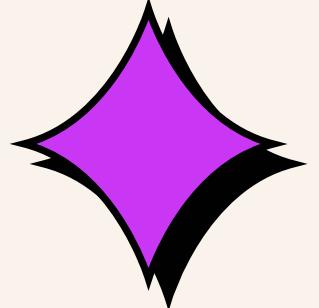
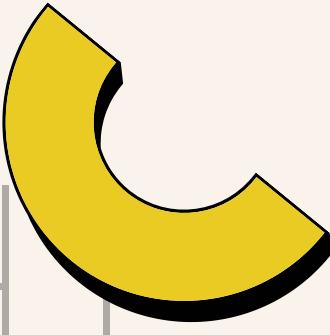


SC663403 DATA PREPARATION AND DATA MINING

HW3

Data Reduced_Marketbasket

กลุ่ม 7



MEMBERS

กัญรัตน์ อินกร์สำราญ 653020567-7

ธนากรณ์ ดาษณิม 653020570-8

กนกน้ำดี เครื่องเนตร 653020599-4



DATA CLEANSING

```
# Drop rows with null values in 'CustomerID' column
rmb = rmb.dropna(subset=['CustomerID'])

# Convert 'CustomerID' to integer
rmb['CustomerID'] = rmb['CustomerID'].astype(int)

# Calculate sales for Europe
rmb['Sales'] = rmb['Quantity'] * rmb['UnitPrice']

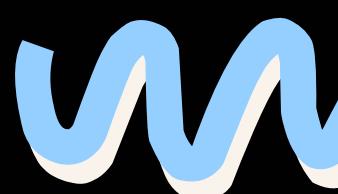
# prompt: InvoiceDate แยกเวลาและวันที่ออกจากกัน
# Assuming 'InvoiceDate' is a datetime column. If not, convert it first:
rmb['InvoiceDate'] = pd.to_datetime(rmb['InvoiceDate'])

# Extract date and time into separate columns
rmb['Date'] = rmb['InvoiceDate'].dt.date
rmb['Time'] = rmb['InvoiceDate'].dt.time

# prompt: แบ่งtimeเป็น 3 ช่วง

# Function to categorize time into three periods
def categorize_time(time):
    if pd.Timestamp(time).time() >= pd.Timestamp('06:00:00').time() and pd.Timestamp(time).time() < pd.Timestamp('12:00:00').time():
        return 'Morning'
    elif pd.Timestamp(time).time() >= pd.Timestamp('12:00:00').time() and pd.Timestamp(time).time() < pd.Timestamp('18:00:00').time():
        return 'Afternoon'
    else:
        return 'Evening'

# Apply the function to create a new 'TimePeriod' column
rmb['TimePeriod'] = rmb['InvoiceDate'].apply(categorize_time)
```

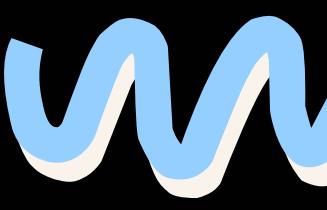


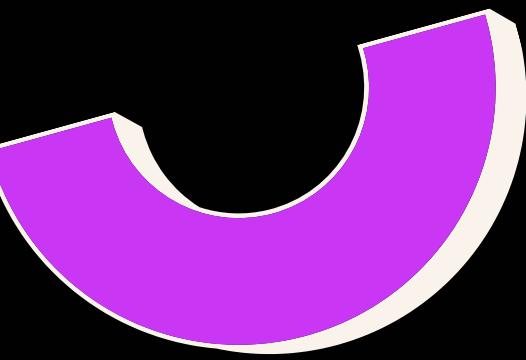


DATA CLEANSING

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Continent	Sales	Date	Time	TimePeriod
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom	Europe	15.30	2010-12-01	08:26:00	Morning
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	Europe	20.34	2010-12-01	08:26:00	Morning
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom	Europe	22.00	2010-12-01	08:26:00	Morning
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	Europe	20.34	2010-12-01	08:26:00	Morning
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom	Europe	20.34	2010-12-01	08:26:00	Morning
...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	2011-12-09 12:50:00	0.85	12680	France	Europe	10.20	2011-12-09	12:50:00	Afternoon
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	2011-12-09 12:50:00	2.10	12680	France	Europe	12.60	2011-12-09	12:50:00	Afternoon
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	2011-12-09 12:50:00	4.15	12680	France	Europe	16.60	2011-12-09	12:50:00	Afternoon
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	2011-12-09 12:50:00	4.15	12680	France	Europe	16.60	2011-12-09	12:50:00	Afternoon
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	2011-12-09 12:50:00	4.95	12680	France	Europe	14.85	2011-12-09	12:50:00	Afternoon

406829 rows × 13 columns





**Q1. ວິເຄຣາະໜໍພຸຕິກຣມຜູ້ບຣິໂຄດ
(GROUP BY CUSTOMER ID) ແຕ່ລະກວີປ**



ແບ່ງກວົປອກເປັນ 7 ກວົປກລຸ່ມ ໄດ້ແກ່

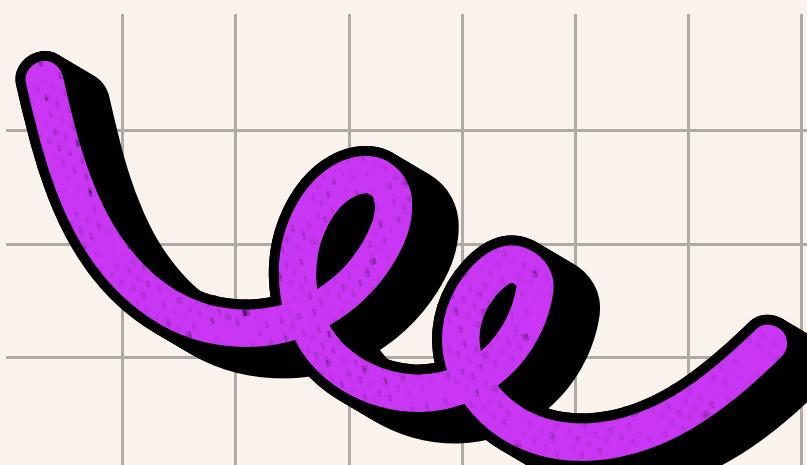
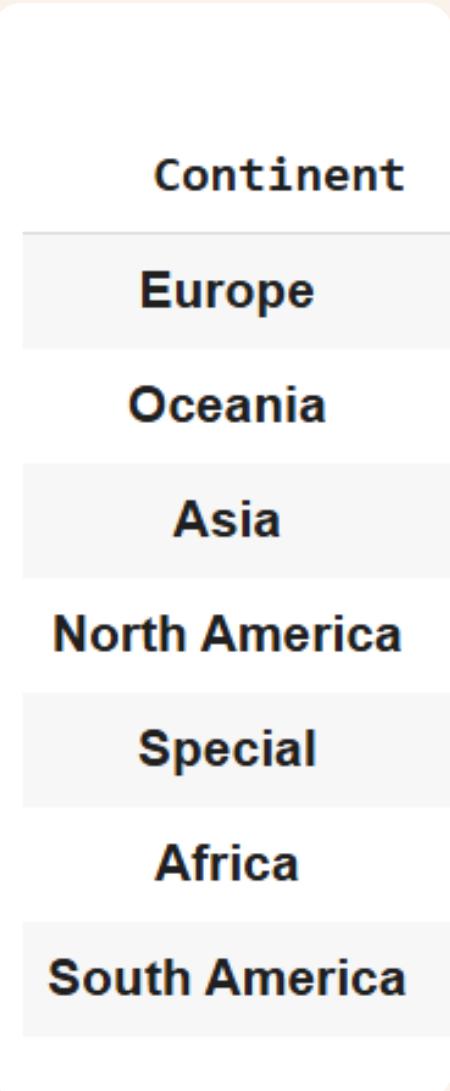
```
# ສ້າງຝຶກຊື່ເພື່ອກໍາທັດທວປ
def assign_continent(country):
    europe = ['United Kingdom', 'France', 'Netherlands', 'Germany',
              'Norway', 'EIRE', 'Switzerland', 'Spain', 'Poland',
              'Portugal', 'Italy', 'Belgium', 'Lithuania',
              'Iceland', 'Channel Islands', 'Denmark',
              'Sweden', 'Austria', 'Finland', 'Greece',
              'Cyprus', 'Malta', 'Czech Republic',
              'European Community']
    asia = ['Japan', 'Israel', 'Bahrain', 'Hong Kong',
            'Singapore', 'Lebanon',
            'United Arab Emirates', 'Saudi Arabia']
    north_america = ['Canada', 'USA']
    south_america = ['Brazil']
    oceania = ['Australia']
    africa = ['RSA']
    special = ['Unspecified']

    if country in europe:
        return 'Europe'
    elif country in asia:
        return 'Asia'
    elif country in north_america:
        return 'North America'
    elif country in south_america:
        return 'South America'
    elif country in oceania:
        return 'Oceania'
    elif country in africa:
        return 'Africa'
    elif country in special:
        return 'Special'
    else:
        return 'Unknown'

# ເພີ່ມຄອລິມນີ້ Continent
rmb['Continent'] = rmb['Country'].apply(assign_continent)

# ບັນທຶກໄຟລັບໄປທ໌ path ເດີມ
rmb.to_csv(os.path.join(path, 'reduced_marketbasket.csv'),
           index=False, encoding='latin-1')

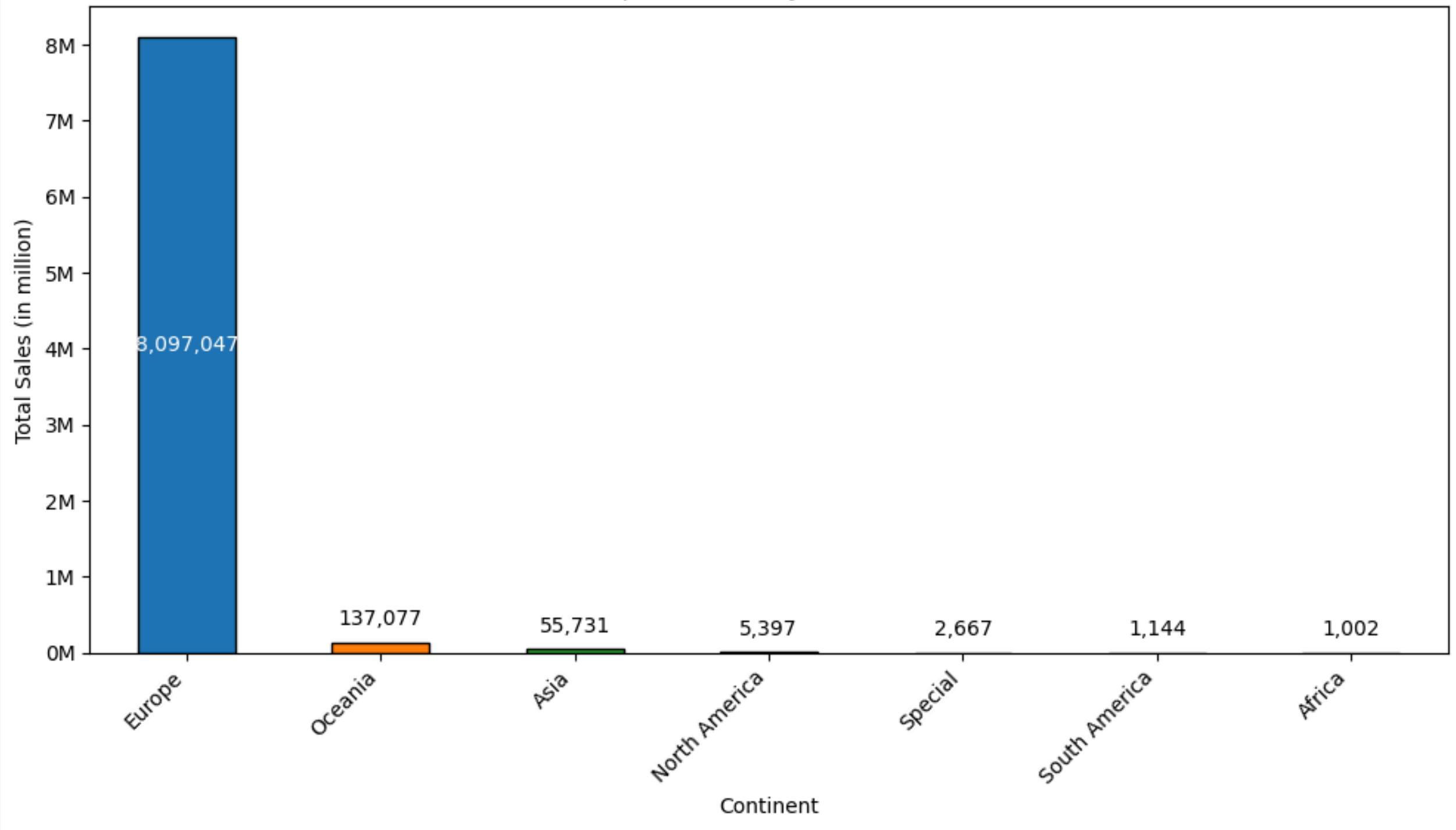
rmb
```





ກວົປກີ່ມືຍອດຂາຍຮວມສູງສຸດ

Top Continent by Total Sales

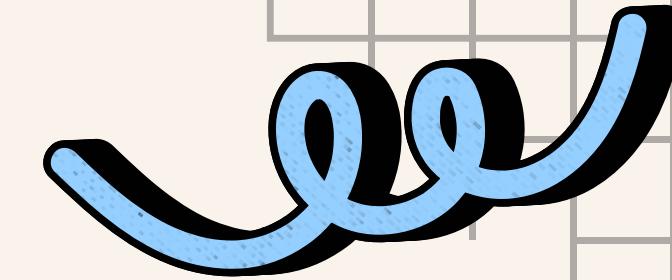


Continent	Sales
Europe	8097047.274
Oceania	137077.270
Asia	55730.990
North America	5397.300
Special	2667.070
South America	1143.600
Africa	1002.310

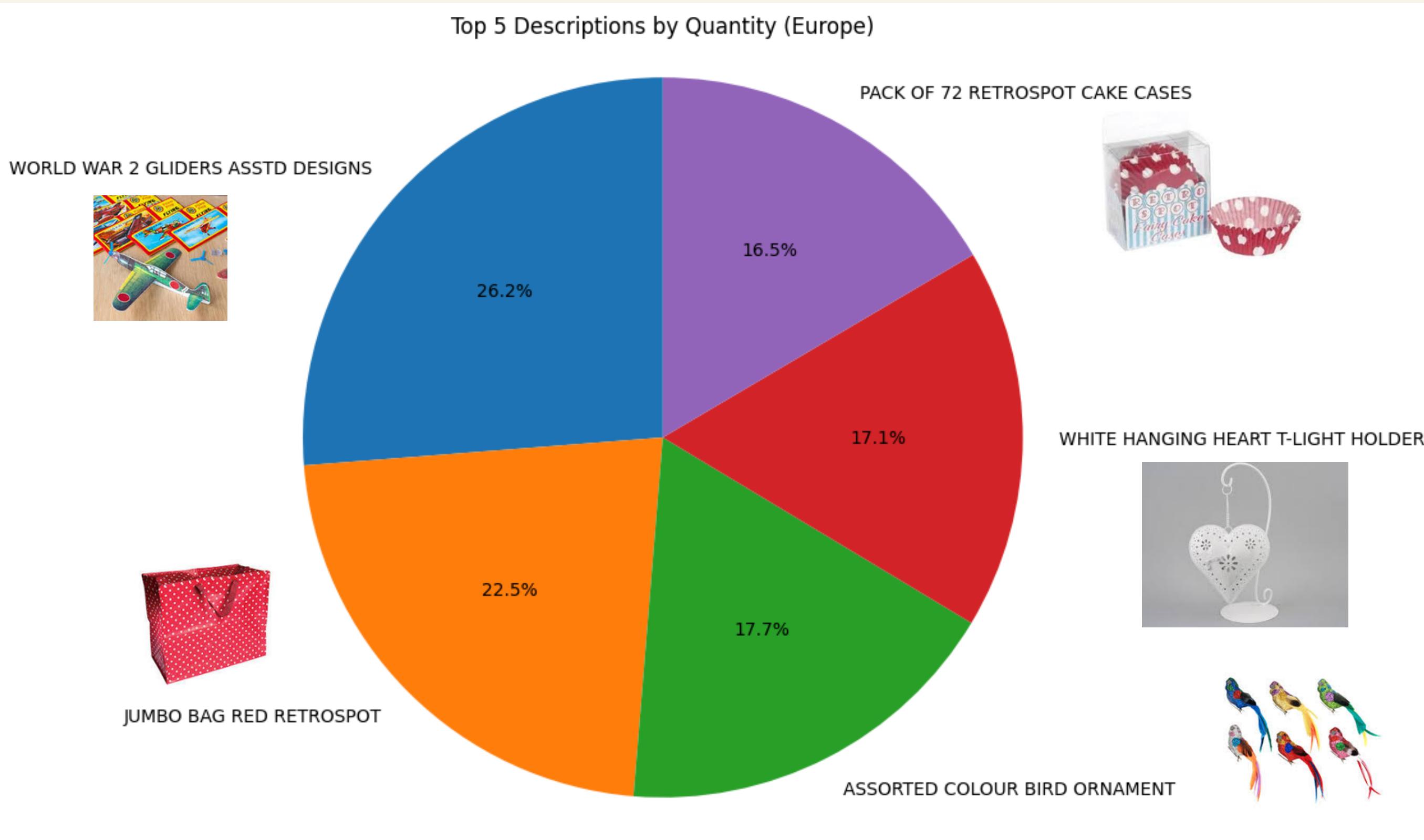




รายการสินค้าขายดีที่สุดในทวีป EUROPE

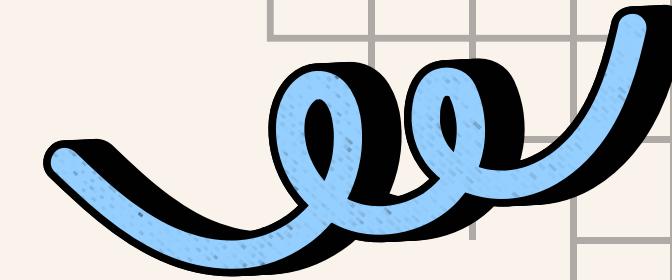


Top 5 Descriptions by Quantity (Europe)

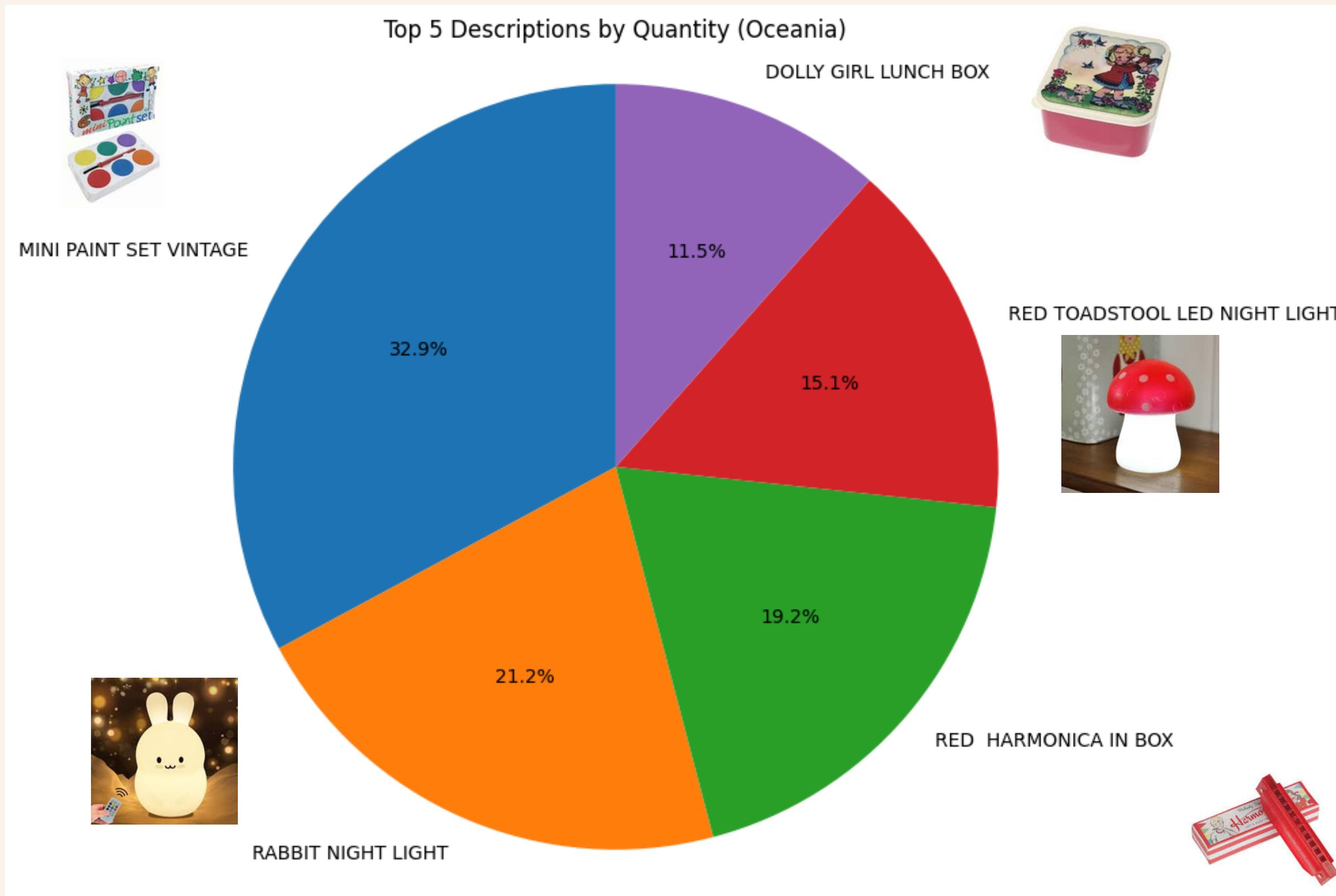




รายการสินค้าขายดีที่สุดในภูมิภาค OCEANIA

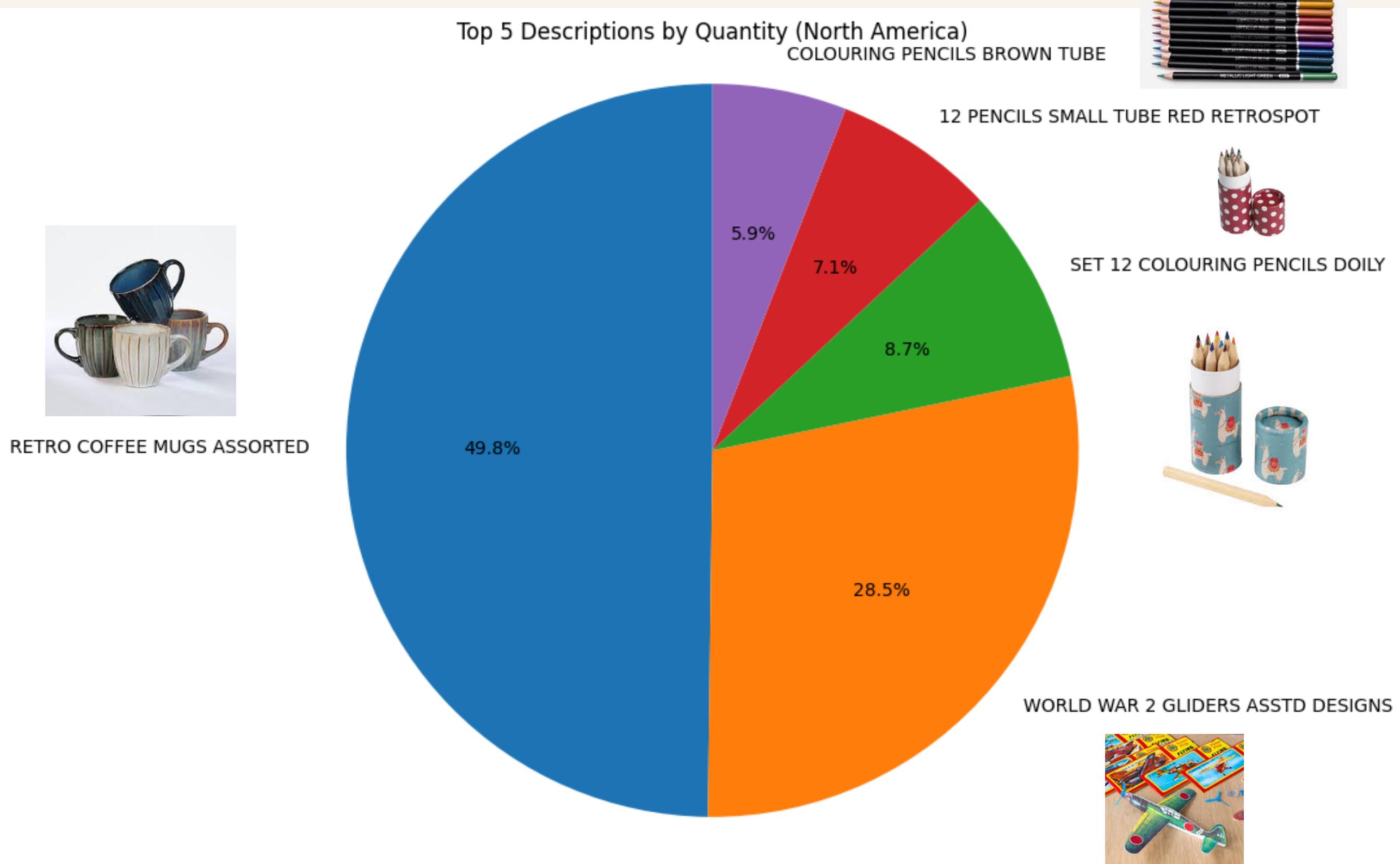


Top 5 Descriptions by Quantity (Oceania)



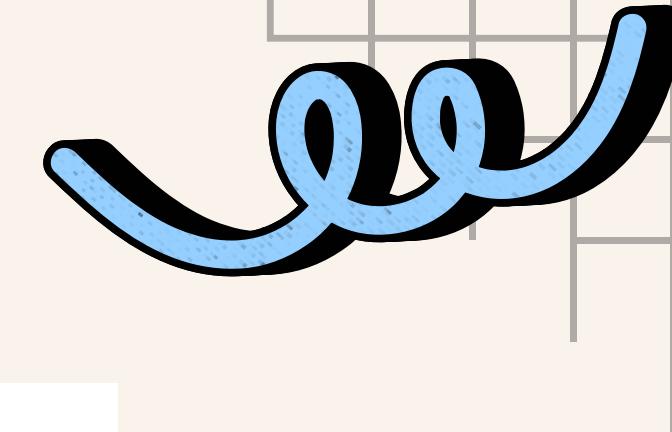


รายการสินค้าขายดีที่สุดในกวีป NORTH AMERICA





รายการสินค้าขายดีที่สุดในทวีป SOUTH AMERICA



Top 5 Descriptions by Quantity (South America)

ROSES REGENCY TEACUP AND SAUCER



SET OF 6 SPICE TINS PANTRY DESIGN



20.0%

20.0%

20.0%

20.0%

20.0%

SET/3 RED GINGHAM ROSE STORAGE BOX



SET OF 4 PANTRY JELLY MOULDS

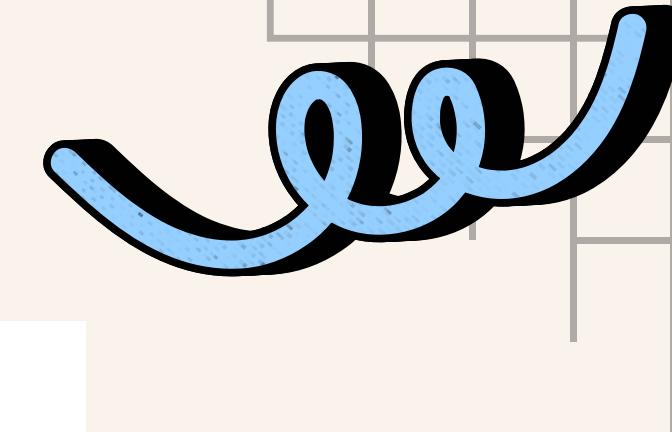


GREEN REGENCY TEACUP AND SAUCER

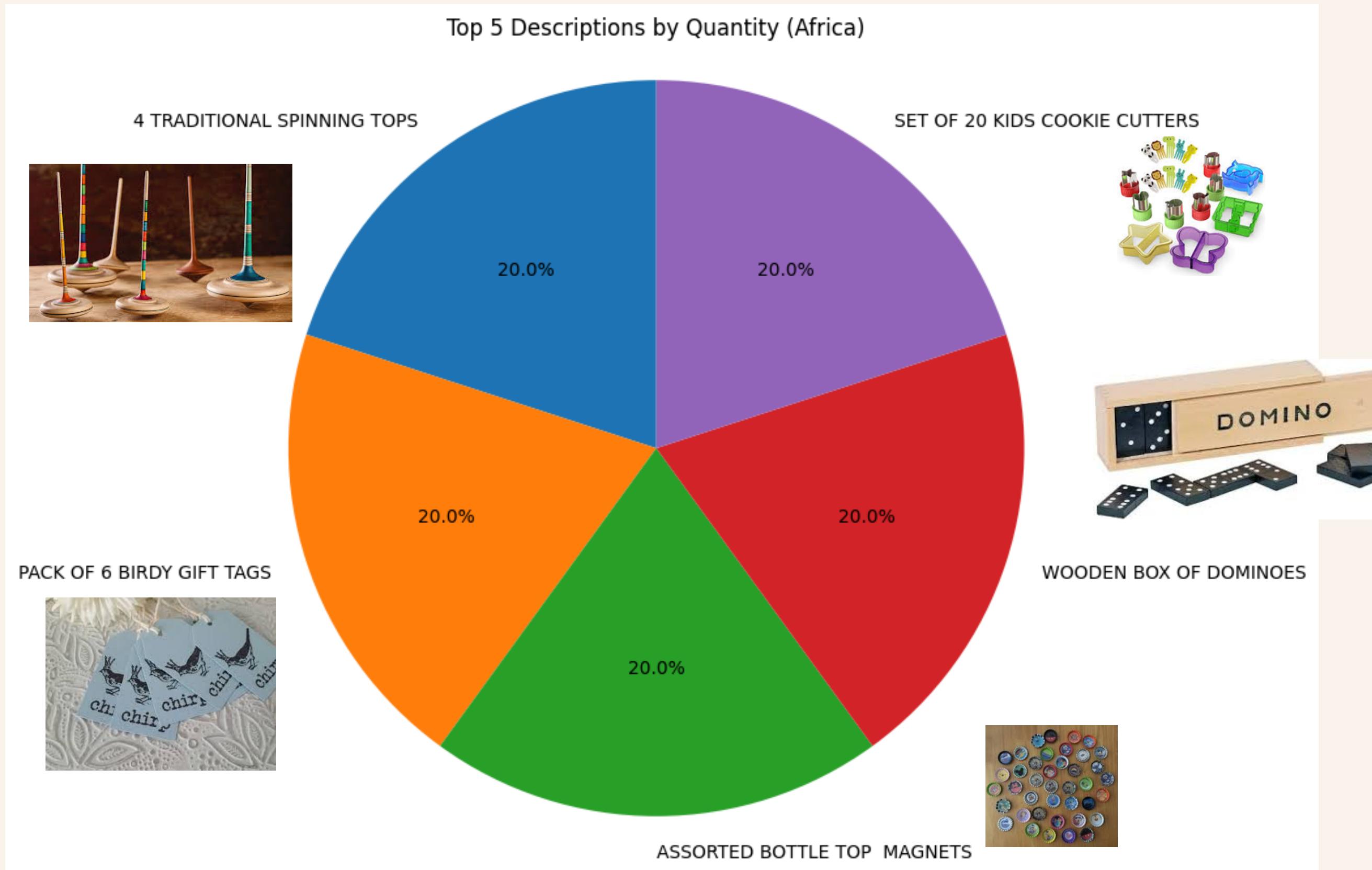




รายการสินค้าขายดีที่สุดในทวีป AFRICA

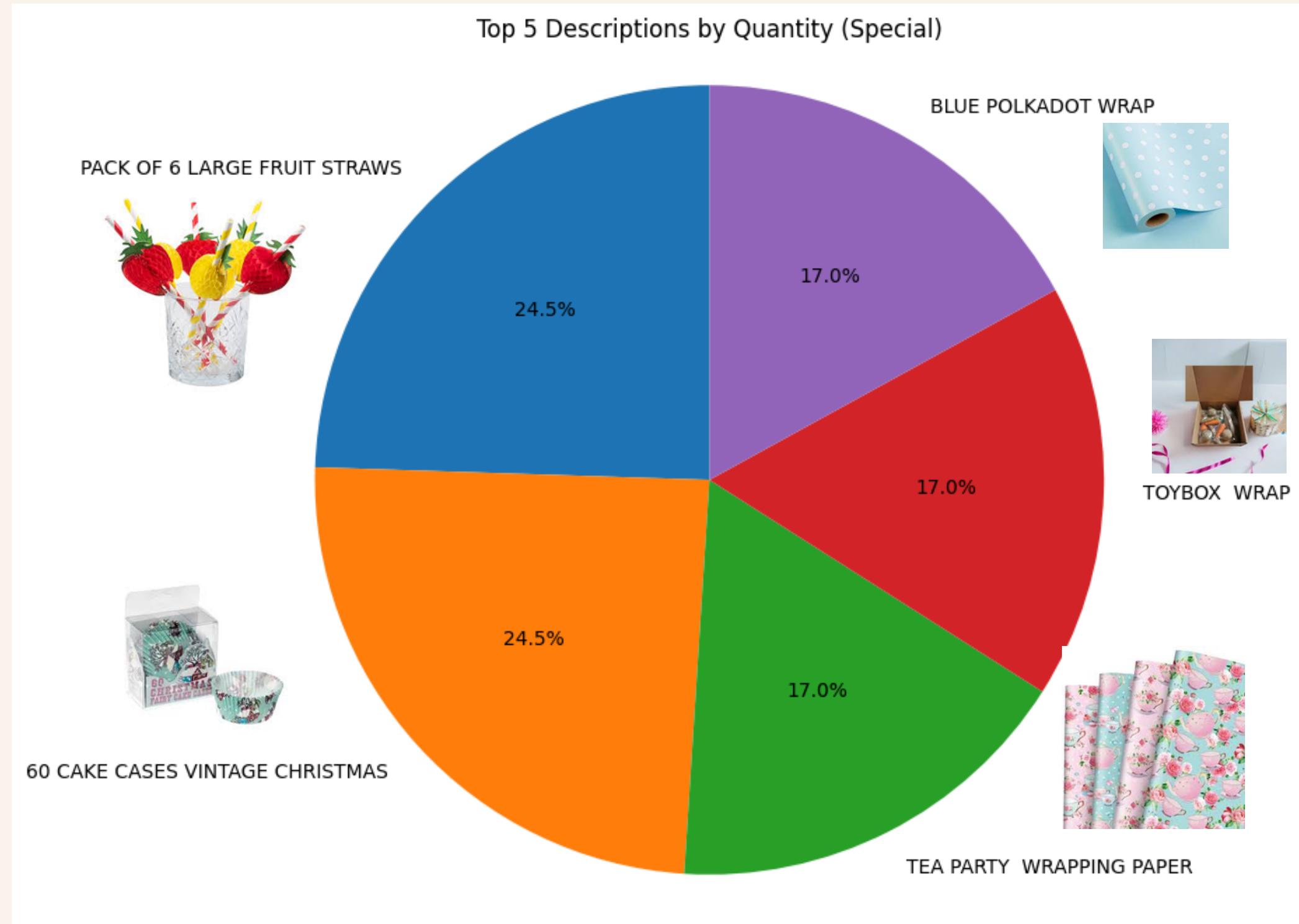
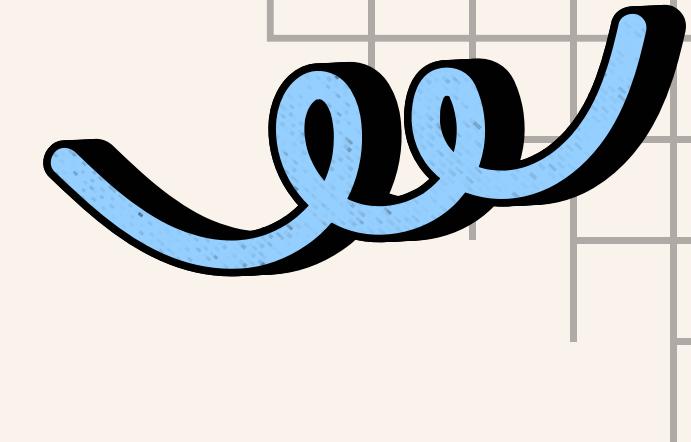


Top 5 Descriptions by Quantity (Africa)





รายการสินค้าขายดีที่สุดในกลุ่ม SPECIAL





TOP 3 MOST FREQUENT ITEM COMBINATIONS IN EUROPE

❖ สินค้าที่ซื้อคู่กันของทวีป EUROPE



GREEN REGENCY TEACUP AND SAUCER

ASSOCIATION RULES

Confidence : 0.8139
Lift : 8.1859



ROSES REGENCY TEACUP AND SAUCER



TOP 3 FREQUENT ITEM COMBINATION IN OCEANIA

❖ สินค้าที่ซื้อคู่กันของทวีป OCEANIA



↔ 38 ครั้ง



MINI PAINT SET VINTAGE



↔ 36 ครั้ง



SET OF 3 CAKE TINS PANTRY DESIGN



↔ 36 ครั้ง



DOLLY GIRL LUNCH BOX

MINI PAINT SET VINTAGE



TOP 3 FREQUENT ITEM COMBINATION IN ASIA

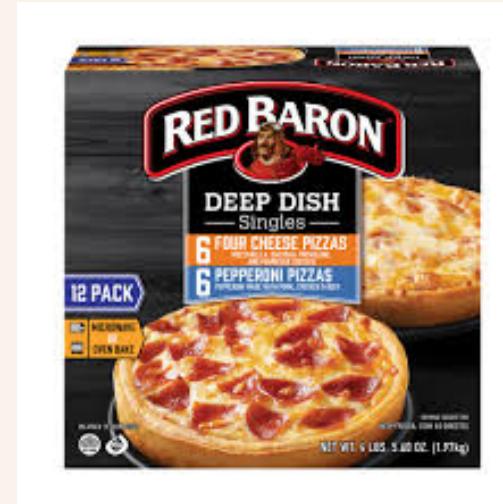
❖ สินค้าที่ซื้อคู่กันของทวีป ASIA



↔ 92 ครั้ง



↔ 56 ครั้ง



Manual

Manual

Manual

SET OF 3 CAKE TINS PANTRY DESIGN



↔ 42 ครั้ง



Manual

DOORMAT RED RETROSPOT



TOP 3 FREQUENT ITEM COMBINATION IN NORTH AMERICA

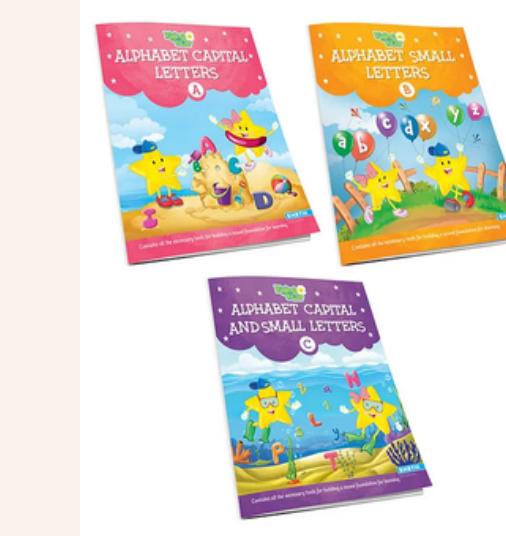
❖ สินค้าที่ซื้อคู่กันของทวีป AMERICA



↔ 8 ครั้ง



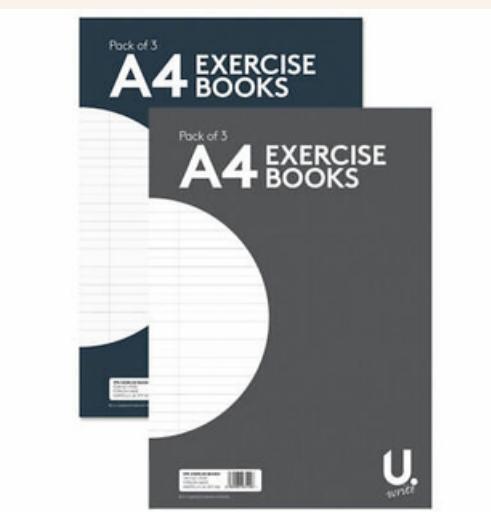
CARD DOLLY GIRL



↔ 6 ครั้ง



COLOURING PENCILS BROWN TUBE



↔ 6 ครั้ง



BUNDLE OF 3 SCHOOL EXERCISE BOOKS

COLOURING PENCILS BROWN TUBE



TOP 3 FREQUENT ITEM COMBINATION IN AFRICA

❖ สินค้าที่ซื้อคู่กันของทวีป AFRICA



↔ 1 ครั้ง



PINK POLKADOT PLATE



RED RETROSPOT CUP

↔ 1 ครั้ง



SET OF 4 KNICK KNACK TINS DOILY

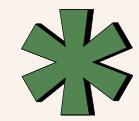


↔ 1 ครั้ง



LUNCH BAG VINTAGE DOILY

RED RETROSPOT CUP

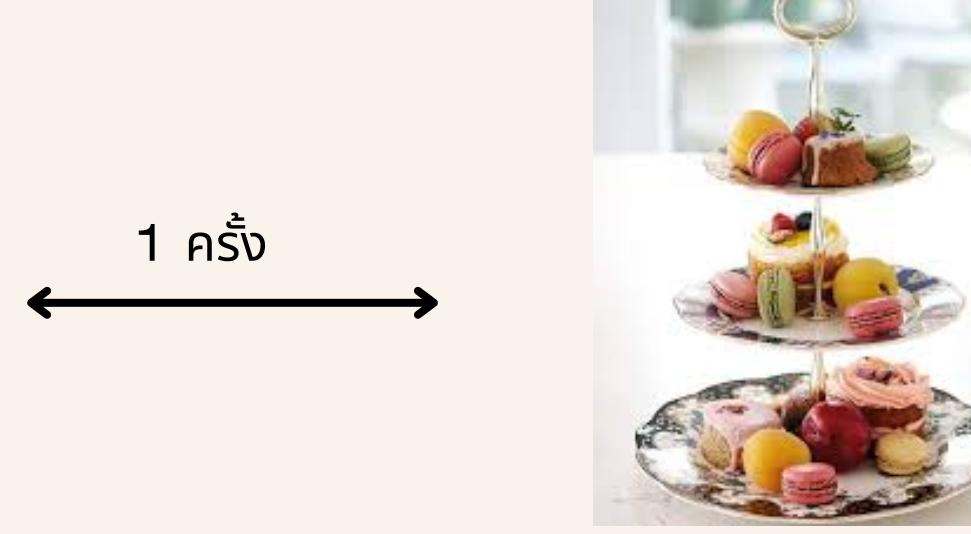


TOP 3 FREQUENT ITEM COMBINATION IN SOUTH AMERICA

❖ สินค้าที่ซื้อคู่กันของทวีป AMERICA



↔ 1 ครั้ง



REGENCY CAKESTAND 3 TIER

ROSES REGENCY TEACUP AND SAUCER

GREEN REGENCY TEACUP AND SAUCER

REGENCY CAKESTAND 3 TIER

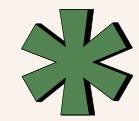


↔ 1 ครั้ง



LUNCH BAG VINTAGE DOILY

REGENCY CAKESTAND 3 TIER



TOP 3 FREQUENT ITEM COMBINATION IN SPECIAL

❖ สินค้าที่ซื้อคู่กันของวีปอีนๆ



↔ 6 ครั้ง



↔ 6 ครั้ง



3D SHEET OF DOG STICKERS

COSY HOUR CIGAR BOX MATCHES

COSY HOUR CIGAR BOX MATCHES

WHITE ROCKING HORSE HAND PAINTED



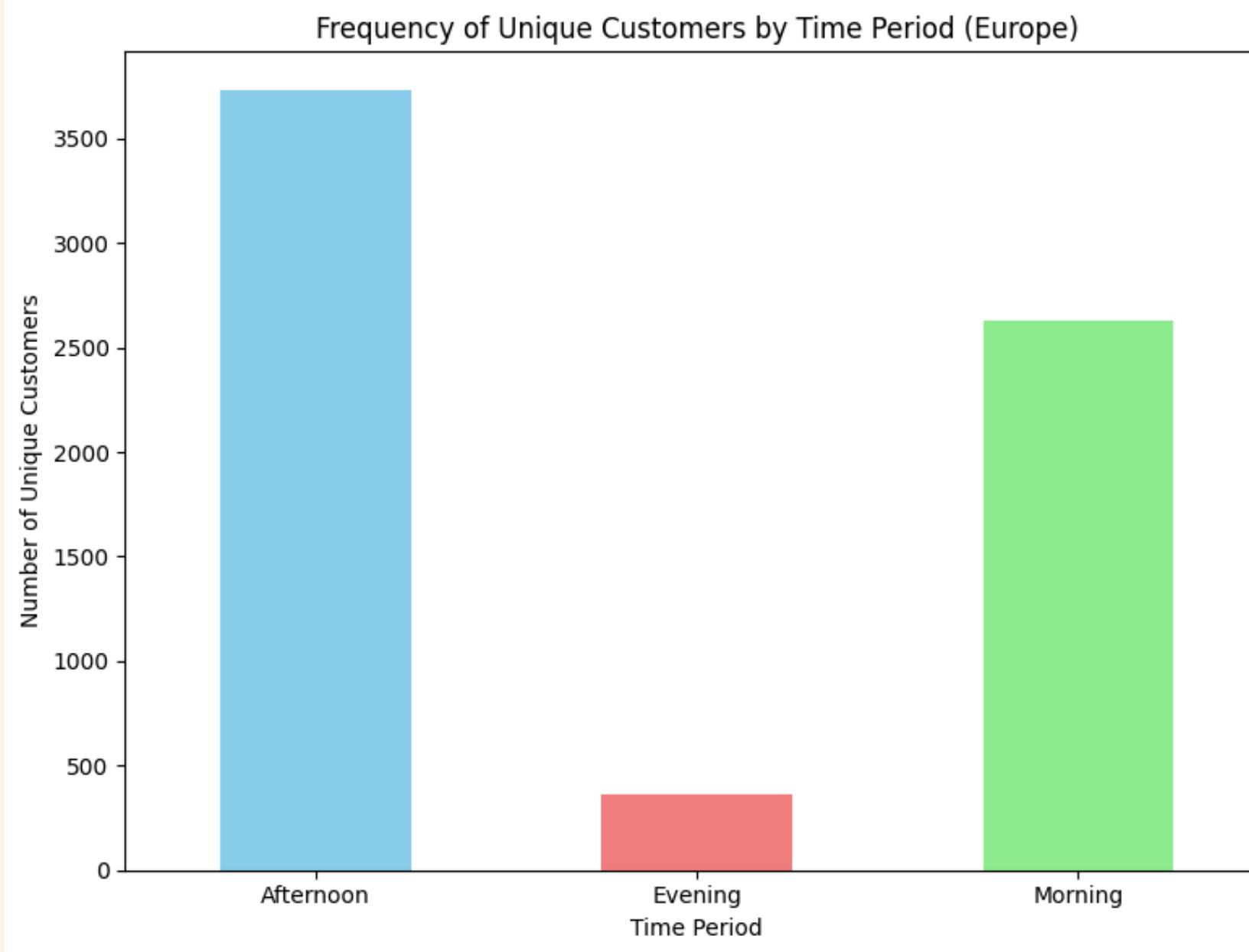
↔ 6 ครั้ง



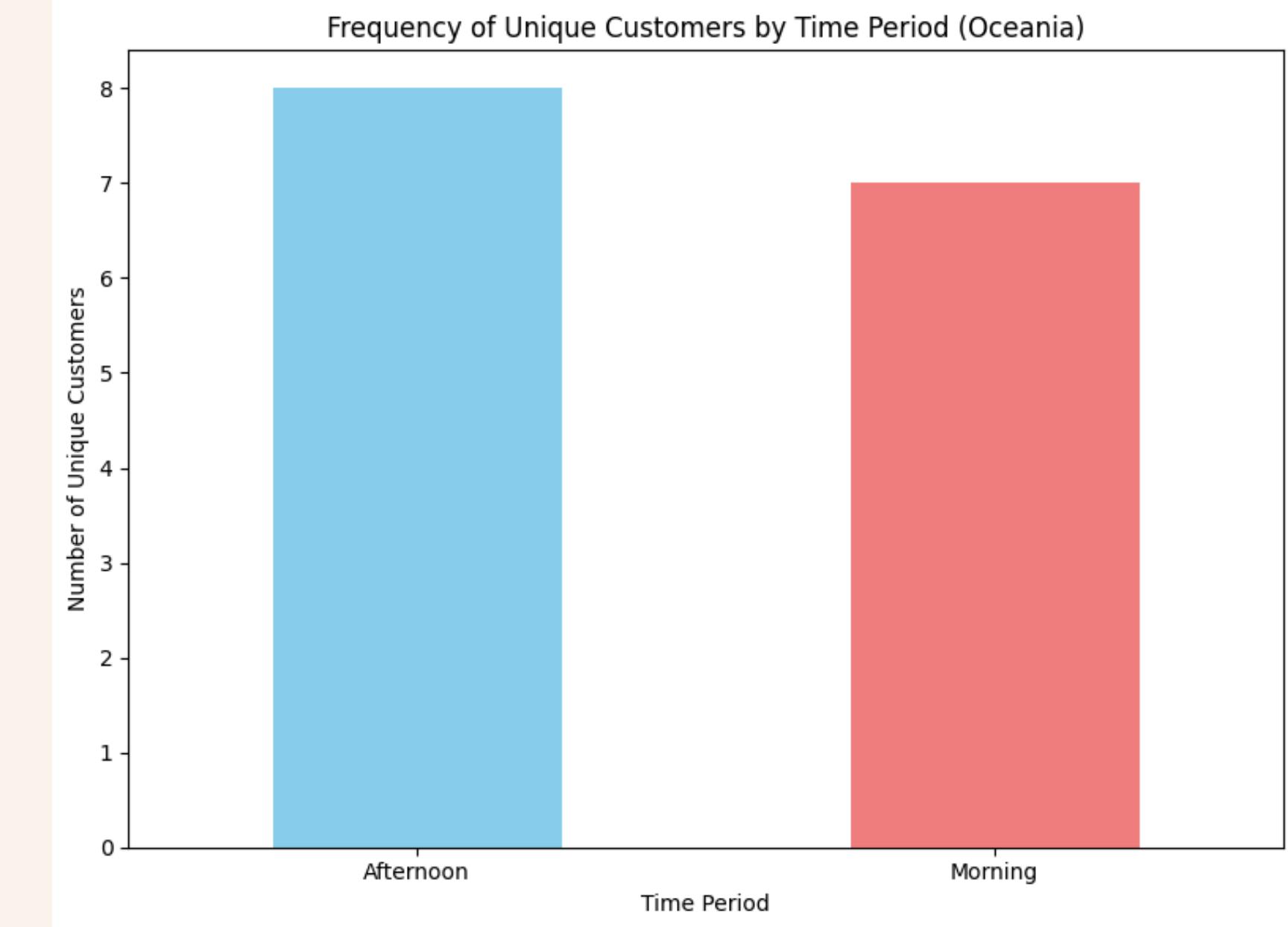
ASSORTED COLOUR BIRD ORNAMENT

COSY HOUR CIGAR BOX MATCHES

วิเคราะห์พฤติกรรมการซื้อในแต่ละช่วงเวลา

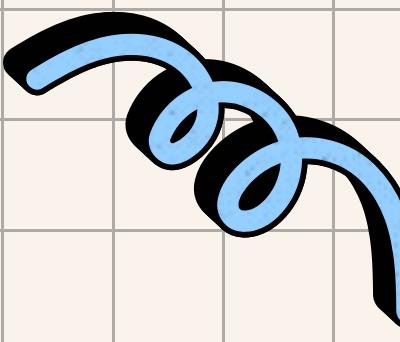


EUROPE

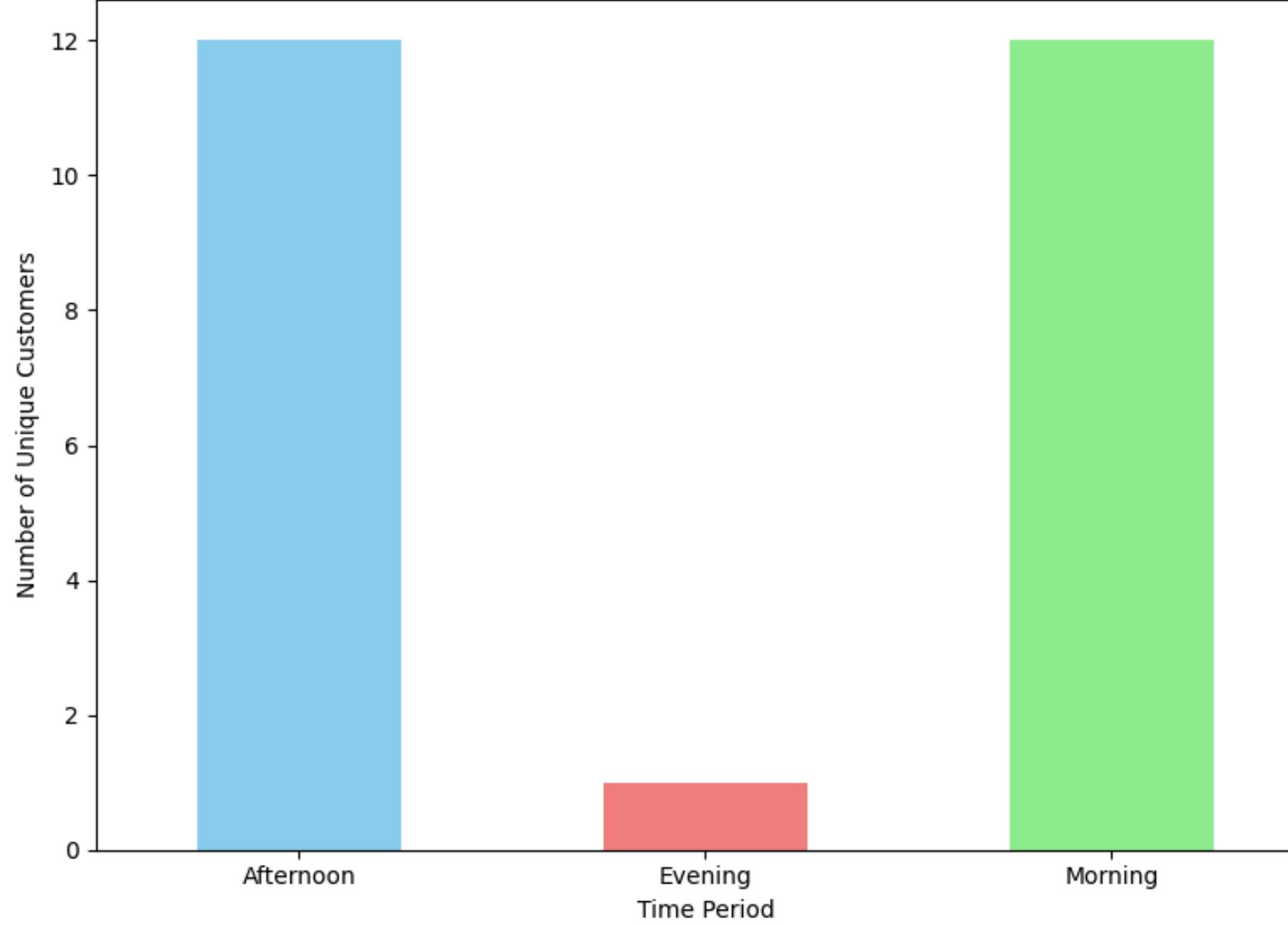


OCEANIA

วิเคราะห์พฤติกรรมการซื้อในแต่ละช่วงเวลา

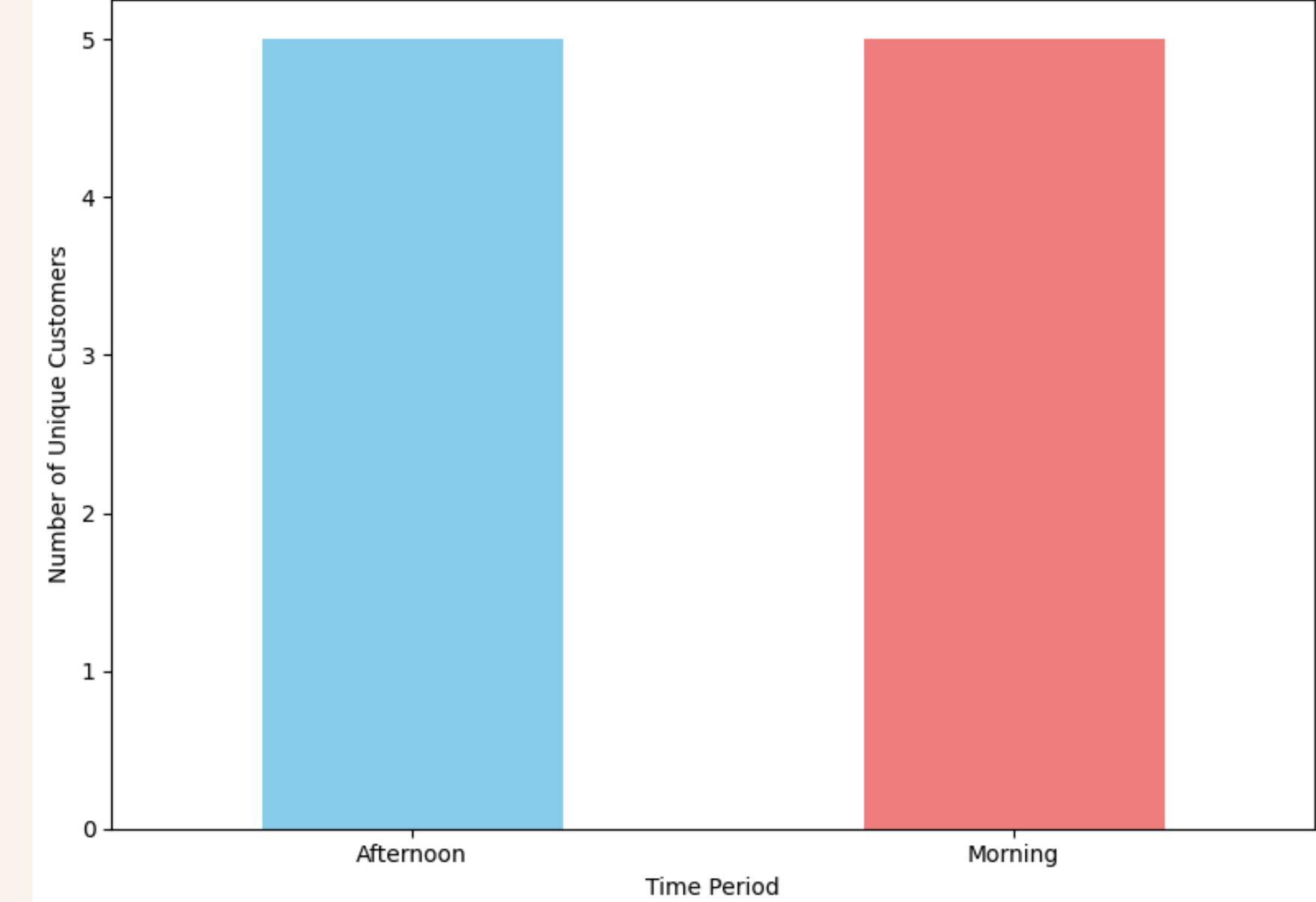


Frequency of Unique Customers by Time Period (Asia)



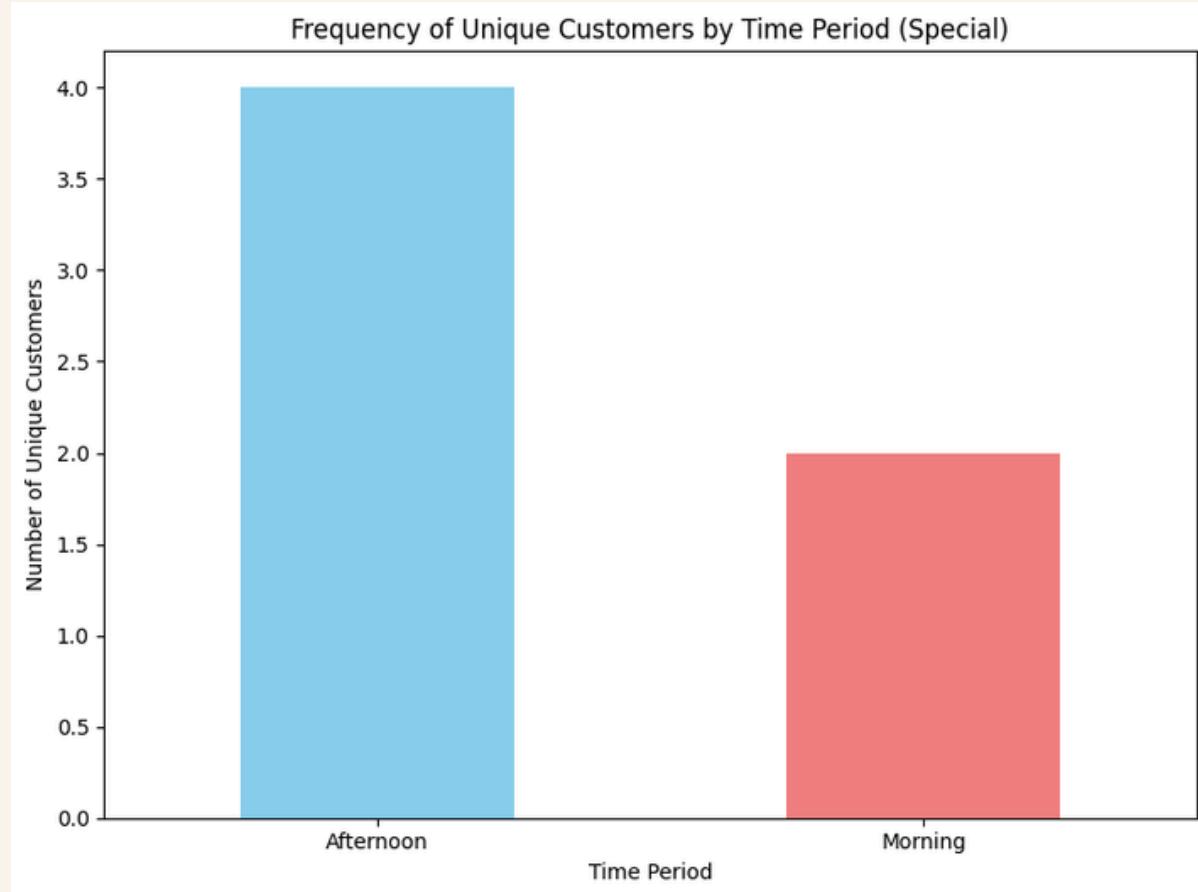
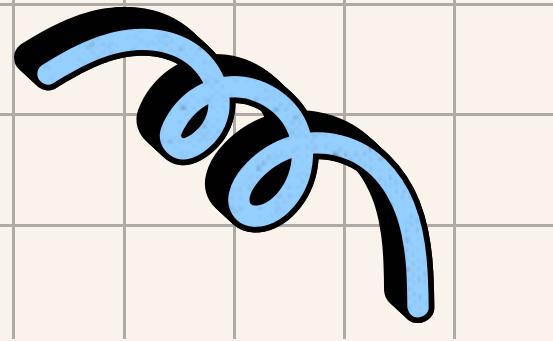
ASIA

Frequency of Unique Customers by Time Period (North America)

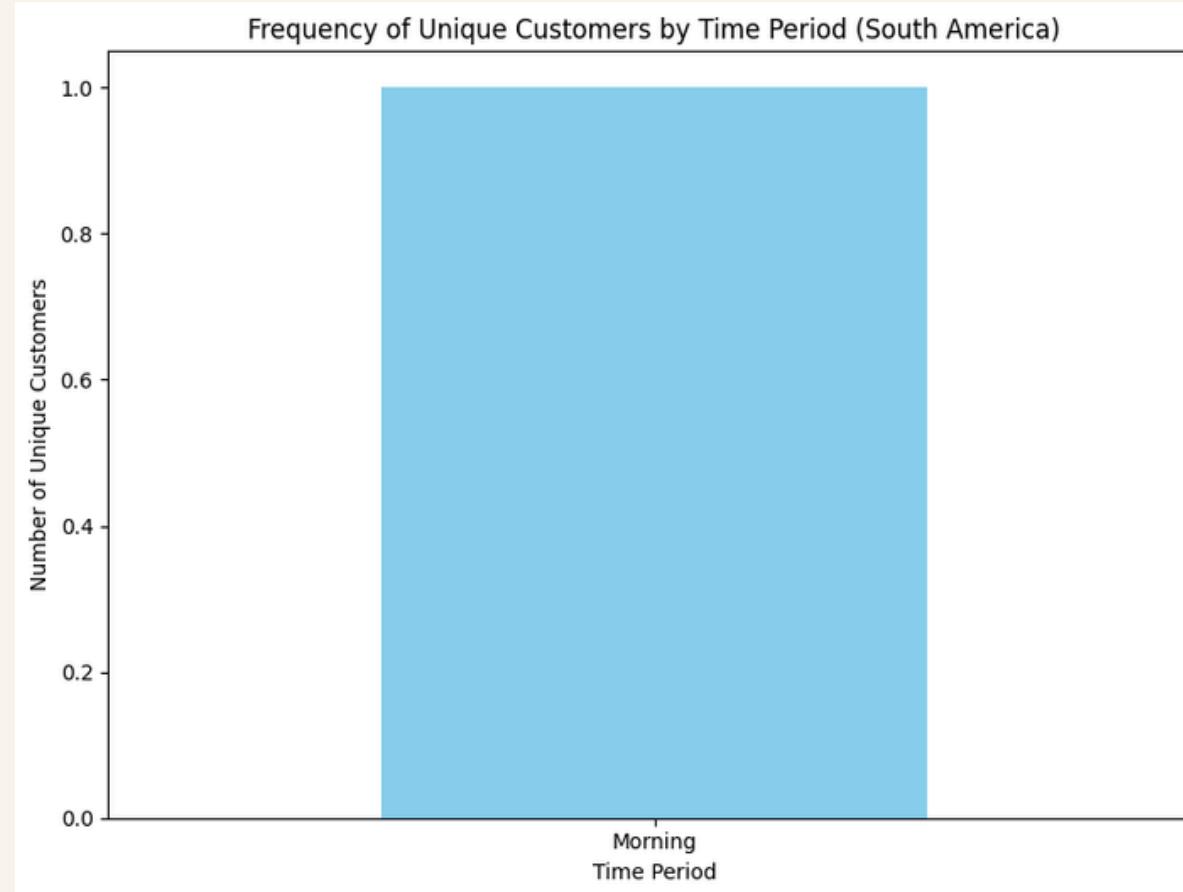


NORTH AMERICA

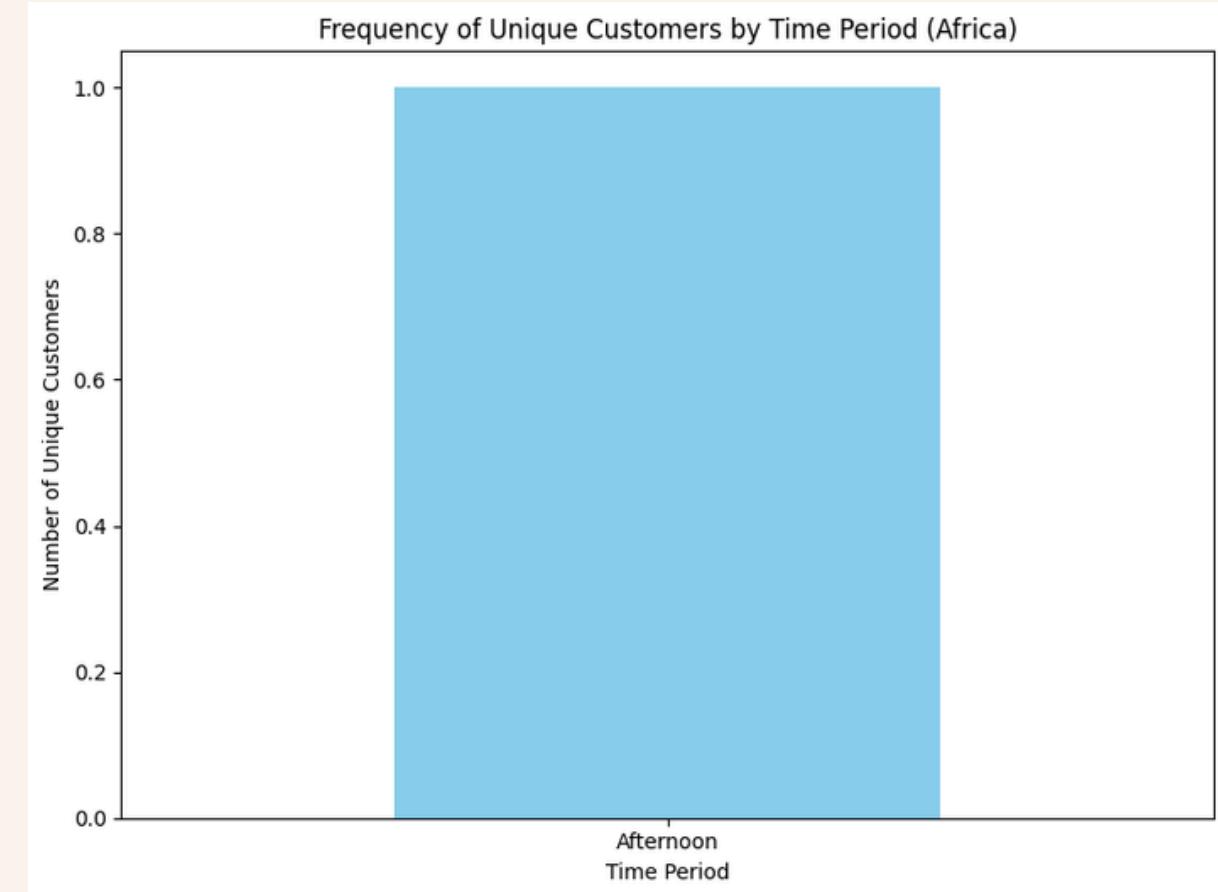
ວິເຄາະໜີ້ພົກຕິກະຮມກາຮ້ອໃນແຕ່ລະບ່ວງເວລາ



SPECIAL



SOUTH AMERICA



AFFRICA



**Q2. ASSOCIATION RULES ของการซื้อสินค้าของ
ลูกค้าในทวีปยุโรป ในแต่ละช่วงเวลา**

```

1 # prompt: แบ่งเวลาเป็น 3 ช่วง
2
3 # Function to categorize time into three periods
4 def categorize_time(time):
5     if pd.Timestamp(time).time() >= pd.Timestamp('06:00:00').time() and pd.Timestamp(time).time() < pd.Timestamp('12:00:00').time():
6         return 'Morning'
7     elif pd.Timestamp(time).time() >= pd.Timestamp('12:00:00').time() and pd.Timestamp(time).time() < pd.Timestamp('18:00:00').time():
8         return 'Afternoon'
9     else:
10        return 'Evening'
11
12 # Apply the function to create a new 'TimePeriod' column
13 rmb['TimePeriod'] = rmb['InvoiceDate'].apply(categorize_time)
14
15 rmb

```

Time	TimePeriod
08:26:00	Morning
...	...
12:50:00	Afternoon



TOP 3 ASSOCIATION RULES FOR MORNING



Confidence = 2.4000
→

← Confidence = 9.0000



Confidence = 2.4000
→

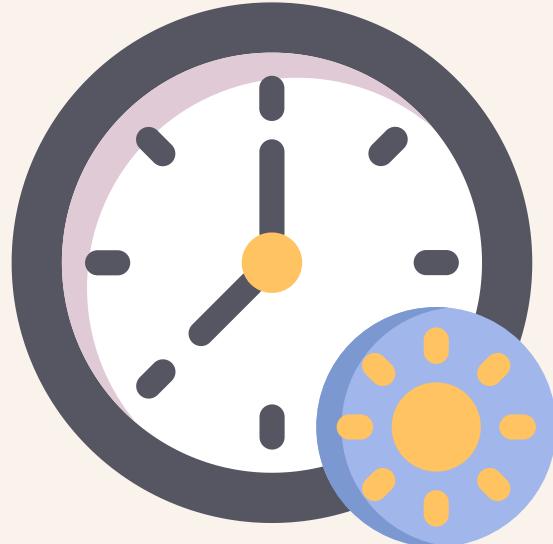
← Confidence = 9.0000



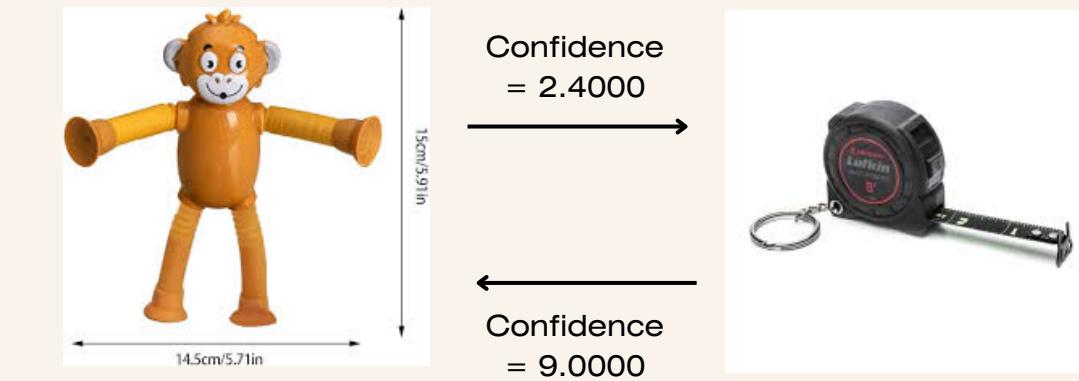
Confidence = 0.5862
→

← Confidence = 34.0000





Morning



คู่กีเด่น = “ASSORTED MONKEY SUCTION CUP HOOK' กับ 'BLACK MINI TAPE MEASURE”
“BLACK MINI TAPE MEASURE" กับ "POLYESTER FILLER PAD 30CMx30CM”
Confidence (A->B) = 2.4 และ (B->A) = 9 แสดงถึงความน่าจะเป็นที่สินค้าหนึ่งจะถูกซื้อเมื่ออีกสินค้าหนึ่งถูกซื้อ

คู่นี้มี Support ที่ค่อนข้างต่ำ (0.0137) แสดงถึงการเกิดขึ้นร่วมกันในธุรกรรมประมาณ 1.37% ของธุรกรรมทั้งหมด ในช่วงเช้า แต่มี Confidence และ Lift ที่สูงมาก โดยเฉพาะในเบื้องต้นการกระตุ้นการซื้อสินค้าที่เกี่ยวข้องกัน (Lift สูงถึง 1576.8) ซึ่งแสดงว่าเมื่อมีการซื้อสินค้าชิ้นหนึ่ง จะมีแนวโน้มที่สูงมากในการซื้อสินค้าชิ้นที่สอง เช่นกัน

ข้อเสนอแนะ

- อาจจับโอกาสในการโปรโมทสินค้าชิ้นเหล่านี้ร่วมกันในช่วงเช้า เช่น ในการจัดโปรโมชันหรือ bundle deals เพื่อกระตุ้นให้ลูกค้าซื้อสินค้าทั้งคู่
- หากต้องการปรับปรุง ควรพิจารณาเพิ่มการแนะนำสินค้าในช่วงเช้าให้ตรงกลุ่มลูกค้ามากขึ้น เช่น เพิ่มสินค้าที่เป็นอุปกรณ์หรือของใช้ในบ้านในช่วงเช้า



TOP 3 ASSOCIATION RULES FOR AFTERNOON



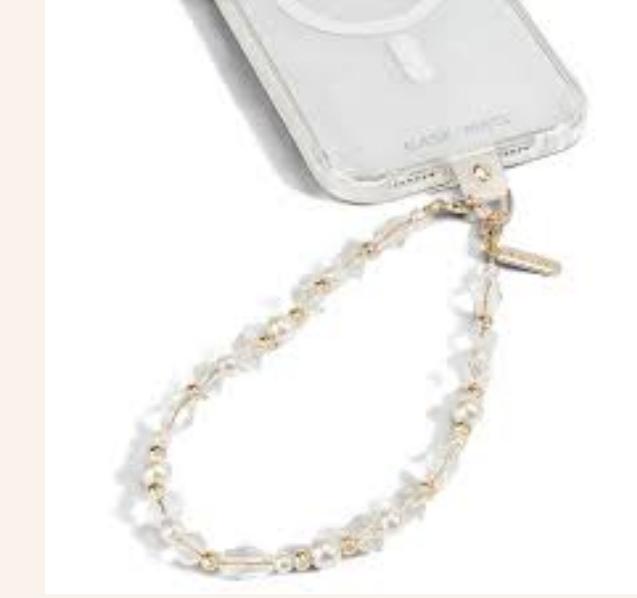
Confidence
= 14.0000



Confidence
= 14.0000



Confidence
= 3.0000



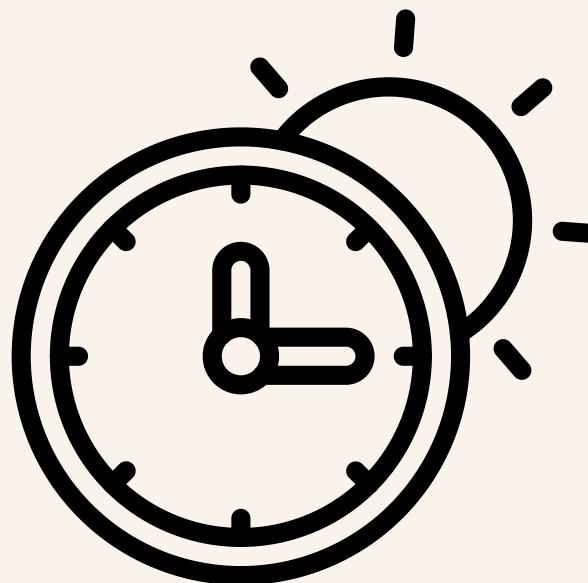
Confidence
= 14.0000



Confidence
= 14.0000



Confidence
= 3.0000



Afternoon



Confidence = 14.0000 →
← Confidence = 3.0000



คู่ที่เด่น = 'CRYSTAL CZECH CROSS PHONE CHARM' กับ 'DOTCOM POSTAGE'
Confidence (A->B) = 14.0 และ Confidence (B->A) = 3.0
แสดงถึงความน่าจะเป็นที่สินค้าหนึ่งจะถูกซื้อเมื่ออีกสินค้าหนึ่งถูกซื้อ



ช่วงบ่ายมีความสัมพันธ์ที่เด่นชัดและตรงไปตรงมา เนื่องจาก Confidence และ Lift ของคู่ 'CRYSTAL CZECH CROSS PHONE CHARM' กับ 'DOTCOM POSTAGE' สูงมากเมื่อเทียบจากกึ่งหนด โดยเฉพาะกับ 'DOTCOM POSTAGE' ที่มี Lift สูงถึง 3728 ซึ่งแสดงถึงความสัมพันธ์ที่แน่นแฟ้นในการซื้อสินค้าเหล่านี้ร่วมกัน



ข้อเสนอแนะ

- อาจจำพิจารณาเพิ่มสินค้าประเภทการจัดส่งหรือบริการที่เกี่ยวข้องในช่วงบ่าย เพื่อเสริมความสะดวกในการซื้อสินค้า เช่น การแนะนำบริการจัดส่งที่รวดเร็ว
- การสร้างໂປຣໂນชັນທີ່ເກື່ອງຂ້ອງກັບ 'DOTCOM POSTAGE' 亦或 'CRUK Commission' ในช่วงบ่ายอาจ เป็นໂວກສາທີ່ດີໃນກະຕຸນການຊົ້ວ



TOP 3 ASSOCIATION RULES FOR EVENING



Confidence
= 2.0000



Confidence
= 2.0000



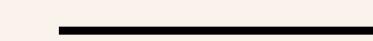
Confidence
= 2.0000



Confidence
= 3.0000

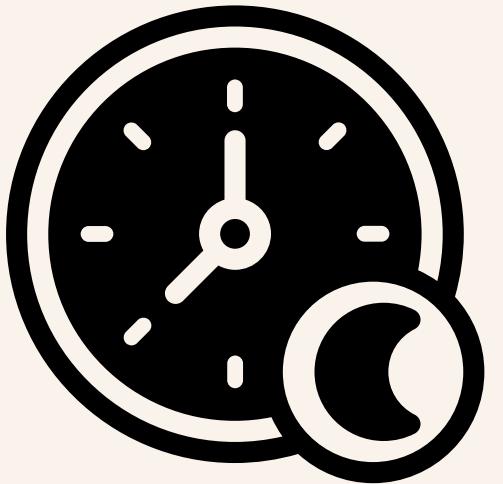


Confidence
= 2.0000



Confidence
= 2.0000





Evening



Confidence = 2.0000
←
Confidence = 2.0000



คู่ที่เด่น = 'ROLL WRAP 50'S RED CHRISTMAS' กับ 'TOOL BOX SOFT TOY'

Confidence (A->B) = 2.0 และ Confidence (B->A) = 2.0 แสดงถึงความน่าจะเป็นที่สินค้าหนึ่งจะถูกซื้อเมื่ออีกสิบค้าหนึ่งถูกซื้อ



ช่วงเย็นพบว่ามีสินค้าที่เกี่ยวข้องกับของขวัญและของเล่น (เช่น 'TOOL BOX SOFT TOY') ปรากฏอยู่ในหลายๆ คู่ โดยเฉพาะการมี Support ที่ค่อนข้างสูงในบางคู่ (เช่น 0.0167) แต่ Lift ค่อนข้างต่ำ (359) ซึ่งแสดงถึงความสัมพันธ์ที่ไม่แข็งแกร่งเท่าช่วงอื่นๆ



ข้อเสนอแนะ



- ควรเพิ่มการโปรโมทสินค้าในกลุ่มของขวัญหรือของเล่นในช่วงเย็นให้มากขึ้น โดยอาจเพิ่มสินค้าที่เป็นของขวัญในช่วงเทศกาลหรือลดราคาสินค้าที่มีความสัมพันธ์สูง
- อาจจับการจัดกลุ่มสินค้าหรือการวางแผนโปรโมชันในช่วงเย็นให้มีความหลากหลายและเน้นไปที่ของขวัญเพื่อเพิ่มการซื้อ



Thank You