Intro to Bus Data Analyt Tools

Matt Warner

Chapter 1

Probability

1 Introduction

- Uncertainty is an ever-present fact of life for decision makers
- Much time and effort are spent trying to plan for and respond to uncertainty.
- **Probability** is the numerical measure of the likelihood that an event will occur.
- This measure of uncertainty is often communicated through a probability distribution.
  - Extermely helpful in providing additional information about an event.
  - Can be used to help a decision maker evaluate possible actions and determine best course of action.

2 Events and Probabilities

- A **radom experiment** is a process that generates well-defined outcomes.
- By specifying all possible outcomes, we identify the **sample space** for a random experiment; examples:
  - A coin toss.
  - Rolling a die.
- An **Event** is defined as a collection of outcomes.

Random Experiment	**Experimental Outcomes**
Toss a coin	Head, tail
Roll a die	1, 2, 3, 4, 5, 6
Conduct a sales call	Purchase, no purchase
Hold a particular share of stock for one year	Price of stock goes up, price of stock goes down, no change in stock price
Reduce price of product	Demand goes up, demand goes down, no change in demand

Table 1.1: Random Experiments and Their Outcomes

Example: California Power & Light Company (CP&L)

CP&L is starting a project designed to increase the generating capacity of one of its plants in southern California.
Analysis of similar construction projects indicates that the possibe completion times for the project are:
8, 9, 10, 11, and 12 months

Completion Time (months)	**No. of Past Projects Having This Completion Time**	**Probability of Outcome**
8	6	$6/40 = 0.15$
9	10	$10/40 = 0.25$
10	12	$12/40 = 0.30$
11	6	$6/40 = 0.15$
12	6	$6/40 = 0.15$
Total	**40**	**1.00**

Table 1.2: Project Completion Times and Probabilities

- The **probability of an event** is equal to the sum of Probabilities of outcomes for the event
- CP&L example: let C denote the event that the project is completed in 10 months or less, $C = \{8, 9, 10\}$
- The probability of event C, denoted $P(C)$, is given by

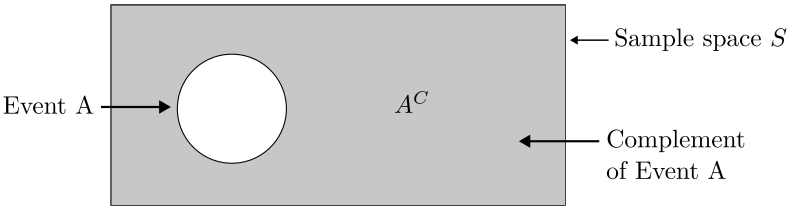
$$P(C) = P(8) + P(9) + P(10) + 0.15 + 0.25 + 0.30 = 0.70$$

- We can tell CP&L management that there is a 0.70 probability that the project will be completed in 10 months or less.

3 Some Basic Relationships of Probability

3.1 Completion of an Event:

- Given an event A, the **complement of A** is defined to be the event consisting of all outcomes that are *not* in a.
- The figure below shows what is known as a **Venn diagram**, which illustrates the concept of a complement:
  - Rectangular area represents the sample space for the random experiment and contains all possible outcomes.
  - Circle represents event A and contains only the outcomes that belong to A
  - Shaded region of the rectangle contains all outcomes not in event A



In any probability application, either event A or its complement A^C must occur.

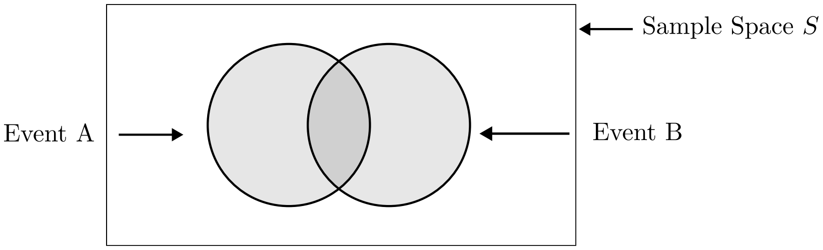
Solving for $P(A)$, we obtain the following result:

$$P(A) = 1 - P(A^C)$$

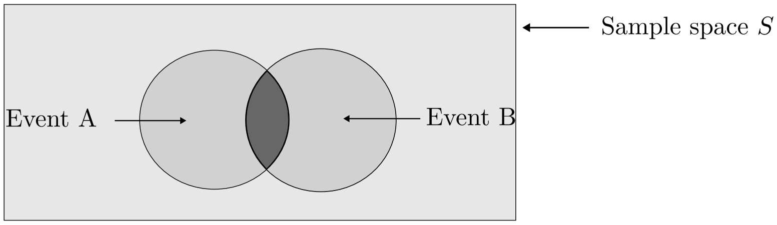
The probability of an event A can be computed easily if the probability of its complement is known.

3.2 Addition Law

- The addition law is helpful when we are interested in knowing the probability that at least one of two events will occur.
- Concepts related to the combination of events
  - The union of events
  - The intersection of events.
- Given two events A and B, the **union of A and B** is defined as the event containing all outcomes belonging to A or B or both.
- The union of A and B is denoted by AB
- The Venn diagram in the figure below depicts the union of A and B:
  - One circle contains all the outcomes of A
  - The other circle contains all the outcomes of B



- The definition of the **intersection of A and B** is the event containing the outcomes that belong to both A and B
- The intersection of A and B is denoted by A
- The Venn diagram below depicts the intersection of A and B
  - The area in which the two circles overlap is the intersection
  - It contains outcomes that are in both A and B

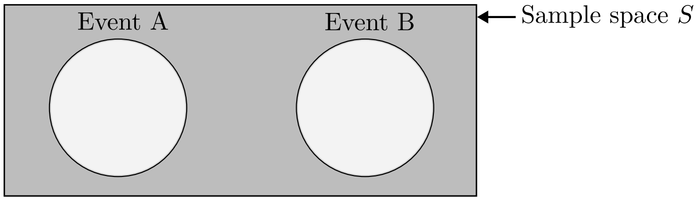


- The **addition law** provides a way to compute the probability that event A or event B or both will occur
- Used to compute the probability of the union of two events

ADDITION LAW:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- A special case arises for **mutually exclusive events:**
  - If the occurrences of one event precludes the occurence of the other.
  - If the events have no outcomes in common.



ADDITION LAW FOR MUTUALLY EXCLUSIVE EVENTS:

$$P(A \cup B) = P(A) + P(B)$$

4 Conditional Probability

Chapter 2

Descriptive Data Mining

1 Introduction

Data mining can be described as the process of discovering patterns, trends, insights, and useful information from large datasets. it involves the use of various techniques and algorithms to extract knowledge and valuable patterns from raw data

The increase in the use of data-mining techniques in business has been caused largely by three events:

- The explosion in the amount of data being produced and electronically tracked
- The ability to electronically warehouse these data
- The affordability of computer power to analyze the data

1.1 Unsupervised Learning

unsupervised learning is a category of machine learning where the algorithm is trained on a dataset **without** explicit supervision or **labeled outcomes**

Unlike **supervised learing**, where the algorithm learns to make predictions based on labeled examples, **unsupervised learning** aims to find patterns, relationships, and structures within the data witout any predefined target or output variable.

In short, the primary goal of unsupervised learning is to discover inherent structures within the data itself

Example 1: Clustering (Unsupervised Learning)

- Imagine you have a dataset containing customer purchasing behavior, but you don't have predefined categories or labels.
- In unsupervised learning, you can use clustering algorithms like k-means to group similar customers together based on their purchasing patterns
- The algorithm identifies natural clusters within the data without being told in advance what those clusters should be.

Example 2: Predicting Customer Churn (Supervised Learning)

- Suppose you have another dataset with customer information, including whether each customer churned or not (a binary label: churned or not churned)
- In supervised learning, you can train a classification algorithm (e.g., logistic regression) using this labeled data.
- The algorithm learns to predict whether new customers are likely to churn based on features such as usage patterns, customer support interactions, etc.
- Here, the algorithm relies on the labeled outcomes to make predictions.

2 Cluster Analysis

The goal of clustering is to segment observations into similar groups based on observed variables.

This can be employed during the data-preparation step to identify variables or observations that can be aggregated or removed from consideration.

Cluster analysis is commonly used in marketing to divide customers into different homogenous groups; known as **market segmentation**

Note:-

Also used to identify outliers

2.1 Clustering methods

- Bottom-up **hierarchical clustering** starts with each observation belonging to its own cluster and then sequentially merges the most similar cluster to create a series of nested clusters
- ***k*-means clustering** assigns each observation to one of k clusters in a manner such that the observations assigned to the same cluster are as similar as possible.

Both methods depend on how two observations are similar - hence, we have to measure similarity between observations.

2.2 Measuring similarity Between Observations

When observations include numeric variables, **Euclidean distance** is the most common method to measure dissimilarity between observations.

Let observations $u = (u_1, u_2, \ldots, u_q)$ and $v = (v_1, v_2, \ldots, v_q)$ each comprise measurements of q variables.

The Euclidean distance between observations u and v is:

$$d_{uv} = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2 + \ldots + (u_q - v_q)^2}$$

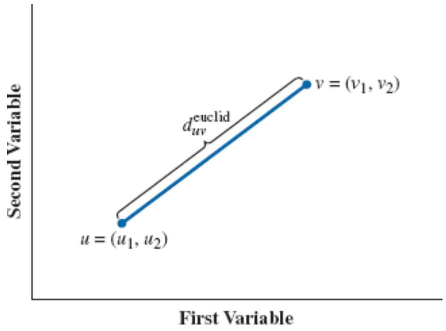
Illustration:

- KTC is a financial advising company that provides personalized financial advice to its clients.
- KTC would like to segment its customers into several groups (or clusters) so that the customers within a group are similar and dissimilar with respect to key characteristics
- For each customer, KTC has an observation of seven variables: Age, Female, Income, Married, Children, Car Loan, Mortgage.

The observation $u = (61, 0, 57881, 1, 2, 0, 0)$ corresponds to a 61-year old male with an annual income of $57,881, married with two children, but no car loan and no mortgage.

Note:-

Euclidean distance becomes smaller as a pair of observations becomes more similar with respect to their variable values.



- Euclidean distance is highly influenced by the scale on which variables are measured.
- We need to standardize the units of each variable j of each observation u.

Example: u_j, the value of variable j, in observation u, is replaced with its z-score z_j

- The conversion to z-score also makes it easier to identify outlier measurements, which can distort the Euclidean distance between observations.
- When clustering observations solely on the basic of categorical variables encoded as binary values (0, or 1) , a **better measure of similarity** between two observations can be achieved by **counting** the number of variables with matching values.
- Matching values indicates agreement between observations, while differing values represent disagreement.
- By counting matching values, you are effectively measuring how many categorical variables the two observations have in common.
- This count can serve as a similarity measure for clustering, with a higher count indicating greater similarity.
- The simplest overlap measure is called the **matching coefficient** and is computed as:

$$\frac{\text{number of variables with matching value for observations } u \text{ and } v}{\text{total number of variables}}$$

A weakness of the matching coefficient is that if two observations both have a 0 entry for a cetegorical variable, this is counted as a sign of similarity between the two observations.

To avoid misstating similarity due to the absence of a feature, a similarity measure called **Jaccard's coefficient** does not count matching zero entries and is computed as:

$$\frac{\text{number of variables with matching nonzero value for observation } u \text{ and } v}{\text{(total number of variables) - (number of variables with matching zero values for observations } u \text{ and } v\text{)}}$$

Observation	Female	Married	Loan	Mortgage
1	1	0	0	0
2	0	1	1	1
3	1	1	1	0
4	1	1	0	0
5	1	1	0	0

Table 2.1: Sample data table

- Similarity Based on Matching coefficient

Observation	1	2	3	4	5
1	1				
2	0	1			
3	0.5	0.5	1		
4	0.75	0.25	0.75	1	
5	0.75	0.25	0.75	1	1

- Similarity Matrix Based on Jaccard's Coefficient

Observation	1	2	3	4	5
1	1				
2	0	1			
3	0.333	0.5	1		
4	0.5	0.25	0.667	1	
5	0.5	0.25	0.667	1	1

2.3 Hierarchical Clustering

- Determines the similarity of two clusters by considering the similarity between the observations composing either cluster
- Starts with each observation in its own cluster and then iteratively combines the two clusters that are the most similar into a single cluster
- Given a way to measure similarity between observations, there are several clustering method alternatives for comparing observations in two clusters to obtain a cluster similarity measure:
  - Single linkage
  - Complete linkage
  - Group average linkage
  - Median linkage
  - Centroid linkage
- **Single linkage:** The similarity between two clusters is defined by the similarity of the pair of observations (one from each cluster) that are the most similar
- **Complete linkage**: This clustering method defines the similarity between two clusters as the similarity of the pair of observations (one from each cluster) that are the most different
- **Group Average linkage**: Defines the similarity between two clusters to be the average similarity computed over all pairs of observations between the two clusters
- **Median linkage**: Analogous to group average linkage except that it uses the median of the similarities computer between all pairs of observations between the two clusters
- **Centroid linkage** uses the averaging concept of cluster centroids to define between-cluster similarity.

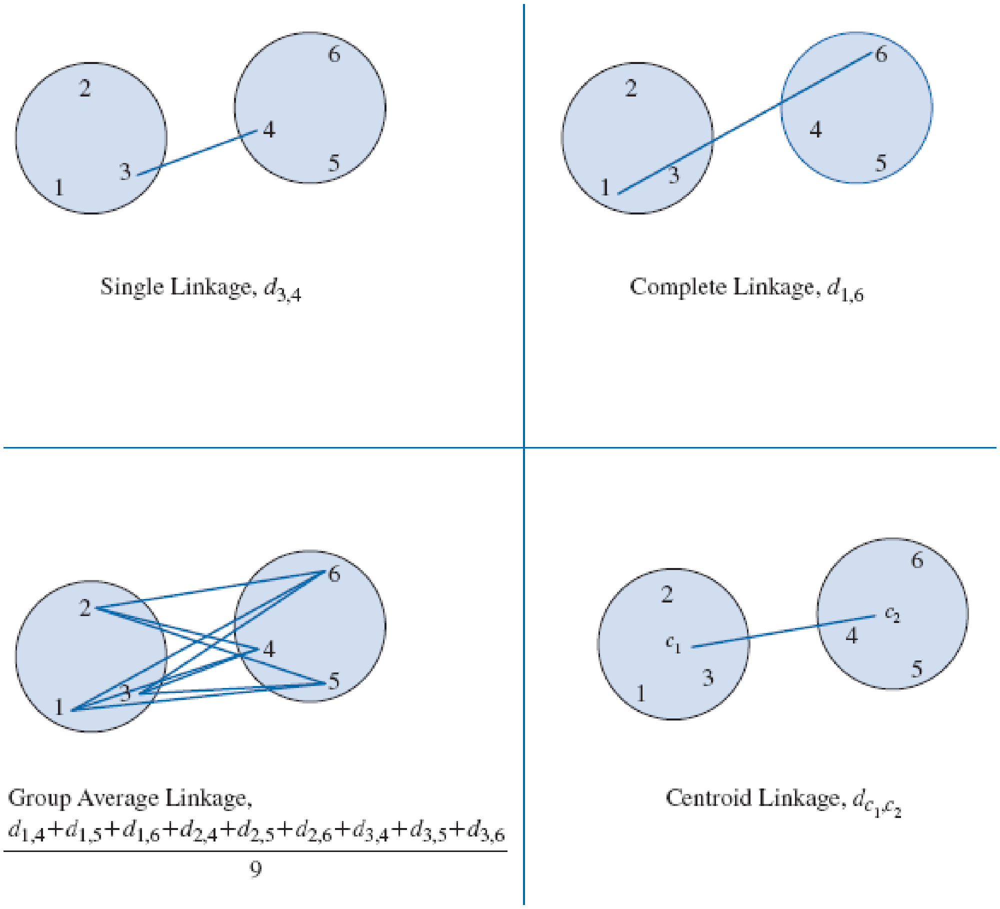

2
3
1
4
6
5
Single Linkage, $d_{3,4}$
2
1
3
6
4
5
Complete Linkage, $d_{1,6}$
2
4
6
3
1
5
Group Average Linkage,
$\frac{d_{1,4}+d_{1,5}+d_{1,6}+d_{2,4}+d_{2,5}+d_{2,6}+d_{3,4}+d_{3,5}+d_{3,6}}{9}$
2
c_1
3
1
6
c_2
4
5
Centroid Linkage, d_{c_1,c_2}

Figure 2.1: Measuring Similarity Between Clusters

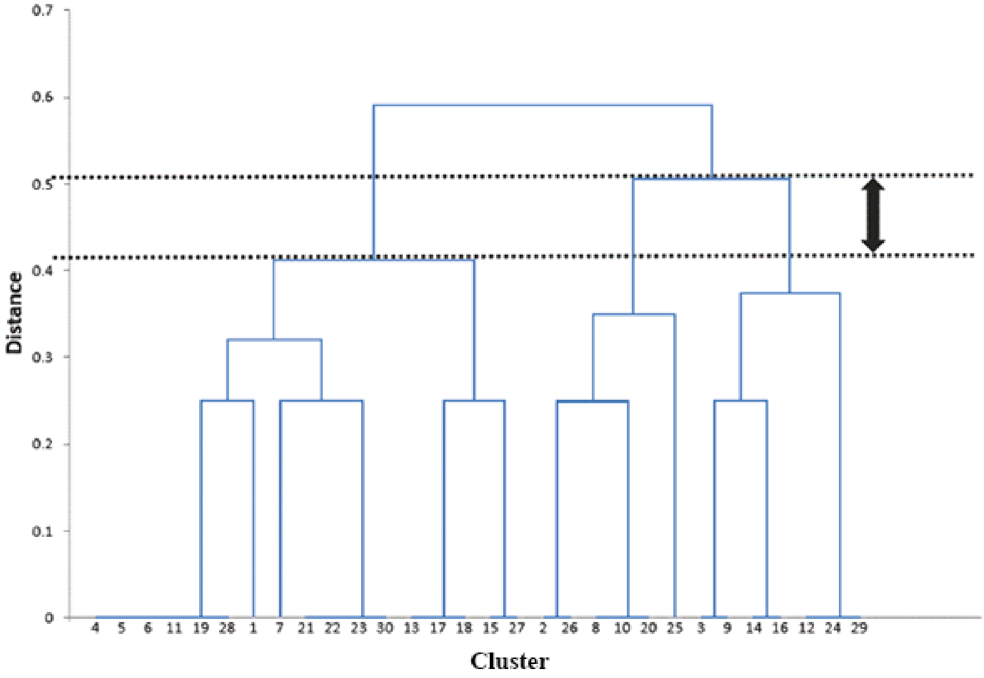


Figure 2.2: Dendrogram for KTC Using Matching Coefficients and Group Average Linkage

2.4 k-Means Clustering

- Given a value of k, the k-means algorithm randomly assigns each observation to one of the k clusters
- After all observations have been assigned to a cluster, the resulting cluster centroid are calculated.
- Using the updated cluster centroids, all observations are reassigned to the cluster with the closest centroid.

	No. of Observations	Average Distance Between Observations in Cluster
Cluster 1	12	0.622
Cluster 2	8	0.739
Cluster 3	10	0.520

Table 2.2: Cluster Observations and Distances

	Cluster 1	Cluster 2	Cluster 3
Cluster 1	0	2.784	1.529
Cluster 2	2.784	0	1.964
Cluster 3	1.529	1.964	0

Table 2.3: Distances Between Clusters

2.5 Hierarchical Clustering vs k-Means Clustering

Hierarchical Clustering

Suitable when we have a small data set (e.g., fewer than 500 observations) and want to easily examine solutions with increasing numbers of clusters.

Convenient method if you want to observe how clusters are nested .

***k*-Means Clustering**

Suitable when you know how many clusters you want and you have a larger data set (e.g., more than 500 observations)

Paritions the observations, which is appropriate if trying to summarize the data with k "average" observations that describe the data with the minimum amount of error. In other words, you're looking to group similar data points together so that each cluster can be represented by a single "average" data point, known as the cluster centroid.

3 Association Rules

- **Association rules:** if-then statements which convey the likelihood of certain items being purchased together.
- Although association rules are an important tool in **market basket analysis**, they are also applicable to other disciplines.
- **Antecendent**: The collection of tiems (or item set) corresponding to the *if* portion of the rule.
- **Consequent**: The item set corresponding to the *then* portion of the rule.
- **Support count** of an item set: **N**umber of transactions in the data that include that item set.

Table 2.4: Shopping Cart Transactions

Transaction	Shopping Cart
1	bread, peanut butter, milk, fruit, jelly
2	bread, jelly, soda, potato chips, milk, fruit, vegetables, peanut butter
3	whipped cream, fruit, chocolate sauce, beer
4	steak, jelly, soda, potato chips, bread, fruit
5	jelly, soda, peanut butter, milk, fruit
6	fruit, soda, potato chips, milk, bread, fruit
7	fruit, soda, potato chips, milk
8	fruit, soda, peanut butter, milk
9	fruit, cheese, yogurt
10	yogurt, vegetables, beer

Confidence: Helps identify reliable association rules:

$$\frac{\text{support of \{antecedent and consequent\}}}{\text{support of antecedent}}$$

Lift ratio: Measure to evalulate the efficiency of a rule:

$$\frac{\text{confidence}}{\text{support of consequent/total number of transactions}}$$

- For the data in Table 1.4, the rule "if {bread, jelly}, then {peanut butter}" has confidence

$$= \frac{2}{4} = 0.5 \text{ and a lift ratio } = \frac{0.5}{\left(\frac{4}{10}\right)} = 1.25$$

Antecedent (A)	Consequent (C)	Support for A	Support for C	Support for A & C	Confidence (%)	Lift Ratio
Bread	Fruit, Jelly	4	5	4	100.0	2.00
Bread	Jelly	4	5	4	100.0	2.00
Bread, Fruit	Jelly	4	5	4	100.0	2.00
Fruit, Jelly	Bread	5	4	4	80.0	2.00
Jelly	Bread	5	4	4	80.0	2.00
Jelly	Bread, Fruit	5	4	4	80.0	2.00
Fruit, Potato Chips	Soda	4	6	4	100.0	1.67
Peanut Butter	Milk	4	6	4	100.0	1.67
Peanut Butter	Milk, Fruit	4	6	4	100.0	1.67
Peanut Butter, Fruit	Milk	4	6	4	100.0	1.67
Potato Chips	Fruit, Soda	4	6	4	100.0	1.67
Fruit, Soda	Potato Chips	6	4	4	66.7	1.67
Milk	Peanut Butter, Fruit	6	4	4	66.7	1.67
Milk, Fruit	Peanut Butter	6	4	4	66.7	1.67
Soda	Potato Chips	6	4	4	66.7	1.67
Fruit, Soda	Milk	6	6	5	83.3	1.39
Milk	Fruit, Soda	6	6	5	83.3	1.39
Milk, Fruit	Soda	6	6	5	83.3	1.39
Soda	Milk	6	6	5	83.3	1.39
Milk	Soda	6	6	5	83.3	1.39
Soda	Milk, Fruit	6	6	5	83.3	1.39

3.1 Evaluating Association Rules:

- An association rule is ultimately judged on how actionable it is and how well it explains the relationship between item sets.
- For example, Walmart mined its transactional data to uncover strong evidence of the association rule, “if a customer purchases a Barbie doll, then a customer also purchased a candy bar.”
- An association rule is useful if it is well supported and explains an important previously unknown relationship.

4 Text Mining

- Text, like numerical data, may contain information that can help sove problems and lead to better decisions.
- **Text mining** is the process of extracting useful information from text data.
- Text data is often referred to as **unstructured data** becuase in its raw form, it cannot be stored in a traditional structured database (rows and columns).
- Audio and video data are also examples of unstructured data.
- Data mining with text data is more challenging than data mining with traditional numerical data, because it requires more preprocessing to conver the text to a format amenable for analysis.

Voice of the Customer at Triad Airline

- Triad solicits feedback from its customers through a follow-up e-mail the day after the customer has completed a flight
- Survey asks the customer to rate various aspects of the flight and asks the respondent to type comments into a dialog box in the e-mail;
  includes:
    - Quantiative feedback from the ratings
    - Comments entered by the respondents which need to be analyzed.

Note:-

A collection of text documents to be analyzed is called a **corpus**

Concerns

The wi-fi service was horrible. It was slow and cut off several times.

My seat was uncomfortable

My flight was delayed 2 hours for no apparent reason.

My seat would not recline.

The man at the ticket counter was rude. Service was horrible.

The flight attendant was rude. Service was bad.

My flight was delayed with no explaination.

My drink spilled when the guy in from of me reclined his seat.

My flight was canceled.

The arm rest of my seat was nasty.

To be analyzed, text data needs to be converted to structured data (rows and columns of numerical data) so that the tools of descriptive statistics, data visualization and data mining can be applied.

Think of converting a group of documents into a matrix of rows and columns where the rows correspond to a document and the columns correspond to a particular word.

A **presence/absence or binary term-document matrix** is a matrix with the rows representing documents and the columns representing words.

- Entries in the columns indicate either the presence or the absence of a particular word in a particular document

Creating the list of terms to use in the presence/absence matrix can be a complicated matter:

- Too many terms results in a matrix with many columns, which may be difficult to manage and could yield meaningless results.
- Too few terms may miss important relationships

Term frequency along with the problem context are often used as a guide.

In Triad's case, management used word frequency and the contex of having a goal of statisfied customers to come up with the following list of terms they feel are relevant for categorizing the respondent's comments: delayed, flight, horrible recline, rude, seat, and service. In Triad's case, management used word frequency and the context of having a goal of statisfied customers to come up with the following list of terms they feel are relevant for categorizing the respondent's comments: delayed, flight, horrible recline, rude, seat, and service.

Document	Delayed	Flight	Horrible	Recline	Rude	Seat	Service
1	0	0	1	0	0	1	1
2	0	0	0	0	0	0	1
3	1	1	0	0	0	0	0
4	0	0	1	0	0	1	1
5	0	0	0	1	1	0	0
6	0	0	1	0	0	1	1
7	0	1	0	0	0	0	0
8	1	1	0	0	0	0	0
9	0	0	0	0	0	1	0
10	0	0	0	0	0	1	0