Clustering, Association Rule, Classification

For the scenario provided, indicate which type of data mining algorithm (i.e., clustering, classification, or association rules) would be most suitable to complete the following task. Justify your choice in 1-3 sentences.

Option 1

Adaptive testing is a way to increase or decrease an exam's difficulty based on whether someone has gotten previous questions right or wrong. Suppose we had three possible difficulty levels for an exam (easy, medium, and hard). Which type of algorithm would you use to match students to the right next question?

Option 2

Karin wants to help organize the students in CPSC 100 into study groups for the final exam. To prepare for this task, she has transformed each exam grade into a letter grade (e.g., if someone scored 85%, that translates to an A). Her considerations when forming study groups are time zone, midterm 1 letter grade, and midterm 2 letter grade. Which type of algorithm should she use to create these study groups?

Option 3

Exam analytics are pieces of information about the exam that are not necessarily related to the exam's content. For example, analytics could be how long someone spent on a particular question or the average amount of time someone has scrolled away from a question or left the page. The CPSC 100 teaching team is interested in identifying students who may have provided each other with unauthorized help during an exam. Which type of algorithm should they use to do so?

Option 4 [Alternate Sitting]

Exam analytics are pieces of information about the exam that are not necessarily related to the content of the exam. For example, analytics could be how long someone spent on a particular question or the average amount of time someone has scrolled away from a question or left the page. The CPSC 100 teaching team wants to use these analytics to determine if students who

perform well on an exam complete it more quickly than average. Which type of algorithm should they use to do so?

Parts of a URL

Use the URL given to answer the following questions

1. What is the highest or (top-level) domain?
2. What is the top-level folder?
3. Which folder contains the image?
4. What is the name of the image file?

Option 1

https://cdn.i-scmp.com/sites/default/files/styles/768x768/public/d8/images/methode/2020/02/26/94f0ea-821_image_hires_19.jpg?itok=7Wj2_kY2&v=1582707444

1. What is the highest or (top-level) domain?
2. What is the top-level folder?
3. Which folder contains the image?
4. What is the name of the image file?

Option 2

https://hips.hearstapps.com/ell.h-cdn.co/assets/16/38/2560x1280/landscape-141-s-gettyimages-930-master-lead.jpg?resize=980

1. What is the highest or (top-level) domain?
2. What is the top-level folder?
3. Which folder contains the image?
4. What is the name of the image file?

Option 3

https://img.republicworld.com/republic-prod/stories/promolrge/xxdpi/jdoj0in-1558.jpeg?tr=w-812,h-464

1. What is the highest or (top-level) domain?

2. What is the top-level folder?
3. Which folder contains the image?
4. What is the name of the image file?

Option 4

https://cdn.i-scmp.com/sites/default/files/styles/768x768/public/d8/images/methode/2020/03/06/df6-5d2b-11ea-image-hires-132.jpg?itok=gS0MlhnL&v=1583469879

1. What is the highest or (top-level) domain?
2. What is the top-level folder?
3. Which folder contains the image?
4. What is the name of the image file?

Option 5

https://images.squarespace-cdn.com/content/v1/51b3dc8ee4b05196ceb10de/159278-M1OLAH/48nE-K-QXox0-W7i2zEA/4543-453b-image-asset.jpeg?format=2500w

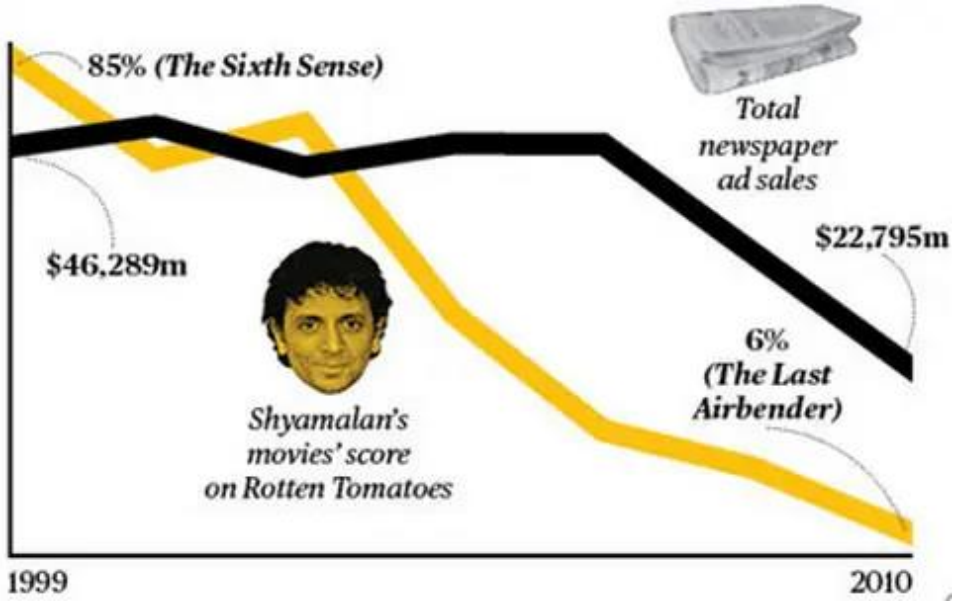
1. What is the highest or (top-level) domain?
2. What is the top-level folder?
3. Which folder contains the image?
4. What is the name of the image file?

Faulty Representation

Option 1

What are two things wrong with this visual representation

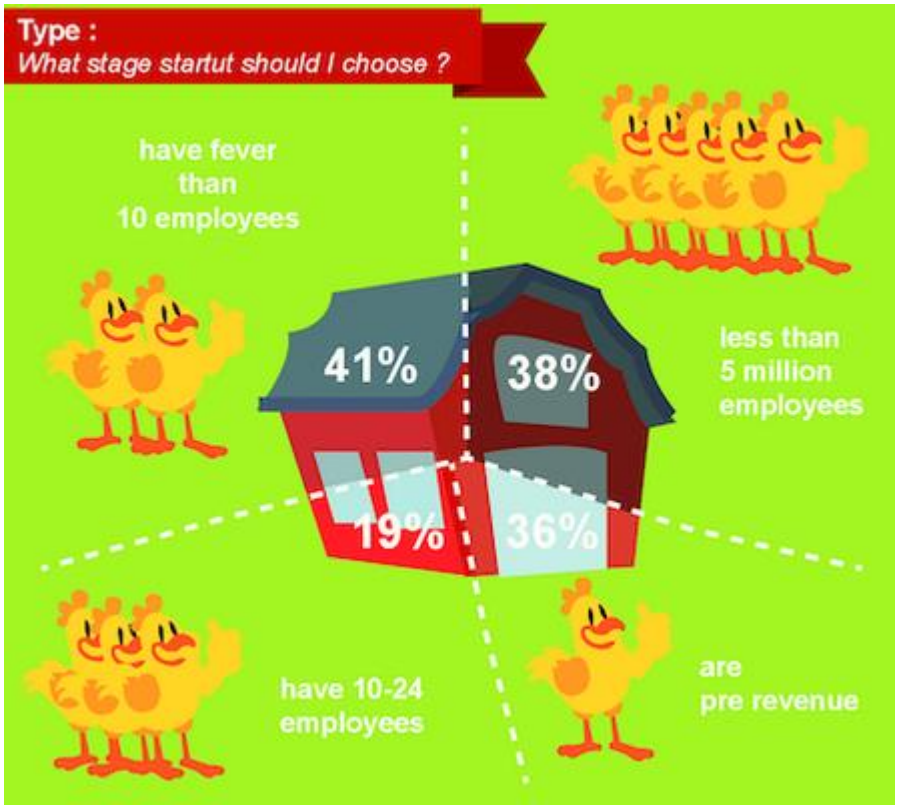
Note that this is NOT an infographic, it is just a static visual representation so we are not asking you to talk about the 5 principles of infographics here.



Option 2 [Alternate Sitting]

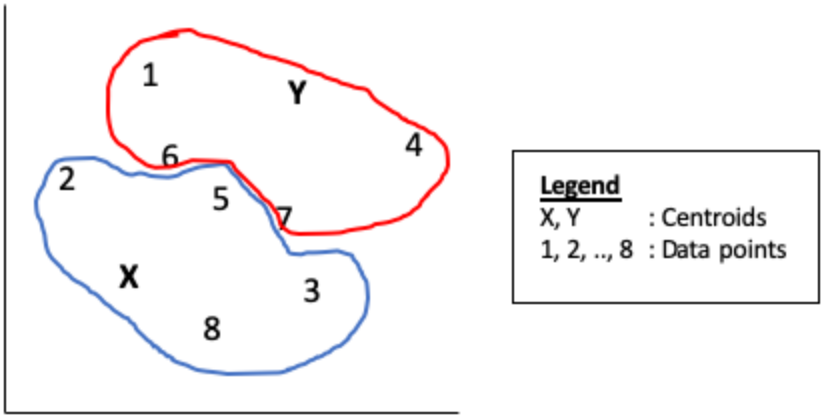
What are three things wrong with this visual representation

Note that this is NOT an infographic, it is just a static visual representation so we are not asking you to talk about the 5 principles of infographics here.

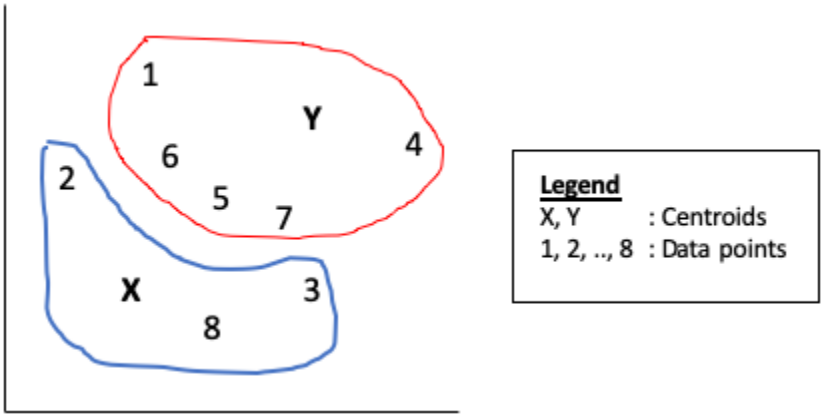


K-Means

You have been tasked with creating two clusters from eight data points. Thinking back to your CPSC 100 class, you recall the k-means algorithm and decide to use it to accomplish your task. After running your algorithm some number of times, you produce the following clustering.



For good measure, you decide to run the algorithm one more time, and you produce the following clustering.



You decide to check the distance between each data point and the centroids X and Y and produce the following table. What should your next step be? You can safely assume that the values shown in the table are correct.

Option 1

Data Point	Distance to Centroid X	Distance to Centroid Y
1	4.8	3.2
2	2.1	3.5
3	5.8	5.9
4	6.9	4.8
5	4.1	4.0
6	3.8	3.7
7	4.6	4.4
8	2.4	6.1

Option 2

Data Point	Distance to Centroid X	Distance to Centroid Y
1	4.8	3.2
2	2.1	3.5
3	5.8	5.9
4	6.9	4.8
5	4.0	4.1
6	3.8	3.7
7	4.6	4.4
8	2.4	6.1

Option 3

Data Point	Distance to Centroid X	Distance to Centroid Y
1	3.2	4.8
2	2.1	3.5
3	5.8	5.9
4	4.8	6.9
5	4.2	4.1
6	3.2	3.7
7	4.6	4.4
8	2.4	6.1

Security Issues, Phishing Scams, Fake News

With a real-world example, explain in your own words how IP addresses, URLs, and routers play a role in fake news websites. Note that an answer that merely defines these terms will be marked as incorrect. Also, give one practical step that individuals can do to identify whether a news story on the website is legitimate.

Alternate Sitting:

With a real-world example, explain in your own words how IP addresses, Email headers, and routers play a role in phishing scams. Note that an answer that merely defines these terms will be marked as incorrect. Also, give one practical step that individuals can do to identify whether an email is part of a phishing scam.

Best Representation Medium and Visualization

For this problem, you must do the following:

- Detail the visual medium (i.e., static visualization, interactive visualization, infographic) you would use and justify why that medium is the most appropriate.
- For the data (or datasets) that need(s) to be visualized, specify what type(s) of visual representation (e.g., bar chart, heatmap, scatter plot) is most appropriate and give a clear description of why the visual representation is the best.

So, for instance, your answer should be of the form.

Medium: Static visualization because .....
Representation: Bar chart because ....

Note that merely stating the definition of the terms will not be awarded any points.

Option 1

As a real estate agent, you have been tasked with presenting a seminar on Vancouver's changing home prices. You will have **an hour** to explain the complexity of the Vancouver real estate industry. **One of the ideas** you would like to convey is the state of home prices in the city over the last 20 years for different types of properties (e.g., townhouse, condo, single-family detached, etc.).

Option 2

As a social anthropologist, you have been involved in research that seeks to understand how people lived during the Ming dynasty. You have read over 400 articles on the subject. You have explored datasets that quantify the dynasty's economic structure, the religious beliefs of the population, the agricultural output of different regions across time, and the trade routes that existed. After four years of research, you want to share some of your findings with members of the research community. Your goal is to convey, at the height of the dynasty, their wealth through the lens of regional strength and trade routes

Option 3

As a high school nutritionist, you have to communicate with students the importance of a balanced diet on mental, physical, and emotional health. Your research indicates that high school students love eating junk and fast food. Your goal is to compare and contrast the calories in healthy and unhealthy options and to discuss the impact on the various aspects of their health. You hope that after they encounter the information, they will make healthier food choices.

Option 4

As an environmental engineer, you work for a firm that ensures the proper disposal of hazardous waste material. Recently, your firm has been hired to educate local factories on the danger of hazardous waste in the environment. Your firm has data that details the cancer incidence and mortality relating to improper disposal of hazardous waste for the last 15 years across Canada. In addition, you have data on the impact of toxic waste on the environment (e.g., water pollution, mercury in fish, etc.). As a member of the team, you must conduct a webinar for factory owners.

Option 5 [Alternate Sitting]

As a senior data analyst for a health insurance firm, you have been charged with ensuring that every suspected cause of insurance fraud has been investigated. Insurance claims datasets include the individual's name, their history of claims, their illnesses (if applicable), their age, gender, income, and history of insurance premiums. You also have a list of all the doctors with whom clients can see. For each doctor, you have data relating to their specialty, the number of claims submitted, and the details of each claim (i.e., date, individual, amount). You have a team of analysts who you need to train to detect health insurance fraud.

Impact question

The ability for your car to drive itself means your destinations and routes have to be stored and sent somewhere. For example, if you want your self-driving car to go to UBC, the car will ping the GPS satellites to get a suitable route. It will also likely periodically send information to the car company to help the car company gain some extra information and analytics on you**. In the context of privacy, is this something society should be concerned about?** Why or why not? When answering this question, it is important that your explanation be grounded in the concepts explored so far in this course. Your explanation should accurately bring together as many concepts as possible.

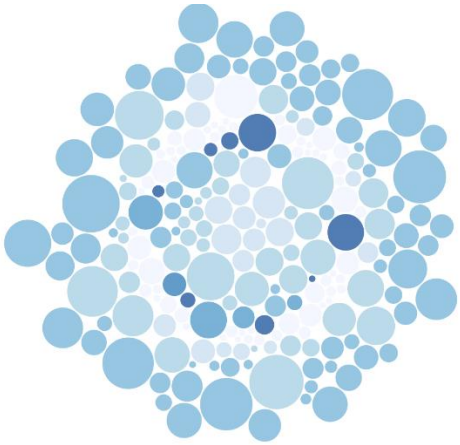
Your answer should not exceed 5 to 8 sentences.

Visual Encoding

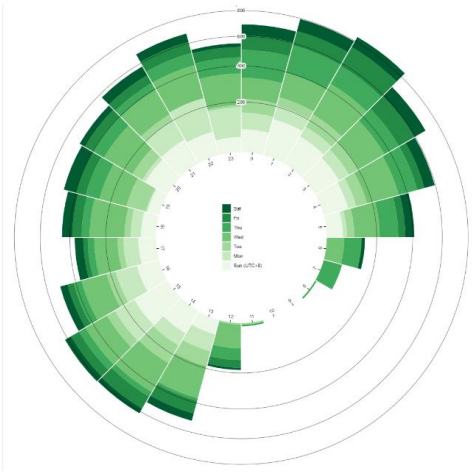
For the provided visualization. Answer the following questions

- List all the visual cues that are being used to encode data
- Describe how the visualization capitalizes on at least one gestalt principle
- Describe a real-world situation that would benefit from using this representation (in other words, what kind of task would benefit from the use of this representation)
- Describe one limitation of the representation

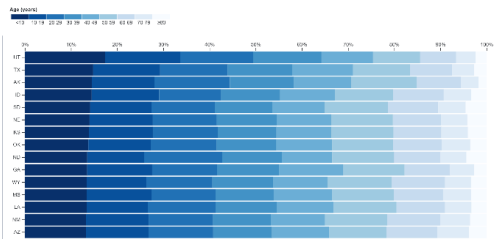
Option 1



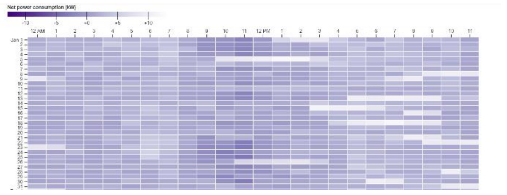
Option 2



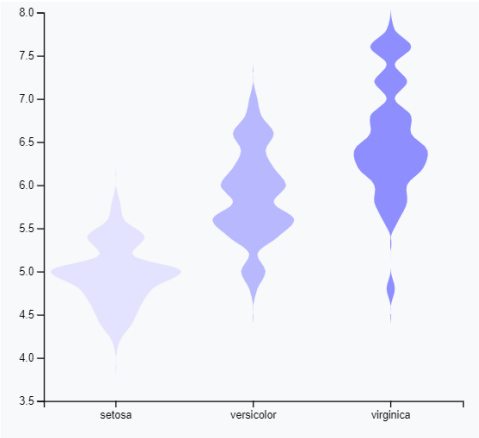
Option 3



Option 4



Option 5 [Alternate Sitting]



Association Rules

Using the Apriori algorithm described in class, for the given dataset, determine the list of frequent itemset of size 1 and size 2 when the minimum support is X%.

After determining the frequent itemsets, use the itemsets of size 2 to calculate the confidence for possible rules that may exist. For instance, if you have frequent itemset of {A, B} you must calculate the confidence for A -> B and B -> A

To receive full points, you must do the following:

- Do not calculate the support for itemsets of size 1 and 2 if you do not have to.
- Clearly indicate the itemsets of size 2 that do not need to have their support calculated.
- Calculate the confidence rules that exist based on the items in the frequent itemsets of size 2.

Here is an example to demonstrate the format of your answer. We reserve the right to deduct marks for answers that are not in this format. The transactions table given here serves as an

example and should not be considered when you are completing your answer. In addition, the answers below are possibly NOT correct. The only thing correct is the template for your solution.

Transactions	Items
T1	apple, dates, rice, corn
T2	corn, dates, tuna
T3	apple, corn, dates, tuna
T4	corn, tuna

Support for itemsets of size 1:
{apple} = 2/4
...

Itemsets of size 2 that do not need to have their support calculated:
{apple, rice}
...

Support for itemsets of size 2 that need to be calculated:
{apple, corn} = 2/4
...

Confidence for the rules that exist:
apple -> corn = 2/4
…

Option 1

Minimum Support = 60%

Transaction ID	Items
T1	notebook, sharpie, ruler, sticky note
T2	sharpie, notebook, ruler, eraser
T3	pencil, notebook, eraser, pen
T4	eraser, pen, sharpie, notebook
T5	pen, eraser, sharpie, ruler, notebook
T6	pencil, ruler

Option 2

Minimum Support = 60%

Transaction ID	Items
T1	pencil
T2	ruler, pencil, notebook, sticky note
T3	ruler, sharpie, notebook, eraser, sticky note, pen
T4	sharpie, pen, sticky note, pencil
T5	eraser, ruler, sticky note, sharpie
T6	sticky note, sharpie, ruler, pencil

Option 3

Minimum Support = 50%

Transaction ID	Items
T1	notebook, sticky note, pencil, eraser, sharpie
T2	sticky note, ruler, notebook, pen
T3	eraser, sticky note, sharpie
T4	sticky note, pencil, ruler
T5	notebook, eraser, pencil
T6	pencil, pen, eraser, sharpie

Option 4

Minimum Support = 50%

Transaction ID	Items
T1	ruler, notebook, pencil
T2	eraser, ruler, sharpie
T3	sticky note, notebook, pencil, eraser
T4	notebook, eraser
T5	ruler, pencil, eraser, sticky note
T6	notebook, ruler, pen, pencil

Option 5 [Alternate Sitting]

Minimum Support = 40%

Movie ID	Actors
M1	clooney, aniston, jolie
M2	roberts, aniston, sorkin, jolie
M3	aniston, damon
M4	damon, roberts, sorkin, clooney
M5	jolie, clooney, sorkin, aniston, roberts, damon
M6	damon, pitt, aniston, clooney, jolie
M7	jolie, aniston
M8	roberts, sorkin