

# Pandas - statistical visualization

Pandas is especially useful for handling large datasets and performing operations like filtering, grouping, merging, reshaping, and more. features: Pandas can read data from various file formats like CSV, Excel, JSON, SQL databases, and more. It can also write data back to these formats. Data Cleaning -- like ex handling missing values Data Manipulation -- ex filter, merging , join rows Statistical Analysis -- (mean,mode,median variance etc.,)

```
In [7]: import pandas as pd
```

```
In [9]: pd.__version__
```

```
Out[9]: '2.2.2'
```

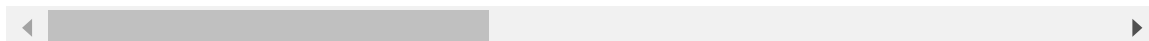
```
In [11]: df=pd.read_csv(r"C:\Users\user\Documents\Sample - Superstore_Orders.csv")
```

```
In [13]: df
```

Out[13]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	20103
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	20112
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	20112
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	04-01-2020	20112
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	20141
...	...	...	...	...	...	...	...
10189	Office Supplies	New York City	United States	Patrick O'Donnell	Wilson Jones	30-12-2023	20143
10190	Office Supplies	Fairfield	United States	Erica Bern	GBC	30-12-2023	20115
10191	Office Supplies	Loveland	United States	Jill Matthias	Other	30-12-2023	20156
10192	Technology	New York City	United States	Patrick O'Donnell	Other	30-12-2023	20143
10193	Office Supplies	Charlottetown	Canada	Harry Olson	Wilson Jones	30-12-2023	20143

10194 rows × 19 columns



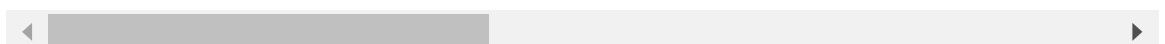
In [20]: store=pd.read\_csv(r"C:\Users\user\Documents\Sample - Superstore\_Orders.csv")

In [28]: store

Out[28]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	20103
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	20112
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	20112
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	04-01-2020	20112
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	20141
...	...	...	...	...	...	...	...
10189	Office Supplies	New York City	United States	Patrick O'Donnell	Wilson Jones	30-12-2023	20143
10190	Office Supplies	Fairfield	United States	Erica Bern	GBC	30-12-2023	20115
10191	Office Supplies	Loveland	United States	Jill Matthias	Other	30-12-2023	20156
10192	Technology	New York City	United States	Patrick O'Donnell	Other	30-12-2023	20143
10193	Office Supplies	Charlottetown	Canada	Harry Olson	Wilson Jones	30-12-2023	20143

10194 rows × 19 columns



In [30]: `len(store) # to know noof rowss`

Out[30]: 10194

In [48]: `id(store)`

Out[48]: 2538442744048

In [32]: `store.columns`

Out[32]: 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 [34]: `store.shape`

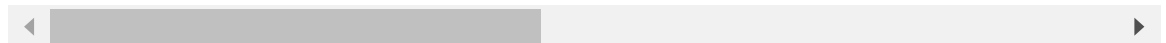
Out[34]: (10194, 19)

In [19]: `store.isnull()`

Out[19]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Postal Code
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 [21]: `store.isnull().sum()`

```
Out[21]: 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 [40]: store.dtypes
```

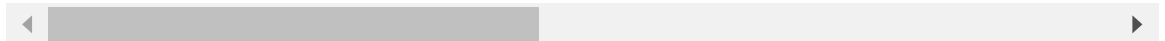
```
Out[40]: Category      object
          City          object
          Country/Region object
          Customer Name object
          Manufacturer   object
          Order Date     object
          Order ID       object
          Postal Code     object
          Product Name    object
          Region         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 [23]: store.isna()
```

Out[23]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Postal Code
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 [42]: `store.isna().sum()`

```
Out[42]: 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 [46]: `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 [52]: store['City']
```

```

Out[52]: 0          Houston
         1      Naperville
         2      Naperville
         3      Naperville
         4      Philadelphia
         ...
        10189  New York City
        10190      Fairfield
        10191      Loveland
        10192  New York City
        10193  Charlottetown
        Name: City, Length: 10194, dtype: object

```

```
In [56]: store[['City', 'Category']]
```

Out[56]:

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

10194 rows × 2 columns

In [58]: store[['City', 'Category', 'Order ID']]

Out[58]:

	City	Category	Order ID
0	Houston	Office Supplies	US-2020-103800
1	Naperville	Office Supplies	US-2020-112326
2	Naperville	Office Supplies	US-2020-112326
3	Naperville	Office Supplies	US-2020-112326
4	Philadelphia	Office Supplies	US-2020-141817
...	...	...	...
10189	New York City	Office Supplies	US-2023-143259
10190	Fairfield	Office Supplies	US-2023-115427
10191	Loveland	Office Supplies	US-2023-156720
10192	New York City	Technology	US-2023-143259
10193	Charlottetown	Office Supplies	CA-2023-143500

10194 rows × 3 columns

```
In [60]: store_text = store[['Category', 'City', 'Country/Region', 'Customer Name', 'Manager',
'Order Date', 'Order ID', 'Postal Code', 'Product Name', 'Region',
'Segment', 'Ship Date', 'Ship Mode', 'State/Province', 'Sub-Category' ]]
```

In [62]: store\_text



Out[62]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	20103
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	20112
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	20112
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	04-01-2020	20112
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	20141
...	...	...	...	...	...	...	...
10189	Office Supplies	New York City	United States	Patrick O'Donnell	Wilson Jones	30-12-2023	20143
10190	Office Supplies	Fairfield	United States	Erica Bern	GBC	30-12-2023	20115
10191	Office Supplies	Loveland	United States	Jill Matthias	Other	30-12-2023	20156
10192	Technology	New York City	United States	Patrick O'Donnell	Other	30-12-2023	20143
10193	Office Supplies	Charlottetown	Canada	Harry Olson	Wilson Jones	30-12-2023	20143

10194 rows × 15 columns



In [64]: store.columns

```
Out[64]: 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 [66]: store_text.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10194 entries, 0 to 10193
Data columns (total 15 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
dtypes: object(15)
memory usage: 1.2+ MB
```

```
In [68]: store_number = store[['Discount', 'Profit', 'Quantity', 'Sales']]
store_number
```

```
Out[68]:
```

	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 [70]: store.columns
```

```
Out[70]: 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 [72]: len(store.columns)
```

```
Out[72]: 19
```

```
In [74]: store_text.columns
```

```
Out[74]: 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'],  
             dtype='object')
```

```
In [76]: len(store_text.columns)
```

```
Out[76]: 15
```

```
In [78]: store_number.columns
```

```
Out[78]: Index(['Discount', 'Profit', 'Quantity', 'Sales'], dtype='object')
```

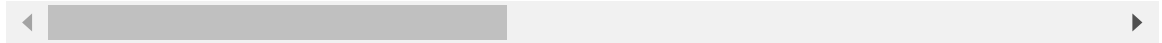
```
In [80]: len(store_number.columns)
```

```
Out[80]: 4
```

```
In [27]: store.head() # fetch by default 1st 5 rows
```

Out[27]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Pos
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	US-2020-103800	77
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	US-2020-112326	60
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	US-2020-112326	60
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	04-01-2020	US-2020-112326	60
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	US-2020-141817	19

In [29]: `store.tail() # by default botttom 5rows`

Out[29]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Post Code
10189	Office Supplies	New York City	United States	Patrick O'Donnell	Wilson Jones	30-12-2023	20143	
10190	Office Supplies	Fairfield	United States	Erica Bern	GBC	30-12-2023	20115	
10191	Office Supplies	Loveland	United States	Jill Matthias	Other	30-12-2023	20156	
10192	Technology	New York City	United States	Patrick O'Donnell	Other	30-12-2023	20143	
10193	Office Supplies	Charlottetown	Canada	Harry Olson	Wilson Jones	30-12-2023	20143	

In [33]: store.head(2) # 1st 2 rows

Out[33]:

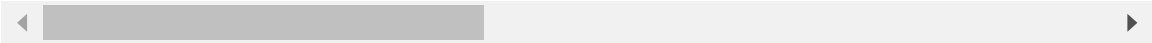
	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Post Code
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	US-2020-103800	77058
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	US-2020-112326	60540

In [35]: store[:]

Out[35]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	20103
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	20112
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	20112
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	04-01-2020	20112
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	20141
...	...	...	...	...	...	...	...
10189	Office Supplies	New York City	United States	Patrick O'Donnell	Wilson Jones	30-12-2023	20143
10190	Office Supplies	Fairfield	United States	Erica Bern	GBC	30-12-2023	20115
10191	Office Supplies	Loveland	United States	Jill Matthias	Other	30-12-2023	20156
10192	Technology	New York City	United States	Patrick O'Donnell	Other	30-12-2023	20143
10193	Office Supplies	Charlottetown	Canada	Harry Olson	Wilson Jones	30-12-2023	20143

10194 rows × 19 columns

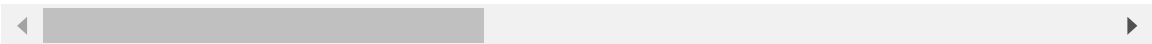


In [37]: store[:, :-1]

Out[37]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID
10193	Office Supplies	Charlottetown	Canada	Harry Olson	Wilson Jones	30-12-2023	20143
10192	Technology	New York City	United States	Patrick O'Donnell	Other	30-12-2023	20143
10191	Office Supplies	Loveland	United States	Jill Matthias	Other	30-12-2023	20156
10190	Office Supplies	Fairfield	United States	Erica Bern	GBC	30-12-2023	20115
10189	Office Supplies	New York City	United States	Patrick O'Donnell	Wilson Jones	30-12-2023	20143
...	...	...	...	...	...	...	...
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	20141
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	04-01-2020	20112
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	20112
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	20112
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	20103

10194 rows × 19 columns



In [39]: store[1:10:3]

Out[39]:

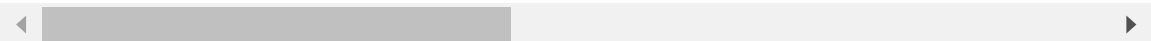
	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Po C
1	Office Supplies	Naperville	United States	Phyllina Ober	GBC	04-01-2020	US-2020-112326	60
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	US-2020-141817	19
7	Office Supplies	Athens	United States	Jack O'Briant	Dixon	06-01-2020	US-2020-106054	30



In [82]: store[10:11]

Out[82]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Po C
10	Office Supplies	Henderson	United States	Maria Etezadi	Southworth	06-01-2020	US-2020-167199	42



In [84]: store[:4]



Out[84]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Post Coc
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	US-2020-103800	7705
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	US-2020-112326	6054
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	US-2020-112326	6054
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	04-01-2020	US-2020-112326	6054

In [41]: `store.describe()` # here discount profit quantity sales are the only numerical da

Out[41]:

	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 [43]: `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 [57]: len(store.dtypes)
```

```
Out[57]: 19
```