home_work_test

Baodong Zhang

Introduction

Cardiovascular diseases (CVD) are a leading cause of mortality worldwide. 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

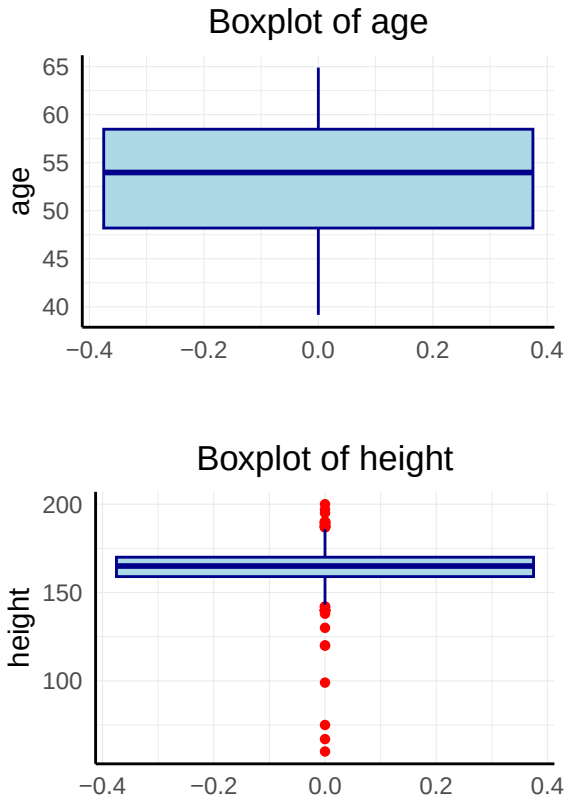
Chapter 2: Outlier Detection

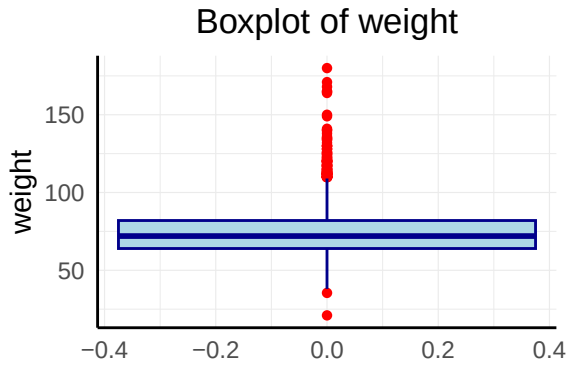
Task 2: Check the continuous variables for outliers and remove implausible values.

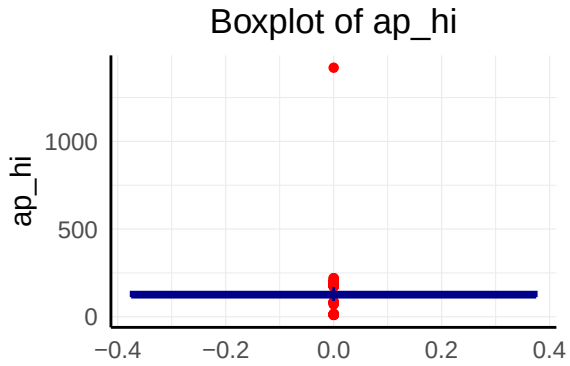
Table 2: Summary of Continuous Variables

	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

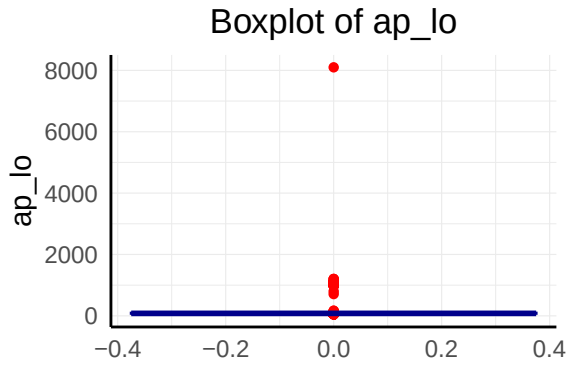
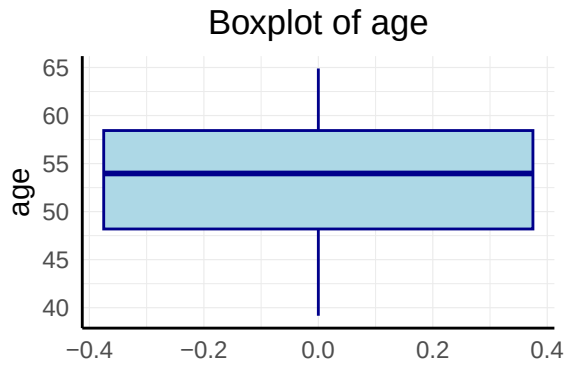
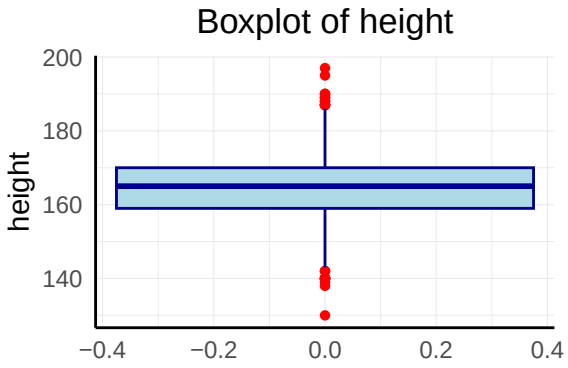

Boxplot of ap_lo
ap_lo
8000
6000
4000
2000
0
−0.4
−0.2
0.0
0.2
0.4

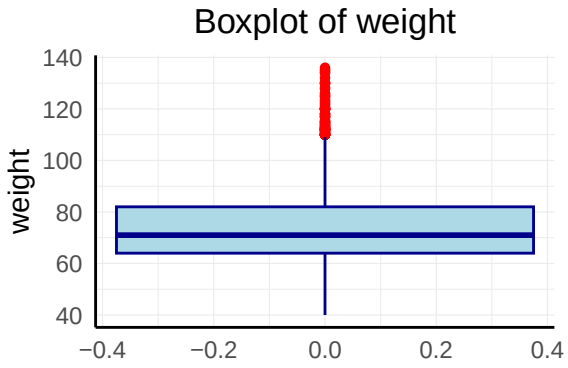
Table 3: Summary of Continuous Variables (outliers are removed)

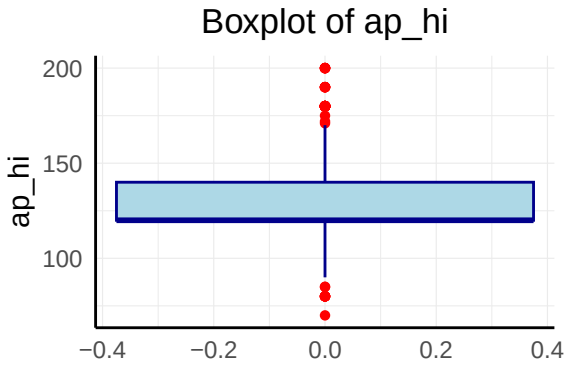
	age	height	weight	ap_hi	ap_lo
	Min. :39.16	Min. :130.0	Min. : 40.00	Min. : 70.0	Min. : 53.00
	1st Qu.:48.19	1st Qu.:159.0	1st Qu.: 64.00	1st Qu.:120.0	1st Qu.: 80.00
	Median :53.97	Median :165.0	Median : 71.00	Median :120.0	Median : 80.00
	Mean :53.30	Mean :164.4	Mean : 73.71	Mean :126.4	Mean : 81.25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

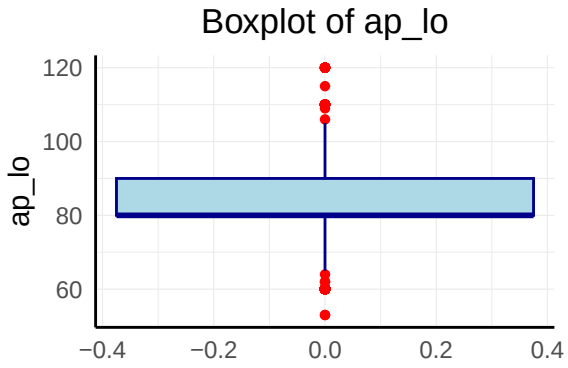
Boxplot after outliers are removed








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

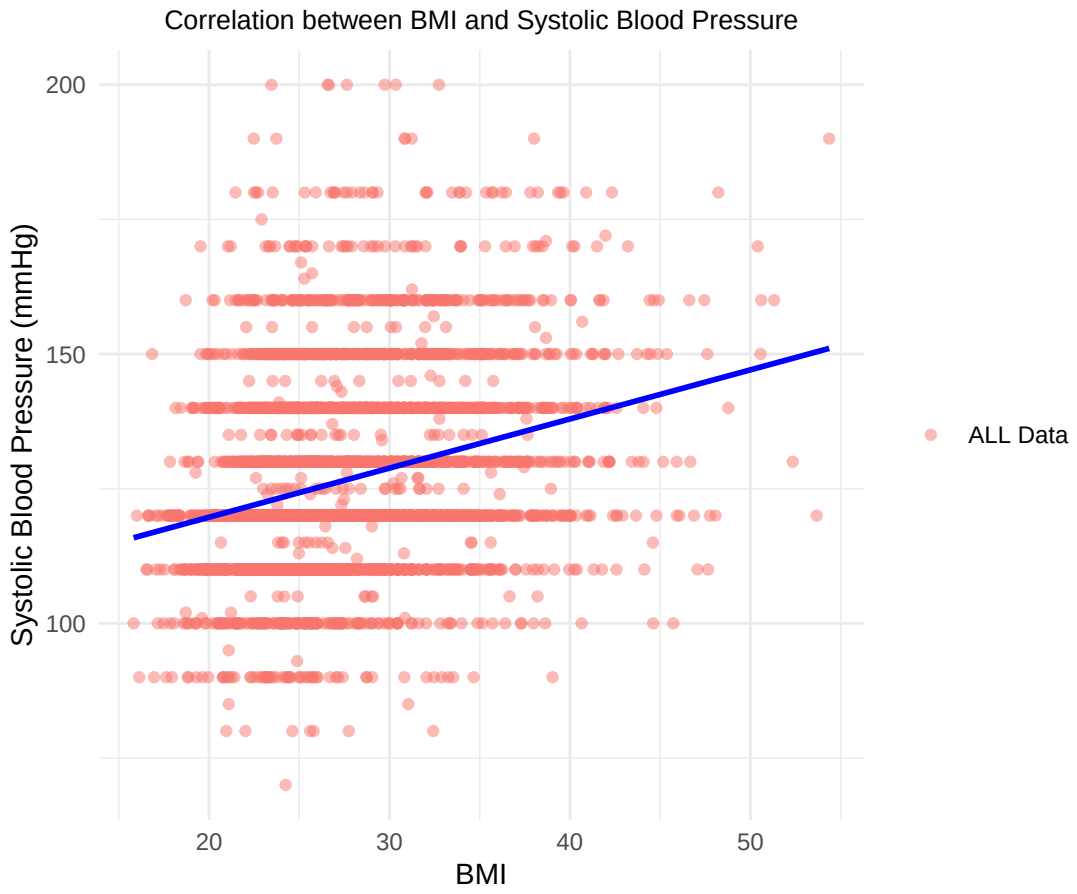
Task 3: Create a new variable BMI and provide a summary table for the variable BMI for both cardio groups.

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

Chapter 4: Correlation Between BMI and Systolic Blood Pressure

Task 4: How does the systolic blood pressure 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 is 0.279. This correlation indicates a weak positive correlation between BMI and systolic blood pressure. That means as BMI increases, systolic blood pressure (ap_hi) tends to increase slightly. However, this relationship is not very strong.

To answer the question whether there is any difference between the two classes of cardiovascular disease, we need to calculate the correlation for the two classes separately and perform a Fisher's Z-test.

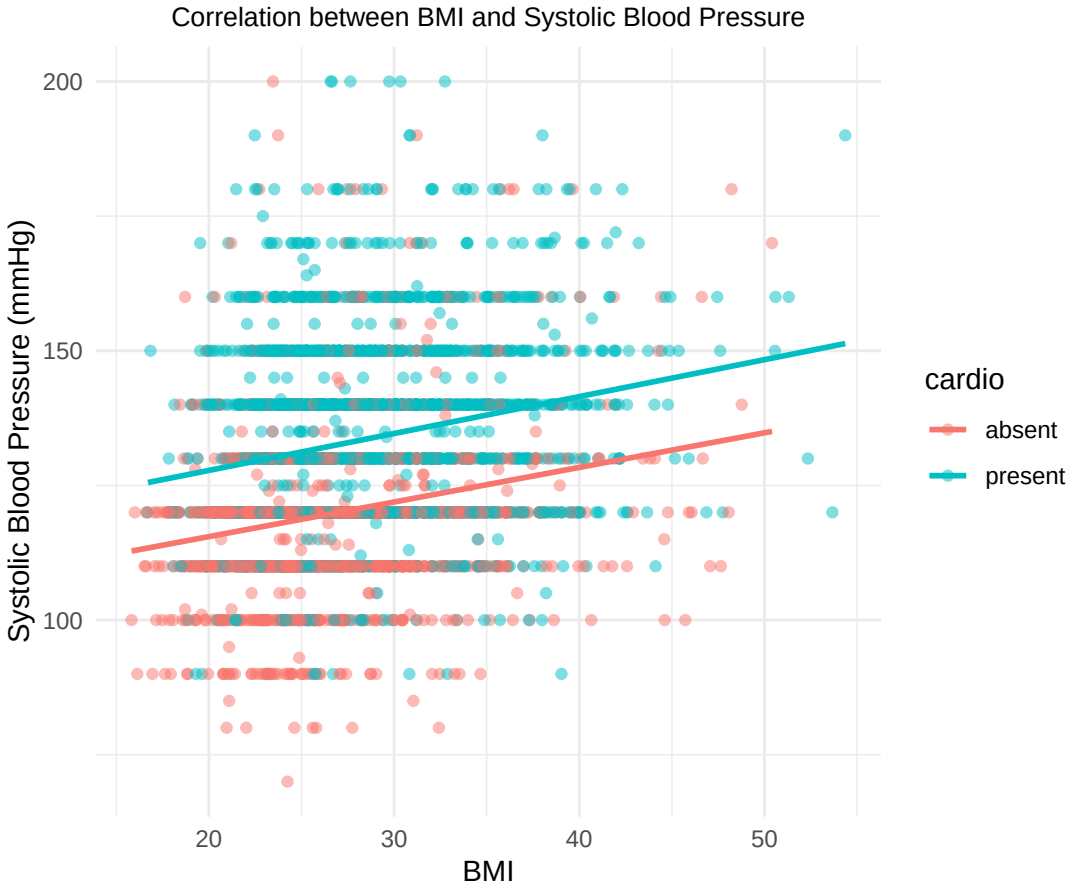


Table 5: Correlation between BMI and Systolic blood pressure grouped by cardio

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 0.2542.

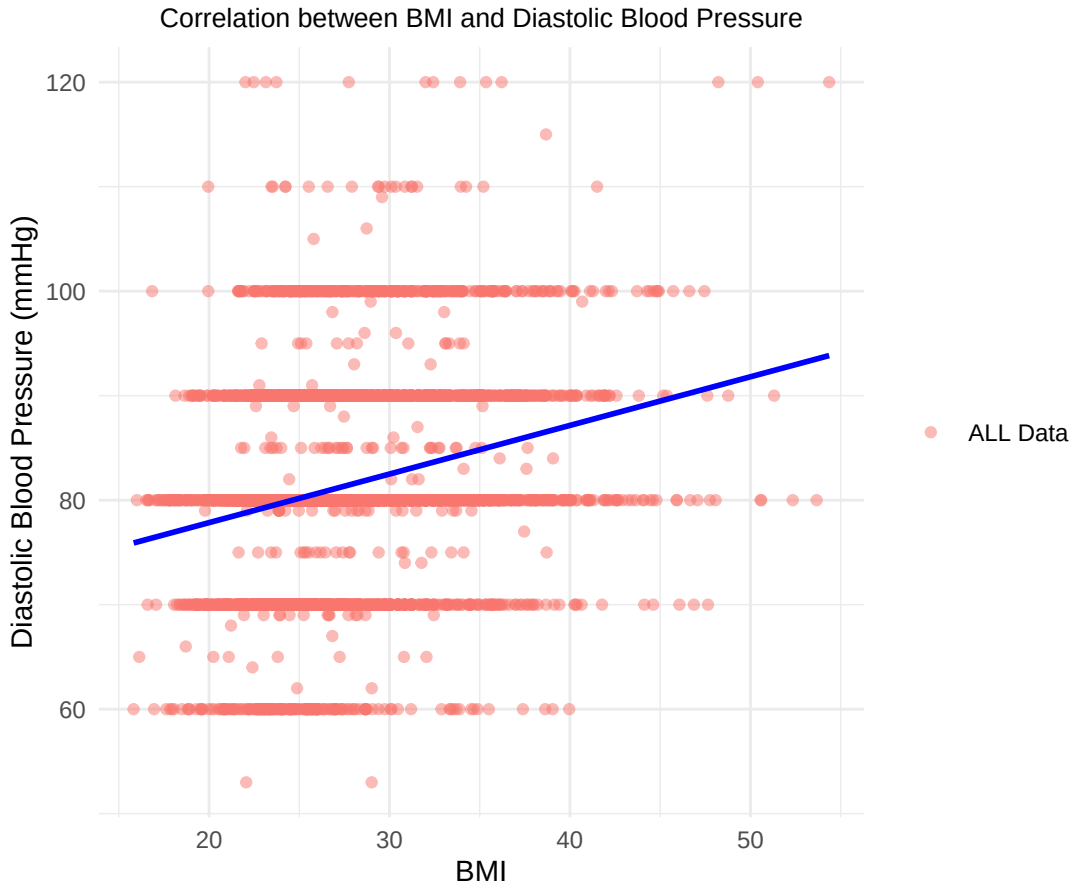


Table 7: Correlation between BMI and Diastolic blood pressure grouped by cardio

cardio	Correlation
absent	0.2233765
present	0.1869794

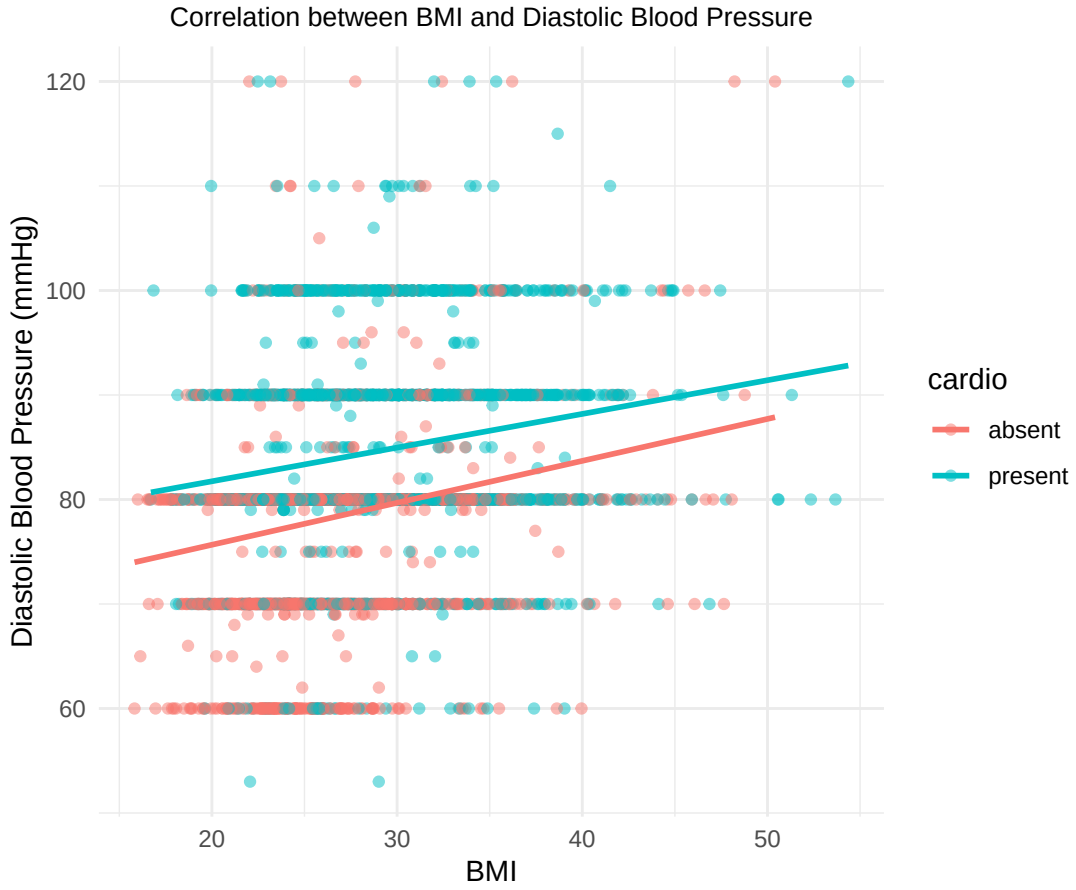


Table 8: Fisher's Z-test for Correlations between BMI and Diastolic blood pressure grouped by cardio

Metric	Value
Z_score	1.3260487
P_value	0.1848236

Answer: The correlation coefficient between BMI and Diastolic Blood Pressure (ap_lo) is 0.2542, indicating a weak positive correlation. This suggests that as BMI increases, Diastolic Blood Pressure tends to rise slightly, though the relationship is not strong.

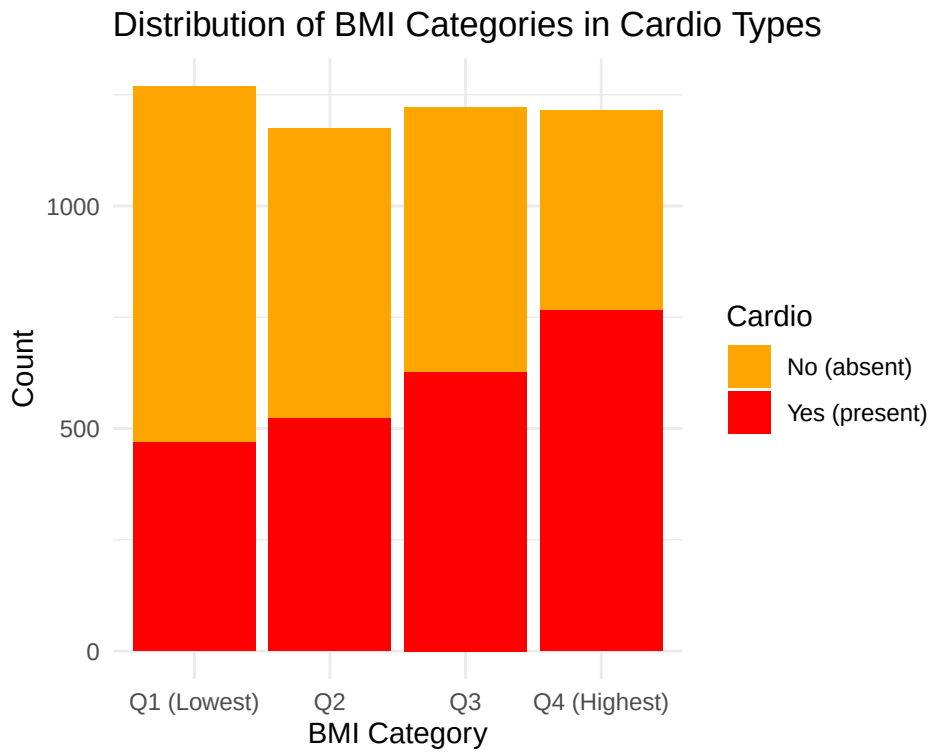
Based on Fisher's Z-test, the Z-score is 1.3260487, and the P-value is 0.1848236. These results show no statistically significant difference in the correlation coefficients between BMI and Diastolic Blood Pressure across the two groups of individuals with/without cardiovascular disease. The observed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

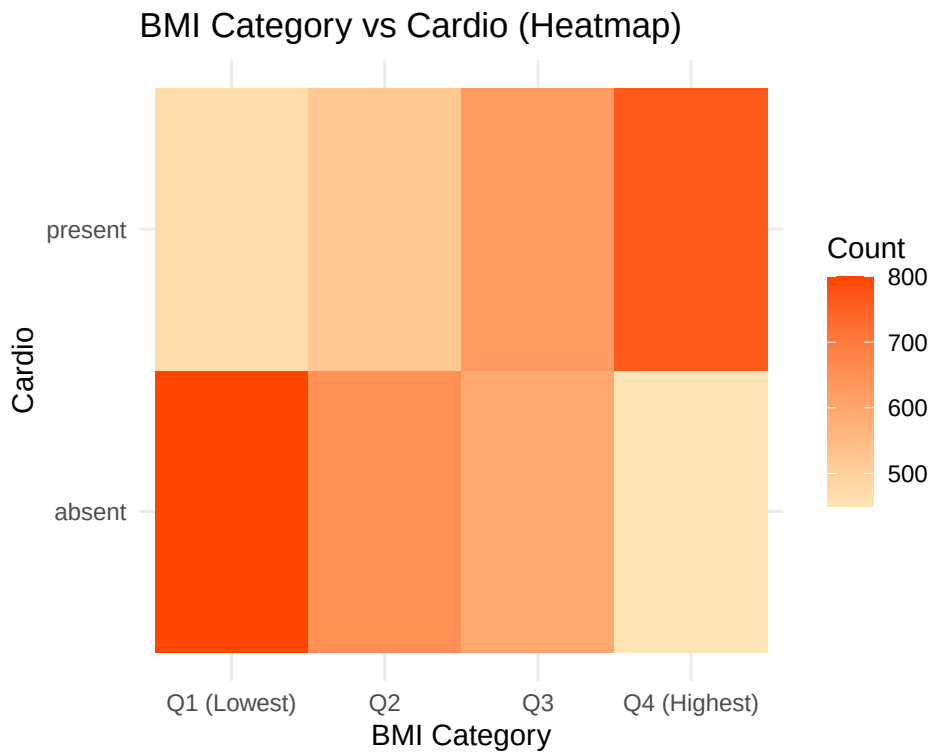
Chapter 6: Categorize BMI into Quartiles and Visualize its Distribution Across Cardio Types

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)

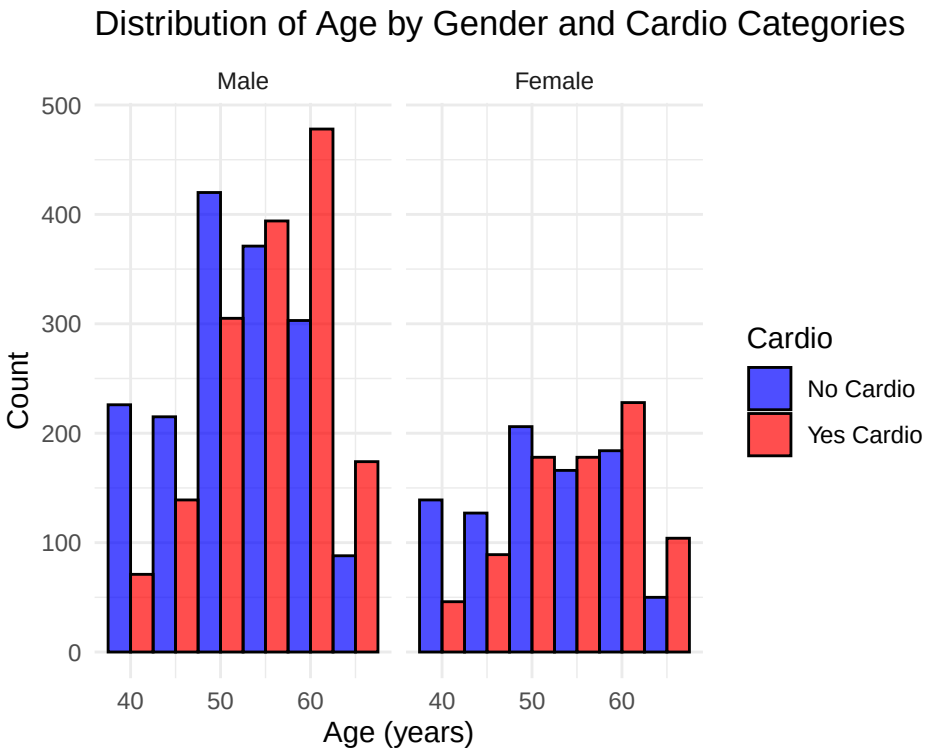
Task 7: How is age distributed in the different categories of cardio? Display age in years.

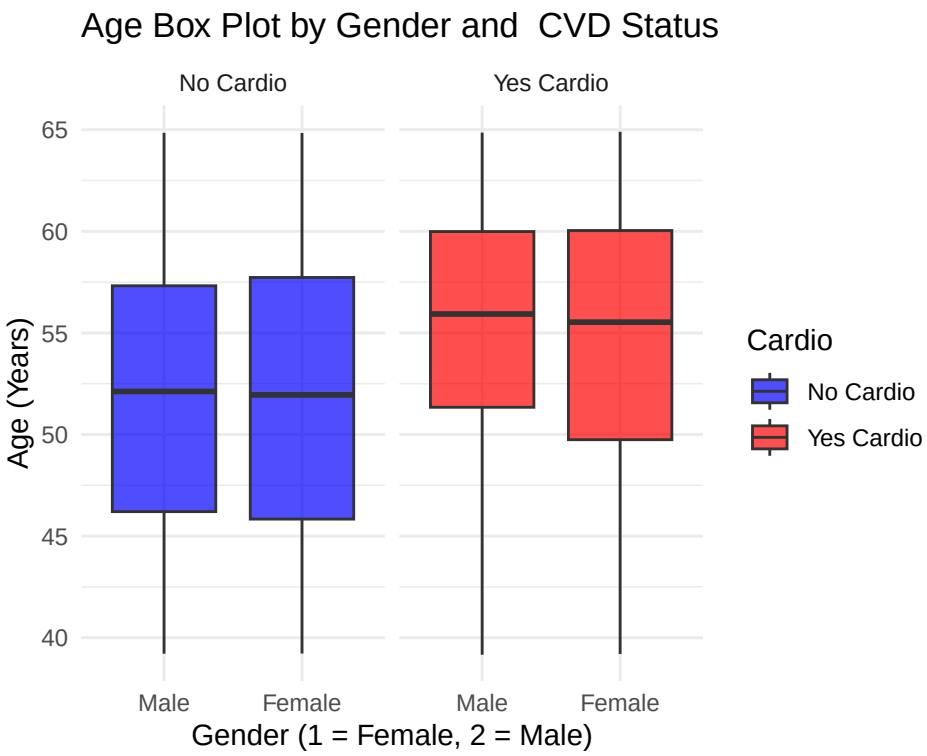
Table 10: Summarize age distribution for each cardio group

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

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

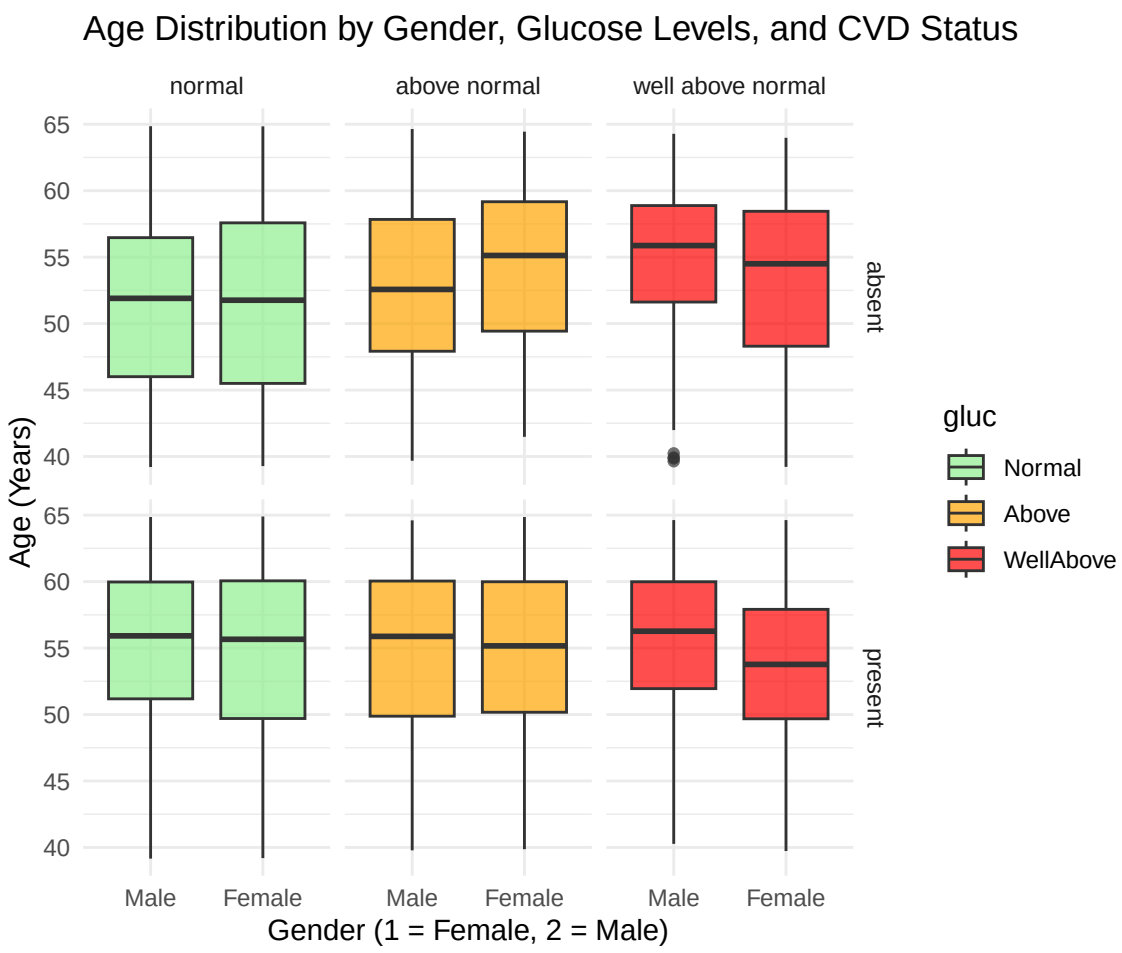
Task 8: Create a plot that shows the distribution of age for both types of gender and both types of cardio



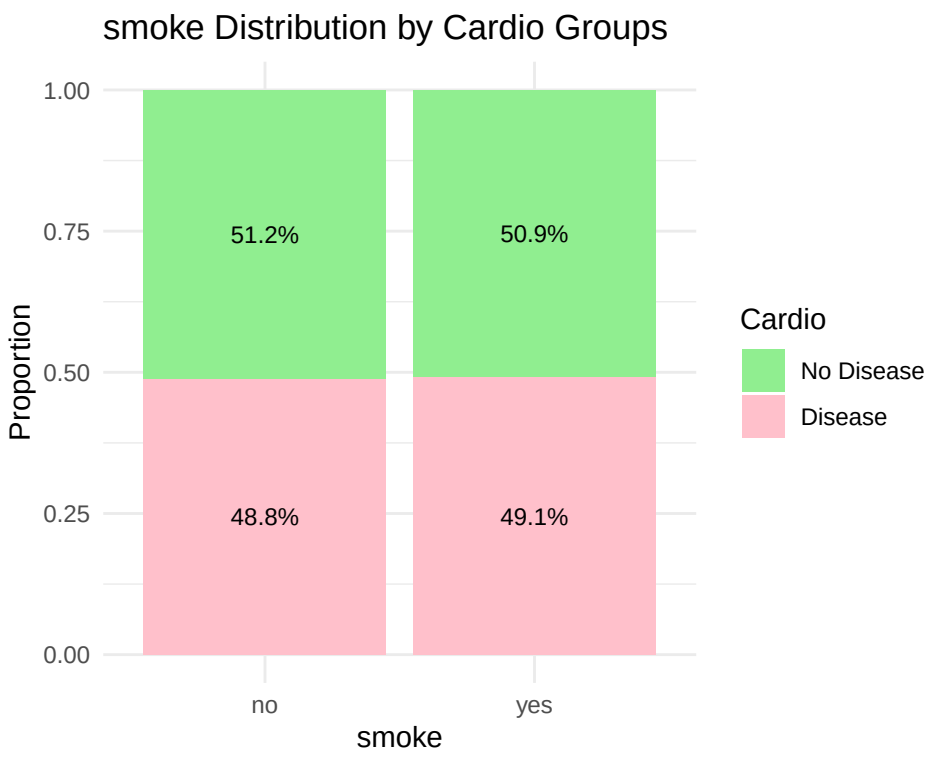

Age Box Plot by Gender and CVD Status
No Cardio
Yes Cardio
65
60
55
50
45
40
Age (Years)
Male
Female
Male
Female
Gender (1 = Female, 2 = Male)
Cardio
No Cardio
Yes Cardio

Chapter 9: Age Distribution with Glucose Levels

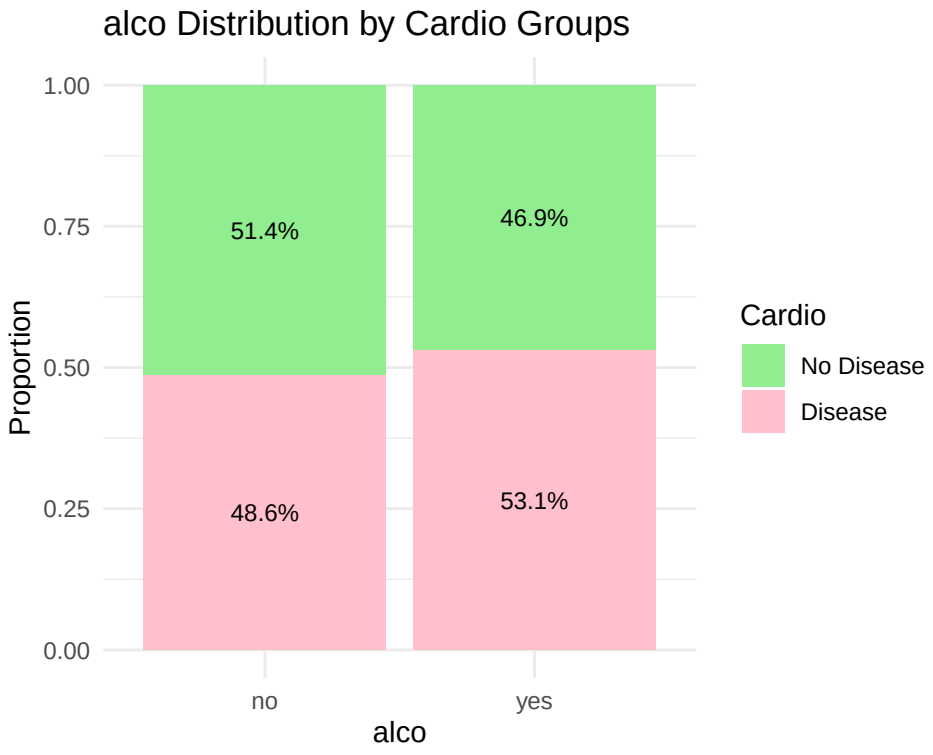
Task 9. Extend this plot by taking the different types of glucose into account.



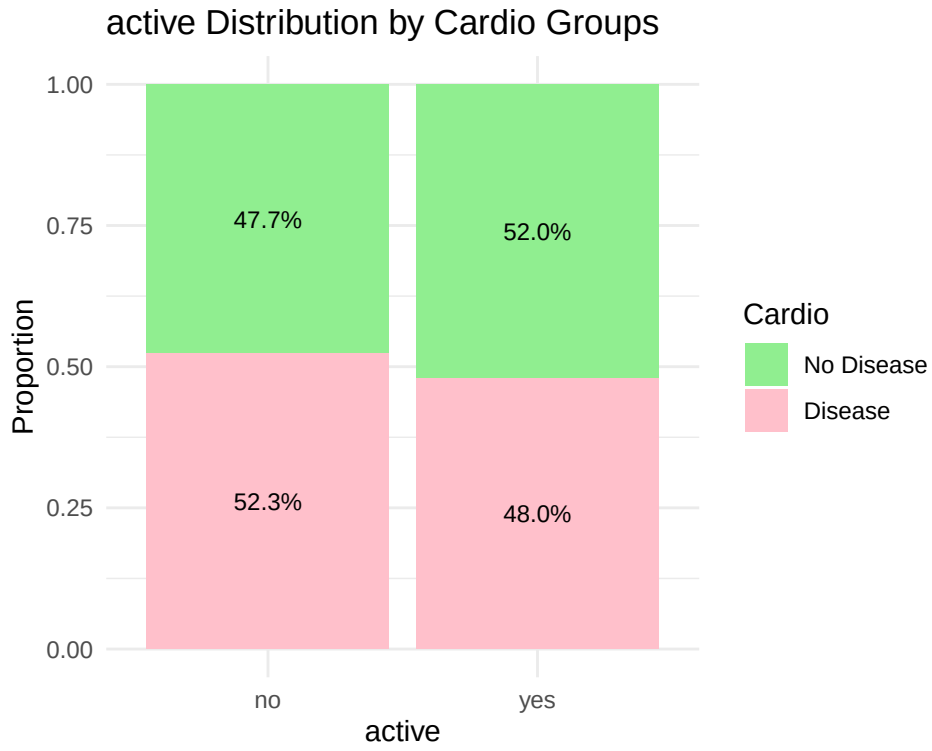

```
[[1]]
```


smoke Distribution by Cardio Groups
1.00
0.75
0.50
0.25
0.00
Proportion
51.2%
50.9%
48.8%
49.1%
no
yes
smoke
Cardio
No Disease
Disease

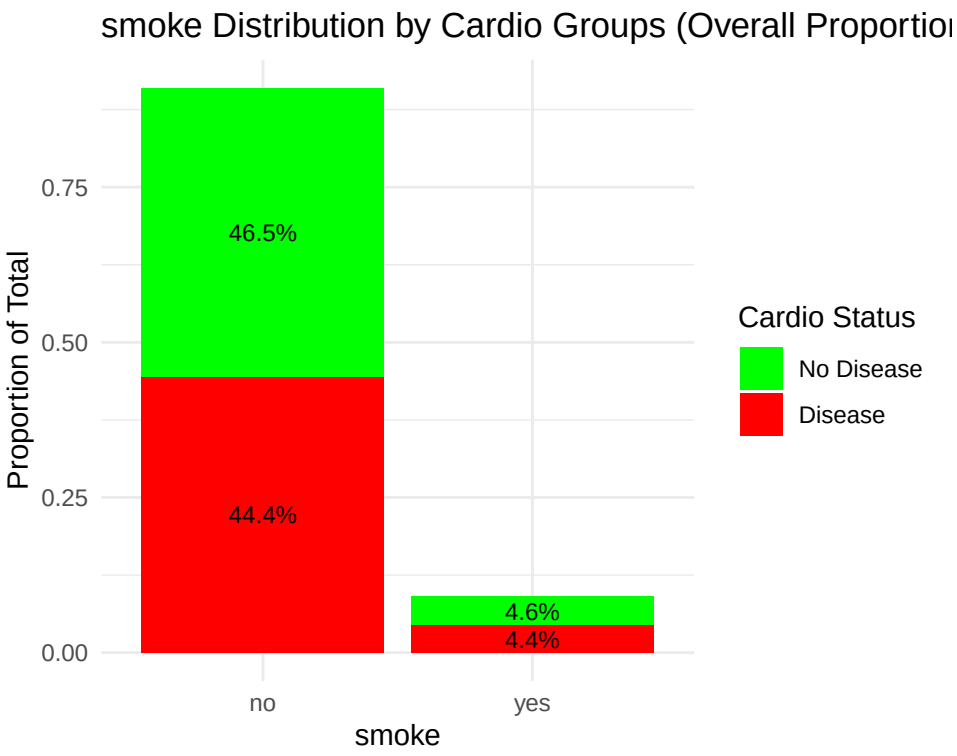
[[2]]


alco Distribution by Cardio Groups
1.00
0.75
0.50
0.25
0.00
Proportion
51.4%
46.9%
48.6%
53.1%
no
yes
alco
Cardio
No Disease
Disease

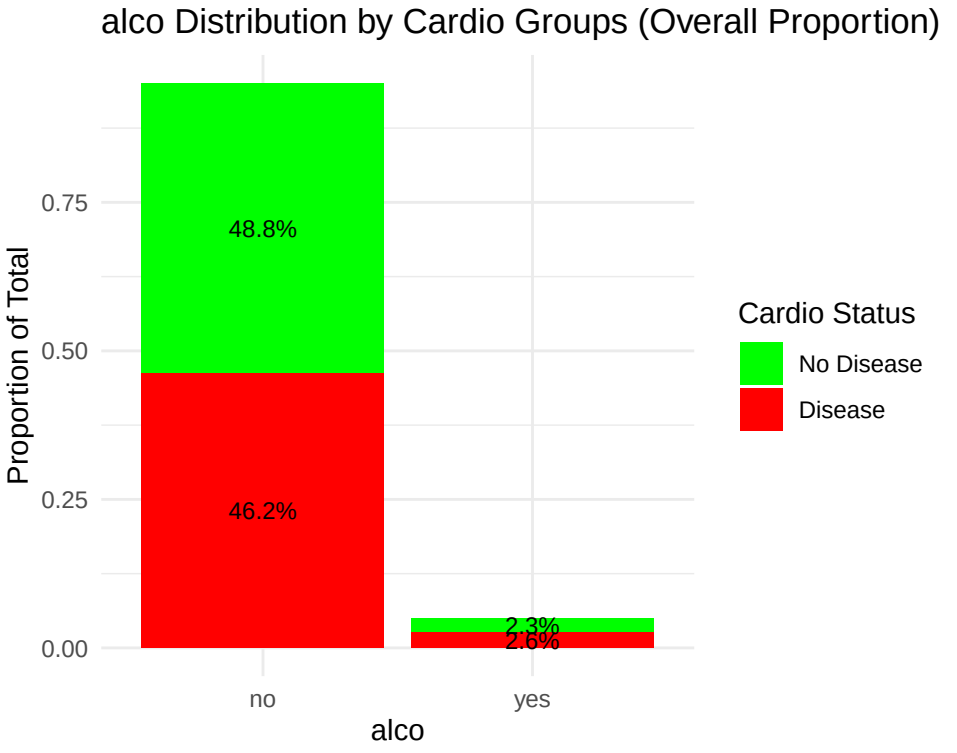
```
[[3]]
```


active Distribution by Cardio Groups
1.00
0.75
0.50
0.25
0.00
Proportion
47.7%
52.0%
52.3%
48.0%
no
yes
active
Cardio
No Disease
Disease

```
[[1]]
```


smoke Distribution by Cardio Groups (Overall Proportio
Proportion of Total
0.75
0.50
0.25
0.00
46.5%
44.4%
4.6%
4.4%
no
yes
smoke
Cardio Status
No Disease
Disease

[[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]]
```

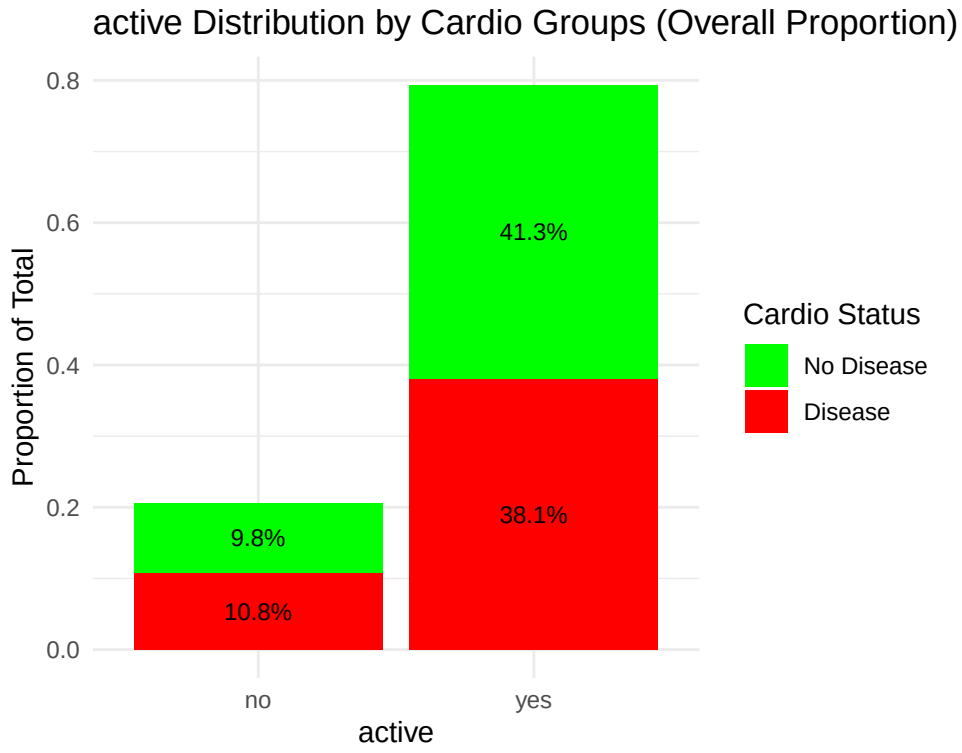


Table 11: propotion for each cardio group

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

```
$Smoking_Test

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

data:  smoking_table
X-squared = 0.0025508, df = 1, p-value = 0.9597


$Alcohol_Test

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

data:  alcohol_table
X-squared = 1.6577, df = 1, p-value = 0.1979
```

```
$Activity_Test

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

data:  activity_table
X-squared = 5.945, df = 1, p-value = 0.01476
```