**SVKM’s NMIMS**

**Mukesh Patel School of Technology Management & Engineering**

Program: BTI Computer Engineering

**Course: Data Mining**

**Experiment No.02**

PART A

**A.1 Aim:** Introduction to Data Mining and Pandas library

**Task:** Perform exploratory data analysis on Indian cuisine dataset and write the inferences for each question.

1. Read the indianfood1.csv file into a DataFrame.
2. Explore size, shape, data types of each column in the dataset.
3. How many total Indian dishes are there?
4. How many different set of features are there in every Indian dish?
5. Using Describe function, view the basic statistics of all columns. What Inference you can make out form that?
6. Are there any missing values in the dataset? If Yes, replace the missing values with the NaN values.
7. How many numeric features and categorical features are there in the dataset?
8. Display the number of unique values in each column.
9. Add a new column in the dataset to calculate the total time taken to make every dish.
10. Add a new column in the dataset that will count the number of ingredients from the ingredients column for each dish.

**A.2 Prerequisite:**

Python Programming, Pandas library

**A.3 Outcome:**

**After successful completion of this experiment students will be able to:**

* 1. Differentiate applications of supervised and unsupervised learning
  2. Read different types of data files (csv, excel, text file etc.)
  3. Obtain metadata of given dataset

**A.4 Theory:**

**Exploratory Data Analysis:**

Exploratory Data Analysis (EDA) is an open-ended process where we calculate statistics and make figures to find trends, anomalies, patterns, or relationships within the data. The goal of EDA is to learn what our data can tell us. It generally starts out with a high level overview, then narrows in to specific areas as we find intriguing areas of the data. The findings may be interesting in their own right, or they can be used to inform our modeling choices, such as by helping us decide which features to use.

**Pandas Library:**

**DataFrame** is a 2-dimensional labeled data structure with columns of potentially different types. You can think of it like a spreadsheet or SQL table, or a dict of Series objects. It is generally the most commonly used pandas object. Like Series, DataFrame accepts many different kinds of input:

* Dict of 1D ndarrays, lists, dicts, or Series
* 2-D numpy.ndarray
* [Structured or record](http://docs.scipy.org/doc/numpy/user/basics.rec.html) ndarray
* A Series
* Another DataFrame

PART B

(PART B : TO BE COMPLETED BY STUDENTS)

***(Students must submit the soft copy as per following segments within two hours of the practical.)***

|  |  |
| --- | --- |
| Roll No.: C026 | Name: Anirbaan Ghatak |
| Class : B | Batch : B1 |
| Date of Experiment: 25/07/2023 | Date of Submission |
| Grade : |  |

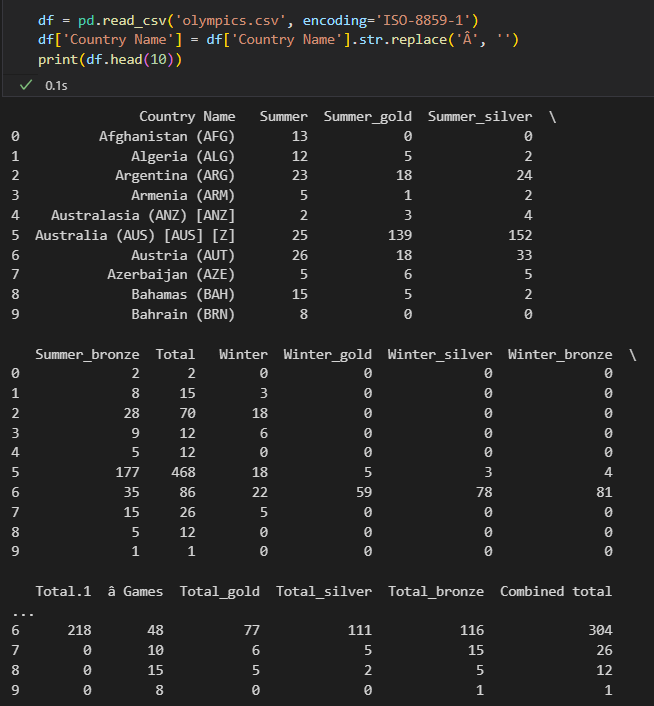
**B.1 Task1**

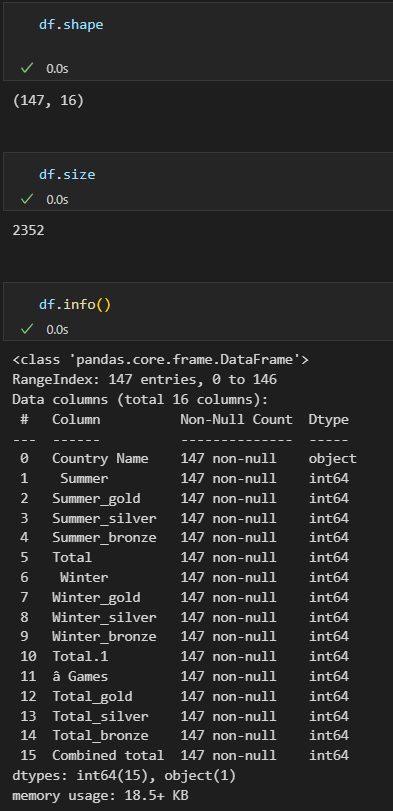
1. Which country has won the most gold medals in summer games?

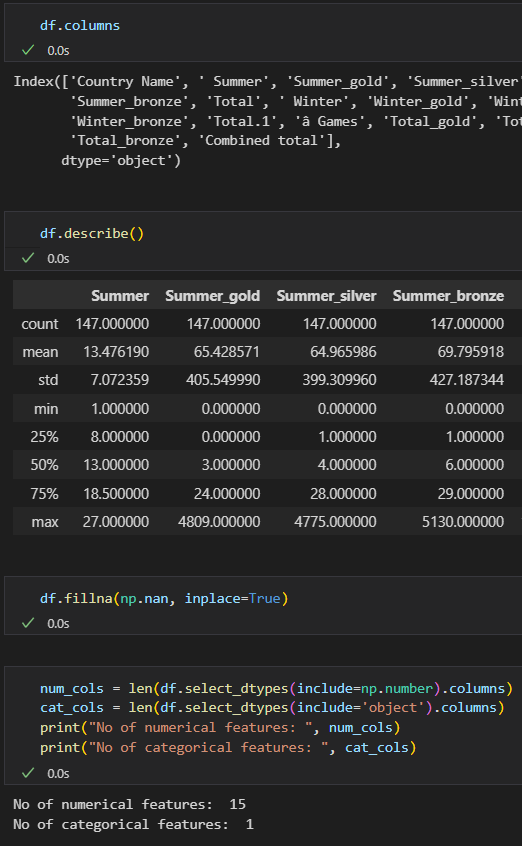
2. Which country had the biggest difference between their summer and winter gold medal counts?

3. Which country has the biggest difference between their summer and winter gold medal counts relative to their total gold medal count? Only include countries that have won at least 1 gold in both summer and winter.

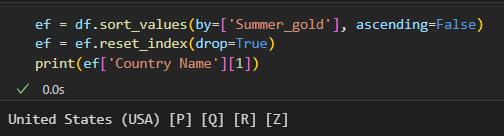
4. Write a function to update the dataframe to include a new column called "Points" which is a weighted value where each gold medal counts for 3 points, silver medals for 2 points, and bronze medals for 1 point. The function should return only the column (a Series object) which you created

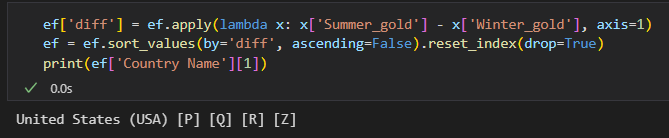
****

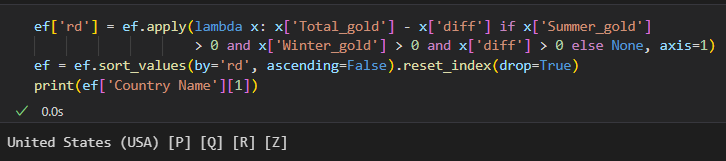
****

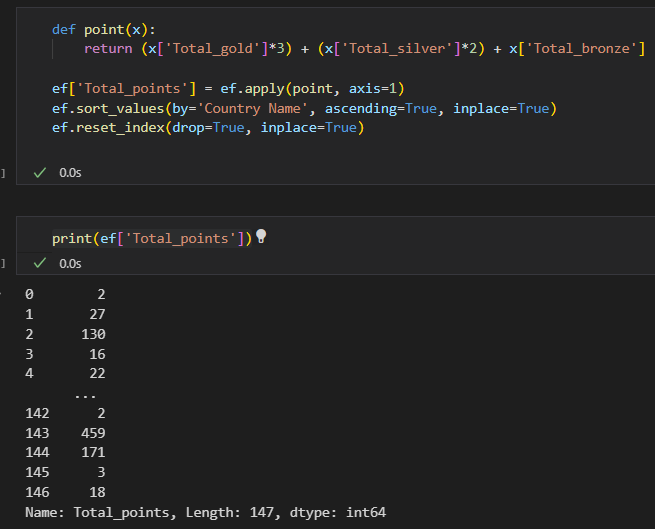
****

****

****

****

****

****

**B.4 Conclusion:**

*Dived a bit deeper into the concepts of pandas*