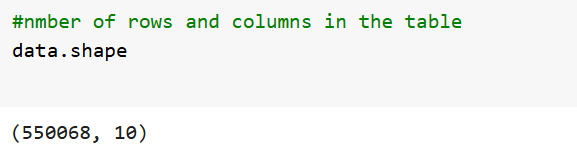
**Business Case: Walmart - Confidence Interval and CLT**

**Problem Statement**

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?

Q) Observation on the shape of data?

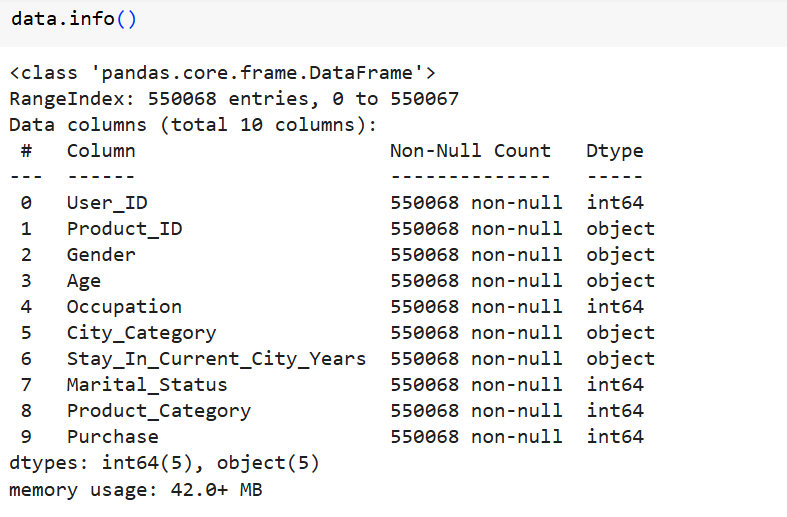
Ans)



Insights : There are 550068 rows for the data set regarding the purchase history on black Friday. There are 10 different columns or attributes that are recorded for these individual records / rows.

Q) What are the datatypes for the columns in the table?

Ans)



Insights : There are 5 columns which have integer behaviour ( UserID, Occupation, Marital\_Status, Product\_Category and Purchase) and rest of the 4 columns are object in nature of the total 9 columns recorded for the various products bought by various customers. The column Age is an object data type since the age is not represented as age of the individual , but, it represents the age group of the individual. Marital\_Status column has the values 0 and 1 which represents False and True for the Married Scenario. 0 means person is unmarried and 1 represents married person. Product\_Category column represents the values from 1 to 20 which represents a variety of categories of the products (masking the original values) that were sold during the black Friday at Walmart Stores. Similarly, column Occupation also has integer value as its masked value. This is internally linked to various actual occupation values.

Q) What are the missing values in the dataset ?

A screenshot of a computer

Description automatically generatedAns)

Insights : None of the columns has any missing values.All the columns has desired values as per their datatype.

Q) Describe the outliers for continuous variables?

Ans)

A screenshot of a graph

Description automatically generated

Insights : as per the box plot representation, we can see that the outliers for the values of purchase amount is in between 20,000 and 25,000 for all the age groups that has purchased during the black Friday. Most of the purchases has been made for an amount between 5000 and a little over 10,000.

A blue rectangular object with white lines

Description automatically generated

Insights : As per this box plot, the majority of the age group who has purchased the items during black Friday falls under the age group 26-55 years old. The other groups have very minimal complaints. There is a data point above the upper whisker, indicating an age value that is significantly higher than the rest of the data. There is a data point below the lower whisker, indicating an age value that is significantly lower than the rest of the data.

Recommendations : The Walmart stores can increase the number of items that falls between the price range of 5000-13000 which is the most preferred purchase amount for the customers. Also any good offers that could attract the people in the age group 26-55 will also be beneficial for the store. As mostly these age groups people would be having an individual life / family life, items that would help them improve their day to day life or items that would ease their efforts for daily chores would be highly appreciated and welcomed by them.

Q) Describe the number of years the tenants are residing in the current location ?

Ans)

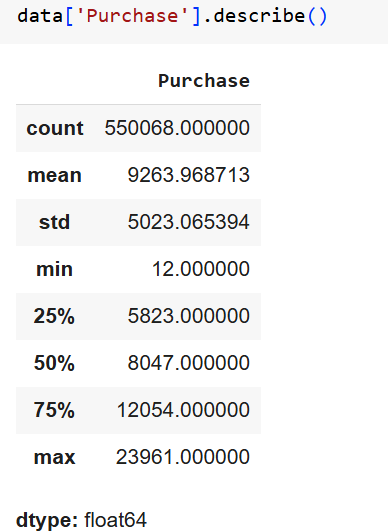
A screenshot of a computer

Description automatically generated

Insights :There are 5 different values for the number of years customers has been living in the current city. It should be noted that the customers that are high in count are those who has been living in the city for an year. The other values for the same are 0(just / recently arrived ), 2, 3 and more than 4. There are about 193821 customers who has the top value of the total count of customers.   
Recommendations : This would mean that people who have just got settled in the city are the one who are mostly the customers during the Black Friday. This would suggest that they would like to make their stay in the city more convenient and comfortable. Items that would aid in these scenarios could be given more promotion and offers so that they would be more attracted to visit the store and buy them all.

Q) Describe the purchase amount made by the customers ?

Ans)

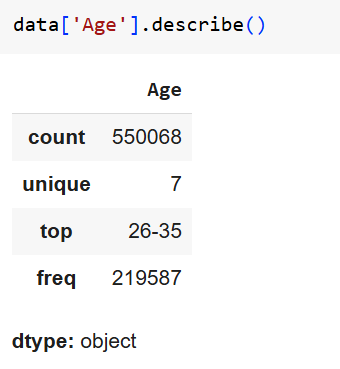


*Insights* : The average purchase amount by the customers is 9263.97/- Looking at the data, the plot says that 50% of the customers buy things that total upto 8047 while 75% of them buys at 12054 and 25% of them buys at 5823.

*Recommendations* : more focus could be given to the purchase amount of less than 8000 so that the lower band of customers could also be added to get to the 50% that contributes to a substantial amount of income.

Q) Describe the age variable of the customers?

Ans)

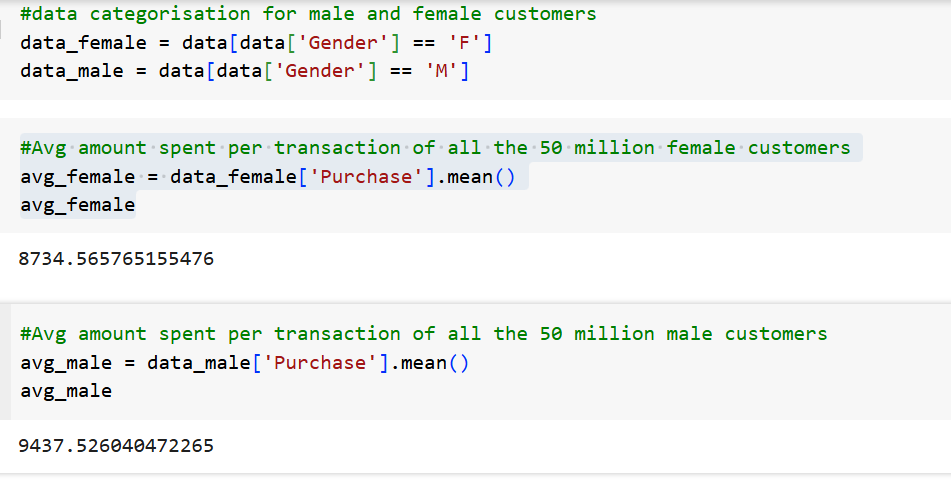


Insights : This info shows that there are 7 different values for age group – namely – 0-17,18-25,26-35,36-45,46-50,51-55, 56+. Of these seven different groups of age, the most occurring or repeated value among them is 26-35. This shows that most of the customers for the black Friday falls under the age limit of 26-35 years old with a total frequency of 219587.

Recommendations : It can inferred that most of the users might be having a family as their age is within the group of 26-35 years old. But this is not a concrete categorization since we don’t have the accurate data regarding the same. However based on the general notion and nature of the society it can be inferred so. This will help us to place products that more family friendly under major promotions and offers so that these group of customers will be more attracted.

Q) What is the average amount spent by the female and male customers ?

Ans)



*Insights* : The average amount spent by the female customers is 8734.56 and that for male customers is 9437.52. This shows that male customers do more purchases during the black Friday than the female customers. Another reason could be Male customers might have bought less number of items, but their price would be sufficiently large enough to get the average value at 9437.

*Recommendations* : More items or products that are appealing to male customers could be given more discounts / offers / combo offers so that they have the incentive to buy the products. There is hardly a difference for 1000 between the mean values of female and male customers.So Walmart could introduce high value products at a discounted rates along with combo offers or otherwise so that more of female customers are moved to make the purchases.

Q) Are there preferred categories for different genders?

Ans)

A screenshot of a graph

Description automatically generated

*Insights* : Product category 10 seems to have more purchase among both the genders. For both the genders it comes around 20,000/-This makes it the more preferred category among both genders. The least preferred category for both genders would be 13 and 20. They have a purchase valuation of less than 2500. The next preferred category would be 6,7,9 which are having comparable purchase amounts. 13,14,15 category are also among the top values when compared to the category 9.

*Recommendations :* Since the product categories 6,7,9,13,14 and 15 have comparable purchases, combos or offers could be given to these so that the sales in these categories can be improved. This would help increase the revenue since these categories are already in the range of 15000-17000. As these categories show similar trend in both the genders, these might be unisex items. Hence by focusing on these categories would help in bringing both the genders to the store as no separate promotions has to be made.

Along with these we need to focus on category 10 which is already in top position. Store can also look into combo offers with the category 10 so that any other category which has lower number of purchases can also be improved.

Q) What is the relationship between age, marital status and purchase amount ?

Ans)



Insights : It should be noted that here 0 (blue dots) represents unmarried customers while 1 (orange dots) represents married customers. We can infer that people in the age group of 46-55 who are married have a more purchase history during the black Friday. While It is mixed for the age group of 26-35 and 36-45. It should be noted that in the band 18-25 most of them are unmarried while there are notable married customers too especially in the lower range of purchase amounts. It should be noted that the 0-17 band has only unmarried customers. The 55+ age group represents a mix of sufficiently large enough unmarried and married customers, with married customers more towards the higher amount of purchase

Recommendations : We can do targeted promotions with the help of above data. The 0-17 age group can be focused with all range of products while the 46-55 age group should be focussed more on the family side of products. Things like what would be more into making of a better home to live in including furnishing section. While for the groups with mixed group of customers store can promote both the items. Costlier ones would be accessible for the married group due to additional income of partner too. While the less costlier items can be made affordable to the unmarried section. We could focus the principle of quantity over individual product prioritsation in this section. While for the married customers and aged customer store could trim the promotions to prioritize on the quality of the individual products. E.g. comfort of a reclining chair. If the comfort is good enough, customers would be ready to trade it for a higher price unlike the unmarried lower age group who focusses more on affordability than finesse of the product.

Q) How does gender affect the amount spent ?

Ans)

A screenshot of a graph

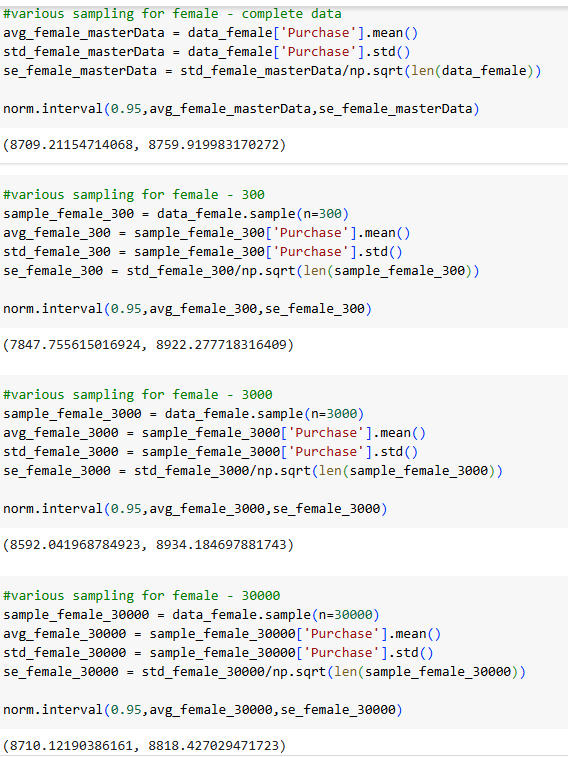
Description automatically generated

*Insights* : There is a small change in the purchase amount spent with respect to the different genders. For Female its slightly over the amount of 8000 and for the male its slightly more than the value of female. From this data we can infer that more or less there is similar amount being spent by different gender

*Recommendations :*  Store can give out offers and promotional discounts so that due to similar shopping pattern, store can get in touch with both female and males customers through the same advertisements. More impact with lesser resource usage.

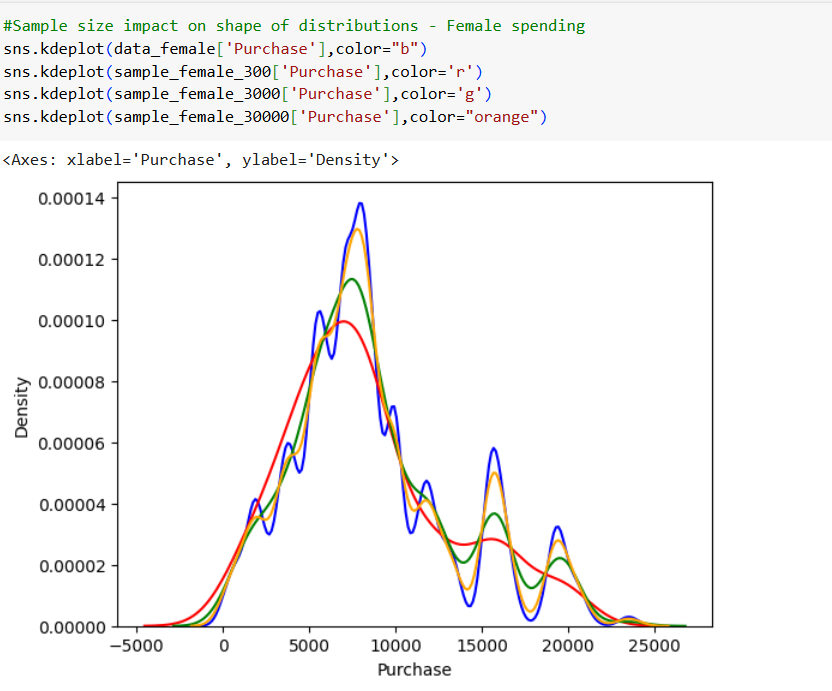
**Q) Confidence intervals and distribution of the mean of the expenses by female customers**

Ans)



Insights : There are various sample sizes being checked for the values of the whole data set. The Interval for the whole data set for female customers are : [8709.21,8759.91]. The sample size has been varied from 300 to 3000 and finally to 30000. Upon inspection, we can see that for these sample sizes ( in order) the lower range that starts at 7847.755 and increasing to 8592.41 and to 8710.121 – in order . The upper range starts at 8922.277 and increasing to 8934.184 and to 8818.4270 – in order. From this it is evident that as the sample size increases the values of the confidence interval is getting more and more closer to the value with the master data for female customers.

Q) **what is the impact of the various sample size on the shape of distribution mean of female spending.**

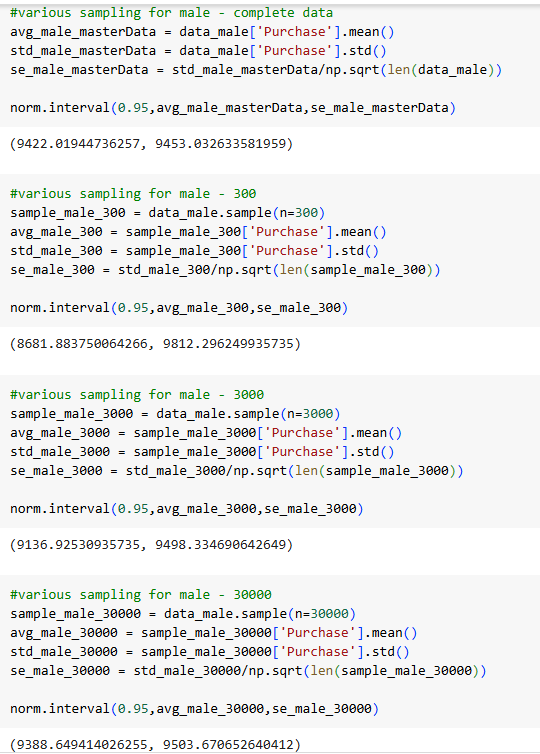


Insights : The blue lines refers to the complete data related to female customers. The red line refers to the sample size for 300, green for sample size 3000 and orange for sample size 30000. The base of the distribution is larger for the smaller sample size and as it increases the distribution becomes lean and closer to the complete data set for female customers.

This also helps understand that the majority of the female customers has the purchase amount interval between 5000 and little above 10000.

**Q) Confidence intervals and distribution of the mean of the expenses male customers**

Ans)



Insights : There are various sample sizes being checked for the values of the whole data set. The Interval for the whole data set for male customers are : [9422.019,9453.032]. The sample size has been varied from 300 to 3000 and finally to 30000. Upon inspection, we can see that for these sample sizes ( in order) the lower range that starts at 8681.883 and increasing to 9136.925 and to 9388.649 – in order . The upper range starts at 9812.296 and increasing to 9498.334 and to 9503.670 – in order. From this it is evident that as the sample size increases the values of the confidence interval is getting more and more closer to the value with the master data for male customers.

Q) **what is the impact of the various sample size on the shape of distribution mean of male spending?**

Ans)

A graph of a line graph

Description automatically generated with medium confidence

Insights : The blue lines refers to the complete data related to male customers. The red line refers to the sample size for 300, green for sample size 3000 and orange for sample size 30000. The base of the distribution is larger for the smaller sample size and as it increases the distribution becomes lean and closer to the complete data set for male customers.

This also helps understand that the majority of the male customers has the purchase amount interval between 5000 and little above 10000.

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

Ans)

A graph of a person and person

Description automatically generated with medium confidence

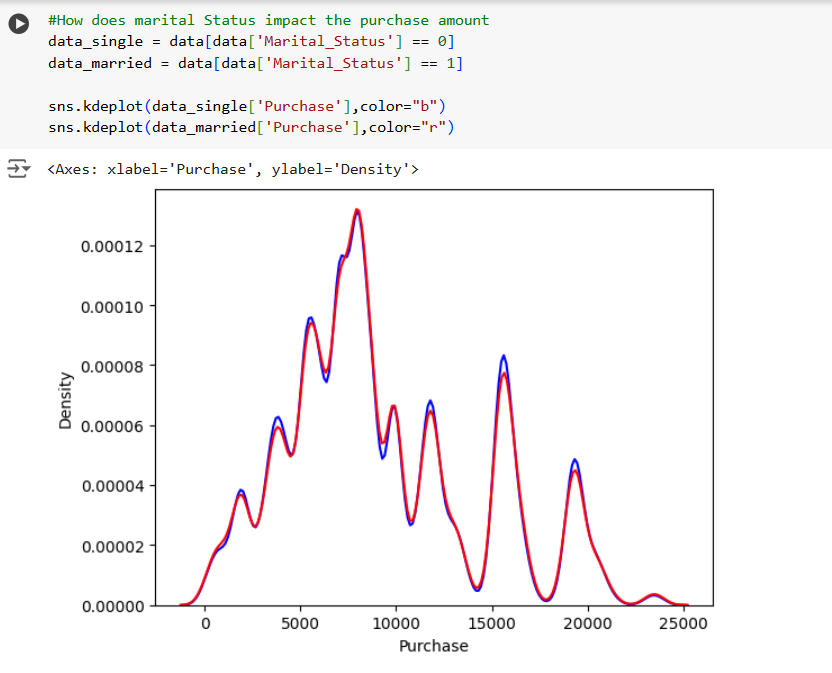
Insights : for the lower valuation of the purchase amounts, the confidence intervals are overlapping at many places if not being congruent. while for the higher amounts beyond 10000, the intervals are not as overlapping as the former one. High differences can be noticed for the ranges of male and female customer spending.

Recommendations : Since the purchase spending for the lower amounts are overlapping more or less for both the genders, the promotional offers and discounts that can be made for those items will definitely cater to both the genders. This would be an efficient use of resources. For the higher amount this the overlapping is not as much for the initial part, it does travel through the same path. i.e. the higher limit for the density would be different while the purchase range remains almost similar for both the genders. Here too higher valuation items can be stocked more for the black Friday. Though the gender that buys the product might differ ultimately they get bought by the customers.

Better offers for the amounts less than 8000 would also be a good move for the store.

Q) Impact of Martial status on purchase

Ans )

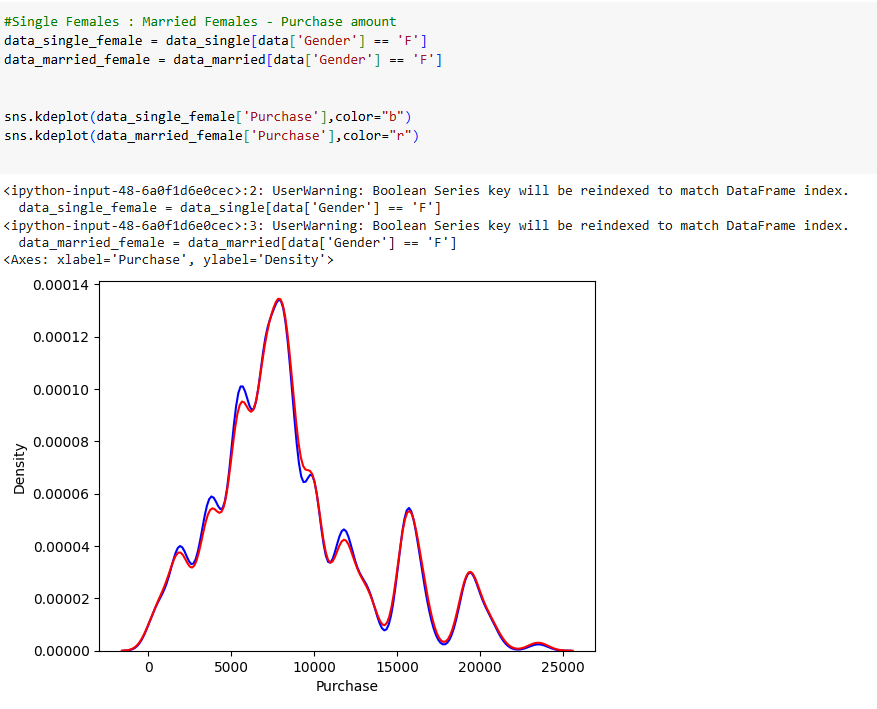


Insights : We are looking at the distribution mean for the unmarried and married customers. This shows that despite the marital status of the customers, the purchase amount is not affected to a great degree. IF we look at the higher density area, both married and unmarried customers range to same values. And for the lower values of purchase also it is having a similar trend. The higher valuation of purchase from 10000 onwards shows similar trend with a very few deviation for the married to unmarried customers.

Recommendations: The above distribution shows that we need not change anything per se at large for the customers based on their marital status. Irrespective of the status, the black Friday sales are going good. However, if we could create some promotional offers for the couples, that would definitely be appreciated by the customers. Customers when treated special are always appreciated and they would definitely increase the chances for being as return customers as well as loyal customers.

Q) Impact of Marital Status of Male and female separately ?

Ans)



Insights : Looking at the distribution mean, among females, for the purchase amounts more than 10000 we have a congruent line for both married and unmarried female customers. For the amounts less than 10000, though there is a slight variation for some price points, most of the part they are overlapping.

A screen shot of a graph

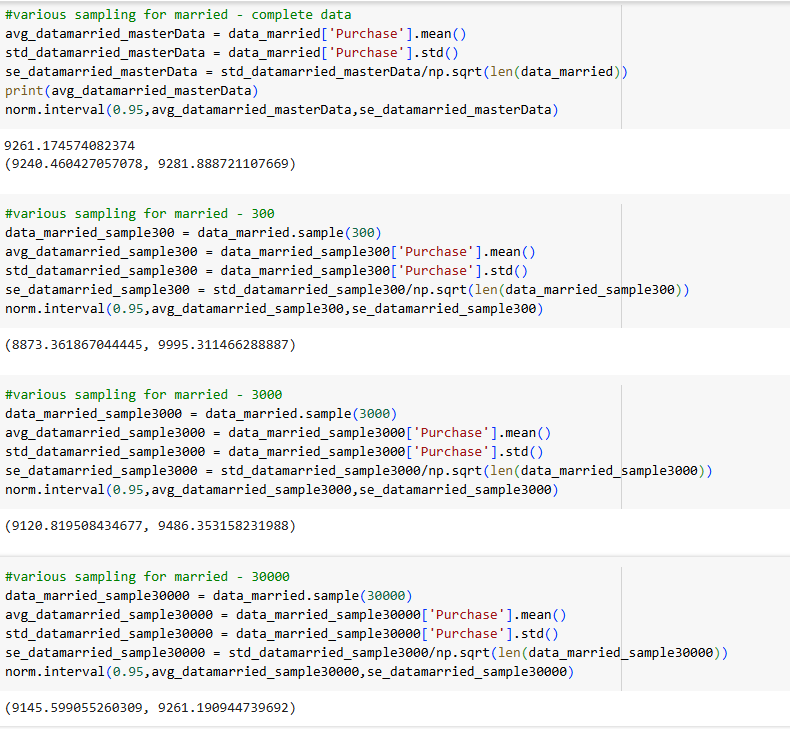
Description automatically generated

Insights : Looking at the distribution mean, among males, for the purchase amounts slightly less than mean and the peaks more than the mean values are slightly off the overlap scenario. However we can deduce that more or less same purchase amount as stated for unmarried and married customers.

Q) What is the impact of various sampling size on unmarried and married customers ?

Ans )





Insights : On changing the sample size from 300 to 3000 to 30000, the CI intervals are changing and they are coming near to the values of the total population of unmarried / married customers.

Q) is there any impact on product categories where purchase is made based on the age group?

Ans)

A screenshot of a graph

Description automatically generated

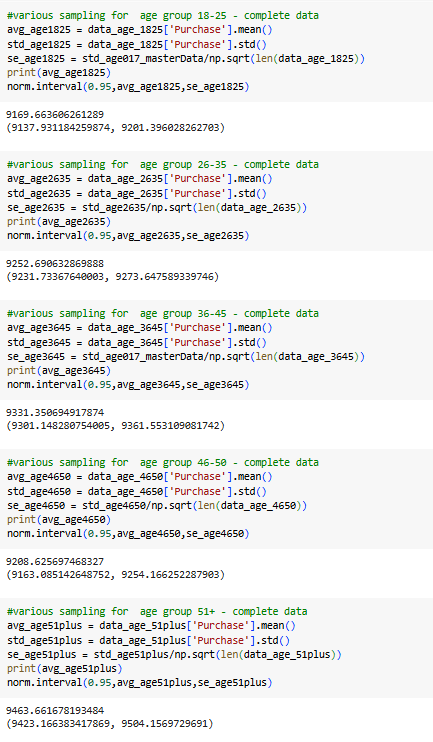
Insights : for almost all the age groups, product category #10 is the most purchased items. This is followed by categories 7,6. This is in similar pattern with that of the product category selection based upon gender. The purchase amount of these categories – cumulative for each age groups is around 20,000, around17,000 and a bit over 15000 in order. So the volume of revenue is sufficiently large enough to make difference to the profitability of the store.

Recommendations : These categories can be given promotional offers which would again increases the number and thus volumetric increase in sales and positive impact on revenue generation. It should also be noted that coupling of products of these category with other low purchase items can sufficiently improve the number of the sluggish category. Even if we are trying to give sluggish items at nominal price or free, it can be offset by the increase in revenue caused by the volumetric increase in the high sales product.

Q) How does the CI interval changes for different age groups ?

Ans) A screenshot of a computer program

Description automatically generated

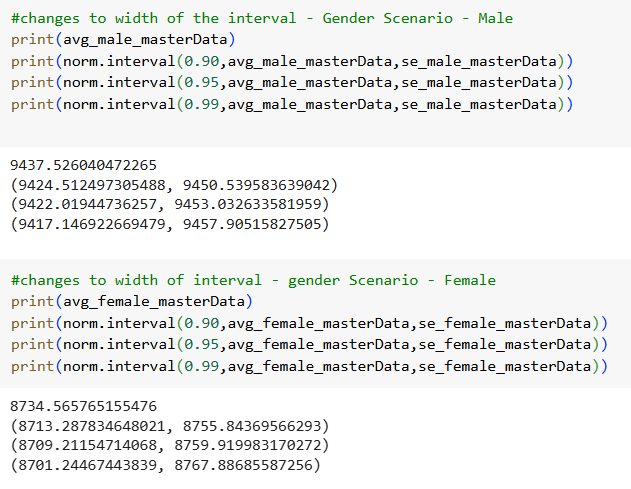


Insights : for the age group 0 – 17 age group, the CI is marked at (8851.947970542686, 9014.981310347262). For the age group 18-25, the CI is marked at (9137.931184259874, 9201.396028262703). For the age group 26-35, the CI is marked at (9231.73367640003, 9273.647589339746). For the age group 36-45, the CI is marked at (9301.148280754005, 9361.553109081742). For the age group 46-50, the CI is marked at (9163.085142648752, 9254.166252287903). For the age group 51+, the CI is marked at (9423.166383417869, 9504.1569729691)

If we look at the figures for the mean values – 8933.464, 9169.663, 9252.690,9331.350, 9208.625, 9463.661- we can notice that there is an increase in value of the mean which steadily goes towards the actual mean values.

Q) what happens when the confidence is changed to 90%, 95% and 99% for gender based spending scenario ?

Ans)



Insights : it can be noted that as the Confidence increases from 90% to 99% the base band of the interval is also increasing – in case of males the differences for three confidence are 26,31 and 40 respectively. While for females the difference in range of CI would be 42, 50 and 66 respectively. The interval change for 90 to 95 is comparable while the change from 95 to 99 is almost double of the earlier interval range. IT should be noted that the confidence shows us the probability of getting the mean in the given interval.

Q) what happens when the confidence is changed to 90%, 95% and 99% for marital status based spending scenario ?

Ans) Insights : it can be noted that as the Confidence increases from 90% to 99% the base band of the interval is also increasing – in case of married customers, the differences for three confidences are 35,41 and 55 respectively. While for unmarried customers, the difference in range of CI would be 29, 35 and 45 respectively. The interval change for 90 to 95 is comparable while the change from 95 to 99 is almost double of the earlier interval range. IT should be noted that the confidence shows us the probability of getting the mean in the given interval.

Q) Are women spending more money per transaction than men? Why or Why not?

Ans)

A screenshot of a computer code

Description automatically generated

Looking at the mean value of the transactions made by the male and female customers, it is evident that men are spending more money per transaction than women.

As mentioned earlier one reason could be because black Friday gives lot of offers and discounts to the customers. So they might be making their year long purchases during the black Friday.It could also be that men are more into electronic items, which have a higher price range when compared to other items bought by women in general. So high price range of items bought could also be another reason to have a larger average shown for men.

*Recommendations* : More items or products that are appealing to male customers could be given more discounts / offers / combo offers so that they have the incentive to buy the products. There is hardly a difference for 1000 between the mean values of female and male customers.So Walmart could introduce high value products at a discounted rates along with combo offers or otherwise so that more of female customers are moved to make the purchases.

**Final Insights :**

Based on the exploration and the CLT values discovered, it can be stated that on a black Friday both men and women spend almost equally on the purchase amount. So is the case with the marital status too. It should also be noted that certain categories of product like #10, #6 #7 are most favored products across the age groups as well as across the genders.

It can also be noted that the CI for higher confidence takes up almost double the range when compared to +/-10 points of Confidence.

**General recommendations :** In total Walmart can give promotional offers on a targeted method for the aged customers who would prefer quality over affordability. While the younger group could be made attracted to the affordability feature. Gender based promotions are not mandatory as most of the purchase trends shows that both men and women purchase on similar bands of cost. Hence general promotion would help in saving money but at the same time cater to all the target audience. Along with that more stocks could be increased for the product categories like 10,6,7, which are purchased across the age groups and gender.