# STATISTICS FOR DATA ANALYTICS
# PIZZA SALES DATASET

## POPULATION & SAMPLING

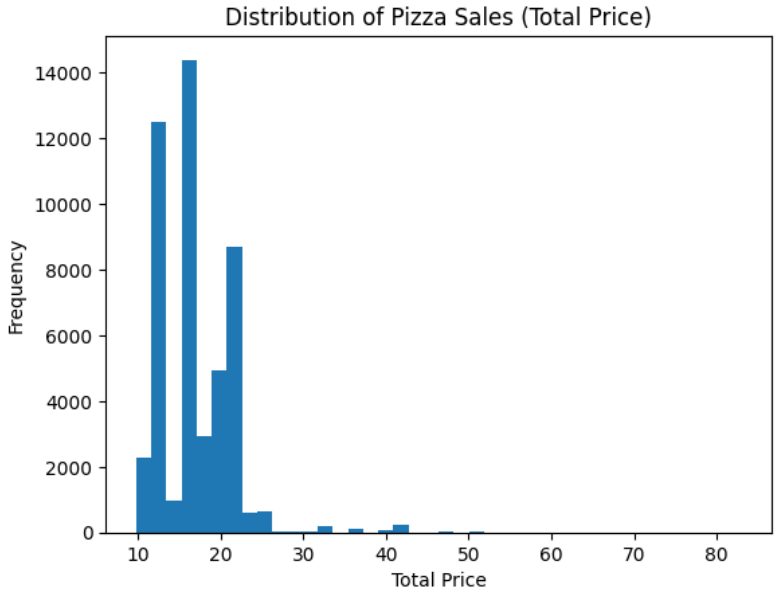
The complete pizza sales dataset was treated as the population.
**Population Size: 48,620 records**
**Sample Size: 4,862 records** (10% Simple Random Sample)
This sampling approach ensures that each record has an equal chance of being selected, making the sample representative of the population.

## SAMPLING TECHNIQUES

**Population Mean (Total Price): 16.82**
**Random Sample Mean: 16.38**
**Systematic Sample Mean: 16.77**
The sample means are close to the population mean. Minor differences occur due to sampling variation. Systematic sampling provided a slightly closer estimate in this case.

## CENTRAL LIMIT THEOREM (CLT)

Multiple random samples of size 30 were taken from the dataset, and the mean of each sample was calculated. The distribution of these sample means formed a bell-shaped curve, demonstrating the Central Limit Theorem. This confirms that sample means follow a normal distribution even when the original data is skewed.



## NORMAL DISTRIBUTION ANALYSIS

**Mean of Total Price: 16.82**
**Standard Deviation: 4.44**

The histogram of total price shows a right-skewed distribution. Applying the 68–95–99.7 rule:

- 78.45% of values lie within 1 standard deviation
- 98.56% of values lie within 2 standard deviations
- 98.62% of values lie within 3 standard deviations

The data approximately follows a normal distribution, a common characteristic in real-world sales data.

## Z-SCORE & OUTLIERS

Z-scores were calculated for the total price column to identify unusual values.
**Number of Outliers Identified: 673**
These outliers represent extremely high-value orders, such as bulk purchases or special events. They are important for business insights rather than data errors.

## BUSINESS INSIGHTS

1. **Why is sampling required in real-world data analysis?**

   In real life, datasets are very large. Analysing all the data takes a lot of time and cost. Sampling helps us work with a small part of the data that represents the whole dataset. This makes analysis faster and easier while still giving good results.

2. **How does the Central Limit Theorem (CLT) help in analytics?**

   CLT helps us understand that when we take many samples of size 30 or more, the average of those samples follows a normal distribution. This is useful because it allows us to apply statistical methods even if the original data is not normally distributed.

3. **Why is normal distribution important before hypothesis testing?**

   Many statistical tests work properly only when data is normally distributed. Checking the normal distribution helps us trust the results of these tests. It makes sure that our analysis and conclusions are accurate.

4. **How does Z-Score help in identifying unusual values?**

   Z-score shows how far a value is from the average. If a value is very far from the mean, it is considered unusual. Z-score helps us easily find such extreme values, which may represent special cases or errors in data.