1. *Visual Data Analysis.* Given the dataset "visual-dataset.csv" (available in TeachCenter), which comprises a number of features taken from Climate Watch. Provide a number of meaningful visualisations (4 visualisations) that show key properties of the dataset and dependencies. Based on the visualisations provide your interpretation. Ideally, one would like to identify some dependency on the yearly values, or the change in yearly values.

   (a) What pre-processing did your do? (e.g., Did you create new features? Did you normalise the data? Did you filter the dataset? Extended with another dataset?)

   (b) What are the most relevant dependencies between the features (selection of the figures)?

   (c) What types of dependency/relationship are there (per figure)?

   (d) Provide a summary of the main dependencies

   **Answer (a)** - Preprocessing steps:

   - First preprocesing step
   - Second step
   - ...

   **Answer (b) and (c)**

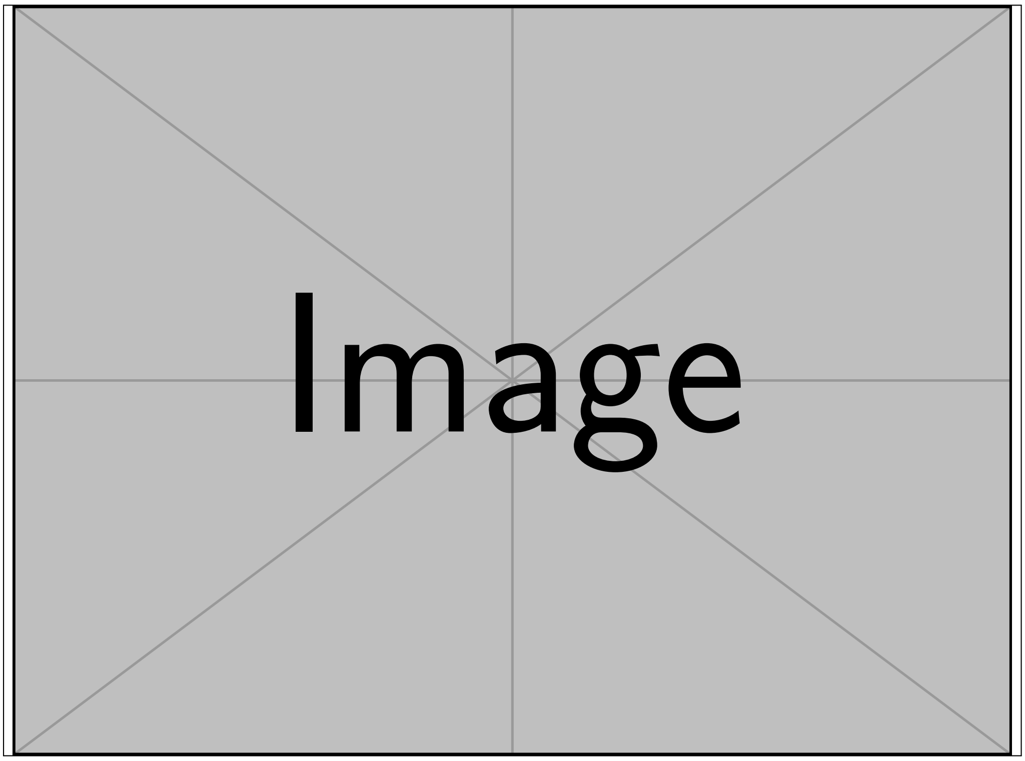


Figure 1: Please provide an explanation for the visualisation - (i) Why this kind of visualisation? (ii) What kind of dependency is being shown? (iii) What are potential interpretations?

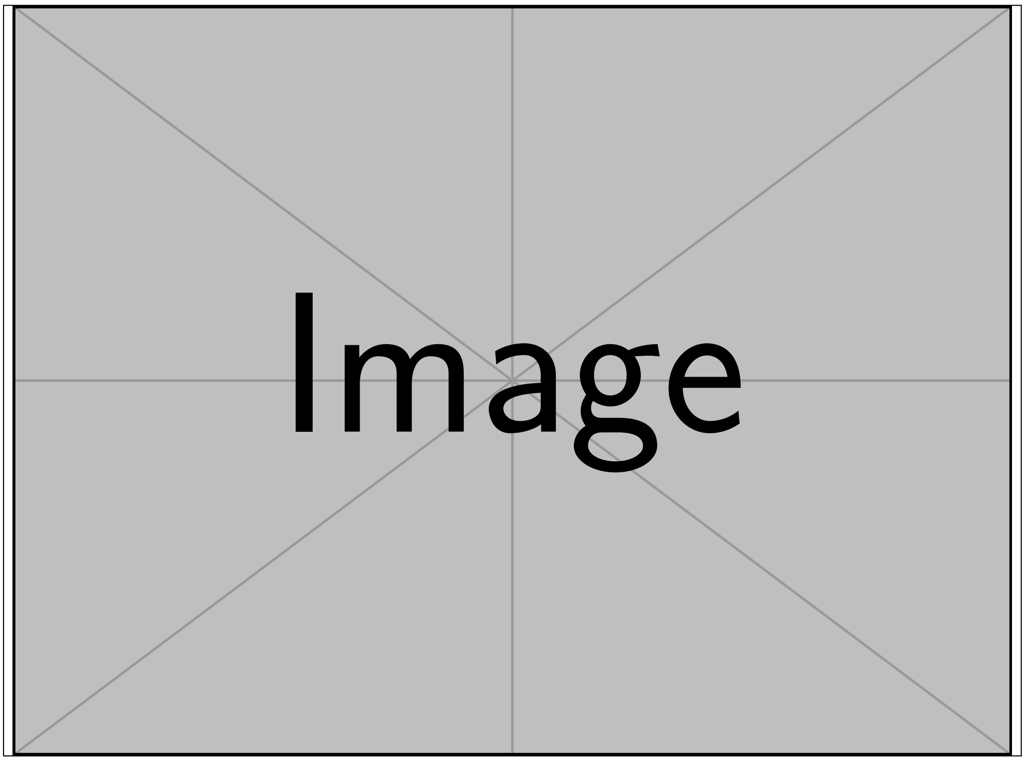


Figure 2: Please provide an explanation for the visualisation

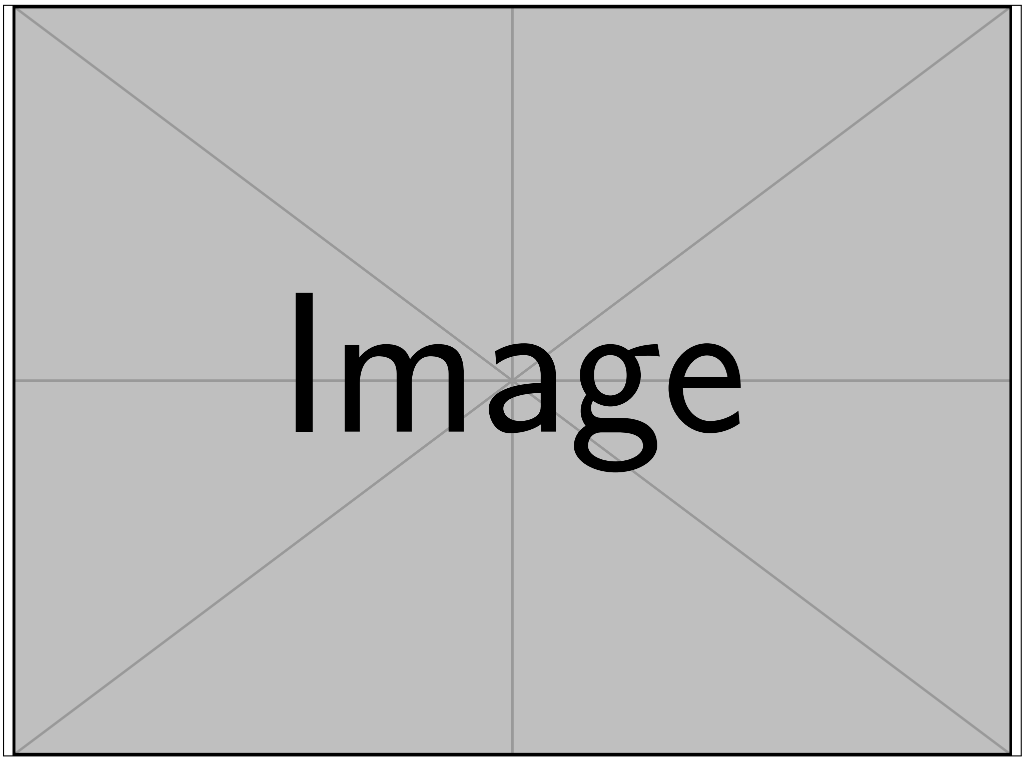


Figure 3: Please provide an explanation for the visualisation

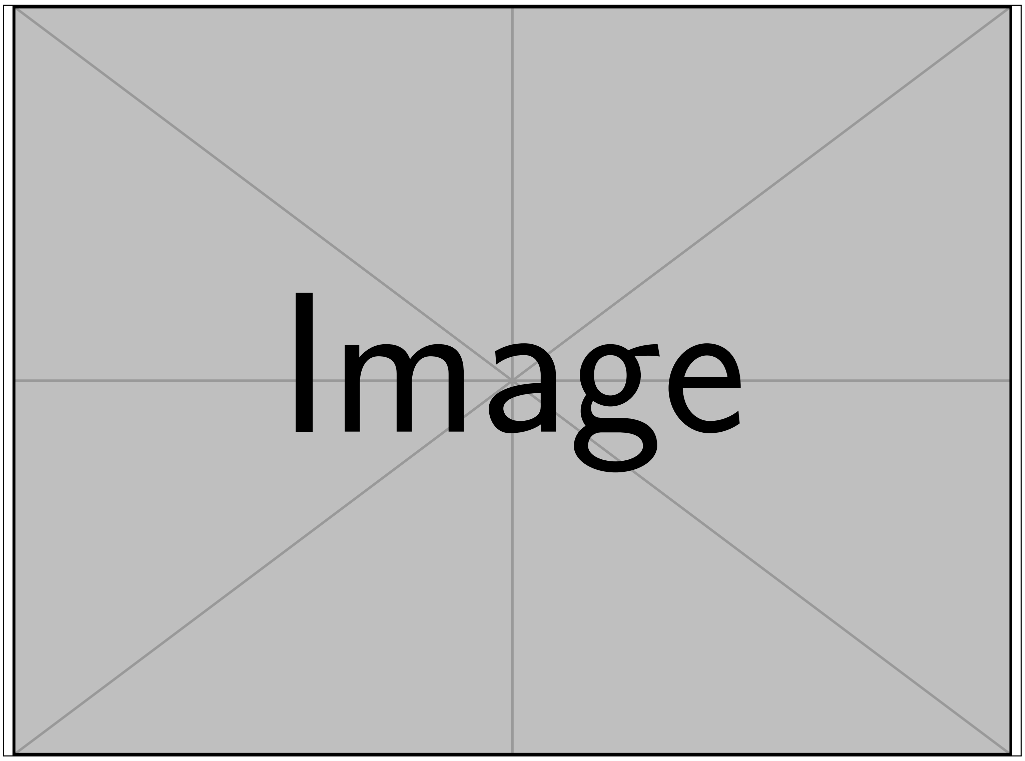


Figure 4: Please provide an explanation for the visualisation

**Answer (d)** - Short summary of the main findings

- Finding #1
- ...

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 “Munchen” to “Delmenhorst” is linear with correlation found via method X of 0.9)?

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

   **Answer (a)** - Method and motivation:

   (a) First Method - Why chosen (one sentence)

   (b) Second Method

   (c) ...

   **Answer (b)**

| Variable 1 | Variable 2 | Type of dependency | Method | Value |
|---|---|---|---|---|
| × | × | × | × | × |
| × | × | × | × | × |

   **Answer (c)**

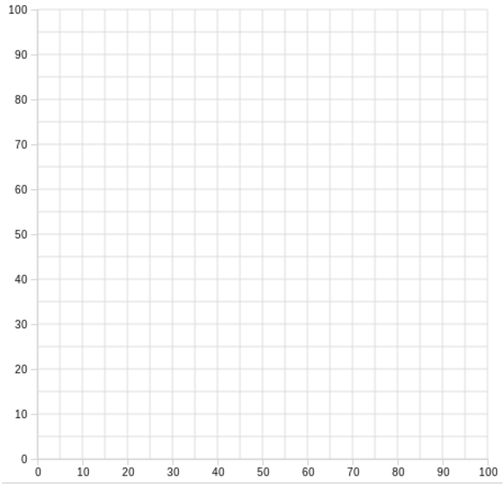
3. *Outliers/Anomalies.* Given two types of anomalies: (1) anomalies are defined to be in low density regions, (2) anomalies are regions of low density.

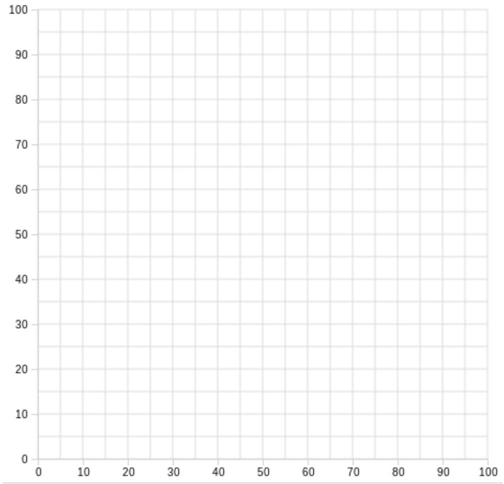
   (a) For both anomalies please create/draw a dataset, with 3 anomalies and many normal data points (the normal datapoints should be marked, e.g., green colour)

   (b) Name the algorithms or describe the algorithmic way of how to identify this anomalous behaviour (you may also describe any necessary preprocessing)

   (c) Name the assumptions made by your algorithms

**Answer (a)** - Draw two datasets





**Answer (b)** - Describe the algorithms

**Dataset 1** ...

**Dataset 2** ...

**Answer (c)** - Describe the main assumptions

| **Algorithm** | **Assumption** |
|---|---|
| × | × |
| × | × |

4. *Missing Values.* The dataset “missing-values-dataset.csv” (available on TeachCenter) contains a number of missing values. Try to reconstruct why the missing values are missing? What could be an explanation?

   (a) What are the dependencies in the dataset?

   (b) What could be reasons for the missingness?

   (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 (before and after applying the strategies to deal with missing values)?

**Answer (a)** - Describe the dependencies in the dataset

| X | Y | Type of dependency |
|---|---|---|
| × | × | × |
| × | × | × |

**Answer (b)** - Describe the reason for missingness

| Variable | Reason |
|---|---|
| × | × |
| × | × |

**Answer (c)** - Describe the strategies for dealing with missing values

| Variable | Strategy |
|---|---|
| × | × |
| × | × |

**Answer (d)** - Arithmetic mean of original dataset, and the one after applying the strategies

| Variable | Before Strategy | After Strategy |
|---|---|---|
| × | × | × |
| × | × | × |