Diwali Sales Analysis

In [16]:
```
import numpy as np
import pandas as pd
import matplotlib as mt
import matplotlib.pyplot as plt
import seaborn as sns
```

Import Data

In [26]:
```
df = pd.read_excel('D:\\2. Professional Data\\4. Data Analytics\\Shubham Deshmukh\\Python Project\\Diwali_Sales_Data.xlsx')
df.head(10)
```

Out[26]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Category	Orders	Amount	St
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto	1	23952.00	
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	Auto	3	23934.00	
2	1001990	Bindu	P00118542	F	26-35	35	1	Uttar Pradesh	Central	Automobile	Auto	3	23924.00	
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	Auto	2	23912.00	
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	Auto	2	23877.00	
5	1000588	Joni	P00057942	M	26-35	28	1	Himachal Pradesh	Northern	Food Processing	Auto	1	23877.00	
6	1001132	Balk	P00018042	F	18-25	25	1	Uttar Pradesh	Central	Lawyer	Auto	4	23841.00	
7	1002092	Shivangi	P00273442	F	55+	61	0	Maharashtra	Western	IT Sector	Auto	1	NaN	
8	1003224	Kushal	P00205642	M	26-35	35	0	Uttar Pradesh	Central	Govt	Auto	2	23809.00	
9	1003650	Ginny	P00031142	F	26-35	26	1	Andhra Pradesh	Southern	Media	Auto	4	23799.99	

Data Cleaning

In [24]:
```
df.shape
```

Out[24]:
```
(11251, 15)
```

In [25]:
```
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 [27]:
```
df.drop(['Status', 'unnamed1'], axis = 1, inplace = True)
```

In [35]:
```
df.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
```

In [38]:
```
df.isnull().sum()
```

Out[38]:
```
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 [41]:
```
df.shape
```

Out[41]:
```
(11251, 13)
```

In [42]:
```
df.dropna(inplace = True)
```

In [43]:
```
df.shape
```

Out[43]:
```
(11239, 13)
```

In [44]:
```
df.isnull().sum()
```

Out[44]:
```
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 [45]:
```
df['Amount'] = df['Amount'].astype('int')
```

In [46]:
```
df['Amount'].dtype
```

Out[46]:
```
dtype('int32')
```

In [48]:
```
df.columns
```

Out[48]:
```
Index(['User_ID', 'Cust_name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

In [55]:
```
df[['Age', 'Orders','Amount']].describe()
```

Out[55]:

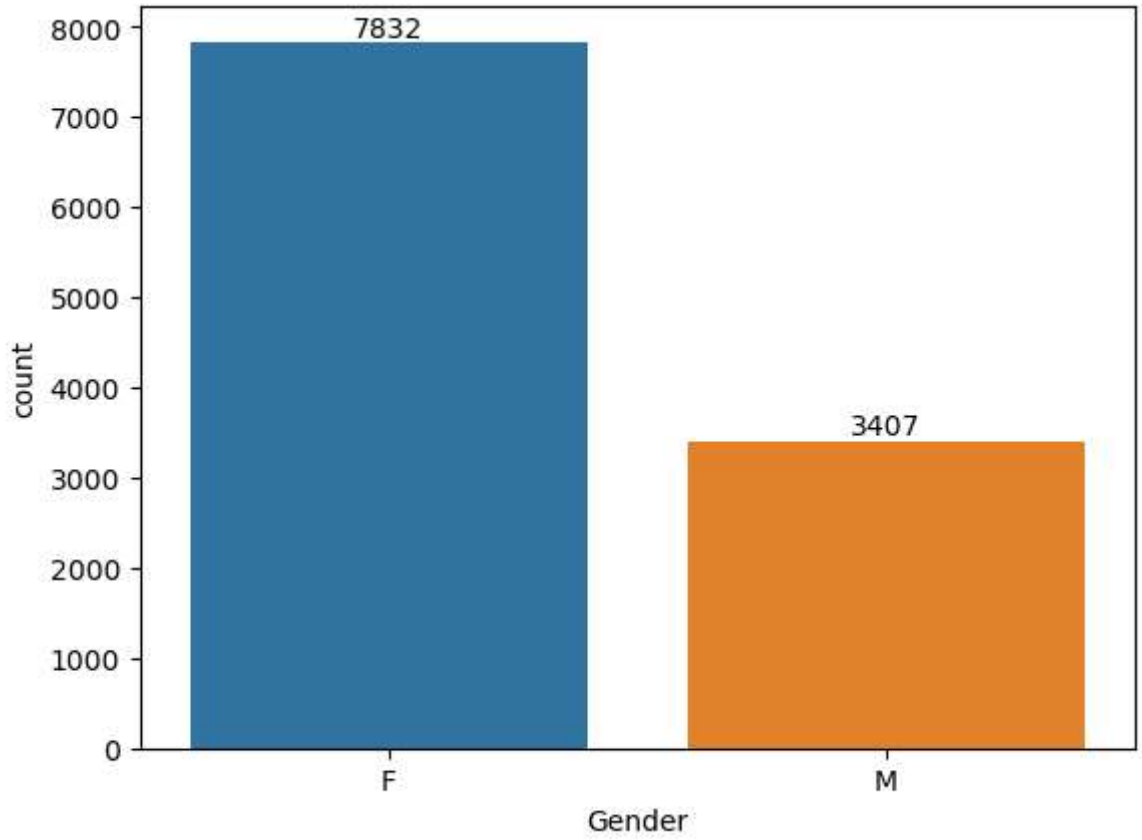
	Age	Orders	Amount
count	11239.000000	11239.000000	11239.000000
mean	35.410357	2.489634	9453.610553
std	12.753866	1.114967	5222.355168
min	12.000000	1.000000	188.000000
25%	27.000000	2.000000	5443.000000
50%	33.000000	2.000000	8109.000000
75%	43.000000	3.000000	12675.000000
max	92.000000	4.000000	23952.000000

Data Analysis

Gender

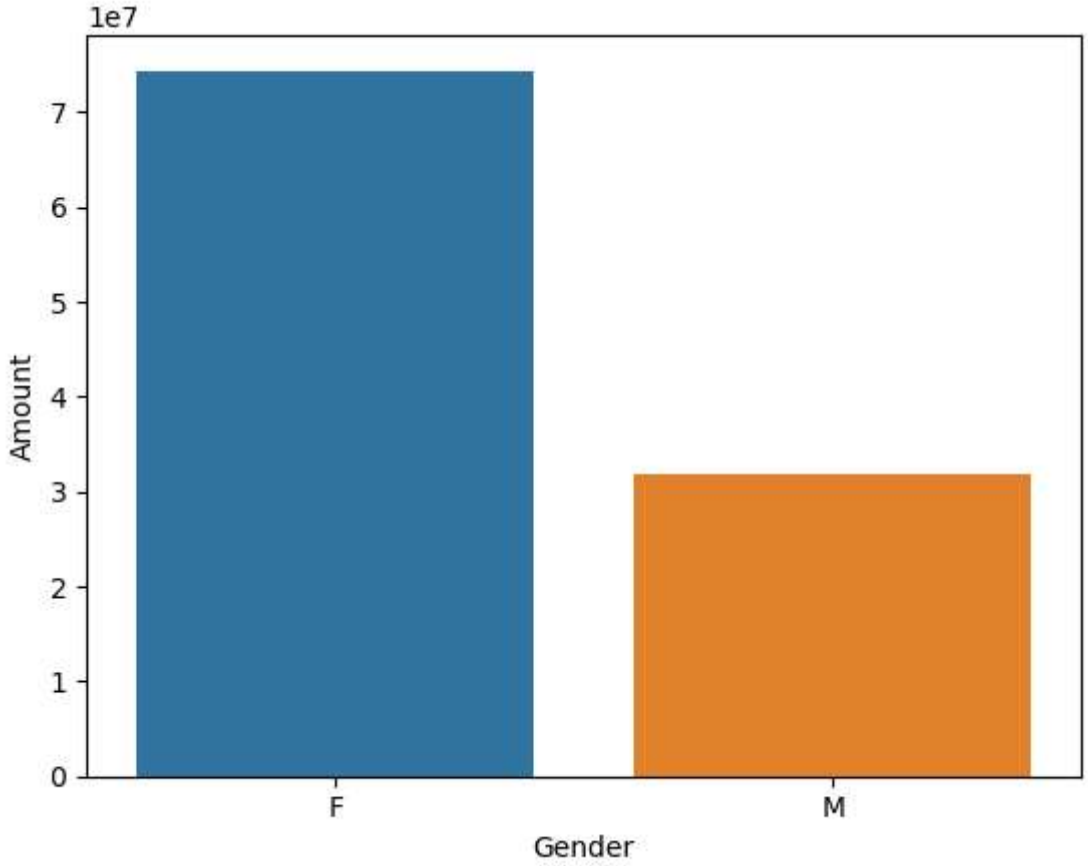
In [59]:

```
a = sns.countplot(x = 'Gender', data = df)

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



In [71]:
```
sales_gender = df.groupby(['Gender'], as_index = False)['Amount'].sum().sort_values(by = 'Amount', ascending = False)

sns.barplot(x = 'Gender', y = 'Amount', data = sales_gender)
```

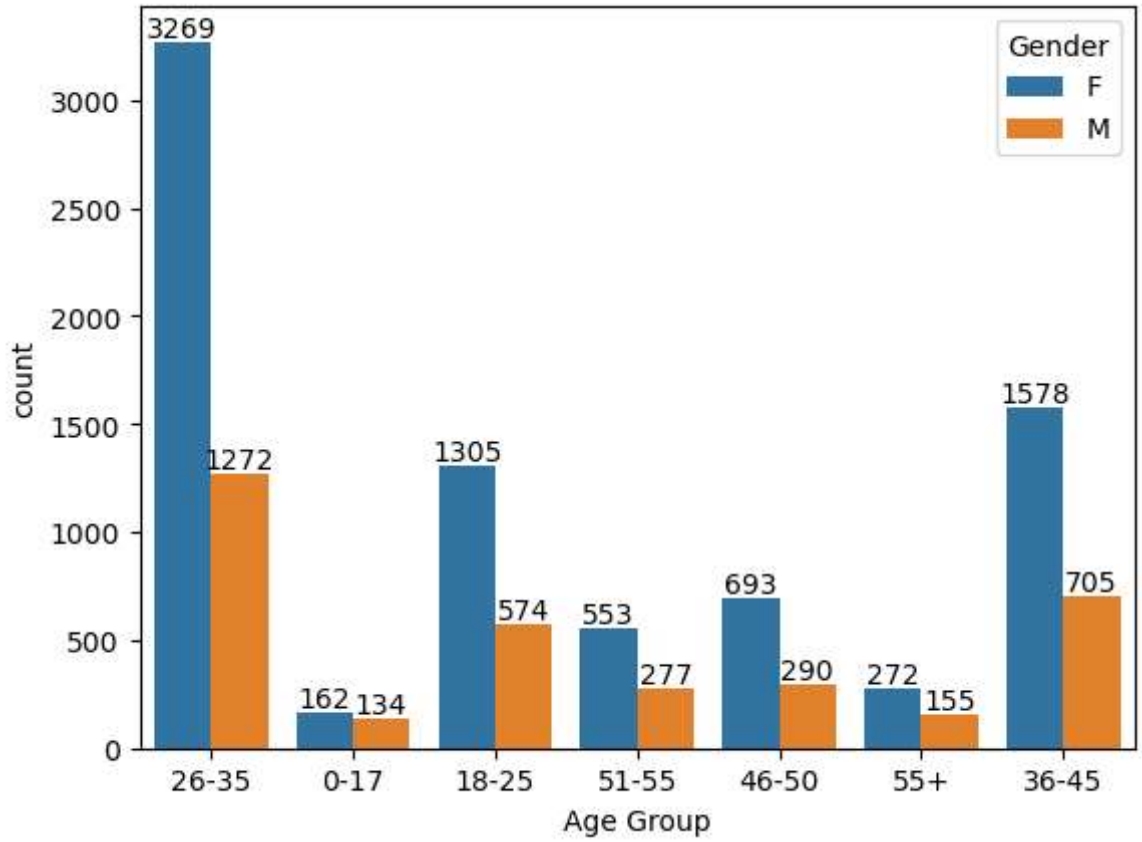
Out[71]:
```
<Axes: xlabel='Gender', ylabel='Amount'>
```



Here, we can observe that female customers are more than male customers and purchasing power of female is more than male.

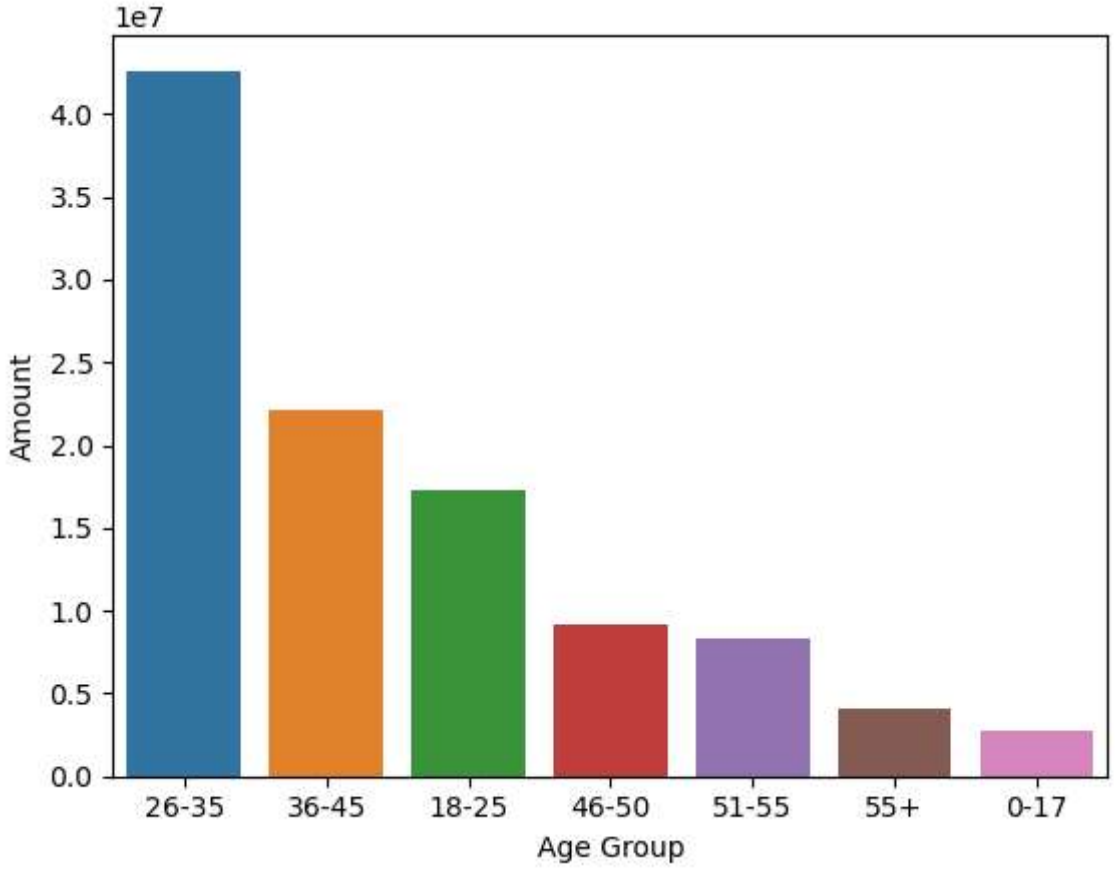
Age

In [74]:
```
age = sns.countplot(data = df, x = 'Age Group', hue = 'Gender')

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



In [75]:
```
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[75]:
```
<Axes: xlabel='Age Group', ylabel='Amount'>
```

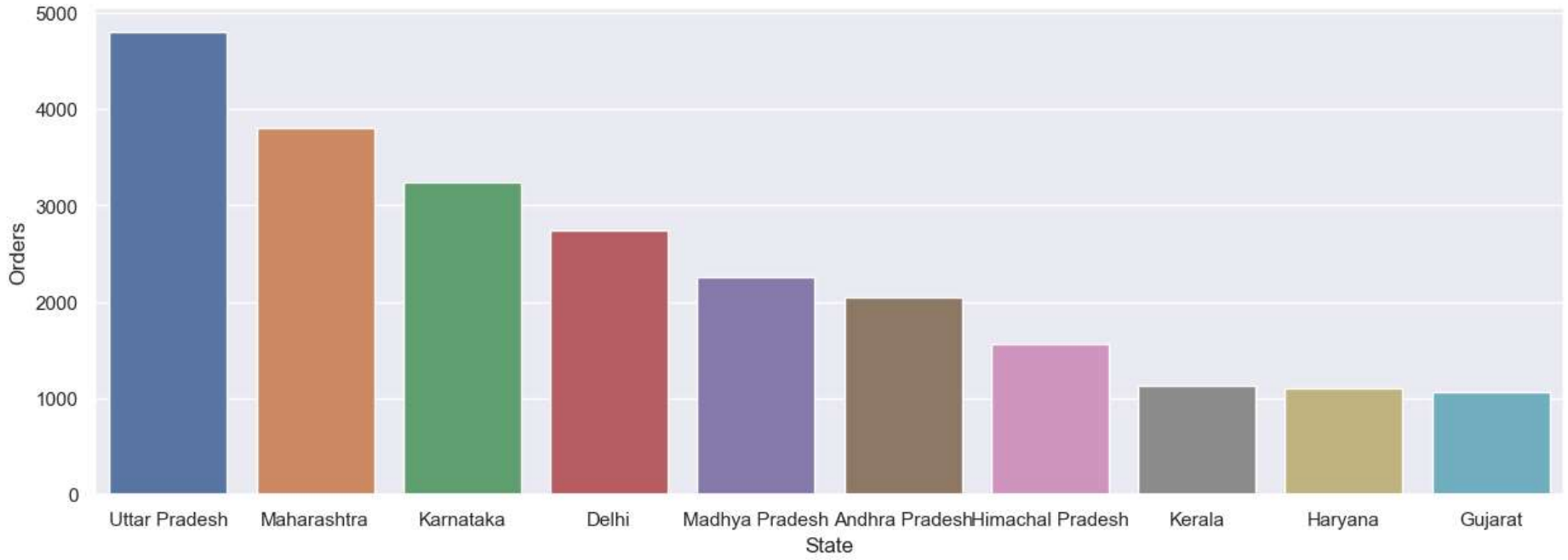


Here, we can observe that most of the customers are female of age group 26-35 years.

State

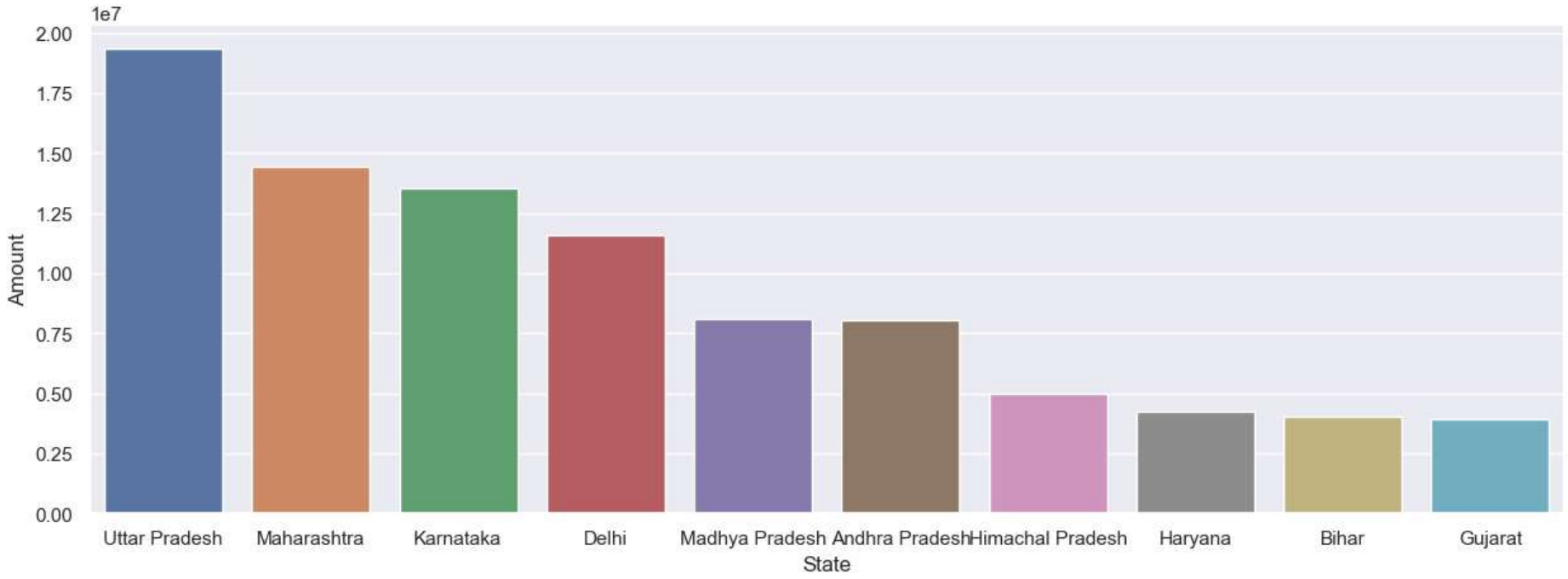
In [77]:
```
order_state = df.groupby(['State'], as_index = False)['Orders'].sum().sort_values(by = 'Orders', ascending = False).head(10)
sns.set(rc={'figure.figsize' : (15,5)})
sns.barplot(x = 'State', y = 'Orders', data = order_state)
```

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



In [78]:
```
amount_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(x = 'State', y = 'Amount', data = amount_state)
```

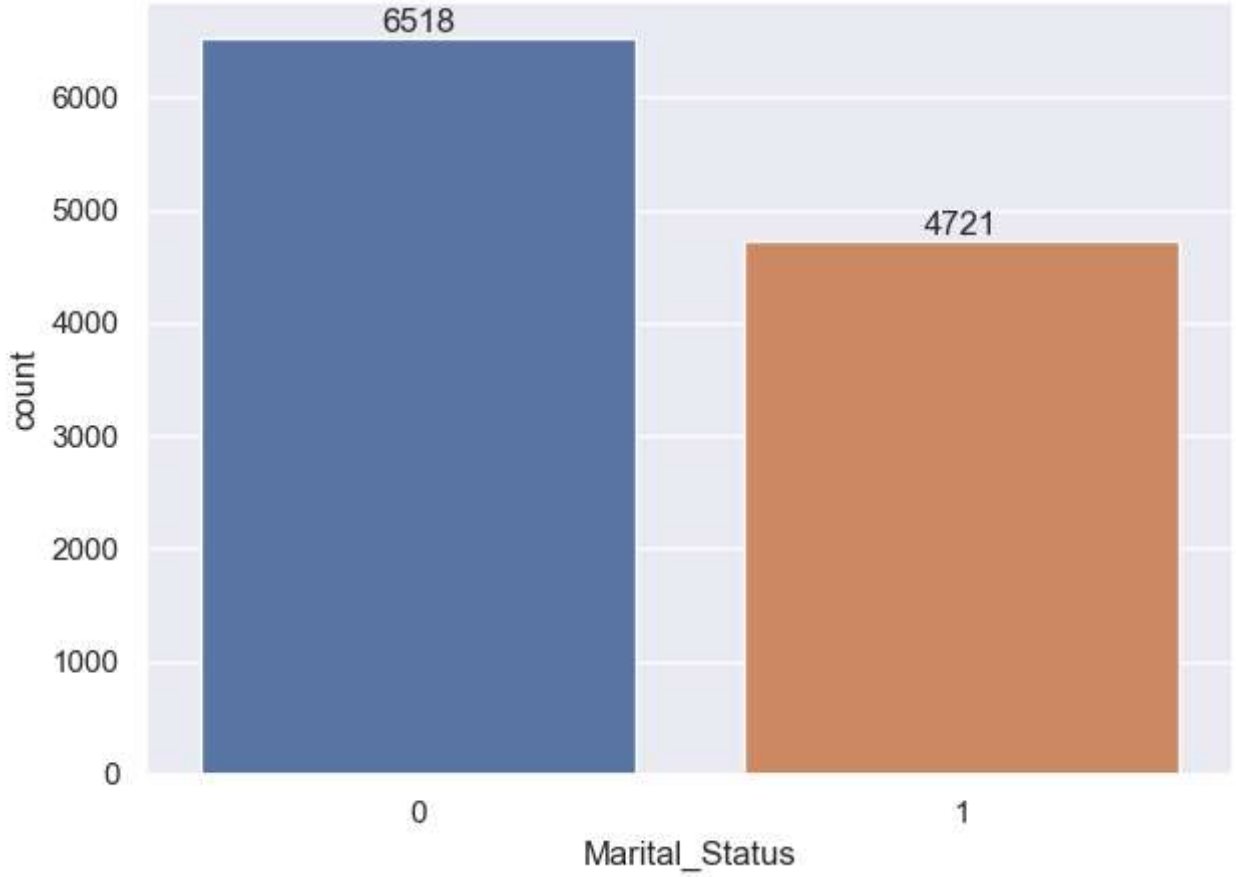
Out[78]:
```
<Axes: xlabel='State', ylabel='Amount'>
```



Here, we can observe that most of the orders and total sales are from Uttar Pradesh, Maharashtra and Karnataka.

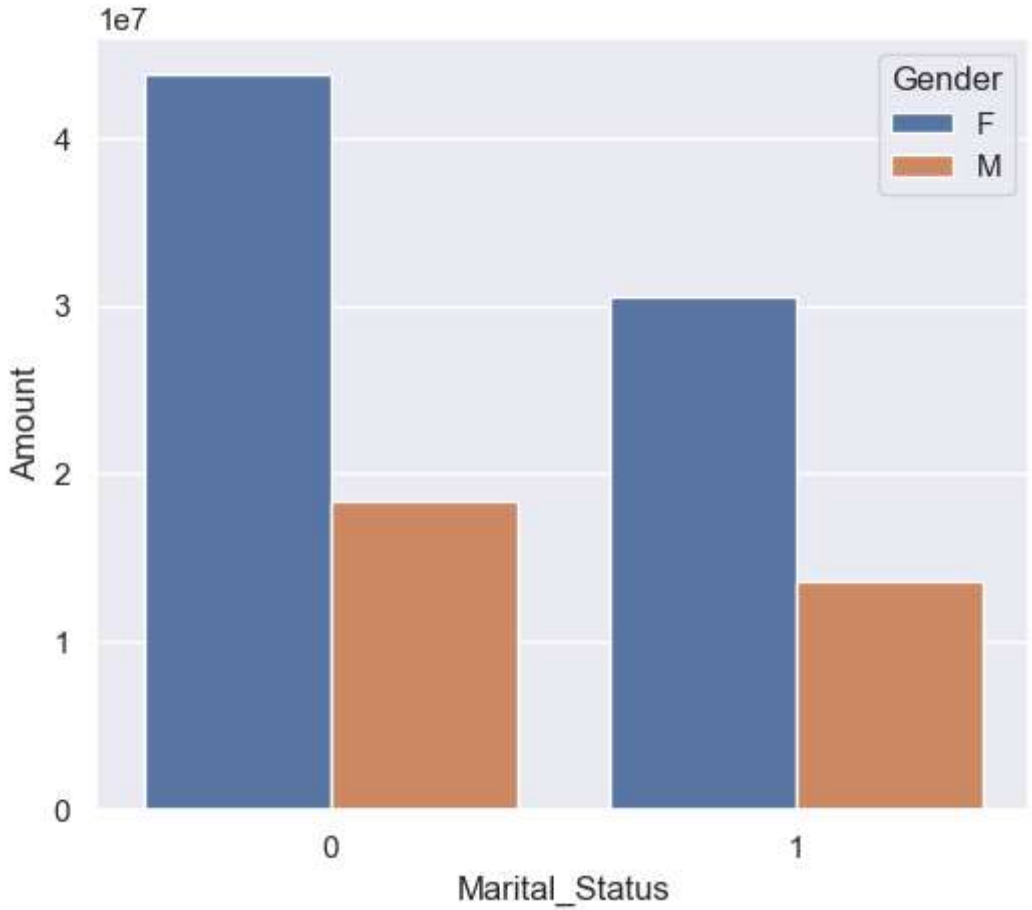
Maratial Status

In [82]:
```
mstatus = sns.countplot(data = df, x = 'Marital_Status')
sns.set(rc={'figure.figsize' : (6,5)})
for bars in mstatus.containers:
    mstatus.bar_label(bars)
```



In [85]:
```
amount_mstatus = df.groupby(['Marital_Status', 'Gender'], as_index = False)['Amount'].sum().sort_values(by = 'Amount', ascending
sns.set(rc={'figure.figsize' : (6,5)})
sns.barplot(x = 'Marital_Status', y = 'Amount', data = amount_mstatus, hue='Gender')
```

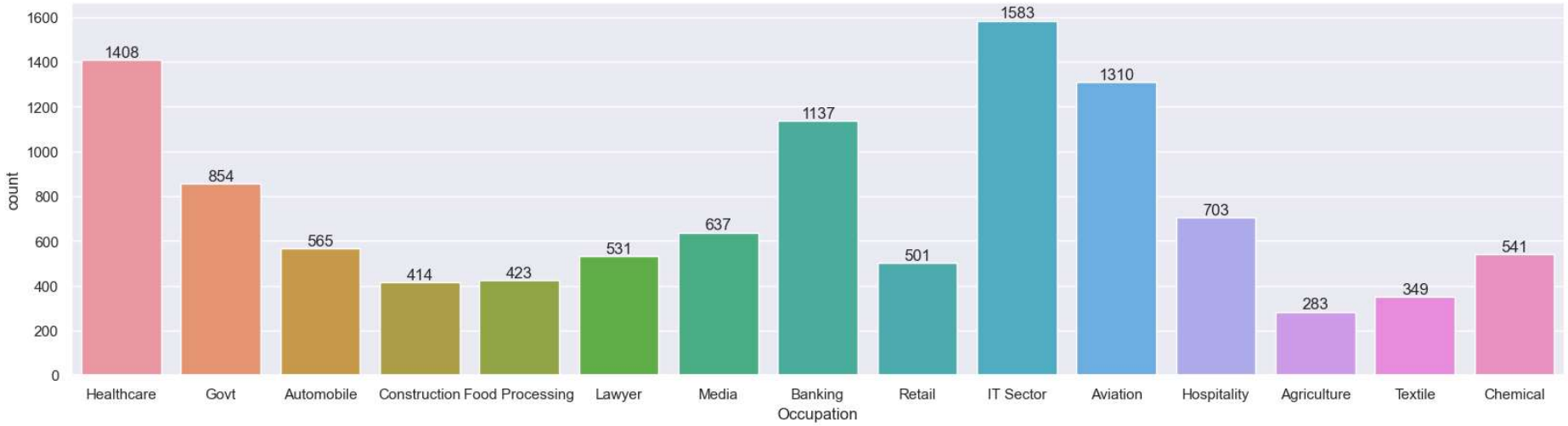
Out[85]:
```
<Axes: xlabel='Marital_Status', ylabel='Amount'>
```



Here, we can observe that most of the customers are married but unmarried female has more purchasing power.

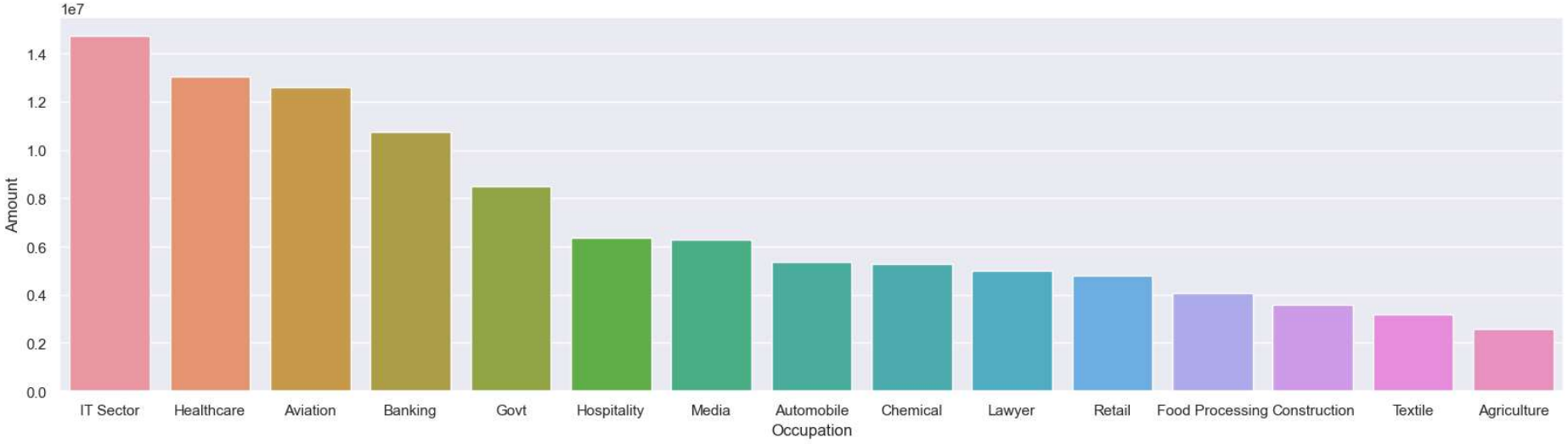
Occupation

In [88]:
```
ostatus = sns.countplot(data = df, x = 'Occupation')
sns.set(rc={'figure.figsize' : (20,5)})
for bars in ostatus.containers:
    ostatus.bar_label(bars)
```



In [89]:
```
amount_ostatus = df.groupby(['Occupation'], as_index = False)['Amount'].sum().sort_values(by = 'Amount', ascending = False)
sns.set(rc={'figure.figsize' : (20,5)})
sns.barplot(x = 'Occupation', y = 'Amount', data = amount_ostatus)
```

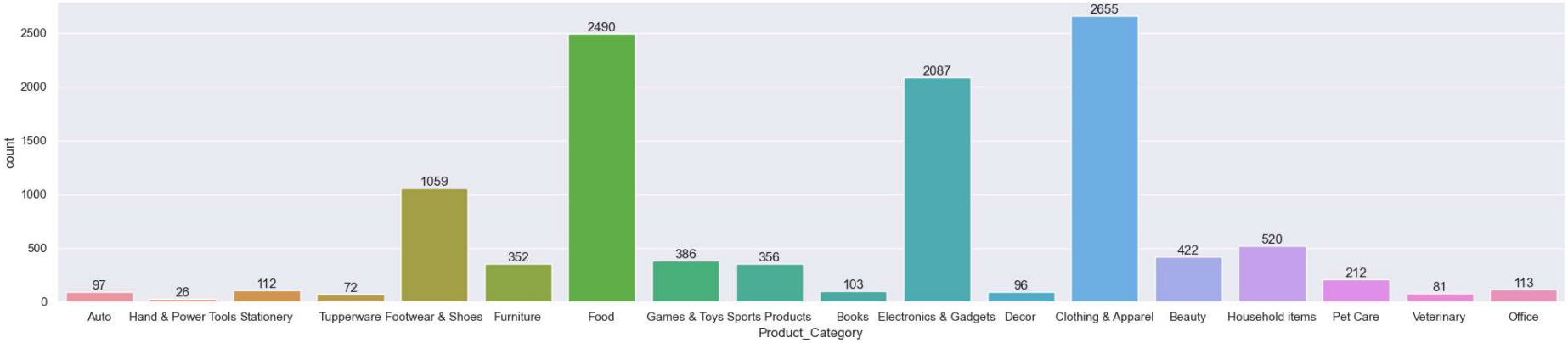
Out[89]:
```
<Axes: xlabel='Occupation', ylabel='Amount'>
```



Here, we can observe that most of the buyers are working in IT, Healthcare and Aviation sector.

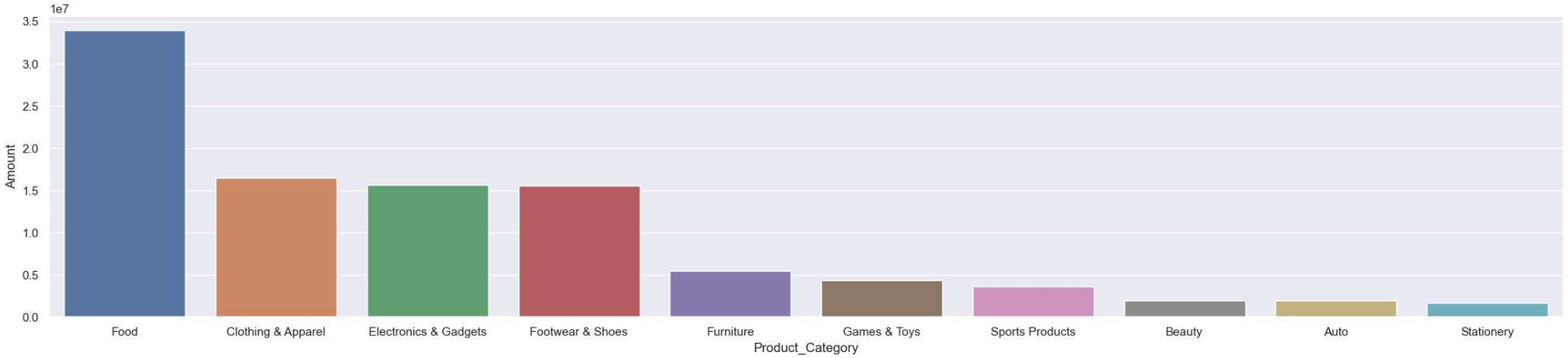
Product Category

In [98]:
```
pcstatus = sns.countplot(data = df, x = 'Product_Category')
sns.set(rc={'figure.figsize' : (25,5)})
for bars in pcstatus.containers:
    pcstatus.bar_label(bars)
```



In [96]:
```
amount_pcstatus = df.groupby(['Product_Category'], as_index = False)['Amount'].sum().sort_values(by = 'Amount',
                                                                                                  ascending = False).head(10)
sns.set(rc={'figure.figsize' : (25,5)})
sns.barplot(x = 'Product_Category', y = 'Amount', data = amount_pcstatus)
```

Out[96]:
```
<Axes: xlabel='Product_Category', ylabel='Amount'>
```



Here, we can observe that most of the sold products are from Food, Clothing and Electronics category.

Conclusion:

1. Female customers are more than male customers and purchasing power of female is more than male.

2. Most of the customers are female of age group 26-35 years.

3. Most of the orders and total sales are from Uttar Pradesh, Maharashtra and Karnataka.

4. Most of the customers are married but unmarried female has more purchasing power.

5. Most of the buyers are working in IT, Healthcare and Aviation sector.

6. Most of the sold products are from Food, Clothing and Electronics category.

Thank You

Shubham Deshmukh

Data Analyst (SQL, Tableau, Adv Excel, Power Query, Python)

Email ID: deshmukhsv3@gmail.com

Phone: +91 9096159803

LinkedIN: www.linkedin.com/in/shubhamdeshmukh3