BÀI THỰC HÀNH SỐ 6 (6 tiết)

Thư viện Numpy, Pandas và Matplotlib

I. Mục tiêu

- Sinh viên làm quen và sử dụng được thư viện Numpy, Pandas và Matplotlib

II. Hướng dẫn và bài tập

1. Thư viện Numpy (2 tiết)

a. Hướng dẫn

- Cài đặt Numpy: https://www.w3schools.com/python/numpy/numpy_getting_started.asp
- Sinh viên đọc phần cơ bản (Basic) ở trang https://www.w3schools.com/python/numpy/default.asp

b. Bài tập

- Sinh viên làm bài tập cơ bản ở trang https://www.w3schools.com/python/numpy/exercise.asp

2. Thư viện Pandas (2 tiết)

a. Hướng dẫn

- Cài đặt Pandas: https://www.w3schools.com/python/pandas/pandas_getting_started.asp
- Sinh viên học khóa học cơ bản về Pandas ở trang: https://www.w3schools.com/python/pandas/default.asp
- Sinh viên làm bài tập cơ bản ở trang: https://www.w3schools.com/python/pandas/exercise.asp

b. Bài tập 1

Trong bài tập này, chúng ta sử dụng tập dữ liệu ô tô để phân tích dữ liệu. Tập dữ liệu này có các đặc điểm khác nhau của ô tô như kiểu thân xe, cơ sở bánh xe, loại động cơ, giá cả, quãng đường đi được, mã lực, v.v.

- *Hiển thị dữ liệu đọc từ file*: có thể hiển thị mặc định hoặc tùy chỉnh số dòng dữ liệu cần xuất dựa vào phương thức head và tail như ví dụ sau đây

```python
import pandas as pd
df = pd.read_csv("D:\\Automobile_data.csv")

# xuất dữ liệu đọc từ tập tin "Automobile_data.csv"
# mặc định sẽ hiển thị 5 dòng đầu và 5 dòng cuối
print(df)

# xuất 6 dòng đầu tiên
print (df.head(6))

# xuất 7 dòng cuối cùng
print (df.tail(7))
```

Kết quả của câu lệnh ở dòng 6 "***print(df)***": mặc định 5 dòng đầu tiên và 5 dòng cuối cùng của tập dữ liệu

	index	company	body-style	wheel-base	length	engine-type	num-of-cylinders	horsepower	average-mileage	price
0	0	alfa-romero	convertible	88.6	168.8	dohc	four	111	21	13495.0
1	1	alfa-romero	convertible	88.6	168.8	dohc	four	111	21	16500.0
2	2	alfa-romero	hatchback	94.5	171.2	ohcv	six	154	19	16500.0
3	3	audi	sedan	99.8	176.6	ohc	four	102	24	13950.0
4	4	audi	sedan	99.4	176.6	ohc	five	115	18	17450.0
..	...	...	...	...	...	...	...	...	...	...
56	81	volkswagen	sedan	97.3	171.7	ohc	four	85	27	7975.0
57	82	volkswagen	sedan	97.3	171.7	ohc	four	52	37	7995.0
58	86	volkswagen	sedan	97.3	171.7	ohc	four	100	26	9995.0
59	87	volvo	sedan	104.3	188.8	ohc	four	114	23	12940.0
60	88	volvo	wagon	104.3	188.8	ohc	four	114	23	13415.0

Kết quả của dòng lệnh số 9 "***print(df.head(6))***): xuất 6 dòng đầu tiên của tập dữ liệu

	index	company	body-style	wheel-base	length	engine-type	num-of-cylinders	horsepower	average-mileage	price
0	0	alfa-romero	convertible	88.6	168.8	dohc	four	111	21	13495.0
1	1	alfa-romero	convertible	88.6	168.8	dohc	four	111	21	16500.0
2	2	alfa-romero	hatchback	94.5	171.2	ohcv	six	154	19	16500.0
3	3	audi	sedan	99.8	176.6	ohc	four	102	24	13950.0
4	4	audi	sedan	99.4	176.6	ohc	five	115	18	17450.0
5	5	audi	sedan	99.8	177.3	ohc	five	110	19	15250.0

Kết quả của dòng lệnh số 12 "***print(df.tail(7))***): xuất 7 dòng cuối của tập dữ liệu

	index	company	body-style	wheel-base	length	engine-type	num-of-cylinders	horsepower	average-mileage	price
54	79	toyota	wagon	104.5	187.8	dohc	six	156	19	15750.0
55	80	volkswagen	sedan	97.3	171.7	ohc	four	52	37	7775.0
56	81	volkswagen	sedan	97.3	171.7	ohc	four	85	27	7975.0
57	82	volkswagen	sedan	97.3	171.7	ohc	four	52	37	7995.0
58	86	volkswagen	sedan	97.3	171.7	ohc	four	100	26	9995.0
59	87	volvo	sedan	104.3	188.8	ohc	four	114	23	12940.0
60	88	volvo	wagon	104.3	188.8	ohc	four	114	23	13415.0

- *Hãy xuất thông tin của dữ liệu như sau:*

```
RangeIndex: 61 entries, 0 to 60
Data columns (total 11 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Unnamed: 0        61 non-null     int64
 1   index             61 non-null     int64
 2   company           61 non-null     object
 3   body-style        61 non-null     object
 4   wheel-base        61 non-null     float64
 5   length            61 non-null     float64
 6   engine-type       61 non-null     object
 7   num-of-cylinders  61 non-null     object
 8   horsepower        61 non-null     int64
 9   average-mileage   61 non-null     int64
 10  price             58 non-null     float64
dtypes: float64(3), int64(4), object(4)
memory usage: 5.4+ KB
```

- *Hãy thực hiện thanh lọc dữ liệu (clean) và cập nhật tập tin csv*
- *Xuất ra màn hình tên công ty có xe ô tô đắt nhất*

```
df = df [['company','price']][df.price==df['price'].max()]
print(df)
```

- *Xuất ra thông tin chi tiết của tất cả các xe thuộc hãng Toyota*

```
#df = df [['company','price']][df.price==df['price'].max()]
#print(df)

car_Manufacturers = df.groupby('company')
toyotaDf = car_Manufacturers.get_group("toyota")
print(toyotaDf)
```

- *Đếm số xe của từng hãng*

```
21    print(df['company'].value_counts())
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL JUPYTE

```
toyota           7
bmw              6
mazda            5
nissan           5
audi             4
mercedes-benz    4
mitsubishi       4
volkswagen       4
alfa-romero      3
chevrolet        3
honda            3
isuzu            3
jaguar           3
porsche          3
dodge            2
volvo            2
```

- *Hãy hiển thị giá xe cao nhất của mỗi hãng xe như sau*

```
                  company    price
company
alfa-romero      alfa-romero  16500.0
audi                   audi  18920.0
bmw                     bmw  41315.0
chevrolet         chevrolet   6575.0
dodge                dodge   6377.0
honda                honda  12945.0
isuzu                isuzu   6785.0
jaguar              jaguar  36000.0
mazda                mazda  18344.0
mercedes-benz  mercedes-benz  45400.0
mitsubishi       mitsubishi   8189.0
nissan               nissan  13499.0
porsche            porsche  37028.0
toyota              toyota  15750.0
volkswagen       volkswagen   9995.0
```

- *Hiển thị giá xe trung bình của mỗi hãng xe*

```
22  car_Manufacturers = df.groupby('company')
23  priceDf = car_Manufacturers[['company','price']].mean('price')
24  print(priceDf)
```

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  JUPYTER

mobile_dataset.py
                      price
company
alfa-romero    15498.333333
audi           16392.500000
bmw            27213.333333
chevrolet       6007.000000
dodge           6303.000000
honda          10195.000000
isuzu           6785.000000
jaguar         34600.000000
mazda           9654.800000
mercedes-benz  35040.000000
mitsubishi      6689.000000
nissan          8289.000000
porsche        35528.000000
toyota          8216.857143
volkswagen      8435.000000
volvo          13177.500000
```

c. *Bài tập 2*

Sử dụng pandas đọc tập tin dữ liệu sales_data.csv và thực hiện các yêu cầu sau:

1) Hiển thị thông tin của dữ liệu, xem và trả lời các câu hỏi sau:
   - Dữ liệu này có bao nhiêu cột, tên của các cột là gì?
   - Kiểu dữ liệu của các cột
   - Có bao nhiêu cột có chứa giá trị null
2) Hiển thị nội dung toàn bộ dữ liệu
3) Xuất hàng dữ liệu của tháng có lợi nhuận cao nhất như sau:

	month_number	facecream	facewash	toothpaste	bathingsoap	shampoo	moisturizer	total_units	total_profit
10	11	2340	2100	7300	13300	2400	2100	41280	412800

4) Xuất hàng dữ liệu của tháng bán nhiều mặt hàng nhất
5) Xuất hàng dữ liệu của tháng bán nhiều kem đánh răng nhất
6) Cho biết tổng lợi nhuận của cả năm
7) Cho biết tổng số lượng đã bán theo mặt hàng
8) Hiển thị số lượng các mặt hàng bán trong tháng 2
9) Số lượng mặt hàng bán chạy nhất tháng 2 (6100)
10) Tìm mặt hàng bán chạy nhất trong năm (*bathingsoap*)

3. Thư viện Matplotlib (2 tiết)

a. *Hướng dẫn*

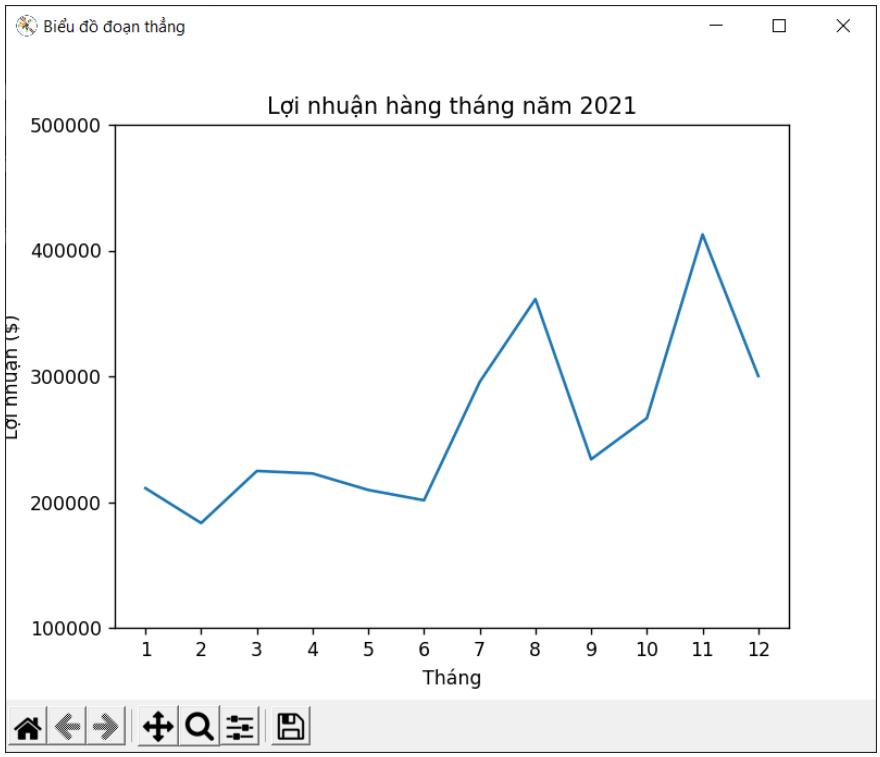
- Xem và làm các bài tập cơ bản của Matplotlib tại trang: https://www.w3schools.com/python/matplotlib_intro.asp

b. Bài tập

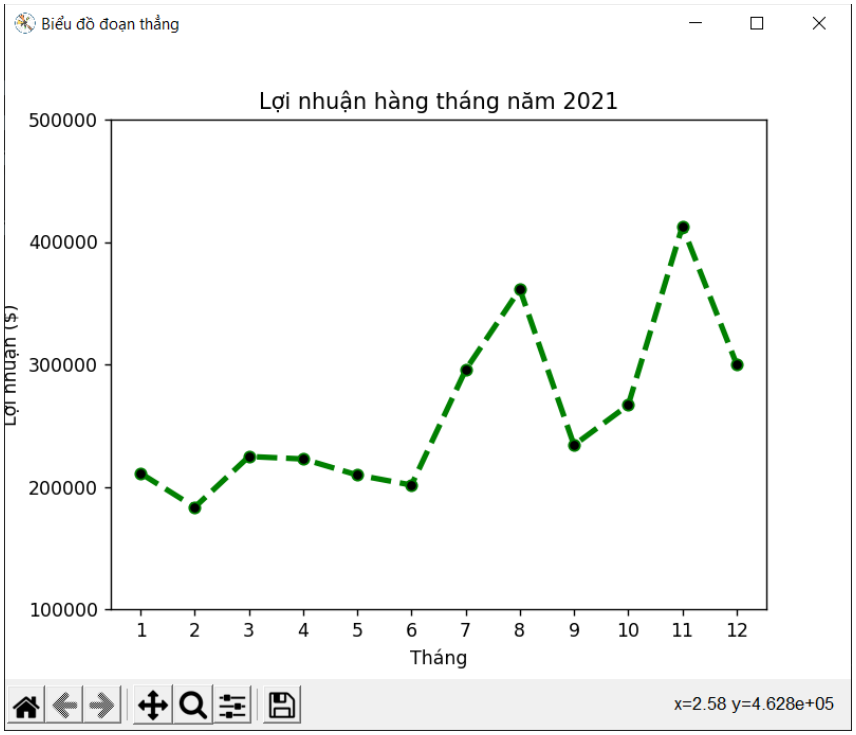
Bài tập này sử dụng dữ liệu trong file "***sales.csv***". Dữ liệu được đọc lên sử dụng Pandas

1) Đọc tổng lợi nhuận (total profit) của tất cả các tháng và hiển thị nó bằng cách sử dụng biểu đồ đường thẳng.

```python
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("D:\\sales_data.csv")
print (df.info())


profitList = df ['total_profit'].tolist()
monthList  = df ['month_number'].tolist()

plt.figure("Biểu đồ đoạn thẳng")
plt.plot(monthList, profitList)
plt.xlabel('Tháng')
plt.ylabel('Lợi nhuận ($)')
plt.xticks(monthList)
plt.title('Lợi nhuận hàng tháng năm 2021')
plt.yticks([100000, 200000, 300000, 400000, 500000])
plt.show()
```

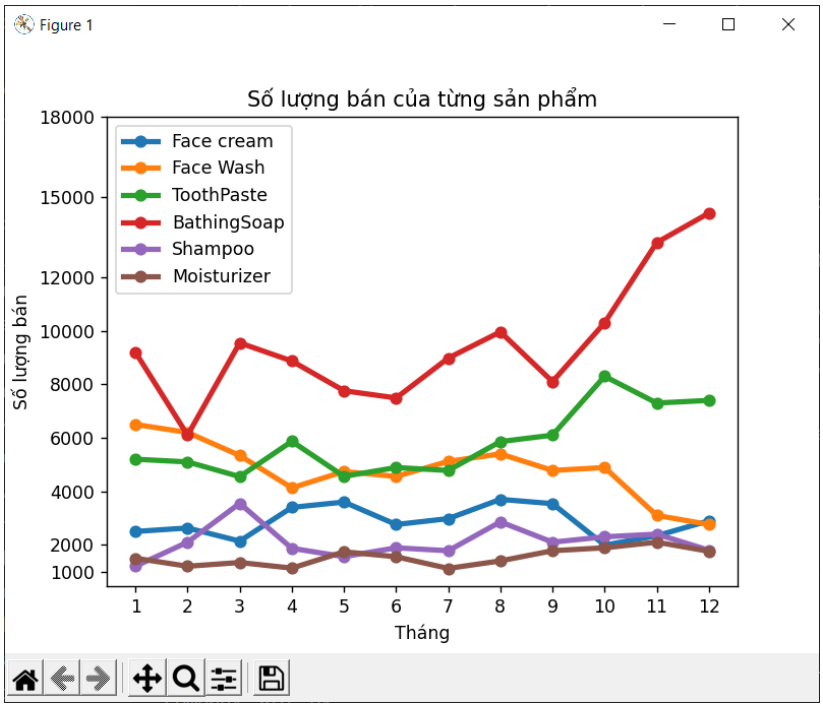


2) Chỉnh sửa biểu đồ ở câu trên thành dạng sau:

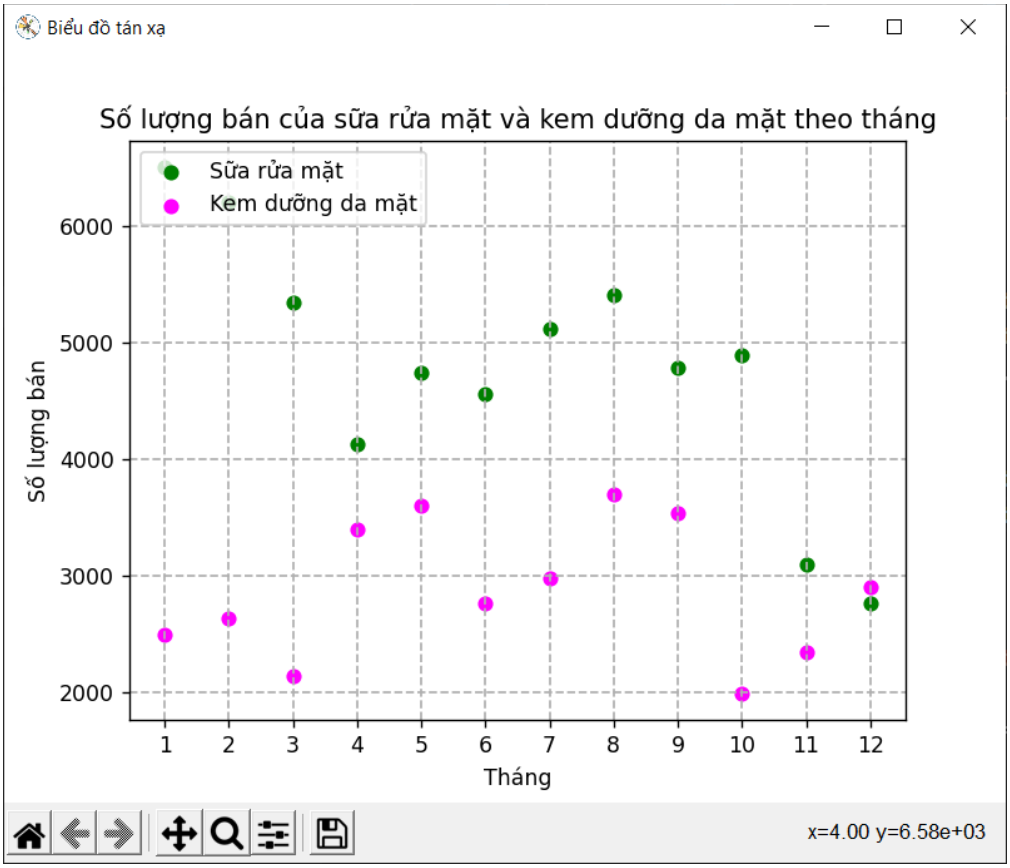


Nhìn vào biểu đồ, cho biết tháng có lợi nhuận cao nhất

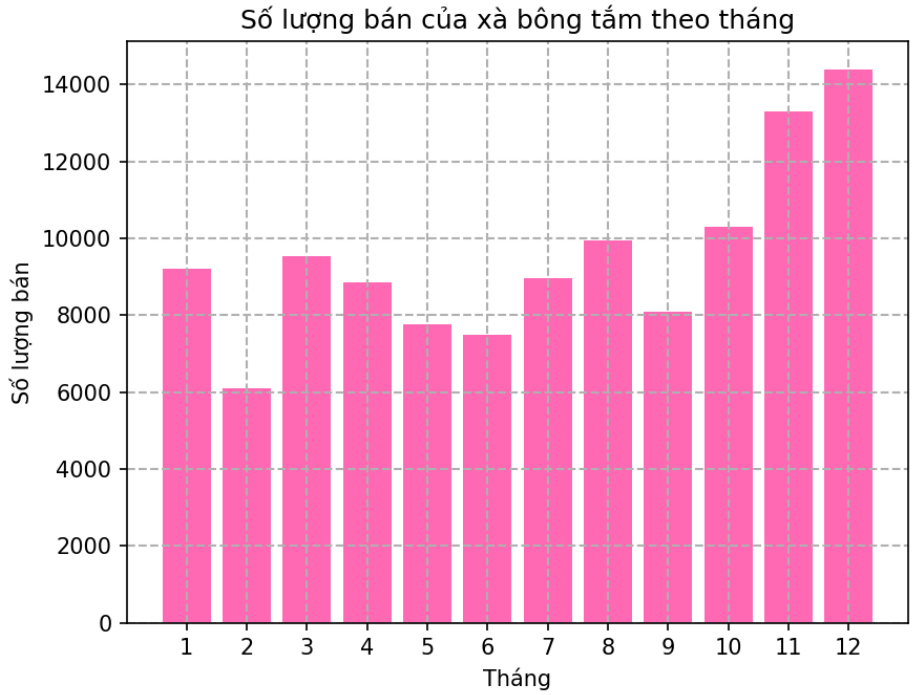
3) Vẽ biểu đồ đường thể hiện số lượng bán của từng mặt hàng trong năm như sau:



4) Hiển thị số lượng bán của mặt hàng sữa rửa mặt và kem dưỡng da mặt theo tháng bằng biểu đồ Điểm (biểu đồ tán xạ) như sau:

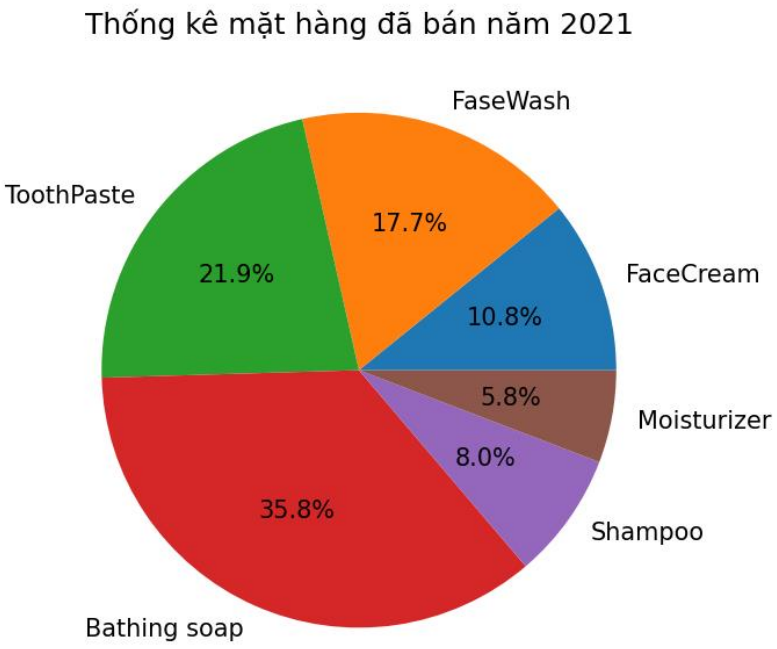


5) Vẽ biểu đồ cột thể hiện số lượng xà bông tắm đã bán như sau

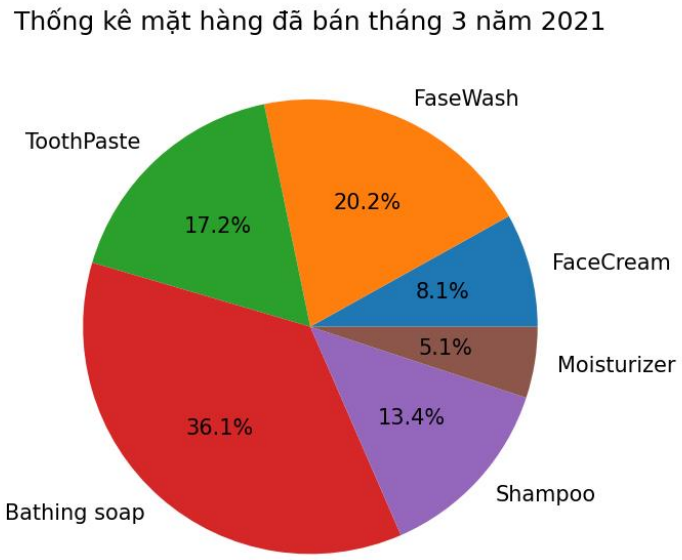


6) Hiển thị số lượng bán của mặt hàng sữa rửa mặt và kem dưỡng da mặt theo tháng bằng biểu đồ cột như sau:

7) Vẽ biểu đồ tròn thể hiện tỉ lệ sản phẩm bán trong năm như sau:



8) Vẽ biểu đồ tròn thể hiện tỉ lệ sản phẩm bán trong tháng 3/2021 như sau:



9) Vẽ dạng 2 hay nhiều biểu đồ con như sau: