**Name:** Aloysius Lobo

**Batch:** Morning – DSML BeginnerFeb23

**Case-Study: Walmart – Confidence Interval and CLT**

**IMPORTING BASIC LIBRARIES**

import pandas as pd

import numpy as np

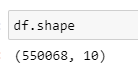
import matplotlib.pyplot as plt

import seaborn as sns

from scipy.stats import norm, binom, ttest\_ind, ttest\_1sample

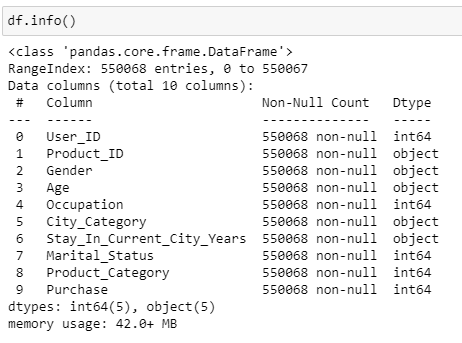
**# Applying basic steps**

1. Checking total rows and columns in the dataset



* There are total **550068 rows** and **10 columns** in the entire dataset.

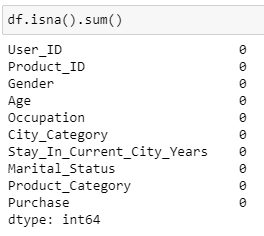
1. Acquiring dataset info like columns and their data type, total null-values in each column, etc.



**Observations:**

1. Product\_ID, Gender, Age, City\_Category and Stay\_In\_Current\_City\_Years have ‘object’ datatype i.e. they have categorical values. Remaining columns have int64 datatype i.e. they have numerical values.
2. The dataset does not contain any null values.

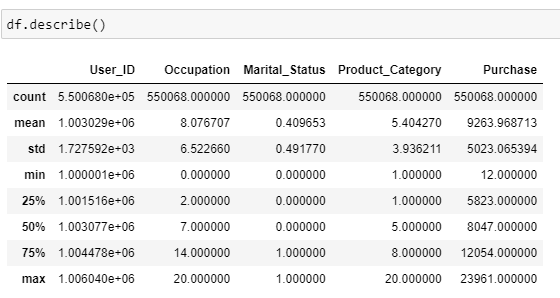
**Missing/Null values**

****

**Observation –**

* No missing values found in the dataset.

**# Getting statistical info about the dataset**

****

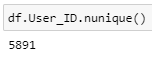
* Average amount of purchase done by a customer is $ 9263.96 where the min purchase is $ 12.00 and max purchase is $ 23961.00.

**# Number of unique products in the dataset**

****

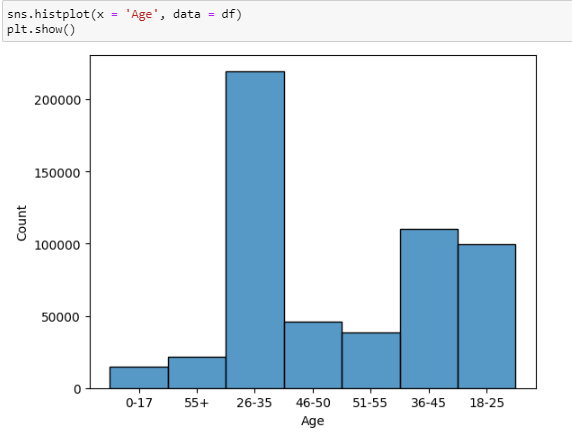
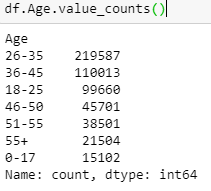
There are 20 unique products in the dataset.

**# Number of users in the dataset**

****

There are 5891 unique users in the given Walmart dataset.

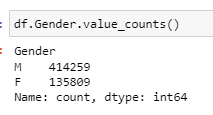
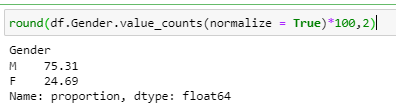
**# Calculating value counts of every data**

1. **Age**

**Observations –**

* Age group of 0-17 yrs are the least contributors to sales in Walmart as their total count is 15102 while max purchases are done by age group of 26-35 years.

1. **Gender – *Customer purchases based on gender***

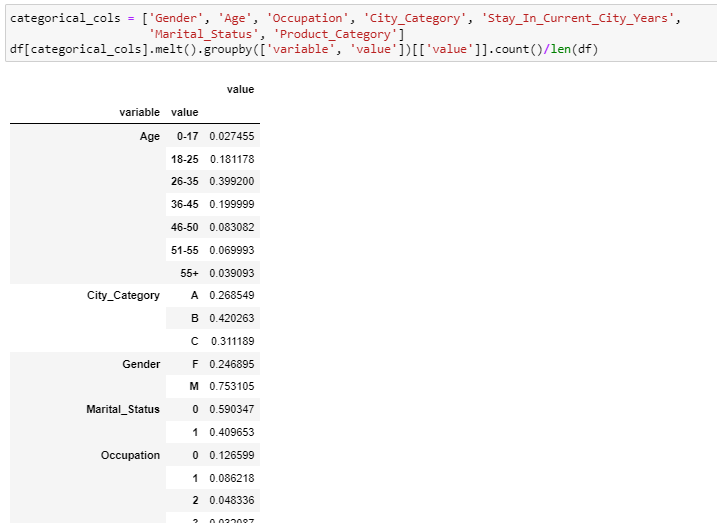
****

**Observation –**

Above analysis show that Male customers purchase more than female customers.

Around 75.31 % Males customers do purchasing in Walmart compared to Female which is 24.69% only.

1. **Value\_counts**

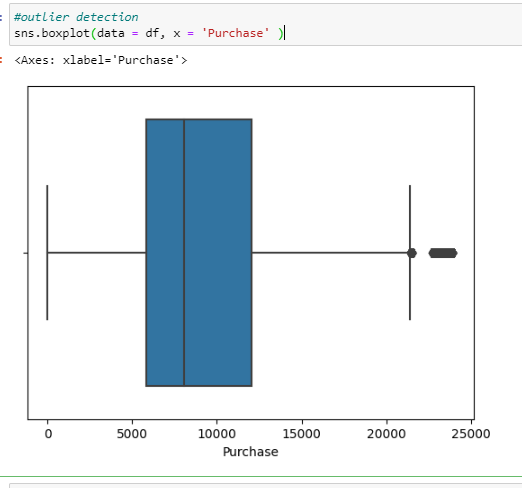


**Observations –**

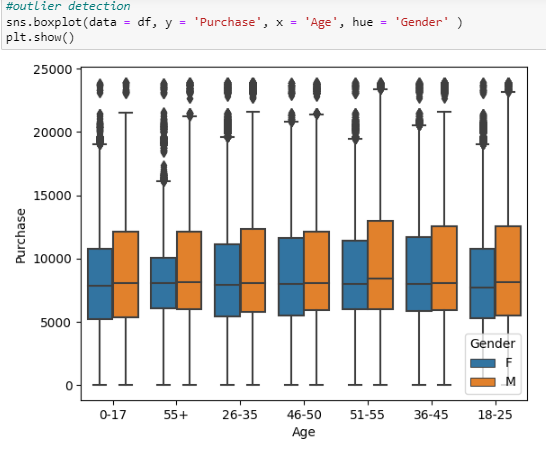
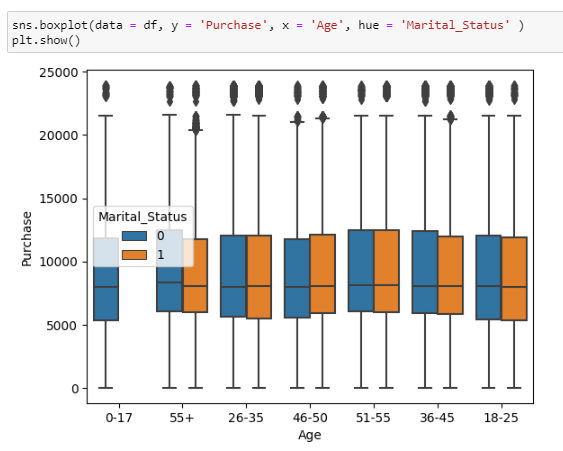
* ~18% users fall in the age-group 18-25.
* 40% users fall in the age-group 26-35
* 20% users fall in the age-group 36-45
* ~75% purchases are done by male customers
* ~25% purchases are done by female customers
* ~60% purchases are done by married customers
* ~40% purchases are done by unmarried customers
* There are total 20 product categories in the dataset.

**# Outlier detection using boxplots**

1. Purchase



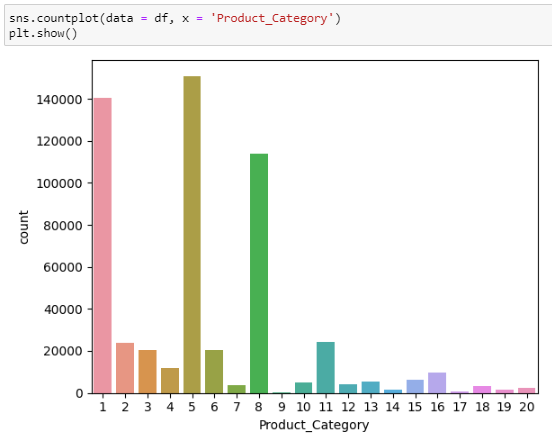
1. **Purchase vs Age w.r.t Gender and Purchase vs Age w.r.t Marital\_Status**



**Observations –**

Outliers are seen in both the plots, where maximum outliers are observed for purchases of female customers of the age-group 55+ and the age-group 18-45.

**# Frequency of purchases based on Product\_Category**

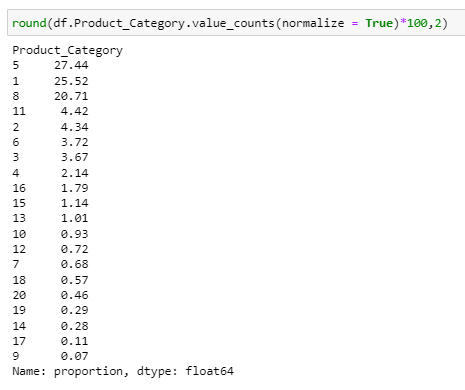


* **Categories 1, 5 and 8 have maximum purchases.**

**Recommendations –**

Walmart must focus on the sales of these product categories.

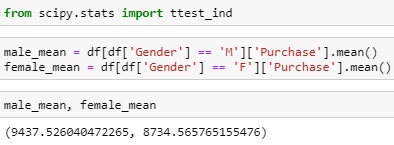
**# No of customer purchases based on Product-type**

****

**Insights:**

* Product Categories 1,5 and 8 are the maximum purchases.
* Thus, least popular product among the customers is Product\_Category 9.

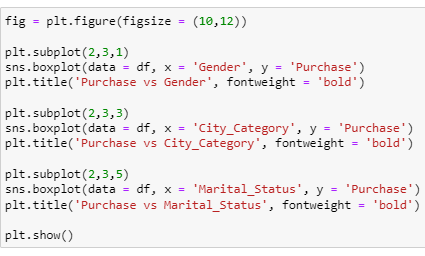
**# Average spending of Male and Female customers**

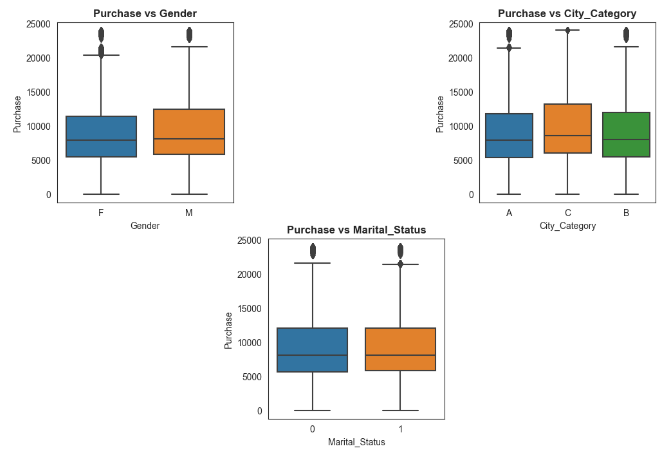
****

**Observations –**

The average mean purchase of male customers is large compared to Female customers.

**Bivariate Analysis**

****

****

**Observations –**

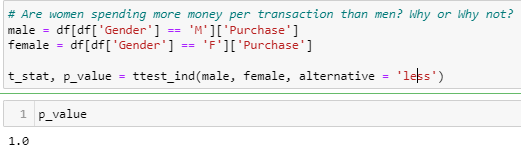
* The above bivariate plots give the spending of customers based on their Gender, City they live and their Marital\_Status.
* The median spending done by Male and Female customers are similar. However, male customers are seen to spend more than female customers.
* Also, the city category C has the highest spending compared to other two categories.
* Observations on spending based on Marital\_Status infer that spending does not depend on the marital status of the customer i.e. both married and unmarried customers spend equally.

**Recommendations –**

Walmart must focus more over the sales acquired by Male customers in City - C compared to others.

**Checking if the data is true using null hypothesis with a 95% confidence interval.**

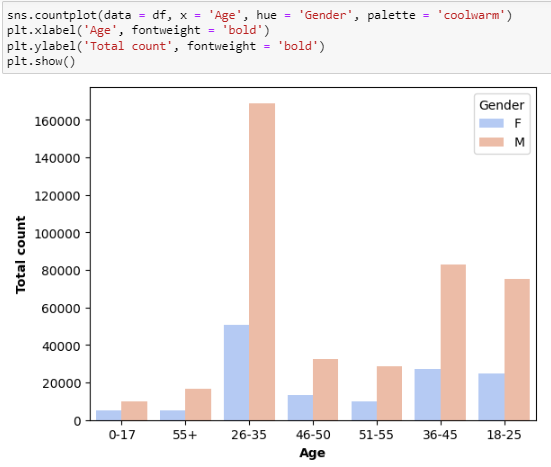
**Q. Are women spending more money per transaction than men? Why or why not?**

****

**Observations –**

As the p\_value > significance (0.05), hence we fail to reject the null hypothesis i.e. women are not spending more money than men.

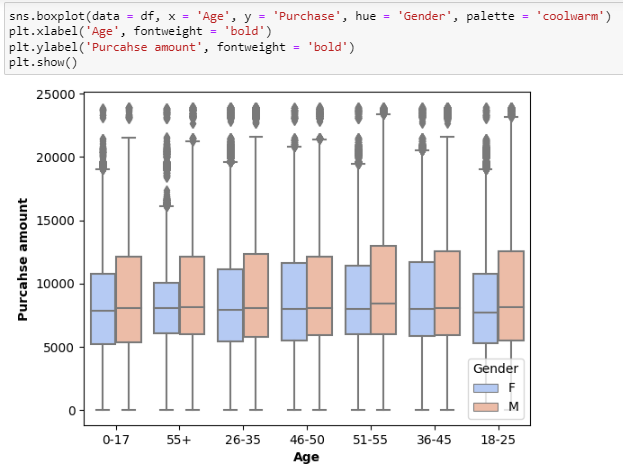
**# Customer purchase based on Age and Gender**

****

**Observations:**

* The above plot shows that Male customers of every age group are the maximum spending ones compared to females.
* Maximum purchases are done by age group 26-35.

**# Analysis on purchase amount based on age and gender of customers**



**Observations –**

* The box plot shows almost same median in the amount of purchases done by both Male and Female customers of every age group which means that total spending done by male and female are same and in the range upto 22K.
* Some outliers exist in every age group that spend upto 25K max.
* Male customers are seen to have fewer outliers compared to females. i.e. only fewer female customers spend more than 20K.

**Recommendation –**

* Walmart is doing well as the average spending of male and female are similar. Thus, product category in Walmart are not biased and it caters the need for both male and female.
* Walmart must increase offers and discounts on purchases of most female products to increase their total purchase.

**Q] Confidence Interval**

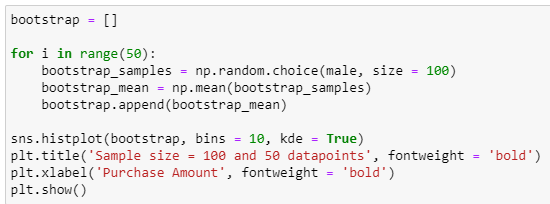
To find the population mean of the expenses for both male and female customers, we take samples from purchase data of both male and female and check whether the population mean lie within the interval for 90%, 95% and 99% confidence level.

**Checking the distribution of the mean spending –**

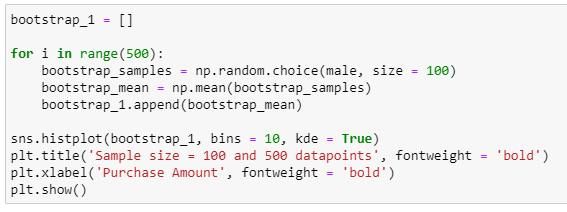
1. **Based on Gender**

**Male customers**

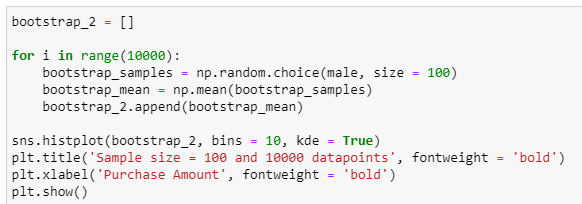
1. Taking sample size = 100 and checking for 50 datapoints.



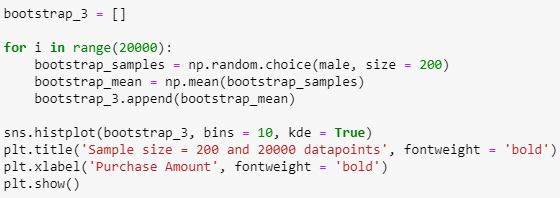
1. Taking sample size = 100 and 500 datapoints



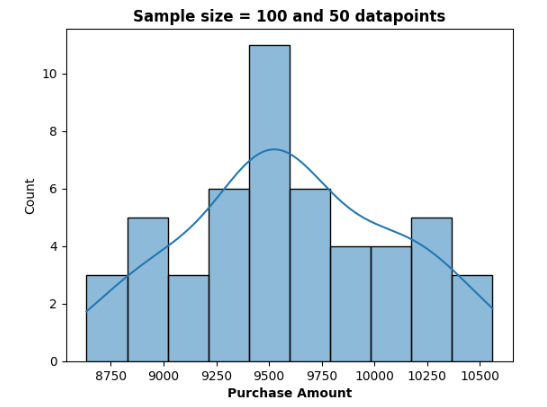
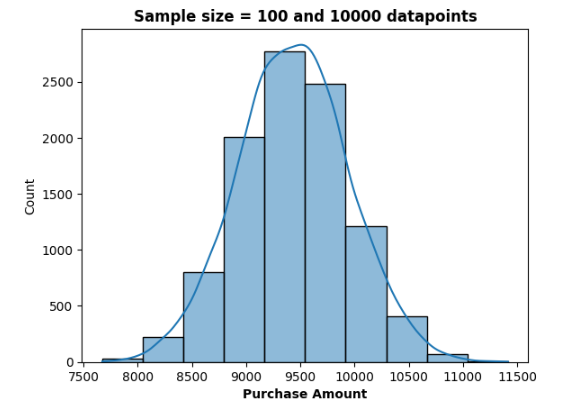
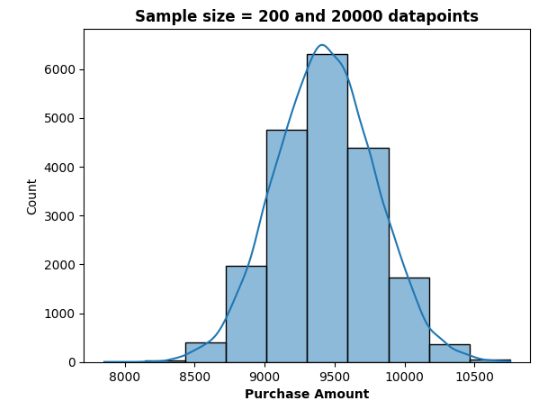
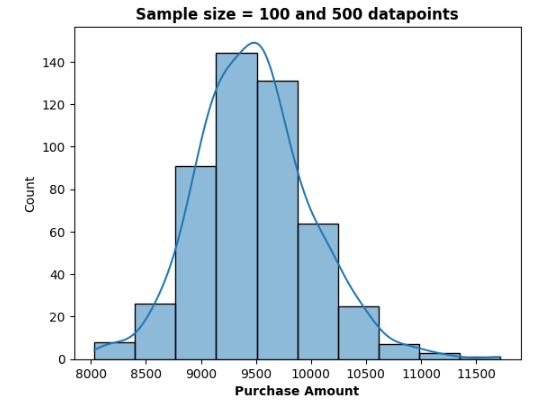
1. Taking sample size = 100 and 10,000 datapoints



1. Taking sample size = 200 and 20,000 datapoints



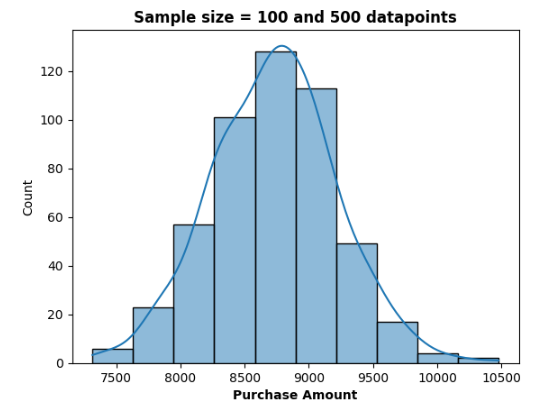
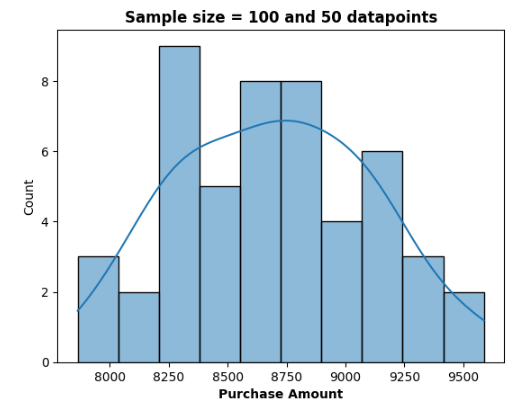
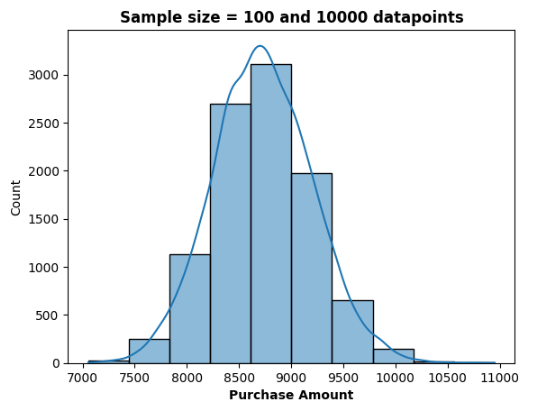
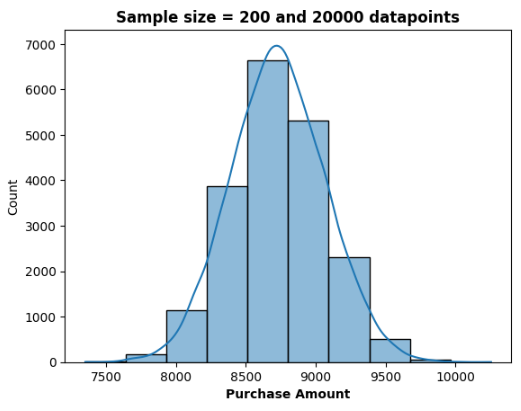
**Visualization of the Distribution of mean spending done by male customers**



**Observations –**

Variance (spread of data) decreases as sample size and datapoints increase.

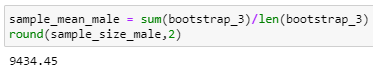
**Visualization of the Distribution of mean spending done by female customers**

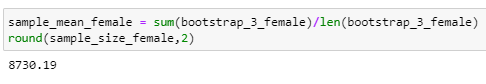
**Observations –**

Comparison between both the distributions infer that the mean spending of Male customers is more compared to the female customers.

**As the variance is less for sample size 200 and 20,000 datapoints, we calculate the confidence intervals for 90%, 95% and 99% confidence levels for the specified sample size and sample mean.**

Sample\_mean for sample size = 200:



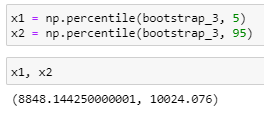


From the above sample means, we can say that the average amount spent by male customers is 9434.45 and the average amount spent by female customers is 8730.19.

**Validating the difference in the mean spending with different confidence interval–**

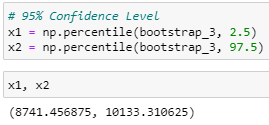
**For Male customers**

1. For Male customer with 90% Confidence Level



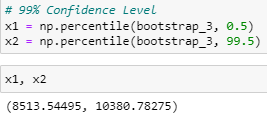
With 90% Confidence Interval, the mean spending of male customers lie in the range (8848.144, 10024.076).

1. For Male customer with 95% Confidence Level



With 95% Confidence Interval, the mean spending of male customers lie in the range (8741.456, 10133.311).

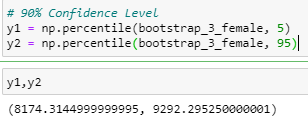
1. For Male customer with 99% Confidence Level



With 99% Confidence Interval, the mean spending of male customers lie in the range (8513.545, 10380.782).

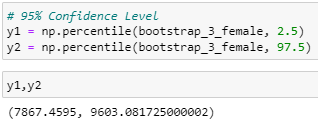
**For Female customers**

1. For Female customer with 90% Confidence Level



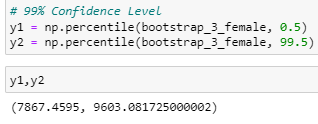
With 90% Confidence Interval, the mean spending of female customers lie in the range (8174.314, 9292.295).

1. For Female customer with 95% Confidence Level



With 95% Confidence Interval, the mean spending of female customers lie in the range (7867.459, 9603.082).

1. For Female customer with 99% Confidence Level



With 99% Confidence Interval, the mean spending of female customers lie in the range (7867.4595, 9603.917).

|  |  |  |  |
| --- | --- | --- | --- |
|  | **90% CI** | **95% CI** | **99% CI** |
| **MALE** | (8848.144, 10024.076) | (8741.456, 10133.311) | (8513.545, 10380.782) |
| **FEMALE** | (8174.314, 9292.295) | (7867.459, 9603.082) | (7867.4595, 9603.917) |

**Observations –**

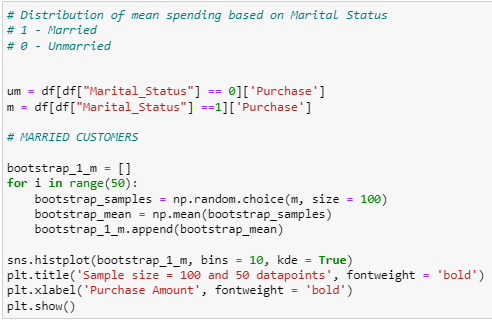
Confidence intervals are found to be overlapping. This concludes that there is no significant difference in the mean spending by male and female customers.

1. **Marital Status**

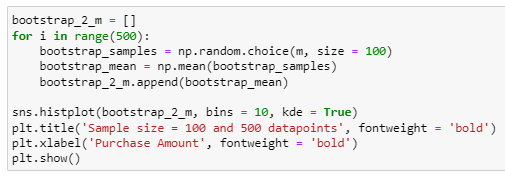
**Distribution of mean spending based on Marital Status –**

**For Married customers**

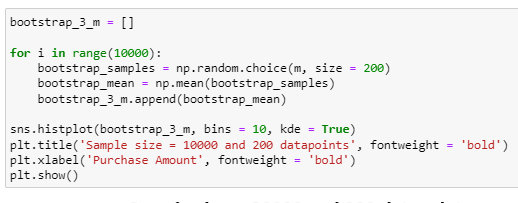
1. Taking sample size = 100 and checking for 50 datapoints.



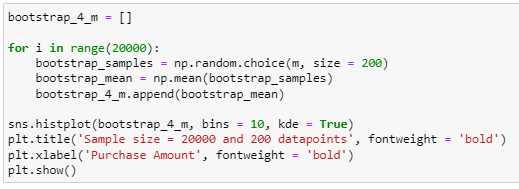
1. Taking sample size = 100 and checking for 500 datapoints.



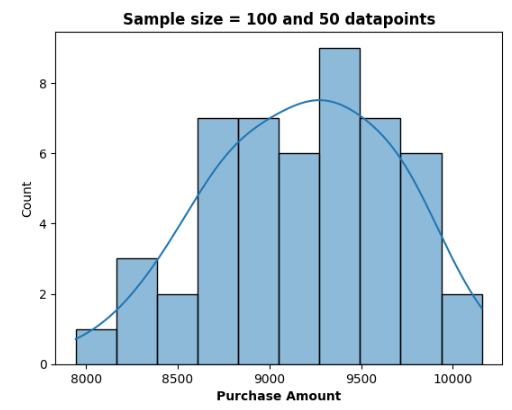
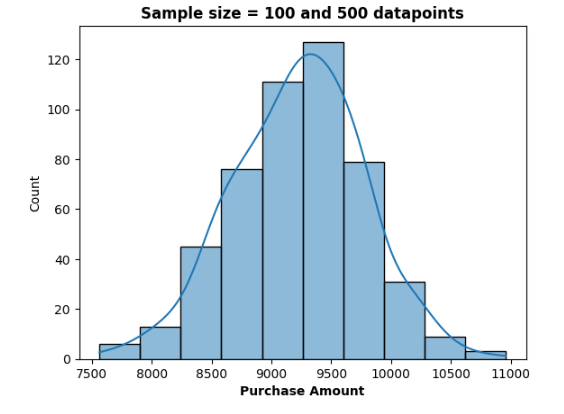
1. Taking sample size = 200 and checking for 10000 datapoints

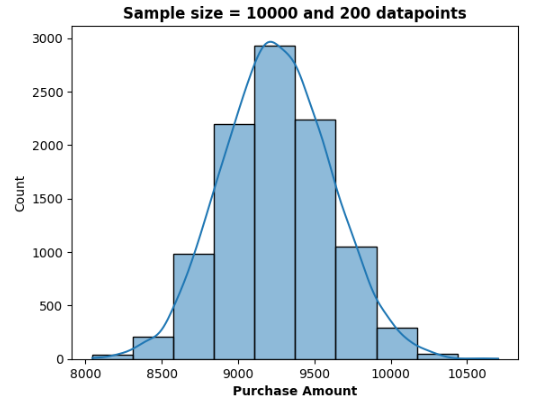
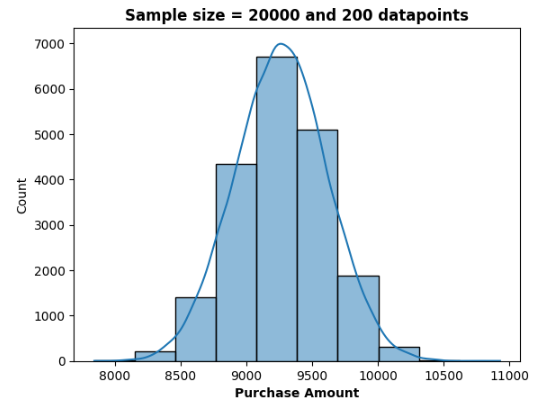


1. Taking sample size = 200 and checking for 10000 datapoints

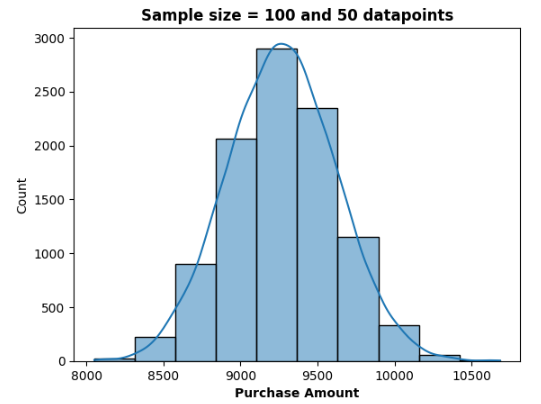
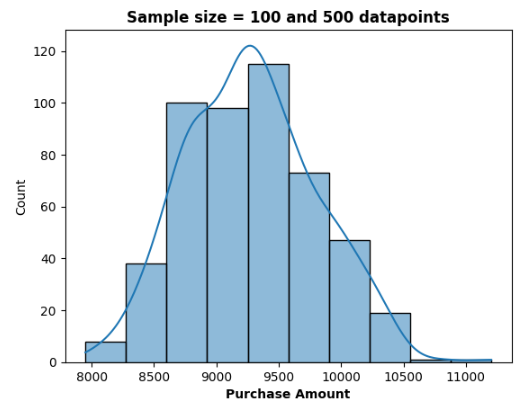
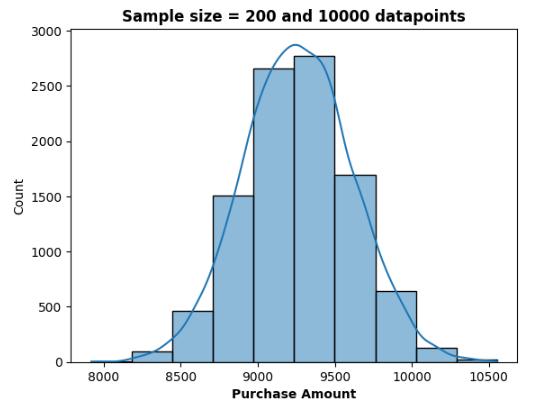
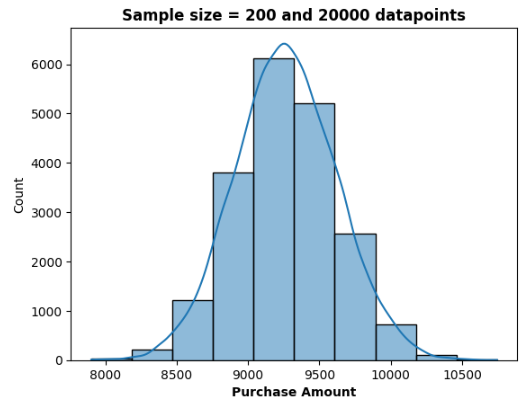


**Visualization of the Distribution of mean spending done by Married customers**

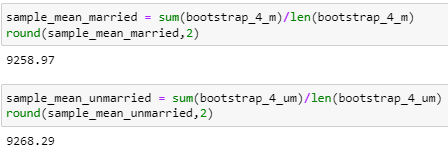




**Visualization of the Distribution of mean spending done by Unmarried customers**



**Validating the difference in the mean spending based on Marital Status –**

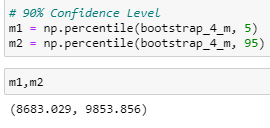
****

**Observations –**

The average spending of the samples taken for both married as well as unmarried customers are not significantly different.

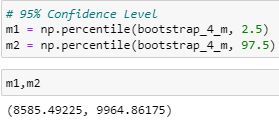
**For Married customers**

1. For Married customers with 90% Confidence Level



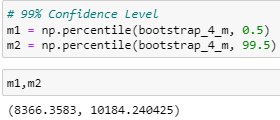
With 90% Confidence Interval, the mean spending of married customers lie in the range (8683.029, 9853.856).

1. For Married customers with 95% Confidence Level



With 95% Confidence Interval, the mean spending of married customers lie in the range (8585.4922, 9964.8617).

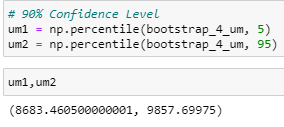
1. For Married customers with 99% Confidence Level



With 99% Confidence Interval, the mean spending of married customers lie in the range (8366.3583, 10184.2404).

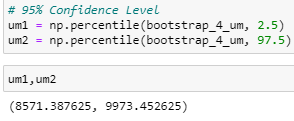
**For Unmarried customers**

1. For unmarried customers with 90% Confidence Level



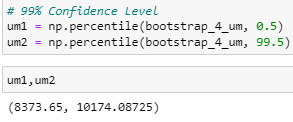
With 90% Confidence Interval, the mean spending of unmarried customers lie in the range (8683.4605, 9857.699).

1. For unmarried customers with 95% Confidence Level



With 95% Confidence Interval, the mean spending of unmarried customers lie in the range (8571.3876, 9973.4526).

1. For unmarried customers with 99% Confidence Level



With 99% Confidence Interval, the mean spending of unmarried customers lie in the range (8373.65, 10174.087).

|  |  |  |  |
| --- | --- | --- | --- |
|  | **90% CI** | **95% CI** | **99% CI** |
| **MARRIED** | (8683.029, 9853.856) | (8585.4922, 9964.8617) | (8366.3583, 10184.2404) |
| **UNMARRIED** | (8683.4605, 9857.699) | (8571.3876, 9973.4526) | (8373.65, 10174.087) |

**Observations –**

Confidence intervals are found to be overlapping. This concludes that there is no significant difference in the mean spending by married and unmarried customers.

**Recommendations**

1. Men spent more money than women, so company should focus on retaining the male customers and getting more male customers.
2. Product\_Category - 1, 5 & 8 have highest purchasing frequency. It means these are the products in these categories are liked more by customers. Company can focus on selling more of these products or selling more of the products which are purchased less.
3. Customers in the age 18-45 spend more money than the others, so company should focus on acquisition of customers who are in the age 18-45
4. Male customers living in City\_Category C spend more money than other male customers living in B or C, Selling more products in the City\_Category C will help the company increase the revenue.

**# Checking the variation in the mean/ average of customers using Hypothesis Testing.**

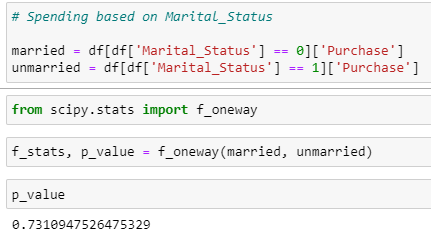
Using ANOVA, we can find the variation in the mean spending for 5% significance.

1. **Spending based on Marital Status**

H0: All means are similar

Ha: Means are different.

Alpha = 0.05 (95% Confidence Level)

****

Since, p\_value > alpha, We fail to reject the Null Hypothesis (H0).

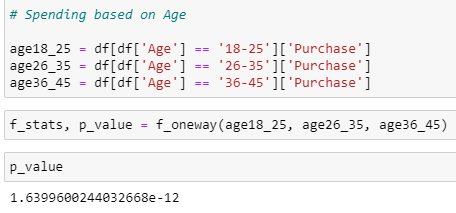
Hence, we can conclude that with a 95% Confidence Level, there is no difference in the average spending of customers based on their Marital Status.

1. **Spending based on Age**

H0: All means are similar

Ha: Means are different.

Alpha = 0.05 (95% Confidence Level)



Since, p\_value < alpha, We reject the Null Hypothesis (H0).

Hence, we can conclude that with a 95% Confidence Level, there is a significant difference in the average spending of customers based in the age-groups 18-25, 26-35, 36-45.

**Recommendations –**

Walmart must focus to retain the customers in these age-groups by providing discount offers and special contests to attract more customers belonging to these age-groups to increase their revenue.