Darshan UNIVERSITY
योग: कर्मसु कौशलम्

Data Mining

🧺 Online Retail Dataset (Apriori Preprocessing) - TASK 2

137 | Vishal Baraiya | 23010101014

Dataset: https://www.kaggle.com/datasets/vijayuv/onlineretail

Step 1: Load Data

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In [2]:
```
OnlineRetail = pd.read_csv("D:\\VS_CODES\\DataMining\\ProjectDataMining\\Dataset\\O
OnlineRetail
```

Out[2]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0
...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680.0
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680.0
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680.0
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680.0
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680.0

541909 rows × 8 columns

Step 2: Drop Missing Invoices or Descriptions

In [3]:
```
OnlineRetail.dropna(subset=['InvoiceNo', 'Description'], inplace=True)
OnlineRetail
```

Out[3]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0
...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680.0
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680.0
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680.0
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680.0
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680.0

540455 rows × 8 columns

Step 3: Remove Canceled Invoices Start with letter 'c'

In [4]:
```
OnlineRetail = OnlineRetail[~OnlineRetail['InvoiceNo'].astype(str).str.startswith('
OnlineRetail
```

Out[4]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0
...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680.0
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680.0
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680.0
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680.0
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680.0

531167 rows × 8 columns

Step 4: Keep Only Positive Quantities

In [5]:
```
OnlineRetail = OnlineRetail[OnlineRetail['Quantity'] > 0]
OnlineRetail
```

Out[5]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0
...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680.0
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680.0
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680.0
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680.0
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680.0

530693 rows × 8 columns

Step 5: Standardize Item Descriptions | Lower case and remove space

In [6]:
```
# r'\s+' => Space ( ) , Tab (\t) , Newline (\n)
OnlineRetail['Description'] = OnlineRetail['Description'].str.lower().str.strip().s
OnlineRetail
```

```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_15132\492253459.py:2: SettingWithCopyWarn
ing:
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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  OnlineRetail['Description'] = OnlineRetail['Description'].str.lower().str.strip().
str.replace(r'\s+', ' ', regex=True)
```

Out[6]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID
0	536365	85123A	white hanging heart t-light holder	6	12/1/2010 8:26	2.55	17850.0
1	536365	71053	white metal lantern	6	12/1/2010 8:26	3.39	17850.0
2	536365	84406B	cream cupid hearts coat hanger	8	12/1/2010 8:26	2.75	17850.0
3	536365	84029G	knitted union flag hot water bottle	6	12/1/2010 8:26	3.39	17850.0
4	536365	84029E	red woolly hottie white heart.	6	12/1/2010 8:26	3.39	17850.0
...	...	...	...	...	...	...	...
541904	581587	22613	pack of 20 spaceboy napkins	12	12/9/2011 12:50	0.85	12680.0
541905	581587	22899	children's apron dolly girl	6	12/9/2011 12:50	2.10	12680.0
541906	581587	23254	childrens cutlery dolly girl	4	12/9/2011 12:50	4.15	12680.0
541907	581587	23255	childrens cutlery circus parade	4	12/9/2011 12:50	4.15	12680.0
541908	581587	22138	baking set 9 piece retrospot	3	12/9/2011 12:50	4.95	12680.0

530693 rows × 8 columns

Step 6: Filter for a Single Country (e.g., United Kingdom)

In [7]:
```
OnlineRetail[OnlineRetail["Country"] == "United Kingdom"]
```

Out[7]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID
0	536365	85123A	white hanging heart t-light holder	6	12/1/2010 8:26	2.55	17850.0
1	536365	71053	white metal lantern	6	12/1/2010 8:26	3.39	17850.0
2	536365	84406B	cream cupid hearts coat hanger	8	12/1/2010 8:26	2.75	17850.0
3	536365	84029G	knitted union flag hot water bottle	6	12/1/2010 8:26	3.39	17850.0
4	536365	84029E	red woolly hottie white heart.	6	12/1/2010 8:26	3.39	17850.0
...	...	...	...	...	...	...	...
541889	581585	22466	fairy tale cottage night light	12	12/9/2011 12:31	1.95	15804.0
541890	581586	22061	large cake stand hanging strawbery	8	12/9/2011 12:49	2.95	13113.0
541891	581586	23275	set of 3 hanging owls ollie beak	24	12/9/2011 12:49	1.25	13113.0
541892	581586	21217	red retrospot round cake tins	24	12/9/2011 12:49	8.95	13113.0
541893	581586	20685	doormat red retrospot	10	12/9/2011 12:49	7.08	13113.0

485694 rows × 8 columns

Step 7: Remove Duplicates

In [8]:
```
OnlineRetail = OnlineRetail.drop_duplicates()
OnlineRetail
```

Out[8]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID
0	536365	85123A	white hanging heart t-light holder	6	12/1/2010 8:26	2.55	17850.0
1	536365	71053	white metal lantern	6	12/1/2010 8:26	3.39	17850.0
2	536365	84406B	cream cupid hearts coat hanger	8	12/1/2010 8:26	2.75	17850.0
3	536365	84029G	knitted union flag hot water bottle	6	12/1/2010 8:26	3.39	17850.0
4	536365	84029E	red woolly hottie white heart.	6	12/1/2010 8:26	3.39	17850.0
...	...	...	...	...	...	...	...
541904	581587	22613	pack of 20 spaceboy napkins	12	12/9/2011 12:50	0.85	12680.0
541905	581587	22899	children's apron dolly girl	6	12/9/2011 12:50	2.10	12680.0
541906	581587	23254	childrens cutlery dolly girl	4	12/9/2011 12:50	4.15	12680.0
541907	581587	23255	childrens cutlery circus parade	4	12/9/2011 12:50	4.15	12680.0
541908	581587	22138	baking set 9 piece retrospot	3	12/9/2011 12:50	4.95	12680.0

525462 rows × 8 columns

Step 8: Add TotalPrice column , TotalPrice = Quantity * UnitPrice

In [9]:
```
OnlineRetail['TotalPrice'] = OnlineRetail['Quantity'] * OnlineRetail['UnitPrice']
OnlineRetail
```

```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_15132\2077981908.py:1: SettingWithCopyWar
ning:
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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  OnlineRetail['TotalPrice'] = OnlineRetail['Quantity'] * OnlineRetail['UnitPrice']
```

Out[9]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID
0	536365	85123A	white hanging heart t-light holder	6	12/1/2010 8:26	2.55	17850.0
1	536365	71053	white metal lantern	6	12/1/2010 8:26	3.39	17850.0
2	536365	84406B	cream cupid hearts coat hanger	8	12/1/2010 8:26	2.75	17850.0
3	536365	84029G	knitted union flag hot water bottle	6	12/1/2010 8:26	3.39	17850.0
4	536365	84029E	red woolly hottie white heart.	6	12/1/2010 8:26	3.39	17850.0
...	...	...	...	...	...	...	...
541904	581587	22613	pack of 20 spaceboy napkins	12	12/9/2011 12:50	0.85	12680.0
541905	581587	22899	children's apron dolly girl	6	12/9/2011 12:50	2.10	12680.0
541906	581587	23254	childrens cutlery dolly girl	4	12/9/2011 12:50	4.15	12680.0
541907	581587	23255	childrens cutlery circus parade	4	12/9/2011 12:50	4.15	12680.0
541908	581587	22138	baking set 9 piece retrospot	3	12/9/2011 12:50	4.95	12680.0

525462 rows × 9 columns

Step 9 : Check for Top 10 Selling Products

In [10]:
```
top_10_products = OnlineRetail.groupby('Description')["Quantity"].sum().sort_values
print(top_10_products)
```

```
Description
paper craft , little birdie           80995
medium ceramic top storage jar        78033
world war 2 gliders asstd designs     54951
jumbo bag red retrospot               48375
white hanging heart t-light holder    37876
popcorn holder                        36749
pack of 72 retrospot cake cases       36396
assorted colour bird ornament         36362
rabbit night light                    30739
mini paint set vintage                26633
Name: Quantity, dtype: int64
```

Step 10 : Identify Most Active Customers

In [11]:
```
# most_active_customers = (
#     OnlineRetail.groupby('CustomerID')['InvoiceNo']
#     .nunique()
#     .sort_values(ascending=False)
#     .head(10)
# )
# print(most_active_customers)

most_active_customers = OnlineRetail.groupby("CustomerID")["InvoiceNo"].nunique().s
print(most_active_customers)
```

```
CustomerID
12748.0    210
14911.0    201
17841.0    124
13089.0     97
14606.0     93
15311.0     91
12971.0     86
14646.0     74
16029.0     63
13408.0     62
Name: InvoiceNo, dtype: int64
```

Step 11: Time-Based Analysis | Convert InvoiceDate to datetime and check popular months/hours.

In [12]:
```
OnlineRetail['InvoiceDate'] = pd.to_datetime(OnlineRetail['InvoiceDate'])
OnlineRetail
```

```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_15132\3988525607.py:1: SettingWithCopyWar
ning:
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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  OnlineRetail['InvoiceDate'] = pd.to_datetime(OnlineRetail['InvoiceDate'])
```

Out[12]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID
0	536365	85123A	white hanging heart t-light holder	6	2010-12-01 08:26:00	2.55	17850.0
1	536365	71053	white metal lantern	6	2010-12-01 08:26:00	3.39	17850.0
2	536365	84406B	cream cupid hearts coat hanger	8	2010-12-01 08:26:00	2.75	17850.0
3	536365	84029G	knitted union flag hot water bottle	6	2010-12-01 08:26:00	3.39	17850.0
4	536365	84029E	red woolly hottie white heart.	6	2010-12-01 08:26:00	3.39	17850.0
...	...	...	...	...	...	...	...
541904	581587	22613	pack of 20 spaceboy napkins	12	2011-12-09 12:50:00	0.85	12680.0
541905	581587	22899	children's apron dolly girl	6	2011-12-09 12:50:00	2.10	12680.0
541906	581587	23254	childrens cutlery dolly girl	4	2011-12-09 12:50:00	4.15	12680.0
541907	581587	23255	childrens cutlery circus parade	4	2011-12-09 12:50:00	4.15	12680.0
541908	581587	22138	baking set 9 piece retrospot	3	2011-12-09 12:50:00	4.95	12680.0

525462 rows × 9 columns

In [13]:
```
# Extract Month and Hour
OnlineRetail['InvoiceMonth'] = OnlineRetail['InvoiceDate'].dt.month
OnlineRetail['InvoiceHour'] = OnlineRetail['InvoiceDate'].dt.hour
OnlineRetail
```

```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_15132\3889572408.py:2: SettingWithCopyWar
ning:
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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  OnlineRetail['InvoiceMonth'] = OnlineRetail['InvoiceDate'].dt.month
C:\Users\ASUS\AppData\Local\Temp\ipykernel_15132\3889572408.py:3: SettingWithCopyWar
ning:
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/u
ser_guide/indexing.html#returning-a-view-versus-a-copy
  OnlineRetail['InvoiceHour'] = OnlineRetail['InvoiceDate'].dt.hour
```

Out[13]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID
0	536365	85123A	white hanging heart t-light holder	6	2010-12-01 08:26:00	2.55	17850.0
1	536365	71053	white metal lantern	6	2010-12-01 08:26:00	3.39	17850.0
2	536365	84406B	cream cupid hearts coat hanger	8	2010-12-01 08:26:00	2.75	17850.0
3	536365	84029G	knitted union flag hot water bottle	6	2010-12-01 08:26:00	3.39	17850.0
4	536365	84029E	red woolly hottie white heart.	6	2010-12-01 08:26:00	3.39	17850.0
...	...	...	...	...	...	...	...
541904	581587	22613	pack of 20 spaceboy napkins	12	2011-12-09 12:50:00	0.85	12680.0
541905	581587	22899	children's apron dolly girl	6	2011-12-09 12:50:00	2.10	12680.0
541906	581587	23254	childrens cutlery dolly girl	4	2011-12-09 12:50:00	4.15	12680.0
541907	581587	23255	childrens cutlery circus parade	4	2011-12-09 12:50:00	4.15	12680.0
541908	581587	22138	baking set 9 piece retrospot	3	2011-12-09 12:50:00	4.95	12680.0

525462 rows × 11 columns

In [14]:
```
# Popular Months
popular_months = (
    OnlineRetail.groupby('InvoiceMonth')['Quantity']
    .sum()
    .sort_values(ascending=False)
```

```
)
print(popular_months)
```

```
InvoiceMonth
11    766844
12    673511
10    624545
9     571399
8     422330
7     403939
5     397099
6     392822
1     387366
3     380158
4     309418
2     283041
Name: Quantity, dtype: int64
```

In [15]:
```
# Popular Purchase Hours
popular_hours = (
    OnlineRetail.groupby('InvoiceHour')['Quantity']
    .sum()
    .sort_values(ascending=False)
)
print(popular_hours)
```

```
InvoiceHour
12    839312
10    813971
13    706660
11    673214
15    645185
14    602616
9     519985
16    334827
17    193072
8     158685
18     67137
19     32929
7      15370
20      9508
6          1
Name: Quantity, dtype: int64
```

Step 12: Convert to Basket Format

Example

Transaction	Milk	Bread	Butter
T1	1	1	1
T2	0	1	1
T3	1	0	0

Transaction	Milk	Bread	Butter
T4	1	1	0
T5	0	1	1

In [16]:
```
# Pick a small sample of 5 invoices
sample_invoices = OnlineRetail['InvoiceNo'].unique()[:5]  # first 5 unique invoices

# Filter the dataset to include only those invoices
sample_data = OnlineRetail[OnlineRetail['InvoiceNo'].isin(sample_invoices)]

# Preview sample
sample_data[['InvoiceNo', 'Description', 'Quantity']].head()
```

Out[16]:

	InvoiceNo	Description	Quantity
0	536365	white hanging heart t-light holder	6
1	536365	white metal lantern	6
2	536365	cream cupid hearts coat hanger	8
3	536365	knitted union flag hot water bottle	6
4	536365	red woolly hottie white heart.	6

In [17]:
```
# Create basket format from sample
mini_basket = (
    sample_data
    .groupby(['InvoiceNo', 'Description'])['Quantity']
    .sum()
    .unstack()
    .fillna(0)
    .applymap(lambda x: 1 if x > 0 else 0)
)

# Optional: Rename index to "Transaction" like the image
mini_basket.index.name = 'Transaction'

# Display result
mini_basket.head()
```

```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_15132\2170959364.py:8: FutureWarning: Dat
aFrame.applymap has been deprecated. Use DataFrame.map instead.
  .applymap(lambda x: 1 if x > 0 else 0)
```

Out[17]:

Description	assorted colour bird ornament	bath building block word	blue coat rack paris fashion	box of 6 assorted colour teaspoons	box of vintage alphabet blocks	box of vintage jigsaw blocks	cream cupid hearts coat hanger	doormat new england
Transaction								
536365	0	0	0	0	0	0	1	0
536366	0	0	0	0	0	0	0	0
536367	1	0	0	1	1	1	0	1
536368	0	0	1	0	0	0	0	0
536369	0	1	0	0	0	0	0	0

5 rows × 26 columns

Step:13 : Save Basket Format into CSV file

In [18]:
```
mini_basket.to_csv("mini_basket.csv")
```

In [ ]:
```

```