Sireesha

PROJECT FOR DATA-CLEANING

Project Title:

Data Cleaning and Data Analysis on Diwali Sales Dataset

In [8]:

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

In [9]:

```
Diwali_dataset=pd.read_csv(r"C:\Users\kastu\OneDrive\Desktop\Diwali dataset.csv",encoding="unicode_escape")
Diwali_dataset
```

Out[9]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Cat
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	
2	1001990	Bindu	P00118542	F	26-35	35	1	Uttar Pradesh	Central	Automobile	
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	
...	...	...	...	...	...	...	...	...	...	...	
11246	1000695	Manning	P00296942	M	18-25	19	1	Maharashtra	Western	Chemical	
11247	1004089	Reichenbach	P00171342	M	26-35	33	0	Haryana	Northern	Healthcare	Vete
11248	1001209	Oshin	P00201342	F	36-45	40	0	Madhya Pradesh	Central	Textile	
11249	1004023	Noonan	P00059442	M	36-45	37	0	Karnataka	Southern	Agriculture	
11250	1002744	Brumley	P00281742	F	18-25	19	0	Maharashtra	Western	Healthcare	

11251 rows × 15 columns

In [10]:

```
Diwali_dataset.shape
```

Out[10]:

```
(11251, 15)
```

In [11]:

```
Diwali_dataset.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 [12]:

```
Diwali_dataset.count()
```

Out[12]:
```
User_ID             11251
Cust_name           11251
Product_ID          11251
Gender              11251
Age Group           11251
Age                 11251
Marital_Status      11251
State               11251
Zone                11251
Occupation          11251
Product_Category    11251
Orders              11251
Amount              11239
Status                  0
unnamed1                0
dtype: int64
```

In [13]:
```
#Dropping Columns which have no values
```

In [14]:
```
Diwali_dataset.drop(["Status", "unnamed1"], axis=1, inplace=True,errors="ignore")
```

In [15]:
```
#checking if the columns are dropped
```

In [16]:
```
Diwali_dataset.head()
```

Out[16]:

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

In [17]:
```
Diwali_dataset.sample(6)
```

Out[17]:

	User_ID	Cust_name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Cate
10897	1003411	Gastineau	P00210042	F	18-25	22	1	Madhya Pradesh	Central	Food Processing	Veter
8576	1005376	Chand	P00086342	F	26-35	32	0	Himachal Pradesh	Northern	Textile	Clothing & Ap
8451	1004979	Kennedy	P00174442	F	36-45	38	0	Telangana	Southern	IT Sector	Clothing & Ap
11069	1001579	Victor	P00003942	F	26-35	26	1	Madhya Pradesh	Central	IT Sector	Veter
5286	1004884	Divyeta	P00007542	F	36-45	38	0	Andhra Pradesh	Southern	Banking	Clothing & Ap
1089	1001266	Emily	P00084442	F	26-35	28	0	Karnataka	Southern	Hospitality	Footwear & S

In [18]:
```
#Renaming cust_name to customer name
Diwali_dataset.rename(columns={"Cust_name":"Customer name"},inplace=True)
```

In [19]:
```
#checking if the name has been changed
```

In [20]:
```
Diwali_dataset.iloc[:,:2].head()
```

Out[20]:

	User_ID	Customer name
0	1002903	Sanskriti
1	1000732	Kartik
2	1001990	Bindu
3	1001425	Sudevi
4	1000588	Joni

In [21]:
```
#finding for null values
```

In [22]:
```
Diwali_dataset.isnull()
```

Out[22]:

	User_ID	Customer name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Category	Orders
0	False	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...	...
11246	False	False	False	False	False	False	False	False	False	False	False	False
11247	False	False	False	False	False	False	False	False	False	False	False	False
11248	False	False	False	False	False	False	False	False	False	False	False	False
11249	False	False	False	False	False	False	False	False	False	False	False	False
11250	False	False	False	False	False	False	False	False	False	False	False	False

11251 rows × 13 columns

In [23]:
```
#finding the null values count for each column
```

In [24]:
```
Diwali_dataset.isnull().sum()
```

Out[24]:
```
User_ID              0
Customer 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 [25]:
```
#dropping the null values
```

In [26]:
```
Diwali_dataset.dropna(inplace=True)
```

In [27]:
```
#Checking for the null values if they are dropped
Diwali_dataset.isnull().sum()
```

Out[27]:
```
User_ID             0
Customer 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 [28]:
```
Diwali_dataset.columns
```

Out[28]:
```
Index(['User_ID', 'Customer name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

In [29]:
```
Diwali_dataset.dtypes
```

Out[29]:

```
User_ID               int64
Customer name        object
Product_ID           object
Gender               object
Age Group            object
Age                   int64
Marital_Status        int64
State                object
Zone                 object
Occupation           object
Product_Category     object
Orders                int64
Amount              float64
dtype: object
```

In [30]:

```
#changing the data type of amount from float to int
```

In [31]:

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

In [32]:

```
Diwali_dataset["Amount"].dtype
```

Out[32]:

```
dtype('int32')
```

In [33]:

```
Diwali_dataset.head()
```

Out[33]:

	User_ID	Customer name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Category
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	Auto
2	1001990	Bindu	P00118542	F	26-35	35	1	Uttar Pradesh	Central	Automobile	Auto
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	Auto
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	Auto

In [34]:

```
#Treatment of duplicates if any
```

In [35]:

```
Diwali_dataset.duplicated()
```

Out[35]:

```
0        False
1        False
2        False
3        False
4        False
         ...
11246    False
11247    False
11248    False
11249    False
11250    False
Length: 11239, dtype: bool
```

In [36]:

```
Diwali_dataset.drop_duplicates().head()
```

Out[36]:

	User_ID	Customer name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Category
0	1002903	Sanskriti	P00125942	F	26-35	28	0	Maharashtra	Western	Healthcare	Auto
1	1000732	Kartik	P00110942	F	26-35	35	1	Andhra Pradesh	Southern	Govt	Auto
2	1001990	Bindu	P00118542	F	26-35	35	1	Uttar Pradesh	Central	Automobile	Auto
3	1001425	Sudevi	P00237842	M	0-17	16	0	Karnataka	Southern	Construction	Auto
4	1000588	Joni	P00057942	M	26-35	28	1	Gujarat	Western	Food Processing	Auto

In [37]:

```
#checking for the shape if any duplicates are dropped
```

In [38]:

```
Diwali_dataset.shape
```

Out[38]:

```
(11239, 13)
```

In [39]:

```
#Filter any invalid rows (e.g., negative purchase amounts).
```

In [40]:

```
Diwali_dataset[Diwali_dataset["Amount"]<0]
```

Out[40]:

	User_ID	Customer name	Product_ID	Gender	Age Group	Age	Marital_Status	State	Zone	Occupation	Product_Category	Orders	Amou

In [41]:

```
#Checking for the statistical summary
Diwali_dataset.describe()
```

Out[41]:

	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

In [42]:

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

Out[42]:

	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

2. Data Visualization

Create at least 5 visualizations using tools like Matplotlib, Seaborn, or Plotly. Suggested charts:

In [45]:

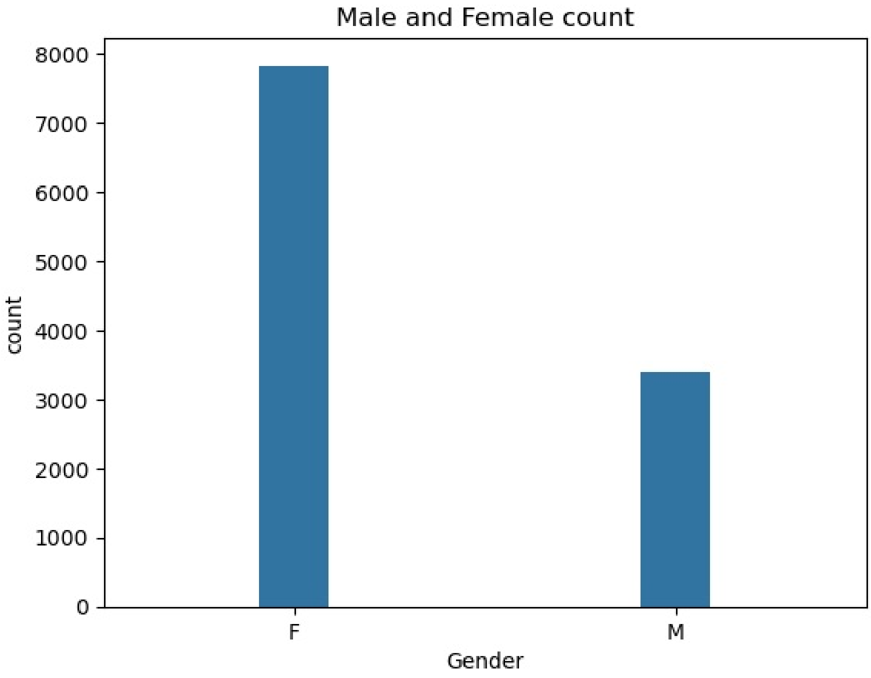
```
#Bar chart for which gender is buying more
#which age group is spending more
#Distribution of amount by product category
#Zone wise average amount
#Distribution of orders by Marital status
#Distribution of Amounts by state
#Age group wise orders
#Count of each occupation
#Amount by Occupation
#Majority Amount spent on buying
```

In [46]:

```
#Bar chart for which gender is buying more
```

In [47]:

```
sn.countplot(x="Gender",data=Diwali_dataset,width=0.18)
plt.title("Male and Female count")
plt.show()
```



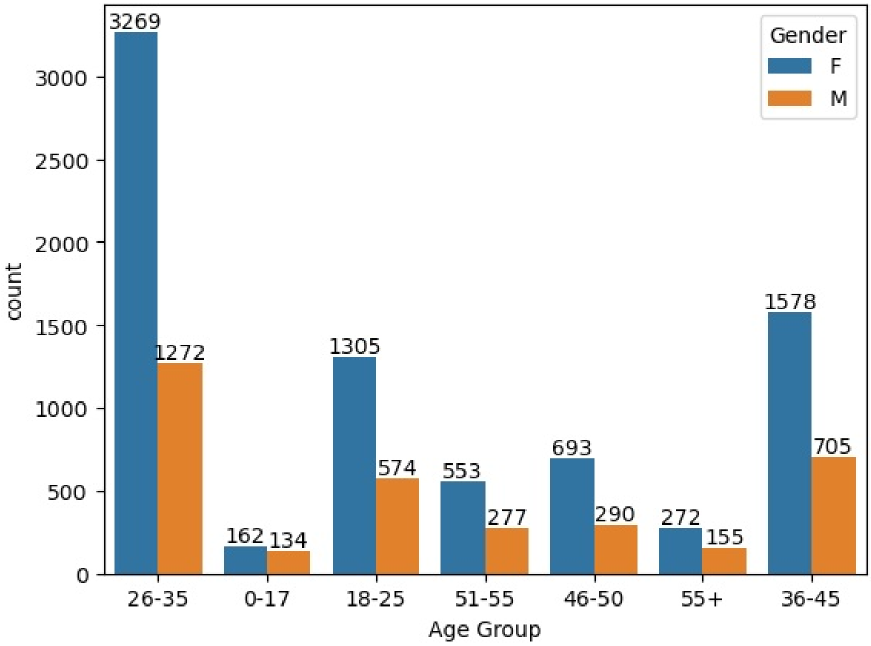
Female is buying more than men

```
In [49]: Diwali_dataset.columns
```

```
Out[49]: Index(['User_ID', 'Customer name', 'Product_ID', 'Gender', 'Age Group', 'Age',
                'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
                'Orders', 'Amount'],
               dtype='object')
```

```
In [50]: #which age group is spending more
```

```
In [51]: Agegroup=sn.countplot(x="Age Group",data=Diwali_dataset,hue="Gender")
         for bars in Agegroup.containers:
             Agegroup.bar_label(bars)
```



Age group of 25-35 is buying more items when compare to other age groups.

Female are buying more in all the age groups.

Age group of 0-17 is buying less products.

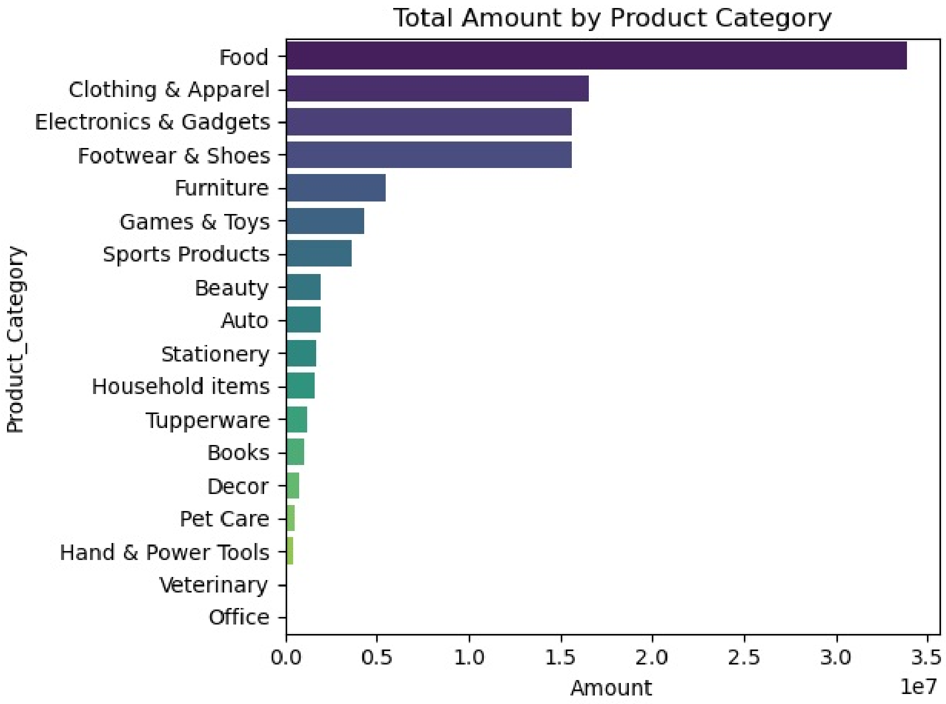
```
In [53]: #Distribution of amount by product category
         Amount_pcat=Diwali_dataset.groupby(['Product_Category'],as_index=False)["Amount"].sum().sort_values(by="Amount"
         Amount_pcat
```

Out[53]:

	Product_Category	Amount
6	Food	33933883
3	Clothing & Apparel	16495019
5	Electronics & Gadgets	15643846
7	Footwear & Shoes	15575209
8	Furniture	5440051
9	Games & Toys	4331694
14	Sports Products	3635933
1	Beauty	1959484
0	Auto	1958609
15	Stationery	1676051
11	Household items	1569337
16	Tupperware	1155642
2	Books	1061478
4	Decor	730360
13	Pet Care	482277
10	Hand & Power Tools	405618
17	Veterinary	112702
12	Office	81936

In [54]:

```
sn.barplot(x='Amount', y='Product_Category', data=Amount_pcat, hue='Product_Category',palette='viridis')
plt.xlabel('Amount')
plt.title('Total Amount by Product Category')
plt.tight_layout()
plt.show()
```



People are spending more amount on Food products

And they are spending less amount on Office products

In [56]:

```
#Zone wise average amount
```

In [57]:

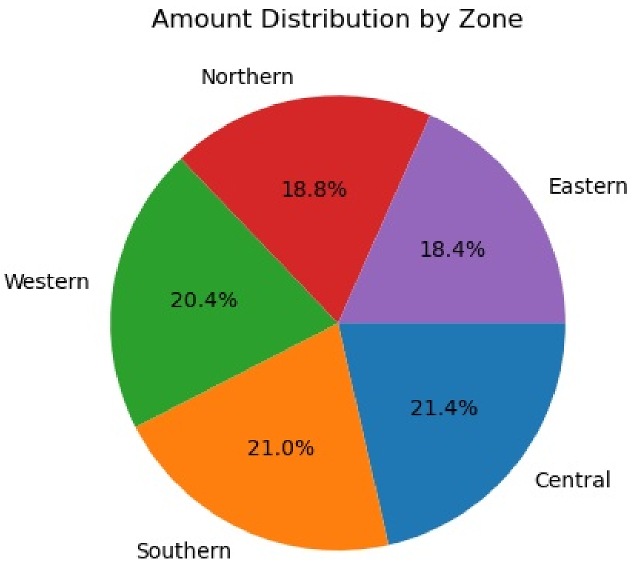
```
Amount_Zone=Diwali_dataset.groupby(['Zone'],as_index=False)["Amount"].mean().sort_values(by="Amount",ascending
Amount_Zone
```

Out[57]:

	Zone	Amount
3	Southern	9879.935759
0	Central	9699.433901
4	Western	9412.717725
1	Eastern	8659.966830
2	Northern	8463.281019

In [58]:

```
zones = ['Central', 'Southern', 'Western', 'Northern', 'Eastern']
amounts = [9879.936495,9699.434239,9412.717725,8659.966830,8463.281019]
plt.pie(amounts, labels=zones, autopct='%1.1f%%', counterclock=False)
plt.title('Amount Distribution by Zone')
plt.show()
```



Average Amount spent on purchasing products by Zone

Central Zone spent more amount on procuring

Eastern Zone procures less items

In [60]:

```
Maitalstatus_Orders=Diwali_dataset.groupby(["Marital_Status"],as_index=False)["Orders"].sum().sort_values(by="O
Maitalstatus_Orders
```

Out[60]:

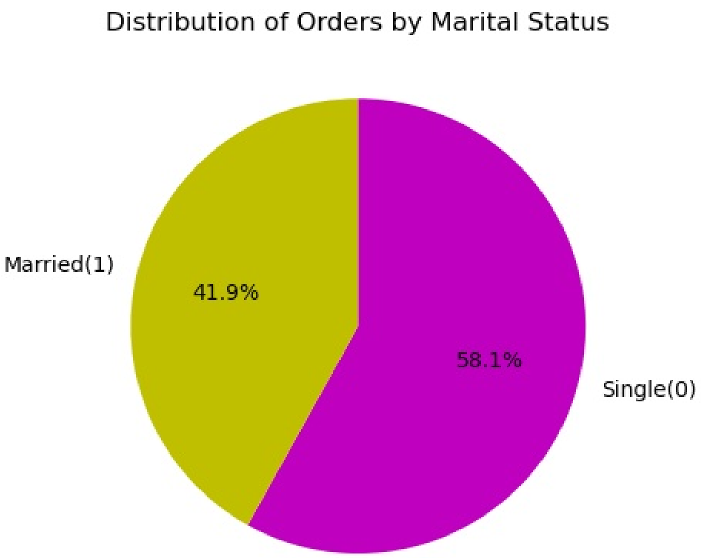
	Marital_Status	Orders
0	0	16249
1	1	11732

In [61]:

```
#Distribution of orders by Marital status
```

In [62]:

```
labels = ['Single(0)', 'Married(1)']
orders = [16249, 11732]
plt.pie(orders, labels=labels, autopct='%1.1f%%', startangle=90, counterclock=False, colors=['m', 'y'])
plt.title('Distribution of Orders by Marital Status')
plt.show()
```



When compared to married, singles are ordering more products.

In [64]:

```
State_Amount=Diwali_dataset.groupby(["State"],as_index=False)["Amount"].sum().sort_values(by="Amount",ascending
State_Amount
```

Out[64]:

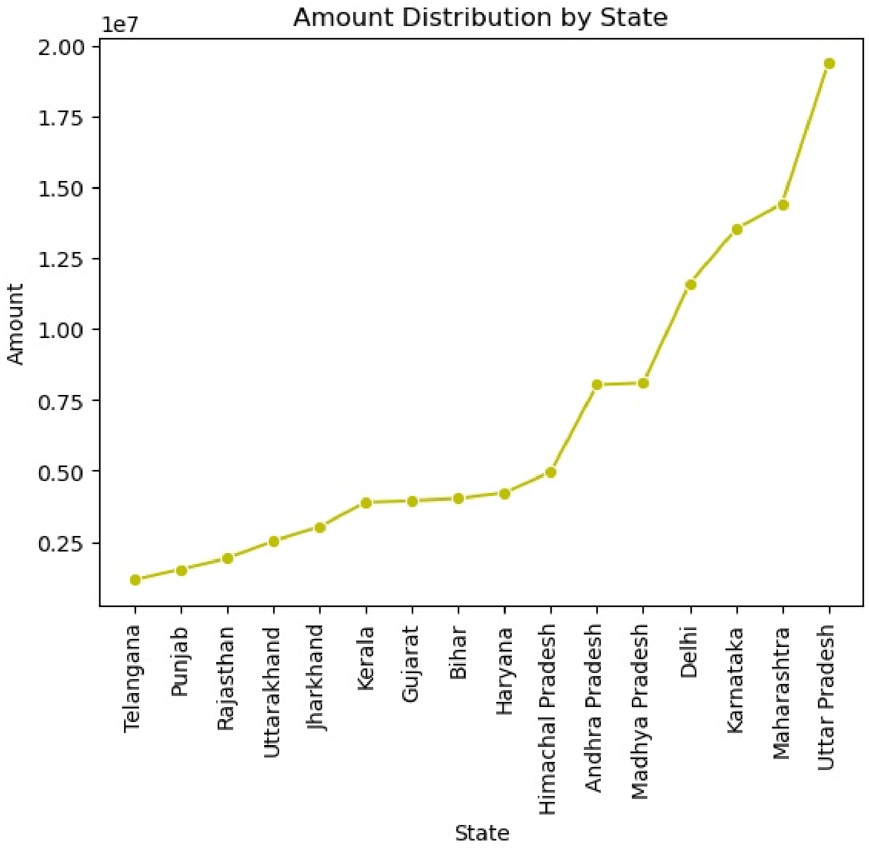
	State	Amount
13	Telangana	1151490
11	Punjab	1525800
12	Rajasthan	1909409
15	Uttarakhand	2520944
6	Jharkhand	3026456
8	Kerala	3894491
3	Gujarat	3946082
1	Bihar	4022757
4	Haryana	4220175
5	Himachal Pradesh	4963368
0	Andhra Pradesh	8037146
9	Madhya Pradesh	8101142
2	Delhi	11603818
7	Karnataka	13523540
10	Maharashtra	14427543
14	Uttar Pradesh	19374968

In [65]:

```
#Distribution of Amounts by state
```

In [66]:

```
sn.lineplot(x='State',y='Amount',data=State_Amount,marker='o', color='y')
plt.title('Amount Distribution by State')
plt.xlabel('State')
plt.ylabel('Amount')
plt.xticks(rotation=90)
plt.show()
```



Telangana state spent less amount on buying

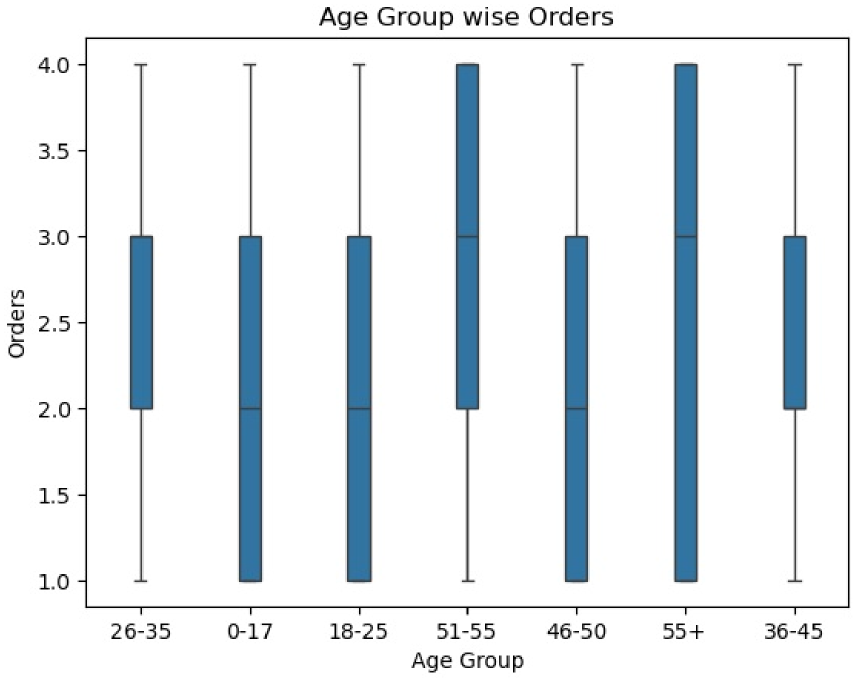
Uttar pradesh spent more amount on buying

```
In [68]: Diwali_dataset.columns
```

```
Out[68]: Index(['User_ID', 'Customer name', 'Product_ID', 'Gender', 'Age Group', 'Age',
                'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
                'Orders', 'Amount'],
               dtype='object')
```

```
In [69]: #Age group wise orders
```

```
In [70]: sn.boxplot(x="Age Group",y="Orders",data=Diwali_dataset,width=0.2)
         plt.title('Age Group wise Orders')
         plt.show()
```

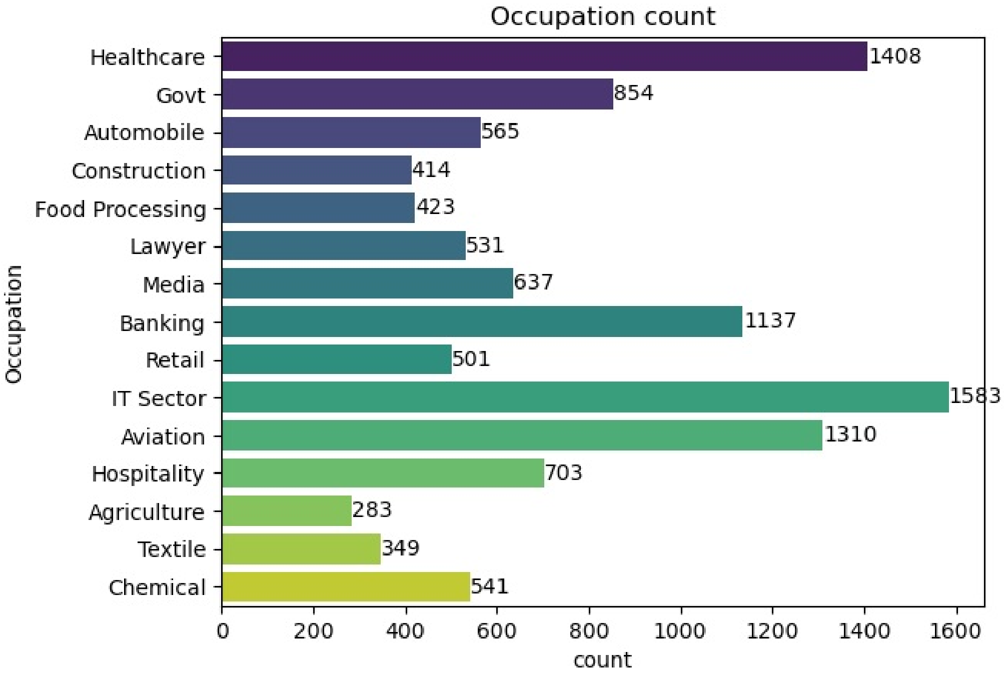


Min count of orders are 1 and max count of orders are 4 in this chart

Age group of 55+ orders more

In [72]:
```
#Count of each occupation
```

In [73]:
```
occ_Amount=sn.countplot(y="Occupation",data=Diwali_dataset,width=0.8,hue='Occupation',palette='viridis')
plt.title('Occupation count')
for bars in occ_Amount.containers:
    occ_Amount.bar_label(bars)
plt.show()
```



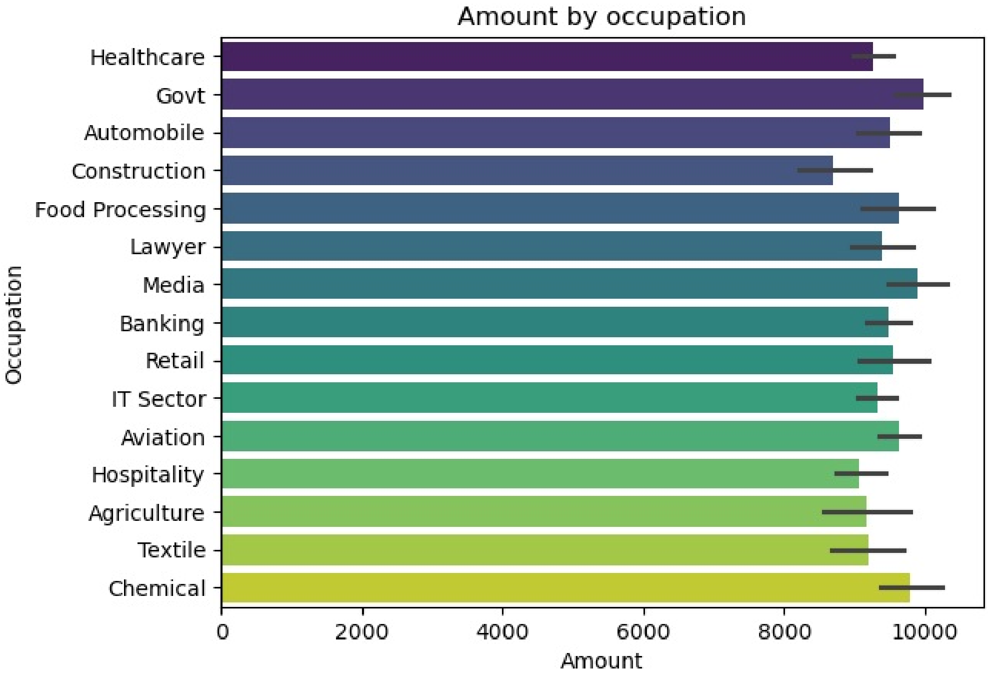
Count of IT Sector is more

IT Sector, Healthcare, Aviation order more items

Agriculture orders less items

In [75]:
```
#Amount by Occupation
```

In [76]:
```
sn.barplot(y='Occupation',x='Amount',data=Diwali_dataset,width=0.8,hue='Occupation',palette='viridis')
plt.title('Amount by occupation')
plt.show()
```



Govt spends most amount on purchasing

Automobile,construction,IT sector and chemical are also spending more comparatively a bit less than govt.

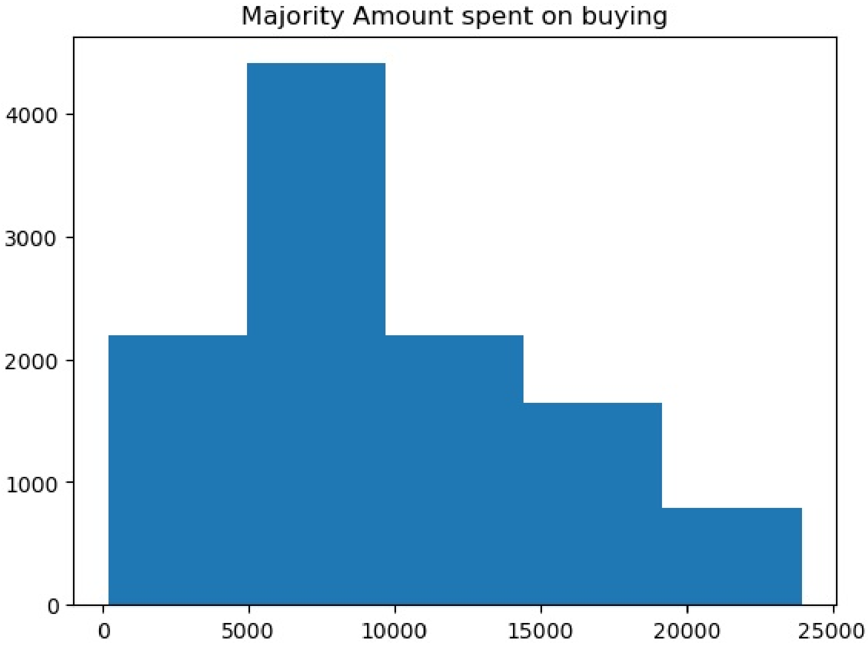
All the occupations are spending more than 8000

In [78]:
```
Diwali_dataset.columns
```

Out[78]:
```
Index(['User_ID', 'Customer name', 'Product_ID', 'Gender', 'Age Group', 'Age',
       'Marital_Status', 'State', 'Zone', 'Occupation', 'Product_Category',
       'Orders', 'Amount'],
      dtype='object')
```

In [79]:
```
#Majority Amount spent on buying
```

In [80]:
```
plt.hist(Diwali_dataset['Amount'],bins=5)
plt.title('Majority Amount spent on buying')
plt.show()
```



Majority Amount spent on purchasing is between 5000 and 10000

Conclusion:

1.Most active customers are from age group of 26-35 and Female are spending more on purchasing.

2.People are spending more on Food products

3.Singles are purchasing more

4.Uttarpradesh state has more buying rate

5.Age group of 55+ have more no.of orders

6.Govt spends most on purchasing

7.IT Sector Employees order more items

8.Majority Amount spend on purchasing is between 5000 to 10000

9.The highest Average percentage of customers come from central zone

In [ ]:

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js