Q1. Business Case: Walmart - Confidence Interval and CLT

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).

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.

WALMART DATA ANALYSIS

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.

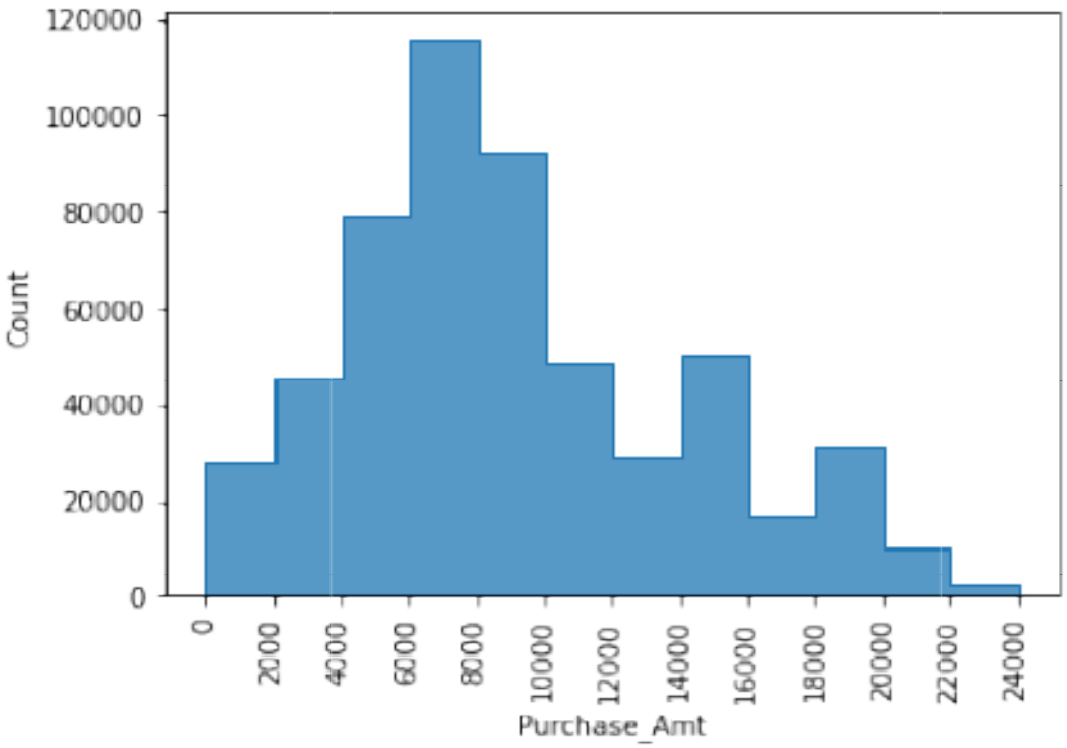
As per the business problem, the Management team at Walmart Inc. wants to know, how attributes of a person (eg. Gender, Age etc.) affect the purchase at Walmart.

It was a simple data frame which consisted data of individuals who purchased different goods from the Walmart stores during Black Friday. I found the data very clean and properly structured. There were no missing values found.

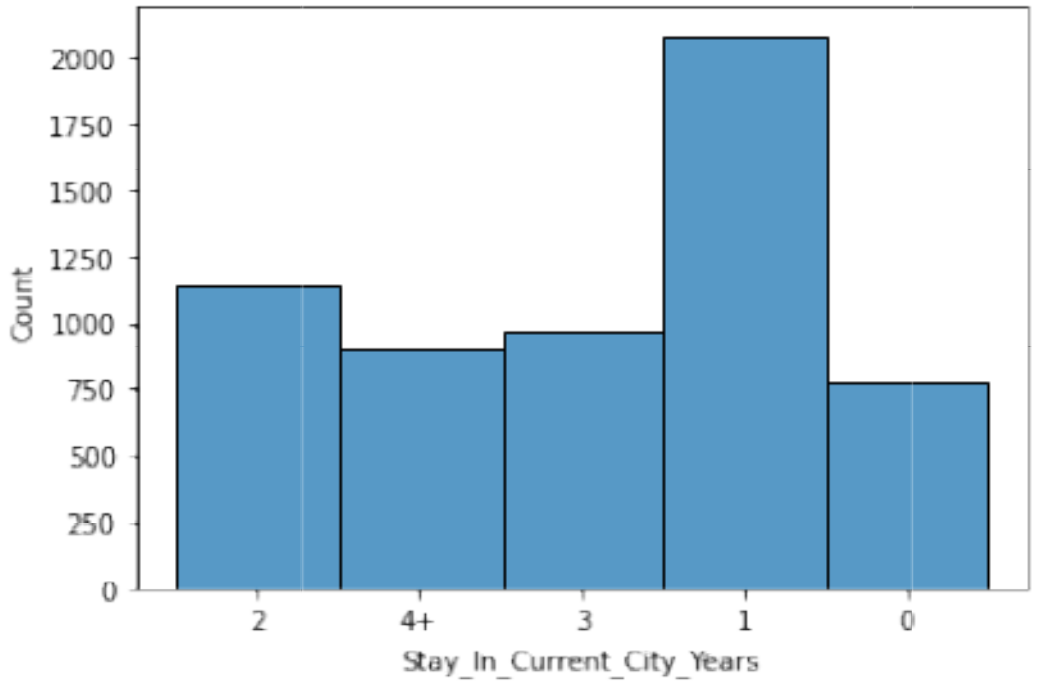
I removed the outliers from the purchase amount data by using IQR method.

With the given data, I was able to draft certain observations which can help improve Walmart's business. Please find them given below:

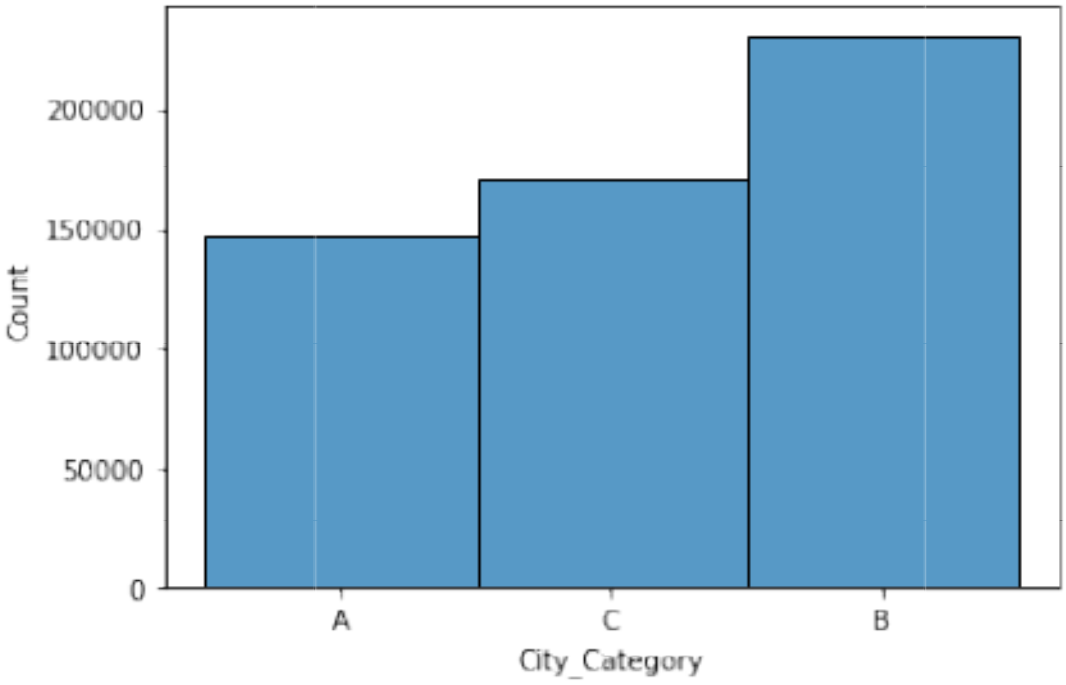
- Starting with the product purchase count based on price, we can see that higher the price, lesser is the purchase count. Products with prices ranging from $6000 to $8000 are bought by most of the people. We can also see certain unevenness in the graph (Ups and downs) from S12000 to S24000. Products with prices ranging from $14000 to $16000 and $18000 and $20000 are bought the most as compared to its nearby counterparts. There has to be some important reason for more people to buy the products in those range. Hence, we can we give more discount for those products and increase its sales count. By doing this we can increase the net business revenue.



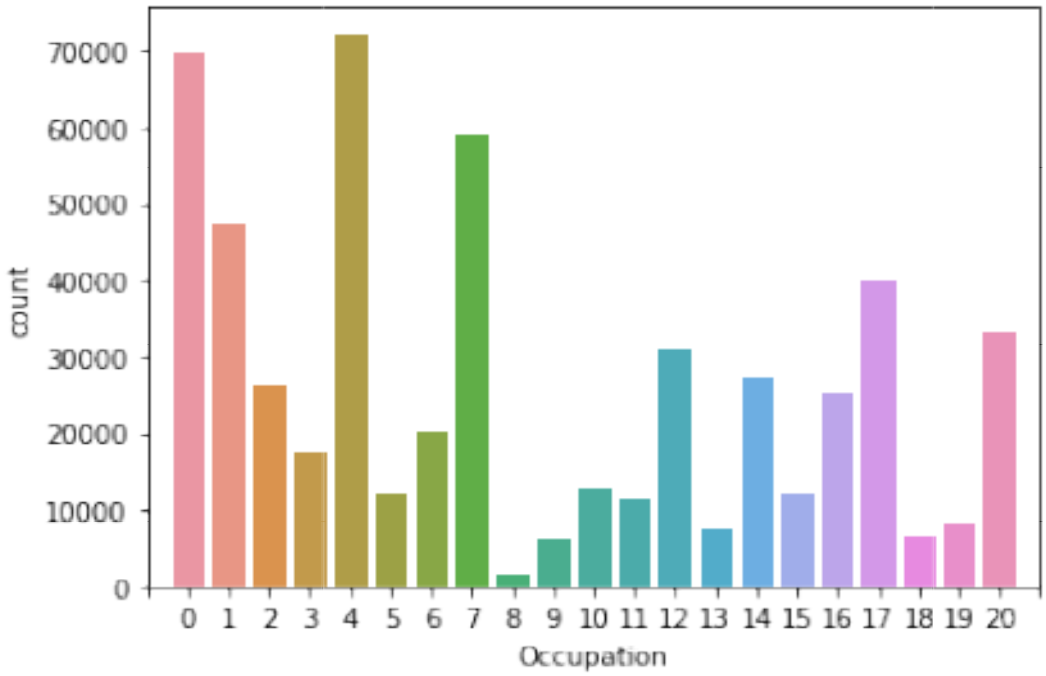
- I would consider this as one of the most important insights of my analysis. If we take a look at the graph, we can see that lesser people purchasing products at Walmart who are staying for more than 2 years. This means, the people initially purchased at the store due to its brand name, but later stopped due to some unknown reason. I would advise the management team to provide more discounts or coupons for people who have been purchasing consistently for more than 2years. I could also see that the amount spent on an average does not deviate much with change in stay years.



- Purchases from 'B' category city are higher as compared to 'A' and 'C'. We need to attract more people to buy in A and C cities by advertising, giving discounts etc. Again, there is not much change in average amount spent among all three cities.



- Mean spend of each and every occupation is almost same. But, if we take a look at purchase count, we can see that there are few occupations (Eg. 8, 18 etc.) where the count is very less. To improve the business, I would advise the management team to procure more products related to that occupation and keep it for sales in retail stores.



- As per the requirement of this business case, I used central limit theorem to calculate C.Is for mean on purchase amounts w.r.t gender, age and Marital status. I was not able to find any difference in 90%, 99% and 95% C.Is for all the attributes. Hence, I went with 95% C.Is for my analysis. From my calculations, we can see that 95% of men spent around 0.688m to 0.725m. However, 95 % of Women spent around 0.548m to 0.599m. There is no overlap and men are clearly dominating based on mean purchases. To increase female purchases, we have to start selling more products which are useful to women.

- There is very less difference in 95% C.Is of married and single. 95 % C.I for mean on purchase by a married is [0.629m, 0.675m] and 95 % C.I for mean on purchase by a single is [0.659m, 0.7m]. There is overlap and we can still go ahead and say that single spending is higher compared to married.

- If we consider age, 95 % C.I for mean on purchase by a person in age group 0-17yrs is [0.48m, 0.6m], 18-25yrs is [0.644m, 0.716m], 26-35yrs is [0.71m, 0.766m], 36-45yrs is [0.638m, 0.706m], 46-50yrs is [0.554m, 0.645m], 51-55yrs is [0.57m, 0.67m] and 55+yrs is [0.44m, 0.531m]. 26-35yrs purchase amounts dominate. We can keep more products which interests teenagers (0-17yrs). I would advise the management team to add more fitness related equipment (Mostly cardio related like treadmills), so that purchase amounts might increase for 36+ years.

I have also created a cross tab in line 20 of my Jupyter notebook for each and every attribute w.r.t purchase amount. Please go through it for in detail analysis.

To whoever reads this, I hope my insights from this case study were meaningful.
Thank you,
Krishna