In [2]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

## Problem Statement:

```
The Management team at Walmart  aims to analyze customer purchase behavior, specifically the
purchase amount, with respect to the customer's gender and other factors. The objective is to
gain insights that can assist the business in making informed decisions. The team specifically
wants to determine whether there are differences in spending habits between male and female
customers,married unmarried or with different age groups.
```

In [4]:
```
df = pd.read_csv(r'C:\Users\suryawaa\OneDrive - TomTom\2022\Scaler\Walmart_Case\Walmart_data.csv'
```

In [3]:
```
df
```

Out[3]:

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Produ
0	1000001	P00069042	F	0-17	10	A	2	0	
1	1000001	P00248942	F	0-17	10	A	2	0	
2	1000001	P00087842	F	0-17	10	A	2	0	
3	1000001	P00085442	F	0-17	10	A	2	0	
4	1000002	P00285442	M	55+	16	C	4+	0	
...	...	...	...	...	...	...	...	...	
550063	1006033	P00372445	M	51-55	13	B	1	1	
550064	1006035	P00375436	F	26-35	1	C	3	0	
550065	1006036	P00375436	F	26-35	15	B	4+	1	
550066	1006038	P00375436	F	55+	1	C	2	0	
550067	1006039	P00371644	F	46-50	0	B	4+	1	

550068 rows × 10 columns

In [4]:
```
df.shape # has 550068 Rows and 10 columns
```

Out[4]:
```
(550068, 10)
```

In [5]:

```
df.info() # 5 column are having data int type and 5 have object type
```

```
<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
```

In [6]:

```
df.isnull().sum() # No Null values. Dataset is clear to perform further analysis.
```

Out[6]:

```
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
```

statistical summary

In [7]:

```
df.describe().round(2)
```

Out[7]:

	User_ID	Occupation	Marital_Status	Product_Category	Purchase
count	550068.00	550068.00	550068.00	550068.00	550068.00
mean	1003028.84	8.08	0.41	5.40	9263.97
std	1727.59	6.52	0.49	3.94	5023.07
min	1000001.00	0.00	0.00	1.00	12.00
25%	1001516.00	2.00	0.00	1.00	5823.00
50%	1003077.00	7.00	0.00	5.00	8047.00
75%	1004478.00	14.00	1.00	8.00	12054.00
max	1006040.00	20.00	1.00	20.00	23961.00

In [8]:

```
df.describe(include=object)
```

Out[8]:

	Product_ID	Gender	Age	City_Category	Stay_In_Current_City_Years
count	550068	550068	550068	550068	550068
unique	3631	2	7	3	5
top	P00265242	M	26-35	B	1
freq	1880	414259	219587	231173	193821

In [9]:

```
df.nunique() # count of unique values present in each columns
```

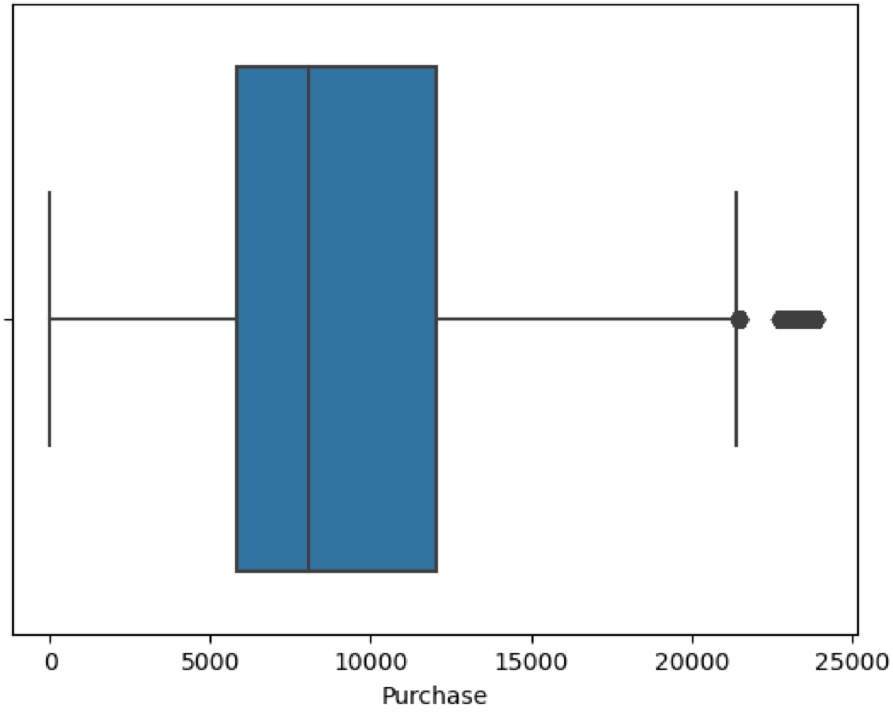
Out[9]:

```
User_ID                        5891
Product_ID                     3631
Gender                            2
Age                               7
Occupation                       21
City_Category                     3
Stay_In_Current_City_Years        5
Marital_Status                    2
Product_Category                 20
Purchase                      18105
dtype: int64
```

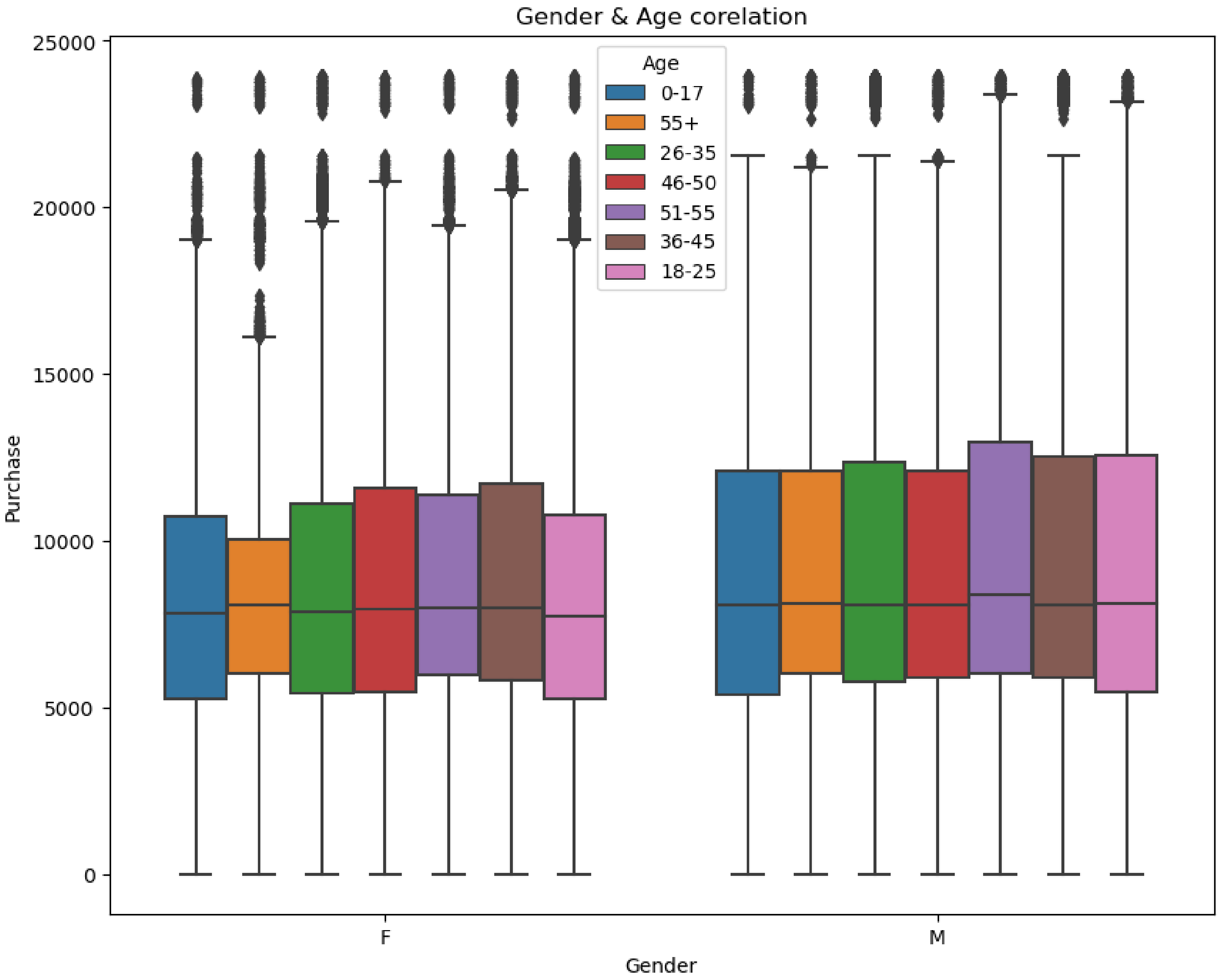
Outlier Detection:

In [10]:

```
sns.boxplot(data=df,x=df['Purchase'])
plt.show()
# Purchases values greater than 21K are less and more number of purchases are below 21K
```



In [11]:
```
plt.figure(figsize=(10,8))
sns.boxplot(x="Gender",y='Purchase',hue='Age',data=df)
plt.title("Gender & Age corelation")
plt.show()
```



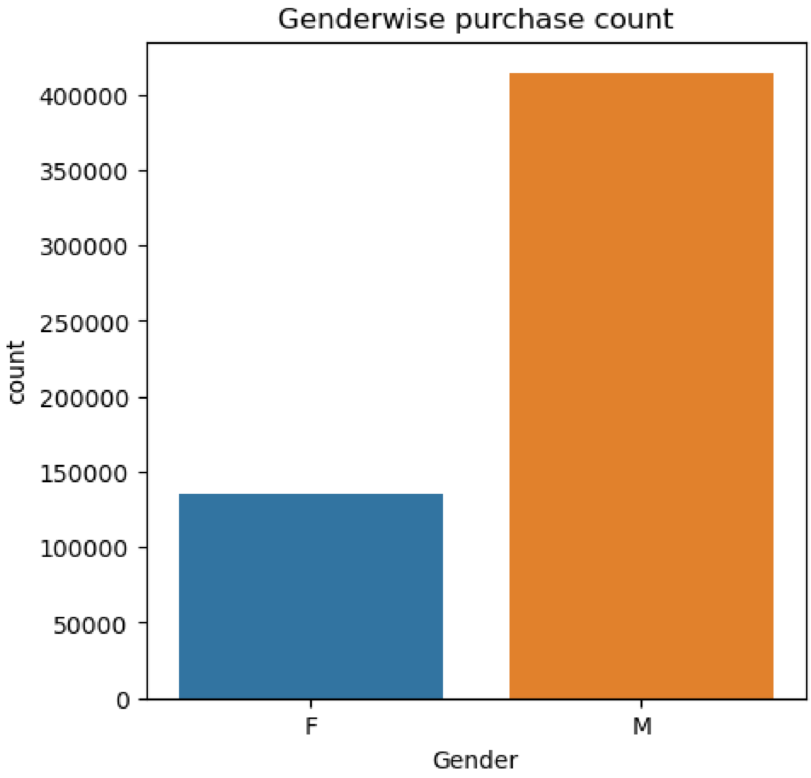
In [12]:
```
# Mosy Purchase values in males is roughly below 21K in overall, however in Age group 51-55 it is
# i.e this group shows more higher amount purchase values
# For female it is below 20K to 18K overall but in 55+ age group it is near to 16 K
# i.e. purchases values above 16K in this Age group is lesser
```

Gender and Distribution:

In [13]:
```
df['Gender'].value_counts() # Count of Male and female records showing male have purchased more tl
```

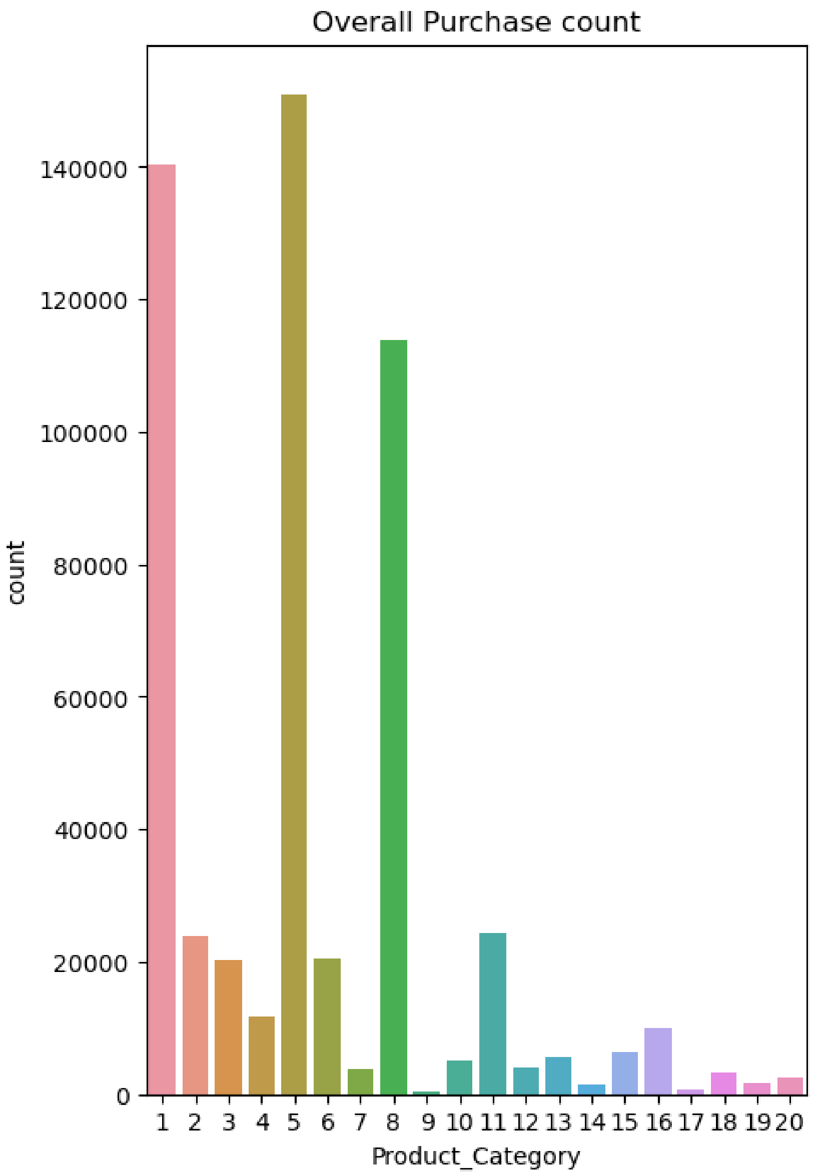
Out[13]:
```
M    414259
F    135809
Name: Gender, dtype: int64
```

In [14]:
```
plt.figure(figsize=(5,5))
sns.countplot(data=df,x='Gender')
plt.title("Genderwise purchase count")
plt.show()
```



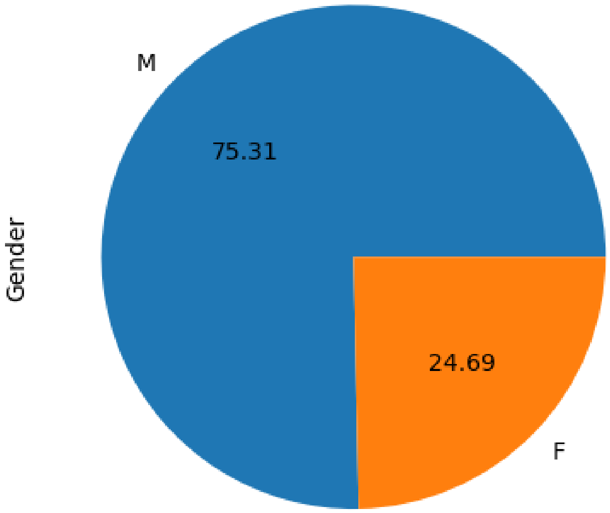
In [15]:
```
# Overall Gender and purchase count distribution
```

In [7]:
```
plt.figure(figsize=(5,8))
sns.countplot(data=df,x='Product_Category')
plt.title("Overall Purchase count")
plt.show()
```



In [ ]:
```
# Product Categories 1,5,8 are the customer's most bought product and rest 17 categories are less
```

```
In [16]: df['Gender'].value_counts().plot(kind='pie',autopct="%.2f")
         plt.show()
```



```
In [17]: # 74.31 % purchases are done by male and 24.69 are done by females
```

```
In [18]: df.groupby('Gender')['Purchase'].mean().round(2)
```

```
Out[18]: Gender
         F    8734.57
         M    9437.53
         Name: Purchase, dtype: float64
```

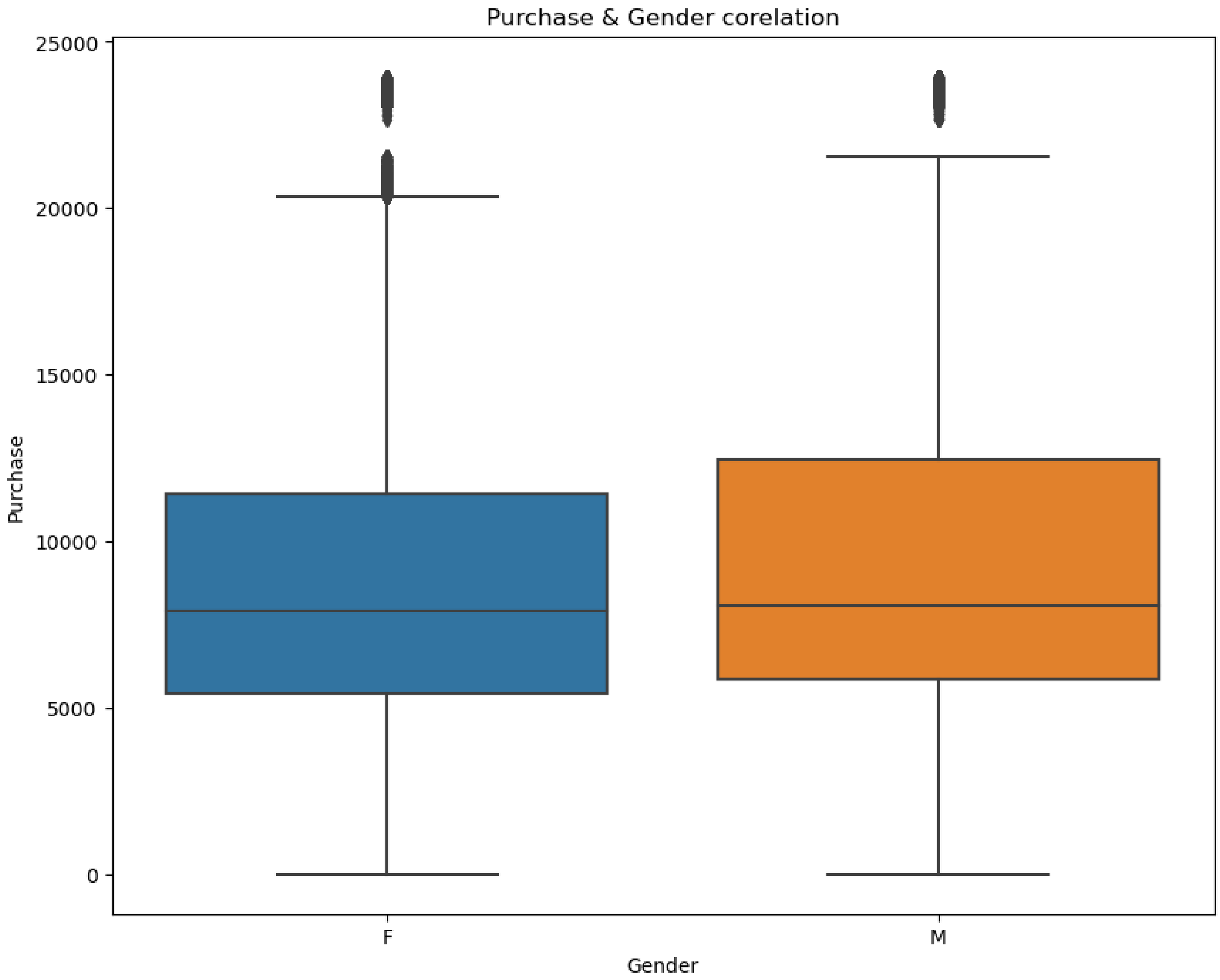
##### Although overall purchase count is higher in males in given data the average purchase of Males is only slightly more than females

```
In [20]: df.groupby('Gender')['User_ID'].nunique() # there are total 1666 unique female customers and 4225
```

```
Out[20]: Gender
         F    1666
         M    4225
         Name: User_ID, dtype: int64
```

In [21]:

```
plt.figure(figsize=(10,8))
sns.boxplot(x='Gender', y='Purchase', data=df)
plt.title("Purchase & Gender corelation")
plt.show()
```



In [22]:

```
# mean of the both Gender are close to each other. Hence we cannot simply say one gender makes mor
```

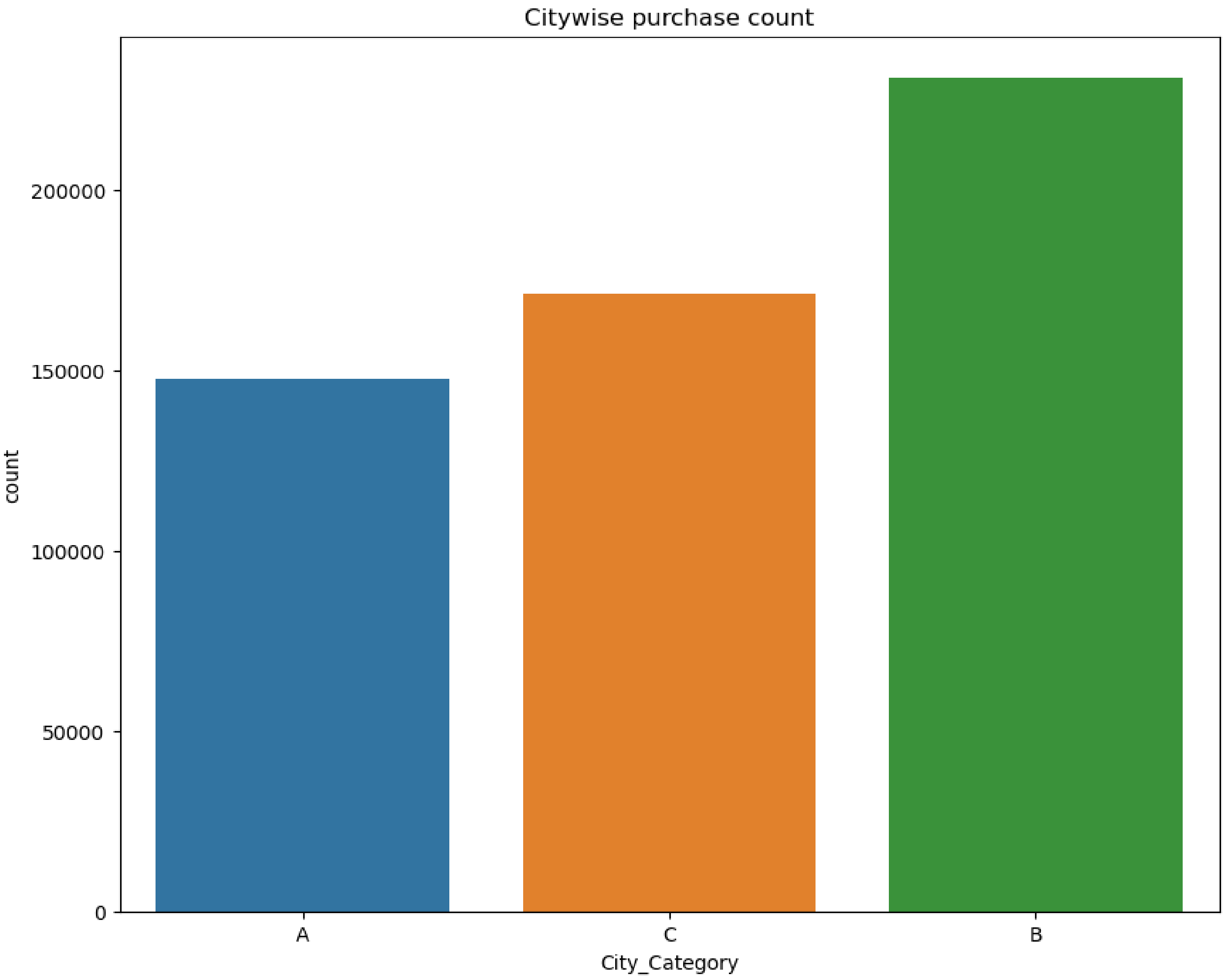
In [23]:

```
df['City_Category'].value_counts()
# Count of 3 city categories shows city catogory B has maximum purchase count while city category
```

Out[23]:

```
B    231173
C    171175
A    147720
Name: City_Category, dtype: int64
```

In [24]:
```
plt.figure(figsize=(10,8))
sns.countplot(data=df,x='City_Category')
plt.title("Citywise purchase count")
plt.show()
```



In [25]:
```
df.groupby('City_Category')['Purchase'].mean().round(2) # Average purchase value is higher in C c
```

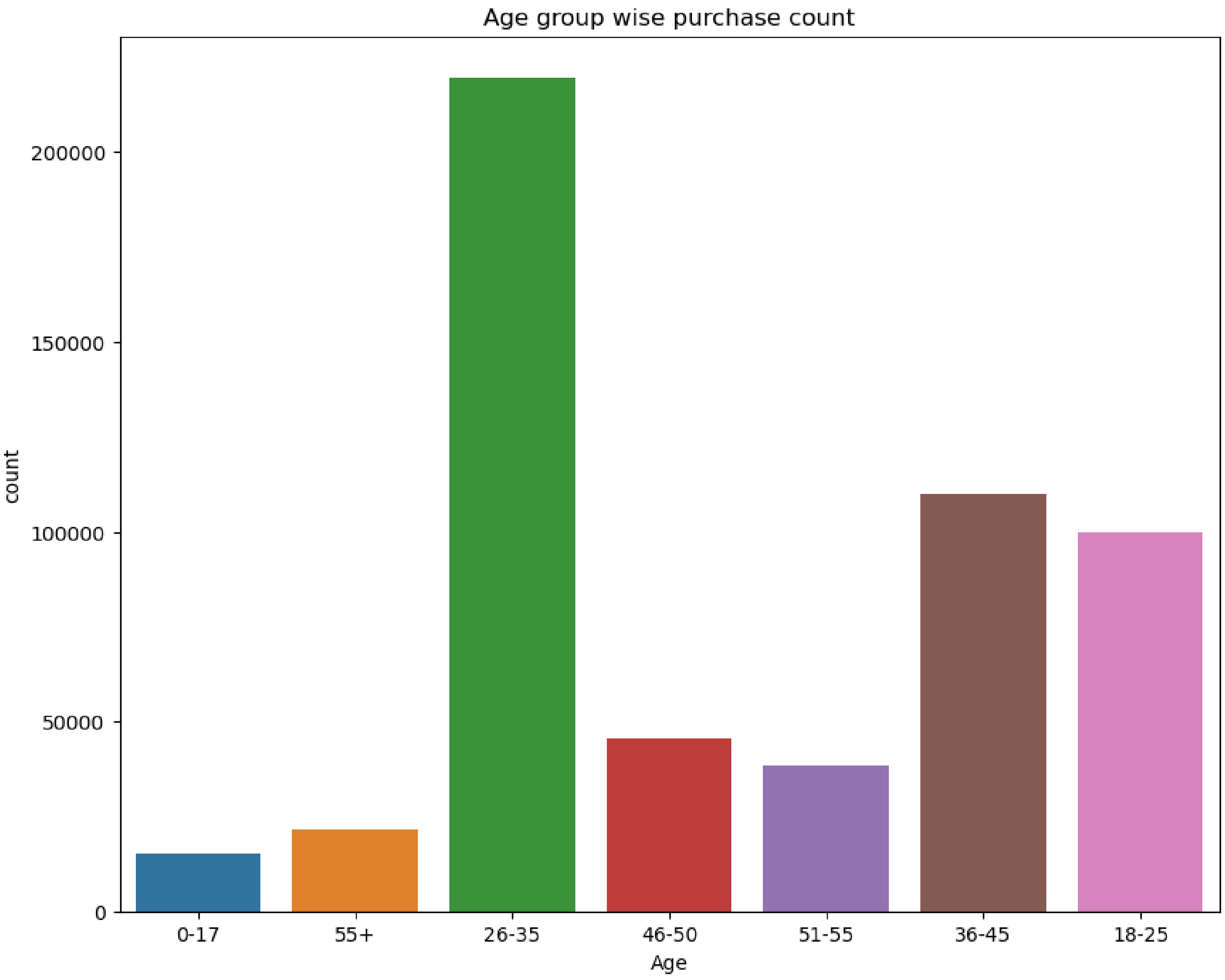
Out[25]:
```
City_Category
A    8911.94
B    9151.30
C    9719.92
Name: Purchase, dtype: float64
```

In [ ]:

In [26]:
```
df['Age'].value_counts()
# Count of 7 different Age group shows that Age group 26-35 makes  way more purchases than any ot
```

Out[26]:
```
26-35    219587
36-45    110013
18-25     99660
46-50     45701
51-55     38501
55+       21504
0-17      15102
Name: Age, dtype: int64
```

In [27]:
```
plt.figure(figsize=(10,8))
sns.countplot(data=df,x='Age')
plt.title("Age group wise purchase count")
plt.show()
```

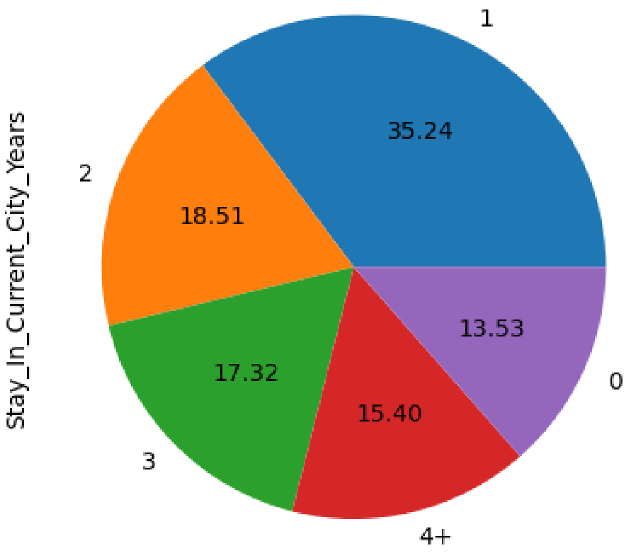


In [28]:
```
df.groupby('Age')['Purchase'].mean().round(2) #
```

Out[28]:
```
Age
0-17     8933.46
18-25    9169.66
26-35    9252.69
36-45    9331.35
46-50    9208.63
51-55    9534.81
55+      9336.28
Name: Purchase, dtype: float64
```

In [29]:
```
# Average of the purchases among all the age group is more or less same overall.
# Age group 51-55 being highest with average purchase value 9534 and 0-17 being lowest with averag
```

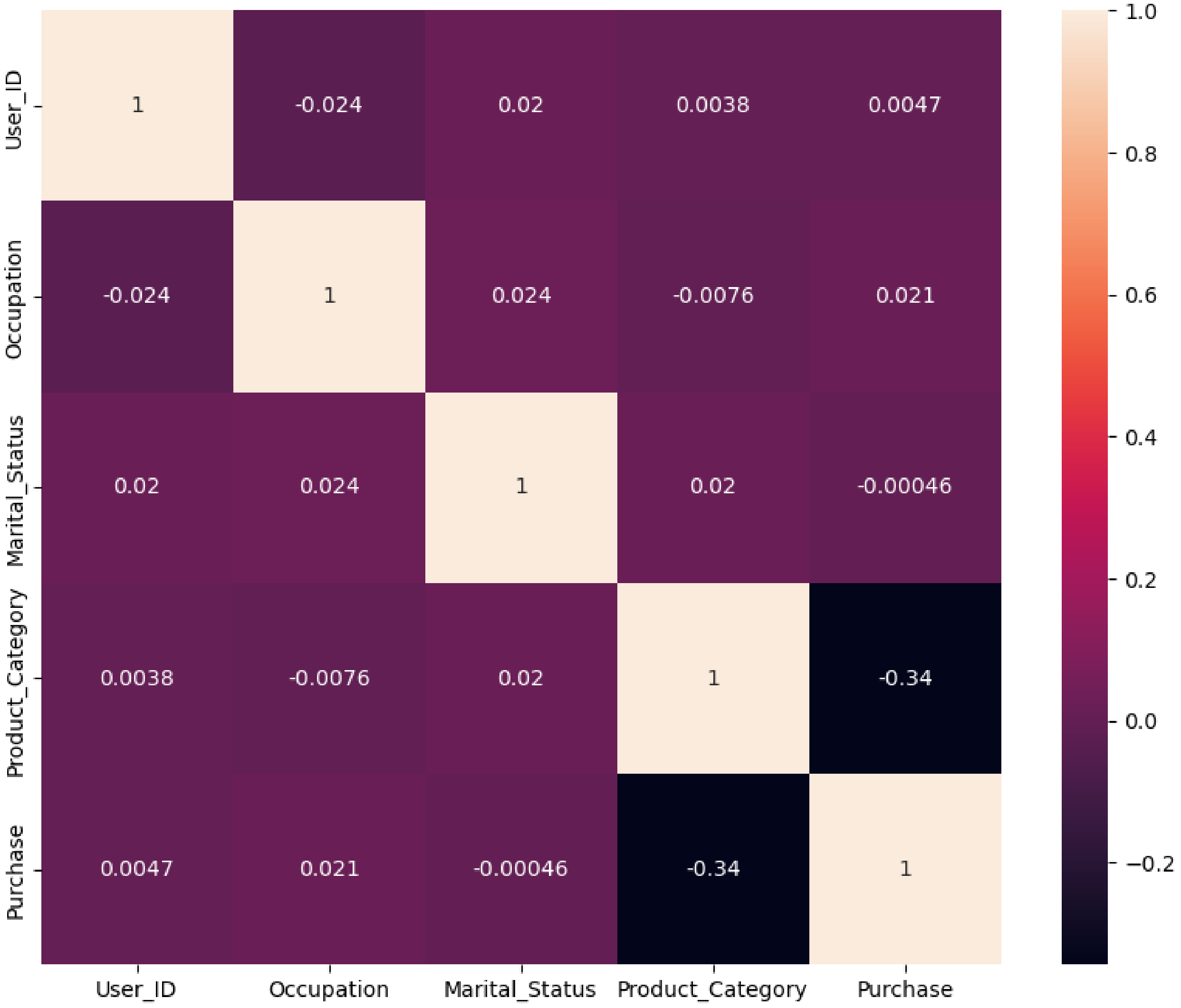
```
In [8]: df['Stay_In_Current_City_Years'].value_counts().plot(kind='pie',autopct="%.2f")
        plt.show()
```



```
In [ ]: # People who are in the city for less than 2 years comprise more number of purchases alltogether.
```

Corelation Analysis with heatmap

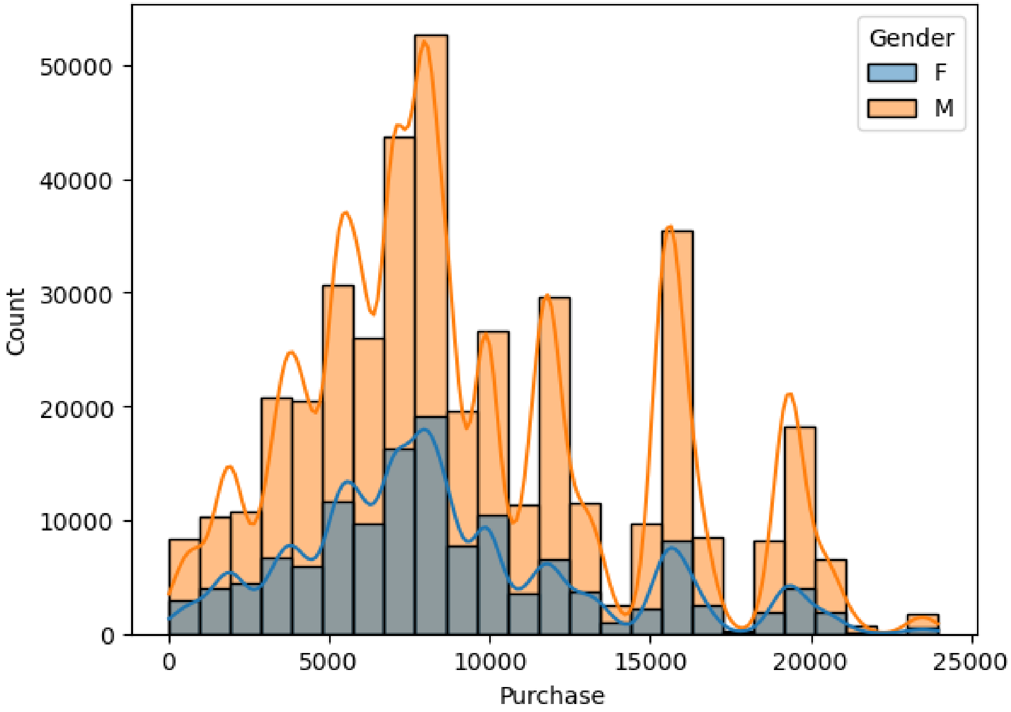
In [30]:
```
plt.figure(figsize=(10,8))
sns.heatmap(df.corr(numeric_only=True),annot=True)
plt.show()
```



In [31]:
```
# All the columns in the dataframe are very weakly corelated
```

In [32]:
```
sns.histplot(x='Purchase',data=df,hue='Gender',bins=25,kde=True)
```

Out[32]:
```
<Axes: xlabel='Purchase', ylabel='Count'>
```



In [33]:
```
# On the basis of count gender wise it is not much inferencing.
```

Gender Wise:

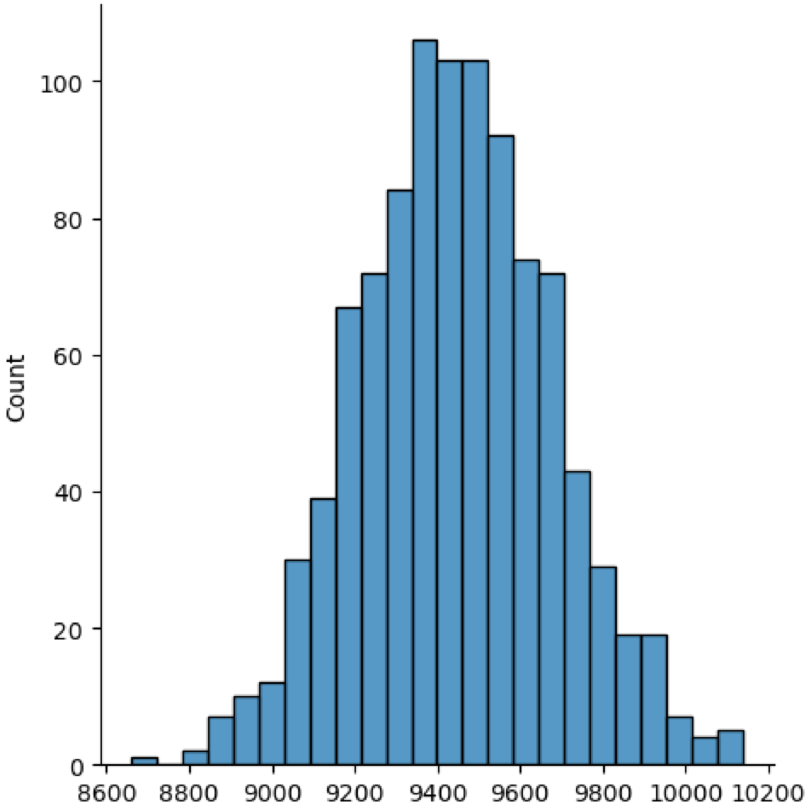
Sampling :

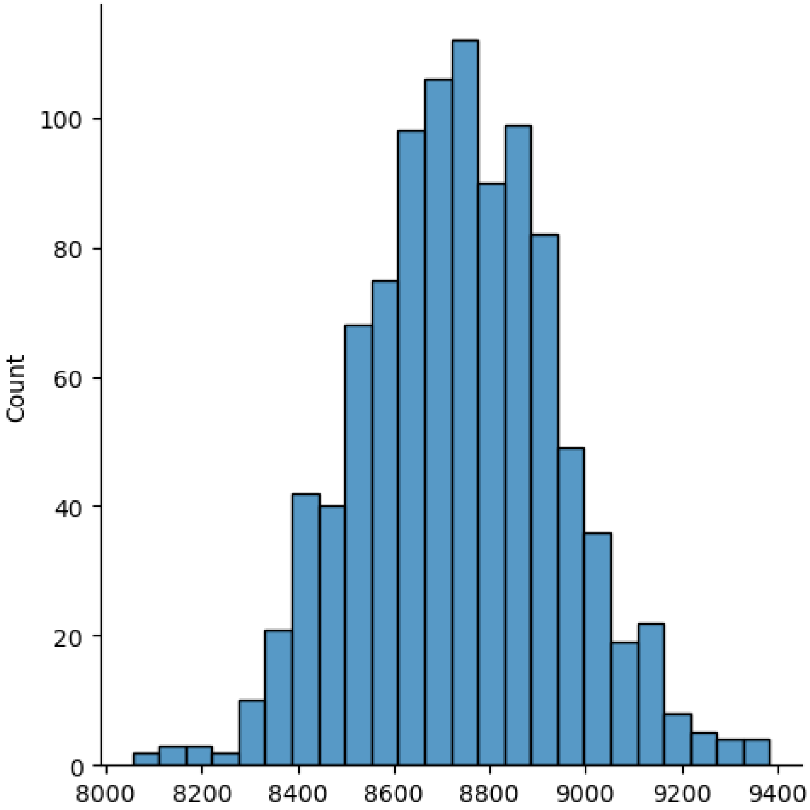
In [34]:
```
n = 500
```

In [35]:
```
male_samp_mean = [df[df['Gender']=='M'].sample(n,replace=True)['Purchase'].mean()for i in range(1
```

In [36]:
```
female_samp_mean = [df[df['Gender']=='F'].sample(n,replace=True)['Purchase'].mean()for i in range
```

In [37]:
```
sns.displot(male_samp_mean)
sns.displot(female_samp_mean)
plt.show()
```





```
In [38]: # Both Male and female sampling shows good symmetric Normal distribution
```

95% Confidence Interval

Males:

```
In [39]: Male_upp_limit = (np.mean(male_samp_mean)+1.96*np.std(male_samp_mean)).round(2)
```

```
In [40]: Male_low_limit = (np.mean(male_samp_mean)-1.96*np.std(male_samp_mean)).round(2)
```

```
In [41]: Male_conf_interval = [Male_low_limit,Male_upp_limit]
```

```
In [42]: Male_conf_interval
```

```
Out[42]: [8986.33, 9900.92]
```

Females:

```
In [43]: Female_upp_limit = (np.mean(female_samp_mean)+1.96*np.std(female_samp_mean)).round(2)
```

```
In [44]: Female_low_limit = (np.mean(female_samp_mean)-1.96*np.std(female_samp_mean)).round(2)
```

```
In [45]: Female_conf_interval = [Female_low_limit,Female_upp_limit]
```

```
In [46]: Female_conf_interval
```

```
Out[46]: [8329.44, 9142.78]
```

```
In [47]: CI_95_percent = [Male_conf_interval,Female_conf_interval]
```

```
In [48]: CI_95_percent
```

```
Out[48]: [[8986.33, 9900.92], [8329.44, 9142.78]]
```

CI at 95 % is overlapping between average male and female spending purchases with sample size of 500, hence lets check with increased sample size

```
In [49]: n = 800
```

```
In [50]: male_samp_mean = [df[df['Gender']=='M'].sample(n,replace=True)['Purchase'].mean()for i in range(1
```

```
In [51]: female_samp_mean = [df[df['Gender']=='F'].sample(n,replace=True)['Purchase'].mean()for i in range
```

Males:

```
In [52]: Male_upp_limit = (np.mean(male_samp_mean)+1.96*np.std(male_samp_mean)).round(2)
In [53]: Male_low_limit = (np.mean(male_samp_mean)-1.96*np.std(male_samp_mean)).round(2)
```

```
In [54]: Male_conf_interval = [Male_low_limit,Male_upp_limit]
```

```
In [55]: Male_conf_interval
```

```
Out[55]: [9087.58, 9812.09]
```

Females:

```
In [56]: Female_upp_limit = (np.mean(female_samp_mean)+1.96*np.std(female_samp_mean)).round(2)
```

```
In [57]: Female_low_limit = (np.mean(female_samp_mean)-1.96*np.std(female_samp_mean)).round(2)
```

```
In [58]: Female_conf_interval = [Female_low_limit,Female_upp_limit]
```

```
In [59]: Female_conf_interval
```

```
Out[59]: [8407.02, 9061.8]
```

95% Confidence Interval with CLT

```
In [60]: CI_95_percent = [Male_conf_interval,Female_conf_interval]
```

```
In [61]: CI_95_percent
```

```
Out[61]: [[9087.58, 9812.09], [8407.02, 9061.8]]
```

90% Confidence Interval with CLT

```
In [62]: male_90_CI = np.percentile(male_samp_mean,[5,95])
         male_90_CI
```

```
Out[62]: array([9141.7461875, 9745.760125 ])
```

```
In [63]: female_90_CI = np.percentile(female_samp_mean,[5,95])
         female_90_CI
```

```
Out[63]: array([8453.3211875, 9004.5326875])
```

99% Confidence Interval with CLT

```
In [64]: male_99_CI = np.percentile(male_samp_mean,[0.5,99.5])
         male_99_CI
```

```
Out[64]: array([9009.992875  , 9916.44220625])
```

```
In [65]: female_99_CI = np.percentile(female_samp_mean,[0.5,99.5])
         female_99_CI
```

```
Out[65]: array([8318.25379375, 9192.5443    ])
```

```
In [66]: # As 99 % shows overlapp its better to go with 95% or 90 % confidence Intervals
```

Age Wise:

```
In [67]: n = 700
```

```
In [68]: samp_mean_0to17 = [df[df['Age']=='0-17'].sample(n,replace=True)['Purchase'].mean()for i in range(
```

```
In [69]: samp_mean_18to25 = [df[df['Age']=='18-25'].sample(n,replace=True)['Purchase'].mean()for i in rang
```

```
In [70]: samp_mean_26to35 = [df[df['Age']=='26-35'].sample(n,replace=True)['Purchase'].mean()for i in rang
```

```
In [71]: samp_mean_36to45 = [df[df['Age']=='36-45'].sample(n,replace=True)['Purchase'].mean()for i in rang
```

```
In [72]: samp_mean_46to50 = [df[df['Age']=='46-50'].sample(n,replace=True)['Purchase'].mean()for i in rang
```

```
In [73]: samp_mean_51to55 = [df[df['Age']=='51-55'].sample(n,replace=True)['Purchase'].mean()for i in rang
```

```
In [74]: samp_mean_55plus = [df[df['Age']=='55+'].sample(n,replace=True)['Purchase'].mean()for i in range(
```

95% Confidence Interval with CLT

```
In [75]: Age_0to17_95_CI = np.percentile(samp_mean_0to17,[2.5,97.5])
         Age_0to17_95_CI

Out[75]: array([8551.96157143, 9318.26571429])
```

```
In [76]: Age_18to25_95_CI = np.percentile(samp_mean_18to25,[2.5,97.5])
         Age_18to25_95_CI

Out[76]: array([8781.89639286, 9562.74785714])
```

```
In [77]: Age_26to35_95_CI = np.percentile(samp_mean_26to35,[2.5,97.5])
         Age_26to35_95_CI

Out[77]: array([8877.19775, 9639.98575])
```

```
In [78]: Age_36to45_95_CI = np.percentile(samp_mean_36to45,[2.5,97.5])
         Age_36to45_95_CI

Out[78]: array([8953.95617857, 9662.95714286])
```

```
In [79]: Age_46to50_95_CI = np.percentile(samp_mean_46to50,[2.5,97.5])
         Age_46to50_95_CI

Out[79]: array([8849.42692857, 9568.87960714])
```

```
In [80]: Age_51to55_95_CI = np.percentile(samp_mean_51to55,[2.5,97.5])
         Age_51to55_95_CI

Out[80]: array([9151.11064286, 9899.2745    ])
```

In [81]:
```
Age_55plus_95_CI = np.percentile(samp_mean_55plus,[2.5,97.5])
Age_55plus_95_CI
```

Out[81]:
```
array([8957.36792857, 9728.77775   ])
```

In [82]:
```
# Above are the Confidence Intervals stating each age group will have average purchase values witl
```

Marital Status Wise:

In [83]:
```
n = 1500
```

In [84]:
```
samp_mean_married = [df[df['Marital_Status']==1].sample(n,replace=True)['Purchase'].mean()for i i
```

In [85]:
```
samp_mean_unmarried = [df[df['Marital_Status']==0].sample(n,replace=True)['Purchase'].mean()for i
```

90% Confidence Interval with CLT

In [88]:
```
married_90_CI = np.percentile(samp_mean_married,[5,95])
married_90_CI
```

Out[88]:
```
array([9048.76356667, 9472.48196667])
```

In [89]:
```
unmarried_90_CI = np.percentile(samp_mean_unmarried,[5,95])
unmarried_90_CI
```

Out[89]:
```
array([9052.77673333, 9478.70716667])
```

Even with 90% confidence interval and with bigger sample size of 1500 with 10K iterations we can get confidence interval for married custormer = [9048.76356667, 9472.48196667] and for unmarried customer = [9052.77673333, 9478.70716667]

```
# Buisness Insights:
* The purchase amount distribution in the dataset is skewed, with a majority of purchases being
below $21,000.

* Male customers generally have purchase values below $21,000, except for the 51-55 age group
where it can go up to $23,000.

* Female customers tend to have purchase values below $20,000 to $18,000 overall, but in the 55+
age group, it is closer to $16,000.

* The count of male customers making purchases is higher than female customers.

* Approximately 74.31% of purchases are made by males, while females account for 24.69%.

* The average purchase amount for males is slightly higher than females, despite the higher
count of male customers.

* There are 1,666 unique female customers and 4,225 unique male customers who made purchases.

* The distribution of purchases across different city categories shows that Category B has the
highest purchase count, while Category A has the lowest.

* Product Categories 1,5,8 are the customer's most bought product and rest 17 categories are
less likely purchased.

* People who are in the city for less than 2 years comprise more number of purchases
alltogether. Natives are less likely to purchase.
```

* The average purchase value is higher in Category C cities.

* The age group 26-35 has the highest purchase count compared to other age groups.

* The average purchase amount is similar across all age groups, with the 51-55 age group having the highest average value
  ($9,534) and the 0-17 age group having the lowest ($8,933).

* There is no strong correlation between any of the columns in the dataset.

* The gender-wise count does not provide significant insights.

* Both male and female samples exhibit a good symmetric normal distribution, indicating a balanced distribution of purchase
  amounts within each gender.

* The 95% confidence interval for average male spending is [8986.33, 9900.92] with a sample size of 500, while the 95%
  confidence interval for average female spending is [8329.44, 9142.78] with the same sample size.

* The 95% confidence intervals of average male and female spending overlap, suggesting that there may not be a significant
  difference in spending habits between genders based on the given sample size.

* To further investigate, the analysis is repeated with an increased sample size of 800. The 95% confidence interval for males
  is [9087.58, 9812.09], and for females, it is [8407.02, 9061.8].

* Additionally, the analysis is performed with a 90% confidence interval, resulting in a male confidence interval of [9141.7461875, 9745.760125] and a female confidence interval of [8453.3211875, 9004.5326875].

* A 99% confidence interval is also calculated, yielding a male interval of [9009.992875, 9916.44220625] and a female interval of [8318.25379375, 9192.5443].

* Confidence intervals for different age groups at a 95% confidence level are calculated. The intervals are as follows:
Age group 0-17: [8551.96157143, 9318.26571429]
Age group 18-25: [8781.89639286, 9562.74785714]
Age group 26-35: [8877.19775, 9639.98575]
Age group 36-45: [8953.95617857, 9662.95714286]
Age group 46-50: [8849.42692857, 9568.87960714]
Age group 51-55: [9151.11064286, 9899.2745]
Age group 55+: [8957.36792857, 9728.77775]

* With a larger sample size of 1500 and a 90% confidence interval, the confidence interval for married customers is [9048.76356667, 9472.48196667], while for unmarried customers, it is [9052.77673333, 9478.70716667]

# Buisness Recommendation:

* Conducted statistical analysis suggests with 95% confidence ,that the mean purchase amounts between males is 8986 to 9900 and female customers spending amount is between 8329 to 9142 stating that men spend more money per transaction than women.

* Overlapping confidence intervals of customers from different Age group and with marital statuses are observed suggesting that there may not be a significant difference in spending habits between different Age group & marital status. This implies that such specific marketing strategies may not be necessary, and efforts can be focused on broader customer segments.

* Company can tailor marketing strategies and promotions based on age-specific patterns identified in the analysis to better cater to the needs and preferences of different age groups like , it is observed that Age groups with greater the age are more likely to make big purchases and with young age customers having comparitively less average spending due to may be less income , hence walmart can provide some systematic plans to make payments in installments if customer wishes to.

* Age group 26-35 shows more overall purchase count may be due to stable income, hence other age groups like below 17 , 46-50, 55+, 51-55 can be targeted by marketing and providing specific installment plans to increase their purchase count.

* Cities where average spending is low there can be launched some marketing strategies to attract the customers. And further analysis on their.

* Customers who are native or living more than 3 to 4 yrs in the city can be targeted to increase sell as they are less likely to make purchases.

* As per product categories more marketing strategies can be formed to increase sale of product categories other than 1,5 & 8.

* Validate the findings and consider the distribution and relationship between variables when generalizing the analysis to the overall population. Ensure the findings are representative of the larger customer base.

* Consider conducting additional surveys or targeted studies to gather more comprehensive data on customer spending habits, or income range details can also add more value in future enabling more accurate insights.

* Considering external factors such as cultural and regional differences may influence spending habits additional research can be conducted to generalize the results to the larger population