Gaussian Mixture Model (GMM) Clustering

Renato R. Maaliw III, *DIT*
College of Engineering
Southern Luzon State University
Lucban, Quezon, Philippines

What is a Gaussian (Normal) Distribution?

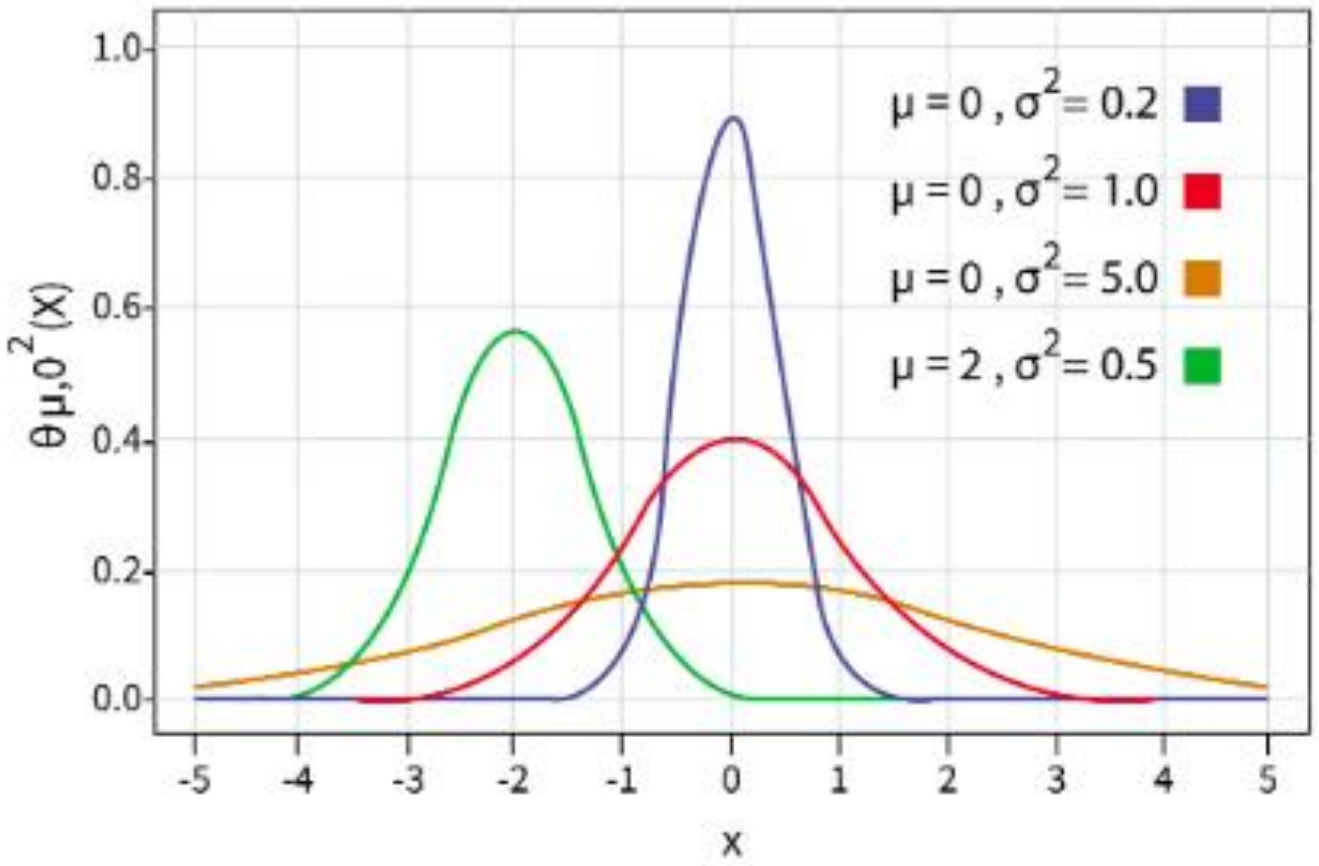
A Gaussian distribution, also known as a normal distribution, is a **bell-shaped curve** that is defined by two key parameters:

1. **Mean (μ)**: This is the center of the bell curve (where most of the data is).
2. **Variance (σ^2)**: This describes how spread out the data is around the mean (how wide the bell is).

Mixture of Gaussians

Instead of assuming that all the data comes from one bell-shaped curve (one Gaussian distribution), a **Gaussian Mixture Model** assumes that the data is generated from a mixture of several bell curves (each representing a different group or cluster).

Cognate/Professional Electives



For example, if we have three clusters of data, we assume there are three **different Gaussian distributions**, each with its own mean and variance. The GMM will try to figure out:

1. How many Gaussian distributions (clusters) there are.
2. The mean and variance for each distribution.
3. How likely it is that a given data point belongs to each distribution.

For example, if we have three clusters of data, we assume there are three **different Gaussian distributions**, each with its own mean and variance. The GMM will try to figure out:

1. How many Gaussian distributions (clusters) there are.
2. The mean and variance for each distribution.
3. How likely it is that a given data point belongs to each distribution.

How GMM Clustering Works?

1. Initial Guess:

The GMM starts with a **random guess** about how many clusters (Gaussians) there are and what their parameters (means and variances) are.

2. Expectation Step:

For each data point, it calculates the probability that it belongs to each of the clusters (each Gaussian distribution).

This is called a **"soft assignment,"** meaning a point might belong to multiple clusters with different probabilities.

3. Maximization Step:

Based on those **probabilities**, the model updates its estimates of the parameters (mean and variance) for each Gaussian to better fit the data.

4. Repeat:

The E-step and M-step repeat **until the model converges** (i.e., stops improving).

Difference Between GMM and K-Means

K-Means clustering assigns each data point strictly to one cluster (*hard clustering*).

GMM gives a probability that a data point belongs to each cluster (*soft clustering*).
So, a point can belong to multiple clusters, but with different probabilities.

Key Points:

1. GMM is useful when clusters are not necessarily spherical, or when there is overlap between clusters.

Key Points:

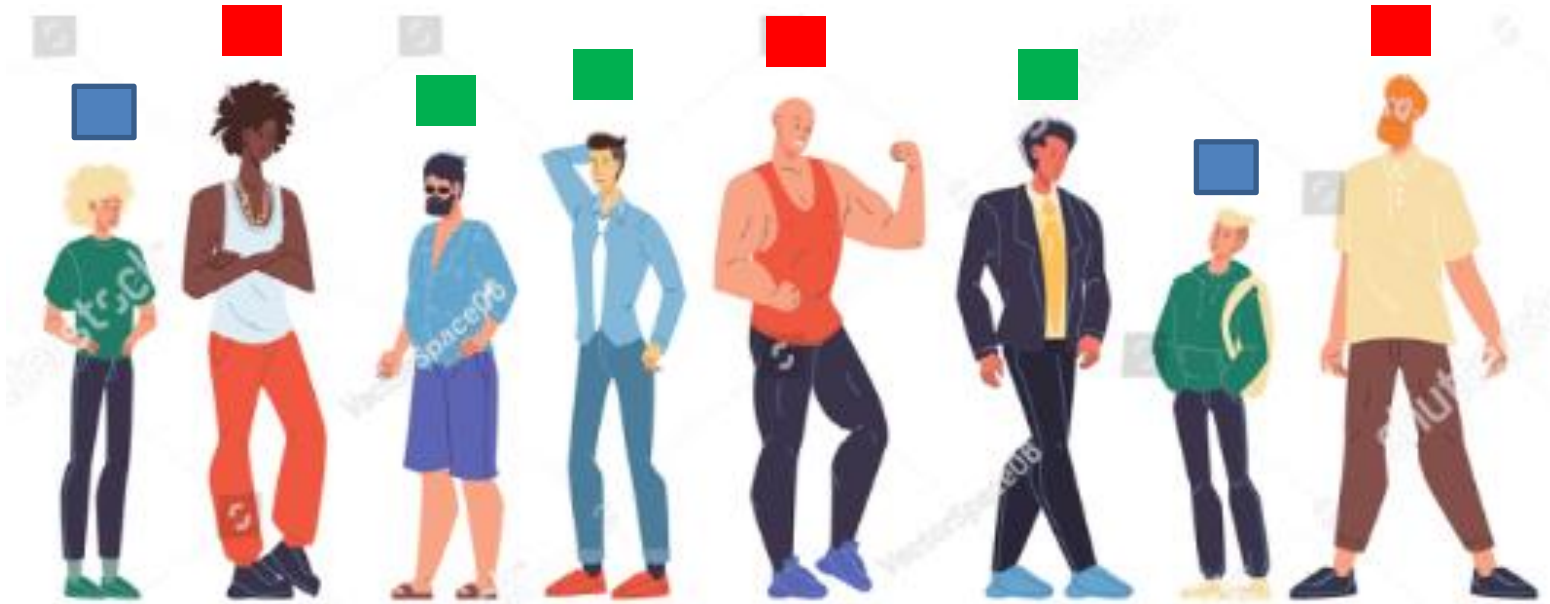
2. It models data as a combination of several Gaussian distributions, allowing for more flexibility than simpler methods like K-Means.

A Simple Explanation

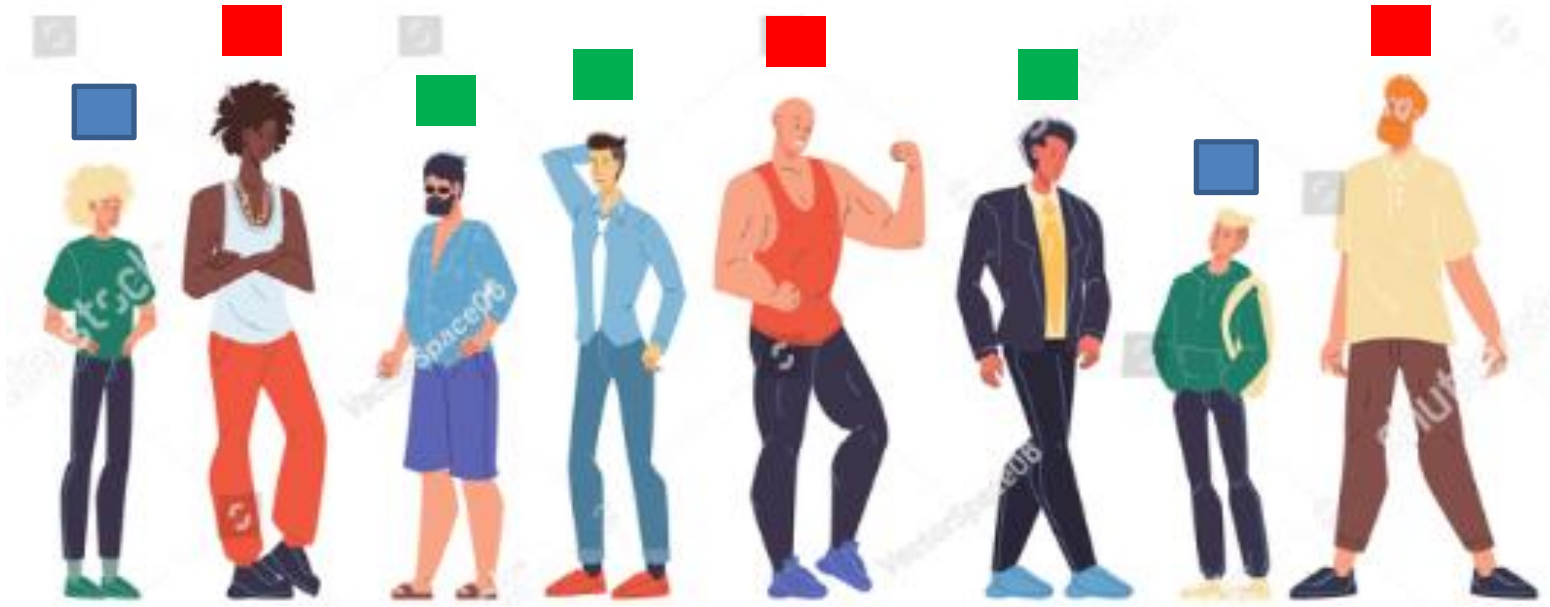
Imagine you have a **group of friends**, and you know that some of them are from different cities. You want to figure out which city each friend belongs to, but no one has told you their city.

1. Clusters (Groups): GMM assumes that each group of friends (city) can be described by a pattern, like the average height and weight of people in that city.

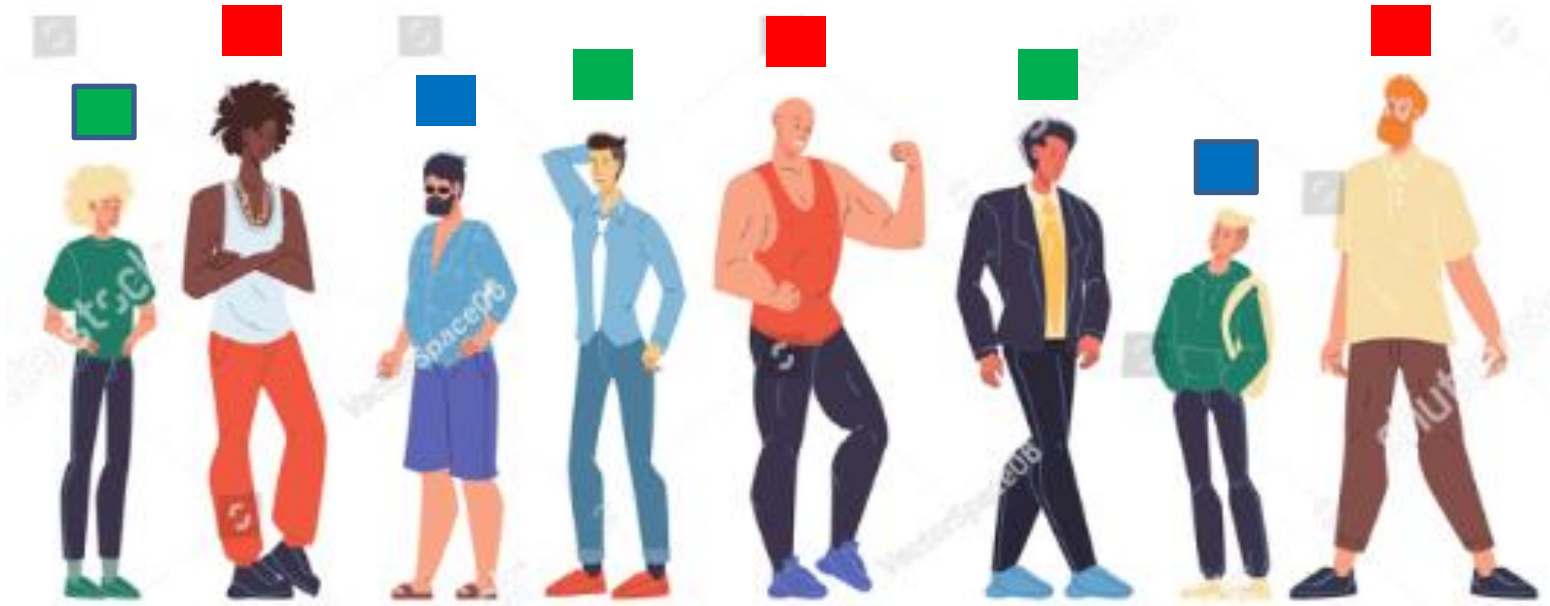
These patterns are shaped like **bell curves** (also called a **Gaussian distribution**). This means most people in a group will have heights and weights near the average, but some will be taller or shorter.



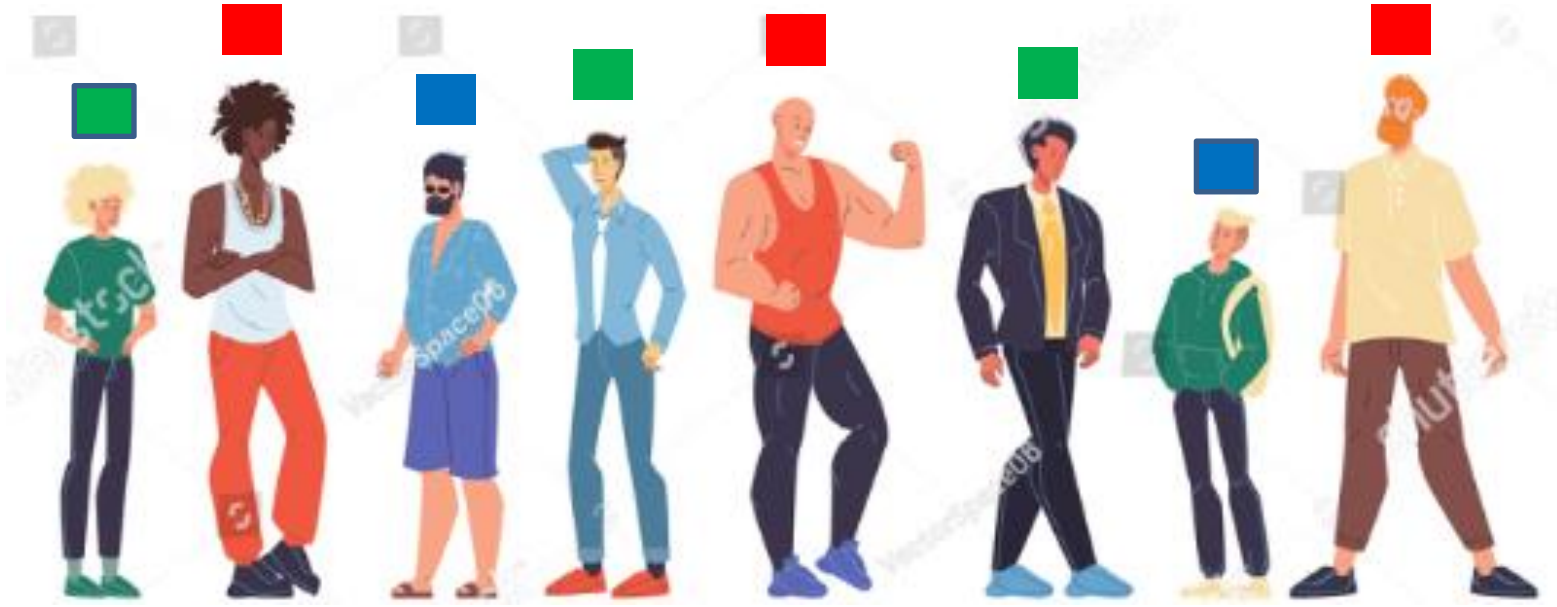
2. Guessing the Groups: Since we don't know the exact city (group) for each friend, GMM will try to **guess** by looking at their heights and weights.



It will try different ways to divide your friends into groups, checking if they fit the **bell curve pattern** for each group.

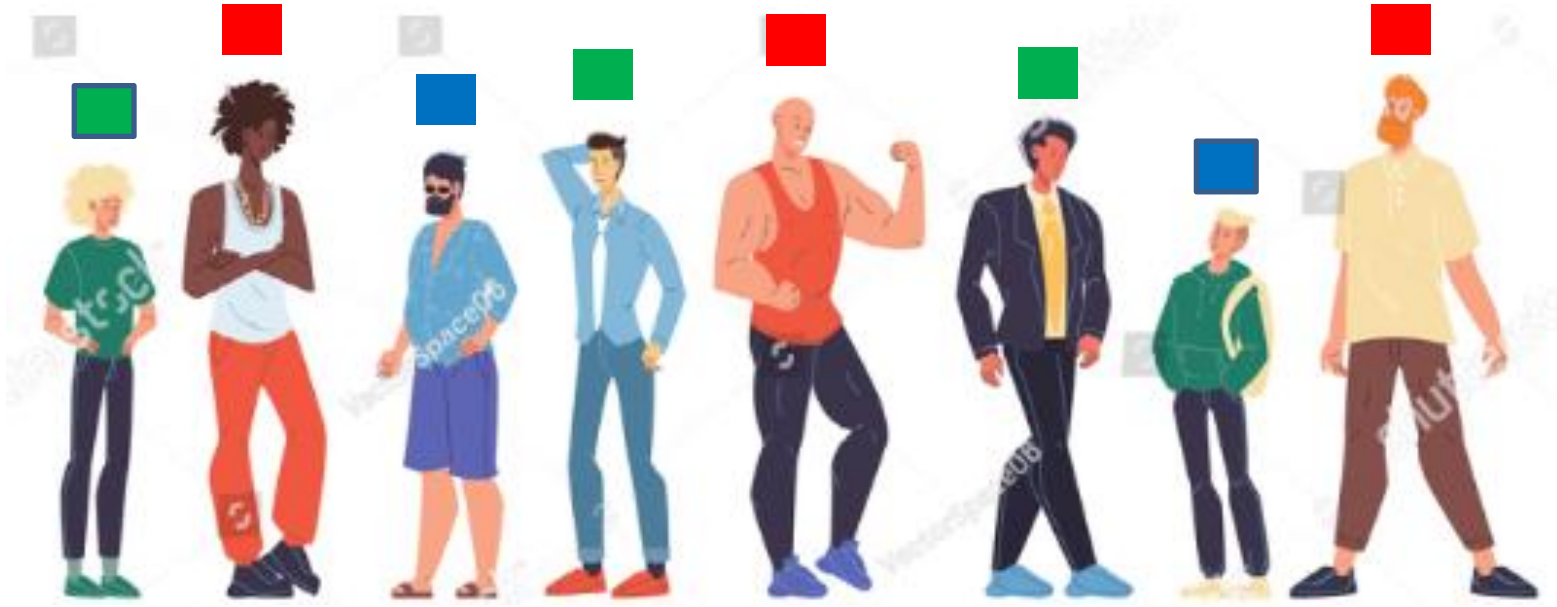


3. Probabilities: Instead of making a hard decision, GMM says, "I think this person is 70% from City A and 30% from City B," meaning each person can have a **probability** of belonging to different groups.



City C (60%) City A (40%)	City B (80%) City C (20%)	City A(90%) City B (10%)	City C (60%) City A (40%)	City B (75%) City C (25%)	City C(85%) City A (15%)	City A (70%) City C (30%)	City B (95%) City A(5%)

This is called **soft clustering** because we aren't forcing someone into just one group — they can belong to more than one group with certain probabilities.



City C (60%)
City A (40%)

City B (80%)
City C (20%)

City A(90%)
City B (10%)

City C (60%)
City A (40%)

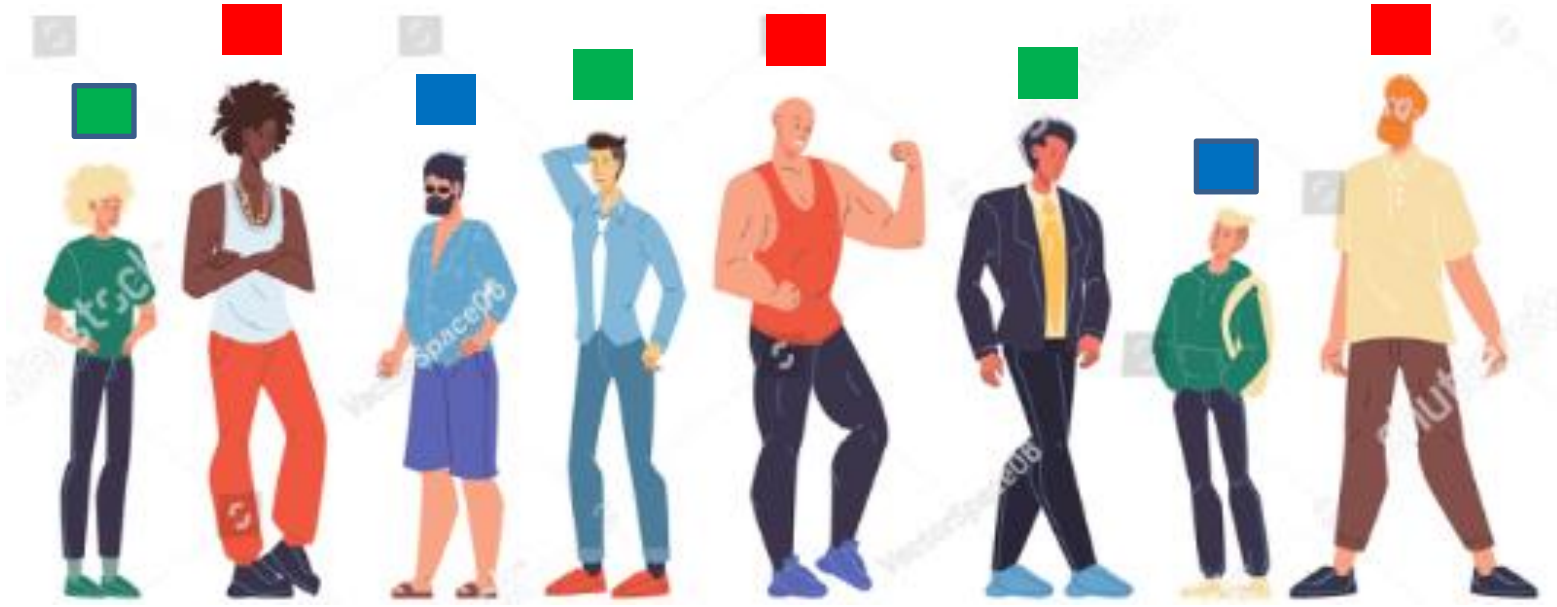
City B (75%)
City C (25%)

City C(85%)
City A (15%)

City A (70%)
City C (30%)

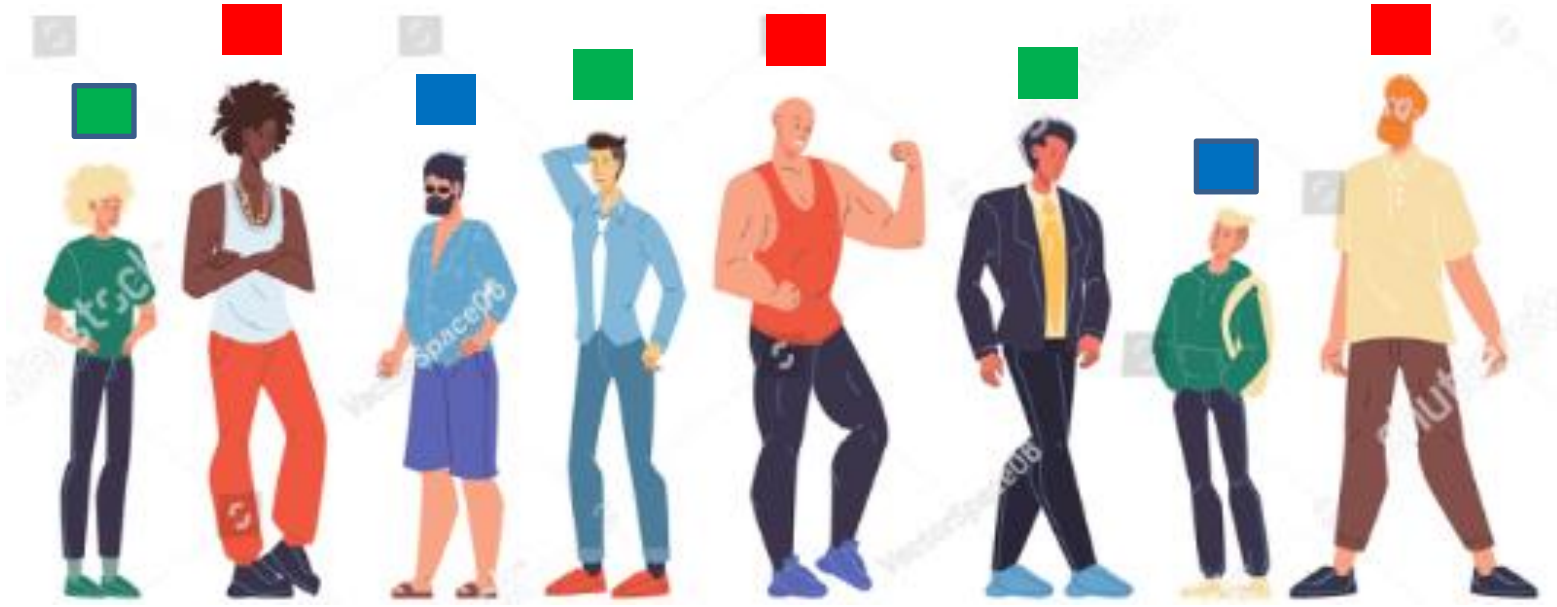
City B (95%)
City A(5%)

4. Improving the Guess: GMM starts with a random guess for the groups. Then it tries to improve by adjusting the groups step by step. It keeps doing this until the groups fit the data well.



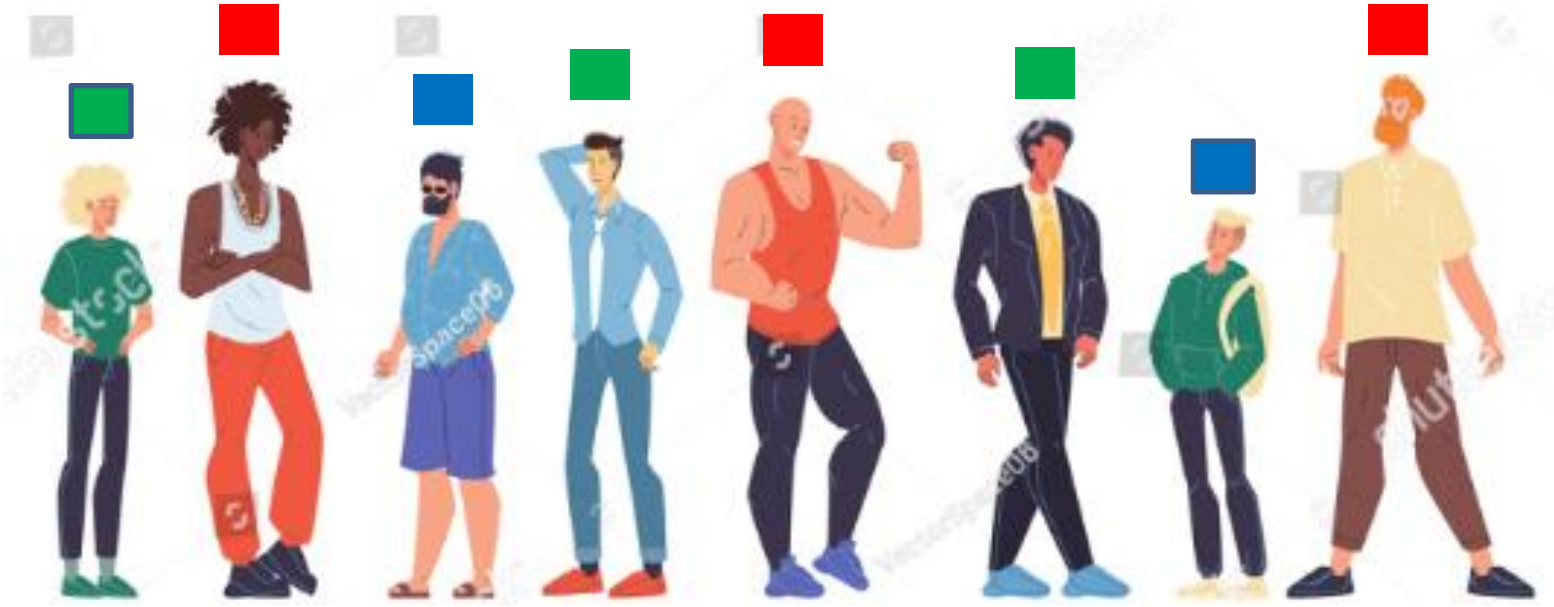
City C (90%) City A (10%)	City B (90%) City C (10%)	City A (95%) City B (5%)	City C (90%) City A (10%)	City B (95%) City C (5%)	City C (90%) City A (10%)	City A (90%) City C (10%)	City B (90%) City A (10%)

It uses something called the **Expectation-Maximization (EM)** algorithm to find the best way to group people based on their heights and weights.



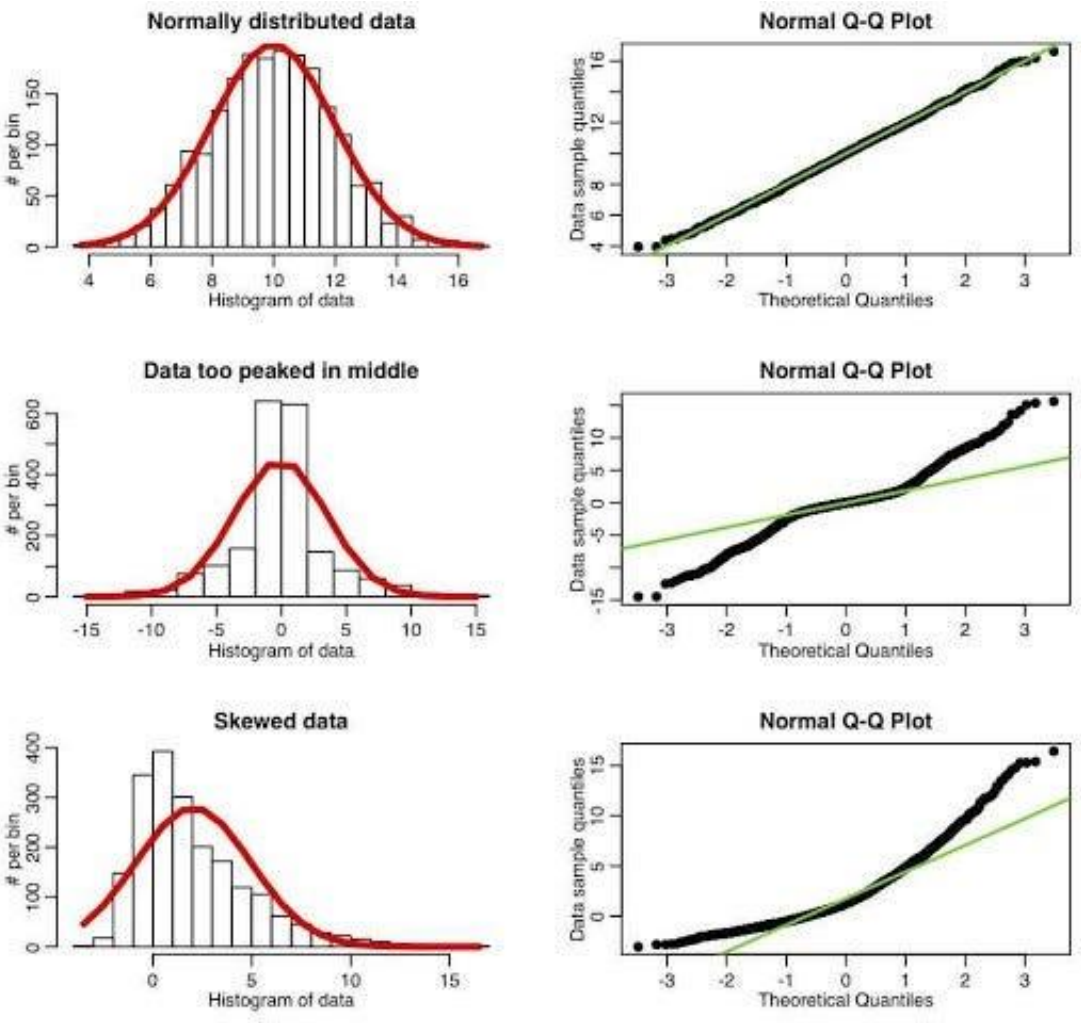
City C (90%) City A (10%)	City B (90%) City C (10%)	City A (95%) City B (5%)	City C (90%) City A (10%)	City B (95%) City C (5%)	City C (90%) City A (10%)	City A (90%) City C (10%)	City B (90%) City A (10%)

5. Result: In the end, GMM gives you a good idea of which group (city) each friend is most likely to belong to, even though it didn't know that at the start!



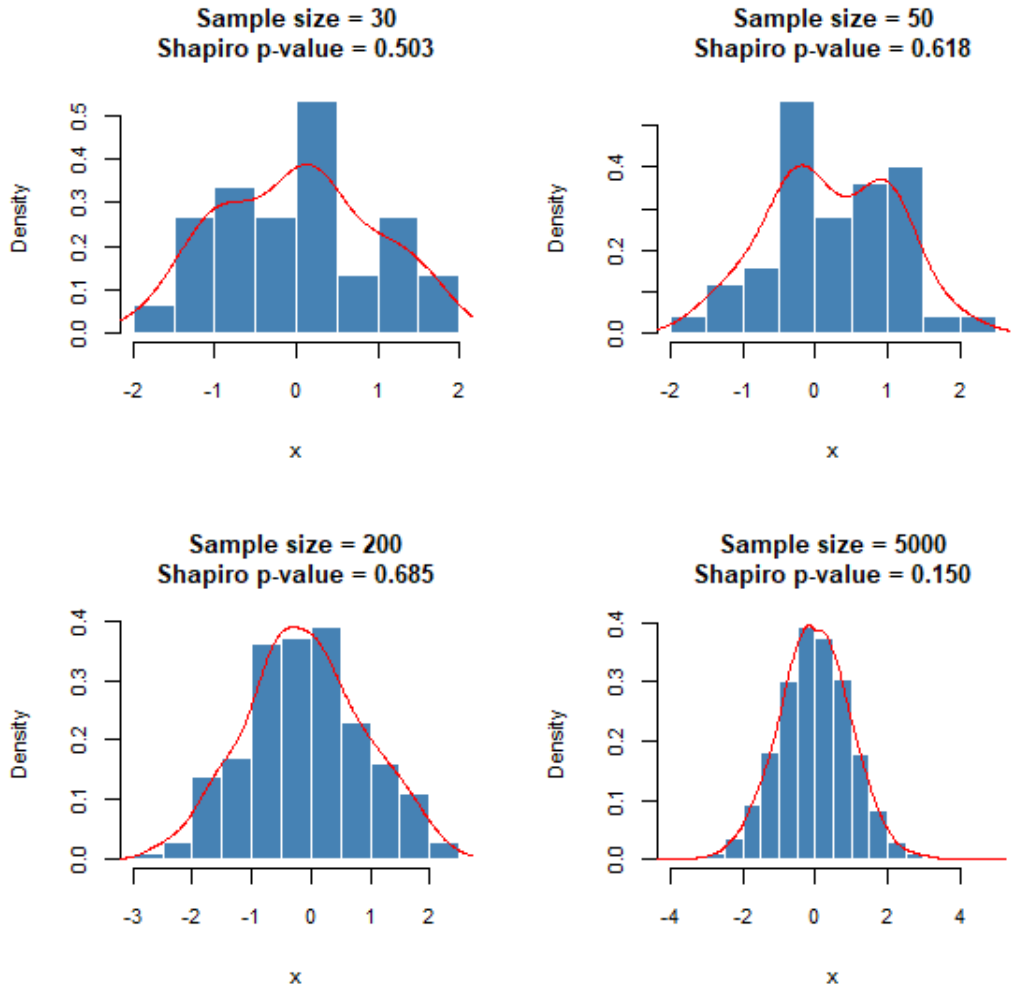
Tests for Normality

Q-Q Plot


Normally distributed data
per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles
Data too peaked in middle
per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles
Skewed data
per bin
Histogram of data
Normal Q-Q Plot
Data sample quantiles
Theoretical Quantiles

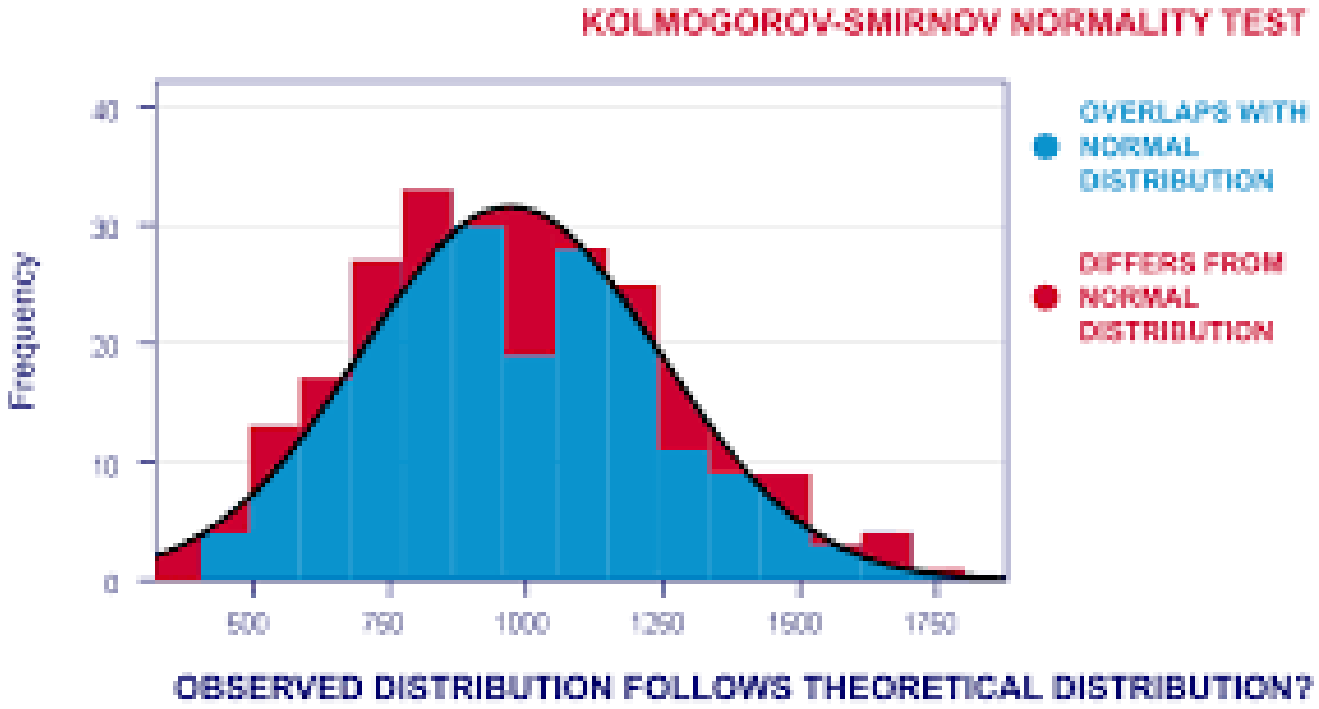
Shapiro – Wilk Test

if p-value <= 0.05 (Non Gausssian)
if p-values > 0.05 (Gaussian)



Kolmogorov-Smirnov Test

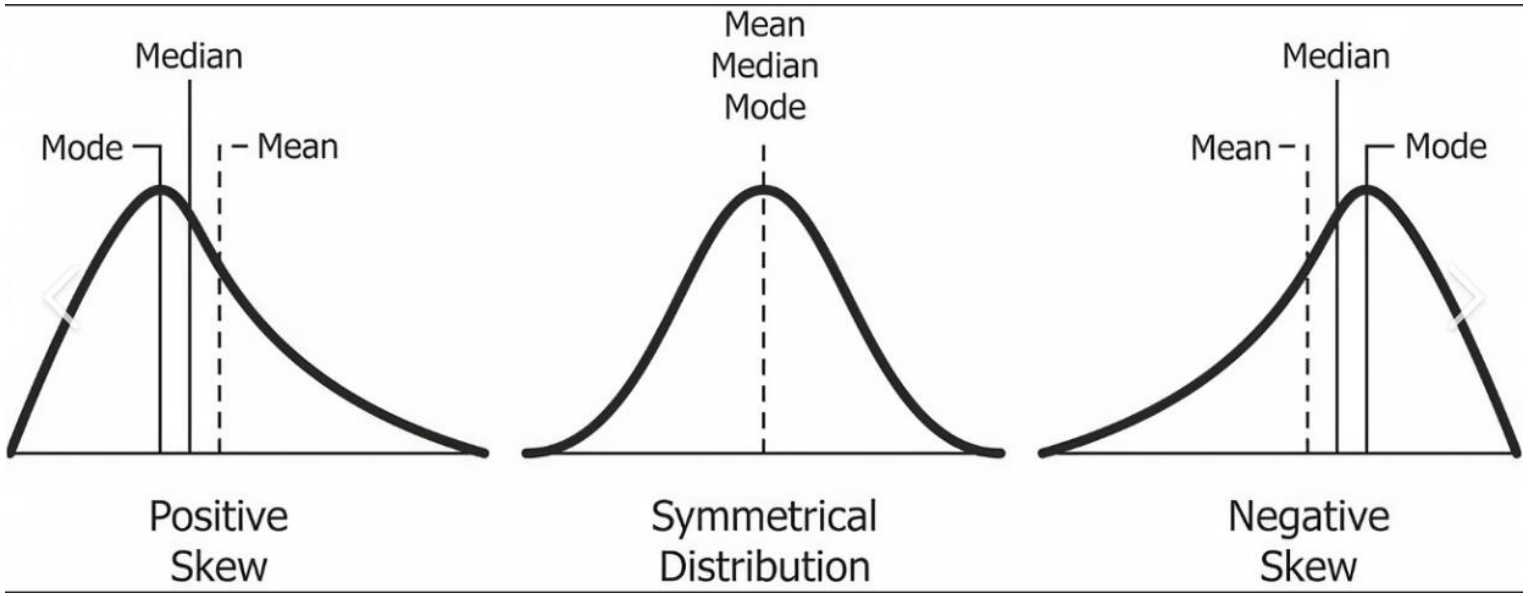
if p-value <= 0.05 (Non Gausssian)
if p-values > 0.05 (Gaussian)



Skewness Test

A skewness value between -1 and +1 is excellent, while -2 to +2 is generally acceptable.

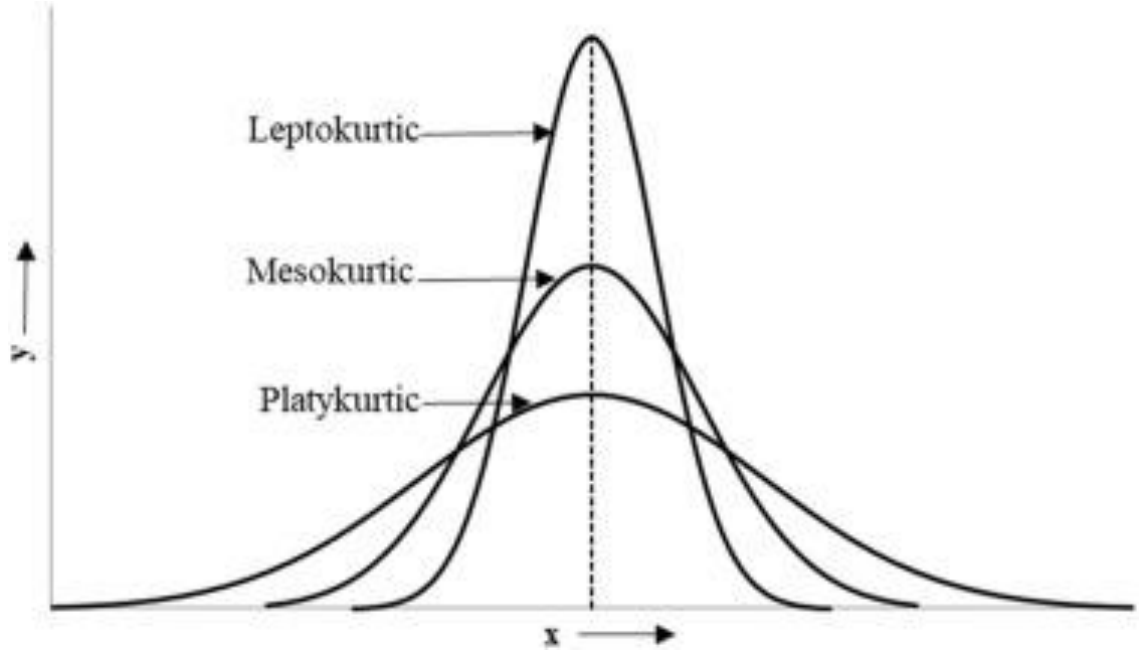
Values beyond -2 and +2 suggest substantial nonnormality (Hair et al., 2022, p. 66).



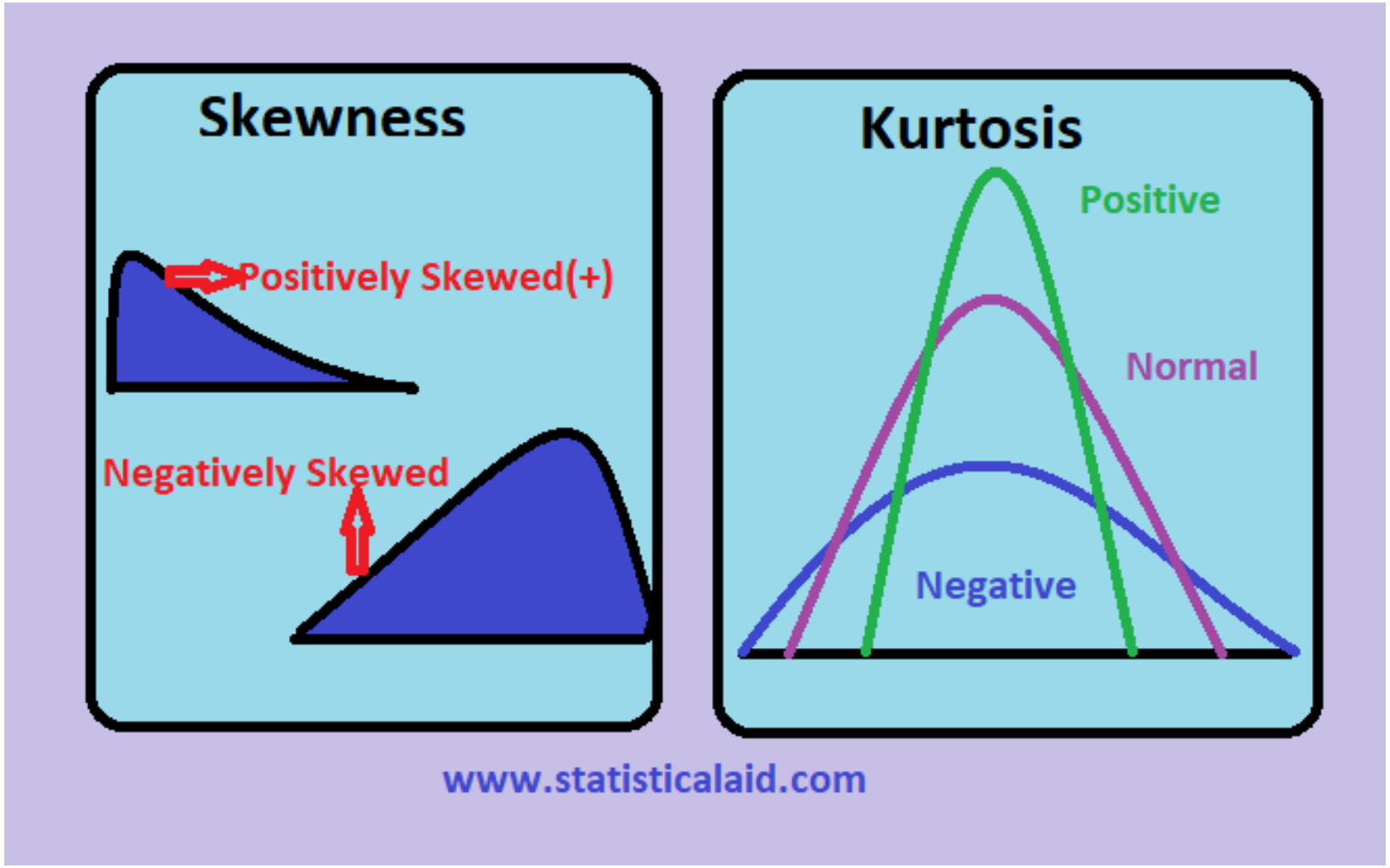
Kurtosis Test

The values for asymmetry and kurtosis between -2 and +2 are considered

Acceptable in order to prove normal univariate distribution (George & Mallery, 2010).



Cognate/Professional Electives



Thank you very much for listening.