1. (a) (3 points) Describe what a five-number summary of a distribution is.

   The five number summary of a distribution consists of the minimum value, first quartile, median value, third quartile, and maximum value.

   (b) (3 points) Describe what boxplots are and explain how boxplots incorporate the five-number summary.

   A boxplot incorporates the five-number summary as follows:

   - Typically, the ends of the box are at the quartiles so that the box length is the interquartile range.
   - The median is marked by a line within the box.
   - Two lines (called $whiskers$) outside the box extend to the smallest ($Minimum$) and largest ($Maximum$) observations.

   (c) Consider a distribution without outliers. If they have the exact same boxplot, only the minimum value, first quartile, median value, third quartile, and maximum value are the same. However, we do not know the distributions between those numbers, which means that we do not know the distribution between the minimum value and first quartile, the distribution between first quartile and the median, etc. Hence, consider two distributions without outliers. They have the same the minimum value, first quartile, median value, third quartile, and maximum value but different distributions between the minimum value and first quartile. Clearly, those different distributions have the same boxplot.

   (d) A quantile plot is a graphical method used to show the approximate percentage of values below or equal to the independent variable in a univariate distribution. It displays quantile information for all the data, where the values measured for the independent variable are plotted against their corresponding quantile.

   (e) The quantile-quantile plot is a graphical technique for determining if two data sets come from populations with a common distribution. It graphs the quantiles of one univariate distribution against the corresponding quantiles of another univariate distribution. Both axes display the range of values measured for their corresponding distribution, and points are plotted that correspond to the two distributions' quantile values. A line ($y = x$) can be added to the graph along with points representing where the first, second, and third quantiles lie to increase the graph's informational value. Points that lie above such a line indicate a correspondingly higher value for the distribution plotted on the y-axis than for the distribution plotted on the x-axis at the same quantile. The opposite effect is true for points lying below this line.

   (f) A quantile plot displays quantile information for all the data, where the values measured for the independent variable are plotted against their corresponding quantile. However, a quantile-quantile plot graphs the quantiles of one univariate distribution against the corresponding quantiles of another univariate distribution. Both axes display the range of values measured for their corresponding distribution, and points are plotted that correspond to the quantile values of the two distributions. A line ($y = x$) can be added to the graph along with points representing where the first, second and third quantiles lie, in order to increase the graph's informational value. Points that lie above such a line indicate a correspondingly higher value for the distribution plotted on the y-axis, than for the distribution plotted on the x-axis at the same quantile. The opposite effect is true for points lying below this line.

2. (a) (3 points) Under the null hypothesis, i.e., 'Buy Beer' and 'Buy Diaper' are independent, what is the expected number for 'Buy Beer' and 'Buy Diaper'?

The expected number for 'Buy Beer' and 'Buy Diaper' is

$$E = \frac{\text{row total} \times \text{column total}}{\text{grand total}} = \frac{(100+400)\times(100+300)}{100+400+300+200} = 200.$$

(b) (3 points) Under the null hypothesis, i.e., 'Buy Beer' and 'Buy Diaper' are independent, what is the expected number for 'Buy Beer' and 'Not But Diaper'?

The expected number for 'Buy Beer' and 'Not Buy Diaper' is

$$E = \frac{\text{row total} \times \text{column total}}{\text{grand total}} = \frac{(100+400)\times(400+200)}{100+400+300+200} = 300.$$

(c) (4 points) What is the χ^2 statistic for the contingency table? Show steps of your calculation.

The expected number for 'Buy Diaper' and 'Not Buy Beer' is

$$E = \frac{\text{row total} \times \text{column total}}{\text{grand total}} = \frac{(100+300)\times(300+200)}{100+400+300+200} = 200.$$

The expected number for 'Not Buy Diaper' and 'Not Buy Beer' is

$$E = \frac{\text{row total} \times \text{column total}}{\text{grand total}} = \frac{(300+200)\times(400+200)}{100+400+300+200} = 300.$$

The observed value for 'Buy Beer' and 'Buy Diaper' is 100 and the expected value for 'Buy Beer' and 'Buy Diaper' is 200. Then

$$\frac{(100-200)^2}{200} = 50.$$

The observed value for 'Buy Beer' and 'Not Buy Diaper' is 400 and the expected value for 'Buy Beer' and 'Not Buy Diaper' is 300. Then

$$\frac{(400-300)^2}{300} = 100/3.$$

The observed value for 'Buy Beer' and 'Not Buy Beer' is 300 and the expected value for 'Buy Beer' and 'Not Buy Diaper' is 200. Then

$$\frac{(300-200)^2}{200} = 50.$$

The observed value for 'Not Buy Beer' and 'Not Buy Beer' is 200 and the expected value for 'Buy Beer' and 'Not Buy Diaper' is 300. Then

$$\frac{(200-300)^2}{300} = 100/3.$$

Hence, the χ^2 statistic for the contingency table is $50 + 100/3 + 50 + 100/3 = 500/3$.

(d) (4 points) At a significance level of $\alpha = 0.05$, are these two variables 'Buy Beer' and 'Buy Diaper' independent? Explain your answer.

The degree of freedom is

$$\text{df} = \text{number of rows - 1} \times \text{number of columns - 1} = (2-1) \times (2-1) = 1.$$

and the significance level is $\alpha = 0.05$.
Then from the table, we know that $\chi^2_{0.05}(1) = 3.841$.
Since $500/3 > 3.841$, we reject the null hypothesis that 'Buy Beer' and 'Buy Diaper' are independent.

(e) (4 points) Consider an updated contingency table where the entry for 'Not Buy Beer' and 'Not Buy Diaper' is 20,000 instead of 200, and all other entries are the same. What is the χ^2 statistic for this updated contingency table? Show steps of your calculation.
The expected number for 'Buy Diaper' and 'Buy Beer' is

$$E = \frac{\text{row total} \times \text{column total}}{\text{grand total}} = \frac{(100+400) \times (100+300)}{100+400+300+20000} = 9.6154.$$

The observed value for 'Buy Beer' and 'Buy Beer' is 100 and the expected value for 'Buy Beer' and 'Not Buy Diaper' is 9.6154. Then

$$\frac{(100-9.6154)^2}{9.6154} \approx 849.6137.$$

The expected number for 'Not Buy Diaper' and 'Buy Beer' is

$$E = \frac{\text{row total} \times \text{column total}}{\text{grand total}} = \frac{(100+400) \times (400+20000)}{100+400+300+20000} = 490.3846.$$

The observed value for 'Not Buy Beer' and 'Buy Beer' is 400 and the expected value for 'Buy Beer' and 'Not Buy Diaper' is 490.2846. Then

$$\frac{(400-490.3846)^2}{490.2846} \approx 16.6591.$$

The expected number for 'Buy Diaper' and 'Not Buy Beer' is

$$E = \frac{\text{row total} \times \text{column total}}{\text{grand total}} = \frac{(300+20000) \times (100+300)}{100+400+300+20000} = 390.3846.$$

The observed value for 'Buy Diaper' and 'Not Buy Beer' is 300 and the expected value for 'Buy Beer' and 'Not Buy Diaper' is 390.3846. Then

$$\frac{(300-390.3846)^2}{390.3846} \approx 20.9265.$$

The expected number for 'Not Buy Beer' and 'Not Buy Diaper' is

$$E = \frac{\text{row total} \times \text{column total}}{\text{grand total}} = \frac{(300+20000) \times (400+20000)}{100+400+300+20000} = 19909.6154.$$

The observed value for 'Not Buy Beer' and 'Not Buy Beer' is 20000 and the expected value for 'Buy Beer' and 'Not Buy Diaper' is 300. Then

$$\frac{(20000 - 19909.6154)^2}{19909.6154} \approx 0.4103.$$

Hence, the chi_2 statistics for this updated contingency table is $849.6137 + 16.6591 + 20.9265 + 0.4103 = 887.6096$.

(f) (4 points) For the updated contingency table, at a significance level of $\alpha = 0.05$, are these two variables 'Buy Beer' and 'Buy Diaper' independent? Explain your answer.

The degree of freedom is

$$\text{df} = \text{number of rows - 1} \times \text{number of columns - 1} = (2-1) \times (2-1) = 1.$$

and the significance level is $\alpha = 0.05$.

Then from the table, we know that $\chi^2_{0.05}(1) = 3.841$.

Since $887.5762 > 3.841$, we reject the null hypothesis that 'Buy Beer' and 'Buy Diaper' are independent.

3. (a) i. (6 points) What is the frequent k-itemset for the largest k? Explain your answer. If there are more than one, it is sufficient to mention (and explain) only one.

We use Apriori algorithm to find all frequent itemsets.

$C_1 = \{A:4, B:3, C:4, D:4, E:2, F:1, G:1, H:2\}$

$L_1 = \{A, B, C, D\}$

$C_2 = \{AB:2, AC:3, AD:3, BC:3, BD:3, CD:3\}$

$L_2 = \{AC, AD, BC, BD, CD\}$

$C_3 = \{ACD:2, BCD:3\}$

$L_3 = \{BCD\}$

$C_4 = \emptyset$

$L_4 = \emptyset$

We observe that the largest k is 3 since $\{BCD\}$ is a 3-itemset and there is no frequent superset of $\{BCD\}$.

ii. Since a strong association rule satisfies both a minimum support threshold and a minimum confidence threshold, we need sup{item1, item2, item3} $\geq 0.6 \times 5 = 3$, which indicates that we should only consider $\{B, C, D\}$.

When item1 is B, item2 is C, and item3 is D, sup{{B,C}, D}/sup{B,C} = 5/5 = $1 \geq min\_conf$. Note that the support is $3/5 = 0.6 > min\_sup$.

When item1 is B, item2 is D, and item3 is C, sup{{B,D}, C}/sup{B,D} = 5/5 = $1 \geq min\_conf$. Note that the support is $3/5 = 0.6 > min\_sup$.

When item1 is C, item2 is D, and item3 is B, sup{{C,D}, B}/sup{C,D} = 5/5 = $1 \geq min\_conf$. Note that the support is $3/5 = 0.6 > min\_sup$.

We conclude that there are three strong association rules, which are

$\forall x \in transaction,\ \ buys(x, B) \wedge buys(x, C) \Rightarrow buys(x, D)$. [60%, 100%]

$\forall x \in transaction,\ \ buys(x, B) \wedge buys(x, D) \Rightarrow buys(x, C)$. [60%, 100%]

$\forall x \in transaction,\ \ buys(x, C) \wedge buys(x, D) \Rightarrow buys(x, B)$. [60%, 100%]

(b) (i) (6 points) The average price of all the items in each pattern is greater than \$50. The constraint is convertible. It can be mined efficiently using FP-growth as follows.

- All the frequent items are listed in price descending order.
- If the average price of items in the pattern is less than or equal to \$50, we prune the pattern as well as its conditional pattern base.

(ii) (6 points) The sum of the price of all the items with profit over \$5 in each pattern is at least \$200.

The constraint "profit over \$5" is succinct, whereas the constraint "sum of the prices is at least \$200 in total" is monotonic and data antimonotonic. It can be mined efficiently using FP-growth as follows.

- Start with only items whose prices < 5, and remove the patterns with only low prices.
- Only derive conditional pattern bases and FP-trees for frequent items from the global FP-tree and mine them recursively. Other items should be excluded from these conditional pattern bases and FP-trees.
- Once a pattern with sum of the price is at least \$200, no further constraint checking for total price is needed in recursive mining.
- If the sum of the price of items in the pattern and the frequent ones in the pattern base is less than \$200, we prune the pattern as well as its conditional pattern base.
- If the sum of the price of items in the pattern is greater than \$200, we prune the pattern as well as its conditional pattern base.

4. (a) (12 points) Frequent subsequences starting with `a`.
From the projected databases of $\langle a \rangle$, we find that $\langle b \rangle : 4, \langle c \rangle : 4$ are frequent.
From the projected databases of $\langle ab \rangle$, we find that there is no frequent items. Stop.
From the projected databases of $\langle ac \rangle$, we find that $\langle b \rangle : 3, \langle c \rangle : 3$ are frequent.
From the projected databases of $\langle acb \rangle$, we find that there is no frequent items. Stop.
From the projected databases of $\langle acc \rangle$, we find that there is no frequent items. Stop.
Hence, the frequent subsequences are $\langle a \rangle$: $\langle a \rangle$, $\langle ab \rangle$, $\langle ac \rangle$, $\langle acc \rangle$, $\langle acb \rangle$.

(b) (6 points) Frequent subsequences starting with `c`.
From the projected databases of $\langle c \rangle$, we find that $\langle b \rangle : 3, \langle c \rangle : 3$ are frequent.
From the projected databases of $\langle cb \rangle$, we find that there is no frequent items. Stop.
From the projected databases of $\langle cc \rangle$, we find that there is no frequent items. Stop.
Hence, the frequent subsequences are $\langle c \rangle$: $\langle c \rangle$, $\langle cb \rangle$, $\langle cc \rangle$.

(c) (6 points) Frequent subsequences starting with `d`.
From the projected databases of $\langle d \rangle$, we find that $\langle c \rangle : 3$ are frequent.
From the projected databases of $\langle dc \rangle$, we find that there is no frequent items. Stop.
Hence, the frequent subsequences are $\langle d \rangle$: $\langle d \rangle$, $\langle dc \rangle$.

(d) (3 points) Frequent subsequences starting with `b`.
From the projected databases of $\langle b \rangle$, we find that $\langle c \rangle : 3$ are frequent.

From the projected databases of $\langle bc \rangle$, we find that there is no frequent items. Stop.

Hence, the frequent subsequences are $\langle b \rangle$: $\langle b \rangle$, $\langle bc \rangle$.

(e) (3 points) Frequent subsequences starting with `e`.

From the projected databases of $\langle e \rangle$, we find that there is no frequent items. Stop.

Hence, the frequent subsequences are $\langle e \rangle$: $\langle e \rangle$.

(f) (3 points) Frequent subsequences starting with `f`.

From the projected databases of $\langle f \rangle$, we find that there is no frequent items. Stop.

Hence, the frequent subsequences are $\langle f \rangle$: $\langle f \rangle$.

(g) (3 points) Are there other frequent subsequences in the database not covered by the above? If your answer is 'yes', list one such frequent subsequence. If your answer is 'no', clearly explain why not.

No. If there is some subsequence that is not covered by the above sequence, then all of its prefixes are not covered by the above since we use PrefixSpan.

Then frequent length-1 prefix ($\langle a \rangle$,$\langle c \rangle$,$\langle d \rangle$,$\langle b \rangle$,$\langle e \rangle$,$\langle f \rangle$) in this case will also not be covered. However, since we covered those frequent length-1 prefix, we conclude there is no frequent subsequences in the database not covered by the above.

To further illustrate why this is related to length-1 prefix, take $\{acc\}$ as an example. If $\{acc\}$ is not covered by the above, then $\{ac\}$ is not covered by the above since if $\{ac\}$ is covered, we will check the frequency of $\{acc\}$ and cover it. Similarly, $\{a\}$ is not covered, which is a contradiction.