In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [2]:
```
df = pd.read_csv('Diwali Sales Data.csv', encoding='unicode_escape')
```

In [3]:
```
df.shape
```

Out[3]:
```
(11251, 15)
```

In [4]:
```
df.head()
```

Out[4]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	State
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh
2	1001990	Bindu	P00118542	F	26-35	35	1	Uttar Pradesh
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11251 entries, 0 to 11250
Data columns (total 15 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   User_ID           11251 non-null  int64
 1   Cust_name         11251 non-null  object
 2   Product_ID        11251 non-null  object
 3   Gender            11251 non-null  object
 4   Age Group         11251 non-null  object
 5   Age               11251 non-null  int64
 6   Marital_Status    11251 non-null  int64
 7   State             11251 non-null  object
 8   Zone              11251 non-null  object
 9   Occupation        11251 non-null  object
 10  Product_Category  11251 non-null  object
 11  Orders            11251 non-null  int64
 12  Amount            11239 non-null  float64
 13  Status            0 non-null      float64
 14  unnamed1          0 non-null      float64
dtypes: float64(3), int64(4), object(8)
memory usage: 1.3+ MB
```

In [6]:
```
df.drop(['Status','unnamed1'],axis=1, inplace=True)
```

In [7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11251 entries, 0 to 11250
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype  
---  ------            --------------  -----  
 0   User_ID           11251 non-null  int64  
 1   Cust_name         11251 non-null  object 
 2   Product_ID        11251 non-null  object 
 3   Gender            11251 non-null  object 
 4   Age Group         11251 non-null  object 
 5   Age               11251 non-null  int64  
 6   Marital_Status    11251 non-null  int64  
 7   State             11251 non-null  object 
 8   Zone              11251 non-null  object 
 9   Occupation        11251 non-null  object 
 10  Product_Category  11251 non-null  object 
 11  Orders            11251 non-null  int64  
 12  Amount            11239 non-null  float64
dtypes: float64(1), int64(4), object(8)
memory usage: 1.1+ MB
```

In [8]:
```
pd.isnull(df)
```

Out[8]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone
0	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...
11246	False	False	False	False	False	False	False	False	False
11247	False	False	False	False	False	False	False	False	False
11248	False	False	False	False	False	False	False	False	False
11249	False	False	False	False	False	False	False	False	False
11250	False	False	False	False	False	False	False	False	False

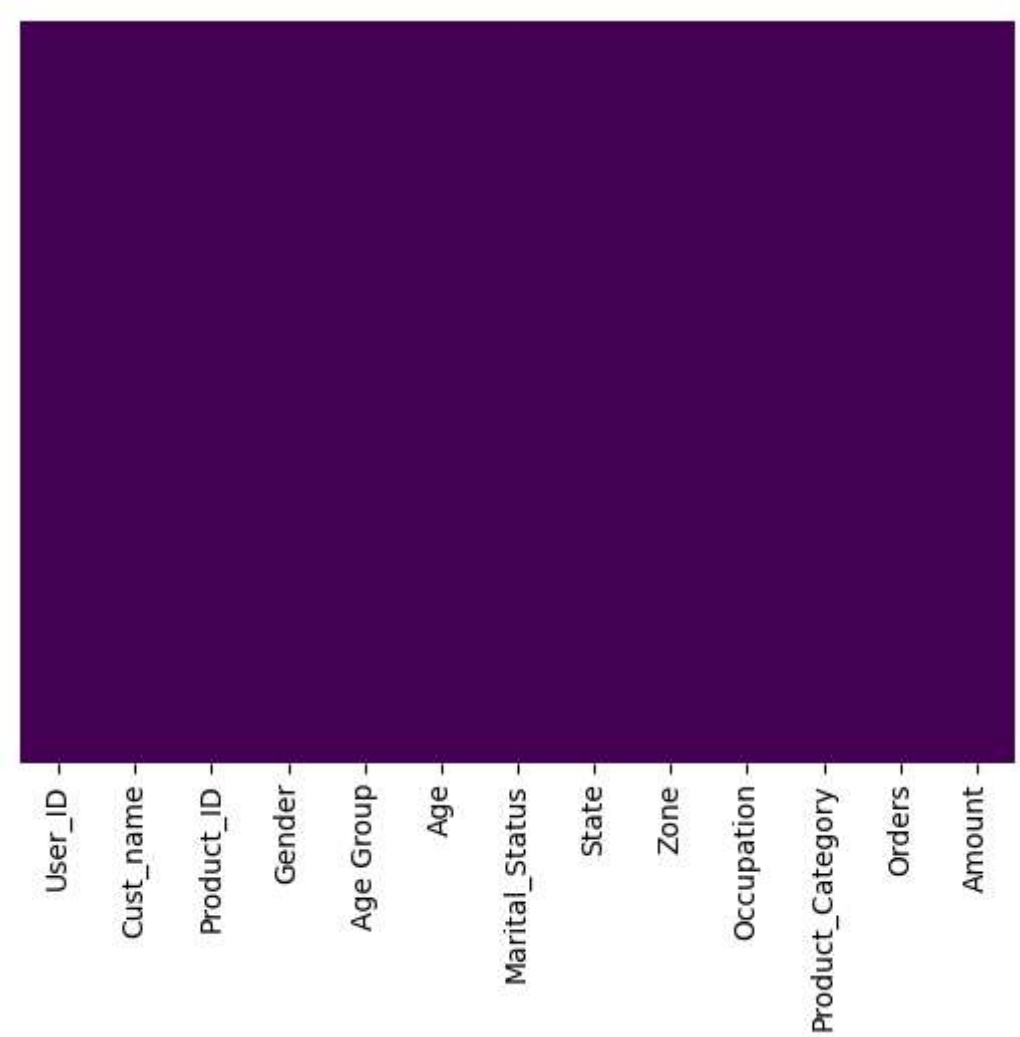
11251 rows × 13 columns

In [9]:
```
df.isnull().sum()
```

Out[9]:
```
User_ID             0
Cust_name           0
Product_ID          0
Gender              0
Age Group           0
Age                 0
Marital_Status      0
State               0
Zone                0
Occupation          0
Product_Category    0
Orders              0
Amount             12
dtype: int64
```

In [10]:
```
sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

Out[10]: <Axes: >



In [12]:
```
df.shape
```

Out[12]:
```
(11251, 13)
```

In [13]:
```
df.dropna(inplace=True)
```

In [14]:
```
df.shape
```

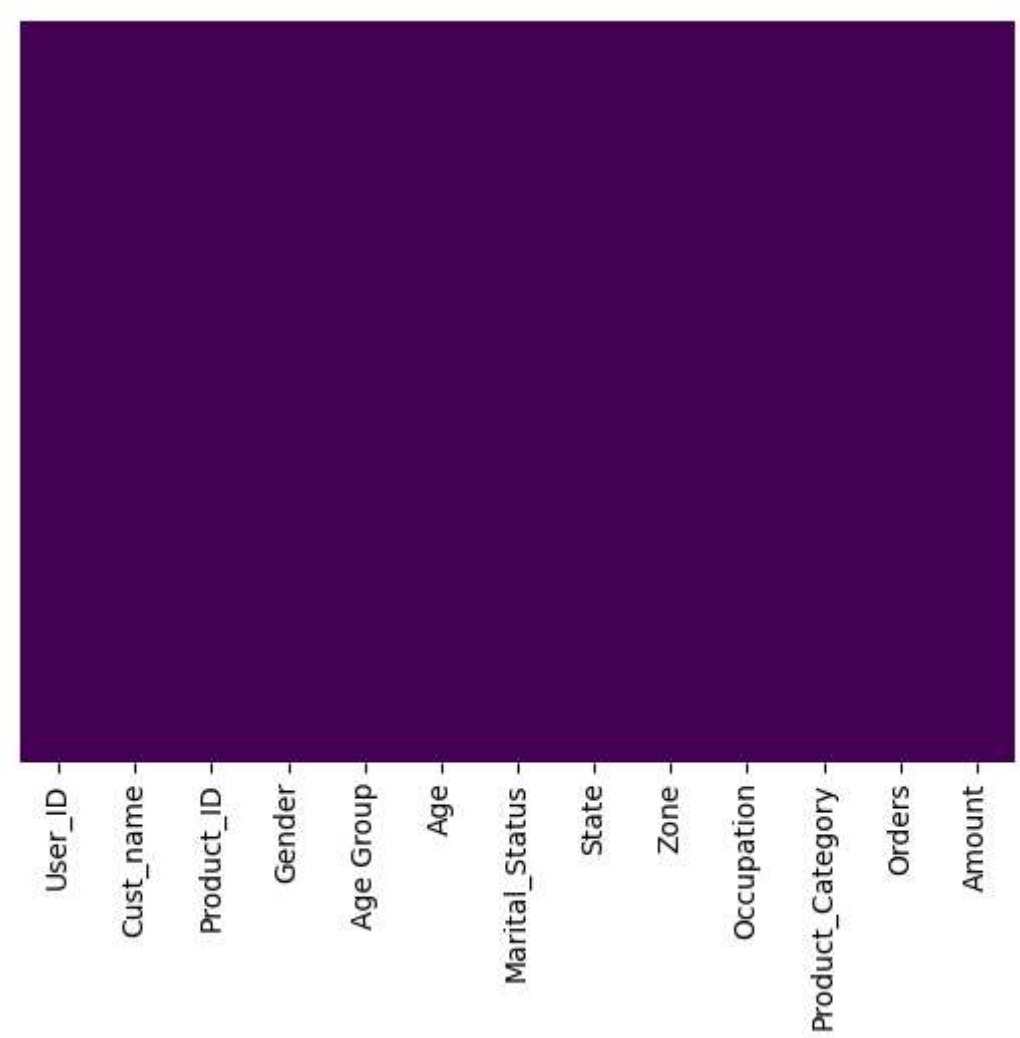
Out[14]:
```
(11239, 13)
```

In [15]:
```
df.isnull().sum()
```

Out[15]:
```
User_ID             0
Cust_name           0
Product_ID          0
Gender              0
Age Group           0
Age                 0
Marital_Status      0
State               0
Zone                0
Occupation          0
Product_Category    0
Orders              0
Amount              0
dtype: int64
```

In [16]:
```
sns.heatmap(df.isnull(),yticklabels=False,cbar=False,cmap='viridis')
```

Out[16]:
```
<Axes: >
```



In [17]:
```
df['Amount'] = df['Amount'].astype(int)
```

In [18]:
```
df['Amount'].dtype
```

Out[18]:
```
dtype('int32')
```

In [19]:
```
df.columns
```

Out[19]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Ag
e',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Categor
y',
       'Orders', 'Amount'],
      dtype='object')
```

In [22]:
```
df.describe()
```

Out[22]:

	User_ID	Age	Marital_Status	Orders	Amount
count	1.123900e+04	11239.000000	11239.000000	11239.000000	11239.000000
mean	1.003004e+06	35.410357	0.420055	2.489634	9453.610553
std	1.716039e+03	12.753866	0.493589	1.114967	5222.355168
min	1.000001e+06	12.000000	0.000000	1.000000	188.000000
25%	1.001492e+06	27.000000	0.000000	2.000000	5443.000000
50%	1.003064e+06	33.000000	0.000000	2.000000	8109.000000
75%	1.004426e+06	43.000000	1.000000	3.000000	12675.000000
max	1.006040e+06	92.000000	1.000000	4.000000	23952.000000

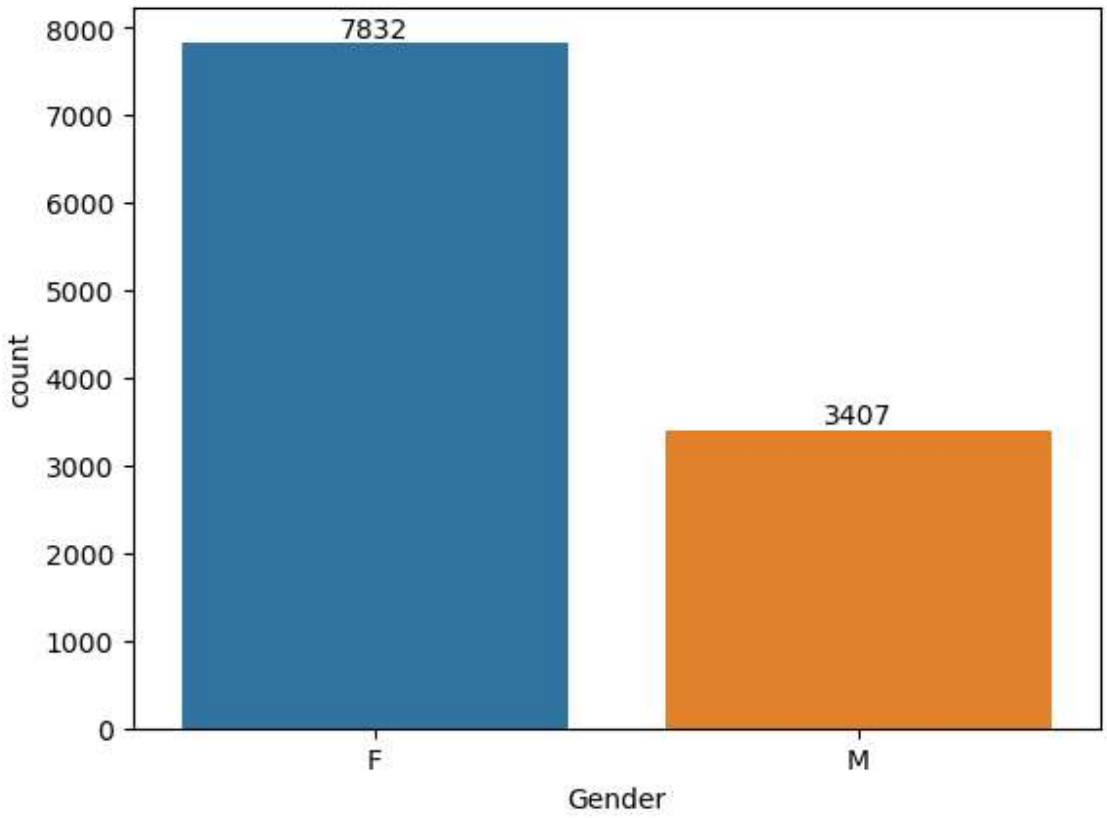
In [23]:
```
df[['Age','Orders','Amount']].describe()
```

Out[23]:

	Age	Orders	Amount
count	11239.000000	11239.000000	11239.000000
mean	35.410357	2.489634	9453.610553
std	12.753866	1.114967	5222.355168
min	12.000000	1.000000	188.000000
25%	27.000000	2.000000	5443.000000
50%	33.000000	2.000000	8109.000000
75%	43.000000	3.000000	12675.000000
max	92.000000	4.000000	23952.000000

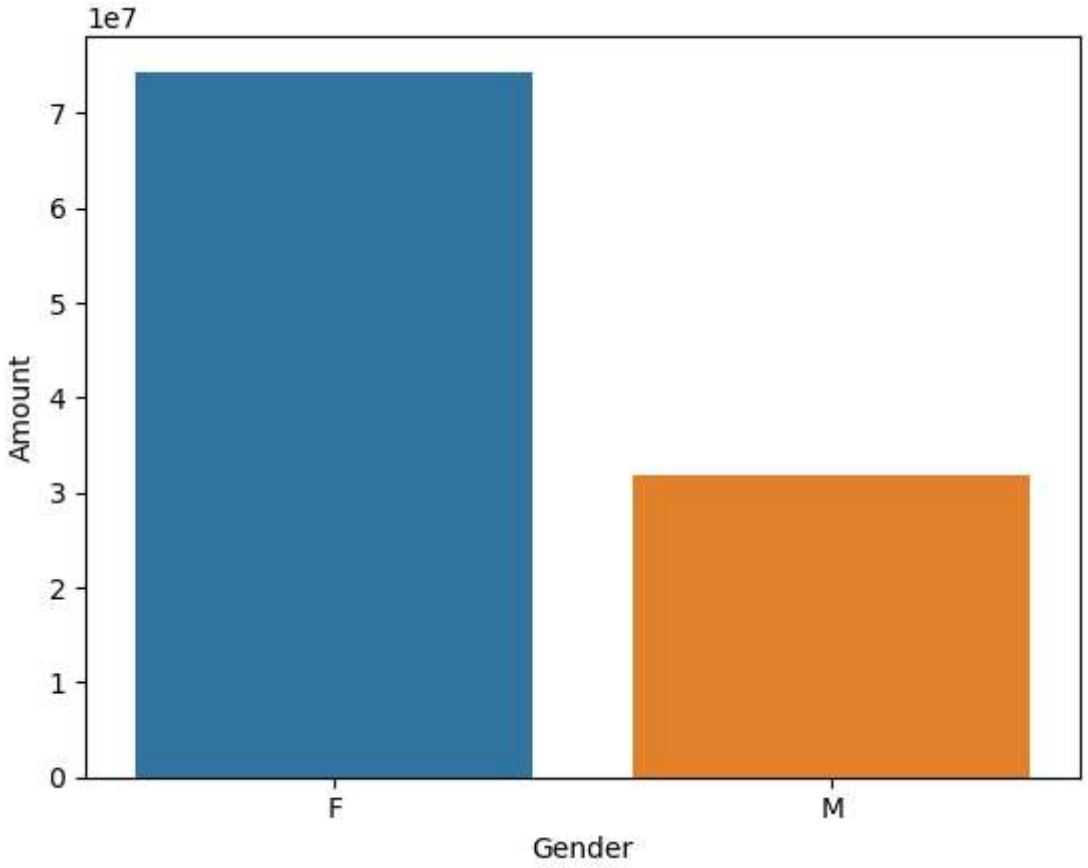
EXPLORATORY ANALYSIS

In [25]:
```
ax = sns.countplot(x='Gender',data=df)

for bars in ax.containers:
    ax.bar_label(bars)
```



In [26]:
```
sales_gen = df.groupby(['Gender'], as_index=False)['Amount'].sum().sort_va

sns.barplot(x='Gender',y='Amount',data=sales_gen)
```

Out[26]: <Axes: xlabel='Gender', ylabel='Amount'>

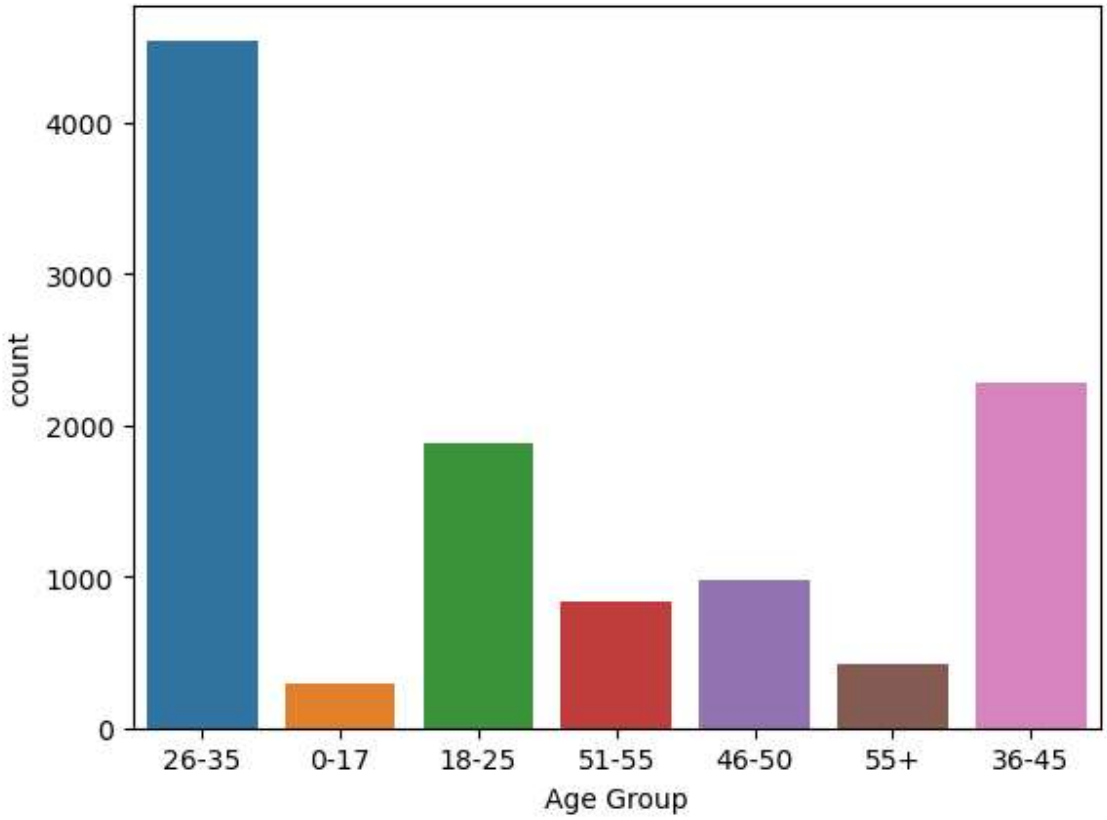


Most purchasers are the women.

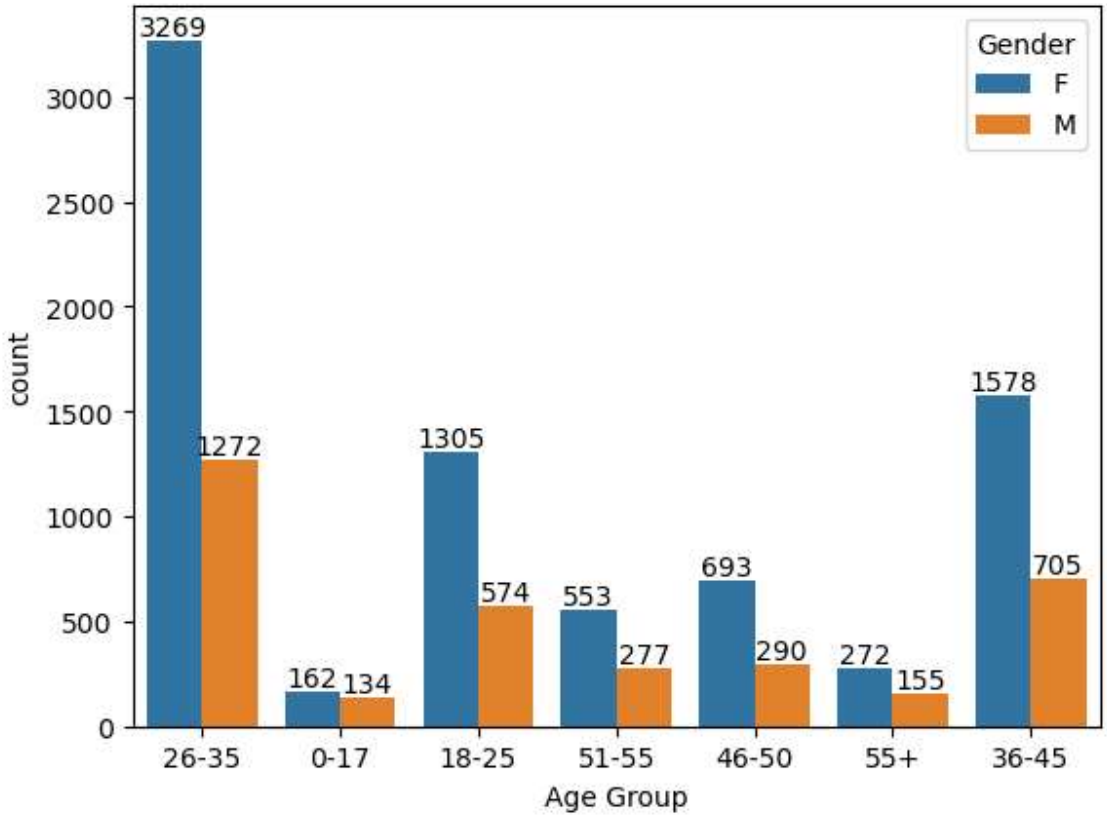
AGE

In [27]:
```
sns.countplot(x='Age Group',data=df)
```

Out[27]:
```
<Axes: xlabel='Age Group', ylabel='count'>
```

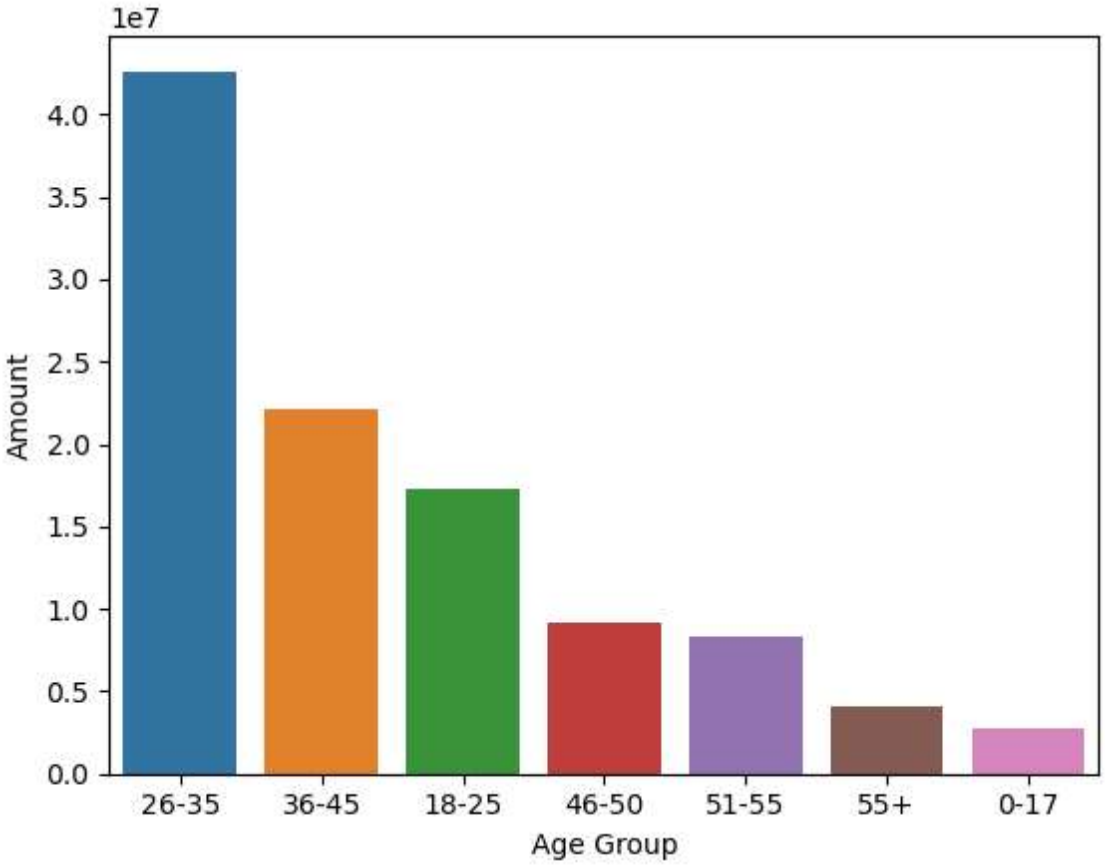


In [30]:
```
ax = sns.countplot(data=df, x='Age Group', hue='Gender')

for bars in ax.containers:
    ax.bar_label(bars)
```



In [31]:
```
sales_age = df.groupby(['Age Group'], as_index=False)['Amount'].sum().sort

sns.barplot(x='Age Group',y='Amount',data=sales_age)
```

Out[31]: <Axes: xlabel='Age Group', ylabel='Amount'>

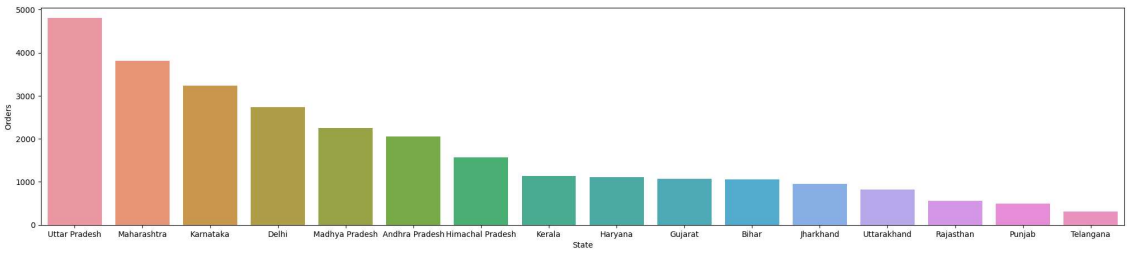


Most of the buyers are between the age of 26 to 35 and the most are females

State

In [32]:
```
sales_state = df.groupby(['State'], as_index=False)['Orders'].sum().sort_v

#sns.set(rc={'figure.figsize' : (6.4,4.8)})
plt.figure(figsize=(25,5))
sns.barplot(data=sales_state,x='State',y='Orders')
```

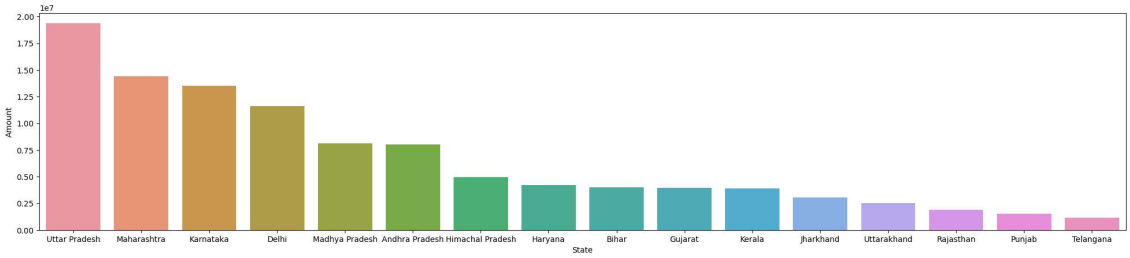
Out[32]: <Axes: xlabel='State', ylabel='Orders'>



Most of the orders are from Uttar Pradesh, then Maharashtra and then Karnataka

In [33]:
```
sales_state = df.groupby(['State'], as_index=False)['Amount'].sum().sort_v
plt.figure(figsize=(25,5))
sns.barplot(data=sales_state,x='State',y='Amount')
```

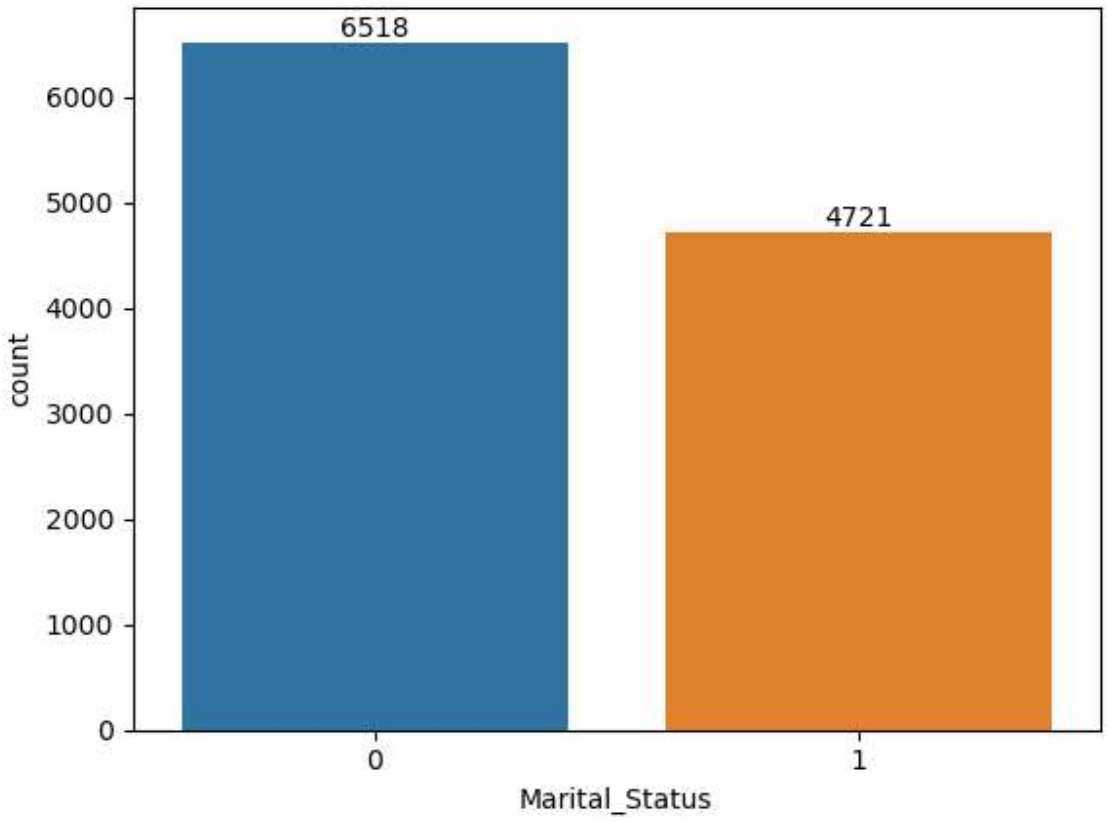
Out[33]: <Axes: xlabel='State', ylabel='Amount'>



The most of Amount was collected by these three states mentioned above too.

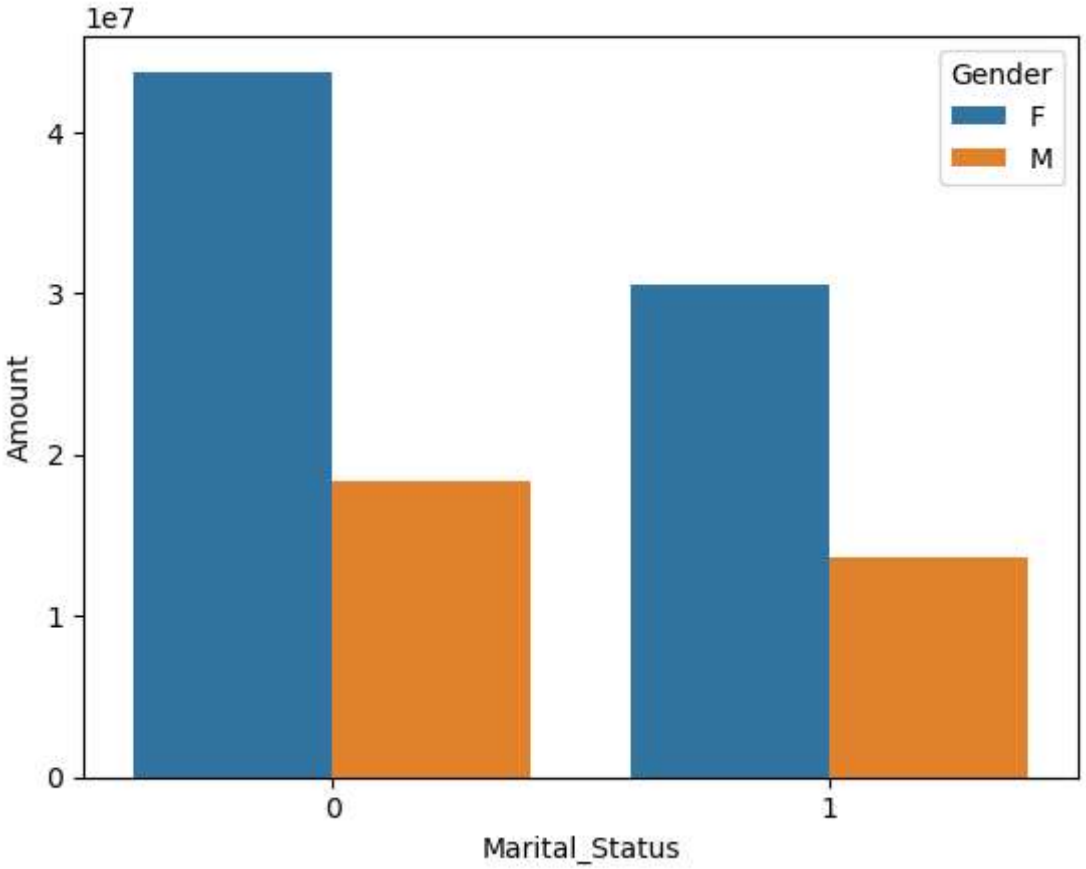
Marital Status

In [34]:
```
#plt.figure(figsize=(7,5))
ax = sns.countplot(x='Marital_Status',data=df)

for bars in ax.containers:
    ax.bar_label(bars)
```



In [35]:
```
sales_state = df.groupby(['Marital_Status','Gender'], as_index=False)['Amo
sns.barplot(data=sales_state,x='Marital_Status',y='Amount',hue='Gender')
```

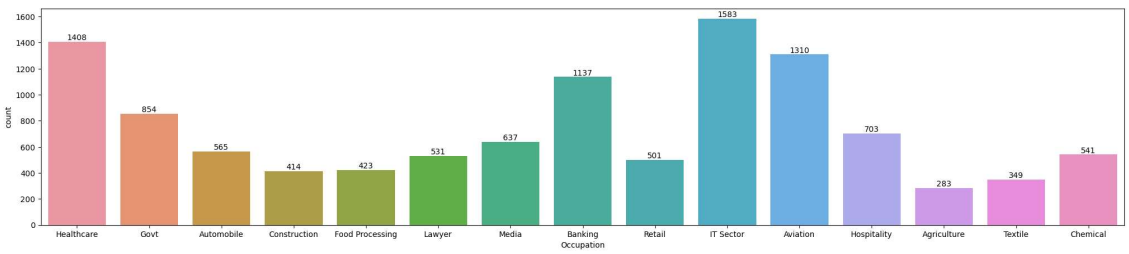
Out[35]: <Axes: xlabel='Marital_Status', ylabel='Amount'>



From above graph we can say that the most of the buyers are married(women) and they have high purchasing power

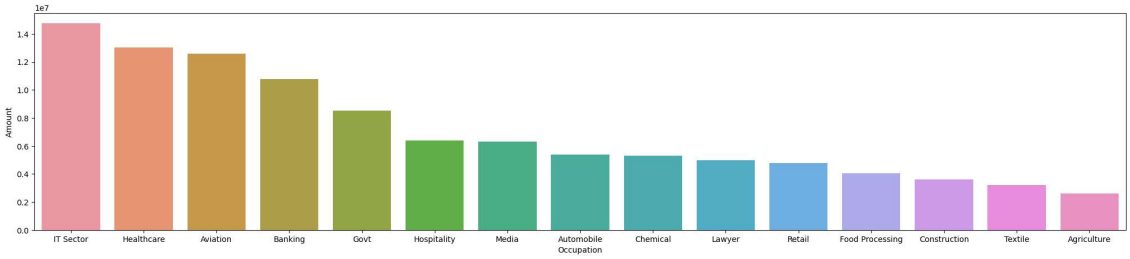
Occupation

In [36]:
```
plt.figure(figsize=(25,5))
ax = sns.countplot(data=df,x='Occupation')

for bars in ax.containers:
    ax.bar_label(bars)
```



In [37]:
```
plt.figure(figsize=(25,5))
sales_occu = df.groupby(['Occupation'], as_index=False)['Amount'].sum().so
sns.barplot(data=sales_occu,x='Occupation',y='Amount')
```

Out[37]:
```
<Axes: xlabel='Occupation', ylabel='Amount'>
```



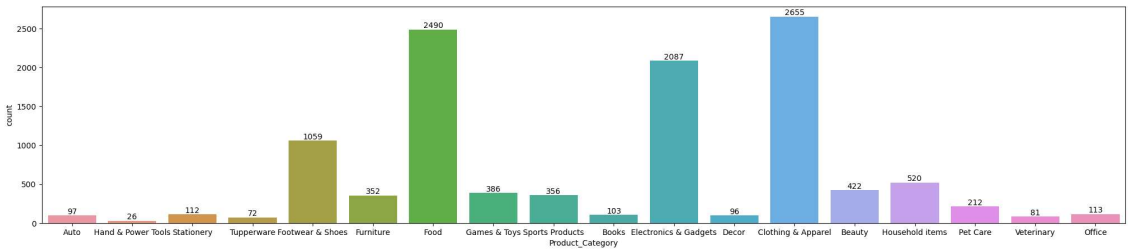
From above graph we can say that most of the buyers are from IT, HealthCare and Aviation Sector

In [38]:
```
df.columns
```

Out[38]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Ag
e',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Categor
y',
       'Orders', 'Amount'],
      dtype='object')
```

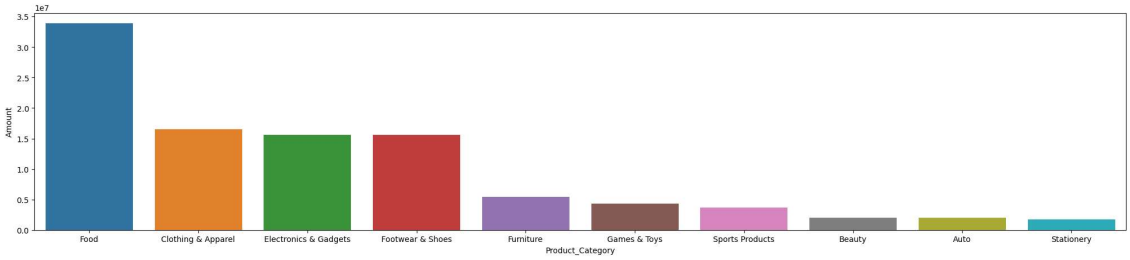
Product Category

In [39]:
```
plt.figure(figsize=(25,5))
ax = sns.countplot(data=df,x='Product_Category')
for bars in ax.containers:
    ax.bar_label(bars)
```



In [40]:
```
plt.figure(figsize=(25,5))
sales_cat = df.groupby(['Product_Category'], as_index=False)['Amount'].sum
sns.barplot(data=sales_cat,x='Product_Category',y='Amount')
```

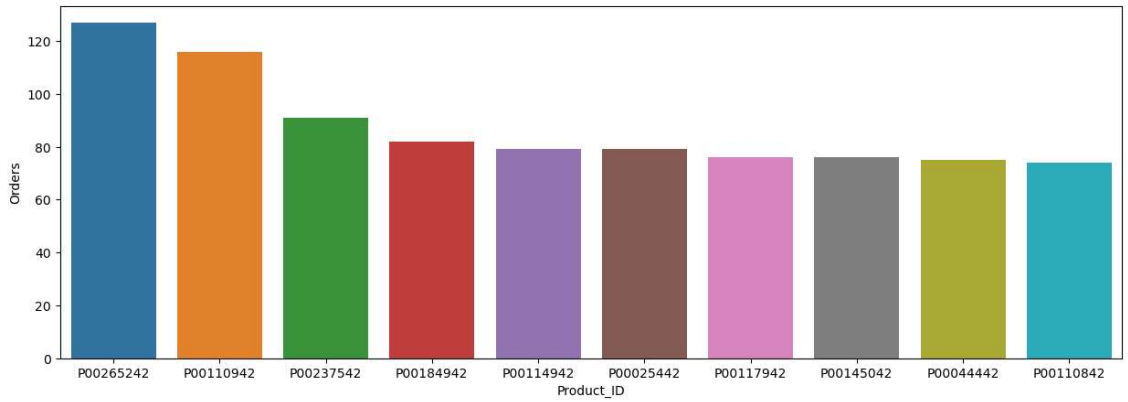
Out[40]: <Axes: xlabel='Product_Category', ylabel='Amount'>



From above graphs we can say that most of the sold products are from food, clothing, and Electronics catagory.

In [41]:
```
plt.figure(figsize=(15,5))
sales_cat = df.groupby(['Product_ID'], as_index=False)['Orders'].sum().sor
sns.barplot(data=sales_cat,x='Product_ID',y='Orders')
```

Out[41]: <Axes: xlabel='Product_ID', ylabel='Orders'>



Conclusion

Married women age group btw 26-35 years from UP, Maharashtra, and Karnataka working in IT, Healthcare, and Aviation sector are more likely to buy products from Food, Clothing, and Electronics category.