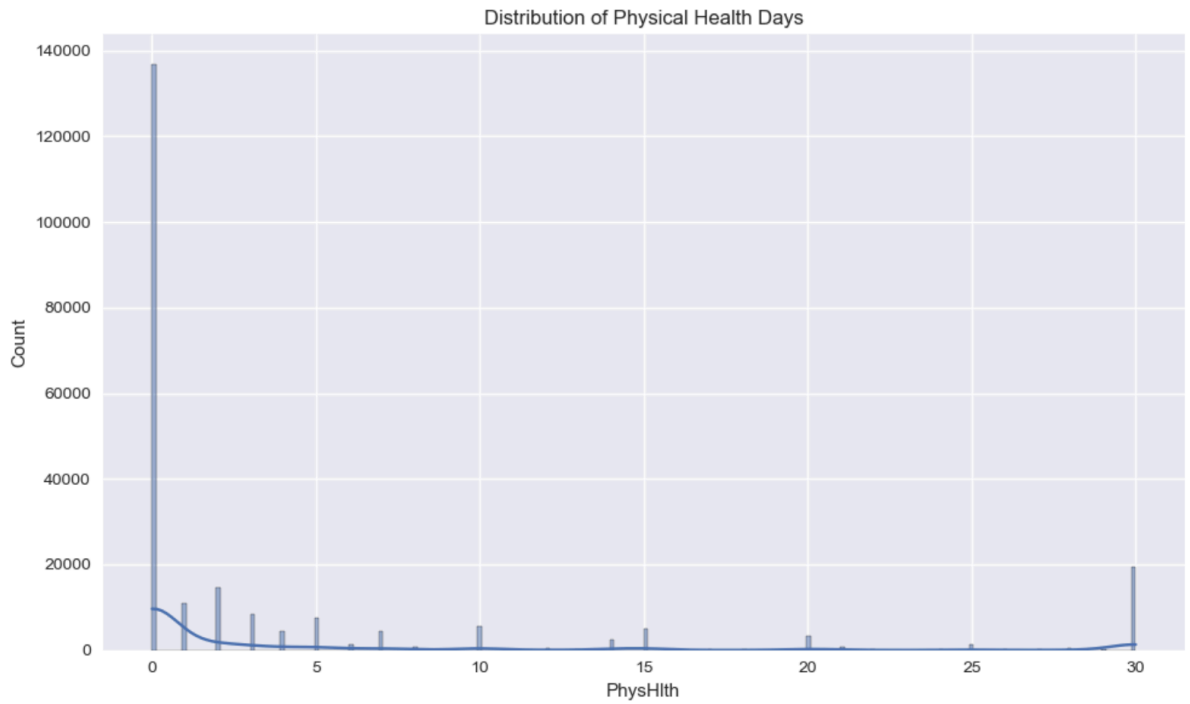
```
Summary Statistics for HighChol:
                         count  unique  top    freq
HeartDiseaseorAttack
0                       206064       2    0  121153
1                        23717       2    1   16597
```

d) Mental and Physical Health:

- Mental Health: Mean of 3.51 days, median of 0 days
- Physical Health: Mean of 4.68 days, median of 0 days
- Both show heavily right-skewed distributions.




```
Mental Health Days statistics:
count    229781.000000
mean          3.505373
std           7.713725
min           0.000000
25%           0.000000
50%           0.000000
75%           2.000000
max          30.000000
Name: MentHlth, dtype: float64
```




```
Physical Health Days statistics:
count    229781.000000
mean          4.675178
std           9.046568
min           0.000000
25%           0.000000
50%           0.000000
75%           4.000000
max          30.000000
Name: PhysHlth, dtype: float64
```

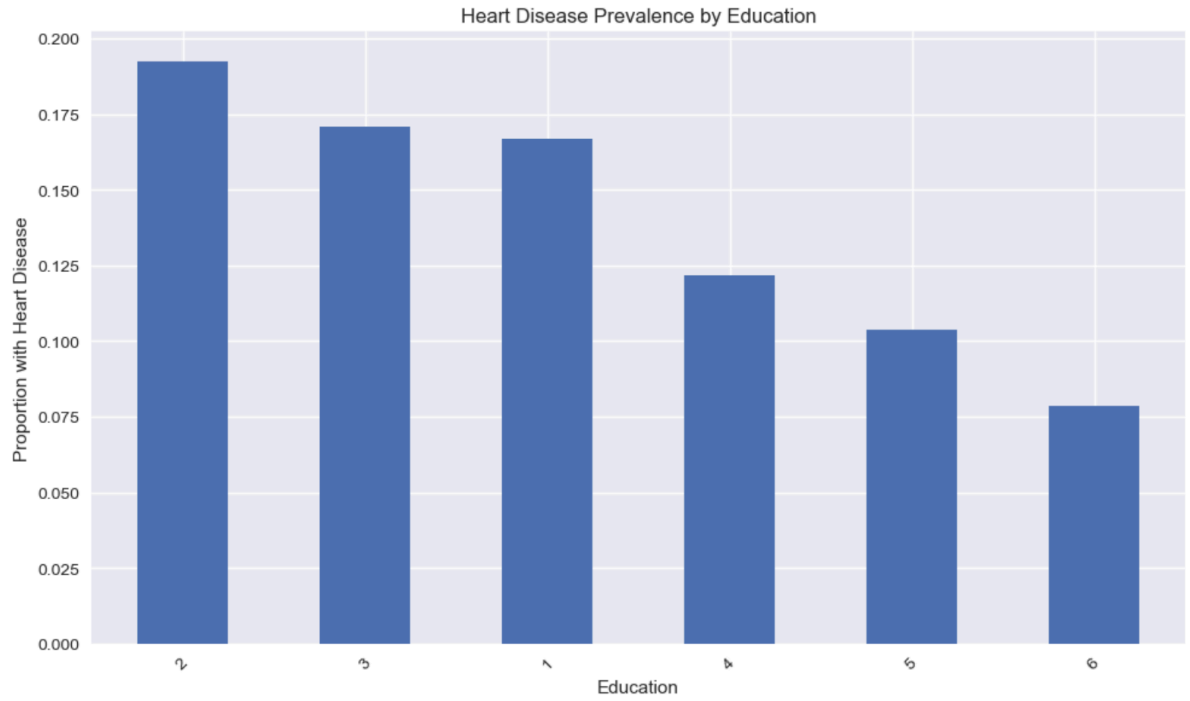
**3.4 Socioeconomic Factors:**

a) Income:

Lower income is often associated with higher risk due to various factors including access to healthcare and healthy lifestyle options

b) Education Level:

- Highest heart disease prevalence in Education group 2 (19.26%)
- Lowest in Education group 6 (7.86%)




```
Heart Disease Prevalence by Education:
Education
2    0.192574
3    0.170804
1    0.166667
4    0.121848
5    0.103611
6    0.078593
Name: HeartDiseaseorAttack, dtype: float64
```

## 4. Importance of Data-Driven Insights

Healthcare analytics provides several benefits in understanding and managing heart disease:

**1. Early Detection:** By analyzing patterns in patient data, analytics can help identify individuals at high risk before symptoms appear.

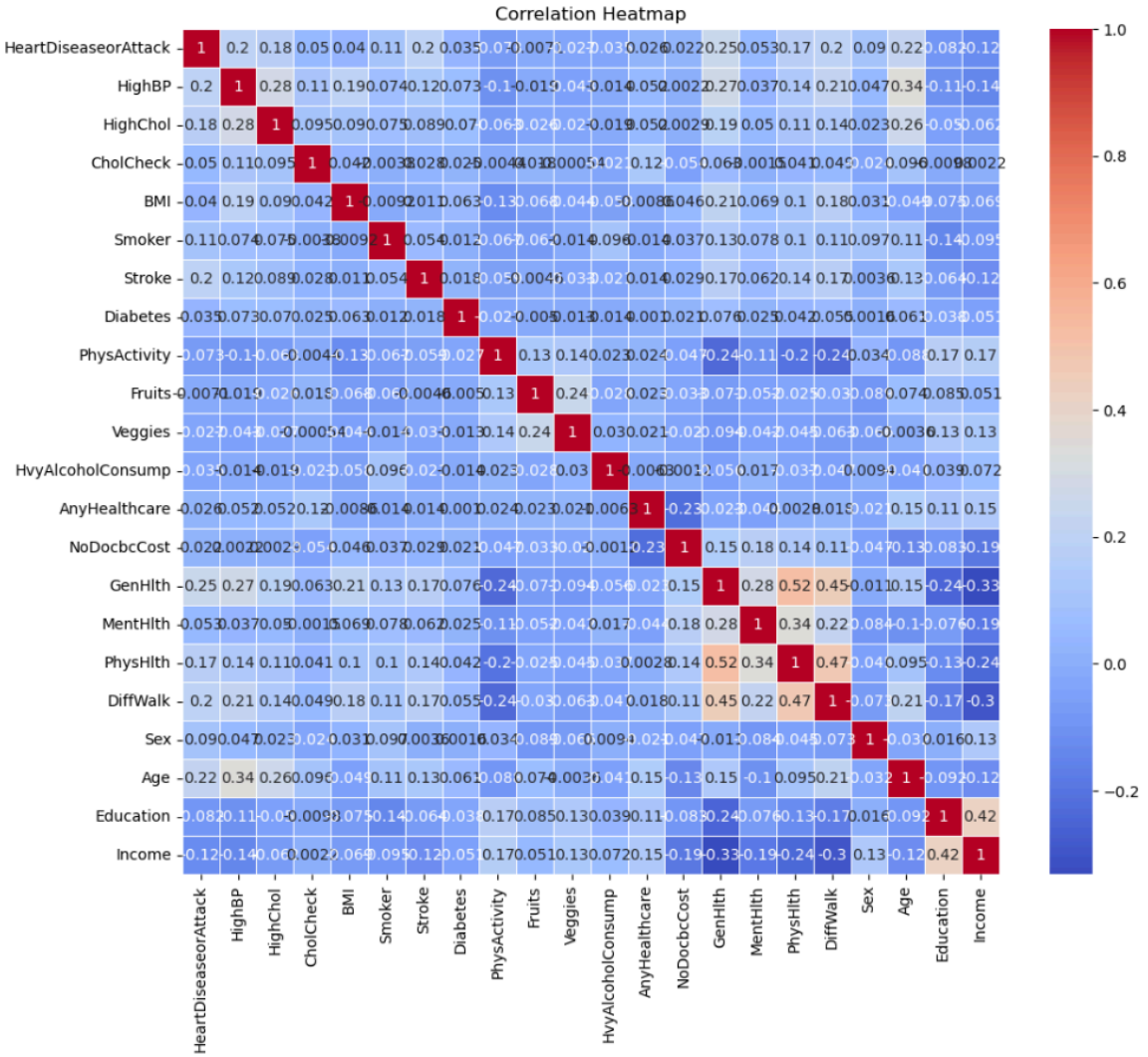
**2. Treatment Efficacy:** Analytics can assess the effectiveness of different treatments across various patient subgroups.

**3. Cost Reduction:** By focusing on prevention and early intervention, analytics can help reduce the overall cost of heart disease management.

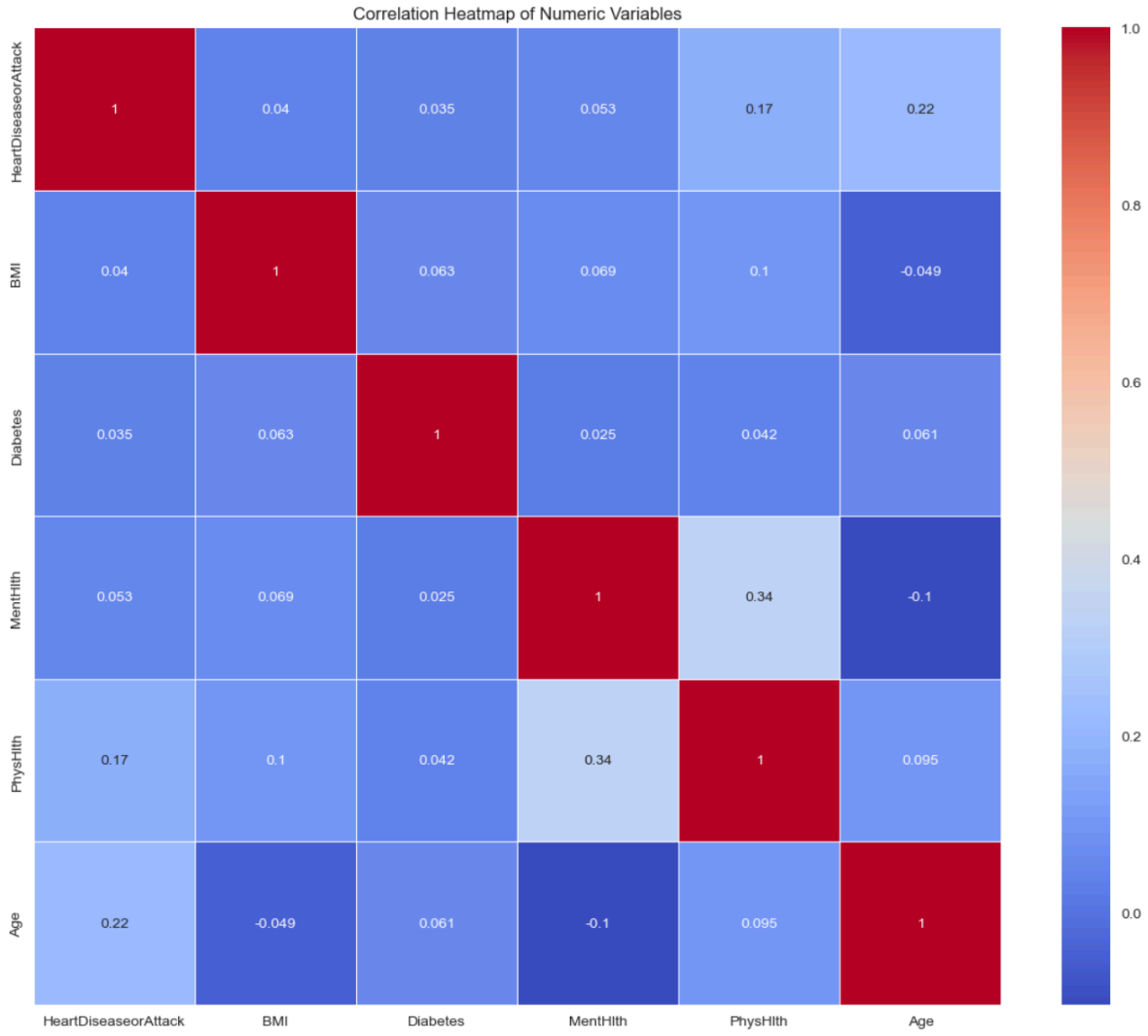
**4. Continuous Improvement:** As more data is collected and analyzed, our understanding of heart disease risk factors and effective interventions continues to improve.

## 5. Correlation Analysis

### → *Correlation Heatmap of Heart Disease Factors:*



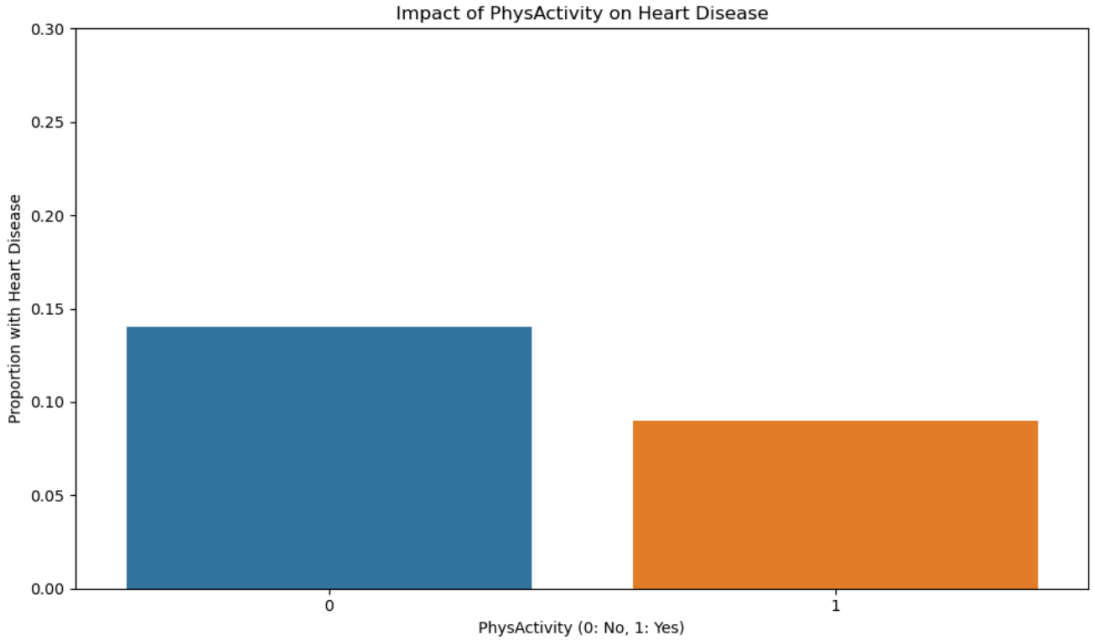

```
Top 5 factors correlated with heart disease:
GenHlth      0.246411
Age          0.223626
DiffWalk     0.202779
HighBP       0.201271
Stroke       0.198863
```



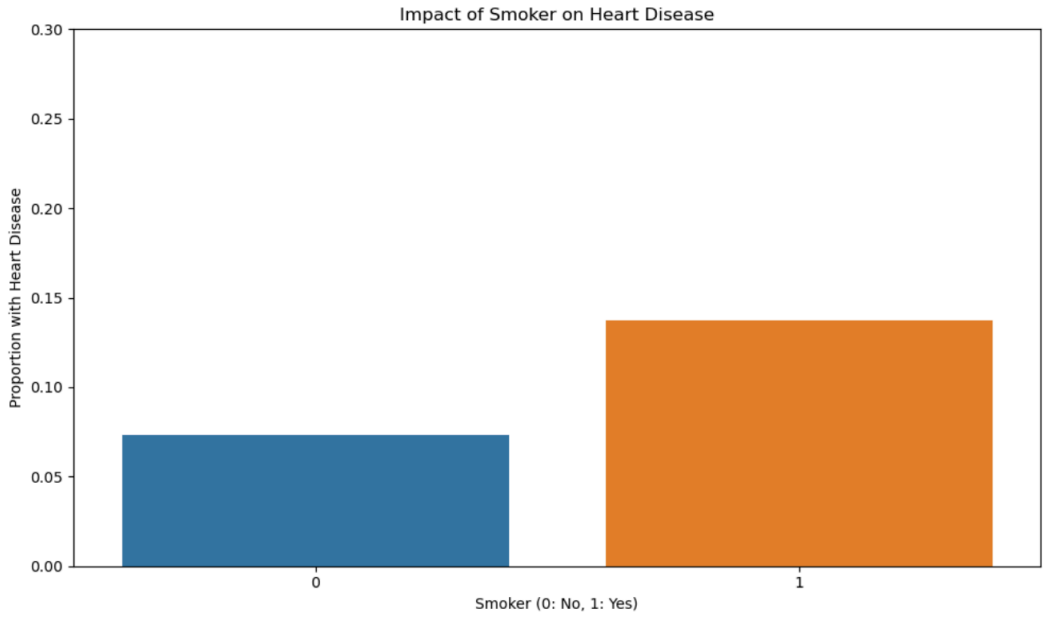

**Key correlations:**

- Strong positive correlations:
    * Age and High Blood Pressure
    * High Blood Pressure and High Cholesterol
- Moderate positive correlations:
    * Age and Heart Disease
    * High Blood Pressure and Heart Disease

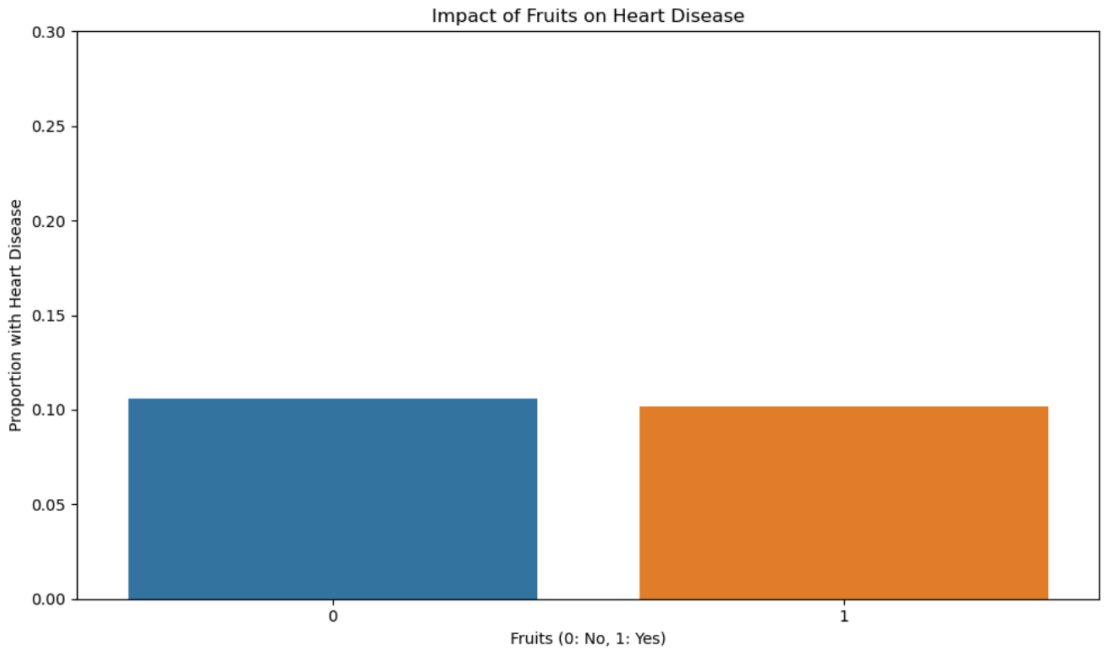
## 6. Detailed view



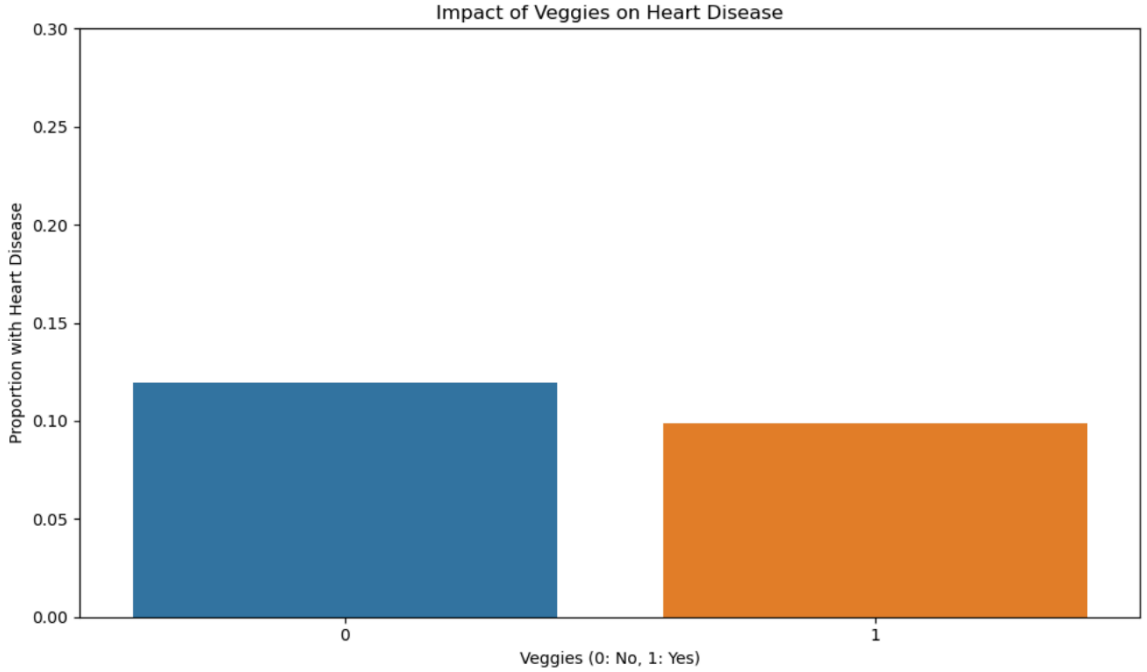

```
Data for PhysActivity:
HeartDiseaseorAttack         0         1
PhysActivity
0                     0.859817  0.140183
1                     0.910225  0.089775
Correlation with Heart Disease: -0.07326704665540902
```



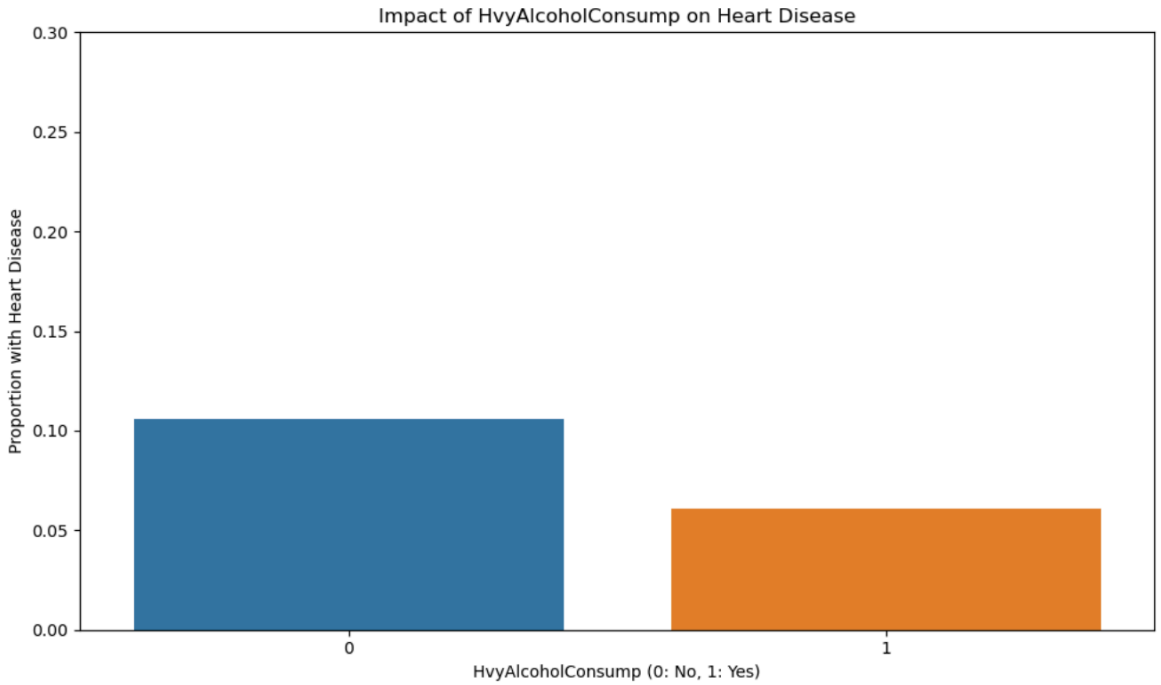

```
Data for Smoker:
HeartDiseaseorAttack         0         1
Smoker
0                     0.926650  0.073350
1                     0.862514  0.137486
Correlation with Heart Disease: 0.10515441785668599
```




```
Data for Fruits:
HeartDiseaseorAttack         0         1
Fruits
0                     0.894055  0.105945
1                     0.898508  0.101492
Correlation with Heart Disease: -0.007128256208859767
```




```
Data for Veggies:
HeartDiseaseorAttack         0         1
Veggies
0                     0.880419  0.119581
1                     0.901009  0.098991
Correlation with Heart Disease: -0.027330471912583577
```




```
Data for HvyAlcoholConsump:
HeartDiseaseorAttack         0         1
HvyAlcoholConsump
0                     0.894042  0.105958
1                     0.939211  0.060789
Correlation with Heart Disease: -0.03545333757997951
```

**6.1 Summary:**

```
Crosstab for Smoker:
HeartDiseaseorAttack         0         1
Smoker
0                     0.926650  0.073350
1                     0.862514  0.137486


Correlation with Heart Disease: 1.0

Crosstab for Veggies:
HeartDiseaseorAttack         0         1
Veggies
0                     0.880419  0.119581
1                     0.901009  0.098991
```

```
Correlation with Heart Disease: 1.0

Crosstab for HvyAlcoholConsump:
HeartDiseaseorAttack         0         1
HvyAlcoholConsump
0                     0.894042  0.105958
1                     0.939211  0.060789

Correlation with Heart Disease: 1.0
```

**6.2 Final Check:**

```
Summary of Lifestyle Factors:
                   Proportion with Factor  \
Smoker                           0.465661
Fruits                           0.612966
Veggies                          0.794813
PhysActivity                     0.733355
HvyAlcoholConsump                0.060710

                   Heart Disease Rate when Factor Present  \
Smoker                                           0.137486
Fruits                                           0.101492
Veggies                                          0.098991
PhysActivity                                     0.089775
HvyAlcoholConsump                                0.060789

                   Heart Disease Rate when Factor Absent
Smoker                                          0.073350
Fruits                                          0.105945
Veggies                                         0.119581
PhysActivity                                    0.140183
HvyAlcoholConsump                               0.105958
```

**6.3 Findings**

1. **Smoking:**
    a. 46.6% of the population are smokers.
    b. Smokers have the highest heart disease rate (13.7%), compared to non-smokers (7.3%).

   c. This suggests a strong positive association between smoking and heart disease risk.

2. **Fruit Consumption:**
   a. 61.3% of the population consume fruits regularly.
   b. Those who consume fruits have a slightly lower heart disease rate (10.1%) compared to those who don't (10.6%).
   c. The difference is small, suggesting a weak protective effect of fruit consumption.
3. **Vegetable Consumption:**
   a. 79.5% of the population consume vegetables regularly.
   b. Those who consume vegetables have a lower heart disease rate (9.9%) compared to those who don't (11.9%).
   c. This suggests a protective effect of vegetable consumption.
4. **Physical Activity:**
   a. 73.3% of the population engage in physical activity.
   b. Those who are physically active have a significantly lower heart disease rate (9.0%) compared to those who aren't (14.0%).
   c. This indicates a strong protective effect of physical activity.
5. **Heavy Alcohol Consumption:**
   a. Only 6.1% of the population are heavy alcohol consumers.
   b. Surprisingly, heavy alcohol consumers have a lower heart disease rate (6.1%) compared to non-heavy drinkers (10.6%).
   c. This counterintuitive result warrants further investigation. It could be due to confounding factors or how heavy alcohol consumption is defined in this dataset.

Our data-driven analysis has revealed several key insights into the factors associated with heart disease:

## 7. Key Insights

1. Age is a critical factor, with heart disease risk increasing substantially in older age groups.

2. High blood pressure and high cholesterol are the most significant modifiable risk factors, present in a large proportion of heart disease cases.

3. BMI shows a moderate association with heart disease risk, with higher BMI correlated with increased risk.

4. There's a notable gender disparity in heart disease risk, with one gender (likely male) showing higher prevalence.

5. Smoking appears to be the most significant lifestyle risk factor for heart disease.

6. Physical activity shows the strongest protective effect against heart disease.

7. Diet, particularly vegetable consumption, shows a protective effect against heart disease.

8. The relationship between heavy alcohol consumption and heart disease is unexpected and needs further investigation.

9. Mental and physical health issues, while not affecting the majority, show right-skewed distributions indicating significant challenges for a subset of the population.

10. Education level shows a complex relationship with heart disease risk, possibly influenced by other socioeconomic factors.

## 8. Recommendations

1. Targeted Screening: Implement more frequent cardiovascular screenings for individuals over 50, especially those with high blood pressure or cholesterol.

2. Blood Pressure and Cholesterol Management: Develop community-wide programs for regular checks and management strategies.

3. Lifestyle Intervention Programs: Focus on smoking cessation, encouraging physical activity, and promoting heart-healthy diets rich in vegetables and fruits.

4. Weight Management: Implement BMI awareness campaigns and support weight management initiatives.

5. Mental Health Support: Integrate mental health screening and support into heart disease prevention programs.

6. Gender-Specific Interventions: Develop targeted prevention strategies addressing the gender disparity in heart disease risk.

7. Education and Socioeconomic Factors: Investigate the complex relationship between education, socioeconomic status, and heart disease risk to develop targeted interventions.

## 9. Limitations and Future Work

1. The dataset may not be representative of all populations, limiting generalizability.

2. Some variables (e.g., alcohol consumption) show counterintuitive relationships that require further investigation.

3. The study is cross-sectional, limiting our ability to infer causality.

**Future research should:**

1. Conduct longitudinal studies to better understand the progression of risk factors over time.

2. Develop and validate predictive models for heart disease risk based on these findings.

3. Investigate potential interactions between risk factors for a more nuanced understanding of heart disease risk.

4. Explore the underlying causes of the observed gender disparity in heart disease risk.

5. Integrate real-time health data, genetic information, and environmental factors to create more robust predictive models.

## 10. Conclusion

This comprehensive analysis provides valuable insights into the multifaceted nature of heart disease risk. By leveraging healthcare analytics, we've identified key risk factors and protective elements, enabling more targeted and effective prevention strategies. The complex interplay of demographic, lifestyle, health, and socioeconomic factors underscores the need for personalized approaches in both prevention and treatment of heart disease.

As technology advances and more comprehensive datasets become available, the role of healthcare analytics in managing heart disease is likely to become even more significant. Our study contributes to this ongoing effort, providing a foundation for more informed decision-making in heart disease prevention and treatment.

Ultimately, the goal of healthcare analytics in heart disease management is not just to analyze data, but to translate these insights into actionable strategies that improve patient outcomes, reduce healthcare costs, and enhance overall population health. By continuing to refine our understanding through data-driven approaches, we can make significant strides in combating one of the world's leading causes of mortality.