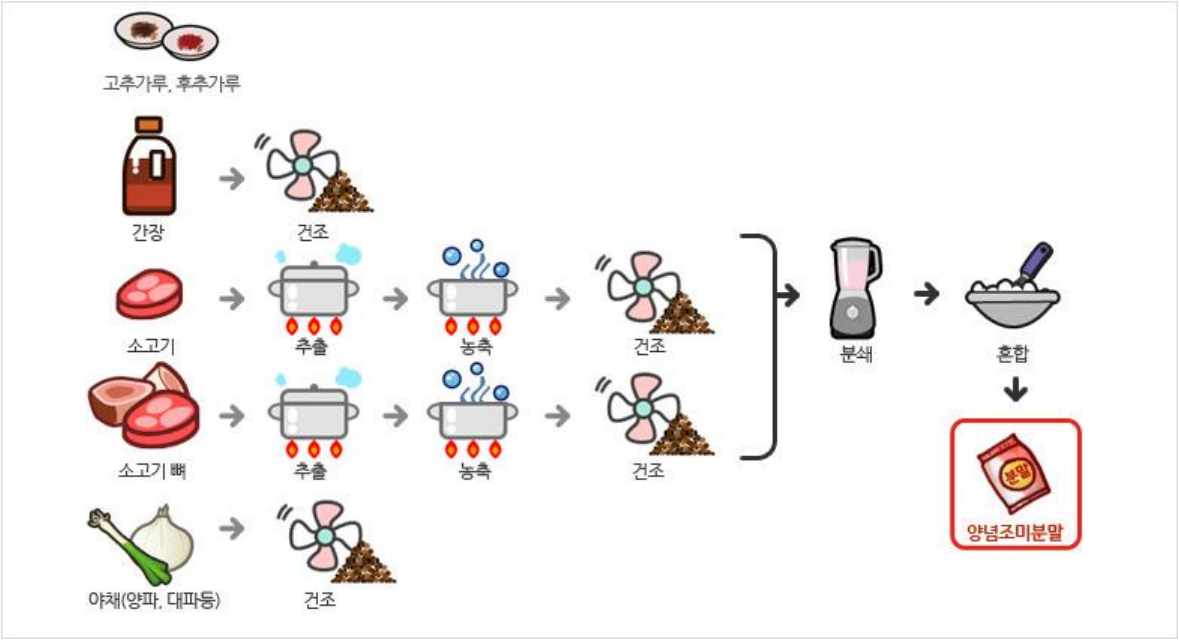
Data Quality

Hyerim Bae
Department of Industrial Engineering, Pusan National University
hrbae@pusan.ac.kr



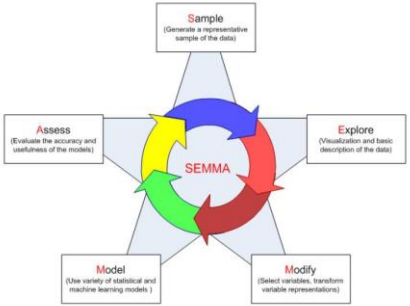
http://www.nongshim.com/ramyun/show_knowledge?groupCode=004&groupId=7

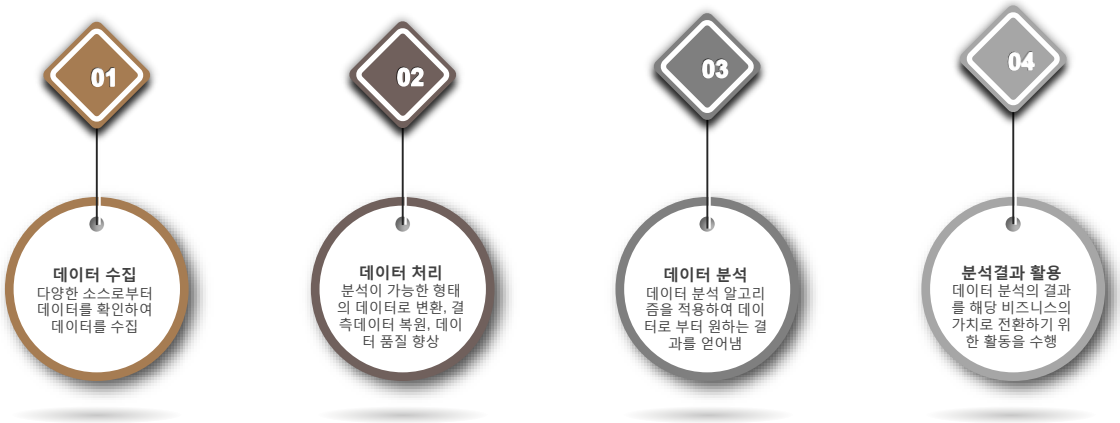
Contents

Data quality improvement

Data analytics process

- 절차를 정의하는 방식
  - SEMMA
    - Sampling, Explore, Modify, Modeling, Assesment
  - CRISP-DM:
    - B-understanding, D-understanding, D-preparation
    - Modeling, evaluation, deployment
  - KDD
    - Selection, Preprocessing, Transformation
    - Data mining, Interpretation(Evaluation)





Garbage In Garbage Out !

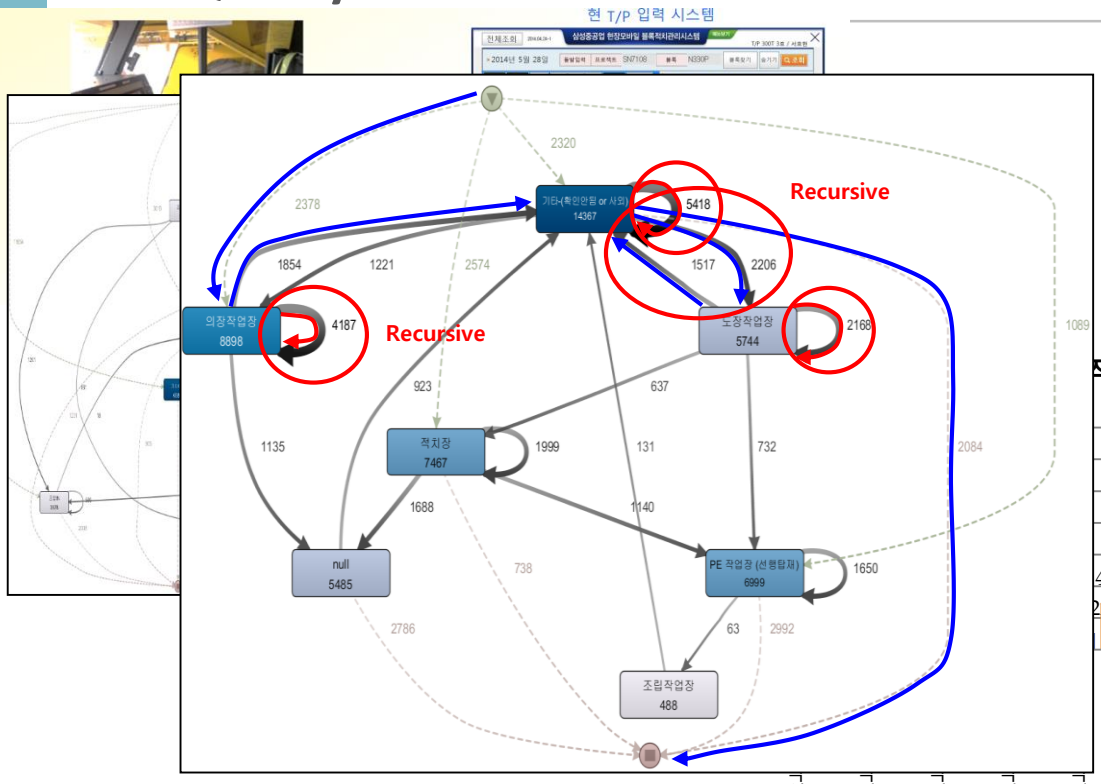
Data imperfection

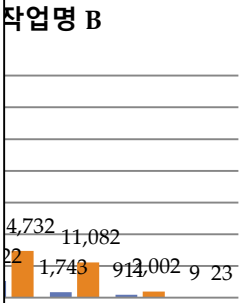
- Missing data
  - 자료 누락
- Incorrect data
  - 잘못된 코드
- Imprecise data
  - 잘못된 측정 데이터
- Irrelevant data
  - 예측에 사용하기 힘든 자료

Table 1. Manifestation of quality issues in event log entities [6].

		Event log entities								
		Case	Event	Relationship	Case attrs.	Position	Activity name	Timestamp	Resource	Event attrs.
Event log quality issues	Missing data	I1	I2	I3	I4	I5	I6	I7	I8	I9
	Incorrect data	I10	I11	I12	I13	I14	I15	I16	I17	I18
	Imprecise data			I19	I20	I21	I22	I23	I24	I25
	Irrelevant data	I26	I27							

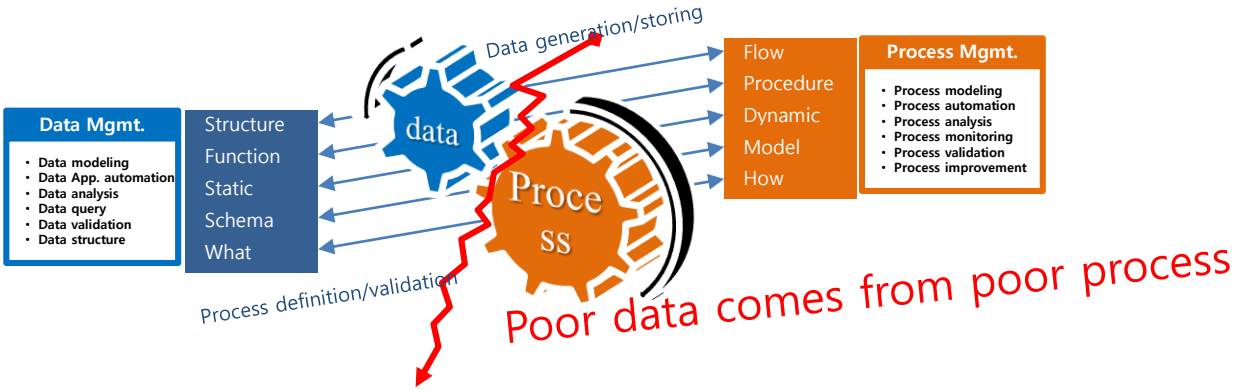
Data Quality의 중요성



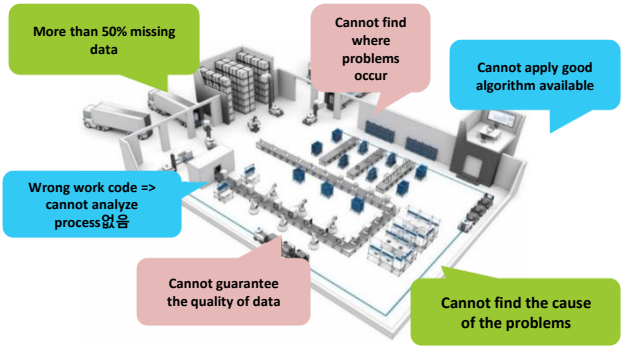


작업 A 작업 B 작업 C 작업 D 작업 E 작업 F 미확인 작업 null

Big-data vs. Big-process

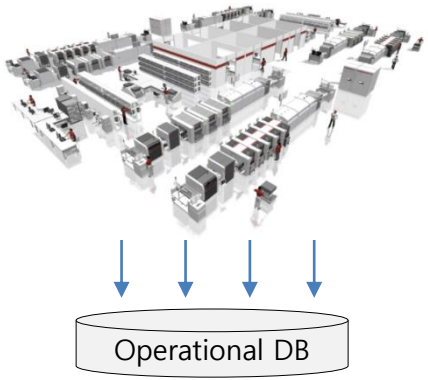


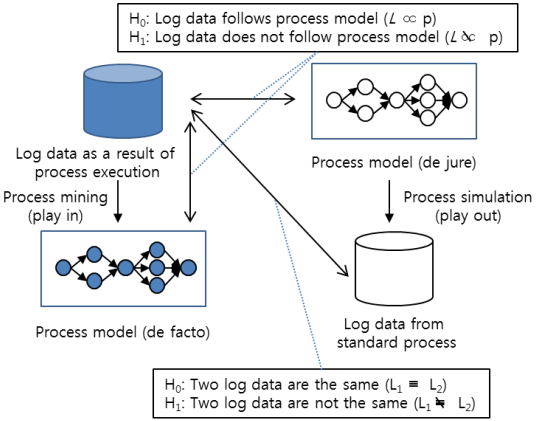
Data-Process Separation	
Undefined process	Missing data
Error in process definition	Data error
Unstandardized process	Low data quality
Process that are not digitalized	Low data value



Data-Process compliance

- 만약 data-process separation을 없앨 수 있다면?





Data Quality

1.Consistency - logical relations
- two similar IDs for two different employees
- a non-existent entry in another table

2.Accuracy - the real state of things
- All calculations based on such data show the true result.

3.Completeness - all needed elements
- lots of sensor data but there's no info about the exact sensor locations

4.Auditability - maintenance and control
- data quality audits regularly or on demand will help to ensure a higher level of data adequacy

5.Orderliness - structure and format
- the temperature in the oven has to be measured in Fahrenheit and can't be -14 °F.

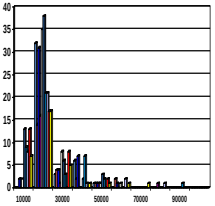
Data Reduction

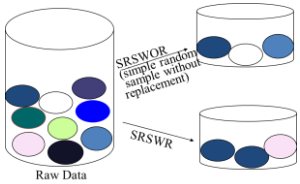
Data Reduction Strategies

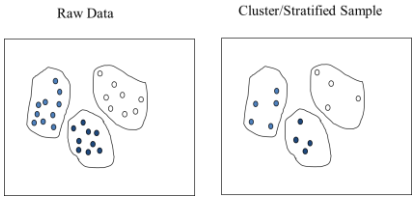
- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- Data reduction strategies
  - Dimensionality reduction, e.g., remove unimportant attributes
    - Wavelet transforms
    - Principal Components Analysis (PCA)
    - Feature subset selection, feature creation
  - Numerosity reduction (some simply call it: Data Reduction)
    - Regression and Log-Linear Models
    - Histograms, clustering, sampling
    - Data cube aggregation
  - Data compression

Data Reduction : Numerosity Reduction (NOT dimension reduction)

- Reduce data volume by choosing alternative, *smaller forms* of data representation
- **Parametric methods** (e.g., regression)
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - Ex.: Log-linear models—obtain value at a point in *m*-D space as the product on appropriate marginal subspaces
- **Non-parametric** methods
  - Do not assume models
  - Major families: histograms, clustering, sampling, …

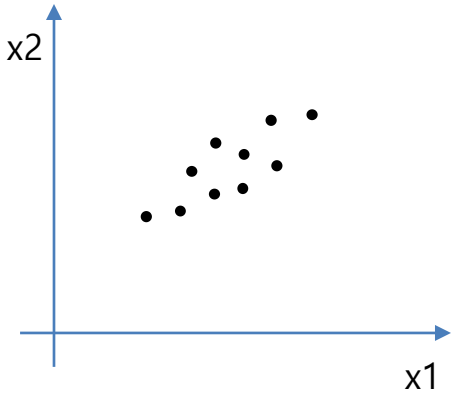






Dimensional reduction

- Visualization
- Reduce noise
- Preserve only useful information
- Less time/space complexity



Principal Components Analysis

Goal: Reduce a set of numerical variables.
상관관계가 높은 측정치들이 존재할 때 유리 => 변수 축소의 여지가 많다.

The idea: Remove the overlap of information between these variable. ["Information" is measured by the sum of the variances of the variables.]

Final product: A smaller number of numerical variables that contain most of the information

For Quantitative variable
What about categorical variable? => 대응분석(Correspondence anlysis)

Principal Components Analysis

How does PCA do this?

•Create new variables that are (weighted) linear combinations of the original variables (i.e., they are weighted averages of the original variables).

•These linear combinations are uncorrelated (no information overlap), and only a few of them contain most of the original information.

•The new variables are called *principal components*.

Example – Breakfast Cereals

Name: name of cereal
mfr: manufacturer
type: cold or hot
calories: calories per serving
protein: grams
fat: grams
sodium: mg.
fiber: grams
Carbo(복합탄수화물): grams complex carbohydrates
sugars: grams
Potass(칼륨): mg.
vitamins: % FDA rec
shelf: display shelf (진열대 높이)
weight: oz. 1 serving
cups: in one serving
rating: consumer reports

name	mfr	type	calories	protein	…	rating
100%_Bran	N	C	70	4	…	68
100%_Natural_Bran	Q	C	120	3	…	34
All-Bran	K	C	70	4	…	59
All-Bran_with_Extra_Fiber	K	C	50	4	…	94
Almond_Delight	R	C	110	2	…	34
Apple_Cinnamon_Cheerios	G	C	110	2	…	30
Apple_Jacks	K	C	110	2	…	33
Basic_4	G	C	130	3	…	37
Bran_Chex	R	C	90	2	…	49
Bran_Flakes	P	C	90	3	…	53
Cap'n'Crunch	Q	C	120	1	…	18
Cheerios	G	C	110	6	…	51
Cinnamon_Toast_Crunch	G	C	120	1	…	20

Covariance and correlation

- Covariance

$$Cov(X,Y) = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})$$

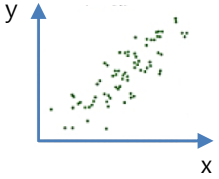
where

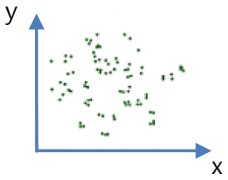
$$\bar{X} = \frac{1}{N}\sum_{i=1}^{n} X_i \text{ and } \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

are the means of X, Y

- Correlation

$$\frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Yz)}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})/(n-1)}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2/(n-1)}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2/(n-1)}}$$





$$Cor(\text{calories, ratings}) = \frac{-188.68}{\sqrt{379.63}\sqrt{197.2}} = -0.69$$

  - 69% of variance is shared between the two

Cov. matrix

- Covariance matrix

$$C=\begin{pmatrix} cov(x,x) & cov(x,y) \\ cov(x,y) & cov(y,y) \end{pmatrix}$$

$$=\begin{pmatrix} \frac{1}{n}\sum(x_i-m_x)^2 & \frac{1}{n}\sum(x_i-m_x)(y_i-m_y) \\ \frac{1}{n}\sum(x_i-m_x)(y_i-m_y) & \frac{1}{n}\sum(y_i-m_y)^2 \end{pmatrix}$$

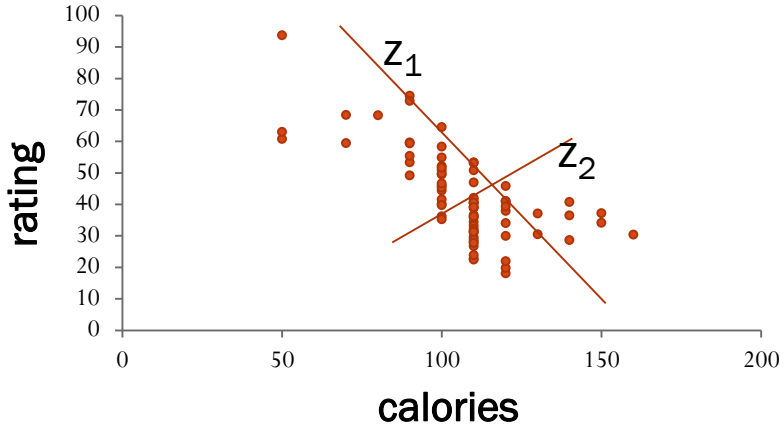
	calories	ratings
calories	379.63	-189.68
ratings	-189.68	197.32

- Total variance (="information") is sum of individual variances: 379.63 + 197.32
- Calories accounts for 379.63/(197.32+379.63) = 66%
- PCA
  - Want to find a new variable explain most of the variance

Principal Components

Z_1 and Z_2 are two linear combinations.

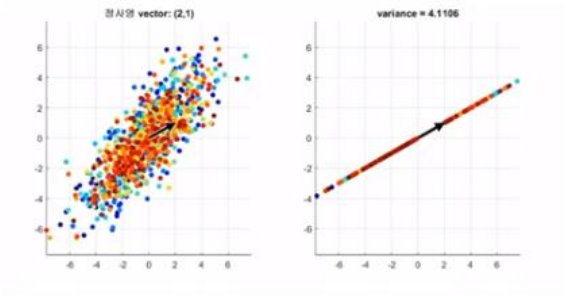
- Z_1 has the highest variation (spread of values)
- Z_2 has the lowest variation



Eigen vector

- PCA is to find a linear combination of the variables
  - Which should explain most of the variances
  - Eigen vector
    - a nonzero vector with the same direction (change of the length) after linear transformation

- PCA
  - Projecting data onto eigenvector of Cov. matrix



Generalization

X_1, X_2, X_3, … X_p, original *p* variables

Z_1, Z_2, Z_3, … Z_p, weighted averages of original variables

All pairs of Z variables have 0 correlation

Order Z's by variance (z_1 largest, Z_p smallest)

Usually the first few Z variables contain most of the information, and so the rest can be dropped.

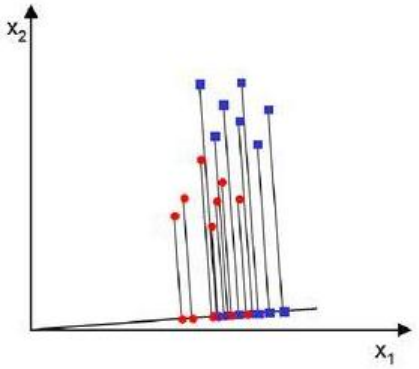
PCA on full data set

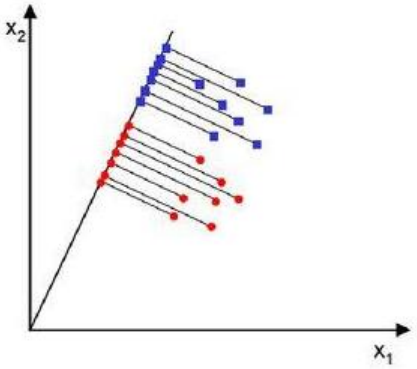
- First 6 components shown
- First 2 capture 93% of the total variation
- Note: data differ slightly from text

Variable	1	2	3	4	5	6
calories	0.07624155	-0.01066097	0.61074823	-0.61706442	0.45754826	0.12601775
protein	-0.00146212	0.00873588	0.00050506	0.0019389	0.05533375	0.10379469
fat	-0.00013779	0.00271266	0.01596125	-0.02595884	-0.01839438	-0.12500292
sodium	0.98165619	0.12513085	-0.14073193	-0.00293341	0.01588042	0.02245871
fiber	-0.00479783	0.03077993	-0.01684542	0.02145976	0.00872434	0.271184
carbo	0.01486445	-0.01731863	0.01272501	0.02175146	0.35580006	-0.56089228
sugars	0.00398314	-0.00013545	0.09870714	-0.11555841	-0.29906386	0.62323487
potass	-0.119053	0.98861349	0.03619435	-0.042696	-0.04644227	-0.05091622
vitamins	0.10149482	0.01598651	0.7074821	0.69835609	-0.02556211	0.01341988
shelf	-0.00093911	0.00443601	0.01267395	0.00574066	-0.00823057	-0.05412053
weight	0.0005016	0.00098829	0.00369807	-0.0026621	0.00318591	0.00817035
cups	0.00047302	-0.00160279	0.00060208	0.00095916	0.00280366	-0.01087413
rating	-0.07615706	0.07254035	-0.30776858	0.33866307	0.75365263	0.41805118

Variance	7204.161133	4833.050293	498.4260864	357.2174377	72.47863007	4.33980322
Variance%	55.52834702	37.25226212	3.84177661	2.75336623	0.55865192	0.0334504
Cum%	55.52834702	92.78060913	96.62238312	99.37575531	99.93440247	99.96785736

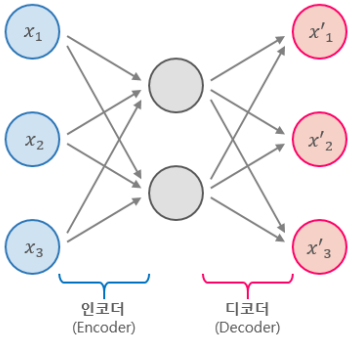
Data reduction by LDA





Data reduction using NN: Autoencoder

- Purpose of autoencoder
  - Data reduction



Summary

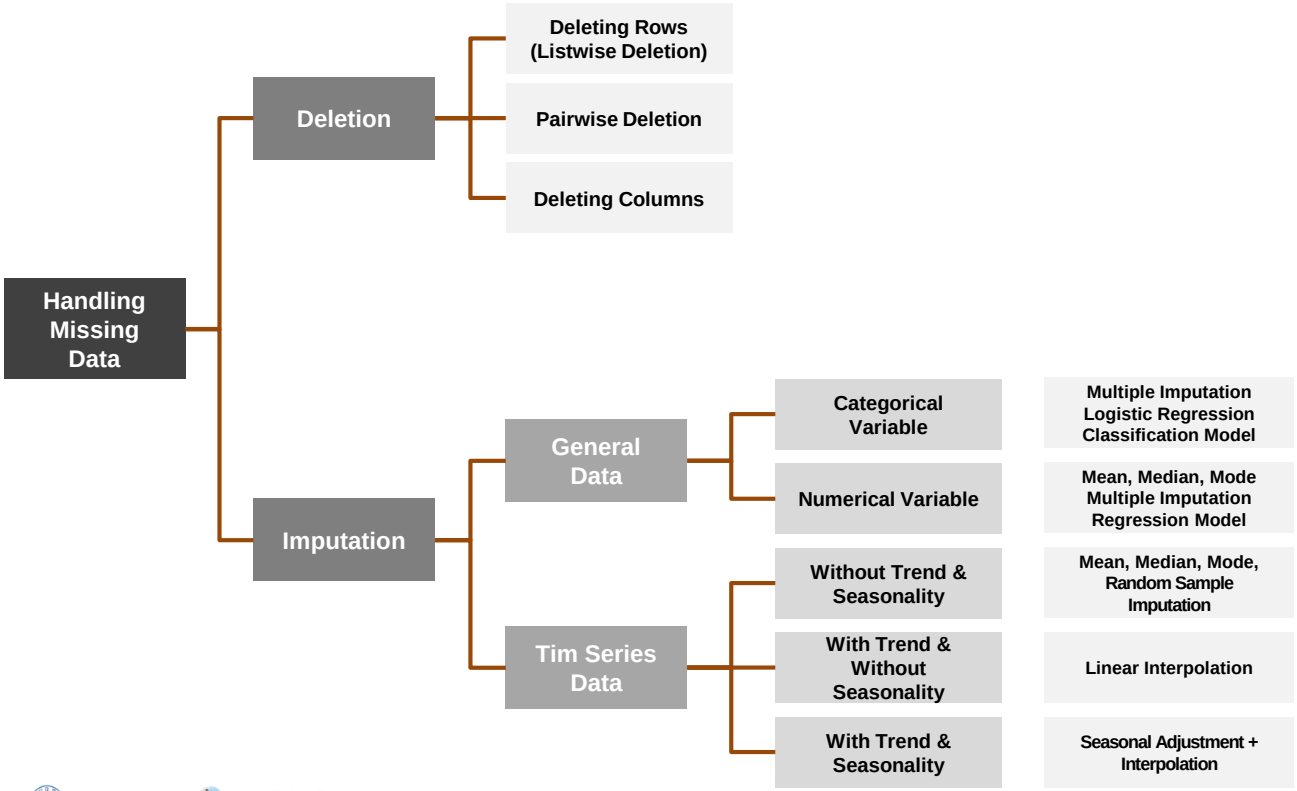
- **Data summarization** is an important for data exploration
- **Data summaries** include numerical metrics (average, median, etc.) and graphical summaries
- **Data reduction** is useful for compressing the information in the data into a smaller subset
  - Categorical variables can be reduced by combining similar categories
  - Principal components analysis transforms an original set of numerical data into a smaller set of weighted averages of the original data that contain most of the original information in less variables.

Data Imputation

Understanding about Missing Data

- One of the most common problems we have faced in Data Cleaning/Exploratory Analysis is handling the missing values.
- The type of missing values was firstly classified by Rubin and the missing values have been classified into the following three kinds.
  - **Missing at Random (MAR) :** Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data (자료 내의 다른 변수와 관련)
  - **Missing Completely at Random (MCAR) :** The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.
  - **Missing not at Random (MNAR) :** Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable's value (e.g. Let's assume that females generally don't want to reveal their ages! Here the missing value in age variable is impacted by gender variable) (Missing여부가 해당변수의 값에 의해서 결정)

Method for handling missing data



Method for handling missing data

- Listwise deletion (complete-case analysis)
  - By far the most common approach to the missing data is to simply omit those cases with the missing data and analyse the remaining data.
  - This approach is known as the complete case (or available case) analysis or list-wise deletion.

	Mobile Package	Download Speed	Data Limit Usage
y_1	Fast	157	80%
y_2	Lite	99	70%
y_3	Fast	167	10%
y_4	Fast	NA	80%
y_5	Lite	76	70%
y_6	Fast	155	10%
y_7	NA	NA	95%
y_8	Lite	76	77%
y_9	Fast	180	NA

	Mobile Package	Download Speed	Data Limit Usage
y_1	Fast	157	80%
y_2	Lite	99	70%
y_3	Fast	167	10%
y_5	Lite	76	70%
y_6	Fast	155	10%
y_8	Lite	76	77%

Method for handling missing data

- Pairwise deletion (available-case analysis)
  - Only the missing observations are ignored and analysis is done on variables present.
  - If there is missing data elsewhere in the data set, the existing values are used. Since a pairwise deletion uses all information observed, it preserves more information than the listwise deletion.

	Mobile Package	Download Speed	Data Limit Usage
y_1	Fast	157	80%
y_2	Lite	99	70%
y_3	Fast	167	10%
y_4	Fast	NA	80%
y_5	Lite	76	70%
y_6	Fast	155	10%
y_7	NA	NA	95%
y_8	Lite	76	77%
y_9	Fast	180	NA

	Mobile Package	Download Speed	Data Limit Usage
y_1	Fast	157	80%
y_2	Lite	99	70%
y_3	Fast	167	10%
y_4	Fast		80%
y_5	Lite	76	70%
y_6	Fast	155	10%
y_7			95%
y_8	Lite	76	77%
y_9	Fast	180	

Method for handling missing data

- Column deletion (available-variable analysis)
  - If there are too many data missing for a variable it may be an option to delete the variable or the column from the dataset.
  - This should be the last option and need to check if model performance improves after deletion of variable.

	Mobile Package	Download Speed	Data Limit Usage
y_1	Fast	NA	80%
y_2	Lite	NA	70%
y_3	Fast	167	10%
y_4	Fast	NA	80%
y_5	Lite	NA	70%
y_6	Fast	155	10%
y_7	Fast	NA	95%
y_8	Lite	NA	77%
y_9	Fast	180	80%

	Mobile Package	Data Limit Usage
y_1	Fast	80%
y_2	Lite	70%
y_3	Fast	10%
y_4	Fast	80%
y_5	Lite	70%
y_6	Fast	10%
y_7	Fast	95%
y_8	Lite	77%
y_9	Fast	80%

Method for handling missing data

- Simple Imputation(Mean, Median and Mode)
  - In this simple imputation technique goal is to replace missing data with statistical estimates of the missing values. Mean, Median or Mode can be used as imputation value.
  - Ex) Mean = 130, Median = 155, Mode = 200

	Mobile Package	Download Speed	Data Limit Usage
y_1	Fast	157	80%
y_2	Lite	99	70%
y_3	Fast	167	10%
y_4	Fast	NA	80%
y_5	Lite	76	70%
y_6	Fast	155	10%
y_7	Fast	NA	95%
y_8	Lite	76	77%
y_9	Fast	180	80%

	SI with Mean	SI with Median	SI with Mode
y_1	157	157	157
y_2	99	99	99
y_3	167	167	167
y_4	130	155	76
y_5	76	76	76
y_6	155	155	155
y_7	130	155	76
y_8	76	76	76
y_9	180	180	180

Method for handling missing data

- Time Series Specific Method (LOCF, NOCB and Linear Interpolation)
  - Last Observation Carried Forward(LOCF)
    - If data is time-series data, one of the most widely used imputation methods.
    - Whenever a value is missing, it is replaced with the last observed value.

	Date	Download Speed	Data Limit Usage
y_1	1-MAR	157	80%
y_2	2-MAR	99	70%
y_3	3-MAR	167	10%
y_4	4-MAR	NA	80%
y_5	5-MAR	76	70%
y_6	6-MAR	155	10%
y_7	7-MAR	NA	95%
y_8	8-MAR	76	77%
y_9	9-MAR	NA	80%

	Date	Download Speed	Data Limit Usage
y_1	1-MAR	157	80%
y_2	2-MAR	99	70%
y_3	3-MAR	167	10%
y_4	4-MAR	167	80%
y_5	5-MAR	76	70%
y_6	6-MAR	155	10%
y_7	7-MAR	155	95%
y_8	8-MAR	76	77%
y_9	9-MAR	76	80%

Method for handling missing data

- Time Series Specific Method (LOCF, NOCB and Linear Interpolation)
  - Next Observation Carried Backward(NOCB)
    - A similar approach like LOCF which works in the opposite direction by taking the first observation after the missing value and carrying it backward

	Date	Download Speed	Data Limit Usage
y_1	1-MAR	157	80%
y_2	2-MAR	99	70%
y_3	3-MAR	167	10%
y_4	4-MAR	NA	80%
y_5	5-MAR	76	70%
y_6	6-MAR	NA	10%
y_7	7-MAR	NA	95%
y_8	8-MAR	76	77%
y_9	9-MAR	180	80%

	Date	Download Speed	Data Limit Usage
y_1	1-MAR	157	80%
y_2	2-MAR	99	70%
y_3	3-MAR	167	10%
y_4	4-MAR	76	80%
y_5	5-MAR	76	70%
y_6	6-MAR	76	10%
y_7	7-MAR	76	95%
y_8	8-MAR	76	77%
y_9	9-MAR	180	80%

Method for handling missing data

- Time Series Specific Method (LOCF, NOCB and Linear Interpolation)
  - Linear Interpolation
    - Interpolation is a mathematical method that adjusts a function to data and uses this function to extrapolate the missing data.
    - The simplest type of interpolation is the linear interpolation, that makes a mean between the values before the missing data and the value after.

	Date	Download Speed	Data Limit Usage
y_1	1-MAR	157	80%
y_2	2-MAR	99	70%
y_3	3-MAR	167	10%
y_4	4-MAR	NA	80%
y_5	5-MAR	76	70%
y_6	6-MAR	NA	10%
y_7	7-MAR	150	95%
y_8	8-MAR	76	77%
y_9	9-MAR	180	80%

	Date	Download Speed	Data Limit Usage
y_1	1-MAR	157	80%
y_2	2-MAR	99	70%
y_3	3-MAR	167	10%
y_4	4-MAR	121.5	80%
y_5	5-MAR	76	70%
y_6	6-MAR	113	10%
y_7	7-MAR	150	95%
y_8	8-MAR	76	77%
y_9	9-MAR	180	80%

(167+76)/2 = 121.5

(76+150)/2 = 113

Method for handling missing data

- Time Series Specific Method (LOCF, NOCB and Linear Interpolation)
  - Linear Interpolation
    - Interpolation is a mathematical method that adjusts a function to data and uses this function to extrapolate the missing data.
    - The simplest type of interpolation is the linear interpolation, that makes a mean between the values before the missing data and the value after.

	Date	Download Speed	Data Limit Usage
y_1	1-MAR	157	80%
y_2	2-MAR	99	70%
y_3	3-MAR	167	10%
y_4	4-MAR	NA	80%
y_5	5-MAR	76	70%
y_6	6-MAR	NA	10%
y_7	7-MAR	150	95%
y_8	8-MAR	76	77%
y_9	9-MAR	180	80%

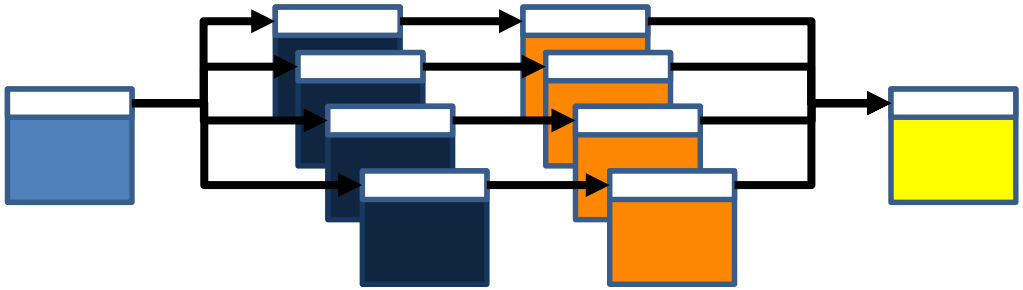
	Date	Download Speed	Data Limit Usage
y_1	1-MAR	157	80%
y_2	2-MAR	99	70%
y_3	3-MAR	167	10%
y_4	4-MAR	121.5	80%
y_5	5-MAR	76	70%
y_6	6-MAR	113	10%
y_7	7-MAR	150	95%
y_8	8-MAR	76	77%
y_9	9-MAR	180	80%

(167+76)/2 = 121.5

(76+150)/2 = 113

Method for handling missing data

- Multiple Imputation
  - Multiple imputation (MI) is a statistical technique for dealing with missing data.
  - The Multiple imputation includes the following 3 components.
    - Generate missing value: To use the distribution of the observed data to estimate a set of plausible values for the missing data.
    - Imputation and Analysis: Missing values are replaced by several estimated values, and are analyzed separately equally to obtain parameter estimates.
    - Pooling: The estimates are combined to obtain a set of parameter estimates.



  - MI technique has various methods depending on how data is imputed, and there are various imputation method such as MICE (Multiple Imputation by Chain Equation), Random Forest Imputation, KNN Imputation, Expectation-Maximization Imputation.

Method for handling missing data

Understanding for MICE – Single Iteration

Age	Income	Gender
33	NA	F
18	12,000	NA
NA	13,542	M

Method for handling missing data

Understanding for MICE – Single Iteration

Age	Income	Gender
33	NA	F
18	12,000	NA
NA	13,542	M

Age	Income	Gender
33	12.771	F
18	12,000	F
25.5	13,542	M

Method for handling missing data

Understanding for MICE – Single Iteration

Age	Income	Gender
33	NA	F
18	12,000	NA
NA	13,542	M

Simple Imputation using mean

Age	Income	Gender
33	12.771	F
18	12,000	F
25.5	13,542	M

Age back to NA

Age	Income	Gender
33	12.771	F
18	12,000	F
NA	13,542	M

Method for handling missing data

Understanding for MICE – Single Iteration

Age	Income	Gender
33	NA	F
18	12,000	NA
NA	13,542	M

Simple Imputation using mean

Age	Income	Gender
33	12.771	F
18	12,000	F
25.5	13,542	M

Age back to NA

Age	Income	Gender
33	12.771	F
18	12,000	F
NA	13,542	M

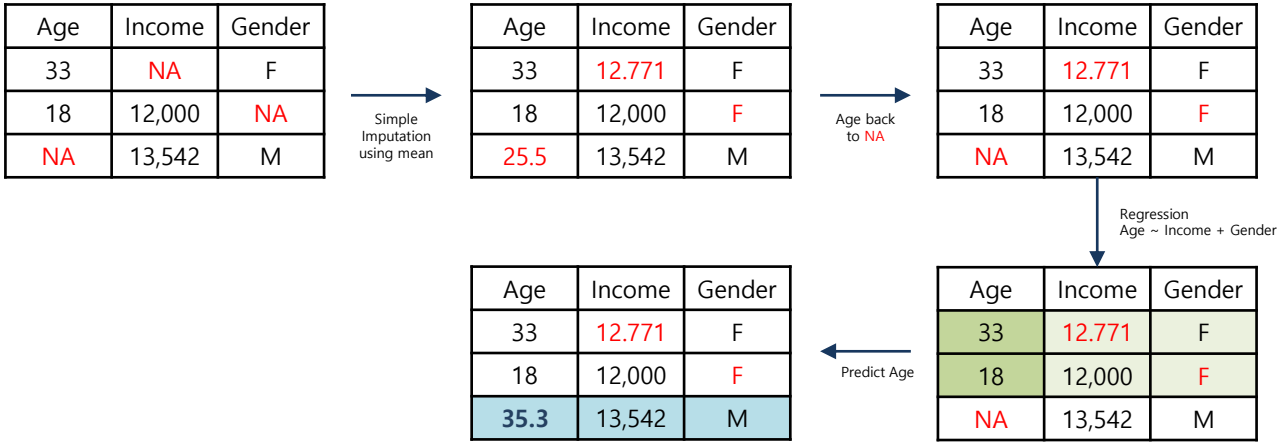
Regression
Age ~ Income + Gender

Predict Age

Age	Income	Gender
33	12.771	F
18	12,000	F
NA	13,542	M

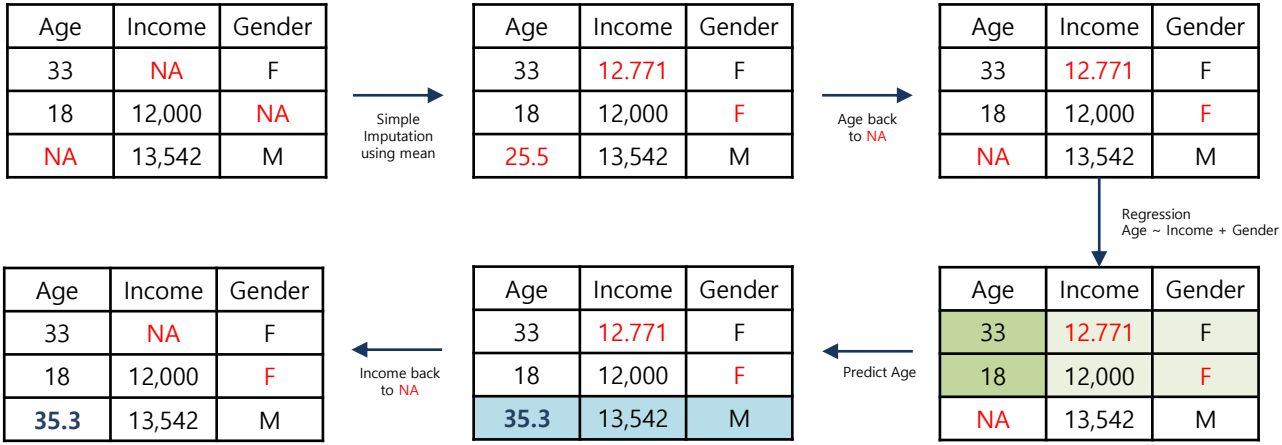
Method for handling missing data

Understanding for MICE – Single Iteration



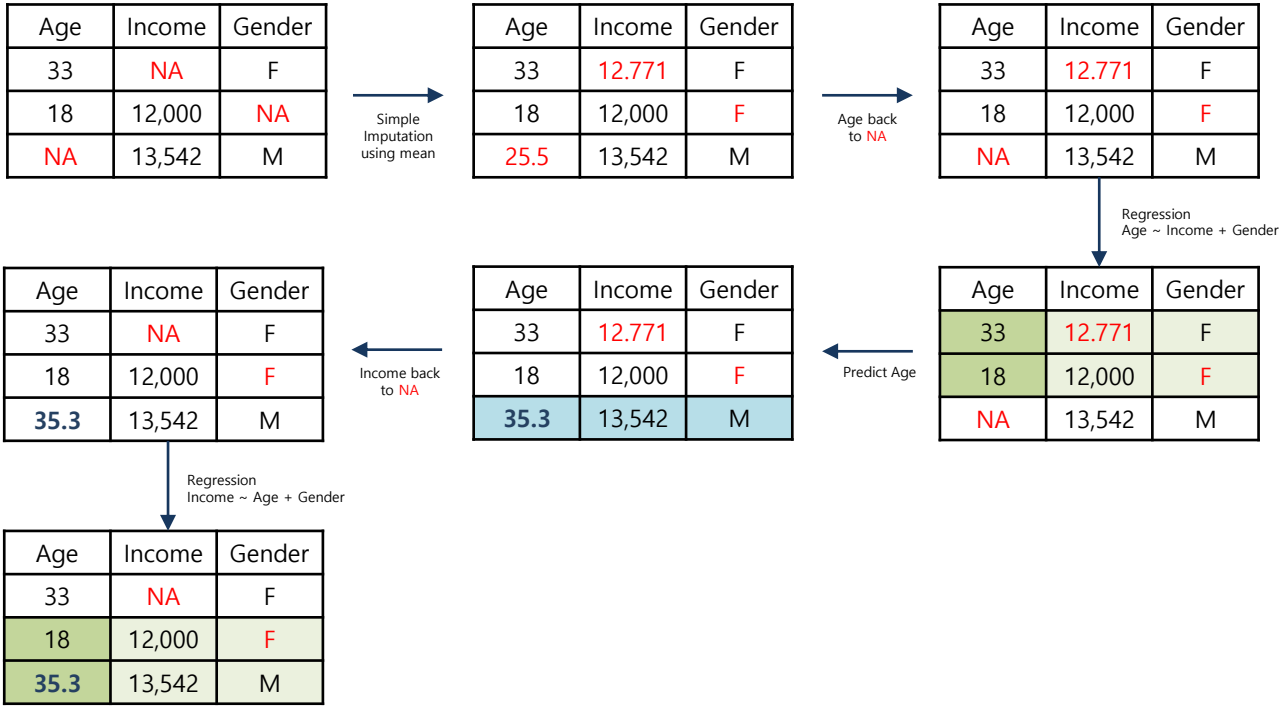
Method for handling missing data

Understanding for MICE – Single Iteration



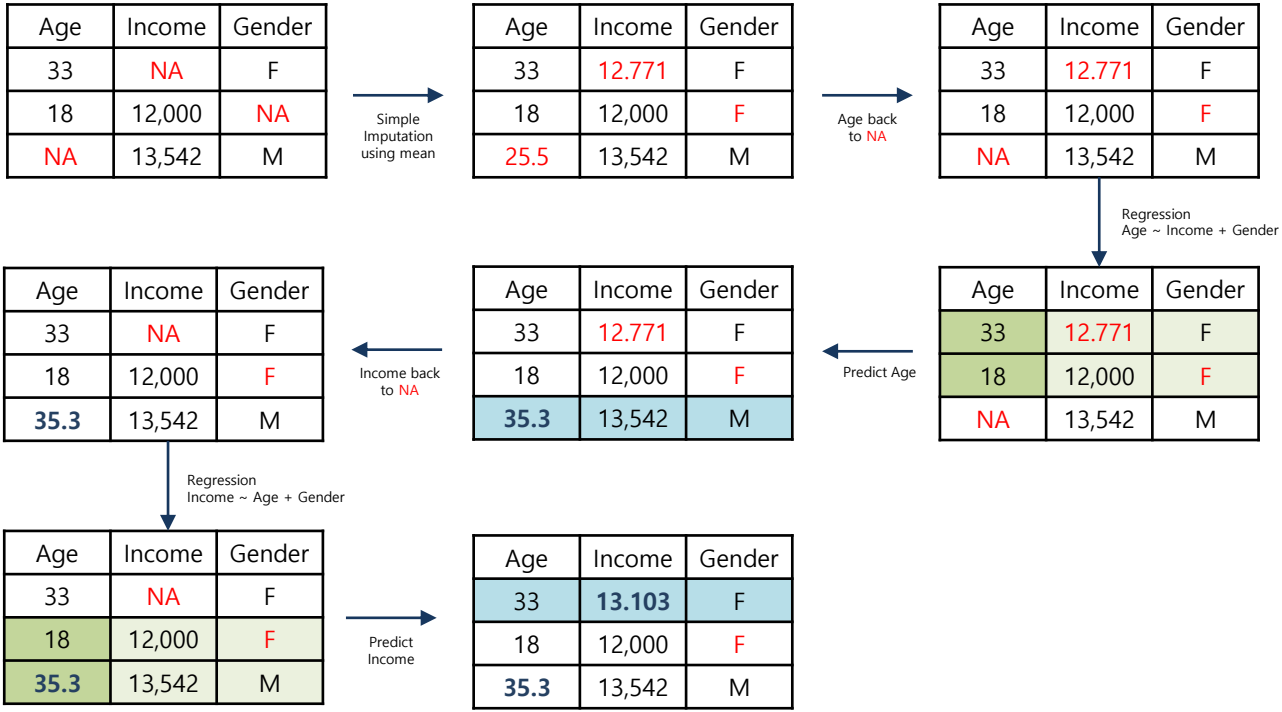
Method for handling missing data

Understanding for MICE – Single Iteration



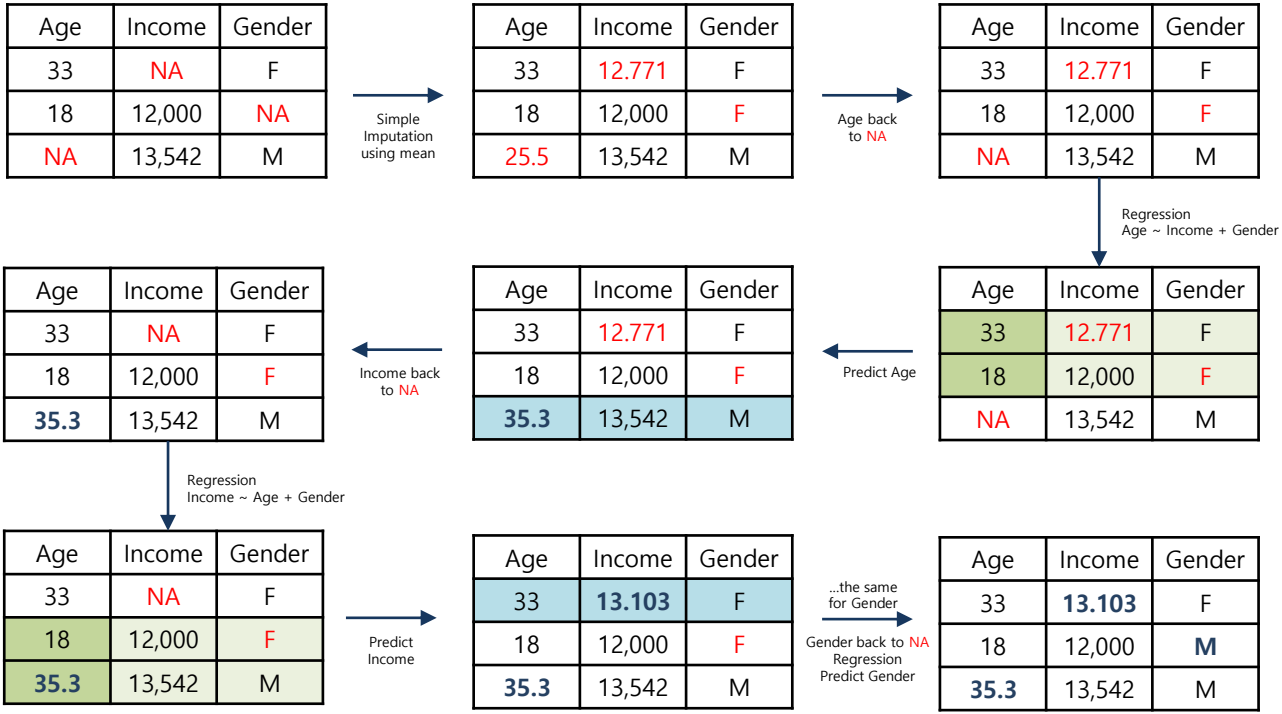
Method for handling missing data

Understanding for MICE – Single Iteration



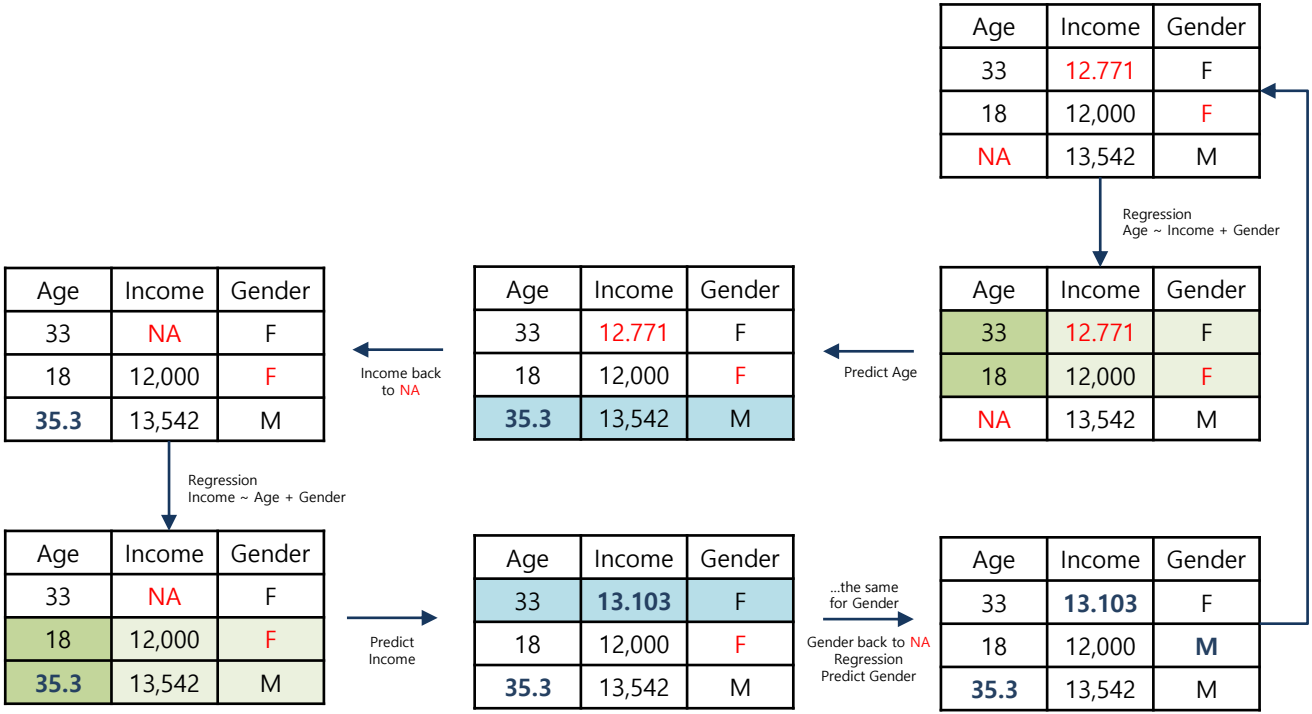
Method for handling missing data

Understanding for MICE – Single Iteration



Method for handling missing data

Understanding for MICE – Single Iteration



Method for handling missing data

- Other Multiple Imputation Techniques
  - Random Forest Imputation
    - Random Forest is a nonparametric replacement method that can be applied to various types of variables that work well on both randomly missing and non-randomly missing data.
    - One caveat is that random forests work best on large datasets, and using random forests on small datasets risks over-consensus.
  - K Nearest Neighbor Imputation
    - k-NN replaces missing attribute values based on the nearest *K* neighbor and neighbors are determined by distance measurements.
    - When K neighbors are determined, the missing value is replaced by taking the mean / medium or mode of the known attribute value of the missing attribute.
  - Expectation-Maximization Imputation
    - EM (Expectation-Maximization) is a type of maximum likelihood method that can be used to create a new data set, and all missing values are replaced with values estimated by the maximum likelihood method
    - The EM algorithm consists of 3 phase
      - 1) Expected phase: Estimated various parameters(e.g. variance, covariance and mean) using list-specific deletions.
      - 2) Imputation phase: Use these estimates to create a regression equation that predicts missing data.
      - 3) Maximizing phase: Uses these equations to fill in the missing data.
    - Then repeat the expected step with the new parameters. The new regression equation is determined to "fill" the missing data. Expectation and maximization steps are repeated until the system stabilizes.

Multiple Imputation Method for handling special types of data

- Likelihood based Multiple Imputation by Event chain for Repairing Event Log

	Case	Activity	Resource	Part Desc.		Start Time	End Time
e_1	$Case_1$	Turning & Milling	Machine 4	Cable Head		2012-01-29 23:24	2012-01-30 05:43
e_2	$Case_1$	Turning & Milling	Machine 4	Cable Head		2012-01-30 05:44	2012-01-30 06:42
e_3	$Case_1$	Turning & Milling	Machine 4	Cable Head		2012-01-30 06:59	2012-01-30 07:21
e_4	$Case_1$	Turning & Milling	Machine 4	Cable Head		2012-01-30 07:21	2012-01-30 10:58
e_5	$Case_1$	Turning & Milling Q.C	Quality Check 1	Cable Head		2012-01-31 13:20	2012-01-31 14:50
e_6	$Case_1$	Laser Marking	Machine 7	Cable Head		2012-02-01 08:18	2012-02-01 08:27
e_7	$Case_1$	Lapping	Machine 1	Cable Head		2012-02-14 00:00	2012-02-14 01:15
e_8	$Case_1$	Lapping	Machine 1	Cable Head		2012-02-14 00:00	2012-02-14 01:15
e_9	$Case_1$	Lapping	Machine 1	Cable Head		2012-02-14 09:05	2012-02-14 10:20
e_{10}	$Case_1$	Lapping	Machine 1	Cable Head	...	2012-02-14 09:05	2012-02-14 09:38
e_{11}	$Case_1$	Round Grinding	Machine 3	Cable Head		2012-02-14 09:13	2012-02-14 13:37
e_{12}	$Case_1$	Round Grinding	Machine 3	Cable Head		2012-02-14 13:37	2012-02-14 15:27
e_{13}	$Case_1$	Final Inspection Q.C.	Quality Check 1	Cable Head		2012-02-16 06:59	2012-02-16 07:59
e_{14}	$Case_1$	Final Inspection Q.C.	Quality Check 1	Cable Head		2012-02-16 12:11	2012-02-16 16:12
e_{15}	$Case_1$	Final Inspection Q.C.	Quality Check 1	Cable Head		2012-02-16 12:43	2012-02-16 13:58
e_{16}	$Case_1$	Packing	Packing	Cable Head		2012-02-17 00:00	2012-02-17 01:00
e_{17}	$Case_2$	Turning & Milling	Machine 9	Spur Gear		2012-01-17 07:01	2012-01-17 11:05
e_{18}	$Case_2$	Turning Q.C.	Quality Check 1	Spur Gear		2012-01-17 11:06	2012-01-17 11:15
e_{19}	$Case_2$	Turning & Milling	Machine 9	Spur Gear		2012-01-17 19:24	2012-01-17 20:01
e_{20}	$Case_2$	Turning & Milling	Machine 9	Spur Gear		2012-01-17 20:01	2012-01-17 23:43
...	...	...	...	...		...	...

Factory Operation Event Log Example

Multiple Imputation Method for handling special types of data

- Likelihood based Multiple Imputation by Event chain for Repairing Event Log

	Mobile Package	Download Speed	Data Limit Usage
y_1	NA	157	80%
y_2	Lite	99	NA
y_3	Fast	167	10%
y_4	Fast	NA	80%

Each included in a case are dependent.

General Data Set with missing

VS

	Case	Activity	Resource	Part Desc.	…	Start Time	End Time
e_1	$Case_1$	NA	Machine 4	Cable Head		2012-01-29 23:24	2012-01-30 05:43
e_2	$Case_1$	Turning & Milling	NA	Cable Head		2012-01-30 05:44	2012-01-30 06:42
e_3	$Case_1$	Turning & Milling	Machine 4	NA	…	2012-01-30 06:59	2012-01-30 07:21
e_4	$Case_1$	Turning & Milling	Machine 4	Cable Head		2012-01-30 07:21	2012-01-30 10:58
e_5	$Case_1$	NA	NA	Cable Head		2012-01-31 13:20	2012-01-31 14:50
…	…	…	…	…		…	…

Observation included in a case are dependent.

Event Log Structure with missing

Multiple Imputation Method for handling special types of data

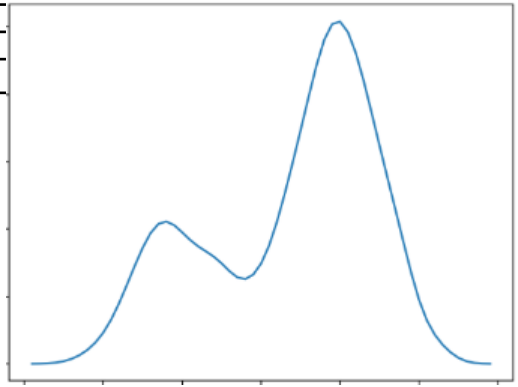
	Case	Activity
e_1	$Case_1$	Turning & Milling
e_2	$Case_1$	Turning & Milling
e_3	$Case_1$	Turning & Milling
e_4	$Case_1$	Turning & Milling
e_5	$Case_1$	Turning & Milling Q.C
e_6	$Case_1$	Laser Marking
e_7	$Case_1$	Lapping
e_8	$Case_1$	Lapping
e_9	$Case_1$	Lapping
e_{10}	$Case_1$	Lapping
e_{11}	$Case_1$	Round Grinding
e_{12}	$Case_1$	Round Grinding
e_{13}	$Case_1$	Final Inspection Q.C.
e_{14}	$Case_1$	Final Inspection Q.C.
e_{15}	$Case_1$	Final Inspection Q.C.
e_{16}	$Case_1$	Packing
...	...	...

Event

	Prior Event	Current Event	Posterior Event
ec_1	Start	Turning & Milling	Turning & Milling
ec_2	Turning & Milling	Turning & Milling	Turning & Milling
ec_3	Turning & Milling	Turning & Milling	Turning & Milling
ec_4	Turning & Milling	Turning & Milling	Turning & Milling Q.C
ec_5	Turning & Milling	Turning & Milling Q.C	Laser Marking
ec_6	Turning & Milling Q.C	Laser Marking	Lapping
ec_7	Laser Marking	Lapping	Lapping
ec_8	Lapping	Lapping	Lapping
ec_9	Lapping	Lapping	Lapping
ec_{10}	Lapping	Lapping	Round Grinding
ec_{11}	Lapping	Round Grinding	Round Grinding
ec_{12}	Round Grinding	Round Grinding	Final Inspection Q.C.
ec_{13}	Round Grinding	Final Inspection Q.C.	Final Inspection Q.C.
ec_{14}	Final Inspection Q.C.	Final Inspection Q.C.	Final Inspection Q.C.
ec_{15}	Final Inspection Q.C.	Final Inspection Q.C.	Packing
ec_{16}	Final Inspection Q.C.	Packing	End
...	...	...	...

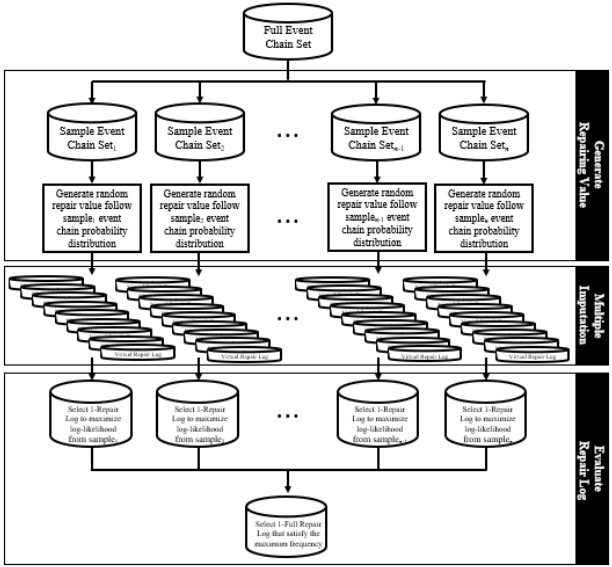
Event Chain

Fitting Target Distribution Using Event Chain



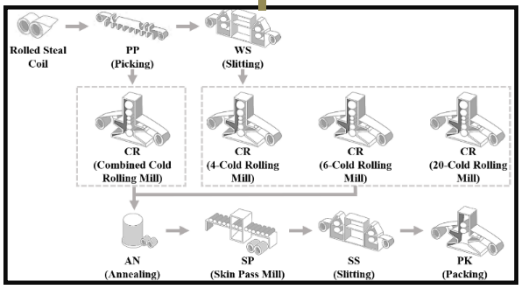
Multiple Imputation Method for handling special types of data

- We developed the concept of an event chain that can reflect sequential information contained in one case for dealing with missing events.
- By converting the events included in the event into an event chain, and replacing the event value that can approximate the distribution of the event chain instead of missing, we developed a restoration method that can restore the restored event log to the original log.



Multiple Imputation Method for handling special types of data

- Likelihood based Multiple Imputation using Event chain for Repairing Event Log
  - Case Study: Korea Steel Company Event Log Data with MIEC(Multiple Imputation by Chained Equations, Expectation-Maximizing Imputation, Random Forest Imputation, K Nearest Neighbor Imputation)



COIL_NO	PRC_CD	PRC_CD1	THK	WDT	WGT	SDT	EDT	PLNPRC_CD	ORD_NO	DRTCOI_NO	Machine
15KM121911111A	PP21	PP	5.99	1132	22900	01/04/2016 16:16:00	01/04/2016 16:38:00	PP21092	KD16090327	15KM1219111	PP2
15KM121911111A	CR21	CR	3.5	1132	22900	01/05/2016 10:10:00	01/05/2016 10:50:00	PP21092	KD16090327	15KM1219111	CR2
15KM121911111A	RC11	RC	3.5	1132	22900	01/05/2016 10:45:00	01/05/2016 11:10:00	PP21092	KD16090327	15KM1219111	RC1
15KM121911111A	WS31	WS	3.5	186	1717	01/07/2016 16:40:00	01/07/2016 17:50:00	PP21092	KD16090327	15KM1219111	WS3
15KM121911111A	PK41	PK	3.5	186	1736	01/07/2016 17:55:00	01/07/2016 17:55:00	PP21092	KD16090327	15KM1219111	PK4
15KM121911111A	PR11	PR	3.5	186	1736	01/07/2016 17:30:00	01/07/2016 17:30:00	PP21092	KD16090327	15KM1219111	PR1
15KM121911112A	PP21	PP	5.99	1132	22900	01/04/2016 16:16:00	01/04/2016 16:38:00	PP21092	KD16090327	15KM1219111	PP2
15KM121911112A	CR21	CR	3.5	1132	22900	01/05/2016 10:10:00	01/05/2016 10:50:00	PP21092	KD16090327	15KM1219111	CR2
15KM121911112A	RC11	RC	3.5	1132	22900	01/05/2016 10:45:00	01/05/2016 11:10:00	PP21092	KD16090327	15KM1219111	RC1
15KM121911112A	WS31	WS	3.5	186	1717	01/07/2016 16:40:00	01/07/2016 17:50:00	PP21092	KD16090327	15KM1219111	WS3
15KM121911112A	PK41	PK	3.5	186	1736	01/07/2016 17:55:00	01/07/2016 17:55:00	PP21092	KD16090327	15KM1219111	PK4
15KM121911112A	PR11	PR	3.5	186	1736	01/07/2016 17:30:00	01/07/2016 17:30:00	PP21092	KD16090327	15KM1219111	PR1
15KM121911113A	PP21	PP	5.99	1132	22900	01/04/2016 16:16:00	01/04/2016 16:38:00	PP21092	KD16090327	15KM1219111	PP2
15KM121911113A	CR21	CR	3.5	1132	22900	01/05/2016 10:10:00	01/05/2016 10:50:00	PP21092	KD16090327	15KM1219111	CR2
15KM121911113A	RC11	RC	3.5	1132	22900	01/05/2016 10:45:00	01/05/2016 11:10:00	PP21092	KD16090327	15KM1219111	RC1
15KM121911113A	WS31	WS	3.5	186	1717	01/07/2016 16:40:00	01/07/2016 17:50:00	PP21092	KD16090327	15KM1219111	WS3
15KM121911113A	PK41	PK	3.5	186	1743	01/07/2016 17:54:00	01/07/2016 17:54:00	PP21092	KD16090327	15KM1219111	PK4
15KM121911113A	PR11	PR	3.5	186	1743	01/07/2016 17:30:00	01/07/2016 17:30:00	PP21092	KD16090327	15KM1219111	PR1
15KM121911113B	PP21	PP	5.99	1132	22900	01/04/2016 16:16:00	01/04/2016 16:38:00	PP21092	KD16080281	15KM1219111	PP2

Missing Rate	MICE	EMBI	RFI	KNNI	MIEC
5%	48.9%	51.3%	60.1%	44.4%	**93.2%**
10%	44.3%	45.5%	59.8%	42.1%	**91.0%**
15%	34.7%	36.9%	50.3%	40.3%	**88.7%**
20%	28.5%	30.1%	47.7%	32.2%	**80.4%**

Performance Event Imputation method using MIEC (Korea Steel Company Event log)