

▼ Wholesale Dataset

ไฟล์ "Wholesale customers data.csv" (download จาก <https://archive.ics.uci.edu/ml/machine-learning-databases/00292/>) เป็นข้อมูลเกี่ยวกับการขายสินค้าประเภทต่างๆ ในแต่ละภูมิภาค ผ่าน 2 ช่องทาง คือ Horeca (Hotel/Restaurant/Cafe) หรือ Retail จาก UCI Machine Learning Repository โดยแถวแรกเป็นชื่อคอลัมน์ และ แถวที่ 2 เป็นต้นไปเก็บข้อมูล

ประเภทของสินค้า คือ fresh products, milk products, grocery products, forzen products, detergents and papers และ delicatessen

คอลัมน์ CHANNEL เป็นช่องทางการขาย ที่มีค่าเป็น Horeca (Hotel/Restaurant/Cafe) หรือ Retail (1,2) คอลัมน์ REGION: เป็นภูมิภาค ได้แก่ Lisnon, Oporto หรือ Other (1,2,3) คอลัมน์ FRESH, MILK, GROCERY, FROZEN, DETERGENTS_PAPER, DELICATESSEN เป็นยอดขายของสินค้าแต่ละประเภทในภูมิภาคนั้น ผ่านช่องทางนั้น ในหนึ่งปี

- 1) เขียน โปรแกรมเพื่อแปลงข้อมูลช่องทางการขายและภูมิภาค จากตัวเลขให้เป็นชื่อของแต่ละช่องทางการขายและภูมิภาคนั้น ๆ จากนั้นบันทึกเป็นไฟล์ใหม่ที่มีชื่อว่า Rename_Wholesale.csv โดยจะใช้ไฟล์นี้เป็นหลักในข้อถัดไป
- 2) เขียน โปรแกรมเพื่อหายอดขายรวมจากทุกช่องทาง ในทุกภูมิภาค สำหรับสินค้าแต่ละประเภทแล้ววาด pie chart
- 3) เขียน โปรแกรมเพื่อหายอดขายรวมของสินค้าทุกประเภท ในทุกช่องทาง สำหรับแต่ละภูมิภาคแล้ววาด pie chart
- 4) เขียน โปรแกรมเพื่อหายอดขายรวมของสินค้าสำหรับทำความสะอาดในแต่ละช่องทางการขาย โดยแบ่งเป็นของแต่ละภูมิภาค พร้อมพล็อตกราฟของแต่ละภูมิภาค
- 5) เขียน โปรแกรมเพื่อหาจำนวน record ที่ยอดขายของอาหารสดรวมกับอาหารแช่แข็งผ่านช่องทาง Horeca(Hotel/Restaurant/Cafe) ไม่เกิน 1,000 ขึ้น
- 6) เขียน โปรแกรมเพื่อหาภูมิภาคที่มีการบริโภคนมมากที่สุด และ ในภูมิภาคนั้นมีการบริโภคนมมากกว่าอาหารสดทั้งหมดที่ record

- 1) เขียน โปรแกรมเพื่อแปลงข้อมูลช่องทางการขายและภูมิภาค จากตัวเลขให้เป็นชื่อของแต่ละช่องทางการขาย และภูมิภาคนั้น ๆ จากนั้นบันทึกเป็นไฟล์ใหม่ที่มีชื่อว่า Rename_Wholesale.csv โดยจะใช้ไฟล์นี้เป็นหลักในข้อถัดไป
- ▼

```
1 from google.colab import drive
2 drive.mount('/content/drive')
```

Go to this URL in a browser: https://accounts.google.com/o/oauth2/auth?client_id=947318989803-6bn6๓k8๑d๑f4n๑3pfee6491hc0brc4i.apps.googleusercontent.com&redirect_uri=urn%3aietf%3awg%3aoauth%3a2.0%3aob&response_type=code&scope=email%20https%3a%2f%2fw

Enter your authorization code:
.....
Mounted at /content/drive

```
1 #Solution
2 import pandas as pd
```

```
1 d = pd.read_csv('/content/drive/My Drive/Python for Data science/Assignment1/Wholesale customers data.csv')
2 df = pd.DataFrame(d)
3 df
```

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	2	3	12669	9656	7561	214	2674	1338
1	2	3	7057	9810	9568	1762	3293	1776
2	2	3	6353	8808	7684	2405	3516	7844
3	1	3	13265	1196	4221	6404	507	1788
4	2	3	22615	5410	7198	3915	1777	5185
...
435	1	3	29703	12051	16027	13135	182	2204
436	1	3	39228	1431	764	4510	93	2346
437	2	3	14531	15488	30243	437	14841	1867
438	1	3	10290	1981	2232	1038	168	2125
439	1	3	2787	1698	2510	65	477	52

440 rows x 8 columns

```
1 #Replace ใน Channel แต่ละ Index
2 for i, channel in df.iterrows():
3     if (df.loc[i,"Channel"] == 1):
4         df.loc[i, "Channel"] = "Horeca"
5     else:
6         df.loc[i, "Channel"] = "Retail"
7 df
```

🔗

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	Retail	3	12669	9656	7561	214	2674	1338
1	Retail	3	7057	9810	9568	1762	3293	1776
2	Retail	3	6353	8808	7684	2405	3516	7844
3	Horeca	3	13265	1196	4221	6404	507	1788
4	Retail	3	22615	5410	7198	3915	1777	5185
...
435	Horeca	3	29703	12051	16027	13135	182	2204
436	Horeca	3	39228	1431	764	4510	93	2346
437	Retail	3	14531	15488	30243	437	14841	1867
438	Horeca	3	10290	1981	2232	1038	168	2125
439	Horeca	3	2787	1698	2510	65	477	52

440 rows x 8 columns

```
1 #Replace ใน Region แต่ละ Index
2 for i, region in df.iterrows():
3     if (df.loc[i, "Region"] == 1):
4         df.loc[i, "Region"] = "Lisnon"
5     elif (df.loc[i, "Region"] == 2):
6         df.loc[i, "Region"] = "Oporto"
7     else:
8         df.loc[i, "Region"] = "Other"
9 df
```

🔗

	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	Retail	Other	12669	9656	7561	214	2674	1338
1	Retail	Other	7057	9810	9568	1762	3293	1776
2	Retail	Other	6353	8808	7684	2405	3516	7844
3	Horeca	Other	13265	1196	4221	6404	507	1788
4	Retail	Other	22615	5410	7198	3915	1777	5185
...
435	Horeca	Other	29703	12051	16027	13135	182	2204
436	Horeca	Other	39228	1431	764	4510	93	2346
437	Retail	Other	14531	15488	30243	437	14841	1867
438	Horeca	Other	10290	1981	2232	1038	168	2125
439	Horeca	Other	2787	1698	2510	65	477	52

440 rows x 8 columns

```
1 #save file
2 df.to_csv("/content/drive/My Drive/Python for Data science/Assignment1/Rename_Wholesale.csv")
```

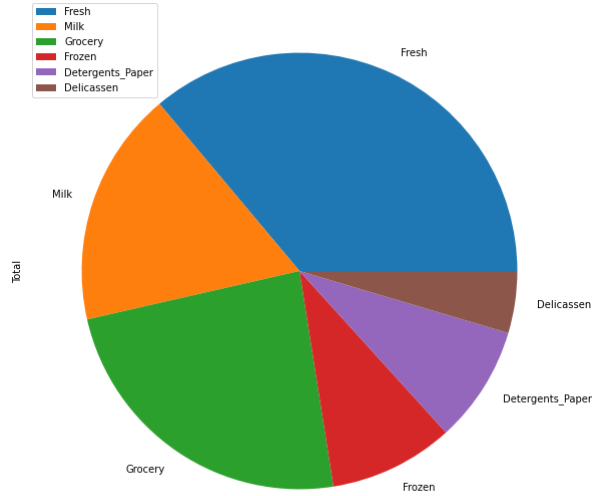
2) เขียนโปรแกรมเพื่อหายอดขายรวมจากทุกช่องทาง ในทุกภูมิภาค สำหรับสินค้าแต่ละประเภทแล้ววาด pie chart

```
1 #Solution
2 product = ["Fresh","Milk","Grocery","Frozen","Detergents_Paper","Delicassen"]
3 productplot = df[product].sum()
4 productplot = pd.DataFrame(productplot,columns=["Total"])
5 productplot
```



	Total
Fresh	5280131
Milk	2550357
Grocery	3498562
Frozen	1351650
Detergents_Paper	1267857
Delicassen	670943

```
1 plot = productplot.plot.pie(y="Total", figsize=(10, 10))
```



▼ 3) เขียนโปรแกรมเพื่อหายอดขายรวมของสินค้าทุกประเภท ในทุกช่องทาง สำหรับแต่ละภูมิภาคแล้ววาด pie chart

```
1 #Solution
2 totalproduct = df["Fresh"]+df["Milk"]+df["Grocery"]+df["Frozen"]+df["Detergents_Paper"]+df["Delicassen"]
3 df["Total_Product"] = totalproduct
4 df
```



	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	Total_Product
0	Retail	Other	12669	9656	7561	214	2674	1338	34112
1	Retail	Other	7057	9810	9568	1762	3293	1776	33266
2	Retail	Other	6353	8808	7684	2405	3516	7844	36610
3	Horeca	Other	13265	1196	4221	6404	507	1788	27381
4	Retail	Other	22615	5410	7198	3915	1777	5185	46100
...
435	Horeca	Other	29703	12051	16027	13135	182	2204	73302
436	Horeca	Other	39228	1431	764	4510	93	2346	48372
437	Retail	Other	14531	15488	30243	437	14841	1867	77407
438	Horeca	Other	10290	1981	2232	1038	168	2125	17834
439	Horeca	Other	2787	1698	2510	65	477	52	7589

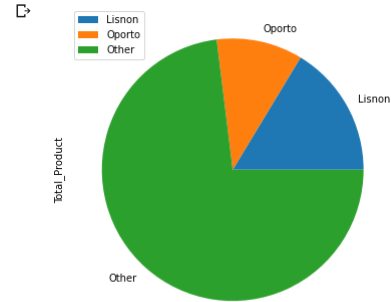
440 rows x 9 columns

```
1 regionplot = df.groupby("Region")[["Region", "Total_Product"]].sum()
2 regionplot
```



Total_Product	
Region	
Lisnon	2386813
Oporto	1555088
Other	10677599

```
1 plot = regionplot.plot.pie(y= "Total_Product", figsize=(6, 6))
```



▼ 4) เขียนโปรแกรมเพื่อหายอดขายรวมของสินค้าสำหรับทำความสะอาดในแต่ละช่องทางการขาย โดยแบ่งเป็นของแต่ละภูมิภาค พร้อมพล็อตกราฟของแต่ละภูมิภาค

```
1 #Solution
2 df2 = df.groupby(["Region", "Channel"])[["Region", "Channel", "Detergents_Paper"]].sum()
3 df2
```



Detergents_Paper		
Region	Channel	
Lisnon	Horeca	56081
	Retail	148055
Oporto	Horeca	13516
	Retail	159795
Other	Horeca	165990
	Retail	724420

```
1 df3 = df2.unstack(level=0)
2 df3
```

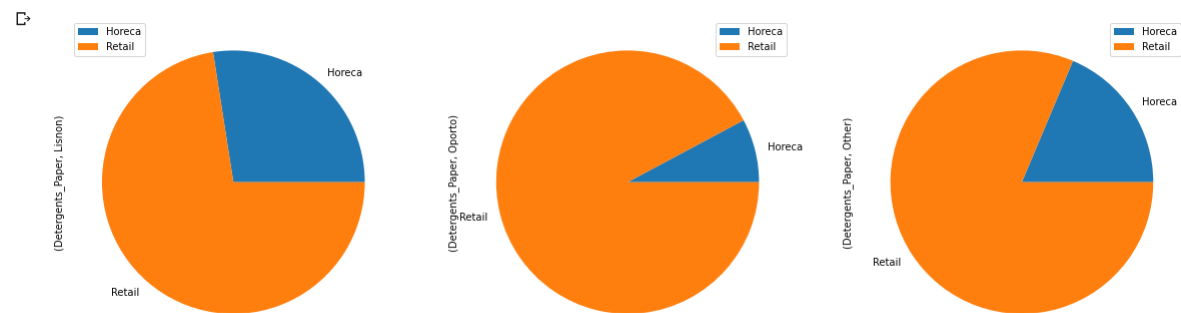


Detergents_Paper			
Region	Lisnon	Oporto	Other
Channel			
Horeca	56081	13516	165990
Retail	148055	159795	724420

```
1 df3.index
```

```
Copy Index(['Horeca', 'Retail'], dtype='object', name='Channel')
```

```
1 plot = df3.plot.pie(subplots=True, figsize=(20, 20))
```



5) เขียนโปรแกรมเพื่อหาจำนวนปีที่ยอดขายของอาหารสดรวมกับอาหารแช่แข็งผ่านช่องทาง Horeca(Hotel/Restaurant/Cafe) ไม่เกิน 1,000 ชิ้น

```
1 #Solution
2 food = df["Fresh"] + df["Frozen"]
3 df["Food"] = food
4 countyear = df[(df.Food <= 1000) & (df.Channel == "Horeca")][["Channel", "Food"]]
5 countyear
```

	Channel	Food
66	Horeca	184
95	Horeca	443
128	Horeca	282
154	Horeca	697
184	Horeca	401
299	Horeca	708
327	Horeca	956
342	Horeca	604
355	Horeca	435
412	Horeca	195

```
1 countyear.count()
```

```
Channel      10
Food         10
dtype: int64
```

6) เขียนโปรแกรมเพื่อหาภูมิภาคที่มีการบริโภคนมมากที่สุด และในภูมิภาคนั้นมีการบริโภคนมมากกว่าอาหารสดทั้งหมดกี่ปี

```
1 #Solution
2 df.groupby("Region")["Milk"].sum()
```

```
Region
Lisnon      422454
Oporto       239144
Other       1888759
Name: Milk, dtype: int64
```

```
1 diff = df["Milk"] - df["Fresh"]
2 df["Diff"] = diff
3 diffyear = df[(df.Region == "Other") & (df.Diff > 0)][["Region", "Diff"]]
4 diffyear
```

	Region	Diff
1	Other	2753
2	Other	2455
9	Other	5087
10	Other	2037
16	Other	7796
...
420	Other	810
428	Other	2923
430	Other	1133
433	Other	1236
437	Other	957

```
1 diffyear.count()
```

 Region 107
Diff 107
dtype: int64