1. *Visual Data Analysis.* Given the dataset “visual-dataset.csv” (available in TeachCenter), which comprises a number of features and a binary target feature.

   (a) Provide a number of meaningful visualisations (3-5 visualisations)

   (b) Based on the visualisations provide your key interpretation on

      i. Are there noteworthy dependencies between the features?

      ii. What types of dependency/relationship are there?

      iii. Are we expecting the prediction to work well?

2. *Correlation.* Given a dataset, which consists of 1,000 variables (hint: most of them are just random), the goal is to find the relationships between variables, i.e., which and how do the variables relate to each other; what are the dependencies. The dataset “correlation-dataset.csv” can be downloaded from TeachCenter.

   (a) Which methods did you apply to find the relationships, and why?

   (b) Which relationships did you find and how do you characterise the relationships (e.g., variable “Michael” to “Christopher” is linear)?

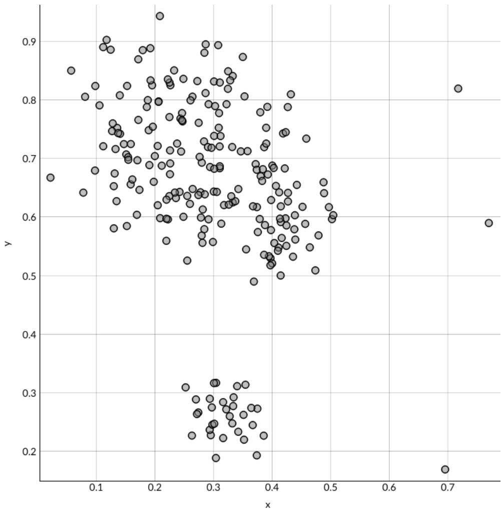
   (c) Which causal relationships between the variables can you find (e.g., variable “Jessica” causes “Matthew”)?

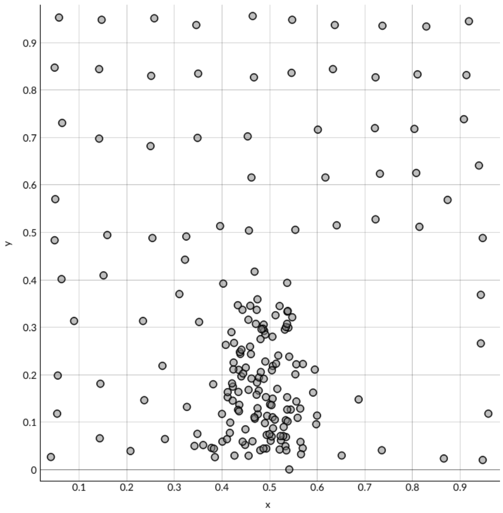
3. *Outliers/Anomalies.*

   (a) For both data sets shown below define yourself, what is the normal behaviour and what are the outliers/anomalies, please indicate in the image the anomalous behaviour

   (b) Name the algorithms or describe the algorithmic way of how to identify this anomalous behaviour (you may choose different algorithms for each data set, also describe any necessary preprocessing)

   (c) Name the assumptions made by your algorithms





4. *Missing Values.* The dataset "missing-values-dataset.csv" (available on TeachCenter) contains a number of missing values.

   (a) Try to reconstruct why the missing values are missing? What could be an explanation?

   (b) What methods do you apply?

   (c) What strategies are applicable for the features to deal with the missing values?

   (d) For each feature provide an estimate of the arithmetic mean (of the version of the dataset without missing values)?