Diwali Sales Data Analysis

In [66]:
```
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import pandas as pd
import numpy as np
```

In [45]:
```
df = pd.read_csv(r"C:\Users\RONI\Desktop\New folder\Sales Data of Diwali.csv", encoding= 'unicode_escape')
# to avoid encoding error, use 'unicode escpae'
```

In [46]:
```
df.shape
```

Out[46]:
```
(11251, 15)
```

In [47]:
```
df.head()
```

Out[47]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Categ
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	A
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	A
2	1001990	Bindu	P00118542	F	26-35	35	1	Uttar Pradesh	Central	Automobile	A
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	A
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	A

In [48]:
```
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 [50]:
```
# drop blank column
df.drop(['Status','unnamed1'], axis =1 , inplace = True)
```

In [52]:
```
#check for null values
pd.isnull(df).sum()
```

Out[52]:
```
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 [55]:
```
#drop null values
df.dropna(inplace = True)
```

In [58]:
```
#change data type
df['Amount'] = df['Amount'].astype('int')
```

In [61]:
```
df.columns
```

Out[61]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

use describe() for specific columns df[['Age', 'Orders','Amount']].describe()

Exploratory Data Analysis

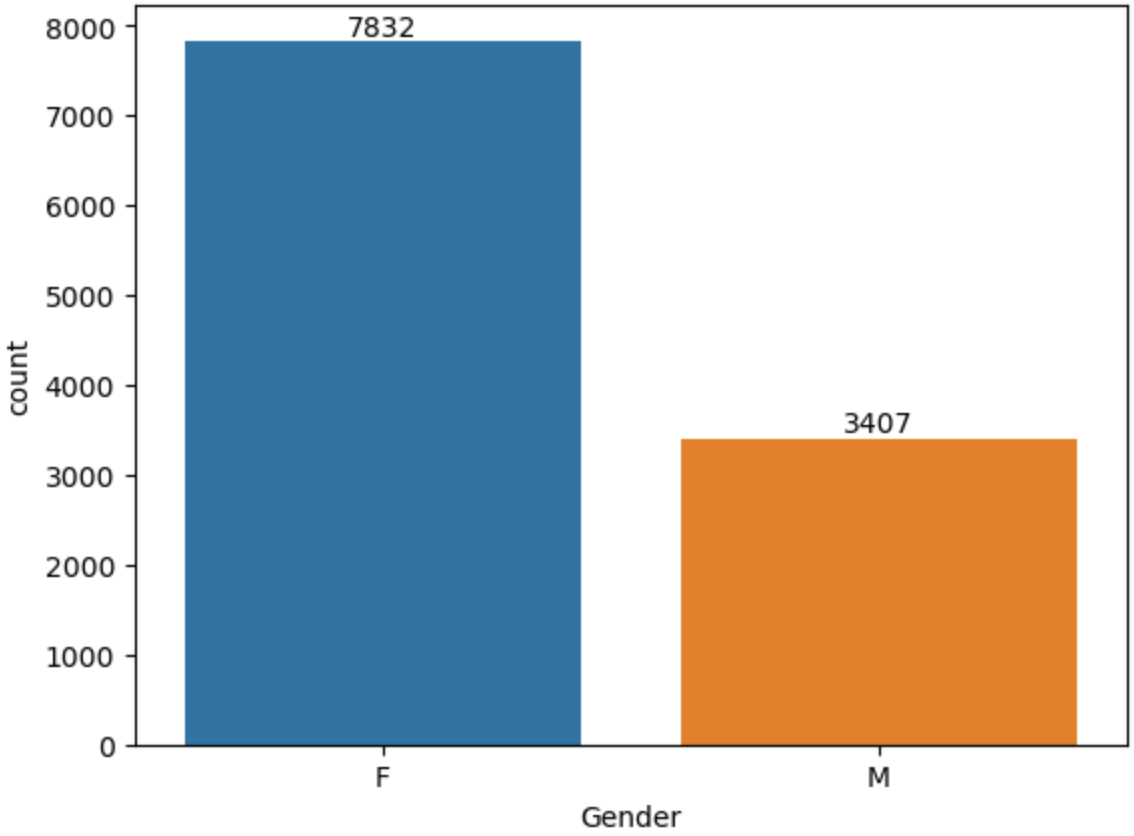
Gender

In [75]:
```
df.columns
```

Out[75]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

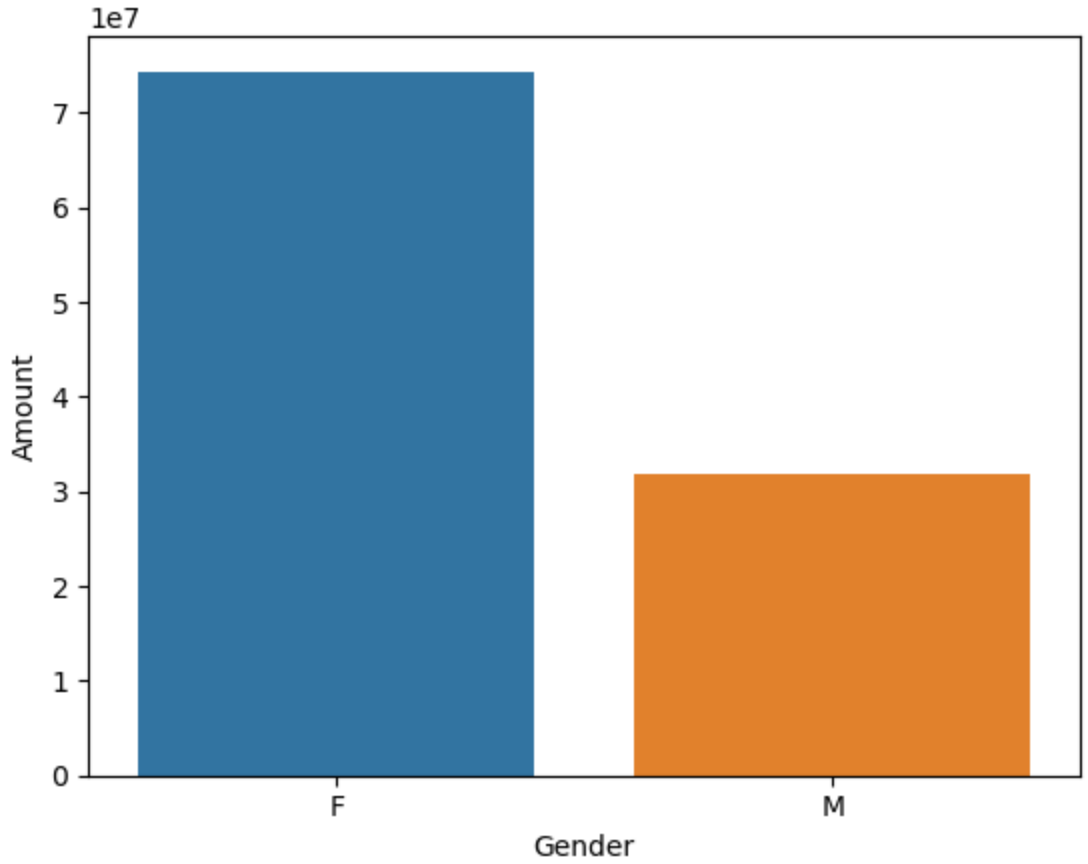
In [78]:
```
ax = sns.countplot(x = 'Gender', data = df)

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



In [80]:
```
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[80]:
```
<Axes: xlabel='Gender', ylabel='Amount'>
```



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

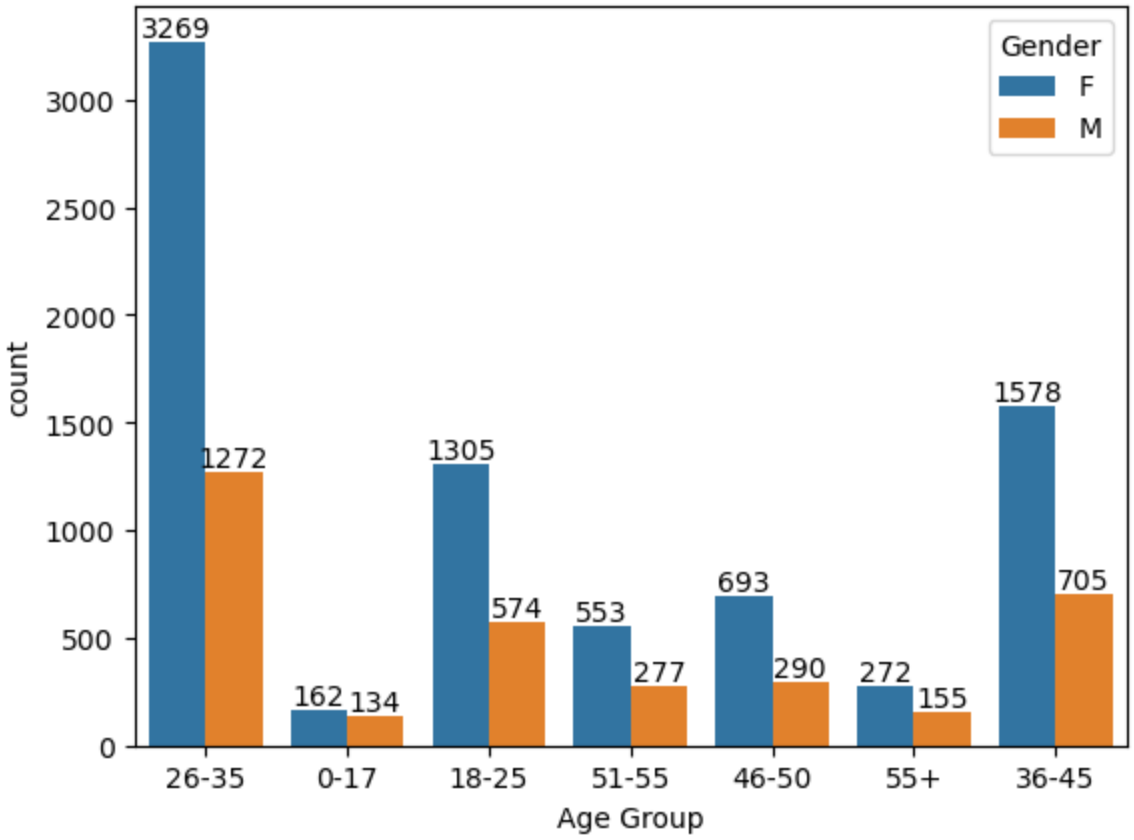
Age

In [83]:
```
df.columns
```

Out[83]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

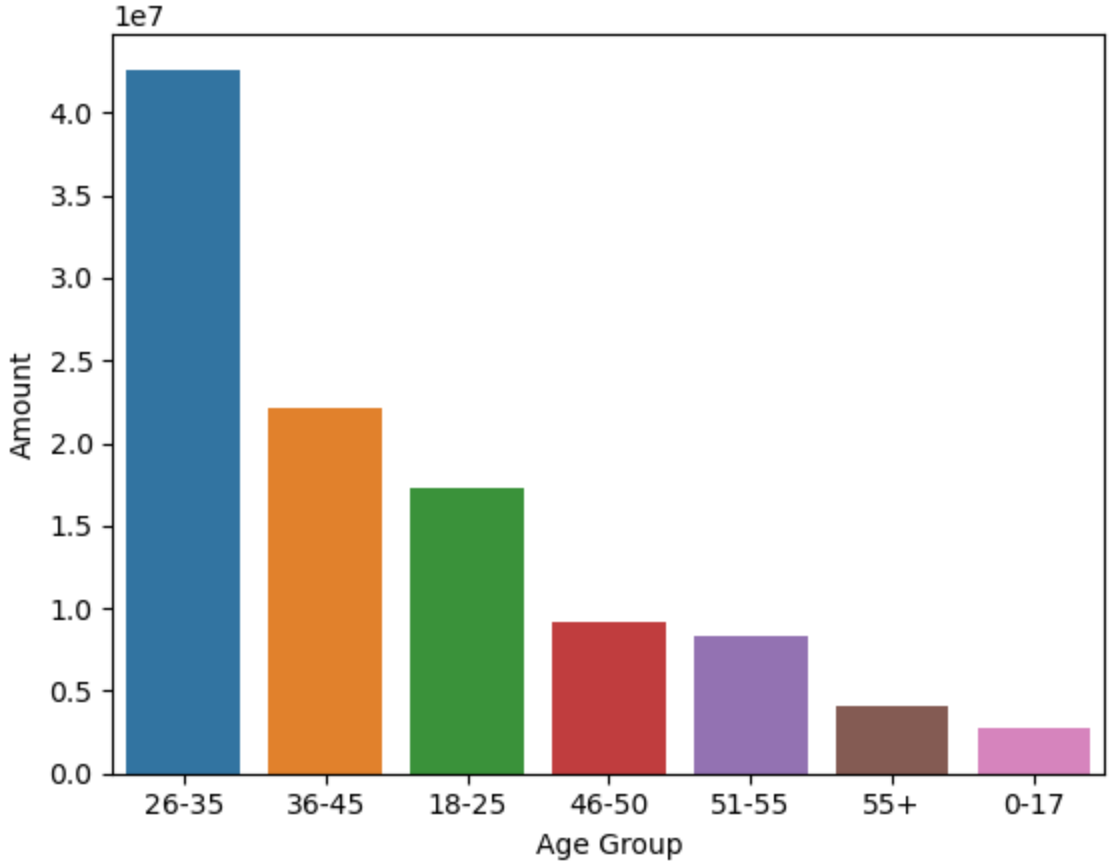
In [84]:
```
ax = sns.countplot(x = 'Age Group', data = df, hue = 'Gender')

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



In [91]:
```
# Total Amount vs Age Group
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[91]:
```
<Axes: xlabel='Age Group', ylabel='Amount'>
```



From above graph we can see that most of the buyers are of age group between 26-35 years Female

State

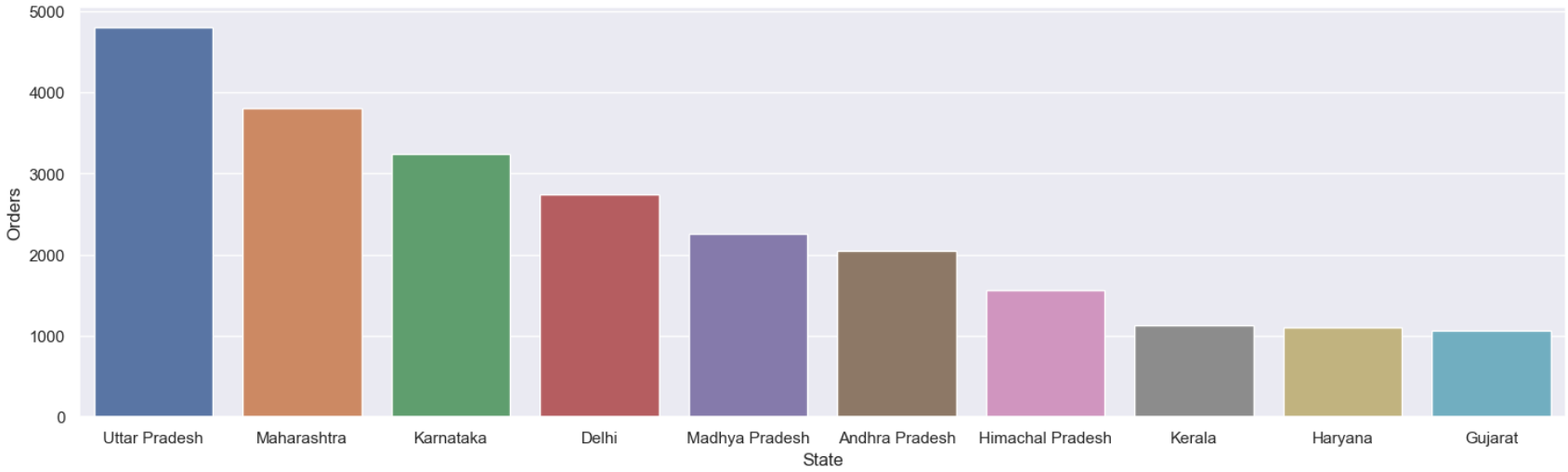
In [99]:
```
df.columns
```

Out[99]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

In [105…
```
# total number of orders from top 10 states

sales_state = df.groupby(['State'], as_index = False)['Orders'].sum().sort_values(by='Orders',ascending=False).head(1
sns.set(rc={'figure.figsize':(18,5)})
sns.barplot(x = 'State', y = 'Orders', data = sales_state)
```

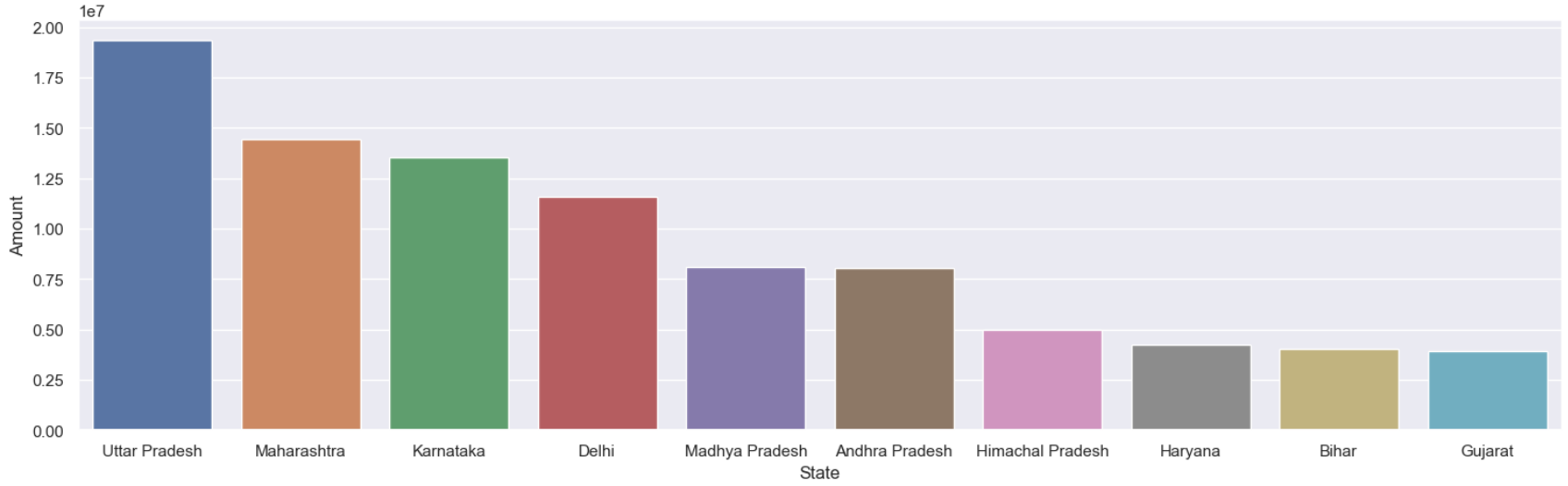
Out[105]: <Axes: xlabel='State', ylabel='Orders'>



In [106…
```
# total amount/sales from top 10 states

sales_state = df.groupby(['State'], as_index = False)['Amount'].sum().sort_values(by='Amount',ascending=False).head(1
sns.set(rc={'figure.figsize':(18,5)})
sns.barplot(x = 'State', y = 'Amount', data = sales_state)
```

Out[106]: <Axes: xlabel='State', ylabel='Amount'>



From above graphs we can see the most of the orders & total sales/amount from UP, Maharashtra and Karnatake respectively

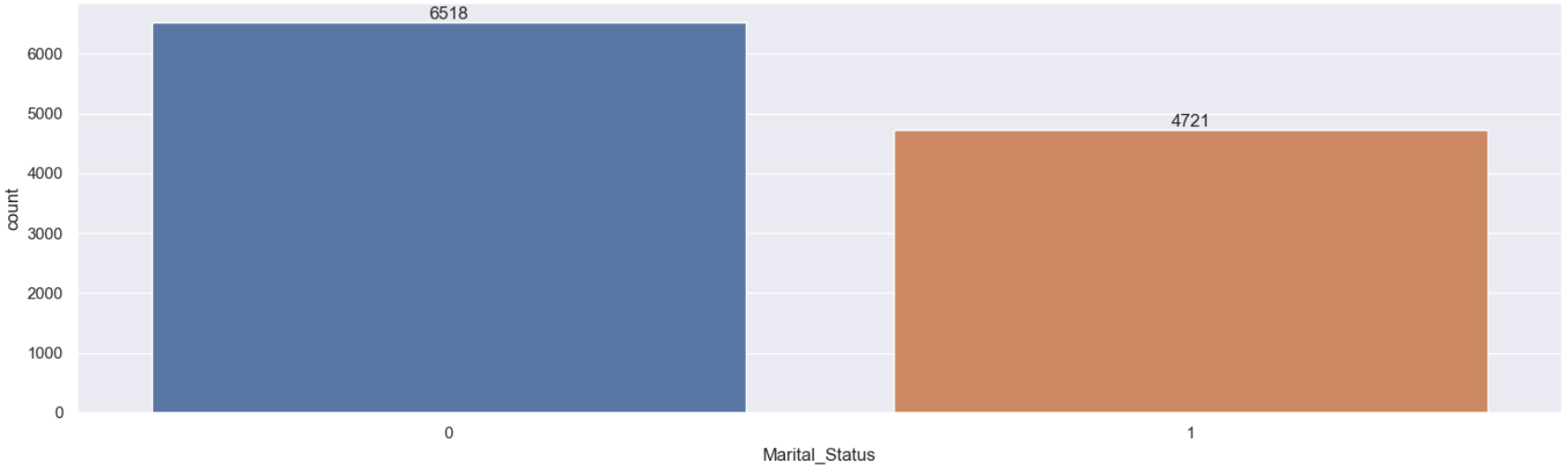
Marital Status

In [111…
```
df.columns
```

Out[111]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

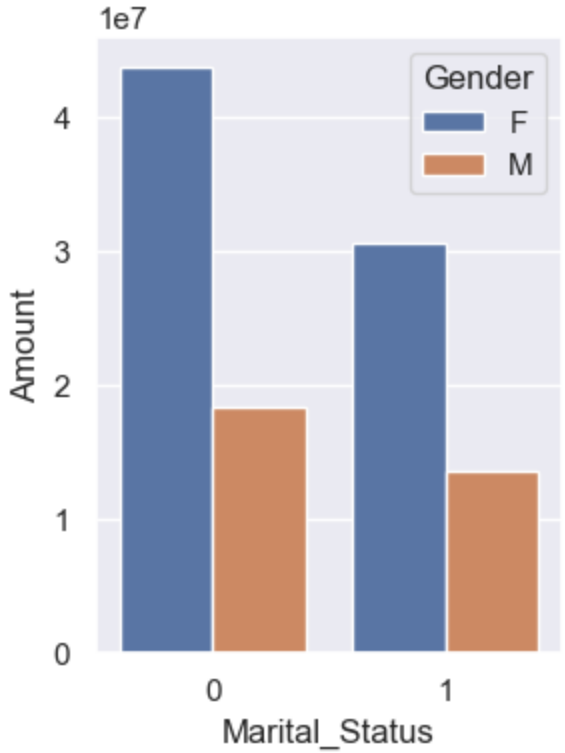
In [114…
```
ax = sns.countplot(data = df, x = 'Marital_Status')

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



In [116…
```
sales_state = df.groupby(['Marital_Status', 'Gender'], as_index = False)['Amount'].sum().sort_values(by='Amount',asce
sns.set(rc={'figure.figsize':(3,4)})
sns.barplot(x = 'Marital_Status', y = 'Amount', data = sales_state, hue = 'Gender')
```

Out[116]:
```
<Axes: xlabel='Marital_Status', ylabel='Amount'>
```



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

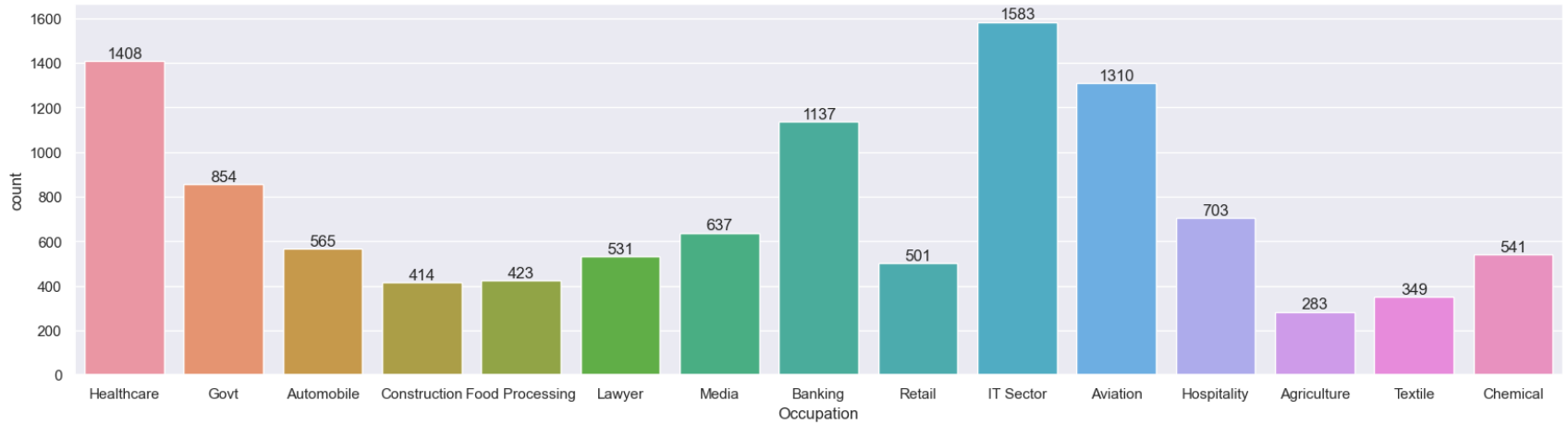
Occupation

In [117…
```
df.columns
```

Out[117]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

In [119…
```
sns.set(rc={'figure.figsize':(20,5)})
ax = sns.countplot(data = df, x = 'Occupation')
```

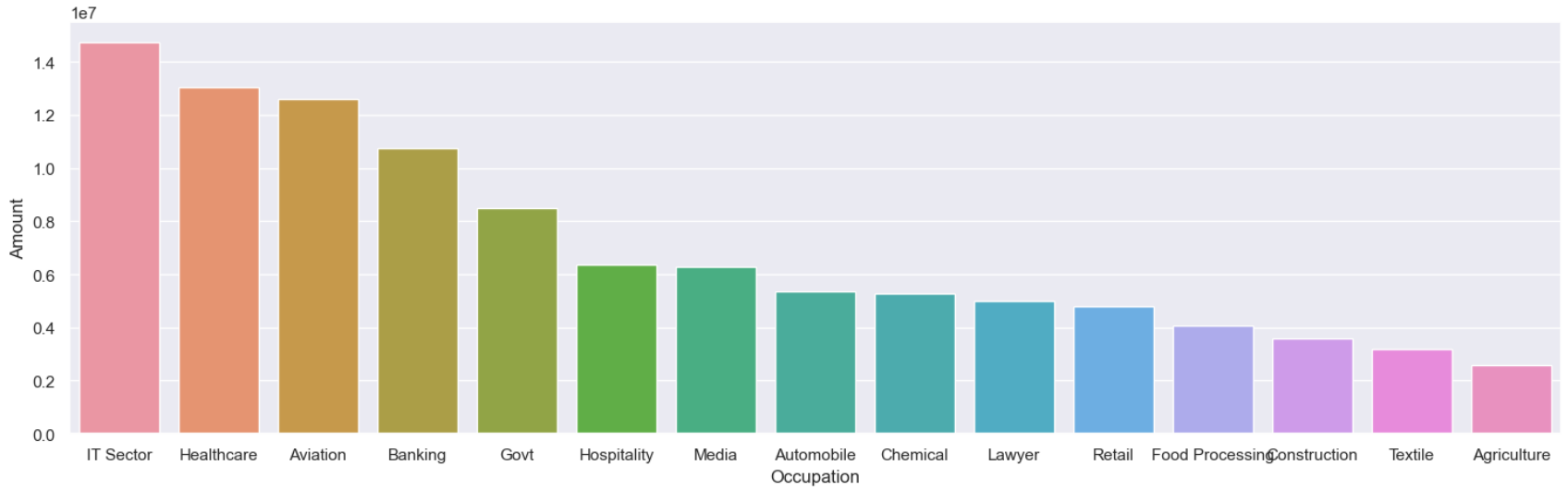
```
for bars in ax.containers:
    ax.bar_label(bars)
```



In [124…

```
sales_state = df.groupby(['Occupation'], as_index = False)['Amount'].sum().sort_values(by='Amount',ascending=False)
sns.set(rc={'figure.figsize':(18,5)})
sns.barplot(x = 'Occupation', y = 'Amount', data = sales_state)
```

Out[124]: <Axes: xlabel='Occupation', ylabel='Amount'>



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

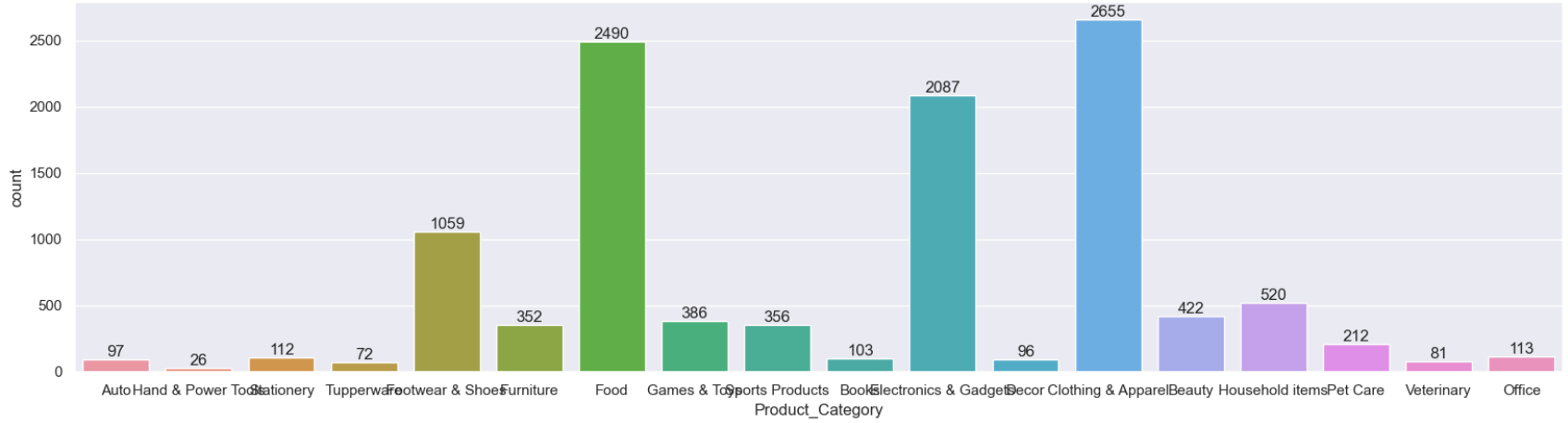
Product Category

In [121…
```
df.columns
```

Out[121]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

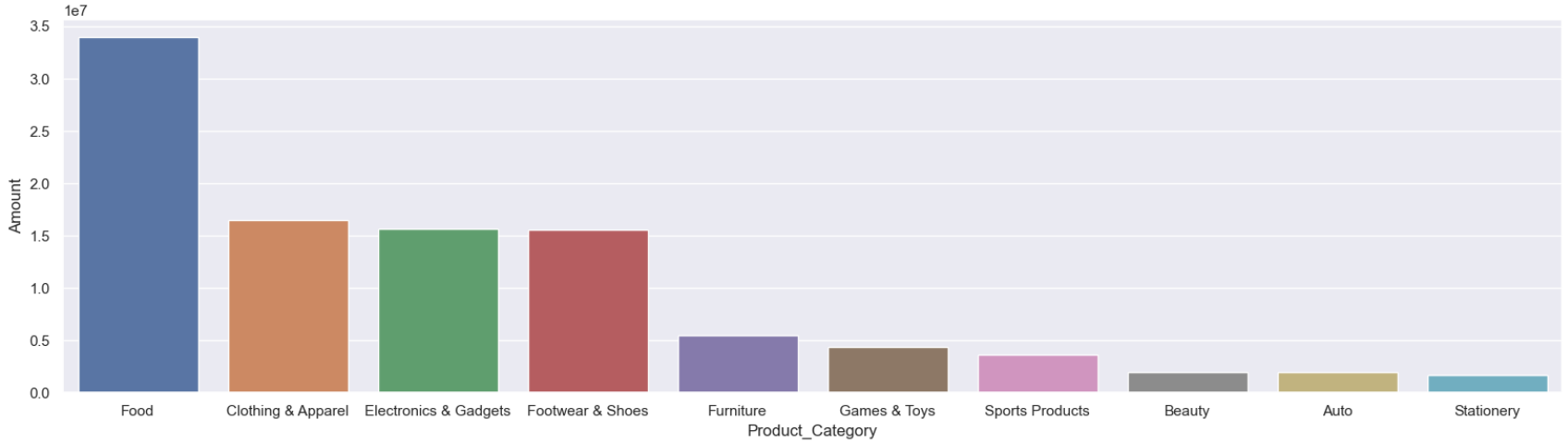
In [125…
```
sns.set(rc={'figure.figsize':(20,5)})
ax = sns.countplot(data = df, x = 'Product_Category')

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



In [133…
```
sales_state = df.groupby(['Product_Category'], as_index = False)['Amount'].sum().sort_values(by='Amount',ascending=Fa
sns.set(rc={'figure.figsize':(20,5)})
sns.barplot(x = 'Product_Category', y = 'Amount', data = sales_state)
```

Out[133]:
```
<Axes: xlabel='Product_Category', ylabel='Amount'>
```

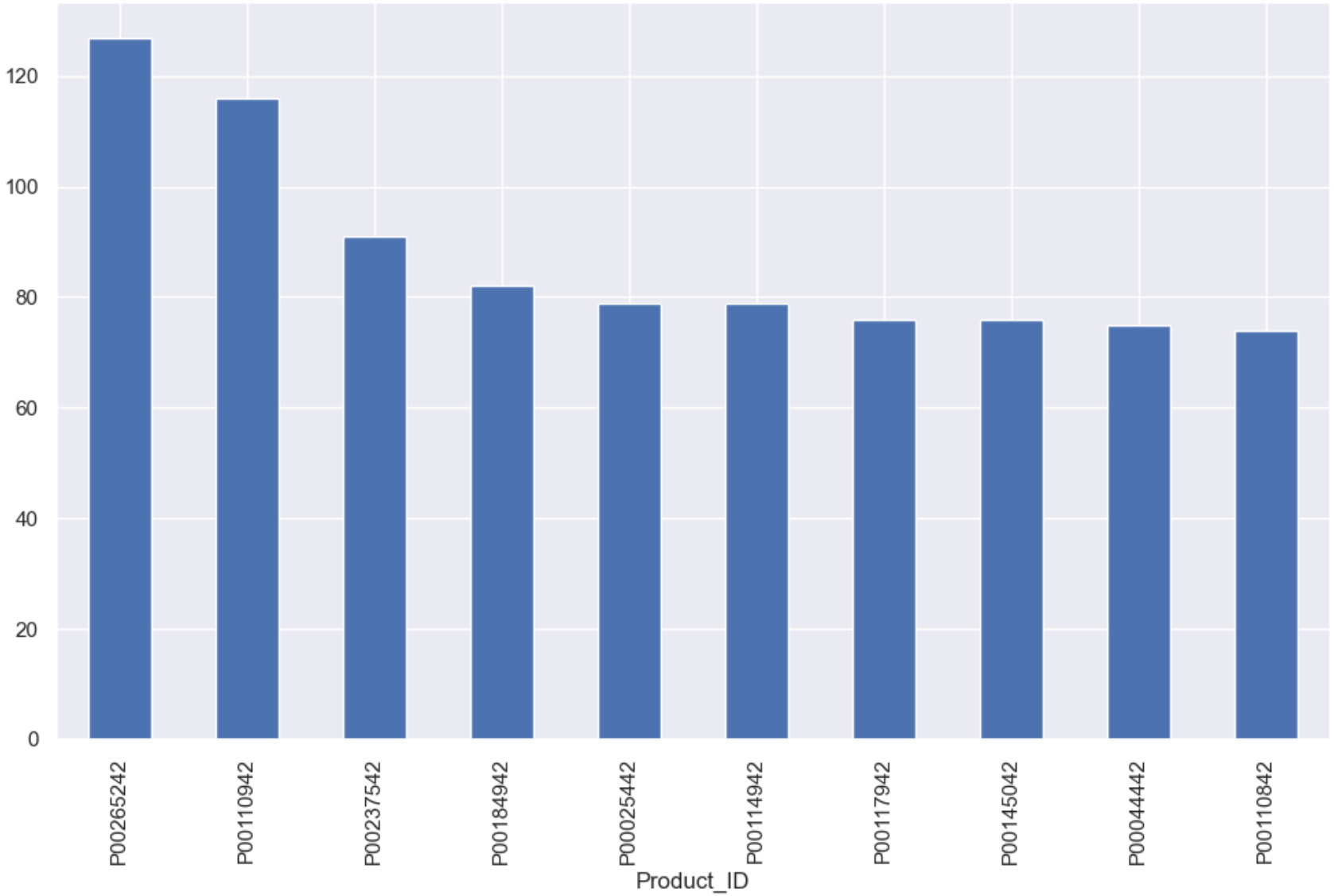


From above graph we can see that most of the sold products are from food, clothing & elctronics

In [135…
```
# top 10 most sold products (same thing as above)

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

Out[135]:
```
<Axes: xlabel='Product_ID'>
```


120
100
80
60
40
20
0
P00265242
P00110942
P00237542
P00184942
P00025442
P00114942
P00117942
P00145042
P00044442
P00110842
Product_ID

Conclusion

In [ ]:
```
* Married women age group 26-35 years from UP, Maharastra and Karnataka working in IT, Healthcare and Aviation are mo
buy products from food, clothing and electronics category.
```