Dr. SNS RAJALAKSHMI COLLEGE OF ARTS AND SCIENCE

(AUTONOMOUS)

Accredited by NAAC (Cycle- III) with 'A+' Grade

DEPARTMENT OF B.SC CS (GCD)

22UDA501 – INTRODUCTION TO DATA ANALYTICS
UNIT- II

Data Quality

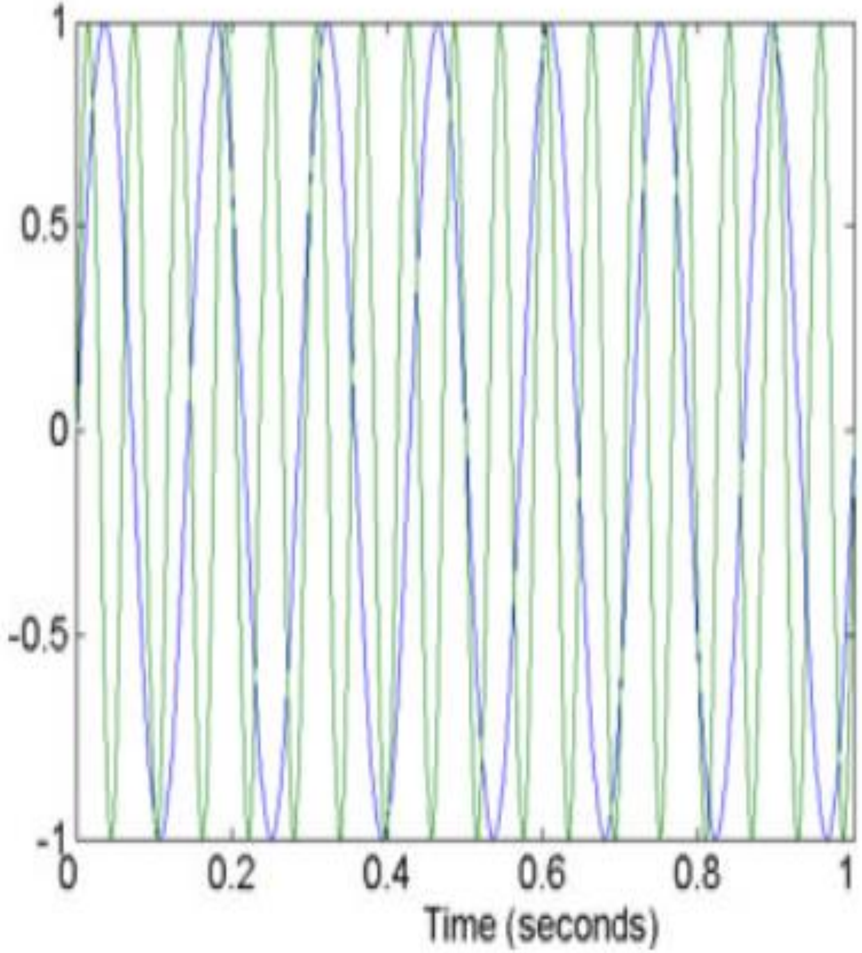
- Data Quality Analysis is the process of analyzing the quality of data in datasets to determine potential issues, shortcomings, and errors.
- The purpose is to identify these and resolve them before using the data for analysis or modeling.

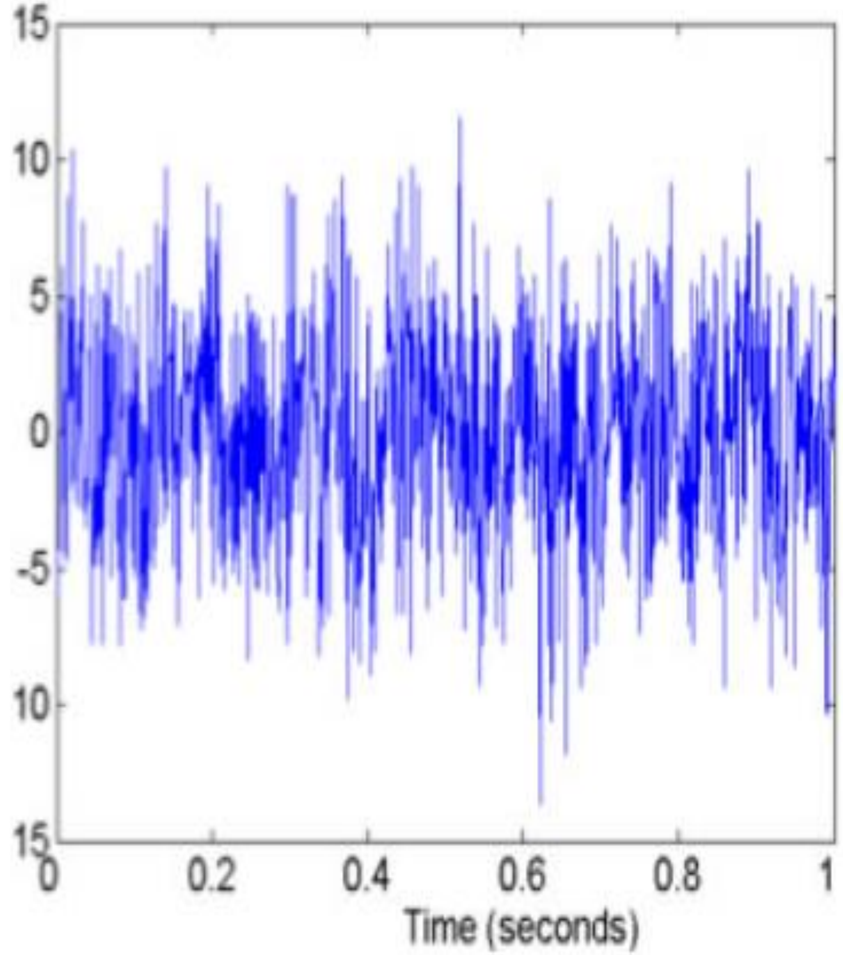
Examples of data quality problems

- Noise and outliers
- Missing values
- Duplicate data
- Wrong data

Noisy data

- For objects, noise is considered an extraneous object.
- For attributes, noise refers to modification of original values.
- Examples: distortion of a person’s voice when talking on a poor phone and “snow” on television screen
- We can talk about **signal to noise ratio.**
  Left image of 2 sine waves has low or zero SNR; the right image are the two waves combined with noise and has high SNR


1
0.5
0
-0.5
-1
0
0.2
0.4
0.6
0.8
1
Time (seconds)


15
10
5
0
-5
10
15
0
0.2
0.4
0.6
0.8
1
Time (seconds)

Origins of noise

- **outliers --** values seemingly out of the normal range of data
- **duplicate records --** good database design should minimize this (use DISTINCT on SQL retrievals)
- **incorrect attribute values --** again good db design and integrity constraints should minimize this
- **numeric only**, deal with rogue strings or characters where numbers should be.
- **null handling** for attributes (nulls=missing values)

Missing Data Handling

Many causes: malfunctioning equipment, changes in experimental design, collation of different data sources, measurement not possible. Information is not applicable (children don't have annual income)

- **Discard** records with missing values
- **Ordinal-continuous** data, could **replace with attribute means**
- **Substitute** with a value from a similar instance
- **Ignore** missing values, i.e., just proceed and let the tools deals with them
- **Treat** missing values **as equals** (all share the same missing value code)
- **Treat** missing values **as unequal values**

Missing completely at random (MCAR)

- Missingness of a value is independent of attributes
- Fill in values based on the attribute as suggested above (e.g. attribute mean)
- Analysis may be unbiased overall

Missing at Random (MAR)

- Missingness is related to other variables
- Fill in values based other values (e.g., from similar instances)
- Almost always produces a bias in the analysis

Missing Not at Random (MNAR)

- Missingness is related to unobserved measurements
- Informative or non-ignorable missingness

Inaccurate values

- Data may not been collected for mining purposes
- Errors and omissions don't affect original purpose of data (e.g. age of customer)
- Typographical errors in nominal attributes => values need to be checked for consistency
- Typographical and measurement errors in numeric attributes => outliers need to be identified
- Errors may be deliberate (e.g. wrong postal codes, birthdates)
- Other problems: duplicates, stale data

Duplicate Data

Data set may include data objects that are duplicates, or almost duplicates of one another

- A major issue when merging data from multiple, heterogeneous sources
- Examples: Same person with multiple email addresses

Data Quality Dimensions

- **Accuracy:** The data should reflect actual, real-world scenarios; the measure of accuracy can be confirmed with a verifiable source.
- **Completeness:** Completeness is a measure of the data’s ability to effectively deliver all the required values that are available.

- **Consistency:** Data consistency refers to the uniformity of data as it moves across networks and applications. The same data values stored in difference locations should not conflict with one another.
- **Validity:** Data should be collected according to defined business rules and parameters, and should conform to the right format and fall within the right range.

- **Uniqueness:** Uniqueness ensures there are no duplications or overlapping of values across all data sets. Data cleansing and deduplication can help remedy a low uniqueness score.
- **Timeliness:** Timely data is data that is available when it is required. Data may be updated in real time to ensure that it is readily available and accessible.

Thank you