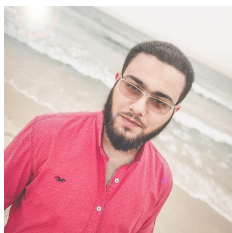




Developer Student Club  
HTI October Branch

# Data Preprocessing



**Mohammed El-Agha**  
DSC Data Science Instructor & Mentor



# Data Preprocessing

- **Data Preprocessing** is preparation of data to be ready for the processing itself.
  - Change format
  - Change datatype
  - Resolve a problem
  - Change in data itself
  - Add something
  - remove/reduce something

# Some Issues

- Missing
- Error in data
- Unreasonable data
- Huge amount of data
- Very high dimensionality
- Very wide range of numeric data
- Unneeded data
- Unneeded features

# Data Preprocessing in DS Model

- Collection -> Understanding -> Preprocessing -> Modeling
- Garbage in = Garbage out
- It take 40-80% of project time
- Complete preprocessing may be pusher of model accuracy

# Data Preprocessing

## Data Cleaning

### Missing Data

- 1.Ignore The Tuple
- 2.Fill The Missing Values(manually,by mean or by most probable value)

### Noisy Data

- 1.Binning Method
- 2.Regression
- 3.Clustering

## Data Transformation

### Normalization

### Attribute Selection

### Discretization

### Concept Hiererchy Generation

## Data Reduction

### Data Cube Aggregation

### Attribute Subset Selection

### Numerosity Reduction

### Dimensionality Reduction



# Data Preprocessing Types

- Data Cleaning
- Data Transformation
- Data Reduction

# Data Preprocessing Types

- **Data Cleaning**
  - Related to fixing problem/error in the data
  - Like: missing, errors, faults, noisy, unreasonable

# Data Preprocessing Types

- **Data Transformation**
- Related to
  - Data be appropriate for DS task
  - Best formatting
  - May used for enhancing performance (accuracy)
- Like: datatype conversion, change numeric range



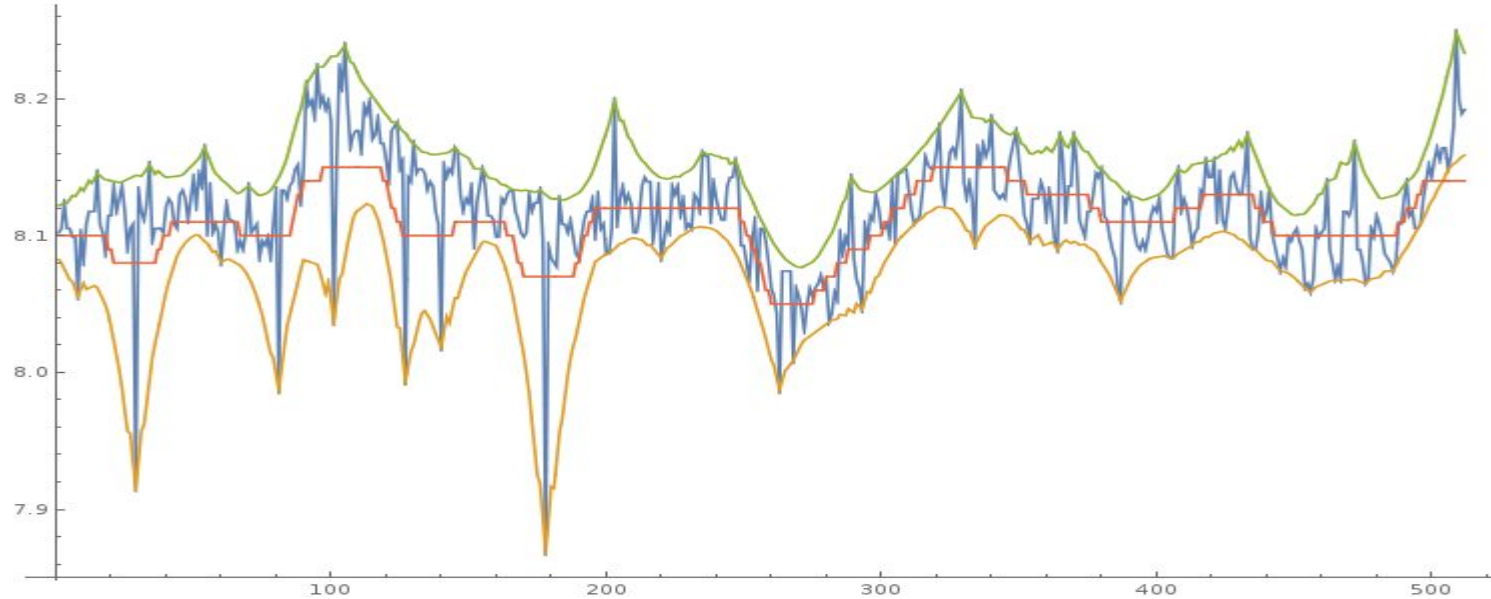
# Data Preprocessing Types

- **Data Reduction**
- Related to
  - Reduction of data size, on row or column level
  - May be use to improve efficiency (time or space performance)
- Like: sampling, feature reduction

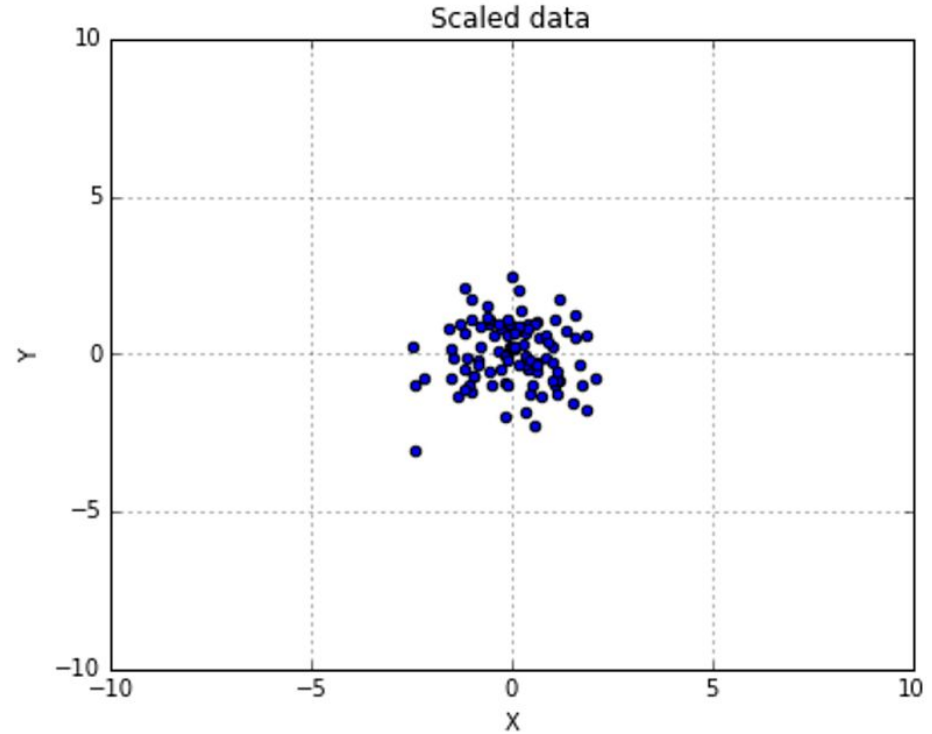
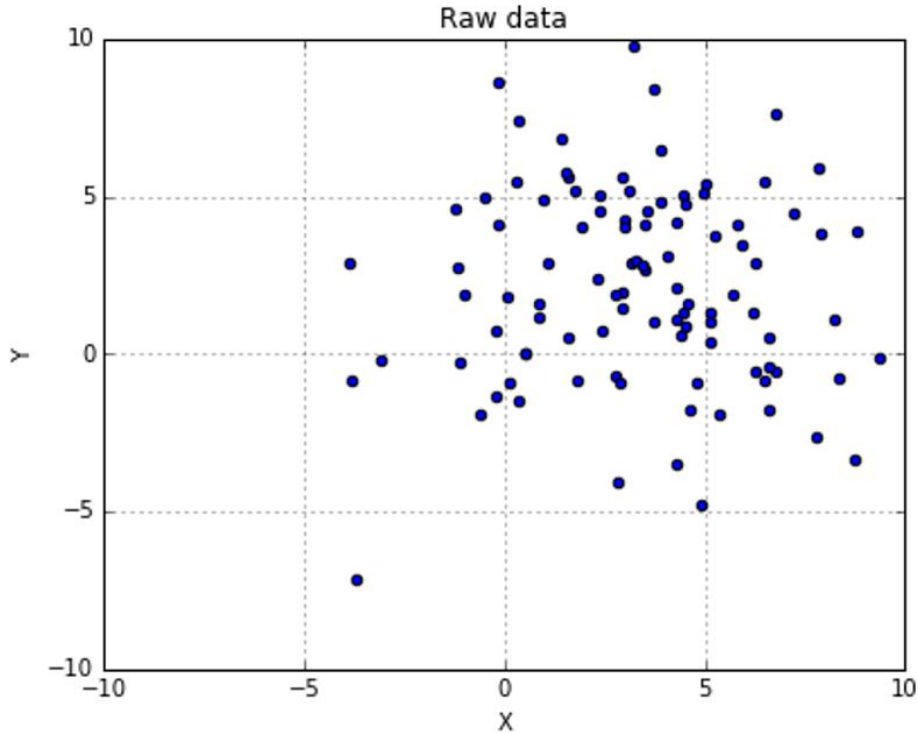
# Example - Missing Data

ID	Color	Weight	Broken	Class
1	Black	80	Yes	1
2	Yellow	100	No	2
3	Yellow	120	Yes	2
4	Blue	90	No	2
5	Blue	85	No	2
<b>6</b>	<b>?</b>	60	No	1
<b>7</b>	Yellow	100	<b>?</b>	2
<b>8</b>	<b>?</b>	40	<b>?</b>	1

# Example - Noise Data



# Example - Normalization



# Data Preprocessing

- **Good Resources**

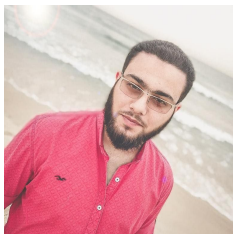
- (Book) Feature Engineering and Selection: A Practical Approach for Predictive Models (Chapman & Hall/CRC Data Science Series)
- (Webpage)

<https://www.geeksforgeeks.org/data-preprocessing-in-data-mining>

# Any Questions



# Thank You!



**Mohammed El-Agha**  
**DSC Data Science**  
**Instructor & Mentor**

