# Heart failure prediction based on clinical features

Žiga Kljun

10/8/2021

## Contents

# Introduction

This report is part of a Capstone assignment from *"HarvardX Profesional Certificate Data Science Program"*. Goal of this assignment is to choose our own project, where we will find the data and perform a machine learning task. I decided to develop prediction model for heart failure.

*Cardiovascular diseases (CVD)* kills annually approximately 18 million people and globally represent number one cause for death - 31% of all deaths worldwide. Four out of five cardiovascular deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. *Heart failure (HF)* occurs when the heart cannot pump enough blood to meet the needs of the body. With modern technology electronics, we can measure patients quantify symptoms, body features, and clinical laboratory test values, which can be used to perform biostatistics analysis that can show different patterns and correlations otherwise undetectable by doctors. Early detection of possible heart attack is crucial for people with cardiovascular disease or who are at high cardiovascular risk. This is where machine learning can help. Machine learning model prediction allows us to make highly accurate guesses, based on historical data. We know many different machine learning algorithms for predictions, some of them will be use in our project.

In our project we will use *Heart Failure Prediction Dataset*, provided by Federico Soriano Palacios on Kaggle. He created dataset by combining five different datasets, already publicly available from UCI Machine Learning Repository. With combination of 11 different features, he created the largest heart disease dataset available for public research purposes. Dataset contain information from the following locations:

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations

Total: 1190 observations

Duplicated: 272 observations

We will first analyze the data to gain a better understandings. We will try to discover some interesting insights as well as validate some health assumptions After that, we will prepare the data to be ready for machine learning process (standardization, feature engineering etc.) and split it by the 90/10 rule. We will use 90% of the data for training our model and 10% for its validation. We will also test different alghoritms and try to optimize them, in order to get the best results possible. We will validate our results with confusion matrix, where we will try to achieve the best accuracy.

# Methods/Analysis

In this section, I will present my development process and methods used. It will be split into two major parts - *Exploratory data analysis* and *Model development*. In the first part I will focus on understanding our dataset, and in the second part I will focus on developing the best model for predicting heart failure.

## Exploratory data analysis

Prior to model development, it is important, that we understand the problem and our data. In this section, I will analyze the dataset, what features are available, what are theirs' distribution, how are they correlated with eachother etc. After this section, we should have a good overview of our data which will be crucial for the next phase - model development.

### Basic information about the dataset

To start with, we need to understand, what is our data structure. As mentioned in the introduction, we have 11 common attributes which can be used to predict heart failure, which is presented in our dataset as 12th column - *HeartDisease*. It can have two values - "1" for heart disease and "0" for normal. Other columns are:

- Age: age of the patient [years]
- Sex: sex of the patient [M: Male, F: Female]
- ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
- RestingBP: resting blood pressure [mm Hg]
- Cholesterol: serum cholesterol [mm/dl]
- FastingBS: fasting blood sugar [1: if FastingBS $>$ 120 mg/dl, 0: otherwise]
- RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or * depression of $>$ 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
- MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
- ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
- Oldpeak: oldpeak = ST [Numeric value measured in depression]
- ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]

((attribute informations are copyed from the dataset description section))

In the table bellow we can see structure and column types of our dataset.

| variable | classe | first_values |
|---|---|---|
| Age | double | 40, 49, 37, 48, 54, 39 |
| Sex | character | M, F, M, F, M, M |
| ChestPainType | character | ATA, NAP, ATA, ASY, NAP, NAP |
| RestingBP | double | 140, 160, 130, 138, 150, 120 |
| Cholesterol | double | 289, 180, 283, 214, 195, 339 |
| FastingBS | double | 0, 0, 0, 0, 0, 0 |
| RestingECG | character | Normal, Normal, ST, Normal, Normal, Normal |
| MaxHR | double | 172, 156, 98, 108, 122, 170 |
| ExerciseAngina | character | N, N, N, Y, N, N |
| Oldpeak | double | 0, 1, 0, 1.5, 0, 0 |
| ST_Slope | character | Up, Flat, Up, Flat, Up, Up |
| HeartDisease | double | 0, 1, 0, 1, 0, 0 |

In the dataset we have 918 rows and 12 columns. In the table bellow we can see preview of the first five rows.

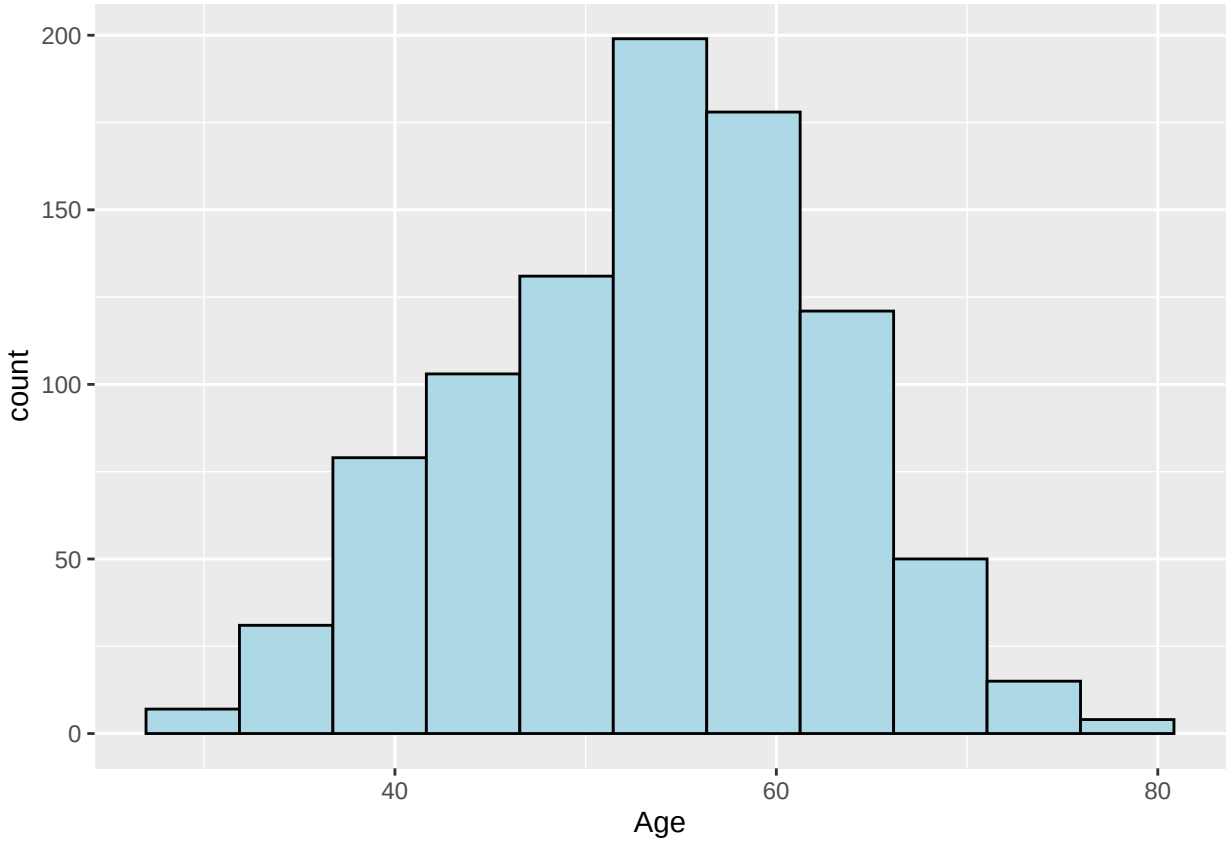
| Age | Sex | ChestPainType | RestingBP | Cholesterol | FastingBS | RestingECG | MaxHR |
|---|---|---|---|---|---|---|---|
| 40 | M | ATA | 140 | 289 | 0 | Normal | 172 |
| 49 | F | NAP | 160 | 180 | 0 | Normal | 156 |
| 37 | M | ATA | 130 | 283 | 0 | ST | 98 |
| 48 | F | ASY | 138 | 214 | 0 | Normal | 108 |
| 54 | M | NAP | 150 | 195 | 0 | Normal | 122 |

| ExerciseAngina | Oldpeak | ST_Slope | HeartDisease |
|---|---|---|---|
| N | 0.0 | Up | 0 |
| N | 1.0 | Flat | 1 |
| N | 0.0 | Up | 0 |
| Y | 1.5 | Flat | 1 |
| N | 0.0 | Up | 0 |

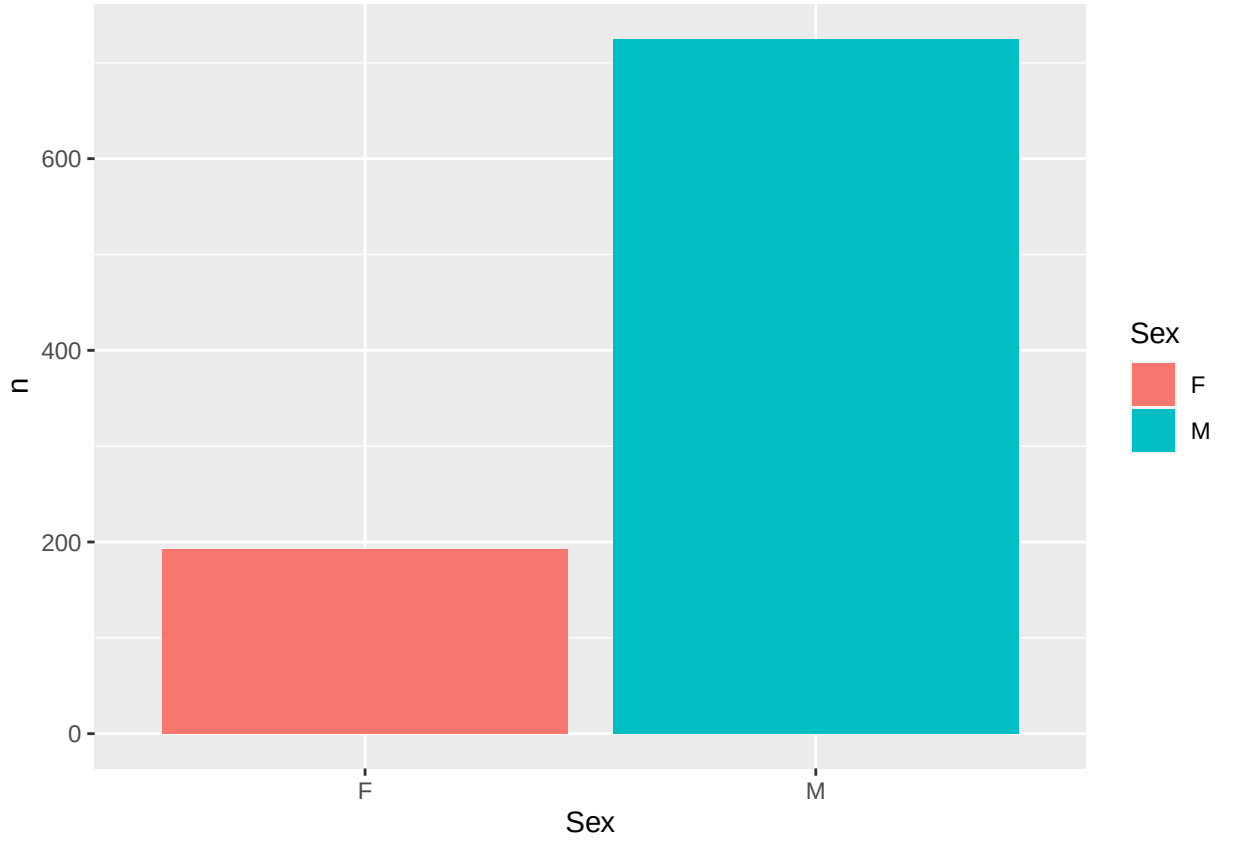
**Attribute Distribution**

We continue our exploration with looking at attribute distributions. The distribution of a statistical dataset is the spread of the data which shows all possible values or intervals of the data and how they occur. Sampling distributions are important for statistics because we need to collect the sample and estimate the parameters of the population distribution. Hence distribution is necessary to make inferences about the overall population.

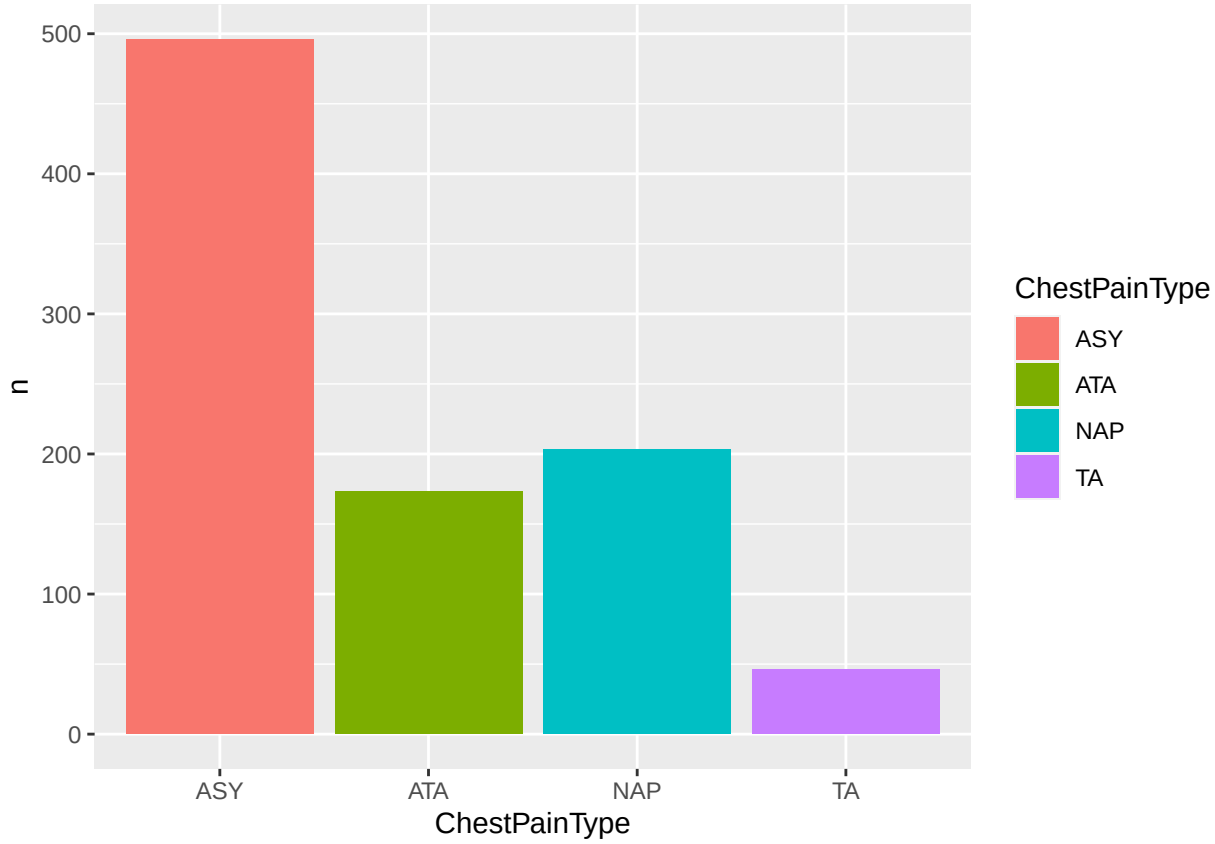
**Age Distribution** We will start with age of our patients which vary from 28 to 77. On the plot we can see, that the data is almost normally distributed, skewed a little to the left. The average value is 53.51 and the median is a little higher at 54 years of age. Our patient are on average a bit older than is the average age of the total population, but that makes sense, since the younger people are on average healthier, thus less often cardiovascular patients.



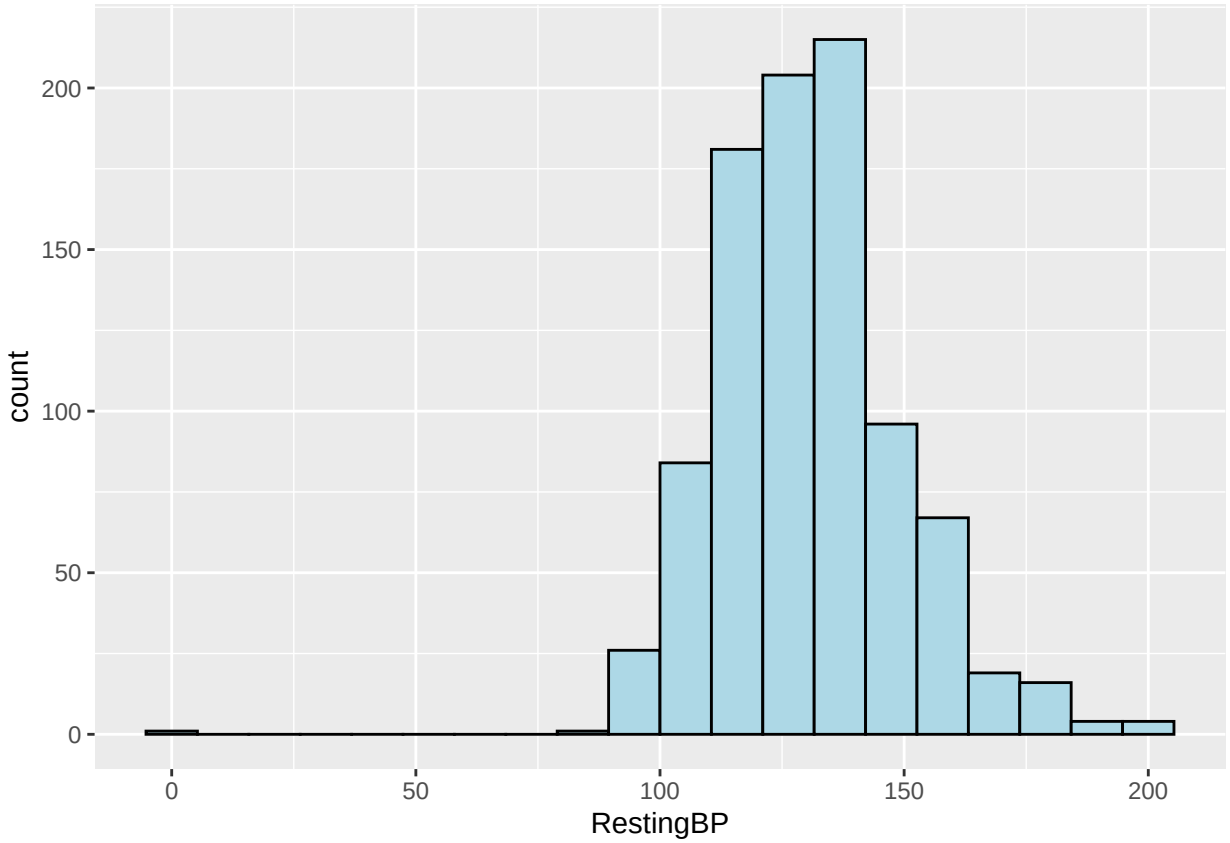
**Sex Distribution** Looking at the sex distribution, we noticed, that more than 75% of our patients are males. Later it will be interesting to see, if males are also more likely to have a heart failure.



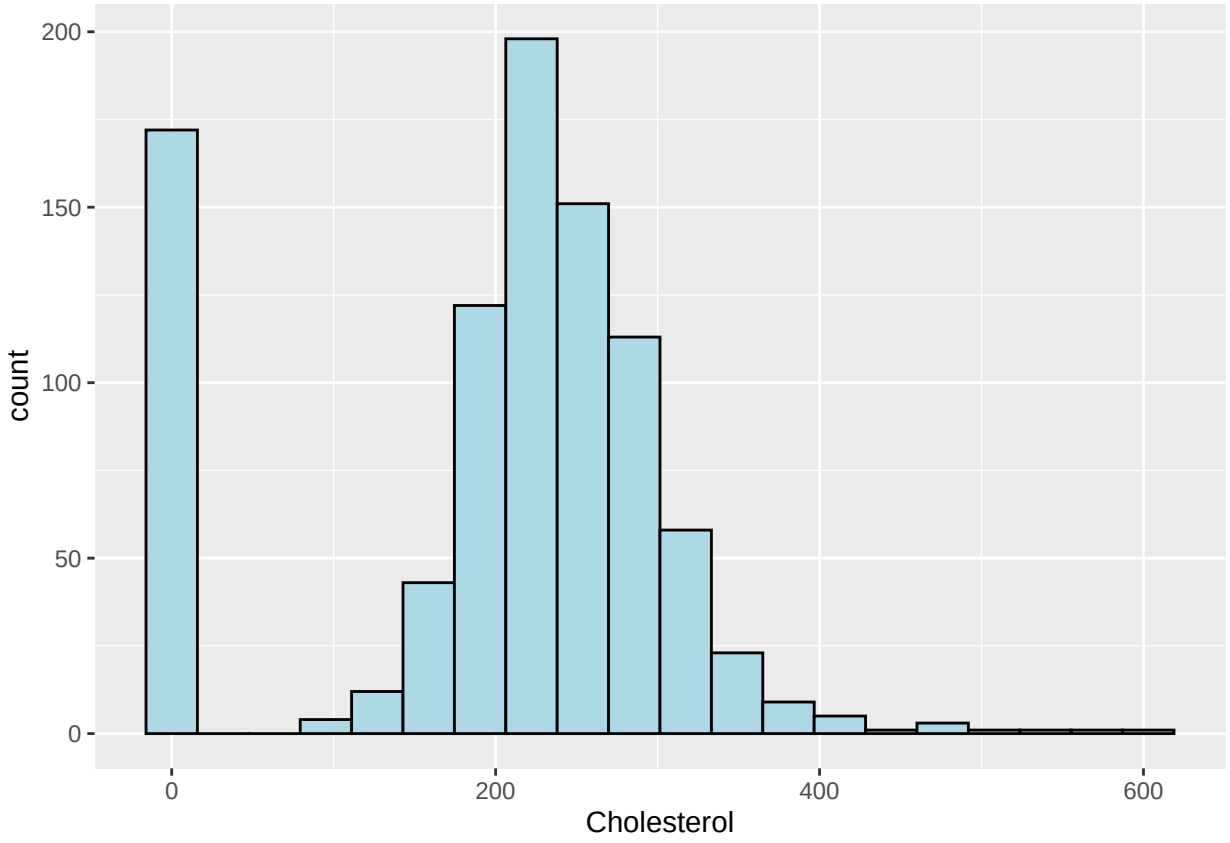
**Chest Pain Type Distribution** In our dataset, patient are separated by three different types of chest pain: typical angina (TA), atypical angina (ATA), non-anginal pain (NAP) and asymptomatic (ASP), which means, that they are not showing any symptoms. As expected, most of our patients are labeled as *Asymptomatic*. Atypical Angina and Non-Anginal pain are almost equally often where typical Angina is being the most rare.



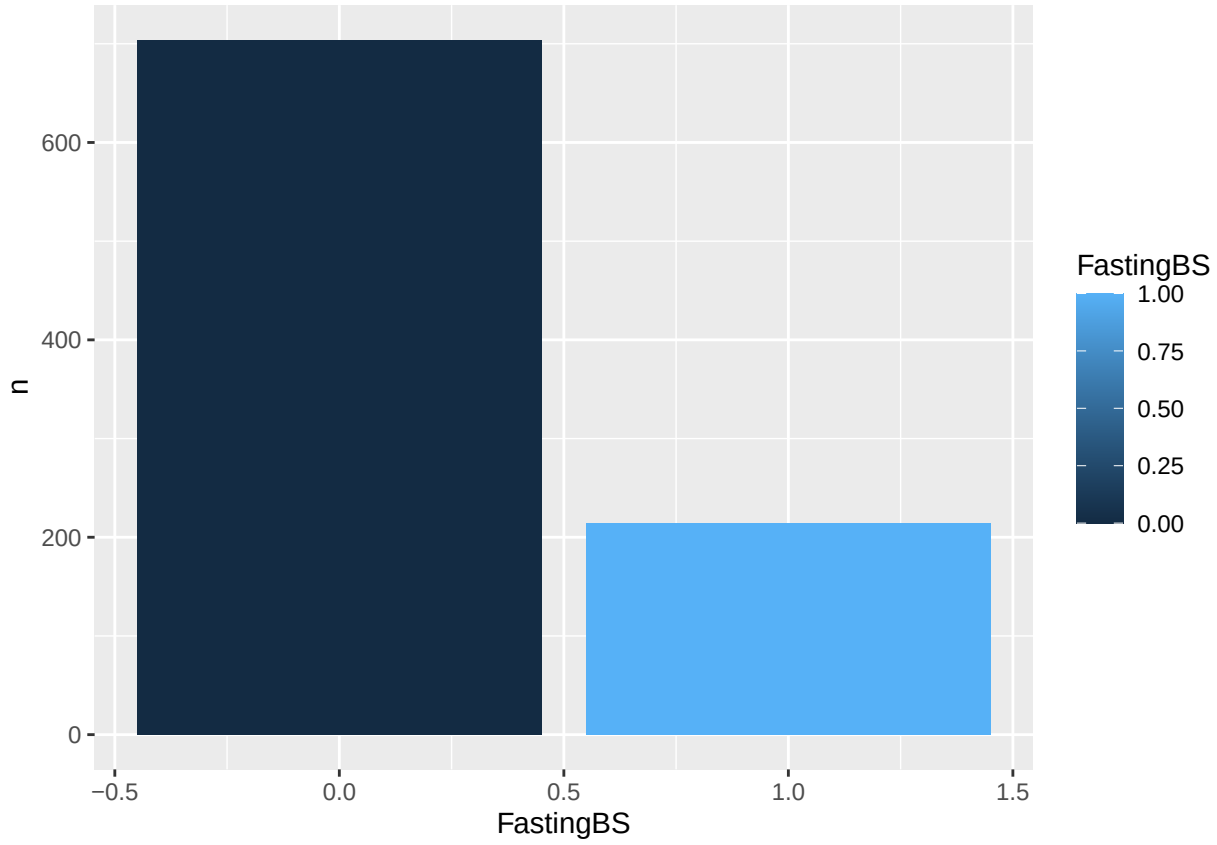
**Resting Blood Pressure Distribution** Resting blood pressure is a measure of the force that your heart uses to pump blood around your body when you are resting. It is measured in millimetres of mercury (mm Hg). Similar as age, resting blood pressure is also almost normally distributed - this time skewed a little into the right. Average value is 132.40 mm Hg with the standard deviation of 18.51 mm Hg, while the median is 130 mm Hg.



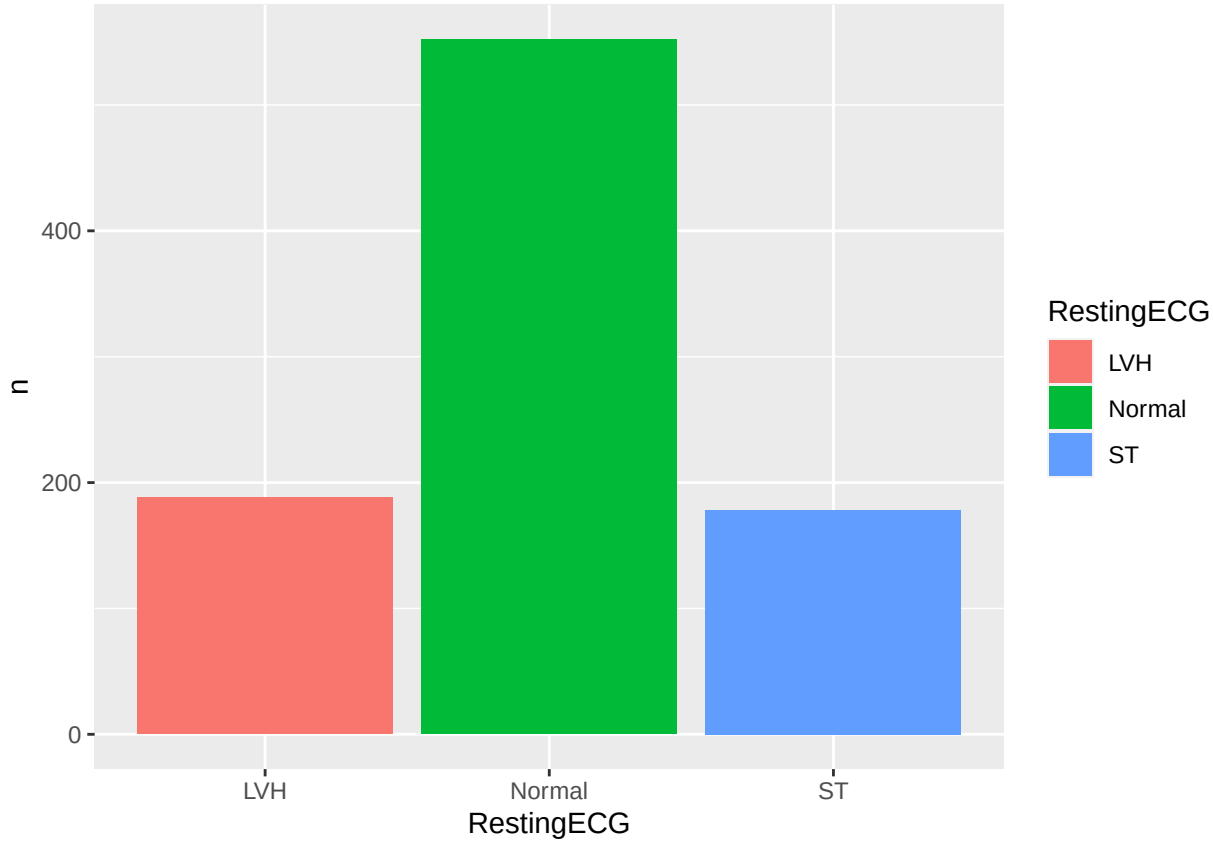
**Cholesterol Distribution** Cholesterol is a waxy, fat-like substance and is present in all the cells of human body. We need it to generate substance for food digesting, vitamin D etc. But when the level of cholesterol rises, so does the risk of cardiovascular diseases, such as heart disease and stroke. Normally, cholesterol values under 200 mm/dl are treated as healthy. The distribution of cholesterol in our data is a bit more interesting than with our previous attributes. We could treat it as two different groups - patients who does not have problems with the cholesterol and the ones, who does. First group represents While the first group represent 18% of our patients, the patients in second group are normally distributed. Medium for cholesterol is 223 mm/dl, while average is "just" 198.80 mm/dl with high standard deviation of 109.38 mm/dl. If we ignore the patients without cholesterol problems, these numbers of mean and medium rise significantly, especially the average which is now 245 mm/dl. Even standard deviation fell to 59.2 mm/dl. Median is 237 mm/dl which is showing, that most of our patients have rather high levels of cholesterol.



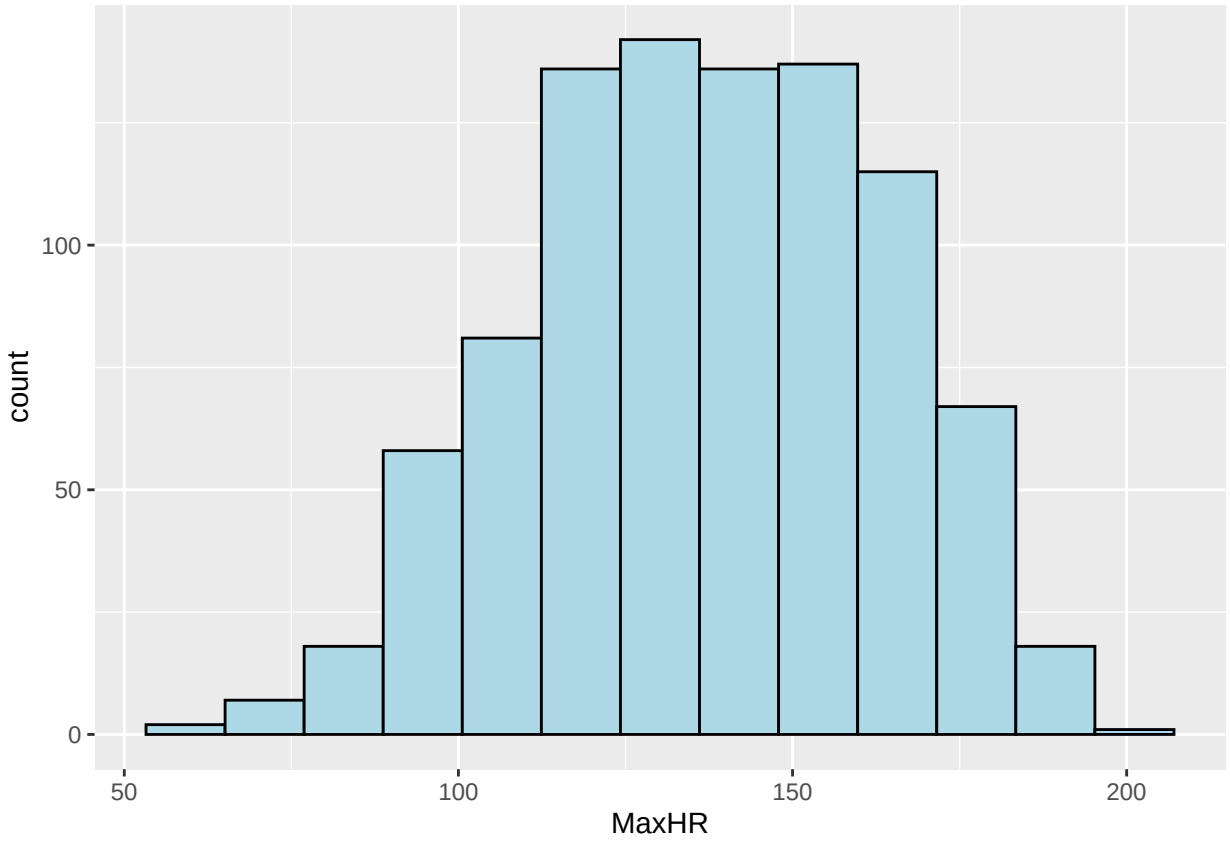
**Fasting Blood Sugar Pressure Distribution** Fasting Blood Sugar Pressure is used to detect diabetes. A blood sample will be taken after an overnight fast. A fasting blood sugar level less than 100 mg/dl is normal. A fasting blood sugar level from 100 to 125 mg/dl is considered prediabetes. If it's 126 mg/dl or higher on two separate tests, you have diabetes. In our data, the value for fasting blood sugar is "1", if it is more than 120 mg/dl and otherwise "0". As expected, patients with less than 120 mg/dl are in the majority. However, I was quite surprise, that the number of patients with 120 mg/dl or higher represent almost one quarter of all patients. However it is probably due to the fact, that we are looking at people who already has some kind of a problem.



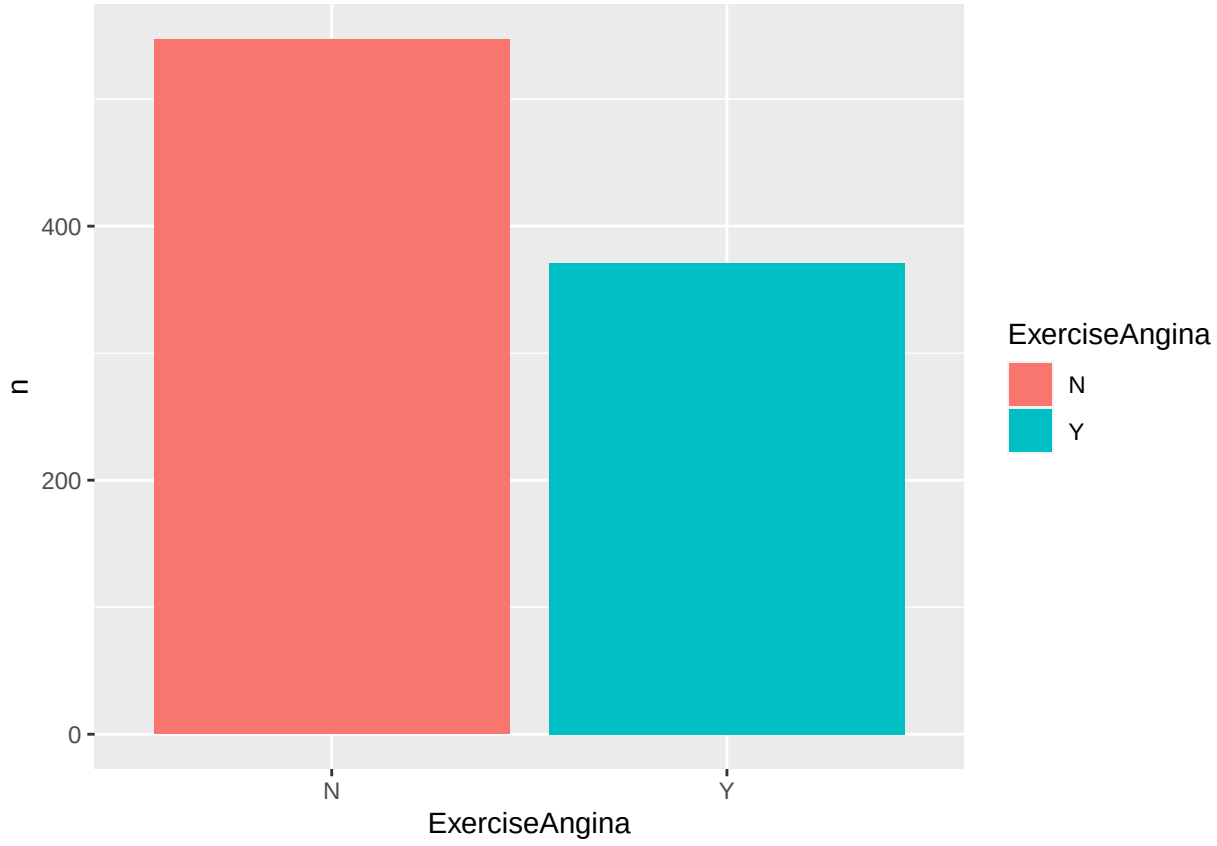
**Resting Electrocardiogram (ECG) Results Distribution** The resting Electrocardiogram (ECG) is measuring electrical activity of the heart. Looking at the resting ECG results in our dataset we can see, that as expected, the most patient have Normal resting ECG while around 20% od patients are having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of $> 0.05$ mV) as well as 20% are showing probable or definite left ventricular hypertrophy by Estes' criteria.



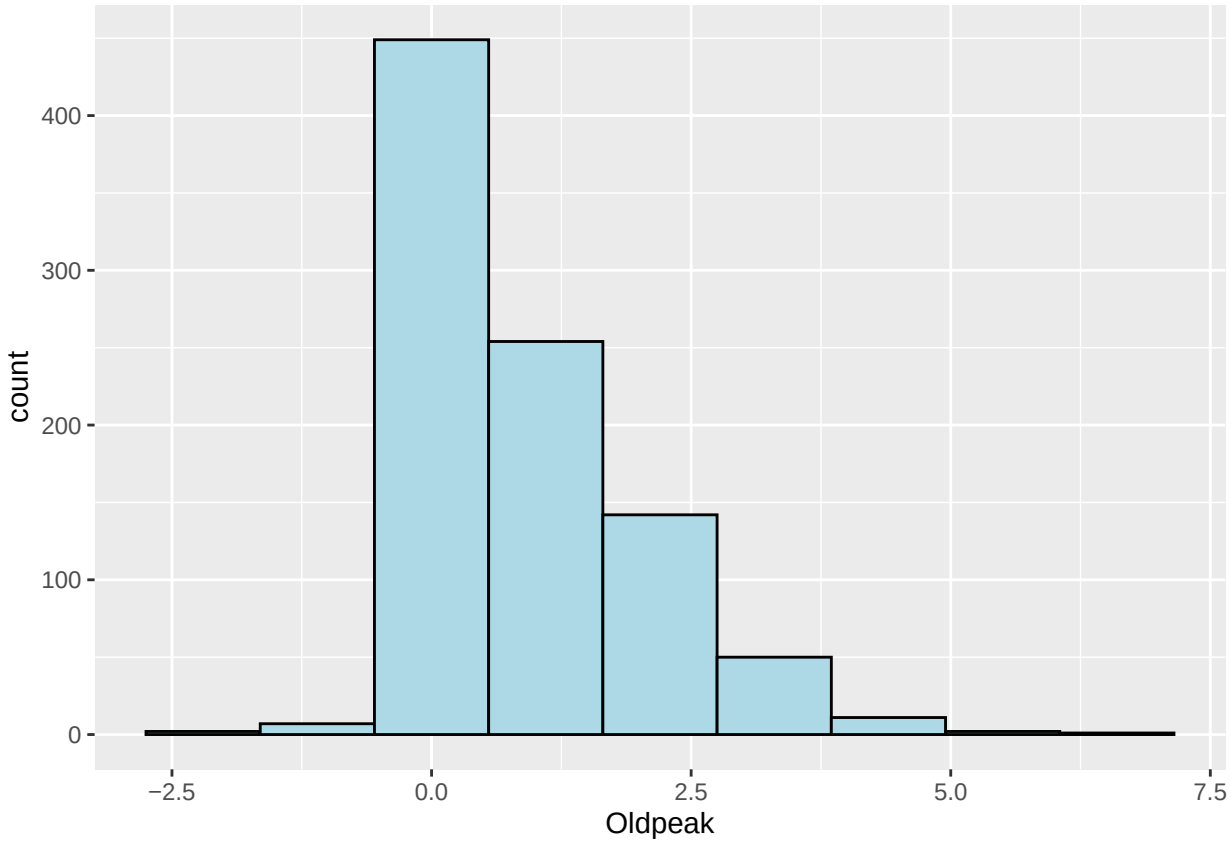
**Maximum Heart Rate Achieved Distribution** Maximum Heart Rate Achieved Distribution represent the highest heart rate achieved when testing and it varies between 60 and 202. As most of our continous attributes, maximum heart rate achieved is also almost perfectly normally distributed. It is just a little skewed to the left. Average value is 136.81 with the standard deviation of 25.46, while median is 138



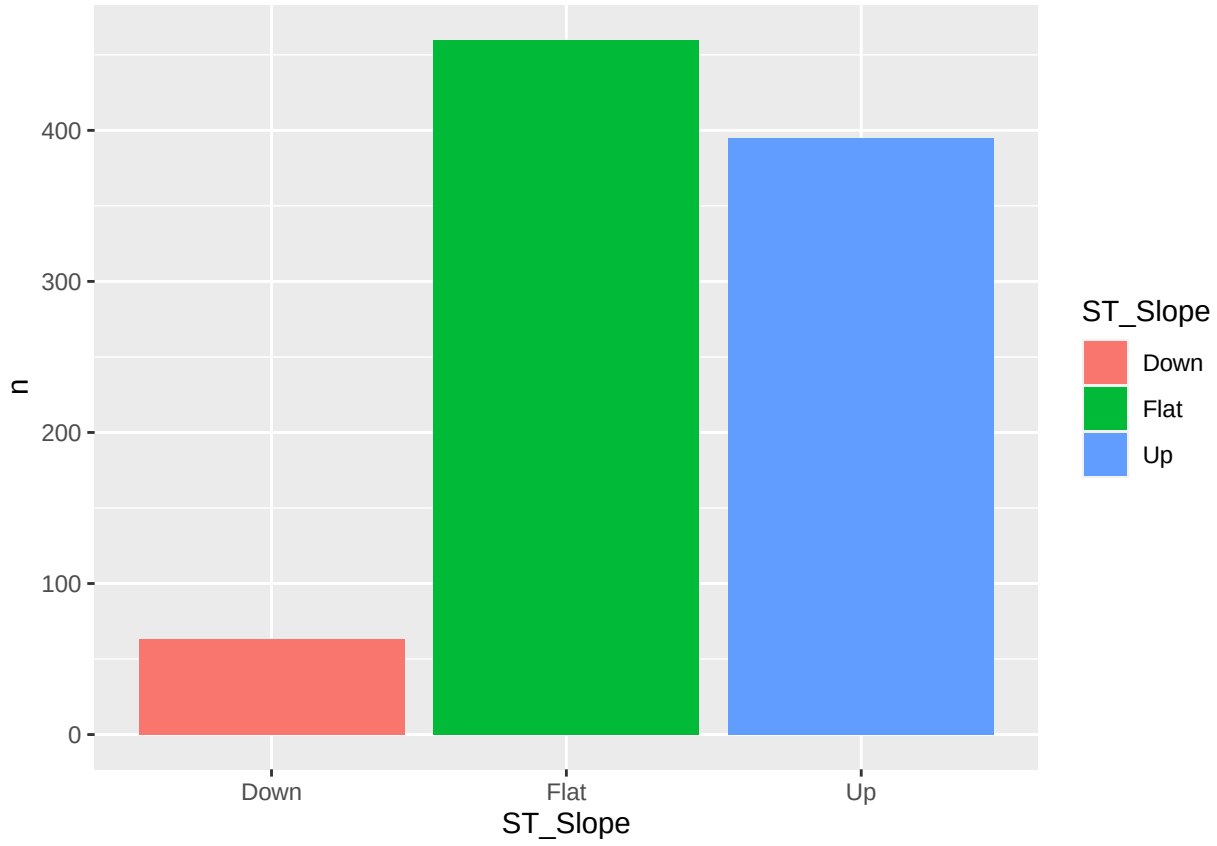
**Exercise-Induced Angina Distribution** Exercise-induced angina is a common complaint of cardiac patients, particularly when exercising in the cold. It is also quite high in our dataset, since it is present with 40% of patients.



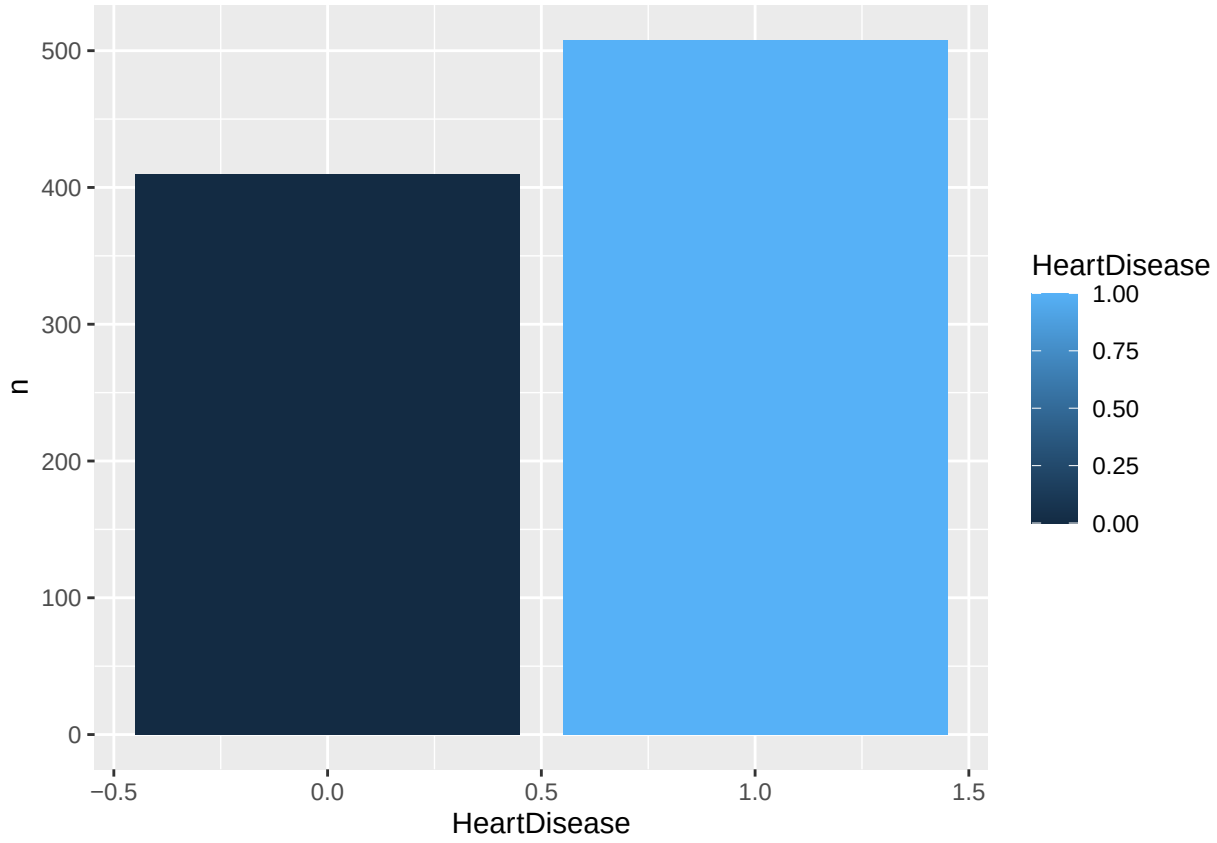
**Old Peak Distribution** Old peak tells us how is ST depression induced by exercise relative to rest. In our data it varies from -2.6 to 6.2. Its most frequent values are around "0", but due to being highly right skewed, average is 0.89 with standard deviation of 1.07 and median 0.6.



**The Slope of the Peak Exercise ST Segment Distribution** As expected, most patients have the slope of the peak exercise ST segment flat. But what I did not expected is, that over 40% of patients are upsloping while just under 7% of patients have downsloping.

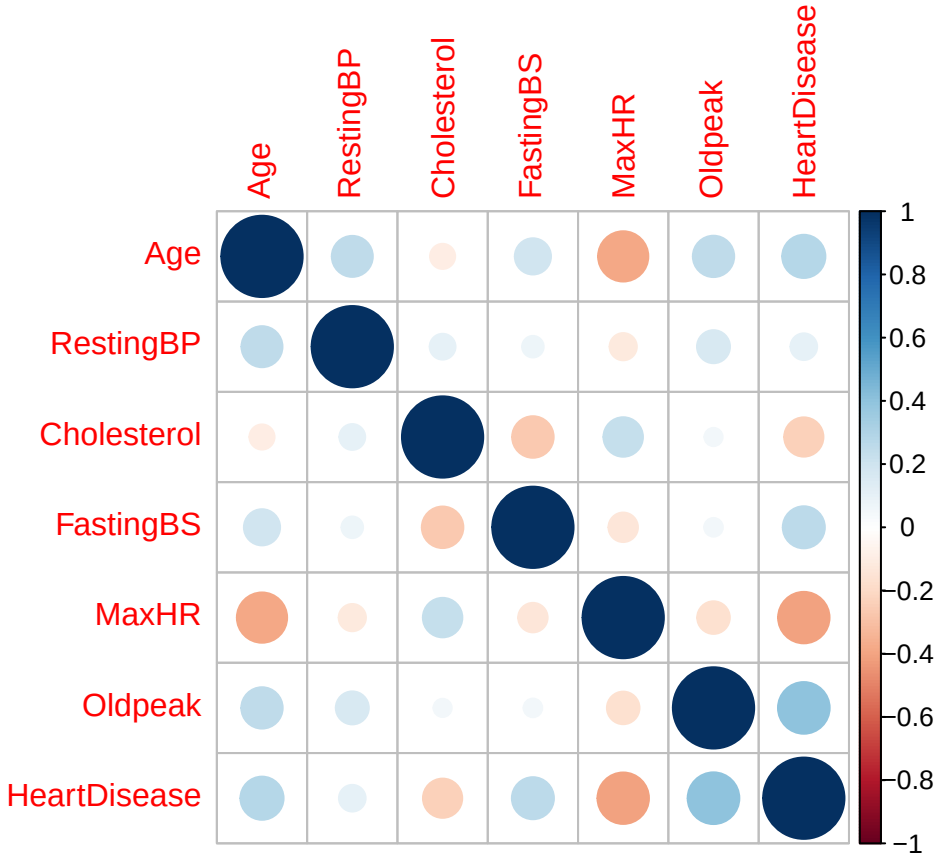


**Heart Disease** Finally we came to the column, we are trying to predict. As we can see, both values - disease and no disease are quite equally distributed, but there are still 55% of patients who has heart disease. Later in our analysis, we will try to discover which attributes have the greatest impact on heart disease.



### Correlations

Correlation is a statistical measure that shows to what extend are two variables related. Correlations are useful for describing simple relationships among the data. Its value ranges from -1 to 1, where 1 represent very strong positive correlation, -1 very strong negative and 0 non-existing one. However, we need to consider, that correlation does not necessarily imply causation. Two variables could have high correlation, even though there is not direct causation between them. In the plot bellow, you can see correlations between our attributes. We can see, that in general, attributes has greater correlation with heart disease, than with each other, which is good for our model. From the plot we can also see, that the strongest positive correlation of heart disease is with old peak, while the strongest negative correlation is with maximum heart rate achieved, which is something we expected.



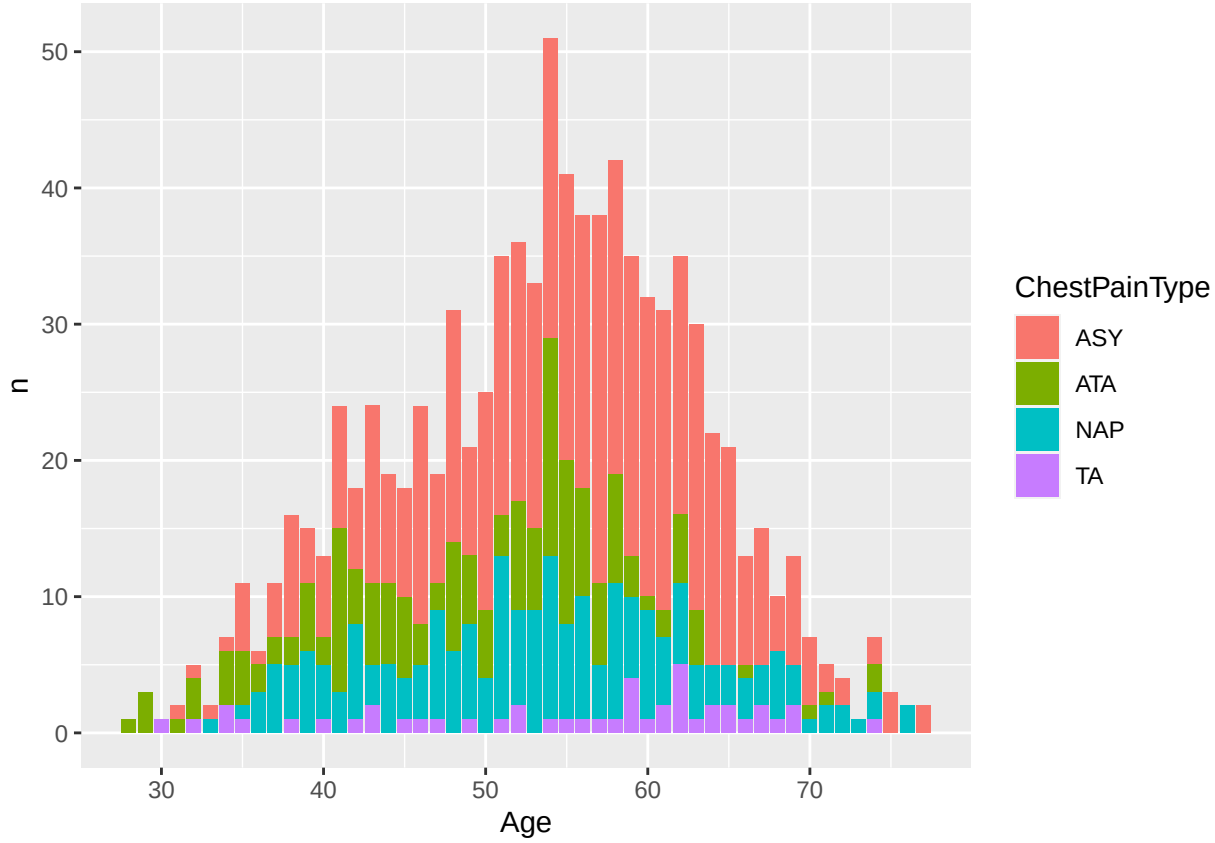

### Testing assumptions

Based on our previous knowledge and correlation plot, I decided to set a few assumptions about the data which I would like to validate whether they are true or false. Some of them are important directly for our modeling and some of them are just out of my curiosity. Assumptions are:

- Age has an impact on the chest pain type - the higher the age, the lower percentage of patients with ASY
- Age has an impact on the resting blood pressure - the higher the age, the higher the resting blood pressure
- Age has an impact on the cholesterol - the higher the age, the higher the cholesterol
- Age has an impact on the maximum heart rate achieved - the higher the age, the lower maximum heart rate achieved

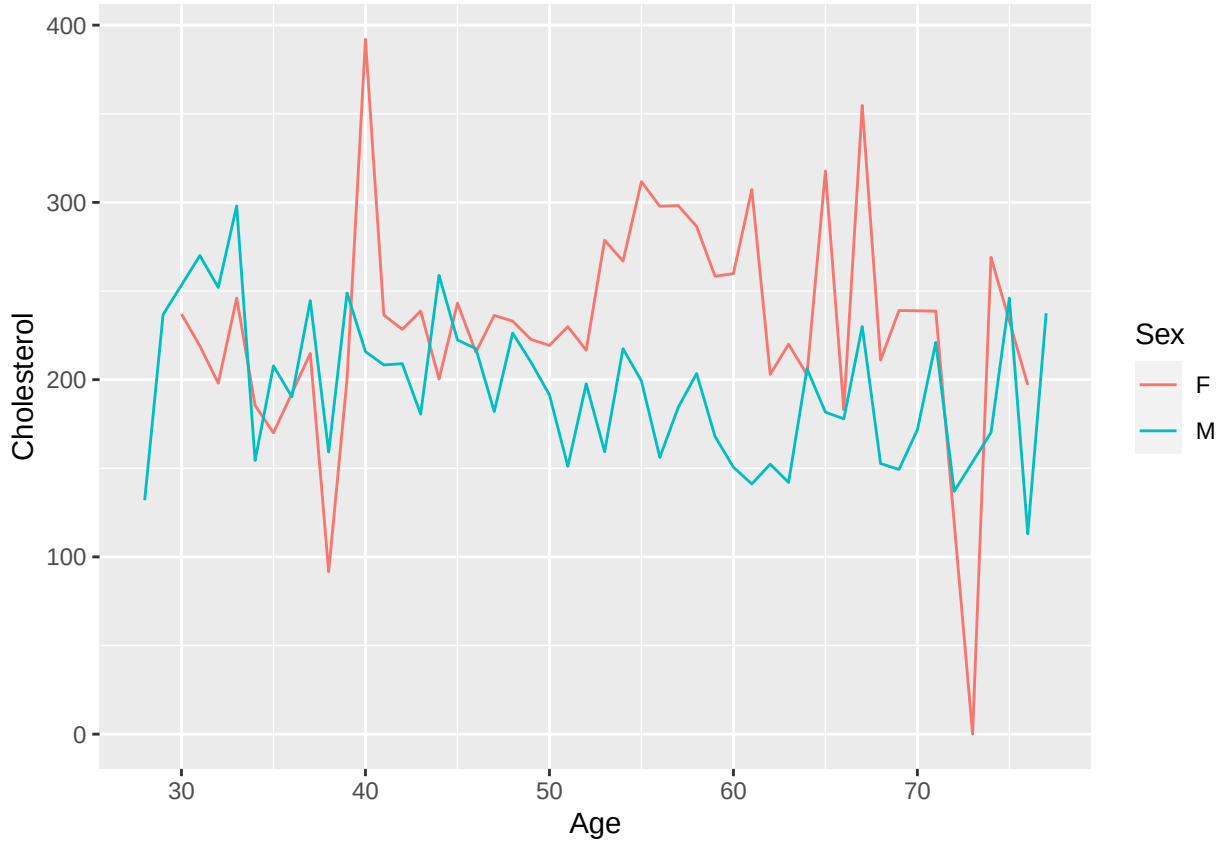
- Age has an impact on the slope of the peak exercise - the higher the age, the higher percentage of downslope, the lower percentage of upslope and diamond-shaped flatslope
- Age has an impact on the heart disease - the higher the age, the higher percentage of patients with heart disease
- Sex has no impact on the heart disease
- Chest pain type has an impact on the resting blood pressure - different groups are having different average resting blood pressure
- Chest pain type has an impact on the cholesterol - different groups are having different average cholesterol levels
- Maximum heart rate achieved has an impact on heart disease - the higher the maximum heart rate achieved, the lower the chance of heart disease

**Age impact on the chest pain type** Plot bellow shows how chest pain is distributed depending on age. We can see, that age has an impact. Over the time, the percentage of patients without symptoms raise and the percentage of the ones with atypical angina declines. I actually predicted the exact opposite - that the percentage of ASY will fall with age. But if we think about it it actually makes sense, since older people have in general more health problems, which means, that even though they do not have chest pain, they can still have some other issues which lead them to the hospital. We need to keep in mind, that this is not data on the whole population, but just on the patient who already has some health problems.



**Age impact on the resting blood pressure** Plot bellow shows how resting blood pressure changes depending on the age and sex. As we predicted, it is rising over time, even though there are some exception which we can attributed to lack of data in our dataset (specially on both ends of the age spectrum). Regarding sex, we can see, that the males are a bit more stable, but this is due to more male data in our dataset.

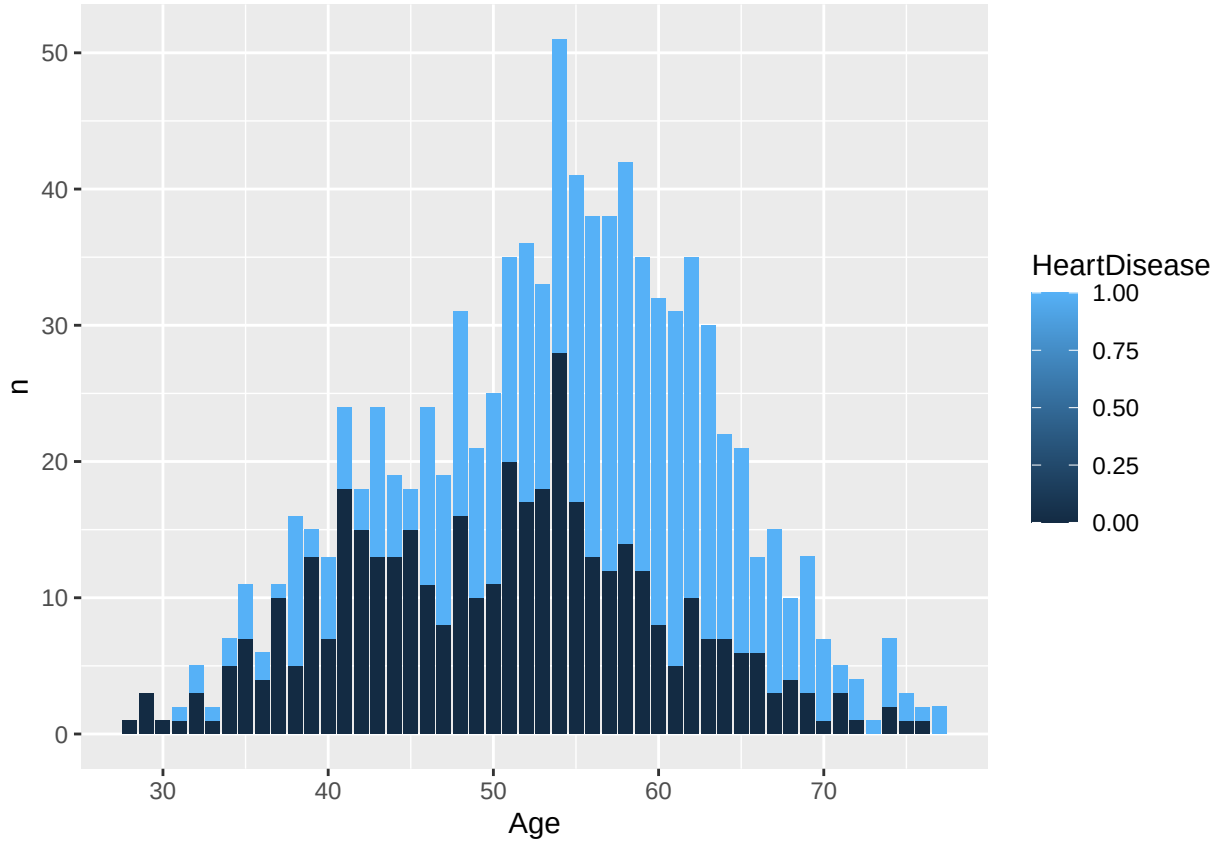
**Age impact on the cholesterol** Plot bellow shows how cholesterol changes depending on the age and sex. We predicted, that with the age, cholesterol would rise. But as we can see, this is not the case. There is close to zero correlation, as it was also seen on the correlation plot. I still wanted to see how is data actually distributed. Sex also do not have an impact, but we can see again, that the males are a bit more stable due to more data in our dataset.



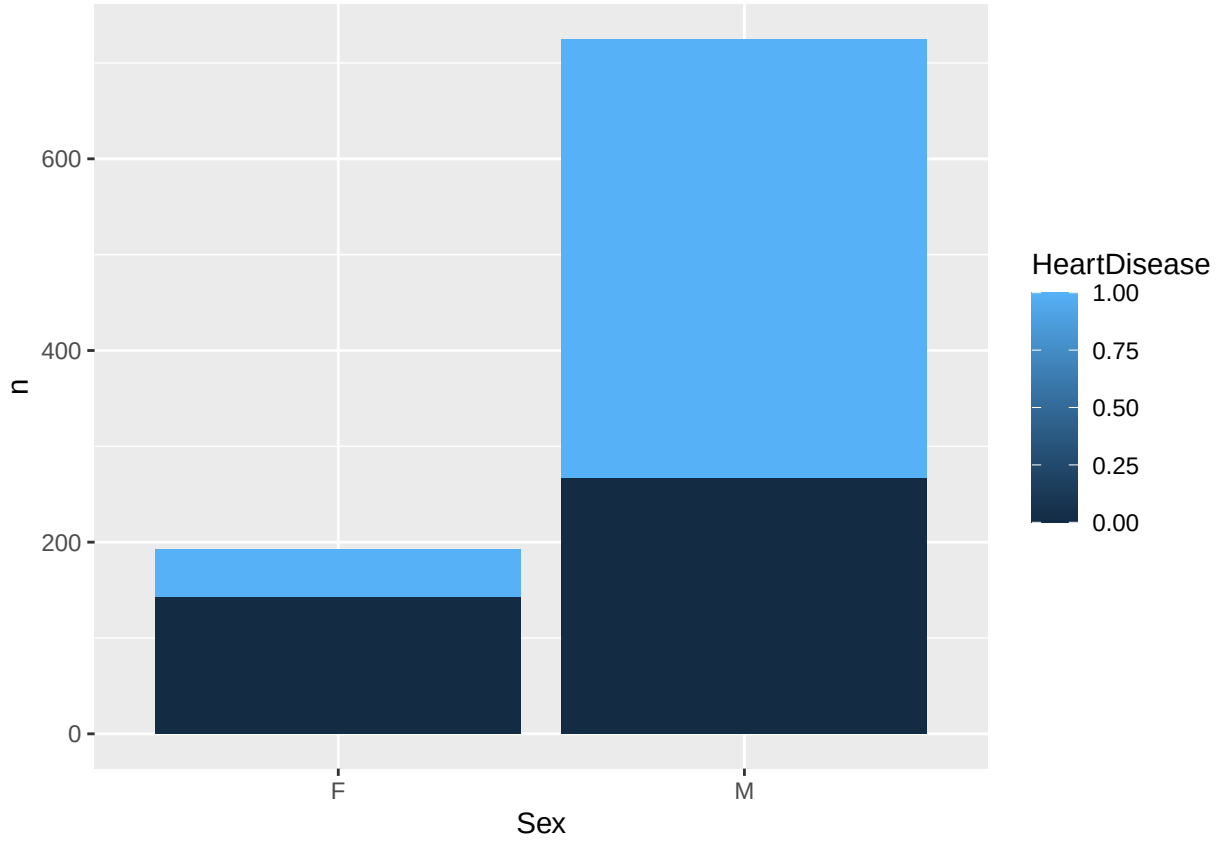
**Age impact on the maximum heart rate achieved** Plot bellow shows how maximum heart rate achieved changes depending on the age and sex. We predicted, that with the age, the maximum heart rate achieved would fall. In the our plot we can see, that this is true. At around 30 years of age the average maximum heart rate achieved was around 175 and then over the time, it fell to around 125. Sex does not have much impact.

**Age impact on slope of the peak exercise** I was expecting that when younger, most people would have upslope, which over time would be converting to flatslope. Then, when reaching higher age, the downslope would become more more frequent. Looking at the plot bellow, which is showing frequency of slope types depending on age, we can see, that our predictions were correct.

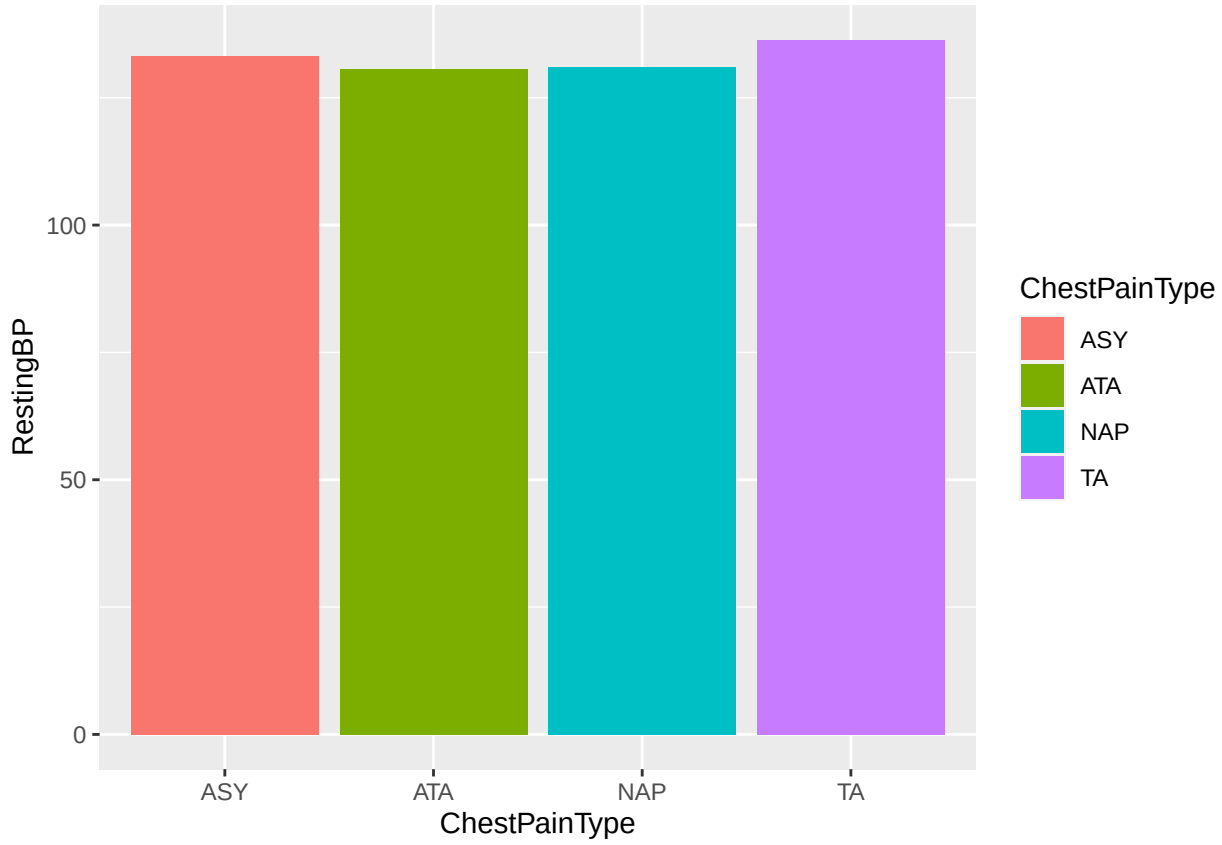
**Age impact on the heart disease** Plot bellow shows how number of heart disease changes depending on the age. We predicted, that with the age, more patients will have a heart disease. Our plot shows, that with aging, the percentage of people with heart disease is raising. It goes even that far, that most of people over 55 in our dataset have a heart disease, even though at around 30, there was barely anyone. This confirms our assumption.



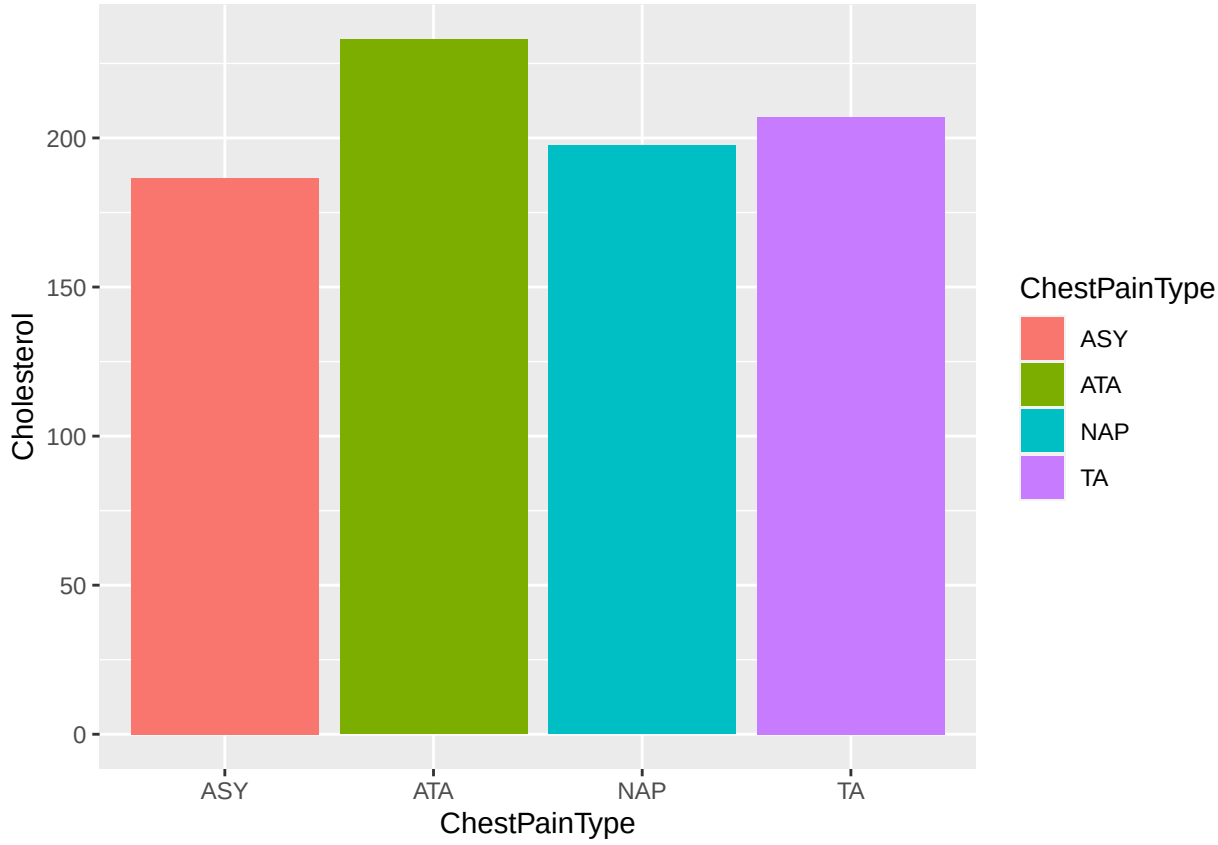
**Sex impact on the heart disease** I was curious whether sex has any impact on the heart disease, but my assumption was still, that there is not really much impact. Plot bellow is representing how many male and female patients has heart disease and how mony does not. We find out, that my assumption was wrong, since almost two thirds of male patients in our dataset have heart disease, while only quarter of female does. For me this was very surprising, so I went and read some article about heart diseases and heart attacks to validate if this is some strange effect in our data or this is more common fact in the population. Jama Internam Medicine posted in 2016, that in a study with 34,000 people (around half of them were females) from Norway between 1979 and 2012, researchers found that men were about twice as likely to have a heart attack, but for now, they do not know haw to explain this gap between sexes.



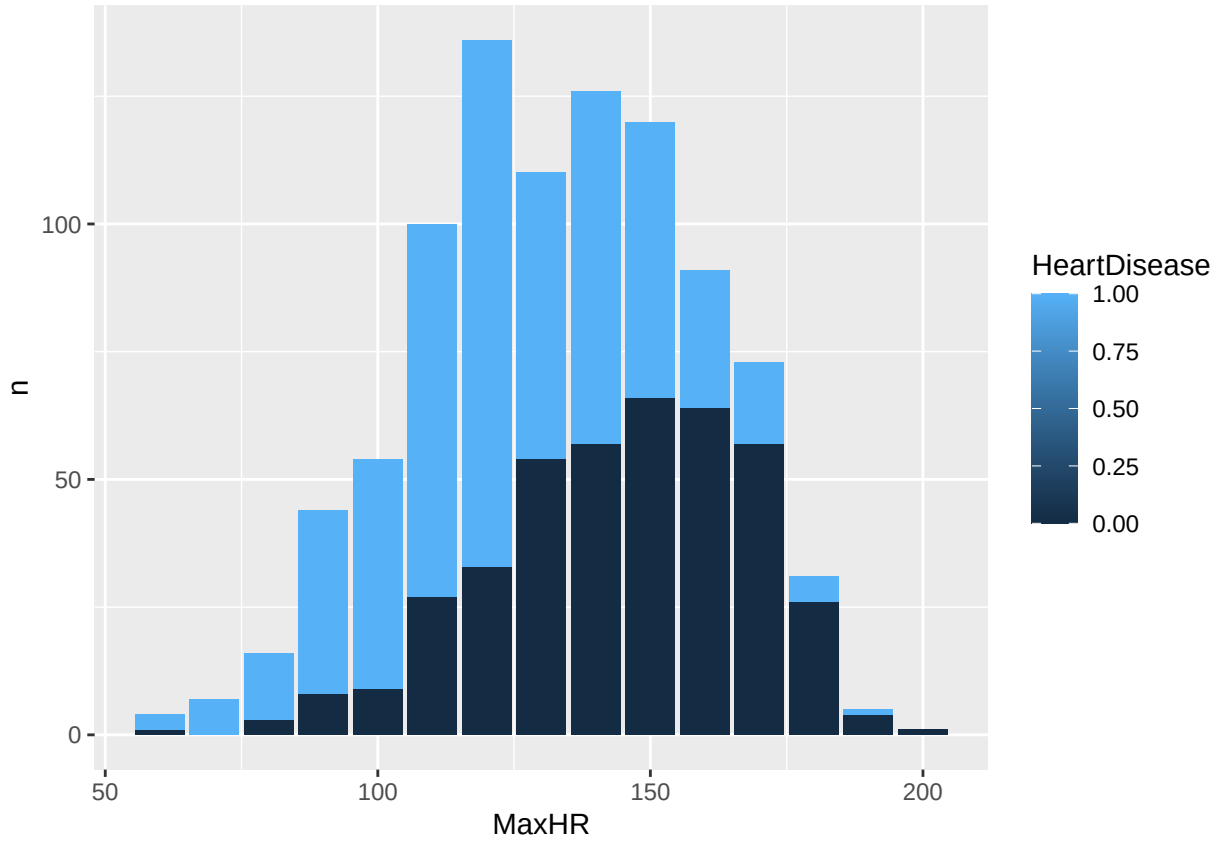
**Chest pain type impact on the resting blood pressure** My assumption was, that different chest pain types has effect on resting blood pressure, thus different types having different average resting blood pressures. However, our assumption was wrong, since average resting blood pressure is almost identical for all four chest pain types.



**Chest pain type impact on cholesterol** I predicted, that different chest pain types will have different levels of cholesterol. But in this case, it could also go the other way around, which means, that level of cholesterol have impact on chest type so there is really a question of causation. However, when looking at the plot bellow, showing cholesterol level depending on chest pain types, we can see, that atypical angina (ATA) has the highest average, while other three have quite similar. It is worth mentioning, that cholesterol levels under 200 mm/dl are healthy, which means, that ATA is showing problems, asymptomatic and non-anginal pain are healthy, while typical angina is just over 200 mm/dl.



**Maximum heart rate achieved impact on heart disease** The assumption was, that with increase of maximum heart rate achieved, the chance of heart disease will decrease. Plot bellow is showing total number of patients with heart disease versus total number of patients without depending on maximum heart rate achieved. We can se, that with lower maximum heart rate, heart disease is almost certain while with higher values of maximum heaert rate it is very rare. This confirms our assumption, that with increase of maximum heart rate achieved, the chance of heart disease decreases.



## Model development

In this section, I will focus on model development. It will cover the whole process, from preparing the data, to implementing algorithms and model validation. We will start with data standardization, following with splitting it. After that we will implement and optimize three different alghoritms - logistic regression, k-nearest neighbors and random forest. For validation we will use confusion matrix. With confusion matrix we will mainly focus on three results: sensitivity (the number of correct positive predictions divided by the total number of positives), specificity (the number of correct negative predictions divided by the total number of negatives) and most importantly accuracy (the number of all correct predictions divided by the total number of the dataset).

### Standardization

Standardization is very popular way of data cleaning, leading to a more clear, consistently defined attributes and in general better quality data. It is the process of transforming data to a common format, giving us the ability to process and analyze it. It is technique in which all the features are centered around zero and have roughly unit variance. We performed standardization on our continious variables: age, resting blood pressure, cholesterol and maximum heart rate.

| Age | RestingBP | Cholesterol | MaxHR |
|---:|---:|---:|---:|
| -1.4323590 | 0.4106850 | 0.8246208 | 1.3821748 |
| -0.4782229 | 1.4909396 | -0.1718674 | 0.7537463 |
| -1.7504044 | -0.1294423 | 0.7697682 | -1.5243071 |
| -0.5842380 | 0.3026596 | 0.1389638 | -1.1315393 |
| 0.0518527 | 0.9508123 | -0.0347360 | -0.5816643 |

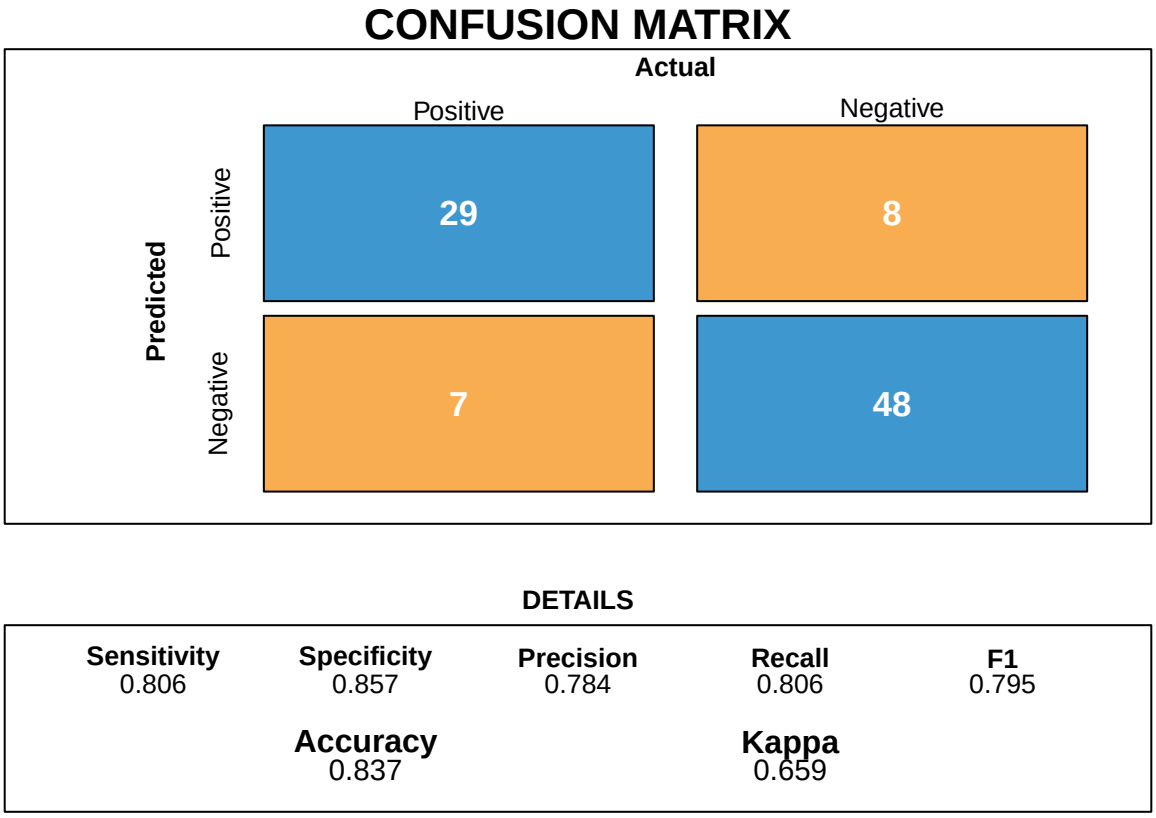
### Spliting the data

Now when we standardized our continuous variables, we can split the data into training and test sets. Training set will be use to train our models and test set will be used to test our model. This means, that our model will learn based on all the data in train set, and then will predict heart disease value in test data, based on the other attributes. We will use 90/10 approach, which means that 90% of our dataset will be used for training and 10% will be used for testing purposes.

**Logistic regression**

Logistic regression is analytical approach, which helps us predict likelihood of an event happening, thus it is very popular in predictive analytics and modeling. It uses logistic function to model dependent categorical variable (in our case it will be binary (heart disease 1 or 0)), based on independent variables (predictors). It is estimating the parameters of a logistic model. For that specific case, I believe, that the logistic regression is not robust enough, so I expect other algorithms to predict better. Nevertheless we will still create our first iteration with logistic regression, which will serve us as a baseline for future development.

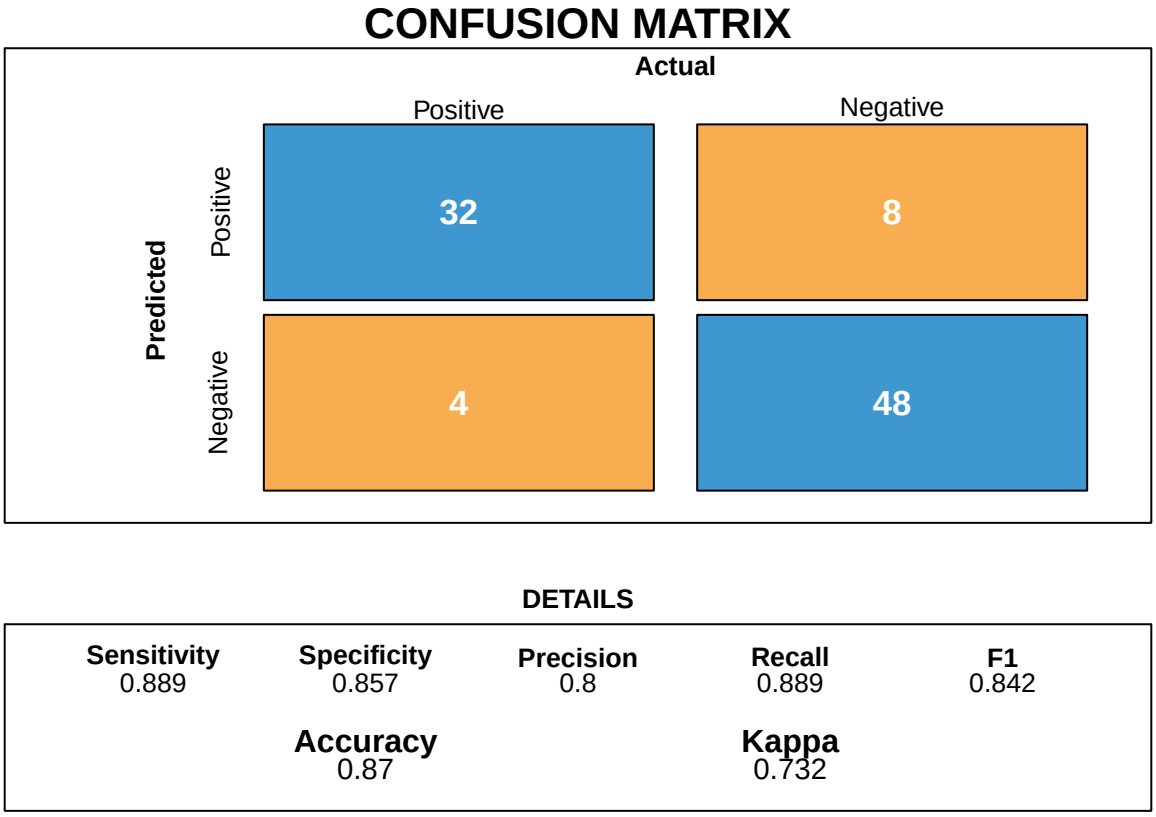
We created logistic regression model, where we used all attributes besides HearDisease as predictors, to predict HeartDisease. When running model on the test set, we assigned all values greater than the 0.5 to 1, and all values bellow that as 0. With confusion matrix, we got sensitivity 0.806, specificity 0.857 and accuracy 0.837, which is great baseline for future model development.



**K-nearest neighbours**

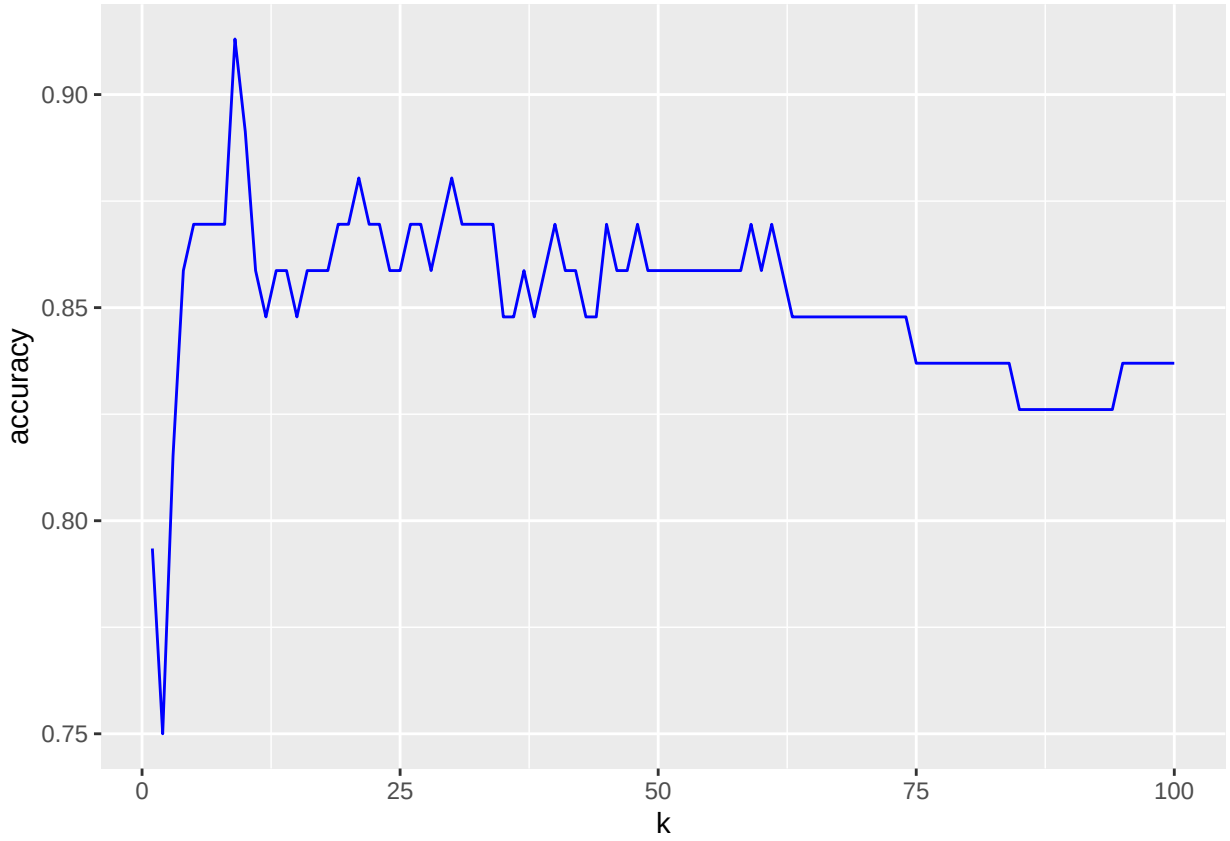
Our next alghoritms is K-nearest neighbours (kNN). KNN is one of the simplest machine learning algorithms based on Supervised Learning technique and can be used for classification use cases. It assumes the similarity between the new CASE and available cases and put the new case into the category that is most similar to the available categories. For that it uses the Euclidean distance of *K number of neighbours*. It does not learn from the data in the process of training, but it stores the data and perform an action in the process of classification.

When modeling kNN model there is however a limitation with our data - kNN is working just with non-categorical data, but in our dataset, we have five categorical attributes - sex, exercise angina, slope of the peak exercise, resting ECG and chest pain type. Prior algorithm implementation, we need to adjust our data in a way, that we will have just the numerical data. To achieve this, we will code our categorical data into numeric one when it is possible (when values could be arranged like "Low"s and "High"s). In the first four cases there is no problem, but in the case of ChestPainType, it would not make sense to code values into 1, 2, 3 and 4, since they cannot be really arranged. The problem could be solved with addition of N-dimensional space such as (1,0,…),(0,1,…) but for know, we will just skip this column since it is the only "problematic" one and it would complicated the model a lot. We will rather focus on improving other aspects of the model. After adjusting attributes, we implemented kNN algorithms with default parameters and got sensitivity 0.889, specificity 0.857 and accuracy of 0.87, which is already an improvement over the logistic regression.



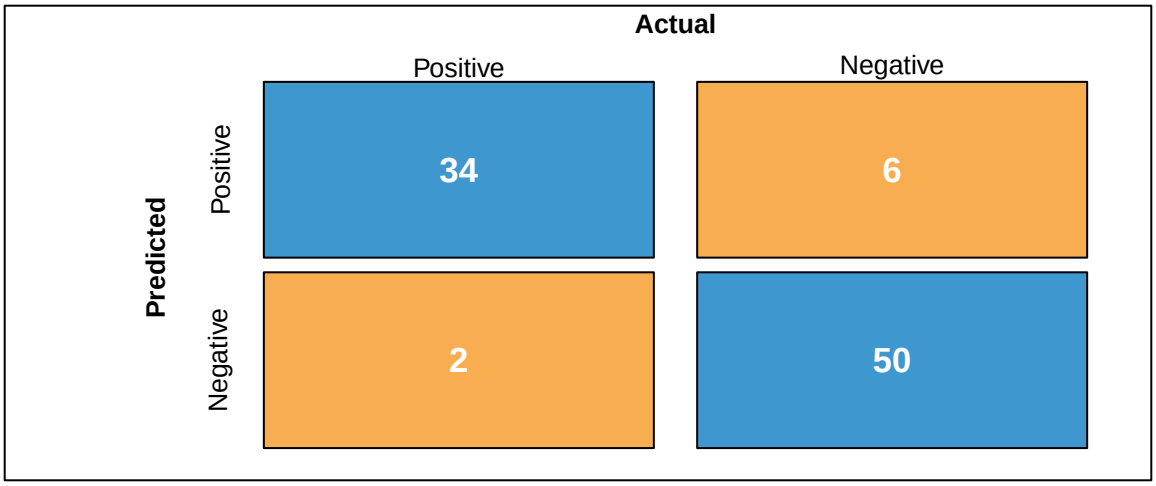

Since we use for our first kNN model just the default settings, we can still improve our model. By default, kNN algorithm is using k value 5 (object being assigned to the class most common among its 5 nearest neighbors), but it can be any positive integer. To find the best performing one, we created function, which tested all k values from 1 to 100 and return the best k. The plot bellow shows, how accuracy has been changing with different values of k. At the beginning, accuracy increase with any increase of value k, but

after a certain value k, it starts to slowly decline. This value of k is 9 and is also the best performing one.




We run kNN again, this time with our optimal value k 9. We managed to drastically improve our model. Our sensitivity is now 0.944, specificity is 0.893 and accuracy is 0.913, which is more than 0.04 better of our initial kNN model.

## CONFUSION MATRIX




**DETAILS**

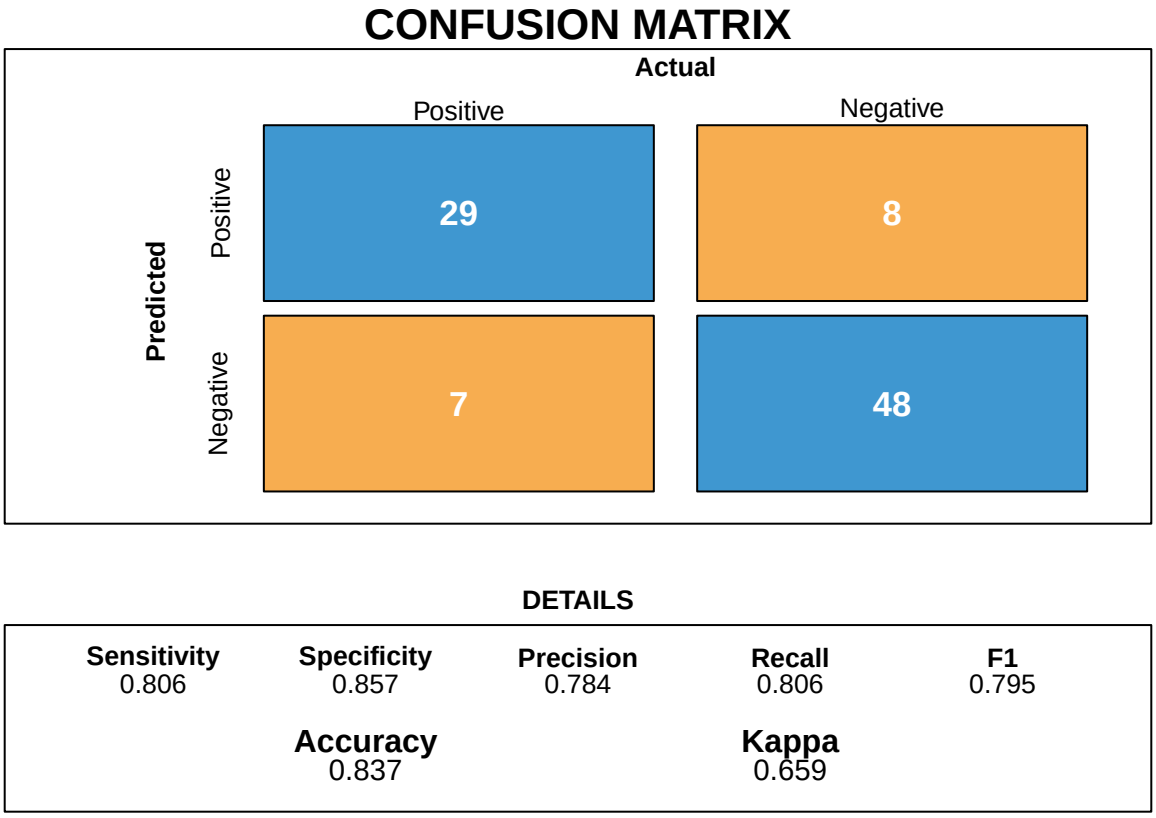
| Sensitivity | Specificity | Precision | Recall | F1 |
|---|---|---|---|---|
| 0.944 | 0.893 | 0.85 | 0.944 | 0.895 |

| Accuracy | Kappa |
|---|---|
| 0.913 | 0.821 |

**Random forest**

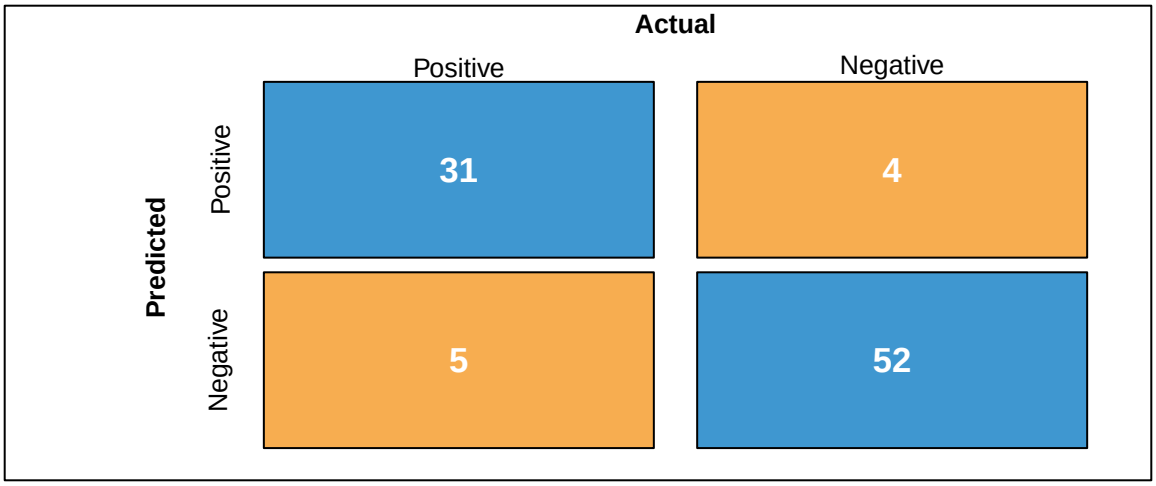
If we want to understand random forest, we first need to know about decision trees. Decision Trees are a supervised learning method used for classification. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A random forest consists of multiple individual decision trees that works as an ensemble. It is a machine learning technique that is used to solve classification problems. Each individual decision tree create its own prediction and the most frequent prediction becomes our model's prediction (wisdom of crowds).

Even though we already achieve quite accurate results with our kNN model, we will still develop a random forest algorithm to see, how do they compare. Similar to kNN development, we start with the default parameters. With that we got sensitivity 0.806, specificity 0.857 and accuracy 0.837, which is the wors so far. Even worse than the logistic regression. We will improve our random forest model with arguments optimization to see, how big of an impact different parameters can be.




Two of the most important arguments in random forest are ntree and mtry. Ntree is the number of trees to grow and mytry is the number of variables randomly sampled as candidates at each split. We will use double for loop function to test all the different combinations of ntree between 1 and 100 (only odd numbers) and mtry between 1 and 15. Going with double for loop is not the most optimal approach, since it consume a lot of resources, but in our case this was not the problem as it took just a few seconds to run. Surprisingly we got quite small optimal arguments with number of trees equals to 3, and number of variables randomly sampled as candidates at each split equals to 1. Results however are much better than before: sensitivity 0.861, specificity 0.929 and accuracy 0.902, which is significantly better than before (almost for 0.07 better) and almost as good as our best kNN model.

## CONFUSION MATRIX



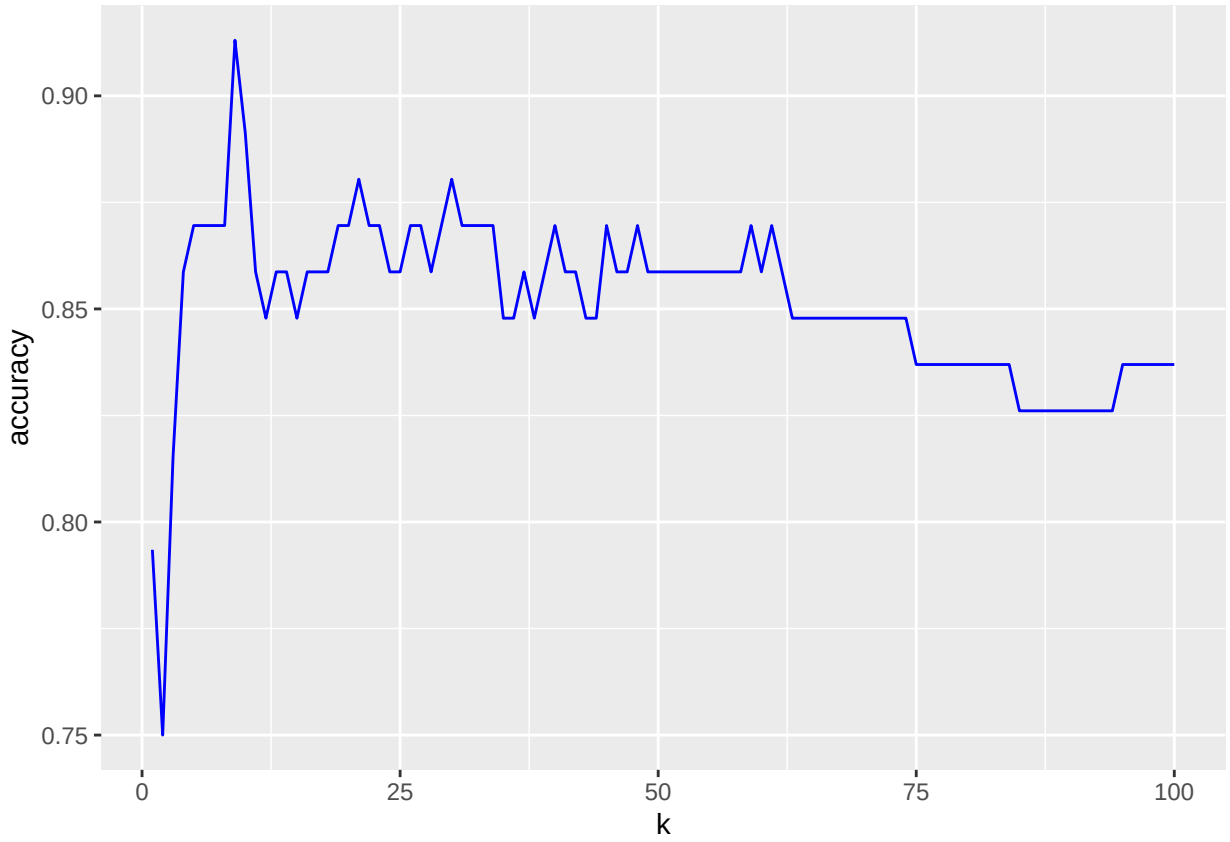

**DETAILS**

| Sensitivity | Specificity | Precision | Recall | F1 |
|---|---|---|---|---|
| 0.861 | 0.929 | 0.886 | 0.861 | 0.873 |

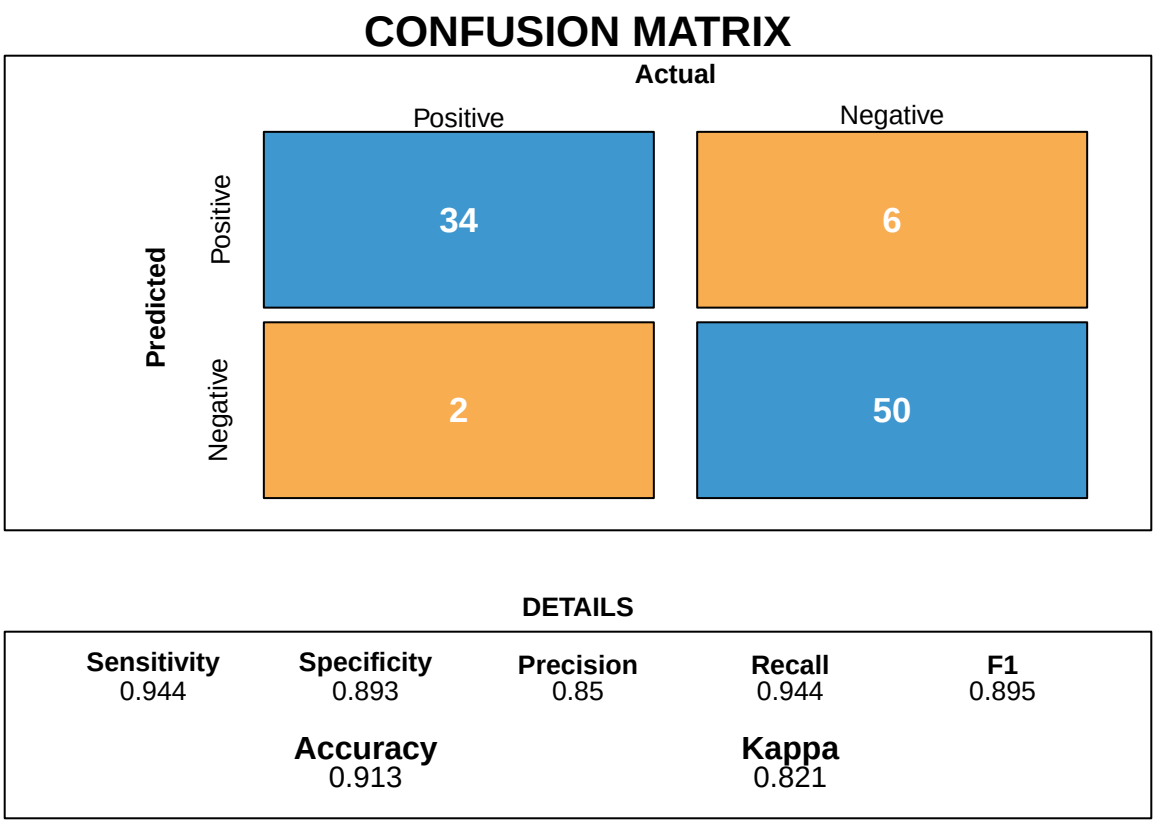
| Accuracy | Kappa |
|---|---|
| 0.902 | 0.794 |

## Results

As presented in the introduction, our goal was to predict heart failure, number one death cause globally, by using 11 features, provided by the dataset from Federico Soriano. In order to achieve the best results, we try three different algorithms and their optimization - logistic regression, k-nearest neighbours and random forest. We managed to achieve the highest accuracy with k-nearest neighbors algorithm (knn). However, our initial knn was not as good and we needed to improve it. To achieve better accuracy of our model, we used an approach, where we tested results for different k value. On the plot bellow is shown, how the accuracy has been changing depending on k value.




As mentioned above, we rated our model based on accuracy. Accuracy is the easiest way to immediately tell whether a model is being trained correctly and how it may perform generally. It is calculated as (true positives + true negatives) / total examples. In our case it was the highest when k = 9: (34 + 50) / (34 + 6 + 2 + 50) = 0.913




When looking at the sensitivity, specificity, precision and recall, it is easier, if we think about our results as positives and negatives:

- true positives (TP) - predicted heart disease (HD), actually has HD
- true negatives (TN) - predicted not to have HD (in our case, it is hard to say "healthy"), actually do not have HD
- false positives (FP) - predicted heart disease (HD), actually do not have HD
- false negative (FN) - predicted not to have HD, actually has HD

Our sensitivity is 0.861, which represents true positive rate or the proportion of heart disease predictions on our test set out of those who actually have the heart disease. These values are on the left side of the matrix: 34 / (34 + 2) = 0.944. Specificity in our case is 0.929, which represents the true negative rate or the proportion of the non-heart disease predictions out of those who do not actually have the heart disease. The values are in the right side of our confusion matrix: 50 / (6 + 50) = 0.893. Both, sensitivity and specificity are quite high in our case. Out of all the heart disease predictions, 85% are truly positive, which means that precision is 0.85 (upper side of the matrix): 34 / (34 + 6) = 0.85. Out of the total positive, 94.4% are predicted positive. It is the same as TPR (true positive rate) or recall (left side of the matrix): 34 / (34 + 2) = 0.944. F1 score is the harmonic mean of precision and recall. It takes both, FP and FN into account, thus, it has good performance also on the imbalanced datasets. F1 score gives the same weightage to recall and precision. It is calculated as 2 * precission * recall / (precission + recall). In our case it is 2 * 0.85 * 0.944 / (0.85 + 0.944) = 0.895. Similar to other metrics, perfect F1 score is 1 and total failure represents 0. Our F1 score is pretty good.

The Cohen's Kappa value is a metric that compares an Observed Accuracy with an Expected Accuracy (random chance). It tries to remove the evaluation bias by taking into account the correct classification by a random guess. It is calculated as overall (accuracy of the model - measure of the agreement between the model

predictions and the actual class values as if happening by chance) / (1 - measure of the agreement between the model predictions and the actual class values as if happening by chance). Calculating Kappa value is a bit more complicated, thus we will not calculate it by hand for our case, but there is a lot of materials online, which describe the procedure step by step. We got Cohen's Kappa value of 0.821. According to Cohen's original article, this is the actual scale:

- less than or 0: indicating no agreement
- 0.01–0.20: none to slight agreement
- 0.21–0.40: fair agreement
- 0.41– 0.60: moderate agreement
- 0.61–0.80: substantial agreement
- 0.81–1.00 almost perfect agreement

Our calculated value is in the latter category, which is almost perfect.

## Conclusion

Goal of our assignment was to find the machine learning use case, find the data and develop a model. We decided to predict heart disease, since it is one of the biggest causes for death globally. Prior model development, we needed to acquire the knowledge about the problem and our data. To gain a better understanding of heart disease topic, I read publicly available articles. Not only to know about how big the problem it is, but also to understand the factors that could impact the possibilities of developing the disease. I also did a data exploration to understand our patients data, how are different features distributed and how are they correlated between each other. When developing a model, I tried a few different approaches - I used logistic regression, k-nearest neighbors (knn) and random forest algorithms. I tried to tune each model in a way, that it gave us the most accurate predictions. We got the best results with knn algorithm, which was accurate in 91.3% cases.

Even though I think, that we managed to achieve great results, I am still aware, that model is far from being perfect. For starters, we could further improve our knn model, adding weights to our predictors etc. We could also test more algorithms, which could possibly gave us better performing model. It would be also possible to test more CPU or GPU-demanding machine learning techniques, which could give us better results. But for that, our personal computer would probably not be enough. However it is not necessary for us to focus just on the machine learning part. We could also perform better feature engineering, or try to collect more and better data, maybe with more parameters. But it would take more time and other resources. However, I am sure that there are people in the world, who are doing just that and I really hope, that with data science, they will be able to make a world just a little bit better.

# References


Albrektsen, G & Heuch, I & Lochen, M (2016). Lifelong Gender Gap in Risk of Incident Myocardial Infarction. Avaiblable at: https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2548254 (Accessed 8 October 2021)

American Heart Association editorial staff (2020). What is Cholesterol?. Available at: https://www.heart.org/en/health-topics/cholesterol/about-cholesterol (Accessed 8 October 2021)

Brown, D & Oldridge, N (1985). Exercise-induced angina in the cold. Available at: https://pubmed.ncbi.nlm.nih.gov/4068968/ (Accessed 8 October 2021)

Chicco, D & Jurman, J (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. Available at: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5 (Accessed 8 October 2021)

Great Learning Team (2020). Understanding Distributions in Statistics. Available at: https://www.mygreatlearning.com/blog/understanding-distributions-in-statistics/ (Accessed 8 October 2021)

Irizarry, R. (2020). Introduction to Data Science: Data Analysis and Prediction Alghoritms with R. CRC Press

Java Point (2021). K-Nearest Neighbor(KNN) Algorithm for Machine Learning. Available at: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning (Accessed 8 October 2021)

Mayo Clinic Staff (2020). Diabetes. Available at: https://www.mayoclinic.org/diseases-conditions/diabetes/diagnosis-treatment/drc-20371451 (Accessed 8 October 2021)

Palacios, F (2021). Heart Failure Prediction Dataset. Available at: https://www.kaggle.com/fedesoriano/heart-failure-prediction (Accessed 8 October 2021)

Sisense (2021). What is data standardization?. Available at: https://www.sisense.com/glossary/data-standardization/ (Accessed 8 October 2021)