Acta Scientific COMPUTER SCIENCES

Volume 4 Issue 3 March 2022

Research Article

Data Analysis Using Pandas Library of Python

Rupal Snehkunj[1]* and Khushboo Vachiyatwala[2]

[1]*Department of Computer Science, Sarvajanik University, India*

[2]*Department of Computer Science, VNSG University, India*

***Corresponding Author:** Rupal Snehkunj, Department of Computer Science, Sarvajanik University, India.

Received: November 25, 2021

Published: February 25, 2022

Abstract

This research paper mainly focuses on usage of Pandas library of python. This rich library provides various integrated support for analysis of data. It is useful for grouping queries, graphical design of data in tabular format. This library is foundational layer for future statistical computing of data in python through various Pandas API. The work is researched with structure data set file accessing various formats as xls, csv, pdf and many more. The work is implemented on randomly created employee database for performing various operations and data visualization in Python using pandas library.

Keywords: Pandas; NumPy; Matplotlib; Scipy

Introduction

The *P*ython panda is used to work with data structure in efficient manner. It supports to Matplotlib, NumPy library. Matplotlib library is use for graphical performance and NumPy library is use only performance to numerical array. The Pandas library provides better platform for data analytics and statistical computing. Pandas library also has support for SQL tools for data manipulation such as merge of data using various joins (inner, right and left). Pandas library name was evolved from Panel data analytics which aims to provide equivalent functionality and has implemented many features such as automatic data alignment and hierarchical indexing.

Pandas library is modern object oriented high level programming library which contain large collection of add-on-package. Pandas have inbuilt library called Numpy use for numerical data. Numpy works with arraydata type for the operations such as indexing, sorting, reshaping etc. Numpy supports homogenous data due to array datatype. Pandas also supports Matplotlib library for displaying data in graphical format and allows to save files in excel, csv,json and many more.

Pandas support mainly three type of data structure: (1) Series (2) Data Frame (3) Panel. Series data structure contains one dimensional array. It supports homogenous type of data. Data Frame data structure contains two-dimensional array and supports heterogeneous type of data and also use to size and data mutable. Panel contains three-dimensional array.

Literature Review

Stančin., *et al*. [1] considers more than 20 libraries and separate them into six groups: core libraries, data preparation, data visualization, machine learning, deep learning and big data. The authors recommends the libraries such as pandas for data preparation; Matplotlib, seaborn or Plotly for data visualization; scikit-learn for machine leraning; TensorFlow, Keras and PyTorch for deep learning; and Hadoop Streaming and PySpark for big data. McKinney., *et al*. [2] described that Pandas library provide labelled and structure based data for grouping and aggregation of data. This paper focused on how to contain multiple tables in each to other. Kumar., *et al*. [3] paper focused on Python how to work NumPy (numerical array). It explained how to reshape of numerical array and also

how to numerical data in display in graphical way. Hoyer., *et al*. [4] revealed about python panda's library increase performance of label and structure based data and files. It is use to grouping of records. Mitrpanont., *et al*. [5] Python panda's library use to data analysis as well as weka. Weka also work on data manipulation and data analysis. In this Panda's Data Frame function set any time of data like json format data set in tabular format using function. It's store in structure format. Python is support Data science using any type of extension file call in tabular or structure format data and search top and bottom data. Panda's support Matplotlib library it's support graphics and 3-D animation. It's use files data display in graph format [6-8]. Van Der Walt., *et al*. [9]. NumPy is inbuilt in panda's. Any numerical work in panda's also use of NumPy library. Panda's library provides files and API format data in tabular or structure format data list. Sessa., *et al*. [10] deals with the real data with missing values. Panda's library provides fill missing data of -files, database and data frame. The author considers three phases : Feature selection, Filling the missing values and Correcting the missing values that have been filled in. The result reveals that that both imputation methods are efficient and yield more or less the same accuracy.

Proposed work

The proposed work is undertaken in Python Pandas library. The work is researched with structure data set file accessing various formats as xls, csv, pdf and many more. The work is implemented on randomly created employee database for performing various operations. In this work, the manually created structure data set will be used and visualize in Python using various pandas libraries.

Implementation

The current work is done by creating employee database manually to work with various features of python panda's library.

Structure data set

Pandas library provide data frame to create two dimensional structured dataset for storing the data as seen in figure. Data frame collects heterogenous data with data size and value mutable and displayed in tabular data format.

```
>>> import pandas as pd
>>> d=[["khushbu",22,25000],["Nikhil",27,70000],["Harshil",19,15000]]
>>> f=pd.DataFrame(d,columns=["Name","Age","Salary"])
>>> print(f)
      Name  Age  Salary
0  khushbu   22   25000
1   Nikhil   27   70000
2  Harshil   19   15000
>>>
```

```
>>> import pandas as pd
>>> d=[["khushbu",22,25000],["Nikhil",27,70000],["Harshil",19,15000],["Meet",25,17000],["Kiran",27,25000],["Kinjal",24,22000]]
>>> f=pd.DataFrame(d,columns=["Name","Age","Salary"])
>>> print(f)
      Name  Age  Salary
0  khushbu   22   25000
1   Nikhil   27   70000
2  Harshil   19   15000
3     Meet   25   17000
4    Kiran   27   25000
5   Kinjal   24   22000
>>> print("Rows where age between 15 and 25 (inclusive):")
Rows where age between 15 and 25 (inclusive):
>>> print(f[f["Age"].between(15, 25)])
      Name  Age  Salary
0  khushbu   22   25000
2  Harshil   19   15000
3     Meet   25   17000
5   Kinjal   24   22000
```

Figure 1: Structured dataset depicts the employee table data with query of age between 15 to 25.

File access in pandas

Pandas library work with different type of files with extension like. xlsx, csv, pdf format. The below shown practical import excel file in pandas and display the content of file in structured format using pandas frame class library.

To also add another important feature of python panda's library to search top ten and bottom ten record from database is display in figure 3.

```
>>> import pandas as pd
>>> import numpy as np
>>> df = pd.read_excel('C:\Workbook.xlsx')
>>> print(df)
    Sr no       Name      Last name  Age  Gender  Salary
0       1    Khushbu  Vachiyatwala   22  Female   25000
1       2     Nikhil          Iwala   27    Male   70000
2       3   Anupriya          Patel   22  Female   18000
3       4    Vandana        Tekwani   23  Female   15000
4       5    Pragati      Chaudhari   22  Female   10000
5       6      Hiren          Patel   22    Male   17000
6       7      Nilam        Tekwani   22  Female   18000
7       8      Kajal          Patel   25  Female   19000
8       9      Pooja          Patel   25  Female   10000
9      10      Nimita       Rasaniya   22  Female   15000
10     11    Ashmita         Ariwala   27  Female   14000
11     12      Khushi      Lakdawala   25  Female   13000
12     13       Divya          Patel   28  Female   14000
13     14   Narayani           Patel   21  Female   15000
14     15       Asha          Patel   23  Female   25000
15     16       Piya        tekwani   26  Female   27000
16     17      Pinki        Panigar   25  Female   25000
17     18     Vaishu         Jagtap   24  Female   65000
18     19     Sheema          Patel   25  Female   25000
19     20       Nita          Patel   29  Female   22000
20     21       Raxa          Patel   20  Female   24000
21     22       Daxa          Patel   19  Female   26000
22     23       Jiva          Patel   22  Female   28000
23     24      Niyati        Panigar   24  Female   12000
24     25        Het  Vachiyatwala   25    Male   15000
25     26       Niya  Vachiyatwala   26    Male   17000
26     27    Nishant       Rasaniya   27    Male   25000
27     28       Meet         Ariwala   28    Male   24000
```

	Name	Last name	Age	Gender	Salary
1	Khushbu	Vachiyatwal	22	Female	25000
2	Nikhil	Iwala	27	Male	70000
3	Anupriya	Patel	22	Female	18000
4	Vandana	Tekwani	23	Female	15000
5	Pragati	Chaudhari	22	Female	10000
6	Hiren	Patel	22	Male	17000
7	Nilam	Tekwani	22	Female	18000
8	Kajal	Patel	25	Female	19000
9	Pooja	Patel	25	Female	10000
10	Nimita	Rasaniya	22	Female	15000
11	Ashmita	Ariwala	27	Female	14000
12	Khushi	Lakdawala	25	Female	13000
13	Divya	Patel	28	Female	14000
14	Narayani	Patel	21	Female	15000
15	Asha	Patel	23	Female	25000
16	Piya	tekwani	26	Female	27000
17	Pinki	Panigar	25	Female	25000
18	Vaishu	Jagtap	24	Female	65000
19	Sheema	Patel	25	Female	25000
20	Nita	Patel	29	Female	22000
21	Raxa	Patel	20	Female	24000
22	Daxa	Patel	19	Female	26000

Figure 2: Reading the Excel file in Python having multiple data and that data read Data Frame format.

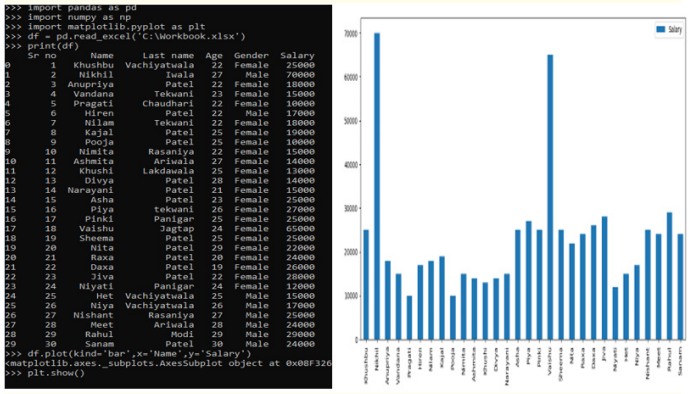

Figure 4: Plotting the file graphically.

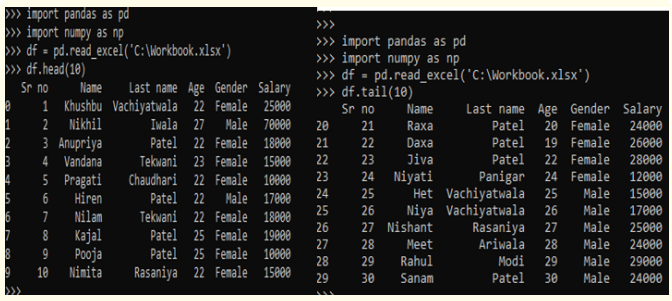

Figure 3: library to search top ten and bottom ten record from database.

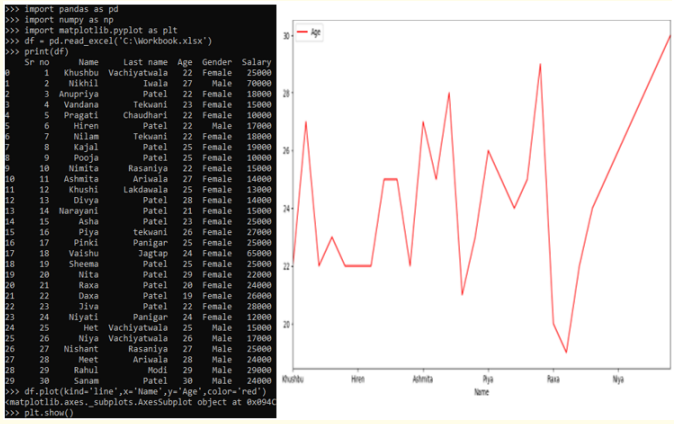

Figure 5: Plotting the content graphically.

Graphics using panda's library

Panda's library provides graphics using Matplotlib library. Matplotlib use to create 2D graphs and plot using script. It's supported many types of graph bar, chart, line etc. Matplotlib also read files extension and using their record display in graphical way in show figure 4.

Matplotlib library also draw line graph use of function "kind" and also change the graph color use of function "color" as show figure 5.

Merge with pandas

Python panda's library provide feature to work with multiple table using join function. In doing so associating observations from one data set with another via a merge key of some kind. For similarly 2D data. The row returns as join function. The figure 9 highlighted on merge two table and their result.

API with pandas

Pandas brings a robust, full –featured and Interred data analysis toolset of python. In python use panda's we will simply and easily work on API.API through json format data can easily get in tabular format using panda's library.

The Figure highlighted on API through json format data display in Data structure format.

```
>>> import pandas as pd
>>> left = {
...     'id':[1,2,3,4,5],
...     'Name': ["khushbu","Nikhil","Harshil","kinjal","meet"],
...     'department_id':[1,2,3,2,2]}
>>> s=pd.DataFrame(left)
>>> print(s)
   id     Name  department_id
0   1  khushbu              1
1   2   Nikhil              2
2   3  Harshil              3
3   4   kinjal              2
4   5     meet              2
>>> right ={
...     'department_id':[1,2,3],
...     'department_name': ["software","hardware","it"]
...     }
>>> d=pd.DataFrame(right)
>>> print(d)
   department_id department_name
0              1        software
1              2        hardware
2              3              it
>>> print (pd.merge(s, d, on='department_id', how='inner'))
   id     Name  department_id department_name
0   1  khushbu              1        software
1   2   Nikhil              2        hardware
2   4   kinjal              2        hardware
3   5     meet              2        hardware
4   3  Harshil              3              it
>>>
```

Figure 6: Merging of two tables.

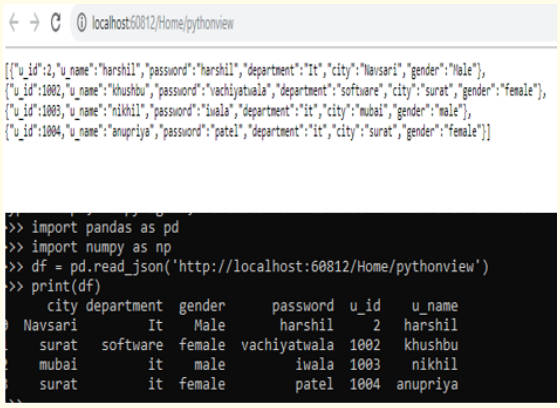

Figure 7: Working with json content.

```
>>> import pandas as pd
>>> d={"col1":[1,2,3],"col2":[4,5,6],"col3":[7,8,9]}
>>> s=pd.DataFrame(d)
>>> print(s)
   col1  col2  col3
0     1     4     7
1     2     5     8
2     3     6     9
>>> print("change dataframe")
change dataframe
>>> s=s[["col3","col2","col1"]]
>>> print(s)
   col3  col2  col1
0     7     4     1
1     8     5     2
2     9     6     3
>>>
>>>
```

Figure 8: Reshaping the data.

```
>>> import pandas as pd
>>> import numpy as np
>>> df = pd.read_excel('C:\Workbook.xlsx')
>>> print(df)
    Sr no      Name     Last name  Age  Gender  Salary
0       1    Khushbu  Vachiyatwala   22  Female   25000
1       2     Nikhil         Iwala   27    Male   70000
2       3   Anupriya         Patel   22  Female   18000
3       4    Vandana       Tekwani   23  Female   15000
4       5    Pragati     Chaudhari   22  Female   10000
5       6      Hiren         Patel   22    Male   17000
6       7      Nilam       Tekwani   22  Female   18000
7       8      Kajal         Patel   25  Female   19000
8       9      Pooja         Patel   25  Female   10000
9      10     Nimita      Rasaniya   22  Female   15000
10     11    Ashmita       Ariwala   27  Female   14000
11     12     Khushi     Lakdawala   25  Female   13000
12     13      Divya         Patel   28  Female   14000
13     14   Narayani         Patel   21  Female   15000
14     15       Asha         Patel   23  Female   25000
15     16       Piya       tekwani   26  Female   27000
16     17      Pinki       Panigar   25  Female   25000
17     18     Vaishu        Jagtap   24  Female   65000
18     19     Sheema         Patel   25  Female   25000
19     20       Nita         Patel   29  Female   22000
20     21       Raxa         Patel   20  Female   24000
21     22       Daxa         Patel   19  Female   26000
22     23       Jiva         Patel   22  Female   28000
23     24     Niyati       Panigar   24  Female   12000
24     25        Het  Vachiyatwala   25    Male   15000
25     26       Niya  Vachiyatwala   26    Male   17000
26     27    Nishant      Rasaniya   27    Male   25000
27     28       Meet       Ariwala   28    Male   24000
28     29      Rahul         Modi   29    Male   29000
29     30      Sanam         Patel   30    Male   24000
>>> print (df.groupby(['Gender']).groups)
{'Female': Int64Index([0, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20,
            21, 22, 23],
           dtype='int64'), 'Male': Int64Index([1, 5, 24, 25, 26, 27, 28, 29], dtype='int64')}
>>>
```

Figure 9: Grouping of data.

Other features of pandas

Use Pandas create multidimensional array so easily read, write and search the array value.Join two data frames using "merge". Easily integrated to graphical library. Any API call to pandas and that API data read in structured format. Easily data integrated, grouping and manipulation of data. Pandas provide resize of data.

Results and Discussion

We have believed that in coming years' great opportunity to analysis tools for development pandas is first chosen because of Matplotlib support, API integrated, merge data, easily data analysis. Pandas is very powerful data analysis, low-cost application. In this paper, the manually created structure data set was plotted in Python using various pandas libraries.

Conclusion

This research focused on various functionalities of Pandas library of Python. This library is foundational layer for data analytics and statistical computing. Pandas library offers data structures and operations for manipulating numerical tables and time series. The research worked with manual created structure data set that was plotted ,analyzed and visualized in Python using various pandas libraries.

Bibliography

1. Stančin Igor and Alan Jović. "An overview and comparison of free Python libraries for data mining and big data analysis". 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE (2019).

2. McKinney Wes. "Pandas: a foundational Python library for data analysis and statistics". *Python for High Performance and Scientific Computing* 14.9 (2011).

3. Kumar Rakesh. "Future for scientific computing using Python". *International Journal of Engineering Technologies and Management Research* 2 (2015): 30-41.

4. Hoyer Stephan and Joe Hamman. "xarray: ND labeled arrays and datasets in Python". *Journal of Open Research Software* 5.1 (2017).

5. Mitrpanont Jarernsri., *et al*. "A study on using Python vs Weka on dialysis data analysis". 2017 2nd International Conference on Information Technology (INCIT). IEEE (2017).

6. https://rabernat.github.io/research_computing/pandas.html

7. https://www.simplilearn.com/why-python-is-essential-for-data-analysis-article

8. https://towardsdatascience.com/a-guide-to-pandas-and-matplotlib-for-data-exploration-56fad95f951c

9. S van der Walt., *et al*. "The NumPy Array: A Structure for Efficient Numerical Computation". in Computing in Science and Engineering 13.2 (2011): 22-30.

10. Sessa Jadran and Dabeeruddin Syed. "Techniques to deal with missing data". 2016 5th international conference on electronic devices, systems and applications (ICEDSA). IEEE (2016).