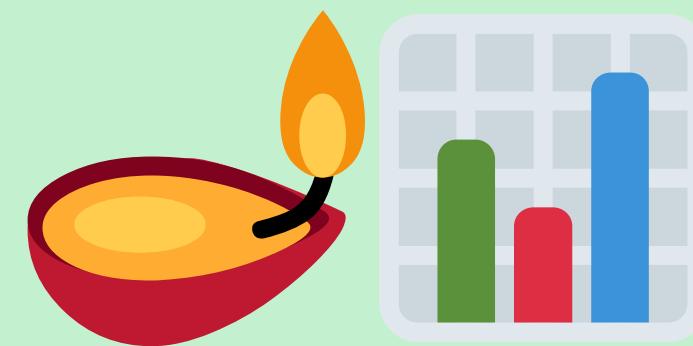


# DIWALI SALES ANALYSIS



Key Insights from Customer Purchase Data


By Aryan Anand

# OBJECTIVE

ANALYZE DIWALI SALES  
DATA TO UNCOVER:

- Top performing states
- Buying behavior by gender & age
- Popular product categories





# Importing Libraries

```
# importing libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt # visualizing data
%matplotlib inline
import seaborn as sns
```

## uploading csv files

```
# importing CSV file
df = pd.read_csv("Diwali Sales Data.csv")
```

```
df.shape ## Returns a tuple (number of rows, number of columns)
```

```
(11251, 15)
```



# Print First 5 rows

```
df.head(5) # print first 5 rows and columns
```

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Category	Orders	Amount	Status	unnamed:0
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto	1	23952.0	NaN	NaN
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	Auto	3	23934.0	NaN	NaN
2	1001990	Bindu	P00118542	F	26-35	35	1	Uttar Pradesh	Central	Automobile	Auto	3	23924.0	NaN	NaN
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	Auto	2	23912.0	NaN	NaN
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	Auto	2	23877.0	NaN	NaN

# Print Last 5 rows

```
df.tail(5) # print last 5 rows and columns
```

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Category	Orders	Amount	Status	unnamed:0
11246	1000695	Manning	P00296942	M	18-25	19	1	Maharashtra	Western	Chemical	Office	4	370.0	NaN	NaN
11247	1004089	Reichenbach	P00171342	M	26-35	33	0	Haryana	Northern	Healthcare	Veterinary	3	367.0	NaN	NaN
11248	1001209	Oshin	P00201342	F	36-45	40	0	Madhya Pradesh	Central	Textile	Office	4	213.0	NaN	NaN
11249	1004023	Noonan	P00059442	M	36-45	37	0	Karnataka	Southern	Agriculture	Office	3	206.0	NaN	NaN
11250	1002744	Brumley	P00281742	F	18-25	19	0	Maharashtra	Western	Healthcare	Office	3	188.0	NaN	NaN

# Data Cleaning

```
df.info()
```

```
# drop unrelated/blank columns  
df.drop(["Status", "unnamed1"], axis=1, inplace=True)  
df # df is modified
```

```
# drop null values  
df.dropna(inplace = True)  
pd.isnull(df).sum()
```

- Used df.info() to inspect data types and structure before and after cleaning.
- Removed unnecessary columns like "Status" and "unnamed1" to streamline the dataset.
- Dropped all rows containing missing values to ensure clean and accurate analysis.



# Change the DataType/Rename Column

```
df["Amount"] = df["Amount"].astype("int")  
df["Amount"].dtypes  
  
dtype('int64')
```

- Converted the "Amount" column to integer (int64) to enable numerical operations and visualizations.

```
df.rename(columns = {"Marital_Status" : "Shadi"})
```

- Renamed the "Marital\_Status" column to "Shadi" to simplify naming and reflect a more relatable label.

```
# dreturns description of the data in the dataframe  
df.describe()
```

Generated statistical summary for numerical columns using df.describe().

	User_ID	Age	Marital_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

- Average Age: ~35 years
- Average Orders per user: ~2.5
- Amount Range: ₹188 to ₹23,952

# Expploratory Data Analysis

## Gender:

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

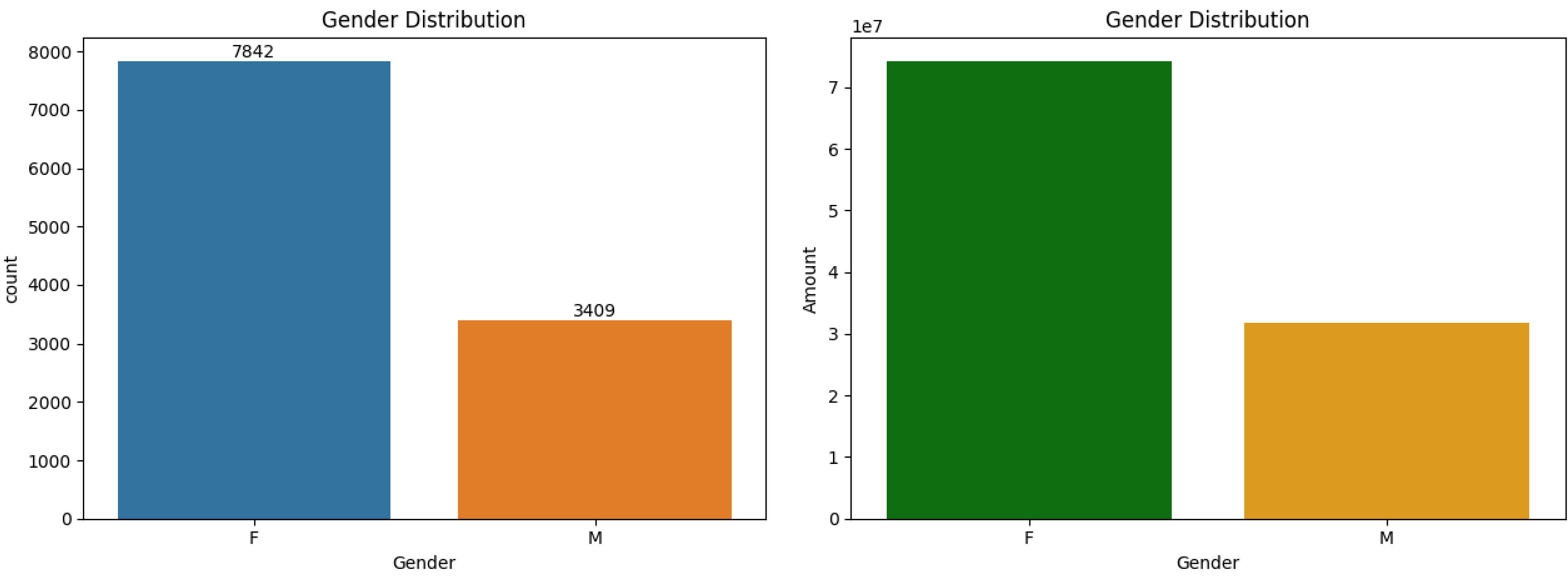
for bars in ax.containers:
    ax.bar_label(bars)
plt.title("Gender Distribution")
plt.tight_layout()
plt.savefig('Gender_Distribution.png')
```

```
df.groupby(["Gender"] , as_index=False )["Amount"].sum().sort_values(by = "Amount" , ascending= False)
```

	Gender	Amount
0	F	74335856.43
1	M	31913276.00

```
sales_gen = df.groupby(["Gender"] , as_index=False )["Amount"].sum().sort_values(by = "Amount" , ascending= False )
sns.barplot(x = "Gender" , y = "Amount", hue="Gender" , data = sales_gen,palette= {"M":"orange" , "F" : "green"} )
plt.title("Gender Distribution")
```





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

# Age:

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

*# Total Amount vs Age Group with Gender-wise split*

```
sales_age_gender = df.groupby(['Age Group', 'Gender'], as_index=False)['Amount'].sum().sort_values(by='Amount', ascending=False)
```

*# Barplot with hue*

```
sns.barplot(x='Age Group', y='Amount', hue='Gender', data=sales_age_gender)
```

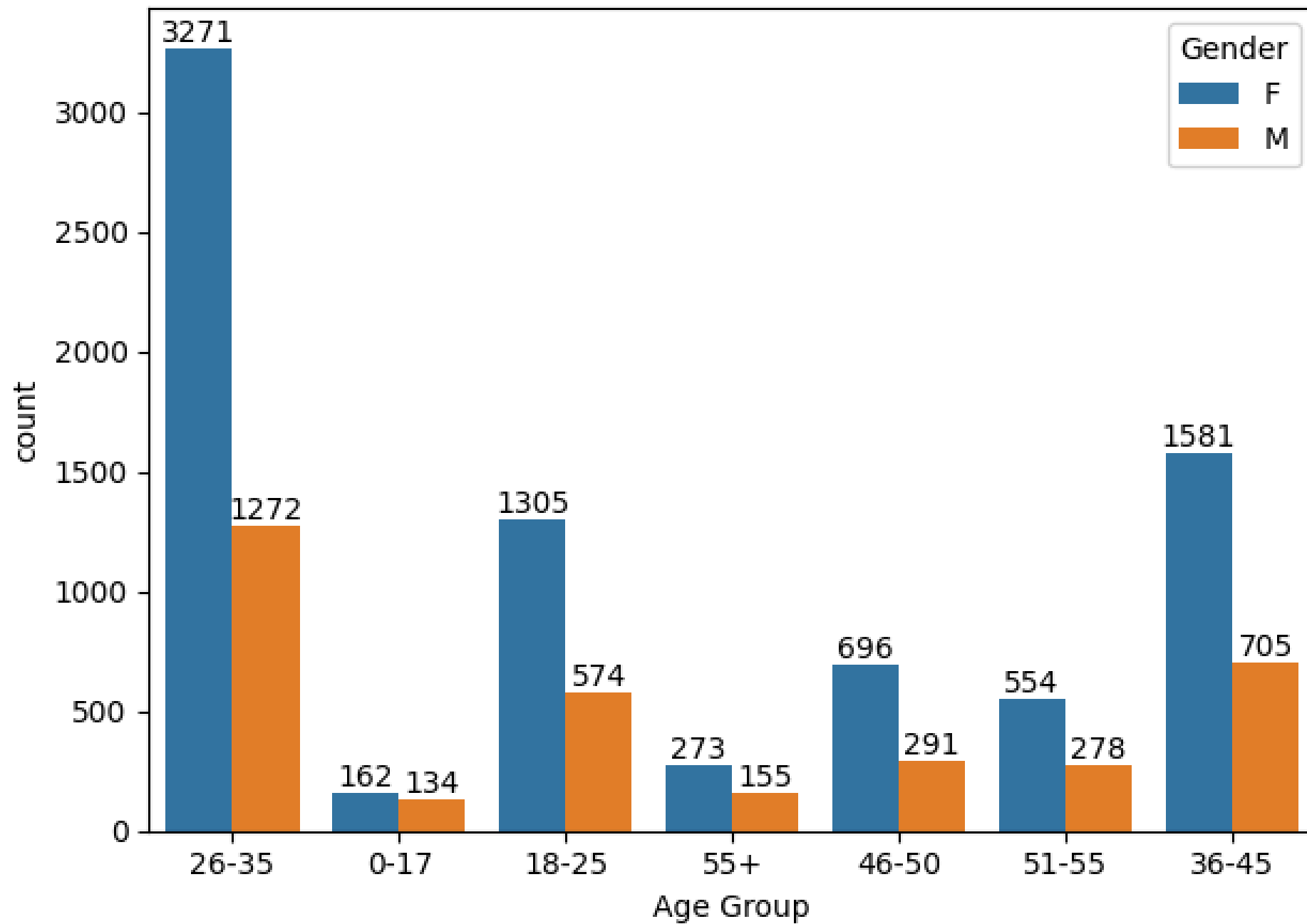
*# Add title*

```
plt.title("Total Amount by Age Group and Gender")
```

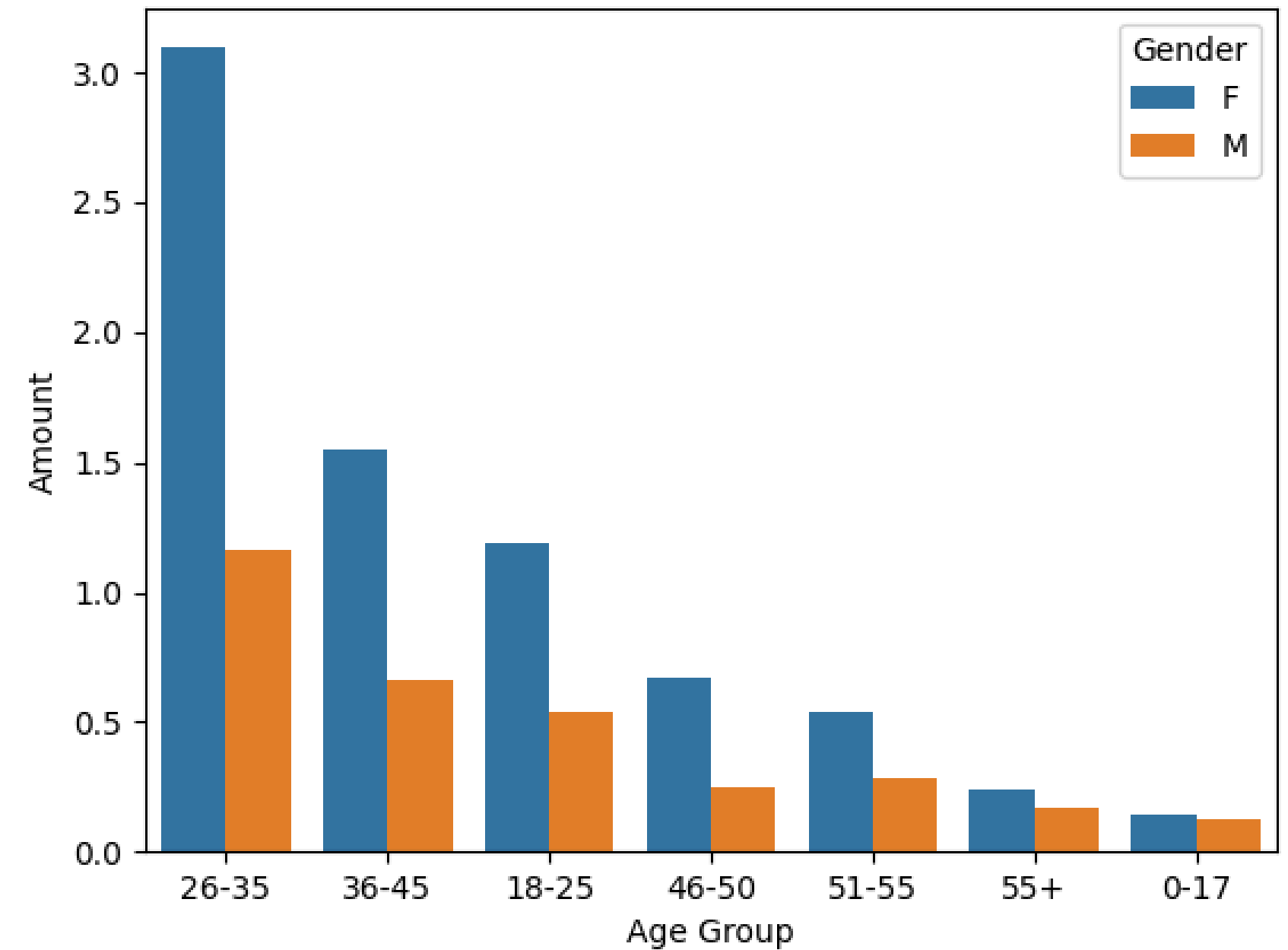
```
plt.savefig('Total_Amount_Age_Group.png')
```

```
plt.show()
```

Total Age Group



Total Amount by Age Group and Gender

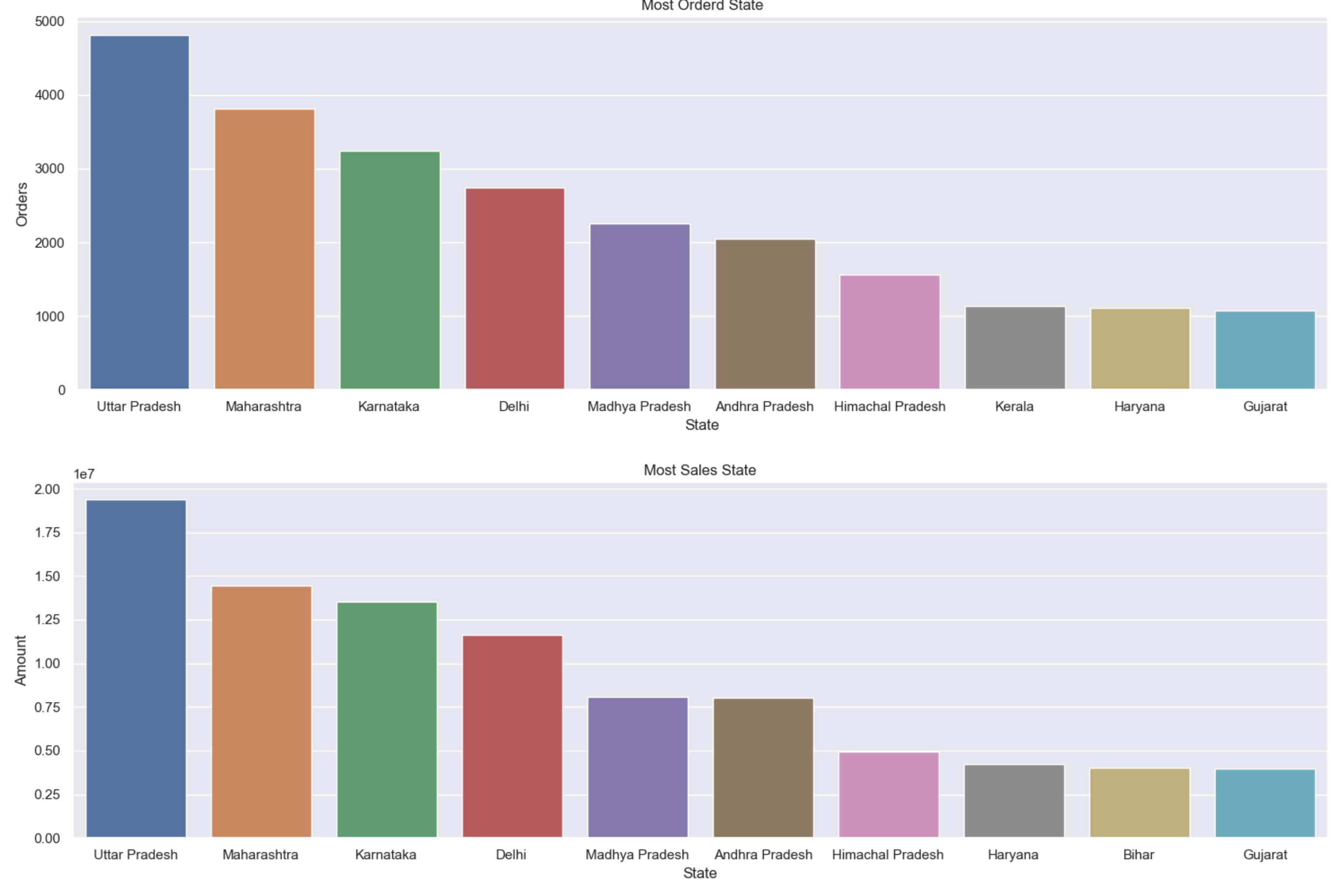


**From above graphs we can see that most of the buyers are of age group between 26-35 yrs female**

# STATE :

```
# total numbers of orders from top 10 states
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' , hue = 'State' , data = Order_state)
plt.tight_layout()
plt.title("Most Orderd State")
plt.savefig('State_BY_Order.png')
```

```
# total amount/sales from top 10 states
sales_sate = 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',hue = 'State' , data = sales_sate)
plt.title("Most Sales State")
plt.tight_layout()
plt.savefig('State_BY_Order2.png')
```



From above graphs we can see that most of the orders & total sales/amount are from Uttar Pradesh, Maharashtra and Karnataka respectively

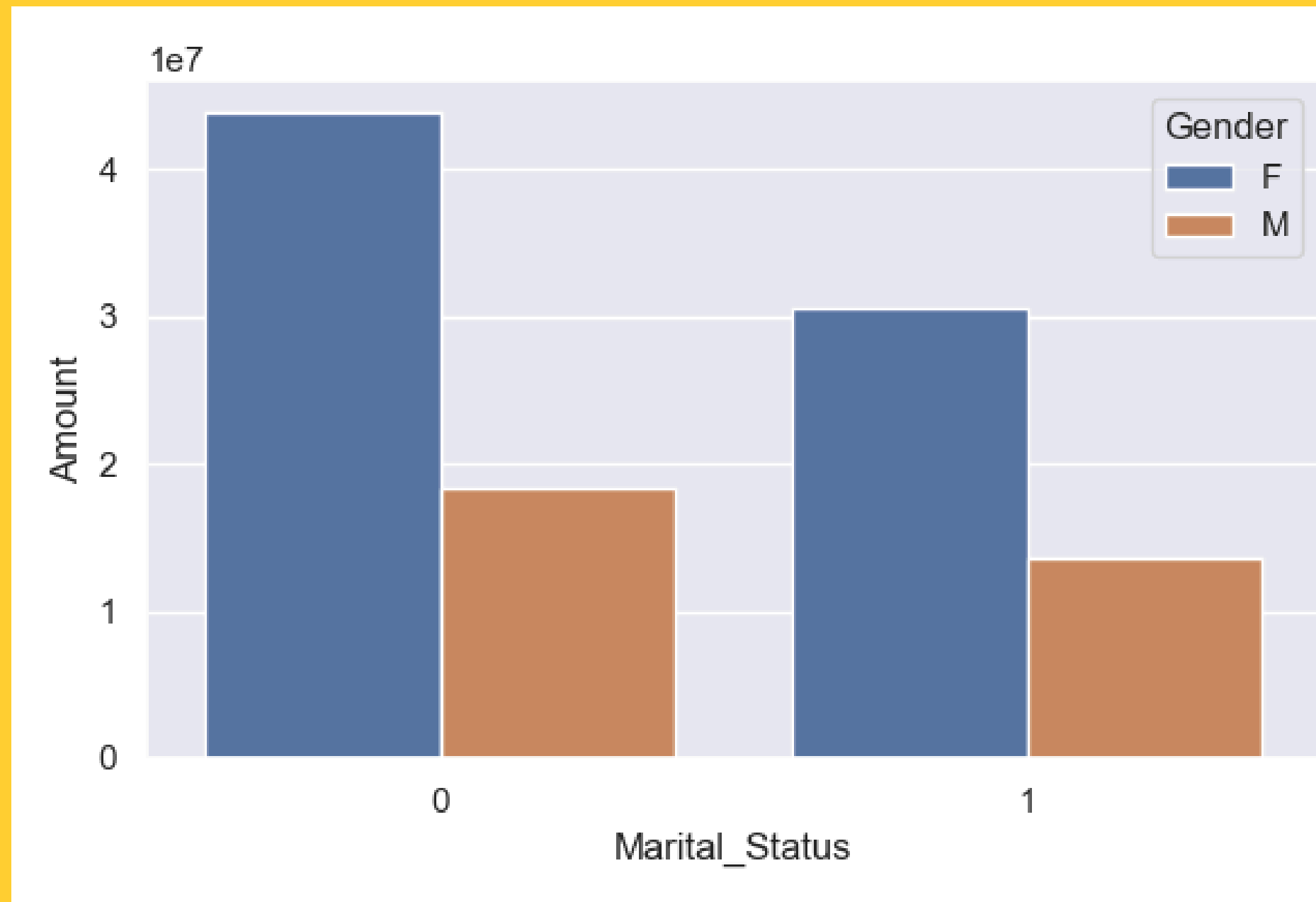
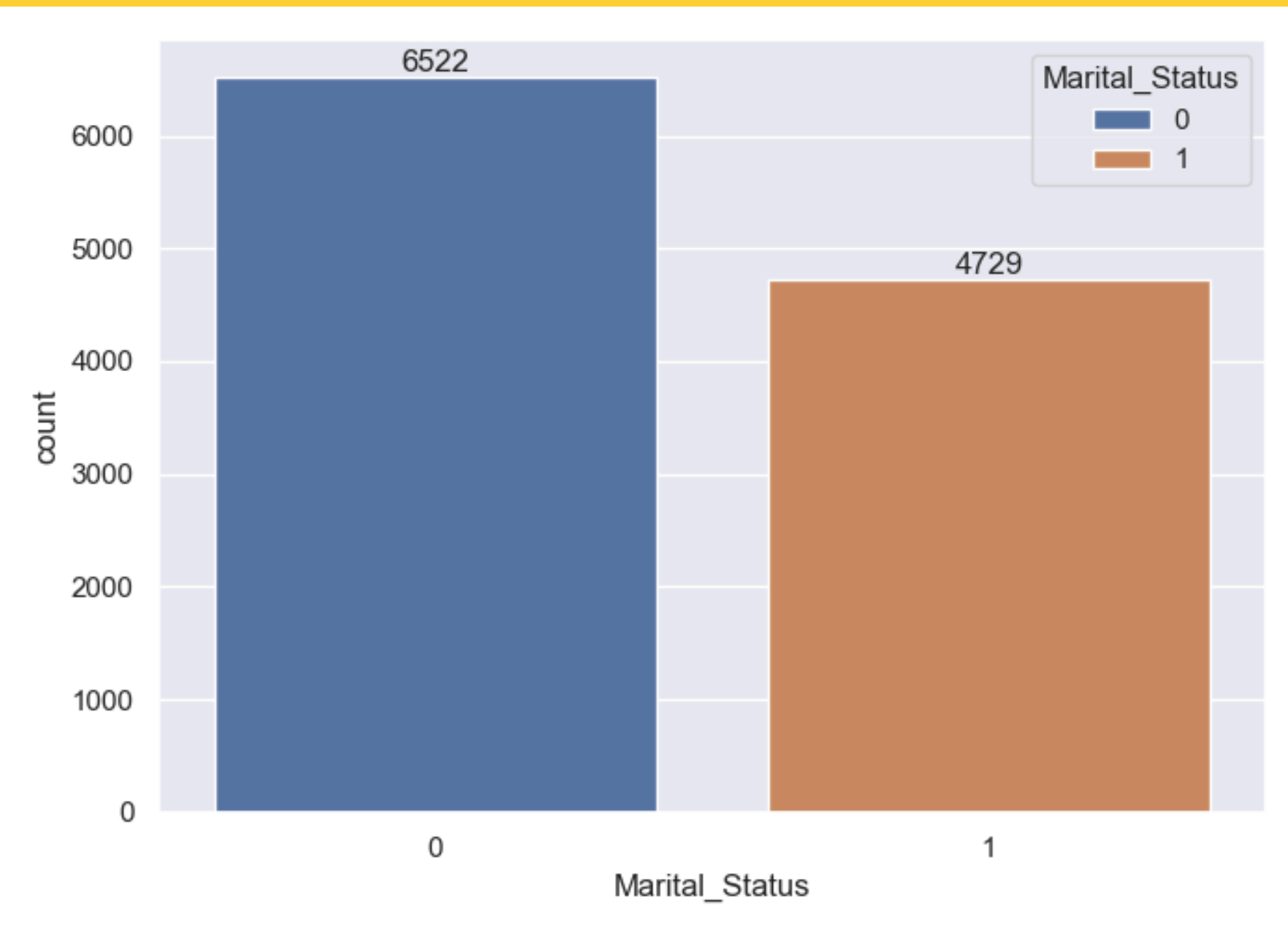
# Marital Status :

```
ax = sns.countplot(x='Marital_Status', data=df, hue = 'Marital_Status')

sns.set(rc={'figure.figsize': (7, 5)})

for bars in ax.containers:
    ax.bar_label(bars)
plt.tight_layout()
plt.savefig('Marital_Status.png')
```

```
marraige_sales = df.groupby(['Marital_Status' , 'Gender'] , as_index= False)['Amount'].sum().sort_values(by = 'Amount' , ascending=False)
sns.set(rc = {'figure.figsize': (6,4)})
sns.barplot(x = 'Marital_Status' , y = 'Amount' , data = marraige_sales , hue = 'Gender')
plt.tight_layout()
plt.savefig('Marital_Status2.png')
```



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

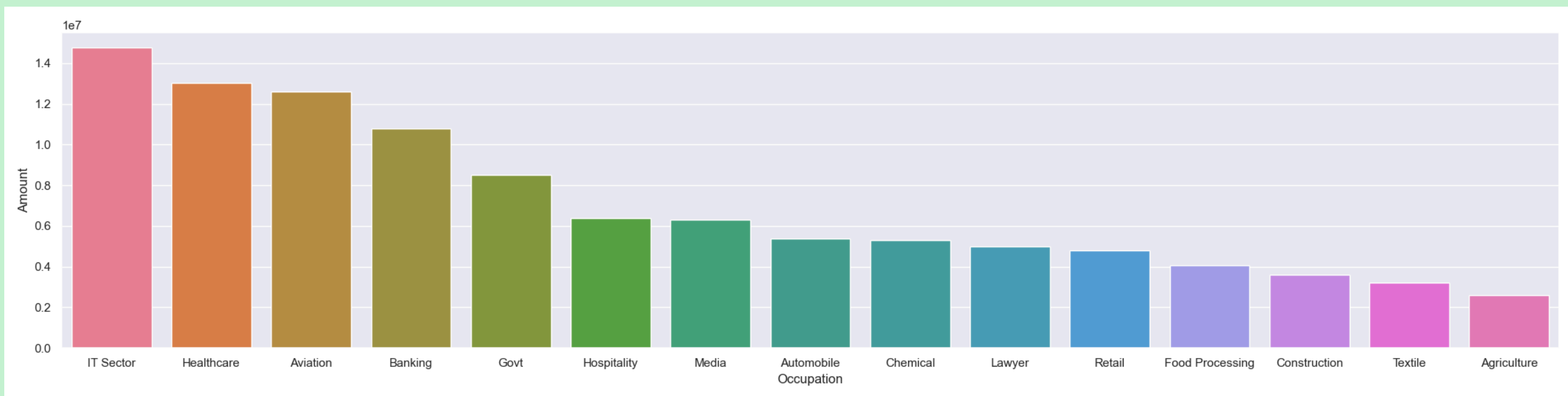
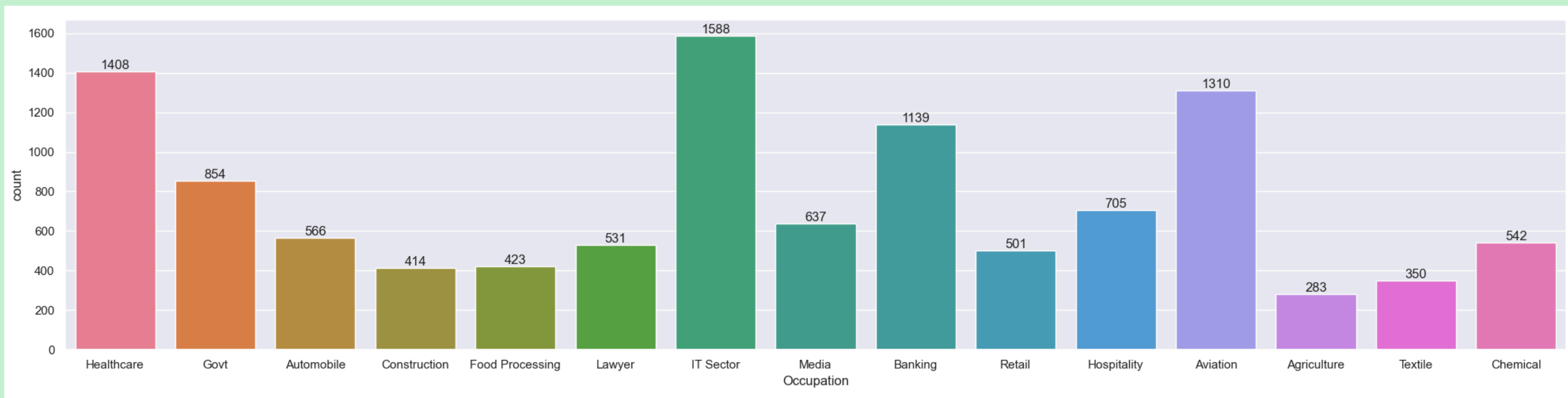
# Occupation :

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

for bars in ax.containers:
    ax.bar_label(bars)
plt.tight_layout()
plt.savefig('Occupation.png')
```

```
sales_occ = 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 = sales_occ , hue = 'Occupation')
plt.tight_layout()
plt.savefig('Occupation2.png')
```





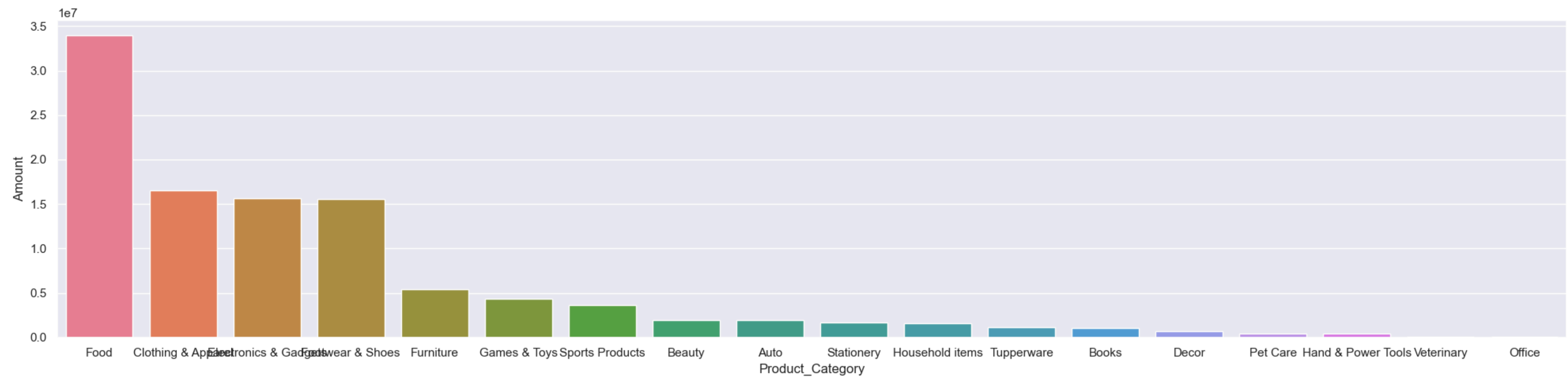
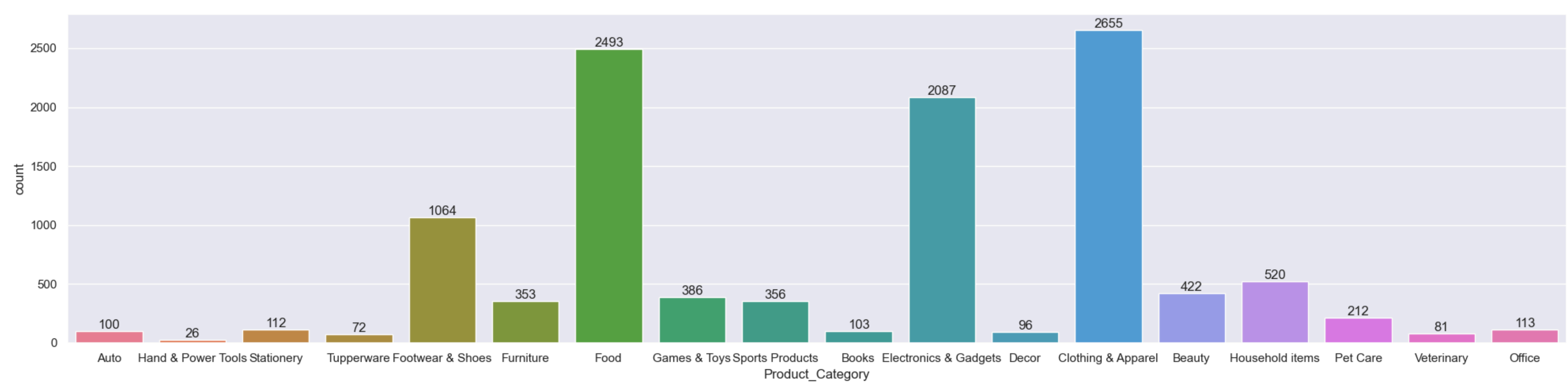
**From above graphs we can see that most of the buyers are working in IT, Healthcare and Aviation sector**

# Product Category:

```
#custom_colors = sns.color_palette("Set2", df['Product_Category'].nunique())
ax = sns.countplot(x = 'Product_Category' ,hue = 'Product_Category' , data = df, )

for bars in ax.containers:
    ax.bar_label(bars)
plt.tight_layout()
plt.savefig('Product_Category.png')
```

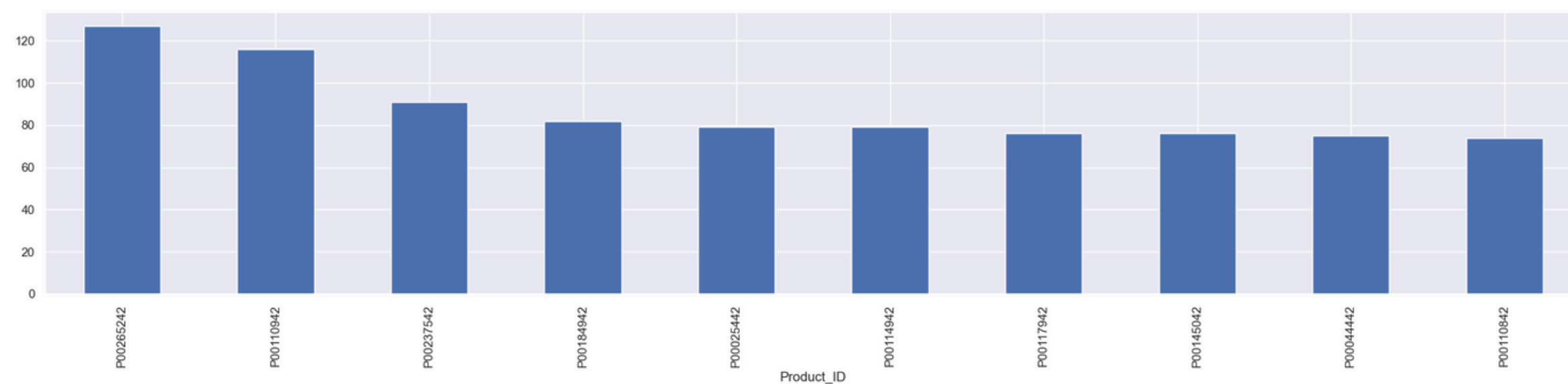
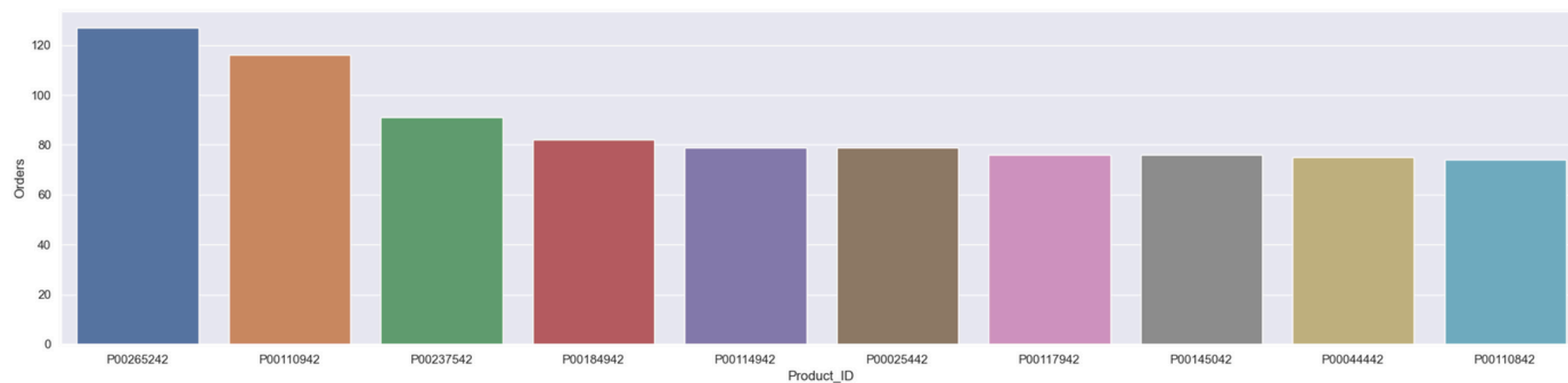
```
sales_pro = df.groupby(['Product_Category'] , as_index= False)['Amount'].sum().sort_values(by = 'Amount' , ascending= False)
sns.set(rc = {'figure.figsize':(20,5)})
sns.barplot(x = 'Product_Category' , y = 'Amount' , data = sales_pro , hue = 'Product_Category')
plt.tight_layout()
plt.savefig('Product_Category2.png')
```



**From above graphs we can see that most of the sold products are from Food, Clothing and Electronics category**

```
sales_state = df.groupby(['Product_ID'], as_index=False)['Orders'].sum().sort_values(by='Orders', ascending=False).head(10)
sns.barplot(data = sales_state, x = 'Product_ID',y= 'Orders' , hue = 'Product_ID')
plt.tight_layout()
```

```
# top 10 product most sold
df.groupby('Product_ID')['Orders'].sum().nlargest(10).sort_values(ascending=False).plot(kind='bar')
plt.tight_layout()
plt.savefig('Top_10_Product2.png')
```



top 10 product  
most sold

# Conclusion :

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



Thankyou...

