به نام خدا

مرضیه علیدادی – ۸۱۰۱۰۱۲۳۶ **تکلیف عملی ۲ داده کاوی**

❖ تمرین‌های تشریحی:

سوال ۱:

- **الف:**
  - مقدار support:

$$support(X \Rightarrow Y) = P(X \cup Y) = \frac{support\_count(X \cup Y)}{N}$$

✓ $\{(1 \le A \le 2), B = 1\} \rightarrow \{C = 1\}$:

X: $\{(1 \le A \le 2), B = 1\}$, Y: $\{C = 1\}$

$$support(X \Rightarrow Y) = P(X \cup Y) = \frac{support\_count(X \cup Y)}{12}$$
$$= \frac{2}{12} = 0.166 \cong 17\%$$

✓ $\{(5 \le A \le 8), B = 1\} \rightarrow \{C = 1\}$:

X: $\{(5 \le A \le 8), B = 1\}$, Y: $\{C = 1\}$

$$support(X \Rightarrow Y) = P(X \cup Y) = \frac{support\_count(X \cup Y)}{12}$$
$$= \frac{2}{12} = 0.166 \cong 17\%$$

  - مقدار confidence:

$$confidence(X \Rightarrow Y) = P(Y|X) = \frac{support(X \cup Y)}{support(X)} = \frac{support\_count(X \cup Y)}{support\_count(X)}$$

:$\{(1 \le A \le 2), B = 1\} \rightarrow \{C = 1\}$ ✓

X: $\{(1 \le A \le 2), B = 1\}$, Y: $\{C = 1\}$

$$confidence(X \Rightarrow Y) = P(Y|X) = \frac{support(X \cup Y)}{support(X)} = \frac{support\_count(X \cup Y)}{support\_count(X)}$$
$$= \frac{2}{2} = 100\%$$

:$\{(5 \le A \le 8), B = 1\} \rightarrow \{C = 1\}$ ✓

X: $\{(5 \le A \le 8), B = 1\}$, Y: $\{C = 1\}$

$$confidence(X \Rightarrow Y) = P(Y|X) = \frac{support(X \cup Y)}{support(X)} = \frac{support\_count(X \cup Y)}{support\_count(X)}$$
$$= \frac{2}{2} = 100\%$$

- **ب:** صورت این سوال دارای ابهام است. من دو برداشت از آن داشتم که هر دو را در زیر استفاده کرده‌ام:

برداشت اول

- Bin width= 2

1. bin 1: $(1 \le A \le 2)$

برای این bin، مقدار A=1 را در نظر می‌گیریم.

2. bin 2: $(3 \le A \le 4)$

برای این bin، مقدار A=2 را در نظر می‌گیریم.

3. bin 3: $(5 \le A \le 6)$

برای این bin، مقدار A=3 را در نظر می‌گیریم.

4. bin 4: $(7 \le A \le 8)$

برای این bin، مقدار A=4 را در نظر می‌گیریم.

5. bin 5: $(9 \le A \le 10)$

برای این bin، مقدار A=5 را در نظر می‌گیریم.

6. bin 6: $(11 \le A \le 12)$

برای این bin، مقدار A=6 را در نظر می‌گیریم.

مجموعه داده‌های جدید:

A	B	C
1	1	1
1	1	1
2	1	0
2	1	0
3	1	1
3	0	1
4	0	0
4	1	1
5	0	0
5	0	0
6	0	0
6	0	1

شروط جدید به فرم زیر می‌شوند:

1. {A = 1, B = 1} → {C = 1}
2. {(A = 3 or 4), B = 1} → {C = 1}

- مقدار support هر یک از 1-itemset ها:

$$support(A = 1) = \frac{2}{12} = 0.166 \cong 17\%$$
$$support(B = 1) = \frac{6}{12} = 0.5 = 50\%$$
$$support(C = 1) = \frac{6}{12} = 0.5 = 50\%$$
$$support(A = 3\ or\ 4) = \frac{4}{12} = 0.333 = 33\%$$

همه‌ی این 1-itemset ها، مکرر هستند.

- مقدار support هر یک از 2-itemset ها:

$$support(A = 1\ and\ B = 1) = \frac{2}{12} = 0.166 \cong 17\%$$
$$support(A = 1\ and\ C = 1) = \frac{2}{12} = 0.166 \cong 17\%$$
$$support((A = 3\ or\ 4)\ and\ B = 1) = \frac{2}{12} = 0.166 \cong 17\%$$
$$support((A = 3\ or\ 4)\ and\ C = 1) = \frac{3}{12} = 0.25 = 25\%$$
$$support(B = 1\ and\ C = 1) = \frac{4}{12} = 0.333 \cong 33\%$$

همه‌ی این 2-itemset ها، مکرر هستند.

- مقدار support هر یک از 3-itemset ها:

$$support(A = 1\ and\ B = 1\ and\ C = 1) = \frac{2}{12} = 0.166\ \cong 17\%$$

$$support((A = 3\ or\ 4)\ and\ B = 1\ and\ C = 1) = \frac{2}{12} = 0.166\ \cong 17\%$$

در نتیجه، مقدار support هر دو rule برابر ۱۷٪ است. و هر دو مکرر هستند.

- مقدار confidence هر یک از rule ها:

$$confidence((A = 1\ and\ B = 1) \rightarrow\ C = 1) = \frac{2}{2} = 100\%$$

$$confidence\left(((A = 3\ or\ 4)\ and\ B = 1) \rightarrow\ C = 1\right) = \frac{2}{2} = 100\%$$

هر دو rule همچنان قوی هستند.

- Bin width= 3

1. bin 1: $(1 \le A \le 3)$

برای این bin، مقدار A=1 را در نظر می‌گیریم.

2. bin 2: $(4 \le A \le 6)$

برای این bin، مقدار A=2 را در نظر می‌گیریم.

3. bin 3: $(7 \le A \le 9)$

برای این bin، مقدار A=3 را در نظر می‌گیریم.

4. bin 4: $(10 \le A \le 12)$

برای این bin، مقدار A=4 را در نظر می‌گیریم.

مجموعه داده‌های جدید:

A	B	C
1	1	1
1	1	1
1	1	0
2	1	0
2	1	1
2	0	1
3	0	0
3	1	1
3	0	0
4	0	0
4	0	0
4	0	1

شروط جدید به فرم زیر می‌شوند:

1. {A = 1, B = 1} → {C = 1}
2. {(A = 2 or 3), B = 1} → {C = 1}

- مقدار support هر یک از 1-itemset ها:

$$support(A = 1) = \frac{3}{12} = 0.25 = 25\%$$
$$support(B = 1) = \frac{6}{12} = 0.5 = 50\%$$
$$support(C = 1) = \frac{6}{12} = 0.5 = 50\%$$
$$support(A = 2\ or\ 3) = \frac{6}{12} = 0.50 = 50\%$$

همه‌ی این 1-itemset ها، مکرر هستند.

- مقدار support هر یک از 2-itemset ها:

$$support(A = 1\ and\ B = 1) = \frac{3}{12} = 0.25 = 25\%$$
$$support(A = 1\ and\ C = 1) = \frac{2}{12} = 0.166\ \cong 17\%$$
$$support((A = 2\ or\ 3)\ and\ B = 1) = \frac{3}{12} = 0.25 = 25\%$$
$$support((A = 2\ or\ 3)\ and\ C = 1) = \frac{3}{12} = 0.25 = 25\%$$
$$support(B = 1\ and\ C = 1) = \frac{4}{12} = 0.333 \cong 33\%$$

همه‌ی این 2-itemset ها، مکرر هستند.

- مقدار support هر یک از 3-itemset ها:

$$support(A = 1\ and\ B = 1\ and\ C = 1) = \frac{2}{12} = 0.166\ \cong 17\%$$
$$support((A = 2\ or\ 3)\ and\ B = 1\ and\ C = 1) = \frac{2}{12} = 0.166\ \cong 17\%$$

در نتیجه، مقدار support هر دو rule برابر ۱۷٪ است. و هر دو مکرر هستند.

- مقدار confidence هر یک از rule ها:

$$confidence((A = 1\ and\ B = 1) \rightarrow\ C = 1) = \frac{2}{3} = 0.666 \cong 67\%$$
$$confidence\left(((A = 2\ or\ 3)\ and\ B = 1) \rightarrow\ C = 1\right) = \frac{2}{3} = 0.666 \cong 67\%$$

هر دو rule همچنان قوی هستند.

- Bin width= 4

1. bin 1: $(1 \leq A \leq 4)$

برای این bin، مقدار A=1 را در نظر می‌گیریم.

2. bin 2: $(5 \leq A \leq 8)$

برای این bin، مقدار A=2 را در نظر می‌گیریم.

3. bin 3: $(9 \leq A \leq 12)$

برای این bin، مقدار A=3 را در نظر می‌گیریم

مجموعه داده‌های جدید:

A	B	C
1	1	1
1	1	1
1	1	0
1	1	0
2	1	1
2	0	1
2	0	0
2	1	1
3	0	0
3	0	0
3	0	0
3	0	1

شروط جدید به فرم زیر می‌شوند:

1. {A = 1, B = 1} → {C = 1}
2. {A = 2, B = 1} → {C = 1}

- مقدار support هر یک از 1-itemset ها:

$$support(A = 1) = \frac{4}{12} = 0.333 \cong 33\%$$

$$support(B = 1) = \frac{6}{12} = 0.5 = 50\%$$

$$support(C = 1) = \frac{6}{12} = 0.5 = 50\%$$

$$support(A = 2) = \frac{4}{12} = 0.333 \cong 33\%$$

همه‌ی این 1-itemset ها، مکرر هستند.

- مقدار support هر یک از 2-itemset ها:

$$support(A = 1\ and\ B = 1) = \frac{4}{12} = 0.333 \cong 33\%$$
$$support(A = 1\ and\ C = 1) = \frac{2}{12} = 0.166\ \cong 17\%$$
$$support(A = 2\ and\ B = 1) = \frac{2}{12} = 0.166\ \cong 17\%$$
$$support(A = 2\ and\ C = 1) = \frac{3}{12} = 0.25 = 25\%$$
$$support(B = 1\ and\ C = 1) = \frac{4}{12} = 0.333 \cong 33\%$$

همه‌ی این 2-itemset ها، مکرر هستند.

- مقدار support هر یک از 3-itemset ها:

$$support(A = 1\ and\ B = 1\ and\ C = 1) = \frac{2}{12} = 0.166\ \cong 17\%$$
$$support(A = 2\ and\ B = 1\ and\ C = 1) = \frac{2}{12} = 0.166\ \cong 17\%$$

در نتیجه، مقدار support هر دو rule برابر ۱۷٪ است. و هر دو مکرر هستند.

- مقدار confidence هر یک از rule ها:

$$confidence((A = 1\ and\ B = 1) \rightarrow\ C = 1) = \frac{2}{4} = 0.5 \cong 50\%$$
$$confidence\big((A = 2\ and\ B = 1) \rightarrow\ C = 1\big) = \frac{2}{2} = 100\%$$

فقط rule دوم، با این اندازه‌ی bin همچنان قوی است.

برداشت دوم

- Bin width= 2

شروط جدید به فرم زیر می‌شوند:

1. $\{\{(1 \le A \le 2), B = 1\} \rightarrow \{C = 1\}$
2. $\{\{(5 \le A \le 6), B = 1\} \rightarrow \{C = 1\}$
3. $\{\{(7 \le A \le 8), B = 1\} \rightarrow \{C = 1\}$

- مقدار support هر یک از این 3-itemset ها:

$$support(1 \le \mathrm{A} \le 2\ and\ B = 1\ and\ C = 1) = \frac{2}{12} = 0.166\ \cong 17\%$$

$$support(5 \leq \mathrm{A} \leq 6\ and\ B = 1\ and\ C = 1) = \frac{1}{12} = 0.083 \cong 8\%$$

$$support(7 \leq \mathrm{A} \leq 8\ and\ B = 1\ and\ C = 1) = \frac{1}{12} = 0.083 \cong 8\%$$

در نتیجه، مقدار support مربوط به rule دوم و سوم، کم‌تر از ۱۵٪ شد و به عنوان itemset های مکرر شناخته نشدند. که اشتباه است.

- مقدار confidence هر یک از rule ها:

$$confidence((1 \leq \mathrm{A} \leq 2\ \ and\ B = 1) \rightarrow\ C = 1) = \frac{2}{2} = 100\%$$

$$confidence\big((5 \leq \mathrm{A} \leq 6\ \ and\ B = 1) \rightarrow\ C = 1\big) = \frac{1}{1} = 100\%$$

$$confidence((7 \leq \mathrm{A} \leq 8\ \ and\ B = 1) \rightarrow\ C = 1) = \frac{1}{1} = 100\%$$

فقط rule اول، با این اندازه‌ی bin همچنان قوی است. rule دوم و سوم شرط confidence را دارند، ولی چون شرط support را ندارند، قوی نیستند. و کشف نمی‌شوند.

- Bin width= 3

شروط جدید به فرم زیر می‌شوند:

1. $\{\{(1 \leq A \leq 3), B = 1\} \rightarrow \{C = 1\}$
2. $\{\{(4 \leq A \leq 6), B = 1\} \rightarrow \{C = 1\}$
3. $\{\{(7 \leq A \leq 9), B = 1\} \rightarrow \{C = 1\}$

- مقدار support هر یک از این 3-itemset ها:

$$support(1 \leq \mathrm{A} \leq 3\ and\ B = 1\ and\ C = 1) = \frac{2}{12} = 0.166\ \cong 17\%$$

$$support(4 \leq \mathrm{A} \leq 6\ and\ B = 1\ and\ C = 1) = \frac{1}{12} = 0.083 \cong 8\%$$

$$support(7 \leq \mathrm{A} \leq 9\ and\ B = 1\ and\ C = 1) = \frac{1}{12} = 0.083 \cong 8\%$$

در نتیجه، مقدار support مربوط به rule دوم و سوم، کم‌تر از ۱۵٪ شد و به عنوان itemset های مکرر شناخته نشدند. که اشتباه است.

- مقدار confidence هر یک از rule ها:

$$confidence((1 \leq \mathrm{A} \leq 3\ \ and\ B = 1) \rightarrow\ C = 1) = \frac{2}{3} = 0.67 \cong 70\%$$

$$confidence\big((4 \leq \mathrm{A} \leq 6\ \ and\ B = 1) \rightarrow\ C = 1\big) = \frac{1}{2} = 50\%$$

$$confidence((7 \leq \mathrm{A} \leq 9 \ and\ B = 1) \rightarrow\ C = 1) = \frac{1}{1} = 100\%$$

فقط rule اول، با این اندازه‌ی bin همچنان قوی است. rule سوم شرط confidence را دارد، ولی چون شرط support را ندارد، قوی نیست. و هر دوی rule دوم و سوم کشف نمی‌شوند.

- Bin width= 4

شروط جدید به فرم زیر می‌شوند:

1. $\{\{(1 \leq A \leq 4), B = 1\} \rightarrow \{C = 1\}$
2. $\{\{(5 \leq A \leq 8), B = 1\} \rightarrow \{C = 1\}$

- مقدار support هر یک از این 3-itemset ها:

$$support(1 \leq \mathrm{A} \leq 4\ and\ B = 1\ and\ C = 1) = \frac{2}{12} = 0.166\ \cong 17\%$$

$$support(5 \leq \mathrm{A} \leq 8\ and\ B = 1\ and\ C = 1) = \frac{2}{12} = 0.166\ \cong 17\%$$

در نتیجه، مقدار support مربوط به هر دو rule، بیشتر از ۱۵٪ شد و به عنوان itemset های مکرر شناخته شدند.

- مقدار confidence هر یک از rule ها:

$$confidence((1 \leq \mathrm{A} \leq 4\ \ and\ B = 1) \rightarrow\ C = 1) = \frac{2}{4} = 50\%$$

$$confidence\big((5 \leq \mathrm{A} \leq 8\ \ and\ B = 1) \rightarrow\ C = 1\big) = \frac{2}{2} = 100\%$$

فقط rule دوم، با این اندازه‌ی bin همچنان قوی است. rule اول شرط support را داشت، ولی چون شرط confidence را ندارد، قوی نیست. و کشف نمی‌شود.

بیشتر این به نظر می‌رسد که برداشت دوم مدنظر طراح سوال بوده‌است. پس بخش بعدی با در نظر گرفتن این برداشت، پاسخ داده شد.

- **ج:** استفاده از این روش مناسب نیست؛ زیرا ممکن است شرطی که بررسی می‌شود، در دو یا چند bin متفاوت قرار گیرد و نتایج به درستی محاسبه نشود. برای مثال، وقتی اندازه‌ی bin ها برابر ۲ یا ۳ در نظر گرفته‌شد، rule دوم کشف نشد. و وقتی اندازه‌ی bin ها برابر ۴ در نظر گرفته‌شد، rule اول کشف نشد.

هیچ bin-width ای که برای پیدا کردن هر دو rule مناسب باشد، در این روش وجود ندارد. زیرا بازه‌های بررسی این دو rule با هم متفاوت است. و binning موجب شکستن یکی یا بزرگ شدن

یکی یا هر دو می‌شود. که این مناسب نیست و موجب جواب ناکامل می‌شود.

- بهتر است از الگوریتم **FPGrowth** استفاده کرد:

۱. Rule اول:

	Itemset number
A	$1 \leq A \leq 2$
B	1, 2
C	1, 2

	frequency
B	2
C	2

A	Frequent itemsets
1	B, C
2	B, C

۲. Rule دوم:

	Itemset number
A	$5 \leq A \leq 8$
B	5, 8
C	5, 6, 8

	frequency
B	2
C	3

A	Frequent itemsets
5	C, B
6	C
7	-
8	C, B

بدین ترتیب، با این روش می‌توان هر دو rule را کشف کرد.

سوال ۲:

- **الف:** برای پیدا کردن بیشترین k، می‌توان از k=1 شروع کرد. و به صورت تکراری، با افزایش k تا حد ممکن، به بیشترین k ممکن رسید:

۱. اگر k=1 در نظر بگیریم، 1-itemset ها عبارتند از:

A, B, C, D, E, F, G, H

مقدار support هر کدام برای پیدا کردن 1-itemset ها برای بررسی مکرر بودن/نبودن آن‌ها، محاسبه می‌شود:

$$support(A) = \frac{3}{5} = 60\%$$
$$support(B) = \frac{4}{5} = 80\%$$
$$support(C) = \frac{4}{5} = 80\%$$
$$support(D) = \frac{4}{5} = 80\%$$
$$support(E) = \frac{2}{5} = 40\%$$
$$support(F) = \frac{1}{5} = 20\%$$
$$support(G) = \frac{1}{5} = 20\%$$
$$support(H) = \frac{2}{5} = 40\%$$

با توجه به این‌که min_support= 60% است، 1-itemset های مکرر عبارتند از:

A, B, C, D

۲. حال k=2 را در نظر می‌گیریم و با در نظر گرفتن 1-itemset های مکرر، 2-itemset هایی که ممکن است مکرر باشند، عبارتند از:

AB, AC, AD, BC, BD, CD

مقدار support هر کدام برای پیدا کردن 2-itemset ها برای بررسی مکرر بودن/نبودن آن‌ها، محاسبه می‌شود:

$$support(AB) = \frac{2}{5} = 40\%$$
$$support(AC) = \frac{2}{5} = 40\%$$

$$support(AD) = \frac{3}{5} = 60\%$$
$$support(BC) = \frac{4}{5} = 80\%$$
$$support(BD) = \frac{3}{5} = 60\%$$
$$support(CD) = \frac{3}{5} = 60\%$$

با توجه به این‌که min_support= 60% است، 2-itemset های مکرر عبارتند از:
AD, BC, BD, CD

۳. حال k=3 را در نظر می‌گیریم و با در نظر گرفتن 2-itemset های مکرر، 3-itemset ای که ممکن است مکرر باشد، عبارت است از:

BCD

مقدار support آن برای بررسی مکرر بودن/نبودن آن، محاسبه می‌شود:

$$support(BCD) = \frac{3}{5} = 60\%$$

با توجه به این‌که min_support= 60% است، **BCD** یک 3-itemset مکرر است.

- بنابراین، بزرگ‌ترین k برای k-itemset های مکرر، برابر **۳** است. و یک 3-itemset مکرر به شرح زیر وجود دارد:

{B, C, D}

- **ب:** itemset زیر را در نظر بگیرید:

{A, B}

این itemset دارای support(AB)= 40% است و مکرر نیست. اما اگر A و B که زیرمجموعه‌های آن هستند را در نظر بگیریم، support(A)= 60% و support(B)= 80% هستند. و هر دوی آن‌ها مکرر هستند.

- **ج:** closed pattern ها عبارتند از:

{B, C, D}: 60%
{A, D}: 60%
{B, C}: 80%
{D}: 80%

- **د:** max pattern ها عبارتند از:

{B, C, D}: 60%
{A, D}: 60%

- **ه:** هدف، به دست آوردن association rule های قوی است که با metarule زیر مطابقت دارند:

$x \in \{001, 002, \ldots, 005\}, buys(x, item1) \wedge buys(x, item2) \Rightarrow buys(x, item3).[s, c]$

این association rule ها باید حداقل دارای support= 60% و confidence= 70% باشند. و همچنین، دارای ۳ آیتم هستند. با توجه به این‌که itemset های مکرر (دارای support= 60%) در بخش (الف) محاسبه شده‌اند، و مکرر بودن، یکی از شروط لازم برای association rule قوی بودن است، 3-itemset های مکرر همان بخش را برای بررسی بقیه‌ی شرایط در نظر می‌گیریم:

{B, C, D}

حال تمام شکل‌های ممکن association rule های مربوط به این itemset را بررسی می‌کنیم:

1. $x \in \{001, 002, \ldots, 005\}, buys(x, \mathrm{B}) \wedge buys(x, \mathrm{C}) \Rightarrow buys(x, \mathrm{D}).[60\%, c]$
2. $x \in \{001, 002, \ldots, 005\}, buys(x, \mathrm{B}) \wedge buys(x, \mathrm{D}) \Rightarrow buys(x, \mathrm{C}).[60\%, c]$
3. $x \in \{001, 002, \ldots, 005\}, buys(x, \mathrm{C}) \wedge buys(x, \mathrm{D}) \Rightarrow buys(x, \mathrm{B}).[60\%, c]$

مقدار confidence این association rule ها برای بررسی قوی بودن/نبودن آن‌ها بررسی می‌شود:

$$confidence(rule1) = \frac{3}{4} = 75\%$$

$$confidence(rule2) = \frac{3}{3} = 100\%$$
$$confidence(rule3) = \frac{3}{3} = 100\%$$

مشخص است که برای هر سه rule، مقدار confidence آن‌ها از min_confidence= 70% بیشتر است.

بنابراین، association rule های قوی، عبارتند از:

1. **{B, C} ⇒ {D}.[60%, 75%]**
2. **{B, D} ⇒ {C}.[60%, 100%]**
3. **{C, D} ⇒ {B}.[60%, 100%]**

- **و:** الگوهای مکرر single به ترتیب نزولی support آن‌ها:

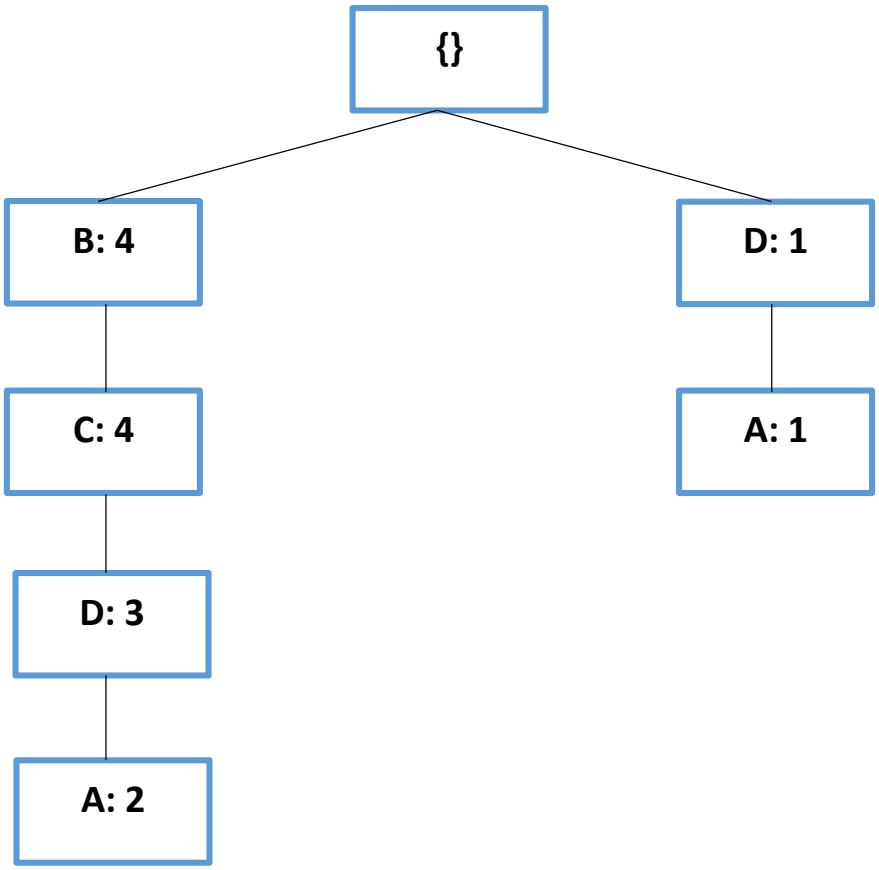
$$support(B) = \frac{4}{5} = 80\%$$
$$support(C) = \frac{4}{5} = 80\%$$
$$support(D) = \frac{4}{5} = 80\%$$
$$support(A) = \frac{3}{5} = 60\%$$

ترتیب خواسته شده در سوال برای شکستن پیوندها، از چپ به راست:

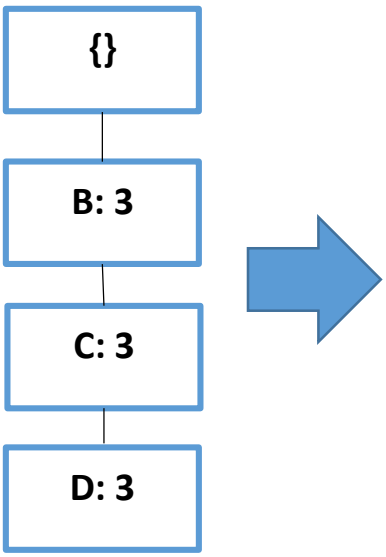
B, C, D, A

FP-Tree متناظر:

Transaction_id	Item_bought	Ordered frequent items
001	{H, A, D, B, C}	{B, C, D, A}
002	{D, A, E, F}	{D, A}
003	{C, D, B, E}	{B, C, D}
004	{B, A, C, H, D}	{B, C, D, A}
005	{B, G, C}	{B, C}



- ز: پایگاه داده‌ی A's conditional:
  - A-conditional pattern base: **BCD:2, D:1**
  - A-conditional pattern tree:



  - all frequent patterns related to A:

A,
BA, CA, DA,
BCA, BDA, CDA,
BCDA

سوال ۳:

- **الف:** با توجه به این‌که min-support در این سوال، مانند سوال قبل، همچنان برابر ٪۶۰ است، و در سوال قبل در بخش الف، الگوهای مکرر پیدا شده‌اند، از همان الگوهای مکرر گزارش شده استفاده می‌کنیم:

1. A, 2. B, 3. C, 4. D, 5. AD,
6. BC, 7. BD, 8. CD, 9. BCD

حال باید شرط "مجموع قیمت اقلام بزرگتر از ۴۵" را برای هر یک از این الگوهای مکرر بررسی کرد:

$sum(1.\ price)$ = A.price = 10 < 45
$sum(2.\ price)$ = B.price = 20 < 45
$sum(3.\ price)$ = C.price = 40 < 45
$sum(4.\ price)$ = D.price = 30 < 45
$sum(5.\ price)$ = A.price + D.price = 10 + 30 = 40 < 45
$sum(6.\ price)$ = B.price + C.price = 20 + 40 = 60 ≥ 45
$sum(7.\ price)$ = B.price + D.price = 20 + 30 = 50 ≥ 45
$sum(8.\ price)$ = C.price + D.price = 40 + 30 = 70 ≥ 45
$sum(9.\ price)$ = B.price + C.price + D.price = 20 + 40 + 30 = 90 ≥ 45

بنابراین، الگوهای مکرر دارای شرط بیان شده، عبارتند از:

BC, BD, CD, BCD

- **ب:** یک شرط، در صورتی anti-monotone است، که اگر یک itemset داشته‌باشیم که این شرط را satisfy می‌کند، هر زیرمجموعه از آن نیز آن شرط را satisfy کند.
و یک شرط در صورتی monotone است، که اگر یک itemset داشته‌باشیم که این شرط را satisfy می‌کند، هر superset از آن نیر آن شرط را satisfy کند.
شرط $sum(S.\ price) \geq 45$ ، در صورتی که بر روی یک itemset برقرار باشد و مجموع قیمت آیتم‌های موجود در آن itemset بیشتر از ۴۵ باشد، روی هر superset از آن نیز برقرار است. اما لزوماً روی هر زیرمجموعه‌ی آن برقرار نیست. در نتیجه، این شرط، monotone است.
شرط $sum(S.\ price) \leq 45$ ، در صورتی که بر روی یک itemset برقرار باشد و مجموع قیمت آیتم‌های موجود در آن itemset کم‌تر از ۴۵ باشد، روی هر زیرمجموعه از آن نیز برقرار است. اما لزوماً روی هر superset آن برقرار نیست. در نتیجه، این شرط، anti-monotone است.

با توجه به این‌که شرط $sum(S.price) \leq 45$ ، anti-monotone است، از الگوریتم ECLAT برای استخراج الگوهای مکرر با این شرط، استفاده می‌شود:
این الگوریتم، ابتدا تراکنش‌ها را در فرم عمودی بازنویسی می‌کند:

A: 001, 002, 004
B: 001, 003, 004, 005
C: 001, 003, 004, 005
D: 001, 002, 003, 004
E: 002, 003
F: 002
G: 005
H: 001, 004

مقدار support و price هر یک از آیتم‌ها برای بررسی شرایط، محاسبه می‌شود:

A: $\text{support} = \frac{3}{5} = 60\% \geq 60\%$, price = $10 \leq 45$ → ✓
B: $\text{support} = \frac{4}{5} = 80\% \geq 60\%$, price = $20 \leq 45$ → ✓
C: $\text{support} = \frac{4}{5} = 80\% \geq 60\%$, price = $40 \leq 45$ → ✓
D: $\text{support} = \frac{4}{5} = 80\% \geq 60\%$, price = $30 \leq 45$ → ✓
E: $\text{support} = \frac{2}{5} = 40\% < 60\%$, price = $90 > 45$ → ✗
F: $\text{support} = \frac{1}{5} = 20\% < 60\%$, price = $90 > 45$ → ✗
G: $\text{support} = \frac{1}{5} = 20\% < 60\%$, price = $30 \leq 45$ → ✗
H: $\text{support} = \frac{2}{5} = 40\% < 60\%$, price = $50 > 45$ → ✗

سپس، با توجه به anti-monotone بودن شرط اعمال شده روی itemset ها، هر itemset ای که این شرط را satisfy نکند، superset آن نیز، آن را satisfy نمی‌کند (این مسئله، از عکس نقیض تعریف شروط anti-monotone که در بالا گفته شد، برداشت می‌شود.). در نتیجه، با اشتراک گرفتن از سطرهای مربوط به 1-itemset های مکرری که شرط anti-monotone اعمال شده را satisfy می‌کنند، فرم عمودی مربوط به 2-itemset هایی که ممکن است مکرر باشند و آن شرط را satisfy کنند، تشکیل داده می‌شود:

AB: 001, 004
AC: 001, 004
AD: 001, 002, 004
BC: 001, 003, 004, 005
BD: 001, 003, 004
CD: 001, 003, 004

مقدار support و price هر یک از این itemset ها برای بررسی شرایط، محاسبه می‌شود:

AB: $\text{support} = \frac{2}{5} = 40\% < 60\%$, price = 10+20 = 30 ≤ 45 → ✗

AC: $\text{support} = \frac{2}{5} = 40\% < 60\%$, price = 10+40 = 50 > 45 → ✗

AD: $\text{support} = \frac{3}{5} = 60\% \geq 60\%$, price = 10+30 = 40 ≤ 45 → ✓

BC: $\text{support} = \frac{4}{5} = 80\% \geq 60\%$, price = 20+40 = 60 > 45 → ✗

BD: $\text{support} = \frac{3}{5} = 60\% \geq 60\%$, price = 20+30 = 50 > 45 → ✗

CD: $\text{support} = \frac{3}{5} = 60\% \geq 60\%$, price = 40+30 = 70 > 45 → ✗

در نتیجه، الگوهای مکرر دارای شرط قید شده، با استفاده از این الگوریتم محاسبه شدند، و عبارتند از:

A, B, C, D, AD

- **ج:** بله؛ هر دو این شروط convertible هستند.

اگر آیتم‌ها را به صورت نزولی قیمت آن‌ها مرتب کنیم و به همین ترتیب هم itemset ها را تشکیل بدهیم، آنگاه شرط $\text{avg(S.price)} \geq 30$ به یک شرط anti-monotone تبدیل می‌شوند:

1. E: 90	2. F: 90	3. H: 50	4. C: 40
5. D: 30	6. G: 30	7. B: 20	8. A: 10

زیرا برای مثال اگر itemset ای به فرم B وجود داشته باشد که این شرط را satisfy نمی‌کند، هر itemset دیگری که این itemset را به عنوان پیشوند داشته باشد، مانند BA نیز، این شرط را satisfy نمی‌کند. و دیگر بررسی نمی‌شود.

اگر آیتم‌ها را به صورت صعودی قیمت آن‌ها مرتب کنیم، آنگاه شرط $\text{avg(S.price)} \leq 30$ به یک شرط anti-monotone تبدیل می‌شوند:

1. A: 10	2. B: 20	3. G: 30	4. D: 30
5. C: 40	6. H: 50	7. F: 90	8. E: 90

زیرا برای مثال اگر itemset ای به فرم DC وجود داشته باشد که این شرط را satisfy نمی‌کند، هر itemset دیگری که این itemset را به عنوان پیشوند داشته باشد، مانند DCH نیز، این شرط را satisfy نمی‌کند. و دیگر بررسی نمی‌شود.

❖ تمرین‌های عملی:

سوال ۱:

- ۵ سطر اول دیتاست movies:

	movieId	title	genres
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	2	Jumanji (1995)	Adventure|Children|Fantasy
2	3	Grumpier Old Men (1995)	Comedy|Romance
3	4	Waiting to Exhale (1995)	Comedy|Drama|Romance
4	5	Father of the Bride Part II (1995)	Comedy

- ۵ سطر اول دیتاست ratings:

	userId	movieId	rating	timestamp
0	1	1	4.0	964982703
1	1	3	4.0	964981247
2	1	6	4.0	964982224
3	1	47	5.0	964983815
4	1	50	5.0	964982931

- در هیچ یک از دیتاست‌ها، مقدار گم‌شده وجود ندارد:

Movies:

	Total	Percent
movieId	0	0.0
title	0	0.0
genres	0	0.0

Ratings:

	Total	Percent
userId	0	0.0
movieId	0	0.0
rating	0	0.0
timestamp	0	0.0

- ستون‌های دیتاست‌ها از نظر عدم وجود داده‌های غیرمرتبط، بررسی شدند. داده‌ی غیرمرتبطی در هیچ یک از ستون‌ها یافت نشد.

- دو دیتاست با استفاده از ستون movieId باهم ترکیب شدند:

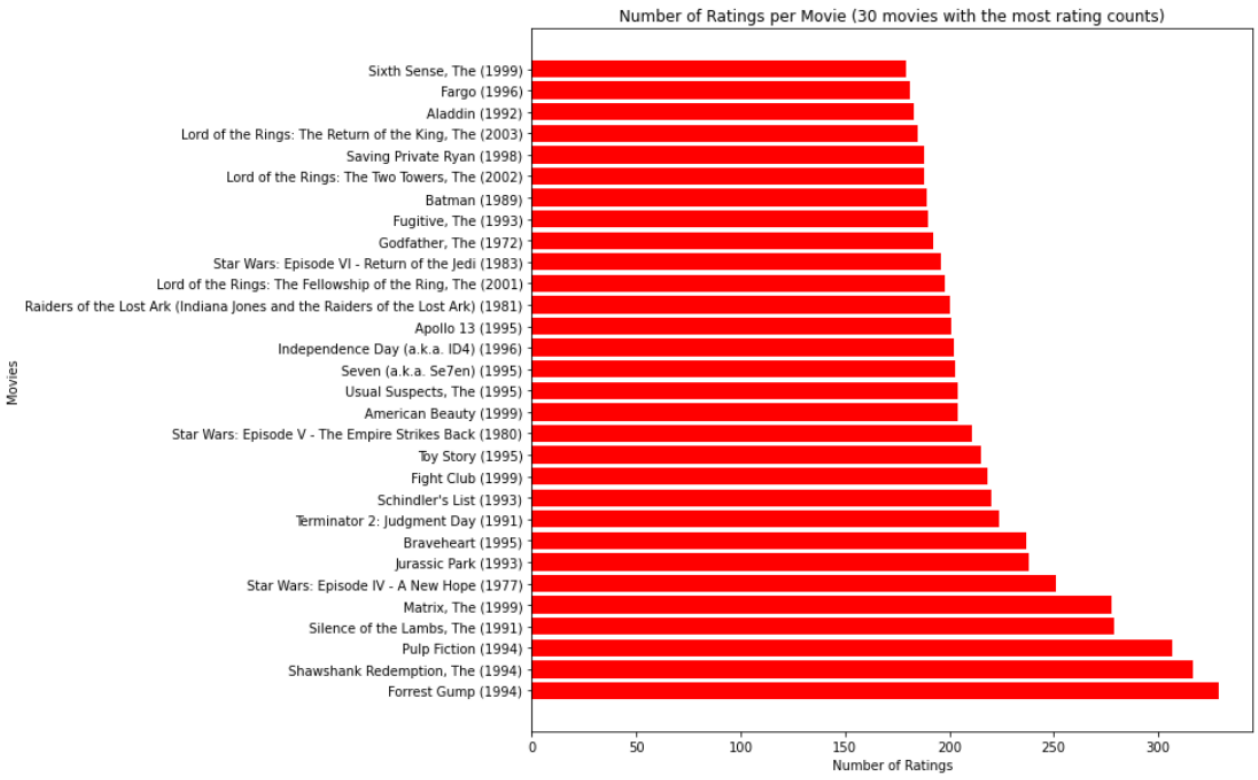
	movieId	title	genres	userId	rating	timestamp
0	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	1.0	4.0	9.649827e+08
1	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	5.0	4.0	8.474350e+08
2	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	7.0	4.5	1.106636e+09
3	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	15.0	2.5	1.510578e+09
4	1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy	17.0	4.5	1.305696e+09
...	...	...	...	...	...	...
100849	193581	Black Butler: Book of the Atlantic (2017)	Action|Animation|Comedy|Fantasy	184.0	4.0	1.537109e+09
100850	193583	No Game No Life: Zero (2017)	Animation|Comedy|Fantasy	184.0	3.5	1.537110e+09
100851	193585	Flint (2017)	Drama	184.0	3.5	1.537110e+09
100852	193587	Bungo Stray Dogs: Dead Apple (2018)	Action|Animation	184.0	3.5	1.537110e+09
100853	193609	Andrew Dice Clay: Dice Rules (1991)	Comedy	331.0	4.0	1.537158e+09

100854 rows × 6 columns

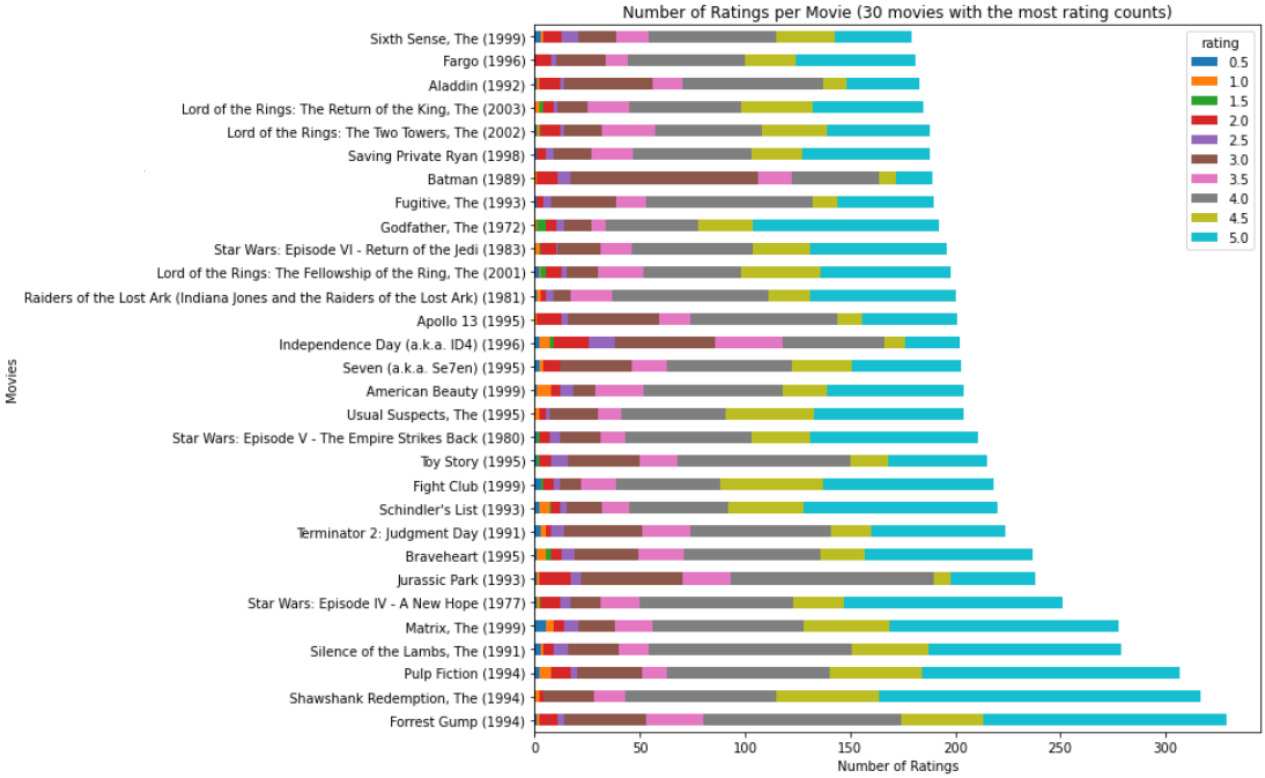
- ۱۸ مورد از فیلم‌ها، هیچ نظری برای آن‌ها ثبت نشده‌است: (آن‌ها را فعلاً از دیتاست حذف نکردم؛ ولی این موضوع را برای تحلیل‌های بعدی در نظر خواهیم داشت.)

	Total	Percent
userId	18	0.000178
rating	18	0.000178
timestamp	18	0.000178
movieId	0	0.000000
title	0	0.000000
genres	0	0.000000

- نمودار:
  - ۳۰ فیلم با بیشترین تعداد نظر ثبت شده برای آن‌ها، در نمودار bar plot زیر نمایش داده شده‌اند:



- یک بار دیگر، این نمودار به صورت رنگ شده به تفکیک امتیازهای داده‌شده، به شکل زیر نمایش داده‌شد:

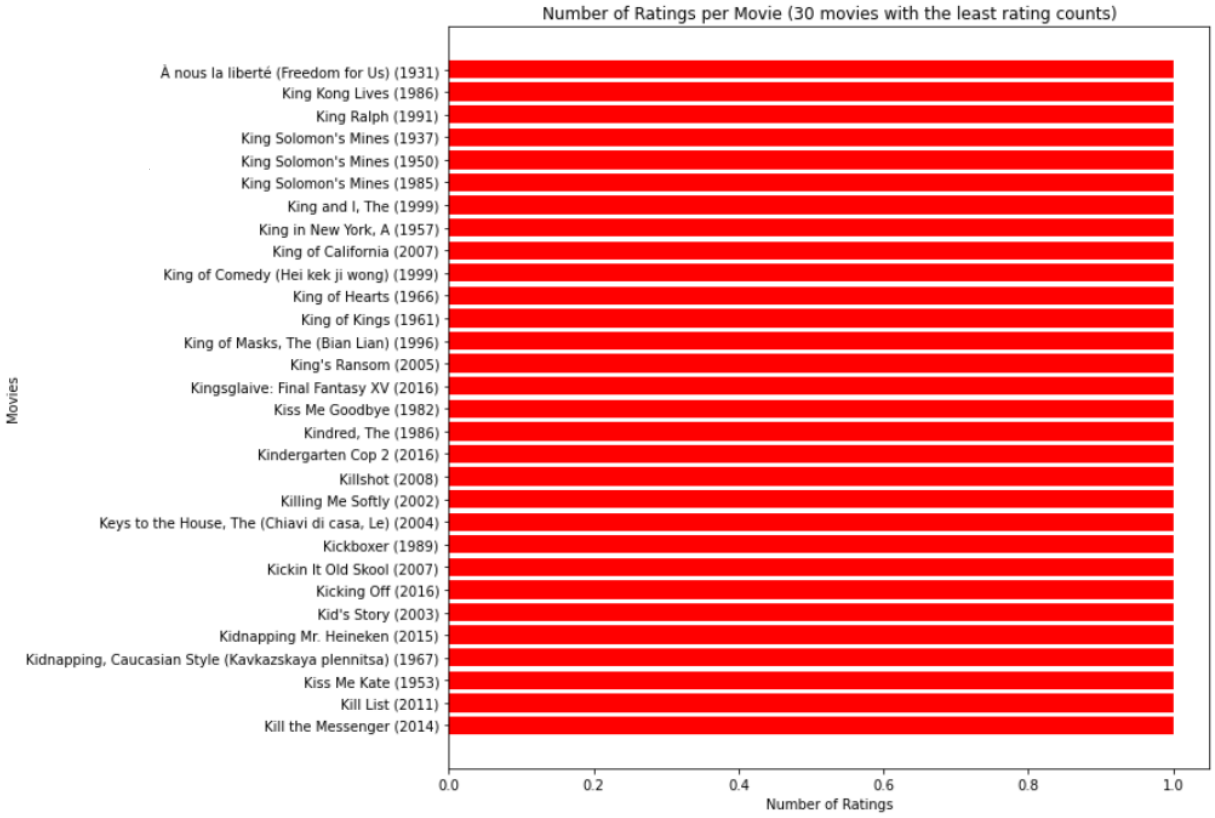


به نظر می‌رسد، فیلم‌هایی که بیشترین نظرات برای آن‌ها ثبت شده، غالباً دارای امتیازات بالایی (بیشتر از ۴) هستند. فیلم‌های مربوط به ۱۰ سال آخر قرن ۲۰ ام (مخصوصاً سال ۱۹۹۴)، نسبت به بقیه‌ی سال‌ها، توجه بینندگان را بیشتر به خود جلب کرده‌اند. و مشخصاً در آن سال‌ها، فیلم‌های خوبی ساخته شده‌است. همچنین، هرچقدر که فیلم‌ها قدیمی‌تر باشند، با توجه به این‌که سال‌های زیادی وجود داشته که کاربران به آن‌ها نظر بدهند، نظرات بیشتری برای آن‌ها ثبت شده، نسبت به فیلم‌های جدیدتر.

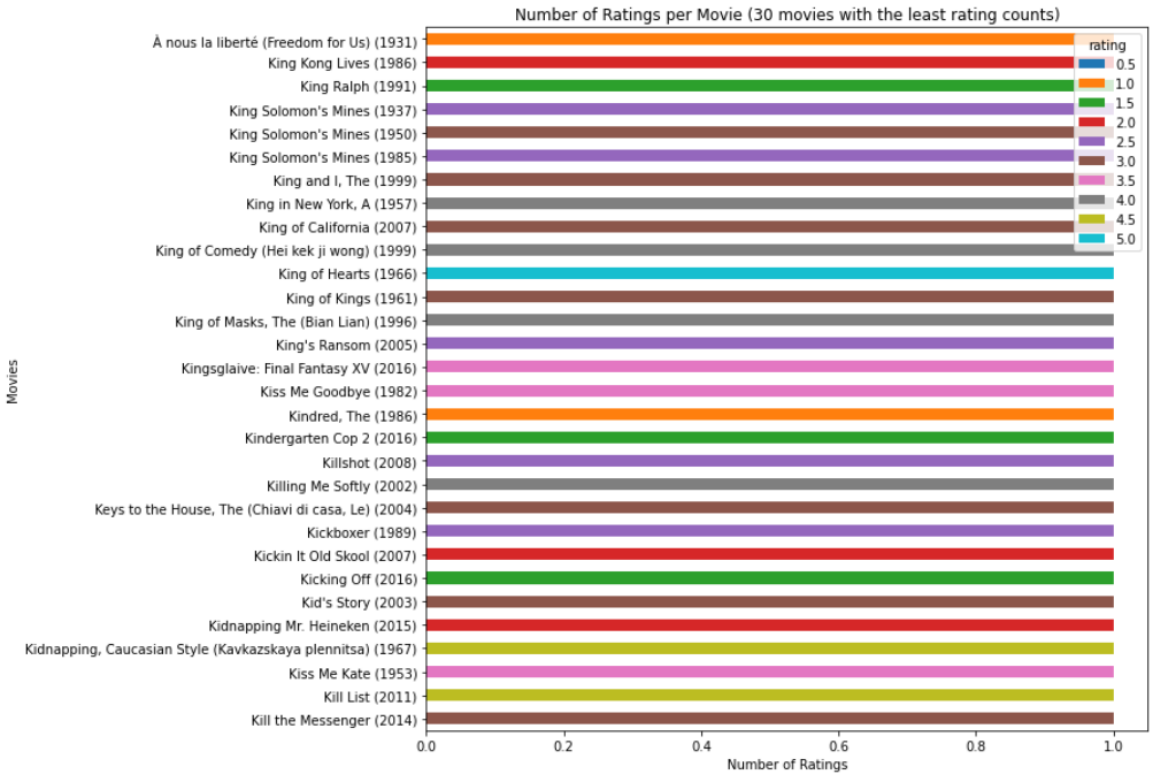
۳۰ فیلمی که بیشترین تعداد نظرات برای آن‌ها ثبت شده، از حدوداً ۲۰۰ تا ۳۲۹ نظر (فیلم forrest gump) به آن‌ها تخصیص یافته. ۱۰ فیلم با بیشترین تعداد نظرات، به شرح زیر هستند:

1. Forrest Gump (1994)
2. Shawshank Redemption, The (1994)
3. Pulp Fiction (1994)
4. Silence of the Lambs, The (1991)
5. Matrix, The (1999)
6. Star Wars: Episode IV - A New Hope (1977)
7. Jurassic Park (1993)
8. Braveheart (1995)
9. Terminator 2: Judgment Day (1991)
10. Schindler's List (1993)

- ۳۰ فیلم با کمترین تعداد نظر ثبت شده برای آن‌ها، در نمودار bar plot زیر نمایش داده شده‌اند: (آن ۱۸ فیلم که نظری برای آن‌ها ثبت نشده بود، در نظر گرفته نشده‌اند.)



- یک بار دیگر، این نمودار به صورت رنگ شده به تفکیک امتیازهای داده‌شده، به شکل زیر نمایش داده‌شد:



به نظر می‌رسد، فیلم‌هایی که کم‌ترین نظرات برای آن‌ها ثبت شده، غالباً دارای امتیازات ۳ و پایین‌تر هستند. و مشخصاً فیلم‌های پرطرفداری نیستند. برای همگی این فیلم‌ها، ۱ نظر ثبت شده‌است. و همچنان علاوه بر این ۳۰ فیلم، فیلم‌های دیگری هم هستند که تنها یک نظر برای آن‌ها ثبت شده‌است.

سوال ۲: (کد مربوط به این گزارشات، در فایل مربوطه ضمیمه شده‌است.)

- تعداد نظرات: به تعداد سطرهای دیتاست ratings است: **۱۰۰۸۳۶**

- تعداد فیلم‌های متمایز: **۹۷۴۲** مورد Id متمایز در دیتاست movies وجود دارد. اما **۹۷۳۷** عنوان متمایز برای فیلم‌ها وجود دارد و ۵ مورد از فیلم‌ها دوبار با دو Id مختلف در دیتاست ذخیره شده‌اند:

	movieId	title	genres
5601	26958	Emma (1996)	Romance
6932	64997	War of the Worlds (2005)	Action|Sci-Fi
9106	144606	Confessions of a Dangerous Mind (2002)	Comedy|Crime|Drama|Romance|Thriller
9135	147002	Eros (2004)	Drama|Romance
9468	168358	Saturn 3 (1980)	Sci-Fi|Thriller

- ۱۰ فیلمی که بیشترین نظر را داشته‌اند:

```
title
Forrest Gump (1994)                          329
Shawshank Redemption, The (1994)             317
Pulp Fiction (1994)                          307
Silence of the Lambs, The (1991)             279
Matrix, The (1999)                           278
Star Wars: Episode IV - A New Hope (1977)    251
Jurassic Park (1993)                         238
Braveheart (1995)                            237
Terminator 2: Judgment Day (1991)            224
Schindler's List (1993)                      220
Name: rating, dtype: int64
```

- تعداد کاربرانی که به فیلم forrest gump نظر داده‌اند: **۳۲۹** کاربر

سوال ۳:

- **الف:**
  - در حالتی که min_support=0.1 و min_length=2 قرار داده‌شد، ۵۰۶۱ مورد 2-itemset ، ۲۱۵۵۶ مورد 3-itemset، ۳۸۱۶۰ مورد 4-itemset، ۳۳۸۳۷ مورد 5-itemset، ۱۶۴۱۹ مورد 6-itemset، ۴۳۶۵ مورد 7-itemset، ۶۰۷ مورد 8-itemset، ۳۵ مورد 9-itemset و ۱ مورد 10-itemset کشف شد:

```
There are 5061 frequent itemsets with length 2 and minimum support of 0.1.
There are 21556 frequent itemsets with length 3 and minimum support of 0.1.
There are 38160 frequent itemsets with length 4 and minimum support of 0.1.
There are 33837 frequent itemsets with length 5 and minimum support of 0.1.
There are 16419 frequent itemsets with length 6 and minimum support of 0.1.
There are 4365 frequent itemsets with length 7 and minimum support of 0.1.
There are 607 frequent itemsets with length 8 and minimum support of 0.1.
There are 35 frequent itemsets with length 9 and minimum support of 0.1.
There are 1 frequent itemsets with length 10 and minimum support of 0.1.
```

  - در حالتی که min_support=0.2 و min_length=2 قرار داده‌شد، ۲۲۶ مورد 2-itemset ، ۹۳ مورد 3-itemset و ۱۰ مورد 4-itemset کشف شد:

```
There are 226 frequent itemsets with length 2 and minimum support of 0.2.
There are 93 frequent itemsets with length 3 and minimum support of 0.2.
There are 10 frequent itemsets with length 4 and minimum support of 0.2.
```

  - در حالتی که min_support=0.3 و min_length=2 قرار داده‌شد، ۱۱ مورد 2-itemset کشف شد:

```
There are 11 frequent itemsets with length 2 and minimum support of 0.3.
```

  - در حالتی که min_support=0.5 و min_length=2 قرار داده‌شد، هیچ itemset مکرری کشف نشد.

  - مقایسه: شروط min_support=0.5 یا min_support=0.3، محدودیت بالایی برای کشف itemsetها قرار داده‌اند. و موجب شده‌اند که تعداد کمی itemset مکرر کشف شود، و یا اصلاً کشف نشود. و شرط min_support=0.1 محدودیت پایینی برای کشف itemsetها قرار داده‌است. و این موجب شده تا تعداد بسیار زیادی از itemsetها به عنوان مکرر شناخته شوند. اما شرط min_support=0.2 محدودیت معقول‌تری را لحاظ کرده، و این موجب شده تا تعداد مناسبی از itemsetها به عنوان itemsetهای مکرر کشف شوند.

- **ب:**
  - در حالتی که min_support=0.5 یا min_support=0.3 قرار داده‌شد، تعداد itemsetهای مکرر قابل توجهی کشف نشد.
  - و در حالتی که min_support=0.1 قرار داده‌شد، تعداد بسیار زیادی از itemsetها، مکرر شناخته شدند. و این هم مناسب نیست؛ چرا که در این حالت، نمی‌توان نکته‌ی قابل توجهی، از این itemset ها دریافت کرد.
  - اما در حالتی که min_support=0.2 قرار داده‌شد، تعداد مناسبی itemset کشف شد. و می‌توان این min_support=0.2 را به عنوان مناسب ترین حالت برای min_support دانست.

- **ج:** با این روش نیز، تعداد itemset های مکرر دقیقاً مشابه روش قبل بدست آمد. فقط با توجه به این‌که این‌جا شرط min_length=2 لحاظ نشد، تعداد 1-itemset ها نیز گزارش شده‌است.

  - min_support=0.1:

	support	itemsets
0	0.539344	(Forrest Gump (1994))
1	0.503279	(Pulp Fiction (1994))
2	0.457377	(Silence of the Lambs, The (1991))
3	0.455738	(Matrix, The (1999))
4	0.411475	(Star Wars: Episode IV - A New Hope (1977))
...	...	...
120364	0.101639	(Jerry Maguire (1996), Independence Day (a.k.a...
120365	0.106557	(Jerry Maguire (1996), Star Wars: Episode IV -...
120366	0.104918	(Jerry Maguire (1996), Silence of the Lambs, T...
120367	0.114754	(Jerry Maguire (1996), Forrest Gump (1994))
120368	0.103279	(Jerry Maguire (1996), Star Wars: Episode VI -...

120369 rows × 2 columns

```
There are 328 frequent itemsets with length 1 and minimum support of 0.1.
There are 5061 frequent itemsets with length 2 and minimum support of 0.1.
There are 21556 frequent itemsets with length 3 and minimum support of 0.1.
There are 38160 frequent itemsets with length 4 and minimum support of 0.1.
There are 33837 frequent itemsets with length 5 and minimum support of 0.1.
There are 16419 frequent itemsets with length 6 and minimum support of 0.1.
There are 4365 frequent itemsets with length 7 and minimum support of 0.1.
There are 607 frequent itemsets with length 8 and minimum support of 0.1.
There are 35 frequent itemsets with length 9 and minimum support of 0.1.
There are 1 frequent itemsets with length 10 and minimum support of 0.1.
```

- min_support=0.2:

	support	itemsets
0	0.539344	(Forrest Gump (1994))
1	0.503279	(Pulp Fiction (1994))
2	0.457377	(Silence of the Lambs, The (1991))
3	0.455738	(Matrix, The (1999))
4	0.411475	(Star Wars: Episode IV - A New Hope (1977))
...	...	...
408	0.216393	(Star Wars: Episode IV - A New Hope (1977), Go...
409	0.216393	(Godfather, The (1972), Pulp Fiction (1994))
410	0.209836	(Forrest Gump (1994), Godfather, The (1972))
411	0.208197	(Godfather, The (1972), Silence of the Lambs, ...
412	0.204918	(Godfather: Part II, The (1974), Godfather, Th...

413 rows × 2 columns

```
There are 84 frequent itemsets with length 1 and minimum support of 0.2.
There are 226 frequent itemsets with length 2 and minimum support of 0.2.
There are 93 frequent itemsets with length 3 and minimum support of 0.2.
There are 10 frequent itemsets with length 4 and minimum support of 0.2.
```

- min_support=0.3:

	support	itemsets
0	0.539344	(Forrest Gump (1994))
1	0.503279	(Pulp Fiction (1994))
2	0.457377	(Silence of the Lambs, The (1991))
3	0.455738	(Matrix, The (1999))
4	0.411475	(Star Wars: Episode IV - A New Hope (1977))
5	0.390164	(Jurassic Park (1993))
6	0.388525	(Braveheart (1995))
7	0.360656	(Schindler's List (1993))
8	0.357377	(Fight Club (1999))
9	0.352459	(Toy Story (1995))
10	0.345902	(Star Wars: Episode V - The Empire Strikes Bac...
11	0.334426	(American Beauty (1999))
12	0.334426	(Usual Suspects, The (1995))
13	0.332787	(Seven (a.k.a. Se7en) (1995))
14	0.331148	(Independence Day (a.k.a. ID4) (1996))
15	0.327869	(Raiders of the Lost Ark (Indiana Jones and th...

```
There are 28 frequent itemsets with length 1 and minimum support of 0.3.
There are 11 frequent itemsets with length 2 and minimum support of 0.3.
```

- min_support=0.5:

	support	itemsets
0	0.539344	(Forrest Gump (1994))
1	0.503279	(Pulp Fiction (1994))
2	0.519672	(Shawshank Redemption, The (1994))

```
There are 3 frequent itemsets with length 1 and minimum support of 0.5.
```

سوال ۴:

- **الف:**
  - association rule های با min_support=0.3 و min_confidence=0.6، **۲۰** مورد هستند:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Braveheart (1995))	(Forrest Gump (1994))	0.388525	0.539344	0.300000	0.772152	1.431649	0.090451	2.021767
1	(Forrest Gump (1994))	(Jurassic Park (1993))	0.539344	0.390164	0.324590	0.601824	1.542489	0.114157	1.531573
2	(Jurassic Park (1993))	(Forrest Gump (1994))	0.390164	0.539344	0.324590	0.831933	1.542489	0.114157	2.740902
3	(Matrix, The (1999))	(Forrest Gump (1994))	0.455738	0.539344	0.318033	0.697842	1.293871	0.072233	1.524551
4	(Forrest Gump (1994))	(Pulp Fiction (1994))	0.539344	0.503279	0.377049	0.699088	1.389068	0.105609	1.650720
5	(Pulp Fiction (1994))	(Forrest Gump (1994))	0.503279	0.539344	0.377049	0.749186	1.389068	0.105609	1.836640
6	(Forrest Gump (1994))	(Shawshank Redemption, The (1994))	0.539344	0.519672	0.378689	0.702128	1.351097	0.098406	1.612529
7	(Shawshank Redemption, The (1994))	(Forrest Gump (1994))	0.519672	0.539344	0.378689	0.728707	1.351097	0.098406	1.697998
8	(Forrest Gump (1994))	(Silence of the Lambs, The (1991))	0.539344	0.457377	0.326230	0.604863	1.322461	0.079546	1.373253
9	(Silence of the Lambs, The (1991))	(Forrest Gump (1994))	0.457377	0.539344	0.326230	0.713262	1.322461	0.079546	1.606537
10	(Matrix, The (1999))	(Star Wars: Episode IV - A New Hope (1977))	0.455738	0.411475	0.300000	0.658273	1.599788	0.112475	1.722209
11	(Star Wars: Episode IV - A New Hope (1977))	(Matrix, The (1999))	0.411475	0.455738	0.300000	0.729084	1.599788	0.112475	2.008968
12	(Shawshank Redemption, The (1994))	(Pulp Fiction (1994))	0.519672	0.503279	0.363934	0.700315	1.391506	0.102395	1.657481
13	(Pulp Fiction (1994))	(Shawshank Redemption, The (1994))	0.503279	0.519672	0.363934	0.723127	1.391506	0.102395	1.734831
14	(Pulp Fiction (1994))	(Silence of the Lambs, The (1991))	0.503279	0.457377	0.339344	0.674267	1.474204	0.109156	1.665852
15	(Silence of the Lambs, The (1991))	(Pulp Fiction (1994))	0.457377	0.503279	0.339344	0.741935	1.474204	0.109156	1.924795
16	(Shawshank Redemption, The (1994))	(Silence of the Lambs, The (1991))	0.519672	0.457377	0.326230	0.627760	1.372522	0.088543	1.457724
17	(Silence of the Lambs, The (1991))	(Shawshank Redemption, The (1994))	0.457377	0.519672	0.326230	0.713262	1.372522	0.088543	1.675143
18	(Star Wars: Episode V - The Empire Strikes Bac...	(Star Wars: Episode IV - A New Hope (1977))	0.345902	0.411475	0.311475	0.900474	2.188403	0.169145	5.913271
19	(Star Wars: Episode IV - A New Hope (1977))	(Star Wars: Episode V - The Empire Strikes Bac...	0.411475	0.345902	0.311475	0.756972	2.188403	0.169145	2.691454

این جدول، بدین شکل تفسیر می‌شود که هر سطر، مربوط به یک association rule به فرم زیر است:

$X \rightarrow Y$

که فیلد antecedents، همان X بوده، و فیلد consequents همان Y است.

- سه association rule برتر از نظر lift عبارتند از:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
19	(Star Wars: Episode IV - A New Hope (1977))	(Star Wars: Episode V - The Empire Strikes Bac...	0.411475	0.345902	0.311475	0.756972	2.188403	0.169145	2.691454
18	(Star Wars: Episode V - The Empire Strikes Bac...	(Star Wars: Episode IV - A New Hope (1977))	0.345902	0.411475	0.311475	0.900474	2.188403	0.169145	5.913271
11	(Star Wars: Episode IV - A New Hope (1977))	(Matrix, The (1999))	0.411475	0.455738	0.300000	0.729084	1.599788	0.112475	2.008968

- **ب:**
  - association rule های با min_support=0.3 و min_confidence=0.8، **۲** مورد هستند:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(Jurassic Park (1993))	(Forrest Gump (1994))	0.390164	0.539344	0.324590	0.831933	1.542489	0.114157	2.740902
1	(Star Wars: Episode V - The Empire Strikes Bac...	(Star Wars: Episode IV - A New Hope (1977))	0.345902	0.411475	0.311475	0.900474	2.188403	0.169145	5.913271

این جدول، همچنان، به همان شکلی که در بخش قبل گفته‌شد، تفسیر می‌شود.

  - این association rule ها از نظر lift عبارتند از:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
1	(Star Wars: Episode V - The Empire Strikes Bac...	(Star Wars: Episode IV - A New Hope (1977))	0.345902	0.411475	0.311475	0.900474	2.188403	0.169145	5.913271
0	(Jurassic Park (1993))	(Forrest Gump (1994))	0.390164	0.539344	0.324590	0.831933	1.542489	0.114157	2.740902

  - مقایسه: تعداد association rule ها در این حالت نسبت به حالت قبل، (یک دهم) کم‌تر شد. دلیل این اتفاق این است که محدودیت min_confidence را از ۰٫۶ به ۰٫۸ افزایش دادیم و محدودیت بیشتری را برای کشف association rule ها در نظر گرفتیم.
اگر یک association rule به فرم $X \rightarrow Y$ در نظر بگیریم، در بخش قبل، باید ۶۰ درصد از مواقع، اگر X در یک تراکنش ظاهر شده‌بود، Y هم ظاهر می‌شد، تا آن association rule کشف شود. اما در این بخش، باید ۸۰ درصد از مواقع، اگر X در یک تراکنش ظاهر شده‌بود، Y هم ظاهر شود، تا آن association rule کشف شود.

❖ تمرین تشریحی امتیازی:

سوال ۱:

۱. chi-squared:

- expected for cell [i][j] $= \frac{row\ i\ total * column\ i\ total}{\text{total of total}}$
  - ✓ expected for cell [date][$milk$] $= \frac{200 * 850}{10000} = \frac{170000}{10000} = 17$
  - ✓ expected for cell [date][$not\ milk$] $= \frac{9800 * 850}{10000} = \frac{8330000}{10000} = 833$
  - ✓ expected for cell [not date][$milk$] $= \frac{200 * 9150}{10000} = \frac{1830000}{10000} = 183$
  - ✓ expected for cell [not date][$not\ milk$] $= \frac{9800 * 9150}{10000} = \frac{89670000}{10000} = 8967$

- $\chi 2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \frac{(50-17)^2}{17} + \frac{(800-833)^2}{833} + \frac{(150-183)^2}{183}$

$$+\frac{(9000-8967)^2}{8967} = \frac{(33)^2}{17} + \frac{(33)^2}{833} + \frac{(33)^2}{183} + \frac{(33)^2}{8967} = \frac{1089}{17}$$

$$+\frac{1089}{833} + \frac{1089}{183} + \frac{1089}{8967} = 64.058 + 1.307 + 5.950 + 0.121 = 71.428$$

درجه آزادی این contingency table به شرح زیر است:

Degree of freedom = (number of rows – 1) * (number of columns – 1)
= (2-1)*(2-1) = 1

مقدار آلفا را برابر ۰,۰۵ در نظر می‌گیریم.

با در نظر گرفتن این دو پارامتر، با استفاده از جدول توزیع chi-squared، مقدار بحرانی را بدست می‌آوریم: ۳,۸۴۱

حال با استفاده از مقایسه‌ی chi-squared بدست آمده و نقطه‌ی بحرانی، می‌توان درباره‌ی رابطه‌ی این خرید خرما و شیر نظر داد:

در این مسائل، ما ابتدا یک فرض اولیه در نظر می‌گیریم. فرض اولیه در این سوال عبارت است از: هیچ ارتباطی میان خریدن خرما و شیر وجود ندارد.

با توجه به این‌که مقدار chi-squared بزرگ‌تر از مقدار بحرانی است، فرض اولیه رد می‌شود. و نمی‌توان ادعا کرد که هیچ ارتباطی میان این دو وجود ندارد. اما لزوماً نمی‌توان گفت که رابطه‌ای میان این دو وجود دارد.

۲. lift:

- ✓ $P(date) = \frac{850}{10000} = 0.085$
- ✓ $P(milk) = \frac{200}{10000} = 0.02$
- ✓ $P(\text{date} \cup \text{milk}) = \frac{50}{10000} = 0.005$

- $\text{lift}(\text{date}.\text{milk}) = \frac{P(\text{date} \cup \text{milk})}{P(\text{date})\,P(\text{milk})} = \frac{0.005}{0.02*0.085} = \frac{0.005}{0.0017} = 2.941$

با توجه به این‌که مقدار lift بزرگ‌تر از ۱ بدست آمد، این را می‌توان نتیجه گرفت که یک رابطه‌ی مثبت بین خریدن شیر و خریدن خرما وجود دارد. یعنی احتمال این‌که وقتی یکی از این دو خریده می‌شود، دیگری هم خریداری شده‌باشد، بالاست.

۳. all-confidence:

- ✓ $sup(\{date\}) = \frac{transactions\ containing\ date}{\text{Total}} = \frac{850}{10000} = 0.085$
- ✓ $sup(\{milk\}) = \frac{transactions\ containing\ milk}{\text{Total}} = \frac{200}{10000} = 0.02$
- ✓ $sup(\{date.milk\}) = \frac{transactions\ containing\ date\ and\ milk}{\text{Total}} = \frac{50}{10000} = 0.005$

فرمول بیان شده در صورت سوال به طور کامل توضیح داده نشده‌است. و در اینترنت هم منبعی برای خواندن درباره‌ی این پارامتر پیدا نکردم. برداشت من این بود که باید X را برابر با itemset ای در نظر بگیریم که هر دو خرما و شیر را شامل می‌شود. مطابق با این برداشت، فرمول زیر محاسبه شد:

- $\text{all} - \text{conf}(\text{X}) = \frac{\sup(\text{X})}{\max-\text{item}-\sup(\text{X})} = \frac{\sup(\text{X})}{\max\{\sup(i_j) \mid i_j \in \text{X}\}}$

$$= \frac{\sup(\{\text{date}.\text{milk}\})}{\max\{\sup(\{\text{date}\}).\sup(\{\text{milk}\}).\sup(\{\text{date}.\text{milk}\})\}}$$

$$= \frac{0.005}{\max\{0.085 \text{ and } 0.02 \text{ and } 0.005\ \}} = \frac{0.005}{0.085} = 0.058$$

اگر این عدد نزدیک به ۱ باشد، نشان‌گر این است که ارتباطی بین این دو وجود دارد. و اگر نزدیک به ۰ باشد، نشان‌گر ارتباط نداشتن این دو با یک‌دیگر است. این‌جا این عدد بسیار نزدیک به ۰ است؛ پس می‌توان نتیجه گرفت که خرید شیر و خرید خرما ارتباطی با یکدیگر ندارند.

- در کل، خریدن خرما و خریدن شیر ارتباطی با یکدیگر ندارند.