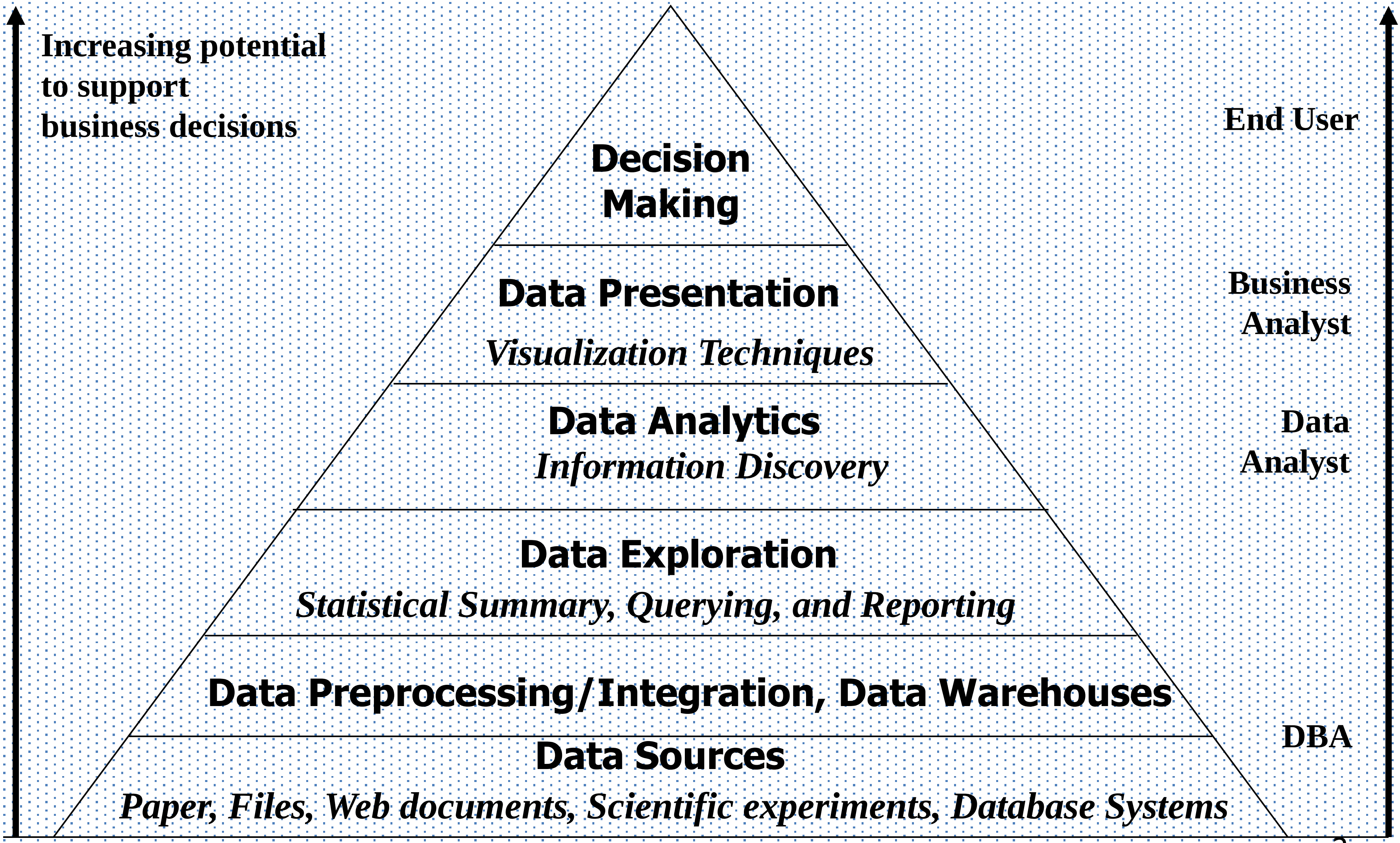
# Matters of Discussion

## Introduction to Data & Analytics

- Getting to Know your data and dataset.
- Analytic case

# Data Analytics in Business Intelligence



Increasing potential to support business decisions

**Decision Making** — End User

**Data Presentation**

*Visualization Techniques* — Business Analyst

**Data Analytics**

*Information Discovery* — Data Analyst

**Data Exploration**

*Statistical Summary, Querying, and Reporting*

**Data Preprocessing/Integration, Data Warehouses** — DBA

**Data Sources**

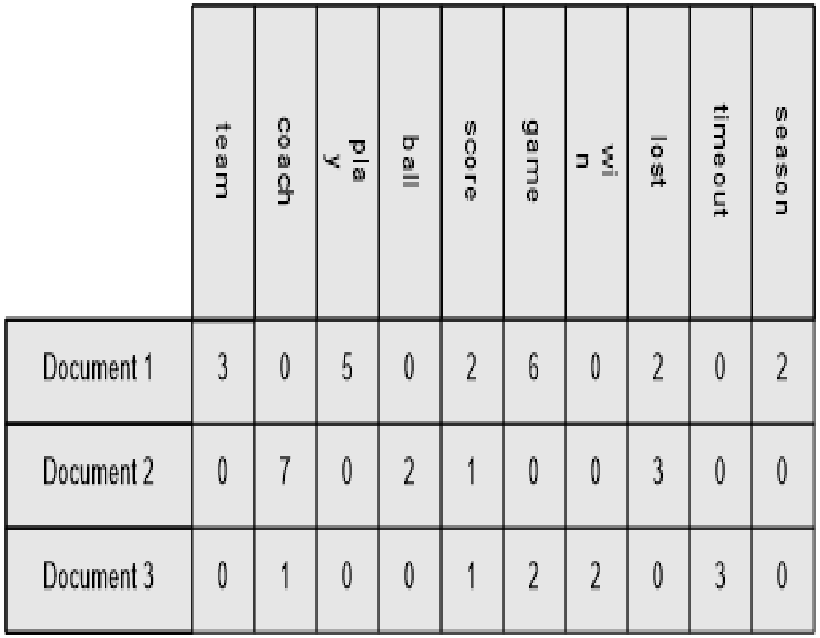
*Paper, Files, Web documents, Scientific experiments, Database Systems*

# Data Analytics: On What Kinds of Data?

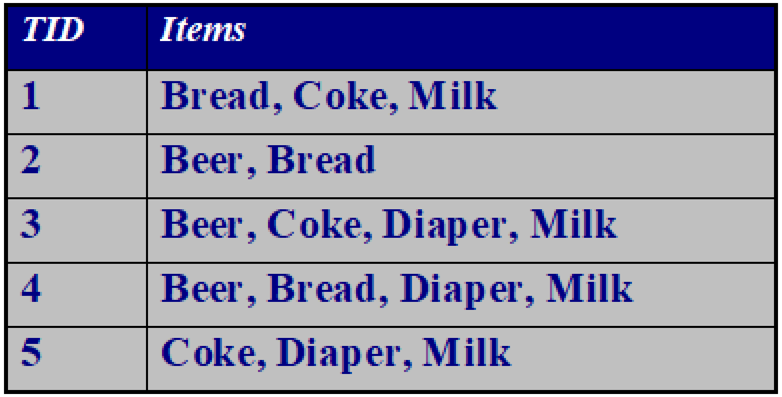
- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. bio-sequences)
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

# Types of Data Sets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video data:



| | team | coach | play | ball | score | game | win | 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 |



| *TID* | *Items* |
|---|---|
| 1 | Bread, Coke, Milk |
| 2 | Beer, Bread |
| 3 | Beer, Coke, Diaper, Milk |
| 4 | Beer, Bread, Diaper, Milk |
| 5 | Coke, Diaper, Milk |

# Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
  - sales database: customers, store items, sales
  - medical database: patients, treatments
  - university database: students, professors, courses
- Also called *samples* , *examples, instances, data points, objects, tuples*.
- Data objects are described by **attributes**.
- Database rows -> objects; columns ->attributes.

# Attributes

- **Attribute (**or **dimensions, features, variables)**: a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

# Attribute Types

- **Nominal:** categories, states, or “names of things”
  - *Hair_color = {auburn, black, blond, brown, grey, red, white}*
  - marital status, occupation, zip codes
  - A variable with values which have no numerical value
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size = {small, medium, large}*, grades, army rankings
  - A variable with values which have no numerical value

# Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C°or F°, calendar dates*
  - No true zero-point
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K° is twice as high as 5 K°).
    - e.g., *temperature in Kelvin, length, counts, monetary quantities*

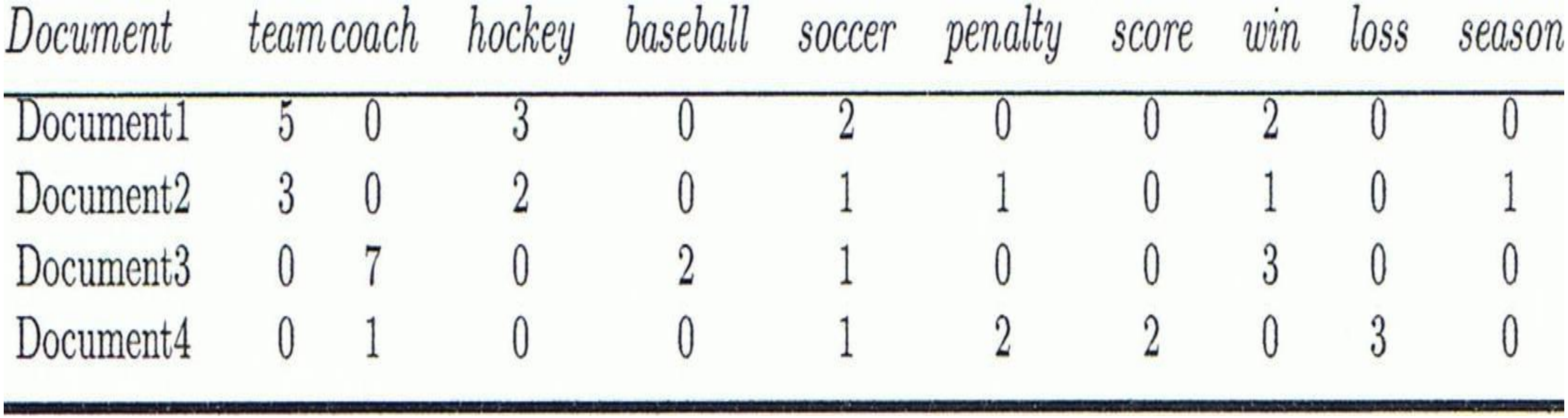
# Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

- **Document database**

- A **document** can be represented by thousands of attributes, each recording the *frequency* of a particular word (such as keywords) or phrase in the document.
- Term frequency (TF) means how often a term occurs in a document.

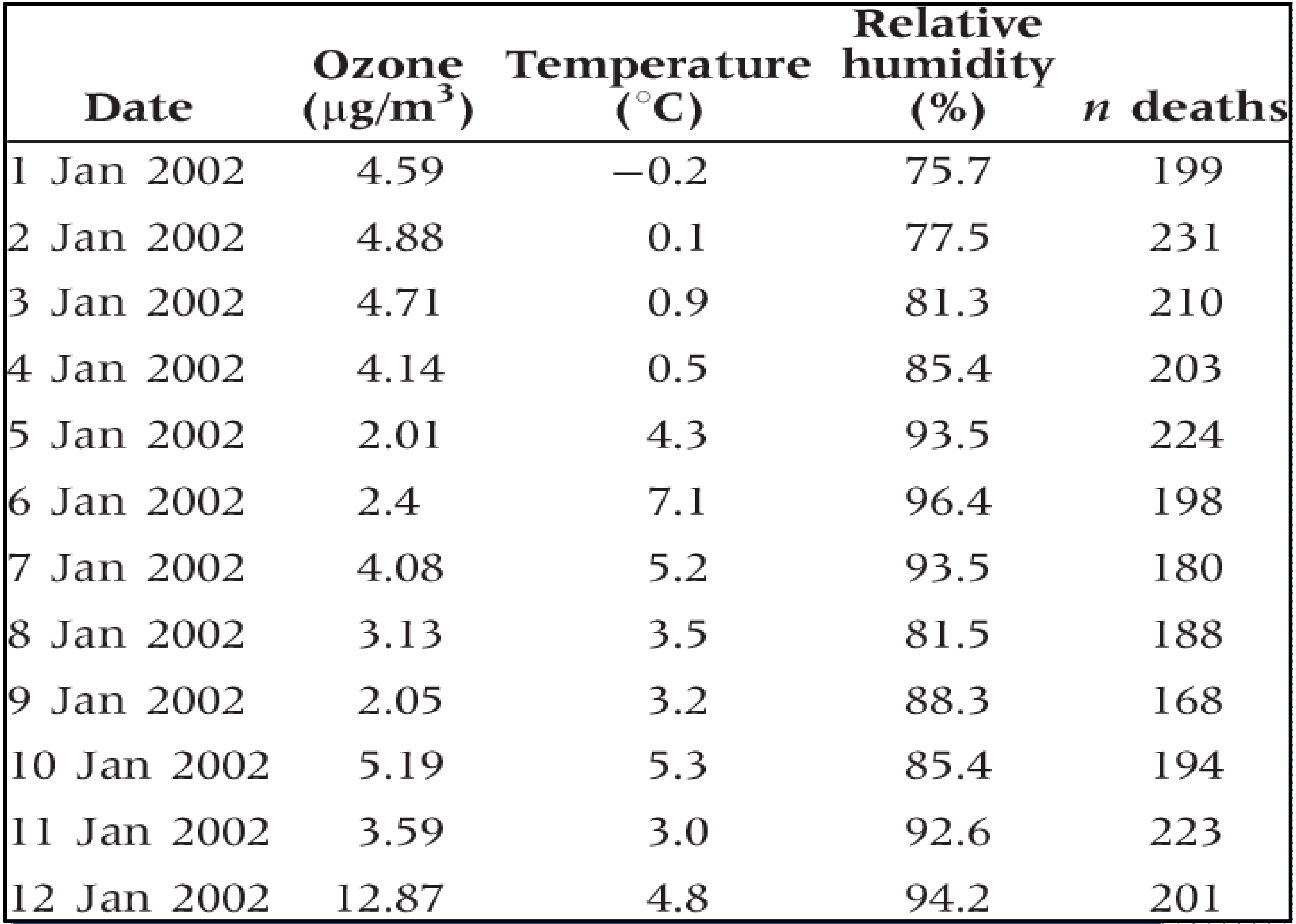
- Term frequency dataset



| Document | team | coach | hockey | baseball | soccer | penalty | score | win | loss | season |
|---|---|---|---|---|---|---|---|---|---|---|
| Document1 | 5 | 0 | 3 | 0 | 2 | 0 | 0 | 2 | 0 | 0 |
| Document2 | 3 | 0 | 2 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| Document3 | 0 | 7 | 0 | 2 | 1 | 0 | 0 | 3 | 0 | 0 |
| Document4 | 0 | 1 | 0 | 0 | 1 | 2 | 2 | 0 | 3 | 0 |

# Time series Data

- A time series is a series of data points indexed (or listed or graphed) in time order.
- Most commonly, a time series is a sequence taken at successive equally spaced points in time.
- Thus it is a sequence of discrete-time data.
- Examples of time series are heights of ocean tides, USD value, market share value, many more..

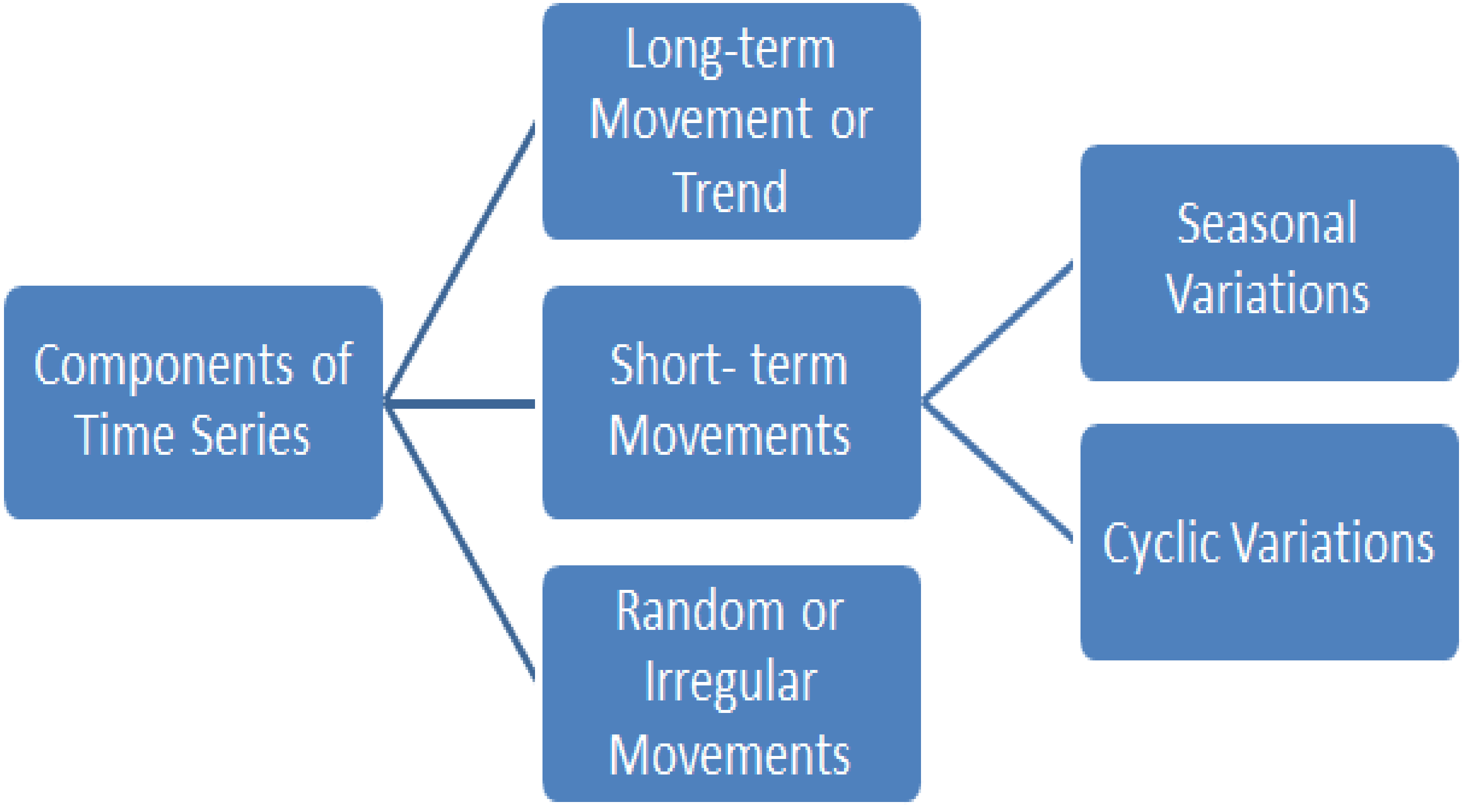


| Date | Ozone ($\mu g/m^3$) | Temperature (°C) | Relative humidity (%) | $n$ deaths |
|---|---|---|---|---|
| 1 Jan 2002 | 4.59 | −0.2 | 75.7 | 199 |
| 2 Jan 2002 | 4.88 | 0.1 | 77.5 | 231 |
| 3 Jan 2002 | 4.71 | 0.9 | 81.3 | 210 |
| 4 Jan 2002 | 4.14 | 0.5 | 85.4 | 203 |
| 5 Jan 2002 | 2.01 | 4.3 | 93.5 | 224 |
| 6 Jan 2002 | 2.4 | 7.1 | 96.4 | 198 |
| 7 Jan 2002 | 4.08 | 5.2 | 93.5 | 180 |
| 8 Jan 2002 | 3.13 | 3.5 | 81.5 | 188 |
| 9 Jan 2002 | 2.05 | 3.2 | 88.3 | 168 |
| 10 Jan 2002 | 5.19 | 5.3 | 85.4 | 194 |
| 11 Jan 2002 | 3.59 | 3.0 | 92.6 | 223 |
| 12 Jan 2002 | 12.87 | 4.8 | 94.2 | 201 |

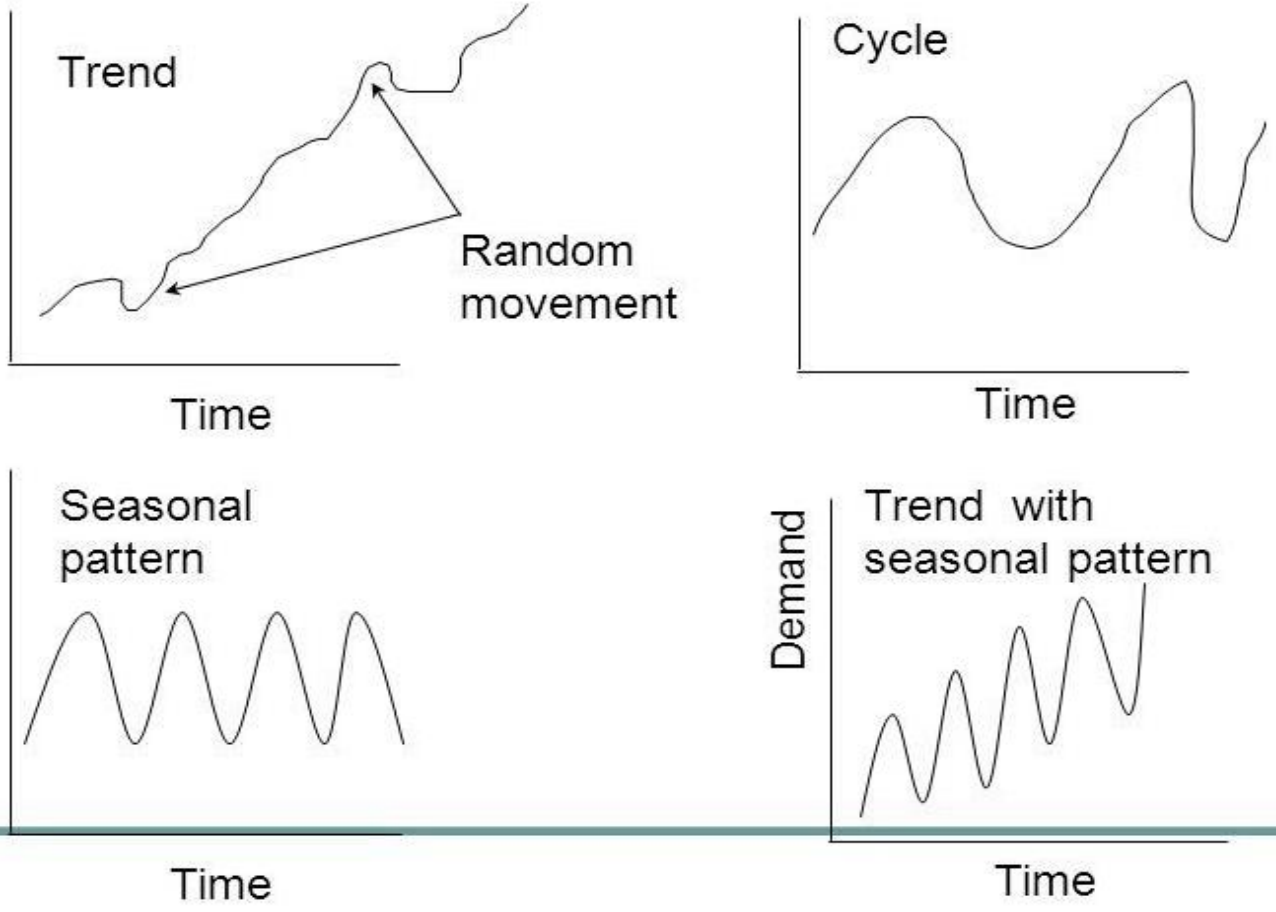
# Example of Time Series Data

| Field | Example topics |
|---|---|
| Economics | Gross Domestic Product (GDP), Consumer Price Index (CPI), S&P 500 Index, and unemployment rates |
| Social sciences | Birth rates, population, migration data, political indicators |
| Epidemiology | Disease rates, mortality rates, mosquito populations |
| Medicine | Blood pressure tracking, weight tracking, cholesterol measurements, heart rate monitoring |
| Physical sciences | Global temperatures, monthly sunspot observations, pollution levels. |

# Time Series Components


Components of Time Series
Long-term Movement or Trend
Short- term Movements
Random or Irregular Movements
Seasonal Variations
Cyclic Variations

# Time Series Components – cont..

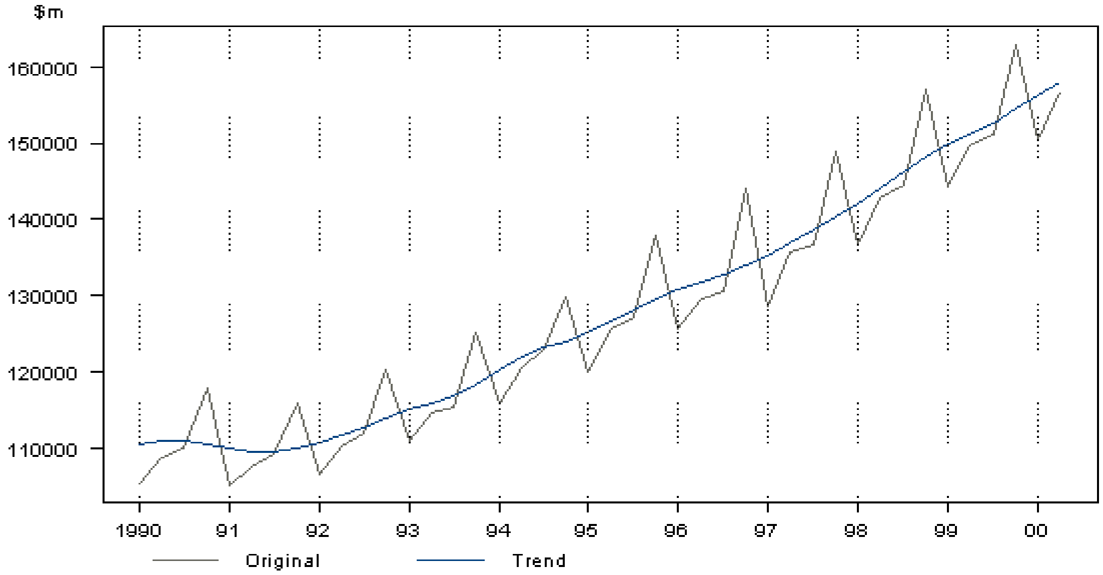


# Time Series Components---cont..

- **Trend**:- general tendency of the data to increase or decrease during a long period of time.

- 'long term' movement in a time series without calendar related and irregular effects.

- population growth, price inflation and general economic changes.



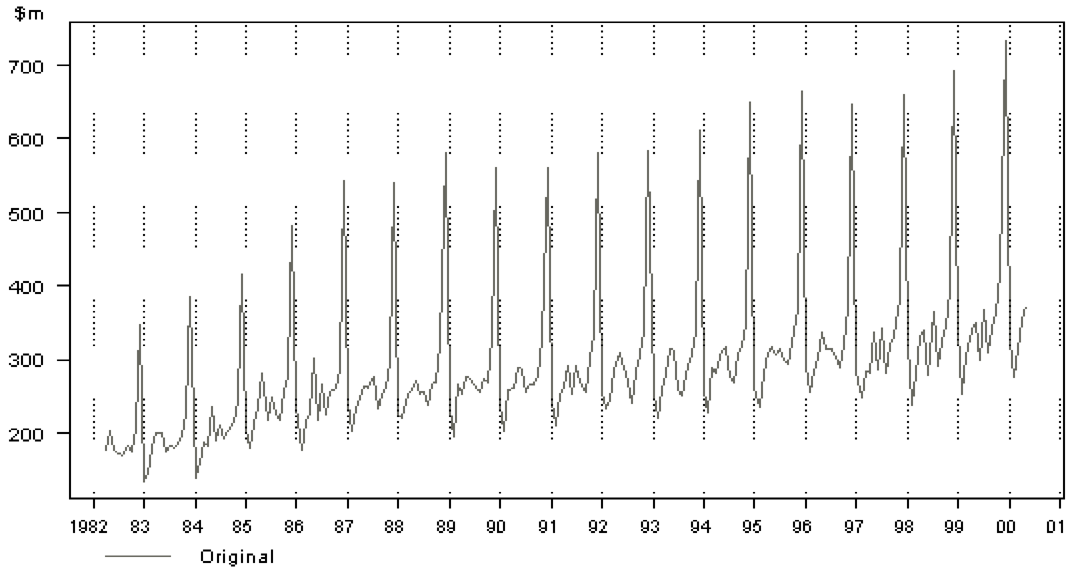
Compiled By: Dr. Nilamadhab Mishra [(PhD- CSIE) Taiwan]

# Time Series Components—cont..

How do we identify **seasonality or seasonal pattern**?

- ❖ With respect to calendar related effects.
- ❖ large seasonal increase in December retail sales due to Christmas shopping.
- ❖ magnitude of the seasonal component increases over time , as the trend does.

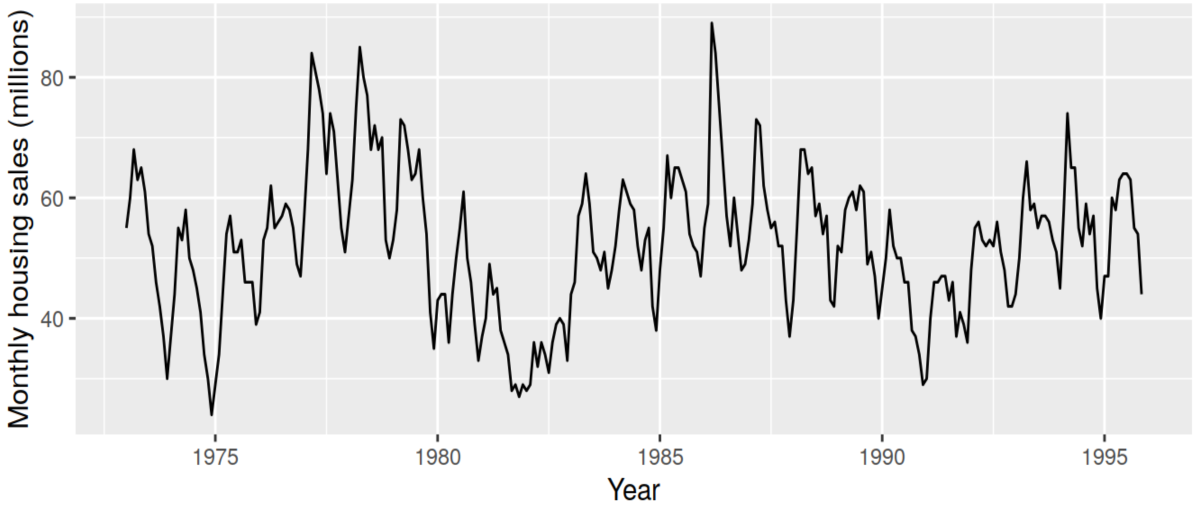
periodic time series



# Time Series Components--Cycle

- A cyclic pattern exists when data exhibit rises and falls that are not of fixed period.
- The duration of these fluctuations is usually of at least 2 years.
- **If the fluctuations are not of fixed period then they are cyclic else seasonal.**

**Within the time interval,
the fluctuations
are not fixed**



# Time Series Components-summary

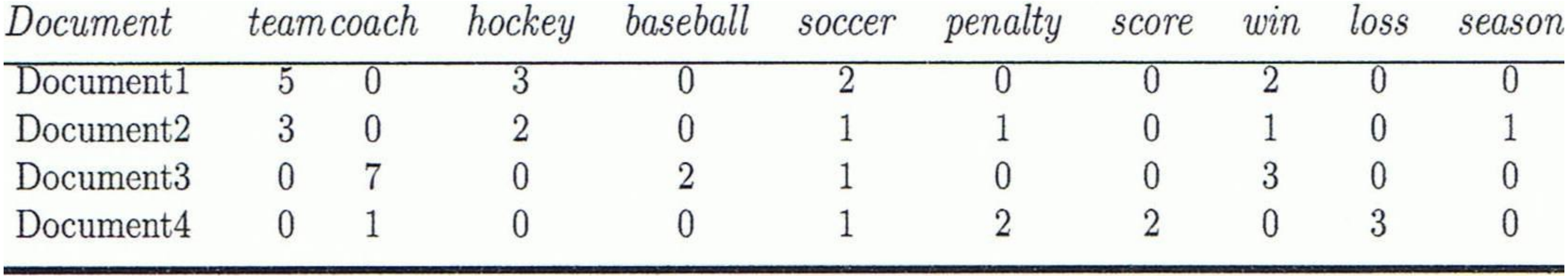
**These components are defined as follows:**

- Level: The average value in the series.
- Trend: The increasing or decreasing value in the series.
- Seasonality: The repeating short-term cycle in the series.
- Cyclic: data exhibit rises and falls that are not of fixed period
- Noise: The random variation in the series.

# Specific Data Analytic case

# Cosine Similarity in Data Analytic Apps

- A **document** can be represented by thousands of attributes, each recording the *frequency* of a particular word (such as keywords) or phrase in the document.



| *Document* | *team* | *coach* | *hockey* | *baseball* | *soccer* | *penalty* | *score* | *win* | *loss* | *season* |
|---|---|---|---|---|---|---|---|---|---|---|
| Document1 | 5 | 0 | 3 | 0 | 2 | 0 | 0 | 2 | 0 | 0 |
| Document2 | 3 | 0 | 2 | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| Document3 | 0 | 7 | 0 | 2 | 1 | 0 | 0 | 3 | 0 | 0 |
| Document4 | 0 | 1 | 0 | 0 | 1 | 2 | 2 | 0 | 3 | 0 |

- Other vector objects: gene features in micro-arrays, …
- **Applications: information retrieval, biologic taxonomy, gene feature mapping, ...**
- Cosine measure: If $d_1$ and $d_2$ are two vectors (e.g., term-frequency vectors), then

$$\mathbf{cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1||\ ||d_2||}\ ,$$

where • indicates vector dot product, $||d||$: the length of vector $d$

# Example: Cosine Similarity

- $\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2||$ ,
  where • indicates vector dot product, $||d|$: the length of vector $d$

- Ex: Find the **similarity** between documents 1 and 2.

$d_1 =$ (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)
$d_2 =$ (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

$d_1 \bullet d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$

$||d_1|| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5}=(42)^{0.5} = 6.481$

$||d_2|| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5}=(17)^{0.5} = 4.12$

$\cos(d_1, d_2) = 0.94$

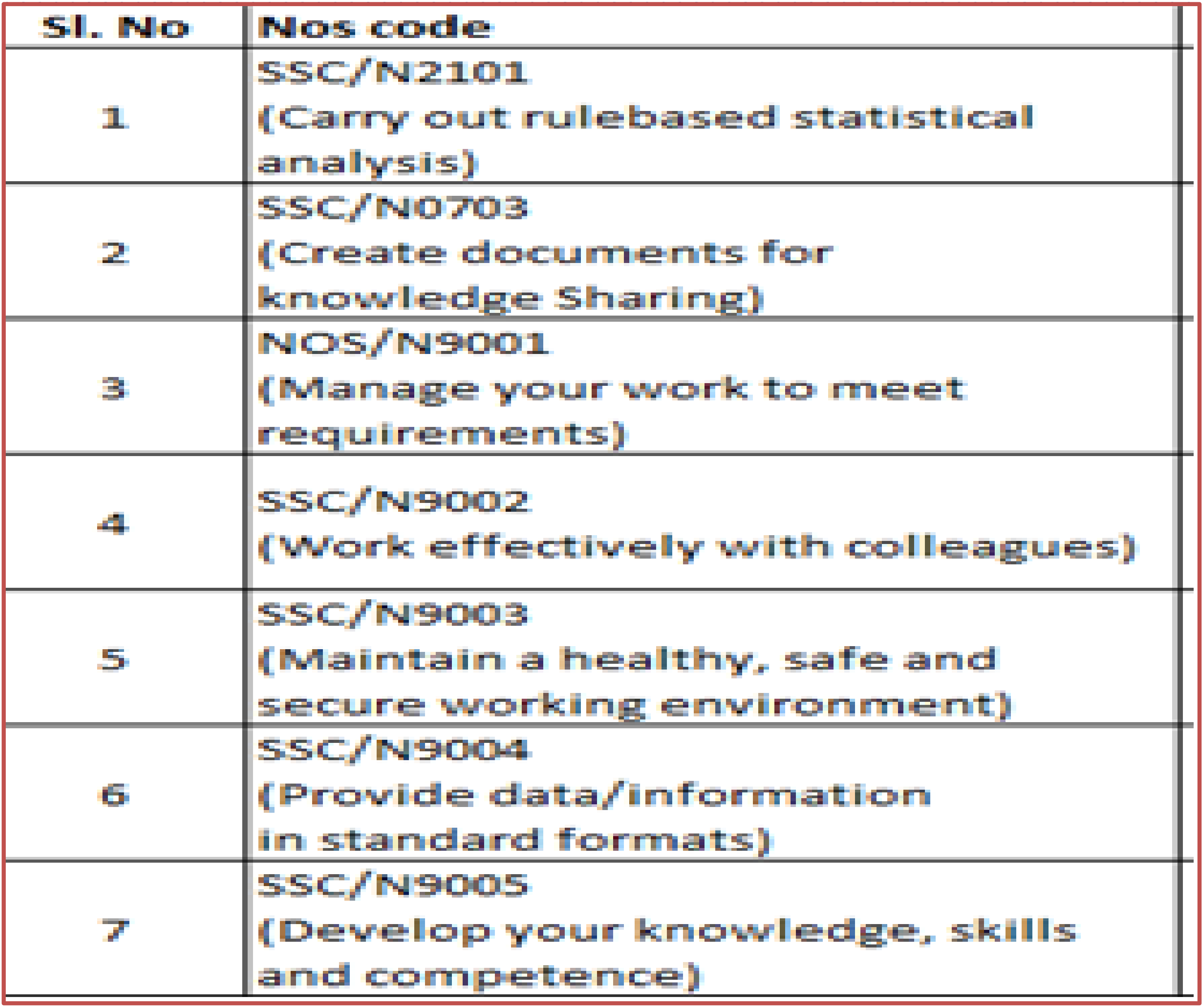
# TASK FOR YOU—A2

1. Investigate the Attribute or dimensions or features or variables with a suitable scenario and prepare your critical report?

- Nominal
- Binary
- ordinal
- Numeric: quantitative

# V.V.I

1. Investigate the numerous time series components in context to the business analytics and application modeling.
2. Types of Data Sets and Data Object concept
3. Document database- Term frequency dataset
4. Cosine Similarity in Data Analytic Apps.



| Sl. No | Nos code |
|---|---|
| 1 | SSC/N2101<br>(Carry out rulebased statistical analysis) |
| 2 | SSC/N0703<br>(Create documents for knowledge Sharing) |
| 3 | NOS/N9001<br>(Manage your work to meet requirements) |
| 4 | SSC/N9002<br>(Work effectively with colleagues) |
| 5 | SSC/N9003<br>(Maintain a healthy, safe and secure working environment) |
| 6 | SSC/N9004<br>(Provide data/information in standard formats) |
| 7 | SSC/N9005<br>(Develop your knowledge, skills and competence) |

# Cheers For the Great Patience!

Query Please?