home_work_test

Baodong Zhang

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, blood pressure, 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

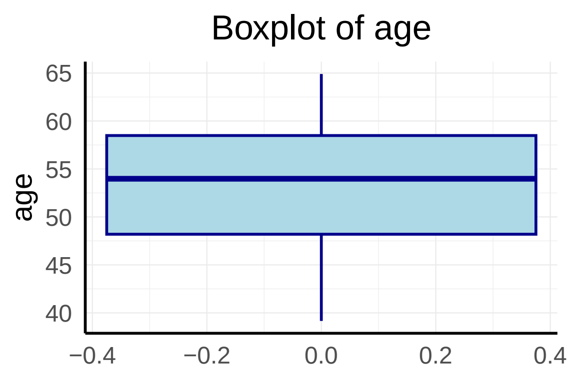
Chapter 2: Outlier Detection

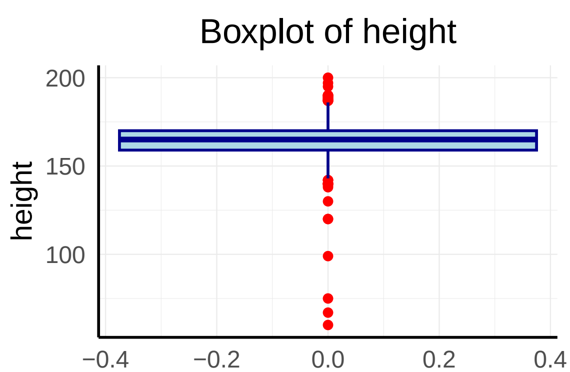
Task 2: Check the continuous variables for outliers and remove implausible values.

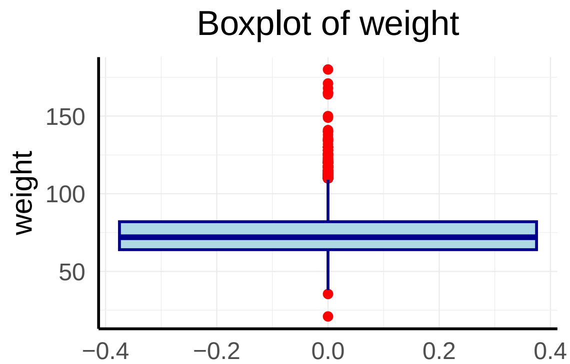
Table 2: Summary of Continuous Variables

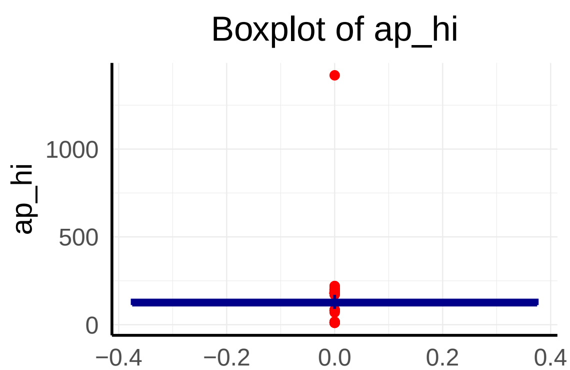
	age	height	weight	ap_hi	ap_lo
	Min. :39.16	Min. : 60.0	Min. : 21.00	Min. : 10.0	Min. : 40.0
	1st Qu.:48.20	1st Qu.:159.0	1st Qu.: 64.00	1st Qu.: 120.0	1st Qu.: 80.0
	Median :53.98	Median :165.0	Median : 72.00	Median : 120.0	Median : 80.0
	Mean :53.31	Mean :164.3	Mean : 74.01	Mean : 126.7	Mean : 96.1
	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
weight
150
100
50
−0.4
−0.2
0.0
0.2
0.4


Boxplot of ap_hi
ap_hi
1000
500
0
−0.4
−0.2
0.0
0.2
0.4

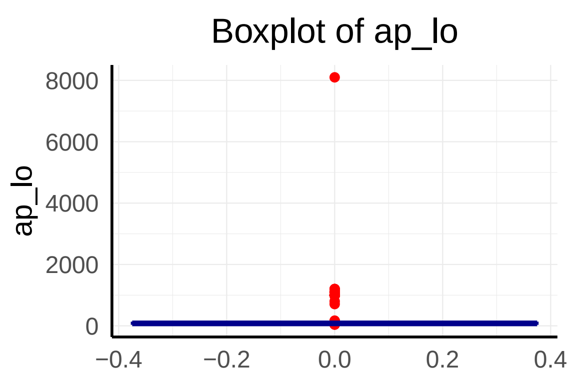
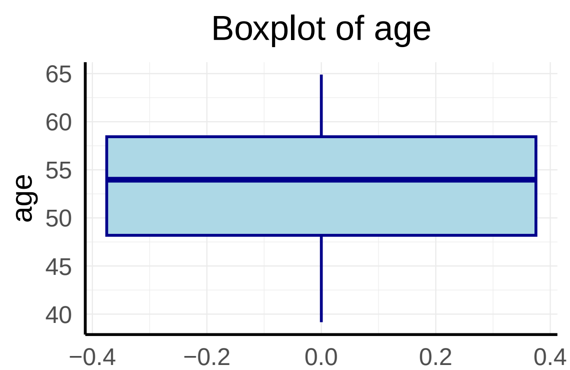
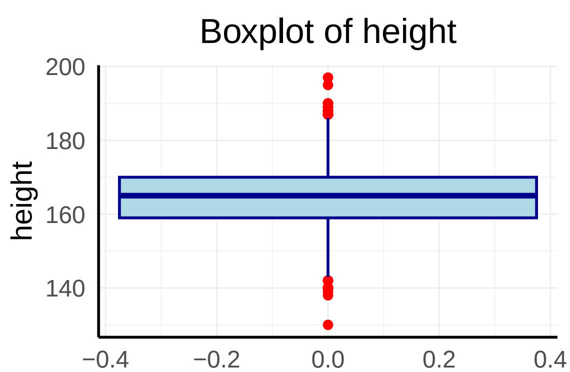

Boxplot of ap_lo
ap_lo
8000
6000
4000
2000
0
−0.4
−0.2
0.0
0.2
0.4

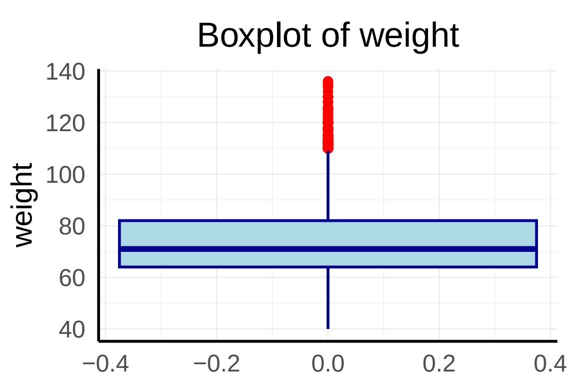
Table 3: Summary of Continuous Variables (outliers are removed)

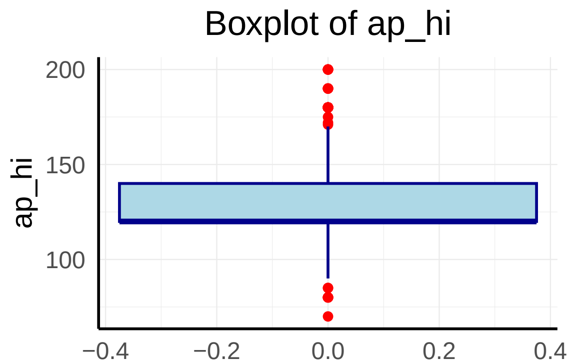
	age	height	weight	ap_hi	ap_lo
	Min. :39.16	Min. :130.0	Min. : 40.00	Min. : 70.0	Min. : 53.00
	1st Qu.:48.19	1st Qu.:159.0	1st Qu.: 64.00	1st Qu.:120.0	1st Qu.: 80.00
	Median :53.97	Median :165.0	Median : 71.00	Median :120.0	Median : 80.00
	Mean :53.30	Mean :164.4	Mean : 73.71	Mean :126.4	Mean : 81.25
	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

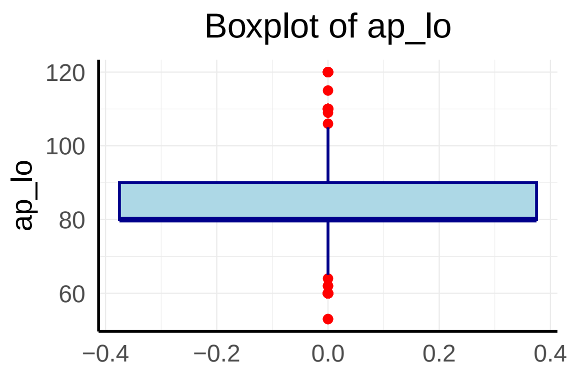
Boxplot after outliers are removed








Boxplot of ap_hi
ap_hi
200
150
100
−0.4
−0.2
0.0
0.2
0.4


Boxplot of ap_lo
ap_lo
120
100
80
60
−0.4
−0.2
0.0
0.2
0.4

Chapter 3: BMI Calculation

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

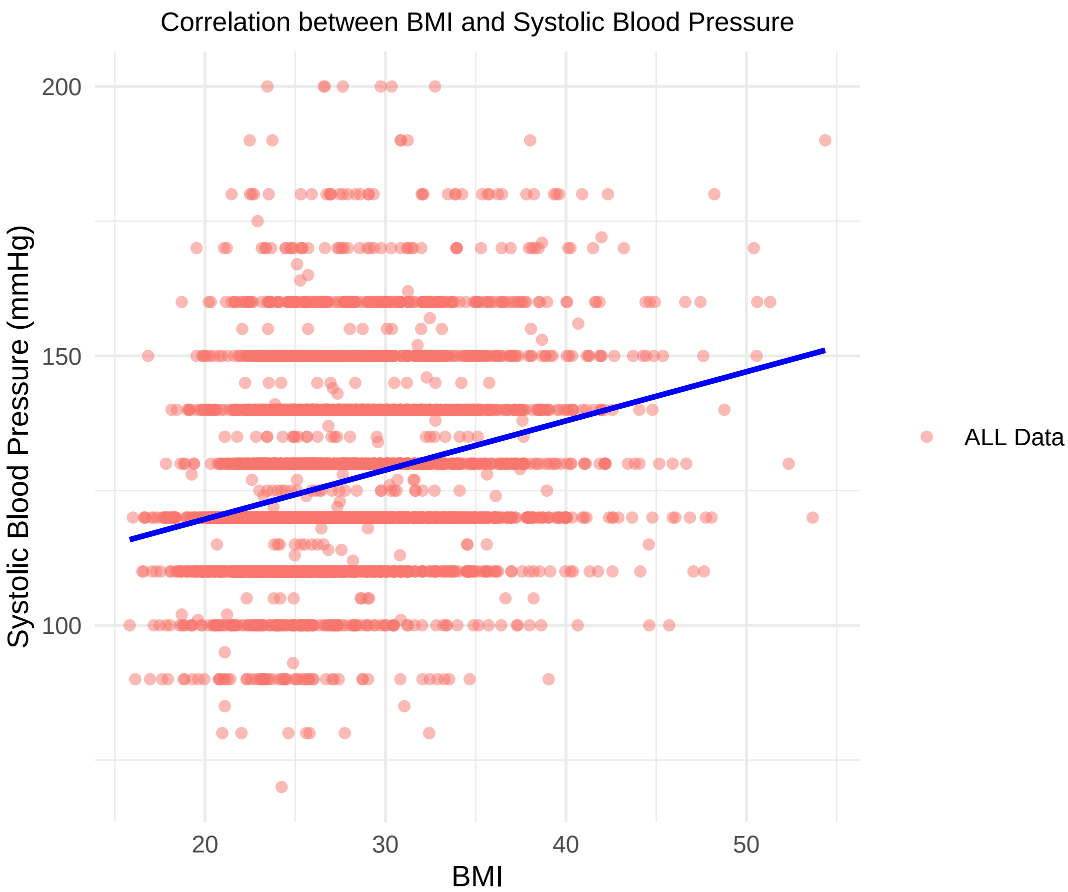
Table 4: Summary table for BMI grouped by cardio

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.

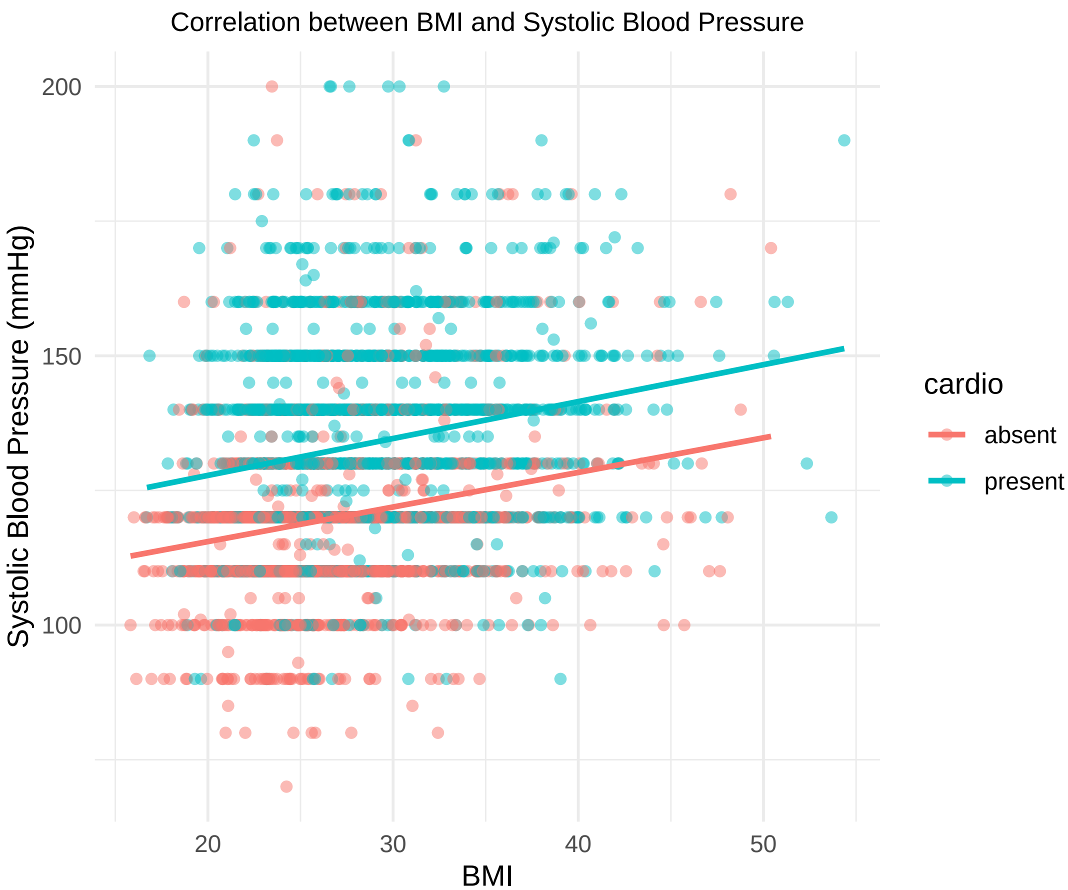


Table 5: Correlation between BMI and Systolic blood pressure grouped by cardio

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 $8.1781399 \times 10^{-73}$.

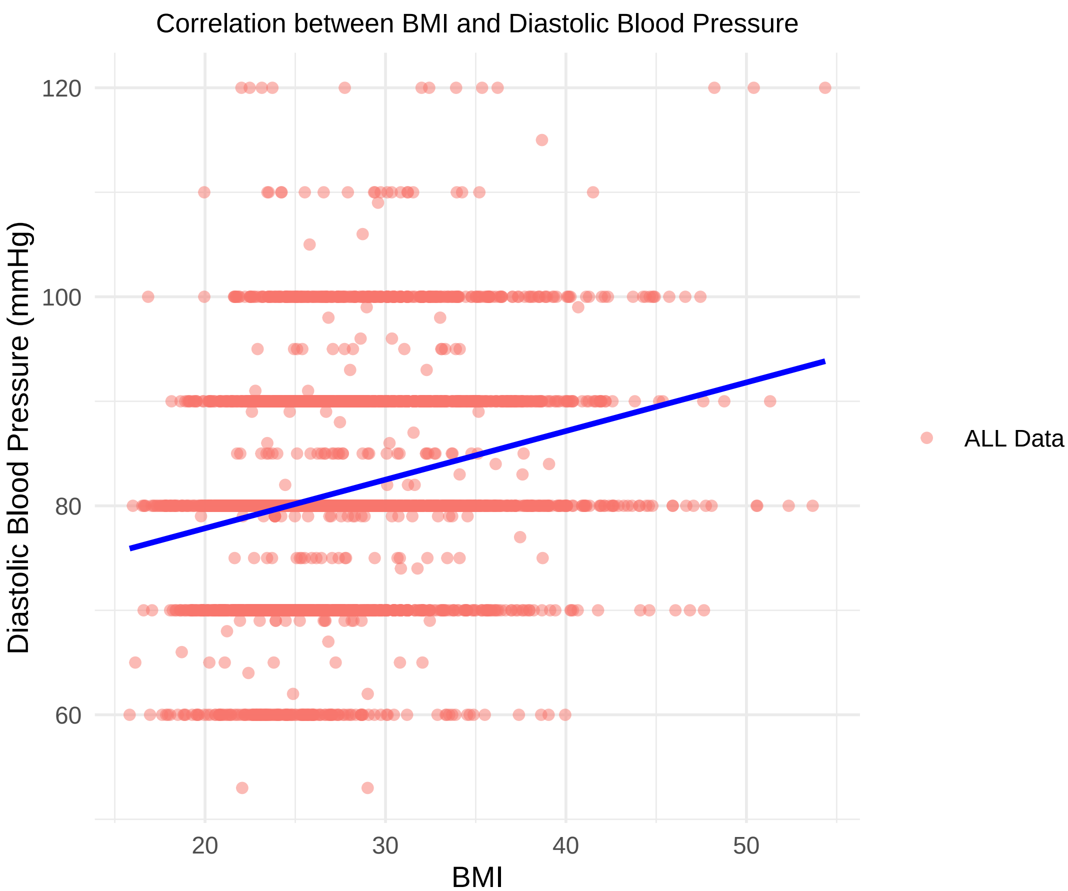


Table 7: Correlation between BMI and Diastolic blood pressure grouped by cardio

cardio	Correlation
absent	0.2233765
present	0.1869794

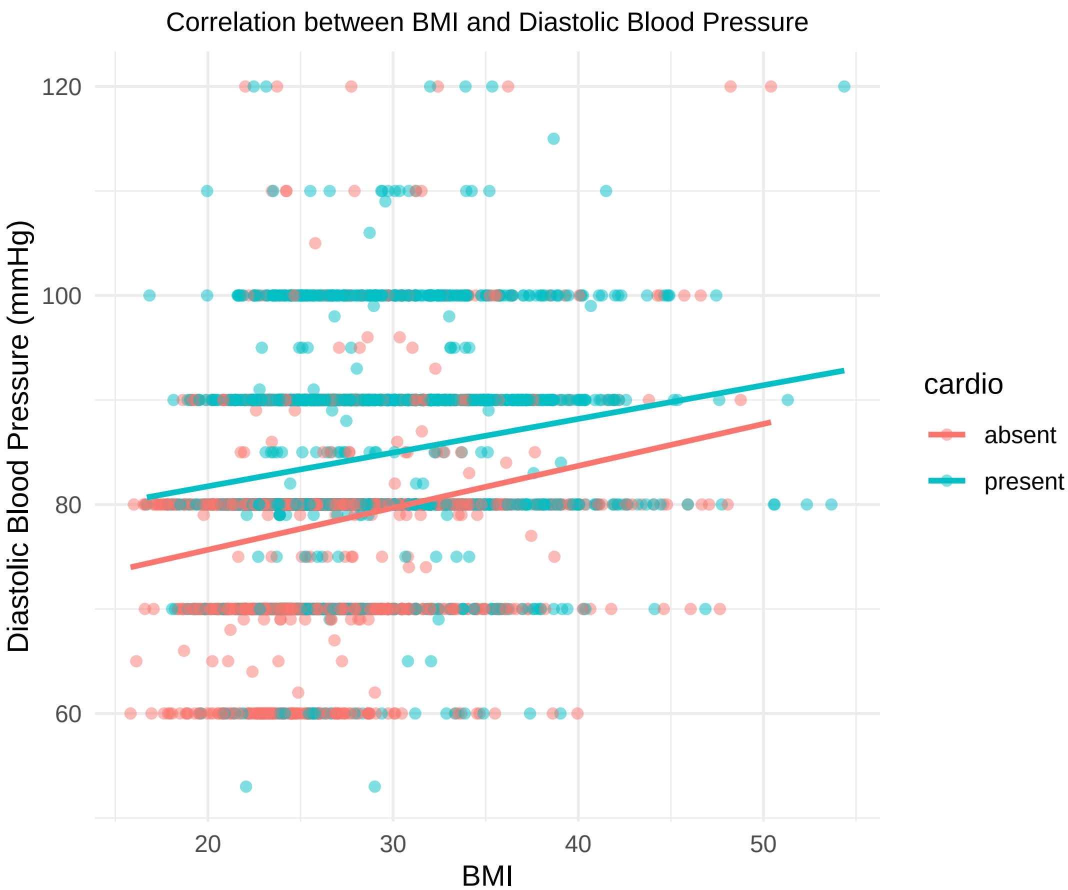


Table 8: Fisher's Z-test for Correlations between BMI and Diastolic blood pressure grouped by cardio

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.

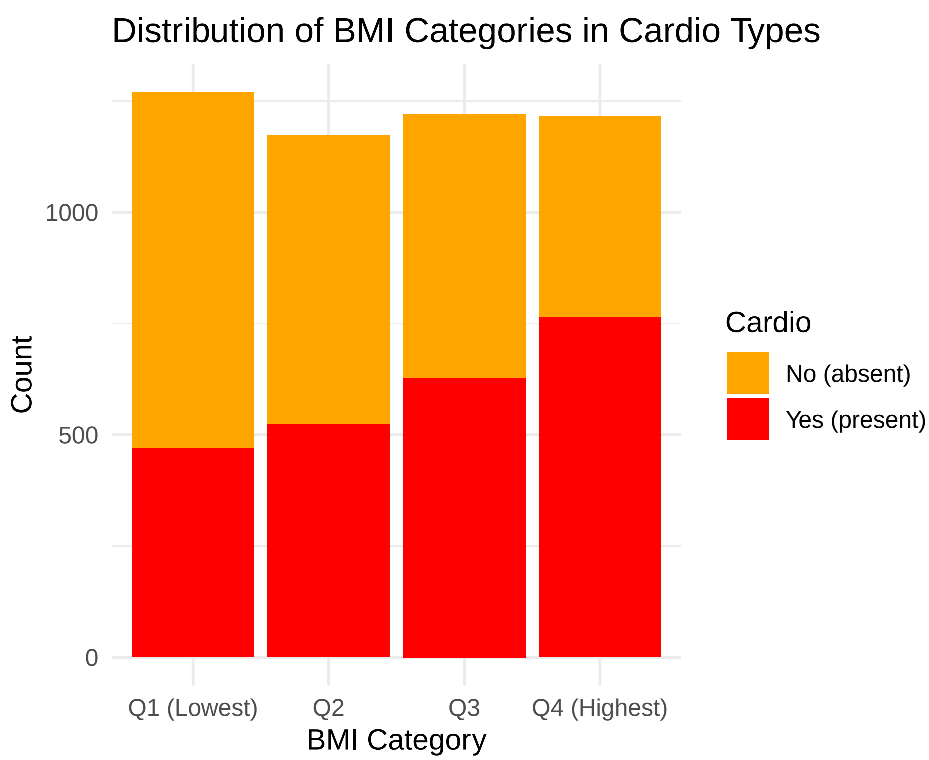
Based on Fisher's Z-test, the Z-score is 1.3260487, and the P-value is 0.1848236. These results show no statistically significant difference in the correlation coefficients between BMI and Diastolic Blood Pressure across the two groups of individuals with/without cardiovascular disease. The observed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

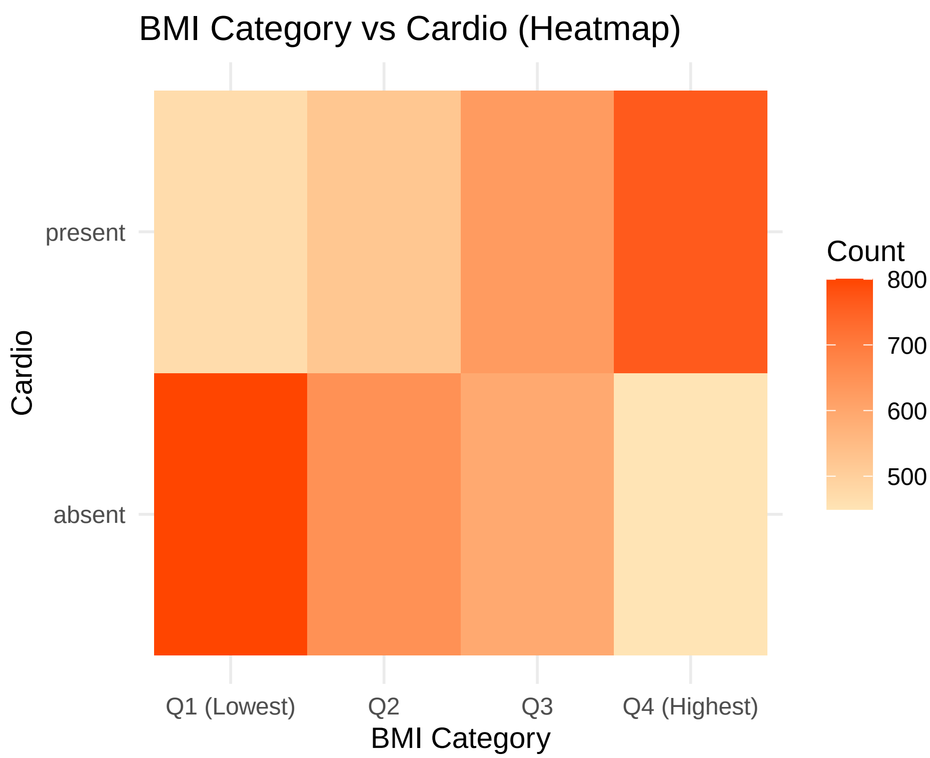
Chapter 6: Categorize BMI into Quartiles and Visualize its Distribution Across Cardio Types

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.

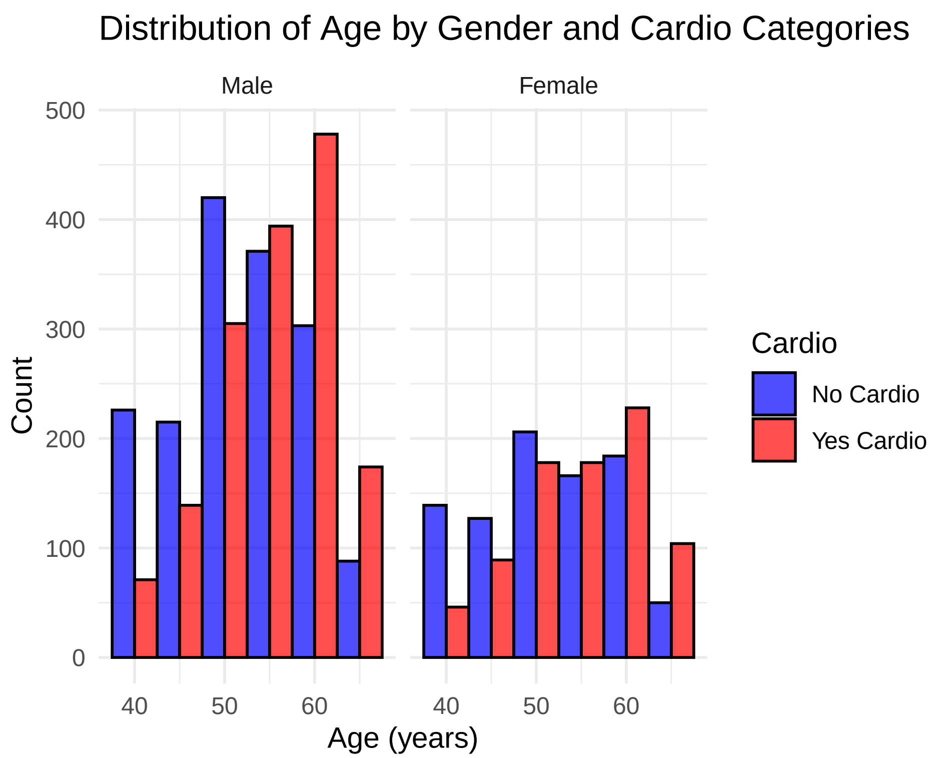
Table 10: Summarize age distribution for each cardio group

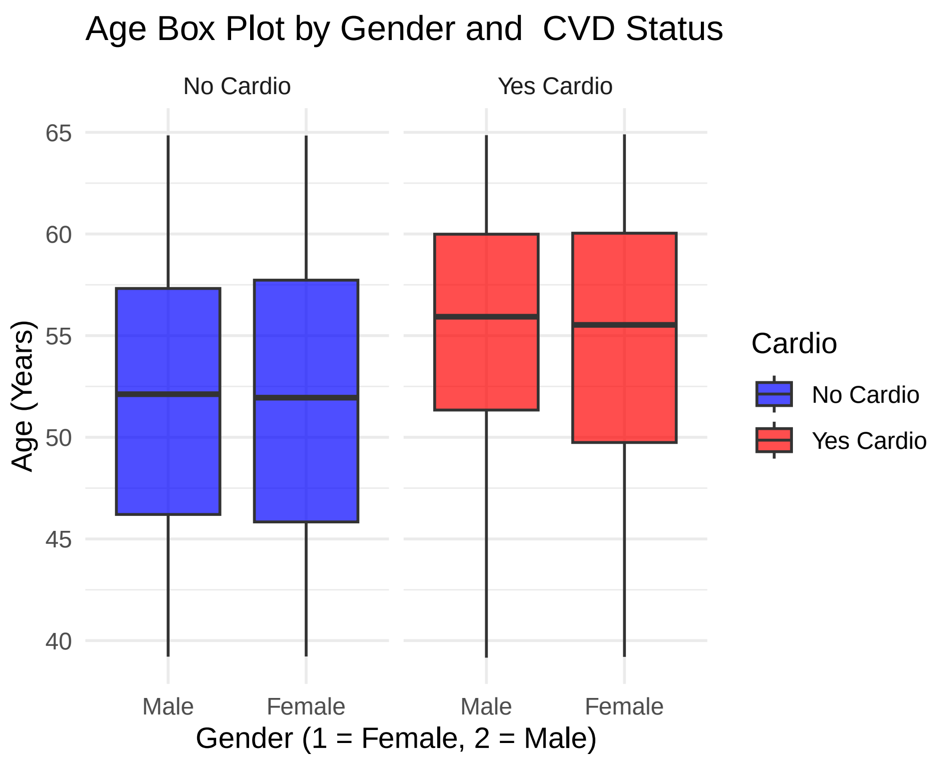
cardio	Count	Mean_Age	Median_Age	Min_Age	Max_Age	SD_Age
absent	2495	51.77784	52.070	39.21	64.85	6.901031
present	2384	54.89690	55.835	39.16	64.90	6.428545

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

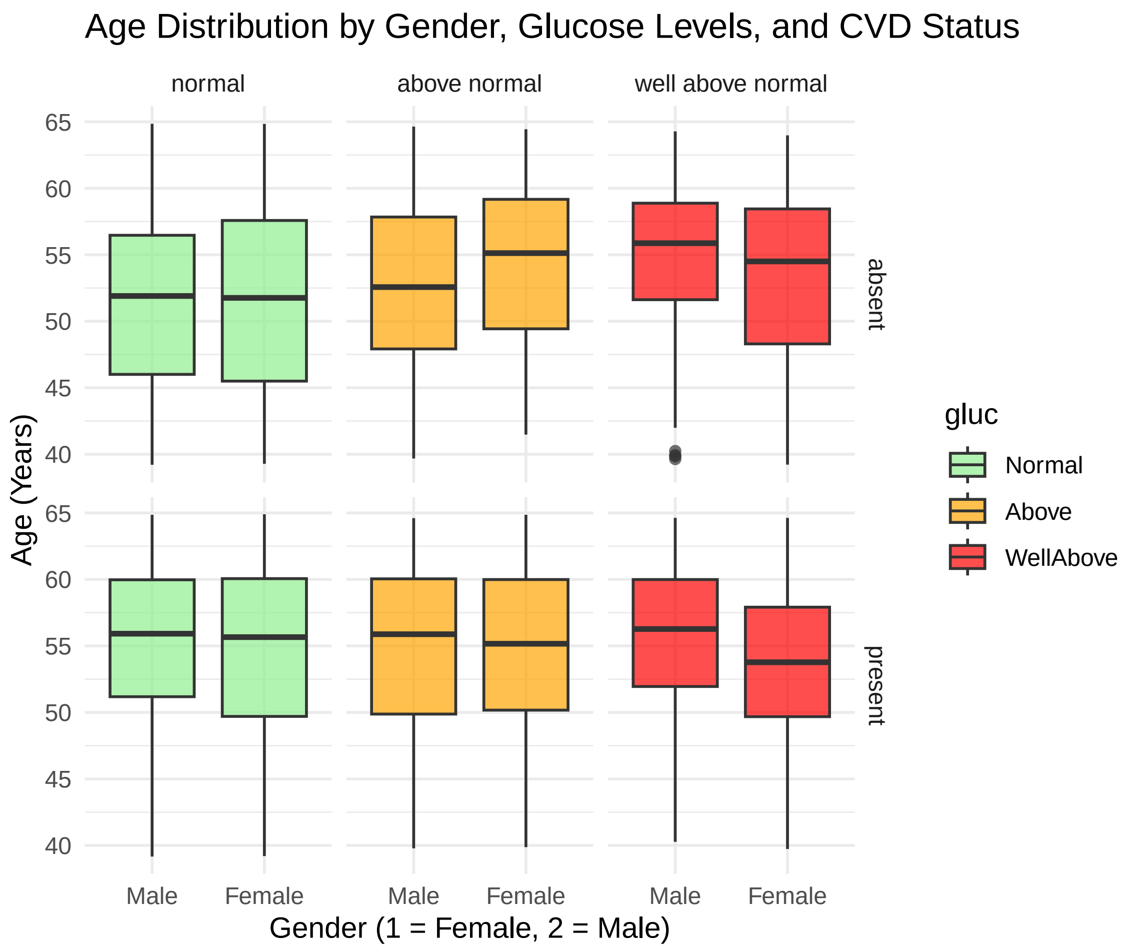




Conclusion: Age distribution shows that individuals with cardiovascular disease are generally older than those without the disease, and this pattern is consistent across both genders. Both males and females with cardiovascular disease have higher ages compared to their counterparts without the disease. 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.



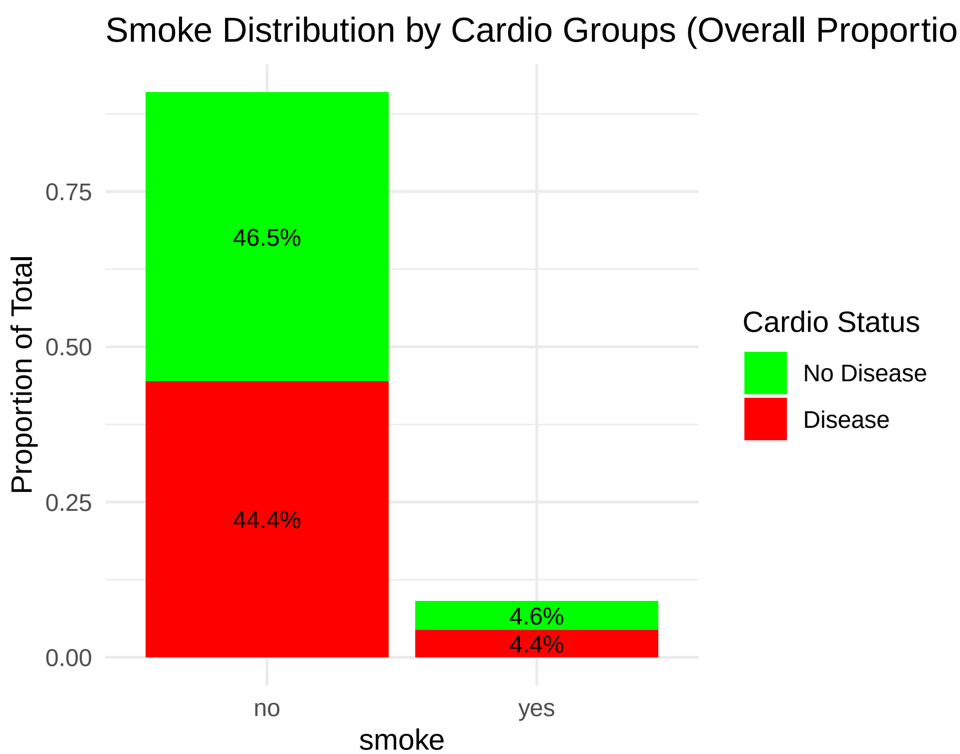
Age and Glucose Levels: Across all glucose levels, individuals with cardiovascular disease are older than those without, for both genders.

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.

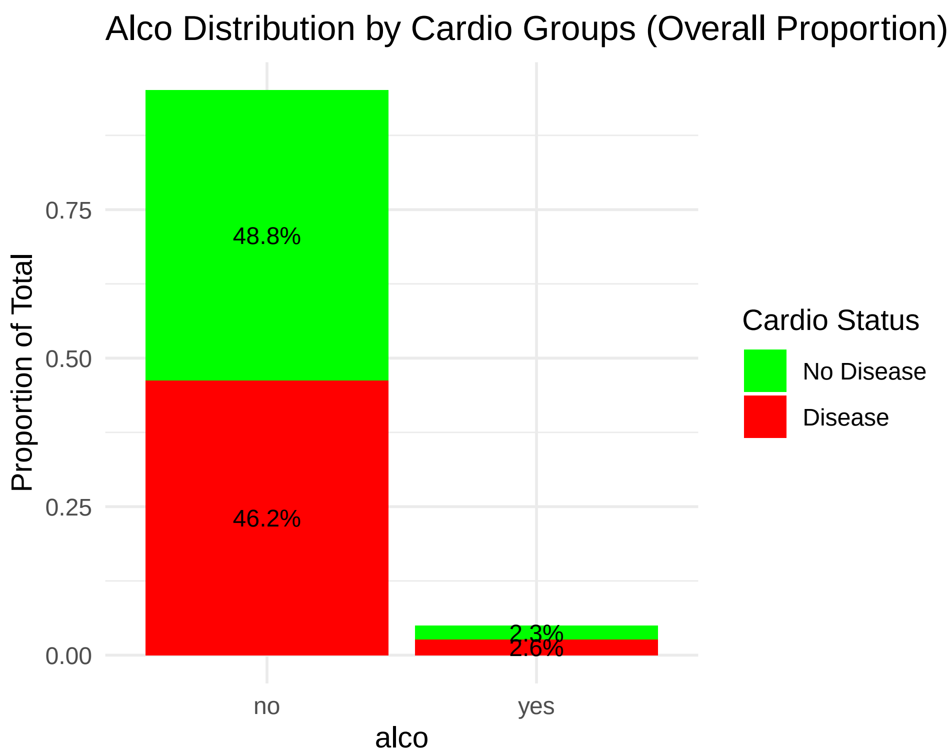
Chapter 10: Risk Factors for CVD: Comparing Lifestyle Parameters Using ^{2} Tests

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
0.75
0.50
0.25
0.00
48.8%
46.2%
2.3%
2.6%
no
yes
alco
Cardio Status
No Disease
Disease

[[3]]

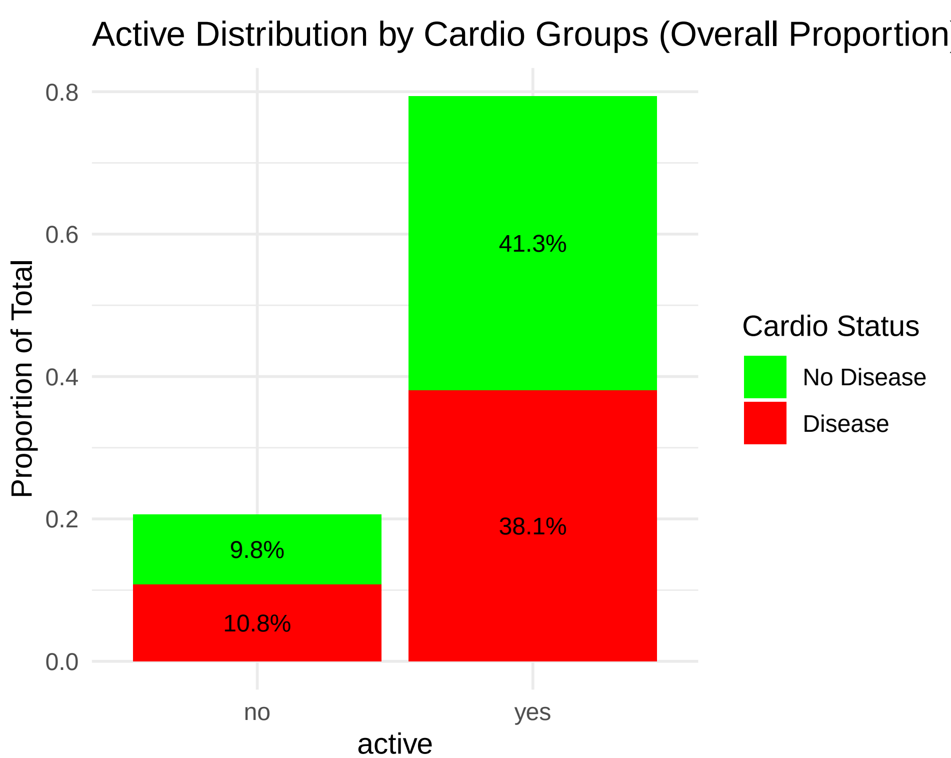

Active Distribution by Cardio Groups (Overall Proportion)
Proportion of Total
0.8
0.6
0.4
0.2
0.0
41.3%
38.1%
9.8%
10.8%
no
yes
active
Cardio Status
No Disease
Disease

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.

Analysis Steps and Results

1. Data Preparation

- Variables were transformed into appropriate types, and categorical variables were labeled (e.g., gender, cholesterol, smoking habits). The dataset included health metrics such as age, blood pressure, BMI, and lifestyle factors.

2. Outlier Detection

- Continuous variables like age, height, weight, systolic blood pressure (ap_hi), and diastolic blood pressure (ap_lo) were analyzed for outliers.
- Implausible values (e.g., extreme blood pressure readings) were removed, and summaries were recalculated.

3. BMI Calculation

- BMI was computed and summarized across two groups: individuals with and without cardiovascular disease (CVD). Results showed higher average BMI in the CVD group (28.31) compared to the non-CVD group (26.34).

4. Correlation Analysis

- A weak positive correlation was found between BMI and both systolic (0.279) and diastolic blood pressure (0.2542). However, no statistically significant differences in correlations were found between individuals with or without CVD.

5. BMI Quartile Categorization

- BMI was divided into quartiles to analyze distribution across CVD groups. Higher BMI quartiles were associated with a greater proportion of CVD cases.

6. Age Distribution

- CVD patients were, on average, older (mean age ~55 years) compared to non-CVD individuals (~52 years). Age distribution was visualized by gender and glucose levels, revealing slight variations across groups.

7. Lifestyle Risk Factors

- Smoking, alcohol consumption, and physical inactivity were compared between CVD groups using ^{2} tests. Results indicated no significant association for smoking (p=0.96) or alcohol (p=0.20).
- Insufficient physical activity showed a significant association with CVD (p=0.015), highlighting it as a potential risk factor.

Conclusion

Key findings suggest that higher BMI, older age, and physical inactivity are associated with cardiovascular disease. Lifestyle interventions targeting physical activity could be a priority for reducing CVD risk.

Analysis Steps and Results

Analysis Steps

1. **Data Preparation**
   Variables were transformed into appropriate types, and categorical variables were labeled. Key health metrics were categorized for analysis.

2. **Outlier Detection**
   Continuous variables like age, blood pressure, height, and weight were checked for outliers. Implausible values were removed to ensure the integrity of the analysis.

3. **BMI Calculation**
   BMI was computed, and its distribution was summarized across groups with and without cardiovascular disease (CVD). It was observed that individuals with CVD had a higher mean BMI (28.31) than those without CVD (26.34). T test suggests that individuals with CVD tend to have significantly higher BMI than those without.

4. **Correlation Analysis between BMI and blood pressure**
   Weak positive correlations were found between BMI and blood pressure (both systolic and diastolic). However, there was no statistically significant difference in these correlations between CVD and non-CVD groups based on Fisher's Z-tests.

5. **Categorization and Distribution Analysis of BMI**
   BMI was categorized into quartiles. The analysis revealed that higher BMI quartiles were associated with an increased prevalence of CVD, as demonstrated by the distribution in quartiles and visualizations like heatmaps.

6. **Lifestyle Risk Factors**
   Chi-square tests were conducted to assess the association between smoking, alcohol consumption, and physical activity with CVD. The results indicated:

- Physical inactivity was significantly associated with CVD (p=0.015).
- Smoking and alcohol consumption showed no statistically significant associations.

7. **Age Distribution**
   CVD patients were, on average, older (mean age: ~55 years) compared to non-CVD individuals (~52 years).

Key Findings

- Higher BMI is associated with cardiovascular disease. Individuals with CVD have a higher mean BMI and are more represented in the higher BMI quartiles consistent with findings from previous studies emphasizing that cardiovascular disease risk was increased among those who were overweight. (Hubert et al. 1983).
- Older age is also associated with CVD, as evidenced by the higher mean and median ages of CVD patients compared to non-CVD individuals.
- Physical inactivity is identified as a significant risk factor for cardiovascular disease (p=0.015 in the chi-square test), aligning with research highlighting sedentary behavior as a modifiable CVD risk factor (Lee et al. 2012).

Conclusion:
This study highlights BMI, age, and physical inactivity as critical factors associated with cardiovascular disease. The results emphasize the importance of managing weight and promoting physical activity to reduce CVD risk. Further inferential testing is needed to establish causality.

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).