# IBM Machine Learning Peer-Reviewed Assignment 1

Evangelos Tselentis - Straitouris

2021-04-07

## Contents



## Introduction

The heart is an amazing organ. It continuously pumps oxygen and nutrient-rich blood throughout your body to sustain life. This fist-sized powerhouse beats (expands and contracts) 100,000 times per day pumping 23,000 liters (5,000 gallons) of blood every day. To work properly, the heart (just like any other muscle) needs a good blood supply.

A heart attack (also known as myocardial infarction; MI) is defined as the sudden blockage of blood flow to a portion of the heart. Some of the heart muscle begins to die during a heart attack, and without early medical treatment, the loss of the muscle could be permanent.

Conditions such as high blood pressure, high blood cholesterol, obesity, and diabetes (Martini (2004)) can raise the risk of a heart attack. Behaviors such as an unhealthy diet, low levels of physical activity, smoking, and excessive alcohol consumption can contribute to the conditions that can cause heart attacks. Some factors, such as age and family history of heart disease, cannot be modified but are associated with a higher risk of a heart attack.

## Dataset

For the exploration of the risk a person has to develop a heart attack, the Heart Attack Analysis & Prediction Dataset from *kaggle.com* was utilized. It consists of:

- Age of the patient (age in years)
- Sex of the patient (sex; 1 = male, 0 = female)
- Exercise induced angina (exng; 1 = yes, 0 = no)
- Number of major vessels (caa; 0-3)
- Chest pain type (cp; Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
- Resting blood pressure (trtpbs; in mm/Hg on admission to the hospital)
- Cholesterol levels (chol; in mg/dl)
- Fasting blood sugar (fbs; if > 120 mg/dl, 1 = true; 0 = false)
- Resting electrocardiographic results (rest_ecg; 0 = normal, 1 = having ST-T wave abnormality, 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
- Maximum heart rate achieved (thalachh)
- Chance of heart attack (output; target: Heart disease)
- A blood disorder called thalassemia (thall; 1 = normal; 2 = fixed defect; 3 = reversable defect)
- Previous peak (oldpeak; ST depression induced by exercise relative to rest - 'ST' relates to positions on the ECG plot)
- Slope (slp; the slope of the peak exercise ST segment, Value 1: upsloping, Value 2: flat, Value 3: downsloping)

## Exploratory Data Analysis

According to the report created by the *pandas profiler* package, the dataset has 303 observations and 14 (13 features plus the output label) variables with 5 out of them being numerical; and the rest nine being categorical. Another important finding is that there is one duplicate in the dataset (Figure 1). By using the pandas *drop_duplicate* function, the description of the dataset can be summarized as the Figure 2 shows.

| | age | sex | cp | trtbps | chol | fbs | restecg | thalachh | exng | oldpeak | slp | caa | thall | output |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **164** | 38 | 1 | 2 | 138 | 175 | 0 | 1 | 173 | 0 | 0.0 | 2 | 4 | 2 | 1 |

**Figure 1:** The dataset that was used, has 1 duplicate observation

| | age | sex | cp | trtbps | chol | fbs | restecg | thalachh | exng | oldpeak | slp | caa | thall | output |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 302.00000 | 302.000000 | 302.000000 | 302.000000 | 302.000000 | 302.000000 | 302.000000 | 302.000000 | 302.000000 | 302.000000 | 302.000000 | 302.000000 | 302.000000 | 302.000000 |
| **mean** | 54.42053 | 0.682119 | 0.963576 | 131.602649 | 246.500000 | 0.149007 | 0.526490 | 149.569536 | 0.327815 | 1.043046 | 1.397351 | 0.718543 | 2.314570 | 0.543046 |
| **std** | 9.04797 | 0.466426 | 1.032044 | 17.563394 | 51.753489 | 0.356686 | 0.526027 | 22.903527 | 0.470196 | 1.161452 | 0.616274 | 1.006748 | 0.613026 | 0.498970 |
| **min** | 29.00000 | 0.000000 | 0.000000 | 94.000000 | 126.000000 | 0.000000 | 0.000000 | 71.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 48.00000 | 0.000000 | 0.000000 | 120.000000 | 211.000000 | 0.000000 | 0.000000 | 133.250000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 | 2.000000 | 0.000000 |
| **50%** | 55.50000 | 1.000000 | 1.000000 | 130.000000 | 240.500000 | 0.000000 | 1.000000 | 152.500000 | 0.000000 | 0.800000 | 1.000000 | 0.000000 | 2.000000 | 1.000000 |
| **75%** | 61.00000 | 1.000000 | 2.000000 | 140.000000 | 274.750000 | 0.000000 | 1.000000 | 166.000000 | 1.000000 | 1.600000 | 2.000000 | 1.000000 | 3.000000 | 1.000000 |
| **max** | 77.00000 | 1.000000 | 3.000000 | 200.000000 | 564.000000 | 1.000000 | 2.000000 | 202.000000 | 1.000000 | 6.200000 | 2.000000 | 4.000000 | 3.000000 | 1.000000 |

**Figure 2:** Description of the new dataset without the duplicate

In order to find if there are any missing information from the dataset, a heatmap (Figure 3) of the *isnull* Boolean values was created. Fortunately, the dataset is clean and tidy, and there are no null values. Hence, there is no need of transforming the dataset for handling missing data.

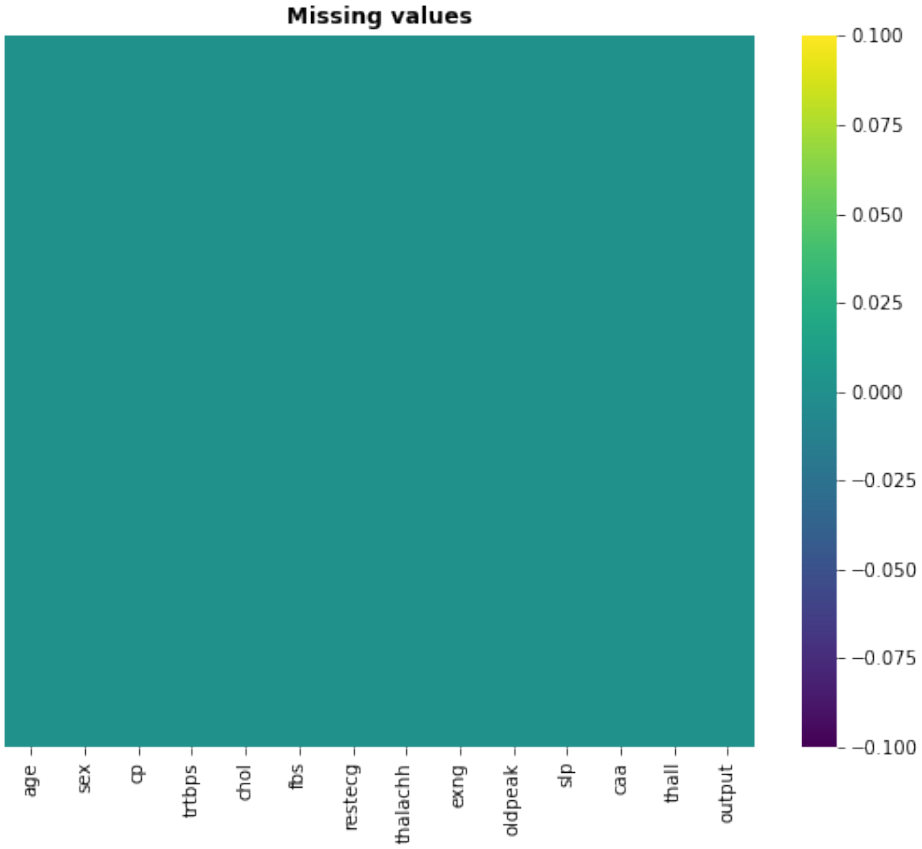


**Figure 3:** No missing values from the dataset. All 302 observations are valid and have no null values.

After the investigation for any missing (null) values, there is the examination of the data for outliers. To

find if there are any outliers in the dataset, boxplots for the numerical features were created. As Figure 4 shows, it is evident that there are outliers for the ‘trtbps,’ ‘chol,’ ‘thalachh’ and ‘oldpeak’ features.

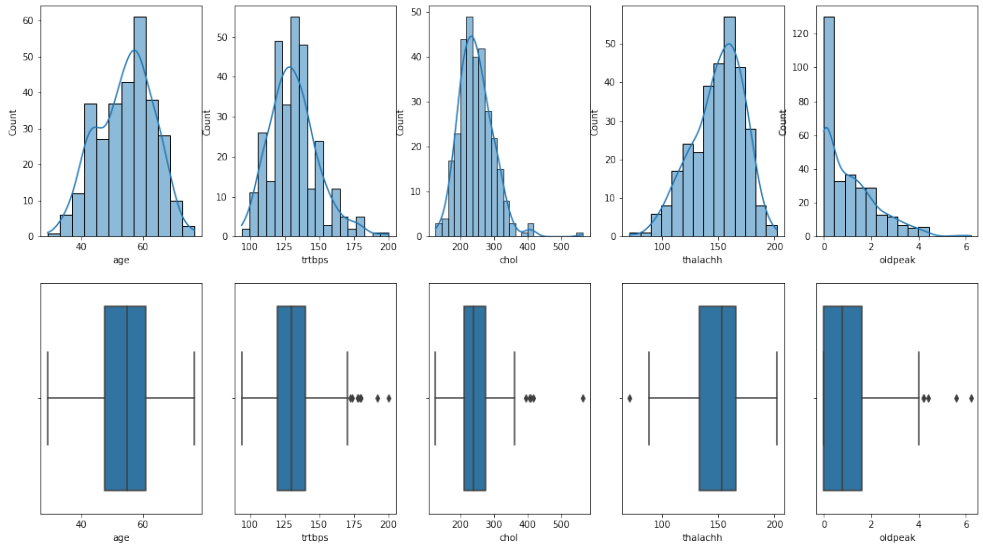


**Figure 4:** Outliers

There are several ways to handle outliers in a dataset. However, since the particular dataset consists of only 302 observations (plus a duplicate), it was decided instead of dropping them to change the outliers with the features medians. For instance, the outlier values for the ‘trtbps’ feature (Figure 5), were changed to its median, i.e. 130. The same procedure was followed for all features with outliers. Figure 6, shows the effect that replacing the outliers with the median values had on the dataset, by creating both histograms and box plots.

| | age | sex | cp | trtbps | chol | fbs | restecg | thalachh | exng | oldpeak | slp | caa | thall | output |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **8** | 52 | 1 | 2 | 172 | 199 | 1 | 1 | 162 | 0 | 0.5 | 2 | 0 | 3 | 1 |
| **101** | 59 | 1 | 3 | 178 | 270 | 0 | 0 | 145 | 0 | 4.2 | 0 | 0 | 3 | 1 |
| **110** | 64 | 0 | 0 | 180 | 325 | 0 | 1 | 154 | 1 | 0.0 | 2 | 0 | 2 | 1 |
| **202** | 68 | 1 | 2 | 180 | 274 | 1 | 0 | 150 | 1 | 1.6 | 1 | 0 | 3 | 0 |
| **222** | 56 | 0 | 0 | 200 | 288 | 1 | 0 | 133 | 1 | 4.0 | 0 | 2 | 3 | 0 |
| **240** | 59 | 0 | 0 | 174 | 249 | 0 | 1 | 143 | 1 | 0.0 | 1 | 0 | 2 | 0 |
| **247** | 54 | 1 | 1 | 192 | 283 | 0 | 0 | 195 | 0 | 0.0 | 2 | 1 | 3 | 0 |
| **259** | 66 | 0 | 0 | 178 | 228 | 1 | 1 | 165 | 1 | 1.0 | 1 | 2 | 3 | 0 |
| **265** | 55 | 0 | 0 | 180 | 327 | 0 | 2 | 117 | 1 | 3.4 | 1 | 0 | 2 | 0 |

**Figure 5:** Outlier values for the resting blood pressure observations. The values that were larger than the maximum (Q3 + 1.5*IQR) were replaced with the median.

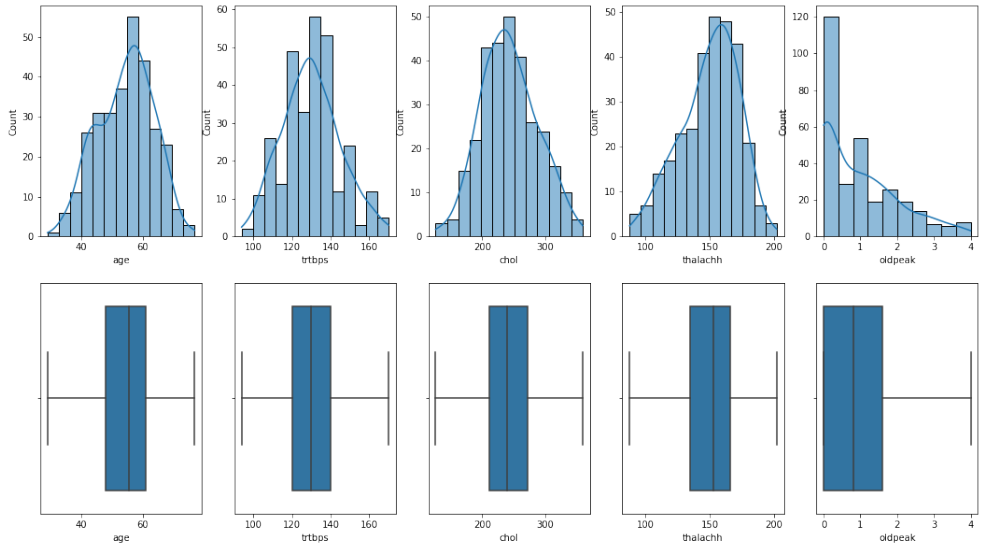


**Figure 6:** The histograms and boxplots of the numerical variables after all outliers were replaced by the median values.

Figure 7 depicts the pairwise relationships between age, resting blood pressure, levels of cholesterol, and the previous peak for the different outputs (chance of a heart attack).

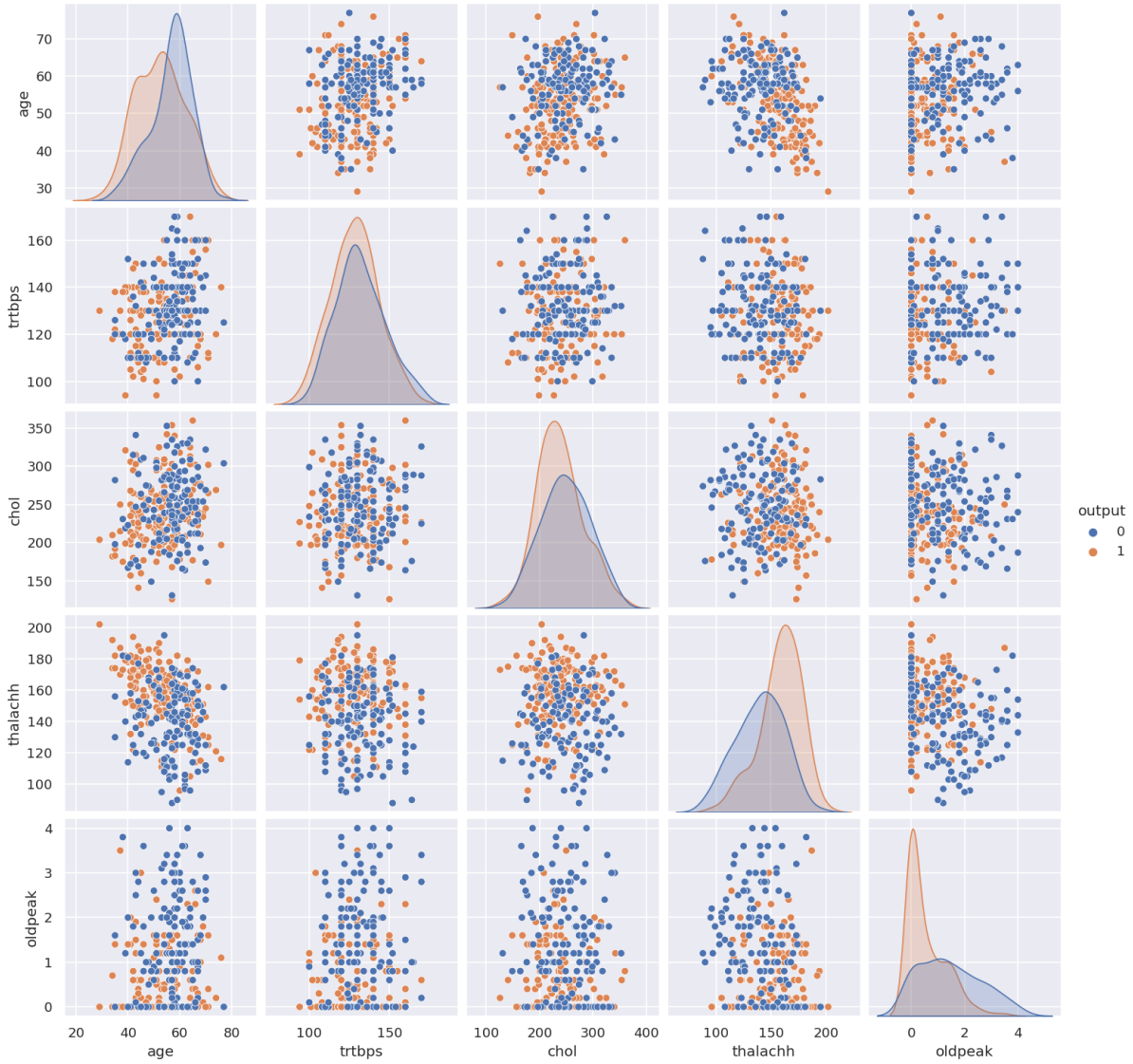


**Figure 7:** Pairplot, using output as hue

Following the analysis, detection, and handling of the outliers, the next step is to explore some of the dataset's features. Using *value_counts* one could deduce that the dataset is quite balanced since there are 164 (54.3%) observations (Figure 8; left pie chart) of people with a high chance to develop a heart attack and 138 (45.7%) with a lesser chance. On the other hand though, in terms of the male: female sex ratio, the dataset shows an imbalance (Figure 8; right pie chart), approximately 2 to 1.

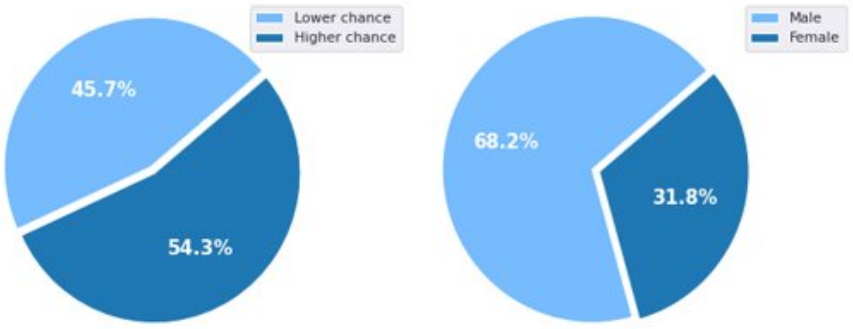


**Figure 8:** Pie plots of both output and sex observations

However, the above pie charts do not give lots of information about the dataset, or the relationship between the chance of a heart attack and the sex of the patient. To solve that, a simple *groupby* function can be used to find which sex has a higher chance of developing a heart attack. According to Figure 9, male individuals have a higher likelihood of evolving heart attack (44.7%; approx 1:1 lower to higher chance) rather than women (25%; approx 3:1 lower to higher chance).

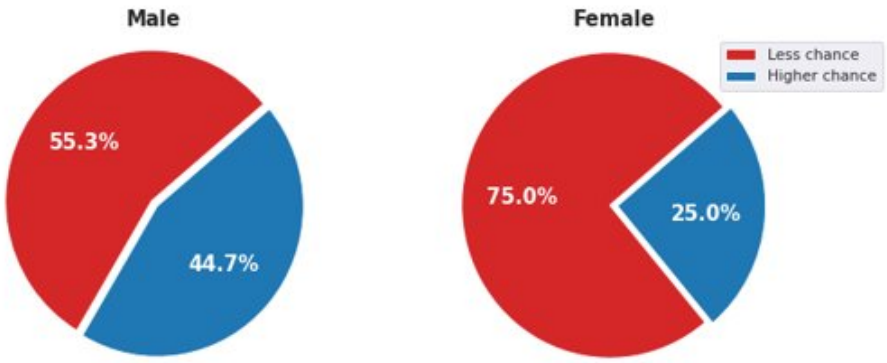


**Figure 9:** Pie plots after grouping sexes, to find which gender shows higher probability of developing serious heart attack. **Left**: Pie plot for males. **Right**: Pie plot for females. Comparing the results, it is evident that men are more likely to show myocardial infarction.

After exploring the relationship between the output and the sex of the individual, it is important to investigate how age can result in a possible heart attack. Figure 10 shows the number of patients concerning different age bands (lower bar plot). Approximately 68% of the patients are older than 50 years old, with the people of age band 50 ~ 60 having the highest hospital admissions; i.e., 125. Near 55% of the total admissions, show a high chance of developing myocardial infarction, with the upper plot displaying that the lower the age, the more elevated is the possibility of a serious heart attack for the individual. Nonetheless, it must be pointed out that only one person of the age band between 20

and 30 years old was admitted based on the specific dataset. Hence it would be helpful if more data were given for the specific age group.

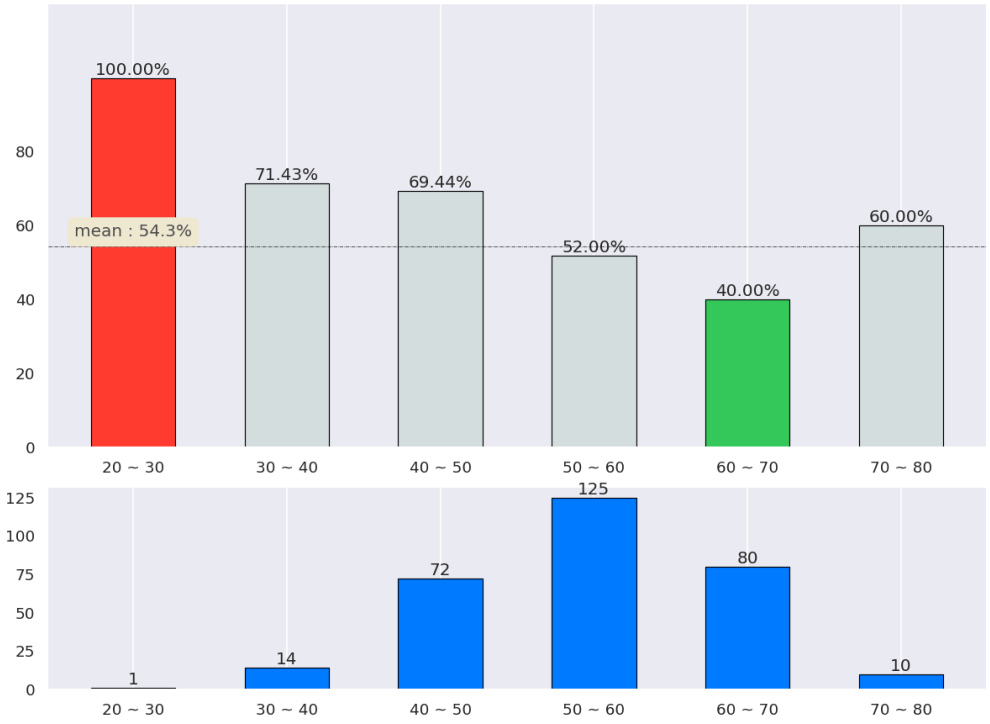


**Figure 10:** Bar plots for different age bands, and their responding heart attack chances.

Subsequent to the exploration of some important numerical variables, like age and sex of the patient, it is prudent to examine the corellation between the given features. By using the *correlation* function, with the Pearson method as input, the heatmap (Figure 11) of the dataset will be as follows:

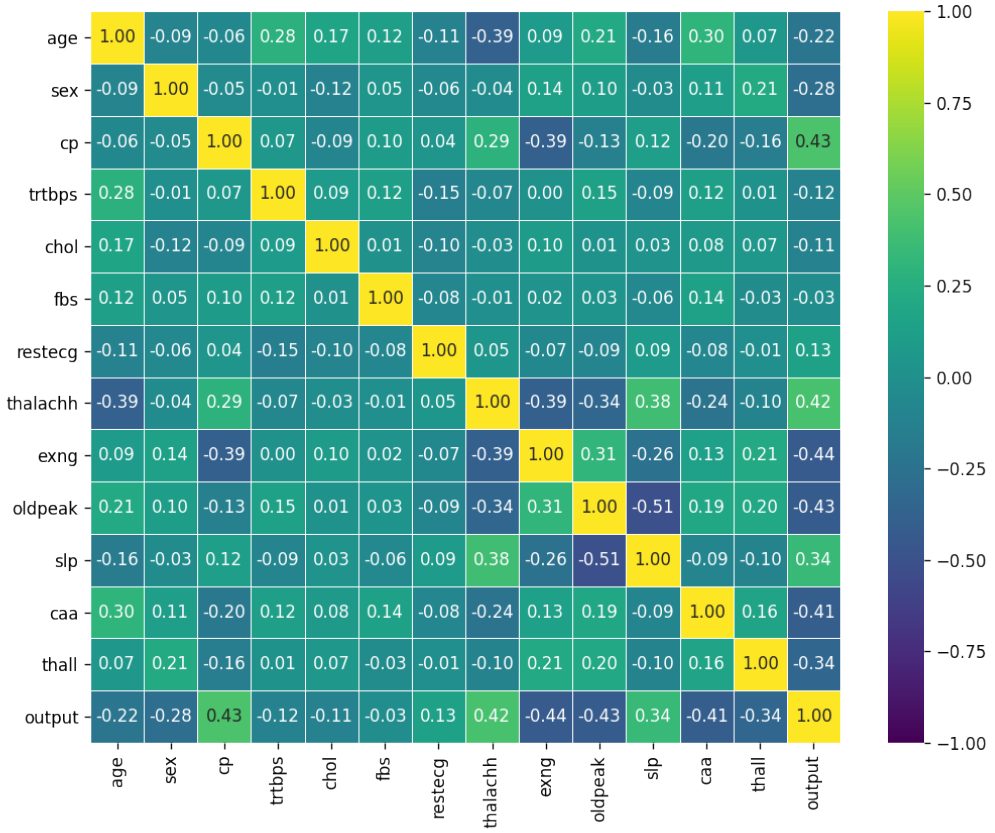


**Figure 11:** Features correlating with heart attack

However, even though the above chart is helpful for the exploration of the relationship between the features, it is neither intelligible nor visually appealing. Therefore, the heatmap was transformed, by removing some of the duplicate information; i.e., creating a triangle correlation heatmap. By sorting the output (heart attack) values in descending order, it is easier to find which feature has a higher correlation with the heart attack output. Figure 12 underlines that the type of chest pain and the maximum heart rate that was achieved have the highest positive correlation with a possible heart attack. On the contrary, exercise-induced angina and the previous peak show the lowest negative relationship with the output. The aforementioned findings are valuable for the creation of the three required hypotheses.

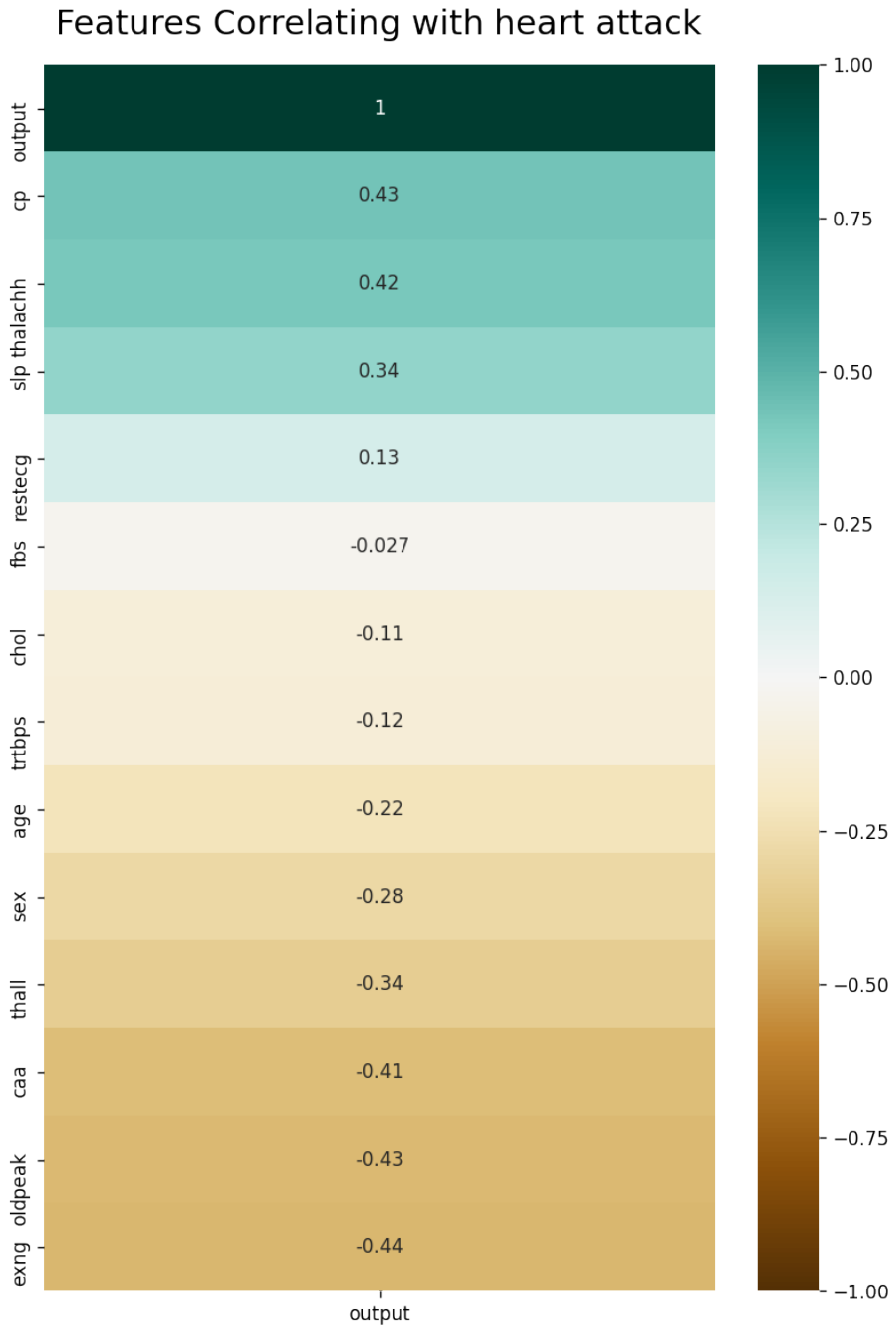


**Figure 12:** Features correlating with heart attack

## Hypotheses

In inferential statistics, the **null hypothesis** $H_o$ is a general statement or default position that there is no relationship between two measured phenomena or no association among groups. The **alternative hypothesis** $H_a$ is the hypothesis used in hypothesis testing that is contrary to the null hypothesis. It is usually taken to be that the observations are the result of a real effect.

Based on the findings from the dataset's heatmap, and the upper definitions for both $H_o$ and $H_a$, three different hypotheses can be formulated.

### First

*Question:* Does induced angina has as an effect a higher risk of a serious heart attack?

$H_o$: There is no relationship between angina and heart attack. All individuals have the same chance of developing myocardial infarction.

$H_a$: Exercise induced angina could result to higher chances of developing myocardial infarction

### Second

Question: Does the type of the chest pain an individual develops, has any link with heart attack?

$H_o$: There is no relationship between chest pain type and heart attack. All individuals have the same chance of developing myocardial infarction.

$H_a$: Chest pain type is responsible for serious heart attack.

### Third

Question: Does the maximum heart rate that was achieved relates with the output; the chance a person has to develop serious heart attack?

$H_o$: There is no relationship between the maximum heart rate achieved and heart attack. All individuals have the same chance of developing myocardial infarction.

$H_a$: Maximum heart rate (thalachh) associates with the heart attack.

### Hypothesis testing

For the hypothesis testing, the third scenario was preferred. To find which hypothesis is appropriate to choose, a t-test was employed, by using the *stats.ttest_rel* function from the *scipy* package. A t-test is a

type of inferential statistic which is used to determine if there is a significant difference between the means of two groups that may be related to certain features. A T-test is used as a hypothesis testing tool, which allows testing of an assumption applicable to a population.

```
statistic, p_value = stats.ttest_rel(heart['thalachh'],
                                     heart['output'])
print('p-value: ',p_value)
```

```
p-value:  1.18980325468209e-250
```

Assuming that the significance level $(a)$ is 0.01, then $p < a$. Therefore, we can reject the null hypothesis. So, by accepting the alternative one, the maximum heart rate has a positive correlation with serious myocardial infarction; the higher the maximum rate, the higher the chance of serious heart implications.

## Conclusion

In this report, a simple exploratory analysis of the heart attack risk dataset was conducted. After some feature engineering of the acquired dataset, and the removal of duplicate and outlier values, some numerical variables were explored (i.e., sex and age) by creating the corresponding plots (Figure 9 and 10 respectfully). A correlation heatmap was then created to find the most appropriate hypothesis. Following the statistical t-test results, the alternative hypothesis was accepted, that the chance of a heart attack is proportional to the maximum heart rate of the individual.

Overall, the dataset didn't require a lot of change and is of good quality. Nonetheless, it would be beneficial if there were more than 302 observations. Further suggestions for improving the quality of this dataset is to add features that are responsible for myocardial infarctions, like the diet, smoking information of the patient, and type of diabetes (if eligible).

### Next Steps

The next step would be a more thoroughly exploratory data analysis, by investigating and creating plots for other variables like the type of chest pain and the previous peak. After the EDA is the creation of training and test splits for establishing a machine learning (ML) logistic regression model.

## References

Martini, Frederic. 2004. *Fundamentals of Anatomy & Physiology / Frederic h. Martini with William c. Ober ... [Et Al.].* 6th ed. San Francisco, Calif.: Benjamin Cummings.