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

In [2]:
```python
df =pd.read_csv("Diwali Sales Data.csv")
```

In [3]:
```python
df.head()
```

Out[3]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Category	Orders	Amount	Status	unnamed1
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto	1	23952.0	NaN	NaN
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	Auto	3	23934.0	NaN	NaN
2	1001990	Bindu	P00118542	F	26-35	35	1	Uttar Pradesh	Central	Automobile	Auto	3	23924.0	NaN	NaN
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	Auto	2	23912.0	NaN	NaN
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	Auto	2	23877.0	NaN	NaN

In [4]:
```python
df.shape
```

Out[4]:
```
(11251, 15)
```

In [5]:
```python
df.isnull().sum()
```

Out[5]:
```
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
Status              11251
unnamed1            11251
dtype: int64
```

In [6]:
```python
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 [7]:
```python
df.drop(['Status','unnamed1'],axis=1,inplace=True)
```

In [8]:
```python
df.isnull().sum()
```

Out[8]:
```
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]:
```python
df.shape
```

Out[9]:
```
(11251, 13)
```

In [10]:
```python
df.dropna(inplace=True)
```

In [11]:
```python
df.shape
```

Out[11]:
```
(11239, 13)
```

In [12]:
```python
df['Amount']=df['Amount'].astype('int')
```

In [13]:
```python
df.dtypes
```

Out[13]:
```
User_ID              int64
Cust_name           object
Product_ID          object
Gender              object
Age Group           object
Age                  int64
Marital_Status       int64
State               object
Zone                object
Occupation          object
Product_Category    object
Orders               int64
Amount               int64
dtype: object
```

In [14]:
```python
df.columns
```

Out[14]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

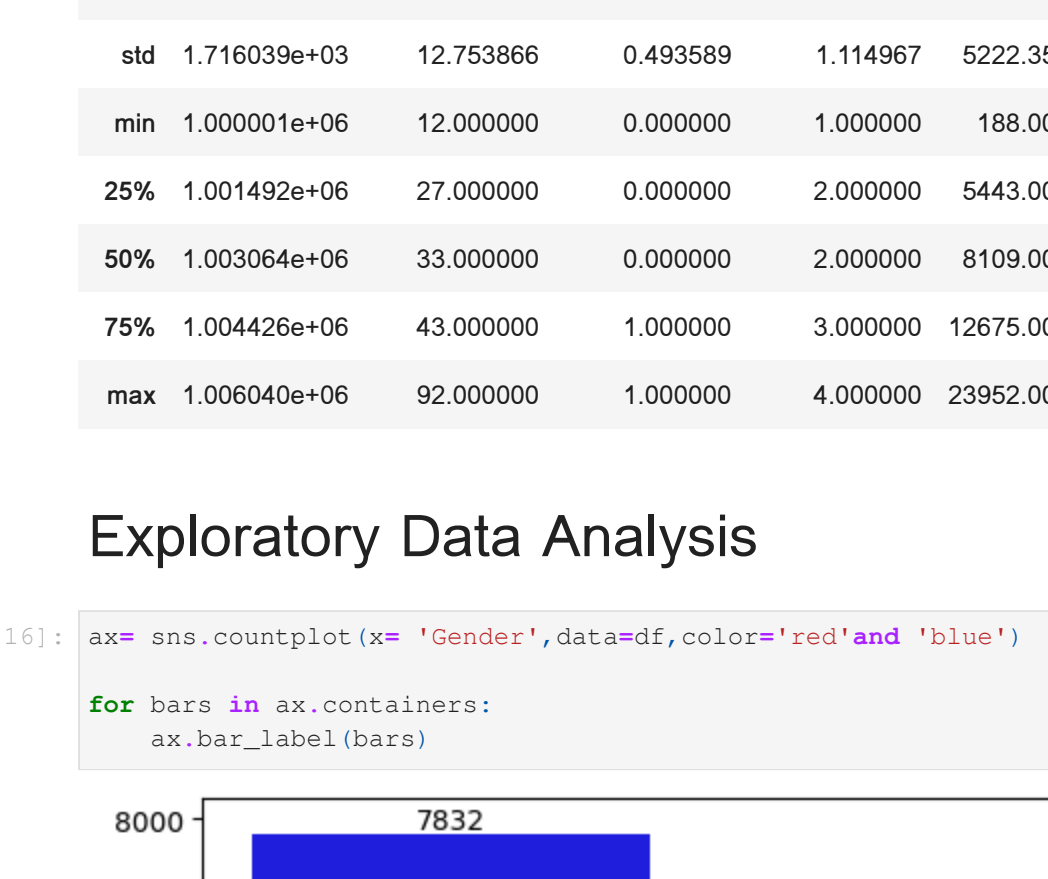
In [15]:
```python
df.describe()
```

Out[15]:

	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

Exploratory Data Analysis

In [16]:
```python
ax= sns.countplot(x= 'Gender',data=df,color='red'and 'blue')

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



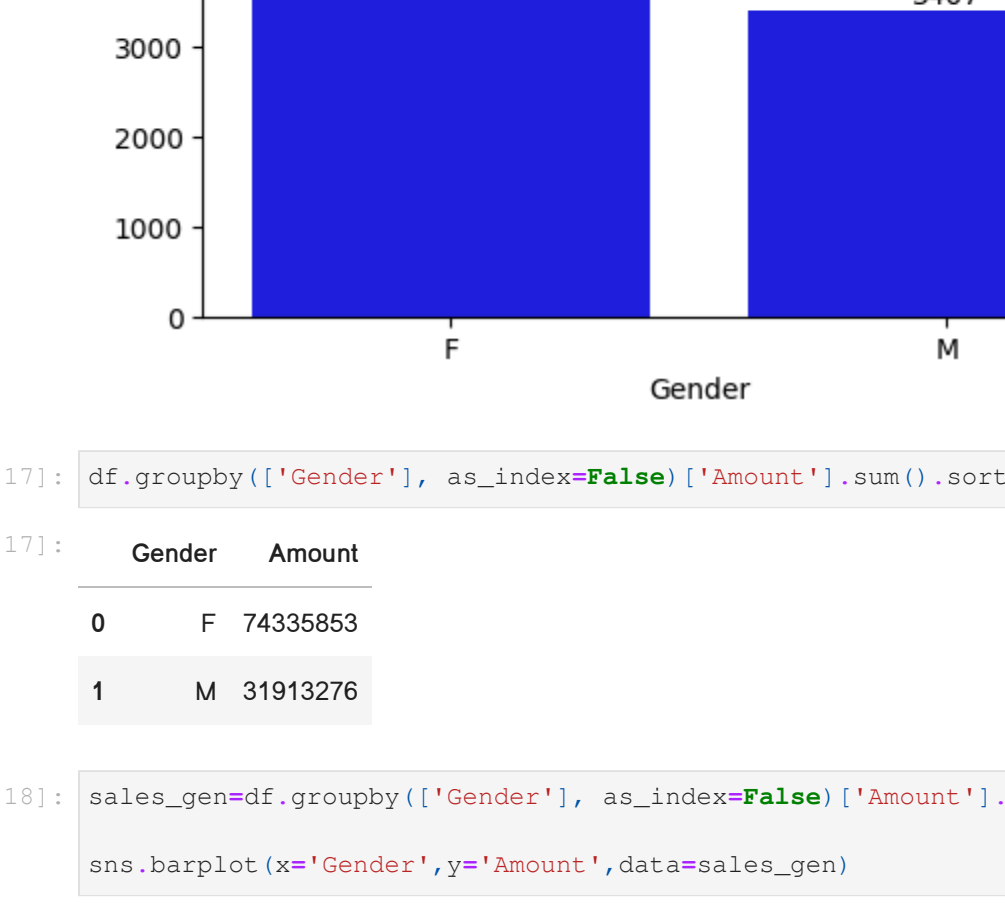
In [17]:
```python
df.groupby(['Gender'], as_index=False)['Amount'].sum().sort_values(by='Amount',ascending=False)
```

Out[17]:

	Gender	Amount
0	F	74335853
1	M	31913276

In [18]:
```python
sales_gen=df.groupby(['Gender'], as_index=False)['Amount'].sum().sort_values(by='Amount',ascending=False)

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

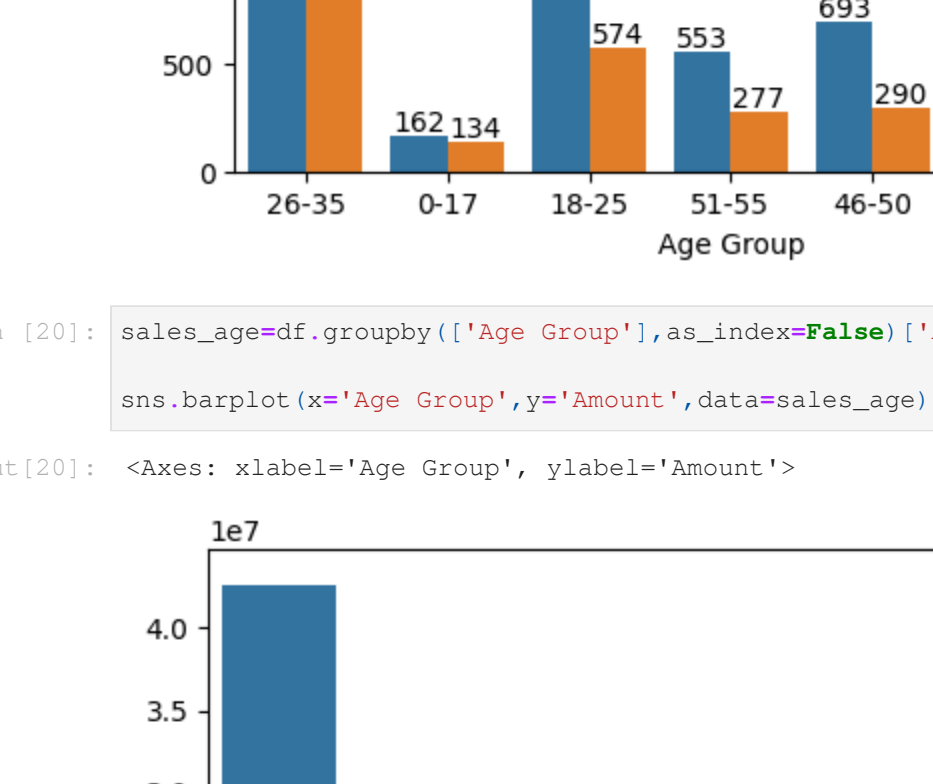
Out[18]:
```
<Axes: xlabel='Gender', ylabel='Amount'>
```



from above graphs we can see the most of the buyers are females and even the purchasing power of females are greater than man.

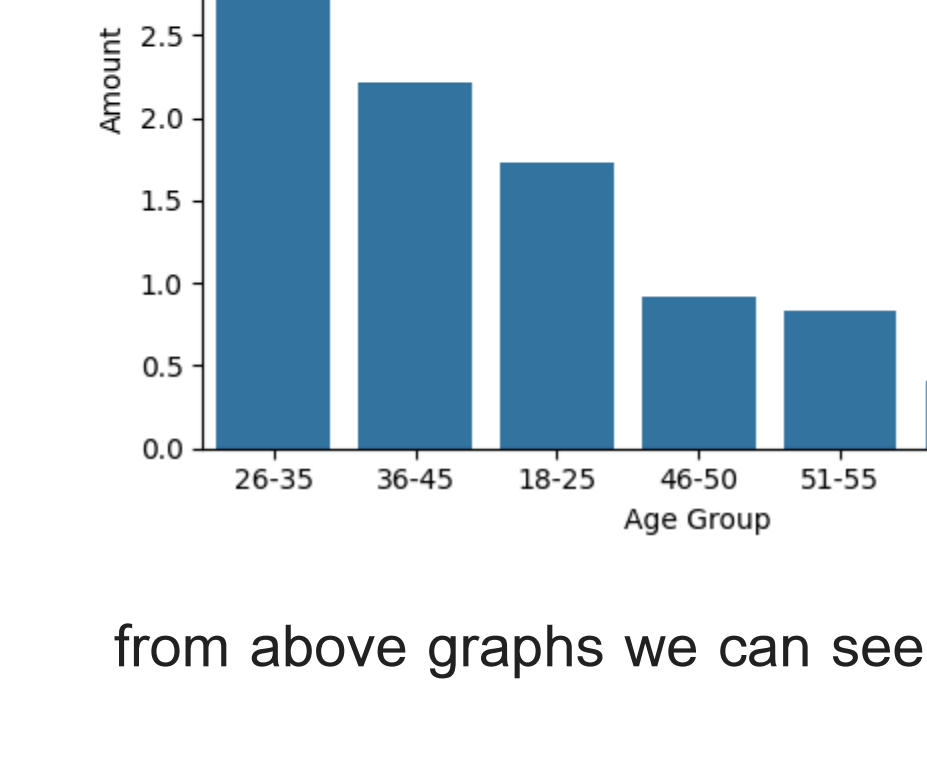
Age

In [19]:
```python
ax=sns.countplot(data=df,x='Age Group',hue ='Gender')

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



In [20]:
```python
sales_age=df.groupby(['Age Group'],as_index=False)['Amount'].sum().sort_values(by='Amount',ascending=False)

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

Out[20]:
```
<Axes: xlabel='Age Group', ylabel='Amount'>
```



from above graphs we can see that most of the buyers are of age 26-35 yrs female

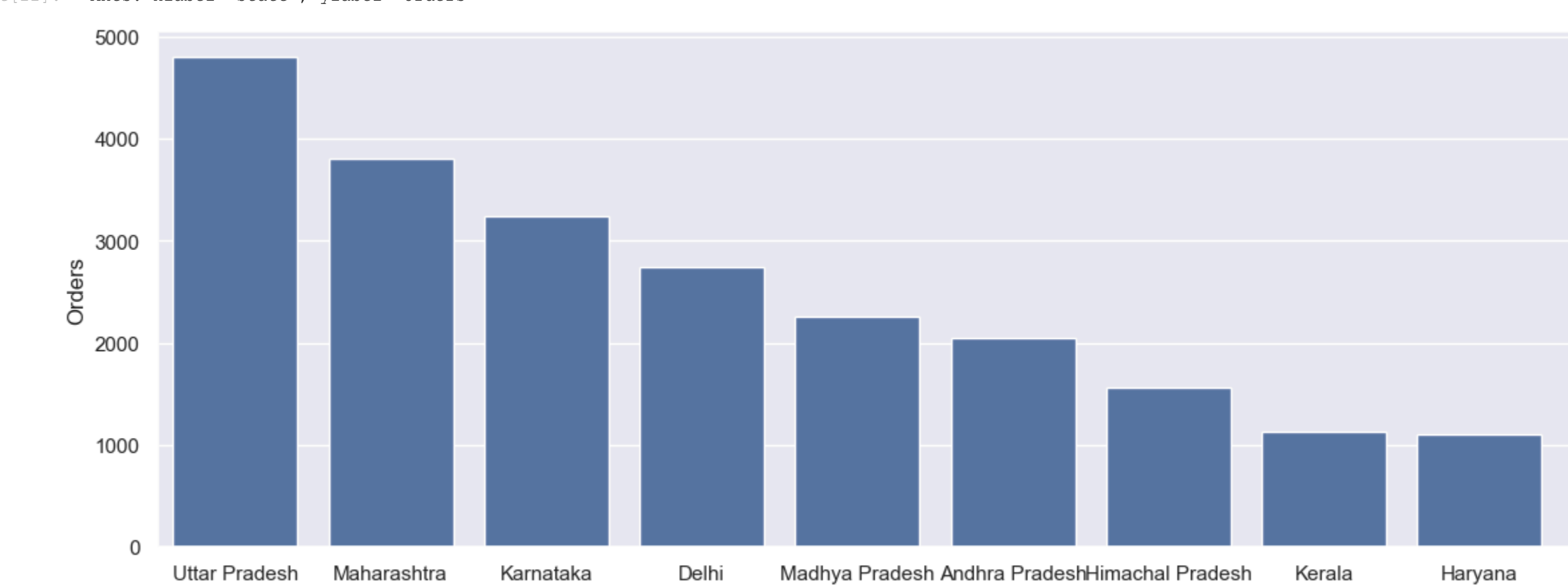
state

In [21]:
```python
df.columns
```

Out[21]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

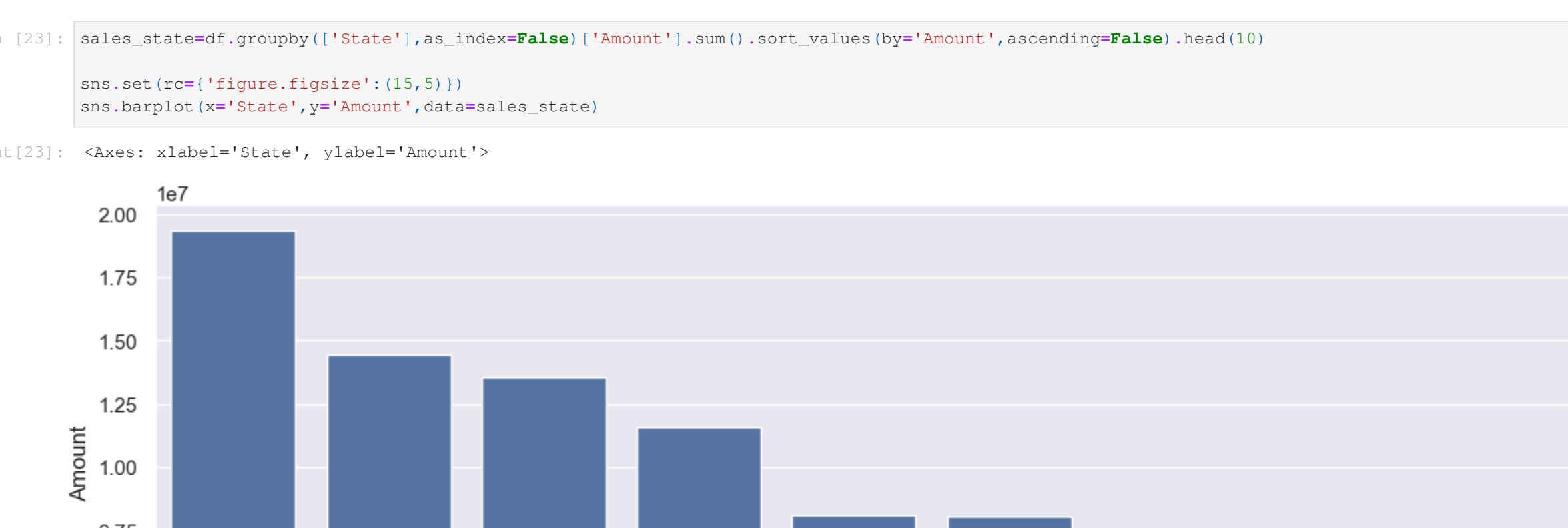
In [22]:
```python
sales_state=df.groupby(['State'],as_index=False)['Orders'].sum().sort_values(by='Orders',ascending=False).head(10)

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

Out[22]:
```
<Axes: xlabel='State', ylabel='Orders'>
```



In [23]:
```python
sales_state=df.groupby(['State'],as_index=False)['Amount'].sum().sort_values(by='Amount',ascending=False).head(10)

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

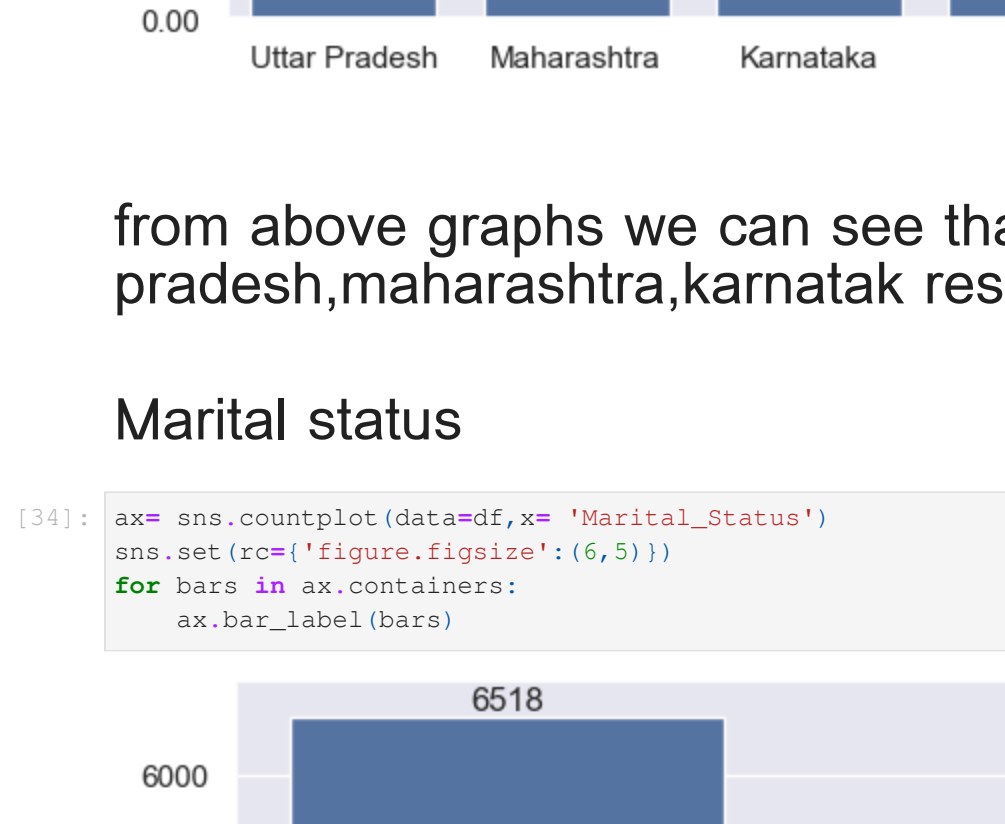
Out[23]:
```
<Axes: xlabel='State', ylabel='Amount'>
```



from above graphs we can see that most of the orders and total amount are from uttar pradesh,maharashtra,karnatak respectively

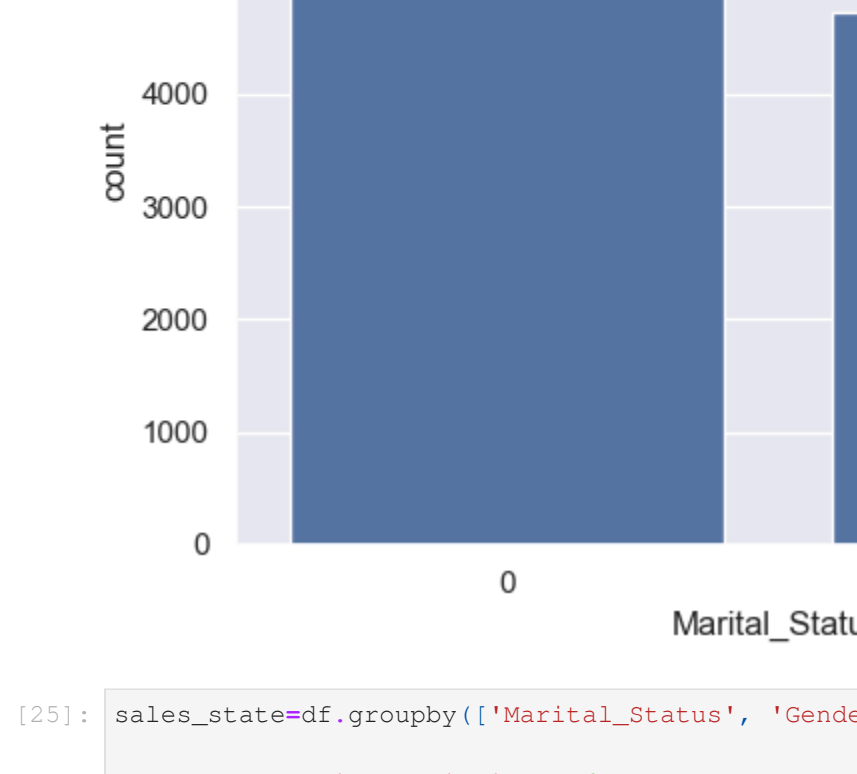
Marital status

In [24]:
```python
ax= sns.countplot(data=df,x= 'Marital_Status')
sns.set(rc={'figure.figsize':(6,5)})
for bars in ax.containers:
    ax.bar_label(bars)
```



In [25]:
```python
sales_state=df.groupby(['Marital_Status', 'Gender'],as_index=False)['Amount'].sum().sort_values(by='Amount',ascending=False)

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

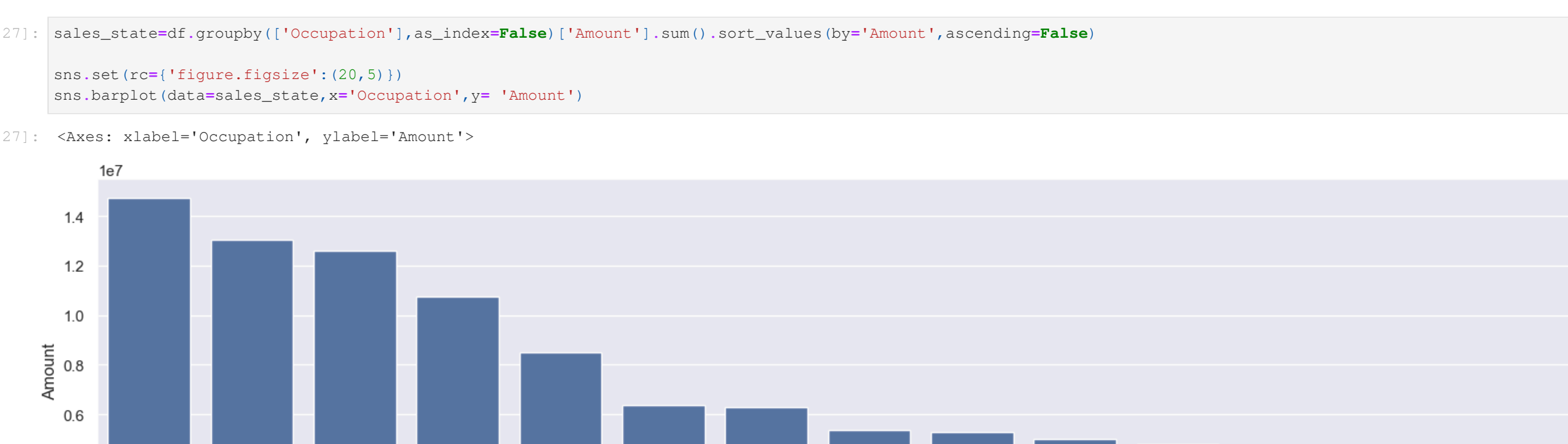
Out[25]:
```
<Axes: xlabel='Marital_Status', ylabel='Amount'>
```



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

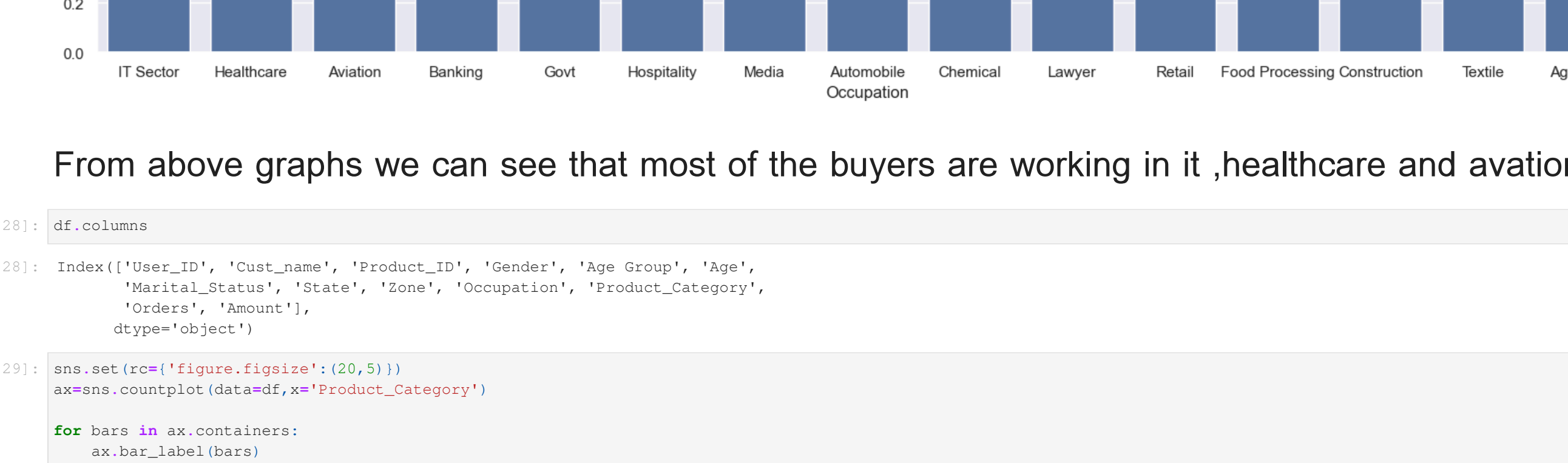
Occupation

In [26]:
```python
sns.set(rc={'figure.figsize':(20,5)})
ax=sns.countplot(data=df,x='Occupation')

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



In [27]:
```python
sales_state=df.groupby(['Occupation'],as_index=False)['Amount'].sum().sort_values(by='Amount',ascending=False)

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

Out[27]:
```
<Axes: xlabel='Occupation', ylabel='Amount'>
```



From above graphs we can see that most of the buyers are working in it ,healthcare and avation.

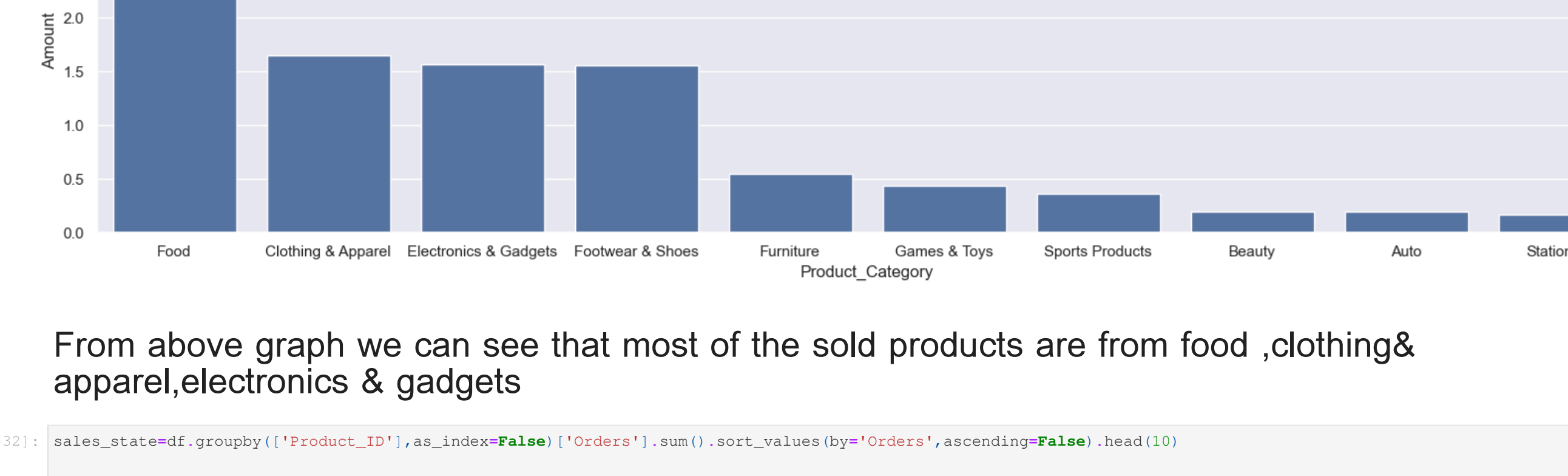
In [28]:
```python
df.columns
```

Out[28]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

In [29]:
```python
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 [30]:
```python
sales_state=df.groupby(['Product_Category'],as_index=False)['Amount'].sum().sort_values(by='Amount',ascending=False).head(10)

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

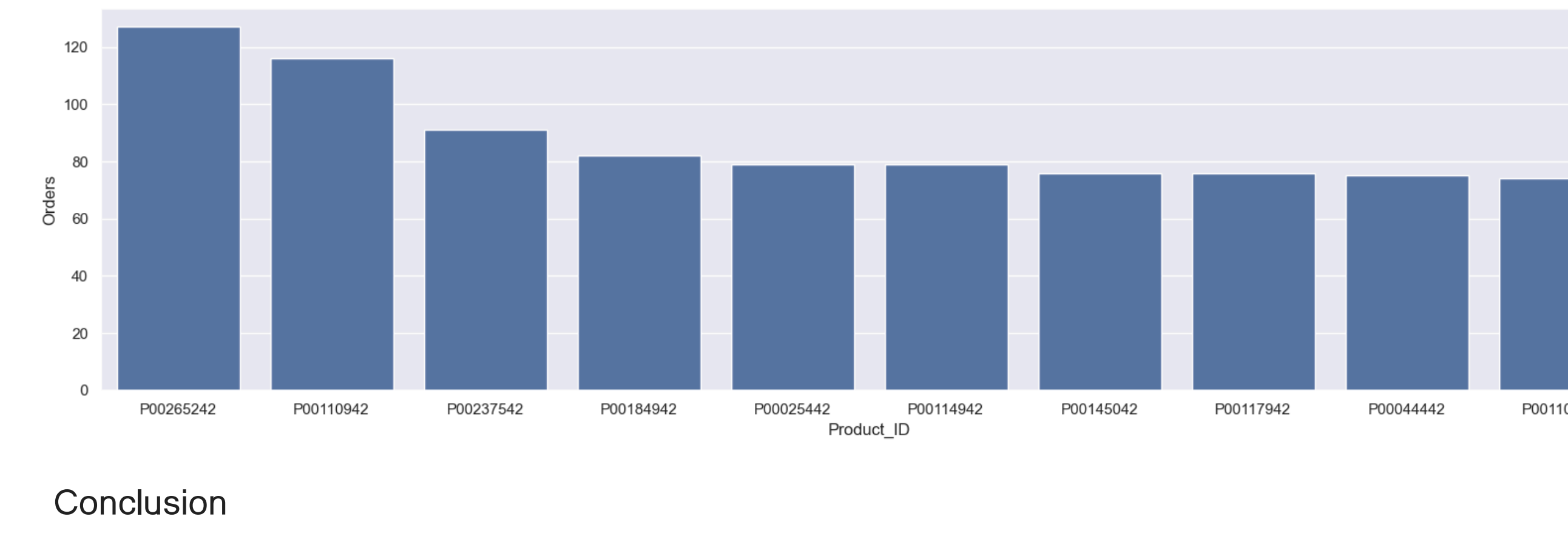
Out[30]:
```
<Axes: xlabel='Product_Category', ylabel='Amount'>
```



From above graph we can see that most of the sold products are from food ,clothing& apparel,electronics & gadgets

In [32]:
```python
sales_state=df.groupby(['Product_ID'],as_index=False)['Orders'].sum().sort_values(by='Orders',ascending=False).head(10)

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

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



Conclusion

Married woman age group 26-35 from up,maharastra,karnatak working on it,healthcare and aviation are more buyer likely to buy products from food clothing and electronics category.

In [ ]: