# 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>

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

Project Expectation

## 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 chestpain, 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.

## 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.