**Project 2: Putting Everything We have learned So far into a Project**

**Assignment: Statistical Analysis of Ugandan Restaurant Data (UGX)**

**Objective:** Analyze the gb\_restaurant\_data\_ugx.csv dataset to practice data cleaning, descriptive statistics, cross-tabulations, and hypothesis testing, deriving insights for a Ugandan restaurant chain. All monetary values are in Ugandan Shillings (UGX). Create a presentation with findings and recommendations.

**Dataset:** Restaurant data with issues (missing values, outliers, misaligned categories, bad column names).

* **Columns:**CustomerID, Total\_Bill\_Amt (bill in UGX), Tip\_Amount\_UGX (tip in UGX), Gender\_Cat (Male/Female), Location\_City (Kampala, Entebbe, Gulu, Arua, Mbarara), Meal\_Type (Kids Meal, Vegetarian, Non-Vegetarian, Special), Visit\_Date (date), Party\_Size (people).
* **Currency:** All amounts in UGX (e.g., Total\_Bill ~100,000-500,000 UGX).

**Tasks:**

1. Clean the data (fix missing values, outliers, categories, column names).
2. Compute descriptive statistics (numerical and categorical).
3. Create cross-tabulations and group-by summaries.
4. Test hypotheses (t-tests, ANOVA, Chi-Squared).
5. Prepare a presentation with findings and recommendations.

**Tools:** Python (Jupyter), pandas, scipy, statsmodels, numpy, seaborn, matplotlib**.**

**Introduction:**

1. **Data Cleaning**:
   * Handle missing values
   * Detect and address outliers
   * Standardize categories (e.g., unify “M”, “Male” to “Male”; fix “Kampalla” to “Kampala”).
   * Rename columns where necessary
   * Convert Visit\_Date to consistent datetime format.
2. **Descriptive Statistics**:
   * Calculate measures for Total\_Bill, Tip\_Amount, Party\_Size (mean, median, mode, std, min, max, quartiles).
   * Summarize categorical variables (Gender, Location, Meal\_Type) with frequency counts and percentages.
   * Visualize distributions (histograms for continuous, bar charts for categorical).
3. **Cross-Tabulations and Group-By**:
   * Create cross-tabs (e.g., Gender vs. Location, Meal\_Type vs. Location).
   * Group by Location or Gender to compute mean Total\_Bill, Tip\_Amount.
   * Visualize with stacked bar charts or heatmaps.
4. **Hypothesis Testing**:
   * **One-Sample t-Test**: Test if average Total\_Bill equals 200,000 UGX
   * **Paired t-Test**: Compare Total\_Bill vs. Tip\_Amount for paired observations.
   * **Independent t-Test**: Compare Total\_Bill between Males and Females.
   * **ANOVA**: Test if Total\_Bill differs across Location (5 cities).
   * **Chi-Squared Test**: Test if Gender and Location are independent.
   * Check assumptions (normality via Shapiro-Wilk, equal variances via Levene’s).
   * ***NOTE: Remember if any of those assumptions are violated for any case, we then opt for the non-parametric counterpart.***
5. **Presentation**:
   * Summarize data cleaning, descriptive stats, cross-tabs, and hypothesis test results.
   * Present findings with visualizations (e.g., histograms, bar charts, boxplots).
   * Provide actionable recommendations (e.g., target marketing, menu adjustments).
   * You will make a presentation for this work on Friday. You will be given 8 minutes each to make your presentation
   * Please rehearse prior to ensure that this fits within your 8 minutes.

**Presentation Outline (Guidance for Learners)**

**Slide Structure** (10-12 slides):

1. **Title Slide**: Assignment title, learner’s name, date (June 18, 2025).
2. **Introduction**: Dataset overview, assignment goals (analyze restaurant data, derive insights).
3. **Data Cleaning**:
   * Issues found: Missing values (~5-10%), outliers (e.g., $1000 bills), misaligned categories (e.g., “M” vs. “Male”), bad column names.
   * Actions: Imputed medians, capped outliers, standardized categories, renamed columns.
4. **Descriptive Statistics**:
   * Numerical: Mean/median Total\_Bill, Tip\_Amount, Party\_Size.
   * Categorical: Gender (~50% Male), Location (Kampala most frequent), Meal\_Type (Non-Vegetarian dominant).
   * Visuals: Include numerical\_distributions.png, categorical\_distributions.png.
5. **Cross-Tabs/Group-By**:
   * Gender vs. Location: Similar distribution across cities.
   * Mean Total\_Bill by Location:
   * Visuals: Include gender\_location\_crosstab.png, mean\_bill\_by\_location.png.
6. **Hypothesis Tests**:
   * **One-Sample t-Test**: H₀: Mean Total\_Bill = 200,000 UGX (e.g., p < 0.05, reject H₀).
   * **Paired t-Test**: H₀: Total\_Bill = Tip\_Amount (e.g., p < 0.05, bills higher).
   * **Independent t-Test**: H₀: Total\_Bill same for Males/Females (e.g., p < 0.05, Males spend more).
   * **ANOVA**: H₀: Total\_Bill same across Locations (e.g., p < 0.05, differences exist; Tukey’s shows Kampala vs. Gulu).
   * **Chi-Squared**: H₀: Gender and Location independent (e.g., p > 0.05, independent).
   * Assumptions: Normality (Shapiro-Wilk, p > 0.05 for Total\_Bill), equal variances (Levene’s, p > 0.05 for Gender).
7. **Findings**:
   * Bills below $50 benchmark, suggesting price adjustments.
   * Males spend more, indicating targeting opportunities.
   * Kampala has higher bills, Gulu lower.
   * Gender and Location unrelated, but meal preferences vary by city.
8. **Recommendations**:
   * Recommendation 1
   * Recommendation 2
   * Recommendation 3
9. **Conclusion**: Summary of skills learned (cleaning, stats, testing), business impact.
10. **References**: Python libraries, statistical concepts.