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

In [3]:
```
df=pd.read_csv("C:\Data Analytics\Diwali_Sales\Diwali Sales Data.csv",encod
```

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

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

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
```

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

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 [8]:
```
df.dropna(inplace=True)
```

In [9]:
```
pd.isnull(df).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              0
dtype: int64
```

In [10]:
```
pd.isnull(df)
```

Out[10]:

	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

11239 rows × 13 columns

In [11]:
```
df['Amount']=df['Amount'].astype('int')
```

In [12]:
```
df['Amount'].dtype
```

Out[12]:
```
dtype('int32')
```

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

Out[13]:

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

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

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

In [15]:
```
df.rename(columns={'Marital_Status':'Are you Married?'})
```

Out[15]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Are you Married?	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
...	...	...	...	...	...	...	...	...
11246	1000695	Manning	P00296942	M	18-25	19	1	Maharashtra
11247	1004089	Reichenbach	P00171342	M	26-35	33	0	Haryana
11248	1001209	Oshin	P00201342	F	36-45	40	0	Madhya Pradesh

In [16]:
```
df.describe()
```

Out[16]:

	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 [17]:
```
df[['Age','Orders','Amount']].describe()
```

Out[17]:

	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

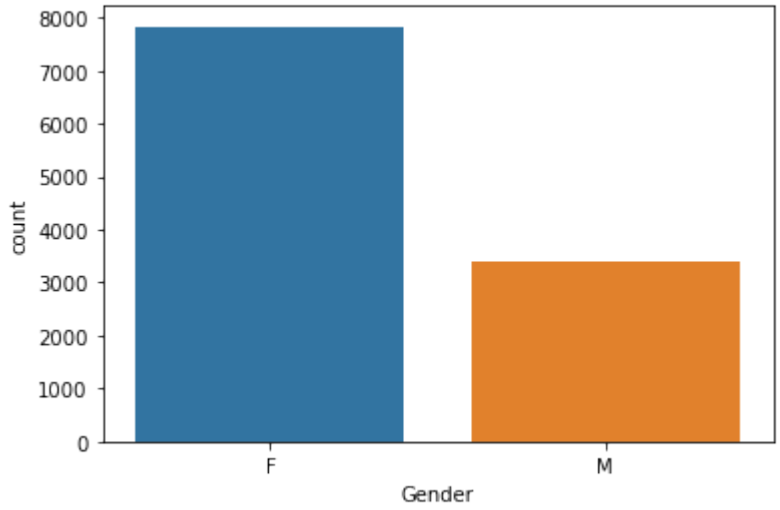
Exploring Data

In [18]:
```
df.columns
```

Out[18]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Categor
y',
       'Orders', 'Amount'],
      dtype='object')
```

In [19]:
```
sns.countplot(x='Gender',data=df)
```

Out[19]:
```
<AxesSubplot:xlabel='Gender', ylabel='count'>
```



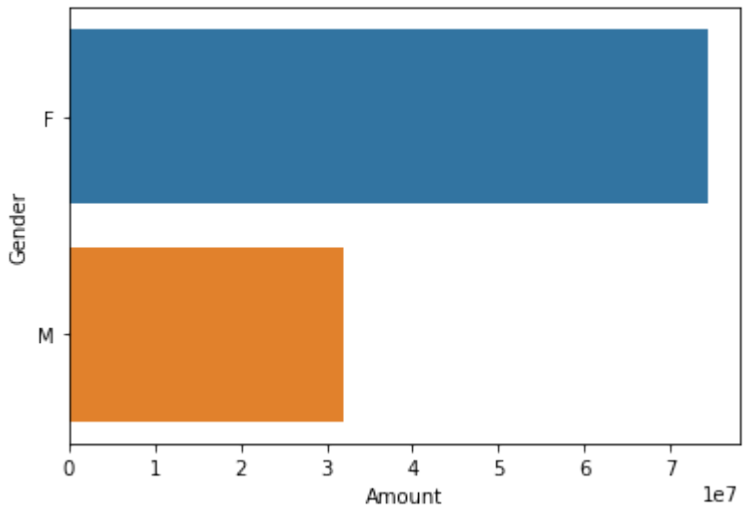
In [20]:
```
gen=df.groupby(['Gender'],as_index=False)['Amount'].sum().sort_values(by='A
gen
```

Out[20]:

	Gender	Amount
0	F	74335853
1	M	31913276

In [21]:
```
sns.barplot(y='Gender',x='Amount',data=gen)
```

Out[21]:
```
<AxesSubplot:xlabel='Amount', ylabel='Gender'>
```



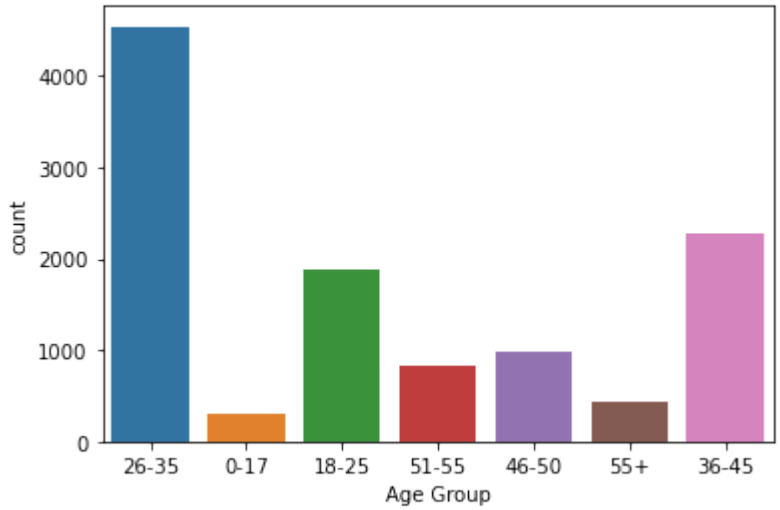
Female has dominated over male in spending money.Also, there are more females buyers than male.

In [22]:
```
df.columns
```

Out[22]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Categor
y',
       'Orders', 'Amount'],
      dtype='object')
```

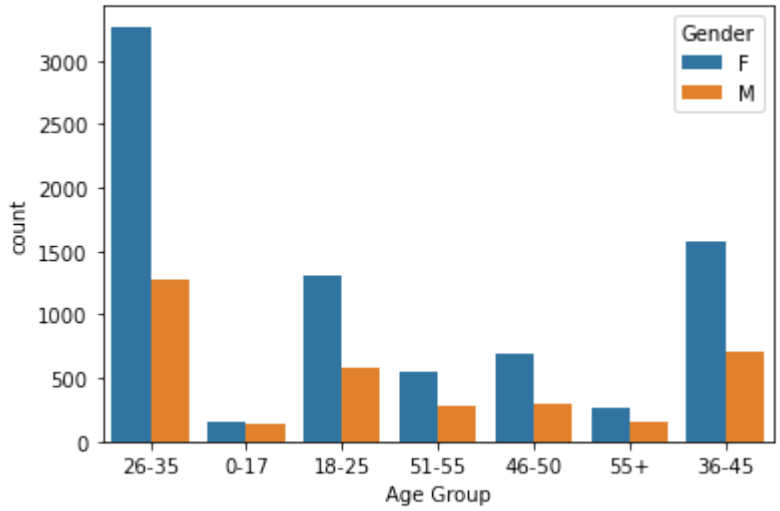
In [23]:
```
sns.countplot(x='Age Group',data=df)
```

Out[23]:
```
<AxesSubplot:xlabel='Age Group', ylabel='count'>
```



In [24]:
```
sns.countplot(x='Age Group',data=df,hue='Gender')
```

Out[24]:
```
<AxesSubplot:xlabel='Age Group', ylabel='count'>
```



In every age group, Female has dominated.

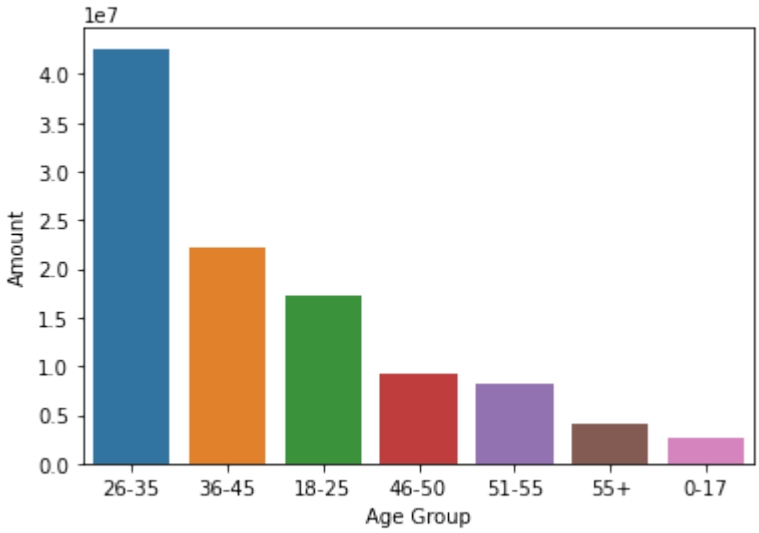
In [25]:
```
agrp=df.groupby(['Age Group'],as_index=False)['Amount'].sum().sort_values(b
agrp
```

Out[25]:

	Age Group	Amount
2	26-35	42613442
3	36-45	22144994
1	18-25	17240732
4	46-50	9207844
5	51-55	8261477
6	55+	4080987
0	0-17	2699653

In [26]:
```
sns.barplot(x='Age Group',y='Amount',data=agrp)
```

Out[26]:
```
<AxesSubplot:xlabel='Age Group', ylabel='Amount'>
```



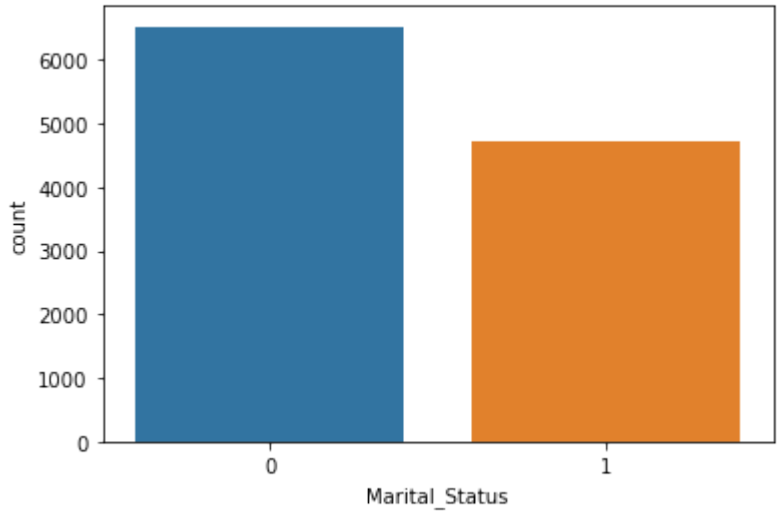
Most of the buyers are from the age group 26-35

In [27]:
```
df.columns
```

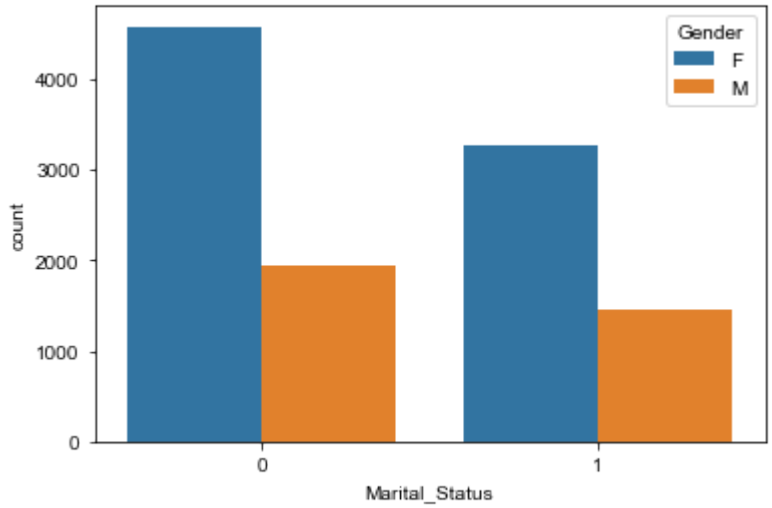
Out[27]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Categor
y',
       'Orders', 'Amount'],
      dtype='object')
```

In [28]:
```
sns.countplot(x='Marital_Status',data=df)
```

Out[28]:
```
<AxesSubplot:xlabel='Marital_Status', ylabel='count'>
```



In [29]:
```
sns.countplot(x='Marital_Status',data=df,hue='Gender')
sns.set(rc={'figure.figsize':(7,5)})
```



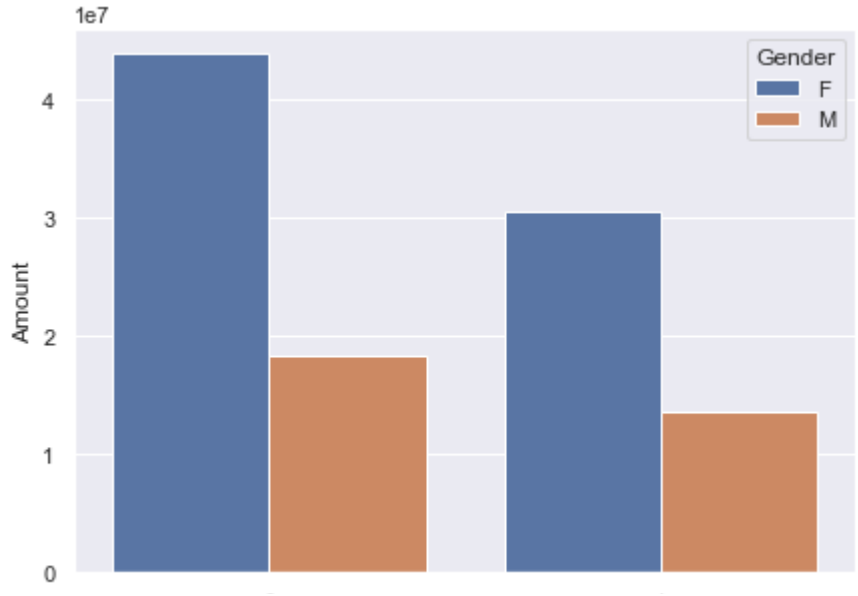
In [30]:
```
msts=df.groupby(['Marital_Status','Gender'],as_index=False)['Amount'].sum()
msts
```

Out[30]:

	Marital_Status	Gender	Amount
0	0	F	43786646
2	1	F	30549207
1	0	M	18338738
3	1	M	13574538

In [31]:
```
sns.barplot(x='Marital_Status',y='Amount',data=msts,hue='Gender')
```

Out[31]:
```
<AxesSubplot:xlabel='Marital_Status', ylabel='Amount'>
```

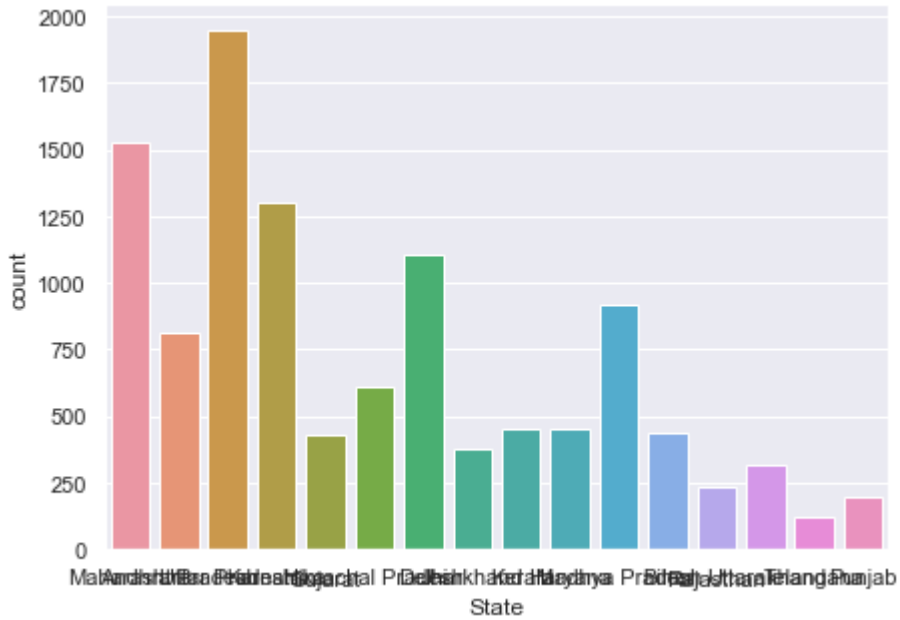


Most of the buyers are Married Female. Also, they have dominated in purchasing.

In [32]:
```
df.columns
```

Out[32]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Categor
y',
       'Orders', 'Amount'],
      dtype='object')
```

In [33]:
```
sns.countplot(x='State',data=df)
sns.set(rc={'figure.figsize':(35,10)})
```

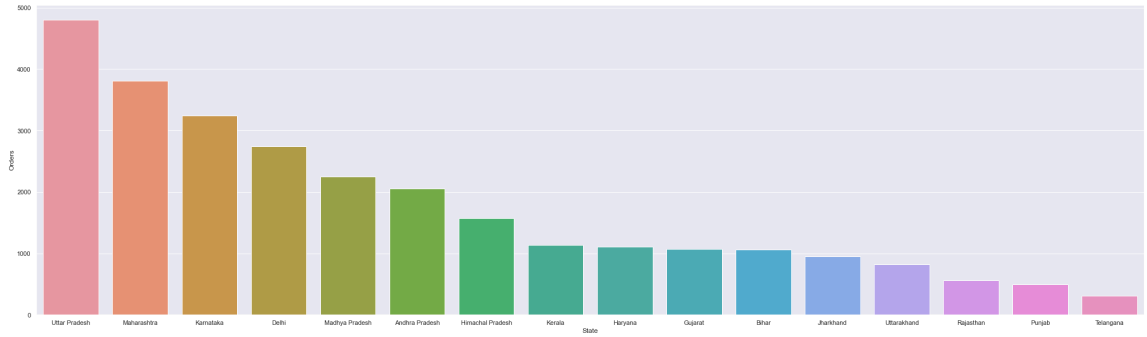


In [34]:
```
st=df.groupby(['State'],as_index=False)['Orders'].sum().sort_values(by='Ord
st
```

Out[34]:

	State	Orders
14	Uttar Pradesh	4807
10	Maharashtra	3810
7	Karnataka	3240
2	Delhi	2740
9	Madhya Pradesh	2252
0	Andhra Pradesh	2051
5	Himachal Pradesh	1568
8	Kerala	1137
4	Haryana	1109
3	Gujarat	1066
1	Bihar	1062
6	Jharkhand	953

In [35]:
```
sns.barplot(x='State',y='Orders',data=st)
sns.set(rc={'figure.figsize':(10,5)})
```



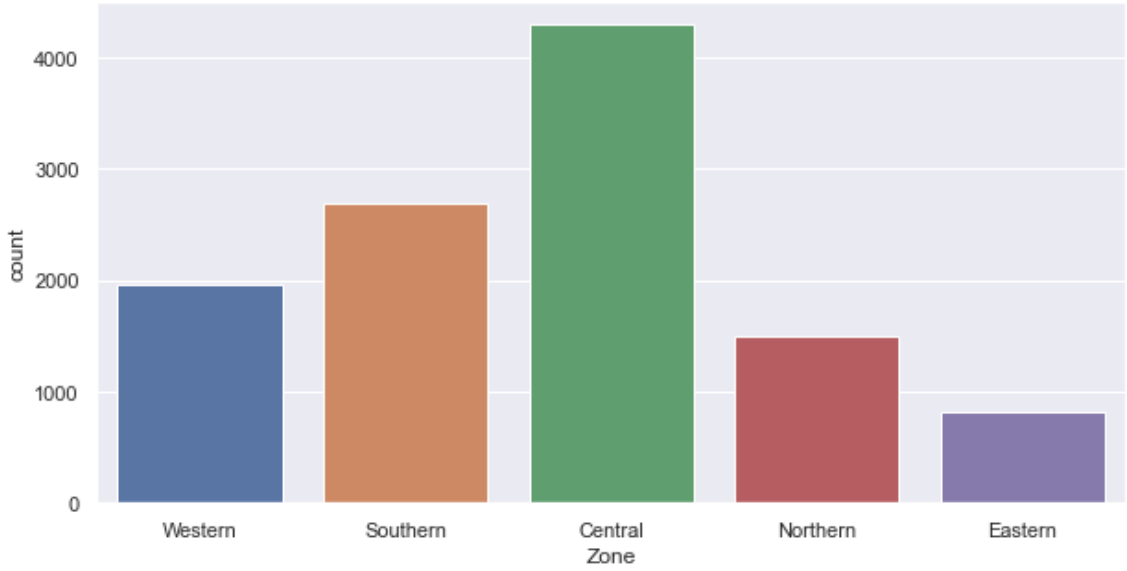
Top 4 States on the basis of Orders & Sales:

1.Uttar Pradesh 2.Maharastra 3.Karnataka 4.Delhi

In [36]:
```
df.columns
```

Out[36]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Categor
y',
       'Orders', 'Amount'],
      dtype='object')
```

In [37]:
```
sns.countplot(x='Zone',data=df)
sns.set(rc={'figure.figsize':(5,2)})
```



In [38]:
```
df.groupby(['Zone'],as_index=False)['Orders'].sum().sort_values(by='Orders'
```

Out[38]:

	Zone	Orders
0	Central	10623
3	Southern	6740
4	Western	4876
2	Northern	3727
1	Eastern	2015

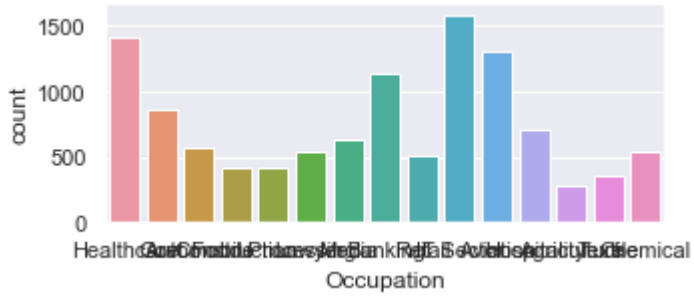
Central Zone has the highest number of orders.

In [39]:
```
df.columns
```

Out[39]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Categor
y',
       'Orders', 'Amount'],
      dtype='object')
```

In [40]:
```
sns.countplot(x='Occupation',data=df)
```

Out[40]:
```
<AxesSubplot:xlabel='Occupation', ylabel='count'>
```



In [41]:
```
sns.countplot(x='Occupation',data=df,hue='Gender')
```

Out[41]:
```
<AxesSubplot:xlabel='Occupation', ylabel='count'>
```

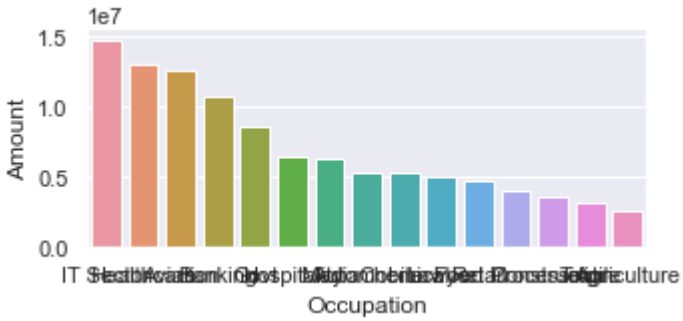
In [42]:
```
oc=df.groupby(['Occupation'],as_index=False)['Amount'].sum().sort_values(by
oc
```

Out[42]:

	Occupation	Amount
10	IT Sector	14755079
8	Healthcare	13034586
2	Aviation	12602298
3	Banking	10770610
7	Govt	8517212
9	Hospitality	6376405
12	Media	6295832
1	Automobile	5368596
4	Chemical	5297436
11	Lawyer	4981665
13	Retail	4783170
6	Food Processing	4070670

In [44]:
```
sns.barplot(x='Occupation',y='Amount',data=oc)
```

Out[44]:
```
<AxesSubplot:xlabel='Occupation', ylabel='Amount'>
```



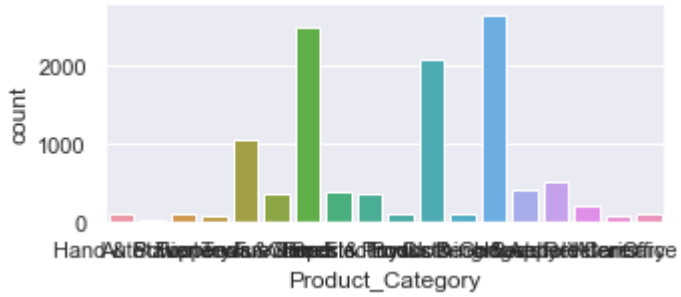
Most of the buyers are from IT Ssector, Healthcare, Aviation and Banking.

In [45]:
```
df.columns
```

Out[45]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Categor
y',
       'Orders', 'Amount'],
      dtype='object')
```

In [46]:
```
sns.countplot(x='Product_Category',data=df)
```

Out[46]:
```
<AxesSubplot:xlabel='Product_Category', ylabel='count'>
```



In [47]:
```
df.groupby(['Product_Category'],as_index=False)['Orders'].sum().sort_values
```

Out[47]:

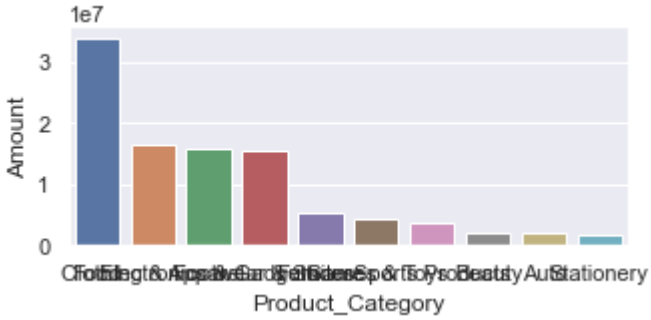
	Product_Category	Orders
3	Clothing & Apparel	6634
6	Food	6110
5	Electronics & Gadgets	5226
7	Footwear & Shoes	2646
11	Household items	1331

In [48]:
```
prod_a=df.groupby(['Product_Category'],as_index=False)['Amount'].sum().sort
prod_a
```

Out[48]:

	Product_Category	Amount
6	Food	33933883
3	Clothing & Apparel	16495019
5	Electronics & Gadgets	15643846
7	Footwear & Shoes	15575209
8	Furniture	5440051
9	Games & Toys	4331694
14	Sports Products	3635933
1	Beauty	1959484
0	Auto	1958609
15	Stationery	1676051

In [49]:
```
sns.barplot(x='Product_Category',y='Amount',data=prod_a)
sns.set(rc={'figure.figsize':(10,5)})
```



Top 3 sell products

1.Clothing & Apparel
2.Food
3. Electronic & gadgets

In [55]:
```
df.groupby(['Product_Category','Gender'],as_index=False)['Orders'].sum().so
```

Out[55]:

	Product_Category	Gender	Orders
6	Clothing & Apparel	F	4648
12	Food	F	4406
10	Electronics & Gadgets	F	3682
7	Clothing & Apparel	M	1986
14	Footwear & Shoes	F	1925

In the top sold products, Female is dominated

In [56]:
```
df.columns
```

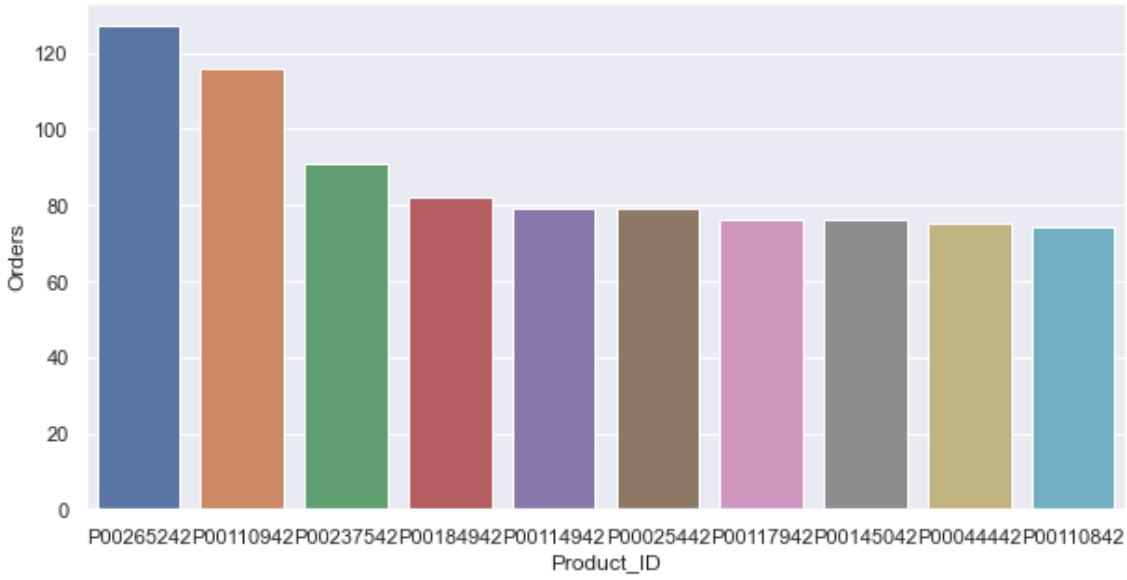
Out[56]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Categor
y',
       'Orders', 'Amount'],
      dtype='object')
```

In [57]:
```
prod_id=df.groupby(['Product_ID'],as_index=False)['Orders'].sum().sort_valu
prod_id
```

Out[57]:

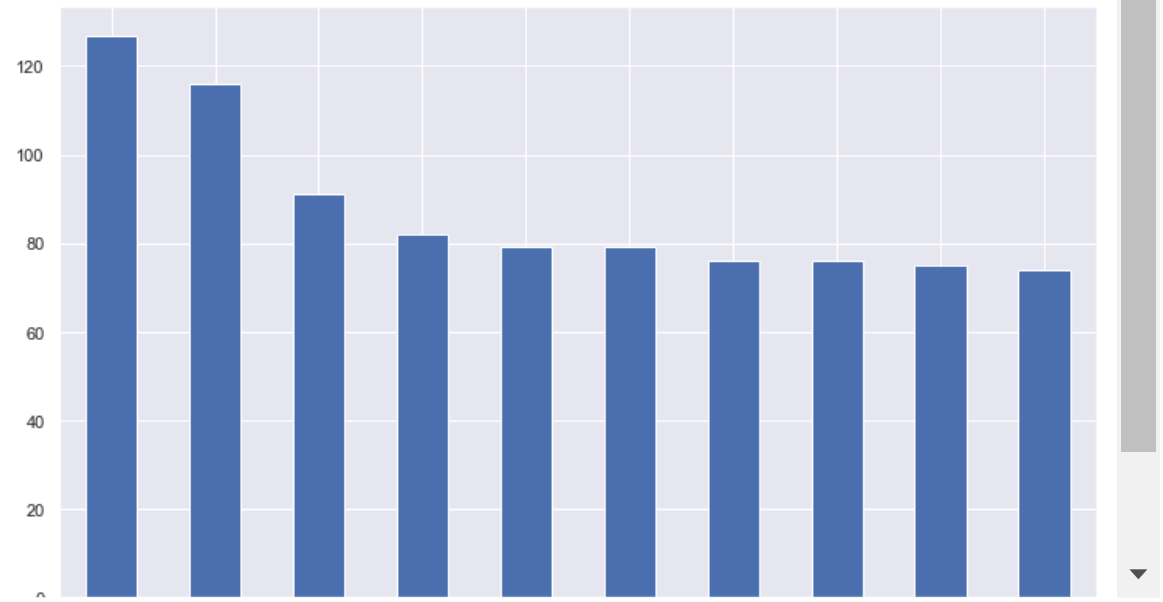
	Product_ID	Orders
1679	P00265242	127
644	P00110942	116
1504	P00237542	91
1146	P00184942	82
679	P00114942	79
171	P00025442	79
708	P00117942	76
888	P00145042	76
298	P00044442	75
643	P00110842	74

In [58]:
```
sns.barplot(x='Product_ID',y='Orders',data=prod_id)
sns.set(rc={'figure.figsize':(12,7)})
```



In [59]:
```
df.groupby('Product_ID')['Orders'].sum().nlargest(10).sort_values(ascending
```

Out[59]:
```
<AxesSubplot:xlabel='Product_ID'>
```



Conclusion

People of age group 26-35 years who are married are from UP, Maharastra and Karnataka working in IT Sector, Healthcare and Aviation are most likely to buy products from Clothing & Apparel, Food and Electronic & Gadgets Category.