# Applied Statistical Analysis Project

## \*\*Project I: Descriptive Statistics\*\*

\* \*\*Topic\*\*: Statistical Analysis of Diet Preference Data

\* \*\*Project Objective\*\*: To understand the distributions of data within a diet preference dataset.

\* \*\*Project Description\*\*: This project requires performing the first phase of data analysis on a dataset related to diet choices. It involves cleaning the data, calculating summary statistics, and performing initial visualizations to understand its basic features.

\* \*\*Project Outcome\*\*: The project will deliver a complete data set parameter analysis, including summary statistics (mean, median, etc.), visualizations (histograms, box plots), and pattern analysis through correlation matrices.

### \*\*1. Data Preprocessing\*\*

Data preprocessing is the initial phase involving the cleaning and preparation of the raw data. The `synthetic\_diet\_responses.csv` dataset was loaded into RStudio. A procedure was established to handle potential missing values, and numerical columns were standardized using the Z-score method to prepare the data for analysis.

### \*\*2. Exploratory Data Analysis (EDA)\*\*

This section focuses on summarizing the main characteristics of the dataset through descriptive statistics and initial visualizations.

#### \*\*2.1 Basic Statistics\*\*

The fundamental statistical properties of the key numerical variables were calculated.

| Statistic | What is your age? | How confident are you...? (1-5) | How important are ethical concerns...? (1-5) |

| :--- | :--- | :--- | :--- |

| \*\*Mean\*\* | 40.88 | 3.656 | 3.717 |

| \*\*Median\*\* | 41.00 | 4.000 | 4.000 |

| \*\*Minimum\*\* | 18.00 | 1.000 | 1.000 |

| \*\*Maximum\*\* | 65.00 | 5.000 | 5.000 |

\*\*Observation\*\*: The average age of participants is \*\*40.88 years\*\*. The mean scores for both nutritional confidence (\*\*3.66\*\*) and the importance of ethical concerns (\*\*3.72\*\*) are above the neutral midpoint of 3.

#### \*\*2.2 Outlier Identification\*\*

Outliers were identified using standardized Z-scores. No data points were found to fall outside the ±3 range, indicating the absence of significant outliers.

#### \*\*2.3 Frequency Distribution and Histograms\*\*

To understand the distribution of each variable, frequency tables and visualizations were prepared.

\*\*Variable: Which of the following best describes your current diet?\*\*

This variable shows the dietary patterns of the participants.

| Diet Type | Frequency (Count) |

| :--- | :--- |

| Non-Vegetarian | 47 |

| Vegetarian | 40 |

| Eggetarian | 16 |

| Vegan | 11 |

\*\*

\*\*Observation\*\*: The most prevalent diet among participants is Non-Vegetarian (47 responses), making it the largest group.

\*(Add other frequency tables and observations here as needed)\*

### \*\*3. Visualization\*\*

Visualizations were created to explore the relationships between variables.

\*[Insert Scatter Plot, Box Plot, and Correlation Matrix Heatmap here with their observations]\*

---

## \*\*Project II: Statistical Testing\*\*

\* \*\*Topic\*\*: Statistical Hypothesis Testing on Diet Preference Data

\* \*\*Project Objective\*\*: To understand the dependencies and test for significant differences among various data attributes.

\* \*\*Project Description\*\*: This project examines the variances and relationships between different variables using formal statistical hypothesis tests.

\* \*\*Project Outcome\*\*: The project will deliver the results and interpretation of t-tests, ANOVA, and a predictive analysis using simple regression.

### \*\*1. Statistical Testing\*\*

A series of statistical hypothesis tests were conducted. The significance level for all tests was set at α = 0.05.

\*(Insert the consolidated summary of the T-test, ANOVA, Chi-Square Test, Correlation Analysis, and Regression Analysis here.)\*

### \*\*2. Results Visualization\*\*

This section presents the findings from the statistical tests in both tabular and graphical formats.

#### \*\*2.1 Summary of Statistical Tests\*\*

| Test Performed | Variables Analyzed | Key Statistic | P-value | Decision (at α = 0.05) |

| :--- | :--- | :--- | :--- | :--- |

| \*\*Independent T-test\*\* | Age by Gender | t = -0.378 | 0.7058 | Fail to Reject Null Hypothesis |

| \*\*ANOVA\*\* | Age by Diet Type | F = 0.381 | 0.822 | Fail to Reject Null Hypothesis |

| \*\*Chi-Square Test\*\* | Gender and Diet Type | X² = 7.9146 | 0.4419 | Fail to Reject Null Hypothesis |

| \*\*Simple Regression\*\* | Confidence predicted by Age | R² = 0.0002 | 0.6454 | Fail to Reject Null Hypothesis |

#### \*\*2.2 Graphical Analysis\*\*

\*[Insert the Scatter Plot with Trend Line and the Residual Plot here with their observations]\*

### \*\*3. Conclusion\*\*

A series of statistical tests were conducted, and the consistent result across all tests was a \*\*lack of statistically significant findings\*\*. For every test, we failed to reject the null hypothesis. The primary finding is the absence of significant relationships among the tested variables (age, gender, diet type, etc.). A key limitation is that the analysis was performed on synthetic data. Future research could include a wider range of variables on a larger, real-world dataset.