Retail Order Analysis Project

Data Cleaning

In [1]:
```
#libraries
import pandas as pd
import mysql.connector
from sqlalchemy import create_engine


file_path = r"C:\Users\nicho\Downloads\archive (7)\orders.csv"

orders = pd.read_csv(file_path)
```

In [2]:
```
#viewing dataset
orders.head(10)
```

Out[2]:

	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	Not Available	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	Not Available	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

In [3]:
```
orders.dtypes
```

Out[3]:
```
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
dtype: object
```

In [4]:
```
orders['Ship Mode'].unique()
```

Out[4]:
```
array(['Second Class', 'Standard Class', 'Not Available', 'unknown',
       'First Class', nan, 'Same Day'], dtype=object)
```

In [5]:
```
#removing Not Available  and unkowon from ship mode replacing with null
orders = pd.read_csv(file_path,na_values=['Not Available','unknown'])
orders['Ship Mode'].unique()
```

Out[5]:
```
array(['Second Class', 'Standard Class', nan, 'First Class', 'Same Day'],
      dtype=object)
```

In [6]:
```
#column formating
orders.columns = orders.columns.str.lower()
orders.columns = orders.columns.str.replace(' ','_')
orders.columns
```

Out[6]:
```
Index(['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'],
      dtype='object')
```

In [7]:
```
orders.head(3)
```

Out[7]:

	order_id	order_date	ship_mode	segment	country	city	state	postal_code	region	category	sub_category	product_id	cost_price	list_p
0	1	2023-03-01	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	FUR-BO-10001798	240	
1	2	2023-08-15	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	FUR-CH-10000454	600	
2	3	2023-01-10	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	OFF-LA-10000240	10	

In [8]:
```
#Converting 'order_date' to datetime format
orders['order_date'] = pd.to_datetime(orders['order_date'], format="%Y-%m-%d")
orders.dtypes
```

Out[8]:
```
order_id                   int64
order_date          datetime64[ns]
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
dtype: object
```

In [9]:
```
# Deriving new columns: discount, sale_price, and profit
orders['discount'] = orders['list_price'] * orders['discount_percent'] * 0.01
orders['sale_price'] = orders['list_price'] - orders['discount']
orders['profit'] = orders['sale_price'] - orders['cost_price']
orders.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_p
0	1	2023-03-01	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	FUR-BO-10001798	240	
1	2	2023-08-15	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	FUR-CH-10000454	600	
2	3	2023-01-10	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	OFF-LA-10000240	10	
3	4	2022-06-18	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	FUR-TA-10000577	780	
4	5	2022-07-13	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	OFF-ST-10000760	20	

Moving data to Sql database

In [10]:
```
engine = create_engine('mysql+mysqlconnector://root:NJohnson884@127.0.0.1/retail_orders')
orders.to_sql('orders', con=engine, if_exists='replace', index=False)
```

Out[10]:
```
9994
```

In [11]:
```
connection = mysql.connector.connect(
    host='127.0.0.1',
    user='root',
    password='NJohnson884',
    database='retail_orders'
)

cursor = connection.cursor()

#select all from orders
query = "SELECT * FROM orders"

cursor = connection.cursor()
cursor.execute(query)
columns = [desc[0] for desc in cursor.description]
results = cursor.fetchall()
```

In [12]:
```
orders_in_sql = pd.DataFrame(results, columns=columns)
orders_in_sql
```

Out[12]:

	order_id	order_date	ship_mode	segment	country	city	state	postal_code	region	category	sub_category	product_id	cost_price
0	1	2023-03-01	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	FUR-BO-10001798	240
1	2	2023-08-15	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Chairs	FUR-CH-10000454	600
2	3	2023-01-10	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies	Labels	OFF-LA-10000240	10
3	4	2022-06-18	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Furniture	Tables	FUR-TA-10000577	780
4	5	2022-07-13	Standard Class	Consumer	United States	Fort Lauderdale	Florida	33311	South	Office Supplies	Storage	OFF-ST-10000760	20
...	...	...	...	...	...	...	...	...	...	...	...	...	...
9989	9990	2023-02-18	Second Class	Consumer	United States	Miami	Florida	33180	South	Furniture	Furnishings	FUR-FU-10001889	30
9990	9991	2023-03-17	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Furniture	Furnishings	FUR-FU-10000747	70
9991	9992	2022-08-07	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Technology	Phones	TEC-PH-10003645	220
9992	9993	2022-11-19	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Office Supplies	Paper	OFF-PA-10004041	30
9993	9994	2022-07-17	Second Class	Consumer	United States	Westminster	California	92683	West	Office Supplies	Appliances	OFF-AP-10002684	210

9994 rows × 19 columns

In [13]:
```
connection.close()
```