



جامعة مصر للمعلوماتية
EGYPT UNIVERSITY
OF INFORMATICS

Egypt University of Informatics
Computer and Information Systems
Data Analysis Course

The Analysis of Different Health Factors Related to Heart Attack

Submitted by: Mohamed Ibrahim 22-101058

Mohamed Mostafa 22-101203

Mohamed Karim 22-101273

Seif El Islam 22-101172

Abdullah Moussa 22-101114

25/5/2024

Introduction

This report investigates the dataset "Personal Key Indicators of Heart Disease" from 2022, aimed at understanding the factors influencing heart disease prevalence. The analysis includes univariate analysis, bivariate analysis, and applying hypotheses testing techniques.

Research Question

Is Heart Attack related to other health factors?

Hypothesis (all questions are included in the notebook as null hypothesis format)

1. Is there a significant difference in the BMI between individuals who have had a heart attack and those who haven't?
2. Is there a significant difference in the Physical Health Days between individuals who have had a heart attack and those who haven't?
3. Is there a significant difference in the Sleeping Hours between individuals who have had a heart attack and those who haven't?
4. Is there a significant difference in the Mental Health Days between individuals who have had a heart attack and those who haven't?

Population of Interest:

The population of interest includes individuals from the dataset who have been assessed for various health indicators and heart disease status in the USA.

Sampling Method:

The dataset includes a comprehensive collection of individuals without the explicit sampling methods mentioned.

Bias Identification:

Potential biases include:

- Self-reported health indicators may be inaccurate.
- Imbalance in the dataset where non-heart disease cases outnumber heart disease cases.

Dataset:

Number of samples used: 246000 (cleaned data)

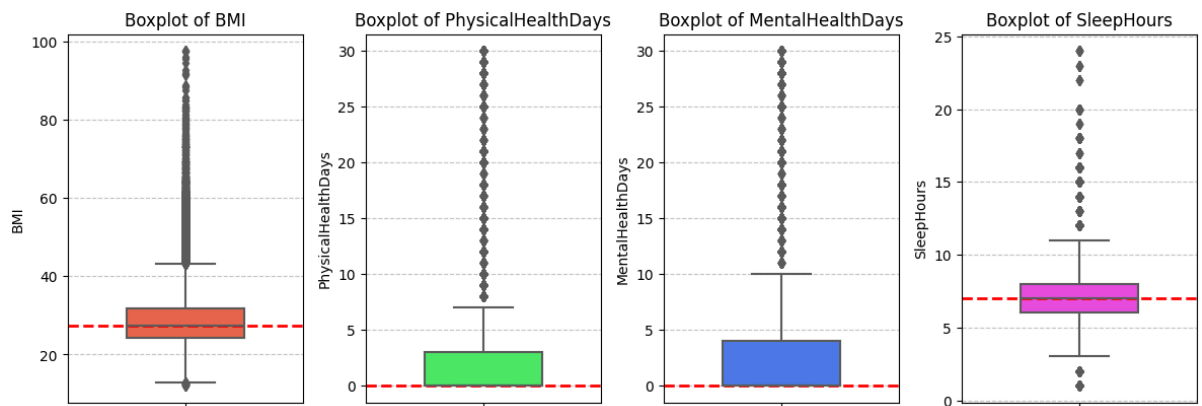
The dataset "heart_2022_no_nans.csv" includes multiple health-related features and the target variable indicating heart disease status.

The dataset includes information on individuals' state of residence, gender, general health status, physical and mental health days, healthcare utilization such as last checkup time, lifestyle factors like physical activity and sleep hours, as well as medical history including occurrences of heart attacks, strokes, asthma, skin cancer, COPD, depressive disorders, kidney disease, arthritis, and diabetes, among others. Additionally, the dataset contains data on sensory impairments, difficulties with daily activities, smoking and e-cigarette usage, chest scans, race/ethnicity, age, height, weight, BMI, alcohol consumption, HIV testing, flu vaccination, pneumococcal vaccination, tetanus vaccination, high-risk status, and COVID-19 positivity.

Analysis:

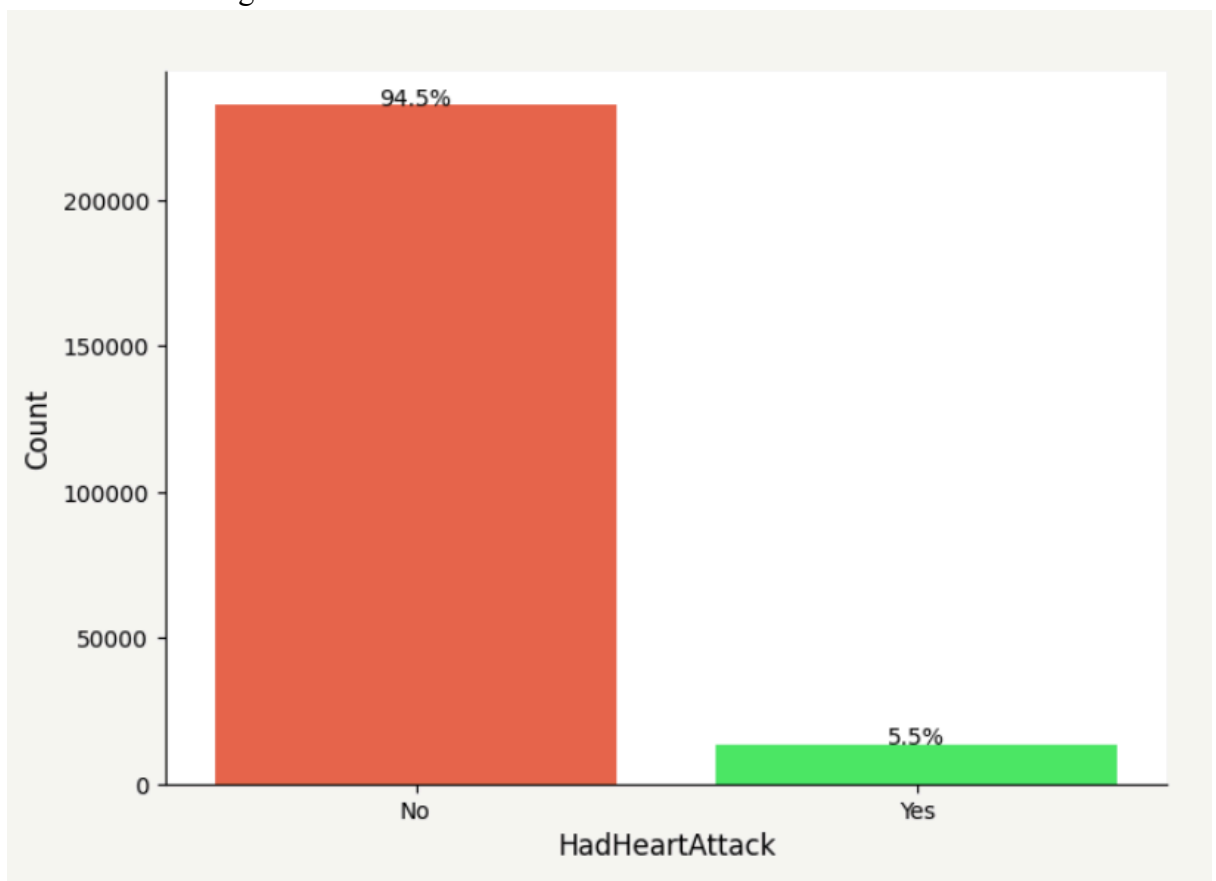
1. **Data Cleaning:** Removed duplicates.
2. **Data Summarization:**
 - Numerical summary statistics using **df.describe()**.
 - Count of distinct values in each column.
3. **Feature Identification:**
 - Numerical columns were identified for detailed analysis.
 - Categorical columns are listed for potential encoding.

4. Univariate Analysis:



Purpose: boxplots for each numerical feature to understand distributions and identify outliers.

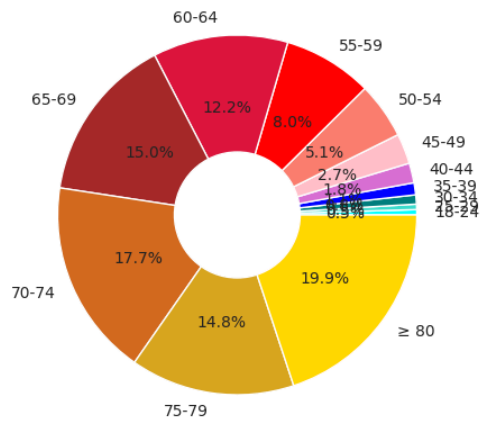
Bar charts for categorical features:



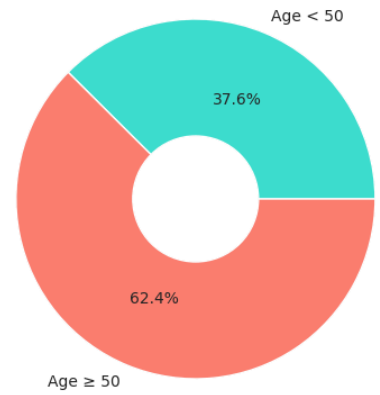
Insights: Only five percent of the American population has a heart attack.

5. Bivariate Analysis:

% Age Category of People with HD



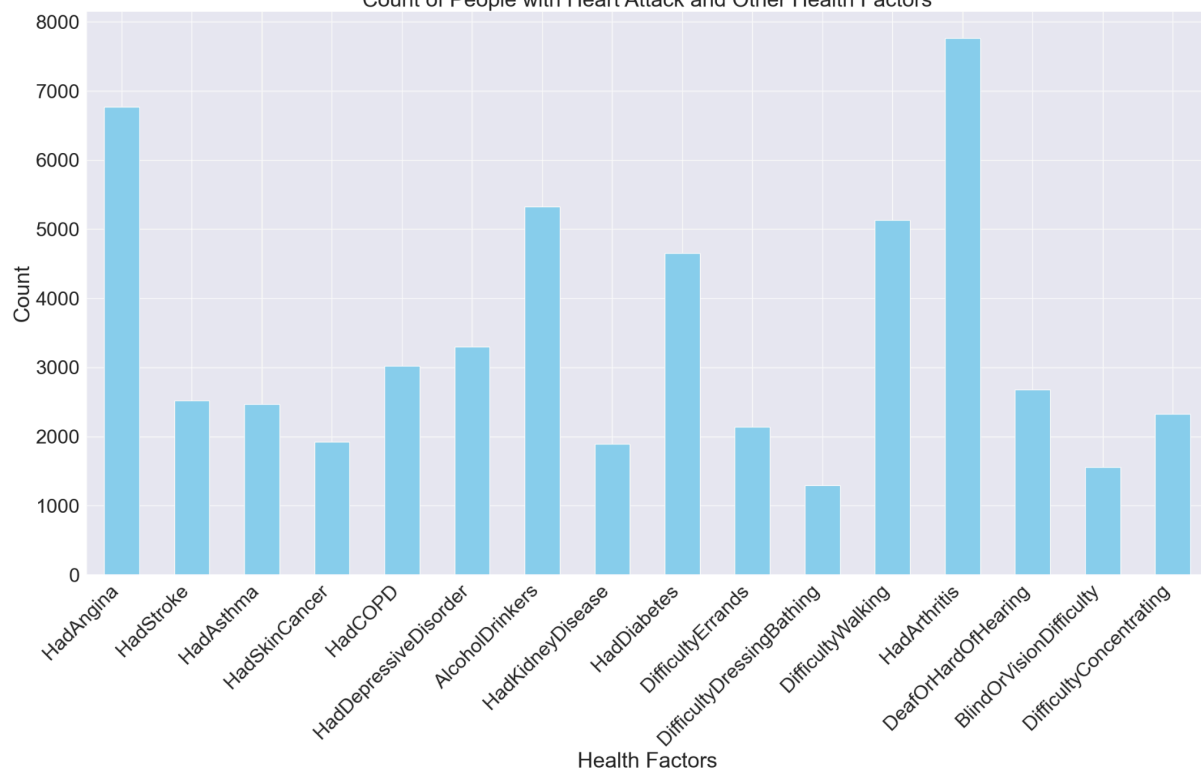
% Age of Less Than 50 and the Complement of People with HD



Insights: Older individuals are more expected to have a heart attack

6. Multivariate Analysis:

Count of People with Heart Attack and Other Health Factors



Insights: Arthritis, Angina, Alcohol, Difficulty in Walking, and Diabetes are the most common diseases among heart attack patients

Hypothesis Testing Steps

- Z-test is used to calculate the P-value for each hypothesis question provided above
- General Steps for Each Test:
 1. Stating the null and alternative hypotheses.
 2. Collecting the data for individuals who had a heart attack and those who did not have.
 3. Conducting a two-sample z-test using the **z-test** function to compare the means of the two groups.
 4. Obtaining the z-statistic and p-value from the z-test.
 5. Set the significance level (alpha: α) at 0.05
 6. Compare the p-value to the alpha:
 - If the p-value is less than 0.05, the null hypothesis is rejected
 - If the p-value is greater than 0.05, the null hypothesis is not rejected (fail to reject it)

Conclusion


The Biggest Contributors to Heart Attack according to the bivariate and multivariate analyses provided above:

1. Arthritis
2. Angina
3. Alcohol
4. Difficulty in Walking
5. Diabetes
6. Aging


The Biggest Contributors to HEART DISEASE

Heart disease is the #1 cause of death in the United States


1. Arthritis




2. Angina




3. Alcohol




4. Difficulty in Walking





5. Diabetes



6. Aging



Now is the time to raise awareness
around heart health and wellness



Z Test for Independence Results:

1. There is a significant difference in the BMI between individuals who have had a heart attack and those who haven't.
2. There is a significant difference in the Physical health days between individuals who have had a heart attack and those who haven't.
3. There is a significant difference in the Mental Health Days between individuals who have had a heart attack and those who haven't.
4. There is no significant difference in the Sleep Hours between individuals who have had a heart attack and those who haven't.

Any potential issues

1. Imbalanced Groups:

If the number of individuals who have had a heart attack is much smaller or larger than those who haven't, it can affect the reliability of the statistical tests.

2. Confounding Variables:

There may be other variables not accounted for in the analysis that could influence both the likelihood of having a heart attack and the variables being tested (e.g., Lifestyle Factors, Gender, Depressive Disorder, Socioeconomic Status, Chronic Conditions, Medication Usage).

Workload Distribution:

Mohamed Ibrahim: **Jupyter Notebook Code & Report**

Mohamed Mostafa: **Jupyter Notebook Code & Report**

Mohamed Karim: **Poster**

Seif El Islam: **Nothing**

Abdullah Moussa: **Nothing**