DATA SCIENCE ANALYSIS

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data science is related to data mining, machine learning and big data.

REPORT

In this video we use Python (Jupyter Note Book) And Import Pandas , Matplotlib , Os and seaborn libraries to analyze and answer business questions about 4 years worth of sales data of super store. The data contains thousands of store purchases broken down by yearly and month year, product type, cost, purchase address, etc.

We have answered these 5 questions through our data analysis mainly using pandas and matplotlib library.

I. What is the overall sales trend?

II. Which are the Top 10 products by sales?

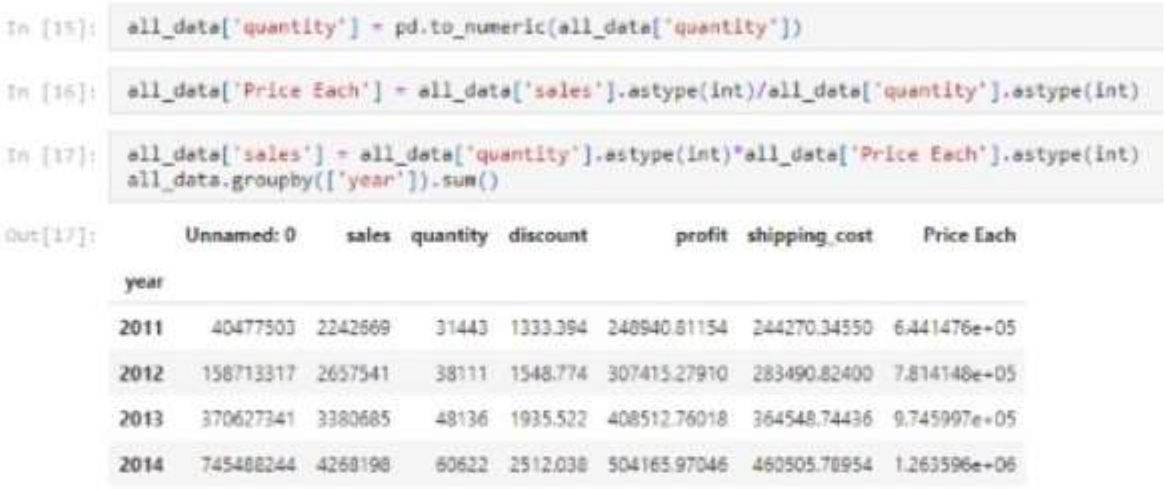
III. Which are the Most Selling Products?

IV. Which is the most preferred Ship Mode?

V. Which are the Most Profitable Category and Sub-Category?

VI. What products are most often sold together? and we can make packages easily?

i. What is the overall sales trend?

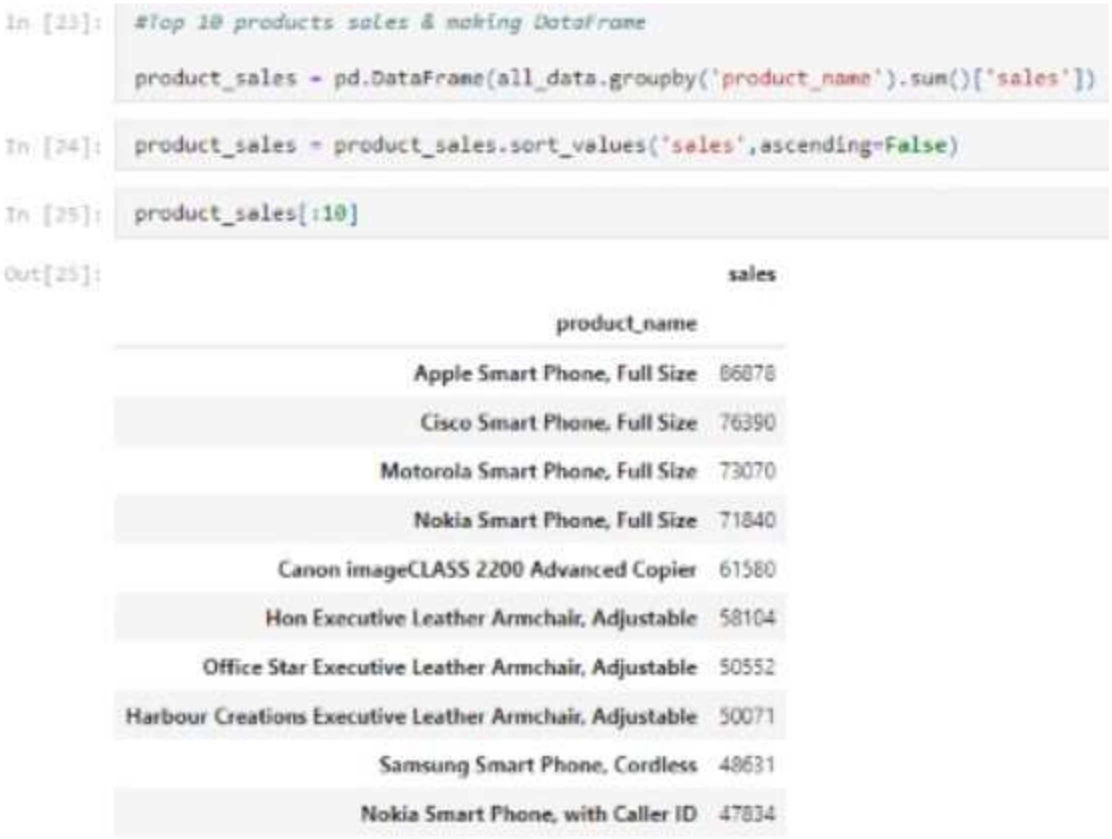


```
In [15]: all_data['quantity'] = pd.to_numeric(all_data['quantity'])

In [16]: all_data['Price Each'] = all_data['sales'].astype(int)/all_data['quantity'].astype(int)

In [17]: all_data['sales'] = all_data['quantity'].astype(int)*all_data['Price Each'].astype(int)
         all_data.groupby(['year']).sum()
```

Out[17]:

	Unnamed: 0	sales	quantity	discount	profit	shipping_cost	Price Each
year							
2011	40477503	2242669	31443	1333.394	248940.81154	244270.34550	6.441476e+05
2012	158713317	2657541	38111	1548.774	307415.27910	283490.82400	7.814148e+05
2013	370627341	3380685	48136	1935.522	408512.76018	364548.74436	9.745997e+05
2014	745488244	4268198	60622	2512.038	504165.97046	460505.78954	1.263596e+06

ii. Which are the Top 10 products by sales?



```
In [23]: #Top 10 products sales & making DataFrame

         product_sales = pd.DataFrame(all_data.groupby('product_name').sum()['sales'])
```

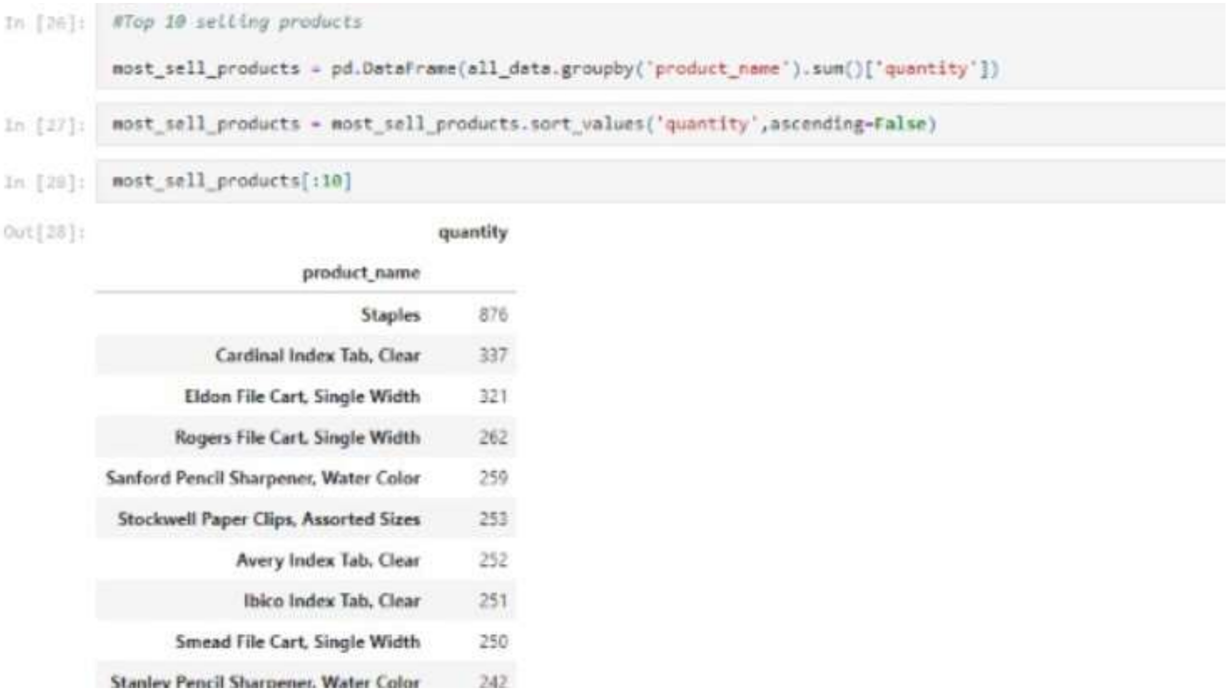
```
In [24]: product_sales = product_sales.sort_values('sales',ascending=False)
```

```
In [25]: product_sales[:10]
```

Out[25]:

	sales
product_name	
Apple Smart Phone, Full Size	86878
Cisco Smart Phone, Full Size	76390
Motorola Smart Phone, Full Size	73070
Nokia Smart Phone, Full Size	71840
Canon imageCLASS 2200 Advanced Copier	61580
Hon Executive Leather Armchair, Adjustable	58104
Office Star Executive Leather Armchair, Adjustable	50552
Harbour Creations Executive Leather Armchair, Adjustable	50071
Samsung Smart Phone, Cordless	48631
Nokia Smart Phone, with Caller ID	47834

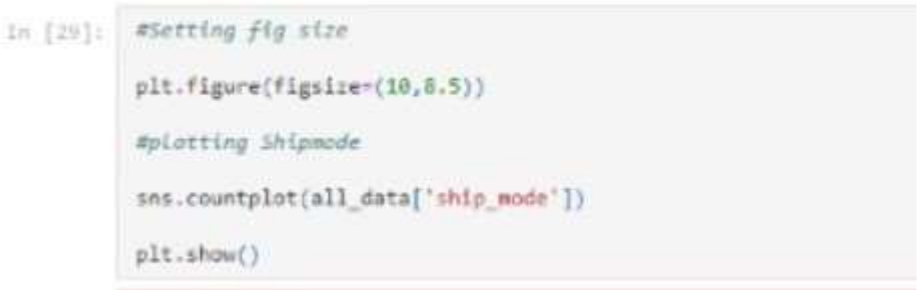
iii. Which are the Most Selling Products?



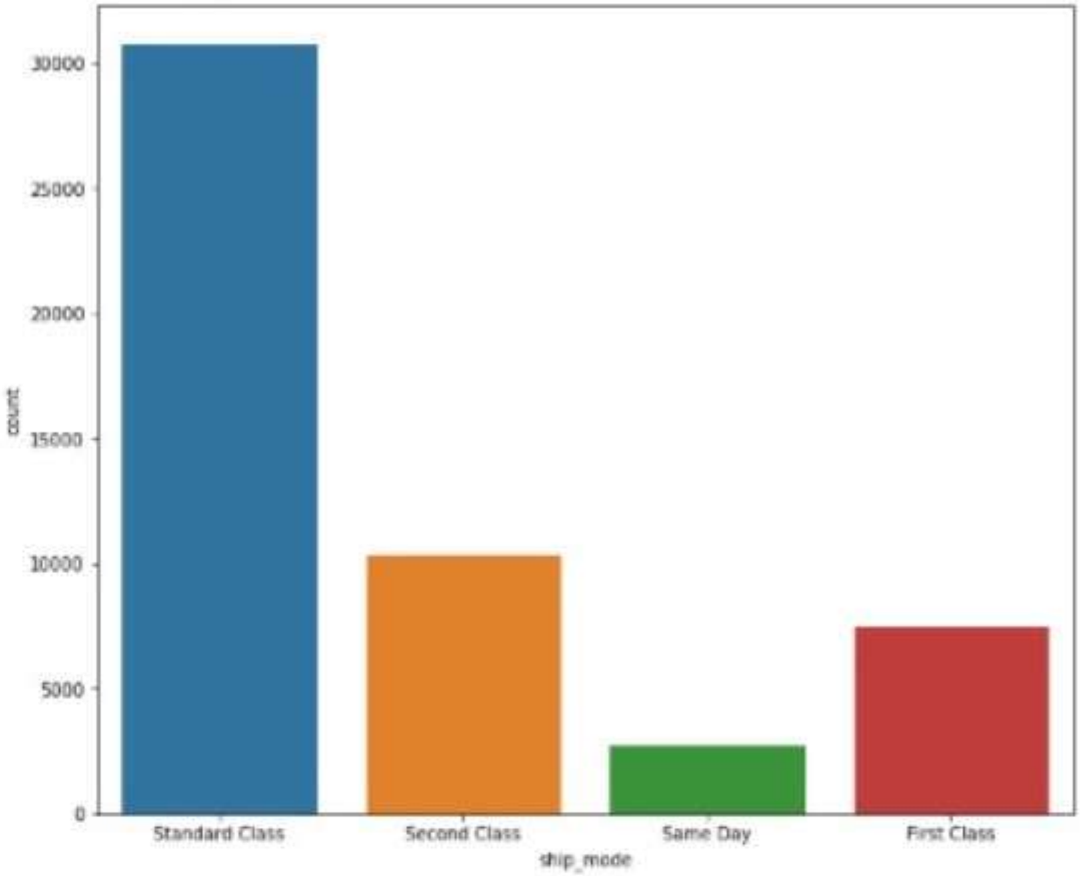
```
In [26]: #Top 10 selling products

         most_sell_products = pd.DataFrame(all_data.groupby('product_name').sum()['quantity'])

In [27]: most_sell_products = most_sell_products.sort_values('quantity',ascending=False)

In [28]: most_sell_products[:10]

Out[28]:
```

	quantity
product_name	
Staples	876
Cardinal Index Tab, Clear	337
Eldon File Cart, Single Width	321
Rogers File Cart, Single Width	262
Sanford Pencil Sharpener, Water Color	259
Stockwell Paper Clips, Assorted Sizes	253
Avery Index Tab, Clear	252
Ibico Index Tab, Clear	251
Smead File Cart, Single Width	250
Stanley Pencil Sharpener, Water Color	242

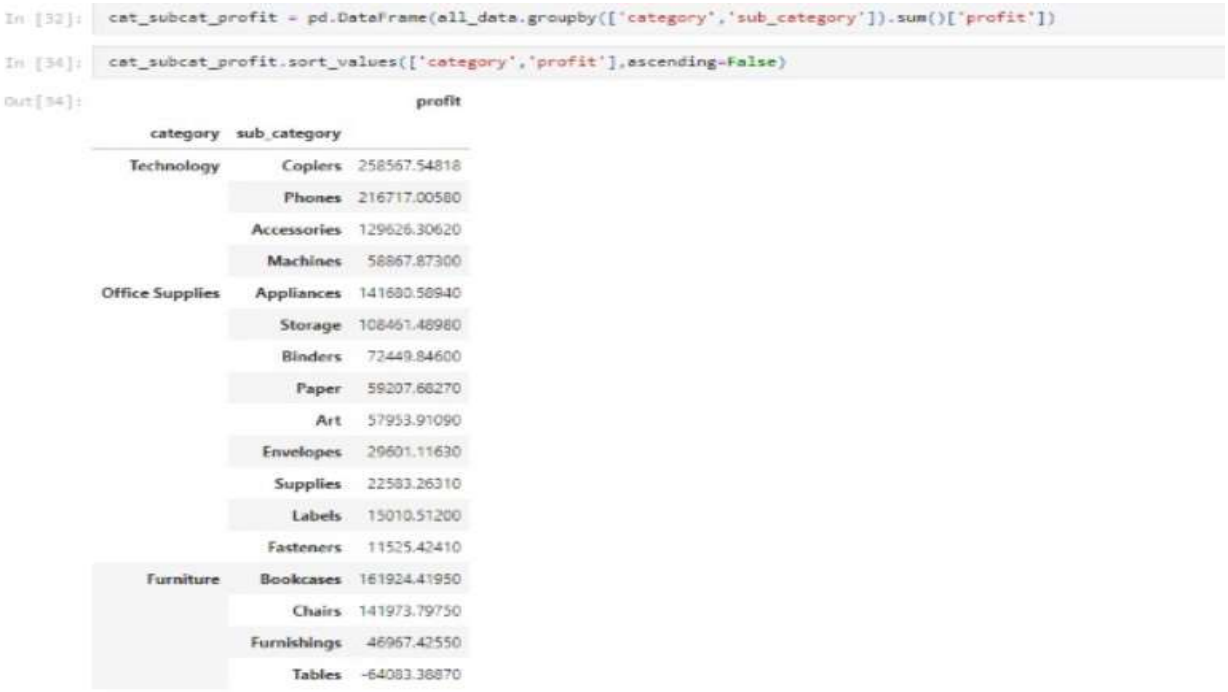
iv. Which is the most preferred Ship Mode?



```
In [29]:  #Setting fig size

           plt.figure(figsize=(10,8.5))

           #plotting Shipmode

           sns.countplot(all_data['ship_mode'])

           plt.show()
```



v. Which are the Most Profitable Category and Sub-Category?



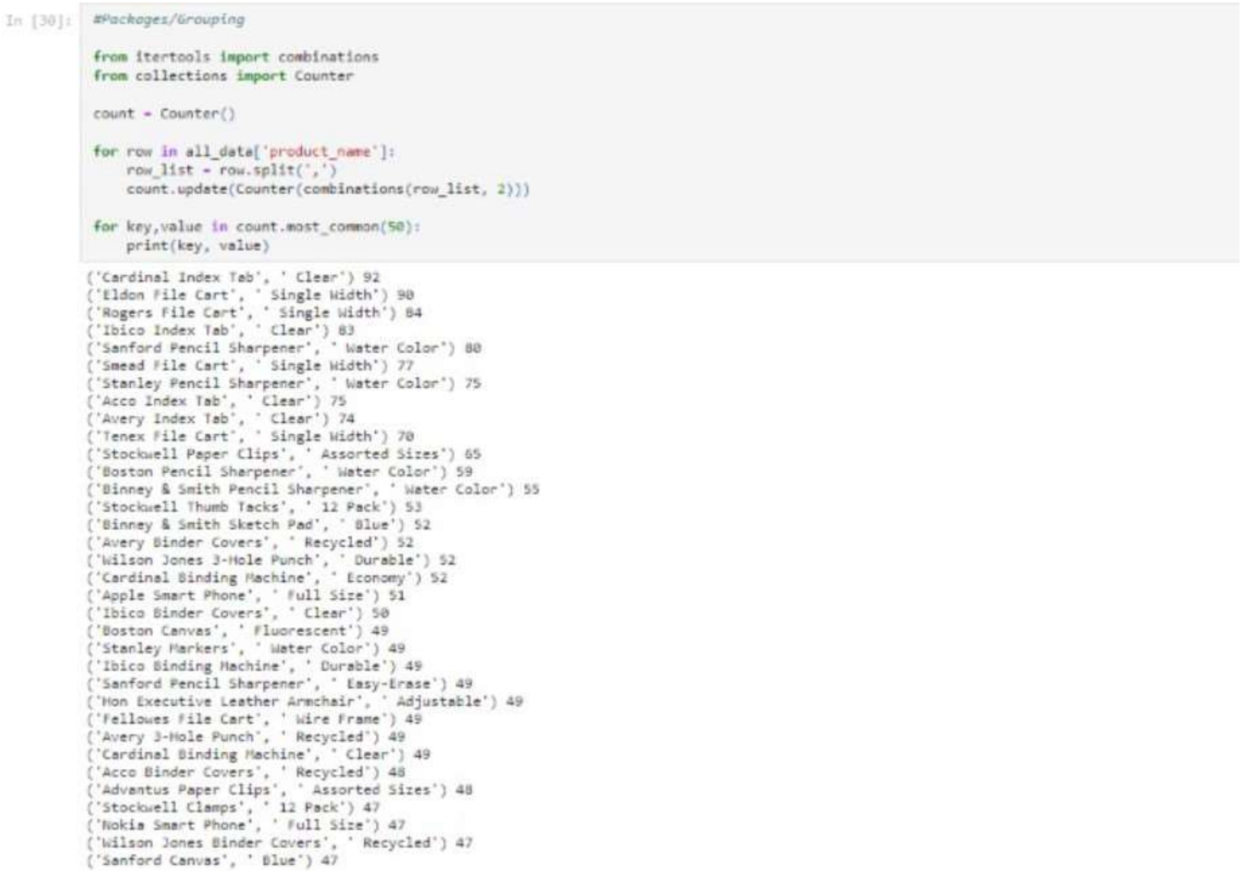
```
In [32]: cat_subcat_profit = pd.DataFrame(all_data.groupby(['category','sub_category']).sum()['profit'])
```

```
In [34]: cat_subcat_profit.sort_values(['category','profit'],ascending=False)
```

Out[34]:

category	sub_category	profit
Technology	Copiers	258567.54818
	Phones	216717.00580
	Accessories	129626.30620
	Machines	58867.87300
Office Supplies	Appliances	141680.58940
	Storage	108461.48980
	Binders	72449.84600
	Paper	59207.68270
	Art	57953.91090
	Envelopes	29601.11630
	Supplies	22583.26310
	Labels	15010.51200
	Fasteners	11525.42410
Furniture	Bookcases	161924.41950
	Chairs	141973.79750
	Furnishings	46967.42550
	Tables	-64083.38870

vi. What products are most often sold together? and we can make packages easily ?



In [30]:

```
#Packages/Grouping

from itertools import combinations
from collections import Counter

count = Counter()

for row in all_data['product_name']:
    row_list = row.split(',')
    count.update(Counter(combinations(row_list, 2)))

for key,value in count.most_common(50):
    print(key, value)
```

```
('Cardinal Index Tab', ' Clear') 92
('Eldon File Cart', ' Single Width') 90
('Rogers File Cart', ' Single Width') 84
('Ibico Index Tab', ' Clear') 83
('Sanford Pencil Sharpener', ' Water Color') 80
('Smead File Cart', ' Single Width') 77
('Stanley Pencil Sharpener', ' Water Color') 75
('Acco Index Tab', ' Clear') 75
('Avery Index Tab', ' Clear') 74
('Tenex File Cart', ' Single Width') 70
('Stockwell Paper Clips', ' Assorted Sizes') 65
('Boston Pencil Sharpener', ' Water Color') 59
('Binney & Smith Pencil Sharpener', ' Water Color') 55
('Stockwell Thumb Tacks', ' 12 Pack') 53
('Binney & Smith Sketch Pad', ' Blue') 52
('Avery Binder Covers', ' Recycled') 52
('Wilson Jones 3-Hole Punch', ' Durable') 52
('Cardinal Binding Machine', ' Economy') 52
('Apple Smart Phone', ' Full Size') 51
('Ibico Binder Covers', ' Clear') 50
('Boston Canvas', ' Fluorescent') 49
('Stanley Markers', ' Water Color') 49
('Ibico Binding Machine', ' Durable') 49
('Sanford Pencil Sharpener', ' Easy-Erase') 49
('Hon Executive Leather Armchair', ' Adjustable') 49
('Fellowes File Cart', ' Wire Frame') 49
('Avery 3-Hole Punch', ' Recycled') 49
('Cardinal Binding Machine', ' Clear') 49
('Acco Binder Covers', ' Recycled') 48
('Advantus Paper Clips', ' Assorted Sizes') 48
('Stockwell Clamps', ' 12 Pack') 47
('Nokia Smart Phone', ' Full Size') 47
('Wilson Jones Binder Covers', ' Recycled') 47
('Sanford Canvas', ' Blue') 47
```