DATA ANALYSIS USING PYTHON PART -I

1)Importing python libraries

2]:

```
1 import numpy as np
2 import pandas as pd
3 import matplotlib as plt
4 import seaborn as sns
```

2)Reading the dataset

```
df=pd.read_csv("D:\DOWNLOAD\salesdata (1).csv")
df
```

3)HEAD(),TAIL(),SAMPLE()

```
1 df.head(2)
```

	Order Date	Customer Name	State	Category
0	2014-01-03	Darren Powers	Texas	Office Supplies
1	2014-01-04	Phillina Ober	Illinois	Office Supplies

```
1 df.tail(2)
```

	Order Date	Customer Name	State	Category
9992	2017-12-30	Erica Bern	California	Office Supplies
9993	2017-12-30	Jill Matthias	Colorado	Office Supplies

```
1 df.sample(2)
```

	Order Date	Customer Name	State	Category
5161	2016-07-22	Dionis Lloyd	California	Office Supplies
5664	2016-09-20	Greg Maxwell	California	Office Supplies

4)Check for duplicates if present then drop it

```
: 1 df.duplicated().sum()
```

```
: 1
```

```
1 df.drop_duplicates(inplace=True)
2 df
```

5)CHECKING FOR DATATYPES AND NO. OF ROWS AND COLUMNS (DF.INFO())

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Order Date      9994 non-null   object
 1   Customer Name   9994 non-null   object
 2   State           9994 non-null   object
 3   Category        9994 non-null   object
 4   Sub-Category    9994 non-null   object
 5   Sales           9994 non-null   float64
 6   Quantity        9994 non-null   int64
 7   Profit          9994 non-null   float64
dtypes: float64(2), int64(1), object(5)
memory usage: 624.8+ KB
```

```
1 df.shape
```

```
(9994, 8)
```

6)Describe()

```
1 df.describe(include="all")
```

	Order Date	Customer Name	State	Category
count	9994	9994	9994	9994
unique	1237	793	49	3
top	05-09-2016	William Brown	California	Office Supplies
freq	38	37	2001	6026
mean	NaN	NaN	NaN	NaN
std	NaN	NaN	NaN	NaN
min	NaN	NaN	NaN	NaN
25%	NaN	NaN	NaN	NaN
50%	NaN	NaN	NaN	NaN
75%	NaN	NaN	NaN	NaN
max	NaN	NaN	NaN	NaN

7)checking for no of null values in each column

```
1 df.isnull().sum()
```

```
Order Date       0
Customer Name    0
State            0
Category         0
Sub-Category     0
Sales            0
Quantity         0
Profit           0
dtype: int64
```

8)If null values are present drop it

```
1 df.dropna(how="all",inplace=True)
```

9)Change datatype of columns if necessary

```
1 df["Order Date"]=pd.to_datetime(df["Order Date"],format="%d-%m-%Y")
```

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9994 entries, 0 to 9993
Data columns (total 8 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Order Date     9994 non-null   datetime64[ns]
 1   Customer Name  9994 non-null   object
 2   State          9994 non-null   object
 3   Category       9994 non-null   object
 4   Sub-Category   9994 non-null   object
 5   Sales          9994 non-null   float64
 6   Quantity       9994 non-null   int64
 7   Profit         9994 non-null   float64
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 624.8+ KB
```

10)Change the name of column if necessary

```
1  df.rename(columns={"Order Date":"Order_Date"})
```

Order_Date	Customer Name	State	Category