UNIT

03

DATA OBJECTS AND ATTRIBUTE TYPES

This unit will cover about data and dataset being used as an entry to the different statistical analysis and reports. It will discuss the kind of attributes a data can have in order to use it for a particular application. In descriptive statistics, attributes of data can be nominal, ordinal, interval and ratio in terms of value. On the latter part of this unit, it will outline the different datasets of an object.

LESSON 1:

ATTRIBUTES/VARIABLES

OBJECTIVES:

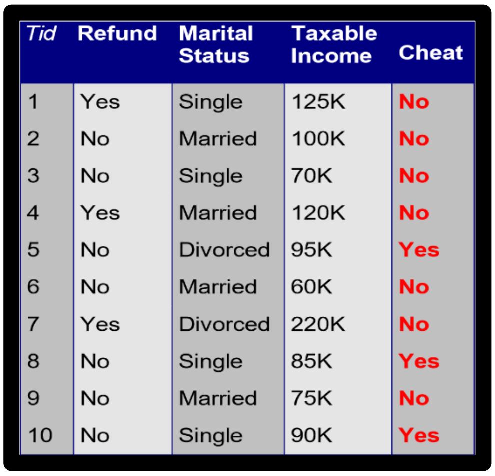
At the end of this lesson, the student will be able to:

- Define attributes as an important part in statistical analysis.
- Differentiate each attribute types based on their usage and functions.
- Characterize the different attribute types as a level of measurement.

Duration: 2 hours

Data objects represents an object in a particular database or record. An object could be a customer, an item or sales. In an enrollment database an object could be a student, courses or professors. Data objects are usually describe by its attributes. Data objects can also be treated as instances, data points, samples, examples or objects. If they are stored in a database, they are termed as tuples, which means that rows in a database corresponds to data objects and the column corresponds to its attributes.

Data or dataset are made up of data objects. A data object embodies an entity - in a sales database, the objects may be customers, store items, and sales; in a medical database, the objects may be patients; in a university database, sample of objects may be students, professors, and courses. Data objects are typically described by attributes. Data objects can also be referred to as samples, examples, instances, data points, or objects. If the data objects are stored in a database, they are data tuples. That is, the rows of a database correspond to the data objects, and the columns correspond to the attributes. In this section, we define attributes and look at the various attribute types.



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Figure 3.1. Sample Dataset

Attributes are data field describing characteristics or properties of a data object. Some of the terms used to in place of attributes are dimension, variable or feature. The term dimension is commonly used in data warehousing. Machine learning literature tends to use the term feature, while statisticians prefer the term variable. Data mining and database professionals commonly use the term attribute, and we do here as well. Attributes describing a customer object can include, for example, customer ID, name, and address. Observed values for a given attribute are known as observations. The type of an attribute is determined by the set of possible values. An attribute may be nominal, ordinal, interval or ratio.

Attributes of data in Descriptive Analytics

1. Nominal
2. Ordinal
3. Interval
4. Ratio

Nominal Attributes

Nominal means “relating to names.” The values of a nominal attribute are symbols or names of things. Each value represents some kind of category, code, or state, and so nominal attributes are also referred to as categorical. The values do not have any meaningful order.

A nominal attributes is a type of attribute that is used to name, label or categorize particular attributes that are being measured. It takes qualitative values representing different categories, and there is no intrinsic ordering of these categories.

You can code nominal variables with numbers, but the order is arbitrary and arithmetic operations cannot be performed on the numbers. This is the case when a person’s phone number, National Identification Number postal code, etc. are being collected.

A nominal variable is one of the 2 types of categorical variables and is the simplest among all the measurement variables. Some examples of nominal variables include gender, Name, phone, etc.

Ordinal Attributes

An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known. Ordinal attribute is a type of measurement attribute that takes values with an order or rank. They are built upon nominal scales by assigning numbers to objects to reflect a rank or ordering on an attribute. Also, there is no standard ordering in the ordinal variable scale.

In another sense, we could say the difference in the rank of an ordinal variable is not equal. It is mostly classified as one of the 2 types of categorical variables, while in some cases it is said to be a midpoint between categorical and numerical variables.

Interval Attributes

Interval-scaled attributes are measured on a scale of equal-size units. The values of interval-scaled attributes have order and can be positive, 0, or negative. Thus, in addition to providing a ranking of values, such attributes allow us to compare and quantify the difference between values. The interval attribute is a measurement attribute that is used to define values measured along a scale, with each point placed at an equal distance from one another.

Unlike ordinal variables that take values with no standardized scale, every point in the interval scale is equidistant. Arithmetic operations can also be performed on the numerical values of the interval variable. These arithmetic operations are, however, just limited to addition and subtraction. Examples of interval variables include; temperature measured in Celsius or Fahrenheit, time, generation age range, etc.

Ratio Attributes

A ratio-scaled attribute is a numeric attribute with an inherent zero-point. That is, if a measurement is ratio-scaled, we can speak of a value as being a multiple (or ratio) of another value. In addition, the values are ordered, and we can also compute the difference between values, as well as the mean, median, and mode.

The ratio variable is one of the 2 types of continuous variables, where the interval variable is the 2nd. It is an extension of the interval variable and is also the peak of the measurement variable types.

The only difference between the ratio variable and interval variable is that the ratio variable already has a zero value. For example, temperature, when measured in Kelvin is an example of ratio variables. The presence of a zero-point accommodates the measurement in Kelvin. Also, unlike the interval variable multiplication and division operations can be performed on the values of a ratio variable.

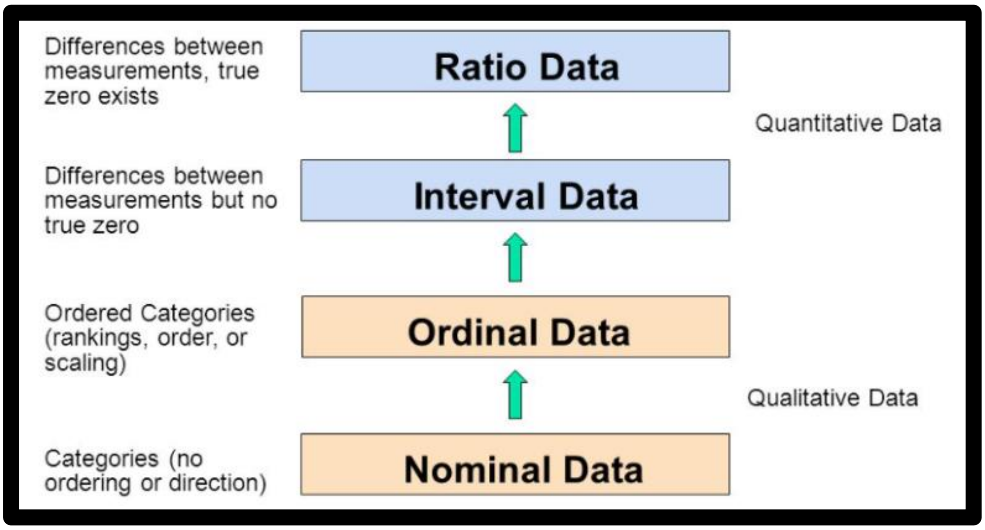

Differences between measurements, true zero exists
Ratio Data
Quantitative Data
Differences between measurements but no true zero
Interval Data
Ordered Categories (rankings, order, or scaling)
Ordinal Data
Qualitative Data
Categories (no ordering or direction)
Nominal Data

Figure 3.2. Data Objects and Attributes

Source: https://www.graphpad.com/support/faq/what-is-the-difference-between-ordinal-interval-and-ratio-variables-why-should-i-care/

There are other ways of classifying attributes in statistics. One is qualitative vs. quantitative. Qualitative variables are descriptive or categorical. In other statistical computation, such as mean and standard deviation, cannot compute using qualitative variables. Quantitative variables numerical, so computing means and standard deviation is possible.

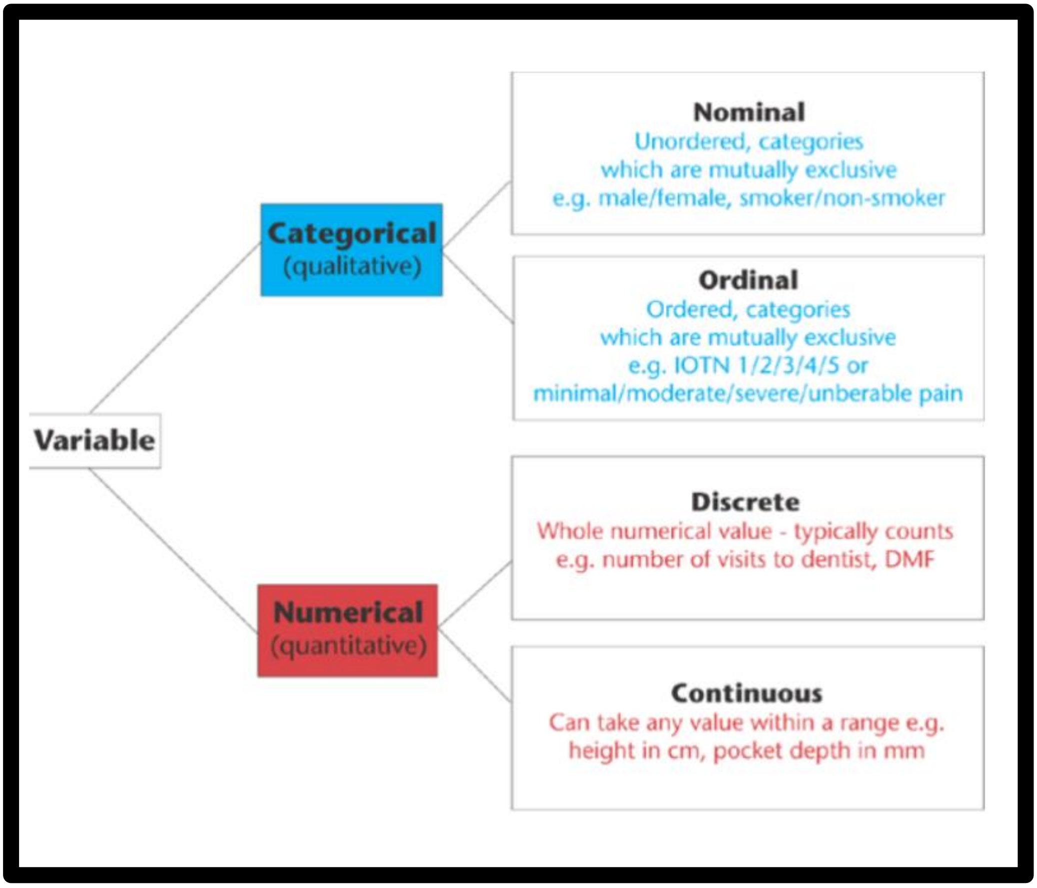

Variable
Categorical
(qualitative)
Numerical
(quantitative)
Nominal
Unordered, categories
which are mutually exclusive
e.g. male/female, smoker/non-smoker
Ordinal
Ordered, categories
which are mutually exclusive
e.g. IOTN 1/2/3/4/5 or
minimal/moderate/severe/unberable pain
Discrete
Whole numerical value - typically counts
e.g. number of visits to dentist, DMF
Continuous
Can take any value within a range e.g.
height in cm, pocket depth in mm

Figure 3.3. Qualitative VS Qualitative Categories of Data

Source: https://www.graphpad.com/support/faq/what-is-the-difference-between-ordinal-interval-and-ratio-variables-why-should-i-care/

Discrete versus Continuous Attributes

Attributes has been organized into nominal, binary, ordinal, and numeric types. There are many ways to organize attribute types. Classification algorithms developed from the field of machine learning often talk of attributes as being either discrete or continuous. Each type may be processed differently. A discrete attribute has a finite or countably infinite set of values, which may or may not be represented as integers. The attributes hair color, smoker, medical test, and drink size each have a finite number of values, and so are discrete. Note that discrete attributes may have numeric values, such as 0 and 1 for binary attributes or, the values 0 to 110 for the attribute age. An attribute is countably infinite if the set of possible values is infinite but the values can be put in a one-to-one correspondence with natural numbers. For example, the attribute customer ID is countably infinite. The number of customers can grow to infinity, but in reality, the actual set of values is countable (where the values can be put in one-to-one correspondence with the set of integers). Zip codes are another example.

If an attribute is not discrete, it is continuous. The terms numeric attribute and continuous attribute are often used interchangeably in the literature. (This can be confusing because, in the classic sense, continuous values are real numbers, whereas numeric values can be either integers or real numbers.) In practice, real values are represented using a finite number of digits. Continuous attributes are typically represented as floating-point variables.

APPLICATION ACTIVITY FOR LESSON 1

HOW DO YOU APPLY WHAT YOU HAVE LEARNED?

Name: ________________________________ Date: _______________________

Course & Section: ______________________ Result: ______________________

Discuss the different type of data being used in descriptive analytics in terms of their functions and usage. Give examples for each type of data. State your answer in 180 to 200 words.

Rubrics:

Your answer will be graded based on the following criteria, the final score will be computed as the average of the four items.

No.	Items	Weight %	Actual Score
1.	The different type of data set are completely stated.	25	
2.	Examples given are accurate	25	
3.	Constructions of statements are understandable	25	
4.	The statement is organized and made use of applicable and direct to the point wordings	25	

LESSON 2:

TYPES OF DATA SET

OBJECTIVES:

At the end of this lesson, the student will be able to:

- Describe dataset as a representation of data.
- Determine the type of dataset that will be used on a given objects in the analysis.
- Understant the concept of the different data sets.

Duration: 3 hours

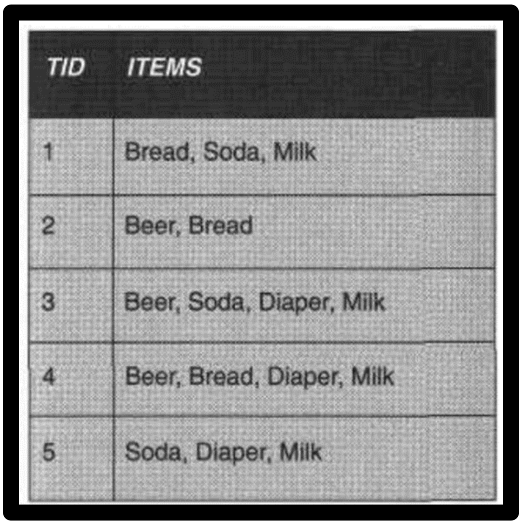
Types of Data Set

Record

The most basic form of record data has no explicit relationship among records or data fields, and every record (object) has the same set of attributes. Record data is usually stored either in flat files or in relational databases.

There are a few variations of Record Data, which have some characteristic properties.

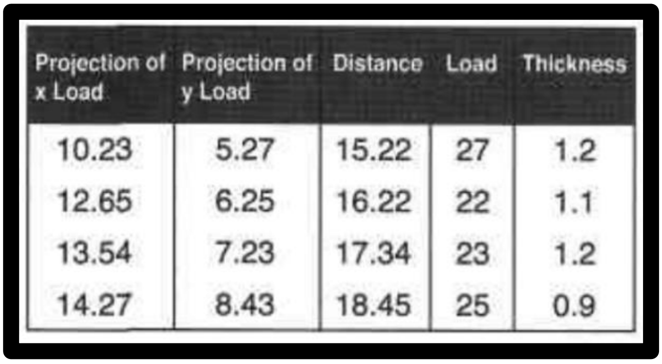
1. Transaction or Market Basket Data: It is a special type of record data, in which each record contains a set of items. For example, shopping in a supermarket or a grocery store. For any particular customer, a record will contain a set of items purchased by the customer in that respective visit to the supermarket or the grocery store. This type of data is called Market Basket Data. Transaction data is a collection of sets of items, but it can be viewed as a set of records whose fields are asymmetric attributes. Most often, the attributes are binary, indicating whether or not an item was purchased or not.
2. The Data Matrix: If the data objects in a collection of data all have the same fixed set of numeric attributes, then the data objects can be thought of as points (vectors) in a multidimensional space, where each dimension represents a distinct attribute describing the object. A set of such data objects can be interpreted as an m X n matrix, where there are n rows, one for each object, and n columns, one for each attribute. Standard matrix operation can be applied to transform and manipulate the data. Therefore, the data matrix is the standard data format for most statistical data.
3. The Sparse Data Matrix: A sparse data matrix (sometimes also called document-data matrix) is a special case of a data matrix in which the attributes are of the same type and are asymmetric; i.e., only non-zero values are important.



TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

Figure 3.4. Transaction data

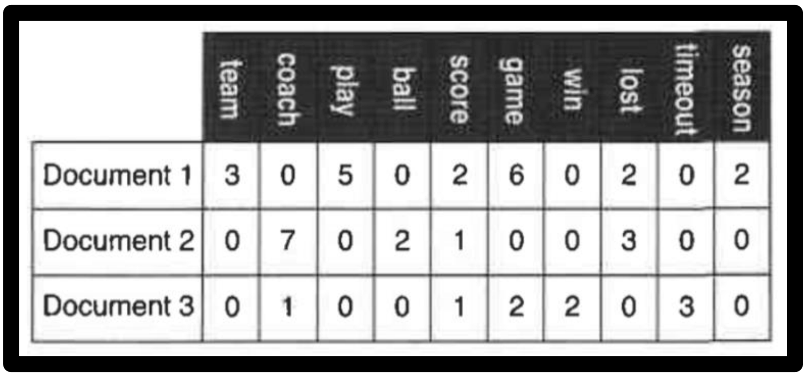
Source: https://towardsdatascience.com/types-of-data-sets-in-data-science-data-mining-machine-learning-eb47c80af7a



Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

Figure 3.5. Data Matrix

Source: https://towardsdatascience.com/types-of-data-sets-in-data-science-data-mining-machine-learning-eb47c80af7a



	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

Figure 3.6. Document-data matrix

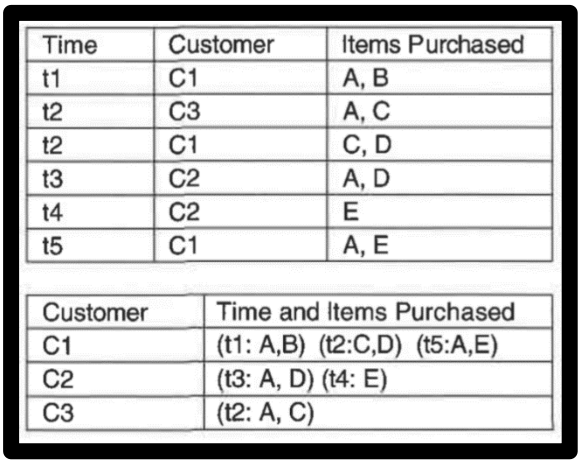
Source: https://towardsdatascience.com/types-of-data-sets-in-data-science-data-mining-machine-learning-eb47c80af7a

Ordered

For some types of data, the attributes have relationships that involve order in time or space. As you can see in the picture above, it can be segregated into four types:

1. Sequential Data: Also referred to as temporal data, can be thought of as an extension of record data, where each record has a time associated with it. Consider a retail transaction data set that also stores the time at which the transaction took place
2. Sequence Data: Sequence data consists of a data set that is a sequence of individual entities, such as a sequence of words or letters. It is quite similar to sequential data, except that there are no time stamps; instead, there are positions in an ordered sequence. For example, the genetic information of plants and animals can be represented in the form of sequences of nucleotides that are known as genes.
3. Time Series Data: Time series data is a special type of sequential data in which each record is a time series, i.e., a series of measurements taken over time. For example, a financial data set might contain objects that are time series of the daily prices of various stocks.
4. Spatial Data: Some objects have spatial attributes, such as positions or areas, as well as other types of attributes. An example of spatial data is weather data

(precipitation, temperature, pressure) that is collected for a variety of geographical locations.

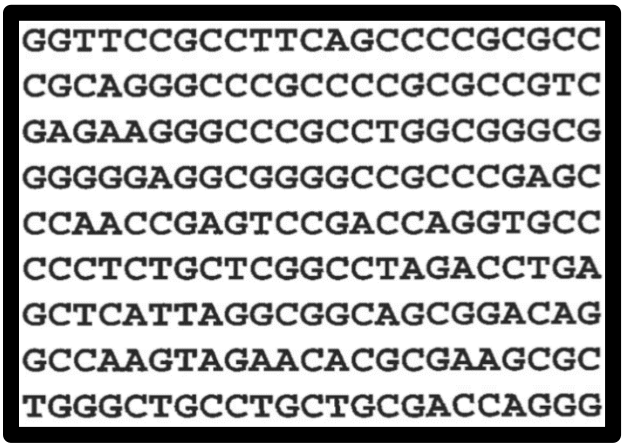


Time	Customer	Items Purchased
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

Customer	Time and Items Purchased
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

Figure 3.7. Sequential data

Source: https://towardsdatascience.com/types-of-data-sets-in-data-science-data-mining-machine-learning-eb47c80af7a



```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Figure 3.8. Genomic sequence data

Source: https://towardsdatascience.com/types-of-data-sets-in-data-science-data-mining-machine-learning-eb47c80af7a

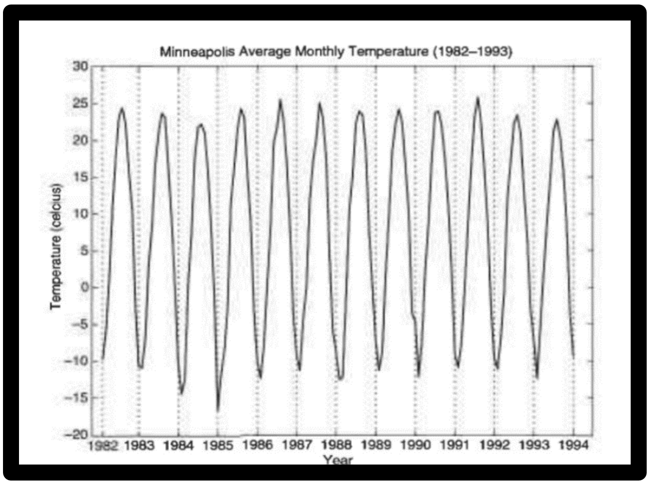


Figure 3.9. Temperature time-series

Source: https://towardsdatascience.com/types-of-data-sets-in-data-science-data-mining-machine-learning-eb47c80af7a

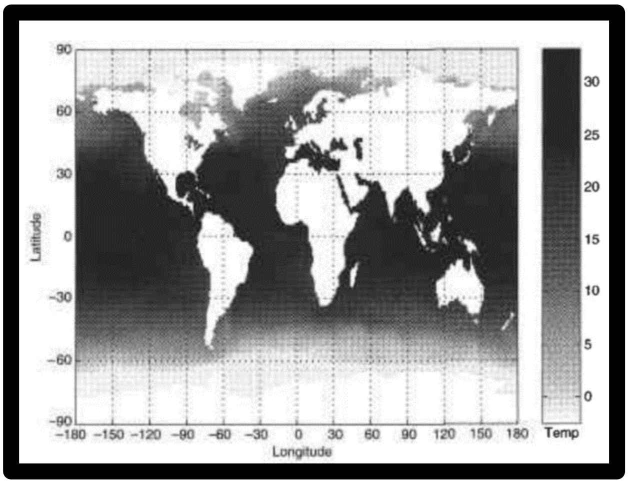


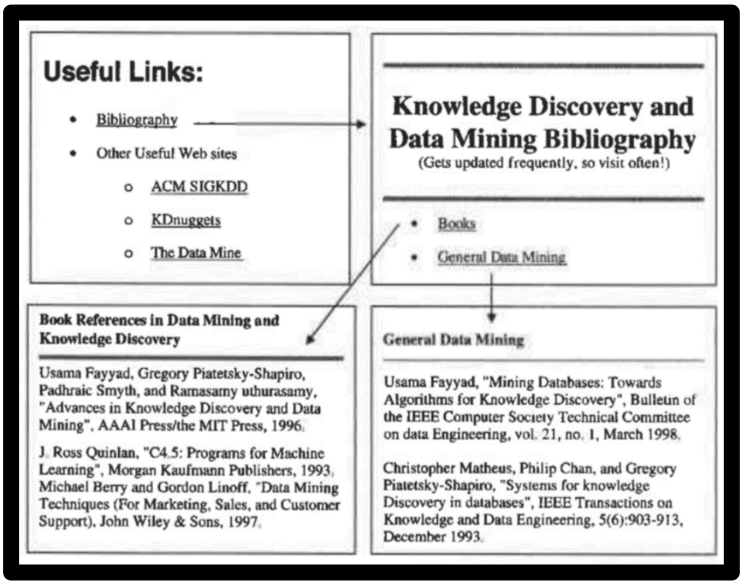
Figure 3.10. Spatial temperature data

Source: https://towardsdatascience.com/types-of-data-sets-in-data-science-data-mining-machine-learning-eb47c80af7a

Graph

This can be further divided into types:

1. Data with Relationships among Objects: The data objects are mapped to nodes of the graph, while the relationships among objects are captured by the links between objects and link properties, such as direction and weight. Consider Web pages on the World Wide Web, which contain both text and links to other pages. In order to process search queries, Web search engines collect and process Web pages to extract their contents.
2. Data with Objects That Are Graphs: If objects have structure, that is, the objects contain sub objects that have relationships, then such objects are frequently represented as graphs. For example, the structure of chemical compounds can be represented by a graph, where the nodes are atoms and the links between nodes are chemical bonds.



Useful Links:

- Bibliography
- Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - The Data Mine

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- Books
- General Data Mining

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993. Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

Figure 3.11. Linked Web pages

Source: https://towardsdatascience.com/types-of-data-sets-in-data-science-data-mining-machine-learning-eb47c80af7a

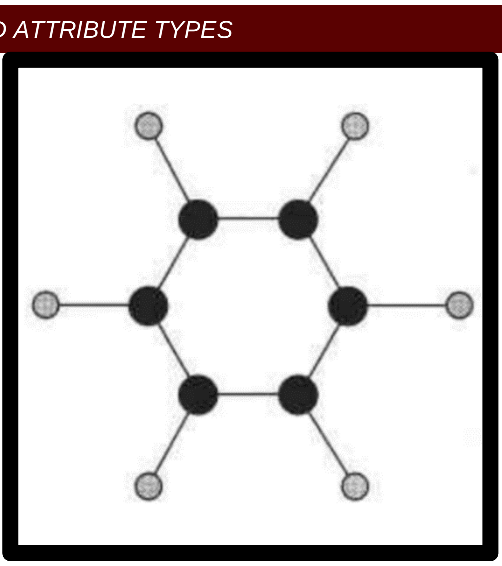


Figure 3.12. Benzene Molecule

Source: https://towardsdatascience.com/types-of-data-sets-in-data-science-data-mining-machine-learning-eb47c80af7a

APPLICATION ACTIVITY FOR LESSON 2

HOW DO YOU APPLY WHAT YOU HAVE LEARNED?

Name: ______________________________ Date: ______________________

Course & Section: ____________________ Result: _____________________

Discuss in 180 to 200 words as to when it is best to use record data, ordered data and graph data. Cite an example to support your answer

____

____

____

____

____

____

____

____

____

____

____

____

____

____

____

Rubrics:

Your answer will be graded based on the following criteria, the final score will be computed as the average of the four items.

No.	Items	Weight %	Actual Score
1.	The different type of data set are completely stated.	25	
2.	Examples given are accurate	25	
3.	Constructions of statements are understandable	25	
4.	The statement is organized and made use of applicable and direct to the point wordings	25	