sales

Report

BY Rizwan Shah

2.0Retail_order_ETL(python+SQL)

September 22, 2024

[1]:
```
import opendatasets as od
```

[2]:
```
link=r'https://www.kaggle.com/datasets/ankitbansal06/retail-orders/orders.csv'
od.download(link)
```

```
Please provide your Kaggle credentials to download this dataset. Learn more:
http://bit.ly/kaggle-creds
Your Kaggle username:

  shahrizwan52

Your Kaggle Key:

  ........

Dataset URL: https://www.kaggle.com/datasets/ankitbansal06/retail-orders
Downloading retail-orders.zip to .\retail-orders

100%|
   | 200k/200k [00:00<00:00, 205kB/s]
```

[1]:
```
import os
os.listdir(r'.\retail-orders')
```

[1]:
```
['orders.csv', 'retail_orders.pbix']
```

[1]:
```
import pandas as pd
df=pd.read_csv(r'C:\Users\Rizwan\Downloads\Shah Rizwan\Python data analyst␣
 ↪project\Retail_orders_ETL(Python+SQL)_project\retail-orders\orders.
 ↪csv',encoding='unicode_escape')
```

[2]:
```
df
```

[2]:

	Order Id	Order Date	Ship Mode	Segment	Country	\
0	1	2023-03-01	Second Class	Consumer	United States	
1	2	2023-08-15	Second Class	Consumer	United States	
2	3	2023-01-10	Second Class	Corporate	United States	

```
3            4  2022-06-18  Standard Class   Consumer  United States
4            5  2022-07-13  Standard Class   Consumer  United States
…          …          …              …          …            …
9989      9990  2023-02-18    Second Class   Consumer  United States
9990      9991  2023-03-17  Standard Class   Consumer  United States
9991      9992  2022-08-07  Standard Class   Consumer  United States
9992      9993  2022-11-19  Standard Class   Consumer  United States
9993      9994  2022-07-17    Second Class   Consumer  United States

                City       State  Postal Code Region         Category  \
0          Henderson    Kentucky        42420  South        Furniture
1          Henderson    Kentucky        42420  South        Furniture
2        Los Angeles  California        90036   West  Office Supplies
3     Fort Lauderdale     Florida        33311  South        Furniture
4     Fort Lauderdale     Florida        33311  South  Office Supplies
…               …          …          …    …               …
9989           Miami     Florida        33180  South        Furniture
9990      Costa Mesa  California        92627   West        Furniture
9991      Costa Mesa  California        92627   West       Technology
9992      Costa Mesa  California        92627   West  Office Supplies
9993     Westminster  California        92683   West  Office Supplies

     Sub Category       Product Id  cost price  List Price  Quantity  \
0       Bookcases  FUR-BO-10001798         240         260         2
1          Chairs  FUR-CH-10000454         600         730         3
2          Labels  OFF-LA-10000240          10          10         2
3          Tables  FUR-TA-10000577         780         960         5
4         Storage  OFF-ST-10000760          20          20         2
…           …              …          …          …        …
9989  Furnishings  FUR-FU-10001889          30          30         3
9990  Furnishings  FUR-FU-10000747          70          90         2
9991       Phones  TEC-PH-10003645         220         260         2
9992        Paper  OFF-PA-10004041          30          30         4
9993   Appliances  OFF-AP-10002684         210         240         2

      Discount Percent
0                    2
1                    3
2                    5
3                    2
4                    5
…                  …
9989                 4
9990                 4
9991                 2
9992                 3
9993                 3
```

```
[9994 rows x 16 columns]
```

[3]:
```
df.columns
```

[3]:
```
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')
```

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

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

[6]:
```
#replacing 'Not Available', 'unknown' with nan
df=pd.read_csv(r'C:\Users\Rizwan\Downloads\Shah Rizwan\Python data analyst␣
 ↪project\Retail_orders_ETL(Python+SQL)_project\retail-orders\orders.
 ↪csv',encoding='unicode_escape',na_values=['Not Available', 'unknown'])
```

[7]:
```
df.columns=df.columns.str.lower().str.replace(' ','_')
df.columns
```

[7]:
```
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')
```

[8]:
```
df.count()
```

[8]:
```
order_id            9994
order_date          9994
ship_mode           9988
segment             9994
country             9994
city                9994
state               9994
postal_code         9994
region              9994
category            9994
sub_category        9994
product_id          9994
cost_price          9994
list_price          9994
quantity            9994
discount_percent    9994
```

```
dtype: int64
```

[9]:
```
#there are some null values in ship_mode column replacing them with mode
```

[10]:
```
mode=df['ship_mode'].mode()[0]
df['ship_mode']=df['ship_mode'].fillna(mode)
```

[11]:
```
df.count()
```

[11]:
```
order_id            9994
order_date          9994
ship_mode           9994
segment             9994
country             9994
city                9994
state               9994
postal_code         9994
region              9994
category            9994
sub_category        9994
product_id          9994
cost_price          9994
list_price          9994
quantity            9994
discount_percent    9994
dtype: int64
```

[12]:
```
df['ship_mode'].unique()
```

[12]:
```
array(['Second Class', 'Standard Class', 'First Class', 'Same Day'],
      dtype=object)
```

[15]:
```
#df['discount']=df['list_price']*df['discount_percent']*0.01
#df['sale_price']=df['list_price']-df['discount']
#df['profit']=df['cost_price']-df['sale_price']
df.head()
```

[15]:

	order_id	order_date	ship_mode	segment	country	city	state	postal_code	region	category
0	1	2023-03-01	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture
1	2	2023-08-15	Second Class	Consumer	United States	Henderson	Kentucky	42420	South	Furniture
2	3	2023-01-10	Second Class	Corporate	United States	Los Angeles	California	90036	West	Office Supplies

```
3  Fort Lauderdale     Florida        33311  South        Furniture
4  Fort Lauderdale     Florida        33311  South  Office Supplies

  sub_category       product_id  cost_price  list_price  quantity  \
0    Bookcases  FUR-BO-10001798         240         260         2
1       Chairs  FUR-CH-10000454         600         730         3
2       Labels  OFF-LA-10000240          10          10         2
3       Tables  FUR-TA-10000577         780         960         5
4      Storage  OFF-ST-10000760          20          20         2

   discount_percent  discount  sale_price  profit
0                 2       5.2       254.8   -14.8
1                 3      21.9       708.1  -108.1
2                 5       0.5         9.5     0.5
3                 2      19.2       940.8  -160.8
4                 5       1.0        19.0     1.0
```

[16]:
```
df.dtypes
```

[16]:
```
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
```

[17]:
```
df['order_date']=pd.to_datetime(df['order_date'],format="%Y-%m-%d")
#df['order_date']=pd.to_datetime(df['order_date'])

#df.dtypes
df.head()
```

[17]:

	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

[18]:
```
!pip install psycopg2-binary
```

```
Collecting psycopg2-binary
  Downloading psycopg2_binary-2.9.9-cp312-cp312-win_amd64.whl.metadata (4.6 kB)
Downloading psycopg2_binary-2.9.9-cp312-cp312-win_amd64.whl (1.2 MB)
   ---------------------------------------- 0.0/1.2 MB ? eta -:--:--
   ---------------------------------------- 0.0/1.2 MB ? eta -:--:--
   -- ------------------------------------- 0.1/1.2 MB 825.8 kB/s eta 0:00:02
   ------- -------------------------------- 0.2/1.2 MB 1.8 MB/s eta 0:00:01
   -------------- ------------------------- 0.4/1.2 MB 2.8 MB/s eta 0:00:01
   -------------------- ------------------- 0.6/1.2 MB 3.3 MB/s eta 0:00:01
   -------------------- ------------------- 0.6/1.2 MB 2.5 MB/s eta 0:00:01
   ------------------------- ------------ 0.8/1.2 MB 2.8 MB/s eta 0:00:01
   ------------------------- ------------ 0.8/1.2 MB 2.8 MB/s eta 0:00:01
   ---------------------------------- --- 1.1/1.2 MB 2.8 MB/s eta 0:00:01
   ---------------------------------------  1.2/1.2 MB 2.8 MB/s eta 0:00:01
   ---------------------------------------- 1.2/1.2 MB 2.5 MB/s eta 0:00:00
Installing collected packages: psycopg2-binary
Successfully installed psycopg2-binary-2.9.9
```

```
[notice] A new release of pip is available: 24.0 -> 24.2
[notice] To update, run: python.exe -m pip install --upgrade pip
```

[19]:
```
from sqlalchemy import create_engine
conn=create_engine(r'mysql+mysqlconnector://root:RizSQL819@localhost/
 ↪retail_orders')
df.to_sql('df_orders', con=conn, index=False, if_exists='append')
```

[19]: 9994

[19]:
```
from sqlalchemy import create_engine

# PostgreSQL connection string
conn = create_engine('postgresql+psycopg2://postgres:Rizpostsql@localhost:5432/
 ↪retail_orders')

# Insert data into the PostgreSQL database
df.to_sql('df_orders', con=conn, index=False, if_exists='append')
```

[19]: 994

[ ]:

```sql
#create database retail_orders;
#select * from retail_orders.df_orders;
truncate retail_orders.df_orders;
create table  retail_orders.df_orders(
order_id int primary key,
order_date datetime,
ship_mode varchar(20),
segment varchar(20),
country varchar(20),
city    varchar(20),
state   varchar(20),
postal_code varchar(20),
region  varchar(20),
category varchar(20),
sub_category  varchar(20),
product_id    varchar(50),
cost_price    int,
list_price  int,
quantity    int,
discount_percent     decimal(4,2),
discount             decimal(9,2),
sale_price           decimal(9,2),
profit               decimal(9,2) );
```

```sql
--#Q1-find top 10 highest revenue generating products
#select * from df_orders;
select product_id,category,sub_category,sum(sale_price) as total_sales
from df_orders
group by product_id,category,sub_category
order by total_sales desc limit 10;

--#Q2-find top 5 highest selling product in each region
#select * from df_orders;
#select distinct(region) from df_orders;

with cte as (
Select region,product_id,category,sub_category,sum(sale_price) as total_sales
from df_orders
group by region,product_id
)
select * from (
select *,
row_number() over (partition by region order by total_sales desc ) as rn
from cte ) as a
where rn<=5;
```

```sql
--#Q3.find month over month sales comparison for 2022 vs 2023 eg.jan 2022 vs jan 2023
with cte as (
select month(order_date) as order_month ,year(order_date) as order_year ,sum(sale_price) as total_sales
from df_orders
group by order_month, order_year
)
select order_month,
sum(case when order_year=2022 then total_sales else 0 end) as 2022_sales,
sum(case when order_year=2023 then total_sales else 0 end) as 2023_sales
from cte
group by order_month
order by order_month  ;
```

```sql
--#Q4.for each category which month had highest sales
#select * from df_orders;

with cte as(
select category,date_format(order_date,"%Y%M") as order_year_month,sum(sale_price) as total_sales
from df_orders
group by category, order_year_month
order by total_sales desc)
select * from (
select *,
row_number() over (partition by category order by total_sales desc ) as rn
from cte ) as a
where rn=1;
```

```sql
--#Q5.which subcategory has highes growth by profit   in 2023 compare to 2022

select  * from df_orders;
with cte as (
select sub_category,sum(profit) as total_profit,year(order_date) as order_year
from df_orders
group by sub_category,order_year
)
,cte2 as (
select sub_category,
sum(case when order_year=2023 then total_profit else 0 end) as 2022_profit,
sum(case when order_year=2022 then total_profit else 0 end) as 2023_profit
from cte
group by sub_category )
select *,
(2023_profit-2022_profit)*100/2022_profit as profit_growth_2022_to_2023
from cte2
order by profit_growth_2022_to_2023 desc
limit 1 ;
```

SALES INSIGHT

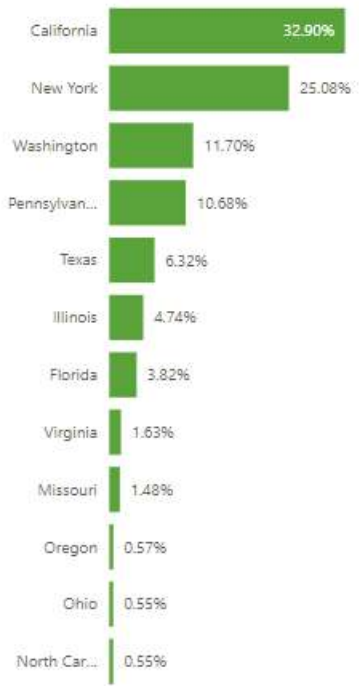
Revenue	Profit Margin	Sales Qauntity
2,216K	-205K	38K

Year
- 2022
- 2023

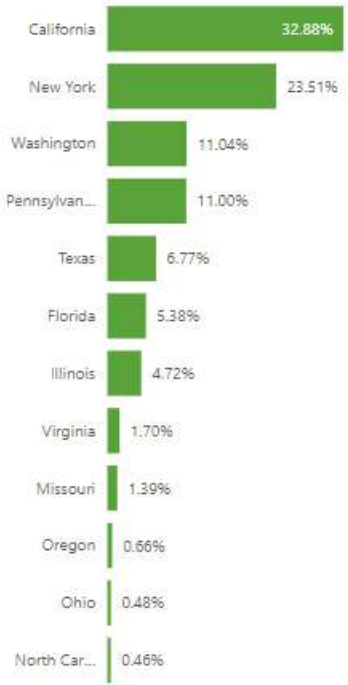
ship mode
First Class | Same Day | Second Class | Standard Class

segment
- Consumer
- Corporate
- Home Office

% Revenue by state



% Profit by state



% Profit margin by sub-category

% Profit by Region