Data Mining

Data Characteristics

What is Data?

- Collection of data objects and their attributes

- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature

- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

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**

Attribute Values

- Attribute values are numbers or symbols assigned to an attribute

- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

Types of Attributes

- Nominal
  - Examples: ID numbers, eye color, zip codes
- Ordinal
  - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Sense of something higher or lower than something else
- Interval
  - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Does not have a true zero - zero degrees or 0 years is considered more of a label rather than the absence of a temperature or time
  - Ratios, such as 20 degrees F/2 degrees F, August 2, 2017/August 2, 2018, or 3 minutes/2 minutes, do not make sense
    - Can divide 20 degrees by 2 to get 10 degrees
  - Multiplications, such as August 2, 2017 * August 2, 2018 or 0 degrees F * 10 degrees F, do not make sense
    - Can multiply 20 degrees by 2 to get 40 degrees

Types of Attributes

- Ratio
  - Examples: temperature in Kelvin, length, time, counts
    - 0 degrees Kelvin is the point at which atoms stop moving (-459.7 degrees F or -273.15 degrees C)
  - Has a true zero meaning indicating absence of something

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Addition: + -
  - Multiplication: * /

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties

Attribute Type	**Description**	**Examples**	**Operations**
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

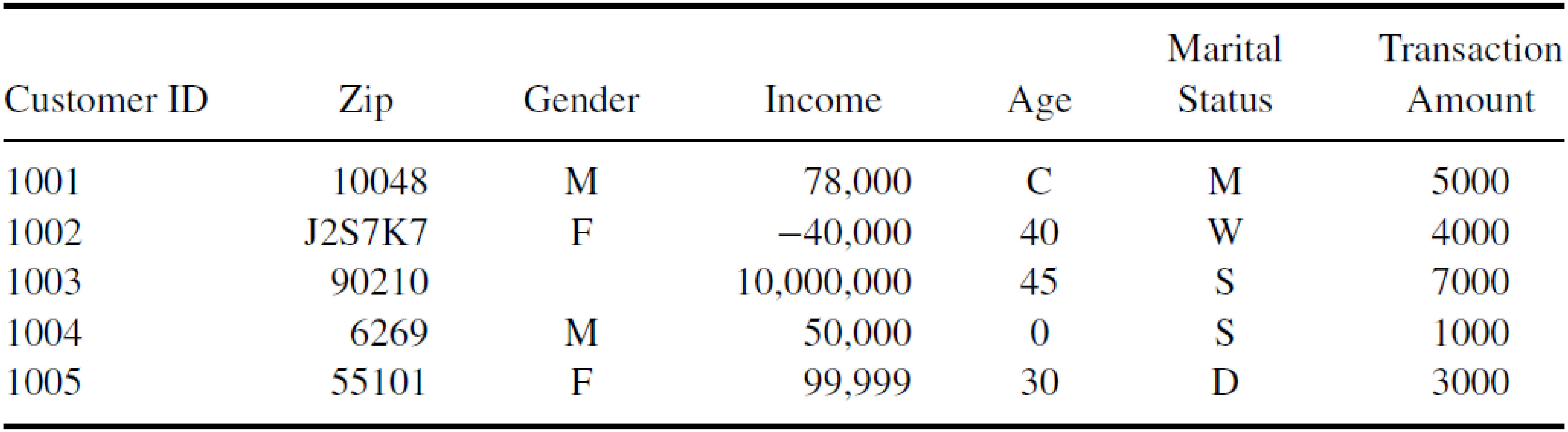
Attribute Level	**Transformation**	**Comments**
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function (entirely nonincreasing or nondecreasing function).	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes

- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: 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.

Exercise - A Small Dataset



Customer ID	Zip	Gender	Income	Age	Marital Status	Transaction Amount
1001	10048	M	78,000	C	M	5000
1002	J2S7K7	F	−40,000	40	W	4000
1003	90210		10,000,000	45	S	7000
1004	6269	M	50,000	0	S	1000
1005	55101	F	99,999	30	D	3000

Let’s ask some questions about this dataset.

- **What are the nominal, ordinal, interval, and ratio values?**
- **What are the discrete and continuous values?**
- **Do you see any issues with the Customer ID values?**
- **Do you see any issues with the zip code values?**
- **Do you see any issues with the Gender values?**
- **Do you see any issues with the Income values?**
- **Do you see any issues with the age values?**
- **Do you see any issues with the marital status values?**
- **Do you see any issues with the transaction amount values?**
- **Are the units of measure clear?**

References

- Chantal D. Larose, Daniel T. Larose, Data Science Using Python and R, Wiley, 2019.
- Daniel T. Larose, Chantal D. Larose, Discovering Knowledge in Data, an Introduction to Data Mining, Wiley, 2014.