Assignment 2

Language Used: Python (version Python 3.9.4)

Submission by Ritwik Babu

In [1]:
```
import pandas as pd
```

Question 1: Some of the orders are stored in another csv file named `bigkart_newsales`. Read the csv file, store it in a data frame and add it to the `bigkart_sales` data frame. Find the total sales value of the category 'Office Supplies' after combining the dataframes

In [2]:
```
df1 = pd.read_csv('bigkart_sales.csv')
```

In [3]:
```
df2 = pd.read_csv('bigkart_newsales.csv')
```

In [4]:
```
df1 = df1.append(df2, ignore_index=True)
```

In [5]:
```
df1
```

Out[5]:

	Order ID	Product Name	Discount	Sales	Profit	Quantity	Category	Sub-Category
0	AZ-2011-1029887	Novimex Color Coded Labels, 5000 Label Set	0.0	26	7	2	Office Supplies	Labels
1	AZ-2011-107716	Deflect-O Door Stop, Erganomic	0.0	85	15	2	Furniture	Furnishings
2	AZ-2011-1087704	Belkin Flash Drive, Bluetooth	0.0	294	109	7	Technology	Accessories
3	AZ-2011-1372644	Panasonic Printer, Durable	0.0	800	168	3	Technology	Machines
4	AZ-2011-1362199	Sanford Pens, Fluorescent	0.5	25	-11	4	Office Supplies	Art
...	...	...	...	...	...	...	...	...
63	AZ-2011-1967754	Logitech Numeric Keypad, USB	0.0	93	40	2	Technology	Accessories
64	AZ-2011-1976919	Boston Markers, Blue	0.0	132	54	5	Office Supplies	Art
65	AZ-2011-2001312	Avery Binding Machine, Clear	0.0	97	12	2	Office Supplies	Binders
66	AZ-2011-2002251	SanDisk Computer Printout Paper, 8.5 x 11	0.0	136	15	4	Office Supplies	Paper
67	AZ-2011-201891	Cameo Clasp Envelope, with clear poly window	0.0	52	19	4	Office Supplies	Envelopes

68 rows × 8 columns

Question 2: There are some duplicate rows in the data frame. Drop these rows and calculate the total sales value of the category Office Supplies.

In [6]:

```
# drop duplicate rows
df = df1.drop_duplicates()
len(df)
```

Out[6]:

```
61
```

In [7]:

```
# total sales value of category 'Office Supplies'
df[df['Category'] == 'Office Supplies']['Sales'].sum()
```

Out[7]:

```
6964
```

Question 3: Find the most profitable category and sub category combination based on the net profit.

In [8]:

```
# group data by combination of 'Category' and 'Sub-Category'
# sum all the rows for each grouped combination to get net profit
df3 = df.groupby(['Category', 'Sub-Category']).sum()
```

In [9]:

```
# get the row having the highest profit
df3[df3['Profit'] == df3['Profit'].max()]
```

Out[9]:

		Discount	Sales	Profit	Quantity
Category	**Sub-Category**				
Technology	**Phones**	0.0	5199	1618	26

Question 4: How many invalid order IDs are there in the data frame. An order id is of the form AZ-2011-Y where Y represents a whole number. A Order ID is said to be valid only if Y consists of 7 digits. Find the number of invalid order order IDs in the data frame.

In [10]:

```
# match pattern for 'Order ID' using regular expression and filter out unmatched rows
df4 = df[df['Order ID'].str.match('AZ-2011-[0-9]{7}') == False]
df4
```

Out[10]:

	Order ID	Product Name	Discount	Sales	Profit	Quantity	Category	Sub-Category
1	AZ-2011-107716	Deflect-O Door Stop, Erganomic	0.0	85	15	2	Furniture	Furnishings
9	AZ-2011-122598	Avery Removable Labels, Alphabetical	0.0	32	6	3	Office Supplies	Labels
17	AZ-2011-130330	Office Star Chairmat, Adjustable	0.1	307	99	5	Furniture	Chairs
31	AZ-2011-144325	Bush Stackable Bookrack, Pine	0.0	630	132	5	Furniture	Bookcases
34	AZ-2011-145488	Rogers File Cart, Industrial	0.4	255	-98	3	Office Supplies	Storage
58	AZ-2011-176674	Hoover Microwave, Red	0.1	1667	185	6	Office Supplies	Appliances
67	AZ-2011-201891	Cameo Clasp Envelope, with clear poly window	0.0	52	19	4	Office Supplies	Envelopes

Question 5: Find the top 25 orders based on sales value and find the number of orders which belong to furniture category.

In [11]:

```
# top 25 orders based on sales value
df5 = df.nlargest(25,['Sales'])
df5
```

Out[11]:

	Order ID	Product Name	Discount	Sales	Profit	Quantity	Category	Sub-Category
30	AZ-2011-1410648	Nokia Smart Phone, Full Size	0.0	1908	820	3	Technology	Phones
58	AZ-2011-176674	Hoover Microwave, Red	0.1	1667	185	6	Office Supplies	Appliances
8	AZ-2011-1174243	Nokia Audio Dock, with Caller ID	0.0	1334	200	8	Technology	Phones
20	AZ-2011-1322840	Motorola Headset, with Caller ID	0.0	957	316	12	Technology	Phones
3	AZ-2011-1372644	Panasonic Printer, Durable	0.0	800	168	3	Technology	Machines
18	AZ-2011-1406494	Fellowes Lockers, Industrial	0.1	748	283	4	Office Supplies	Storage
39	AZ-2011-1536006	Logitech Keyboard, Programmable	0.0	666	66	9	Technology	Accessories
33	AZ-2011-1445262	Apple Smart Phone, Cordless	0.0	636	140	1	Technology	Phones
31	AZ-2011-144325	Bush Stackable Bookrack, Pine	0.0	630	132	5	Furniture	Bookcases
14	AZ-2011-1260928	Eldon File Cart, Single Width	0.1	576	51	5	Office Supplies	Storage
12	AZ-2011-1253407	Safco Stackable Bookrack, Pine	0.1	541	156	4	Furniture	Bookcases
48	AZ-2011-1672552	Binney & Smith Sketch Pad, Blue	0.0	510	132	11	Office Supplies	Art
41	AZ-2011-1584049	Brother Ink, Laser	0.0	442	0	3	Technology	Copiers
52	AZ-2011-1722024	Cisco Audio Dock, VoIP	0.0	364	142	2	Technology	Phones
59	AZ-2011-1902971	Wilson Jones Binding Machine, Clear	0.0	339	102	7	Office Supplies	Binders
17	AZ-2011-130330	Office Star Chairmat, Adjustable	0.1	307	99	5	Furniture	Chairs
2	AZ-2011-1087704	Belkin Flash Drive, Bluetooth	0.0	294	109	7	Technology	Accessories
60	AZ-2011-1916360	Dania 3-Shelf Cabinet, Mobile	0.0	288	20	2	Furniture	Bookcases
11	AZ-2011-1240916	Boston Canvas, Water Color	0.0	284	43	5	Office Supplies	Art
34	AZ-2011-145488	Rogers File Cart, Industrial	0.4	255	-98	3	Office Supplies	Storage
6	AZ-2011-1116129	Avery Binding Machine, Durable	0.0	252	15	5	Office Supplies	Binders
42	AZ-2011-	Tenex File Cart, Industrial	0.1	241	24	2	Office	Storage

	1584987						Supplies	
38	AZ-2011-1499597	Boston Markers, Fluorescent	0.0	193	29	7	Office Supplies	Art
46	AZ-2011-1655349	Fiskars Trimmer, Easy Grip	0.0	176	65	4	Office Supplies	Supplies
43	AZ-2011-1589827	Novimex Steel Folding Chair, Red	0.6	164	-70	5	Furniture	Chairs

In [12]:

```
# number of orders which belong to furniture category
len(df[df['Category']=='Furniture'])
```

Out[12]:

```
11
```

Question 6: Among the orders with sales>250 and profit>50, find the product name of the fourth highest order based on sales value.

In [13]:

```
# filter in rows with sales>250 and profit>50
df6 =df[df.apply(lambda x: x['Sales']>250 and x['Profit']>50, axis=1)]
```

In [14]:

```
# order data based on highest sales value
df6 = df6.sort_values('Sales', ascending=False)
```

In [15]:

```
# get product name of the fourth highest order
df6.iloc[3]['Product Name']
```

Out[15]:

```
'Motorola Headset, with Caller ID'
```

Question 7: Remove the orders with negative profit by dropping the corresponding rows with negative `Profit`. Find the product that makes the lowest profit per Quantity in the Technology category.

In [16]:

```
# filter out orders with negative profit
df7 = df[df.apply(lambda x: x['Profit']>=0, axis=1)]
len(df7)
```

Out[16]:

```
53
```

In [17]:

```
# fitler in orders in 'Technology' category
df7=df7[df7['Category']=='Technology']
df7
```

Out[17]:

	Order ID	Product Name	Discount	Sales	Profit	Quantity	Category	Sub-Category
2	AZ-2011-1087704	Belkin Flash Drive, Bluetooth	0.0	294	109	7	Technology	Accessories
3	AZ-2011-1372644	Panasonic Printer, Durable	0.0	800	168	3	Technology	Machines
8	AZ-2011-1174243	Nokia Audio Dock, with Caller ID	0.0	1334	200	8	Technology	Phones
20	AZ-2011-1322840	Motorola Headset, with Caller ID	0.0	957	316	12	Technology	Phones
30	AZ-2011-1410648	Nokia Smart Phone, Full Size	0.0	1908	820	3	Technology	Phones
33	AZ-2011-	Apple Smart Phone, Cordless	0.0	636	140	1	Technology	Phones

	1445262							
39	AZ-2011-1536006	Logitech Keyboard, Programmable	0.0	666	66	9	Technology	Accessories
41	AZ-2011-1584049	Brother Ink, Laser	0.0	442	0	3	Technology	Copiers
52	AZ-2011-1722024	Cisco Audio Dock, VoIP	0.0	364	142	2	Technology	Phones
63	AZ-2011-1967754	Logitech Numeric Keypad, USB	0.0	93	40	2	Technology	Accessories

In [18]:

```
# product that makes lowest profit per quantity
df7['ratio']= df7.apply(lambda x: x.Profit/x.Quantity,axis=1)
df7.nsmallest(1,'ratio')
```

Out[18]:

	Order ID	Product Name	Discount	Sales	Profit	Quantity	Category	Sub-Category	ratio
41	AZ-2011-1584049	Brother Ink, Laser	0.0	442	0	3	Technology	Copiers	0.0

<< End >>