Data Science :

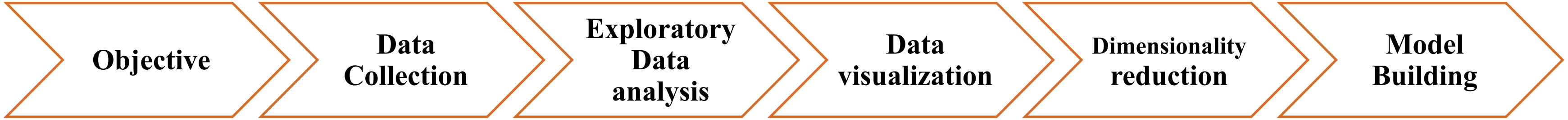
How to deal with missing data

Umang Patel

Topic

- Why – there is need of deal with missing data?
- How to deal with missing data?
  - Two Ways of handling missing data (Drop / Fill)
  - Different approaches to fill the missing data
    - Simple
    - Statistical
    - Analytical
    - Model based

Data Science Process Model



Objective

Data Collection

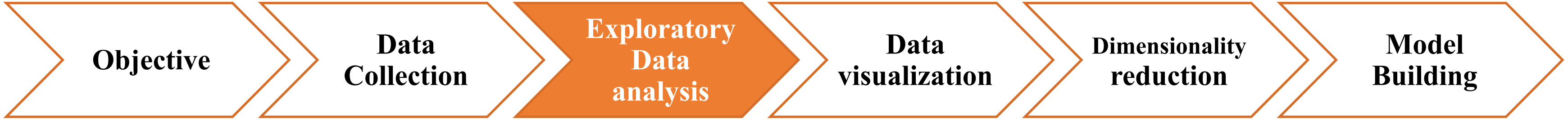
Exploratory Data analysis

Data visualization

Dimensionality reduction

Model Building

Data Science Process Model



Objective

Data Collection

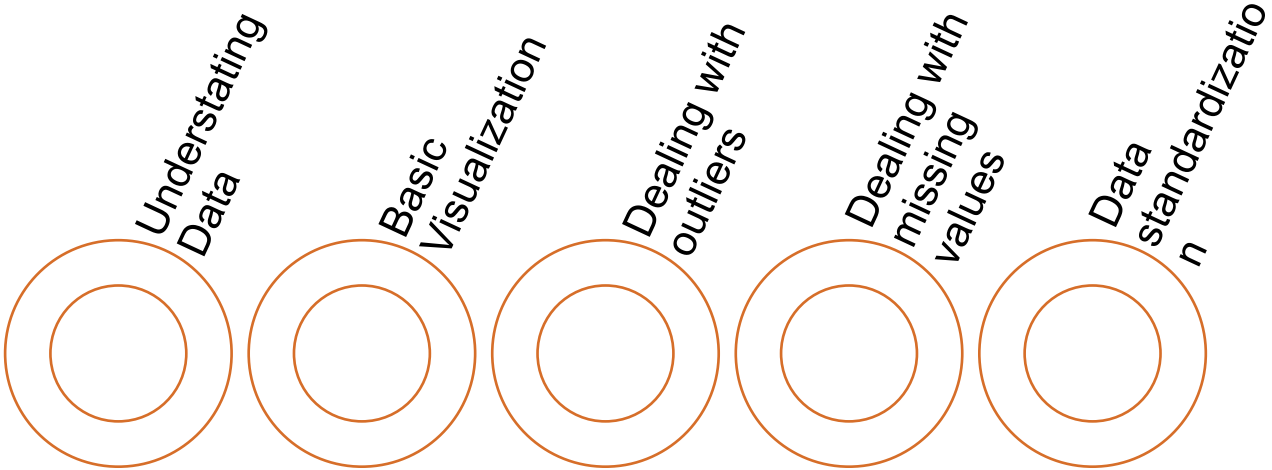
Exploratory Data analysis

Data visualization

Dimensionality reduction

Model Building

Sub-task in EDA



Understating Data

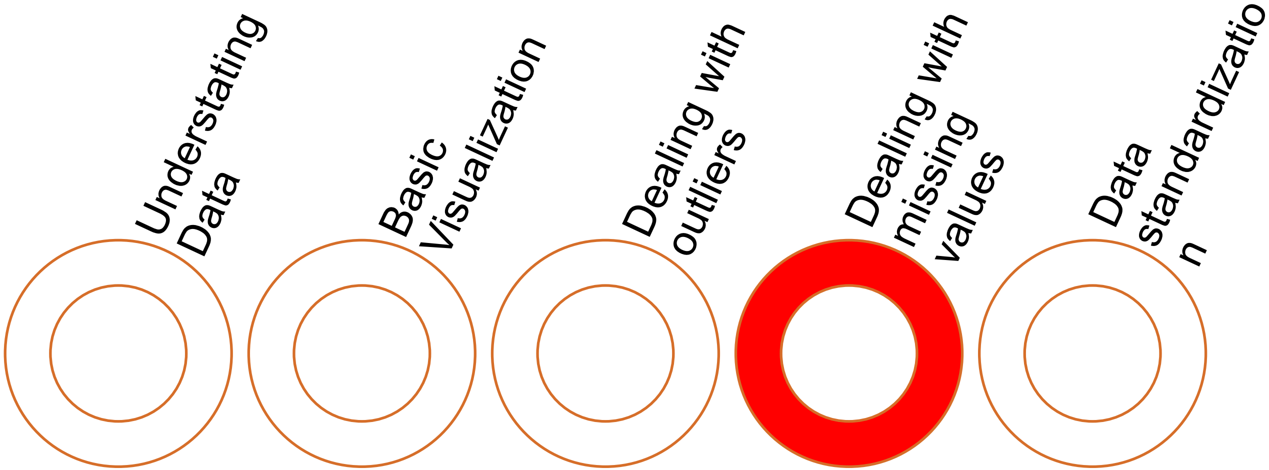
Basic Visualization

Dealing with outliers

Dealing with missing values

Data standardization

Sub-task in EDA



Why – there is need to deal with missing data?

- Improper visualization
- Can not train the model
- Can not have clear analysis of objective
- May draw wrong conclusion

How it will look?

name	gender	age	nos of call	nos of sms	bill amount	complaint
abc	M	23	450	18	518	NaN
def	M	45	200	10	250	3
ghi	NaN	60	180	52	NaN	NaN
pqr	F	NaN	NaN	NaN	400	NaN
xyz	F	18	400	40	502	3

How to deal with the missing data?

DROP
DELETE

FILL

Analysis of data

- Study about size or shape of data.
  - Rows – no of individual record
  - Columns – no of features/variables
- Application – final objective
- Analysis missing data - rows & columns wise – *df.isnull()*

Analysis of missing data - row & column wise

Data size: 5000 x 7

Feat	Name	Gender	Age	Nos of calls	Nos of SMS	Bill amount	complaint
Count	100	25	50	2	2	37	100

Decision for Drop or fill???

Record	Count
R1	3
R2	0
R3	1
R4	5
R5	2
.	
.	

Analysis missing data - row & column wise

Data size: 5000 x 6

Feat	Gender	Age	Nos of calls	Nos of SMS	Bill amount	complaint
Count	25	50	2	2	37	100

Decision for Drop or fill???

Huge data ??

For rows:
Count >= 3/4th (Nos of feat)

Record	Count
R1	3
R2	0
R3	0
R4	5
R5	2
.	
.	

Methods to fill the data

- **Structure of data:**

Features	Type	Remarks
Name	text	Already removed
Gender	binary data	male / female – (0/1)
Age	int	0 to 100
Nos of calls	int	0 to …
Nos of SMS	int	0 to …
Bill amount	float	-…. to +…
Complaint	categorical	0 -> No problem 1-> Recharge issue 2 -> problems in offer 3 -> Network problem 4 -> Any others

Method 1: Simple

- Fill with previous value - *df.fillna(method = ‘pad’)*
- Fill with next value - *df.fillna(method = ‘bfill’)*
- Fill with left side value - *df.fillna(method = ‘pad’, axis = 1)*
- Fill with right side value - *df.fillna(method = ‘bfill’, axis = 1)*
- Fill with any specific value - *df.fillna(value = 0)*
- Fill with specific value to each features - *df.fillna({ ‘age’ : 20 , ‘gender’ : ‘male’})*

Method 2: Statistical

- Fill with following statistical value (when data-type is *int* or *float*)
  - Mean
  - Median
  - Mode
  - Min
  - Max
  - Interpolate - linear

Features	Type
Age	int
Nos of calls	int
Nos of SMS	int
Bill amount	float

- When data-type – categorical or binary
  - Probability

Features	Type	Remarks
Gender	binary data	male / female – (0/1)
Complaint	categorical	0 -> No problem 1-> Recharge issue 2 -> problems in offer 3 -> Network problem 4 -> Any others

Method 3: Analytical

- Need to understand and analysis dependencies with other columns or rows

- Examples:
  - **Bill amount** can depends on **Nos of call** and **Nos of SMS**
  - **Age** cab depends on **Bill amount**

Method 4: Model based

F1	F2	F3	F4	F5	
NaN	M	25	35	0	□Test data
4	F	22	34	0	
2	M	44	30	1	
4	F	32	28	0	
NaN	F	25	26	1	□Test data

Output data

Input data

Conclusion

- What is need of deal with missing values
- What to choose? DROP / FILL
- Different methods to FILL these values
- How to choose these methods
- Drawback

Any Questions?