Members:
Alakhjot Singh; Alexandre Feres; Mya Nandar May; Pinky Sangarmangkang; Shadi Hamidiaval

DANA Proposal

Population: The population is made up of all patients with potential heart disease in USA, Hungary, and Switzerland.
Objective: To perform statistical investigation analysis in R and predict if a person is prone to a heart attack or not.

Variables: Categorical | Numerical

1. **Cp**: chest pain type: typical angina, atypical angina, non-anginal, asymptomatic
2. **Restecg** (resting electrocardiographic results): normal, stt abnormality, lv hypertrophy
3. **Origin / dataset** (place of study): Cleaveland, Hungry, Switzerland, VA Long Beach
4. **Sex**: Male/Female
5. **Exang**: exercise-induced angina (True/ False)
6. **Chol**: serum cholesterol in mg/dl
7. **Age**: Age of the patient in years
8. **Thalch**: maximum heart rate achieved

5 statements:

Three for univariate analysis

Categorical: **Chest pain type (cp)** & **Ecg observation at resting condition (restecg)**

1. Describe the distribution of **chest pain types (cp)** in patients with potential heart disease.
    - Frequency / relative frequency
2. Describe the distribution of **ecg observation at resting condition (restecg)** in patients with potential heart disease.
    - Pie chart

Numerical: **Cholesterol measure (chol)**

3. Describe the distribution of **cholesterol level (chol)** in patients with potential heart disease.

Three for bivariate analysis

4. Is there a correlation between **age** and **maximum heart rate (thalch)** in patients?
    - Correlation analysis
5. Is there a relationship between **cholesterol levels (chol)** and **geographic regions (origin)**?
    - Box-plot comparison between geographic regions
6. Are **sex** and presence of **exercise induced angina (exang)** independent of each other?
    - Use Chi-square analysis

Sample size: 920

Source of data set: Kaggle
Data source: Source 1 (Kaggle) / Source 2 (UCI)