Q4) Given a hetrogneous data scenario,

where $X \in (\mathbb{R}^{n \times dr} \times I^{n \times di} \times C^{n \times dc})$

and $d = d_r + d_i + d_c$

where $\mathbb{R}$ → Set of real numbers

I → Set of integers

C → Categorical values $\{c_1, c_2 \dots c_K\}$

→ While learning a graph, learning with Real numbers and Integers will not create much significant problem.

→ But there is a need to handle categorical values explictly

→ For eg, if we map categories $\{c_1, c_2 \cdot \cdot c_{10}\}$ to ~~na~~ numerical values $\{1, 2 \dots 10\}$, can create issues while learning graph.

→ The issue is in this mapping. Category 9 (c_9) is very 'far' from category 1 (c_1) than it is from c_8.

So we need to do some encoding according to the given data setting so that we can map it to a justifiable encoding

There are mainly 2 kinds of encoding in the data science literature

1) Nominal encoding

This encoding is done when there is no particular order to any category

eg) Place names, Object names, ~~etc~~ location etc

2) Ordinal encoding

Used when ~~to~~ the categorical variable has some ordering

eg) {~~Good~~ Excellent, Good, Bad}

Example of Nominal encoding: One hot encoding

⇒ So after all these pre processing on X, we will get all numerical ~~values~~ attributes and we can learn graph on this transformed X