DATA ANALYST OR DATASCIENCE MOST OF DATASET (EXCELSHEET) WE WILL WORK WITH CSV COMMA SEPARATED VALUE len() in pd .columns .shape .isnull() .isnull().sum() .isna() = .isnull() .head() .tail() .dtype() .categorical()

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

In [13]:
```
pd.__version__
```

Out[13]:
```
'2.2.2'
```

In [21]:
```
df= pd.read_csv(r'C:\Data science\20th- basic python\20th- basic python\Sample - Su
```

In [23]:
```
df
```

Out[23]:

	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 [27]:
```
len(df)
```

Out[27]:
```
10194
```

In [39]:
```
df.columns
```

Out[39]:
```
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 [41]:
```
df.shape
```

Out[41]:
```
(10194, 19)
```

In [43]:
```
df.isnull  #
```

Out[43]:
```
<bound method DataFrame.isnull of              Category           City Country/Re
gion      Customer Name  \
0      Office Supplies        Houston  United States      Darren Powers
1      Office Supplies     Naperville  United States      Phillina Ober
2      Office Supplies     Naperville  United States      Phillina Ober
3      Office Supplies     Naperville  United States      Phillina Ober
4      Office Supplies   Philadelphia  United States         Mick Brown
...                ...            ...            ...                ...
10189  Office Supplies  New York City  United States  Patrick O'Donnell
10190  Office Supplies      Fairfield  United States         Erica Bern
10191  Office Supplies       Loveland  United States      Jill Matthias
10192       Technology  New York City  United States  Patrick O'Donnell
10193  Office Supplies  Charlottetown         Canada        Harry Olson

       Manufacturer  Order Date        Order ID Postal Code  \
0      Message Book  03-01-2020  US-2020-103800       77095
1               GBC  04-01-2020  US-2020-112326       60540
2             Avery  04-01-2020  US-2020-112326       60540
3             SAFCO  04-01-2020  US-2020-112326       60540
4             Avery  05-01-2020  US-2020-141817       19143
...             ...         ...             ...         ...
10189  Wilson Jones  30-12-2023  US-2023-143259       10009
10190           GBC  30-12-2023  US-2023-115427       94533
10191         Other  30-12-2023  US-2023-156720       80538
10192         Other  30-12-2023  US-2023-143259       10009
10193  Wilson Jones  30-12-2023  CA-2023-143500         C0A

                                          Product Name   Region  \
0      Message Book, Wirebound, Four 5 1/2" X 4" Form...  Central
1             GBC Standard Plastic Binding Systems Combs  Central
2                                             Avery 508  Central
3                          SAFCO Boltless Steel Shelving  Central
4      Avery Hi-Liter EverBold Pen Style Fluorescent ...     East
...                                                 ...      ...
10189              Wilson Jones Legal Size Ring Binders     East
10190                                GBC Binding covers     West
10191                               Bagged Rubber Bands     West
10192                          Gear Head AU3700S Headset     East
10193                        Wilson Jones Impact Binders     East

           Segment   Ship Date       Ship Mode        State/Province  \
0         Consumer  07-01-2020  Standard Class                 Texas
1      Home Office  08-01-2020  Standard Class              Illinois
2      Home Office  08-01-2020  Standard Class              Illinois
3      Home Office  08-01-2020  Standard Class              Illinois
4         Consumer  12-01-2020  Standard Class          Pennsylvania
...            ...         ...             ...                   ...
10189     Consumer  03-01-2024  Standard Class              New York
10190    Corporate  03-01-2024  Standard Class            California
10191     Consumer  03-01-2024  Standard Class              Colorado
10192     Consumer  03-01-2024  Standard Class              New York
10193     Consumer  03-01-2024  Standard Class  Prince Edward Island

      Sub-Category  Discount   Profit  Quantity    Sales
0            Paper       0.2   5.5512         2   16.448
1          Binders       0.8  -5.4870         2    3.540
```

```
2           Labels       0.2   4.2717         3   11.784
3          Storage       0.2 -64.7748         3  272.736
4              Art       0.2   4.8840         3   19.536
...            ...       ...      ...       ...      ...
10189      Binders       0.2  19.7910         3   52.776
10190      Binders       0.2   6.4750         2   20.720
10191    Fasteners       0.2  -0.6048         3    3.024
10192       Phones       0.0   2.7279         7   90.930
10193      Binders       0.2  -0.6048         3    3.024

[10194 rows x 19 columns]>
```

In [45]:
```
df.isnull()   #missing values
```

Out[45]:

	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 [47]:
```
df.isnull() .sum()
```

Out[47]:
```
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 [49]:
```
df.isna()
```

Out[49]:

	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 [55]:
```
df.head () #top 5
```

Out[55]:

	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 [57]:
```
df.tail()   #last 5 row in column
```

Out[57]:

	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 [59]:
```
df.head(2) #top 5 row in column
```

Out[59]:

	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 [69]:
```
df[:]
```

Out[69]:

	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 [71]:
```
df[::-1]
```

Out[71]:

	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 [73]:
```
df[1:10:3]
```

Out[73]:

	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 [77]:
```
df.describe()
```

Out[77]:

	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 [83]:
```
df.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 [95]:
```
df['Category']
```

Out[95]:
```
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 [97]:
```
df[['Category','City']]
```

Out[97]:

	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 [103…

```
df.dtypes
```

Out[103…

```
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 [121…

```
len(df.dtypes)
```

Out[121…

```
19
```

In [129…

```
len(df.columns)
```

Out[129…

```
19
```

In [151…
```
df.head()
```

Out[151…

	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 [141…
```
print(len(df.columns))
```

```
19
```

In [153…
```
df['Category']
```

Out[153…
```
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 [169…
```
df ['Category']
```

Out[169…

```
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 [173…

```
df[['Customer Name','City']]
```

Out[173…

	Customer Name	City
0	Darren Powers	Houston
1	Phillina Ober	Naperville
2	Phillina Ober	Naperville
3	Phillina Ober	Naperville
4	Mick Brown	Philadelphia
...	...	...
10189	Patrick O'Donnell	New York City
10190	Erica Bern	Fairfield
10191	Jill Matthias	Loveland
10192	Patrick O'Donnell	New York City
10193	Harry Olson	Charlottetown

10194 rows × 2 columns

In [182…

```
df_categorical = df[['Category', 'City', 'Country/Region', 'Customer Name', 'Manufa
       'Order Date', 'Order ID', 'Postal Code', 'Product Name', 'Region',
       'Segment', 'Ship Date', 'Ship Mode', 'State/Province', 'Sub-Category']]
```

In [186…

```
df_categorical
```

Out[186…

	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 × 15 columns

In [192…
```
df_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 [196…
```
df_numerical = df[['Discount','Profit','Quantity','Sales']]
df_numerical
```

Out[196…

	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 [ ]:

```

```