Data Mining

Data: Part 2

Exploratory Data Analysis (EDA)

Mohammed Brahimi & Sami Belkacem

Exploratory Data Analysis (EDA)

- EDA is a set of **statistical** and **visualization** techniques.

- Used for seeing what the data can tell us before the **preprocessing** and **modeling**:
  - Understand the data and summarize its keys properties.
  - Discover noisy data and outliers.
  - Comprehend the distribution of the data.
  - Decide which set of data cleaning techniques to be applied.

- EDA is cross-classified in two ways:
  1. The method is either **non-graphical** or **graphical**.
  2. The method is either **univariate** or **multivariate** (usually just bivariate).

Exploratory Data Analysis (EDA)

1. Statics of Data

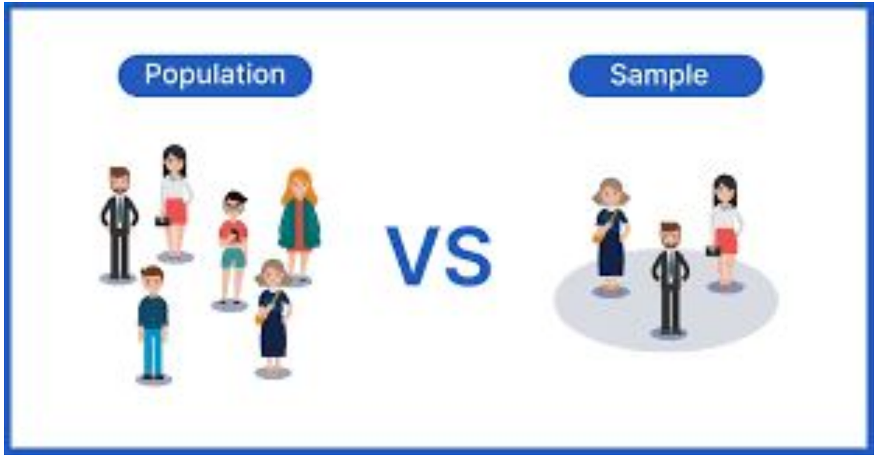
2. Data visualization

1: Statics of Data

- Population vs Sample
- Measuring the Central Tendency
  - Mean, Median, and Mode
- Measuring the distribution of data
  - Variance and Standard deviation
- Analysis of two variables
  - Covariance and Correlation

Population vs Sample

- **Population**
  - The entire group that you want to draw conclusions about.
- **Sample**
  - Subset of the population, used when the population size is too large to analyze
  - A sample is an unbiased subset that best represents the entire population.



Measuring the Central Tendency: (1) Mean

- <u>Mean (algebraic measure) (sample vs. population):</u>

  n is sample size and ***N*** is population size.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mu = \frac{\sum x}{N}$$

- **Weighted arithmetic mean:**

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- **Trimmed mean:** Chopping extreme values (e.g., olympics gymnastics score computation)

Measuring the Central Tendency: (2) Median

- **Median** is the middle value in a data set when values are ordered.
- **How to calculate**
  - Sort data in ascending order.
  - Repeat values according to their frequency.
  - If there's an odd number of data points, the median is the middle value.
  - If there's an even number of data points, the median is the average of the two middle values.
- **Why use the Median**
  - Resistant to extreme outliers.
  - Useful for skewed distributions.

Example1: Calculating the median

Number of cars	Frequency
0	4
1	5
2	3
3	1

0	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**3**

- N = 4+5+3 +1 = 13
- Median is in the position **ceil(13/2)= 7**
- **The median is the value 1**

Example 2: Calculating the median

	Age	Frequency
0	5-10	4
1	10-20	5
2	20-50	3
3	50-55	1

0	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**2**	**2**	**2**	**3**

- N = 4+5+3 +1 = 13
- Median is in the position **ceil(13/2)= 7**
- **The median bin is the value 10-20**

How to find the median value ?

Median calculation for grouped data

$$\textbf{Estimated Median} = L + \frac{(n/2) - B}{G} \times w$$

- **L :** the lower class boundary of the median bin.
- **n:** the total number of values.
- **B**: cumulative frequency of the bins before the median bin.
- **G**: the frequency of the median bin.
- **W**: the group width.

	Age	Frequency
0	5-10	4
1	10-20	5
2	20-50	3
3	50-55	1

Example 2: Calculating the median for grouped data

$$\text{Estimated Median} = L + \frac{(n/2) - B}{G} \times w$$

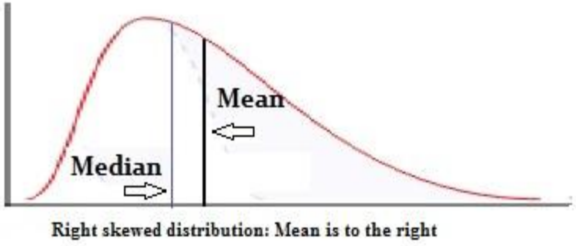
	Age	Frequency
0	5-10	4
1	10-20	5
2	20-50	3
3	50-55	1

- **The median bin is the value 10-20**
  - L = 10
  - n=13
  - B =4 and G = 5
  - w = 20-10 =10

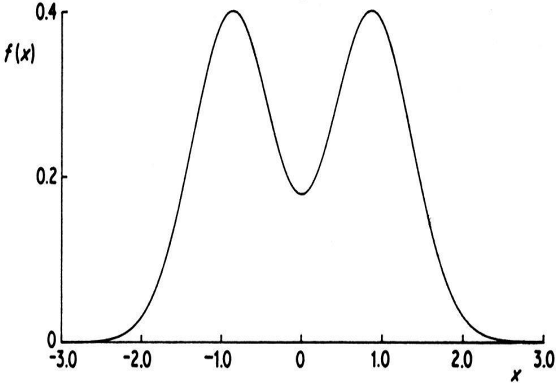
Median value = 10 + ((6.5-4)/5) *10 = 15

Measuring the Central Tendency: (3) Mode

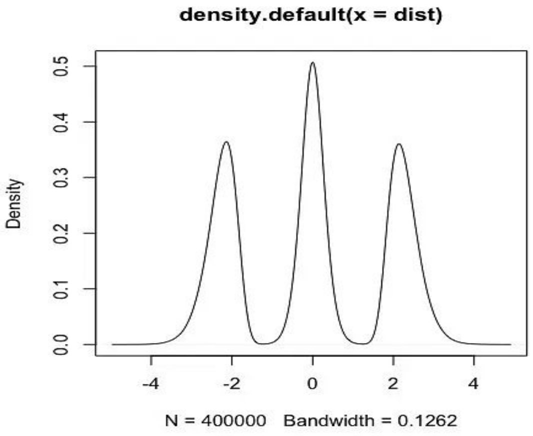
- **Mode:** Value that occurs most frequently in the data
- **Unimodal**
  - Empirical formula: $mean - mode = 3 \times (mean - median)$



- **Multimodal**
  - Bimodal (a)
  - Trimodal (b)



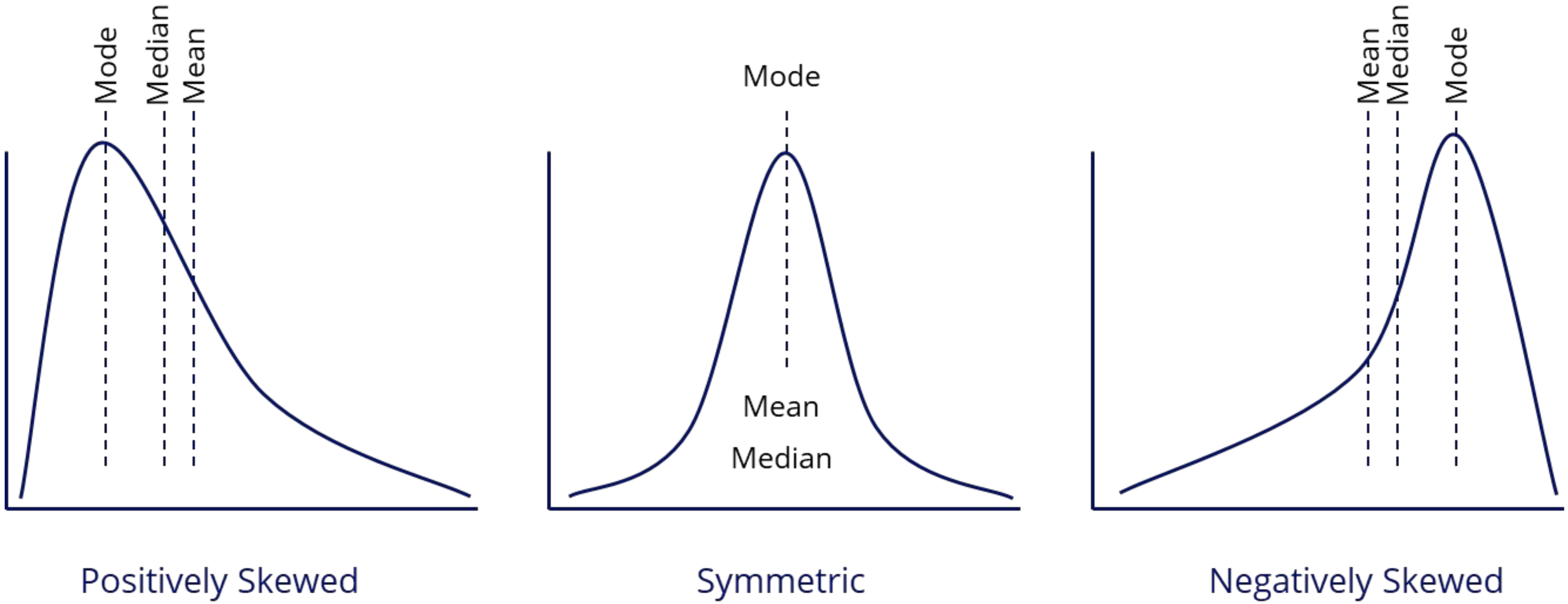
(a)



(b)

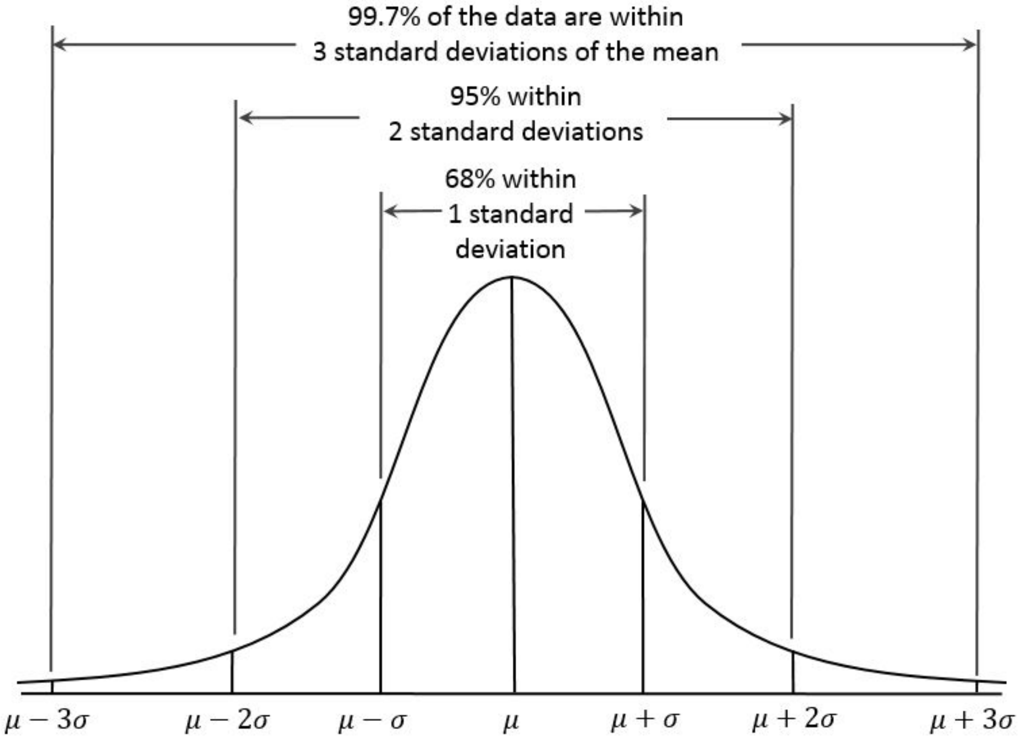
Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data



Properties of Normal Distribution Curve

←— —————Represent data dispersion, spread — —————→



Represent central tendency

Measuring Data Distribution: Variance and Standard Deviation

- *sample: **s**, population: **σ***
  - **Variance**: Measure dispersion around the mean

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right]$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^2 - \mu^2$$

Note: The subtle difference of formulae for sample vs. population
- **n** : the size of the sample
- **N** : the size of the population

  - **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or σ^2)*
  - Measures dispersion around the mean, but in the same units as the

Covariance for Two Numerical Variables

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

- The goal is to understand relationships between two variables.
- Covariance shows how two variables change together.
  - **Positive Covariance:** both variables move together.
  - **Negative Covariance:** variables move in opposite directions.
  - **Zero Covariance:** bo clear pattern in variable movements.

Covariance for Two Numerical Variables

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

- Understanding relationships between variables.
- Covariance shows how two variables change together.

Covariance is useful but sensitive to scale, while correlation addresses this issue by standardizing the measurement.

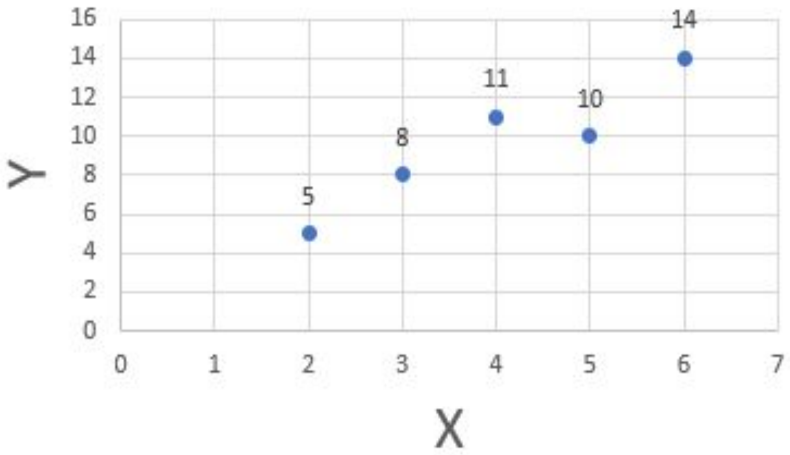
Covariance for Two Numerical Variables

$$\mathrm{cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$$

- Two Stocks - X1 and X2 Weekly Stock Prices:

  X: (2, 3, 5, 4, 6)   Y: (5, 8, 10, 11, 14)

- Calculate the Covariance
  - E(X) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4
  - E(Y) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6
  - E(XY) = (2×5 + 3×8 + 5×10 + 4×11 + 6×14)/5 = 42.4
- Cov(X,Y) = 42.4 - 4*9.6 = 4
- Thus, X and Y rise together since Cov(X,Y)>0



Covariance Matrix

- The variance and covariance information for the two variables can be summarized as 2X2 covariance matrix as:

$$\Sigma_{i,j} = \begin{bmatrix} \sigma_{ii}^2 & \sigma_{ij}^2 \\ \sigma_{ji}^2 & \sigma_{jj}^2 \end{bmatrix}$$

- Generalizing it to ***d*** dimensions:

$$\mathbf{\Sigma} = E[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T] = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix}$$

Correlation between Two Numerical Variables

- **Correlation** between two variables X_1 and X_2 is the standard covariance, obtained by normalizing the covariance with the standard deviation of each variable

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^{\,2} \sigma_2^{\,2}}}$$

- **Sample correlation** for two attributes X_1 and X_2: $\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^{n}(x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^{n}(x_{i2} - \hat{\mu}_2)^2}}$

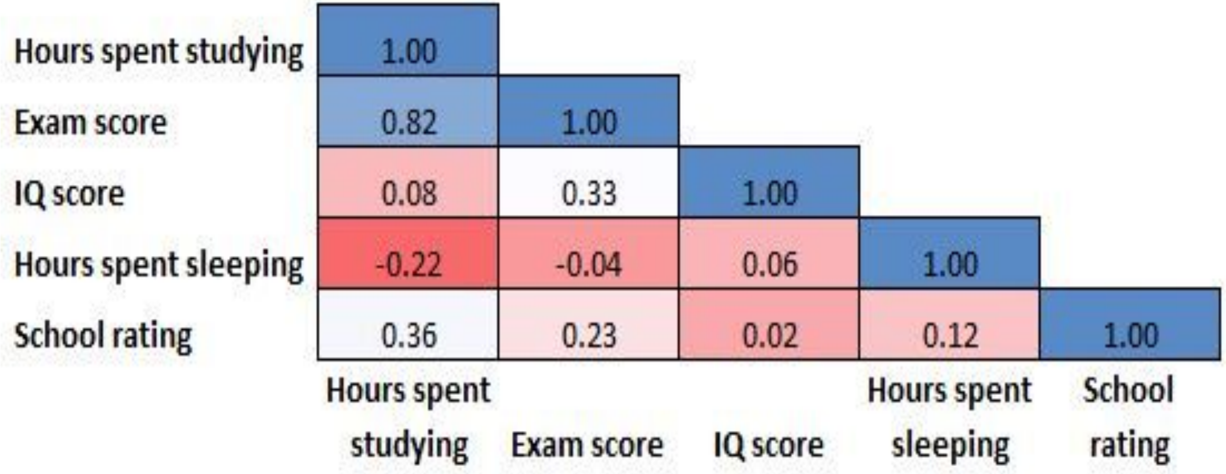
n is the number of tuples, $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ are the respective means of $\mathbf{X_1}$ and $\mathbf{X_2}$,

$\boldsymbol{\sigma_1}$ and $\boldsymbol{\sigma_2}$ are the respective standard deviation of X_1 and X_2

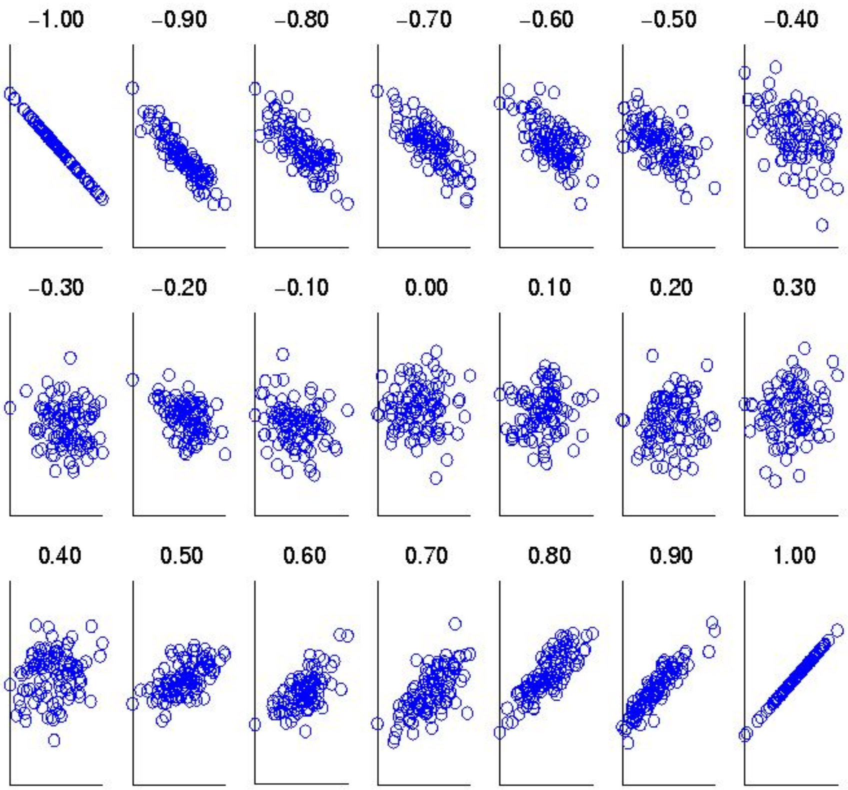
- **If $\boldsymbol{\rho_{12} > 0}$:** A and B are positively correlated (X_1's values increase as X_2's)
- **If $\boldsymbol{\rho_{12} = 0}$:** independent (under the same assumption as discussed in co-variance)
- **If $\boldsymbol{\rho_{12} < 0}$:** negatively correlated

Correlation Matrix (Correlation Heatmap)

- The correlation matrix is a matrix that shows the correlations between each pair of variables in a dataset



Visualizing Changes of Correlation Coefficient



- Correlation coefficient value range: [−1, 1]
- A set of scatter plots shows sets of points and their correlation coefficients changing from −1 to 1

Exploratory Data Analysis (EDA)

1. Statics of Data

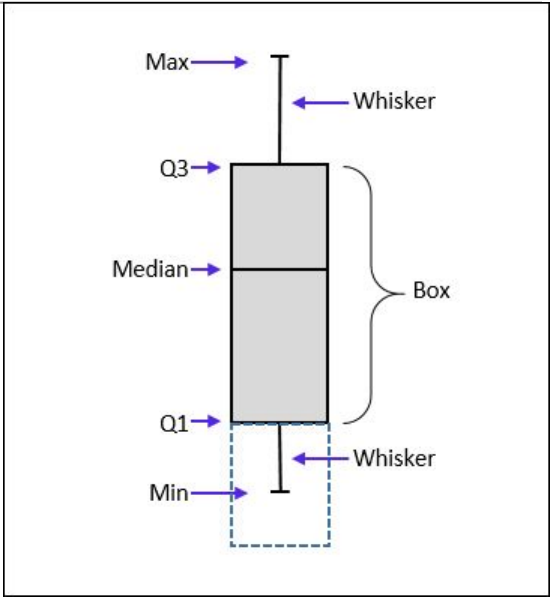
2. **Data visualization**

2- Data Visualization

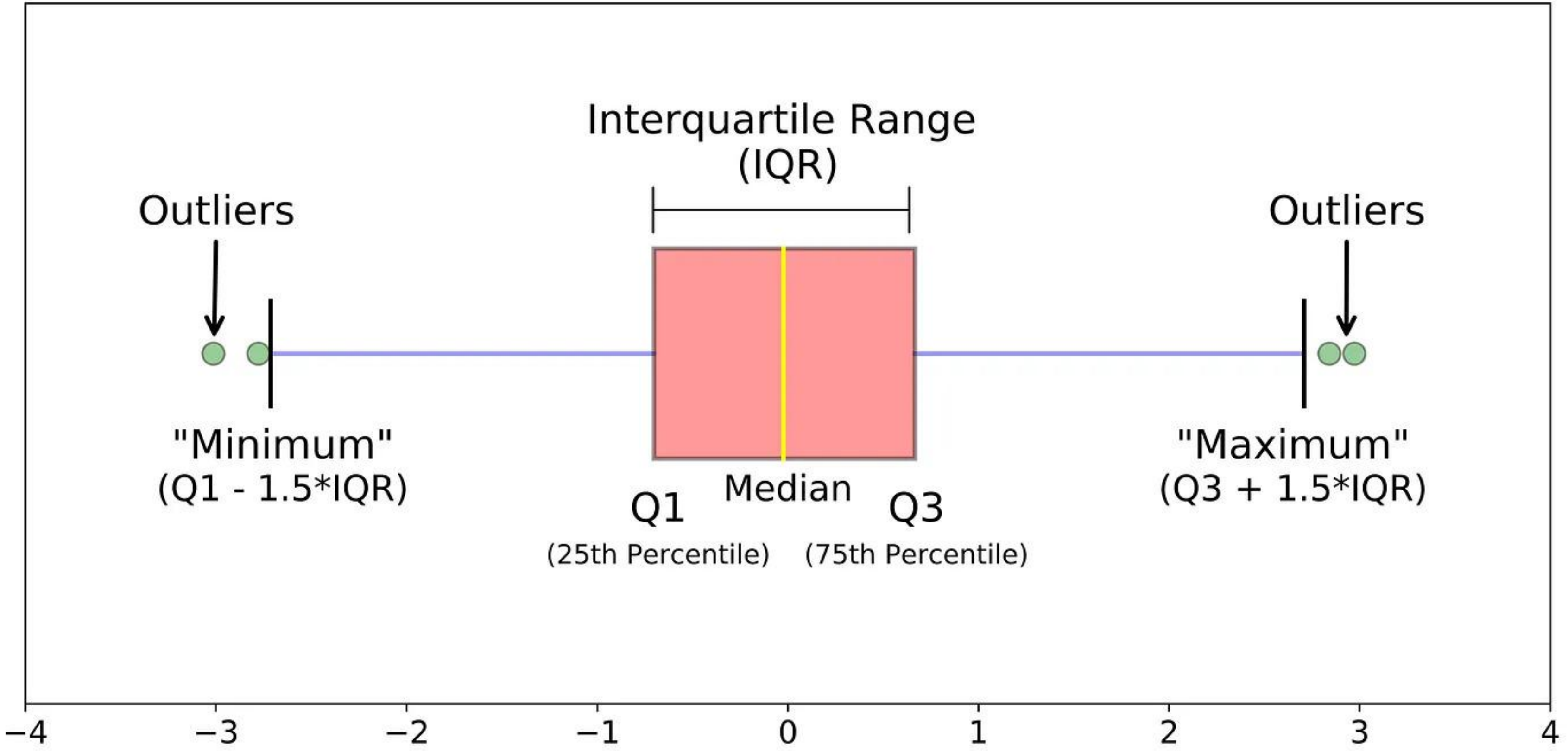
- Boxplot
- Histogram and Bar chart
- Quantile plot
- Quantile-quantile (Q-Q) plot
- Scatter plot
- Line chart
- Parallel Coordinates plot

Measuring the Dispersion of Data: Quartiles & Boxplots

- **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
- **Interquartile range**: IQR = $Q_3 - Q_1$
- **Five number summary**: min, Q_1, median, Q_3, max
- **Box plot**: Data is represented with a box
  - **Q_1, Q_3, IQR**: The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - **Median (Q_2)** is marked by a line within the box
  - **Whiskers:** two lines outside the box extended to Minimum and Maximum
  - **Outliers:** points beyond a specified threshold (e.g. value higher/lower than 1.5 x IQR)

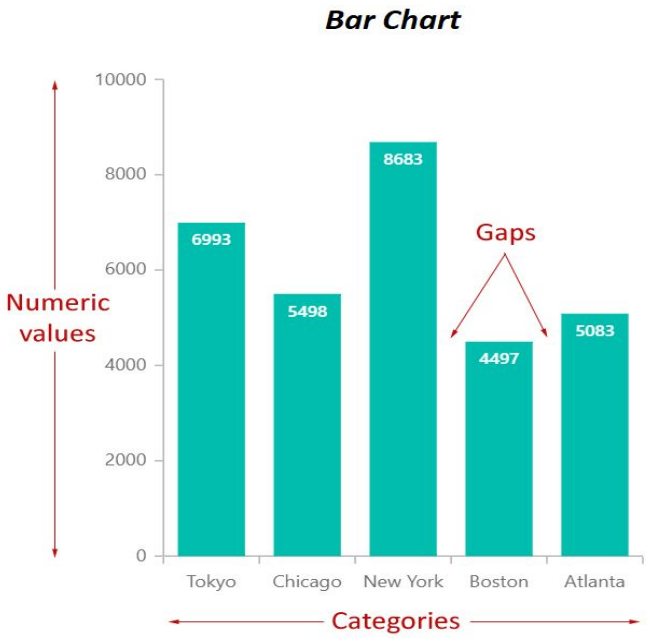


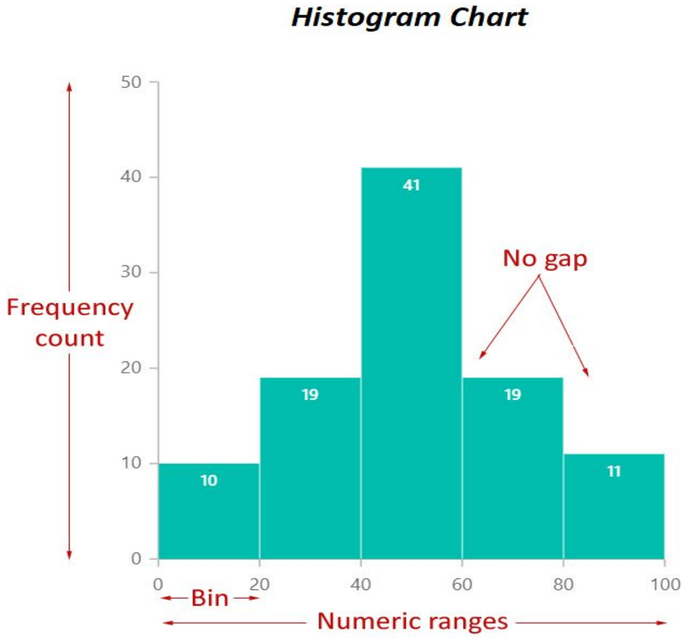
Measuring the Dispersion of Data: detect Outliers



Histograms vs Bar charts

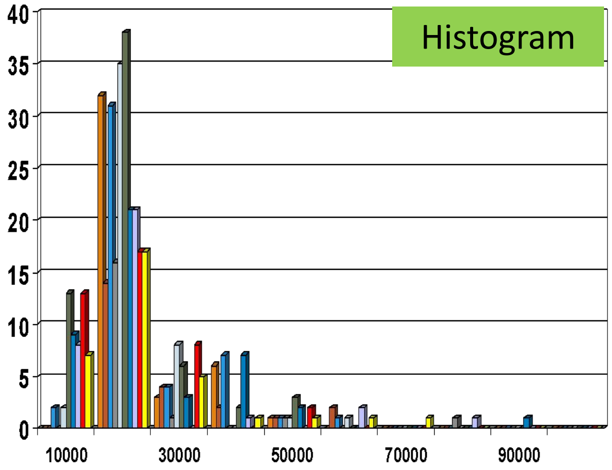
- **Histogram:** Tabulated frequencies represented by bars.
- **Bar chart:** Categorical data with bars proportional to the values they represent.

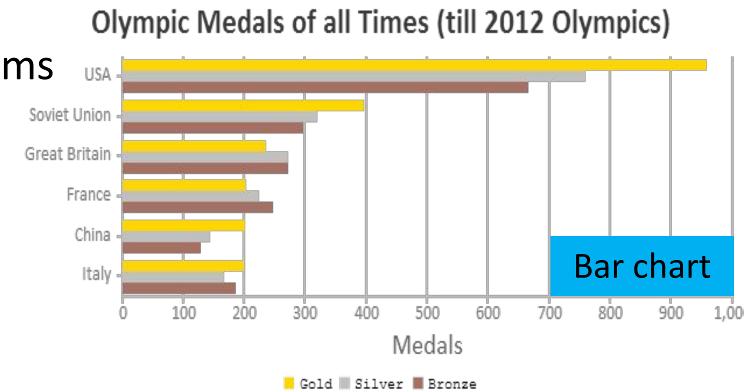




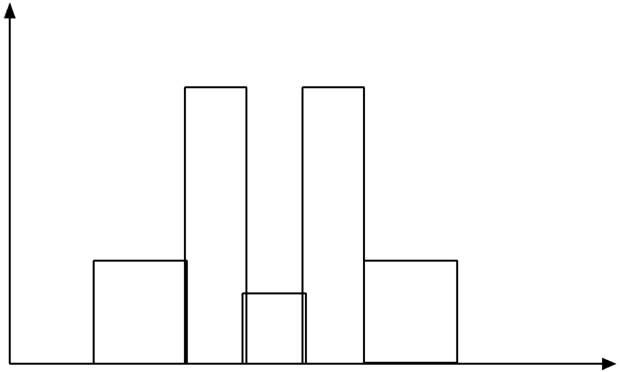
Differences between Histograms and Bar charts:

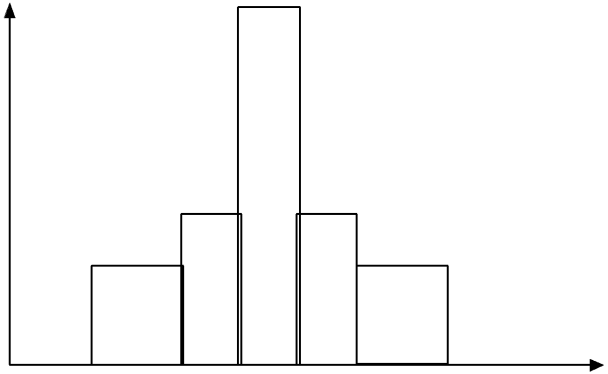
- Histograms show **distributions** of variables, while bar charts compare **variables**
- Histograms plot binned **quantitative/categorical** data, while bar charts only plot **categorical** data
- Bars can be **reordered** in bar charts, but not in histograms
- In histograms, it is the area of the bar that denotes the value, not the height as in bar charts.





Histograms Often Tell More than Boxplots

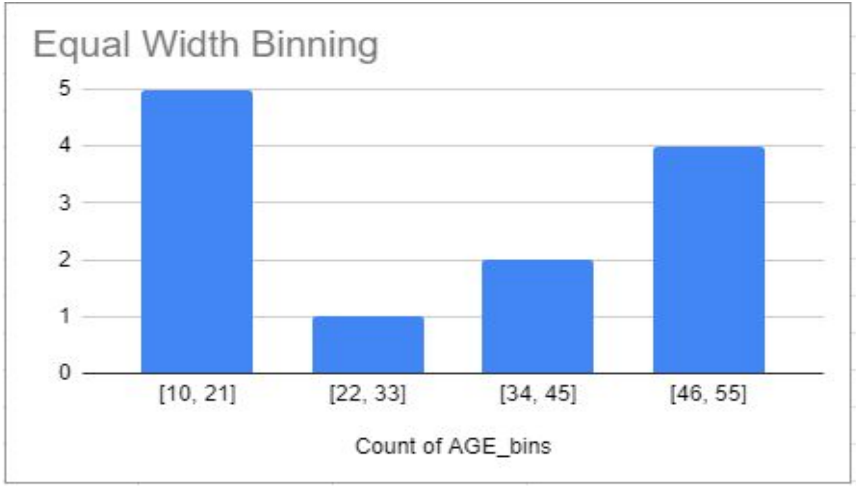


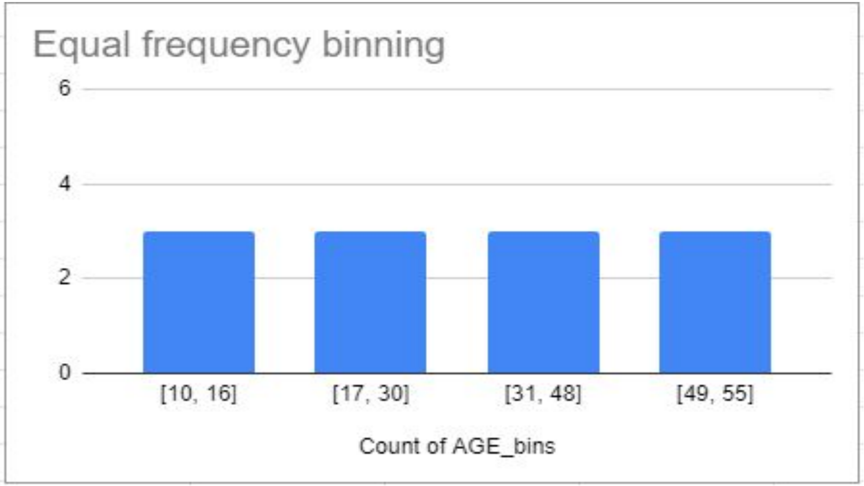


- The two histograms shown in the left may have the same box plot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

Histogram Analysis

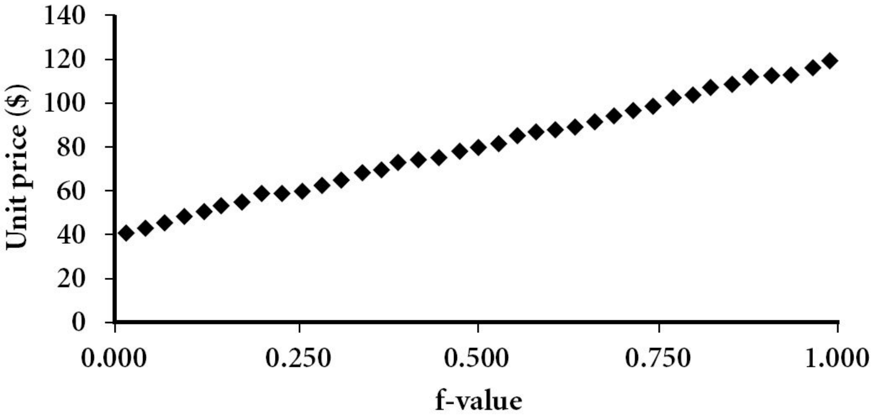
- Divide data into buckets and store average (sum) for each bucket.
- Partitioning rules:
  - **Equal-width:** equal bucket range
  - **Equal-frequency** (or equal-depth)





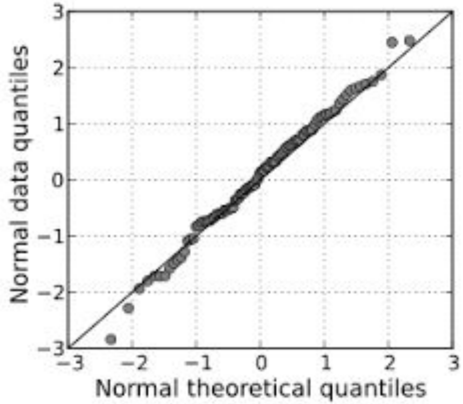
Quantile Plot

- **Purpose**: Visualizes all quantile information for a specific attribute
- **Benefits**
  - Provides a comprehensive view of the attribute's distribution.
  - Helps identify both general trends and outliers.
- **Construction**
  - For a data {X1, X2, …, XN} sorted in increasing order.
  - f_i indicates that approximately f_i of the data point have values ≤ x_i

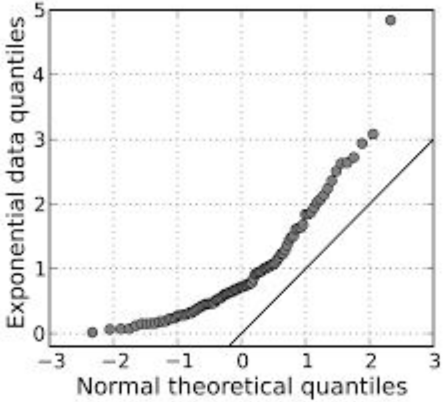


Quantile-Quantile (Q-Q) Plot

- **Purpose**
  - Assess the distributional similarity between an attribute and either **another attribute** or a **theoretical distribution.**

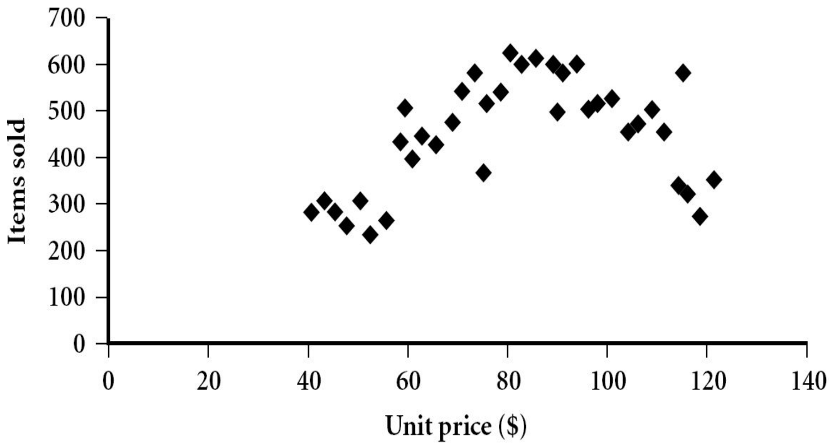


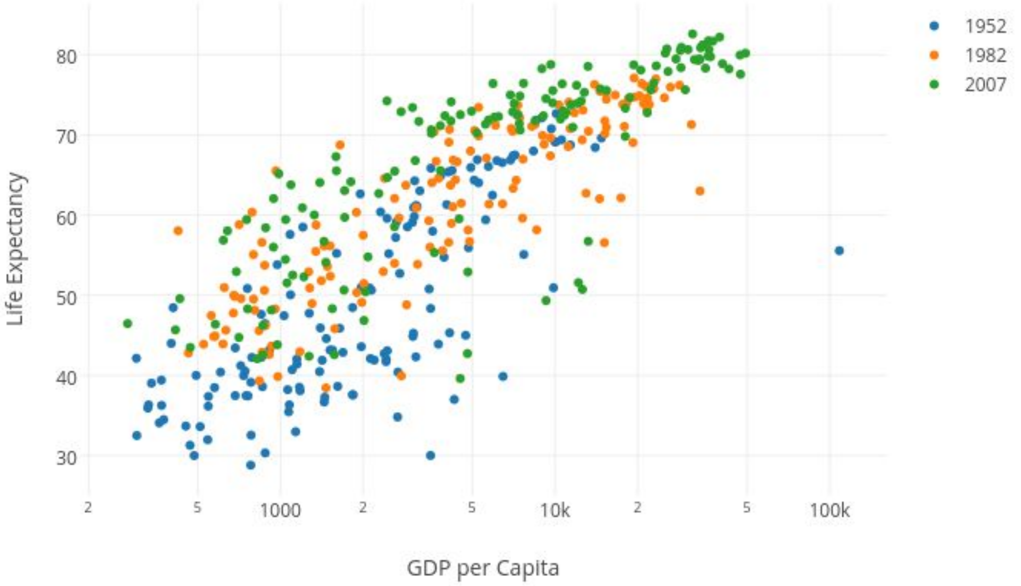
- **Interpreting a Q-Q Plot**
  - If the points closely **follow a straight line** ⇒ The two distributions are similar.
  - Deviations from the line indicate differences in distribution.



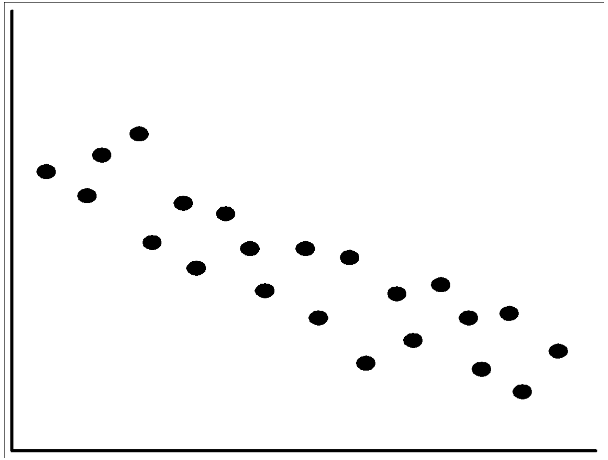
Scatter plot

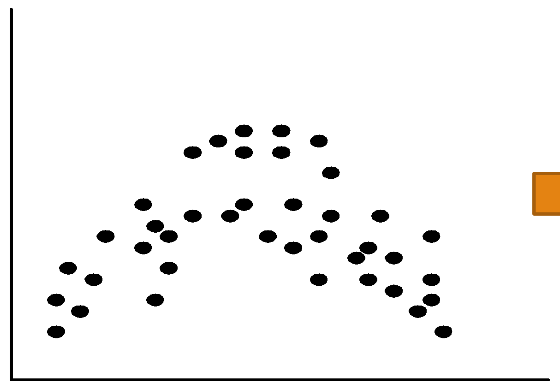
- Provides a first look at bivariate data to see clusters of points, outliers, etc.
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane.





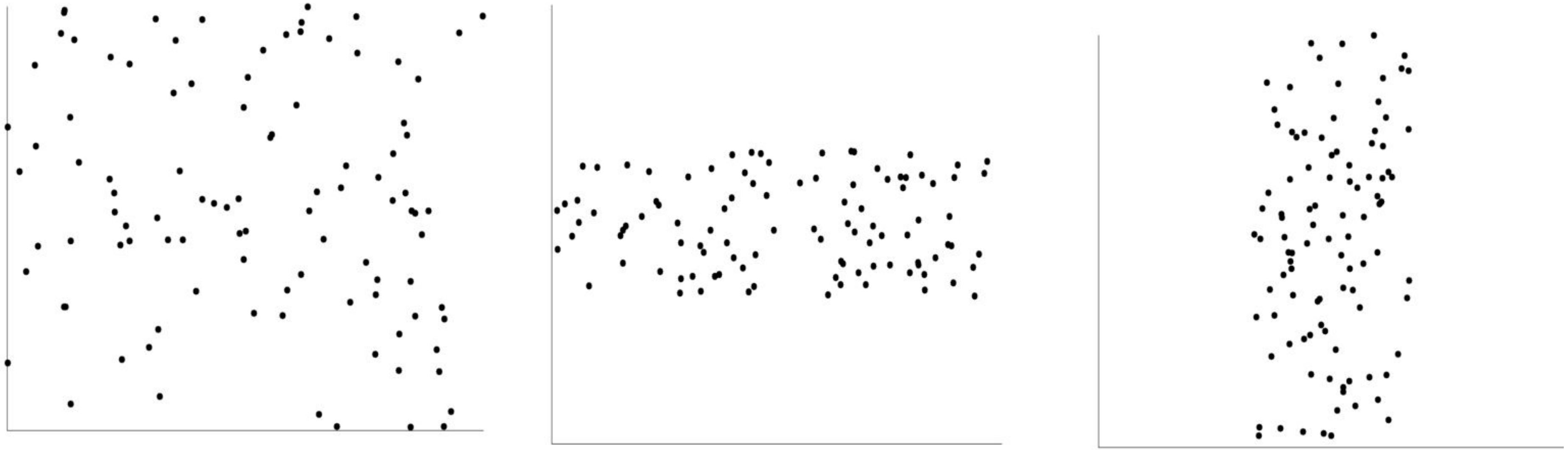
Positively and Negatively Correlated Data





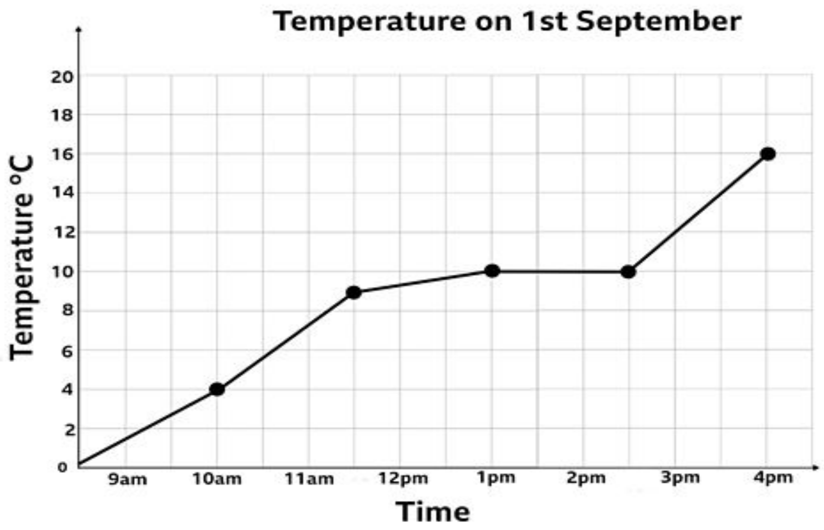
- The left half fragment is positively correlated
- The right half is negative correlated

Uncorrelated Data



Line chart

- A line chart displays information as a series of data points called 'markers'.
- The markers are connected to each other by straight line segments



Parallel Coordinates Plots of Iris Data

- A parallel coordinate plot maps each row in the data table as a line, or profile.
- Each attribute of a row is represented by a point on the line.