1.(a) A five-number summary of a distribution contains minimum, Q1, median, Q3, maximum, where Q1 is 25th percentile, Q3 is 75th percentile. It gives a rough description of the distribution of data.

(b) Data represented with a box is called box plot. It is used in explanatory data analysis and could visually show the distribution of numerical data and skewness.

Box plot is graphic display of five-number summary.

(i)Q1, Q3, IQR: Box starts at Q1(25th percentile) and ends at Q3(75th percentile). The height of box is IQR=Q3-Q1.

(ii)Median is marked by a line within the box.

(iii)Two lines outside the box extended to Minimum and Maximum.

(iv)Outliers: points beyond a specified outlier threshold, plotted individually, usually higher||lower than 1.5IQR

(c)Yes, they can.

Reason: Box plot gives a rough description of data distribution, which only contains five-number summary. If two different datasets have the same five-number summary, then the box plot of them are the same.

Example: A={-2,-2,-1,0,0,1,2,2} B={-2,-1.5,-1,0,0,1,2,2} For A and B, min=-2, Q1=-1, median=0, Q3=2, max=2. Therefore, the box plot of A and B are the same. However, A and B are different. Therefore, two different distributions can have the exact same box plot.

(d)(i)Yes, it can.

Reason: Q-Q plot graphs the quantiles of one univariate distribution against the corresponding quantiles of another. If the price of A is always larger than the price of B at the same percentile, then the line is below y=x.

Example: A follows the distribution of $p_a = (a+1)^{per} - 1$, where per is the percentile variable, ranging from 0 to 1. B follows the distribution of $p_b = log_{a+1}(per+1)$. The Q-Q plot in this case stays entirely below the y = x line except for the endpoints (0,0) and (a,a).

(ii)No, I disagree.

Reason: If (m_a, m_b) lives above the y=x line, then $m_a < m_b$. Median and mean are different: median is the value of the middle distribution, which is highly connected with percentile; mean of x is $\frac{1}{n}\sum_{i=1}^{i=n} x_i$. If the data of lower or higher percentile of A is mainly around higher of B, then the hypothesis is wrong.

Example: A={$\frac{a}{4}$,$\frac{a}{4}$,$\frac{a}{4}$,a,a} B={0,0,$\frac{a}{2}$,a,a}, where $\mu_B = \frac{a}{2} = \frac{10a}{20} < \frac{11a}{20} = \mu_A$ and $m_B = \frac{a}{2} > \frac{a}{4} = m_A$. (m_A, m_B) lives above the y=x line but $\mu_B < \mu_A$, which disobeys the hypothesis.

2.(a)

	BUY DIAPER	NOT BUY DIAPER	TOTAL
BUY BEER	200	400	600
NOT BUY BEER	200	20,000	20,200
TOTAL	400	20,400	20,800

$$Lift(A,B) = \frac{s(A \cup B)}{s(A) \times s(B)} = \frac{200/20800}{400/20800 \times 600/20800} \approx 17.333$$

$$Jaccard(A,B) = \frac{s(A \cup B)}{s(A) + s(B) - s(A \cup B)} = \frac{200/20800}{200/20800 + 200/20800 + 400/20800} = 0.$$

$$Cosine(A,B) = \frac{s(A \cup B)}{\sqrt{(s(A) \times s(B))}} = \frac{200/20800}{\sqrt{400/20800 \times 600/20800}} \approx 0.408$$

$$Kulczynski(A,B) = \frac{1}{2}(\frac{s(A \cup B)}{s(A)} + \frac{s(A \cup B)}{s(B)}) = \frac{1}{2}(\frac{200/20800}{400/20800} + \frac{200/20800}{600/20800}) \approx 0.417$$

(b)Yes, I agree.

Suppose A,B obey the following rules.

	B	B'	TOTAL
A	a	b	a+b
A'	c	d	c+d
TOTAL	a+c	b+d	a+b+c+d

$$P(B|A) = \frac{s(A \cup B)}{s(A)} = \frac{a/(a+b+c+d)}{(a+b)/(a+b+c+d)} = \frac{a}{a+b}$$

$$P(A|B) = \frac{s(A \cup B)}{s(B)} = \frac{a/(a+b+c+d)}{(a+c)/(a+b+c+d)} = \frac{a}{a+c}$$

Both P(B|A) and P(A|B) don't contain any "d" . Therefore, they are null-invariant.

Inputs of the function $I(A, B) = f(P(B|A), P(A|B))$ are $\frac{a}{a+b}$ and $\frac{a}{a+c}$, which don't contain any "d".

Since the total number (a+b+c+d) is also unknown, d could not be created by minus: (a+b+c+d)-a-b-c.

Therefore, d should never appears in $I(A, B)$. Therefore, I is null-invariant.

3.(a)

1st scan

C1

ITEMSET	RELATIVE SUPPORT
{A}	0.8
{B}	0.6
{C}	0.8
{D}	0.8
~~{E}~~	~~0.4~~
~~{F}~~	~~0.2~~
~~{G}~~	~~0.2~~
~~{H}~~	~~0.4~~
~~{J}~~	~~0.4~~
~~{K}~~	~~0.2~~

F1

ITEMSET	RELATIVE SUPPORT
{A}	0.8
{B}	0.6
{C}	0.8

ITEMSET	RELATIVE SUPPORT
{D}	0.8

2nd scan

C2

ITEMSET	RELATIVE SUPPORT
~~{A,B}~~	~~0.2~~
{A,C}	0.6
{A,D}	0.6
{B,C}	0.6
{B,D}	0.6
{C,D}	0.6

F2

ITEMSET	RELATIVE SUPPORT
{A,C}	0.6
{A,D}	0.6
{B,C}	0.6
{B,D}	0.6
{C,D}	0.6

3rd scan

C3

ITEMSET	RELATIVE SUPPORT
~~{A,C,D}~~	~~0.4~~
{B,C,D}	0.6

F3

ITEMSET	RELATIVE SUPPORT
{B,C,D}	0.6

(ii)

for (b,c)->(d), $confidence = \frac{sup(former \bigcap latter)}{sup(former)} = \frac{sup(b,c,d)}{sup(b,c)} = 1$

for (b,d)->(c), $confidence = \frac{sup(former \bigcap latter)}{sup(former)} = \frac{sup(b,c,d)}{sup(b,d)} = 1$

for (c,d)->(b), $confidence = \frac{sup(former \bigcap latter)}{sup(former)} = \frac{sup(b,c,d)}{sup(c,d)} = 1$

(b)(i) Constrain 1: average(price)>50 is Convertible.

Reason: After proper ordering the prices of all items, data can either be monotone or anti-monotone. For example, if the prices of items are ascending order(price increases as number of items increases), then it is anti-monotone. If a set violates constrain, then all supersets violates.

Mine measures: 1. sort data 2. scan from the smallest set, if any of the average value is less than $50, then stop the scan

(ii)Constrain 2: profit>$10 is data Succinct.

A data succinct constraint means that data space can be pruned at the initial pattern mining process. Found the profit over $10 is very fast and easy.

Mine measure: if(profit>10) then(keep) else(delete)

Constrain 3: sum(price)>100 is pattern monotonic and data anti-monotonic constraint.

Reason: Pattern monotone: if and itemset S satisfies the constraint c, so does any of its superset. Data anti-monotone constraint means that in the mining process, if a data entry t cannot contribute to a pattern p satisfying c, t cannot contribute to p's superset either.

Mine measures: scan and add the profit whose price is greater than 10 , if the sum is less than 100, then stop.

4.(a)

The algorithm in slides is much faster. When deriving deriving C_k from F_{k-1}. Algorithm in slides **first sort items according to their (k-2) prefix**, then do the prefix match. While the code only do the prefix match according to the length of item.

code: compare abc and ade--abcde(not match)

compare abc and acd----abcd(match)

slides: compare abc and acd---abcd(not match)

compare abc and abd----abcd(match)

(b)

```
for every i in F_(k-1):
    for every j in F_(k-1):
        if(i.item[1]=j.item[2],...,i.item[k-2]=i.item[k-
2],i.item[k-1]<j.item[k-1]){
        c=join(i,j)
        if(has_infrequent_subset(c,F_(k-1)))
            continue
        else
            add c to C_k
        }
```

5.(a)

PATTERN	SUPPORT
a	3
b	5
c	4
d	3
e	3
f	2

(b)

PREFIX	PROJECTED DATABASE
b	<(ac)>,<(fg)>,< f >,<cb(ade)>

bb is a length-2 frequent pattern in SDB1.

The support of bb is 4.

(c)

PREFIX	PROJECTED DATABASE
<(bd) >	<cb(ac)>,<bcb(ade)>

<(bd)> is a length-2 frequent pattern in SDB1.

The support of <(bd)> is 2.

(d)

PREFIX	PROJECTED DATABASE
< b >	<(_d)cb(ac)>,<(_ f)(ce)b(fg)>,<(_f)abf>,<(_e)(ce)d>,<(_d)bcb(ade)>

(e)

PREFIX	SEQUENTIAL PATTERN
< b >	< b >, < ba >, < bb >, < bba >, < bbc >, < bbf >, < bc >, < bca >, < bcb >, < bcba >, < bcd>, <b(ce)>, < bd>, <(bd)>, <(bd)a>, <(bd)b>, <(bd)ba>, <(bd)bc>, <(bd)c>, <(bd)ca>, <(bd)cb>, <(bd)cba>, , , <(bf)>, <(bf)b>, <(bf)bf>, <(bf)f>