Machine Learning

| Unsupervised Learning

Prof. Dr. Mostafa Elhosseini

Professor of Smart Systems Engineering

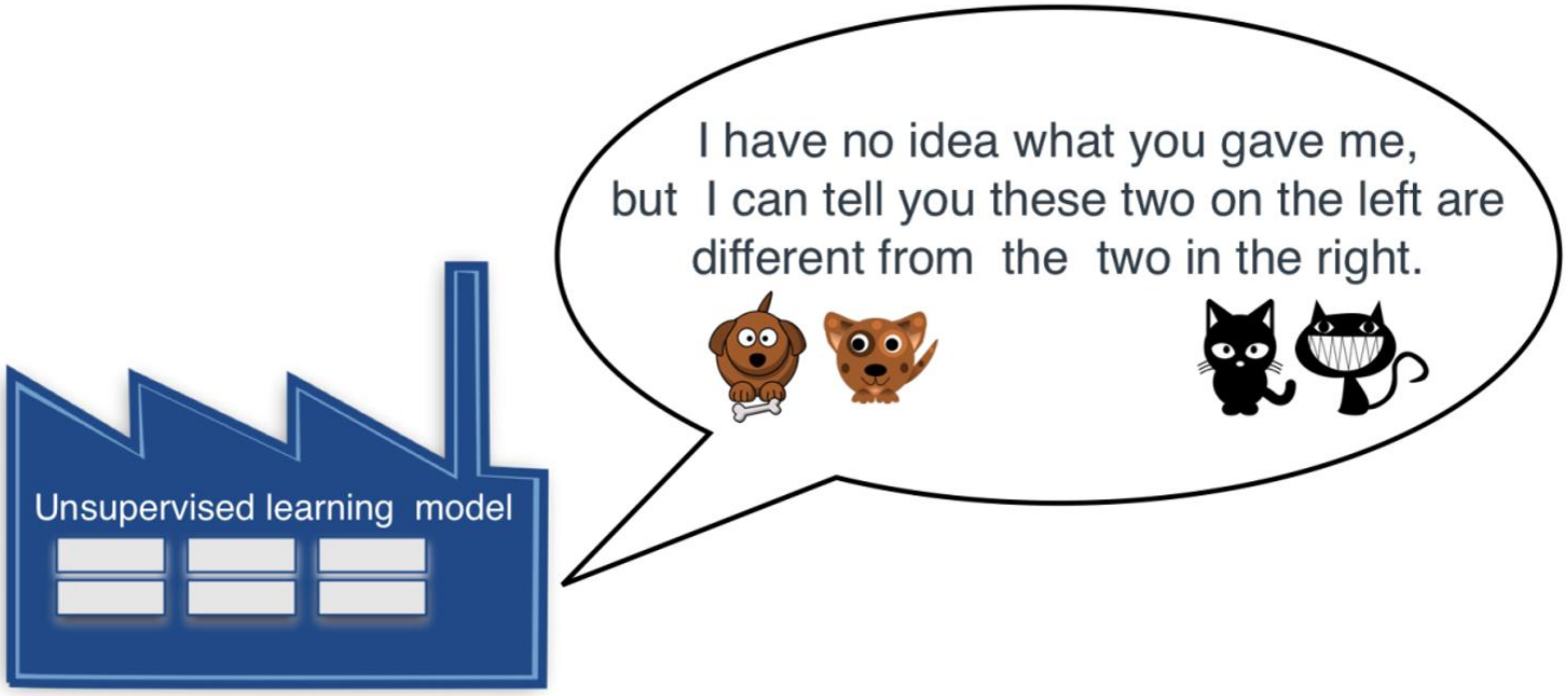
https://youtube.com/drmelhosseini

Types of ML

Machine learning systems can be classified into several broad categories based on specific criteria:

- Supervision during Training: This includes:
  - **Supervised Learning**: Where models are trained with labeled data.
  - **Unsupervised Learning**: Where models work with unlabeled data.
  - **Semi-Supervised Learning**: A mix of labeled and unlabeled data is used.
  - **Self-Supervised Learning**: Where the model generates its own labels from the input data.
  - **Reinforcement learning**: where the model learns through rewards and penalties.

Unsupervised Learning



Unsupervised Learning

- The training data is unlabeled.
- The system tries to learn without a teacher.
- The algorithm must find structure and patterns in the data on its own, without any guidance on what to look for.

Unsupervised Learning

Key aspects of unsupervised learning include

- Clustering
- Dimensionality Reduction
- Association Rule Learning
- Anomaly Detection
- Autoencoders and Generative Models

Clustering

- Consider a **popular blog** that receives traffic from various types of visitors interested in different topics.
  - To segment these visitors into distinct groups based on their behavior and interests to tailor the content better, recommend articles, or for targeted advertising
- Imagine a **magazine** that covers a wide range of topics such as technology, health, finance, and travel. The magazine has an archive of thousands of articles.
  - To organize these articles into distinct groups so that similar articles are placed together. This would help in better content management and enhance the reader's experience by easily finding related articles.

Unsupervised Learning

Email	Size	Recipients
1	8	1
2	12	1
3	43	1
4	10	2
5	40	2
6	25	5
7	23	6
8	28	6
9	26	7

Unsupervised Learning

Email	Size	Recipients
1	8	1
2	12	1
3	43	1
4	10	2
5	40	2
6	25	5
7	23	6
8	28	6
9	26	7

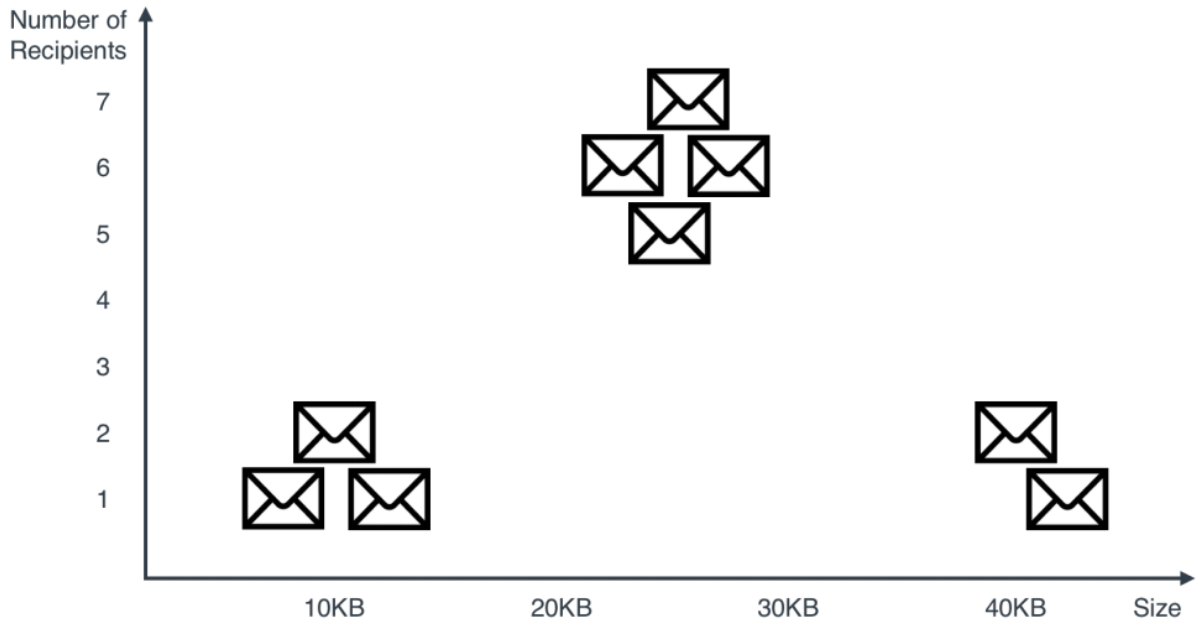
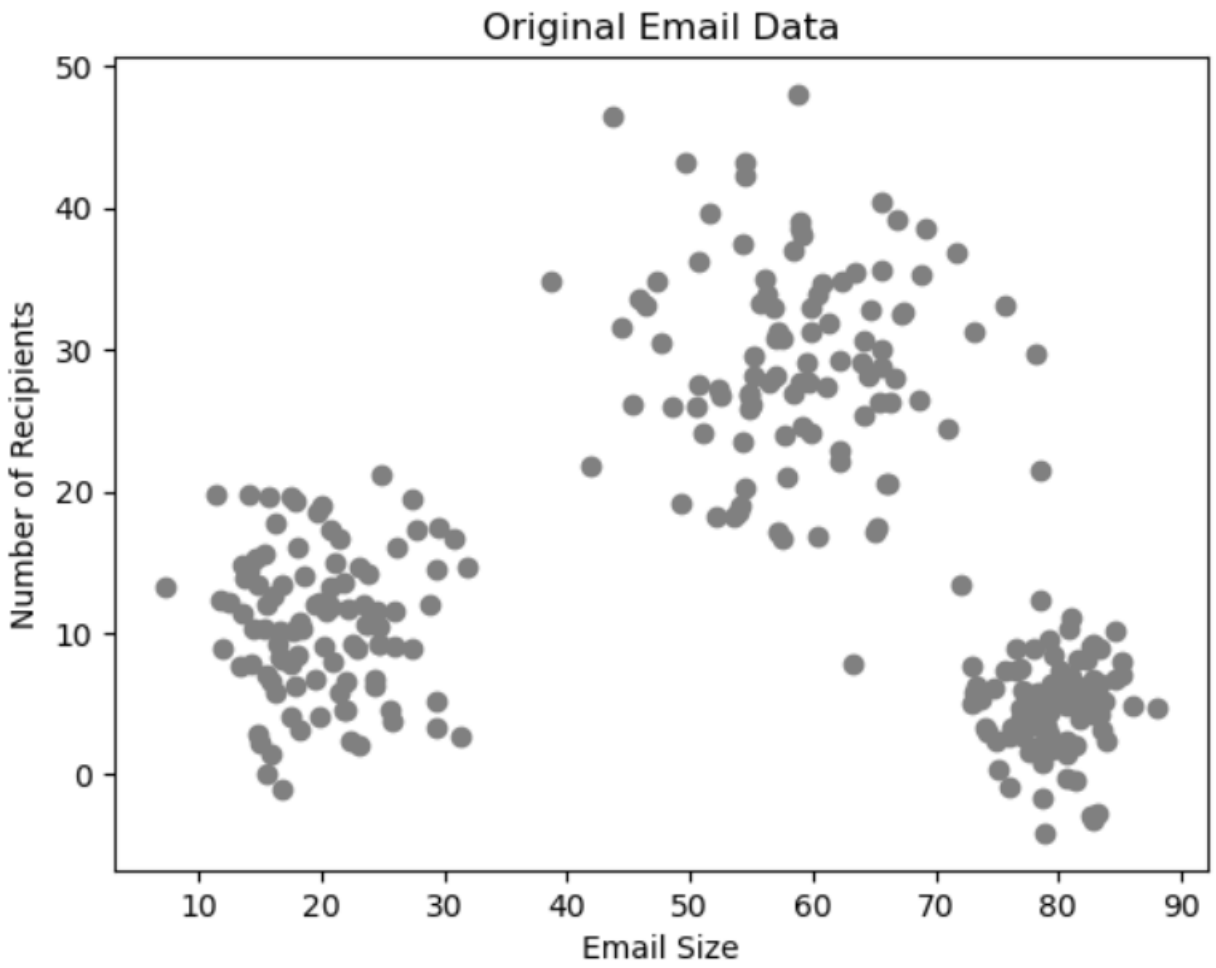
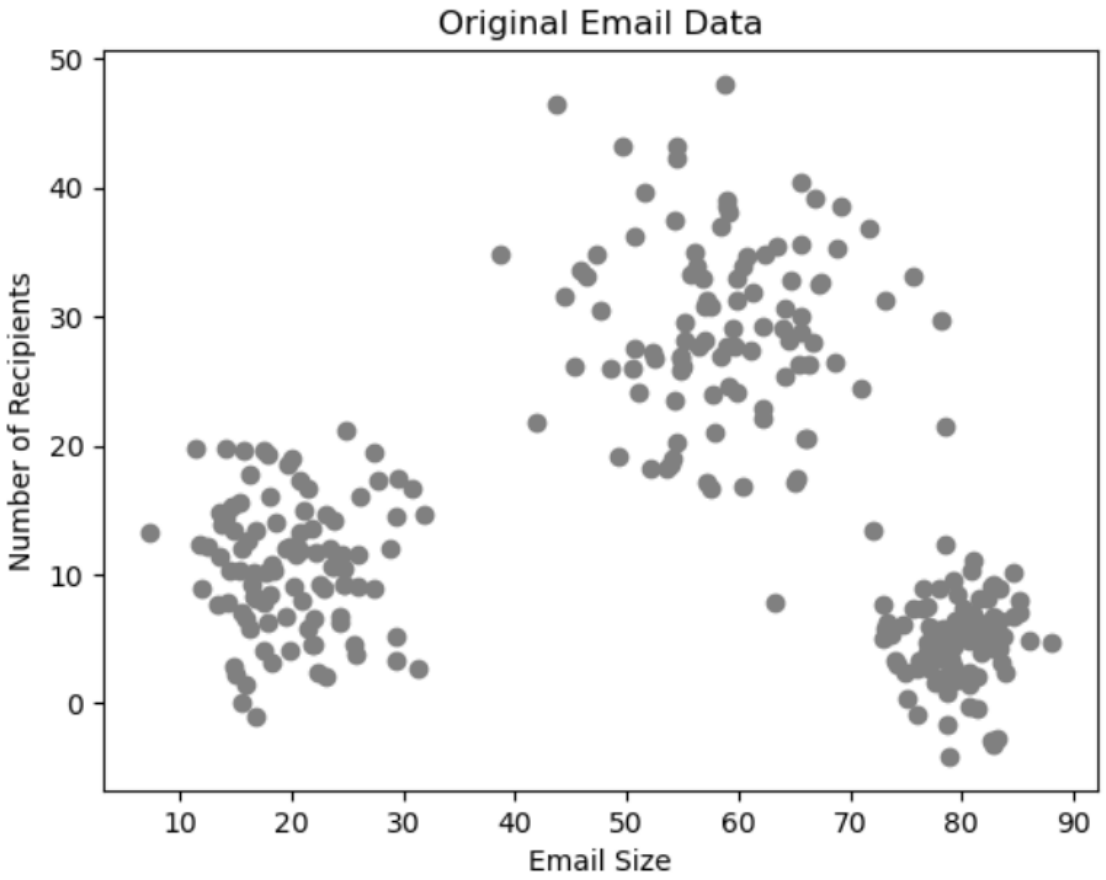


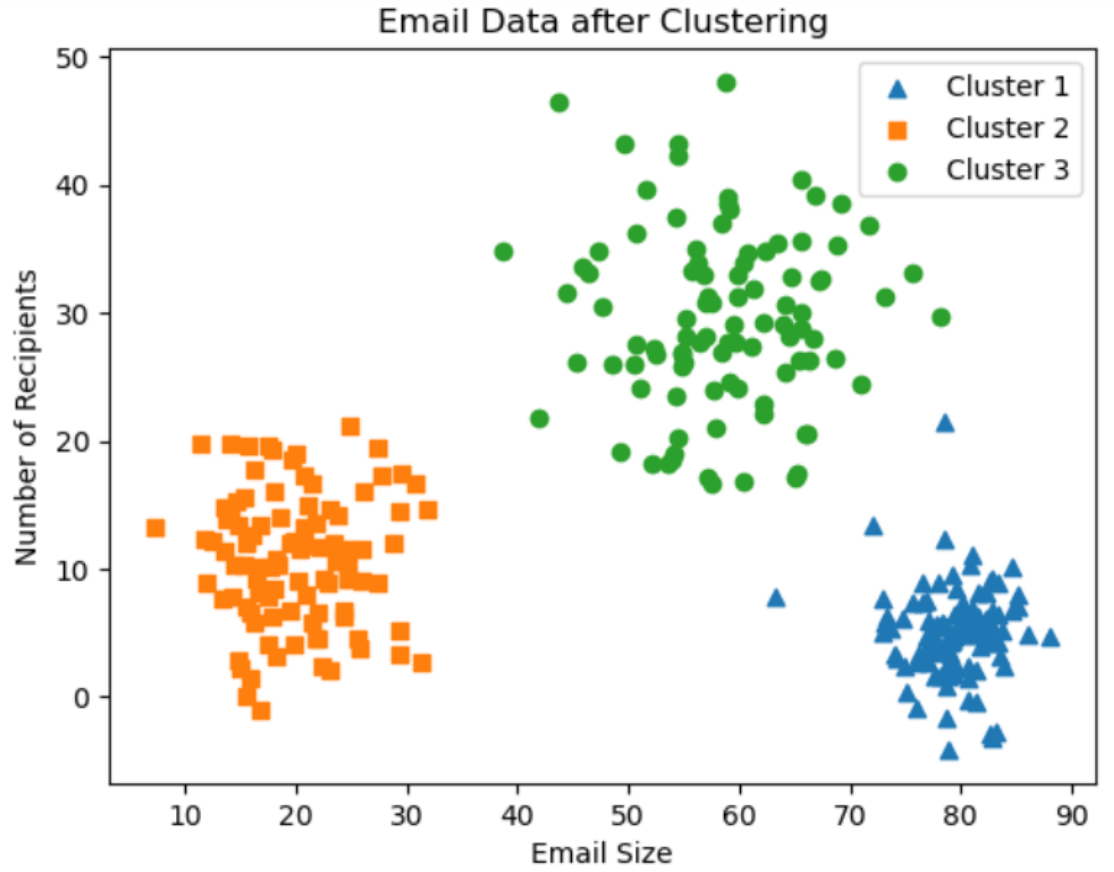
Figure 2.6. A plot of the emails with size on the horizontal axis and number of recipients on the vertical axis. Eyeballing it, it is obvious that there are three distinct types of emails.

Unsupervised Learning



Unsupervised Learning





Dimensionality Reduction

- Simplify the data without losing too much information
- Reduce the number of variables under consideration and can be divided into feature selection and feature extraction.
- Techniques like Principal Component Analysis (PCA) are used to reduce the complexity of data and make it easier to analyze.
- When dealing with high-dimensional data (i.e., data with many features or variables), it can be challenging to analyze and visualize the data, and it may also lead to issues like overfitting in machine learning models.

Dimensionality Reduction

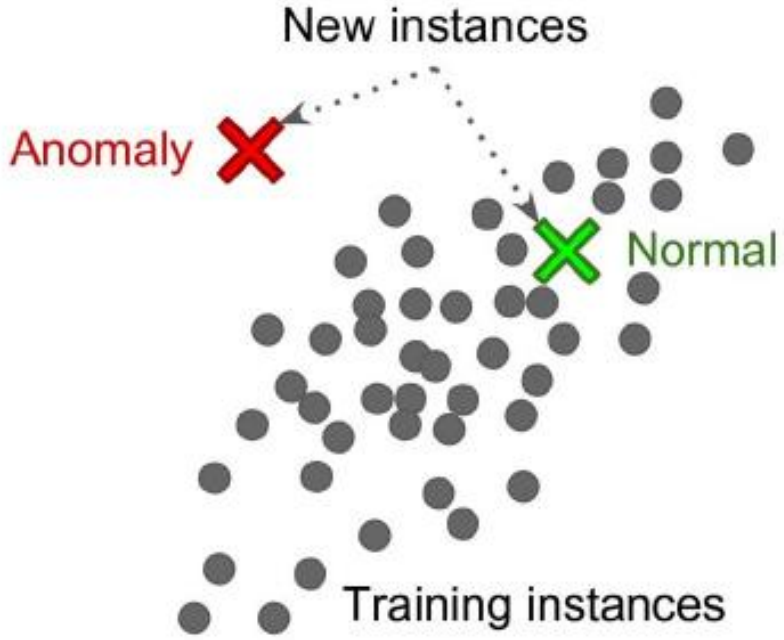
- Dimensionality reduction helps to simplify the dataset while retaining as much of the significant information as possible.
- There are two main types of dimensionality reduction:
  - **Feature Selection**: This involves selecting a subset of the most important features (variables) from the original dataset.
  - **Feature Extraction**: This involves transforming the original data into a new set of features. This new feature set should capture most of the important information in a smaller number of features

Dimensionality Reduction

- It is often a good idea to try to reduce the dimension of your training data
  - It will run much faster,
  - The data will take up less disk and memory space, and in some cases
  - It may also perform better

Anomaly Detection

- Anomaly detection, also known as outlier detection, is a process in machine learning and statistics used to identify unusual patterns or observations in data that do not conform to a well-defined notion of normal behavior.
- These anomalies or outliers can be indicative of critical incidents, such as a problem with the system, fraud, or errors.



Anomaly Detection

- The importance of anomaly detection varies across different domains:
  - Finance: Identifying fraudulent transactions.
  - Cybersecurity: Detecting intrusions and security breaches.
  - Healthcare: Monitoring patient vitals and identifying unusual readings that could indicate a medical issue.
  - Industrial: Monitoring machinery and detecting faults or failures.
  - Internet of Things (IoT): Detecting abnormal behavior in connected devices.

Novelty Detection

- It's widely used in areas like fraud detection in banking, intrusion detection in network security, system health monitoring, and fault detection in manufacturing processes.
- Novelty detection aims to discover new patterns for adaptation and learning, while anomaly detection is about identifying outliers that may indicate problems.
- In novelty detection, the new data is not inherently bad and might represent a valid new pattern, whereas, in anomaly detection, the anomalies are often indicative of a problem or an error.

Association Rule Learning

- To discover interesting relations between variables in large databases.
- It's a rule-based machine learning technique used to find patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories.

Association Rule Learning - Example

Imagine a small dataset of transactions recorded by a supermarket. Each transaction lists items purchased by a customer:

- Transaction 1: Bread, Milk
- Transaction 2: Bread, Diapers, Juice, Eggs
- Transaction 3: Milk, Diapers, Juice, Cola
- Transaction 4: Bread, Milk, Diapers, Juice
- Transaction 5: Bread, Milk, Diapers, Cola.

- Transaction 1: Bread, Milk
- Transaction 2: Bread, Diapers, Juice, Eggs
- Transaction 3: Milk, Diapers, Juice, Cola
- Transaction 4: Bread, Milk, Diapers, Juice
- Transaction 5: Bread, Milk, Diapers, Cola.

Example

Discovering Association Rules:

- **Identifying Itemsets**: First, we identify frequent itemsets (sets of items that appear frequently together). For example, {**Bread, Milk**} appears in three out of the five transactions.
- **Calculating Support**: The support for {Bread, Milk} is calculated as the number of transactions containing both items divided by the total number of transactions. So, Support({Bread, Milk}) = 3/5 = 60%.
- **Generating Rules**: From the frequent itemsets, we can generate rules. For instance, one rule could be Bread ⇒ Milk, suggesting that when customers buy bread, they also tend to buy milk.

Example

- Transaction 1: Bread, Milk
- Transaction 2: Bread, Diapers, Juice, Eggs
- Transaction 3: Milk, Diapers, Juice, Cola
- Transaction 4: Bread, Milk, Diapers, Juice
- Transaction 5: Bread, Milk, Diapers, Cola.

Discovering Association Rules:

- **Identifying Itemsets**: {Bread, Milk}
- **Calculating Support**: 3/5 = 60%.
- **Generating Rules**: Bread ⇒ Milk
- **Calculating Confidence**: The confidence for the rule Bread ⇒ Milk is calculated as the number of transactions containing both Bread and Milk divided by the number of transactions containing Bread. So, Confidence(Bread ⇒ Milk) = Support({Bread, Milk}) / Support({Bread}) = 60% / (4/5) = 75%.

- Transaction 1: Bread, Milk
- Transaction 2: Bread, Diapers, Juice, Eggs
- Transaction 3: Milk, Diapers, Juice, Cola
- Transaction 4: Bread, Milk, Diapers, Juice
- Transaction 5: Bread, Milk, Diapers, Cola.

Example

Discovering Association Rules:

- **Identifying Itemsets**: {Bread, Milk}
- **Calculating Support**: 3/5 = 60%.
- **Generating Rules**: Bread ⇒ Milk
- **Calculating Confidence**: 75%.
- **Calculating Lift**: Lift measures the likelihood of Bread and Milk being bought together compared to being bought independently. Lift(Bread ⇒ Milk) = Confidence(Bread ⇒ Milk) / Support({Milk}) = 75% / (3/5) = 125%. A lift value greater than 1 indicates that Bread and Milk are more likely to be bought together than separately

Example

- Transaction 1: Bread, Milk
- Transaction 2: Bread, Diapers, Juice, Eggs
- Transaction 3: Milk, Diapers, Juice, Cola
- Transaction 4: Bread, Milk, Diapers, Juice
- Transaction 5: Bread, Milk, Diapers, Cola.

Discovering Association Rules:

- **Identifying Itemsets**: {Bread, Milk}
- **Calculating Support**: 3/5 = 60%.
- **Generating Rules**: Bread ⇒ Milk
- **Calculating Confidence**: 75%.
- **Calculating Lift**: 125%. A lift value greater than 1 indicates that Bread and Milk are more likely to be bought together than separately

Example

Applications in Supermarket

- **Product Placement**: Items with strong associations, like pasta and tomato sauce, could be placed closer together in store aisles to encourage joint purchases.
- **Promotional Bundling**: The supermarket might consider discounts or bundle offers for items like bread and milk or chicken and barbecue sauce, encouraging customers to buy them together.
- **Inventory Management**: Understanding these patterns can help in predicting demand for certain items based on the sale of associated items.