Cardiovascular_Risk_Analysis_Report:Home_Work

Baodong Zhang

2025-01-12

Introduction

Cardiovascular diseases (CVD) are a leading cause of mortality worldwide (Organization 2021). This report investigates whether cardiovascular disease (variable ‘cardio‘) can be explained by other variables such as age, gender, glucose level, BMI, and lifestyle factors like smoking, alcohol consumption, and physical activity. The analysis is based on a dataset containing various health metrics.

Chapter 1: Data Preparation

Task 1: Transform the variables of the data set to appropriate data types and assign factor labels for the categorical variables.

Table 1: Data Overview

	variable	class	unique__values	example__values
id	id	integer	5000	24628, 66016, 36566, 30609, 53555
age	age	numeric	1753	58.68, 55.89, 60.11, 47.87, 52.16
gender	gender	factor	2	Male, Female
height	height	integer	61	159, 167, 169, 163, 165
weight	weight	numeric	115	59, 89, 78, 75, 73
ap__hi	ap__hi	integer	65	120, 140, 12, 110, 150
ap__lo	ap__lo	integer	54	80, 90, 79, 70, 69
cholesterol	cholesterol	factor	3	normal, above normal, well above normal
gluc	gluc	factor	3	normal, above normal, well above normal
smoke	smoke	factor	2	no, yes
alco	alco	factor	2	no, yes
active	active	factor	2	yes, no
cardio	cardio	factor	2	absent, present

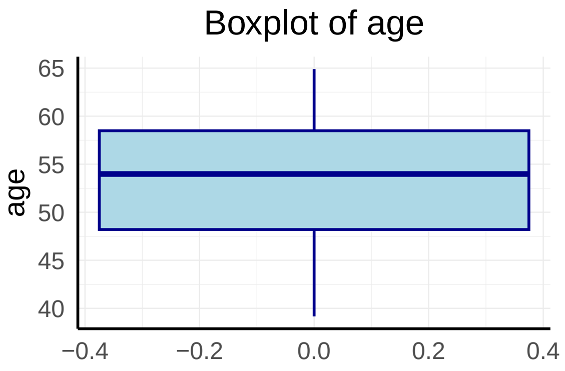
Chapter 2: Outlier Detection

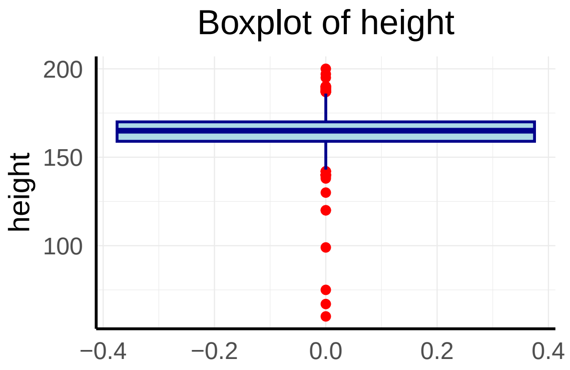
Task 2: Check the continuous variables for outliers and remove implausible values.

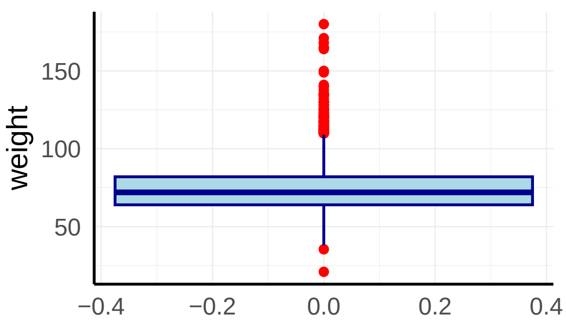
Table 2: Summary of Continuous Variables

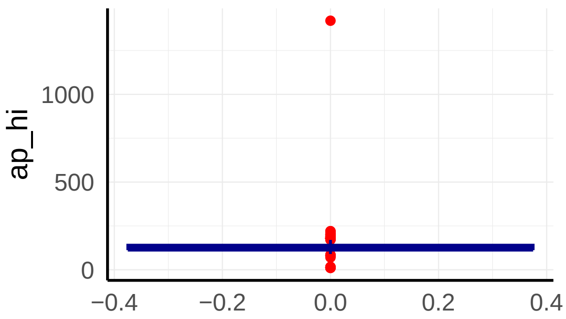
	age	height	weight	ap_hi	ap_lo
	Min. :39.16	Min. : 60.0	Min. : 21.00	Min. : 10.0	Min. : 40.0
	1st Qu.:48.20	1st Qu.:159.0	1st Qu.: 64.00	1st Qu.: 120.0	1st Qu.: 80.0
	Median :53.98	Median :165.0	Median : 72.00	Median : 120.0	Median : 80.0
	Mean :53.31	Mean :164.3	Mean : 74.01	Mean : 126.7	Mean : 96.1
	3rd Qu.:58.48	3rd Qu.:170.0	3rd Qu.: 82.00	3rd Qu.: 140.0	3rd Qu.: 90.0
	Max. :64.90	Max. :200.0	Max. :180.00	Max. :1420.0	Max. :8099.0

Boxplot before outliers are removed:






Boxplot of weight


Boxplot of ap_hi

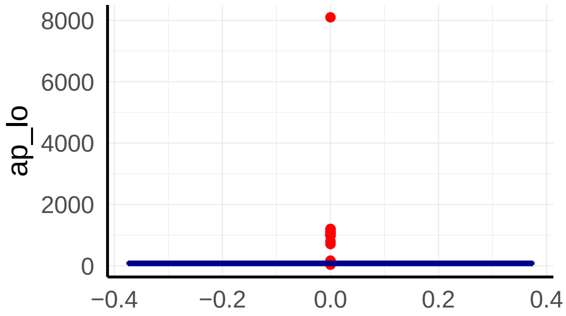
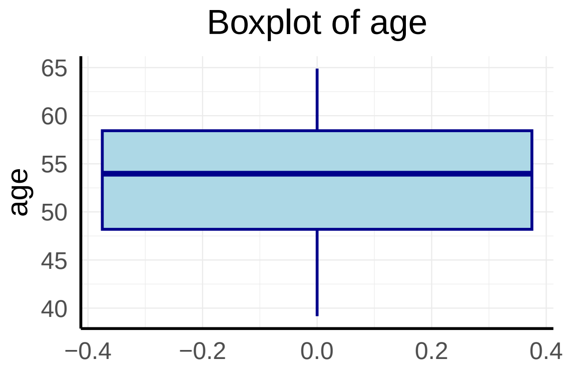
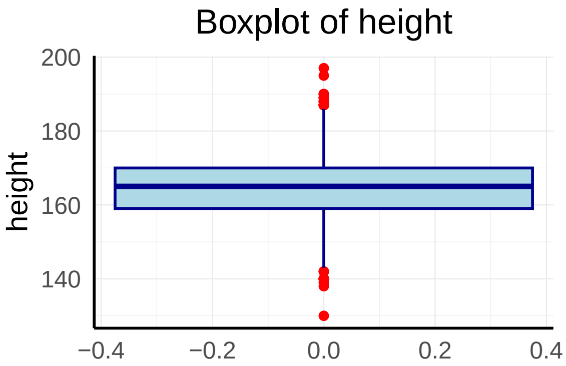

Boxplot of ap_lo

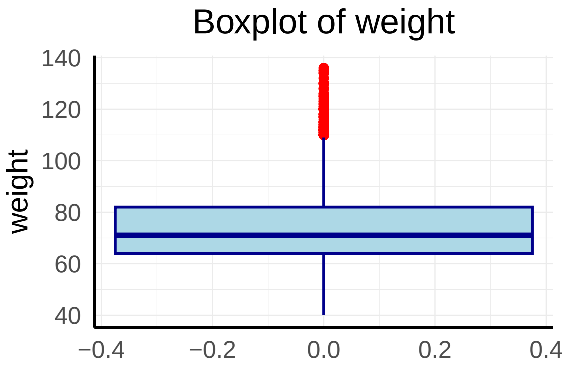
Table 3: Summary of Continuous Variables (outliers are removed)

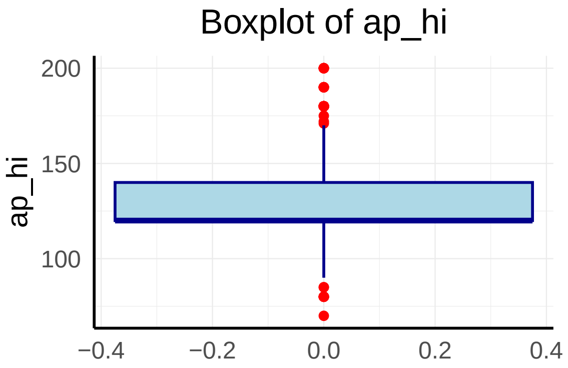
	age	height	weight	ap_hi	ap_lo
	Min. :39.16	Min. :130.0	Min. : 40.00	Min. : 70.0	Min. : 53.00
	1st Qu.:48.19	1st Qu.:159.0	1st Qu.: 64.00	1st Qu.:120.0	1st Qu.: 80.00
	Median :53.97	Median :165.0	Median : 71.00	Median :120.0	Median : 80.00
	Mean :53.30	Mean :164.4	Mean : 73.71	Mean :126.4	Mean : 81.25
	3rd Qu.:58.44	3rd Qu.:170.0	3rd Qu.: 82.00	3rd Qu.:140.0	3rd Qu.: 90.00
	Max. :64.90	Max. :197.0	Max. :136.00	Max. :200.0	Max. :120.00

Boxplot after outliers are removed

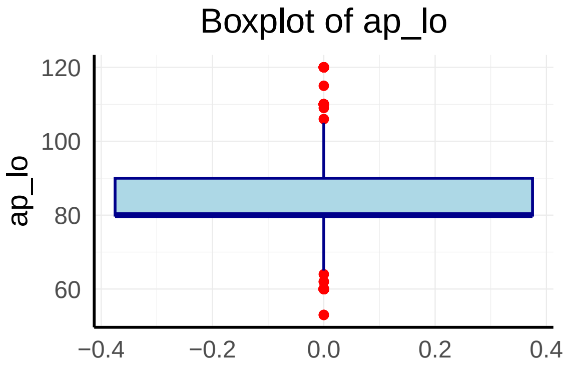

Boxplot of age


Boxplot of height


Boxplot of weight



Boxplot of ap_hi



Boxplot of ap_lo

Chapter 3: BMI Calculation

Task 3: Create a new variable BMI and provide a summary table for the variable BMI for both cardio groups.

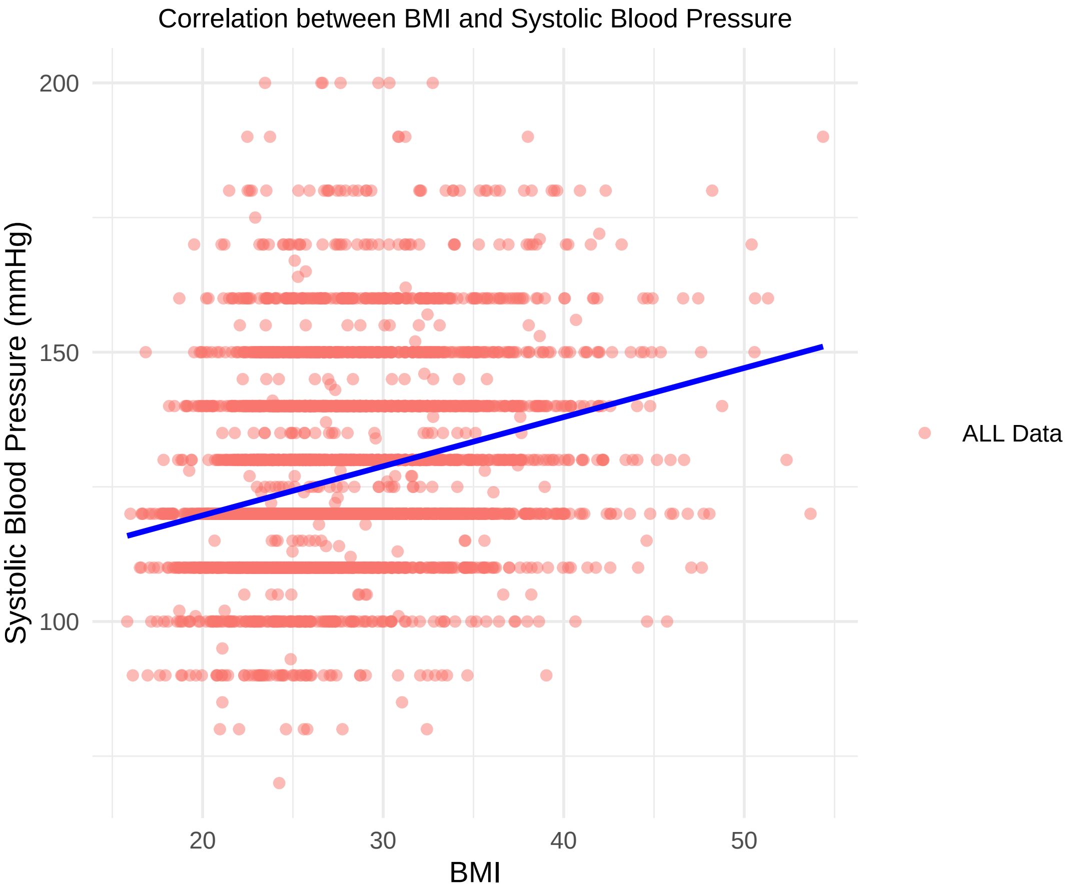
Table 4: Summary table for BMI grouped by cardio

cardio	count	mean_BMI	median_BMI	sd_BMI	min_BMI	max_BMI
absent	2495	26.34	25.34	4.62	15.82	50.41
present	2384	28.31	27.18	5.23	16.71	54.36

The two-sample t-test, under the assumptions of normality and independence of observations, indicated a significant difference in mean BMI between individuals without cardiovascular disease and those with the disease. Assuming equal variances, the test yielded a a p-value $1.7377518 \times 10^{-43'}$. This suggesting that individuals with CVD tend to have significantly higher BMI than those without.

Chapter 4: Correlation Between BMI and Systolic Blood Pressure

Task 4: How does the systolic blood pressure (SBP) and the BMI correlate to each other? Is there any difference between the two classes of cardiovascular disease?



Answer: The correlation coefficient between BMI and Systolic Blood Pressure (SBP) is 0.2789966, indicating a weak positive correlation. The p-value is $6.1991978 \times 10^{-88}$, showing that this weak positive correlation is statistically significant across the dataset.

To determine whether there is a difference between the two classes of cardiovascular disease, we will calculate the correlation coefficients for each class separately and conduct a Fisher's Z-test.

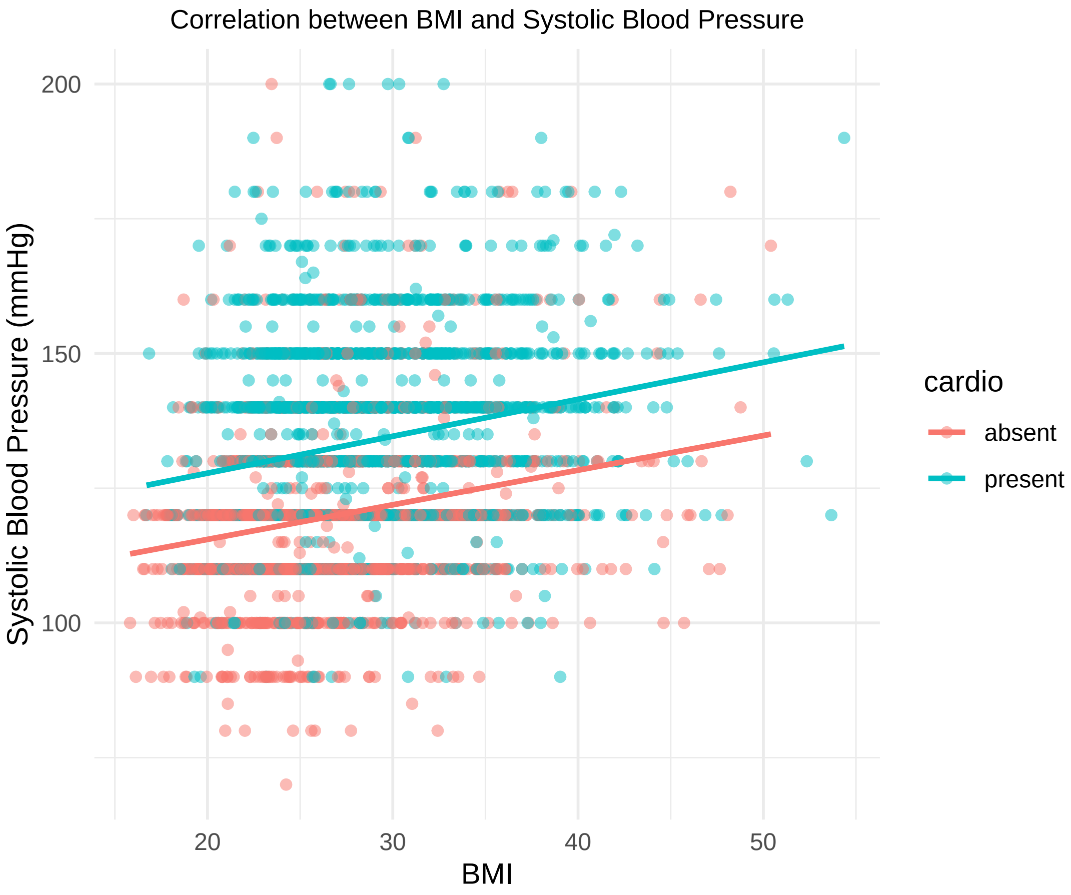


Table 5: Correlation between BMI and Systolic blood pressure grouped by cardio

cardio	Correlation
absent	0.2317877
present	0.2136453

Table 6: Fisher's Z-test for Correlations between BMI and Systolic blood pressure grouped by cardio

Metric	Value
Z_score	0.666130
P_value	0.505328

Based on the Fisher's Z-test, the Z score is 0.66613 and the P value is 0.505328. The correlation between BMI and Systolic Blood Pressure shows no statistically significant difference between the two AVD groups, suggesting the results may be due to random variation rather than a meaningful effect.

Chapter 5: Correlation between BMI and Diastolic blood pressure.

Task 5: Answer the same question for the diastolic blood pressure.

The Correlation Between BMI and Diastolic blood pressure is 0.2541928 and p-value is $8.1781399 \times 10^{-73}$.

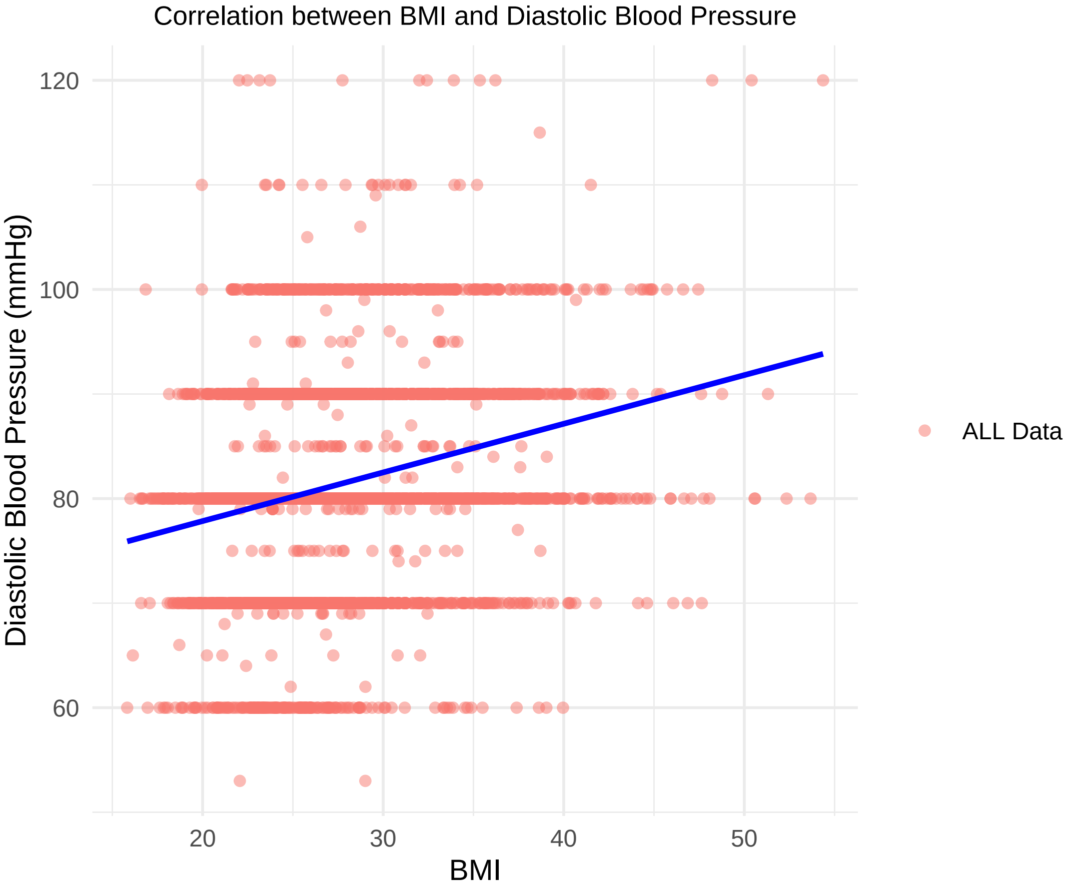


Table 7: Correlation between BMI and Diastolic blood pressure grouped by cardio

cardio	Correlation
absent	0.2233765
present	0.1869794

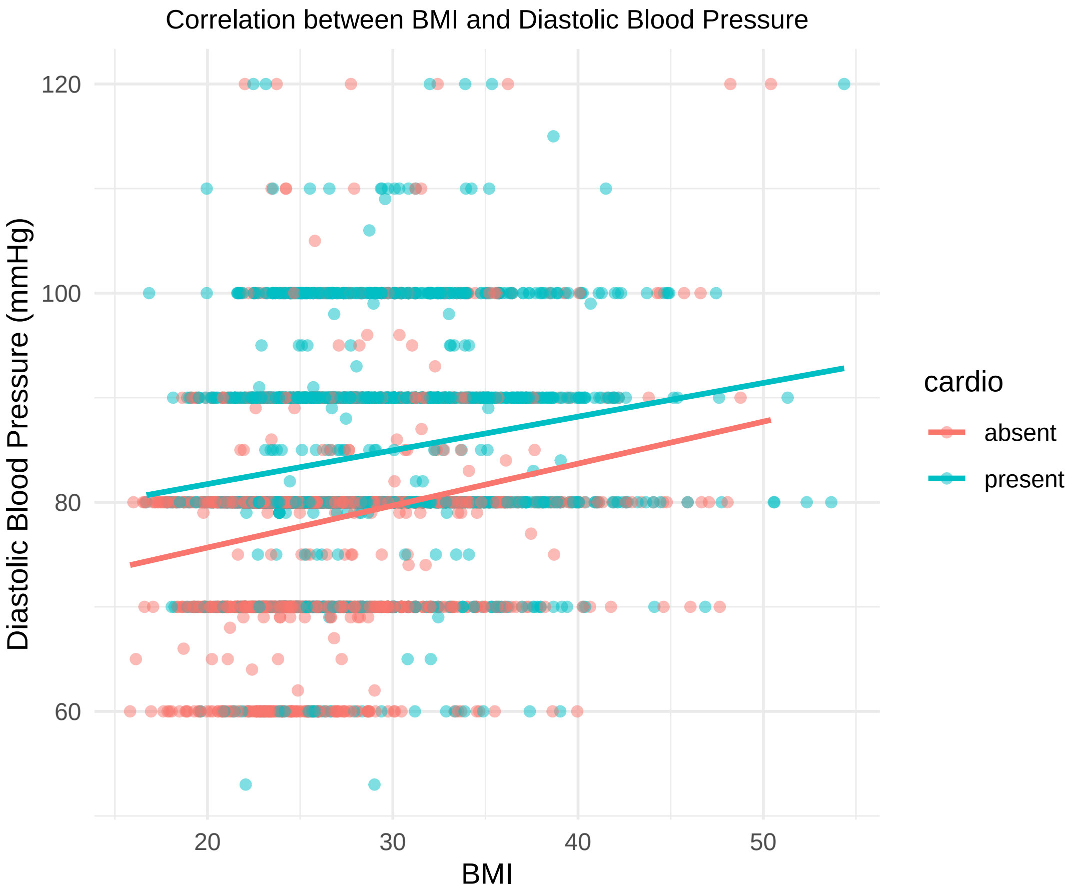


Table 8: Fisher's Z-test for Correlations between BMI and Diastolic blood pressure grouped by cardio

Metric	Value
Z_score	1.3260487
P_value	0.1848236

Answer: The correlation coefficient between BMI and Diastolic Blood Pressure (DBP) is 0.2541928, indicating a weak positive correlation. The p-value is $8.1781399 \times 10^{-73}$, suggesting that this correlation is statistically significant across the dataset.

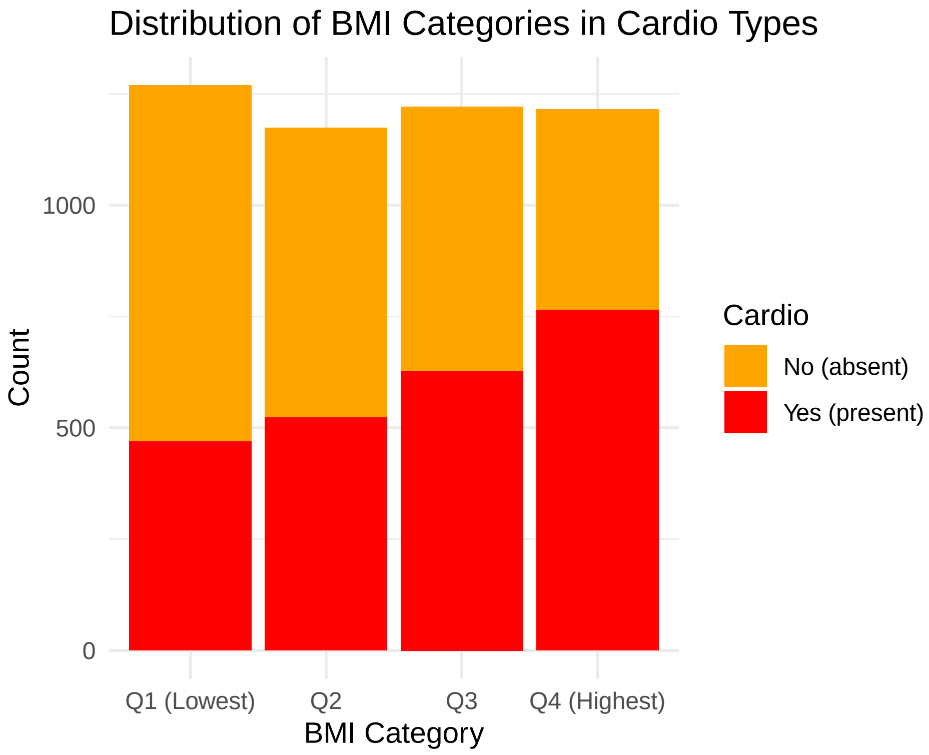
Based on Fisher's Z-test, the Z-score is 1.3260487, and the P-value is 0.1848236. These results show no statistically significant difference in the correlation coefficients between BMI and Diastolic Blood Pressure across the two groups of individuals with/without cardiovascular disease. The observed outcomes may be attributed to random variation rather than a meaningful effect.

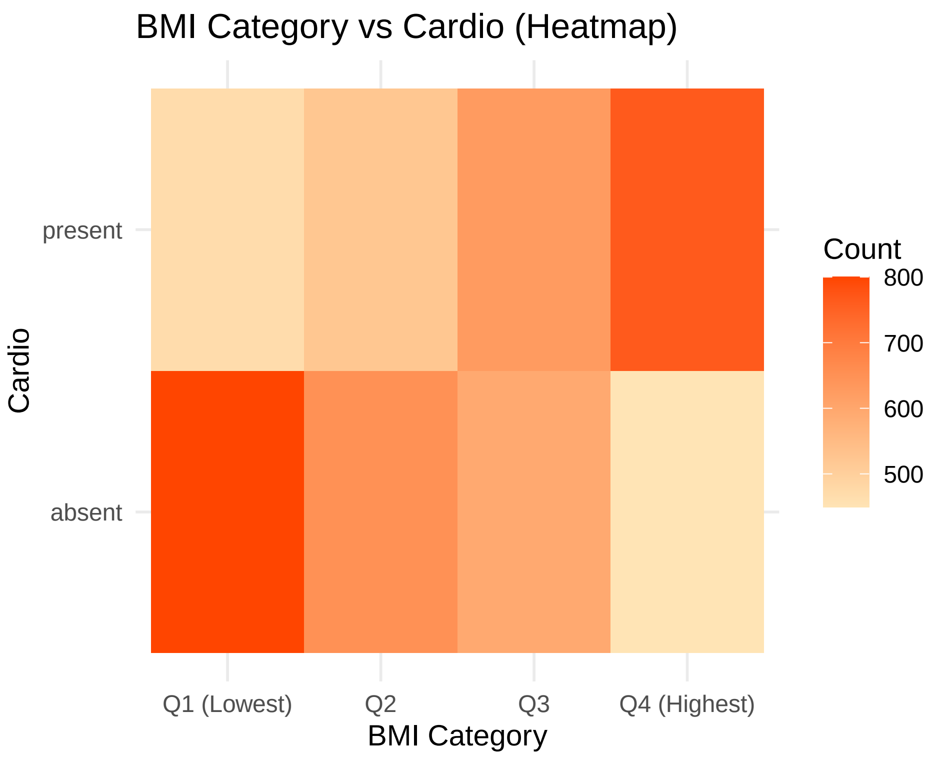
Chapter 6: Categorize BMI into Quartiles and Visualize its Distribution Across Cardio Types

Task 6: Categorize the variable BMI according to its quartiles and visualize the distribution of the BMI categories in both cardio types.

Table 9: BMI Quartiles Distribution Across Cardio Types

BMI_category	cardio	count
Q1 (Lowest)	absent	800
Q1 (Lowest)	present	469
Q2	absent	651
Q2	present	523
Q3	absent	594
Q3	present	627
Q4 (Highest)	absent	450
Q4 (Highest)	present	765





Conclusion from the Heatmap:The heatmap shows a positive association between BMI and cardiovascular disease. Higher BMI quartiles have more individuals with cardiovascular conditions ("present") and fewer without ("absent"). This suggests that as BMI increases, the likelihood of cardiovascular disease also rises.

Chapter 7: Age Distribution Across Cardio Categories (in Years)

Task 7: How is age distributed in the different categories of cardio? Display age in years.

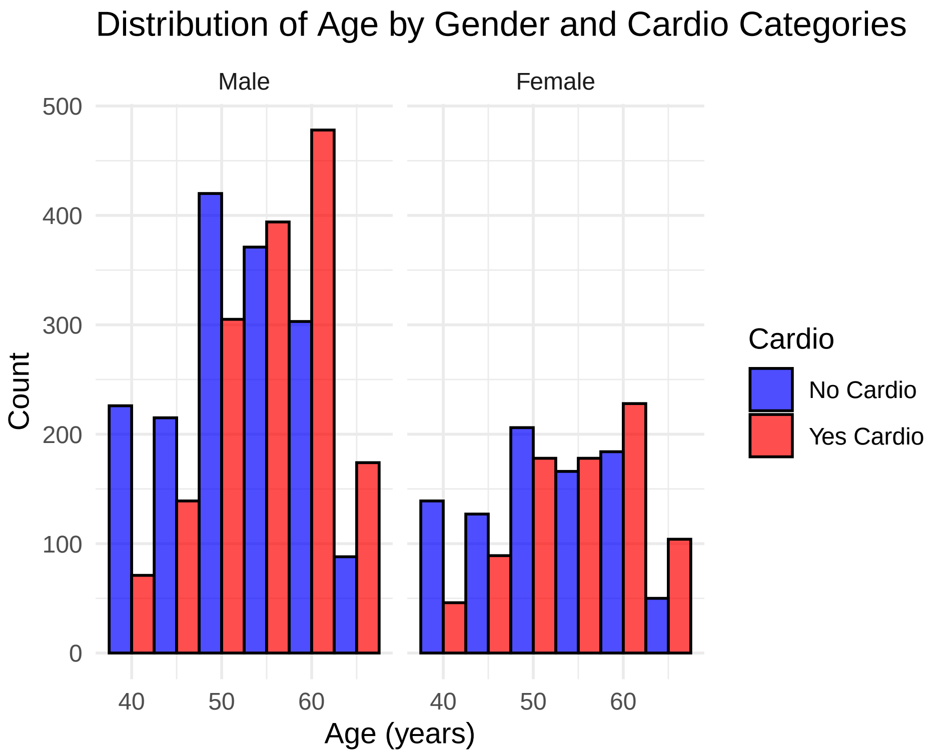
Table 10: Summarize age distribution for each cardio group

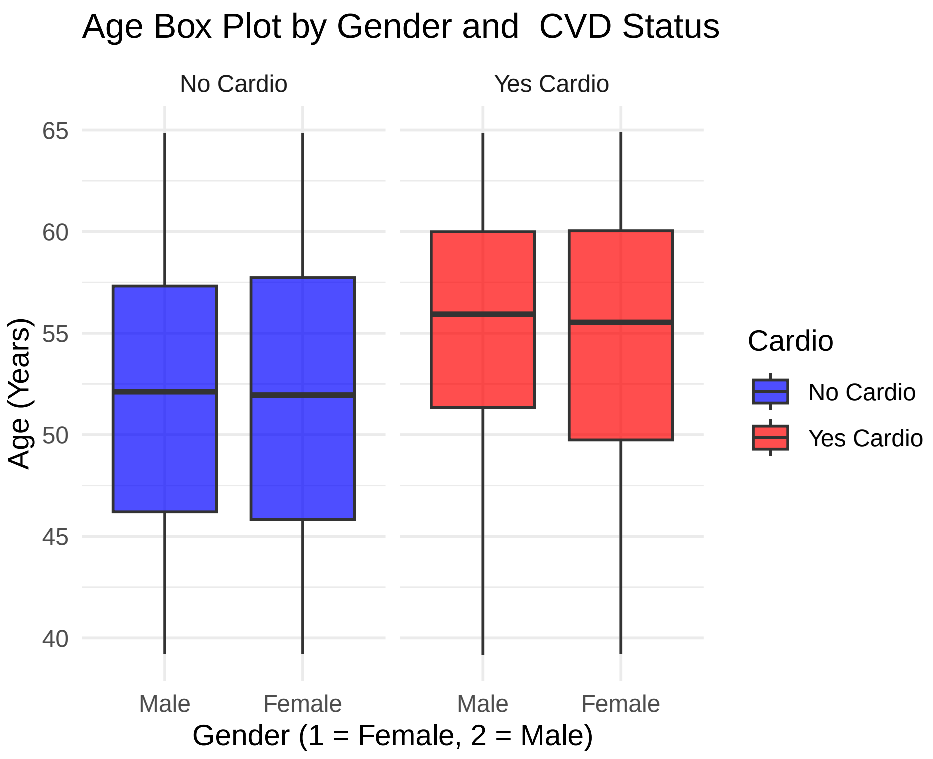
cardio	Count	Mean_Age	Median_Age	Min_Age	Max_Age	SD_Age
absent	2495	51.77784	52.070	39.21	64.85	6.901031
present	2384	54.89690	55.835	39.16	64.90	6.428545

The two-sample t-test, assuming normality, independence, and equal variances, showed a significant difference in mean age between individuals with and without cardiovascular disease (p-value $2.5609992 \times 10^{-58}$), indicating that those with CVD tend to have elder age.

Chapter 8: Visualizing the Age Distribution by Gender and CVD Status

Task 8: Create a plot that shows the distribution of age for both types of gender and both types of cardio

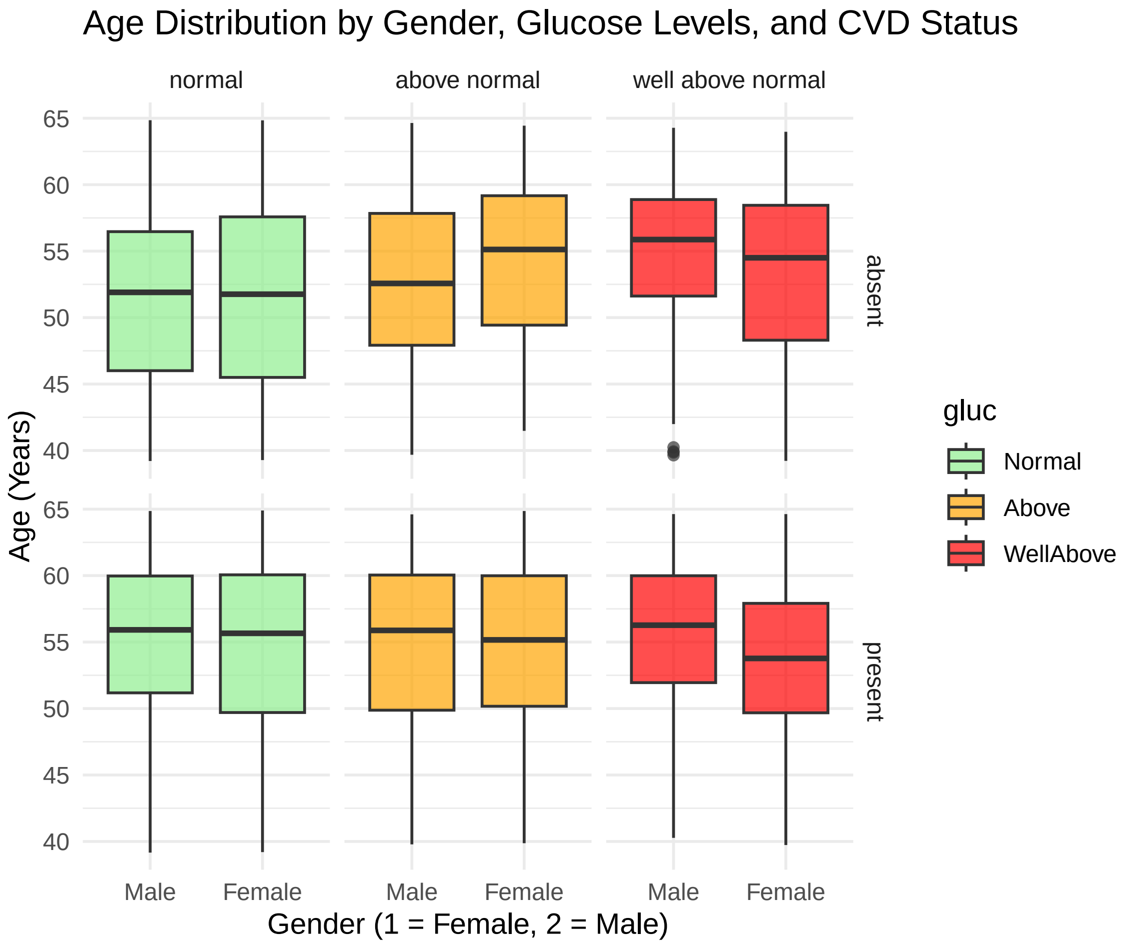




Conclusion: Age distribution shows that individuals with cardiovascular disease are generally older than those without the disease, and this pattern is consistent across both genders. Both males and females with cardiovascular disease have higher ages compared to their counterparts without the disease. This indicates that age is a significant risk factor for cardiovascular disease, regardless of gender.

Chapter 9: Age Distribution with Glucose Levels

Task 9. Extend this plot by taking the different types of glucose into account.



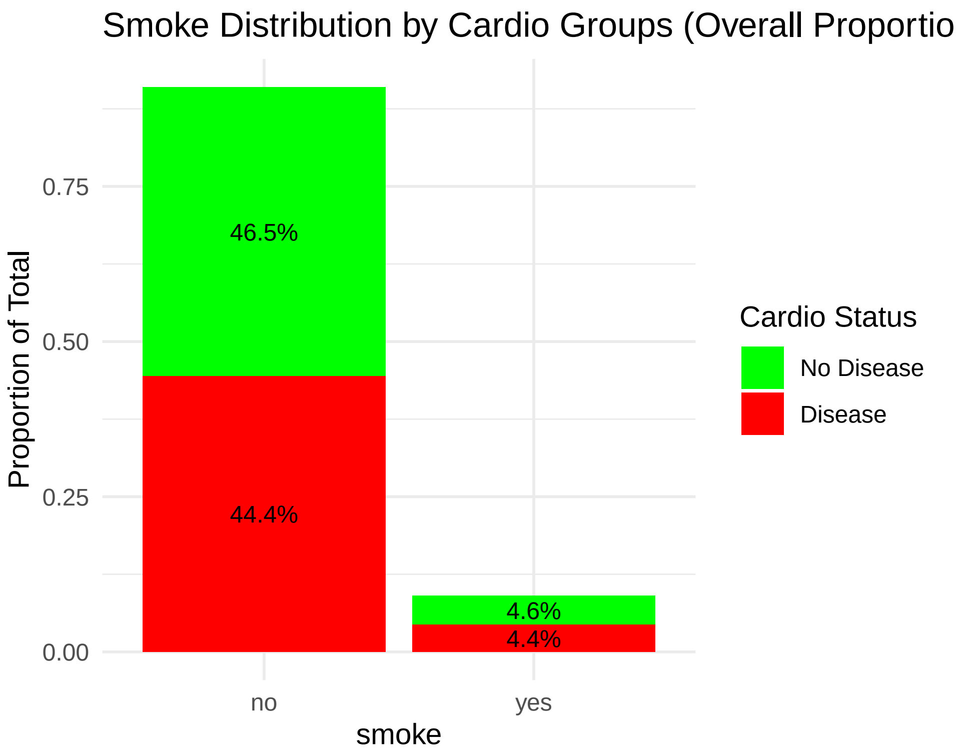
Age and Glucose Levels: Among individuals with normal blood glucose levels, those with cardiovascular disease generally have a higher median age compared to those without. However, in groups with elevated or very high blood glucose levels, the difference in median age between individuals with and without cardiovascular disease is less pronounced, particularly among females.

Gender-Specific Observations: We observe that in the population with very high glucose levels (glucose level = “high”), females are generally younger than males, indicating that women may develop blood sugar issues at a younger age.

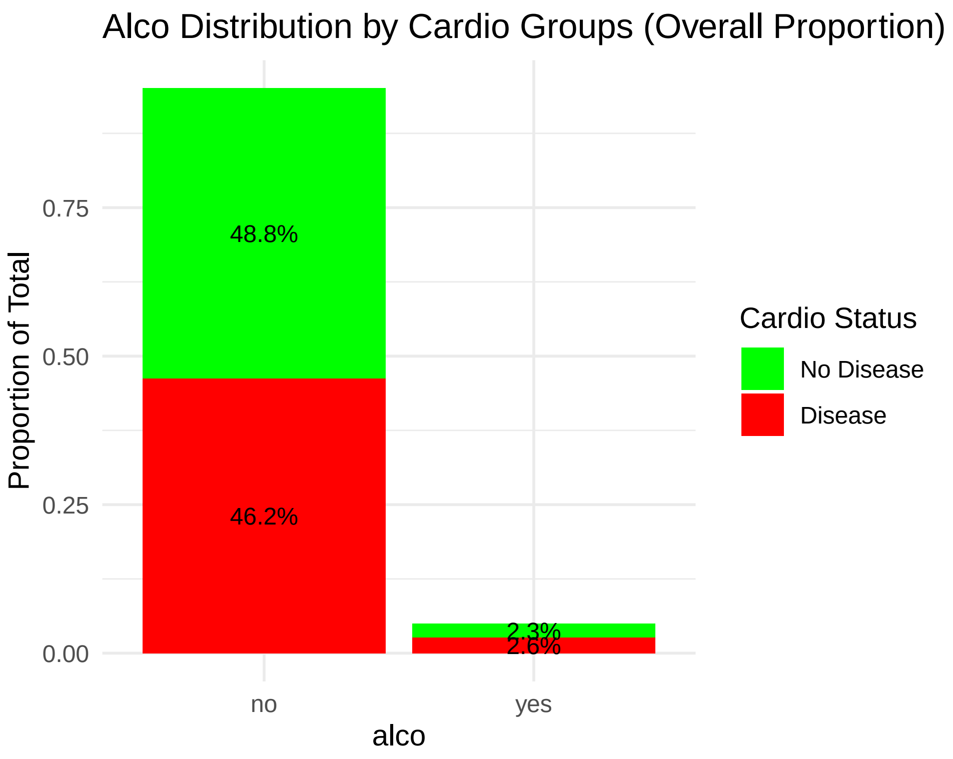
Chapter 10: Risk Factors for CVD: Comparing Lifestyle Parameters Using ^{2} Tests

Task 10: Further risk factors for a cardiovascular disease may be smoking, alcohol, and insufffcient physical activity. Create plots and an overview table of how these three parameters are distributed between the two types of cardio and compare all three with a 2 -test, respectively. Draw a conclusion about which of these parameters may be risk factors for cardiovascular diseases.

```
[[1]]
```



```
[[2]]
```


Alco Distribution by Cardio Groups (Overall Proportion)
Proportion of Total
0.00
0.25
0.50
0.75
48.8%
46.2%
2.3%
2.6%
no
yes
alco
Cardio Status
🟩 No Disease
🟥 Disease

```
[[3]]
```

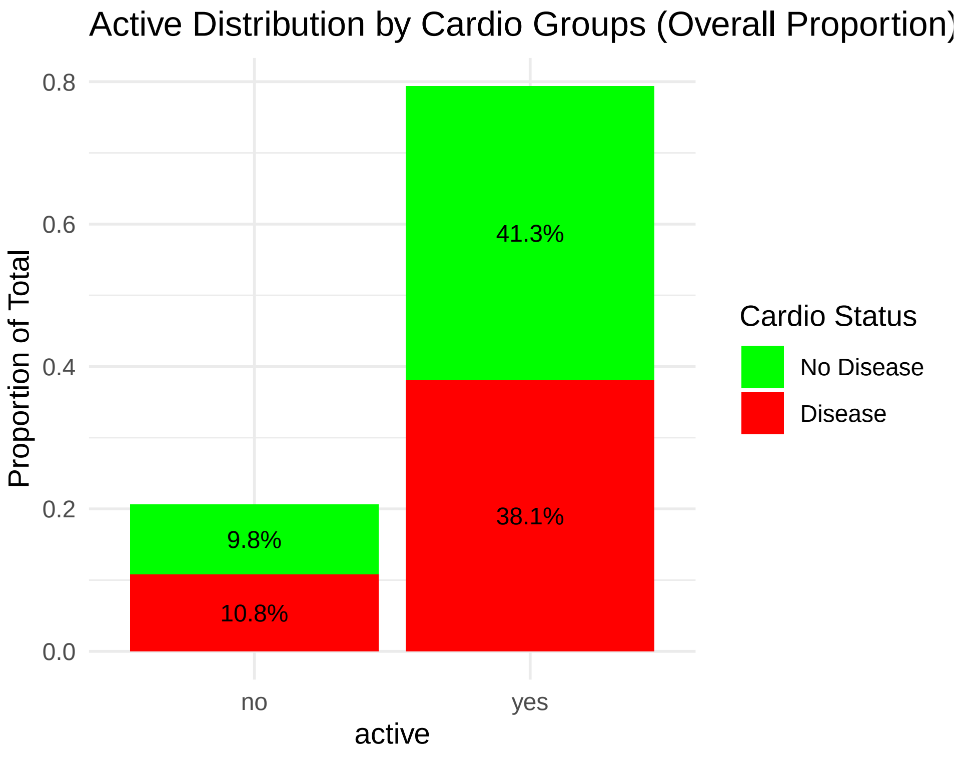

Active Distribution by Cardio Groups (Overall Proportion)
Proportion of Total
0.0
0.2
0.4
0.6
0.8
9.8%
10.8%
41.3%
38.1%
no
yes
active
Cardio Status
🟩 No Disease
🟥 Disease

Table 11: Summary of Lifestyle Factors Counts by Cardiovascular Risk Group

cardio	Category	Count
absent	Active_No	480
present	Active_No	527
absent	Active_Yes	2015
present	Active_Yes	1857
absent	Alcohol_No	2382
present	Alcohol_No	2256
absent	Alcohol_Yes	113
present	Alcohol_Yes	128
absent	Smoking_No	2271
present	Smoking_No	2168
absent	Smoking_Yes	224
present	Smoking_Yes	216

Table 12: Proportions of Lifestyle Factors by Cardiovascular Risk Group

cardio	Category	Proportion
absent	Active_No	0.1924
present	Active_No	0.2211
absent	Active_Yes	0.8076
present	Active_Yes	0.7789
absent	Alcohol_No	0.9547
present	Alcohol_No	0.9463
absent	Alcohol_Yes	0.0453
present	Alcohol_Yes	0.0537
absent	Smoking_No	0.9102
present	Smoking_No	0.9094
absent	Smoking_Yes	0.0898
present	Smoking_Yes	0.0906

Table 13: Chi-Square Test Results for Risk Factors and Cardiovascular Risk

Test	Chi_Square_Statistic	Degrees_of_Freedom	p_Value
Smoking	0.0025508	1	0.9597197
Alcohol	1.6577325	1	0.1979098
Physical Activity	5.9450162	1	0.0147590
Glucose Level	61.4048369	2	0.0000000

Based on the p-values:

1. **Smoking** and **Alcohol Consumption** : No significant association with cardiovascular diseases.

2. **Physical Activity** and **Glucose Level** : Significant association, suggesting that insufficient activity and high glucose level are potential risk factors.

Chapter 11: Summary

Analysis Steps:

1. Data Preparation: Variables were transformed into appropriate data types, and categorical variables were labeled for clarity.
2. Outlier Detection: Continuous variables (age, height, weight, systolic blood pressure, and diastolic blood pressure) were analyzed for outliers, and missing values were excluded to ensure data quality.
3. BMI and CVD: BMI was calculated and analyzed. Individuals with CVD had significantly higher BMI values compared to those without, as confirmed by a t-test.
4. Correlation Analysis: Weak positive correlations were observed between BMI and both systolic and diastolic blood pressure. Fisher's Z-tests showed no significant differences in these correlations between CVD and non-CVD groups.
5. BMI Distribution: BMI quartile analysis revealed a higher prevalence of CVD in the upper quartiles, supporting the link between increased BMI and cardiovascular disease.
6. Age and CVD: CVD patients were older on average, with a statistically significant difference confirmed by a t-test.
7. Lifestyle and Glucose Levels: Chi-square tests highlighted physical inactivity and elevated glucose levels as significant risk factors for CVD, while smoking and alcohol consumption showed no significant associations.

Key Findings:

- BMI and CVD: Higher BMI is linked to cardiovascular disease, with individuals in higher BMI quartiles at greater risk, consistent with previous studies showing increased CVD risk in overweight individuals (Hubert et al. 1983).
- Glucose Levels and Age: Elevated glucose levels and older age also emerged as key factors linked to CVD.
- Physical Inactivity: Insufficient physical activity was identified as a significant modifiable risk factor for CVD, consistent with research emphasizing the same conclusion(Lee et al. 2012).

Next Steps:

This analysis identified BMI, glucose levels, age, and physical inactivity as significant predictors of CVD. To better understand their combined impact, further steps include:

1. Advanced Modeling: Use logistic regression or other classification models to quantify each predictor's contribution.
2. Performance Metrics: Assess model effectiveness using metrics like pseudo R^2, AUC, and F1 score.

3. Future Research: Incorporate additional variables, such as cholesterol, blood pressure, dietary habits, genetics, and socioeconomic factors, to provide a more comprehensive understanding of CVD risk.

These approaches will help evaluate how well these factors collectively explain and predict cardiovascular disease.

References

Hubert, H. B., M. Feinleib, P. M. McNamara, and W. P. Castelli. 1983. “Obesity as an Independent Risk Factor for Cardiovascular Disease: A 26-Year Follow-up of Participants in the Framingham Heart Study.” *Circulation* 67 (5): 968–77. https://doi.org/10.1161/01.cir.67.5.968.

Lee, I-Min, Eric J. Shiroma, Felipe Lobelo, Pekka Puska, Steven N. Blair, and Peter T. Katzmarzyk. 2012. “Effect of Physical Inactivity on Major Non-Communicable Diseases Worldwide: An Analysis of Burden of Disease and Life Expectancy.” *The Lancet* 380 (9838): 219–29. https://doi.org/10.1016/S0140-6736(12)61031-9.

Organization, World Health. 2021. “Cardiovascular Diseases (CVDs).” *World Health Organization*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).