Preparation for a Clustering Project:

Vectorization:

- Bag-of-Words (BoW): Convert text data into a matrix of word frequencies. Each row corresponds to a document and each column corresponds to a unique word
- TF-IDF (Term Frequency-Inverse Document Frequency): Weight words based on their importance in a document and across a corpus. Helps to identify key terms that distinguish documents from each other
- Word Embeddings: Use techniques like Word2Vec or GloVe to represent words as dense vectors in a continuous space. Captures semantic relationships between words and enables more nuanced analysis of text data.
- Image Feature Extraction: Extract features from images. Techniques like Histogram of Oriented Gradients (HOG) or Convolutional Neural Networks (CNNs) extract meaningful features from images, which can then be used for tasks like object detection, image classification, or image retrieval. HOG focuses on gradients and edge orientations, while CNNs leverage deep learning to automatically learn hierarchical representations of visual features.

Normalization: used to preprocess data and bring it into a standardized format, which is often necessary for effective machine learning models

- Min-Max Scaling: Scale features to a range, often between 0 and 1. Preserves the relationships between data points but sensitive to outliers
- Standardization (Z-score normalization): Transform features to have a mean of 0 and a standard deviation of 1, making data more amenable to algorithms sensitive to varying scales.
- Robust Scaling: Scale features using statistics robust to outliers.
- Unit Vector Scaling: Scale feature vectors to have unit norm, useful for cosine similarity in high-dimensional spaces.

Similarity or Distance Metrics: fundamental in various machine learning tasks, such as clustering, classification, and recommendation systems

- Euclidean Distance: Measure straight-line distance between points in a multidimensional space, often used in k-nearest neighbors algorithms.
- Cosine Similarity: Measure the cosine of the angle between two vectors, often used for text or high-dimensional data.
- Jaccard Similarity: Measure similarity between sets, useful for binary data like presence/absence of features.

- Manhattan Distance: Measure distance by summing the absolute differences between coordinates along each dimension, suitable for grid-like structures or city-block-style distances.

Clustering Algorithms: used to group similar data points together based on certain criteria, facilitating tasks like customer segmentation, anomaly detection, and recommendation systems

- K-Means: Partition data into K clusters by minimizing within-cluster variance, often used in customer segmentation to identify groups with similar behaviors or preferences
- Hierarchical Clustering: Build a hierarchy of clusters by merging or splitting them recursively.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Identify clusters based on density and separate noise points.
- Gaussian Mixture Models (GMM): Model data as a mixture of Gaussian distributions, accommodating complex data distributions and enabling probabilistic cluster assignments.

Evaluation Metrics or Methods: employed to assess the performance and quality of clustering algorithms

- Silhouette Score: Measure cluster cohesion and separation, providing an overall assessment of cluster quality.
- Davies-Bouldin Index: Evaluate clustering quality based on intra-cluster and inter-cluster distances, helping to identify well-separated and compact clusters.
- Purity: Measure the agreement between true class labels and cluster assignments, assessing the accuracy of clustering results.
- Adjusted Rand Index (ARI): Measure similarity between true and predicted clusters, adjusted for chance.

Applications of Clustering:

- Customer Segmentation: Identify groups of customers with similar behaviors or preferences, enabling targeted marketing strategies.
- Anomaly Detection: Detectsoutliers or unusual patterns in data, essential for fraud detection or system monitoring.
- Image Segmentation: Segment images into meaningful regions, facilitating tasks like object recognition or medical image analysis
- Recommendation Systems: Cluster items or users to provide personalized recommendations and enhance user experience.

Stories:

Story 1: Market Segmentation for a Tech Startup

> (prompt: We are preparing for a text/data clustering project and are researching methods and how to apply them. Write a quick (paragraph or so) fiction story on how you would use clustering to solve a real-life challenging problem.)

In a bustling city, a group of tech enthusiasts founded a startup that revolutionized how people interact with urban spaces. To understand their diverse user base better, they embarked on a clustering project. By vectorizing user interactions and normalizing data, they applied k-means clustering, revealing distinct user segments: the early adopters craving cutting-edge features, the pragmatists valuing utility, and the social connectors seeking community engagement. Armed with these insights, the startup tailored their marketing strategies, product updates, and community events, fostering a vibrant ecosystem around their innovative platform.

Story 2: Website Improvement

The owners of a small company want to improve their website to tailor to their users. To understand various aspects of how the users use the site they decide to utilize clustering to reveal helpful information. After collecting user interaction data such as page views, time spent per page, clicks, and user demographics, they normalize and vectorize the data so clustering algorithms can be applied. The algorithms aim to reveal frequent/occasional/bounce users and the differences between them. From there, they aim to implement new enhancements to improve the overall user experience. For frequent users they might choose to add personalized recommendations. Occasional users may have their key sections of the website highlighted for ease of access. For bounce users the site homepage may be redesigned to be more engaging and easy to navigate. The owners intend on continuously monitoring these metrics and improving their website as more feedback is generated.