```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

```
df=pd.read_csv('/content/sales_data.csv')
```

```
df.head()
```

	date	product	category	price	quantity	revenue
0	2022-01-01	Smartphone	Electronics	600.0	10.0	6000.0
1	2022-01-01	Laptop	Electronics	1200.0	5.0	6000.0
2	2022-01-02	T-Shirt	Clothing	20.0	50.0	1000.0
3	2022-01-03	Headphones	Electronics	100.0	20.0	2000.0
4	2022-01-04	T-Shirt	Clothing	20.0	25.0	500.0

```
df.tail()
```

	date	product	category	price	quantity	revenue
364	2022-12-27	Watch	Accessories	150.0	5.0	750.0
365	2022-12-28	Coat	Clothing	100.0	5.0	500.0
366	2022-12-29	Headphones	Electronics	100.0	10.0	1000.0
367	2022-12-30	Smartphone	Electronics	600.0	11.0	6600.0
368	2022-12-31	Hoodie	Clothing	40.0	30.0	1200.0

```
df['date']=pd.to_datetime(df['date']) # to convert all date in right format
```

```
df[['date']] # to veiw date as dataframe
```

	date
0	2022-01-01
1	2022-01-01
2	2022-01-02
3	2022-01-03
4	2022-01-04
...	...
364	2022-12-27
365	2022-12-28
366	2022-12-29
367	2022-12-30
368	2022-12-31

369 rows × 1 columns

```
# to see if there's  anull value
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 369 entries, 0 to 368
Data columns (total 6 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   date      369 non-null    datetime64[ns]
 1   product   369 non-null    object
 2   category  369 non-null    object
 3   price     367 non-null    float64
 4   quantity  368 non-null    float64
 5   revenue   368 non-null    float64
dtypes: datetime64[ns](1), float64(3), object(2)
memory usage: 17.4+ KB
```

```
# lets see if there is duplicate
df.duplicated().sum()
```

```
1
```

```
df.loc[df.duplicated()]
```

	date	product	category	price	quantity	revenue
276	2022-10-01	Hoodie	Clothing	40.0	30.0	1200.0

```
# lets drop duplicates
df.drop_duplicates(inplace=True)
```

```
# lets check
df.duplicated().sum()
```

```
0
```

```
# lets see null values
df.isnull().sum()
```

```
date        0
product     0
category    0
price       2
quantity    1
revenue     1
dtype: int64
```

```
df.loc[df['price'].isnull()]
```

	date	product	category	price	quantity	revenue
193	2022-07-11	Watch	Accessories	NaN	15.0	2250.0
320	2022-11-13	Wallet	Accessories	NaN	35.0	1050.0

```
# since revenue = price * quantity , price = revenue / quantity
df['price'].fillna(df['revenue']/df['quantity'],inplace=True)
```

```
df.loc[df['quantity'].isnull()]
```

	date	product	category	price	quantity	revenue
122	2022-05-01	Smartphone	Electronics	600.0	NaN	6600.0

```
# quantity = revenue / price
df['quantity'].fillna(df['revenue']/df['price'],inplace=True)
```

```
df.loc[df['revenue'].isnull()]
```

	date	product	category	price	quantity	revenue
96	2022-04-05	Smartwatch	Accessories	200.0	10.0	NaN

```
# revenue = price * quantity
df['revenue'].fillna(df['price']*df['quantity'],inplace=True)
```

```
# lets check
df.isnull().sum()
```

```
date        0
product     0
category    0
price       0
quantity    0
revenue     0
dtype: int64
```

```
# lets do some statistical tests
df.describe()
```

	price	**quantity**	**revenue**
count	368.000000	368.000000	368.000000
mean	211.032609	14.513587	2062.853261
std	227.068797	8.559765	1910.403972
min	20.000000	3.000000	300.000000
25%	50.000000	8.000000	800.000000
50%	100.000000	12.000000	1200.000000
75%	250.000000	20.000000	2400.000000
max	1200.000000	50.000000	7200.000000