

DATE : 13.11.2024

DT/NT : NT

LESSON : PANDAS

SUBJECT: INTRODUCTION

BATCH : B304 / B305



techproeducation.com



+1 (585) 304 29 59



TECHPRO
EDUCATION





Pandas - Introduction

Data Science Program
Session-1

Pandas Lesson Content



**Pandas
derslerinde ne
öğreneceğiz?**

- Series & Data Frames
- Indexing, Slicing & Selection
- Groupby & Useful Operations
- Handling with Missing_Values
- Combining Data Frames
- Text & Time Data

Session-1 Content



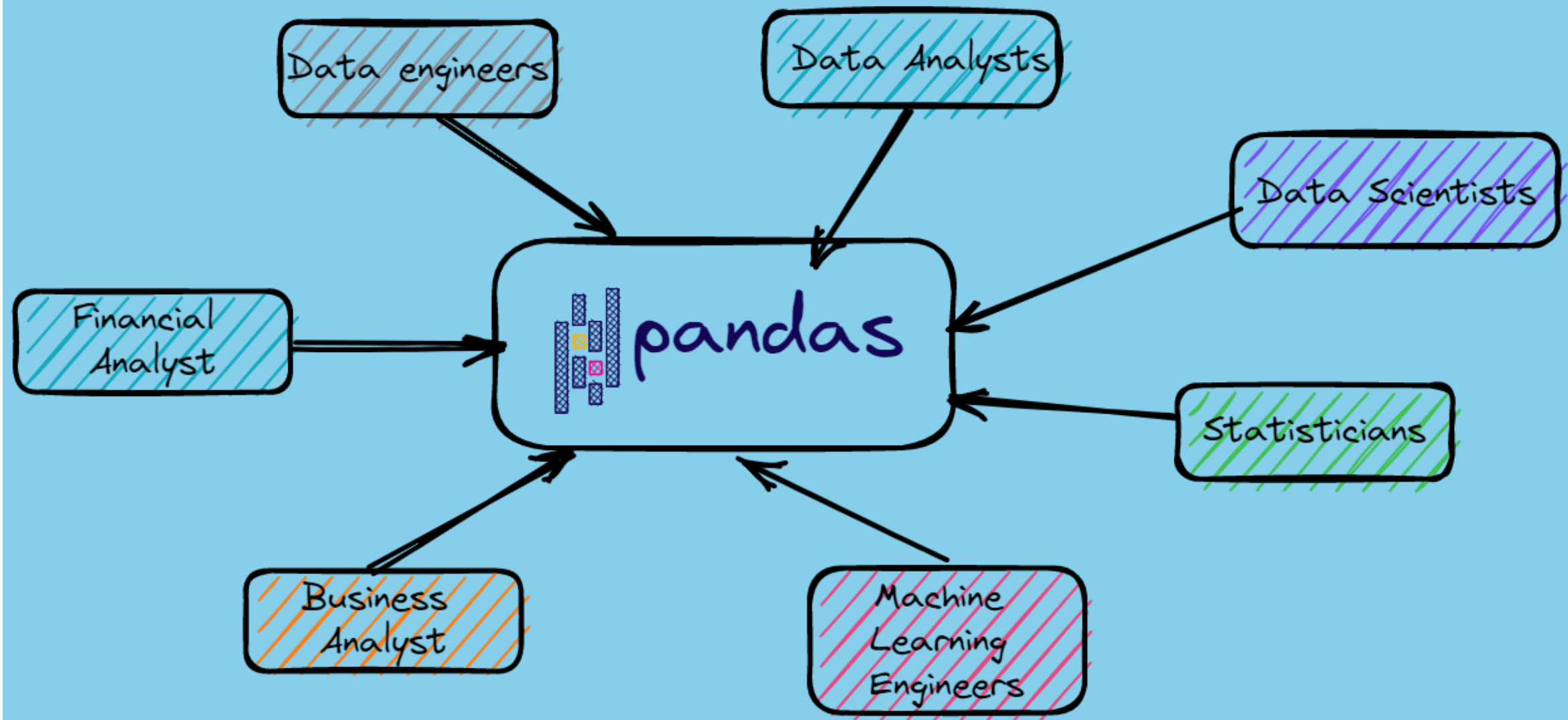
**Bugün ne
öğreneceğiz?**

- Pandas Introduction
- Pandas Data Types
- Pandas Series
- Pandas DataFrame
- Creating Series- Notebook

Why Pandas



Why Pandas



Pandas

- Pandas, Python programlama dilinde veri manipülasyonu ve analizi için kullanılan açık kaynaklı bir kütüphanedir.
- Pandas'ın temel veri yapıları olan Series (Tek boyutlu veriler için) ve DataFrame (İki boyutlu veriler için) son derece kullanışlı ve esnek araçlardır.



Pandas

- **Veri temizleme ve ön işleme:** Eksik verilerin doldurulması, hatalı verilerin çıkarılması, verilerin formatının değiştirilmesi, veri dönüştürme işlemleri vb.
- **Veri analizi:** Verilerin analiz edilmesi ve özetlenmesi, istatistiksel analizlerin yapılması, özellik mühendisliği vb.
- **Veri görselleştirme:** Verilerin çeşitli grafiklerle görselleştirilmesi, verinin anlaşılmasını ve yorumlanmasını kolaylaştırmak için.
- **Makine öğrenmesi:** Veri setini makine öğrenmesi algoritmalarına beslemek üzere veriyi ön işleme ve hazırlama konusunda çok kullanışlıdır.



Pandas

Pandas = Panel Data System

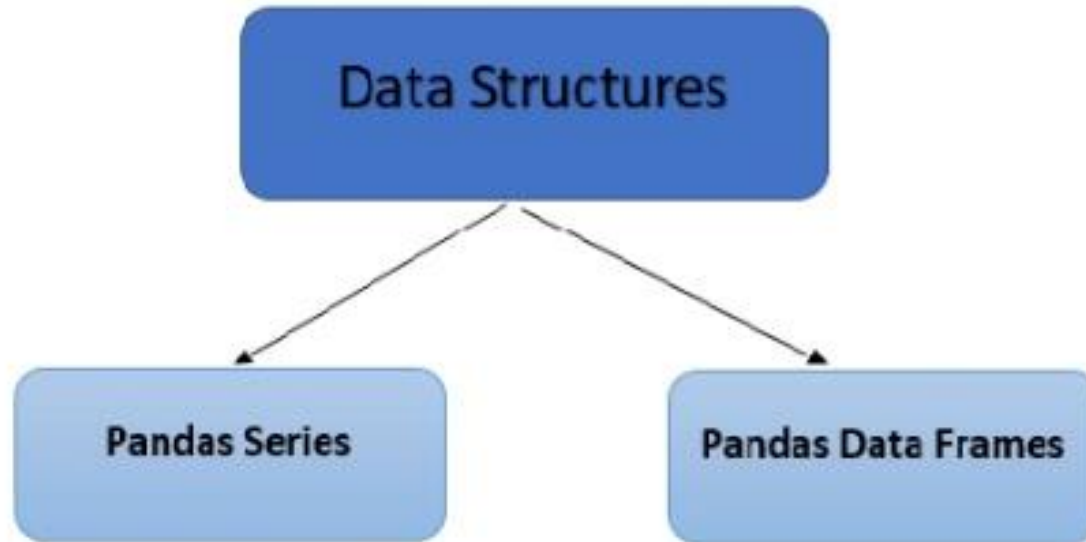
- Pandas kütüphanesi, Python programlama dili için geliştirilmiş, panel veriler dahil olmak üzere birçok veri türünü işlemek için kullanılan etkili bir araçtır.
- Panel data, ekonometri ve istatistiksel analizde kullanılan bir veri türüdür.
- Panel veri, birden fazla öznenen veya birimden (örneğin bireyler, firmalar, ülkeler vs.) zaman içerisinde toplanan verileri ifade eder.



Pandas

Pandas = Panel Data System

- Ancak, pratikte Pandas'ın en yaygın kullanımı, DataFrame adlı yapı ile tablosal veriler üzerinde işlem yapmaktır.
- Pandas ile veri analizi yaparken kullanacağımız temel veri yapıları; **Series** ve **DataFrames**



Pandas



Pandas



Data Structure	Dimensionality	Format	View																														
Series	1D	Column	<table><thead><tr><th></th><th>name</th></tr></thead><tbody><tr><td>0</td><td>Rukshan</td></tr><tr><td>1</td><td>Prasadi</td></tr><tr><td>2</td><td>Gihan</td></tr><tr><td>3</td><td>Hansana</td></tr></tbody></table> <table><thead><tr><th></th><th>age</th></tr></thead><tbody><tr><td>0</td><td>25</td></tr><tr><td>1</td><td>25</td></tr><tr><td>2</td><td>26</td></tr><tr><td>3</td><td>24</td></tr></tbody></table> <table><thead><tr><th></th><th>marks</th></tr></thead><tbody><tr><td>0</td><td>85</td></tr><tr><td>1</td><td>90</td></tr><tr><td>2</td><td>70</td></tr><tr><td>3</td><td>80</td></tr></tbody></table>		name	0	Rukshan	1	Prasadi	2	Gihan	3	Hansana		age	0	25	1	25	2	26	3	24		marks	0	85	1	90	2	70	3	80
	name																																
0	Rukshan																																
1	Prasadi																																
2	Gihan																																
3	Hansana																																
	age																																
0	25																																
1	25																																
2	26																																
3	24																																
	marks																																
0	85																																
1	90																																
2	70																																
3	80																																
DataFrame	2D	Single Sheet	<table><thead><tr><th></th><th>name</th><th>age</th><th>marks</th></tr></thead><tbody><tr><td>0</td><td>Rukshan</td><td>25</td><td>85</td></tr><tr><td>1</td><td>Prasadi</td><td>25</td><td>90</td></tr><tr><td>2</td><td>Gihan</td><td>26</td><td>70</td></tr><tr><td>3</td><td>Hansana</td><td>24</td><td>80</td></tr></tbody></table>		name	age	marks	0	Rukshan	25	85	1	Prasadi	25	90	2	Gihan	26	70	3	Hansana	24	80										
	name	age	marks																														
0	Rukshan	25	85																														
1	Prasadi	25	90																														
2	Gihan	26	70																														
3	Hansana	24	80																														
Panel	3D	Multiple Sheets	<table><thead><tr><th></th><th>name</th><th>age</th><th>marks</th></tr></thead><tbody><tr><td>0</td><td>Rukshan</td><td>25</td><td>85</td></tr><tr><td>1</td><td>Prasadi</td><td>25</td><td>90</td></tr><tr><td>2</td><td>Gihan</td><td>26</td><td>70</td></tr><tr><td>3</td><td>Hansana</td><td>24</td><td>80</td></tr></tbody></table>		name	age	marks	0	Rukshan	25	85	1	Prasadi	25	90	2	Gihan	26	70	3	Hansana	24	80										
	name	age	marks																														
0	Rukshan	25	85																														
1	Prasadi	25	90																														
2	Gihan	26	70																														
3	Hansana	24	80																														

Pandas Series

- Pandas Serisi, NumPy dizi nesnelerinin üzerine inşa edilmiştir ve çok benzerler.
- Herhangi bir veri tipinde veri tutabilen tek boyutlu etiketli bir dizidir.
- Etiket değerlerine ise indeks denir.
- Verinin kendisi sayılar, dizeler veya başka Python objelerinden oluşabilir.
- Serileri oluşturmak için ise listeler, sıralı diziler ya da sözlükler kullanılabilir.



Pandas Series



Series Index

	A
1	1
2	2
3	3
4	4

**Series
Name**

**Series
Values**

Pandas Series



```
import pandas as pd  
series1 = pd.Series([10, 20, 30])  
print(series1)
```

giving an alias name to pandas

Series
object

List

index

Output

0	10
1	20
2	30

dtype: int64

Pandas Series



	First Name
0	Lois
1	Brenda
2	Joe
3	Diane
4	Benjamin
5	Patrick
6	Nancy
7	Carol
8	Frances
9	Diana

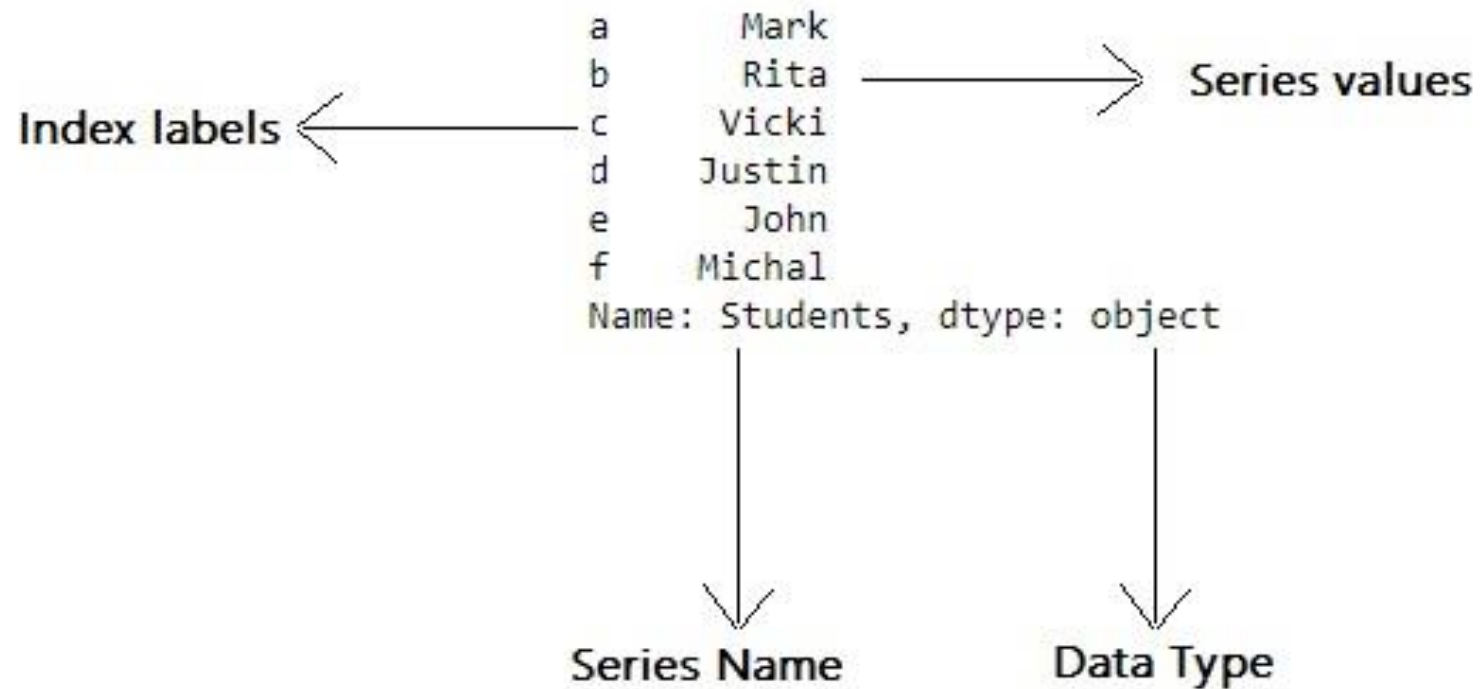
Diagram illustrating a Pandas Series structure. The left side is labeled "Index" and the right side is labeled "Data". The data is represented as a list of names corresponding to the index values 0 through 9.

PANDAS
SERIES

Pandas Series



Pandas Series



Pandas Series



Series 1

INDEX	DATA
0	A
1	B
2	C
3	D
4	E
5	F

Series 2

INDEX	DATA
A	1
B	2
C	3
D	4
E	5
F	6

Series 3

INDEX	DATA
0	[1, 2]
1	A
2	1
3	(4, 5)
4	{"a": 1}
5	6

Series 4

INDEX	DATA
Jan-18	11
Feb-18	23
Mar-18	43
Apr-18	21
May-18	17
Jun-18	6

Pandas Series



Creating Series

```
import pandas as pd  
s1 = pd.Series([1, 2, 3, 4])
```

```
s2 = pd.Series([1, 2, 3, 4], index=['A', 'B', 'C', 'D'])
```

0	1
1	2
2	3
3	4

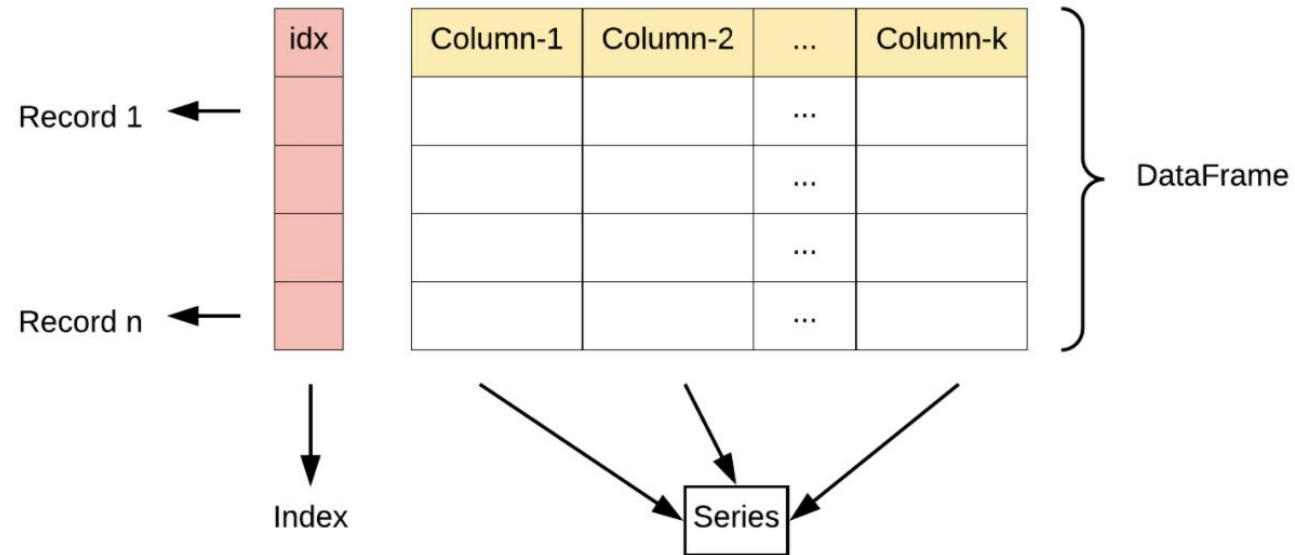
A	1
B	2
C	3
D	4

Pandas DataFrames

- Pandas Dataframe, satırları ve sütunları olan iki boyutlu etiketli veri yapısıdır.
- Pandas DataFrame'deki her sütun bir Pandas Serisidir.
- Verinin kendisi sayılar, dizeler veya başka Python objelerinden oluşabilir.
- Serileri oluşturmak için ise listeler, sıralı diziler ya da sözlükler kullanılabilir.



Pandas DataFrames



Series 1			Series 2			Series 3			DataFrame			
Mango			Apple			Banana			Mango	Apple	Banana	
0	4		0	5		0	2		0	4	5	2
1	5		1	4		1	3		1	5	4	3
2	6		2	3		2	5		2	6	3	5
3	3		3	0		3	2		3	3	0	2
4	1		4	2		4	7		4	1	2	7

https://blog.csdn.net/weixin_39887317

Pandas DataFrames



Series

	apples
0	3
1	2
2	0
3	1

Series

	oranges
0	0
1	3
2	7
3	2

DataFrame

	apples	oranges
0	3	0
1	2	3
2	0	7
3	1	2

Pandas DataFrames



Column Label/ Header

Index Label

	0	1	2	3	4	
Label	Name	Age	Marks	Grade	Hobby	
0	S1	Joe	20	85.10	A	Swimming
1	S2	Nat	21	77.80	B	Reading
2	S3	Harry	19	91.54	A	Music
3	S4	Sam	20	88.78	A	Painting
4	S5	Monica	22	60.55	B	Dancing

Column Index

Row

Row Index

Column

Element/ Value/ Entry

DataFrame



- DataFrame is a two-dimensional array with heterogeneous data. For example,

Name	Age	Gender	Rating
Steve	32	Male	3.45
Lia	28	Female	4.6
Vin	45	Male	3.9
Katie	38	Female	2.78

Data Type of Columns

Column	Type
Name	String
Age	Integer
Gender	String
Rating	Float

Pandas DataFrames



Pandas Data Structures

Series

<i>index</i>	<i>values</i>
A	6
B	3.14
C	-4
D	0

DataFrame

<i>index</i>	<i>columns</i>		
	foo	bar	baz
A	x	6	True
B	y	10	True
C	z	NaN	False

Pandas DataFrames



Diagram illustrating a Pandas DataFrame structure with rows and columns.

		Columns			
		Name	Score	Attempts	Qualify
Rows	0	Anastasia	12.5	1	yes
	1	Dima	9.0	3	no
	2	Katherine	16.5	2	yes
	3	James	NaN	3	no
	4	Emily	9.0	2	no

The diagram shows a table with 5 rows and 5 columns. The columns are labeled Name, Score, Attempts, and Qualify. The rows are indexed 0 to 4. The data values are: Anastasia (12.5, 1, yes), Dima (9.0, 3, no), Katherine (16.5, 2, yes), James (NaN, 3, no), and Emily (9.0, 2, no). The word 'Data' is written below the table.

Pandas DataFrame

Pandas DataFrames



```
df = pd.DataFrame ( { 'month' : [ 2, 5, 8, 10 ],  
    'year' : [ 2017, 2019, 2018, 2019 ],  
    'sale' : [ 60, 45, 90, 36 ] } )
```



DataFrame

df	month	year	sale
0	2	2017	60
1	5	2019	45
2	8	2018	90
3	10	2019	36

```
pd.DataFrame ( np.array ( ( [ 2, 3, 4 ], [ 5, 6, 7 ] ) ),  
    index = [ 'tiger', 'lion' ],  
    columns = [ 'one', 'two', 'three' ] )
```



DataFrame	one	two	three
tiger	2	3	4
lion	5	6	7

© w3resource.com

Pandas DataFrames



df.values



DataFrame

	name	max_speed	rank
0	sparrow	30.0	second
1	tiger	90.5	1
2	fox	NaN	None

dataframe contains
mixed values



Array

sparrow	30.0	second
tiger	90.5	1
fox	NaN	None

array also contains
mixed values

w3resource.com

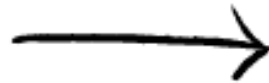
Pandas DataFrames



DATA SCIENCE
PARICHAY



```
{  
    "Name": ["Jim",  
            "Dwight", "Angela",  
            "Tobi"],  
    "Age": [26, 28, 27,  
            32],  
    "Department": ["Sales",  
                  "Sales", "Accounting",  
                  "Human Resources"]  
}
```



	Name	Age	Department
0	Jim	26	Sales
1	Dwight	28	Sales
2	Angela	27	Accounting
3	Tobi	32	Human Resources

DataFrame from Dictionary

Pandas DataFrames



IMPORTANT METHODS IN PANDAS PACKAGE

@MUKESH NAGAR

DATA IMPORTING

- `pd.read_csv ()`
- `pd.read_table ()`
- `pd.read_excel ()`
- `pd.read_sql ()`
- `pd.read_json ()`
- `pd.read_html ()`
- `pd.read_clipboard ()`
- `pd.DataFrame ()`
- `pd.concat ()`
- `pd.Series ()`
- `pd.date_range ()`

DATA CLEANING

- `df.dropna ()`
- `df.fillna ()`
- `df.describe ()`
- `df.sort_values ()`
- `df.groupby ()`
- `df.apply ()`
- `df.append ()`
- `df.join ()`
- `df.rename ()`
- `df.set_index ()`
- `df.to_csv ()`

DATA STATISTICS

- `df.head ()`
- `df.tail ()`
- `df.info ()`
- `df.describe ()`
- `df.mean ()`
- `df.median ()`
- `df.std ()`
- `df.corr ()`
- `df.count ()`
- `df.max ()`
- `df.min ()`