```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('Diwali Sales Data.csv',encoding = 'unicode_escape')

df.shape
```

```
(11251, 15)
```

```
df.head(15)
```


```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 11251,\n  \"fields\":
[\n    {\n      \"column\": \"User_ID\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 1716,\n        \"min\":
1000001,\n        \"max\": 1006040,\n        \"num_unique_values\":
3755,\n        \"samples\": [\n          1005905,\n          1003730,\
n          1005326\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Cust_name\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 1250,\n
\"samples\": [\n          \"Nida\",\n          \"Lacy\",\n
\"Caudle\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Product_ID\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 2351,\n
\"samples\": [\n          \"P00224442\",\n          \"P00205242\",\n
\"P00347442\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Gender\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 2,\n        \"samples\":
[\n          \"M\",\n          \"F\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Age Group\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 7,\n        \"samples\": [\n          \"26-
35\",\n          \"0-17\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Age\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 12,\n        \"min\": 12,\n
\"max\": 92,\n        \"num_unique_values\": 81,\n        \"samples\":
[\n          18,\n          28\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Marital_Status\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 0,\n        \"min\": 0,\n
\"max\": 1,\n        \"num_unique_values\": 2,\n        \"samples\":
[\n          1,\n          0\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"State\",\n      \"properties\": {\n        \"dtype\":
```

```
\"category\",\n        \"num_unique_values\": 16,\n
\"samples\": [\n          \"Maharashtra\",\n          \"Andhra\\
u00a0Pradesh\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Zone\",\n      \"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 5,\n        \"samples\": [\n
\"Southern\",\n          \"Eastern\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Occupation\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 15,\n        \"samples\": [\n
\"Retail\",\n          \"Aviation\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Product_Category\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 18,\n        \"samples\": [\n
\"Auto\",\n          \"Hand & Power Tools\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Orders\",\n      \"properties\":
{\n        \"dtype\": \"number\",\n        \"std\": 1,\n
\"min\": 1,\n        \"max\": 4,\n        \"num_unique_values\": 4,\n
\"samples\": [\n          3,\n          4\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Amount\",\n      \"properties\":
{\n        \"dtype\": \"number\",\n        \"std\":
5222.355869186444,\n        \"min\": 188.0,\n        \"max\":
23952.0,\n        \"num_unique_values\": 6584,\n        \"samples\":
[\n          19249.0,\n          13184.0\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Status\",\n      \"properties\":
{\n        \"dtype\": \"number\",\n        \"std\": null,\n
\"min\": null,\n        \"max\": null,\n        \"num_unique_values\":
0,\n        \"samples\": [],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"unnamed1\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": null,\n        \"min\": null,\n
\"max\": null,\n        \"num_unique_values\": 0,\n
\"samples\": [],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"df"}
```


Data Cleaning

```
df.info()
# in this info we will see we have 2 blank columns so we have to drop
them
```

```
<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 Blank Columns
df.drop(['Status', 'unnamed1'], axis = 1 , inplace  = True)
```

```
# To Check Null Values
pd.isnull(df).sum()
```

```
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
```

```
# To Drop Null Values
df.dropna(inplace = True)
```

```
# To correct values in Marital_Status column
df['Marital_Status'] =
df['Marital_Status'].replace({0:'Married',1:'Single'})
```

```
# To Change Data Type of 'Amount' to int
df['Amount'] = df['Amount'].astype('int')
```

```
df.columns
```

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

```
# To Change Column Name
df.rename(columns = {'Cust_name':'Customer_Name'},inplace = True)
```

```
# To get Basic Descibtion of numeric Columns
df[['Age','Orders','Amount']].describe()
```


```
{"summary":"{\n  \"name\": \"df[['Age','Orders','Amount']]\",\n
\"rows\": 8,\n  \"fields\": [\n    {\n      \"column\": \"Age\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
3960.7779927819724,\n        \"min\": 12.0,\n        \"max\":
11239.0,\n        \"num_unique_values\": 8,\n        \"samples\": [\n
35.41035679330901,\n          33.0,\n          11239.0\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Orders\",\n      \"properties\":
{\n        \"dtype\": \"number\",\n        \"std\":
3972.7985251346995,\n        \"min\": 1.0,\n        \"max\": 11239.0,\
n        \"num_unique_values\": 7,\n        \"samples\": [\n
11239.0,\n          2.4896343091022333,\n          3.0\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Amount\",\n      \"properties\":
{\n        \"dtype\": \"number\",\n        \"std\":
7024.070687950828,\n        \"min\": 188.0,\n        \"max\":
23952.0,\n        \"num_unique_values\": 8,\n        \"samples\": [\n
9453.610552540262,\n          8109.0,\n          11239.0\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe"}
```


Data Filtering

```
# Number of Customer as per Zone
df['Zone'].value_counts()
```

```
Zone
Central     4289
Southern    2693
Western     1952
Northern    1491
Eastern      814
Name: count, dtype: int64
```

```
# Top 10 Purchase by amount
df.sort_values(by = 'Amount' , ascending = False).head(10)
```


```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 10,\n  \"fields\": [\n
{\n      \"column\": \"User_ID\",\n      \"properties\": {\n
\"dtype\": \"number\",\n        \"std\": 1293,\n        \"min\":
1000588,\n        \"max\": 1003829,\n        \"num_unique_values\":
9,\n        \"samples\": [\n          1003650,\n          1000732,\n
1001132\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Customer_Name\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 9,\n        \"samples\":
[\n          \"Ginny\",\n          \"Kartik\",\n          \"Balk\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"Product_ID\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 9,\n        \"samples\": [\n
\"P00031142\",\n          \"P00110942\",\n          \"P00018042\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"Gender\",\n      \"properties\":
{\n        \"dtype\": \"category\",\n        \"num_unique_values\":
2,\n        \"samples\": [\n          \"M\",\n          \"F\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"Age Group\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 3,\n        \"samples\": [\n          \"26-
35\",\n          \"0-17\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Age\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 6,\n        \"min\": 16,\n
\"max\": 35,\n        \"num_unique_values\": 6,\n        \"samples\":
[\n          28,\n          35\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Marital_Status\",\n      \"properties\": {\n
\"dtype\": \"category\",\n        \"num_unique_values\": 2,\n
\"samples\": [\n          \"Single\",\n          \"Married\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"State\",\n      \"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 7,\n        \"samples\": [\n
\"Maharashtra\",\n          \"Andhra\\u00a0Pradesh\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Zone\",\n      \"properties\": {\n
\"dtype\": \"category\",\n        \"num_unique_values\": 4,\n
\"samples\": [\n          \"Southern\",\n          \"Northern\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"Occupation\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 8,\n        \"samples\": [\n
\"Govt\",\n          \"Lawyer\"\n        ],\n
```

\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"Product_Category\",\n
\"properties\": {\n   \"dtype\": \"category\",\n
\"num_unique_values\": 1,\n   \"samples\": [\n   \"Auto\"\
n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"Orders\",\n   \"properties\": {\n   \"dtype\": \"number\",\n
\"std\": 1,\n   \"min\": 1,\n   \"max\": 4,\n
\"num_unique_values\": 4,\n   \"samples\": [\n   3\n
],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   },\n   {\n   \"column\": \"Amount\",\n   \"properties\":
{\n   \"dtype\": \"number\",\n   \"std\": 62,\n
\"min\": 23770,\n   \"max\": 23952,\n
\"num_unique_values\": 9,\n   \"samples\": [\n   23799\n
],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   }\n]\n}","type":"dataframe"}

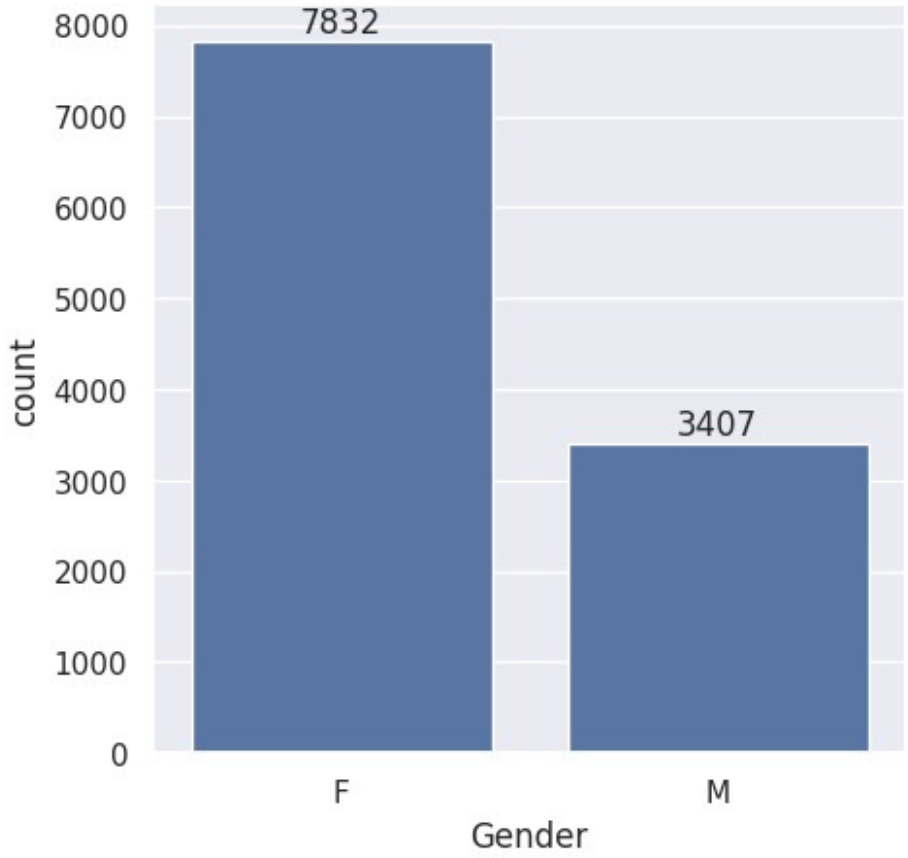
```
# list of purchase where number of order is above 3
df[df['Orders']>3]
```

{"summary":"{\n \"name\": \"df[df['Orders']>3]\",\n \"rows\": 2773,\
n \"fields\": [\n   {\n   \"column\": \"User_ID\",\n
\"properties\": {\n   \"dtype\": \"number\",\n   \"std\":
1745,\n   \"min\": 1000008,\n   \"max\": 1006037,\n
\"num_unique_values\": 1815,\n   \"samples\": [\n
1001611,\n   1003206,\n   1002136\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"Customer_Name\",\n
\"properties\": {\n   \"dtype\": \"category\",\n
\"num_unique_values\": 1087,\n   \"samples\": [\n
\"Tate\",\n   \"Fein\",\n   \"Apsingekar\"\n   ],\n
\"semantic_type\": \"\",\n   \"description\": \"\"\n   }\
n   },\n   {\n   \"column\": \"Product_ID\",\n
\"properties\": {\n   \"dtype\": \"category\",\n
\"num_unique_values\": 1359,\n   \"samples\": [\n
\"P00113442\",\n   \"P00251442\",\n   \"P00006142\"\n
],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   },\n   {\n   \"column\": \"Gender\",\n   \"properties\":
{\n   \"dtype\": \"category\",\n   \"num_unique_values\":
2,\n   \"samples\": [\n   \"M\",\n   \"F\"\n
],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   },\n   {\n   \"column\": \"Age Group\",\n
\"properties\": {\n   \"dtype\": \"category\",\n
\"num_unique_values\": 7,\n   \"samples\": [\n   \"18-
25\",\n   \"26-35\"\n   ],\n   \"semantic_type\":
\"\",\n   \"description\": \"\"\n   }\n   },\n   {\n
\"column\": \"Age\",\n   \"properties\": {\n   \"dtype\":
\"number\",\n   \"std\": 12,\n   \"min\": 12,\n
\"max\": 92,\n   \"num_unique_values\": 80,\n   \"samples\":
[\n   27,\n   25\n   ],\n   \"semantic_type\":

```
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Marital_Status\",\n      \"properties\": {\n
\"dtype\": \"category\",\n        \"num_unique_values\": 2,\n
\"samples\": [\n          \"Married\",\n          \"Single\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"State\",\n      \"properties\": {\n        \"dtype\": \"category\",\
n        \"num_unique_values\": 16,\n        \"samples\": [\n
\"Uttar Pradesh\",\n          \"Andhra\\u00a0Pradesh\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Zone\",\n      \"properties\": {\n
\"dtype\": \"category\",\n        \"num_unique_values\": 5,\n
\"samples\": [\n          \"Southern\",\n          \"Eastern\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"Occupation\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 15,\n        \"samples\": [\n
\"Agriculture\",\n          \"Textile\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Product_Category\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 18,\n        \"samples\": [\n
\"Auto\",\n          \"Stationery\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Orders\",\n      \"properties\":
{\n        \"dtype\": \"number\",\n        \"std\": 0,\n
\"min\": 4,\n        \"max\": 4,\n        \"num_unique_values\": 1,\n
\"samples\": [\n          4\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Amount\",\n      \"properties\": {\n        \"dtype\":
\"number\",\n        \"std\": 5215,\n        \"min\": 213,\n
\"max\": 23841,\n        \"num_unique_values\": 2358,\n
\"samples\": [\n          16209\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe"}
```

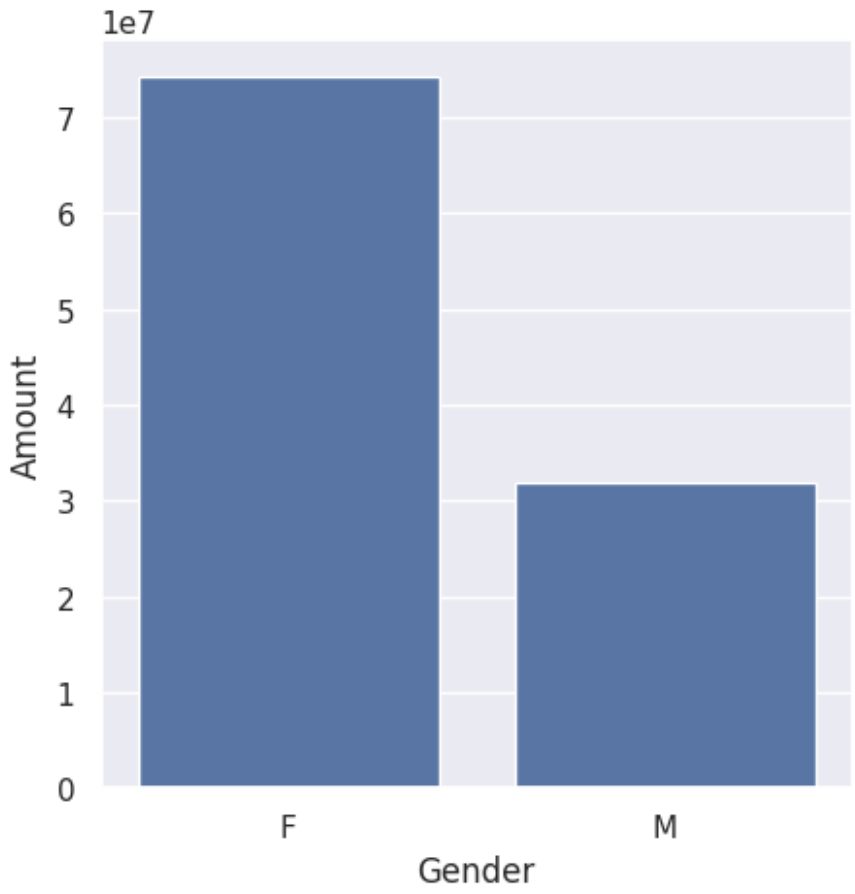

(EDA) Exploratory Data Analysis

```
plt.figure(figsize=(5, 5))
xa = sns.countplot(x='Gender', data=df)
for bars in xa.containers:
    xa.bar_label(bars)
plt.show()
```




```
# Sales amount by gender
plt.figure(figsize=(5, 5))
sales_gen = df.groupby(['Gender'] , as_index = False)
['Amount'].sum().sort_values( by = 'Amount' , ascending = False)

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

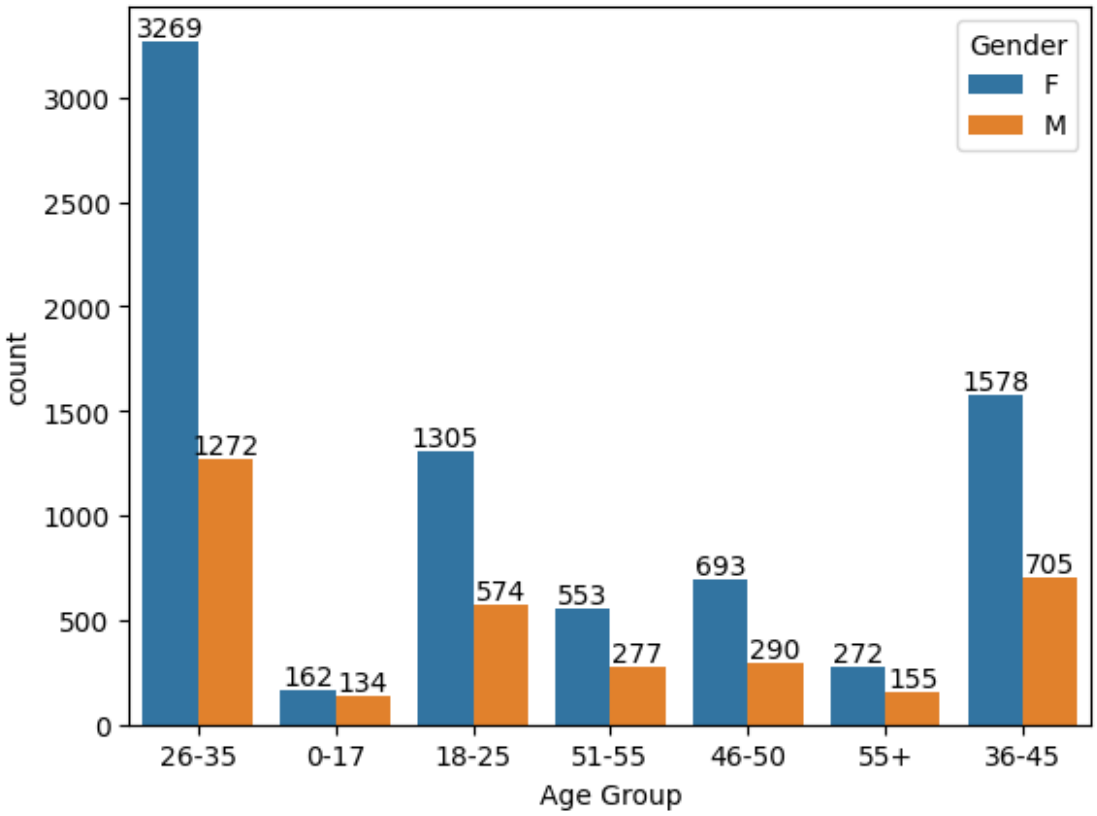
```
<Axes: xlabel='Gender', ylabel='Amount'>
```



Conclusion: From above Graphs we can see that most of the buyers are females and even Spendings of females are greater then men

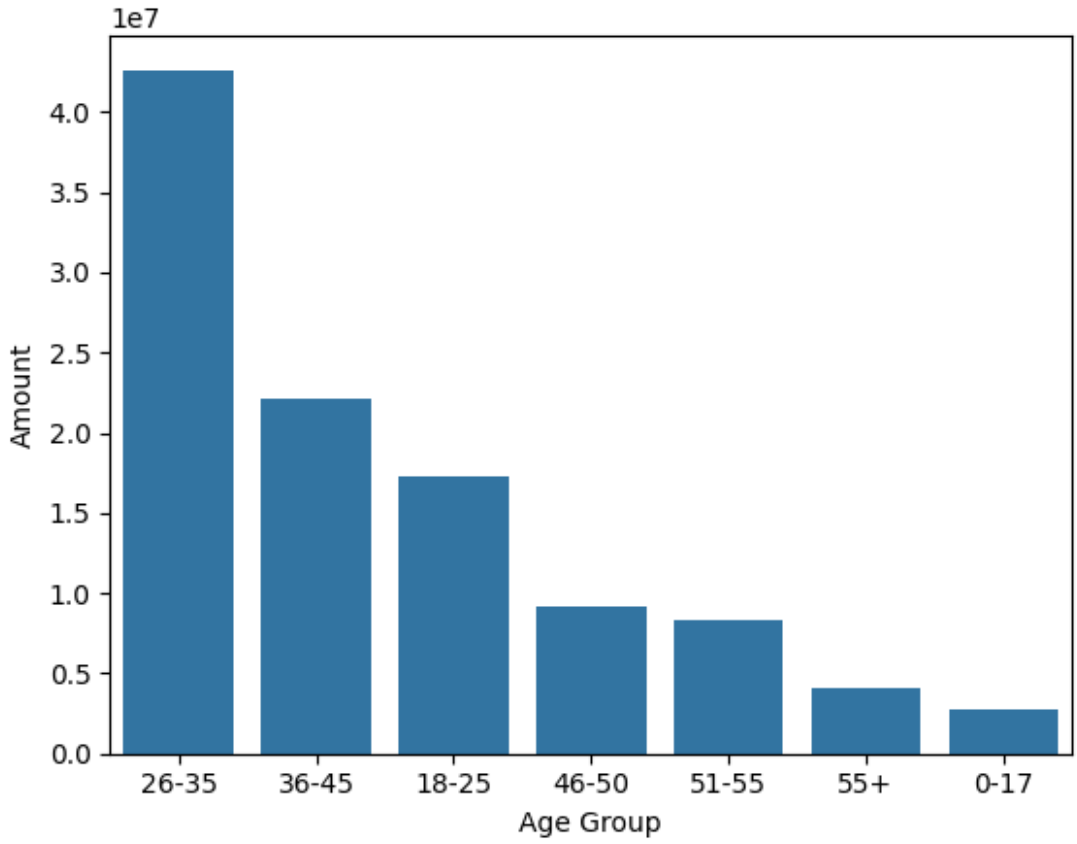
Age

```
# Customers by Age Group and Gender
ax = sns.countplot(data = df, x = "Age Group", hue = 'Gender')

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




```
# Total sales amount by Age Group
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)
```

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

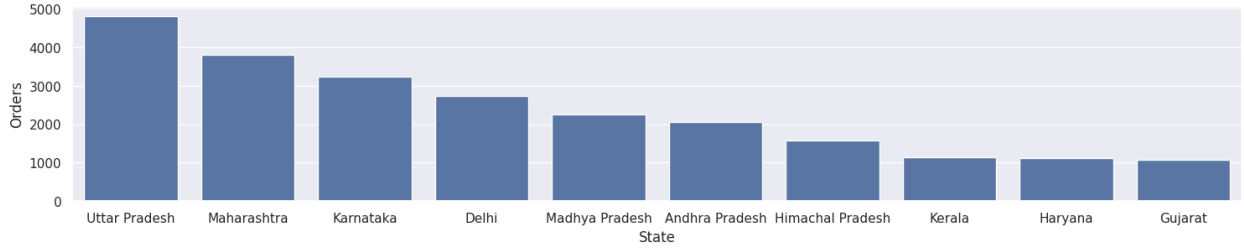


Conclusion: From above Graphs we can see that most of the buyers are from age group 26-35 and are females

State

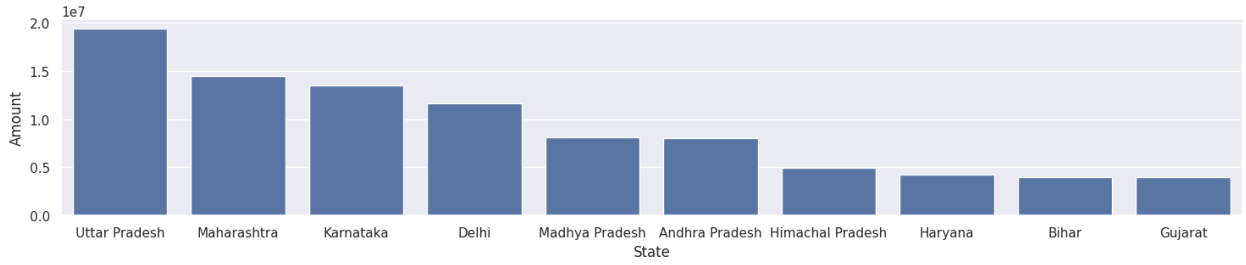
```
# Top 10 State as Number of Orders
sales_state = df.groupby(['State'] , as_index = False)
['Orders'].sum().sort_values( by = 'Orders' , ascending =
False).head(10)

sns.set(rc={'figure.figsize':(18,3)})

sns.barplot(x = 'State', y = 'Orders' , data = sales_state)
```

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



```
# Top 10 State as per Sum of sales
sales_state = df.groupby(['State'] , as_index = False)
['Amount'].sum().sort_values( by = 'Amount' , ascending =
False).head(10)

sns.set(rc={'figure.figsize':(18,3)})

sns.barplot(x = 'State', y = 'Amount' , data = sales_state)
```

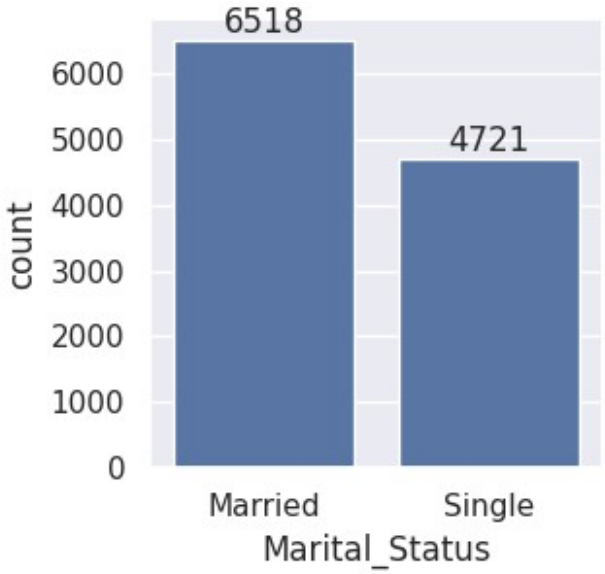
```
<Axes: xlabel='State', ylabel='Amount'>
```



Conclusion: From above graphs we can see that most of the orders and total sales amount is from uttar pradesh, maharashtra, karnataka respectively

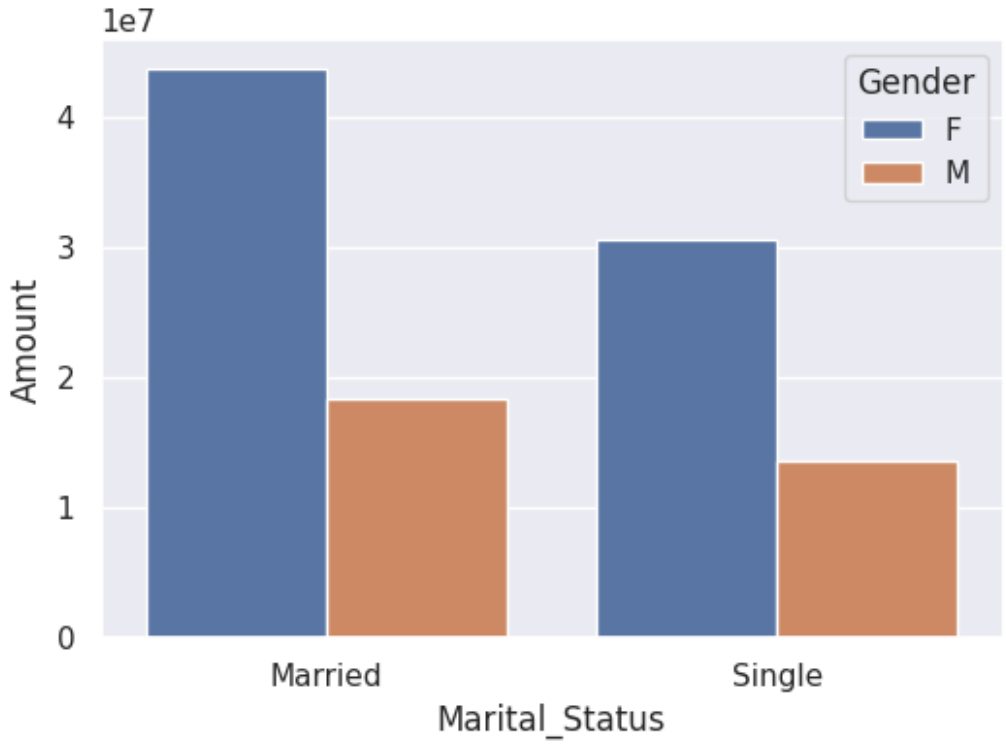
Marital Status

```
# Number of customers as per marital status
xa = sns.countplot(data = df , x = 'Marital_Status')

sns.set(rc={'figure.figsize':(3,2)})

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



```
# Total sales as per marital status and gender
sales_ms = 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 = sales_ms, hue =
'Gender')
```

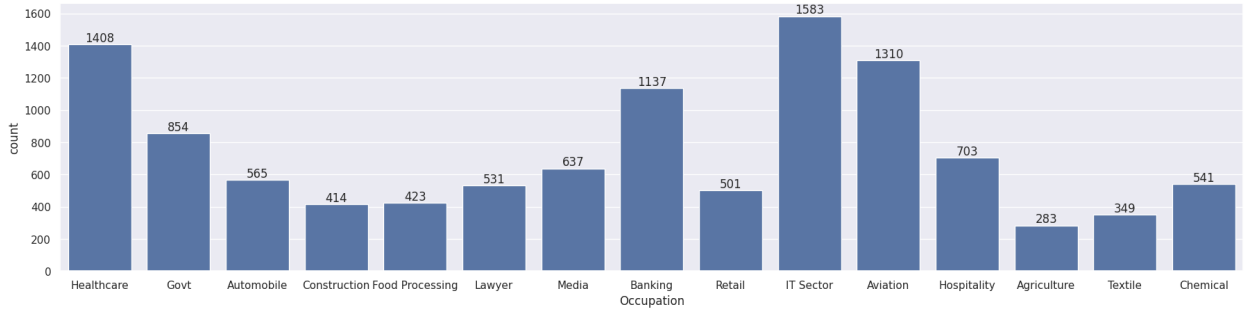
```
<Axes: xlabel='Marital_Status', ylabel='Amount'>
```



Most of the buyers are Married womens and they also contributes to highest amounts of sales

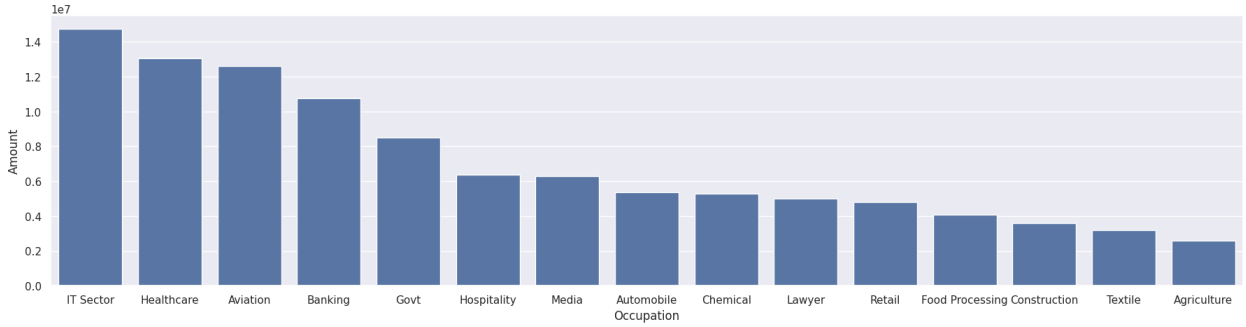
Occupation

```
# Number of customer as per Occupation
sns.set(rc={'figure.figsize':(22,5)})

ax = sns.countplot(x = 'Occupation', data = df)

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




```
# Total sales amoont as per customer occupation
sales_occu = df.groupby(['Occupation'], as_index = False)
['Amount'].sum().sort_values(by = 'Amount', ascending = False)

sns.set(rc={'figure.figsize':(22,5)})

sns.barplot(x = 'Occupation' , y = 'Amount' , data = sales_occu)
```

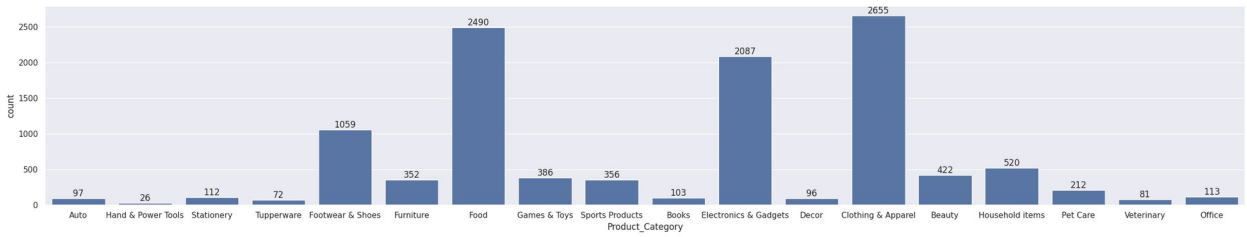
```
<Axes: xlabel='Occupation', ylabel='Amount'>
```



Conclusion: Here we see that most of they buyers are working in IT, Healthcare and Aviation sector

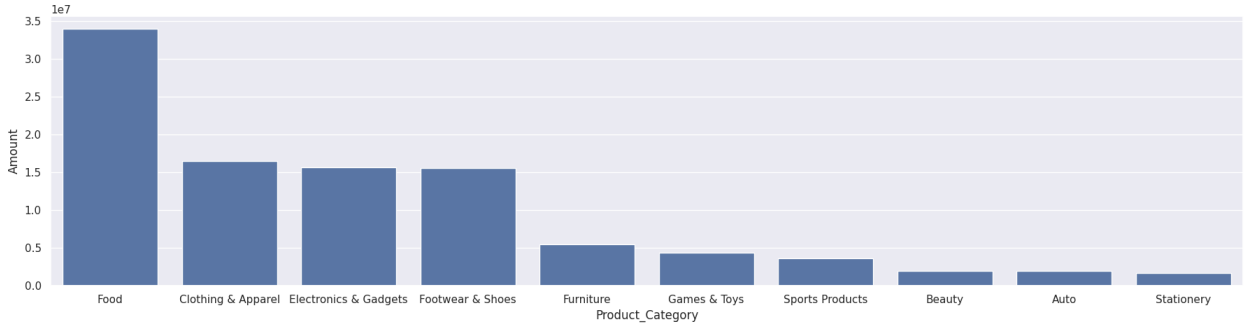
Product_Category

```
# Count of customer as per Product category
plt.figure(figsize=(30, 5))
ax = sns.countplot(x = 'Product_Category', data = df,)

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



```
# Top 10 Product category as per sales amount
sales_Pc = df.groupby(['Product_Category'], as_index = False)
['Amount'].sum().sort_values(by = 'Amount', ascending =
False).head(10)

sns.set(rc={'figure.figsize':(22,5)})

sns.barplot(x = 'Product_Category' , y = 'Amount' , data = sales_Pc)
```

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



Conclusion: From they above graphs we can see that most of the sold products are from food, clothing and electronics category

Short Conclusion: Married Women from age-group of 26-25 yrs from UP, Maharashtra and Karnataka working in IT, Healthcare and Aviation are more likely to buy products from Food, Clothing adn electronics category