CSE 40647/60647 Data Science (Spring 2018)
Lecture 2: Data Preprocessing: Data Description

Goals:

- Understand what is data object and attribute/feature;
- **Understand different attribute types;**
- **Understand different data set types;**
- Describe basic statistical descriptions
  - Describe and calculate central tendency
    - **Mean, Median, Mode, Frequency, Percentiles**
    - Population and sample
  - Describe and calculate outlier-ness
    - **Variance, Standard Deviation, Z-score**
    - Biased/Unbiased sample variance
    - Z-score normalization vs min-max normalization

Attribute types:

- Nominal (e.g., "red", "blue")
- Binary (e.g., yes/no, true/false)
- Ordinal (e.g., {freshman, sophomore, junior, senior}, {XS, S, M, L, XL, XXL})
- Numeric (e.g., ranking score, #faculty)

Data set type:

- Record data such as transaction data and document term-frequency data. Relational records in (highly structured) relational tables.
- Graph and networks such as transportation networks, World Wide Web, molecular structures, social or information networks.
- Ordered data such as video data, time-series data, sequential data and genetic sequence data.
- Spatial data (GIS)
- Image and multimedia data

In-class practice:

Given College CS Department's ***CS-Ranking Score*** *and* ***Faculty Number****, Calculate* ***Mean, Median, Mode, Frequency*** *and* ***Variance, Standard Deviation, Z-score****. Use* ***Min-Max Normalization*** *and* ***Z-Score Normalization*** *to normalize the attribute values.*

Discussion — Project:

What data objects do you want to study? What is the data mining task? What are the attributes?

Draw lines to match each attribute on the left side to an attribute type on the right side.

Hair color	Nominal
Marital status	Binary
Occupation	Ordinal
Driver license ID	Numeric
Zip code	
Gender	
Medical test	
Course grade	
Army ranking	
Age	
Body weight	

Name: **NetID:**

Please write down whatever question you have about this course: