|  |
| --- |
| Cartoon Heart Drawing - How To Draw A Cartoon Heart Step By Step  comp5310 – project Stage 2  Data Summarisation and Analysis |

­

**Section 1: Problem and Overview of Data**

In this report, the problem we will be tackling is the Heart Disease Prediction dataset on the diagnosis of heart disease. The goal is to develop a model that can accurately predict the presence or absence of heart disease based on patient information such as age, sex, blood pressure and cholesterol level. This problem is of great importance as heart disease is the leading cause of death worldwide and early diagnosis can lead to better treatment and improved outcomes. The dataset contains 12 attributes such as age, sex, blood pressure, cholesterol levels and maximum heart rate. The outcome variable is binary, with 1 indicating the presence of heart disease and 0 indicating the absence of heart disease. We will be performing an Exploratory Data Analysis (EDA) on the dataset to better understand the data and identify potential predictors of heart disease. Model selection and evaluation can be the corresponding methods used during later stages of the project.

The data analysis began with importing the dataset and exploring the data by visualising the distributions of the variables using histograms and box plots. The EDA also included correlation analysis to identify any linear relationships between the variables. The results of the EDA signified that age, sex, and maximum heart rate had the strongest correlation with the presence of heart disease. Additionally, the analysis showed that some variables, such as serum cholesterol and resting electrocardiographic results did not have a strong correlation with heart disease. The data analysis also included data pre-processing steps such as dealing with missing data and categorical variables. The categorical variables were encoding using one-hot encoding technique to allow for the use of machine learning models that require numerical inputs.

**Section 2: Data Pre-processing**

Once we have obtained the data, it is stored in a csv format consisting of 12 columns and 918 rows, including numerical and categorical data. The next step is to pre-process the data in Jupyter notebook using the pandas library, as demonstrated in appendix 1.1. Upon examining the variables, we find that the outcome variable is HeartDisease, which is binary, while the other predictors consist of quantitative and qualitative variables. The numerical and categorical variables can be found in appendix 1.2. To gain an initial understanding of the data, we examine the basic descriptive statistics, revealing the FastingBS and HeartDisease are binary variables, with minimum and maximum values of 0 and 1, respectively. Meanwhile, the other predictors are numerical, and categorical variables such as Sex and ExerciseAngina are not shown. The predictor with the highest mean value is Cholesterol, while FastingBS has the lowest.

In appendix 1.3, a correlation analysis was conducted, which revealed that most predictors are not highly correlated. This is evident from the R-score values, which did not exceed 0.7, indicating that there will be no issues with multicollinearity. However, it is worth noting that MaxHR and HeartDisease exhibit a negative correlation with an R-score of -0.4, which requires further investigation through linear regression to quantify parameter estimations and assess the relationship between predictors and the outcome. Furthermore, to check for null values, we used the .isna().sum()command, and as shown in appendix 1.4, no null values were found. To confirm the absence of duplicate values, we used the .drop\_duplicates()command, and after checking the dataset’s shape in appendix 1.5, it was confirmed that no duplicate values exist.

**Section 3: Exploratory Data Analysis**

The EDA conducted in this report will be instrumental for healthcare professionals to gain insights into the factors that contribute to heart failure and identify patients at high risk of developing the condition, thereby facilitating early intervention and improved patient outcomes. We will be able to gather data8 and evidence to assist in answering the following research question: **what are the most important factors in predicting the presence of heart disease, and how can these factors be incorporated into a predictive model?**

Initial exploration of our data will require some visualisations as shown in appendix 2.1 to 2.4, showcasing a range of different histograms. Other histograms of variables are provided in the jupyter notebook attached in the submission, however we observe there are more males and females in this sample. The most frequent chest pain type experienced by patients is Asymptomatic, whereas Typical Angina is the least common. Scrutinising the quantitative variables in this data set such as RestingBP, we observe a left-tail distribution with most of the data revolving around 120-140 resting blood pressure. Moreover, analysing the Cholesterol levels in the histogram shown in appendix 2.4 indicates that most patients have 0 Cholesterol which is not possible as research shows that Cholesterol is produced by the liver and present in the cells of the human body naturally. Therefore, there is an incentive to remove this data in order to produce more accurate results.

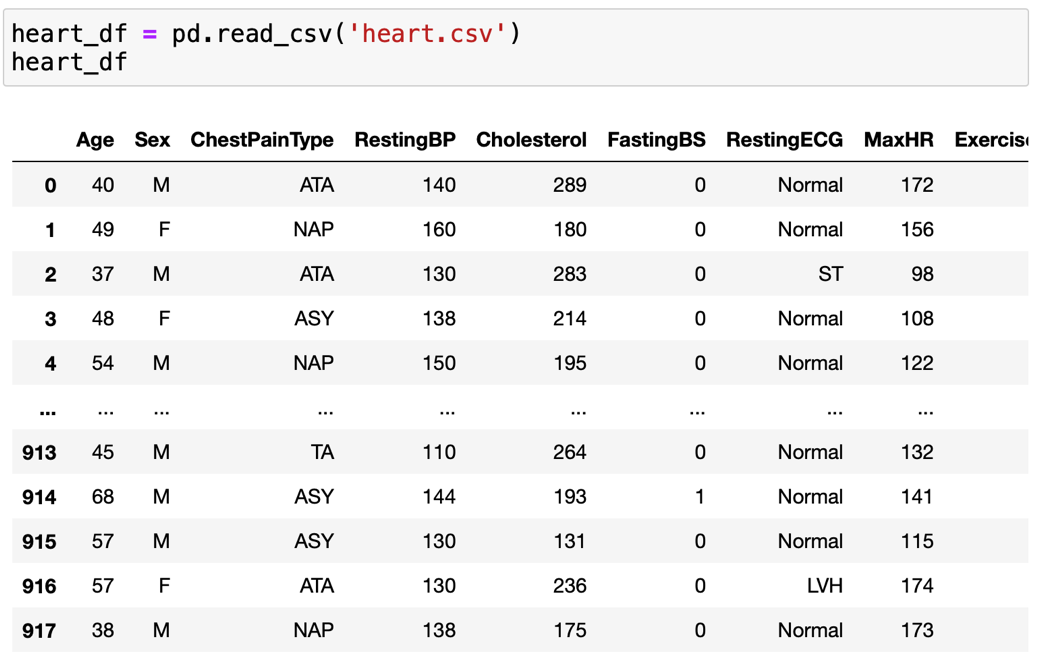
Further data visualisations have been provided in the form of pie charts, displayed in appendix 2.5 and 2.6. Here, we have provided pie charts for categorical columns, where the first pie chart highlights the percentage of patients with heart disease and patients without it, approximately 55% and 45% respectively. The pie chart which displays the RestingECG is the resting electrocardiogram results of the patients. We observe that most patients have normal RestingECG of 60.13%, followed by the least populated segment of ST-T wave abnormality. Extending our correlation analysis in section 2, we’ve added pair plots of the columns in the data set in appendix 2.7. It is noteworthy to state that all the variables included in this pair plot are quantitative and the categorical variables have been omitted as we cannot provide insights on linear relationships. Inspecting these plots alludes that there is only one bivariate relationship that indicates a linear relationship, whereas the other variables are mostly clustered. Mainly the relationship between MaxHR and Age, which has a negative relationship as shown by the data points following a negative trend. As a result, we conclude that as the age of each patient increases, their maximum heart rate achieved is lower. Although there could be confounding variables which may influence this result, such as the dietary choices and exercising levels of each patient, thus we cannot prove causality. Interpreting this relationship does not provide us with great confidence as their influence on the dependent variable HeartDisease will require more statistical analysis in linear regression more specifically that will be conducted during model deployment.

To extend our analysis of bivariate relationships in the heart dataset, we’ve created join plots in appendix 2.8 and 2.9. The first plot as discussed above shows a negative and downward relationship, however this plot incorporates a line of best fit which confirms this interpretation. Further, the relationship between MaxHR and Cholesterol has a positive relationship, indicating that patients with higher maximum heart rate generally tend to have higher Cholesterol levels as well. Although, we cannot confirm that this correlation is strong due to the various outliers with patients having around 500 to 600 Cholesterol levels in mm/dl. More importantly, the misclassified data of patients having Cholesterol levels at 0 is incorrect.

Prolonging our EDA section, let’s observe the potential outliers in our dataset, starting with boxplots. In appendix 3.1 to 3.5 we have included various boxplots to grasp a better understanding of the potential outliers which may affect the accuracy of our results later in the project. We’ve removed various outliers and displayed them in these boxplots, more specifically the boxplot of HeartDisease over RestingBP in appendix 3.2 which contained various data points where patients’ RestingBP was 0. Analogously, patients’ with 0 Cholesterol has also been omitted as displayed as appendix 3.3, where the Cholesterol levels now begin above 0. An EDA summary has been provided in the jupyter notebook using the pandas\_profiling library, however in appendix 4.1 we have provided a few insights on the basic information on the independent variables in our dataset such as the percentage of distinct values of ExerciseAngina with 0.2% and the percentage of missing values being 0.0%, confirming our previous data pre-processing section of removing these missing values as successful. ExerciseAngina is also a Boolean variable with 40.4% of the patients in this dataset testing positive for this symptom, meanwhile 59.6% do not have this. This variable has provided us with a new insight that although there are technically 7 qualitative variables and 5 quantitative variables in this dataset, there is a Boolean variable, however due to their represetnations in datasets, we do not require one-hot encoding to convert them to numeric values. Another valuable discernment we can speculate are the Alerts provided to us in the overview as shown in appendix 4.2. ChestPainType and ST\_Slope is highly overall correlated with HeartDisease, with Oldpeak having 368 zeros. Despite removing all the zeros in Cholesterol, based on research we are unable to remove all the zeros because it is possible for patients to have 0 as the ST depression induced by exercise relative to rest ranges from 0 to 6, therefore we need to remove all negative values which are incorrect, this is shown in appendix 4.3.

**Appendix**

**1.1: Reading the data in Jupyter Notebook**



**1.2: Numerical and Catagorical predictors in heart dataset**

A screenshot of a computer program

Description automatically generated with medium confidence

**1.3: Correlation matrix of all the numerical variables**

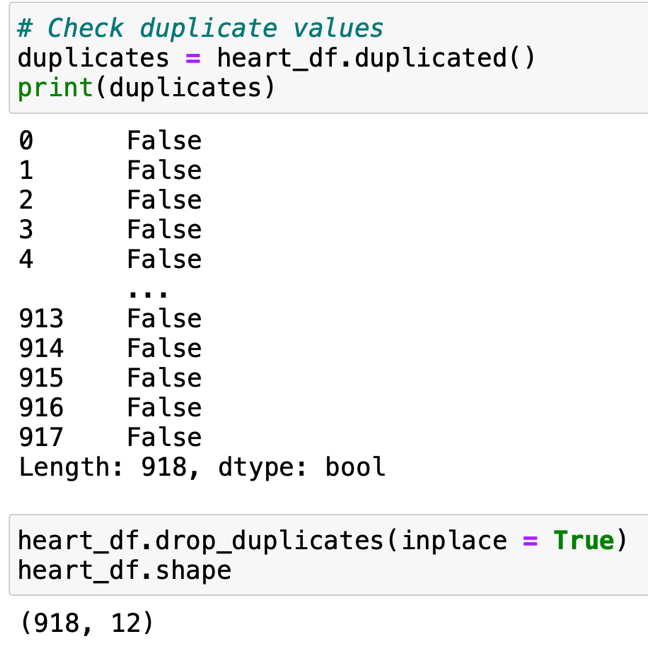
A screenshot of a computer

Description automatically generated with medium confidence

**1.4: Checking for null values**



**1.5: Checking for duplicate values**



**1.6: Encoding categorical variables to numerical variables**



**2.1: Histogram of Sex in heart dataset**

A picture containing text, screenshot, diagram, plot

Description automatically generated

**2.2: Histogram of ChestPainType in heart dataset**

A picture containing text, diagram, plot, screenshot

Description automatically generated

**2.3: Histogram of RestingBP in heart dataset**

A picture containing text, screenshot, plot, diagram

Description automatically generated

**2.4: Histogram of Cholesterol in heat dataset**

A picture containing text, screenshot, diagram, plot

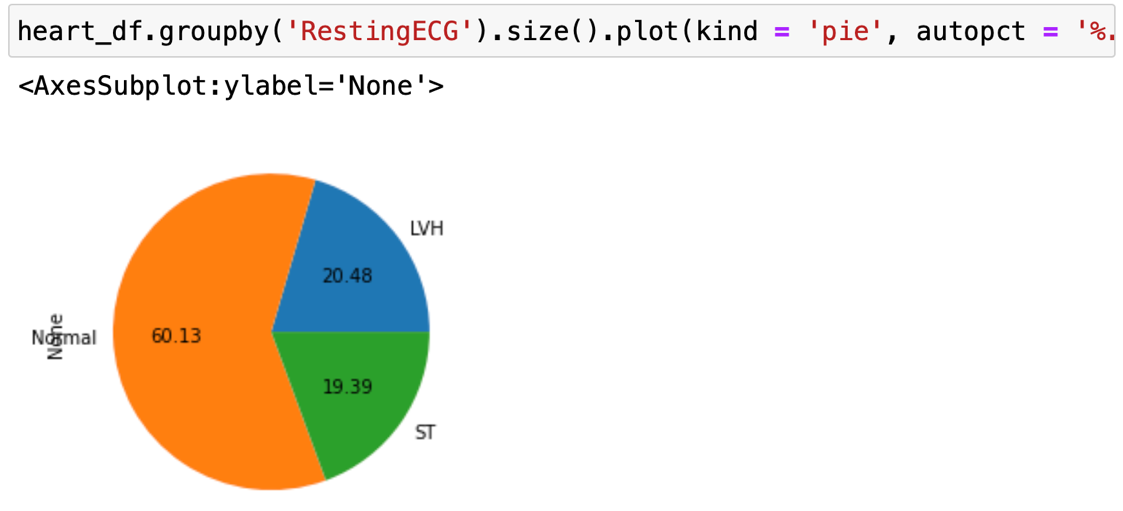
Description automatically generated

**2.5: Pie chart of patients with and without heart disease**

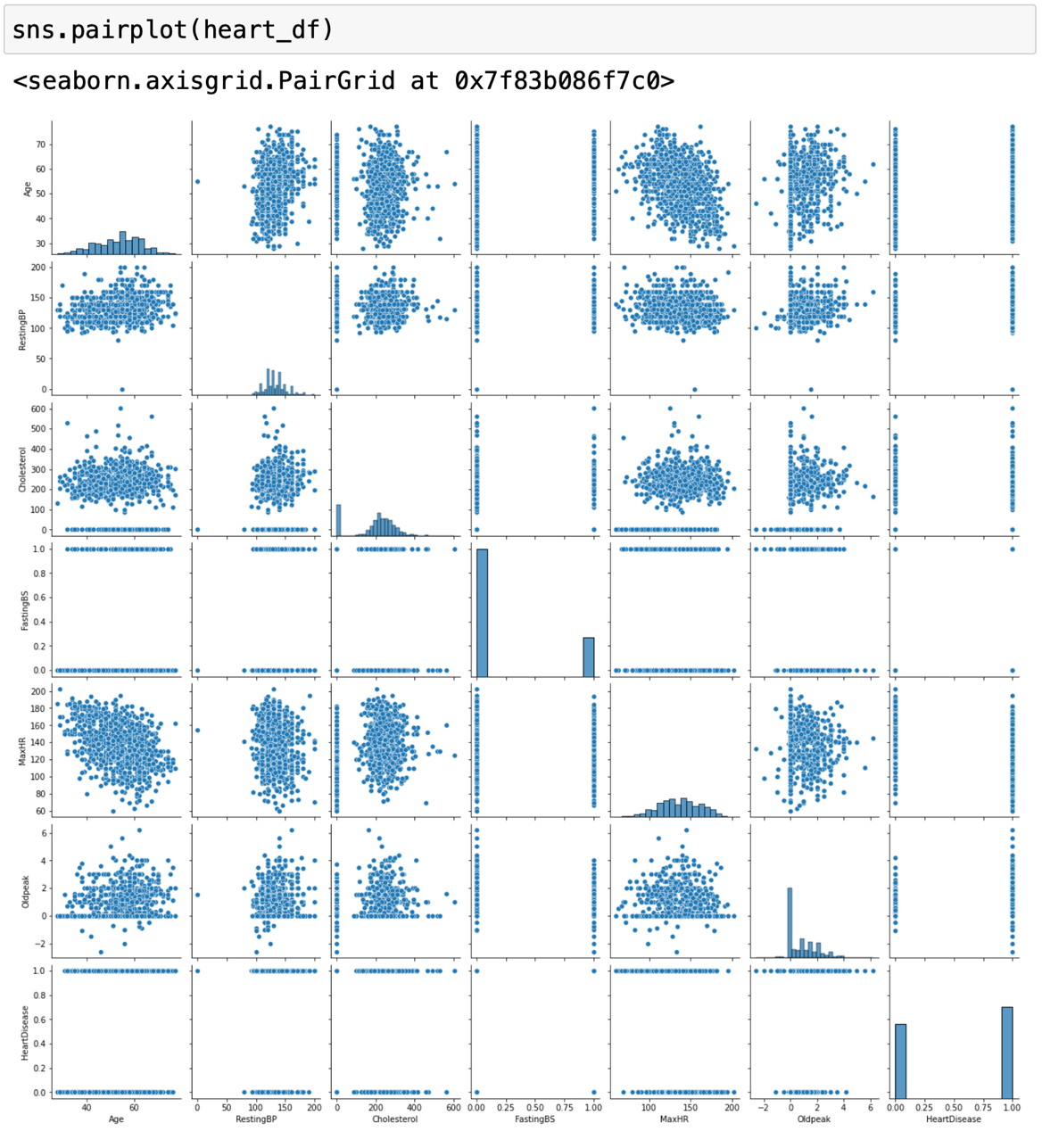
A screen shot of a graph

Description automatically generated with low confidence

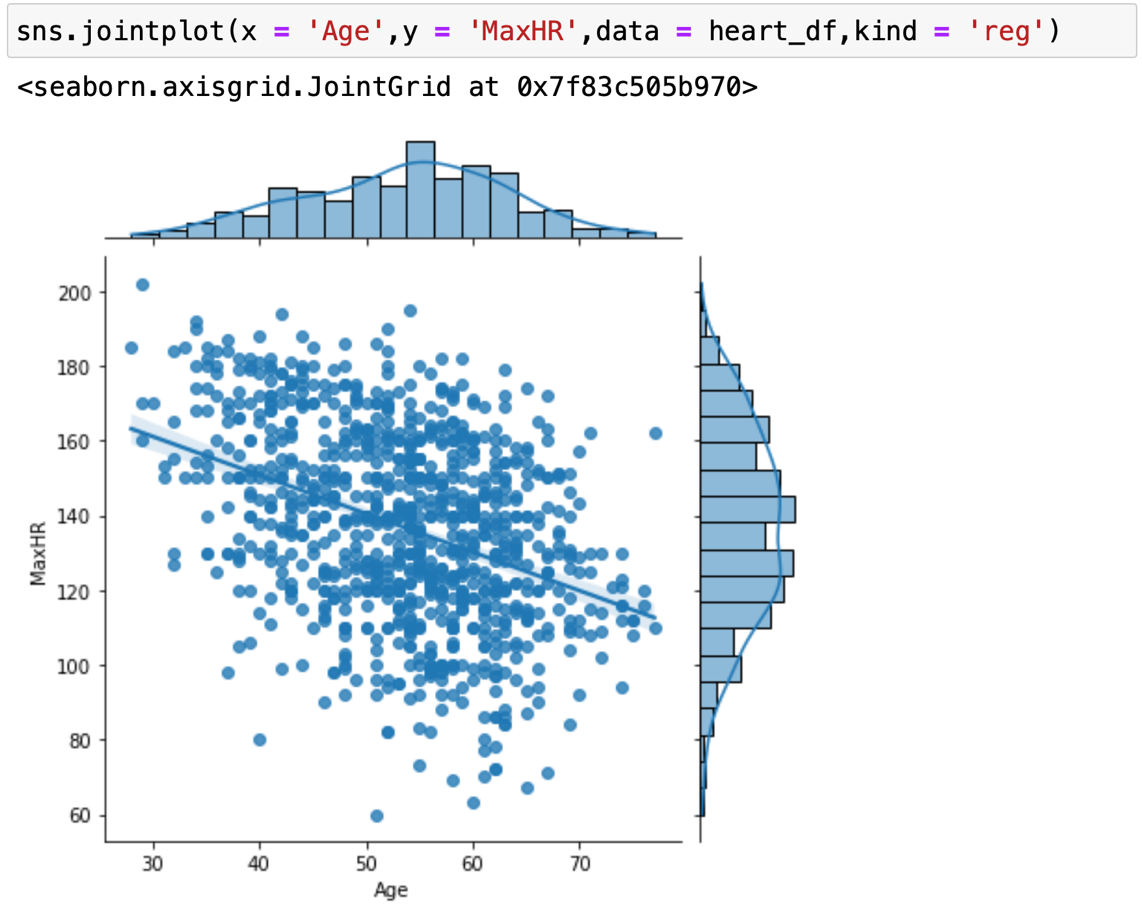
**2.6: Pie chart of patients with different levels of RestingECG**



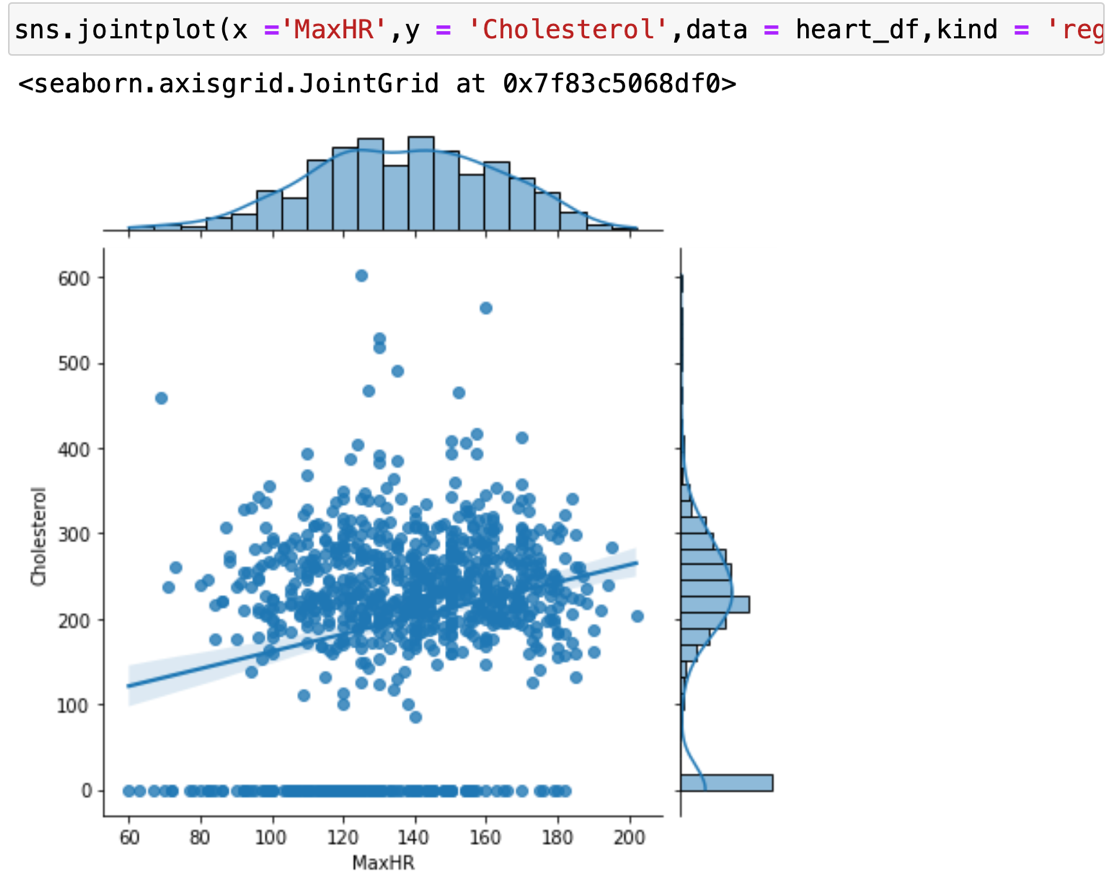
**2.7: Pair plots of variables in the heart dataset**



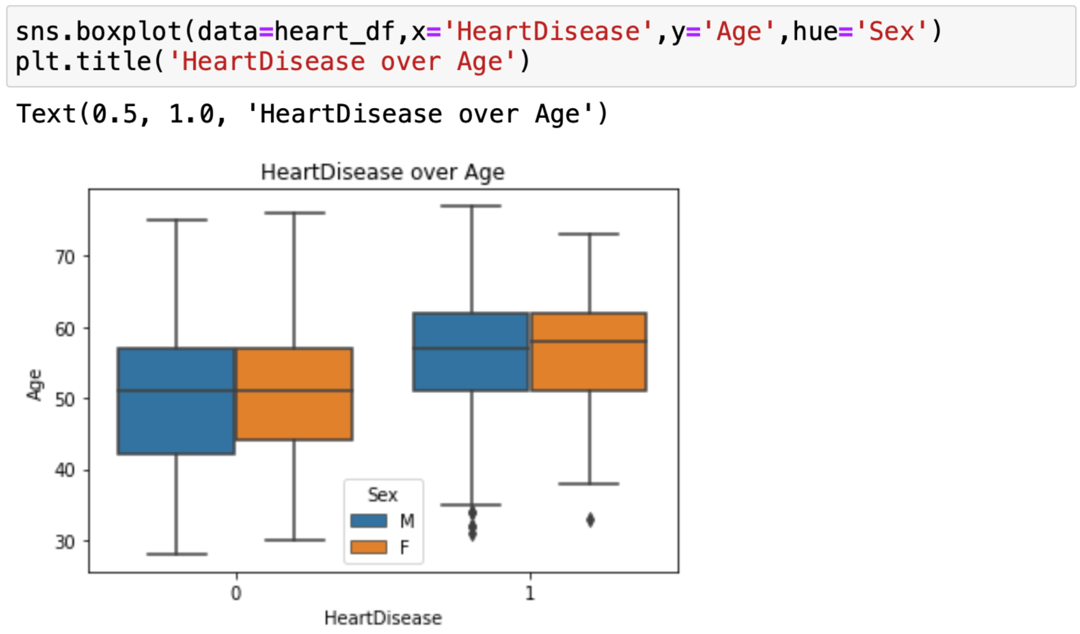
**2.8: Join plot of MaxHR against Age in heart dataset**



**2.9: Join plot of MaxHR against Cholesterol in heart dataset**



**3.1: Boxplot of HeartDisease over age**

****

**3.2: Boxplot of HeartDisease over RestingBP (without outliers)**

****

**A screenshot of a graph

Description automatically generated with low confidence**

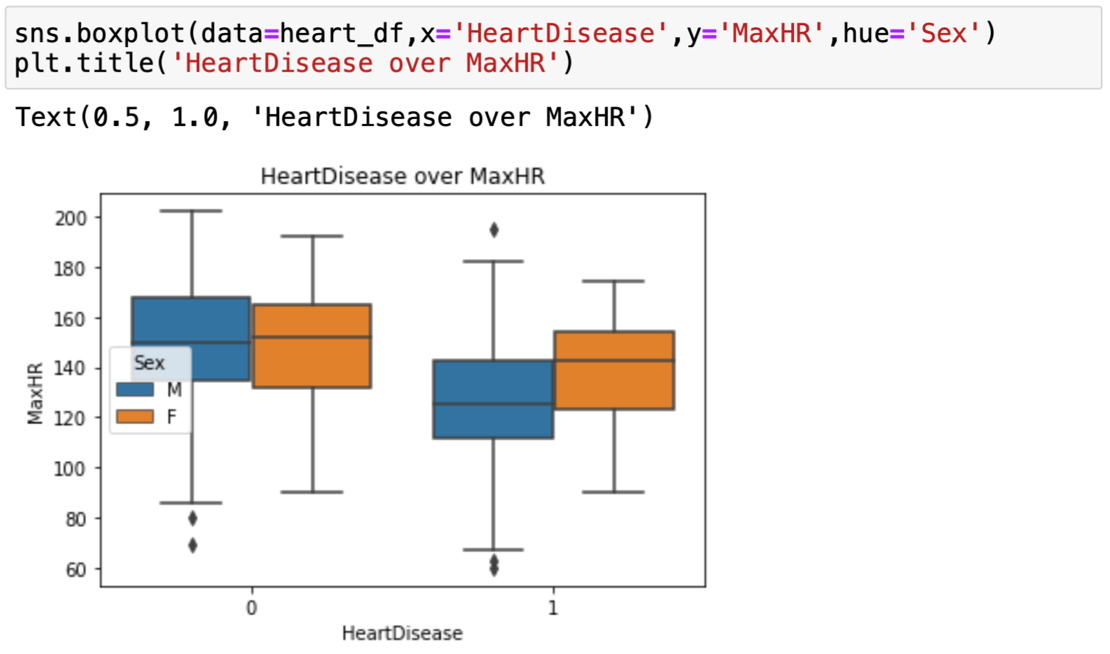
**3.3: Boxplot of HeartDisease over Cholesterol**

****

**A screenshot of a computer screen

Description automatically generated with low confidence**

**3.4: Boxplot of HeartDisease over MaxHR**

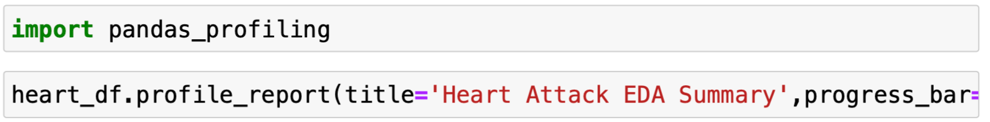
****

**3.5: Boxplot of HeartDisease over Oldpeak**

**A screenshot of a graph

Description automatically generated with medium confidence**

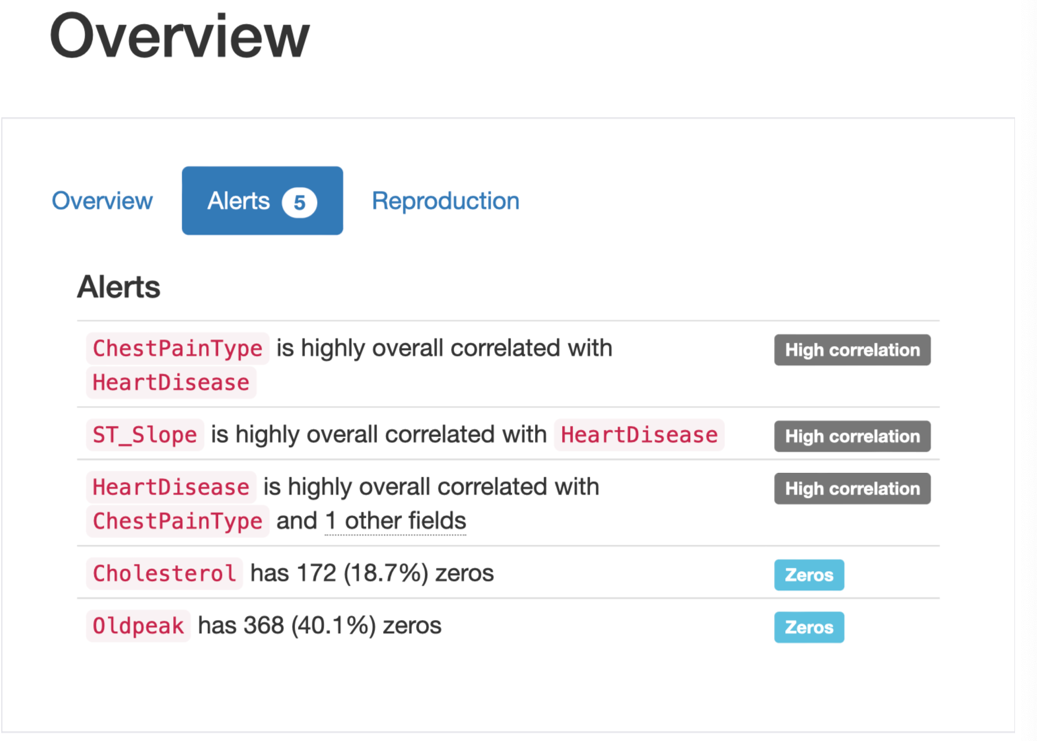
**4.1: EDA Summary of heart\_df using pandas\_profiling library**

****

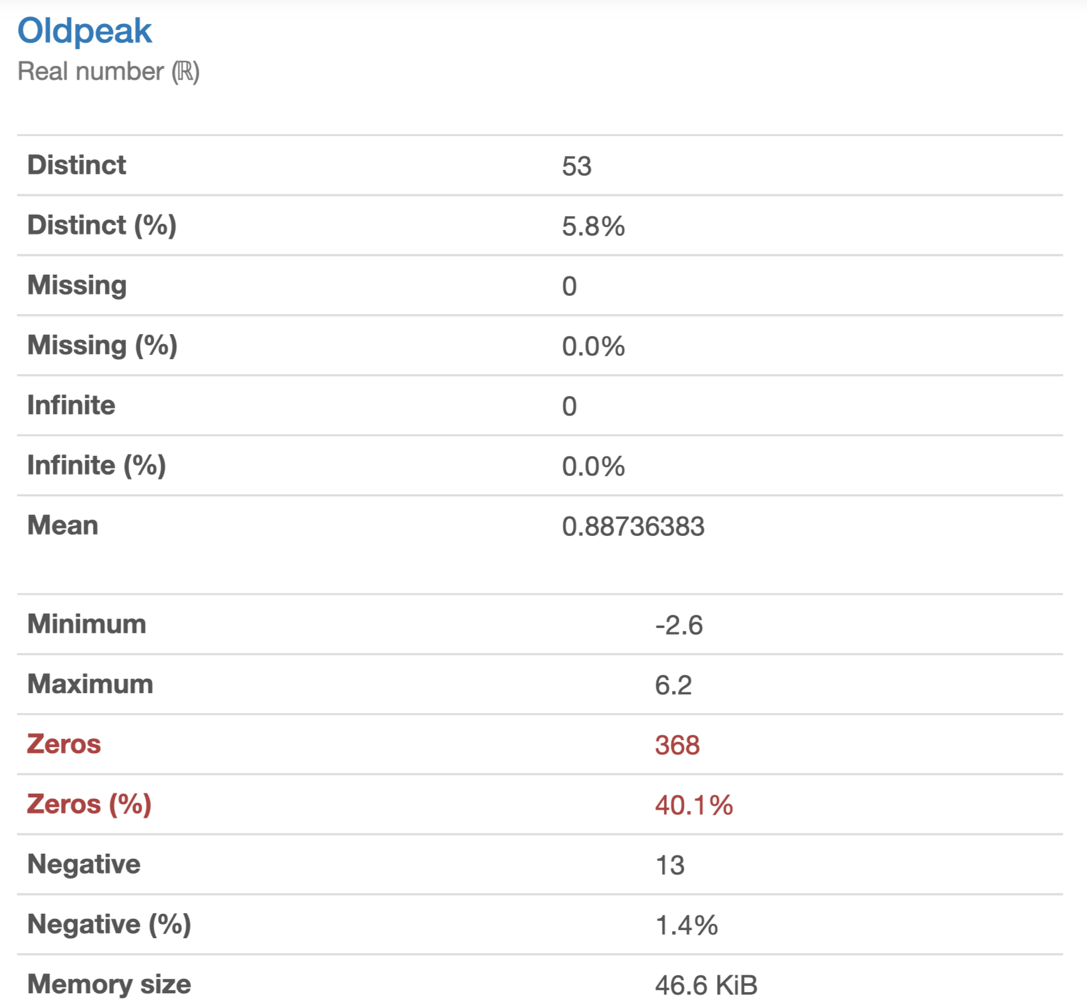
**A screenshot of a computer

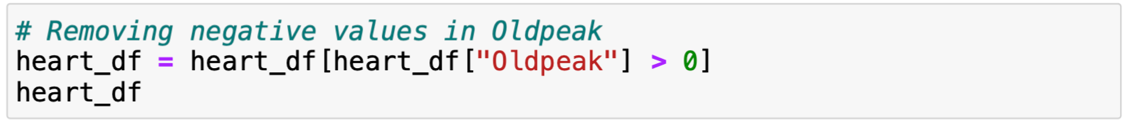
Description automatically generated with low confidence**

**4.2: Overview of Alerts in EDA Summary**

****

**4.3: Negative values present in Oldpeak and removing them**

****

****