Name: Raghu
Business Case: Walmart – Confidence Interval and
Date: 05/07/2023

Business Problem

The Management team at Walmart Inc. wants to analyze the customer purchase behaviour (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) & same for material status & different age groups.

```
[29] import pandas as pd
     import numpy as np
     import seaborn as sns
     import matplotlib.pyplot as plt
```

```
[30] df = pd.read_csv("https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/293/original/walmart_data.csv?1641285094")
```

```
df.head()
```

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category	Purchase
0	1000001	P00069042	F	0-17	10	A	2	0	3	8370
1	1000001	P00248942	F	0-17	10	A	2	0	1	15200
2	1000001	P00087842	F	0-17	10	A	2	0	12	1422
3	1000001	P00085442	F	0-17	10	A	2	0	12	1057
4	1000002	P00285442	M	55+	16	C	4+	0	8	7969

```
[34] df.shape

(550068, 10)
```

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 550068 entries, 0 to 550067
Data columns (total 10 columns):
 #   Column                      Non-Null Count   Dtype 
---  ------                      --------------   ----- 
 0   User_ID                     550068 non-null  int64 
 1   Product_ID                  550068 non-null  object
 2   Gender                      550068 non-null  object
 3   Age                         550068 non-null  object
 4   Occupation                  550068 non-null  int64 
 5   City_Category               550068 non-null  object
 6   Stay_In_Current_City_Years  550068 non-null  object
 7   Marital_Status              550068 non-null  int64 
 8   Product_Category            550068 non-null  int64 
 9   Purchase                    550068 non-null  int64 
dtypes: int64(5), object(5)
memory usage: 42.0+ MB
```

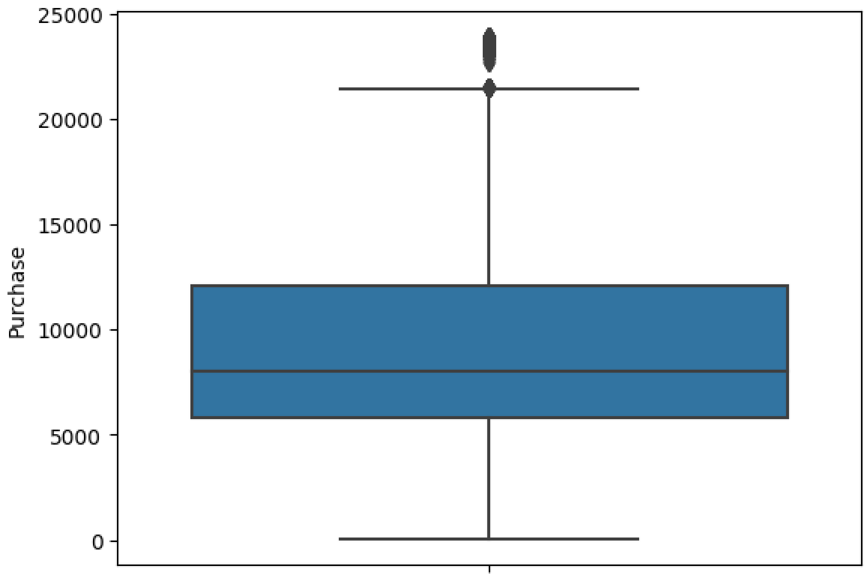
```
[36] # Count of Null values in each column

     df.isnull().sum()
```

```
User_ID                       0
Product_ID                    0
Gender                        0
Age                           0
Occupation                    0
City_Category                 0
Stay_In_Current_City_Years    0
Marital_Status                0
Product_Category              0
Purchase                      0
dtype: int64
```

Name: Raghu
Business Case: Walmart – Confidence Interval and

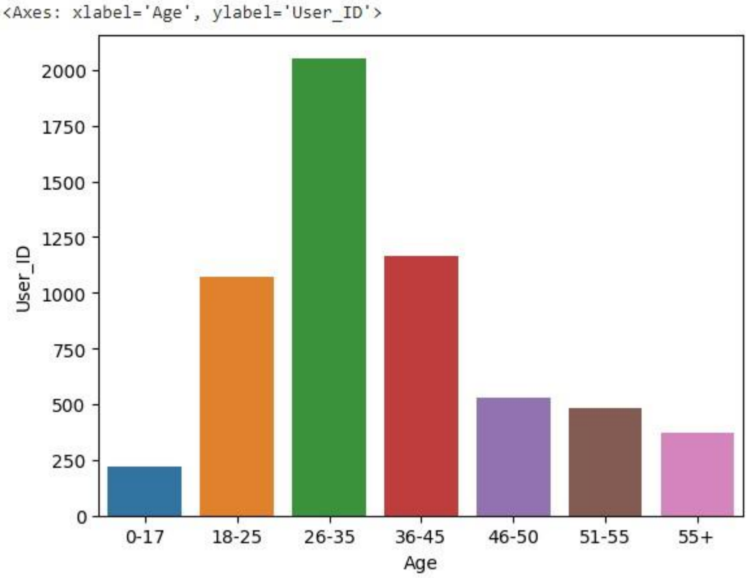
```
[39] # checking outliers in purchase amount

     sns.boxplot(data = df, y = "Purchase")
```

```
<Axes: ylabel='Purchase'>
```




```
[56] # Count of unique users in each age group

     age_df = pd.DataFrame(df.groupby("Age")["User_ID"].nunique()).reset_index()
     sns.barplot(data = age_df, x = 'Age', y = 'User_ID')
```

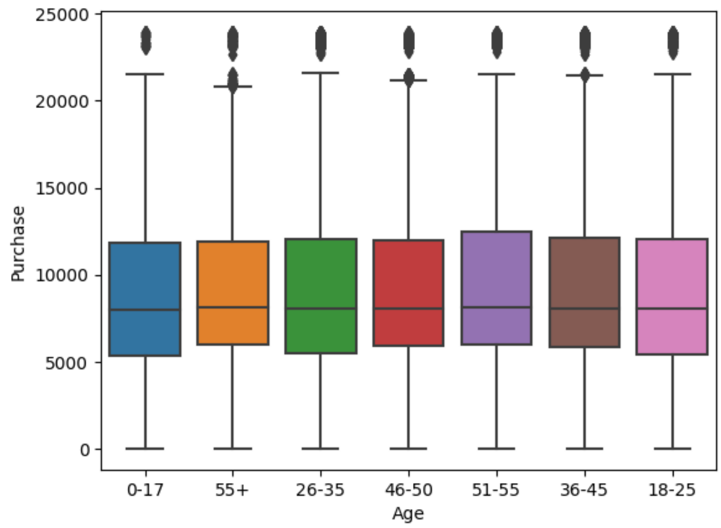
```
<Axes: xlabel='Age', ylabel='User_ID'>
```



People of age group btween 18 & 45 are willing to purchase more. So targeting people of this age group can enhance the number of sales especially age group of 26-35.

```
[57] # Minimum, Maximum and mean of purchase amount/Spending habits of each age group

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

```
<Axes: xlabel='Age', ylabel='Purchase'>
```



Minimum, maximum & aveage amount spending across each age group is almost same.Hence spending habits/pattern of any particular age group does not have exterior impact on buisness' growth

Constructing confidence intervals for spending of Male & Female

```
[15] # Avearge spending of each male & Female users

     Male_avg_spending = df[df["Gender"] == 'M'].groupby("User_ID")["Purchase"].mean()
     Female_avg_spending = df[df["Gender"] == 'F'].groupby("User_ID")["Purchase"].mean()
```

```
[21] # Generating 10000 samples from from Male_avg_spending using bootstrap

     bootstrap_male_samples_mean = []
     for i in range (10000):
       bootstrap_male_samples = np.random.choice(Male_avg_spending, size = 150)
       bootstrap_male_mean = np.mean(  bootstrap_male_samples)
       bootstrap_male_samples_mean.append( bootstrap_male_mean)
```

```
[22] # Generating 10000 samples from from Female_avg_spending using bootstrap

     bootstrap_Female_samples_mean = []
     for i in range (10000):
       bootstrap_Female_samples = np.random.choice(Female_avg_spending, size = 150)
       bootstrap_Female_mean = np.mean(bootstrap_Female_samples)
       bootstrap_Female_samples_mean.append( bootstrap_Female_mean)
```

Name: Raghu
Business Case: Walmart – Confidence Interval and

Constructing 90% Confidence interval for Gender

```
[ ] # 90% Confidence interval for male
    x1 = np.percentile(bootstrap_male_samples_mean, 5)
    x2 = np.percentile(bootstrap_male_samples_mean, 95)
    confidence_interval_95_perc_male = [x1,x2]
    print(confidence_interval_95_perc_male)
```

```
[9704.836238251994, 9897.662672384728]
```

```
[ ] # 90% Confidence interval for Female

    x1 = np.percentile(bootstrap_Female_samples_mean, 5)
    x2 = np.percentile(bootstrap_Female_samples_mean, 95)
    confidence_interval_95_perc_Female = [x1,x2]

    print(confidence_interval_95_perc_Female)
```

```
[8876.44348434495, 9053.530573409798]
```

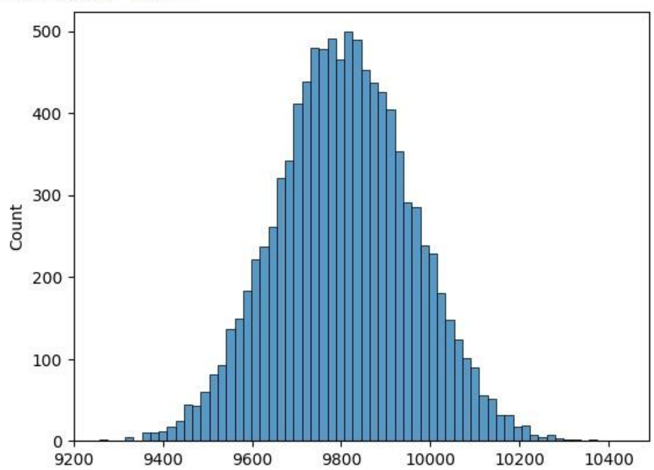
Above two results show where average spending of 50 million male and 50 million female customers may lie respectively with 90% confident

```
[23] #95% Confidence interval for Male customers

     x1 = np.percentile(bootstrap_male_samples_mean, 2.5)
     x2 = np.percentile(bootstrap_male_samples_mean, 97.5)
     confidence_interval_95_perc_male = [x1,x2]

     print(np.mean(bootstrap_male_samples_mean))
     print(confidence_interval_95_perc_male)
```

```
9807.171414801314
[9509.200069956547, 10105.55971532105]
```

```
[24] sns.histplot(bootstrap_male_samples_mean)
```

```
<Axes: ylabel='Count'>
```



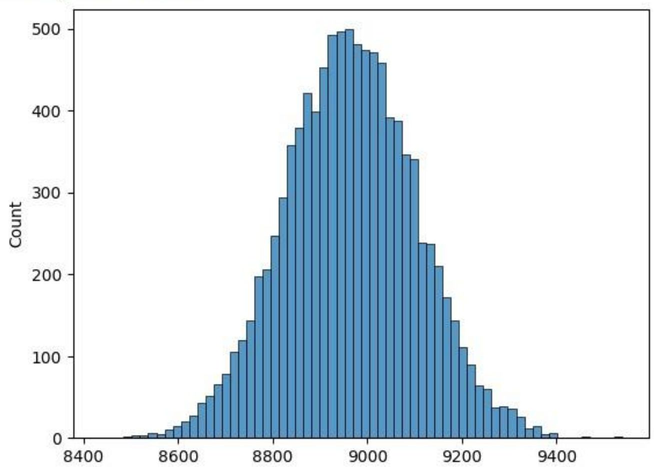
From the above confidence interval, it is concluded that average spending of Male of population lie in [9509.200069956547, 10105.55971532105] with 95% confident and average spending is 9807.17

```
[27] #95% Confidence interval for Female customers

x1 = np.percentile(bootstrap_Female_samples_mean, 2.5)
x2 = np.percentile(bootstrap_Female_samples_mean, 97.5)
confidence_interval_95_perc_Female = [x1,x2]

print(np.mean(bootstrap_Female_samples_mean))
print(confidence_interval_95_perc_Female)
```

```
8965.12662296556
[8691.32669226636, 9243.767080631997]
```

```
[28] sns.histplot(bootstrap_Female_samples_mean)
```

```
<Axes: ylabel='Count'>
```



From the above confidence interval, it is concluded that average spending of Male of population lie in [8691.32669226636, 9243.767080631997] with 95% confident and average spending is 8965.12

From the above, Since 95% Confidence intervals for spending of Male & Female are not overlapping, it is concluded that average spending of Male is greater than that of female. Company should come up with reason for the low average spending of female compared to male and new marketing strategy to increase the average spending of female

Name: Raghu
Business Case: Walmart – Confidence Interval and

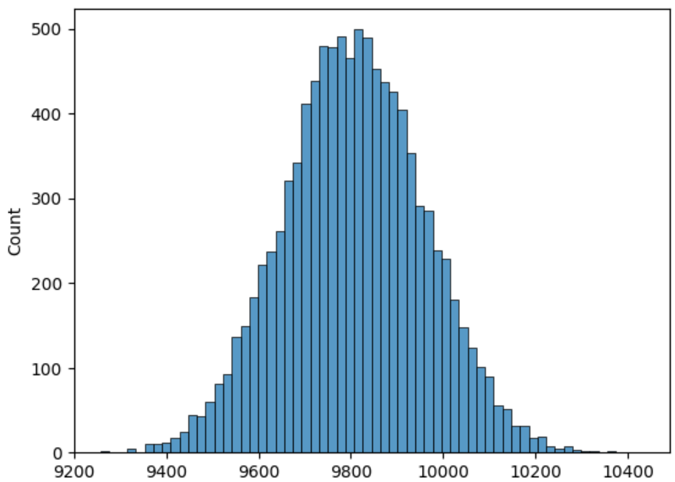
90% Confidence intervals

```
# 90% Confidence interval for male
x1 = np.percentile(bootstrap_male_samples_mean, 5)
x2 = np.percentile(bootstrap_male_samples_mean, 95)
confidence_interval_95_perc_male = [x1,x2]

print(f"mean of male population-->{np.mean(bootstrap_male_samples_mean)}")
print(confidence_interval_95_perc_male)
```

```
mean of male population-->9807.171414801314
[9555.77366640113, 10060.153198606573]
```

```
[34] sns.histplot(bootstrap_male_samples_mean)
```

```
<Axes: ylabel='Count'>
```



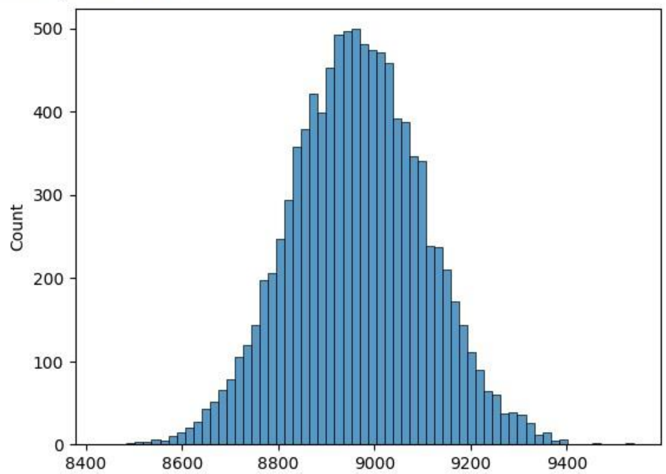
Name: Raghu
Business Case: Walmart – Confidence Interval and

```
# 90% Confidence interval for Female

x1 = np.percentile(bootstrap_Female_samples_mean, 5)
x2 = np.percentile(bootstrap_Female_samples_mean, 95)
confidence_interval_95_perc_Female = [x1,x2]

print(f"mean of male population-->{np.mean(bootstrap_Female_samples_mean)}")
print(confidence_interval_95_perc_Female)
```

```
mean of male population-->8965.12662296556
[8734.888411179081, 9194.574121684527]
```

```
[35] sns.histplot(bootstrap_Female_samples_mean)
```

```
<Axes: ylabel='Count'>
```



From the above two 90% confidential intervals for male & female, 90% confidential interval for male is [9555.77366640113, 10060.153198606573] and for female is [8734.888411179081, 9194.574121684527].

Constructing confidence interval for spending of Married and unmarried customers:

```
[ ] df["Marital_Status"].value_counts()
```

```
0    324731
1    225337
Name: Marital_Status, dtype: int64
```

```
[37] df_married = df[df['Marital_Status'] == 1].groupby("User_ID")["Purchase"].mean()
     df_unmarried = df[df['Marital_Status'] == 0].groupby("User_ID")["Purchase"].mean()
```

```
[38] # Genearing 10000 samples for married & unmarried customers using bootstrap

     bootstrap_married_samples_mean = []
     bootstrap_unmarried_samples_mean = []
     for i in range (10000):
       bootstrap_married_samples = np.random.choice(df_married, size = 150)
       bootstrap_married_mean = np.mean(bootstrap_married_samples)
       bootstrap_married_samples_mean.append( bootstrap_married_mean)

       bootstrap_unmarried_samples = np.random.choice(df_unmarried, size = 150)
       bootstrap_unmarried_mean = np.mean(bootstrap_unmarried_samples)
       bootstrap_unmarried_samples_mean.append( bootstrap_unmarried_mean)
```

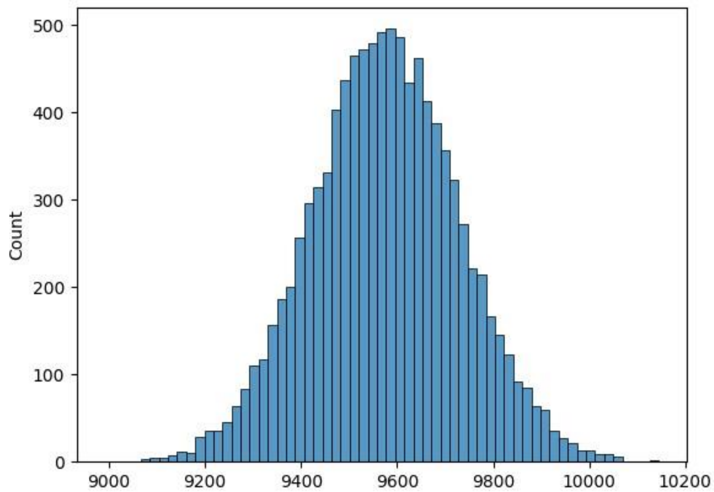
Name: Raghu
Business Case: Walmart – Confidence Interval and

```
[40] # 95% Confidence interval for married

x1 = np.percentile(bootstrap_married_samples_mean, 2.5)
x2 = np.percentile(bootstrap_married_samples_mean, 97.5)
confidence_interval_95_perc_married = [x1,x2]

print(f"Avearge spending of married--->{np.mean(bootstrap_married_samples_mean)}")
print(confidence_interval_95_perc_married)
```

```
Avearge spending of married--->9575.553733639006
[9274.548913013039, 9883.403051659669]
```

```
sns.histplot(bootstrap_married_samples_mean)
```

```
<Axes: ylabel='Count'>
```



From the above confidence interval, it is concluded that average spending of Mrried customers of population lie in [9274.548913013039, 9883.403051659669] with 95% confident and average spending is 9575.553733639006

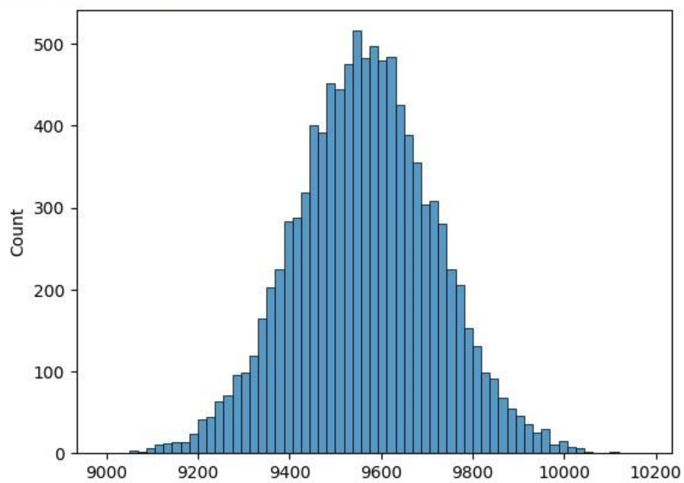
Name: Raghu
Business Case: Walmart – Confidence Interval and

```
# 95% Confidence interval for unmarried

x1 = np.percentile(bootstrap_unmarried_samples_mean, 2.5)
x2 = np.percentile(bootstrap_unmarried_samples_mean, 97.5)
confidence_interval_95_perc_unmarried = [x1,x2]

print(f"Avearge spending of married--->{np.mean(bootstrap_unmarried_samples_mean)}")
print(confidence_interval_95_perc_unmarried)
```

```
Avearge spending of married--->9564.813887277192
[9258.498494372996, 9870.272470760434]
```

```
[43] sns.histplot(bootstrap_unmarried_samples_mean)
```

```
<Axes: ylabel='Count'>
```



From the above confidence interval, it is concluded that average spending of Unmarried customers of population lie in [9258.498494372996, 9870.272470760434] with 95% confident and average spending is 9564.813887277192

From the above 95% confidence intervals for spending of married & unmarried customers, it is concluded that two intervals are overlapping, confidence intervals are almost same and mean of spending of married & unmarried customers are almost same, hence material status does not have any impact on purchase

Name: Raghu
Business Case: Walmart – Confidence Interval and

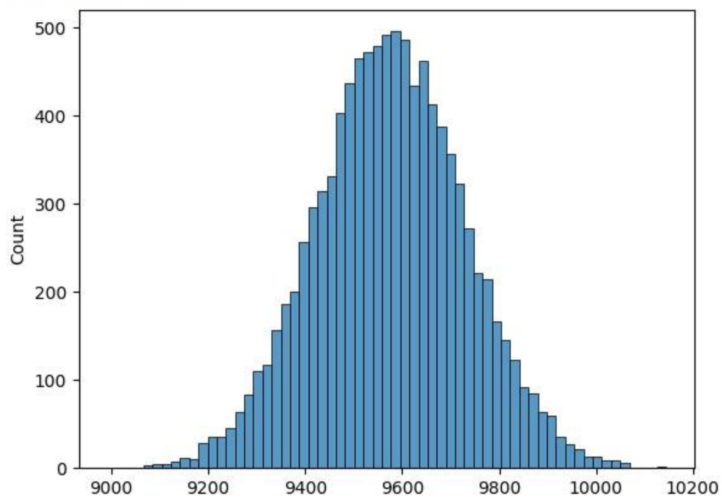
90% Confidence intervals - Married Vs Unmarried

```
[46] # 90% Confidence interval for married

     x1 = np.percentile(bootstrap_married_samples_mean, 5)
     x2 = np.percentile(bootstrap_married_samples_mean, 95)
     confidence_interval_90_perc_married = [x1,x2]

     print(f"Avearge spending of married--->{np.mean(bootstrap_married_samples_mean)}")
     print(confidence_interval_90_perc_married)
```

```
Avearge spending of married--->9575.553733639006
[9320.834743553214, 9830.855932759692]
```

```
[47] sns.histplot(bootstrap_married_samples_mean)
```

```
<Axes: ylabel='Count'>
```



From the above confidence interval, it is concluded that average spending of Unmarried customers of population lie in [9320.834743553214, 9830.855932759692]with 90% confident and average spending is 9575.553733639006

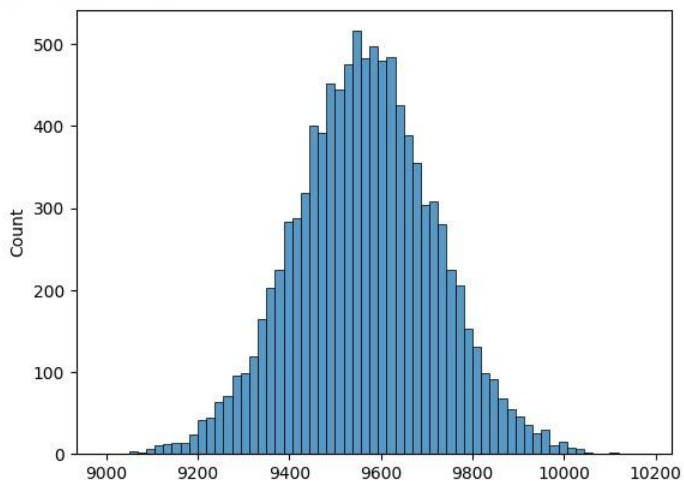
Name: Raghu
Business Case: Walmart – Confidence Interval and

```
[51] # 90% Confidence interval for Unmarried

x1 = np.percentile(bootstrap_unmarried_samples_mean, 5)
x2 = np.percentile(bootstrap_unmarried_samples_mean, 95)
confidence_interval_90_perc_unmarried = [x1,x2]

print(f"Avearge spending of unmarried--->{np.mean(bootstrap_unmarried_samples_mean)}")
print(confidence_interval_90_perc_unmarried)
```

```
Avearge spending of unmarried--->9564.813887277192
[9311.596499382862, 9818.099171393029]
```

```
sns.histplot(bootstrap_unmarried_samples_mean)
```

```
<Axes: ylabel='Count'>
```



From the above confidence interval, it is concluded that average spending of Unmarried customers of population lie in [9311.596499382862, 9818.099171393029]with 90% confident and average spending is 9564.813887277192

Constructing confidential intervals for spending of users of different age groups

```
[54] # Average spending of each user for different age groups

df_age_0to17 = df[df["Age"] == '0-17'].groupby("User_ID")["Purchase"].mean()
df_age_18t025 = df[df["Age"] == '18-25'].groupby("User_ID")["Purchase"].mean()
df_age_26to35 = df[df["Age"] == '26-35'].groupby("User_ID")["Purchase"].mean()
df_age_46to50 = df[df["Age"] == '46-50'].groupby("User_ID")["Purchase"].mean()
df_age_51to55 = df[df["Age"] == '51-55'].groupby("User_ID")["Purchase"].mean()
df_age_55plus = df[df["Age"] == '55+'].groupby("User_ID")["Purchase"].mean()
```

95% confidential interval - age

```
[60] #  Generating 95% confidential intervals for different age groups
var_list = ["df_age_0to17", "df_age_18t025", "df_age_26to35", "df_age_46to50","df_age_51to55", "df_age_55plus"]
j = 0
for i in [df_age_0to17, df_age_18t025, df_age_26to35, df_age_46to50,df_age_51to55, df_age_55plus]:

  bootstrap_age_means_samples = []
  for k in range(10000):
    bootstrap_sample = np.random.choice(i,size = 100)
    bootstrap_sample_mean = np.mean(bootstrap_sample)
    bootstrap_age_means_samples.append(bootstrap_sample_mean)

  x1 = np.percentile(bootstrap_age_means_samples, 2.5)
  x2 = np.percentile(bootstrap_age_means_samples, 97.5)
  confidence_interval_95_perc = [x1,x2]

  print(f"Avearge spending of customers of{var_list[j]} --->{np.mean(bootstrap_age_means_samples)}")
  print(f"95% confidence interval for {var_list[j]} --> {confidence_interval_95_perc}")
  print('-------------------------------------')
  j = j+1
```

```
Avearge spending of customers ofdf_age_0to17 --->8986.72848095367
95% confidence interval for df_age_0to17 --> [8614.300223107963, 9361.346915684746]
-------------------------------------
Avearge spending of customers ofdf_age_18t025 --->9515.485979223677
95% confidence interval for df_age_18t025 --> [9125.518394569192, 9916.710043098492]
-------------------------------------
Avearge spending of customers ofdf_age_26to35 --->9607.36521486867
95% confidence interval for df_age_26to35 --> [9250.557599560332, 9961.595398958218]
-------------------------------------
Avearge spending of customers ofdf_age_46to50 --->9563.898660593859
95% confidence interval for df_age_46to50 --> [9205.637465568461, 9923.641059567131]
-------------------------------------
Avearge spending of customers ofdf_age_51to55 --->9627.759006287091
95% confidence interval for df_age_51to55 --> [9262.370652895197, 10007.162363225108]
-------------------------------------
Avearge spending of customers ofdf_age_55plus --->9404.807285298186
95% confidence interval for df_age_55plus --> [9012.342841902377, 9811.377765104697]
-------------------------------------
```

From the above 95% confidence intervals for different age groups, it is concluded that average spending of age group 0-17 of population are lower compared to other age groups and there is no noticeable difference in avearge spending of other age groups Low average spending of age group 0 to 17 may be because they are students that they are not earning.

Name: Raghu
Business Case: Walmart – Confidence Interval and