МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЕТ

по лабораторной работе №4

по дисциплине «Ассоциативный анализ»

Тема: Машинное обучение

Студент гр. 6304 _____ Виноградов К.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2020

Загрузка данных.

Загрузим данные из csv таблицы. Результат представлен на рис.1.

	Item(s)	Item 1	Item 2	...	Item 30	Item 31	Item 32
0	4	citrus fruit	semi-finished bread	...	NaN	NaN	NaN
1	3	tropical fruit	yogurt	...	NaN	NaN	NaN
2	1	whole milk	NaN	...	NaN	NaN	NaN
3	4	pip fruit	yogurt	...	NaN	NaN	NaN
4	4	other vegetables	whole milk	...	NaN	NaN	NaN
...	...	...	...	...	...	...	...
9830	17	sausage	chicken	...	NaN	NaN	NaN
9831	1	cooking chocolate	NaN	...	NaN	NaN	NaN
9832	10	chicken	citrus fruit	...	NaN	NaN	NaN
9833	4	semi-finished bread	bottled water	...	NaN	NaN	NaN
9834	5	chicken	tropical fruit	...	NaN	NaN	NaN

[9835 rows x 33 columns]

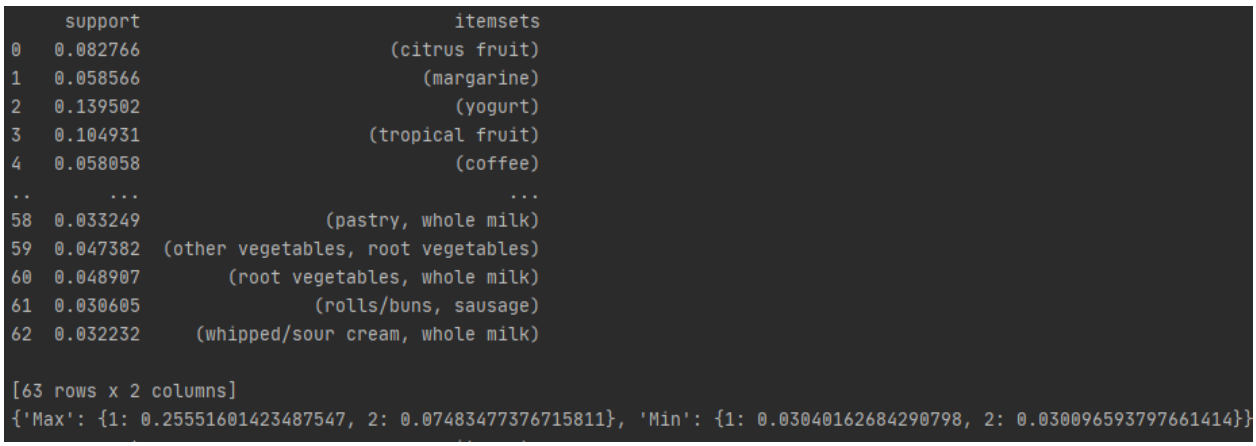
Рисунок 1 – Данные с NaN

Найдем список товаров и их количество. Результаты на рис. 2.

Рисунок 2 – Уникальные товары

FPGrowth и FPMax.

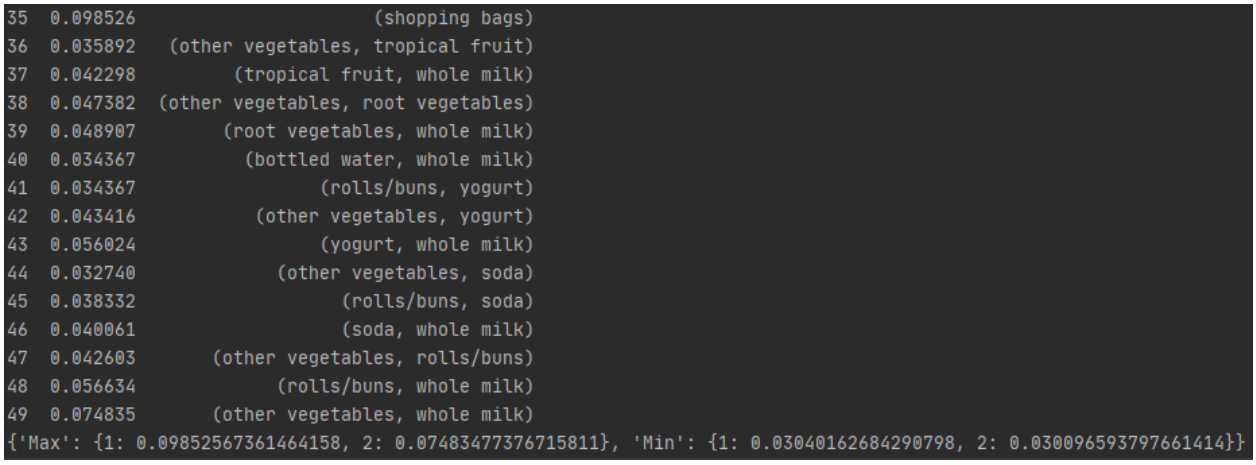
Проведем ассоциативный анализ сначала с помощью алгоритма FPGrowth, а затем с помощью алгоритма FPMax. Также найдем минимальные и максимальные значения поддержки для набора каждого уровня. Результаты представлены на рис. 3 и 4.



```
     support                           itemsets
0   0.082766                      (citrus fruit)
1   0.058566                         (margarine)
2   0.139502                            (yogurt)
3   0.104931                     (tropical fruit)
4   0.058058                            (coffee)
..       ...                                 ...
58  0.033249                 (pastry, whole milk)
59  0.047382  (other vegetables, root vegetables)
60  0.048907       (root vegetables, whole milk)
61  0.030605                (rolls/buns, sausage)
62  0.032232     (whipped/sour cream, whole milk)

[63 rows x 2 columns]
{'Max': {1: 0.25551601423487547, 2: 0.07483477376715811}, 'Min': {1: 0.03040162684290798, 2: 0.030096593797661414}}
```

Рисунок 3 – Результат применения FPGrowth



```
35  0.098526                      (shopping bags)
36  0.035892   (other vegetables, tropical fruit)
37  0.042298         (tropical fruit, whole milk)
38  0.047382  (other vegetables, root vegetables)
39  0.048907        (root vegetables, whole milk)
40  0.034367          (bottled water, whole milk)
41  0.034367                (rolls/buns, yogurt)
42  0.043416           (other vegetables, yogurt)
43  0.056024                (yogurt, whole milk)
44  0.032740             (other vegetables, soda)
45  0.038332                   (rolls/buns, soda)
46  0.040061                   (soda, whole milk)
47  0.042603       (other vegetables, rolls/buns)
48  0.056634             (rolls/buns, whole milk)
49  0.074835       (other vegetables, whole milk)
{'Max': {1: 0.09852567361464158, 2: 0.07483477376715811}, 'Min': {1: 0.03040162684290798, 2: 0.030096593797661414}}
```

Рисунок 4 – Результат применения FPMax

Можно отметить что FPMax отбирает максимальные наборы удовлетворяющие уровню поддержки и не отображает их поднаборы, в отличие от FPGrowth, который отображает все наборыудовлетворяющие уровню поддержки. Поэтому у результатов при наличие одинакового количества уровней могут отличаться значения максимальной поддержки по уровню так как чем набор меньше тем больше значение поддержки, при том минимальные значения поддержки всегда будут одинаковыми.

Построим гистограммы для первых 10 товаров по уровню поддержки. Результаты представлены на рис. 5.

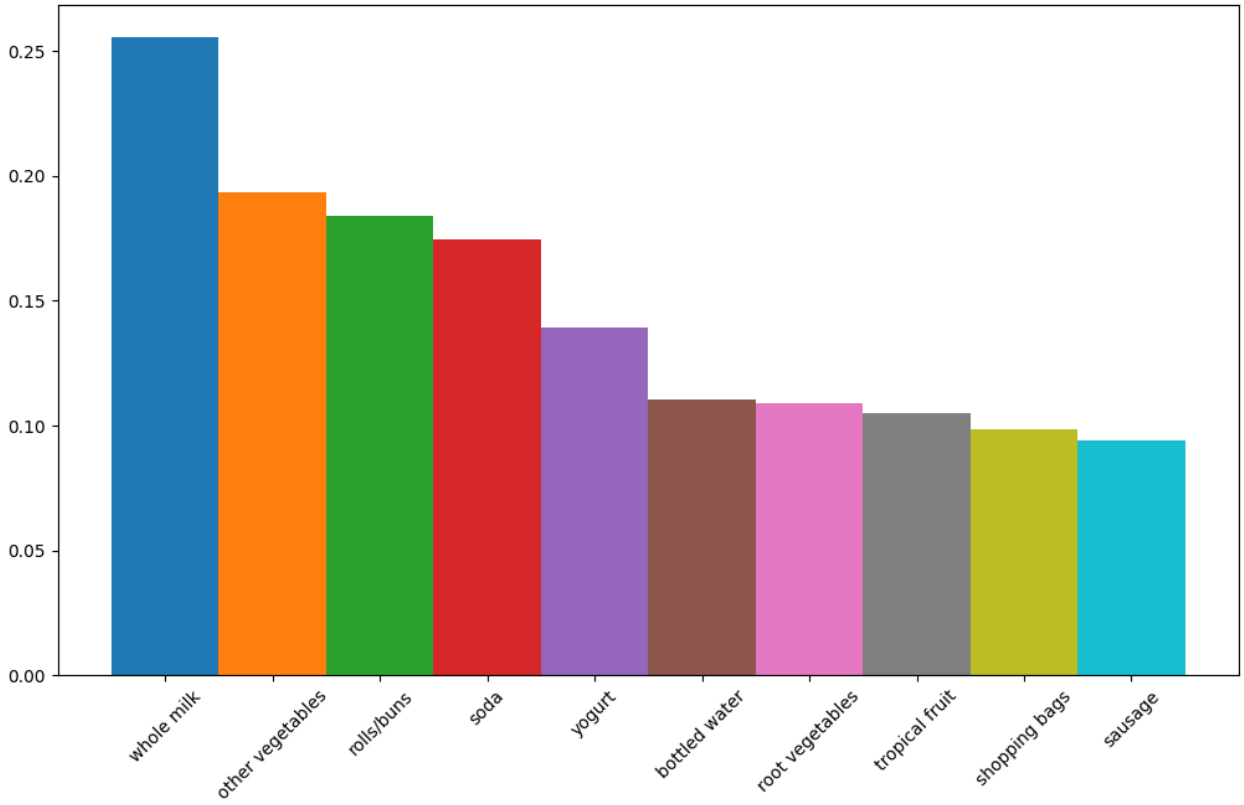


Рисунок 5 – Гистограммы самых частых товаров

Можно заметить что частота самого частого товара соответствует максимальной частоте товара уровня 1 при анализе с помощью алгоритма FPGrowth.

Отберем только транзакции с определенными продуктами и проведем анализ с помощью алгоритма FPGrowth, а затем с помощью алгоритма FPMax. Результаты представлены на рис. 6 и 7.

```
19  0.043416           (yogurt, other vegetables)
20  0.035892   (tropical fruit, other vegetables)
21  0.042298         (tropical fruit, whole milk)
22  0.074835       (whole milk, other vegetables)
23  0.042603       (rolls/buns, other vegetables)
24  0.056634             (whole milk, rolls/buns)
25  0.034367          (whole milk, bottled water)
26  0.038332                   (rolls/buns, soda)
27  0.040061                   (whole milk, soda)
28  0.032740            (soda, other vegetables)
29  0.033249                (whole milk, pastry)
30  0.047382  (root vegetables, other vegetables)
31  0.048907        (whole milk, root vegetables)
32  0.030605                (sausage, rolls/buns)
33  0.032232     (whipped/sour cream, whole milk)
{'Max': {1: 0.25551601423487547, 2: 0.07483477376715811}, 'Min': {1: 0.05765124555160142, 2: 0.030503304524656837}}
```

Рисунок 6 – Результат применения FPGrowth к ограниченному набору

```
7   0.098526                     (shopping bags)
8   0.035892   (tropical fruit, other vegetables)
9   0.042298         (tropical fruit, whole milk)
10  0.047382  (root vegetables, other vegetables)
11  0.048907        (whole milk, root vegetables)
12  0.034367          (whole milk, bottled water)
13  0.034367                (yogurt, rolls/buns)
14  0.043416           (yogurt, other vegetables)
15  0.056024                (whole milk, yogurt)
16  0.032740            (soda, other vegetables)
17  0.038332                   (rolls/buns, soda)
18  0.040061                   (whole milk, soda)
19  0.042603       (rolls/buns, other vegetables)
20  0.056634             (whole milk, rolls/buns)
21  0.074835       (whole milk, other vegetables)
{'Max': {1: 0.09852567361464158, 2: 0.07483477376715811}, 'Min': {1: 0.05765124555160142, 2: 0.030503304524656837}}
```

Рисунок 7 – Результат применения FPMax к ограниченному набору

Можно заметить что из за ограничений по товарам наборов в целом стало меньше. Из-за того что оставленные товары были преимущественно из самых часто встречающихся, в частности присутствовал самый часто встречающийся товар – whole milk – максимальное значение частоты товара сохранилось. Однако так как самый редко встречающийся товар в список не попал поднялся уровень минимальной поддержки в обоих случаях для одномерного набора.

Построим графики зависимости количества наборов определенной длины от уровня поддержки в промежутке [0.005, 0.4] с шагом 0.005 для 4 измерений: FPGrowth рассчитанных для полного набора данных и для выборочного набора данных и FPMax рассчитанных для тех же наборов. Результаты на рис. 8 – 11.

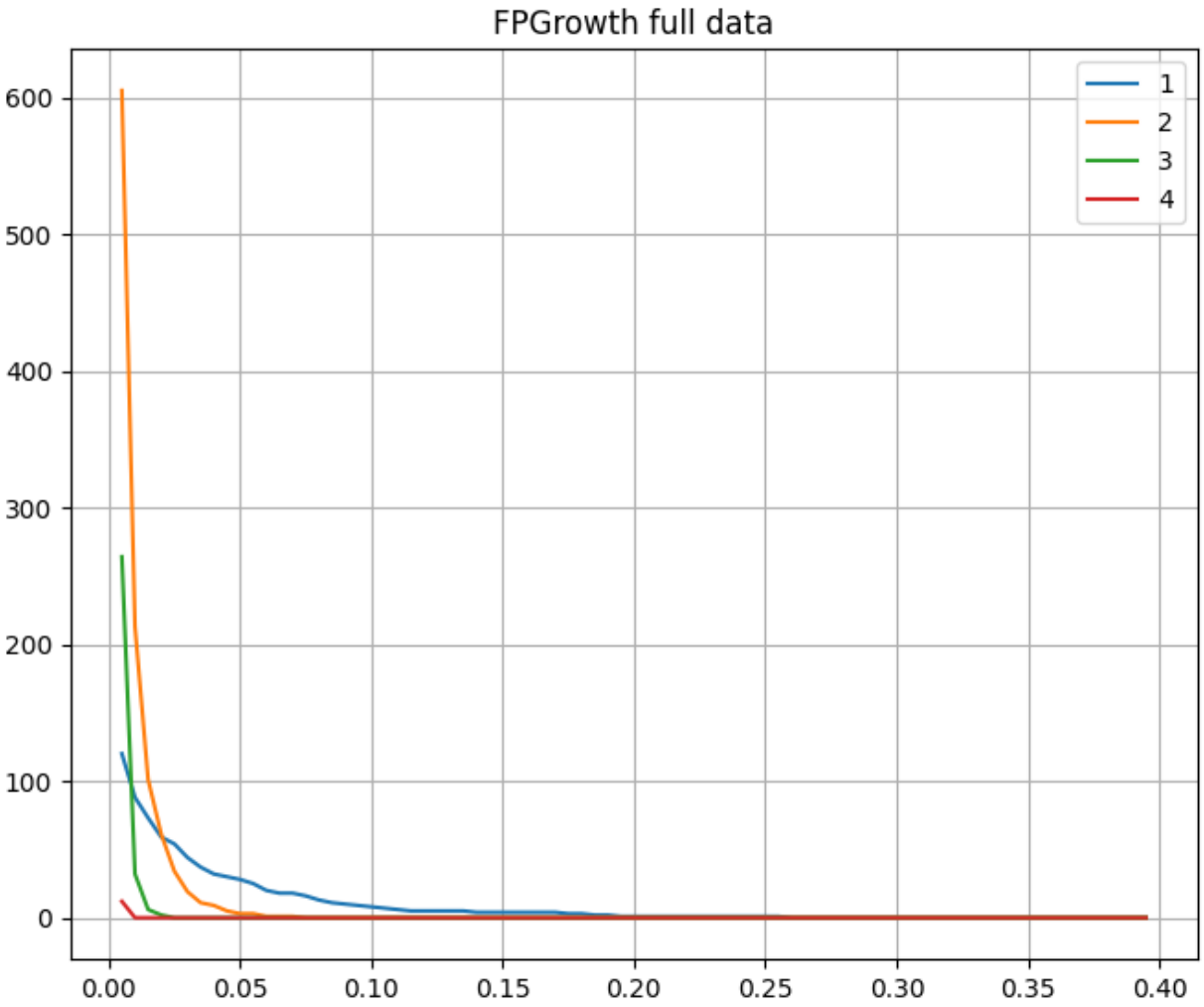


Рисунок 8 – График зависимости наборов определенной длины от уровня поддержки для полных данных обработанных FPGrowth

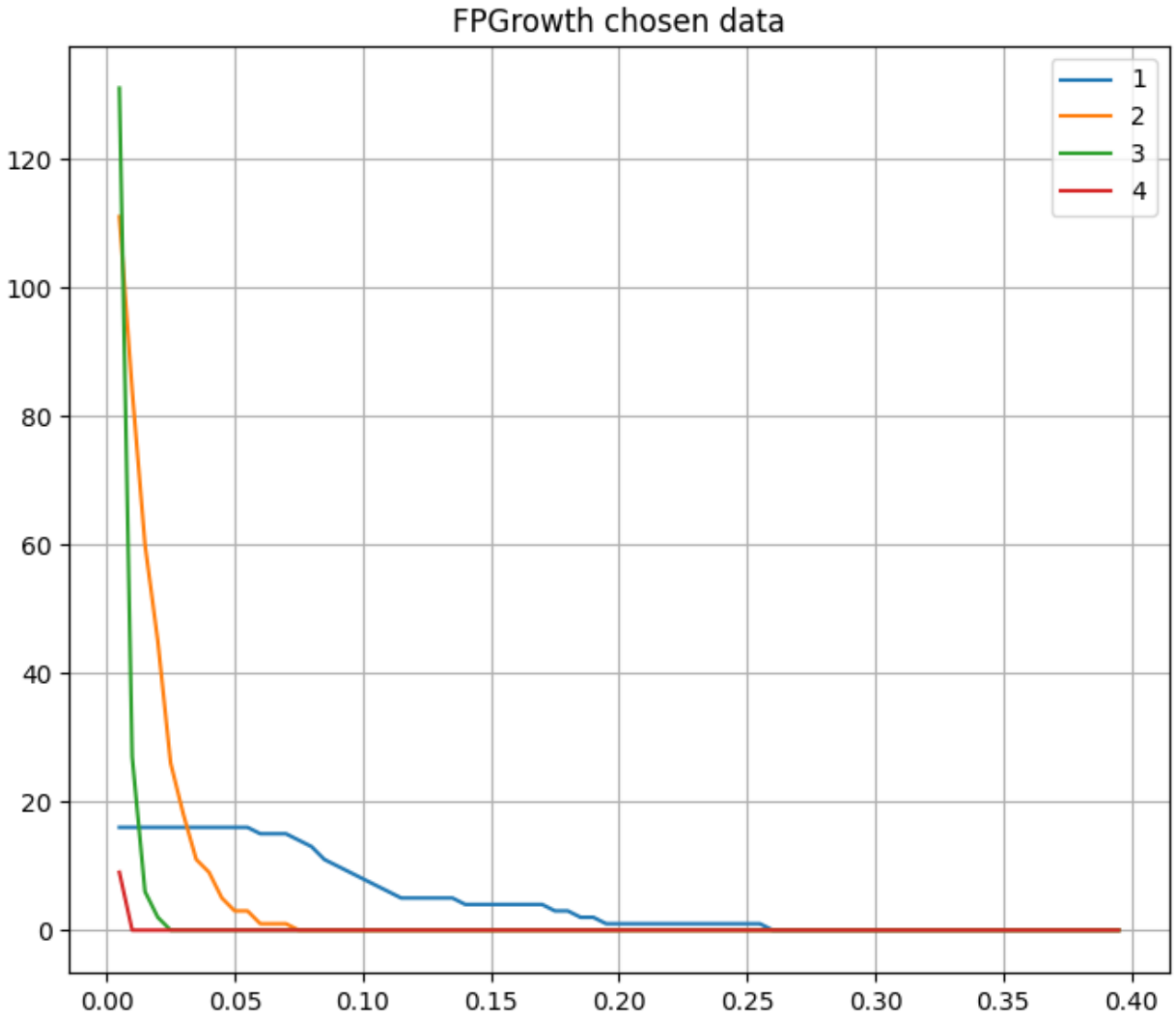


Рисунок 9 – График зависимости наборов определенной длины от уровня поддержки для выборных данных обработанных FPGrowth

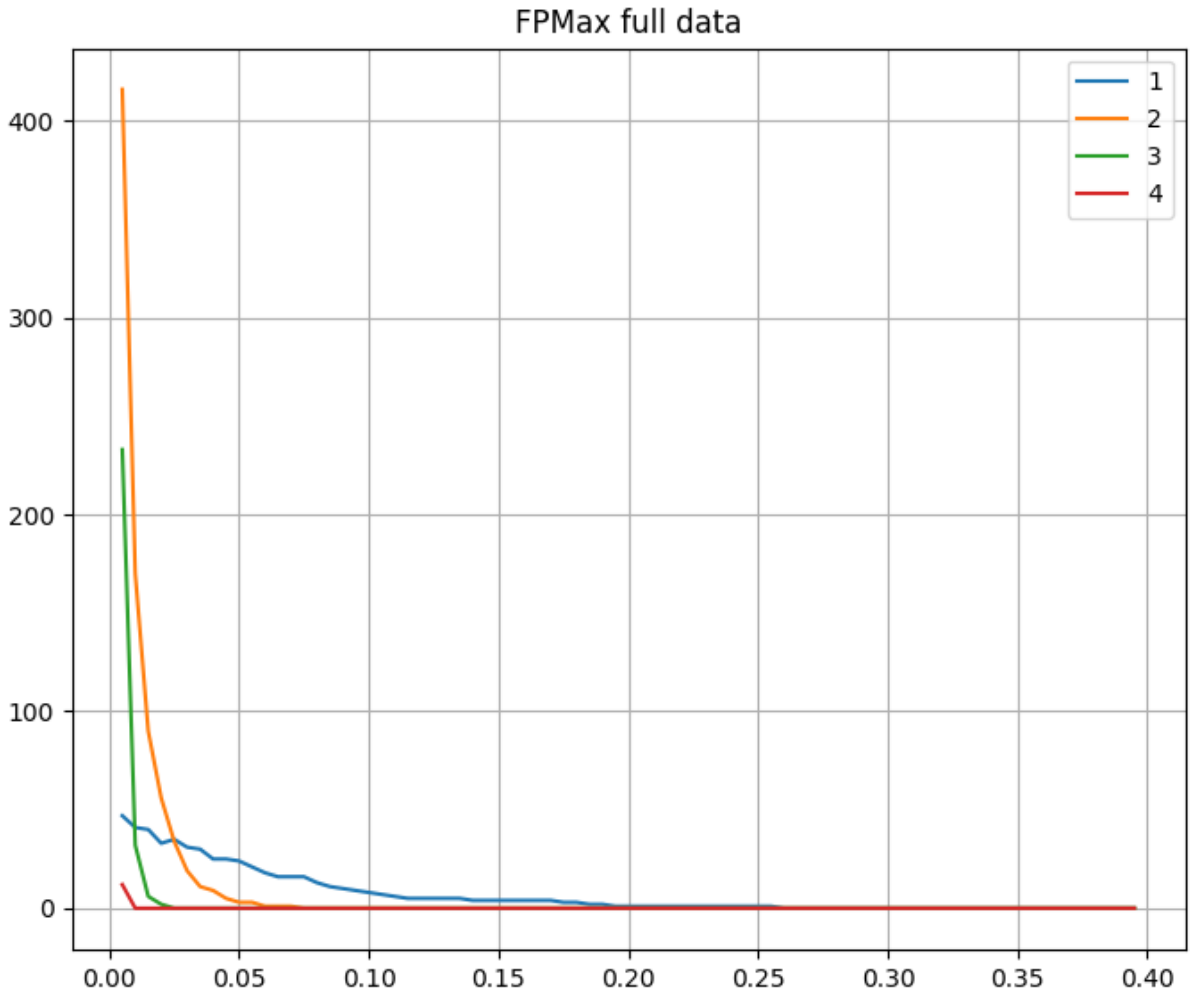


Рисунок 10 – График зависимости наборов определенной длины от уровня поддержки для полных данных обработанных FPMax

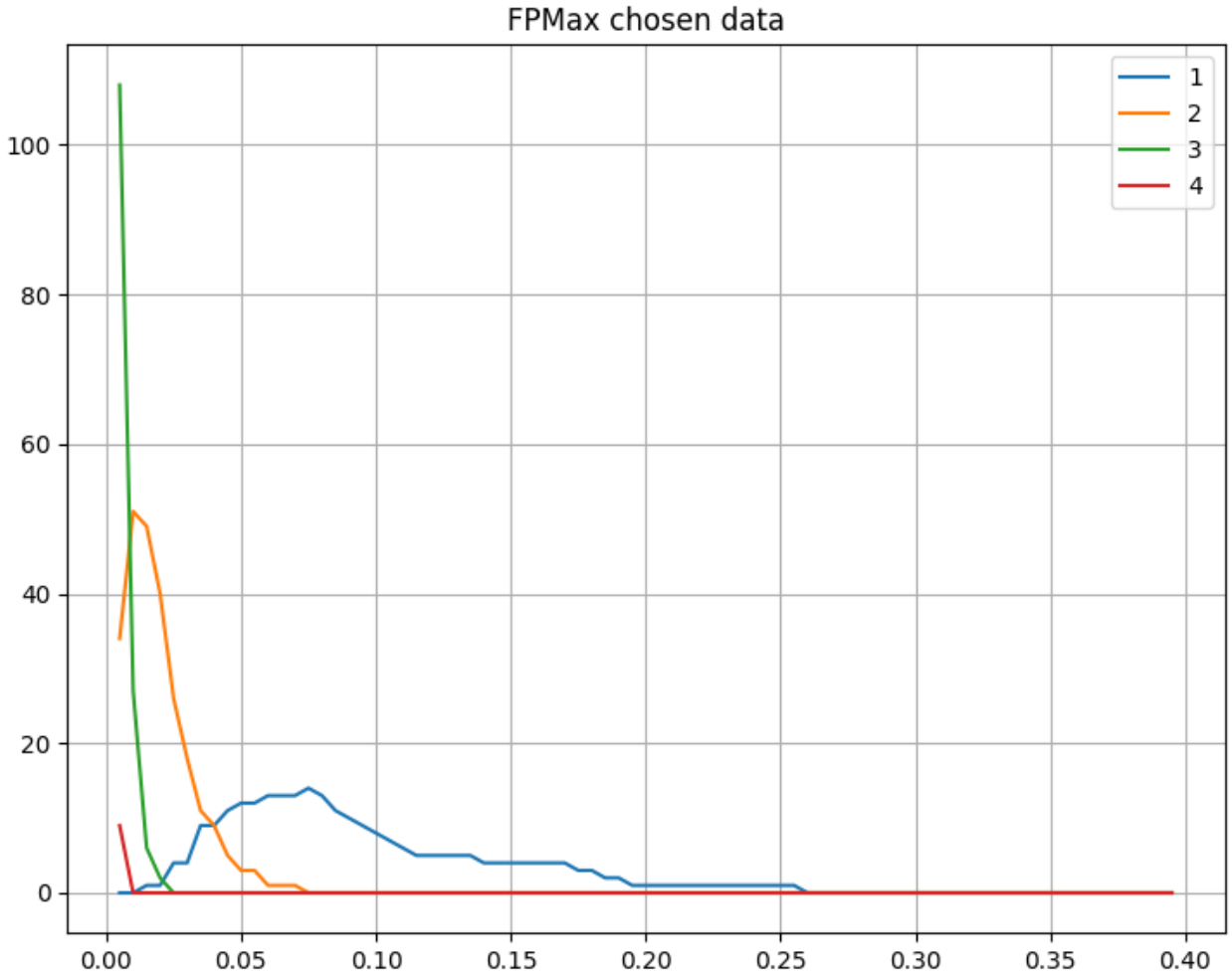


Рисунок 11 – График зависимости наборов определенной длины от уровня поддержки для выборных данных обработанных FPMax

Можно заметить, что применение алгоритма FPGrowth дает менее отличные результаты при применении к полному и выборочному наборам.

Ассоциативный анализ.

Проведем ассоциативный анализ полного набора данных. Результат на рис. 12.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	frozenset({'yogurt'})	frozenset({'whole milk'})	0.13950	0.25552	0.05602	0.40160	1.57174	0.02038	1.24413
1	frozenset({'other vegetables'})	frozenset({'whole milk'})	0.19349	0.25552	0.07483	0.38676	1.51363	0.02539	1.21401
2	frozenset({'rolls/buns'})	frozenset({'whole milk'})	0.18393	0.25552	0.05663	0.30790	1.20503	0.00964	1.07570

Рисунок 12 – Таблица зависимостей товаров

Рассмотрим столбцы исходного датафрейма:

- antecedent – товар от которого зависят зависимые товары, товар-причина

- consequent – товар, зависящий от налияия товара-причины, товар-следствие
- antecendent support – шанс наличия в транзакции первого товара, находится по формуле:

$$support(A) = \frac{N(A)}{N(All)}$$

- consequent support – шанс наличия в транзакции зависимого товара, находится по идентичной формуле
- support – шанс наличия в транзакции обоих товаров одновременно, находится по формуле:

$$support(A \to C) = \frac{N(A \cup C)}{N(All)}$$

- confidence – вероятность нахождения консеквента в транзакции в которой есть антецедент, находится по формуле:

$$confidence(A \to C) = \frac{support(A \to C)}{support(A)}$$

- lift – значение, показывающее насколько более часто консеквент с антецедентом присутствуют вместе в транзакциях сравнительно с тем условием если бы они были независимы, при независимости значене равняется 1:

$$lift(A \to C) = \frac{confidence(A \to C)}{support(C)}$$

- leverage – значение, показывающее разницу между наблюдаемой частотой появления транзакций в которых консеквент с антецедентом присутствуют вместе сравнительно с тем условием если бы они были независимы, вычисляется по формуле:

$$leverage(A \to C) = support(A \to C) - support(A) \times support(C)$$

- conviction – значение, уровень зависимости между наблюдаемой частотой появления транзакций в которых консеквент с антецедентом

присутствуют вместе сравнительно с тем условием если бы они были нанезависимы, т.е. похоже на lift, однако оперирует понятием частоты ошибок в отличие от lift, вычисляется по формуле:

$$conviction(A \to C) = \frac{1 - support(C)}{1 - confidence(A \to C)}$$

По умолчанию параметром metrics является confidence, следовательно все наборы подбираются по уровню min_threshold > confidence.

Подберем для метрик 'confidence', 'lift', 'leverage', 'conviction' такие значения параметров чтобы количество пар было больше 10 и рассчитаем для каждой метрики датафремов среднее значение, медиану и среднеквадратичное отклонение. Результаты на рис. 13 – 14.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	frozenset({'yogurt'})	frozenset({'whole milk'})	0.13950	0.25552	0.05602	0.40160	1.57174	0.02038	1.24413
1	frozenset({'whole milk'})	frozenset({'yogurt'})	0.25552	0.13950	0.05602	0.21926	1.57174	0.02038	1.10216
2	frozenset({'yogurt'})	frozenset({'other vegetables'})	0.13950	0.19349	0.04342	0.31122	1.60846	0.01642	1.17093
3	frozenset({'other vegetables'})	frozenset({'yogurt'})	0.19349	0.13950	0.04342	0.22438	1.60846	0.01642	1.10944
4	frozenset({'tropical fruit'})	frozenset({'whole milk'})	0.10493	0.25552	0.04230	0.40310	1.57759	0.01549	1.24725
5	frozenset({'whole milk'})	frozenset({'tropical fruit'})	0.25552	0.10493	0.04230	0.16554	1.57759	0.01549	1.07263
6	frozenset({'other vegetables'})	frozenset({'whole milk'})	0.19349	0.25552	0.07483	0.38676	1.51363	0.02539	1.21401
7	frozenset({'whole milk'})	frozenset({'other vegetables'})	0.25552	0.19349	0.07483	0.29288	1.51363	0.02539	1.14055
8	frozenset({'rolls/buns'})	frozenset({'other vegetables'})	0.18393	0.19349	0.04260	0.23162	1.19705	0.00701	1.04962
9	frozenset({'other vegetables'})	frozenset({'rolls/buns'})	0.19349	0.18393	0.04260	0.22018	1.19705	0.00701	1.04648
10	frozenset({'rolls/buns'})	frozenset({'whole milk'})	0.18393	0.25552	0.05663	0.30790	1.20503	0.00964	1.07570
11	frozenset({'whole milk'})	frozenset({'rolls/buns'})	0.25552	0.18393	0.05663	0.22165	1.20503	0.00964	1.04845
12	frozenset({'soda'})	frozenset({'whole milk'})	0.17438	0.25552	0.04006	0.22974	0.89911	-0.00450	0.96653
13	frozenset({'whole milk'})	frozenset({'soda'})	0.25552	0.17438	0.04006	0.15678	0.89911	-0.00450	0.97914
14	frozenset({'root vegetables'})	frozenset({'other vegetables'})	0.10900	0.19349	0.04738	0.43470	2.24660	0.02629	1.42669
15	frozenset({'other vegetables'})	frozenset({'root vegetables'})	0.19349	0.10900	0.04738	0.24488	2.24660	0.02629	1.17994
16	frozenset({'root vegetables'})	frozenset({'whole milk'})	0.10900	0.25552	0.04891	0.44869	1.75603	0.02106	1.35040
17	frozenset({'whole milk'})	frozenset({'root vegetables'})	0.25552	0.10900	0.04891	0.19140	1.75603	0.02106	1.10191

Рисунок 13 – Таблица пар для метрики confidence

```
Metric confidence stats:
confidence  -   Mean:  0.2829052726691693   Median:  0.23824809507666894    MSD:  0.09174103324706069
lift  -     Mean:  1.5083608149781373   Median:  1.5717351405345266     MSD:  0.36290651632916177
leverage  -     Mean:  0.015242803782607332     Median:  0.016423804156482313   MSD:  0.009282670551621401
conviction  -   Mean:  1.140331125249048    Median:  1.1057966943491302     MSD:  0.11799021582624744
```

Рисунок 14 – Статистика для каждой метрики

Построим граф с помощью библиотеки NetworkX по анализу по метрике confidence. Результат на рис. 15.

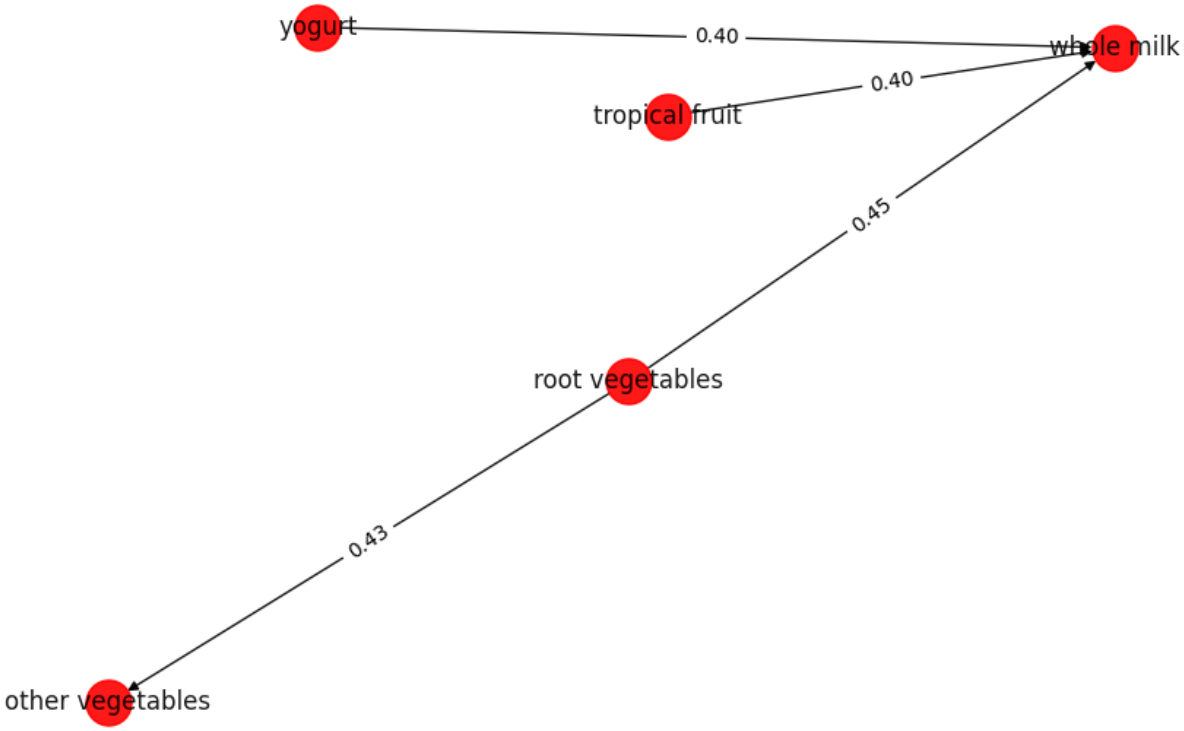


Рисунок 15 – Граф ассоциаций

Из графа можно понять что среди антецедентов половину ассоциаций занимают корнеплоды а самым частым консеквентом является цельное молоко присутствуя в 75% от всех транзакций. Также можно отметить что значение confidence у всех ассоциаций примерно одинаково.

Также ассоциации можно представлять с помощью визуализации матриц смежности и инцидентности.

Выводы

В ходе выполнения данной лабораторной работы было произведено знакомство с ассоциативным анализом. Были произведены трансформации транзакций с помощью TransactionEncoder. Были проведены исследования алгоритмов FPGrowth и FPMax библиотеки MLxtend на тестовых данных. Также был построен граф ассоциаций с помощью библиотеки NetworkX.