DIAN VITIANA NINGRUM
0621154000020

ETS DATA MINING TAKE HOME

N0. 4. Lakukan nalisis Association Rule menggunakan alogaritma apriori pada data Online Retail. Diperbolehkan menggunakan bantuan software. Data tersedia di link berikut. https://intip.in/etsdatmin2

Dalam data Online Retail tersebut terdapat 8 Variabel, yaitu Invoice No, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer dan Country. Dengan keterangan sebagai berikut:

a. Invoice No merupakan nomor nota pada setiap pembelian,
b. Stock Code merupakan kode stok dari sebuah barang,
c. Decription yaitu deskripsi dari barang,
d. Quantity merupakan jumlah barang yang dibeli pada sebuah transaksi,
e. Invoice Date adalah tanggal pembelian,
f. Unit Price adalah harga per unit barang,
g. Customer ID adalah ID pelanggan,
h. Country adalah negara pelanggan.

Berikut merupakan output dari Jupyter Notebook beserta Syntax nya

```
import pandas as pd

df = pd.read_excel(r'E:/ets2.xlsx', sheet_name='Sheet1')

df.head()
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-01-12 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-01-12 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-01-12 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-01-12 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-01-12 08:26:00	3.39	17850.0	United Kingdom

Gambar di atas merupakan tampilan data Online Retail yang terdiri dari 8 Variabel yaitu Invoice No, Stock Code, Description, Quantity, Invoice Date, Unit Price, Customer dan Country.

```
df.info
```

```
[541909 rows x 8 columns]
```

Data Online Retail terdiri dari 8 kolom dan 541909 baris. Kemudian akan dilihat apakah terdapat missing value atau tidak.

DIAN VITIANA NINGRUM
0621154000020

```
## imputation missing value

import numpy as np

np.sum(df.isnull())
```

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

Berdasarkan output di atas, diketahui bahwa terdapat 1454 missing value pada variabel Description dan sebanyak 135080 pada Variabel Customer ID.

```
from mlxtend.frequent_patterns import apriori

from mlxtend.frequent_patterns import
```

```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
<ipython-input-10-73c97be96c5f> in <module>()
----> 1 from mlxtend.frequent_patterns import apriori
      2 from mlxtend.frequent_patterns import association_rules

ModuleNotFoundError: No module named 'mlxtend'
```

Dilakukan import Package atau import library, namun library mlxtend tidak dapat diinstal/ tidak kompantibel dengan python yang digunakan.

```
# Clean up spaces in description and remove any rows that don't
have a valid invoice

df['Description'] = df['Description'].str.strip()

df.dropna(axis=0, subset=['InvoiceNo'], inplace=True)

df['InvoiceNo'] = df['InvoiceNo'].astype('str')

df = df[~df['InvoiceNo'].str.contains('C')]

basket = (df[df['Country'] =="France"]

        .groupby(['InvoiceNo', 'Description'])['Quantity']

        .sum().unstack().reset_index().fillna(0)

        .set_index('InvoiceNo'))

basket.head()
```

DIAN VITIANA NINGRUM
0621154000020

Description	10 COLOUR SPACEBOY PEN	12 COLOURED PARTY BALLOONS	12 EGG HOUSE PAINTED WOOD	12 MESSAGE CARDS WITH ENVELOPES	12 PENCIL SMALL TUBE WOODLAND	12 PENCILS SMALL TUBE RED RETROSPOT	12 PENCILS SMALL TUBE SKULL
InvoiceNo							
536370	0.0	0.0	0.0	0.0	0.0	0.0	0.0
536852	0.0	0.0	0.0	0.0	0.0	0.0	0.0
536974	0.0	0.0	0.0	0.0	0.0	0.0	0.0
537065	0.0	0.0	0.0	0.0	0.0	0.0	0.0
537463	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5 rows × 1563 columns

Pada gambar di atas telah dilakukan pembersihan data pada variabel description dan menghilangkan data yang memiliki invoice tidak valid

```
# Show a subset of columns
basket.iloc[:,[0,1,2,3,4,5,6, 7]].head()
```

Description	10 COLOUR SPACEBOY PEN	12 COLOURED PARTY BALLOONS	12 EGG HOUSE PAINTED WOOD	12 MESSAGE CARDS WITH ENVELOPES	12 PENCIL SMALL TUBE WOODLAND	12 PENCILS SMALL TUBE RED RETROSPOT	12 PENCILS SMALL TUBE SKULL
InvoiceNo							
536370	0.0	0.0	0.0	0.0	0.0	0.0	0.0
536852	0.0	0.0	0.0	0.0	0.0	0.0	0.0
536974	0.0	0.0	0.0	0.0	0.0	0.0	0.0
537065	0.0	0.0	0.0	0.0	0.0	0.0	0.0
537463	0.0	0.0	0.0	0.0	0.0	0.0	0.0

```
# Convert the units to 1 hot encoded values
def encode_units(x):
    if x <= 0:
        return 0
    if x >= 1:
        return 1
basket_sets = basket.applymap(encode_units)
# No need to track postage
basket_sets.drop('POSTAGE', inplace=True, axis=1)
basket_sets.head()
```

DIAN VITIANA NINGRUM
0621154000020

Description	10 COLOUR SPACEBOY PEN	12 COLOURED PARTY BALLOONS	12 EGG HOUSE PAINTED WOOD	12 MESSAGE CARDS WITH ENVELOPES	12 PENCIL SMALL TUBE WOODLAND	12 PENCILS SMALL TUBE RED RETROSPOT	12 PENCILS SMALL TUBE SKULL
InvoiceNo							
536370	0	0	0	0	0	0	0
536852	0	0	0	0	0	0	0
536974	0	0	0	0	0	0	0
537065	0	0	0	0	0	0	0
537463	0	0	0	0	0	0	0

5 rows × 1562 columns

Kemudian dilakukan pengkodean, jika x <= 0, maka dikoding 0 dan jika x >= 1 maka dikoding 1

```
# Build up the frequent items# Build

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

frequent_itemsets.head()
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-22-f20b5b441eb6> in <module>()
      1 # Build up the frequent items# Build
----> 2 frequent_itemsets = apriori(basket_sets, min_support=0.07, use_colname
s=True)

NameError: name 'apriori' is not defined
```

Frekuensi itemset tidak dapat diketahui sebab Apriori tidak dapat digunakan karena package mlextend tidak dapat diinstall.

```
# Create the rules# Create

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

rules
```

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-24-7f5690a745b0> in <module>()
      1 # Create the rules# Create
----> 2 rules = association_rules(frequent_itemsets, metric="lift", min_thresh
old=1)
      3 rules

NameError: name 'association_rules' is not defined
```

DIAN VITIANA NINGRUM
0621154000020

Karena frekuensi itemsets tidak diketahui, maka association rules juga tidak dapat dilakukan.

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

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

```
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
<ipython-input-25-f4d9642da78b> in <module>()
----> 1 rules[ (rules['lift'] >= 6) &
      2        (rules['confidence'] >= 0.8) ]

NameError: name 'rules' is not defined
```

Lift dan Confidence pun tidak dapat diketahui.

DIAN VITIANA NINGRUM
0621154000020