

key notes:

1. Column name --> Attribute / feature
2. excel sheet --> Dataset
3. RAW Data --> Dataset with some missing/uncleaned data
4. Clean Data --> Dataset without any missing value
5. TEXT DATA --> CATEGORICAL DATA
6. NUMBERS --> NUMERICAL DATA
7. RAW string --> contains many special symbols & characters
8. RAW string --> r'C:\Users\TANISHQ\Sample - Superstore_Orders.csv'
9. string --> " Tanishq hello"
10. .csv --> comma seprated-values

```
In [1]: import pandas as pd
```

```
In [2]: pd.__version__
```

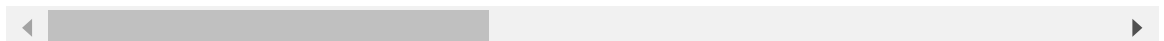
```
Out[2]: '2.2.2'
```

```
In [7]: df = pd.read_csv(r'C:\Users\lenovo\Desktop\New folder (3)\Sample - Superstore_Or  
df
```

Out[7]:

	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



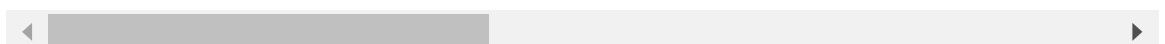
In [4]: store = pd.read_csv(r'C:\Users\lenovo\Desktop\New folder (3)\Sample - Superstor

In [5]: store

Out[5]:

	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 [8]: len(store)
```

```
Out[8]: 10194
```

```
In [9]: store.columns
```

```
Out[9]: 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 [11]: store.shape
```

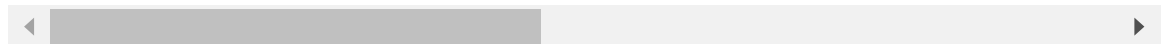
```
Out[11]: (10194, 19)
```

```
In [12]: store.isnull()
```

```
Out[12]:
```

	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 [13]: store.isnull().sum()
```

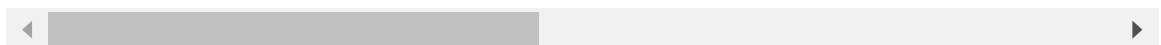
```
Out[13]: Category      0
City      0
Country/Region      0
Customer Name      0
Manufacturer      0
Order Date      0
Order ID      0
Postal Code      0
Product Name      0
Region      0
Segment      0
Ship Date      0
Ship Mode      0
State/Province      0
Sub-Category      0
Discount      0
Profit      0
Quantity      0
Sales      0
dtype: int64
```

```
In [15]: store.isna()
```

```
Out[15]:
```

	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 × 9 columns



```
In [16]: store.isna().sum()
```

```
Out[16]: 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 [17]: store.head()
```

```
Out[17]:
```

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Po C
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 [18]: store.tail()
```

Out[18]:

	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 [19]: store.head(2)

Out[19]:

	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	7706
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	US-2020-112326	6054

In [20]: store.tail(2)

Out[20]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID
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

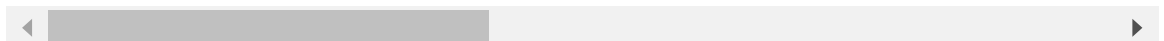
In [21]:

```
store[:]
```


Out[21]:

	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

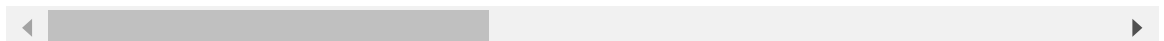


In [22]: store[:, :-1]

Out[22]:

	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

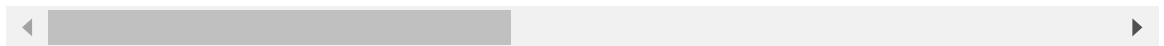
10194 rows × 19 columns



In [52]: store[1:10:3] # 1 to 10 with step of 3

Out[52]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID	Pos
1	Office Supplies	Naperville	United States	Phillina 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



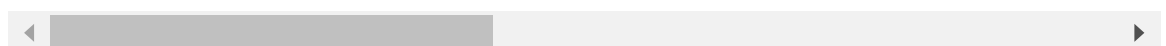
In [49]:

store[2:-1]

Out[49]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order ID
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	U11232
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	04-01-2020	U11232
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	U14187
5	Furniture	Henderson	United States	Maria Etezadi	Global	06-01-2020	U16719
6	Office Supplies	Henderson	United States	Maria Etezadi	Rogers	06-01-2020	U16719
...
10188	Office Supplies	Fairfield	United States	Erica Bern	Cardinal	30-12-2023	U11542
10189	Office Supplies	New York City	United States	Patrick O'Donnell	Wilson Jones	30-12-2023	U14321
10190	Office Supplies	Fairfield	United States	Erica Bern	GBC	30-12-2023	U11542
10191	Office Supplies	Loveland	United States	Jill Matthias	Other	30-12-2023	U15672
10192	Technology	New York City	United States	Patrick O'Donnell	Other	30-12-2023	U14321

10191 rows × 19 columns

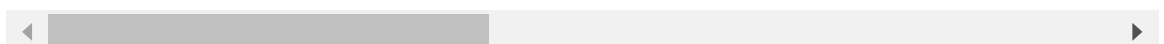


In [50]: `store[::-1] #Reverse rows`

Out[50]:

	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 [51]: store[::-3] #Arrange in reverse with 3 steps
```

Out[51]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Order Number
10193	Office Supplies	Charlottetown	Canada	Harry Olson	Wilson Jones	30-12-2023	2145
10190	Office Supplies	Fairfield	United States	Erica Bern	GBC	30-12-2023	2115
10187	Office Supplies	Columbus	United States	Chuck Clark	Eureka	30-12-2023	2126
10184	Furniture	Quebec City	Canada	Bruce Galang	Other	29-12-2023	2147
10181	Office Supplies	Grand Rapids	United States	Ken Brennan	Xerox	29-12-2023	2158
...
14	Furniture	Huntsville	United States	Vivek Sundaresam	Howard Miller	07-01-2020	2105
11	Office Supplies	Los Angeles	United States	Lycoris Saunders	Xerox	06-01-2020	2136
8	Office Supplies	Henderson	United States	Maria Etezadi	Ibico	06-01-2020	2167
5	Furniture	Henderson	United States	Maria Etezadi	Global	06-01-2020	2167
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	2112

3398 rows × 19 columns

In [27]: `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 [28]: `store['Category']`

```
Out[28]: 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 [29]: `store[['Category', 'City']]`

Out[29]:

	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 [31]: `store.dtypes`

```
Out[31]: 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 [32]: `len(store.dtypes)`

Out[32]: 19

In [33]: `store.columns`

```
Out[33]: 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_categorical = store[['Category', 'City', 'Country/Region', 'Customer Name',
    'Order Date', 'Order ID', 'Postal Code', 'Product Name', 'Region',
    'Segment', 'Ship Date', 'Ship Mode', 'State/Province', 'Sub-Category']]
```

```
In [38]: store_categorical.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 [43]: store_categorical.dtypes
```

```
Out[43]: 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
dtype: object
```

```
In [39]: store_numerical = store[['Discount', 'Profit', 'Quantity', 'Sales']]
```

```
In [40]: store_numerical.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 10194 entries, 0 to 10193  
Data columns (total 4 columns):  
#   Column      Non-Null Count  Dtype  
---  ---  
0   Discount    10194 non-null  float64  
1   Profit      10194 non-null  float64  
2   Quantity    10194 non-null  int64  
3   Sales       10194 non-null  float64  
dtypes: float64(3), int64(1)  
memory usage: 318.7 KB
```

```
In [45]: store_numerical.dtypes
```

```
Out[45]: Discount    float64  
Profit      float64  
Quantity     int64  
Sales       float64  
dtype: object
```

```
In [48]: print(len(store.columns))  
print(len(store_categorical.columns))  
print(len(store_numerical.columns))
```

```
19  
15  
4
```