**Business Case: Walmart - Confidence Interval and CLT**

1. Import the dataset and do usual data analysis steps like checking the structure & characteristics of the dataset.
2. Detect Null values & Outliers (using boxplot, “describe” method by checking the difference between mean and median, isnull etc.)
3. Do some data exploration steps like:
   * Tracking the amount spent per transaction of all the 50 million female customers, and all the 50 million male customers, calculate the average, and conclude the results.
   * Inference after computing the average female and male expenses.
   * Use the sample average to find out an interval within which the population average will lie. Using the sample of female customers you will calculate the interval within which the average spending of 50 million male and female customers may lie.
4. Use the Central limit theorem to compute the interval. Change the sample size to observe the distribution of the mean of the expenses by female and male customers.
   * The interval that you calculated is called Confidence Interval. The width of the interval is mostly decided by the business: Typically 90%, 95%, or 99%. Play around with the width parameter and report the observations.
5. Conclude the results and check if the confidence intervals of average male and female spends are overlapping or not overlapping. How can Walmart leverage this conclusion to make changes or improvements?
6. Perform the same activity for Married vs Unmarried and Age
   * For Age, you can try bins based on life stages: 0-17, 18-25, 26-35, 36-50, 51+ years.
7. Give recommendations and action items to Walmart.

**Evaluation Criteria**

1. Defining Problem Statement and Analyzing basic metrics **(10 Points)**
   1. Observations on shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), statistical summary
   2. Non-Graphical Analysis: Value counts and unique attributes ​
   3. Visual Analysis - Univariate & Bivariate
      * For continuous variable(s): Distplot, countplot, histogram for univariate analysis
      * For categorical variable(s): Boxplot
      * For correlation: Heatmaps, Pairplots
2. Missing Value & Outlier Detection **(10 Points)**
3. Business Insights based on Non- Graphical and Visual Analysis **(10 Points)**
   1. Comments on the range of attributes
   2. Comments on the distribution of the variables and relationship between them
   3. Comments for each univariate and bivariate plot
4. Answering questions (50 Points)
   1. Are women spending more money per transaction than men? Why or Why not? **(10 Points)**
   2. Confidence intervals and distribution of the mean of the expenses by female and male customers **(10 Points)**
   3. Are confidence intervals of average male and female spending overlapping? How can Walmart leverage this conclusion to make changes or improvements? **(10 Points)**
   4. Results when the same activity is performed for Married vs Unmarried **(10 Points)**
   5. Results when the same activity is performed for Age **(10 Points)**
5. Final Insights **(10 Points)** - Illustrate the insights based on exploration and CLT
   1. Comments on the distribution of the variables and relationship between them
   2. Comments for each univariate and bivariate plots
   3. Comments on different variables when generalizing it for Population
6. Recommendations **(10 Points)**
   1. Actionable items for business. No technical jargon. No complications. Simple action items that everyone can understand