Chan Myae Kyaw
SID: 1155170386

BMSB3105 Data Analytics for Personalized Genomics and Precision Medicine
Lecture 6 - Data Exploration

Data Cleaning

Potential data quality problems -

- **Noise and outliers**
- **Missing values**
- **Duplicate data**
- **Unormalized data**
- **Categorical data**

What is noise data?

- Modifications in original data values

(basically alterations in original data, might be due to interference etc,.)

- Example: Distorted voice during phone call

What are outliers?

- Data objects that are significantly different thatn most other data objects in the dataset

What are missing values?

- Just as the name says, they are missing values
- Occurs when information is not collected, or when the attribute may not be applicable to all cases such as numbers of coke bottles sold is not applicable to students height.

What can we do about missing values?

- Eliminate data objects
- Estimate missing values
- Ignore the missing values when performing analysis
- Replace them with possible values (weighted by thier probabilities)

What is duplicate data?

- Just as the name says, they are duplicate data of one another.
- It has major problem when merging data from heterogeneous sources such as the same person with multiple email addresses

What about unormalized data?

- These are attributes not on the similar level of measurement

Then what is normalization?

- These are attributes ont he similar level of measurement

There is Min-Max normalization :

$$\triangleright v' = \frac{v - v^{min}}{v^{max} - v^{min}}$$

And Z-score normalization : Assumes the dataset follows Gaussian distribution.

$$\triangleright v' = \frac{v - \boldsymbol{Mean}(v)}{\mathbf{Std}(v)}$$

What is categorical data?

- Data that can be categorized

Note: Computers are better on handling numbers for categorical data. For that we can use one-hot coding.

What do we do when we have these type of data quality problems?

We do data cleaning
- Denoise data
- Remove outliers
- Handling missing data
- Remove duplicates
- Categorical data encoding
- Data normalization

Data exploration

Summary Statistics

- numbers that summarize properties of data.
- summerized properties include frequency, location and spread such as location, mean spread, and standard deviation.
- most summary statistics can be calculated in a single pass through the data

Measures of location

Mean

$$\triangleright mean(x) = \frac{1}{m}\sum_{i=1}^{m} x_i$$

- most common measure of location

- but very sensitive to outliers

Median (or) trimed mean

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}\left(x_{(r)} + x_{(r+1)}\right) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

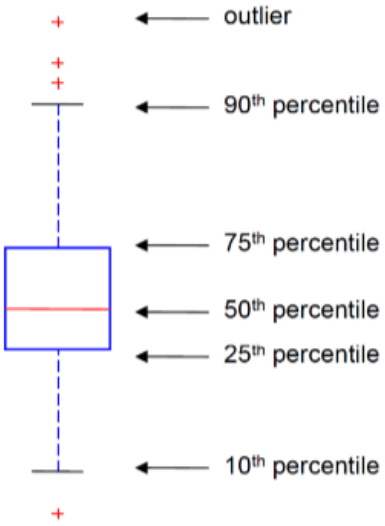
Range

- difference betwen max and min.

Variance (or) Standard deviation

$$\text{variance}(x) = \frac{1}{m-1}\sum_{i=1}^{m}\left(x_i - \text{mean}(x)\right)^2$$

- most common measure of spread
- sensitive to outliers



Other measures

- Median absolute deviation (MAD)

$$\text{median}(|x_1 - \text{mean}(x)|, \ldots, |x_m - \text{mean}(x)|)$$

- Interquartile range >> $X_{75\%}$- $X_{25\%}$

Percentiles

- a score below which a given percentage of scores in its frequency distribution falls

Chan Myae Kyaw
SID: 1155170386

- if there is p-th percentile, the p-th percentile refers to the x-value that is lower than x_p

Frequency and mode

Frequency - percentage of time the value occurs in the dataset
Mode - most frequent attribute value

Note: frequency and mode are typically used in categorical data

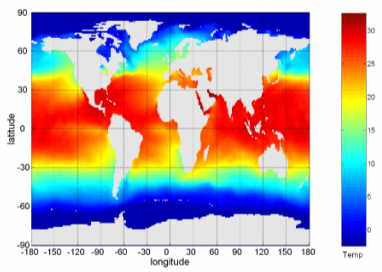
---

Exploratory visualization

Visualization

- conversion of data into visual format (tabular format) for analysis
- can detect general patterns and trends
- can detect outliers and unusal patterns

Can see the following as example: Sea surface temperature



Visualization Techniques
Histograms

- usually shows distribution of values of a single variable

  How?

  - divide the values into bins and show a bar plot of the nubmer of objects in each bin
  - the height of each bar indicates the number of objects
  - shape of histogram depends on the number of bins

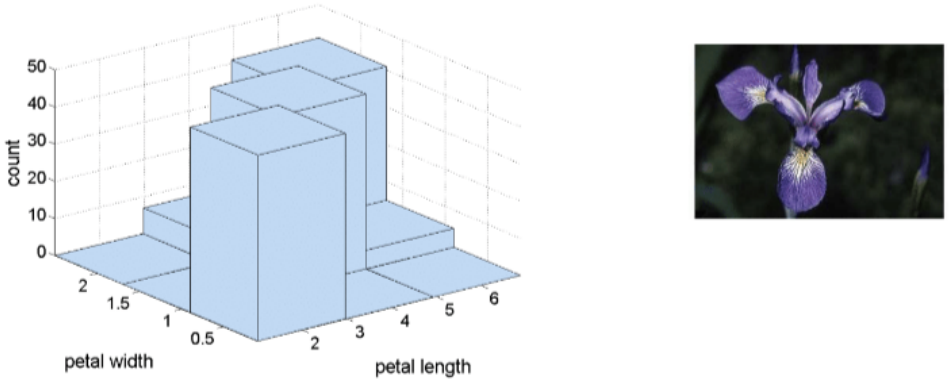
Can see example here: Petal width of Iris Plant data set

Chan Myae Kyaw
SID: 1155170386

Two-dimensional histograms

- shows the joint distribution of the values of two attributes

Another example in comparison to the last example: Petal width and length



Box Plots

- Another way of displaying and comparing data distribution.

An example box plot with percentiles: