

Walmart Inc.

Business Case

Scaler

Topic: Central limit theorem & Confidence interval

Duration: 1 week

Mindset

1. Evaluation will be kept lenient, so make sure you attempt this case study.
2. It is understandable that you might struggle with getting started on this. Just brainstorm, discuss with peers, or get help from TAs.
3. There is no right or wrong answer. We have to become comfortable with dealing with uncertainty in business. This is exactly the skill we want to develop.

About Walmart

Walmart is an American multinational retail corporation that operates a chain of supercenters, discount departmental stores, and grocery stores from the United States. Walmart has more than 100 million customers worldwide.

Business Problem

The Management team at Walmart Inc. wants to analyze the customer purchase behavior (specifically, purchase amount) against the customer's gender and the various other factors to help the business make better decisions. They want to understand if the spending habits differ between male and female customers: *Do women spend more on Black Friday than men?* (Assume 50 million customers are male and 50 million are female).

Dataset

The company collected the transactional data of customers who purchased products from the Walmart Stores during Black Friday. The dataset has the following features:

Dataset link: [walmart_data.csv](#)

- **User_ID:** User ID
- **Product_ID:** Product ID
- **Gender:** Sex of User
- **Age:** Age in bins
- **Occupation:** Occupation(Masked)

- **City_Category:** Category of the City (A,B,C)
- **StayInCurrentCityYears:** Number of years stay in current city
- **Marital_Status:** Marital Status
- **ProductCategory:** Product Category (Masked)
- **Purchase:** Purchase Amount

What good looks like?

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)**
 - a. Observations on shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), statistical summary

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

Submission Process:

- Type your insights and recommendations in the text editor.
- Convert your jupyter notebook into PDF (Save as PDF using Chrome browser's Print command), upload it in your Google Drive (set the permission to allow public access), and paste that link in the text editor.

- Optionally, you may add images/graphs in the text editor by taking screenshots or saving matplotlib graphs using `plt.savefig(...)`.
- After submitting, you will not be allowed to edit your submission.

50M population

1000 samples

Each sample 300

Each of 1000 samples - mean