EDA Results

Heart Failure Prediction Dataset

Sander J. Bouwman

10/5/2021

Contents

1 Results

1.1 Cleaning

Datacleaning was done on multiple variables. Column ChestPainType originally categorical data was split out into 4 new columns containing logical data. This was also performed for the ST_Peak column which was split into 3 new colums containing logical data. There were no missing datafields.

1.2 Heart disease (HD)

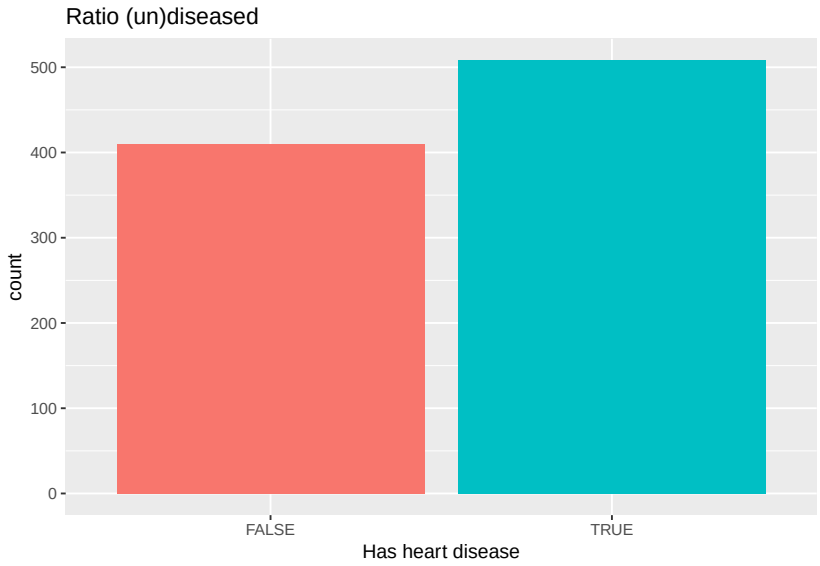

Figure 1: Count of heart disease

The ratio patients having HD is 55% (n=508) 1 while 45% (n=410) has no HD, altough there is a slight balance but is still acceptable.

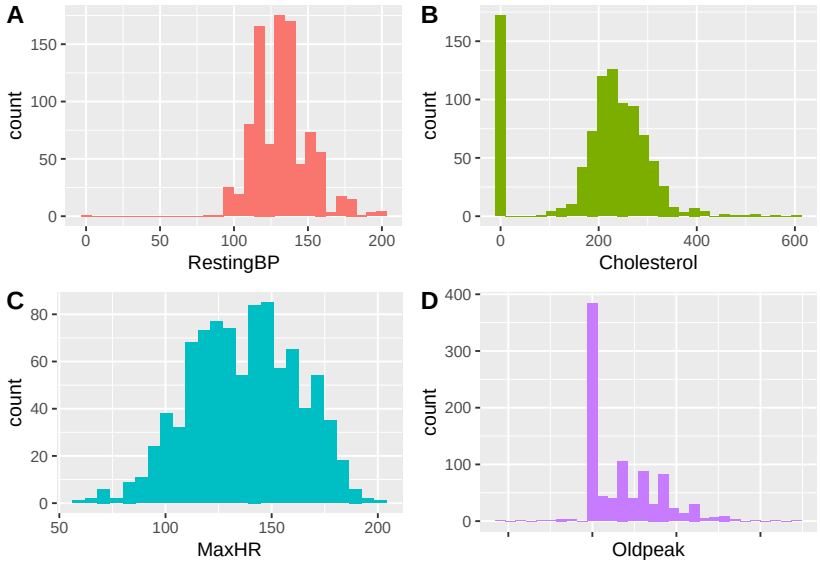

Figure 2: Four histograms of restingbp, cholesterol, maxhr and oldpeak

Figure 2 shows that data is quite normally distributed. Altough there are some values with very high counts. Notable cholesterol (B) with a very high count of 0 mm/dl (n = 172) and oldpeak (D) with a very high count of 0 ST.

1.3 Excercise angina

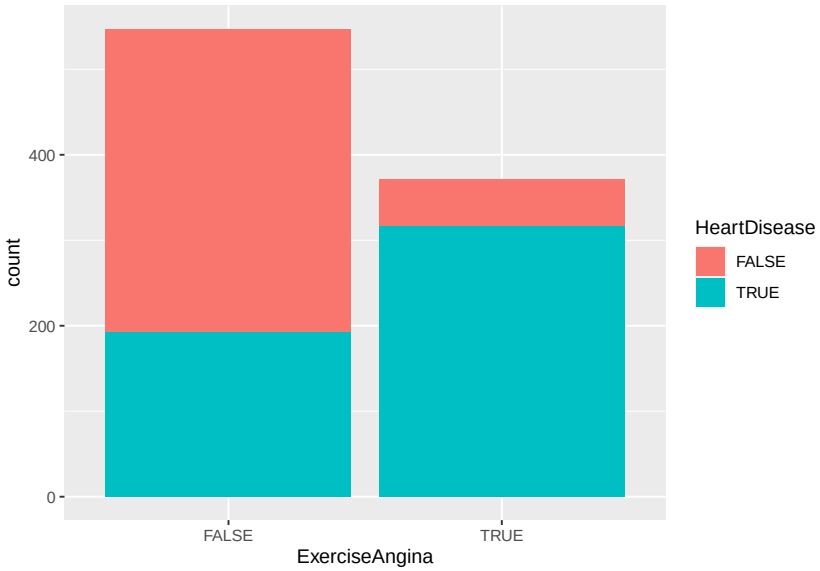

Figure 3: Effect of exercise angina on HD

In figure 3 it is visible that having excersise angina has a much higher chance of having HD in comparison to not having excercise angina.

1.4 ST-Slope

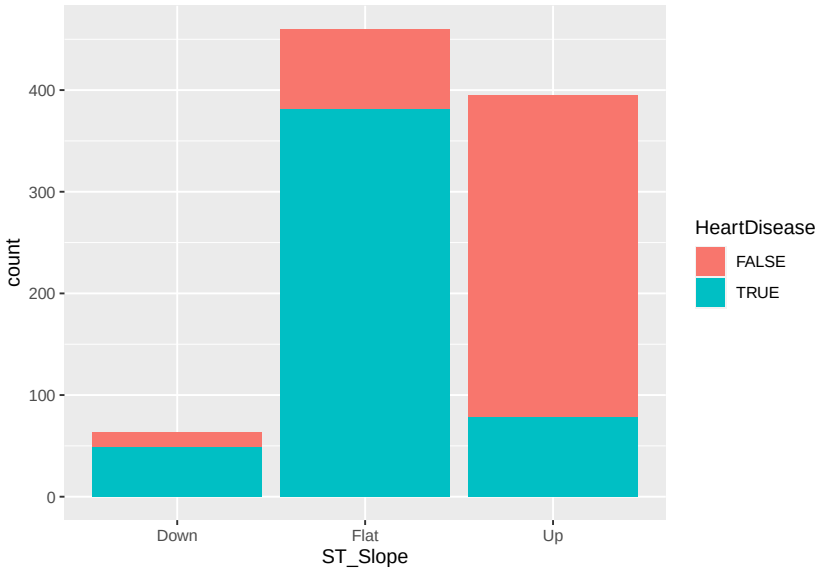

Figure 4: Relationship between heart slope and heart disease

Figure 4 shows that having an UP ST slope depression significantly lowers the chance of HD, when comparing this with a down or flat slope. Both down and flat ST slopes give a much heigher chance of having HD. ST slope down is fairly rare in this dataset.

1.5 Age

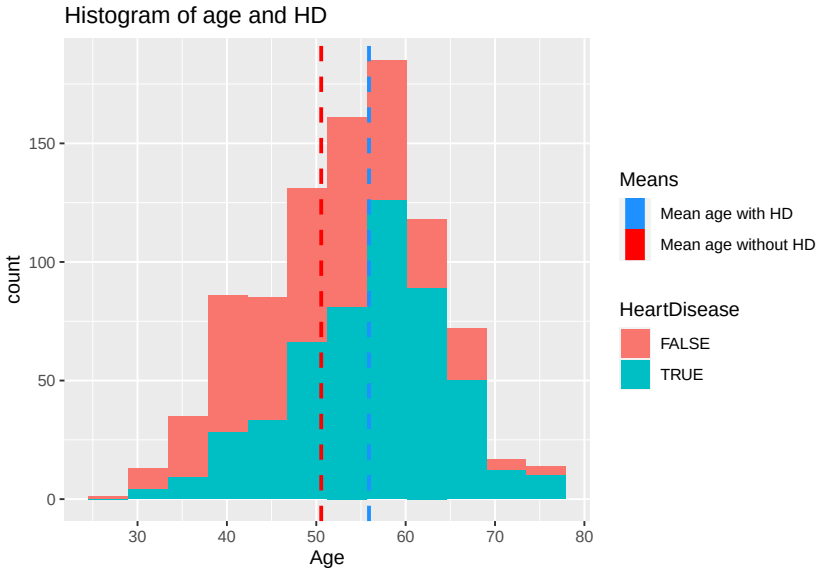

Figure 5: Histogram of age grouped by HD diagnoses

In the figure above (figure 5) it is visible that the percentage of patients with HD shifts to the majority with age. After the age of around 70, this percentage seems to shrink again. Using the dotted vertical lines we can see that there is a difference in mean age in persons with or without HD. The mean age of persons with HD is ~56 years whereas the mean age of persons without HD is ~51 years. The histogram is itself shows a descent normal distribution in age in the dataset. To further explore the possible relation/correlation between age and the prevelance of HD another plot is used.

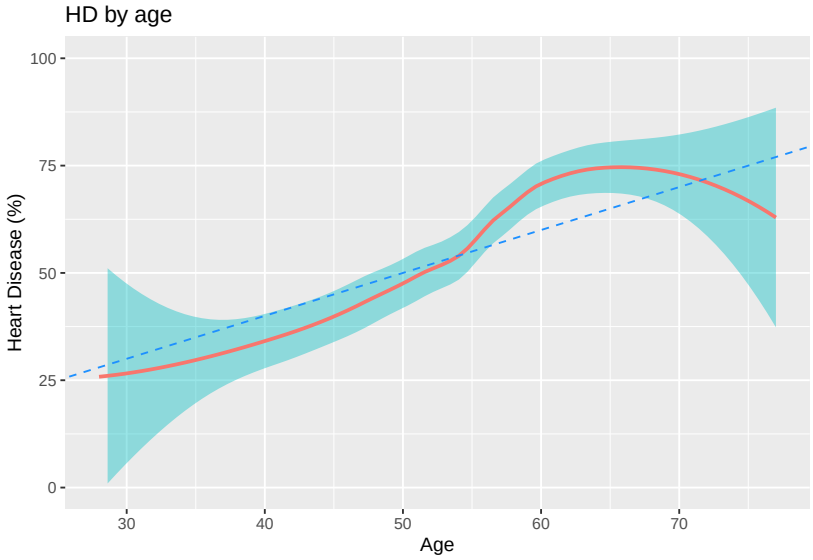

Figure 6: Percentage of patients affected by HD by age

Figure 6 shows the relation between age and prevalance of HD. Both ends show a major wide area of heart disease (percentage) while the overal trend shows a clear increase of HD with an increase of age. The percentage of patients with a positive HD diagnosis at age 30 is around 25% wheras patients age 65 have a 75% positive diagnosis rate. As visible both figure 1 and in figure 2 it is visible that the percentage of positive diagnosis lowers after age 65.

1.6 Sex

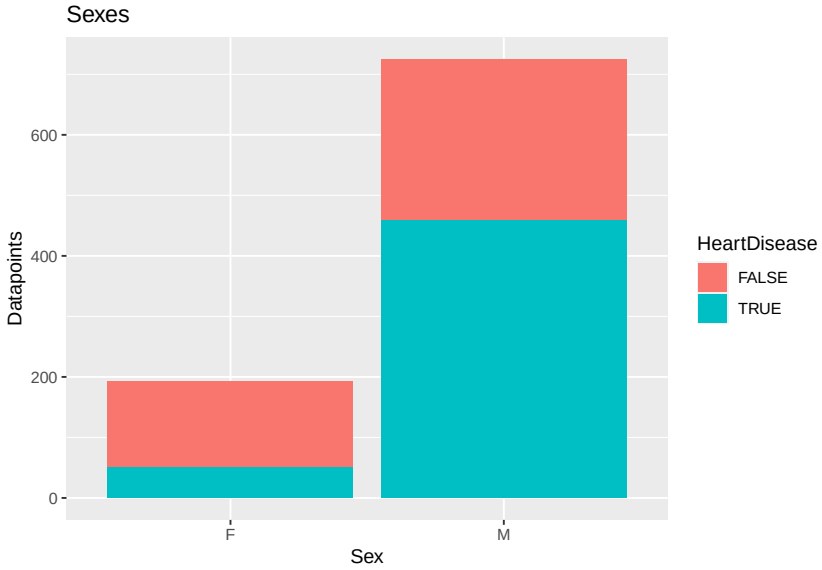

Figure 7: Amount of instances by sex grouped by HD diagnosis

As visible in the above figure (figure 7) the ratio of female(21.02%) to male(78.98%) is very skewed towards males. The ratio of a positive diagnosis is obviously different. Males tend to have on average a positive HD diagnosis whereas females have a negative diagnosis.

To further visiualize if the density of age with sex is roughtly the same the underlying figure was generated.

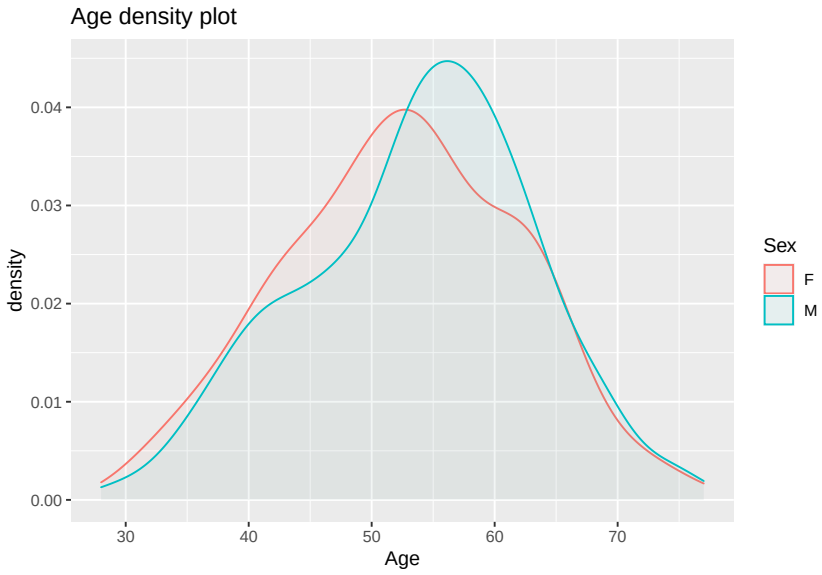

Figure 8: Density plot of ages grouped by sex

Figure 8 shows that both sexes have an about even amount of age groups. Where males have a higher representation in the ages 55+ years, and females have a larger (relative for females) representation lower than 55 years.

1.7 Pain types

To bring an easier understanding of different chestpain types (CPT), the underlying table can be used. Pain type:

Code	Type	Explanation
TA	Typical Angina	Meets all three of the following characteristics: 1. Substernal chest discomfort of characteristic quality and duration 2. Provoked by extortion or emotional stress 3. Relieved by rest and/or nitrates within minutes.
ATA	Atypical Angina	Meets two of the above characteristics
NAP	Non-Anginal Pain	Lacks or meets only one or none of the characteristics
ASY	Asymptomatic Pain	

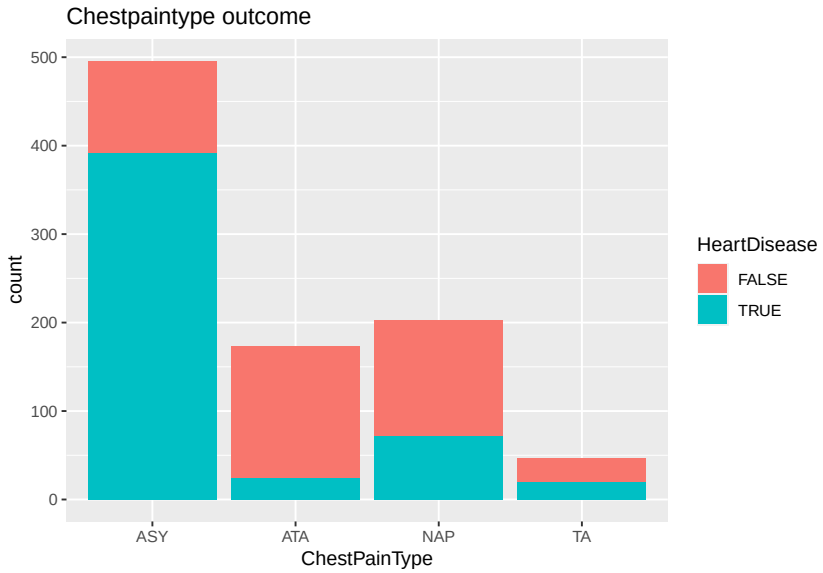

Figure 9: Ratio of diagnosis grouped by chestpain type

Note: For the creation of the PCA plots the categorical variable ChestPainType was split into multiple numerical values containting columns.

As visible see in the above figure (figure 9) the uttermost CPT with HD is asymptomic pain. After that Non-Angical Pain is the heighest altough more patients with NAP have no diagnosed HD, the same can be seen in TA where the ratio between having and not having HD is rougly the same as in NAP. ATA has by far the lowest ratio diagnosed HD.

1.8 PCA plotting

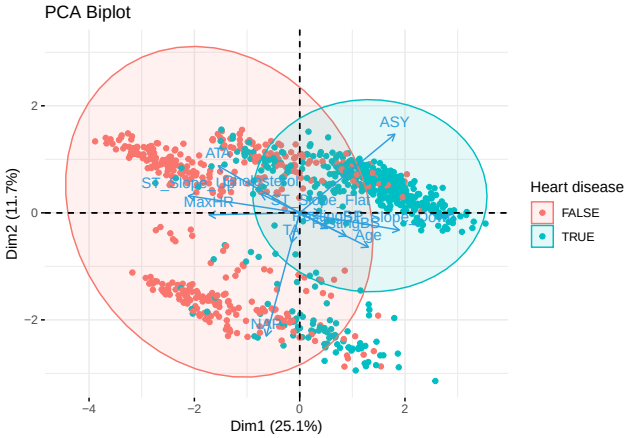

Figure 10: PCA Biplot of all non-categoral variables

In the above figure (figure 10) two clusters are visible. The postive (turqoise) cluster seems to be in correlation with ASY (asymptomic pain). Whereas there are two distinct negative clusters. The left upper cluster is affected by ATA, Max heartrate and cholesterol, while the left down clusters is effected by NAP.

To further see the contribution of each variable the underlying figure was generated.

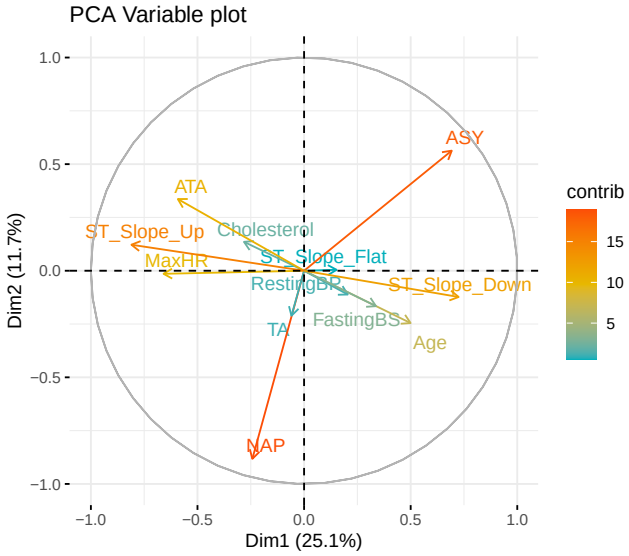

Figure 11: PCA Variable plot of all non-categoral variables

Figure 11 shows that both ASY and NAP are the biggest contributors between a positive or negative diagnoses. As previously mentioned, it is also visible here that an ST_Slope up lowers the chance of HD while ST_Slope down has a positive correlation with HD.

2 Discussion

The undiseased to diseased ratio is not fully balanced (0.45 to 0.55 ratio), but this is not critical. There seems to be quite a lot of patients where cholesterol is not measured and defaulted to 0 mm/dl (n=172) (see figure 2). Currently, asymptomatic pain is the major point of differentiation. Asymptomatic pain is not something usable in itself, as it is not measurable by patient nor caregiver. If asymptomatic pain is removed, it might still be possible to create a reliable prediction of HD if an appropriate algorithm is used. But using asymptomatic pain in combination with other variables is might still be used to increase prediction accuracy. Furthermore the data is highly skewed towards males. Of these males a very large percentage (63%) have HD while females have a far lower ratio of heartdisease (26%). In the real-world males have a slighty higher chance of heartdisease but as big as this dataset implies. For further research, it would benefit if more female data would be added to the dataset with a better HD ratio then the current situation. Looking at figure 10, it seems that cholesterol has a positive effect on HD, a higher cholesterol generally increases the chance of heartdisease, which seems counterintuitive with this dataset. The sudden lowering of patients with HD above the age of 65 might be explained by the fact that patients older then 65 simply die due their HD.

The current dataset seems to be cleaned well enough, depending on the used algorithm some changes might be necessary, factorizing certain variables such as age, cholesterol, maxhr and oldpeak might be needed. During EDA it was found that certain variables such as cholesterol had a heigh count of the same values, for example cholesterol had a very large count of 0 mg/dl, the dataset source doesn't provide a reason, but it is suspected that this is due to no measurements available and defaulting to 0, the same could be said for the variable oldpeak where there are also a lot of instances with the value 0.

3 Conclusion

Age seems a decent indicator in the diagnosis of heart disease. This is confirmed by the PCA plots (figures 10 & 11) where it shows that it is a small but meaningful indicator. Asymptomatic pain seems to be the major indicator in the diagnosis of HD As visible in figure 9, the precense of asymptomic pain is a large indicator of a positive HD diagnosis as 79% percent of asymptomic pain patients have diagnosed HD The PCA variable plot (figure 11) also confirms this as NAP has a large contribution. NAP has a high contribution to a negative diagnosis according to figure 11. Figure 9 confirms this as a large amount of patients experiencing NAP (64.5%) has a negative HD diagnosis. ATA has a low (13.9%) positivity rate of HD, this also shows in the PCA plot (figure 11) as ATA has relative smaller overall contribution. Having an UP ST slope depression lessens the chance of HD significantly. It also seems that non-numerical variables have a higher contribution than numerical variables.

As visible by the data it is probable that it is possible to predict heart disease with fair accuracy on this dataset, although it is not probable that it is accurate on real-world data as there are significant differences between the used dataset and real world data.