REAL DATASET

- Scaling labels usually essential in a multi-feature model
- Split data set into two subsets: training vs. test. The test set should be large enough to yield meaningful results and be representative of data as a whole
- Features (even synthetic ones) may not correlate well with the labels. May need trial and error, or...

Correlation Matrix: shows how each attribute (feature) relates to the others (i.e. their values).

1.0 = perfect positive (both rise together)
-1.0 = perfect negative (one ↗ one ↘ together)
0.0 = no correlation, not linearly related

Higher abs. value ⇒ higher predictive power

★ Some features may raise ethics and fairness issues!