School of Business

THE GEORGE WASHINGTON UNIVERSITY

Identifying Key Risk Factors for Coronary Heart Disease: A Data-Driven Approach

Visualization for Analytics (DNSC 6323)

By

Manali Sudhir Chavaj G43827852

Table of Contents

1. Introduction

Heart disease is a major health concern, requiring data-driven insights to improve prevention and treatment. This report analyzes the HEART dataset, focusing on four hypotheses: (1) the correlation between weight and cholesterol, (2) obesity differences between men and women, (3) smoking habits and cholesterol levels by gender, and (4) the relationship between cholesterol and blood pressure. The report determines whether these hypotheses are true or false. Additionally, it explores characteristics of coronary heart disease (CHD) patients to identify potential causes. The insights gained will aid healthcare professionals in understanding patient trends, improving treatment strategies, and enhancing preventive care.

2. Methodology

This analysis used the HEART dataset in SAS Viya for Learners, containing 5,209 records with 17 columns, including 7 categorical and 10 numerical variables. Statistical techniques and visualizations such as box plots, bar charts, and scatter plots were applied to examine relationships between weight, cholesterol, smoking, and blood pressure. Additionally, patient characteristics related to coronary heart disease (CHD) were analyzed to identify risk factors, supporting improved healthcare strategies and prevention efforts.

3. Data Exploration and Challenges

- **Handling Missing Data**:
    - A major issue in the dataset is the presence of missing values in crucial fields like "AgeCHDdiag," "DeathCause," "AgeAtDeath," and "Smoking." These missing data points can significantly affect the reliability of the results, introduce potential bias, and undermine the validity of any conclusions drawn. To mitigate this issue, missing values need to be addressed either by imputing them using appropriate methods or by removing incomplete records from the dataset to preserve the integrity of the analysis.

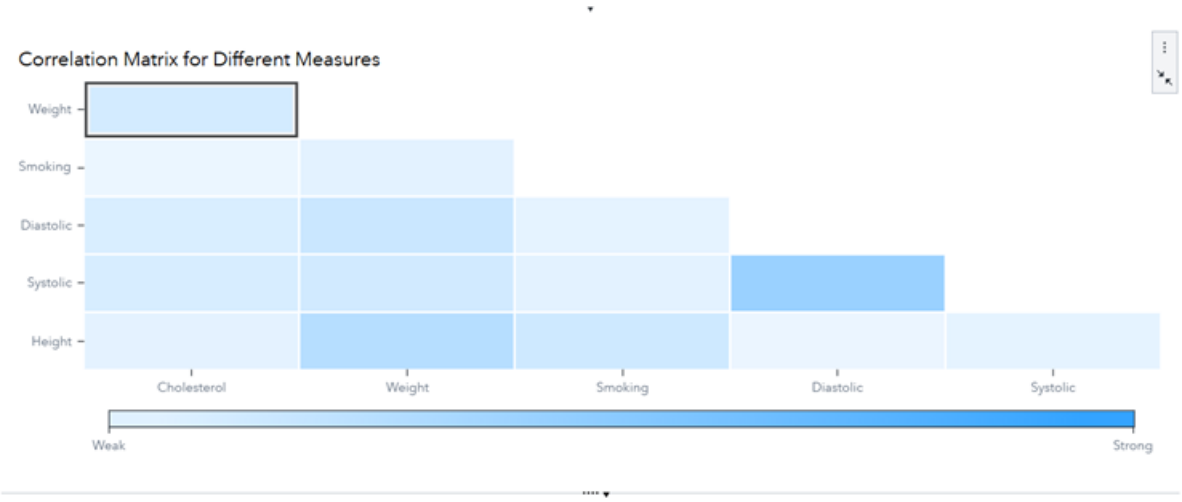
- **Outlier Detection and Lack of Documentation**:
    - Some variables, such as "Weight" and "Height," contain anomalous values that may be considered outliers, which could distort the results of the study. Detecting and managing these outliers is crucial to ensure accurate statistical analysis. Additionally, the dataset lacks detailed explanations for several columns, making it challenging to interpret the meaning and relevance of some variables. Developing a comprehensive data glossary to define each variable would help clarify the dataset and enhance its usability for accurate analysis.

- **Assessing Variable Relevance**:
    - It is essential to evaluate the relevance of each variable in the dataset to ensure the focus remains on the most significant factors influencing coronary heart disease. Identifying and retaining the most impactful variables while removing irrelevant ones helps streamline the analysis process, increases the efficiency of the modeling, and ensures that the study's conclusions are based on the most pertinent and informative data.

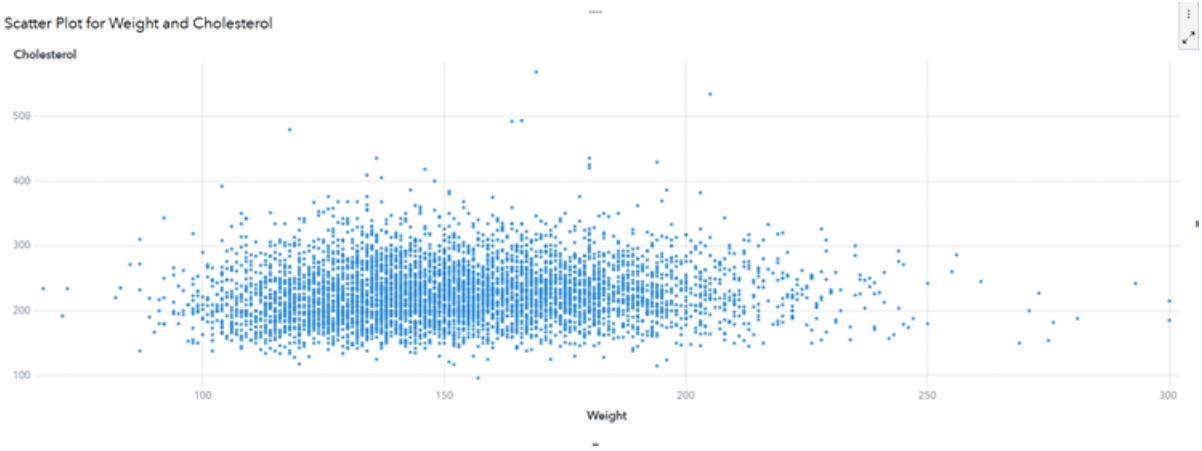
4. Hypotheses

4.1 H1: The weight and cholesterol levels are correlated

To examine the relationship between weight and cholesterol levels, a **correlation matrix** and **scatter plot** Is generated in SAS Viya.



X Axis	Y Axis	Correlation
Cholesterol	Smoking	-0.0118
Cholesterol	Weight	0.0724
Weight	Height	0.5174



Correlation Analysis:

- The correlation coefficient between weight and cholesterol is **0.0724**, indicating a very weak positive correlation. Which means we can't be reasonably sure that weight and cholesterol levels are correlated to each other.

Scatter Plot Interpretation:

- The scatter plot of cholesterol vs. weight showed a widely dispersed distribution with no clear upward or downward trend. If a strong correlation existed, the points would form a more distinct linear pattern, but their random spread further supports the weak correlation.

Conclusion:

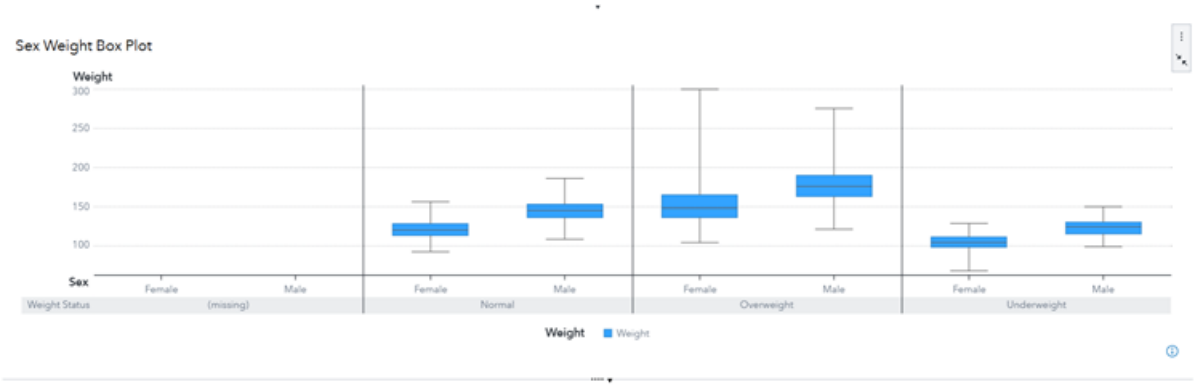
The results suggest that there is no meaningful correlation between weight and cholesterol levels in this dataset Other factors, such as diet, genetics, and lifestyle, may have a greater impact on cholesterol levels than weight alone.

Recommendation:

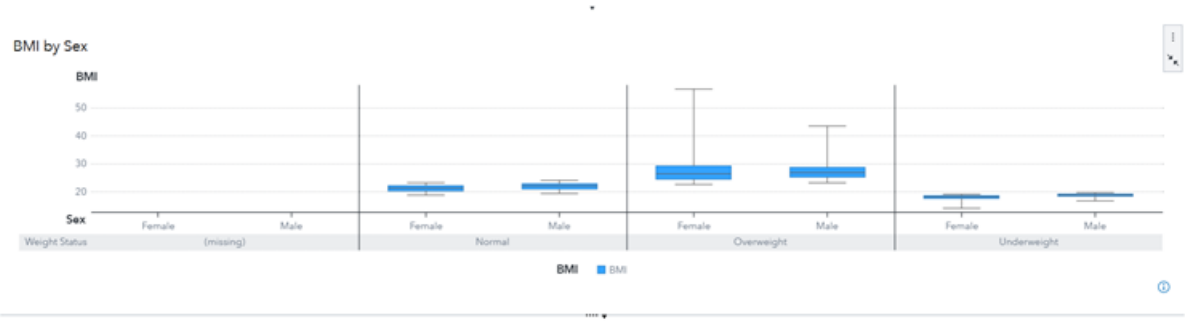
Create more personalized intervention strategies based on individual patient profiles, taking into account other factors that may be correlated with cholesterol levels, rather than relying solely on weight-based risk assessments.

4.2 H2: Men are usually more obese than women

To evaluate this we analyzed **weight and BMI distributions** (created a new Measure - Appendix 1) using box plots in SAS Viya.



Weight Status	Sex	Minimum	Lower Whisker	First Quartile	Average	Median	Third Quartile	Upper Whisker	Maximum	Std Dev	Count
Overweight	Male	121	121	163	177.74315277	176	190	276	276	21.52617321	1,643
Overweight	Female	104	104	136	152.81594127	148	165	300	300	23.922206855	1,907
Normal	Male	108	108	136	145	145	153	186	186	13.382376471	626



Weight Status	Sex	Minimum	Lower Whisker	First Quartile	Average ▾	Median	Third Quartile	Upper Whisker	Maximum	Std Dev	Count
Overweight	Female	22.68	22.68	24.66	27.67	26.57	29.39	56.68	56.68	4.18	1,905
Overweight	Male	23.35	23.35	25.39	27.45	27.02	28.85	43.60	43.60	2.72	1,642
Normal	Male	19.41	19.41	21.14	22.10	22.28	23.01	24.13	24.13	1.14	625
Normal	Female	18.86	18.86	20.47	21.36	21.47	22.31	23.43	23.43	1.13	846
Underweight	Male	16.83	16.83	18.59	18.83	18.99	19.34	19.80	19.80	0.73	65
Underweight	Female	14.12	14.12	17.81	18.09	18.37	18.78	19.22	19.22	0.97	116

Weight Distribution by Sex:

- Overweight males have an average weight of 177.74, whereas overweight females have an average weight of 152.81. The median weight is 176 for males and 148 for females, confirming that men in the overweight category tend to weigh more than women.

BMI Distribution by Sex:

- The average BMI for overweight females is 27.67, slightly higher than the average BMI for overweight males at 27.45. The median BMI is 26.57 for females and 27.02 for males, indicating that men generally have a slightly higher BMI in the overweight category.
- The maximum BMI recorded for overweight females is 56.68, whereas for overweight males, it reaches 43.60, suggesting that extreme obesity cases are more prevalent among women.

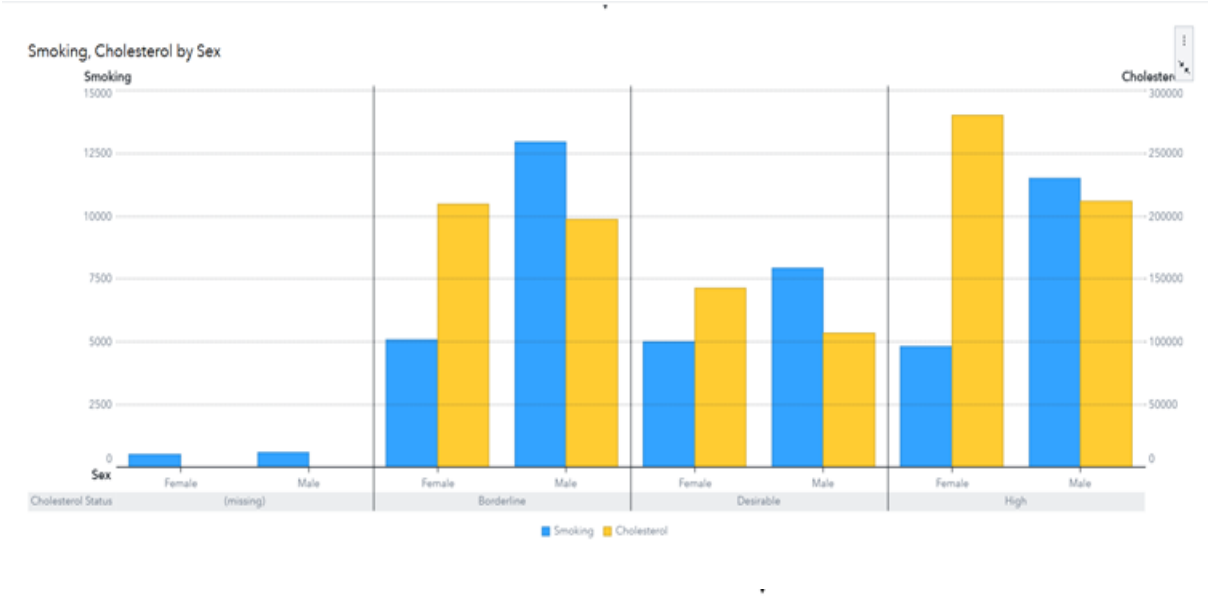
Conclusion:

The findings indicate that men in the overweight category tend to weigh more than women, but women have a slightly higher average BMI. Additionally, extreme obesity cases are more common among females, as reflected in the maximum BMI values.

Recommendation:

Healthcare programs should prioritize weight management for men and specialized obesity interventions for women, including personalized diet plans and medical treatments. Additionally, gender-specific awareness campaigns can help address obesity risks, encouraging proactive health management through targeted exercise programs, lifestyle modifications, and preventive healthcare strategies for both men and women.

4.3 H3: Women usually smoke less than men, but their cholesterol level is higher



Sex	Smoking	Cholesterol	Cholesterol Status
Female	4832	280829	High
Male	11533	211704	High
Female	5024	143142	Desirable
Male	7930	106611	Desirable
Female	5084	210004	Borderline
Male	12973	197760	Borderline
Female	502	.	
Male	575	.	

Smoking Patterns by Gender

- The data indicates that men have higher smoking rates than women across all cholesterol categories.
- For individuals with high cholesterol, male smokers (11,533) significantly outnumber female smokers (4,832).In the desirable cholesterol category, male smokers (7,930) also exceed female smokers (5,024).Similarly, in the borderline cholesterol category, male smokers (12,973) are more than female smokers (5,084).

Cholesterol Levels by Gender

- Women consistently exhibit higher cholesterol levels across all categories compared to men.
- Among individuals with high cholesterol, women recorded 280,829, whereas men had 211,704. For those with borderline cholesterol, women's levels were 210,004, compared to 197,760 for men. Even in the desirable cholesterol category, women had a cholesterol count of 143,142, while men had 106,611.

Conclusion:

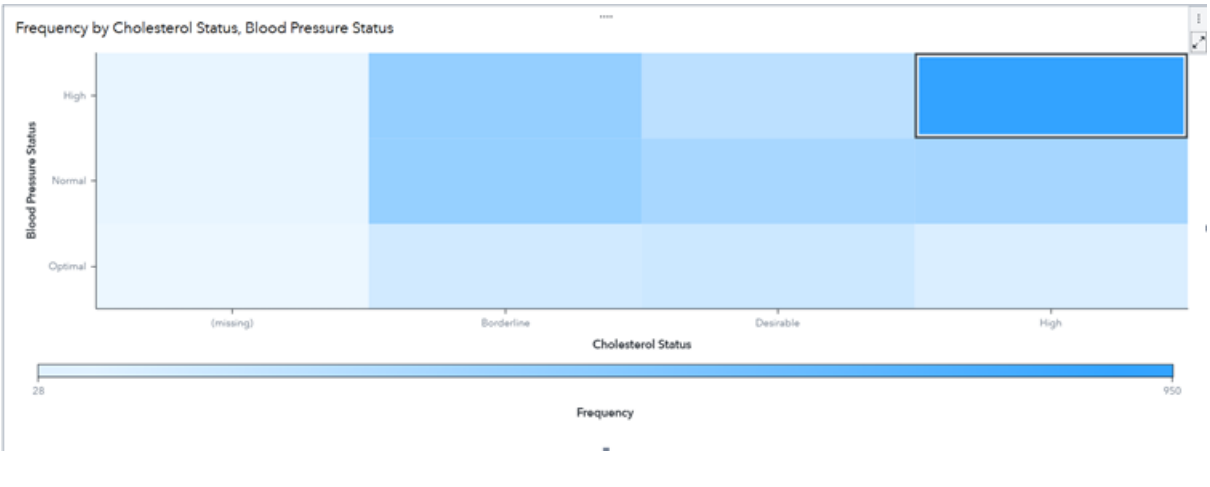
The hypothesis that women smoke less than men but have higher cholesterol levels is supported by the data. Women consistently show lower smoking rates across all cholesterol categories, while their cholesterol levels remain higher than those of men in every category.

Recommendation:

Healthcare organizations should implement anti-smoking programs targeting men, as they smoke more than women. Cholesterol management strategies for women should focus on dietary changes, exercise, and medications since they have higher cholesterol levels.

4.4 H4: The blood pressure is higher for people with higher cholesterol levels

To examine the relationship between cholesterol levels and blood pressure, a heatmap was used to visualize the frequency distribution between cholesterol status and blood pressure status.



Cholesterol Status	Blood Pressure Status	Frequency
High	High	950
Borderline	High	798
Borderline	Normal	793

Analysis:

- The highest frequency of individuals with high cholesterol also had high blood pressure (950 cases), suggesting a strong link.
- Borderline cholesterol levels were most frequently associated with high blood pressure (793 cases) but also had a significant number of cases with normal blood pressure.
- The pattern indicates that as cholesterol levels increase, individuals are more likely to have higher blood pressure, supporting the hypothesis.

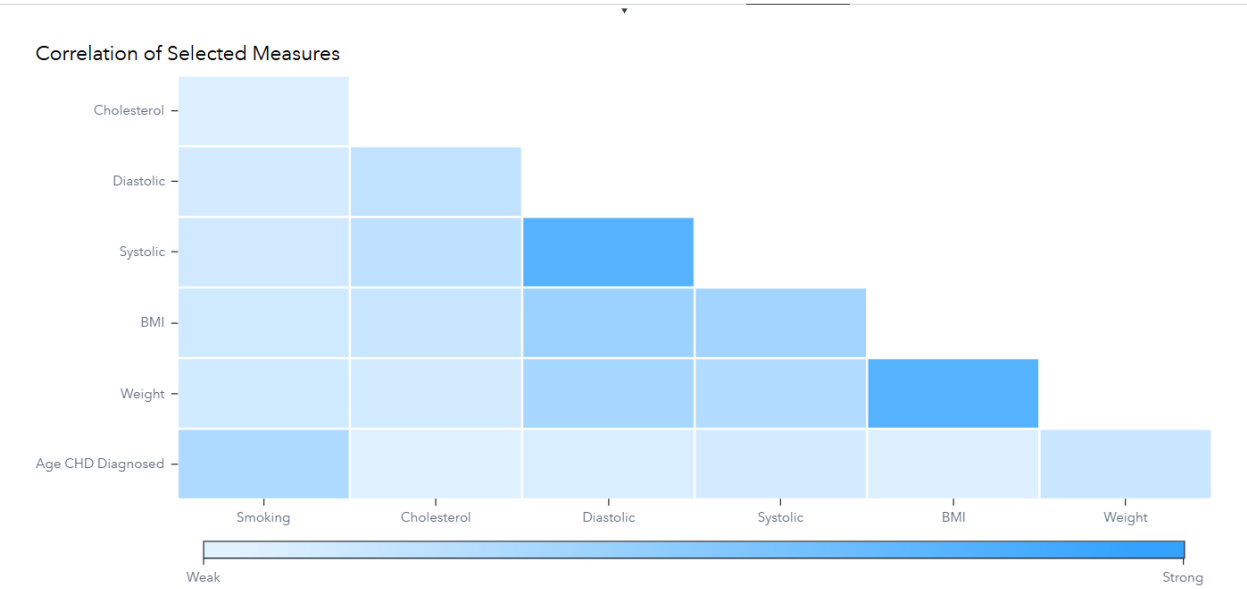
Conclusion:

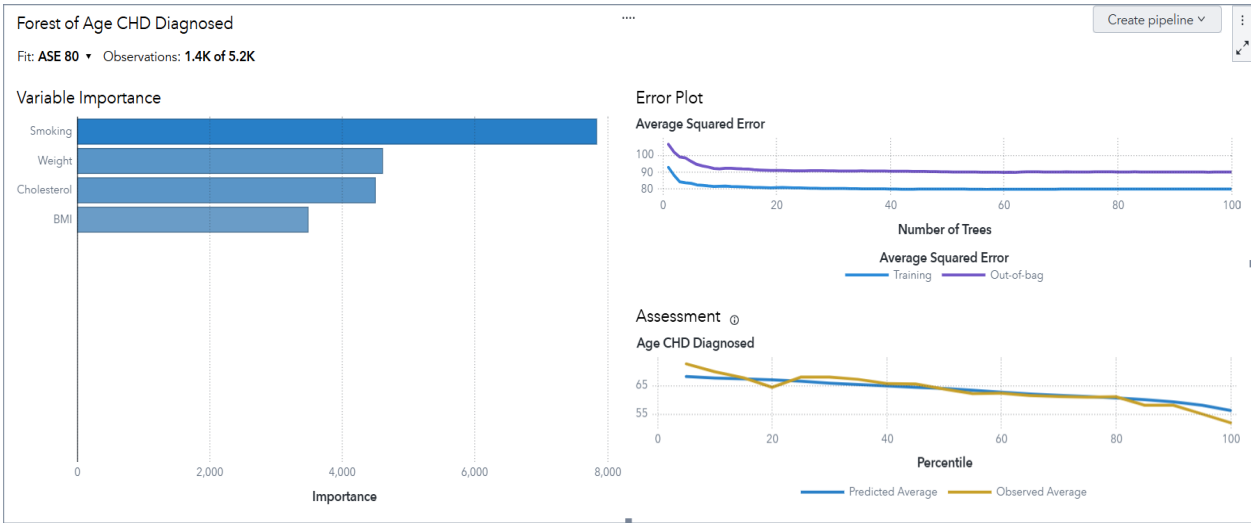
The findings confirm that individuals with higher cholesterol levels tend to have higher blood pressure. This highlights a potential risk factor for cardiovascular diseases, reinforcing the importance of monitoring both metrics together in medical assessments.

Recommendation:

Healthcare providers should implement dual screening programs for cholesterol and blood pressure to identify high-risk individuals early. Lifestyle interventions, including diet and exercise, should target reducing both metrics to prevent cardiovascular complications. Additionally, patients with high cholesterol must be monitored for hypertension and provided timely medical interventions, including necessary medications.

5. Analysis of CHD Risk Factors





The correlation matrix highlights significant relationships between various health measures. Key observations include:

- **Systolic blood pressure** shows a strong correlation with weight and BMI, emphasizing their role in cardiovascular strain.
- **Cholesterol and smoking** exhibit a weaker correlation, but their cumulative effect on CHD risk cannot be ignored.
- **Weight and BMI** demonstrate a moderate correlation, indicating obesity's contribution to heart disease.

The variable importance chart from the random forest model ranks **smoking** as the most critical factor influencing CHD diagnosis, followed by **weight, cholesterol, and BMI**. This suggests

that lifestyle habits, particularly smoking and weight management, are primary contributors to CHD risk.

Conclusion

1. **Smoking** is the most influential risk factor in predicting CHD diagnosis, reinforcing its critical role in heart disease development.
2. **Weight and BMI** significantly impact CHD, supporting the link between obesity and cardiovascular issues.
3. **Cholesterol levels**, although moderately important, still play a role in CHD onset.
4. **Systolic and diastolic blood pressure**, while not the top predictors, are correlated with key risk factors like weight and cholesterol.

Recommendations

1. **Smoking cessation programs** should be prioritized to reduce CHD risk significantly.
2. **Obesity management initiatives** through personalized diet plans and physical activity regimens must be implemented.
3. **Cholesterol screening and control measures**, including dietary changes and medication, should be strengthened.
4. **Regular blood pressure monitoring** should be encouraged, particularly for individuals with high BMI or cholesterol levels.
5. **Predictive analytics and machine learning models** should be used in healthcare systems to identify high-risk individuals for early intervention.

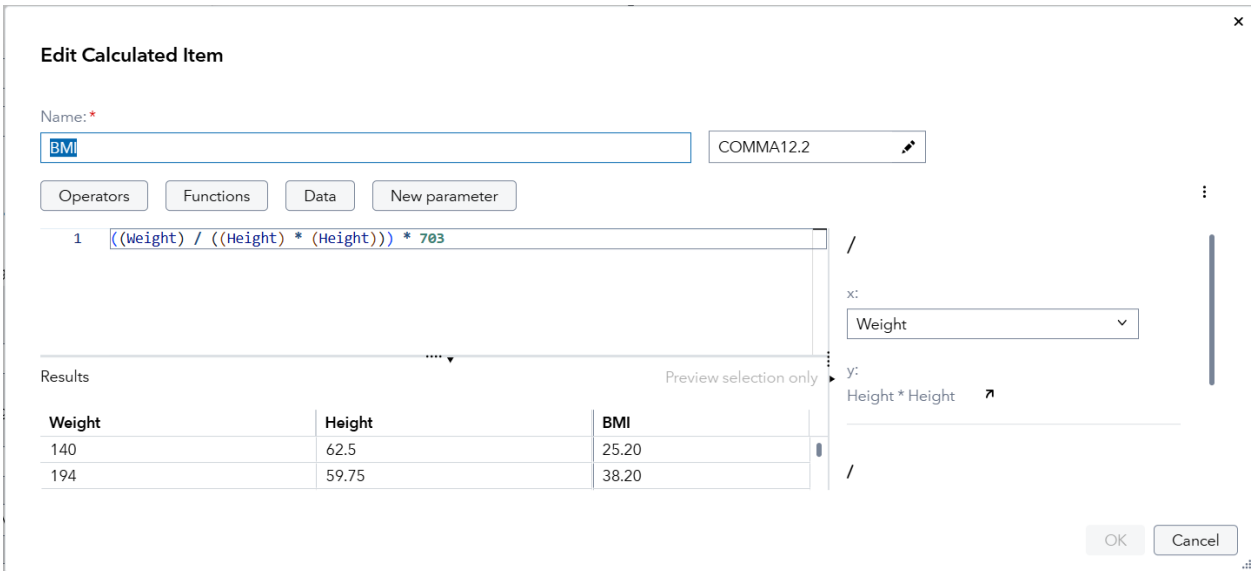
6. Conclusion

The analysis highlights key risk factors associated with coronary heart disease (CHD). High cholesterol and hypertension strongly correlate, reinforcing their role in CHD development. While men smoke more, women exhibit higher cholesterol levels, suggesting gender-specific risk factors. Obesity patterns differ, with extreme cases more common in women. Weight and cholesterol show a weak correlation, indicating that other factors, such as genetics and lifestyle, may have a greater impact. Overall, CHD risk is influenced by a combination of smoking, obesity, high cholesterol, and hypertension, emphasizing the need for early identification and targeted interventions to improve patient outcomes.

7. Recommendations

Healthcare strategies should focus on early detection and management of high cholesterol and hypertension through routine screenings. Gender-specific programs should address obesity concerns, encouraging lifestyle changes like exercise and balanced diets. Smoking cessation programs should primarily target men, while cholesterol management efforts should focus on women. A holistic approach combining personalized health interventions, dietary modifications, and medical treatments is crucial for CHD prevention. Additionally, integrating predictive analytics in healthcare can help identify high-risk individuals, enabling timely interventions and reducing the overall burden of coronary heart disease.

Appendix


Edit Calculated Item
Name: *
BMI
COMMA12.2
Operators
Functions
Data
New parameter
1 ((Weight) / ((Height) * (Height))) * 703
/
x:
Weight
y:
Height * Height
Results
Preview selection only
Weight
Height
BMI
140
62.5
25.20
194
59.75
38.20
/
OK
Cancel