Research School of Computer Science, Australian National University

COMP3420, Advanced Databases and Data Mining

Assignment 1

Due: Thursday 31 March 2016, 5 pm

Front Matters:

- There are 7 problems in this assignment. There are 6 written questions and 1 data analysis question. A calculator or computer may be needed for the written questions, the data analysis question can be done in any tool of your choice – spreadsheet, Rattle, R, Python, or other programming languages.
- This assignment will be graded out of a total of 100 points, it is worth 20% of the final marks.
- This assignment must be submitted in electronic form via Wattle. We only accept PDF documents (in one single file, typesetting or scanned – with typset preferred), no other formats such as Word or OpenOffice documents, etc. On Wattle, there will be a link called "Assignment 1 Submission" – follow instructions there to upload your assignment. It is your responsibility to ensure that the uploaded PDF file is clearly legible and printable.
- Please clearly put your **student ID** at the top of the first page of your submission. **Do not** put your name on the assignment sheet since grading will be blind.
- Note that you are required to show all your major working steps for all calculation questions. In other words, if you just write down the final result for a question, you would not receive credit for that question.
- Note that the questions are sorted by theme and example, and not sorted according to their difficulty.
- Late policy: we adhere to the standard ANU policy of special considerations. Late penalties are assessed as 5% of obtained mark per every 24 hours late, or part thereof. The **submission cut-off is at 4 days**, i.e. an assignment submitted more than four days late after the submission deadline will receive no marks.
- Exceptions: Should there be a medical condition or other unfortunate circumstances beyond the student's control, it is the student's responsibility to get in touch with the lecturers *before* the original deadline to agree on an alternative arrangement, and be prepared to show proof.

Question 1 [20 marks] Data Mining in an Online Game

In this problem we examine data from *Magic League*, an online *role-playing game* (RPG), where each player assumes a virtual identity, aims to advance in a virtual landscape, and earn scores by defeating virtual monsters. Even for the most prolific geeks, spending hours shooting virtual monsters can get lonely. Therefore, game developers have introduced online multiplayer options, in which gamers register accounts, they chat with other users, they form online alliances and clans, and they share virtual quality moments together fending minions.

Your mission is to study the user and games data and help improve the social aspect of undead shooting. The user data is stored in the following tables:

1. **Accounts**. These include: information about each user such as their screen name, chosen character, date they sign up, best scores, unlocked achievements, held artifacts, experience level and hit points.
2. **Alliances**. This database includes information about user alliance composition, recording for each alliance the id of its members.
3. **User Interactions**. This table contains information about user-to-user interactions. The information include: the time of the interaction, id of the two users involved in the interaction, the type and attribute of the interaction (e.g., artifact exchange, fight, or chat).
4. **Battle Sessions**. This contains information about each game session. Including: the time that a user logs in; the time the session starts and ends; each user's actions, losses and achievements; and the outcome of the battle.

Your overall task is to build a data warehouse from these databases to analyse user behaviour, with the goal of building a game that is even more fun to play.

(a) A gamer, Tintin, logs into his account and together with a few friends from his alliance started to conquer a forest occupied by monsters. In this battle, Tintin was bitten by a unicorn and temporarily lost his ability to fight. He recovered by drinking a life potion offered by an ally, Pinocchio. The battle ended with the alliance taking control of the forest. Which databases tables (among 1–4 above) are changed during this session? And if changed, are entries added or updated? Please indicate your answer by circling the appropriate option for each DB below. **[4 marks]**

DB(1) Accounts: **Changed / Unchanged**; (if changed) **Added / Updated**

DB(2) Alliances: **Changed / Unchanged**; (if changed) **Added / Updated**

DB(3) User interactions: **Changed / Unchanged**; (if changed) **Added / Updated**

DB(4) Battle sessions: **Changed / Unchanged**; (if changed) **Added / Updated**

Answer:

Accounts:	Changed / Updated	*(1 mark)*
Alliances:	Unchanged	*(1 mark)*
User interactions:	Changed / Added	*(1 mark)*
Battle sessions:	Changed / Added	*(1 mark)*

(b) Tintin and Pinocchio's team needs to assign a team member to fight a new monster, *Basilisk*, in the *reptile* category. You need to supply a function that helps them decide who has had the most number of wins against *reptiles*. Which database(s) do you need to use to get the information? **[2 marks]**

Answer:
We need **Battle Sessions** to count the number of victories of each member. If we don't know who is in Tintin and Pinocchio's team, we would need to extract member information from **Alliances**. **User Interactions** might be needed if we, for example, want to put a different weight to battles in which a member shares a victory with another member.

(Mentioning any 2 of the above 3 databases will give you 2 marks.)

(Mentioning Accounts will cost you 1 mark, unless you can justify your choice.)

(c) What is metadata of a database? Please provide a brief description. Also give two examples of metadata about the **Battle Sessions** database above. **[4 marks]**

Note: 2 marks for the right definition and 1 mark for each example.

Answer:
Metadata is literally "data about data". Metadata is about the properties of the data themselves – for example, the data type, the domain, the acceptable values of a data item, where it originally comes from, what transformation it has been subjected to, where it has been used or what it means, etc. *(2 marks)*

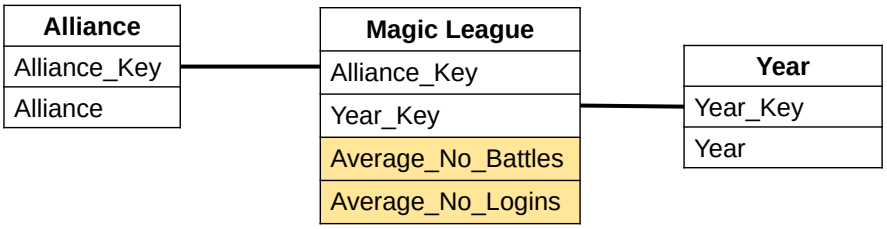
Examples:

- the format of the attribute *time*
- the allowed values for the attribute *outcome*
- information about the foreign keys and where they connect to (e.g. IDs for users and monsters)

(2 marks for any 2 examples of metadata. Examples of normal or derived data get 0 marks.)

(d) You are to generate a data warehouse, containing the average number of battles users participated in, and the average number of logins they made of in May 2014; tabulated by the alliances they are in, and the year they signed up for the game (2011, 2012, 2013). Draw a star schema of this table. Clearly annotate where each measure and each dimension come from in databases (1–4). **[4 marks]**

Answer:
In the star schema, we have 2 dimensions (alliance and year) and 2 measures (average number of battles and logins):



(1 mark for the correct dimensions and 1 mark for the correct measures)

(Providing irrelevant information (e.g. Accounts table) will cost you 1 to 2 marks.)

Alliance comes from **Alliances**. *(1/2 marks)*
Year comes from **Accounts**. *(1/2 marks)*
Average number of battles comes from **Battle Sessions**. *(1/2 marks)*
Average number of logins comes from **Battle Sessions**. *(1/2 marks)*

(e) You analysed the game achievements versus user interactions for a number of prominent alliances in May 2014. Compute the average size of alliances in this sample. **[1 mark]**

Alliance Name	Size (# Members)	Number of Monsters Defeated per Member	Chat Messages Exchanged per Member
Justice League	7	150	2.5
Gryffindor	150	4.2	300
Dumbledore's Army	12	83.0	18.5
Fellowship of the Ring	6	0.0	260
Southern Airbenders	45	23.5	25

Answer:
Average is $(7 + 150 + 12 + 6 + 45)/5 = 44$ members. *(1 mark)*

(f) In the game achievements table in the previous part, how does the average number of defeated monsters related to the average number of chat messages exchanged? Answer this question by computing the Pearson correlation coefficient of these two quantities. Show your workings. **[3 marks]**

Answer:
Let x be the number of monsters defeated and y be the number of chat messages. First we need to find the mean and standard deviation of each of the attributes:

$$\begin{aligned}
\overline{x} &= (150 + 4.2 + 83 + 0 + 23.5)/5 = 52.14 \\
\overline{y} &= (2.5 + 300 + 18.5 + 260 + 25)/5 = 121.2 \\
\overline{x^2} &= (150^2 + 4.2^2 + 83^2 + 0^2 + 23.5^2)/5 = 5991.778 \\
\overline{y^2} &= (2.5^2 + 300^2 + 18.5^2 + 260^2 + 25^2)/5 = 31714.7 \\
s_x &= \sqrt{\frac{5}{5-1}(5991.778 - 52.14^2)} = 63.96 \\
s_y &= \sqrt{\frac{5}{5-1}(31714.7 - 121.2^2)} = 145.88
\end{aligned}$$

If we don't use Bessel's correction (i.e. we divide by n instead of of $(n-1)$), then $s_x = 57.21$ and $s_y = 130.48$. It turns out that it doesn't actually matter whether we use Bessel's correction or not – the final answer for the correlation is unaffected.

The dot product of x and y is also useful:

$$\begin{aligned}
\sum x_i y_i &= (150)(2.5) + (4.2)(300) + (83)(18.5) + (0)(260) + (23.5)(25) \\
&= 3758
\end{aligned}$$

The Pearson correlation is:

$$\begin{aligned}
r &= \frac{\sum x_i y_i - n\overline{x}\,\overline{y}}{(n-1)s_x s_y} \\
&= \frac{3758 - (5)(52.14)(121.2)}{(5-1)(63.96)(145.88)} \\
&= -0.7458
\end{aligned}$$

If not corrected for the bias, then the calculation would be:

$$\begin{aligned} r &= \frac{\sum x_i y_i - n\overline{x}\,\overline{y}}{n s_x s_y} \\ &= \frac{3758 - (5)(52.14)(121.2)}{5(57.21)(130.48)} \\ &= -0.7458 \end{aligned}$$

Correct means: *1 mark*

Correct standard deviations: *1 mark*

Correct correlation: *1 mark*

(If no workings are shown, award a maximum of 1 mark.)
(If the bias correction is applied inconsistently, take away 1 mark.)

(g) Assuming the data in part (e) is collected correctly, what do you think *could* lead to the observed behaviour in each alliance? Provide one example reason. **[2 marks]**

Answer:
There is a strong negative correlation between the two attributes. Perhaps in alliances with more chat messages exchanged, members care more about the social interactions with friends than playing the game itself. These alliances might not take the game seriously, which leads to lower performance.

(2 marks for any sensible justification)

Question 2 [10 marks] Understanding User Data

Consider the following set of 6 game users.

ID	Name	Gender	Age	Skill Level	Motto
u001	Harry	Male	17yrs	Black	"Never tickle a sleeping dragon"
u002	Hermione	Female	16.5yrs	Brown	"When in doubt, go to the library"
u003	Katara	Female	15yrs	Black	"Never turn my back on people who need me"
u012	Frodo	Male	16yrs	Green	"One ring to rule them all"
u011	Superman	Male	25yrs	Black	"An acceptable face of invading realities"
u066	Astro	Male	15mo	Doggy	"Happiness is a warm puppy"

(a) What is the number of datum n and the number of attributes p in this table? **[2 marks]**

Answer:
Each row is a datum, so $n = 6$ *(1 mark)*
The best answer for p is 5. We normally don't count the primary key as an attribute. (No marks are lost if you say 6 though). *(1 mark)*

(b) Among the p different attributes, identify one binary and one numeric attribute. **[2 marks]**

Answer:
Gender is a binary attribute. *(1 mark)*
Strictly speaking, there are currently no numeric attributes, but we can convert *age* to a numeric attribute. *(1 mark)*

(c) You need to apply a data mining algorithm which only accepts binary attributes on this user profile dataset. Explain how the age and skill level variables can be transformed into a binary attribute, or a set of binary attributes **without loosing any information contained the original dataset**. Write out the transformed binary attributes for the skill level attribute for each user. **[6 marks]**

Answer:
To preserve the original information, let us transform the skill level attribute into four binary attributes, each of which is a boolean test to see if a member is at a particular skill level. The transformed attributes are:

ID	Name	Black	Brown	Green	Doggy
u001	Harry	1	0	0	0
u002	Hermione	0	1	0	0
u003	Katara	1	0	0	0
u012	Frodo	0	0	1	0
u011	Superman	1	0	0	0
u066	Astro	0	0	0	1

(2 marks for the explanation and 2 marks for the table.)

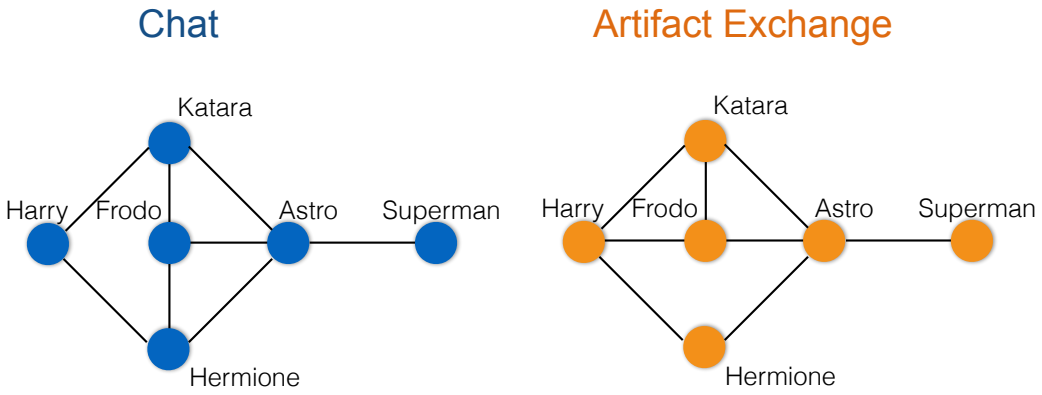
(Subtract 3 marks if only 1 binary attribute is used.)

We can do a similar transformation to the attribute. However if there are more ages, it might be better to bin the ages first to reduce the number of categories.

(2 marks for a sensible approach to transform the 'age' attribute.)

Question 3 [12 marks] Constructing User Graphs

We construct two graphs among these six users, by making an (unweighted) edge between two users when they have exchanged at least 5 messages in May 2014 (on the left, the *Chat* graph); and when they have exchanged at least 5 virtual artifacts in May 2014 (on the right, the *Artifact Exchange* graph).



(a) How many edges are there in each graph? **[2 marks]**

Answer:
In **Chat**, there are 8 edges. *(1 mark)*

In **Artifact Exchange**, there are also 8 edges. *(1 mark)*

(b) Which node(s) have the highest degree in each graph? **[2 marks]**

Answer:
In **Chat**, Astro has the highest degree (of 4). *(1 mark)*

In **Artifact Exchange**, Astro also has the highest degree (of 4). *(1 mark)*

(c) Which node(s) have the second highest degree in each graph? **[2 marks]**

Answer:
In **Chat**, Katara, Frodo, and Hermione have the second highest degree (of 3). *(1 mark)*

In **Artifact Exchange**, Katara, Frodo, and Harry have the highest degree (of 3). *(1 mark)*

(d) What is the closeness centrality of Frodo in each graph? **[2 marks]**

Answer:
In both **Chat** and **Artifact Exchange**, Frodo's closeness centrality is:

$$C_c(\text{Frodo}) = 5(1+1+1+2+2)^{-1} = 5/7 = 0.714$$

(The un-normalized closenes centrality would be $1/7$*.)*

(e) What is the (un-normalized) betweenness centrality of Katara in the *Chat* graph? i.e. the number of shortest paths from all users to all others users that pass through Katara. Is this the same with her betweenness centrality in the *Artifact Exchange* graph, why or why not? **[4 marks]**

Answer:
In **Chat**, it's 1.5 – Katara is on the shortest paths between Harry-Astro, Harry-Frodo, Harry-Superman, each of which is tied with Hermione. *(2 marks)*

In **Artifact Exchange**, it's 2/3 – Now Katara has the same "position" as both Hermione and Frodo. They are all on the shortest paths of Harry-Astro and Harry-Superman. *(2 marks)*

Question 4 [8 marks] Data cubes and OLAP

The *Magic League* game provides a feature for users to build virtual pets, take them along as battle companions, or give to each other as gifts.

1. `Species` ten possible values: Dog, Cat, Pig, Lizard, Horse, Stag, Otter, Swan, Hare, Phoenix.
2. `Gender` two possible values: Male/Female.
3. `Color` seven different values.
4. `Size` five different values.
5. `Intelligence`, five different values.

You are to study users pet-keeping behaviour.

(a) The first task is to construct a data cube. How many cells are in the base cuboid? **[3 marks]**

Answer:
$10 * 2 * 7 * 5 * 5 = 3,500$ *(2 marks for working, 1 mark for the correct final answer)*

(b) How many cells are there in total if you were to compute all cuboids? **[3 marks]**

Answer:
Each cell is indexed by a 5-dimensional vector $x = (x_1, \cdots, x_6)$, where $x_i \in \{0, 1, \cdots, L_i\}$. Here L_i represents the different values that dimension i can take, and 0 indicates that this dimension is missing. Thus the total number of cells is just the number of ways to construct vector x:

$$\prod_{i=1}^{5}(1 + L_i) = 11 \times 3 \times 8 \times 6 \times 6 = 9,504$$

(2 marks for justification of the approach, 1 mark for the correct final answer)

(c) One summary measure in the datacube is the number of pets. Given a cuboid with dimensions `Species`, `Colour` and `Intelligence`, what OLAP operations do you use to get the number of pets that is *purple*, and tabulated by their `Intelligence`? **[2 marks]**

Answer:
Roll-up on species *(1 mark)* and then slice for purple *(1 mark)*.

(Subtract 1 mark if unnecessary operations are included.)

(Subtract 1 mark if the dimensions being operated on are not mentioned.)

Question 5 [15 marks] Structure of a Network

Consider the set of 18 Web pages drawn in the following figure, whose links forming a directed graph.

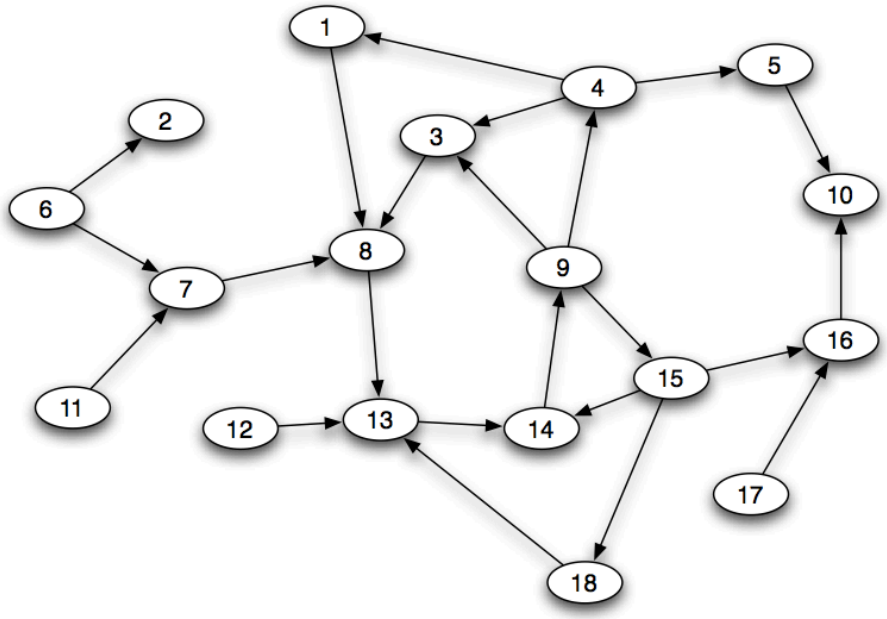


Figure 1: A directed network of 18 web pages.

(a) Which nodes constitute the largest strongly connected component (SCC) in this graph? Taking this as the giant SCC, which nodes then belong to the sets IN and OUT as defined in the lectures? Which nodes belong to the tendrils of the graph? Explain all of your answers. **[9 marks]**

Answer:
SCC = {1, 3, 4, 8, 9, 13, 14, 15, 18} *(1 mark)*

- Look for directed cycles in the graph as follows.
  - 9, 14, 15 form a closed triangle, hence they are strongly connected.
  - 13 and 18 are part of another directed cycle with 9, 14 and 15.
  - 3, 4, 8 are part of another directed cycle with 13, 14, 9.
  - 1 has a directed cycle with 4, 8, 13, 14, 9.

  (2 marks for any reasonable explanation)

IN = {6, 7, 11, 12} *(1 mark)*

- Nodes that have a directed path to (but not from) the SCC. *(1 mark)*

OUT = {5, 10, 16} *(1 mark)*

- Nodes that have a direct path from (but not to) the SCC. *(1 mark)*

TENDRILS = {2, 17} *(1 mark)*

- Nodes that can neither reach the SCC nor can we get to them from the SCC. *(1 mark)*

(b) As new links are created and old ones are removed among an existing set of Web pages, the pages move between different parts of the bow-tie structure.

Name an edge you could add or delete from the graph in the above figure so as to increase the size of the largest strongly connected component. Explain why you named this edge. **[3 marks]**

Answer:
There are many examples, some of which are:

- Add edge 5 → any node in the SCC. This will move node 5 to SCC.
- Add edge 10 → any node in the SCC. This will move nodes 5, 10, and 16 to SCC.
- Add edge 16 → any node in the SCC. This will move node 16 to SCC.
- Add edge 8 → 11. This will move nodes 7 and 11 to SCC.
- Add edge 8 → 6. This will move nodes 7 and 6 to SCC.
- Add edge 8 → 12. This will move node 12 to SCC.

(1 mark for adding an appropriate edge, 2 marks for explaining the effect of such addition)

(c) Name an edge you could add or delete from the graph in the above figure so as to increase the size of the set IN. Explain why you named this edge. **[3 marks]**

Answer:
Again, there are many possibilities:

- Adding edge 2 → any node in the SCC will move node 2 to IN.
- Adding edge 17 → any node in the SCC will move node 17 to IN.
- Deleting edge 15 → 18 will move node 18 from SCC to IN.
- Deleting edge 4 → 1 will move node 1 from SCC to IN.
- Deleting edge 9 → 4 will move nodes 4 and 1 from SCC to IN.
- Deleting edge 9 → 15 will move nodes 15 and 18 from SCC to IN.

(1 mark for deleting/adding an appropriate edge, 2 marks for explaining the effect)

Question 6 [15 marks] Chi-Square Test

Hogwarts owlry keeps a large number of owls with varying magic capacity. We examine their *feather colour* - black or white; and *beak colour* - red or yellow, along with a critical magic property: *ability to locate the recipient* - strong or weak. The table below contains the number of owls that possess two qualities simultaneously, e.g. there are 10 black-feathered owls that have strong localization ability.

	black feather	white feather	red beak	yellow beak
weak localization	45	30	60	15
strong localization	10	15	20	5

(a) How many owls have black feather? white feather? How about red or yellow beak? **[4 marks]**

Answer:
There are 55 owls with black feathers. *(1 mark)*
There are 45 owls with white feathers. *(1 mark)*
There are 80 owls with red beaks. *(1 mark)*
There are 20 owls with yellow beaks. *(1 mark)*

(b) Is the magic localization ability correlated with feather colour or beak colour? Which feather or beak colour seems to produce highly capable owls? Answer this question by manually computing χ^2 tests on feather/beak colours and localization ability. Show all of your workings. **[7 marks]**

Answer:
Define the following notations:

W – having weak localization

S – having strong localization

B – having black feather

H – having white feather

R – having red beak

Y – having yellow beak

First we compute the probability of having each of the features:

$$\mathbb{P}(W) = \frac{75}{100} \qquad \mathbb{P}(S) = \frac{25}{100}$$
$$\mathbb{P}(B) = \frac{55}{100} \qquad \mathbb{P}(H) = \frac{45}{100}$$
$$\mathbb{P}(R) = \frac{80}{100} \qquad \mathbb{P}(Y) = \frac{20}{100}$$

Next we compute the expected count in each category under the independence assumption:

$$e(WB) = 100\frac{75}{100}\frac{55}{100} = 41.25 \qquad e(WH) = 100\frac{75}{100}\frac{45}{100} = 33.75$$
$$e(SB) = 100\frac{25}{100}\frac{55}{100} = 13.75 \qquad e(SH) = 100\frac{25}{100}\frac{45}{100} = 11.25$$
$$e(WR) = 100\frac{75}{100}\frac{80}{100} = 60 \qquad e(WY) = 100\frac{75}{100}\frac{20}{100} = 15$$
$$e(SR) = 100\frac{25}{100}\frac{80}{100} = 20 \qquad e(SY) = 100\frac{25}{100}\frac{20}{100} = 5$$

We can now compute the χ^2 test statistic to test the independence between the magic localization ability and the feather colour:

$$\begin{aligned}\chi^2 &= \frac{(45-41.25)^2}{41.25} + \frac{(30-33.75)^2}{33.75} + \frac{(10-13.75)^2}{13.75} + \frac{(15-11.25)^2}{11.25} \\ &= 3.03\end{aligned}$$

(2 marks for the correct workings and 1 mark for the correct answer)

The χ^2 test statistic to test the independence between the magic localization ability and the beak colour is:

$$\begin{aligned}\chi^2 &= \frac{(60-60)^2}{60} + \frac{(15-15)^2}{60} + \frac{(20-20)^2}{20} + \frac{(5-5)^2}{5} \\ &= 0\end{aligned}$$

(2 marks for the correct workings and 1 mark for the correct answer)

For a 2 by 2 contingency table, the degrees of freedom are $(2-1)(2-1) = 1$. The cut-off point for a χ^2 distribution with 1 degree of freedom at the 5% significance level is 3.841. Since both of our χ^2 test statistics are less than 3.841, there is ***not*** enough evidence to reject the null hypotheses that the localization ability is independent of the feather colour and the beak colour. This would probably mean that the localization ability is not correlated with the feather or the beak colour.
(1 mark for having a reasonable conclusion)

(c) If there is a third attribute, having *sulphur crest*, found to be highly correlated with strong localization ability in owls, with $\chi^2 = 20$. Is it correct to say that *sulphur crest* causes improved localization ability? Why or why not? **[4 marks]**

Answer:
No. *(1 mark)*

Correlation between two variables does not necessarily imply causality. *(1 mark)*

These traits could be, for example, the result of a common genetic cause. *(2 marks)*

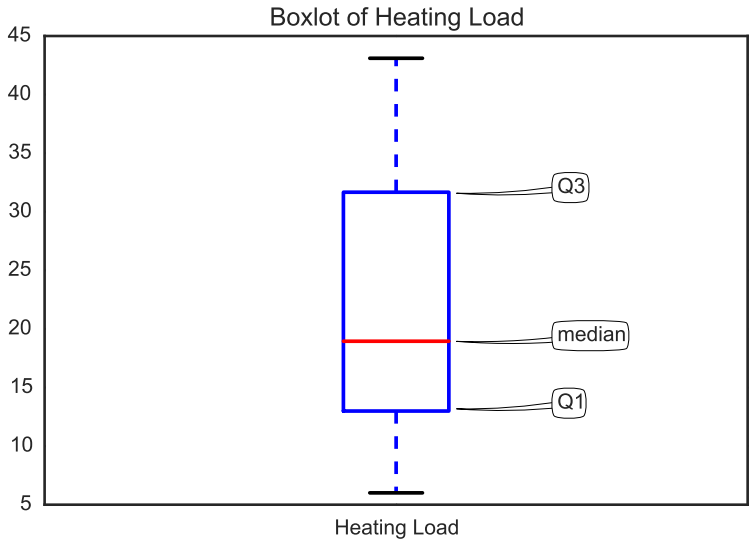
Question 7 [20 marks] Hands-on Analysis of a Real-World Dataset

Take the UCI Energy efficiency dataset `http://archive.ics.uci.edu/ml/datasets/Energy+efficiency`, complete the following analysis and compute the designated metrics.

(a) Draw a boxplot of dimension $y1$: `heating load`, annotate all key landmarks on the box plot. What is the median of $y1$, what is the mean of $y1$, what are the values of $Q1$ and $Q3$, how large is the inter-quartile range, are there any outliers?

[8 marks]

Answer:



(1 mark for the correct plot)
(1 mark for the correct annotation of key landmarks – Q1, median, and Q3)

The median of $y1$ is 18.95. *(1 mark)*

The mean of $y1$ is 22.31. *(1 mark)*

Q1 is 12.99. *(1 mark)*

Q3 is 31.67. *(1 mark)*

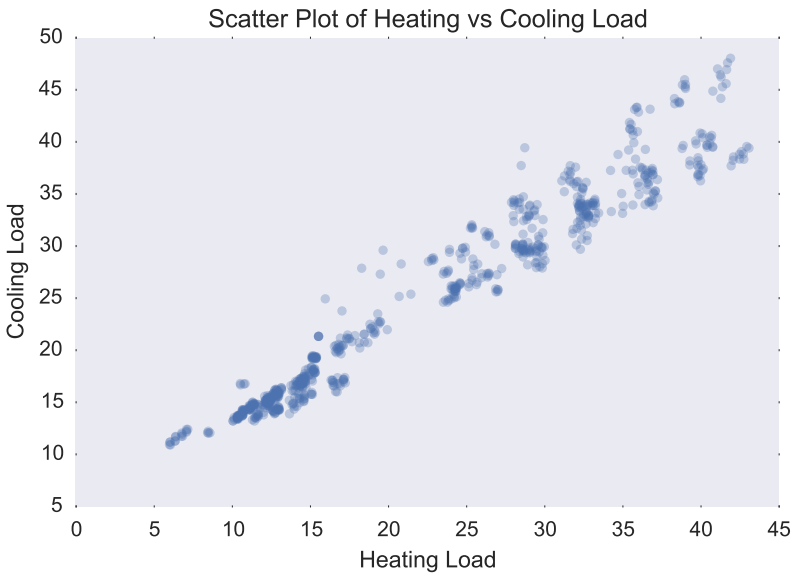
IQR is $31.67 - 12.99 = 18.68$. *(1 mark)*

In the above plot, the maximum length of the whiskers is (1.5×IQR). Data that are beyond the whiskers are considered as outliers and are plotted as individual points. Given this criteria, we can see no outliers on the plot. *(1 mark)*

(b) Draw a scatter plot of dimension $Y1$: `heating load` vs $Y2$: `cooling load`. What is the minimum and maximum of $Y1$ and $Y2$, respectively? Are $Y1$ and $Y2$ positively correlated, negatively correlated, or appear to be un-correlated?

[5 marks]

Answer:



(1 mark for the correct plot and 1 mark for labelling the axes)

	Minimum	Maximum
`Y1 heating load`	6.01	43.10
`Y2 cooling load`	10.90	48.03

(2 marks for correct max and min values)

From the plot, it appears that heating load is strongly positively correlated with cooling load. In fact, the Pearson correlation is 0.9758.

(1 free mark for everyone since we accidentally revealed the solution in the question sheet.)

(c) Compute the Pearson correlation of the 8 attributes $X1 \ldots X8$ with $Y1$ `heating load`.

[4 marks]

Answer:

Attribute	ρ with `Y1`
`X1 relative compactness`	0.6223
`X2 surface area`	-0.6581
`X3 wall area`	0.4557
`X4 roof area`	-0.8618
`X5 overall height`	0.8894
`X6 orientation`	-0.0026
`X7 glazing area`	0.2698
`X8 glazing area distribution`	0.0874

(1/2 marks for each correct value)

(d) Rank the attributes with respect to their strength of correlation with $y2$ `cooling load` – from the least to the most correlated.

[3 marks]

Answer:

Attribute	ρ with `Y2`
`X1 relative compactness`	0.6343
`X2 surface area`	-0.6730
`X3 wall area`	0.4271
`X4 roof area`	-0.8625
`X5 overall height`	0.8958
`X6 orientation`	0.0143
`X7 glazing area`	0.2075
`X8 glazing area distribution`	0.0505

If the correlations are ranked by absolute value, award 3 marks:

$$X6 < X8 < X7 < X3 < X1 < X2 < X4 < X5$$

If the correlations are ranked from the most negative to the most positive, award 1 mark:

$$X4 < X2 < X6 < X8 < X7 < X3 < X1 < X5$$

Appendix: Python Code for Question 7

It's best to run the following code in IPython Notebook. If you don't have Python yet, you can install the Anaconda distribution and, for pretty visualisations, the seaborn library.

To run the code, copy each of the blocks below in a separate code cell in IPython Notebook and press `Ctr-Enter`.

```
# some essential imports
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# you might need to install the 'seaborn' package first
import seaborn as sns

# we want the plots to appear on the same page (instead of in a
#    separate window)
%matplotlib inline

# download the excel spreadsheet from source and load in the dataset
energy = pd.io.excel.read_excel('http://archive.ics.uci.edu/' +
         'ml/machine-learning-databases/00242/ENB2012_data.xlsx')

# preview the first few rows
energy.head()
```

Part a: Make the boxplot

```
# pick a theme and initialise the figure
sns.set_style('white')
fig = plt.figure(figsize=(6, 5))

# create a new axis and make the scatter plot
ax = fig.add_subplot(111)
ax.boxplot(energy['Y1'], labels=['Heating Load'])
ax.set_title('Boxlot of Heating Load')

# annotation in matplotlib takes a bit of work
# first we customise the shape of the arrow and text box
arrowprops = dict(arrowstyle='wedge', fc='w', ec='k',
                  connectionstyle="arc3,rad=-0.05")
bbox=dict(boxstyle="round4", fc="w")

# add annotation
ax.annotate('Q3', xy=(0.58, 0.665), xytext=(50, 0),
            xycoords=ax.transAxes, textcoords='offset points',
            arrowprops=arrowprops, bbox=bbox)
ax.annotate('median', xy=(0.58, 0.349), xytext=(50, 0),
            xycoords=ax.transAxes, textcoords='offset points',
            arrowprops=arrowprops, bbox=bbox)
ax.annotate('Q1', xy=(0.58, 0.205), xytext=(50, 0),
            xycoords=ax.transAxes, textcoords='offset points',
            arrowprops=arrowprops, bbox=bbox)

# export plot to pdf and display plot on screen
fig.savefig('energy_boxplot.pdf')
plt.show()
```

Part b: Get some statistics on Y1 and Y2

```
# general statistics
energy[['Y1', 'Y2']].describe()
```

```
# median of Y1 and Y2
energy[['Y1', 'Y2']].median()
```

Part b: Make the scatter plot

```
# pick a theme and initialise the figure
sns.set_style("dark")
fig = plt.figure(figsize=(8, 5))

# create an axis and make the scatter plot
ax = fig.add_subplot(111)
ax.scatter(energy['Y1'], energy['Y2'], c=sns.color_palette()[0],
           alpha=0.3, lw = 0)
ax.set_title('Scatter Plot of Heating vs Cooling Load')
ax.set_xlabel('Heating Load')
ax.set_ylabel('Cooling Load')

# save to pdf and display the plot
fig.savefig('energy_scatter.pdf')
plt.show()
```

Part b: Correlation between Y1 and Y2

```
np.corrcoef(energy['Y1'], energy['Y2'])
```

Part c: Correlation between the attributes and Y1

```
attributes = energy[['X1', 'X2', 'X3', 'X4', 'X5', 'X6', 'X7', 'X8']]
attributes.corrwith(energy['Y1'])
```

Part d: Rank correlation with Y2

```
corr_y2 = attributes.corrwith(energy['Y2'])
corr_y2_df = pd.DataFrame({'corr_y2': corr_y2,
                           'corr_y2_mag': corr_y2.abs()})
corr_y2_df.sort('corr_y2_mag')['corr_y2']
```