About Dataset

- The dataset I used for this analysis is titled FAO Food Loss 2000-2021. It contains information about food loss across various countries, years, and commodities. Key columns include:

Attributes	Description
country	Name of the country where the data was collected.
year	Year of data collection.
commodity	Type of food or agricultural product.
loss_percentage	Percentage of food loss for the given commodity and year.
loss_quantity	Quantity of food lost (in relevant units).

- I selected the columns loss_percentage, loss_quantity, commodity, and year because they provide numeric and categorical information essential for statistical techniques like regression, clustering, and hypothesis testing. Understanding these features allows me to perform meaningful analysis and derive actionable insights from the data.

Descriptive Statistics

- Descriptive statistics involve summarizing and describing the essential characteristics of a dataset. These techniques help us understand the general trends and patterns in the data. The key measures include:

- Mean: This represents the average value, calculated by summing all data points and dividing by the number of points.

- Median: This is the middle value when data points are arranged in ascending or descending order, providing a measure of central tendency that is less sensitive to outliers.

- Mode: The most frequently occurring value in the dataset, highlighting common data points.

- Variance: Measures the spread of data points around the mean, indicating how dispersed the data is.

- Standard Deviation: The square root of variance, it quantifies the typical deviation of data points from the mean.

- Descriptive statistics give me a snapshot of the data, highlighting the average trends, variability, and anomalies. This foundation is crucial for deeper analysis and helps guide decisions on what methods to apply next

Probability Distributions

- A probability distribution is a mathematical function that describes the likelihood of different outcomes in a dataset. The normal distribution, often called the bell curve, is one of the most widely recognized. It is symmetric, with most data points clustering around the mean and fewer occurring as they move away from it.

- Understanding probability distributions helps me model real-world processes and predict future outcomes. For example, by fitting a normal distribution to loss_percentage, I can identify the most likely range of values and detect outliers or unusual patterns.

Hypothesis Testing

- Hypothesis testing is a statistical method used to determine if there is enough evidence to reject a null hypothesis. The null hypothesis typically states that there is no effect or difference between groups. A common technique is the t-test, which compares the means of two groups.

- I use hypothesis testing to validate assumptions and ensure that observed differences in the dataset are statistically significant rather than due to random chance. For instance, comparing the loss_percentage of two commodities can help identify significant differences in their loss behaviors.

Chi-Square Test

- The chi-square test evaluates whether there is a significant association between two categorical variables. It compares observed frequencies in a contingency table to expected frequencies under the assumption of independence.

- This test helps me understand relationships between variables, like whether certain commodities are more likely to fall into specific loss categories. Identifying such patterns can aid in targeted interventions.

Regression Analysis

- Regression analysis is a predictive modeling technique used to estimate the relationship between an independent variable (predictor) and a dependent variable (outcome). Linear regression, for example, assumes a straight-line relationship between variables.

- I use regression analysis to quantify how changes in loss_percentage affect loss_quantity. This helps predict outcomes and evaluate the strength of the relationship between variables, providing actionable insights for stakeholders.

Clustering

- Clustering is an unsupervised learning method that groups data points based on their similarities. K-means clustering, for example, divides data into predefined clusters, ensuring that points within the same cluster are more similar to each other than to those in other clusters.

- Clustering helps me identify patterns and group commodities with similar loss characteristics. This segmentation is valuable for tailoring strategies to reduce food loss in specific clusters.