IT496: Introduction to Data Mining

Lecture - 02

Statistics for Data Mining - I

[Attribute Types and Measures of Central Tendency]

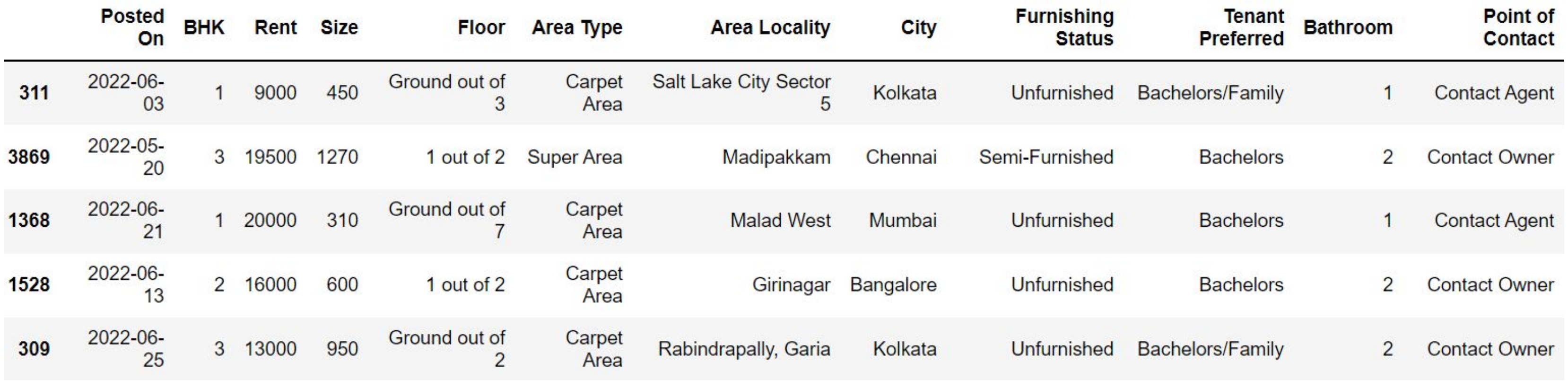
Arpit Rana

27th July 2023

Attribute Types

Terminology of Structured Data

House Rent Prediction Dataset (from Magicbricks, India)



	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
311	2022-06-03	1	9000	450	Ground out of 3	Carpet Area	Salt Lake City Sector 5	Kolkata	Unfurnished	Bachelors/Family	1	Contact Agent
3869	2022-05-20	3	19500	1270	1 out of 2	Super Area	Madipakkam	Chennai	Semi-Furnished	Bachelors	2	Contact Owner
1368	2022-06-21	1	20000	310	Ground out of 7	Carpet Area	Malad West	Mumbai	Unfurnished	Bachelors	1	Contact Agent
1528	2022-06-13	2	16000	600	1 out of 2	Carpet Area	Girinagar	Bangalore	Unfurnished	Bachelors	2	Contact Owner
309	2022-06-25	3	13000	950	Ground out of 2	Carpet Area	Rabindrapally, Garia	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner

Dataset Glossary (Column-Wise): 4746 Records

- **BHK**: Number of Bedrooms, Hall, Kitchen.
- **Floor**: Ground out of 2, 3 out of 5, etc.
- **Size**: Size of property in Square Feet.
- **Area** Type: Super Area/Carpet Area/Built Area.
- **Furnishing Status**: Furnished/Semi-Furnished/ Unfurnished.
- **Bathroom**: Number of Bathrooms.

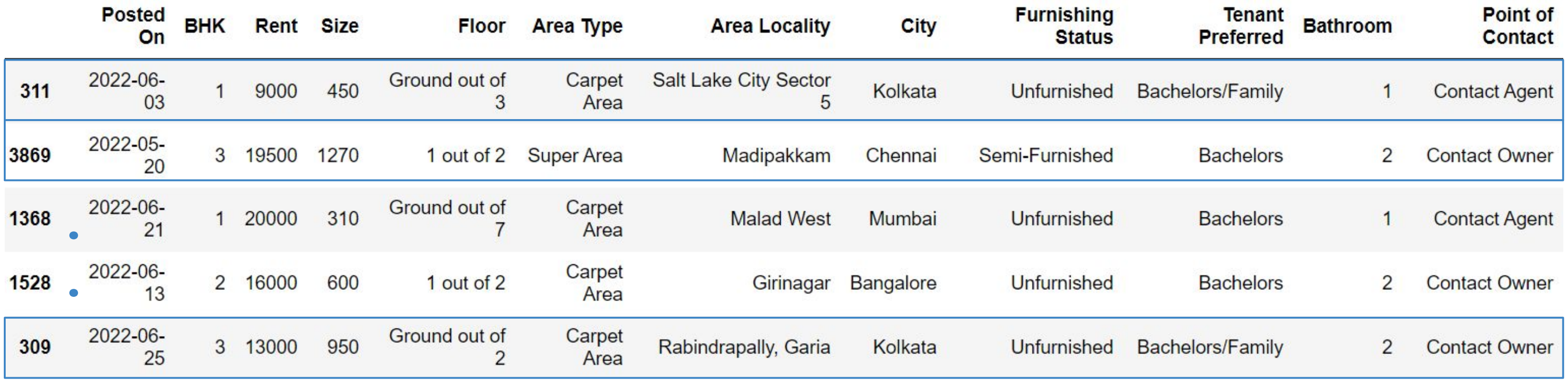
- **Area Locality**: Locality of the property
- **City**: City where the property is Located.
- **Tenant Preferred**: Family/Bachelor
- **Point of Contact**: Agent / Owner
- **Rent**: Price of the property

Data Objects

A data object represents an *entity* (*i.e. a row in a database*).

- Customers, store items in a sales database
- Professors, students, and courses in a university database
- Patients in a medical database

Data objects are also referred to as *samples*, *examples*, *instances*, *data points*, *domain points or simply objects.*



	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
311	2022-06-03	1	9000	450	Ground out of 3	Carpet Area	Salt Lake City Sector 5	Kolkata	Unfurnished	Bachelors/Family	1	Contact Agent
3869	2022-05-20	3	19500	1270	1 out of 2	Super Area	Madipakkam	Chennai	Semi-Furnished	Bachelors	2	Contact Owner
1368	2022-06-21	1	20000	310	Ground out of 7	Carpet Area	Malad West	Mumbai	Unfurnished	Bachelors	1	Contact Agent
1528	2022-06-13	2	16000	600	1 out of 2	Carpet Area	Girinagar	Bangalore	Unfurnished	Bachelors	2	Contact Owner
309	2022-06-25	3	13000	950	Ground out of 2	Carpet Area	Rabindrapally, Garia	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner

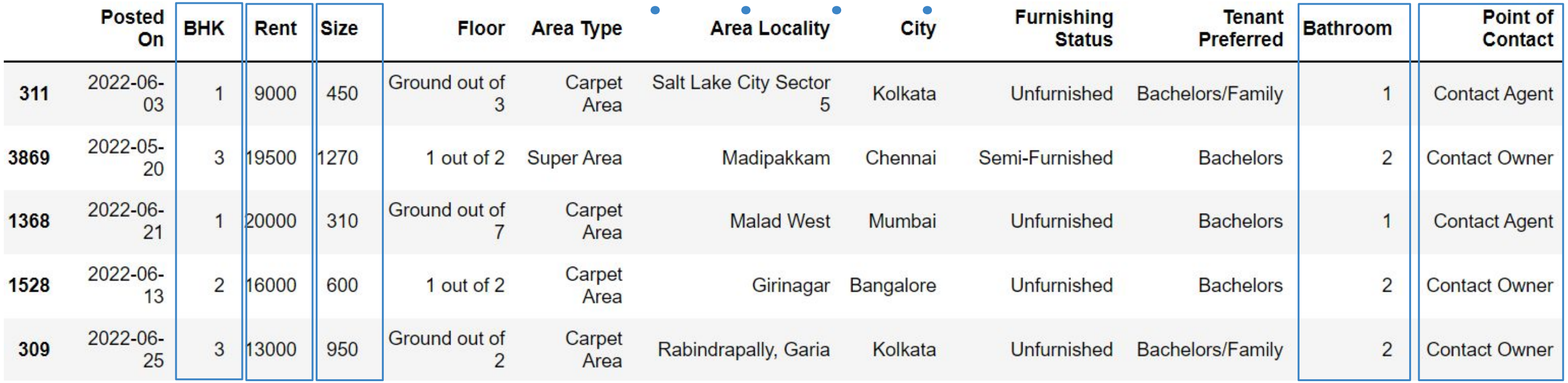
Attributes

An attribute is a data field *representing a characteristic or a feature* of a data object.

- A customer object → customer_ID, name, address, age, and gender
- A course object → course_ID, credit_structure, slot, and instructor

Attributes are also referred to as *dimensions*, *features*, and *variables*.

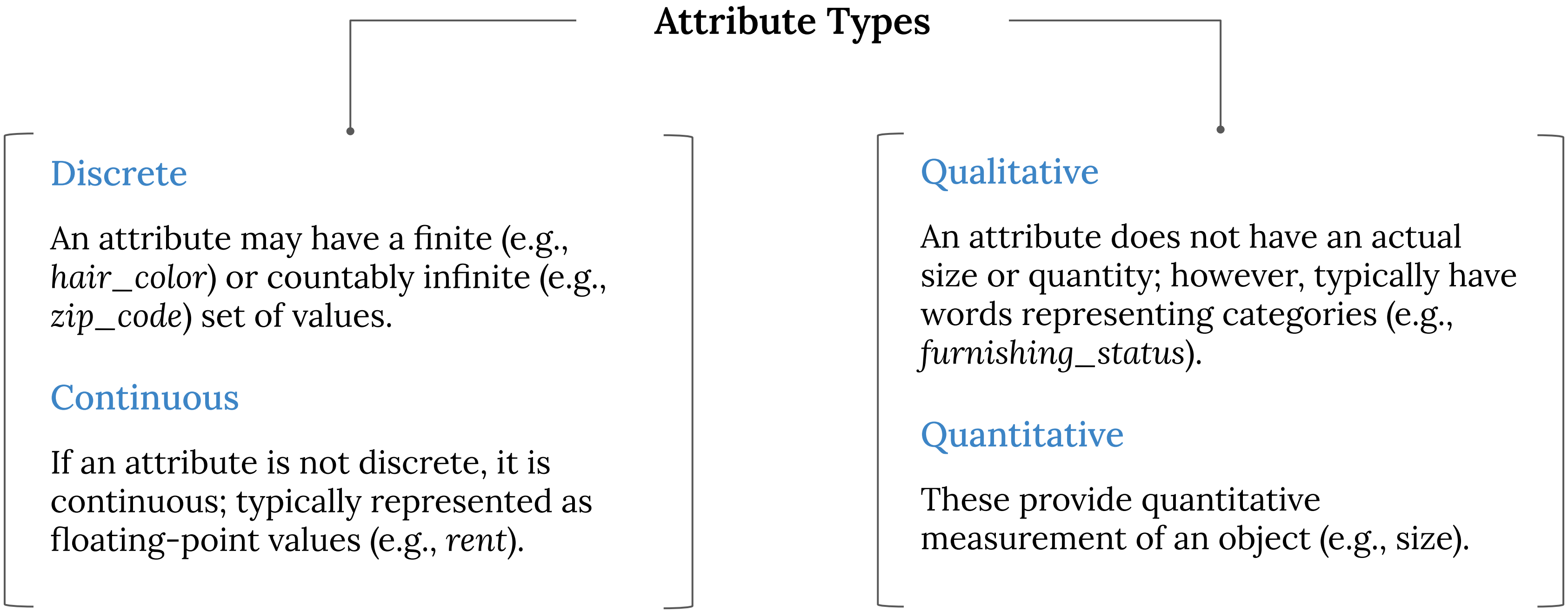
- A set of attributes used to describe a given object is called an ***attribute vector***.



	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
311	2022-06-03	1	9000	450	Ground out of 3	Carpet Area	Salt Lake City Sector 5	Kolkata	Unfurnished	Bachelors/Family	1	Contact Agent
3869	2022-05-20	3	19500	1270	1 out of 2	Super Area	Madipakkam	Chennai	Semi-Furnished	Bachelors	2	Contact Owner
1368	2022-06-21	1	20000	310	Ground out of 7	Carpet Area	Malad West	Mumbai	Unfurnished	Bachelors	1	Contact Agent
1528	2022-06-13	2	16000	600	1 out of 2	Carpet Area	Girinagar	Bangalore	Unfurnished	Bachelors	2	Contact Owner
309	2022-06-25	3	13000	950	Ground out of 2	Carpet Area	Rabindrapally, Garia	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner

Attribute Types

The type of an attribute is determined by the set of possible values it can have. They can be categorized in two ways:



Attribute Types

Discrete

An attribute may have a finite (e.g., *hair_color*) or countably infinite (e.g., *zip_code*) set of values.

Continuous

If an attribute is not discrete, it is continuous; typically represented as floating-point values (e.g., *rent*).

Qualitative

An attribute does not have an actual size or quantity; however, typically have words representing categories (e.g., *furnishing_status*).

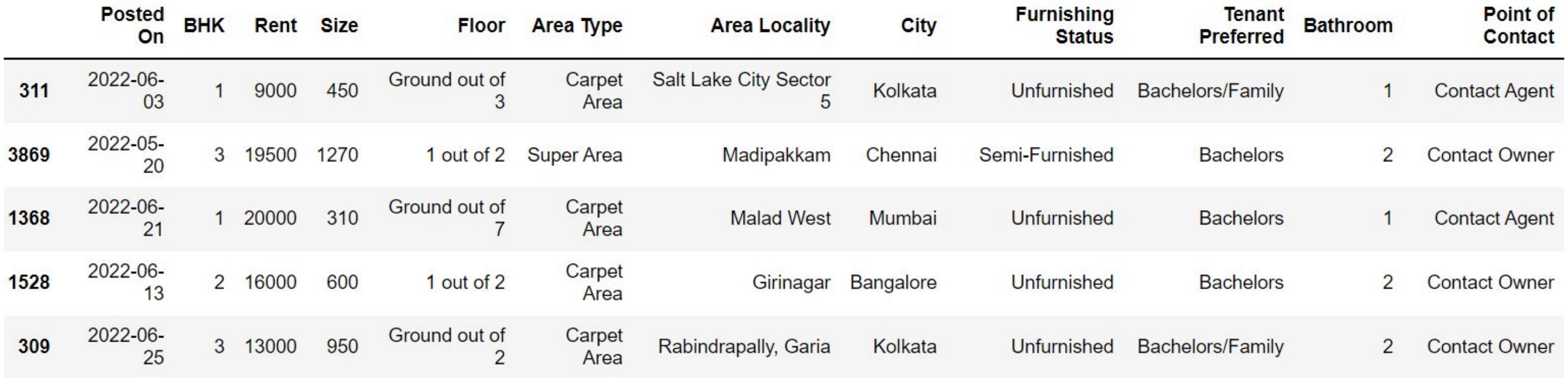
Quantitative

These provide quantitative measurement of an object (e.g., size).

Nominal Attributes (*a.k.a. categorical*) | Qualitative | Discrete

Its values can be *symbols or names of things* representing a category, code, or state.

- *hair_color* (black, brown, blond, etc.)
- *marrital_status* (single, married, divorced, and widowed)
- *occupation* (doctor, programmer, teacher, etc.)

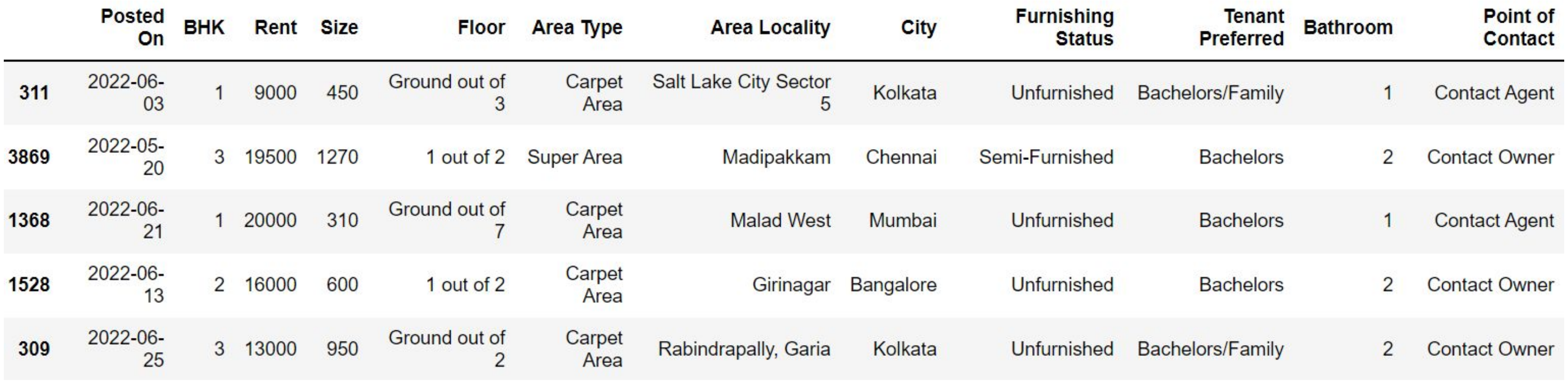


	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
311	2022-06-03	1	9000	450	Ground out of 3	Carpet Area	Salt Lake City Sector 5	Kolkata	Unfurnished	Bachelors/Family	1	Contact Agent
3869	2022-05-20	3	19500	1270	1 out of 2	Super Area	Madipakkam	Chennai	Semi-Furnished	Bachelors	2	Contact Owner
1368	2022-06-21	1	20000	310	Ground out of 7	Carpet Area	Malad West	Mumbai	Unfurnished	Bachelors	1	Contact Agent
1528	2022-06-13	2	16000	600	1 out of 2	Carpet Area	Girinagar	Bangalore	Unfurnished	Bachelors	2	Contact Owner
309	2022-06-25	3	13000	950	Ground out of 2	Carpet Area	Rabindrapally, Garia	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner

Binary Attributes | Qualitative | Discrete

These are nominal attributes with *only two categories or states* (0 or 1 / true or false).

- *gender* (make or female | *symmetric* i.e. both the outcomes are equally important)
- *medical_test* (positive or negative | *asymmetric*)

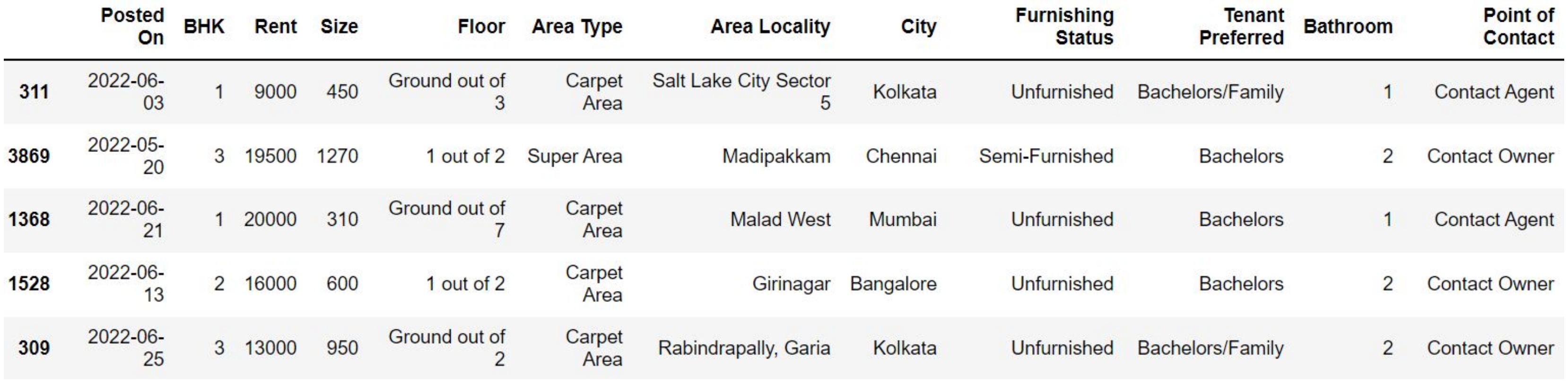


	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
311	2022-06-03	1	9000	450	Ground out of 3	Carpet Area	Salt Lake City Sector 5	Kolkata	Unfurnished	Bachelors/Family	1	Contact Agent
3869	2022-05-20	3	19500	1270	1 out of 2	Super Area	Madipakkam	Chennai	Semi-Furnished	Bachelors	2	Contact Owner
1368	2022-06-21	1	20000	310	Ground out of 7	Carpet Area	Malad West	Mumbai	Unfurnished	Bachelors	1	Contact Agent
1528	2022-06-13	2	16000	600	1 out of 2	Carpet Area	Girinagar	Bangalore	Unfurnished	Bachelors	2	Contact Owner
309	2022-06-25	3	13000	950	Ground out of 2	Carpet Area	Rabindrapally, Garia	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner

Ordinal Attributes | Qualitative | Discrete

These are nominal attributes with values that have a *meaningful order or ranking among them*, however, the magnitude between successive values is not known.

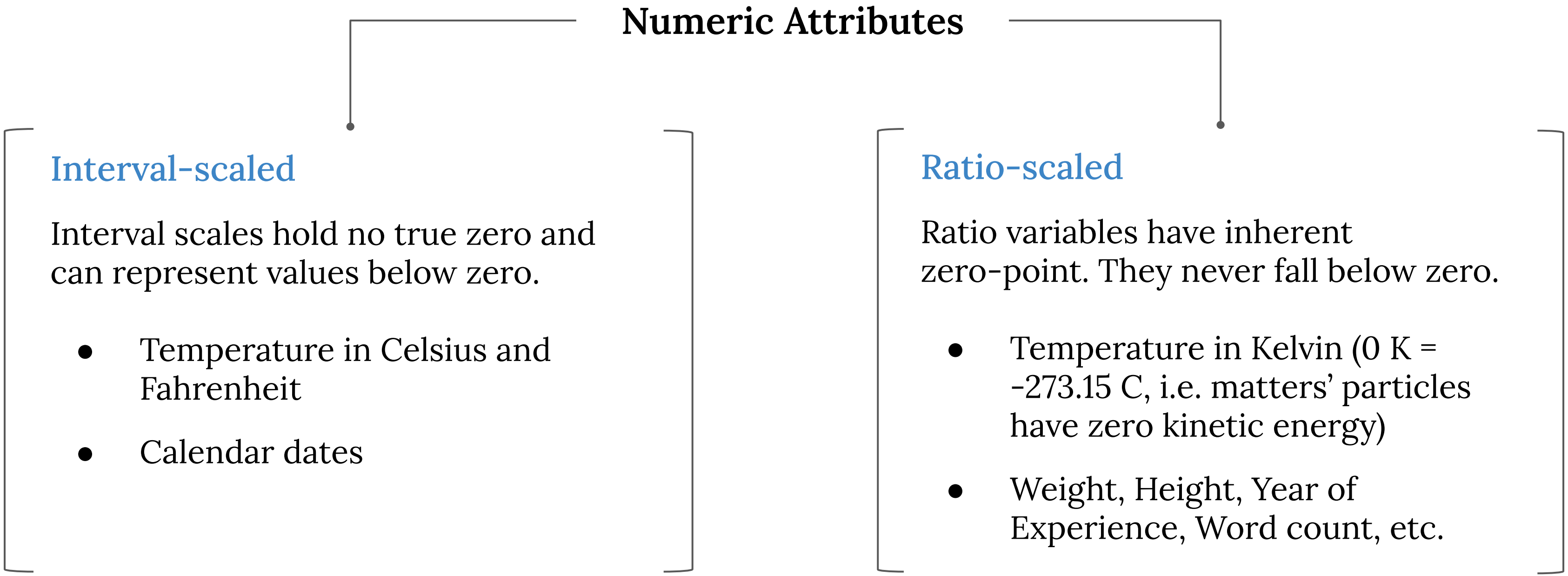
- Professional ranks (assistant, associate, and full professor)
- Likert scale (Below average, average, Above average)
- A finite number of ordered categories, e.g., Age <= 45, 45 < Age <= 60, Age > 60



	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
311	2022-06-03	1	9000	450	Ground out of 3	Carpet Area	Salt Lake City Sector 5	Kolkata	Unfurnished	Bachelors/Family	1	Contact Agent
3869	2022-05-20	3	19500	1270	1 out of 2	Super Area	Madipakkam	Chennai	Semi-Furnished	Bachelors	2	Contact Owner
1368	2022-06-21	1	20000	310	Ground out of 7	Carpet Area	Malad West	Mumbai	Unfurnished	Bachelors	1	Contact Agent
1528	2022-06-13	2	16000	600	1 out of 2	Carpet Area	Girinagar	Bangalore	Unfurnished	Bachelors	2	Contact Owner
309	2022-06-25	3	13000	950	Ground out of 2	Carpet Area	Rabindrapally, Garia	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner

Numeric Attributes | Quantitative | Continuous/Discrete

These are measurable quantity *represented in integers or real values*. These attributes can be *interval-scaled* or *ratio-scaled.*



Numeric Attributes

Interval-scaled

Interval scales hold no true zero and can represent values below zero.

- Temperature in Celsius and Fahrenheit
- Calendar dates

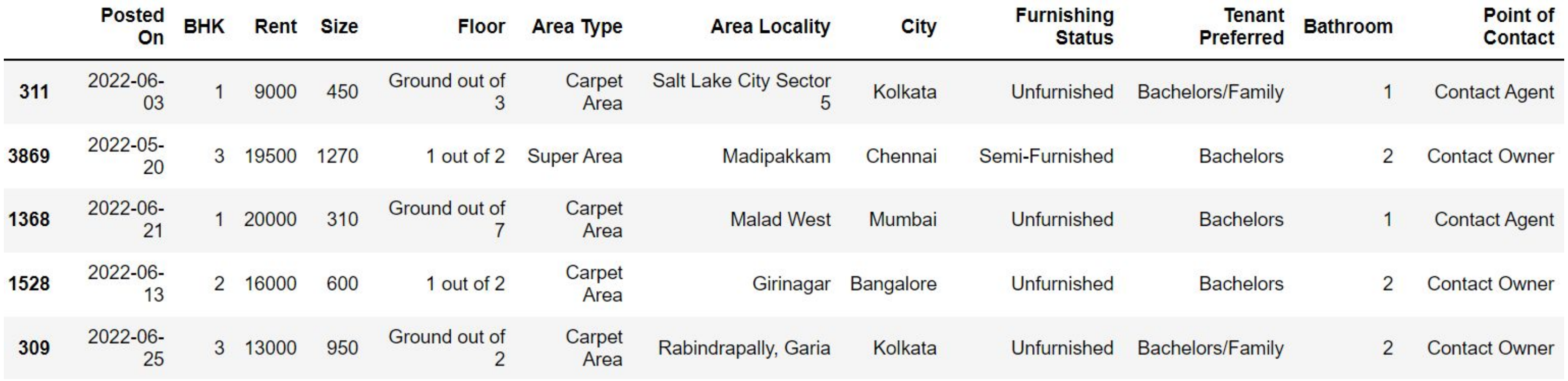
Ratio-scaled

Ratio variables have inherent zero-point. They never fall below zero.

- Temperature in Kelvin (0 K = -273.15 C, i.e. matters' particles have zero kinetic energy)
- Weight, Height, Year of Experience, Word count, etc.

Numeric Attributes | Quantitative | Continuous/Discrete

Find out *interval-scaled* or *ratio-scaled* attributes from the database below.



	Posted On	BHK	Rent	Size	Floor	Area Type	Area Locality	City	Furnishing Status	Tenant Preferred	Bathroom	Point of Contact
311	2022-06-03	1	9000	450	Ground out of 3	Carpet Area	Salt Lake City Sector 5	Kolkata	Unfurnished	Bachelors/Family	1	Contact Agent
3869	2022-05-20	3	19500	1270	1 out of 2	Super Area	Madipakkam	Chennai	Semi-Furnished	Bachelors	2	Contact Owner
1368	2022-06-21	1	20000	310	Ground out of 7	Carpet Area	Malad West	Mumbai	Unfurnished	Bachelors	1	Contact Agent
1528	2022-06-13	2	16000	600	1 out of 2	Carpet Area	Girinagar	Bangalore	Unfurnished	Bachelors	2	Contact Owner
309	2022-06-25	3	13000	950	Ground out of 2	Carpet Area	Rabindrapally, Garia	Kolkata	Unfurnished	Bachelors/Family	2	Contact Owner

Measures of Central Tendency

. . .where would the most of the values fall. . .

Mean (*the average value*)

Let x_1 , x_2 , . . . , x_N be the set of N observed values or observations for an attribute X.

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

Suppose that each value x_i is associated with a weight w_i for i = 1,...,N.

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

Highly sensitive to extreme values (e.g. outliers)!

Trimmed Mean:
After chopping off extreme values

Too much chopping may lead to LoI.

Median (*the middle value*)

Let $x_1 , x_2 , . . . , x_N$ be the set of N observed values or observations for an attribute X.

- Sort $x_1 , x_2 , . . . , x_N$ in an increasing order.
  - If N is *odd*, select the middle value.
  - If N is *even* and type of X is *numeric*, take the average of the two middlemost values.
  - If N is *even* and type of X is *ordinal*, the two middlemost values and any value in between.

For *skewed* data, median is a better measure of central tendency.

Mode (*the most frequent value*)

- A distribution may have more than one most frequent values (modes), are called *multimodal*.
    - The *bimodal* and *trimodal* distributions are special cases of multimodal.
- It can be determined for both *qualitative* and *quantitative* attributes.

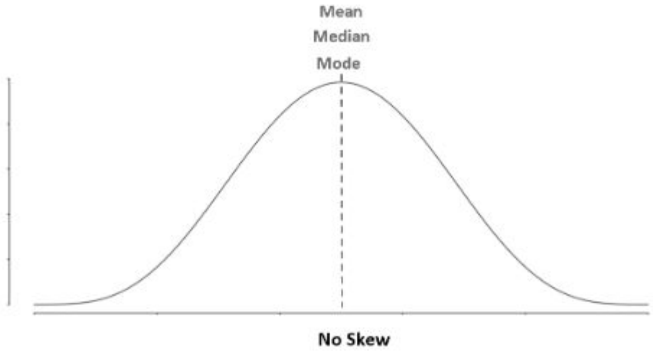
Please Indicate How You Feel About Capital Punishment?					
Frequency of Responses	Strongly oppose	Somewhat oppose	Neither	Somewhat support	Strongly support
	42	6	3	4	45

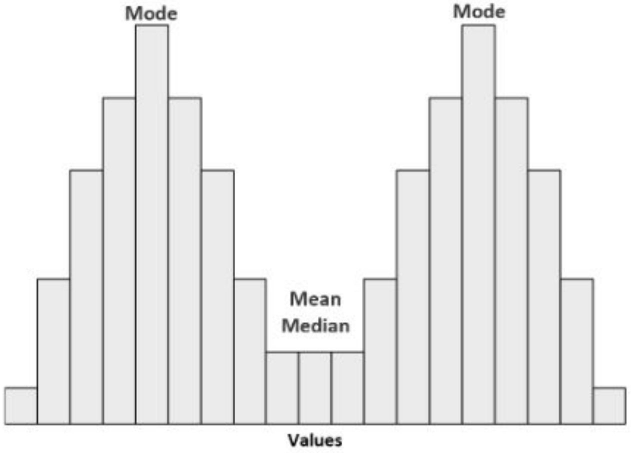
Example of Bimodal Distribution – controversial questions tend to polarize the public.

Measures of Central Tendency

Symmetric Data

- Values are spreaded symmetrically about its mean, i.e. *Skewness* = 0
- For *unimodal* distribution, Mean = Mode = Median
- For bimodal distribution, Mean = Median, and there will be two *Modes*.

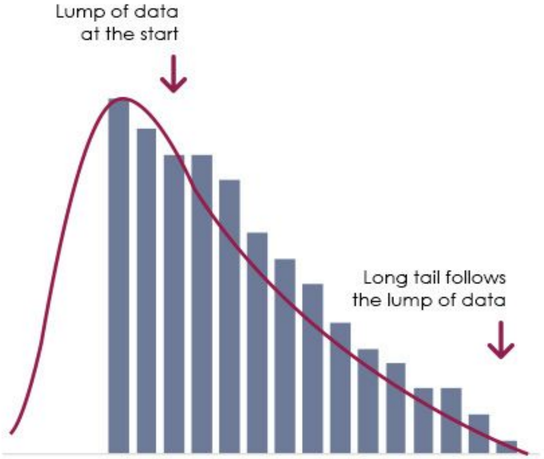




Measures of Central Tendency

Skewed data

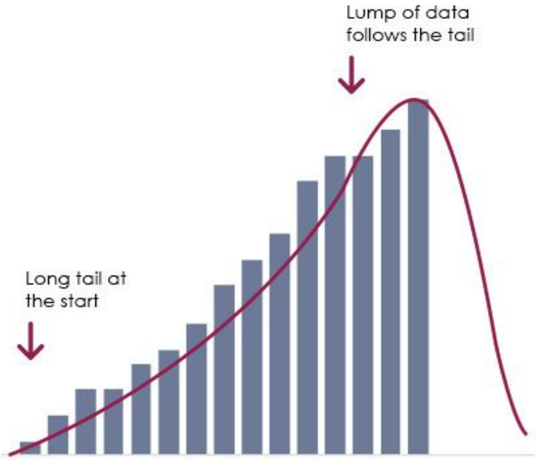
- If the values of a (random) variable are spreaded *asymmetrically about its mean.*



**Positive skew
(right-skewed distribution)**

Mode < Median < Mean

Long tail pulls the Mean towards its end.



**Negative skew
(left-skewed distribution)**

Mode > Median > Mean

Exercises

1. Does all data have a median, mode and mean?
2. In a normally distributed data set, which is greatest: mode, median or mean?
3. For any data set, which measures of central tendency have only one value?
4. Fill the entries in the following table.

Attribute Type	Measure(s) Defined	Best Measure	Why is it the best?
Nominal			
Ordinal			
Numeric (symmetric)			
Numeric (skewed)			

Next lecture

Statistics for Data Mining

28th July 2023