# Data Science Survival Skills

Exercise 5 - Data visualization and statistics

# Agenda

- PCA
- t-SNE
- UMAP
- Statistics

# Problem with high-dimensional data

- **Curse of Dimensionality:**
  - Increase number of features (dimensions) ➔ volume of feature space grows exponentially
  - Leads to a sparsity problem ➔ available data points become more scattered and sparse in higher dimensions
  - Sparsity makes it challenging to generalize patterns and relationships from the data
- **Increased computational complexity:**
  - Analyzing and processing high-dimensional data require more computational resources and time
  - Complexity makes ML tasks computational expensive
- **Diminishing returns in performance:**
  - Adding more features does not always lead to better model performance
  - At some point, the inclusion of additional features may not provide significant improvements in model performance
- Difficulty in visualization
- Issues in model interpretability (which individual feature contributed the most?)

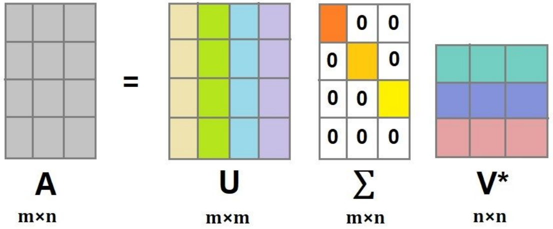
# PCA - Principal Component Analysis

- Statistical method that simplifies the complexity in high-dimensional data while retaining trends and patterns
- PCA transforms data into a new coordinate system → the axes represent the principal components
- PCA helps overcome the challenge of high-dimensional data by capturing the most significant variation in the data

# PCA

- **Principal components:**
  - Principal Components are linear combinations of the original features in a dataset
  - The first principal component (PC1) captures the most significant variance in the data, the second (PC2) captures the second most, and so on
  - **Mathematically**: PC1 is the linear combination of the original features that maximizes the variance. Subsequent principal components are orthogonal to the previous ones and capture the remaining variance
- **Covariance matrix:**
  - Square matrix that summarizes the covariances between different features in a dataset
  - Used to obtain **eigenvectors** and **eigenvalues**
- **Eigenvectors:**
  - Directions along which the data varies the most. Each eigenvector corresponds to a principal component
- **Eigenvalues:**
  - Magnitude of the variance in the data along the corresponding eigenvector. Higher eigenvalues mean more significant variance

# PCA



- Alternative to eigendecomposition: **SVD**
- SVD is generally more numerically stable
- SVD can be applied to rectangular matrices, whereas eigendecomposition is defined only for square matrices (PCA: data matrix is often rectangular → more features than samples)
- SVD more efficient than eigendecomposition
- Singular values in Σ in SVD are related to the eigenvalues in eigendecomposition (square of the singular values is equal to the eigenvalues of the covariance matrix. The left and right singular vectors in SVD correspond to the eigenvectors in PCA)

# PCA

- **Applications:**
  - Image compression
  - Feature selection
  - Noise reduction
- **Limitations:**
  - Linearity Assumption: PCA assumes linear relationships between features
  - Sensitive to Outliers: Outliers can significantly impact the results
  - Selecting the Number of Components: Trade-off between dimensionality reduction and retaining enough information

# t-SNE - t-Distributed Stochastic Neighbor Embedding

- Nonlinear dimensionality reduction technique
- Used in machine learning for exploratory data analysis and pattern recognition
- Map high-dimensional data points to a low-dimensional space (usually 2D or 3D) in such a way that similar instances in the high-dimensional space are modeled as nearby points in the low-dimensional space, while dissimilar instances are modeled as distant points

# t-SNE

- Defines two probability distributions over pairs of high-dimensional data points
- One distribution represents pairwise similarities in the input space, and the other represents pairwise similarities in the low-dimensional space
- For each pair of data points x_i and x_j in the high-dimensional space, t-SNE defines a conditional probability p_j_i that represents the similarity of x_i to x_j → using a Gaussian distribution based on the Euclidean distance between data points
- Similar in the low-dimensional space → using a Student’s t-distribution
- Minimizes the Kullback-Leibler (KL) divergence (gradient descent)

# Statistic

- Important terms:
  - Mean
  - Standard deviation:
    - Measure of the spread or dispersion of a set of values from their mean
  - Variance:
    - Square of the standard deviation (average of the squared differences from the mean)
  - Standard error of the mean:
    - Estimate of how much the sample mean is expected to vary from the true population mean
  - Confidence interval:
    - Range of values that likely contains the true population parameter
  - t-statistic:
    - Measure for comparing means (accounts for variability and sample size)
  - p-value:
    - Probability of obtaining a result as extreme as, or more extreme than, the observed result, assuming the null hypothesis is true

# Statistic

- QQ-Plot:
  - Used for distributional diagnostics
  - Assess the goodness of fit between observed data and a theoretical distribution
  - Theoretical quantiles are calculated based on a chosen distribution
  - Perfect Fit: Points lie along a straight line, indicating close conformity to the theoretical distribution
  - S-Shaped Curve: Deviations upwards or downwards suggest heavier tails or skewness
  - Outliers: Points significantly deviating from the straight line may indicate outliers
  - Use cases:
    - Checking normality assumptions for statistical tests
    - Assessing the fit of data to a specific distribution

# Statistical Testing

- **Parametric tests:**
  - T-test, ANOVA, Pearson correlation
  - Assumptions: Normality, homogeneity of variances
- **Non-parametric:**
  - Analyze data when parametric assumptions are not met or for ordinal/nominal data
  - Mann-Whitney U, Wilcoxon signed-rank, Kruskal-Wallis, Spearman correlation
  - Robustness to non-normality, suitability for small sample sizes
- **Hypothesis testing:**
  - Evaluate whether observed differences are statistically significant.
  - Formulate null and alternative hypotheses, choose a significance level, conduct the test, interpret results