TITLE

Analysis of Clinical Trial data – Identifying factors affecting angiographic disease status of patients.

TABLE OF CONTENTS

ABSTRACT

Introduction

This report delves into the analysis of the heart disease dataset obtained from the UCI Machine Learning Repository, featuring noninvasive test results from 120 patients undergoing angiography at the Cleveland Clinic in 1988. The primary objective of this study is to identify factors influencing the presence of heart disease in patients.

Method

Utilizing the R programming language, various appropriate methods are employed for data analysis. The results are visualized through data visualizations such as bar charts, boxplots and others, aiding in the interpretation of findings related to factors affecting development of heart disease. Furthermore, a logistic regression model is fitted on the data to determine the most significant features for heart disease.

Result

The result showed that number of major vessels, type of chest pain, ST depression and a disease called thalassemia are the main significant features affecting the development of heart disease.

Discussion

The insights derived from this analysis contribute to a better understanding of the influential factors in heart disease, providing valuable information for medical research and patient care.

INTRODUCTION

Heart disease, a collective term encompassing a range of cardiovascular conditions, stands as a leading cause of morbidity and mortality worldwide. Characterized by disorders affecting the heart and blood vessels, heart disease poses a formidable challenge to global public health (Lopez et al, 2023). The multifaceted nature of these conditions necessitates a thorough understanding of contributing factors, risk elements, and diagnostic markers to develop effective prevention and intervention strategies.

In the pursuit of unraveling the complexities surrounding heart disease, this introduction focuses on a specific dataset obtained from the UCI Machine Learning Repository. Originating from a study conducted at the Cleveland Clinic in 1988, this dataset comprises noninvasive test results from 120 patients undergoing angiography. The primary objective of this investigation is to identify factors influencing the 'num' attribute, a critical parameter in assessing the severity and nature of heart disease.

Through the lens of data analysis using the R programming language, this research endeavors to contribute to the broader understanding of heart disease. By examining patterns, correlations, and variables within the dataset, we aim to derive insights that may enhance diagnostic precision, treatment modalities, and overall cardiovascular health outcomes. As we embark on this exploration, the ultimate goal is to leverage data-driven methodologies to inform medical practices, improve patient care, and advance the collective efforts in mitigating the impact of heart disease on global health.

METHOD

A. **Setting up R environment**

The dataset was loaded into Rstudio and the required libraries were loaded to aid data manipulation and visualization. Some of the packages used in carrying out the analysis include tidyr and dplyr (for data manipulation), and ggplot2 for data visualization.

B. **Investigating the dataset**

The dataset structure was investigated using the str() function. There are 120 records with 14 features which include age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal and num. The description of the variables was provided in Appendix I. The 'num' is the target attribute for the analysis.

C. **Exploratory Data Analysis**

The effect of each feature on the final attribute "num" was explored using appropriate method and data visualization technique. Influence of categorical variables such as sex, cp, fbs, restecg, slope, thal, ca, and exang, on the variable 'num' were investigated using frequency bar chart and Chi-Square statistic was used to test the significance of the observed relationship as suggested by Hazra and Gogtay, (2016).

For scale variables such as age, chol, oldpeak, thalach, and trestbps, the effect of these variables on the 'num' attribute was evaluated using boxplot visualization and t-test was used to determine if the difference in the mean of the two groups of 'num' attribute is significant (Armitage and Berry, 1994).

D. **Regression Analysis**

A logistic regression model was performed on the data to determine the most significant features for predicting the final attribute 'num'. The accuracy of the model was determined. The p-value of the predicting variables were extracted and significant features (predicting variables with p-value less than 0.05) were identified. A bar chart was produced to visualize the significant variables.

RESULT AND DISCUSSION

Exploratory Data Analysis

1. Distribution of Age of the Patients

The summary of the data revealed that the minimum age of the patients used in the study is 29 while the maximum age is 76. The mean age of the patients is 54 years. The distribution of the patients' age was visualized using histogram and presented in figure 1.

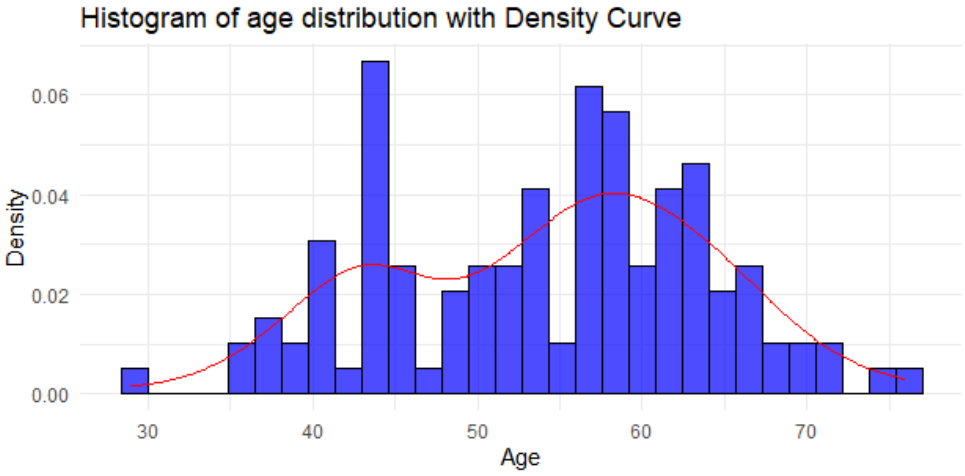


Fig. 1: Histogram of age distribution among patients.

The age distribution between the two groups of patients based on angiographic disease status ('num' attribute) was explored using a boxplot shown in figure 2. The plot revealed that the mean age (represented by thick black line inside the box) of patient positive for 'num' attribute is greater than the mean age of those without the disease.

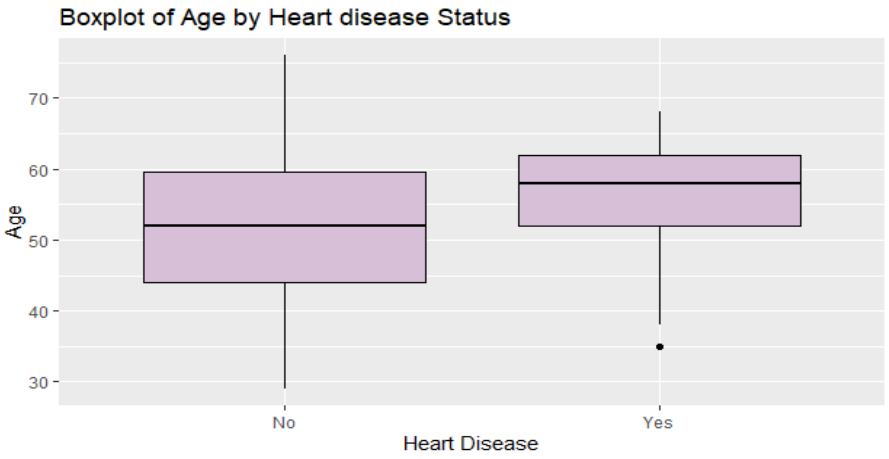


Fig. 2: Boxplot to compare the age distribution of the two categories of patients

2. Sex of the Patients

The sex of the patients as a factor affecting the heart disease status was explored using a double column bar chart. The result is presented in figure 3. It was observed that there are more females than are male patients. However, the proportion of female patients with positive status for the disease is higher than that of the male patients.

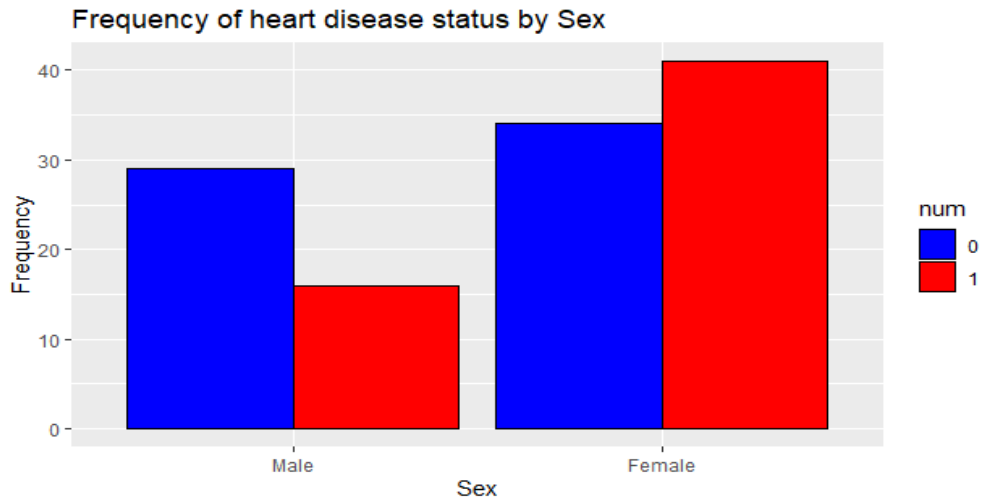


Fig. 3: Frequency of male and female patients based on the disease status.

3. Effects of Chest Pain on the heart disease 'num' attribute

The four categories of chest pain were considered a factor affecting the development of heart disease. This was investigated using a frequency bar chart as shown in figure 4. It was observed that majority of the patients, positive for the angiographic disease status have type 4 chest pain.

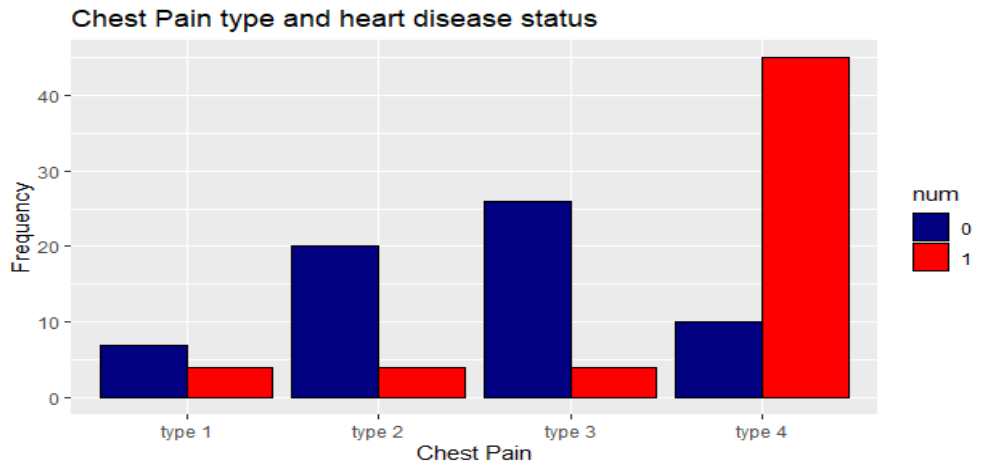


Fig. 4: A column bar chart showing the frequency of patients based on chest pain type.

4. Fasting blood sugar and ‘num’ attribute

The effect of fasting blood sugar on angiographic disease status of the patient was investigated using a frequency bar chart presented in figure 5. The visual revealed no difference in the proportion of positive and negative patients based on the status of their blood sugar, although there are more patients with fasting blood sugar less than 120mg/dl.

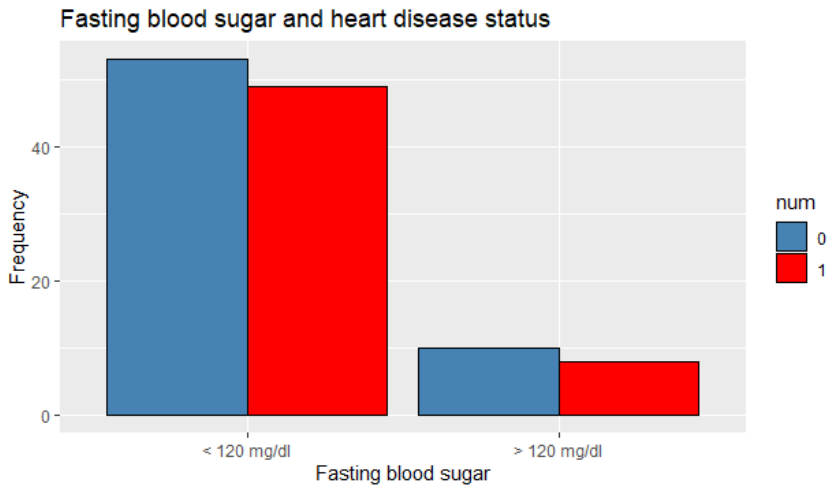


Fig. 5: A column bar chart showing the frequency of patients based on fasting blood sugar.

5. Resting ECG and ‘num’ attribute

As seen in the figure 6 below, higher proportion of patients with normal resting ECG are negative for ‘num’ attribute. Only few patients have ST-T wave abnormality while higher proportion of patients with left ventricular hypertrophy are positive for angiographic disease status.

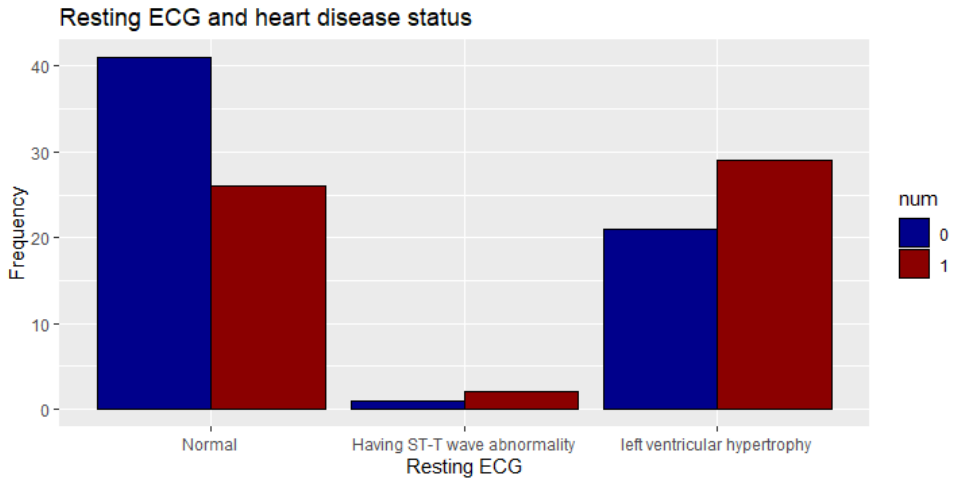


Fig. 6: A column bar chart showing the frequency of patients based on resting ECG.

6. Effects of other categorical variables on the 'num' attribute

Importantly, the other features such as slope, induced angina, number of major vessels and presence of disease called thalassemia were found to have significant influence on the angiographic disease status. It was revealed that majority of patients with flat slope are positive for the disease status (as shown in Figure 7). Also, figure 8 indicated that higher proportion of patients that are positive for exercise induced angina are positive for angiographic disease status.

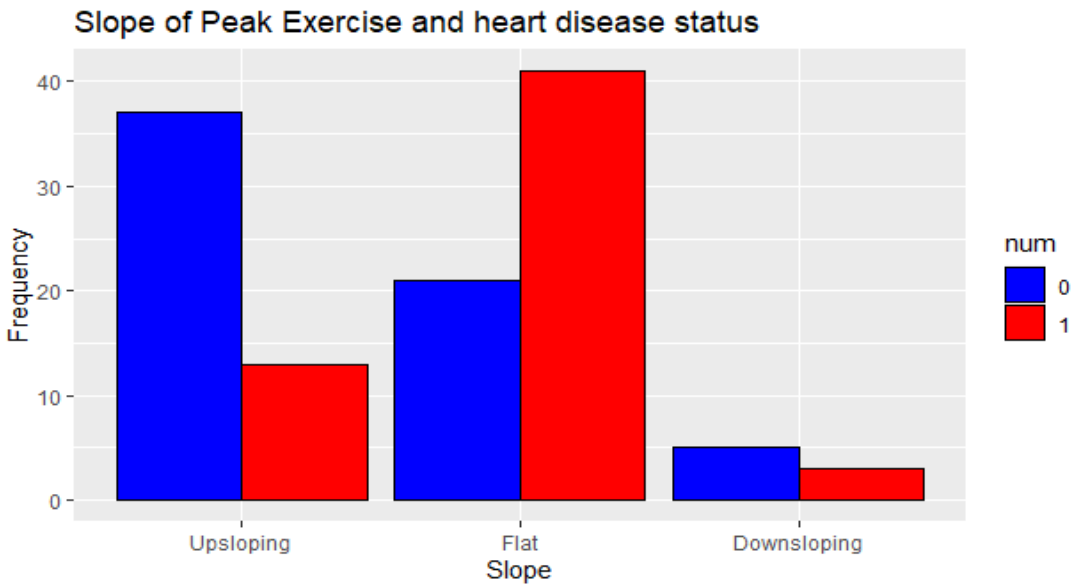


Fig. 7: A column bar chart showing the frequency of patients based on Slope of peak exercise.

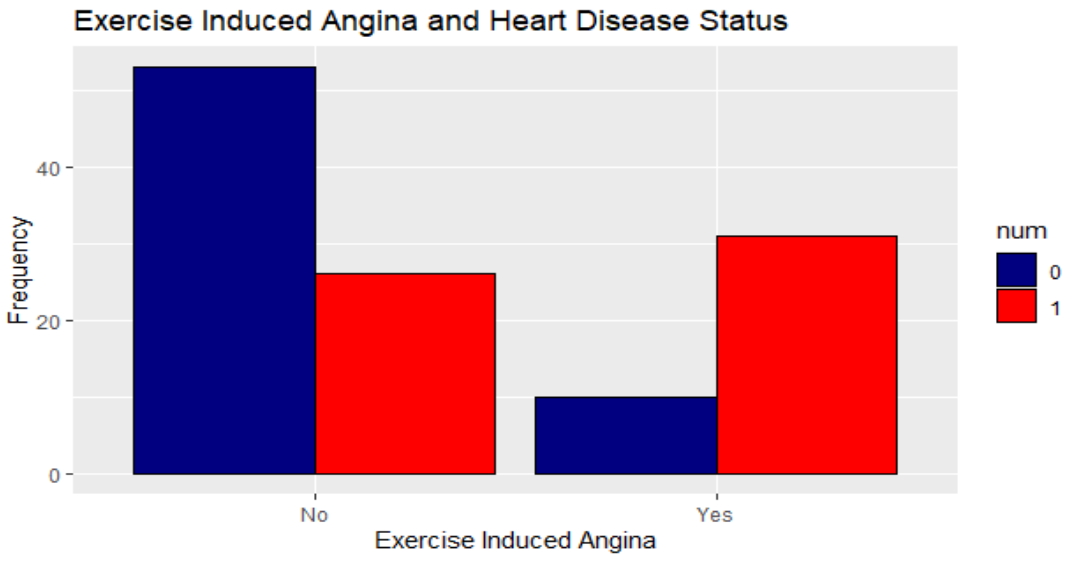


Fig. 8: A column bar chart showing the frequency of patients based on exercise induced angina.

Similarly, number of major vessels was observed to be an influencing factor for ‘num’ attribute. A shown in figure 9, majority of patients with no major vessels were observed to be negative for angiographic disease status while higher proportions of patients with one, two, or three major vessels are positive for the disease status.

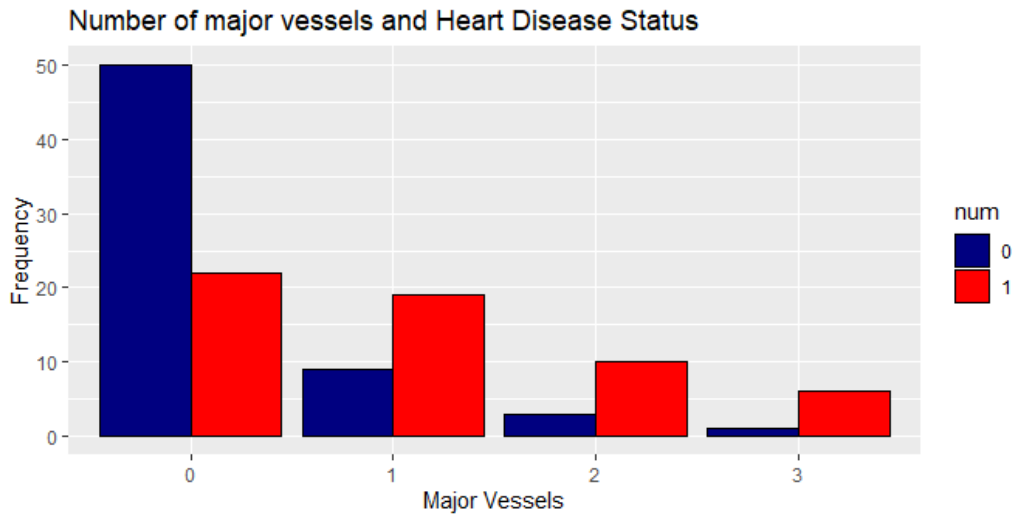


Fig. 9: A column bar chart showing the frequency of patients based on number of major vessels.

Thalassemia, a disease of the heart, was also observed to be an important factor affecting the angiographic disease status. As seen from figure 10, majority of patients that are normal for the thalassemia condition are negative for the ‘num’ attribute while higher proportion of others with thalassemia condition are also positive for angiographic disease status.

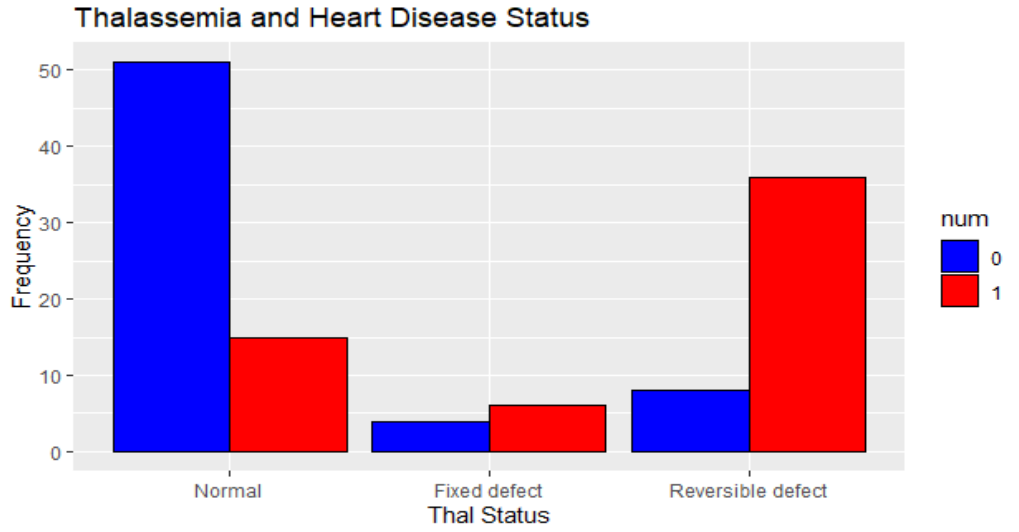


Fig. 10: A column bar chart showing the frequency of patients based on chest pain type.

Chi-Square Statistics

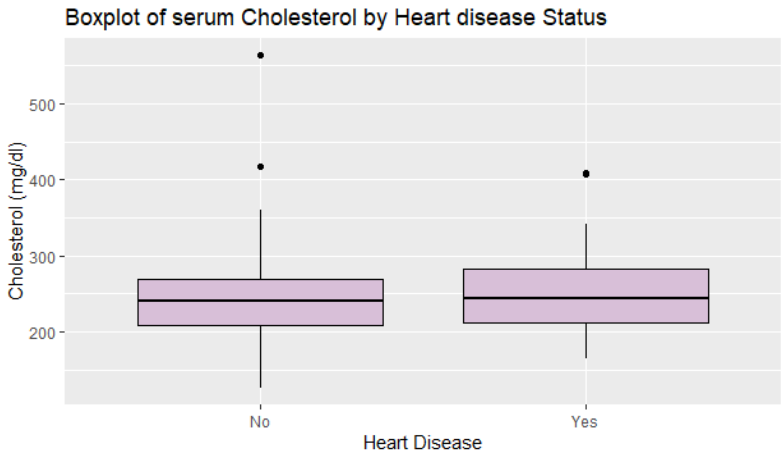
Chi-Square statistic was performed to validate the observed relationship between the categorical features and the 'num' attribute. The result of the test was presented in table 1. The null hypothesis is that there is no significant association between the features and the 'num' attribute. The result revealed that chest pain, slope, exercise induced angina, no of vessels and thalassemia are significant factors affecting the 'num' attribute. These features significantly influence the angiographic disease status.

Table 1. Results of Chi-Square statistics of the categorical features and 'num' attribute.

Features	X-squared	df	p-value	Conclusion on the null hypothesis
Sex	3.38845	1	0.065	**Retain**
Chest Pain	49.715	3	9.187e-11	**Rejected**
Fasting blood sugar	0.0006552	1	0.9796	**Retain**
Resting ECG	4.6833	2	0.09617	**Retain**
Slope	18.217	2	0.0001107	**Rejected**
Exercise Induced angina	18.058	1	2.143e-05	**Rejected**
No of major vessels	21.555	3	8.073e-05	**Rejected**
Thalassemia	37.649	2	6.679e-09	**Rejected**

7. Effects of Cholesterol, resting blood pressure, and oldpeak on the 'num' attribute

These features are continuous scale variables. Their relationships with the angiographic disease status were explored using boxplots and the hypotheses were tested using t-test. It was observed that oldpeak is the only feature that is shown to significantly affect the angiographic disease status. The boxplot along with the t-test results were shown in figure 11, 12, and 13 respectively. Oldpeak has a p-value less than 0.05 and hence indicated that there is significant difference in the mean oldpeak between patients that are positive for 'num' attribute and those that are negative, with positive patients having a higher mean value for oldpeak.



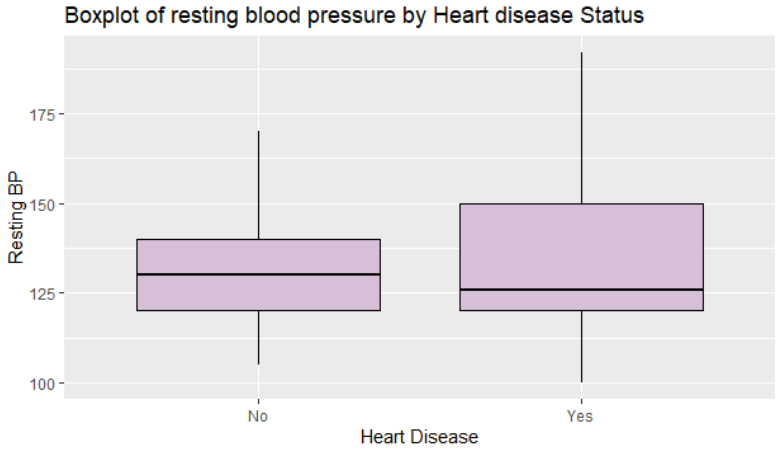
t = 0.02322

p-value = 0.9815

Mean of chol in group 0 = 249.84

Mean of chol in group 1 = 249.57

Fig. 11. A boxplot comparing the mean of cholesterol among the 'num' group.



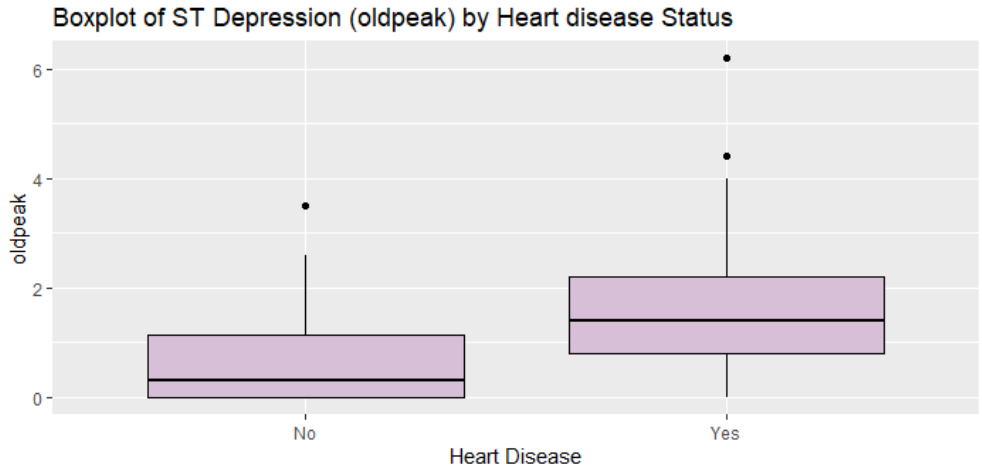
t = -1.5559

p-value = 0.1229

Mean of trestbps in group 0 = 128.19

Mean of trestbps in group 1 = 133.42

Fig. 12. A boxplot comparing the mean of resting bps among the 'num' group.



t = -4.7862

p-value = 6.465e-06

Mean of oldpeak in group 0 = 0.6619

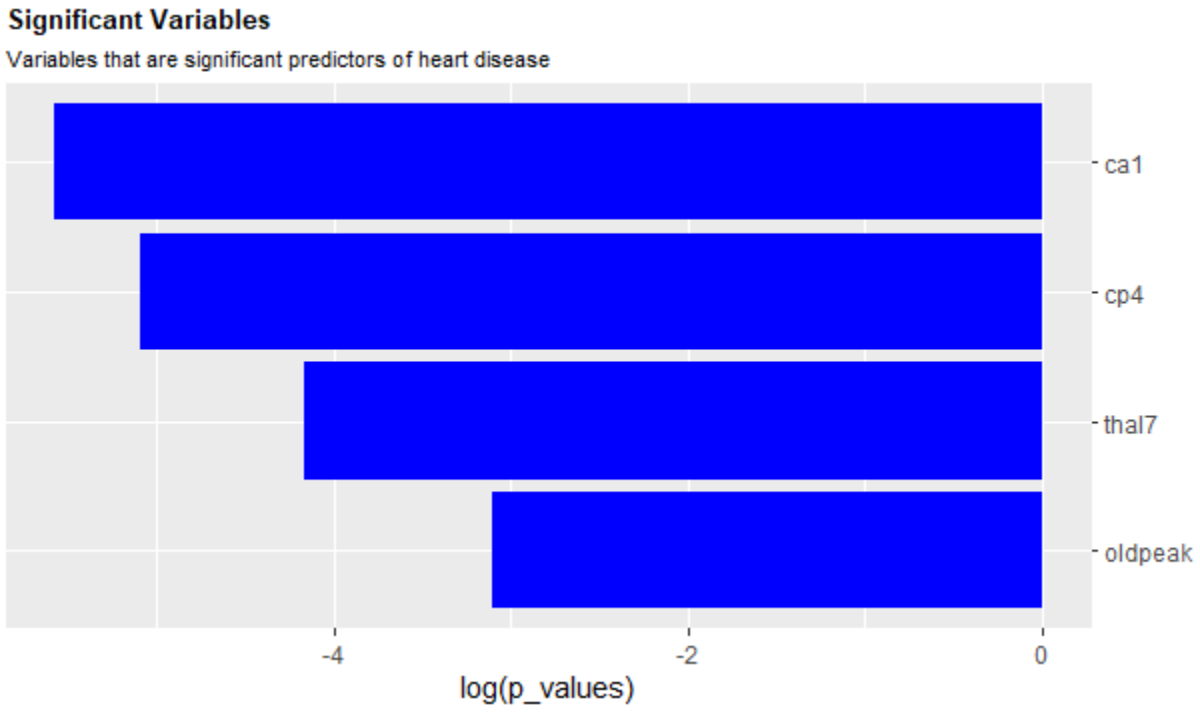
Mean of oldpeak in group 1 = 1.6228

Fig. 13. A boxplot comparing the mean of oldpeak among the 'num' group.

Logistic Regression Model

A logistic regression model was fitted on the data to identify the most significant features influencing the development of of angiographic disease status. The formula and output of the model fitting with the number of fisher iteration was provided in Appendix II. The model gave an accuracy of 81.1%. We filtered the features and extract only features with p-values less than 0.05 as these are the significant features that determine the final 'num' attribute.

The significant features were visualized using a bar chart as shown in figure 14. From the chart, the identified factors affecting the 'num' attribute are ca1 (having one major vessel), cp4 (chest pain type 4), thal7 (a revesible defect thalassemia) and oldpeak (ST depression induced by exercise relative to rest).



CONCLUSION

In conclusion, the analysis of the heart disease dataset revealed significant insights into factors influencing the 'num' attribute, representing angiographic disease status. Patient's age, sex, chest pain type, and various categorical variables were explored, demonstrating their impact on heart disease.

Chi-Square tests confirmed the significance of chest pain, slope, exercise-induced angina, the number of major vessels, and thalassemia. Cholesterol, resting blood pressure and oldpeak were also assessed, with oldpeak identified as a significant predictor. The logistic regression model achieved 81.1% accuracy, highlighting one major vessel, chest pain type 4, reversible defect thalassemia, and oldpeak as key influencers of the 'num' attribute.

These findings advance our understanding of heart disease, providing valuable insights for targeted interventions and personalized patient care, bridging statistical analyses and machine learning for a comprehensive approach.

REFERENCES

Armitage P, Berry G. Statistical Methods in Medical Research. 3rd ed. Oxford: Blackwell Scientific Publications, 1994:112-13.

Hazra, A., and Gogtay, N. (2016). Biostatistics Series Module 4: Comparing Groups - Categorical Variables. *Indian journal of dermatology*, *61*(4), 385–392. https://doi.org/10.4103/0019-5154.185700

Olvera Lopez E, Ballard BD, Jan A. Cardiovascular Disease. [Updated 2023 Aug 22]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2023 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK535419/

UC Irvine Machine Learning Repository - Heart Disease Dataset. Available at: https://archive.ics.uci.edu/dataset/45/heart+disease

APPENDIX

Appendix I – Description of dataset variables

- Age
- Sex: 0 – male; 1 - female
- cp: chest pain type (1 - 4)
- trestbps: resting blood pressure (in mm Hg on admission to the hospital)
- chol: serum cholestoral in mg/dl
- fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- restecg: resting electrocardiographic results

  – value 0: normal

  – value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation

  or depression of > 0.05 mV)

  – value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- thalach: maximum heart rate achieved
- exang: exercise induced angina (1 = yes; 0 = no)
- oldpeak = ST depression induced by exercise relative to rest
- slope: the slope of the peak exercise ST segment

  – value 1: upsloping

  – value 2: flat

  – value 3: downsloping
- ca: number of major vessels (0-3) coloured by flourosopy
- thal: 3 = normal; 6 = fixed defect; 7 = reversible defect
- num: diagnosis of heart disease (angiographic disease status)

  – value 0: < 50% diameter narrowing

  – value 1: > 50% diameter narrowing

  (in any major vessel)

Appendix II – Logistic Regression Model Output

```
Call:
glm(formula = num ~ ., family = binomial, data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.2254468  6.4210813  -0.970  0.33228
age         -0.0438243  0.0623651  -0.703  0.48224
sex1         0.4871684  0.9945542   0.490  0.62425
cp2          1.1004769  1.5847612   0.694  0.48742
cp3         -0.4692223  1.2562274  -0.374  0.70876
cp4          4.3294443  1.5796743   2.741  0.00613 **
trestbps     0.0344408  0.0205812   1.673  0.09425 .
chol         0.0004686  0.0062530   0.075  0.94026
fbs1        -1.2779356  1.1021128  -1.160  0.24624
restecg1     0.6707901  5.8531654   0.115  0.90876
restecg2     2.0299098  1.1112908   1.827  0.06776 .
thalach     -0.0243791  0.0235727  -1.034  0.30104
exang1       0.9541616  0.8100305   1.178  0.23882
oldpeak      1.3720391  0.6846677   2.004  0.04508 *
slope2       1.6491399  1.0200918   1.617  0.10595
slope3      -0.0950635  2.5108594  -0.038  0.96980
ca1          3.9867488  1.3774332   2.894  0.00380 **
ca2          0.5169122  1.5582203   0.332  0.74009
ca3          1.1368646  1.8823147   0.604  0.54586
thal6       -0.3475133  1.6842893  -0.206  0.83654
thal7        2.7624778  1.1407588   2.422  0.01545 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 166.055  on 119  degrees of freedom
Residual deviance:  56.279  on  99  degrees of freedom
AIC: 98.279

Number of Fisher Scoring iterations: 8
```