

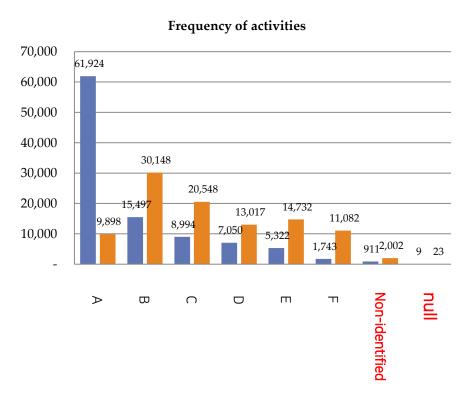


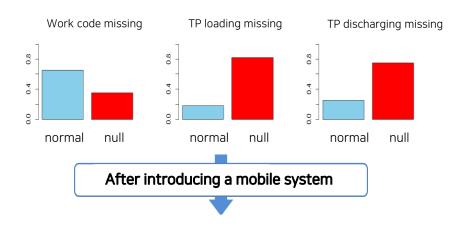
#### **Data quality for process improvement**

## **Overview**



• Is technology the only issue for Industry 4.0?

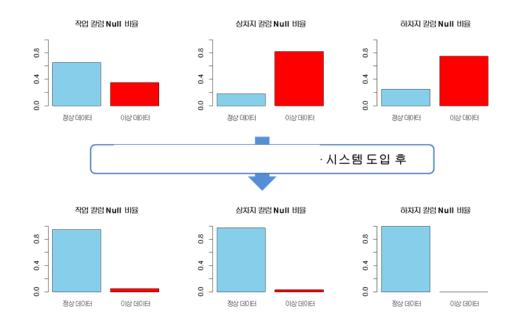








## Garbage in garbage out







## **Data imperfection**

- Missing data
  - 자료누락
- Incorrect data
  - 잘못된코드
- Imprecise data
  - 잘못된 측정 데이터
- Irrelevant data
  - 예측에 사용하기 힘든 자료

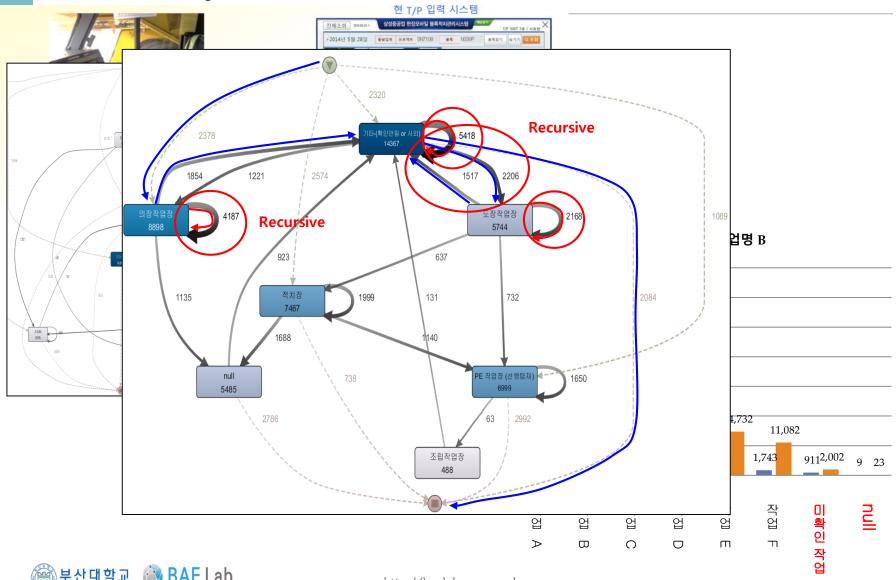
Table 1. Manifestation of quality issues in event log entities [6].

		Event log entities								
		Case	Event	Relationship	Case attrs.	Position	Activity name	Timestamp	Resource	Event attrs.
Event log	Missing data	I1	12	13	14	15	16	17	18	19
quality issues	Incorrect data	I10	l11	112	I13	114	115	I16	I17	l18
	Imprecise data			119	120	I21	122	123	124	125
	Irrelevant data	126	127							





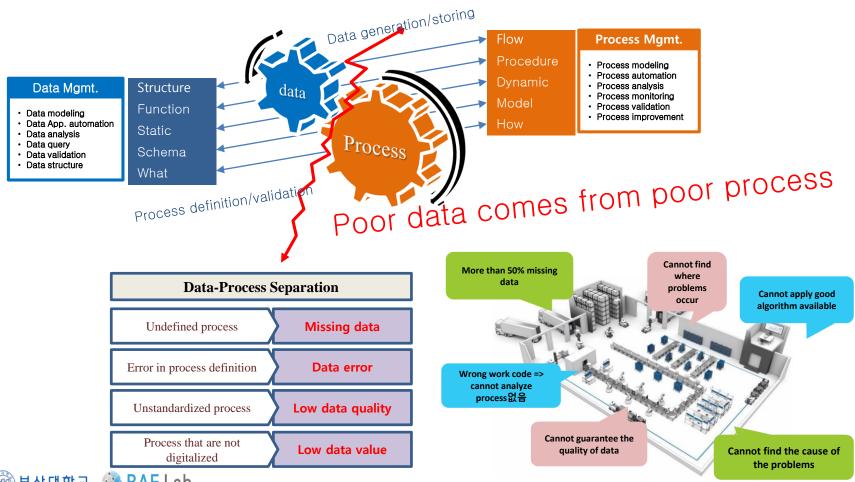
## Data Quality의 중요성







## Big-data vs. process improvement

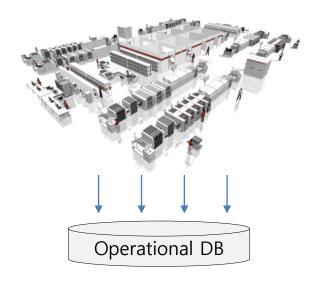


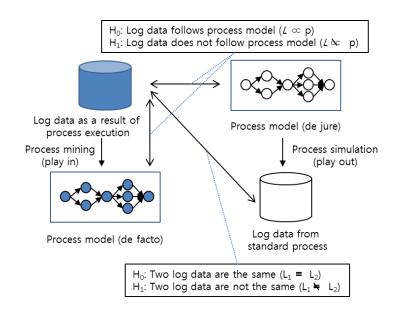




#### **Data-Process compliance**

• 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.











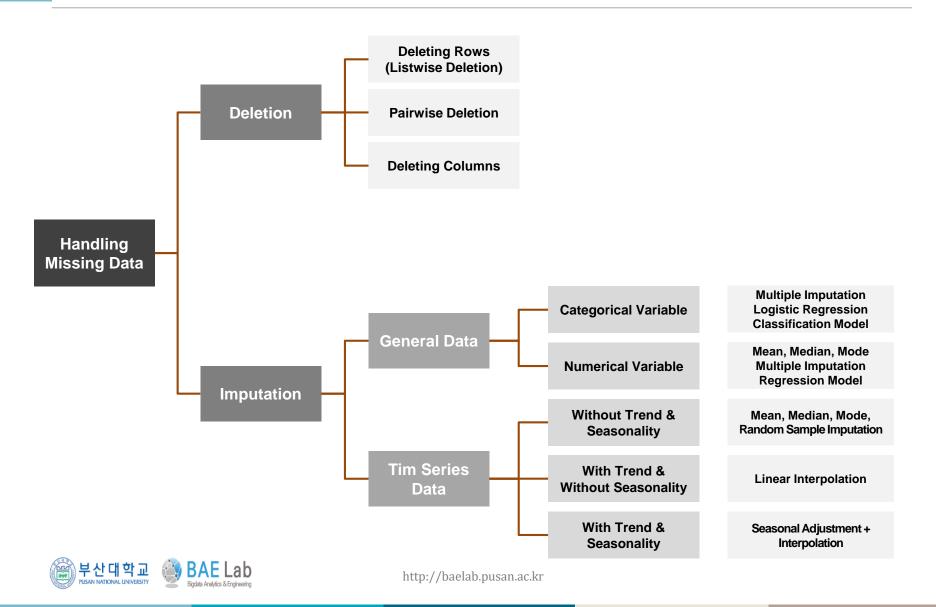
## **Data Imputation**

#### **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)

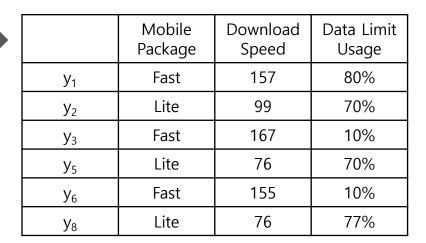






- Listwise deletion (complete-case analysis)
  - By far the most common approach to the missing data is to simply omit those cases with the missing data and analyse the remaining data.
  - This approach is known as the complete case (or available case) analysis or list-wise deletion.

	Mobile Package	Download Speed	Data Limit Usage
y <sub>1</sub>	Fast	157	80%
<b>y</b> <sub>2</sub>	Lite	99	70%
<b>y</b> <sub>3</sub>	Fast	167	10%
<b>y</b> <sub>4</sub>	Fast	NA	80%
<b>y</b> <sub>5</sub>	Lite	76	70%
У <sub>6</sub>	Fast	155	10%
<b>y</b> <sub>7</sub>	NA	NA	95%
y <sub>8</sub>	Lite	76	77%
<b>y</b> <sub>9</sub>	Fast	180	NA







- 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
У <sub>1</sub>	Fast	157	80%
<b>y</b> <sub>2</sub>	Lite	99	70%
<b>y</b> <sub>3</sub>	Fast	167	10%
<b>y</b> <sub>4</sub>	Fast	NA	80%
<b>y</b> <sub>5</sub>	Lite	76	70%
<b>y</b> <sub>6</sub>	Fast	155	10%
<b>y</b> <sub>7</sub>	NA	NA	95%
y <sub>8</sub>	Lite	76	77%
<b>y</b> <sub>9</sub>	Fast	180	NA



	Mobile Package	Download Speed	Data Limit Usage
y <sub>1</sub>	Fast	157	80%
<b>y</b> <sub>2</sub>	Lite	99	70%
<b>y</b> <sub>3</sub>	Fast	167	10%
y <sub>4</sub>	Fast		80%
<b>y</b> <sub>5</sub>	Lite	76	70%
y <sub>6</sub>	Fast	155	10%
<b>y</b> <sub>7</sub>			95%
y <sub>8</sub>	Lite	76	77%
<b>y</b> <sub>9</sub>	Fast	180	



- 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
У <sub>1</sub>	Fast	NA	80%
<b>y</b> <sub>2</sub>	Lite	NA	70%
<b>y</b> <sub>3</sub>	Fast	167	10%
<b>y</b> <sub>4</sub>	Fast	NA	80%
<b>y</b> <sub>5</sub>	Lite	NA	70%
У <sub>6</sub>	Fast	155	10%
<b>y</b> <sub>7</sub>	Fast	NA	95%
y <sub>8</sub>	Lite	NA	77%
<b>y</b> <sub>9</sub>	Fast	180	80%



	Mobile Package	Data Limit Usage
y <sub>1</sub>	Fast	80%
<b>y</b> <sub>2</sub>	Lite	70%
<b>y</b> <sub>3</sub>	Fast	10%
y <sub>4</sub>	Fast	80%
<b>y</b> <sub>5</sub>	Lite	70%
y <sub>6</sub>	Fast	10%
<b>y</b> <sub>7</sub>	Fast	95%
y <sub>8</sub>	Lite	77%
<b>y</b> <sub>9</sub>	Fast	80%



- 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 = 76

	Mobile Package	Download Speed	Data Limit Usage
<b>У</b> 1	Fast	157	80%
<b>y</b> <sub>2</sub>	Lite	99	70%
<b>y</b> <sub>3</sub>	Fast	167	10%
y <sub>4</sub>	Fast	NA	80%
<b>y</b> <sub>5</sub>	Lite	76	70%
<b>y</b> <sub>6</sub>	Fast	155	10%
<b>y</b> <sub>7</sub>	Fast	NA	95%
<b>y</b> <sub>8</sub>	Lite	76	77%
<b>y</b> <sub>9</sub>	Fast	180	80%

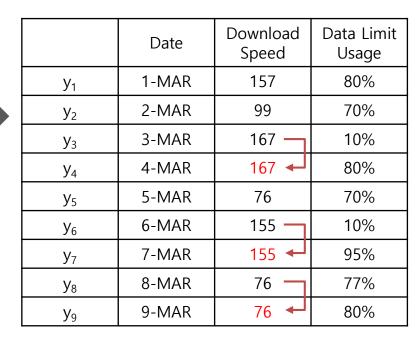
		SI with Mean	SI with Median	SI with Mode
	У <sub>1</sub>	157	157	157
•	<b>y</b> <sub>2</sub>	99	99	99
	<b>y</b> <sub>3</sub>	167	167	167
	<b>y</b> <sub>4</sub>	130	155	76
	<b>y</b> <sub>5</sub>	76	76	76
	<b>y</b> <sub>6</sub>	155	155	155
	<b>y</b> <sub>7</sub>	130	155	76
	<b>y</b> <sub>8</sub>	76	76	76
	<b>y</b> <sub>9</sub>	180	180	180





- 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
У <sub>1</sub>	1-MAR	157	80%
<b>y</b> <sub>2</sub>	2-MAR	99	70%
<b>y</b> <sub>3</sub>	3-MAR	167	10%
<b>y</b> <sub>4</sub>	4-MAR	NA	80%
<b>y</b> <sub>5</sub>	5-MAR	76	70%
<b>y</b> <sub>6</sub>	6-MAR	155	10%
<b>y</b> <sub>7</sub>	7-MAR	NA	95%
<b>y</b> <sub>8</sub>	8-MAR	76	77%
<b>y</b> <sub>9</sub>	9-MAR	NA	80%

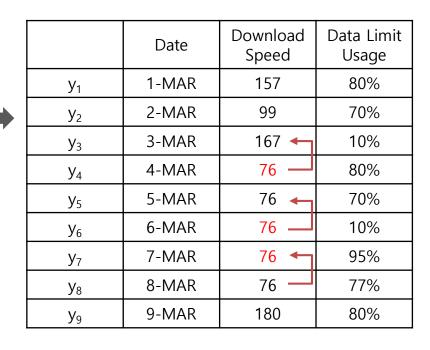






- Time Series Specific Method (LOCF, NOCB and Linear Interpolation)
  - Next Observation Carried Backward(NOCB)
    - A similar approach like LOCF which works in the opposite direction by taking the first observation after the missing value and **c**arrying it backward

	Date	Download Speed	Data Limit Usage
У <sub>1</sub>	1-MAR	157	80%
<b>y</b> <sub>2</sub>	2-MAR	99	70%
<b>y</b> <sub>3</sub>	3-MAR	167	10%
У <sub>4</sub>	4-MAR	NA	80%
<b>y</b> <sub>5</sub>	5-MAR	76	70%
У <sub>6</sub>	6-MAR	NA	10%
<b>y</b> <sub>7</sub>	7-MAR	NA	95%
<b>y</b> <sub>8</sub>	8-MAR	76	77%
<b>y</b> <sub>9</sub>	9-MAR	180	80%







- 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
У <sub>1</sub>	1-MAR	157	80%
<b>y</b> <sub>2</sub>	2-MAR	99	70%
<b>y</b> <sub>3</sub>	3-MAR	167	10%
<b>y</b> <sub>4</sub>	4-MAR	NA	80%
<b>y</b> <sub>5</sub>	5-MAR	76	70%
<b>y</b> <sub>6</sub>	6-MAR	NA	10%
<b>y</b> <sub>7</sub>	7-MAR	150	95%
<b>y</b> <sub>8</sub>	8-MAR	76	77%
<b>y</b> <sub>9</sub>	9-MAR	180	80%

		Date	Download Speed		D	Data Limit Usage	
	y <sub>1</sub>	1-MAR		157			80%
	<b>y</b> <sub>2</sub>	2-MAR		99			70%
ĺ	<b>y</b> <sub>3</sub>	3-MAR		167			10%
	<b>y</b> <sub>4</sub>	4-MAR	121.5			80% 70%	
	<b>y</b> <sub>5</sub>	5-MAR		76			
	<b>y</b> <sub>6</sub>	6-MAR		113			10%
	<b>y</b> <sub>7</sub>	7-MAR		150			95%
	y <sub>8</sub>	8-MAR		76			77%
	<b>y</b> <sub>9</sub>	9-MAR		180			80%
•					_		

(167+76)/2 = 121.5

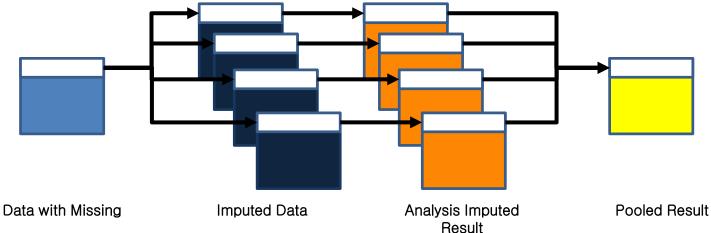
(76+150)/2 = 113





#### 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.





Understanding for MICE – Single Iteration

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



Understanding for MICE – Single Iteration

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

Simple Imputation using mean

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





Understanding for MICE – Single Iteration.

-		$\omega$ $\omega$ $\omega$
Age	Income	Gender
33	NA	F
18	12,000	NA
NA	13,542	М

Simple Imputation using mean

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

Age back to

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



Understanding for MICE – Single Iteration

т.		1314111111	$U \cup U \cup U \cup U$
	Age	Income	Gender
	33	NA	F
	18	12,000	NA
	NA	13,542	М

Simple Imputation using mean

<u>lieralion</u>		
Age	Income	Gender
33	12.771	F
18	12,000	F
25.5	13,542	М

Age back to NA

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

Regression Age  $\sim$  Income + Gender

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

Predict Age





Understanding for MICE – Single Iteration

 UIIUE	1314111111	$U \cup U \cup U \cup U$
Age	Income	Gender
33	NA	F
18	12,000	NA
NA	13,542	М

Simple Imputation using mean

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



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

Regression Age ~ Income + Gender

Age	Income	Gender
33	12.771	F
18	12,000	F
35.3	13,542	М



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





Understanding for MICE – Single Iteration.

٠.		1314111111	
	Age	Income	Gender
	33	NA	F
	18	12,000	NA
	NA	13,542	М

Simple Imputation using mean

<u>lieralion</u>		
Age	Income	Gender
33	12.771	F
18	12,000	F
25.5	13,542	М

Age back to NA

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

Regression Age ~ Income + Gender

Age	Income	Gender
33	NA	F
18	12,000	F
35.3	13,542	М

Income back to NA

Age	Income	Gender
33	12.771	F
18	12,000	F
35.3	13,542	М

Predict Age

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



Understanding for MICE - Single Iteration Gender Age Income 33 NA F 18 12,000 NA 13,542 NA М

Simple Imputation using mean

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

Age back to

Predict Age

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

Regression Age ~ Income + Gender

Age	Income	Gender
33	NA	F
18	12,000	F
35.3	13,542	М

Income back to NA

Age	Income	Gender
33	12.771	F
18	12,000	F
35.3	13,542	М

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

Regression	
Income ~ Age	+ Gender

Age	Income	Gender
33	NA	F
18	12,000	F
35.3	13,542	М





•	<ul> <li>Understanding for MIC</li> </ul>		
	Age	Income	Gender
	33	NA	F
	18	12,000	NA
	NA	13,542	М

CE – Single Iteration			
	Age	Income	Gender
	33	12.771	F
Simple	18	12,000	F
Imputation using mean	25.5	13,542	М

Age back to

l	Age	Income	Gender
	33	12.771	F
	18	12,000	F
	NA	13,542	М
-			

Regression Age ~ Income + Gender

Age	Income	Gender
33	NA	F
18	12,000	F
35.3	13,542	М

Income back to NA

Age	Income	Gender
33	12.771	F
18	12,000	F
35.3	13,542	М

	3
Predict Age	

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

Regression Income ~ Age + Gender

Age	Income	Gender
33	NA	F
18	12,000	F
35.3	13,542	М

Predict Income

Age	Income	Gender	
33	13.103	F	
18	12,000	F	
35.3	13,542	М	





Understanding for MICE - Single Iteration Gender Age Income 33 F NA 18 12,000 NA 13,542 NA Μ

Simple Imputation using mean

)	iteration		
	Age	Income	Gender
	33	12.771	F
	18	12,000	F
	25.5	13,542	М

Age back to

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

Regression Age ~ Income + Gender

Age	Income	Gender
33	NA	F
18	12,000	F
35.3	13,542	М

Income back to NA

Age	Income	Gender
33	12.771	F
18	12,000	F
35.3	13,542	М

Predict Age

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

Regression Income ~ Age + Gender

Age	Income	Gender
33	NA	F
18	12,000	F
35.3	13,542	М

Predict Income

Age	Income	Gender
33	13.103	F
18	12,000	F
35.3	13,542	М

...the same for Gender Gender back to NA Regression Predict Gender

Age	Income	Gender
33	13.103	F
18	12,000	M
35.3	13,542	М





• Understanding for MICE – Multiple Iterations

Age	Income		Gender	
33	12.771		F	
18	12,000		F	
NA	13,	542	М	
		_	ession ~ Income + Gen	<b>d</b> er

Age	Income	Gender
33	NA	F
18	12,000	F
35.3	13,542	М

Income back to NA

Age	Income	Gender
33	12.771	F
18	12,000	F
35.3	13,542	М

Predict Age

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

Age	Income	Gender
33	NA	F
18	12,000	F
35.3	13,542	М

Predict Income

Age	Income	Gender
33	13.103	F
18	12,000	F
35.3	13,542	М

	Age	Income	Gender	
	33	13.103	F	
1A	18	12,000	М	
	35.3	13,542	М	





Regression

Income ~ Age + Gender

#### 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 overconsensus.
- 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.	
$e_1$	Case <sub>1</sub>	Turning & Milling	Machine 4	Cable Head	
$e_2$	Case <sub>1</sub>	Turning & Milling	Machine 4	Cable Head	
$e_3$	Case <sub>1</sub>	Turning & Milling	Machine 4	Cable Head	
$e_4$	Case <sub>1</sub>	Turning & Milling	Machine 4	Cable Head	
$e_5$	Case <sub>1</sub>	Turning & Milling Q.C	Quality Check 1	Cable Head	
$e_6$	Case <sub>1</sub>	Laser Marking	Machine 7	Cable Head	
$e_7$	Case <sub>1</sub>	Lapping	Machine 1	Cable Head	
$e_8$	Case <sub>1</sub>	Lapping	Machine 1	Cable Head	
$e_{g}$	Case <sub>1</sub>	Lapping	Machine 1	Cable Head	
$e_{10}$	Case <sub>1</sub>	Lapping	Machine 1	Cable Head	
e <sub>11</sub>	Case <sub>1</sub>	Round Grinding	Machine 3	Cable Head	
$e_{12}$	Case <sub>1</sub>	Round Grinding	Machine 3	Cable Head	
e <sub>13</sub>	Case <sub>1</sub>	Final Inspection Q.C.	Quality Check 1	Cable Head	
e <sub>14</sub>	Case <sub>1</sub>	Final Inspection Q.C.	Quality Check 1	Cable Head	
e <sub>15</sub>	Case <sub>1</sub>	Final Inspection Q.C.	Quality Check 1	Cable Head	
$e_{16}$	Case <sub>1</sub>	Packing	Packing	Cable Head	
e <sub>17</sub>	Case <sub>2</sub>	Turning & Milling	Machine 9	Spur Gear	
e <sub>18</sub>	Case <sub>2</sub>	Turning Q.C.	Quality Check 1	Spur Gear	
$e_{19}$	Case <sub>2</sub>	Turning & Milling	Machine 9	Spur Gear	
$e_{20}$	Case <sub>2</sub>	Turning & Milling	Machine 9	Spur Gear	

Start Time	End Time
2012-01-29 23:24	2012-01-30 05:43
2012-01-30 05:44	2012-01-30 06:42
2012-01-30 06:59	2012-01-30 07:21
2012-01-30 07:21	2012-01-30 10:58
2012-01-31 13:20	2012-01-31 14:50
2012-02-01 08:18	2012-02-01 08:27
2012-02-14 00:00	2012-02-14 01:15
2012-02-14 00:00	2012-02-14 01:15
2012-02-14 09:05	2012-02-14 10:20
2012-02-14 09:05	2012-02-14 09:38
2012-02-14 09:13	2012-02-14 13:37
2012-02-14 13:37	2012-02-14 15:27
2012-02-16 06:59	2012-02-16 07:59
2012-02-16 12:11	2012-02-16 16:12
2012-02-16 12:43	2012-02-16 13:58
2012-02-17 00:00	2012-02-17 01:00
2012-01-17 07:01	2012-01-17 11:05
2012-01-17 11:06	2012-01-17 11:15
2012-01-17 19:24	2012-01-17 20:01
2012-01-17 20:01	2012-01-17 23:43







# Multiple Imputation Method for handling special types of data

Likelihood based Multiple Imputation by Event chain for Repairing Event Log

data included in a case (row) are dependent.

	Mobile Package	Download Speed	Data Limit Usage
y <sub>1</sub>	NA	157	80%
<b>y</b> <sub>2</sub>	Lite	99	NA
<b>y</b> <sub>3</sub>	Fast	167	10%
<b>y</b> <sub>4</sub>	Fast	NA	80%

General Data Set with missing

#### VS

	Case Activity Resource			Part Desc.		
$e_1$	Case <sub>1</sub>	NA	Machine 4	Cable Head		
$e_2$	Case <sub>1</sub> Turning & Milling		NA	Cable Head		
$e_3$	Case <sub>1</sub>	Turning & Milling	Machine 4	NA		
$e_4$	Case <sub>1</sub>	Turning & Milling	Machine 4	Cable Head		
$e_5$	Case <sub>1</sub>	NA	NA	Cable Head		
•••		•••				

Start Time	End Time	
2012-01-29 23:24	2012-01-30 05:43	
2012-01-30 05:44	2012-01-30 06:42	
2012-01-30 06:59	2012-01-30 07:21	
2012-01-30 07:21	2012-01-30 10:58	
2012-01-31 13:20	2012-01-31 14:50	

Observations included in a case are dependent.

Event Log Structure with missing

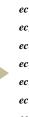




## Multiple Imputation Method for handling special types of data

Likelihood based Multiple Imputation by Event chain for Repairing Event Log

	Case	Activity		
$e_1$	Case <sub>1</sub>	Turning & Milling		
$e_2$	Case <sub>1</sub>	Turning & Milling		
$e_3$	Case <sub>1</sub>	Turning & Milling		
$e_4$	Case <sub>1</sub>	Turning & Milling		
$e_5$	Case <sub>1</sub>	Turning & Milling Q.C		
$e_6$	Case <sub>1</sub>	Laser Marking		
e <sub>7</sub>	Case <sub>1</sub>	Lapping		
$e_8$	Case <sub>1</sub>	Lapping		
$e_g$	Case <sub>1</sub>	Lapping		
$e_{10}$	Case <sub>1</sub>	Lapping		
e <sub>11</sub>	Case <sub>1</sub>	Round Grinding		
$e_{12}$	Case <sub>1</sub>	Round Grinding		
e <sub>13</sub>	Case <sub>1</sub>	Final Inspection Q.C.		
e <sub>14</sub>	Case <sub>1</sub>	Final Inspection Q.C.		
e <sub>15</sub>	Case <sub>1</sub>	Final Inspection Q.C.		
e <sub>16</sub>	Case <sub>1</sub>	Packing		



**Prior Event Current Event** Posterior Event Start Turning & Milling Turning & Milling  $ec_1$ Turning & Milling Turning & Milling Turning & Milling  $ec_2$ Turning & Milling Turning & Milling Turning & Milling  $ec_3$ Turning & Milling Turning & Milling Turning & Milling Q.C  $ec_{4}$  $ec_5$ Turning & Milling Turning & Milling Q.C Laser Marking Turning & Milling Q.C  $ec_6$ Laser Marking Lapping Laser Marking Lapping Lapping  $ec_8$ Lapping Lapping Lapping Lapping Lapping Lapping Round Grinding  $ec_{10}$ Lapping Lapping Round Grinding Round Grinding  $ec_{11}$ Lapping Final Inspection Q.C. Round Grinding Round Grinding  $ec_{12}$  $ec_{13}$ Round Grinding Final Inspection Q.C.  $ec_{14}$ Final Inspection Q.C. Final Inspection Q.C. Packing  $ec_{15}$ Final Inspection Q.C. Packing End  $ec_{16}$ 

**Fitting Target Distribution Using Event Chain** 

**Event Chain** 

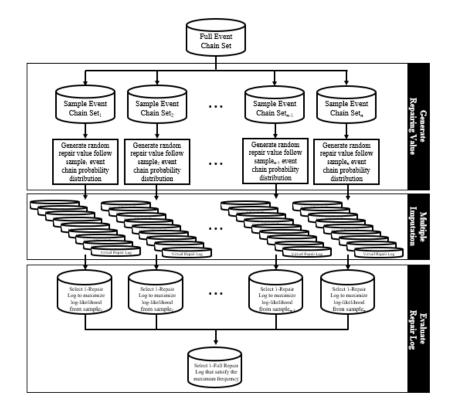
**Event** 





## 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.



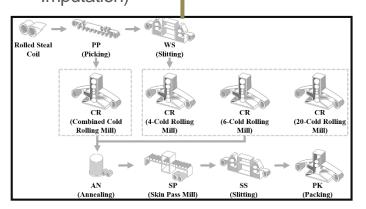




## Multiple Imputation Method for handling special types of data

Likelihood based Multiple Imputation using Event chain for Repairing Event Log

 Case Study: Korea Steel Company Event Log Data with MICE(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
15KM12191111A	PP21	PP	5.99	1132	22900	01/04/2016 16:16:00	01/04/2016 16:38:00	PP21092	KD16090327	15KM1219111	PP2
15KM12191111A	CR21	CR	3.5	1132	22900	01/05/2016 10:10:00	01/05/2016 10:50:00	PP21092	KD16090327	15KM1219111	CR2
15KM12191111A	RC11	RC	3.5	1132	22900	01/05/2016 10:45:00	01/05/2016 11:10:00	PP21092	KD16090327	15KM1219111	RC1
15KM12191111A	WS31	WS	3.5	186	1717	01/07/2016 16:40:00	01/07/2016 17:50:00	PP21092	KD16090327	15KM1219111	WS3
15KM12191111A	PK41	PK	3.5	186	1736	01/07/2016 17:55:00	01/07/2016 17:55:00	PP21092	KD16090327	15KM1219111	PK4
15KM12191111A	PR11	PR	3.5	186	1736	01/07/2016 17:30:00	01/07/2016 17:30:00	PP21092	KD16090327	15KM1219111	PR1
15KM12191112A	PP21	PP	5.99	1132	22900	01/04/2016 16:16:00	01/04/2016 16:38:00	PP21092	KD16090327	15KM1219111	PP2
15KM12191112A	CR21	CR	3.5	1132	22900	01/05/2016 10:10:00	01/05/2016 10:50:00	PP21092	KD16090327	15KM1219111	CR2
15KM12191112A	RC11	RC	3.5	1132	22900	01/05/2016 10:45:00	01/05/2016 11:10:00	PP21092	KD16090327	15KM1219111	RC1
15KM12191112A	WS31	WS	3.5	186	1717	01/07/2016 16:40:00	01/07/2016 17:50:00	PP21092	KD16090327	15KM1219111	WS3
15KM12191112A	PK41	PK	3.5	186	1736	01/07/2016 17:55:00	01/07/2016 17:55:00	PP21092	KD16090327	15KM1219111	PK4
15KM12191112A	PR11	PR	3.5	186	1736	01/07/2016 17:30:00	01/07/2016 17:30:00	PP21092	KD16090327	15KM1219111	PR1
15KM12191113A	PP21	PP	5.99	1132	22900	01/04/2016 16:16:00	01/04/2016 16:38:00	PP21092	KD16090327	15KM1219111	PP2
15KM12191113A	CR21	CR	3.5	1132	22900	01/05/2016 10:10:00	01/05/2016 10:50:00	PP21092	KD16090327	15KM1219111	CR2
15KM12191113A	RC11	RC	3.5	1132	22900	01/05/2016 10:45:00	01/05/2016 11:10:00	PP21092	KD16090327	15KM1219111	RC1
15KM12191113A	WS31	WS	3.5	186	1717	01/07/2016 16:40:00	01/07/2016 17:50:00	PP21092	KD16090327	15KM1219111	WS3
15KM12191113A	PK41	PK	3.5	186	1743	01/07/2016 17:54:00	01/07/2016 17:54:00	PP21092	KD16090327	15KM1219111	PK4
15KM12191113A	PR11	PR	3.5	186	1743	01/07/2016 17:30:00	01/07/2016 17:30:00	PP21092	KD16090327	15KM1219111	PR1
15KM12191113B	PP21	PP	5.99	1132	22900	01/04/2016 16:16:00	01/04/2016 16:38:00	PP21092	KD16080281	15KM1219111	PP2

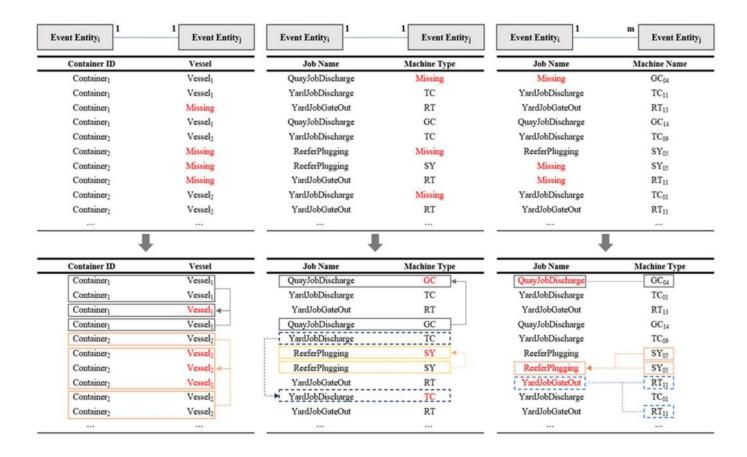
Missing Rate	MICE	ЕМВІ	RFI	KNNI	MIEC
5%	48.9%	51.3%	60.1%	44.4%	93.2%
10%	44.3%	45.5%	59.8%	42.1%	91.0%
15%	34.7%	36.9%	50.3%	40.3%	88.7%
20%	28.5%	30.1%	47.7%	32.2%	80.4%

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





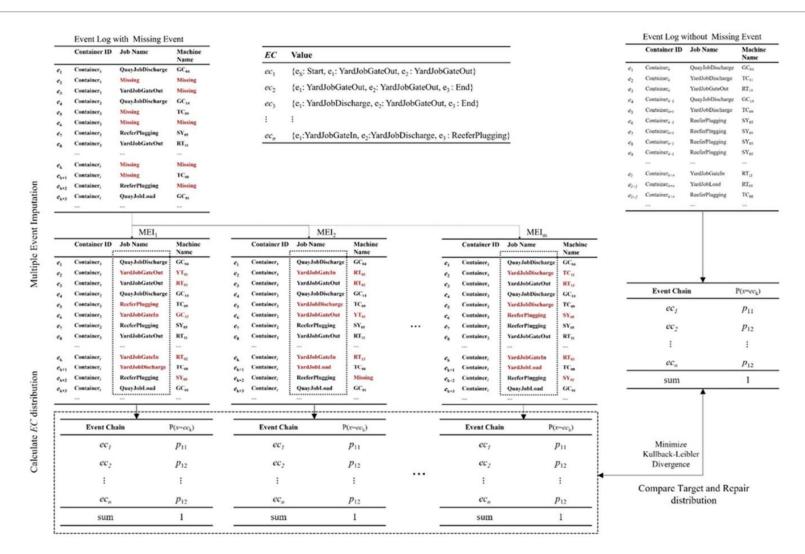
### Single imputation vs multiple imputation







#### Multiple impuptation







#### **Recurrent method**

Container Id	Job Name	Machine Type	Machine Name	Vessel Name	Working Location	Cardina	dity	SER		М	ER								
Container <sub>1</sub> Container <sub>1</sub> Container <sub>1</sub>	QuayJobDischarge YardJobDischarge Missing	Missing Missing Missing	GC <sub>04</sub> Missing TC <sub>01</sub>	Vessel <sub>01</sub> Missing Missing	Missing 11-Block Missing	one-to-	one	Vessel Na	me	No	o exist				_				
Container <sub>2</sub> Container <sub>2</sub> Container <sub>2</sub> Container <sub>3</sub>	YardJobGateOut Missing Missing ReeferPlugging	Missing GC YT Missing	RT <sub>13</sub> GC <sub>02</sub> Missing SY <sub>05</sub>	Missing Missing Vesselot Vesselot	Missing 1-Berth 14-Block 27-Block	one-to-many		Job Name Machine Type Working Location			Machine Type (one) and Job Name (many) Machine Type (one) and Machine Name (many)								
Container <sub>2</sub> Container <sub>3</sub> Container <sub>3</sub>	Missing Missing QuayJobDischarge Missing	SY RT GC Missing	Missing Missing GC <sub>01</sub> Missing	Missing Vesselot Vesselot Missing	27-Block Missing 1-Berth Missing			No exist		Jo	Job Name and Machine Name Job Name and Working Location Machine Name and Working Location								
Container <sub>4</sub> Container <sub>4</sub>	YardJobCoateOut YardJobLoad QuayJobLoad	Missing TC GC	RT <sub>11</sub> TC <sub>12</sub> GC <sub>02</sub>	Missing Vesselor Missing	Missing 12-Block 1-Berth			E	vent car	dinality	and rela	ition							
		SEI by	vertical	<b></b>										If multi	ple iteration procedu	re,			
Container Id	Job Name	Machine Type	Machine Name	Vessel Name	Working Location									it return	as to the SEI stage.	- !			
Container <sub>1</sub> Container <sub>1</sub>	QuayJobDischarge YardJobDischarge	Missing Missing	Missing	Vessel <sub>01</sub> Vessel <sub>01</sub>	Missing 11-Block		Container Id	Job Name	Machine Type	Machine Name	Vessel Name	Working Location				i			
Container <sub>1</sub> Container <sub>1</sub>	Missing YardJobGateOut	Missing Missing	TC <sub>01</sub> RT <sub>13</sub>	Vessel <sub>01</sub> Vessel <sub>01</sub>	Missing		Container;	QuaylobDucharge	oc .	GC <sub>(a</sub>	Vesselg	1-Bech				i			
Container <sub>2</sub>	Missing	GC	GC <sub>02</sub>	Vesselo	1-Berth		Container <sub>1</sub>	YardJohDischarge	TC	TCm	Vessel <sub>et</sub>	11-Block				1			
Container <sub>2</sub>	Missing	YT	Missing	Vesselat	14-Block		Container;	Shuffling	TC	TCm	Vessel <sub>G</sub>	11-Block							
Container <sub>2</sub>	ReeferPlugging	Missing	SY <sub>05</sub>	Vessel <sub>01</sub>	27-Block		Container:	YardJobGateOut QuarJobDischarge	RT GC	RT <sub>IJ</sub> GC <sub>ID</sub>	Vessel <sub>ci</sub> Vessel <sub>ci</sub>	Gate 1-Berth							
Container <sub>2</sub>	Missing	SY	Missing	Vessel <sub>01</sub>	27-Block		Container;	YardiobDischarge	YT	TCia	Vessela	14-Block							
Container <sub>2</sub>	Missing QuayJobDischarge	RT GC	Missing	Vesselos Vesselos	Missing 1-Berth	$\rightarrow$	Container <sub>2</sub>	ReeferPlugging	SY	SYm	Vessel	27-Block +				- 1			
Container <sub>3</sub> Container <sub>3</sub>	Missing	Missing	GC01 Missing	Vesselos Vesselos	Missing		Container <sub>2</sub>	RevierPlugging	SY	SYm	Vessel <sub>it</sub>	27-Block				- 1			
Container <sub>1</sub>	YardJobGateOut	Missing	RT <sub>11</sub>	Vessel <sub>01</sub>	Missing	- 1	Container <sub>2</sub>	YardlobGateOut	RT	RTI	Vessel	Gate				i			
Container,	YardJobLoad	TC	TC <sub>12</sub>	Vesselor	12-Block	- 1	Container,	QuaylobDocharge YardlobDocharge	Moung	GC <sub>tt</sub>	Vessel <sub>ct</sub> Vessel <sub>ct</sub>	1-Berth 14-Block				i			
Container <sub>4</sub>	QuayJobLoad	GC	$GC_{02}$	Vessel <sub>01</sub>	1-Berth	- 1	Container,	YardJobGateOut	RT	RTin	Vessel	Ciate				i			
***	***		***	***		- 1	Container,	YardJobLoad	TC	TC <sub>12</sub>	Vessela	12-Block				i			
		$\neg$				- 1	Container <sub>a</sub>	Quay fobLoad	oc	GC <sub>III</sub>	Vesseli	1-Berth				1			
		SEI by	horizontal				in .	_			-					1			
		+				<b>→</b>			:				-						
Container Id	Job Name	Machine Type	Machine Name	Vessel Name	Working Location	-	Container Id	Job Name	Machine	Machine	Vessel	Working		Container Id	Job Name	Machine Type	Machine Name	Vessel Name	Working Location
Container <sub>1</sub>	QuayJobDischarge	GC	GC <sub>04</sub>	Vessel <sub>01</sub>	Missing				Type	Name	Name	Location		Container <sub>1</sub>	QuayJobDischarge	GC	$GC_{04}$	Vessel <sub>01</sub>	1-Berth
Container <sub>1</sub>	YardJobDischarge	TC	Missing	Vessel <sub>01</sub>	11-Block		Container,	QuayJobDischarge YardJobDischarge	GC TC	GC <sub>te</sub>	Vessel <sub>et</sub>	1-Berth 11-Block		Container <sub>1</sub>	YardJobDischarge	TC	TC <sub>01</sub>	Vessel <sub>01</sub>	11-Block
Container <sub>1</sub>	Missing YardJobGateOut	SY RT	SY <sub>01</sub>	Vessel <sub>01</sub>	Missing Missing		Container,	Shaffling	TC	TCm	Vessela	13-Block		Container <sub>1</sub>	Shuffling YardJobGateOut	TC RT	TC <sub>01</sub>	Vessel <sub>01</sub>	11-Block Gate
Container <sub>1</sub> Container <sub>2</sub>	Missing	GC	RT <sub>13</sub> GC <sub>02</sub>	Vessel <sub>01</sub> Vessel <sub>01</sub>	1-Berth		Container,	YardfobGserOut	RT	RT <sub>10</sub>	Vessels	Gare		Container <sub>1</sub> Container <sub>2</sub>	QuayJobDischarge	GC	RT <sub>13</sub> GC <sub>02</sub>	Vesselor Vesselor	1-Berth
Container:	Missing	YT	Missing	Vessel <sub>01</sub> Vessel <sub>01</sub>	14-Block		Container <sub>2</sub>	QuayJobDischarge	GC	GC <sub>62</sub>	Vessela	1-Berth		Containers	YardJobDischarge	YT	TC <sub>14</sub>	Vessel <sub>01</sub> Vessel <sub>01</sub>	14-Block
Container,	ReeferPlugging	SY	SYos	Vessel	27-Block —		Container	YardJohDucharge	YT	GCya	Vessels	14-Block	$\rightarrow$	Container	ReeferPlugging	SY	SYos	Vesselat	27-Block
Container <sub>2</sub>	Missing	SY	Missing	Vessel <sub>01</sub>	27-Block		Container <sub>2</sub>	ReefecPlugging YardJobDischurge	SY	SYm	Vessel <sub>G</sub>	27-Block 27-Block	Evaluate	Container:	ReeferPlugging	SY	SYos	Vessel <sub>01</sub>	27-Block
Container <sub>2</sub>	Missing	RT	Missing	Vessel <sub>01</sub>	Missing		Container <sub>2</sub> Container <sub>2</sub>	YardiokGurdurge YardiokGurdur	SY RT	SY <sub>10</sub> ST <sub>10</sub>	Vessel <sub>et</sub> Vessel <sub>et</sub>	27-Block Cate		Container <sub>2</sub>	YardJobGateOut	RT	RT <sub>11</sub>	Vessel <sub>01</sub>	Gate
Container <sub>3</sub>	QuayJobDischarge	GC	GC <sub>01</sub>	Vesselot	1-Berth		Container:	Quay Job Discharge	GC C	GC <sub>11</sub>	Vessela	1-Berth		Container <sub>3</sub>	QuayJobDischarge	GC	$GC_{01}$	Vessel <sub>01</sub>	1-Berth
Container <sub>3</sub>	Missing	Missing	Missing	Vessel <sub>01</sub>	Missing		Container	YardlobGonrOut	Moning	RTon	Venela	14-Block		Container <sub>3</sub>	YardJobDischarge	Missing	TC01	Vessel <sub>01</sub>	14-Block
		RT	RT <sub>11</sub>	Vessela	Missing		Container,	YardJobGateOut	RT	RTH	Venela	Cute		Container <sub>3</sub>	YardJobGateOut	RT	RT <sub>11</sub>	Vessel <sub>01</sub>	Gate
Container <sub>3</sub>	YardJobGateOut											1 5 1 7 1 7 1 7 1 7 1 7		All the second second	**	7800		**	4.0 ml
	YardJobLoad QuayJobLoad	TC GC	TC <sub>12</sub> GC <sub>62</sub>	Vesselot Vesselot	12-Block 1-Berth		Container, Container,	Yardiobl.oad Quaylobl.oad	TC GC	TC <sub>13</sub>	Vessel <sub>st</sub> Vessel <sub>st</sub>	12-Block 1-Berth		Container <sub>4</sub> Container <sub>4</sub>	YardJobLoad QuayJobLoad	TC GC	TC <sub>12</sub> GC <sub>02</sub>	Vesselot Vesselot	12-Block 1-Berth





Data	Missingness	BREI	AE	MICE	ЕМВІ	RFI	KNNI
	30%	94.68% (0.26)	90.92% (0.26)	56.13% (1.77)	48.67% (1.12)	67.80% (1.78)	65.96% (0.15)
	35%	93.31% (0.15)	87.28% (0.35)	49.38% (1.82)	45.31% (2.31)	66.57% (2.09)	65.91% (0.15)
	40%	92.86% (0.30)	84.49% (0.28)	40.11% (1.34)	42.01% (2.49)	65.10% (2.48)	64.71% (1.37)
LD1	45%	91.59% (0.18)	81.75% (0.43)	38.54% (1.47)	40.23% (2.55)	64.44% (2.76)	61.44% (1.21)
	50%	91.02% (0.21)	78.99% (0.47)	33.97% (1.65)	39.23% (2.67)	63.29% (2.91)	59.21% (2.61)
	55%	90.64% (0.20)	76.51% (0.38)	29.67% (2.15)	36.78% (3.16)	62.48% (3.13)	55.77% (2.71)
	60%	89.45% (0.31)	74.50% (0.45)	27.77% (2.21)	32.43% (3.89)	60.12% (3.59)	50.62% (3.12)
	30%	95.31% (0.25)	84.12% (0.38)	61.21% (0.98)	35.88% (1.86)	45.21% (1.69)	37.41% (1.37)
	35%	94.83% (0.29)	80.87% (0.42)	60.99% (1.11)	34.27% (1.96)	43.33% (1.97)	35.21% (1.99)
	40%	94.75% (0.21)	78.83% (0.44)	58.98% (1.15)	34.01% (2.01)	40.67% (1.98)	34.12% (2.13)
LD2	45%	94.00% (0.27)	76.66% (0.47)	57.47% (1.21)	33.69% (2.43)	38.67% (2.04)	33.89% (2.56)
	50%	93.78% (0.26)	72.09% (0.49)	56.19% (1.87)	31.27% (2.59)	38.01% (2.05)	33.17% (2.78)
	55%	93.09% (0.26)	68.58% (0.57)	55.44% (1.93)	29.90% (2.64)	37.89% (2.33)	32.34% (3.31)
	60%	92.28% (0.39)	65.99% (0.63)	53.28% (2.00)	29.16% (2.67)	35.67% (2.65)	29.87% (4.01)
	30%	93.86% (0.17)	84.31% (0.44)	58.98% (0.78)	44.54% (0.57)	69.21% (0.98)	47.19% (2.10)
	35%	92.84% (0.19)	81.83% (0.49)	57.81% (0.82)	43.36% (0.67)	68.11% (1.41)	45.27% (2.17)
	40%	92.55% (0.28)	79.07% (0.59)	56.23% (0.99)	42.56% (0.91)	67.67% (1.64)	41.68% (2.12)
LD3	45%	92.11% (0.23)	77.06% (0.66)	54.45% (1.06)	40.24% (1.21)	66.26% (1.78).	38.10% (2.71)
	50%	91.35% (0.24)	76.20% (0.69)	52.77% (1.29)	39.97% (1.33)	65.79% (1.99)	35.12% (3.01)
	55%	90.99% (0.24)	75.48% (0.78)	48.31% (1.37)	38.21% (1.34)	65.21% (1.99)	33.39% (3.21)
	60%	90.25% (0.26)	74.59% (0.77)	45.65% (1.58)	37.75% (1.67)	64.98% (2.01)	29.94% (3.43)
	30%	96.87% (0.21)	85.67% (0.34)	41.21% (1.54)	55.56% (1.45)	58.76% (1.34)	39.48% (1.15)
	35%	95.28% (0.25)	85.37% (0.29)	39.56% (1.86)	51.32% (1.56)	55.31% (1.54)	36.01% (1.41)
	40%	94.49% (0.19)	84.50% (0.47)	35.11% (2.23)	47.47% (1.63)	53.29% (1.56)	33.81% (2.03)
LD4	45%	93.52% (0.25)	83.10% (0.55)	34.98% (2.31)	44.09% (2.05)	49.00% (1.60)	32.11% (2.36)
	50%	92.65% (0.21)	82.71% (0.60)	33.34% (2.47)	40.29% (2.41)	47.21% (1.71)	28.45% (2.78)
	55%	91.96% (0.29)	80.28% (0.69)	32.97% (2.51)	35.21% (3.33)	45.98% (1.75)	26.78% (3.13)
	60%	89.89% (0.36)	79.21% (0.71)	31.01% (2.83)	29.27% (4.65)	44.44% (2.12)	23.86% (3.35)
	30%	95.77% (0.15)	85.67% (0.34)	69.86% (1.08)	59.92% (1.25)	68.92% (1.64)	52.83% (1.39)
	35%	94.98% (0.13)	85.37% (0.29)	68.95% (1.29)	56.10% (1.27)	66.10% (1.76)	51.73% (1.58)
	40%	94.29% (0.17)	84.50% (0.47)	67.04% (1.41)	55.12% (1.53)	63.12% (1.91)	51.68% (2.17)
LD5	45%	93.57% (0.19)	83.10% (0.55)	66.54% (1.76)	48.27% (1.95)	59.27% (2.21)	50.22% (2.48)
	50%	92.30% (0.20)	82.71% (0.60)	65.72% (1.95)	44.69% (2.60)	51.69% (2.79)	49.03% (2.56)
	55%	91.71% (0.20)	80.28% (0.69)	63.25% (2.29)	39.67% (3.11)	48.67% (3.55)	47.89% (2.60)
	60%	91.26% (0.22)	79.21% (0.71)	61.09% (2.31)	39.57% (4.07)	41.57% (4.03)	46.69% (2.73)
	30%	93.14% (0.07)	80.67% (0.34)	64.67% (1.28)	67.99% (1.72)	63.24% (2.02)	64.72% (1.44)
	35%	92.37% (0.10)	79.37% (0.29)	59.32% (1.59)	64.72% (1.76)	62.65% (2.68)	63.65% (1.71)
LDC	40%	91.75% (0.18)	76.50% (0.47)	50.18% (2.16)	58.11% (1.99)	58.08% (3.15)	60.48% (1.82)
LD6	45%	91.09% (0.20)	75.10% (0.55)	45.03% (2.69)	54.80% (2.25)	54.26% (3.13)	58.58% (1.90)
	50%	90.61% (0.22)	73.71% (0.60)	40.52% (2.86)	51.78% (2.46)	52.32% (3.32)	56.58% (2.35)
	55%	90.10% (0.35)	71.28% (0.69)	38.08% (3.08)	44.81% (2.64)	49.05% (3.35)	55.98% (2.89)
	60%	89.71% (0.58)	70.21% (0.71)	32.51% (3.35)	35.65% (2.75)	45.33% (3.56)	52.87% (3.00)



