**DATASET 1: SAheart.data**

The dataset SAheart.data is taken from the link below

<http://www-stat.stanford.edu/~tibs/ElemStatLearn/datasets/SAheart.data>

The dataset contains records of males in a heart-disease high-risk region of the Western Cape, South Africa. There are roughly two controls per case of CHD. Many of the CHD positive men have undergone blood pressure reduction treatment and other programs to reduce their risk factors after their CHD event. In some cases, the measurements were made after these treatments. These data are taken from a larger dataset, described in Rousseauw et al. (1983), South African Medical Journal. It is a tab separated file (.csv) and contains the following columns:

� sbp – Systolic blood pressure

� tobacco – Cumulative tobacco (kg)

� ldl – Low densiity lipoprotein cholesterol

� adiposity

� famhist – Family history of heart disease (Present, Absent)

� typea – type-A behavior

� obesity

� alcohol – Current alcohol consumption

� age – Age at onset

� chd – Response, coronary heart disease

To read into Python:

SA\_df = pd.read\_csv("https://web.stanford.edu/~hastie/ElemStatLearn/datasets/SAheart.data" )

**DATASET 2: customerspends.csv**

An online grocery store has captured amount spent per annum (in Indian rupees) by 20 customers on apparel and beauty and healthcare products and given in the fle customerspends.csv. It contains the following records.

� Customer − Customer ID

� Apparel − Amount spent in apparel products

� Beauty and Healthcare − Amount spent in beauty and healthcare products

**DATASET 3: Online Retail.xlsx**

The dataset Online Retail.xlsx and the description of the data is taken from https://archive.ics.uci.edu/ml/datasets/online+retailOnline Retail.xlsx contains records of transactions that occurred between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers. The attributes in the dataset are:

InvoiceNo − Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter ‘c’, it indicates a cancellation.

StockCode − Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.

Description − Product (item) name. Nominal.

Quantity − The quantities of each product (item) per transaction. Numeric.

InvoiceDate − Invoice Date and time. Numeric, the day and time when each transaction was generated.

UnitPrice − Unit price. Numeric, Product price per unit in sterling.

CustomerID − Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

Country − Country name. Nominal, the name of the country where each customer resides.

**DATASET 4: ratings.csv and movies.csv**

We will use MovieLens dataset (see https://grouplens.org/datasets/movielens/) for ﬁnding similar users based on common movies the users have watched and how they have rated those movies. The ﬁle ratings.csv in the dataset contains ratings given by users. Each line in this ﬁle represents a rating given by a user to a movie. The ratings are on the scale of 1 to 5. The dataset has the following features:

1. userId

2. movieId

3. rating

4. timestamp

Movie information is contained in the ﬁle movies.csv. Each line of this ﬁle contains

* the movieid,
* the movie name, and
* the movie genre.

Movie titles are entered manually or imported from https://www.themoviedb.org/ and include the year of release in parentheses. Errors and inconsistencies may exist in these titles. The movie can be loaded using the following codes: