In [3]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
```

In [4]:
```
df = pd.read_csv('C:\\Users\\kushw\\Downloads\\Diwali Sales Data.csv', encoding='unicode_escape')
```

In [5]:
```
df.shape
```

Out[5]: (11251, 15)

In [6]:
```
df.head(10)
```

Out[6]:

User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Category	Orders	A
1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto	1	23
1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	Auto	3	23
1001990	Bindu	P00118542	F	26-35	35	1	Uttar Pradesh	Central	Automobile	Auto	3	23
1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	Auto	2	23
1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	Auto	2	23
1000588	Joni	P00057942	M	26-35	28	1	Himachal Pradesh	Northern	Food Processing	Auto	1	23
1001132	Balk	P00018042	F	18-25	25	1	Uttar Pradesh	Central	Lawyer	Auto	4	23
1002092	Shivangi	P00273442	F	55+	61	0	Maharashtra	Western	IT Sector	Auto	1	
1003224	Kushal	P00205642	M	26-35	35	0	Uttar Pradesh	Central	Govt	Auto	2	23
1003650	Ginny	P00031142	F	26-35	26	1	Andhra Pradesh	Southern	Media	Auto	4	23

In [7]:
```
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 [8]:
```
df.drop(['Status','unnamed1'],axis=1,inplace=True)
```

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

Out[9]:

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

11251 rows × 13 columns

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

Out[10]:
```
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 [12]:
```
df.shape
```

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

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

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

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

In [15]:
```
pd.isnull(df).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]:
```
df['Amount'] = df['Amount'].astype('int')
```

In [17]:
```
df['Amount'].dtypes
```

Out[17]:
```
dtype('int32')
```

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

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

In [19]:
```
df.rename(columns={'Marital_Status':'Shaadi'})
```

Out[19]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Shaadi	State	Zone	Occupation	Product_Category	Orders
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto	1
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	Auto	3
2	1001990	Bindu	P00118542	F	26-35	35	1	Uttar Pradesh	Central	Automobile	Auto	3
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	Auto	2
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	Auto	2
...	...	...	...	...	...	...	...	...	...	...	...	...
11246	1000695	Manning	P00296942	M	18-25	19	1	Maharashtra	Western	Chemical	Office	4
11247	1004089	Reichenbach	P00171342	M	26-35	33	0	Haryana	Northern	Healthcare	Veterinary	3
11248	1001209	Oshin	P00201342	F	36-45	40	0	Madhya Pradesh	Central	Textile	Office	4
11249	1004023	Noonan	P00059442	M	36-45	37	0	Karnataka	Southern	Agriculture	Office	3
11250	1002744	Brumley	P00281742	F	18-25	19	0	Maharashtra	Western	Healthcare	Office	3

11239 rows × 13 columns

In [20]:
```
df.describe()
```

Out[20]:

	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 [21]:
```
df[['Age','Orders','Amount']].describe()
```

Out[21]:

	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

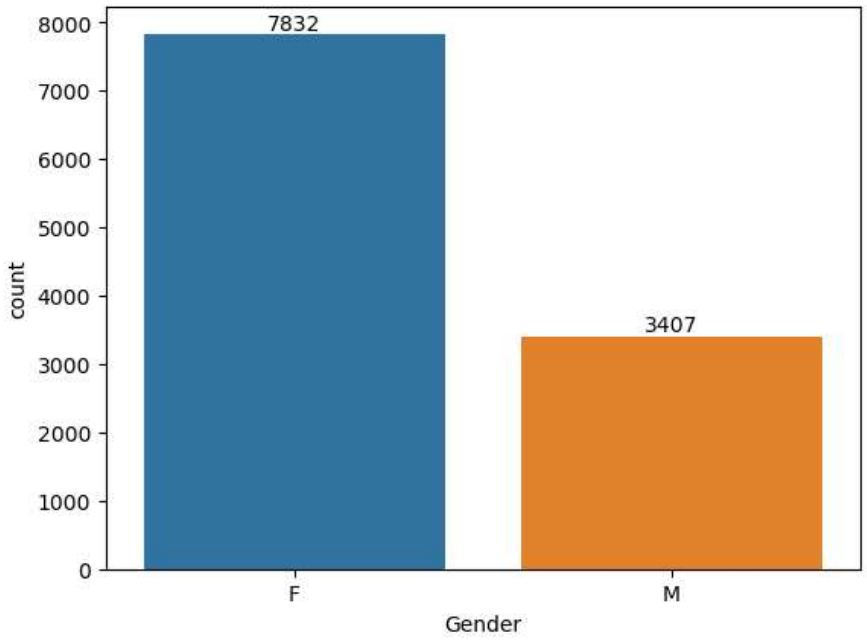
Exploratory Data Analysis

Gender

In [26]:
```
df.columns
```

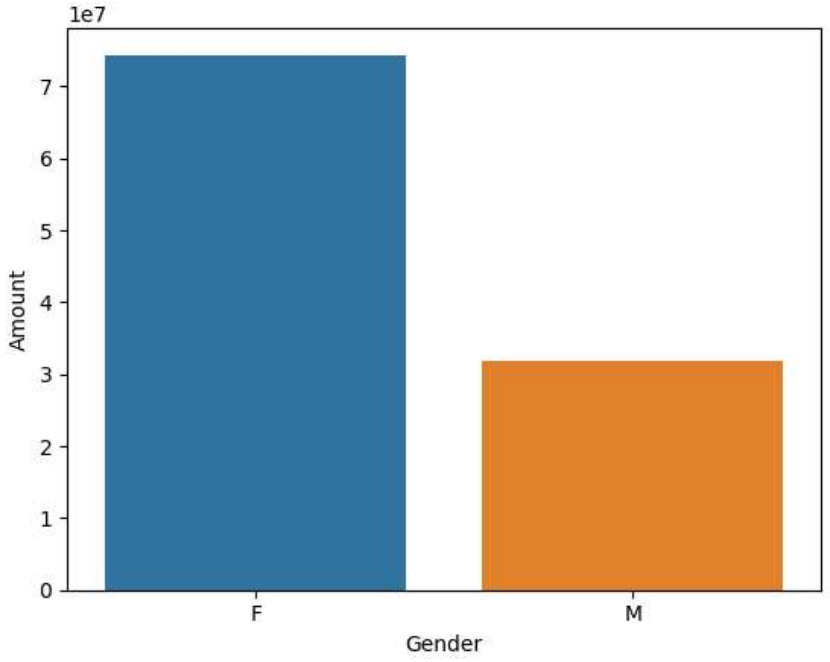
Out[26]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

In [29]:
```
ax = sns.countplot(x = 'Gender', data = df)
for bars in ax.containers:
    ax.bar_label(bars)
```



In [32]:
```
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[32]:
```
<Axes: xlabel='Gender', ylabel='Amount'>
```



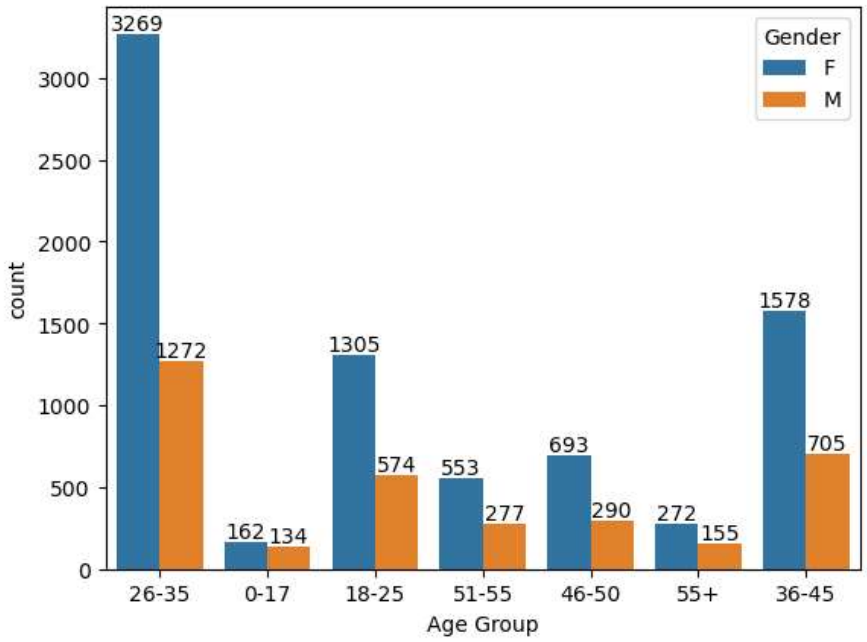
Insights from the data : Most of the buyers are female and even the purchasing power of females are greater than men

Age

In [33]:
```
df.columns
```

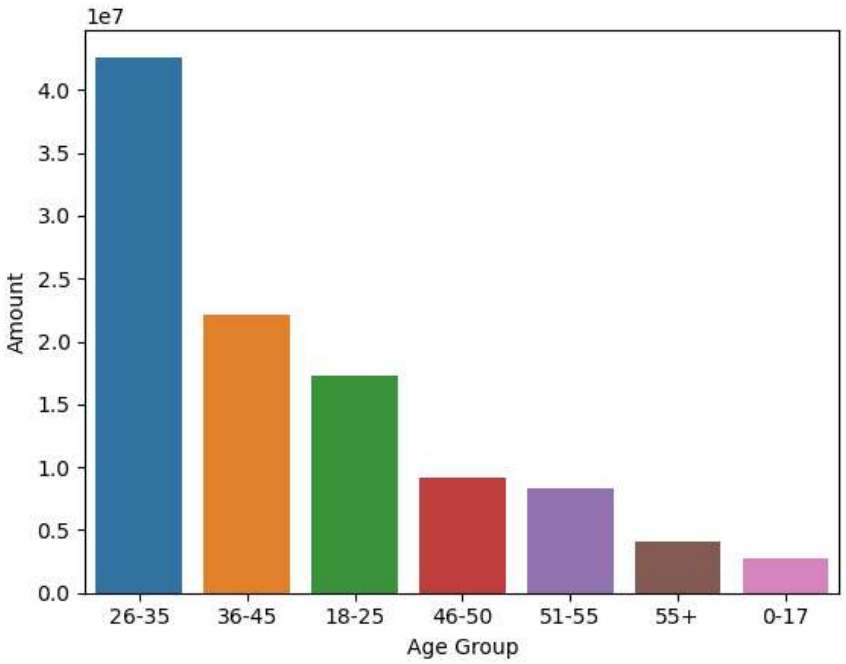
Out[33]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

In [36]:
```
ax = sns.countplot(data=df, x = 'Age Group', hue = 'Gender')
for bar in ax.containers:
    ax.bar_label(bar)
```



In [39]:
```
sales_age = df.groupby(['Age Group'], as_index=False)['Amount'].sum().sort_values(by='Amount', ascending = Fals

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

Out[39]: <Axes: xlabel='Age Group', ylabel='Amount'>



Insights from the data : Most of the buyers are of the age group between 26-35 years female

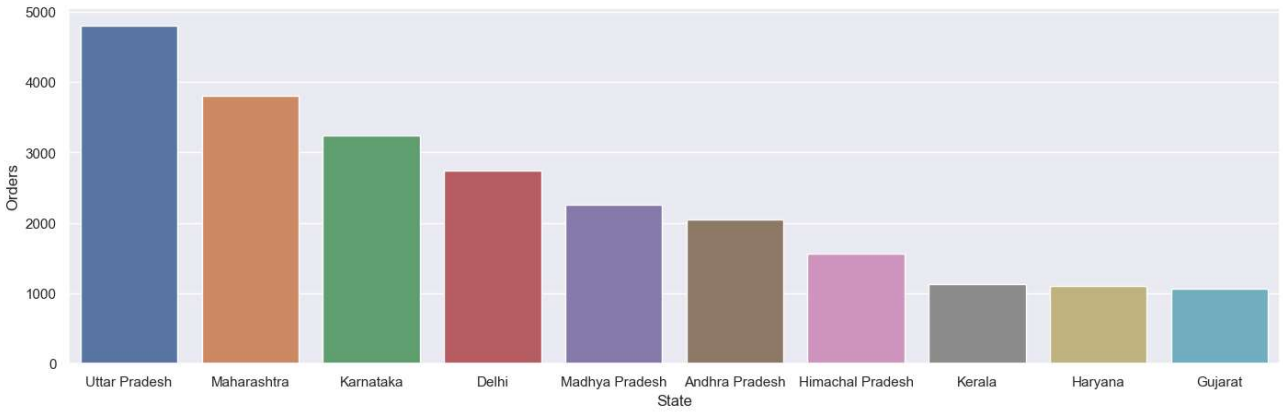
State

In [42]:
```
df.columns
```

Out[42]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

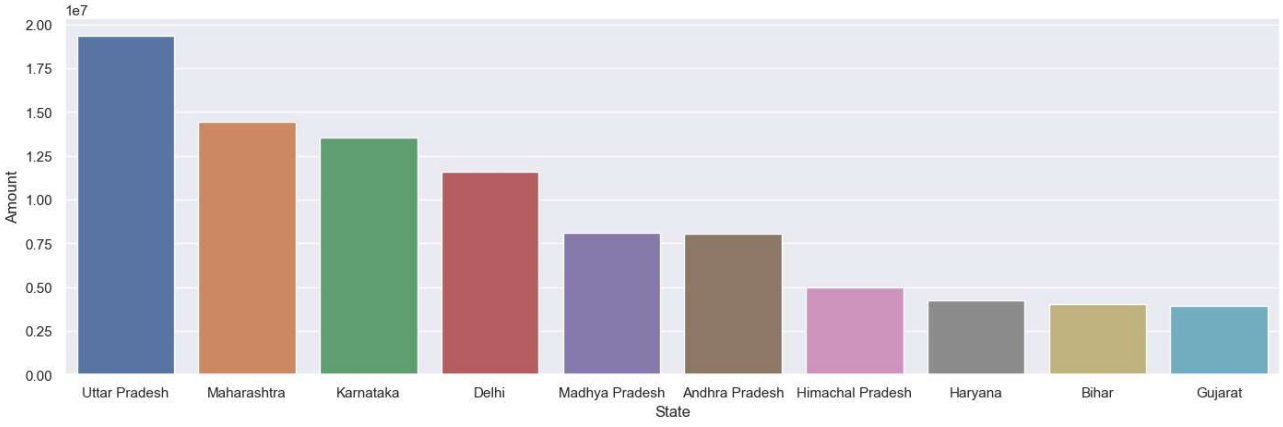
In [47]:
```
sales_state = df.groupby(['State'], as_index = False)['Orders'].sum().sort_values(by='Orders', ascending = Fals
sns.set(rc = {'figure.figsize':(17,5)})
sns.barplot(data = sales_state, x = 'State', y='Orders')
```

Out[47]:
```
<Axes: xlabel='State', ylabel='Orders'>
```



In [48]:
```
sales_state = df.groupby(['State'], as_index = False)['Amount'].sum().sort_values(by='Amount', ascending = Fals
sns.set(rc = {'figure.figsize':(17,5)})
sns.barplot(data = sales_state, x = 'State', y='Amount')
```

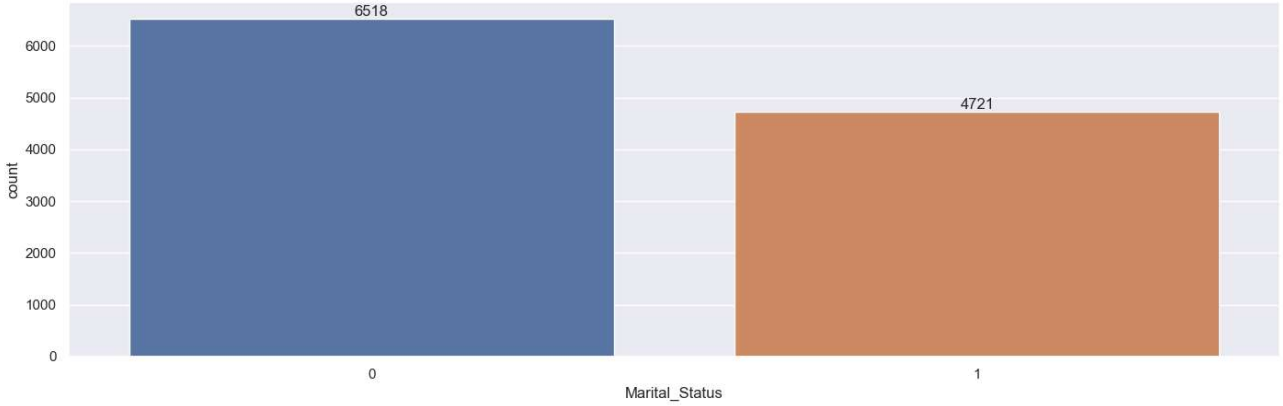
Out[48]:
```
<Axes: xlabel='State', ylabel='Amount'>
```



Insights from the data : Most of the orders & total sales/amount are from Uttar Pradesh, Maharashtra & Karnataka.

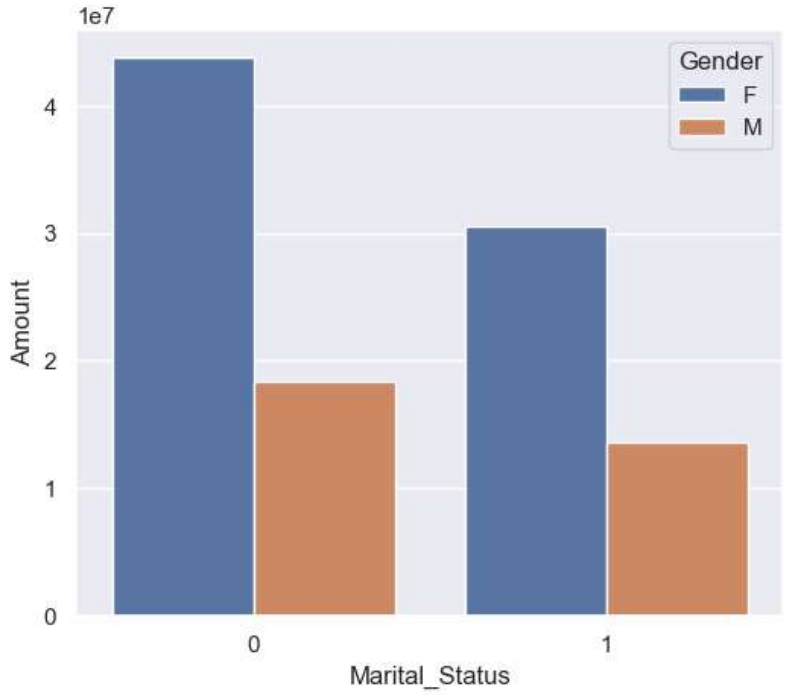
Marital Status

In [49]:
```
ax = sns.countplot(data = df, x = 'Marital_Status')

sns.set(rc={'figure.figsize':(8,5)})
for bars in ax.containers:
    ax.bar_label(bars)
```



In [50]:
```
sales_state = df.groupby(['Marital_Status', 'Gender'], as_index=False)['Amount'].sum().sort_values(by='Amount',

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

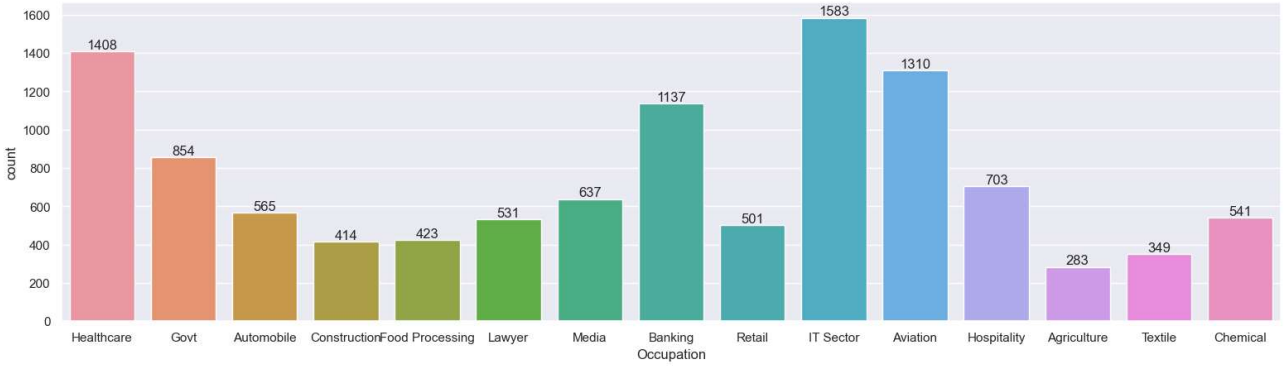
Out[50]: <Axes: xlabel='Marital_Status', ylabel='Amount'>



Insights from the data : Most of the buyers are married (women) & they have high purchasing power.

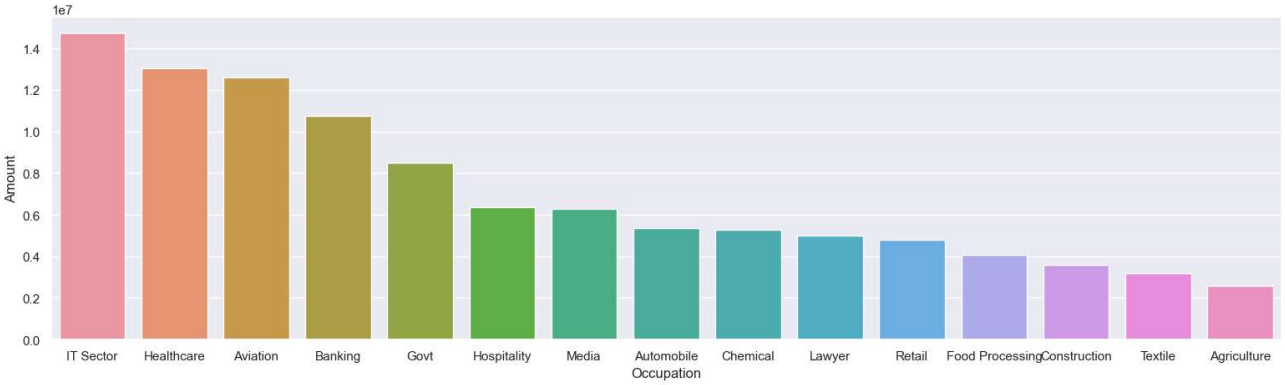
Occupation

In [53]:
```
sns.set(rc={'figure.figsize':(19,5)})
ax = sns.countplot(data = df, x = 'Occupation')

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



In [54]:
```
sales_state = df.groupby(['Occupation'], as_index=False)['Amount'].sum().sort_values(by='Amount', ascending=Fal

sns.set(rc={'figure.figsize':(19,5)})
sns.barplot(data = sales_state, x = 'Occupation',y= 'Amount')
```

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



Insights from the data : Most of the buyers are working in the IT Sector, Aviation & Healthcare.

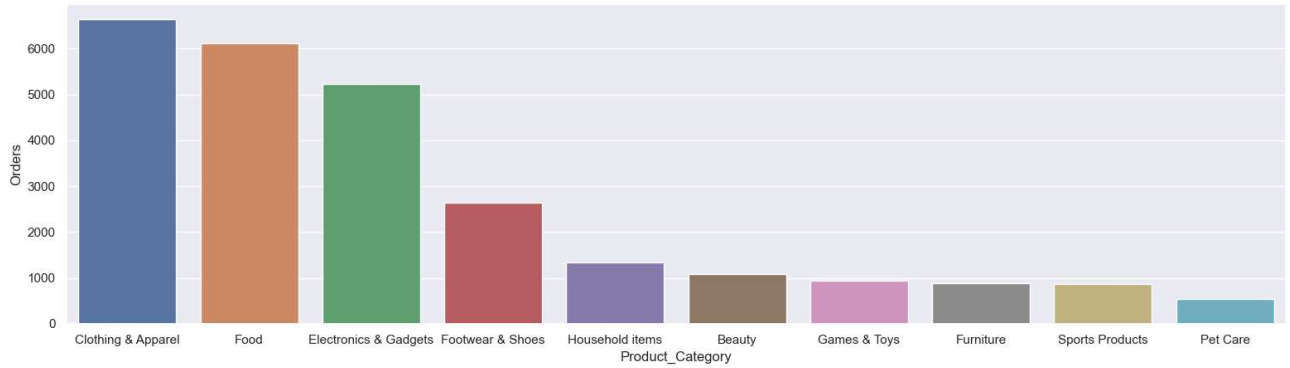
Product Category

In [63]:
```
df.columns
```

Out[63]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

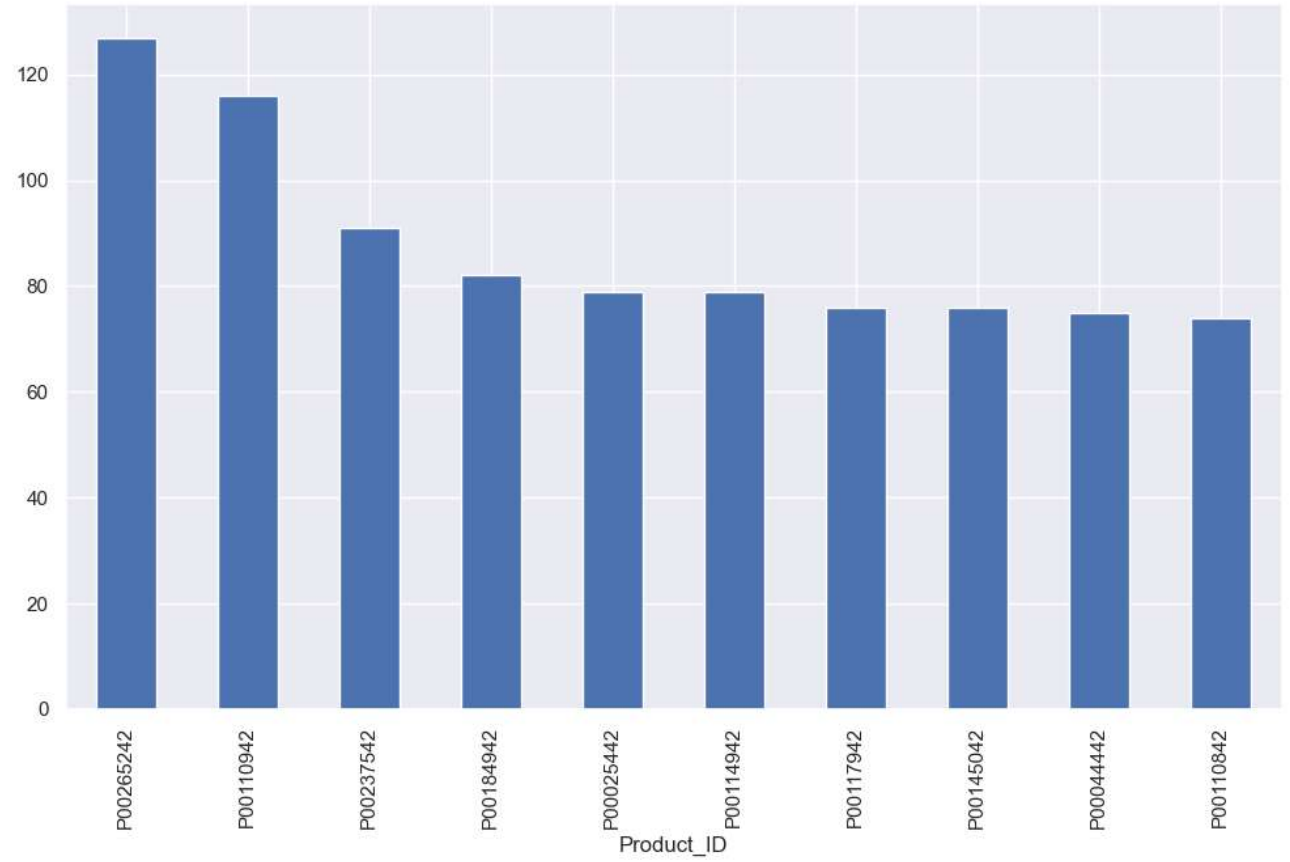
In [70]:
```
product = df.groupby(['Product_Category'], as_index = False)['Orders'].sum().sort_values(by='Orders', ascending
sns.set(rc = {'figure.figsize':(19,5)})
sns.barplot(data = product, x = 'Product_Category', y='Orders')
```

Out[70]: <Axes: xlabel='Product_Category', ylabel='Orders'>



In [68]:
```
fig1, ax1 = plt.subplots(figsize=(12,7))
df.groupby('Product_ID')['Orders'].sum().nlargest(10).sort_values(ascending=False).plot(kind='bar')
```

Out[68]: <Axes: xlabel='Product_ID'>



Conclusion

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