Fourth Industrial Revolution (4IR) Summer School

Module **2** – **D**ay **1** exercises

Pandas Data Structures

Task 1: Series Manipulation

Given two series: first one created form a list e.g. "list('abcedfghijklmnopqrstuvwxyz'), while the second one is created using NumPy array.

Write a Pandas program to combine two series to form a DataFrame, and rename the labels for the DataFrame's columns to be "Char" and "Num"

Task 2: Filtering

Write a Pandas program to create a subset of a given series based on value and condition.

Sample Series: [2, 4, 6, 8, 10]
Sample condition: the value > 5.

Task 3: DataFrame manipulation

Write a Pandas program to create a DataFrame using the data given in the following:

Sample DataFrame:

```
exam_data = {'Name': ['Anna', 'Dima', 'Alex', 'James', 'Emily', 'Tom', 'Matt',
                      'Laura', 'Kevin', 'Jonas'],
        'Score': [12.5, 9, 16.5, np.nan, 9, 20, 14.5, np.nan, 8, 19],
        'Attempt': [1, 3, 2, 3, 2, 3, 1, 1, 2, 1],
        'Qualify': ['yes', 'no', 'yes', 'no', 'no', 'yes', 'yes', 'no', 'no', 'yes']}
labels = ['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j']
```

a) Get a list from DataFrame column headers
b) Display the datatypes of columns of a DataFrame
c) Use the Name column to be the index of the DataFrame

Task 4: DataFrame manipulation

Extend the previous task to perform the following operations on the created DataFrame:

a) Calculate the sum of the examination attempts by the students.
b) Return the 'complete data' of the row having the maximum score
c) Find the average of the provided scores

Task 5: DataFrame manipulation

Utilize the DataFrame created above to perform the following tasks:

a) Update the score obtained by 'James' to 11.5.
b) Change the name 'James' to 'Sam' in name column of the data frame.
c) Append a new row 'k' to DataFrame with given values for each column.
d) Add new column “Training” that specify the eligibility of training for each student if his qualify is Yes.

Additional Task:

Consider a DataFrame called "sales_data" that contains information about sales transactions. The DataFrame has the following columns: "Product", "Quantity", "Price", and "Total".

Sample DataFrame:

```
data = {
   "Product": ["A", "B", "C", "D", "E", "F", "G", "H", "I", "J"],
   "Quantity": [10, 15, 8, 12, 20, 5, 18, 9, 14, 7],
   "Price": [25, 30, 20, 15, 35, 40, 22, 28, 32, 18]
}
```

a) Create a DataFrame called "top_selling_products" that includes only the top 5 products based on the highest quantity sold.
b) Calculate the total revenue generated from each product by multiplying the "Quantity" and "Price" columns, and add a new column called "Revenue" to the DataFrame.