Vương Thanh Linh

MSSV: 21521082

1. Đọc dữ liệu

In [1]:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline

from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules
from mlxtend.frequent_patterns import fpgrowth
```

In [2]:

```
df = pd.read_excel('Online Retail.xlsx', engine='openpyxl')
df.head()
```

Out[2]:

	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. Tiền xử lý dữ liệu

In [3]:

```
df['Description'] = df['Description'].str.strip()
df.dropna(axis=0, subset=['InvoiceNo'], inplace=True)
df['InvoiceNo'] = df['InvoiceNo'].astype('str')
df.head(10)
```

Out[3]:

	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
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.65	17850.0	United Kingdom
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.25	17850.0	United Kingdom
7	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom
8	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	1.69	13047.0	United Kingdom

3. Xóa hóa đơn tín dụng (Chứa kí tự 'C')

In [4]:

```
df[df.InvoiceNo.str.contains('C', na=False)].head(10)
```

Out[4]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
141	C536379	D	Discount	-1	2010-12-01 09:41:00	27.50	14527.0	United Kingdom
154	C536383	35004C	SET OF 3 COLOURED FLYING DUCKS	-1	2010-12-01 09:49:00	4.65	15311.0	United Kingdom
235	C536391	22556	PLASTERS IN TIN CIRCUS PARADE	-12	2010-12-01 10:24:00	1.65	17548.0	United Kingdom
236	C536391	21984	PACK OF 12 PINK PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom
237	C536391	21983	PACK OF 12 BLUE PAISLEY TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom
238	C536391	21980	PACK OF 12 RED RETROSPOT TISSUES	-24	2010-12-01 10:24:00	0.29	17548.0	United Kingdom
239	C536391	21484	CHICK GREY HOT WATER BOTTLE	-12	2010-12-01 10:24:00	3.45	17548.0	United Kingdom
240	C536391	22557	PLASTERS IN TIN VINTAGE PAISLEY	-12	2010-12-01 10:24:00	1.65	17548.0	United Kingdom
241	C536391	22553	PLASTERS IN TIN SKULLS	-24	2010-12-01 10:24:00	1.65	17548.0	United Kingdom
939	C536506	22960	JAM MAKING SET WITH JARS	-6	2010-12-01 12:38:00	4.25	17897.0	United Kingdom

File failed to load: /extensions/MathZoom.js

In [5]:

```
df = df[~df['InvoiceNo'].str.contains('C')]
df[df.InvoiceNo.str.contains('C', na=False)].head()
```

Out[5]:

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country

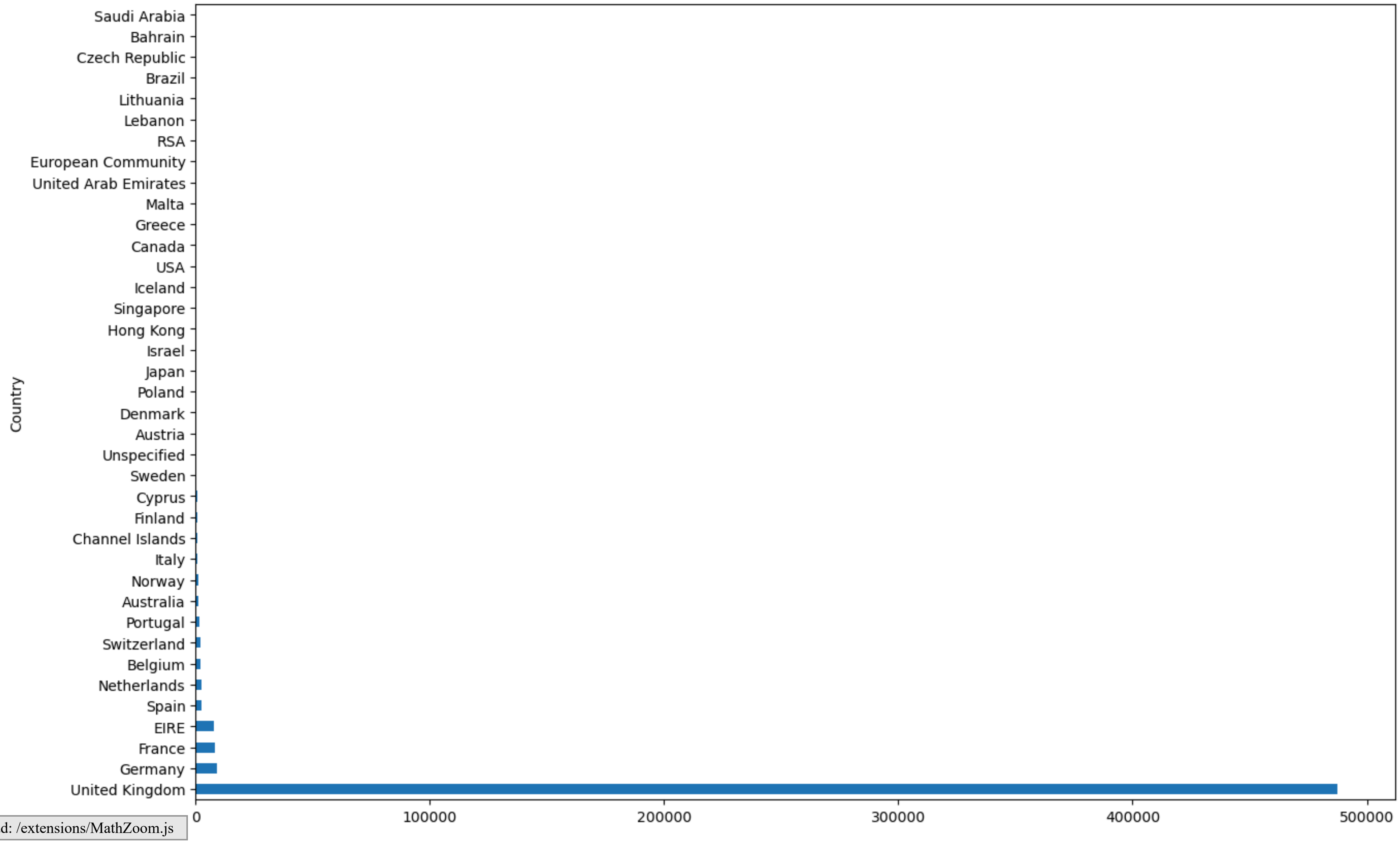
4. Thống kê dữ liệu theo từng quốc gia

In [6]:

```
df['Country'].value_counts().plot(kind='barh', figsize=(15,10))
```

Out[6]:

```
<Axes: ylabel='Country'>
```



File failed to load: /extensions/MathZoom.js

5. Xét hóa đơn tại nước Đức theo InvoiceNo và Tên mặt hàng

In [7]:

```
basket = df[df.Country == "Germany"].groupby(['InvoiceNo', 'Description'])['Quantity']
```

6. Chuyển đổi dữ liệu về hot-encoding

In [8]:

```
basket = basket.sum().unstack().reset_index().fillna(0).set_index('InvoiceNo')
basket.head(10)
```

Out[8]:

Description	10 COLOUR SPACEBOY PEN	12 COLOURED PARTY BALLOONS	12 IVORY ROSE PEG PLACE SETTINGS	12 MESSAGE CARDS WITH ENVELOPES	12 PENCIL SMALL TUBE WOODLAND	12 PENCILS SMALL TUBE RED RETROSPOT	12 PENCILS SMALL TUBE SKULL	12 PENCILS TALL TUBE POSY	12 PENCILS TALL TUBE RED RETROSPOT	12 PENCILS TALL TUBE SKULLS	...	YULETI IMAG GI WRA S
InvoiceNo												
536527	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0
536840	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0
536861	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0
536967	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0
536983	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0
537197	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0
537198	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0
537201	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0
537212	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0
537250	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0

10 rows × 1695 columns

7. Tạo hàm biến đổi dữ liêu có số lượng (Quantity)

File failed to load: /extensions/MathZoom.js

In [9]:
```
def encode_data(data):
    if data <= 0:
        return 0
    if data >= 1:
        return 1
```

In [10]:
```
basket = basket.map(encode_data)
basket.head(10)
```

Out[10]:

Description	10 COLOUR SPACEBOY PEN	12 COLOURED PARTY BALLOONS	12 IVORY ROSE PEG PLACE SETTINGS	12 MESSAGE CARDS WITH ENVELOPES	12 PENCIL SMALL TUBE WOODLAND	12 PENCILS SMALL TUBE RED RETROSPOT	12 PENCILS SMALL TUBE SKULL	12 PENCILS TALL TUBE POSY	12 PENCILS TALL TUBE RED RETROSPOT	12 PENCILS TALL TUBE SKULLS	...	YULETII IMAGI GII WRA SI
InvoiceNo												
536527	0	0	0	0	0	0	0	0	0	0	...	
536840	0	0	0	0	0	0	0	0	0	0	...	
536861	0	0	0	0	0	0	0	0	0	0	...	
536967	0	0	0	0	0	0	0	0	0	0	...	
536983	0	0	0	0	0	0	0	0	0	0	...	
537197	0	0	0	0	0	0	0	0	0	0	...	
537198	0	0	0	0	0	0	0	0	0	0	...	
537201	0	0	0	0	0	0	0	0	0	0	...	
537212	0	0	0	0	0	0	0	0	0	0	...	
537250	0	0	0	0	0	0	0	0	0	0	...	

10 rows × 1695 columns

8. Xóa cột 'POSTAGE'

In [11]:
```
basket.drop('POSTAGE', inplace=True, axis=1)
```

File failed to load: /extensions/MathZoom.js

9. Apriori với min_support = 5%

In [12]:

```
itemsets = apriori(basket.astype('bool'), min_support=0.05, use_colnames=True)
itemsets.head(10)
```

Out[12]:

	support	itemsets
0	0.102845	(6 RIBBONS RUSTIC CHARM)
1	0.070022	(ALARM CLOCK BAKELIKE PINK)
2	0.065646	(CHARLOTTE BAG APPLES DESIGN)
3	0.050328	(CHILDRENS CUTLERY DOLLY GIRL)
4	0.061269	(COFFEE MUG APPLES DESIGN)
5	0.063457	(FAWN BLUE HOT WATER BOTTLE)
6	0.072210	(GUMBALL COAT RACK)
7	0.056893	(IVORY KITCHEN SCALES)
8	0.063457	(JAM JAR WITH PINK LID)
9	0.091904	(JAM MAKING SET PRINTED)

10. Tạo luật kết hợp với min_conf = 50%

In [13]:

```
rules = association_rules(itemsets, metric="confidence", min_threshold=0.5)
rules.info()
```

File failed to load: /extensions/MathZoom.js

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8 entries, 0 to 7
Data columns (total 10 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   antecedents         8 non-null      object 
 1   consequents         8 non-null      object 
 2   antecedent support  8 non-null      float64
 3   consequent support  8 non-null      float64
 4   support             8 non-null      float64
 5   confidence          8 non-null      float64
 6   lift                8 non-null      float64
 7   leverage            8 non-null      float64
 8   conviction          8 non-null      float64
 9   zhangs_metric       8 non-null      float64
dtypes: float64(8), object(2)
memory usage: 772.0+ bytes
```

In [14]:

```
rules['antecedents'] = rules['antecedents'].apply(lambda x: list(x)[0]).astype("unicode")
rules['consequents'] = rules['consequents'].apply(lambda x: list(x)[0]).astype("unicode")
for i in range(len(rules)):
    print(rules.loc[i, 'antecedents'], ' ==> ', rules.loc[i, 'consequents'],
          ' [', rules.loc[i, 'support'], ', ', rules.loc[i, 'confidence'], ']')
```

```
PLASTERS IN TIN CIRCUS PARADE  ==>  PLASTERS IN TIN WOODLAND ANIMALS  [ 0.06783369803063458 ,  0.5849056603773585 ]
PLASTERS IN TIN SPACEBOY  ==>  PLASTERS IN TIN WOODLAND ANIMALS  [ 0.061269146608315096 ,  0.5714285714285714 ]
PLASTERS IN TIN WOODLAND ANIMALS  ==>  ROUND SNACK BOXES SET OF4 WOODLAND  [ 0.07439824945295405 ,  0.5396825396825397 ]
RED RETROSPOT CHARLOTTE BAG  ==>  WOODLAND CHARLOTTE BAG  [ 0.05908096280087528 ,  0.8437500000000001 ]
ROUND SNACK BOXES SET OF 4 FRUITS  ==>  ROUND SNACK BOXES SET OF4 WOODLAND  [ 0.13129102844638948 ,  0.8333333333333333 ]
ROUND SNACK BOXES SET OF4 WOODLAND  ==>  ROUND SNACK BOXES SET OF 4 FRUITS  [ 0.13129102844638948 ,  0.5357142857142857 ]
SPACEBOY LUNCH BOX  ==>  ROUND SNACK BOXES SET OF4 WOODLAND  [ 0.0700218818380744 ,  0.6808510638297872 ]
WOODLAND CHARLOTTE BAG  ==>  ROUND SNACK BOXES SET OF4 WOODLAND  [ 0.06345733041575492 ,  0.5 ]
```

11. Lấy giá trị độ hỗ trợ và độ tin cậy

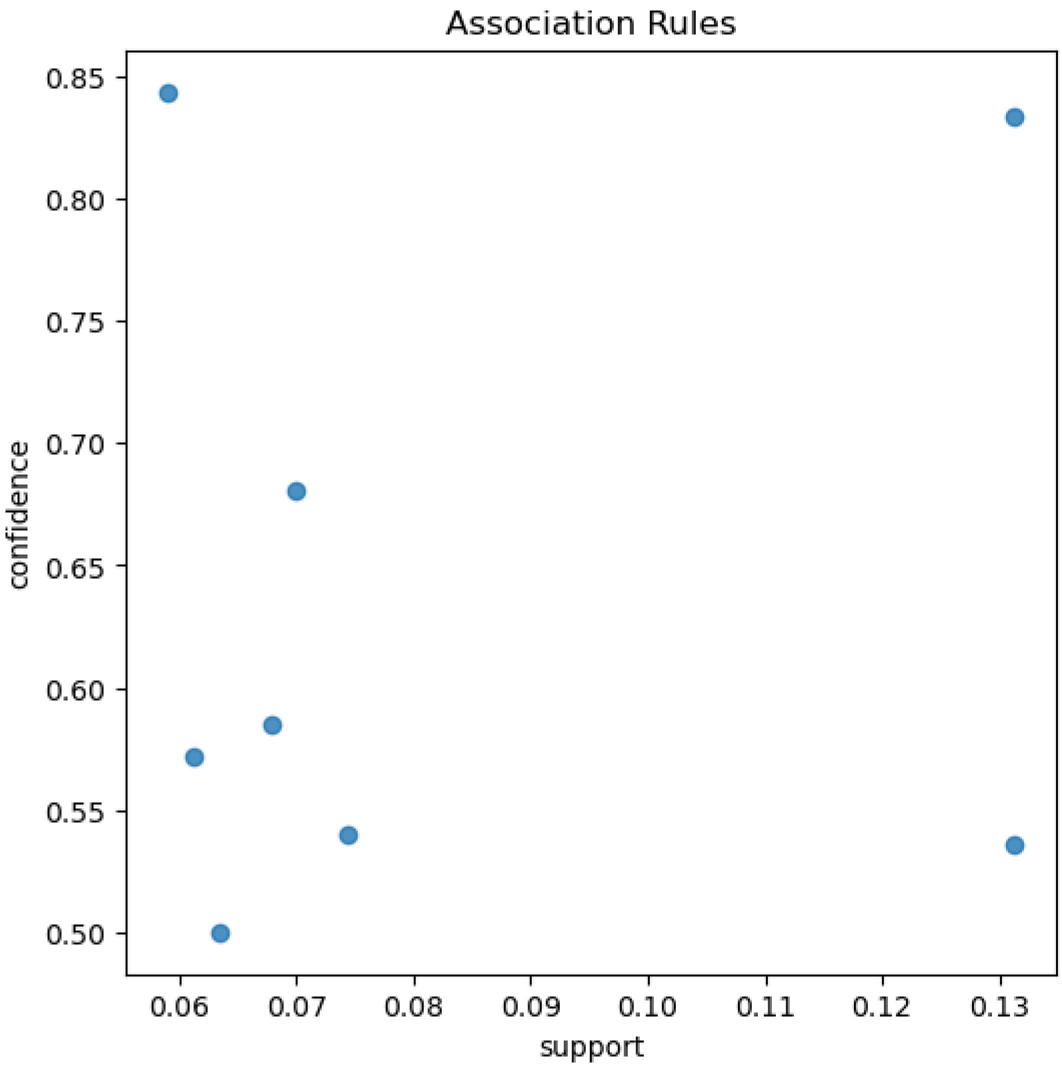
In [15]:

```
support = rules['support'].values
confidence = rules['confidence'].values
# Plot
plt.figure(figsize=(6,6))
plt.title('Association Rules')
plt.xlabel('support')
plt.ylabel('confidence')
sns.regplot(x=support, y=confidence, fit_reg=False)
```

Out[15]:

```
<Axes: title={'center': 'Association Rules'}, xlabel='support', ylabel='confidence'>
```

File failed to load: /extensions/MathZoom.js



12. Tìm tập phổ biến bằng FP-Growth

```
In [18]: itemsets_fp = fpgrowth(basket.astype('bool'), min_support=0.05, use_colnames=True)
         itemsets_fp.tail(10)
```

Out[18]:

	support	itemsets
49	0.056893	(SET OF 60 PANTRY DESIGN CAKE CASES)
50	0.131291	(ROUND SNACK BOXES SET OF 4 FRUITS, ROUND SNAC...
51	0.063457	(WOODLAND CHARLOTTE BAG, ROUND SNACK BOXES SET...
52	0.056893	(PLASTERS IN TIN CIRCUS PARADE, ROUND SNACK BO...
53	0.050328	(ROUND SNACK BOXES SET OF 4 FRUITS, PLASTERS I...
54	0.067834	(PLASTERS IN TIN CIRCUS PARADE, PLASTERS IN TI...
55	0.070022	(SPACEBOY LUNCH BOX, ROUND SNACK BOXES SET OF4...
56	0.059081	(WOODLAND CHARLOTTE BAG, RED RETROSPOT CHARLOT...
57	0.074398	(PLASTERS IN TIN WOODLAND ANIMALS, ROUND SNACK...
58	0.061269	(PLASTERS IN TIN SPACEBOY, PLASTERS IN TIN WOO...

In [17]:

```
itemsets.tail(10)
```

Out[17]:

	support	itemsets
49	0.067834	(WOODLAND PARTY BAG + STICKER SET)
50	0.067834	(PLASTERS IN TIN CIRCUS PARADE, PLASTERS IN TI...
51	0.050328	(ROUND SNACK BOXES SET OF 4 FRUITS, PLASTERS I...
52	0.056893	(PLASTERS IN TIN CIRCUS PARADE, ROUND SNACK BO...
53	0.061269	(PLASTERS IN TIN SPACEBOY, PLASTERS IN TIN WOO...
54	0.074398	(PLASTERS IN TIN WOODLAND ANIMALS, ROUND SNACK...
55	0.059081	(WOODLAND CHARLOTTE BAG, RED RETROSPOT CHARLOT...
56	0.131291	(ROUND SNACK BOXES SET OF 4 FRUITS, ROUND SNAC...
57	0.070022	(SPACEBOY LUNCH BOX, ROUND SNACK BOXES SET OF4...
58	0.063457	(WOODLAND CHARLOTTE BAG, ROUND SNACK BOXES SET...

File failed to load: /extensions/MathZoom.js