KLE TECHNOLOGICAL UNIVERSITY
Creating Value, Leveraging Knowledge
DR. M. S. SHESHGIRI COLLEGE OF ENGINEERING AND TECHNOLOGY
Belagavi Campus

HEART DISEASE HEALTH INDICATORS

21ECSC210
Exploratory Data Analysis
Course Project: Phase - I Review

Department of Computer Science and Engineering,
KLE Technological University's Dr. M. S. Sheshgiri College of Engineering and Technology, Belagavi

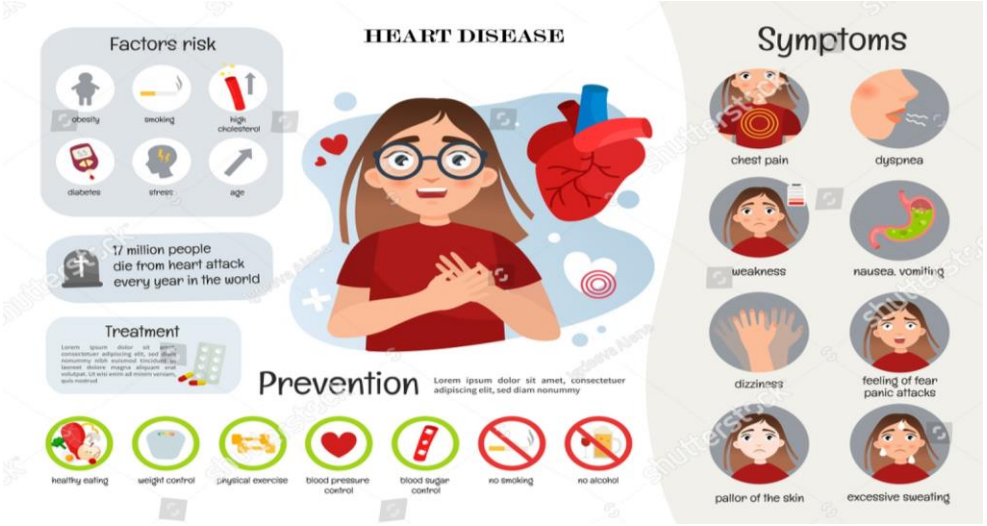
Details of the Team

Div:	

Sl. No.	Name	SRN.
1	Chandrakala B Walake	02FE22BCS027
2	Rutika Wagalekar	02FE22BCS088
3	Rohan Mushannavar	02FE23BCS423
4	Affan Mujawar	02FE22BCS008

BACKGROUND

- The 1st heart disease health indicators was likely compiled and studied in the mid 20th century.
- Heart disease annually claims approximately 647,000 lives in the United States, rendering it the leading cause of death.
- Roughly half of Americans are affected by at least one key risk factor for heart disease, such as high blood pressure, high blood cholesterol, or smoking.



PROBLEM STATEMENT

This project aims to develop a efficient model to assess the risk of heart disease based on various health indicators such as age, gender, blood pressure, cholesterol levels, and other relevant factors. Utilize a dataset containing historical patient data to train and validate the model, aiming for high accuracy and reliability in predicting the likelihood of heart disease occurrence.

Dataset Details:

There are 253680 rows and 22 columns.

SOURCE LINK OF THE DATASET:-https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset/data

```
import pandas as pd
data = pd.read_csv(r"C:\Users\CHANDRAKALA\OneDrive\Desktop\modified_dataset2
data
```

	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes
0	0	1	1	1	40	1	0	0
1	0	0	0	0	25	1	0	0
2	0	1	1	1	28	0	0	0
3	0	1	0	1	27	0	0	0

Feature Set Description:

Name Of The Feature	Number of Null Values	Number of distinct values
HeartDiseaseorAttack	0	253680.00
HighBP	0	253680.00
HighChol	0	253680.00
CholCheck	0	253680.00
BMI	0	253680.00
Smoker	0	253680.00
Stroke	0	253680.00
Diabetes	0	253680.00
PhysActivity	0	253680.00
Fruits	200	253680.00
Veggies	0	253680.00

Name Of The Feature	Number of Null Values	Number of distinct values
HvyAlcoholConsump	0	253680.00
AnyHealthcare	200	253680.00
NoDocbcCost	0	253680.00
GenHLTH	0	253680.00
MentHLTH	0	253680.00
PhysHLTH	0	253680.00
DiffWalk	0	253680.00
Sex	0	253680.00
Age	0	253680.00
Education	0	253680.00
Income	0	253680.00

Knowing the Dataset

Features of the dataset:

- 1. HighBP: Indicates if the person has been told by a health professional that they have High Blood Pressure.
- 2. HighChol: Indicates if the person has been told by a health professional that they have High Blood Cholesterol.
- 3. CholCheck: Cholesterol Check, if the person has their cholesterol levels checked within the last 5 years.
- 4. BMI: Body Mass Index, calculated by dividing the persons weight (in kilogram) by the square of their height (in meters).
- 5. Smoker: Indicates if the person has smoked at least 100 cigarettes.
- 6. Stroke: Indicates if the person has a history of stroke.
- 7. Diabetes: Indicates if the person has a history of diabetes, or currently in pre-diabetes, or suffers from either type of diabetes.
- 8. PhysActivity: Indicates if the person has some form of physical activity in their day-to-day routine.
- 9. Fruits: Indicates if the person consumes 1 or more fruit(s) daily.
- 10. Veggies: Indicates if the person consumes 1 or more vegetable(s) daily.
- 11. HvyAlcohol Consump: Indicates if the person has more than 14 drinks per week.
- 12. AnyHealthcare: Indicates if the person has any form of health insurance.

Features of the dataset:

13. NoDocbcCost: Indicates if the person wanted to visit a doctor within the past 1 year but couldn't, due to cost.

14. GenHith: Indicates the persons response to how well is their general health, ranging from 1 (excellent) to 5 (poor).

15. Menthith: Indicates the number of days, within the past 30 days that the person had bad mental health.

16. PhysHith: Indicates the number of days, within the past 30 days that the person had bad physical health.

17. DiffWalk: Indicates if the person has difficulty while walking or climbing stairs.

18. Sex: Indicates the gender of the person, where 0 is female and 1 is male.

19. Age: Indicates the age class of the person, where 1 is 18 years to 24 years up till 13 which is 80 years or older, each interval between has a 5-year increment.

20. Education: Indicates the highest year of school completed, with 0 being never attended or kindergarten only and 6 being, having attended 4 years of college or more.

Feature Set Description

Numeric Attributes:

- BMI
- Income
- education
- age
- GenHlth
- MentHlth
- PhysHlth

Feature Set Description

Binary Attributes:

- HeartDiseaseorAttack
- HighBP
- HighChol
- CholCheck
- Smoker
- Stroke
- Diabetes
- PhysActivity
- Fruits
- Veggies
- HvyAlcoholConsump
- AnyHealthcare
- NoDocbcCost
- DiffWalk
- Sex

SDG 3:

The Sustainable Development Goals (**SDGs**), adopted by the United Nations in 2015, provide a comprehensive framework to address global challenges, including health-related issues. A heart disease project aligns primarily with SDG 3: Good Health and Well-Being, particularly with Target 3.4, which aims to reduce by one third premature mortality from non-communicable diseases (**NCDs**) by 2030 through prevention, treatment, and promotion of mental health and well-being.

Domain Understanding:

Domain understanding in the context of a project focusing on heart disease and health indicators involves a comprehensive grasp of the following key aspects:

1. Epidemiology of Heart Disease: Understanding the prevalence, incidence, risk factors, and outcomes associated with heart disease. This includes knowledge of demographic trends, such as how age, gender, ethnicity, and socioeconomic factors influence heart disease risk.

2. Risk Factors: Awareness of the major modifiable and non-modifiable risk factors for heart disease. Modifiable factors include lifestyle choices (smoking, diet, physical activity) and conditions like hypertension and high cholesterol. Non-modifiable factors include age, family history, and genetics.

3. Public Health Policies and Interventions: Familiarity with public health strategies aimed at preventing and managing heart disease. This includes policies promoting healthy lifestyles, screening programs, educational campaigns, and healthcare initiatives targeting high-risk populations.

4. Healthcare Systems and Data Sources: Knowledge of healthcare systems and data sources used for monitoring and assessing heart disease trends. This includes understanding data collection methods (such as surveys like BRFSS), electronic health records, and national health databases.

5. Clinical Aspects: Basic understanding of cardiac anatomy, physiology, and pathophysiology related to heart disease. This includes knowledge of diagnostic procedures, treatment options (medications, surgery, interventions), and patient management strategies.

6. Impact and Burden: Awareness of the socioeconomic impact and burden of heart disease on individuals, families, healthcare systems, and society at large. This encompasses healthcare costs, productivity loss, and quality of life implications for affected individuals.

7. Research and Innovations: Stay informed about current research trends, advancements in cardiac care, and emerging technologies (like AI and machine learning) used in predictive analytics, personalized medicine, and precision health approaches for heart disease prevention and management..

Data Preprocessing:

Handling Null Values:

Columns with Null Values and their counts:

- HeartDiseaseorAttack 0
- HighBP 0
- HighChol 0
- CholCheck 0
- BMI 0
- Smoker 0
- Stroke 0
- Diabetes 0
- PhysActivity 0
- Fruits 200
- Veggies 0
- HvyAlcoholConsump 0
- AnyHealthcare 200
- NoDocbcCost 0
- GenHlth 0
- MentHlth 0
- PhysHlth 0
- DiffWalk 0
- Sex 0
- Age 0
- Education 0
- Income 0

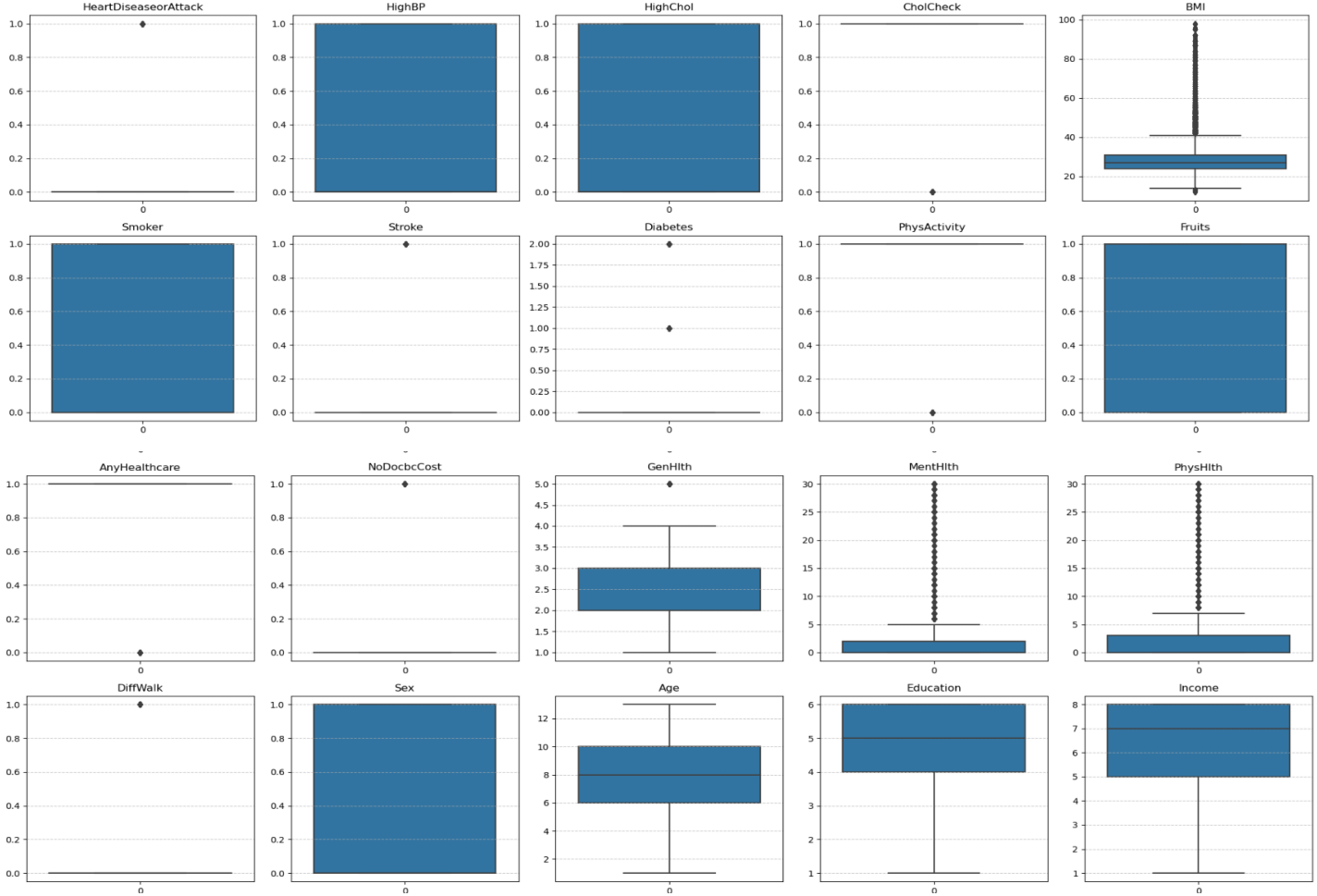
Detecting Outliers:

Detecting outliers involves identifying observations in a dataset that significantly deviate from the rest of the data. Outliers can occur due to various reasons such as data entry errors, measurement errors, or genuine anomalies in the data.

	Attribute	Outlier Count
0.	HeartDiseaseorAttack	23893
1.	HighBP	0
2.	HighChol	0
3.	CholCheck	9470
4.	BMI	9847
5.	Smoker	0
6.	Stroke	10292
7.	Diabetes	39977
8.	PhysActivity	61760
9.	Fruits	0
10.	AnyHealthcare	12407

Attributes	Outliers count
11. NoDocbcCost	21354
12. GenHlth	12081
13. MentHlth	36208
14. PhysHlth	40949
15. DiffWalk	42675
16. Sex	0
17. Age	0
18. Education	0
19. Income	0

Outliers representation using box plot:



Data Cleaning:

Removing all the NULL values:

- HeartDiseaseorAttack 0
- HighBP 0
- HighChol 0
- CholCheck 0
- BMI 0
- Smoker 0
- Stroke 0
- Diabetes 0
- PhysActivity 0
- Fruits 0
- AnyHealthcare 0
- NoDocbcCost 0

Removing all the NULL values:

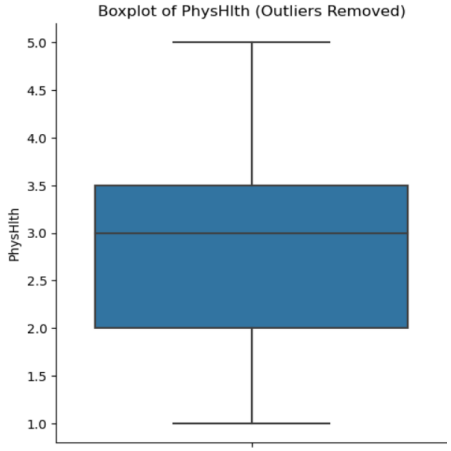
- GenHlth 0
- MentHlth 0
- PhysHlth 0
- DiffWalk 0
- Sex 0
- Age 0
- Education 0
- Income 0
- dtype: int64

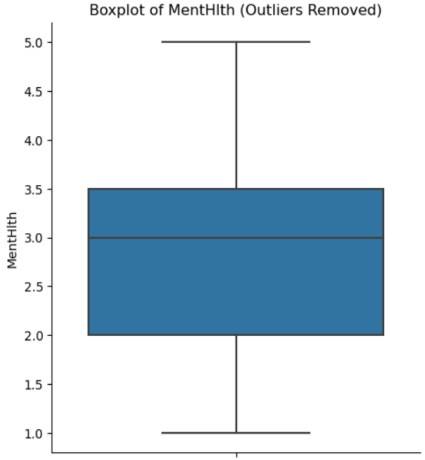
Removing all the outliers:

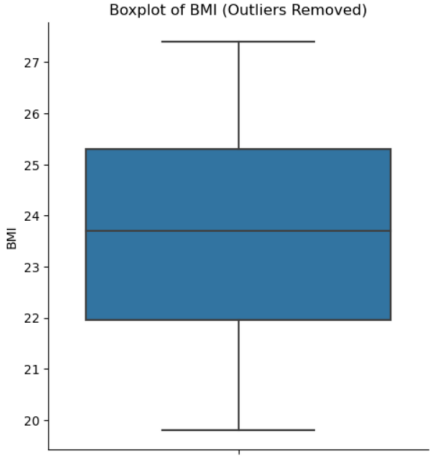
The method used to remove outliers is based on the z-score method.

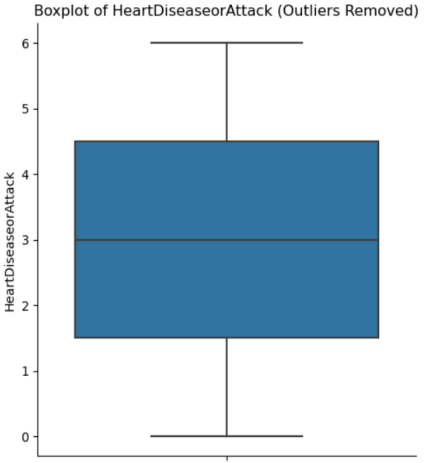
	Attribute	Outlier Count
0	HeartDiseaseorAttack	0
1	HighBP	0
2	HighChol	0
3	CholCheck	0
4	BMI	0
5	Smoker	0
6	Stroke	0
7	Diabetes	0
8	PhysActivity	0
9	Fruits	0
10	AnyHealthcare	0
11	NoDocbcCost	0
12	GenHlth	0
13	MentHlth	0
14	PhysHlth	0
15	DiffWalk	0
16	Sex	0
17	Age	0
18	Education	0
19	Income	0

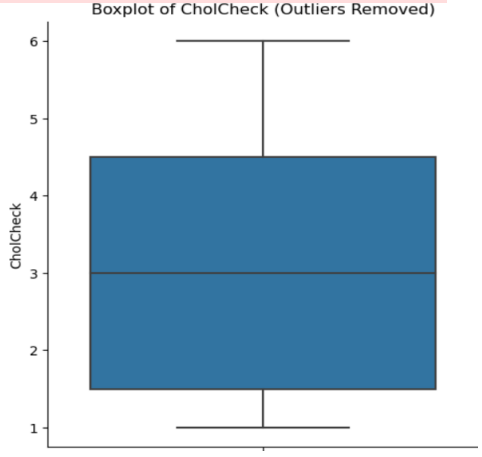
Representation of removed outliers using Box plot:

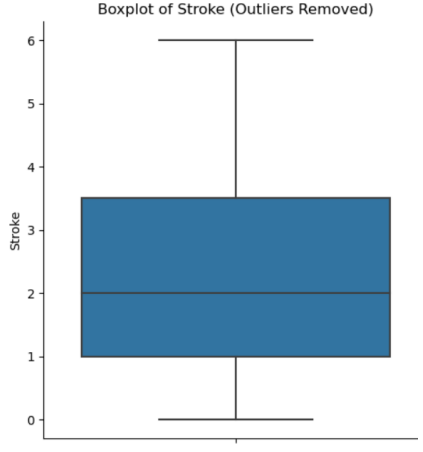


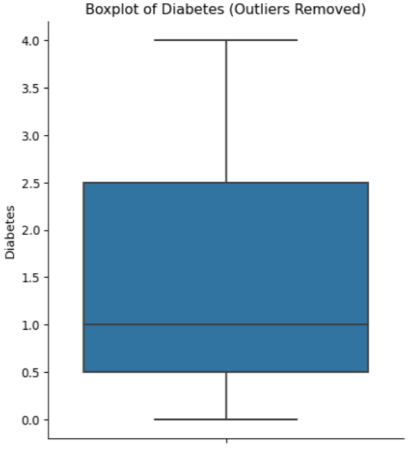


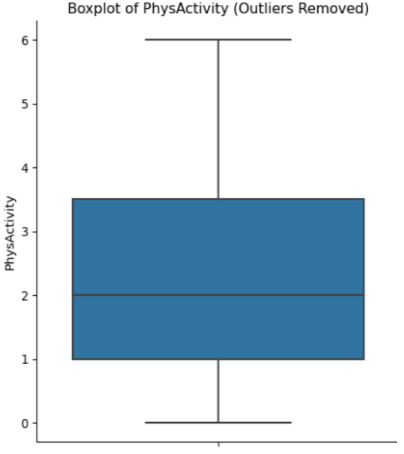




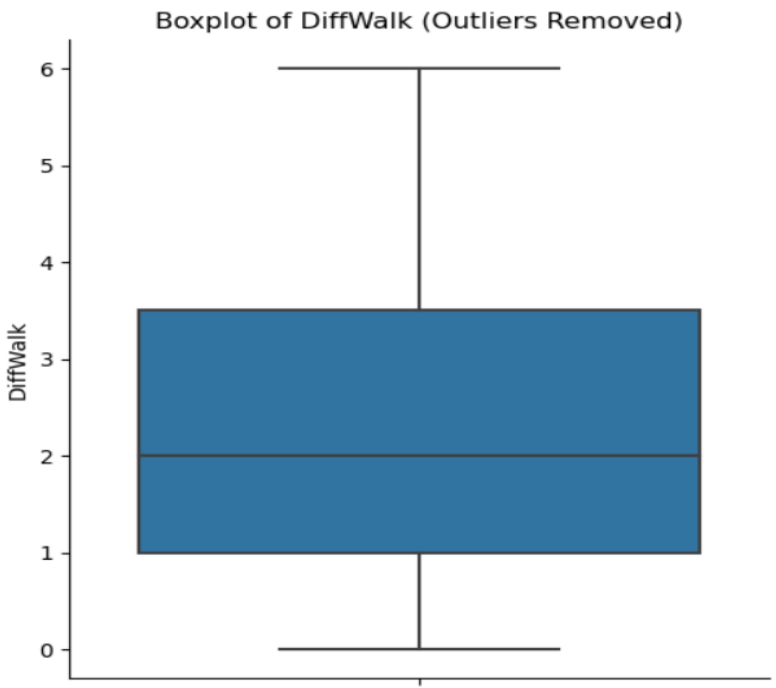


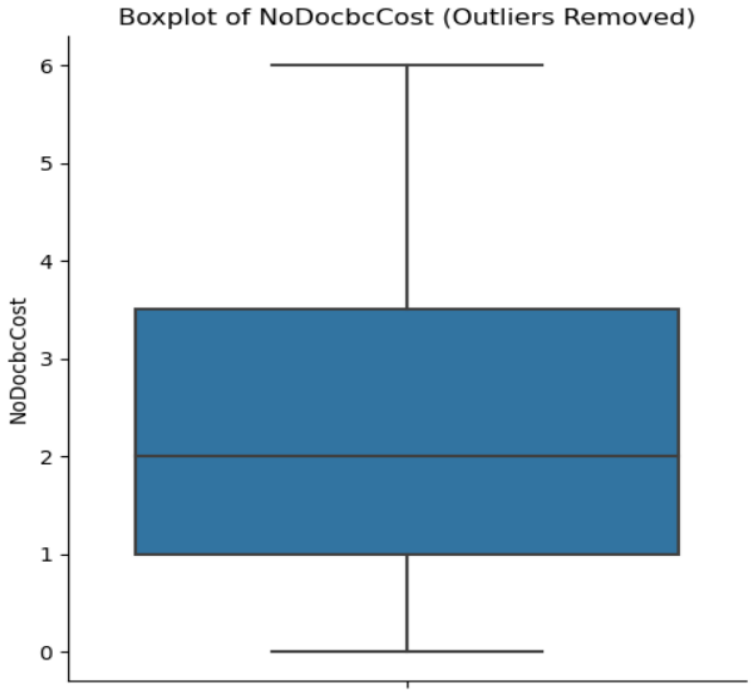






Representation of removed outliers using Box plot:





Skewness:

Finding the skewness of all the features:

Skewness measures the degree and direction of asymmetry in a distribution. A distribution is symmetric if its left and right tails are mirror images of each other. Skewness measures how much a distribution deviates from this symmetry.

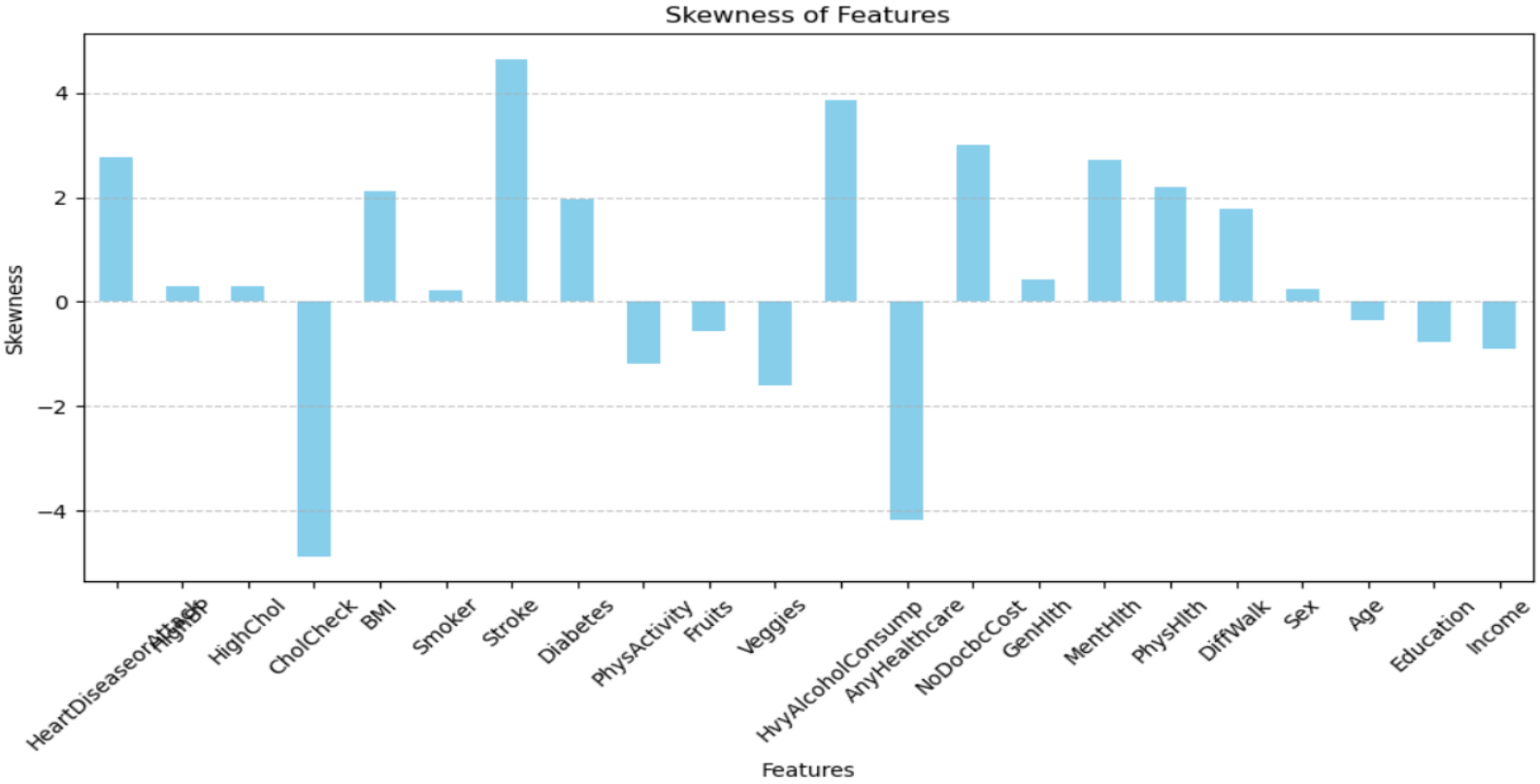
Skewness of each numeric feature:

Attributes	Skewness
HeartDiseaseorAttack	2.778742
HighBP	0.286904
HighChol	0.307075
CholCheck	-4.881271
BMI	2.122004
Smoker	0.228810
Stroke	4.657340
Diabetes	1.976390
PhysActivity	-1.195546
Fruits	-0.557515
Veggies	-1.592239

Finding the skewness of all the features:

- HvyAlcoholConsump 3.854132
- AnyHealthcare -4.181157
- NoDocbcCost 2.995290
- GenHlth 0.422867
- MentHlth 2.721148
- PhysHlth 2.207395
- DiffWalk 1.773907
- Sex 0.240350
- Age -0.359903
- Education -0.777255
- Income -0.891345
- dtype: float64

Graphical Representation:

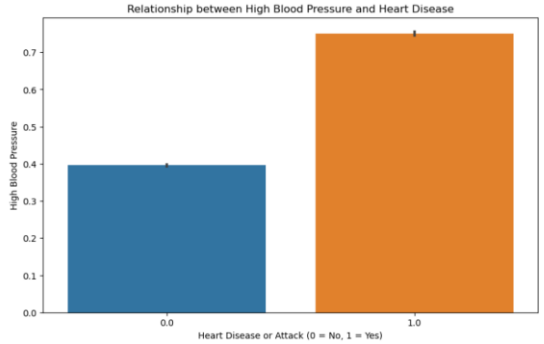


Implement Framework

- **Data Cleaning :** Remove irrelevant features and handle missing values.
- **Data Distribution Analysis :** Explore the distribution of physiological and psychological indicators.
- **Statistical Data Analysis :** Identify outliers and measure central tendency , range for features and we will perform data pre-processing.
- **Feature category:** We will be then perform univariate and multivariate analysis on the features to analyse and answer the hypothesis.
- **Feature Engineering :** Create new features and find correlation between
- between different features..
- **Machine Learning Modeling :** Select appropriate classification algorithms
- For heart disease health indicators.

Proposed Hypothesis

1: People with High Blood Pressure are more likely to have Heart Disease



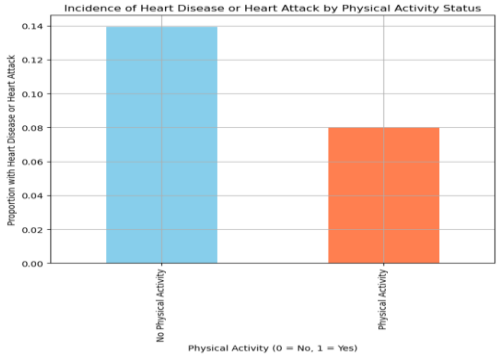
HighBP: A binary feature indicating if an individual has high blood pressure (1) or not (0).

HeartDiseaseorAttack: The target variable indicating if an individual has heart disease or has had a heart attack (1) or not (0).

This plot shows the average occurrence of high blood pressure in individuals with and without heart disease. If individuals with heart disease have a significantly higher average of high blood pressure, it supports the hypothesis.

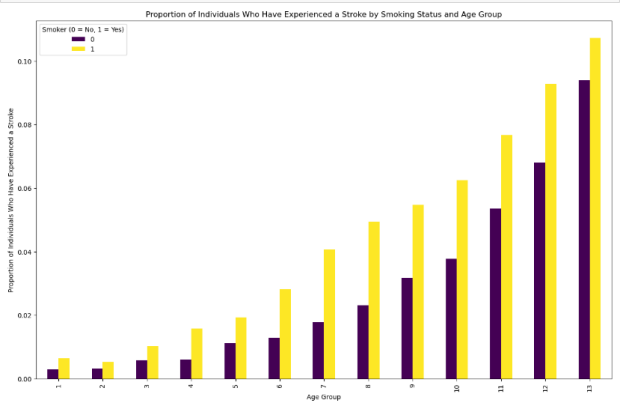
Proposed Hypothesis

2. Is there a difference in the incidence of heart disease or heart attack between individuals who engage in regular physical activity and those who don't?



Proposed Hypothesis

3.Are smokers more likely to have experienced a stroke, and does this relationship vary by age group?



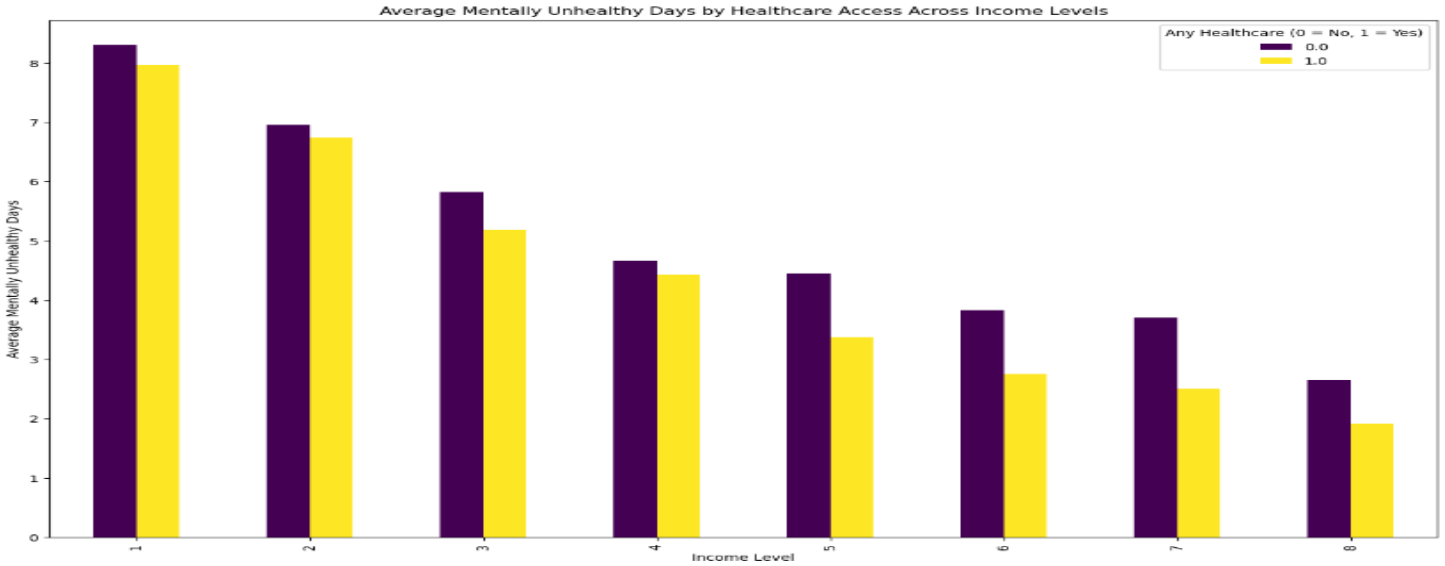
Age Groups 1-4 (Younger Age Groups): In these age groups, the proportion of individuals who have experienced a stroke is relatively low for both smokers and non-smokers. However, smokers still show a slightly higher likelihood of having had a stroke.

Age Groups 5-8 (Middle Age Groups): The difference in stroke prevalence between smokers and non-smokers becomes more apparent. Smokers in these age groups have a noticeably higher proportion of strokes compared to non-smokers.

Age Groups 9-12 (Older Age Groups): The trend is most pronounced in the older age groups. Smokers in these groups have a significantly higher likelihood of having experienced a stroke compared to their non-smoking counterparts.

Proposed Hypothesis

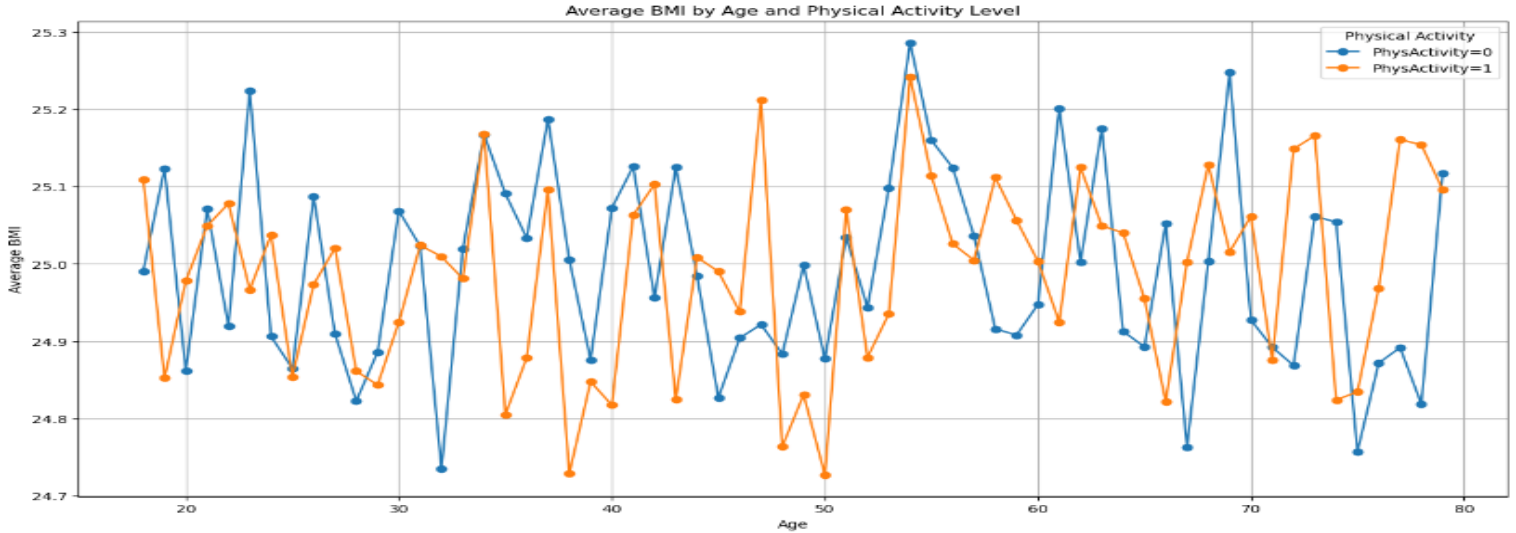
4. How the average number of mentally unhealthy days by income level and healthcare coverage using a DataFrame?



The visualization of average mentally unhealthy days by income level and healthcare coverage reveals significant insights into mental health disparities. Generally, individuals with higher income levels experience fewer mentally unhealthy days, highlighting the positive correlation between higher income and better mental health. Additionally, having healthcare coverage is associated with fewer mentally unhealthy days across almost all income levels, underscoring the critical role of healthcare access in improving mental health outcomes. This suggests that policies aimed at increasing income and expanding healthcare coverage could be effective in reducing mental health issues.

Proposed Hypothesis

5.How does the average body mass index (BMI) vary with age, stratified by physical activity level?



Interpretation:

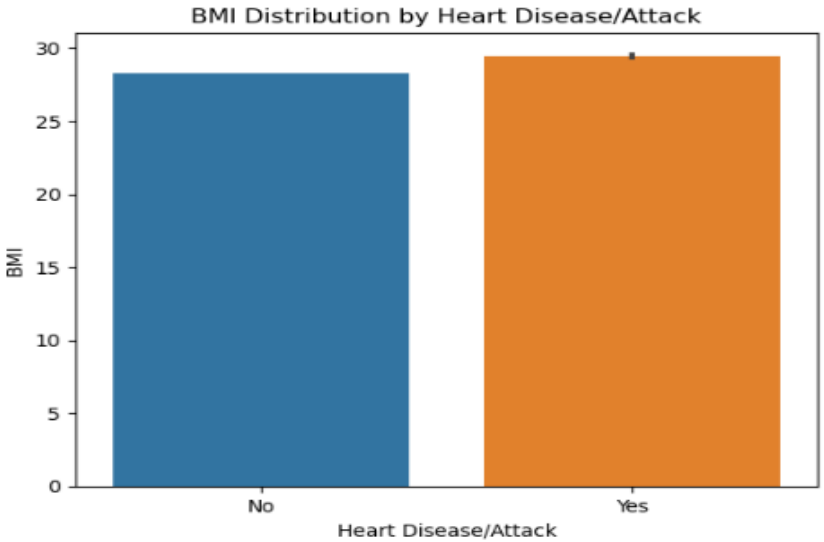
Trend by Age: The plot allows us to observe how average BMI changes across different age groups. Impact of Physical Activity: Comparing the lines for different physical activity levels shows whether and how physical activity influences BMI at different ages. Insights into Health Patterns: Higher BMI averages may indicate potential health risks associated with age and physical activity levels.

Implications:

Understanding these patterns can inform health interventions and policies aimed at promoting physical activity to manage BMI and potentially reduce health risks. Monitoring BMI across age groups helps in identifying trends and patterns that may require targeted health interventions or support.

Proposed Hypothesis

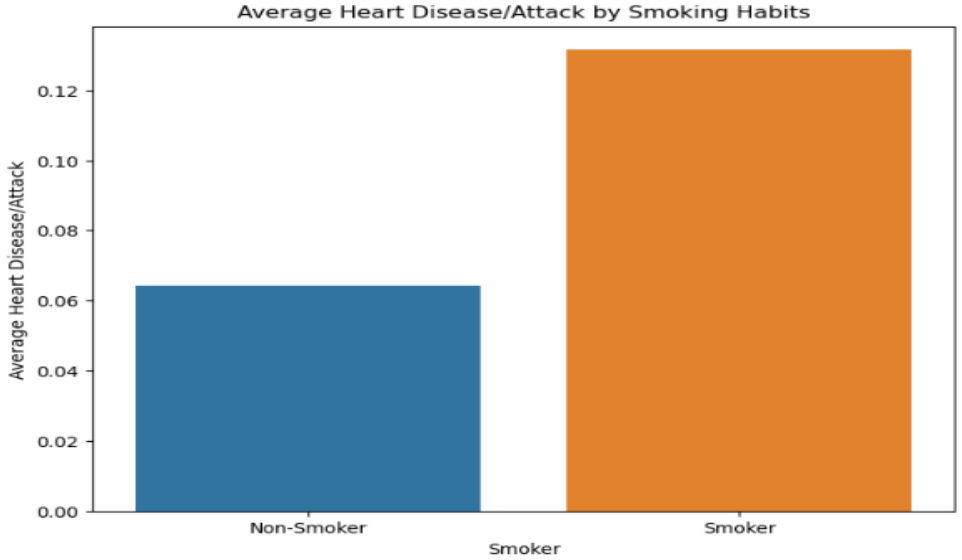
6.Is there a correlation between body mass index (BMI) and the occurrence of heart disease or heart attack?



The resulting bar plot will display the distribution of BMI for individuals with and without heart disease or heart attack. Observing any differences in the distributions can give insights into the correlation between BMI and the occurrence of heart disease or heart attack. If there's a notable difference, it may suggest a correlation between higher BMI and a higher risk of heart disease or heart attack.

Proposed Hypothesis

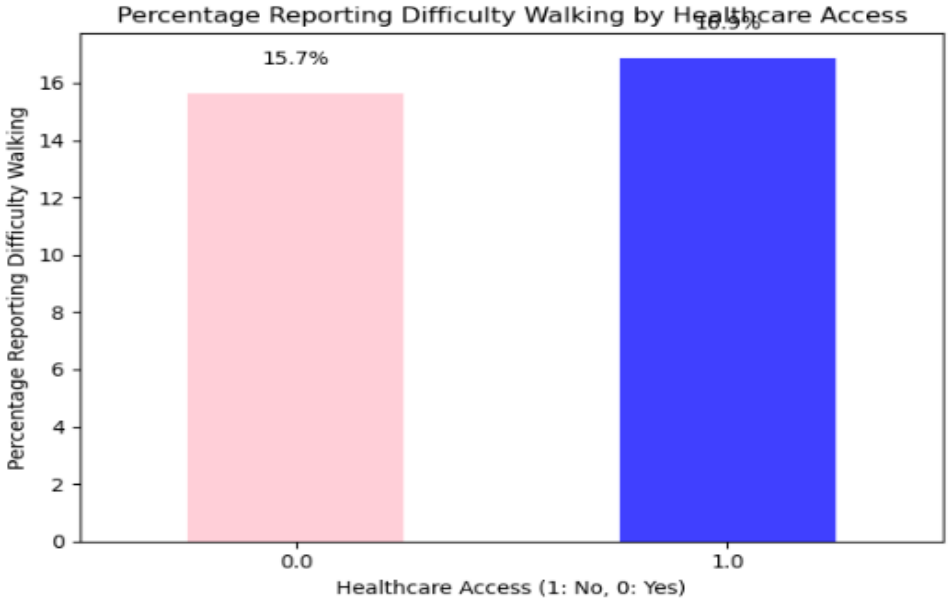
7.Is there a relationship between smoking habits and the likelihood of having heart disease or a heart attack?



This code will provide us with insights into the relationship between smoking habits and the likelihood of having heart disease or a heart attack. The visualization will help in understanding the average heart disease/attack among smokers and non-smokers, answering our hypothesis question.

Proposed Hypothesis

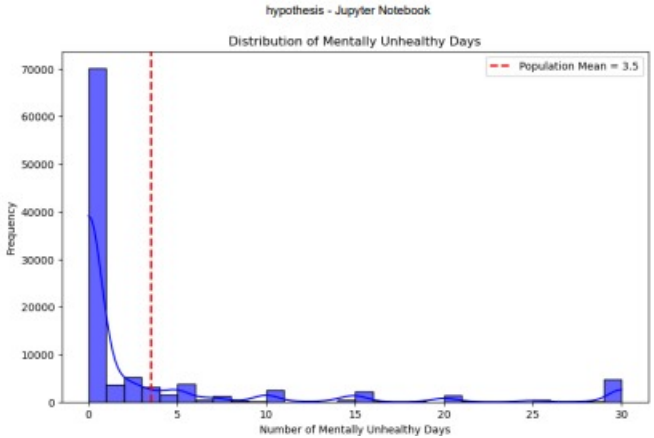
8.Individuals with healthcare access are less likely to report difficulty walking (DiffWalk).



Calculate the percentage of individuals reporting difficulty walking (DiffWalk = 1) in both groups (with and without healthcare access). A lower percentage in the healthcare access group would support the hypothesis that healthcare access correlates with fewer mobility issues.

Proposed Hypothesis

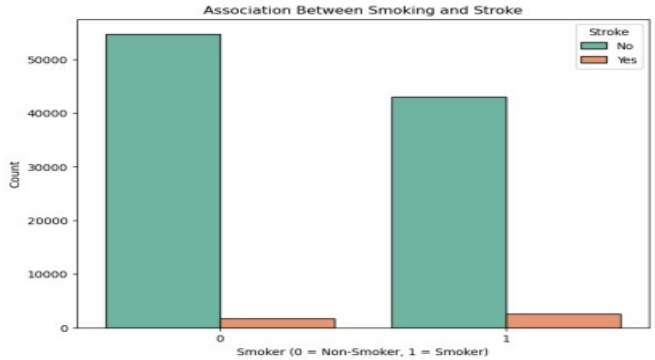
9.The average number of mentally unhealthy days (MentHlth) reported by individuals in the dataset is significantly different from the population average



Inference: The hypothesis suggests that the average number of mentally unhealthy days reported by individuals in the dataset differs significantly from the population average, potentially indicating unique mental health characteristics or conditions among the dataset's sample. Further analysis aims to confirm this difference and its implications for mental health interventions.

Proposed Hypothesis

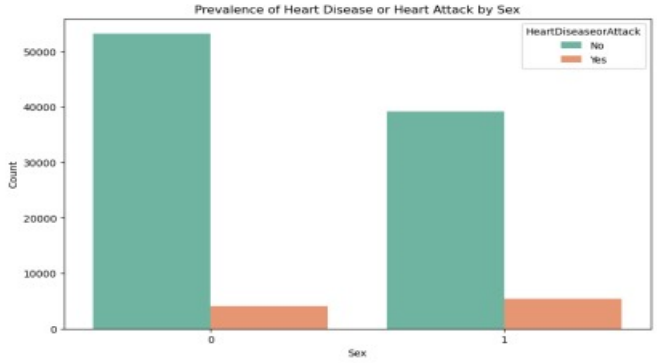
- 10.There is a significant association between smoking status (Smoker) and the occurrence of stroke (Stroke).



Inference: Smoking significantly increases the likelihood of stroke occurrence, highlighting smoking cessation as crucial for reducing stroke risk. This association underscores the importance of targeted public health efforts to promote smoking cessation and improve cardiovascular health outcomes.

Proposed Hypothesis

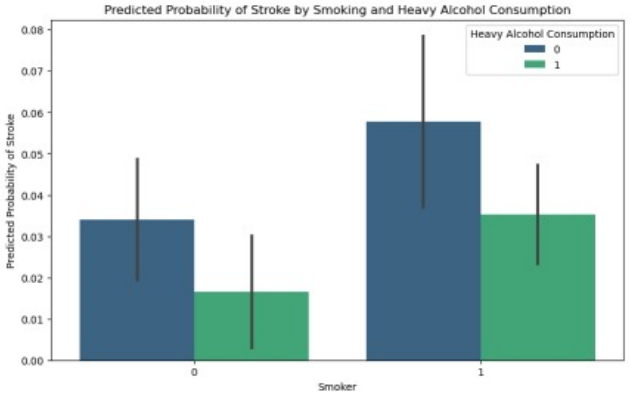
- 11.There is a significant difference in the prevalence of HeartDiseaseorAttack between males and females.



Inference: Men and women show statistically significant differences in heart disease prevalence, highlighting gender's role in heart health disparities and necessitating gender-specific health interventions.

Proposed Hypothesis

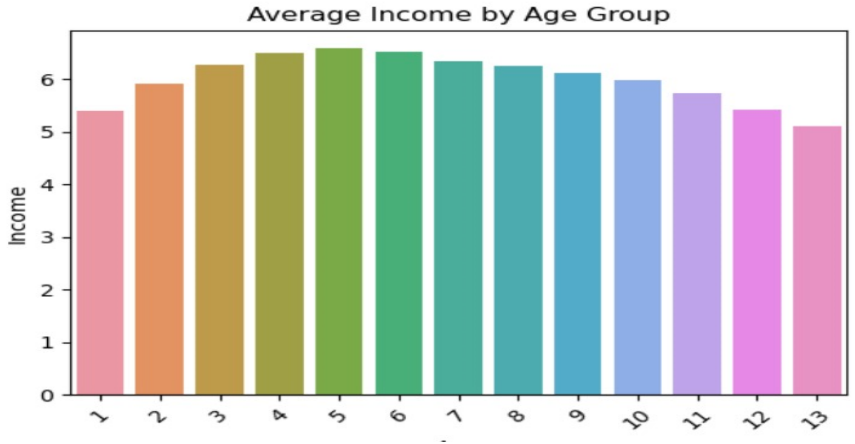
- 12.Smoking and heavy alcohol consumption (HvyAlcoholConsump) collectively increase the risk of stroke, and this relationship is moderated by cholesterol levels (HighChol). Analysis: Investigate the combined effect of smoking and heavy alcohol consumption on stroke risk, considering cholesterol levels as a moderating factor.



Inference: Smoking and heavy alcohol consumption synergistically increase stroke risk, with cholesterol levels moderating this relationship, suggesting higher risks for individuals with elevated cholesterol.

Proposed Hypothesis

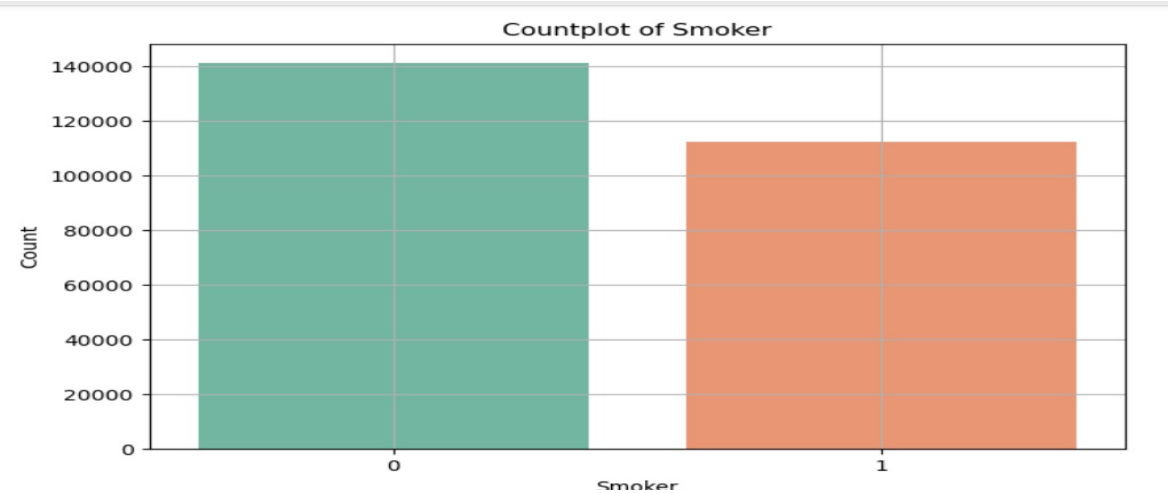
13.Does income vary significantly across different age groups in the dataset?



This represents the relationship between age groups and average income . Each bar shows the average income within a specific age group, providing insights into income distribution . This visualization enables observation of trends such as whether older age groups tend to have higher average incomes compared to younger age groups. This analysis is crucial for understanding income disparities .

Proposed Hypothesis

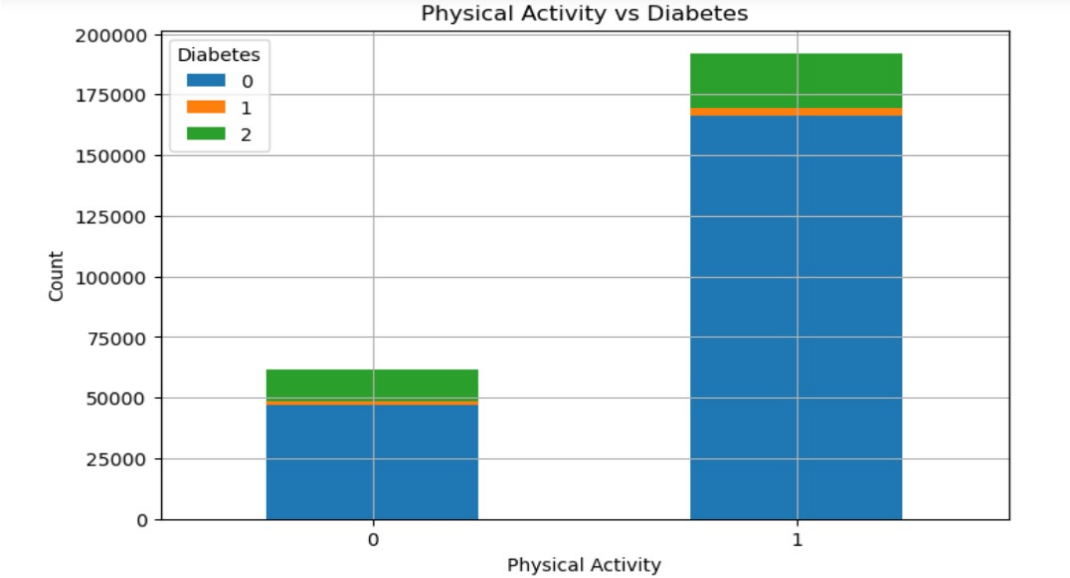
14.How many individuals are smokers?



Inference:The dataset suggests a minority of individuals are smokers. This visualization helps in understanding how smoking behaviors vary within dataset, Each bar represents the number of smokers within a specific age group. Taller bars indicate more smokers in that age group.

Proposed Hypothesis

15.Is there an association between physical activity levels and the incidence of diabetes among individuals in the dataset?



Inference:This plot helps in exploring whether age influences BMI levels and provides a basis for further statistical analysis to quantify the strength and significance of this relationship. Further analysis could include correlation coefficient calculation and regression modeling to better understand how age contributes to BMI variation in the dataset. it's evident that there is a discernible trend where older individuals generally tend to have higher BMIs. This suggests that age may play a role in influencing BMI levels, potentially indicating that as individuals age, there may be a tendency towards higher body mass indices.

MOOC Course Details

Team No. : EDA-A2

Div: C

Sl. No.	Name	SRN.	Course Name	Course Link	Status
1	Rutika W	02FE22BCS088	Infosys Springboard	https://infyspringboard.us.onwingspan.com/	Completed
2	Chandrakala B Walake	02FE22BCS027	Infosys Springboard	https://infyspringboard.us.onwingspan.com/	Completed
3	Rohan M	02FE23BCS423	Infosys Springboard	https://infyspringboard.us.onwingspan.com/	Comleted
4	Affan M	02FE22BCS008	Infosys Springboard	https://infyspringboard.us.onwingspan.com/	Completed

Thank you !