Research Methodology for PhDs

GBSB GLOBAL
BUSINESS SCHOOL

Session 12_1 Topics

- Non-Standard and Complex Methods:
  - Factor Analysis
  - Heteroscedasticity, auto- and multicorrelation
  - Cluster Analysis

Factor Analysis

Factor analysis is a statistical method used to analyze the relationships among a set of observed variables by explaining the correlations or covariances between them in terms of a smaller number of unobserved variables called factors.

In the context of factor analysis, a “factor” refers to an underlying, unobserved variable or latent construct that represents a common source of variation among a set of observed variables.

These observed variables, also known as indicators or manifest variables, are the measurable variables that are directly observed or measured in a study.

©https://www.geeksforgeeks.org/introduction-to-factor-analytics/

Purposes of Factor Analysis

Simplify Your Data: Imagine a giant ball of yarn – that's your complex data. Factor analysis untangles it, revealing a smaller number of core threads (factors) that make up the whole thing.

Find Hidden Connections: Beyond just fewer threads, factor analysis reveals how these core threads are secretly connected. It spots hidden patterns that explain why some variables move together.

Understand the Bigger Picture: By seeing these hidden connections, you can understand the underlying forces at play in your data. It helps you move from "what" (variables) to "why" (factors) that truly influence your results.

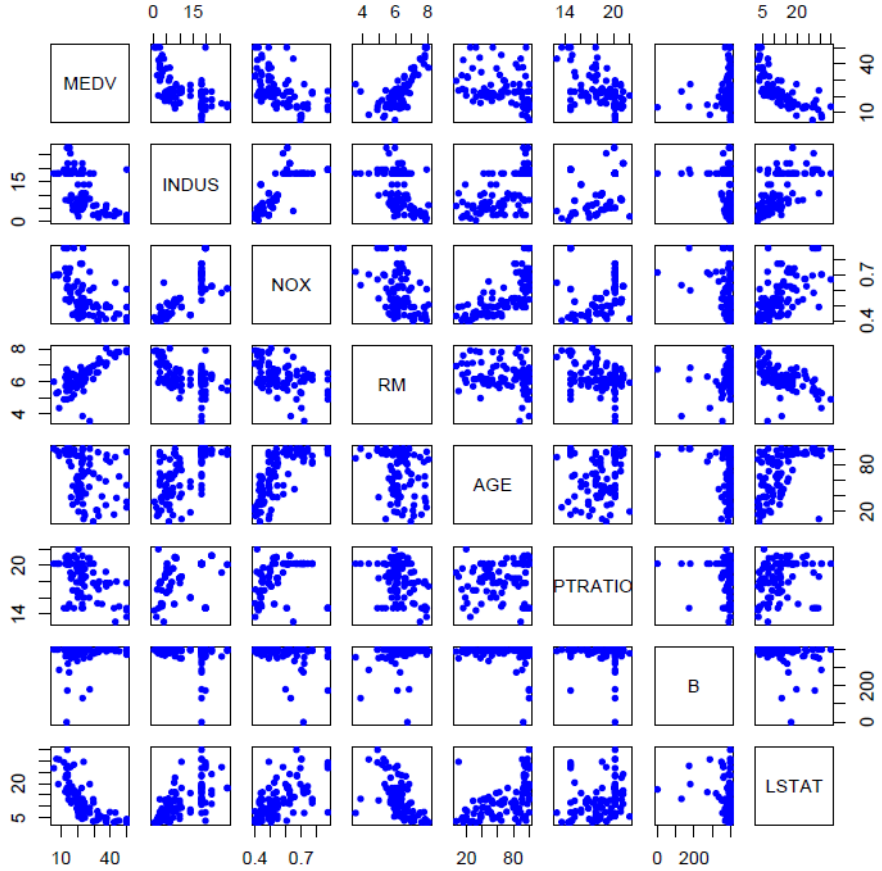
©https://www.geeksforgeeks.org/introduction-to-factor-analytics/

Example – Real Estate Costs Analysis

- 1. CRIM— Crime rate,
- 2. ZN —Share of residential development,
- 3. INDUS— share of business development,
- 4. CHAS— Availability of a recreation area,
- 5. NOX— Ecological condition of the site,
- 6. RM— Number of rooms,
- 7. AGE—Year of construction,
- 8. DIS— Remoteness from business centers,
- 9. RAD— Remoteness from highways,
- 10. LSTAT – Share of residents of low social status

$$\text{MEDV} = \beta_0 + \beta_1 \cdot \text{CRIM} + \beta_2 \cdot \text{ZN} + \ldots + \beta_{13} \cdot \text{LSTAT}.$$

Example – Real Estate Costs Analysis



Terms

The eigen (factor) number is a value that reflects how important a factor can be extracted from the data set. The factor number is calculated based on eigenvectors, and the higher the number, the more important the corresponding factor.

Factor space is a multidimensional space in which each variable is represented by a factor. The factor space shows how strong the correlation between different indicators is and allows you to see which variables are closer to each other and which are farther away.

Terms

Factor burden (or workload) is a coefficient that shows how much each indicator affects a given factor. The factor load can be positive or negative, which indicates the direction of the influence. The higher the factor load, the greater the role of a particular indicator in the formation of this factor.

Utility (or communality factor) is a coefficient that shows how much a separate indicator is explained by a common factor. Utility is close to one if the variable is a good representative of this factor. The closer the utility is to zero, the weaker the relationship of this variable with the general factor.

Model Types in Factor Analysis

The general model of factor analysis assumes the existence of a general factor and specific factors that affect the variables. For example, this model is used to identify a separate or general factor that affects profits, expenses, and sales volumes.

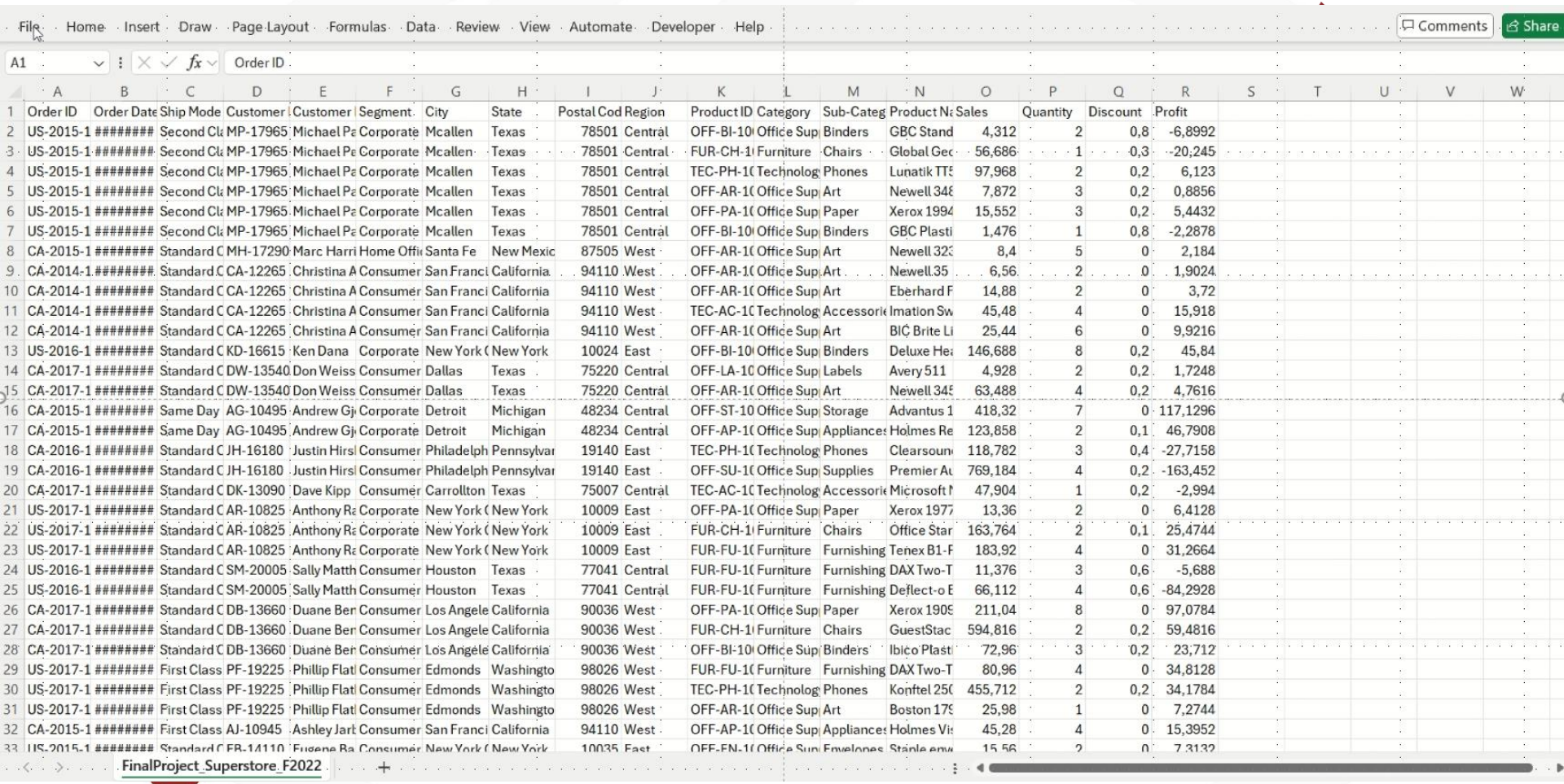
Factor Load Analysis Model – Determines which factors have the greatest impact on a particular outcome or variable. For example, the CFO can use this model in budgeting to take into account the impact of resource prices, inflation, changes in expenses on future profits, and other financial indicators.

Model Types in Factor Analysis

Principal Component Method (PCA) is a factor analysis technique that is used to isolate the most important factors from a large number of variables. It is based on the search for linear combinations of variables that explain the largest share of data variability. These linear combinations are called principal components.

Partial Least Squares (PLS): This method is widely used to solve multiple regression problems in large dimensions of data. It allows you to reduce a large set of dimensions to a smaller number of meaningful components that best describe the relationships between them.

Maximum Likelihood Estimation (MLE) - This method aims to evaluate the parameters of a statistical model based on available data. It allows you to determine the value of the parameters that most likely corresponds to the observed data



File Home Insert Draw Page Layout Formulas Data Review View Automate Developer Help Comments Share

A1 Order ID

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Order ID	Order Date	Ship Mode	Customer	Customer	Segment	City	State	Postal Cod	Region	Product ID	Category	Sub-Categ	Product Na	Sales	Quantity	Discount	Profit
2	US-2015-1	########	Second Cla	MP-17965	Michael Pa	Corporate	Mcallen	Texas	78501	Central	OFF-BI-10	Office Sup	Binders	GBC Stand	4,312	2	0,8	-6,8992
3	US-2015-1	########	Second Cla	MP-17965	Michael Pa	Corporate	Mcallen	Texas	78501	Central	FUR-CH-1	Furniture	Chairs	Global Gec	56,686	1	0,3	-20,245
4	US-2015-1	########	Second Cla	MP-17965	Michael Pa	Corporate	Mcallen	Texas	78501	Central	TEC-PH-1	Technolog	Phones	Lunatik TT	97,968	2	0,2	6,123
5	US-2015-1	########	Second Cla	MP-17965	Michael Pa	Corporate	Mcallen	Texas	78501	Central	OFF-AR-1	Office Sup	Art	Newell 348	7,872	3	0,2	0,8856
6	US-2015-1	########	Second Cla	MP-17965	Michael Pa	Corporate	Mcallen	Texas	78501	Central	OFF-PA-1	Office Sup	Paper	Xerox 1994	15,552	3	0,2	5,4432
7	US-2015-1	########	Second Cla	MP-17965	Michael Pa	Corporate	Mcallen	Texas	78501	Central	OFF-BI-10	Office Sup	Binders	GBC Plasti	1,476	1	0,8	-2,2878
8	CA-2015-1	########	Standard C	MH-17290	Marc Harri	Home Offi	Santa Fe	New Mexic	87505	West	OFF-AR-1	Office Sup	Art	Newell 323	8,4	5	0	2,184
9	CA-2014-1	########	Standard C	CA-12265	Christina A	Consumer	San Franci	California	94110	West	OFF-AR-1	Office Sup	Art	Newell 35	6,56	2	0	1,9024
10	CA-2014-1	########	Standard C	CA-12265	Christina A	Consumer	San Franci	California	94110	West	OFF-AR-1	Office Sup	Art	Eberhard F	14,88	2	0	3,72
11	CA-2014-1	########	Standard C	CA-12265	Christina A	Consumer	San Franci	California	94110	West	TEC-AC-1	Technolog	Accessorie	Imation Sw	45,48	4	0	15,918
12	CA-2014-1	########	Standard C	CA-12265	Christina A	Consumer	San Franci	California	94110	West	OFF-AR-1	Office Sup	Art	BIC Brite Li	25,44	6	0	9,9216
13	US-2016-1	########	Standard C	KD-16615	Ken Dana	Corporate	New York (	New York	10024	East	OFF-BI-10	Office Sup	Binders	Deluxe Hea	146,688	8	0,2	45,84
14	CA-2017-1	########	Standard C	DW-13540	Don Weiss	Consumer	Dallas	Texas	75220	Central	OFF-LA-1	Office Sup	Labels	Avery 511	4,928	2	0,2	1,7248
15	CA-2017-1	########	Standard C	DW-13540	Don Weiss	Consumer	Dallas	Texas	75220	Central	OFF-AR-1	Office Sup	Art	Newell 345	63,488	4	0,2	4,7616
16	CA-2015-1	########	Same Day	AG-10495	Andrew Gi	Corporate	Detroit	Michigan	48234	Central	OFF-ST-10	Office Sup	Storage	Advantus 1	418,32	7	0	117,1296
17	CA-2015-1	########	Same Day	AG-10495	Andrew Gi	Corporate	Detroit	Michigan	48234	Central	OFF-AP-1	Office Sup	Appliances	Holmes Re	123,858	2	0,1	46,7908
18	CA-2016-1	########	Standard C	JH-16180	Justin Hirsl	Consumer	Philadelph	Pennsylvar	19140	East	TEC-PH-1	Technolog	Phones	Clearsoun	118,782	3	0,4	-27,7158
19	CA-2016-1	########	Standard C	JH-16180	Justin Hirsl	Consumer	Philadelph	Pennsylvar	19140	East	OFF-SU-1	Office Sup	Supplies	Premier Au	769,184	4	0,2	-163,452
20	CA-2017-1	########	Standard C	DK-13090	Dave Kipp	Consumer	Carrollton	Texas	75007	Central	TEC-AC-1	Technolog	Accessorie	Microsoft N	47,904	1	0,2	-2,994
21	US-2017-1	########	Standard C	AR-10825	Anthony Ra	Corporate	New York (	New York	10009	East	OFF-PA-1	Office Sup	Paper	Xerox 1977	13,36	2	0	6,4128
22	US-2017-1	########	Standard C	AR-10825	Anthony Ra	Corporate	New York (	New York	10009	East	FUR-CH-1	Furniture	Chairs	Office Star	163,764	2	0,1	25,4744
23	US-2017-1	########	Standard C	AR-10825	Anthony Ra	Corporate	New York (	New York	10009	East	FUR-FU-1	Furniture	Furnishing	Tenex B1-F	183,92	4	0	31,2664
24	US-2016-1	########	Standard C	SM-20005	Sally Matth	Consumer	Houston	Texas	77041	Central	FUR-FU-1	Furniture	Furnishing	DAX Two-T	11,376	3	0,6	-5,688
25	US-2016-1	########	Standard C	SM-20005	Sally Matth	Consumer	Houston	Texas	77041	Central	FUR-FU-1	Furniture	Furnishing	Deflect-o E	66,112	4	0,6	-84,2928
26	CA-2017-1	########	Standard C	DB-13660	Duane Ber	Consumer	Los Angele	California	90036	West	OFF-PA-1	Office Sup	Paper	Xerox 1909	211,04	8	0	97,0784
27	CA-2017-1	########	Standard C	DB-13660	Duane Ber	Consumer	Los Angele	California	90036	West	FUR-CH-1	Furniture	Chairs	GuestStac	594,816	2	0,2	59,4816
28	CA-2017-1	########	Standard C	DB-13660	Duane Ber	Consumer	Los Angele	California	90036	West	OFF-BI-10	Office Sup	Binders	Ibico Plasti	72,96	3	0,2	23,712
29	US-2017-1	########	First Class	PF-19225	Phillip Flat	Consumer	Edmonds	Washingto	98026	West	FUR-FU-1	Furniture	Furnishing	DAX Two-T	80,96	4	0	34,8128
30	US-2017-1	########	First Class	PF-19225	Phillip Flat	Consumer	Edmonds	Washingto	98026	West	TEC-PH-1	Technolog	Phones	Konftel 250	455,712	2	0,2	34,1784
31	US-2017-1	########	First Class	PF-19225	Phillip Flat	Consumer	Edmonds	Washingto	98026	West	OFF-AR-1	Office Sup	Art	Boston 179	25,98	1	0	7,2744
32	CA-2015-1	########	First Class	AJ-10945	Ashley Jarl	Consumer	San Franci	California	94110	West	OFF-AP-1	Office Sup	Appliances	Holmes Vis	45,28	4	0	15,3952
33	US-2015-1	########	Standard C	FB-14110	Eugene Ba	Consumer	New York (	New York	10035	East	OFF-EN-1	Office Sup	Envelopes	Staple env	15,56	2	0	7,3132

FinalProject_Superstore_F2022

Factor Analysis Advantages

Uncover hidden relationships between variables to better understand data trends.

create models and predict risks;

to detect bottlenecks in the work of the enterprise;

influence the results of activities.

Factor Analysis Challenges & Limitations

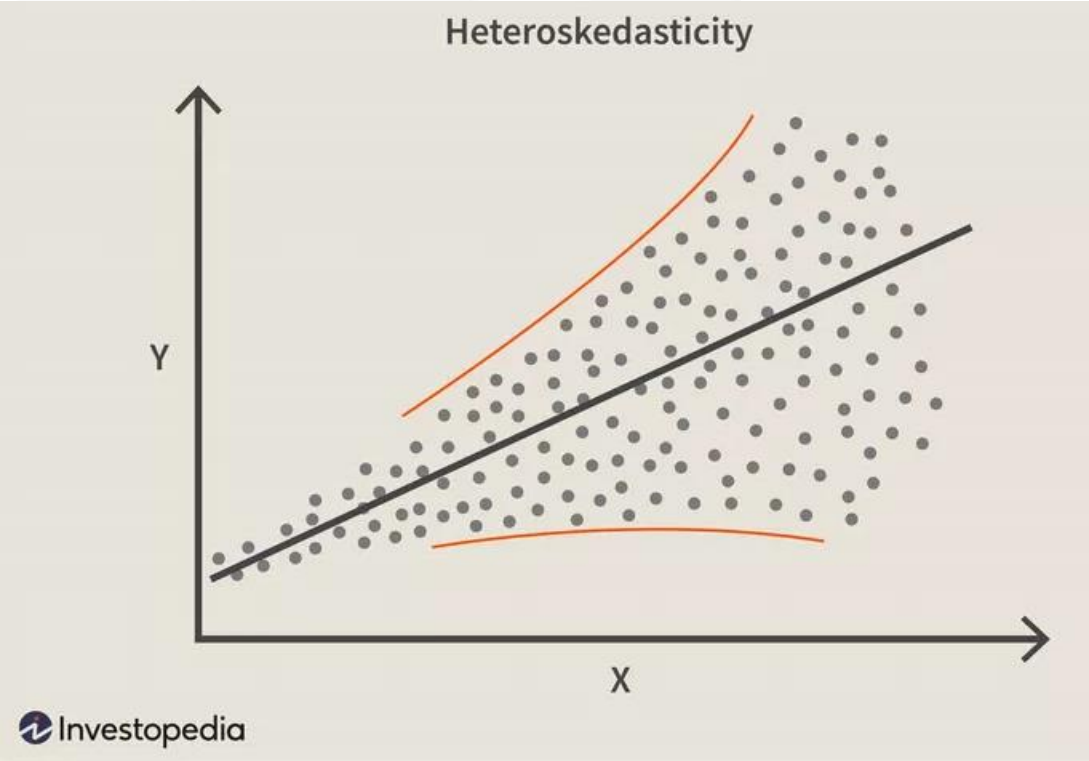
- Factors are defined subjectively
- Most Software models apply linear analysis
- Data are inter-dependent

Heterscedacity

In statistics, heteroskedasticity (or heteroscedasticity) happens when the standard deviations of a predicted variable, monitored over different values of an independent variable or as related to prior time periods, are non-constant.

With heteroskedasticity, the tell-tale sign upon visual inspection of the residual errors is that they will tend to fan out over time,

Heteroscedacity



Cluster analysis

- **Cluster analysis** is used for automatic identification of natural groupings
- of things. It is also known as the segmentation technique. In this technique,
- data instances that are similar to (or near) each other are categorized
- into one cluster.

Example:

- Age: 0-10 – Kids
- 11 - 20 – teenagers
- 21-45 – young
- 46 – 60 – middle aged
- 61 – 75 – elderly
- Above 75 - old

Business Applications and Tasks

- Market segmentation (geographic, vertical, or any other)
- Product segmentation, i.e. Basic, Business, “Gold” VIP services
- Text mining
- Loan borrower segmentation
- Customer segments
- Social net groups

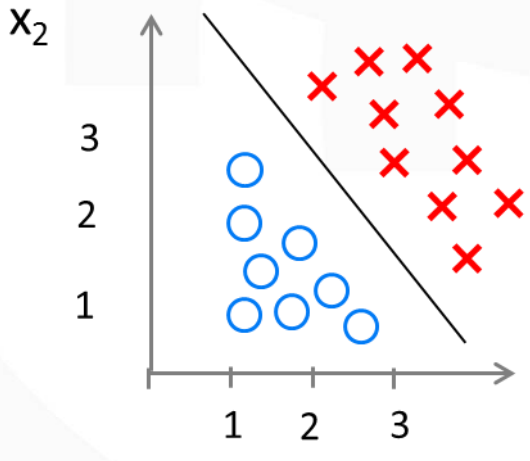
- Give your examples!

Cluster Definition

An ideal **cluster** can be defined as a set of points that is compact and isolated.

In reality, a cluster is a subjective entity whose significance and interpretation requires domain knowledge.

A cluster can be defined as the “centroid” of the collection of points belonging to it.



Cluster analysis techniques

Cluster analysis is a machine learning task, which often uses ANN methods.

The task is to find clusters (or centroids), which correctly represent real-world features (for example, responsible borrowers; typical customers, etc.)

Example: Cluster “The best borrower”:

- middle-aged
- Has a house
- Married with 2-3 children
- Annual income between $k100 - $k500

General algorithm for clustering (by Maheswari)

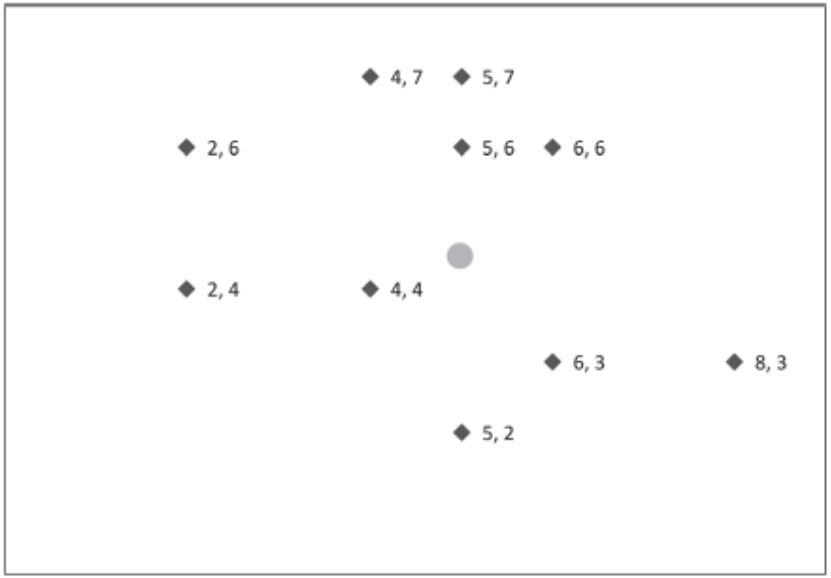
1. Pick an arbitrary number of groups/segments to be created.
2. Start with some initial randomly chosen center values for groups.
3. Classify instances to closest groups.
4. Compute new values for the group centers.
5. Repeat Steps 3 and 4 till groups converge.
6. If clusters are not satisfactory, go to Step 1 and pick a different number of groups/segments.

- Maheswari, p. 103/118

Clustering Example by Maheswari

X	Y
2	4
2	6
5	6
4	7
8	3
6	6
5	2
5	7
6	3
4	4

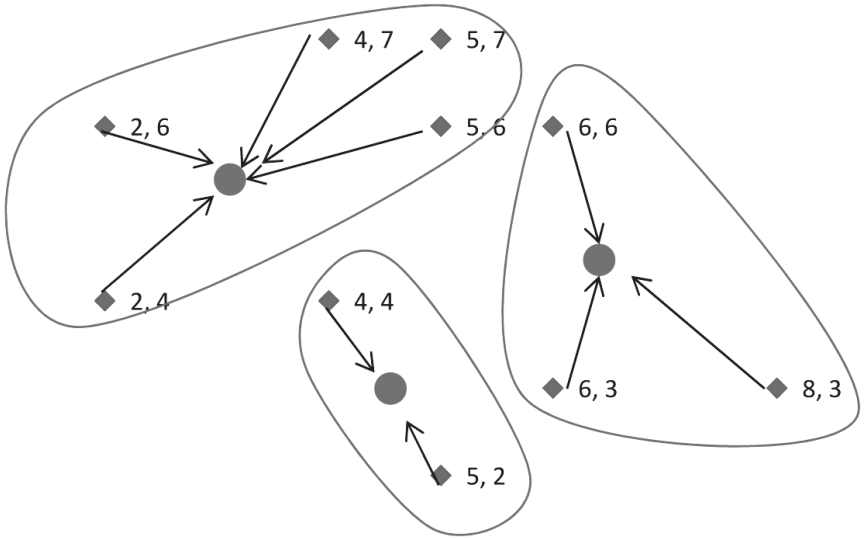
Initial Data sets



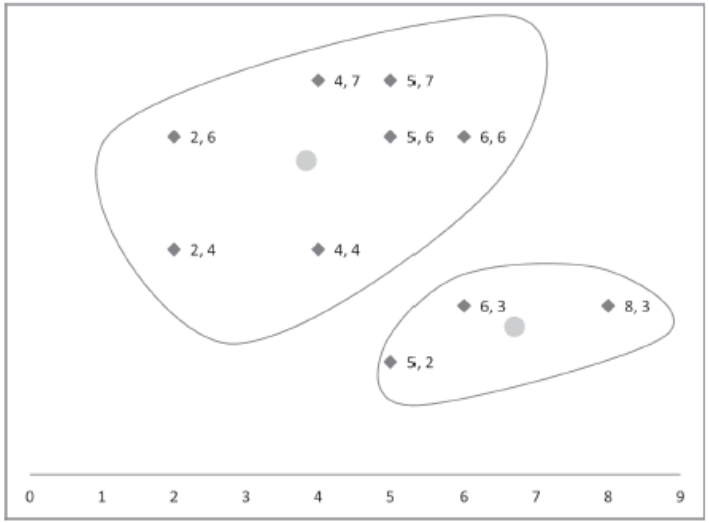
Initial step

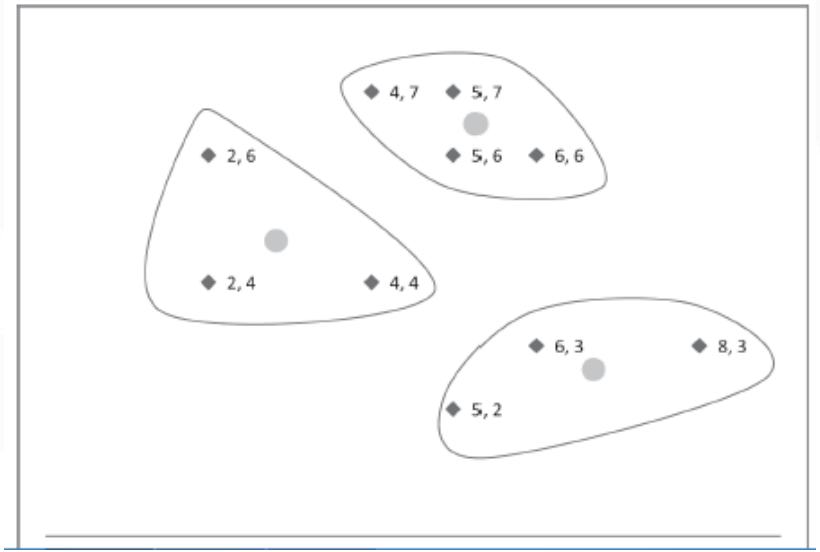
K-mean algorithm

- (K number of clusters, D list of data points)
- 1. Choose K number of random data points as initial centroids (cluster centers).
- 2. Repeat till cluster centers stabilize:
- a. Allocate each point in D to the nearest of K centroids.
- b. Compute centroid for the cluster using all points in the cluster.



Clustering Example by Maheswari (2)





The following steps in building clusters

Clustering advantages

1. K-means algorithm is simple, easy to understand, and easy to implement.

2. It is also efficient, in which the time taken to cluster K-means rises linearly with the number of data points.

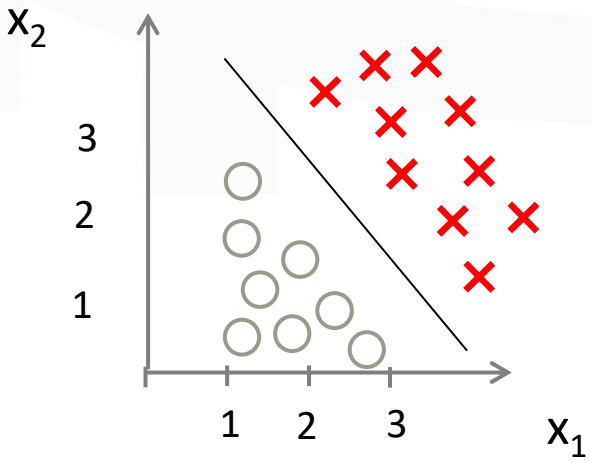
Session 12-2 Topics

- Non-Standard and Complex Methods:
  - Dummy variables
  - Fuzzy logic analysis

Dummy Variables

In regression analysis, a dummy variable (also known as indicator variable or just dummy) is one that takes a binary value (0 or 1) to indicate the absence or presence of some categorical effect that may be expected to shift the outcome

©Wikipedia



Purpose of Dummy Variables in Data Analysis

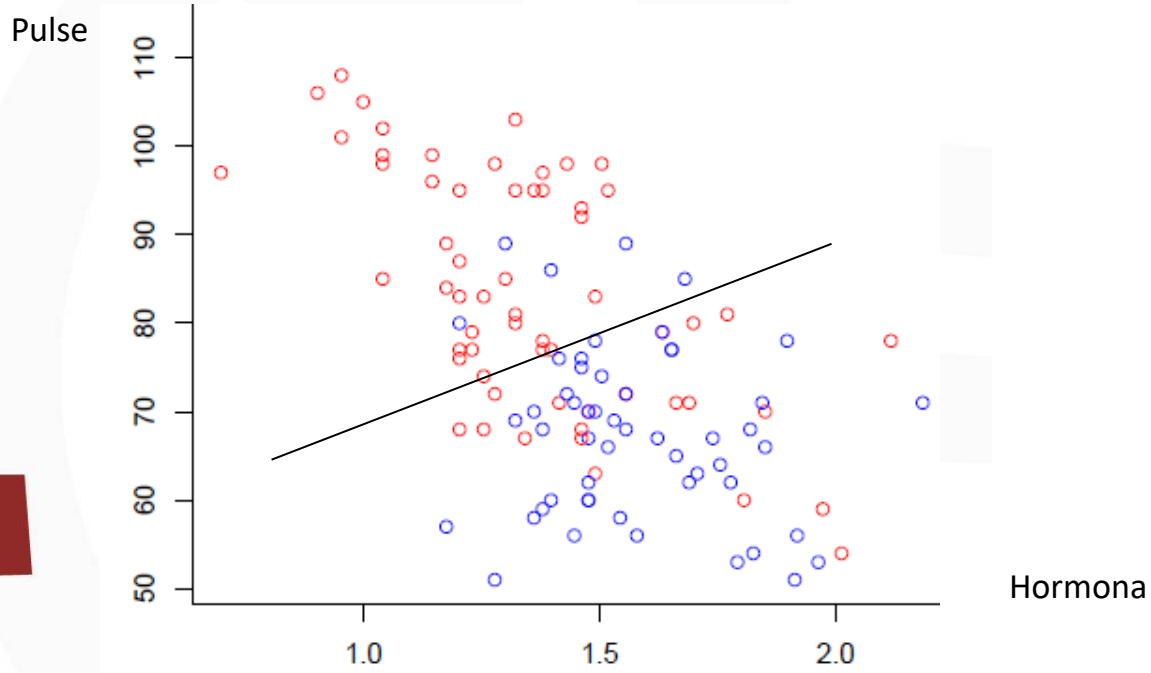
The primary purpose of dummy variables is to facilitate the analysis of categorical data within regression models.

By converting categorical variables into a series of binary variables, analysts can assess the impact of different categories on the dependent variable.

For example, if a dataset includes a categorical variable such as “Color” with three categories (Red, Blue, Green), three dummy variables would be created: one for Red, one for Blue, and one for Green.

This transformation allows the model to evaluate how each color influences the outcome while maintaining the integrity of the categorical information

Example – Regression (Support Vector) as a border between sets



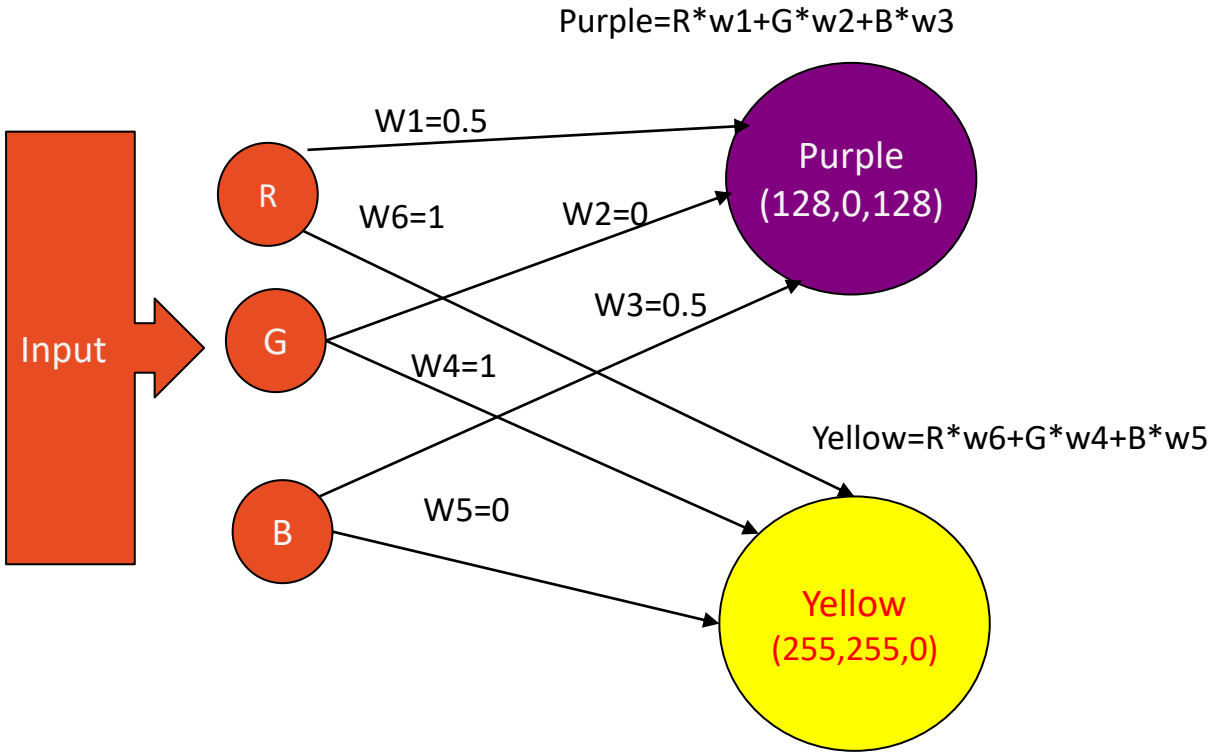
Task: Select patients with a certain disease

Dummy Variables in Machine Learning

In machine learning, dummy variables play a critical role in preparing data for algorithms that require numerical input.

Many machine learning models, such as linear regression, logistic regression, and support vector machines, necessitate the use of dummy variables to handle categorical data effectively.

Example – Defining Weights in Afrtificial Neuro Nets (ANN)



Example – Excel Analysis

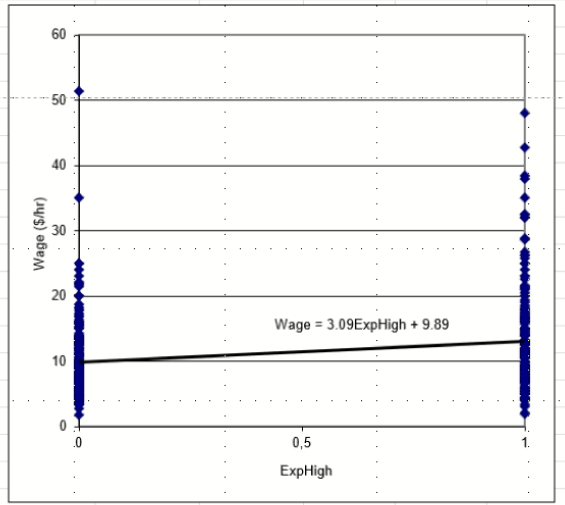
File Home Insert Draw Page Layout Formulas Data Review View Automate Developer Help Comments

L6 =IF(J6>44;1;0)

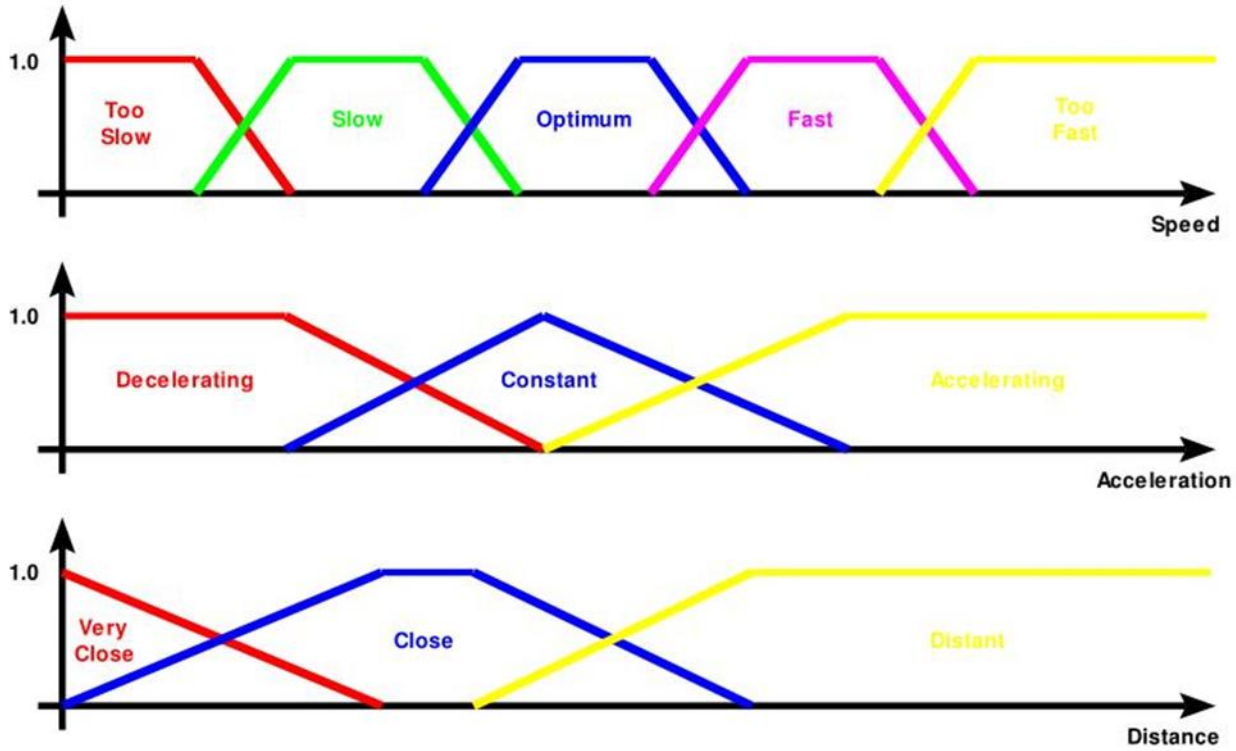
See CPS90ExpWorkers.xls in Chapter12/Excel Files for full documentation.

March 1990 CPS.
276 25-29 years olds
202 45-49 year olds

Education	Nonwhite	Hispanic	Female	Married	Age	Wage	ExpHigh
16	0	0	0	0	45	24,02	1
14	0	0	1	0	45	14,4	1
13	0	0	1	1	45	11,43	1
16	1	0	0	1	45	7	1
12	0	1	0	1	45	6,25	1
12	0	0	0	1	45	12,5	1
13	0	0	1	1	45	7	1
18	0	0	0	1	45	38,46	1
12	0	0	1	1	45	5,5	1
10	0	0	0	1	45	16	1
12	0	0	0	0	45	11,15	1
14	0	0	0	1	45	11,25	1
12	1	0	1	1	45	14	1
17	0	0	1	1	45	8	1
12	0	0	0	1	45	7,45	1
15	1	0	0	1	45	15	1
12	0	0	0	1	45	13	1
12	0	0	0	1	45	12	1
12	0	0	0	1	45	9	1
13	0	0	1	0	45	9,4	1
13	0	0	1	1	45	4,07	1
18	1	0	1	0	45	8,75	1
12	0	0	1	0	45	13,16	1
12	0	0	1	1	45	5	1
18	0	0	0	1	45	12,5	1
10	0	0	1	1	45	5,23	1
8	0	0	0	1	45	5	1
16	0	0	0	1	45	4,89	1
12	0	0	0	1	45	22,5	1



Fuzzy Logic Analysis



Fuzzy Logic Concept

The fundamental concept of Fuzzy Logic is the membership function, which defines the degree of membership of an input value to a certain set or category.

The membership function is a mapping from an input value to a membership degree between 0 and 1, where 0 represents non-membership and 1 represents full membership.

Fuzzy Logic is implemented using Fuzzy Rules, which are if-then statements that express the relationship between input variables and output variables in a fuzzy way.

The output of a Fuzzy Logic system is a fuzzy set, which is a set of membership degrees for each possible output value.

©https://www.geeksforgeeks.org/fuzzy-logic-introduction/

What is Fuzzy Control?

It is a technique to embody human-like thinkings into a control system.

It may not be designed to give accurate reasoning but it is designed to give acceptable reasoning.

It can emulate human deductive thinking, that is, the process people use to infer conclusions from what they know.

Any uncertainties can be easily dealt with the help of fuzzy logic.

©https://www.geeksforgeeks.org/fuzzy-logic-introduction/

Fuzzy Variable Example

- Fuzzy Variable α=age (young, adult, mature, elderly, old)
- X={0; 120}
- x(young) € A={15;50}
- x(mature) € A={30;65}
  - X (young)=30, μA(x)=0,8
  - x(mature)=30, μA(x)=0,3
  - x(young)=45, μA(x)=0,1
  - x(mature)=45, μA(x)=0,75

Example for Research

- «Respondent A» =

0.1	0.8	0.75	0.8	0.25
15	18	25	30	35

-

0.1	0.9	0.6	0.7	0.6	0.45
15	20	25	30	35	40

Compare Opinions in Fuzzy Logic

- «Respondent A» =

0.1	0.8	0.75	0.3	0.25
15	20	25	30	35

- «Respondent B» =

0.1	0.9	0.8	0.7	0.6
15	20	25	30	35

- Equality rank between A and B –

$I(s)=1-\max|(\mu A(x)-\mu B(x)|=1-0.4=0.6$

Unite Operation (Logical OR)

- «Respondent A» =

0.1	0.8	0.75	0.3	0.25
15	20	25	30	35

-

0.1	0.75	0.8	0.7	0.6	0.4
18	20	25	30	35	40

-

0.1	0.1	0.8	0.8	0.7	0.6	0.4
15	18	20	25	30	35	40

Intersection (Logical AND)

- «Respondent A» =

0.1	0.8	0.75	0.3	0.25
15	20	25	30	35

- «Respondent B» =

0.1	0.75	0.8	0.7	0.6	0.4
18	20	25	30	35	40

- $A \cap B$ =

0.75	0.75	0.3	0.25
20	25	30	35

Difference of fuzzy sets

DIFFERENCE: the resulting set consists of identical elements of the original sets.

- «Respondent A» =

0.1	0.8	0.75	0.3	0.25
15	20	25	30	35

- «Respondent B» =

0.1	0.8	0.75	0.7	0.6	0.4
18	20	25	30	35	40

- A - B =

0.8	0.75
20	25

Extended Operands

Not t	$1 - \mu_t(u)$
Very t	$(\mu_t(u))^2$
More or Less t	$\sqrt{\mu_t(u)}$
Super t	$(\mu_t(u))^3$

- X (young)=30, μA(x)=0,8
- X (very young)=30, μA(x)=0,64
- x(mature)=30, μA(x)=0,3
- X(not mature)=30, μA(x)=1-0,3=0,7
- x(more or less mature)=45, μA(x)=0,54

Advantages of Fuzzy Logic System

- This system can work with any type of inputs whether it is imprecise, distorted or noisy input information.
- The construction of Fuzzy Logic Systems is easy and understandable.
- Fuzzy logic comes with mathematical concepts of set theory and the reasoning of that is quite simple.
- It provides a very efficient solution to complex problems in all fields of life as it resembles human reasoning and decision-making.
- The algorithms can be described with little data, so little memory is required.

©https://www.geeksforgeeks.org/fuzzy-logic-introduction/

Challenges of Fuzzy Logic System

Many researchers proposed different ways to solve a given problem through fuzzy logic which leads to ambiguity. There is no systematic approach to solve a given problem through fuzzy logic.

Proof of its characteristics is difficult or impossible in most cases because every time we do not get a mathematical description of our approach.

As fuzzy logic works on precise as well as imprecise data so most of the time accuracy is compromised.

©https://www.geeksforgeeks.org/fuzzy-logic-introduction/

Application Areas for Fuzzy Logic System

- It is used in the aerospace field for altitude control of spacecraft and satellites.
- It has been used in the automotive system for speed control, traffic control.
- It is used for decision-making support systems and personal evaluation in the large company business.
- It has application in the chemical industry for controlling the pH, drying, chemical distillation process.
- Fuzzy logic is used in Natural language processing and various intensive applications in Artificial Intelligence.
- Fuzzy logic is extensively used in modern control systems such as expert systems.
- Fuzzy Logic is used with Neural Networks as it mimics how a person would make decisions, only much faster. It is done by Aggregation of data and changing it into more meaningful data by forming partial truths as Fuzzy sets.

©https://www.geeksforgeeks.org/fuzzy-logic-introduction/

Thank You!
Read the Recommended Readings
You're welcome with your discussions and questions in VLE!

Please note, that since the recordings are done, some Readings may become unavailable. Inform us immediately in VLE, so we can offer substitutions