Retail_Corporation_Sale_Data_Analysis_by_Diptyajit_Das

March 22, 2024

0.0.1 About the Retail Corporation

The retail corporation is an American multinational entity that operates a network of supercenters, discount departmental stores, and grocery outlets across the United States. With a vast customer base exceeding 100 million globally, the corporation has established itself as a prominent player in the retail industry.

0.0.2 Business Problem

The management team aims to delve into customer purchase behavior, particularly focusing on the purchase amount, relative to various demographic factors, including gender. The primary objective is to gain insights into potential differences in spending patterns between male and female customers during Black Friday. Given an equal distribution of 50 million male and 50 million female customers, the team seeks to answer the question: Do women exhibit higher spending levels on Black Friday compared to men?

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
```

[2]:
```
df=pd.read_csv('Black_Friday.txt')
```

0.1 1.Structure and basic characteristics

[3]:
```
df.head()
```

[3]:

	User_ID	Product_ID	Gender	Age	Occupation	City_Category	Stay_In_Current_City_Years	Marital_Status	Product_Category	Purchase
0	1000001	P00069042	F	0-17	10	A	2	0	3	8370
1	1000001	P00248942	F	0-17	10	A	2	0	1	15200
2	1000001	P00087842	F	0-17	10	A	2	0	12	1422
3	1000001	P00085442	F	0-17	10	A				
4	1000002	P00285442	M	55+	16	C				

```
3                            2                 0                 12       1057
4                           4+                 0                  8       7969
```

[4]:
```
df.shape
```

[4]:
```
(550068, 10)
```

0.1.1 a) 550068 rows and 10 columns

0.1.2 Converting appropriate columns to string type.

[5]:
```
cols=['Gender','Age','Product_Category','Stay_In_Current_City_Years','Marital_Status','City_Ca
df[cols]=df[cols].astype('object')
```

[6]:
```
df.dtypes
```

[6]:
```
User_ID                      int64
Product_ID                  object
Gender                      object
Age                         object
Occupation                  object
City_Category               object
Stay_In_Current_City_Years  object
Marital_Status              object
Product_Category            object
Purchase                     int64
dtype: object
```

0.1.3 b)

0.1.4 String Data Type:

- Product_ID
- Gender
- Age
- City_Category
- Occupation
- Marital_Status
- Product_Category
- Stay_In_Current_City_Years

0.1.5 Integer Data Type:

- User_ID
- Purchase

[7]:
```
df.isna().sum()
```

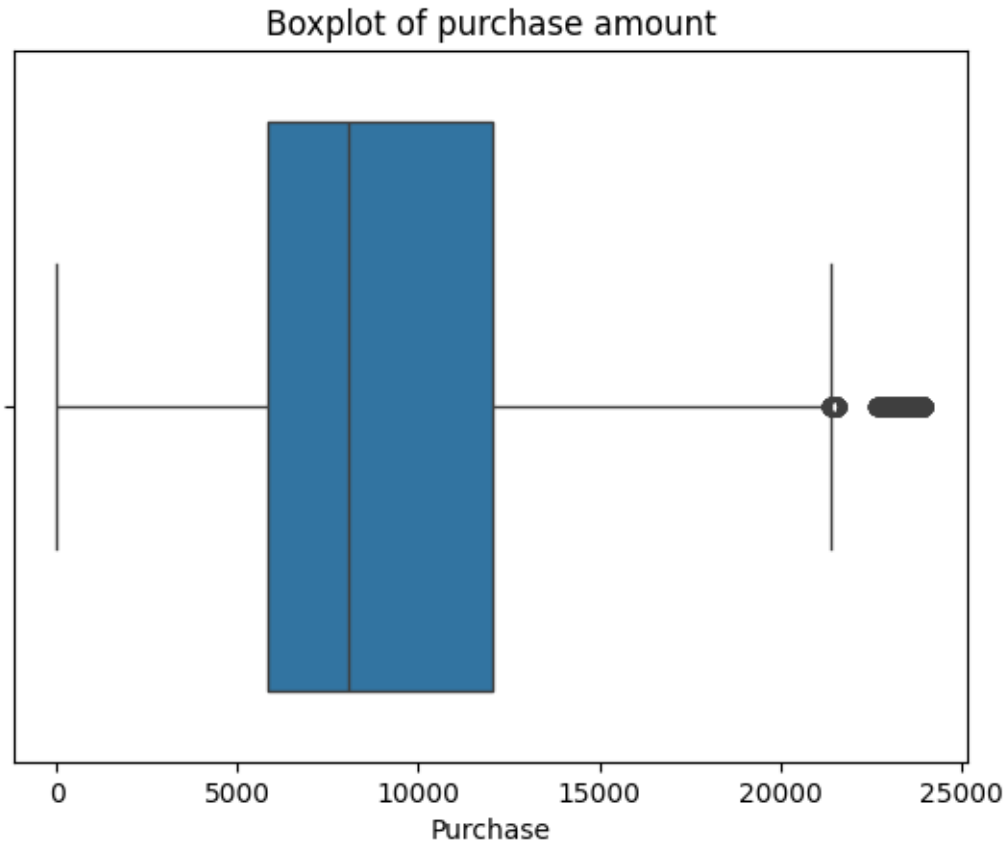
[7]:
```
User_ID                       0
Product_ID                    0
Gender                        0
Age                           0
Occupation                    0
City_Category                 0
Stay_In_Current_City_Years    0
Marital_Status                0
Product_Category              0
Purchase                      0
dtype: int64
```

0.1.6 c) No missing data.

1 Unit of purchase is dollars throughout.

1.1 2.Detecting and removing outliers

[8]:
```
sns.boxplot(data=df,x='Purchase')
plt.title('Boxplot of purchase amount')
plt.show()
```

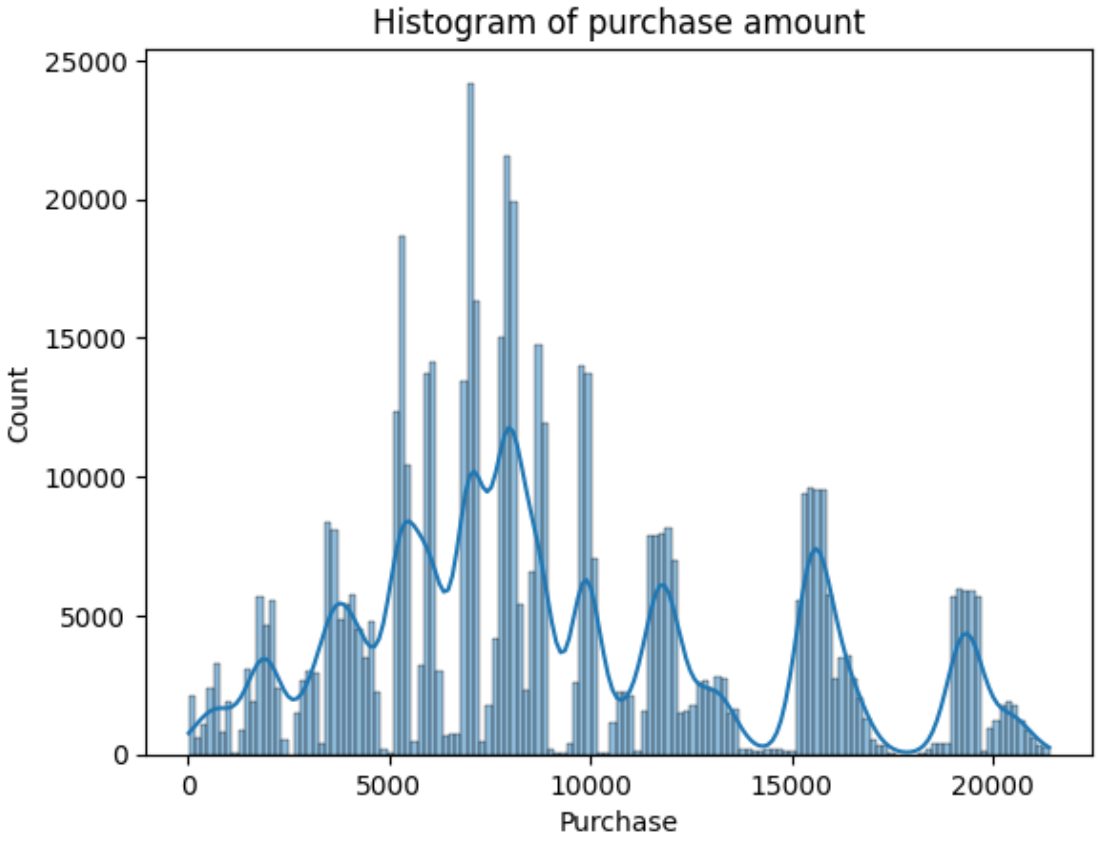


[9]:
```
columns = ['Purchase']
def filter_data_by_iqr(df, column):
    iqr = np.quantile(df[column], 0.75) - np.quantile(df[column], 0.25)
    q1 = np.quantile(df[column], 0.25)
    q3 = np.quantile(df[column], 0.75)
    return df[df[column].between(q1 - 1.5 * iqr, q3 + 1.5 * iqr)]
for column in columns:
    df = filter_data_by_iqr(df, column)
df.shape
```

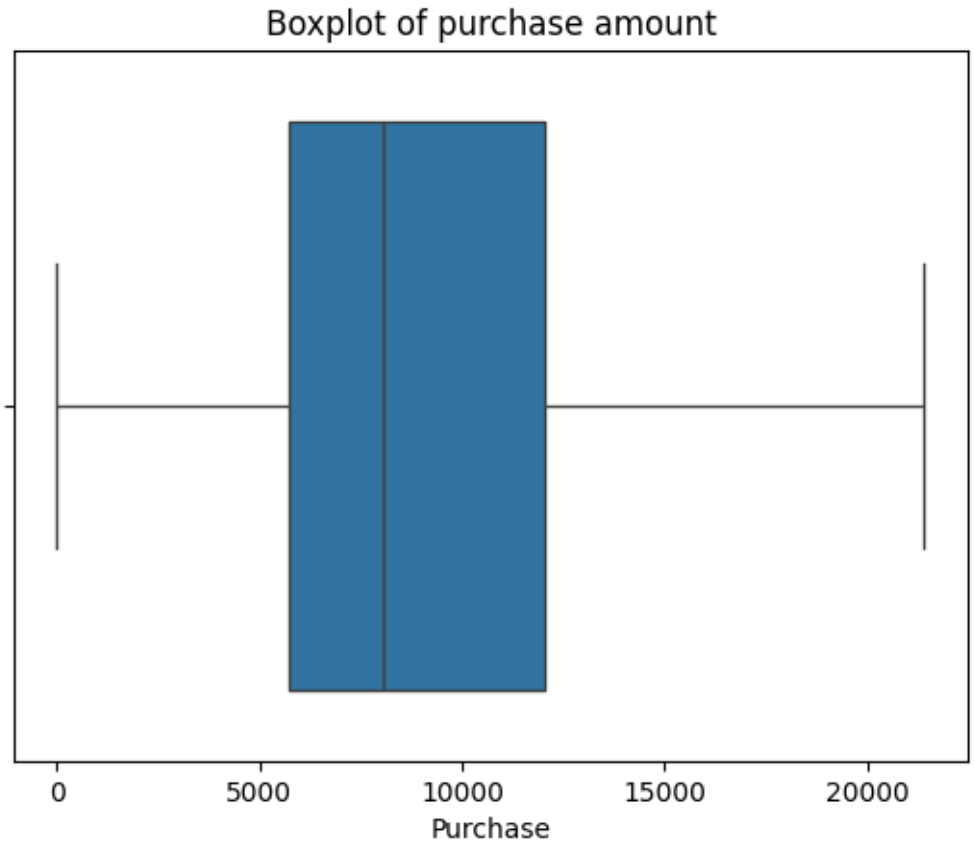
[9]:
```
(547391, 10)
```

1.2 2677 outlier records removed

[10]:
```
sns.histplot(data=df,x='Purchase',kde=True)
plt.title('Histogram of purchase amount')
plt.show()
```



[11]:
```
sns.boxplot(data=df,x='Purchase')
plt.title('Boxplot of purchase amount')
plt.show()
```



1.3 3. Data Exploration

[12]:
```
df['Purchase'].describe()
```

[12]:
```
count    547391.000000
mean       9195.627195
std        4938.872953
min          12.000000
25%        5721.000000
50%        8038.000000
75%       12019.000000
max       21399.000000
Name: Purchase, dtype: float64
```

1.3.1 Insight

- **Mean purchase:** 9196
- **Purchase range:** 12-21399

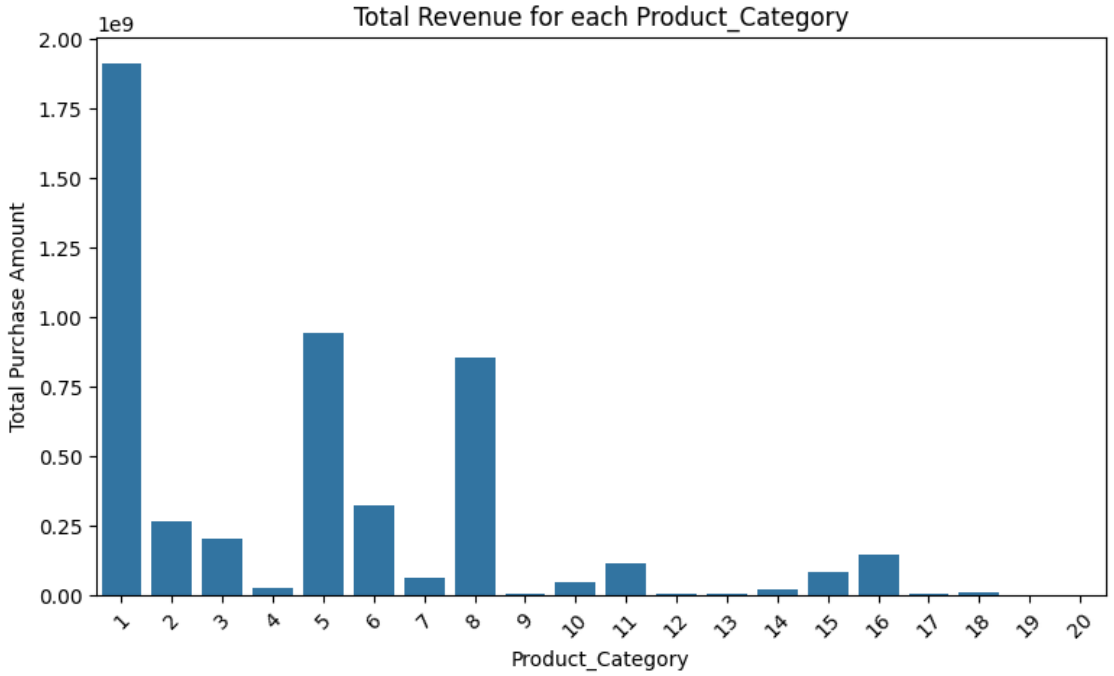
[13]:
```
cols=['Age','Product_Category','Stay_In_Current_City_Years','City_Category','Occupation','Prod
print(list(df[cols].nunique()))
```

```
[7, 20, 5, 3, 21, 3631, 5891]
```

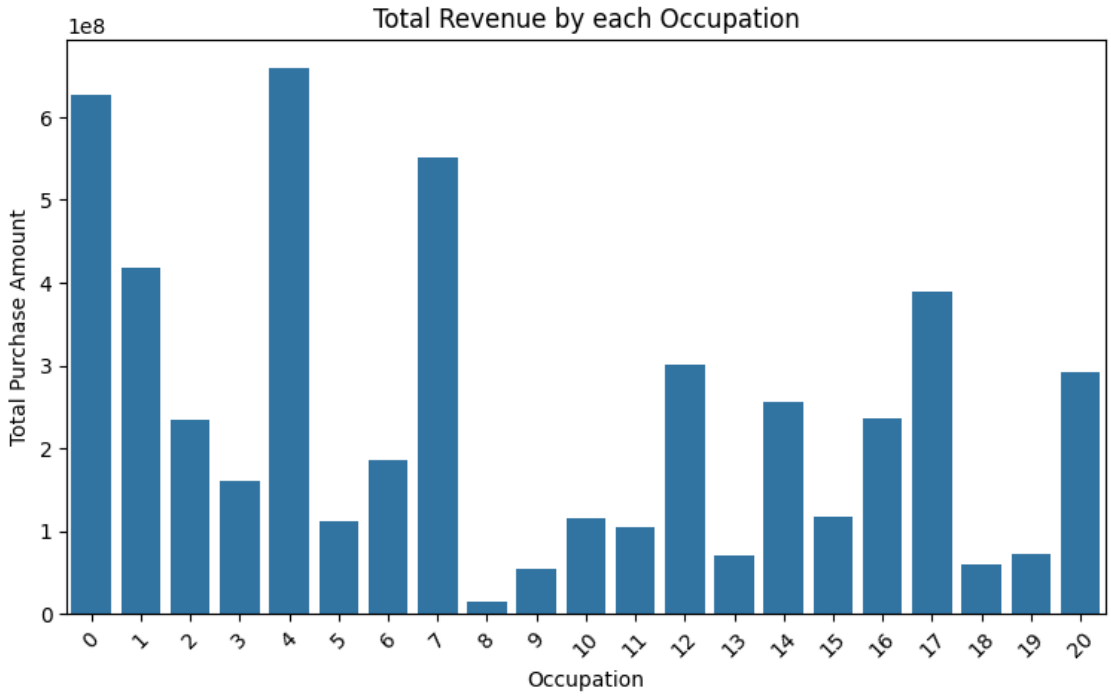
1.3.2 Insight

The dataset contains 7 unique age groups, 20 distinct product categories, data for 5 different durations of stay in the current city, 3 city categories, records for 21 unique occupation types, 3631 unique product IDs, and data from 5891 unique user IDs.

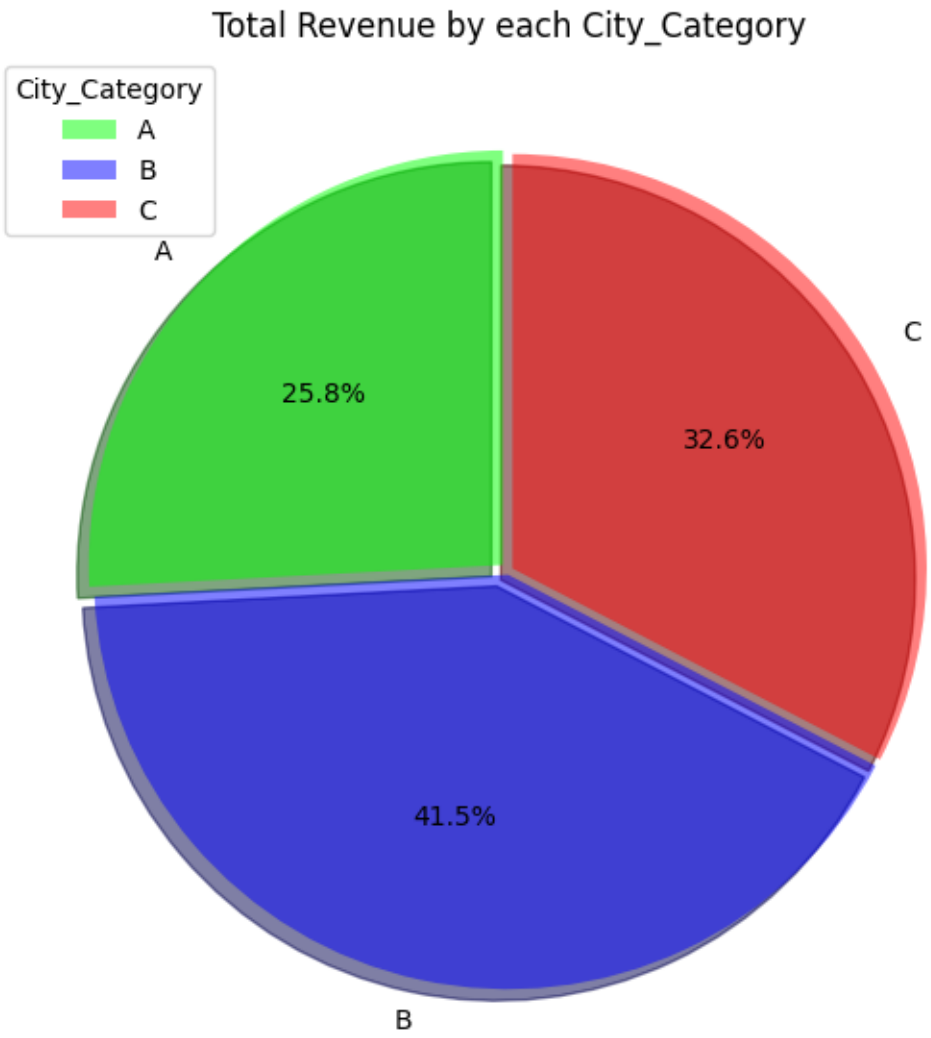
[14]:
```
product_cat_data = df.groupby('Product_Category', as_index=False)['Purchase'].
↪sum()
plt.figure(figsize=(9, 5))
sns.barplot(data=product_cat_data, x='Product_Category', y='Purchase')
plt.title('Total Revenue for each Product_Category')
plt.xlabel('Product_Category')
plt.ylabel('Total Purchase Amount')
plt.xticks(rotation=45)
plt.show()
```



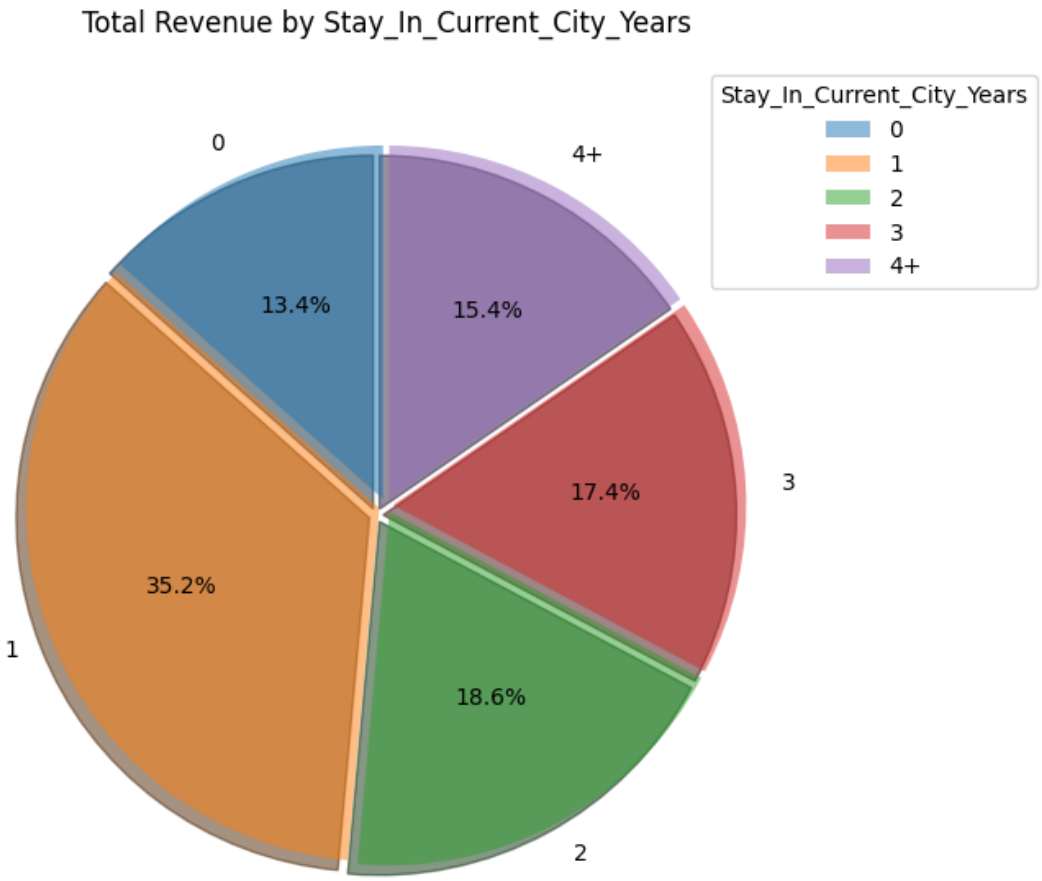
[15]:
```
occupation_data = df.groupby('Occupation', as_index=False)['Purchase'].sum()
plt.figure(figsize=(9,5))
sns.barplot(data=occupation_data, x='Occupation', y='Purchase')
plt.title('Total Revenue by each Occupation')
plt.xlabel('Occupation')
plt.ylabel('Total Purchase Amount')
plt.xticks(rotation=45)
plt.show()
```



[16]:
```
city_category_data = df.groupby('City_Category')['Purchase'].sum()
explode = [0.02, 0.01, 0.01]
colors = [(0, 1, 0, 0.5), (0, 0, 1, 0.5), (1, 0, 0, 0.5)]  # (R, G, B, Alpha)
plt.figure(figsize=(7, 7))
plt.pie(city_category_data, labels=city_category_data.index, autopct='%1.1f%%',␣
↪startangle=90, colors=colors, explode=explode,shadow=True)
plt.title('Total Revenue by each City_Category')
plt.legend(title='City_Category', loc='upper left')
plt.show()
```



[17]:
```
Stay_In_Current_City_Years_data = df.
 ↪groupby('Stay_In_Current_City_Years')['Purchase'].sum()
explode = [0.02, 0.02, 0.02, 0.02, 0.02]
plt.figure(figsize=(7, 7))
plt.pie(Stay_In_Current_City_Years_data, labels=Stay_In_Current_City_Years_data.
 ↪index, autopct='%1.1f%%', startangle=90, explode=explode, shadow=True,␣
 ↪wedgeprops={'alpha':0.5})
plt.title('Total Revenue by Stay_In_Current_City_Years', y=1.02)
plt.legend(title='Stay_In_Current_City_Years', loc='upper right',␣
 ↪bbox_to_anchor=(1.25, 1))
plt.show()
```



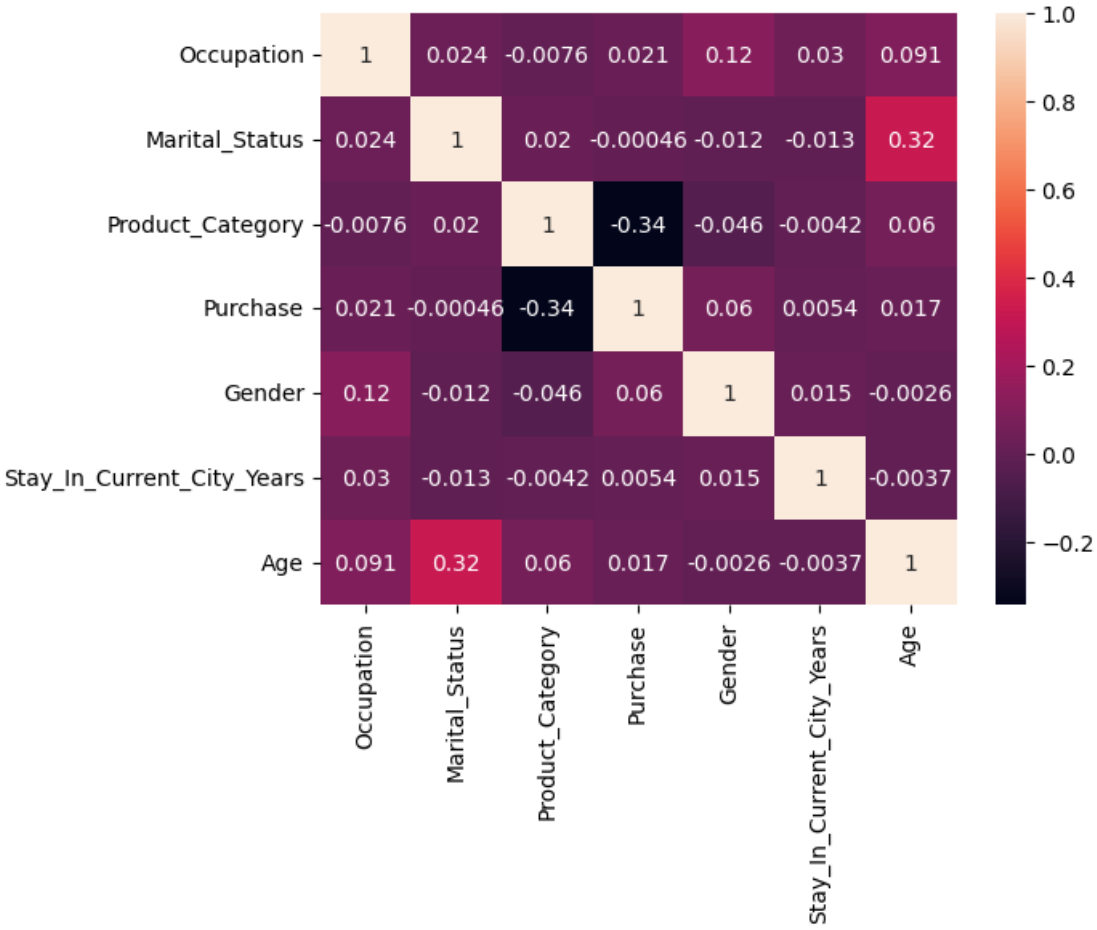
1.3.3 Insight

- Most profit comes from Product_Category 1,5,8.
- Most profit comes from Occupation 0,4,7.
- Most profit comes from City_Category B,C.
- Most profit comes from Stay_In_Current_City_Years=1,2 or 3 years.

1.3.4 Correlation

[18]:
```
corr=pd.read_csv('Black_Friday.txt')
corr['Gender']=corr['Gender'].map({'M':'1','F':'0'}).astype(int)
stay_mapping = {'0': 0,'1': 1,'2': 2,'3': 3,'4+': 4}
corr['Stay_In_Current_City_Years'] = corr['Stay_In_Current_City_Years'].
 ↪map(stay_mapping)
age_mapping = {'0-17': (0 + 17) / 2,'55+': 55,'26-35': (26 + 35) / 2,'46-50':␣
 ↪(46 + 50) / 2,'51-55': (51 + 55) / 2,'36-45': (36 + 45) / 2,'18-25': (18 +␣
 ↪25)/2}
```

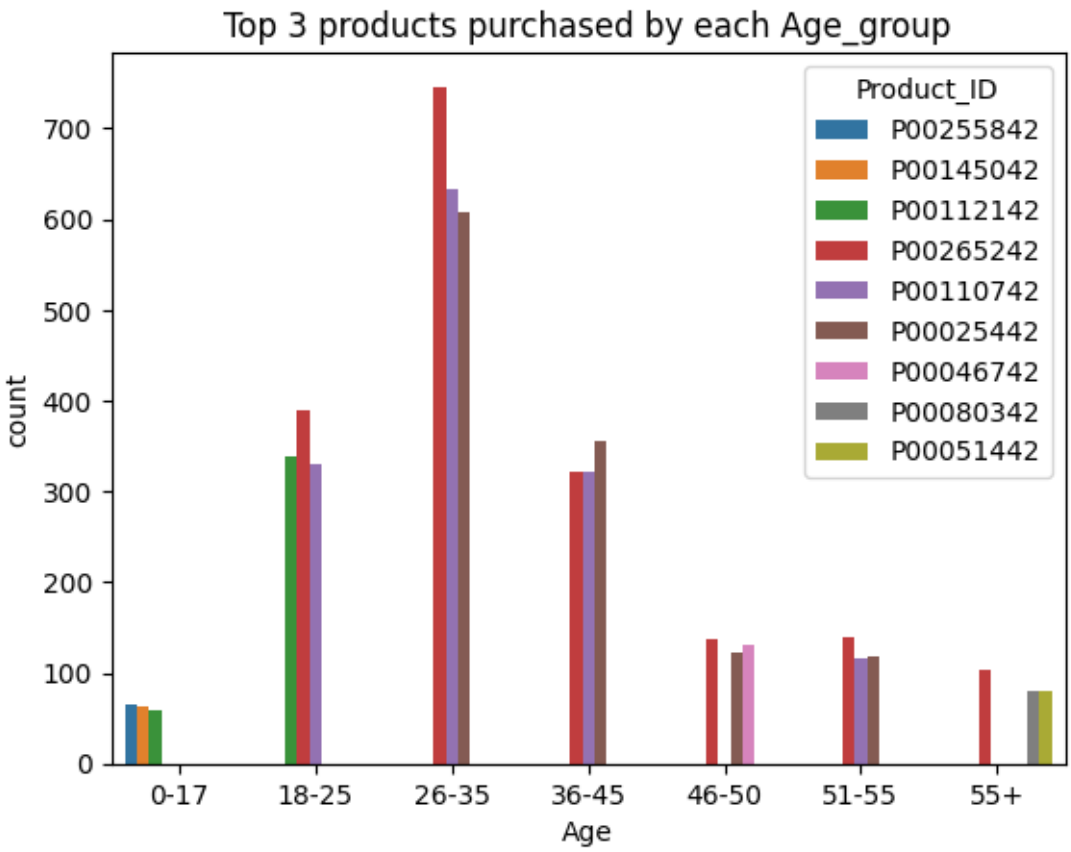
```
corr['Age'] = corr['Age'].map(age_mapping)
corr=corr[['Occupation','Marital_Status','Product_Category','Purchase','Gender','Stay_In_Curre
sns.heatmap(corr.corr(),annot=True)
plt.show()
```



- No real findings or correlation as most of them are categorical columns.

1.3.5 a) What products are different age groups buying?

[19]:
```
g=df.groupby('Age',as_index=False)['Product_ID'].value_counts()
g['r']=g.groupby('Age',as_index=False)['count'].
  ↪rank(method='dense',ascending=False)
g=g[g['r']<=3]
sns.barplot(data=g,x='Age',y='count',hue='Product_ID')
plt.title('Top 3 products purchased by each Age_group')
plt.show()
```

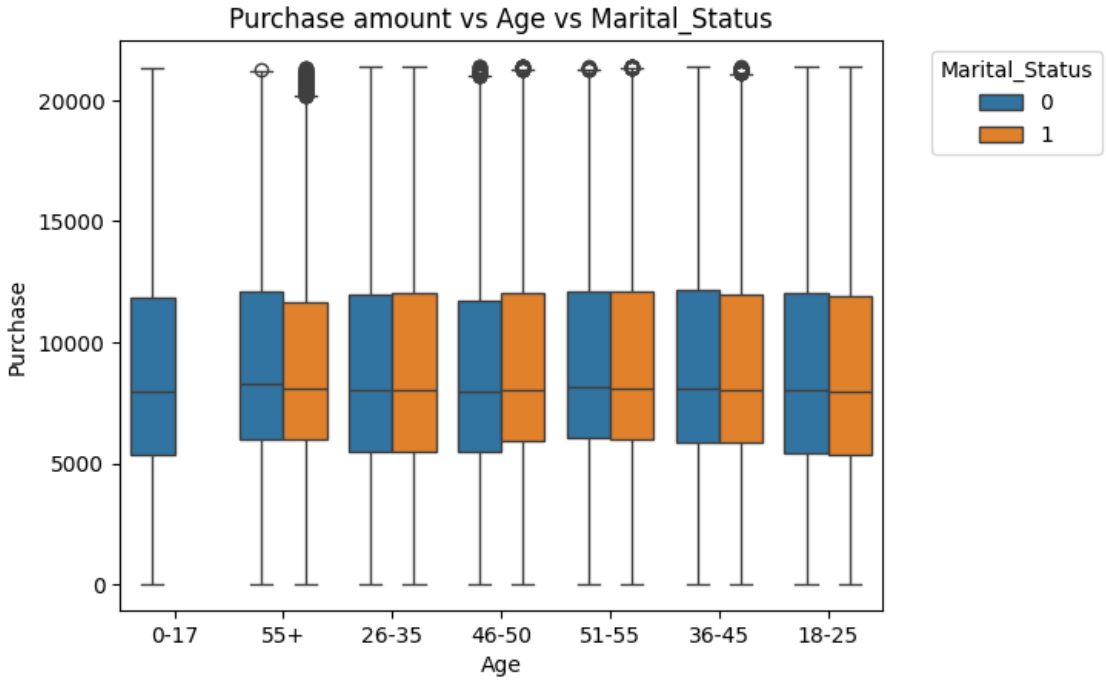


1.3.6 Insight

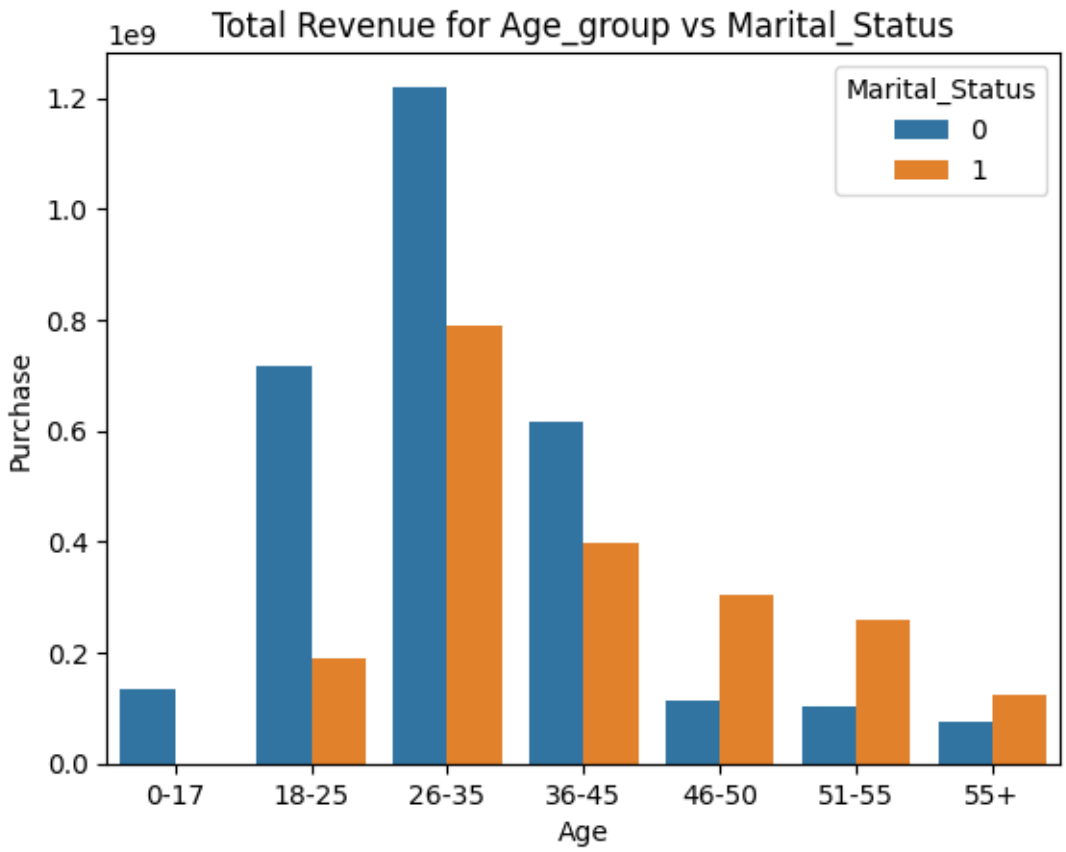
- P00265242 is one common product frequently purchased across all age groups
- P00110742 and P00025442 is also common.

1.3.7 b) Is there a relationship between age, marital status, and the amount spent?

[59]:
```
sns.boxplot(data=df,x='Age',y='Purchase',hue='Marital_Status')
plt.title('Purchase amount vs Age vs Marital_Status')
plt.legend(title='Marital_Status', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()
```



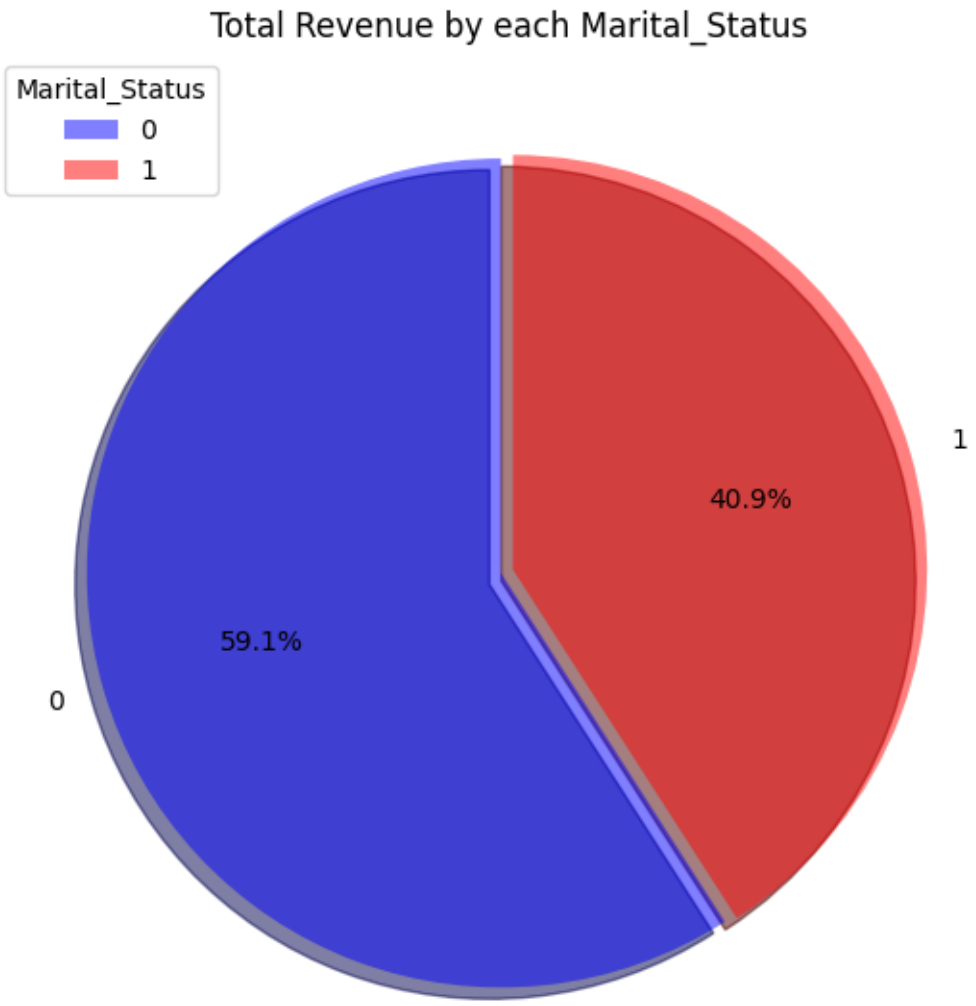
[53]:
```
g=df.groupby(['Age','Marital_Status'],as_index=False)['Purchase'].sum()
sns.barplot(data=g, x='Age', y='Purchase',hue='Marital_Status')
plt.title('Total Revenue for Age_group vs Marital_Status')
plt.show()
```



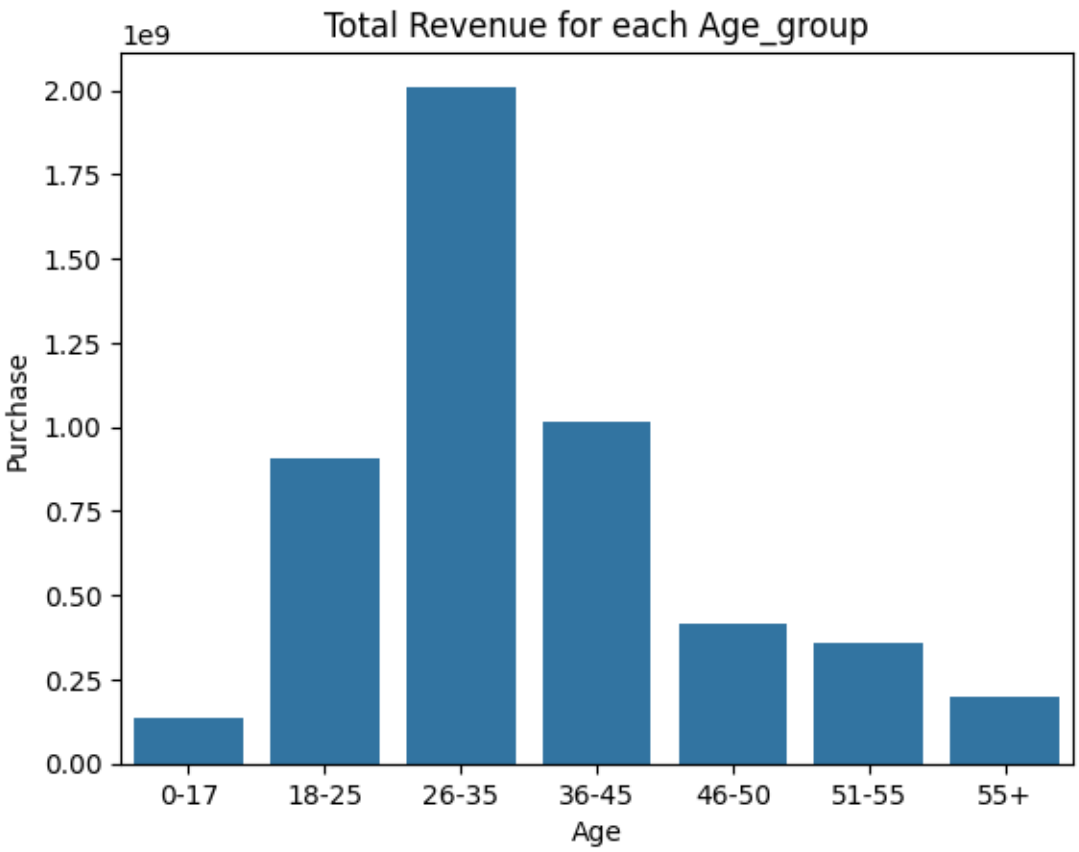
1.3.8 Insight

- The boxplot does not show much conclusive relationship between purchase amount,age and marital_status.
- When we aggregate the purchase amount then we can see that unmarried people of age 18-45 are generating the most revenue.

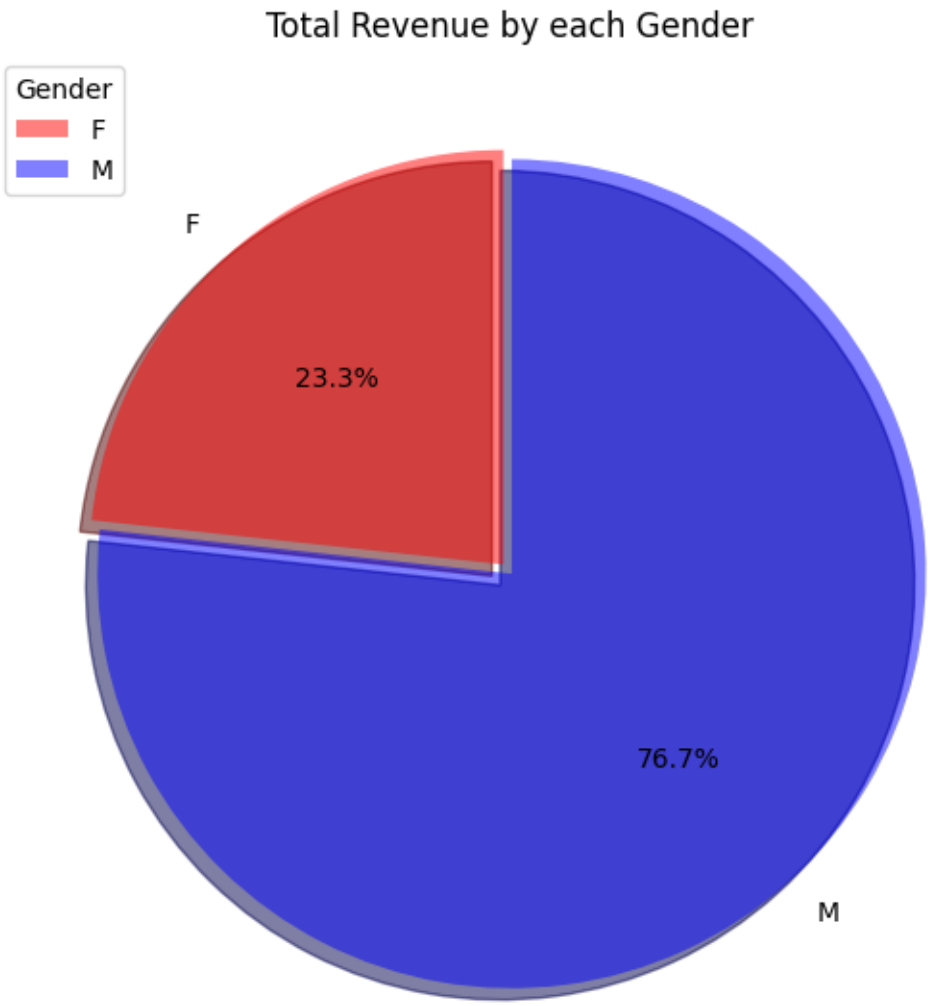
[22]:
```
marital_data = df.groupby('Marital_Status')['Purchase'].sum()
explode = [0.02, 0.01]
colors = [(0, 0, 1, 0.5), (1, 0, 0, 0.5)]  # (R, G, B, Alpha)
plt.figure(figsize=(7, 7))
plt.pie(marital_data, labels=marital_data.index, autopct='%1.1f%%',␣
 ↪startangle=90, colors=colors, explode=explode,shadow=True)
plt.title('Total Revenue by each Marital_Status')
plt.legend(title='Marital_Status', loc='upper left')
plt.show()
```



[60]:
```
g=df.groupby('Age',as_index=False)['Purchase'].sum()
sns.barplot(data=g, x='Age', y='Purchase')
plt.title('Total Revenue for each Age_group')
plt.show()
```



[61]:
```
gender_data = df.groupby('Gender')['Purchase'].sum()
explode = [0.02, 0.01]
colors = [(1, 0, 0, 0.5), (0, 0, 1, 0.5)] # (R, G, B, Alpha)
plt.figure(figsize=(7, 7))
plt.pie(gender_data, labels=gender_data.index, autopct='%1.1f%%',startangle=90,␣
 ↪colors=colors, explode=explode,shadow=True)
plt.title('Total Revenue by each Gender')
plt.legend(title='Gender', loc='upper left')
plt.show()
```



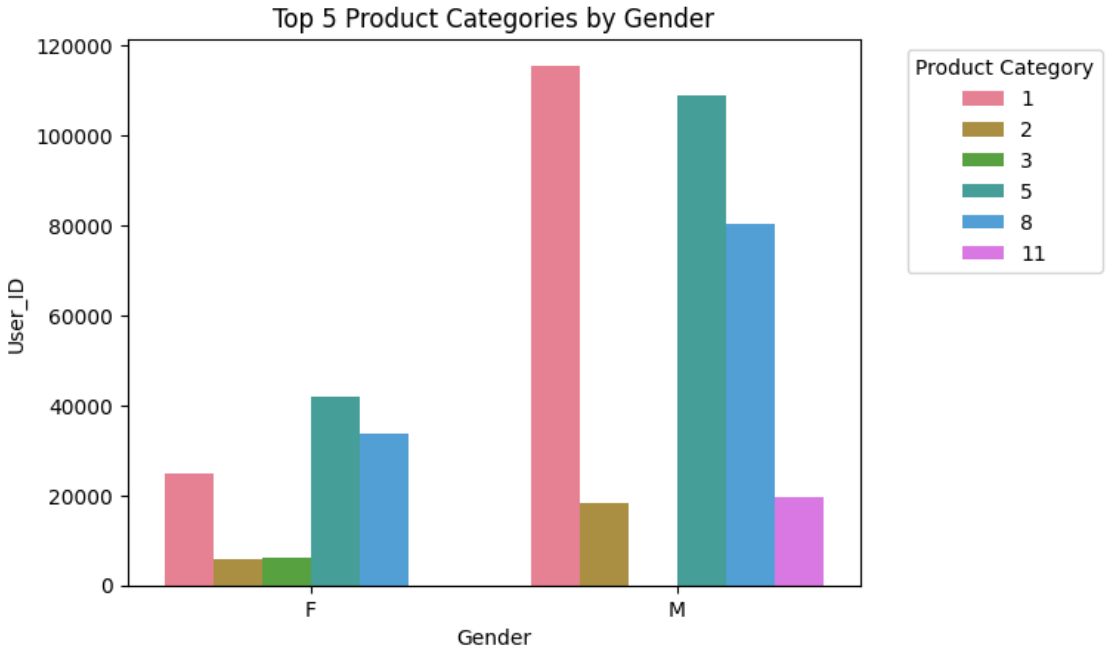
1.3.9 Insight

- Most revenue is generated by males.
- People with age 26-35 and 36-45 are spending the most.
- Unmarried people are purchasing more than married people.

1.3.10 c) Are there preferred product categories for different genders?

[23]:
```
g = df.groupby(['Gender', 'Product_Category'], as_index=False)['User_ID'].
  ↪count()
g['r'] = g.groupby('Gender')['User_ID'].rank(method='dense', ascending=False)
g = g[g['r'] <= 5]
custom_palette = sns.color_palette("husl", 6)
```

```
sns.barplot(data=g, x='Gender', y='User_ID', hue='Product_Category',␣
 ↪palette=custom_palette)
plt.title('Top 5 Product Categories by Gender')
plt.legend(title='Product Category', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()
```

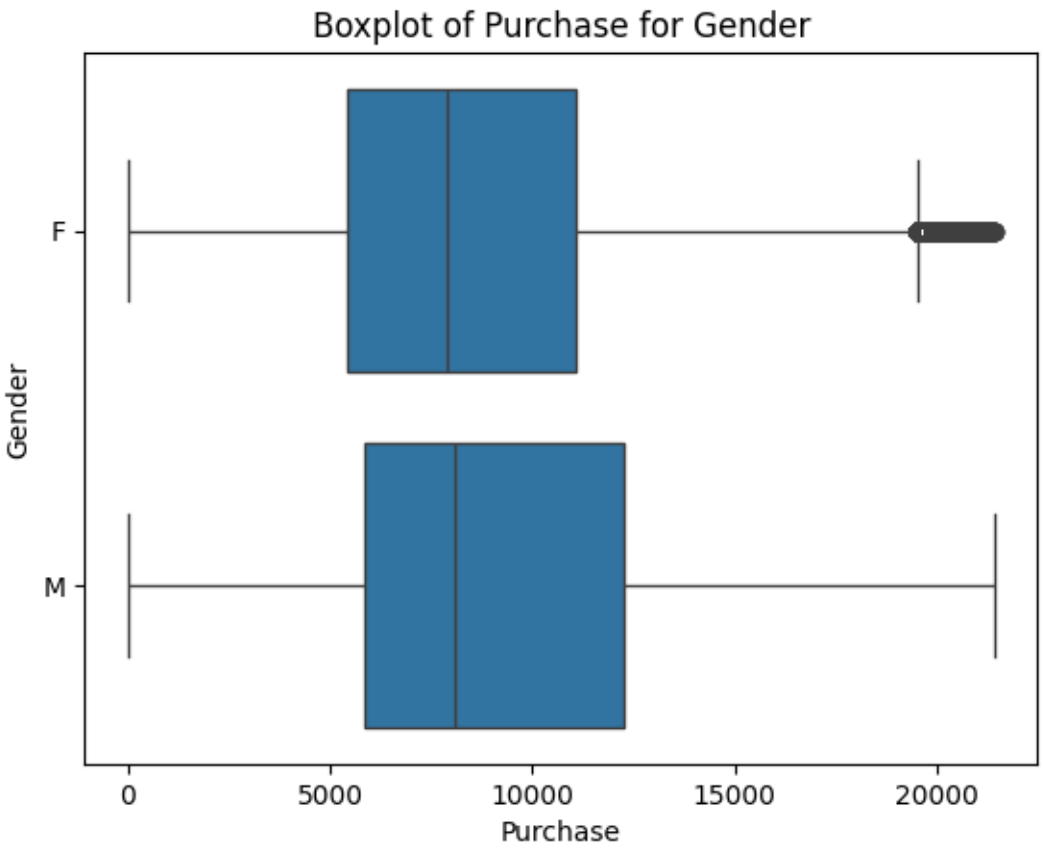


1.3.11 Inisight

- 1,2,5,8 are commonly popular product categories purchased by both females and males.
- 3 is only popular in females whereas 11 is only popular in males.

1.4 4.How does gender affect the amount spent?

[24]:
```
sns.boxplot(data=df,y='Gender',x='Purchase')
plt.title('Boxplot of Purchase for Gender')
plt.show()
```



[25]:
```
df.groupby('Gender',as_index=False)['User_ID'].nunique()
```

[25]:

	Gender	User_ID
0	F	1666
1	M	4225

1.4.1 Insight

- Boxplot shows that male users tend to spend more.
- More male users are present.

[26]:
```
female=df[df['Gender']=='F']['Purchase']
male=df[df['Gender']=='M']['Purchase']
df.groupby('Gender')['Purchase'].describe()
```

[26]:

Gender	count	mean	std	min	25%	50%	75%	\
F	135220.0	8671.049039	4679.058483	12.0	5429.0	7906.0	11064.0	
M	412171.0	9367.724355	5009.234088	12.0	5852.0	8089.0	12247.0	

```
            max
Gender
F       21398.0
M       21399.0
```

[27]:
```
female.shape
```

[27]:
```
(135220,)
```

[28]:
```
male.shape
```

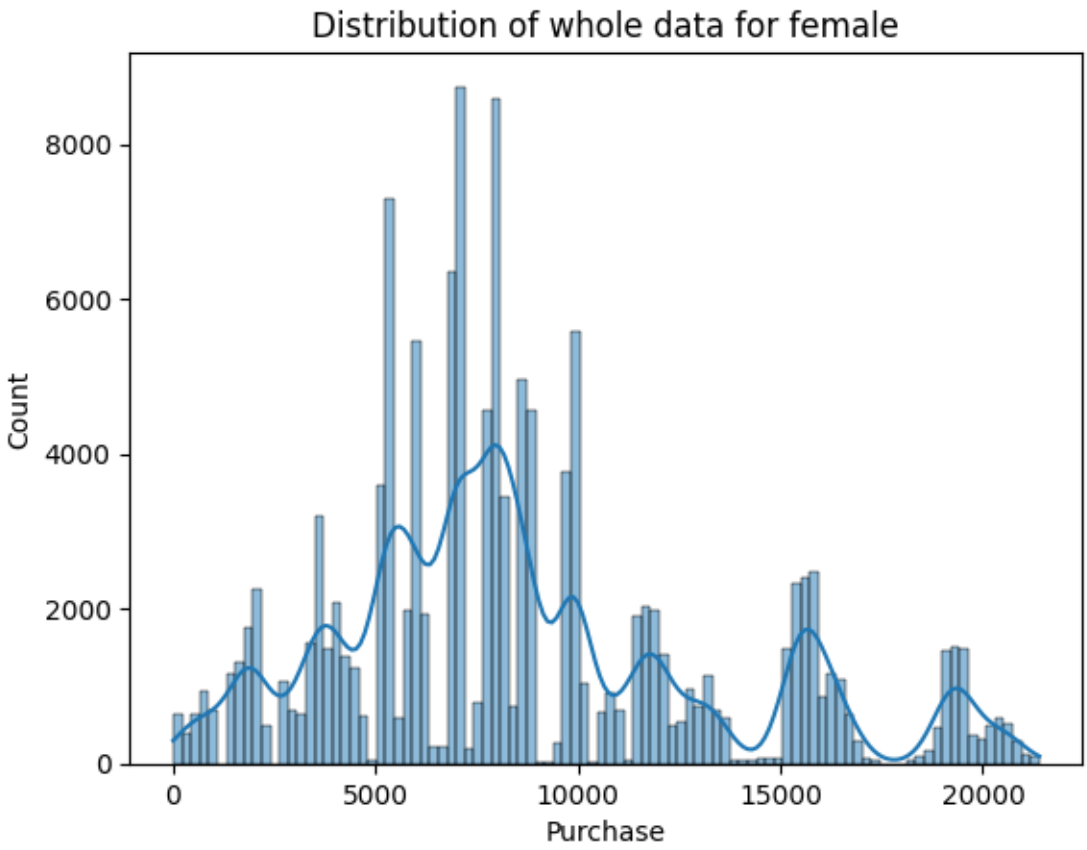
[28]:
```
(412171,)
```

1.4.2 Clearly dataset is large enough to apply CLT (>30)

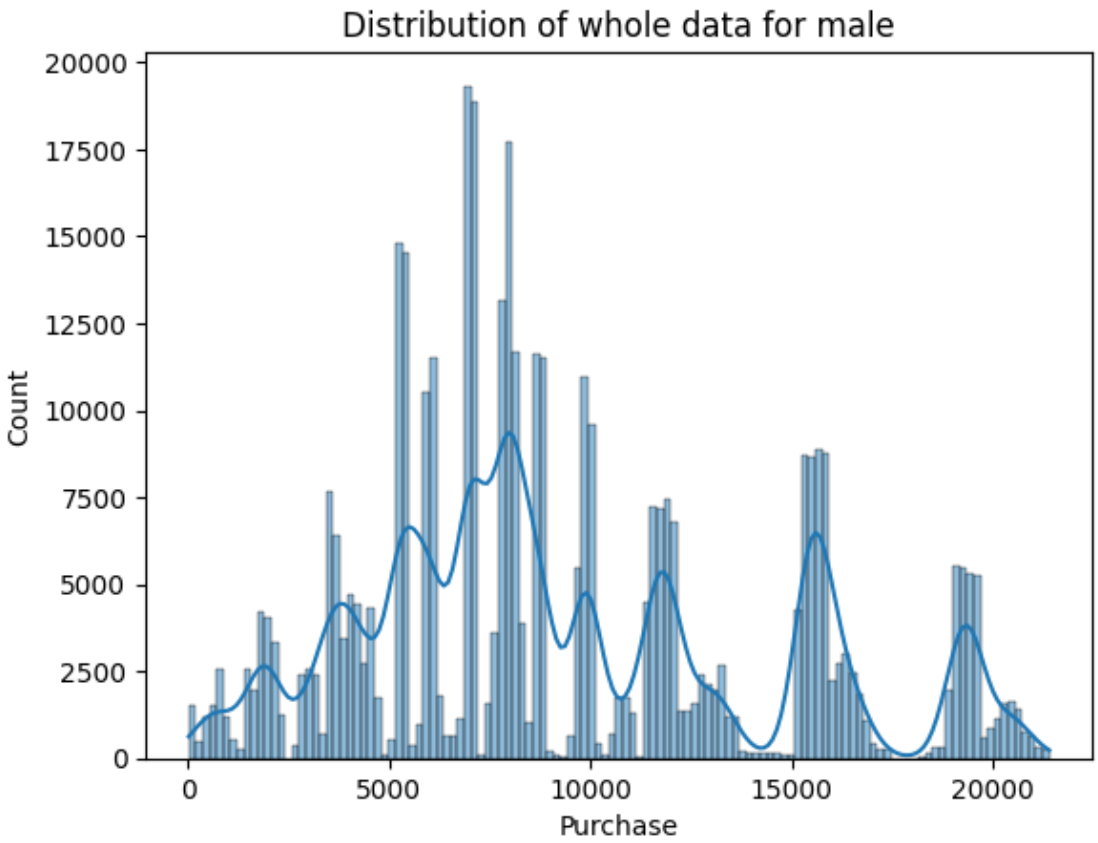
1.4.3 i) CLT on whole dataset

[29]:
```
def construct95ciwhole(series,label):
    n=len(series)
    mean,std=series.mean(),np.std(series,ddof=1)/np.sqrt(n)
    lower_bound,upper_bound=mean+stats.norm.ppf(.025)*std,mean+stats.norm.ppf(.
 ↪975)*std
    age_groups = ['51-55', '36-45', '55+', '26-35', '46-50', '18-25', '0-17']
    if label not in age_groups:
        sns.histplot(series,kde=True)
        plt.title(f'Distribution of whole data for {label}')
        plt.show()
    print(f"The 95% confidence interval for whole data for'Purchase' amount for␣
 ↪{label} is: [{lower_bound:.2f}, {upper_bound:.2f}]")
```

[30]:
```
construct95ciwhole(female,'female')
construct95ciwhole(male,'male')
```




```
The 95% confidence interval for whole data for'Purchase' amount for female is:
[8646.11, 8695.99]
```




```
The 95% confidence interval for whole data for'Purchase' amount for male is:
[9352.43, 9383.02]
```

1.4.4 Insight

- Female confidence interval is slightly wider reflecting slightly higher variability.

[31]:
```
np.random.seed(95)
def construct_ci(n,series,label,a):
    sample=[]
    for i in range(1000):
        sample.append(np.mean(np.random.choice(series,size=n,replace=True)))
    mean,std=np.mean(sample),np.std(sample,ddof=1)
    lower_bound,upper_bound=mean+stats.norm.ppf(a/2)*std,mean+stats.norm.
 ↪ppf(1-a/2)*std
    if a==.1:
        print(f"The 90% confidence interval for 'Purchase' amount for {label}␣
 ↪for n={n} is: [{lower_bound:.2f}, {upper_bound:.2f}]")
    else:print(f"The 95% confidence interval for 'Purchase' amount for {label}␣
 ↪for n ={n} is: [{lower_bound:.2f}, {upper_bound:.2f}]")
```

1.4.5 ii) n=300

[32]:
```
construct_ci(300,female,'female',.05)
construct_ci(300,male,'male',.05)
```

```
The 95% confidence interval for 'Purchase' amount for female for n =300 is:
[8109.57, 9229.04]
The 95% confidence interval for 'Purchase' amount for male for n =300 is:
[8802.20, 9923.32]
```

1.4.6 iii) n=3000

[33]:
```
construct_ci(3000,female,'female',.05)
construct_ci(3000,male,'male',.05)
```

```
The 95% confidence interval for 'Purchase' amount for female for n =3000 is:
[8511.84, 8841.72]
The 95% confidence interval for 'Purchase' amount for male for n =3000 is:
[9186.64, 9554.44]
```

1.4.7 iv) n=30000

[34]:
```
construct_ci(30000,female,'female',.05)
construct_ci(30000,male,'male',.05)
```

```
The 95% confidence interval for 'Purchase' amount for female for n =30000 is:
[8617.46, 8724.31]
The 95% confidence interval for 'Purchase' amount for male for n =30000 is:
[9313.23, 9424.12]
```

1.4.8 v) Increasing alpha to .1

[35]:
```
construct_ci(30000,female,'female',.1)
construct_ci(30000,male,'male',.1)
```

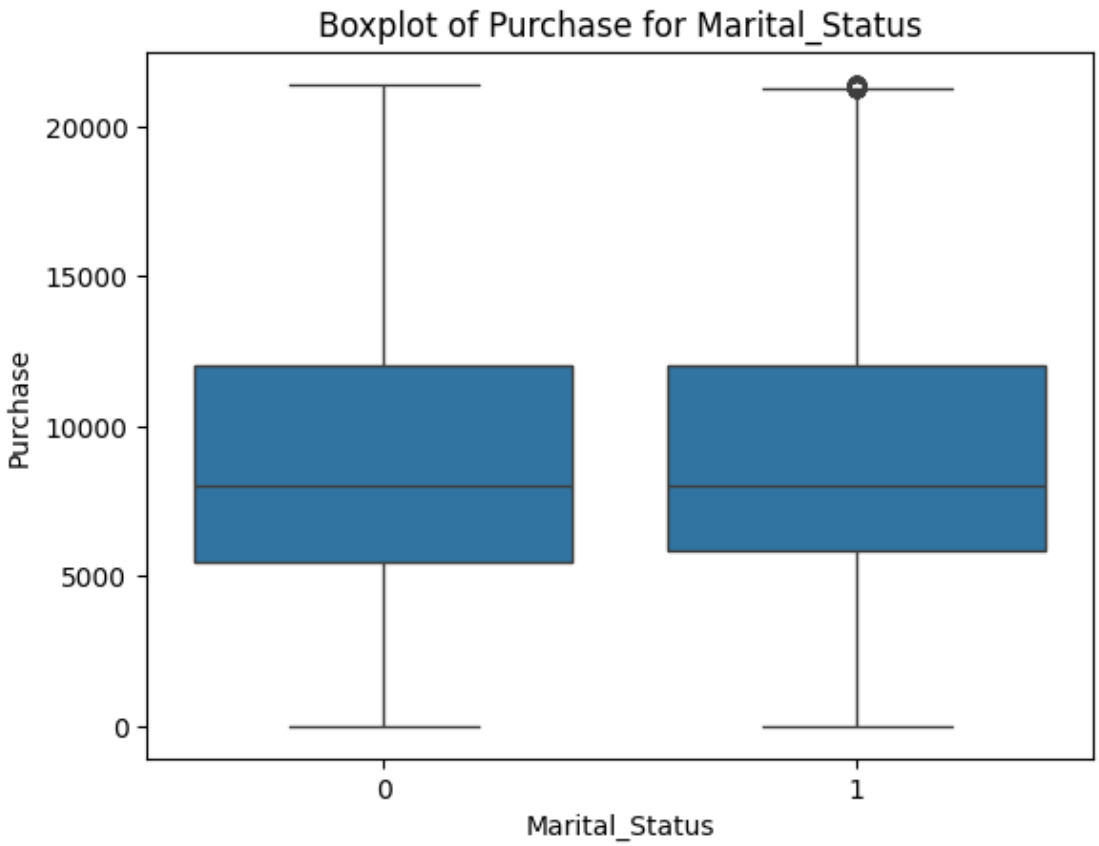
```
The 90% confidence interval for 'Purchase' amount for female for n=30000 is:
[8628.03, 8714.03]
The 90% confidence interval for 'Purchase' amount for male for n=30000 is:
[9319.03, 9415.17]
```

1.4.9 Insight

- With increase in sample size the confidence intervals are getting more concentrated and converging towards the actual mean.
- We can clearly see for larger sample size or slightly higher alpha, males spend more than female.

1.5 5.How does marital status affect the amount spent?

[36]:
```
sns.boxplot(data=df,x='Marital_Status',y='Purchase')
plt.title('Boxplot of Purchase for Marital_Status')
plt.show()
```



[37]:
```
df.groupby('Marital_Status',as_index=False)['User_ID'].nunique()
```

[37]:

	Marital_Status	User_ID
0	0	3417
1	1	2474

1.5.1 Insight

- Boxplot is almost equal.
- More unmarried users are present.

[38]:
```
married=df[df['Marital_Status']==1]['Purchase']
unmarried=df[df['Marital_Status']==0]['Purchase']
df.groupby('Marital_Status')['Purchase'].describe()
```

[38]:

Marital_Status	count	mean	std	min	25%	50%	75%	max
0	323242.0	9201.581849	4948.327397	12.0	5480.0	8035.0	12028.0	21399.0
1	224149.0	9187.040076	4925.205232	12.0	5833.0	8042.0	12006.0	21398.0

[39]:
```
married.shape
```

[39]: (224149,)

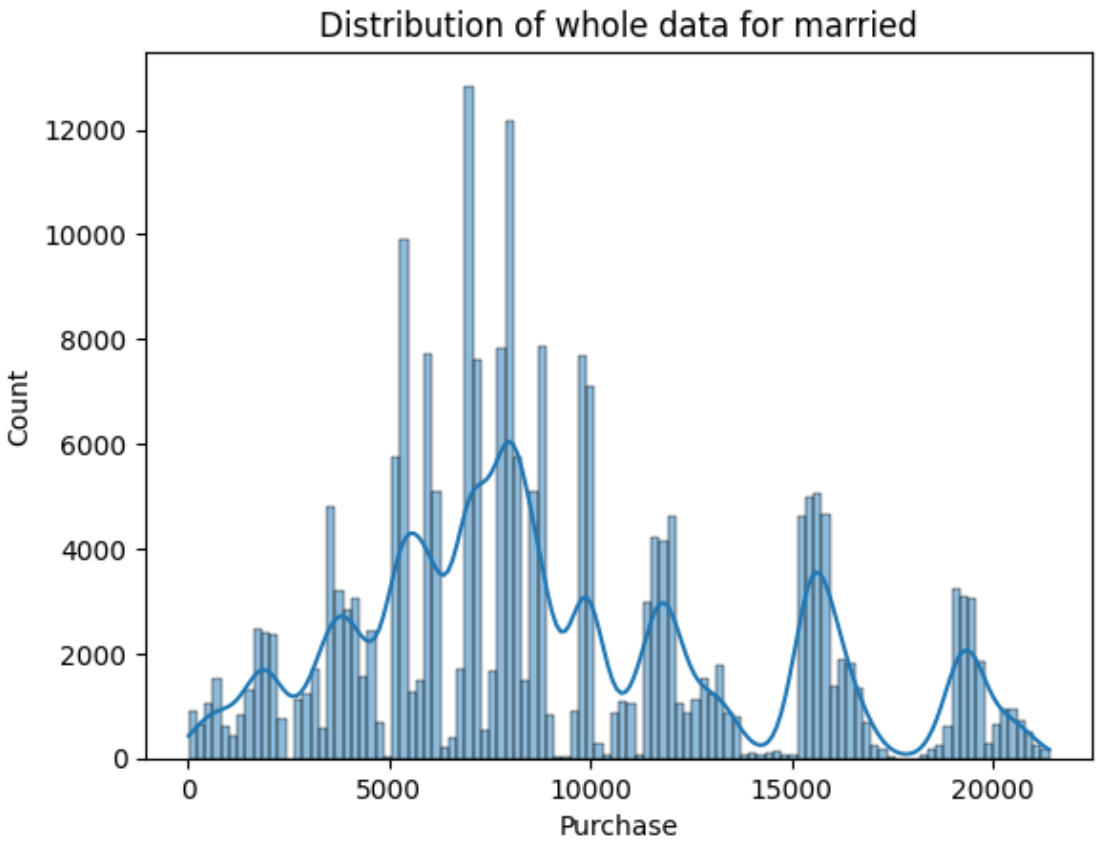
[40]:
```
unmarried.shape
```

[40]: (323242,)

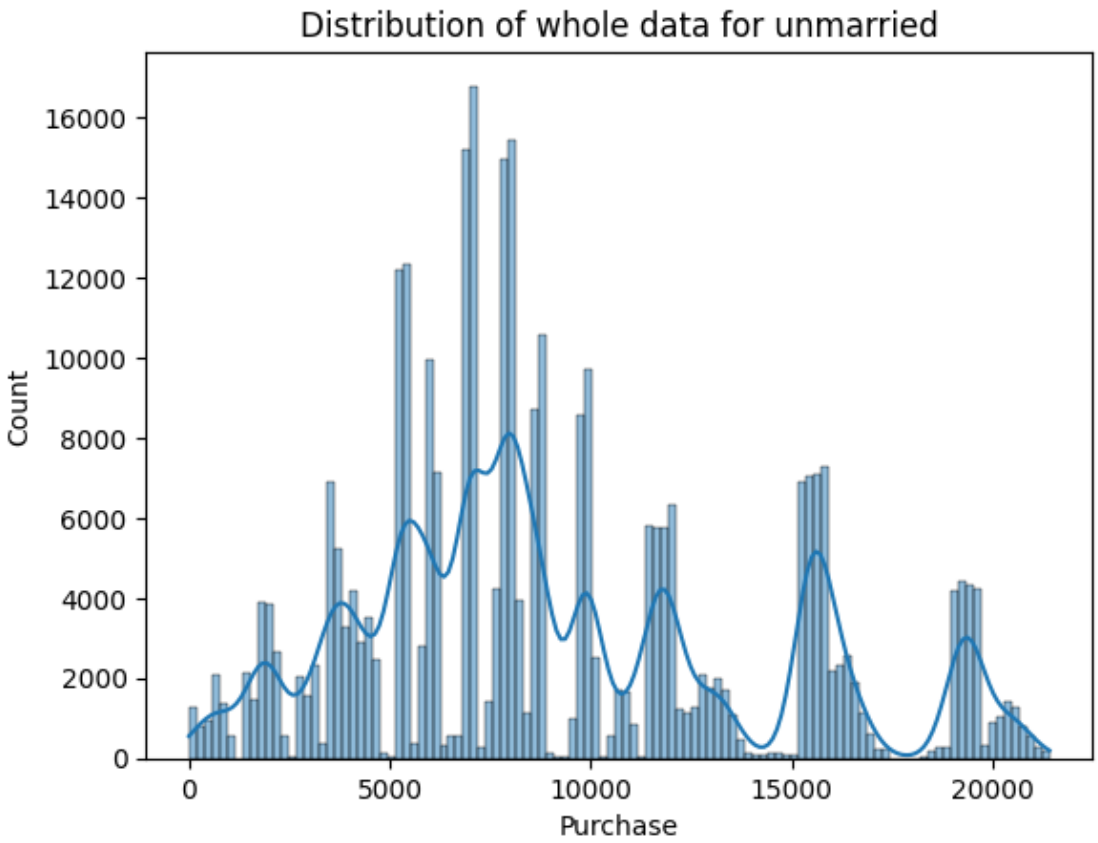
1.5.2 Clearly dataset is large enough to apply CLT (>30)

1.5.3 i) CLT on whole dataset

[41]:
```
construct95ciwhole(married,'married')
construct95ciwhole(unmarried,'unmarried')
```



The 95% confidence interval for whole data for'Purchase' amount for married is:
[9166.65, 9207.43]




```
The 95% confidence interval for whole data for'Purchase' amount for unmarried
is: [9184.52, 9218.64]
```

1.5.4 ii) n=300

[42]:
```
construct_ci(300,married,'married',.05)
construct_ci(300,unmarried,'unmarried',.05)
```

```
The 95% confidence interval for 'Purchase' amount for married for n =300 is:
[8633.83, 9760.12]
The 95% confidence interval for 'Purchase' amount for unmarried for n =300 is:
[8654.64, 9754.73]
```

1.5.5 iii) n=3000

[43]:
```
construct_ci(3000,married,'married',.05)
construct_ci(3000,unmarried,'unmarried',.05)
```

```
The 95% confidence interval for 'Purchase' amount for married for n =3000 is:
[9016.09, 9364.30]
```

```
The 95% confidence interval for 'Purchase' amount for unmarried for n =3000 is:
[9026.22, 9378.78]
```

1.5.6 iv) n=30000

[44]:
```
construct_ci(30000,married,'married',.05)
construct_ci(30000,unmarried,'unmarried',.05)
```

```
The 95% confidence interval for 'Purchase' amount for married for n =30000 is:
[9129.94, 9240.55]
The 95% confidence interval for 'Purchase' amount for unmarried for n =30000 is:
[9144.52, 9256.77]
```

1.5.7 v) Increasing alpha to .1

[45]:
```
construct_ci(30000,married,'married',.1)
construct_ci(30000,unmarried,'unmarried',.1)
```

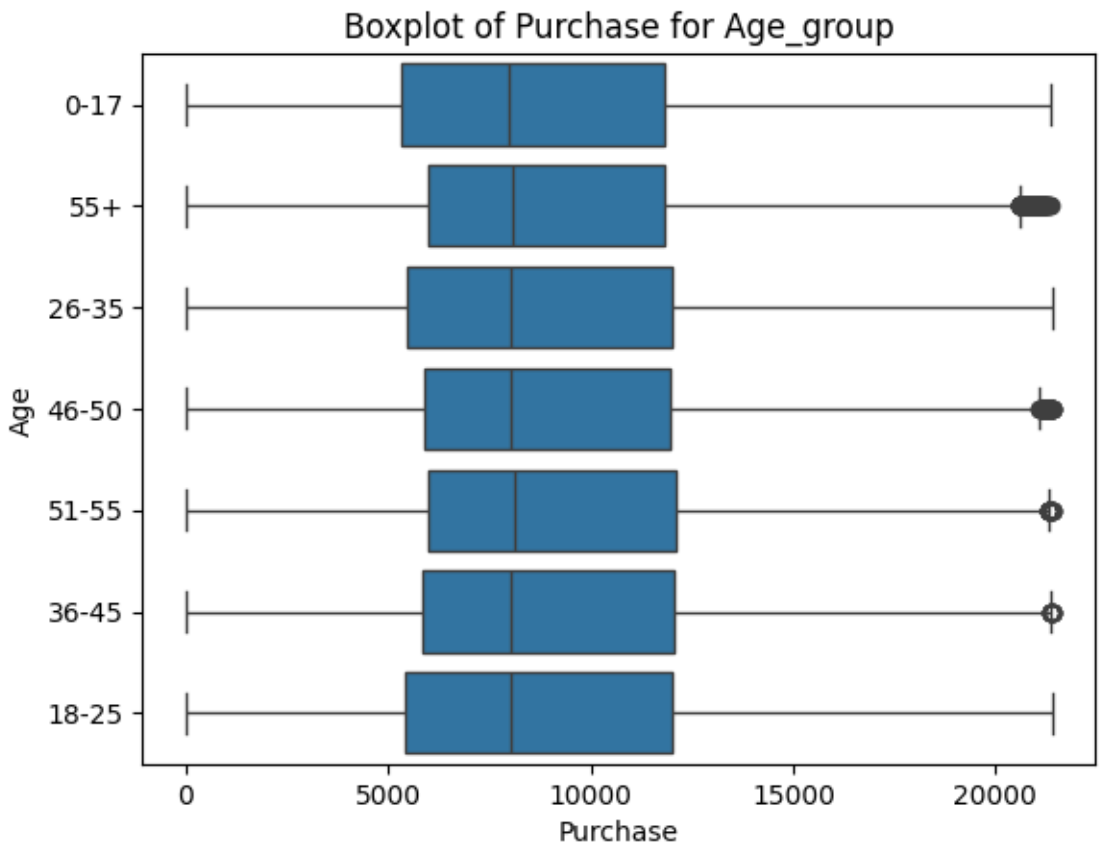
```
The 90% confidence interval for 'Purchase' amount for married for n=30000 is:
[9137.53, 9235.78]
The 90% confidence interval for 'Purchase' amount for unmarried for n=30000 is:
[9156.39, 9248.02]
```

1.5.8 Insight

- With increase in sample size the confidence intervals are getting more concentrated and converging towards the actual mean.
- Even with increase in alpha still we see overlapping intervals so cannot conclude anything.

1.6 6.How does age affect the amount spent?

[46]:
```
sns.boxplot(data=df,x='Purchase',y='Age')
plt.title('Boxplot of Purchase for Age_group')
plt.show()
```



[47]:
```
df.groupby('Age',as_index=False)['User_ID'].nunique()
```

[47]:

	Age	User_ID
0	0-17	218
1	18-25	1069
2	26-35	2053
3	36-45	1167
4	46-50	531
5	51-55	481
6	55+	372

1.6.1 Insight

- Cannot conclude much from Boxplot.
- Most users are of age 26-35.

[48]:
```
df.groupby('Age')['Purchase'].describe()
```

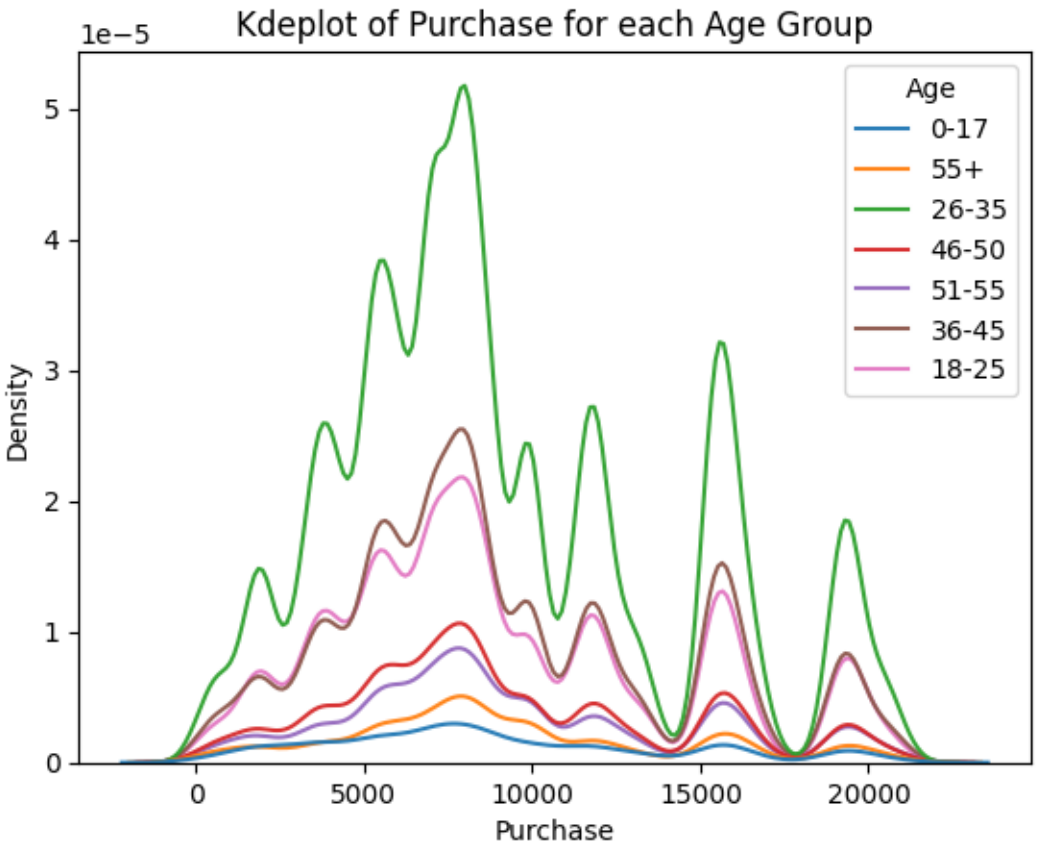
[48]:

Age	count	mean	std	min	25%	50%	75%	\

```
0-17    15032.0  8867.447046  5030.052846  12.0  5324.0  7974.5  11833.25
18-25   99334.0  9124.031731  4978.831062  12.0  5412.0  8020.0  12004.00
26-35  218661.0  9193.469924  4937.410901  12.0  5471.0  8021.0  12018.00
36-45  109409.0  9254.202214  4927.744433  12.0  5866.0  8051.0  12065.00
46-50   45442.0  9128.985080  4867.413951  12.0  5879.0  8025.0  11958.00
51-55   38191.0  9423.121704  4953.644650  12.0  6007.0  8118.0  12123.00
55+     21322.0  9216.650220  4861.626596  12.0  6007.0  8092.5  11837.75
```

```
           max
Age
0-17   21342.0
18-25  21398.0
26-35  21398.0
36-45  21399.0
46-50  21391.0
51-55  21388.0
55+    21345.0
```

[49]:
```
sns.kdeplot(data=df,x='Purchase',hue='Age',alpha=.95)
plt.title('Kdeplot of Purchase for each Age Group')
plt.show()
```



[50]:
```
age_groups = ['51-55', '36-45', '55+', '26-35', '46-50', '18-25', '0-17']
for age_group in age_groups:
    age_data = df[df['Age'] == age_group]['Purchase']
    print(f"\nAge Group: {age_group}")
    assert len(age_data) > 30
    print('We can apply CLT on the dataset as length of dataset is greater than␣
 ↪30')
    construct95ciwhole(age_data, age_group)
    print()
    construct_ci(300, age_data, age_group, 0.05)
    construct_ci(3000, age_data, age_group, 0.05)
    construct_ci(30000, age_data, age_group, 0.05)
    construct_ci(30000, age_data, age_group, 0.1)
```

```
Age Group: 51-55
We can apply CLT on the dataset as length of dataset is greater than 30
The 95% confidence interval for whole data for'Purchase' amount for 51-55 is:
[9373.44, 9472.80]

The 95% confidence interval for 'Purchase' amount for 51-55 for n =300 is:
[8867.87, 10005.47]
The 95% confidence interval for 'Purchase' amount for 51-55 for n =3000 is:
[9253.01, 9608.82]
The 95% confidence interval for 'Purchase' amount for 51-55 for n =30000 is:
[9367.31, 9477.39]
The 90% confidence interval for 'Purchase' amount for 51-55 for n=30000 is:
[9377.03, 9473.05]

Age Group: 36-45
We can apply CLT on the dataset as length of dataset is greater than 30
The 95% confidence interval for whole data for'Purchase' amount for 36-45 is:
[9225.00, 9283.40]

The 95% confidence interval for 'Purchase' amount for 36-45 for n =300 is:
[8690.19, 9839.01]
The 95% confidence interval for 'Purchase' amount for 36-45 for n =3000 is:
[9064.70, 9438.75]
The 95% confidence interval for 'Purchase' amount for 36-45 for n =30000 is:
[9198.89, 9307.60]
The 90% confidence interval for 'Purchase' amount for 36-45 for n=30000 is:
[9210.97, 9299.24]

Age Group: 55+
We can apply CLT on the dataset as length of dataset is greater than 30
The 95% confidence interval for whole data for'Purchase' amount for 55+ is:
```

```
[9151.39, 9281.91]

The 95% confidence interval for 'Purchase' amount for 55+ for n =300 is:
[8656.46, 9768.01]
The 95% confidence interval for 'Purchase' amount for 55+ for n =3000 is:
[9039.11, 9395.80]
The 95% confidence interval for 'Purchase' amount for 55+ for n =30000 is:
[9162.13, 9273.73]
The 90% confidence interval for 'Purchase' amount for 55+ for n=30000 is:
[9172.42, 9261.71]

Age Group: 26-35
We can apply CLT on the dataset as length of dataset is greater than 30
The 95% confidence interval for whole data for'Purchase' amount for 26-35 is:
[9172.78, 9214.16]

The 95% confidence interval for 'Purchase' amount for 26-35 for n =300 is:
[8664.26, 9733.31]
The 95% confidence interval for 'Purchase' amount for 26-35 for n =3000 is:
[9017.57, 9368.20]
The 95% confidence interval for 'Purchase' amount for 26-35 for n =30000 is:
[9137.07, 9250.46]
The 90% confidence interval for 'Purchase' amount for 26-35 for n=30000 is:
[9147.88, 9239.28]

Age Group: 46-50
We can apply CLT on the dataset as length of dataset is greater than 30
The 95% confidence interval for whole data for'Purchase' amount for 46-50 is:
[9084.23, 9173.74]

The 95% confidence interval for 'Purchase' amount for 46-50 for n =300 is:
[8567.54, 9676.85]
The 95% confidence interval for 'Purchase' amount for 46-50 for n =3000 is:
[8949.45, 9308.81]
The 95% confidence interval for 'Purchase' amount for 46-50 for n =30000 is:
[9074.08, 9185.10]
The 90% confidence interval for 'Purchase' amount for 46-50 for n=30000 is:
[9084.63, 9173.09]

Age Group: 18-25
We can apply CLT on the dataset as length of dataset is greater than 30
The 95% confidence interval for whole data for'Purchase' amount for 18-25 is:
[9093.07, 9154.99]

The 95% confidence interval for 'Purchase' amount for 18-25 for n =300 is:
[8586.08, 9678.17]
The 95% confidence interval for 'Purchase' amount for 18-25 for n =3000 is:
[8942.37, 9302.26]
```

```
The 95% confidence interval for 'Purchase' amount for 18-25 for n =30000 is:
[9066.65, 9183.29]
The 90% confidence interval for 'Purchase' amount for 18-25 for n=30000 is:
[9076.53, 9170.73]

Age Group: 0-17
We can apply CLT on the dataset as length of dataset is greater than 30
The 95% confidence interval for whole data for'Purchase' amount for 0-17 is:
[8787.04, 8947.86]

The 95% confidence interval for 'Purchase' amount for 0-17 for n =300 is:
[8276.40, 9421.17]
The 95% confidence interval for 'Purchase' amount for 0-17 for n =3000 is:
[8687.96, 9046.68]
The 95% confidence interval for 'Purchase' amount for 0-17 for n =30000 is:
[8813.33, 8925.21]
The 90% confidence interval for 'Purchase' amount for 0-17 for n=30000 is:
[8819.59, 8916.91]
```

1.7 7.Report

1.7.1 a)Gender

1. **Whole Dataset Analysis:**
   - The 95% confidence interval for the purchase amount of female customers ranges from 8646.11 to 8695.99, while for male customers, it ranges from 9352.43 to 9383.02.
   - Insight: The confidence interval for females is slightly wider, reflecting slightly higher dispersion in their purchase behavior.
2. **Effect of Sample Size (n=300, 3000, 30000):**
   - As the sample size increases, the width of the confidence intervals narrows for both genders, indicating greater precision in estimating the population mean purchase amount.
3. **Impact of Significance Level (=0.05, 0.1):**
   - Reducing the significance level to 0.1 results in slightly narrower confidence intervals for both genders, with the 90% confidence interval for females ranging from 8628.03 to 8714.03 and for males ranging from 9319.03 to 9415.17.
4. **Conclusion:**
   - There is a significant difference in the purchase behavior between male and female customers, as evidenced by non-overlapping confidence intervals across different sample sizes and significance levels.Males tend to spend more.

1.7.2 b)Marital Status

1. **Whole Dataset Analysis:**
   - The 95% confidence interval for the purchase amount of married individuals ranges from 9166.65 to 9207.43, while for unmarried individuals, it ranges from 9184.52 to 9218.64.
2. **Effect of Sample Size (n=300, 3000, 30000):**
   - As the sample size increases, the width of the confidence intervals narrows, indicating greater precision in estimating the population mean purchase amount for both married and unmarried individuals.

3. **Impact of Significance Level (=0.05, 0.1):**
    - Lowering the significance level to 0.1 results in slightly narrower confidence intervals, with the 90% confidence intervals for married and unmarried individuals ranging from 9137.53 to 9235.78 and 9156.39 to 9248.02, respectively.
4. **Conclusion:**
    - There is no significant difference in the purchase behavior between married and unmarried individuals, as evidenced by overlapping confidence intervals across different sample sizes and significance levels.

1.7.3 c)Age

From the 90% confidence interval results for n=30000:

- **51-55 Age Group:** The mean purchase amount is estimated to be between 9377.03 and 9473.05.
- **36-45 Age Group:** The mean purchase amount falls within the range of 9210.97 to 9299.24.
- **55+ Age Group:** The mean purchase amount ranges from 9172.42 to 9261.71.
- **26-35 Age Group:** The mean purchase amount is estimated to be between 9147.88 and 9239.28.
- **46-50 Age Group:** The mean purchase amount falls within the range of 9084.63 to 9173.09.
- **18-25 Age Group:** The mean purchase amount ranges from 9076.53 to 9170.73.
- **0-17 Age Group:** The mean purchase amount is estimated to be between 8819.59 and 8916.91.

These intervals provide a narrower range compared to the 95% confidence intervals, reflecting increased precision in the estimates at a slightly lower confidence level.51-55 and 36-45 age groups spend higher than the rest.But we cannot conclude much as many of them are overlapping.Need more data to analyze further and conclude.

1.8 8.Recommendations

- **Targeted Marketing for High-Spending Demographics:** Tailor marketing campaigns towards demographics showing higher purchase amounts, such as the 26-35 age group and unmarried individuals. Highlight popular products like P00265242 and P00110742 to resonate with their preferences.Product categories 1,5,8 Occupations 0,4,7 to be also targeted.
- **Gender-Tailored Promotions:** Develop promotions tailored to specific gender preferences, with a focus on items favored by male shoppers as they spend more. Utilize insights on popular products, including P00265242, to create offers aimed at engaging male customers effectively.
- **Product Placement and Promotion:** Optimize product placement strategies to feature high-margin items preferred by demographics such as unmarried individuals and male consumers. Utilize strategic displays and signage to enhance visibility and drive impulse purchases, maximizing revenue potential.
- **Online Targeting:** Leverage digital channels to target demographics like unmarried individuals or males or of 26-35 years age with personalized marketing messages and exclusive

online offers. Utilize platforms like social media to showcase product assortments aligned with their preferences, driving online sales.

- **Customer Engagement Initiatives:** Implement engagement programs designed to incentivize repeat purchases and foster brand loyalty among unmarried individuals or male demographics or of 18-45 years age. Offer loyalty rewards and referral bonuses, along with targeted promotions on preferred products, to encourage ongoing engagement. Can engage people from city categories B and C or who have stayed in the city for 1,2 or 3 years.
- **Maximizing Revenue from Female Users:** Explore opportunities to increase revenue from female users by introducing gender-specific promotions and product bundles tailored to their preferences. Use identified product categories with high female engagement and develop marketing strategies to capture their attention and drive sales effectively.
- **Studying different Age groups:** With more data we can further investigate why for certain age groups the confidence interval is higher.For example if we know why age group of 51-55 years has higher purchase average then we can boost sales.

[ ]: