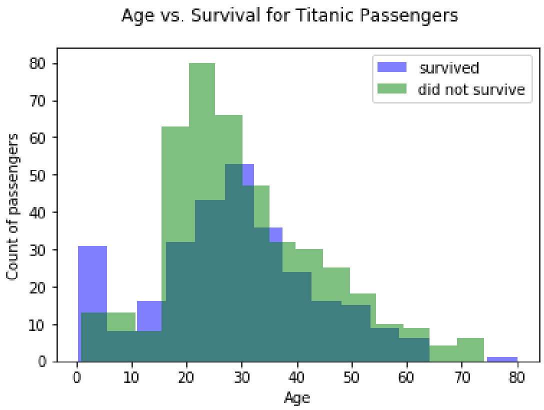
CS4092D Machine Learning Lab
Module1 (EDA): Exercise
Exploratory Data Analysis
(5 Marks)

Date: 19-Jan-2023

Q. **Download the Dermatology dataset 'dermatology.csv' (taken from the UCI machine learning repository). The dermatology.NAMES file (in text format) contains the names and descriptions of all the 35 features(columns) present in the dataset. Perform the following operations on this dataset:**

1. Read the dataset into a data-frame, and **rename the columns to a human-readable form:** You can use the 'dermatology.NAMES' file to get the corresponding column names. **(0.5 Marks)**

2. **Develop the histogram between features:** Develop a histogram by plotting between 'Age (linear)' (a numeric value) against the target feature 'Class code' (a categorical value). See the following example from the classical Titanic dataset. For better visualization, you may restrict 'Class code' to a few classes(2 or 3) even though class code varies from 1 to 6. **(0.5 Marks)**



3. **Impute the missing values** in the 'Age (linear)' attribute using the mean value of the column. Compute the mean of the column 'Age (linear)' after imputation. **(1 Mark)**

4. Create a new column in the dataframe titled 'Normalised_Age' whose values are the values of the **age column normalized between 0 and 1**. Calculate the mean value of the column 'Normalised_Age'. **(1 Mark)**

5. **Compute the Covariance Matrix** for the first 4 columns in the data-frame (*viz.*, 'erythema', 'scaling', 'definite borders', and, 'itching'). **(1 Mark)**

6. **Using a box plot find the anomalous data-points** in the feature 'Age (linear)'. Name the index of the rows corresponding to the anomalies, if any. **(1 Mark)**