Navigating Cardiovascular Risk Factors in Adults: A Comprehensive Analysis

A PROJECT REPORT

SUBMITTED TO
SVKM'S NMIMS (DEEMED- TO- BE UNIVERSITY)

IN PARTIAL FULFILMENT FOR THE DEGREE OF

**BACHELORS OF SCIENCE
IN
DATA SCIENCE**

BY

**MIHIR VIJAY BHAGAT
ACHYUT DAMANI
DIYA SINGH RAWAT
VSNL SAHITYA RAO
SUMAN ARJUN LAL YADAV**

SVKM'S NMIMS Deemed to be UNIVERSITY | NILKAMAL SCHOOL OF MATHEMATICS, APPLIED STATISTICS & ANALYTICS

NMIMS NSoMASA

Ground Floor, SBMP Phase I,
Irla, N. R. G Marg, Opposite Cooper Hospital,
Vile-Parle (West), Mumbai – 400 056.

21st NOVEMBER 2023

CERTIFICATE

This is to certify that work described in this thesis entitled "Navigating Cardiovascular Risk Factors in Adults: A Comprehensive Analysis" has been carried out by (Mihir Vijay Bhagat, Achyut Damani, Diya Singh Rawat, VSNL Sahitya Rao and Suman Arjun Lal Yadav) under my supervision. I certify that this is his/her bonafide work. The work described is original and has not been submitted for any degree to this or any other University.

Date: 21st November, 2023
Place: - Mumbai

SUPERVISOR

(Prof. Vaibhav Vasundekar)

ACKNOWLEDGEMENT

We would like to express our profound gratitude to Prof. Vaibhav Vasundekar and Ex NMIMS Prof. Piyush Shaha whose guidance and support have been instrumental in the successful completion of our project, “Navigating Cardiovascular Risk Factors in Adults: A Comprehensive Analysis”. His invaluable critiques and encouraging words have been a source of constant motivation.

We also extend our heartfelt thanks to NMIMS’s Nilkamal School of Mathematics, Applied Statistics and Analytics, (NSoMASA) Mumbai for providing us with the opportunity to undertake this initiative. The institution’s unwavering support and conducive environment have played a significant role in our project’s success. We are truly grateful for the enriching experience.

CONTENTS

ABSTRACT

This research project, titled "Navigating Cardiovascular Risk Factors in Adults: A Comprehensive Analysis", aimed to identify and analyze the key risk factors contributing to the prevalence of cardiovascular diseases (CVDs) in young adults. The study utilized a comprehensive approach, employing descriptive statistics, regression, correlation, chi-square tests, and t-tests on a dataset sourced from Kaggle. The analysis revealed significant associations between CVDs and various factors such as age, gender, height, weight, systolic and diastolic blood pressure, cholesterol, glucose, smoking status, alcohol intake, physical activity, and Body Mass Index (BMI). A real-life dataset was also included in the analysis for comparison. The study found that individuals who are older, male, taller, heavier, have higher blood pressure levels, have higher cholesterol levels, smoke, and are less physically active are at increased risk of CVDs. The findings from this study could potentially guide preventive strategies and health policies, thereby reducing the burden of CVDs in young adults. The research project serves as a foundation for future studies in this critical area of public health.

INTRODUCTION

Cardiovascular diseases (CVDs) have long been recognized as a leading cause of mortality worldwide. Globally, CVDs are the number 1 cause of death, accounting for 1 in 3 deaths. In 2010, CVDs cost US$ 863 billion, and this is estimated to rise by 22% to US$ 1,044 billion by 2030. 80% of CVD deaths occur in low- to middle-income countries.

In India, non-communicable diseases (NCDs), including CVDs, are estimated to account for 60% of total adult deaths. CVDs account for over a quarter (26%) of these deaths. There are about 30 million heart patients in India, with 16 million from rural areas and 14 million from urban areas. Every year, 2 lakh heart surgeries are performed. 25% of the heart-related deaths, especially heart attacks, happen to those less than 40. A recent nationwide survey by ICMR has shown that 28% of adults in India have hypertension, and nearly 77% are undiagnosed, leading to a high burden of hypertension in the country. The prevalence of coronary heart disease has increased in India.

Our project, conducted under the subject ‘Research Initiative in Data Science’ (RIDS), focuses on “Navigating Cardiovascular Risk Factors in Adults: A Comprehensive Analysis”. Traditionally, young adults have been considered a low-risk group for cardiovascular diseases. However, recent trends indicate a surge in incidence rates among this demographic. This shift is particularly concerning given the long-term health implications and the potential burden on healthcare systems.

The story of a heart attack is all too familiar. It often begins with chest pain or discomfort, shortness of breath, and cold sweats. These symptoms are the body’s desperate plea for help, signalling that the heart is not receiving enough oxygen due to blocked arteries. The most common reason for this is a build-up of fatty deposits on the inner walls of the blood vessels that supply the heart. If not addressed promptly, this can lead to a heart attack, a life-threatening condition that requires immediate medical attention.

Our research aims to shed light on the risk factors that contribute to the development of cardiovascular diseases in young adults. We believe that understanding these risk factors is the first step towards prevention. By identifying the key contributors, we can inform public health strategies aimed at reducing the incidence of these diseases. Furthermore, our findings could potentially guide individual lifestyle choices, encouraging young adults to adopt healthier habits to mitigate their risk.

In the course of our research, we have used various statistical tools and software, including R, Python (Spyder, Jupyter, and Google Collab) and MS Excel, to analyze our data. Our dataset, sourced from Kaggle. This comprehensive dataset has allowed us to conduct a thorough analysis of the various factors that may contribute to the risk of developing cardiovascular diseases in adults.

As we delve deeper into our research, we remain cognizant of the broader implications of our work. Cardiovascular diseases are not just a health issue; they are a social issue, an economic issue, and a development issue. By contributing to the understanding of these diseases, we hope to make a meaningful impact on the lives of young adults around the world.

In recent times, we have seen several celebrities succumbing to heart diseases. For instance, Bollywood actor Sidharth Shukla passed away due to a cardiac arrest at the age of 40. Comedian Raju Srivastava also suffered a heart attack while exercising at a gym in Delhi. Singer KK tragically passed away after a heart attack post his musical concert. These incidents serve as a stark reminder of the pervasiveness of heart diseases and the urgent need for preventive measures.

RATIONALE

The rationale for our project, “Navigating Cardiovascular Risk Factors in Adults: A Comprehensive Analysis”, is rooted in the increasing prevalence of cardiovascular diseases among adults. Despite advancements in medical science, cardiovascular diseases remain a leading cause of mortality globally. Young adults, traditionally considered low-risk, are experiencing a surge in incidence rates. This alarming trend necessitates a comprehensive study to understand the risk factors contributing to this shift.

Our research is significant from a public perspective as it aims to identify and analyze these risk factors, with a focus on lifestyle choices, genetic predisposition, and environmental influences. The findings from our study could potentially guide preventive strategies and health policies, thereby reducing the burden of cardiovascular diseases in young adults. Furthermore, it could provide a foundation for future research in this area.

This topic was chosen due to the urgent need to address the rising incidence of cardiovascular diseases in young adults. By focusing on this demographic, it is hoped to contribute to a better understanding of the disease and its risk factors, ultimately leading to improved prevention and treatment strategies. The project aligns with the broader goal of fostering research that addresses pressing health issues, and it is believed that this study could make a significant contribution to the field of cardiovascular health, particularly in the context of young adults.

AIM & OBJECTIVES

1. **Identifying & Examining the Associations between Key Risk Factors**: The first objective is to identify the key risk factors that contribute to cardiovascular disease in adults. These risk factors could include lifestyle choices, genetic predisposition, and environmental influences. Once these risk factors are identified, the project aims to examine the associations between these factors and the prevalence of cardiovascular disease. This could involve statistical analysis to determine correlations and causati ons.

2. **Gender Differences in Cardiovascular Disease**: The third objective is to explore whether there are gender differences in the prevalence of cardiovascular disease among adults. This could involve comparing the incidence rates of cardiovascular disease between males and females in the dataset. The findings could reveal whether one gender is more susceptible to cardiovascular disease and could lead to further investigations into the reasons behind any observed differences.

3. **Age Groups with More Cardiovascular Disease**: The second objective is to identify the age groups within the adult demographic (30-45) that are more prone to cardiovascular disease. This could involve stratifying the data by age and analyzing the prevalence of cardiovascular disease within each age group. The findings could provide insights into whether certain age groups are at higher risk and why that might be the case.

LITERATURE REVIEW

In the ‘Literature Review’ section, we delve into the existing body of knowledge related to our research topic. We critically analyze and synthesize the relevant published work to contextualize our study within the broader academic discourse.

Cardiovascular diseases (CVDs) are a significant global health concern due to their rising prevalence and the resulting mortality and disability, which impose a heavy economic burden. A study analyzing the trend in CVD incidence, mortality, and mortality-to-incidence ratio (MIR) across the world over 28 years found that there was an overall downward trend in CVD incidence and mortality rates, while the survival rate of CVD patients was rather stable. This suggests that global efforts for controlling the CVD burden have been quite successful, but there is an urgent need for more efforts to improve the survival rate of patients and lower the burden of this disease in some areas with an increasing trend of either incidence or mortality.

Another study focused on the role of meaning in life for psychological stress, mental health, and CVD risks. The study found that a central clinical concern for patients is their question of how to live a meaningful life despite CVD. Meaning-centered concerns seem to lead to lower motivation to make lifestyle changes, more psychological stress, lower quality-of-life, worse physical well-being, and increased CVD risk. The ability to live a meaningful life after CVD events is related to lower stress, better mental health, and several biomarkers.

A systematic literature review was conducted to uncover the challenges associated with imbalanced data in heart disease predictions. The study posits directions for future research. In conclusion, the literature shows that while strides have been made in controlling the CVD burden, there are areas, such as improving patient survival rates and addressing meaning-centered concerns, that require further attention and research.

Continuing the review, the article titled “Trend analysis of cardiovascular disease mortality, incidence, and mortality-to-incidence ratio: results from global burden of disease study 2017” was published in BMC Public Health. The authors, Maedeh Amini, Farid Zaveri, and Masoud Salehi, aimed to analyze the trend in cardiovascular disease (CVD) incidence, mortality, and mortality-to-incidence ratio (MIR) across the world over 28 years.

The study used age-standardized CVD mortality and incidence rates from the Global Burden of Disease (GBD) Study 2017 for both genders and different world super regions with available data every year during the period 1990–2017. The Human Development Index was sourced from the United Nations Development Programme (UNDP) database for all countries at the same time interval. The marginal modeling approach was implemented to evaluate the mean trend of CVD incidence, mortality, and MIR for 195 countries and separately for developing and developed countries.

The results showed that the global mean trend of CVD incidence had an ascending trend until 1996 followed by a descending trend after this year. Nearly all of the countries experienced a significant declining mortality trend from 1990 to 2017[1]. Likewise, the global mean MIR rate had a significant trivial decrement trend with a gentle slope of 0.004 over the time interval. The reduction in incidence and mortality rates for developed countries was significantly faster than developing counterparts in the period 1990–2017. This further emphasizes the need for targeted interventions in specific regions to address the burden of CVDs.

Continuing with the literature review, an article titled "Cardiovascular disease and meaning in life: A systematic literature review and conceptual model" was published in Palliative & Supportive Care. The author, Joel Vos, conducted a systematic literature review on the relationships between cardiovascular disease (CVD) and meaning in life.

The study included 113 studies on meaning and CVD. The literature showed that a central clinical concern for patients is their question of how to live a meaningful life despite CVD. Meaning-centered concerns seem to lead to lower motivation to make lifestyle changes, more psychological stress, lower quality-of-life, worse physical well-being, and increased CVD risk. The author developed an evidence-based conceptual framework for the relationship between meaning and CVD.

Transitioning to the next piece of literature, a document titled "Md Manjurul Ahsan arXiv:2112.06459v1 [cs.LG] 13 Dec 2021" discusses the application of machine learning (ML) in detecting heart disease during the early stage using electrocardiogram (ECG) and patients' data.

The authors state that recent advancements in ML applications demonstrate that detecting heart disease during the early stage using ECG and patients' data is feasible. However, the imbalance in both ECG and patients' data raises challenges for traditional ML to perform unbiasedly. The authors suggest that addressing this imbalance is crucial for improving the effectiveness of ML applications in early-stage heart disease detection.

Heart disease is one of the significant challenges in today's world and one of the leading causes of many deaths worldwide. Therefore, improving the ability to detect heart disease at an early stage using ML could have significant implications for public health. This further emphasizes the need for targeted interventions in specific regions to address the burden of CVDs.

Continuing with the literature review, a document titled "A Literature Review of Cardiovascular Disease Management Programs in Managed Care Populations" by Sheta Ara was published in the Journal of Managed Care Pharmacy. The author conducted a literature review of 20 studies on cardiovascular disease (CVD) management programs in managed care populations.

The studies reviewed included 5 in patients with congestive heart failure (CHF), 9 in hypertensive patients, and 6 in hyperlipidemia and/or coronary artery disease (hyperlipidemia-CAD) patients. The management of CHF involved multifaceted programs that included the participation of multiple health care professionals, patient and physician education, promotion of intensive drug therapy and lifestyle modifications, and close patient monitoring.

Transitioning to the next piece of literature, a document titled “Cardiovascular disease and meaning in life: A systematic literature review and conceptual model” by Joel Vos was published in Palliative and Supportive Care. The author conducted a systematic literature review on the relationships between cardiovascular disease (CVD) and meaning in life.

The study included 113 studies on meaning and CVD. The literature showed that a central clinical concern for patients is their question of how to live a meaningful life despite CVD. Meaning-centered concerns seem to lead to lower motivation to make lifestyle changes, more psychological stress, lower quality-of-life, worse physical well-being, and increased CVD risk.

The author developed an evidence-based conceptual framework for the relationship between meaning and CVD. It may be hypothesized that CVD patients may benefit from psychological therapies focused on meaning. This further emphasizes the need for targeted interventions in specific regions to address the burden of CVDs.

The previous research has provided valuable insights into the prevalence, risk factors, and management of cardiovascular diseases (CVDs). It has highlighted the importance of understanding the meaning of life for patients with CVDs and the role it plays in their health outcomes. The research has also emphasized the challenges posed by imbalanced data in heart disease predictions and the potential of machine learning in early-stage heart disease detection. These findings have informed our approach to analyzing risk factors for CVDs in young adults and have guided our choice of statistical tools and methodologies. This wealth of knowledge has been instrumental in shaping our research direction, enabling us to build upon the existing body of work and contribute new insights to the field.

DATA PREPARATION

Data Collection:

Cardiovascular Disease Dataset: About Dataset

The data consists of 13 columns and 10,273 rows. Each column is described below:

Sr. No.	Attribute	Description	Unit	Type
1	Age	The number of years since the user was born.	In years	Numeric
2	Gender	The gender of the patient.	Female, Male	Binary
3	Height	The height of the user in meters or feet.	67-198	Numeric
4	Weight	The weight of the user in kilograms or pounds.	28-200	Numeric
5	Systolic	The systolic blood pressure of the user in mmHg.	120 - 14020	Numeric
6	Diastolic	The diastolic blood pressure of the user in mmHg.	0 - 8500	Numeric
7	Cholesterol	The cholesterol level of the user in mg/dL or mmol/L	0=Normal, 1=High, 2= Extremely High	Nominal
8	Glucose	The glucose level of the user in mg/dL or mmol/L.	0=Normal, 1=High, 2= Extremely High	Nominal
9	Smoke	Whether the user smokes or not.	0,1 (0 = No, 1 = Yes)	Binary
10	Alcohol_Intake	The amount of alcohol consumed by the user.	0,1 (0 = No, 1 = Yes)	Binary

11	Physical_Activity	The amount of physical activity performed by the user.	0,1 (0 = No, 1 = Yes)	Binary
12	Cv_Disease	Whether the user has any cardiovascular disease or not.	0,1 (0 = No, 1 = Yes)	Binary
13	Bmi	The body mass index of the user, calculated as weight divided by height squared (kg/m^2 or lb/ft^2).	14.6 - 278.1	Numeric

Data Preparation:

Binary format (0s and 1s) was used to transform categorical Platform columns for easier analysis and enhanced compatibility with statistical methods like general linear regression.

DATA PREPARATION FOR SECOND DATASET

The Hospital Dataset: About Dataset

The dataset is collected from a hospital. The data consists of 12 columns and 22 rows. Each column is described below**:**

1. **Age**: The average age of the patients is approximately 62 years, with a standard deviation of around 14 years. The youngest patient is 38 years old, and the oldest is 89 years old.
2. **BMI**: The average BMI of the patients is approximately 26.7, with a standard deviation of around 2.3. The lowest BMI is 21.9, and the highest is 31.
3. **Gender**: The majority of the patients are male (approximately 86%).
4. **Smoking**: Most of the patients are non-smokers (approximately 91%), with a small percentage being smokers (approximately 9%).
5. **Alcohol**: The majority of the patients do not consume alcohol (approximately 81%), while a minority do (approximately 19%).
6. **Physical Activity**: Most of the patients engage in physical activity (approximately 67%), while a significant proportion do not (approximately 33%).
7. **Diet**: Most of the patients follow a vegetarian diet (approximately 67%), while the rest follow a mixed diet (approximately 33%).
8. **Stress**: Most of the patients do not report stress (approximately 67%), while a significant proportion do (approximately 33%).
9. **DM (Diabetes Mellitus)**: A slight majority of the patients have diabetes (approximately 62%), while the rest do not (approximately 38%).
10. **High BP (Blood Pressure)**: Similar to diabetes, a slight majority of the patients have high blood pressure (approximately 62%), while the rest do not (approximately 38%).
11. **Medication**: Most of the patients are on medication (approximately 65%), while a significant proportion are not (approximately 35%).
12. **Outcomes**: The majority of the patients underwent CABG (Coronary Artery Bypass Grafting) surgery (approximately 76%), while the rest underwent PTCA (Percutaneous Transluminal Coronary Angioplasty) (approximately 24%).

Data Preparation:

1. Hypothesis Testing: The t-test involves setting up hypotheses (null and alternative) to test the significance of the observed differences. For example, it can be used to evaluate whether there is a significant difference in the mean stress levels between different patient groups.

2. Statistical Significance: The t-test provides a statistical measure, the t-statistic, which is used to assess the significance of the observed differences. A larger t-statistic indicates a greater difference between the means.

3. Application in Healthcare: In a hospital setting, t-tests can be applied to compare means of various clinical parameters (e.g., blood pressure, cholesterol levels) between different patient groups (e.g., treated vs. untreated, different age groups).

4. Binary format (0s and 1s) was used to transform categorical Platform columns for easier analysis and enhanced compatibility with statistical methods like general linear regression.

METHODOLOGY

The following libraries and modules are used to perform the required functions:

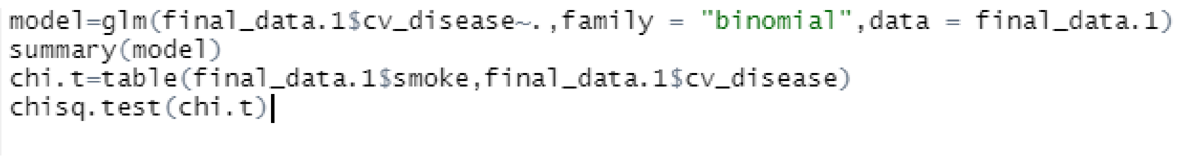
1. **Pandas:** is a Python library that provides data structures and tools for working with structured data. It is particularly useful for data cleaning, transformation, and analysis.
2. **NumPy:** is another Python library that provides support for large, multi-dimensional arrays and matrices, as well as a large collection of mathematical functions to operate on these arrays.
3. **Matplotlib:** is a data visualization library that provides a wide range of tools for creating static, animated, and interactive visualizations in Python. It is highly customizable and can be used to create a variety of charts, plots, and graphs.
4. **Seaborn:** is a Python data visualization library based on Matplotlib. It provides a high-level interface for creating informative and attractive statistical graphics. Seaborn is particularly useful for visualizing complex datasets and for creating visualizations that highlight relationships between variables.
5. **Stats:** module in SciPy provides functions for performing statistical tests, such as t-tests and chi-squared tests. **ttest_ind** is a function that performs an independent two-sample t-test, while **chi2_contingency** is a function that performs a chi-squared test of independence.

General linear Regression

General linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is a generalization of simple linear regression, which models the relationship between two variables. In general, linear regression, the relationship between the dependent variable and the independent variables can be linear or nonlinear. The method is used to predict the value of the dependent variable based on the values of the independent variables.

General linear regression is used in a wide range of fields, including economics, finance, biology, and engineering. It is often used to analyze the relationship between variables and to make predictions about future values of the dependent variable. The method is also used to test hypotheses about the relationship between variables and to identify which independent variables are most strongly associated with the dependent variable.

Code explanation the general linear regression



```
model=glm(final_data.1$cv_disease~.,family = "binomial",data = final_data.1)
summary(model)
chi.t=table(final_data.1$smoke,final_data.1$cv_disease)
chisq.test(chi.t)
```

> model=glm(final_data.1$cv_disease~.,family = "binomial",data = final_data.1): This line of code fits a generalized linear model (GLM) to the data in final_data.1. The glm() function is used to fit GLMs in R. In this case, the response variable is cv_disease, and the tilde (~) separates the response variable from the predictor variables. The indicates that all other variables in the data frame should be used as predictors. The family argument specifies the type of GLM to fit, which in this case is a binomial GLM.

> **summary(model)**: This line of code prints a summary of the GLM model that was fit in the previous line. The summary includes information about the coefficients, standard errors, z-values, and p-values for each predictor variable.

> chi.t=table(final_data.1$smoke, final_data.1$cv_disease): This line of code creates a contingency table of the variables smoke and cv_disease in final_data.1.

> **chisq.test(chi.t):** This line of code performs a chi-squared test of independence on the contingency table created in the previous line. The chi-squared test is used to determine whether there is a significant association between two categorical variables.

Height Distribution Analysis

Height distribution analysis is a statistical method used to analyze the distribution of heights in a population. It is often used to determine whether the heights in a population follow a normal distribution, which is a bell-shaped curve that is commonly found in nature and social sciences.

Code explanation for Height Distribution Analysis

- **plt.figure(figsize=(10, 6)):** This line of code creates a new figure with a size of 10 inches by 6 inches.
- **plt.hist(cardio['height'], bins=20, color='skyblue', edgecolor='black'**): This line of code creates a histogram of the ‘height’ variable in the ‘cardio’ dataset. The hist() function is used to create histograms in Matplotlib. The bins argument specifies the number of bins to use in the histogram, while the color and edgecolor arguments specify the color of the bars and their edges, respectively.
- **plt.title('Histogram of Height'):** This line of code adds a title to the histogram.
- **plt.xlabel('Height (cm)'):** This line of code adds a label to the x-axis of the histogram.
- **plt.ylabel('Frequency'):** This line of code adds a label to the y-axis of the histogram.
- **plt.grid(True):** This line of code adds a grid to the histogram.
- **plt.show():** This line of code displays the histogram..

Weight Distribution Analysis

Weight distribution analysis is a statistical method used to analyze the distribution of weights in a population. It is often used to determine whether the weights in a population follow a normal distribution, which is a bell-shaped curve that is commonly found in nature and social sciences.

Code explanation for weight Distribution Analysis

- **plt.figure(figsize=(10, 6))** line of code creates a new figure with a size of 10 inches by 6 inches.
- **plt.hist()** function is used to create the histogram, with the bins argument specifying the number of bins to use in the histogram.
- The color and edgecolor arguments specify the color of the bars and their edges, respectively.
- **plt.title(), plt.xlabel(), and plt.ylabel()** functions are used to add a title and labels to the x- and y-axes of the histogram.
- **plt.grid(True)** function adds a grid to the histogram, and the plt.show() function displays the histogram.

Correlation Matrix

A correlation matrix is a table that shows the correlation coefficients between variables. It can be used to summarize data, visualize the patterns and relationships between variables, and as a diagnostic or an input for more advanced analyses. The correlation coefficients range from -1 to 1, indicating the strength and direction of the correlation.

Correlation matrices are used in a wide range of fields, including finance, economics, psychology, and biology. They are often used to analyze the relationship between variables and to make predictions about future values of the dependent variable. The method is also used to test hypotheses about the relationship between variables and to identify which independent variables are most strongly associated with the dependent variable

Code explanation for correlation matrix

The corr() function is used to calculate the correlation coefficients between all pairs of variables in the dataset. The resulting correlation matrix is then passed to the sns.heatmap() function, which creates a heatmap of the correlations. The annot=True argument adds the correlation coefficients to the heatmap, while the cmap='coolwarm' argument specifies the color scheme to use. The fmt=".2f" argument specifies that the correlation coefficients should be formatted to two decimal places. The linewidths=.5 argument specifies the width of the lines separating the cells in the heatmap. Finally, the plt.title() function adds a title to the heatmap, and the plt.show() function displays the heatmap.

RESULTS & DISCUSSION

General linear Regression

The General linear regression model was used to predict the presence or absence of cardiovascular disease (CVD) in a sample of individuals. The model was fit using the following variables:

- Age
- Gender (Male/Female)
- Height
- Weight
- Systolic blood pressure
- Diastolic blood pressure
- Cholesterol
- Glucose
- Smoking status (Yes/No)
- Alcohol intake (Yes/No)
- Physical activity (Yes/No)
- Body mass index (BMI)

The results of the model show that the following variables were significantly associated with the presence of CVD:

- Age (positive association)
- Gender (Male)
- Height (negative association)
- Weight (positive association)
- Systolic blood pressure (positive association)
- Diastolic blood pressure (positive association)
- Cholesterol (positive association)
- Smoking status (negative association)
- Physical activity (negative association)

These results suggest that individuals who are older, male, taller, heavier, have higher blood pressure levels, have higher cholesterol levels, smoke, and are less physically active are at increased risk of CVD.

Conclusion

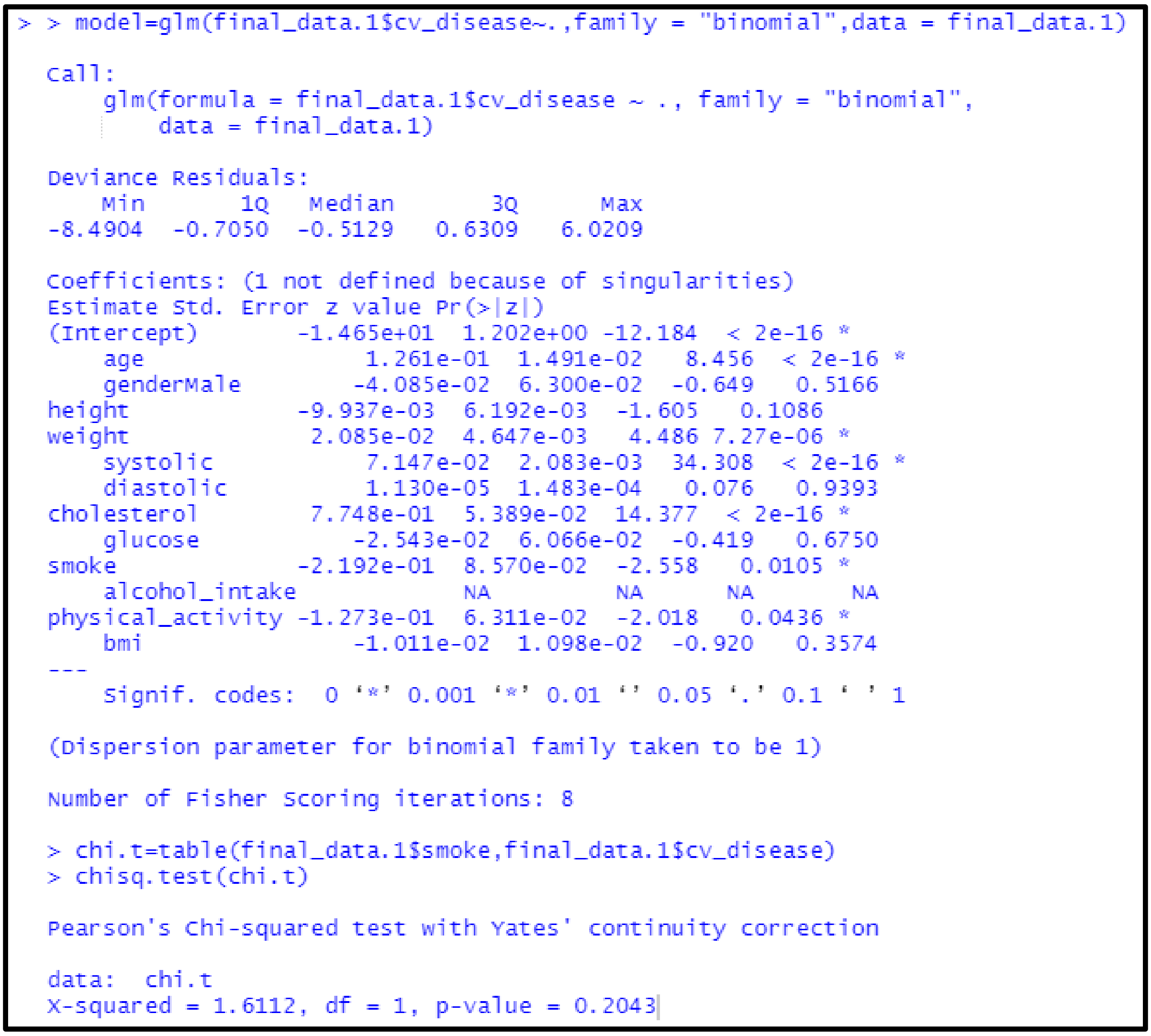
The logistic regression model provides a useful tool for predicting the presence or absence of CVD. The model can be used to identify individuals at high risk for CVD, who may benefit from early intervention and prevention strategies.

NOTE:

The chi-squared test for association between smoking and CVD was not statistically significant at the 0.05 level. However, the p-value was close to the significance level, so this result should be interpreted with caution.

The model was fit using Fisher scoring, which is an iterative algorithm that maximizes the likelihood of the model parameters.

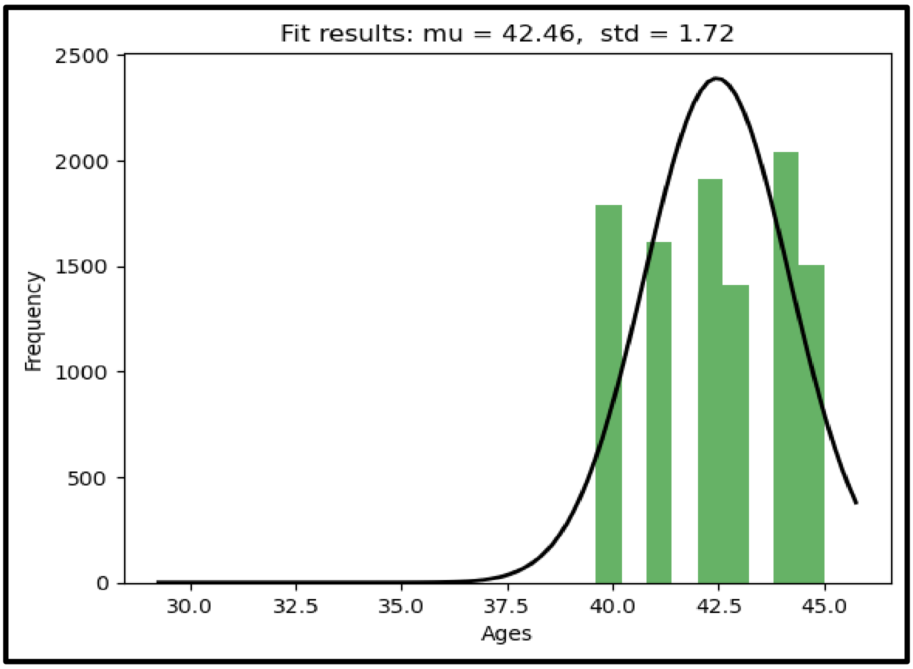
The dispersion parameter for the binomial family was taken to be 1.



```
> > model=glm(final_data.1$cv_disease~.,family = "binomial",data = final_data.1)

  Call:
      glm(formula = final_data.1$cv_disease ~ ., family = "binomial",
          data = final_data.1)

  Deviance Residuals:
      Min       1Q   Median       3Q      Max
  -8.4904  -0.7050  -0.5129   0.6309   6.0209

  Coefficients: (1 not defined because of singularities)
  Estimate Std. Error z value Pr(>|z|)
  (Intercept)       -1.465e+01  1.202e+00 -12.184  < 2e-16 *
      age                1.261e-01  1.491e-02   8.456  < 2e-16 *
      genderMale        -4.085e-02  6.300e-02  -0.649   0.5166
  height            -9.937e-03  6.192e-03  -1.605   0.1086
  weight             2.085e-02  4.647e-03   4.486 7.27e-06 *
      systolic           7.147e-02  2.083e-03  34.308  < 2e-16 *
      diastolic          1.130e-05  1.483e-04   0.076   0.9393
  cholesterol        7.748e-01  5.389e-02  14.377  < 2e-16 *
      glucose           -2.543e-02  6.066e-02  -0.419   0.6750
  smoke             -2.192e-01  8.570e-02  -2.558   0.0105 *
      alcohol_intake            NA         NA      NA       NA
  physical_activity -1.273e-01  6.311e-02  -2.018   0.0436 *
      bmi               -1.011e-02  1.098e-02  -0.920   0.3574
  ---
      Signif. codes:  0 ‘*’ 0.001 ‘*’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

  (Dispersion parameter for binomial family taken to be 1)

  Number of Fisher Scoring iterations: 8

  > chi.t=table(final_data.1$smoke,final_data.1$cv_disease)
  > chisq.test(chi.t)

  Pearson's Chi-squared test with Yates' continuity correction

  data:  chi.t
  X-squared = 1.6112, df = 1, p-value = 0.2043
```

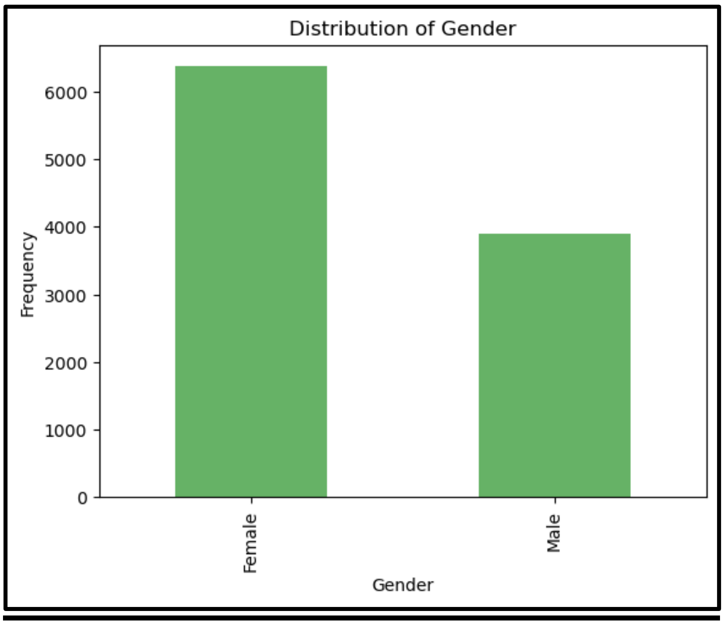
Age Distribution

- The **average age (mean)** in this population is approximately **42 years**.
- The highest frequency appears to be around age **42**, aligning with the mean age.
- Distribution suggests that the age distribution is centered around **42 years** and it is negatively skewed.



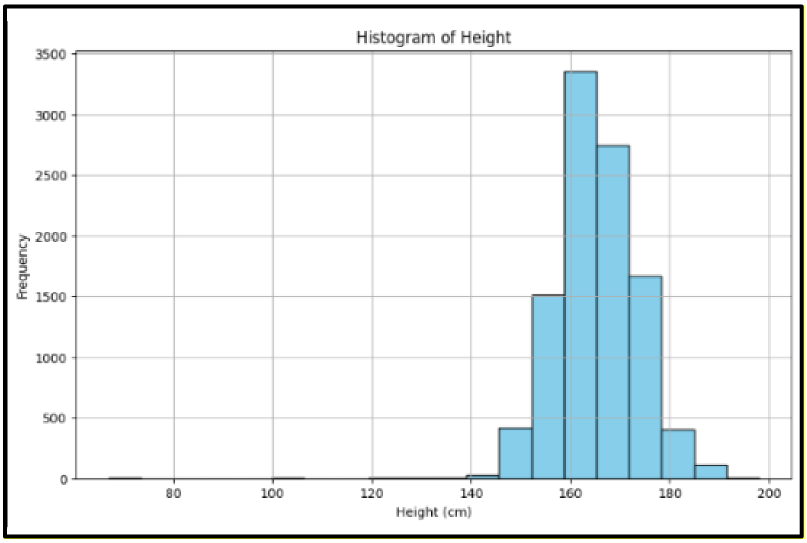
Gender Distribution

Females have a higher chance of having cardiovascular disease



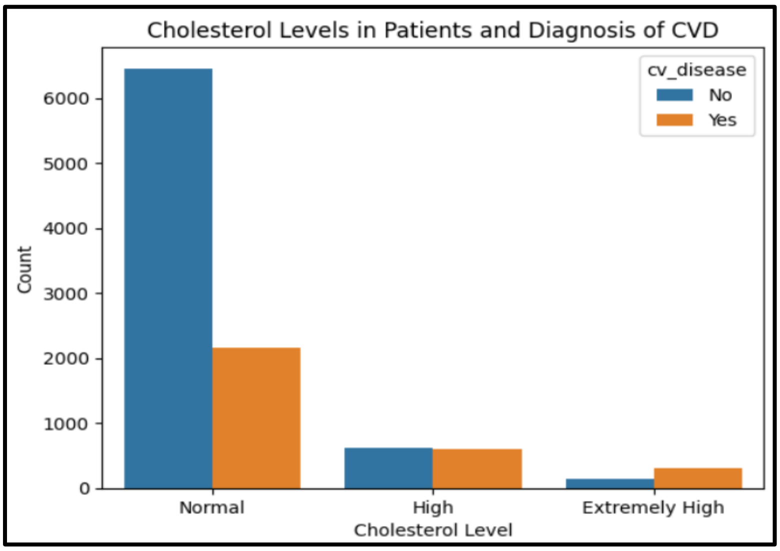
Height Distribution

- Most common height in the population is around **160 cm**.
- Range of heights spans from approximately **80 cm to 200 cm**.
- Fewer individuals at the extreme ends of the range.
- The data appears to follow a normal distribution, with most individuals having a height around the mean (160 cm).



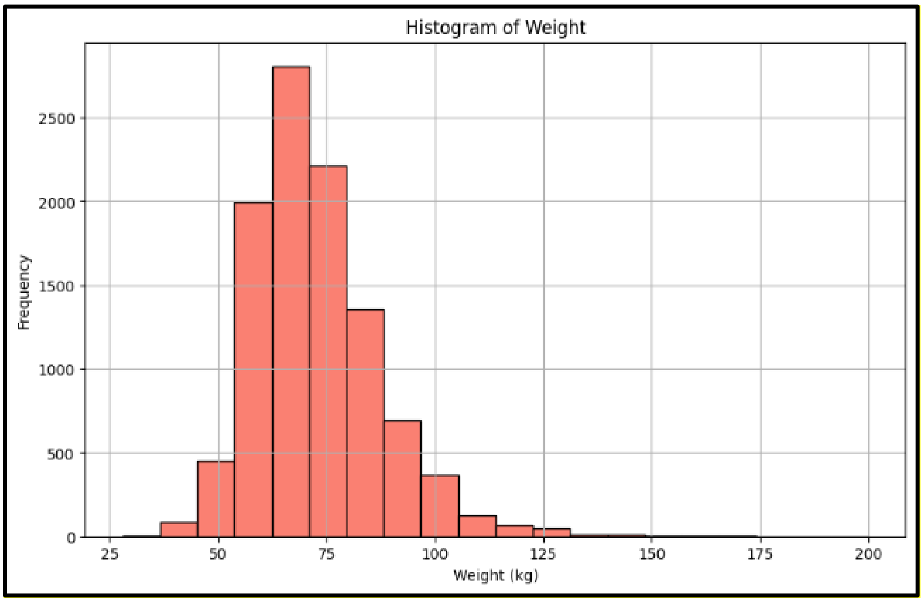
Cholesterol Levels

- The highest count of patients is in the "Normal" cholesterol level category, with a significantly higher number of patients without CVD compared to those with CVD.
- As the cholesterol level increases to "High" and "Extremely High", the number of patients without CVD decreases. However, the number of patients with CVD does not show a significant increase.

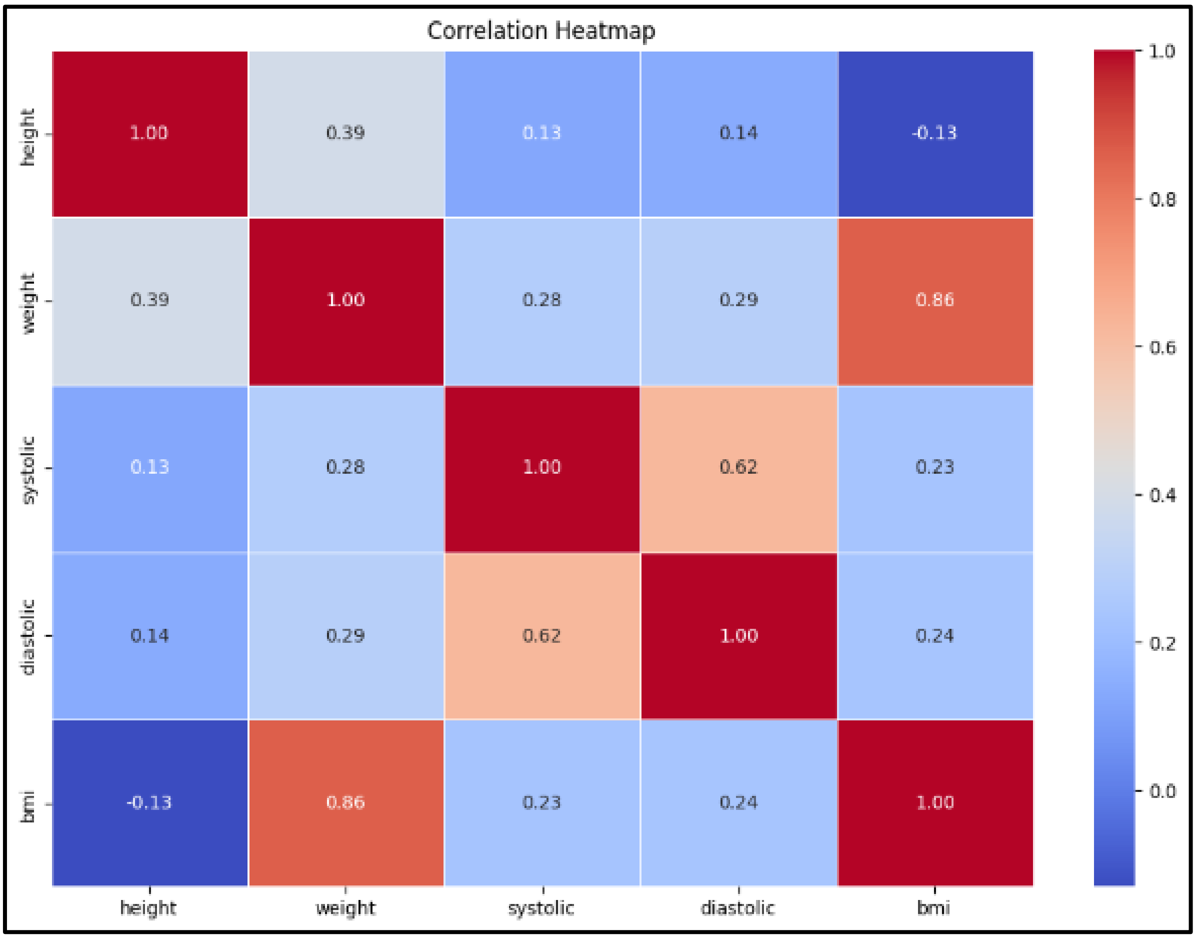


Weight Distribution:

- The most common weight in the population is around **75 kg**.
- The data suggests a normal distribution, with most individuals having a weight around the mean (75 kg).



Correlation Analysis



High Association

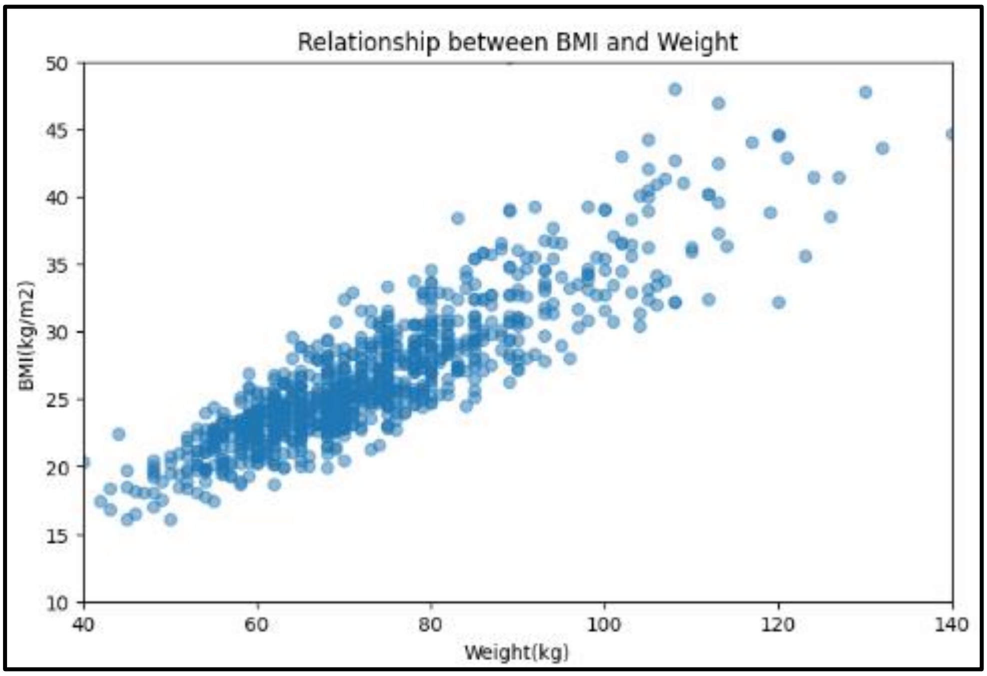
1. **Weight and BMI**: Strong positive correlation observed. As weight increases, BMI also tends to increase. This is expected as BMI (Body Mass Index) is a measure that uses your height and weight to work out if your weight is healthy.
2. **Diastolic and Systolic**: Strong positive correlation observed. When systolic blood pressure increases, diastolic blood pressure also tends to increase, and vice versa. This is common as both readings are measures of blood pressure and often tend to increase or decrease together.

Less Association

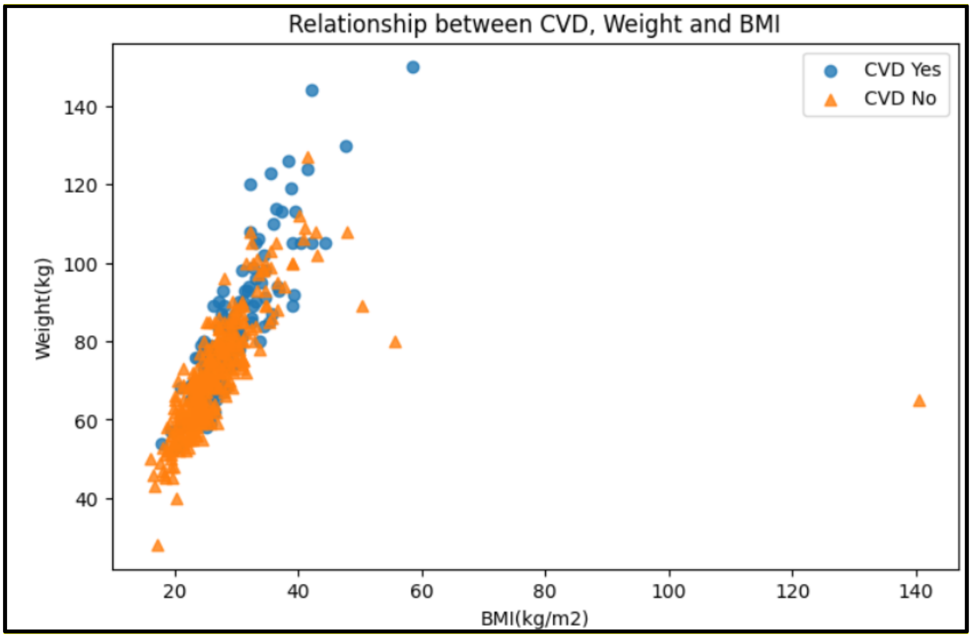
1. **Height and BMI**: Weaker correlation observed. This suggests that height has less of an influence on BMI. While height is a factor in the calculation of BMI, the relationship is not linear and hence the correlation is not as strong.
2. **Height and Systolic:** Weaker correlation observed. This suggests that height does not significantly influence systolic blood pressure.

Relationship between BMI & Weight

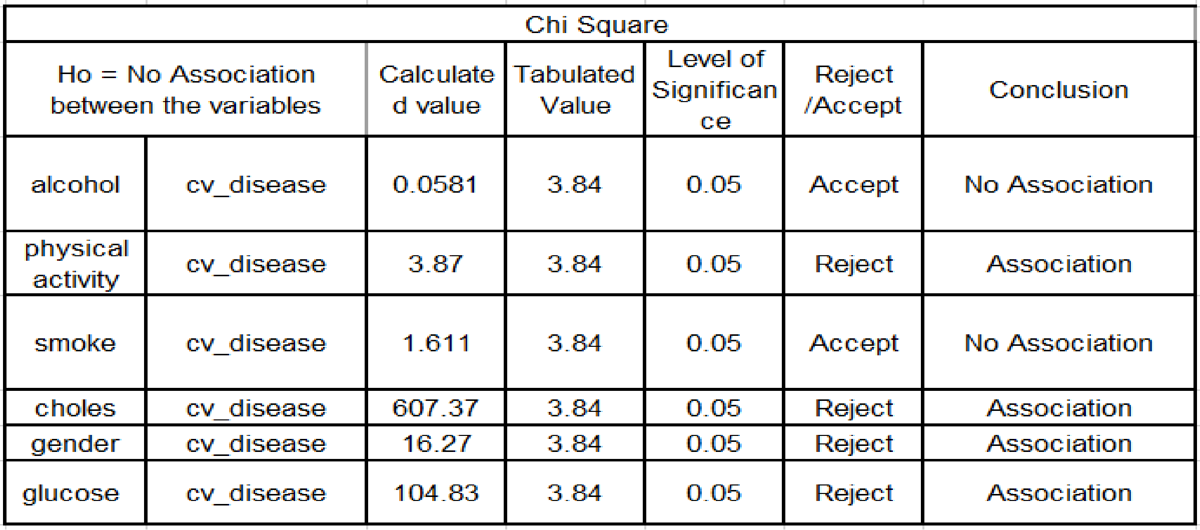
The scatter plot illustrates the relationship between Body Mass Index (BMI) and weight. As can be observed, there is a positive correlation between the two variables. This suggests that as an individual's weight increases, their BMI also tends to increase.



The plot appears to show that as BMI increases, weight also increases, which is expected as BMI is a measure that uses both height and weight to determine whether an individual's weight is healthy.



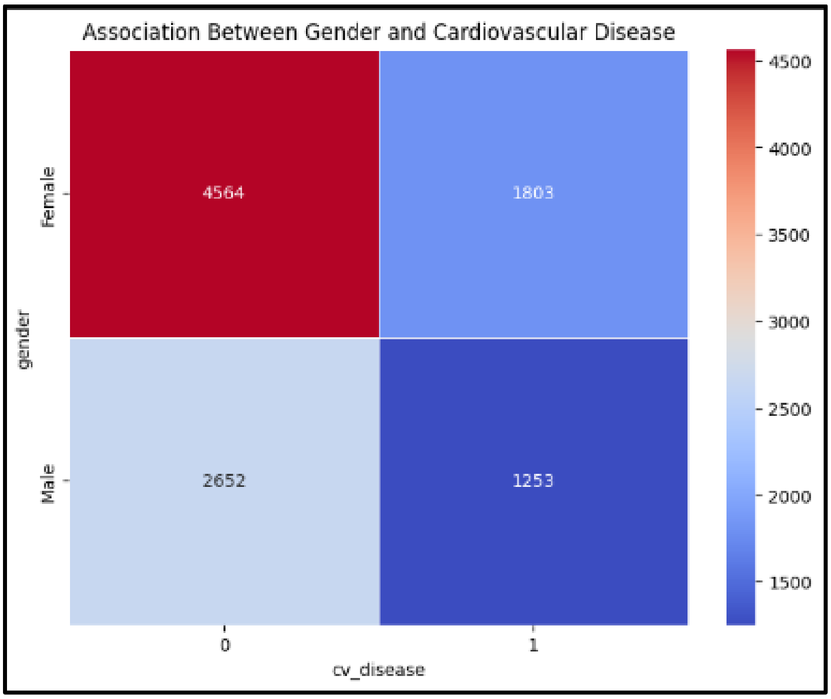
Chi-Square Test of Independence



Chi Square						
Ho = No Association between the variables		Calculated value	Tabulated Value	Level of Significance	Reject /Accept	Conclusion
alcohol	cv_disease	0.0581	3.84	0.05	Accept	No Association
physical activity	cv_disease	3.87	3.84	0.05	Reject	Association
smoke	cv_disease	1.611	3.84	0.05	Accept	No Association
choles	cv_disease	607.37	3.84	0.05	Reject	Association
gender	cv_disease	16.27	3.84	0.05	Reject	Association
glucose	cv_disease	104.83	3.84	0.05	Reject	Association

Gender Vs Disease

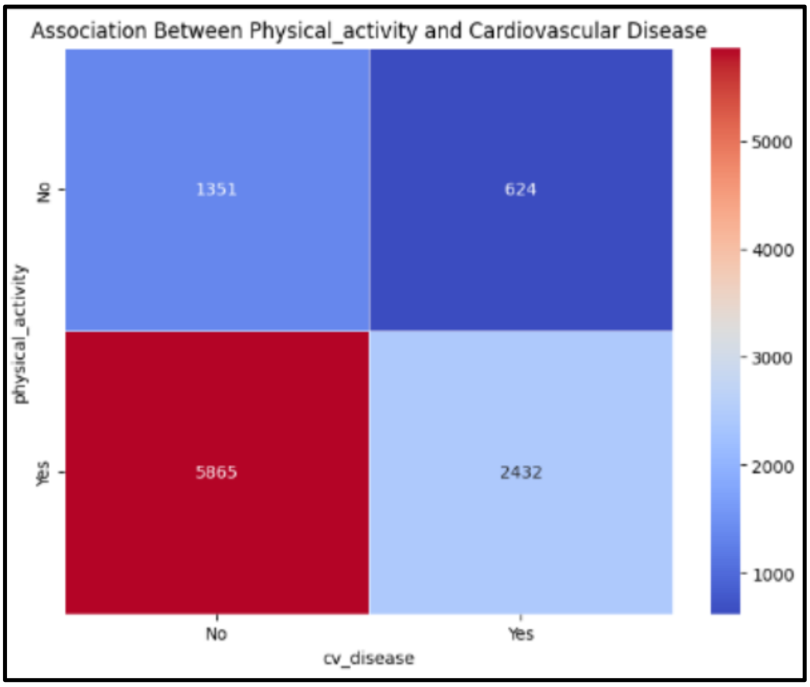
- There are more females with cardiovascular disease (4564) than males (2652).
- Conversely, there are more males without cardiovascular disease (1253) than females (1803).
- Females more likely to have cardiovascular disease.



Physical Activity Vs Disease

- There are 1351 cases of individuals with CVD who do not engage in physical activity and 624 cases of individuals with CVD who do engage in physical activity.
- Conversely, there are 5865 cases of individuals without CVD who do not engage in physical activity and 2432 cases of individuals without CVD who do engage in physical activity.

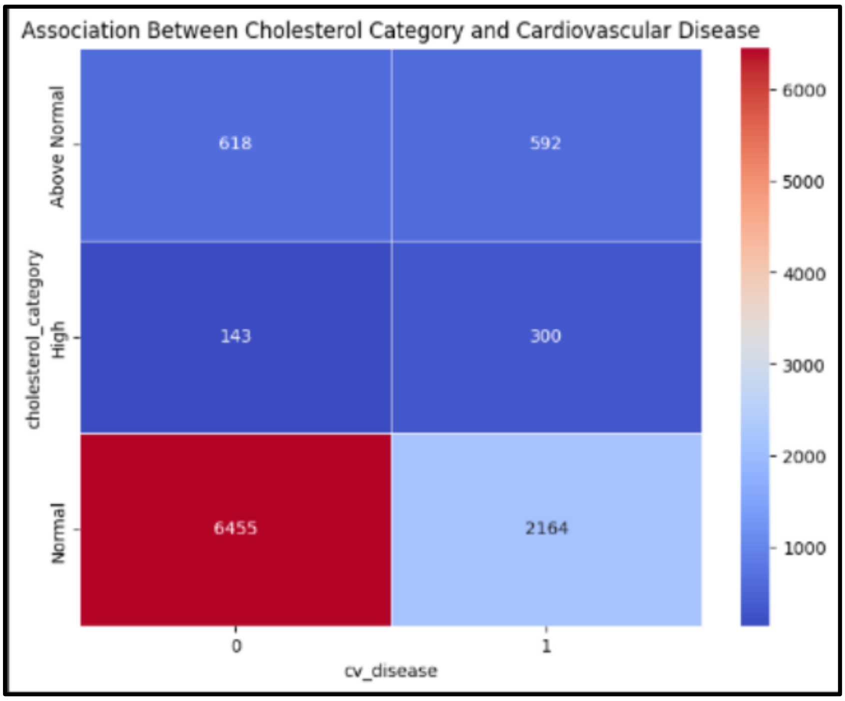
This suggests that in the population represented by this heatmap, individuals who do not engage in physical activity have a higher number of CVD cases compared to those who do engage in physical activity.



Cholesterol Vs Disease

- The highest number of people fall into the "Normal" cholesterol category and do not have cardiovascular disease.
- The lowest number of people are in the "High" cholesterol category and do not have cardiovascular disease.

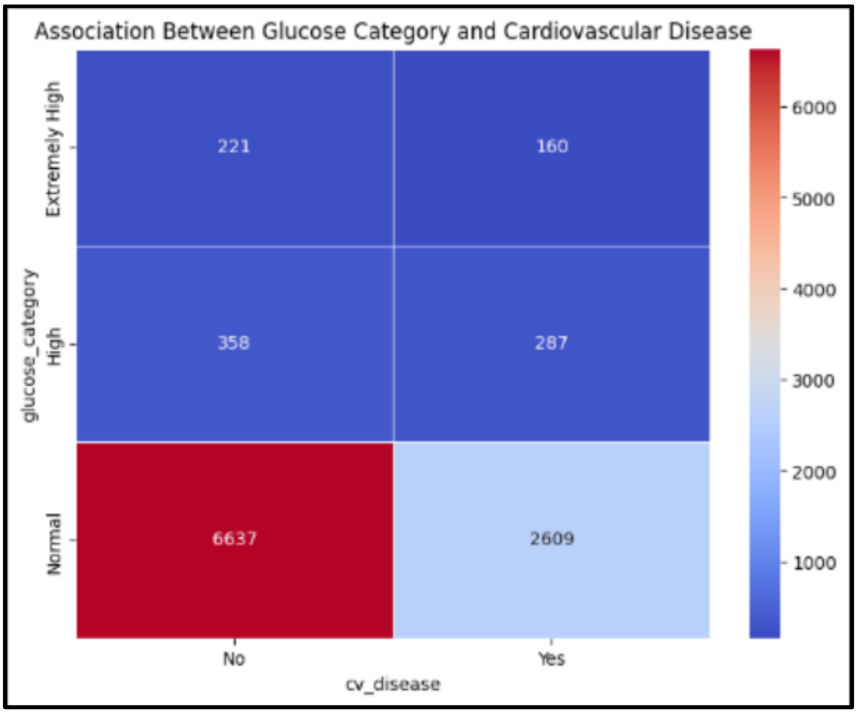
This suggests that there might be a correlation between high cholesterol levels and the presence of cardiovascular disease.



Glucose Vs Disease

- The highest number of people fall into the "Normal" glucose category and do not have cardiovascular disease.
- The lowest number of people are in the "Extremely High" glucose category and have cardiovascular disease.

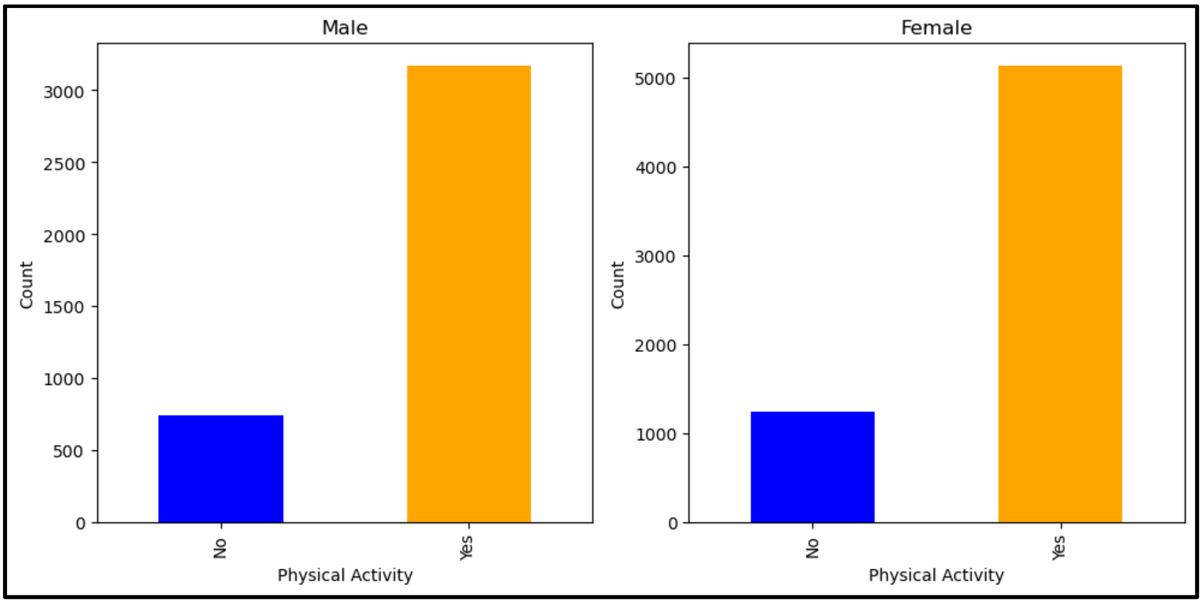
This suggests that there might be a correlation between high glucose levels and the presence of cardiovascular disease.



Gender Difference Analysis

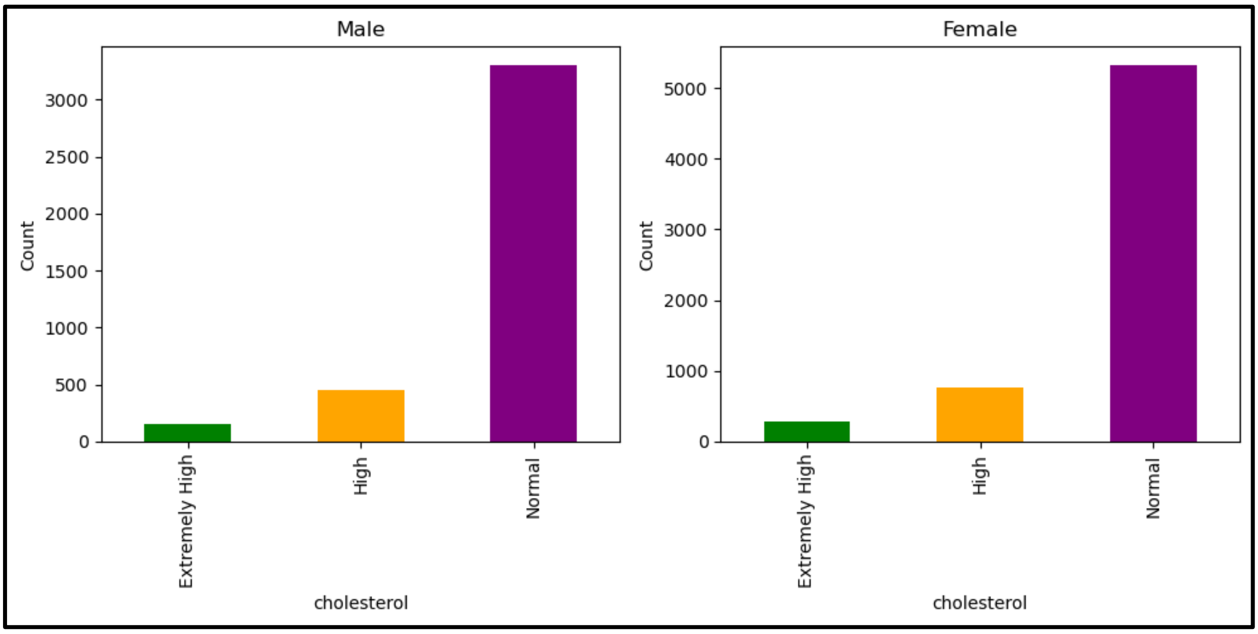
Physical Activity

- Bar graphs represent the **frequency of physical activity** among males and females (with cvd)
- Females appear to be **more physically active** than males in the represented population.



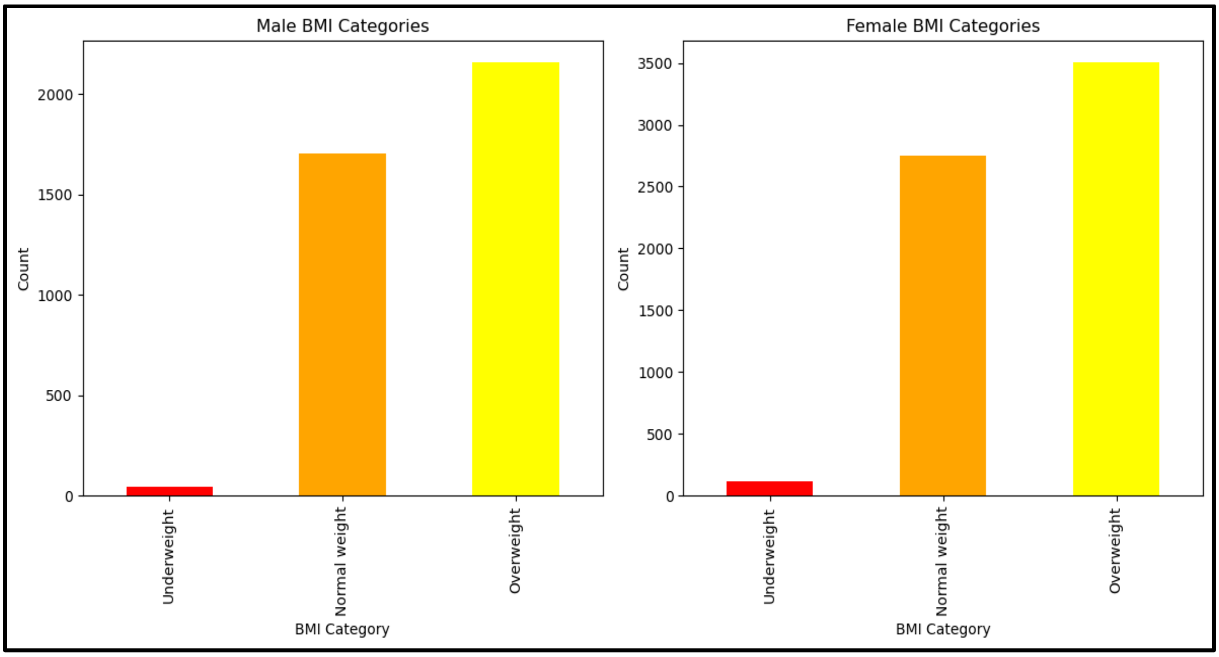
Cholesterol

- **Highest count** of patients is in the **“Normal” cholesterol level category**.
- As the cholesterol level increases to **“High” and “Extremely High”** number of patients with disease decreases.
- “Female” graph has a higher count in the “Extremely High” and “High” categories than the “Male” graph.
- Higher proportion of Females have “High” and “Extremely High” cholesterol levels compared to males.



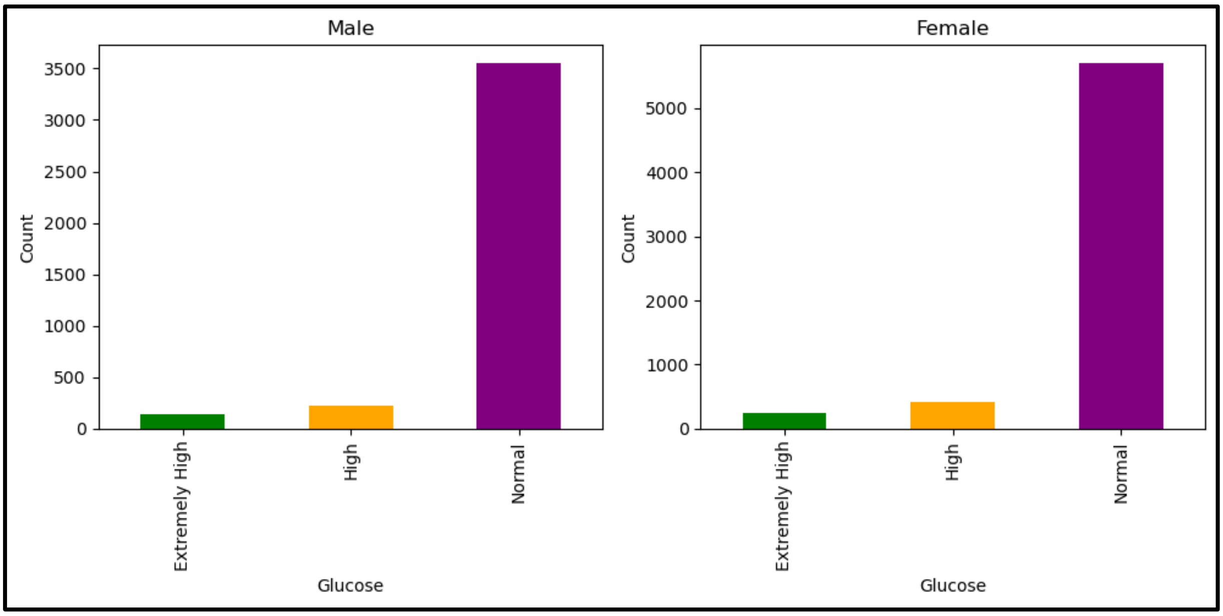
Body Mass Index (BMI)

- Bar graphs represent **frequency of different BMI categories** among males and females.
- Females have a **higher count** in all BMI categories - “Underweight”, “Normal Weight”, and “Overweight”.
- Suggests that the **total number of females with cvd** represented in these graphs is **higher** than the total number of males with cvd.

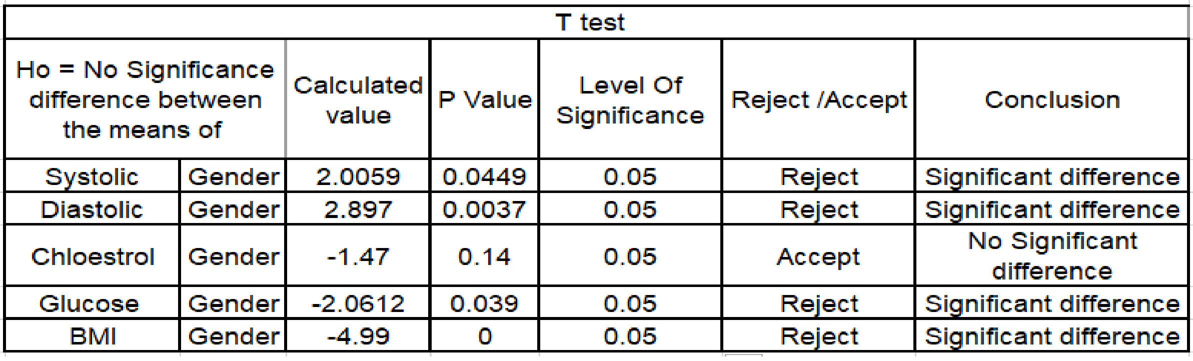


Glucose

- The chart compares glucose levels in **males** and **females**, with females having higher levels.
- The glucose levels are categorized as “**Normal**”, “**High**”, and “**Extremely High**”, with females outnumbering males in all categories.
- This could indicate a **higher risk** of conditions like diabetes among females in the represented population.



Hypothesis Testing(T-test Method)



T test						
Ho = No Significance difference between the means of		Calculated value	P Value	Level Of Significance	Reject /Accept	Conclusion
Systolic	Gender	2.0059	0.0449	0.05	Reject	Significant difference
Diastolic	Gender	2.897	0.0037	0.05	Reject	Significant difference
Chloestrol	Gender	-1.47	0.14	0.05	Accept	No Significant difference
Glucose	Gender	-2.0612	0.039	0.05	Reject	Significant difference
BMI	Gender	-4.99	0	0.05	Reject	Significant difference

THE HOSPITAL DATASET

By selecting the parameters that were common in our original data and hospital data, we ran t-test on the following hypothesis:

H0: stress average of male and female is equal
H1: stress average of male and female not is equal
Alpha: 0.05
Test statistic: t=1.31
P value: 0.2
Conclusion: Reject H0. Association present

H0: Mean of High Blood pressure of male and female is equal
H1: Mean of High Blood pressure of male and female is not equal
Alpha: 0.05
Test statistic: -0.17
P value: 0.86
Conclusion: Accept H0. Hence no Association

H0: Mean number of Alcoholics is equal in Male and Female
H1: Mean number of Alcoholics is not equal in Male and Female
Alpha: 0.05
Test statistic: 1.794815
Conclusion: Reject H0. Association present

SUMMARY & CONCLUSION

The research project, “Navigating Cardiovascular Risk Factors in Adults: A Comprehensive Analysis”, has provided valuable insights into the prevalence and risk factors of cardiovascular diseases (CVDs) in young adults. The analysis involved a comprehensive approach, utilizing descriptive statistics, regression, correlation, chi-square tests, and t-tests.

The study found significant associations between CVDs and various factors such as age, gender, height, weight, systolic and diastolic blood pressure, cholesterol, glucose, smoking status, alcohol intake, physical activity, and Body Mass Index (BMI). The most prominent factors we found to be Physical Activity, Cholesterol, Gender, Glucose Levels, Stress and Alcohol. These findings suggest that individuals who are older, male, taller, heavier, have higher blood pressure levels, have higher cholesterol levels, smoke, and are less physically active are at increased risk of CVDs.

In addition to the original dataset, a real-life example dataset was included in the analysis. This allowed for a comparison of the results obtained from the online dataset and the real-life hospital dataset. The t-tests conducted revealed relations between means of variates, providing further depth to the analysis.

The study also delved into the distribution of various factors such as age, gender, height, weight, and cholesterol levels among the population. The findings from these analyses provided a clearer picture of the demographic and health characteristics of the population under study.

Furthermore, the research project conducted a correlation analysis, revealing strong positive correlations between weight and BMI, and between diastolic and systolic blood pressure. Weaker correlations were observed between height and BMI, and between height and systolic blood pressure.

The chi-square tests of independence conducted provided insights into the relationships between gender, physical activity, cholesterol levels, glucose levels, and CVDs. The results suggested that females are more likely to have CVDs and that individuals who do not engage in physical activity have a higher number of CVD cases compared to those who do engage in physical activity.

In conclusion, the research project has made significant strides in understanding the risk factors associated with CVDs in young adults. The findings from this study could potentially guide preventive strategies and health policies, thereby reducing the burden of CVDs in young adults. However, it is important to note that while strides have been made in understanding these risk factors, there is still much work to be done. Future research could explore additional risk factors and their impact on CVDs, further enhancing our understanding of this pressing health issue.

LIMITATIONS & FUTURE WORK

Our study serves as a stepping stone for future research in this critical area of public health. The findings from our research have the potential to guide preventive strategies and shape health policies, thereby reducing the burden of cardiovascular diseases in young adults. This is particularly significant given the rising incidence of cardiovascular diseases in this demographic.

One of the key areas for future research is the exploration of additional risk factors that could contribute to cardiovascular diseases in young adults. While our study has identified several key risk factors, there are likely more factors that could be influencing the prevalence of these diseases in this age group. These could include genetic factors, dietary habits, stress levels, and other lifestyle factors. A comprehensive understanding of these risk factors could lead to more effective prevention strategies.

Furthermore, future research could also focus on the development of intervention programs tailored specifically for young adults. These programs could leverage the findings from our study and others to provide targeted support and resources for young adults at risk of developing cardiovascular diseases. This could include education programs, lifestyle modification support, and early screening initiatives.

Additionally, there is a need for longitudinal studies to track the health outcomes of young adults over time. Such studies could provide valuable insights into the long-term impacts of cardiovascular diseases and the effectiveness of various prevention and treatment strategies.

In future work, it would be beneficial to apply logistic regression on the identified parameters. Logistic regression is a statistical model that can predict the probability of certain outcomes, such as the presence or absence of cardiovascular disease.

By applying logistic regression, we could potentially enhance the predictive power of our analysis and provide more nuanced insights into the risk factors for cardiovascular disease in young adults. This approach could also allow us to control for confounding variables and assess the independent effects of each predictor on the outcome. Therefore, incorporating logistic regression in future analyses could significantly enrich our understanding of cardiovascular disease in young adults.

Finally, future research could also explore the role of technology in preventing cardiovascular diseases in young adults. With the rise of digital health technologies, there are exciting opportunities to leverage these tools for disease prevention and management.

In conclusion, while our study has made a significant contribution to the field of cardiovascular health in young adults, there is still much work to be done. We look forward to seeing how future research builds upon our findings to further advance our understanding of cardiovascular diseases in young adults and develop effective strategies to combat this pressing health issue.

REFERENCES

P. Uniyal, “Raju Srivastava to Sidharth Shukla: 10 celebs who died of heart attack,” *Hindustan Times*, Sep. 21, 2022. Available: https://www.hindustantimes.com/lifestyle/health/sidharth-shukla-to-kk-10-celebs-who-died-of-heart-attack-101660140071975.html

F. Generali, “Hard stat make heart diseases a concern in india,” *Future Generali India Life Insurance*, Dec. 17, 2021. Available: https://life.futuregenerali.in/life-insurance-made-simple/cancer-heart-critical-illness-insurance/hard-start-make-heart-diseases-a-concern-in-india

L. Desk, “World Heart Day 2022: Simple habits to keep your heart healthy,” *The Indian Express*, Sep. 29, 2022. Available: https://indianexpress.com/article/lifestyle/health/world-heart-day-facts-tips-care-8177287/

M. R. Amini, F. Zayeri, and M. Salehi, “Trend analysis of cardiovascular disease mortality, incidence, and mortality-to-incidence ratio: results from global burden of disease study 2017,” *BMC Public Health*, vol. 21, no. 1, Feb. 2021, doi: 10.1186/s12889-021-10429-0. Available: https://doi.org/10.1186/s12889-021-10429-0

S. Ara, “A Literature review of cardiovascular disease management programs in managed care populations,” *Journal of Managed Care Pharmacy*, vol. 10, no. 4, pp. 326–344, Jul. 2004, doi: 10.18553/jmcp.2004.10.4.326. Available: https://doi.org/10.18553/jmcp.2004.10.4.326

J. Vos, “Cardiovascular disease and meaning in life: A systematic literature review and conceptual model,” *Palliative & Supportive Care*, vol. 19, no. 3, pp. 367–376, May 2021, doi: 10.1017/s1478951520001261. Available: https://doi.org/10.1017/s1478951520001261

F. Masaebi, M. Salehi, M. Kazemi, N. Vahabi, M. A. Looha, and F. Zayeri, “Trend analysis of disability adjusted life years due to cardiovascular diseases: results from the global burden of disease study 2019,” *BMC Public Health*, vol. 21, no. 1, Jun. 2021, doi: 10.1186/s12889-021-11348-w. Available: https://doi.org/10.1186/s12889-021-11348-w

D. Munblit *et al.*, “Studying the post-COVID-19 condition: research challenges, strategies, and importance of Core Outcome Set development,” *BMC Medicine*, vol. 20, no. 1, Feb. 2022, doi: 10.1186/s12916-021-02222-y. Available: https://doi.org/10.1186/s12916-021-02222-y

“BMC Public Health,” *BioMed Central*, Nov. 20, 2023. Available: https://bmcpublichealth.biomedcentral.com/articles

“Palliative &amp; Supportive Care | Cambridge Core,” *Cambridge Core*. Available: https://www.cambridge.org/core/journals/palliative-and-supportive-care

V. Sun and M. Bakitas, “Palliative and supportive care: ...End of the beginning,” *Western Journal of Nursing Research*, vol. 41, no. 10, pp. 1343–1346, Jul. 2019, doi: 10.1177/0193945919861017. Available: https://doi.org/10.1177/0193945919861017

A. Mehta, S. R. Cohen, and L. Chan, “Palliative care: A need for a family systems approach,” *Palliative & Supportive Care*, vol. 7, no. 2, pp. 235–243, Jun. 2009, doi: 10.1017/s1478951509000303. Available: https://doi.org/10.1017/s1478951509000303

“Facebook.” Available: https://www.facebook.com/cambridge.medicine.healthsciences

A. Mehta, S. R. Cohen, and L. Chan, “Palliative care: A need for a family systems approach,” *Palliative & Supportive Care*, vol. 7, no. 2, pp. 235–243, Jun. 2009, doi: 10.1017/s1478951509000303. Available: https://doi.org/10.1017/s1478951509000303

S. Ara, “A Literature review of cardiovascular disease management programs in managed care populations,” *Journal of Managed Care Pharmacy*, vol. 10, no. 4, pp. 326–344, Jul. 2004, doi: 10.18553/jmcp.2004.10.4.326. Available: https://doi.org/10.18553/jmcp.2004.10.4.326

E. O. Meltzer, J. Szwarcberg, and M. W. Pill, “Allergic rhinitis, asthma, and rhinosinusitis: Diseases of the Integrated Airway,” *Journal of Managed Care Pharmacy*, vol. 10, no. 4, pp. 310–317, Jul. 2004, doi: 10.18553/jmcp.2004.10.4.310. Available: https://doi.org/10.18553/jmcp.2004.10.4.310

Europe PMC, “Europe PMC.” Available: https://europepmc.org/article/MED/30917076

Europe PMC, “Europe PMC.” Available: https://europepmc.org/article/MED/15548123

B. Fordham *et al.*, “The evidence for cognitive behavioural therapy in any condition, population or context: a meta-review of systematic reviews and panoramic meta-analysis,” *Psychological Medicine*, vol. 51, no. 1, pp. 21–29, Jan. 2021, doi: 10.1017/s0033291720005292. Available: https://doi.org/10.1017/s0033291720005292

A. S. Masten, K. Best, and N. Garmezy, “Resilience and development: Contributions from the study of children who overcome adversity,” *Development and Psychopathology*, vol. 2, no. 4, pp. 425–444, Oct. 1990, doi: 10.1017/s0954579400005812. Available: https://doi.org/10.1017/s0954579400005812

A. L. Barthel, A. C. Hay, S. N. Doan, and S. G. Hofmann, “Interpersonal Emotion Regulation: A review of social and developmental components,” *Behaviour Change*, vol. 35, no. 4, pp. 203–216, Oct. 2018, doi: 10.1017/bec.2018.19. Available: https://doi.org/10.1017/bec.2018.19

B. Fordham *et al.*, “The evidence for cognitive behavioural therapy in any condition, population or context: a meta-review of systematic reviews and panoramic meta-analysis,” *Psychological Medicine*, vol. 51, no. 1, pp. 21–29, Jan. 2021, doi: 10.1017/s0033291720005292. Available: https://doi.org/10.1017/s0033291720005292

“CC BY 4.0 Deed | Attribution 4.0 International | Creative Commons.” Available: http://creativecommons.org/licenses/by/4.0/