Working on CSV file

In [96]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [97]:
```
df = pd.read_csv('Data/data.csv')
```

In [98]:
```
df.head(3)
```

Out[98]:

	Car	Model	Volume	Weight	CO2
0	Toyoty	Aygo	1000	790	99
1	Mitsubishi	Space Star	1200	1160	95
2	Skoda	Citigo	1000	929	95

In [99]:
```
df.tail(3)
```

Out[99]:

	Car	Model	Volume	Weight	CO2
33	BMW	216	1600	1390	108
34	Opel	Zafira	1600	1405	109
35	Mercedes	SLK	2500	1395	120

In [100…
```
df.shape
```

Out[100… (36, 5)

In [101…
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36 entries, 0 to 35
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   Car     36 non-null     object
 1   Model   36 non-null     object
 2   Volume  36 non-null     int64 
 3   Weight  36 non-null     int64 
 4   CO2     36 non-null     int64 
dtypes: int64(3), object(2)
memory usage: 1.5+ KB
```

In [102…
```
df.describe()
```

Out[102…

	Volume	Weight	CO2
count	36.000000	36.000000	36.000000
mean	1611.111111	1292.277778	102.027778
std	388.975047	242.123889	7.454571
min	900.000000	790.000000	90.000000
25%	1475.000000	1117.250000	97.750000
50%	1600.000000	1329.000000	99.000000
75%	2000.000000	1418.250000	105.000000
max	2500.000000	1746.000000	120.000000

In [103…
```
df.isnull().sum()
```

Out[103…
```
Car       0
Model     0
Volume    0
Weight    0
CO2       0
dtype: int64
```

Working on Excel file

In [105…
```
ex = pd.read_excel('Data/Financial Sample.xlsx')
```

In [106…
```
ex.head(3)
```

Out[106…

	Segment	Country	Product	Discount Band	Units Sold	Manufacturing Price	Sale Price	Gross Sales	Disco
0	Government	Canada	Carretera	NaN	1618.5	3	20	32370.0	
1	Government	Germany	Carretera	NaN	1321.0	3	20	26420.0	
2	Midmarket	France	Carretera	NaN	2178.0	3	15	32670.0	

In [107…
```
ex.tail(3)
```

Out[107…

	Segment	Country	Product	Discount Band	Units Sold	Manufacturing Price	Sale Price	Gross Sales	Disc
697	Government	Mexico	Montana	High	1368.0	5	7	9576.0	1
698	Government	Canada	Paseo	High	723.0	10	7	5061.0	
699	Channel Partners	United States of America	VTT	High	1806.0	250	12	21672.0	3

In [108…
```
ex.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 700 entries, 0 to 699
Data columns (total 16 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Segment              700 non-null    object
 1   Country              700 non-null    object
 2   Product              700 non-null    object
 3   Discount Band        647 non-null    object
 4   Units Sold           700 non-null    float64
 5   Manufacturing Price  700 non-null    int64
 6   Sale Price           700 non-null    int64
 7   Gross Sales          700 non-null    float64
 8   Discounts            700 non-null    float64
 9    Sales               700 non-null    float64
 10  COGS                 700 non-null    float64
 11  Profit               700 non-null    float64
 12  Date                 700 non-null    datetime64[ns]
 13  Month Number         700 non-null    int64
 14  Month Name           700 non-null    object
 15  Year                 700 non-null    int64
dtypes: datetime64[ns](1), float64(6), int64(4), object(5)
memory usage: 87.6+ KB
```

In [109…
```
ex.shape
```

Out[109…
```
(700, 16)
```

In [110…
```
ex.describe()
```

Out[110…

	Units Sold	Manufacturing Price	Sale Price	Gross Sales	Discounts	Sales
count	700.000000	700.000000	700.000000	7.000000e+02	700.000000	7.000000e+02
mean	1608.294286	96.477143	118.428571	1.827594e+05	13150.354629	1.696091e+05
min	200.000000	3.000000	7.000000	1.799000e+03	0.000000	1.655080e+03
25%	905.000000	5.000000	12.000000	1.739175e+04	800.320000	1.592800e+04
50%	1542.500000	10.000000	20.000000	3.798000e+04	2585.250000	3.554020e+04
75%	2229.125000	250.000000	300.000000	2.790250e+05	15956.343750	2.610775e+05
max	4492.500000	260.000000	350.000000	1.207500e+06	149677.500000	1.159200e+06
std	867.427859	108.602612	136.775515	2.542623e+05	22962.928775	2.367263e+05

In [111…

```
ex.describe(include="object")
```

Out[111…

	Segment	Country	Product	Discount Band	Month Name
count	700	700	700	647	700
unique	5	5	6	3	12
top	Government	Canada	Paseo	High	October
freq	300	140	202	245	140

In [112…

```
ex.isnull().sum()
```

Out[112…
```
Segment                0
Country                0
Product                0
Discount Band         53
Units Sold             0
Manufacturing Price    0
Sale Price             0
Gross Sales            0
Discounts              0
 Sales                 0
COGS                   0
Profit                 0
Date                   0
Month Number           0
Month Name             0
Year                   0
dtype: int64
```

in this dataset Discount Band is containing null value

to handel this

first check data type

according to datatype fill the value

mean median Mode

In [113…
```
ex.dtypes
```

Out[113…
```
Segment                      object
Country                      object
Product                      object
Discount Band                object
Units Sold                  float64
Manufacturing Price           int64
Sale Price                    int64
Gross Sales                 float64
Discounts                   float64
 Sales                      float64
COGS                        float64
Profit                      float64
Date                 datetime64[ns]
Month Number                  int64
Month Name                   object
Year                          int64
dtype: object
```

In [117…
```
(ex.isnull().sum()/ex.shape[0])*100
```

Out[117…
```
Segment                0.000000
Country                0.000000
Product                0.000000
Discount Band          7.571429
Units Sold             0.000000
Manufacturing Price    0.000000
Sale Price             0.000000
Gross Sales            0.000000
Discounts              0.000000
 Sales                 0.000000
COGS                   0.000000
Profit                 0.000000
Date                   0.000000
Month Number           0.000000
Month Name             0.000000
Year                   0.000000
dtype: float64
```

hence we see that here only 7% of data is missing

so we can fill it or drop it

it will not much effect on dataset

In [118…
```
ex.dropna(inplace=True)
```

In [119…
```
ex.isnull().sum()
```

Out[119…
```
Segment                0
Country                0
Product                0
Discount Band          0
Units Sold             0
Manufacturing Price    0
Sale Price             0
Gross Sales            0
Discounts              0
 Sales                 0
COGS                   0
Profit                 0
Date                   0
Month Number           0
Month Name             0
Year                   0
dtype: int64
```

In [121…
```
ex.isnull().sum()
```

Out[121…

```
Segment                0
Country                0
Product                0
Discount Band          0
Units Sold             0
Manufacturing Price    0
Sale Price             0
Gross Sales            0
Discounts              0
 Sales                 0
COGS                   0
Profit                 0
Date                   0
Month Number           0
Month Name             0
Year                   0
dtype: int64
```

In [ ]: