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?

Classification because there are three possible outcomes that we are sorting students into.

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?

Classification because there are a limited number of possible outcomes.

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?

Clustering because you do not know the number of resulting groups. There is no labelled data that shows us the result (i.e., there is no test data that tells us who was cheating). Since we don’t already know who may have cheated, classification can’t help us.

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?

Association rules because you want to see the relationship between two things.

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? com
2. What is the top-level folder? sites
3. Which folder contains the image? 26
4. What is the name of the image file? 94f0ea-821_image_hires_19.jpg

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? com
2. What is the top-level folder? ell.h-cdn.co
3. Which folder contains the image? 2560x1280
4. What is the name of the image file? landscape-141-s-gettyimages-930-master-lead.jpg

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? com
2. What is the top-level folder? republic-prod
3. Which folder contains the image? xxdpi
4. What is the name of the image file? jdoj0in-1558.jpeg

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? com
2. What is the top-level folder? sites
3. Which folder contains the image? 06
4. What is the name of the image file? df6-5d2b-11ea-image-hires-132.jpg

Option 5

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

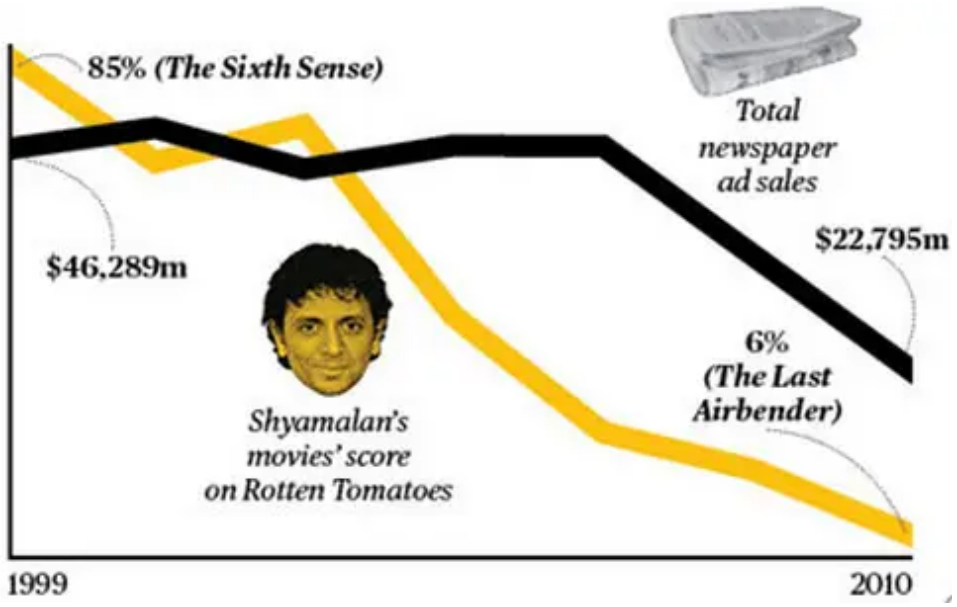
1. What is the highest or (top-level) domain? com
2. What is the top-level folder? content
3. Which folder contains the image? 48nE-K-QXox0-W7i2zEA
4. What is the name of the image file? 4543-453b-image-asset.jpeg

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.



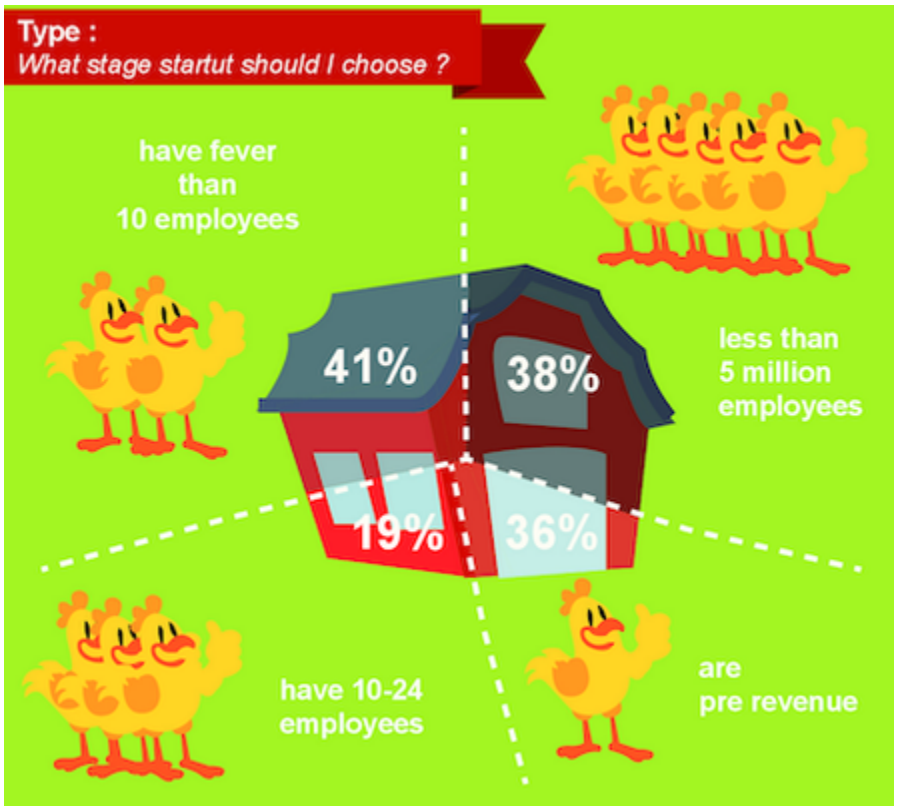
1. Correlation between Shyamalan's movies and total newspapers ad sales
2. No labeling of the y axis (either of them)

3.

Option 2 [Alternate Sitting]

What are three things wrong with this visual representation

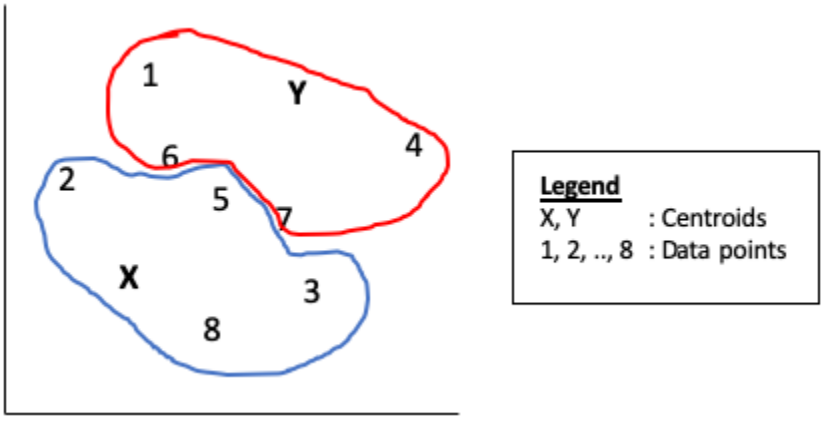
Note that this is NOT an infographic, it is just a static visual representation so we are not asking you to talk about the 5 principles of infographics here.



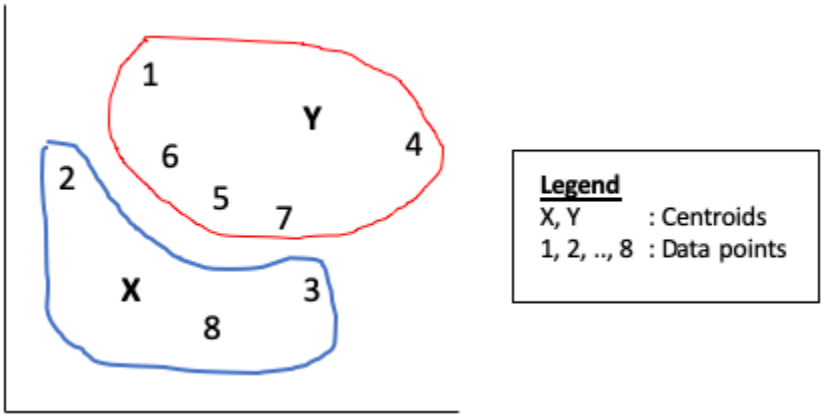
1. The proportions of the house
2. Why are they using chickens
3. Why is the 38% greyed out
4.

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

We stop as the clusters have not changed.

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

We stop as we are stuck in a loop.

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

We have to do another round of k-means as the clustering has changed.

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.

- Fake news websites often use URLs that look similar to, but are not the same as, well known media organizations. The example we discussed in class was with cnn.com vs. cnn.com.de.

- IP addresses are a way for us to indicate which specific computer we are referring to/want to connect to. The IP address for a fake news website like cnn.com.de would be different from the IP address for cnn.com.

  OR

  IP addresses are unique numbers that identify devices connected to the internet. They can be used to track users and identify roughly where they are. With this info, fake news websites can target demographics that are more vulnerable to misinformation.

- Routers are responsible for passing packets from one computer to another (e.g., from a server to a client). The return address and destination of a packet is indicated by the IP address. When you access a fake news website, the destination address is different so you are receiving data from a different server than what you might expect.

- To look for legitimacy in a website, you can do things like look at the URL and compare that to the URL you see from other sources. For example, social media accounts often list websites in the biography section; look to see if the social media account is verified and then look at the biography.

  You can Google the new organization's name to see if the first Google result has a URL that matches what you are looking at. Note this is different from the idea that fake news websites often have high ranking search results when you Google a specific news event. When you google a well known new organization's name (e.g., cnn, reuters, etc.), the first search result will not be the fake news version of that website.

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.

- Phishing Scams typically have websites that look similar to the organization they are trying to pose as. They may also try to send emails that look very similar to emails you would expect from real organizations, like your bank or a social media platform. In the lectures, we covered an email from Apple. The scams typically try to get you to reveal some personal information or download some file that can harvest your data from your computer

- IP addresses are a way for us to indicate which specific computer we are referring to/want to connect to. The IP address for a phishing scam for a bank would be tdbank.ca vs td.com/ca. Phishing scams typically hide the destination by providing users with a similar domain name they may be familiar with. It may try to take advantage of first perceptions such as misspelling errors (e.g., google and goggle) or similar looking characters (e.g., O instead of a 0).

- Email headers help reveal the real destination and sender of the email.

- Routers are responsible for passing packets from one computer to another (e.g., from a server to a client). The return address and destination of a packet is indicated by the IP address. When you receive a phishing email, they typically have you send your data (or reply to the email) but the destination is different from the place it appeared to come from.

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.).

Medium:

- Static visualization as I am presenting to the community and the prices are just one part of my report
- **Alternate Answer**: Interactive visualization because I will be in control and I will use the visualization to emphasize certain aspects.

  NOTE: Giving the users an interactive visualization is not suitable. They will not have time to learn how to use the viz, they will also not follow along with your presentation and miss key ideas.

Representation:

- Stacked bar or grouped bar chart because it will allow me to show the cost has changed for different types of properties. Multi-line charts for the change across time.

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**

Assumption: I am disseminating the research through mail

Medium:

- If you assume that you are disseminating the research through mail, then an Infographic would be appropriate as there is one main topic I wish to convey and the members of the research community don't need access to the raw data for exploration.
- If you assume that you are presenting research at a conference or you are publishing an article, then you would use a static visualization as we are only sharing some of our findings.

Representation: Chrolopleth maps because I am focused on regional differences 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.

Medium:

- If you assume that you will not present to the students but are distributing the information via leaflet, use an infographic. We use an infographic when we want to elicit some type of behavior. The end result is that we want students to make healthier food choices so an infographic around the subject of balanced diet would be the way to go. In addition, the attention span of teenagers is short, we want to get the message across as fast as possible.
- If you assume that you will be giving a lecture, then use a static visualization.
- If you assume that you will be doing some sort of booth (e.g., in a hallway for a long period of time) or a website, then use an interactive visualization.

Representation:

- For comparison of calories in healthy and unhealthy options I would use a bar chart. I will also use some type of relational diagram (Radar Chart, Parallel Coordinates) to show how a balanced diet impacts the various aspects of their health.

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 l**ast 15 year**s 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.

Medium:

- Either a static visualization or an infographic can be used.
- Static visualizations would be used if you assume that the webinar includes other topics that have not been fully laid out in the explanation.
- Infographics would be used on an assumption that I want all the factory owners to leave with a detailed understanding of the impact of hazardous waste on the environment

Representation:

- Choropleth map since we are dealing with regional differences of pollution
- Map with embedded charts
- Multiline graph
- Multiarea chart

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.

Medium: Interactive VIsualization because they need to work with the data

Representation: chord diagram, parallel coordinates, radar chart, grouped bar chart, because we need to be able to see the relationships between the doctors and the clients.

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.

Yes this is something that society should be concerned about. Some concerns would be:

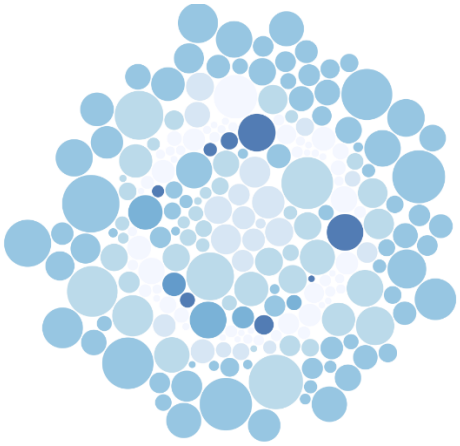
- Data ownership: Who should own your data? You or the car company?
- Privacy and data storage: Where is the data stored? Depending on who Your everyday movements can tell a lot about you like where you work, who you associate with, where you live, what hours you are active in a day. Are these pieces of information something you are comfortable storing on a server in a foreign country?
- Bias: How will data about your daily whereabouts be used? Will you be classified into a particular group based on what types of places you like to go to? Is this something that may cause you to be treated differently depending on marketing (e.g., if you live in an expensive part of town, will you be treated differently by the car company since they estimate you have more money?)
- Control over your car: Another company can now control where you can and cannot go. If they decide (either the company or the company has followed the request from a government agency) that you are not allowed to go to a particular place, you are forced to comply unless you bike or walk. Some locations can be so far that biking or walking is not a realistic method to use for travel.
- Encryption: Is the data sent between your car and the car company encrypted? If it is not, someone could take the data and use it for nefarious reasons like stalking or burglary (they can see when you aren't home). Also it's possible for your car to be remotely hacked and someone can take control of where your car is going without you knowing.
- Server Security: Are the servers where your data resides protected well? Do they have contingency plans for when the servers are attacked by hackers? Are their staff trained to not open random documents in case they download ransomware?
-

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



- Visual cues being used:
  - Area
  - Colour saturation (shading is NOT correct)
- Gestalt principles:
  - Proximity: All the circles are bunched together so they must represents parts of a whole
  - Similarity: Data of the same group has the same shade of blue
- Real-world situation: Sharing the number of calories of every food item in a vending machine. I would use size for the various calories groups (0 ~ 50) (50 ~ 100) (100~150) etc, circle for each snack, and the color for the type of snack (vegan, gluten free, etc)
- Limitation: This viz works for groups but is not really effective for the comparison of actual quantitative values.

Option 2

- Visual cues being used:
  - Position common scale
  - Position along nonaligned scale
  - Color saturation

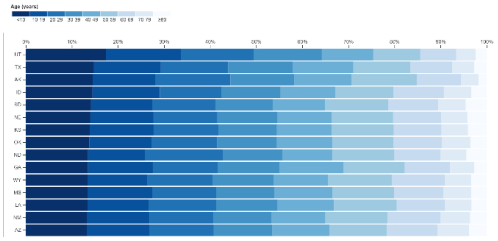
- Gestalt principles:
  - Proximity: Time is arranged next to each other 1am, 2am, etc.
  - Similarity: Data of the same day of the week has the same shade of green)
  - Connection/symmetry: All of the arcs stacked on top of each other represent a similar idea. For this viz, it would be what is happening at 2am
  - Closure: circle structure of the viz
  - Figure and Ground: axes are the background and the figure on top would be the stacked arcs

- Real-world situation: Communicating the internet traffic for every hour of every day in the week.

  Note: The answer should contain something about time and days of the week.

- Limitation
  - This viz is not the best for comparing the quantity across different hours in a day. For instance, I don't know if there are more people using the internet at 1am or at 2am on Saturday.
  - Saturation: the colors may make it hard to perceive where it starts or stops, so it hard to tell the days apart

Option 3



- Visual cues being used:
  - Color saturation
  - Position common scale
  - Position non-aligned scale

- Gestalt principles:
  - Similarity: age groups have the same color
  - Connection: all of the rectangles stacked on top of each other represent a similar idea. This shows that they are all related to the same state
  - Symmetry: The fact that they are all headed in the same direction

- Real-world situation: The percentage of the entire population for each state that falls within certain age groups.

  OR

  The proportion of individuals in each state that eat pizza for all the given age groups

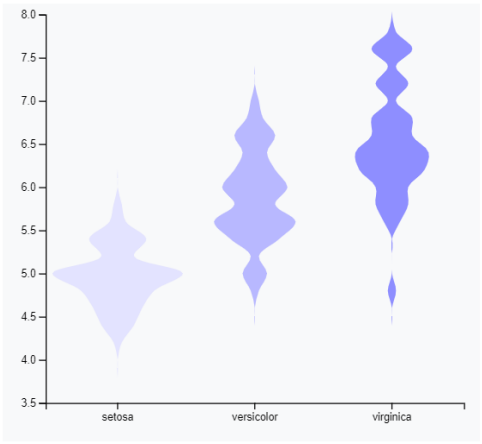
  Note: The answer should contain something about age and states.

- Limitation: It is hard to compare what is happening for a specified age group across states.

Option 4

- Visual cues being used:
  - Color saturation
  - Position on common scale
- Gestalt principles:
  - Similarity: Colour for net power consumption
  - Proximity: Those in the same row are related or the same column
  - Symmetry: Axes
- Real-world situation: Representing how much power was being used for the city of Vancouver for January 2021
- Limitation:
  - It is impossible to determine the exact value for 10am on January 1st.
  - Hard to navigate, dense vis.

Option 5 [Alternate Sitting]



- Visual cues being used:
  - Colour saturation
  - Position common axis
  - Area
- Gestalt principles:
  - Symmetry: Axis
  - Closure: See the parts as together
  - Similarity: The human eye views areas with the same color as representing the same data. For instance, for virginica, the small part at the bottom is not connected to the bigger portion and yet we will see them as part of the same data.
- Real-world situation: The distribution of heights or weights for different flower types
- Limitation:
  - What do the different areas mean
  - It is hard to determine if the versicolor weight is bigger than the virginica

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

Support for itemsets of size 1:

{notebook} = 5/6
{sharpie} = 4/6
{ruler} = 4/6
{pen} = 3/6
{sticky note} = 1/6
{pencil} = 2/6
{eraser} = 4/6

Itemsets of size 2 that do not need to have their support calculated:

{pencil, notebook}
{pencil, sharpie}
{pencil, ruler}
{pencil, sticky note}
{pencil, eraser}
{pencil, pen}
{sticky note, notebook}
{sticky note, sharpie}
{sticky note, ruler}
{sticky note, eraser}
{sticky note, pen}
{pen, notebook}
{pen, sharpie}
{pen, ruler}
{pen, eraser}

Support for itemsets of size 2 that need to be calculated:

{notebook, sharpie} = 4/6
{notebook, ruler} = 3/6
{notebook, eraser} = 4/6
{ruler, sharpie} = 3/6
{ruler, eraser} = 2/6
{sharpie, eraser} = 3/6

Confidence for the rules that exist:

Notebook → Sharpie = 4/5
Sharpie → Notebook = 4/4
Notebook → Eraser = 4/5
Eraser → Notebook = 4/4

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

Support for itemsets of size 1:

{pencil} = 4/6
{ruler} = 4/6
{notebook} = 2/6
{pen} = 2/6
{sticky note} = 5/6
{sharpie} = 4/6
{eraser} = 2/6

Itemsets of size 2 that do not need to have their support calculated:

{notebook, pencil}
{notebook, ruler}
{notebook, pen}
{notebook, sticky note}
{notebook, sharpie}
{notebook, eraser}
{eraser, pencil}
{eraser, ruler}
{eraser, sticky note}
{eraser, sharpie}
{eraser, pen}
{pen, pencil}
{pen, ruler}
{pen, sticky note}
{pen, sharpie}

Support for itemsets of size 2 that need to be calculated:

{pencil, ruler} = 2/6
{pencil, sticky note} = 3/6
{pencil, sharpie} = 2/6
{ruler, sticky note} = 4/6
{ruler, sharpie} = 2/6
{sticky note, sharpie} = 4/6

Confidence for the rules that exist:

Ruler → Sticky Note = 4/4
Sticky Note → Ruler = 4/5
Sticky Note → Sharpie = 4/5
Sharpie → Sticky Note = 4/4

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

Support for itemsets of size 1:

{notebook} = 3/6
{sticky note} = 4/6
{pencil} = 4/6
{pen} = 2/6
{eraser} = 4/6
{sharpie} = 3/6
{ruler} = 2/6

Itemsets of size 2 that do not need to have their support calculated:

{notebook, ruler}
{sticky note, ruler}
{pencil, ruler}
{eraser, ruler}
{sharpie, ruler}
{sharpie, pen}
{pen, notebook}
{pen, sticky note}
{pen, pencil}
{pen, eraser}
{pen, sharpie}
{pen, ruler}

Support for itemsets of size 2 that need to be calculated:

{notebook, sticky note} = 2/6
{notebook, pencil} = 2/6
{notebook, eraser} = 2/6
{notebook, sharpie} = 1/6
{eraser, sharpie} = 3/6
{sticky note, pencil} = 2/6
{sticky note, eraser} = 2/6
{sticky note, sharpie} = 2/6
{pencil, eraser} = 3/6
{pencil, sharpie} = 2/6

Confidence for the rules that exist:

pencil → eraser = 3/4
eraser → pencil = 3/4
eraser → sharpie = 3/4
sharpie → eraser = 3/3

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

Support for itemsets of size 1:

{notebook} = 4/6
{sticky note} = 2/6
{pencil} = 4/6
{pen} = 1/6
{eraser} = 4/6
{sharpie} = 1/6
{ruler} = 4/6

Itemsets of size 2 that do not need to have their support calculated:

{sticky note, sharpie}
{sticky note, notebook
{sticky note, pencil}
{sticky note, eraser}
{sticky note, ruler}
{sticky note, pen}
{sharpie, notebook}
{sharpie, pencil}
{sharpie, eraser}
{sharpie, ruler}
{sharpie, pen}
{pen, notebook}
{pen, pencil}
{pen, eraser}
{pen, ruler}

Support for itemsets of size 2 that need to be calculated:

{notebook, pencil} = 3/6
{notebook, eraser} = 2/6
{notebook, ruler} = 2/6
{pencil, eraser} = 2/6
{pencil, ruler} = 3/6
{eraser, ruler} = 2/6

Confidence for the rules that exist:

ruler → pencil = 3/4
pencil → ruler = 3/4
notebook → pencil = 3/4
pencil → notebook = 3/4

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

Support for itemsets of size 1:

{clooney} = 4/8
{aniston} = 6/8
{jolie} = 5/8
{pitt} = 1/8
{roberts} = 4/8
{sorkin} = 4/8
{damon} = 4/8

Itemsets of size 2 that do not need to have their support calculated:

{pitt, clooney}
{pitt, aniston}
{pitt, jolie}
{pitt, roberts}
{pitt, sorkin}
{pitt, damon}

Support for itemsets of size 2 that need to be calculated:

{clooney, aniston} = 3/8
{clooney, jolie} = 3/8
{clooney, roberts} = 2/8
{clooney, sorkin} = 2/8
{clooney, damon} = 3/8
{roberts, sorkin} = 4/8
{roberts, damon} = 2/8
{sorkin, damon} = 2/8
{aniston, jolie} = 5/8
{aniston, roberts} = 2/8
{aniston, sorkin} = 2/8
{aniston, damon} = 3/8
{jolie, roberts} = 2/8
{jolie, sorkin} = 2/8
{jolie, damon} = 2/8

Confidence for the rules that exist:

aniston → jolie = 5/6
jolie → aniston = 5/5
roberts → sorkin = 4/4
sorkin → roberts = 4/4