In [ ]:
```
Name : Mahesh Darakhe

Batch : Nov22 beginner DSML
```

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

## Business Problem

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?
(Assume 50 million customers are male and 50 million are female).

In [3]:
```
df=pd.read_csv('https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/293/original/walmart_data.csv?16412
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
```

In [6]:
```
df.nunique()
```

Out[6]:
```
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
```

In [7]:
```
df.describe(include='all')
```

Out[7]:

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category
count	5.500680e+05	550068	550068	550068	550068.000000	550068	550068	550068.000000	550068.000000
unique	NaN	3631	2	7	NaN	3	5	NaN	NaN
top	NaN	P00265242	M	26-35	NaN	B	1	NaN	NaN
freq	NaN	1880	414259	219587	NaN	231173	193821	NaN	NaN
mean	1.003029e+06	NaN	NaN	NaN	8.076707	NaN	NaN	0.409653	5.404270
std	1.727592e+03	NaN	NaN	NaN	6.522660	NaN	NaN	0.491770	3.936211
min	1.000001e+06	NaN	NaN	NaN	0.000000	NaN	NaN	0.000000	1.000000
25%	1.001516e+06	NaN	NaN	NaN	2.000000	NaN	NaN	0.000000	1.000000
50%	1.003077e+06	NaN	NaN	NaN	7.000000	NaN	NaN	0.000000	5.000000
75%	1.004478e+06	NaN	NaN	NaN	14.000000	NaN	NaN	1.000000	8.000000
max	1.006040e+06	NaN	NaN	NaN	20.000000	NaN	NaN	1.000000	20.000000

## Initial Observations:

1. There are no missing values in the data.

2. There are 3631 unique product IDs in the dataset. P00265242 is the most sold Product ID.

3. There are 7 unique age groups and most of the purchase belongs to age 26-35 group.

4. There are 3 unique citi categories with category B being the highest.

5. 5 unique values for Stay_in_current_citi_years with 1 being the highest.

6. The difference between mean and median seems to be significant for purchase that suggests outliers in the data.

7. Minimum & Maximum purchase is 12 and 23961 suggests the purchasing behaviour is quite spread over a aignificant range of values. Mean is 9264 and 75% of purchase is of less than or equal to 12054. It suggest most of the purchase is not more than 12k.

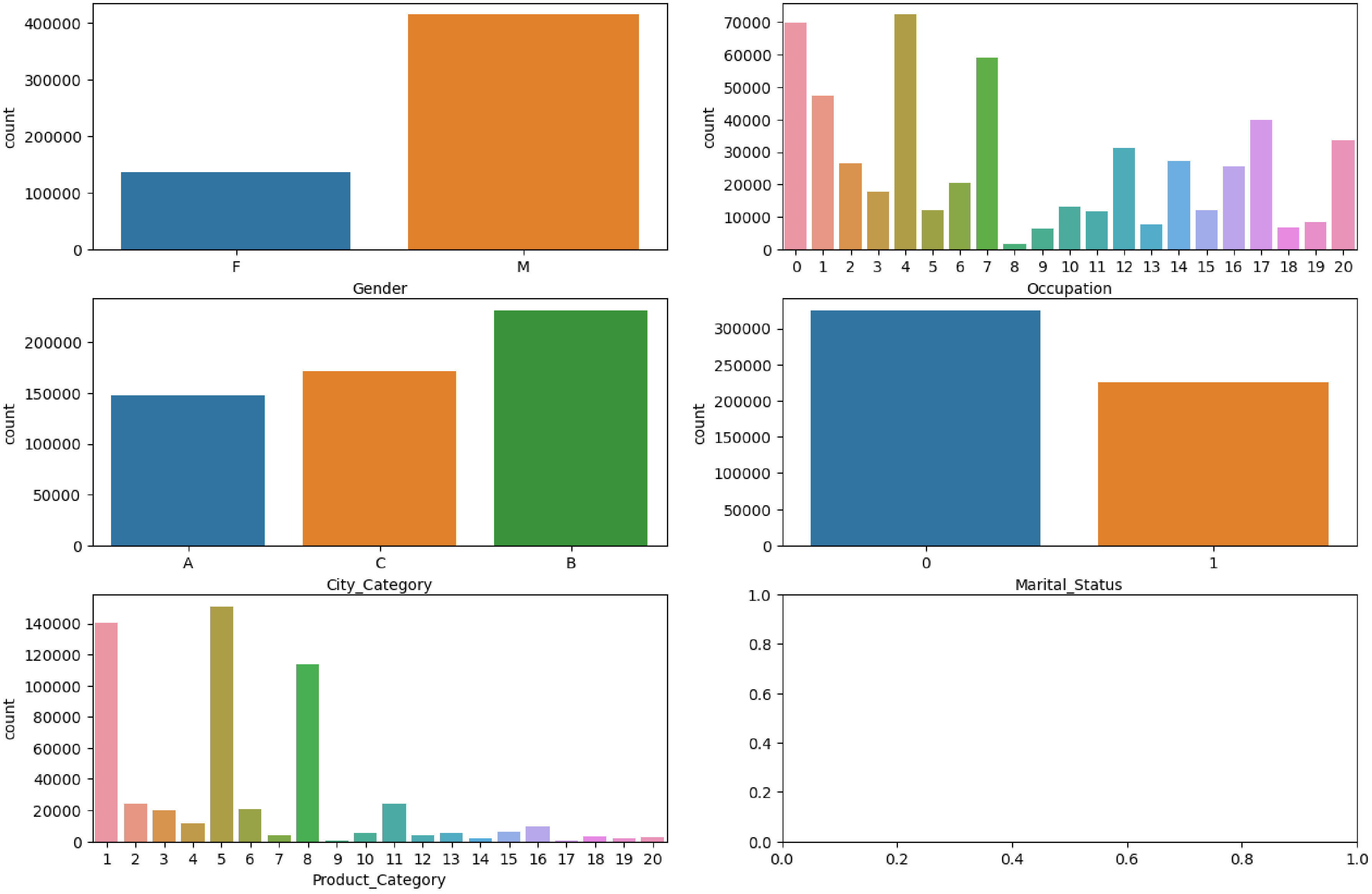
8. Few categorical variable are of integer data type. It can be converted to character type.

9. Out of 550068 data points, 414259 are Male and rest are the female. Male purchase count is much higher than female.

10. Standard deviation for purchase have significant value which suggests data is more spread out for this attribute.

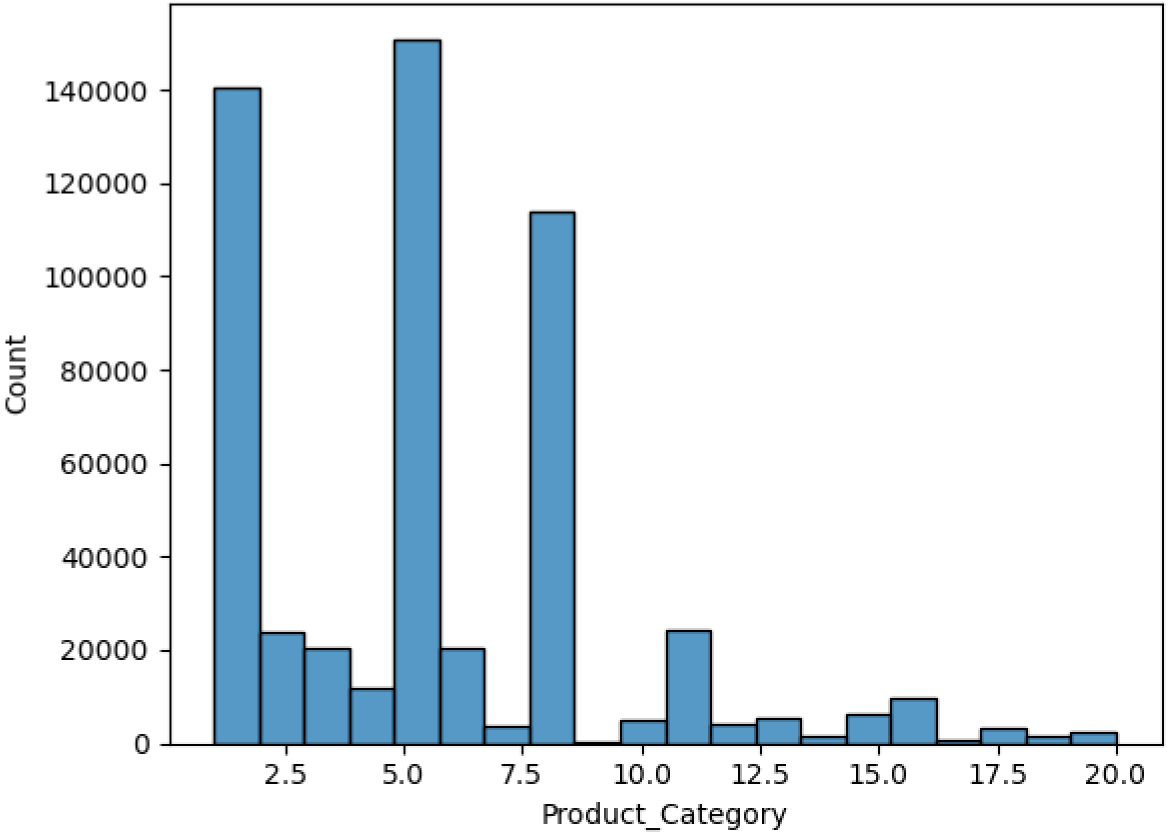
In [9]:

```
fig, axs = plt.subplots(nrows=3, ncols=2, figsize=(15, 10))
sns.countplot(data=df, x='Gender', ax=axs[0,0])
sns.countplot(data=df, x='Occupation', ax=axs[0,1])
sns.countplot(data=df, x='City_Category', ax=axs[1,0])
sns.countplot(data=df, x='Marital_Status', ax=axs[1,1])
sns.countplot(data=df, x='Product_Category', ax=axs[2,0])
plt.show()
```


count
0
100000
200000
300000
400000
F
M
Gender
count
0
10000
20000
30000
40000
50000
60000
70000
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Occupation
count
0
50000
100000
150000
200000
A
C
B
City_Category
count
0
50000
100000
150000
200000
250000
300000
0
1
Marital_Status
count
0
20000
40000
60000
80000
100000
120000
140000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Product_Category
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

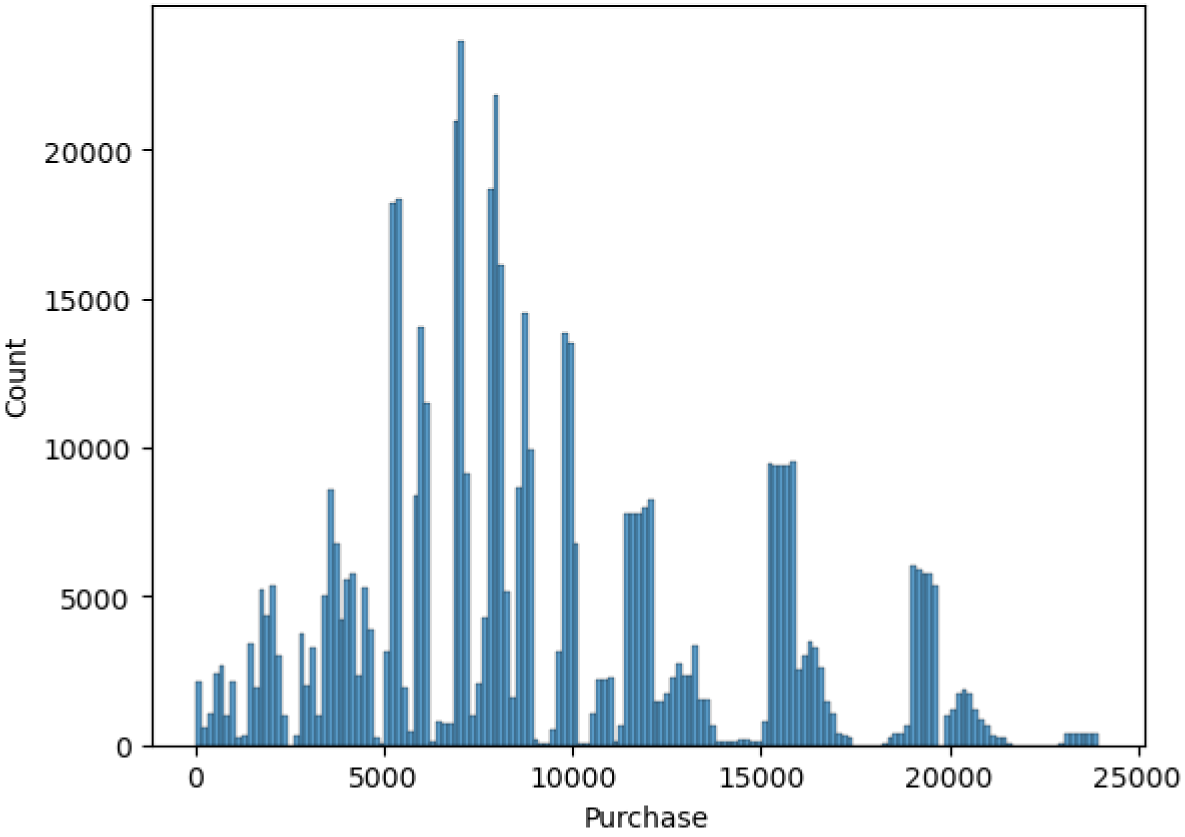
```
In [10]: sns.histplot(x='Product_Category', data=df,bins=20)
```

```
Out[10]: <Axes: xlabel='Product_Category', ylabel='Count'>
```



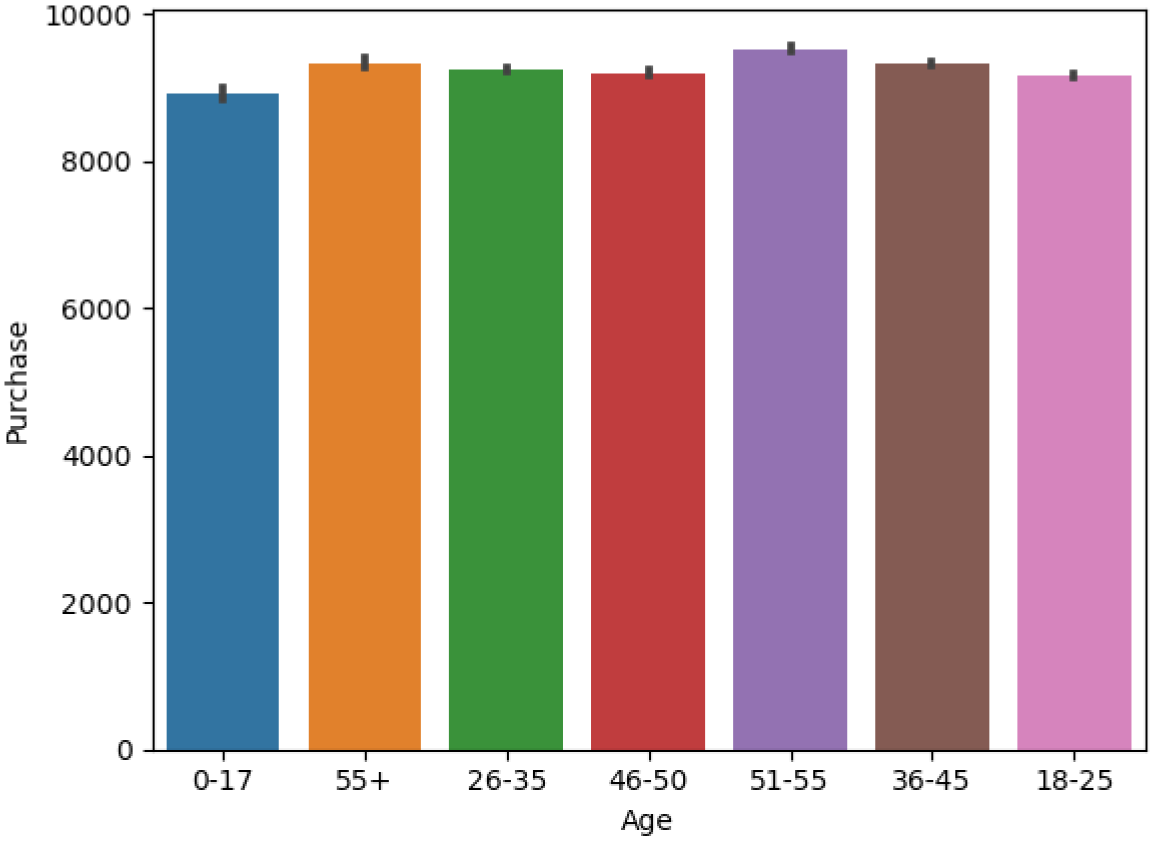
In [11]:
```
sns.histplot(x='Purchase',data=df)
```

Out[11]:
```
<Axes: xlabel='Purchase', ylabel='Count'>
```



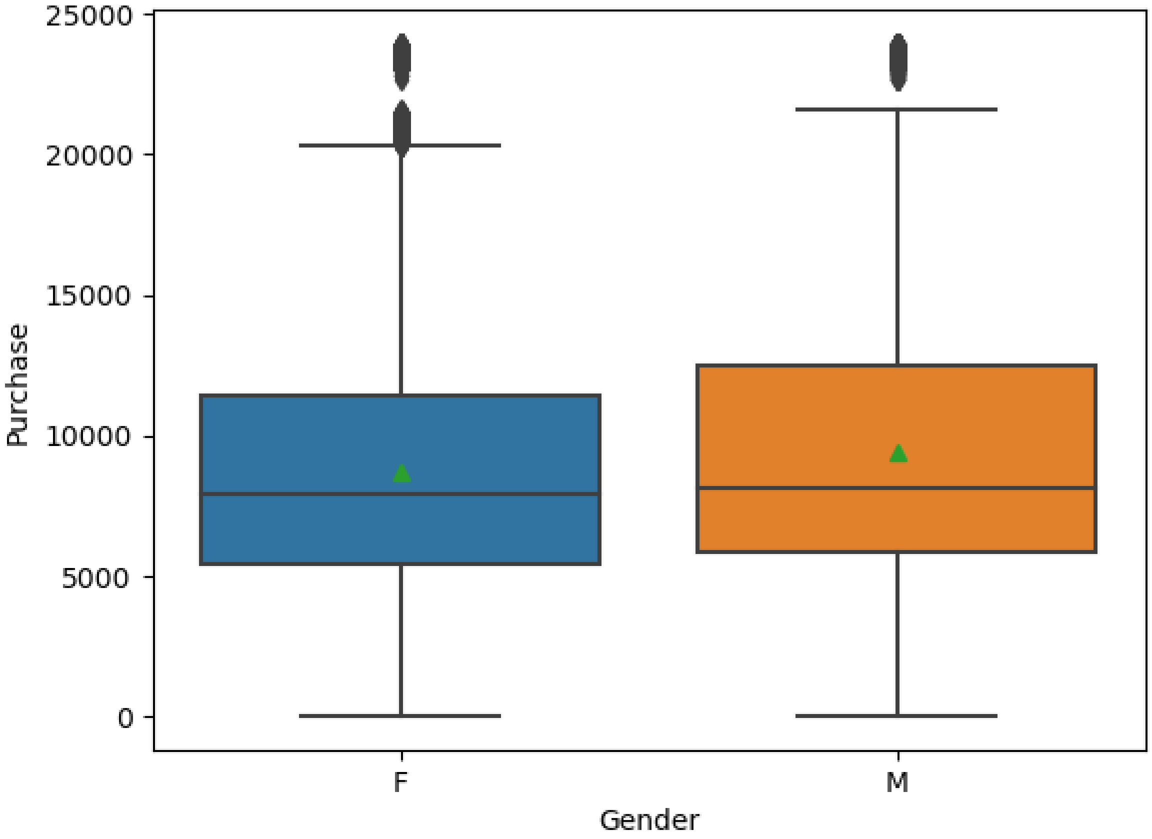
In [12]:
```
sns.barplot(df,x='Age',y='Purchase',estimator='mean')
```

Out[12]:
```
<Axes: xlabel='Age', ylabel='Purchase'>
```



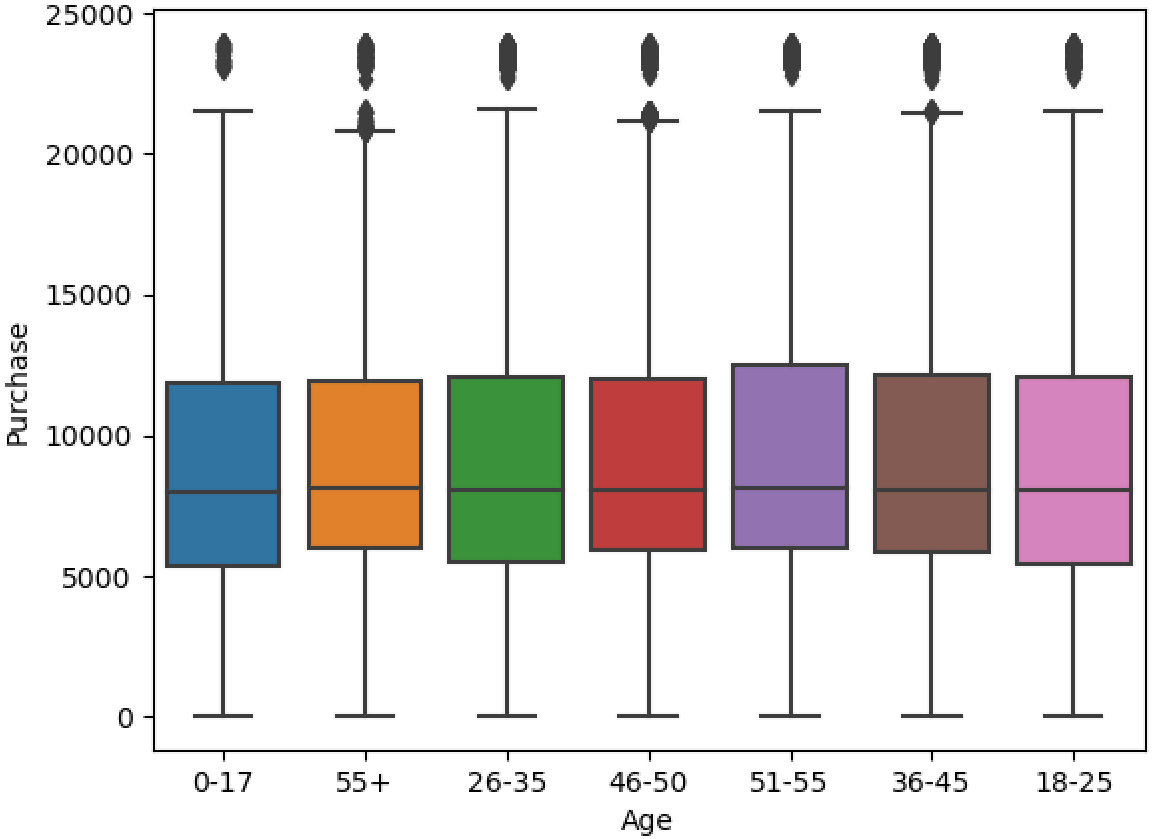
```
In [13]: sns.boxplot( x='Gender',y='Purchase', data=df,showmeans=True)
```

```
Out[13]: <Axes: xlabel='Gender', ylabel='Purchase'>
```

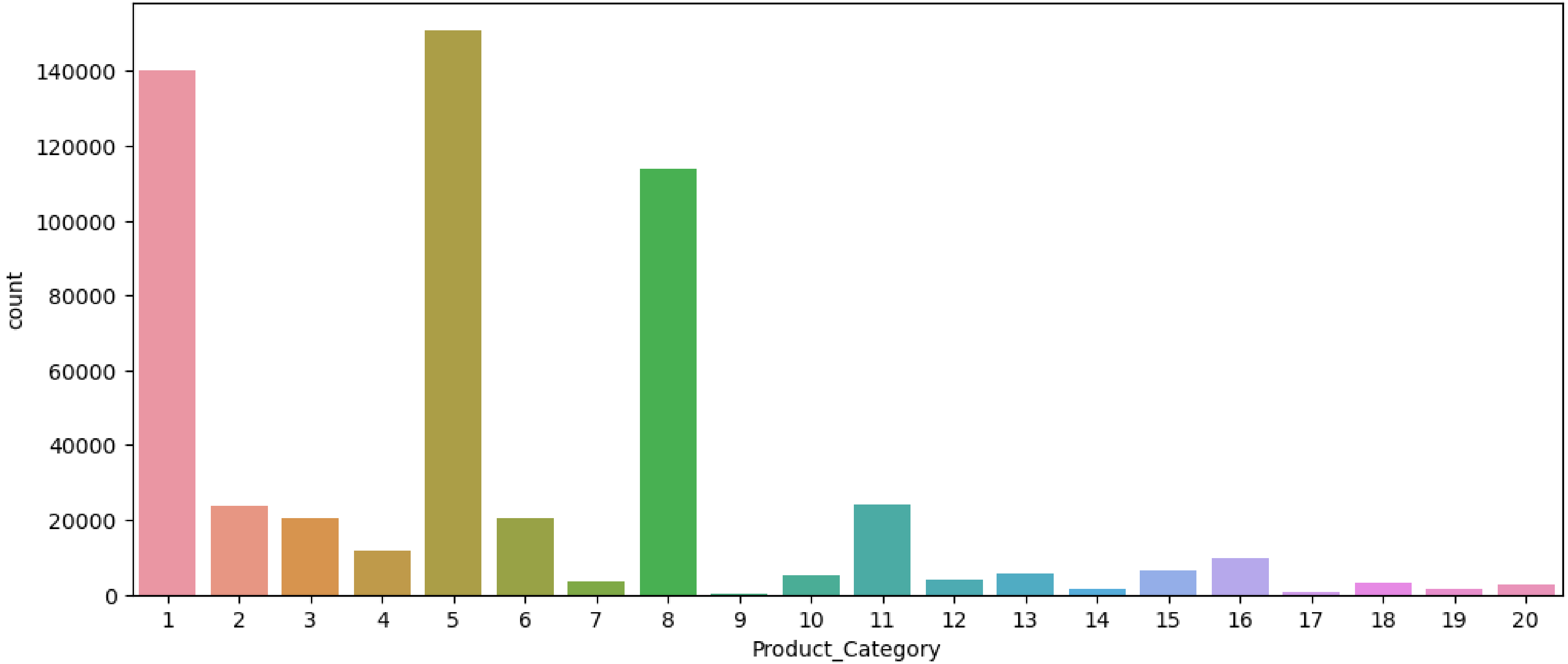


In [14]:
```
sns.boxplot(df,x='Age',y='Purchase')
```

Out[14]:
```
<Axes: xlabel='Age', ylabel='Purchase'>
```



In [15]:
```
plt.figure(figsize=(12, 5))
sns.countplot(data=df, x='Product_Category')
plt.show()
```



### Observations:

1. There are 20 product categories with product category 1, 5 and 8 having higher purchasing frequency.

2. Outliers are present in Purchase catagory.

3. We can clearly see from the graphs above the purchases done by males are much higher than females.

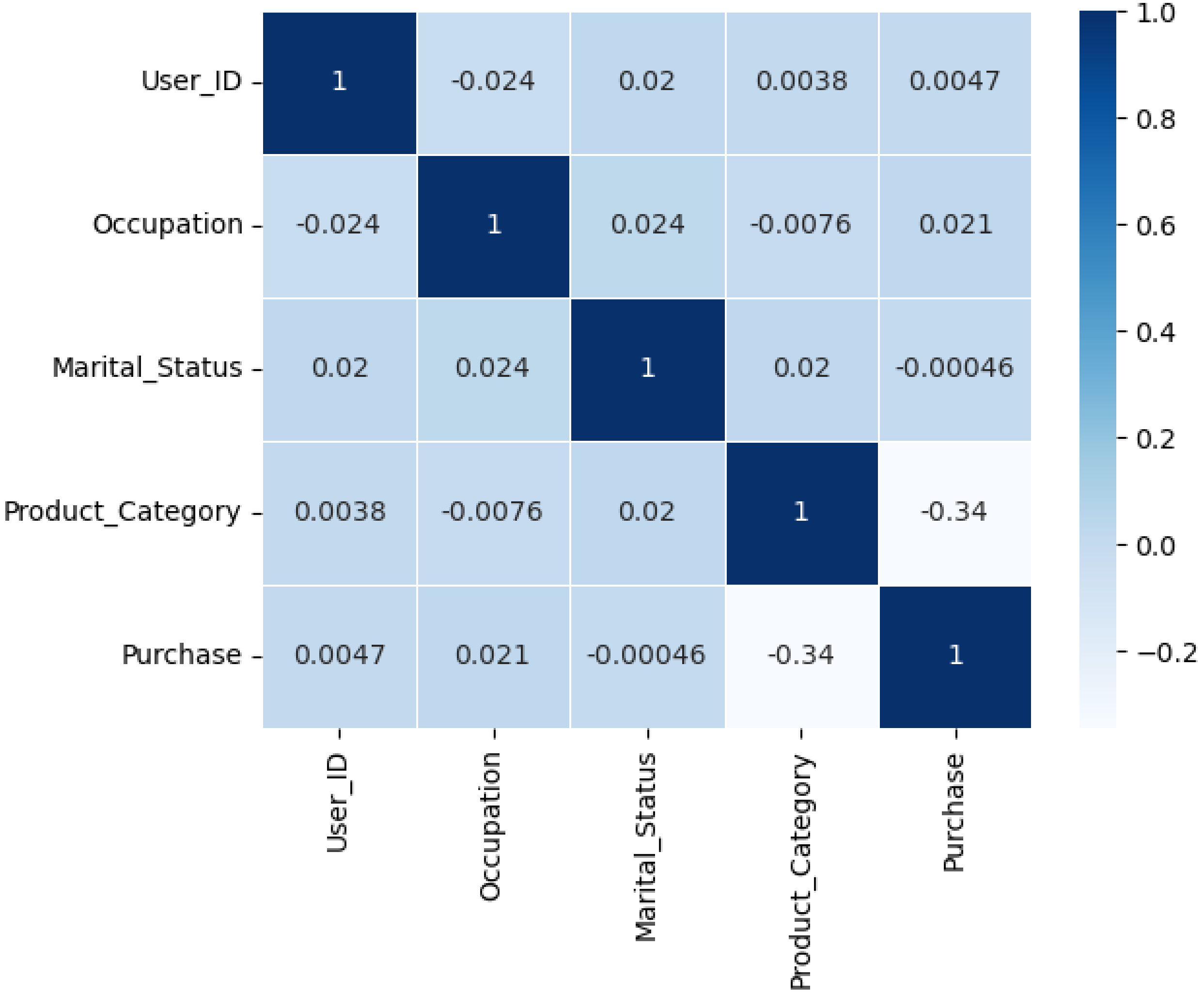
4. We have 21 occupations categories. Occupation category 4, 0, and 7 are with higher number of purchases and category 8 with the lowest number of purchaes.

5. The purchases are highest from City category B.

```
6. Single customer purchases are higher than married users.
```

In [4]:
```
sns.heatmap(df.corr(), annot=True, cmap="Blues", linewidth=.5)
plt.show()
```

```
C:\Users\Mahesh\AppData\Local\Temp\ipykernel_2912\3470554049.py:1: FutureWarning: The default value of numeric_only i
n DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify t
he value of numeric_only to silence this warning.
  sns.heatmap(df.corr(), annot=True, cmap="Blues", linewidth=.5)
```



In [38]:
```
n=200
n_trials=1000
bs_total_mean=np.random.choice(df['Purchase'],size=(n_trials,n),replace=True).mean(axis=1).mean()
bs_total_std=np.random.choice(df['Purchase'],size=(n_trials,n),replace=True).mean(axis=1).std()
bs_total_mean,bs_total_std
```

Out[38]:
```
(9249.918265, 351.6725619861707)
```

In [41]:
```
# 95% Confidence Interval for total population
se=(bs_total_std/np.sqrt(n))
print(f'Interval within which the 95% average amount spent for total population will lie between{bs_total_mean-(se*1.9

#99% CI
print(f'Interval within which the 99% average amount spent for total population will lie between{bs_total_mean-(se*2.5
```

```
Interval within which the 95% average amount spent for total population will lie between9201.178934545818 & 9298.6575
95454182
Interval within which the 99% average amount spent for total population will lie between9186.01006129222 & 9313.82646
870778
```

In [33]:
```
#Confidence Interval for male population
n=200
n_trials=1000
bs_male_mean=np.random.choice(df[df['Gender']=='M']['Purchase'],size=(n_trials,n),replace=True).mean(axis=1).mean()
bs_male_std=np.random.choice(df[df['Gender']=='M']['Purchase'],size=(n_trials,n),replace=True).mean(axis=1).std()
bs_male_mean,bs_male_std
```

Out[33]:
```
(9420.007515000001, 367.0129719765708)
```

In [48]:

```
#95% CI
se1=(bs_male_std/np.sqrt(n))
print(f'Interval within which the 95% average amount spent for male population will lie between {bs_male_mean-(se1*1.9

#99% CI
print(f'Interval within which the 99% average amount spent for male population will lie between {bs_male_mean-(se1*2.5
```

```
Interval within which the 95% average amount spent for male population will lie between 9369.142112191461 & 9470.8729
17808541
Interval within which the 99% average amount spent for male population will lie between 9353.31155315411 & 9486.70347
6845893
```

In [49]:

```
#Confidence Interval for female population
n=200
n_trials=1000
bs_female_mean=np.random.choice(df[df['Gender']=='F']['Purchase'],size=(n_trials,n),replace=True).mean(axis=1).mean()
bs_female_std=np.random.choice(df[df['Gender']=='F']['Purchase'],size=(n_trials,n),replace=True).mean(axis=1).std()
bs_female_mean,bs_female_std
```

Out[49]:

```
(8715.85499, 347.45592301501074)
```

In [50]:

```
#95% CI
se2=(bs_female_std/np.sqrt(n))
print(f'Interval within which the 95% average amount spent for female population will lie between {bs_female_mean-(se2

#99% CI
print(f'Interval within which the 99% average amount spent for female population will lie between {bs_female_mean-(se2
```

```
Interval within which the 95% average amount spent for female population will lie between 8667.70005589184 & 8764.009
924108159
Interval within which the 99% average amount spent for female population will lie between 8652.713061092873 & 8778.99
6918907127
```

there is a significant difference in spending between male and female customers. In this case explore strategies to target each group more effectively. For example, they might design gender-specific marketing campaigns, tailor product offerings, or adjust store layouts to better cater to the preferences of each group.

In [58]:

```
#Confidence Interval for married
n=200
n_trials=1000
bs_married_mean=np.random.choice(df[df['Marital_Status']==1]['Purchase'],size=(n_trials,n),replace=True).mean(axis=1).
bs_married_std=np.random.choice(df[df['Marital_Status']==1]['Purchase'],size=(n_trials,n),replace=True).mean(axis=1).s
bs_married_mean,bs_married_std
```

Out[58]:

```
(9254.759184999999, 343.40684789532804)
```

In [53]:

```
#95% CI
se_m=(bs_married_std/np.sqrt(n))
print(f'Interval within which the 95% average amount spent for married population will lie between {bs_married_mean-(s

#99% CI
print(f'Interval within which the 99% average amount spent for married population will lie between {bs_married_mean-(s
```

```
Interval within which the 95% average amount spent for married population will lie between 9210.171267885962 & 9306.8
2258211404
Interval within which the 99% average amount spent for married population will lie between 9195.131139906593 & 9321.8
6271009341
```

In [57]:

```
#Confidence Interval for unmarried
n=200
n_trials=1000
bs_unmarried_mean=np.random.choice(df[df['Marital_Status']==0]['Purchase'],size=(n_trials,n),replace=True).mean(axis=1
bs_unmarried_std=np.random.choice(df[df['Marital_Status']==0]['Purchase'],size=(n_trials,n),replace=True).mean(axis=1)
bs_unmarried_mean,bs_unmarried_std
```

Out[57]:

```
(9275.50427, 348.3390422394221)
```

In [56]:

```python
#95% CI
se_unm=(bs_unmarried_std/np.sqrt(n))
print(f'Interval within which the 95% average amount spent for married population will lie between{bs_unmarried_mean-(

#99% CI
print(f'Interval within which the 99% average amount spent for married population will lie between{bs_unmarried_mean-(
```

```
Interval within which the 95% average amount spent for married population will lie between9217.77448636021 & 9318.012
303639789
Interval within which the 99% average amount spent for married population will lie between9202.176254589664 & 9333.61
0535410335
```

In [5]:

```python
age_intervals = ['0-17', '26-35', '36-45', '18-25', '46-50', '51-55', '55+']
m_and_s=[]
n=200
n_trials=1000
for i in age_intervals:
    bs_age_mean=np.random.choice(df[df['Age']==i]['Purchase'],size=(n_trials,n),replace=True).mean(axis=1).mean()
    bs_age_std=np.random.choice(df[df['Age']==i]['Purchase'],size=(n_trials,n),replace=True).mean(axis=1).std()
    m_and_s.append([bs_age_mean,bs_age_std])
```

In [75]:

```
#95% CI for age groups

#for 0-17 age group
sample_se=(m_and_s[0][1]/np.sqrt(n))
print(f'Confidence Intervals for 0-17 age group are {m_and_s[0][0]-1.96*sample_se} & {m_and_s[0][0]+1.96*sample_se}')

#for 26-35 age group
sample_se=(m_and_s[1][1]/np.sqrt(n))
print(f'Confidence Intervals for 26-35 age group are {m_and_s[1][0]-1.96*sample_se} & {m_and_s[1][0]+1.96*sample_se}')

#for 36-45 age group
sample_se=(m_and_s[2][1]/np.sqrt(n))
print(f'Confidence Intervals for 36-45 age group are {m_and_s[2][0]-1.96*sample_se} & {m_and_s[2][0]+1.96*sample_se}')

#for 18-25 age group
sample_se=(m_and_s[3][1]/np.sqrt(n))
print(f'Confidence Intervals for 18-25 age group are {m_and_s[3][0]-1.96*sample_se} & {m_and_s[3][0]+1.96*sample_se}')

#for 46-50 age group
sample_se=(m_and_s[4][1]/np.sqrt(n))
print(f'Confidence Intervals for 46-50 age group are {m_and_s[4][0]-1.96*sample_se} & {m_and_s[4][0]+1.96*sample_se}')

#for 51-55 age group
sample_se=(m_and_s[5][1]/np.sqrt(n))
print(f'Confidence Intervals for 51-55 age group are {m_and_s[5][0]-1.96*sample_se} & {m_and_s[5][0]+1.96*sample_se}')

#for 55+ age group
sample_se=(m_and_s[6][1]/np.sqrt(n))
print(f'Confidence Intervals for 55+ age group are {m_and_s[6][0]-1.96*sample_se} & {m_and_s[6][0]+1.96*sample_se}')
```

```
Confidence Intervals for 0-17 age group are 8865.616377820876 & 8965.631452179121
Confidence Intervals for 26-35 age group are 9210.322353330701 & 9309.301446669297
Confidence Intervals for 36-45 age group are 9294.000912128451 & 9391.695317871547
Confidence Intervals for 18-25 age group are 9125.415110879303 & 9220.3589491207
Confidence Intervals for 46-50 age group are 9165.992860861616 & 9262.873569138383
Confidence Intervals for 51-55 age group are 9487.528992731855 & 9587.015757268146
Confidence Intervals for 55+ age group are 9294.026872081704 & 9390.169267918296
```

## Questions

1. Are women spending more money per transaction than men? Why or Why not?

Ans: No. CI of male and female do not overlap and upper limits of female purchase CI are lesser than
lower limits of male purchase CI. This proves that men usually spend more than women
(NOTE: as per data 77% contibutions are from men and only 23% purchases are from women).

The reason for less purchase by women could have several factors:

Males might be doing the purchase for females.
Salary can be a factor in less purchase.
We also need to see whether male-based products were sold more than women-based products to clearly identify difference
in spending pattern.
If the female based products quality/quantity needs to be improved for women purchasing.

2. Confidence intervals and distribution of the mean of the expenses by female and male customers.

Interval within which the 95% average amount spent for male population will lie between 9369.142112191461 & 9470.872917808541
Interval within which the 99% average amount spent for male population will lie between 9353.31155315411 & 9486.703476845893

Interval within which the 95% average amount spent for female population will lie between 8667.70005589184 & 8764.009924108159
Interval within which the 99% average amount spent for female population will lie between 8652.713061092873 & 8778.996918907127

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

Ans: No. Confidence intervals of average male and female spending are not overlapping. This trend can be changed
via introducing female centric marketing strategies by Walmart so that more female customers are attracted to
increase female purchases to achieve comparable statistics close to 50%.

4. Results when the same activity is performed for Married vs Unmarried

Interval within which the 95% average amount spent for married population will lie between 9210.171267885962 & 9306.82258211404
Interval within which the 99% average amount spent for married population will lie between 9195.131139906593 & 9321.86271009341

Interval within which the 95% average amount spent for married population will lie between9217.77448636021 & 9318.012303639789

```
Interval within which the 99% average amount spent for married population will lie between9202.176254589664 &
9333.610535410335

5. Results when the same activity is performed for Age

At 99% Confidence Interval with sample size 200

Confidence Intervals for 0-17 age group are 8865.616377820876 & 8965.631452179121
Confidence Intervals for 26-35 age group are 9210.322353330701 & 9309.301446669297
Confidence Intervals for 36-45 age group are 9294.000912128451 & 9391.695317871547
Confidence Intervals for 18-25 age group are 9125.415110879303 & 9220.3589491207
Confidence Intervals for 46-50 age group are 9165.992860861616 & 9262.873569138383
Confidence Intervals for 51-55 age group are 9487.528992731855 & 9587.015757268146
Confidence Intervals for 55+ age group are 9294.026872081704 & 9390.169267918296
```

## Recommendations:

```
1. Men spent more money than women, company can focus on retaining the male customers and getting more male
customers.

2. Product_Category - 1, 5, 8 have highest purchasing frequency. it means these are the products in these categories
are
   in more demand. Company can focus on selling more of these products.

3. Unmarried customers spend more money than married customers, So company should focus on acquisition of Unmarried
customers.

4. Customers in the age 26-35 spend more money than the others, So company should focus on acquisition of customers
who are
   in the age 26-35.

5. We have more customers aged 26-35 in the city category B and A, company can focus more on these customers for
these
   cities to increase the business.

6. Male customers living in City_Category C spend more money than other male customers living in B or C, Selling more
   products in the City_Category C will help the company increase the revenue.

7. Some of the Product category like 19,20,13 have very less purchase. Company can think of dropping it.
```

```
8. The top 10 users who have purchased more company should give more offers and discounts so that they can be
retained
   and can be helpful for companies business.

9. The occupation which are contributing more company can think of offering credit cards or other benefits to those
   customers by liasing with some financial partners to increase the sales.

10. The top products should be given focus in order to maintain the quality in order to further increase the sales
    of those products.

11. People who are staying in city for an year have contributed to 35% of the total purchase amount. Company can
focus
    on such customer base who are neither too old nor too new residents in the city.

12. We have highest frequency of purchase order between 5k and 10k, company can focus more on these mid range
products
    to increase the sales.
```