ETL Retail

July 9, 2025

[54]:
```
import numpy as np
import pandas as pd
```

[55]:
```
df = pd.read_csv(r"C:\Users\Ryota Kohama\Documents\Sales Performance Analysis␣
↪Project\DataSet\OnlineRetail.csv", encoding='ISO-8859-1')
df.head()
```

[55]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

[56]:
```
df.columns
```

[56]:
```
Index(['InvoiceNo', 'StockCode', 'Description', 'Quantity', 'InvoiceDate',
       'UnitPrice', 'CustomerID', 'Country'],
      dtype='object')
```

[57]:
```
df.describe()
```

[57]:

	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

[58]:
```
df.shape
```

[58]:
```
(541909, 8)
```

[59]:
```
df.isnull().sum()
```

[59]:
```
InvoiceNo          0
StockCode          0
Description     1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID     135080
Country            0
dtype: int64
```

[60]:
```
df.dtypes
```

[60]:
```
InvoiceNo       object
StockCode       object
Description     object
Quantity         int64
InvoiceDate     object
UnitPrice      float64
CustomerID     float64
Country         object
dtype: object
```

[61]:
```
#Data Cleaning
cdf = df.copy()
cdf['Description'] = cdf['Description'].fillna('No Description')
```

[62]:
```
cdf = cdf.dropna(subset=['CustomerID'])
```

[63]:
```
cdf['InvoiceDate'] = pd.to_datetime(cdf['InvoiceDate'])
```

[64]:
```
cdf['CustomerID'] = cdf['CustomerID'].astype('Int64')
```

[65]:
```
cdf = cdf.drop_duplicates()
```

[66]:
```
cdf = cdf[(cdf['Quantity'] > 0) & (cdf['UnitPrice'] > 0)]
```

[67]:
```
cdf['Description'] = cdf['Description'].str.strip()
cdf['Country'] = cdf['Country'].str.strip()
```

[68]:
```
cdf.head(10)
```

[68]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.65	17850	United Kingdom
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.25	17850	United Kingdom
7	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.85	17850	United Kingdom
8	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	1.85	17850	United Kingdom
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	1.69	13047	United Kingdom

[69]:
```
cdf.isnull().sum()
```

[69]:
```
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
```

[70]:
```
cdf.dtypes
```

[70]:
```
InvoiceNo              object
StockCode              object
Description            object
Quantity                int64
InvoiceDate    datetime64[ns]
UnitPrice             float64
CustomerID              Int64
Country                object
dtype: object
```

[71]:
```
# Product Table
product_df = df[['StockCode', 'Description']].drop_duplicates()
product_df.rename(columns={'StockCode': 'ProductID', 'Description':␣
↪'ProductName'}, inplace=True)
#product_df.to_csv('Product.csv', index=False)
```

[72]:
```
#customer_df.to_csv('Customer.csv', index=False)
```

[73]:
```
# GeographyKey Table
geography_df = df[['Country']].drop_duplicates().reset_index(drop=True)
geography_df['GeographyKey'] = geography_df.index + 1
geography_df = geography_df[['GeographyKey', 'Country']]
#geography_df.to_csv('Geography.csv', index=False)
```

[ ]:
```

```

[74]:
```
unique_dates = cdf['InvoiceDate'].dt.date.unique()
date_df = pd.DataFrame(unique_dates, columns=['Date'])
date_df['DateKey'] = date_df['Date'].apply(lambda x: int(x.strftime('%Y%m%d')))
date_df['Day'] = date_df['Date'].apply(lambda x: x.day)
date_df['Month'] = date_df['Date'].apply(lambda x: x.month)
date_df['MonthName'] = date_df['Date'].apply(lambda x: x.strftime('%B'))
date_df['Quarter'] = date_df['Date'].apply(lambda x: (x.month-1)//3 + 1)
date_df['Year'] = date_df['Date'].apply(lambda x: x.year)
date_df['WeekOfYear'] = date_df['Date'].apply(lambda x: x.isocalendar()[1])
date_df = date_df[['DateKey', 'Date', 'Day', 'Month', 'MonthName', 'Quarter',␣
↪'Year', 'WeekOfYear']]
#date_df.to_csv('Date.csv', index=False)
```

[75]:
```
cdf['DateKey'] = cdf['InvoiceDate'].dt.strftime('%Y%m%d').astype(int)
sales_df = cdf[['InvoiceNo', 'InvoiceDate', 'StockCode', 'CustomerID',␣
↪'Quantity', 'UnitPrice','DateKey']]
sales_df['TotalAmount'] = sales_df['Quantity'] * sales_df['UnitPrice']
sales_df = sales_df.rename(columns={
    'InvoiceNo': 'SalesID',
    'StockCode': 'ProductID'
})
sales_df.to_csv('SalesDetails.csv', index=False)
```

```
C:\Users\Ryota Kohama\AppData\Local\Temp\ipykernel_11316\649051004.py:3:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  sales_df['TotalAmount'] = sales_df['Quantity'] * sales_df['UnitPrice']
```

[76]:
```
product_df.head()
```

[76]:
```
  ProductID                         ProductName
0    85123A   WHITE HANGING HEART T-LIGHT HOLDER
1     71053                  WHITE METAL LANTERN
2    84406B       CREAM CUPID HEARTS COAT HANGER
3    84029G  KNITTED UNION FLAG HOT WATER BOTTLE
4    84029E       RED WOOLLY HOTTIE WHITE HEART.
```

[77]:
```
date_df.head()
```

[77]:
```
    DateKey        Date  Day  Month MonthName  Quarter  Year  WeekOfYear
0  20101201  2010-12-01    1     12  December        4  2010          48
1  20101202  2010-12-02    2     12  December        4  2010          48
2  20101203  2010-12-03    3     12  December        4  2010          48
3  20101205  2010-12-05    5     12  December        4  2010          48
4  20101206  2010-12-06    6     12  December        4  2010          49
```

[78]:
```
sales_df.head()
```

[78]:
```
  SalesID         InvoiceDate ProductID  CustomerID  Quantity  UnitPrice  \
0  536365 2010-12-01 08:26:00    85123A       17850         6       2.55
1  536365 2010-12-01 08:26:00     71053       17850         6       3.39
2  536365 2010-12-01 08:26:00    84406B       17850         8       2.75
3  536365 2010-12-01 08:26:00    84029G       17850         6       3.39
4  536365 2010-12-01 08:26:00    84029E       17850         6       3.39

    DateKey  TotalAmount
0  20101201        15.30
1  20101201        20.34
2  20101201        22.00
3  20101201        20.34
4  20101201        20.34
```

[ ]:
```

```

[79]:
```
geography_df.head()
```

[79]:
```
   GeographyKey         Country
0             1  United Kingdom
1             2          France
2             3       Australia
3             4     Netherlands
4             5         Germany
```

[ ]:
```

```