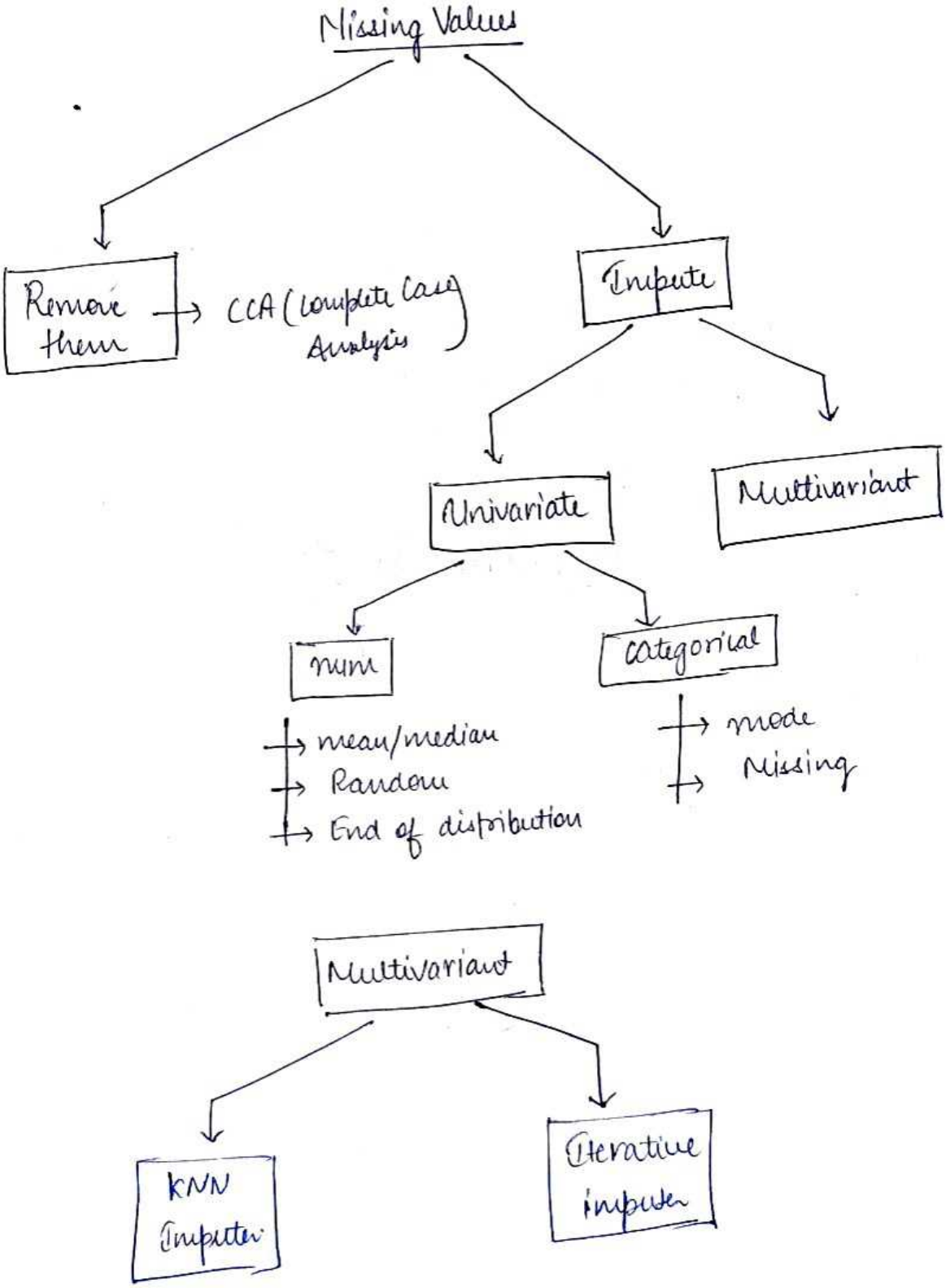
Handling Missing Data



Complete Case Analysis

Complete-case analysis (CCA) also called "list-wise deletion" of cases, consists in discarding observations where values in any of the variable are missing.

Complete case Analysis means literally analyzing only those observations for which there is information in all of the variables in the datasets.

Assumption for CCA

1) Missing Completely at Random

1000, 4↴

Age → 50 → missing value (Randomly)

↳ 950, 4

Advantage / Disadvantage

Advantage

1. Easy to implement as no data manipulation required.
2. Preserves variable distribution (if data is MCAR, then the distribution of the variables of the reduced dataset should match the distribution in the original dataset.

Disadvantages

1) It can exclude a large fraction of the original dataset (if missing data is abundant)

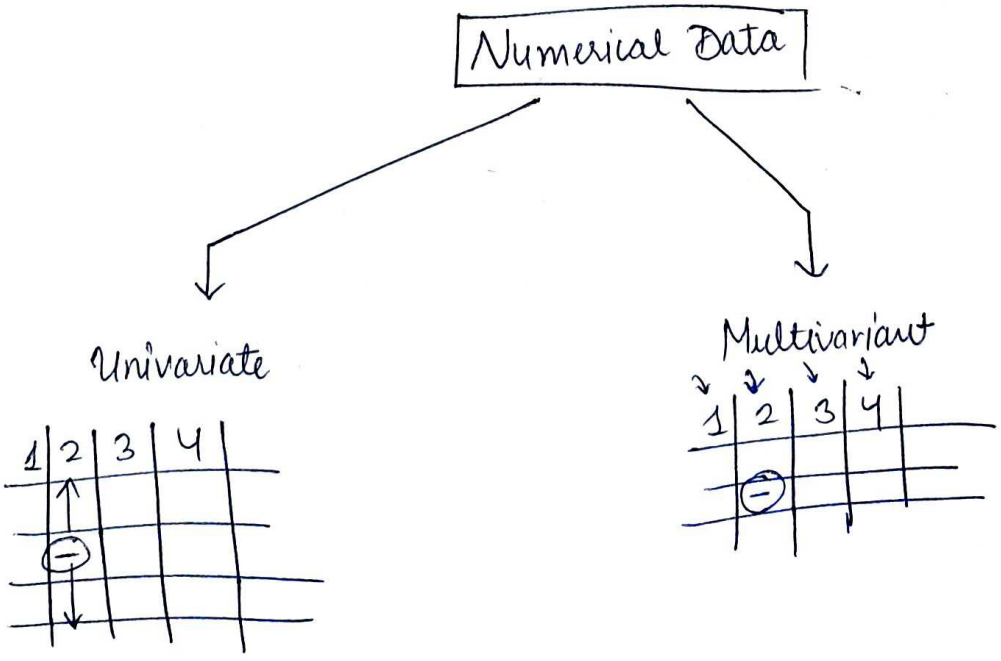
2) Exclude information could be informative for the analysis (if data is not missing at random)

3) When using our model in production, the ~~production~~ model will not know how to handle missing data.

When to use CCA?

1. MCAR
2. 5% < (less than 5%)

Handling Missing Numerical Data



Mean / Median Imputation

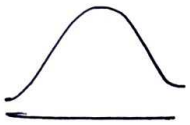
Age
27
32
Na
27

Na → Mean
Na → Median

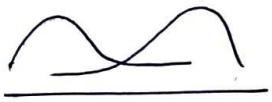
Benefit

1. Simple
2. 5%>

Mean →

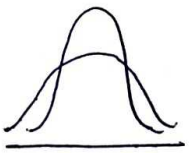


Median →



Disadvantage:

1.) Distribution shape →



2.) Outliers

3.) Covariance / Corelation changes

3. Arbitrary Value Imputation

↳ category → NA → Missing

Num → -1
→ 0
→ 99.999
→ Any Random Number

Data is not missing at Random

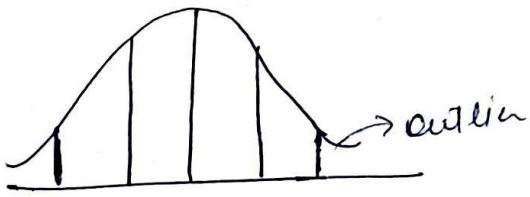
Benfit → easy to apply

Disadvantage:

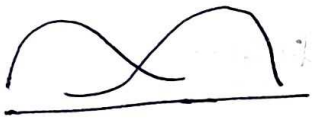
→ Distribution shape
→ Variance
→ Covariance Changes

End of Distribution Imputation

End of Distribution



- **Normality**
  - (mean + 3σ)
  - (mean − 3σ)
- **Skewed** (IQR priority)



$Q_3 - Q_1$

- 75th
- 25th

- $Q_1 - 1.5\ IQR$
- $Q_3 - 1.5\ IQR$

Benefit → Easy to apply

disadvantage:
1. Distribution shape
2. Outliers
3. Covariance / corelation changes

Handling Categorical Missing Data

- Most frequent
- Missing
  - Random

Most Frequent Value Imputation

MCAR
5%

Num → Mean / Median

Cat → Mode → frequency

Benifit → easy to apply

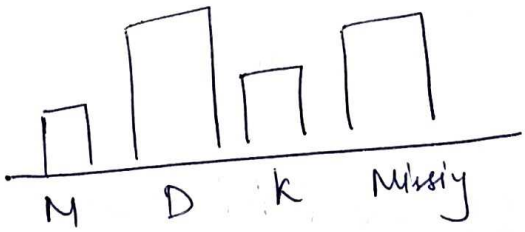
Disadvantage
↳ Distribution Shape (change)

Missing Category Imputation

City
Mumbai
Delhi
Kolkata

10% → Outlier

New column → Missing



Random Imputation

Age
26
32
56
41
NA
-
-
-
NA

Random Numbers

Gender
M
F
F
M
NA

Benefit → easy to apply
→ Data distribution and variance same

Random Imputation → Linear Model
Logistic Model

Disadvantage

1) Memory heavy for deployment, as we need to store the original training set to extract values from oneplace the NA is coming Observation.

2) Well suited for linear models as it does not distort the distribution, regardless of the % of NA.

Missing Indicator

Age	Fare	Age - NA
27	31	F
41	35	F
NA	41	T
62	32	F

Automatically select value for Imputation

Use Grid Search CV

↳ Try different combination and select prefed one.

Multivariant

KNN Imputer

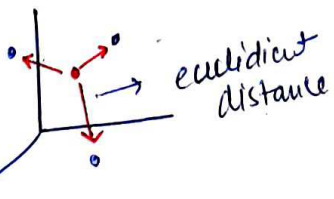
Working : There are 4 columns first column have Null value. So find value of null value on the basis of which row similar to Null value of row. Similarity found with the help of euclidient distant.

S.No	Feature 1	Feature 2	Feature 3	Feature 4
1	33	—	67	21
2	—	45	68	12
3	23	51	71	18
4	40	—	81	—
5	35	60	79	—

Let we have 3-D data

(33, —, 67)

K = No. of Neighbour



* In Euclidean distance, we use (x,y) $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$ for 2D and for 3D (X,Y,Z) ↗

* But sometimes in data have more than 1 Null value. So, we use Nan-euclidean distance

Formula →

dist(x,y) = sqrt(weight * sq. distance from present coordinates)

Where,
weight = (Total # of coordinates)/(# of present coordinates)

Let's calculate distance

Ignore bcz of NULL value present

SNo	Feature 1	Feature 2	Feature 3	Feature 4
1	33	—	67 (2)	21 (3)
2	—	45 (1)	68	12
3	23	51	71	18
4	40	—	81	—
5	35	60	79	—

$distance = \sqrt{\frac{3}{2}(67-68)^2+(21-12)^2}$

2 → weight

Total No. of Coordinate = 3

Total No. of Present Co-ordinate = 2

* We choose Row1 and Row2.

We can also check Row 2 and Row 3, or Row4, Row5

Distance bet^n Row2 and Row1

$$d = \sqrt{\frac{3}{2}(68-67)^2 + (12-21)^2}$$

$$\boxed{d = 11.29}$$

Distance bet^n Row2 and Row3

$$d = \sqrt{\frac{3}{3}(151-45)^2 + (71-68)^2 + (18-12)^2}$$

$$\boxed{d = 9}$$

Dist^n bet^n Row2 and Row4

$$d = \sqrt{\frac{3}{1}((81-68))^2}$$

$$\boxed{d = 5.19}$$

Distance bet^n Row2 and Row5

$$d = \sqrt{\frac{3}{2}((60-45)^2 + (79-68)^2}$$

$$\boxed{d = 14.79}$$

Closest 2 Number → (9) and (5.19)

(9) → Row3

(5.19) → Row4 → (40)

Row3 → (23)

(23) and (40) → Value in column first where Null value present

$$\frac{23+40}{2} = \frac{63}{2} = 31.5$$ in Value of Null Value in column first.

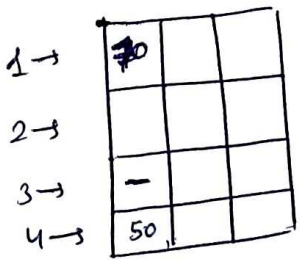
* If I consider 3 value or closst 3 value than

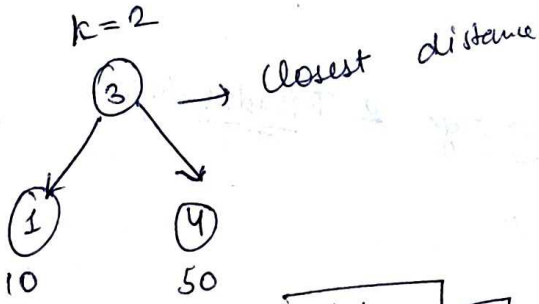
$$\frac{23+40+33}{2} = \text{Null Value}$$

Advantages & Disadvantages

1. More accurate
2. More Number of Calculation (Calculate distance with every point)
3. Upload Train Data on server.

Sklearn → weight = Uniform, distance





Uniform

$\frac{70+50}{2} = 60$ → Null Value

distance

$\frac{\frac{1}{10} \times 70 + \frac{1}{50} \times 50}{2} = 4$ → Null Value

Iterative Imputer / MICE

MICE stands for Multivariate Imputation by chained Equation.

There are four types of missing data

1. Missing Completely at Random (MCAR)
2. Missing At Random (MAR)
3. Missing Not At Random (MNAR)

Assumption → When Missing data in MAR.

↓

Means when optionally data Missing (ex: option → you have to fill date of birth some of people fill some of not

missig data relate to Other

MNAR

→ intentionally data Missing

Not relate

Advantage & Disadvantage

↳ accurate

↳ slow process

↳ Train Data on server

How it works?

1. Actual Dataset

	R&D spend	Administration	Marketing Spend	Profit
21	8.0	15.0	30.0	11.0
37	4.0	5.0	20.0	9.0
2	15.0	10.0	41.0	19.0
14	12.0	16.0	26.0	13.0
44	2.0	15.0	3.0	7.0

2. Removing the target column

	R&D spend	Administration	Marketing spend
21	8.0	15.0	30.0
37	4.0	5.0	20.0
2	15.0	10.0	41.0
14	12.0	16.0	26.0
44	2.0	15.0	3.0

3. Introduced some fake nan values

	R&D Spend	Administration	Marketing Spend
21	8.0	15.0	30.0
37	NAN	5.0	20.0
2	15.0	10.0	41.0
14	12.0	NAN	26.0
44	2.0	15.0	NAN

Step1→ Fill all the NAN values with mean of respective col

	R&D spend	Administration	Marketing Spend
21	8.00	15.00	30.00
37	9.25 mean	5.00	20.00
2	15.00	10.00	41.00
14	12.00	mean→ 11.25	26.00
44	2.0	15.00	mean→ 29.25

Step2:→ Remove all column missing Values

	R&D	Admins	MS
21	8.0	15.00	30.0
37	NAN	5.00	20.0 → Test data
2	15.0	10.00	41.00
14	12.0	11.25	26.00
44	2.0	15.00	29.25

	Admins	MS
21	15.00	30.00
2	10.00	41.00
14	11.25	26.00
44	15.00	29.25

Training Data

Output Data

Input data

Step3:- Predict the missing value of col1 using other cols

	R&D Spend	Administration	Marketing Spend
21	8.00	15.00	30.00
37	23.14 (predicted value)	5.00	20.00
2	15.00	10.00	41.00
14	12.00	11.06	26.00
44	2.00	15.00	29.25

Step4:- Remove all col2 missing Value

	R&D	Admin	MS
21	8.0	15.0	30.0
37	23.14	5.0	20.0
2	15.0	10.0	41.0
14	12.0	NAN	26.0
44	2.0	15.0	29.25

Input Value

Output Value

Test Value

Step5:- Predict the missing values of col2 using other cols

	R&D Spend	Administration	Marketing Spend
21	8.0	15.0	30.0
37	23.14	5.0	20.0
2	15.0	10.0	41.0
14	12.0	11.06	26.0
44	2.0	15.0	29.25

predicted Value

Step 6: Remove all col3 values

	R&D	Admin	MS
21	8.0	15.0	30.0
37	23.14	5.0	20.0
2	15.0	10.0	41.0
14	12.0	11.06	26.0
44	2.0	15.0	NAN

	R&D	Admin	MS
21	8.0	15.0	30.0
37	23.14	5.0	20.0
2	15.00	10.0	41.0
14	12.0	11.06	26.0
44	2.0	15.0	31.56

Predicted Value

Iteration 0 (Mean)

	R&D	Admin	MS
21	8.0	15.0	30.0
37	9.25	5.0	20.0
2	15.0	10.0	41.0
14	12.0	11.25	26.0
44	2.0	15.0	29.25

Iteration 1 (Predict)

R&D	Admin	MS
8.0	15.0	30.0
23.14	5.0	20.0
15.0	10.0	41.0
12.0	11.06	26.0
2.0	15.0	31.56

Difference

R&D	Admin	MS
0.0	0.0	0.0
13.89	0.0	0.0
0.0	0.0	0.0
0.0	0.19	0.0
0.0	0.0	2.31

Iteration 1 (Predict)

R&D	Admin	MS
8.0	15.0	30.0
23.14	5.0	20.0
15.0	10.0	41.0
12.0	11.06	26.0
2.0	15.0	31.56

Iteration 2 (New Predict)

R&D	Admin	MS
8.0	15.0	30.0
25.78	5.0	20.0
15.0	10.0	41.0
12.0	11.22	26.0
2.0	15.00	31.56

Difference

R&D	Admin	MS
0.0	0.0	0.0
0.64	0.0	0.0
0.0	0.0	0.0
0.0	0.16	0.0
0.0	0.0	0.0

* This process → continous → Difference (all 0.0)

In Iteration 2 we do same process → first Remove the value into NAN. and then predict the value and then find the difference beth Iteration 1 and Iteration 2. Until all difference be 0.0