- **Google Colab Link:** Walmart Data Analysis

1. Defining Problem Statement and Analyzing basic metrics **(10 Points)**
    1. Observations on shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), statistical summary

Assumptions:

- I have taken sample size as 5 million instead of 50 million, since the run time is very high for 50 million sample for Marital_status and Gender confidence interval.
- For Age- confidence interval, I have considered the sample size to be 1 million.

Observations:

- The given dataset contains **550068** rows and **10** columns.
- User_ID – corresponds to unique ID of the customer.
- Product ID – corresponds to unique ID of the product.
- Gender – corresponds to the Gender of the customer.
- Age- corresponds to the age bucket that the customer belongs to.
- Occupation – corresponds to occupation of the customer.
- City_category – corresponds to category of the city.
- Stay_in_city_in_years – corresponds to the totals no.of years, that particular customer is reciding in that particular city.
- Marital_Status – corresponds to whether the customer is married or not.
- Product Category – corresponds to the product purchased by the customer.
- Purchase – corresponds to the sales amount of that product.

statistical summary:

- The minimum purchase value is **$12** , where as the maximum purchase value is **$23961** and the mean purchase value is **$9263.9**

2. Non-Graphical Analysis: Value counts and unique attributes

The dataset consists of:

- 5891 unique customers out of which 1666(28.2%) are Female and 4225(71.7%) are male.
- 3631 products.
- There are no null values in the given dataset.
- There age column is divided into 7 bins, and 39.9% of customers belong to 26-35, only 2.7% of customers belong to 0-17.

- There are 21 occupations. 12.5% of the customers belong to occupation-4 , only 0.28% belong to occupation-8.
- There are 3 city categories, where 53% customers belong to city C, 17% customers belong A.
- The Stay_in_current_city_in_years column consists of 5 unique values, where 35% have the value as 1 and 13% has value 0.
- There are 58% unmarried customers and 41.9% of married customers.
- There are 18 product categories, of which 1st category is most sold and 9th category is least sold.
- Total purchase value is $5095812742.

3. Visual Analysis - Univariate & Bivariate

▪ For continuous variable(s): Distplot, countplot, histogram for univariate analysis

Univariate Analysis:

- There are more men than women in the customers.
- The customers in the age group of (26-35) are more in number , followed by (36-45) and (18-25).
- Even though there are less no.of customers in city B, their purchases are more and the corresponding records are more, hence the histogram is showing B has high record count.
- There are more no.of customers who are residing in the current city for 1 year.
- There are more no.of unmarried customers than married.
- 1st category product is the most sold one , followed by 5 and 8….

▪ For categorical variable(s): Boxplot

Bivariate Analysis:

- For males the size of boxplot is bigger than that of female, where the median value is almost equal and female plot has more outliers than males.
- For Age Vs Purchase, almost all the age groups have outliers and the spread and median is also almost same, where 55+ age group has more no.of outliers.
- 12th and 17th occupations have more spread, and the median value of all the occupations is almost same, except 17th occupation, all the occupations have outliers.
- The median value of purchases is almost same for all the cites, where C category has more spread, and B has more outliers, but overall purchases from city B is more.
- For stay_in_current_city_in_years, the spread and median value of all the categories is almost same and all the plots have outliers.
- For married and unmarried, the median purchase value, spread of the plot is almost same and both the plots have outliers.

- The 10th product category has the highest median purchase value and more no.of outliers, where as 13th category has the least spread and least median value.

    - For correlation: Heatmaps, Pairplots

Heatmaps and Pairplots:

- From the heat map and pair plot , it is visible that there is almost no correlation between the columns.

2. Missing Value & Outlier Detection **(10 Points)**

Missing Value:

There are no null values in the given dataset.

Outlier Detection:

- For males the size of boxplot is bigger than that of female, where the median value is almost equal and female plot has more outliers than males.
- For Age Vs Purchase, almost all the age groups have outliers and the spread and median is also almost same, where 55+ age group has more no.of outliers.
- 12th and 17th occupations have more spread, and the median value of all the occupations is almost same, except 17th occupation, all the occupations have outliers.
- The median value of purchases is almost same for all the cites, where C category has more spread, and B has more outliers, but overall purcases from city B is more.
- For stay_in_current_city_in_years, the spread and median value of all the categories is almost same and all the plots have outliers.
- For married and unmarried, the median purchase value, spread of the plot is almost same and both the plots have outliers.
- The 10th product category has the highest median purchase value and more no.of outliers, whereas 13th category has the least spread and least median value.

3. Business Insights based on Non- Graphical and Visual Analysis **(10 Points)**

Comments on the range of attributes

- The User ID column has 5891 users in total.

- The Gender column has 1666 female customers and 4225 male customers in the dataset.
- Highest no.of purchases made by the customer is 1026 and lowest no.of purchases made by customer is 6.
- On an average a customer is purchasing 93-94 times in the given dataset.
- Product category -1 has highest no.of sold units, Product category -9 has least no.of sold units
- Product category -1 has highest no.of sales, product category -19 has lowest sales.

    - Comments on the distribution of the variables and relationship between them

- The highest no.of purchases made by male is 1026 and that of female is 752 for the given dataset.
- No.of purchases made by many customers are in the the range of 0-200.
- Only few customers had purchased more than 800 times for males and no females are purchasing more than 800 times.
- Product category 10 has more sales for male and females.
- Product 19 has the least sales among male and female.

    - Comments for each univariate and bivariate plot.

- Even though there are more no.of customers in city C, the sales in city C are less compared to city B.
- Customers of city B are spending more.
- Among all the age groups , the product category wise purchase shows similar pattern.
- There are more no.of customers in 26-35 years age bucket and are contributing to the highest sales.

4. Answering questions (50 Points)
    1. Are women spending more money per transaction than men? Why or Why not? **(10 Points)**

- The no.of males in the dataset are more compared to that of no.of female and women are spending less money than men per transaction.
- One reason could be women are less financially independent in comparison to that of men.

    2. Confidence intervals and distribution of the mean of the expenses by female and male customers **(10 Points)**

Note: the below confidence intervals correspond to 5million male population and 5million female population that have been calculated using the given sample.

Confidence intervals and distribution of mean of expenses by female:
Mean: 8735
The 90% confidence interval is 6365.2 ,11320.0
The 95% confidence interval is 5967.9 ,11852.6
The 99% confidence interval is 5231.3 ,12910.0
Confidence intervals and distribution of mean of expenses by male:
Mean: 9438.5
The 90% confidence interval is 6872.5 ,12170.3
The 95% confidence interval is 6431.1 ,12717.1
The 99% confidence interval is 5619.3 ,13785.4

3. Are confidence intervals of average male and female spending overlapping? How can Walmart leverage this conclusion to make changes or improvements? **(10 Points)**

- The confidence interval of men and women is overlapping.
- This indicates, that on 90% of times men are spending in the range of $6800 to $12000.
- 90% of times women are spending in the range of $6300 to $11000.
- It indicates that lower limit for women is less than that of men by $500.
- Upper limit for women is less than that of men by $1000.
- To increase the spending limit of women, special offers and discounts could be provided to women such as coupons.

4. Results when the same activity is performed for Married vs Unmarried **(10 Points)**

Confidence intervals and distribution of mean of expenses by Married:
Mean: 9260.7
The 90% confidence interval is 6741.6 ,11959.8
The 95% confidence interval is 6309.5 ,12508.402499999944
The 99% confidence interval is 5511.7 ,13586.800499999988
Confidence intervals and distribution of mean of expenses by Unmarried:
Mean: 9265.9
The 90% confidence interval is 6741.3 ,11967.8
The 95% confidence interval is 6311.5 ,12514.1
The 99% confidence interval is 5517.1 ,13589.0

Inference:

- The 90% confidence interval of married and unmarried customers is almost same.
- So same kind of offers could be given to both married and unmarried customers to increase the sales.

5. Results when the same activity is performed for Age **(10 Points)**

Age	Mean
51-55	9534.808031
55+	9336.280459
36-45	9331.350695
26-35	9252.690633
46-50	9208.625697
18-25	9169.663606
0-17	9534.808031

Confidence intervals and distribution of mean of expenses for Age (0 – 17):
The 90% confidence interval is 6366.2 ,11680.7
The 95% confidence interval is 5931.9 ,12240.9
The 99% confidence interval is 5118.3 ,13333.7005
Confidence intervals and distribution of mean of expenses for Age (18 – 25):
The 90% confidence interval is 6640.1 ,11876.0
The 95% confidence interval is 6209.9 ,12430.7
The 99% confidence interval is 5416.0995 ,13486.8005
Confidence intervals and distribution of mean of expenses for Age (26 – 35):
The 90% confidence interval is 6740.3 ,11946.5
The 95% confidence interval is 6312.597500000001 ,12495.3
The 99% confidence interval is 5532.5995 ,13567.301500000001
Confidence intervals and distribution of mean of expenses for Age (36– 45):
The 90% confidence interval is 6805.495000000001 ,12025.904999999993
The 95% confidence interval is 6371.8 ,12573.7
The 99% confidence interval is 5573.0 ,13640.3005
Confidence intervals and distribution of mean of expenses for Age (46 – 50):
The 90% confidence interval is 6720.7 ,11879.9
The 95% confidence interval is 6293.097500000001 ,12426.3
The 99% confidence interval is 5516.898999999999 ,13504.5005
Confidence intervals and distribution of mean of expenses for Age (51 – 55):
The 90% confidence interval is 6972.8 ,12265.4
The 95% confidence interval is 6529.8 ,12820.6
The 99% confidence interval is 5713.699 ,13890.6
Confidence intervals and distribution of mean of expenses for Age (55+):
The 90% confidence interval is 6821.7 ,12036.8
The 95% confidence interval is 6394.9 ,12586.702500000003

The 99% confidence interval is 5581.5995 ,13676.9

Inferences:
The 90% confidence interval is almost overlapping for all the categories.
Therefore, all the categories equally likely to purchase the products with almost same probability.
So, all the categories can be given same importance while giving discounts and offers.

5. Final Insights **(10 Points)** - Illustrate the insights based on exploration and CLT
    - Comments on the distribution of the variables and relationship between them

- According to the heatmap and pairplot, we can conclude that there exists almost no correlation between the attributes in the dataset.
- All most all the categories have similar purchase patterns with respect to age, marritals status and gender.
- Most of the customers are have the no.of purchases in 0-100 range.

    - Comments for each univariate and bivariate plots

- There are more unmarried customers.
- There are more men than that of women.
- Most no.of customers are in the age bucket of 26-35 and least are in the age bucket 0-17.
- 51-55 age bucket has the highest mean purchase value.

    - Comments on different variables when generalizing it for Population

- When generalized for the population, all the mean values are almost overlapping.
- There is no much deviation between the mean purchase value for the customers of different categories.
- Almost 85% of customers in the dataset are purchasing only 0-50 times. They can be given more offers so that they can purchase more no.of times.

6. Recommendations **(10 Points)**
    - Actionable items for business. No technical jargon. No complications. Simple action items that everyone can understand

- The sales are almost equal among all types of customers.
- The buying pattern is almost same and same types of discounts and coupons can be given to all the customers to increase the sales.
- At regular intervals , campaigns could be run across various platforms like youtube, google to take suggestions from the customers to know the services and products that the customers are looking for.