1. *Visual Data Analysis.* Given the dataset "task1-dataset.ods" (available in TeachCenter), which comprises a number of features. Provide a number of meaningful visualisations (4 visualisations) that show key properties of the dataset and dependencies. Based on the visualisations provide your interpretation and insights.

   (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) Provide a series of meaningful plots that show a specific relationship (dependency) or characteristic of the dataset

   (d) Provide a summary of the main insights

   Answer (a) - Preprocessing steps:

   - First preprocesing step
   - Second step
   - ...

   Answer (b) - List of main dependencies:

   - First dependency
   - ...

   Answer (b) and (c)

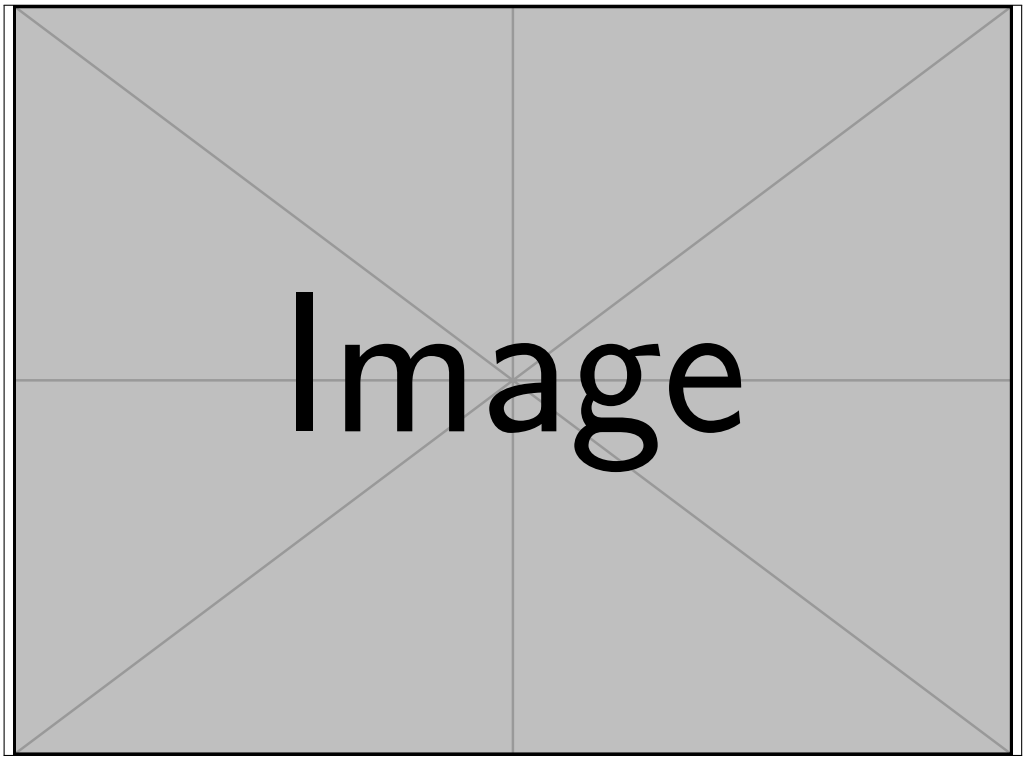


Figure 1: Please provide an explanation for the visualisation - (i) Why this kind of visualisation? (ii) What kind of dependency is being shown? (iii) What are potential interpretations?

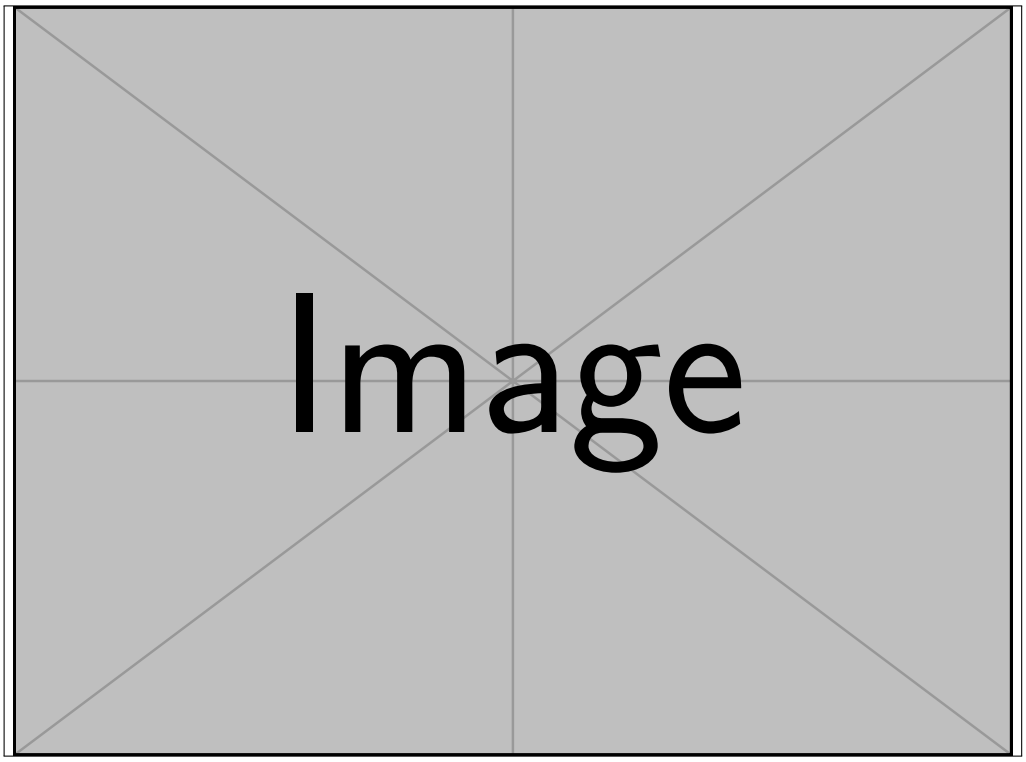


Figure 2: Please provide an explanation for the visualisation

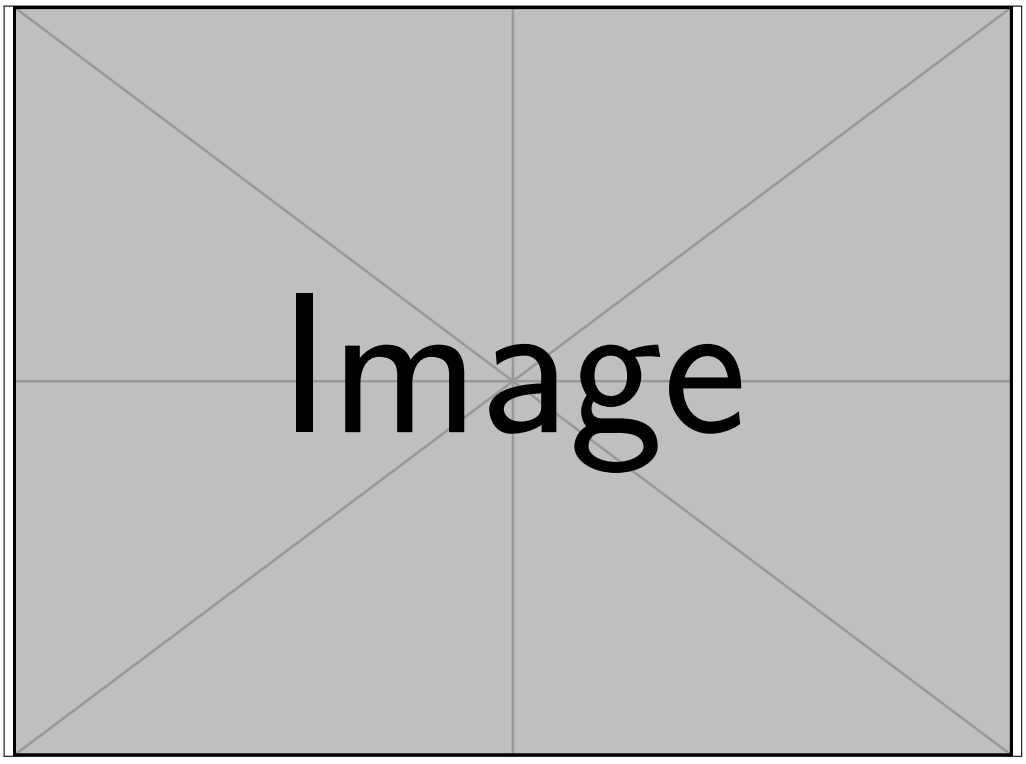


Figure 3: Please provide an explanation for the visualisation

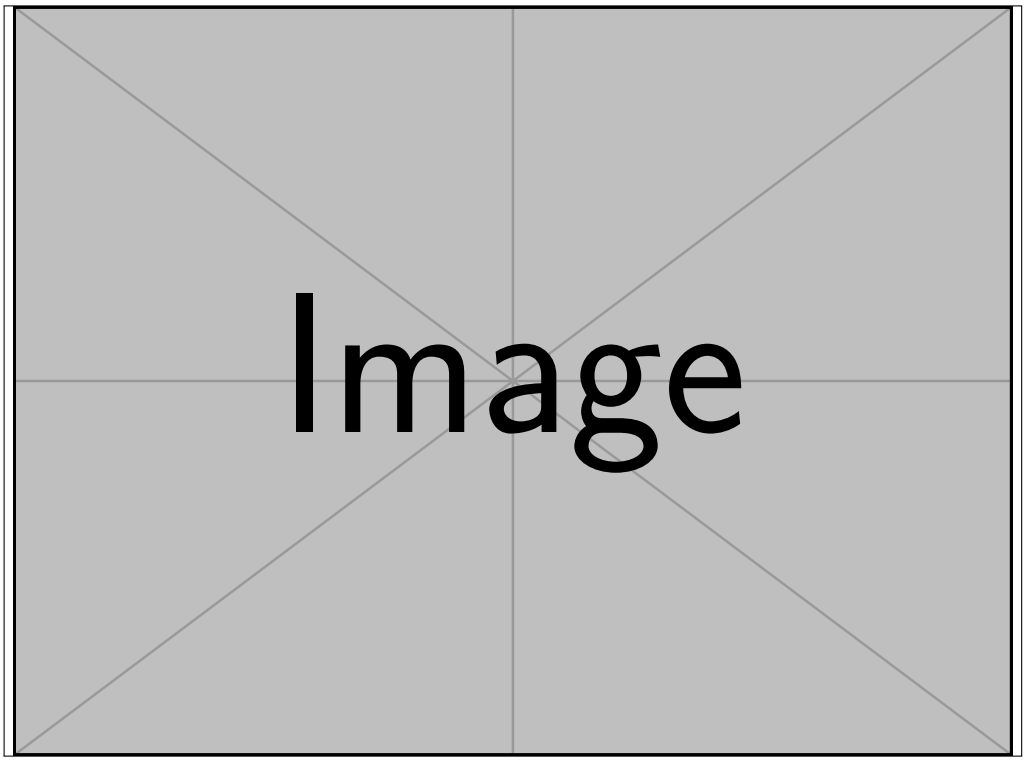


Figure 4: Please provide an explanation for the visualisation

Answer (d) - Short summary of the main insights (with references to the corresponding image)

- Finding #1, as illustrated in Figure 1 one ...
- ...

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 "task2-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 "Lurkowl (Strix umbra #1068)" to "Frosthawk (Accipiter glacies #1064)" is linear with correlation found via method X of 0.9)?

   (c) Which causal relationships between the variables can you find (e.g., variable "Rattlepuff (Lynx rattleus #1067)" causes "Slingshark (Carcharodon slingus #1068)")?

Answer (a) - Method and motivation:

(a) **Pearson Correlation**: Used to detect linear relationships between variables.

(b) **Spearman Correlation**: Used to capture monotonic relationships, which includes non-linear but ordered patterns

(c) **Mutual Information (MI)**: Used to detect both linear and non-linear dependencies by measuring the shared information between variables (but this showed no good result)

Answer (b)

Variable 1	Variable 2	Type of dependency	Method	Value
Puffpounce	Danglefawn	Linear	Pearson	0.9999
Slingshark	Shivershark	Linear	Pearson	0.9938
Crunchbeetle	Chirpsnail	Linear	Pearson	0.9922
Frogsnapper	Blurpglider	Linear	Pearson	0.9766
Mothglow	Munchsnail	Linear	Pearson	0.9665
Sparklequail	Scruffpaws	Linear	Pearson	0.9564
Golden Eagle	Chinese Dragon	Linear	Pearson	0.9544
Mossbeard	Windleopard	Linear	Pearson	0.9538
Fluffernox	Driftwolf	Linear	Pearson	0.9503
Moonlight Giraffe	Whiskerflare	Linear	Pearson	0.9447
Jellyfish	Mongoose	Monotonic	Spearman	0.8256
Goose	Dromedary Camel	Monotonic	Spearman	0.8049
Golden Eagle	Lion	Monotonic	Spearman	0.7749
Chinese Dragon	Golden Eagle	Monotonic	Spearman	0.7674
Ladybug	Mongoose	Monotonic	Spearman	0.5932
Chinese Dragon	Lion	Monotonic	Spearman	0.5909
Jellyfish	Ladybug	Monotonic	Spearman	0.5273

Answer (c) No significant Granger causality was found for any tested pair (all p-values > 0.05). This is likely due to the dataset's mostly random nature, which reduces the likelihood of true causal links

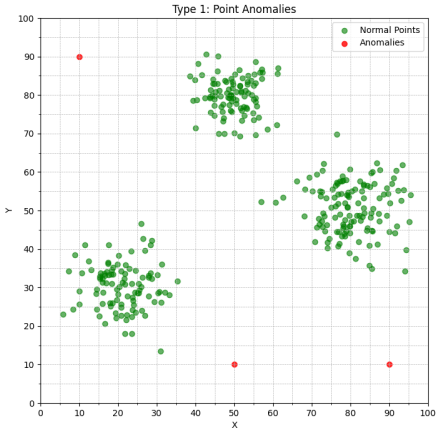
3. *Outliers/Anomalies.* Given two types of anomalies: (1) anomalies are defined to be **datapoints** in low density regions, (2) anomalies are **regions** of low density.

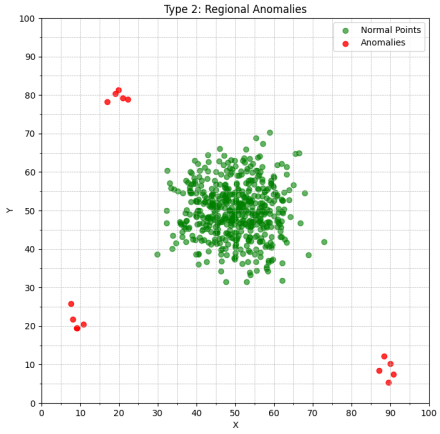
   (a) For both anomalies please create/draw a dataset of 2 features (x and y axis), with 3 anomalies and many normal data points (the normal datapoints should be marked, e.g., green colour)

   (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 Local Outlier Factor: It's good at finding single points that stick out. It works by looking at how crowded a point's neighborhood is compared to how crowded its friends' neighborhoods are. This is useful for dataset 1 because it doesn't care if the data groups are perfectly round, so even though some points are slightly outside crowded regions, only points that are really far away (the red points) get marked as anomalies. Preprocessing: Normalize the x and y features to [0,1]

Dataset 2 DBSCAN ((Density-Based Spatial Clustering of Applications with Noise): This identifies regions of low density by labeling points that do not belong to dense clusters as noise, which aligns with the Type 2 dataset's sparse regions. DBSCAN is robust to varying cluster shapes and does not require specifying the number of clusters. Preprocessing: Standardize the x and y features to have zero mean and spread of 1.

Answer (c) - Describe the main assumptions

Algorithm	**Assumption**
Local Outlier Factor	Assumes that normal points are located in regions of relatively high density, while anomalies are in low-density regions with fewer neighbors. Also assumes that the Euclidean distance is a valid measure of similarity between points.
DBSCAN	Assumes that the dataset has a clear distinction between dense clusters and sparse regions, with anomalies not forming significant clusters themselves

4. *Missing Values.* The dataset "task4-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
Age	Semester	Strong dependency, both linear and non-linear relationship
Height	Gender	Moderate dependency, non-linear relationship

Answer (b) - Describe the reason for missingness

Variable	Reason
height	MAR, as missingness strongly depends on gender (1 gender skips giving height info significantly more than the other gender). The weak correlation with age (-0.102) suggests younger students might be slightly more likely to skip this question
likes pinapple on pizza	MCAR, not important questions so student likely skips this. Data shows missingness has no strong correlation with other feature
likes chocolate	MCAR, similar reason with pizza, missingness seems random
english skills	MAR, as missingness depends on study (some study has twice more "missingness" in english skills collumn than some other study)

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

Variable	Strategy
Height	Regression Imputation: predict missing height values based on gender and age. This approach is suitable for MAR data
likes pinapple on pizza	Mode Imputation: Impute missing values with the most frequent value (0 or 1) in the likes-pinapple-on-pizza column. This minimizes distortion of the distribution, compared to assigning values randomly
likes chocolate	Mode Imputation: same as for pinapple on pizza above
english skills	Regression Imputation: Predict missing english-skills values using a regression model with study

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

Variable	Before Strategy	After Strategy
Height	172.1614	172.7154
Likes-pinapple-on-pizza	0.3607	0.2720
Likes-chocolate	0.7910	0.8610
english-skills	87.0097	87.0024