Exercise on the understanding of the basic concept of data preprocessing

1. Data quality can be assessed in terms of accuracy, completeness, and consistency. Propose two other dimensions of data quality.

* Accessibility
* Timeliness

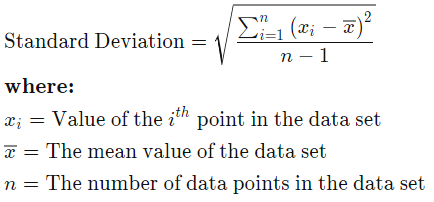
1. Give three additional commonly used statistical measures (i.e., not illustrated in this chapter) for the characterization of data dispersion, and discuss how they can be computed efficiently in large databases.

Mean

Mode

Median

1. The Mean/average deviation
2. The Standard Deviation: *The standard deviation is a statistic that measures the dispersion of a dataset relative to its*[*mean*](https://www.investopedia.com/terms/m/mean.asp)*and is calculated as the square root of the*[*variance*](https://www.investopedia.com/terms/v/variance.asp)*. The standard deviation is*[*calculated*](https://www.investopedia.com/ask/answers/021115/what-difference-between-standard-deviation-and-z-score.asp)*as the square root of variance by determining each data point's deviation relative to the mean. If the data points are further from the mean, there is a higher deviation within the data set; thus, the more spread out the data, the higher the standard deviation.*



How standard deviation can be applied on large database

1. Calculate the mean or average of each data set. To do this, add up all the numbers in a data set and divide by the total number of pieces of data. For example, if you have four numbers in a data set, divide the sum by four. This is the *mean* of the data set.
2. Subtract the *deviance* of each piece of data by subtracting the mean from each number. Note that the variance for each piece of data may be a positive or negative number.
3. Square each of the deviations.
4. Add up all of the squared deviations.
5. Divide this number by one less than the number of items in the data set. For example, if you had four numbers, divide by three.
6. Calculate the square root of the resulting value. This is the [*sample standard deviation*.](https://www.thoughtco.com/sample-standard-deviation-problem-609528)
7. *The Coefficient of variation*: In [probability theory](https://en.wikipedia.org/wiki/Probability_theory) and [statistics](https://en.wikipedia.org/wiki/Statistics), the coefficient of variation (CV), also known as relative standard deviation (RSD), is a [standardized](https://en.wikipedia.org/wiki/Standardized_(statistics)) measure of [dispersion](https://en.wikipedia.org/wiki/Statistical_dispersion) of a [probability distribution](https://en.wikipedia.org/wiki/Probability_distribution) or [frequency distribution](https://en.wikipedia.org/wiki/Frequency_distribution). It is often expressed as a percentage, and is defined as the ratio of the [standard deviation](https://en.wikipedia.org/wiki/Standard_deviation) {\displaystyle \ \sigma }to the [mean](https://en.wikipedia.org/wiki/Mean) {\displaystyle \ \mu } (or its [absolute value](https://en.wikipedia.org/wiki/Absolute_value), {\displaystyle |\mu |}). The CV or RSD is widely used in [analytical chemistry](https://en.wikipedia.org/wiki/Analytical_chemistry) to express the precision and repeatability of an [assay](https://en.wikipedia.org/wiki/Assay). It is also commonly used in fields such as [engineering](https://en.wikipedia.org/wiki/Engineering) or [physics](https://en.wikipedia.org/wiki/Physics) when doing quality assurance studies

## How to calculate coefficient of variation

. The basic formula used in mathematics sets the coefficient of variation equal to standard of deviation over mean:

**CV = Standard of deviation / Mean x 100%**

**CV = Volatility / Expected return x 100%**

**The are 4 steps for calculating coefficient of valiation**

### 1. Determine volatility

To find volatility or standard deviation, subtract the mean price for the period from each price point. To convert the difference into variance, square, sum and average the answer. The square root of the variance becomes a viable percentage for volatility.

### 2. Determine expected return

To find the expected return, multiply potential outcomes or returns by their chances of occurring. The sum of all collected answers becomes the expected return. At this point, both figures are ready for the formula.

### 3. Divide

With both volatility and expected return figures calculated, divide them by each other. Most answers come in the form of decimals. However, CV requires a percentage.

### 4. Multiply by 100%

To convert to a percentage, multiply decimals by 100%. It moves the decimal place, creating either a whole number or decimal percentage. The final answer is the coefficient of variation.

THE GINI COEFFICIENT AND THE LORENZ CURVE

The Gini coefficient is equal to the area below the line of perfect equality (0.5 by definition) minus the area below the Lorenz curve, divided by the area below the line of perfect equality. In other words, it is double the area between the Lorenz curve and the line of perfect equality

1. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.

## Mean or Median Imputation: When data is missing at random, we can use list-wise or pair-wise deletion of the missing observations.

## Multivariate Imputation by Chained Equations (MICE)

MICE assumes that the missing data are Missing at Random (MAR). It imputes data on a variable-by-variable basis by specifying an imputation model per variable. MICE uses predictive mean matching (PMM) for continuous variables, logistic regressions for binary variables, bayesian polytomous regressions for factor variables, and proportional odds model for ordered variables to impute missing data.

## 3. Random Forest

Random forest is a non-parametric imputation method applicable to various variable types that works well with both data missing at random and not missing at random. Random forest uses multiple [decision trees](https://www.datascience.com/blog/random-forests-decision-trees-ensemble-methods" \t "_blank) to estimate missing values and outputs OOB (out of bag) imputation error estimates

1. What are the major tasks in data preprocessing?

Data cleaning

Data integration

Data transformation

Data reduction

Data discretization