Business Anaytics

M. Tech QROR – 2nd yr (2024)

Data Pre-Processing

Prepatory Analytics-01

Dr. Prasun Das

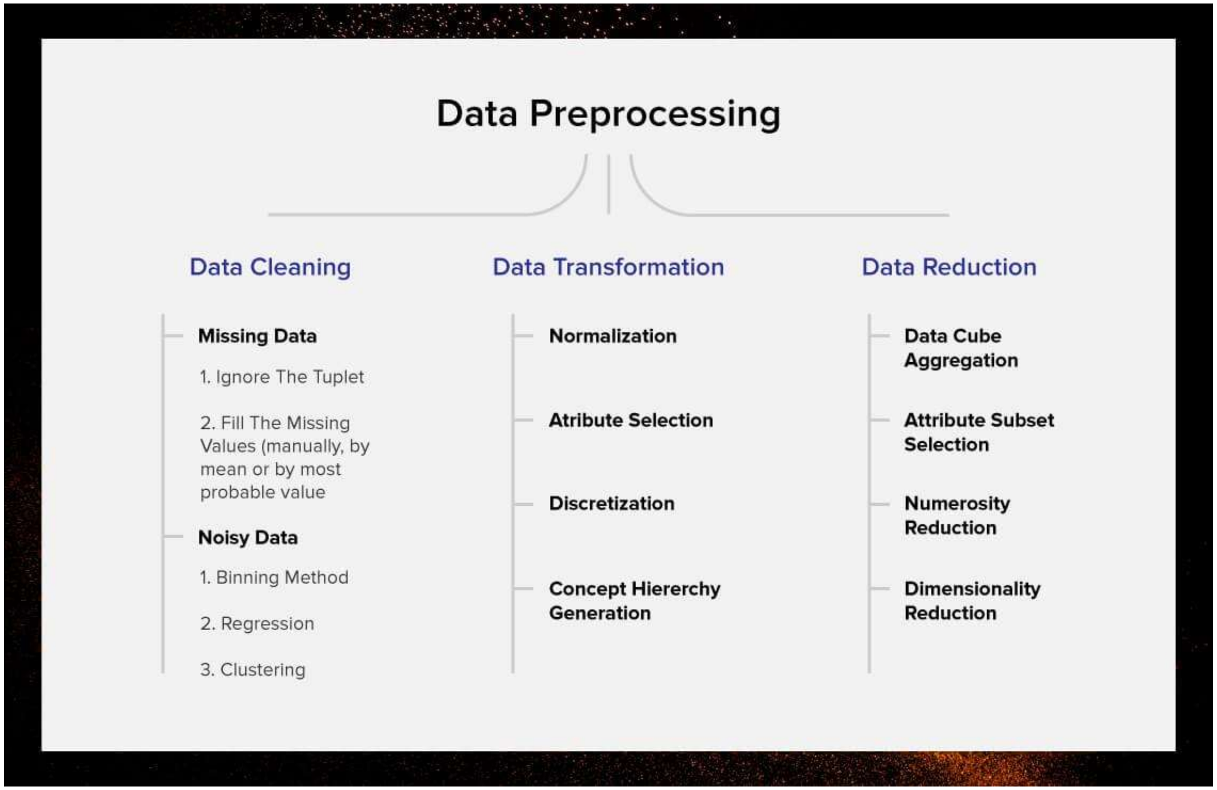
SQC & OR Unit

Indian Statistical Institute

e-mail: prasun@isical.ac.in

Data Pre-Processing

Data Pre-processing describes any type of processing performed on raw data to prepare it for subsequent data processing procedure.



Data Cleaning (Raw Data)

1. **Outlier** **-** ian observation which deviates significantly from the other observations. The general suspicion is that it was generated by a different mechanism [***Ref.:*** Hawkins, D. 1980. *Identification of Outliers*. Chapman and Hall.]

2. **Missing Data** - occur when no data value is stored for the variable in an observation. Data missing can have a significant effect on the conclusions. It's usually common, particularly for Big Data scenario [Ref.: – Rubin & Little , 2019. *Statistical Analysis with Missing Data*].

3. **Noisy Data –** through distribution fitting (Binning / Frequency distribution / Histogram), Regression (supervised), Clustering (unsupervised)

Techniques for Outlier/Anomaly Detection

a) **Statistical Tests:** Dixon's Q test, Mann-Kendall test, Young's test, mean & sd based, tail bounds of arbitrary distributions
b) **Depth-based** (indept. of stat. distrn.)**:** ISODEPTH [Ruts and Rousseeuw 1996], FDC [Johnson et al. 1998]
c) **Deviation-based** (minimum variance based): [Arning et al. 1996]: Naïve solution is in O(2*n) for n data objects*
d) **Distance-based** (ED, MD, kNN): Basic model [Knorr & Ng 1997]; Index-based [Knorr and Ng 1998]; Nested-loop based [Knorr and Ng 1998]; Grid-based [Knorr and Ng 1998]; Deriving intentional knowledge [Knorr and Ng 1999]; outlier score [Ramaswamy et al. 2000; Angiulli and Pizzuti 2002]; Resolution-based outlier factor (ROF) [Fan et al. 2006]; In-degree Number [Hautamaki et al. 2004]; ORCA [Bay and Schwabacher 2003]; RBRP [Ghoting et al. 2006]; reference points [Pei et al. 2006]; micro clusters [McCallum et al 2000; Tao et al. 2006]
e) **Density-based**: LOF [Breunig et al. 1999; Breunig et al. 2000]; top-n local outliers [Jin et al. 2001]; COF [Tang et al. 2002]; INFLO [Jin et al. 2006]; LOCI [Papadimitriou et al. 2003]
f) **High dimensional Approaches** (subspace, spatial, sequential, kernel); ABOD [Kriegel et al. 2008]; GBS [Aggarwal and Yu 2000]; SOD [Kriegel et al. 2009]

How Data are found "Missing" !!

- Missing data can occur because of <u>nonresponse</u>: no information is provided for one or more items or for a whole unit ("subject").

- Some items are more likely to generate a <u>nonresponse</u> than others: for ex., items about "private" subjects such as *income/salary*.

- <u>Attrition</u> ("Drop out") is a type of missingness that can occur in longitudinal studies - for ex., studying development of a patient's health condition over every week where a measurement is missing at certain time points.

- Missingness occurs when participants <u>drop out</u> before the test ends and one or more measurements are missing.

- Many more such experimental situations…

More Examples of "Missing Data" !!

case	A	B	C	D	E	F
1	a_1	b_1	$*$	d_1	e_1	$*$
2	a_2	$*$	c_2	d_2	e_2	$*$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	a_n	b_n	c_n	$*$	$*$	$*$

- non-reply in surveys (leading questions → avoidance);
- non-reply for specific questions: ”missing” ~ don’t know
- just not recorded (e.g. too expensive / not reqd. technically/casually)

Different types of missing-ness demand different treatments.

Example 1. Six-Cities data

- Consider the data from the Six Cities longitudinal study of the health effects of respiratory function in children (Ware et al., 1984). This is a well known environmental dataset that has been analyzed extensively in the literature.
- The binary response is the wheezing status (no wheeze, wheeze) of a child at age 11.
- The wheezing status is modeled as a function of the city of residence (x_1) and smoking status of the mother (x_2).
- The covariate x_1 is a binary covariate which equals 1 if the child lived in Kingston-Harriman, Tennessee, the more polluted city, and 0 if the child lived in Portage, Wisconsin.
- The covariate x_2 is maternal cigarette smoking measured in number of cigarettes per day.
- There are $n = 2394$ subjects in the dataset. The covariate x_1 is missing for 32.8% of the cases, and x_2 is missing for 3.3% of the cases.

Table 1.1: Summary of the Six-Cities Data

y		x_1		x_2	
0	$N = 1827(76.3\%)$	0	$N = 862(36.0\%)$	Obs'ved	mean 7.2 (s.d. 11.3)
1	$N = 567(23.7\%)$	1	$N = 747(31.2\%)$	NA	$N = 79(3.3\%)$
		NA	$N = 785(32.8\%)$		

Example 2. Liver cancer data

- Consider data on $n = 191$ patients from two Eastern Cooperative Oncology Group clinical trials, EST 2282 (Falkson et al., 1990) and EST 1286 (Falkson et al., 1994).
- Here, we are primarily interested in the patient's status as he/she enters the trials.
- In particular, we are interested in how the number of cancerous liver nodes (y) when entering the trials is predicted by six other baseline characteristics: time since diagnosis of the disease in weeks (x_1), two biochemical markers (each classified as normal or abnormal): Alpha fetoprotein (x_2), and Anti Hepatitis B antigen (x_3); associated jaundice (yes, no) (x_4), body mass index (x_5) (defined as weight in kilograms divided by the square of height in meters), and age in years (x_6).
- Table 1.2 shows that 28.8% of the patients have at least one covariate missing. The biochemical marker Anti-hepatitis B antigen, which is not easy to obtain, has the highest proportion missing.

Table 1.2: Missingness summary of the liver cancer data

Variable	Missing $N(\%)$
Time Since Diagnosis	17 (8.9%)
Alpha Fetoprotein	11 (5.8%)
Anti Hepatitis B	35 (18.3%)
Overall	55 (28.8%)

Notations: Missing Data

Data Matrix: $Y = \{Y_{ij}\}$

Notation: $Y = (Y_{obs}, Y_{mis})$

Missing data matrix: M = $\{M_{ij}\}$:

$$M_{ij} = \begin{cases} 1 \; if \; Y_{ij} \; is \, missing \\ 0 \; if \; Y_{ij} \; is \, observed \end{cases}$$

Patterns of "Missing Data" !!

Let us illustrate with a case of cross-classification of Sex (S), Race (R), Admission (A) and Department (D).

Univariate: $M_{ij} = 0$ unless $j=j^*$, e.g. an unmeasured response. Ex.: R unobserved for some, but data otherwise complete.

Multivariate: $M_{ij} = 0$ unless $j \in J \subset V$, as above, just with multivariate response, e.g. in surveys. Ex.: For some subjects, both R and S unobserved.

Patterns of "Missing Data" !!

Monotone: There is an ordering of V so $M_{ik} = 0$ implies $M_{ij} = 0$ for $j < k$, e.g. drop-out in longitudinal studies. Ex.: For some, A is unobserved, others neither A nor R, but data otherwise complete.

Disjoint: Two subsets of variables never observed together. Controversial (appears in Rubin's causal model). Ex.: S and R never both observed.

General: none of the above. Haphazardly scattered missing values. Ex.: R unobserved for some, A unobserved for others, S,D for some.

Latent: A certain variable is never observed. May be it is even unobservable. Ex.: S never observed, but believed to be important for explaining the data.

Type-1: Missing Completely at Random (MCAR)

If any particular data-item being missing are <u>independent both</u> of observable variables (sample) and of unobservable parameters (population) of interest, and occur entirely at random (i.e., probability of being missing is the same for all cases, no <u>systematic</u> pattern).

This indicates that the probability of a datum being missing is independent of both the variable itself, as well as of all other variables included in the data set (King et. al. 2001, Allison 2009).

When data are MCAR, the analyses performed on the data are unbiased; however, data are **rarely** MCAR.

Type-2: Missing at Random (MAR)

In this case, failure to observe a value does not depend on the unobserved value, given the observed data. This indicates that the probability that a specific datum is missing is independent of the value of the datum itself, but may depend on other variables which are included in the data set (Allison 2009, Schafer 1997).

Ex.: males are less likely to fill in a depression survey but this has nothing to do with their level of depression, after accounting for maleness.
Ex.: older people are more likely to respond to question #10 of a survey than younger people.

It is more realistic than MCAR, and conditioned on other variables.

These data can still induce parameter bias in analysis due to the contingent emptyness of cells [(male, very high depression) or (younger people, question # 10) may have zero entries].

Type-3: Missing Not at Random (MNAR)

In this case, failure to observe a value depends on the value that could have been observed. This means that whether a specific value for ***Y*** is missing or observed is dependent on the value of ***Y*** itself, and that it is not possible to predict this value from the values of the observed set of variables ***X*** (Allison 2009, King et. al. 2001).

MNAR (also known as non-ignorable nonresponse) analysis are problematic because the distribution of the missing observations do not only depend on the observed values but also the unobserved values as well.

Ex.: people with lowest education are missing on education; most of the absenteeism (missing attendance) are from the sickest people; few specific questions on a survey tend to be skipped deliberately.

To extend the *previous example*, this would occur if men failed to fill in a depression survey because of their level of depression.

Techniques of Dealing with Missing Data

Use methods of data analysis that are robust to missingness.

a) Partial deletion
b) Imputation
c) Full analysis (ML/EM)
d) Interpolation

Partial Deletion - MCAR

Methods which involve reducing the data available to a dataset having no missing values include:

- Listwise / Casewise deletion
- Pairwise deletion (*multicollinearity* issue)

Imputation

1. **Mean / Median Imputation (**MCAR case, preserve unbiased estimates for mean, but underestimates std. error and might introduce bias for relationship among variables**) :**
2. **Regression Imputation (**predict the missing value of a variable and add a random component**):**
3. **Predictive Mean Matching (or, Hot Deck Imputation):**

   It calculates the predicted value of target variable ***Y*** according to the specified imputation model. For each missing entry, the method forms a small set of candidate donors (typically with 3, 5 or 10 members) from all complete cases that have predicted values closest to the predicted value for the missing entry. One donor is randomly drawn from the candidates, and the observed value of the donor is taken to replace the missing value. Imputations outside the observed data range will not occur, thus evading problems with meaningless imputations (e.g., negative body height). The **assumption** is the distribution of the missing cell is the same as the observed data of the candidate donors. It is fairly robust to transformations of the target variable.

Imputation

5. **Multiple Imputation (**repeatable approach: MAR case → unbiased estimates**):**

   Rubin (1987) argued that even a small number (5 or fewer) of repeated imputations capture most of the relative efficiency and enormously improves the quality of estimation. However, a too-small number of imputations can lead to a substantial loss of statistical power.

   Alison, P. (2000). *Sociological methods and research*, 28, 301-309

6. **Cold Deck, Deductive Imputation: ???**

ASSIGNMENT

Summary: Missing Data Methods

Method	Bias	Data Use	Variability	Speed	Ease-of-use
Listwise Deletion	Unbiased for MCaR	Minimal	Too Large	Instant	Very Easy
Pairwise Deletion	Unbiased for MCaR	Medium	Unclear	Instant	Difficult
Mean Imputation	Biased for all but μ_{MCaR}	Medium	Too Small	Instant	Very Easy
Regression Imputation	Unbiased for MaR	High	Too Small	Fast	Easy
Regression Imputation w. uncertainty	Unbiased for MaR	Maximal	Appropriate	Fast	Medium
Predictive Mean Matching	Unbiased for MaR	Maximal	Appropriate	Fast	Medium
Data Augmentation	Unbiased for MaR	Maximal	Appropriate	Slow	Difficult
MICE	Unbiased for MaR	Maximal	Appropriate	Very Slow	Difficult
EM Bootstrapping	Unbiased for MaR	Maximal	Appropriate	Medium	Medium

Full Analysis - EM Algorithm

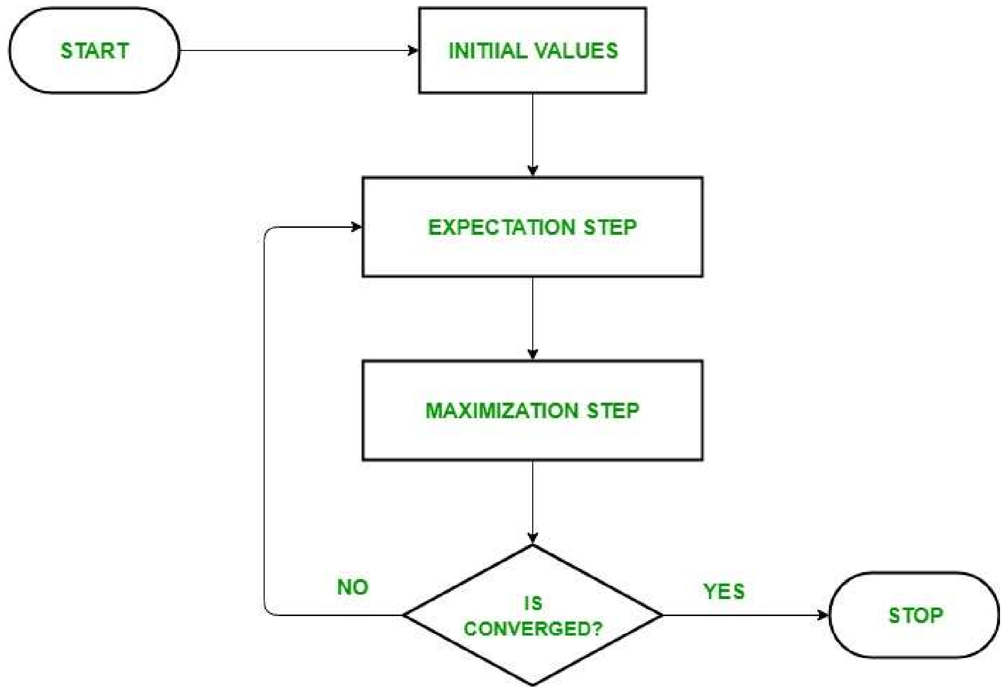
Full Analysis takes full account of all information available, **without** the distortion resulting from using imputed values as if they were actually observed:

- **E**xpectation-**M**aximization (EM) algorithm

The EM algorithm is one of the most commonly used method in machine learning to obtain maximum likelihood estimates (MLE) of parameters that are sometimes observable and sometimes not. However, it is also applicable to unobserved data or sometimes called latent.

Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society, Series B*. **39** (1): 1–38.

EM Algorithm (Iterative)



Step-1: Expectation (E) - creates a log-likelihood function using the current estimate for the parameters.

Step-2: Maximization (M) – computes parameters maximizing the expected log-likelihood function of **Step-1**. These parameter-estimates are then used to determine the distribution of the latent variables in the next E-step.

EM Algorithm – MCAR Case

Let, $Pr\ (M_Y = 1/\boldsymbol{X}, Y)$ be the probability that a specific datum is missing given the value of itself and all other variables in the data set.

Then, $Pr\ (M_Y = 1/\boldsymbol{X}, Y) = Pr\ (M_Y = 1) \Rightarrow$ probability of a specific value of Y being missing, conditioned on both $\boldsymbol{X}$ and Y itself, is the same as the <u>unconditional</u> probability that the same specific value of Y being missing.

If there is another set of variables, say $\boldsymbol{Z}$, entirely unobserved (latent), which affect the missingness of the data, and which are uncorrelated with both $\boldsymbol{X}$ and Y, such that

$$Pr\ (M_Y = 1/\boldsymbol{X}, Y, \boldsymbol{Z}) = \mathrm{Pr}\ (M_Y = 1/\boldsymbol{Z}) \neq \mathrm{Pr}\ (M_Y = 1), \text{ and}$$
$$Cov(Y, \boldsymbol{Z}) = Cov(\boldsymbol{X}, \boldsymbol{Z}) = 0$$

then, the data are still considered MCAR, primarily because of no correlation of $\boldsymbol{Z}$ with both Y and $\boldsymbol{X}$.

EM Algorithm - MAR Case

$Pr\ (M_Y = 1/\boldsymbol{X}, Y) = \Pr\ (M_Y = 1/\boldsymbol{X}) \Rightarrow$ missingness of Y is independent of Y itself, when conditioned on $\boldsymbol{X}$.

For ex., if in a survey, the sample is divided into different groups (***X***-variables) and the probability that the individuals answer a specific question (*Y*-variable) varies between the groups, but does not vary (w.r.t *Y*) within the groups, then the missing data mechanism would generate data <u>missing at random</u> on *Y* (Allison 2009, Graham 2009).

It is easy to see from the above expression that <u>MCAR is a special case of MAR</u>, since if the missingness of *Y* is independent of ***X***, the expression reduces to

$$Pr\ (M_Y = 1/\boldsymbol{X}, Y) = \Pr\ (M_Y = 1/\boldsymbol{X}) = \Pr\ (M_Y = 1)$$

EM Algorithm - MAR Case

The MAR assumption is in many cases relatively reasonable since if ***X*** and ***Y*** are correlated, conditioning the missingness of ***Y*** on ***X*** may account for a pattern of missingness.

$$Pr\ (M_Y = 1/\boldsymbol{X}, Y) = \Pr\ (M_Y = 1/\boldsymbol{X}) \neq \Pr\ (M_Y = 1)$$

If now, the missingness of ***Y*** is dependent on some set of variables ***Z*** which are included in the data set, independent of **X**, and observed, but ***Y*** and ***Z*** are correlated, then

$$Pr\ (M_Y = 1/\boldsymbol{X}, Y, \boldsymbol{Z}) = \Pr\ (M_Y = 1/Y, \boldsymbol{Z}) = \Pr\ (M_Y = 1/\boldsymbol{Z})$$

which means that the missingness of ***Y*** is MAR when ***Z*** is included in the data. This indicates that if the data are **MNAR**, it may be possible to correct this by adding more variables to the data set (Graham 2009, Graham & Donaldson 1993).

EM Algorithm - MNAR Case

There are several different methods for dealing with MNAR data (King *et al*. 2001). since each MNAR situation is unique, and it is therefore difficult to create generalized approaches to MNAR data. It is often possible to improve the MNAR situation by including more variables in the data set. This may not necessarily turn the MNAR missing data into MAR missing data, but it may reduce the severity of the MNAR problem such that

$$Pr\,(M_Y = 1/\boldsymbol{X}, Y, \boldsymbol{Z}) \approx \mathrm{Pr}\,(M_Y = 1/\boldsymbol{X}, \boldsymbol{Z})$$

which may allow for an approximate MAR assumption under mild MNAR conditions (Allison 2012).

EM Algorithm

We begin by assuming that the complete data-set consists of $\mathcal{Z} = (\mathcal{X}, \mathcal{Y})$ but that only $\mathcal{X}$ is observed. The complete-data log likelihood is then denoted by $l(\theta; \mathcal{X}, \mathcal{Y})$ where θ is the unknown parameter vector for which we wish to find the MLE.

E-Step: The E-step of the EM algorithm computes the expected value of $l(\theta; \mathcal{X}, \mathcal{Y})$ given the observed data, $\mathcal{X}$, and the current parameter estimate, θ_{old} say. In particular, we define

$$\begin{aligned} Q(\theta; \theta_{old}) &:= \mathsf{E}\left[l(\theta; \mathcal{X}, \mathcal{Y}) \mid \mathcal{X}, \theta_{old}\right] \\ &= \int l(\theta; \mathcal{X}, y)\, p(y \mid \mathcal{X}, \theta_{old})\, dy \end{aligned} \tag{1}$$

where $p(\cdot \mid \mathcal{X}, \theta_{old})$ is the conditional density of $\mathcal{Y}$ given the observed data, $\mathcal{X}$, and assuming $\theta = \theta_{old}$.

M-Step: The M-step consists of maximizing over θ the expectation computed in (1). That is, we set

$$\theta_{new} := \max_{\theta} Q(\theta; \theta_{old}).$$

We then set $\theta_{old} = \theta_{new}$.

The two steps are repeated as necessary until the sequence of θ_{new}'s converges. Indeed under very general circumstances convergence to a local maximum can be guaranteed and we explain why this is the case below. If it is suspected that the log-likelihood function has multiple local maximums then the EM algorithm should be run many times, using a different starting value of θ_{old} on each occasion. The ML estimate of θ is then taken to be the best of the set of local maximums obtained from the various runs of the EM algorithm.

EM Algorithm: Advantages & Disadvantages

Advantages

- The basic two steps, i.e, E-step and M-step are often easy for many of the machine learning problems in terms of implementation.
- The solution to the M-steps often exists in the closed-form.
- It is always guaranteed that the value of likelihood will increase after each iteration.

Disadvantages

- It has **slow convergence**.
- It converges to the **local optimum** only.
- It takes both forward and backward probabilities into account. This thing is in contrast to that of numerical optimization which considers only **forward probabilities**.

Interpolation - Assignment

In the mathematical field of numerical analysis, interpolation is a method of constructing new data points within the range of a discrete set of known data points.

- Linear Interpolation
- Spline Interpolation
- Exponential Smoothing
- Polynomial Interpolation
- Combined Function Interpolation

Case Studies

Case Study-01

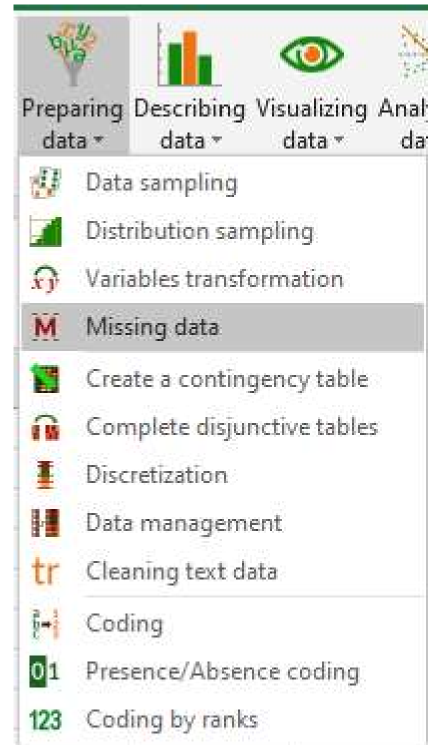
Original Data: Fisher's IRIS Data Set

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
50	33	14	2	Setosa
67	31	56	24	Virginica
46	34	14	3	Setosa
69	31	51	23	Virginica
59	32	48	18	Versicolor
46	36	10	2	Setosa
61	30	46	14	Versicolor
65	30	52	20	Virginica
65	30	55	18	Virginica
68	32	59	23	Virginica
51	33	17	5	Setosa
62	34	54	23	Virginica
77	38	67	22	Virginica
63	33	47	16	Versicolor
67	33	57	25	Virginica
76	30	66	21	Virginica
55	35	13	2	Setosa
67	30	52	23	Virginica

Missing data at Random

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
50		14	2	Setosa
67	31		24	Virginica
46	34		3	Setosa
69			23	Virginica
				Versicolor
	36		2	Setosa
61		46		Versicolor
	30	52	20	Virginica
65	30			Virginica
	32	59	23	Virginica
	33		5	Setosa
62	34	54	23	Virginica
77	38	67	22	Virginica
63	33		16	Versicolor
	33			Virginica
76	30	66	21	Virginica
55	35	13	2	Setosa
67		52	23	Virginica

Select (MS Excel):
XLSTAT/ Preparing data / Missing data



Missing Data Dialogue Box

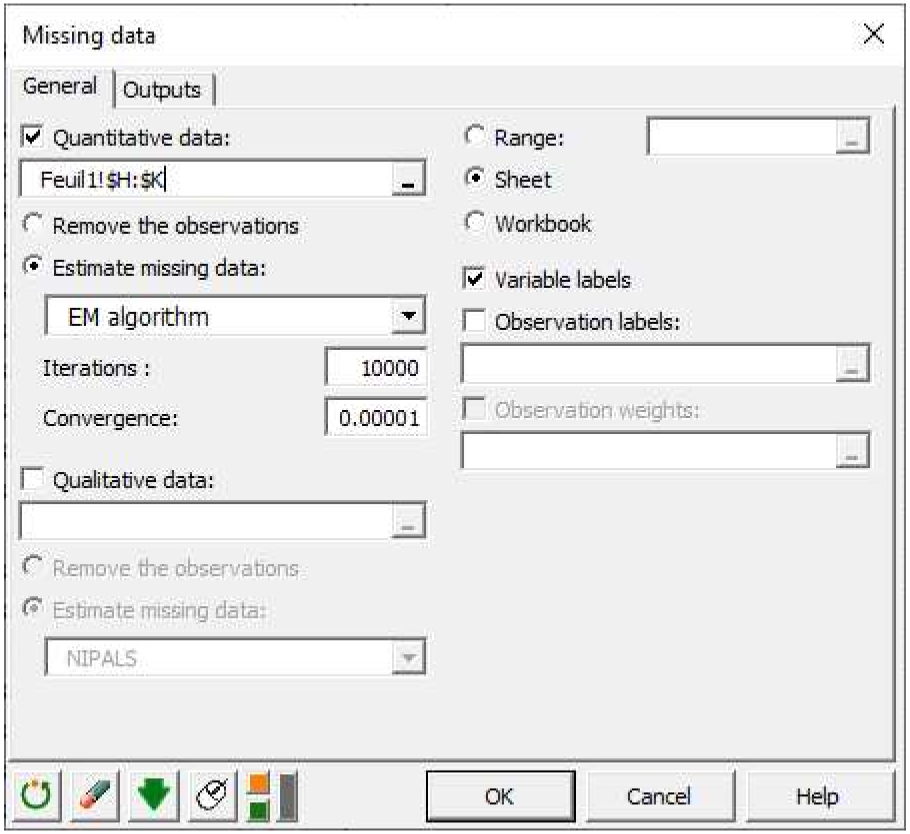
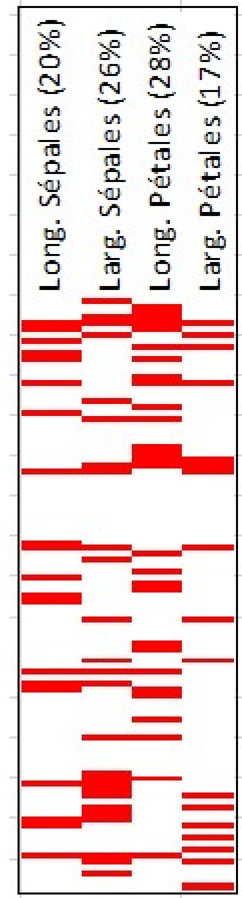


Chart representing missing data in red.
No particular patterns found.



Descriptive Statistics

Summary statistics (Before treatment):

Variable	Observations	Obs. with missing	Obs. without	Minimum	Maximum	Mean	Std. deviation
Long. Sépale	100	20	80	43,000	79,000	58,050	8,388
Larg. Sépales	100	26	74	20,000	44,000	30,892	4,452
Long. Pétale	100	28	72	11,000	69,000	38,931	17,695
Larg. Pétales	100	17	83	1,000	24,000	11,614	7,659

Summary statistics (Post treatment):

Variable	Observations	Obs. with missing	Obs. without	Minimum	Maximum	Mean	Std. deviation
Long. Sépale	100	0	100	43,000	79,000	57,950	8,093
Larg. Sépales	100	0	100	20,000	44,000	31,026	4,511
Long. Pétale	100	0	100	5,873	69,000	37,632	17,959
Larg. Pétales	100	0	100	0,014	24,000	11,873	7,560

Completed data (Imputed through EM Algo):

	Sepal Length	Sepal Width	Petal Length	Petal Width
Obs1	50.000	**32.848**	14.000	2.000
Obs2	67.000	31.000	**65.503**	24.000
Obs3	46.000	34.000	**17.079**	3.000
Obs4	69.000	**24.737**	**64.392**	23.000
Obs5	**56.743**	**32.128**	**48.696**	**15.883**
Obs6	**50.120**	36.000	**15.854**	2.000
Obs7	61.000	**31.942**	46.000	**16.295**
Obs8	**59.766**	30.000	52.000	20.000
Obs9	65.000	30.000	**55.631**	**16.855**
Obs10	**68.258**	32.000	59.000	23.000
Obs11	**59.156**	33.000	**25.471**	5.000
Obs12	62.000	34.000	54.000	23.000
Obs13	77.000	38.000	67.000	22.000
Obs14	63.000	33.000	**49.980**	16.000
Obs15	**56.814**	33.000	**20.351**	**6.238**
Obs16	76.000	30.000	66.000	21.000
Obs17	55.000	35.000	13.000	2.000
Obs18	67.000	**33.772**	52.000	23.000
Obs19	70.000	32.000	**52.791**	14.000
Obs20	**60.322**	32.000	45.000	15.000

Original Data: Fisher's IRIS Data Set

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
50	33	14	2	Setosa
67	31	56	24	Virginica
46	34	14	3	Setosa
69	31	51	23	Virginica
59	32	48	18	Versicolor
46	36	10	2	Setosa
61	30	46	14	Versicolor
65	30	52	20	Virginica
65	30	55	18	Virginica
68	32	59	23	Virginica
51	33	17	5	Setosa
62	34	54	23	Virginica
77	38	67	22	Virginica
63	33	47	16	Versicolor
67	33	57	25	Virginica
76	30	66	21	Virginica
55	35	13	2	Setosa
67	30	52	23	Virginica

Note: imputed values are close to the true ones. For ex., we get 32.8 instead of 33 for the first observation.

More unbiased estimate than using any other type of imputation.

Case Study-02:

Example: Suppose a telecom company wants to analyze the performance of its circles based on the following parameters

1. Current Month's Usage
2. Last 3 Month's Usage
3. Average Recharge
4. Projected Growth

The data set is given in next slide. (Missing_Values_Telecom Data)

Example: Circle wise Data

SL No.	Current Month's Usage	Last 3 Month's Usage	Average Recharge	Projected Growth	Circle
1	5.1	3.5	99.4	99.2	A
2	4.9	3	98.6	99.2	A
3		3.2		99.2	A
4	4.6	3.1	98.5	9..2	A
5	5		98.4	99.2	A
6	5.4	3.9	98.3	99.4	A
7	7	3.2	95.3	98.4.	B
8	6.4	3.2	95.5	98.5	B
9	6.9	3.1	95.1	98.5	B
10		2.3	96	98.3	B
11	6.5	2.8	95.4	98.5	B
12	5.7		95.5	98.3	B
13	6.3	3.3		98.6	B
14	6.7	3.3	94.3	97.5	C
15	6.7	3	94.8	97.3	C
16	6.3	2.5	95	98.9	C
17		3	94.8	98	C
18	6.2	3.4	94.6	97.3	C
19	5.9	3	94.9	98.8	C

Option 1: Discard all records with missing values

```
>newdata = na.omit(mydata)
>write.csv(newdata,"E:/ISI/newdata.csv")
```

SL.No.	Current.Month.s.Usage	Last.3.Month.s.Usage	Average.Recharge	Projected.Growth	Circle
1	5.1	3.5	99.4	99.2	A
2	4.9	3	98.6	99.2	A
4	4.6	3.1	98.5	9..2	A
6	5.4	3.9	98.3	99.4	A
7	7	3.2	95.3	98.4.	B
8	6.4	3.2	95.5	98.5	B
9	6.9	3.1	95.1	98.5	B
11	6.5	2.8	95.4	98.5	B
14	6.7	3.3	94.3	97.5	C
15	6.7	3	94.8	97.3	C
16	6.3	2.5	95	98.9	C
18	6.2	3.4	94.6	97.3	C
19	5.9	3	94.9	98.8	C

Option 2: Replace the missing values with variable mean, median, etc

Replacing the missing values with men

SL No	cmusage	l3musage	avrecharge	Proj Growth	Circle
1	5.1	3.5	99.4	11	1
2	4.9	3	98.6	11	1
3	5.975	3.2	96.14117647	11	1
4	4.6	3.1	98.5	1	1
5	5	3.105882353	98.4	11	1
6	5.4	3.9	98.3	12	1
7	7	3.2	95.3	6	2
8	6.4	3.2	95.5	7	2
9	6.9	3.1	95.1	7	2
10	5.975	2.3	96	5	2
11	6.5	2.8	95.4	7	2
12	5.7	3.105882353	95.5	5	2
13	6.3	3.3	96.14117647	8	2
14	6.7	3.3	94.3	3	3
15	6.7	3	94.8	2	3
16	6.3	2.5	95	10	3
17	5.975	3	94.8	4	3
18	6.2	3.4	94.6	2	3
19	5.9	3	94.9	9	3

What about EM Algorithm ??