Analyzing Gender-Based Purchase Behavior on Black Friday: A Study of Walmart Inc. Customers

***Abstract*—This research endeavors to unveil crucial insights into Walmart Inc. by analyzing a substantial dataset which consist of information about customer's age, gender, marital status, product code, occupation and their city category. By performing visualizations using matplotlib, seaborn and other libraries the study highlights distinct patterns in purchasing habits. Notably, the findings underscore gender-based variations, revealing that men tend to spend more than women. Additionally, it sheds light on age and marital status dynamics, emphasizing that the highest spenders predominantly belong to the 50+ age group and are unmarried. These insights contribute to a comprehensive understanding of consumer behavior, providing valuable considerations for strategic decision-making in the retail landscape.**

I. Introduction

In the dynamic world of retail, understanding how customers make purchases is crucial for businesses looking to improve customer experiences and streamline operations. Walmart, being one of the world's largest retailers, serves millions of customers daily, highlighting the importance of examining the patterns that shape their buying habits. This project aims to explore customer purchase behavior at Walmart through univariate, bivariate, and central limit theorem analyses. Using transactional data obtained from Kaggle.com, we aim to uncover patterns, preferences, and trends in customer purchasing habits based on age, gender, and marital status. By grasping the factors influencing customer choices, purchase frequency, and the potential impact of external factors, this research seeks to offer insights for better targeted marketing strategies, and overall business improvement.

II. Previous work

With the relentless increase in retail industries and consumerism, several research institutes and retailers place heavy focus on the study and analysis of gender-based purchasing behavior. Different approaches are available to analyze spending patterns. For instance, an exploratory study aimed at understanding the spending patterns of customers of natural cosmetics is carried out by (Tengli and Srinivasan, 2022), who apply the theory of planned behavior to obtain evidence of similarity between men and women in natural cosmetics purchasing behavior. Meanwhile, hypothesis testing and independent sample t-tests are utilized by (Pakasi and Tumiwa, 2016), revealing significant differences in purchasing behavior between genders. Finally, several statistical methods like the Cronbach test are used to establish a significant relationship between gender and consumer behavior in a case study on mobile phones by (Ronaghi, Danae, et al.,2013). Regardless of the industry, these studies highlight the importance of understanding consumer spending patterns, providing valuable insights for product inventory, marketing, and overall profit enhancement.

III. Methodology

A. Preprocessing:

As with any dataset, understanding the attributes and its characteristics is vital prior to the analysis. With appropriate preprocessing steps, we can familiarize and distinguish the importance of the 10 attributes spanning 550068 records in the data. For this, we rely heavily on the pandas module which not only provides meaningful functions that help understand the dynamics of the dataset but also gives several statistical figures that further help us understand the composition of records to the corresponding attributes. One of the most pivotal aspects of preprocessing, especially when dealing with a large dataset is checking for outliers. These extreme values can greatly affect the analysis of data if not addressed properly before the start of the analysis. There are few known methods that deal with the presence of outliers, one of which is simply ignoring the outliers, which we can only do after confirming that no harm is inflicted by the extreme values on the overall structure of the dataset. For this, we use the Interquartile range method. In statistics, the interquartile range is a measure of spread of the data. This technique gives us values for calculating the range (Max–Min) and the inter quartile IQR(Q3 - Q1) which can provide us with a boundary for distinguishing normal data from outliers (Abir, 2020). Any data points 1.5 times above or below the measured IQR will be considered outliers. The mean, standard deviation, and median are measured before and after the removal of outliers and if there is no significant impact, the outliers can be ignored.

More preprocessing steps, apart from inspecting null values and outliers, include checking the data types of the attributes, generating a summary of descriptive statistics of eligible attributes, type encoding if required, and checking for duplicates in the data.

B. *Univariate Analysis*

To better understand the distribution of features in the data we start by exploring single variables. Here we are looking at the composition of male and female customers, the age distribution of customers, and finding out which products are popular among both the genders and different age groups. The main objective of performing univariate analysis is to try and visualize the data in order to find any patterns.

a) Customer Gender Proportion

To understand the distribution of genders among Walmart Inc. customers on Black Friday, a bar graph was generated using the matplotlib library. The x-axis represents gender categories (Male, Female), and the y-axis depicts the count of customers. In addition to that, a pie chart showcasing the percentage of male and female population out of the total customer is also used.

b) Customer Age Distribution

The age distribution of customers using the pie chart and count plot will help us identify whether the customer base is predominantly composed of a specific age group or if it is more evenly distributed across different age ranges. Additionally the pie chart and count plot would also show that if there is a peak in a particular age group, it suggests that marketing strategies and product offerings tailored to that age group may be particularly effective.

c) Marital Status Distribution

Marital status can influence purchasing behavior, especially for products and deals that may be attractive to families. For instance, married couples with children might be more interested in family-oriented promotions, while single individuals may have different preferences. We analyze the Marital Status distribution using the Pie chart and Count plot to see the percentage and count of married and unmarried customers, respectively.

d) Popular Products Among Male and Female

The bar plot allows for a direct comparison of the most popular product categories between male and female customers. It will help us identify whether certain product categories are more appealing to one gender over the other. Furthermore, the popularity of product categories informs inventory planning. Retailers can ensure that they have sufficient stock of products that are in high demand among both genders.

C. *Bivariate Analysis*

The aim of conducting bivariate analysis is to explore how the distribution of age groups varies between male and female customers and based on the marital status of customers. Analysing age, gender, and marital status together helps find points where they all affect customers' behaviour. This overall understanding will help stores make shopping more personal, manage products better, and improve ads to meet the different needs of customers on Black Friday. Furthermore, a correlation analysis was considered to see what attributes impact purchase behaviour, however, with the dataset predominantly having categorical values, the correlation analysis seemed fruitless as it could not generate any insight related to gender, age or marital status affecting the purchase.

a) Purchase habits related to gender among different age groups

In order to understand purchase value pertaining to different age groups based on gender, we make use of the Box plot and Line graph that effectively display trends of purchase amount within each age group. On the x-axis of the Box Plot and Line graph age group is plotted and on the y-axis of the Box plot Purchase Amount and for the Line graph Average Purchase Amount is plotted.

The Box plot visually represents the distribution and spread of purchase values within each age group for both male and female customers. It can reveal not only the median purchase values but also the variation and potential outliers, helping to identify specific age groups where gender-based spending differences are more pronounced. On the other hand, the Line graph would show trends over different age groups, presenting a clear comparison of purchase habits between males and females. By observing the trends in spending behaviour across various age brackets, retailers can identify patterns such as age groups where one gender tends to spend more than the other. Together, these visualizations provide a comprehensive understanding of how gender influences purchase habits within distinct age segments.

b) Purchase habits related to marital status among different age groups

A similar visualization, in the case of understanding the purchase behaviour based on the marital status of customers among different age groups, would help identify specific age groups where marital status plays a notable role in shaping purchasing habits. Simultaneously, the line graph would illustrate trends in spending behaviour across various age groups, segmented by marital status. It provides a clear comparison of how purchase habits differ between married and single individuals within different age brackets.

D. *Statistical Analysis - Central Limit Theorem*

To obtain robust conclusions regarding the spending behaviour among gender, age groups and marital status, it is beneficial to implement a statistical method that allows us to make more reliable inferences about spending patterns. In this case, we implement the Central Limit Theorem and calculate Confidence Intervals. The Central Limit Theorem states that the distribution of sample means approaches a normal distribution, even if the original population distribution is not normal, given a sufficiently large sample size. By calculating sample means for spending patterns within different gender, age, and marital status categories, we can rely on the CLT to assume normality. This allows for the construction of confidence intervals around the sample means, providing a range within which the true population mean is likely to fall. Thus, we can estimate the spending differences among gender, age group and marital status which will provide valuable insights for data driven decision making at Walmart.

a) Implementing Central Limit Theorem and Calculating Confidence Intervals Using Bootstrap

We have employed the Central Limit Theorem (CLT) to find the spending patterns across various demographic dimensions, including gender, age, and marital status.
The CLT's role in ensuring normality is important as it enables the construction of Confidence Intervals around sample means for each demographic category.

Procedure:

1. Sampling: We drew random samples within each demographic category, meticulously adhering to the CLT sample size requirements to guarantee statistical reliability.
2. Sample Means: For every sample within gender, age, and marital status groups, we computed the spending pattern means, capturing the central tendency of each subgroup.
3. Bootstrap: To enhance our statistical robustness, we employed Bootstrap resampling. This iterative process created a distribution of sample means, capturing the variability inherent in different demographic segments.
4. Confidence Intervals: Using the Bootstrap-generated distribution, we constructed Confidence Intervals. These intervals provide a statistical range within which we can confidently estimate the likely population mean for each demographic category.

IV. Results and Visualizations

a) Univariate Analysis Results

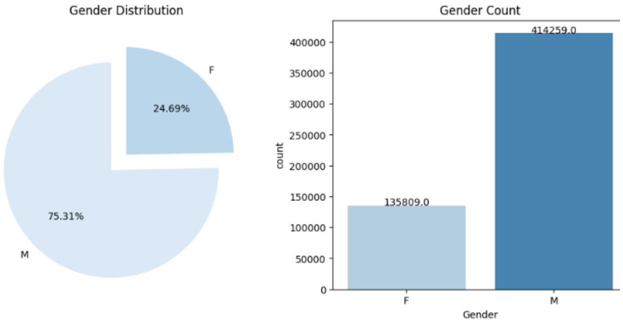


Fig 3.1

On conducting a univariate analysis on the data to examine the gender distribution. The results in *fig 3.1* indicate that the majority of customers, approximately 75%, are men, while the remaining 25% are women. This distribution is visually represented in a bar plot, where the count for women is 135,809 and for men is 414,259.

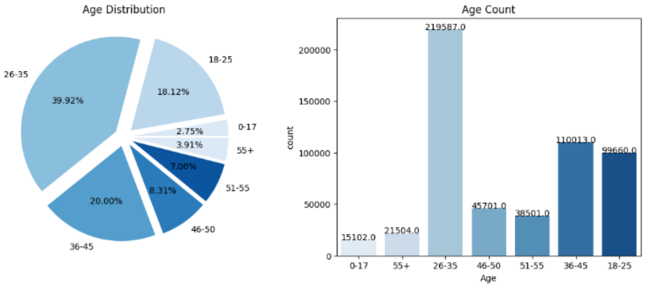


Fig 3.2

The age distribution of all customers was categorized into seven sections: [0-17], [18-25], [26-35], [36-45], [46-50], [51-55], and [55+]. The bar plot *(fig 3.2)* reveals that the majority of customers fall into the age group of 26-35 years old, while the least number of people are in the 0-17 age bracket.

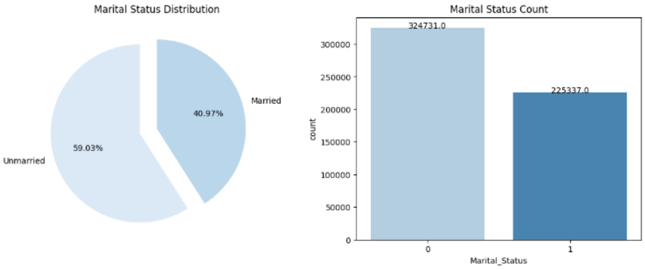


Fig 3.3

This plot *(fig 3.3)* illustrates the marital status of customers. It reveals that most of the customers are unmarried around 59% and the remaining are married around 41%.

b) Bivariate Analysis Results

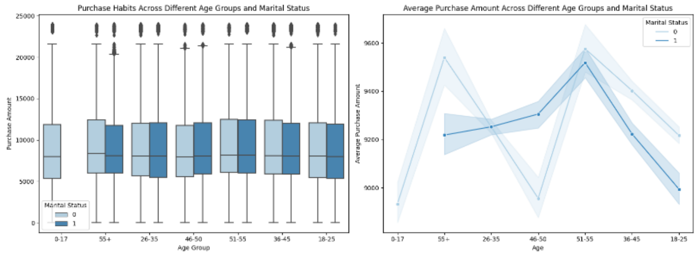


Fig 3.4

We observe a consistent pattern in *fig 3.4* where unmarried individuals tend to spend more than their married counterparts across all age groups. Notably, the age group [50-55] emerges as the highest spending group, regardless of marital status. Conversely, the least spending is observed in the [0-17] and [46-50] age groups, particularly among those who are not married.

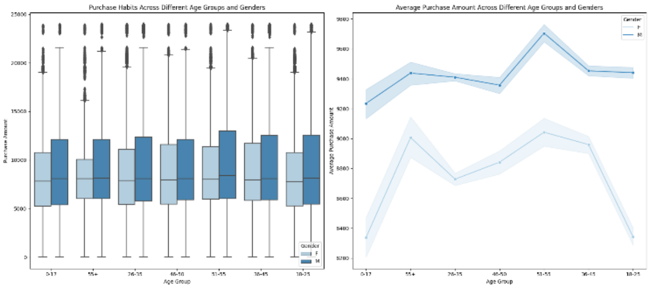
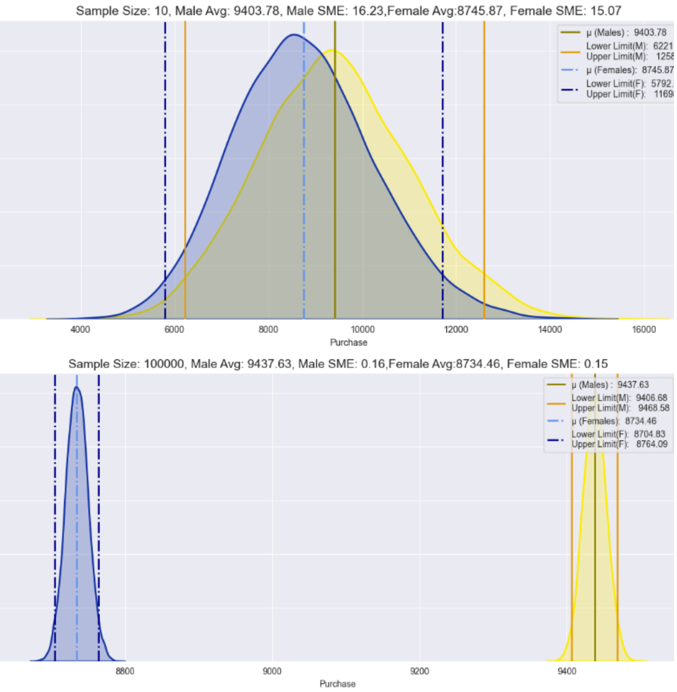


Fig 3.5

A consistent trend emerges as we observe in *fig 3.5* that males consistently exhibit higher spending compared to females across all age groups. Notably, the age group [50-55] stands out as the highest spending bracket for men, while for women, it is [50+]. Conversely, the lowest spending among women is observed in the [0-17] and [18-25] age groups around $8300 and for men the lowest spending is around $9200.

c) CLT Analysis Results

- CLT Analysis with respect to Gender:

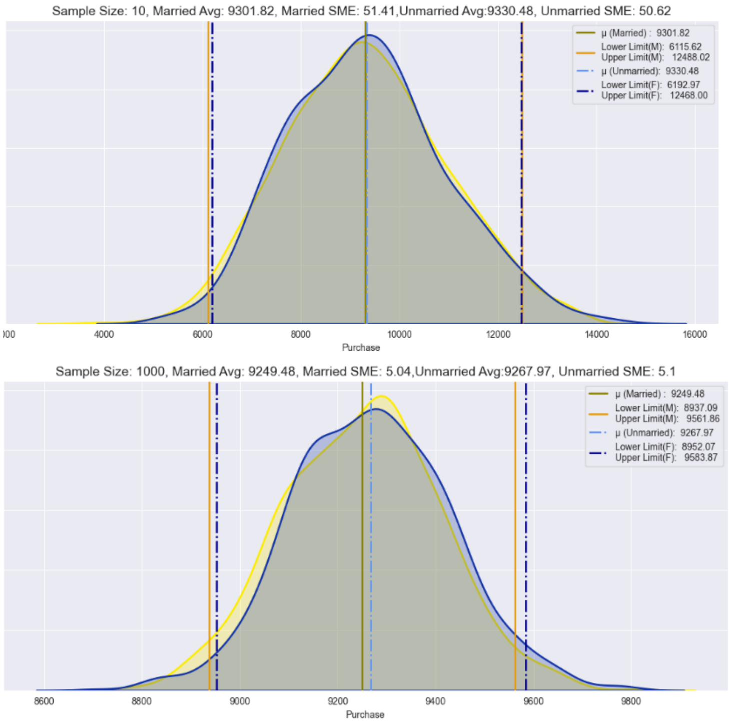

Sample Size: 10, Male Avg: 9403.78, Male SME: 16.23,Female Avg:8745.87, Female SME: 15.07

Sample Size: 100000, Male Avg: 9437.63, Male SME: 0.16,Female Avg:8734.46, Female SME: 0.15

Fig 3.6

We can see from the figures that as the sample size increases the difference between both the gender becomes more pronounced. The 95% confidence interval for the population mean of male spending is relatively narrow, ranging from $9,406.68 to $9,468.58. However, the 95% confidence interval for the population mean of female spending is narrow, spanning from $8,704.83 to $8,764.09. Thus, we can conclude that male spend relatively higher compared to women.

- CLT Analysis with respect to Marital Status:


Sample Size: 10, Married Avg: 9301.82, Married SME: 51.41,Unmarried Avg:9330.48, Unmarried SME: 50.62

Sample Size: 1000, Married Avg: 9249.48, Married SME: 5.04,Unmarried Avg:9267.97, Unmarried SME: 5.1

Fig 3.7

Here, using CLT we can see in *Fig 3.7* that increasing the sample size does not result into strong difference between the values of married and unmarried groups. Unmarried group tends to spend 9267.97$ and the married ones spend 9249.48$. Thus, we can see that the difference between them is not significant.

- CLT Analysis with respect to Age:

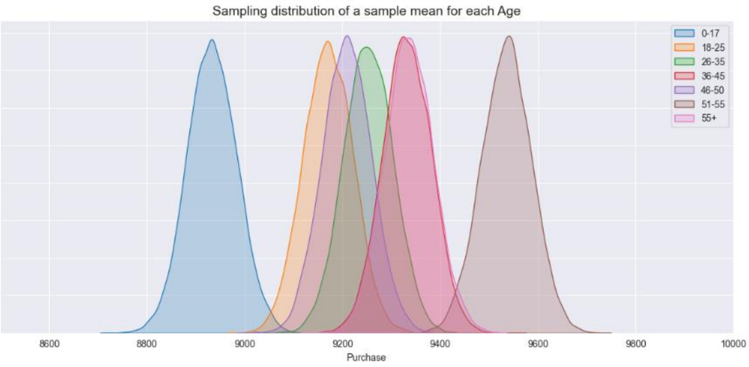

Sampling distribution of a sample mean for each Age

Fig 3.8

Based on the analysis of the Age group which was divided into 7 groups. We can conclude from *fig 3.8* that customers in the age group of 0-17 spent the lowest compared to the age group of 51-55 which spent the highest. The other age groups has almost similar spending.

V. Conclusion

In this research project, we delved into understanding customer purchasing behavior at Walmart using a diverse customer data. The study employed a variety of analytical techniques, including univariate analysis, bivariate analysis, and the application of the Central Limit Theorem (CLT). The univariate analysis revealed that men tend to be more significant spenders than women, with the highest spenders belonging to the 50+ age group and being unmarried. Bivariate analysis showed us that unmarried individuals consistently exhibited higher spending across all age groups, and males tended to spend more than females, especially in [50-55] bracket. The implementation of the Central Limit Theorem allowed for a more robust statistical analysis, providing confidence intervals around the sample means for spending patterns within different demographic categories. The results confirmed that, with an increasing sample size, the spending difference between genders became more pronounced, emphasizing that males generally spend more than females. Furthermore, the analysis of marital status indicated a relatively minor spending difference between married and unmarried individuals. In the age category, customers in the 51-55 age group were identified as the highest spenders, while those in the 0-17 age group spent the least. These findings contribute valuable insights for decision-making and enabling Walmart and similar businesses to tailor marketing strategies, inventory planning, and customer experiences to the specific preferences and behaviors of their customers. Overall, this research provides a comprehensive understanding of consumer behavior which helps in making useful decisions.

VII. References

1. Tengli A, Srinivasan SH. An Exploratory Study to Identify the Gender-Based Purchase Behavior of Consumers of Natural Cosmetics. *Cosmetics*. 2022; 9(5):101. https://doi.org/10.3390/cosmetics9050101
2. Pakasi, A., & Tumiwa, J. (2016). Comparison analysis between male and female of consumer purchase behavior of Yamaha Mio. *Jurnal EMBA: Jurnal Riset Ekonomi, Manajemen, Bisnis dan Akuntansi*, *4*(1).
3. Ronaghi, M., Danae, H., & Haghtalab, H. (2013). Survey of effects of gender on consumer behavior; case study on mobile phone. *International Journal of Advanced Studies in Humanities and Social Science*, *1*(8), 1024-1033.
4. Abir Smiti, A critical overview of outlier detection methods, Computer Science Review, Volume 38,2020, 100306, ISSN 1574-0137, https://doi.org/10.1016/j.cosrev.2020.100306. (https://www.sciencedirect.com/science/article/pii/S1574013720304068).