Retail Orders

This Notebook extracted the data from the kaggle API, for a retail order. Then we clean the data to a neat form and loading the data into SQL Server for solving some of the Business Problems.

In [ ]:

In [3]:
```
!where python
```

```
C:\ProgramData\anaconda3\python.exe
C:\Program Files\Python311\python.exe
C:\Users\ASUS\AppData\Local\Microsoft\WindowsApps\python.exe
```

In [4]:
```
# !pip install kaggle
```

In [5]:
```
import kaggle
print("Kaggle is installed and working!")
```

```
Kaggle is installed and working!
```

In [7]:
```
import kaggle

!kaggle datasets download ankitbansal06/retail-orders -f orders.csv
```

```
'kaggle' is not recognized as an internal or external command,
operable program or batch file.
```

In [ ]:

In [5]:
```
!where kaggle
```

```
INFO: Could not find files for the given pattern(s).
```

In [ ]:

In [8]:
```
from kaggle.api.kaggle_api_extended import KaggleApi

# Authenticate
api = KaggleApi()
api.authenticate()

# Download the file
api.dataset_download_file('ankitbansal06/retail-orders', file_name='orders.csv', path='./')
```

```
Dataset URL: https://www.kaggle.com/datasets/ankitbansal06/retail-orders
```

Out[8]:
```
True
```

In [9]:
```
#extract file from zip file

import zipfile
zip_ref = zipfile.ZipFile('orders.csv.zip')
zip_ref.extractall() # extract file to dir
zip_ref.close() # close file
```

In [26]:
```
# read data from the file and handle the null values

import pandas as pd
df = pd.read_csv('orders.csv', na_values = ['Not Available', 'unknown'])
df['Ship Mode'].unique()
```

Out[26]:
```
array(['Second Class', 'Standard Class', nan, 'First Class', 'Same Day'],
      dtype=object)
```

In [36]:
```
# rename the column names of the dataframe. make them lower case and replace the blank_spaces with "underscore"

df.columns
```

Out[36]:
```
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 [37]:
```
# let's convert all the columns to string and make them all to lower-case

df.columns = df.columns.str.lower()
```

In [38]:
```
df.columns
```

Out[38]:
```
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 [39]:
```
# replace the blank spaces with an undrscore

df.columns = df.columns.str.replace(' ', '_')
```

In [40]:
```
df.columns
```

Out[40]:
```
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 [41]:
```
df.head(5)
```

Out[41]:

	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 [48]:
```
# create new columns: discount, sale price and profit

# df['discount'] = (df['list_price'] * df['discount_percent']) / 100
df['sale_price'] = df['list_price'] - df['discount']
df['profit'] = df['sale_price'] - df['cost_price']
df
```

Out[48]:

	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	sale_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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9989	9990	2023-02-18	Second Class	Consumer	United States	Miami	Florida	33180	South	Furniture	Furnishings	FUR-FU-10001889	30	30	3	4	1.2	28.8	-1.2
9990	9991	2023-03-17	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Furniture	Furnishings	FUR-FU-10000747	70	90	2	4	3.6	86.4	16.4
9991	9992	2022-08-07	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Technology	Phones	TEC-PH-10003645	220	260	2	2	5.2	254.8	34.8
9992	9993	2022-11-19	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Office Supplies	Paper	OFF-PA-10004041	30	30	4	3	0.9	29.1	-0.9
9993	9994	2022-07-17	Second Class	Consumer	United States	Westminster	California	92683	West	Office Supplies	Appliances	OFF-AP-10002684	210	240	2	3	7.2	232.8	22.8

9994 rows × 19 columns

In [50]:
```
# checking out the datetypes of the "date"-related columns and if found any other, simply convert it to date data-type

df.dtypes
```

Out[50]:
```
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
sale_price          float64
profit              float64
dtype: object
```

In [55]:
```
# convert "order_date" column object data-type to "datetime" datatype

df['order_date'] = pd.to_datetime(df['order_date'], format='%Y-%m-%d')
df
```

Out[55]:

	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	sale_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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9989	9990	2023-02-18	Second Class	Consumer	United States	Miami	Florida	33180	South	Furniture	Furnishings	FUR-FU-10001889	30	30	3	4	1.2	28.8	-1.2
9990	9991	2023-03-17	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Furniture	Furnishings	FUR-FU-10000747	70	90	2	4	3.6	86.4	16.4
9991	9992	2022-08-07	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Technology	Phones	TEC-PH-10003645	220	260	2	2	5.2	254.8	34.8
9992	9993	2022-11-19	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Office Supplies	Paper	OFF-PA-10004041	30	30	4	3	0.9	29.1	-0.9
9993	9994	2022-07-17	Second Class	Consumer	United States	Westminster	California	92683	West	Office Supplies	Appliances	OFF-AP-10002684	210	240	2	3	7.2	232.8	22.8

9994 rows × 19 columns

In [56]:
```
df.dtypes
```

Out[56]:
```
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
discount                   float64
sale_price                 float64
profit                     float64
dtype: object
```

In [59]:
```
# drop the columns: 'cost_price', 'list_price', 'discount_percent'

df.drop(columns=['cost_price', 'list_price', 'discount_percent'], inplace=True)
```

In [60]:
```
df
```

Out[60]:

	order_id	order_date	ship_mode	segment	country	city	state	postal_code	region	category	sub_category	product_id	quantity	discount	sale_price	profit
0	1	2023-03-01	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture	Bookcases	FUR-BO-10001798	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	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	2	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	5	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	2	1.0	19.0	-1.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
9989	9990	2023-02-18	Second Class	Consumer	United States	Miami	Florida	33180	South	Furniture	Furnishings	FUR-FU-10001889	3	1.2	28.8	-1.2
9990	9991	2023-03-17	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Furniture	Furnishings	FUR-FU-10000747	2	3.6	86.4	16.4
9991	9992	2022-08-07	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Technology	Phones	TEC-PH-10003645	2	5.2	254.8	34.8
9992	9993	2022-11-19	Standard Class	Consumer	United States	Costa Mesa	California	92627	West	Office Supplies	Paper	OFF-PA-10004041	4	0.9	29.1	-0.9
9993	9994	2022-07-17	Second Class	Consumer	United States	Westminster	California	92683	West	Office Supplies	Appliances	OFF-AP-10002684	2	7.2	232.8	22.8

9994 rows × 16 columns

In [61]:
```
# load data into sql server

import sqlalchemy as sal
engine = sal.create_engine('mssql://shashank\SQLEXPRESS/master?driver=ODBC+DRIVER+17+FOR+SQL+SERVER')
conn = engine.connect()
```

```
<>:4: SyntaxWarning: invalid escape sequence '\S'
<>:4: SyntaxWarning: invalid escape sequence '\S'
C:\Users\ASUS\AppData\Local\Temp\ipykernel_16408\1528994280.py:4: SyntaxWarning: invalid escape sequence '\S'
  engine = sal.create_engine('mssql://shashank\SQLEXPRESS/master?driver=ODBC+DRIVER+17+FOR+SQL+SERVER')
```

In [64]:
```
# create the connection

df.to_sql('df_orders', con=conn, index=False, if_exists='append')
```

Out[64]:
```
38
```

since the datatype, by default were max in size, we need to create the table in mssql manually and "append", again.

In [ ]:

In [ ]:

In [ ]: