Data Mining:
Concepts and Techniques

Data Mining Preprocessing

Presented by

Dr. Siddique Ibrahim

SCOPE

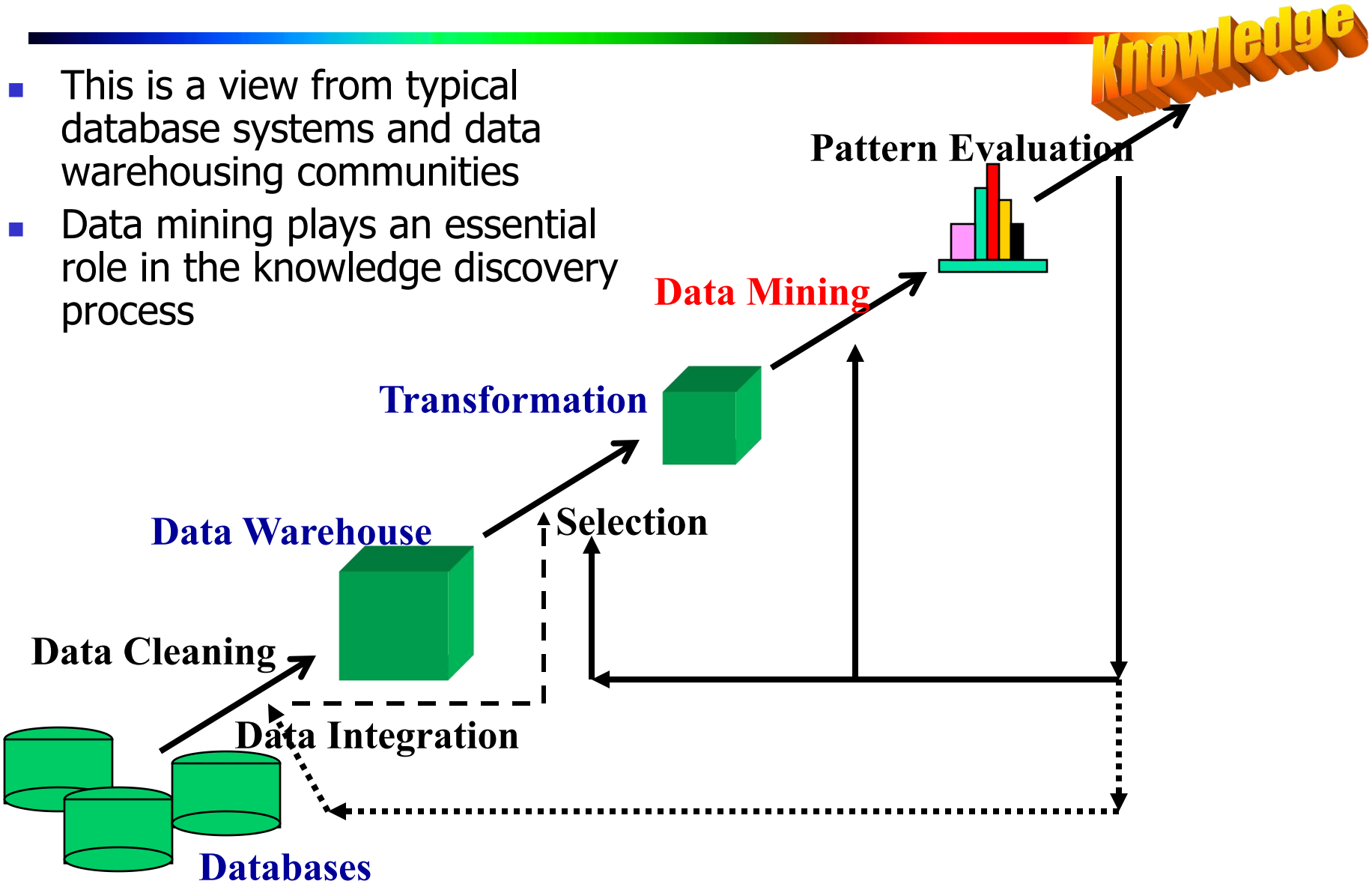
VIT-AP University

Agenta

- Data Mining Steps
- Data Preprocessing

Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



Steps in Data Mining

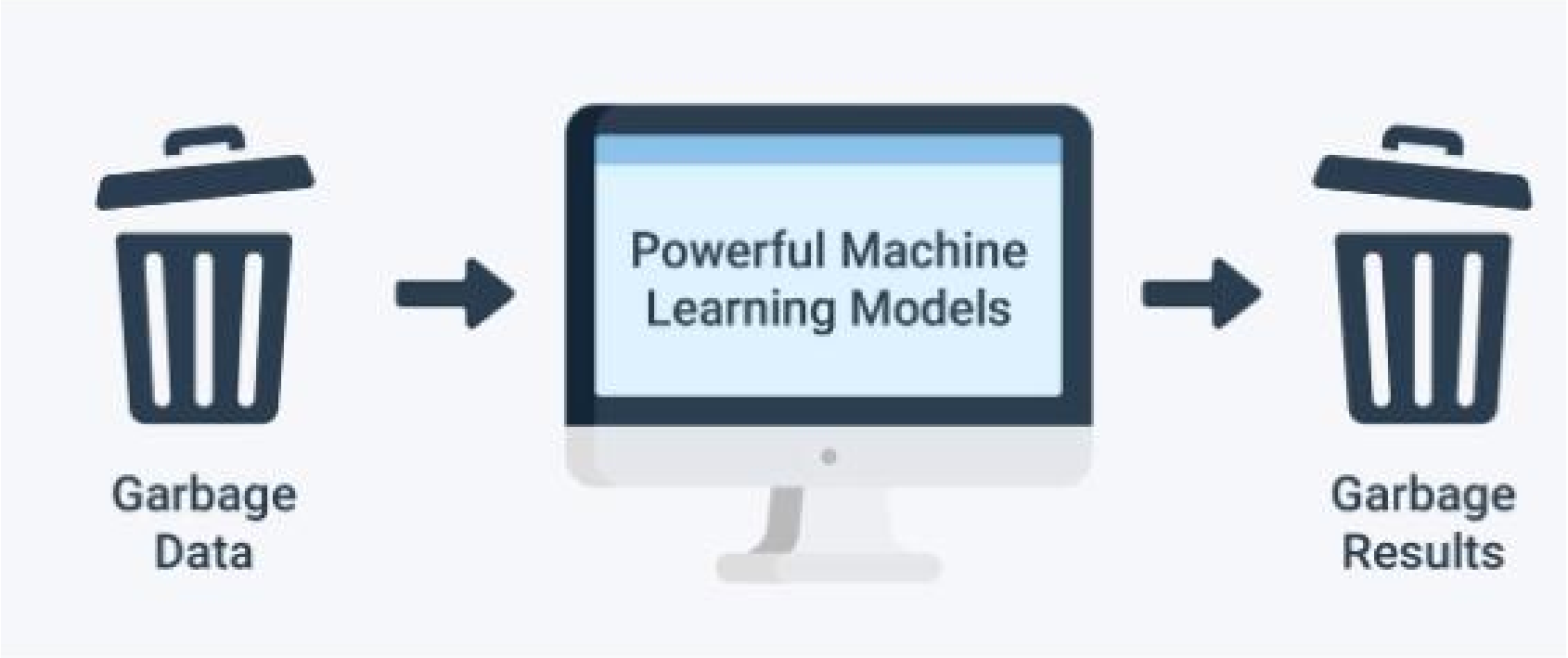
- 1. Data Cleaning - To remove noise and inconsistent data
- 2. Data Integration - Where multiple data sources may be combined
- 3. Data Selection - Where data relevant to the analysis task are retrieved from the database
- 4. Data Transformation - Where data are transformed or consolidated into forms appropriate for mining by performing summery or aggregation operations, for instance
- 5. Data Mining - An essential process where intelligent methods are applied in order to extract data patterns
- 6. Pattern Evaluation - To identify the truly interesting patterns representing knowledge based on some interestingness measures
- 7. Knowledge Presentation - Where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

What Is Data Preprocessing?

- Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning.
- Raw, real-world data in the form of text, images, video, etc., is messy. Not only may it contain errors and inconsistencies, but it is often incomplete, and doesn't have a regular, uniform design.

Data Preprocessing Importance

- When using data sets to train machine learning models, you'll often hear the phrase "garbage in, garbage out"
- This means that if you use bad or "dirty" data to train your model, you'll end up with a bad, improperly trained model that won't actually be relevant to your analysis.
- Good, preprocessed data is even more important than the most powerful algorithms


Powerful Machine
Learning Models
Garbage
Data
Garbage
Results

1. Goal-Setting and Application Understanding

- This is the first step in the process and requires prior understanding and knowledge of the field to be applied in. This is where we decide how the transformed data and the patterns arrived at by data mining will be used to extract knowledge.
- This premise is extremely important which, if set wrong, can lead to false interpretations and negative impacts on the end-user.

2. Data Selection and Integration

- After setting the goals and objectives, the data collected needs to be selected and segregated into meaningful sets based on availability, accessibility importance and quality.
- These parameters are critical for data mining because they make the base for it and will affect what kinds of data models are formed.

Data Integration

- Multiple heterogeneous sources of data are combined into single dataset.
- Two types of data integration:
- 1) Tight coupling: Data is combined together into a physical location. (You can't access the previous data)
- 2) Loosly coupled: Data not integrated)
- Only an interface is created and data combined through the interface and also accessed through interface.
- Data remain in actual database only

3. Data Cleaning and Preprocessing

- This step involves searching for missing data and removing noisy, redundant and low-quality data from the data set in order to improve the reliability of the data and its effectiveness.

- Data cleaning is the process of adding missing data and correcting, repairing, or removing incorrect or irrelevant data from a data set.

- Certain algorithms are used for searching and eliminating unwanted data based on attributes specific to the application.

Reason for noise in data

- Data in the Real World Is Dirty

- Lots of potentially incorrect data
  - Faulty instruments
  - Human or computer error
  - Transmission error

Some examples for noisy data

- **Incomplete:** lacking attribute values, lacking certain attributes of interest, or containing only aggregate data

Ex: Occupation=“ ” (missing data)

- **Noisy:** containing noise, errors, or outliers

Ex: Salary=“−10” (an error)

Some examples for noisy data

- **Inconsistent:** containing discrepancies in codes or names, discrepancy between duplicate records

Ex:

1. Age="42", Birthday="03/07/2010"
2. Was rating "1, 2, 3", now rating "A, B, C"

- **Intentional:** disguised missing data

Ex: Jan. 1 as everyone's birthday?

- Mismatched data types
- Mixed data values:man or male
- Data outliers-Averaging test scores
- Missing data-blank spaces in text, or unanswered survey questions

working with text data, for example, some things you should consider when cleaning your data are

- Remove URLs, symbols, emojis, etc., that aren't relevant to your analysis
- Translate all text into the language you'll be working in
- Remove HTML tags
- Remove boilerplate email text
- Remove unnecessary blank text between words
- Remove duplicate data

i) Handling Missing Values

- Missing values can be filled in two ways
- 1) Manual 2) Automatic

- Replace with ‘NA’

- With Mean values (Data is normally distributed)

- With Median values (Non-Normally distributed)

- Sometimes replaced with most probable values

Data Cleaning (Dealing with Missing Values)

- Ignore the tuple
- Fill in the missing value manually
- Fill in it automatically with
  - a global constant
  - the attribute mean
  - the attribute mean for all samples belonging to the same class
  - the most probable value

i) Handling Missing Values

1) Ignore the tuple - when the class label is missing, not effective, unless having multiple attributes

2) Fill in the missing value manually - Time consuming and may not be feasible with many missing values

3) Use a global constant - “NA” / “Unknown/∞” -not recommeded(Mining generates common pattern) & simple

4) Use the attribute mean - credit risk category

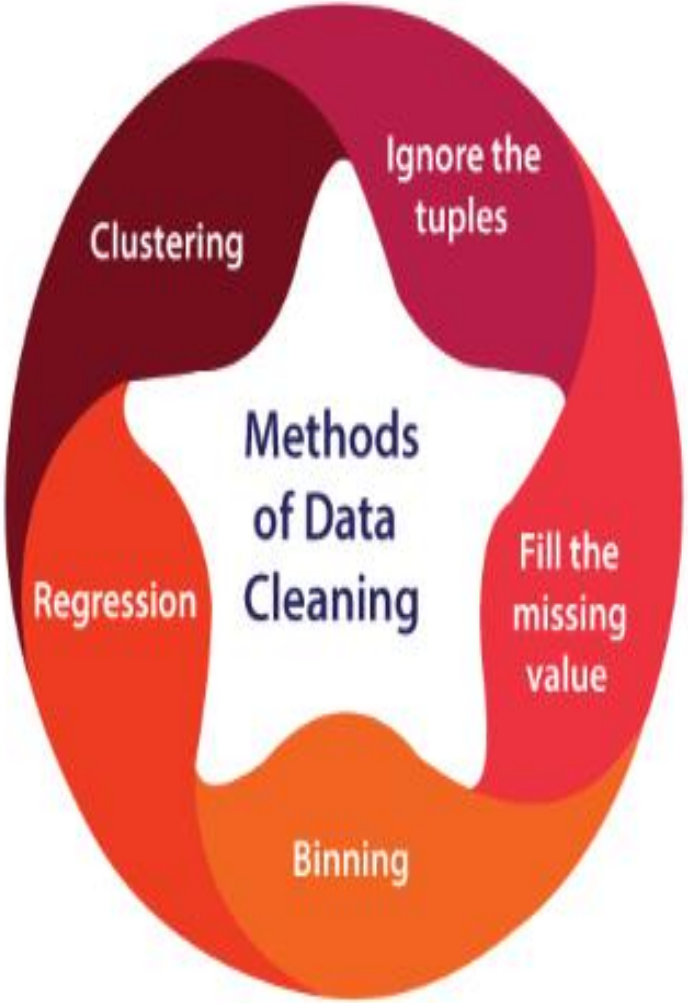
5) Most probable value to fill -regression , inference based tools using decision tree/baysian to fill

ii) Handling Noise Data

- Noise data - inconsistent data/Error data
- 1) Binning Method- smooth a sorted data by considering its neighborhood(value around it)
- First, data will be sorted, then stored in bucket/bins
- Three methods to handle data in bins:
- 4,8,15,21,21,24,25,28,34
- Partion into(Equidepth 3) bins
- Smooting by bin Means
- Smoothing by bin Median
- Smoothing by bin boundary

Methods of Data Cleaning

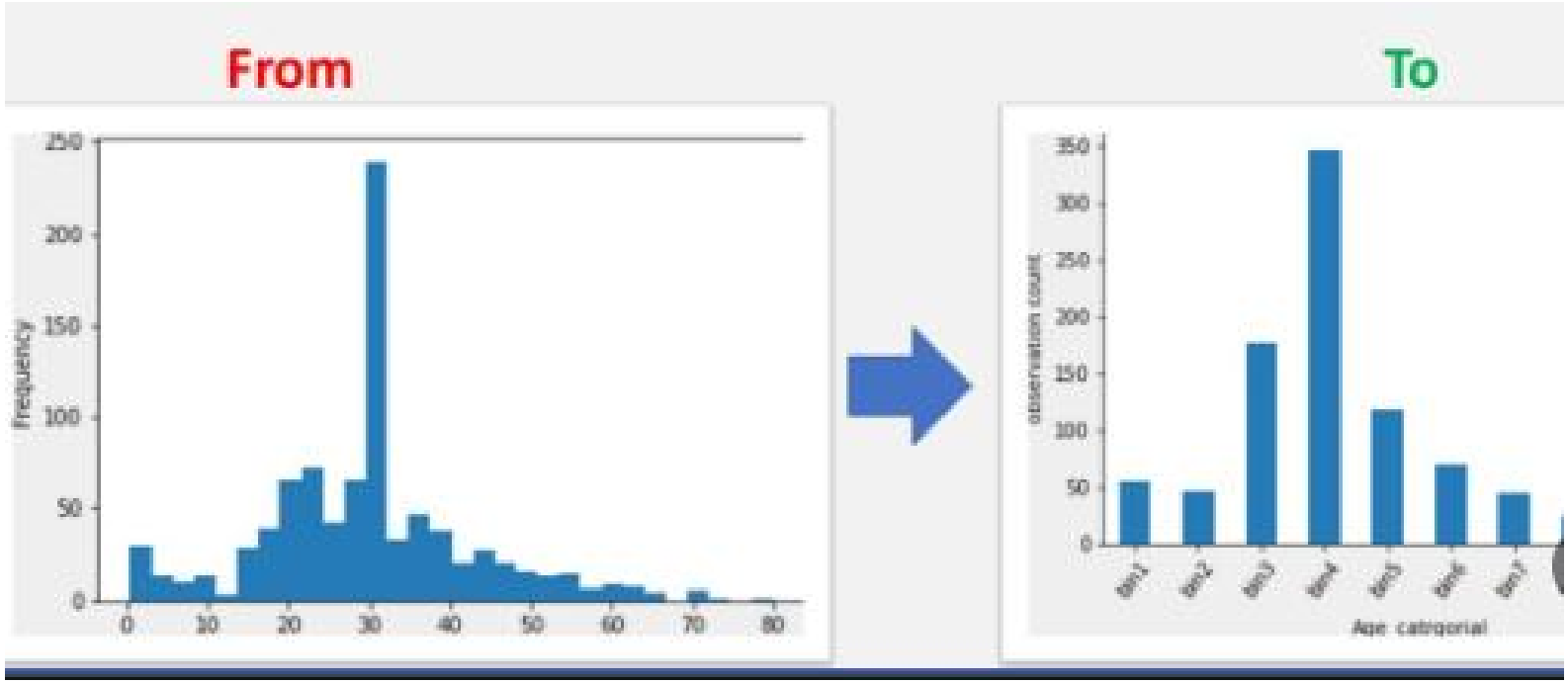
There are many data cleaning methods through which the data should be run. The methods are described below:



Data Cleaning (Dealing with Noise)

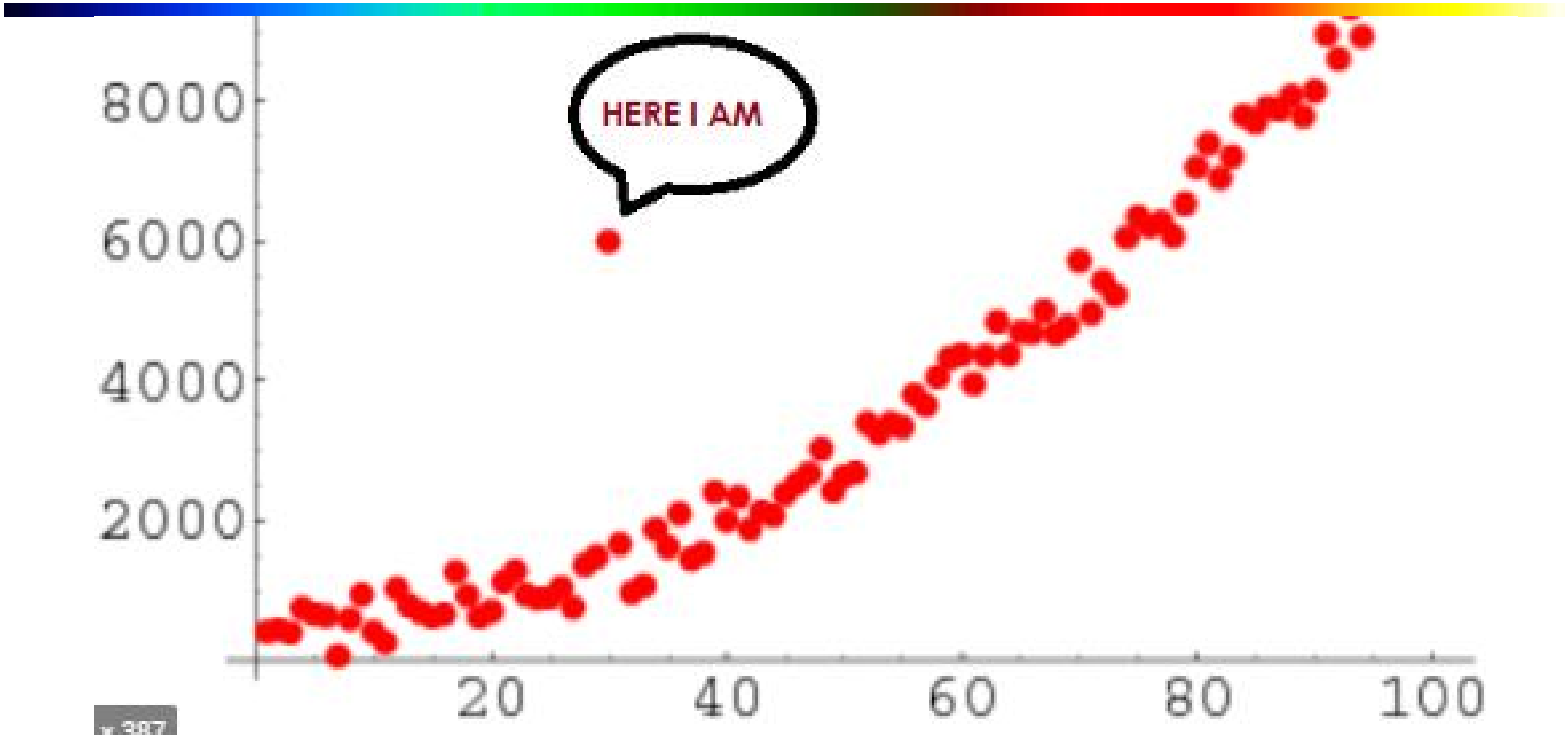
- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions

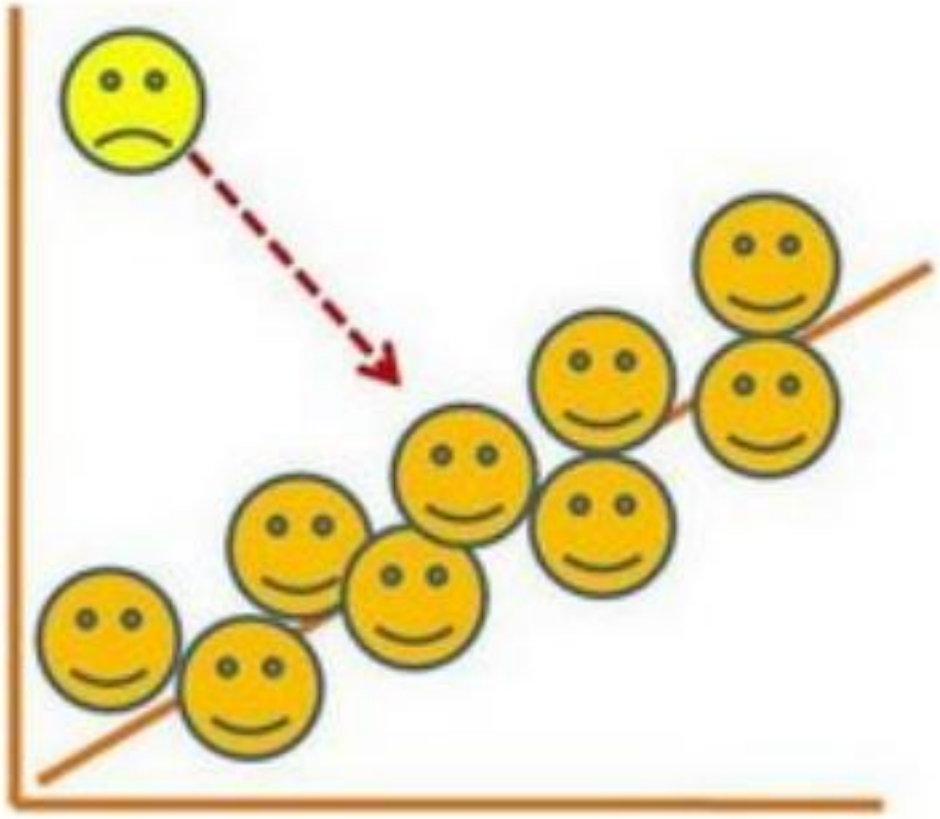
Data Binning

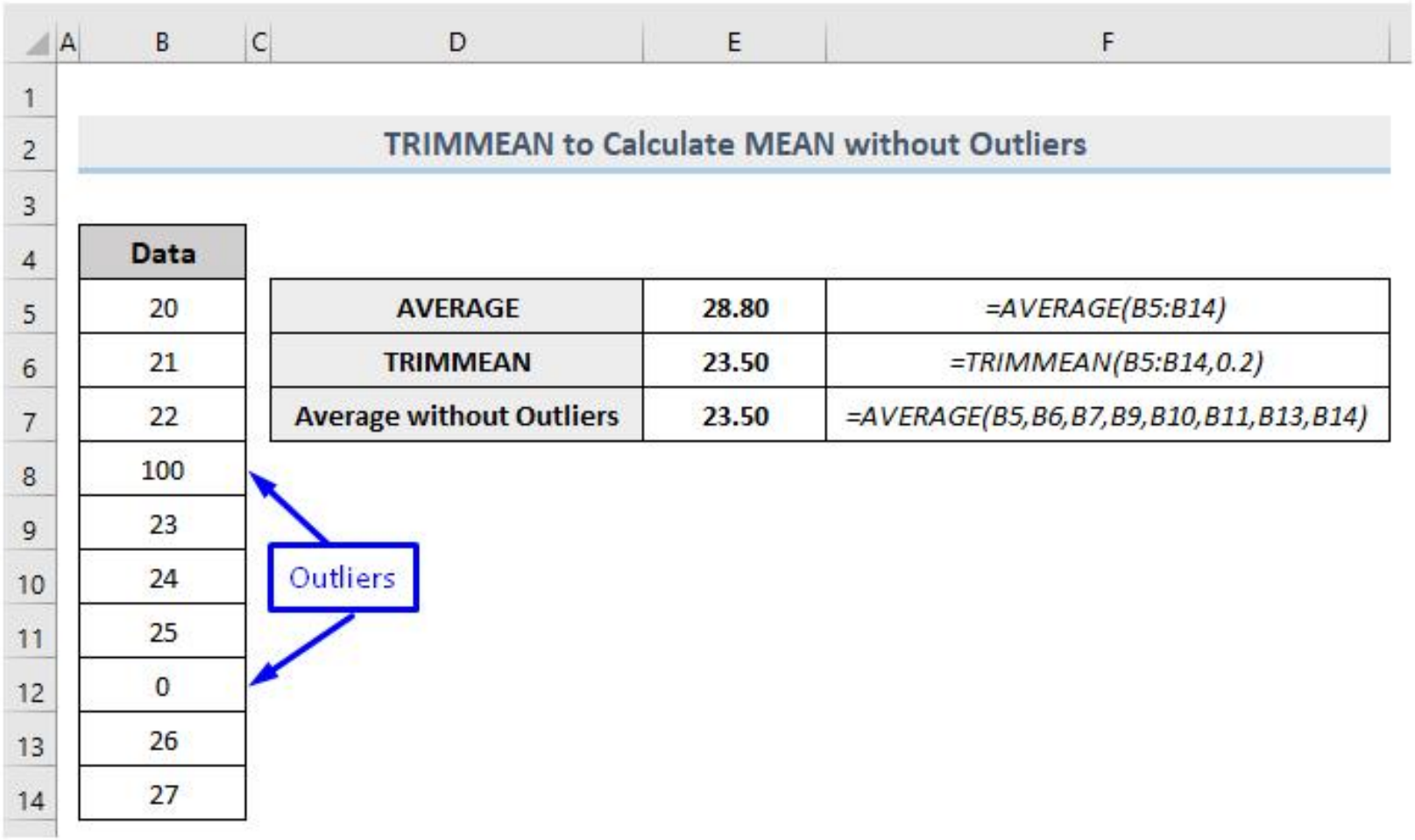


Data Cleaning (Removing Outliers)

- Clustering
  - detect and remove outliers


HERE I AM
8000
6000
4000
2000
20
40
60
80
100




TRIMMEAN to Calculate MEAN without Outliers
Data
20
21
22
100
23
24
25
0
26
27
AVERAGE
28.80
=AVERAGE(B5:B14)
TRIMMEAN
23.50
=TRIMMEAN(B5:B14,0.2)
Average without Outliers
23.50
=AVERAGE(B5,B6,B7,B9,B10,B11,B13,B14)
Outliers

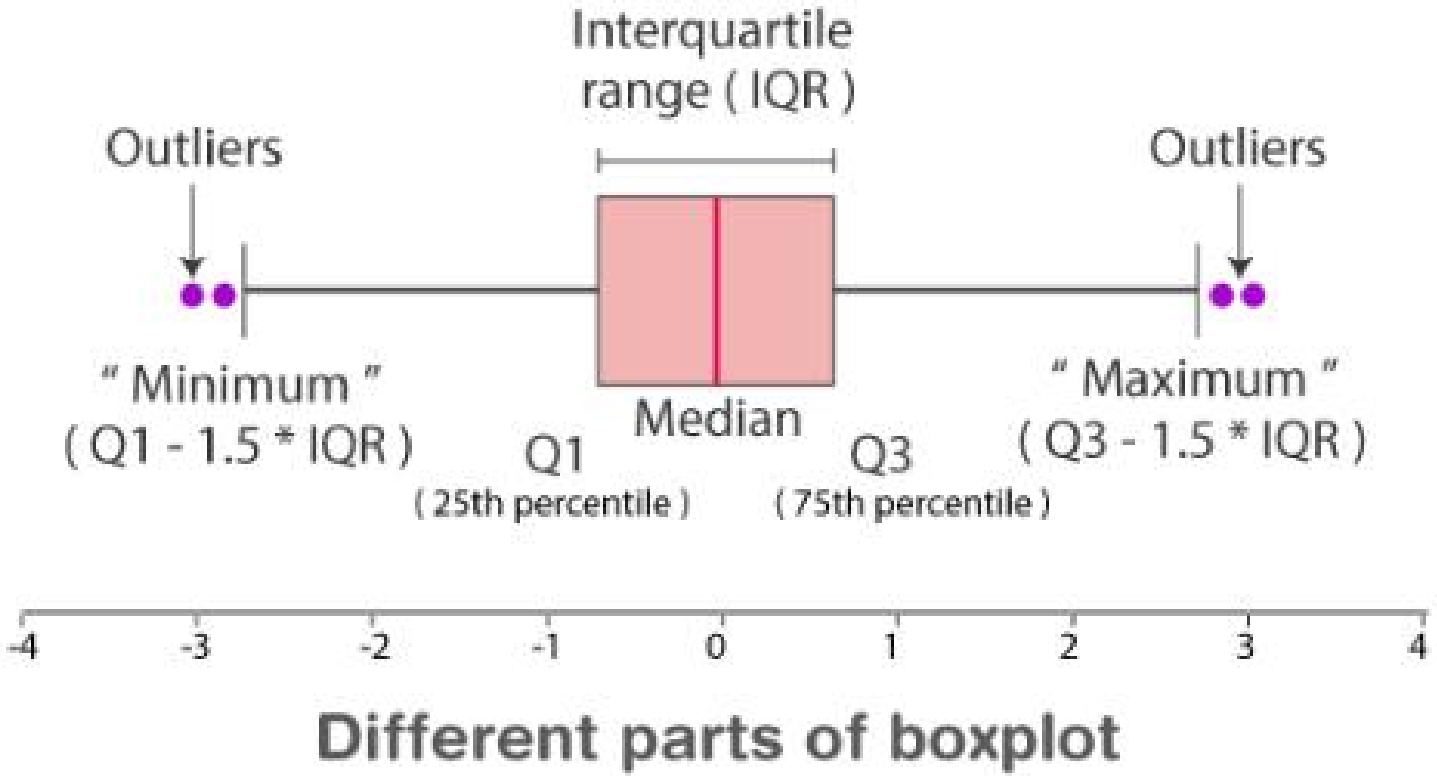
Data Dispersion

- Dispersion of data used to understands the distribution of data.
- It helps to understand the variation of data and provides a piece of information about the distribution data.
- Range, IOR, Variance, and Standard Deviation are the methods used to understand the distribution data.

Boxplots

- Boxplots are generally used in order to measure how well data from a given dataset is distributed and find outlier.
- The box plot is a standardized way to display the distribution of data based on following five number summary.
- Minimum First Quartile
- Median Third Quartile
- Maximum

Parts of Boxplots



Parts of Boxplots

Minimum: The minimum value in the given dataset

First Quartile (Q1): The first quartile is the median of the lower half of the data set.

Median: The median is the middle value of the dataset, which divides the given dataset into two equal parts. The median is considered as the second quartile.

Third Quartile (Q3): The third quartile is the median of the upper half of the data.

Maximum: The maximum value in the given dataset.

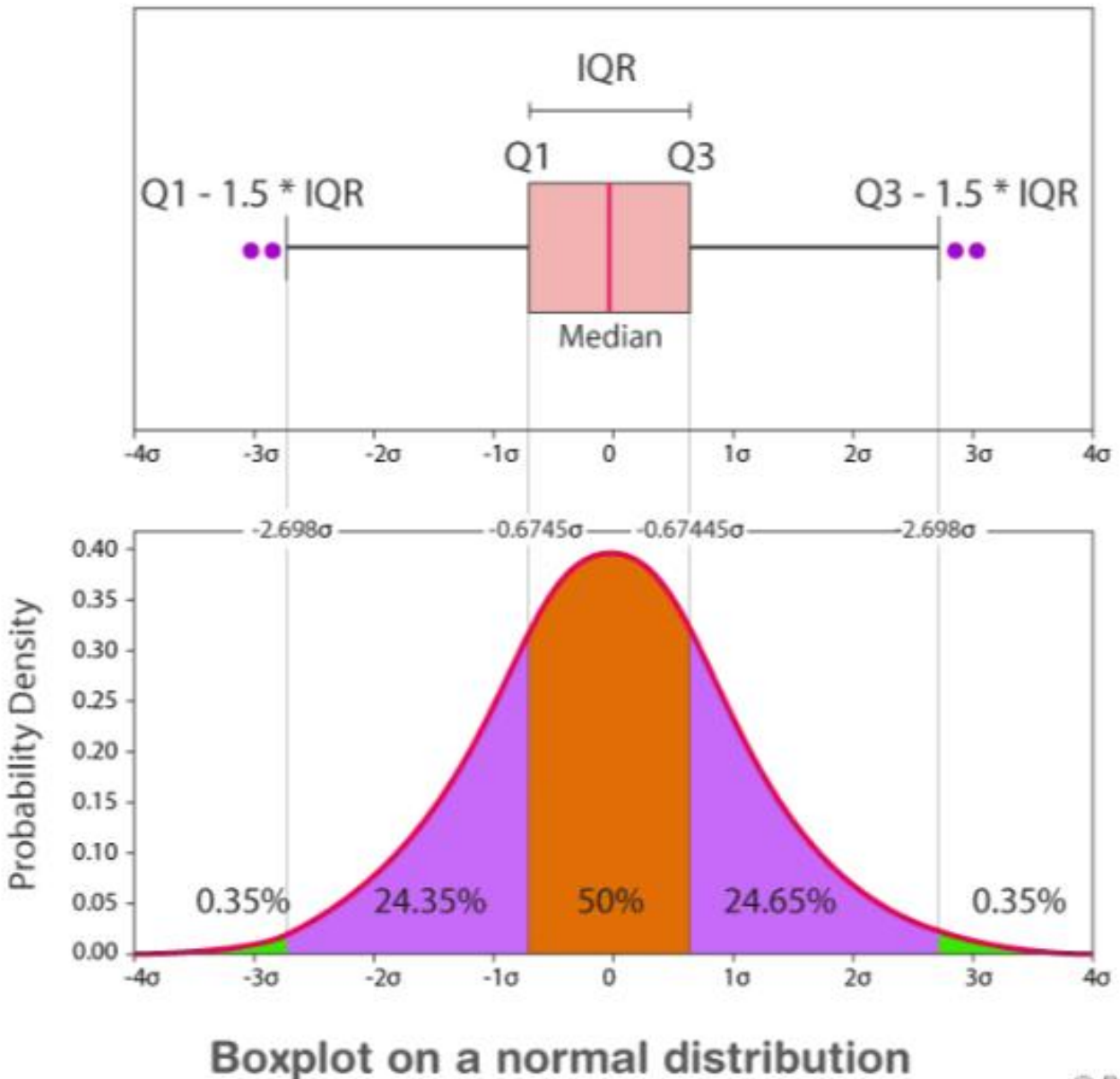
Apart from these five terms, the other terms used in the box plot are:

Interquartile Range (IQR): The difference between the third quartile and first quartile is known as the interquartile range. (i.e.) IQR = Q3-Q1

Outlier: The data that falls on the far left or right side of the ordered data is tested to be the outliers. Generally, the outliers fall more than the specified distance from the first and third quartile.

(i.e.) Outliers are greater than Q3+(1.5 . IQR) or less than Q1-(1.5 . IQR).

Parts of Boxplots



Parts of Boxplots

- Find the maximum, minimum, median, first quartile, third quartile for the given data set: 23, 42, 12, 10, 15, 14, 9.

- 1. Arrange in acenting order
- 2. Calculate Quartiles (Q1, Q2, Q3)
- 3. Identify the minimum and maximum values
- 4. Calculate IQR(Interquatile range) Q3-Q1
- 5. Identify the outliers
- Q1-1.5*IQR-lower bound
- Q3+1.5*IQR-Upper bound

Problem in Boxplots

- 23, 42, 12, 10, 15, 14, 9,28,5,3,50,100

3. Combined computer and Human Inspection

- Cluster may be identified through a combination of computer and human inspection.
- Information-theoretic measure is used

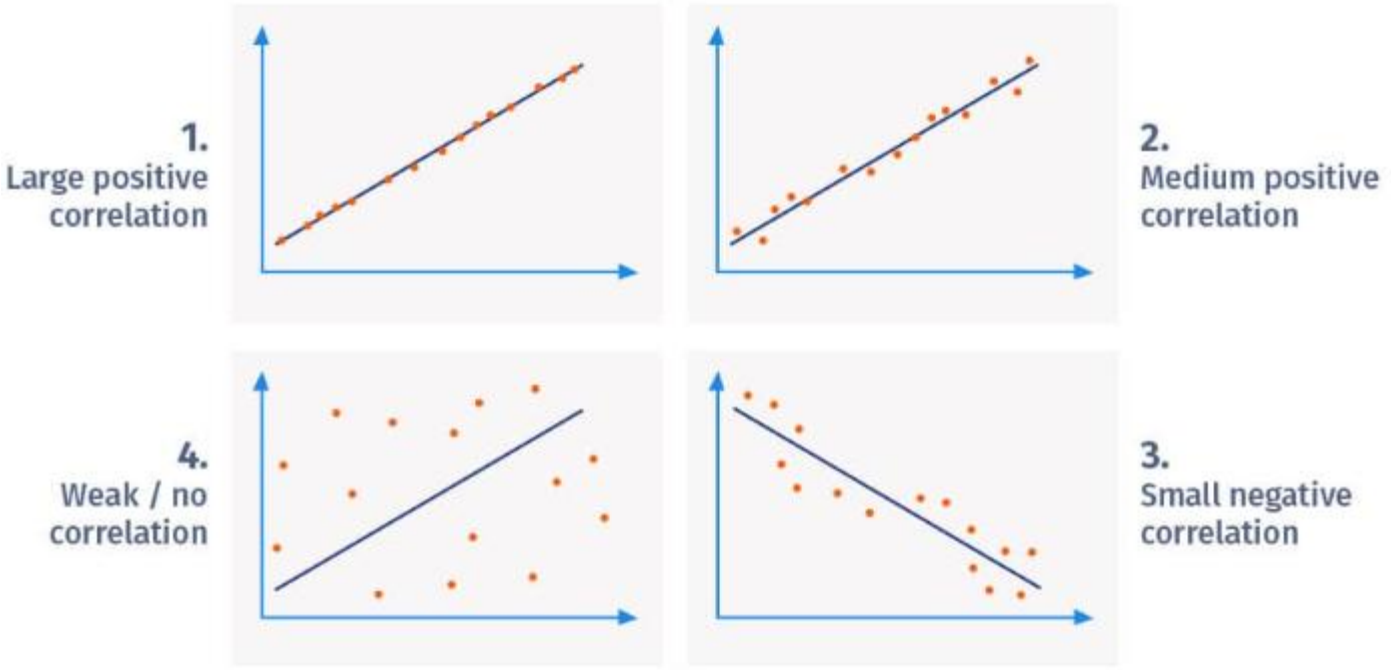
3. Data Integration

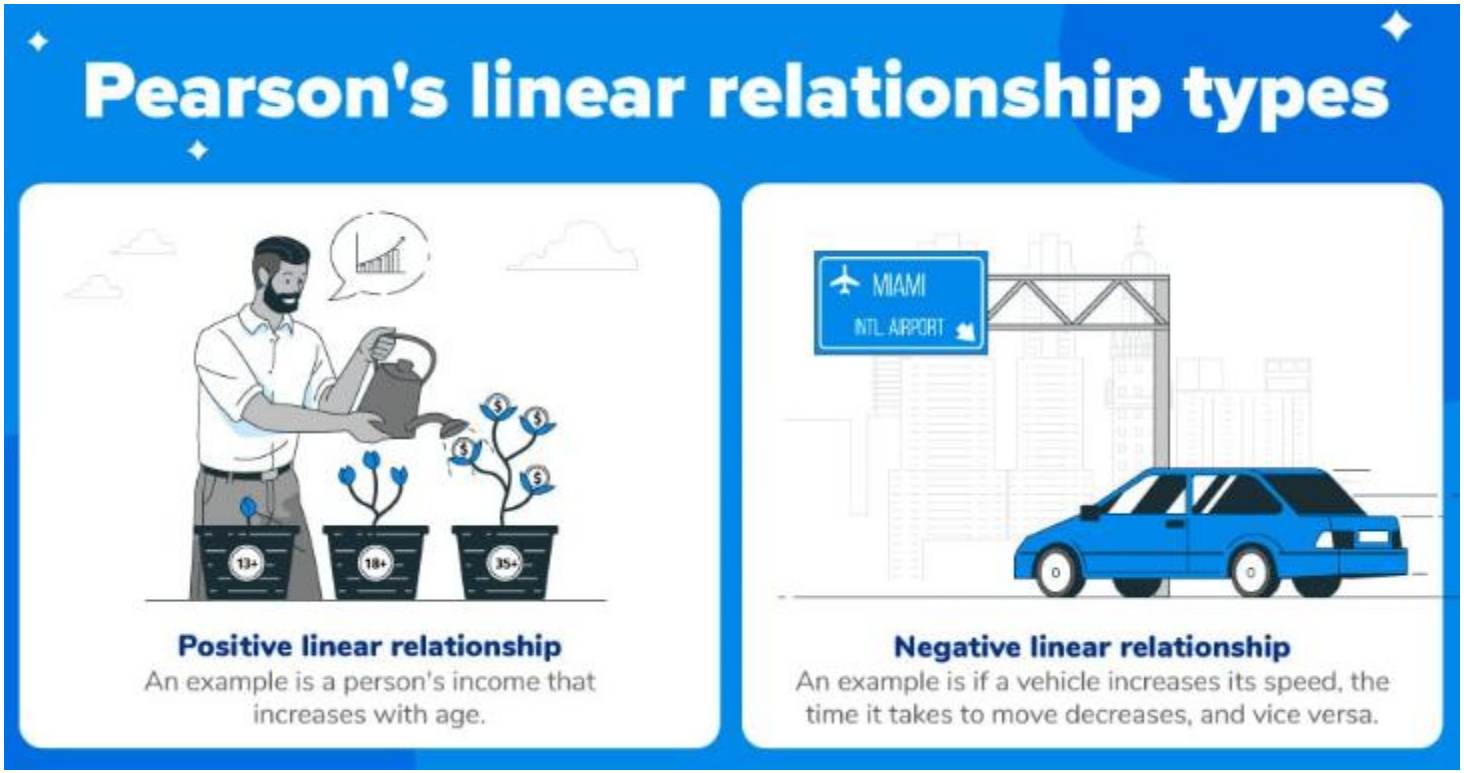
- Entity identification problem - cust_id & cust_number(Meta data can be used to solve schema integration)
- Redundancy- Derived attribute from another table such as annual revenue.
- It can be detected by correlation analysis

Pearson corelation

- Entity identification problem - cust_id & cust_number
- Redundancy- derived attribute from another table such as annual revenue.
- It can be detected by correlation analysis

Pearson corelation




Pearson's linear relationship types
13+
18+
35+
MIAMI
INTL AIRPORT
Positive linear relationship
An example is a person's income that increases with age.
Negative linear relationship
An example is if a vehicle increases its speed, the time it takes to move decreases, and vice versa.

- Find out the number of pairs of variables denoted by n. Suppose x consists of 3 variables : 6, 8, 10. Suppose y consists of corresponding three variables: 12, 10, and 20.

x	y	x*y	x^2	y^2
6	12	72	36	144
8	10	80	64	100
10	20	200	100	400
24	**42**	**352**	**200**	**644**

Pearson Correlation Coefficient

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2]\,[n\Sigma y^2 - (\Sigma y)^2]}}$$

Insert the values found above in the formula and solve it.

r = 3*352-24*42 / √(3*200-24^2)*(3*644-42^2)

= 0.7559

	Coefficient, r	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to 1.0

Detection and resolution of data value conflicts

- **Mean:** To calculate the mean of a given data set, we use the following formula,

$$\text{Mean } (\bar{x}) = \frac{\sum x}{N}$$

- **Median:** In the case of the median, we have two different formulas. If we have an odd number of terms in the data set we use the following formula,

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{observation}$$

If an even number of terms are given in the data set, we use the following formula,

$$\text{Median} = \frac{\left(\frac{n}{2}\right)^{th} observation + \left(\frac{n}{2}+1\right)^{th} observation}{2}$$

- **Variance:** The variance is defined as the total of the square distances from the mean (μ) of each term in the distribution, divided by the number of distribution terms (N).

$$\text{Variance}(\sigma^2) = \frac{\sum(x_i - \mu)^2}{N}$$

- **Standard Deviation:** By evaluating the deviation of each data point relative to the mean, the standard deviation is calculated as the square root of variance.

$$\text{Standard deviation}(\sigma) = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

4. Data Transformation

- This step prepares the data to be fed to the data mining algorithms.
- Hence, the data needs to be in consolidated and aggregate forms.
- The data is consolidated on the basis of functions, attributes, features etc.

This generally happens in one or more of the below:

- Smoothing - Remove noise, Binning, clustering, and regression are the techniques
- Aggregation-Data aggregation combines all of your data together in a uniform format.
- Normalization -It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
- Feature/Attribute construction-In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

Why Normalization

- -Distance algorithms like KNN, K-means, and SVM use distances between data points to determine their similarity. They’re most affected by a range of features. Machine learning algorithms like linear regression and logistic regression use gradient descent for optimization techniques that require data to be scaled

Min-Max Normalization

For example, if the minimum value of a feature was 20, and the maximum value was 40, then 30 would be transformed to about 0.5 since it is halfway between 20 and 40. The formula is as follows:

$$\frac{value - min}{max - min}$$

There are five numeric values: 17, 19, 8, 29, 45, 60, 31

- suppose the supermarket wants to normalize their daily income are min 15,000 and 58,000 respectively.
- Now, the new income i.e 46,600 for income is transformed to normalize value.
- Find the anser?

Z-Score Normalization

New value = (x – μ) / σ

where:

- **x**: Original value
- **μ**: Mean of data
- **σ**: Standard deviation of data

It's a good practice to remove correlated variables during feature selection.

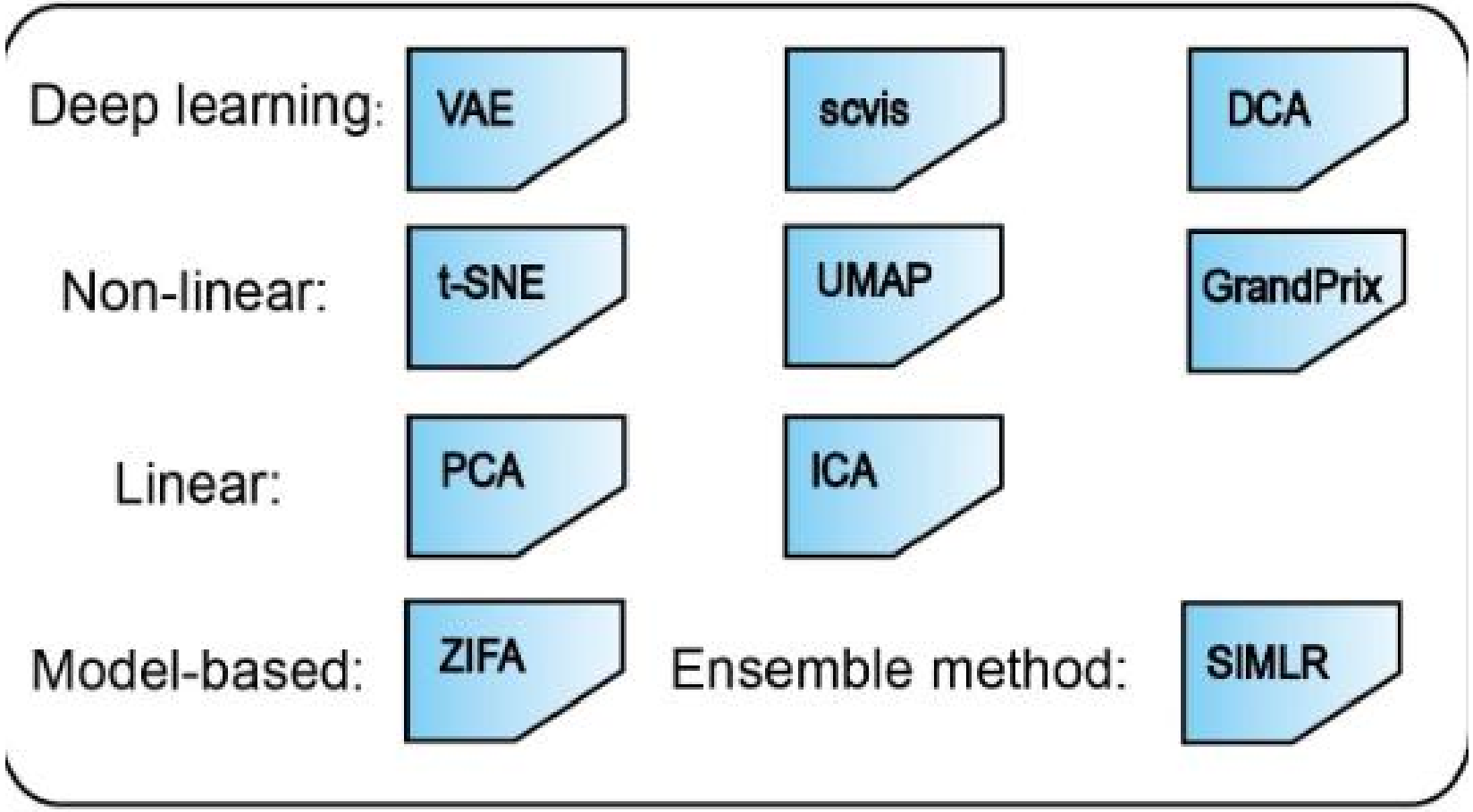
Correlation Matrix

A

5. Data Reduction

- Volume of data is reduced to make a mining/analysis easier.(Image/text conversion)
- 1) Dimensionality reduction: Data is combined to construct a data cube.(Redundant and noise data is removed), where irrelevant, weakly relevant, redundant attributes may be detected and removed.
- 2) Attribute subset/Feature selection:
- Highly relevant attributes shouls be used
- 3) Numerosity reduction/Compression: (.csv file, text college file) Use the model of data instead of entire data(Sampling technique). Encoding mechanisms are used to reduce the data set size.

Dimensionality reduction methods



- Dimensionality reduction is the process of transforming a dataset from a high-dimentional space to a low-dimensional space whilst maintaining it’s informational integrety for predictive modelling

contd.,

- Discreditization-This is done to replace the raw values of numeric attribute by interval levels or conceptual levels. (when calculating average daily exercise, rather than using the exact minutes and seconds, you could join together data to fall into 0-15 minutes, 15-30, etc.)
- Concept hierarchy generation/Generalization-Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

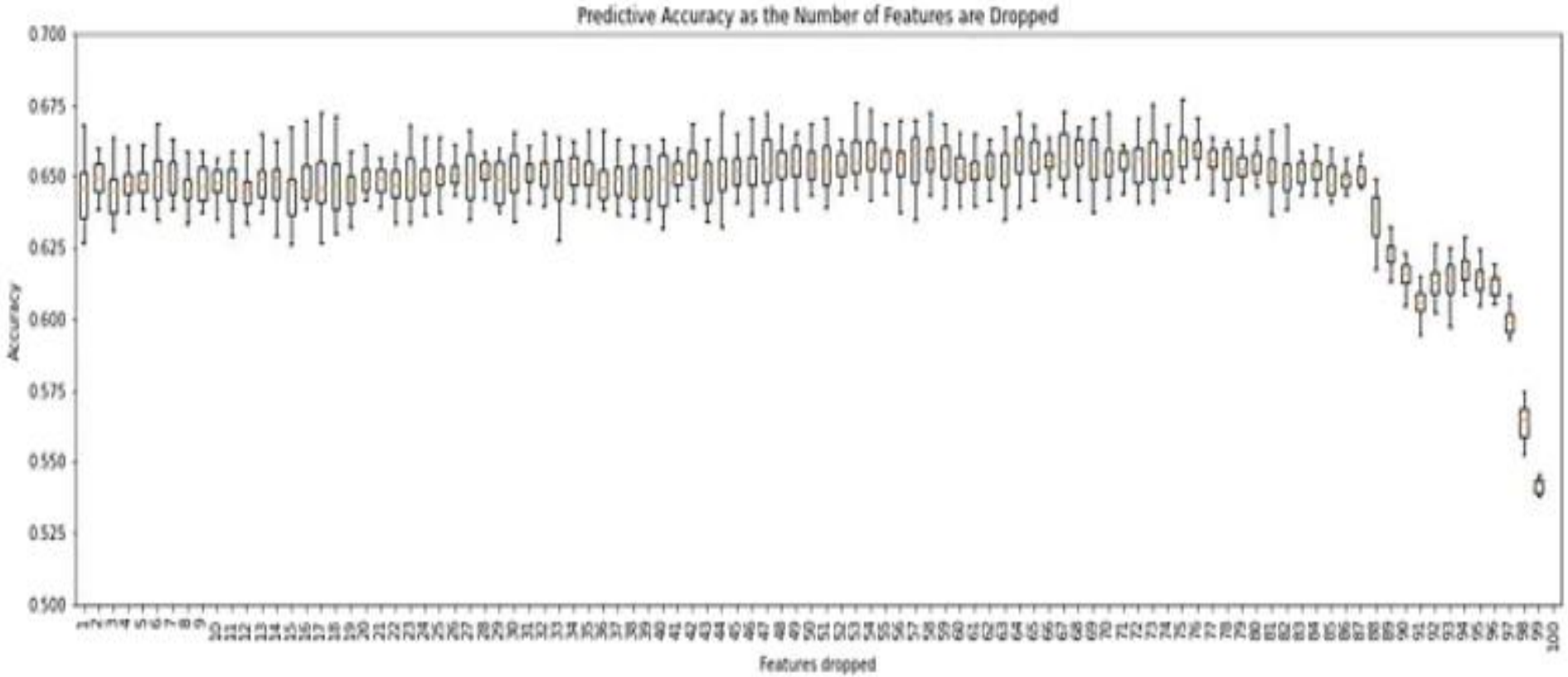


Image by author

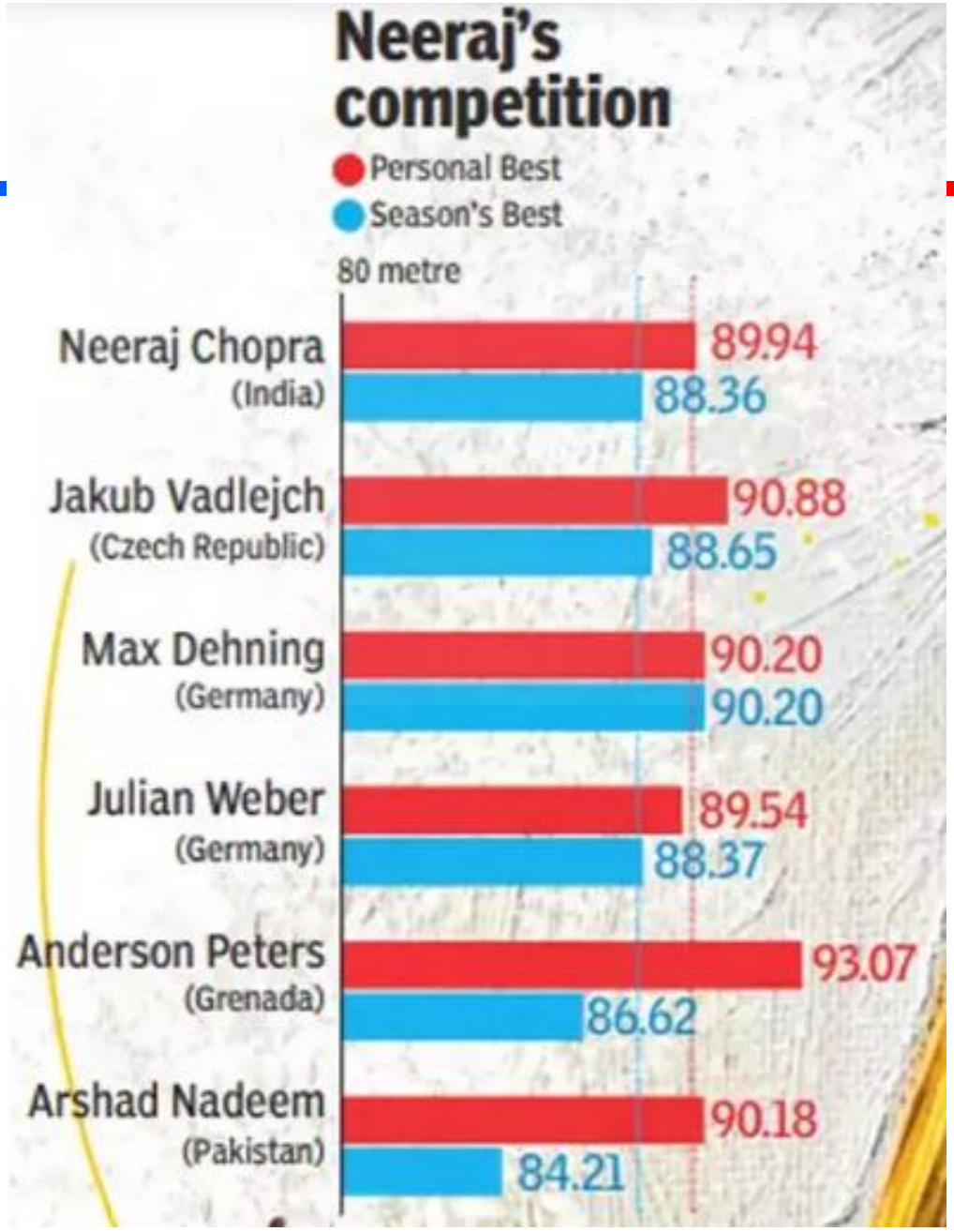
The Figure above shows how the model's predictive accuracy reduces as the number of least important features is dropped from the dataset.

5. Data Mining

- This is the root or backbone process of the whole KDD. This is where algorithms are used to extract meaningful patterns from the transformed data, which help in prediction models.

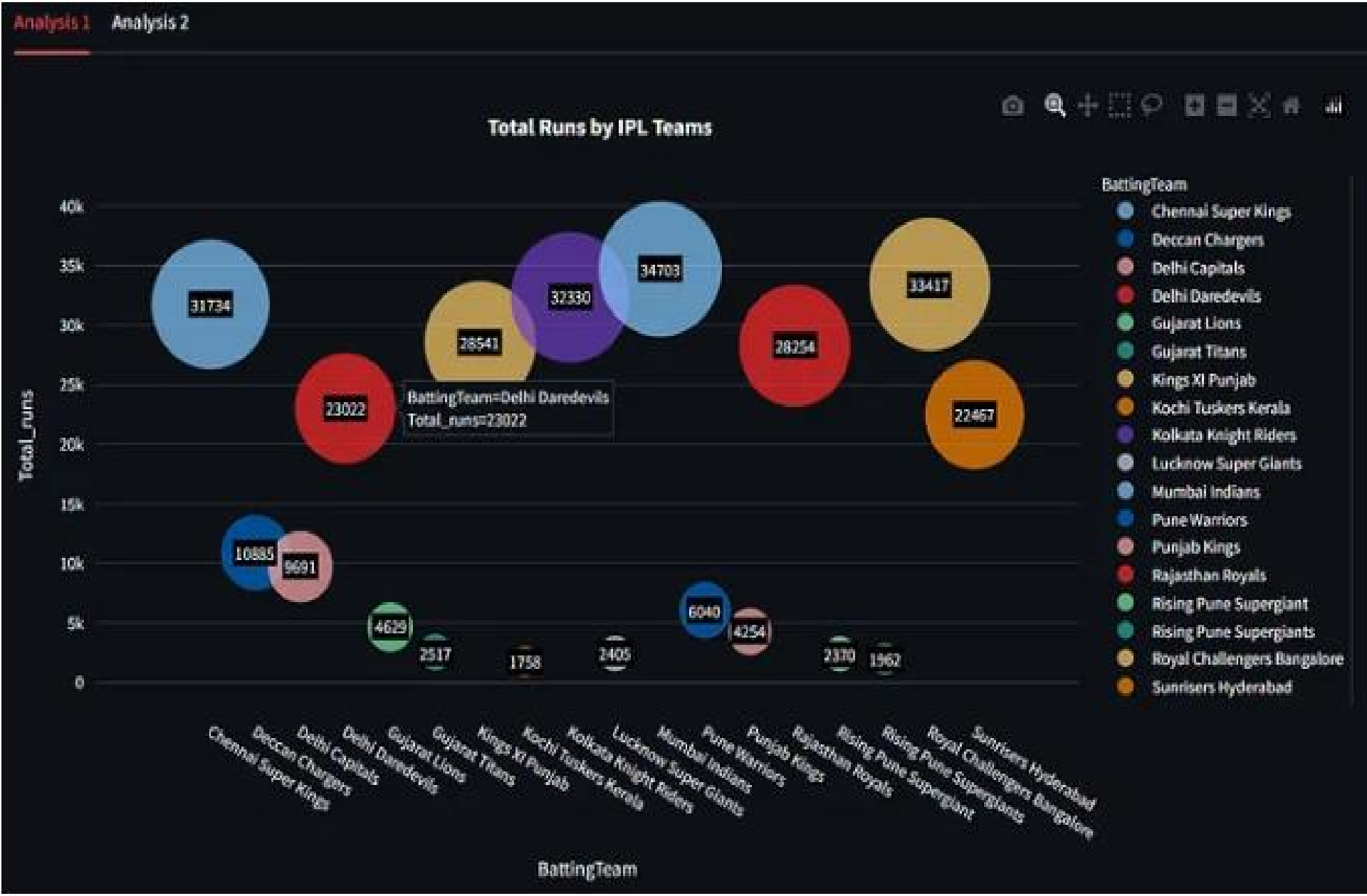
6. Pattern Evaluation/Interpretation

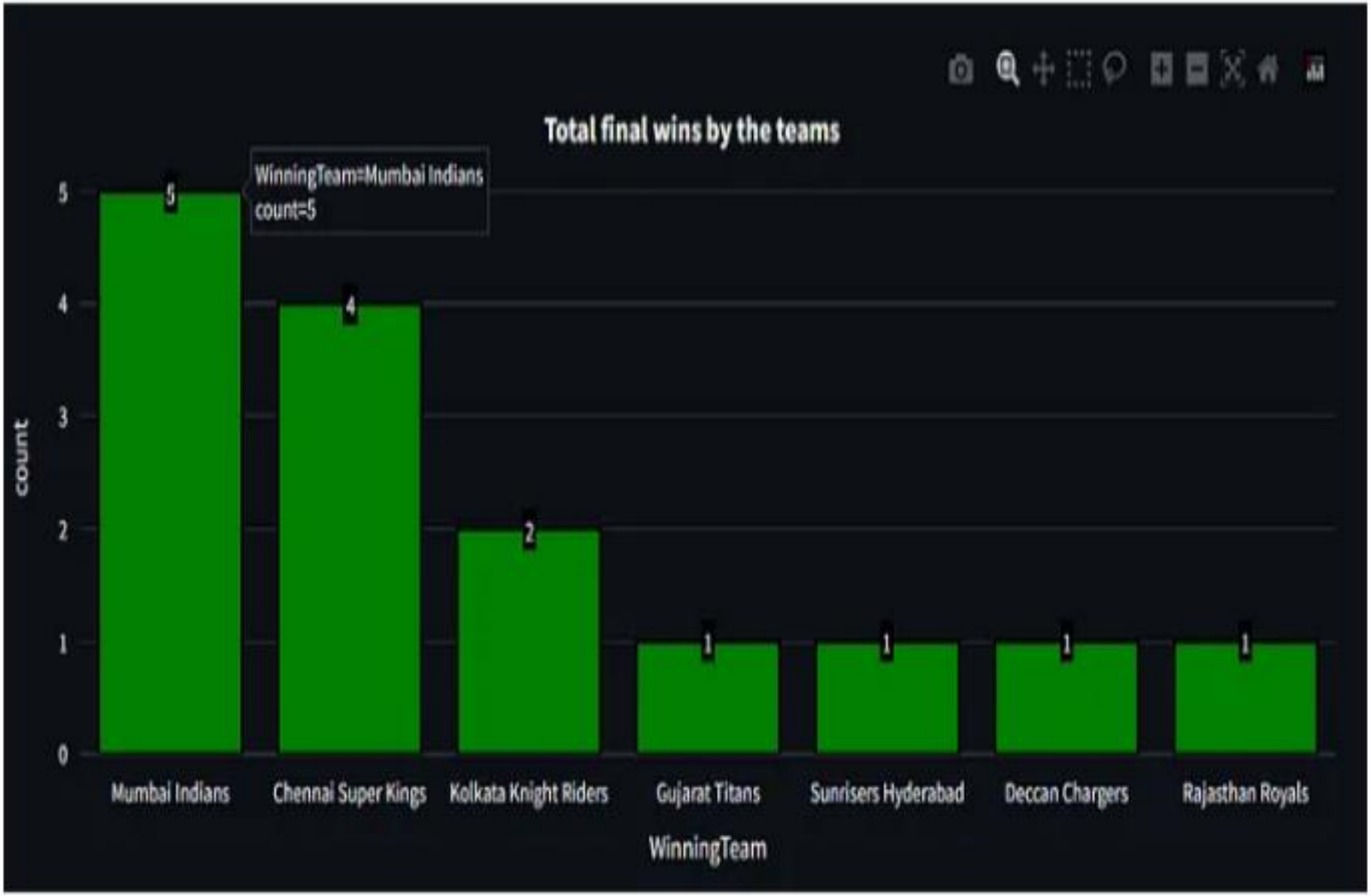
- Once the trend and patterns have been obtained from various data mining methods and iterations, these patterns need to be represented in **discrete forms** such as bar graphs, pie charts, histograms etc. to study the impact of data collected and transformed during previous steps.
- This also helps in evaluating the effectiveness of a particular data model in view of the domain.


Neeraj's competition
Personal Best
Season's Best
80 metre
Neeraj Chopra
(India)
89.94
88.36
Jakub Vadlejch
(Czech Republic)
90.88
88.65
Max Dehning
(Germany)
90.20
90.20
Julian Weber
(Germany)
89.54
88.37
Anderson Peters
(Grenada)
93.07
86.62
Arshad Nadeem
(Pakistan)
90.18
84.21

7. Knowledge Discovery and Use

- This is the final step in the KDD process and requires the 'knowledge' extracted from the previous step to be applied to the specific application or domain in a visualised format such as Visualization, interactive tools, tables, reports etc.
- Presenting the discovered knowledge in a meaningful and understandable way to the end-users or decision-makers.
- This step drives the decision-making process for the said application.

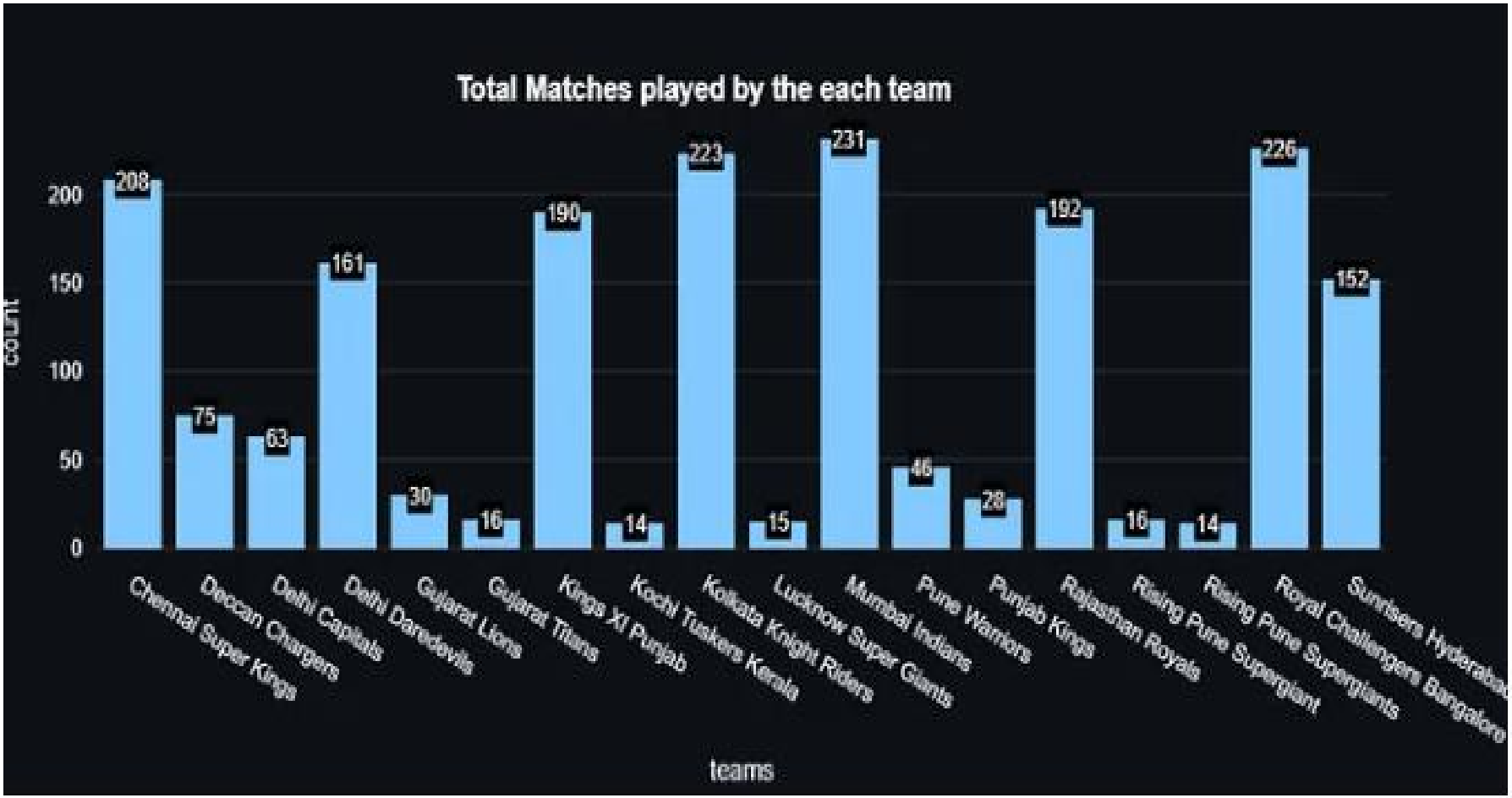

Analysis 1
Analysis 2
Total Runs by IPL Teams
Total_runs
BattingTeam
BattingTeam=Delhi Daredevils
Total_runs=23022
40k
35k
30k
25k
20k
15k
10k
5k
0
31734
23022
28541
32330
34708
28254
33417
22467
10885
9691
4629
2517
1758
2405
6040
4254
2370
1962
Chennai Super Kings
Deccan Chargers
Delhi Capitals
Delhi Daredevils
Gujarat Lions
Gujarat Titans
Kings XI Punjab
Kochi Tuskers Kerala
Kolkata Knight Riders
Lucknow Super Giants
Mumbai Indians
Pune Warriors
Punjab Kings
Rajasthan Royals
Rising Pune Supergiant
Rising Pune Supergiants
Royal Challengers Bangalore
Sunrisers Hyderabad

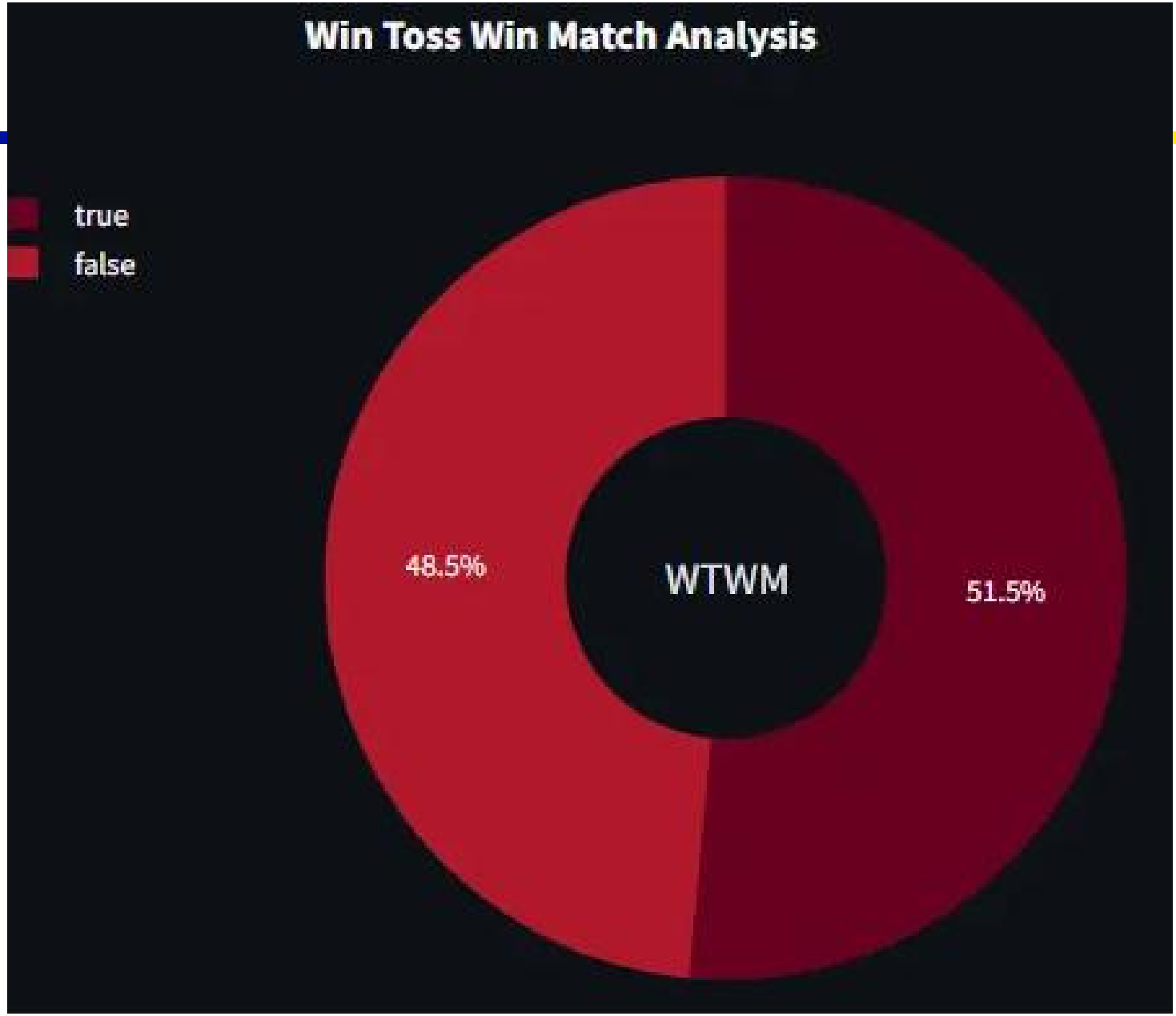


August 21, 2024

Total matches win by the teams

	WinningTeam	count
0	Mumbai Indians	131
1	Chennai Super Kings	121
2	Kolkata Knight Riders	114
3	Royal Challengers Bangalore	109
4	Rajasthan Royals	96
5	Kings XI Punjab	88
6	Sunrisers Hyderabad	75
7	Delhi Daredevils	67
8	Delhi Capitals	36
9	Deccan Chargers	29
10	Gujarat Lions	13
11	Punjab Kings	13
12	Pune Warriors	12
13	Gujarat Titans	12
14	Rising Pune Supergiant	10


Total Matches played by the each team
count
0
50
100
150
200
208
75
63
161
30
16
190
14
223
15
231
46
28
192
16
14
226
152
Chennai Super Kings
Deccan Chargers
Delhi Capitals
Delhi Daredevils
Gujarat Lions
Gujarat Titans
Kings XI Punjab
Kochi Tuskers Kerala
Kolkata Knight Riders
Lucknow Super Giants
Mumbai Indians
Pune Warriors
Punjab Kings
Rajasthan Royals
Rising Pune Supergiant
Rising Pune Supergiants
Royal Challengers Bangalore
Sunrisers Hyderabad
teams


Win Toss Win Match Analysis
true
false
48.5%
WTWM
51.5%

Summary

- Data mining: Discovering interesting patterns and knowledge from massive amount of data
- A natural evolution of database technology, in great demand, with wide applications
- A KDD process includes data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation
- Mining can be performed in a variety of data
- Data mining functionalities: characterization, discrimination, association, classification, clustering, outlier and trend analysis, etc.
- Data mining technologies and applications
- Major issues in data mining

Recommended Reference Books

- S. Chakrabarti. Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data. Morgan Kaufmann, 2002
- R. O. Duda, P. E. Hart, and D. G. Stork, Pattern Classification, 2ed., Wiley-Interscience, 2000
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley & Sons, 2003
- U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996
- U. Fayyad, G. Grinstein, and A. Wierse, Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed., 2011
- D. J. Hand, H. Mannila, and P. Smyth, Principles of Data Mining, MIT Press, 2001
- T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer-Verlag, 2009
- B. Liu, Web Data Mining, Springer 2006.
- T. M. Mitchell, Machine Learning, McGraw Hill, 1997
- G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991
- P.-N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Wiley, 2005
- S. M. Weiss and N. Indurkhya, Predictive Data Mining, Morgan Kaufmann, 1998
- I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2nd ed. 2005