

21 st Aug

PANDAS

pandas: powerful Python data analysis toolkit

1. What is it?

- pandas is a Python package that provides fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis / manipulation tool available in any language. It is already well on its way towards this goal.

2. What is Pandas?

- Pandas is a Python library used for working with data sets.
- It has functions for analyzing, cleaning, exploring, and manipulating data.
- The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

3. Why Use Pandas?

- Pandas allows us to analyze big data and make conclusions based on statistical theories.
- Pandas can clean messy data sets, and make them readable and relevant.
- Relevant data is very important in data science.

4. What Can Pandas Do?

Pandas gives you answers about the data. Like:

- Is there a correlation between two or more columns?
- What is average value?
- Max value?
- Min value?

Pandas are also able to delete rows that are not relevant, or contains wrong values, like empty or NULL values. This is called cleaning the data.

```
In [109... import pandas as pd
In [111... pd.__version__
Out[111... '2.2.2'
In [112... store = pd.read_csv(r"C:\data science and ai\Sample - Superstore_Orders.csv")
In [113... store
```

Out[113...]

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	US-2020-103800
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	US-2020-112326
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	US-2020-112326
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	04-01-2020	US-2020-112326
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	US-2020-141817
...							
10189	Office Supplies	New York City	United States	Patrick O'Donnell	Wilson Jones	30-12-2023	US-2023-143259
10190	Office Supplies	Fairfield	United States	Erica Bern	GBC	30-12-2023	US-2023-115427
10191	Office Supplies	Loveland	United States	Jill Matthias	Other	30-12-2023	US-2023-156720
10192	Technology	New York City	United States	Patrick O'Donnell	Other	30-12-2023	US-2023-143259
10193	Office Supplies	Charlottetown	Canada	Harry Olson	Wilson Jones	30-12-2023	CA-2023-143500

10194 rows × 19 columns

In [114...]: len(store)

Out[114...]: 10194

In [115...]: store.columns

Out[115...]: Index(['Category', 'City', 'Country/Region', 'Customer Name', 'Manufacturer', 'Order Date', 'Order ID', 'Postal Code', 'Product Name', 'Region', 'Segment', 'Ship Date', 'Ship Mode', 'State/Province', 'Sub-Category', 'Discount', 'Profit', 'Quantity', 'Sales'], dtype='object')

In [116...]: store.shape

Out[116...]: (10194, 19)

In [117...]: store.isnull()

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Postal Code	F
0	False	False		False	False	False	False	False	False
1	False	False		False	False	False	False	False	False
2	False	False		False	False	False	False	False	False
3	False	False		False	False	False	False	False	False
4	False	False		False	False	False	False	False	False
...
10189	False	False		False	False	False	False	False	False
10190	False	False		False	False	False	False	False	False
10191	False	False		False	False	False	False	False	False
10192	False	False		False	False	False	False	False	False
10193	False	False		False	False	False	False	False	False

10194 rows × 19 columns



In [118...]: store.isnull().sum()

```
Out[118... Category      0
City          0
Country/Region 0
Customer Name 0
Manufacturer   0
Order Date    0
Order ID      0
Postal Code   0
Product Name  0
Region         0
Segment        0
Ship Date     0
Ship Mode     0
State/Province 0
Sub-Category   0
Discount       0
Profit         0
Quantity       0
Sales          0
dtype: int64
```

In [119... `store.isna()`

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Postal Code	F
0	False	False		False	False	False	False	False	False
1	False	False		False	False	False	False	False	False
2	False	False		False	False	False	False	False	False
3	False	False		False	False	False	False	False	False
4	False	False		False	False	False	False	False	False
...
10189	False	False		False	False	False	False	False	False
10190	False	False		False	False	False	False	False	False
10191	False	False		False	False	False	False	False	False
10192	False	False		False	False	False	False	False	False
10193	False	False		False	False	False	False	False	False

10194 rows × 19 columns



In [120... `store.head()`

Out[120...]

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Postal Code
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	US-2020-103800	77095
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	US-2020-112326	60540
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	US-2020-112326	60540
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	04-01-2020	US-2020-112326	60540
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	US-2020-141817	19143



In [121...]

store.tail()

Out[121...]

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID
10189	Office Supplies	New York City	United States	Patrick O'Donnell	Wilson Jones	30-12-2023	US-2023-143259
10190	Office Supplies	Fairfield	United States	Erica Bern	GBC	30-12-2023	US-2023-115427
10191	Office Supplies	Loveland	United States	Jill Matthias	Other	30-12-2023	US-2023-156720
10192	Technology	New York City	United States	Patrick O'Donnell	Other	30-12-2023	US-2023-143259
10193	Office Supplies	Charlottetown	Canada	Harry Olson	Wilson Jones	30-12-2023	CA-2023-143500



In [122...]

store.head(2)

Out[122...]

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Postal Code
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	US-2020-103800	77095
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	US-2020-112326	60540



In [123...]

store[:]

Out[123...]

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	US-2020-103800
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	US-2020-112326
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	US-2020-112326
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	04-01-2020	US-2020-112326
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	US-2020-141817
...							
10189	Office Supplies	New York City	United States	Patrick O'Donnell	Wilson Jones	30-12-2023	US-2023-143259
10190	Office Supplies	Fairfield	United States	Erica Bern	GBC	30-12-2023	US-2023-115427
10191	Office Supplies	Loveland	United States	Jill Matthias	Other	30-12-2023	US-2023-156720
10192	Technology	New York City	United States	Patrick O'Donnell	Other	30-12-2023	US-2023-143259
10193	Office Supplies	Charlottetown	Canada	Harry Olson	Wilson Jones	30-12-2023	CA-2023-143500

10194 rows × 19 columns

In [124...]

```
store[::-1]
```

Out[124...]

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID
10193	Office Supplies	Charlottetown	Canada	Harry Olson	Wilson Jones	30-12-2023	CA-2023-143500
10192	Technology	New York City	United States	Patrick O'Donnell	Other	30-12-2023	US-2023-143259
10191	Office Supplies	Loveland	United States	Jill Matthias	Other	30-12-2023	US-2023-156720
10190	Office Supplies	Fairfield	United States	Erica Bern	GBC	30-12-2023	US-2023-115427
10189	Office Supplies	New York City	United States	Patrick O'Donnell	Wilson Jones	30-12-2023	US-2023-143259
...
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	US-2020-141817
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	04-01-2020	US-2020-112326
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	US-2020-112326
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	US-2020-112326
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	US-2020-103800

10194 rows × 19 columns

In [125...]

store[1:10:3]

Out[125...]

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Postal Code
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	US-2020-112326	60540
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	US-2020-141817	19143
7	Office Supplies	Athens	United States	Jack O'Briant	Dixon	06-01-2020	US-2020-106054	30605



In [126...]

store.describe()

Out[126...]

	Discount	Profit	Quantity	Sales
count	10194.000000	10194.000000	10194.000000	10194.000000
mean	0.155385	28.673417	3.791838	228.225854
std	0.206249	232.465115	2.228317	619.906839
min	0.000000	-6599.978000	1.000000	0.444000
25%	0.000000	1.760800	2.000000	17.220000
50%	0.200000	8.690000	3.000000	53.910000
75%	0.200000	29.297925	5.000000	209.500000
max	0.800000	8399.976000	14.000000	22638.480000

In [127...]

store.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10194 entries, 0 to 10193
Data columns (total 19 columns):
 #   Column            Non-Null Count  Dtype  
--- 
 0   Category          10194 non-null   object  
 1   City               10194 non-null   object  
 2   Country/Region    10194 non-null   object  
 3   Customer Name     10194 non-null   object  
 4   Manufacturer      10194 non-null   object  
 5   Order Date        10194 non-null   object  
 6   Order ID          10194 non-null   object  
 7   Postal Code       10194 non-null   object  
 8   Product Name      10194 non-null   object  
 9   Region             10194 non-null   object  
 10  Segment            10194 non-null   object  
 11  Ship Date         10194 non-null   object  
 12  Ship Mode         10194 non-null   object  
 13  State/Province    10194 non-null   object  
 14  Sub-Category      10194 non-null   object  
 15  Discount           10194 non-null   float64 
 16  Profit              10194 non-null   float64 
 17  Quantity            10194 non-null   int64   
 18  Sales               10194 non-null   float64 
dtypes: float64(3), int64(1), object(15)
memory usage: 1.5+ MB
```

```
In [128...]: store['Category']
```

```
Out[128...]: 0      Office Supplies
 1      Office Supplies
 2      Office Supplies
 3      Office Supplies
 4      Office Supplies
 ...
 10189  Office Supplies
 10190  Office Supplies
 10191  Office Supplies
 10192      Technology
 10193  Office Supplies
Name: Category, Length: 10194, dtype: object
```

```
In [129...]: store[['Category', 'City']]
```

Out[129...]

	Category	City
0	Office Supplies	Houston
1	Office Supplies	Naperville
2	Office Supplies	Naperville
3	Office Supplies	Naperville
4	Office Supplies	Philadelphia
...
10189	Office Supplies	New York City
10190	Office Supplies	Fairfield
10191	Office Supplies	Loveland
10192	Technology	New York City
10193	Office Supplies	Charlottetown

10194 rows × 2 columns

In [130...]

store[['Category', 'City', 'Manufacturer']]

Out[130...]

	Category	City	Manufacturer
0	Office Supplies	Houston	Message Book
1	Office Supplies	Naperville	GBC
2	Office Supplies	Naperville	Avery
3	Office Supplies	Naperville	SAFCO
4	Office Supplies	Philadelphia	Avery
...
10189	Office Supplies	New York City	Wilson Jones
10190	Office Supplies	Fairfield	GBC
10191	Office Supplies	Loveland	Other
10192	Technology	New York City	Other
10193	Office Supplies	Charlottetown	Wilson Jones

10194 rows × 3 columns

In [131...]

store.dtypes

```
Out[131... Category      object
          City        object
          Country/Region    object
          Customer Name    object
          Manufacturer    object
          Order Date      object
          Order ID        object
          Postal Code     object
          Product Name     object
          Region         object
          Segment        object
          Ship Date      object
          Ship Mode       object
          State/Province   object
          Sub-Category    object
          Discount        float64
          Profit          float64
          Quantity        int64
          Sales           float64
          dtype: object
```

```
In [132... len(store.dtypes)
```

```
Out[132... 19
```

```
In [133... store.columns
```

```
Out[133... Index(['Category', 'City', 'Country/Region', 'Customer Name', 'Manufacturer',
                  'Order Date', 'Order ID', 'Postal Code', 'Product Name', 'Region',
                  'Segment', 'Ship Date', 'Ship Mode', 'State/Province', 'Sub-Category',
                  'Discount', 'Profit', 'Quantity', 'Sales'],
                  dtype='object')
```

```
In [135... store_categorical =store[['Category', 'City', 'Country/Region', 'Customer Name', 'Manufacturer',
                                         'Order Date', 'Order ID', 'Postal Code', 'Product Name', 'Region',
                                         'Segment', 'Ship Date', 'Ship Mode', 'State/Province', 'Sub-Category',
                                         'Discount', 'Profit', 'Quantity', 'Sales']]
```

```
In [137... store_categorical.head()
```

Out[137...]

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Postal Code
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	US-2020-103800	77095
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	US-2020-112326	60540
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	US-2020-112326	60540
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	04-01-2020	US-2020-112326	60540
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	US-2020-141817	19143



In [138...]

store_categorical.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10194 entries, 0 to 10193
Data columns (total 19 columns):
 #   Column            Non-Null Count  Dtype  
--- 
 0   Category          10194 non-null   object  
 1   City               10194 non-null   object  
 2   Country/Region    10194 non-null   object  
 3   Customer Name     10194 non-null   object  
 4   Manufacturer      10194 non-null   object  
 5   Order Date        10194 non-null   object  
 6   Order ID          10194 non-null   object  
 7   Postal Code       10194 non-null   object  
 8   Product Name      10194 non-null   object  
 9   Region             10194 non-null   object  
 10  Segment            10194 non-null   object  
 11  Ship Date         10194 non-null   object  
 12  Ship Mode         10194 non-null   object  
 13  State/Province    10194 non-null   object  
 14  Sub-Category      10194 non-null   object  
 15  Discount           10194 non-null   float64 
 16  Profit              10194 non-null   float64 
 17  Quantity            10194 non-null   int64  
 18  Sales               10194 non-null   float64 
dtypes: float64(3), int64(1), object(15)
memory usage: 1.5+ MB
```

In [139...]: `store_numerical = store[['Discount', 'Profit', 'Quantity', 'Sales']]`

In [140...]: `store_numerical`

Out[140...]:

	Discount	Profit	Quantity	Sales
0	0.2	5.5512	2	16.448
1	0.8	-5.4870	2	3.540
2	0.2	4.2717	3	11.784
3	0.2	-64.7748	3	272.736
4	0.2	4.8840	3	19.536
...
10189	0.2	19.7910	3	52.776
10190	0.2	6.4750	2	20.720
10191	0.2	-0.6048	3	3.024
10192	0.0	2.7279	7	90.930
10193	0.2	-0.6048	3	3.024

10194 rows × 4 columns

```
In [141... store_numerical.dtypes
```

```
Out[141... Discount      float64
Profit        float64
Quantity      int64
Sales         float64
dtype: object
```

```
In [183... print(len(store.columns))
print(len(store_categorical.columns))
print(len(store_numerical.columns))
```

```
19
19
4
```

```
In [185... store.head()
```

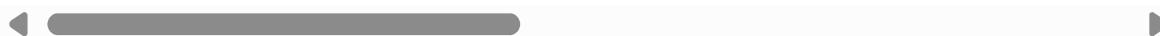
	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Postal Code
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	US-2020-103800	77095
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	US-2020-112326	60540
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	US-2020-112326	60540
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	04-01-2020	US-2020-112326	60540
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	US-2020-141817	19143



```
In [187... store.tail()
```

Out[187...]

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID
10189	Office Supplies	New York City	United States	Patrick O'Donnell	Wilson Jones	30-12-2023	US-2023-143259
10190	Office Supplies	Fairfield	United States	Erica Bern	GBC	30-12-2023	US-2023-115427
10191	Office Supplies	Loveland	United States	Jill Matthias	Other	30-12-2023	US-2023-156720
10192	Technology	New York City	United States	Patrick O'Donnell	Other	30-12-2023	US-2023-143259
10193	Office Supplies	Charlottetown	Canada	Harry Olson	Wilson Jones	30-12-2023	CA-2023-143500



In [189...]

store[:]

Out[189...]

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	US-2020-103800
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	US-2020-112326
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	US-2020-112326
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	04-01-2020	US-2020-112326
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	US-2020-141817
...							
10189	Office Supplies	New York City	United States	Patrick O'Donnell	Wilson Jones	30-12-2023	US-2023-143259
10190	Office Supplies	Fairfield	United States	Erica Bern	GBC	30-12-2023	US-2023-115427
10191	Office Supplies	Loveland	United States	Jill Matthias	Other	30-12-2023	US-2023-156720
10192	Technology	New York City	United States	Patrick O'Donnell	Other	30-12-2023	US-2023-143259
10193	Office Supplies	Charlottetown	Canada	Harry Olson	Wilson Jones	30-12-2023	CA-2023-143500

10194 rows × 19 columns

In [191...]

```
store[::-1]
```

Out[191...]

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID
10193	Office Supplies	Charlottetown	Canada	Harry Olson	Wilson Jones	30-12-2023	CA-2023-143500
10192	Technology	New York City	United States	Patrick O'Donnell	Other	30-12-2023	US-2023-143259
10191	Office Supplies	Loveland	United States	Jill Matthias	Other	30-12-2023	US-2023-156720
10190	Office Supplies	Fairfield	United States	Erica Bern	GBC	30-12-2023	US-2023-115427
10189	Office Supplies	New York City	United States	Patrick O'Donnell	Wilson Jones	30-12-2023	US-2023-143259
...
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	US-2020-141817
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	04-01-2020	US-2020-112326
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	US-2020-112326
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	US-2020-112326
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	US-2020-103800

10194 rows × 19 columns

In [193...]

```
store[0:100:10]
```

Out[193...]

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Pa C
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	US-2020-103800	77
10	Office Supplies	Henderson	United States	Maria Etezadi	Southworth	06-01-2020	US-2020-167199	42
20	Furniture	Dover	United States	Seth Vernon	DAX	11-01-2020	US-2020-130092	19
30	Office Supplies	San Francisco	United States	Brian Dahlen	Tennsco	13-01-2020	US-2020-157147	94
40	Office Supplies	Scottsdale	United States	Toby Swindell	GBC	19-01-2020	US-2020-146591	85
50	Office Supplies	Jonesboro	United States	Hunter Lopez	Newell	20-01-2020	US-2020-147627	72
60	Furniture	Miami	United States	Tom Boeckenhauer	Ultra	21-01-2020	US-2020-110422	33
70	Furniture	Alexandria	United States	Shirley Daniels	GE	26-01-2020	US-2020-155502	22
80	Office Supplies	New York City	United States	Jasper Cacioppo	Pressboard	28-01-2020	US-2020-100328	10
90	Office Supplies	Escondido	United States	Mark Haberlin	GBC	04-02-2020	US-2020-104808	92



In [195...]: store[10:11]

Out[195...]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Postal Code
10	Office Supplies	Henderson	United States	Maria Etezadi	Southworth	06-01-2020	2020-167199	US-42420



In [199...]: store_categorical [:4]

Out[199...]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Postal Code
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	2020-103800	US-77095
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	2020-112326	US-60540
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	2020-112326	US-60540



In []:

In []:

In []:

In []:

In []: