



**BITS Pilani**

Pilani | Dubai | Goa | Hyderabad

# INTRODUCTION TO DATA SCIENCE

## MODULE # 5 : DATA AND DATA QUALITY

IDS Course Team

BITS Pilani

The instructor is gratefully acknowledging  
the authors who made their course  
materials freely available online.

# TABLE OF CONTENTS

---

- 1 DATA
- 2 DATA-SETS
- 3 DATA RETRIEVAL
- 4 DATA PREPARATION
- 5 DATA EXPLORATION
- 6 DATA QUALITY
- 7 OUTLIERS

- Data is a collection of data objects and their attributes.
- An attribute is a property or characteristic of an object.  
Examples: eye color of a person, temperature
- Attribute is also known as variable, field, characteristic, or feature.
- A collection of attributes describe an object.
- Object is also known as record, point, case, sample, entity, or instance.

Attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Objects



# TYPE OF DATA

---

- The type of data determines which tools and techniques can be used to analyze the data.

# QUALITY OF DATA

---

- Data quality issues

- ▶ Noise and outliers;
- ▶ Missing data
- ▶ Inconsistent data
- ▶ Duplicate data
- ▶ Data that is biased
- ▶ Data that is unrepresentative of the phenomenon or population that the data is supposed to describe.

# DATA QUALITY ISSUES

Find the issues in the given data.

Name	Age	Date of Birth	Course ID	CGPA
Amy	24	01-Jan-1995	CS 104	7.4
Ben	23	Dec-01-1996	CS 102	7.5
Cathy	25	01-Nov-1994		6.7
Diana	24	Oct-01-1995	CS 104	7.9
Ben	23	Dec-01-1996	CS 102	7.5
Eden	24		CS 103	87.5
Fischer		01-01-1959	CS 105	7.0

# PREPROCESSING ON DATA

---

- Improve Data Quality
- To better fit a specified data mining or machine learning technique or tool.
- Number of attributes in a data set is often reduced because many techniques are more effective when the data has a relatively small number of attributes.
- Data correction corrects the errors in the data. Data cleansing removes irrelevant data. Data transformation changes data from one format to another. Correction improves the data quality.



# ATTRIBUTE / FEATURE

- An attribute is a property or characteristic of an object.
  - ▶ eye color of a person, temperature
- Attribute is also known as variable, field, characteristic, or feature.
- The values used to represent an attribute may have properties that are not properties of the attribute itself.
  - ▶ Average age of an employee may have a meaning, whereas it makes no sense to talk about the average employee ID.

Attributes

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Objects

# ATTRIBUTE / FEATURE

- The type of an attribute should tell us what properties of the attribute are reflected in the values used to measure it.
  - ▶ For the age attribute, the properties of the integers used to represent age are very much the properties of the attribute. Even so, ages have a maximum while integers do not.
  - ▶ The ID attribute is distinct. The only valid operation for employee IDs is to test whether they are equal.

Attributes

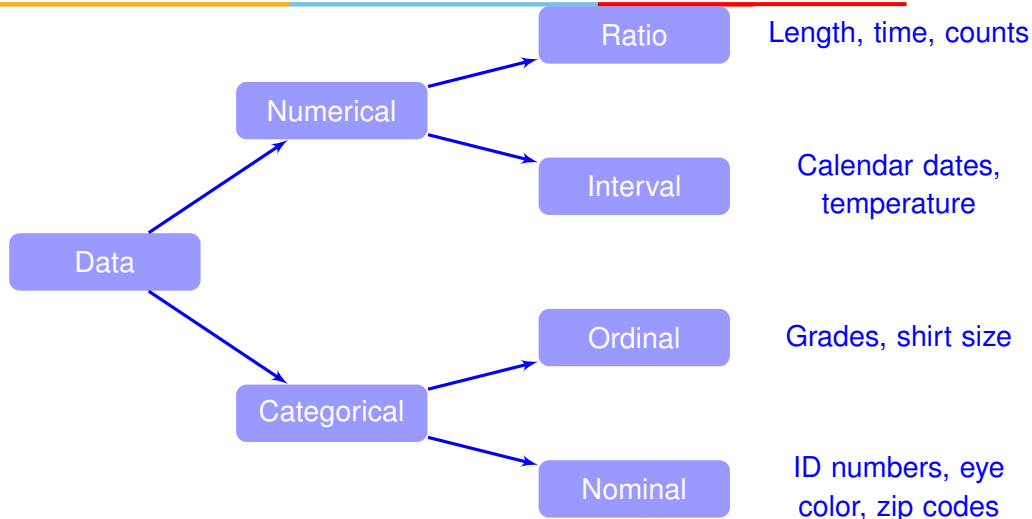
<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Objects

# PROPERTIES OF ATTRIBUTES

- Specify the type of an attribute by identifying the properties of numbers that correspond to underlying properties of the attribute.
- Properties include
  - ▶ Distinctiveness  $=, \neq$
  - ▶ Order  $<, >, \geq, \leq$
  - ▶ Addition  $+, -$
  - ▶ Multiplication  $*, /$
- Based on these properties, we define four types of attributes: nominal, ordinal, interval, and ratio.
- Each attribute type possesses all of the properties and operations of the attribute types above it.

# TYPES OF ATTRIBUTES



# TYPES OF ATTRIBUTES

Attribute Type		Description	Examples	Operations
Categorical (Qualitative)	Nominal	The values of a nominal attribute are just different names; i.e., nominal values provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, gender	mode, entropy, contingency correlation, $\chi^2$ test
	Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {good, better, best}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric (Quantitative)	Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, <i>t</i> and <i>F</i> tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

# TYPES OF ATTRIBUTES EXAMPLE

Identify the types of attributes in the given data.

ID	Age	Gender	Course ID	CGPA	Grade
19001	24	Female	CS 104	7.4	Good
19002	23	Male	CS 102	7.5	Good
19003	25	Female	CS 103	6.7	Fair
19004	24	Female	CS 104	7.9	Good
19005	23	Male	CS 102	7.5	Good
19006	24	Female	CS 103	8.5	Excellent
19007	26	Male	CS 105	7.0	Good

# TYPES OF ATTRIBUTES EXAMPLE

Identify the types of attributes in the given data.

ID	Age	Gender	Course ID	CGPA	Grade
19001	24	Female	CS 104	7.4	Good
19002	23	Male	CS 102	7.5	Good
19003	25	Female	CS 103	6.7	Fair
19004	24	Female	CS 104	7.9	Good
19005	23	Male	CS 102	7.5	Good
19006	24	Female	CS 103	8.5	Excellent
19007	26	Male	CS 105	7.0	Good
Nominal	Ratio	Nominal	Nominal	Ratio	Ordinal

# ATTRIBUTES AND TRANSFORMATIONS

Attribute Type		Transformation	Comment
Categorical (Qualitative)	Nominal	Any one-to-one mapping, e.g., a permutation of values	If all employee ID numbers are reassigned, it will not make any difference.
	Ordinal	An order-preserving change of values, i.e., $new\_value = f(old\_value)$ , where $f$ is a monotonic function.	An attribute encompassing the notion of good, better, best can be represented equally well by the values {1, 2, 3} or by {0.5, 1, 10}.
Numeric (Quantitative)	Interval	$new\_value = a * old\_value + b$ , $a$ and $b$ constants.	The Fahrenheit and Celsius temperature scales differ in the location of their zero value and the size of a degree (unit).
	Ratio	$new\_value = a * old\_value$	Length can be measured in meters or feet.



# ATTRIBUTES BY THE NUMBER OF VALUES

- Discrete Attribute

- ▶ only a finite or countable infinite set of values.
- ▶ zip codes, counts, or the set of words in a collection of documents
- ▶ Often represented as integer variables.
- ▶ Note: binary attributes are a special case of discrete attributes

- Continuous Attribute

- ▶ Real numbers as attribute values.
- ▶ temperature, height, or weight
- ▶ Continuous attributes are typically represented as floating-point variables.

- Asymmetric Attribute

- ▶ only presence a non-zero attribute value-is considered.
- ▶ For a specific student, an attribute has a value of 1 if the student took the course associated with that attribute and a value of 0 otherwise
- ▶ Asymmetric binary attributes.

# TYPES OF ATTRIBUTES EXAMPLE

Identify the types of attributes in the given data.

ID	Age	Gender	Course ID	CGPA	Grade
19001	24	Female	CS 104	7.4	Good
19002	23	Male	CS 102	7.5	Good
19003	25	Female	CS 103	6.7	Fair
19004	24	Female	CS 104	7.9	Good
19005	23	Male	CS 102	7.5	Good
19006	24	Female	CS 103	8.5	Excellent
19007	26	Male	CS 105	7.0	Good
Discrete	Continuous	Discrete	Discrete	Continuous	Discrete

# DATA FORMATS

- Record data
  - ▶ Transaction or Market Basket data – set of items
  - ▶ Data Matrix – record data with only numeric attributes.
  - ▶ Sparse Data Matrix – binary asymmetric data. 0/1 entries.
  - ▶ Document term matrix
- Graph data
  - ▶ Data with relationships among objects – Web pages
  - ▶ Data with objects as graphs – chemical compound
- Ordered data
  - ▶ Sequential data or temporal data – Record data + time.
  - ▶ Sequence data – genome representation
  - ▶ Time series data – temporal autocorrelation
  - ▶ Spatial data – spatial autocorrelation

# RECORD DATA EXAMPLE

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

(a) Record data.

TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

(b) Transaction data.

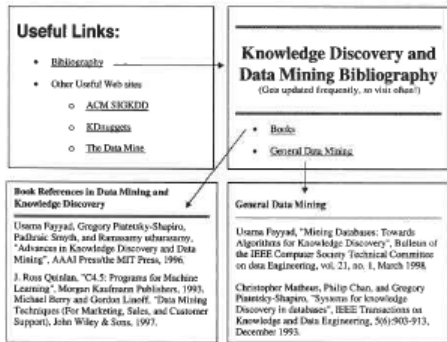
Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

(c) Data matrix.

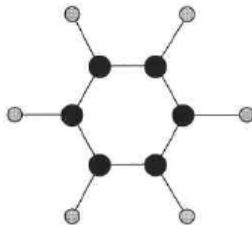
	team	coach	play	ball	score	game	wfr	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

(d) Document-term matrix.

# GRAPH DATA EXAMPLE



(a) Linked Web pages.



(b) Benzene molecule.

# ORDERED DATA EXAMPLE

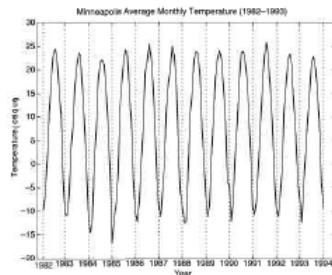
Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

(a) Sequential transaction data.

GGTTCCGCCTTCAGCCCCGCGCC  
 CGCAGGGCCCCGCCCCGCGCCGTC  
 GAGAAGGGCCCCGCCTGGCGGGCG  
 GGGGGAGGCGGGGCCGCCGAGC  
 CCAACCGAGTCCGACCAGGTGCC  
 CCCTCTGCTCGGCCTAGACCTGA  
 GCTCATTAGGCGGCAGCGGACAG  
 GCCAAGTAGAACACGCGAAGCGC  
 TGGGCTGCCTGCTGCGACCAGGG

(b) Genomic sequence data.



(c) Temperature time series.

# TABLE OF CONTENTS

---

- 1 DATA
- 2 DATA-SETS
- 3 DATA RETRIEVAL
- 4 DATA PREPARATION
- 5 DATA EXPLORATION
- 6 DATA QUALITY
- 7 OUTLIERS

# TYPES OF DATA-SETS

- ❶ Structured data
  - ▶ Data containing a defined data type, format and structure.
  - ▶ Example: transaction data, online analytical processing , OLAP data cubes, traditional RDBMS, CSV file and spreadsheets.
- ❷ Semi structured data
  - ▶ Textual data file with discernible pattern that enables parsing
  - ▶ Example: XML data file
- ❸ Quasi structured data
  - ▶ Textual data with erratic data format that can be formatted with effort, tools and time
  - ▶ Example: Web click-stream data
- ❹ Unstructured data
  - ▶ Data that has no inherent structure.
  - ▶ Example: text document, PDF, images and video, email



# TYPES OF DATA-SETS

---

## 5 Natural Language

- ▶ Entity recognition, topic recognition, summarization, text completion, and sentiment analysis
- ▶ Models trained in one domain don't generalize well to other domains.

## 6 Machine generated data

- ▶ Machine-generated data is automatically created by a computer, process, application, or other machine without human intervention.
- ▶ High volume and speed.
- ▶ Examples web server logs, call detail records, network event logs, and telemetry

# TYPES OF DATA-SETS

---

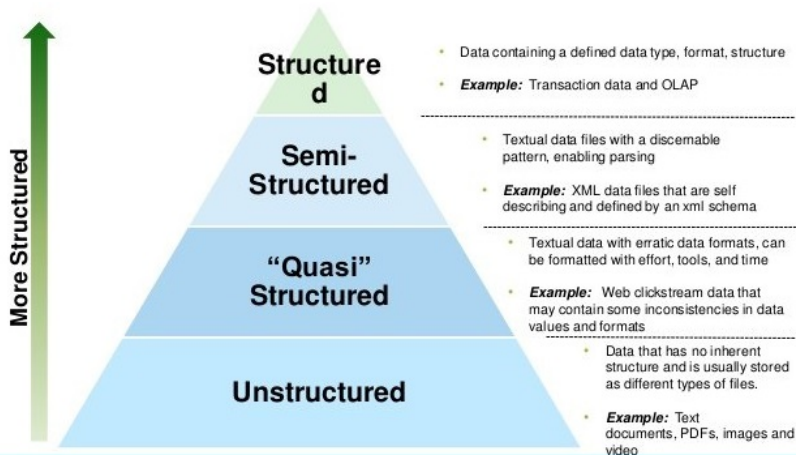
## 6 Graph-based or network data

- ▶ Data can be shown in a graph.
- ▶ A graph is a mathematical structure to model pair-wise relationships between objects.
- ▶ Graph or network data focuses on the relationship or adjacency of objects.
- ▶ Graph databases with specialized query languages such as SPARQL.
- ▶ Example: represent social networks

## 7 Streaming data

- ▶ The data flows into the system when an event happens instead of being loaded into a data store in a batch.
- ▶ Example: live sports or music events, stock market.

# TYPES OF DATA-SETS



EMC<sup>2</sup> PROVEN PROFESSIONAL

# CHARACTERISTICS OF DATA-SETS

## 1 Dimensionality

- ▶ Number of attributes
- ▶ Curse of Dimensionality – the difficulties associated with analyzing high-dimensional data
- ▶ Dimensionality reduction techniques

## 2 Sparsity

- ▶ For some data sets, such as those with asymmetric features, most attributes of an object have values of 0; in many cases, fewer than 1% of the entries are non-zero.
- ▶ Advantage because usually only the non-zero values need to be stored and manipulated.

## 3 Resolution

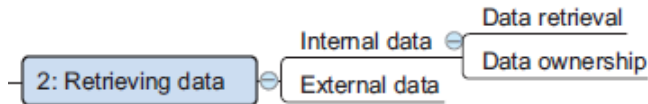
- ▶ The patterns in the data depend on the level of resolution.
- ▶ If the resolution is too fine, a pattern may not be visible or may be buried in noise; if the resolution is too coarse, the pattern may disappear.

# TABLE OF CONTENTS

---

- 1 DATA
- 2 DATA-SETS
- 3 DATA RETRIEVAL**
- 4 DATA PREPARATION
- 5 DATA EXPLORATION
- 6 DATA QUALITY
- 7 OUTLIERS

# RETRIEVING DATA



- Already collected and stored the data in the organization
- Look outside the organization for high-quality data available for public and commercial use. (open-data providers)
- Quality Check while Retrieving Data
  - ▶ Check to see if data is equal to the data in the source document.
  - ▶ Check for the right data types.

# RETRIEVING DATA

---

- Data Storage
  - ▶ Text files
  - ▶ Database tables
  - ▶ Data marts
  - ▶ Data warehouses
  - ▶ Data lakes ( raw data)

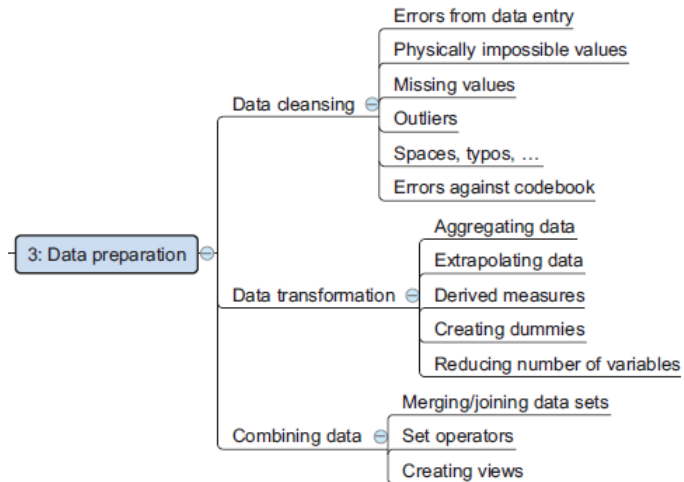
# TABLE OF CONTENTS

---

- 1 DATA
- 2 DATA-SETS
- 3 DATA RETRIEVAL
- 4 DATA PREPARATION**
- 5 DATA EXPLORATION
- 6 DATA QUALITY
- 7 OUTLIERS



# DATA PREPARATION



# DATA CLEANSING

- Focuses on removing errors in your data so your data becomes a true and consistent representation of the processes it originates from.
- Two types of errors
  - ▶ Interpretation error
    - ★ *Age* < 100
    - ★ Height of a person is less than 7feet.
    - ★ Price is positive.
  - ▶ Inconsistencies between data sources or against your company's standardized values.
    - ★ Female and F
    - ★ Feet and meter
    - ★ Dollars and Pounds

# DATA CLEANSING

Error description	Possible solution
<i>Errors pointing to false values within one data set</i>	
Mistakes during data entry	Manual overrules
Redundant white space	Use string functions
Impossible values	Manual overrules
Missing values	Remove observation or value
Outliers	Validate and, if erroneous, treat as missing value (remove or insert)
<i>Errors pointing to inconsistencies between data sets</i>	
Deviations from a code book	Match on keys or else use manual overrules
Different units of measurement	Recalculate
Different levels of aggregation	Bring to same level of measurement by aggregation or extrapolation

# DATA CLEANSING

- Errors from data entry
  - ▶ Cause
    - ★ Typos
    - ★ Errors due to lack of concentration
    - ★ Machine or hardware failure
  - ▶ Detection
    - ★ Frequency table
  - ▶ Correction
    - ★ Simple assignment statements
    - ★ If-then-else rules
- White-spaces and typos
  - ▶ Remove leading and trailing white-spaces.
  - ▶ Change case of the alphabets from upper to lower.

- Physically impossible values

- ▶ Examples

- ★  $Age < 100$
    - ★ Height of a person is less than 7feet.
    - ★ Price is positive.

- ▶ If-then-else rules

- Outliers

- ▶ Use visualization techniques like box plots or scatter plots.
  - ▶ Use statistical summary with minimum and maximum values.

# DATA CLEANSING

- Missing values

Technique	Advantage	Disadvantage
Omit the values	Easy to perform	You lose the information from an observation
Set value to <code>null</code>	Easy to perform	Not every modeling technique and/or implementation can handle <code>null</code> values
Impute a static value such as 0 or the mean	Easy to perform You don't lose information from the other variables in the observation	Can lead to false estimations from a model
Impute a value from an estimated or theoretical distribution	Does not disturb the model as much	Harder to execute You make data assumptions
Modeling the value (nondependent)	Does not disturb the model too much	Can lead to too much confidence in the model Can artificially raise dependence among the variables Harder to execute

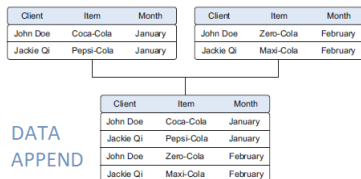
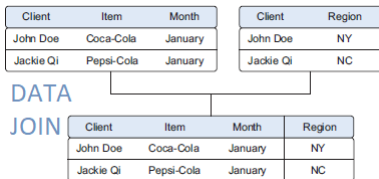
# DATA CLEANSING

- Deviations from code-book
  - ▶ A code book is a description of your data. It contains things such as the number of variables per observation, the number of observations, and what each encoding within a variable means.
  - ▶ Discrepancies between the code-book and the data should be corrected.
- Different units of measurement
  - ▶ Pay attention to the respective units of measurement.
  - ▶ Simple conversion can rectify.
- Different levels of aggregation
  - ▶ Data set containing data per week versus one containing data per work week.
  - ▶ Data summarization will fix it.

# COMBINING DATA



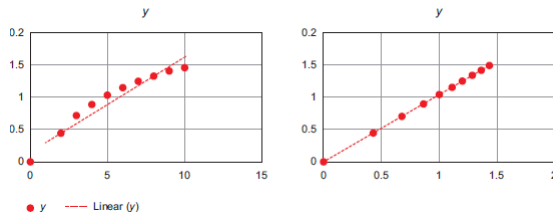
- Two operations to combine information from different data sets.
  - ▶ Joining
    - ★ Enriching an observation from one table with information from another table.
    - ★ Requires primary keys or candidate keys.
    - ★ Use views to virtually combine data.
  - ▶ Appending or stacking
    - ★ Adding the observations of one table to those of another table.





# TRANSFORMING DATA

- Applying mathematical transformation to the input variable.
  - For a relationship of the form,  $y = ae^{bx}$  transforming  $x$  to  $\log x$  makes the relationship between  $x$  and  $y$  linear.

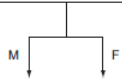


- Reducing number of variables.
- Combining two variables into a new variable.

# TRANSFORMING DATA

- Turning variables into dummies.
  - ▶ Dummy variables can only take two values: true(1) or false(0).
  - ▶ Create separate columns for the classes stored in one variable and indicate it with 1 if the class is present and 0 otherwise.

Customer	Year	Gender	Sales
1	2015	F	10
2	2015	M	8
1	2016	F	11
3	2016	M	12
4	2017	F	14
3	2017	M	13



Customer	Year	Sales	Male	Female
1	2015	10	0	1
1	2016	11	0	1
2	2015	8	1	0
3	2016	12	1	0
3	2017	13	1	0
4	2017	14	0	1

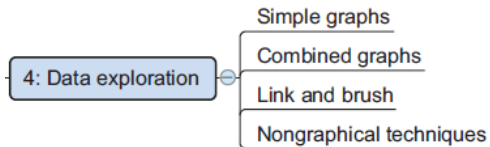
# TABLE OF CONTENTS

---

- 1 DATA
- 2 DATA-SETS
- 3 DATA RETRIEVAL
- 4 DATA PREPARATION
- 5 DATA EXPLORATION**
- 6 DATA QUALITY
- 7 OUTLIERS

# EXPLORATORY DATA ANALYSIS (EDA)

- Use graphical techniques to gain an understanding of the data and the interactions between variables.
- Look at what can be learned from the data.
- Statistical properties like distribution of data, correlation.
- Discover outliers.



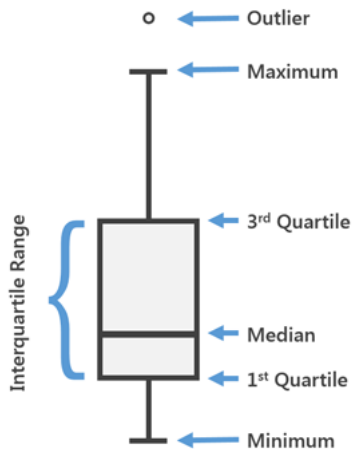


# EXPLORATORY DATA ANALYSIS (EDA)

- Boxplot – can show the maximum, minimum, median, and other characterizing measures at the same time.
- Histogram – In a histogram a variable is cut into discrete categories and the number of occurrences in each category are summed up and shown in the graph.
- Pareto diagram – is a combination of the values and a cumulative distribution.
- Tabulation
- Clustering and other modeling techniques can also be a part of exploratory analysis.

# BOXPLOT

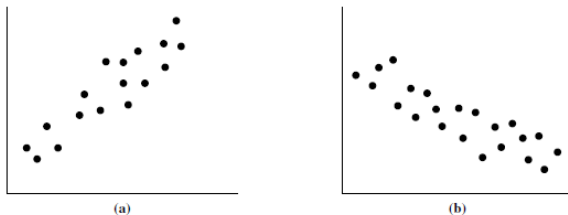
- A boxplot incorporates the five-number summary.
- The ends of the box are at the quartiles.
- The box length is the interquartile range.
- The median is marked by a line within the box.
- The whiskers outside the box extend to the Minimum and Maximum observations.



# SCATTERPLOT



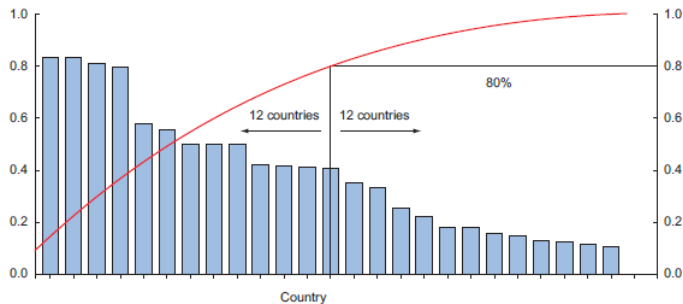
- Determine if there appears to be a relationship, pattern, or trend between two numeric attributes.
- Provide a visualization of bi-variate data to see clusters of points and outliers, or correlation relationships.



Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.

# PARETO DIAGRAM

- Pareto diagram is a combination of the values and a cumulative distribution.





# TABLE OF CONTENTS

---

- 1 DATA
- 2 DATA-SETS
- 3 DATA RETRIEVAL
- 4 DATA PREPARATION
- 5 DATA EXPLORATION
- 6 DATA QUALITY
- 7 OUTLIERS

# DATA QUALITY INDEX

innovate

achieve

lead

## UNIQUENESS

- Existence of unique values for a specific data attribute within a table
- **Example:** Data attribute which has duplicated values will not have the highest score on uniqueness dimension

## CONSISTENCY

- Logical coherence within data of a system that free them from contradiction
- **Example:** 'Order Fulfilment Date' should be after the 'Order Creation Date'

## INTEGRITY

- Existence of data values in reference table(s) from different system(s)
- **Example:** 'Product ID' values should exist in the Product reference table



## COMPLETENESS

- Existence of values in a specific data attribute (data field)
- **Example:** Data attribute with missing values is not complete

## TIMELINESS

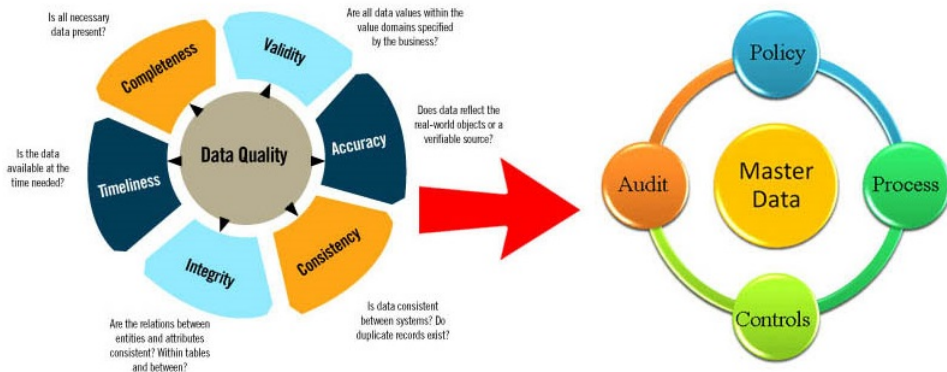
- Degree to which data is representative of current business conditions (updated and available)
- **Example:** A plan price change not updated on the day it was issued creates a breach of timeliness

## CONFORMITY

- Data are valid if it conforms to the syntax (format, type, range) of its definition
- **Example:** 'Landline Number' should be numeric with 8 digits

<https://www.deltapartnersgroup.com/>

# DATA QUALITY INDEX



<http://www.dataintegration.ninja>

# IMPACT OF MISSING DATA

---

- Missing data imputation may distort variable distribution.
- Affect the performance of Machine Learning Models.
- Incompatible in Scikit Learn library of Python.

# MISSING DATA MECHANISMS

---

- Understanding the mechanism of missing data will help us choose appropriate imputation method.
  - ▶ Missing Completely At Random (MCAR)
  - ▶ Missing At Random (MAR)
  - ▶ Not Missing At Random (NMAR)

# MISSING COMPLETELY AT RANDOM (MCAR)

- The probability of missing is same for all the observations.
- There is no relationship between the missing values and any other values in the dataset.
- Removing such missing values will not effect the inferences made.

Gender	Age
Male	42
Male	NA
Male	24
Male	NA
Male	36
Male	57
Female	32
Female	NA
Female	NA
Female	18
Female	NA
Female	23

# MISSING AT RANDOM (MAR)

- The probability of a missing values depends on available information i.e it depends on other variables in the dataset.

Gender	Age
Male	42
Male	NA
Male	24
Male	NA
Male	36
Male	57
Female	32
Female	NA
Female	NA
Female	18
Female	NA
Female	23

33% males

50% Females

# NOT MISSING AT RANDOM (NMAR)

- The missing values exist as an indication of a certain class.
- Depression = yes has more missing values. Hence choose imputation technique appropriately.

No of clinical visits	No of sports classes attended	Depression
1	NA	Yes
NA	NA	Yes
NA	0	Yes
4	2	Yes
NA	1	Yes
3	NA	Yes
0	0	No
NA	5	No
1	2	No
1	1	No
2	1	No
NA	2	No



# IMPUTATION TECHNIQUES

## Categorical Attributes

- Imputation by most frequent category
- Treating NA as an additional category

## Numerical Attributes

- Mean / Median Imputation
- Random Sampling Imputation
- Adding a new variable to indicate missingness
- Imputation of NA by values at the end of distribution
- Imputation of NA by arbitrary values

# IMPUTATION BY MOST FREQUENT CATEGORY

- Used when NMAR.
- Use mode.

Gender	Age
Male	42
Male	43
Male	24
NA	63
Male	36
Male	57
Female	32
NA	33
Female	45
Female	18
NA	55
Female	23

# TREATING NA AS AN ADDITIONAL CATEGORY

- Encode as unique category as unknown or missing
- Use mode to fill missing value.

Gender	Gender_new	Gender_new_value
Male	Male	Male
Male	Male	Male
Male	Male	Male
NA	Missing	Male
Male	Male	Male
Male	Male	Male
Female	Female	Female
NA	Missing	Male
Female	Female	Female
Female	Female	Female
NA	Missing	Male
Female	Female	Female

# MEAN / MEDIAN IMPUTATION

- Used when MCAR / MAR.
- Assumes that the feature follows normal distribution
- Advantages
  - ▶ Easy to implement
  - ▶ Faster way of obtaining complete dataset
- Disadvantages
  - ▶ Mean imputation reduces the variance of the imputed variables.
  - ▶ Mean imputation does not preserve relationships between variables such as correlations.

Gender	Age
Male	42
Male	NA
Male	24
Male	NA
Male	36
Male	57
Female	32
Female	NA
Female	NA
Female	18
Female	NA
Female	23

# RANDOM SAMPLING IMPUTATION

- Used when MCAR / MAR.
- Aim to preserve the statistical parameters of the feature.
- Number of random samples are at least as many as missing values.

Gender	Age
Male	42
Male	NA
Male	24
Male	NA
Male	36
Male	57
Female	32
Female	NA
Female	NA
Female	18
Female	NA
Female	23

# ADDING A NEW FEATURE TO INDICATE MISSINGNESS

- Used when MCAR / MAR.
- Use carefully, as the entire dataset may confuse ML models.

Gender	Age	Age Missing?
Male	42	0
Male	NA	1
Male	24	0
Male	NA	1
Male	36	0
Male	57	0
Female	32	0
Female	NA	1
Female	NA	1
Female	18	1
Female	NA	0
Female	23	1

# IMPUTATION BY VALUES AT THE END OF DISTRIBUTION

- Used when NMAR.
- Values at the tail of the normal distribution is used for imputation.
- To be verified by domain expert.

Gender	Age	Age Missing?
Male	42	0
Male	NA	1
Male	24	0
Male	NA	1
Male	36	0
Male	57	0
Female	32	0
Female	NA	1
Female	NA	1
Female	18	1
Female	NA	0
Female	23	1

# IMPUTATION BY ARBITRARY VALUES

- Used when NMAR.
- Use any value except mean/median value .

Gender	Age
Male	42
Male	NA
Male	24
Male	NA
Male	36
Male	57
Female	32
Female	NA
Female	NA
Female	18
Female	NA
Female	23



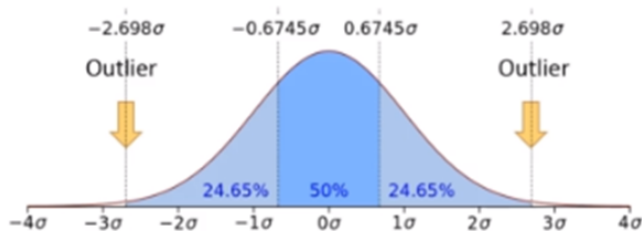
# TABLE OF CONTENTS

---

- 1 DATA
- 2 DATA-SETS
- 3 DATA RETRIEVAL
- 4 DATA PREPARATION
- 5 DATA EXPLORATION
- 6 DATA QUALITY
- 7 OUTLIERS**

- Data objects with behaviors that are very different from expectation are called outliers or anomalies.
- Outliers can significantly skew the distribution of your data.
- Outliers can be identified using summary statistics and plots of the data.
- Algorithms like Linear Regression, K-Nearest Neighbor, Adaboost are sensitive to noise.

# OUTLIER DETECTION USING NORMAL DISTRIBUTION



- 99% of the observations of a variable following a normal distribution lie within  $\mu \pm 3\sigma$ .

# OUTLIER DETECTION IN SKEWED DATA



- Calculate the quantiles and the Inter-quantile range(IQR).
- $IQR = 75^{th} \text{Quantile} - 25^{th} \text{Quantile}$
- $Upperlimit = 75^{th} \text{Quantile} + IQR \times 1.5$
- $Lowerlimit = 25^{th} \text{Quantile} - IQR \times 1.5$
- Data objects that lie outside the range  $[Lowerlimit, Upperlimit]$  are considered as outliers.

- Introduction to Data Mining, by Tan, Steinbach and Vipin Kumar (T3)
- The Art of Data Science by Roger D Peng and Elizabeth Matsui (R1)
- Introducing Data Science by Cielen, Meysman and Ali
- <https://www.deltapartnersgroup.com/managing-data-quality-optimize-value-extraction>
- <http://www.dataintegration.ninja/relationship-between-data-quality-and-master-data-management/>

THANK YOU