РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ

Факультет физико-математических и естественных наук

Кафедра информационных технологий

ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ № 2

Дисциплина: Информационный анализ данных

Студент: Ильин Никита

Группа: НФИбд-01-19

Москва 2022

Вариант №4

Алгоритм: Apriori

День недели (поле order_dow таблицы orders): “2”

Код департамента (поле department_id таблицы products): “2”

Показатель оценки ассоциативных правил: лифт (lift)

Задание

Для закрепленного за Вами варианта лабораторной работы:

1. При помощи модуля sqlite3 откройте базу данных Instacart в файле instacart.db.
2. Загрузите таблицы `departments` и `products` в датафреймы Pandas. При помощи запроса SELECT извлеките из таблицы `order_products__train` записи, соответствующие указанным в индивидуальном задании дню недели (поле `order_dow` таблицы `orders`) и коду департамента (поле `department_id` таблицы `products`) и загрузите в датафрейм Pandas. Определите количество строк в полученном датафрейме и определите количество товаров (стоолбец `product_id`) в транзакциях датафрейма.
3. Определите пять наиболее популярных товаров в датафрейме транзакций и определите количество покупок (транзакций) этих товаров.
4. Постройте транзакционную базу данных из полученного датафрейма, используя в качестве идентификатора транзакции столбец `order_id`, а в качестве названий товаров - поле `product_name` из датафрейма для таблицы `products`, соответствующее столбцу `product_id`. Найдите в транзакционной базе данных три транзакции с наибольшим количеством товаров и выведите их на экран.

5. Постройте по транзакционной базе данных бинарную базу данных в формате датафрейма пакета `mlxtend`. По бинарной базе данных определите пять наиболее популярных товаров и определите количество покупок (транзакций) этих товаров.

6. При помощи указанного в индивидуальном задании метода построения популярных наборов предметов постройте популярный набор предметов с минимальной поддержкой не менее 3, имеющий максимальную длину. При отсутствии таких наборов уменьшите поддержку до 2. В случае нехватки вычислительных ресурсов (слишком долгой работы программы) при построении популярных наборов предметов сокращайте число записей в наборе данных (например, делая выборку половины записей набора).

7. Используя пакет `mlxtend` или реализацию на Python, постройте набор ассоциативных правил для полученного популярного наборов предметов. Используйте уровень достоверности (confidence), равный 0.6.

8. Для построенного набора ассоциативных правил вычислите показатель (меру) оценки ассоциативных правил, указанную в индивидуальном задании, и определите ассоциативные правила с наилучшим значением показателя оценки.

1. При помощи модуля sqlite3 откройте базу данных Instacart в файле instacart.db.

In [1]:
```
import numpy as np
import pandas as pd
import sqlite3

conn = sqlite3.connect('instacart.db')
type(conn)
```

Out[1]:
```
sqlite3.Connection
```

In [2]:
```
cursor = conn.cursor()

type(cursor)
```

Out[2]:
```
sqlite3.Cursor
```

1. Загрузите таблицы `departments` и `products` в датафреймы Pandas. При помощи запроса SELECT извлеките из таблицы `order_products__train` записи, соответствующие указанным в индивидуальном задании дню недели (поле `order_dow` таблицы `orders`) и коду департамента (поле `department_id` таблицы `products`) и загрузите в датафрейм Pandas. Определите количество строк в полученном датафрейме и определите количество товаров (стоолбец `product_id`) в транзакциях датафрейма.

In [3]:
```
pd.read_sql_query("SELECT * FROM departments", conn).tail()
```

Out[3]:

	department_id	department
16	17	household
17	18	babies
18	19	snacks
19	20	deli
20	21	missing

In [4]:
```
pd.read_sql_query("SELECT * FROM products", conn).tail()
```

Out[4]:

	product_id	product_name	aisle_id	department_id
49683	49684	Vodka, Triple Distilled, Twist of Vanilla	124	5
49684	49685	En Croute Roast Hazelnut Cranberry	42	1
49685	49686	Artisan Baguette	112	3
49686	49687	Smartblend Healthy Metabolism Dry Cat Food	41	8
49687	49688	Fresh Foaming Cleanser	73	11

In [5]:
```
order_products__train = pd.read_sql_query(
    """SELECT opt.order_id, opt.product_id, add_to_cart_order, reordered, product_nam
        FROM order_products__train as opt, orders as ord, products as pr
        WHERE ord.order_dow=2 AND pr.department_id=2
        AND ord.order_id = opt.order_id AND pr.product_id = opt.product_id""", conn)
order_products__train.head()
```

Out[5]:

	order_id	product_id	add_to_cart_order	reordered	product_name
0	2316178	32115	2	1	93/7 Ground Beef
1	988	4818	4	1	Classic Vanilla Coffee Creamer
2	245330	11537	5	0	Anti Bug Shake & Spray
3	3398797	26756	4	0	Light CocoWhip! Coconut Whipped Topping
4	1673226	5435	9	0	Melatonin 1 Mg Peppermint Sublingual Tablets

In [6]:
```
len(order_products__train)
```

Out[6]:
```
202
```

In [7]:
```
len(order_products__train['product_id'].unique())
```

Out[7]:
```
121
```

1. Определите пять наиболее популярных товаров в датафрейме транзакций и определите количество покупок (транзакций) этих товаров.

In [8]:
```
order_products__train['product_name'].value_counts()[:5]
```

Out[8]:
```
Roasted Almond Butter                    14
Light CocoWhip! Coconut Whipped Topping   8
Roasted Unsalted Almonds                  6
Pierogi Potato & Cheese                   5
Organic Breakfast Blend Coffee            5
Name: product_name, dtype: int64
```

1. Постройте транзакционную базу данных из полученного датафрейма, используя в качестве идентификатора транзакции столбец `order_id`, а в качестве названий товаров - поле `product_name` из датафрейма для таблицы `products`, соответствующее столбцу `product_id`. Найдите в транзакционной базе данных три транзакции с наибольшим количеством товаров и выведите их на экран.

In [9]:
```
import itertools
dataset = order_products__train.groupby('order_id')['product_name'].apply(list).to_di
dataset
```

Out[9]:

```
{'1002376': ['Natural Calm Anti-Stress Drink'],
 '1002771': ['Plain Salt'],
 '1010544': ['Candle Lighter'],
 '1026937': ['Organic Turmeric Powder'],
 '1032117': ['Tahini Sesame Paste'],
 '1033632': ['Roasted Unsalted Almonds'],
 '103529': ['Roasted Almond Butter'],
 '10362': ['Coconut Flour'],
 '1041996': ['California White Zinfandel'],
 '1045720': ['Carob Spirulina Energy Chunks'],
 '1052847': ['Butt Paste Original Diaper Rash Ointment',
  'Baby Healing Ointment'],
 '1073137': ['Cherry Vanilla Granola'],
 '1082545': ['Pierogi Potato & Cheese'],
 '1111176': ['Recolte Wine'],
 '1115459': ['Healing Baby Diaper Rash Cream'],
 '1130967': ['Roasted Almond Butter'],
 '1137096': ['Roasted Almond Butter'],
 '1162810': ['Coffee Mate French Vanilla Creamer Packets'],
 '1164706': ['Liquid Teething Relief'],
 '1183347': ['Bold & Spicy Bloody Mary Mix'],
 '1202295': ['Baby Healing Ointment'],
 '1209507': ['93/7 Ground Beef'],
 '1247960': ['Black Chia Seeds'],
 '1273758': ['Light CocoWhip! Coconut Whipped Topping'],
 '1275786': ['Creamer'],
 '1287961': ['Coconut Flour'],
 '1290501': ['Roasted Unsalted Almonds'],
 '1295930': ['Max AAA Batteries'],
 '1319144': ['Ultimate Intensive Healing Hand Cream'],
 '1323388': ['Maximum Absorbency L for Women  Underwear'],
 '1356205': ['Light CocoWhip! Coconut Whipped Topping'],
 '1374945': ['Liquid Melatonin Natural Black Cherry Flavor'],
 '1384735': ['Tropical Fruit Electrolyte Solution'],
 '1400434': ['Tranquil Rose Massage Oil'],
 '1454666': ['Baby Tummy Gripe Water Dietary Supplement'],
 '1456960': ['Early Result Pregnancy Test'],
 '1461479': ['Max AAA Batteries'],
 '1467249': ['Sanitizing Wipes'],
 '1467260': ['Queso Fresco Cheese'],
 '1480477': ['Roasted Salted Pistachios'],
 '1492530': ["Children's Chestal Homeopathic Medicine"],
 '1516077': ['Organic Breakfast Blend Coffee'],
 '1524554': ['Greek Style Honey Yogurt'],
 '1563902': ['Pierogi Potato & Cheese'],
 '1618177': ['Kiwifruit'],
 '1626568': ['Recolte Wine'],
 '1649353': ['Pierogi Potato & Cheese'],
 '1673226': ['Melatonin 1 Mg Peppermint Sublingual Tablets'],
 '1688254': ['Condoms, Premium Latex, Ultra Thin, Premium Lubricant'],
 '1710895': ['Light CocoWhip! Coconut Whipped Topping'],
 '1724499': ['Melatonin, 3 mg, Tablets'],
 '1749088': ['Camilia, Single Liquid Doses'],
 '1767154': ['Yogurt Covered Almonds'],
 '1797451': ['Ibuprofen Drops for ages 6-23 Months, White Grape Flavor'],
 '1847238': ['Wellness Herbal Kids Tincture'],
 '1852694': ['Original Bloody Mary Mix'],
 '1864946': ['Oral Electrolyte Powder Assorted Flavors'],
 '1887216': ['Digital Pregnancy Test'],
 '1935022': ['Whispering Angel Rosé'],
 '1949309': ['Roasted Almond Butter'],
 '200538': ['Bloody Mary Mix'],
 '2007234': ['Vitamin Code Kids Cherry Berry Chewables'],
 '2024912': ['Pierogi Potato & Cheese'],
 '2043595': ['Sunset Blush'],
 '2063517': ['Raw Pistachios'],
```

```
 '2091412': ['Multivitamin, Kids Complete, Gummies'],
 '2099956': ['Roasted Almond Butter'],
 '2101195': ['Roasted Unsalted Almonds'],
 '2111359': ['Cuticle Rehab', 'Max AAA Batteries'],
 '2120211': ['Sunflower Seeds'],
 '216859': ['Roasted Almond Butter'],
 '220023': ['Margarita Salt', 'The Original Margarita Mix'],
 '2243739': ['Bold & Spicy Bloody Mary Mix'],
 '2316178': ['93/7 Ground Beef'],
 '2332783': ['Pleasure Pack Lubricated Premium Latex Condoms'],
 '2337649': ['White Zinfandel'],
 '2356929': ['Baby Healing Ointment'],
 '2403964': ['Early Pregnancy Test'],
 '2407797': ['Roasted Almond Butter'],
 '2413407': ['Vitamin Code Kids Chewable Whole Food Multivitamin For Kids Cherry Berr
y'],
 '2427526': ['Roasted Almond Butter'],
 '2435080': ['Spicy Salmon Roll'],
 '2437216': ['Italian Sweet Creme Creamer'],
 '244531': ['Super Colon Cleanse Capsules Psyllum Supplement With Herbs'],
 '2450236': ['Walnuts'],
 '245330': ['Anti Bug Shake & Spray'],
 '2515929': ['BabyRub® Soothing Ointment'],
 '2544357': ['Early Result Pregnancy Test'],
 '2550786': ['Kiwifruit'],
 '2553199': ['Light CocoWhip! Coconut Whipped Topping'],
 '2553216': ['Rainbow Roll', 'Italian Sweet Creme Creamer'],
 '2594544': ['White Zinfandel'],
 '2626769': ['Ibuprofen Drops for ages 6-23 Months, White Grape Flavor'],
 '2628044': ['Rainbow Roll'],
 '2654206': ['5-HTP 100 Mg Vegetarian Capsules'],
 '2659807': ['Light CocoWhip! Coconut Whipped Topping'],
 '2664766': ['Original Liquid Coffee Creamer'],
 '2737380': ['Kiwifruit'],
 '275764': ['Roasted Almond Butter'],
 '2773952': ['Light CocoWhip! Coconut Whipped Topping'],
 '2776395': ['Organic Flax Seed', 'Black Chia Seeds'],
 '278561': ["Pears D'Anjou Kid Size Fruit"],
 '2803718': ['Magnum Thin Lubricated Condoms'],
 '2871756': ['Organic Tea Tree Oil'],
 '2873625': ['93/7 Ground Beef'],
 '2882563': ['Lavender & Chamomile Hand Soap'],
 '2893224': ['Raw Pistachios'],
 '2896656': ['Roasted Salted Pistachios'],
 '2901574': ['Salted Mixed Nuts'],
 '2903292': ['Kitchen Bug Killer 2 Botanical'],
 '2916357': ['Sweet & Sour Mix'],
 '2929036': ['Pinot Noir Rose'],
 '2931409': ['Roasted Almond Butter'],
 '2969967': ['California Blush Pink Champagne'],
 '2971144': ['Fruit Fly Trap'],
 '2988402': ['Pork Back Ribs'],
 '299137': ['Rapid Relief Creamy Diaper Rash Ointment'],
 '2992147': ['Organic Breakfast Blend Coffee'],
 '2997017': ['Black Chia Seeds'],
 '3004791': ['Bloody Mary Mix'],
 '3010023': ['Cherry Vanilla Granola'],
 '3031288': ['Organic Flax Seed'],
 '3074326': ['Sweets Organic Lollipops'],
 '3086823': ['Coffee Mate French Vanilla Creamer Packets'],
 '3093037': ['Queso Fresco Cheese'],
 '312417': ['Organic Supreme Fruit & Nut Mix'],
 '3142996': ['Calms Forte Sleep Aid Tablets - 100 CT'],
 '3143478': ['Plus Pregnancy Test'],
 '3150761': ['Cinnamon Vanilla Creme Liquid Coffee Creamer'],
 '3150842': ['Healing Ointment Advanced Therapy'],
```

```
'3153312': ['Paleo Magazine'],
'3158931': ['Beef Chuck Roast'],
'3167456': ['Ultimate Intensive Healing Hand Cream'],
'317060': ['Light CocoWhip! Coconut Whipped Topping'],
'3179786': ['Light Classic Lime Margarita Mix'],
'319618': ['All Purpose Precision Tip 2 Pack'],
'3246632': ['Baby Powder'],
'3253268': ['Roasted Unsalted Almonds'],
'3256008': ['Carob Spirulina Energy Chunks', 'Cherry Vanilla Granola'],
'3257464': ['Deluxe Nut Mix'],
'327243': ['Original Diaper Rash Ointment & Skin Protectant'],
'3294302': ['Maximum Absorbency XL for Women Underwear'],
'3298705': ['Margarita Salt', 'Light Margarita'],
'3319149': ['Baby Tummy Gripe Water Dietary Supplement'],
'3340032': ['Organic Breakfast Blend Coffee'],
'3348853': ['Cherry Vanilla Granola'],
'3364755': ['The Original Margarita Mix'],
'3367258': ['Facial Tissues with Lotion'],
'3377673': ['Pierogi Potato & Cheese'],
'3382706': ['Emergency Contraceptive'],
'3398797': ['Light CocoWhip! Coconut Whipped Topping'],
'3398895': ['Recolte Wine'],
'3420021': ['Beef Chuck Roast'],
'361824': ['Recolte Wine'],
'375119': ['Maximum Strength Original Paste Diaper Rash Ointment'],
'379665': ['California Madeira'],
'399599': ['Early Result Pregnancy Test'],
'40123': ['Aromatic Bitters'],
'429260': ['Tamari Roasted Pumpkin Seeds'],
'43800': ['Organic Raw Pumpkin Seeds'],
'450908': ['Coffee Mate French Vanilla Creamer Packets'],
'481731': ['Roasted Almond Butter'],
'513619': ['Homeopathic Grape Flavor Baby Gas Drops'],
'521201': ['Facial Tissues with Lotion'],
'540696': ['Classic Vanilla Coffee Creamer'],
'551826': ['Sirloin Tip Roast'],
'554373': ['Milk Chocolate Coconut Bar'],
'563830': ['Beef Flank Steak'],
'599733': ["Children's Grape 24-Hour"],
'629046': ['Max AAA Batteries'],
'644625': ['Nighttime Sleep Aid Caplets'],
'655420': ['Camilia, Single Liquid Doses'],
'688464': ['Baby Healing Ointment'],
'711287': ['Roasted Almond Butter'],
'718849': ['Multivitamin, Kids Complete, Gummies'],
'727578': ['Kiwifruit'],
'73469': ['Roasted Almond Butter'],
'737059': ['Organic Breakfast Blend Coffee'],
'745812': ['Butt Paste Original Diaper Rash Ointment'],
'812903': ['Creamer'],
'81475': ['Roasted Unsalted Almonds'],
'832956': ['Coconut Flour'],
'833029': ['Roasted Almond Butter'],
'85654': ['Creamer'],
'858605': ['Strike On Box Matches'],
'869026': ['Facial Mask Age Defying Hydro Serum'],
'875423': ['Polenta'],
'906289': ['Roasted Unsalted Almonds'],
'908561': ['Organic Breakfast Blend Coffee'],
'915206': ['Rapid Relief Zinc Oxide Diaper Rash Cream'],
'950090': ['Rescue Remedy'],
'954363': ['Organic Tamari Almonds'],
'970849': ['Coconut Almond Granola'],
'98476': ['SleepGels Nighttime Sleep Aid'],
'988': ['Classic Vanilla Coffee Creamer']}
```

In [10]:
```
order_products__train[['order_id', 'product_name']].groupby('order_id')['product_name
```

Out[10]:
```
order_id
2776395    2
220023     2
3298705    2
2111359    2
3256008    2
          ..
2101195    1
2120211    1
216859     1
2243739    1
988        1
Name: product_name, Length: 195, dtype: int64
```

In [11]:
```
dataset.get('2776395')
dataset.get('220023')
dataset.get('3298705')
```

Out[11]:
```
['Margarita Salt', 'Light Margarita']
```

1. Постройте по транзакционной базе данных бинарную базу данных в формате датафрейма пакета `mlxtend`. По бинарной базе данных определите пять наиболее популярных товаров и определите количество покупок (транзакций) этих товаров.

In [12]:
```
from mlxtend.preprocessing import TransactionEncoder

dataset1 = []

for i in dataset.values():
    dataset1.append(i)

keyss = []

for i in dataset.keys():
    keyss.append(i)

te = TransactionEncoder()
dataset_bin = te.fit(dataset1).transform(dataset1)
```

In [13]:
```
df = pd.DataFrame(dataset_bin, columns=te.columns_, index=keyss)
df.head(5)
```

Out[13]:

	5-HTP 100 Mg Vegetarian Capsules	93/7 Ground Beef	All Purpose Precision Tip 2 Pack	Anti Bug Shake & Spray	Aromatic Bitters	Baby Healing Ointment	Baby Powder	Baby Tummy Gripe Water Dietary Supplement	BabyRub® Soothing Ointment
1002376	False	False	False	False	False	False	False	False	False
1002771	False	False	False	False	False	False	False	False	False
1010544	False	False	False	False	False	False	False	False	False
1026937	False	False	False	False	False	False	False	False	False
1032117	False	False	False	False	False	False	False	False	False

5 rows × 121 columns

In [14]:
```
asd= []
for i in df.columns:
    asd.append((df[i]==True).sum())
pd.Series(asd,index=df.columns).sort_values(ascending=False)[:5]
```

Out[14]:
```
Roasted Almond Butter                        14
Light CocoWhip! Coconut Whipped Topping       8
Roasted Unsalted Almonds                      6
Pierogi Potato & Cheese                       5
Organic Breakfast Blend Coffee                5
dtype: int64
```

1. При помощи указанного в индивидуальном задании метода построения популярных наборов предметов постройте популярный набор предметов с минимальной поддержкой не менее 3, имеющий максимальную длину. При отсутствии таких наборов уменьшите поддержку до 2. В случае нехватки вычислительных ресурсов (слишком долгой работы программы) при построении популярных наборов предметов сокращайте число записей в наборе данных (например, делая выборку половины записей набора).

In [15]:
```
from mlxtend.frequent_patterns import apriori

apr = apriori(df, min_support=1/df.shape[0],use_colnames=True)
apr
```

Out[15]:

	support	itemsets
0	0.005128	(5-HTP 100 Mg Vegetarian Capsules)
1	0.015385	(93/7 Ground Beef)
2	0.005128	(All Purpose Precision Tip 2 Pack)
3	0.005128	(Anti Bug Shake & Spray)
4	0.005128	(Aromatic Bitters)
...	...	...
123	0.005128	(Cherry Vanilla Granola, Carob Spirulina Energ...
124	0.005128	(Max AAA Batteries, Cuticle Rehab)
125	0.005128	(Italian Sweet Creme Creamer, Rainbow Roll)
126	0.005128	(Light Margarita, Margarita Salt)
127	0.005128	(The Original Margarita Mix, Margarita Salt)

128 rows × 2 columns

1. Используя пакет `mlxtend` или реализацию на Python, постройте набор ассоциативных правил для полученного популярного наборов предметов. Используйте уровень достоверности (confidence), равный 0.6.

In [16]:
```
from mlxtend.frequent_patterns import association_rules
association_rules(apr, metric="confidence", min_threshold=0.6)
```

Out[16]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	convict
0	(Cuticle Rehab)	(Max AAA Batteries)	0.005128	0.020513	0.005128	1.0	48.75	0.005023	
1	(Light Margarita)	(Margarita Salt)	0.005128	0.010256	0.005128	1.0	97.50	0.005076	

1. Для построенного набора ассоциативных правил вычислите показатель (меру) оценки ассоциативных правил, указанную в индивидуальном задании, и определите ассоциативные правила с наилучшим значением показателя оценки.

In [18]:
```
association_rules(apr, metric="confidence", min_threshold=0.6).sort_values('lift', as
```

Out[18]:

	antecedents	consequents	lift
1	(Light Margarita)	(Margarita Salt)	97.50
0	(Cuticle Rehab)	(Max AAA Batteries)	48.75