APPLIED STATISTICS REPORT

ANALYSIS ON FEATURES CAUSING CARDIOVASCULAR DISEASES

Module Coordinator

Dr. Atif Azad

Co-Guide

Aliyu Sambo

Submitted By

Meera Mohan Mullamkuzhy (22185279)
MSc Big Data Analytics

Faculty of Computing, Engineering and the Built Environment

December 2022

Contents

List of Figures

List of Tables

1 EXECUTIVE SUMMARY

Ischemic heart disease, which accounts for 16% of the world fatalities, is one of the leading causes of death. The second and third leading causes of death are stroke and chronic obstructive pulmonary diseases (WHO, 2020). Identifying the root causes of cardiovascular disease can lower mortality rates globally.
Major question that arises here is to find the factors causing Cardiovascular diseases (CVDs). Age, Body Mass Index, hypertension, smoking habits, alcohol consumption, physical inactivity, abnormal cholesterol, and uncontrolled glucose levels can stimulate Cardiovascular diseases.

2 INTRODUCTION

2.1 PROBLEM DOMAIN

As per World Health Organisation(WHO)(2020), cardiovascular diseases are the major causes of mortality worldwide. WHO (2021) also says that, the Cardiovascular diseases should be detected and treated as early as possible. Cardiovascular disorders are more likely to occur when blood pressure, cholesterol, hyperglycemia, and obesity are not under control (Reyhaneh Rajab Boloukat et al., 2018). WHO has mentioned that "most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol" (WHO, 2021). Edward G. Lakatta (2022) concludes that there is an age-associated increase in the Cardiovascular diseases. A person's gender may also contribute to this disease. Lori Mosca et al. (2011) say that women outnumber men in terms of CVD survival and death. This study statistically examines the factors like age, gender, blood pressure, glucose levels, cholesterol levels, Body Mass Index, smoking habits, alcohol consumption, and physical inactivity and their relationship with Cardiovascular disease.

2.2 STATISTICAL QUESTIONS

The following questions are statistically analysed in this report

1. Does ageing contributes to Cardiovascular diseases?
2. Are Cardiovascular diseases more common in women than men?
3. Is obesity a major contributor to CVD?
4. Are all individuals with Cardiovascular disease obese?
5. Is hypertension a necessary component of Cardiovascular disease?
6. Can a Cardiovascular condition be brought on by abnormal glucose and cholesterol levels?
7. How can alcoholism, smoking, and sedentary lifestyles affect the development of cardiovascular diseases?

3 METHODOLOGY

This section includes details of the statistical techniques used to analyse the data and to solve the questions mentioned in section 2.

1. **Exploratory Data Analysis**
   Descriptive Summary Statistics is the initial step in data analysis.EDA was performed to pre-process the data for analysis as mentioned in section 4.1.
   Boxplots were plotted to summarise the data to identify mean value, IQR and outliers in the data.Histogram graphically represents frequency distribution of the data and is used to visualise the skewness of the data. It was used to identify the relationship of age, blood pressure, and BMI with Cardiovascular diseases.
   Bar charts were used to perform a comparison of values across different features in the dataset. It was used to identify the relationship of gender, cholesterol, glucose levels, smoking habit, alcohol consumption, and physical inactivity with Cardiovascular diseases.

2. **Hypothesis Testing**

   - **Pearson's Chi-Squared Test**
     Chi-Squared test was used to find the statistical significance of categorical features such as gender, cholesterol, glucose levels, and behavioural habits with Cardiovascular diseases.Details of the test done are given in the section 5.2.1.
   - **t-test**
     t-test was used to compare the means of two groups and to gain insightful information on the contribution of obesity and hypertension to Cardiovascular diseases. The details of the tests carried out are given in the section 5.2.2.

3. **Logistic Regression**
   Logistic Regression was used to analyse the significance of age, blood pressure, and BMI with Cardiovascular diseases. Additionally, a model that assesses the aggregate impact of the chosen features on the disease was created using it. The details of the tests carried out are given in the section 5.3.

4 DATASETS

The cardiovascular disease dataset includes data that was gathered from people during physical examinations. Three input features are included in the dataset: factual information, examination results, and patient-provided information. There are 70,000 rows and 13 columns in the dataset

Dataset Field Details			
Feature	**Feature Type**	**Variable name**	**Feature Data Type**
ID	NA	id	int
Age	objective	age	int (days)
Height	objective	height	int (cm)
Weight	objective	weight	float (kg)
Gender	objective	gender	categorical
Systolic blood pressure	examination	ap_hi	int
Diastolic blood pressure	examination	ap_lo	int
Cholesterol	examination	cholesterol	1: normal 2: above normal 3: well above normal
Glucose	examination	gluc	1: normal 2: above normal 3: well above normal
Smoking habit	subjective	smoke	binary
Alcohol consumption	subjective	alco	binary
Physical activity	subjective	active	binary
Presence or absence of cardiovascular disease	target	cardio	binary

Table 1 : *Details of the Dataset Fields*

Table 1 lists the feature type, variable name, and data type of all the features available in the dataset. The dataset has enough data to provide answers to the queries listed in section 2. As given in section 4.1, combination of systolic blood pressure and diastolic blood pressure, as well as a combination of weight and height, were utilised to answer the questions. The features age, gender, cholesterol, glucose, smoking habit, alcohol consumption, and physical activity were directly employed. There are 70,000 rows and 13 columns in the dataset.

4.1 DATA PRE-PROCESSING

The following operations are done to pre-process the data :

- The existence of null values is verified
- The dataset is examined to see if any null values or NAs are present.
- Outliers are eliminated from the fields for height, weight, systolic pressure, and diastolic pressure. Details of the outliers removal are given in the section 5.1.
- To make it easier to comprehend, the feature 'age' which is presented in days, is translated to years. Outliers are also eliminated.

- The first column id is eliminated since it is ineffective.
- Duplicates are removed.
- The features gender, smoke, alco, active, cholesterol, gluc, and cardio are converted to factors.

5 RESULTS AND DISCUSSION

5.1 Exploratory Data Analysis

As mentioned in 3, EDA is used to clean and summarise the dataset features.

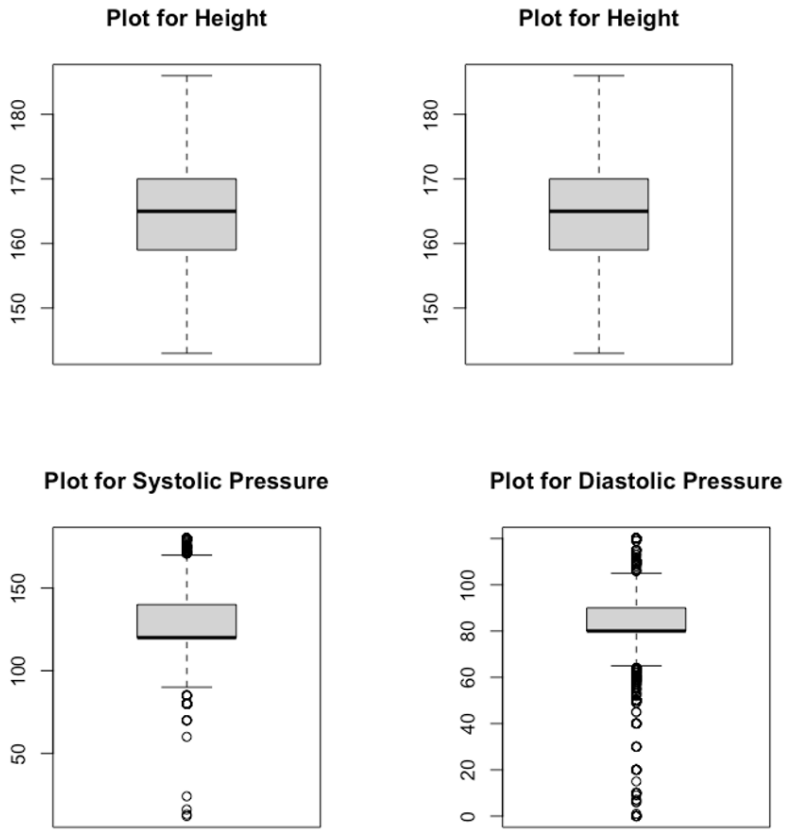

Figure 1: *Boxplot of the features height, weight, systolic blood pressure, and diastolic blood pressure after outlier elimination*

Figure 1 shows the boxplots for features after outlier elimination. Outliers of the features height and weight are eliminated using Interquartile Range approach. The systolic blood pressure (ap_hi) and diastolic blood pressure (ap_lo) fields is cleaned up of outliers. The results for ap_hi and ap_li that are greater than 180 mmHg and 120 mmHg, respectively, is eliminated as outliers (Robin Madell and Erica Hersh, 2021). Since the blood pressure can never be negative, negative readings are also dropped. Systolic pressure will be higher than diastolic pressure. Additional values are dropped.A new feature, BMI, is formed by dividing weight in kilograms by height in meter squared. Extreme obesity is identified as a BMI value greater than 60 kg/m2, and underweight is identified as a case of high risk when it is less than 9 kg/m2 (Wikipedia, 2022). The rare high risk values are removed.

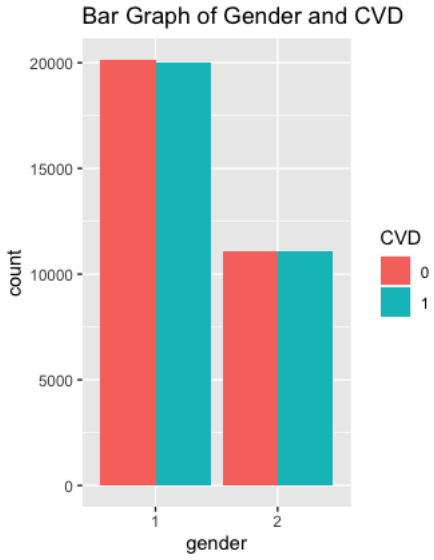

Figure 2: *Plot between Gender and CVD*

Figure 2 shows the count of women (gender = 1) and men (gender = 2) with and without Cardiovascular disease. The percentage of men and women with Cardiovascular diseases is 49.87591 and 49.86312 respectively. In the dataset, 0.013% of men have Cardiovascular disease than women.

From Figure 3, it is clear that the number of individuals with Cardiovascular disease increases with the increase in the cholesterol and glucose levels.

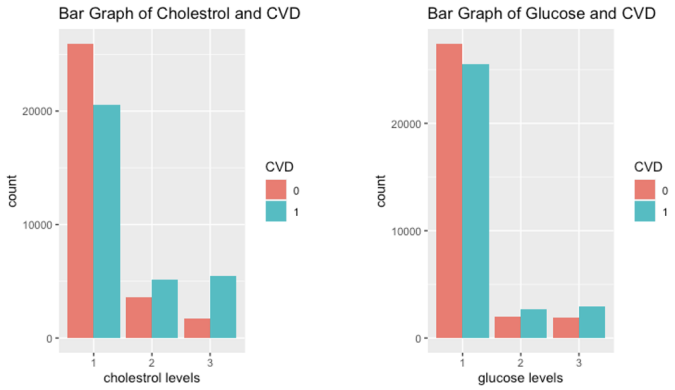

Figure 3: *Plot between Cholesterol, Glucose levels and CVD*

Figure 4 shows that the Cardiovascular diseases are observed to be less in individuals who are physically active.

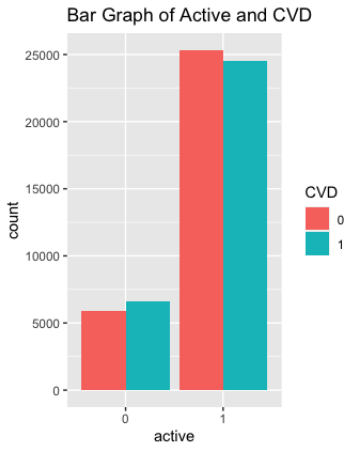

Figure 4: *Plot between Physical Activity and CVD*

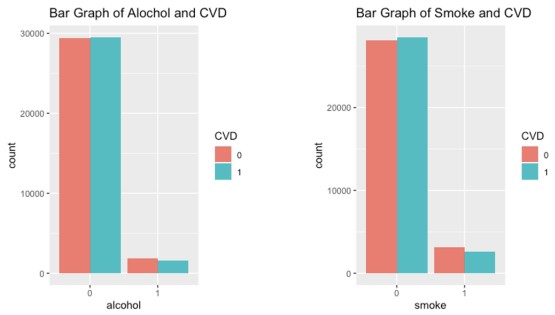

Figure 5: *Plot between Alcohol, Smoking and CVD*

Figure 5 shows that in the selected dataset, the number of patients with smoking habits and alcohol consumption who underwent medical examination is less than the people without these behaviorial risks.

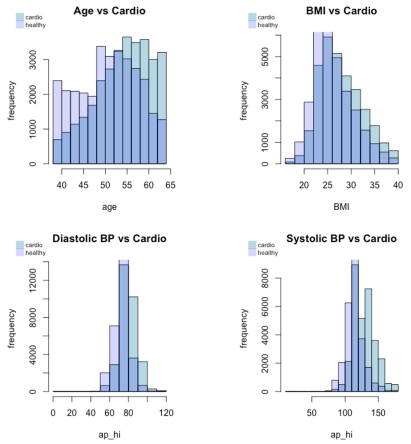

Figure 6: *Plot between Age, BMI, Blood Pressure and CVD*

From figure 6, it is evident that the features age, Body Mass Index, and blood pressure have a relationship with CVD.

5.2 Hypothesis Testing

5.2.1 Pearson's Chi-squared test

As mentioned in section 3, Chi-Squared test is carried out to find the relationship of the features gender, cholesterol levels, glucose levels, smoking habit, alcohol consumption, and physical activity with Cardiovascular diseases. Table 2 lists the p-value of the each of the above mentioned features.

T-test Results	
Feature	**p-value**
Gender	0.9823
Cholesterol	< 2.2e-16
Glucose	< 2.2e-16
Smoking	2.319e-11
Alochol Consumption	2.261e-05
Physical Activity	8.173e-13

Table 2 : *Features and p-value obtained from Chi-Squared Test*

The p-value of all the features except gender is <0.05. The test failed to reject the null hypothesis that there is no relationship between the feature and CVD. The p-value of 0.9823 shows that there is no relationship between the feature gender and Cardiovascular diseases.

5.2.2 t-test

As mentioned in section 3, t-test is performed to confirm whether hypertension and obesity is a necessary component of Cardiovascular disease. As the size of the dataset is >30 t-test can be performed even if the data is not normally distributed.

1. **Test to check if hypertension is a necessary factor of Cardiovascular disease**

   - Systolic blood pressure and CVD
     The systolic blood pressure values for hypertension are > 140 mmHg (William B White et al., 1989).

     Null hypothesis : mean value of ap_hi is >=140
     Alternate hypothesis : mean of ap_hi is < 140

     t-test results shows that the p-value is <2.2e-16 and the mean of the sample is 126.2957. As the p-value is <0.05, the test rejected the null hypothesis. Thus, we can conclude that in the given dataset, there are no enough evidence to accept that systolic blood pressure is a mandatory component of Cardiovascular diseases.

- Diastolic blood pressure and CVD
  The systolic blood pressure values for hypertension are > 140 mmHg (William B White et al., 1989).

  Null hypothesis : mean value of ap_lo is >=90
  Alternate hypothesis : mean of ap_lo is < 90

  t-test results shows that the p-value is <2.2e-16 and the mean of the sample is 81.03521. As the p-value is <0.05, the test rejected the null hypothesis. Thus, we can conclude that in the given dataset, there are no enough evidence to accept that diastolic blood pressure is a mandatory component of Cardiovascular diseases.

2. **Test to check if Obesity is a necessary feature of Cardiovascular disease**
   The BMI values for obesity is > 30 kg/m2 (Jacob C Seidell and Katherine M Flegal, 1997)

   Null hypothesis : mean value of bmi is >=30
   Alternate hypothesis : mean of ap_hi is < 30

   t-test results shows that the p-value is <2.2e-16 and the mean of the sample is 126.2957. As the p-value is <0.05, the test rejected the null hypothesis. Thus, we can conclude that in the given dataset, there are no enough evidence to accept that Systolic blood pressure is a mandatory component of Cardiovascular diseases.

5.3 Logistic Regression

Logistic Regression was performed to find the significance of each feature age, BMI, and blood pressure with Cardiovascular diseases. Table 3 lists the p-value obtained from Logistic Regression model for each of the feature.

Logistic Regression Results		
Feature	**p-value**	**Accuracy**
age	< 2e-16	59.51%
Systolic blood pressure	< 2e-16	70%
Diastolic blood pressure	<2e-16	65.32%
BMI	<2e-16	57.1%

Table 3 : *Features and p-value obtained from Logistic Regression model*

The null hypothesis for the Logistic Regression model is that there is no relationship between the input features. As the p-value of the features evaluated are less than 0.05, the test rejected the null hypothesis. Thus, it can be concluded that there exists a relationship between the above mentioned features and Cardiovascular disease.

Another Logistic Regression model was created considering all the features as input. Figure 7 shows summary of the Logistic Regression model.

```
glm(formula = cardio ~ age + ap_hi + ap_lo + bmi, family = "binomial",
    data = train_reg)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9743  -0.9540  -0.3097   0.9628   3.8189

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.050958   0.162749 -74.046  < 2e-16 ***
age           0.052614   0.001781  29.539  < 2e-16 ***
ap_hi         0.058764   0.001212  48.470  < 2e-16 ***
ap_lo         0.011545   0.001838   6.282 3.35e-10 ***
bmi           0.034582   0.002769  12.487  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 51854  on 37405  degrees of freedom
Residual deviance: 42777  on 37401  degrees of freedom
AIC: 42787

Number of Fisher Scoring iterations: 4
```

Figure 7: *Summary of Logistic Regression model with age, BMI, and blood pressure as features*

Logistic Regression model with the selected features gives an accuracy of 71.85% .

Figure 8 is the ROC curve for the model. Higher the area under the curve, the better the model.

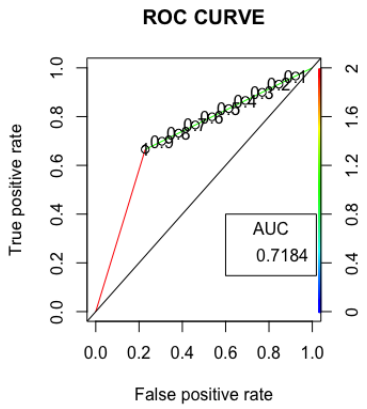

Figure 8: *ROC curve of Logistic Regression model with age, BMI, and blood pressure as features*

6 Conclusion

The primary aim of this statistical was to find the factors causing Cardiovascular diseases. As per the study, the percentage of men and women surviving or dead with Cardiovascular disease is almost similar. Gender cannot be considered as a factor for Cardiovascular diseases. Ageing, high blood pressure, obesity, uncontrolled cholesterol and glucose levels , and BMI can contribute to the disease. Though hypertension and obesity can increase the chances of disease, these factors cannot be considered as a mandatory component. Behavioural habits like alcoholism and smoking may cause Cardiovascular diseases. The absence of physical activity make increase the chances of the disease.

7 References

1. Svetlana Ulianova (2018) *Cardiovascular Disease dataset*. Available at: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset [Accessed 7 December 2022]

2. World Health Organisation (2020) *The top 10 causes of death*. Available at: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death [Accessed 7 December 2022]

3. World Health Organisation (2021) *Cardiovascular diseases (CVDs)*. Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds [Accessed 7 December 2022]

4. Reyhaneh Rajab Boloukat, Azra Ramezankhani, Mitra Hasheminia, Erfan Tasdighi, Fereidoun Azizi, FarzadHadaegh (2018) *Impact of blood pressure, cholesterol and glucose in the association between adiposity measures*

and coronary heart disease and stroke among Iranian population. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0261561417313559 [Accessed 7 December 2022]

5. Edward G. Lakatta (2022) *Age-associated Cardiovascular Changes in Health: Impact on Cardiovascular Disease in Older Persons.* Available at: https://link.springer.com/content/pdf/10.1023/A:1013797722156.pdf?pdf=inline%20link [Accessed 7 December 2022]

6. Lori Mosca, Elizabeth Barrett-Connor Nanette Kass Wenger (2011) *Sex/-Gender Differences in Cardiovascular Disease Prevention.* Available at: https://www.ahajournals.org/doi/full/10.1161/CIRCULATIONAHA.110.968792#d3e488 [Accessed 7 December 2022]

7. Robin Madell Erica Hersh (2021) *https://www.healthline.com/health/high-blood-pressure-hypertension/blood-pressure-reading-explained.* Available at: https://www.healthline.com/health/high-blood-pressure-hypertension/blood-pressure-reading-explained [Accessed 7 December 2022]

8. Wikipedia (2022) *Body mass index.* Available at: https://en.wikipedia.org/wiki/Body_mass_index [Accessed 7 December 2022]

9. William B. White, MD; Peter Schulman, MD; Ellen J. McCabe, RN; et al (1989) *Average Daily Blood Pressure, Not Office Blood Pressure, Determines Cardiac Function in Patients With Hypertension.* Available at: https://jamanetwork.com/journals/jama/article-abstract/376280 [Accessed 7 December 2022]

10. Jacob C Seidell and Katherine M Flegal (1997) *Assessing obesity: classification and epidemiology.* Available at: https://academic.oup.com/bmb/article/53/2/238/600180 [Accessed 7 December 2022]

8 Appendix

```
#---------------------------------REQUIRED LIBRARIES---------------------------------
install.packages("ggplot2")
library(ggplot2)
install.packages("dplyr")
library(dplyr)
install.packages("ltm") #Bi-serial
library(ltm)
install.packages("corrplot")
library(corrplot)
#Hypothesis testing
install.packages("dplyr")
library(dplyr)
install.packages("ggpubr")
library(ggpubr)


#---------------------------------DATA PREPROCESSING---------------------------------
```

```r
#Read the csv file
cardio_data <- read.csv2("/Users/meeramohanm/BCU/Assignments/Applied␣Statistics/card
ncol(cardio_data)
nrow(cardio_data)


#1. Check for Null Values
is.null(cardio_data)

#2. Checking for missing values
table(is.na.data.frame(cardio_data))

#3. Check for duplicate values (no duplicate values if the statement returns the val
sum(duplicated(cardio_data))

#4. Weight and height values check outlier removal through IQR

#height check -smallest value is 54cm and highest is 251cm
boxplot(cardio_data$height)
#so many outliers in boxplot. So check for min and max value
range(cardio_data$height)
Q1 <- quantile(cardio_data$height, .25)
Q3 <- quantile(cardio_data$height, .75)
IQR <- IQR(cardio_data$height)
height_lower_range <- Q1 - (1.5*IQR)
height_higher_range <- Q3 + (1.5*IQR)
cardio_cleaned <- subset(cardio_data, cardio_data$height > (Q1 - 1.5*IQR) & cardio_c
boxplot(cardio_cleaned$height, main="Plot␣for␣Height")

#weight check
summary(cardio_cleaned)
cardio_cleaned$weight <- as.numeric(cardio_cleaned$weight)
summary(cardio_cleaned)
boxplot(cardio_cleaned$weight)
range(cardio_cleaned$weight)
Q1 <- quantile(cardio_cleaned$weight, .25)
Q3 <- quantile(cardio_cleaned$weight, .75)
IQR <- IQR(cardio_cleaned$weight)
weight_lower_range <- Q1 - (1.5*IQR)
weight_higher_range <- Q3 + (1.5*IQR)
cardio_data_cleaned <- subset(cardio_cleaned, cardio_cleaned$weight > (Q1 - 1.5*IQR)
boxplot(cardio_data_cleaned$weight,main="Plot␣for␣Weight")

#5. Outlier detection for blood pressure(ap_hi,ap_lo)
#show references
boxplot(cardio_data_cleaned$ap_hi)
range(cardio_data_cleaned$ap_hi)
boxplot(cardio_data_cleaned$ap_lo)
range(cardio_data_cleaned$ap_lo)
cardio_data <- subset(cardio_data_cleaned,cardio_data_cleaned$ap_lo >= 0 & cardio_da
cardio_data <- subset(cardio_data,cardio_data$ap_lo <= 120 & cardio_data$ap_hi <= 18
cardio_data <- subset(cardio_data,cardio_data$ap_lo < cardio_data$ap_hi)
boxplot(cardio_data$ap_hi,main="Plot␣for␣Systolic␣Pressure")
boxplot(cardio_data$ap_lo,main="Plot␣for␣Diastolic␣Pressure")
```

```r
#5. Outlier detection for age
#show references
boxplot(cardio_data$age)
outliers <- boxplot(cardio_data$age, plot=FALSE)$out
cardio_data <- cardio_data[-which(cardio_data$age %in% outliers),]
boxplot(cardio_data$age)

#6. Addition of column BMI
cardio_data_new <- mutate(cardio_data, bmi=cardio_data$weight/((cardio_data$height/
cardio_data_new <- subset(cardio_data_new,cardio_data_new$bmi < 40 & cardio_data_ne
boxplot(cardio_data_new$bmi)


#show references

#7. id column drop
cardio_data <- cardio_data_new[,-1]
head(cardio_data)

#8. Convert age from days to years
cardio_data$age <- floor(cardio_data$age/365)

#9. na,missing and duplicates values check done. Removal of duplicate values
cardio_data <- unique(cardio_data) #index reset

#Converting binary variables to factors
cardio_data$gender <- as.factor(cardio_data$gender)
cardio_data$smoke <- as.factor(cardio_data$smoke)
cardio_data$alco <- as.factor(cardio_data$alco)
cardio_data$active <- as.factor(cardio_data$active)
cardio_data$cardio <- as.factor(cardio_data$cardio)
cardio_data$cholesterol <- as.factor(cardio_data$cholesterol)
cardio_data$gluc <- as.factor(cardio_data$gluc)
str(cardio_data)


#------------------------------------------------------------------------------
##############################################################################
#---------------------------QUESTIONS AND ANSWERS-----------------------------
#------------------------------------------------------------------------------

#Chi-squared Test
#The chi-squared test in R is used to assess whether there is a statistically signi
#association between two categorical variables. The interpretation of the results o
#chi-squared test in R depends on the p-value that is obtained. If the p-value is l
#0.05, this indicates that there is a statistically significant association between
#On the other hand, if the p-value is greater than 0.05, this indicates that there
#significant association between the two variables.

#The solution for each question is given below

#---------------------------------------*START OF QUESTION 1*--------------------
```

```
#Is cardiovascular disease prevalent among men or women

#::SOLUTION::

#gender is a categorical variable ( 1- Women and 2-Men)
#plot a graph to find the ratio of men and women with CVD
plot <- ggplot(cardio_data, aes(x=cardio_data$gender, fill=cardio_data$cardio)) +
  geom_bar(position="dodge") +
  ggtitle("Bar Graph of Gender and CVD") +
  xlab("gender")
plot + labs(fill = "CVD")


#calculation of percentages of cardiac men and women
women_total_data <- subset(cardio_data, cardio_data$gender == 1)
men_total_data<- subset(cardio_data,cardio_data$gender == 2)
cardio_total_data <- subset(cardio_data, cardio_data$cardio == 1)
women_cardio_data<- subset(cardio_total_data, cardio_total_data$gender == 1)
men_cardio_data <- subset(cardio_total_data,cardio_total_data$gender == 2)
women_with_cardio_percentage = (nrow(women_cardio_data)/nrow(women_total_data)) * 1
men_with_cardio_percentage = (nrow(men_cardio_data)/nrow(men_total_data)) * 100
cat("Percentage of women with cardiac disease is",women_with_cardio_percentage) #P
49.86312
cat("Percentage of men with cardiac disease is ", men_with_cardio_percentage)
#Percentage of men with cardiac disease is
difference <- abs(women_with_cardio_percentage - men_with_cardio_percentage)
cat("Difference in percentage is ", difference) #Difference in percentage is
0.01278534

#find the correlation between feature gender and CVD
chisq_gender <- chisq.test(cardio_data$gender,cardio_data$cardio)
print(chisq_gender) #-squared = 0.00049254, df = 1, p-value = 0.9823

#----------------------------------------*END OF QUESTION 1 *----------------------

#----------------------------------------*START OF QUESTION 2*----------------------
#Does cholestrol levels have any impact on cardiovascular disease

#::SOLUTION::


#plot a graph of cholesterol with CVD
plot <- ggplot(cardio_data, aes(x=cardio_data$cholesterol, fill=cardio_data$cardio)
  geom_bar(position="dodge") +
  ggtitle("Bar Graph of Cholestrol and CVD") +
  xlab("cholestrol levels")
plot + labs(fill = "CVD")

#find the correlation between feature cholesterol and CVD
chisq_cholesterol <- chisq.test(cardio_data$cholesterol,cardio_data$cardio)
print(chisq_cholesterol)
#X-squared = 2810.5, df = 2, p-value < 2.2e-16
#----------------------------------------*END OF QUESTION 2 *----------------------
```

```r
#----------------------------------------*START OF QUESTION 3 *--------------------
#Do glucose levels have any impact on cardiovascular disease

#::SOLUTION::

#plot a graph of glucose with CVD
plot <- ggplot(cardio_data, aes(x=cardio_data$gluc, fill=cardio_data$cardio)) +
  geom_bar(position="dodge") +
  ggtitle("Bar Graph of Glucose and CVD") +
  xlab("glucose levels")
plot + labs(fill = "CVD")

#find the correlation between feature glucose and CVD
chisq_gluc <- chisq.test(cardio_data$gluc,cardio_data$cardio)
print(chisq_gluc)
#X-squared = 414.77, df = 2, p-value < 2.2e-16
#----------------------------------------*END OF QUESTION 3 *----------------------


#----------------------------------------*START OF QUESTION 4*---------------------
#Smoking habit and CVD

#::SOLUTION::

#plot a graph of smoke with CVD
plot <- ggplot(cardio_data, aes(x=cardio_data$smoke, fill=cardio_data$cardio)) +
  geom_bar(position="dodge") +
  ggtitle("Bar Graph of Smoke and CVD") +
  xlab("smoke")
plot + labs(fill = "CVD")

#find the correlation between feature smoke and CVD
chisq_smoke <- chisq.test(cardio_data$smoke,cardio_data$cardio)
print(chisq_smoke)
# X-squared = 44.681, df = 1, p-value = 2.319e-11

#----------------------------------------*END OF QUESTION 4 *----------------------

#----------------------------------------*START OF QUESTION 5*---------------------
#Drinking habit and CVD

#::SOLUTION::

#plot a graph of alcohol with CVD
plot <- ggplot(cardio_data, aes(x=cardio_data$alco, fill=cardio_data$cardio)) +
  geom_bar(position="dodge") +
  ggtitle("Bar Graph of Alochol and CVD") +
  xlab("alcohol")
plot + labs(fill = "CVD")

#find the correlation between feature alcohol and CVD
chisq_alco <- chisq.test(cardio_data$alco,cardio_data$cardio)
print(chisq_alco)
#X-squared = 17.956, df = 1, p-value = 2.261e-05
```

```
#----------------------------------------*END OF QUESTION 5*----------------------

#----------------------------------------*START OF QUESTION 6*---------------------
#Physical activity and CVD

#::SOLUTION::

#plot a graph of active with CVD
plot <- ggplot(cardio_data, aes(x=cardio_data$active, fill=cardio_data$cardio)) +
  geom_bar(position="dodge") +
  ggtitle("Bar Graph of Active and CVD") +
  xlab("active")
plot + labs(fill = "CVD")

#find the correlation between feature active and CVD
chisq_active <- chisq.test(cardio_data$active,cardio_data$cardio)
print(chisq_active)
#X-squared = 51.24, df = 1, p-value = 8.173e-13

#----------------------------------------*END OF QUESTION 6*----------------------
#----End Of Categorical Variables

#*****************************************
#----Start of Continuous Variables


#----------------------------------------*START OF QUESTION 7*---------------------
#Age and CVD

#::SOLUTION::

#plot a graph of age with CVD
cardio_true_data <- subset(cardio_data,cardio_data$cardio == 1)
cardio_false_data <- subset(cardio_data,cardio_data$cardio == 0)

hist(cardio_true_data$age,col='lightblue',xlab="age",ylab = "frequency",main = "Age
hist(cardio_false_data$age,col=rgb(0,0,1,0.2), add=T)
legend("bottomright",legend=c("cardio","healthy"),col=c('lightblue',rgb(0,0,1,0.2))



#----------------------------------------*END OF QUESTION 7*----------------------

#----------------------------------------*START OF QUESTION 8*---------------------
#BMI and CVD

#::SOLUTION::

#plot a graph of BMI with CVD

hist(cardio_true_data$bmi,col='lightblue',xlab="BMI",ylab = "frequency",main = "BMI
hist(cardio_false_data$bmi,col=rgb(0,0,1,0.2), add=T)
legend("bottomright",legend=c("cardio","healthy"),col=c('lightblue',rgb(0,0,1,0.2))
```

```
#----------------------------------------*END OF QUESTION 8*----------------------

#----------------------------------------*START OF QUESTION 9*---------------------
#ap_hi and CVD

#::SOLUTION::

#plot a graph of ap_hi with CVD

hist(cardio_true_data$ap_hi,col='lightblue',xlab="ap_hi",ylab = "frequency",main =
hist(cardio_false_data$ap_hi,col=rgb(0,0,1,0.2), add=T)
legend("bottomright",legend=c("cardio","healthy"),col=c('lightblue',rgb(0,0,1,0.2))


#----------------------------------------*END OF QUESTION 9*----------------------

#----------------------------------------*START OF QUESTION 10*--------------------
#ap_lo and CVD

#::SOLUTION::

#plot a graph of ap_lo with CVD

hist(cardio_true_data$ap_lo,col='lightblue',xlab="ap_hi",ylab = "frequency",main =
hist(cardio_false_data$ap_lo,col=rgb(0,0,1,0.2), add=T)
legend("bottomright",legend=c("cardio","healthy"),col=c('lightblue',rgb(0,0,1,0.2))


#----------------------------------------*END OF QUESTION 11*----------------------

#----------------------------------------*START OF HYPOTHESIS TESTING*-------------
#----------**START OF TEST 1 :to test whether hypertension is mandatory**

#::SOLUTION::

ggdensity(cardio_data$ap_hi,
          main = "Density␣plot␣of␣ap_hi",
          xlab = "ap_hi")

#plot is unreliable.So go for kolmogorov-smirnov test
ks.test(cardio_data$ap_hi, 'pnorm')
#p-value is less. Indicates that the data is not normally distributed.

#hypothesis test => to check if the mean bp is >=140
#H0 : mean ap_hi is >= 140
#H1 : mean ap_hi is < 140

t.test(cardio_data$ap_hi,alternative = "less",mu=140)

#hypothesis test => to check if the mean bp is >=90
#H0 : mean ap_lo is >= 90
#H1 : mean ap_lo is < 90
t.test(cardio_data$ap_lo,alternative = "less",mu=90)
```

```
#------********END OF TEST 1**********-----------

#--------***START OF TEST 2 :to test whether all patients are obese****-------

#hypothesis test => to check if the mean bp is >=30
#H0 : mean bmi is >= 30
#H1 : mean bmi is < 30

t.test(cardio_data$bmi,alternative = "less",mu=30)

#------********END OF TEST 2**********-----------

#----------------------------------------*END OF HYPOTHESIS TESTING*----------------


#----------------------------------------*START OF LOGISTIC REGRESSION*-------------


#***LR model with all features*****
log_data <- cardio_data[c(1,5,6,12,13)]

split <- sample.split(log_data, SplitRatio = 0.7)
split

train_reg <- subset(log_data, split == "TRUE")
test_reg <- subset(log_data, split == "FALSE")

logistic_model <- glm(cardio ~ age + ap_hi + ap_lo + bmi,
                      data = train_reg,
                      family = "binomial")
logistic_model

# Summary
summary(logistic_model)

# Predict test data based on model
predict_reg <- predict(logistic_model,
                       test_reg, type = "response")
predict_reg

# Changing probabilities
predict_reg <- ifelse(predict_reg >0.5, 1, 0)

# Evaluating model accuracy
# using confusion matrix
table(test_reg$cardio, predict_reg)

missing_classerr <- mean(predict_reg != test_reg$cardio)
print(paste('Accuracy␣=', 1 - missing_classerr))

# ROC-AUC Curve
ROCPred <- prediction(predict_reg, test_reg$cardio)
ROCPer <- performance(ROCPred, measure = "tpr",
```

```
                        x.measure = "fpr")

auc <- performance(ROCPred, measure = "auc")
auc <- auc@y.values[[1]]
auc

# Plotting curve
plot(ROCPer)
plot(ROCPer, colorize = TRUE,
     print.cutoffs.at = seq(0.1, by = 0.1),
     main = "ROC␣CURVE")
abline(a = 0, b = 1)

auc <- round(auc, 4)
legend(.6, .4, auc, title = "AUC", cex = 1)
#************LR for age*********************
log_data <- cardio_data[c(1,12)]

split <- sample.split(log_data, SplitRatio = 0.7)
split

train_reg <- subset(log_data, split == "TRUE")
test_reg <- subset(log_data, split == "FALSE")

logistic_model <- glm(cardio ~ age ,
                      data = train_reg,
                      family = "binomial")
logistic_model

# Summary
summary(logistic_model)

# Predict test data based on model
predict_reg <- predict(logistic_model,
                       test_reg, type = "response")
predict_reg

# Changing probabilities
predict_reg <- ifelse(predict_reg >0.5, 1, 0)

# Evaluating model accuracy
# using confusion matrix
table(test_reg$cardio, predict_reg)

missing_classerr <- mean(predict_reg != test_reg$cardio)
print(paste('Accuracy␣=', 1 - missing_classerr))

#Accuracy = 0.595104423983831"

#************LR for ap_hi*********************
log_data <- cardio_data[c(5,12)]

split <- sample.split(log_data, SplitRatio = 0.7)
split
```

```
train_reg <- subset(log_data, split == "TRUE")
test_reg <- subset(log_data, split == "FALSE")

logistic_model <- glm(cardio ~ ap_hi ,
                      data = train_reg,
                      family = "binomial")
logistic_model

# Summary
summary(logistic_model)

# Predict test data based on model
predict_reg <- predict(logistic_model,
                       test_reg, type = "response")
predict_reg

# Changing probabilities
predict_reg <- ifelse(predict_reg >0.5, 1, 0)

# Evaluating model accuracy
# using confusion matrix
table(test_reg$cardio, predict_reg)

missing_classerr <- mean(predict_reg != test_reg$cardio)
print(paste('Accuracy␣=', 1 - missing_classerr))

#Accuracy = 0.706524152367662


#************LR for ap_lo*********************
log_data <- cardio_data[c(6,12)]

split <- sample.split(log_data, SplitRatio = 0.7)
split

train_reg <- subset(log_data, split == "TRUE")
test_reg <- subset(log_data, split == "FALSE")

logistic_model <- glm(cardio ~ ap_lo ,
                      data = train_reg,
                      family = "binomial")
logistic_model

# Summary
summary(logistic_model)

# Predict test data based on model
predict_reg <- predict(logistic_model,
                       test_reg, type = "response")
predict_reg

# Changing probabilities
predict_reg <- ifelse(predict_reg >0.5, 1, 0)
```

```
# Evaluating model accuracy
# using confusion matrix
table(test_reg$cardio, predict_reg)

missing_classerr <- mean(predict_reg != test_reg$cardio)
print(paste('Accuracy␣=', 1 - missing_classerr))

#Accuracy = 0.655385688911444

#************LR for bmi*********************
log_data <- cardio_data[c(13,12)]

split <- sample.split(log_data, SplitRatio = 0.7)
split

train_reg <- subset(log_data, split == "TRUE")
test_reg <- subset(log_data, split == "FALSE")

logistic_model <- glm(cardio ~ bmi ,
                      data = train_reg,
                      family = "binomial")
logistic_model

# Summary
summary(logistic_model)

# Predict test data based on model
predict_reg <- predict(logistic_model,
                       test_reg, type = "response")
predict_reg

# Changing probabilities
predict_reg <- ifelse(predict_reg >0.5, 1, 0)

# Evaluating model accuracy
# using confusion matrix
table(test_reg$cardio, predict_reg)

missing_classerr <- mean(predict_reg != test_reg$cardio)
print(paste('Accuracy␣=', 1 - missing_classerr))
```