Walmart Business Case Study

1. Checking the structure & characteristics of the dataset -Structure and characteristics

2. Descriptive analytics - Descriptive

3. Outlier checking, univariate, bivariate, multivariate analysis - Answer

4. Confidence interval analysis for Male and Female
    a. CI 90% (Male vs Female) - Answer
    b. CI 95% (Male vs Female) - Answer
    c. CI 99% (Male vs Female) - Answer

5. Confidence interval analysis for Marital Status
    a. CI 90% (Marital Status) - Answer
    b. CI 95% (Marital Status) - Answer
    c. CI 99% (Marital Status) - Answer

6. Confidence interval analysis for Age category
    a. 0 - 17
        i. CI for 0-17 group (90%) - Answer
        ii. CI for 0-17 group (95%) - Answer
        iii. CI for 0-17 group (99%) - Answer
    b. CI for 18 - 25 group - Answer
    c. CI for 26 - 35 group - Answer
    d. CI for 36 - 45 group - Answer
    e. CI for 46 - 50 group - Answer
    f. CI for 51 - 55 group - Answer
    g. CI for 55+ group - Answer

ASSUMPTIONS

1. CI calculations have been conducted on sample size = 300,3000,30000, with 1000 iterations. These numbers are considered optimal for the project

OBSERVATIONS

1. The dataset is complete with no missing values.
2. The 26-35 age group constitutes the highest purchasing share.
3. City Category B is the most prevalent category.
4. A significant mean-median discrepancy in purchases suggests potential outliers.
5. Purchases vary widely, ranging from 12 to 23961, with a mean of 9264.
6. 75% of purchases are below 12054, indicating a common trend towards lower amounts
7. Conversion of integer-type categorical variables to character types may be beneficial.
8. Male entries make up 72% of the dataset but contribute 76% of total purchases.
9. Females, at 28% of the population, contribute only 23% of the overall purchase sum.
10. Occupations 0, 4, and 7 exhibit higher total purchase amounts.
11. Product Categories 1, 5, and 8 are more frequently purchased.
12. Multivariate analysis highlights comparable spending behavior between genders
13. Most spending aligns within 5k to 12k range across age, occupation, city, and stay duration.
14. Variation in product categories is evident, with Category 10 products being the costliest.
15. When considering the same sample size, it becomes evident that the range between the upper limit and the lower limit of a confidence interval is directly proportional to the chosen confidence level.
16. As the sample size increases while maintaining a constant confidence interval, the range between the upper and lower limits of the confidence interval tends to decrease.
17. With an increase in sample size, the sample mean tends to converge closer to the population mean.
18. The confidence **intervals do not overlap when analyzing purchase data based on gender**.
19. In contrast, clear overlapping is observed when analyzing purchase data based on marital status. This might be due to the almost 50% distribution of purchase we observed earlier

RECOMMENDATIONS

1. Strategies need to be implemented focusing on closing the gap between female and male purchase. 50 % contribution should be the target
2. Company can focus on selling more products falling into 1,5,8 category , as these are in high demand

3. 26 -35 age group spends more money, which can be classified as youth. This population tends to have disposable income and impulsive buying behavior. This is very important category to concentrate upon
4. 5k-12k is the range where maximum volume is taking place. So it'll be easier to sell those products that fall in this category and attract new customers.
5. Occupations 0, 4, and 7 exhibit higher total purchase amounts, a more detailed analysis needs to be done in the sales data and pushing sales through these channels can increase revenue