1 Hands On: Data Import and Manipulation

1. Introduction to Python for Machine Learning

(a) Create an array of 10 random integers between 1 and 100, and then calculate the mean and standard deviation of the array.

(b) Create a 2-dimensional array of 3 rows and 4 columns with random integer values. Then, calculate the sum of each row and column.

(c) Create a DataFrame with 3 columns: "Name", "Age", and "City". Add at least 5 rows of data to the DataFrame. Then, filter the DataFrame to only include rows where the person's age is greater than or equal to 30.

(d) Load a CSV file into a DataFrame and then calculate the mean, median, and mode of one of the columns in the DataFrame.

(e) Create a scatter plot of random x and y values between 1 and 100.

(f) Load a CSV file into a DataFrame and then create a line chart of one of the columns in the DataFrame.

2. Introduction to R for Machine Learning

(g) Create an array of 10 random integers between 1 and 100, and then calculate the mean and standard deviation of the array.

(h) Create a 2-dimensional array of 3 rows and 4 columns with random integer values. Then, calculate the sum of each row and column.

(i) Create a DataFrame with 3 columns: "Name", "Age", and "City". Add at least 5 rows of data to the DataFrame. Then, filter the DataFrame to only include rows where the person's age is greater than or equal to 30.

(j) Create a DataFrame with 3 columns: "Name", "Age", and "City". Add at least 5 rows of data to the DataFrame. Then, filter the DataFrame to only include rows where the person's age is greater than or equal to 30.

(k) Load a CSV file into a DataFrame and then calculate the mean, median, and mode of one of the columns in the DataFrame.

(l) Create a scatter plot of random x and y values between 1 and 100.

(m) Load a CSV file into a DataFrame and then create a line chart of one of the columns in the DataFrame.

3. Data Import

The Echocardiogram data set in the UCI Machine Learning repository contains information on a set of patients that suffered heart attacks at some point in the past.

(a) Download the Echocardiogram data set and import it to a data frame. Read the information on the data set and find out how missing values are represented and make sure that they are properly represented.

(b) Assign the attributes with meaningful names. You can look for this information on the same webpage.

(c) According to that same information, is there any redundant or irrelevant attribute that you can remove? Remove them.

(d) Is there any data type change that you find useful? Perform it.

4. Data Manipulation

Load the `airquality` data set regarding a set of New York Air Quality Measurements.

(n) For which attributes are there missing values?

(o) Do all the attributes are in the most suitable data type? Make the changes you find necessary.

(p) What period of the year do these records refer to?

(q) Load the package `dplyr` and save the data set in a table data frame format.

(r) Select the days in May with a temperature above 70 Fahrenheit.

(s) Create a new attribute `TempC` which represents the temperature values in Celsius.

(t) Inspect which were the 30 hottest days.

(u) Inspect which were the hottest days, but also with the highest ozone values.

(v) Inspect the number of days for which there was a register for each month.

(w) For each month, obtain the minimum and the maximum temperature registered in Celsius.

(x) Obtain the average of the following parameters by month: temperature in celsius, wind, solar radiation and ozone.

(y) What values did you obtain regarding ozone and solar radiation attributes? Why? Make the necessary change so that you get the average of the registered values.