# Likelihood to get heart disease among patients

**Group No: 4**

Group Members: Names listed below sorted with Matric No.

17028703 - Tan Chee Yu

17096993 - Lu Xian Ding

17201982 - Nicholas Lee Kan

17216654/1 - Saira Rupesh Nanglia

17217542 - Nik Faiz Afiq

17218915 - Muhammad Ikhwan b Nawari

17219953 - Arihant Jain

**Github Link** - <https://github.com/SairaRN/WQD7004-Group4-Project/tree/master/WQD7004-Group%204-Merged%20Code-With%20Cleaned%20DataSet>

**Raw Dataset Link**: <https://www.kaggle.com/johnsmith88/heart-disease-dataset>

## Brief Summary about the data

1. The dataset is from 1988.
2. It has 14 columns and 1025 rows.
3. The data is about the presence of heart disease in the patient. For this, various other parameters such as chest pain, fasting blood sugar, etc have also been considered.
4. In the raw data file you will see the following columns:

Age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, target

1. In the cleaned data set you will see an additional column ‘agegroups’ that contains grouping for the ages, and also cleaning of data for columns "Sex", "fbs", "exang","thal","target", “ca”.
2. Columns “Age” has been worked on using “data.table” package.
3. Columns "Sex", "fbs", "exang","thal","target", “ca” have been worked on using “plyr” and “dplyr” packages.

## Which section of the data do you need to tidy? Justify before implementing it

The section of the data regarding the data value of the table will be the one that needs to be tidied up. It is because 99% of the cleaning part will involve all these data values compared to the column’s name which has little or near to nothing to be tidy. We need to clean each row of the data value in the table because:

1. Some sections of the values only have two types of answers such as numerical type (1, 2) which can be converted to true or false or another type of data for example for gender it would be appropriate if we assign it to male or female. The conversion would lead to the high consistency of data and better data visualization.
2. There are some other sections of the values of the data that have been recorded wrongly that can lead to outlier’s error which can lead the results of our case study to become statistically significant. Therefore we need to apply a few blocks of codes to remove any outliers detected, for example, one of the data values called Slope needed to be clean because it has outliers.
3. Another section of the data including the entire dataset of the data values belongs to the table that we need to pay attention to check if there are any null values or duplicate values detected. This is important to ensure the dataset is accurate to prevent any bad reporting and promote standardization. Null values also needed to be filled accordingly to a suitable type of data replacement.
4. Data has also been arranged in groups, for example column “Age”, as this will assist in data visualization such as which age group is prone to heart diseases.

## Preparing data for analysis by correcting the variables and contents of the data.

## Analysis: Write down the questions, what is the objective/goal of processing this dataset? What answers you are interested to find through this dataset. Must write-down at least two questions.

**Objectives:**

The main objectives of the study are to predict what type of diseases the person would potentially develop based on the attributes that have been collected in the excel file. The objectives would be included with 3 research questions throughout the entire research studies.

1. What is the likelihood of getting heart diseases based on the attributes that we will obtain from the patient?
2. Which techniques are the most accurate and suitable that can be used for this case study?
3. How to apply the data mining technique and approaches in handling a large amount of raw data to effectively obtain a new insight and information from it?

## Project Approach

1. All 7 members connected to discuss, search, and narrow down on a dataset and analyze it.
2. Then we decided on the approach that we would follow in order to minimise chaos and redundancy. Thus we came up with the following plan:

* 3 members who will work on the data cleaning part are - Saira, Nik Faiz, and Nicholas.
* Meanwhile remaining members will plan for the next phase.
* The task distribution among the members for Phase 3 will be listed shortly.