

HW05 Mid-term

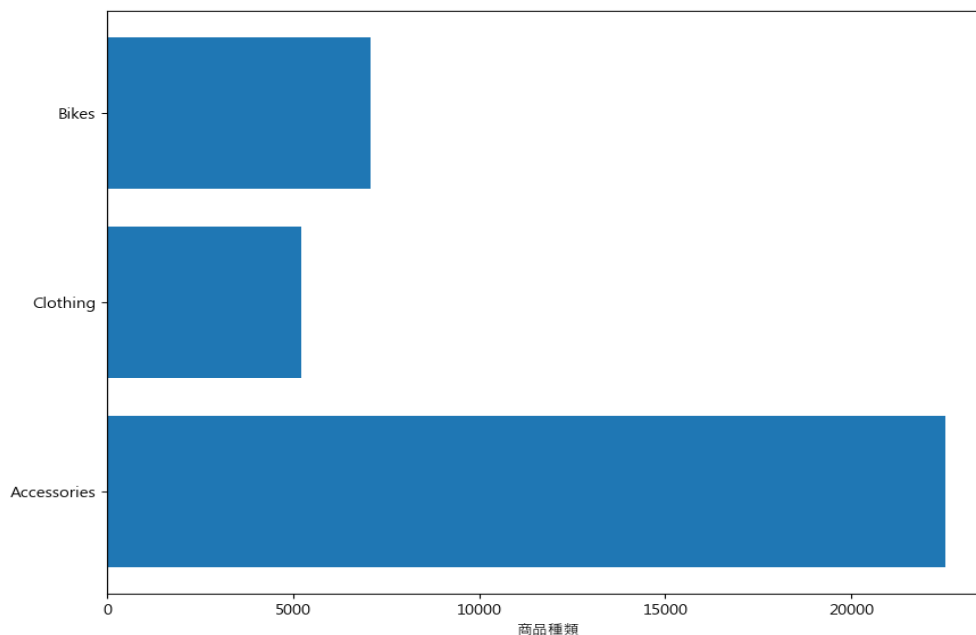
C109152304 許智程

請利用資料集 OnLine_Sale.xlsx 進行以下資料預處理及分析:

1. 新增 3 個欄位，建立每筆資料的銷售成本(單位成本 x 銷售數量)、銷售收入(單位售價 x 銷售數量)及收益(銷售收入 - 銷售成本)。
 - 注：結果以 csv 檔輸出，採 UTF-8 編碼，直接用 excel 開啟會導致亂碼，且檔案規模不宜直接放入報告，於是採用 excel 導入 csv 檔並截圖的方式呈現

1	日期	年	月	年紀	性別	國別	商品種類	商品次分類	銷售數量	單位成本	單位售價	回購	銷售成本	銷售收入	收益
2	2015/8/31	2015	August	17	M	France	Accessories	Bottles and Cages	3	6.67	6.66666667	Y	20.01	20	-0.01
3	2015/8/31	2015	August	17	M	France	Accessories	Bottles and Cages	3	63.33	50	Y	189.99	150	-39.99
4	2015/8/5	2015	August	17	M	Germany	Accessories	Bottles and Cages	3	66.67	95.66666667	Y	200.01	287	86.99
5	2015/8/5	2015	August	17	M	Germany	Accessories	Bottles and Cages	1	20	28	N	20	28	8
6	2015/8/18	2015	August	17	M	United States	Accessories	Bottles and Cages	1	125	149	N	125	149	24
7	2015/9/29	2015	September	17	M	United States	Accessories	Bottles and Cages	3	10	11	Y	30	33	3
8	2015/9/29	2015	September	17	M	United States	Accessories	Bottles and Cages	1	270	266	N	270	266	-4
9	2015/10/23	2015	October	17	M	United States	Accessories	Bottles and Cages	2	7.5	8	Y	15	16	1
10	2015/10/23	2015	October	17	M	United States	Accessories	Bottles and Cages	1	171	165	N	171	165	-6
11	2015/10/15	2015	October	17	M	United States	Accessories	Bottles and Cages	1	50	51	N	50	51	1
12	2015/10/26	2015	October	17	M	United States	Accessories	Bottles and Cages	2	40	43.5	N	80	87	7
13	2015/10/26	2015	October	17	M	United States	Accessories	Bottles and Cages	3	6.67	7.66666667	Y	20.01	23	2.99
14	2015/8/28	2015	August	17	M	United Kingdom	Accessories	Bottles and Cages	2	62.5	71.5	N	125	143	18
15	2015/10/4	2015	October	17	M	United Kingdom	Accessories	Bottles and Cages	3	42	46	Y	126	138	12
16	2015/10/4	2015	October	17	M	United Kingdom	Accessories	Bottles and Cages	3	41.67	47.66666667	Y	125.01	143	17.99
17	2015/10/22	2015	October	17	M	United Kingdom	Accessories	Bottles and Cages	3	60	71.66666667	Y	180	215	35
18	2015/10/22	2015	October	17	M	United Kingdom	Accessories	Bottles and Cages	3	5	6	Y	15	18	3
19	2015/10/27	2015	October	17	M	United Kingdom	Accessories	Bottles and Cages	3	66	71.66666667	Y	198	215	17
20	2015/10/27	2015	October	17	M	United Kingdom	Accessories	Bottles and Cages	3	41.67	48.33333333	Y	125.01	145	19.99
21	2015/12/18	2015	December	17	M	United Kingdom	Accessories	Bottles and Cages	1	170	199	N	170	199	29
22	2015/12/18	2015	December	17	M	United Kingdom	Accessories	Bottles and Cages	2	72.5	79.5	N	145	159	14
23	2015/11/13	2015	November	17	M	United States	Clothing	Caps	3	30	35.66666667	Y	90	107	17
24	2015/10/23	2015	October	17	M	United States	Clothing	Caps	3	33	35.66666667	Y	99	107	8
25	2015/12/25	2015	December	17	M	United Kingdom	Clothing	Caps	3	57	68.66666667	Y	171	206	35

2. 分析商品種類的分佈、特徵、及資料變異等視覺化分析



從上圖看到商品可以大略分成 Accessories、Clothing、Bikes 三大類，每筆交易各個類別的銷售成本、銷售收入、收益的基本統計量如下面的表格所示 (全部數字皆四捨五入到小數點第二位)。

	銷售成本	銷售收入	收益
平均數	268.13	329.33	61.2
中位數	130.0	159.0	25.0
眾數	60.0	6.0	3.0
全距	3598.0	5080.0	2096.99
四分位距	360.0	429.0	67.0
四分位差	180.0	214.5	33.5
平均絕對離差	235.57	291.12	63.49
變異數	90440.48	142287.53	9362.5
標準差	300.73	377.21	96.76
變異係數	1.12	1.15	1.58
P20	46.0	56.0	5.99
P25(Q1)	60.0	70.0	8.0
P50(Q2)	130.0	159.0	25.0
P75(Q3)	420.0	499.0	75.0
P85	623.01	747.0	131.0
偏態係數	1.38	1.35	1.12
峰度係數	5.95	8.43	22.66

Accessories 的銷售成本、銷售收入、收益的基本統計量

	銷售成本	銷售收入	收益
平均數	552.0	656.03	104.03
中位數	300.0	376.0	48.0
眾數	162.0	175.0	6.0
全距	2091.0	3331.0	1929.0
四分位距	753.0	876.0	130.75
四分位差	376.5	438.0	65.38
平均絕對離差	431.92	513.62	108.31
變異數	264114.67	380163.31	24377.73
標準差	513.92	616.57	156.13
變異係數	0.93	0.94	1.5
P20	117.0	142.0	8.0
P25(Q1)	147.0	173.0	13.0
P50(Q2)	300.0	376.0	48.0
P75(Q3)	900.0	1049.0	143.75
P85	1200.0	1409.0	227.0
偏態係數	1.47	1.36	1.08
峰度係數	-0.1	0.25	7.38

Clothing 的銷售成本、銷售收入、收益的基本統計量

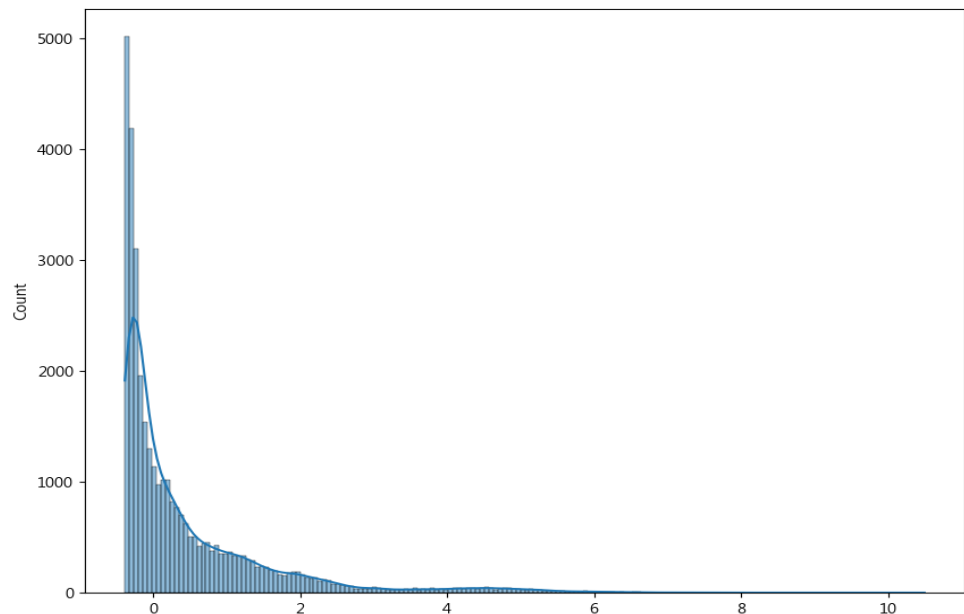
	銷售成本	銷售收入	收益
平均數	1572.04	1619.67	47.63
中位數	1701.0	1732.0	15.0
眾數	540.0	477.0	15.0
全距	1903.0	3346.0	2234.0
四分位距	1550.99	1542.75	245.0
四分位差	775.5	771.38	122.5
平均絕對離差	712.17	741.64	183.7
變異數	567929.19	668365.41	65706.71
標準差	753.61	817.54	256.33
變異係數	0.48	0.5	5.38
P20	742.0	706.0	-120.0
P25(Q1)	769.0	777.0	-86.0
P50(Q2)	1701.0	1732.0	15.0
P75(Q3)	2319.99	2319.75	159.0
P85	2384.0	2522.0	274.0
偏態係數	-0.51	-0.41	0.38
峰度係數	-1.71	-1.37	2.07

Bikes 的銷售成本、銷售收入、收益的基本統計量

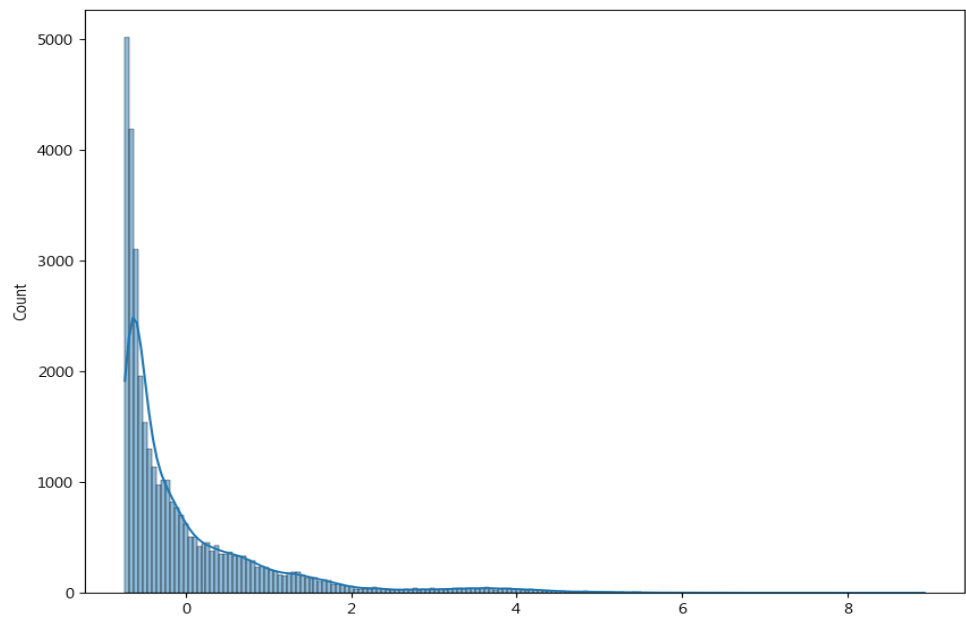
3. 針對”單位售價”與”收益”此兩個欄位，執行下列任意兩種正規化的方法：RobustScaler()、StandardScaler()、MinMaxScaler()、MaxAbsScaler()、Normalizer()。

1. 單位售價

- RobustScaler

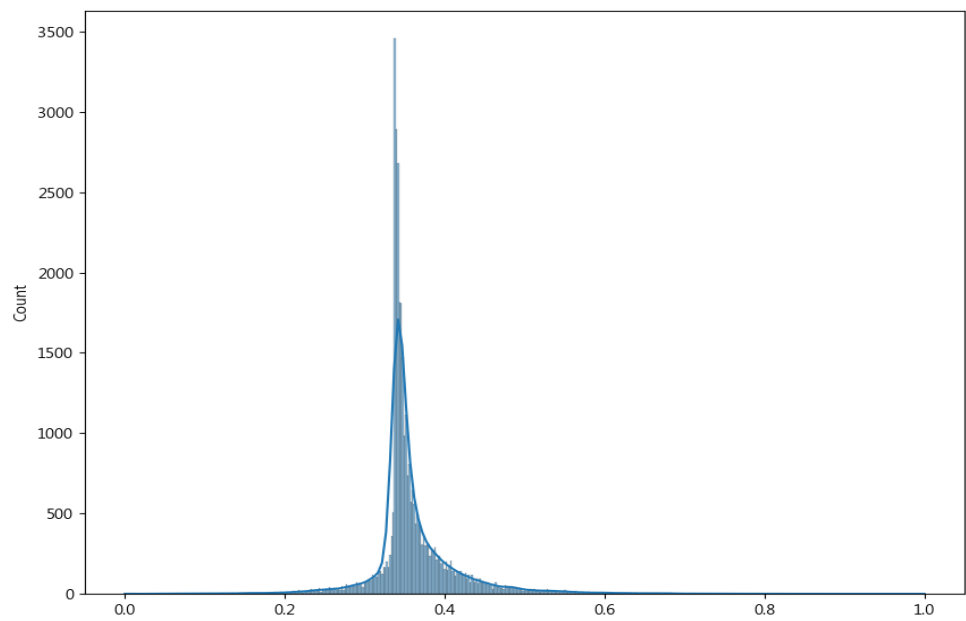


- StandardScaler

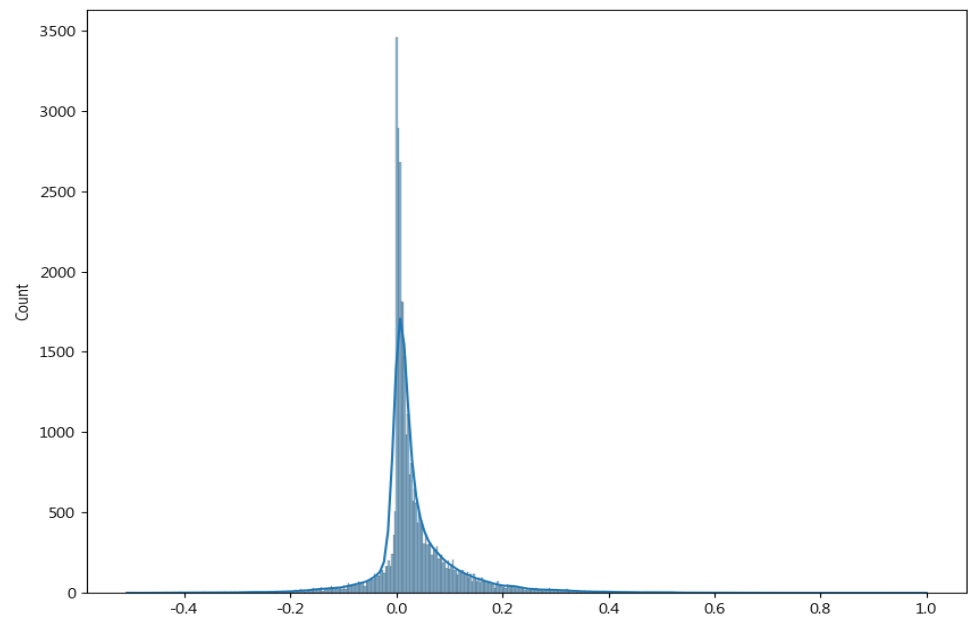


2. 收益

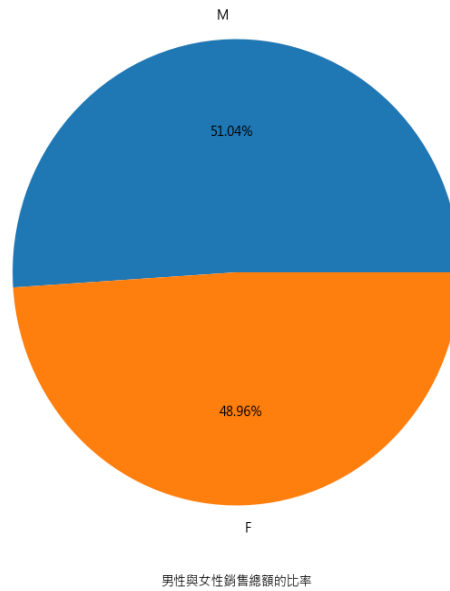
- MinMaxScaler



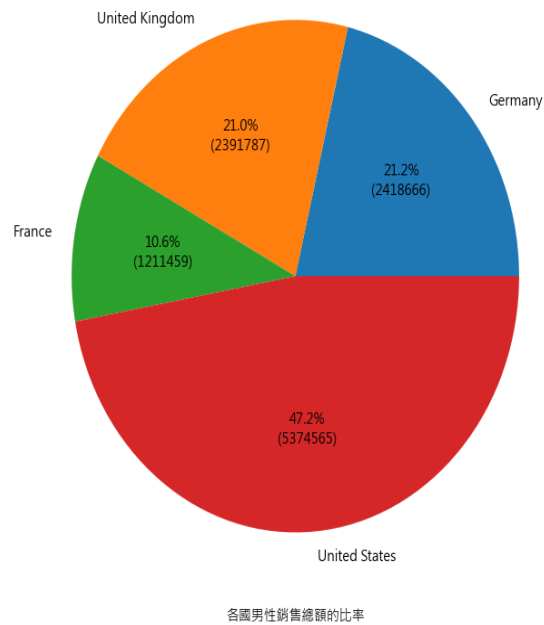
- MaxAbsScaler



4. 針對 a~d，採用圓餅圖呈現
- a. 男性與女性銷售總額的比率為何？

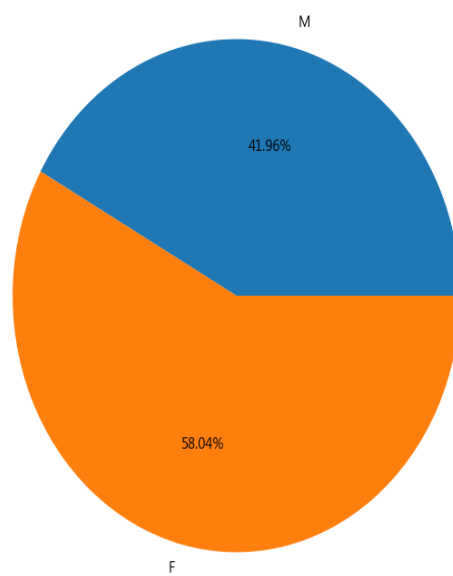


- b. 哪一個國家的男性總銷售金額最高？



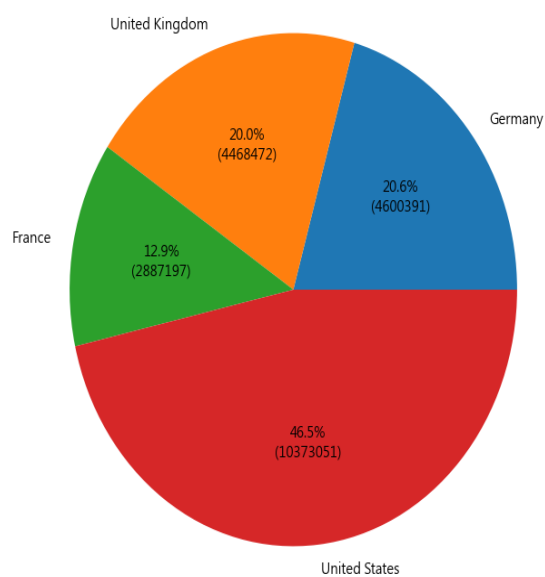
從上圖得知，United State(美國) 的男性的銷售金額最高，有 5374565。

c. 法國男性與女性銷售總額的比率為何？



法國男性與女性銷售總額的比率

d. 哪一個國家的總銷售金額最高？



各國銷售總額的比率

從上圖得知，United State(美國) 的銷售金額最高，有 10373051。

5. 根據顧客是否會回購，使用 DecisionTreeClassifier 建立預測分類器，取得其預測的結果(評估)，在此以 ROC(Receiver Operating Characteristics) 呈現。

