INFO411

Assignment 2 Report

|  |
| --- |
| An overview of the findings and actions throughout Assignment 2 |

**Submitted By**: Jacques Klavs, Daniel Pienaar

**Submission Date**: 07/10/2023

**Contributions**:

- **Exploratory data analysis**: Jacques and Daniel

- **Data imputation**: Jacques and Daniel

- **Modelling**: Jacques

- **Deliverables**: Jacques and Daniel

- **Dashboards**: Jacques and Daniel

- **Report**: Daniel

- **GitHub**: Jacques and Daniel

Contents

[1. Findings: EDA 2](#_Toc147660962)

[2. Findings: Data imputation 6](#_Toc147660963)

[3. Findings: Modelling 11](#_Toc147660964)

[4. Findings: Dashboards 11](#_Toc147660965)

[5. Conclusion 11](#_Toc147660966)

# Findings: EDA

## Understanding the data

Understanding the data, its structure, the element type and how it was first gathered and used is essential. Knowing which variables are categorical (ordered and not) or numerical is key to performing great analyses. A combination of reading the published paper and studying the UCI ML website (and its Variables Table) guided us in understanding what each variable meant, and how it should be used. We believe this is the correct way of doing things, because it informs future procedures.

## Statistical overview

Learning about what could be a determining factor for heart disease was eye-opening. Learning that some conditions may not contribute much to heart disease presence was just as eye-opening. Comparing plots and studying the correlogram helped discern what should and shouldn't be used. The plots and correlogram are shown following this paragraph.

The figure below shows frequencies of males and females according to various measurements:

* Chest Pain Type: Most participants reported chest pain measurements of 4, which is asymptomatic.
* Exercise Induced Angina: Most participants did not experience exercise-induced angina.
* Fasting Blood Sugar: Most participants did not have high fasting blood sugar, above 120 mg/dl.
* Resting Electrocardiographic Results: There was an almost even split between participants who had normal results (0) and possible or definite left ventricular hypertrophy by Estes' criteria (2).

The pattern between males and females remains very similar across all plots.

A screenshot of a graph

Description automatically generated

Figure 1: Male and female statistics from the statlog-heart.data file

The figure that follows dives into the presence of heart disease according to several features. The following can be gathered from the figures as an indication of the presence of heart disease:

* Higher resting blood pressure with lower maximum heart rate achievement.
* Higher resting blood pressure with higher serum cholesterol.
* Lower age and lower maximum heart rate achieved.

There is not a clear indication of how serum cholesterol at different ages affects your chances of getting heart disease. It appears that the factors out of the ones below indicate that resting blood pressure and maximum heart rate achieved during exercise could be best used for determining the presence of heart disease.

A collage of different colored dots

Description automatically generated

Figure 2: Heart disease presence according to various parameters in the statlog-heart.data file

## Key findings

Seeing how much Exercise Induced Depression correlates to the presence of heart disease was frightening. It is a contrast that is more significant than most relationships. This emphasises the need to understand mental health, and how the brain acts and reacts physiologically if we are to prevent heart disease in the future.

The correlogram that follows investigates the correlation each variable has with each other. The following is a summary on what can be discerned.

Low correlation:

* Resting Blood Pressure and Maximum Heart Rate.
* Maximum Heart Rate and Serum Cholesterol.
* Serum Cholesterol and Exercise Induced Depression.

Highest correlation:

* Resting Blood Pressure and Age.
* Exercise Induced Depression and Resting Blood Pressure.
* Serum Cholesterol and Age.
* Exercise Induced Depression and Age.
* Maximum Heart Rate and Age.
* Exercise Induced Depression and Maximum Heart Rate.

What can be inferred from this is that, if we understand that Exercise Induced Depression might be a strong determining factor for the presence of heart disease, the following variables should be avoided in favour of using Depression as a variable for a predictive model:

* Maximum Heart Rate.
* Age.
* Resting Blood Pressure.

The best other determining factor that should be considered is Serum Cholesterol.

A screenshot of a graph

Description automatically generated

Figure 3: A correlogram of select features

The figures below are density plots of age versus resting blood pressure and depression induced by exercise relative to rest. The density plots are separated between those with and without heart disease.

In both instances, age does not seem to influence the presence of heart disease much. On the other hand:

* Resting blood pressure looks to be somewhat positively correlated with the presence of heart disease.
* High measures of depression induced by exercise relative to rest look to be very strongly correlated with the presence of heart disease.

A screenshot of a graph

Description automatically generated

Figure 4: Heart disease vs age, resting blood pressure and depression

The plots below illustrate the frequency of males and females who have and do not have heart disease according to age. The following can be drawn from these figures in the scope of the test study group:

* There are more males who do not have heart disease than those who do not.
* There are more females who have heart disease than those who do not.
* Most of the males and females who do not have heart disease are around 60 years of age.
* The population of females who have heart disease are evenly spread across the ages.

What could be inferred from this is that there are more active 60-year-old females than at any other age, in this study group.

A screenshot of a graph

Description automatically generated

Figure 5: Disease-free populations in males and females

# Findings: Data imputation

## The DS2 dataset

We chose to use the Hungarian, Swiss and VA datasets. Much of the data was missing. Julia does make it very easy to rectify this, however. After some cleaning that was performed in the same way as in the EDA section, a simple application of the Julia Impute package was performed. This filled in all the missing data. EDA could then immediately be performed.

As a reference, a heatmap was generated to show the missing data before and after imputation. Blank spaces in the heatmaps correspond to missing values. This is a visual guide to be used for targeted data cleaning. In all cases, variable 12 is significantly unpopulated, with variable 13 and 11 coming in second.

These variables are:

* 11: Slope of the Peak Exercise ST Segment.
* 12: Number of Major Vessels colored by fluoroscopy.
* 13: Thal: 3 = Normal; 6 = Fixed Defect; 7 = Reversable Defect.

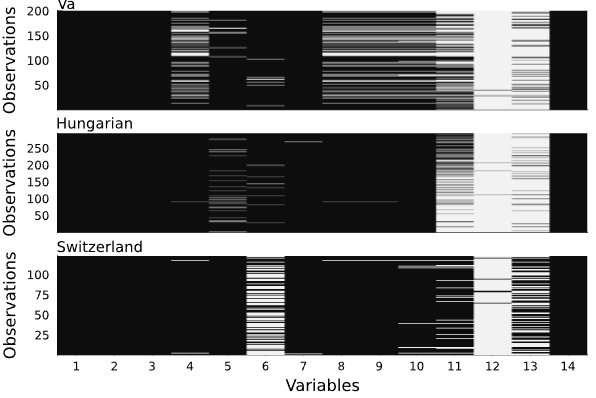


Figure 6: DS2 data sets before imputation

A black rectangular object with white text

Description automatically generated

Figure 7: Data sets after imputation

## Results of the EDA on DS2

Statistical profiles, not surprisingly, varied significantly across the datasets. There were more people in the Hungarian dataset, but older people in the VA dataset. The Hungarian dataset appeared to be filled with healthier people - possibly due to age - but also had a great portion of people experience rank 2 chest pain, the second highest in rank. The Swiss and VA test group participants followed a similar descending profile to that of DS1 from rank 4 to 1.

Seeing how people of different ages in different locations fit within the scope of the presence of heart disease according to blood pressure was interesting: high concentrations of people 10 years apart do not have heart disease in Hungary, but low concentrations do in this same location spread out more evenly across the age groups. There appeared to be some sort of correlation between blood pressure and heart disease routinely.

The figure below shows that the Hungarian population represents the youngest age profile, VA the oldest, and the Swiss in the middle. The VA profile appeared to have far fewer participants, with a very small number of females which are concentrated in the 50–65-year range. The Hungarian profile seems to have a more even distribution for males and females compared to the other two datasets. The females in all cases have smaller age ranges than the males and appear to be smaller in number.

A diagram of different shapes

Description automatically generated

Figure 8: DS2 data set age and sex profiles across locations

For the second figure showing a histogram of Chest Pain Type, a ranking of 4 was most common across all datasets. The frequencies decrease as the ranking moves from 4 to 1, except in the case of Hungary, where rank 2 has the second highest number of reports.

A graph of different types of pain

Description automatically generated

Figure 9: DS2 data set chest pain type across locations

## Key findings

The following figures will explore more measures, how they contribute to the presence of heart disease, and compare. This will be on a basis of Age and Resting Blood pressure, across the three locations. For the first location, we have VA. Here, we can see that older people with higher resting blood pressure more than not have heart disease.



Figure 10: Heart disease in VA

For the second location, we have Switzerland. Here, we can see an interesting relationship. The density of people without heart disease is concentrated, but those without are not.

A graph with blue and orange lines

Description automatically generated

Figure 11: Heart disease in Switzerland

For the third location, we have Hungary. Here, there is no clear relationship between age, resting blood pressure and presence of heart disease.

A graph with blue and orange lines

Description automatically generated

Figure 12: Heart disease in Hungary

# Findings: Modelling

# Findings: Dashboards

# Conclusion