In [ ]:
```
#import libraries
!pip install kaggle
import kaggle
#download dataset using kaggle api
!kaggle datasets download retail-orders -f orders.csv
```

In [15]:
```
#extract file from zip file
import zipfile
zip_ref = zipfile.ZipFile(r'C:\Users\HP\Downloads\archive.zip')
zip_ref.extractall()
zip_ref.close()
```

In [6]:
```
#read data from the file and handle null values
import pandas as pd
df = pd.read_csv(r'C:\Users\HP\Downloads\archive\orders.csv', na_values=['Not Available','unknown'])
df.head(20)
```

Out[6]:

	Order Id	Order Date	Ship Mode	Segment	Country	City	State	Postal Code	Region	Category	Sub Category	Product Id	cost price	List Price	Quantity	Discount Percent
0	1	2023-03-01	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	FUR-BO-10001798	240	260	2	2
1	2	2023-08-15	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	FUR-CH-10000454	600	730	3	3
2	3	2023-01-10	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	OFF-LA-10000240	10	10	2	5
3	4	2022-06-18	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	FUR-TA-10000577	780	960	5	2
4	5	2022-07-13	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	OFF-ST-10000760	20	20	2	5
5	6	2022-03-13	NaN	Consumer	United States	Los Angeles	California	90032	West	Furniture	Furnishings	FUR-FU-10001487	50	50	7	3
6	7	2022-12-28	Standard Class	Consumer	United States	Los Angeles	California	90032	West	Office Supplies	Art	OFF-AR-10002833	10	10	4	3
7	8	2022-01-25	Standard Class	Consumer	United States	Los Angeles	California	90032	West	Technology	Phones	TEC-PH-10002275	860	910	6	5
8	9	2023-03-23	NaN	Consumer	United States	Los Angeles	California	90032	West	Office Supplies	Binders	OFF-BI-10003910	20	20	3	2
9	10	2023-05-16	Standard Class	Consumer	United States	Los Angeles	California	90032	West	Office Supplies	Appliances	OFF-AP-10002892	90	110	5	3
10	11	2023-03-31	NaN	Consumer	United States	Los Angeles	California	90032	West	Furniture	Tables	FUR-TA-10001539	1470	1710	9	3
11	12	2023-12-25	NaN	Consumer	United States	Los Angeles	California	90032	West	Technology	Phones	TEC-PH-10002033	750	910	4	3
12	13	2022-02-11	Standard Class	Consumer	United States	Concord	North Carolina	28027	South	Office Supplies	Paper	OFF-PA-10002365	20	20	3	3
13	14	2023-07-18	Standard Class	Consumer	United States	Seattle	Washington	98103	West	Office Supplies	Binders	OFF-BI-10003656	360	410	3	2
14	15	2023-11-09	NaN	Home Office	United States	Fort Worth	Texas	76106	Central	Office Supplies	Appliances	OFF-AP-10002311	60	70	5	5
15	16	2022-06-18	Standard Class	Home Office	United States	Fort Worth	Texas	76106	Central	Office Supplies	Binders	OFF-BI-10000756	0	0	3	5
16	17	2022-02-04	Standard Class	Consumer	United States	Madison	Wisconsin	53711	Central	Office Supplies	Storage	OFF-ST-10004186	610	670	6	3
17	18	2023-08-04	Second Class	Consumer	United States	West Jordan	Utah	84084	West	Office Supplies	Storage	OFF-ST-10000107	60	60	2	4
18	19	2022-01-23	Second Class	Consumer	United States	San Francisco	California	94109	West	Office Supplies	Art	OFF-AR-10003056	10	10	2	4
19	20	2022-01-11	Second Class	Consumer	United States	San Francisco	California	94109	West	Technology	Phones	TEC-PH-10001949	170	210	3	3

In [7]:
```
df['Ship Mode'].unique()
```

Out[7]:
```
array(['Second Class', 'Standard Class', nan, 'First Class', 'Same Day'],
      dtype=object)
```

In [8]:
```
#rename columns name ..make them lower case and repace sppace with underscore
#df.rename(columns= {'Order Id' : 'order_id', 'City' : 'city'})
df.columns = df.columns.str.lower()
df.columns = df.columns.str.replace(' ', '_' )
df.columns
df.head(5)
```

Out[8]:

	order_id	order_date	ship_mode	segment	country	city	state	postal_code	region	category	sub_category	product_id	cost_price	list_price	quantity	discount_percent
0	1	2023-03-01	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	FUR-BO-10001798	240	260	2	2
1	2	2023-08-15	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	FUR-CH-10000454	600	730	3	3
2	3	2023-01-10	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	OFF-LA-10000240	10	10	2	5
3	4	2022-06-18	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	FUR-TA-10000577	780	960	5	2
4	5	2022-07-13	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	OFF-ST-10000760	20	20	2	5

In [9]:
```
#derive new columns discount, sales_price and proft
df['discount'] = df['list_price']*df['discount_percent']*0.01
df.head(5)
df['sales_price'] = df['list_price']-df['discount']
df
df['profit'] = df['sales_price'] - df['cost_price']
df.head(5)
```

Out[9]:

	order_id	order_date	ship_mode	segment	country	city	state	postal_code	region	category	sub_category	product_id	cost_price	list_price	quantity	discount_percent	discount	sales_price	profit
0	1	2023-03-01	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	FUR-BO-10001798	240	260	2	2	5.2	254.8	14.8
1	2	2023-08-15	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	FUR-CH-10000454	600	730	3	3	21.9	708.1	108.1
2	3	2023-01-10	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	OFF-LA-10000240	10	10	2	5	0.5	9.5	-0.5
3	4	2022-06-18	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	FUR-TA-10000577	780	960	5	2	19.2	940.8	160.8
4	5	2022-07-13	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	OFF-ST-10000760	20	20	2	5	1.0	19.0	-1.0

In [10]:
```
df.dtypes
```

Out[10]:
```
order_id              int64
order_date           object
ship_mode            object
segment              object
country              object
city                 object
state                object
postal_code           int64
region               object
category             object
sub_category         object
product_id           object
cost_price            int64
list_price            int64
quantity              int64
discount_percent      int64
discount            float64
sales_price         float64
profit              float64
dtype: object
```

In [23]:
```
#convert order_date from object data type to datetime
df['order_date'] = pd.to_datetime(df['order_date'], format= "%Y-%m-%d")
```

In [11]:
```
df.dtypes
```

Out[11]:
```
order_id              int64
order_date           object
ship_mode            object
segment              object
country              object
city                 object
state                object
postal_code           int64
region               object
category             object
sub_category         object
product_id           object
cost_price            int64
list_price            int64
quantity              int64
discount_percent      int64
discount            float64
sales_price         float64
profit              float64
dtype: object
```

In [12]:
```
#drop cost_price, list_price and discount_percent columns
df.drop(columns = ['cost_price', 'list_price', 'discount_percent'], axis = 1 , inplace=True, errors = 'ignore')
df.columns
```

Out[12]:
```
Index(['order_id', 'order_date', 'ship_mode', 'segment', 'country', 'city',
       'state', 'postal_code', 'region', 'category', 'sub_category',
       'product_id', 'quantity', 'discount', 'sales_price', 'profit'],
      dtype='object')
```

In [17]:
```
#Load the data into SQL Servrusing replace option
import sqlalchemy as sal
engine = sal.create_engine(r'mssql://DESKTOP-V29FOQO\SQLEXPRESS/sales?driver=ODBC+DRIVER+17+FOR+SQL+SERVER')
conn = engine.connect()
```

In [ ]:
```
#Load the data into sql server using append option
df.to_sql('df_orders', con = conn, index = False, if_exists = 'append')
df.columns
```

In [ ]: