High-dimensional data analysis

In this notebook, we will explore high-dimensional data analysis techniques and learn how to uncover patterns and insights from complex datasets.

Learning objectives

By the end of this train, you should be able to:

- Understand the fundamental concepts and challenges of high-dimensional data analysis.
- Learn various techniques for visualising and simplifying high-dimensional data.

High-Dimensional Data Analysis

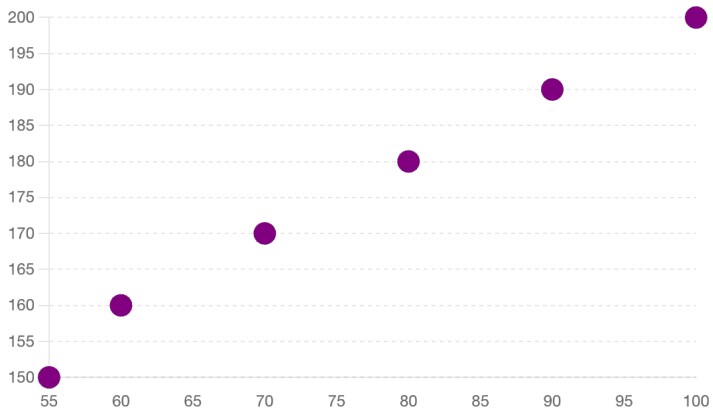
High-dimensional data analysis involves exploring and understanding datasets with a**large number of variables or features**. Traditional analysis techniques often struggle with such data due to the "curse of dimensionality," which refers to various challenges that arise when working with high-dimensional spaces. This complexity makes it **challenging** to visualise, interpret, and extract meaningful insights from the data.

Understanding Dimensionality

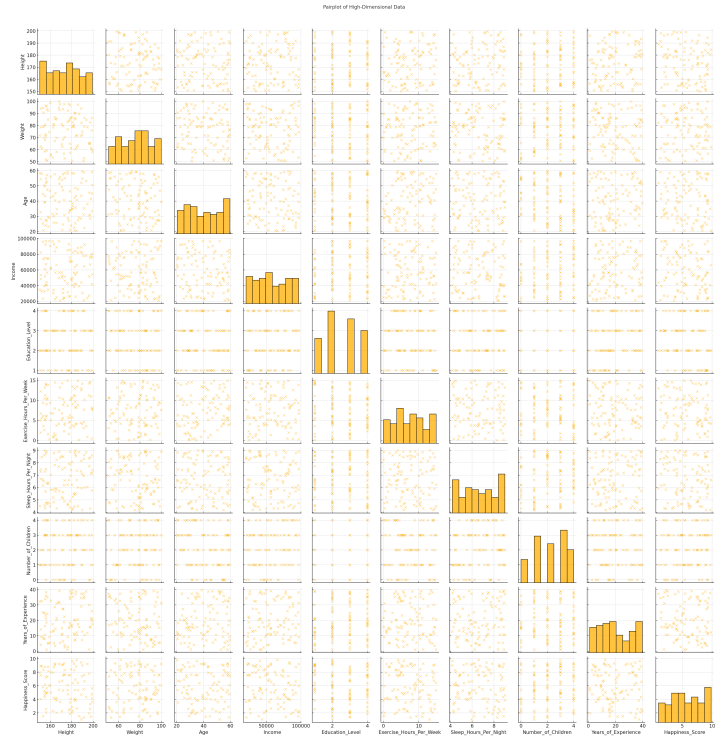
To grasp why specialised techniques are necessary for high-dimensional data, let's first understand how we visualise and analyse data with different numbers of dimensions:

Consider a dataset with a**single variable**, such as the height of individuals. Each data point represents the height of one person, making it a one-dimensional dataset. Visualising this data is **straightforward**, as we only need a single axis.

Now, imagine we have**two variables**: height and weight. Each data point now represents a pair of measurements (height and weight) for one individual. This two-dimensional dataset can be plotted on a 2D graph with height on one axis and weight on the other. Visualising and finding relationships between the two variables is still manageable.



However, as the **number of variables increases**, the **complexity of the dataset** also **increases**. For example, consider a dataset where we have ten variables such as height, weight, age, income, education level, and so on. This ten-dimensional dataset cannot be easily visualised or analysed using traditional methods.



This is already becoming too complex to visualise. Now, imagine we had millions of rows of data with 1000s of features. Advanced techniques are necessary to uncover patterns and insights from this complex data, making high-dimensional data analysis crucial.

Key Techniques in High-Dimensional Data Analysis

To analyse high-dimensional data effectively, several advanced techniques are used. Let's explore some of these key techniques:

Mathematical Distance

Mathematical distance measures the **similarity or dissimilarity between data points** in high-dimensional space. It is crucial for clustering, classification, and various distance-based methods. Common distance metrics include Euclidean, Manhattan, and cosine distances.

Steps:

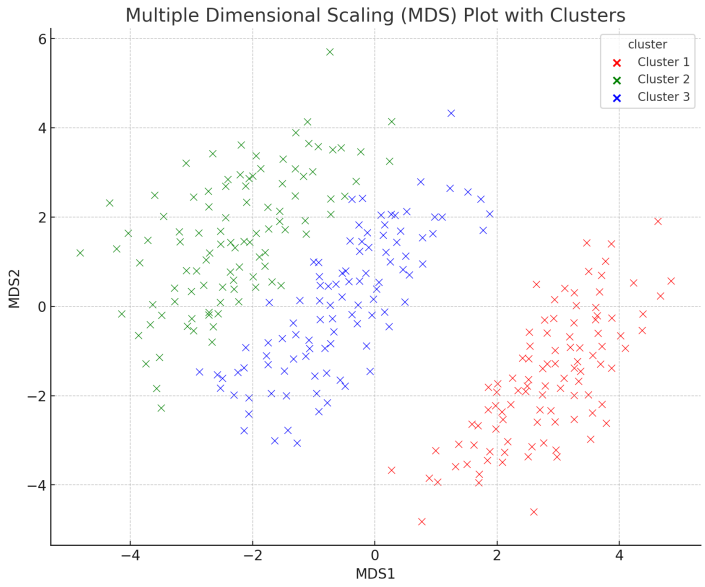
1. **Choose a Distance Metric:** Select a metric like Euclidean or Manhattan distance.
2. **Calculate Distances:** Compute distances between data points.
3. **Analyse Distances:** Group similar profiles to identify patterns.

Multiple Dimensional Scaling (MDS) Plots

MDS visualises high-dimensional data in lower dimensions, **preserving the distances between data points**. This method helps visualise high-dimensional data in a lower-dimensional space, making it easier to identify patterns and clusters.

Steps:

1. **Compute Distance Matrix:** Calculate distances between all pairs of samples.
2. **Apply MDS:** Create a 2D or 3D plot representing the distances.
3. **Interpret Clusters:** Identify clusters of similar profiles.



The axes (MDS1 and MDS2) are dimensions created to best represent the distances in the original high-dimensional space. They don't correspond directly to any of the original variables but rather to combinations of them; this makes the plot much easier to read than the 10-dimensional plot above. Furthermore, points that are closer together in the plot represent individuals whose profiles (across the ten variables) are more similar to each other. Whereas points that are further apart represent individuals with more dissimilar profiles.

Factor Analysis

Factor analysis identifies underlying relationships between observed variables by **grouping them into factors**. This technique reduces the number of variables while retaining most of the information, allowing for a more simplified understanding of complex datasets.

Steps:

1. **Identify Variables:** Select variables for analysis.
2. **Extract Factors:** Use factor analysis to find latent variables (variables that are not directly observed or measured but are inferred from other variables that are observed and measured).
3. **Interpret Factors:** Analyse factors to understand relationships.

Dealing with Batch Effects

Batch effects refer to **non-biological variations in data** that arise from differences in sample processing times, lab conditions, and other factors. Correcting for batch effects is essential to ensure the accuracy of data analysis, typically achieved through normalisation techniques.

Steps:

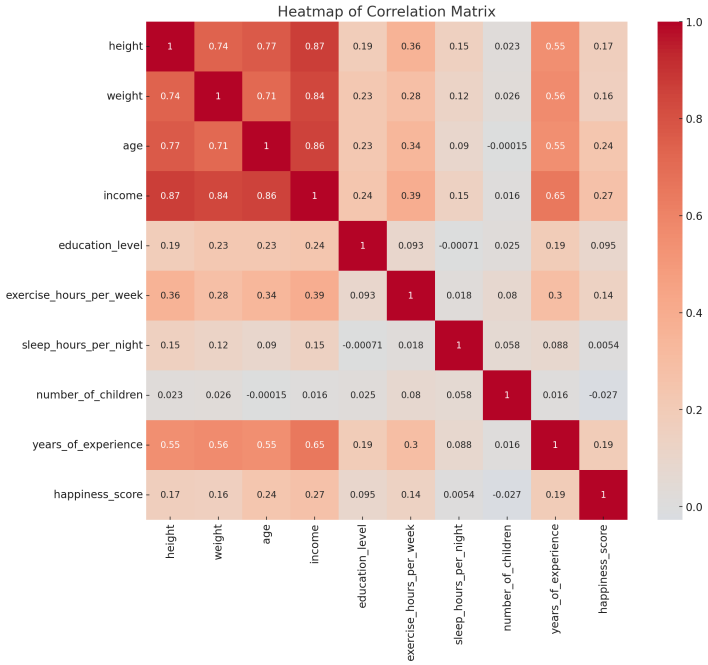
1. **Identify Batch Variables:** Determine which variables represent different batches.
2. **Normalise Data:** Adjust for batch effects using statistical techniques.
3. **Validate Correction:** Ensure the correction preserves significant signals.

Heatmaps

Heatmaps use **colours to represent individual values** in a dataset, making them an effective tool for visualising high-dimensional data. By displaying data in a matrix format, heatmaps help identify patterns, relationships, and clusters within the data.

Steps:

1. **Prepare Data:** Select data for the heatmap.
2. **Generate Heatmap:** Create a heatmap with variables as rows and samples as columns.
3. **Interpret Patterns:** Identify patterns of data, such as clusters of similar values.



The heatmap provides a visual summary of the relationships between the variables in the dataset, by displaying the correlation coefficients between each pair of the ten variables.

- **Positive Correlations:** Variables with positive correlation coefficients (closer to 1) indicate that as one variable increases, the other tends to increase as well. For example, income and years of experience might show a positive correlation.
- **Negative Correlations:** Variables with negative correlation coefficients (closer to -1) indicate that as one variable increases, the other tends to decrease. For example, age and sleep hours per night might show a negative correlation.
- **Weak/No Correlations:** Variables with correlation coefficients close to 0 indicate weak or no linear relationship between them.

Mathematical Tools and Methods

High-dimensional data analysis relies on various mathematical tools and methods, including linear algebra, statistics, and optimisation techniques. Understanding these tools is essential for implementing the techniques mentioned above effectively.

- **Linear Algebra:** Eigenvalues, eigenvectors, and matrix operations are fundamental in factor analysis and MDS.
- **Statistics:** Measures of central tendency, variability, and statistical tests are crucial for analysing data.
- **Optimisation:** Techniques such as gradient descent are used to minimise errors in various models.

Example: Identifying Genetic Markers Associated with Diseases

To illustrate the application of high-dimensional data analysis techniques, let's consider an example in the field of biomedical research. Imagine we have a high-dimensional dataset containing gene expression levels for thousands of genes across multiple patients. Some patients have a specific disease, while others do not. Our goal is to identify genetic markers associated with the disease.

Steps:

1. **Data Collection:** Gather gene expression data for thousands of genes from patients with and without the disease.
2. **Distance Calculation:** Use a suitable distance metric, such as Euclidean or Manhattan distance, to calculate the similarity between the gene expression profiles of different patients.
3. **Apply MDS:** Visualise the high-dimensional gene expression data in 2D or 3D using Multiple Dimensional Scaling (MDS) to identify natural groupings of patients based on their gene expression profiles.
4. **Factor Analysis:** Reduce the number of gene expression variables to simplify the dataset while retaining key information by identifying underlying factors that group genes with similar expression patterns.
5. **Normalisation:** Adjust the gene expression data to correct for any batch effects, such as differences in sample processing times or lab conditions, ensuring the accuracy of the analysis.
6. **Heatmap Creation:** Generate heatmaps to visualise patterns and clusters within the gene expression data, making it easier to identify groups of genes that are upregulated or downregulated in patients with the disease.

By applying these high-dimensional data analysis techniques, we can uncover meaningful patterns in the gene expression data and identify genetic markers that are strongly associated with the disease, providing valuable insights for further biomedical research and potential therapeutic targets.