Introduction to Data Science with Python

WMASDS04

Week 03: Descriptive Statistics for EDA

Descriptive Statistics

- Descriptive statistics
  - helps to simplify large amounts of data in a sensible way.
  - It is simply a way to describe the data.
  - Claim to present quantitative descriptions of it in a manageable form.
  - Statistics, and in particular descriptive statistics, is based on two main concepts:
    - A **population** is a collection of objects, items ("units") about which information is sought.
    - A **sample** is a part of the population that is observed.
- In contrast, in inferential statistics,
  - we draw conclusions beyond the data we are analyzing; we reach a conclusions regarding hypotheses we may make.
  - We try to infer characteristics of the "population" of the data,

Outlines

- Descriptive Statistics
- Exploratory Data Analysis
- Data Preprocessing
- Data Visualization

- Descriptive statistics applies the concepts, measures, and terms that are used to describe the basic features of the samples in a study.
- These procedures are essential to provide summaries about the samples as an approximation of the population.
- Together with simple graphics, they form the basis of every quantitative analysis of data.
- In order to describe the sample data and to be able to infer any conclusion, we should go through several steps:
  - *Data preparation*: Given a specific dataset, we need to prepare the data for generating statistically valid descriptions.
  - *Descriptive statistics*: This generates different statistics to describe and summarize the data concisely and evaluate different ways to visualize them.

Data Preparation

- The most common steps for data preparation involve the following operations.
  - **Obtaining the data:** Data can be read directly from a file or they might be obtained by scraping the web.
  - **Parsing the data:** The right parsing procedure depends on what format the data are in: plain text, fixed columns, CSV, XML, HTML, etc.
  - **Cleaning the data:** Survey responses and other data files are almost always incomplete. Sometimes, there are multiple codes for things such as, not asked, did not know, and declined to answer. And there are almost always errors. A simple strategy is to remove or ignore incomplete records.
  - **Building data structures:** Once you read the data, it is necessary to store them in a data structure that lends itself to the analysis we are interested in.

Exploratory Data Analysis

- One of the main goals of exploratory data analysis is to visualize and summarize the sample distribution, thereby allowing us to make tentative assumptions about the population distribution.
- The data that come from performing a particular measurement on all the subjects in a sample represent our observations for a single characteristic like country, age, education, etc.
- These measurements and categories represent a *sample distribution* of the variable, which in turn approximately represents the *population distribution* of the variable.

Summarizing the Data

- For categorical data, a simple tabulation of the frequency of each category is the best non-graphical exploration for data analysis.
  - For example, we can ask ourselves what is the proportion of high income professionals in our database
- Given a quantitative variable, exploratory data analysis is a way to make preliminary assessments about the population distribution of the variable using the data of the observed samples.
- The characteristics of the population distribution of a quantitative variable are its *mean*, *deviation*, *histograms*, *outliers*, etc.

Mean

One of the first measurements we use to have a look at the data is to obtain *sample statistics* from the data, such as the sample mean [1]. Given a sample of n values, $\{x_i\}$, $i = 1, \ldots, n$, the *mean*, μ, is the sum of the values divided by the number of values,[2] in other words:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \,. \tag{3.1}$$

The terms mean and *average* are often used interchangeably. In fact, the main distinction between them is that the mean of a sample is the summary statistic computed by Eq. (3.1), while an average is not strictly defined and could be one of many summary statistics that can be chosen to describe the central tendency of a sample.

Variance

The mean is not usually a sufficient descriptor of the data. We can go further by knowing two numbers: mean and *variance*. The variance σ^2 describes the spread of the data and it is defined as follows:

$$\sigma^2 = \frac{1}{n}\sum_i (x_i - \mu)^2. \tag{3.2}$$

The term $(x_i - \mu)$ is called the *deviation* from the mean, so the variance is the mean squared deviation. The square root of the variance, σ, is called the *standard deviation*. We consider the standard deviation, because the variance is hard to interpret (e.g., if the units are grams, the variance is in grams squared).

Median

- The mean of the samples is a good descriptor, but it has an important drawback: what will happen if in the sample set there is an error with a value very different from the rest?
- For example, considering hours worked per week, it would normally be in a range between 20 and 80; but what would happen if by mistake there was a value of 1000? An item of data that is significantly different from the rest of the data is called an *outlier*.
- In this case, the mean, μ, will be drastically changed towards the outlier.
- One solution to this drawback is offered by the statistical *median*, μ_{12}, which is an order statistic giving the middle value of a sample.
- In this case, all the values are ordered by their magnitude and the median is defined as the value that is in the middle of the ordered list.
- Hence, it is a value that is much more robust in the face of outliers.

Quantiles and Percentiles

- a fraction p of the data values is less than or equal to x_p and
- the remaining fraction $(1 - p)$ is greater than x_p.

That value, x_p, is the p-*th* quantile, or the $100 \times p$-*th* percentile. For example, a 5-number summary is defined by the values x_{min}, Q_1, Q_2, Q_3, x_{max}, where Q_1 is the $25 \times p$-*th* percentile, Q_2 is the $50 \times p$-*th* percentile and Q_3 is the $75 \times p$-*th* percentile.

Data Distributions

- Summarizing data by just looking at their mean, median, and variance can be dangerous: **very different data can be described by the same statistics**.
- The best thing to do is to validate the data by inspecting them. We can have a look at the data distribution, which describes how often each value appears (i.e., what is its frequency).
- The most common representation of a distribution is a histogram, which is a graph that shows the frequency of each value.
- We can normalize the frequencies of the histogram by dividing/normalizing by n, the number of samples. The normalized histogram is called the Probability Mass Function (PMF).
- The *Cumulative Distribution Function* (CDF), or just distribution function, describes the probability that a real-valued random variable *X* with a given probability distribution will be found to have a value less than or equal to *x*.

Outlier Treatment

- Outliers are data samples with a value that is far from the central tendency.
- Different rules can be defined to detect outliers, as follows:
  - Computing samples that are far from the median.
  - Computing samples whose values exceed the mean by 2 or 3 standard deviations.

3.3.4 Measuring Asymmetry: Skewness and Pearson's Median Skewness Coefficient

For univariate data, the formula for *skewness* is a statistic that measures the asymmetry of the set of n data samples, x_i:

$$g_1 = \frac{1}{n} \frac{\sum_i (x_i - \mu^3)}{\sigma^3}, \tag{3.3}$$

where μ is the mean, σ is the standard deviation, and n is the number of data points.

Negative deviation indicates that the distribution "skews left" (it extends further to the left than to the right). One can easily see that the skewness for a normal distribution is zero, and any symmetric data must have a skewness of zero. Note that skewness can be affected by outliers! A simpler alternative is to look at the relationship between the mean μ and the median μ_{12}.

Continuous Distribution

- The distributions we have considered up to now are based on empirical observations and thus are called *empirical distributions*.
- As an alternative, we may be interested in considering distributions that are defined by a continuous function and are called *continuous distributions*.
- Remember that we defined the PMF, *fX (x)*, of a discrete random variable *X* as *fX (x) = P(X = x)* for all *x*.
- In the case of a continuous random variable *X*, we speak of the *Probability Density Function* (PDF), which

is defined as $F_X(x)$ where this satisfies: $F_X(x) = \int_{\infty}^{x} f_X(t)\delta t$ for all x. There are many continuous distributions; here, we will consider the most common ones: the

- exponential and the normal distributions.

3.3.5.1 The Exponential Distribution

Exponential distributions are well known since they describe the inter-arrival time between events. When the events are equally likely to occur at any time, the distribution of the inter-arrival time tends to an exponential distribution. The CDF and the PDF of the exponential distribution are defined by the following equations:

$$CDF(x) = 1 - e^{-\lambda x}, \quad PDF(x) = \lambda e^{-\lambda x}.$$

The parameter λ defines the shape of the distribution. An example is given in Fig. 3.6. It is easy to show that the mean of the distribution is $\frac{1}{\lambda}$, the variance is $\frac{1}{\lambda^2}$ and the median is $\frac{\ln(2)}{\lambda}$.

Note that for a small number of samples, it is difficult to see that the exact empirical distribution fits a continuous distribution. The best way to observe this match is to generate samples from the continuous distribution and see if these samples match the data. As an exercise, you can consider the birthdays of a large enough group of people, sorting them and computing the inter-arrival time in days. If you plot the CDF of the inter-arrival times, you will observe the exponential distribution.

3.3.5.2 The Normal Distribution

The *normal distribution*, also called the *Gaussian distribution*, is the most common since it represents many real phenomena: economic, natural, social, and others. Some well-known examples of real phenomena with a normal distribution are as follows:

- The size of living tissue (length, height, weight).
- The length of inert appendages (hair, nails, teeth) of biological specimens.
- Different physiological measurements (e.g., blood pressure), etc.

The normal CDF has no closed-form expression and its most common representation is the PDF:

$$\mathrm{PD}F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Kernel Density

- In many real problems, we may **not be interested in the parameters of a particular distribution of data, but just a continuous representation of the data**.
- In this case, we should estimate the distribution non-parametrically (i.e., making no assumptions about the form of the underlying distribution) using kernel density estimation.
- Let us imagine that we have a set of data measurements without knowing their distribution and we need to estimate the continuous representation of their distribution.
- In this case, we can consider a Gaussian kernel to generate the density around the data.

Estimation

- An important aspect when working with statistical data is being able to use estimates to approximate the values of unknown parameters of the dataset.
- In this section, we will review different kinds of estimators (estimated mean, variance, standard score, etc.).

3.4.1.3 Standard Score

In many real problems, when we want to compare data, or estimate their correlations or some other kind of relations, we must avoid data that come in different units. For example, weight can come in kilograms or grams. Even data that come in the same units can still belong to different distributions. We need to normalize them to standard scores. Given a dataset as a series of values, $\{x_i\}$, we convert the data to standard scores by subtracting the mean and dividing them by the standard deviation:

$$z_i = \frac{(x_i - \mu)}{\sigma}.$$

Note that this measure is dimensionless and its distribution has a mean of 0 and variance of 1. It inherits the "shape" of the dataset: if X is normally distributed, so is Z; if X is skewed, so is Z.

3.4.2.1 Covariance

When two variables share the same tendency, we speak about *covariance*. Let us consider two series, $\{x_i\}$ and $\{y_i\}$. Let us center the data with respect to their mean: $dx_i = x_i - \mu_X$ and $dy_i = y_i - \mu_Y$. It is easy to show that when $\{x_i\}$ and $\{y_i\}$ vary together, their deviations tend to have the same sign. The covariance is defined as the mean of the following products:

$$Cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} dx_i dy_i,$$

where n is the length of both sets. Still, the covariance itself is hard to interpret.

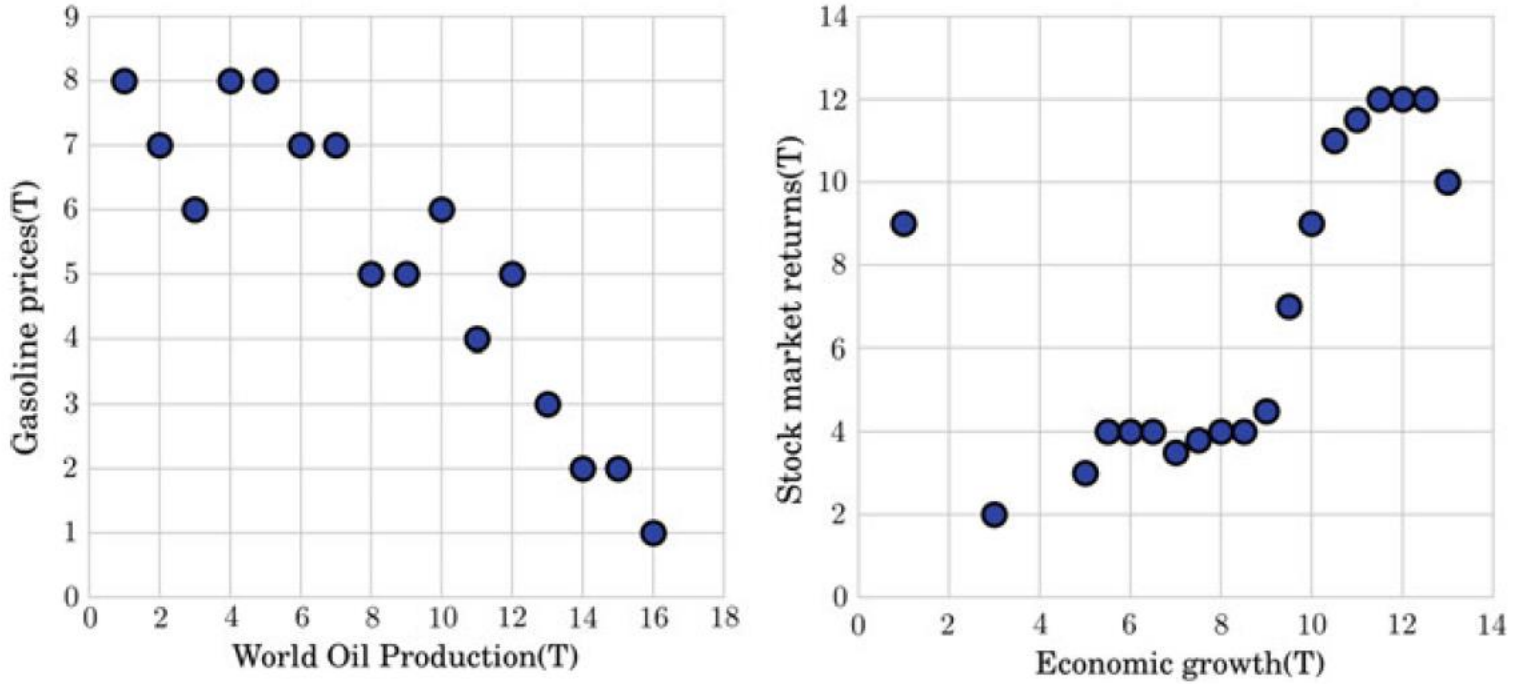


Fig. 3.9 Positive correlation between economic growth and stock market returns worldwide (*left*). Negative correlation between the world oil production and gasoline prices worldwide (*right*)

3.4.2.2 Correlation and the Pearson's Correlation

If we normalize the data with respect to their deviation, that leads to the standard scores; and then multiplying them, we get:

$$\rho_i = \frac{x_i - \mu_X}{\sigma_X} \frac{y_i - \mu_Y}{\sigma_Y}.$$

The mean of this product is $\rho = \frac{1}{n} \sum_{i=1}^{n} \rho_i$. Equivalently, we can rewrite ρ in terms of the covariance, and thus obtain the *Pearson's correlation*:

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}.$$

Note that the Pearson's correlation is always between -1 and $+1$, where the magnitude depends on the degree of correlation. If the Pearson's correlation is 1 (or -1), it means that the variables are perfectly correlated (positively or negatively)

- having $\rho = 0$, does not necessarily mean that the variables are not correlated!
- Pearson's correlation captures correlations of first order, but not nonlinear correlations.
- Moreover, it does not work well in the presence of outliers.

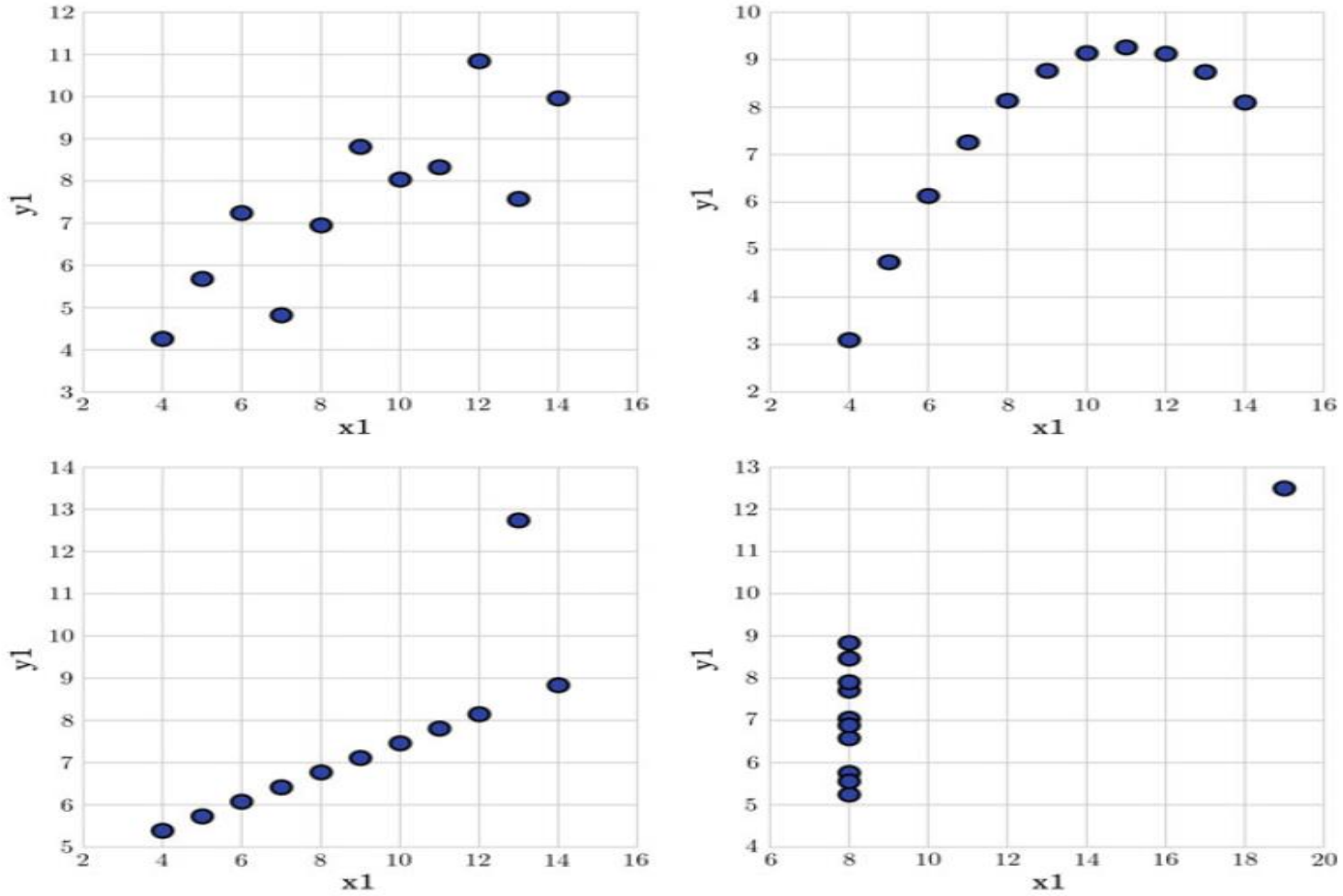


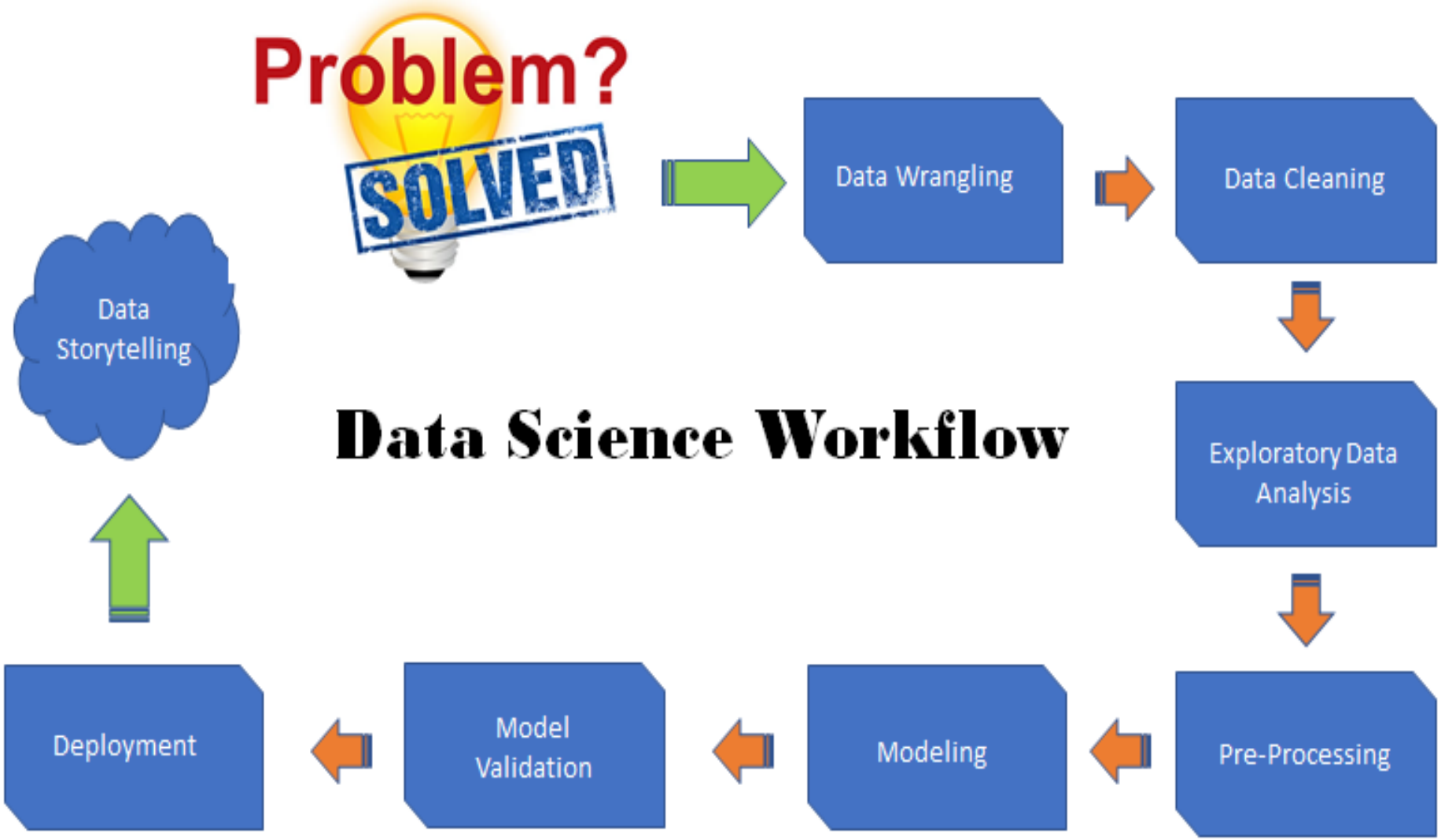
Fig. 3.10 Anscombe configurations

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3		x4	y4
1	10	8.04		10	9.14		10	7.46		8	6.58
2	8	6.95		8	8.14		8	6.77		8	5.76
3	13	7.58		13	8.74		13	12.74		8	7.71
4	9	8.81		9	8.77		9	7.11		8	8.84
5	11	8.33		11	9.26		11	7.81		8	8.47
6	14	9.96		14	8.1		14	8.84		8	7.04
7	6	7.24		6	6.13		6	6.08		8	5.25
8	4	4.26		4	3.1		4	5.39		19	12.5
9	12	10.84		12	9.13		12	8.15		8	5.56
10	7	4.82		7	7.26		7	6.42		8	7.91
11	5	5.68		5	4.74		5	5.73		8	6.89
				Summary Statistics							
N	11	11		11	11		11	11		11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50		9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94		3.16	1.94
r	0.82			0.82			0.82			0.82	

descriptive statistics to explore a dataset

- the central measures of tendency such as the sample mean and median; and measures of variability such as the variance and standard deviation. these measures can be affected by outliers.
- visualizing the dataset, histograms, quantiles, and percentiles.
- when the values are continuous variables, it is convenient to use continuous distributions; the most common of which are the normal and the exponential distributions.
- The advantage of most continuous distributions is that we can have an explicit expression for their PDF and CDF, as well as the mean and variance in terms of a closed formula.
- by using the kernel density, we can obtain a continuous representation of the sample distribution.
- estimate the correlation and the covariance of datasets, where two of the most popular measures are the Pearson's correlations, which are affected in different ways by the outliers of the dataset.

EDA in Data Science Process



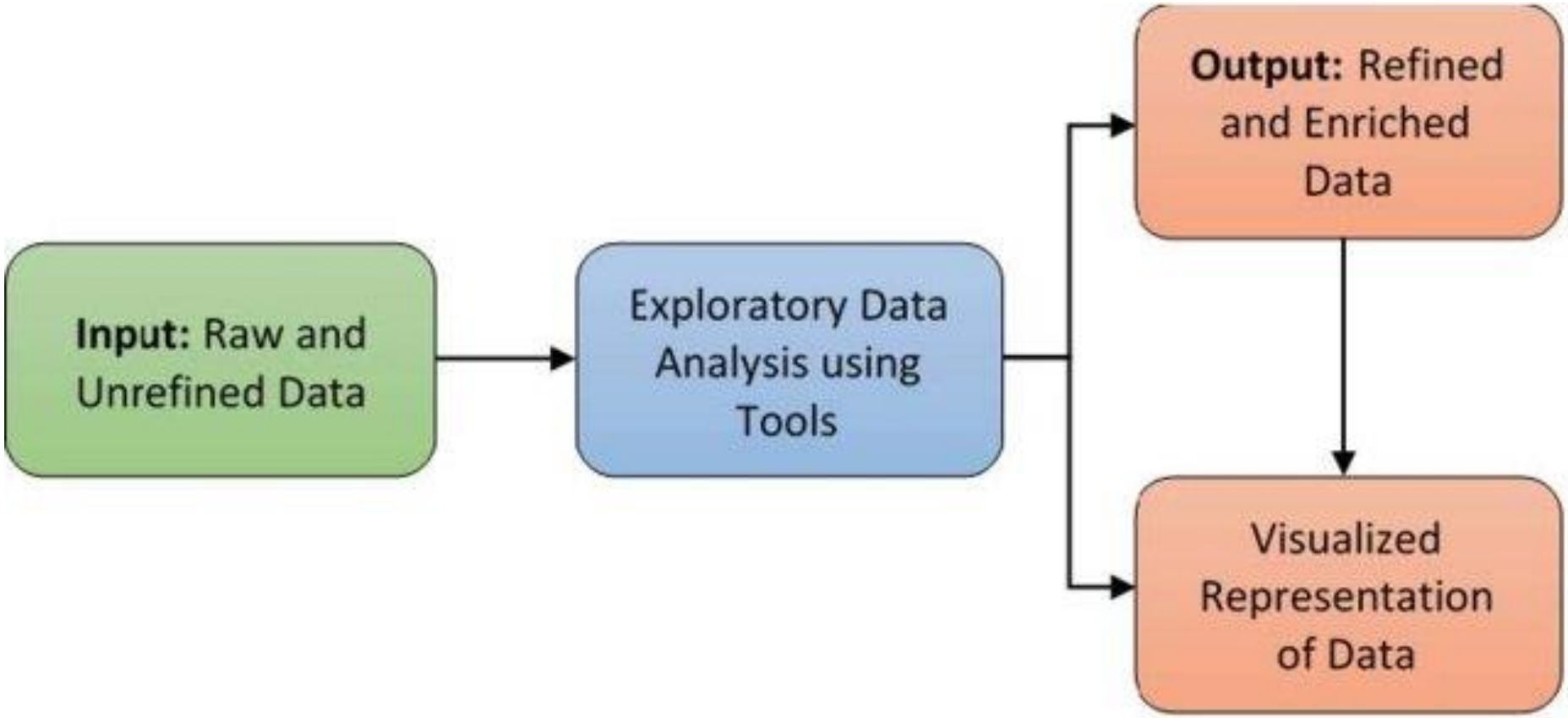
What is Exploratory Data Analysis

- It involves getting a clear understanding of the data and summarizing the data through visuals and summaries.
- an approach to analyze and investigate data sets and to discover trends, patterns, or check assumptions in data with the help of statistical summaries and graphical representations.
- used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them.
- focuses on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed.
- techniques used for extracting vital features and trends used by machine learning and deep learning models in Data Science.

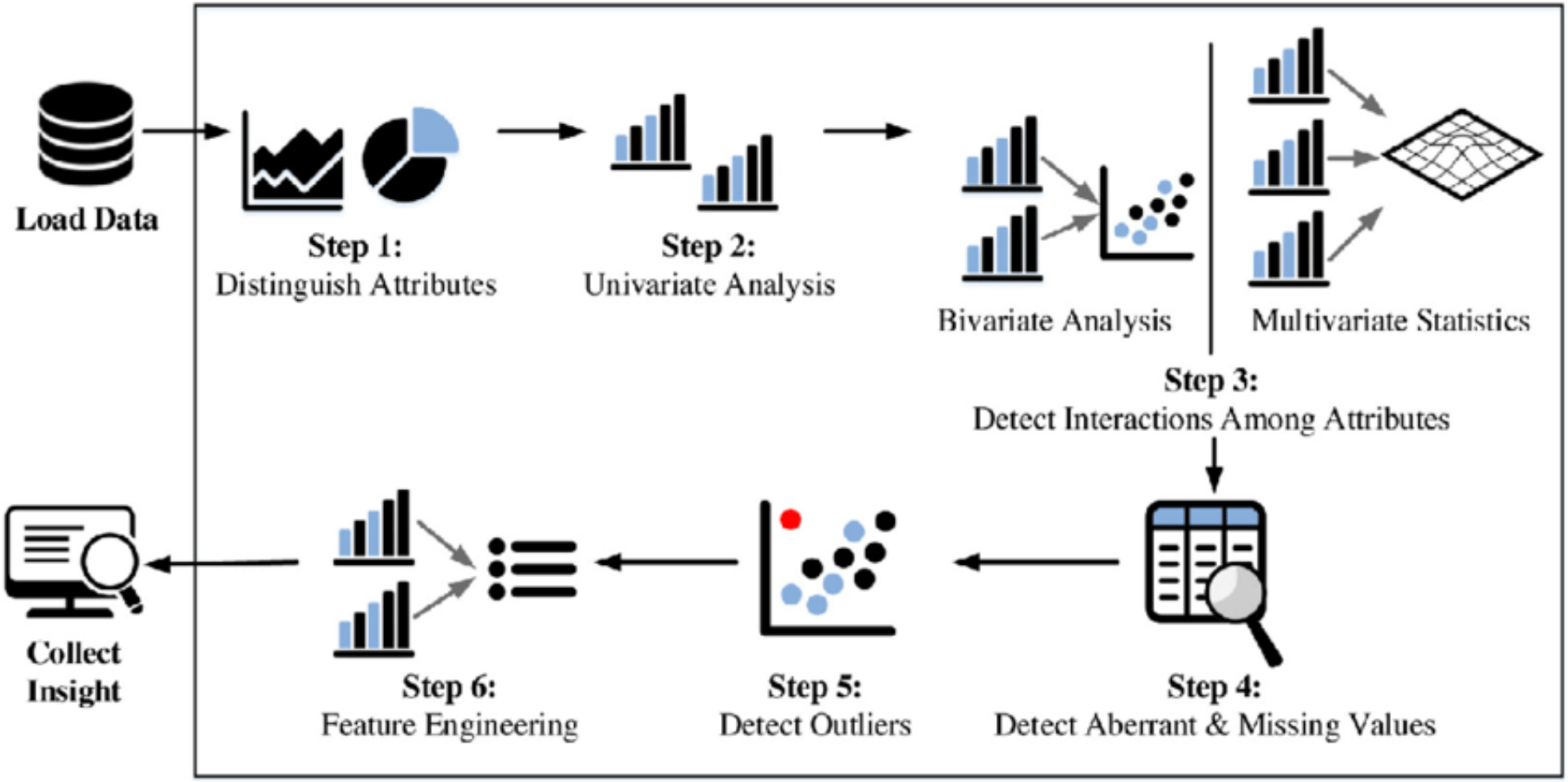
Objectives of EDA

- Develop an understanding of the data and become familiar with data.
- Identify trend and patterns.
- Understand relationship between variables.
- decide future course of action
- Decide on the appropriate models to be executed on the data.
- Find answers to questions relating to the data.
- Test assumptions.

EDA Process


Input: Raw and Unrefined Data
Exploratory Data Analysis using Tools
Output: Refined and Enriched Data
Visualized Representation of Data

What to do in EDA



EDA Tasks

- Data Collection
- Finding all Variables and Understanding Them
- Cleaning the Dataset
- Identify Correlated Variables
- Choosing the Appropriate Statistical Methods
- Visualizing and Analyzing Results

EDA using Descriptive Statistics

EDA Techniques

Objectives	EDA Techniques
Getting an idea of the distribution of a variable	Histogram
Finding Outlier	scatterplots, box-and-whisker plots
Quantify the relationship between two variables (exposure-outcome)	2D scatter plot, Covariance and correlation
Visualize the relationship between two exposure variables and one outcome	Heatmap
Visualize high dimensional data	PCA scatterplot

Univariate Graphical Techniques

Type Of Variable	Graphical Techniques
Continuous Variable	Histogram, KDE, Boxplot and Q-Q plot(specifically for outliers)
Categorical Variable	Bar Plot, Pie Chart, Frequency Table

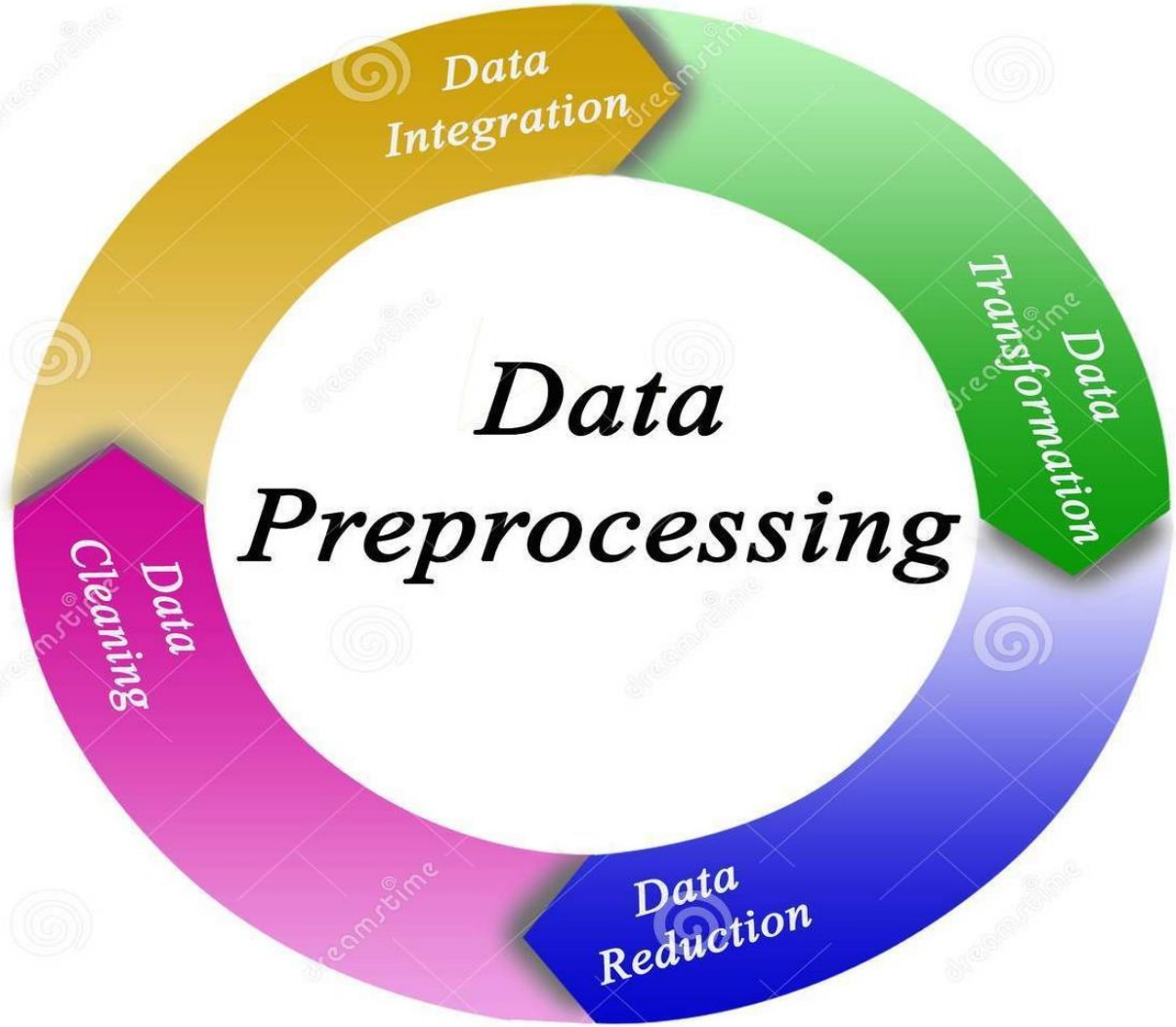
Bivariate Graphical Techniques

Type Of Relationship	Graphical Techniques
Continuous-Continuous	Scatter plot, HeatMap, JointPlot, PairPlot.
Categorical-Contnuous	Factor plot, SwarmMap, ViolinPlot, StripPlot.
Categorical-Categorical	Crosstab, Stacked Bar, Barchart.

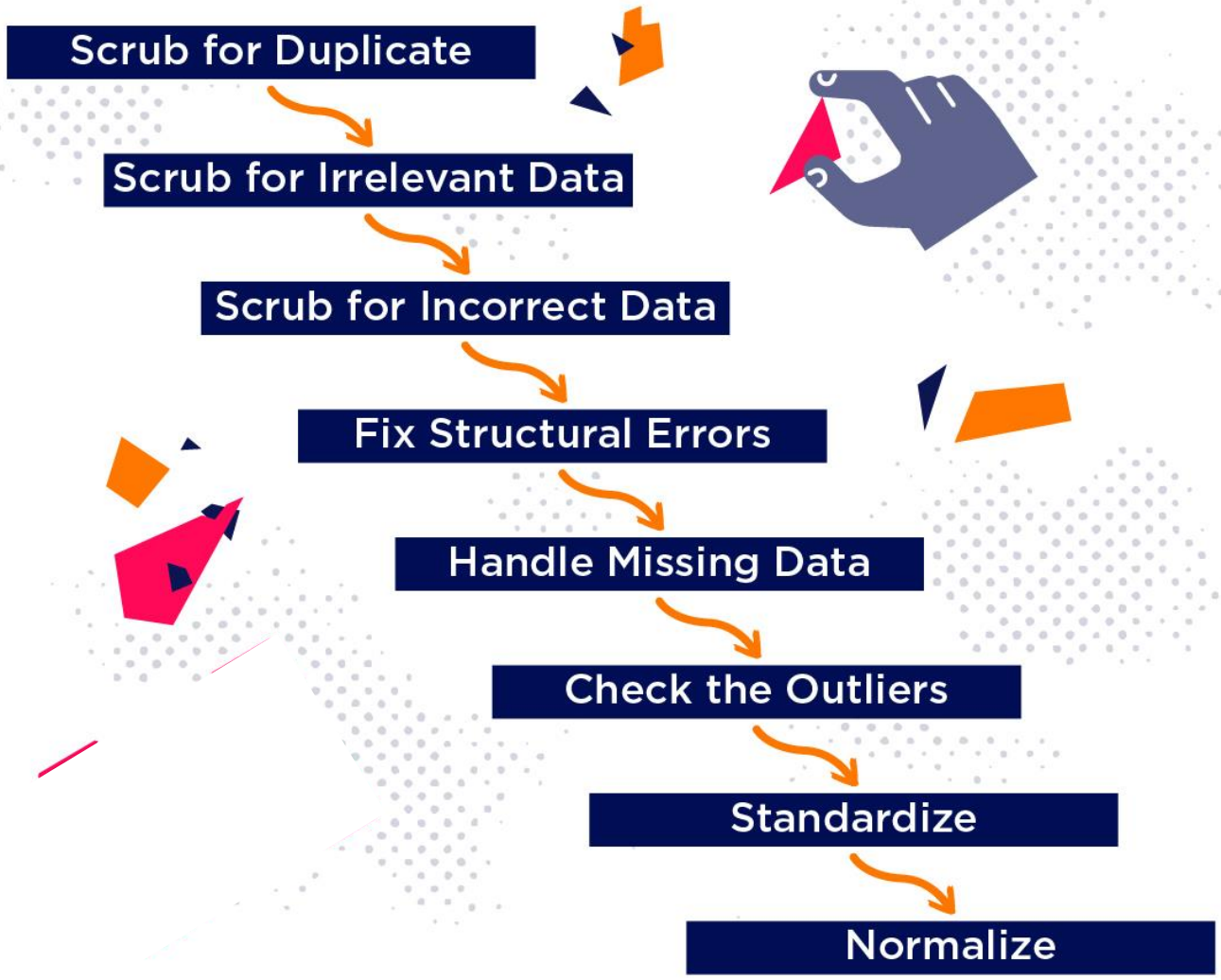
Data Preprocessing

- Data preprocessing is a process of preparing the raw data and making it suitable for further data analysis.
- Data preprocessing transforms the data into a format that is more easily and effectively processed in data mining, machine learning and other data science tasks.
- The techniques are generally used at the earliest stages of the machine learning and AI development pipeline to ensure accurate results.

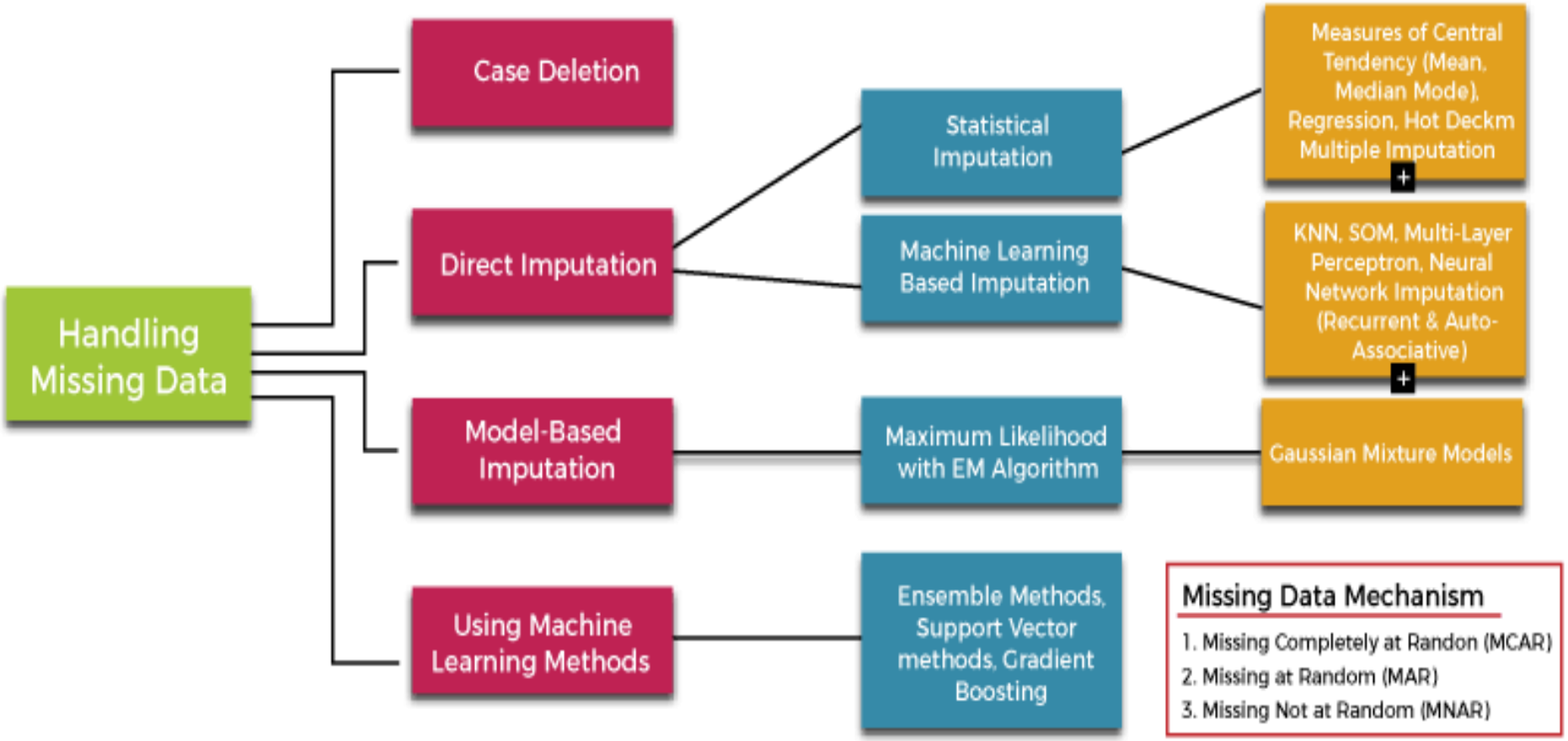
Preprocessing tasks



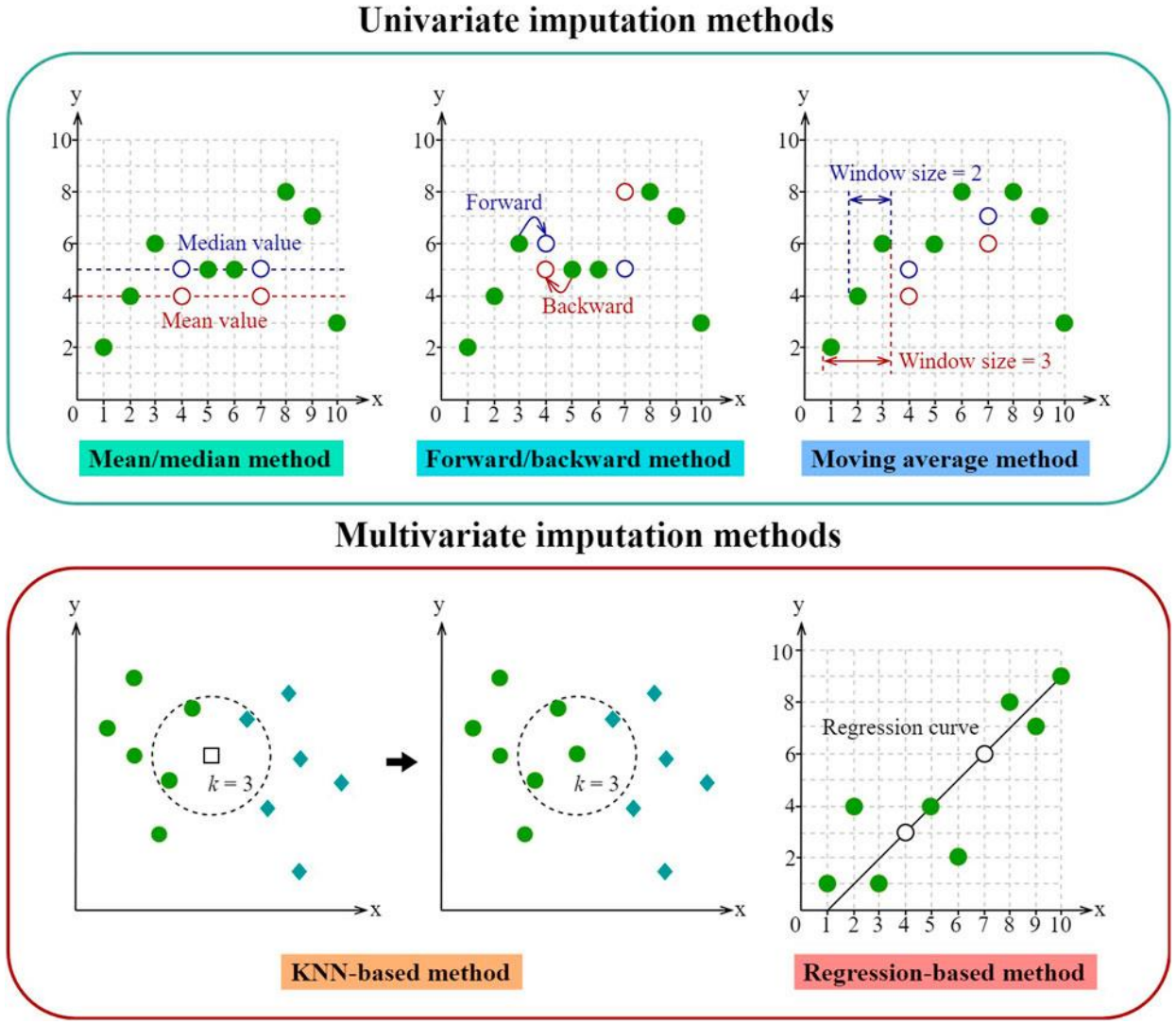
Data Cleaning Techniques



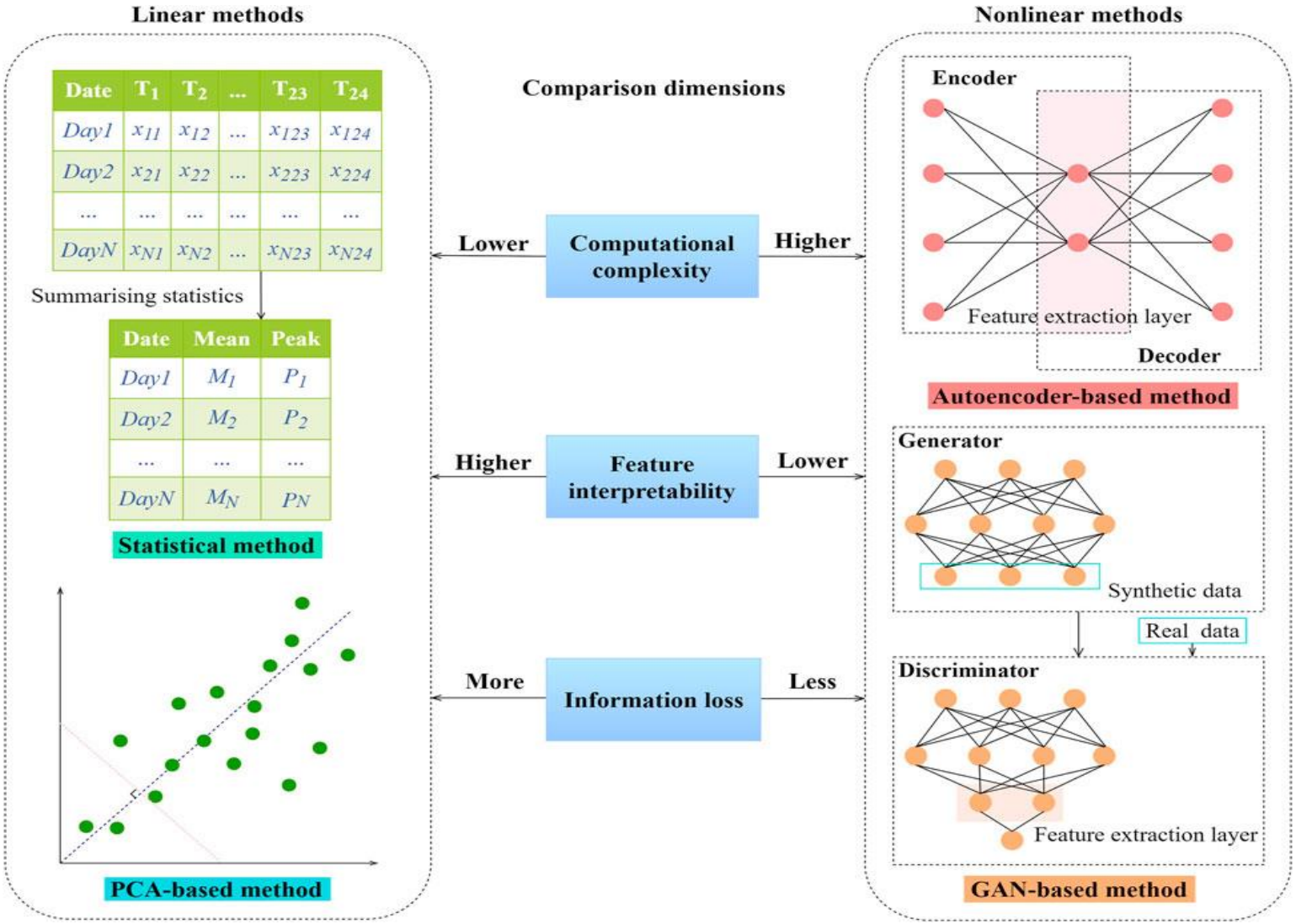
Handling Missing Data



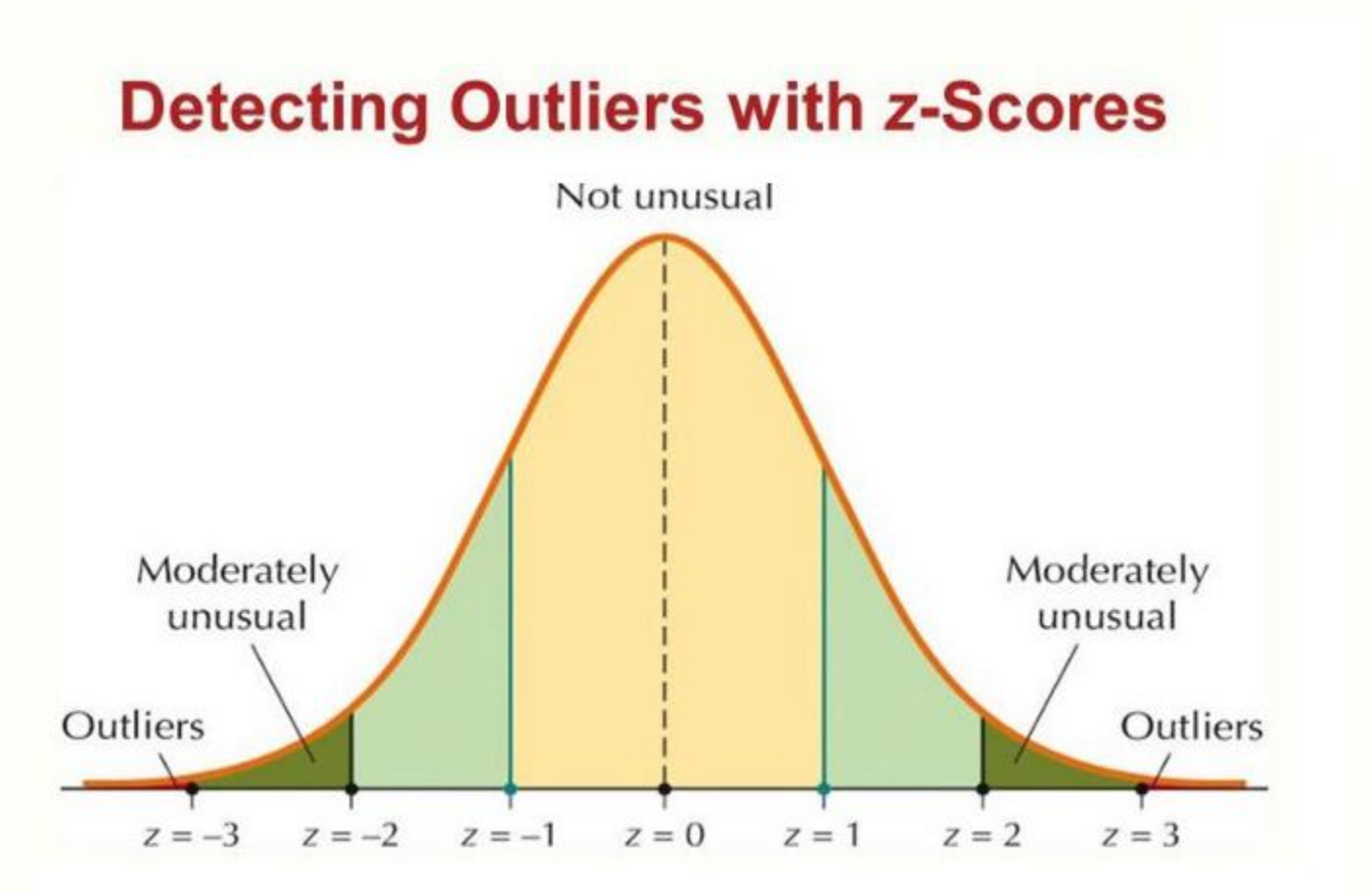
Missing value imputation



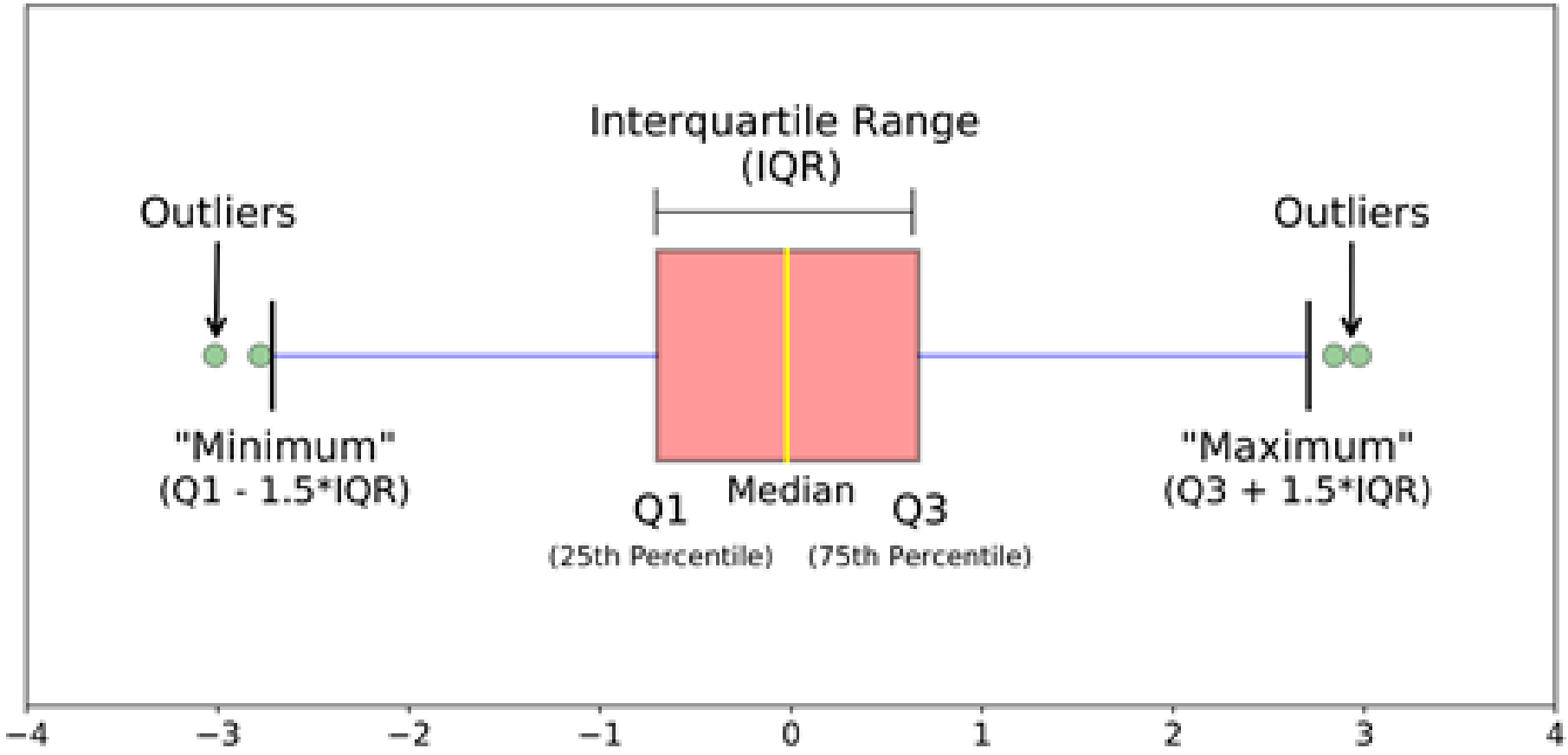
Feature extraction methods



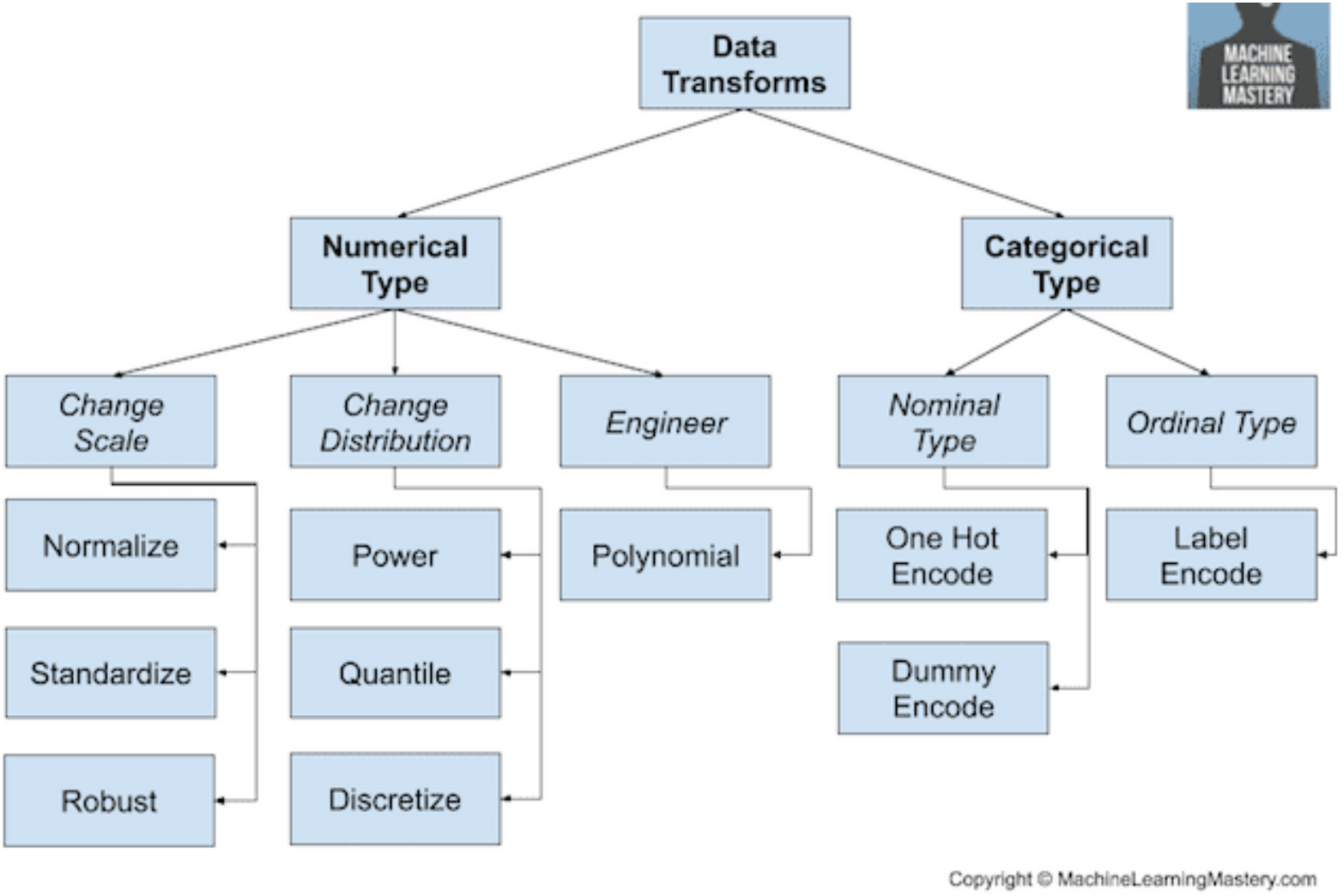
Outlier Detection with z-scores



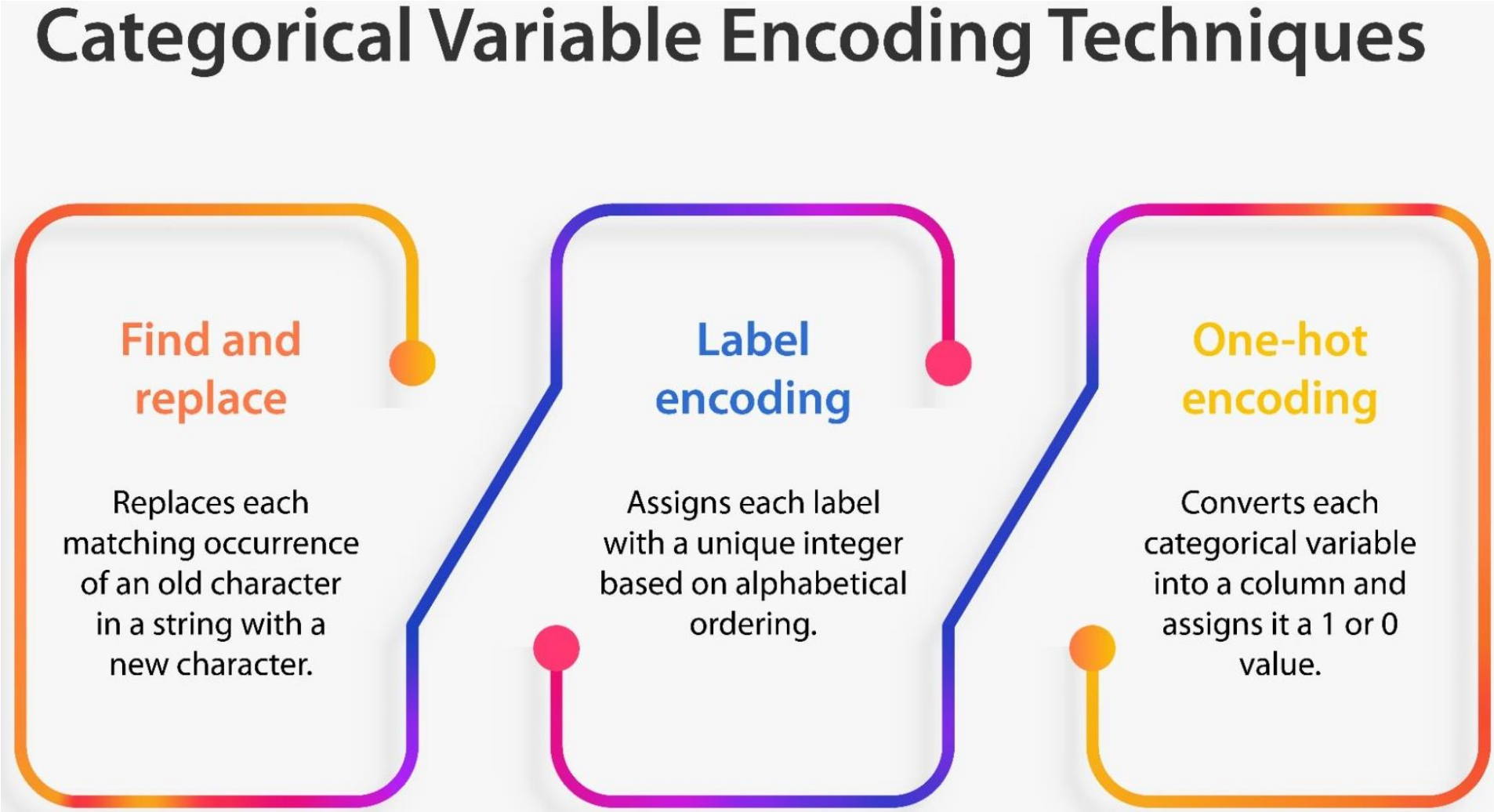
Outlier Detection using IQR



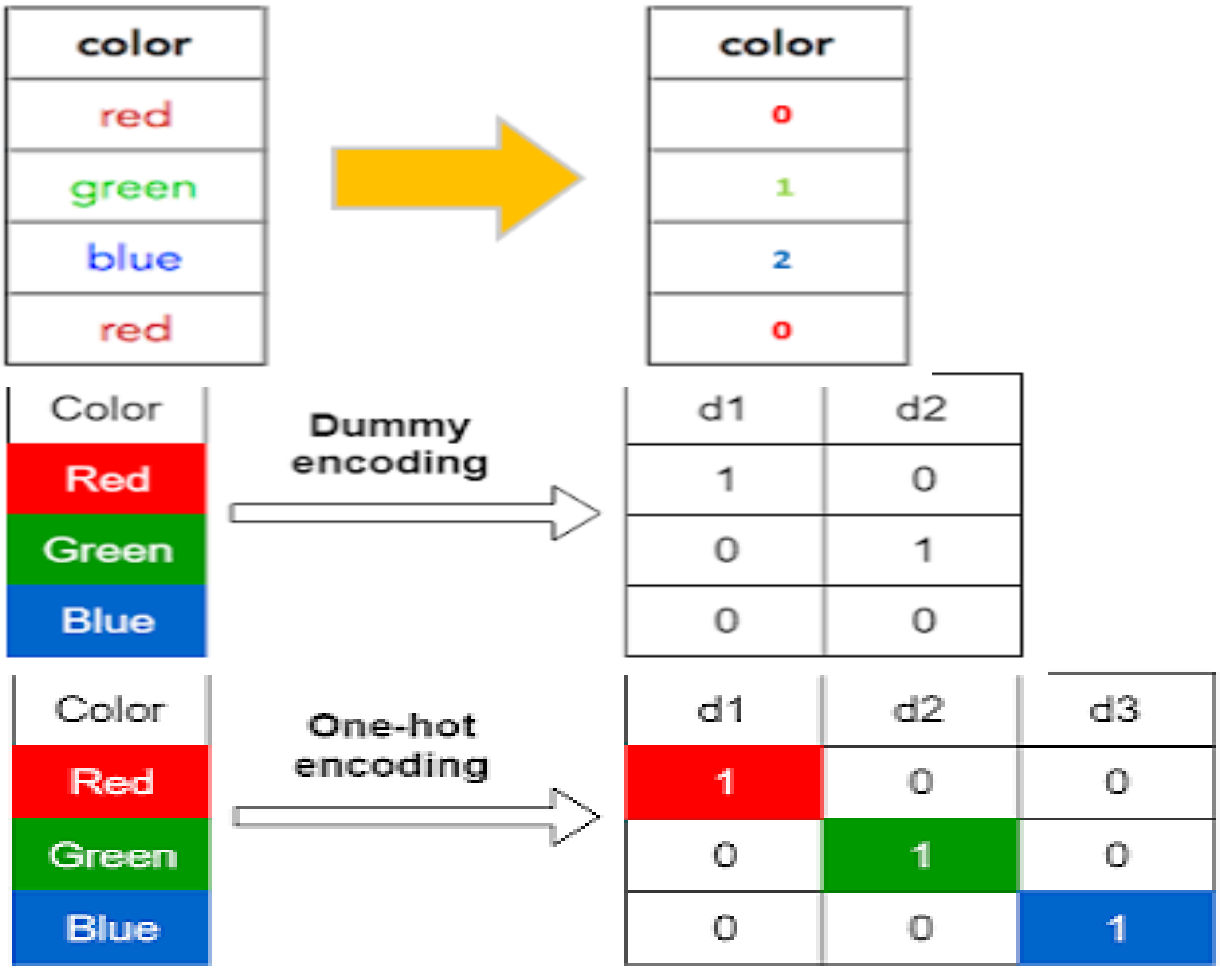
Data Transformation Types



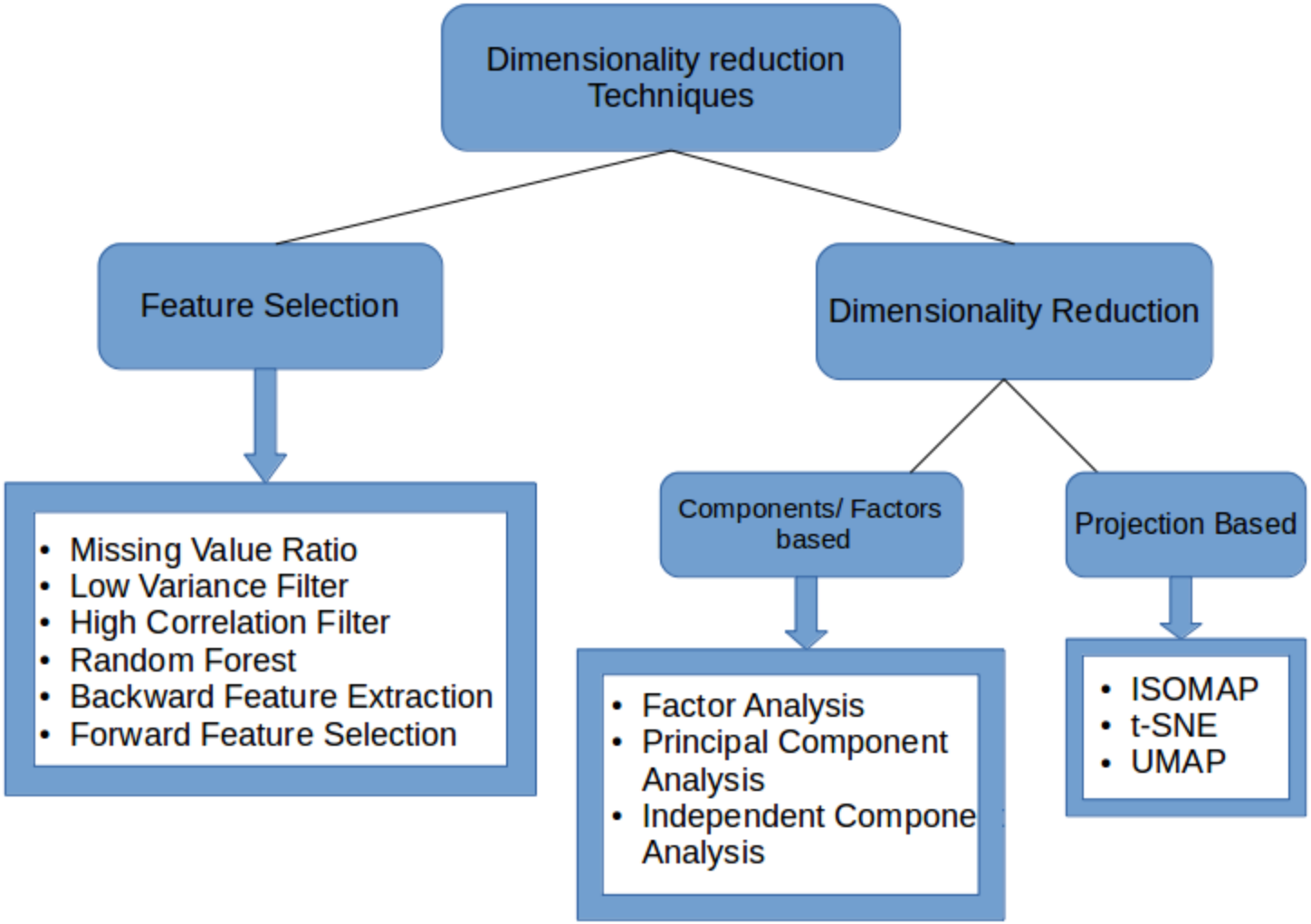
Encoding with Python



Encoding Categorical Variable

- 

Dimension Reduction

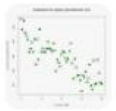


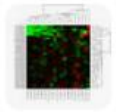
Data Visualization

- Data visualization is the graphical representation of information and data.
- By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
- Data visualization for exploring data
  - it is very difficult to gain knowledge from a large amount of data
  - visualization techniques help see the trends, outliers, or any kind of patterns in the data.
  - All this information helps data scientists to make better decisions to achieve their objective.
- Data visualization for storytelling with data
  - Graphical representations of information and important findings.
  - This process helps the presenter communicate data in a way that's easy for the viewer to interpret and draw conclusions.

Data Visualization Methods


Bar chart


Scatter plot


Heat map


Treemapping


Bubble chart

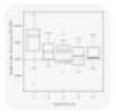

Matrix


Cloud computing


Area


Pie chart

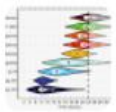

Line chart


Box plot

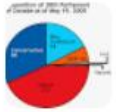

Infographic


Visual analytics


Bullet graph


Violin plot

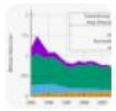

Streamgraph



Chart


Histogram


Table


Area chart

Line graph

FusionCharts


Computer network diagr...


Choropleth map

EDA vs Preprocessing

- EDA - no model trained yet, just exploring the data to see if there are potential problems in the dataset (outliers, mislabeled data, unwanted correlations between variables/samples, etc).
- Preprocessing - the steps required to go from raw data to a format suitable to input to your ML model.
- Data Processing is, generally, the collection and manipulation of items of data to produce meaningful information. Data Preprocessing is the process of preparation of data directly after accessing it from a data source.

Summary

- Descriptive Statistics
  - Central Tendency [Mean, Median, Mode, Quartile, Quantile]
  - Dispersion [Variance, Standard Deviation, Coefficient of Variation]
  - Skewness [Coefficient of Skewness]
- Data Preprocessing
  - Getting the data [Read/Load dataset]
  - Data Cleaning [Outlier, missing values, categorical variable, drop, rename etc]
  - Dimension reduction [Dropping variable/sample, PCA etc]
  - Data Transformation[Scaling, creating new features etc]
- Exploratory Data Analysis
  - Summary Presentation
  - Graphical Presentation
- Data Visualization
  - For exploring data
  - For storytelling with data

THANK YOU