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

In [2]:
```
df = pd.read_csv(r'C:\Users\slgos\Downloads\Python_Diwali_Sales_Analysis-main\Python_D
```

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	Zone	
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	
2	1001990	Bindu	P00118542	F	26-35	35	1	Uttar Pradesh	Central	
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	C
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	

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 [12]:
```
pd.isnull(df)
```

Out[12]:

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

11251 rows × 13 columns

In [7]:
```
pd.isnull(df).sum()
```

Out[7]:
```
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 [9]:
```
df.dropna(inplace=True)
```

In [10]:
```
df['Amount']= df['Amount'].astype('int')
```

In [11]:
```
df['Amount'].dtypes
```

Out[11]:
```
dtype('int32')
```

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

Out[12]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

In [13]:
```
df.describe()
```

Out[13]:

	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 [14]:
```
df[['Orders','Amount']].describe()
```

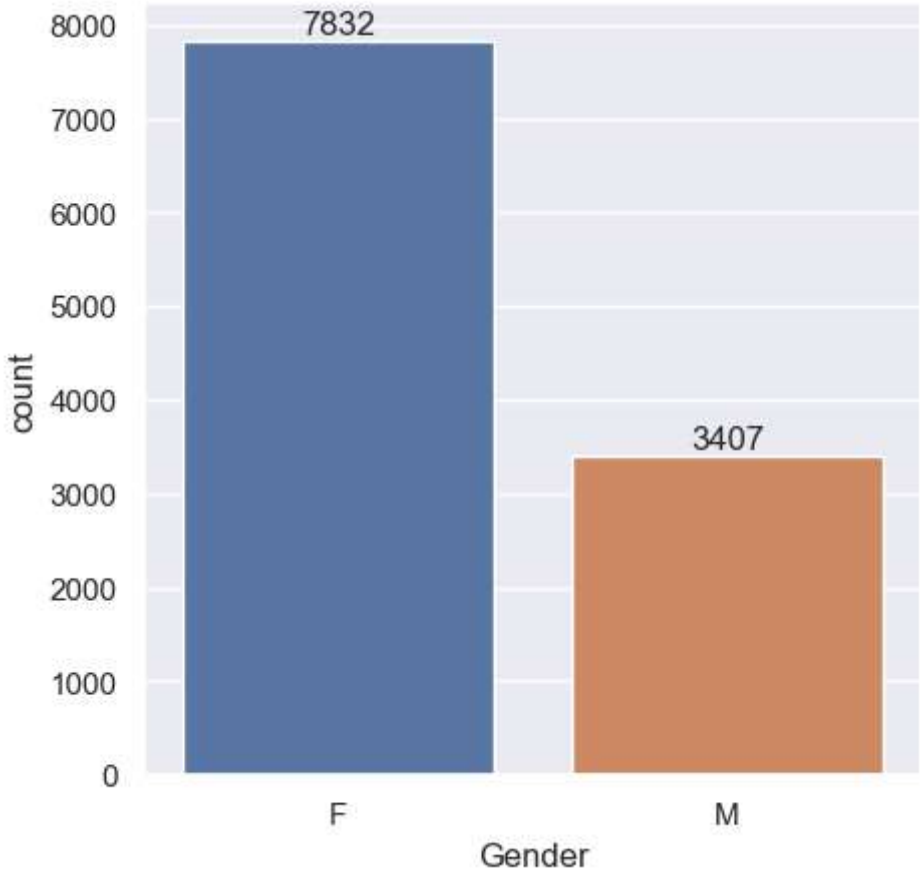
Out[14]:

	Orders	Amount
count	11239.000000	11239.000000
mean	2.489634	9453.610553
std	1.114967	5222.355168
min	1.000000	188.000000
25%	2.000000	5443.000000
50%	2.000000	8109.000000
75%	3.000000	12675.000000
max	4.000000	23952.000000

Exploratory Data Analytics (EDA)

Gender

In [29]:
```
ax = sns.countplot(x= 'Gender', data=df)

for bars in ax.containers:
    ax.bar_label(bars)
    +
sns.set(rc={'figure.figsize':(5,5)})
```

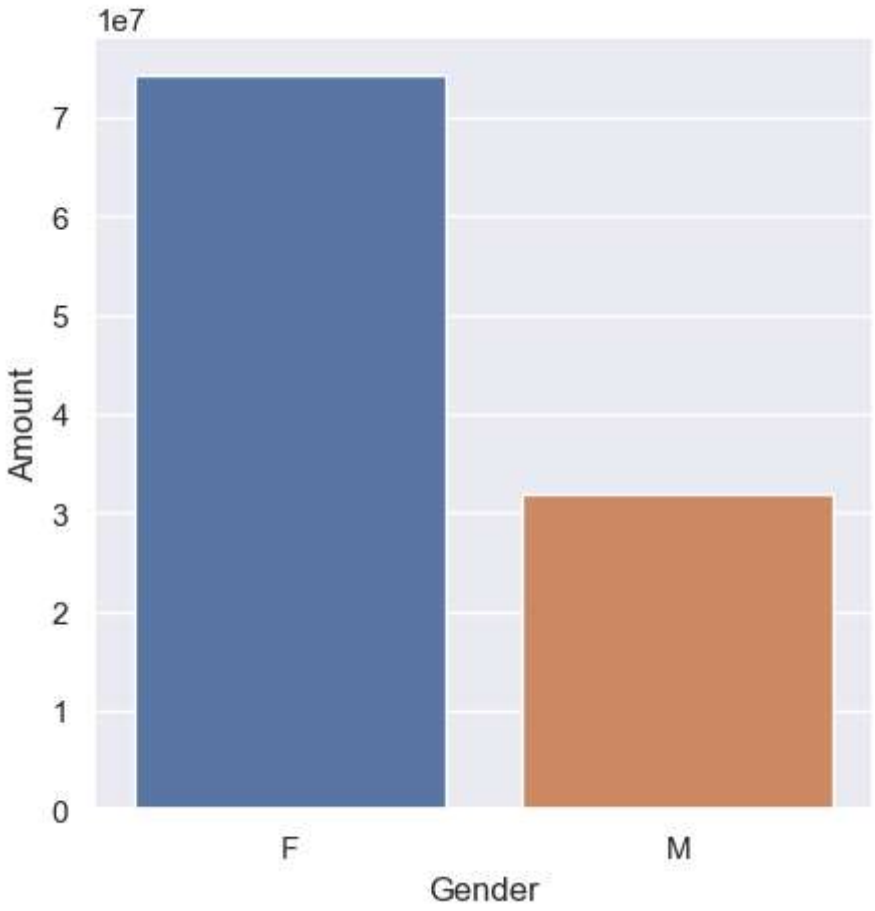


In [19]:
```
df.groupby(['Gender'], as_index=False)['Amount'].sum().sort_values(by='Amount', ascend
```

Out[19]:

	Gender	Amount
0	F	74335853
1	M	31913276

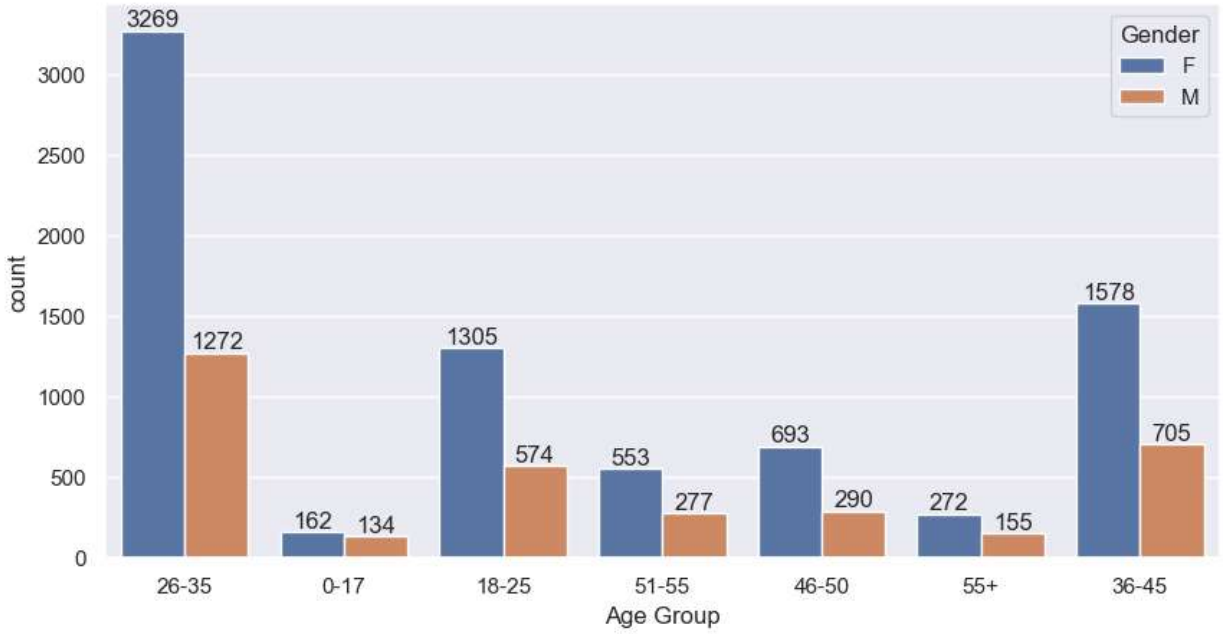
In [30]:
```
sales_gen = df.groupby(['Gender'], as_index=False)['Amount'].sum().sort_values(by='Amo

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

sns.set(rc={'figure.figsize':(5,5)})
```



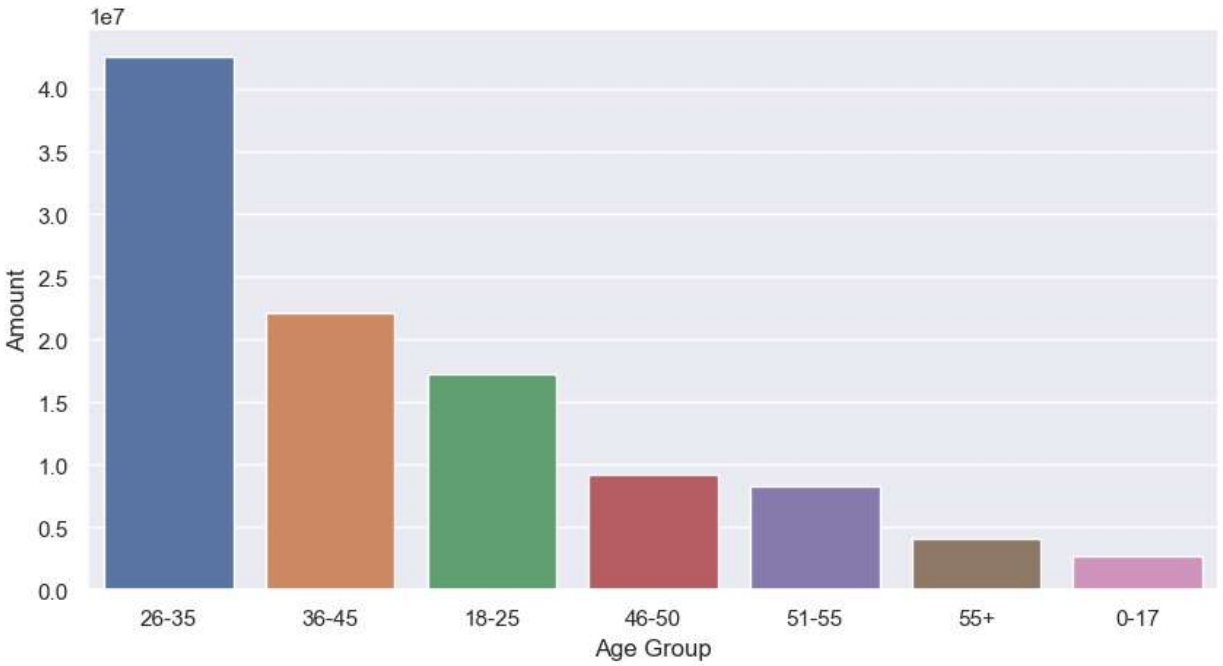
From above graphs we can see that most of the buyers are females and even the purchasing power of females are greater than men.

Age

In [44]:
```
ax = sns.countplot(data=df, x='Age Group', hue='Gender')

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

sns.set(rc={'figure.figsize':(10,5)})
```



In [42]:
```
sales_age = df.groupby(['Age Group'], as_index=False)['Amount'].sum().sort_values(by='

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

sns.set(rc={'figure.figsize':(10,5)})
```

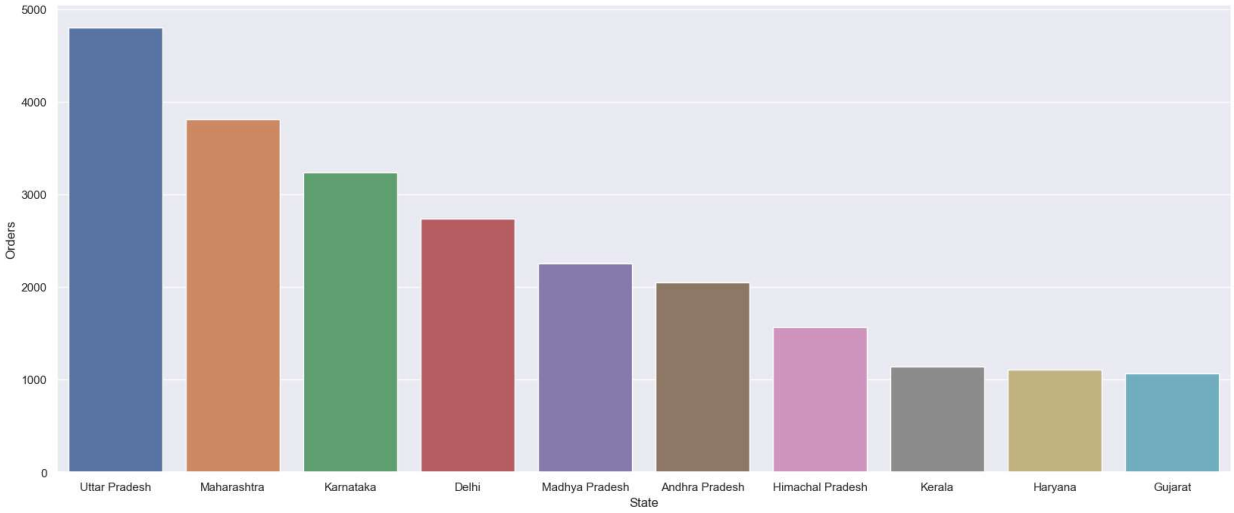


From above graphs we can see that most of the buyers are of age group 26-35 years female.

State

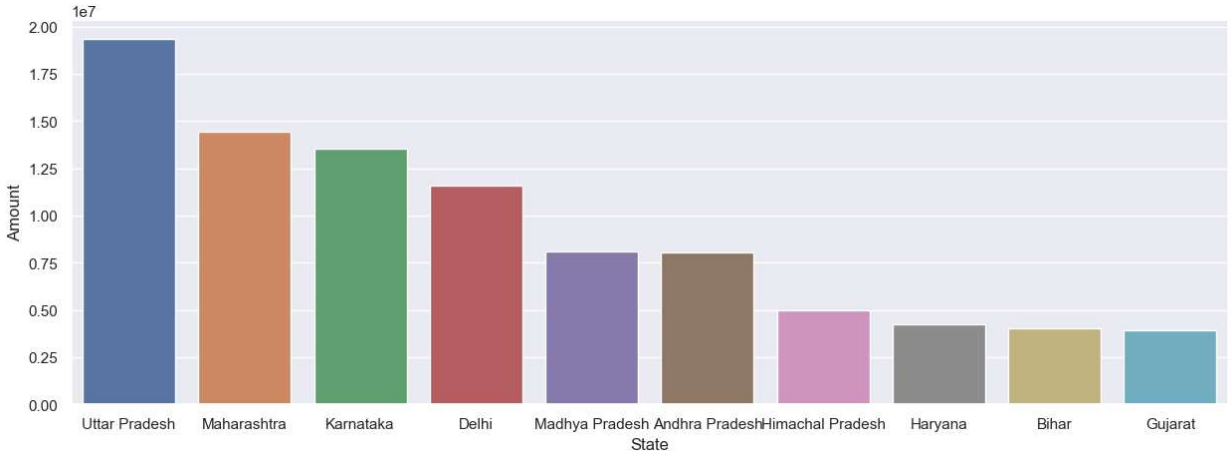
In [50]:
```
sales_state = df.groupby(['State'], as_index=False)['Orders'].sum().sort_values(by='Or

sns.set(rc={'figure.figsize':(20,8)})
sns.barplot(data = sales_state, x = 'State',y= 'Orders')
```

Out[50]:
```
<Axes: xlabel='State', ylabel='Orders'>
```



In [51]:
```
sales_state = df.groupby(['State'], as_index=False)['Amount'].sum().sort_values(by='An

sns.set(rc={'figure.figsize':(15,5)})
sns.barplot(data = sales_state, x = 'State',y= 'Amount')
```

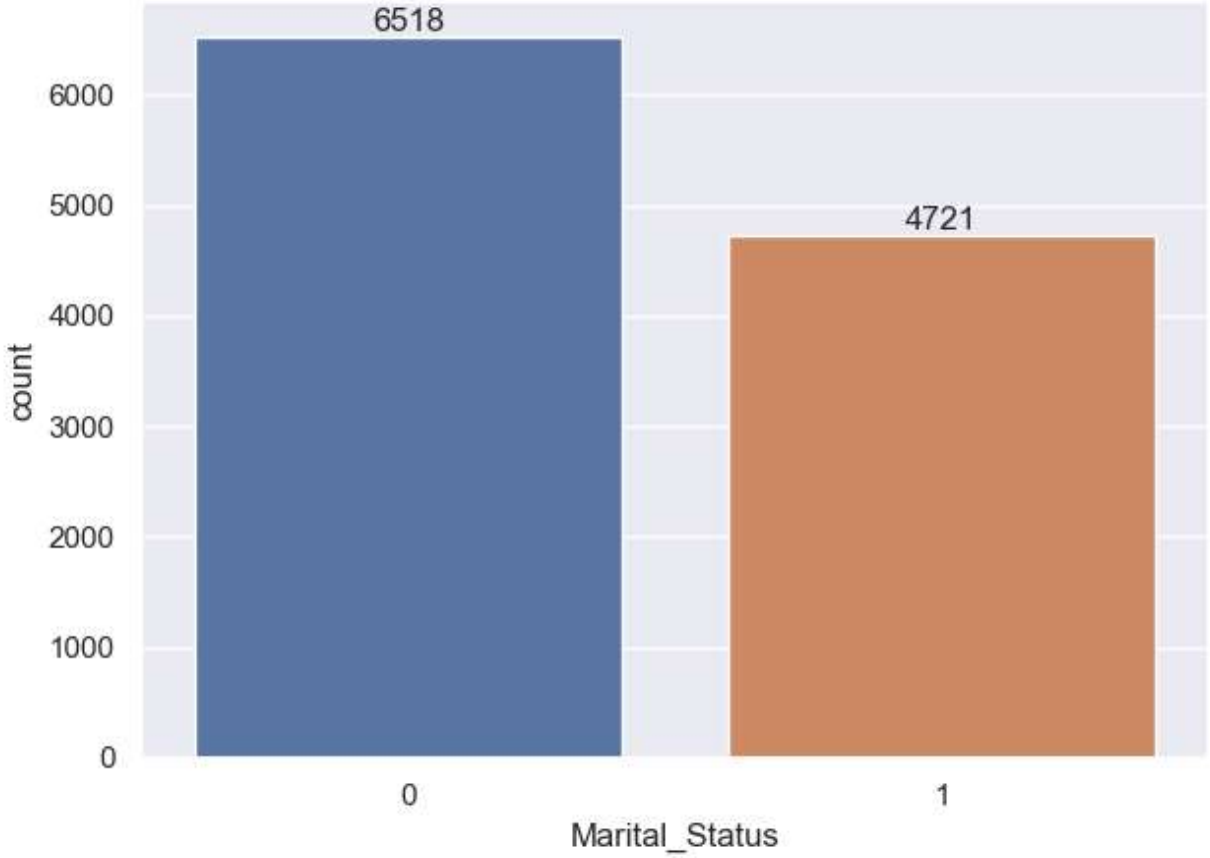
Out[51]:
```
<Axes: xlabel='State', ylabel='Amount'>
```



From above graphs we can see that most of the orders & total sales/amount are from Uttar Pradesh, Maharashtra and Karnataka respectively

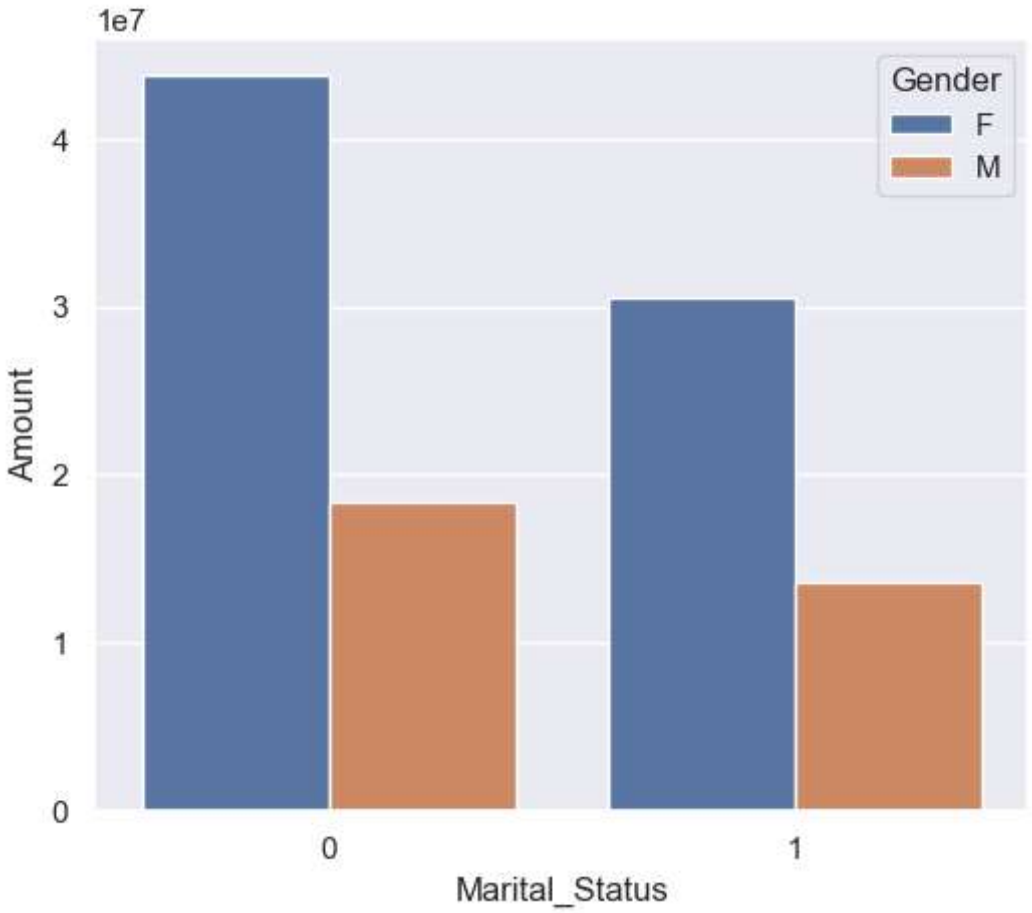
Marital Status

In [53]:
```
ax = sns.countplot(data = df, x = 'Marital_Status')

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

sns.set(rc={'figure.figsize':(7,5)})
```



In [54]:
```
sales_state = df.groupby(['Marital_Status', 'Gender'], as_index=False)['Amount'].sum()

sns.set(rc={'figure.figsize':(6,5)})
sns.barplot(data = sales_state, x = 'Marital_Status',y= 'Amount', hue='Gender')
```

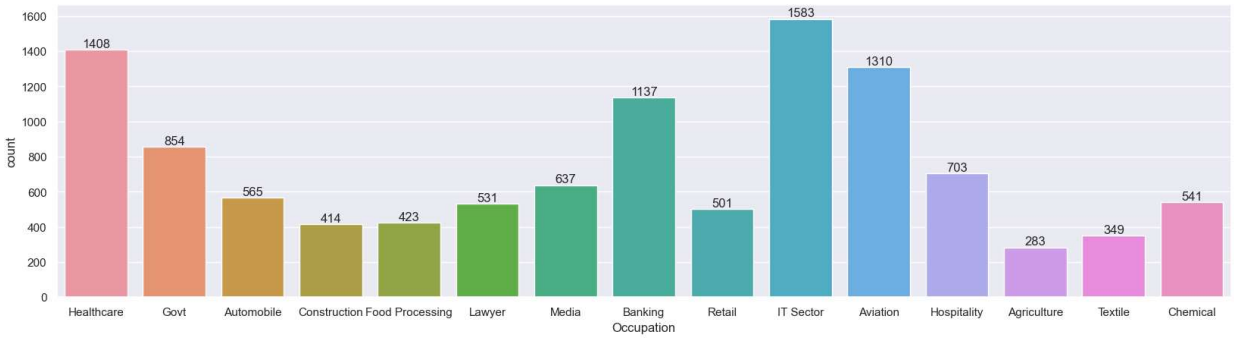
Out[54]:
```
<Axes: xlabel='Marital_Status', ylabel='Amount'>
```



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

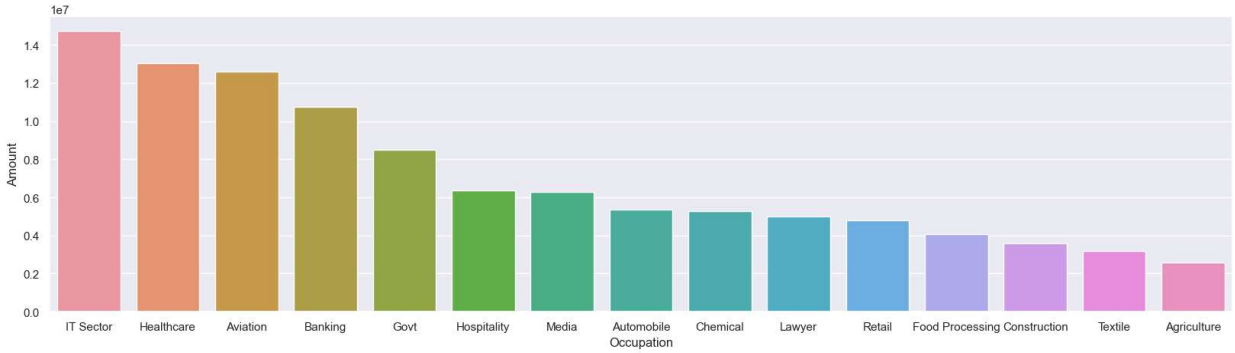
Occupation

In [55]:
```
sns.set(rc={'figure.figsize':(20,5)})
ax = sns.countplot(data = df, x = 'Occupation')

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



In [56]:
```
sales_state = df.groupby(['Occupation'], as_index=False)['Amount'].sum().sort_values(b

sns.set(rc={'figure.figsize':(20,5)})
sns.barplot(data = sales_state, x = 'Occupation',y= 'Amount')
```

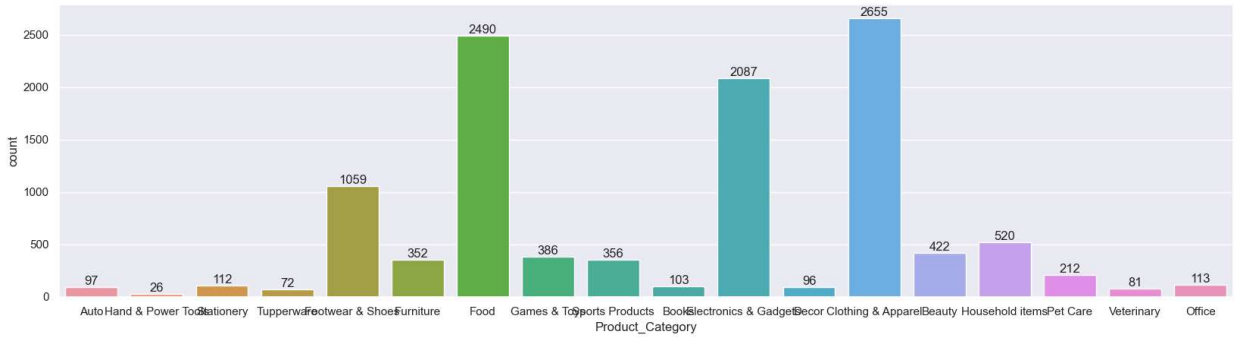
Out[56]:
```
<Axes: xlabel='Occupation', ylabel='Amount'>
```



From above graphs we can see that most of the buyers are working in IT, Healthcare and Aviation sector

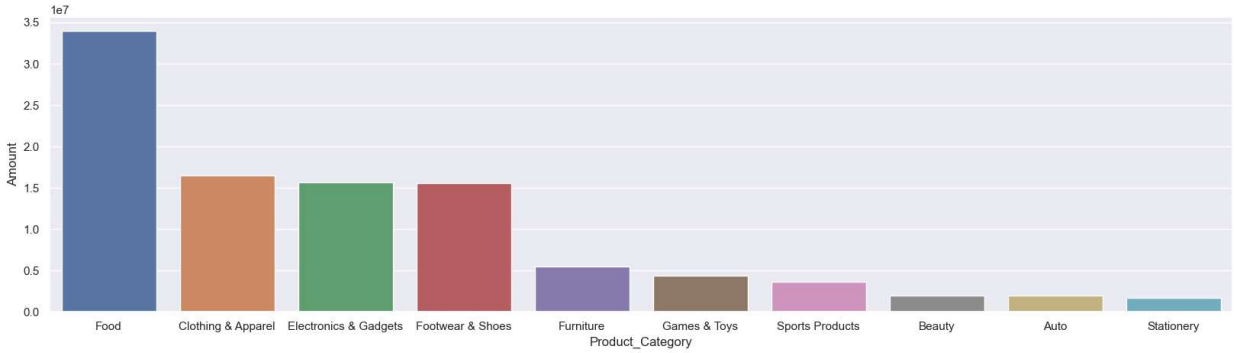
Product Category

In [57]:
```
sns.set(rc={'figure.figsize':(20,5)})
ax = sns.countplot(data = df, x = 'Product_Category')

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



In [58]:
```
sales_state = df.groupby(['Product_Category'], as_index=False)['Amount'].sum().sort_va

sns.set(rc={'figure.figsize':(20,5)})
sns.barplot(data = sales_state, x = 'Product_Category',y= 'Amount')
```

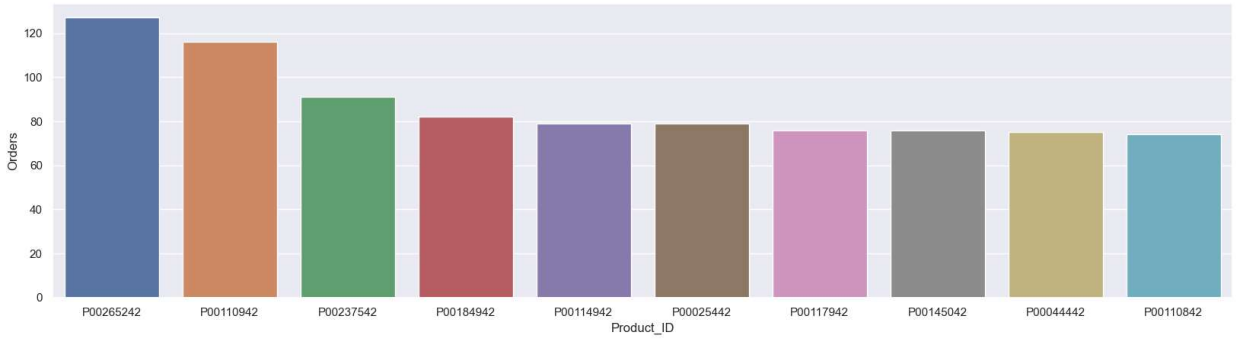
Out[58]:
```
<Axes: xlabel='Product_Category', ylabel='Amount'>
```



From above graphs we can see that most of the sold products are from Food, Clothing and Electronics category

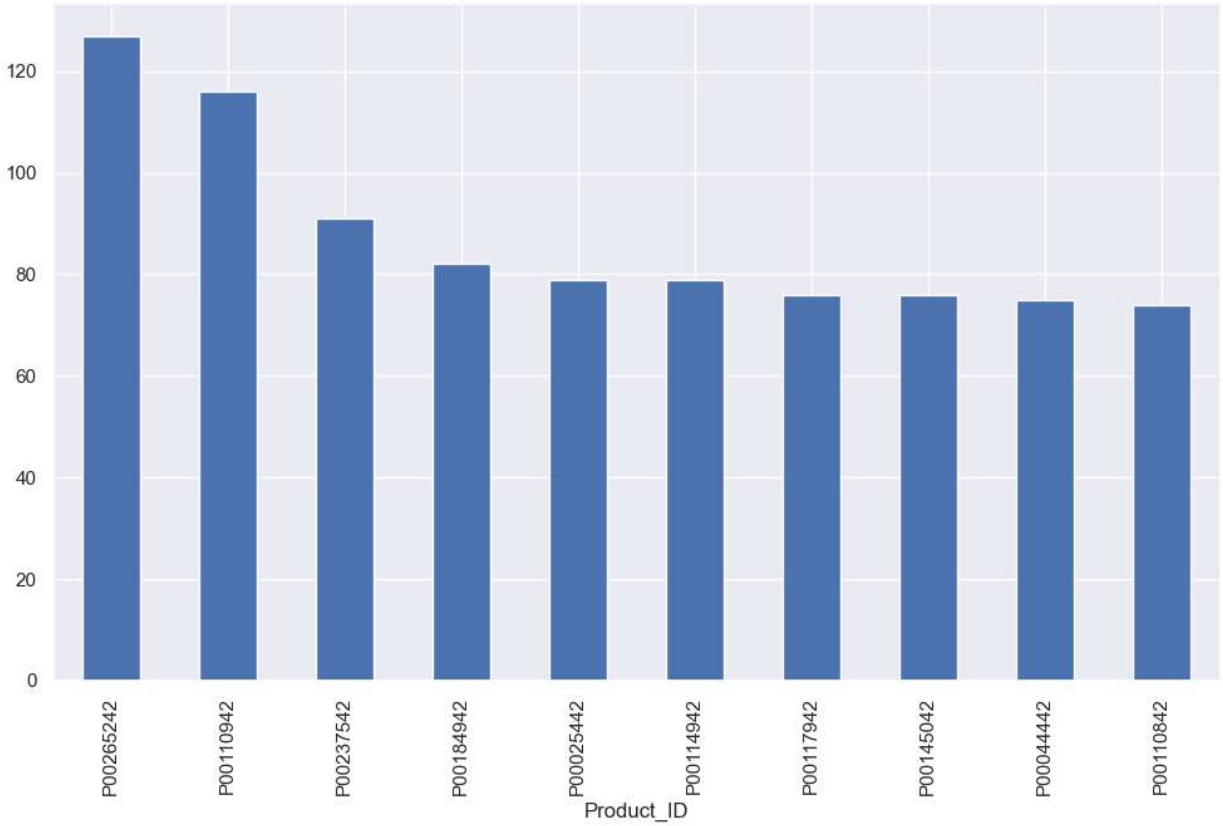
In [59]:
```
sales_state = df.groupby(['Product_ID'], as_index=False)['Orders'].sum().sort_values(b

sns.set(rc={'figure.figsize':(20,5)})
sns.barplot(data = sales_state, x = 'Product_ID',y= 'Orders')
```

Out[59]:
```
<Axes: xlabel='Product_ID', ylabel='Orders'>
```



In [62]:
```
fig1, ax1 = plt.subplots(figsize=(12,7))
df.groupby('Product_ID')['Orders'].sum().nlargest(10).sort_values(ascending=False).plo
```

Out[62]:
```
<Axes: xlabel='Product_ID'>
```



Conclusion

Married women age group 26-35 yrs from UP, Maharastra and Karnataka working in IT, Healthcare and Aviation are more likely to buy products from Food, Clothing and Electronics category