

**Lecture 6 - Exploration and cleaning**

Lecturer: Professor Yu LI

Scriber: Chiu Chi Chung (SID: 1155174790)

1. Introduction

Lecture 6 mainly focused on

- a. Data cleaning: techniques to handle data quality problems
- b. Data exploration: techniques to represent the data as meaningful information.

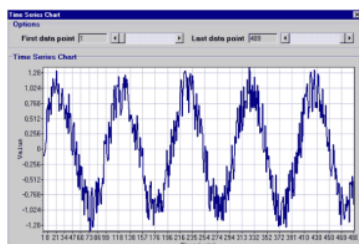
2. Data cleaning

Some common data quality problems were mentioned in the lecture, including noise, outliers, missing values, duplicate data, unnormalized data and categorical data. Their properties and corresponding solutions will be explained in the following.

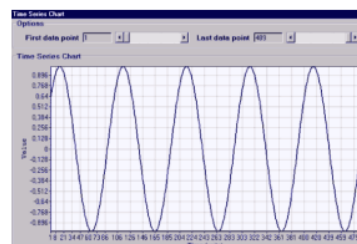
**Noise:**

Noise refers to modification of original values, like distortion of a person's voice when talking on a poor phone and "snow" on television screen. It will usually appear in recording, images or temporal data.

We could handle this problem with some denoising techniques, like rolling windows and auto-encoder models.



**A sine wave with noise**

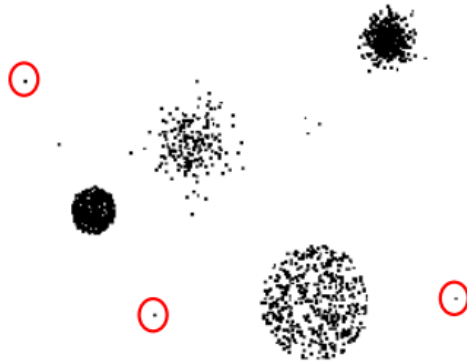


**The denoised sine wave**

**Outliers:**

Outliers are data objects with characteristics that are considerably different from most of the other data objects in the data set. For example, Elon Musk who has a net worth in the billions of dollars would be considered an outlier in terms of annual income.

We could handle this problem by analyzing the data distribution and removing those which are not in the range.

**Missing values:**

Missing values can be a result of uncollected information or inapplicable attributes. They are usually handled by elimination, estimation, ignorance or replacement.

**Duplicate data:**

Missing values can be a result of merging data from heterogeneous sources, like repeat submission from a person with multiple email addresses. They are usually handled by direct removal.

**Unnormalized data:**

Unnormalized data refers to attributes that are not on the similar level of measurement, like height and weight, usually appearing in data matrices.

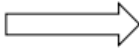
We usually use min-max normalization and z-score normalization which requires the data set to be in Gaussian distribution to prevent domination of particular attributes in difference comparison.

The direction for normalization is also a consideration with reference to the purpose of normalization.

### Categorical data:

Categorical data often refers to discrete and non-numerical data, like gender and nationality. They are usually handled by one-hot encoding which will expand the data matrix.

Person	Height(m)	Weight(kg)	Gender
P1	0.625	0.875	Male
P2	0	0	Female
P3	0.25	0.375	Female
P4	1	1	Male



Person	Height(m)	Weight(kg)	Male	Female
P1	0.625	0.875	1	0
P2	0	0	0	1
P3	0.25	0.375	0	1
P4	1	1	1	0

### Order of data cleaning:

The order of data cleaning matters, as it will affect the final results. For example, the exchange in order between outlier removal and normalization will change the scaling of data seriously. It is suggested you do the data cleaning in the following order:

1. Denoise data (if applicable)
2. Remove outliers
3. Handling missing data
4. Remove duplicates
5. Categorical data encoding
6. Data normalization

## 3. Data exploration

Data exploration mainly includes summary statistics and visualization.

### Summary statistics:

Summary statistics, like CGPA, are numbers that summarize properties of the data, including frequency, location and spread.

For measures of frequency, the percentage of time an attribute value occurs in a data set is usually calculated, like the percentage of females in all kinds of gender. Mode, the most frequent attribute value, is also used in analysis. Meanwhile, both of them are typically used with categorical data.

For measures of location, mean and median are most commonly used, but the former one is sensitive to outliers.

For measures of spread, range, variance, median absolute deviation (MAD), interquartile range and percentiles are all common tools in practice.

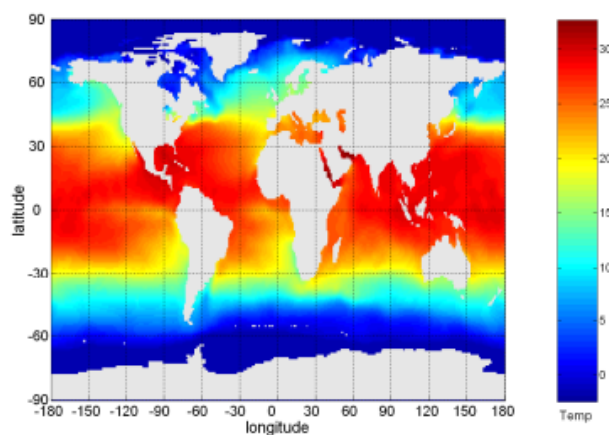
### **Visualization:**

Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported. It is commonly used with a few reasons:

1. Humans have a well-developed ability to analyze large amounts of information that is presented visually.
2. It can detect general patterns and trends.
3. It can detect outliers and unusual patterns.

The commonly used techniques includes:

1. **Heatmap:** a 2-dimensional data visualization technique that represents the magnitude of individual values within a dataset as a color.



2. **2D/3D Histograms:** a distribution of values of a single variable.

It divides the values into bins, showing a bar plot of the number of objects in each bin, with the height of each bar indicating the number of objects. The bin size is an important consideration because it affects the shape of the histograms.



3. **Box plots:** is another way of displaying and comparing the distribution of data.

