Walmart Business case study

- ❖ **Topic:** EDA
- ❖ **Duration:** 1 week

Why this case study?

From the company's perspective:

- Walmart is an American multinational retail corporation that operates a chain of supercenters, discount departmental stores, and grocery stores in the United States. Walmart has more than 100 million customers worldwide.
- The Management team at Walmart Inc. wants to analyze the customer purchase behavior (precisely, 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?*

From the learner's perspective:

- Solving this business case holds immense importance for aspiring data analysts and scientists.
- Through the process of engaging with this case study, individuals acquire practical knowledge and develop skills in Data Analysis, which are essential for deriving meaningful insights from data.
- Engaging in the resolution of this specific case study will aid in cultivating expertise in identifying connections between input variables and output variables.

Dataset link: **walmart_data.csv**

- **User_ID:** User ID
- **Product_ID:** Product ID
- **Gender:** Sex of User
- **Age:** Age in bins

- **Occupation:** Occupation
- **City_Category:** Category of the City (A,B,C)
- **Stay*In*Current*City*Years:** Number of years stay in current city
- **Marital_Status:** Marital Status
- **Product*Category*:** Product Category
- **Purchase:** Purchase Amount

What is expected?

As a data analyst/scientist at Walmart, you have been assigned the responsibility of examining the provided dataset to derive meaningful insights and provide actionable recommendations.

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.

General Guidelines:

- Evaluation will be kept lenient, so make sure you attempt this case study.
- It is understandable that you might struggle with getting started on this or feel stuck at some point.
  In such cases:
  a. Read the question carefully and try to understand what exactly is being asked.
  b. Brainstorm a little. If you're getting an error, remember that Google is your best friend.
  c. You can watch the lecture recordings or go through your lecture notes once again if you feel like you're getting confused over some specific topics.
  d. Discuss your problems with your peers. Make use of the Slack channel and WhatsApp group.
  e. Only if you think that there's a major issue, you can reach out to your Instructor via Slack or Email.

What does 'good' look like?

1. **Import the dataset and do usual data analysis steps like checking the structure & characteristics of the dataset**

    a. The data type of all columns in the “customers” table.

    Hint: We want you to display the data type of each column present in the dataset.

    b. You can find the number of rows and columns given in the dataset

    Hint: You can find the shape of the dataset.

    c. Check for the missing values and find the number of missing values in each column

---

2. **Detect Null values and outliers**

    a. Find the outliers for every continuous variable in the dataset

    Hint: Use boxplots to find the outliers in the given dataset

    b. Remove/clip the data between the 5 percentile and 95 percentile

    Hint: You can use np.clip() for clipping the data

---

3. **Data Exploration**

    a. What products are different age groups buying?

    Hint: You can use histplot to find the relationship between products and age groups

    b. Is there a relationship between age, marital status, and the amount spent?

    Hint: You can do multivariate analysis to find the relationship between age, marital status, and the amount spent

    c. Are there preferred product categories for different genders?

    Hint: You can apply different hist plots for different genders

---

4. **How does gender affect the amount spent?**

   Hint: Use the central limit theorem and bootstrapping to compute the 95% confidence intervals for the average amount spent per gender. First, compute the confidence interval for whatever data is available, and then repeat the same with smaller sample sizes - 300, 3000, and 30000.

   a. From the above calculated CLT answer the following questions.
      i. Is the confidence interval computed using the entire dataset wider for one of the genders? Why is this the case?
      ii. How is the width of the confidence interval affected by the sample size?
      iii. Do the confidence intervals for different sample sizes overlap?
      iv. How does the sample size affect the shape of the distributions of the means?

---

5. **How does Marital_Status affect the amount spent?**

   Hint: Use the central limit theorem and bootstrapping to compute the 95% confidence intervals for the average amount spent per Marital_Status. First, compute the confidence interval for whatever data is available, and then repeat the same with smaller sample sizes - 300, 3000, and 30000.

   a. From the above calculated CLT answer the following questions.
      i. Is the confidence interval computed using the entire dataset wider for one of the genders? Why is this the case?
      ii. How is the width of the confidence interval affected by the sample size?
      iii. Do the confidence intervals for different sample sizes overlap?
      iv. How does the sample size affect the shape of the distributions of the means?

---

6. **How does Age affect the amount spent?**

Hint: Use the central limit theorem and bootstrapping to compute the 95% confidence intervals for the average amount spent per Marital_Status. First, compute the confidence interval for whatever data is available, and then repeat the same with smaller sample sizes - 300, 3000, and 30000.

a. From the above calculated CLT answer the following questions.
    i. Is the confidence interval computed using the entire dataset wider for one of the genders? Why is this the case?
    ii. How is the width of the confidence interval affected by the sample size?
    iii. Do the confidence intervals for different sample sizes overlap?
    iv. How does the sample size affect the shape of the distributions of the means?

---

7. Create a report

a. Report whether the confidence intervals for the average amount spent by males and females (computed using all the data) overlap. How can Walmart leverage this conclusion to make changes or improvements?

   Hint: Check whether the average spending of males and females overlap or not using the CLT that you calculated

b. Report whether the confidence intervals for the average amount spent by married and unmarried (computed using all the data) overlap. How can Walmart leverage this conclusion to make changes or improvements?

   Hint: Check whether the average spending of married and unmarried overlap or not using the CLT that you calculated.

c. Report whether the confidence intervals for the average amount spent by different age groups (computed using all the data) overlap. How can Walmart leverage this conclusion to make changes or improvements?

   Hint: Check whether the average spending of different age groups overlaps or not using the CLT that you calculated.

---

8. **Recommendations**

    a. Write a detailed recommendation from the analysis that you have done.