Nama : Elviyani Mawarni

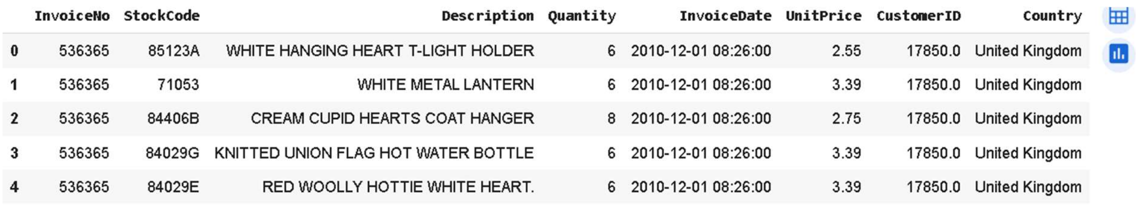
NIM : G.231.22.0077

Mata Kuliah : Data Mining

KODE DAN PENJELASAN PRAKTIKUM 4 MARKET BASKET ANALYSIS

1. Import library dan masukkan data yang akan digunakan serta menampilkannya

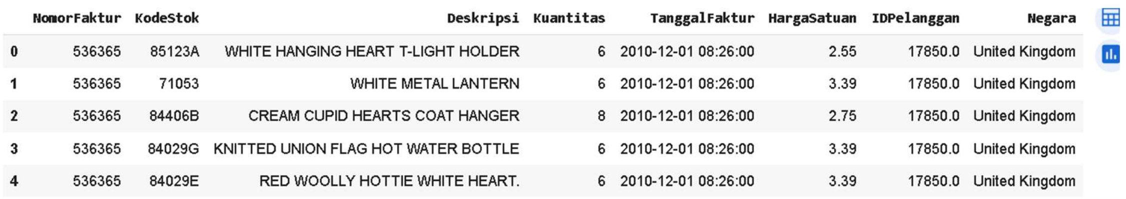
```
import pandas as pd
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules

df = pd.read_excel('http://archive.ics.uci.edu/ml/machine-learning-databases/00352/Online%20Retail.xlsx')
df.head()
```



	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	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

2. Mengganti nama dari data yang di upload.

```
df.rename(columns={
    'InvoiceNo': 'NomorFaktur',
    'StockCode': 'KodeStok',
    'Description': 'Deskripsi',
    'Quantity': 'Kuantitas',
    'InvoiceDate': 'TanggalFaktur',
    'UnitPrice': 'HargaSatuan',
    'CustomerID': 'IDPelanggan',
    'Country': 'Negara'
}, inplace=True)

df.head()
```



	NomorFaktur	KodeStok	Deskripsi	Kuantitas	TanggalFaktur	HargaSatuan	IDPelanggan	Negara
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

3. Menghilangkan spasi pada ‘Deskripsi’, menghapus baris yang kosong pada ‘NomorFaktur’, mengkonversi kolom ‘NomorFaktur’ menjadi tipe data string dan memfilter baris yang mengandung karakter ‘C’ pada NomorFaktur.

```
df['Deskripsi'] = df['Deskripsi'].str.strip()
df.dropna(axis=0, subset=['NomorFaktur'], inplace=True)
df['NomorFaktur'] = df['NomorFaktur'].astype('str')
df = df[~df['NomorFaktur'].str.contains('C')]
```

4. membuat basket

```
df['Deskripsi'] = df['Deskripsi'].str.strip()
df.dropna(axis=0, subset=['NomorFaktur'], inplace=True)
df['NomorFaktur'] = df['NomorFaktur'].astype('str')
df = df[~df['NomorFaktur'].str.contains('C')]
```

5. membuat fungsi encode dan mengubah menjadi biner

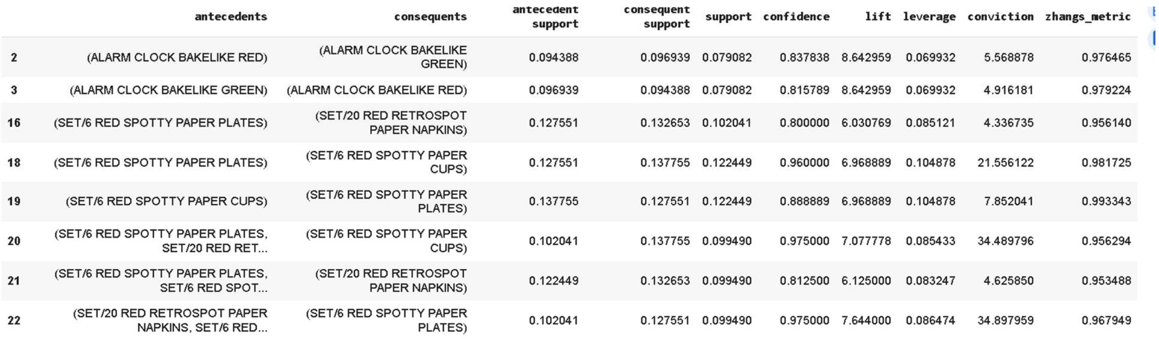
```
def encode_units(x):

    if x <= 0:
        return 0
    if x >= 1:
        return 1

basket_sets = basket.applymap(encode_units)
basket_sets.drop('POSTAGE', inplace=True, axis=1)
```

6. Menggunakan apriori untuk menemukan itemset yang sering muncul

```
frequent_itemsets = apriori(basket_sets, min_support=0.07,
use_colnames=True)
```

```
7. rules = association_rules(frequent_itemsets, metric="lift",
min_threshold=1)

rules.head()
```

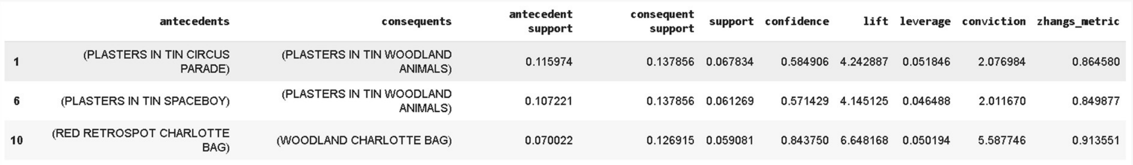
```
8.  rules[ (rules['lift'] >= 6) &

       (rules['confidence'] >= 0.8) ]
```



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
2	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.094388	0.096939	0.079082	0.837838	8.642959	0.069932	5.568878	0.976465
3	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.096939	0.094388	0.079082	0.815789	8.642959	0.069932	4.916181	0.979224
16	(SET/6 RED SPOTTY PAPER PLATES)	(SET/20 RED RETROSPOT PAPER NAPKINS)	0.127551	0.132653	0.102041	0.800000	6.030769	0.085121	4.336735	0.956140
18	(SET/6 RED SPOTTY PAPER PLATES)	(SET/6 RED SPOTTY PAPER CUPS)	0.127551	0.137755	0.122449	0.960000	6.968889	0.104878	21.556122	0.981725
19	(SET/6 RED SPOTTY PAPER CUPS)	(SET/6 RED SPOTTY PAPER PLATES)	0.137755	0.127551	0.122449	0.888889	6.968889	0.104878	7.852041	0.993343
20	(SET/6 RED SPOTTY PAPER PLATES, SET/20 RED RET...	(SET/6 RED SPOTTY PAPER CUPS)	0.102041	0.137755	0.099490	0.975000	7.077778	0.085433	34.489796	0.956294
21	(SET/6 RED SPOTTY PAPER PLATES, SET/6 RED SPOT...	(SET/20 RED RETROSPOT PAPER NAPKINS)	0.122449	0.132653	0.099490	0.812500	6.125000	0.083247	4.625850	0.953488
22	(SET/20 RED RETROSPOT PAPER NAPKINS, SET/6 RED...	(SET/6 RED SPOTTY PAPER PLATES)	0.102041	0.127551	0.099490	0.975000	7.644000	0.086474	34.897959	0.967949

9.
```
basket['ALARM CLOCK BAKELIKE GREEN'].sum()


 340.0

basket['ALARM CLOCK BAKELIKE RED'].sum()

316.0
```

10.
```
basket2 = (df[df['Negara'] =="Germany"]

          .groupby(['NomorFaktur', 'Deskripsi'])['Kuantitas']
          .sum().unstack().reset_index().fillna(0)
          .set_index('NomorFaktur'))

basket_sets2 = basket2.applymap(encode_units)
basket_sets2.drop('POSTAGE', inplace=True, axis=1)
frequent_itemsets2 = apriori(basket_sets2, min_support=0.05,
use_colnames=True)
rules2 = association_rules(frequent_itemsets2, metric="lift",
min_threshold=1)

rules2[ (rules2['lift'] >= 4) &
        (rules2['confidence'] >= 0.5)]
```



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
1	(PLASTERS IN TIN CIRCUS PARADE)	(PLASTERS IN TIN WOODLAND ANIMALS)	0.115974	0.137856	0.067834	0.584906	4.242887	0.051846	2.076984	0.864580
6	(PLASTERS IN TIN SPACEBOY)	(PLASTERS IN TIN WOODLAND ANIMALS)	0.107221	0.137856	0.061269	0.571429	4.145125	0.046488	2.011670	0.849877
10	(RED RETROSPOT CHARLOTTE BAG)	(WOODLAND CHARLOTTE BAG)	0.070022	0.126915	0.059081	0.843750	6.648168	0.050194	5.587746	0.913551