SRI RAMACHANDRA

INSTITUTE OF HIGHER EDUCATION AND RESEARCH

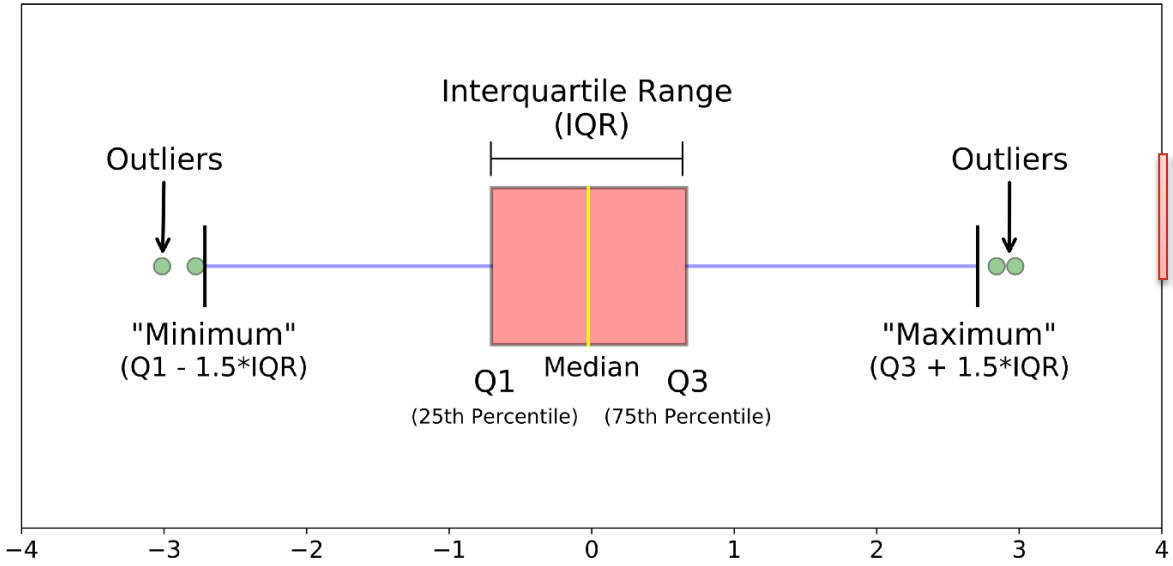
(Category - I Deemed to be University) Porur, Chennai

SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY

CSE 320 Data Mining

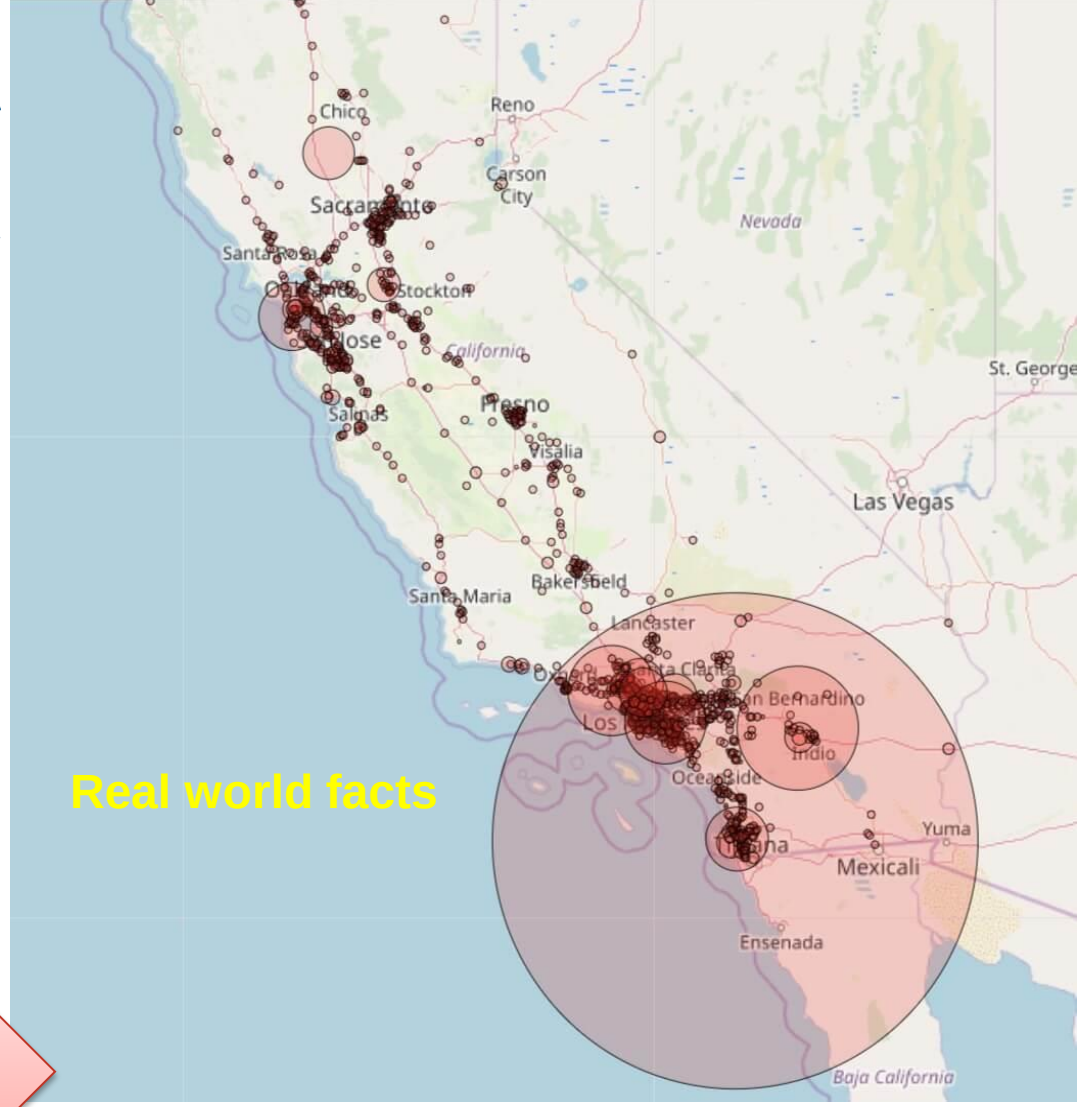
YEAR – III Quarter 1
B. TECH CSE (AIML)

Course Faculty : Dr. Jayanthi G
Assistant Professor,
Faculty of Engineering and Technology



OUTLIER

Data Insight



1.1 Data Objects and Attributes

1.1.1 Types of Datasets

- Record
- Graph and Network
- Ordered
- Spatial
- Image and Multimedia
- Datasets in R

1.1.2 Characteristics of structured data

1.1.3 What is data object?

1.1.4 What is attribute?

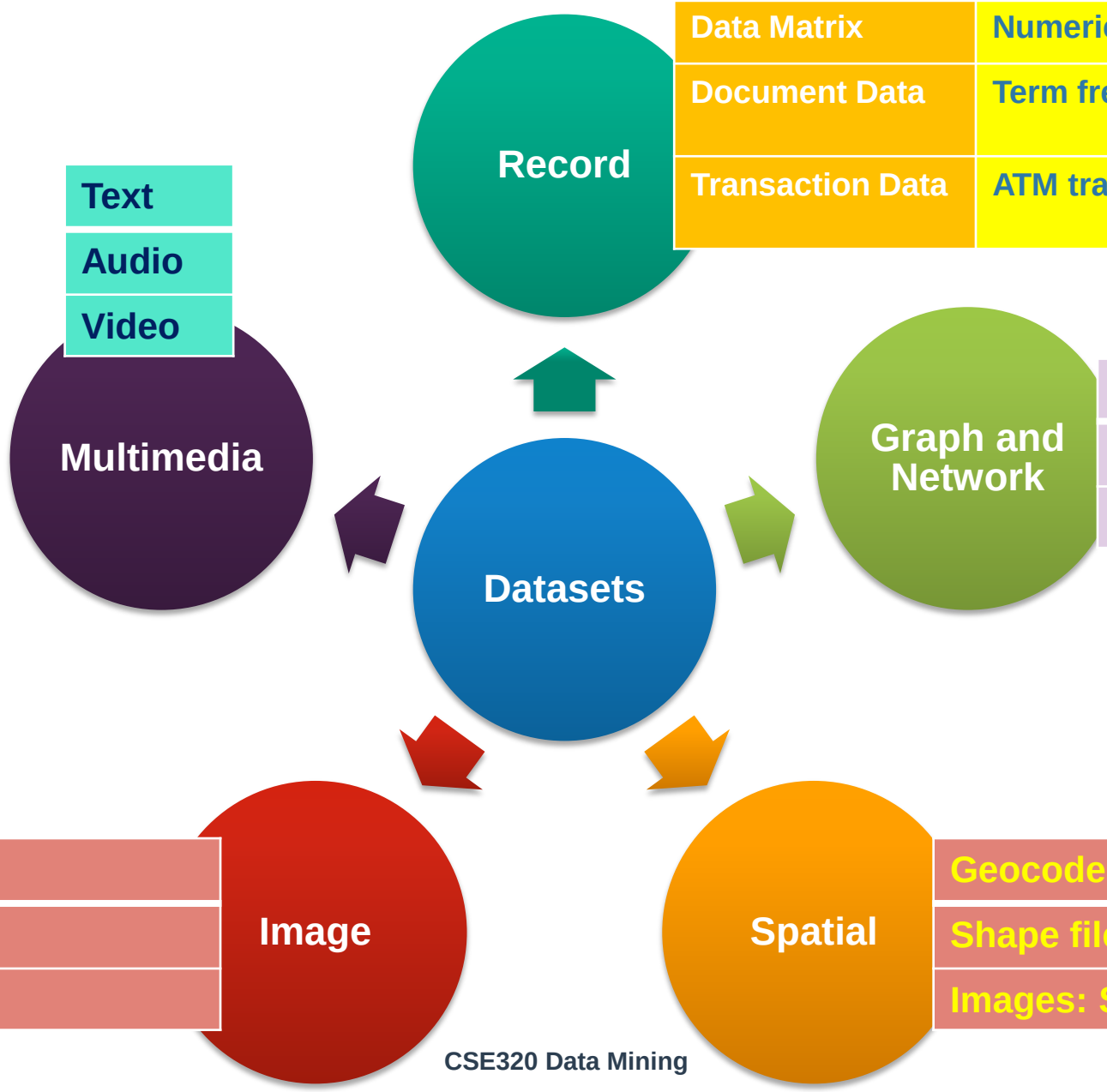
Types of attributes (Nominal, Binary, Ordinal,) –Tutorial

Numeric (Integer or real valued, Interval, ratio) – Tutorial

Discrete Versus Continuous - Tutorial

1.1.1 Types of Data sets

Relational Records	A row in a relational table with fields and data values.
Data Matrix	Numerical matrix, cross tabs, QR code,
Document Data	Term frequency vector
Transaction Data	ATM transaction history



Text
Audio
Video

World Wide Web
Social or Information networks
Molecular Structures

Medical Imaging
Fractal
TIFF, JPEG, PNG, GIF

Geocoded : point coordinates
Shape files: Geometry
Images: Satellite Imagery

1.1.1 Types of Data sets – Record

- **Relational records**

Employee Id	First Name	Last Name	Date of Birth		
			Day	Month	Year
1490	Suraj	Mohan	05	April	1987
1235	Kamala	Vasudev	24	May	1990
1678	Kamal	Prasad	06	Jan	1989
1541	Rajan	Rajesh	25	June	1980
1798	Neela	Kumar	11	July	1978

- **Data matrix, e.g., numerical matrix, crosstabs, QR Code, Data matrix**

Cross Tabulation

	Base	Age					
		Under 18	18-24	25-34	35-44	45-54	55+
Base	204	59 29%	43 21%	38 19%	36 18%	20 10%	8 4%
Frequency of visit							
Daily	18 9%	9 4%	5 2%	4 2%	- -	- -	- -
Twice a week	35 17%	11 5%	8 4%	8 4%	7 3%	- -	1 0%
Weekly	64 31%	16 8%	8 4%	16 8%	16 8%	4 2%	4 2%
Monthly	87 43%	23 11%	22 11%	10 5%	13 6%	16 8%	3 1%

Crosstab

Matrix

$$\begin{matrix} 4 & 5 \\ 2 & 0 \end{matrix}$$

QR Code

Data Matrix

- **Document data: text documents: term-frequency vector**

Document	Team	Coach	Hockey	Baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	2	3	6	4	2	1	0
Document2	3	0	2	6	2	7	0	1	2	0
Document3	0	8	0	4	6	2	3	0	5	1
Document4	0	9	2	5	3	1	2	4	4	2
Document5	1	1	8	4	3	2	2	4	3	1

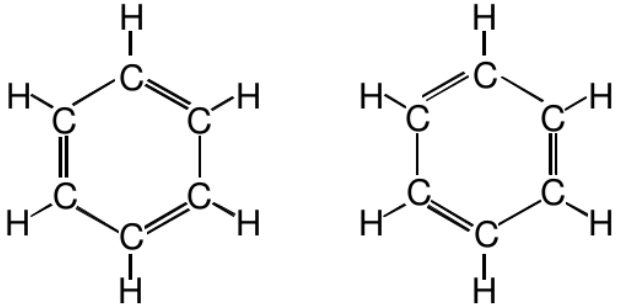
- **Transaction data**

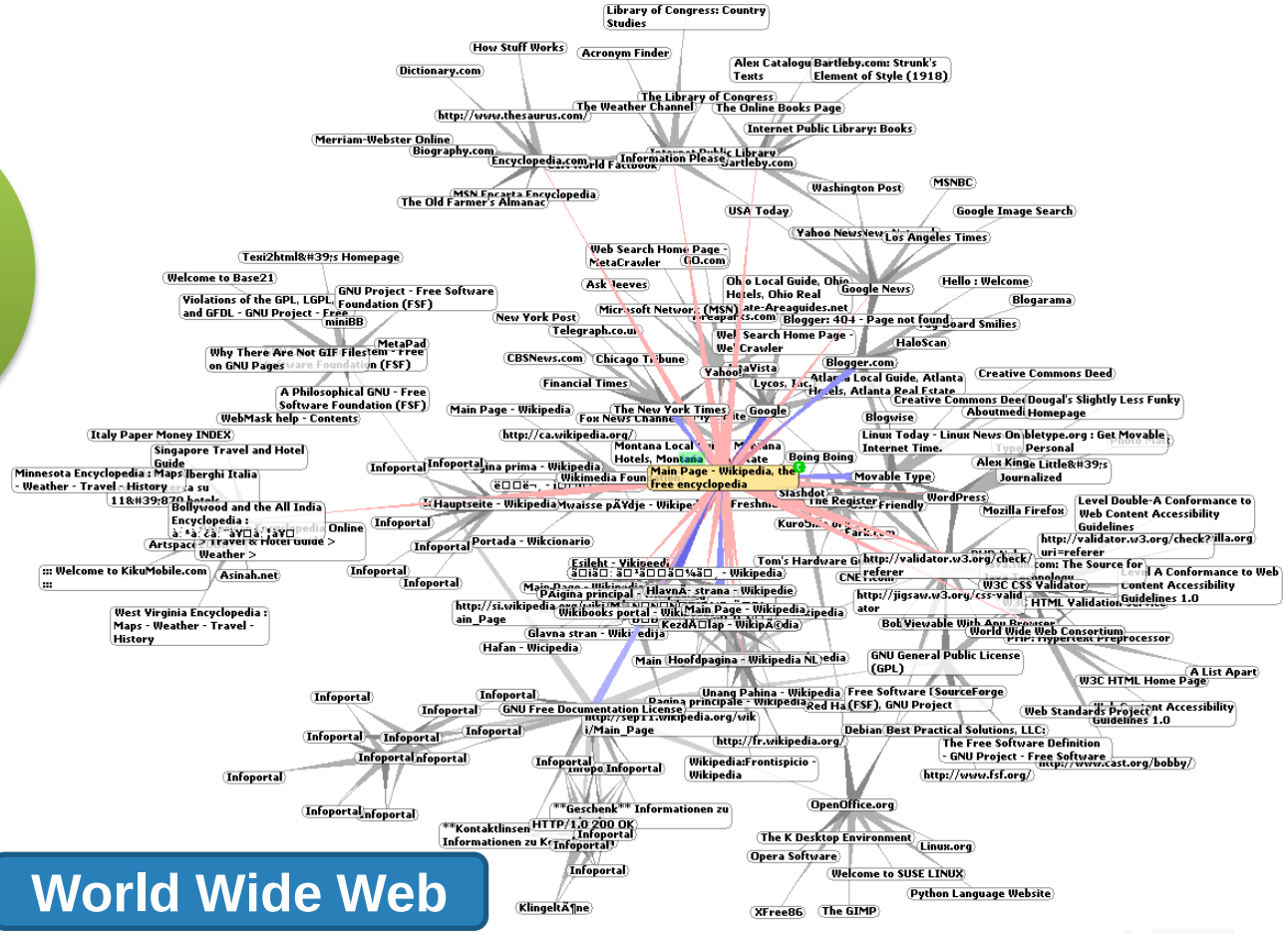
ATM Name	Transaction Date	No Of Withdrawals	No Of XYZ Card Withdrawals	No Of Other Card Withdrawals	Total amount Withdrawn	Amount withdrawn XYZ Card	Amount withdrawn Other Card	Weekday	Festival Religion	Working Day	Holiday Sequence
Big Street ATM	1/1/2011	50	20	30	123800	41700	82100	Saturday	H	H	WHH
Mount Road ATM	1/1/2011	253	67	186	767900	270900	497000	Saturday	C	H	WHH
Airport ATM	1/1/2011	98	56	42	503400	347700	155700	Saturday	C	H	WHH
KK Nagar ATM	1/1/2011	265	159	106	945300	532600	412700	Saturday	C	H	WHH
Christ College ATM	1/1/2011	74	25	49	287700	532600	139500	Saturday	C	H	WHH

1.1.1 Types of Data sets – Graph and Network

Social Networking and Information

Molecular Structures

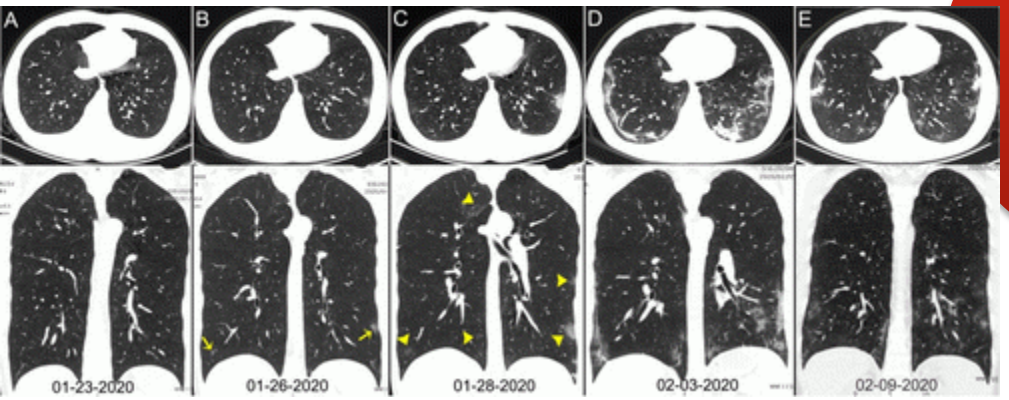




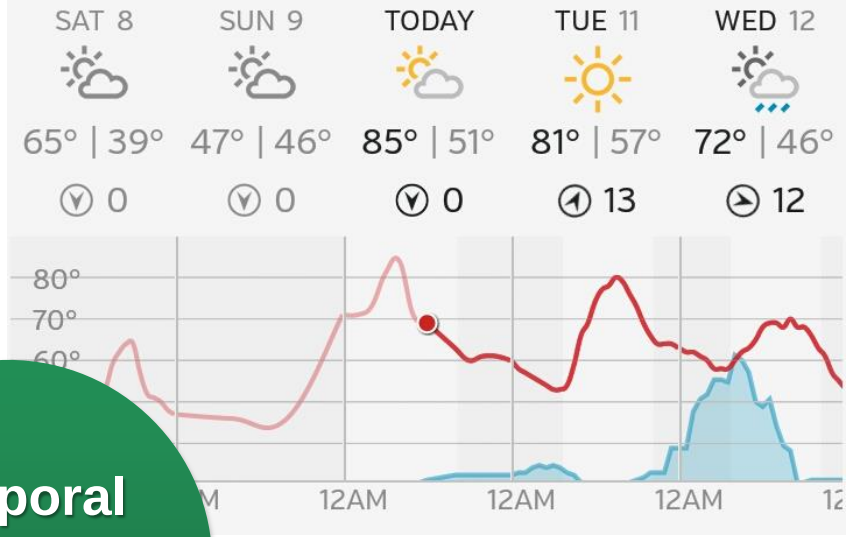
World Wide Web

http://wikipedia.org, we can follow links to eventually reach millions of webpages from across the globe.

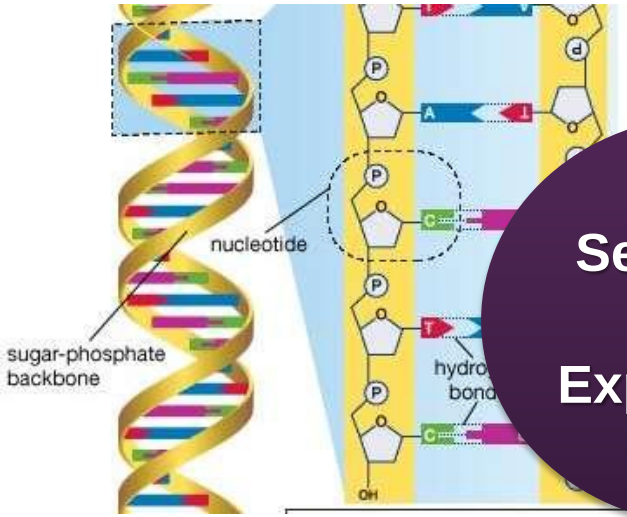
1.1.1 Types of Data sets – Ordered



Imaging [Medical, Geospatial]



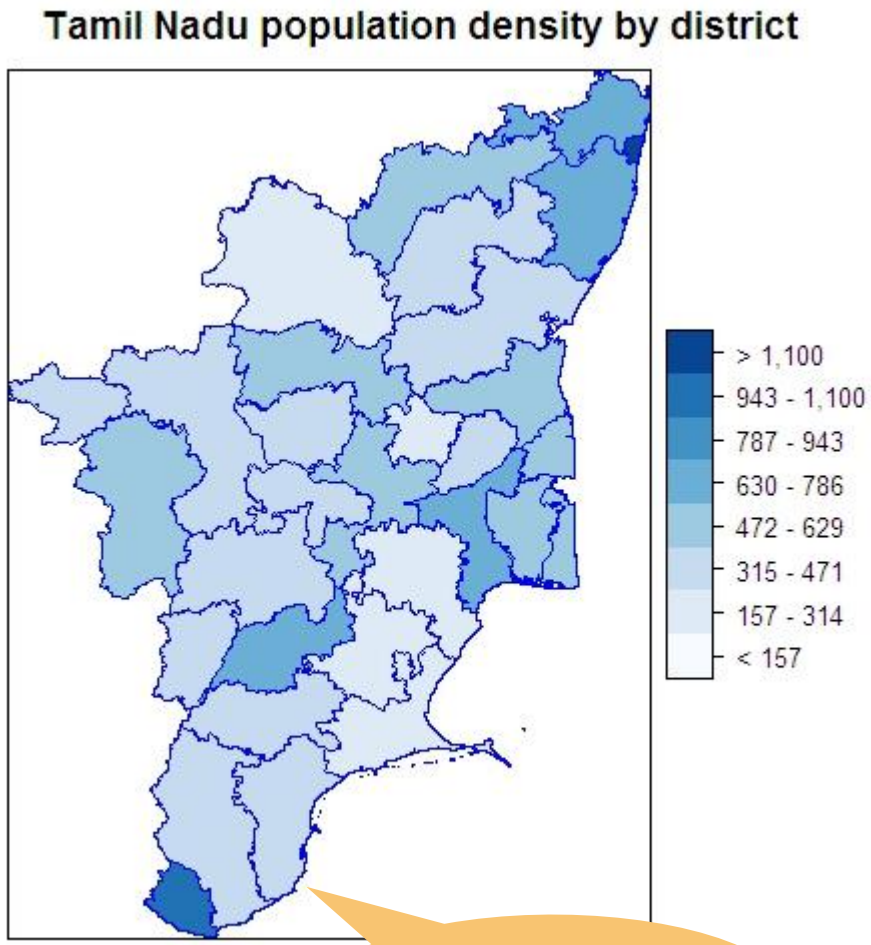
Ordered



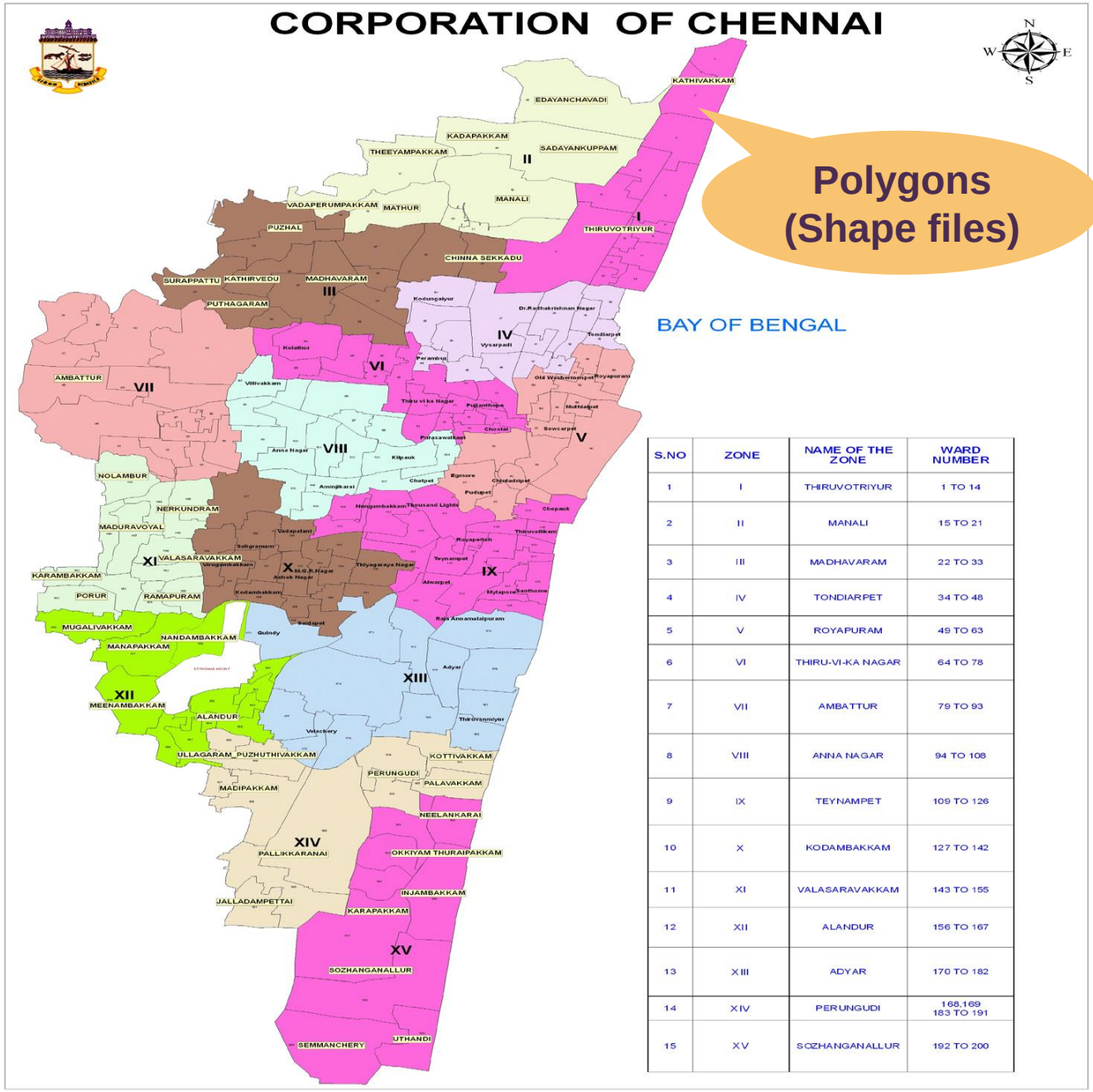
Sequential [Gene Expression]

Temporal [Time Series]

1.1.1 Types of Data sets – Spatial



Spatial Maps
(Shape files)



Polygons
(Shape files)

S.NO	ZONE	NAME OF THE ZONE	WARD NUMBER
1	I	THIRUVOTRIYUR	1 TO 14
2	II	MANALI	15 TO 21
3	III	MADHAVARAM	22 TO 33
4	IV	TONDIARPET	34 TO 48
5	V	ROYAPURAM	49 TO 63
6	VI	THIRU-VI-KA NAGAR	64 TO 78
7	VII	AMBATTUR	79 TO 93
8	VIII	ANNA NAGAR	94 TO 108
9	IX	TEYNAMPET	109 TO 126
10	X	KODAMBAKKAM	127 TO 142
11	XI	VALASARAVAKKAM	143 TO 155
12	XII	ALANDUR	156 TO 167
13	XIII	ADYAR	170 TO 182
14	XIV	PERUNGUDI	168,169 183 TO 191
15	XV	SOZHANGANALLUR	192 TO 200

1.1.1 Datasets in R

```
data("readingSkills")
head(readingSkills)
data("Groceries")
head(Groceries)
```

nativeSpeaker	a
1 yes	5 2
2 yes	6 2
3 no	11 3
4 yes	7 2
5 yes	11 3
6 yes	10 3

"transactions in sparse format with 6 transactions (rows) and 169 items (columns)"

```
data("Income")
head(Income)
```

transactions in sparse f (rows) and 50 items (colu

```
str(Income)
Formal class 'transactions' [package "arules"] with 3
slots
  ..@ data       :Formal class 'ngCMatrix' [package
"Matrix"] with 5 slots
  .. .. ..@ i       : int [1:96264] 1 2 4 10 12 17 23
26 28 30 ...
  .. .. ..@ p       : int [1:6877] 0 14 28 42 56 70 84
98 112 126 ...
  .. .. ..@ Dim     : int [1:2] 50 6876
  .. .. ..@ Dimnames:List of 2
  .. .. .. ..$ : NULL
  .. .. .. ..$ : NULL
  .. .. ..@ factors : list()
  ..@ itemInfo   :'data.frame':   50 obs. of  3
variables:
  .. ..$ labels   : chr [1:50] "income=$0-$40,000"
"income=$40,000+"
  .. ..$ variables: Factor w/ 14 levels "age","dual
incomes",..: 6 6 12 12 8 8 8 8 8 1 ...
  .. ..$ levels   : Factor w/ 48 levels "0","$0-
$40,000",..: 2 10 28 22 29 16 19 47 42 6 ...
  ..@ itemsetInfo:'data.frame':   6876 obs. of  1
variable:
  .. ..$ transactionID: chr [1:6876] "2" "3" "4" "5"
...
```

1.1.2 Characteristics of Structured Data

Dimensionality : Curse of dimensionality

Sparsity : Only presence counts

Resolution: Patterns depend on the scale

Distribution: Centrality and dispersion

```r
data("readingSkills")
head(readingSkills)

data("Groceries")
head(Groceries)

data("Income")
head(Income)
str(Income)
```

```
> data("readingSkills")
> head(readingSkills)
  nativeSpeaker age shoeSize    score
1           yes   5 24.83189 32.29385
2           yes   6 25.95238 36.63105
3            no  11 30.42170 49.60593
4           yes   7 28.66450 40.28456
5           yes  11 31.88207 55.46085
6           yes  10 30.07843 52.83124
> data("Groceries")
> head(Groceries)
transactions in sparse format with
 6 transactions (rows) and
 169 items (columns)
> data("Income")
> head(Income)
transactions in sparse format with
 6 transactions (rows) and
 50 items (columns)
```

```
data("BostonHomicide")
head(BostonHomicide)
head(BostonHousing)

data("scPublications")
data("Epub")
data("AirPassengers")
data("PhillipsCurve")
data("GermanM1")
```

```
> str(Income)
Formal class 'transactions' [package "arules"] with 3 slots
  ..@ data       :Formal class 'ngCMatrix' [package "Matrix"]
with 5 slots
  .. .. ..@ i       : int [1:96264] 1 2 4 10 12 17 23 26 28 30
...
  .. .. ..@ p       : int [1:6877] 0 14 28 42 56 70 84 98 112
126 ...
  .. .. ..@ Dim     : int [1:2] 50 6876
  .. .. ..@ Dimnames:List of 2
  .. .. .. ..$ : NULL
  .. .. .. ..$ : NULL
  .. .. ..@ factors : list()
  ..@ itemInfo   :'data.frame':   50 obs. of  3 variables:
  .. ..$ labels   : chr [1:50] "income=$0-$40,000"
"income=$40,000+" "sex=male" "sex=female" ...
  .. ..$ variables: Factor w/ 14 levels "age","dual
incomes",..: 6 6 12 12 8 8 8 8 8 1 ...
  .. ..$ levels   : Factor w/ 48 levels "0","$0-$40,000",..: 2
10 28 22 29 16 19 47 42 6 ...
  ..@ itemsetInfo:'data.frame':   6876 obs. of  1 variable:
  .. ..$ transactionID: chr [1:6876] "2" "3" "4" "5" ...
```

```
> data("BostonHomicide")
> head(BostonHomicide)
  homicides population populationBM ahomicides25 ahomicides35 unemploy season year
1         2     228465        12977            2            1     20.2    Jan 1992
2         1     228465        12977            1            0     20.2    Feb 1992
3         1     228465        12977            5            1     20.2    Mar 1992
4         1     228465        12977            1            0     20.2    Apr 1992
5         3     228465        12977            3            0     20.2    May 1992
6         3     228465        12977            6            2     20.2    Jun 1992
```

```
> head(BostonHousing)
     crim zn indus chas   nox    rm  age    dis rad tax ptratio      b lstat medv
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33 36.2
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7
> data("scPublications")
> data("Epub")
> data("AirPassengers")
> data("PhillipsCurve")
> data("GermanM1")
>
```

1.1.3 What is data object?

- The raw data sets are the source of data objects.
- A data object is an entity in the real world.
- There are also known as samples, instances, data points, tuples etc.,

Each row is a data object, Columns in the table represents attributes.

Example: Relational Database such as
- **Sales Database**
- **Medical Database**
- **University Database**
- **Geospatial Database**

- **Sales Database : items, Customer transactions, Turnover**
- **Medical Database: Patients, Symptoms, diagnosis,**
- **University Database: Programme, Courses, Students, Professors**
- **Geospatial Database: Spatial coordinate Lat, Long, Geometry, Altitude, demography**

How data objects are represented in R?

1. Array
2. Matrix
3. Data Frame
4. List
5. Factor
6. Vector

In **R**, variables are assigned to objects rather than data types. This enables the use of basic types in a different way during data manipulation. The **class()** function is used to extract attribute information of R objects.

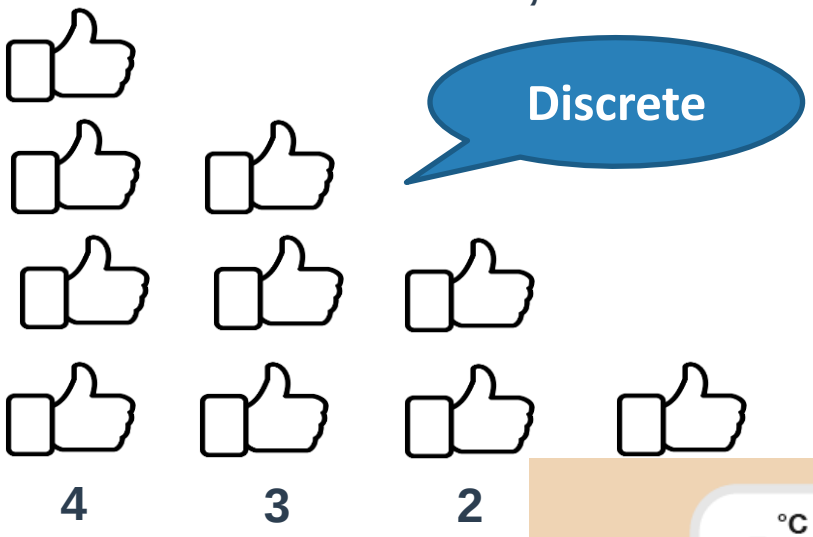
1.1.3 Understanding types of variables?

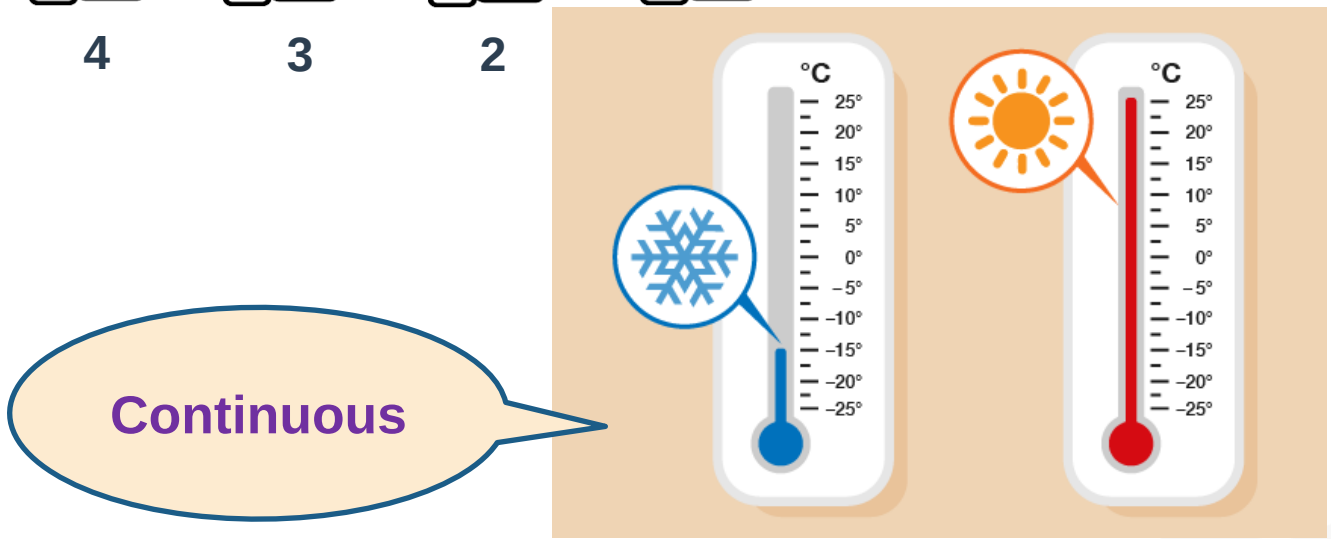
Data is a specific measurement of a variable.

Attribute (or dimensions, features, variables): a data field, representing a specific property or characteristics or feature of a data object. ***E.g., Unique ID. Name. Admission Year***

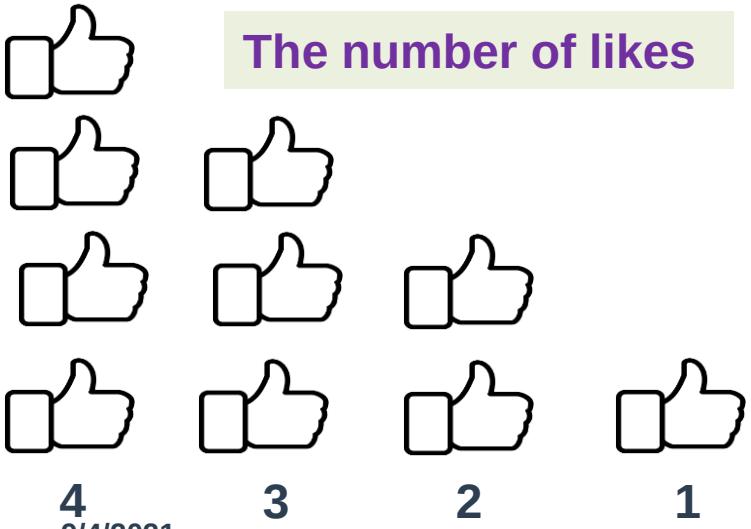
- **Variable**
  - **Quantitative**
    - Discrete
    - Continuous
  - **Categorical**
    - Nominal
    - Ordinal
    - Interval
    - Ratio scaled

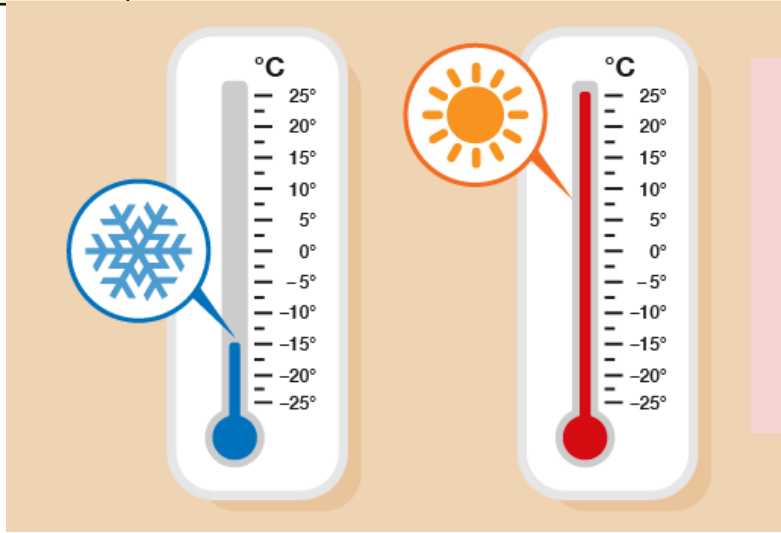
Discrete : Count of individual data items.
Example
Number of students in a class
Number of courses offered
Number of participants in an event
Number of likes, dislikes





Discrete Versus Continuous Variable			
Type of Variable	**Representation**	**Examples**	**Values**
Discrete (integer)	Count of individual items or values Binary attributes are special case	• The strength of a classroom • The different biological species in a Zoological Park • Pin codes • Profession • the set of words in a collection of documents	• The number of students • The number of species • 600 001, 600 116 • Professor, Associate, Assistant • {“the”, “an”, “as”, “of”}
Continuous (Ratio)	Count of non-finite values	• Distance • Speed • Calendar Date • Time • Temperature, Height, Weight	570 meters 40 Km/Hr 12-07-2021 09:00 AM -15°C , 25°C ; 150.5 cm; 45.5 Kg





- Continuous variables are real numbers.
  - E.g., temperature, height, or weight
- Practically, real values can only be measured and represented using a finite number of digits

Qualitative Variable (Categorical)	Attribute	Example
Nominal	Categories, States, names of things	Color = { red, green, yellow} Zip codes, Profession, Martial Status.
Symmetric Nominal	Outcome has only 2 states (1 and 0) equally important	Gender
Asymmetric Nominal	Outcome not equally important (+ve, -ve)	Tested COVID Positive, Tested COVID Negative
Ordinal	Ranking – Meaningful ordering exists while magnitude between successive values not known.	Grade Army ranking Size = { small, medium, large, Extra Large}
Interval [integer or real]	Measured on a scale of equal-sized units Values have order, No true zero-point	E.g., *temperature in C˚or F˚, calendar dates*
Ratio Scaled [integer or real]	Inherent **zero-point, o**rder 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*

NOMINAL

- Nominal level is the most basic level of measurement
- Nominal is also known as categorical or qualitative.
- Data can only be categorised.
- Examples of nominal variables: Gender, preferred type of chocolate, colour. These are descriptions or labels with no sense of order.
- These are descriptions or labels with no sense of order.
- Nominal values can be stored as a word, or text, or can be given a numerical code. However the number does not imply any order.
- To summarise nominal data we use a frequency or percentage.

Variable	Level of measurement : Nominal						
Colour							
Descriptive Statistics							
Frequency	1/7	1/7	1/7	1/7	2/7		1/7
Percentage (%)	14.2857	14.2857	14.2857	14.2857	28.57		14.2857
Mean	Cannot calculate mean or average value of nominal data						

Data Variable

– quantity that varies with measures

- **Continuous**
  - Predict a contiguous target variable (dependent variable) from one or multiple independent variables.
  - **Note : The best choice of regression analysis is with naturally-occurring (non-manipulated) variables, rather than variables that have been manipulated through experimentation.**
- **Discrete**
  - Information that can only take certain values.
- **Categorical**
  - In statistics it is a variable that can take on one of a limited, and usually fixed, a number of possible value.

Activity 1: Match the best choice of graph for the data below.

Chart to show a company's profit over a number of years.
Chart to show favorite drink chosen by customers in a shopping center.
Chart to show the temperature on each day of the week.
Chart to show percentage of each sale of ticket type at a concert.

Variables – Qualitative (Categorical), & Quantitative (Discrete, Continuous)

-The best Chart type and other suitable choices

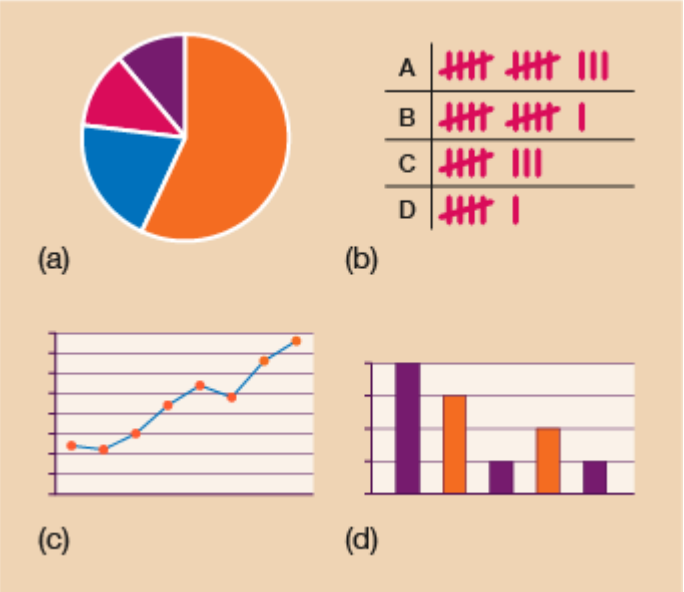
Answer

Chart to show a company's profit over a number of years.
The **best choice** here is (**d**) the bar chart as it can show the profit clearly year by year.

Chart to show favorite drink chosen by customers in a shopping center.
The **best choice** here is (**b**) the tally chart since you can add to this data as each customer makes their choice. A bar or pie chart would also be suitable.

Chart to show the temperature on each day of the week.
The **only choice here** is (**c**) the line graph as it shows how the temperature changes over time.

Chart to show percentage of each sale of ticket type at a concert.
The **best choice** here is probably (**a**) the pie chart since it shows clearly the breakdown of each type of ticket sale. A bar chart would also represent the data suitably.



Quantitative(Continuous, Discrete) and Qualitative(Categorical) - Level of Measurements (scale)

Nominal

- categorical variable scale
- labeling variables into distinct classifications
- No quantitative value or order
- It's simplest measurement scale
- Example: A customer survey
- Which brand of smartphones do you prefer?" Options : "Apple"- 1 , "Samsung"-2, "OnePlus"-3.

Ordinal

- variable measurement scale.
- simply depict the order of variable and not the difference.
- A relative position of variables.
- used in market research.
- Example : How satisfied are you with our services?
- **Very Unsatisfied – 1**
- **Unsatisfied – 2**
- **Neutral – 3**
- **Satisfied – 4**
- **Very Satisfied – 5**

Interval

- A numerical scale. The order of variables is known.
- Data analytics - descriptive statistics, mean, median, or mode.
- correlation, and regression analysis.
- **The only drawback** is there is no pre-decided starting point or a true zero value that represents the absence of the property being measured (**e.g., no money, no behavior, none correct**).
- Example: **Calendar years and time.**
- **What is your family income?**
- **What is the temperature in your city?**

Ratio

- A variable measurement scale.
- The order of variables and the difference between variables are known along with information on the value of true zero. **Example: weight and height.**
- geometric mean, the coefficient of variation, or harmonic mean.
- Example: What is your weight in kilograms?
- Less than 50 kilograms
- 51- 70 kilograms
- 71- 90 kilograms
- 91-110 kilograms
- More than 110 kilograms

Quantitative Variables(Continuous, Discrete, and Categorical) - Summary on Level of Measurements (scale)

Factors influencing data analytics	Nominal	Ordinal	Interval	Ratio
1. Can we sequence the variables?	-	Yes	Yes	Yes
2. The descriptive statistics: Mode	Yes	Yes	Yes	Yes
3. The descriptive statistics: Median	-	Yes	Yes	Yes
4. The descriptive statistics: Mean	-	-	Yes	Yes
5. Can we evaluate the difference between variables ?	-	-	Yes	Yes
6. Can we add or subtract the variables ?	-	-	Yes	Yes
7. How about with Multiplication and Division of variables ?	-	-	-	Yes
8. Absolute zero	-	-	-	Yes

WHY IDENTIFY THE UNITS OF ANALYSIS

- Without units of analysis, there is no measurement.
- Without Measurement, there is no data.
- Without Data, there is no Analysis.
- Without Analysis, there is no Modeling.
- Without Modeling, there is no Explanation and Prediction.
- Without Explanation, there is no Insight.
- Without Prediction, there can be no Optimization.
- Without Insight & Optimization, there is no Management.

References

How to interpret ordinal data (Achilleas Kostoulas)

Solutions to Categorical, Discrete and Continuous Variable Problems (superprof.co.uk)

Quartiles - Definition, Formula, Solved Example Problems (brainkart.com)

Solutions to Quartiles, Deciles and Percentiles Problems (superprof.co.uk)

Categorise the following variable

Brand of Cell phones
Size of Shoes
Body Temperature
Score in math quiz
Ages of children enrolled in a daycare center
Ranking in a mobile legend game
Time required to complete the chess game
Telephone number
Favorite Color
Low, medium, and high income group
Number of siblings
Religion
Weight of a newly born baby
Grade in math 7
Height of your friend
Salary
Blood Type

Answers: Categorise the following variable

Brand of Cell phones – Categorical - Nominal
Size of Shoes - Discrete
Body Temperature - Continuous
Score in math quiz - Discrete
Ages of children enrolled in a daycare center – Continuous
Ranking in a mobile legend game
Time required to complete the chess game
Telephone number – numeric nominal
Favorite Color - Nominal
Low, medium, and high income group - Ordinal
Number of siblings - Discrete
Religion - Categorical (Nominal)
Weight of a newly born baby – Continuous [Interval]
Grade in math - ordinal
Height of your friend – Continuous
Salary – Nominal (Unordered Categorical) under Ambiguous cases, Otherwise Continuous numeric value.
Blood Type : Categorical