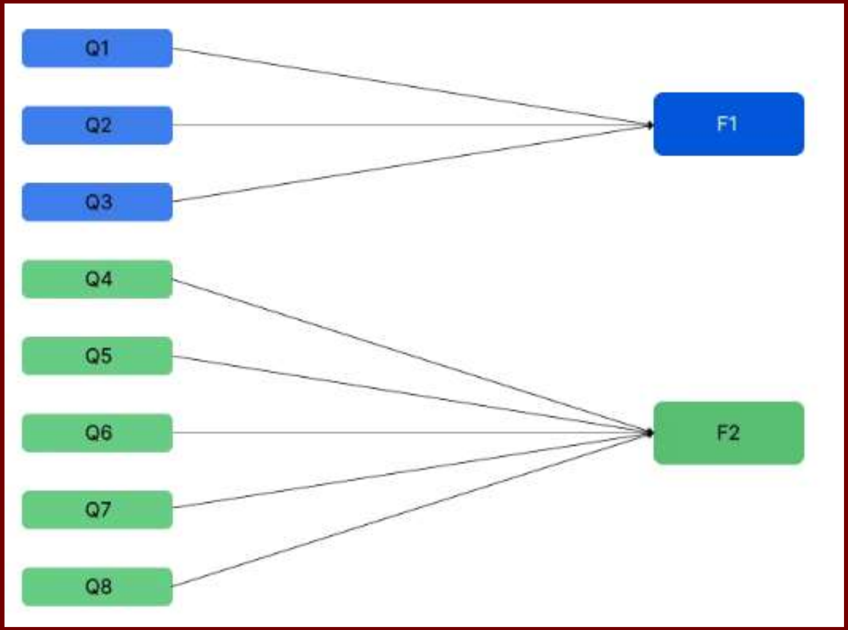
INDEX

1. An introduction to Factor Analysis

Factor analysis can be defined as a technique that is used to condense data of a large number of variables into a smaller number of variables, referred to as factors, so that they can be studied more easily. This technique is generally used to investigate variable relationships that are complex by collapsing several variables into a small number of interpretable factors.



Factor analysis uses the correlation structure amongst observed variables to model a smaller number of unobserved, latent variables known as factors. Researchers use this method when subject-area knowledge suggests that latent factors cause observable variables to covary. Use factor analysis to identify the hidden variables.

Analysts often refer to the observed variables as indicators because they literally indicate information about the factor. Factor analysis treats these indicators as linear combinations of the factors in the analysis plus an error. The procedure assesses how much of the variance each factor explains within the indicators. The idea is that the latent factors create commonalities in some of the observed variables.

Example

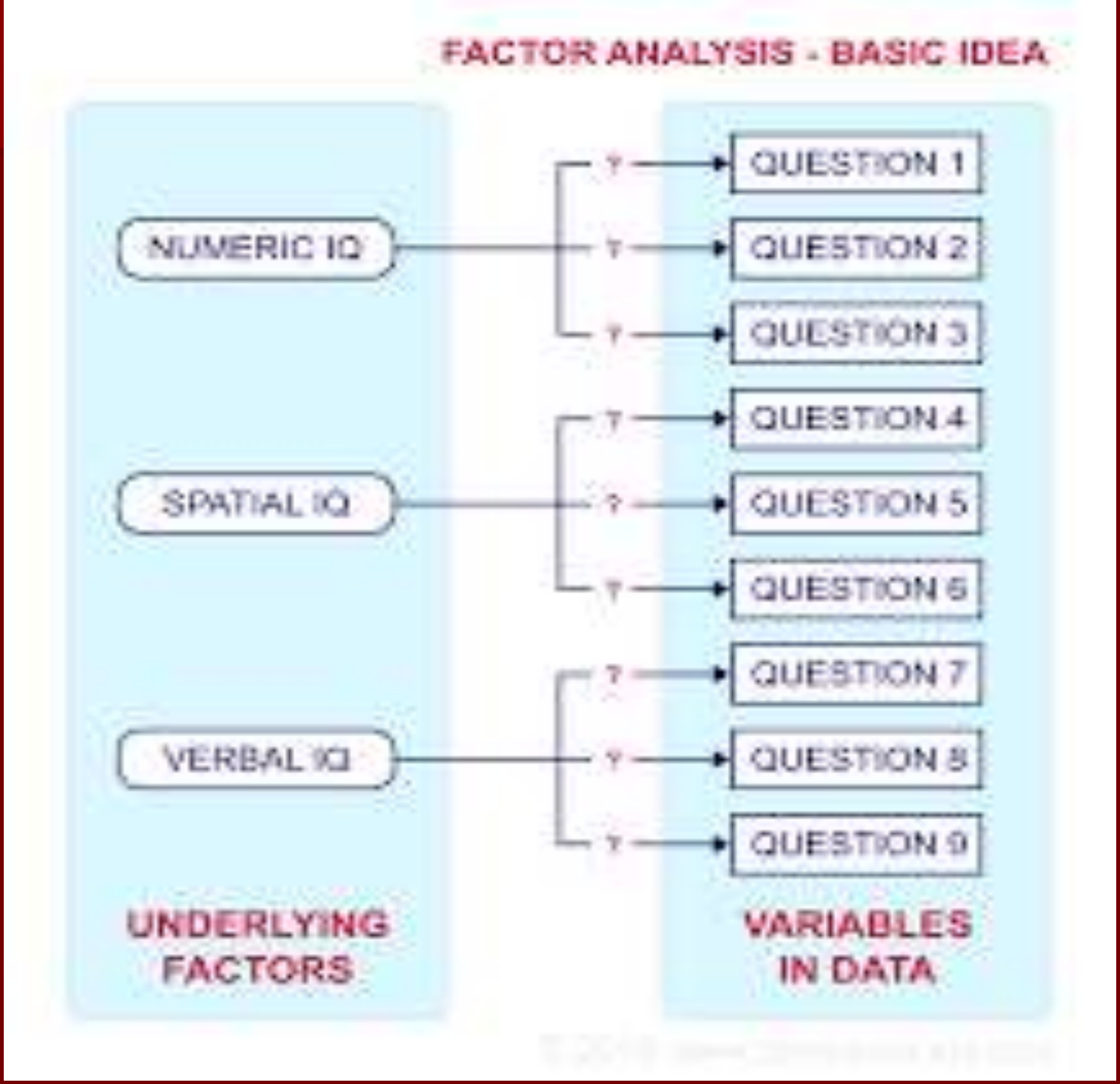
For example, socioeconomic status (SES) is a factor you can't measure directly. However, you can assess occupation, income, and education levels. These variables all relate to socio-economic status. People with a particular socioeconomic status tend to have similar values for the observable variables. If the factor (SES) has a strong relationship with these indicators, then it accounts for a large portion of the variance in the indicators.

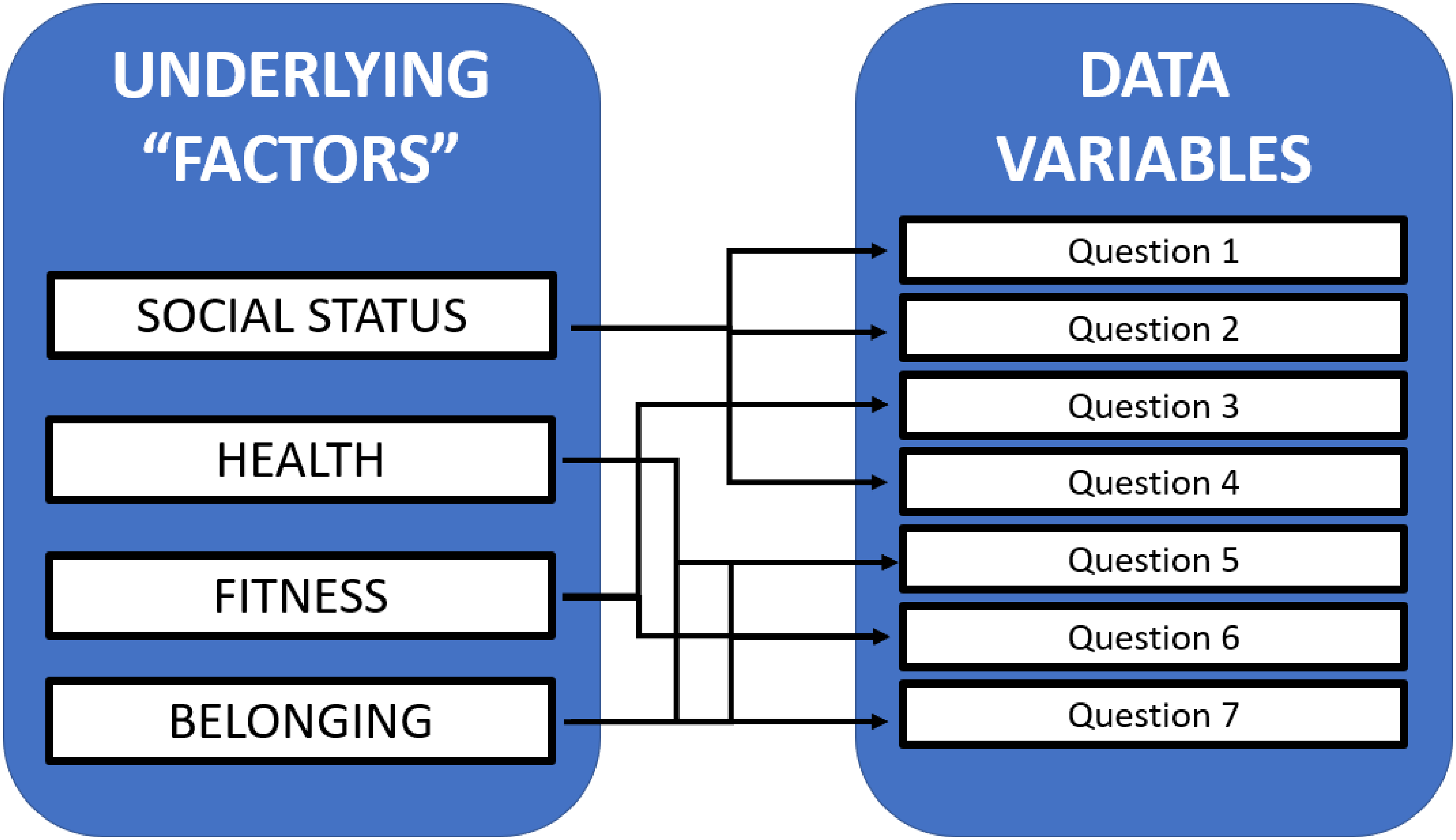
2. Factors

A 'factor' refers to a set of observed variables that have similar response patterns. These variables are associated with a hidden variable known as a confounding variable.

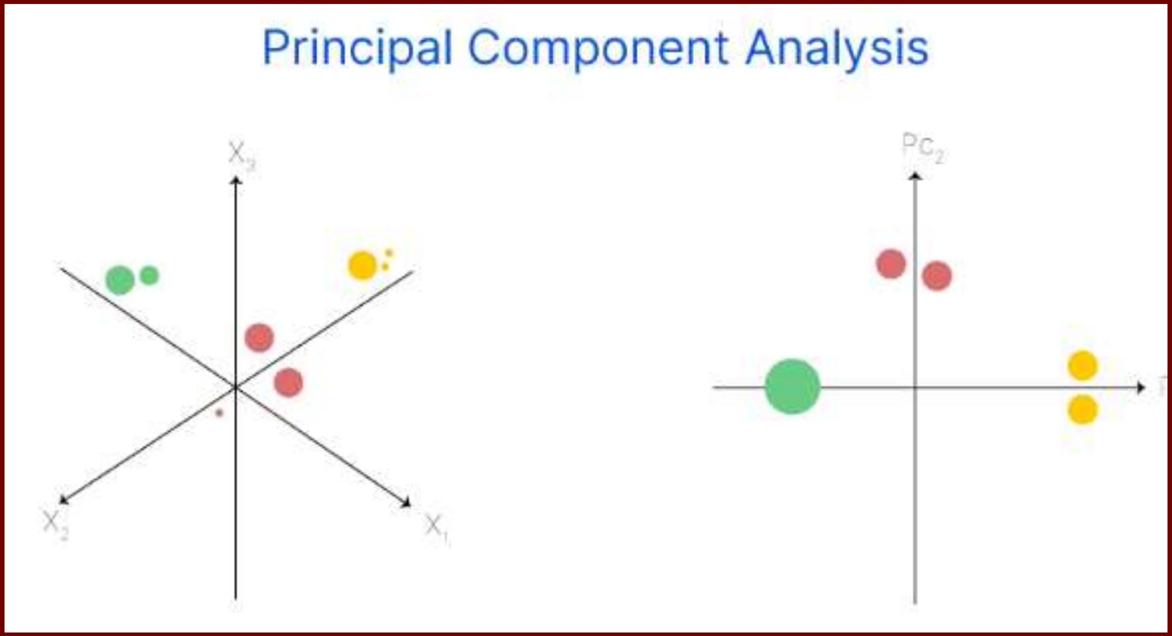
Factor Loadings: Not all factors are created equal, meaning that some have more weight than others. Therefore, some variables will have a stronger association with the underlying latent variables than others. To reflect the difference in the strength of associations, "weights" known as factor loadings are used. Factor loadings are similar to correlation coefficients as they too vary from -1 to +1. The closer the factors are to -1 or +1, the stronger is their association with the latent variable. A factor loading of zero would indicate that the factor has no association with the latent variable and therefore has no effect on it.

3. Basic idea of the concept




UNDERLYING "FACTORS"
SOCIAL STATUS
HEALTH
FITNESS
BELONGING
DATA VARIABLES
Question 1
Question 2
Question 3
Question 4
Question 5
Question 6
Question 7

4. Types of Factor analysis



- **Principal Component Analysis**
- **Maximum Likelihood Method**
- **Common Factor Analysis**

5. Principal component analysis

Principal component analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. PCA starts extracting the maximum variance and puts them into the first factor. After that, it removes that variance explained by the first factors and then starts extracting maximum variance for the second factor. This process goes to the last factor.

6. Maximum likelihood method

It is the statistical method of estimating the parameters of the probability distribution by maximizing the likelihood function. The point in which the parameter value that maximizes the likelihood function is called the maximum likelihood estimate.

7. Common factor analysis

It extracts the common variance and puts them into factors. This method does not include the unique variance of all variables. This method is used in Search Engine Marketing (SEM).

8. Cluster analysis

Cluster analysis is a statistical method used to group similar objects into respective categories. It can also be referred to as segmentation analysis, taxonomy analysis, or clustering.

The goal of performing a cluster analysis is to sort different objects or data points into groups in a manner that the degree of association between two objects is high if they belong to the same group, and low if they belong to different groups.

Cluster analysis differs from many other statistical methods due to the fact that it's mostly used when researchers do not have an assumed principle or fact that they are using as the foundation of their research.

9. Types of Cluster Analysis

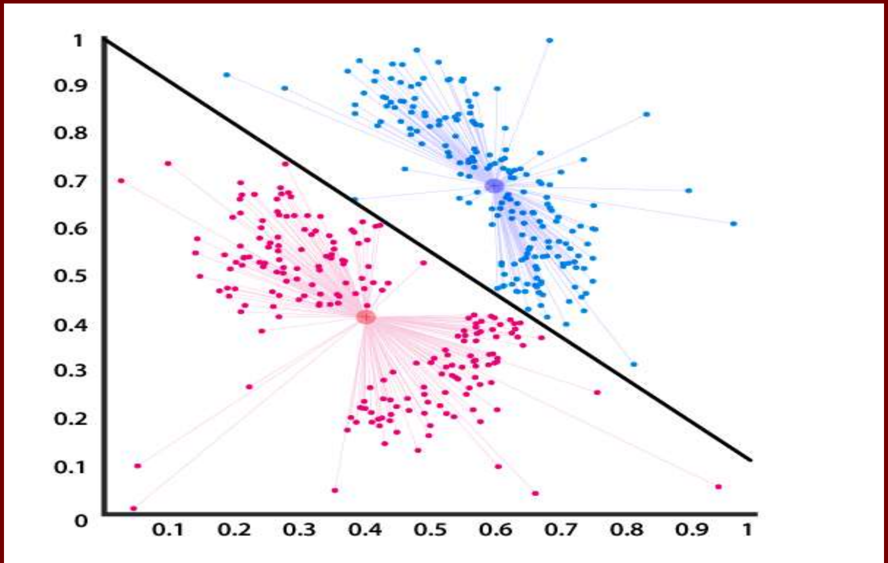
- **Hierarchical clustering**

First, a cluster is made and then added to another cluster (the most similar and closest one) to form one single cluster. This process is repeated until all subjects are in one cluster. This particular method is known as Agglomerative method. Agglomerative clustering starts with single objects and starts grouping them into clusters.

The divisive method is another kind of Hierarchical method in which clustering starts with the complete data set and then starts dividing into partitions.

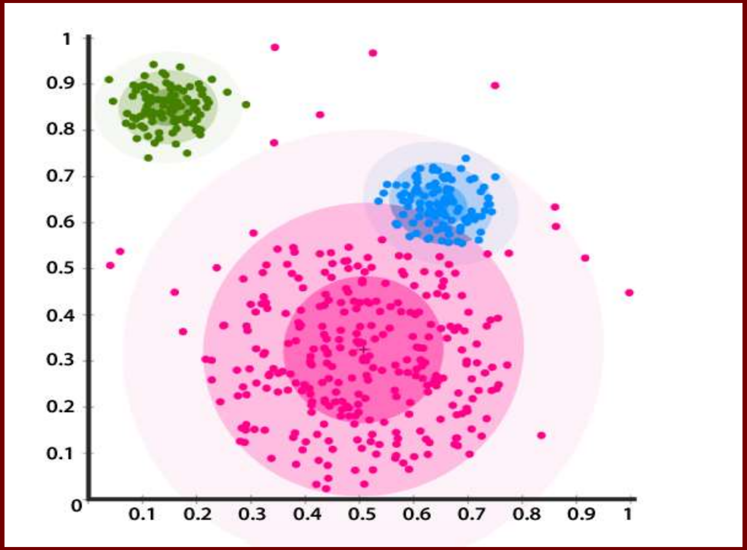
Centroid-based Clustering

Clusters are represented by a central entity, which may or may not be a part of the given data set. K-Means method of clustering is used in this method, where k are the cluster centers and objects are assigned to the nearest cluster centers.



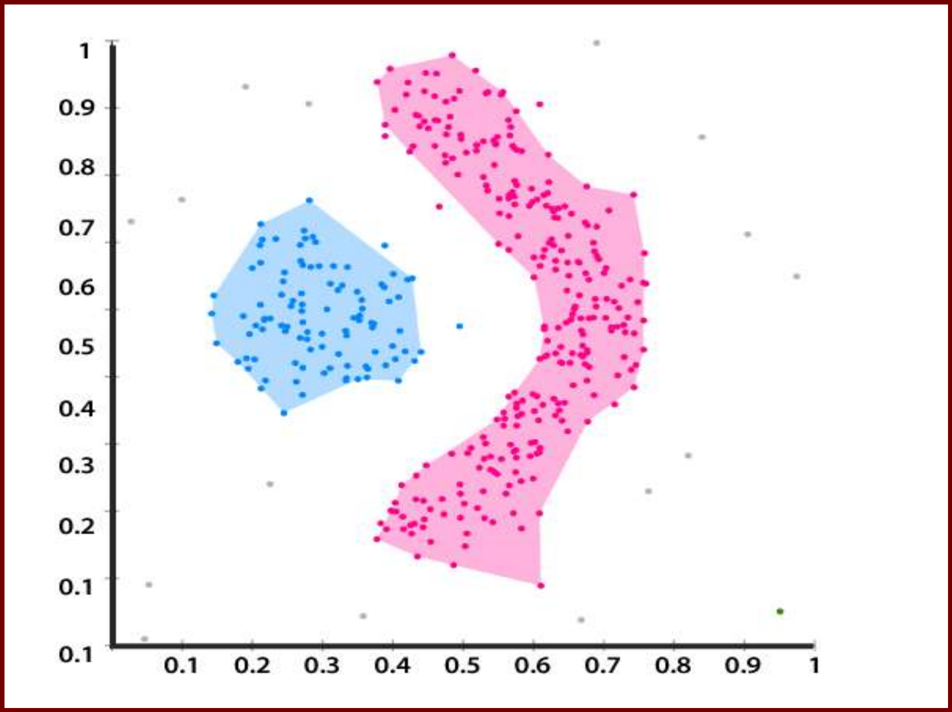
Distribution-based Clustering

It is a type of clustering model closely related to statistics based on the modals of distribution. Objects that belong to the same distribution are put into a single cluster. This type of clustering can capture some complex properties of objects like correlation and dependence between attributes.



Density-based Clustering

In this type of clustering, clusters are defined by the areas of density that are higher than the remaining of the data set. Objects in sparse areas are usually required to separate clusters.



10. Applications of Cluster Analysis

This analysis technique is typically performed during the exploratory phase of research, since unlike techniques such as factor analysis, it doesn't make any distinction between dependent & independent variables. Instead, cluster analysis is leveraged mostly to discover structures in data without providing an explanation/ interpretation.

Cluster analysis discovers structures in data without explaining why those structures exist. For example, when cluster analysis is performed as part of market research, specific groups can be identified within a population. The analysis of these groups can then determine how likely a population cluster is to purchase products or services. If these groups are defined clearly, a marketing team can then target varying cluster with tailored, targeted communication.

Applications of Cluster Analysis

- Marketing
- Insurance
- Geology

11. Benefits of Cluster Analysis

Clustering allows researchers to identify and define patterns between data elements. Revealing these patterns between data points helps to distinguish and outline structures which might not have been apparent before, but which give significant meaning to the data once they are discovered. Once a clearly defined structure emerges from the dataset at hand, informed decision-making becomes much easier.

12. Conjoint Analysis

For a business to run effectively, its leadership needs a firm understanding of the value its products or services bring to consumers. This understanding allows for a more informed strategy across the board—from long-term planning to pricing and sales.

Conjoint analysis is a form of statistical analysis that firms use in market research to understand how customers value different components or features of their products or services. It's based on the principle that any product can be broken down into a set of attributes that ultimately impact users' perceived value of an item or service. It is typically conducted via a specialized survey that asks consumers to rank the importance of the specific features in question. Analyzing the results allows the firm to then assign a value to each one.

- **Choice-Based Conjoint (CBC) Analysis:** It is one of the most common forms of conjoint analysis and is used to identify how a respondent values combinations of features.
- **Adaptive Conjoint Analysis (ACA):** This form of analysis customizes each respondent's survey experience based on their answers to early questions. It's often leveraged in studies where several features or attributes are being evaluated to streamline the process and extract the most valuable insights from each respondent.
- **Full-Profile Conjoint Analysis:** This form of analysis presents the respondent with a series of full product descriptions and asks them to select the one they'd be most inclined to buy.
- **MaxDiff Conjoint Analysis:** This form of analysis presents multiple options to the respondent, which they're asked to organize on a scale of "best" to "worst" (or "most likely to buy" to "least likely to buy").

13. Uses of Conjoint Analysis

The insights a company gleans from conjoint analysis of its product features can be leveraged in several ways. Most often, conjoint analysis impacts pricing strategy, sales and marketing efforts, and research and development plans.

- **Conjoint Analysis in Pricing:** Conjoint analysis works by asking users to directly compare different features to determine how they value each one. When a company understands how its customers value its products' or services' features, it can use the information to develop its pricing strategy.

Uses of Conjoint Analysis

- **Conjoint Analysis in Sales & Marketing:** Conjoint analysis can inform more than just a company's pricing strategy; it can also inform how it markets and sells its offerings. When a company knows which features its customers value most, it can lean into them in its advertisements, marketing copy, and promotions.
- **Conjoint Analysis in Research & Development:** Conjoint analysis can also inform a company's research and development pipeline. The insights gleaned can help determine which new features are added to its products or services, along with whether there's enough market demand for an entirely new product.

14. Discriminant Analysis

Discriminant analysis is a group classification method similar to regression analysis, in which individual groups are classified by making predictions based on independent variables. It is a very popular tool used in statistics & helps companies improve decision making, processes, & solutions across diverse business lines. In marketing, this technique is commonly used to predict customer trends; in finance, it's applied in areas such as bank loan application approval; in image recognition, it can be very accurate in instances of pattern recognition.

15. Methods of Discriminant Analysis

- **Linear**

Linear discriminant analysis is often used in machine learning applications and pattern classification. It's commonly used for dimensionality reduction, which minimizes the number of variables that are being considered.

- **Quadratic**

Similar to linear discrimination analysis, but with observations made from the normal distribution, with each class having its own covariant matrix. In machine learning, it separates the measurements for two or more event categories.

Methods of Discriminant Analysis

- **Canonical**

Measures the connection or correlation between two unique sets of variables, which are split into different groups (X and Y), so that the relationship between the two variables can be further explored.

- **Gaussian**

Also called normal distribution, this method involves a distribution that is dependent on the mean and standard deviation of a data set. Data scientists often use this when working on artificial intelligence (AI) projects.

Discriminant Analysis in business

Businesses use discriminant analysis to help grasp meaning from data sets so that they can drive creative, competitive solutions surrounding the customer experience, personalization, marketing, making predictions, and many other popular strategic purposes. Its applications are only increasing and indicating signs of being more and more useful moving into the future, with new techniques being adapted for many emerging business challenges.

- **Human Resources**

Used to measure potential candidate job performance by using background information to predict how they will perform if hired.

Discriminant Analysis in business

- **Industrial**

Can predict when machine parts might break down or need repairs based on performance indicators.

- **Sales and Marketing**

Can predict market trends that have an effect on new products or services.

16. Benefits of Discriminant Analysis

- Accelerate Trusted Decision Making
- Predict Patterns and Behaviors
- Solve Challenges
- Machine Learning
- Improve Marketing

17. Data Analyst tools

Data analyst tools is a term used to describe software and applications that data analysts use in order to develop and perform analytical processes that help companies to make better, informed business decisions, while decreasing costs and increasing profits.

18. Tools that Data Analysts use

DATA ANALYST TOOLS CATEGORY	SOLUTION EXAMPLE		HEADQUARTERS
Business Intelligence Tools	datapine	datapine	Berlin, Berlin, Germany
Statistical Analysis Tools	R Studio	R-Studio	Boston, Massachusetts, USA
Purpose Programming Languages	python	Python	Wilmington, Delaware, USA
SQL Consoles	MySQL	MySQL Workbench	Austin, Texas, USA
Standalone Predictive Analytics Tools	sas	SAS Forecasting	Cary, North Carolina, USA
Data Modeling Tools	erwin	Erwin Data Modeler	Melville, New York, USA
ETL Tools	talend	Talend	Redwood City, California, USA
Automation Tools	Jenkins	Jenkins	London, London, England
Unified Data Analytics Engines	Spark	Apache Spark	Wilmington, Delaware, USA
Spreadsheet Applications	X	Microsoft Excel	Redmond, Washington, USA
Industry-Specific Analytics Tools	qualtrics	Qualtrics	Provo, Utah, USA
Data Science Platforms	rapidminer	RapidMiner	Boston, Massachusetts, USA
Data Cleansing Tools	OpenRefine	OpenRefine	Melbourne, Victoria, Australia
Data Visualization Tools & Platforms	HIGHCHARTS	Highcharts	Vik i Sogn, Sogn og Fjordane, Norway

ETL-Extract, transform and load

Data Analyst tools

- Microsoft Excel
- Python
- R
- Jupyter Notebook
- Apache Spark
- SAS
- Microsoft Power BI
- Tableau
- KNIME

How to choose a data analysis tool

First, consider that there's no one singular data analytics tool that will address all the data analytics issues you may have. When looking at this list, you may look at one tool for most of your needs, but require the use of a secondary tool for smaller processes.

Data Analyst tools

Second, consider the business needs of your organization and figure out exactly who will need to make use of the data analysis tools. Will they be used primarily by fellow data analysts or scientists, non-technical users who require an interactive and intuitive interface—or both? Many tools on this list will cater to both types of user.

Third, consider the tool's data modeling capabilities.

Fourth, consider the practical aspect of price and licensing. Some of the options are totally free or have some free-to-use features (but will require licensing for the full product). Some tools will be offered on a subscription or licensing basis. In this case, you may need to consider the number of users required or—if you're looking on solely a project-to-project basis—the potential length of the subscription.

Data Analyst tools

a. Microsoft Excel

- **Type of tool:** Spreadsheet software
- **Availability:** Commercial
- **Mostly used for:** Data wrangling and reporting
- **Pros:** Widely-used, with lots of useful functions and plug-ins
- **Cons:** Cost, calculation errors, poor at handling big data

b. Python

- **Type of tool:** Programming language
- **Availability:** Open-source, with thousands of free libraries
- **Used for:** Everything from data scraping to analysis & reporting
- **Pros:** Easy to learn, highly versatile, widely-used
- **Cons:** Memory intensive-doesn't execute as fast as some other languages

Data Analyst tools

c. R

- **Type of tool:** Programming language
- **Availability:** Open-source
- **Mostly used for:** Statistical analysis and data mining
- **Pros:** Platform independent, highly compatible, lots of packages
- **Cons:** Slower, less secure, and more complex to learn than Python

d. Jupyter Notebook

- **Type of tool:** Interactive authoring software
- **Availability:** Open-source
- **Mostly used for:** Sharing code, creating tutorials, presenting work
- **Pros:** Great for showcasing, language-independent
- **Cons:** Not self-contained, nor great for collaboration

Data Analyst tools

e. Apache Spark

- **Type of tool:** Data processing framework
- **Availability:** Open-source
- **Mostly used for:** Big data processing, machine learning
- **Pros:** Fast, dynamic, easy to use
- **Cons:** No file management system, rigid user interface

f. SAS

- **Type of tool:** Statistical software suite
- **Availability:** Commercial
- **Mostly used for:** Business intelligence, multivariate, and predictive analysis
- **Pros:** Easily accessible, business-focused, good user support
- **Cons:** High cost, poor graphical representation

Data Analyst tools

g. Microsoft Power BI

- **Type of tool:** Business analytics suite
- **Availability:** Commercial software (with a free version available)
- **Mostly used for:** Everything from data visualization to predictive analytics
- **Pros:** Great data connectivity, regular updates, good visualizations
- **Cons:** Clunky user interface, rigid formulas, data limits (in the free version)

h. Tableau

- **Type of tool:** Data visualization tool
- **Availability:** Commercial
- **Mostly used for:** Creating data dashboards & worksheets
- **Pros:** Great visualizations, speed, interactivity, mobile support
- **Cons:** Poor version control, no data pre-processing

Data Analyst tools

i. KNIME

- **Type of tool:** Data integration platform
- **Availability:** Open-source
- **Mostly used for:** Data mining and machine learning
- **Pros:** Open-source platform that is great for visually-driven programming
- **Cons:** Lacks scalability, and technical expertise is needed for some functions