# Superstore聚類分析

資訊碩一 鄭智謙
電子碩一 陳立穎
電資四 莊東翰

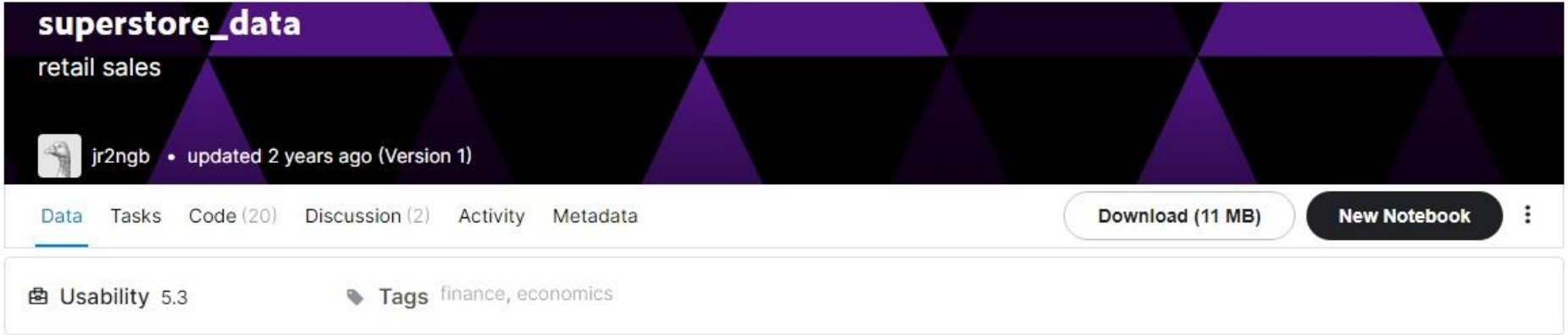
2021/06/15

# 目錄

# 一、資料集來源



- 資料集名稱：superstore_dataset2011-2015
- 來源：https://www.kaggle.com/jr2ngb/superstore-data
- 最近更新日期：2019/01/30



# 一、資料集介紹



- 簡介： superstore的零售數據，時間為2011/01/01 ~ 2014/12/31
- 特徵欄位(共24項)：

  Row.ID, Order.ID, Order.Date, Ship.Date, Ship.Mode, Customer.ID, Customer.Name, Segment, City, State, Country, Postal.Code, Market, Region, Product.ID, Category, Sub.Category, Product.Name, Sales, Quantity, Discount, Profit, Shipping.Cost, Order.Priority
- 原始資料量：共51290筆

# 一、資料集介紹



| Row ID | Order ID | Order Dat | Ship Date | Ship Mode | Customer | Customer | Segment | Category | Product N | Sales | Quantity | Discount | Shipping Cost | Order Priority |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 42433 | AG-2011- | 1-1-2011 | 6-1-2011 | Standard C | TB-11280 | Toby Brau | Consumer | Office Supplies | Tenex Loc | 408.3 | 2 | 0 | 35.46 | Medium |
| 22253 | IN-2011-4 | 1-1-2011 | 8-1-2011 | Standard C | JH-15985 | Joseph Ho | Consumer | Office Supplies | Acme Trir | 120.366 | 3 | 0.1 | 9.72 | Medium |
| 48883 | HU-2011- | 1-1-2011 | 5-1-2011 | Second Cl | AT-735 | Annie Thu | Consumer | Office Supplies | Tenex Box | 66.12 | 4 | 0 | 8.17 | High |
| 11731 | IT-2011-3 | 1-1-2011 | 5-1-2011 | Second Cl | EM-14140 | Eugene M | Home Off | Office Supplies | Enermax N | 44.865 | 3 | 0.5 | 4.82 | High |
| 22255 | IN-2011-4 | 1-1-2011 | 8-1-2011 | Standard C | JH-15985 | Joseph Ho | Consumer | Furniture | Eldon Ligh | 113.67 | 5 | 0.1 | 4.7 | Medium |
| 22254 | IN-2011-4 | 1-1-2011 | 8-1-2011 | Standard C | JH-15985 | Joseph Ho | Consumer | Office Supplies | Eaton Cor | 55.242 | 2 | 0.1 | 1.8 | Medium |
| 21613 | IN-2011-3 | 1-2-2011 | 3-2-2011 | Second Cl | PO-18865 | Patrick O' | Consumer | Technology | Brother Pe | 285.78 | 2 | 0 | 57.3 | Critical |

用k-means計算使用到的欄位:

Sales(售價), Quantity (數量), Shipping Cost (運費)

Category(分類): 內含Furniture(家具), Office Supplies(辦公室用品), Technology(3C產品) ，共3類。

```
> unique(Y$Category)
[1] Furniture       Office Supplies Technology
Levels: Furniture Office Supplies Technology
```

# 二、數據轉換

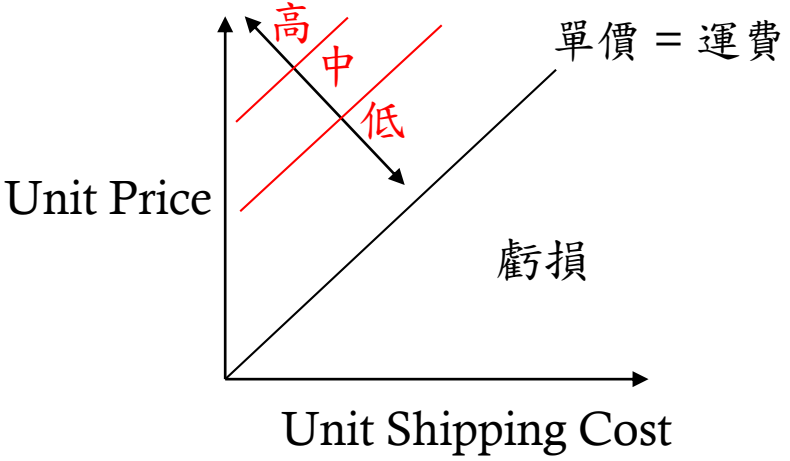


定義:

- Unit Price(單價) = Sales(售價) / Quantity (數量)
- Unit Shipping Cost (單件運費) = Shipping Cost (運費) / Quantity (數量)

預期:

電商經營者角度: 獲利




高單價, 低單件運費: 高獲利

低單價, 高單件運費: 低獲利

# 二、數據轉換



為了讓訓練出來的數據可以與之前比較(plot 函數)，必須先依照Category做排序，知道每種Category的邊界位置，事先紀錄好範圍。

Furniture: 家具, 1至9876

Office Supplies: 辦公室用品, 9877至41149

Technology: 3C產品, 41152至51290

| Unit_price | Unit_shippingCost | Category | row_num |
|---|---|---|---|
| 22.7340 | 0.9400000 | Furniture | 1 |
| 145.3330 | 27.3200000 | Furniture | 2 |
| 51.6000 | 3.2750000 | Furniture | 3 |
| 438.1500 | 50.1333333 | Furniture | 4 |
| 105.6860 | 11.6750000 | Furniture | 5 |
| 61.6200 | 16.4525000 | Furniture | 6 |
| 140.9100 | 12.8800000 | Furniture | 7 |
| 228.7840 | 23.9450000 | Furniture | 8 |
| 125.5030 | 10.9000000 | Furniture | 9 |
| 60.3750 | 4.2033333 | Furniture | 10 |

| | | | |
|---|---|---|---|
| 43.4400 | 3.43000000 | Furniture | 9874 |
| 43.8000 | 9.81000000 | Furniture | 9875 |
| 121.5300 | 1.37333333 | Furniture | 9876 |
| 204.1500 | 17.73000000 | Office Supplies | 9877 |
| 40.1220 | 3.24000000 | Office Supplies | 9878 |
| 16.5300 | 2.04250000 | Office Supplies | 9879 |
| 14.9550 | 1.60666667 | Office Supplies | 9880 |

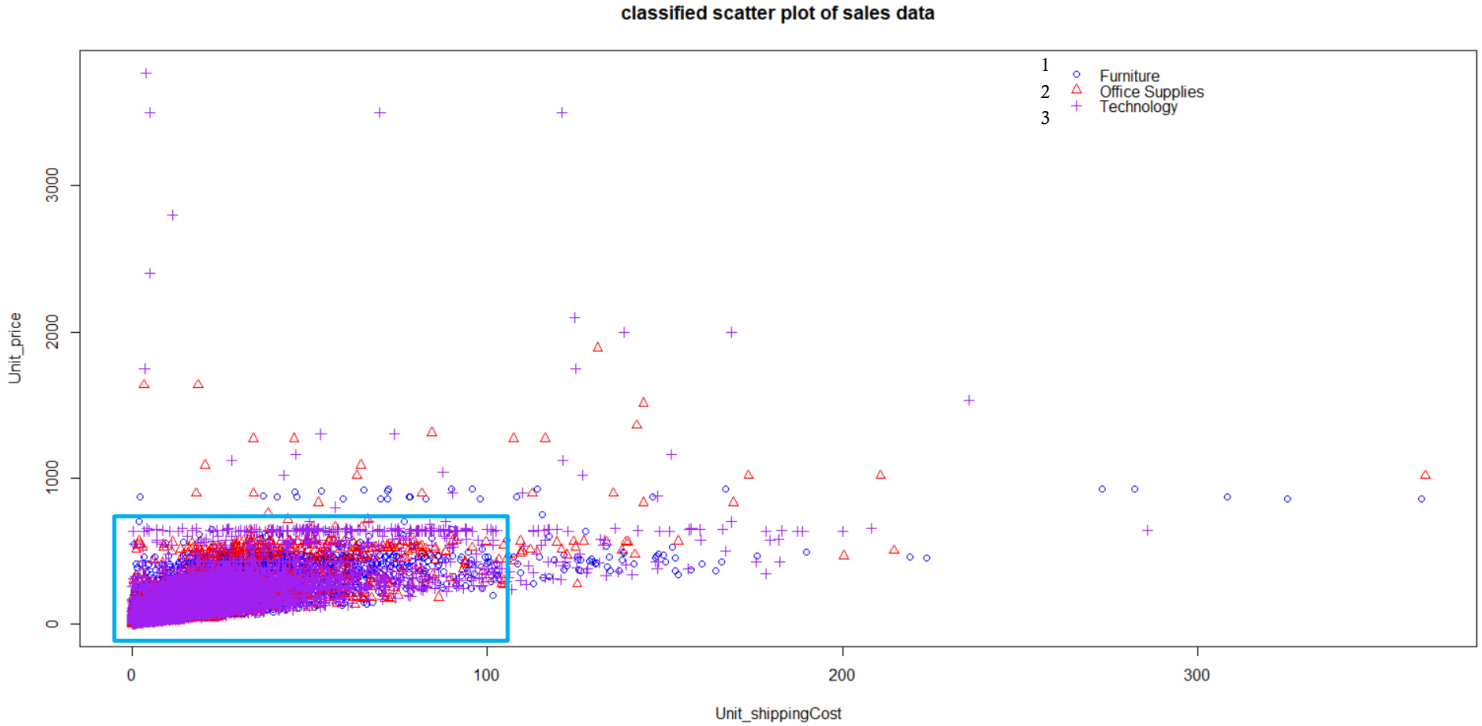
| | | | |
|---|---|---|---|
| 3.9900 | 0.49000000 | Office Supplies | 41146 |
| 8.8000 | 0.11666667 | Office Supplies | 41147 |
| 7.1200 | 0.20000000 | Office Supplies | 41148 |
| 1.0080 | 0.05666667 | Office Supplies | 41149 |
| 142.8900 | 28.65000000 | Technology | 41150 |
| 40.9920 | 3.10500000 | Technology | 41151 |
| 245.1300 | 24.42500000 | Technology | 41152 |

| | | | |
|---|---|---|---|
| 378.30000 | 11.7100000 | Technology | 51285 |
| 12.99000 | 1.3814286 | Technology | 51286 |
| 74.80000 | 7.3100000 | Technology | 51287 |
| 45.32000 | 2.0000000 | Technology | 51288 |
| 32.59200 | 3.4000000 | Technology | 51289 |
| 15.49350 | 0.7950000 | Technology | 51290 |

# 三、資料分布



原始資料量：共51263筆 (尚未做k-means)



# 三、資料分布



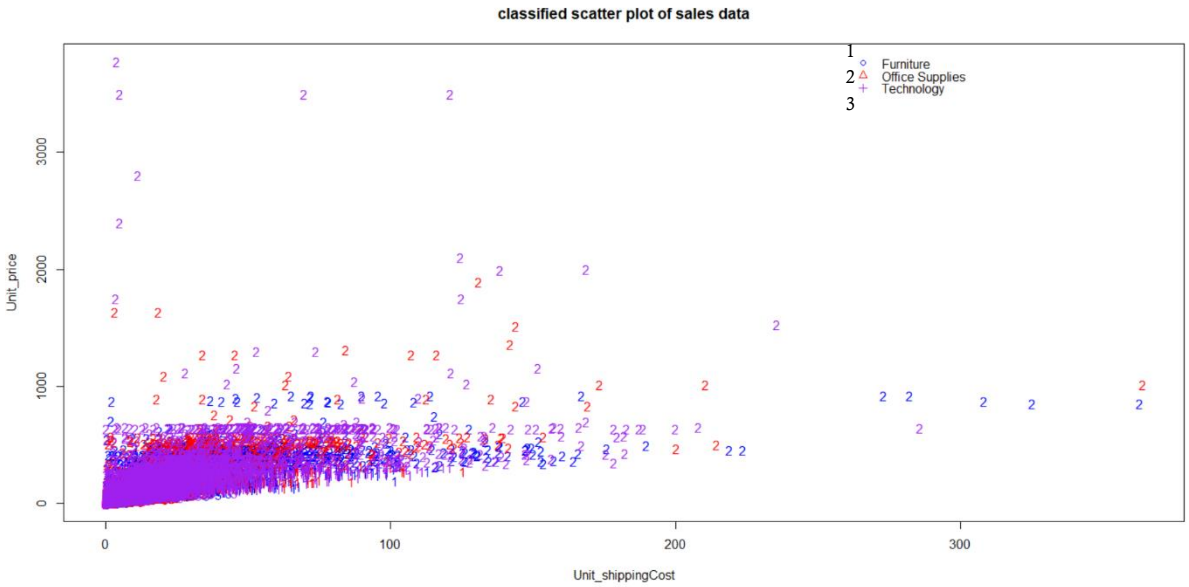
## 原始資料量：共51263筆 (做k-means)

結果不理想

```
Within cluster sum of squares by cluster:
[1]  35153197 103385267  22717395
 (between_SS / total_SS =  76.0 %)
```

```
> kmeans.m$centers #聚類中心
  Unit_price Unit_shippingCost
1  174.58741         18.891643
2  476.39004         51.461800
3   27.91605          3.054779
```

```
> table( kmeans.m$cluster, as.integer(Y$Category) )
        1     2     3
  1  3136  1741  4212
  2   950   404   676
  3  5790 29128  5253
```



# 三、資料分布



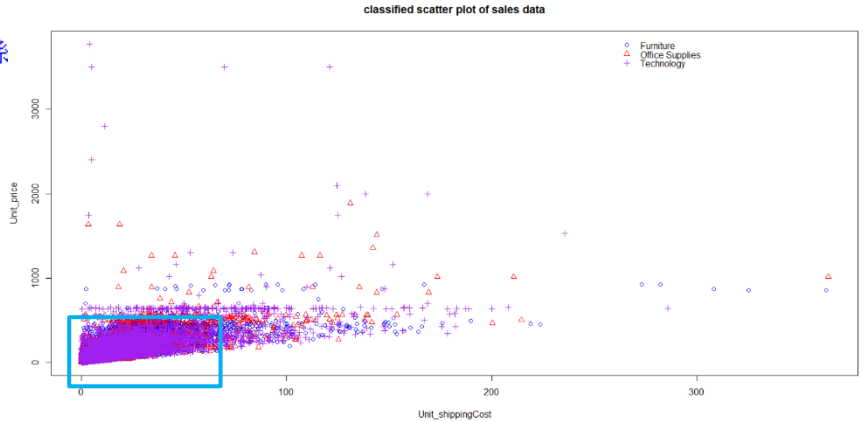
原始資料量：共51263筆

嘗試其他聚類方法:

k-medoids: fpc的pamk花了一小時仍還沒算完。

Hclust: 向量過大

```
> hc_cmp = hclust(dist(Y[,1:2])) # hclust 算出兩兩距離之間的關係
錯誤: 無法配置大小為 9.8 Gb 的向量
```



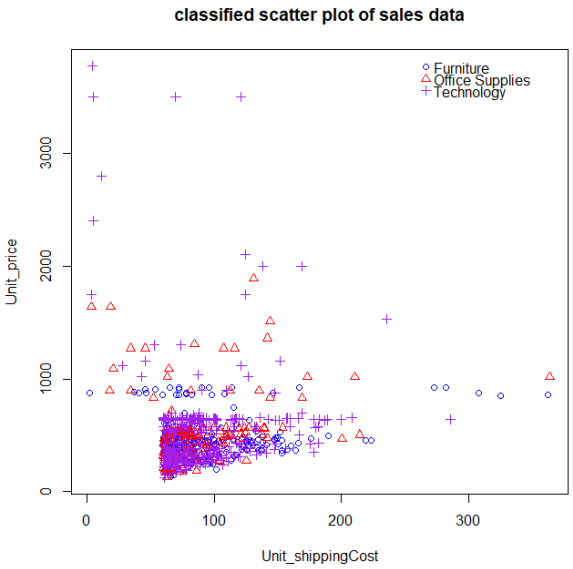

另尋通路:刪減雜訊

# 四、數據整理

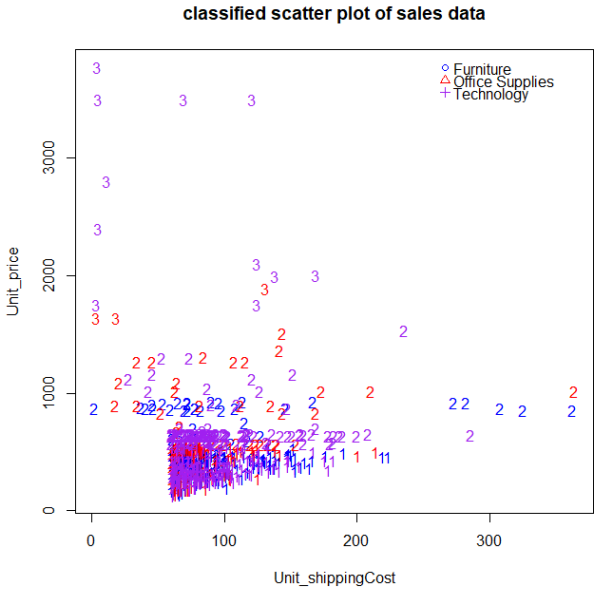


## 修改後資料量：共790筆

### K-means前




### K-means後




kmeans.m = kmeans(Y[,1:2], centers=3 )

```
> kmeans.m$centers #聚類中心
  Unit_price Unit_shippingCost
1   372.4142          86.24944
2   735.8276         103.15781
3  2445.2888          66.42049
```

```
Within cluster sum of squares by cluster:
[1] 5893019 8703729 8381914
 (between_SS / total_SS =  76.2 %)
```

```
> table( kmeans.m$cluster, as.integer(Y$Category) )

     1   2   3
  1 236 128 207
  2  39  39 127
  3   0   3  11
```

# 五、模型解讀

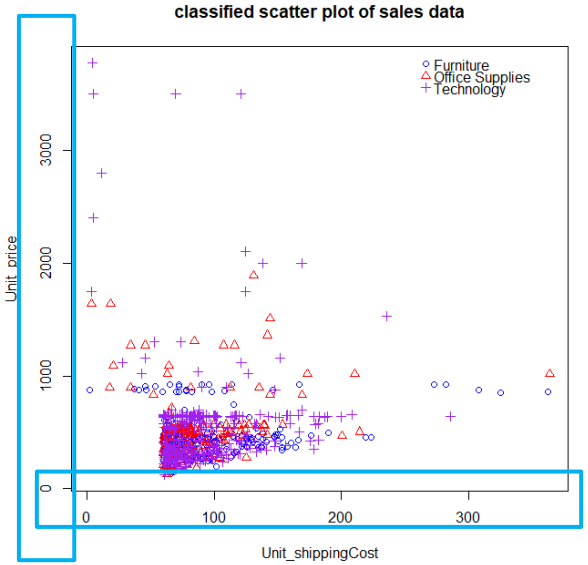


修改後資料量：共790筆

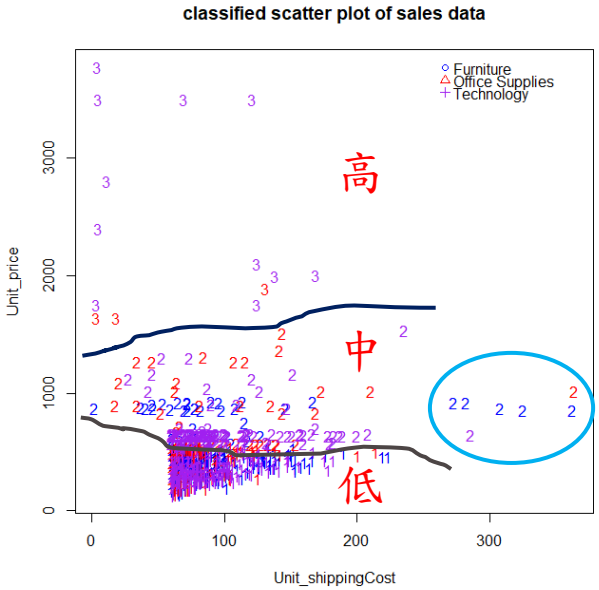
```
> table( kmeans.m$cluster, as.integer(Y$Category) )

     1   2   3
  1 236 128 207
  2  39  39 127
  3   0   3  11
```

K-means前




K-means後




1.符合假設，圖形沒有傾斜是因為單位比例不一致的關係。(y軸:一格1000, x軸:一格100)

2.初始的聚類中心(質心)影響大，容易影響周圍。

3.低獲利佔比(236+128+207)/790 = 72.2%。此公司的行銷策略為薄利多銷。