Lab 1

Dataset analyses

EuroVision Song Contest Dataset

Pros

- There are nearly 50,000 rows.
- The dataset is complete; it covers the number of votes that each country awarded to each other country.
- The dataset covers historical data allowing us to analyse patterns.

Cons

- The data is trivial. By applying machine learning we are unlikely to gain any insights that we could not have gained through the use of more standard statistical techniques.

Dry Bean Dataset

Pros

- The dataset has 13611 instances covering beans of six classes. This is a lot of data for our model.
- The dataset is fairly balanced. The counts for each class are as follows:
  - SEKER: 2027
  - BARBUNYA: 1322
  - BOMBAY: 522
  - CALI: 1630
  - HOROZ: 1928
  - SIRA: 2636
  - DERMASON: 3546
- The species of the bean based on the physical properties is not obvious. This make ML very useful.
- There are 17 attributes (including the class). These mostly pertain to the dimensions of the dry beans as well as other physical qualities such as solidity and compactness. These are all relevant features for identifying the species of different beans.
- The dataset contains both numerical and nominal data

Cons

- It's not obvious why the ML model categorises beans the way it does. We have little bean intuition making it harder to decide on the most sensible approaches.

YouTube Dataset of Different Countries

- There are 24427 rows covering 20 countries. This is a lot of data for our model.

Cons

- There are no errors or outliers.
- There aren't many things we can use ML for.
- There are likely to be many other models for covering similar datasets.