Heart disease and tabular data (Case 1)

Neural Networks of Machine Learning Applications

Spring 2023

Sakari Lukkarinen

Metropolia University of Applied Sciences

Contents

- Heart disease (coronary artery disease)
  - Risk factors for heart disease
  - Selected subset of features from BRFSS 2015 data
  - Heart disease health indicators dataset – Notebook
  - Practise
- Explore the dataset
  - How to read the dataset in Kaggle
  - How to take a sample
  - What features are there
  - How to check missing values
  - Practise
- To do with Case 1
  - Basic skills, Advanced skills

Heart disease (coronary artery disease)

Coronary artery disease (**CAD**), also known as **coronary heart disease** (**CHD**), **ischemic heart disease** (**IHD**), or simply **heart disease**, involves the reduction of blood flow to the heart muscle due to build-up of plaque (atherosclerosis) in the arteries of the heart.

It is the most common of the cardiovascular diseases. Types include stable angina, unstable angina, myocardial infarction, and sudden cardiac death.

A common symptom is chest pain or discomfort which may travel into the shoulder, arm, back, neck, or jaw. Occasionally it may feel like heartburn.

Coronary artery disease (Wikipedia)

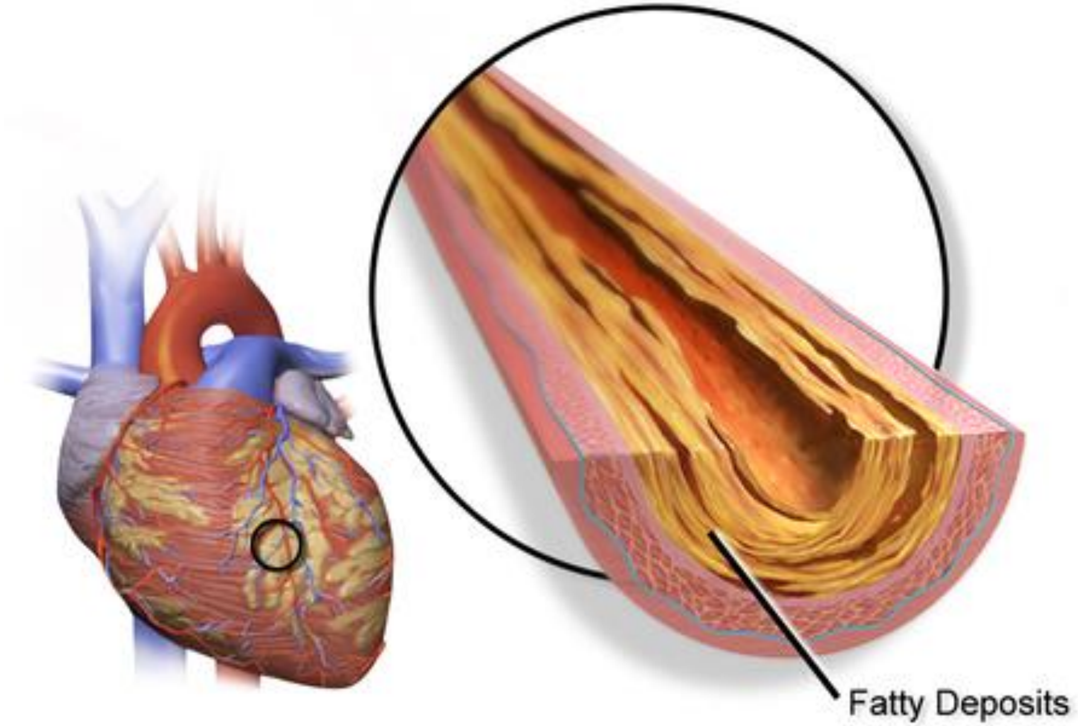


Illustration depicting atherosclerosis in a coronary artery.

Risk factors for heart disease

Risk factors include high blood pressure, smoking, diabetes, lack of exercise, obesity, high blood cholesterol, poor diet, depression, and excessive alcohol.

A number of tests may help with diagnoses including: electrocardiogram, cardiac stress testing, coronary computed tomographic angiography, and coronary angiogram, among others.

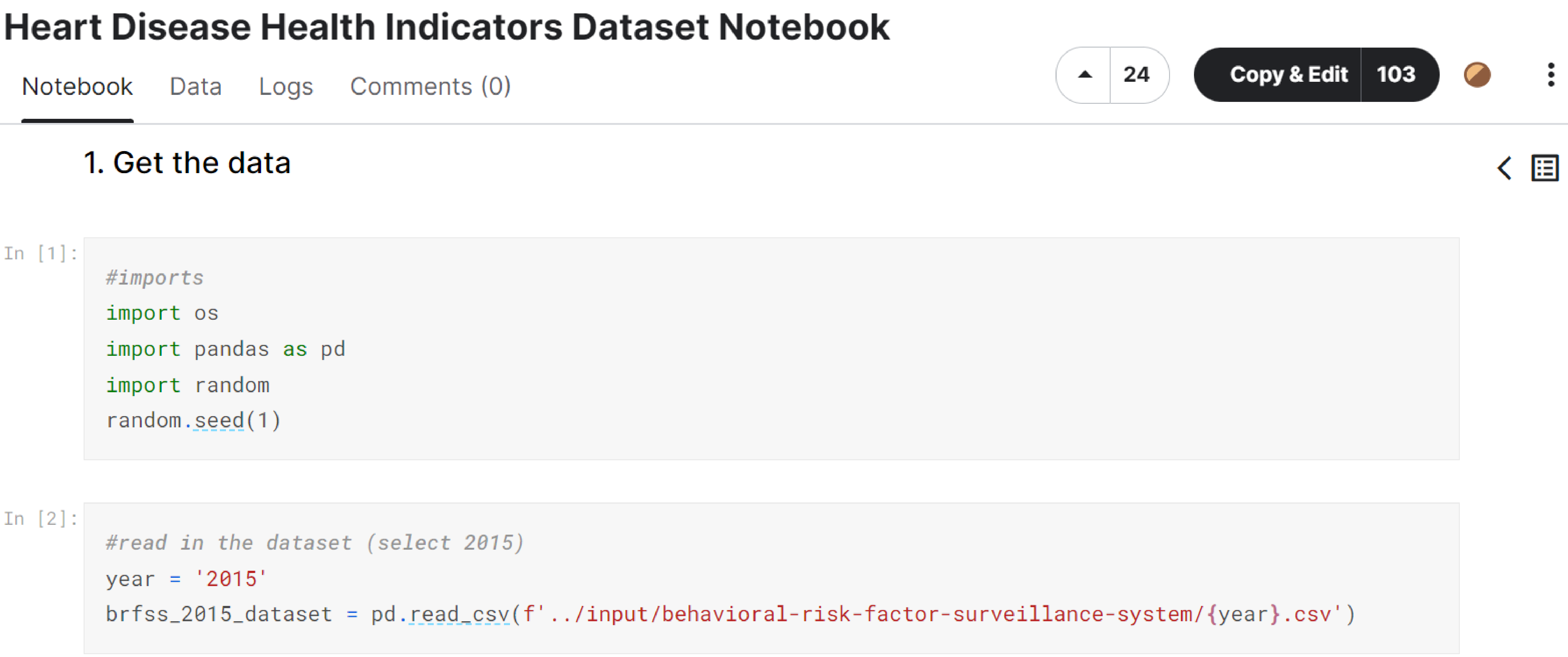
Coronary artery disease (Wikipedia)

A coronary angiogram (an X-ray with radiocontrast agent in the coronary arteries) that shows the left coronary circulation.

Coronary catheterization (Wikipedia)

Selected Subset of Features from BRFSS 2015

- Behavioral Risk Factor Surveillance System (BRFSS) is a premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviours, chronic health conditions, and use of preventive services.
- The author of the dataset for case 1 have selected features (columns/questions) in the BRFSS related to the given risk factors. The Heart Disease Health Indicators Dataset Notebook explain what the columns mean.
- A research paper by Zidian Xie et al (2019) *Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques* has also used some of the same features from BRFSS 2014.
- Diabetes and Heart Disease outcomes are strongly correlated, with the primary cause of death for diabetics being heart disease complications. Given this information, it is a useful starting point.



Heart Disease Health Indicators Dataset Notebook

1. Get the data

In [1]:

```
#imports
import os
import pandas as pd
import random
random.seed(1)
```

In [2]:

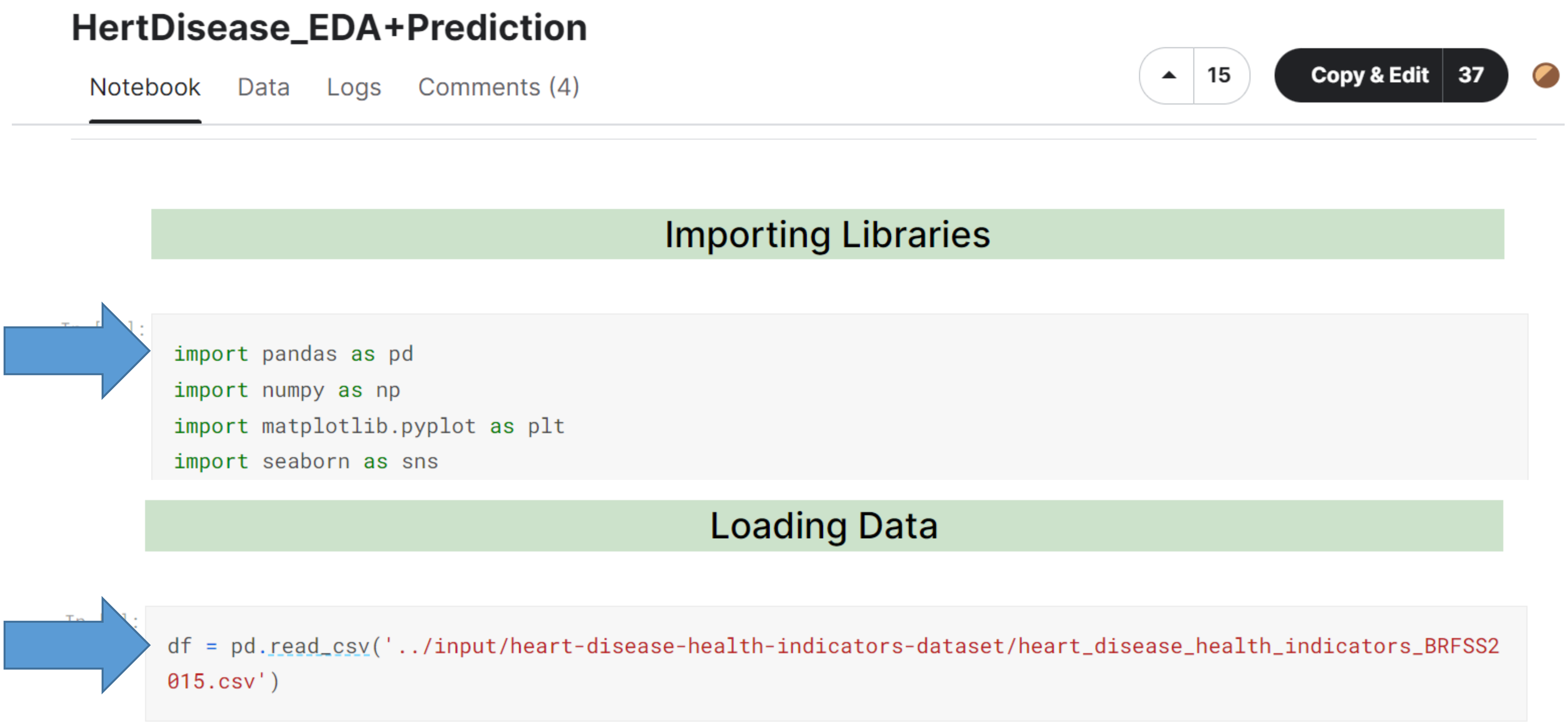
```
#read in the dataset (select 2015)
year = '2015'
brfss_2015_dataset = pd.read_csv(f'../input/behavioral-risk-factor-surveillance-system/{year}.csv')
```

Heart Disease Health Indicators Dataset Notebook | Kaggle

Practise

- Open the Heart Disease Health Indicators Dataset Notebook.
- Study how
    - The raw data is get into the Notebook
    - The data is cleaned
    - The features are renamed to be more readable
    - The cleaned features and labels are saved to a CSV-file

How to read the dataset in Kaggle



HertDisease_EDA+Prediction

Importing Libraries

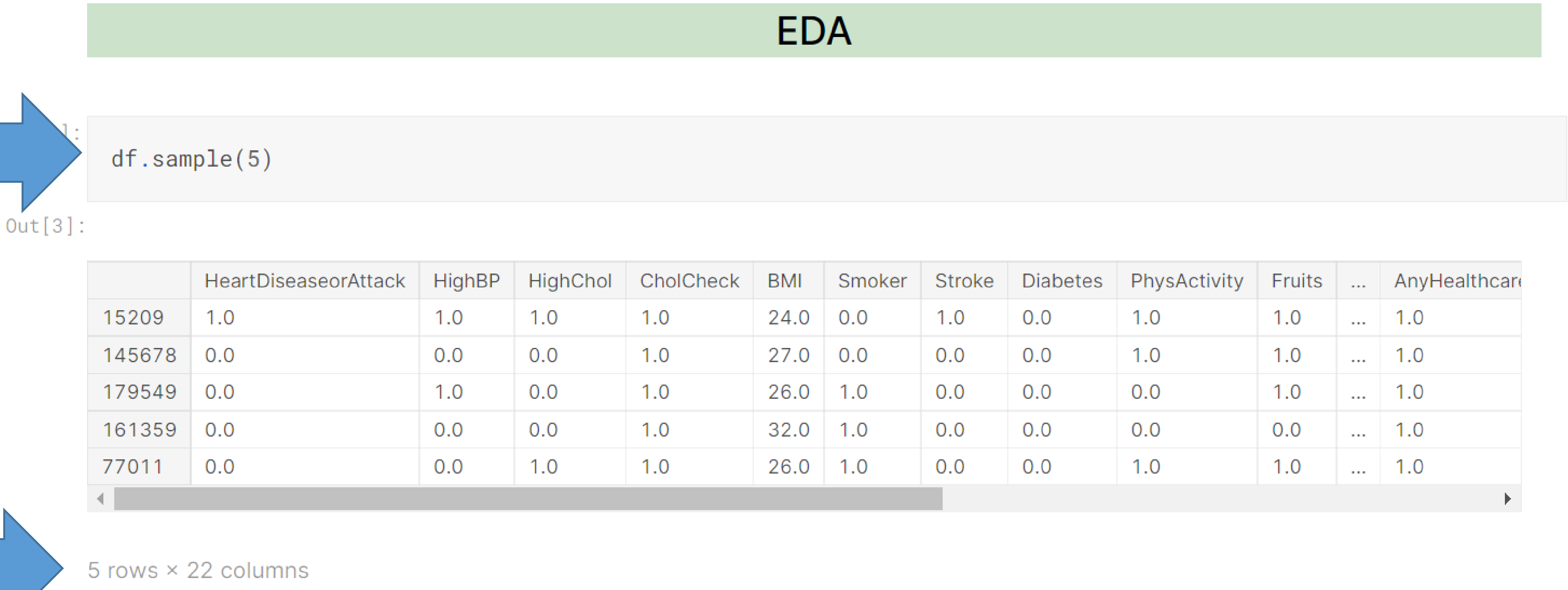
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Loading Data

```
df = pd.read_csv('../input/heart-disease-health-indicators-dataset/heart_disease_health_indicators_BRFSS2015.csv')
```

[HertDisease_EDA+Prediction | Kaggle]

How to take a sample from dataset



EDA

```
df.sample(5)
```

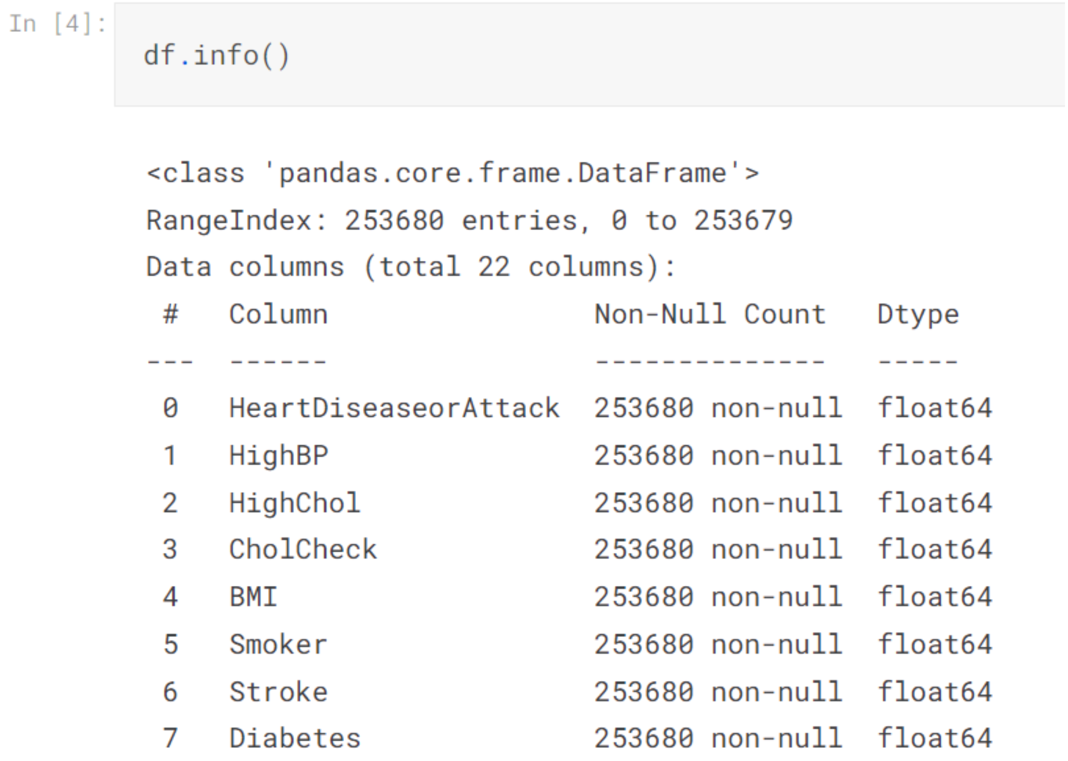
Out[3]:

	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes	PhysActivity	Fruits	...	AnyHealthcar
15209	1.0	1.0	1.0	1.0	24.0	0.0	1.0	0.0	1.0	1.0	...	1.0
145678	0.0	0.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	...	1.0
179549	0.0	1.0	0.0	1.0	26.0	1.0	0.0	0.0	0.0	1.0	...	1.0
161359	0.0	0.0	0.0	1.0	32.0	1.0	0.0	0.0	0.0	0.0	...	1.0
77011	0.0	0.0	1.0	1.0	26.0	1.0	0.0	0.0	1.0	1.0	...	1.0

5 rows × 22 columns

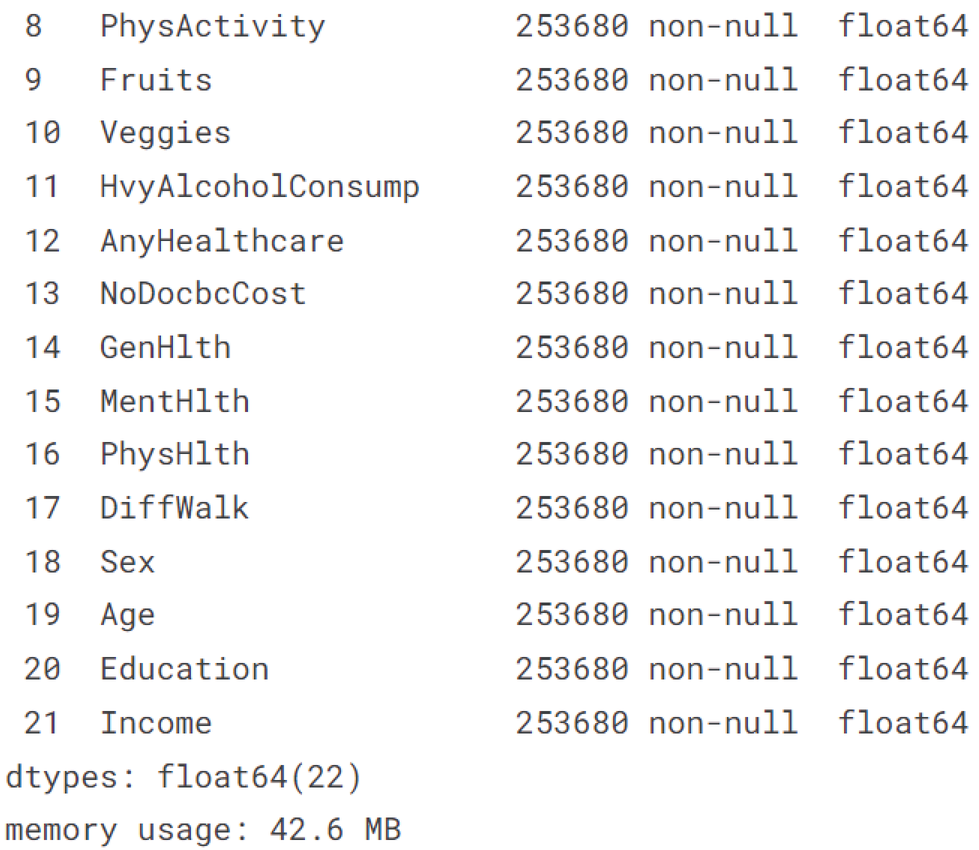
HertDisease_EDA+Prediction | Kaggle

What features (columns) are there?



```
In [4]: df.info()
```

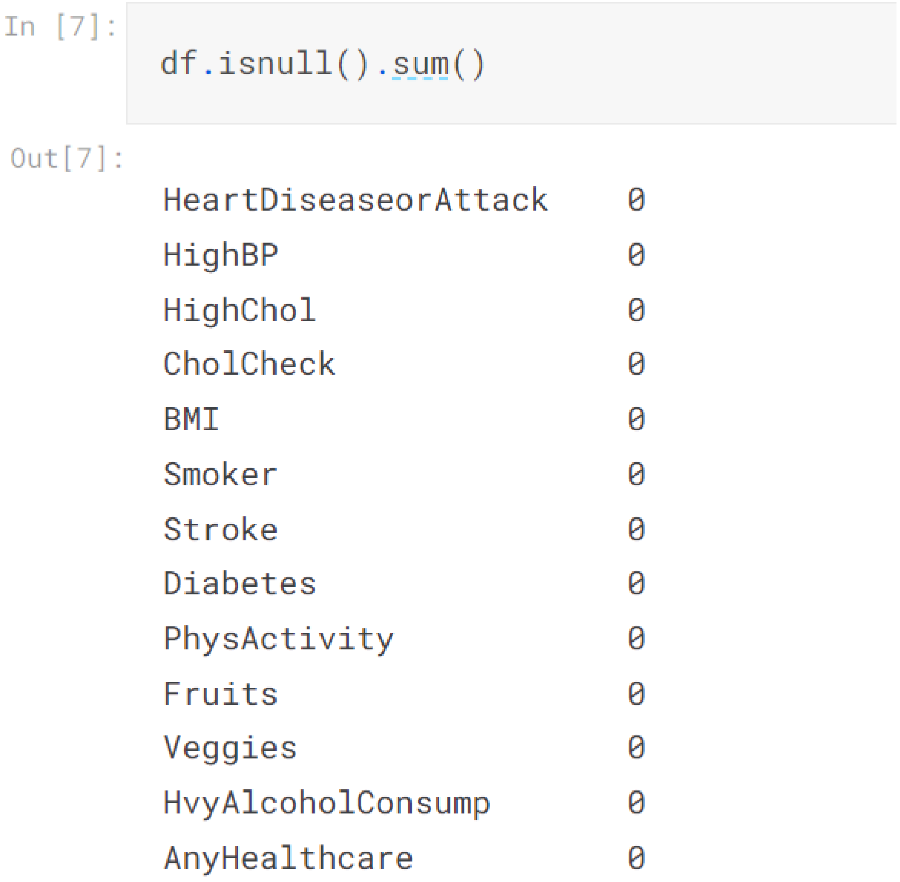
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   HeartDiseaseorAttack  253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   Diabetes              253680 non-null  float64
```



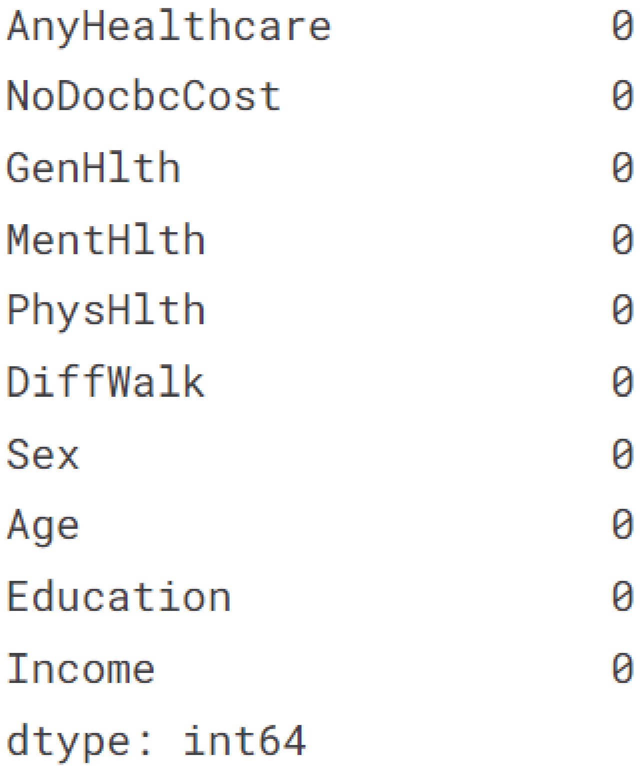
```
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
dtypes: float64(22)
memory usage: 42.6 MB
```

HertDisease_ EDA+Prediction | Kaggle

How many missing values are there?



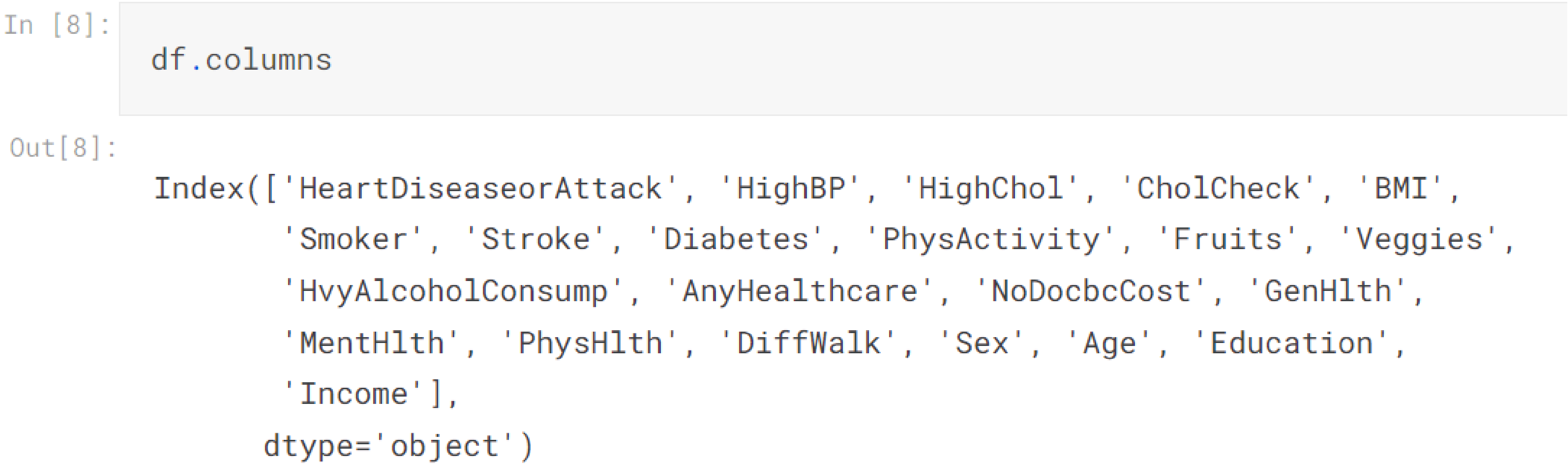
```
In [7]:  df.isnull().sum()

Out[7]:
         HeartDiseaseorAttack    0
         HighBP                  0
         HighChol                0
         CholCheck               0
         BMI                     0
         Smoker                  0
         Stroke                  0
         Diabetes                0
         PhysActivity            0
         Fruits                  0
         Veggies                 0
         HvyAlcoholConsump       0
         AnyHealthcare           0
```



```
AnyHealthcare           0
NoDocbcCost             0
GenHlth                 0
MentHlth                0
PhysHlth                0
DiffWalk                0
Sex                     0
Age                     0
Education               0
Income                  0
dtype: int64
```

[HertDisease_EDA+Prediction | Kaggle]

Short list of the columns



```
In [8]:  df.columns

Out[8]:
Index(['HeartDiseaseorAttack', 'HighBP', 'HighChol', 'CholCheck', 'BMI',
       'Smoker', 'Stroke', 'Diabetes', 'PhysActivity', 'Fruits', 'Veggies',
       'HvyAlcoholConsump', 'AnyHealthcare', 'NoDocbcCost', 'GenHlth',
       'MentHlth', 'PhysHlth', 'DiffWalk', 'Sex', 'Age', 'Education',
       'Income'],
      dtype='object')
```

HertDisease_EDA+Prediction | Kaggle

Practise

- Open the HertDisease_EDA+Prediction | Kaggle
- Study
  - How the data is first explored
    - categories, ordinal data, heart disease vs HighBP, HighCol, etc.
    - education and income vs. different categories
  - How the features are engineered and the data is preprocessed
- Note
  - Forget the last part where different algorithms are tried
  - We are not going to use these algorithms during this course!

To do with Case 1 – Basic skills

- Start with Case 1. Template (see Assignments folder)
- Learn to read the dataset using pandas read_csv-function
  - Use smaller sample from the dataset, for example, try with 20,000 samples

1. Make a straightforward preprocessing step
   - Normalize the dataset
   - Split into training, validation and test sets
2. Make a standard classifier (input, hidden, and output layers)
   - Play with layers and number of neurons
3. Learn to use the performance metrics
   - During training: **Accuracy**
   - After training and during testing:
     - **Classification report** (Sensitivity = Recall, Specificity)**, confusion matrix,** ROC curve

To do with Case 1 – Advanced skills

1. Make a preprocessing plan
   - Study the variables
   - Convert between categorical and numerical values
   - Use one-hot-coding
   - Modify the model accordingly
2. Try cross-validation techniques
3. Use all data