untitled-1

October 27, 2024

```
[ ]: #Download Dataset using kaggle
```

```
[5]: import kaggle
     !kaggle datasets download ankitbansal06/retail-orders -f orders.csv
```

```
Dataset URL: https://www.kaggle.com/datasets/ankitbansal06/retail-orders
License(s): CC0-1.0
orders.csv.zip: Skipping, found more recently modified local copy (use --force
to force download)
```

```
[7]: #Extract file from zip file in the given directory
```

```
[9]: import zipfile
     zip_ref = zipfile.ZipFile('orders.csv.zip')
     zip_ref.extractall(r'C:\Users\User\Desktop\Data Analyst Projects\Order Data␣
       ↪Analysis') # extract file to dir
     zip_ref.close() # close file
```

```
[11]: #Read the csv file and handle null values
```

```
[13]: import pandas as pd
      df=pd.read_csv(r'C:\Users\User\Desktop\Data Analyst Projects\Order Data␣
        ↪Analysis\orders.csv',na_values=['Not Available','unknown'])
      df['Ship Mode'].unique()
```

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

```
[15]: #Rename columns names ..make them lower case and replace space with underscore
```

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

```
[19]:       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]
```

[21]:
```
#derive new columns discount , sale price and profit
```

[26]:
```
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']
```

[28]:
```
df.dtypes
```

[28]:
```
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
```

[30]:
```
#Convert order_date from object data type to date time
df['order_date']=pd.to_datetime(df['order_date'],format="%Y-%m-%d")
```

[32]:
```
#drop cost price list price and discount percent columns
```

[34]:
```
df.drop(columns=['list_price','cost_price','discount_percent'],inplace=True)
```

[76]:
```
#load the data into sql server using replace option
import sqlalchemy as sal

engine = sal.create_engine(r'mssql://DESKTOP-CJTCHDL\SQLEXPRESS/DataAnalystDB?
 ↪driver=ODBC+DRIVER+17+FOR+SQL+SERVER')
conn=engine.connect()
```

[80]:

```
#load the data into sql server using append option
df.to_sql('df_orders', con=conn, index=False, if_exists='append',␣
 ↪chunksize=1000)
```

[80]:

```
824
```

[42]:

```
print(df.shape)
```

```
(9994, 16)
```