**Colab NoteBook Link**  
**https://colab.research.google.com/drive/1LqaMYsYQwKfltvkDXybZC7Am-YMTnY0z?usp=sharing**  
  
**PDF Notebook Link  
  
https://drive.google.com/file/d/1aPN3bNAo\_YDdR0Db0Rcj4vXMHk5b8VyI/view?usp=sharing**

**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).

**Dataset**

The company collected the transactional data of customers who purchased products from the Walmart Stores during Black Friday. The dataset has the following features:

|  |  |
| --- | --- |
| User\_ID: | User ID |
| Product\_ID: | Product ID |
| Gender: | Sex of User |
| Age: | Age in bins |
| Occupation: | Occupation(Masked) |
| City\_Category: | Category of the City (A,B,C) |
| StayInCurrentCityYears: | Number of years stay in current city |
| Marital\_Status: | Marital Status |
| ProductCategory: | Product Category (Masked) |
| Purchase: | Purchase Amount |

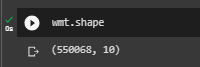
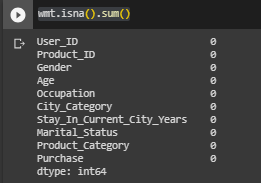
1. **Defining Problem Statement and Analyzing basic metrics** 
   1. **Observations on shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), statistical summary**
   2. **Non-Graphical Analysis: Value counts and unique attributes ​**
   3. **Visual Analysis - Univariate & Bivariate**
      * **For continuous variable(s): Distplot, countplot, histogram for univariate analysis**
      * **For categorical variable(s): Boxplot**
      * **For correlation: Heatmaps, Pairplots**

 We basically have to identify the Population transaction purchase amount against Gender, Marital Status and Age using the sample average of the purchase amount with 90%, 95%, 99% confidence interval. This will help Walmart in understanding the customers spending habits and take some meaningful decisions to further improve sales and their business.

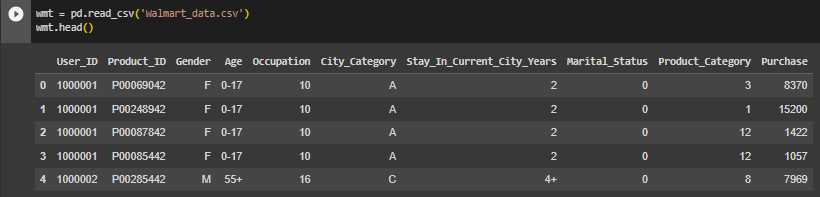
* Below can be an approach:

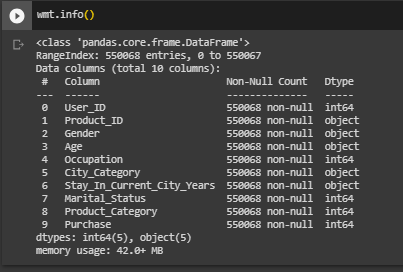
We will calculate the Population Transaction Purchase Amount with 90%, 95%, 99% confidence interval using Central Limit Theorem by utilizing the sample average of the dataset present. Apart from this we can also use Bootstrapping approach to find the range of Population Transaction purchase amount with the same confidence interval mentioned above against gender, marital status and age.

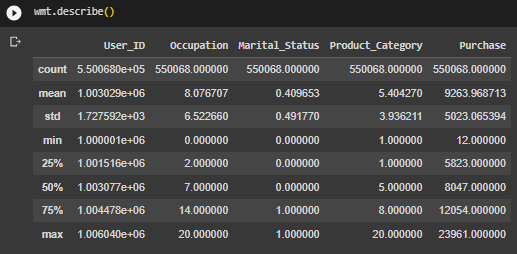
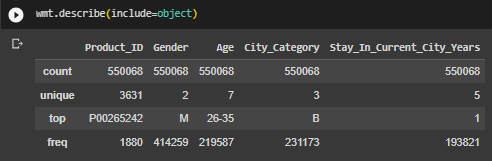
**Basic Metrics**:

* There are 550068 transaction records of customer purchases in the dataset.
* There are 10 columns in the dataset.
* The Shape of the data is (550068, 10).  
  
* There are no-null values present in any column User\_ID, Product\_ID, Gender, Age, Occupation, City\_Category, Stay\_In\_Current\_City\_Years, Marital\_Status, Product\_Category, Purchase  
    
  **wmt.isna().sum()** **np.any(wmt.isna().any(axis=1))**

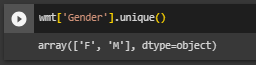
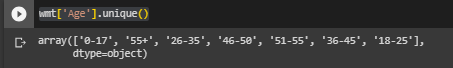
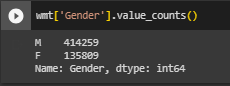
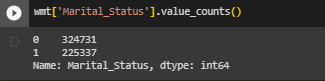
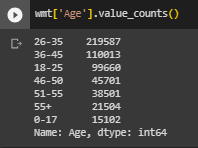


* Top 5 rows in the data set to have a glimpse of data.    
    
  
* Column**Product\_ID, Gender, Age, City\_Category and Stay\_In\_Current\_City\_Years** are **object**types (since they contain **string**), all other columns are**Integer types.**



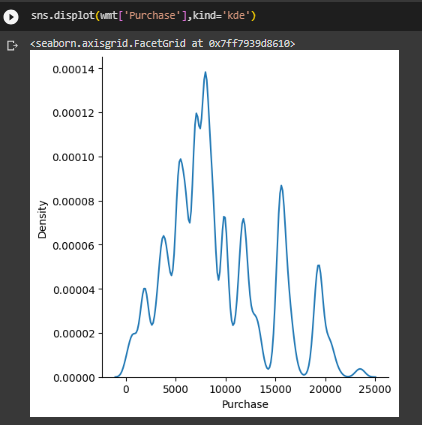
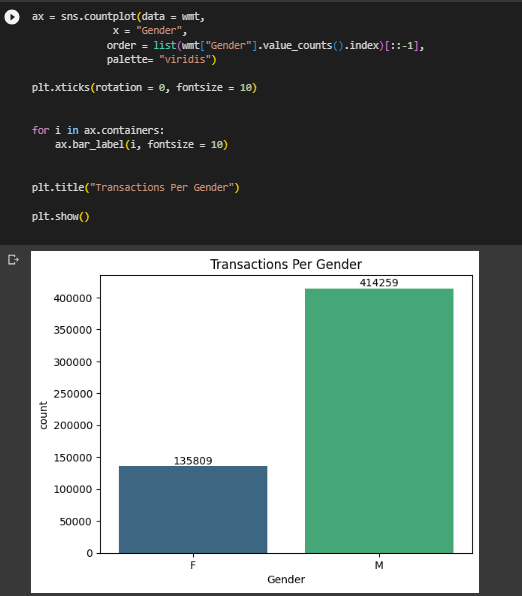
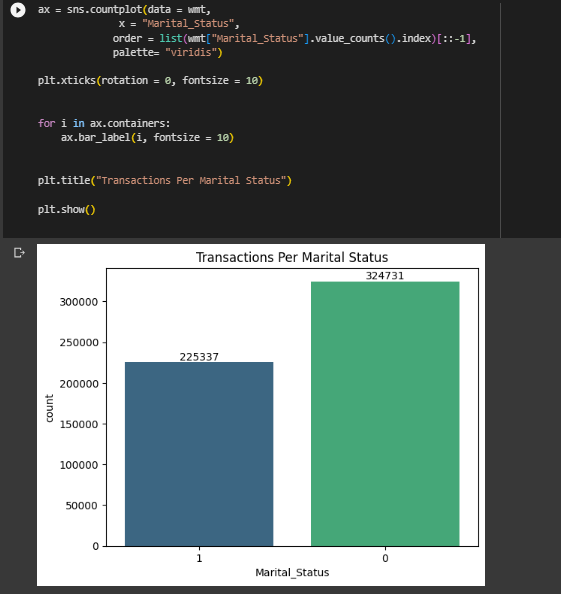
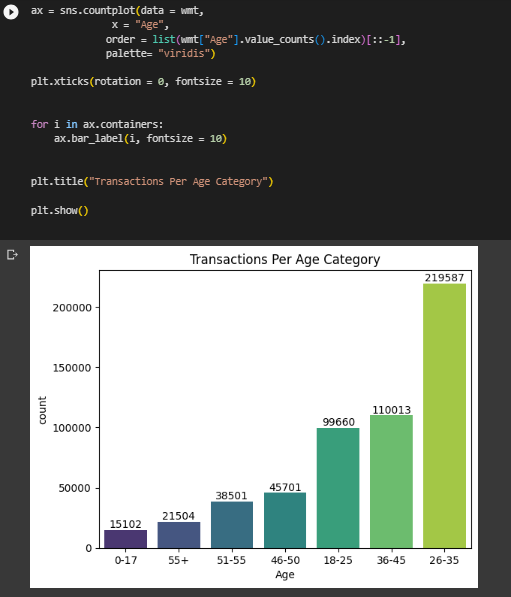
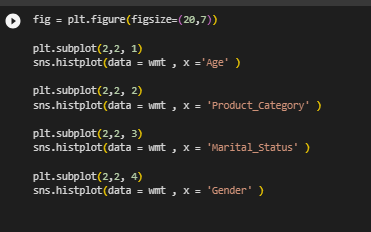
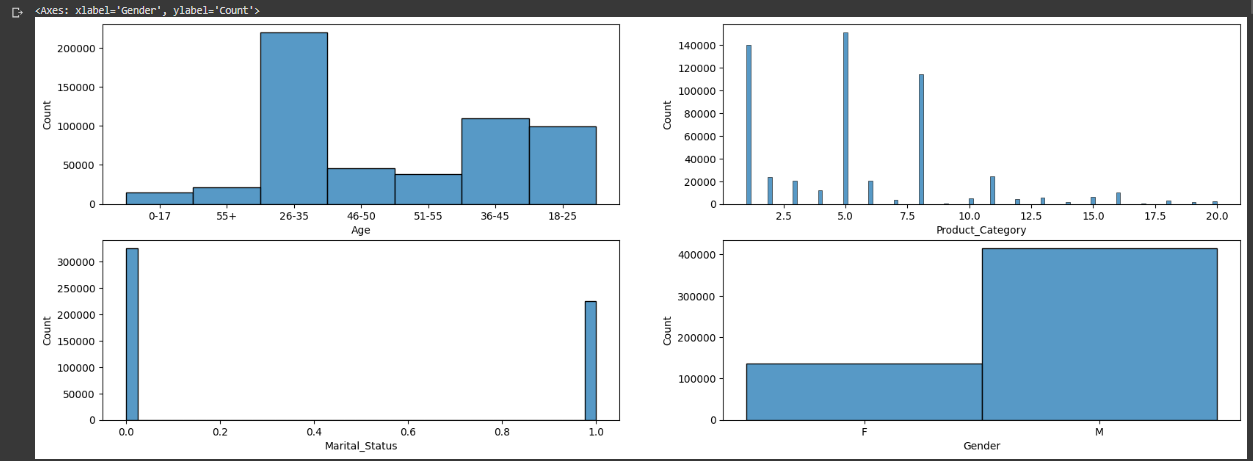
* The statistical summary can be found using **describe**method mentioned below.  
    
    
    
  

​​**1.2** **Non-Graphical Analysis: Value counts and unique attributes**

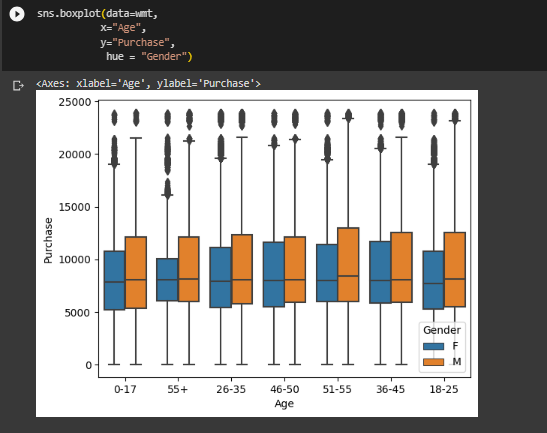
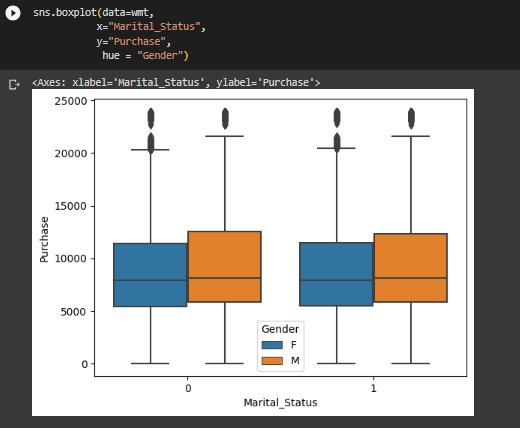
* There are 2 types of masked Marital\_Status 0,1.  
    
  
* There are 2 Unique genders (M and F)  
  
* There are 7 Unique Category of Age (['0-17', '55+', '26-35', '46-50', '51-55', '36-45', '18-25']) ****
* Transaction Value\_Counts Per Gender: There are 414259 transaction for males and 135809 transactions for females.****
* Transaction Value\_Counts Per Marital Status: There are 324731 transaction for unmarried customer and 225337 transactions for married customers.****
* Transaction Value\_Counts Per Age Category:  Maximum Transactions are done by Age group of 26-35 followed by 36-45 and 18-25. ****

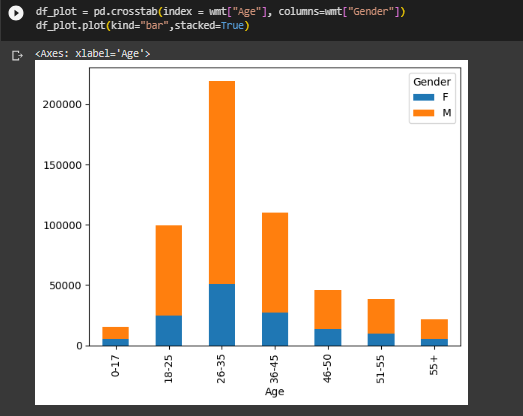
**1.3** **Visual Analysis - Univariate, Bivariate after pre-processing of the data**  
  
For continuous variable(s): Distplot, countplot, histplot for univariate analysis

\*Covering **3.1, 3.2. 3.3** as well.

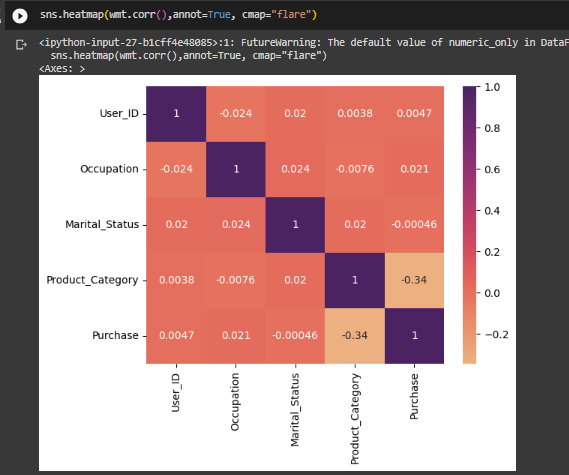
* **Most customers Transaction purchase amount is between 5000-10000 dollars.   
    
  **
* **There are 414259 transactions for males and 135809 transactions for females.   
    
    
  **
* **Unmarried customers have done more transaction.   
  Note : Assumption Marital\_Status 0 denotes unmarried customers  
    
  **
* **Minimum number of transaction is done by customers of Age group 0-17 and maximum by Age group of 26-35**
* **There are more male customers than female customers in the dataset.**
* **  
  **

For categorical variable(s): **Boxplot**Covering **3.1, 3.2. 3.3** as well

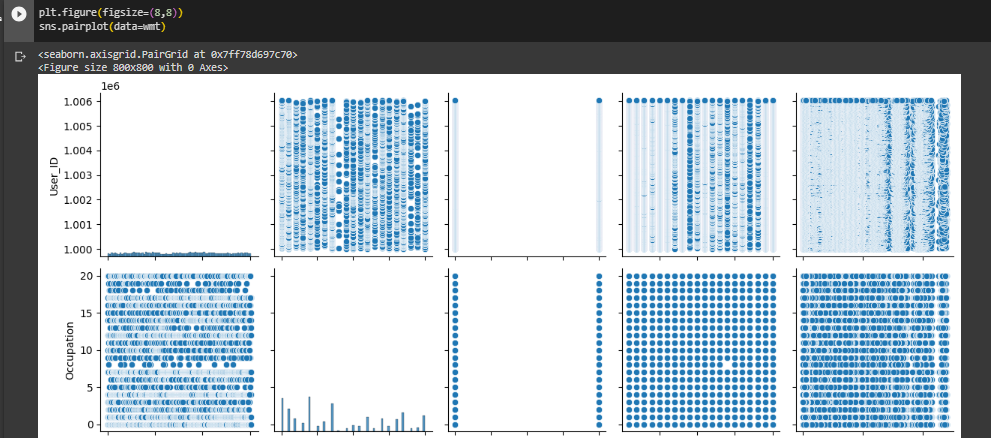
* The median age of male customers is mostly same or slightly higher than for female customers in all age groups.   
    
  
* The median age of male customers is slightly more than for female customers for both Married and Unmarried customers.
* 
* **Categorical Vs Categorical using CountPlot with stacked Columns**
* Male customers transaction records are much more than female customers across all age groups.

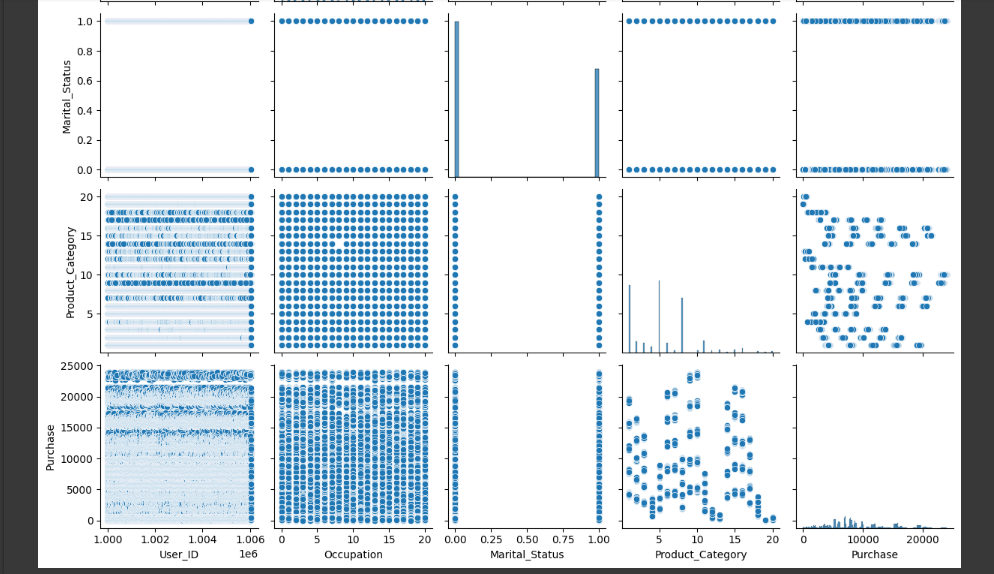


For correlation: Heatmaps, PairplotsCovering **3.1, 3.2. 3.3** as well  
  
**Correlation using Heatmaps  
  
There is not much correlation between columns with integer datatype**

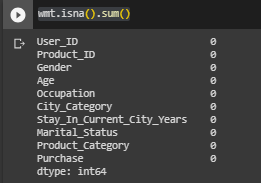


**Correlation using PairPlot**





**2. Missing Value & Outlier check (Treatment optional)**

* There are no-null values present in any column User\_ID, Product\_ID, Gender, Age, Occupation, City\_Category, Stay\_In\_Current\_City\_Years, Marital\_Status, Product\_Category, Purchase  
    
  **wmt.isna().sum()** **np.any(wmt.isna().any(axis=1))**

**Outlier using BoxPlot**

#Outlier

fig = plt.figure(figsize=(20,7))

plt.subplot(2,2, 1)

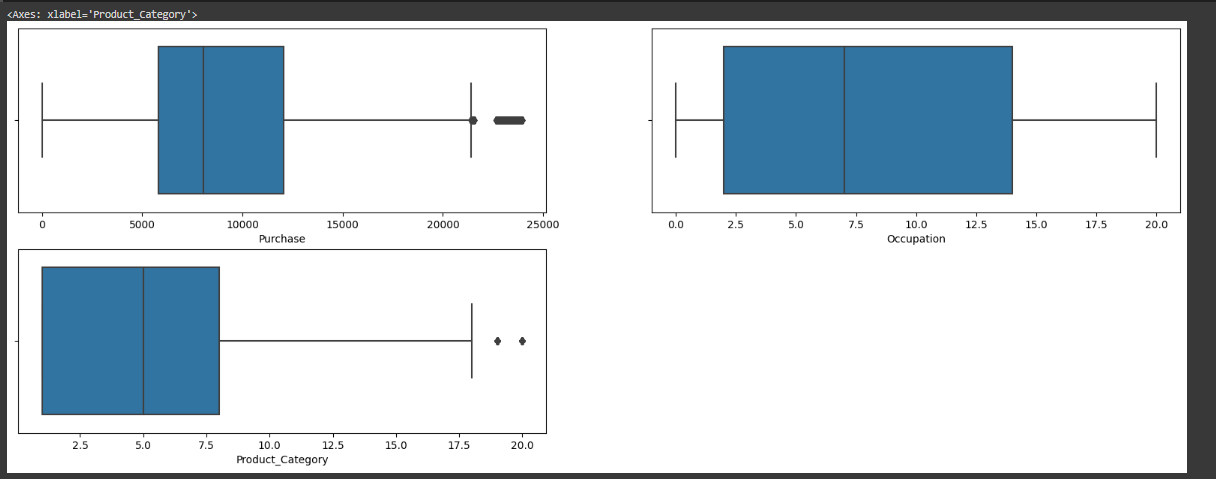
sns.boxplot(data = wmt , x ='Purchase' )

plt.subplot(2,2, 2)

sns.boxplot(data = wmt , x = 'Occupation' )

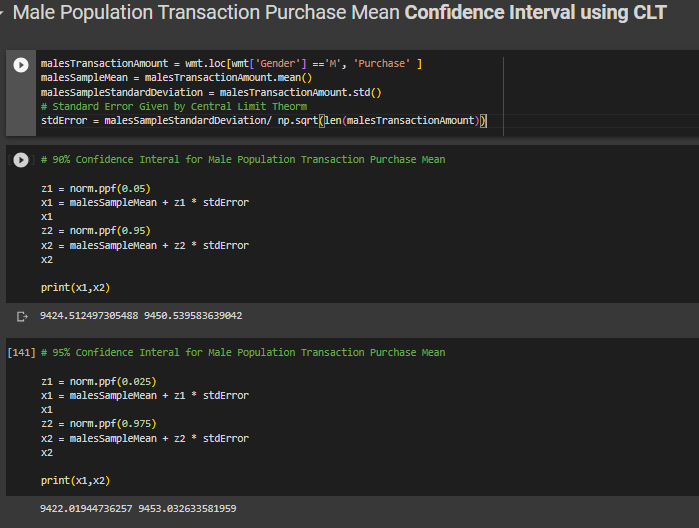
plt.subplot(2,2, 3)

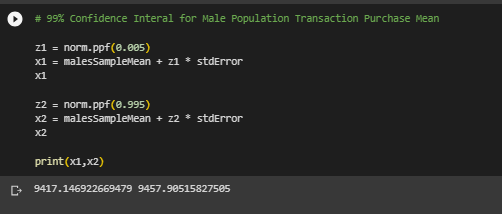
sns.boxplot(data = wmt , x = 'Product\_Category' )

* Most customers are having transaction purchase amount between 5000 and 10000 with median purchase amount of around 7000 dollars.

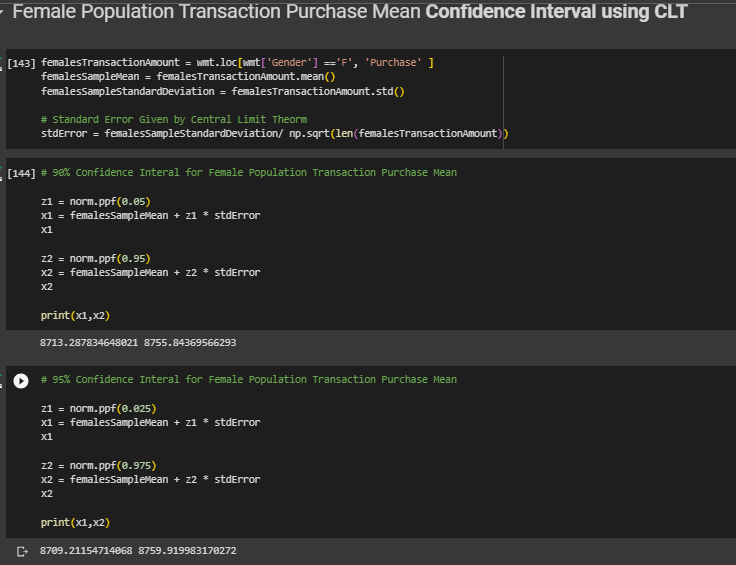
4. Answering questions

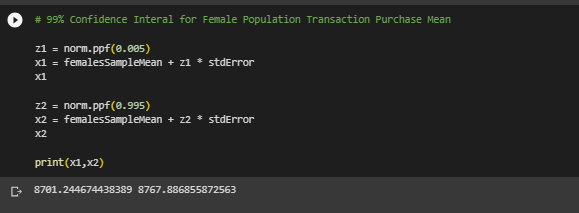
4.1 Are women spending more money per transaction than men?

As per CLT, Male **Population** Transaction Purchase Mean range is (9424.51, 9450.54), (9422.02, 9453.03), (9417.15, 9457.90) for 90%, 95% and 99% Confidence Interval respectively.  
  




As per CLT, Female **Population** Transaction Purchase Mean range is (8713.29, 8755.84), (8709.21, 8759.92), (8701.24, 8767.88) for 90%, 95% and 99% Confidence Interval respectively.

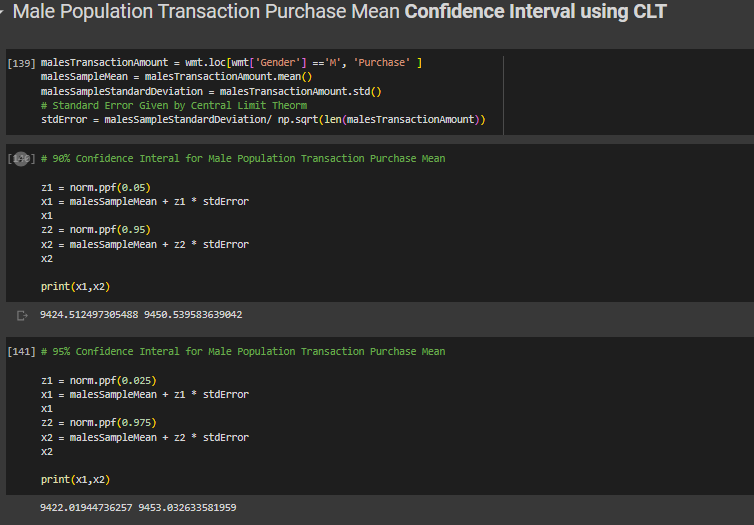


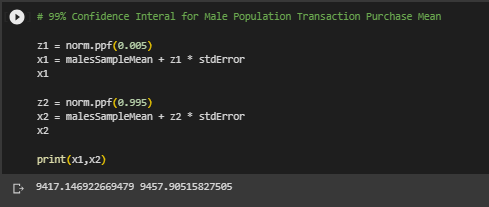


This shows the Average Male Population Transaction Purchase Mean is greater than those of females and males spend more. However, when this analysis is tried using bootstrapping approach with sample size of 10, 30 and 100 with number of samples as 100000 we do not get any conclusive evidence, since the distribution of sample mean for 90%, 95% and 99% confidence interval is overlapping for males and females.

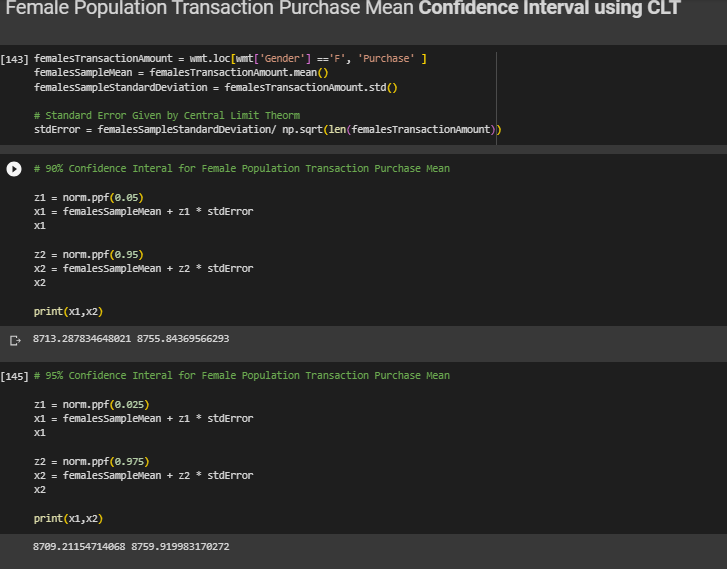
4.2 Confidence intervals and distribution of the mean of the expenses by female and male customers?

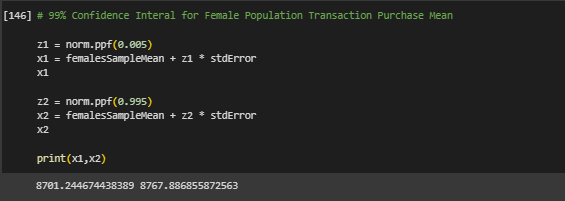
**CONFIDENCE INTERVAL using CLT for Males**

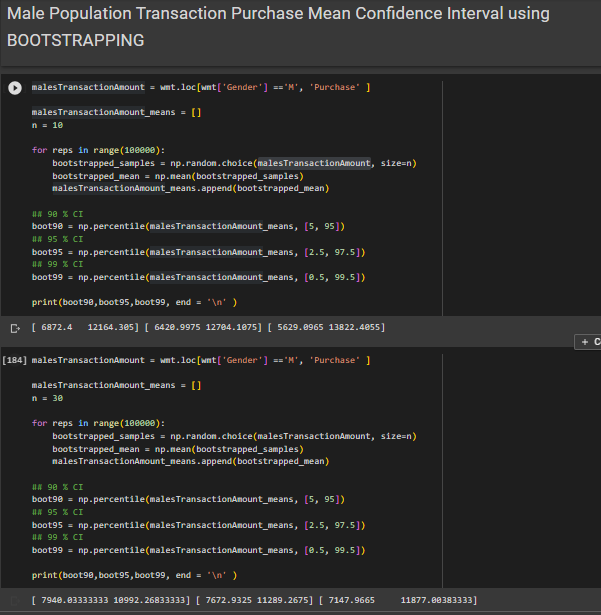
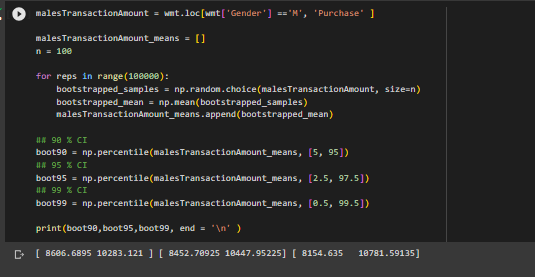
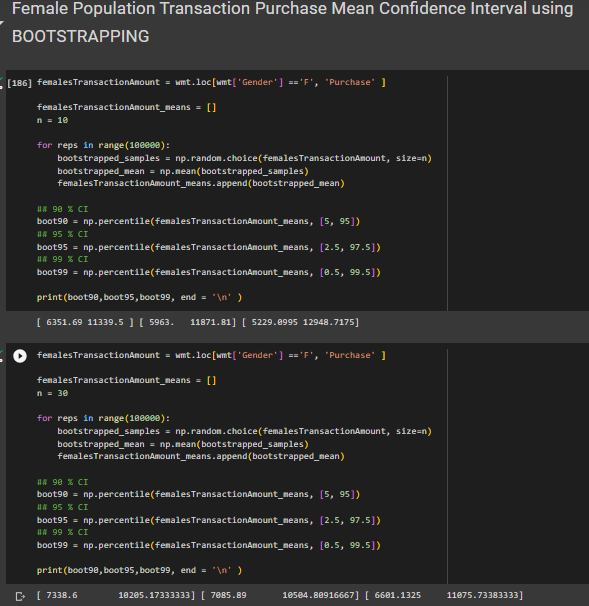
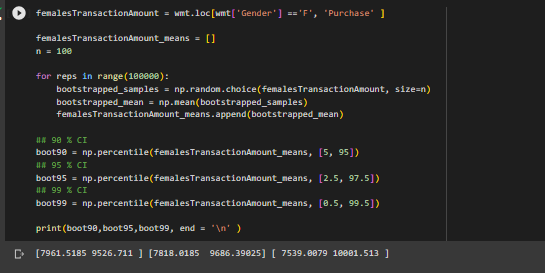
****

****

**CONFIDENCE INTERVAL using CLT for Females**

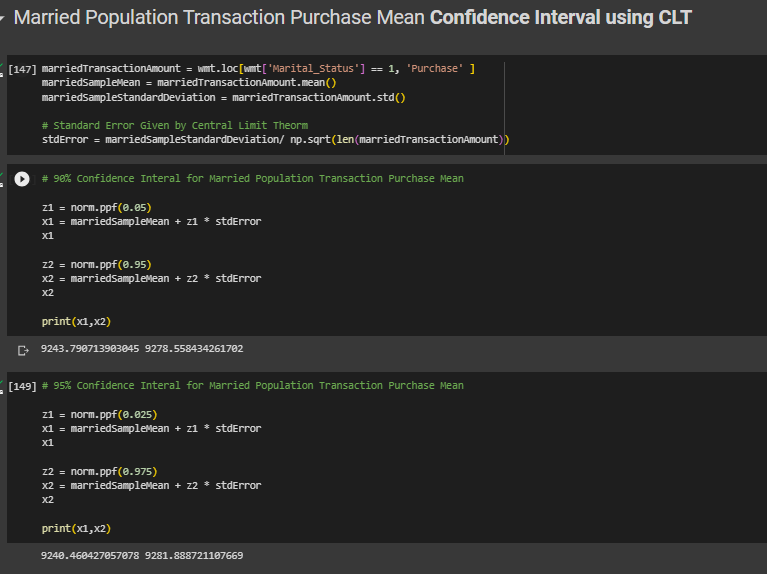
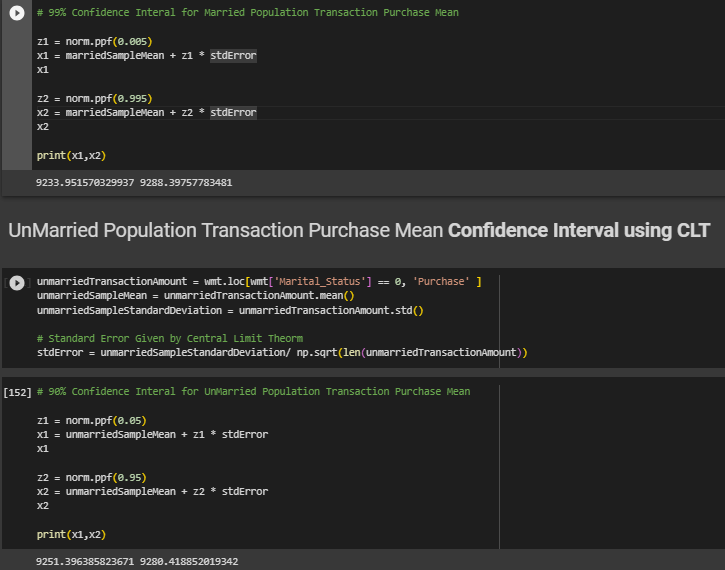
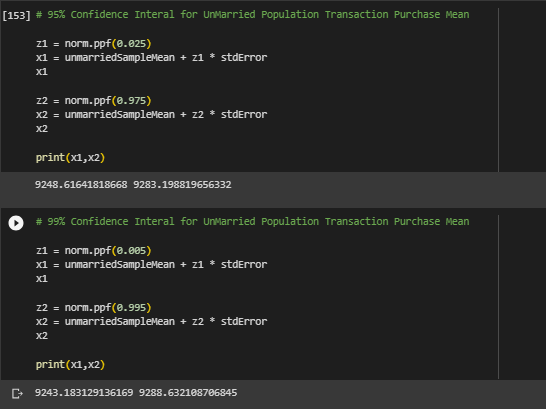
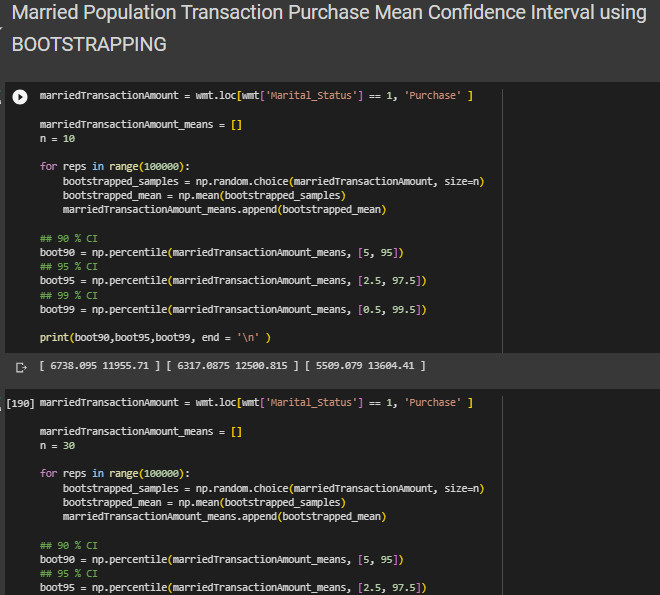
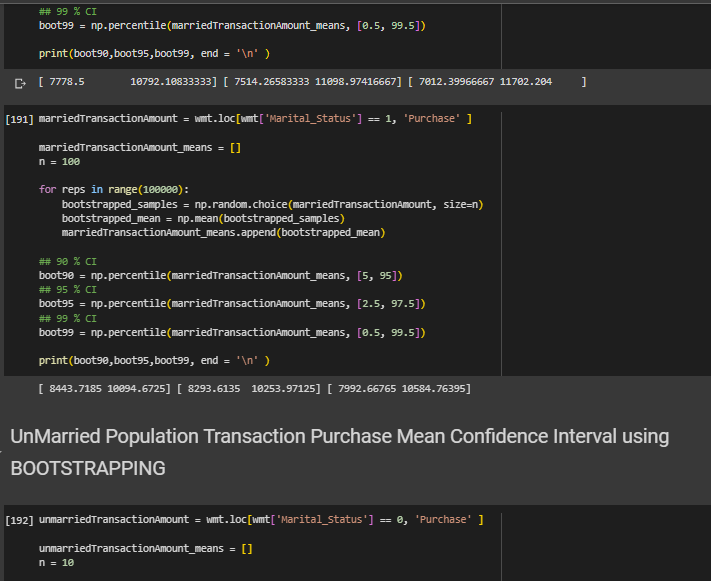
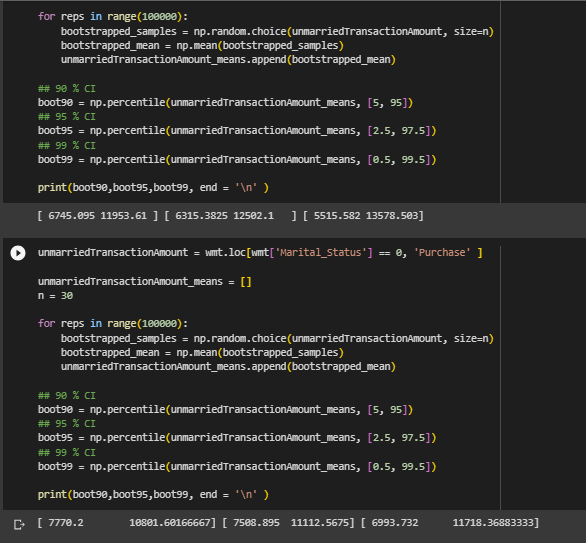
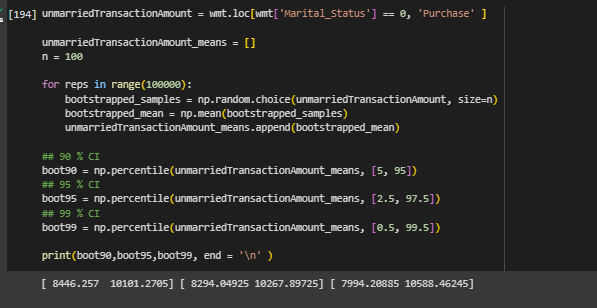
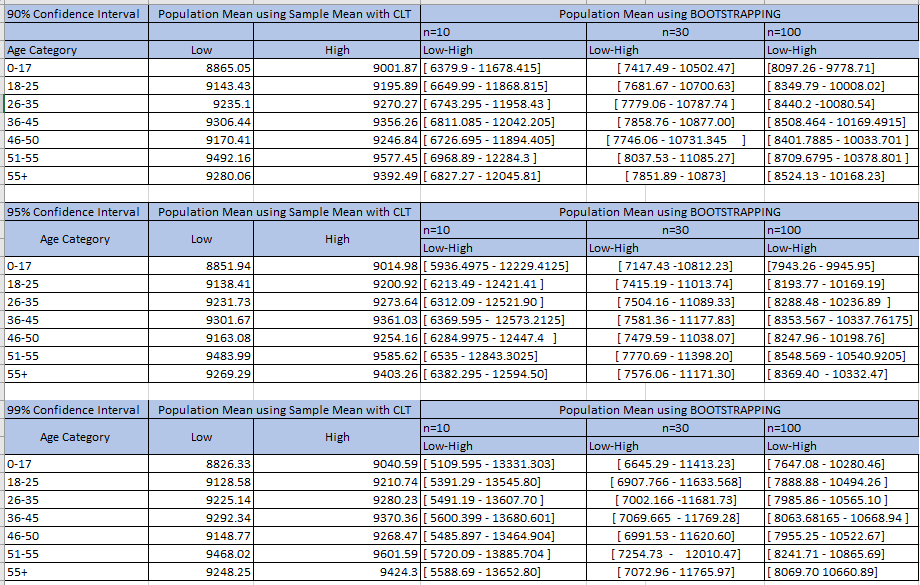
****

****

**CONFIDENCE INTERVAL using BOOTSTRAPPING for Males  
  
  
  
  
  
CONFIDENCE INTERVAL using BOOTSTRAPPING for Females  
  
  
  
**4.3 Are confidence intervals of average male and female spending overlapping? How can Walmart leverage this conclusion to make changes or improvements?

Confidence intervals of average male and female spending are overlapping based on the Bootstrapping approach.  
However, using the provided dataset sample mean and CLT, the average male population mean is greater than those of females.  
  
Walmart can use this insight to create more marketing campaigns/offers to drive more spending from female customers.

4.4 Results when the same activity is performed for **Married vs Unmarried**

* + **Confidence Interval using CLT for Married and Unmarried customers**. (Assumption Marital\_Status 0 is considered Unmarried)  
      
      
      
      
      
    As per CLT, Married **Population** Transaction Purchase Mean range is (9243.79, 9278.55), (9240.46, 9281.88), (9233.95, 9288.39) for 90%, 95% and 99% Confidence Interval respectively.  
      
    As per CLT, Unmarried **Population** Transaction Purchase Mean range is (9251.39, 9280.41), (9248.61, 9283.19), (9243.18, 9288.63) for 90%, 95% and 99% Confidence Interval respectively.
  + **Confidence Interval using BOOTSTRAPPING for Married and Unmarried customers**  
      
      
      
      
      
      
      
    In both scenario of calculating population mean spending for married and unmarried customers, using sample mean with CLT and using bootstrapping approach the distribution range is overlapping for all 3 confidence intervals 90%, 95%, 99%.
* Results when the same activity is performed for Age  
    
    
    
  As we can see the range of distribution of sample mean decreases as we increase the sample size from 10 to 30 to 100.  
  Number of samples considered is 100000 for BOOTSTRAPPING APPROACH.

5. **Business Insights**

* Average Male Population Transaction Purchase Mean is greater than those of females and males spend more. However, when this analysis is tried using bootstrapping approach with sample size of 10, 30 and 100 with number of samples as 100000 we do not get any conclusive evidence, since the distribution of sample mean for 90%, 95% and 99% confidence interval is overlapping for males and females.
* In both scenario of calculating population mean spending for married and unmarried customers, using sample mean with CLT and using bootstrapping approach the distribution range is overlapping for all 3 confidence intervals 90%, 95%, 99%.
* Using sample mean and CLT below are the Population Mean of Purchase Amount based on the Age Groups with 95% Confidence interval.

|  |  |  |
| --- | --- | --- |
| 95% Confidence Interval | Population Mean using Sample Mean with CLT | |
| Age Category | Low | High |
|  |
| 0-17 | 8851.94 | 9014.98 |  |
| 18-25 | 9138.41 | 9200.92 |  |
| 26-35 | 9231.73 | 9273.64 |  |
| 36-45 | 9301.67 | 9361.03 |  |
| 46-50 | 9163.08 | 9254.16 |  |
| 51-55 | 9483.99 | 9585.62 |  |
| 55+ | 9269.29 | 9403.26 |  |

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

* **Recommendation 1:**Confidence intervals of average male and female spending are overlapping based on the Bootstrapping approach. (Sample Size = 30, Number of samples = 100000)  
    
  However, using the provided dataset sample mean and CLT, the average male population mean is greater than those of females.  
    
  Walmart can use this insight to create more marketing campaigns/offers to drive more spending from female customers.
* **Recommendation 2:**With 95% Confidence Interval Age Group from 51-55 spend the most amount per transaction. Special Offers/Campaigns can be created to attract these customers.
* **Recommendation 3**: With 99% Confidence Interval Age Group from 51-55 spend the most amount per transaction. Special Offers/Campaigns can be created to attract these customers.
* **Recommendation 4:**With 95% Confidence Interval Age Group from 0-17 spend the least amount per transaction. No Cost EMI offers/ Marketing campaigns can be created to attract these customers and increase their spending.