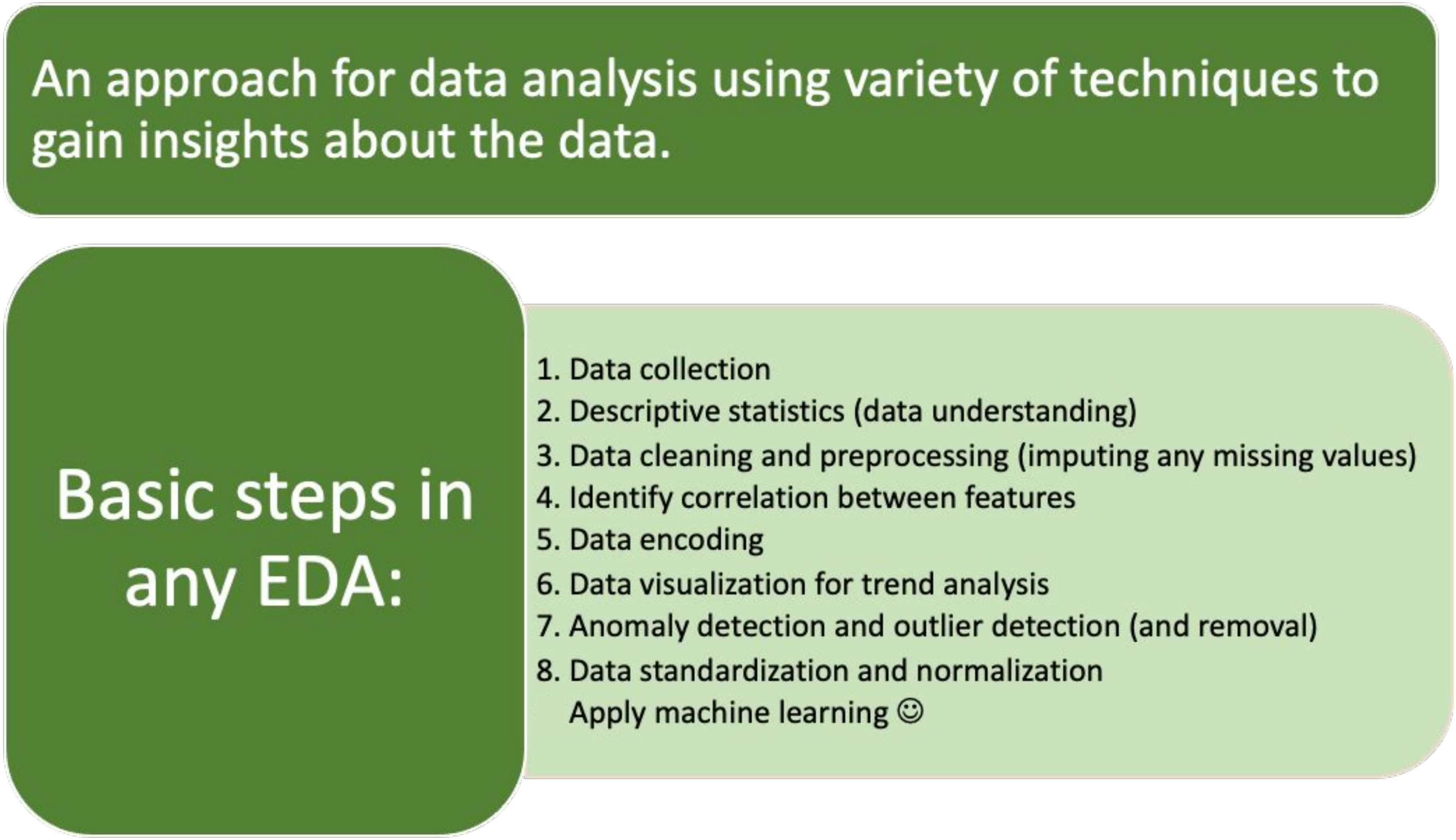
CO 544
Machine Learning and Data Mining

Lab 02 - Exploratory Data Analysis and Visualization

Exploratory Data Analysis (EDA)



An approach for data analysis using variety of techniques to gain insights about the data.

Basic steps in any EDA:

1. Data collection
2. Descriptive statistics (data understanding)
3. Data cleaning and preprocessing (imputing any missing values)
4. Identify correlation between features
5. Data encoding
6. Data visualization for trend analysis
7. Anomaly detection and outlier detection (and removal)
8. Data standardization and normalization

Apply machine learning ☺

Descriptive statistics

- Used to make preliminary assessments about the population distribution of the variable.
- Commonly used statistics:

1. **Central tendency**
   a. Mean – the average value of all the data points.
   b. Median – the middle value when all the data points are put in an ordered list
   c. Mode – the data point which occurs the most in the dataset

2. **Spread**: the measure of how far the data points are away from the mean or median
   a. Variance - the mean of the squares of the individual deviations.
   b. Standard deviation - the square root of the variance.

3. **Skewness**: a measure of asymmetry

Descriptive statistics (cont.)

Quick look on data:

- **Describe()**: summarizes the central tendency, dispersion, and shape of a dataset’s distribution, excluding NaN values.
    - syntax: pandas.dataframe.describe()
- **Info()**: prints a concise summary of the dataframe. This method prints information about a dataframe including the index dtype and columns, non-null values and memory usage.
    - syntax: pandas.dataframe.info()
- **Pandas profiling**

Null and missing values

Detecting

- Detecting Null-values:
  - Isnull(): It is used as an alias for dataframe.isna(). This function returns the dataframe with boolean values indicating missing values.
  - Syntax : dataframe.isnull()

Handling

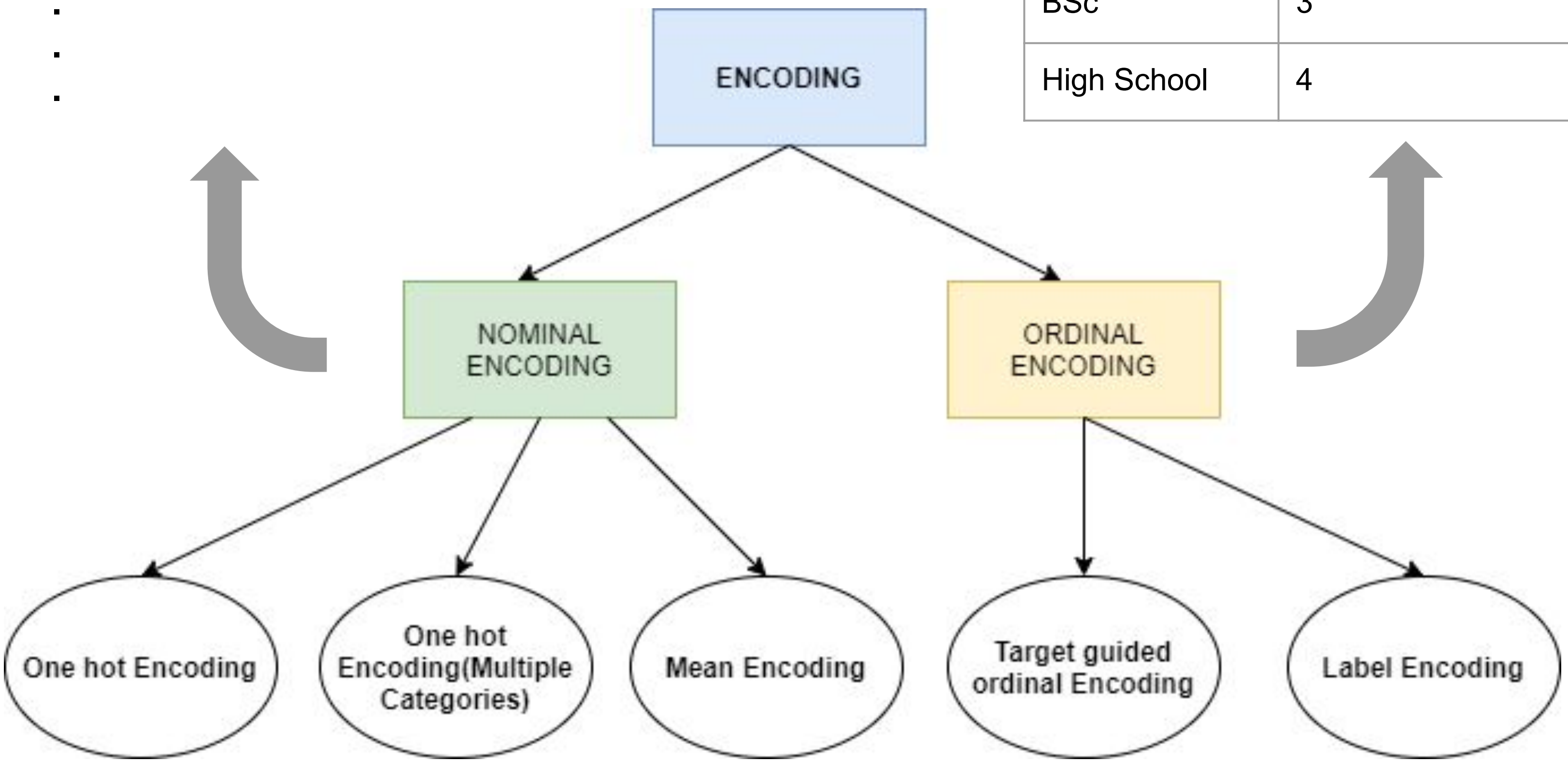
- Handling null values:
  - Dropping the rows with null values: dropna() function is used to delete rows or columns with null values.
  - Replacing missing values: fillna() function can fill the missing values with a special value value like mean or median.

Encoding

Color	Color (encoded)
Red	2
Yellow	3

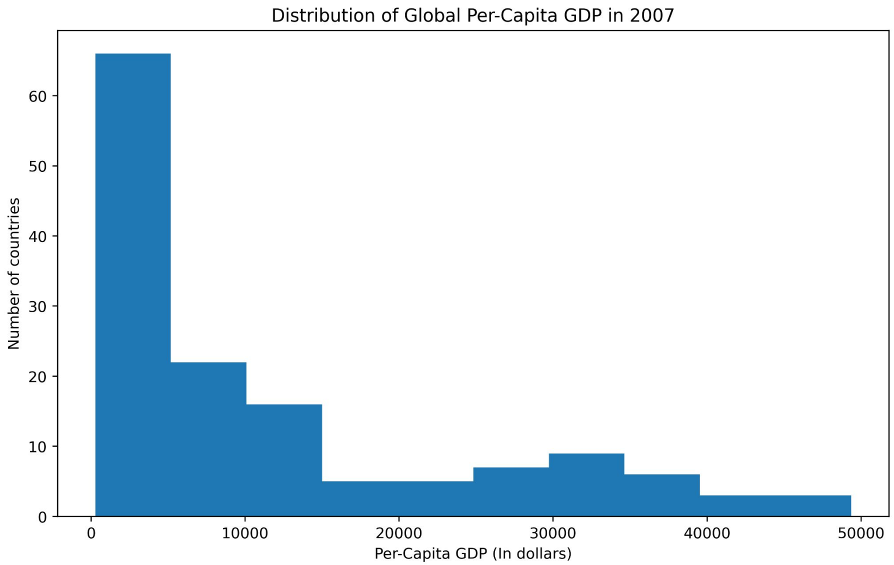
⋮

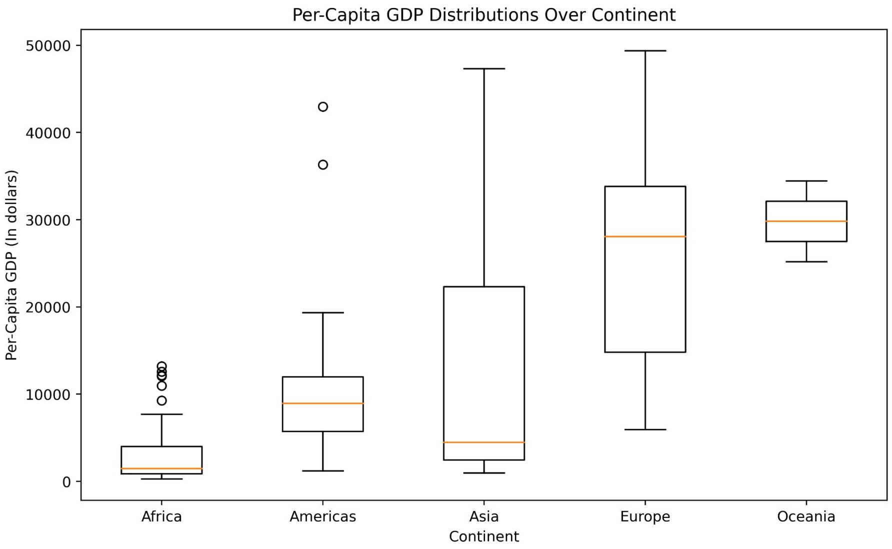
Education	Education (encoded)
PhD	1
MSc	2
BSc	3
High School	4



Visualization

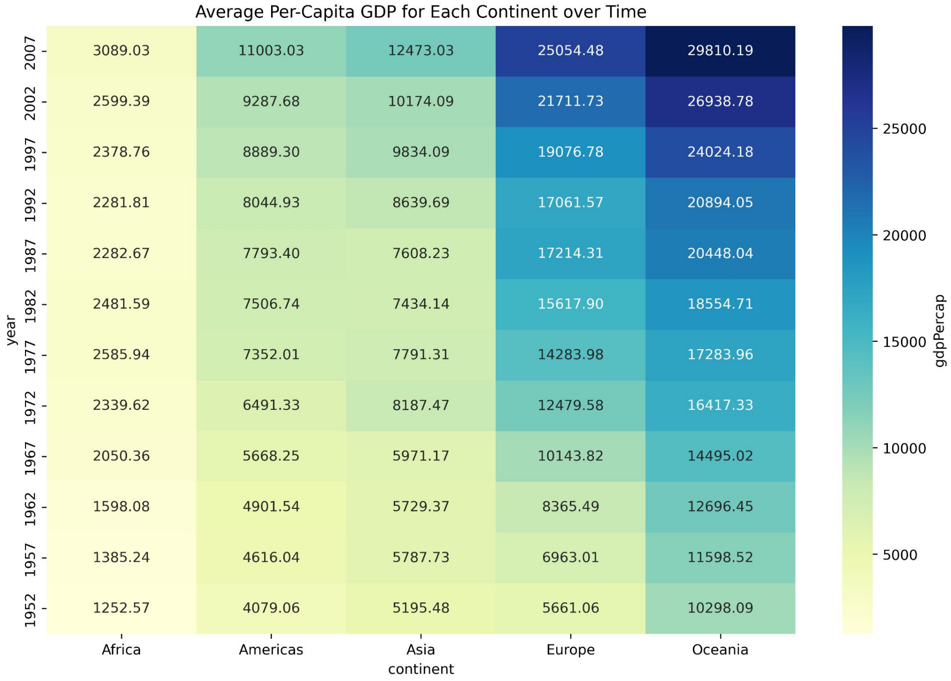
- **Univariate**: looking at one variable/column at a time





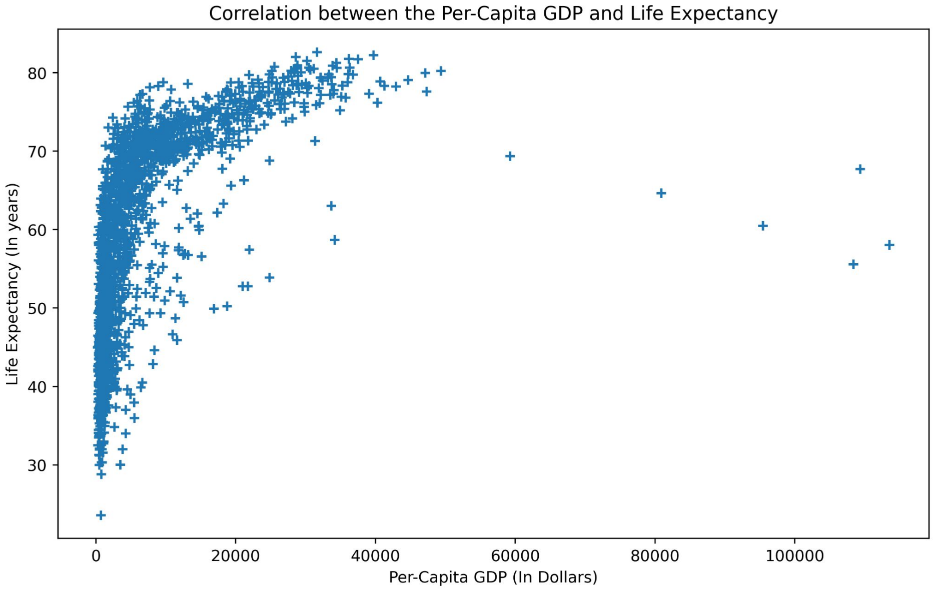
Visualization (cont.)

- **Multivariate:** looking at relationship between two or more variables



Average Per-Capita GDP for Each Continent over Time

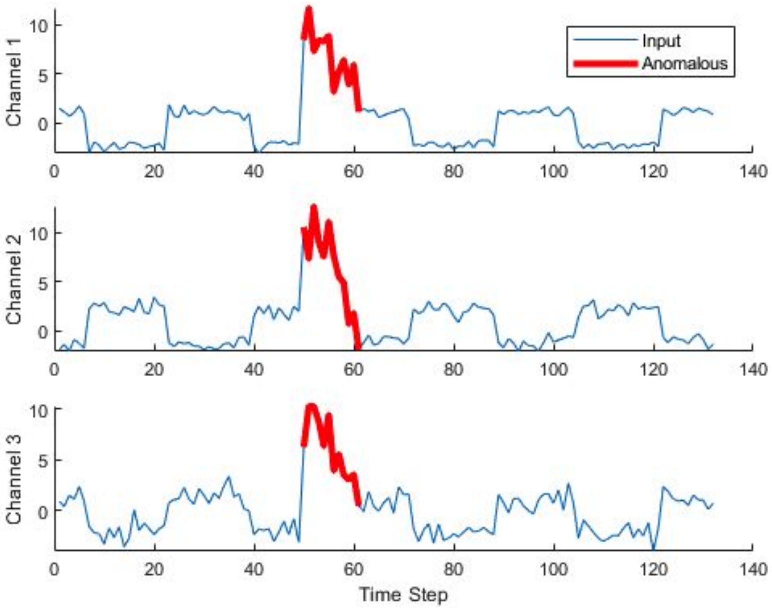
year	Africa	Americas	Asia	Europe	Oceania
2007	3089.03	11003.03	12473.03	25054.48	29810.19
2002	2599.39	9287.68	10174.09	21711.73	26938.78
1997	2378.76	8889.30	9834.09	19076.78	24024.18
1992	2281.81	8044.93	8639.69	17061.57	20894.05
1987	2282.67	7793.40	7608.23	17214.31	20448.04
1982	2481.59	7506.74	7434.14	15617.90	18554.71
1977	2585.94	7352.01	7791.31	14283.98	17283.96
1972	2339.62	6491.33	8187.47	12479.58	16417.33
1967	2050.36	5668.25	5971.17	10143.82	14495.02
1962	1598.08	4901.54	5729.37	8365.49	12696.45
1957	1385.24	4616.04	5787.73	6963.01	11598.52
1952	1252.57	4079.06	5195.48	5661.06	10298.09



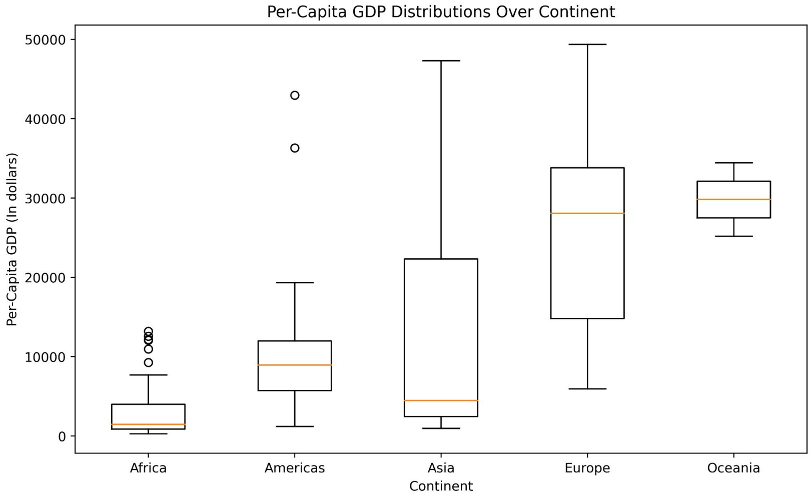
Correlation between the Per-Capita GDP and Life Expectancy

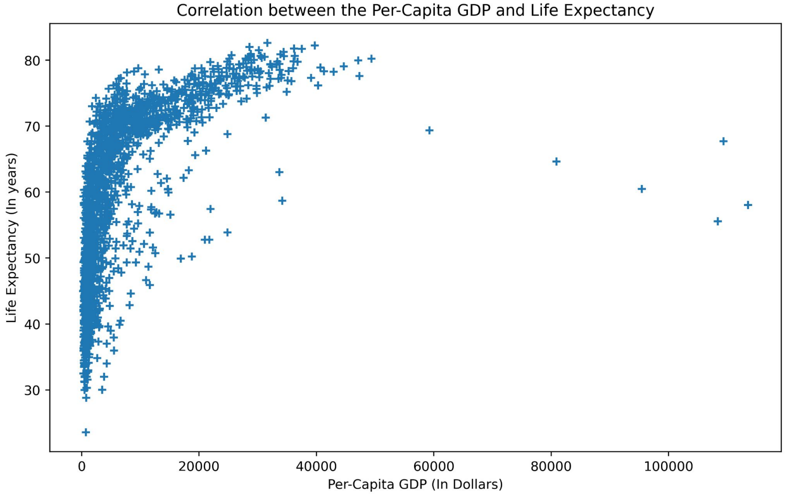
Anomaly / Outlier detection

- Identification of unexpected events, observations, or items that differ significantly from the norm.

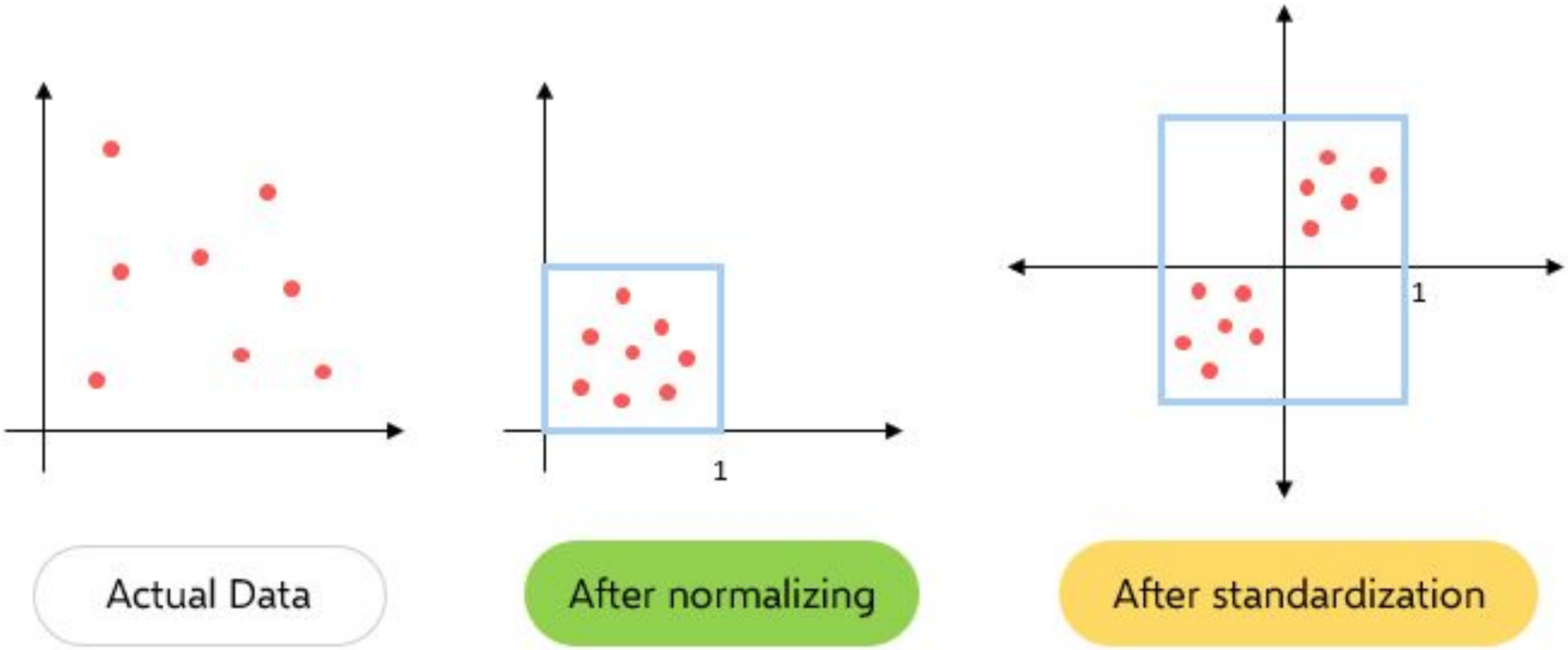


Source: Time Series Anomaly Detection Using Deep Learning (https://www.mathworks.com/help/deeplearning/ug/time-series-anomaly-detection-using-deep-learning.html)





Normalization and Standardization



Source: https://www.someka.net/blog/how-to-normalize-data-in-excel/

Transforms the original values to fit within a certain range, standardization

Transforms them to fit within a distribution that has a mean of 0 and standard deviation of 1 (a.k.a. Mean Centering)

Data repositories

- Google’s Datasets Search Engine (https://toolbox.google.com/datasetsearch)
- .gov Datasets:
  - Indian Government Dataset (https://data.gov.in/)
  - Australian Government Dataset (https://data.gov.au/)
  - EU Open Data Portal (http://data.europa.eu/euodp/en/data/)
  - New Zealand’s Government Dataset (https://data.govt.nz/)
- Kaggle Datasets (https://www.kaggle.com/datasets)
- Amazon Datasets: Registry of Open Data on AWS (https://registry.opendata.aws/)
- UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php)
- Yahoo WebScope (https://webscope.sandbox.yahoo.com/?guccounter=1)
- Datasets subreddit (https://www.reddit.com/r/datasets/)