Currently i am researching into what a BN is, i have discrovered that the data must be discrete continous variables, after asking an ai engine what would be the best means of cleaning my dataset, there are 4 ways:

- **Equal-Width Binning:** Create bins of the same size
- **Equal-Frequency Binning (Quantiles):** Create bins with the same number of data points in each
- **Clustering:** Use an algorithm like K-Means to find natural "clusters" in the data to define the bins.
- **Domain Knowledge:** Use expert knowledge (e.g., a doctor defining "high" blood pressure as > 130)

Some of the attributes require “domain knowledge” for absolute precision, some would be best to use a clustering algorithm and some will require equal binning, i will have to make a python script that sorts the csv file of all 12 attributes using with predetermined sorting means.

New dataset

New dataset: https://archive.ics.uci.edu/dataset/45/heart+disease

Reference: “Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.”

Catagories

When creating a BN it is known best to catagorise what attributes are known:

Only 14 attributes used:

1. #3 (age)

2. #4 (sex)

3. #9 (cp)

4. #10 (trestbps)

5. #12 (chol)

6. #16 (fbs)

7. #19 (restecg)

8. #32 (thalach)

9. #38 (exang)

10. #40 (oldpeak)

11. #41 (slope)

12. #44 (ca)

13. #51 (thal)

14. #58 (num) (the predicted attribute)

Tier Sorting system

I have Separated it into 5 Tiers:

Tier 1: Root Cause

- Age
- Sex

Tier 2: Clinical Factors

- Trestbps (resting bp)
- Chol
- Fbs (fasting blood sugar)
- Restecg (resting ecg)
- Thal (thalassemia)

Tier 3: Exercise test results

- Thalach (max heart rate)
- Oldpeak
- Slope
- Ca (vessels)

Tier4: the disease (target)

- Num (central node, presence of heart disease)

Tier 5: The symptoms

- Cp (chest pains)

- Exang (exercise induced angina)

BN Mock draw up