Problem 4 (Checking independence): Most of the techniques used in checking (conditional) independence between random variables are based on hypotheses testing. Formally, we have following two hypotheses:

$$H_0 = \{\text{Random variables are not (conditionally) independent}\}$$

and

$$H_1 = \{\text{Random variables are (conditionally) independent}\}.$$

By assuming the data follow some types of distribution, we can propose several types of metrics and approximate their distributions, in order to claim which hypothesis is true given our observations. The confidence level of our judgement can be measured using following metric, namely p-value:

$$p = P(H_1 \text{ appears to be true}|data,\ H_0 \text{ is actually true}).$$

Obviously, small p-value implies stronger confidence that H_1 is true. Typically, $p = 0.05$ or $p = 0.1$ is used as thresholds to determine whether accept H_1 or H_0. In this problem, I prefer to use G-test to determine whether two random variables are independent. See https://en.wikipedia.org/wiki/G-test. G-test is basically comparing the actually distribution and the product of two marginal distributions by counting each bin. The divergence between two distribution is measured as:

$$G(p, q) = 2\sum_x p(x)\log(\frac{p(x)}{q(x)}),$$

where $p(x)$ and $q(x)$ are number of observed counts in a bin, instead of a probability mass. Therefore it is different from Kullback-Leibler divergence.

For the implementation I used the code from an existing package: Python gsq package. The code I used in this problem can be found in the corresponding attachment.

The conclusion is summarized as following:

Marginal independence p-value of d0 and d1 is: 0.0
p-value of d0 and d2 is: 0.231658736846
p-value of d0 and d3 is: 0.90331010682
p-value of d0 and d4 is: 0.0
p-value of d1 and d2 is: 0.166504089458
p-value of d1 and d3 is: 0.535798622053
p-value of d1 and d4 is: 0.0
p-value of d2 and d3 is: 0.0253836754723
p-value of d2 and d4 is: 0.804517201563
p-value of d3 and d4 is: 0.218212152554

Therefore I am confident that (d0,d1), (d0,d4), (d1,d4) and (d2,d3) are pairs of independent random variables, and the p-value is the probability that I make such conclusion by mistake given that they are actually dependent.

As for 3-way mutual independence, the only possibility is (d0,d1,d4) since mutual independence implies pairwise independence. In order to verify that, we only need to check whether d1 and d4 are independent conditioned on d0. Because if that statement is true, we have the following:

$$P(d1)P(d4) = P(d1|d0)P(d4|d0) = P(d1, d4|d0) = P(d1, d4, d0)/P(d0),$$

then $P(d1, d4, d0) = P(d1)P(d4)P(d0)$. And the p-value is 0.0 so they are mutually independent.

Conditional independence p-value of d1 and d4 conditioned on d0 is: 0.0
p-value of d0 and d1 conditioned on d2 is: 0.0
p-value of d0 and d1 conditioned on d3 is: 0.0
p-value of d0 and d4 conditioned on d1 is: 0.0
p-value of d0 and d4 conditioned on d2 is: 0.0
p-value of d0 and d4 conditioned on d3 is: 0.0
p-value of d1 and d2 conditioned on d3 is: 0.0916463098153
p-value of d1 and d4 conditioned on d0 is: 0.0
p-value of d1 and d4 conditioned on d2 is: 0.0
p-value of d2 and d3 conditioned on d1 is: 0.0551626358929

There is no conditional independence when conditioned on 2 or more other random variables.