

Lecture 1

Data Mining Introduction

Dr. Amira Rezk

dramirarezk@gmail.com

Sara S. Elhishi

sarashaker161@gmail.com

Information Systems Dept.

DATA MINING

- Vast amounts of data are collected daily
- knowledge mining from data
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data



The Knowledge Discovery Process



DATA CLEANING
remove noise



**DATA
INTEGRATION**
FROM MULTIPLE
SOURCES



DATA SELECTION
DATA RELEVANT
TO THE ANALYSIS
TASK



**DATA
TRANSFORMATION**
INTO FORMS
APPROPRIATE FOR
MINING



DATA MINING
EXTRACT DATA
PATTERNS



**PATTERN
EVALUATION**
INTERESTINGNESS
MEASURES



**KNOWLEDGE
PRESENTATION**
VISUALIZATION
TO USERS

Not Data Mining vs. Data Mining

- Searching for cooking on Google vs. Grouping similar cuisines French, Italian, Arabian ...
- Looking up spa resorts vs. More relevant Spa for curing certain diseases

Data Mining is NOT about searching in a Data, but more about
Implicit meaningful Information

What Kinds of Data Can Be Mined ?

Data
warehouses

Data streams

Time-series data

Structure data,
graphs, social
networks

Heterogeneous
databases

Spatial data and
spatiotemporal
data

Multimedia
database

Text databases

The World-Wide
Web



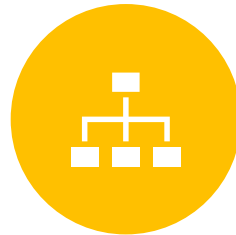
What Kinds of Patterns Can Be Mined?



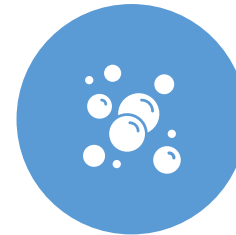
CHARACTERIZATION
AND DISCRIMINATION



ASSOCIATION /
CORRELATION
ANALYSIS



CLASSIFICATION



CLUSTER ANALYSIS



OUTLIER ANALYSIS

Mining Tasks

Descriptive Tasks

- Characterize properties of the data in a target data set.
- e.g., (classification, regression, anomalies/outliers detection)

Predictive Tasks

- Perform induction on the current data in order to make predictions.
- (e.g., clustering, association rule discovery, sequential pattern discovery)

Technologies



STATISTICS



**MACHINE
LEARNING**



**PATTERN
RECOGNITION**



DATABASE



**INFORMATION
RETRIEVAL**



**VISUALIZATION
N**

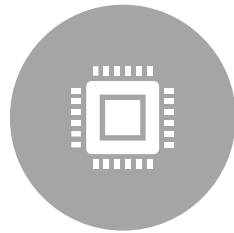


ALGORITHMS

Applications



WEB PAGE
ANALYSIS



RECOMMENDER
SYSTEMS



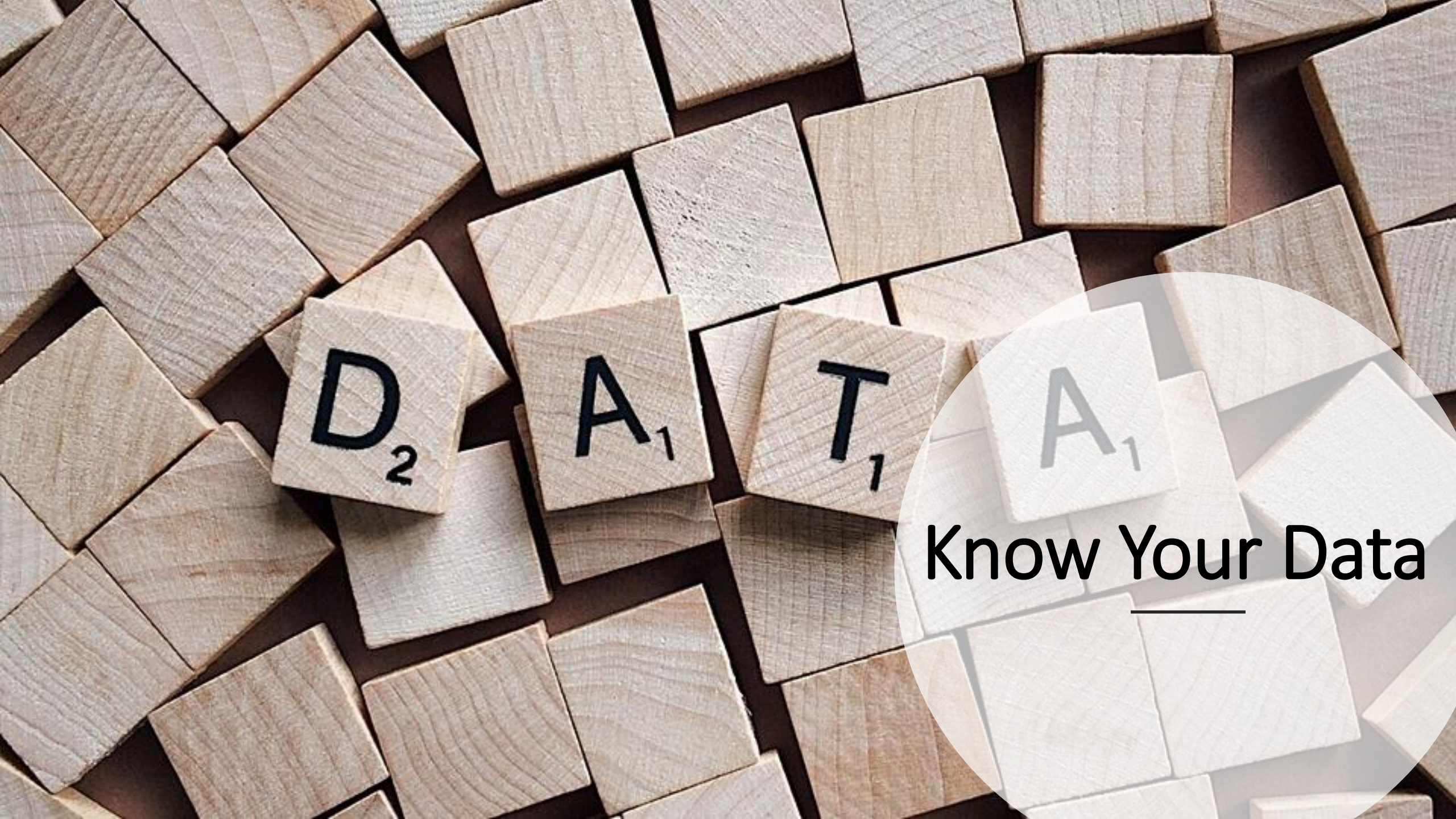
BASKET DATA
ANALYSIS



MEDICAL DATA
ANALYSIS



..ETC



D_2

A_1

T_1



Know Your Data

Data Objects

- Data sets are made up of data objects.
- Also referred as samples, examples, instances, data points .
- e.g. customers, students, patients, books .
- Data objects are typically described by attributes.





Attributes

- A data field, representing a characteristic or feature of a data object.
- attribute, dimension, feature, and variable are often used interchangeably .
- A customer object can include, for example, customer ID, name, and address.
- Observed values for a given attribute are known as observations.

Data,
as the values
taken by its
attributes



CATEGORIAL DATA
QUALITATIVE



NUMERICAL DATA
QUANTITATIVE

Categorical

Nominal Attributes

- The values of a nominal attribute are symbols or names of things
- do not have any meaningful order
- e.g. hair color, marital status, occupation

Binary Attributes

- a nominal attribute with only two categories or states: 0 or 1
- symmetric if both of its states carry the same weight (e.g. gender)

Ordinal Attributes

- an attribute with possible values that have a meaningful order or ranking among them
- e.g. professional rank, grade, customer satisfaction

Numeric

Interval-Scaled

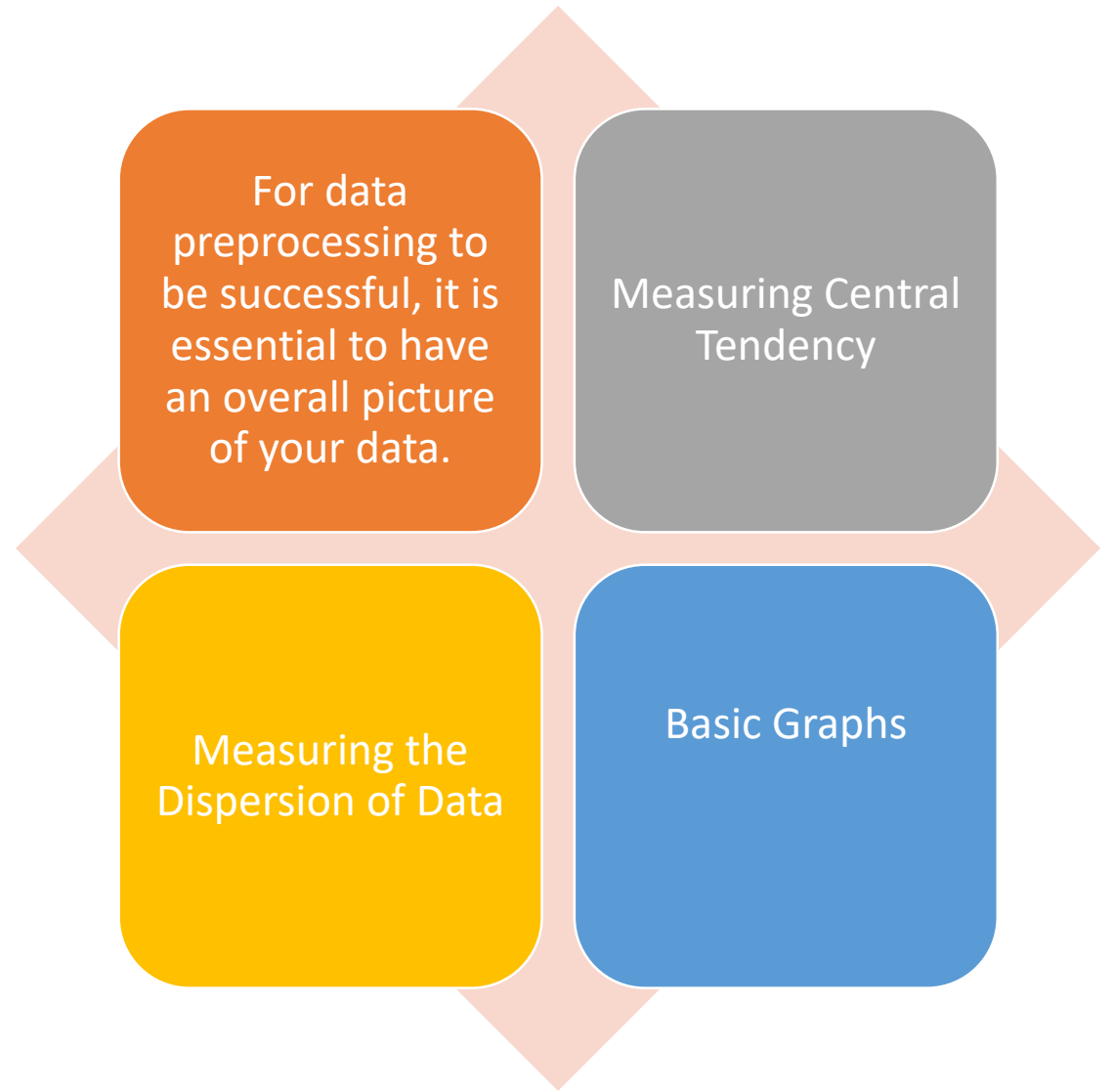
- measured on a scale of equal-size units.
- does not have a true zero-point.
- e.g. temperature, neither 0C nor 0F indicates “no temperature.”

Ratio-Scaled

- a numeric attribute with an inherent zero-point
- e.g. years of experience



Statistical Descriptions of Data



Mean

- Let x_1, x_2, \dots, x_n be a set of N values or observations

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} = \frac{x_1 + x_2 + \dots + x_N}{N}.$$


- Weighted mean:

$$\bar{x} = \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i} = \frac{w_1 x_1 + w_2 x_2 + \dots + w_N x_N}{w_1 + w_2 + \dots + w_N}.$$

- A major problem with the mean is its sensitivity to extreme (e.g., outlier) values.
- we can instead use the trimmed mean, which is the mean obtained after chopping off values at the high and low extremes.



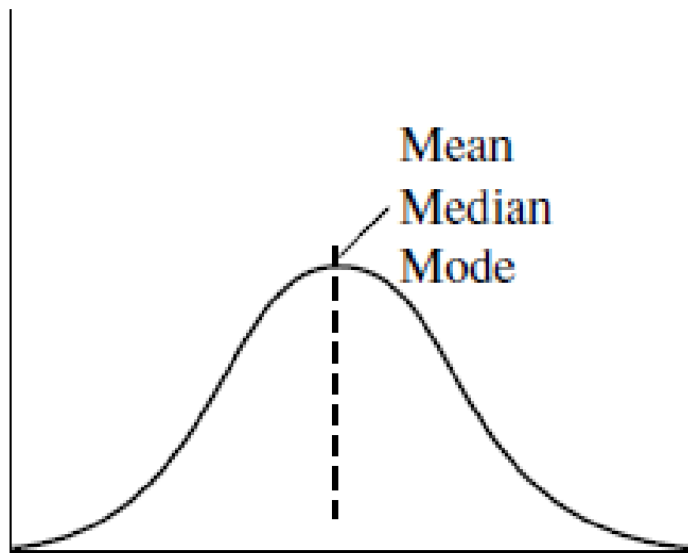
Median

- the middle value in a set of ordered data values.
 - It is the value that separates the higher half of a data set from the lower half.
 - The median is expensive to compute when we have a large number of observations
- 

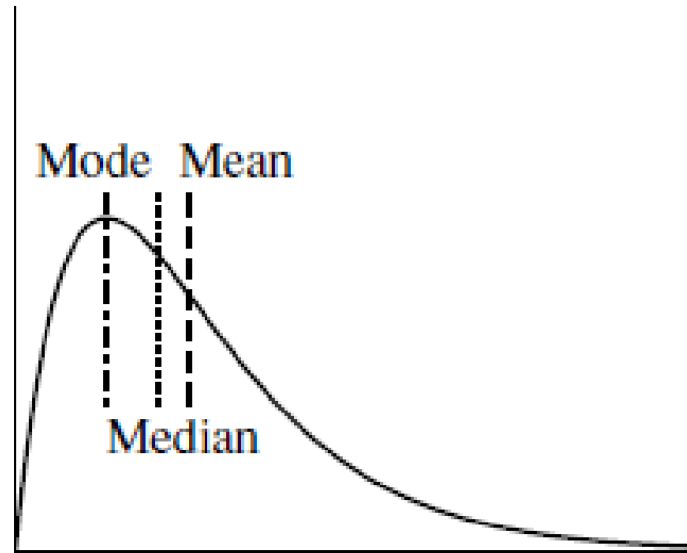


Mode & Mid Range

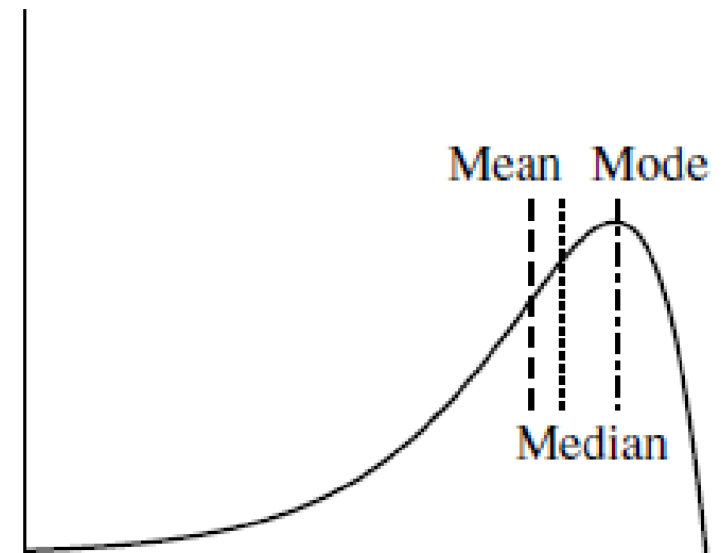
- the value that occurs most frequently in the set
- it can be determined for qualitative and quantitative attributes.
- Data sets with one, two, or three modes are respectively called unimodal, bimodal, and trimodal.
- **Mid Range:** is the average of the largest and smallest values in the set.



(a) Symmetric data



(b) Positively skewed data



(c) Negatively skewed data

Data in most real applications are not symmetric. They may instead be either positively skewed, where the mode occurs at a value that is smaller than the median or negatively skewed, where the mode occurs at a value greater than the median

$$\bar{x} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12}$$
$$= \frac{696}{12} = 58.$$

Mean = 58,000

$$\text{Median} = \frac{52 + 56}{2} = \frac{108}{2} = 54,000$$

Mode = 52,000 and 70,000 – bimodal

$$\text{Midrange} = \frac{30,000 + 110,000}{2} = 70,000$$

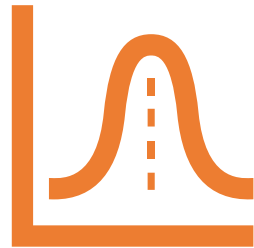
Example

- Salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110

Measuring the Dispersion of Data

The dispersion or spread of numeric data is useful in identifying outliers .

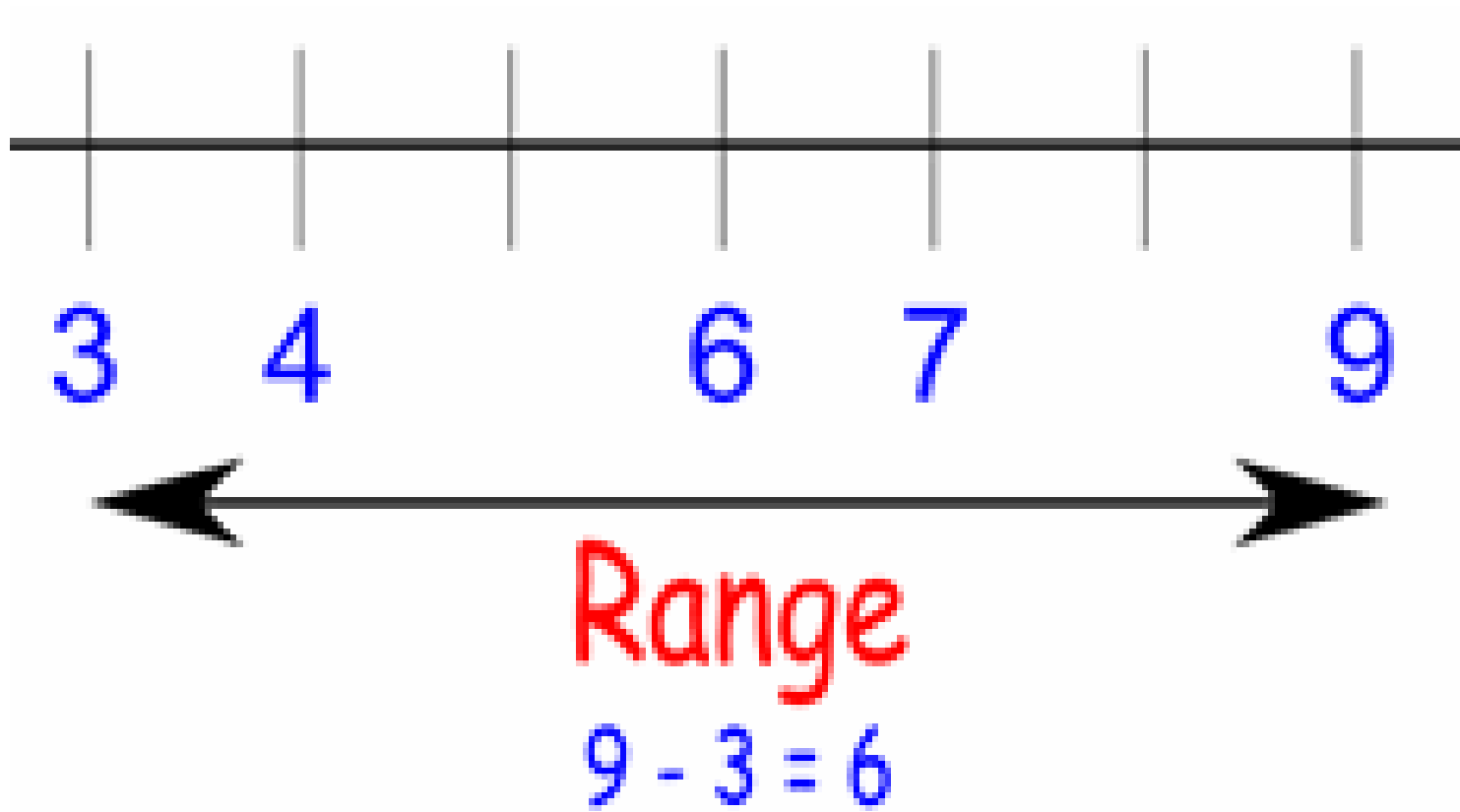
Variance and Standard Deviation



low standard deviation means that the data observations tend to be very close to the mean,

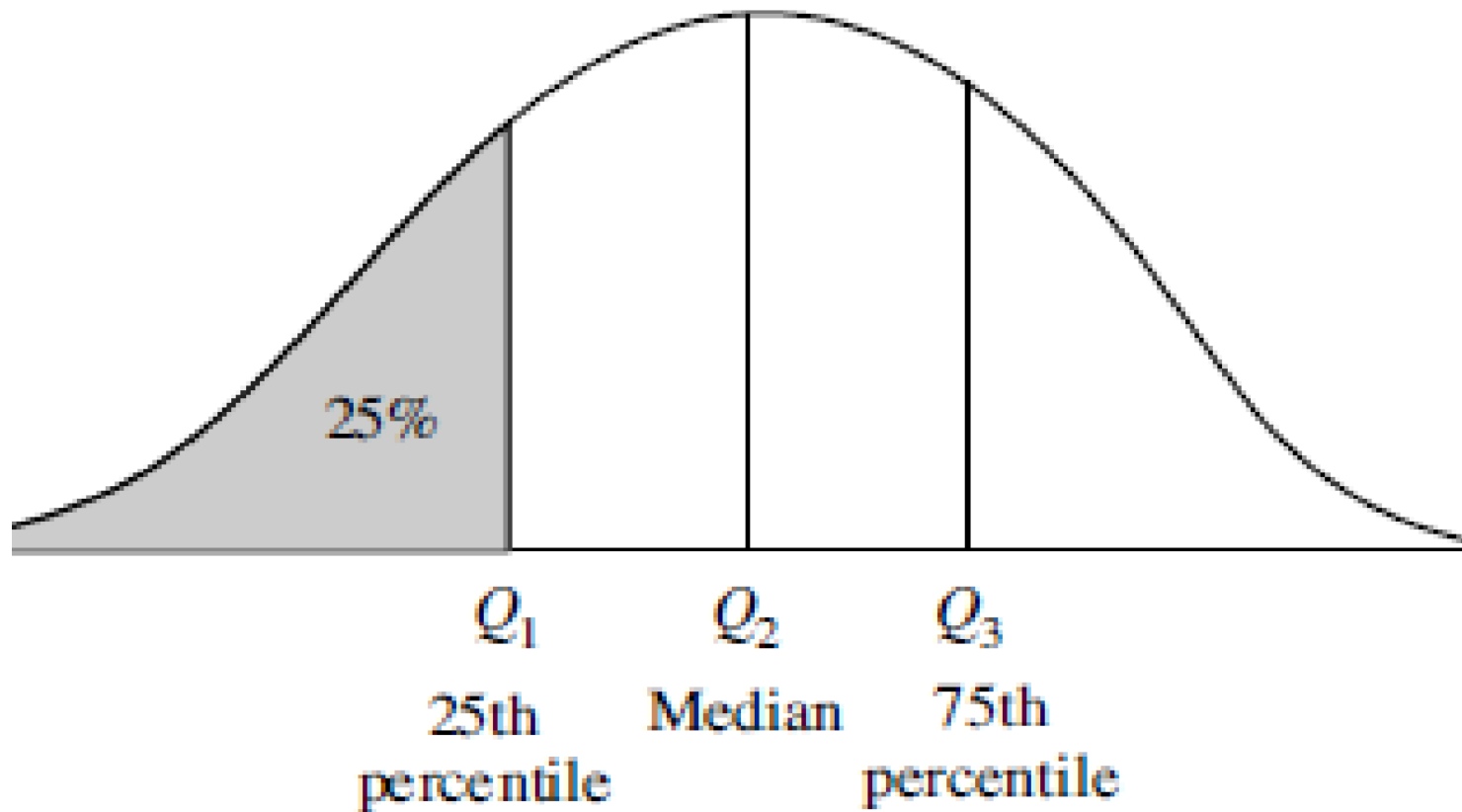


while a high standard deviation indicates that the data are spread out over a large range of values.



- the difference between the largest (`max()`) and smallest (`min()`) values

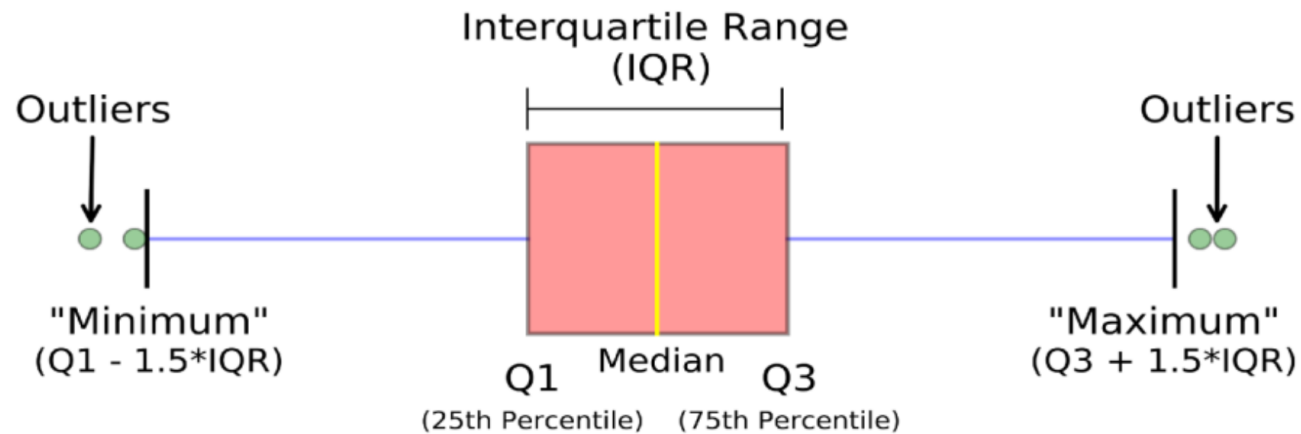
range



- are points taken at regular intervals of a data distribution, dividing it into equal size consecutive sets.
- For example ,3 quantiles shown to the left
- The distance between the first and third quartiles is *interquartile range (IQR)*

Quantiles

Five-Number Summary, Boxplots, and Outliers



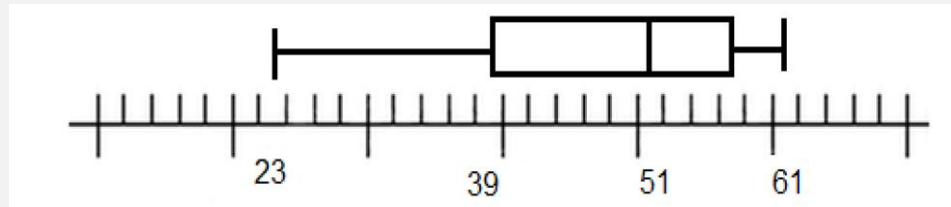
- a standardized way of displaying the distribution of data based on a five numbers summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum").

Example

- Draw the boxplot for the following data sets
- Age: 23, 23, 27, 27, 39, 41, 47, 49, 50, 52, 54, 54, 56, 57, 58, 58, 60, 61
- %fat: 9.5, 26.5, 7.8, 17.8, 31.4, 25.9, 27.4, 27.2, 31.2, 34.6, 42.5, 28.8, 33.4, 30.2, 34.1, 32.9, 41.2, 35.7

First:
order the data
set if it is not
ordered

- Age: 23, 23, 27, 27, 39, 41, 47, 49, 50, 52, 54, 54, 56, 57, 58, 58, 60, 61
- **%fat:** 7.8, 9.5, 17.8, 25.9, 26.5, 27.2, 27.4, 28.8, 30.2, 31.2, 31.4, 32.9, 33.4, 34.1, 34.6, 35.7, 41.2, 42.5

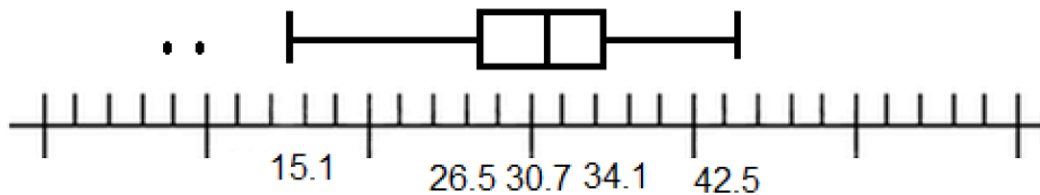


For Age

Q1=39, median= 51, Q3=57,
min=23, max=61

$IQR = 57 - 39 = 18 \rightarrow 1.5 IQR = 27$

newMin= $39 - 27 = 12$, newMax=
 $57 + 27 = 84$



For Fat

Q1=26.5, median= 30.7, Q3=34.1,
min=15.1, max=42.5

$IQR = 34.1 - 26.5 = 7.6$, $1.5 IQR = 11.4$

newMin= $26.5 - 11.4 = 15.1$, newMax=
 $34.1 + 11.4 = 45.5$

Lab

- Setup your workspace by installing the Anaconda Distribution
- Get used to Jupyter Notebook environment
- Python Fundamentals



Thanks