Question 1 Data Mining in an Online Game

1b:
-1 for mentioning the Account and didn't explain how to use it
-1 didn't mention Alliances or User Interaction
1c:
-2 no example of metadata

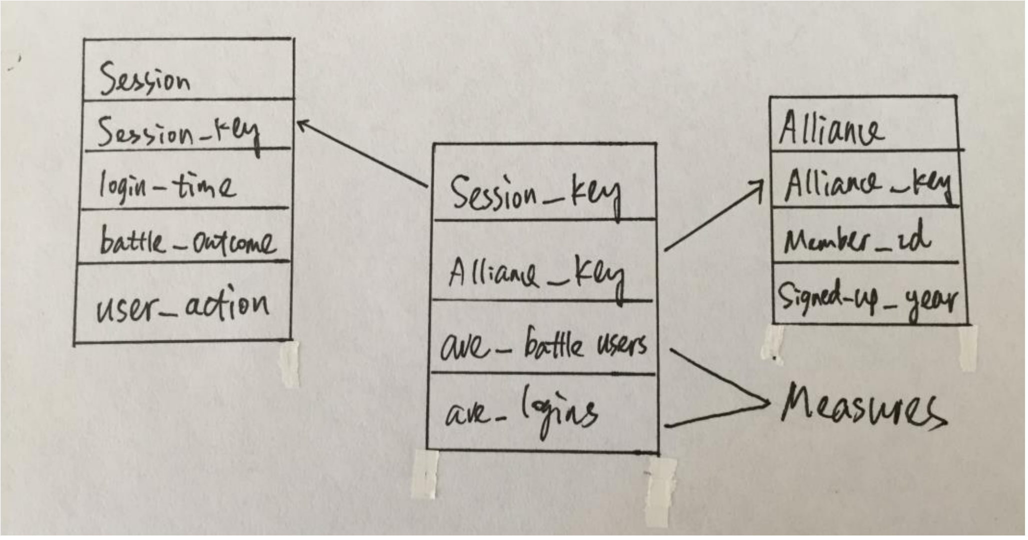
(a) Accounts: Changed: Updated.
Alliances: Unchanged.
User interactions: Changed; Added.
Battle sessions: Changed; Added.

(b) We need Accounts and Battle sessions to get the information.

(c) Metadata provides the information about all the properties of other data. It just likes a kind of electronic catalogue, which makes searching for specific data convenient.

(d)

1d:
-1 wrong dimension
-1 didn't explain which database the two measures come from



The session table is from attribute Battle session and Alliance table is from attribute Alliance and Account.

(e) Average size: (7+150+12+6+45)/5=44.

(f) $\mathrm{E(A)} = (150 + 4.2 + 83 + 0 + 23.5) \div 5 = 52.14$

$\mathrm{E(B)} = (2.5 + 300 + 18.5 + 260 + 25) \div 5 = 121.2$

$$\sigma(\mathrm{A}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - E(A))^2} = 57.212.$$

$$\sigma(\mathrm{B}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(b_i - E(B))^2} = 130.481.$$

$$\mathrm{r(A,B)} = \frac{\sum_{i=1}^{n}(a_i - E(A))(b_i - (E(B))}{n\sigma(A)\sigma(B)} = \frac{-27838.84}{5 \times 57.212 \times 130.481} = -0.746$$

(g) We could find that members who exchanged less message may defeat more monsters. This may lead to users exchange less chat messages while fighting with monsters.

-2 didn't explain why there is a negative correlation

Question 2 Understanding User Data

(a) n=6; p=6, which are ID, Name, Gender, Age, Skill Level, Motto.

(b) Gender is a binary attribute and Age is a numeric attribute.

2c: -6 wrong way to transform attribute to binary attribute

(c) First, we calculated age in to binary. Since there are two people have the same age, we only need 4 digits. Sorting the ages to: 15, 15, 16, 16.5, 17, 25. 15 is the first in the ordering, thus the first digit is 1. 15 is also one of the first two numbers in the ordering, thus, the second digit is 1 as well. Keep on doing this and we can get 1111 for 15. While 16 is not the first one number, thus, the first digit is 0 and it is one of the first two number so the second digits is 1. Therefore, we can get the binary digits:1111, 1111, 0111, 0011, 0001, 0000 for age.
We can use the same method for skill level. Sorting the value: Black, Black, Black, Brown, Doggy, Green. The binary digits are: 111, 111, 111, 011, 001, 000.

Question 3 Constructing User Graphs

(a) Each graph has 8 edges.

(b) In Chat graph, Astro has 4 degrees, which is highest. In Artifact Exchange grap Frodo has 4 degrees.

3b: -1, Astro has highest degree in Artifact Exchange graph too

(c) In Chat graph, Katara, Frodo and Hermione have the second highest degree. In Artifact Exchange graph, Harry, Frodo and Katara have the second highest degree.

(d) In Chat graph:

$$C_C'(A) = \left[\frac{\sum_{i=1}^{N} d(A,i)}{N-1}\right]^{-1} = \frac{6-1}{1+2+1+1+2} = \frac{5}{7} = 0.714$$

In Artifact Exchange graph:

$$C_C'(A) = \left[\frac{\sum_{i=1}^{N} d(A,i)}{N-1}\right]^{-1} = \frac{6-1}{1+2+1+2+1} = \frac{5}{7} = 0.714$$

(e) In Chat graph:
Harry – Frodo: 0.5; Astro – Harry: 0.5; Superman – Harry: 0.5;
Thus, the betweenness centrality of Katara is 0.5+0.5+0.5=1.5.

In Artifact Exchange graph:
Harry – Astro: 1/3; Harry – Superman: 1/3
Thus, the betweenness centrality of Katara is 2/3=0.667.
Therefore, the betweenness centrality in two graphs are different.

Question 4 Data cubes and OLAP

(a) The base cuboid has 5 dimensions, which have 10, 2, 7, 5 and 5 distinct values. Thus, the number of cells is $10 \times 2 \times 7 \times 5 \times 5 = 3500.$

(b) 0-D cuboid: 1

1-D cuboid: $10 + 2 + 7 + 5 + 5 = 29$

2-D cuboid:

$$10 \times 2 + 10 \times 7 + 10 \times 5 \times 2 + 2 \times 7 + 2 \times 5 \times 2 + 7 \times 5 \times 2 + 5 \times 5 = 319$$

3-D cuboid:

$$10 \times 2 \times 7 + 10 \times 2 \times 2 \times 5 + 10 \times 7 \times 5 \times 2 + 10 \times 5 \times 5 + 2 \times 7 \times 5 \times 2 + 2 \times 5 \times 5 + 7 \times 5 \times 5 = 1655$$

4-D cuboid:

$10 \times 2 \times 7 \times 5 \times 2 + 10 \times 2 \times 5 \times 5 + 2 \times 7 \times 5 \times 5 + 10 \times 7 \times 5 \times 5 =$4000

5-D cuboid: $10 \times 2 \times 7 \times 5 \times 5 = 3500$

The total number of cells: $1 + 29 + 319 + 1655 + 4000 + 3500 = 9504.$

(c) Slice/dice the data on color purple and then roll up on intelligence.

-1 roll up on species

Question 5 Structure of a Network

(a) Nodes (1, 3, 4, 8, 9, 13, 14, 15, 18) constitute the largest strongly connected component in the graph.
Nodes (6, 7, 11, 12) belong to the set IN.
Nodes (5, 10, 16) belong to the set OUT.
Nodes (2, 17) belong to tendrils of the graph.

(b) Add node 5 to node 15. Since node 5 is in set OUT, we just need to make it could reach the SCC.

(c) Delete the edge that node 15 to node 18, therefore, node 18 would belong to set IN.

Question 6 Chi-Square Test

(a) 55 owls have black feather and 45 owls have white feather. 80 owls have red beak and 20 owls have yellow beak.

(b) Feather:

$$e_{11} = \frac{55 \times 75}{100} = 41.25, \; e_{12} = \frac{45 \times 75}{100} = 33.75, \; e_{21} = \frac{55 \times 25}{100} = 13.75$$

$$e_{22} = \frac{25 \times 45}{100} = 11.25$$

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$$

$$= \frac{(45-41.25)^2}{41.25} + \frac{(30-33.75)^2}{33.75} + \frac{(10-13.75)^2}{13.75} + \frac{(15-11.25)^2}{11.25}$$

$$= 3.03$$

Beak:

$$e_{11} = \frac{80\times75}{100} = 60,\ e_{12} = \frac{20\times75}{100} = 15,\ e_{21} = \frac{80\times25}{100} = 20,\ e_{22} = \frac{25\times20}{100} = 5$$

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \quad = \frac{(60-60)^2}{60} + \frac{(15-15)^2}{15} + \frac{(20-20)^2}{20} + \frac{(5-5)^2}{5}$$

$$= 0$$

6b: -1 lack of justification in conclusion

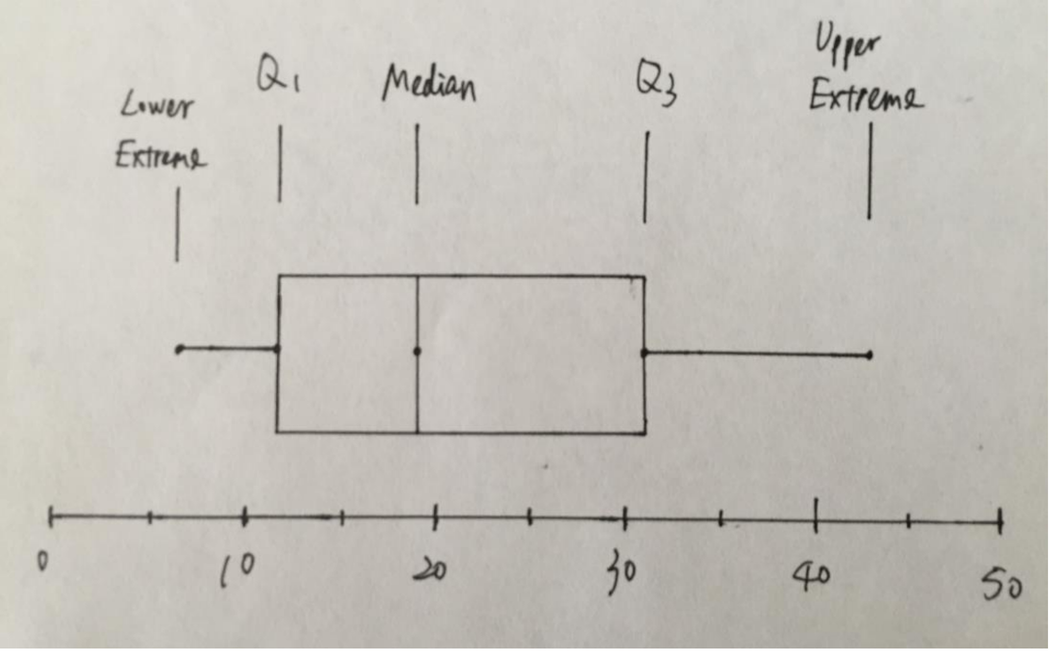
Therefore, the magic localization ability is correlated with feather color and white feather seems to produce highly capable owls.

(c) Since $\chi^2 = 20$, this means sulphur crest is highly correlated with strong abilit owls. However, correlation is different from causation, thus, we cannot say sulphur crest causes improved localization ability.

6c: -2 should provide example of actual cause

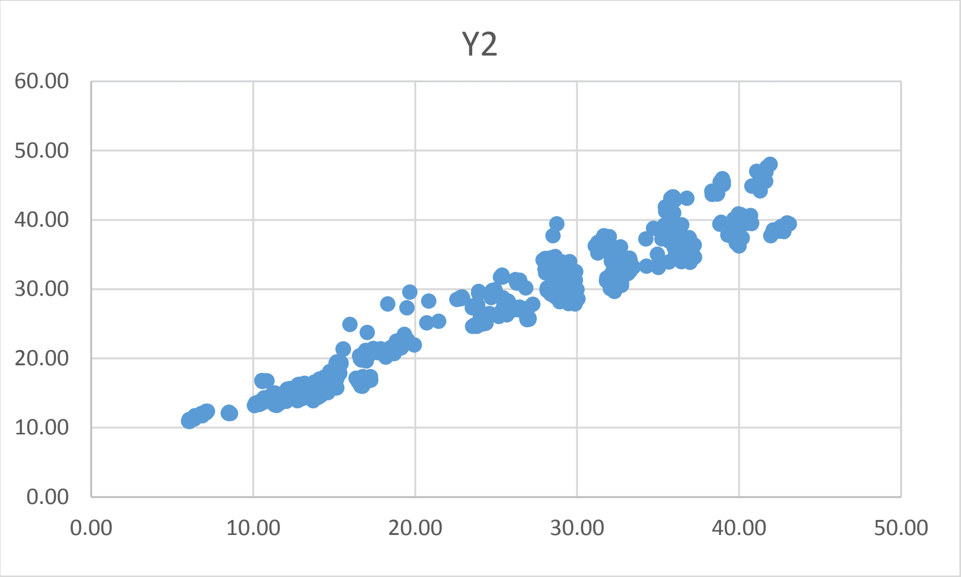
Question 7 Hands-on Analysis of a Real-World Dataset

(a) Median: 18.95, Q1: 12.9775, Q3: 31.6825, Average: 22.31, IQR: 18.705
Maximum: 43.10, Minimum: 6.01.
Since Q1-1.5*IQR<minimum and Q3+1.5*IQR>maximum, therefore, there is no outlier.



(b) In Y1, maximum: 43.10, minimum:6.01.

In Y2, maximum: 48.03, minimum:10.90.
Y1 and Y2 are positively correlated as shown in the graph.



(c) $r(X1, Y1) = 0.602, r(X2, Y1) = -0.648, r(X3, Y1) = 0.42, r(X4, Y1) = -0.843, r(X5, Y1) = 0.874, r(X6, Y1) = -0.003, r(X7, Y1) = 0.099, r(X8, Y1) = 0.021.$
The data is calculated by Excel.

7c: -1 the last two is wrong

(d) $r(X1, Y2) = 0.612, r(X2, Y2) = -0.660, r(X3, Y2) = 0.393, r(X4, Y2) = -0.842, r(X5, Y2) = 0.879, r(X6, Y2) = -0.018, r(X7, Y2) = 0.029, r(X8, Y2) = -0.013.$
Therefore, we can rank the attributes from the least to the most correlated: X8, X6, X7, X3, X1, X2, X4, X5.

7d: -3 the result of r(X6, y2) and r(X7, y2) are wrong, so the final ranking is wrong