Data Visualization and Description

Prof. Hariprasad Kodamana

August 27, 2025

What is a Model?

Definition

A simplified representation of a real-world system, used to understand, predict, or control its behavior.

- It captures the essential features while ignoring irrelevant details.
- In engineering, models are often mathematical, describing the relationships between variables.
- For example: A model might describe how solar irradiance and panel temperature affect a photovoltaic cell's power output.

Three Paradigms of Modeling

First-Principles

- Based on fundamental physical laws (e.g., thermodynamics, kinetics).
- **Analogy:** A "White Box". We know the internal mechanics.
- Requires deep domain expertise.

Data-Driven

- Learns relationships directly from historical data.
- **Analogy:** A "Black Box". We know the inputs and outputs, but not the inner workings.
- Does not assume physical laws.

Hybrid

- Combines known principles with data-driven methods.
- **Analogy:** A "Grey Box". We know some of the mechanics.
- Uses data to fill in the gaps in our knowledge.

From Data-Driven Models to Machine Learning

The Modern Approach

Data-driven modeling is the foundational concept behind Machine Learning.

- Machine Learning provides a powerful set of algorithms to automatically build models by learning patterns from data.
- Instead of a human defining the model's equations, the algorithm discovers the relationships.

The Starting Point

All machine learning begins with data. Therefore, to build these models, we must first understand the structure of the data itself.

Identifying the Machine Learning Problem

The Goal: Generalizing from Data

At its core, Machine Learning is about using data to train a model that can make useful predictions on new, unseen data. The specific task depends on the nature of the problem and the data available

The first step is to categorize the problem. The three main paradigms of machine learning offer different strategies for different types of data.

Problem Type 1: Supervised Learning

Concept

The model learns from a dataset containing labeled inputs and their corresponding correct outputs.

Regression: Predicting a Continuous Value

- The output is a number.
- **Example 1:** Forecasting the power output (in MW) of a solar farm based on weather data (irradiance, temperature).
- **Example 2:** Predicting the energy consumption of a building based on its size, insulation, and occupancy.

Classification: Predicting a Discrete Category

- The output is a label.
- **Example 1:** Predicting if a wind turbine will have a critical fault ("Yes" or "No") based on vibration and temperature sensor data.
- **Example 2:** Classifying an energy source in a smart grid as 'Solar', 'Wind', or 'Coal' based on real-time data.

Problem Types 2 & 3: Unsupervised and Reinforcement Learning

Unsupervised Learning

The model learns from unlabeled data to find hidden patterns or intrinsic structures.

- **Clustering:** Grouping similar data points.

  Example: Segmenting households into different "energy user profiles" based on smart meter data to design targeted efficiency programs.

- **Dimensionality Reduction:** Simplifying data.

  Example: Identifying the few key variables that most influence power grid stability from thousands of sensor readings.

Reinforcement Learning

A model ("agent") learns to make decisions by performing actions in an environment to maximize a cumulative reward.

- It learns through trial and error.
- *Example: Training an AI agent to manage a large-scale battery storage system. The agent learns when to charge the battery (when prices are low) and when to discharge to the grid (when prices are high) to maximize revenue.*

Data

Typically, m data points are present (each data point is a row with $y_1, x_1, x_2, ..., x_n$ in a Table) such that $m > n$.

Sample No.	Actual output, y	First Feature, x_1	Second Feature, x_2	Third Feature, x_3	..	n^{th} Feature, x_n
1	$y^{(1)}$	$x_1^{(1)}$	$x_2^{(1)}$	$x_3^{(1)}$	..	$x_n^{(1)}$
2	$y^{(2)}$	$x_1^{(2)}$	$x_2^{(2)}$	$x_3^{(2)}$	..	$x_n^{(2)}$
:	:	:	:	:	:	:
m	$y^{(m)}$	$x_1^{(m)}$	$x_2^{(m)}$	$x_3^{(m)}$	..	$x_n^{(m)}$

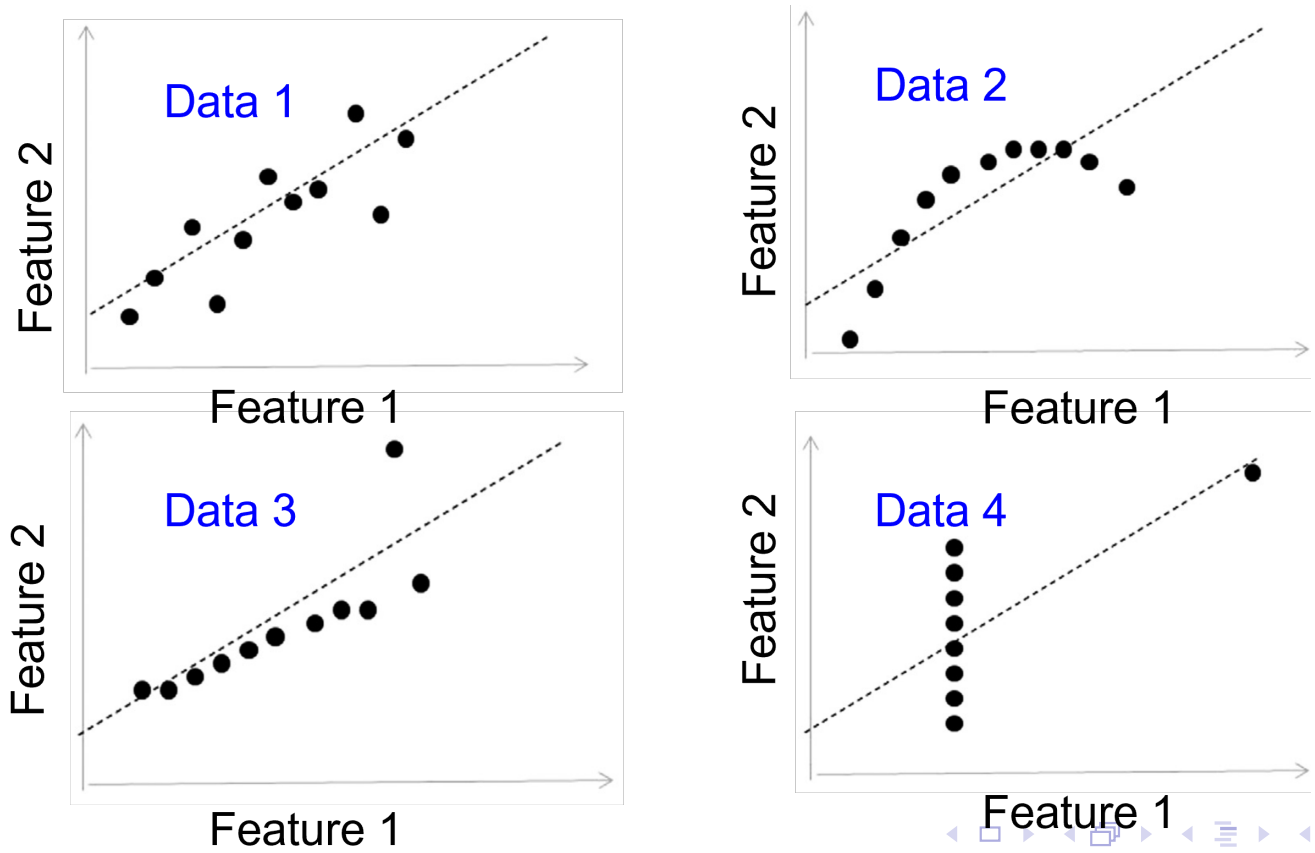
Anscombe's Quartet

Table: Properties of Anscombe's Quartet

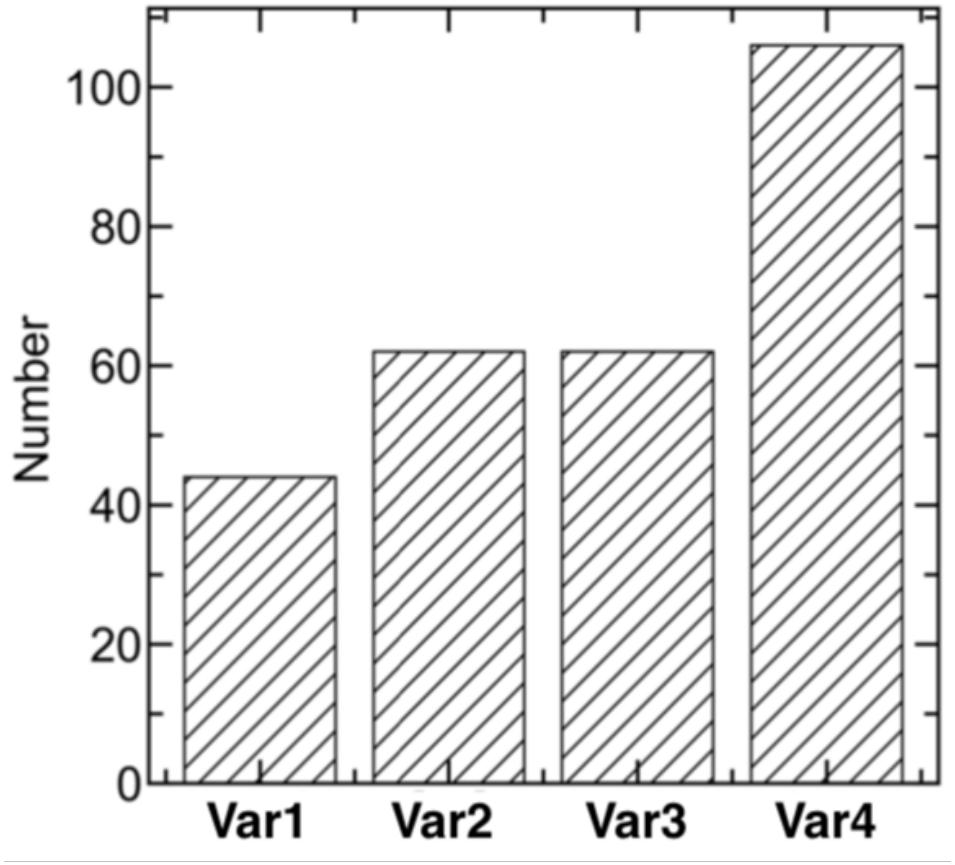
Property	**Dataset I**	**Dataset II**	**Dataset III**	**Dataset IV**
Mean of x	9.0	9.0	9.0	9.0
Variance of x	11.0	11.0	11.0	11.0
Mean of y	7.5	7.5	7.5	7.5
Variance of y	4.12	4.12	4.12	4.12
Correlation (x,y)	0.816	0.816	0.816	0.816
Linear regression line	$y = 3.00 + 0.500x$	$y = 3.00 + 0.500x$	$y = 3.00 + 0.500x$	$y = 3.00 + 0.500x$
R-squared	0.666	0.666	0.666	0.666

Data Visualization-Anscombe's Quartet

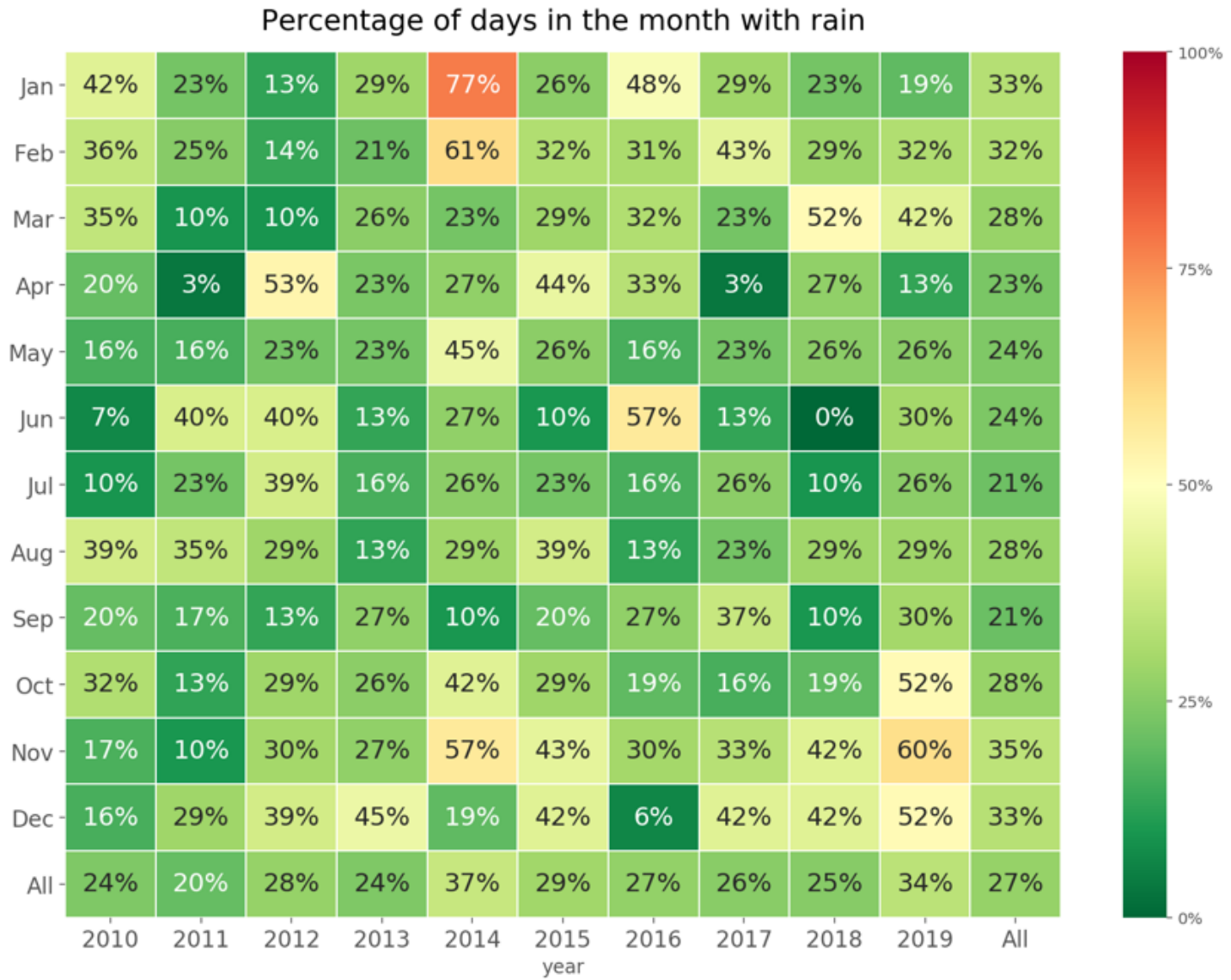
1. The first step in understanding the correlation between the data is visualization. It can identify nonlinear relationships and show any outliers in the dataset.
2. Anscombe's quartet (Anscombe, 1973) consists of four different datasets. All four datasets have the same mean, the same variance, and the same correlation coefficient, but look drastically different when plotted.



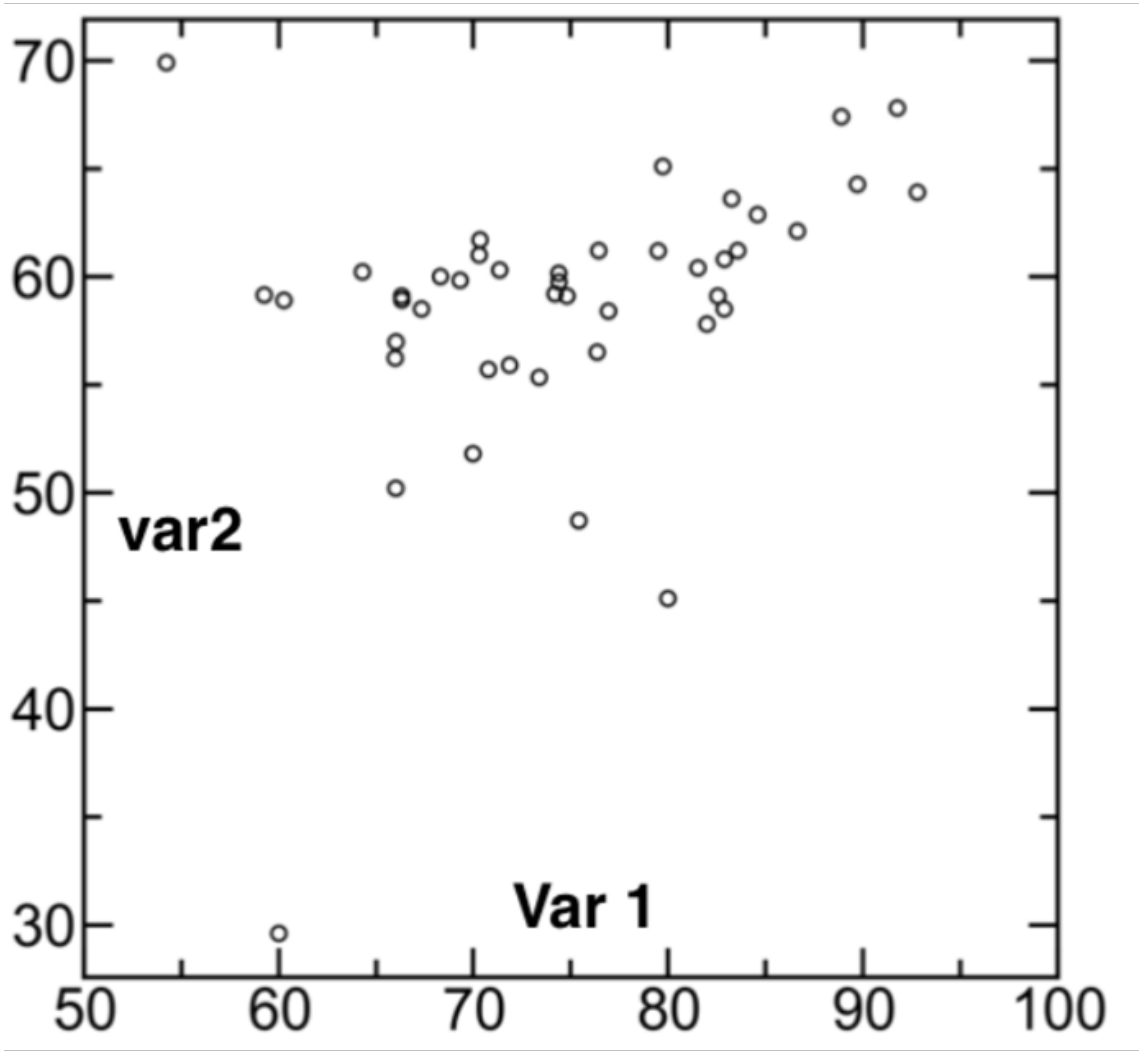
Data Visualization: Bar Plot



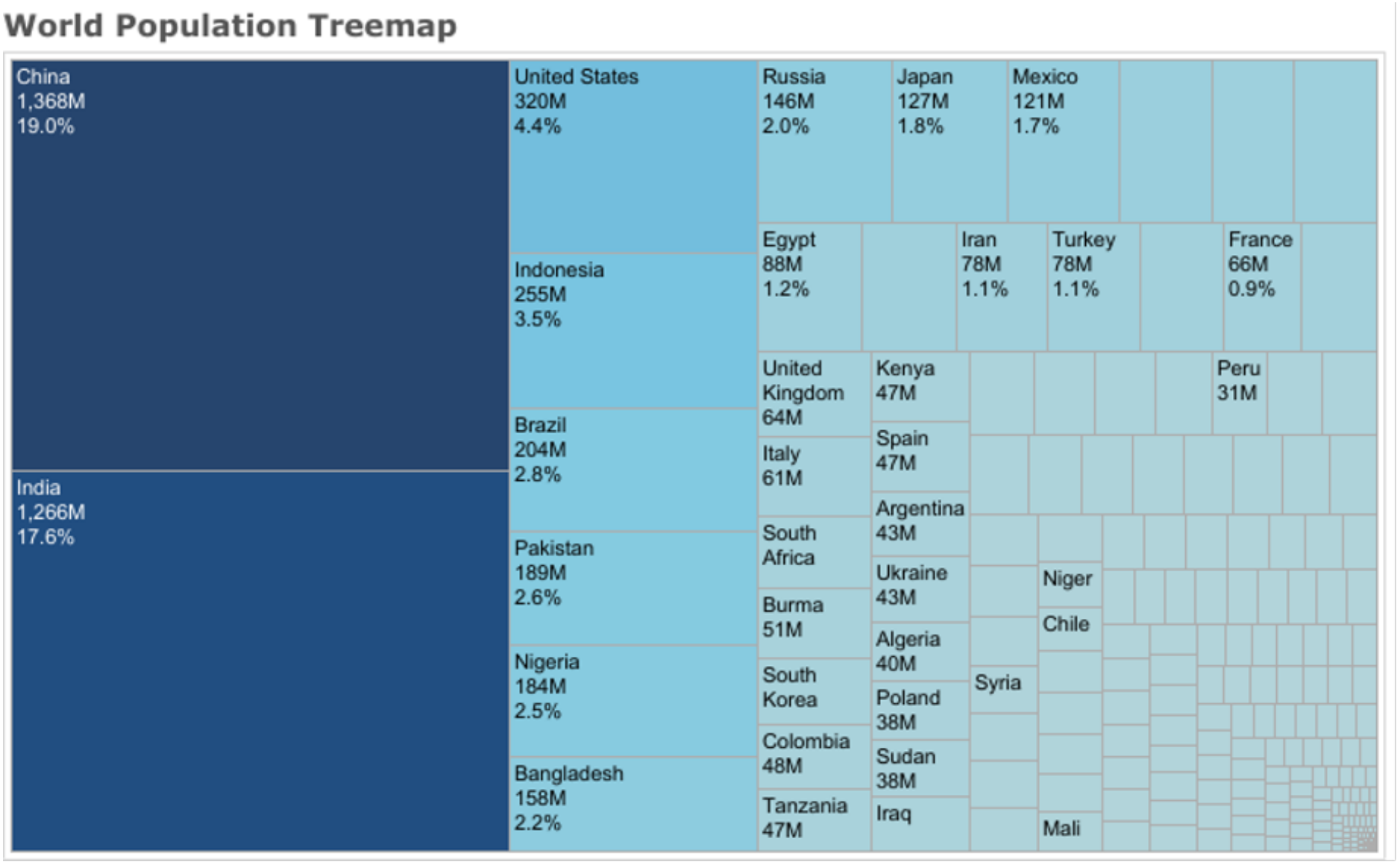
Data Visualization: Heat Map



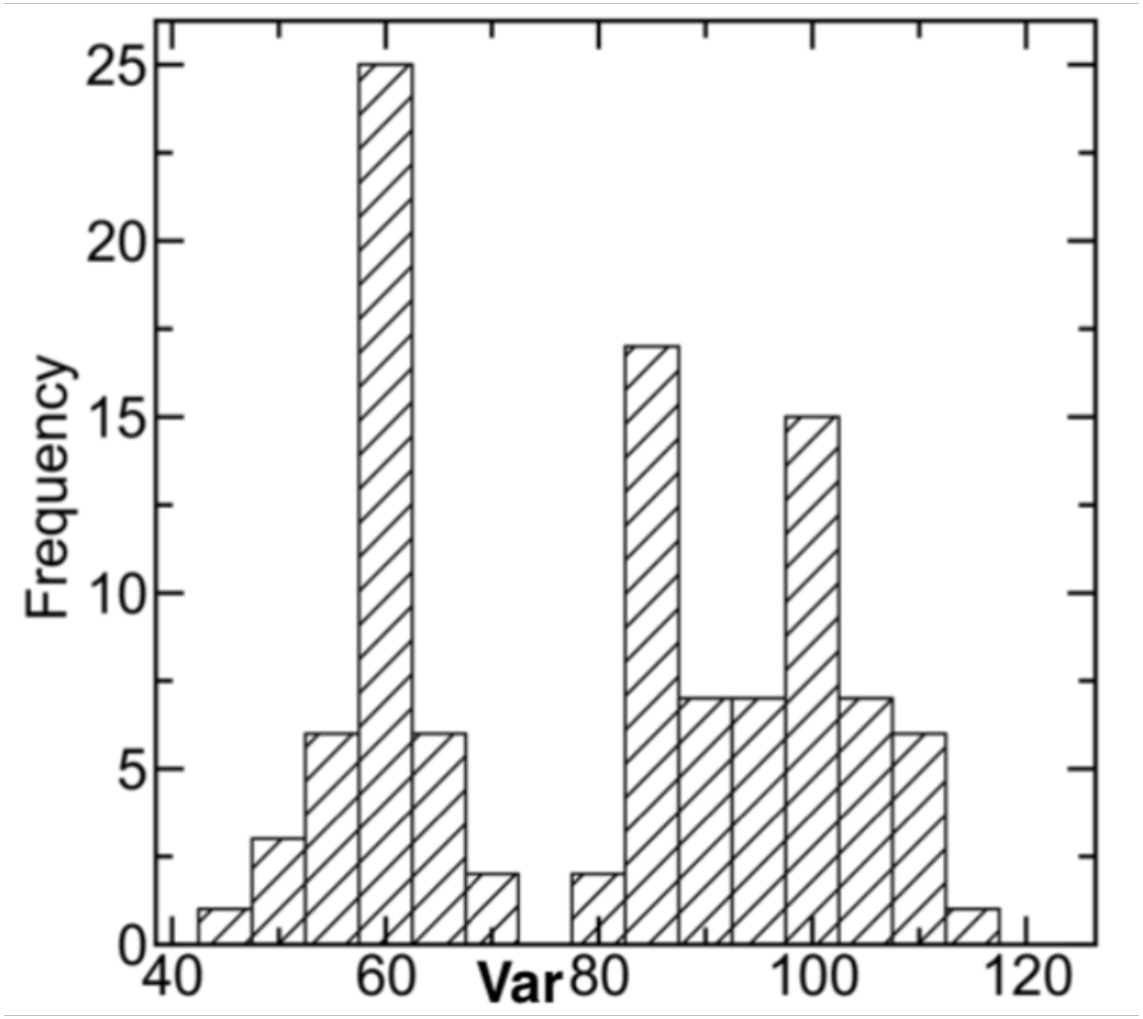
Data Visualization: Scatter Plot



Data Visualization: Tree Map



Data Visualization: Histogram



Data Visualization: Histograms and Binning

A histogram is a powerful graphical representation of the distribution of numerical data. It works by dividing the range of values into a series of intervals, called **bins**, and then counting how many values fall into each bin.

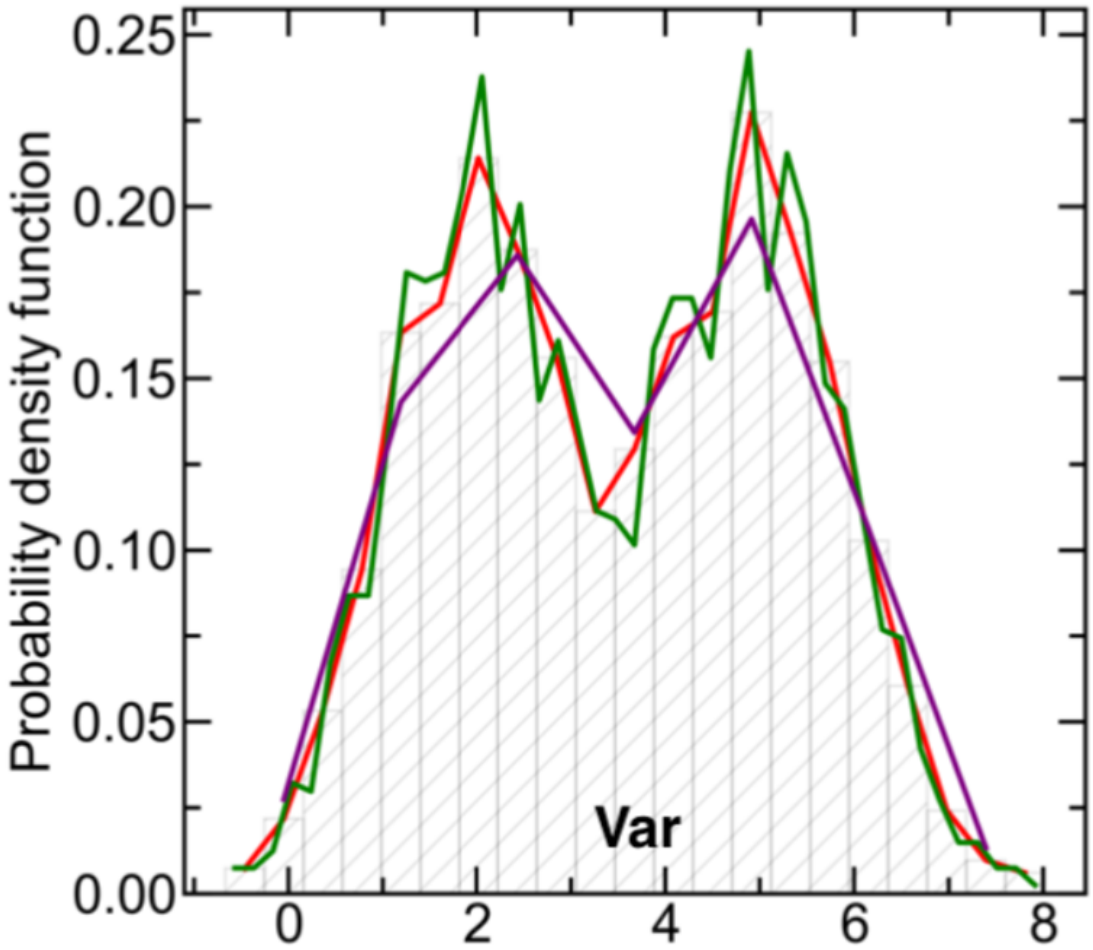
- The number of bins is the most important parameter. It controls the level of detail in the visualization.
- **Too Few Bins:** Can over-simplify the distribution, hiding important features like peaks or gaps.
- **Too Many Bins:** Can create a noisy and jagged plot that obscures the underlying shape of the data.

Common Rules for Choosing the Number of Bins:

- **Square Root Rule:** Number of bins $\approx \sqrt{n}$
- **Sturges' Rule:** Number of bins $= 1 + \log_2(n)$

(where n is the number of data points)

Data Visualization: Distribution Plot



Distribution plots are obtained by connecting the midpoints of histograms. Kernel density plots are also used for representing distributions.

Why Scaling Matters in Visualization

When visualizing two variables, if their scales are dramatically different, the variable with the smaller scale can appear compressed and its patterns can be hidden.

Before Scaling:

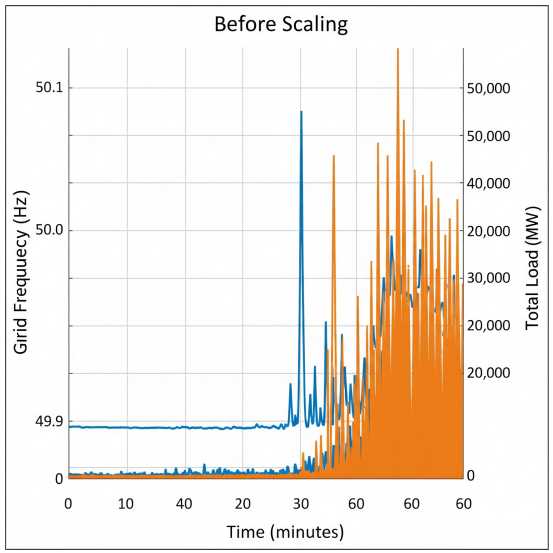
- One axis dominates the plot.
- Visual relationships are distorted.

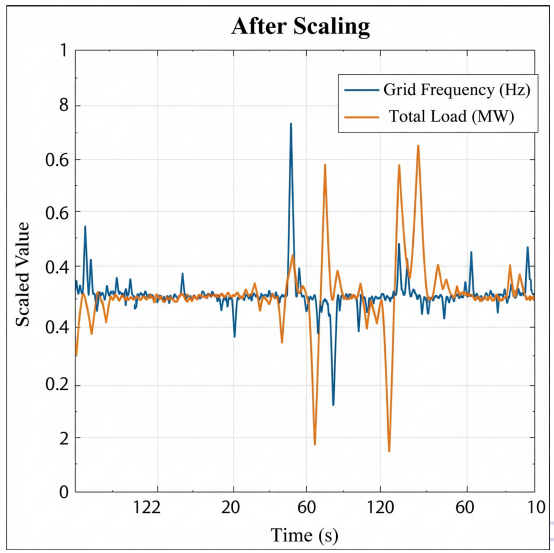
After Scaling (e.g., Min-Max, Std):

$X_normalized = (X_i - X_min)/(X_max - X_min)$

$X_{std} = (X - mean(X)/(mean(X) - std(X))$

- Both axes have a similar range (e.g., 0 to 1).
- The true underlying pattern becomes visible.





Data Description: Population and Sample

1. The qualitative nature of data can be obtained from data visualization.
2. However, a quantitative representation requires the extraction of statistics from the data.
3. A population is the set of all possible data of the characteristics under investigation. This population may be finite or infinite in size.
4. Due to various limitations, a population may not be fully accessible; however, a sample of the population may be accessed.

Measures of Central Tendency and Spread

1. A measure of central tendency of a dataset is a single characteristic that describes the data most appropriately. They are also called summarizing statistics.
2. The mean (or average) is the most widely used central measure, along with the median and the mode.
3. Measures of spread summarize how scattered the data is with respect to the central measure.
4. Common measures of spread include range, percentile, and variance.

Mean

1. Mean is the most common central measure used to represent continuously distributed data.
2. The sample mean is defined as:

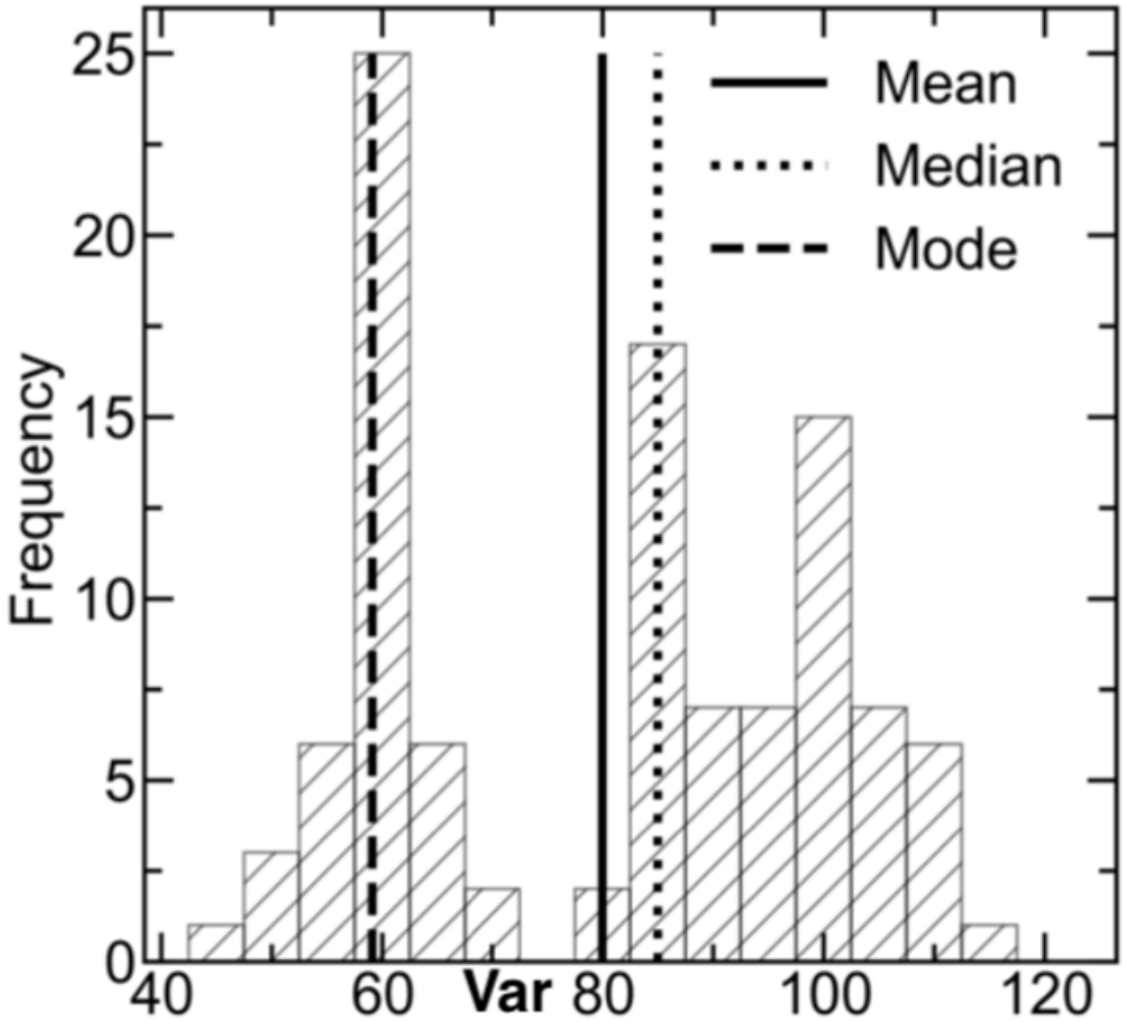
$$\bar{X} = \frac{1}{n} \sum X_i$$

   Where n is the sample size and X_i is the i^{th} sample.
3. Note that the mean may not represent the region where data is most densely distributed, especially if the distribution is asymmetric.
4. In fact, the mean can occur in regions with no or sparse data.
5. Further, as the mean is the weighted sum of all the data points, outliers present in the data can significantly affect the location of the mean.

Median and Mode

1. For n observations, the sample median, M, is the $[(n+1)/2]^{th}$ largest observation for odd values of n, or mean of $(n/2)$ and $[(n/2)+1]^{th}$ largest observations for even values of n.
2. Hence, the median is the value that splits the dataset in half. Due to this property, outliers present in the data may not affect the location of the median significantly.
3. The mode is the value that occurs with the most significant frequency.
4. If each X_i is unique, it is not possible to represent the Mode.
5. Typically, the Mode is the best central measure while dealing with categorical or discrete data.

Mean, Median and Mode



Range, Percentile, and Quartiles

1. The range is the simplest measure of the spread in the dataset. Range is defined as the difference between the largest and smallest observations in a sample set. Although it is easy to compute, the range is highly sensitive to outliers.
2. The p^{th} percentile is that threshold such that $p\%$ of observations are at or below this value. It is $(k+1)^{th}$ largest sample point if $np/100 \neq$ integer, where k is the largest integer less than $np/100$, for a total of n observations. The average of the $(np/100)^{th}$ and $(np/100+1)^{th}$ observations if $np/100$ is an integer.
3. The first quartile (Q_1), the second quartile (Q_2), and the third quartile (Q_3) are defined as $25p = (n+1)/4$, $50p = (n+1)/2$, and $75p = 3(n+1)/4$, respectively.
4. Second quartile is also known as the median (M).

Variance

1. A key measure to express the spread of data is variance.
2. The variance (S^2) and its square root value, the standard deviation (S), are measures of the variability of the data sets around the mean.
3. They give a picture of the distribution of the data around their mean value. If a dataset is highly dispersed, it tends to spread farther away from the mean, leading to a high value of variance and standard deviation and vice versa.
4. The standard deviation of a normal distribution enables us to calculate confidence intervals.
5. In a normal distribution, about 68% of the values lie within one standard deviation either side of the mean and about 95% of the scores are within two standard deviations of the mean, and about 99.5% of the values are within three standard deviations of the mean.
6. The sample variance is computed by

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

7. It is to be noted for the population data, notation used for mean, variance, and standard deviation are μ, σ^2 and σ, respectively.

A Note on Variance: Why Divide by n-1?

When we have data from an entire population, we use N:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} \quad \text{(Population Variance)}$$

But when we only have a sample, we use n-1:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \quad \text{(Sample Variance)}$$

- The sample mean ($\bar{X}$) is calculated *from our specific sample data*. It's a perfect center for that sample.
- The true population mean (μ) is unknown.
- Because of this, the sum of squared differences from our sample mean, $\sum(X_i - \bar{X})^2$, will on average be **smaller** than the sum of squared differences from the true population mean, $\sum(X_i - \mu)^2$.
- Dividing by the smaller number 'n-1' (instead of 'n') corrects for this underestimation. This is known as **Bessel's Correction**.

A Note on Variance: Why Divide by n-1?

The Degrees of Freedom Solution: Degrees of freedom (dof) are the number of independent values available to estimate a parameter.

- We start with ‘n‘ independent data points.
- However, to calculate variance, we first have to "spend" one degree of freedom to estimate the sample mean ($\bar{X}$).
- Once $\bar{X}$ is fixed, only ‘n-1‘ of the data points are free to vary (the last one is determined by the mean).
- Therefore, we divide by the available degrees of freedom, which is ‘n-1‘.

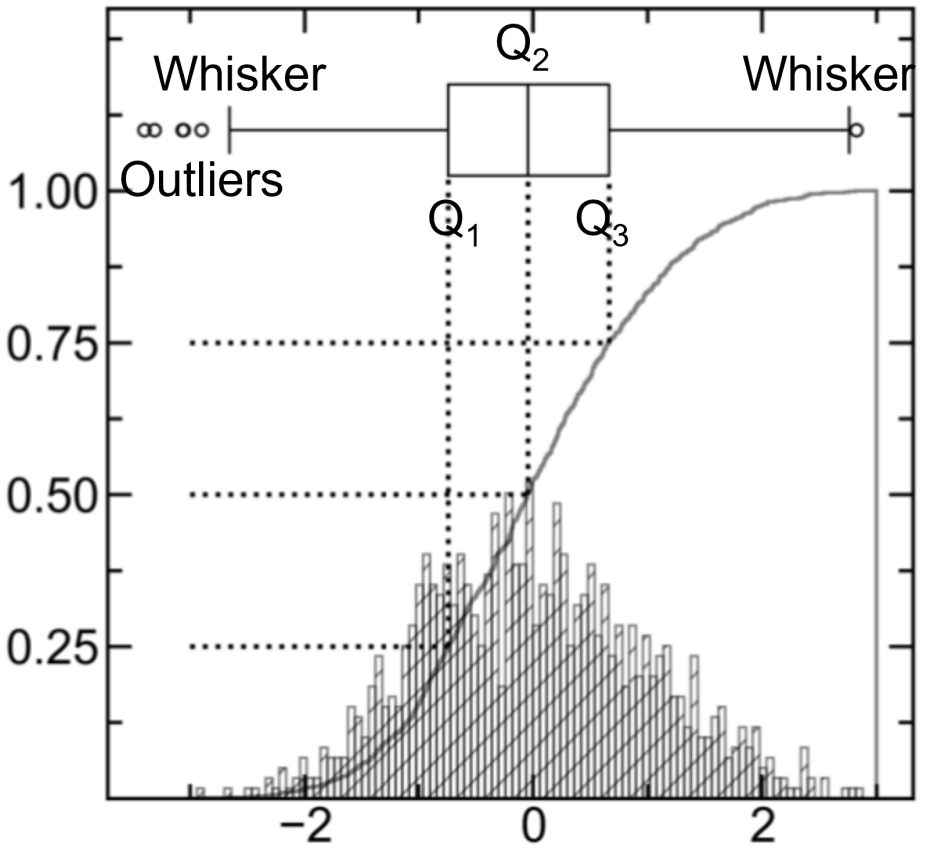
The Result: An Unbiased Estimator

Using ‘n-1‘ makes the sample variance (s^2) an unbiased estimator of the population variance (σ^2). This means that if we took many samples, the average of their sample variances would equal the true population variance.

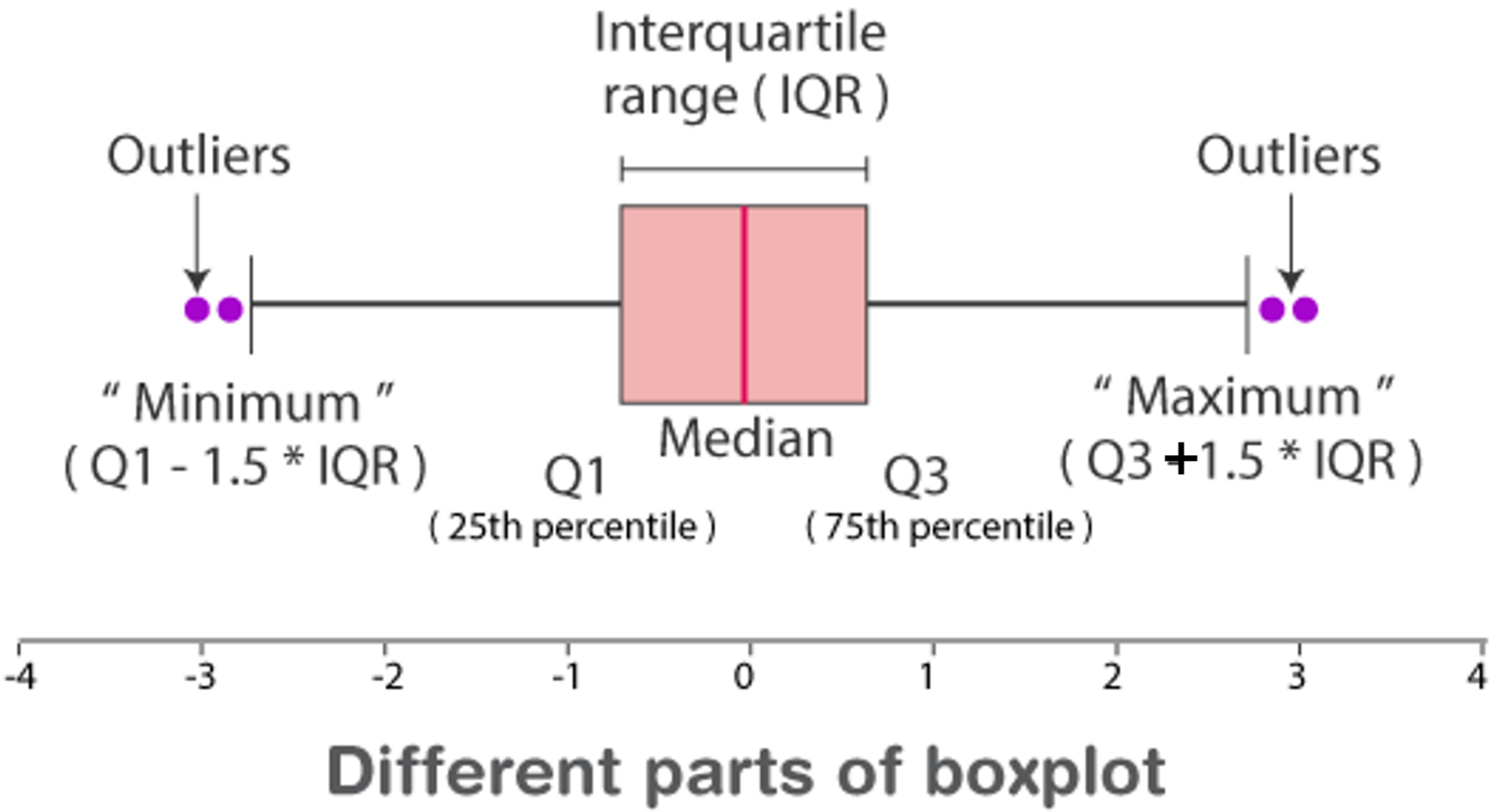
$$E[S^2] = \sigma^2$$

Box Plot

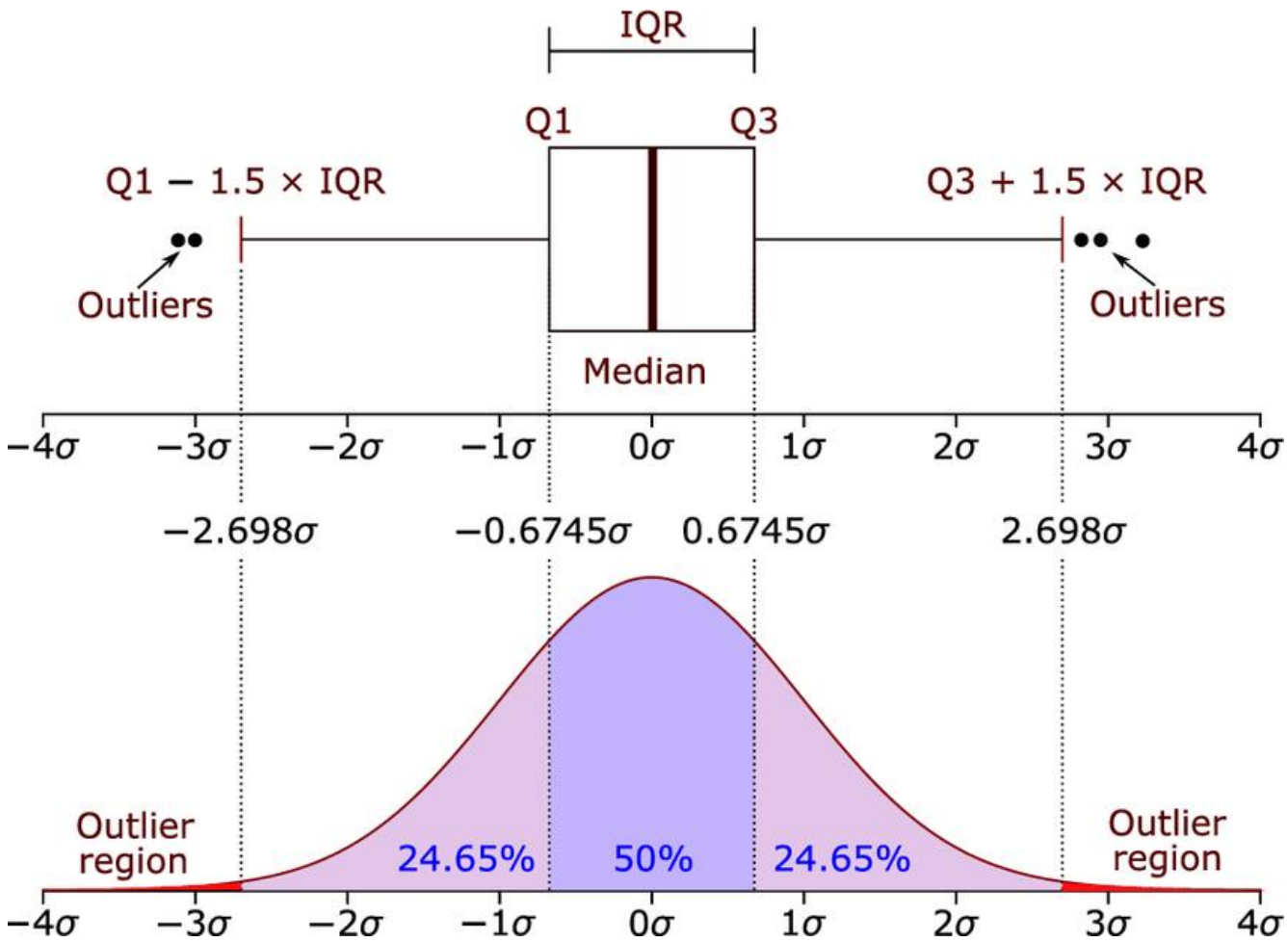
A box plot is a convenient graphic to represent range, median and quartiles. In a box plot, a box is drawn from Q_1 to Q_2, Q_2 (median) is drawn as a vertical line in the box, and outer lines (whiskers) are drawn either up to the outermost points (the length of the line represents the range) or at $1.5 \times (Q_3 - Q_1)$. Therefore, the length of the whiskers can be different.



Different Parts of a Boxplot



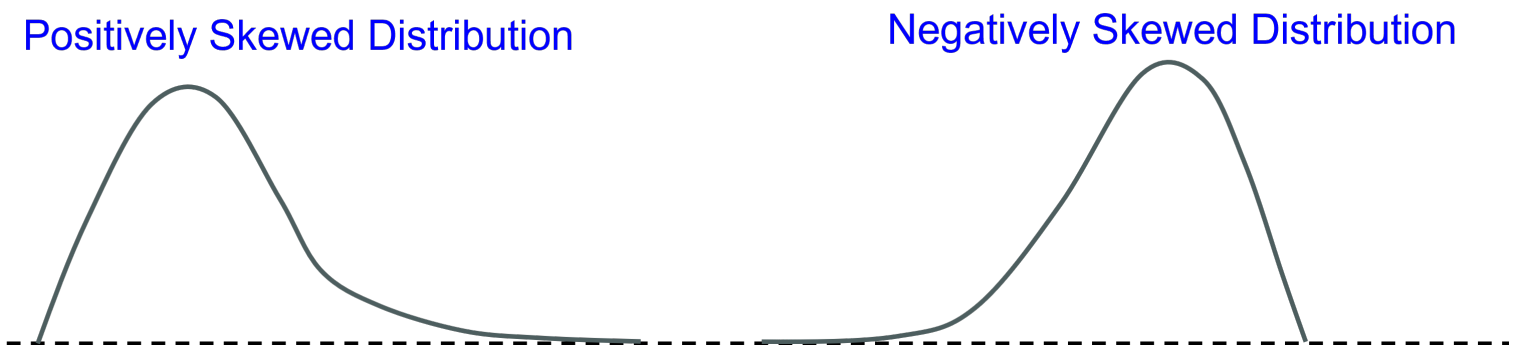
Different Parts of a Boxplot



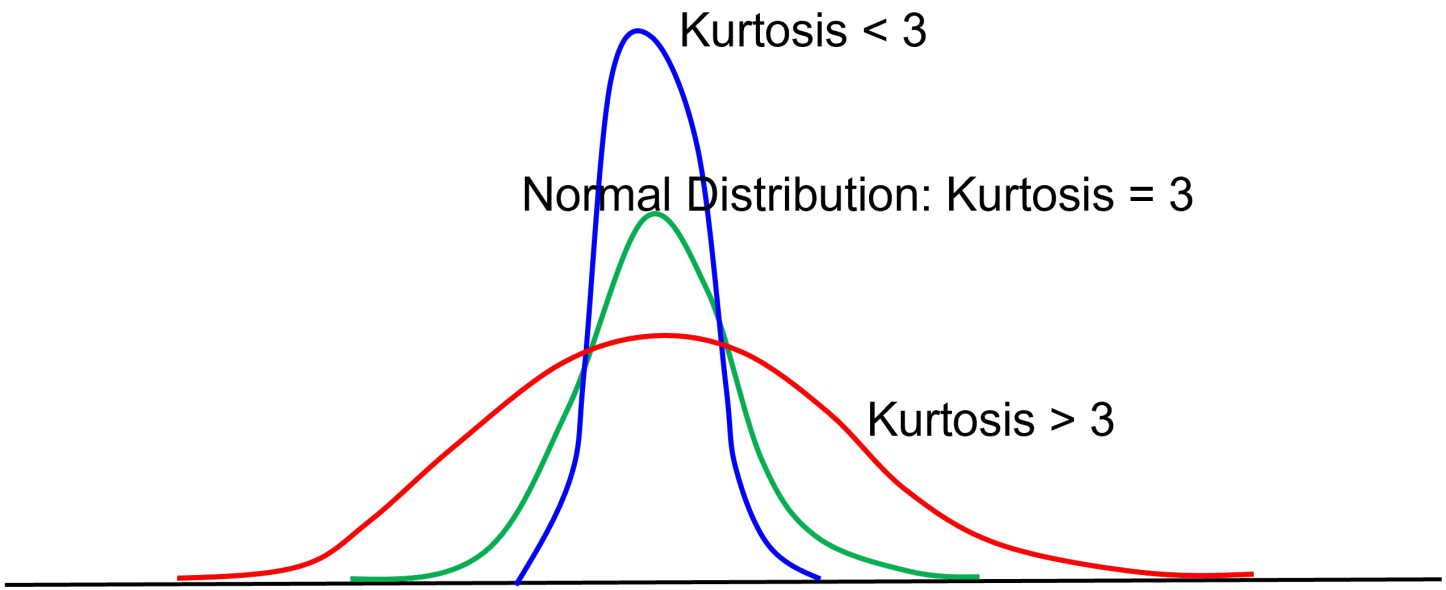
Higher Order Measures: Skewness and Kurtosis

1. Two higher-order measures of data representation are skewness and kurtosis. While skewness is a measure of the distortion, kurtosis is a measure heavy-tailed nature of the data relative to a Normal distribution.
2. Skewness measure the degree of distortion of the data from the normal distribution. A symmetrical distribution will have a skewness of 0. Skewness is calculated by
$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{nS^3}$$
3. If the $-0.5 \leq G_1 \leq 0.5$ then data are fairly symmetrical. If the $G_1 < -0.5$, it is called negatively skewed, while if the $G_1 > 0.5$, it is called positively skewed.
4. Kurtosis is used to describe the extreme values in one versus the other tail, and therefore, it is a measure of outliers present in the distribution. Excess Kurtosis is calculated as:
$$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$$
5. A High value of kurtosis in a data set is an indicator that the data has heavy tails or outliers, and vice versa.

Higher Order Measures: Skewness and Kurtosis



Normal Distribution is Symmetric with zero skewness and kurtosis = 3.



Measuring Joint Variability: Covariance

Covariance is a measure of how much two random variables vary together. It describes the direction of the linear relationship between two variables.

- **Positive Covariance:** Indicates that as one variable deviates from its mean, the other variable tends to deviate in the **same direction**.
  Energy Example: 'Temperature' and 'Air Conditioner Power Draw' typically have a positive covariance.
- **Negative Covariance:** Indicates that as one variable deviates from its mean, the other tends to deviate in the **opposite direction**.
  Energy Example: 'Building Insulation Thickness' and 'Heating Energy Consumption' would have a negative covariance.

The sample covariance is calculated as:

$$\text{Cov}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

Limitation

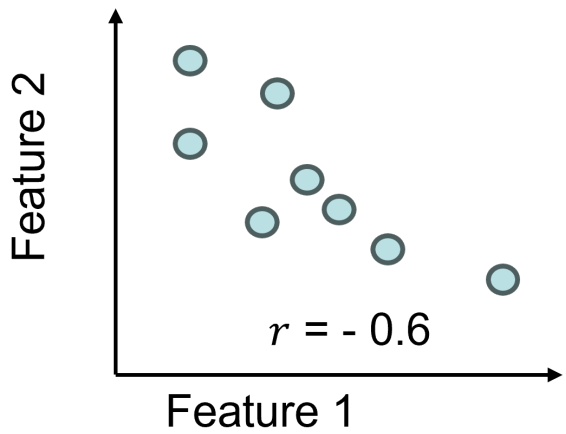
The magnitude of the covariance is not standardized and depends on the scale of the variables, making it difficult to interpret the strength of the relationship.

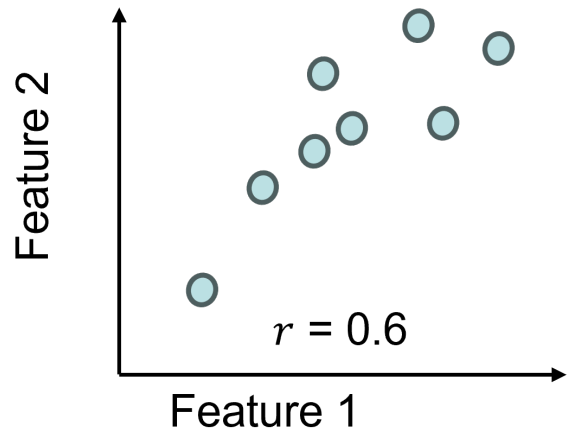
Correlations Between the Features

A correlation coefficient is a statistical measure that quantifies the degree to which two variables are related or associated. The most used correlation coefficient is Pearson's correlation coefficient, denoted by the symbol r.

1. r ranges from -1 to 1. $r = 1$ indicates a perfect positive linear relationship. r = -1 indicates a perfect negative linear relationship. $r = 0$ indicates no linear relationship.
2. The formula for Pearson's correlation coefficient between variables X and Y with observations x_i and y_i is given by:

$$r = \frac{\sum(x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum(x_i - \bar{X})^2 \sum(y_i - \bar{Y})^2}}$$





Spearman's Rank Correlation (ρ or r_s)

Non-parametric alternative to Pearson's that does not assume a linear relationship between variables.

Key Idea

It measures the strength and direction of a **monotonic** relationship by operating on the **ranks** of the data, not the raw values.

Relationship is monotonic if, as one variable increases, the other variable consistently increases or consistently decreases (but not at a constant rate).

How It's Calculated

1. Convert the values of each variable to ranks (from lowest to highest).
2. Calculate the difference in ranks (d_i) for each data pair.
3. Apply the Spearman correlation formula:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between ranks for each pair, and n is the number of observations.

Spearman's Rank Correlation (ρ or r_s)

Spearman's correlation measures a **monotonic** relationship by operating on the **ranks** of the data.

1. For each variable, the data points are replaced by their ranks (1st, 2nd, 3rd, etc.).
2. The difference in ranks (d_i) is calculated for each pair.
3. The formula is applied to these rank differences.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Where $d_i = \text{rank}(X_i) - \text{rank}(Y_i)$ and n is the number of observations.

Example of Ranking:

Raw Data		Ranks			
X	**Y**	**rank(X)**	**rank(Y)**	$\mathbf{d_i}$	$\mathbf{d_i^2}$
10	5	1	1	0	0
35	25	3	4	-1	1
12	8	2	2	0	0
40	20	4	3	1	1
			Sum of $\mathbf{d_i^2}$:		**2**

Comparison: Pearson vs. Spearman

Table: Key Differences Between Correlation Coefficients

Aspect	**Pearson Correlation (r)**	**Spearman Correlation (ρ)**
Measures	Strength and direction of a **linear** relationship.	Strength and direction of a **monotonic** relationship.
Data Type	Requires interval or ratio data.	Can be used with ordinal, interval, or ratio data.
Sensitivity to Outliers	**High**. Extreme values can heavily influence the result.	**Low**. Ranks are less affected by outliers, making it more robust.
Core Assumption	The variables are linearly related.	The variables are monotonically related.

Key Takeaway

Use Pearson for linear relationships. Use Spearman for monotonic relationships, or when your data has significant outliers.

Covariance vs. Correlation: A Summary

Covariance and Correlation both describe the direction of a linear relationship, but only correlation describes its strength in a standardized way.

Table: Covariance vs. Correlation

Aspect	**Covariance**	**Pearson Correlation (r)**
Value Range	Unbounded ($-\infty$ to $+\infty$)	Bounded between -1 and 1
Interpretation	The sign indicates the direction of the relationship. The magnitude is not interpretable.	The sign indicates direction, and the magnitude indicates the strength of the relationship.
Effect of Scale	Highly dependent on the units and scale of the variables.	Scale-invariant. Changing the units of variables does not change the correlation.

In short

Correlation is a standardized version of covariance, making it much more useful for comparing the strength of relationships across different datasets.

Outlier Detection: Standard Deviation Approach

1. We first calculate the mean and standard deviation of the data. A data point is identified as an outlier if it is away from the mean by a pre-specified threshold in terms of the standard deviation.
2. That is, if a data point X_i satisfies calculated the Z score,

$$\frac{|X_i - \bar{X}|}{S} > k; \quad k = 1, 2, \text{ or } 3$$

   then it is detected as an outlier.
3. In other words, a data point is an outlier if it is beyond $\bar{X} \pm kS$. For a normally distributed dataset, 1S, 2S, and 3S represents, 68.27%, 95.45%, and 99.73%, respectively, of the dataset.
4. However, this method can fail to detect outliers if the S is large.

Outlier Detection: Median Absolute Deviation (MAD)

1. Median is a central measure of data that is less susceptible to outliers.
2. Median Absolute Deviation (MAD) is calculated as the median absolute difference between each point and the median as:

$$MAD = \text{median}(|X_i - M|), i = 1, 2, \ldots, n$$

3. The modified Z_M score is calculated using MAD values as

$$Z_M = \frac{0.6745(X_i - M)}{MAD}$$

4. As a rule of thumb, if Z_M is greater than 3, an outlier is detected.

Outlier Detection: Interquartile Approach

1. The interquartile range (IQR) is calculated the same way as the range.
2. IQR is computed by subtracting the first quartile from the third quartile:

$$IQR = Q_3 - Q_1$$

3. IQR can be used to detect outliers as follows: if any data point X_i lies outside the whiskers as given

$$Q_1 - 1.5IQR < X_i < Q_3 + 1.5IQR$$

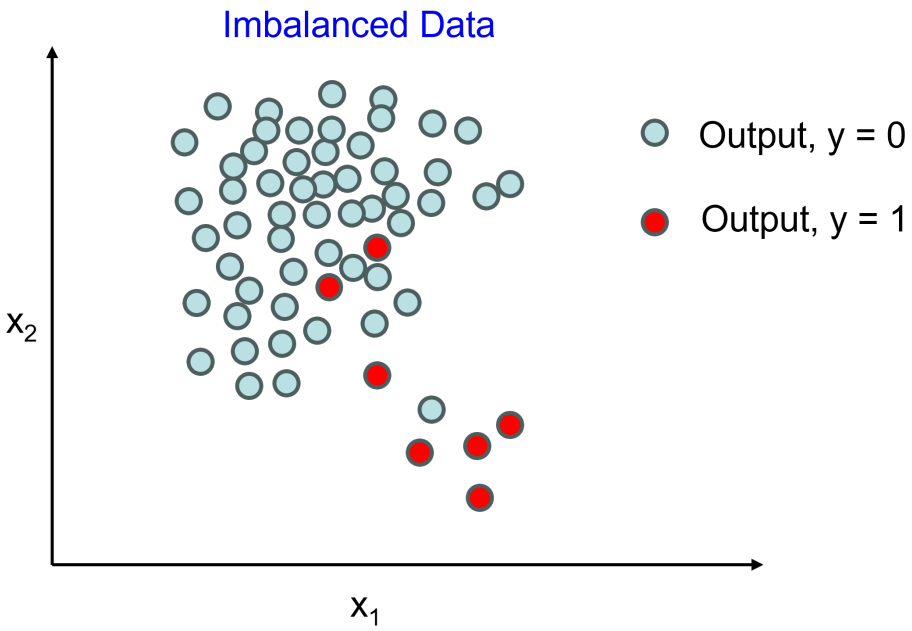
Then the data can be a potential outlier.

Data Augmentation

1. If the data set size is small or unbalanced, there would be difficulty in training a desired model.
2. This is because very little information can be extracted from small data sets, and data-driven modelling generally depends on sufficiently large data sets for information extraction.
3. To address this issue, data augmentation may be performed which generates and includes artificial datapoints to the dataset.
4. There are various approaches, such as the synthetic minority oversampling technique (SMOTE), the adaptive synthetic sampling method (ADASyn), to enable data augmentation. These methods oversample or artificially synthesise data from smaller sample data sets, sufficient for training.
5. To achieve this, SMOTE uses a strategy wherein a line is generated between the neighbours and generates random points on the line.
6. ADAsyn algorithm works on the data generated by SMOTE by also considering uncertainties.

Data Augmentation

Imbalanced Data

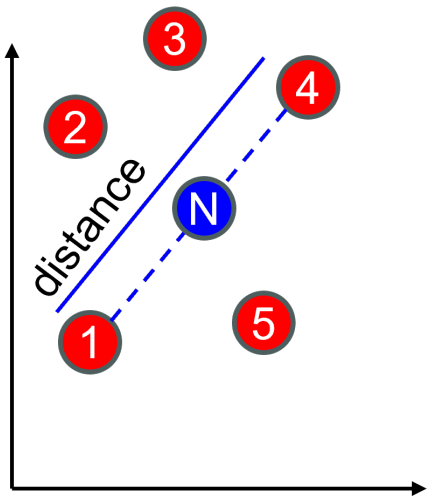


Typical Examples: fraud detection (e.g., insurance fraud), spam filtering, and rare disease discovery. In these, the positive class data are rare, whereas the negative class data are excessively represented.

1. Under-sampling: Remove some of the data $\rightarrow$ Not recommended as we lose some information.
2. Over-sampling: New data is generated $\rightarrow$ Recommended.

Synthetic Minority Oversampling Technique (SMOTE)

1. First the number of oversampling observations, N, need to be generated is set up. Ideally, these are selected in such a way that data becomes balanced (i.e., in binary class the distribution is 1:1). However, in some cases this number could be lower too.
2. Next an instance of a positive class is selected randomly and its k nearest neighbors (default = 5) are selected.
3. At last, N of these K instances is chosen to interpolate new synthetic instances.
4. For this purpose, any distance metric representing the difference in distance between the feature vector and its neighbors is calculated.
5. Now, this difference is multiplied by any random value in [0,1] and is added to the previous feature vector.



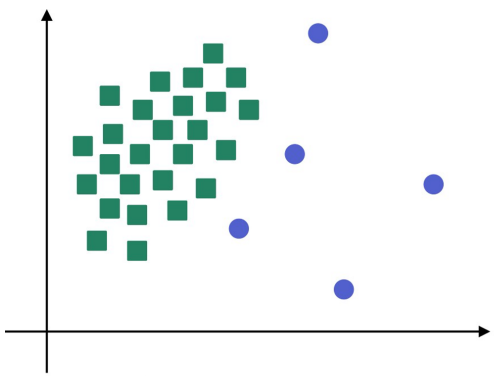
Adaptive Synthetic Sampling (ADAsyn)

1. It is a generalized version of SMOTE.
2. Generate more data for minority examples that are harder to learn.
3. Hardness is measured by the local class imbalance.
4. It adaptively change the decision boundaries based on the samples difficult to learn.
5. **Find its k-Nearest Neighbors:**
   Calculate the distances to all other points in the **entire training set** and identify the k points closest to $\mathbf{x}_i$.
6. **Count Majority Neighbors:**
   Examine the class labels of these k neighbors. Count how many of them belong to the **majority class**. Let this count be N_{maj}.
7. **Calculate the Ratio r_i:**
   Compute the ratio using the formula:

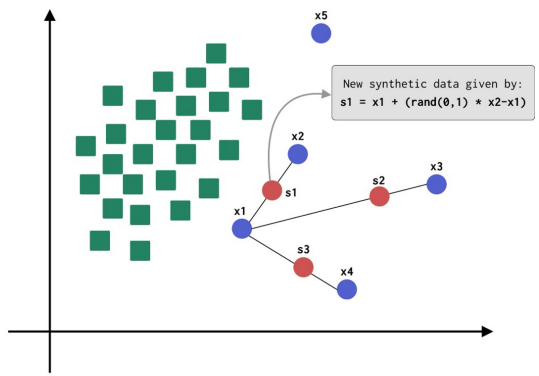
$$r_i = \frac{N_{\text{maj}}}{k}$$

8. A high r_i value (close to 1) means the point is surrounded by majority instances and is **hard to classify**.
9. A low r_i value (close to 0) means the point is in a safe, dense minority region and is **easy to classify**.

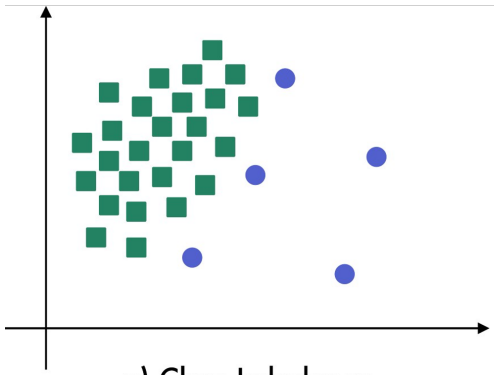
Comparing Oversampling Techniques



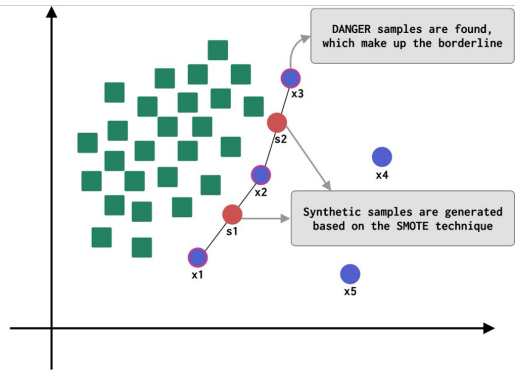
a) Class Imbalance



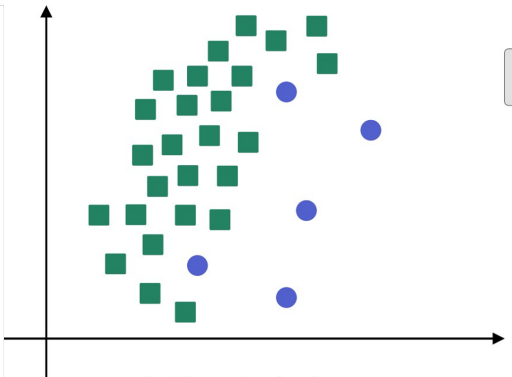
b) SMOTE



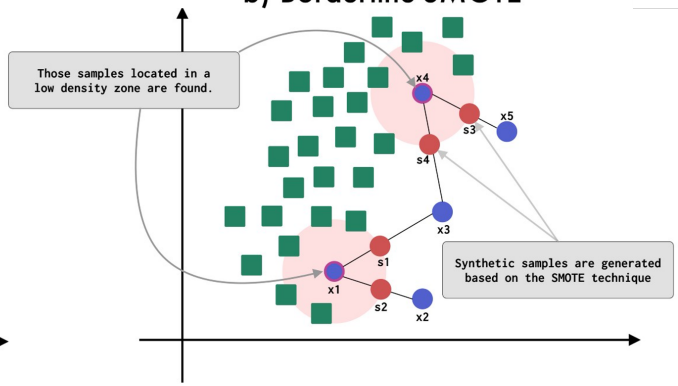
a) Class Imbalance



b) Borderline-SMOTE



a) Class Imbalance



b) ADASYN

Thank You