Sales_project

July 4, 2025

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

```
[5]: df=pd.read_csv('/Users/akula/Desktop/Diwali Sales Data.csv', encoding=␣
       ↪'unicode_escape')
```

```
[6]: df.shape
```

```
[6]: (11251, 15)
```

the table data consists of 11251 rows and 15 columns

```
[7]: df.head(10)
```

```
[7]:    User_ID  Cust_name Product_ID Gender Age Group  Age  Marital_Status  \
     0  1002903  Sanskriti  P00125942      F     26-35   28               0
     1  1000732     Kartik  P00110942      F     26-35   35               1
     2  1001990      Bindu  P00118542      F     26-35   35               1
     3  1001425     Sudevi  P00237842      M      0-17   16               0
     4  1000588       Joni  P00057942      M     26-35   28               1
     5  1000588       Joni  P00057942      M     26-35   28               1
     6  1001132       Balk  P00018042      F     18-25   25               1
     7  1002092   Shivangi  P00273442      F       55+   61               0
     8  1003224     Kushal  P00205642      M     26-35   35               0
     9  1003650      Ginny  P00031142      F     26-35   26               1

                  State      Zone       Occupation Product_Category  Orders  \
     0       Maharashtra   Western       Healthcare             Auto       1
     1    Andhra Pradesh  Southern             Govt             Auto       3
     2     Uttar Pradesh   Central       Automobile             Auto       3
     3         Karnataka  Southern     Construction             Auto       2
     4           Gujarat   Western  Food Processing             Auto       2
     5  Himachal Pradesh  Northern  Food Processing             Auto       1
     6     Uttar Pradesh   Central           Lawyer             Auto       4
     7       Maharashtra   Western        IT Sector             Auto       1
     8     Uttar Pradesh   Central             Govt             Auto       2
```

```
9    Andhra Pradesh  Southern            Media              Auto       4

      Amount  Status  unnamed1
0   23952.00     NaN       NaN
1   23934.00     NaN       NaN
2   23924.00     NaN       NaN
3   23912.00     NaN       NaN
4   23877.00     NaN       NaN
5   23877.00     NaN       NaN
6   23841.00     NaN       NaN
7        NaN     NaN       NaN
8   23809.00     NaN       NaN
9   23799.99     NaN       NaN
```

[8]:

```
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 above data two columns having null values and one colunm is with wrong data type that is amount should be in int but it is in float we have to change it and remove two null rows

[12]:

```
df.shape
```

[12]:

```
(11251, 13)
```

after removing the null columns from the table we have 11251 rows and 13 columns

[16]:

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

```
[16]: 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
```

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

```
[19]: df.shape
```

```
[19]: (11239, 13)
```

after removing all null values we have 11239 rows and 13 columns

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

```
[24]: df['Amount'].dtypes
```

```
[24]: dtype('int64')
```

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

```
[27]:        User_ID    Cust_name Product_ID Gender Age Group  Age  Shaadi  \
      0      1002903    Sanskriti  P00125942      F     26-35   28       0
      1      1000732       Kartik  P00110942      F     26-35   35       1
      2      1001990        Bindu  P00118542      F     26-35   35       1
      3      1001425       Sudevi  P00237842      M      0-17   16       0
      4      1000588         Joni  P00057942      M     26-35   28       1
      …          …            …          …      …         …    …       …
      11246  1000695      Manning  P00296942      M     18-25   19       1
      11247  1004089 Reichenbach  P00171342      M     26-35   33       0
      11248  1001209        Oshin  P00201342      F     36-45   40       0
      11249  1004023       Noonan  P00059442      M     36-45   37       0
      11250  1002744      Brumley  P00281742      F     18-25   19       0

                      State      Zone       Occupation Product_Category  Orders  \
      0         Maharashtra   Western       Healthcare             Auto       1
      1      Andhra Pradesh  Southern             Govt             Auto       3
      2       Uttar Pradesh   Central       Automobile             Auto       3
```

```
3           Karnataka  Southern     Construction             Auto       2
4             Gujarat   Western  Food Processing             Auto       2
…                …       …              …              …     …
11246     Maharashtra   Western         Chemical           Office       4
11247         Haryana  Northern       Healthcare       Veterinary       3
11248  Madhya Pradesh   Central          Textile           Office       4
11249       Karnataka  Southern      Agriculture           Office       3
11250     Maharashtra   Western       Healthcare           Office       3

       Amount
0       23952
1       23934
2       23924
3       23912
4       23877
…       …
11246     370
11247     367
11248     213
11249     206
11250     188

[11239 rows x 13 columns]
```

[28]:

```
df.columns
```

[28]:

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

[41]:

```
Continous=['Age']
discrete_count=['Orders', 'Amount']
discrete_categorical=['Cust_name','Gender','Age␣
 ↪Group','Marital_Status','State','Zone','Occupation','Product_Category','Product_ID']
```

[42]:

```
df[Continous].describe()
```

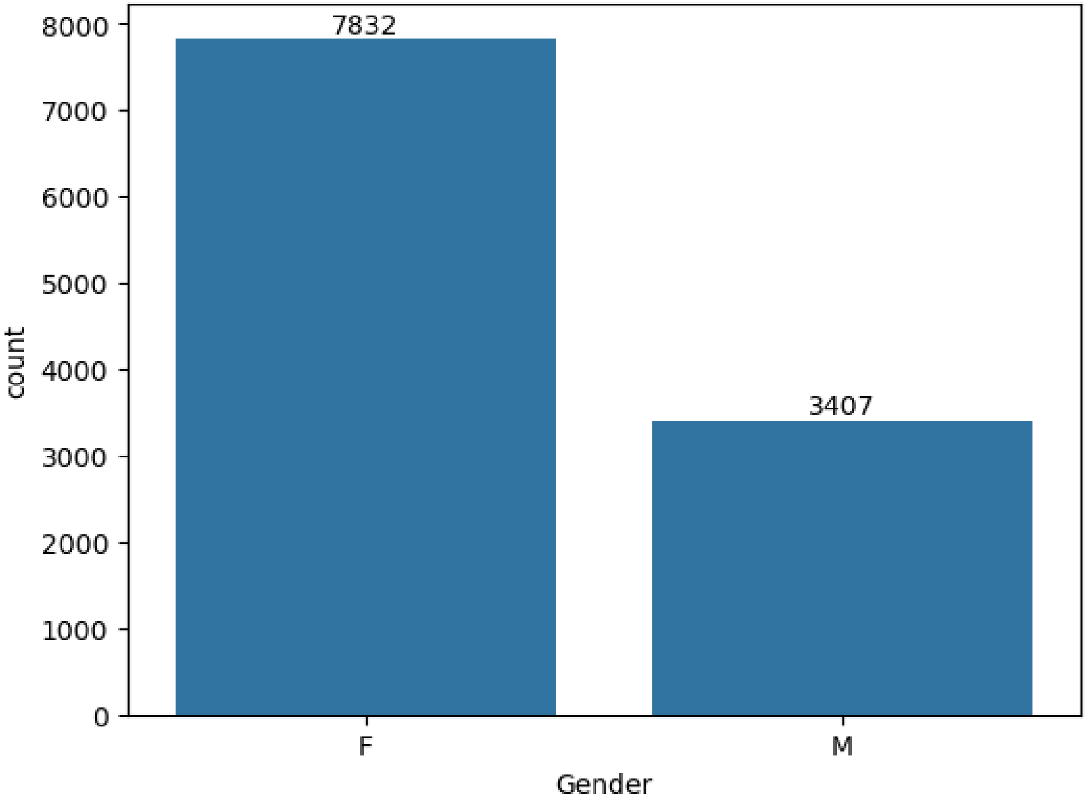
[42]:

```
                Age
count  11239.000000
mean      35.410357
std       12.753866
min       12.000000
25%       27.000000
50%       33.000000
75%       43.000000
max       92.000000
```

```
[43]: df[discrete_count].describe()
```

```
[43]:            Orders        Amount
      count  11239.000000  11239.000000
      mean       2.489634   9453.610553
      std        1.114967   5222.355168
      min        1.000000    188.000000
      25%        2.000000   5443.000000
      50%        2.000000   8109.000000
      75%        3.000000  12675.000000
      max        4.000000  23952.000000
```

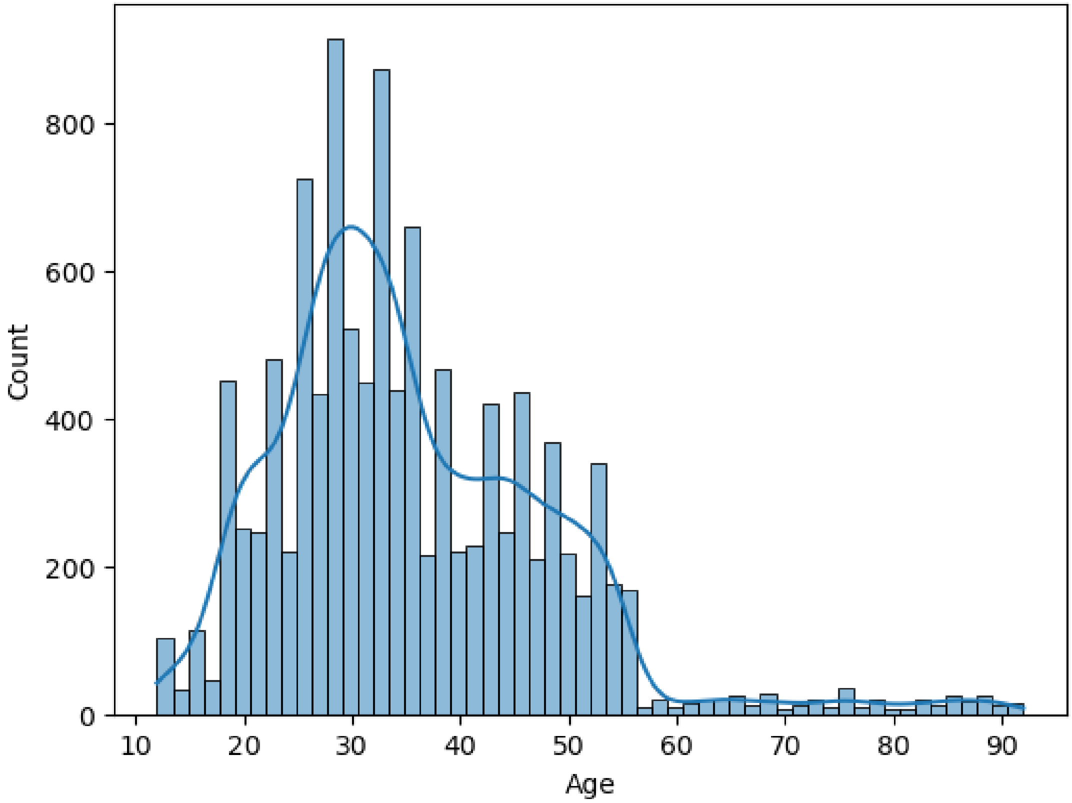
```
[61]: count = sns.countplot(x=df['Gender'])
      for bars in count.containers:
          count.bar_label(bars)
```



from above bar plot there are 7832 females and 3407 males

```
[62]: sns.histplot(df['Age'],kde=True)
```

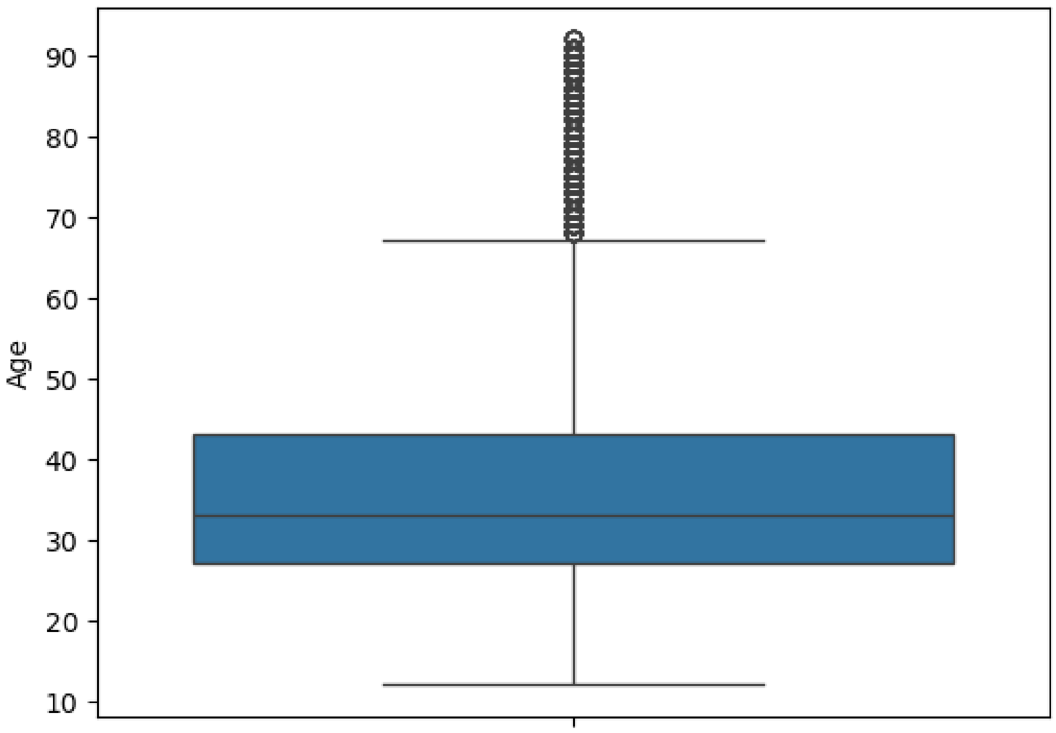
```
[62]: <Axes: xlabel='Age', ylabel='Count'>
```



from above observation it is a right skewed

```
[52]: sns.boxplot(df['Age'])
```

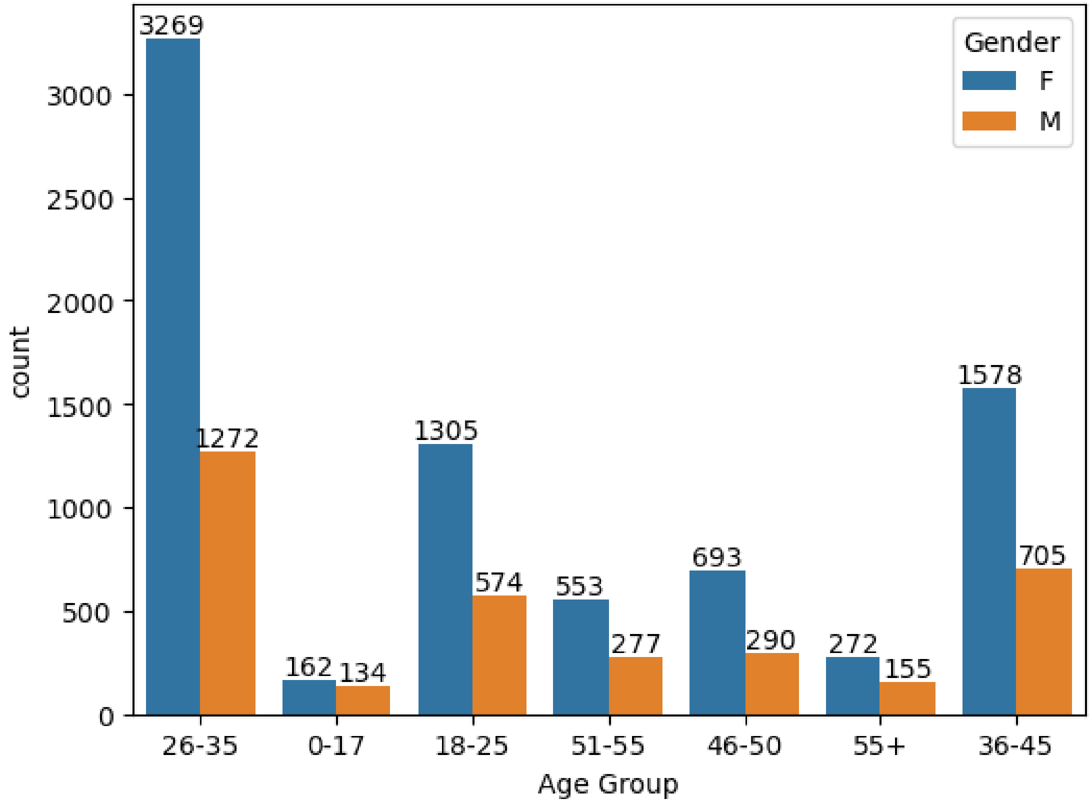
```
[52]: <Axes: ylabel='Age'>
```



```
[63]: df['Age'].unique()
```

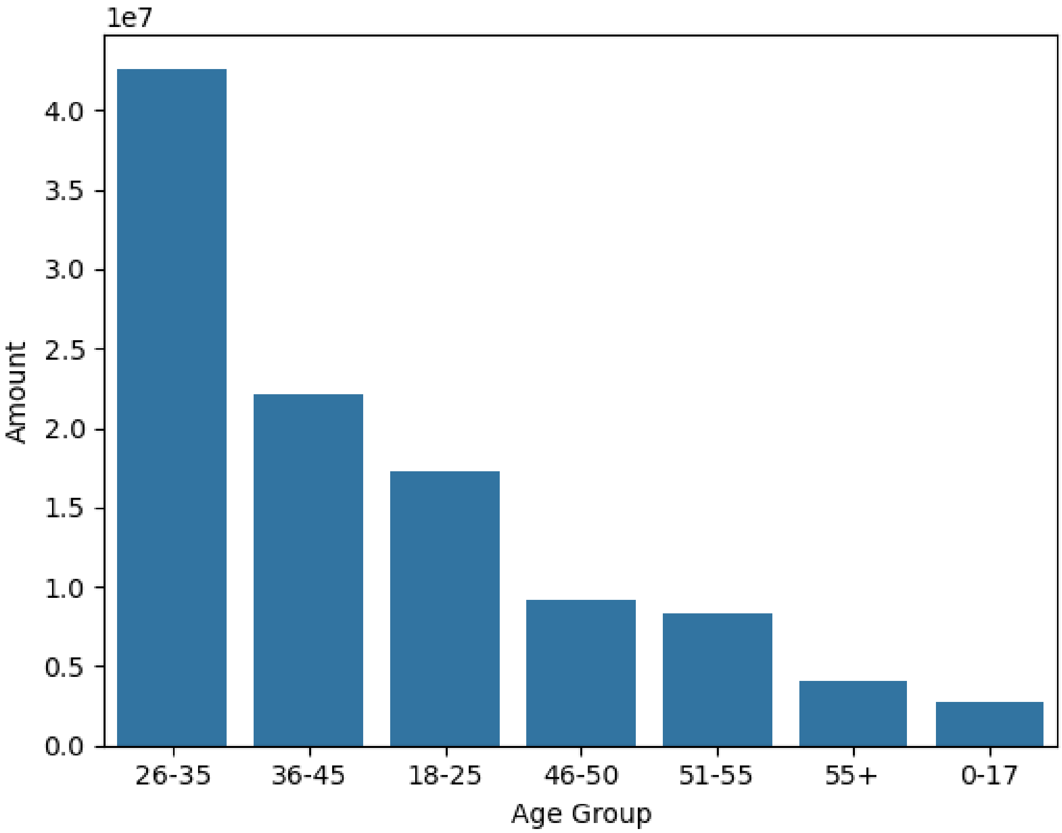
```
[63]: array([28, 35, 16, 25, 26, 34, 20, 24, 29, 54, 19, 46, 30, 53, 83, 33, 40,
             39, 32, 36, 55, 27, 72, 45, 43, 47, 22, 52, 18, 21, 38, 37, 23, 49,
             42, 50, 48, 31, 44, 41, 66, 15, 51, 77, 87, 79, 71, 88, 58, 82, 62,
             92, 12, 63, 17, 13, 67, 90, 56, 75, 81, 64, 73, 84, 14, 76, 86, 89,
             68, 61, 91, 85, 70, 80, 65, 74, 69, 78, 57, 60, 59])
```

```
[66]: counts = sns.countplot(data = df, x = 'Age Group', hue = 'Gender')
      for bars in counts.containers:
          counts.bar_label(bars)
```



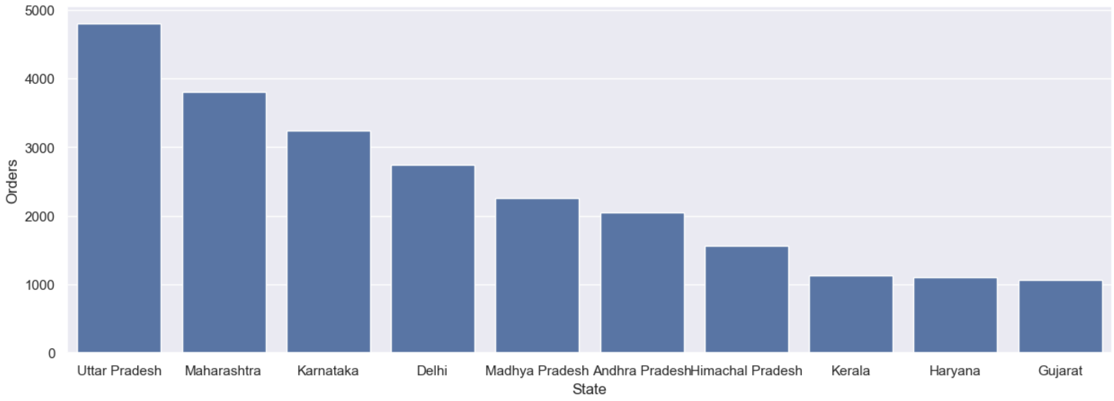
[67]:
```
sales_age = df.groupby(['Age Group'], as_index=False)['Amount'].sum().
 ↪sort_values(by='Amount', ascending=False)
sns.barplot(x = 'Age Group',y= 'Amount' ,data = sales_age)
```

[67]:
```
<Axes: xlabel='Age Group', ylabel='Amount'>
```



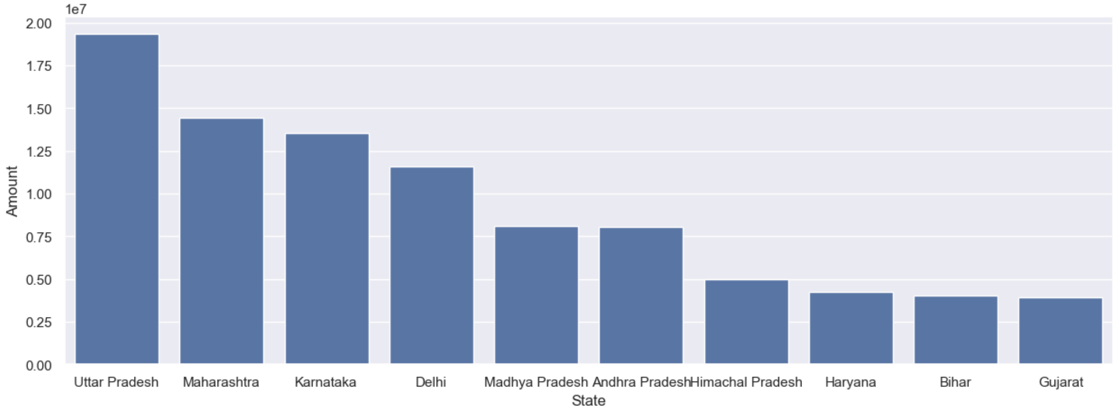
```
[69]: sales_state = df.groupby(['State'], as_index=False)['Orders'].sum().
        ↪sort_values(by='Orders', ascending=False).head(10)
      sns.set(rc={'figure.figsize':(15,5)})
      sns.barplot(data = sales_state, x = 'State',y= 'Orders')
```

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

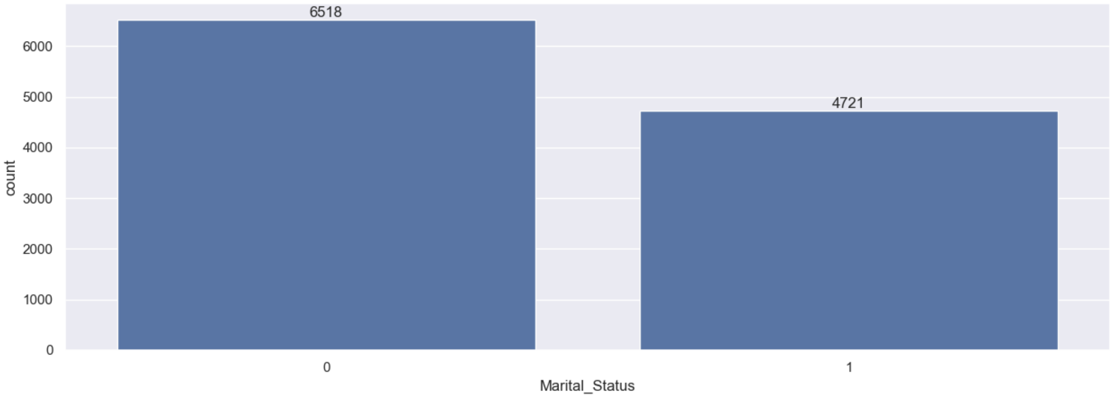


[70]:
```
sales_state = df.groupby(['State'], as_index=False)['Amount'].sum().
  ↪sort_values(by='Amount', ascending=False).head(10)
sns.set(rc={'figure.figsize':(15,5)})
sns.barplot(data = sales_state, x = 'State',y= 'Amount')
```

[70]:
```
<Axes: xlabel='State', ylabel='Amount'>
```



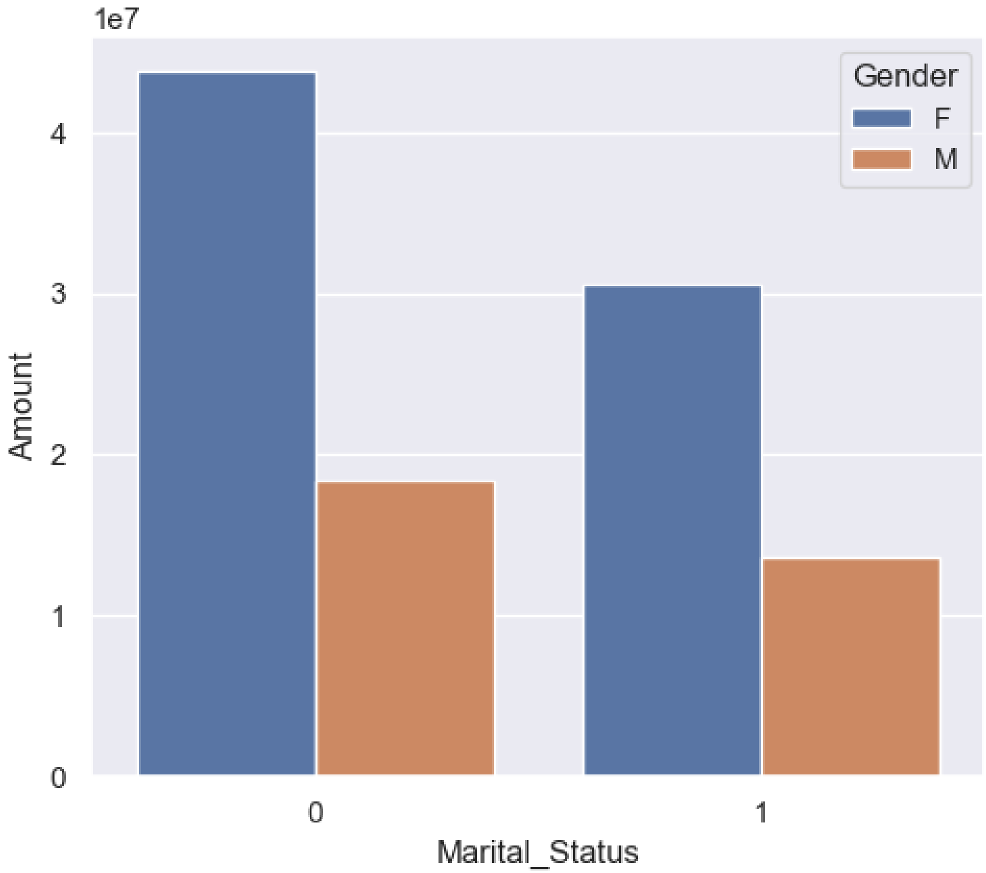
[71]:
```
ax = sns.countplot(data = df, x = 'Marital_Status')

sns.set(rc={'figure.figsize':(7,5)})
for bars in ax.containers:
    ax.bar_label(bars)
```



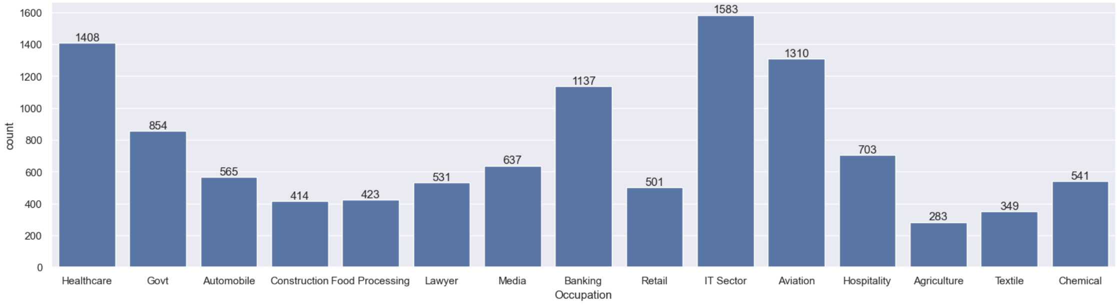
[72]:
```
sales_state = df.groupby(['Marital_Status', 'Gender'],␣
  ↪as_index=False)['Amount'].sum().sort_values(by='Amount', ascending=False)
```

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

[72]:
```
<Axes: xlabel='Marital_Status', ylabel='Amount'>
```



[73]:
```
sns.set(rc={'figure.figsize':(20,5)})
ax = sns.countplot(data = df, x = 'Occupation')

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

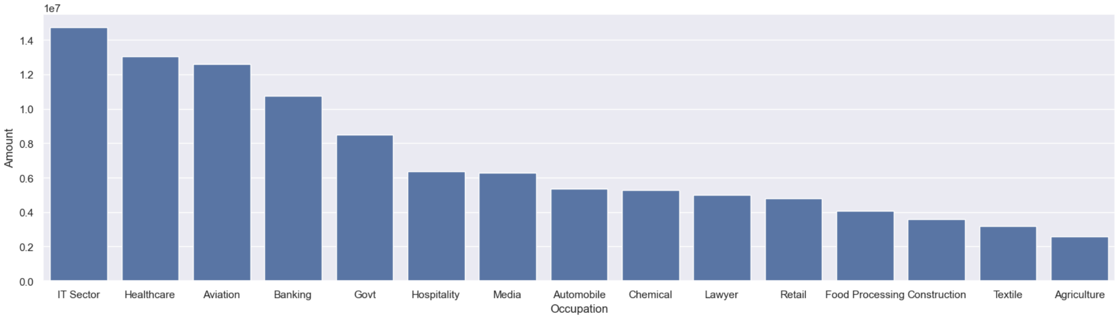


[74]:

```
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')
```

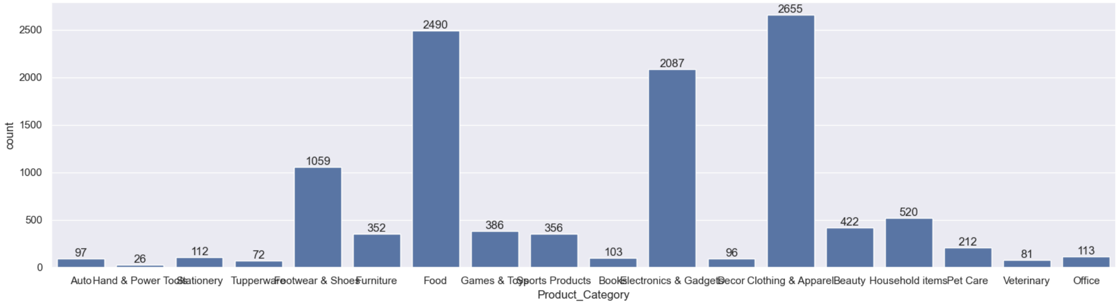
[74]:

```
<Axes: xlabel='Occupation', ylabel='Amount'>
```



[75]:

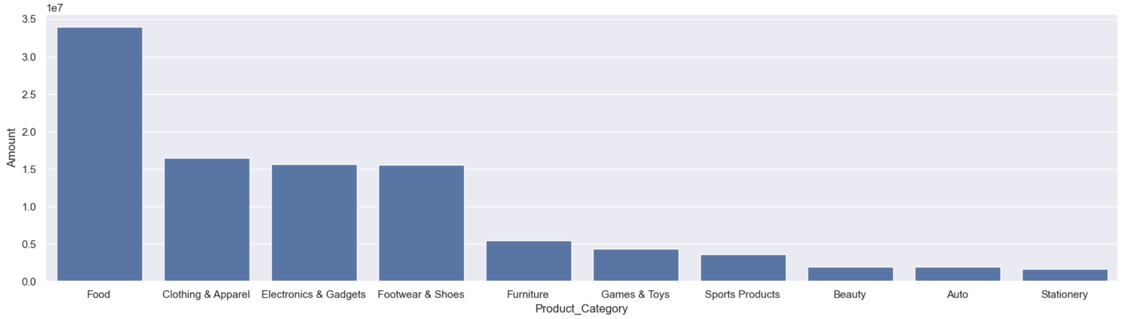
```
sns.set(rc={'figure.figsize':(20,5)})
ax = sns.countplot(data = df, x = 'Product_Category')

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



[76]:

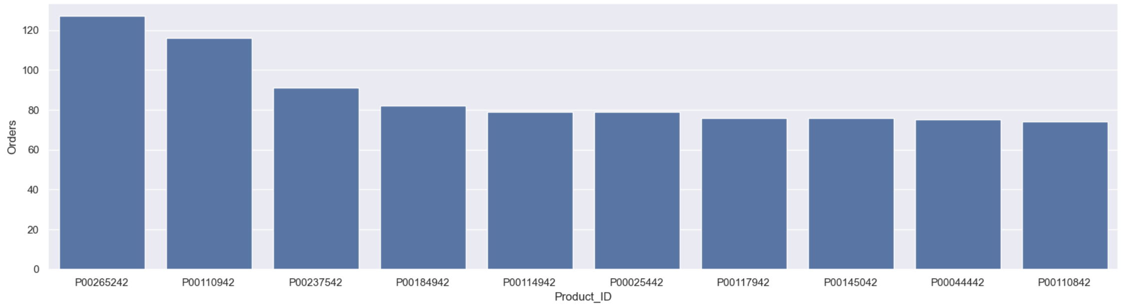
```
sales_state = df.groupby(['Product_Category'], as_index=False)['Amount'].sum().
 ↪sort_values(by='Amount', ascending=False).head(10)

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

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



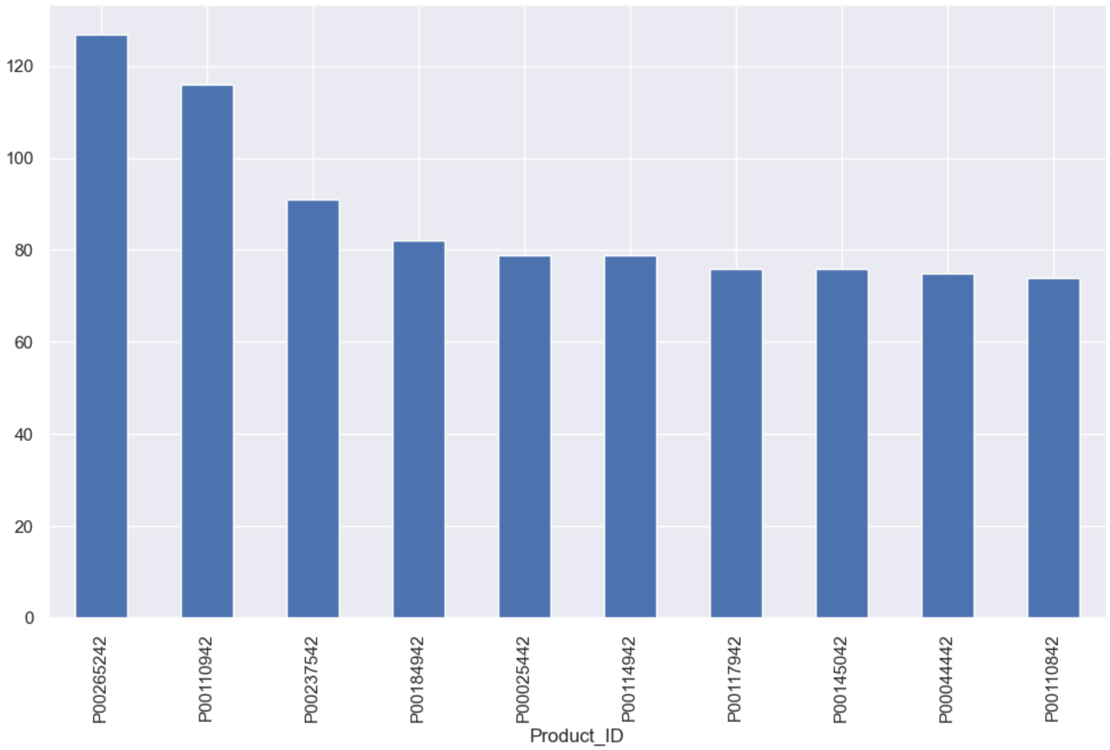
```
[77]: sales_state = df.groupby(['Product_ID'], as_index=False)['Orders'].sum().
        ↪sort_values(by='Orders', ascending=False).head(10)

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

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



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

```
[78]: <Axes: xlabel='Product_ID'>
```



Conclusion:

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

```
[ ]:
```