In [11]:

```python
import pandas as pd
import numpy as np
import matplotlib as plt
import seaborn as sns
```

In [12]:

```python
df = pd.read_csv(r"C:\Users\Sairam\Desktop\python learning\Diwali Sales Data.csv",encoding='ISO-8859-1')
df
```

Out[12]:

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

11251 rows × 15 columns

In [5]:

```python
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_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
5	1000588	Joni	P00057942	M	26-35	28	1	Himachal Pradesh	Northern	Food Processing	A
6	1001132	Balk	P00018042	F	18-25	25	1	Uttar Pradesh	Central	Lawyer	A
7	1002092	Shivangi	P00273442	F	55+	61	0	Maharashtra	Western	IT Sector	A
8	1003224	Kushal	P00205642	M	26-35	35	0	Uttar Pradesh	Central	Govt	A
9	1003650	Ginny	P00031142	F	26-35	26	1	Andhra Pradesh	Southern	Media	A

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 [11]: df.columns
```

```
Out[11]: Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
               'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
               'Orders', 'Amount', 'Status', 'unnamed1'],
              dtype='object')
```

```
In [16]: df[['Age','Orders','Amount']].describe()
```

Out[16]:

	Age	Orders	Amount
count	11251.000000	11251.000000	11239.000000
mean	35.421207	2.489290	9453.610858
std	12.754122	1.115047	5222.355869
min	12.000000	1.000000	188.000000
25%	27.000000	1.500000	5443.000000
50%	33.000000	2.000000	8109.000000
75%	43.000000	3.000000	12675.000000
max	92.000000	4.000000	23952.000000

In [17]:

```
df.drop(['unnamed1', 'Status'], axis=1, inplace=True, errors='ignore')
```

In [18]:

```
df
```

Out[18]:

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

11251 rows × 13 columns

In [20]:

```
df.isnull().sum()
```

```
Out[20]: 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 [21]: df.dropna(inplace=True)
```

```
In [23]: df.isnull().sum()
```

```
Out[23]: 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 [24]: df.shape
```

```
Out[24]: (11239, 13)
```

```
In [25]: df['Amount']=df['Amount'].astype('int')
```

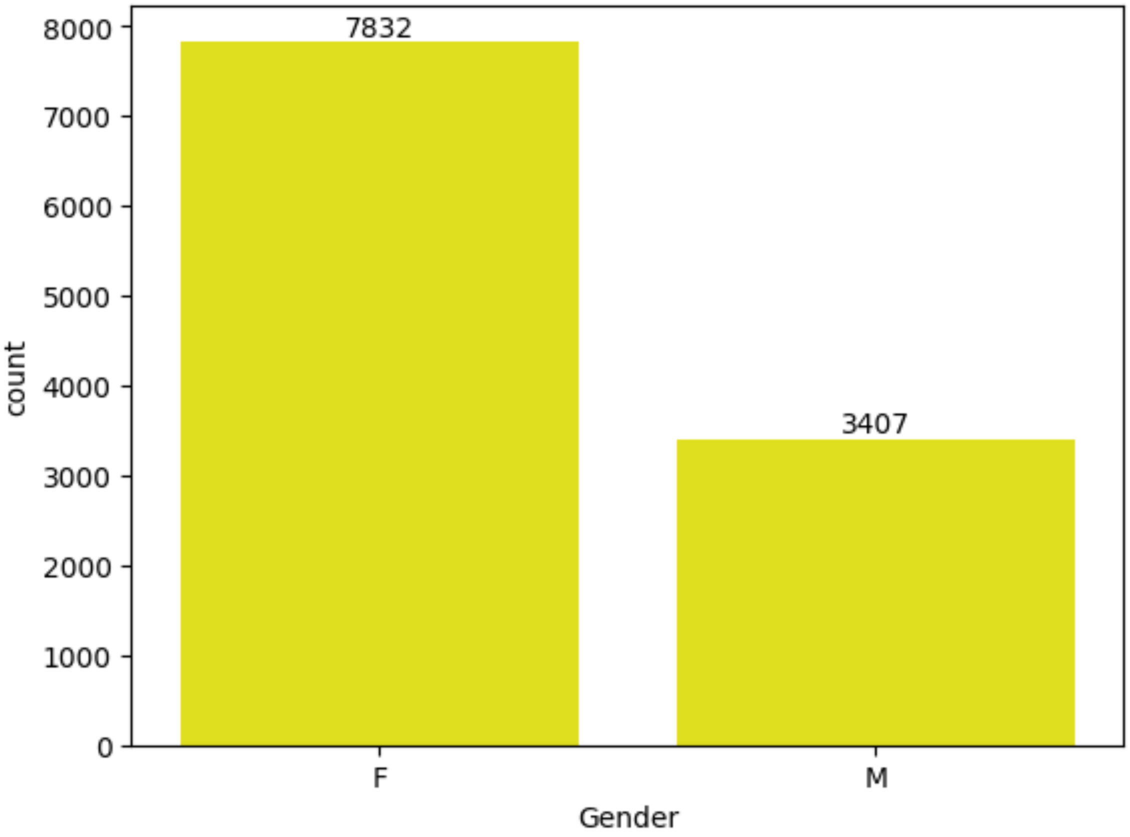
```
In [26]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 11239 entries, 0 to 11250
Data columns (total 13 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   User_ID           11239 non-null  int64 
 1   Cust_name         11239 non-null  object
 2   Product_ID        11239 non-null  object
 3   Gender            11239 non-null  object
 4   Age Group         11239 non-null  object
 5   Age               11239 non-null  int64 
 6   Marital_Status    11239 non-null  int64 
 7   State             11239 non-null  object
 8   Zone              11239 non-null  object
 9   Occupation        11239 non-null  object
 10  Product_Category  11239 non-null  object
 11  Orders            11239 non-null  int64 
 12  Amount            11239 non-null  int64 
dtypes: int64(5), object(8)
memory usage: 1.2+ MB
```

EDA

In [33]:

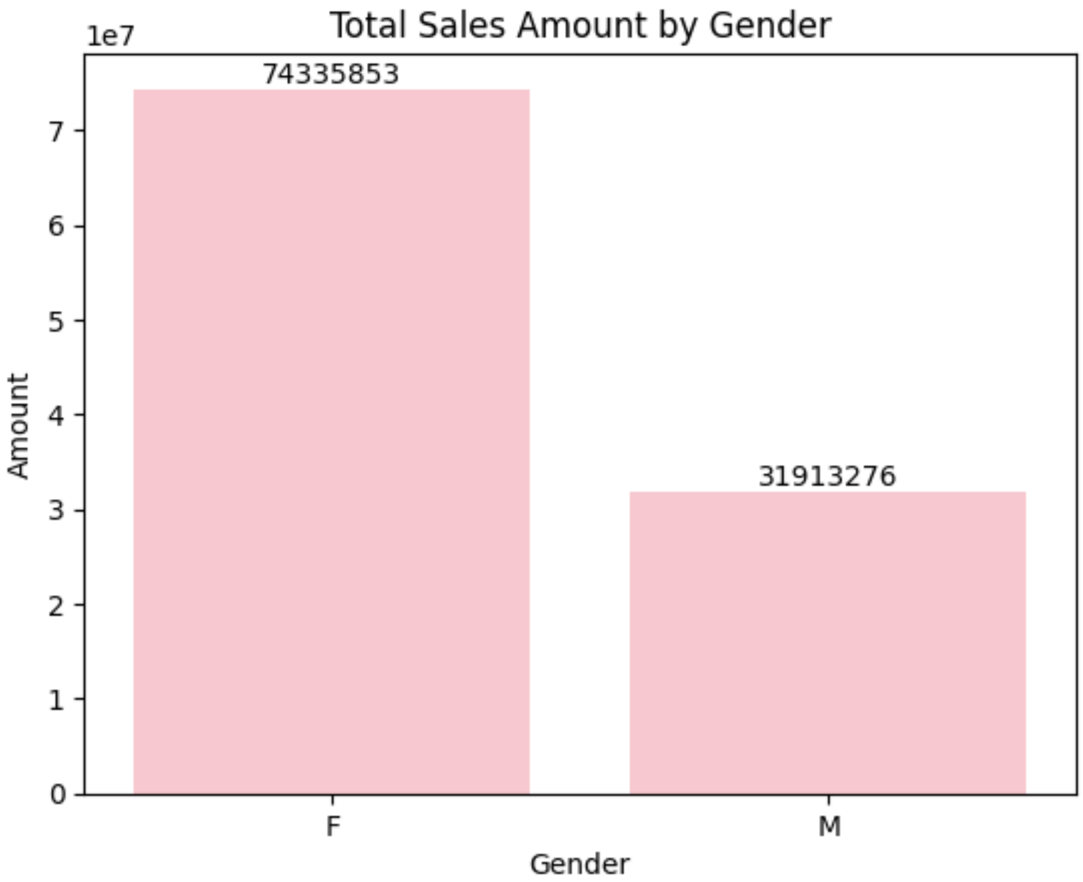
```
new=sns.countplot(x='Gender',data=df,color='yellow')
for bars in new.containers:
    new.bar_label(bars)
```



```
In [38]: gender_sales = df.groupby(['Gender'], as_index=False)['Amount'].sum().sort_values(by='Amount', ascending=False)
         sns.barplot(x='Gender', y='Amount', data=gender_sales, color='pink')
         for index, row in gender_sales.iterrows():
             plt.text(x=index,
                      y=row['Amount'] + 10,
                      s=f"{row['Amount']:.0f}",
                      ha='center',
                      va='bottom')

         plt.title("Total Sales Amount by Gender")
```

```
Out[38]: Text(0.5, 1.0, 'Total Sales Amount by Gender')
```

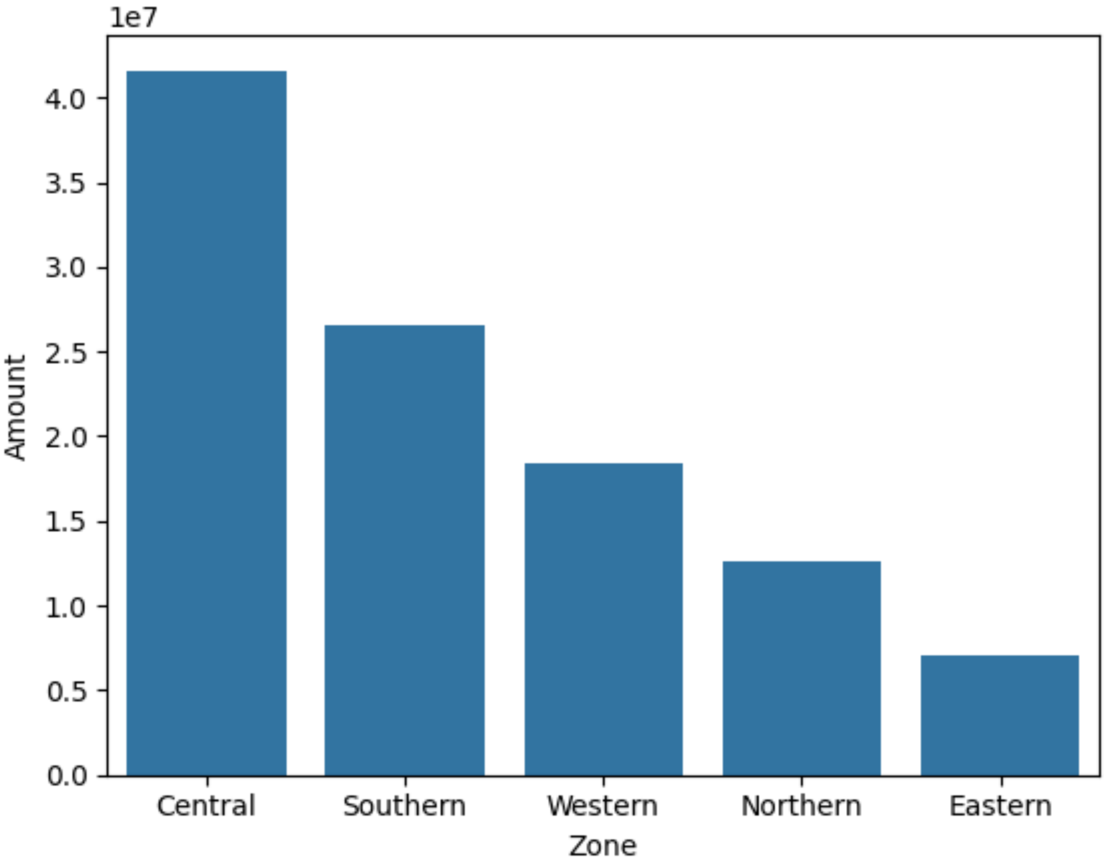


```
In [42]: df['Zone'].unique()
```

```
Out[42]: array(['Western', 'Southern', 'Central', 'Northern', 'Eastern'],
               dtype=object)
```

```
In [43]: zone_sales=df.groupby(['Zone'],as_index=False)['Amount'].sum().sort_values(by='Amount',ascending=False)
         sns.barplot(x='Zone',y='Amount',data=zone_sales)
```

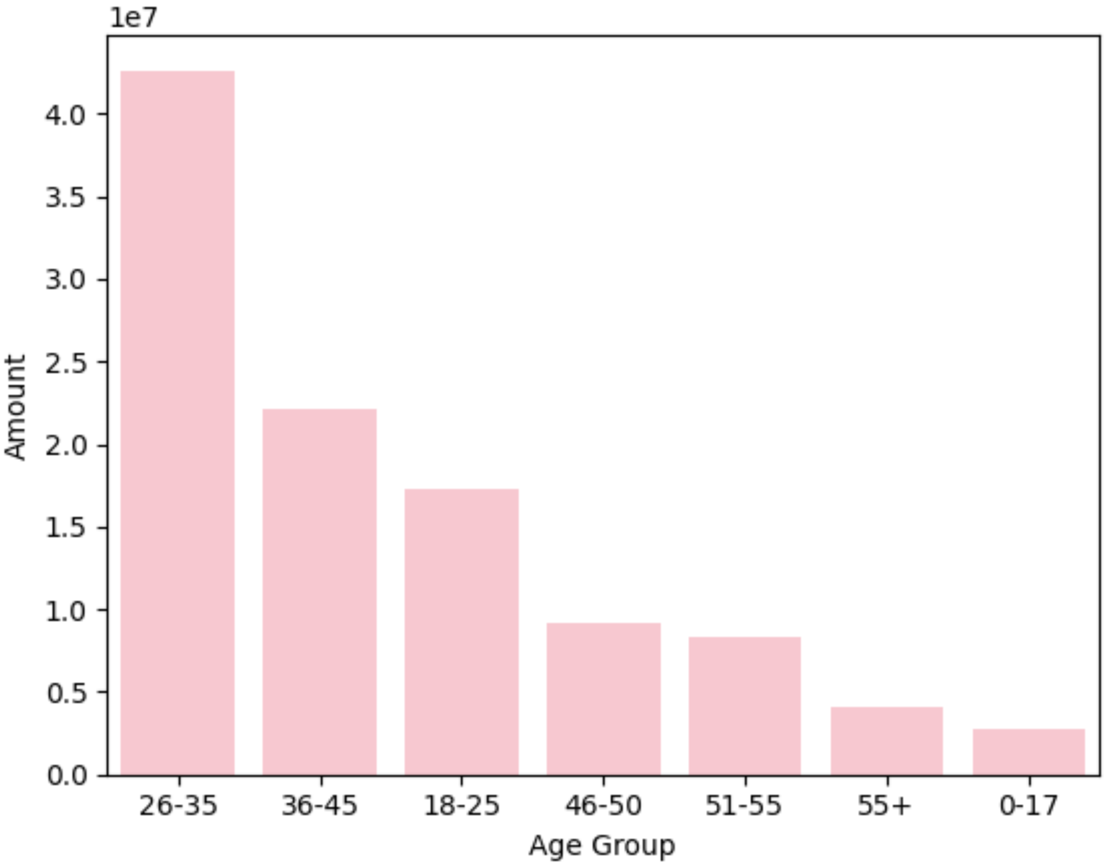
```
Out[43]: <Axes: xlabel='Zone', ylabel='Amount'>
```



In [52]:
```
age_sales=df.groupby(['Age Group'],as_index=False)['Amount'].sum().sort_values(by='Amount',ascending=False)

sns.barplot(x='Age Group',y='Amount',data=age_sales,color='pink')
```

Out[52]:
```
<Axes: xlabel='Age Group', ylabel='Amount'>
```

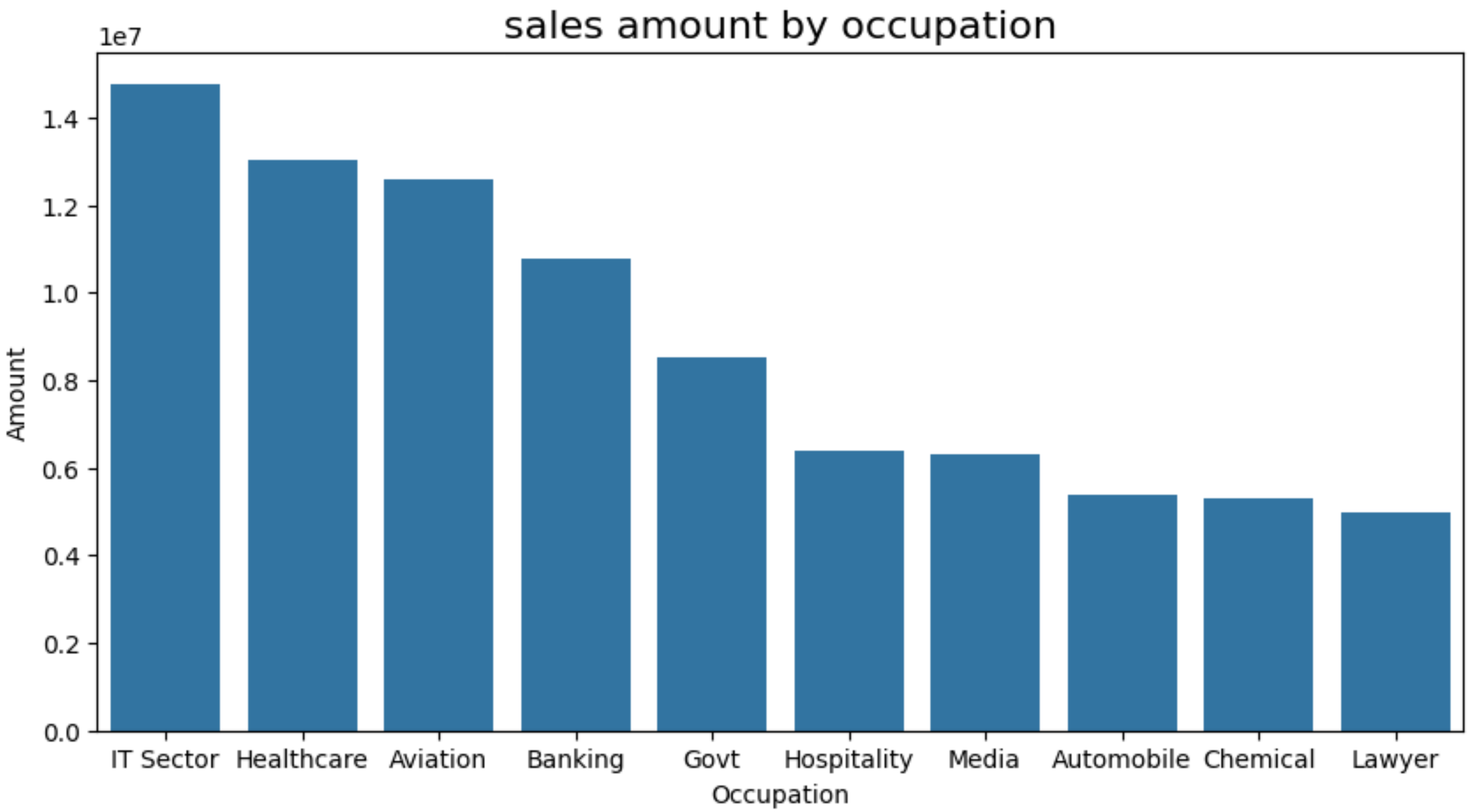


```
In [53]: df['Occupation'].unique()
```

```
Out[53]: array(['Healthcare', 'Govt', 'Automobile', 'Construction',
                'Food Processing', 'Lawyer', 'Media', 'Banking', 'Retail',
                'IT Sector', 'Aviation', 'Hospitality', 'Agriculture', 'Textile',
                'Chemical'], dtype=object)
```

```
In [84]: occup_sales=df.groupby(['Occupation'],as_index=False)['Amount'].sum().sort_values(by='Amount',ascending=False)
         top_10_occu=occup_sales.head(10)
         plt.figure(figsize=(10, 5))
         sns.barplot(x='Occupation',y='Amount',data=top_10_occu)
         plt.title('sales amount by occupation',fontsize=16)
```

```
Out[84]: Text(0.5, 1.0, 'sales amount by occupation')
```

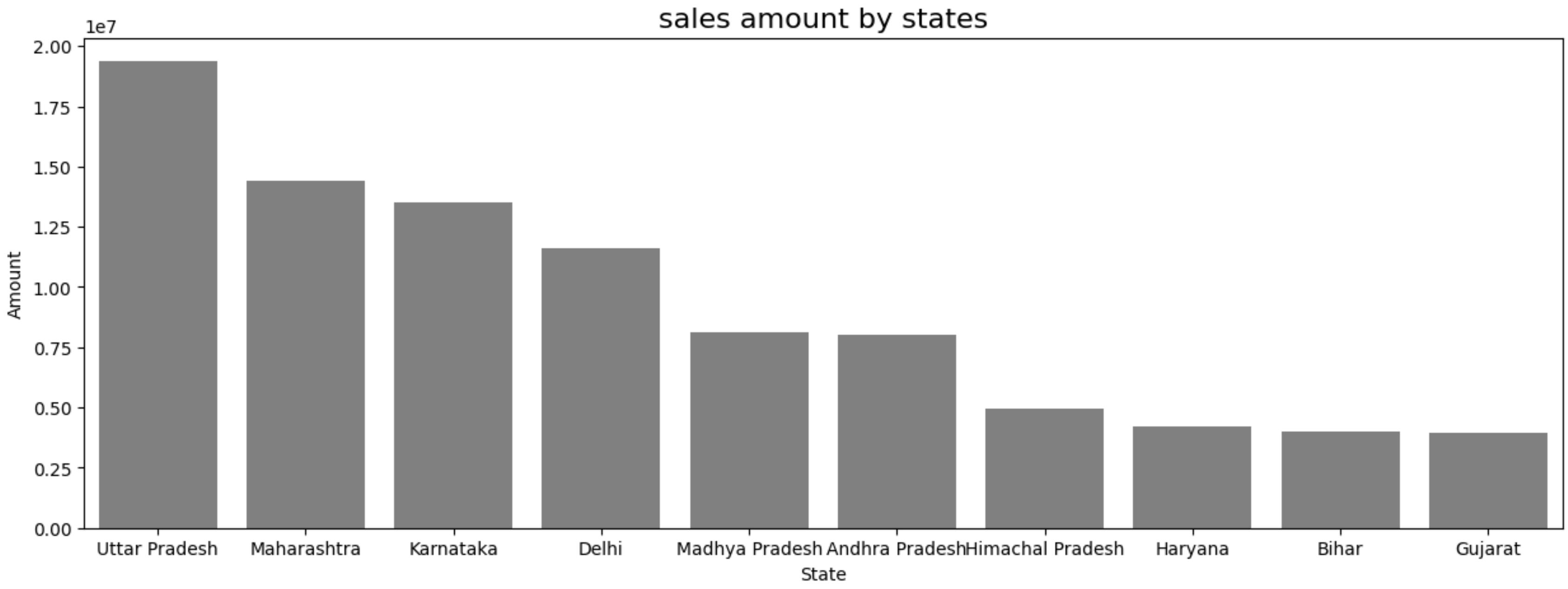


In [83]:

```
stat_sales=df.groupby(['State'],as_index=False)['Amount'].sum().sort_values(by='Amount',ascending=False)
top_10_stat=stat_sales.head(10)
plt.figure(figsize=(15,5))
sns.barplot(x='State',y='Amount',data=top_10_stat,color='grey')
plt.title('sales amount by states',fontsize=16)
```

Out[83]:

```
Text(0.5, 1.0, 'sales amount by states')
```

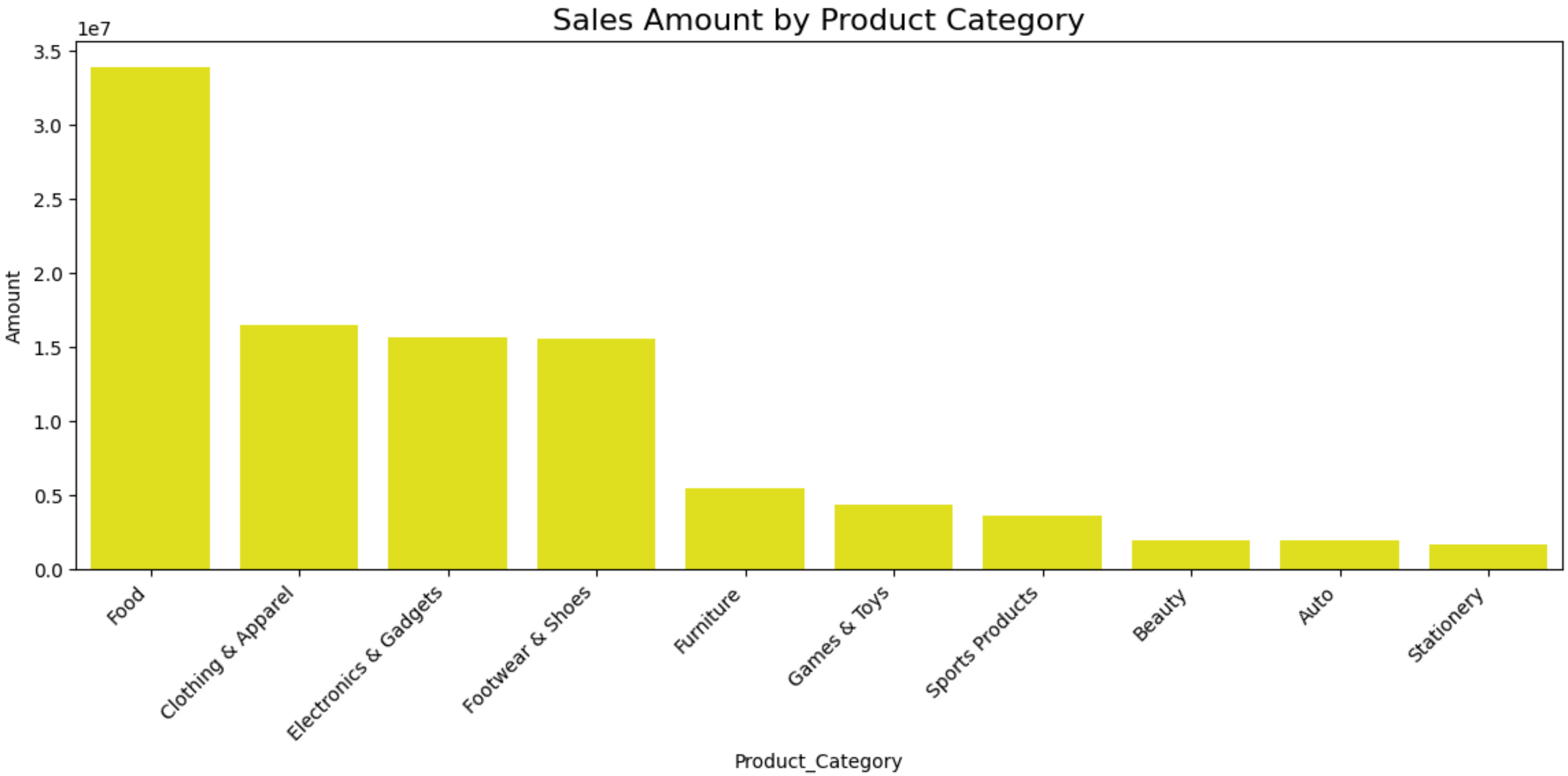


```
In [74]:  df['Product_Category'].unique()
```

```
Out[74]:  array(['Auto', 'Hand & Power Tools', 'Stationery', 'Tupperware',
                 'Footwear & Shoes', 'Furniture', 'Food', 'Games & Toys',
                 'Sports Products', 'Books', 'Electronics & Gadgets', 'Decor',
                 'Clothing & Apparel', 'Beauty', 'Household items', 'Pet Care',
                 'Veterinary', 'Office'], dtype=object)
```

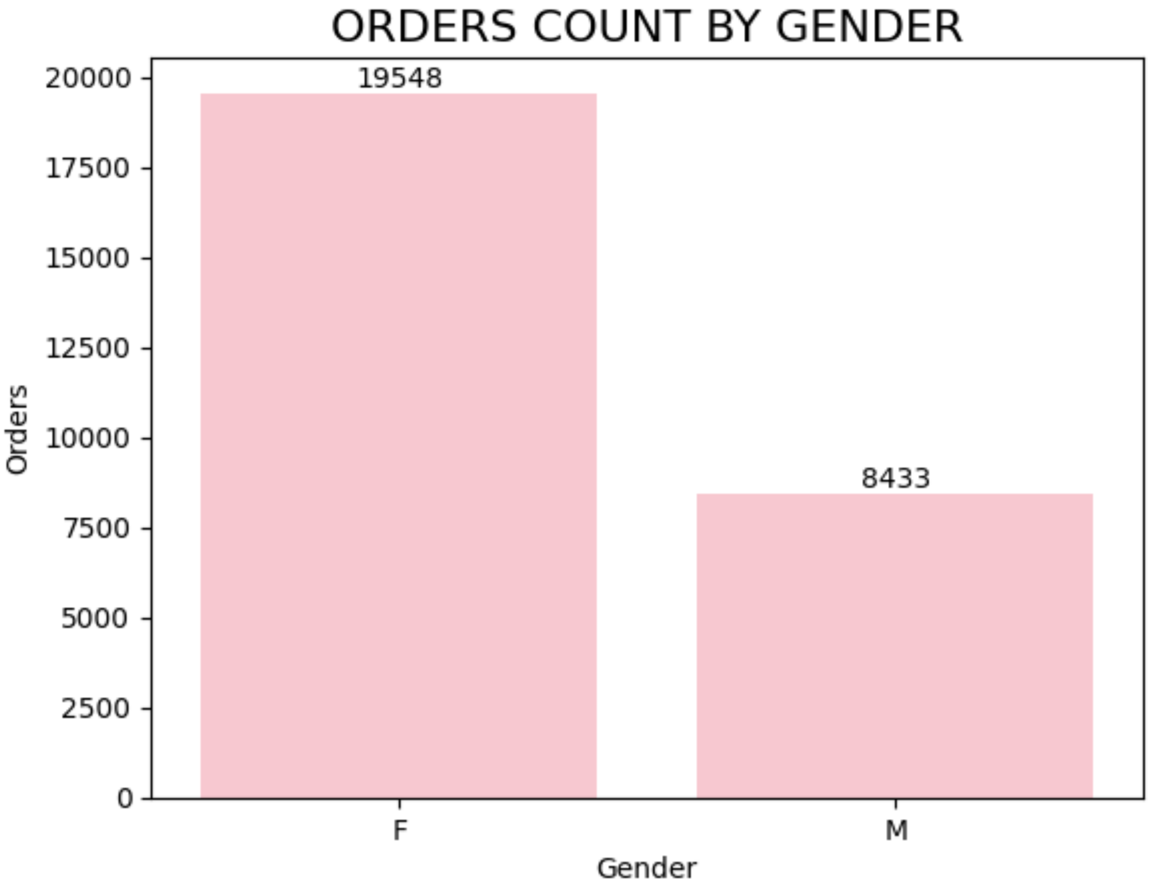
```
In [82]:  cat_sales=df.groupby(['Product_Category'],as_index=False)['Amount'].sum().sort_values(by='Amount',ascending=False)
          top_10_stat=cat_sales.head(10)
          plt.figure(figsize=(14,5))
          sns.barplot(x='Product_Category',y='Amount',data=top_10_stat,color='yellow')
          plt.xticks(rotation=45, ha='right')
          plt.title("Sales Amount by Product Category", fontsize=16)
```

```
Out[82]:  Text(0.5, 1.0, 'Sales Amount by Product Category')
```



```
In [89]: age_sales=df.groupby(['Gender'],as_index=False)['Orders'].sum().sort_values(by='Orders',ascending=False)

         sns.barplot(x='Gender',y='Orders',data=age_sales,color='pink')
         for index, row in age_sales.iterrows():
             plt.text(x=index,
                      y=row['Orders'] + 1,  # Offset a bit above the bar
                      s=row['Orders'],
                      ha='center',
                      va='bottom',
                      fontsize=10)
             plt.title('ORDERS COUNT BY GENDER',fontsize=16)
```



Women are the top buyers and the highest contributors to revenue. Most of the purchases come from the states of Uttar Pradesh and Maharashtra. Among product categories, Food, Clothing & Apparel, and Electronics & Gadgets recorded the highest sales.