NAME: DEVIKA PILLAI

PROJECT II REPORT

Introduction

Globally recognized as the leading cause of mortality, heart diseases encompass a range of conditions affecting the heart's blood vessels, often stemming from the accumulation of fatty deposits in arteries. In this project, I delved into the intricate landscape of cardiovascular health by exploring a comprehensive dataset focused on heart diseases, also known as cardiovascular diseases (CVD). This dataset, sourced from Kaggle, serves as a good repository of information related to individuals' cardiovascular health. Each row corresponds to a specific individual, while the columns encapsulate a diverse set of variables, including demographic factors and clinical measurements such as age, gender, chest pain type, blood pressure, cholesterol levels, and more. This exploration aims to unravel relationships between these variables and the presence or absence of heart disease, fostering insights that may contribute to our understanding of cardiovascular health and inform potential predictive models.
Dataset link: https://www.kaggle.com/datasets/amirmahdiabbootalebi/heart-disease/data

The columns of the dataset are:

1. Age: Numeric variable; Age of individuals
2. Sex: Categorical variable; Represents age of individuals with values M - Male, F - Female
3. ChestPainType: Categorical variable; Describes the type of chest pain experienced by individuals, with categories like "ASY" (Asymptomatic), "NAP" (Non-Anginal Pain), “ATA” (Atypical Angina), “TA” (Typical Angina).
4. RestingBP: Numerical variable; Represents resting blood pressure of individuals.
5. Cholesterol: Numerical variable; Represents serum cholesterol of individuals.
6. FastingBS: Numerical variable; fasting blood sugar with values 1 – fasting blood sugar > 120, 0 - otherwise.
7. RestingECG: Categorical variable; Resting electrocardiogram results with values Normal – Normal ECG, ST - having ST-T wave abnormality, LVH - Left Ventricular Hypertrophy.
8. MaxHR: Numerical variable; maximum heart rate achieved.
9. ExcerciseAngina: Categorical variable; exercise-induced angina with values Y – Yes, N – No.
10. Oldpeak: Numerical variable; oldpeak measured in depression.
11. ST_Slope: Categorical variable; the slope of the peak exercise ST segment with values Up : upsloping, Flat : flat, Down : down sloping.
12. HeartDisease: Numerical variable; output class with values 0 – No heart disease, 1 – heart disease.

```
df <- (heart)
head(df,10)
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
1	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
2	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
3	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
4	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
5	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
6	39	M	NAP	120	339	0	Normal	170	N	0.0	Up	0
7	45	F	ATA	130	237	0	Normal	170	N	0.0	Up	0
8	54	M	ATA	110	208	0	Normal	142	N	0.0	Up	0
9	37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
10	48	F	ATA	120	284	0	Normal	120	N	0.0	Up	0

Fig.1: Heart Disease Dataset

1. Data preparation

b. Data transformation

In the data preprocessing phase, a function has been developed to transform the dataset to enhance the quality and reliability. The function systematically addresses potential issues, starting with a check for missing values. If any missing values are detected, the function provides a summary of the missing values per column and takes a proactive step by removing rows containing any null values. In this dataset however, no null values were found. This approach ensures that the dataset remains free from the potential biases and inaccuracies introduced by missing data.

```
any_missing <- any(is.na(df))
 if (any_missing) {
  missing_per_column <- colSums(is.na(df))
  df <- df[complete.cases(df), ]
  print("Rows with missing values dropped.")
 } else {
  print("No missing values found.")
 }
```

In the process of scrutinizing the dataset, it was identified that certain entries within the Resting Blood Pressure (RestingBP) and Cholesterol variables exhibited unrealistic values, specifically recorded as 0. Such values are inconsistent with physiological norms in a real-world setting, as it is implausible for individuals to have resting blood pressure or cholesterol levels of zero. After doing this, the number of rows in the dataset drops from 918 to 746.

```
df <- subset(df, RestingBP != 0 & Cholesterol != 0)
```

Moreover, selected columns, including Sex, ChestPainType, FastingBS, RestingECG, ExerciseAngina, ST_Slope, and HeartDisease, were converted into factors. The resultant clean dataset is then ready for further exploration and modeling, setting a solid foundation for subsequent stages in the data science pipeline.

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
1	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
2	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
3	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
4	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
5	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
6	39	M	NAP	120	339	0	Normal	170	N	0.0	Up	0
7	45	F	ATA	130	237	0	Normal	170	N	0.0	Up	0
8	54	M	ATA	110	208	0	Normal	142	N	0.0	Up	0
9	37	M	ASY	140	207	0	Normal	130	Y	1.5	Flat	1
10	48	F	ATA	120	284	0	Normal	120	N	0.0	Up	0

Fig.2: Dataset after transformation

e. Univariate Statistics

```
getMode <- function(x1) {
 datavalues <- unique(x1)
 datavalues[which.max(tabulate(match(x1,datavalues)))]
}
#tabulating numerical data summary
```

```
generate_summary_stats <- function(df, columns) {
  stats <- data.frame(
    Statistics = c("Mean", "Median", "Mode", "Min", "Max", "Variance", "Standard Deviation", "0% Quantile",
"25% Quantile", "50% Quantile", "75% Quantile", "100% Quantile", "IQR")
  )
  for (col in columns) {
    col_stats <- c(
      mean(df[[col]]),
      median(df[[col]]),
      getMode(df[[col]]),
      min(df[[col]]),
      max(df[[col]]),
      var(df[[col]]),
      sd(df[[col]]),
      quantile(df[[col]], prob = c(0, 0.25, 0.5, 0.75, 1)),
      IQR(df[[col]])
    )
    stats[[col]] <- col_stats
  }
  return(stats)
}

numerical_columns <- c("Age", "RestingBP", "Cholesterol")
summary_stats <- generate_summary_stats(df, numerical_columns)
print(summary_stats)
```

Statistics	Age	RestingBP	Cholestrol
Mean	52.882038	133.02279	244.63539
Median	54.000000	130.00000	237.00000
Mode	54.000000	120.00000	254.00000
Min	28.000000	92.00000	85.00000
Max	77.000000	200.00000	603.0000
Variance	90.361905	298.69344	3499.13936
Standard Deviation	9.505888	17.28275	59.15352
0% Quantile	28.000000	92.00000	85.00000
25% Quantile	46.000000	120.00000	207.25000
50% Quantile	54.000000	130.00000	237.00000
75% Quantile	59.000000	140.00000	275.00000
100% Quantile	77.000000	200.00000	603.00000
IQR	13.000000	20.00000	67.75000

Tab. 1: Summary statistics

The average age in the dataset is approximately 52.9 years, suggesting a relatively mature population. The median age of 54 years indicates a central tendency, with half the individuals being 54 years or younger. The dataset spans from 28 to 77 years, indicating a diverse representation of age groups. The average resting blood pressure is 133 mm Hg, a value within the normal range. The median of 130 mm Hg aligns with the mean, reflecting a symmetric distribution. The resting blood pressure ranges from 92 to 200 mm Hg, indicating variability in blood pressure levels. The IQR of 20 mm Hg suggests moderate

dispersion around the median. The mean cholesterol level is 244.6 mg/dL, which is relatively high. Cholesterol levels span from 85 to 603 mg/dL, signifying considerable variability. The IQR of 67.8 mg/dL indicates the spread of cholesterol values around the median. While the majority have blood pressure and cholesterol levels within typical ranges, there is variability, and some individuals exhibit higher cholesterol levels.

```
library(ggplot2)
create_bar_plot <- function(data, variable) {
  ggplot(data, aes(x = !!as.symbol(variable), fill = factor(HeartDisease))) +
    geom_bar(position = "dodge") +
    scale_fill_manual(values = c("0" = "green", "1" = "red")) +
    labs(title = paste("Distribution of Heart Disease for", variable),
         x = variable,
         y = "Count",
         fill = "Heart Disease") +
    theme_minimal()
}

# Create bar plots for each categorical variable
categorical_variables <- c("Sex", "ChestPainType", "FastingBS", "RestingECG", "ExerciseAngina", "ST_Slope")
for (variable in categorical_variables) {
  plot_object <- create_bar_plot(df, variable)
  print(plot_object)
}
```

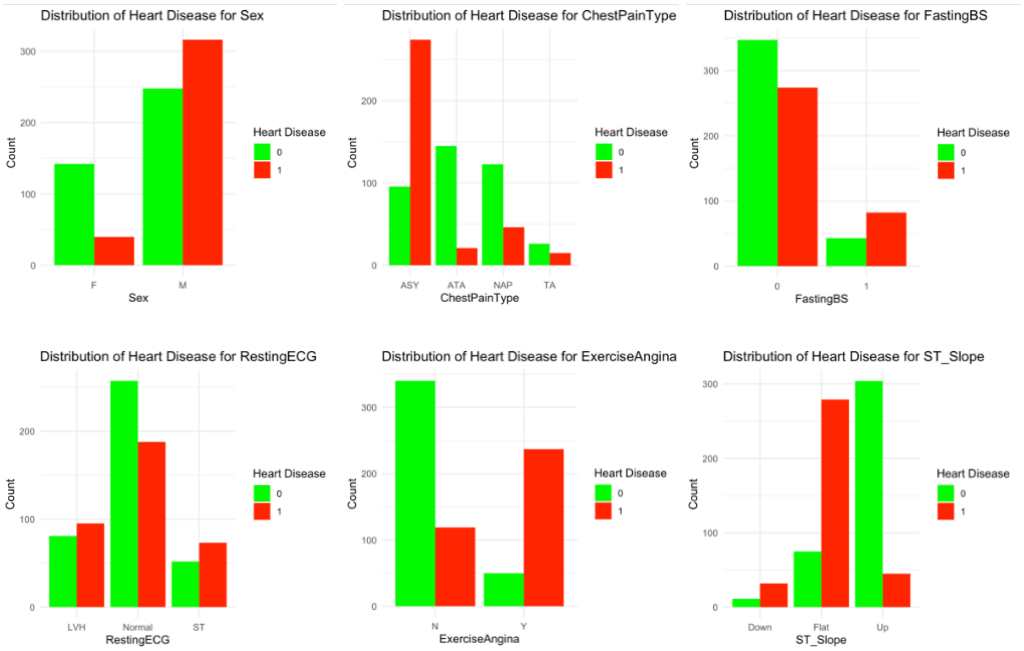

Fig. 3: Graphs for categorical variables

Figure 3, shows graphs that plot the categorical variables in the dataset and figure 4, shows graphs that plot the numerical variables in the dataset. All these graphs have been made using ggplot, functions and loops. It shows that in the dataset there are more instances of

males having heart disease than females. Chest Pain type ASY has the most instances of heart disease and TA has the least. However, fasting blood sugar values less than 120 has the most instances of heart disease. Normal resting ECG has the highest instances of heart disease and ST has the lowest instances. Individuals with exercise induced angina has the highest instances of heart disease. Flat ST Slope has the largest instances of heart disease and up has the largest instances of non-heart disease.

```
create_histogram <- function(data, variable) {
  ggplot(data, aes(x = !!as.symbol(variable), fill = factor(HeartDisease))) +
    geom_histogram(position = "dodge", bins = 20, alpha = 0.7) +
    scale_fill_manual(values = c("0" = "green", "1" = "red")) +
    labs(title = paste("Distribution of Heart Disease for", variable),
         x = variable,
         y = "Count",
         fill = "Heart Disease") +
    theme_minimal()
}

# Create histograms for each numerical variable
numerical_variables <- c("Age", "RestingBP", "Cholesterol", "MaxHR", "Oldpeak")
for (variable in numerical_variables) {
  plot_object <- create_histogram(df, variable)
  print(plot_object)
}
```

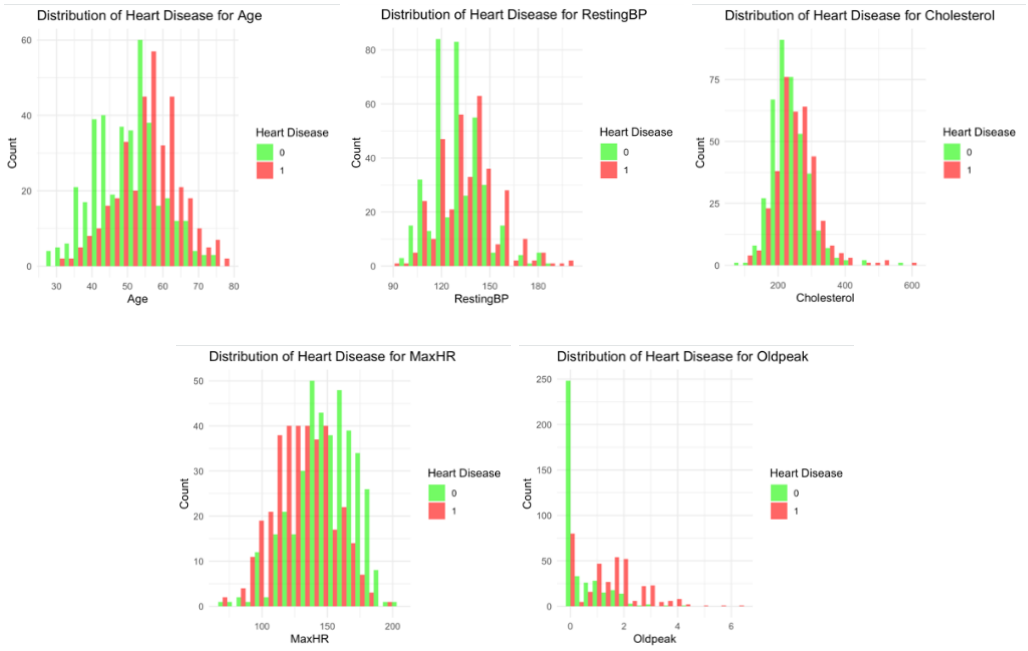

Fig. 4: Graphs for numerical variable

The graphs in figure 4 show that in the dataset approximately the age of 57 has the highest instances of positive heart disease diagnosis and the age of 52 has the highest instances of negative heart disease diagnosis. They keep reducing on both sides of the graph almost forming a normal distribution. Resting BP values of 140 approximately have the highest

positive heart disease diagnosis and the value 120 approximately have the highest negative heart disease diagnosis. Cholesterol values of approximately 220 have the have the highest positive heart disease diagnosis and 210 have the highest negative heart disease diagnosis. Max HR values of approximately 120, 130, 135 and 150 have the have the highest positive heart disease diagnosis and 140 have the highest negative heart disease diagnosis. Oldpeak values of 0 have the have the highest positive and negative heart disease diagnosis but the instances for negative ones are much higher than positive ones.

2. Hypothesis testing

a. One sample

```
t_test_onesample <- t.test(df$RestingBP, mu = 120, alternative = "two.sided")
print(t_test_onesample)
```

Claim is that mean blood pressure is different from 120 (population mean)
Parameter of interest: μ (mean resting blood pressure).
H_0: $\mu = 120$
H_1: $\mu \neq 120$
Level of significance $\alpha = 0.05$
$\bar{x} = 133.0228$
t_0 = 20.581
ν = 745
p – value = 0.00000000000000022
Since p – value is smaller than α, our decision is to reject H_0.
We can conclude that, at 5% level of significance, that the average resting blood pressure differs significantly from the population mean of 120mm Hg. (We have sufficient evidence to reject H_0.)

b. Two sample

```
df_1 <- df[df$HeartDisease == 1, ]
df_0 <- df[df$HeartDisease == 0, ]
t_test_twosamople <- t.test(df_1$Cholesterol, df_0$Cholesterol)
t_test_twosamople
```

Claim is that mean cholesterol level differs significantly between patients with and without heart disease.
Parameter of interest: μ_0 (mean cholesterol for patients without heart disease), μ_0 (mean cholesterol for patients with heart disease).
H_0: $\mu_0 = \mu_1; \mu_0 - \mu_1 = 0$
H_1: $\mu_0 \neq \mu_1; \mu_0 - \mu_1 \neq 0$
Level of significance $\alpha = 0.05$
$\bar{x}_0 = 238.7692$
$\bar{x}_1 = 251.0618$
t_0 = 2.833
ν = 712.54
p – value = 0.004742

Since p – value is smaller than α, our decision is to reject H_0.
We can conclude that, at 5% level of significance, that the average cholesterol for patients with heart disease differs significantly from the average cholesterol for patients without heart disease. (We have sufficient evidence to reject H_0.)

c. ANOVA

The analysis of variance (ANOVA) result tests for the significance of different factors and their interactions in explaining the variation in the response variable. To perform anova on different variables, I used formula object and a list containing all the interaction terms, which would print all 3 variations of anova when run once.

```
ivars2 <- c("ChestPainType * RestingECG * HeartDisease",
        "ChestPainType * RestingECG + FastingBS * HeartDisease",
        "ChestPainType * RestingECG + Sex * HeartDisease")

# Create an empty list to store aov objects
obj2 <- vector(mode = "list", length = length(ivars2))
count2 <- 1
# Loop through the input variables and create aov objects
for (i in ivars2) {
  form2 <- as.formula(paste("Age ~", i))
  obj2[[count2]] <- aov(form2, data = df)
  count2 <- count2 + 1
}
# Print the summaries
for (i in 1:length(ivars2)) {
  cat("\nSummary for ivars2[", ivars2[i], "]:\n")
  print(summary(obj2[[i]]))
}
```

Source of variation	Degrees of freedom	Sum of Squares	Mean Square	F test statistic	P - value
ChestPainType	3	4137	1378.9	17.885	0.0000000000332
RestingECG	2	2833	1416.3	18.369	0.0000000165501
HeartDisease	1	2692	2692.1	34.917	0.0000000053020
Interaction: ChestPainType:RestingECG	6	721	120.2	1.559	0.1565
Interaction: ChestPainType:HeartDisease	3	502	167.4	2.171	0.0901
Interaction:	2	384	191.8	2.487	0.0839

RestingECG:HeartDisease					
ChestPainType:RestingECG:HeartDisease	6	384	64.0	0.830	0.5467
Residuals	722	567	77.1		
Total	745	12220			

Tab 2: ANOVA Table

The F test statistic of 17.885 with an extremely low p-value of indicates that there is a significant difference in means among ChestPainType categories. The F test statistic of 18.369 with a low p-value suggests a significant effect of RestingECG on the response variable. The F test statistic of 34.917 with a low p-value indicates a highly significant effect of HeartDisease on the response variable. The interaction between ChestPainType and RestingECG is not statistically significant as the p-value is greater than α, the interaction between ChestPainType and HeartDisease is marginally significant as the p-value is slightly greater than α, the interaction between ChestPainType and HeartDisease is marginally significant as the p-value is slightly greater than α, the interaction between RestingECG and HeartDisease is marginally significant as the p-value is slightly greater than α. The interaction between ChestPainType, RestingECG HeartDisease is not significant as the p-value is greater than α.

d. ANOVA Variations

(i) Variation 1

Source of variation	**Degrees of freedom**	**Sum of Squares**	**Mean Square**	**F test statistic**	**P - value**
ChestPainType	3	4137	1378.9	18.322	0.000000000018
RestingECG	2	2833	1416.3	18.818	0.000000010733
FastingBS	1	2461	2461.3	32.704	0.000000015645
HeartDisease	1	2075	2075.3	27.576	0.000000198296
Interaction: ChestPainType:RestingECG	6	673	112.2	1.491	0.179
Interaction: FastingBS:HeartDisease	1	125	125.4	1.666	0.197
Residuals	731	55015	75.3		
Total	745	67319			

Tab 3: ANOVA Table with new interaction term

In the 1st variation of anova I have added a new interaction term between FastingBS and HeartDisease. The F test statistic of 18.322 with an extremely low p-value of indicates that

there is a significant difference in means among ChestPainType categories. The F test statistic of 18.818 with a low p-value suggests a significant effect of RestingECG on the response variable. The F test statistic of 32.704 with a low p-value suggests a significant effect of FastingBS on the response variable. The F test statistic of 27.576 with a low p-value indicates a highly significant effect of HeartDisease on the response variable. The interaction between ChestPainType and RestingECG and, FastingBS and HeartDisease is not statistically significant as the p-value is greater than α.

(ii) Variation 2

Source of variation	Degrees of freedom	Sum of Squares	Mean Square	F test statistic	P - value
ChestPainType	3	4137	1378.9	17.761	0.0000000000389
RestingECG	2	2833	1416.3	18.242	0.0000000185690
Sex	1	1	0.7	0.009	0.926
HeartDisease	1	2828	2827.6	36.420	0.0000000025270
Interaction: ChestPainType:RestingECG	6	749	124.9	1.609	0.142
Interaction: Sex:HeartDisease	1	21	20.6	0.265	0.607
Residuals	731	56752	77.6		
Total	745	67321			

Tab 4: ANOVA Table with new interaction term

In the 2nd variation of anova I have added a new interaction term between Sex and HeartDisease. The F test statistic of 17.761 with an extremely low p-value of indicates that there is a significant difference in means among ChestPainType categories. The F test statistic of 18.242 with a low p-value suggests a significant effect of RestingECG on the response variable. The F test statistic of 0.009 with a high p-value suggests that there is no significant effect of Sex on the response variable. The F test statistic of 36.420 with a low p-value indicates a highly significant effect of HeartDisease on the response variable. The interaction between ChestPainType and RestingECG and, Sex and HeartDisease is not statistically significant as the p-value is greater than α.

e (ii). Logistic Regression

The logistic regression model was applied to investigate the relationship between various predictor variables and the binary response variable, HeartDisease. The logistic regression model in this case, was trained on the dataset with predictor variables Age, RestingBP, Cholesterol, MaxHR, and Oldpeak to predict the occurrence of HeartDisease. The coefficients provide valuable insights into the influence of each predictor on the log-odds of HeartDisease.

```
logit_model <- glm(HeartDisease ~ Age + RestingBP + Cholesterol + MaxHR + Oldpeak, family = "binomial", data = df)
summary(logit_model)
```

```
Call:
glm(formula = HeartDisease ~ Age + RestingBP + Cholesterol +
    MaxHR + Oldpeak, family = "binomial", data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8296  -0.7811  -0.3895   0.7386   2.3613

Coefficients:
             Estimate Std. Error z value             Pr(>|z|)
(Intercept) -0.354200   1.101294  -0.322               0.7477
Age          0.023055   0.010772   2.140               0.0323 *
RestingBP    0.007433   0.005512   1.348               0.1775
Cholesterol  0.003127   0.001523   2.053               0.0401 *
MaxHR       -0.026013   0.004131  -6.297       0.000000000303 ***
Oldpeak      1.091319   0.107793  10.124 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1032.63  on 745  degrees of freedom
Residual deviance:  746.47  on 740  degrees of freedom
AIC: 758.47

Number of Fisher Scoring iterations: 4
```

Fig. 5: Logistic model summary

A positive coefficient for Age implies that, holding other variables constant, older individuals might exhibit a higher likelihood of HeartDisease. Similarly, the positive coefficient for RestingBP suggests that elevated blood pressure levels could contribute to an increased likelihood of HeartDisease. Similarly, the positive coefficient for Cholesterol suggests that elevated cholesterol levels could contribute to an increased log-odds of HeartDisease. Conversely, the negative coefficient for MaxHR indicates that higher resting heart rates may be associated with a reduced likelihood of HeartDisease. Additionally, a substantial positive coefficient for Oldpeak implies that a greater extent of ST depression induced by exercise is linked to an increased likelihood of HeartDisease.

The significant coefficients underscore the importance of Age, Cholesterol, MaxHR, and Oldpeak in predicting HeartDisease and these coefficients have p – value less than α. The overall model fit is assessed through the deviance values, with the decrease in deviance from the null model indicating that the inclusion of predictor variables improves the model's explanatory power. The insights gained from this logistic regression analysis help in understanding of the factors influencing the likelihood of HeartDisease, with potential applications in real-world scenarios for risk assessment and informed decision-making.

```
#1
new_data1 <- data.frame(Age = 25, RestingBP = 140, Cholesterol = 150, MaxHR = 200, Oldpeak = 1.5, FastingBS = 0)
predicted_prob1 <- predict.glm(logit_model, new_data1, type = "response")
predicted_prob1
```

```
        1
0.1377777
```

```
#2
new_data2 <- data.frame(Age = 80, RestingBP = 100, Cholesterol = 450, MaxHR = 120, Oldpeak = 2, FastingBS = 1)
predicted_prob2 <- predict.glm(logit_model, new_data2, type = "response")
predicted_prob2
```

```
        1
0.937124
```

In logistic regression, postestimation estimates involve predicting the probability of an event occurring based on the fitted model. After fitting the logistic regression model to the training data, we can use the model to make predictions on new or unseen data. The predict.glm function is employed to obtain these postestimation estimates. The first estimate gives the probability of the instance occurring as 0.137 approximately. The second estimate gives the probability of the instance occurring as 0.937. This shows that the 2nd instance is much more likely to occur than the 1st instance. This could be because of multiple factors for example, age 80 instance has a higher likelihood of occurrence as compared to 25.

Post - production and conclusion

In conclusion, the project's comprehensive analysis of cardiovascular health data has yielded valuable real-world insights with significant implications. The logistic regression model, considering factors such as age, resting blood pressure, cholesterol levels, maximum heart rate, and ST depression induced by exercise, provides a nuanced understanding of the predictors associated with the likelihood of heart disease. The model highlights the importance of specific variables, such as age and cholesterol levels, in influencing the probability of heart disease. These findings have practical applications in healthcare and risk assessment. For instance, the identification of age and cholesterol as significant contributors to heart disease risk underscores the importance of targeted interventions and monitoring for individuals in these demographic categories. We can see that higher the age or higher the cholesterol or higher the ST depression induced by exercise, the higher is the chance of the individual getting a positive diagnosis of heart disease. Moreover, the negative coefficient for maximum heart rate suggests a potential protective effect, informing strategies for promoting cardiovascular health. By uncovering these real-world insights, the project contributes to the development of evidence-based approaches for identifying and mitigating the risk of heart disease, ultimately fostering better-informed decision-making in clinical settings and public health initiatives.