Data Analysis With Python:-

Data Analysis is the technique of collecting, transforming, and organizing data to make future predictions and informed data-driven decisions. It also helps to find possible solutions for a business problem. There are six steps for Data Analysis. They are:

- Ask or Specify Data Requirements
- Prepare or Collect Data
- Clean and Process
- Analyze
- Share
- Act or Report

The collection, transformation, and organization of data to draw conclusions make predictions for the future, and make informed data-driven decisions is called **Data Analysis.** The profession that handles data analysis is called a **Data Analyst.** There is a huge demand for Data Analysts as the data is expanding rapidly nowadays. Data Analysis is used to find possible solutions for a business problem. The advantage of being a Data Analyst is that they can work in any field they love: healthcare, agriculture, IT, finance, business. Data-driven decision-making is an important part of Data Analysis. It makes the analysis process much easier. There are six steps for Data Analysis. They are:

1. **Ask or Specify Data Requirements**
2. **Prepare or Collect Data**
3. **Clean and Process**
4. **Analyze**
5. **Share**
6. **Act or Report**

Each step has its own process and tools to make overall conclusions based on the data.

1. Ask

The first step in the process is to **Ask**. The data analyst is given a problem/business task. The analyst has to understand the task and the stakeholder's expectations for the solution. A stakeholder is a person that has invested their money and resources to a project. The analyst must be able to ask different questions in order to find the right solution to their problem. The analyst has to find the root cause of the problem in order to fully understand the problem. The analyst must make sure that he/she doesn't have any distractions while analyzing the problem. Communicate effectively with the stakeholders and other colleagues to completely understand what the underlying problem is. Questions to ask yourself for the Ask phase are:

- What are the problems that are being mentioned by my stakeholders?
- What are their expectations for the solutions?

2. Prepare

The second step is to **Prepare or Collect the Data.** This step includes collecting data and storing it for further analysis. The analyst has to collect the data based on the task given from multiple sources. The data has to be collected from various sources, internal or external sources. Internal data is the data available in the organization that you work for while external data is the data available in sources other than your organization. The data that is collected by an individual from their own resources is called first-party data. The data that is collected and sold is called second-party data. Data that is collected from outside sources is called third-party data. The common sources from where the data is collected are Interviews, Surveys, Feedbacks, Questionnaires. The collected data can be stored in a spreadsheet or SQL database.

A spreadsheet is a digital worksheet that contains rows and columns while a database contains tables that have functions to manipulate the data. Spreadsheets are used to store some thousands or ten thousand of data while databases are used when there are too many rows to store. The best tools to store the data are MS Excel or Google Sheets in the case of Spreadsheets and there are so many databases like Oracle, Microsoft to store the data.

3. Clean and Process Data

The third step is **Process**. After the data is collected from multiple sources, it is time to **clean** the data. Clean data means data that is free from misspellings, redundancies, and irrelevance. Clean data largely depends on data integrity. There might be duplicate data or the data might not be in a format, therefore the unnecessary data is removed and cleaned. There are different functions provided by SQL and Excel to clean the data. This is one of the most important steps in Data Analysis as clean and formatted data helps in finding trends and solutions. The most important part of the Process phase is to check whether your data is biased or not. Bias is an act of favoring a particular group/community while ignoring the rest. Biasing is a big no-no as it might affect the overall data analysis. The data analyst must make sure to include every group while the data is being collected.

4. Analyze

The fourth step is to **Analyze**. The cleaned data is used for analyzing and identifying trends. It also performs calculations and combines data for better results. The tools used for performing calculations are Excel or SQL. These tools provide in-built functions to perform calculations or sample code is written in SQL to perform calculations. Using Excel, we can create pivot tables and perform calculations while SQL creates temporary tables to perform calculations. Programming languages are another way of solving problems. They make it much easier to solve problems by providing packages. The most widely used programming languages for data analysis are R and Python.

5. Share

The fifth step is Share. Nothing is more compelling than a visualization. The data now transformed has to be made into a visual(chart, graph). The reason for making data visualizations is that there might be people, mostly stakeholders that are non-technical. Visualizations are made for a simple understanding of complex data. Tableau and Looker are the two popular tools used for compelling data visualizations. Tableau is a simple drag and drop tool that helps in creating compelling visualizations. Looker is a data viz tool that directly connects to the database and creates visualizations. Tableau and Looker are both equally used by data analysts for creating a visualization. R and Python have some packages that provide beautiful data visualizations. R has a package named ggplot which has a variety of data visualizations. A presentation is given based on the data findings. Sharing the insights with the team members and stakeholders will help in making better decisions. It helps in making more informed decisions and it leads to better outcomes.

6. Act or Report

The final/sixth step is Act. After a presentation is given based on your findings, the stakeholders discuss whether to move forward or not. If they agreed to your recommendations, they move further with your solutions. If they don’t agree with your findings, you will have to dig deeper to find more possible solutions. Every step has to be re-organized. We have to repeat every step to see whether there are any gaps in there. The data collected must be reviewed to see if there is any bias and identify options. After the gaps are identified and the data is analyzed, a presentation is given again

Analyzing Data Using Pandas :-

Python Pandas Is used for relational or labeled data and provides various data structures for manipulating such data and time series. This library is built on top of the NumPy library. This module is generally imported as:

```
import pandas as pd
```

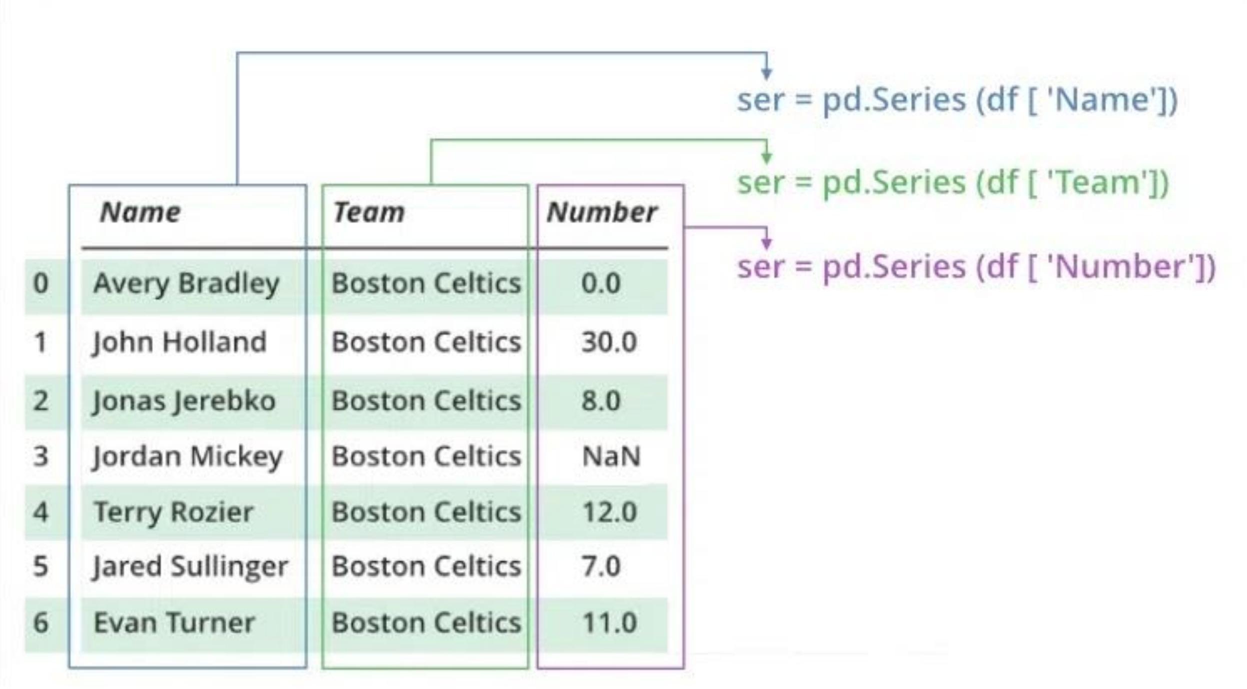
Here, pd is referred to as an alias to the Pandas. However, it is not necessary to import the library using the alias, it just helps in writing less amount code every time a method or property is called. Pandas generally provide two data structures for manipulating data,

They are:-

- Series
- Dataframe

Series: -

Pandas Series is a one-dimensional labeled array capable of holding data of any type (integer, string, float, python objects, etc.).
The axis labels are collectively called indexes.
Pandas Series is nothing but a column in an excel sheet.
Labels need not be unique but must be a hashable type. The object supports both integer and label-based indexing and provides a host of methods for performing operations involving the index.



```
ser = pd.Series (df [ 'Name'])
ser = pd.Series (df [ 'Team'])
ser = pd.Series (df [ 'Number'])
```

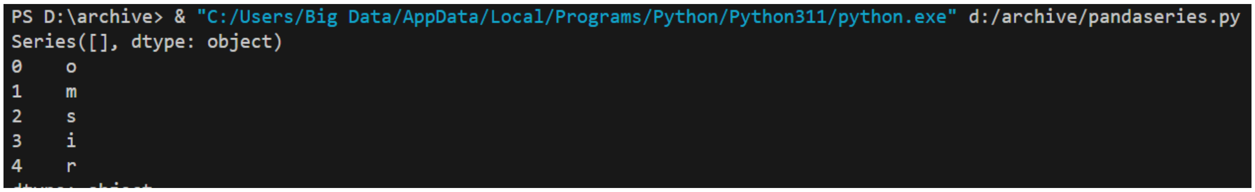
	Name	*Team*	*Number*
0	Avery Bradley	Boston Celtics	0.0
1	John Holland	Boston Celtics	30.0
2	Jonas Jerebko	Boston Celtics	8.0
3	Jordan Mickey	Boston Celtics	NaN
4	Terry Rozier	Boston Celtics	12.0
5	Jared Sullinger	Boston Celtics	7.0
6	Evan Turner	Boston Celtics	11.0

it can be created using the Series() function by loading the dataset from the existing storage like SQL, Database, CSV Files, Excel Files, etc., or from data structures like lists, dictionaries, etc.

Python Pandas Creating Series

```python
import pandas as pd
import numpy as np


# Creating empty series
ser = pd.Series()

print(ser)

# simple array
data = np.array(['o', 'm', 's', 'i', 'r'])

ser = pd.Series(data)
print(ser)
```

Output:-



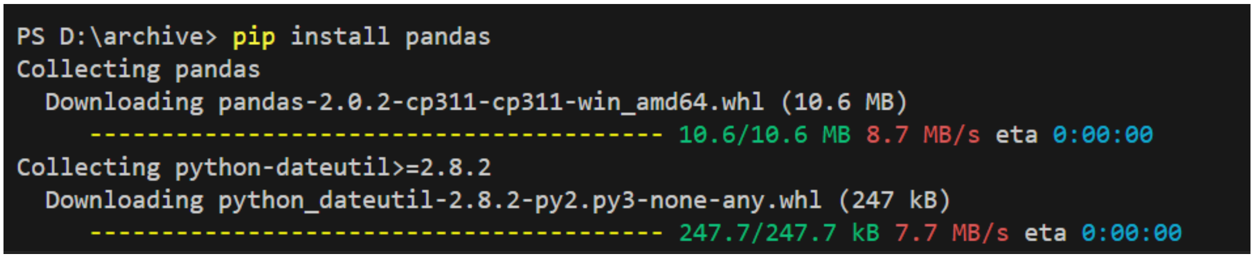
```
PS D:\archive> & "C:/Users/Big Data/AppData/Local/Programs/Python/Python311/python.exe" d:/archive/pandaseries.py
Series([], dtype: object)
0    o
1    m
2    s
3    i
4    r
dtype: object
```

Dataframe:-

Pandas DataFrame is a two-dimensional size-mutable, potentially heterogeneous tabular data structure with labeled axes (rows and columns). A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns. Pandas DataFrame consists of three principal components, the data, rows, and columns.

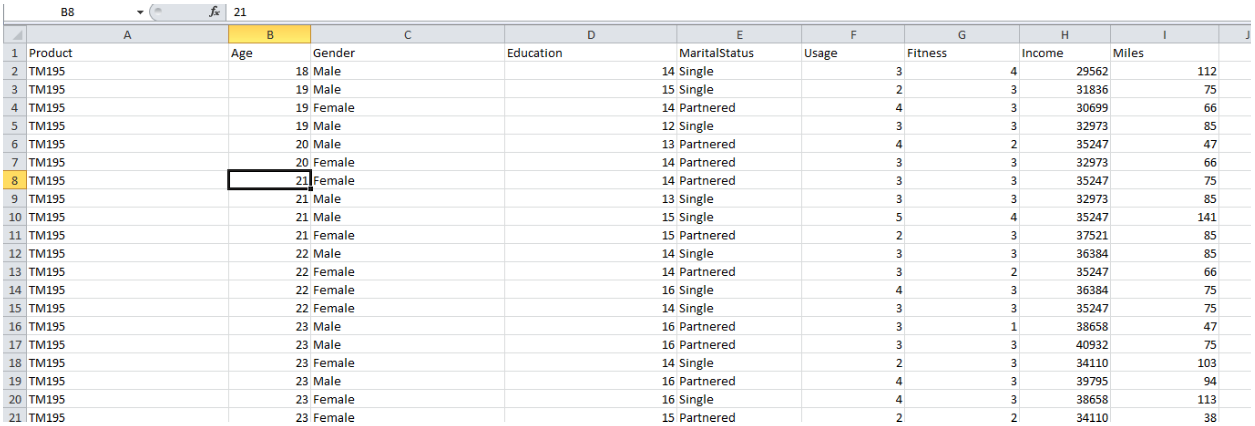
Creating a dataframe using CSV files

First of install pandas: -



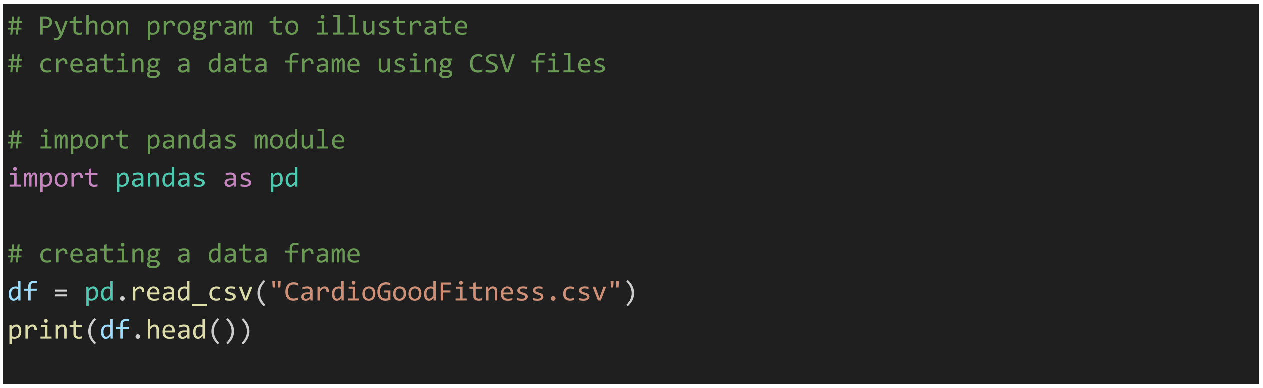
```
PS D:\archive> pip install pandas
Collecting pandas
  Downloading pandas-2.0.2-cp311-cp311-win_amd64.whl (10.6 MB)
     ---------------------------------------- 10.6/10.6 MB 8.7 MB/s eta 0:00:00
Collecting python-dateutil>=2.8.2
  Downloading python_dateutil-2.8.2-py2.py3-none-any.whl (247 kB)
     ---------------------------------------- 247.7/247.7 kB 7.7 MB/s eta 0:00:00
```

Csv data file CardioGoodFitness.csv :-



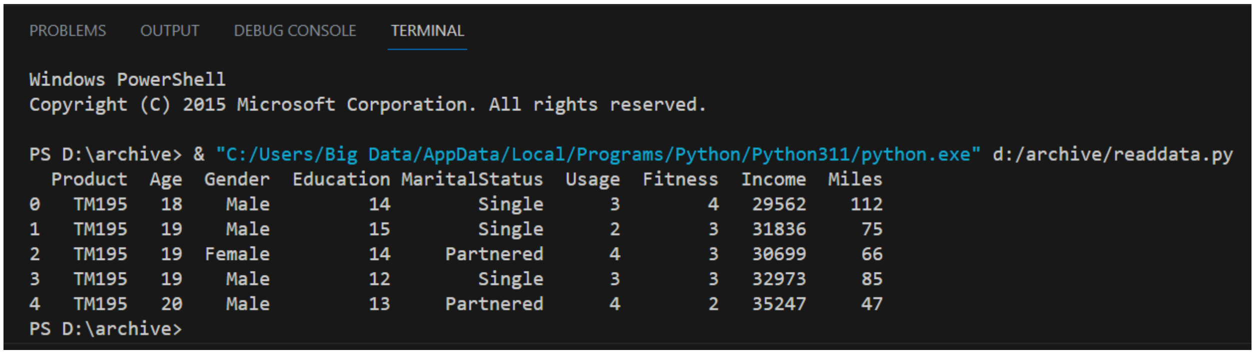
	A	B	C	D	E	F	G	H	I
1	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
2	TM195	18	Male	14	Single	3	4	29562	112
3	TM195	19	Male	15	Single	2	3	31836	75
4	TM195	19	Female	14	Partnered	4	3	30699	66
5	TM195	19	Male	12	Single	3	3	32973	85
6	TM195	20	Male	13	Partnered	4	2	35247	47
7	TM195	20	Female	14	Partnered	3	3	32973	66
8	TM195	21	Female	14	Partnered	3	3	35247	75
9	TM195	21	Male	13	Single	3	3	32973	85
10	TM195	21	Male	15	Single	5	4	35247	141
11	TM195	21	Female	15	Partnered	2	3	37521	85
12	TM195	22	Male	14	Single	3	3	36384	85
13	TM195	22	Female	14	Partnered	3	2	35247	66
14	TM195	22	Female	16	Single	4	3	36384	75
15	TM195	22	Female	14	Single	3	3	35247	75
16	TM195	23	Male	16	Partnered	3	1	38658	47
17	TM195	23	Male	16	Partnered	3	3	40932	75
18	TM195	23	Female	14	Single	2	3	34110	103
19	TM195	23	Male	16	Partnered	4	3	39795	94
20	TM195	23	Female	16	Single	4	3	38658	113
21	TM195	23	Female	15	Partnered	2	2	34110	38

Write code for readdata.py file:-



```python
# Python program to illustrate
# creating a data frame using CSV files

# import pandas module
import pandas as pd

# creating a data frame
df = pd.read_csv("CardioGoodFitness.csv")
print(df.head())
```

Output:-



```
Windows PowerShell
Copyright (C) 2015 Microsoft Corporation. All rights reserved.

PS D:\archive> & "C:/Users/Big Data/AppData/Local/Programs/Python/Python311/python.exe" d:/archive/readdata.py
  Product  Age  Gender  Education MaritalStatus  Usage  Fitness  Income  Miles
0   TM195   18    Male         14        Single      3        4   29562    112
1   TM195   19    Male         15        Single      2        3   31836     75
2   TM195   19  Female         14     Partnered      4        3   30699     66
3   TM195   19    Male         12        Single      3        3   32973     85
4   TM195   20    Male         13     Partnered      4        2   35247     47
PS D:\archive>
```

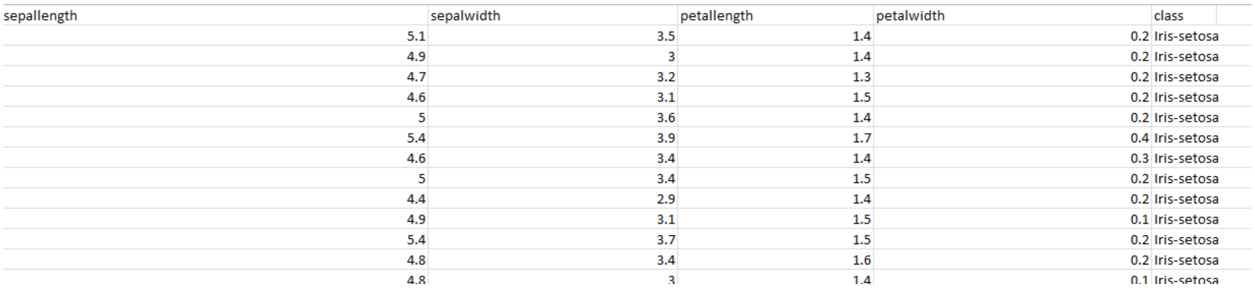
Filtering DataFrame

Pandas dataframe.filter() function is used to Subset rows or columns of dataframe according to labels in the specified index. Note that this routine does not filter a dataframe on its contents. The filter is applied to the labels of the index.

Python Pandas Filter Dataframe

Csv file :-

```
iris_csv.csv
```

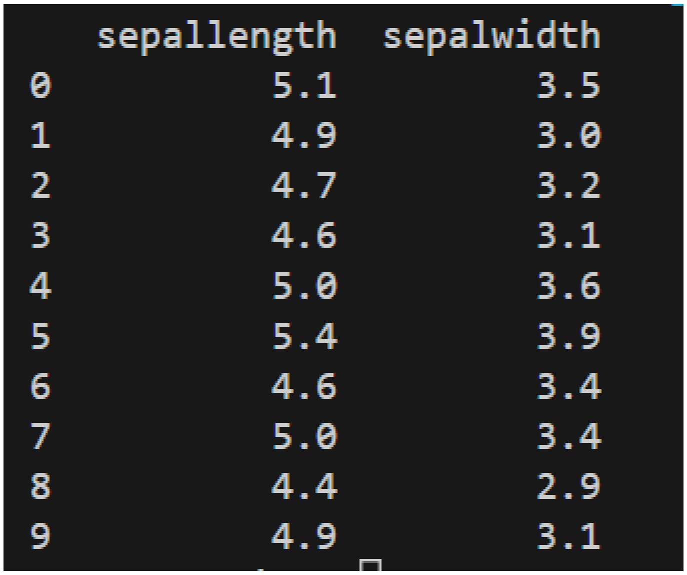


sepallength	sepalwidth	petallength	petalwidth	class
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5	3.6	1.4	0.2	Iris-setosa
5.4	3.9	1.7	0.4	Iris-setosa
4.6	3.4	1.4	0.3	Iris-setosa
5	3.4	1.5	0.2	Iris-setosa
4.4	2.9	1.4	0.2	Iris-setosa
4.9	3.1	1.5	0.1	Iris-setosa
5.4	3.7	1.5	0.2	Iris-setosa
4.8	3.4	1.6	0.2	Iris-setosa
4.8	3	1.4	0.1	Iris-setosa

For example we want to display only 2 columns sepalllength and sepalwidth :-

```
import pandas as pd

df = pd.read_csv("iris_csv.csv")


newdf = df.filter(["sepallength", "sepalwidth"]).head(10)

print(newdf)
```

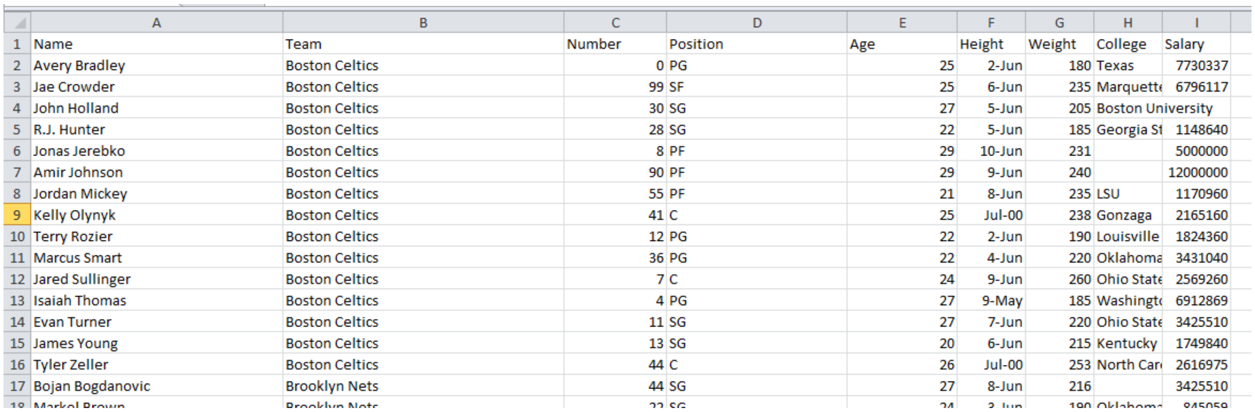
Output :-



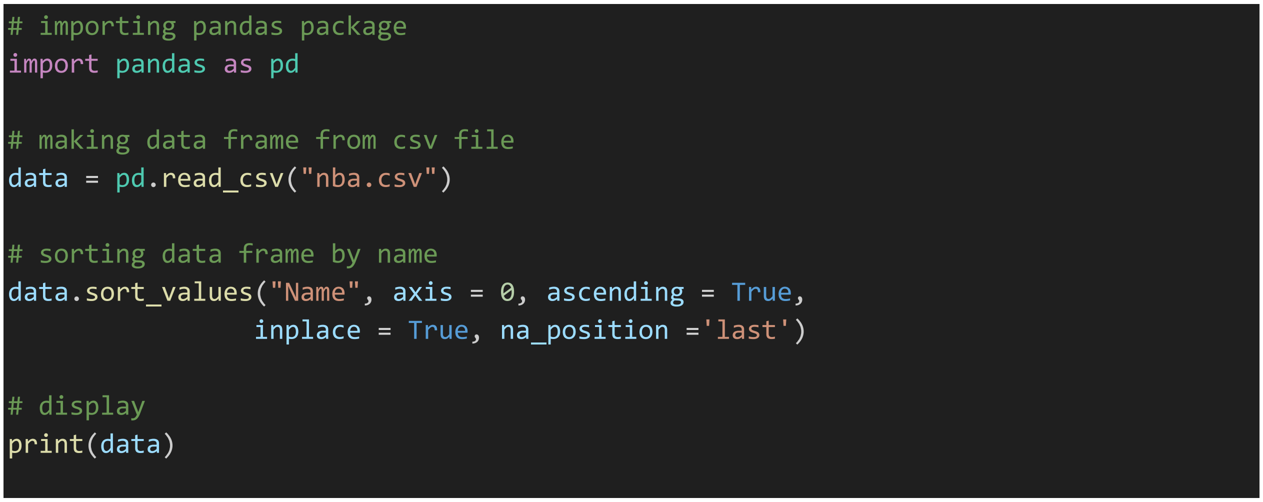
```
   sepallength  sepalwidth
0          5.1         3.5
1          4.9         3.0
2          4.7         3.2
3          4.6         3.1
4          5.0         3.6
5          5.4         3.9
6          4.6         3.4
7          5.0         3.4
8          4.4         2.9
9          4.9         3.1
```

In the following example, A data frame is made from the csv file and the data frame is sorted in ascending order of Names of Players

nba.csv file :-

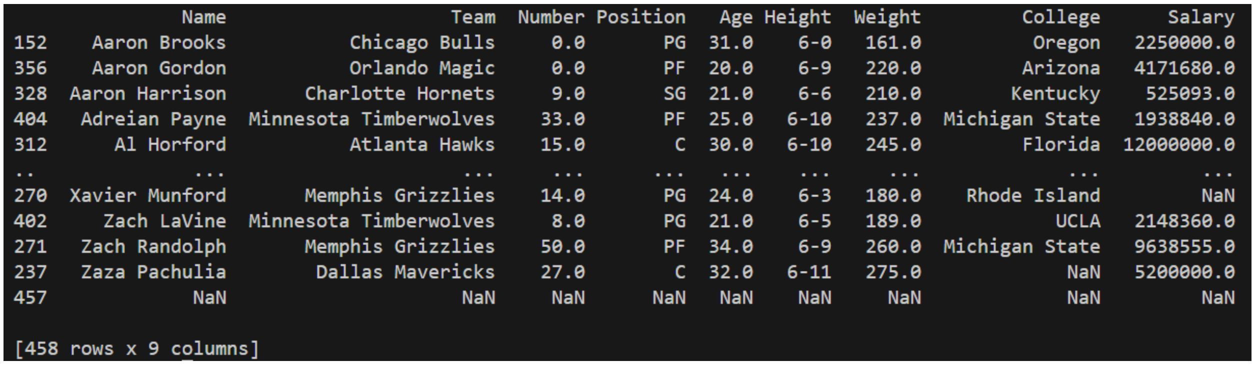


	A	B	C	D	E	F	G	H	I
1	Name	Team	Number	Position	Age	Height	Weight	College	Salary
2	Avery Bradley	Boston Celtics	0	PG	25	2-Jun	180	Texas	7730337
3	Jae Crowder	Boston Celtics	99	SF	25	6-Jun	235	Marquette	6796117
4	John Holland	Boston Celtics	30	SG	27	5-Jun	205	Boston University	
5	R.J. Hunter	Boston Celtics	28	SG	22	5-Jun	185	Georgia St	1148640
6	Jonas Jerebko	Boston Celtics	8	PF	29	10-Jun	231		5000000
7	Amir Johnson	Boston Celtics	90	PF	29	9-Jun	240		12000000
8	Jordan Mickey	Boston Celtics	55	PF	21	8-Jun	235	LSU	1170960
9	Kelly Olynyk	Boston Celtics	41	C	25	Jul-00	238	Gonzaga	2165160
10	Terry Rozier	Boston Celtics	12	PG	22	2-Jun	190	Louisville	1824360
11	Marcus Smart	Boston Celtics	36	PG	22	4-Jun	220	Oklahoma	3431040
12	Jared Sullinger	Boston Celtics	7	C	24	9-Jun	260	Ohio State	2569260
13	Isaiah Thomas	Boston Celtics	4	PG	27	9-May	185	Washingto	6912869
14	Evan Turner	Boston Celtics	11	SG	27	7-Jun	220	Ohio State	3425510
15	James Young	Boston Celtics	13	SG	20	6-Jun	215	Kentucky	1749840
16	Tyler Zeller	Boston Celtics	44	C	26	Jul-00	253	North Car	2616975
17	Bojan Bogdanovic	Brooklyn Nets	44	SG	27	8-Jun	216		3425510
18	Markel Brown	Brooklyn Nets	22	SG	24	3-Jun	190	Oklahoma	845059



```
# importing pandas package
import pandas as pd

# making data frame from csv file
data = pd.read_csv("nba.csv")

# sorting data frame by name
data.sort_values("Name", axis = 0, ascending = True,
                 inplace = True, na_position ='last')

# display
print(data)
```

Output:-



```
                Name                    Team  Number Position   Age Height  Weight            College     Salary
152     Aaron Brooks           Chicago Bulls     0.0       PG  31.0    6-0   161.0             Oregon  2250000.0
356     Aaron Gordon           Orlando Magic     0.0       PF  20.0    6-9   220.0            Arizona  4171680.0
328   Aaron Harrison       Charlotte Hornets     9.0       SG  21.0    6-6   210.0           Kentucky   525093.0
404    Adreian Payne  Minnesota Timberwolves    33.0       PF  25.0   6-10   237.0     Michigan State  1938840.0
312       Al Horford           Atlanta Hawks    15.0        C  30.0   6-10   245.0            Florida 12000000.0
..               ...                     ...     ...      ...   ...    ...     ...                ...        ...
270   Xavier Munford       Memphis Grizzlies    14.0       PG  24.0    6-3   180.0       Rhode Island        NaN
402      Zach LaVine  Minnesota Timberwolves     8.0       PG  21.0    6-5   189.0               UCLA  2148360.0
271    Zach Randolph       Memphis Grizzlies    50.0       PF  34.0    6-9   260.0     Michigan State  9638555.0
237    Zaza Pachulia        Dallas Mavericks    27.0        C  32.0   6-11   275.0                NaN  5200000.0
457              NaN                     NaN     NaN      NaN   NaN    NaN     NaN                NaN        NaN

[458 rows x 9 columns]
```