Data Engineering Batch 1 **Date: 02/02/2024**

Name: Pradip Bochare

Python Coding Challenge

- **Explain Pandas for Data Processing:**
  Pandas is a popular open-source data manipulation and analysis library for the Python programming language. It provides data structures and functions needed to manipulate and analyze structured data, primarily in the form of tabular data like spreadsheets and SQL tables.

❖ Key components of Pandas

1. DataFrame:
   The central data structure in Pandas is the DataFrame, a two-dimensional table with labeled axes (rows and columns). Columns can have different data types (integers, floats, strings) and can be heterogeneous.

2. Series:
   A one-dimensional labeled array capable of holding any data type. A single column of a DataFrame is essentially a Series.

3. Index:
   Pandas uses the Index to label the rows and columns of a DataFrame.It helps in selecting, slicing, and manipulating data.

4. Data Cleaning and Transformation:
   Pandas provides numerous functions for handling missing data, filtering, and cleaning data. It allows reshaping and pivoting data, merging and concatenating datasets, and handling duplicates.

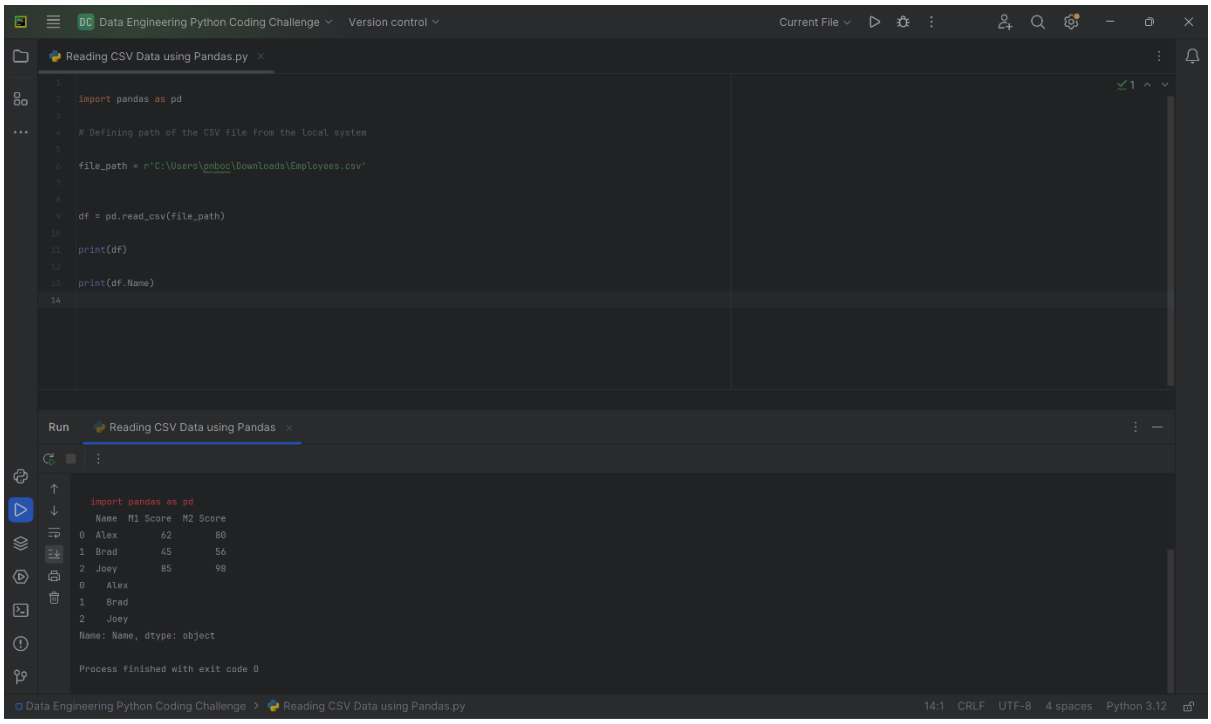
5. Data Selection and Indexing:
   Pandas offers powerful tools for selecting, indexing, and filtering data. It supports both label-based and position-based indexing.

6. Grouping and Aggregation:
   Pandas enables grouping of data based on one or more criteria and then applying a function to each group independently. Functions like groupby() and agg() facilitate these operations.

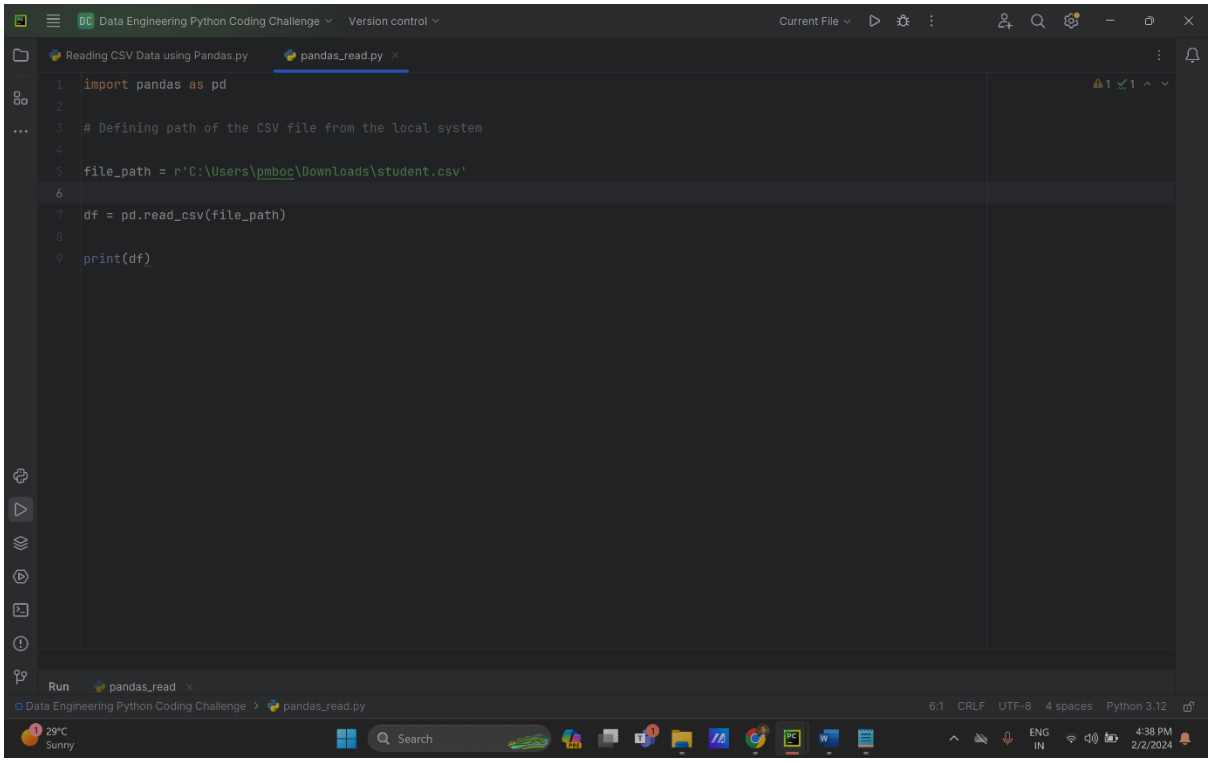
Pandas supports reading and writing data in various formats, such as CSV, Excel, SQL databases

Execute Reading CSV Data using Pandas

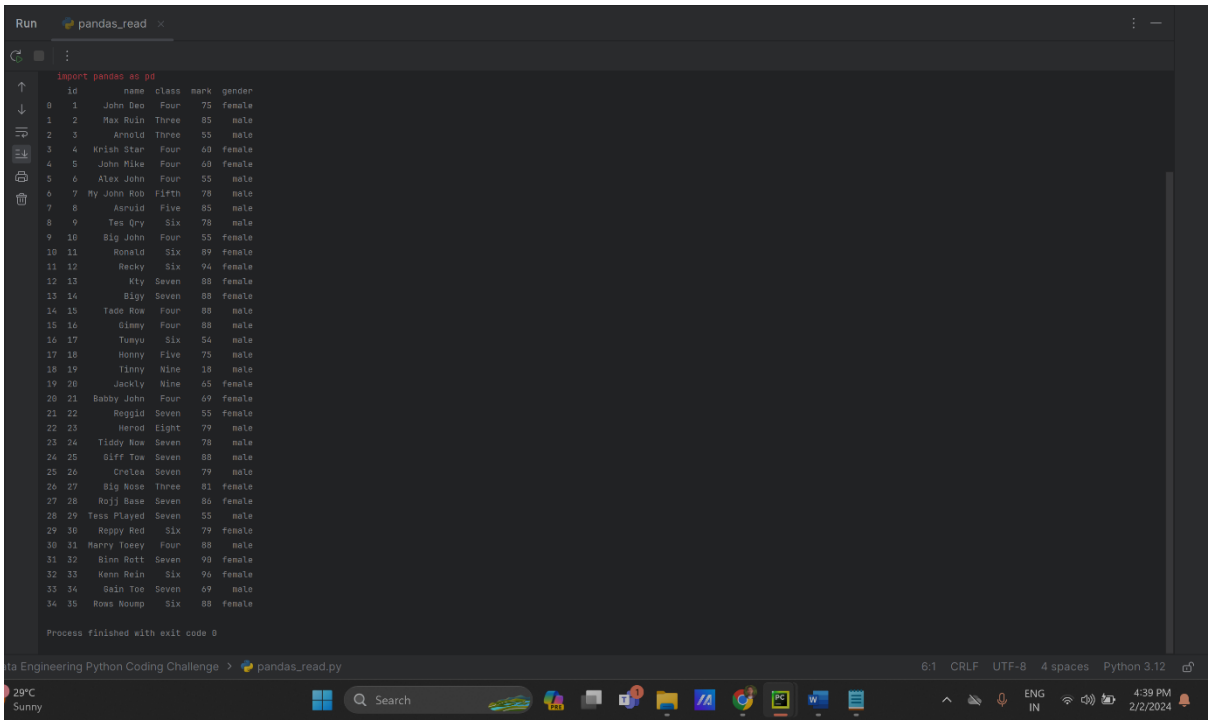
To read CSV file using Pandas, we use the ‘pandas.read_csv()’ function
First, we need to import pandas module after that give file path of the csv file we want to read.
After that we can read csv file by pd.read_csv(file path)



Read Data from CSV Files to Pandas Dataframes

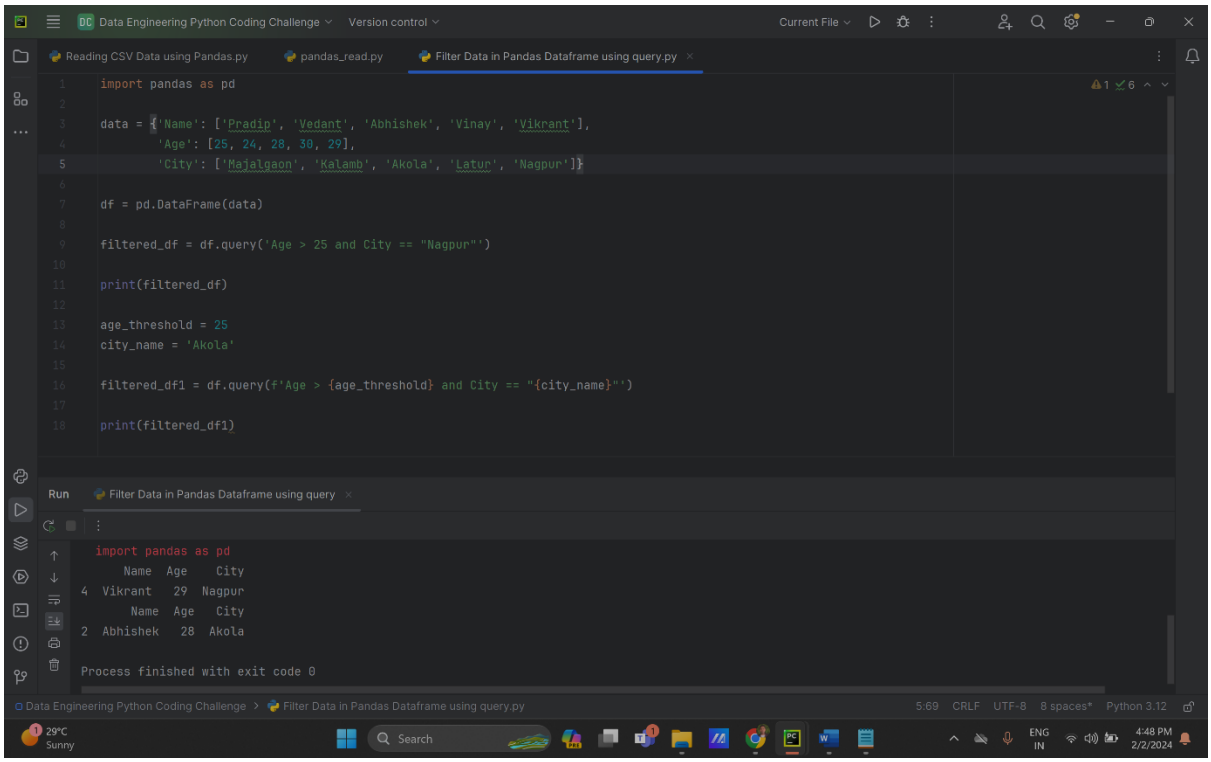


Result



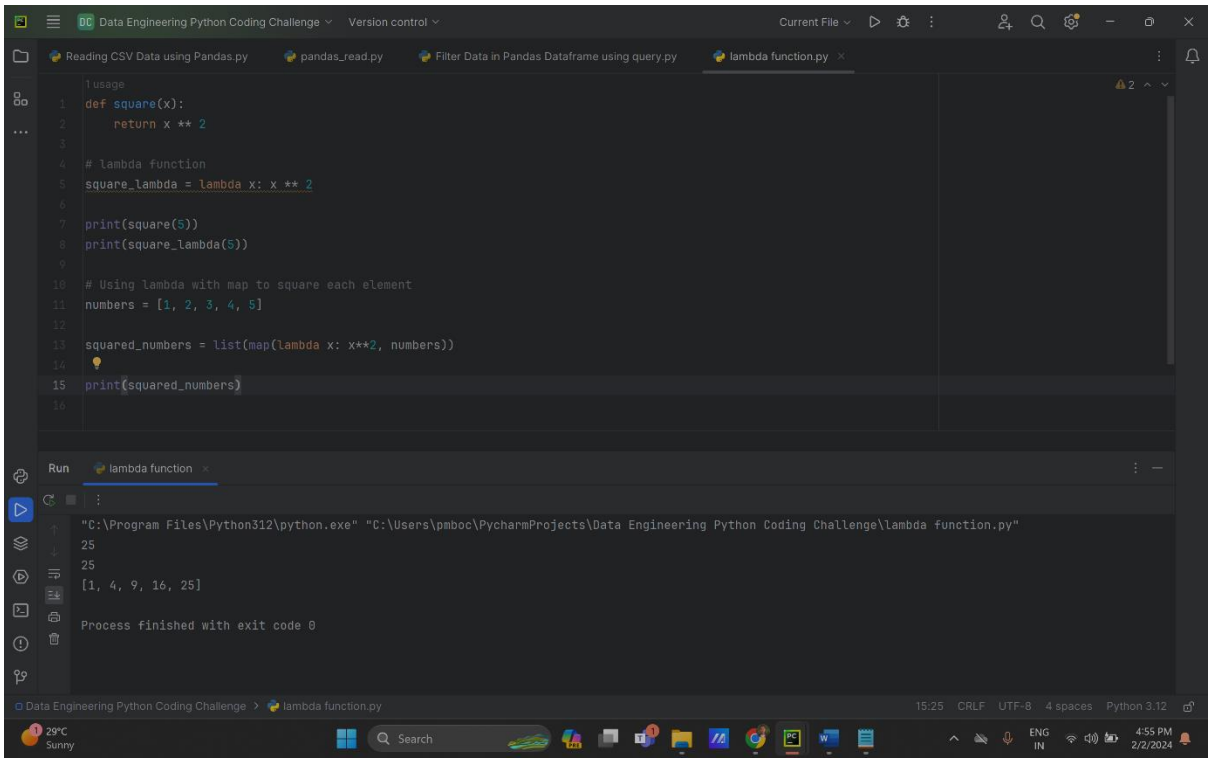
Filter Data in Pandas Dataframe using query.

In pandas we can filter Dataframe using ‘query ()’ method which allows you to express complex filtering conditions using SQL like syntax



Execute with one example Lambda Functions in Python

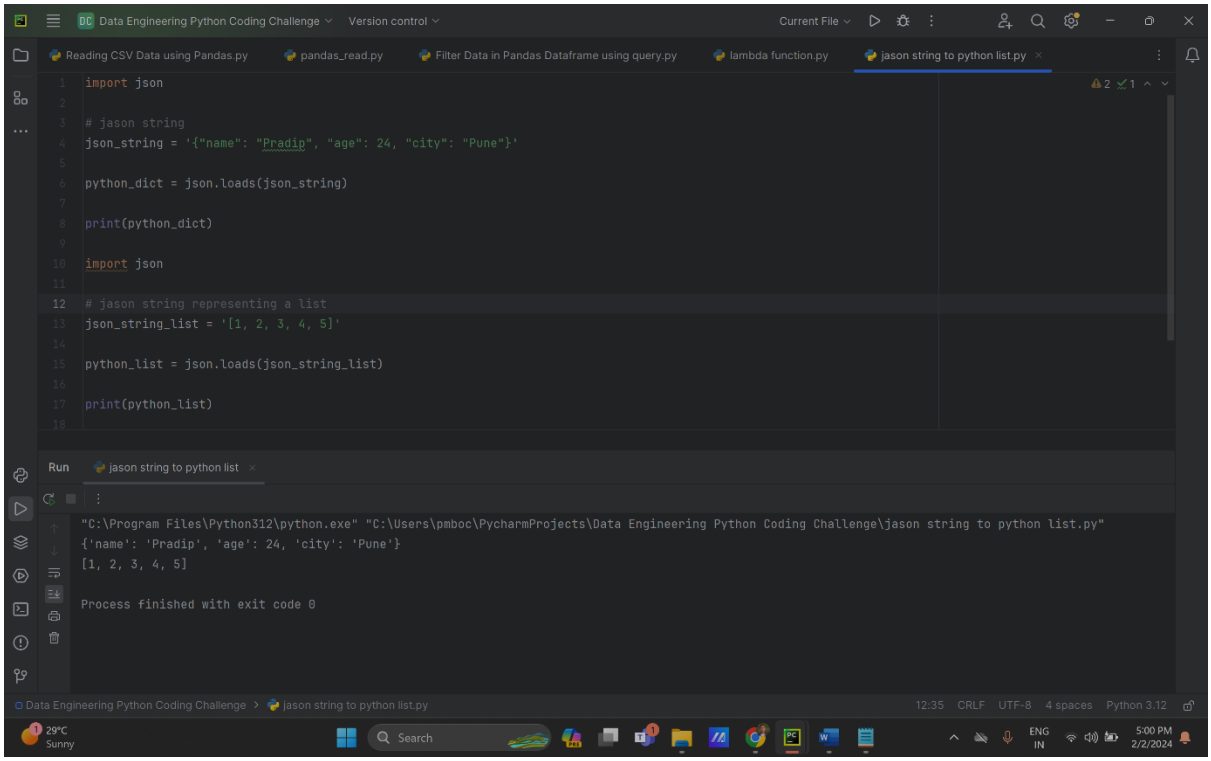
A lambda function is a small anonymous function.
A lambda function can take any number of arguments, but can only have one expression.



Read JSON Strings to Python dicts or lists

In python we can use ‘json’ module to convert JSON strings to python dictionaries or lists. This module provides methods for encoding and decoding JSON data.

‘json.loads()’ is used to parse the json string into python dictionary or list. Resulting python_dict variable contains the equivalent python data structure.



❖ Instead of json string we can use JSON file also here we use ‘json.laod’ function to read directly from file