1. Mention and explain three uses of clustering in data visualization.

Clustering is the process of making a group of abstract objects into classes of similar objects.

Points to Remember

- A cluster of data objects can be treated as one group.
- While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
- The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

Applications of Cluster Analysis

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

Requirements of Clustering in Data Mining

The following points throw light on why clustering is required in data mining −

- **Scalability** − We need highly scalable clustering algorithms to deal with large databases.
- **Ability to deal with different kinds of attributes** − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
- **Discovery of clusters with attribute shape** − The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
- **High dimensionality** − The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
- **Ability to deal with noisy data** − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- **Interpretability** − The clustering results should be interpretable, comprehensible, and usable.

Ref: https://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm

2. Mention and explain three uses of force-directed algorithms in data visualization.

A **Force-Directed Graph**, or Force-Based Graph, is a type of layout commonly used in a variety of application areas: network visualization, large graph visualization, knowledge representation, system management, or mesh visualization.

It is used to visualize the connections between objects in a network. By grouping the objects connected to each other in a natural way, a Force-Directed Graph is visually interesting and also makes it possible to discover subtle relationships between groups.

These graphs have the particularity of being drawn by algorithms called "force-directed graph drawing algorithms" or "force-directed graph drawing algorithms".

Good-quality results

At least for graphs of medium size (up to 50–500 vertices), the results obtained have usually very good results based on the following criteria: uniform edge length, uniform vertex distribution and showing symmetry. This last criterion is among the most important ones and is hard to achieve with any other type of algorithm.

Flexibility

Force-directed algorithms can be easily adapted and extended to fulfill additional aesthetic criteria. This makes them the most versatile class of graph drawing algorithms. Examples of existing extensions include the ones for directed graphs, 3D graph drawing,[6] cluster graph drawing, constrained graph drawing, and dynamic graph drawing.

Intuitive

Since they are based on physical analogies of common objects, like springs, the behavior of the algorithms is relatively easy to predict and understand. This is not the case with other types of graph-drawing algorithms.

Simplicity

Typical force-directed algorithms are simple and can be implemented in a few lines of code. Other classes of graph-drawing algorithms, like the ones for orthogonal layouts, are usually much more involved.

Interactivity

Another advantage of this class of algorithm is the interactive aspect. By drawing the intermediate stages of the graph, the user can follow how the graph evolves, seeing it unfold from a tangled mess into a good-looking configuration. In some interactive graph drawing tools, the user can pull one or more nodes out of their equilibrium state and watch them migrate back into position. This makes them a preferred choice for dynamic and online graph-drawing systems.

Ref: https://www.toucantoco.com/en/glossary/force-directed-graph.html#:~:text=A%20Force%2DDirected%20Graph%2C%20or,between%20objects%20in%20a%20network.

https://en.wikipedia.org/wiki/Force-directed_graph_drawing#Advantages

3. **What are stack (or stacked) graphs? What is their main use? How would you apply them for text visualizations?**

This type of visualisation depicts items stacked one on top (column) of the other or side-by-side (bar), differentiated by coloured bars or strips. Items are "stacked" in this type of graph allowing the user to add up the underlying data points. Stacked graphs should be used when the sum of the values is as important as the individual items.

Uses:
A stacked graph is useful for looking at changes in, for example, expenditures added up over time, across several products or services.

Stacked graphs are commonly used on bars, to show multiple values for individual categories, or lines, to show multiple values over time. Thus, stacked graphs must always work with positive values.

Stacked line graphs often show how quantities have changed over time.

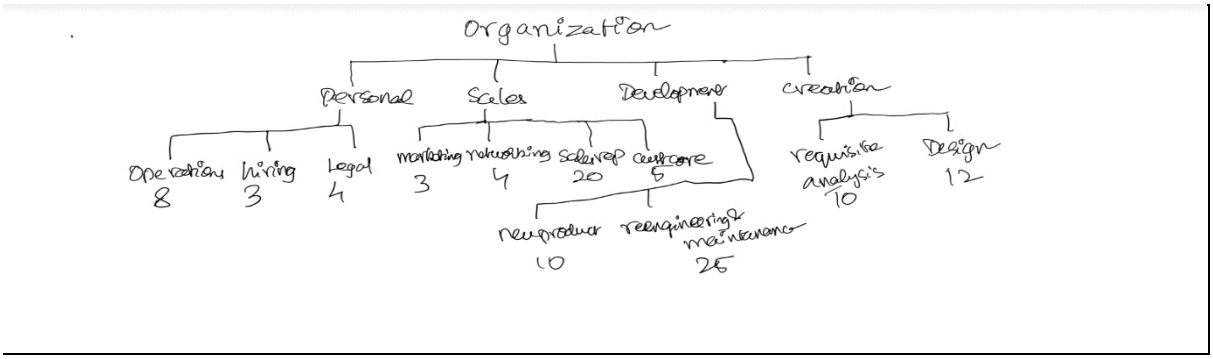
Yet to complete

4. **An analyst is organising a data set for managing the numbers of employees of the various departments of the company. The following text describes the data:**

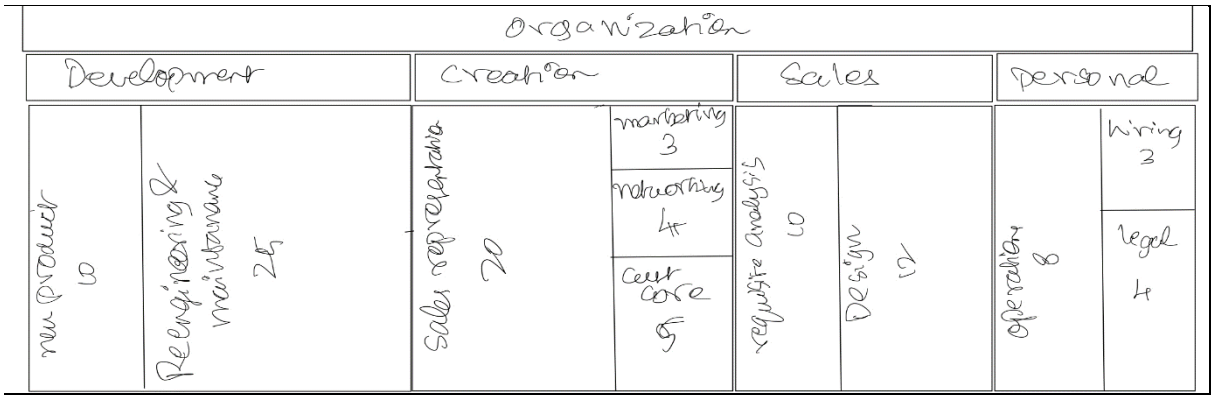
Spread Sheet:

Personal			Sales				Development		Creation	
Operations	Hiring	Legal	Marketing	Networking	Sales Representations	Customer Care	New Product	Reengineering and maintainance	Requisites analysis	Design
8	3	4	3	4	20	5	10	25	10	12

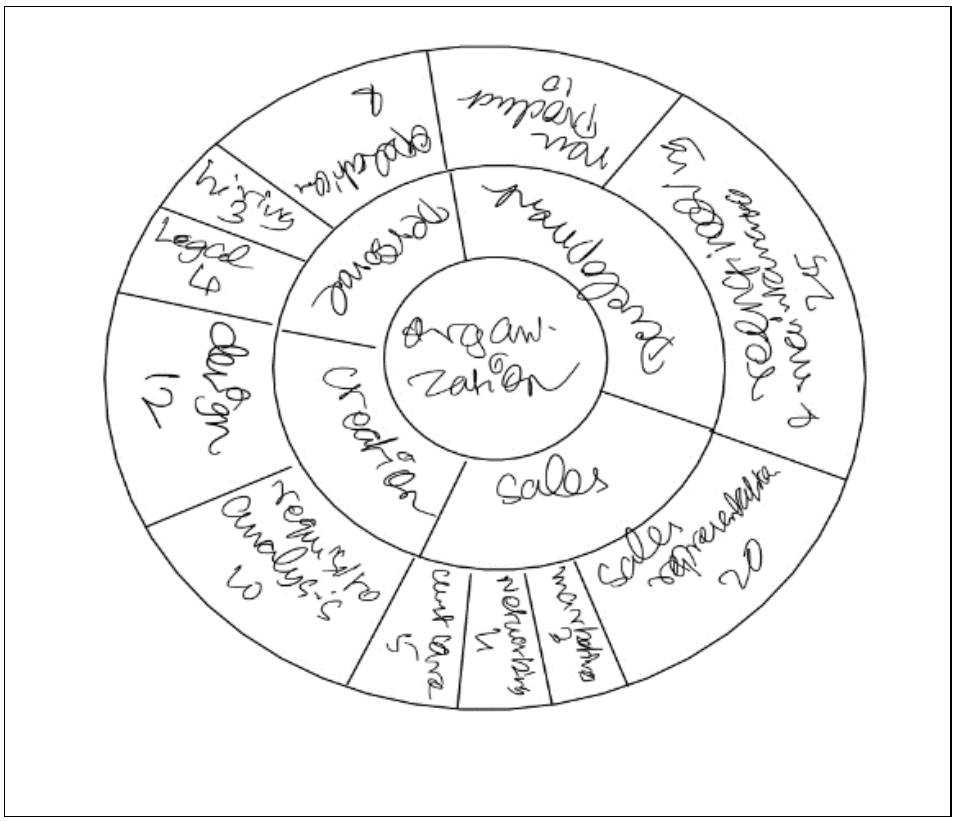
Node Link Tree:



Tree Map:



Sunburst:



5. **An analyst is designing a visualization support real estate agents to keep control of contents of houses they rent. She has built a fake data set to test alternatives. The following text describes the data:**

6. **Draw a comparative table of 5 multidimensional visualization techniques. What are they useful for? What type of observations do they provide? What is their input? What are the advantages and disadvantages? What applications could they support?**
   Sunburst, force-based graph, tree map, node link, Spreadsheet, stack graph
   Yet to do

7. **Visualization Task: Suppose you want to produce a visualization that summarizes and helps as a guide to a web-based course. The course is composed by series of linked web pages plus external links and references. Pages can be classes, tasks or additional information. The user must understand the general structure of the course, verify what pages he/she visited and for how long, and quickly find references given for each class.**

Describe the elements of your visualization and discuss the possible interactive functionalities the user would have in this context. Draw a schema for your visualization. Justify your choices of visual elements. What would change if the visualization were designed to support the teacher in improving the course?

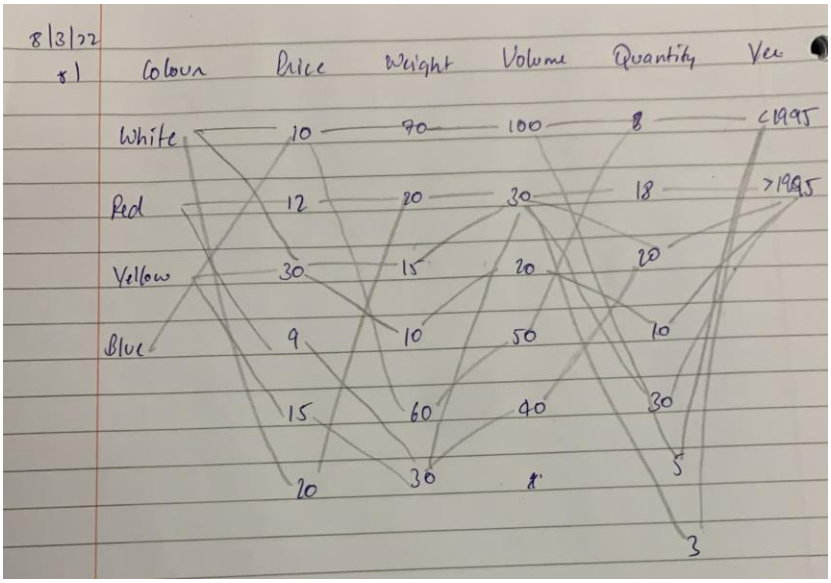
visual layout as in system
pages can be classes, tasks or information
distinction should be made (in dashboard)
associate with the proposed model
what visualisation used
target: find the way to pages
what we see, tree of a page with titles and
user would go hower the mouse above
can draw the scheme
functionalities: what user can do, interacts

Yet to do

8. **Visualization Task: Given the Data set in Table 1, draw the parallel coordinates visualization for it. What can you conclude about the data from the visualization?**

Table 1: Products

Color	Price	Weight	Volume	Quantity	Year Fab.
White	10	70	100	8	Before 1995
Red	12	20	30	18	After 1995
White	30	15	30	20	After 1995
Yellow	30	10	20	10	After 1995
Blue	10	60	50	8	Before 1995
Red	9	30	40	20	After 1995
Yellow	15	30	30	30	After 1995
Blue	10	70	100	5	Before 1995
White	20	20	30	3	Before 1995



9. What is the distinction you make between Information Visualization (InfoVis) and Scientific Visualization (SciVis)? Mention two common points and five distinct features between them.

Yet to Complete

10. What differs InfoVis and SciVis regarding data types?

	Scivis	Infovis
Data domain	spatial $\subset \mathbb{R}^n$	abstract, non-spatial
Attribute types	numeric $\subset \mathbb{R}^m$	any data types
Data points	samples of attributes over domain	tuples of attributes without spatial location
Cells	support interpolation	describe relations
Interpolation	piecewise continuous	can be inexistent

11. What is the distinctions between Direct Volume Rendering and Surface based Rendering?

Surface rendering relies on an assumption about the underlying structures you are visualizing from the data. In other words, you are estimating or making assumptions about what's underneath the surface based on its structure.

In volume rendering, assessing that underlying structure is part of the visualization process, and there is no such assumption. Instead, the nature of the data at each voxel is analyzed. Based on that analysis, colors and opacities are assigned, calculations are made, and the structures are visualized based on optical behavior of the components.

12. What are the distinctions of 2D reconstruction and 3D reconstructions and in what circumstance each one would be applied?

Yet to find

13. What are the main strategies to visualize hierarchies? What distinguishes them?

Tree Diagram

The idea behind the Tree Diagram hierarchical data visualization is to showcase relations of chart objects against each other in ranked order — it starts with a single value residing on top and all other values spiralling from it and their relations with each other.

It is often used in organizations to determine superiors and subordinates in a management system, showcase seniority in family trees, display the hierarchy of items according to their value like in Maslow's pyramid of needs, to fast-forward decision making and track the cause of a certain problem and its possible effects.

Treemap

Treemaps not only show value hierarchy by splitting the whole area into smaller rectangle pieces but also show value relations by obtaining rectangles of different sizes within each split category.

Visualizing hierarchical data using Treemaps is a great way to show items relations to the whole and to each other in a single system.

Sunburst Diagram

Here hierarchical data format is distributed in a way, quite similar to the Treemap, but instead of a rectangle, a whole circle is divided into separate categories that take up space in accordance with its value.

Sunburn Diagrams have multiple levels, which are using a hierarchical data structure and are placed one below the other in correspondence to each other's values.

Circle Packing

Circle Packing displays a whole area with a single circle and the biggest circles representing the main categories, with each smaller circle inside representing a value in proportion to the parent and the entire circle accordingly.

This visualized hierarchical data structure is best suited to display a large area with multiple elements making it up, and their relations to the whole and to each other.

REF: https://insightwhale.medium.com/how-to-show-hierarchy-with-data-visualization-526fb45ee4c2

14. Give an example of an application and a task or question for which a hierarchy is a natural application. Give an example of an application and a task or question for which a hierarchy can be used, but is non-native.

Native:

There is a hierarchical nature in every government organization of a country like parliament, army, etc. For example, in Army, we have Field Marshals -> Generals -> Lieutenant Generals -> Major General -> Brigadier -> Colonel -> Lieutenant Colonel -> Major -> Captain ->

Lieutenant. When we have two entities under same group, they cannot interact with each other as one is not dependent on another. Here, hierarchy is a natural application.

Non-Native:

Documents in an organization gets transferred from one department to another. This flow of documents through various departments can be represented by hierarchy but it is non-native because, the documents can go back to the departments from where it came.

15. Mention and explain three different types of layout for a tree, as well as their advantages and disadvantages.

Similarity trees:
They organize data objects on the visual plane emphasizing their levels of similarity with high capability of detecting and separating groups and subgroups of objects.
Advantages:
ability to decrease point clutter;
high precision;
consistent view of the data set during focusing, offering a very intuitive way to view the general structure of the data set as well as to drill down to groups and subgroups of interest.
Disadvantages:
their computational cost
presence of virtual nodes that utilize too much of the visual space.

Classification Trees:
Tree models where the target variable can take a discrete set of values are
called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.
Advantages:

1. Compared to other algorithms decision trees requires less effort for data preparation during pre-processing.
2. A decision tree does not require normalization of data.
3. A decision tree does not require scaling of data as well.
4. Missing values in the data also do NOT affect the process of building a decision tree to any considerable extent.
5. A Decision tree model is very intuitive and easy to explain to technical teams as well as stakeholders.

Disadvantage:

1. A small change in the data can cause a large change in the structure of the decision tree causing instability.
2. For a Decision tree sometimes calculation can go far more complex compared to other algorithms.
3. Decision tree often involves higher time to train the model.
4. Decision tree training is relatively expensive as the complexity and time has taken are more.
5. The Decision Tree algorithm is inadequate for applying regression and predicting continuous values.

Tree maps:
Treemaps display hierarchical (tree-structured) data as a set of nested rectangles. Each branch of the tree is given a rectangle, which is then tiled with smaller rectangles representing sub-branches. A leaf node's rectangle has an area proportional to a specified dimension of the data.[1] Often the leaf nodes are colored to show a separate dimension of the data.
Advantages:

1. When the color and size dimensions are correlated in some way with the tree structure, one can often easily see patterns that would be difficult to spot in other ways.
2. A second advantage of treemaps is that, by construction, they make efficient use of space.

Disadvantage:

1. Size Distortion: The more pixels you use to show the hierarchy, the more the size distorts. You quickly run into an issue that you can only optimize for size comparisons at one level of the hierarchy at a time.
2. Requires Interactivity: Treemaps are ill-suited for print. Unless you have only a few data points, the labels quickly disappear or become unreadable.

18. What insights can be drawn when visualizing text using: force-directed graphs, word trees and similarity trees? What is the purpose of word clouds?

Word trees:

Word trees show a pre-selected word(s) and how it is connected to other words in text-based data through a visual branching structure. Unlike word clouds, word trees visually display the connection of words in the dataset, providing some context to their use. Words that show up more frequently in combination with the pre-selected word(s) are displayed in larger font size. The visualisation allows users to choose whether they are interested in connections preceding a word or following a word.

Purpose of word cloud:

Word Clouds display the most prominent or frequent words in a body of text. Typically, a Word Cloud will ignore the most common words in the language ("a", "an", "the" etc). The remaining words are displayed in a "cloud" with the font size of the word (and-or the colouring of the characters in the word) depicting the relative frequency of occurrence of each target word in the source material. Word clouds helps in showcasing written data in a visual manner. It provides fast insights into the more relevant keywords in the data, summarises large volumes of text and reveal trends and find patterns in the data.

Yet to complete

19. What is the role of data summarization in InfoVis? Exemplify data summarization for two different types of data.

Summarising techniques like pattern detection, classification and clustering, dimensionality reduction, and aggregation can be used to compress large multi-dimensional datasets into smaller datasets which still retains the principal characteristics of the original data. Summarizing large datasets before visualizing the data, could help in more effective visualizations and support more efficient and accurate visual analysis.

1. The bivariate relationship between two categorical variables is summarized using a contingency table.
2. Quantitative data are grouped into classes. The lower limit of a class equals the smallest value within that class.
3. To summarize the relationship between a quantitative variable and a categorical variable, we calculate summary statistics for the quantitative variable for each level of the categorical variable

The term Data Summarization refers to presenting the summary of generated data in an easily comprehensible and informative manner.

We summarize data to "simplify" the data and quickly identify what looks "normal" and what looks odd. The distribution of a variable shows what values the variable takes and how often the variable takes these values.

The two most useful ways of describing the distribution of data are:

1. The typical: This describes the center–or middle–of the data. This way of describing the center is also called a “measure of central tendency”.
2. The spread of the values around the center: This describes how densely the data is distributed around the center. This is also called a “measure of dispersion”.

These two ways of describing the data are also referred to as descriptive statistics.

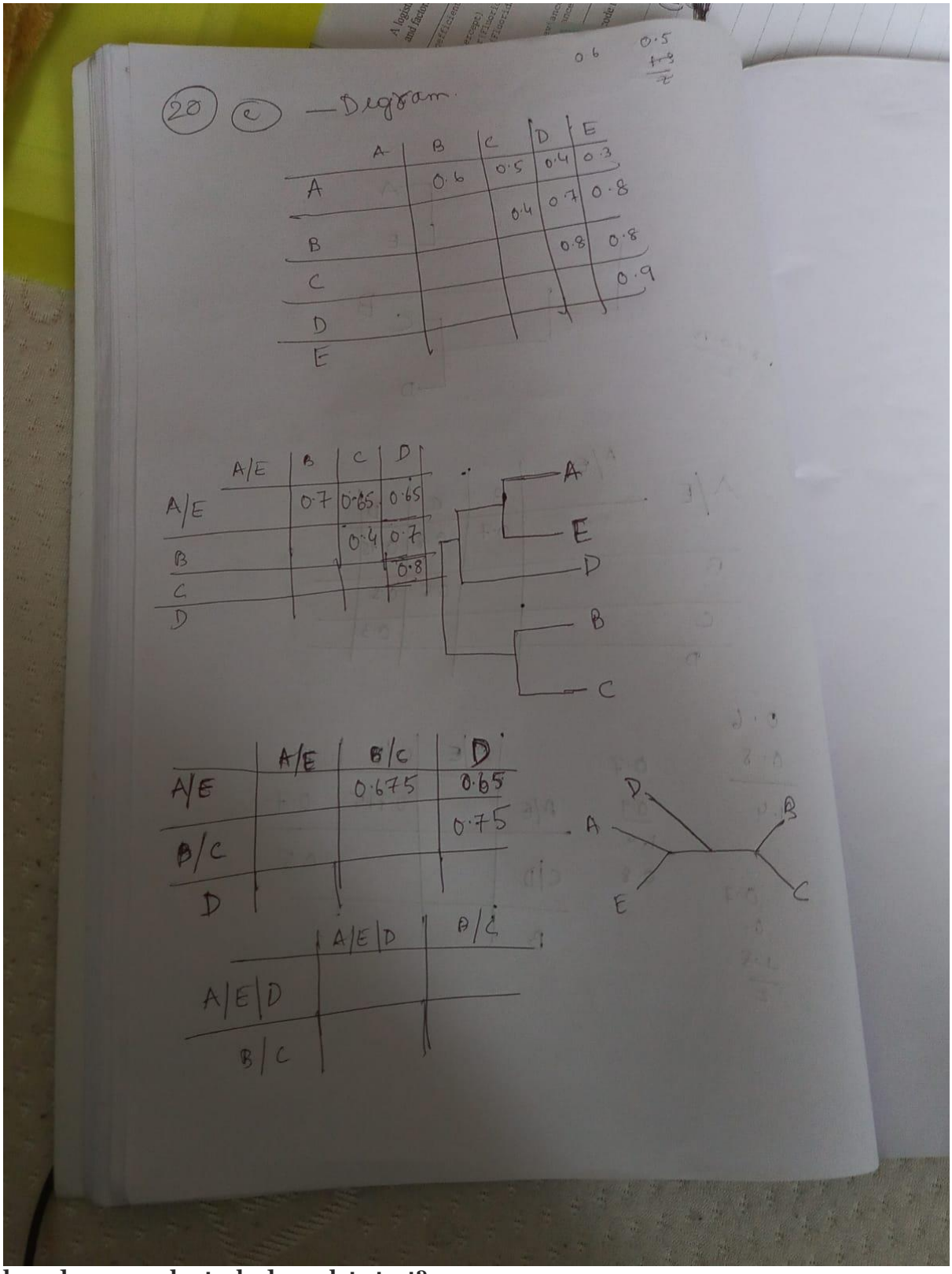
Types of data:

1. Categorical or Nominal
2. Ordinal
3. Continuous

20. (a) For the NJ-tree below, which of the Distance matrices CANNOT be the one that generated it? Why?

(b) Draw an NJ-tree for the answer in 20a

Yet to complete



how do you apply stacked graph to text?
theme river
- each one of these layer is how a topic changed over time
- process a doc along timeline and compile different topics along something
- no of times a doc/ word used in social discussion in a duration
- size of the flow is the count, flow itself(diff color) is a theme/ sub/particular partition of dataset(cluster)
- for eg 10 clusters than change over time

What insights can be drawn when visualizing text using: force-directed graphs, similarity trees?
- similarity tree: groups document (similar docs in same branch)
- graphs represented for text: node can be doc/ concept/ text/ theme/ topic, depending on the representation the connections of node can be
associations btw concepts/ topics
- depending on text applications in graph, using force directed displays will basically be ur similarity on edge then consistent groups of
labeled docs will be similar, in middle the docs will be which cannot be easily distinguishable

should you include nominal data in parralel co-ords? You can right but since it has no order it can give spurious patterns
- depends on nomial data, if too many cannot do as it should be associated with certain catergories
- eg survey data can be shown on parallel cords if wanted to show neighbourhoods, want to associated attributes, if attribute makes sense or think
its influencial in pattern then we should. Maybe we use 100 neighbours it can be too many