Project Report One Group 3

2023-09-12

Contents

Group Three Members Information:

Grace:

ORCID ID: 0009-0007-5950-1669
STUDENT ID: 300579109
EMAIL: *brownlgrac@vuw.ac.nz*

Hannah:

ORCID ID: 0009-0003-8155-9657
STUDENT ID: 300343315
EMAIL: *Colliehann2@vuw.ac.nz*

Thomas:

ORCID ID: 0009-0007-5097-9017
STUDENT ID: 300475577
EMAIL: *rowleythom1@vuw.ac.nz*

Aiden:

ORCID ID: 0009-0006-7202-082X
STUDENT ID: 300561276
EMAIL: *angaida@vuw.ac.nz*

Bridget:

ORCID ID: 0009-0004-8935-8646
STUDENT ID: 300572906
EMAIL: *fijmabrid@myvuw.ac.nz*

Reason for Choice and Availability:

Heart disease is the number one killer in New Zealand [cite heart foundation here]. Consequently, this is an important area of research, as understanding the clinical variables that make people vulnerable to heart disease is crucial to implementing effective prevention strategies and informing effective policy decisions.
We chose to look at specifically at how clinical variables can predict the presence or absence of heart disease.

Our data is freely available through Kaggle, and we are able to share and adapt the data to suit our needs. We accessed the data from https://www.kaggle.com/datasets/thedevastator/predicting-heart-disease-risk-using-clinical-var to the dataset. We will also ensure that we give appropriate credit to our data, and provide a link to the liscence Data files © Original Authors. Our own findings will also be freely available through GitHub in keeping with the policy of our original data.

Number and Type of Variables:

The sample population of the dataset is 270 adults from Cleveland, USA.

The "Heart_Disease_Prediction" dataset has 14 variables.

6 of those variables are numeric and 8 are categorical.

Numeric Variables:

Age: The age of the patient
BP: The blood pressure level of the patient in mmHg
Cholesterol: The cholesterol level of the patient in mg/dl
Max HR: The maximum heart rate levels achieved during exercise testing (bpm)
ST depression: The ST(Stress Test) depression on an Electrocardiogram induced by exercise relative to rest (mm)
A clinically significant ST depression is typically defined as ≥ 1 mm. A significant ST depression, particularly for a patient with chest pain, may be indicative of restriction of blood supply (myocardial ischemia)
Number of vessels fluro: The number of major vessels coloured by fluoroscopy
The major coronary vessels considered here are the left anterior descending artery, the left circumflex artery, and the right coronary artery. The intent of the fluoroscopy is to visualise all three major vessels.

Categorical:

Sex: Sex of the patient(0 = Female, 1 = Male)
Chest pain type: Type of chest pain experienced by the patient (1 = Typical angina, 2 = Atypical angina, 3 = Non-anginal pain, 4 = Asymptomatic)
Typical angina occurs behind the sternum and is often described as a squeezing or tightness in the chest. It is triggered by exertion or stress and is relieved by rest within 20 minutes. Typical angina is highly suggestive of heart disease.
Atypical angina shares some of the characteristics of typical angina but is either not triggered by exertion or stress or is not relieved by rest.
If "asymptomatic" the patient is not experiencing chest pain.
FBS over 120: Fasting blood sugar test results are over 120 mg/dl (0 = False, 1 = True)
EKG results: Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 =

Showing Probable or Definite Left Ventricular Hypertrophy by Estes' Criteria)
ST-T wave abnormalities can include T-wave inversions and or ST-segment elevation or depression of greater than 0.05 millivolts(mV). These can both be indicative of partial or complete blockage of a coronary artery (mysocardial ischemia).

Research Questions:

One:

Can we accurately predict whether heart disease is present based on the variables measured.

Two:

How does sex and age interact and influence whether heart disease is present.

Three:

What manageable characteristics effect rates of heart disease and can people reduce/increase these.

Preliminary EDA:

Histograms, notched box plots, and Q-Q plots will be generated from the heart disease data, considering the numerical variables `Age`, `BP`, `Cholesterol`, `Max.HR`, and `ST.depression`. These plots will be found in Figures 1, 2, and 3 respectively.

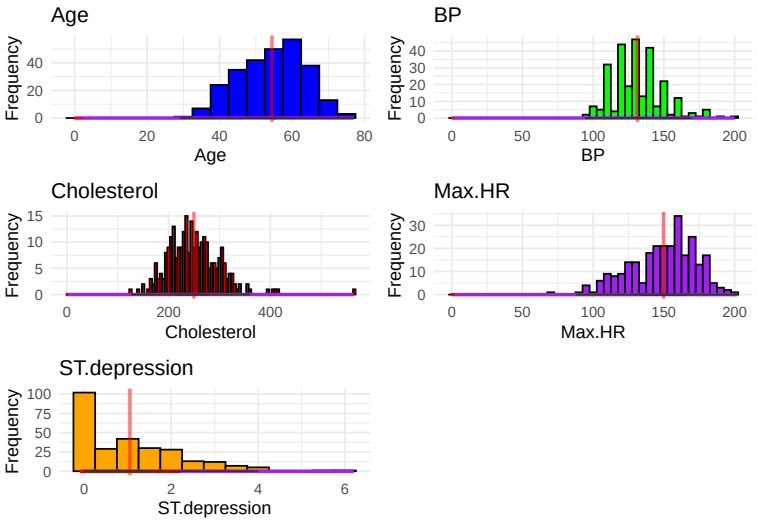

Figure 1: Histograms for the numerical variables of the heart disease dataset.

The age, cholesterol, and maximum heart rate histograms appear to follow a normal distribution, as seen in Figure 1.

The blood pressure histogram has many peaks at the 10 mmHg marks. This is a result of "zero end-digit preference", a phenomenon where blood pressure readings are rounded to the nearest ten. This is known to occur especially around treatment cutoffs (120 mmHg is considered 'elevated', 130mmHg is 'stage 1 hypertension', and 140mmHg is 'stage 2 hypertension') (https://www.cdc.gov/bloodpressure/facts.htm). 150 of the 270 observations, approximately 56%, are given as a multiple of 10. Most blood pressure indicators give

guidelines to measure to the nearest 2 mmHg. The potential rounding of data indicates a lack of normality in the data since many observations may have been rounded up or down significantly.

The ST depression histogram has a tail to the right, which is expected. ST depression is a measure of how far an ECG line passes below the baseline on the graph, and these values tend to be small. ST depression values indicate potential issues with the heart, like restricted blood flow to the heart. Greater values indicate more significantly that issues may be present, but the magnitude of the value alone does not indicate the type of issue; that is the role of the `Slope` variable, which categorises the direction of the slope in order to rule out specific causes for the depression.

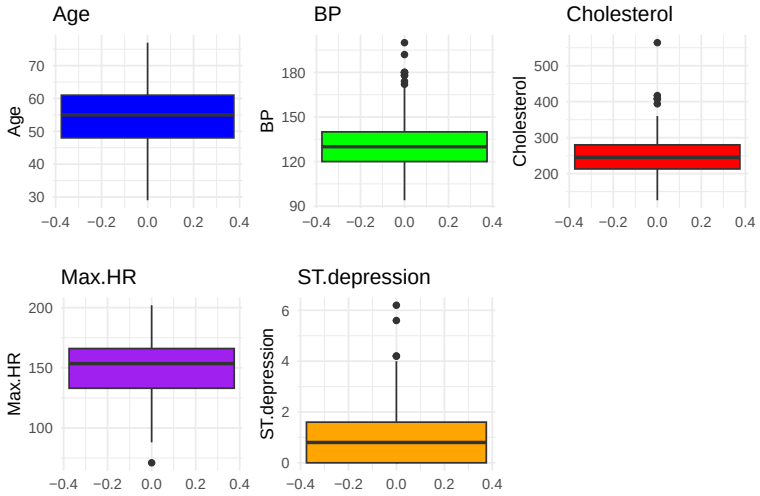

Figure 2: Notched box plots for the numerical variables of the heart disease dataset.

The box plots from Figure 2 shows that `BP`, `Cholesterol`, and `ST.depression` have noted outliers to the higher end of values, where `Max.HR` has one outlier on the lower end. `ST.depression` has a very asymmetrical spread, indicating the data is not normally distributed for this variable.

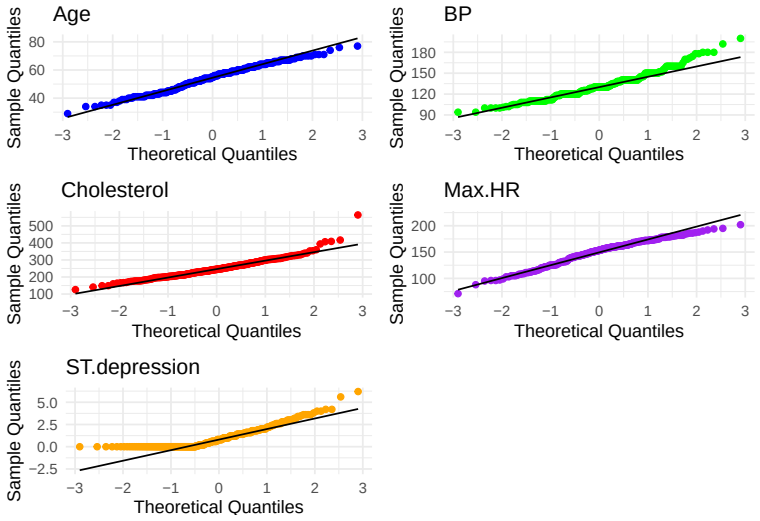

Figure 3: Normal Q-Q plots for the numerical variables of the heart disease dataset.

The normal Q-Q plots from Figure 3 for the variables suggest normality in residuals for all variables except `ST.depression`, where the quantile plot differs significantly from the standard Q-Q line in the negative

theoretical quantiles.

Summary statistics, the covariance matrix, and the correlation matrix are provided below in Tables 1, 2, and 3 respectively.

Table 1: Exploratory data analysis summary for the numerical variables of heart disease data.

	Age	BP	Cholesterol	Max.HR	ST.depression
Sample size	270	270	270	270	270
Minimum	29.00	94.0	126.0	71.0	0.00
1st quartile	48.00	120.0	213.0	133.0	0.00
Median	55.00	130.0	245.0	153.5	0.80
Mean	54.43	131.3	249.7	149.7	1.05
3rd quartile	61.00	140.0	280.0	166.0	1.60
Maximum	77.00	200.0	564.0	202.0	6.20
Skewness	-0.1627	0.7186	1.1771	-0.5248	1.2559
Kurtosis	2.4431	3.884	7.7833	2.8767	4.7048

Table 2: Covariance matrix of the numerical variables of the heart disease dataset.

	Age	BP	Cholesterol	Max.HR	ST.depression
Age	82.975093	44.426394	103.605452	-84.874721	2.026208
BP	44.426394	319.037051	159.731186	-16.193433	4.557435
Cholesterol	103.605452	159.731186	2671.467107	-22.437340	1.640149
Max.HR	-84.874721	-16.193433	-22.437340	536.650434	-9.260037
ST.depression	2.026208	4.557435	1.640149	-9.260037	1.311506

Table 3: Covariance matrix of the numerical variables of the heart disease dataset.

	Age	BP	Cholesterol	Max.HR	ST.depression
Age	1.0000000	0.2730528	0.2200563	-0.4022154	0.1942339
BP	0.2730528	1.0000000	0.1730192	-0.0391357	0.2227998
Cholesterol	0.2200563	0.1730192	1.0000000	-0.0187392	0.0277092
Max.HR	-0.4022154	-0.0391357	-0.0187392	1.0000000	-0.3490454
ST.depression	0.1942339	0.2227998	0.0277092	-0.3490454	1.0000000

A correlogram is generated for correlation between variables and presented in Figure 4.

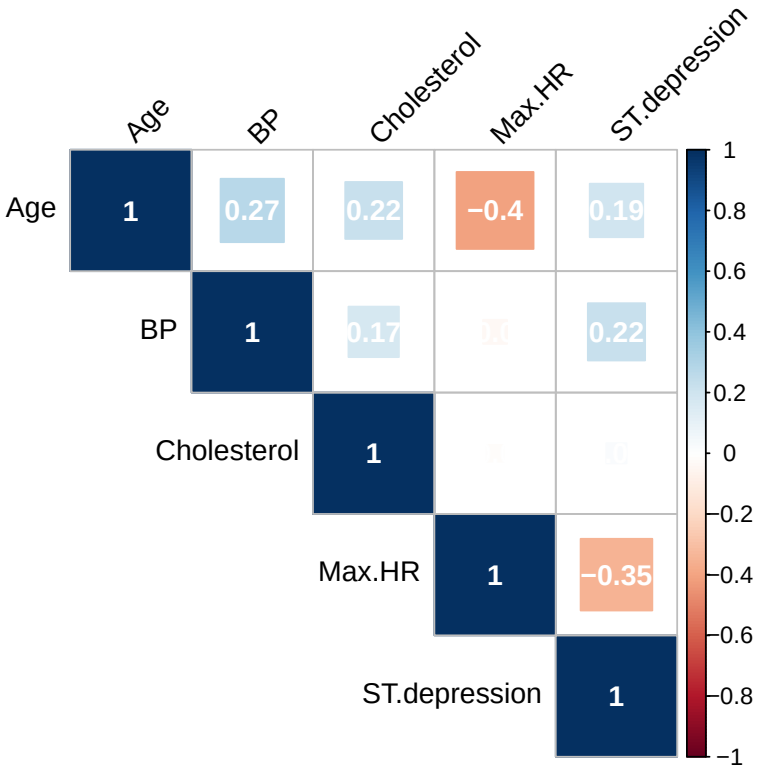

Figure 4: Correlogram of numerical variables in the heart disease dataset.

The only notable correlations are with `Age` and `Max.HR`, and with `Max.HR` and `ST.depression`. These correlations are moderate and negative.

Next the data should be tested for normality using the Anderson-Darling test. The results will be found in Table 4.

Table 4: Anderson-Darling test p-values for normality.

Age	BP	Cholesterol	Max.HR	ST.depression
0.0060111	8.3e-06	0.0008372	1.05e-05	0

There is evidence to reject normality for all numerical variables by the Anderson-Darling test, at the $\alpha = 0.05$ significance level.

Timeline for Project:

Friday 8th September: Group meeting 11-1pm.
- Find dataset.
- Split workload for project report 1.
Grace & Hannah: reason for choice, Availability, number of variables & observations.
Thomas & Aiden: EDA
Bridget: Research Qs and Timeline.

Friday 9th September – Thursday 14th September:
- working individually on assigned parts of the assignment.

Friday 15th September: Group meeting 12-2pm.
- Combine individual efforts, go over everyone parts.
- make report for submission.

Monday 18th September:
- Hand in submission 1.

Friday 22nd September: Group Meeting 11:1pm.
- Focus goals
- Discuss any feedback received as a group and implement.

Friday 6th October:
- Submission 2 hand in.

Friday 13th October: Group Meeting 11-1pm.
- TBC

Friday 20th October:
- Final Submission.