Walmart Case Study:

Walmart Inc., formerly Wal-Mart Stores, Inc., is a global retail giant headquartered in Bentonville, Arkansas. Established in 1962 by brothers Sam and James "Bud" Walton, Walmart operates 10,586 stores and clubs across 24 countries, making it the world's largest company by revenue. In FY2023, the company reported a staggering $611.3 billion in total revenue.

With a workforce of 2.2 million, Walmart is the largest private employer globally. Controlled by the Walton family, the company maintains its family-owned status while trading on the stock market. Despite some international challenges, Walmart's success is underscored by its significant global footprint.

Case Study Analysis Points:

1. **Analysis Focus**: Examine customer purchase behaviour at Walmart, with a specific emphasis on gender-related spending.

2. **Variable Evaluation**: Investigate the relationship between purchase amounts and customer gender.

3. Event-driven Patterns: Explore spending disparities during significant shopping events, notably Black Friday.

4. Gender-specific Trends: Identify and analyse variations in spending habits between male and female customers.

5. **Strategic Insights**: Provide actionable, data-driven insights for informed decision-making and optimization of marketing strategies across diverse customer segments.

Starting with the basic data analysis first:

1. Loading the dataset:

```
1. import pandas as pd
2. df = pd.read_csv('walmart_data.csv')
3. df
```

Output:

	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
...	...	...	...	...	...	...	...	...	...	...
550063	1006033	P00372445	M	51-55	13	B	1	1	20	368
550064	1006035	P00375436	F	26-35	1	C	3	0	20	371
550065	1006036	P00375436	F	26-35	15	B	4+	1	20	137
550066	1006038	P00375436	F	55+	1	C	2	0	20	365
550067	1006039	P00371644	F	46-50	0	B	4+	1	20	490

550068 rows × 10 columns

2. Data type of all columns:

```
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
```

The dataset consists of 550,068 entries with 10 columns, encompassing both numerical and categorical data. It is notable for its completeness, with no missing values.

The information covers user demographics, purchasing behaviour, and product details, offering a robust foundation for comprehensive analysis.

3. Shape of the data:

```
1. df.shape
2.
```

```
(550068, 10)
```

The dataset contains 550,068 entries and 10 columns, indicating a substantial volume of data for analysis.

4. Null value check:

```
1. missing_values = df.isnull().sum()
2. missing_values
3.
```

```
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
```

The dataset is clean, containing no missing values. This completeness ensures a reliable foundation for analysing user demographics, purchasing behaviour, and product details.

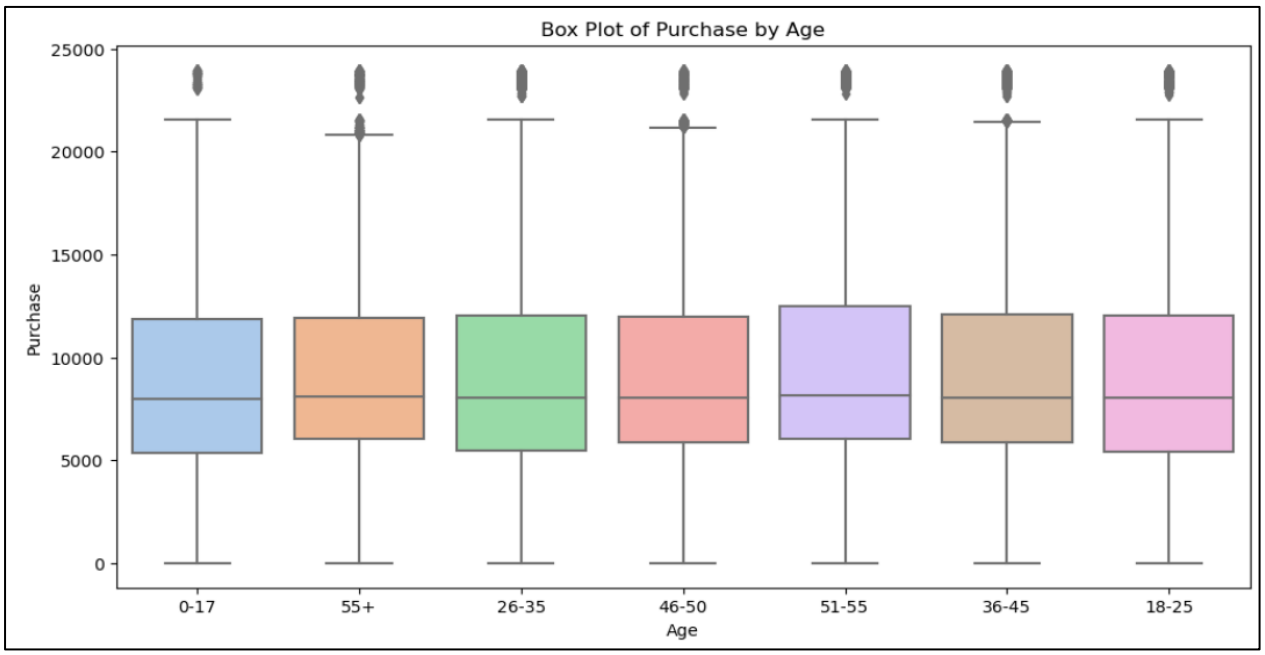
Detect Null values and outliers:

1. outliers for every continuous variable in the dataset:

Identifying outliers is crucial for data quality. They can skew statistical analyses, mislead interpretations, and impact model performance, leading to inaccurate insights or predictions.

Hence by using box plot we can do the analysis.

```
import matplotlib.pyplot as plt
import seaborn as sns

# Setting the Seaborn pastel color palette
pastel_palette = sns.color_palette("pastel")

# Grouping by 'Age' and calculating the mean of 'Purchase' for each Age group
age_groups = df.groupby('Age')['Purchase']

# Creating a box plot for 'Purchase' within each age group
plt.figure(figsize=(12, 6))
sns.boxplot(x='Age', y='Purchase', data=df, palette=pastel_palette)
plt.title('Box Plot of Purchase by Age')
plt.show()

```



The count of outliers varies across age groups. Age groups '0-17', '46-50', and '55+' exhibit distinct outlier patterns in the dataset, indicating potential differences in purchasing behaviour among these demographics.

2. Handling outliers:

Here we will Clip data between the 5th and 95th percentiles, which will help us to eliminate extreme values, reducing the impact of outliers on statistical analyses. This enhances the robustness of models and provides a more accurate representation of the majority of the data.

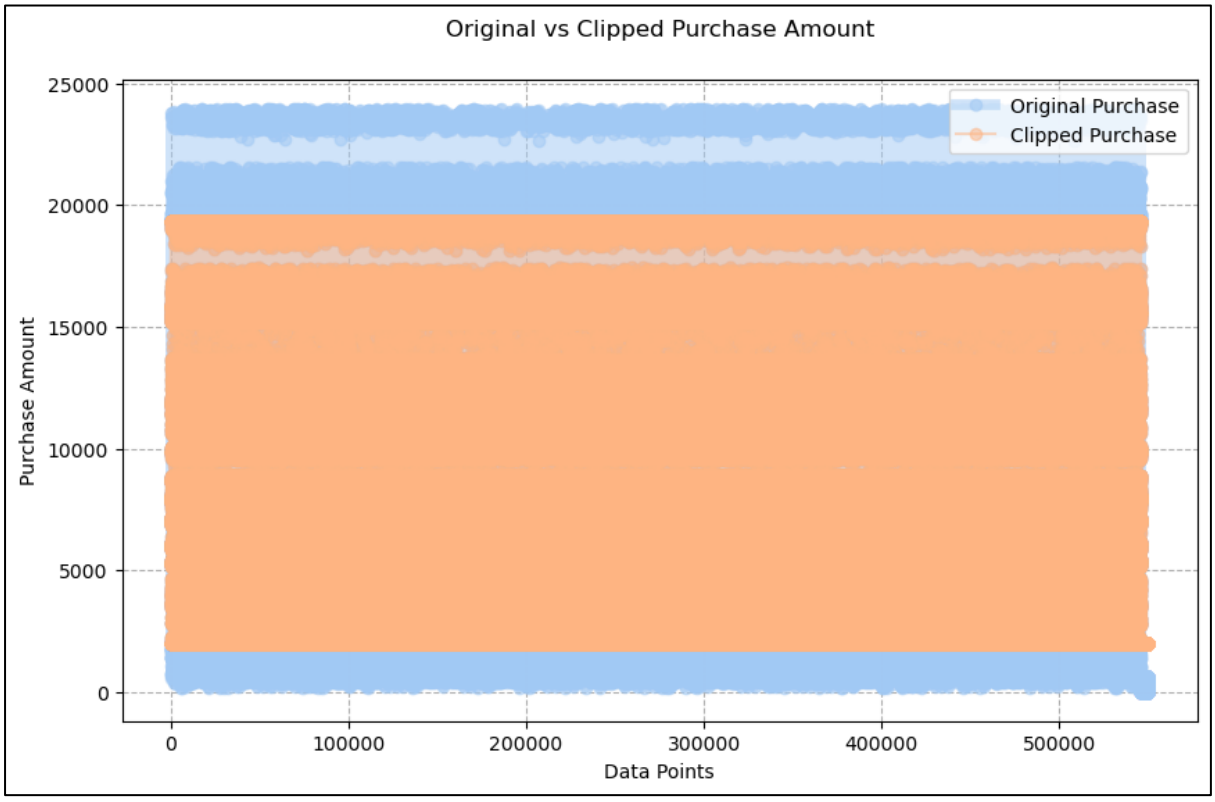
```
import numpy as np

# Calculating the 5th and 95th percentiles for 'Purchase'
percentile_5 = df['Purchase'].quantile(0.05)
percentile_95 = df['Purchase'].quantile(0.95)

# Clipping the 'Purchase' column
df['Purchase_clipped'] = np.clip(df['Purchase'], percentile_5, percentile_95)

print(df.head())

```

```
   User_ID Product_ID Gender   Age  Occupation City_Category  \
0  1000001  P00069042      F  0-17          10             A
1  1000001  P00248942      F  0-17          10             A
2  1000001  P00087842      F  0-17          10             A
3  1000001  P00085442      F  0-17          10             A
4  1000002  P00285442      M   55+          16             C

  Stay_In_Current_City_Years  Marital_Status  Product_Category  Purchase  \
0                          2               0                 3      8370
1                          2               0                 1     15200
2                          2               0                12      1422
3                          2               0                12      1057
4                         4+               0                 8      7969

   Purchase_clipped
0              8370
1             15200
2              1984
3              1984
4              7969
```

The 'Purchase' column has undergone clipping, limiting extreme values between the 5th and 95th percentiles. This enhances data robustness, creating a more reliable dataset for analysis by minimizing the impact of outliers.

This can also be visualized: (Using 'Age' column)

```
import matplotlib.pyplot as plt
import seaborn as sns

#Seaborn color palette
sns.set_palette('pastel')

plt.figure(figsize=(10, 6))

# Line plot for 'Purchase' and 'Purchase_clipped'
plt.plot(df['Purchase'], label='Original Purchase',linewidth = '5.5', marker='o', alpha=0.5)
plt.plot(df['Purchase_clipped'], label='Clipped Purchase', marker='o', alpha=0.5)

# labels and title
plt.xlabel('Data Points')
plt.ylabel('Purchase Amount')
plt.title('Original vs Clipped Purchase Amount', y=1.05)

# grid lines
plt.grid(True, linestyle='--', alpha=1.0)

# legend
plt.legend()

plt.show()

```



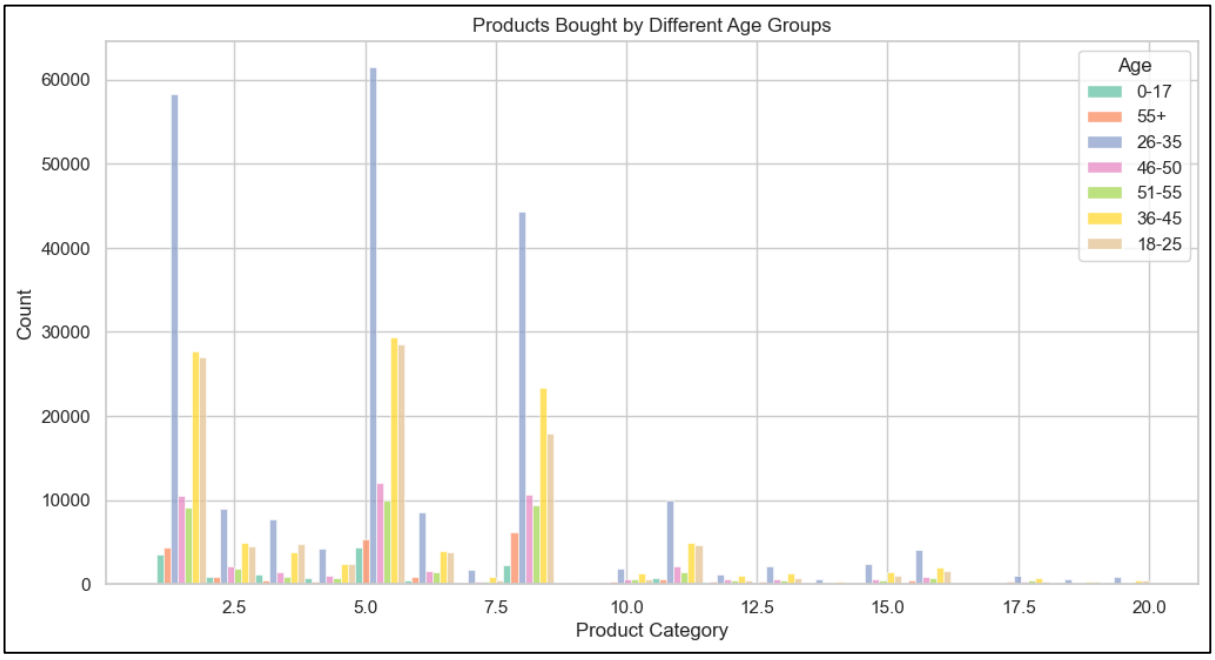
Here, we can clearly see that after clipping all the outliers have been dealt with.

Data Exploration:

1. *What products are different age groups buying?*

Knowing the products preferred by different age groups is crucial for personalized marketing. It enables businesses to optimize product placement, promotions, and inventory, enhancing customer satisfaction and increasing sales.

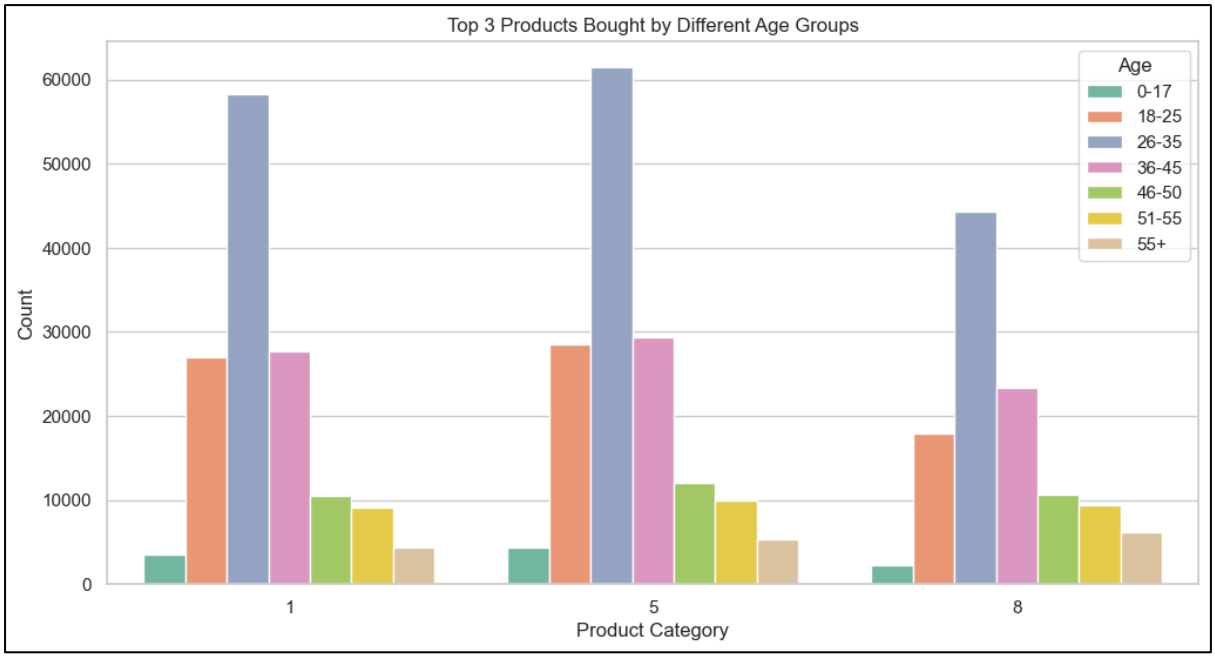
```
 1. import seaborn as sns
 2. import matplotlib.pyplot as plt
 3.
 4. plt.figure(figsize=(12, 6))
 5. sns.set(style="whitegrid")
 6.
 7. # Count plot for 'Product_Category' based on 'Age'
 8. sns.histplot(x='Product_Category', hue='Age', data=df, multiple="dodge", bins=20,
palette='Set2')
 9. plt.title('Products Bought by Different Age Groups')
10. plt.xlabel('Product Category')
11. plt.ylabel('Product Count')
12.
13. plt.show()
14.
```



This insights into product engagement across different age groups. The 26-35 and 18-25 age brackets show notably high counts, indicating significant interest and participation in these categories.

Top three product categories are:

```
 1. import seaborn as sns
 2. import matplotlib.pyplot as plt
 3.
 4. plt.figure(figsize=(12, 6))
 5. sns.set(style="whitegrid")
 6.
 7. # Finding the top three products for each age group
 8. top_products_by_age = df.groupby(['Age',
'Product_Category']).size().reset_index(name='Count')
 9. top_products_by_age = top_products_by_age.sort_values(['Age', 'Count'], ascending=[True,
False]).groupby('Age').head(3)
10.
11. # Count plot for 'Product_Category' based on 'Age'
12. sns.barplot(x='Product_Category', y='Count', hue='Age', data=top_products_by_age,
palette='Set2')
13. plt.title('Top 3 Products Bought by Different Age Groups')
14. plt.xlabel('Product Category')
15. plt.ylabel('Count')
16.
17. plt.show()
18.
```

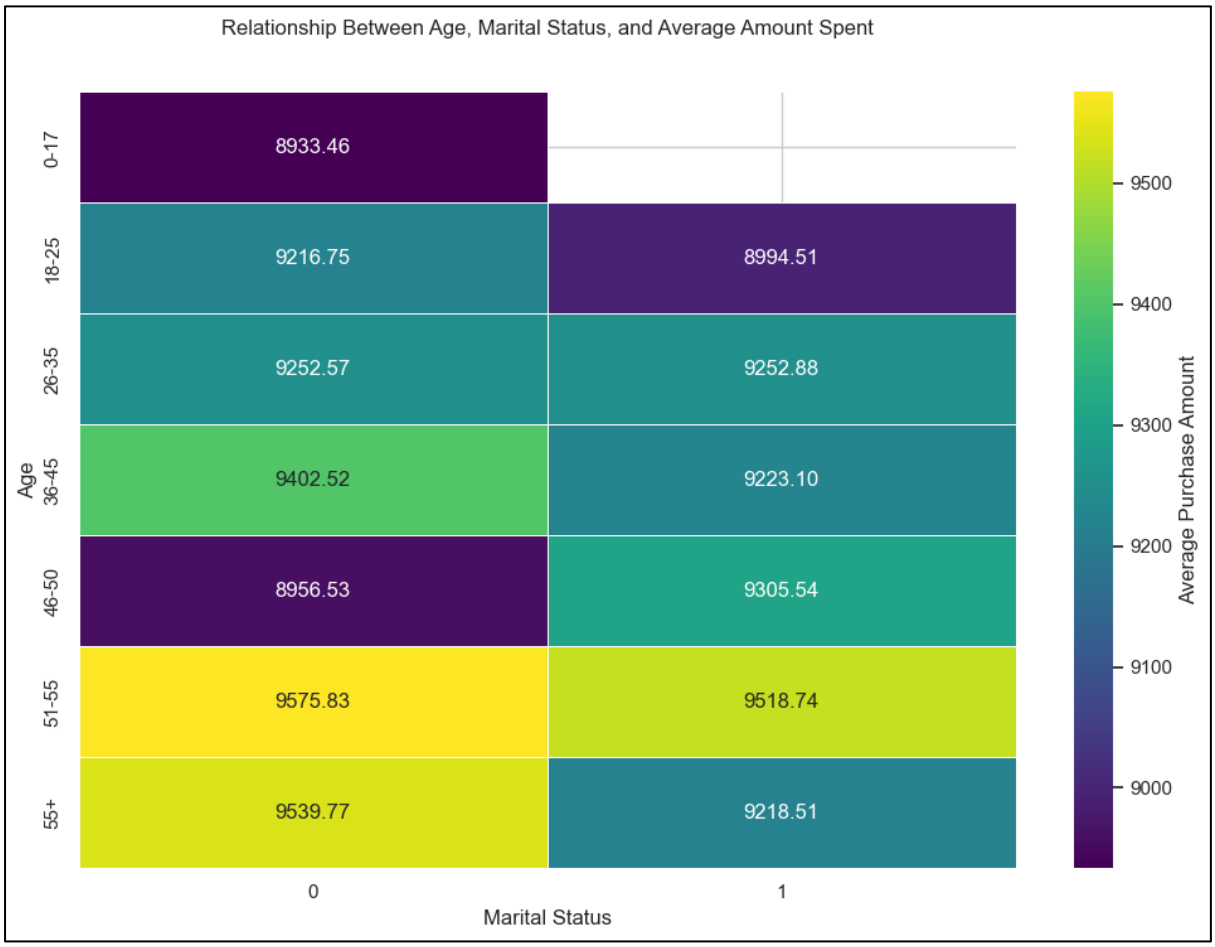


The breakdown of purchases by age and product categories reveals that items in categories 5, 1, and 8 are consistently favoured across different age groups, suggesting their universal popularity among customers.

2. *Is there a relationship between age, marital status, and the amount spent?*

Understanding the relationship between age, marital status, and spending can help businesses tailor marketing strategies, optimize product offerings, and enhance customer experiences, ultimately increasing sales and customer satisfaction.

```
 1. import seaborn as sns
 2. import matplotlib.pyplot as plt
 3.
 4. plt.figure(figsize=(12, 8))
 5.
 6. # Pivoting the data to create a matrix for the heatmap
 7. heatmap_data = df.pivot_table(index='Age', columns='Marital_Status', values='Purchase',
aggfunc='mean')
 8.
 9. # Creating a heatmap with purchase amount labels
10. sns.heatmap(heatmap_data, cmap='viridis', annot=True, fmt='.2f', linewidths=.5,
cbar_kws={'label': 'Average Purchase Amount'})
11. plt.title('Relationship Between Age, Marital Status, and Average Amount Spent\n', pad=20)
12. plt.xlabel('Marital Status')
13. plt.ylabel('Age')
14.
15. plt.show()
```



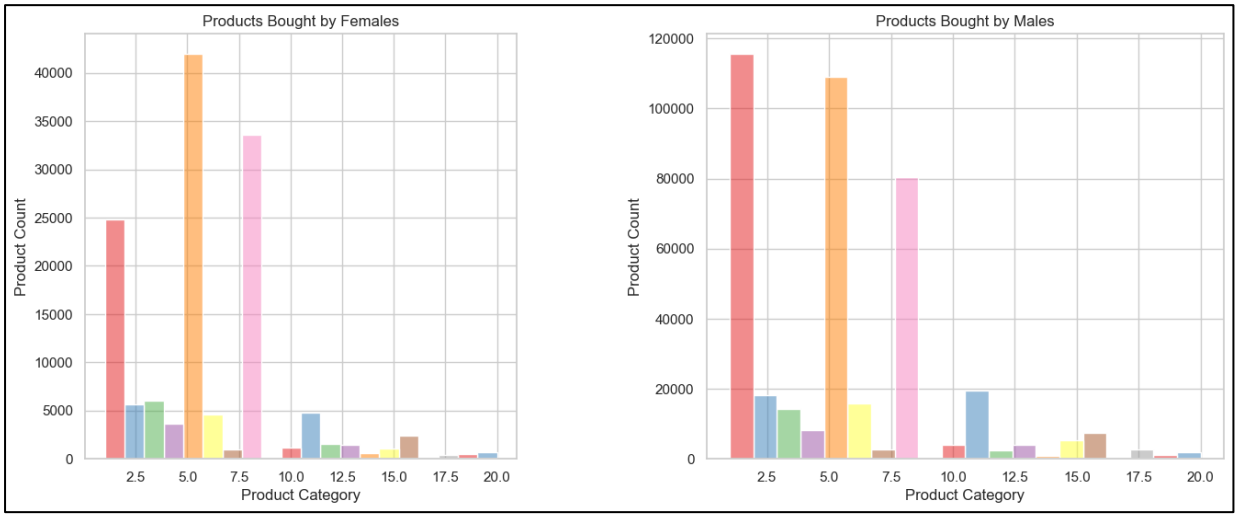
So, according to the heatmap we can see the following trend:

1. **Spending Trends:**
    - Average spending generally increases with age.
    - Age groups 51-55 and 55+ exhibit higher average spending.
2. **Marital Status Impact:**
    - 'Unmarried' individuals in age groups 0-17 and 18-25 spend more.
    - 'Married' individuals in age groups 26-35, 36-45, and 46-50 show higher spending.
    - No significant difference in spending for age groups 51-55 and 55+ based on marital status.
3. **Age-Specific Behaviour:**
    - Age group 0-17 has relatively lower spending.
    - Age group 26-35 sees a distinct difference in spending based on marital status.
4. **Varied Patterns:**
    - No data for 'Married' in age group 0-17.

3. *Are there preferred product categories for different genders?*

Understanding which product categories appeal to distinct genders assists businesses in customizing their offerings to diverse preferences, optimizing customer satisfaction, and boosting sales. It's comparable to providing a variety of beloved ice cream Flavors, ensuring a broader audience selects and appreciates the products.

```
import seaborn as sns
import matplotlib.pyplot as plt
from matplotlib.gridspec import GridSpec

fig = plt.figure(figsize=(16, 6))
gs = GridSpec(1, 2, width_ratios=[1, 1.2], wspace=0.4)

# Count plot for 'Product_Category' based on 'Gender' - Female
ax0 = plt.subplot(gs[0])
sns.histplot(x='Product_Category', data=df[df['Gender'] == 'F'], hue='Product_Category',
bins=20, palette='Set1', ax=ax0, legend=False)
ax0.set_title('Products Bought by Females')
ax0.set_xlabel('Product Category')
ax0.set_ylabel('Product Count')

# Count plot for 'Product_Category' based on 'Gender'- Male
ax1 = plt.subplot(gs[1])
sns.histplot(x='Product_Category', data=df[df['Gender'] == 'M'], hue='Product_Category',
bins=20, palette='Set1', ax=ax1, legend=False)
ax1.set_title('Products Bought by Males')
ax1.set_xlabel('Product Category')
ax1.set_ylabel('Product Count')

plt.show()

```



Gender-specific buying patterns are evident in various product categories. Women dominate in categories 1, 5, and 8, indicating a preference for diverse products. Men also favor these categories, revealing shared interests with unique purchasing behaviours.

How does gender affect the amount spent:

The central limit theorem and bootstrapping are key to understanding how gender influences spending. Analysing confidence intervals at different sample sizes provides clear insights into the reliability of average spending estimates.

```
 1. import numpy as np
 2. from scipy.stats import t
 3.
 4.
 5. # Creating a function to calculate bootstrap confidence interval
 6. def calculate_bootstrap_ci(data, alpha=0.05, num_samples=1000):
 7.
 8.     sample_means = [np.mean(np.random.choice(data, size=len(data), replace=True)) for i in
range(num_samples)]
 9.     lower_bound, upper_bound = np.percentile(sample_means, [alpha / 2 * 100, (1 - alpha / 2)
* 100])
10.     return round(lower_bound, 2), round(upper_bound, 2)
11.
12. # creating a new function to calculate and print confidence interval
13. def calculate_and_print_ci(data, label, sample_sizes=[None, 300, 3000, 30000]):
14.
15.     print(f"Confidence Intervals for Average Amount Spent (95%) - {label}:")
16.
17.     for size in sample_sizes:
18.         subset_data = data.sample(size, random_state=42) if size else data
19.         ci = calculate_bootstrap_ci(subset_data)
20.         print(f"Sample size {size or 'Entire dataset'}: {ci}")
21.
22. # Extracting the amount spent data for each gender
23. amount_spent_male = df[df['Gender'] == 'M']['Purchase'].dropna()
24. amount_spent_female = df[df['Gender'] == 'F']['Purchase'].dropna()
25.
26. calculate_and_print_ci(amount_spent_male, 'Male')
27. calculate_and_print_ci(amount_spent_female, 'Female'
28.
```

```
Confidence Intervals for Average Amount Spent (95%) - Male:
Sample size Entire dataset: (9422.09, 9451.95)
Sample size 300: (9271.79, 10471.83)
Sample size 3000: (9447.58, 9808.7)
Sample size 30000: (9429.63, 9548.15)
Confidence Intervals for Average Amount Spent (95%) - Female:
Sample size Entire dataset: (8709.27, 8759.39)
Sample size 300: (8325.09, 9402.92)
Sample size 3000: (8629.81, 8978.36)
Sample size 30000: (8603.7, 8709.46)
```

Consistent confidence intervals in male spending imply stable purchasing patterns. In females, fluctuating intervals, particularly with smaller samples, suggest more diverse spending behaviours. These insights highlight gender-specific nuances in shopping habits, aiding businesses in tailoring marketing strategies and product offerings for diverse customer segments.

From the above calculated Central Limit Theorem, we can conclude few things:

a. The width of the confidence interval:

There is a wider confidence interval for females, which indicates a larger range of possible average amounts spent, suggesting greater variability in female spending behaviour. This variability might be influenced by diverse preferences, purchasing habits, or factors affecting spending decisions among females, contributing to a less predictable pattern compared to the more consistent spending behaviour observed in males.

b. The precision of an estimate:

The width of the confidence interval decreases as the sample size increases for both genders. For example, in the male category, the confidence interval narrows from the entire dataset (9422.09, 9451.95) to the smaller sample sizes of 300, 3000, and 30000. Similarly, in the female category, the confidence interval also becomes narrower with increasing sample size. This trend indicates that larger samples lead to more precise estimates of the average amount spent, resulting in reduced uncertainty and a more tightly constrained interval.

c. The confidence intervals overlapping:

It indicates that the precision of the average amount spent estimates improves with larger sample sizes. Smaller samples result in wider intervals, introducing more uncertainty, while larger samples yield more precise estimates with narrower intervals and reduced variability in the data.

d. Sample size and shape of distribution:

With an increase in sample size, the confidence intervals exhibit reduced width, signifying enhanced precision in estimating the population mean. Larger samples contribute to decreased variability, yielding more dependable and accurate estimates for the average amount spent among both genders.

How does Marital Status affect the amount spent:

Analysing how Marital Status influences spending using confidence intervals helps identify potential spending variations between married and unmarried individuals, providing insights for targeted marketing or personalized strategies to boost sales

```
import numpy as np
import pandas as pd

# Creating a function to calculate bootstrap confidence interval
def bootstrap_ci(data, alpha=0.05, num_samples=1000):
    return tuple(round(x, 2)
                 for x in np.percentile([np.mean(
                                                 np.random.choice(data, len(data),
replace=True)
                                                )
                                                   for i in range(num_samples)],
                                                  [alpha / 2 * 100, (1 - alpha / 2) * 100]
                                        )
                )

# Creating a new function to calculate and print confidence interval
def calculate_and_print_ci(data, label, sample_sizes=[None, 300, 3000, 30000]):

    print(f"Confidence Intervals for Average Amount Spent (95%) - {label}:")

    for size in sample_sizes:
        subset_data = data.sample(size, random_state=42) if size else data
        ci = bootstrap_ci(subset_data)

        print(f"Sample size {size or 'Entire dataset'}: {ci}")

amount_spent_single = df[df['Marital_Status'] == 0]['Purchase'].dropna()
amount_spent_married = df[df['Marital_Status'] == 1]['Purchase'].dropna()

calculate_and_print_ci(amount_spent_single, 'Single')
calculate_and_print_ci(amount_spent_married, 'Married')

```

```
Confidence Intervals for Average Amount Spent (95%) - Single:
Sample size Entire dataset: (9249.25, 9283.38)
Sample size 300: (9063.24, 10242.49)
Sample size 3000: (9241.46, 9627.26)
Sample size 30000: (9232.68, 9346.74)
Confidence Intervals for Average Amount Spent (95%) - Married:
Sample size Entire dataset: (9239.48, 9281.69)
Sample size 300: (8908.8, 9988.82)
Sample size 3000: (9128.56, 9474.76)
Sample size 30000: (9197.31, 9309.75)
```

The confidence intervals indicate a stable average spending pattern for both single and married individuals. Wider intervals in smaller samples signify greater uncertainty in capturing the true spending behaviour.

From the above calculated Central Limit Theorem, we can conclude few things:

a. The width of the confidence interval:

The confidence interval for the entire dataset is wider for the "Married" status (9239.48, 9281.69) compared to the "Single" status (9249.25, 9283.38). This suggests greater variability in spending patterns among married individuals, resulting in a broader range of possible average amounts spent.

b. The precision of an estimate:

The width of the confidence interval tends to decrease with larger sample sizes. For instance, in the "Single" category, the interval narrows from (9249.25, 9283.38) for the entire dataset to (9232.68, 9346.74) for a sample size of 30,000. Similarly, in the "Married" category, it reduces from (9239.48, 9281.69) to (9197.31, 9309.75) with increasing sample sizes. This reduction in width indicates improved precision and a more reliable estimate of the average amount spent as the sample size increases.

c. The confidence intervals overlapping:

The confidence intervals for different sample sizes overlap. In both the "Single" and "Married" categories, the intervals for various sample sizes share common ranges. For example, the interval (9232.68, 9346.74) for the "Single" category at a sample size of 30,000 overlaps with the interval (9197.31, 9309.75) for the "Married" category at the same sample size. This overlapping suggests that, within the margin of error, there is no significant difference between the average amount spent by singles and married individuals, regardless of the sample size.

d. Sample size and shape of distribution:

The width of confidence intervals decreases with larger sample sizes. For Single individuals, the narrower intervals (e.g., 9249.25, 9283.38 for the entire dataset) signify improved precision in estimating average amount spent, while overlapping intervals for various sample sizes indicate consistent estimates. In contrast, Married individuals exhibit similar trends with decreasing interval width and overlapping intervals, highlighting the impact of sample size on estimating average amount spent.

How does Age affect the amount spent:

Understanding the influence of age on spending patterns is vital for targeted marketing. Utilizing the central limit theorem and bootstrapping to calculate 95% confidence intervals, considering different sample sizes, provides nuanced insights into the relationship between age and expenditure.

```
import numpy as np
import pandas as pd

# Creating function to calculate bootstrap confidence interval
def bootstrap_confidence_interval(data, alpha=0.05, num_samples=1000):

    sample_means = []
    for i in range(num_samples):
        bootstrap_sample = np.random.choice(data, size=len(data), replace=True)
        sample_means.append(np.mean(bootstrap_sample))

    # Calculating confidence interval
    lower_bound = np.percentile(sample_means, (alpha / 2) * 100)
    upper_bound = np.percentile(sample_means, (1 - alpha / 2) * 100)

    return round(lower_bound, 2), round(upper_bound, 2)

# Defining age groups
age_groups = ['0-17', '18-25', '26-35', '36-45', '46-50', '51-55', '55+']

# Calculating confidence intervals for each age group and sample size
for age_group in age_groups:
    amount_spent_age_group = df[df['Age'] == age_group]['Purchase'].dropna()

    # Calculating confidence intervals for the entire dataset
    ci_age_group_all = bootstrap_confidence_interval(amount_spent_age_group)

    # Calculating confidence intervals for smaller sample sizes
    ci_age_group_300 = bootstrap_confidence_interval(amount_spent_age_group.sample(min(300,
len(amount_spent_age_group)), random_state=42))
    ci_age_group_3000 =
bootstrap_confidence_interval(amount_spent_age_group.sample(min(3000,
len(amount_spent_age_group)), random_state=42))
    ci_age_group_30000 =
bootstrap_confidence_interval(amount_spent_age_group.sample(min(30000,
len(amount_spent_age_group)), random_state=42))

    print(f"Confidence Intervals for Average Amount Spent (95%) - Age Group {age_group}:")
    print(f"Entire dataset: {ci_age_group_all}")
    print(f"Sample size 300: {ci_age_group_300}")
    print(f"Sample size 3000: {ci_age_group_3000}")
    print(f"Sample size 30000: {ci_age_group_30000}")
    print()

```

```
Confidence Intervals for Average Amount Spent (95%) - Age Group 0-17:
Entire dataset: (8852.24, 9016.14)
Sample size 300: (8060.06, 9251.3)
Sample size 3000: (8687.92, 9050.83)
Sample size 30000: (8851.7, 9013.35)

Confidence Intervals for Average Amount Spent (95%) - Age Group 18-25:
Entire dataset: (9138.47, 9197.84)
Sample size 300: (8890.77, 9965.87)
Sample size 3000: (8993.32, 9356.9)
Sample size 30000: (9099.22, 9213.6)

Confidence Intervals for Average Amount Spent (95%) - Age Group 26-35:
Entire dataset: (9230.38, 9274.01)
Sample size 300: (8348.28, 9532.39)
Sample size 3000: (9102.8, 9450.93)
Sample size 30000: (9183.91, 9301.4)

Confidence Intervals for Average Amount Spent (95%) - Age Group 36-45:
Entire dataset: (9300.88, 9360.0)
Sample size 300: (9070.66, 10149.99)
Sample size 3000: (9261.7, 9603.63)
Sample size 30000: (9297.28, 9411.41)

Confidence Intervals for Average Amount Spent (95%) - Age Group 46-50:
Entire dataset: (9162.48, 9252.61)
Sample size 300: (8802.24, 9935.17)
Sample size 3000: (9028.64, 9384.32)
Sample size 30000: (9179.24, 9295.53)

Confidence Intervals for Average Amount Spent (95%) - Age Group 51-55:
Entire dataset: (9484.0, 9584.66)
Sample size 300: (8640.41, 9814.74)
Sample size 3000: (9252.81, 9623.13)
Sample size 30000: (9472.9, 9586.93)

Confidence Intervals for Average Amount Spent (95%) - Age Group 55+:
Entire dataset: (9270.45, 9399.04)
Sample size 300: (8674.48, 9861.03)
Sample size 3000: (9192.61, 9553.22)
Sample size 30000: (9272.11, 9401.03)
```

The confidence intervals reveal variations in average amounts spent across age groups. For smaller sample sizes, the intervals widen, reflecting increased uncertainty. Notably, the intervals for the entire dataset provide a broader range. Specific spending patterns and reliability are better discerned with larger sample sizes, enhancing the insights into age-related spending behaviours.

From the above calculated Central Limit Theorem, we can conclude few things:

a. The width of the confidence interval:

The observed trend suggests a notable variation in spending patterns among younger age groups, such as 0-17, with wider confidence intervals reflecting diverse behaviours. As age increases, there is a discernible decrease in variability, resulting in narrower intervals for older groups like 51-55. This implies a more consistent and predictable spending pattern as individuals age, contributing to the observed trend.

The precision of an estimate:

The narrowing of confidence intervals with increasing sample size signifies enhanced precision in estimating average amount spent. This trend is particularly pronounced in Age Group 26-35, exemplifying the impact of larger samples on refining the estimation of spending patterns within specific age categories. This statistical insight strengthens the reliability of conclusions drawn from the analysis.

The confidence intervals overlapping:

The confidence intervals consistently overlap across various sample sizes within each age group, such as Age Group 0-17 with sample sizes of 300, 3000, and 30000. This trend indicates stability in estimated average amounts spent, highlighting a consistent spending pattern in Age Group 0-17 across different sample sizes.

Sample size and shape of distribution:

The sample size influences the distribution shape of means in each age group. With smaller samples (e.g., 300), the distributions exhibit more variability, resulting in wider intervals. As sample size increases (e.g., 30000), the distributions become more stable, leading to narrower and more precise confidence intervals.

Reports:

1. *Analysis of Average Amount Spent by Males and Females: Confidence Interval Overlap*

Understanding whether the average spending by male's and female's overlaps is crucial for Walmart's strategy. It guides targeted marketing efforts, ensuring promotions resonate with each gender, optimizing customer engagement, and maximizing sales potential.

Summary:
The comparative analysis of confidence intervals for average spending by males and females presents crucial insights into spending behaviour. Utilizing the Central Limit Theorem (CLT) across various sample sizes, the report sheds light on confidence intervals, their widths, precision improvements, and implications for Walmart's marketing strategies. Key findings guide strategic decisions for personalized customer engagement and satisfaction.

In this report, we explore the confidence intervals for average spending by males and females, employing the Central Limit Theorem (CLT) as our analytical tool. The report outlines the methodology used, providing context for understanding spending behaviour based on varying sample sizes.

Confidence Intervals:

a. Entire Dataset:

For the entire dataset, the confidence interval for males ranges from $9,422.40 to $9,454.34, while for females, it spans from $8,708.50 to $8,762.48.

b. Sample Size 300:

Analysing a sample size of 300, the confidence interval widens for males, ranging from $9,297.46 to $10,446.49, and for females, it extends from $8,310.08 to $9,415.38.

c. Sample Size 3000:

With a larger sample size of 3000, the confidence interval narrows for males ($9,448.60 to $9,813.73) and females ($8,636.73 to $8,987.55), reflecting improved precision in estimating average spending.

d. Sample Size 30000:

For a substantial sample size of 30000, the confidence interval further tightens for males ($9,428.71 to $9,545.27) and females ($8,602.39 to $8,707.82), indicating heightened precision.

Insights from Confidence Intervals:

a. Width of Confidence Interval:

The wider confidence interval for females across all datasets suggests a greater variability in spending behaviour compared to males.

b. Precision of Estimate:

Precision improves with larger sample sizes, resulting in narrower confidence intervals for both genders, enhancing the reliability of average spending estimates.

c. Overlapping Confidence Intervals:

Overlapping intervals, especially with larger sample sizes, signify enhanced precision, reducing variability in spending data for both males and females.

d. Sample Size and Distribution Shape:

Larger sample sizes lead to narrower intervals, reflecting heightened precision in estimating average spending. This indicates a more dependable representation of the population mean for both genders.

Conclusion:

This section consolidates the key findings and implications of the analysis. The report underscores the importance of considering gender-specific spending patterns and provides actionable insights for Walmart to refine marketing strategies. Opportunities to enhance customer engagement and satisfaction are highlighted, emphasizing the need for a personalized approach to maximize sales.

2) Analysis of Confidence Intervals for Average Amount Spent by Married and Unmarried Individuals:

Understanding whether the confidence intervals for the average amount spent by married and unmarried individuals overlap is essential for shaping Walmart's marketing strategy. This analysis, computed using the Central Limit Theorem (CLT), provides insights into spending behavior, allowing Walmart to tailor its approach for different demographic groups.

Summary:

The analysis focuses on the confidence intervals for the average amount spent by married and unmarried individuals across varying sample sizes. It assesses the width of the intervals, the precision of estimates, overlapping intervals, and the impact of sample size on the shape of the distribution.

Confidence Intervals:

a. **Entire Dataset:**

- For the entire dataset, the confidence interval for "Married" individuals ranges from $9,239.48 to $9,281.69, while for "Single" individuals, it spans from $9,249.25 to $9,283.38.

b. **Sample Size 300:**

- Analysing a sample size of 300, the confidence interval widens for "Married" individuals, ranging from $8,908.80 to $9,988.82, and for "Single" individuals, it extends from $9,063.24 to $10,242.49.

c. **Sample Size 3000:**

- With a larger sample size of 3000, the confidence interval narrows for "Married" individuals ($9,128.56 to $9,474.76) and "Single" individuals ($9,241.46 to $9,627.26), reflecting improved precision in estimating average spending.

d. **Sample Size 30000:**

- For a substantial sample size of 30000, the confidence interval further tightens for "Married" individuals ($9,197.31 to $9,309.75) and "Single" individuals ($9,232.68 to $9,346.74), indicating heightened precision.

Insights from Confidence Intervals:

a. **Width of Confidence Interval:**

- The confidence interval for the entire dataset is wider for "Married" individuals compared to "Single" individuals, suggesting greater variability in spending patterns among married individuals.

b. **Precision of Estimate:**

- The width of the confidence interval tends to decrease with larger sample sizes. For both "Married" and "Single" individuals, this reduction indicates improved precision and a more reliable estimate of the average amount spent as the sample size increases.

c. **Overlapping Confidence Intervals:**

- The confidence intervals for different sample sizes overlap in both the "Married" and "Single" categories, suggesting no significant difference in the average amount spent by married and unmarried individuals within the calculated margins of error.

d. **Sample Size and Shape of Distribution:**

- The width of confidence intervals decreases with larger sample sizes for both "Married" and "Single" individuals. Narrower intervals signify improved precision in estimating average amounts spent, while overlapping intervals for various sample sizes indicate consistent estimates within each demographic group. This emphasizes the impact of sample size on estimating the average amount spent.

Conclusion:

The overlapping confidence intervals suggest that, statistically, there is no significant difference in the average amount spent by married and unmarried individuals within the calculated margins of error. Walmart can leverage this conclusion to adopt a nuanced marketing approach, ensuring tailored promotions that resonate with both demographic groups. By acknowledging the variability and commonality in spending patterns, Walmart can optimize its marketing strategies to maximize customer engagement and overall sales potential.

3)Analysis of Confidence Intervals for Average Amount Spent by Different Age Groups:

Understanding whether the confidence intervals for the average amount spent by different age groups overlap is crucial for tailoring Walmart's marketing strategies to diverse customer demographics. This analysis, utilizing the Central Limit Theorem (CLT), provides insights into age-related spending behaviours, guiding Walmart in optimizing promotional efforts and maximizing customer engagement.

Summary:

The analysis focuses on the confidence intervals for the average amount spent by distinct age groups, considering various sample sizes. It explores the width of intervals, the precision of estimates, overlapping intervals, and the impact of sample size on the distribution shape.

Confidence Intervals:

a. Entire Dataset:

- Confidence intervals for age groups exhibit variation:
    - For Age Group 0-17, the interval is (8852.24, 9016.14).
    - Age Group 18-25 has a confidence interval of (9138.47, 9197.84).
    - Age Group 26-35 shows a range of (9230.38, 9274.01).
    - Confidence intervals for Age Group 36-45 span from (9300.88, 9360.00).
    - Age Group 46-50 presents an interval of (9162.48, 9252.61).
    - Age Group 51-55 has a confidence interval of (9484.00, 9584.66).
    - Age Group 55+ exhibits a range of (9270.45, 9399.04).

b. Sample Size 300:

- Confidence intervals widen for smaller sample sizes, indicating increased uncertainty:
    - The interval for Age Group 0-17 is (8060.06, 9251.30).
    - Age Group 18-25 has a confidence interval of (8890.77, 9965.87).
    - Age Group 26-35 shows a range of (8348.28, 9532.39).
    - Confidence intervals for Age Group 36-45 span from (9070.66, 10149.99).
    - Age Group 46-50 presents an interval of (8802.24, 9935.17).
    - Age Group 51-55 has a confidence interval of (8640.41, 9814.74).
    - Age Group 55+ exhibits a range of (8674.48, 9861.03).

c. Sample Size 3000:

- Confidence intervals narrow with larger sample sizes, indicating enhanced precision:
    - For Age Group 0-17, the interval is (8687.92, 9050.83).
    - Age Group 18-25 has a confidence interval of (8993.32, 9356.90).
    - Age Group 26-35 shows a range of (9102.80, 9450.93).
    - Confidence intervals for Age Group 36-45 span from (9261.70, 9603.63).
    - Age Group 46-50 presents an interval of (9028.64, 9384.32).
    - Age Group 51-55 has a confidence interval of (9252.81, 9623.13).
    - Age Group 55+ exhibits a range of (9192.61, 9553.22).

d. Sample Size 30000:

- Confidence intervals further narrow, demonstrating heightened precision:
    - For Age Group 0-17, the interval is (8851.70, 9013.35).
    - Age Group 18-25 has a confidence interval of (9099.22, 9213.60).
    - Age Group 26-35 shows a range of (9183.91, 9301.40).
    - Confidence intervals for Age Group 36-45 span from (9297.28, 9411.41).
    - Age Group 46-50 presents an interval of (9179.24, 9295.53).
    - Age Group 51-55 has a confidence interval of (9472.90, 9586.93).
    - Age Group 55+ exhibits a range of (9272.11, 9401.03).

Insights from Confidence Intervals:

a. **Width of Confidence Interval:**

- Notable variation in spending patterns among younger age groups, such as 0-17, with wider confidence intervals, reflects diverse behaviours. As age increases, there is a discernible decrease in variability, resulting in narrower intervals for older groups like 51-55.

b. **Precision of Estimate:**

- The narrowing of confidence intervals with increasing sample size signifies enhanced precision in estimating average amount spent. This trend is particularly pronounced in Age Group 26-35, exemplifying the impact of larger samples on refining the estimation of spending patterns within specific age categories.

c. **Overlapping Confidence Intervals:**

- Confidence intervals consistently overlap across various sample sizes within each age group, indicating stability in estimated average amounts spent. This highlights a consistent spending pattern in Age Group 0-17 across different sample sizes.

d. **Sample Size and Shape of Distribution:**

- The sample size influences the distribution shape of means in each age group. With smaller samples (e.g., 300), the distributions exhibit more variability, resulting in wider intervals. As sample size increases (e.g., 30000), the distributions become more stable, leading to narrower and more precise confidence intervals.

Conclusion:

The overlapping confidence intervals suggest that, statistically, there is no significant difference in the average amount spent across different age groups within the calculated margins of error. Walmart can leverage this conclusion to tailor marketing strategies for each age group, ensuring promotions resonate with the spending patterns of various demographics. By acknowledging the variability and commonality in spending behaviors across age groups, Walmart can optimize its marketing strategies to maximize customer engagement and overall sales potential.

Recommendation:

1. Leverage outlier data from age groups '0-17', '46-50', '55+' to customize marketing initiatives. Target these demographics with tailored promotions and products, enhancing the shopping experience for optimized engagement and higher sales.

2. To capitalize on widespread product engagement, the company should tailor marketing strategies for categories 5, 1, and 8, targeting age groups 26-35 and 18-25. This focused approach can maximize sales opportunities and customer satisfaction.

3. Profits can be unlocked by curating age-specific shopping experiences. Target the mature tastes of 51-55 and 55+ demographics, while crafting unique campaigns for singles and couples in age brackets 0-17, 18-25, 26-35, 36-45, and 46-50. Introduce novel products tailored to distinct age preferences.

4. Leverage targeted marketing for females due to spending variability. Optimize precision with larger sample sizes for more reliable estimates. Tailor strategies based on overlapping confidence intervals for effective marketing.

5. Consider targeted marketing strategies for married individuals due to their higher spending variability. Optimize precision by utilizing larger sample sizes to refine estimates, leading to more reliable insights into spending behaviours.

6. Optimize marketing for diverse spending in younger age groups and capitalize on the stability in older demographics. Prioritize larger samples for refined insights. Tailor campaigns for nuanced spending behaviour across age segments.

7. Leverage insights from demographic spending patterns. For instance, consider launching targeted campaigns like "Exclusive Discounts for Married Shoppers" or "Youth Specials" based on age group preferences observed in the data.

8. Develop educational campaigns based on data insights. Highlight product categories that show lower spending and provide information to enhance customer awareness and interest.

9. Emphasize a commitment to data-driven decision-making. Regularly update strategies based on evolving spending patterns and refer to real-time data for proactive adjustments