Business Case: **Walmart**

Walmart is an American multinational retail corporation that operates a chain of supercenters, discount departmental stores, and grocery stores from the United States. The management team wants to analyze the customer purchase behavior (specifically, purchase amount) against the customer’s gender & the other various factors to help the business make better decisions, if the spending habits differ between male & female customers.

**Analysing basic metrics**

1. From google.colab import files
2. df.read\_csv(‘walmart\_data.txt’)
3. df.head(10)
4. df.info()
5. df.isnull().sum()

**Shape & Data Types of all the Attributes**

1. df.shape

2. df.dtypes

**Conversion of Categorical attributes**

df['City\_Category']= df['City\_Category'].astype('category')

df['Gender'] = df['Gender'].astype('category')

**Statistical Summary**

1. df.describe()

**Non- Graphical Analysis**

1. df[‘Product\_Category’].value\_counts()

2. df.nunique()

**Visual Analysis- Univariate & Bivariate**

**For Univariate -Continous Variable**

**(Histogram)**

sns.histplot(x= 'Stay\_In\_Current\_City\_Years', data = df, kde = True)

plt.title('Stay Range')

plt.xlabel('Stay\_In\_Current\_City\_Years', fontsize = 12)

plt.ylabel('Count', fontsize = 12)

plt.show()

**(Distplot)**

sns.displot(df[‘Purchase’],kde = True, bins  = 30)

plt.title('Distribution of Purchase Amount')

plt.xlabel('Purchase Amount', fontsize = 12)

plt.ylabel('Density',fontsize = 12)

plt.show()

**(Countplot)**

sns.countplot(x= 'Age', data = df, order= sorted(df['Age'].unique()))

plt.title('Distribution of Age')

plt.xlabel('Age Groups',fontsize = 12)

plt.ylabel('Count', fontsize = 12)

plt.show()

**For Categorical Variable**

sns.boxplot(x = 'Gender', y = 'Purchase',data = df)

plt.title('Purchase Distribution by Gender')

plt.xlabel('Gender', fontsize = 12)

plt.ylabel('Purchase Amount', fontsize = 12)

plt.show()

sns.boxplot(x='Age', y='Purchase', data=df)

plt.title('Purchase Distribution by Age Group')

plt.xlabel('Age Group', fontsize = 12)

plt.ylabel('Purchase Amount', fontsize = 12)

plt.show()

**For Correlation- Heatmap & Pairplot**

**Heatmap**

corr = df.select\_dtypes(include=['number']).corr()

plt.figure(figsize=(10,6))

sns.heatmap(corr,annot=  True, cmap= 'coolwarm', fmt = '.2f')

plt.title('Correlation Heatmap')

plt.show()

**Pairplot**

sns.pairplot(df, vars=['Purchase', 'Occupation])

plt.show()

**Missing Value & Outlier Detection**

**Missing Value**

missing\_values = df.isnull().sum()

missing\_percentage  = (df.isnull().mean())\*100

missing\_data = pd.DataFrame({'Missing Values': missing\_values, 'Percentage': missing\_percentage})

print(missing\_data)

**Outlier Detection**

numerical = df.select\_dtypes(include = ['float64','int64']).columns

Q1 = df[numerical].quantile(0.25)

Q3 = df[numerical].quantile(0.75)

IQR= Q3-Q1

lower\_bound = Q1-1.5\*IQR

upper\_bound = Q3+1.5\*IQR

outlier = df[(df[numerical]<lower\_bound)|(df[numerical]>upper\_bound)].any(axis = 1)

df[~outlier]

**Business Insights (Non-Graphical & Visual Analysis)**

**Range of Attributes & Distribution:**

1. **Gender:** There are only 2 categories (Male & Female). Male customers are dominant in the dataset.
2. **Age:**  (‘0-17’.’18-25’,’26-35’,’36-45’,’46-50’,’51-55’,’55+’), these are 7 age groups covering all customer demographics. Most customers belongs to **’26-35’** age group.
3. **Occupation:**  It is encoded as numerical values & the range is 0 to certain max value(masked).Some occupations shop more frequently than others.
4. **Marital\_Status:** It is encoded as (0 as Unmarried & 1 as Married) where married customers show slightly higher spending trends.
5. **Product\_Category:** It has various product categories & some categories are more frequently purchased than others.

**Relationship between variables**

**Gender vs Purhcase:** Males generally spend more on average than females.

**Age vs Purchase:** The 26-35 age group spends the most & the older one which is 51+ tend to spend less.

**Occupation vs Purchase:** Certain occupations show higher spending patterns than others.

**Marital Status vs Purchase:** Married customers tend to spend slightly more than unmarried customers.

**Comments on Univariate & Bivariate Plots**

**Histogram:** It shows that more customers have stayed in their current city for 1-2 years. There are fewer customers who have been in the city for 4+ years. Walmart could focus marketing efforts on newer city residents, as they form a large portion of customers.

**Displot:** It has a right skewed distribution, most purchase amounts are lower, with fewer value transactions. There are presence of high-value outliers, suggesting the existence of premium shoppers. Personalized offers may boost revenue.

**Countplot:** The highest number of shoppers is from 26-35 age group & 0-17 & 55+ have the least participation.Walmart should create more promote codes or promotions targeting young adults as they are the most engaged customers.

**Boxplot (Gender vs Purhase):** Males have a wider range of spending & females have lower median spending but fewer extreme outliers. Marketing strategies could include gender based promotions.

**Boxplot (Age vs Purhase):**  The 26-35 age group has the highest median purchase amount & the widest range of spending. On the other hand 18-35 and 36-45 age groups also show moderate spending but the 0-17 & 51+ groups have lower median purchase amounts means they spend less. Walmart could introduce discounts for senior citizens..

**Heatmap**

**(Product\_Category vs Purhase):**  As the product category increases, purchase amount decreases, this could indicate that higher-numbered product categories tend to be lower-priced items.

**(Occupation vs Purhase):**  Occupation does not strongly impact purchase behavior.

**(Marrital\_Status vs Purhase):**  Being married or unmarried does not significantly affect purchase amounts.

**(User ID):**  It has no impact on other variables.

**Answering Questions (4)**

**Ans 1:**  No, Males spend more per transaction on averages than females. The boxplot shows males have a higher median and wider range of spending. T-Test confirms statistical significance (p<0.05). Males might purchase higher-value products (electronic,gadgets) .Females could make more frequent but lower-value purchases.

**Ans 2:**  Confidence Interval for Male Spending: (9422.019402055814, 9453.032678888716) & Confidence Interval for Female Spending: (8709.21132117373, 8759.92020913722). If the confidence intervals do not overlap, men and women have significantly different spreading behaviors but if they overlap, their spreading patterns are not significantly different.

**Ans 3:**  The upper bound of females spending (8759.92) is lower than the lower bound of male spending (9422.02), their confidence intervals do not overlap. This means the difference in spending between males and females is statistically significant.

(Implications for Walmart)

1. Since men spend more, Walmart could introduce premium product recommendations for male customers.
2. Walmart can focus on promotions, discounts or bundling strategies to encourage higher spending from female customers.

**Ans 4:** Married customers spend slightly more per transaction than unmarried ones. Confidence Intervals for married vs unmarried may overlap slightly,but if p<0.05, it is statistically significant. Box plot confirm that married customers have higher spending spread. Married customers may be shopping for family needs, buying in large quantities.

(Implication for Walmart)

1. Offer family-based promotions & bulk discounts for married customers.
2. Introduce ‘Household saver’ plans for married couples.

**Ans 5:** 26-35 age group spends the most, followed by 18-25 and 36-45. 51+ and 0-17 groups spend the least. T-Test and Confidence Intervals confirm significant spending differences. The 26-35 group is the most financially active (working professionals, young families).

(Walmart Implication)

1. Target younger adults with tech and lifestyle products.
2. Introduce senior citizen discounts for 51+ age groups.

**Final Insights: Exploration & CLT-Based Analysis**

**1. Distribution & Relationships Between Variables**

* **Purchase Behavior:** The purchase amount is right-skewed, meaning most transactions are lower, but a few high-value purchases exist.
* **Gender & Purchase:** Males spend more per transaction than females, with a statistically significant difference.
* **Age & Purchase:** The 26-35 age group has the highest spending, while the 0-17 and 55+ groups spend the least.
* **Marital Status & Purchase:** Married customers tend to spend slightly more than unmarried customers.

**Univariate & Bivariate Plot Observations**

* **Histograms & Distplots:** Show that most purchases are at the lower range, with a few outliers making expensive purchases.
* **Countplots:** Show that the highest number of shoppers fall into the 26-35 age category.
* **Boxplots:** Indicate significant spending differences based on gender, age, and marital status.
* **Heatmap:** Correlations are weak between most numerical variables, except for Product\_Category, which has a negative correlation with Purchase.

**Generalization to the Population (Using CLT)**

* The mean purchase amount follows a normal distribution when sample sizes increase.
* Confidence intervals suggest that male spending is consistently higher than female spending.
* Since confidence intervals do not overlap for gender, age, and marital status groups, we can generalize these spending trends to Walmart’s broader customer base.

**Recommendations**

**1. Gender-Based Promotions**

* Offer personalized discounts to **female customers** to encourage higher spending.
* Introduce premium product bundles or loyalty rewards for **male customers**, who already spend more.

**2. Age-Targeted Marketing**

* **Focus on 26-35 age group**, as they are the highest spenders.
* Provide **student or family discounts** for the 0-17 and 55+ age groups to boost their spending.

**3. Product & Category Optimization**

* Identify and **promote high-demand product categories** with targeted advertising.
* Consider **discounting slow-moving categories** to clear inventory.

**4. Marital Status-Based Offers**

* Offer **family-oriented deals** and **household bundles** for married customers.
* Provide **individual-centric promotions** for unmarried customers, such as personal care and electronics.

**5. Store & Online Strategies**

* Use insights from high-spending occupation groups to create **customized ads and promotions**.
* **Leverage e-commerce** by pushing personalized recommendations based on past purchases.