In [4]:
```
import pandas as pd
```

Reading a CSV File

In [3]:
```
df1=pd.read_csv(r"D:\Sid 17-03-2025\SIDDHARTH BOSE\FSDS & GEN AI\March\24th - Ka
```

In [4]:
```
df1
```

Out[4]:

	Category	City	Country/Region	Customer Name	Manufacturer	Order Date	Or
0	Office Supplies	Houston	United States	Darren Powers	Message Book	03-01-2020	20 103
1	Office Supplies	Naperville	United States	Phillina Ober	GBC	04-01-2020	20 112
2	Office Supplies	Naperville	United States	Phillina Ober	Avery	04-01-2020	20 112
3	Office Supplies	Naperville	United States	Phillina Ober	SAFCO	04-01-2020	20 112
4	Office Supplies	Philadelphia	United States	Mick Brown	Avery	05-01-2020	20 141
...	...	...	...	...	...	...	
10189	Office Supplies	New York City	United States	Patrick O'Donnell	Wilson Jones	30-12-2023	20 143
10190	Office Supplies	Fairfield	United States	Erica Bern	GBC	30-12-2023	20 115
10191	Office Supplies	Loveland	United States	Jill Matthias	Other	30-12-2023	20 156
10192	Technology	New York City	United States	Patrick O'Donnell	Other	30-12-2023	20 143
10193	Office Supplies	Charlottetown	Canada	Harry Olson	Wilson Jones	30-12-2023	20 143

10194 rows × 19 columns

Reading an Excel File

In [6]:
```
df2=pd.read_excel(r"D:\Sid 17-03-2025\SIDDHARTH BOSE\Sid's Laptop Data\Sid's Lap
```

In [7]:
```
df2
```

Out[7]:

	Unnamed: 0	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4
0	NaN	For 500 Gram Glass Bottle	NaN	NaN	NaN
1	S.No.	Ketchup Brands	MRP	Cost with Retailer Margin (10%)	Cost with Distributor Margin (3%)
2	NaN	NaN	NaN	NaN	NaN
3	1	Kissan	105	95.45	92.3
4	2	Maggi- Rich Tomato Sauce	105	95.45	92.3
5	3	Maggi- Hot and Sweet	108	98.18	94.94
6	4	Maggi- Chilli Sauce	99	90	87.03
7	5	Maggi- No Onion No Garlic	107	97.27	94.06
8	6	Veeba	93	84.55	81.76
9	NaN	NaN	NaN	NaN	NaN
10	NaN	For 450 Gram Spout Packing	NaN	NaN	NaN
11	S.No.	Ketchup Brands	MRP	Cost with Retailer Margin (10%)	Cost with Distributor Margin (3%)
12	NaN	NaN	NaN	NaN	NaN
13	1	Kissan	NaN	NaN	NaN
14	2	Del Monte	95	86.36	83.51
15	3	Wingreens	75	68.18	65.93
16	NaN	NaN	NaN	NaN	NaN
17	NaN	For 1 KG Spout Packing	NaN	NaN	NaN
18	S.No.	Ketchup Brands	MRP	Cost with Retailer Margin (10%)	Cost with Distributor Margin (3%)
19	NaN	NaN	NaN	NaN	NaN
20	1	Kissan	120	109.09	105.49
21	2	Veeba	129	117.27	113.4
22	3	Del Monte- Classic Blend (950 G)	135	122.73	118.68
23	4	Smith & Jones	130	118.18	114.28
24	5	Cremica	130	118.18	114.28

	Unnamed: 0	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4
25	6	Del Monte- No Onion No Garlic (950 G)	135	122.73	118.68
26	7	Nilon's (900 Gram)	125	113.64	109.89
27	8	Wingreens	125	113.64	109.89

In [8]:
```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 28 entries, 0 to 27
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Unnamed: 0  20 non-null     object
 1   Unnamed: 1  23 non-null     object
 2   Unnamed: 2  19 non-null     object
 3   Unnamed: 3  19 non-null     object
 4   Unnamed: 4  19 non-null     object
dtypes: object(5)
memory usage: 1.2+ KB
```

In [9]:
```
df2.describe()
```

Out[9]:

	Unnamed: 0	Unnamed: 1	Unnamed: 2	Unnamed: 3	Unnamed: 4
count	20	23	19	19	19
unique	9	18	13	13	13
top	S.No.	Kissan	MRP	Cost with Retailer Margin (10%)	Cost with Distributor Margin (3%)
freq	3	3	3	3	3

In [11]:
```
df3=pd.read_xml(r"D:\Sid 17-03-2025\SIDDHARTH BOSE\FSDS & GEN AI\March\24th - Ka
```

In [12]:
```
df3
```

Out[12]:

	id	firstName	lastName	position	department	hireDate	salary
0	1	FirstName1	LastName1	Project Manager	Operations	2019-02-01	71000
1	2	FirstName2	LastName2	Software Engineer	IT	2019-03-01	72000
2	3	FirstName3	LastName3	Project Manager	Operations	2019-04-01	73000
3	4	FirstName4	LastName4	Software Engineer	IT	2019-05-01	74000
4	5	FirstName5	LastName5	Project Manager	Operations	2019-06-01	70000
...	...	...	...	...	...	...	...
12495	12496	FirstName12496	LastName12496	Software Engineer	IT	2019-05-01	71000
12496	12497	FirstName12497	LastName12497	Project Manager	Operations	2019-06-01	72000
12497	12498	FirstName12498	LastName12498	Software Engineer	IT	2019-07-01	73000
12498	12499	FirstName12499	LastName12499	Project Manager	Operations	2019-08-01	74000
12499	12500	FirstName12500	LastName12500	Software Engineer	IT	2019-09-01	70000

12500 rows × 8 columns

In [13]:
```
df3.head()
```

Out[13]:

	id	firstName	lastName	position	department	hireDate	salary	projects
0	1	FirstName1	LastName1	Project Manager	Operations	2019-02-01	71000	NaN
1	2	FirstName2	LastName2	Software Engineer	IT	2019-03-01	72000	NaN
2	3	FirstName3	LastName3	Project Manager	Operations	2019-04-01	73000	NaN
3	4	FirstName4	LastName4	Software Engineer	IT	2019-05-01	74000	NaN
4	5	FirstName5	LastName5	Project Manager	Operations	2019-06-01	70000	NaN

In [14]:
```
df3.describe()
```

Out[14]:

	id	salary	projects
count	12500.000000	12500.000000	0.0
mean	6250.500000	72000.000000	NaN
std	3608.583517	1414.270134	NaN
min	1.000000	70000.000000	NaN
25%	3125.750000	71000.000000	NaN
50%	6250.500000	72000.000000	NaN
75%	9375.250000	73000.000000	NaN
max	12500.000000	74000.000000	NaN

In [15]:
```
df3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12500 entries, 0 to 12499
Data columns (total 8 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   id          12500 non-null  int64  
 1   firstName   12500 non-null  object 
 2   lastName    12500 non-null  object 
 3   position    12500 non-null  object 
 4   department  12500 non-null  object 
 5   hireDate    12500 non-null  object 
 6   salary      12500 non-null  int64  
 7   projects    0 non-null      float64
dtypes: float64(1), int64(2), object(5)
memory usage: 781.4+ KB
```

In [16]:
```
df3.isnull()
```

Out[16]:

	id	firstName	lastName	position	department	hireDate	salary	projects
0	False	False	False	False	False	False	False	True
1	False	False	False	False	False	False	False	True
2	False	False	False	False	False	False	False	True
3	False	False	False	False	False	False	False	True
4	False	False	False	False	False	False	False	True
...	...	...	...	...	...	...	...	...
12495	False	False	False	False	False	False	False	True
12496	False	False	False	False	False	False	False	True
12497	False	False	False	False	False	False	False	True
12498	False	False	False	False	False	False	False	True
12499	False	False	False	False	False	False	False	True

12500 rows × 8 columns

In [17]:
```
df3.isnull().any()
```

Out[17]:
```
id            False
firstName     False
lastName      False
position      False
department    False
hireDate      False
salary        False
projects       True
dtype: bool
```