In [1]:

```
import pandas as pd
import numpy as np
```

In [7]:

```
diwali = pd.read_csv(r"E:\Python_Diwali_Sales_Analysis\Python_Diwali_Sales_Analysis\Diwal
diwali
```

26	1001101	Gibson	P00234742	F	36-45	40	0	Uttar Pradesh	Central
27	1004736	Mahima	P00058042	F	18-25	25	1	Andhra Pradesh	Southern
28	1004037	Etezadi	P00190542	M	51-55	54	1	Andhra Pradesh	Southern
29	1002340	James	P00119642	F	36-45	39	1	Andhra Pradesh	Southern
30	1005664	Dean	P00111642	F	18-25	20	0	Andhra Pradesh	Southern
31	1002523	Aman	P00293342	F	26-35	32	1	Andhra Pradesh	Southern
32	1002503	Mousam	P00220042	F	36-45	36	0	Andhra Pradesh	Southern
33	1002638	Damala	P00346242	F	26-35	35	1	Maharashtra	Western
34	1004505	Daniels	P00080042	F	51-55	55	1	Andhra Pradesh	Southern
35	1004957	Inderpreet	P00111842	M	26-35	27	1	Jharkhand	Eastern
36	1005649	Sweta	P00238542	M	18-25	20	1	Delhi	Central

In [3]:

```
diwali.shape
```

Out[3]:

```
(11251, 15)
```

In [6]:

```
diwali.head(n=1000)
```

105	1004335	Aryan	P00075542	F	36-45	38	0	Karnataka	Southern	H
106	1000280	Kajal	P00216042	F	51-55	55	0	Delhi	Central	I
107	1003311	Neola	P00142742	F	26-35	26	1	Karnataka	Southern	I
108	1004161	Murray	P00345642	F	46-50	46	0	Karnataka	Southern	
109	1005265	Sakshi	P00296242	F	46-50	48	1	Delhi	Central	
110	1004285	Bhishm	P00315842	M	36-45	38	0	Uttar Pradesh	Central	P
111	1005261	Apoorva	P00057942	F	36-45	41	1	Delhi	Central	
112	1000445	Sukruta	P00114042	F	46-50	47	0	Delhi	Central	
113	1003265	Arti	P00184942	F	26-35	35	0	Uttar Pradesh	Central	P
114	1003396	Akshay	P00178242	F	26-35	31	1	Delhi	Central	
115	1002380	Swati	P00124642	F	26-35	26	1	Delhi	Central	

In [5]:

```
pd.set_option('display.max_rows', 11251)
pd.set_option('display.max_columns', 15)
```

In [8]:

```
diwali.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
```

drop the unrelated / blank columns

In [9]:

```
diwali.drop(["Status","unnamed1"],axis=1,inplace=True)
```

In [10]:

```
diwali.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11251 entries, 0 to 11250
Data columns (total 13 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
dtypes: float64(1), int64(4), object(8)
memory usage: 1.1+ MB
```

check for any null values

In [14]:

```
diwali.isnull().sum()
```

Out[14]:

```
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
```

drop the null values

In [15]:

```
diwali.dropna(inplace=True)
```

In [16]:

```
diwali.isnull().sum()
```

Out[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              0
dtype: int64
```

Change the dataype of a columns

In [18]:

```
diwali.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 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  float64
dtypes: float64(1), int64(4), object(8)
memory usage: 1.2+ MB
```

In [22]:

```
diwali["Amount"]=diwali["Amount"].astype("int")
```

In [23]:

```
diwali.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 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  int32 
dtypes: int32(1), int64(4), object(8)
memory usage: 1.2+ MB
```

In [25]:

```
diwali.columns
```

Out[25]:

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

Rename the columns

In [31]:

```
diwali.rename(columns={"Marital_Status":'Marraige_Status'},inplace=True)
```

In [33]:

```
diwali.columns
```

Out[33]:

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

describe about the dataset

In [34]:

```
diwali.describe()
```

Out[34]:

	User_ID	Age	Marraige_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 [35]:

```
diwali[["Age","Orders","Amount"]].describe()
```

Out[35]:

	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

In [36]:

```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [37]:

```
diwali.columns
```

Out[37]:

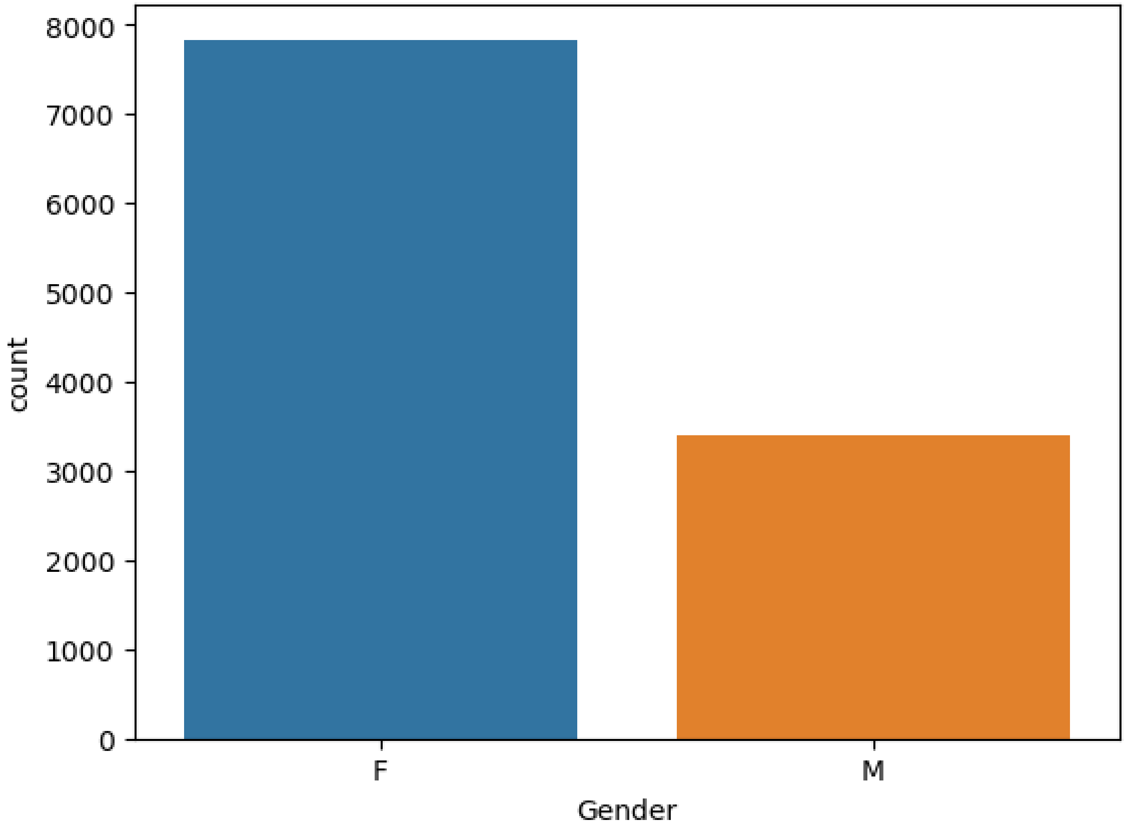
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marraige_Status', 'State', 'Zone', 'Occupation', 'Product_Categor
y',
       'Orders', 'Amount'],
      dtype='object')
```

In [38]:

```
sns.countplot(x="Gender",data=diwali)
```

Out[38]:

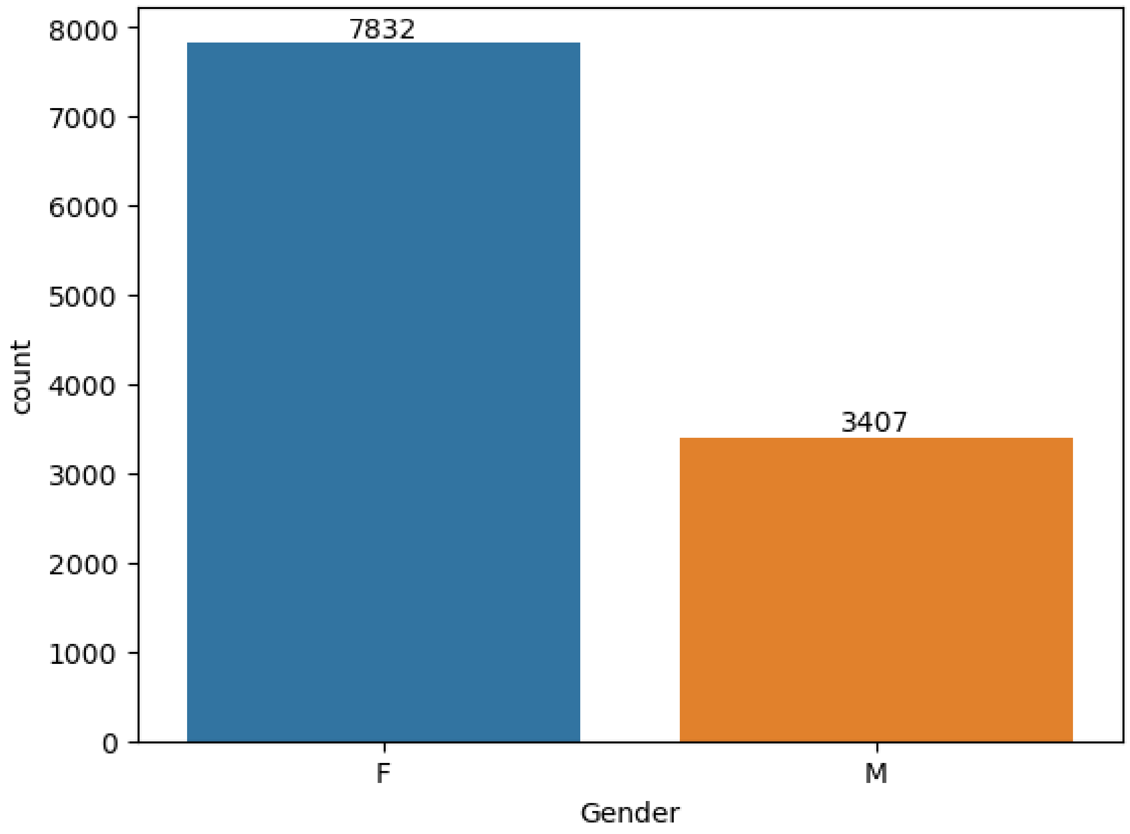
```
<Axes: xlabel='Gender', ylabel='count'>
```



In [39]:

```
ax=sns.countplot(x="Gender",data=diwali)

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



In [59]:

```
xx=diwali.groupby("Gender",as_index=False)["Amount"].sum().sort_values(by="Amount",ascend
xx
```

Out[59]:

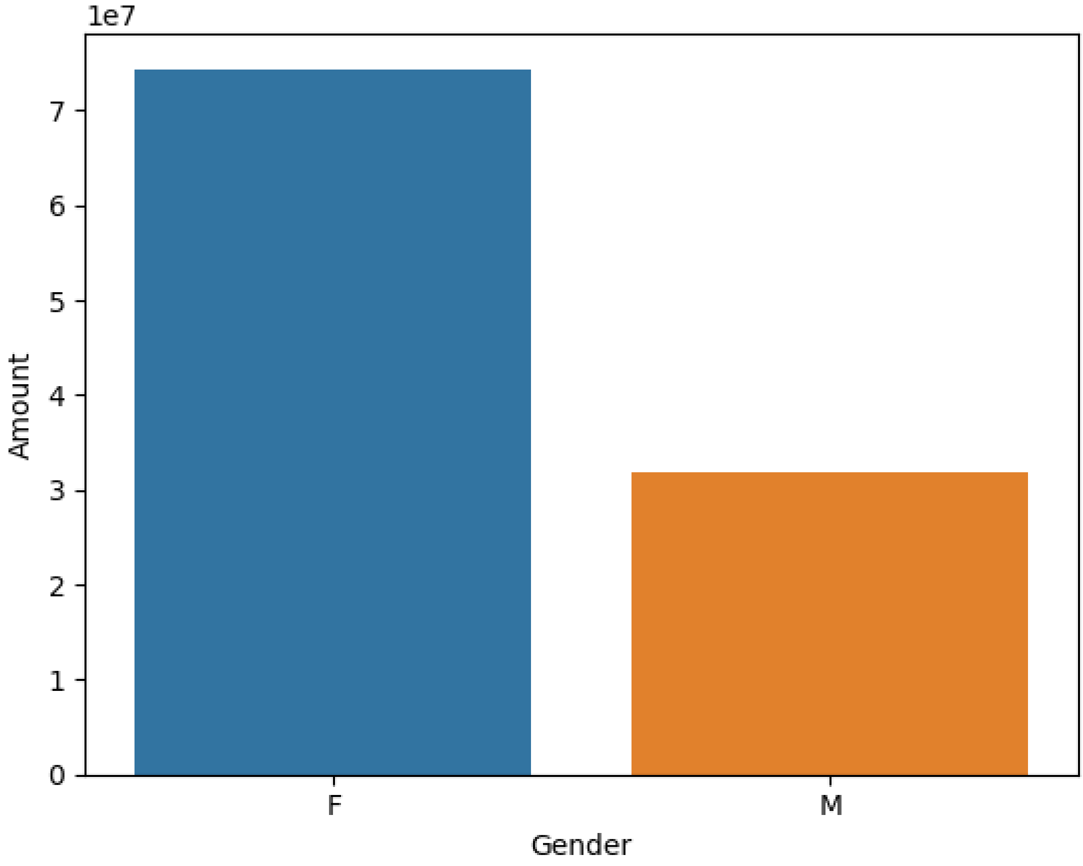
	Gender	Amount
0	F	74335853
1	M	31913276

In [61]:

```
sns.barplot(x="Gender",y="Amount",data=xx)
```

Out[61]:

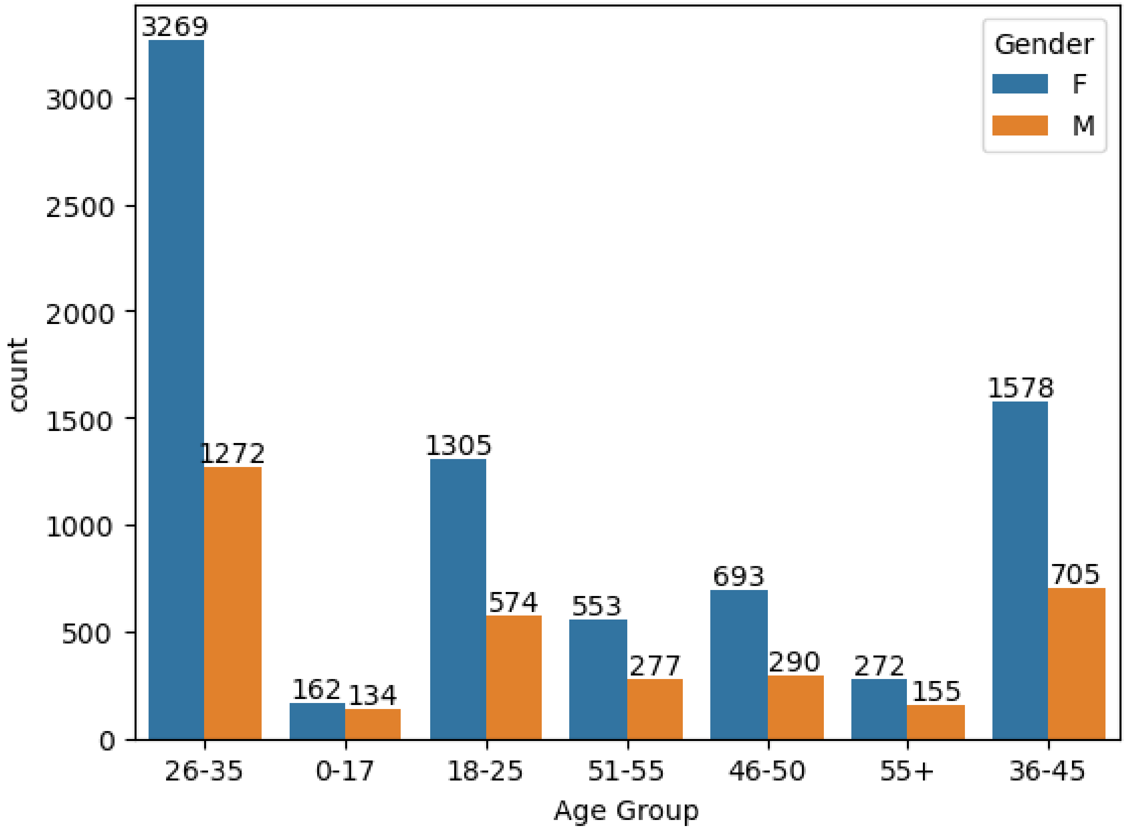
```
<Axes: xlabel='Gender', ylabel='Amount'>
```



age

In [63]:

```
tt=sns.countplot(x="Age Group",hue="Gender",data=diwali)
for bars in tt.containers:
    tt.bar_label(bars)
```



state

In [64]:

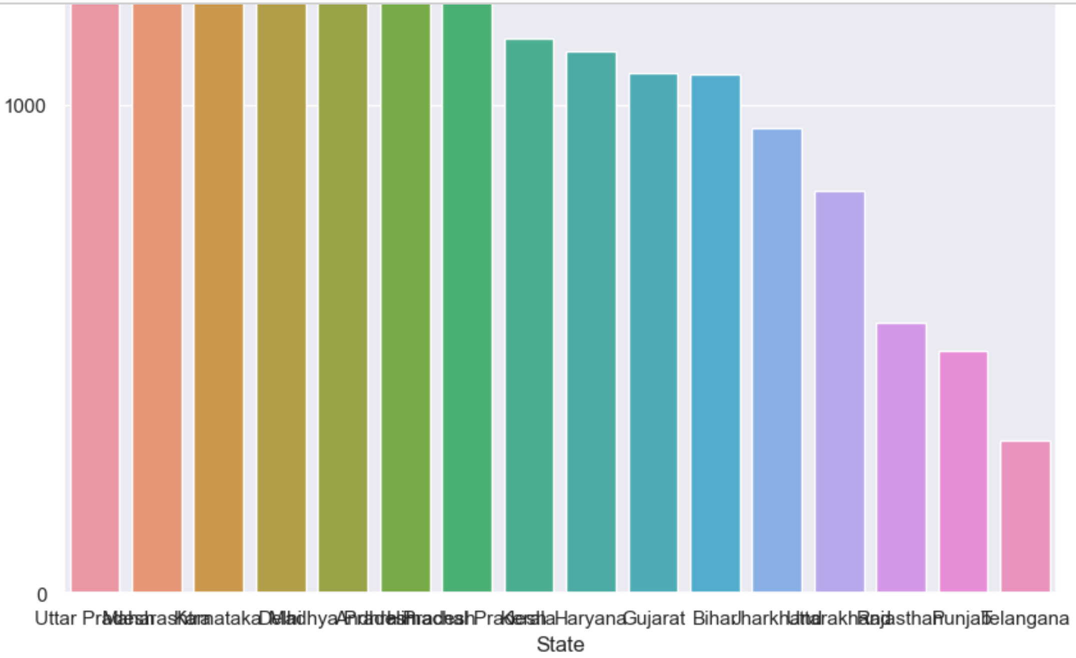
```
diwali.columns
```

Out[64]:

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

In [93]:

```
rr=diwali.groupby("State",as_index=False)["Orders"].sum().sort_values(by="Orders",ascendi
sns.set(rc={"figure.figsize":(10,25)})
sns.barplot(x="State",y="Orders",data=rr)
```

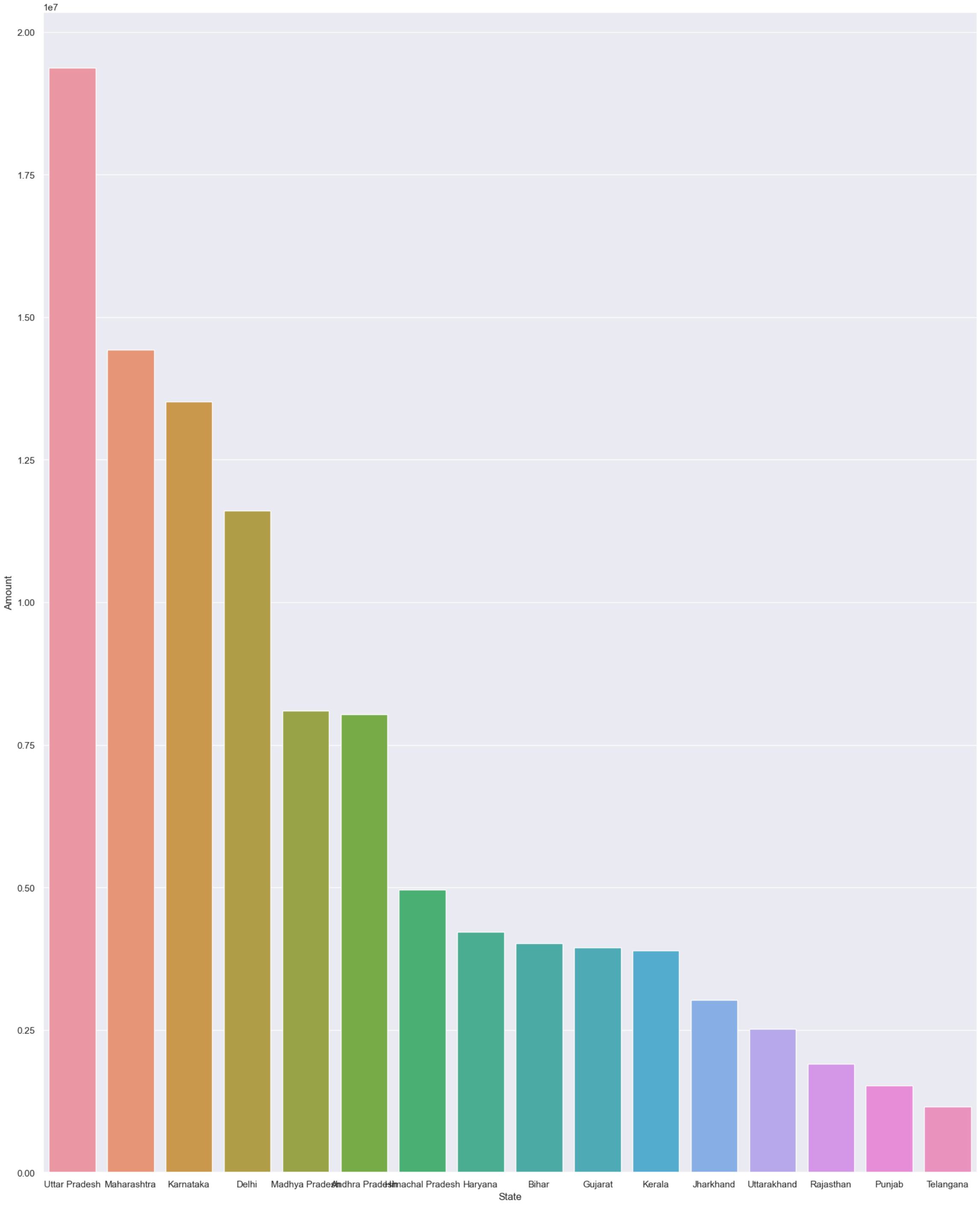


In [96]:

```
gr=diwali.groupby("State",as_index=False)["Amount"].sum().sort_values(by="Amount",ascendi
sns.set(rc={"figure.figsize":(20,25)})
sns.barplot(x="State",y="Amount",data=gr)
```

Out[96]:

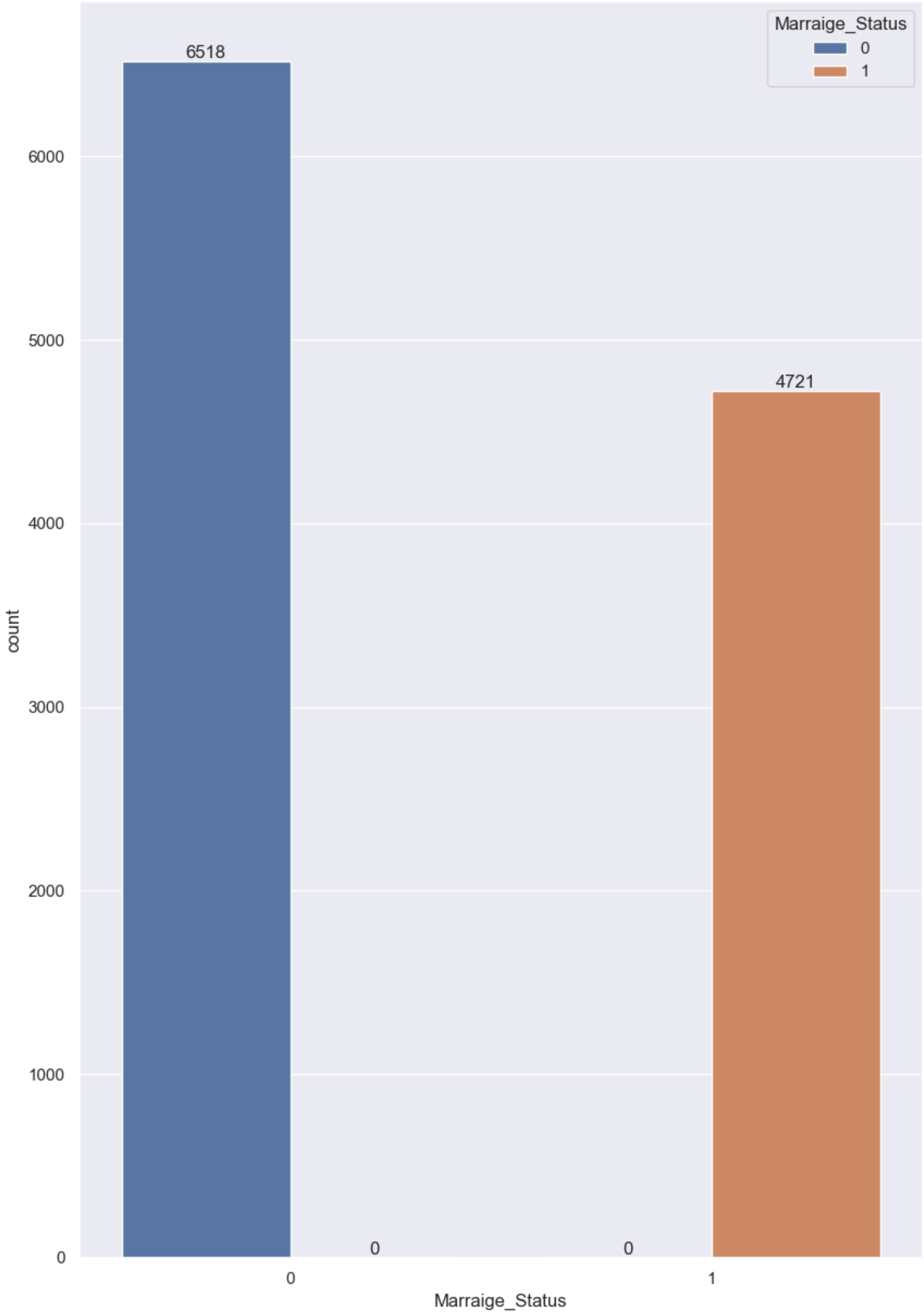
```
<Axes: xlabel='State', ylabel='Amount'>
```



marital status

In [102]:

```
yy=sns.countplot(x="Marraige_Status",data=diwali,hue="Marraige_Status")
sns.set(rc={"figure.figsize":(7,5)})
for bars in yy.containers:
    yy.bar_label(bars)
```

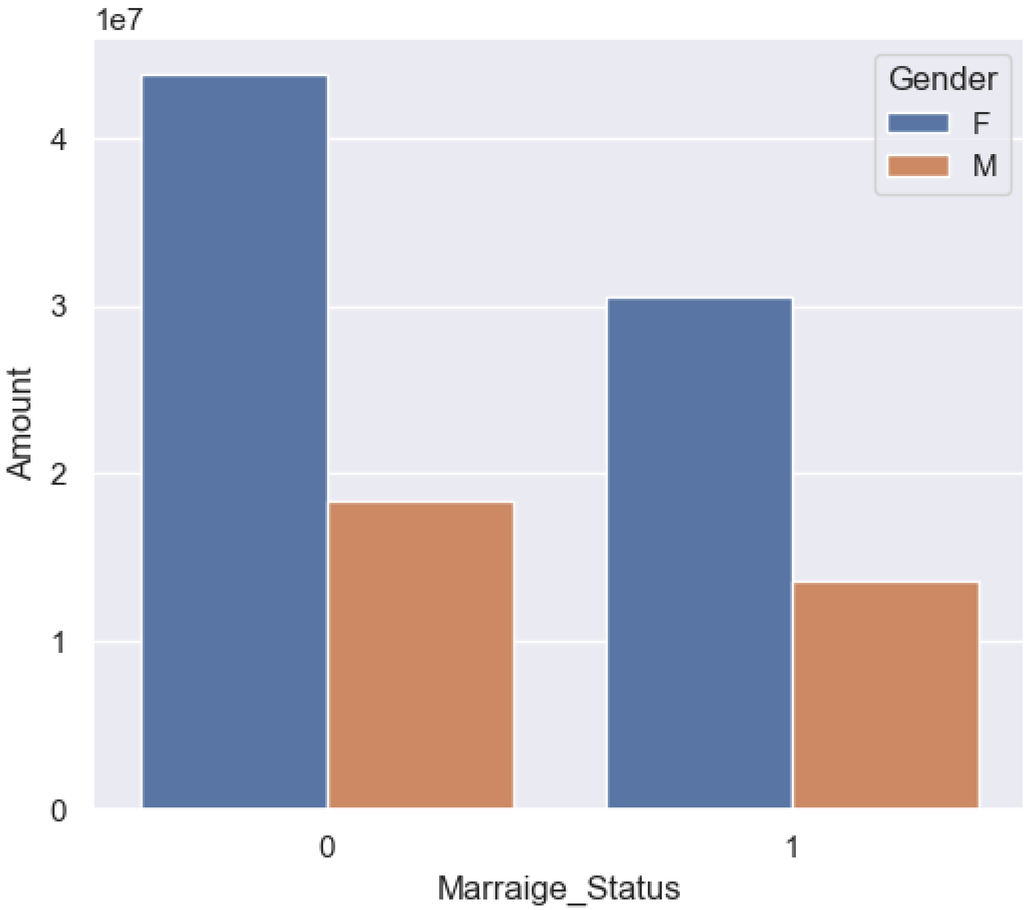


In [106]:

```
gr=diwali.groupby(["Marraige_Status","Gender"],as_index=False)["Amount"].sum().sort_value
sns.set(rc={"figure.figsize":(6,5)})
sns.barplot(x="Marraige_Status",y="Amount",data=gr,hue="Gender")
```

Out[106]:

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



Occupation

In [107]:

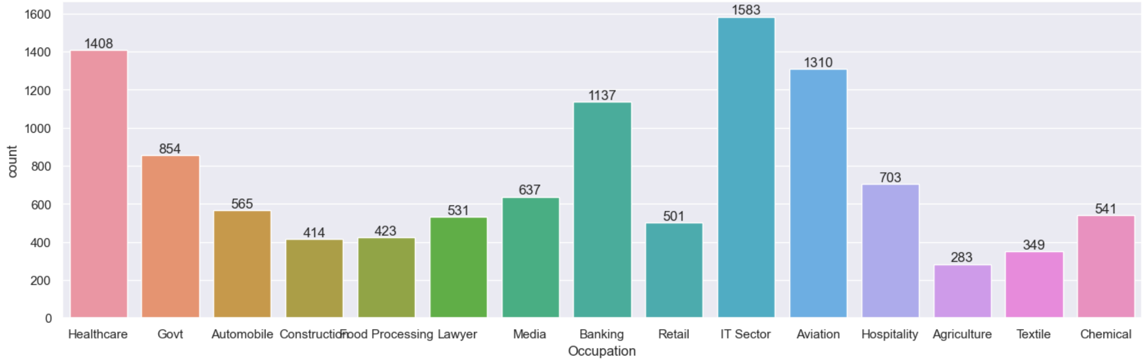
```
diwali.columns
```

Out[107]:

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

In [110]:

```
sy=sns.countplot(x="Occupation",data=diwali)
sns.set(rc={"figure.figsize":(17,15)})
for bars in sy.containers:
    sy.bar_label(bars)
```

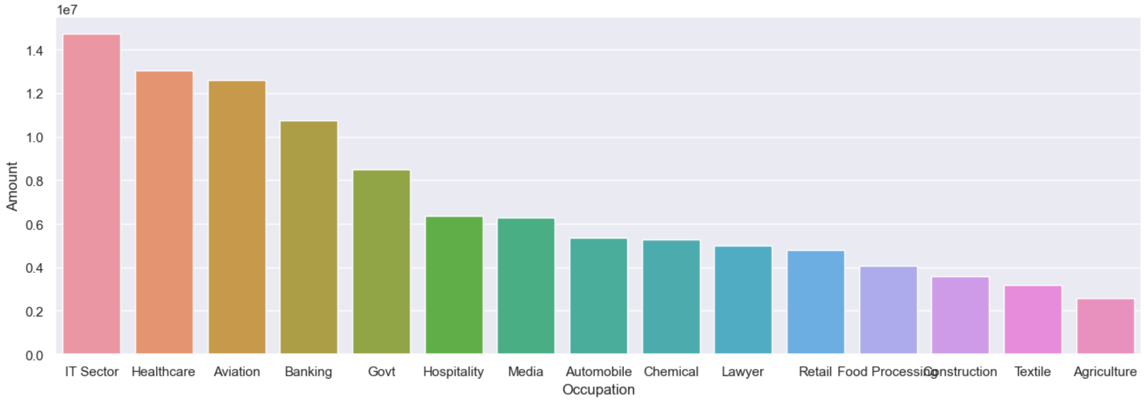


In [114]:

```
frr=diwali.groupby(["Occupation"],as_index=False)["Amount"].sum().sort_values(by="Amount"
sns.set(rc={"figure.figsize":(16,5)})
sns.barplot(x="Occupation",y="Amount",data=frr)
```

Out[114]:

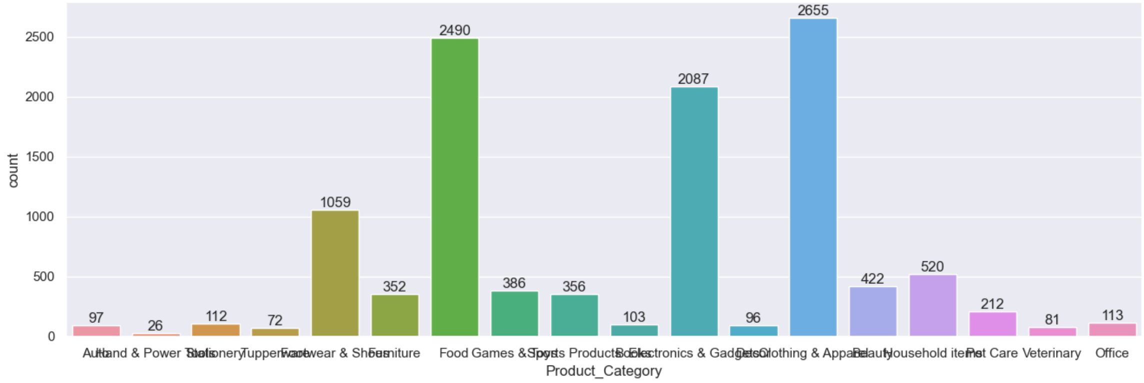
```
<Axes: xlabel='Occupation', ylabel='Amount'>
```



Product category

In [115]:

```
ssy=sns.countplot(x="Product_Category",data=diwali)
sns.set(rc={"figure.figsize":(17,15)})
for bars in ssy.containers:
    ssy.bar_label(bars)
```

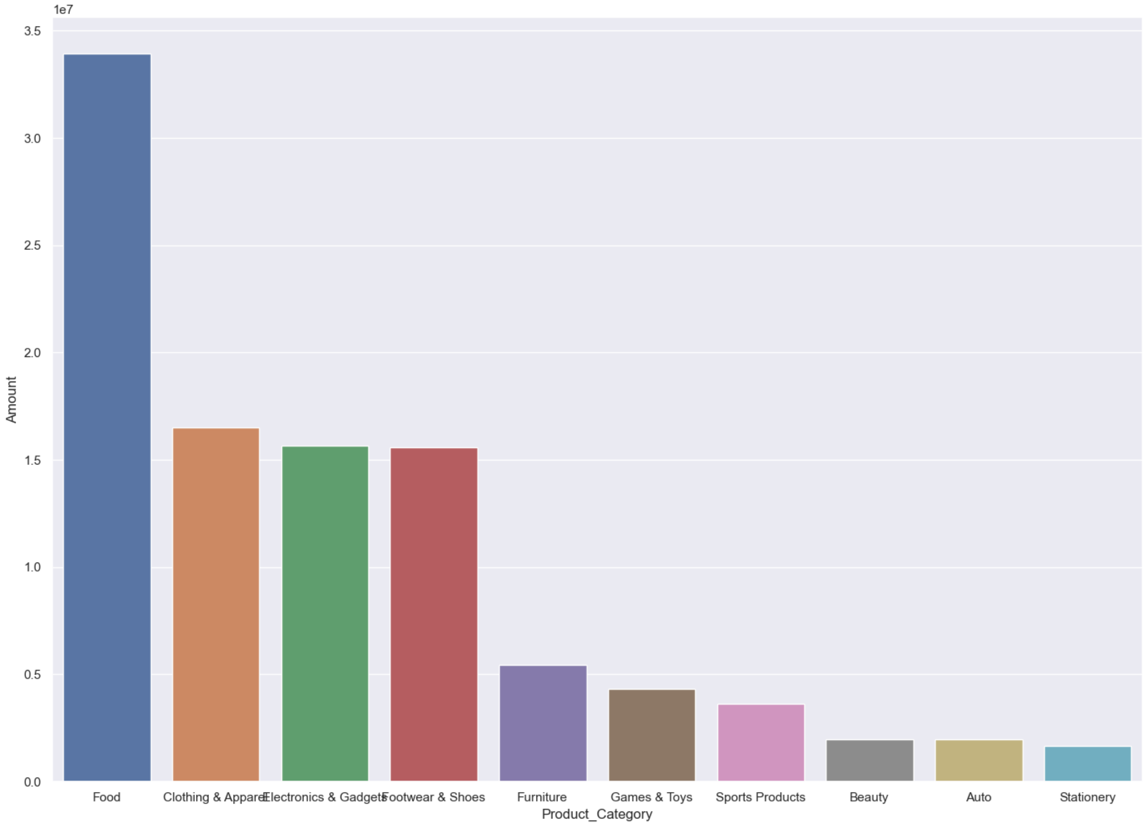


In [124]:

```
frr=diwali.groupby(["Product_Category"],as_index=False)["Amount"].sum().sort_values(by="A
sns.set(rc={"figure.figsize":(17,12)})
sns.barplot(x="Product_Category",y="Amount",data=frr)
```

Out[124]:

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



Product Id's

In [125]:

```
diwali.columns
```

Out[125]:

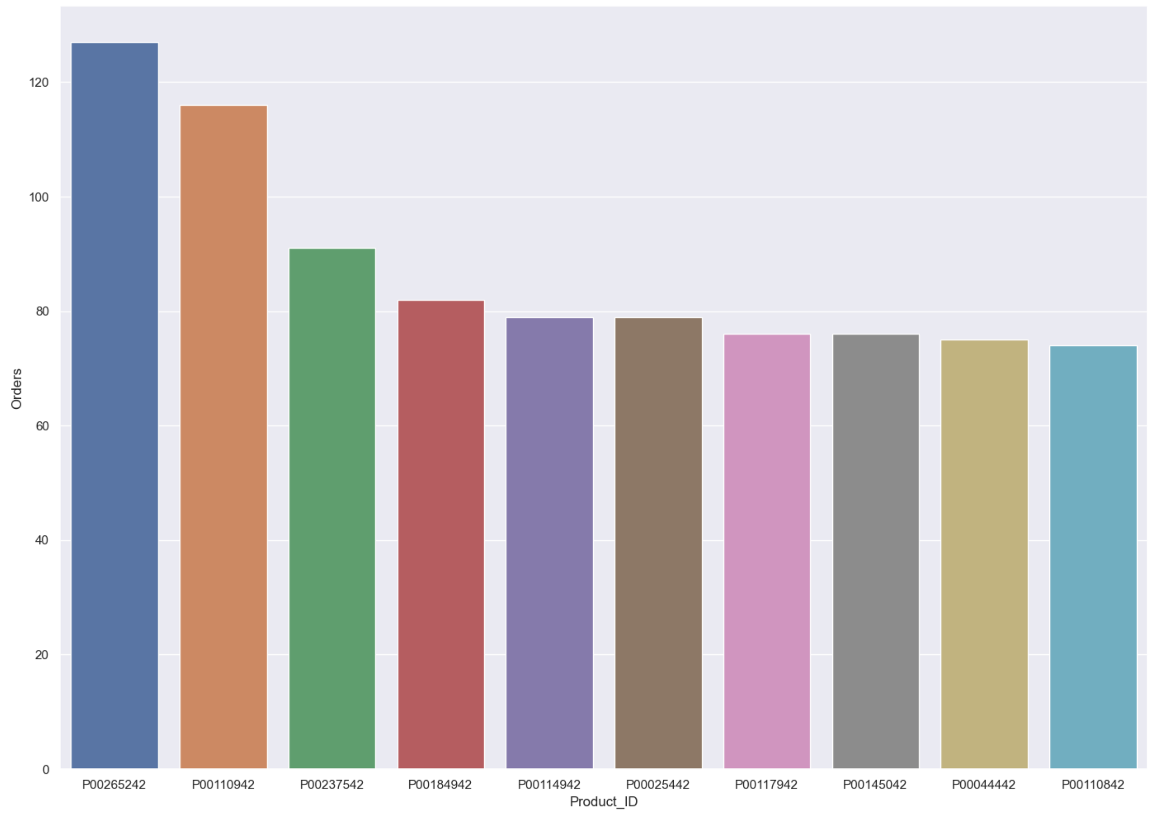
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marraige_Status', 'State', 'Zone', 'Occupation', 'Product_Categor
y',
       'Orders', 'Amount'],
      dtype='object')
```

In [127]:

```
frr=diwali.groupby(["Product_ID"],as_index=False)["Orders"].sum().sort_values(by="Orders'
sns.set(rc={"figure.figsize":(17,12)})
sns.barplot(x="Product_ID",y="Orders",data=frr)
```

Out[127]:

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



In [ ]:

```

```