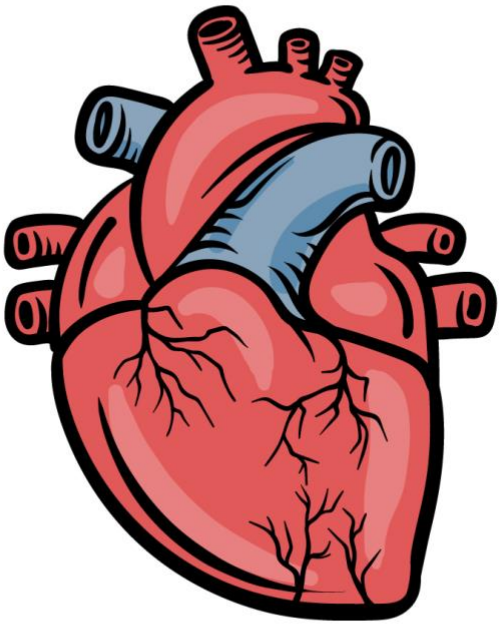


COMP5310 –
PROJECT STAGE 2

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

```
heart_df = pd.read_csv('heart.csv')
heart_df
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	Exercis
0	40	M	ATA	140	289	0	Normal	172	
1	49	F	NAP	160	180	0	Normal	156	
2	37	M	ATA	130	283	0	ST	98	
3	48	F	ASY	138	214	0	Normal	108	
4	54	M	NAP	150	195	0	Normal	122	
...	...	...	...	...	...	...	...	...	
913	45	M	TA	110	264	0	Normal	132	
914	68	M	ASY	144	193	1	Normal	141	
915	57	M	ASY	130	131	0	Normal	115	
916	57	F	ATA	130	236	0	LVH	174	
917	38	M	NAP	138	175	0	Normal	173	

1.2: Numerical and Catagorical predictors in heart dataset

```
# Numerical Variables
numerical = heart_df.select_dtypes(exclude = object).columns
numerical
```

```
Index(['Age', 'RestingBP', 'Cholesterol', 'FastingBS', 'MaxHR', 'Old
peak',
       'HeartDisease'],
      dtype='object')
```

```
# Categorical Variables
categorical = heart_df.select_dtypes(include = object).columns
categorical
```

```
Index(['Sex', 'ChestPainType', 'RestingECG', 'ExerciseAngina', 'ST_S
lope'], dtype='object')
```

1.3: Correlation matrix of all the numerical variables

```
# Descriptive statistics of dataset - mean, std, min, max etc
heart_df.describe().style.background_gradient(cmap = 'Purples')
```

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	53.510893	132.396514	198.799564	0.233115	136.809368	0.887364	0.553377
std	9.432617	18.514154	109.384145	0.423046	25.460334	1.066570	0.497414
min	28.000000	0.000000	0.000000	0.000000	60.000000	-2.600000	0.000000
25%	47.000000	120.000000	173.250000	0.000000	120.000000	0.000000	0.000000
50%	54.000000	130.000000	223.000000	0.000000	138.000000	0.600000	1.000000
75%	60.000000	140.000000	267.000000	0.000000	156.000000	1.500000	1.000000
max	77.000000	200.000000	603.000000	1.000000	202.000000	6.200000	1.000000

1.4: Checking for null values

```
# To check null values
heart_df.isna().sum()
```

```
Age               0
Sex               0
ChestPainType     0
RestingBP         0
Cholesterol       0
FastingBS         0
RestingECG        0
MaxHR             0
ExerciseAngina    0
Oldpeak           0
ST_Slope          0
HeartDisease      0
dtype: int64
```

1.5: Checking for duplicate values

```
# Check duplicate values
duplicates = heart_df.duplicated()
print(duplicates)
```

```
0      False
1      False
2      False
3      False
4      False
       ...
913    False
914    False
915    False
916    False
917    False
Length: 918, dtype: bool
```

```
heart_df.drop_duplicates(inplace = True)
heart_df.shape
```

```
(918, 12)
```

1.6: Encoding categorical variables to numerical variables

```
# Select categorical variables
categ = heart_df.select_dtypes(include=object).columns

# One hot encoding
heart_df = pd.get_dummies(heart_df, columns=categ, drop_first=True)
heart_df.head()
```

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease	Sex_M	ChestPainTy
0	40	140	289	0	172	0.0	0	1	
1	49	160	180	0	156	1.0	1	0	
2	37	130	283	0	98	0.0	0	1	
3	48	138	214	0	108	1.5	1	0	
4	54	150	195	0	122	0.0	0	1	

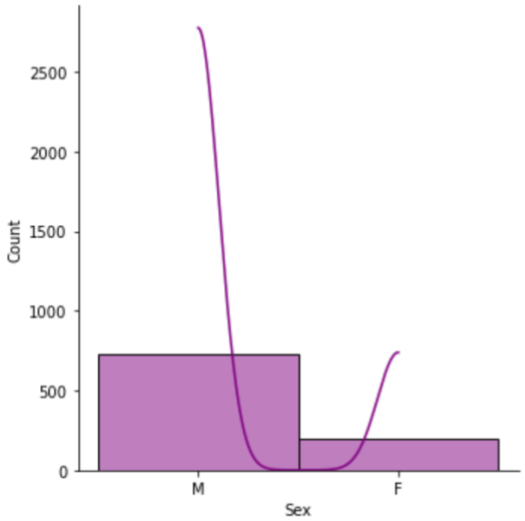
```
heart_df.tail()
```

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease	Sex_M	ChestPain
913	45	110	264	0	132	1.2	1	1	
914	68	144	193	1	141	3.4	1	1	
915	57	130	131	0	115	1.2	1	1	
916	57	130	236	0	174	0.0	1	0	
917	38	138	175	0	173	0.0	0	1	

2.1: Histogram of Sex in heart dataset

```
sns.displot(heart_df['Sex'], color = 'Purple', kde = True)
```

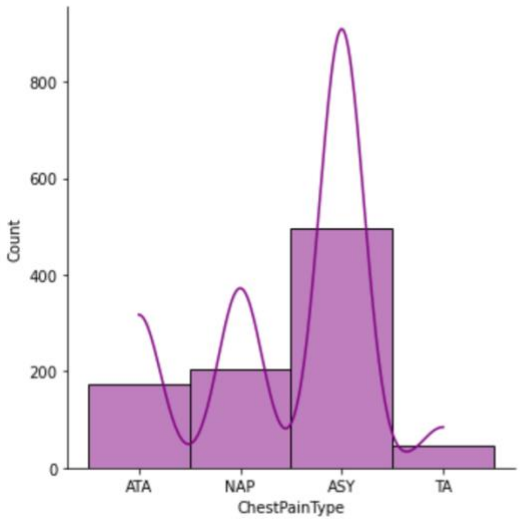
```
<seaborn.axisgrid.FacetGrid at 0x7f83c41923a0>
```



2.2: Histogram of ChestPainType in heart dataset

```
sns.displot(heart_df['ChestPainType'], color = 'Purple', kde = True)
```

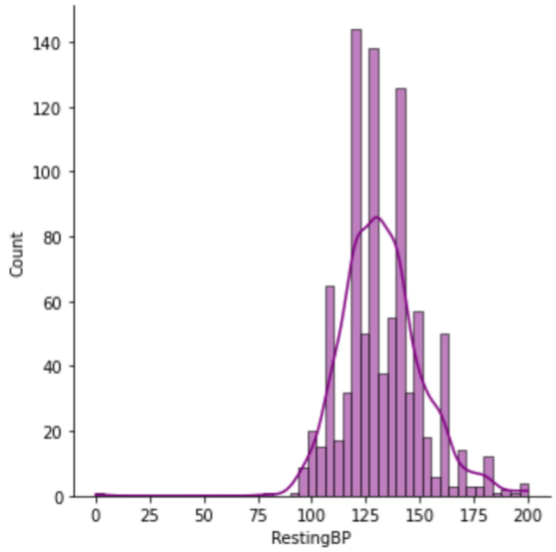
```
<seaborn.axisgrid.FacetGrid at 0x7f83c42ea9a0>
```



2.3: Histogram of RestingBP in heart dataset

```
sns.displot(heart_df['RestingBP'], color = 'Purple', kde = True)
```

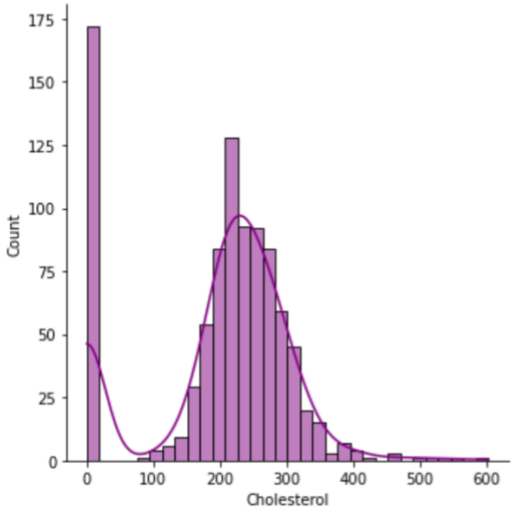
```
<seaborn.axisgrid.FacetGrid at 0x7f83a07531f0>
```



2.4: Histogram of Cholesterol in heat dataset

```
sns.displot(heart_df['Cholesterol'], color = 'Purple', kde = True)
```

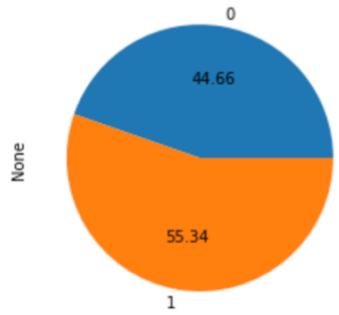
```
<seaborn.axisgrid.FacetGrid at 0x7f83a05a70a0>
```



2.5: Pie chart of patients with and without heart disease

```
heart_df.groupby('HeartDisease').size().plot(kind = 'pie', autopct = '
```

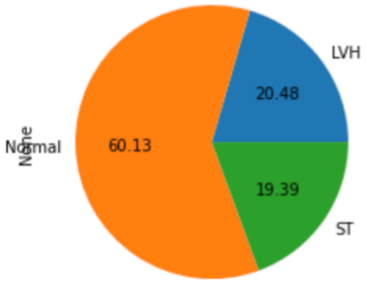
```
<AxesSubplot:ylabel='None'>
```



2.6: Pie chart of patients with different levels of RestingECG

```
heart_df.groupby('RestingECG').size().plot(kind = 'pie', autopct = '%.
```

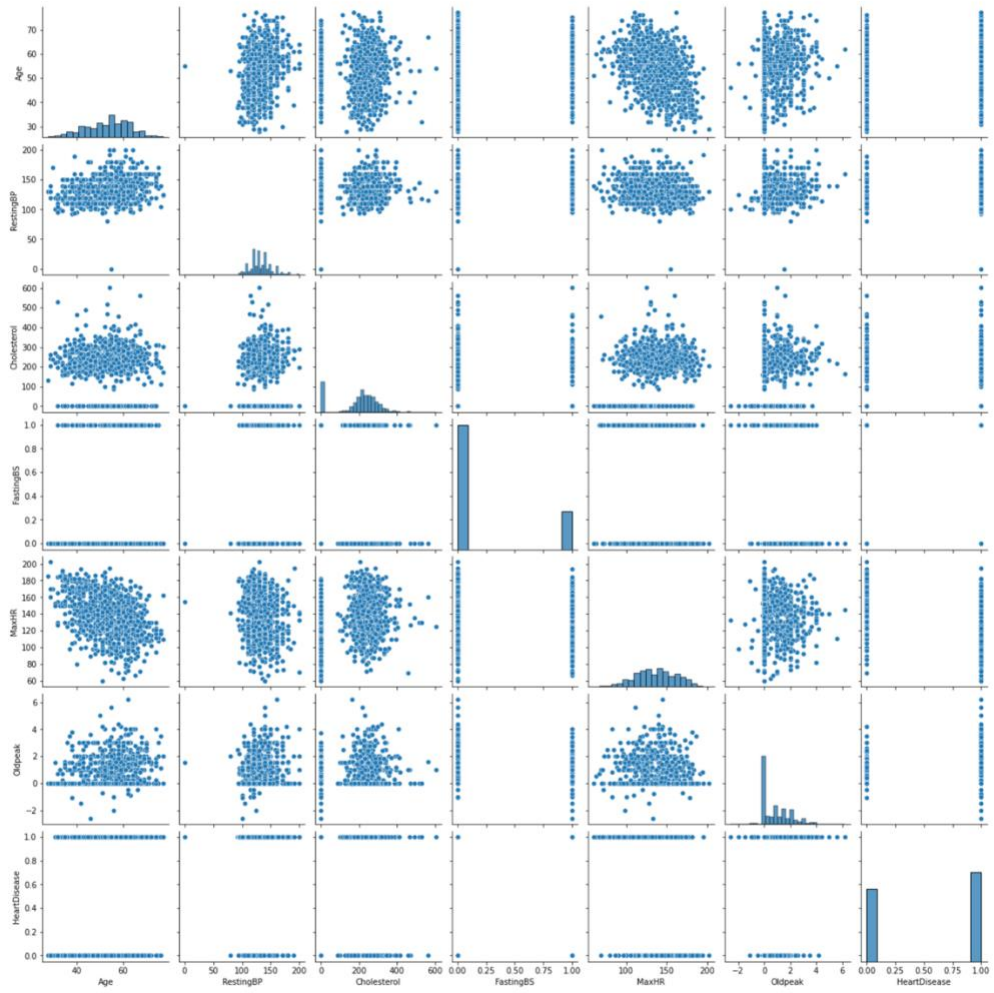
```
<AxesSubplot:ylabel='None'>
```



2.7: Pair plots of variables in the heart dataset

```
sns.pairplot(heart_df)
```

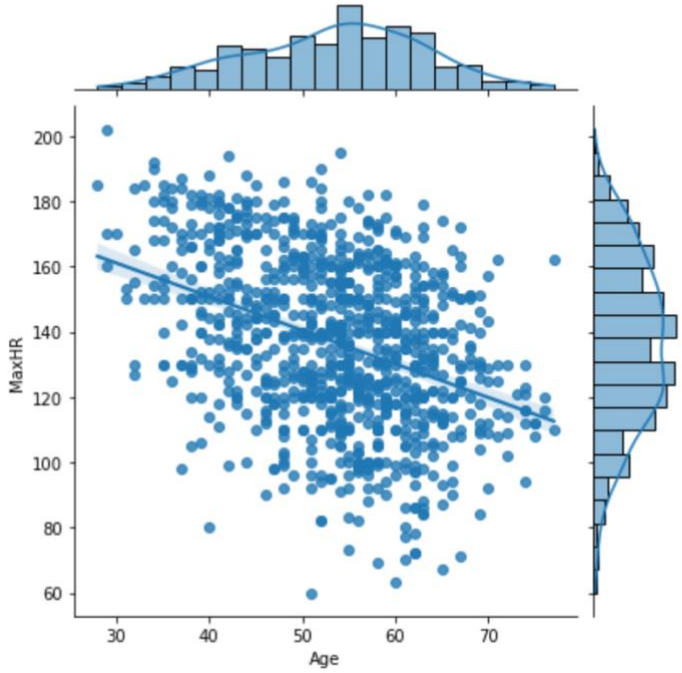
```
<seaborn.axisgrid.PairGrid at 0x7f83b086f7c0>
```



2.8: Join plot of MaxHR against Age in heart dataset

```
sns.jointplot(x = 'Age',y = 'MaxHR',data = heart_df,kind = 'reg')
```

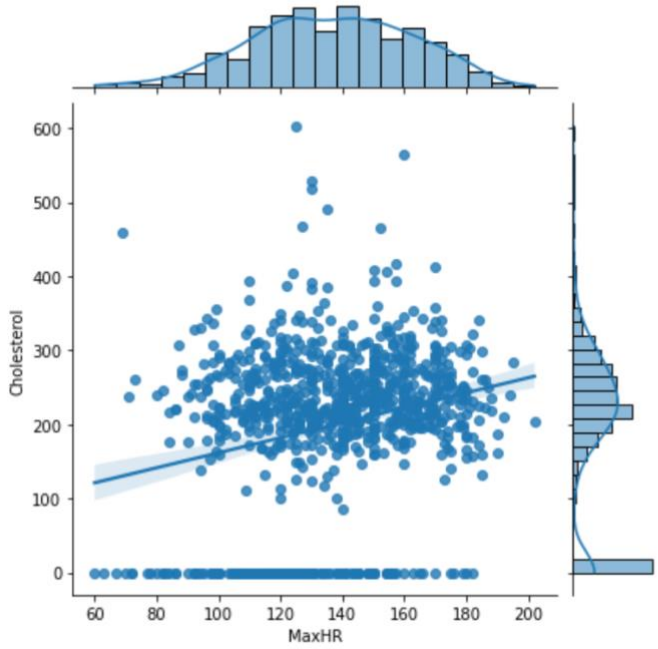
```
<seaborn.axisgrid.JointGrid at 0x7f83c505b970>
```



2.9: Join plot of MaxHR against Cholesterol in heart dataset

```
sns.jointplot(x ='MaxHR',y = 'Cholesterol',data = heart_df,kind = 'reg
```

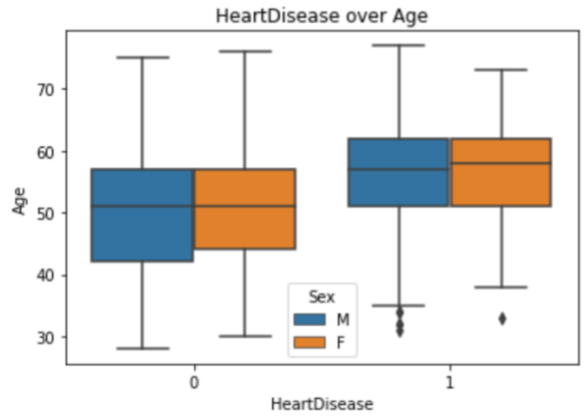
```
<seaborn.axisgrid.JointGrid at 0x7f83c5068df0>
```



3.1: Boxplot of HeartDisease over age

```
sns.boxplot(data=heart_df,x='HeartDisease',y='Age',hue='Sex')
plt.title('HeartDisease over Age')
```

```
Text(0.5, 1.0, 'HeartDisease over Age')
```

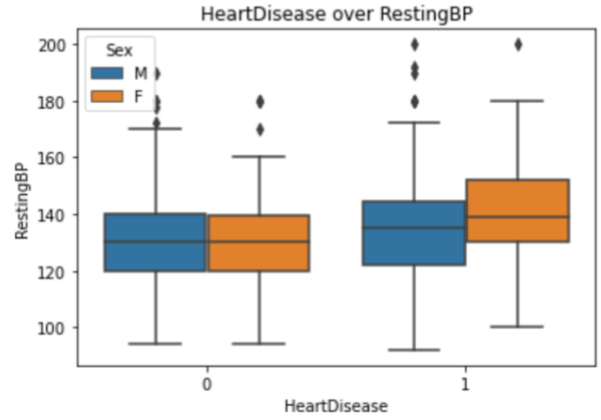


3.2: Boxplot of HeartDisease over RestingBP (without outliers)

```
# Removing zeros in RestingBP
heart_df = heart_df[heart_df["RestingBP"] != 0]
heart_df
```

```
sns.boxplot(data=heart_df,x='HeartDisease',y='RestingBP',hue='Sex')
plt.title('HeartDisease over RestingBP')
```

```
Text(0.5, 1.0, 'HeartDisease over RestingBP')
```

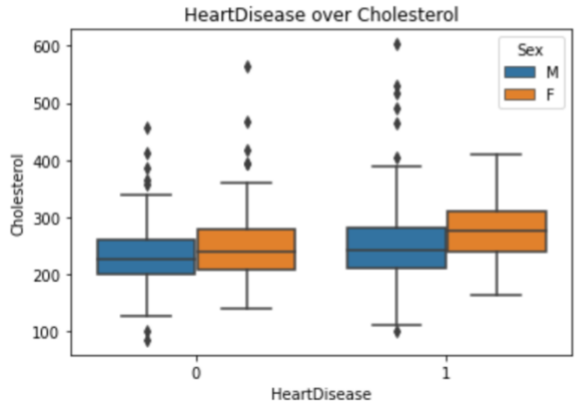


3.3: Boxplot of HeartDisease over Cholesterol

```
# Removing zeros in Cholesterol
heart_df = heart_df[heart_df["Cholesterol"] != 0]
heart_df
```

```
sns.boxplot(data=heart_df,x='HeartDisease',y='Cholesterol',hue='Sex')
plt.title('HeartDisease over Cholesterol')
```

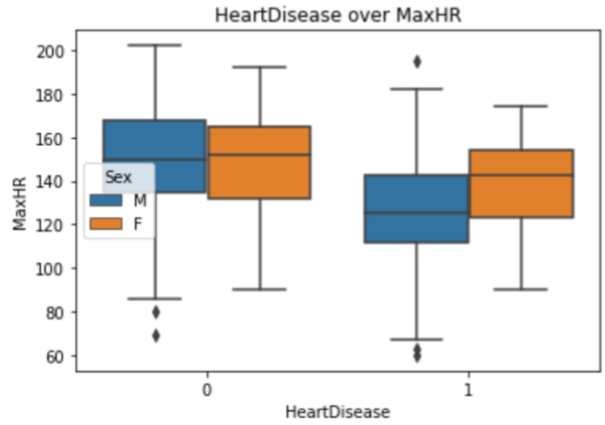
```
Text(0.5, 1.0, 'HeartDisease over Cholesterol')
```



3.4: Boxplot of HeartDisease over MaxHR

```
sns.boxplot(data=heart_df,x='HeartDisease',y='MaxHR',hue='Sex')
plt.title('HeartDisease over MaxHR')
```

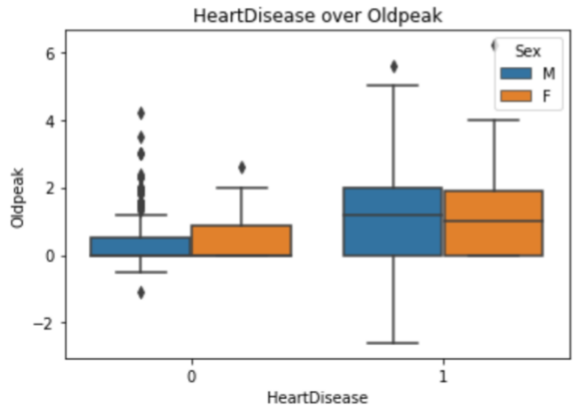
```
Text(0.5, 1.0, 'HeartDisease over MaxHR')
```



3.5: Boxplot of HeartDisease over Oldpeak

```
sns.boxplot(data=heart_df,x='HeartDisease',y='Oldpeak',hue='Sex')
plt.title('HeartDisease over Oldpeak')
```

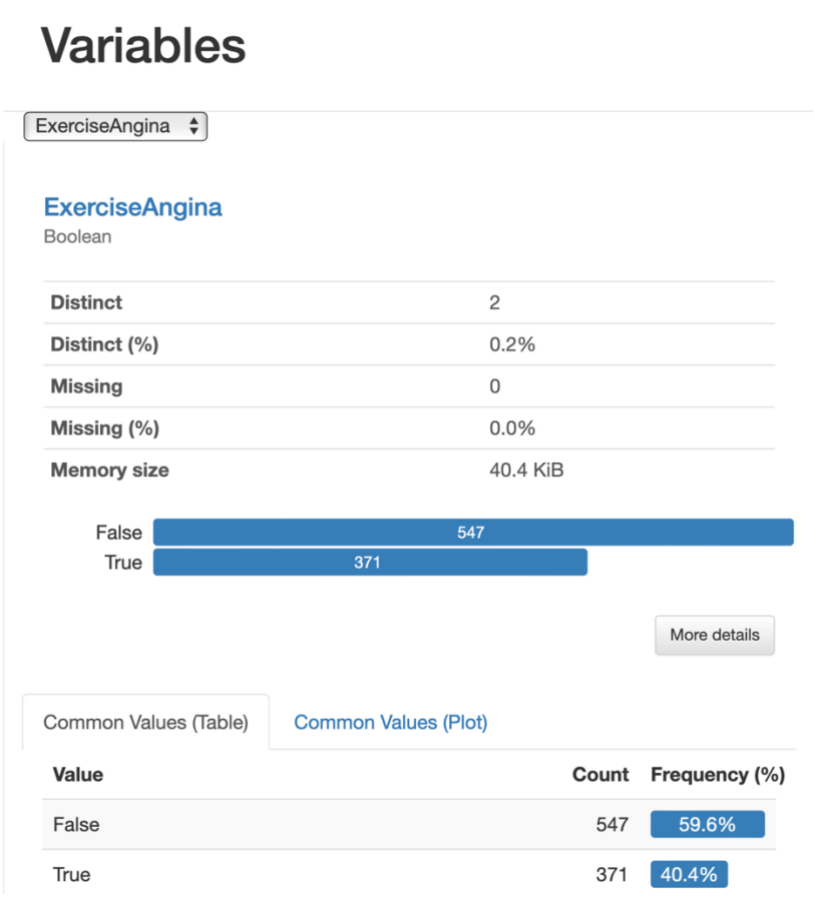
```
Text(0.5, 1.0, 'HeartDisease over Oldpeak')
```



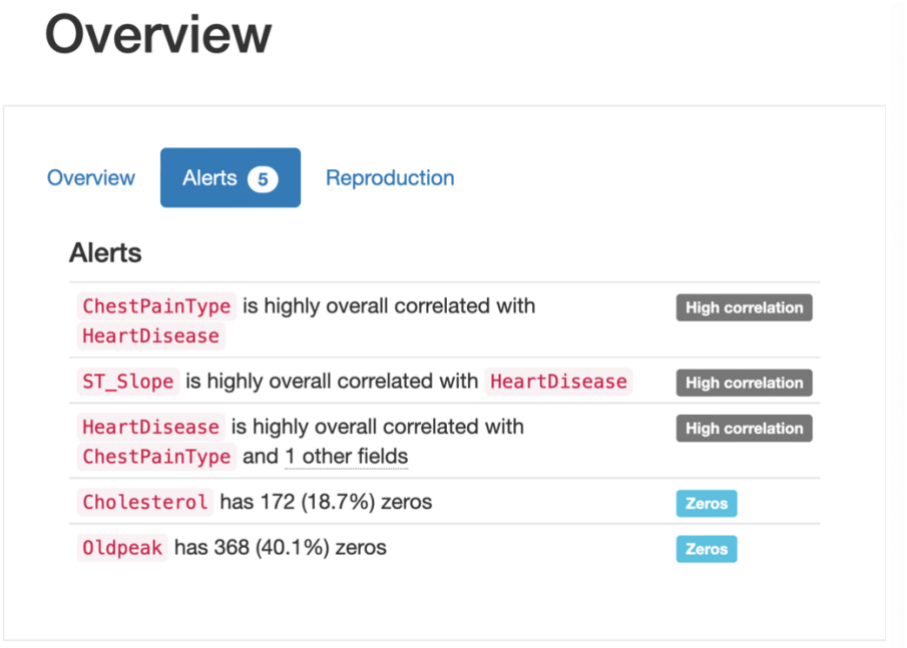
4.1: EDA Summary of heart_df using pandas_profiling library

```
import pandas_profiling
```

```
heart_df.profile_report(title='Heart Attack EDA Summary',progress_bar=
```



4.2: Overview of Alerts in EDA Summary



4.3: Negative values present in Oldpeak and removing them

Oldpeak

Real number (ℝ)

Distinct	53
Distinct (%)	5.8%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%
Mean	0.88736383

Minimum	-2.6
Maximum	6.2
Zeros	368
Zeros (%)	40.1%
Negative	13
Negative (%)	1.4%
Memory size	46.6 KiB

```
# Removing negative values in Oldpeak
heart_df = heart_df[heart_df["Oldpeak"] > 0]
heart_df
```