# Walmart : trip type classification

第六組　陳致希　李咨蓉

# TABLE OF CONTENTS

# EDA

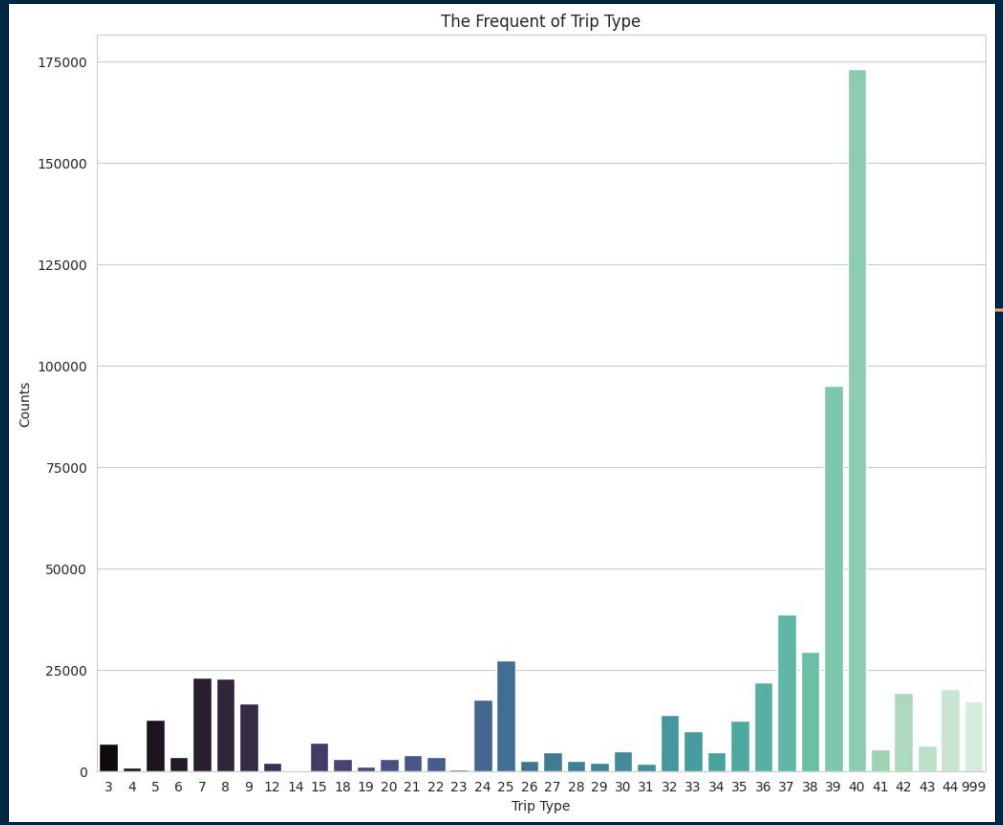
01

# Explore the responses

- 共38種trip type
- trip type 39.40的數量最多

| | |
|---|---|
| count | 38.000000 |
| mean | 16877.631579 |
| std | 30921.087184 |
| min | 34.000000 |
| 25% | 2983.250000 |
| 50% | 6575.500000 |
| 75% | 18912.500000 |
| max | 173031.000000 |

Name: TripType, dtype: float64



# Explore the predictors

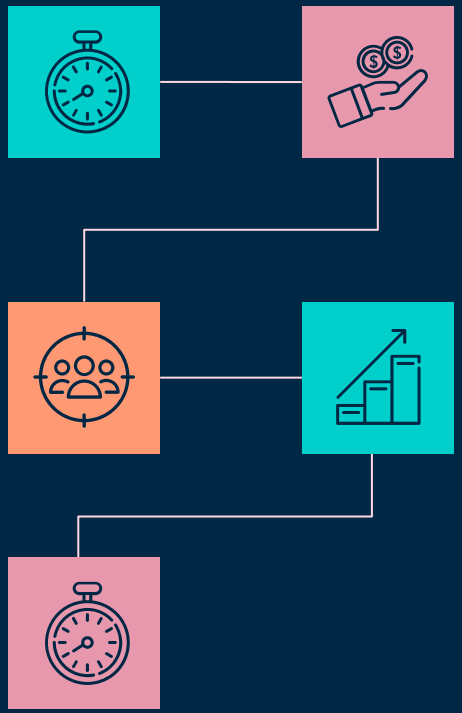
ScanCount
Returned
item

VisitNumber
UPC

Weekday

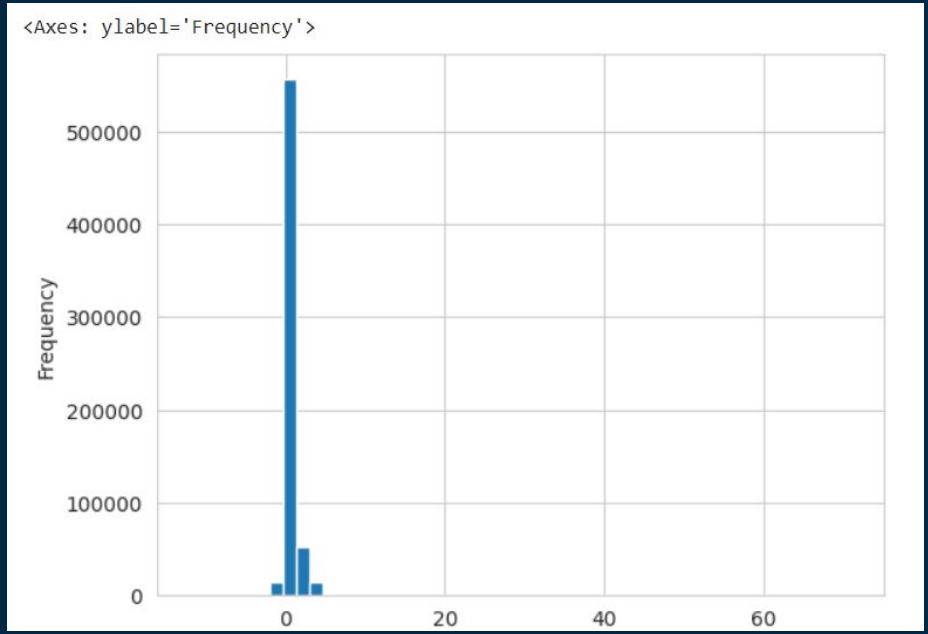
Department
Description

FinelineNumber



# ScanCount

- 我們發現沒有缺失值
- 我們可以得知1和2佔絕大多數，即大部分的採買一次只會買1或2個商品




| | |
|---|---|
| 1 | 556283 |
| 2 | 52839 |
| -1 | 14105 |
| 3 | 9421 |
| 4 | 4530 |
| 5 | 1378 |
| 6 | 865 |
| -2 | 802 |
| 8 | 246 |
| 7 | 228 |
| -3 | 142 |
| 10 | 130 |
| 9 | 88 |
| -4 | 76 |
| 12 | 60 |
| 11 | 45 |
| -5 | 16 |
| 14 | 15 |
| 13 | 13 |
| 15 | 12 |
| -6 | 10 |
| 20 | 7 |
| 16 | 6 |
| 19 | 4 |
| 18 | 4 |
| 24 | 4 |
| 23 | 3 |
| 17 | 3 |
| 25 | 3 |
| 22 | 2 |
| -9 | 2 |
| 71 | 1 |
| 51 | 1 |
| 30 | 1 |
| 31 | 1 |
| -7 | 1 |
| 46 | 1 |
| -12 | 1 |
| -10 | 1 |

Name: ScanCount, dtype: int64

# Returned item

- 已購買並退貨的商品應被刪除
- 刪除後我們發現購買物品數量沒有太大變化

Before

| | |
|---|---|
| 1 | 556283 |
| 2 | 52839 |
| -1 | 14105 |
| 3 | 9421 |
| 4 | 4530 |
| 5 | 1378 |
| 6 | 865 |
| -2 | 802 |
| 8 | 246 |
| 7 | 228 |
| -3 | 142 |
| 10 | 130 |
| 9 | 88 |
| -4 | 76 |
| 12 | 60 |
| 11 | 45 |
| -5 | 16 |
| 14 | 15 |
| 13 | 13 |
| 15 | 12 |
| -6 | 10 |
| 20 | 7 |
| 16 | 6 |
| 19 | 4 |
| 18 | 4 |
| 24 | 4 |
| 23 | 3 |
| 17 | 3 |
| 25 | 3 |
| 22 | 2 |
| -9 | 2 |
| 71 | 1 |
| 51 | 1 |
| 30 | 1 |
| 31 | 1 |
| -7 | 1 |
| 46 | 1 |
| -12 | 1 |
| -10 | 1 |

Name: ScanCount, dtype: int64

After

| | |
|---|---|
| 1 | 561489 |
| 2 | 53002 |
| -1 | 14394 |
| 3 | 9445 |
| 4 | 4536 |
| 5 | 1380 |
| 6 | 866 |
| -2 | 814 |
| 8 | 246 |
| 7 | 228 |
| -3 | 143 |
| 10 | 130 |
| 9 | 88 |
| -4 | 76 |
| 12 | 60 |
| 11 | 45 |
| -5 | 16 |
| 14 | 15 |
| 13 | 13 |
| 15 | 12 |
| -6 | 10 |
| 20 | 7 |
| 16 | 6 |
| 19 | 4 |
| 18 | 4 |
| 24 | 4 |
| 23 | 3 |
| 17 | 3 |
| 25 | 3 |
| 22 | 2 |
| -9 | 2 |
| 71 | 1 |
| 51 | 1 |
| 30 | 1 |
| 31 | 1 |
| -7 | 1 |
| 46 | 1 |
| -12 | 1 |
| -10 | 1 |

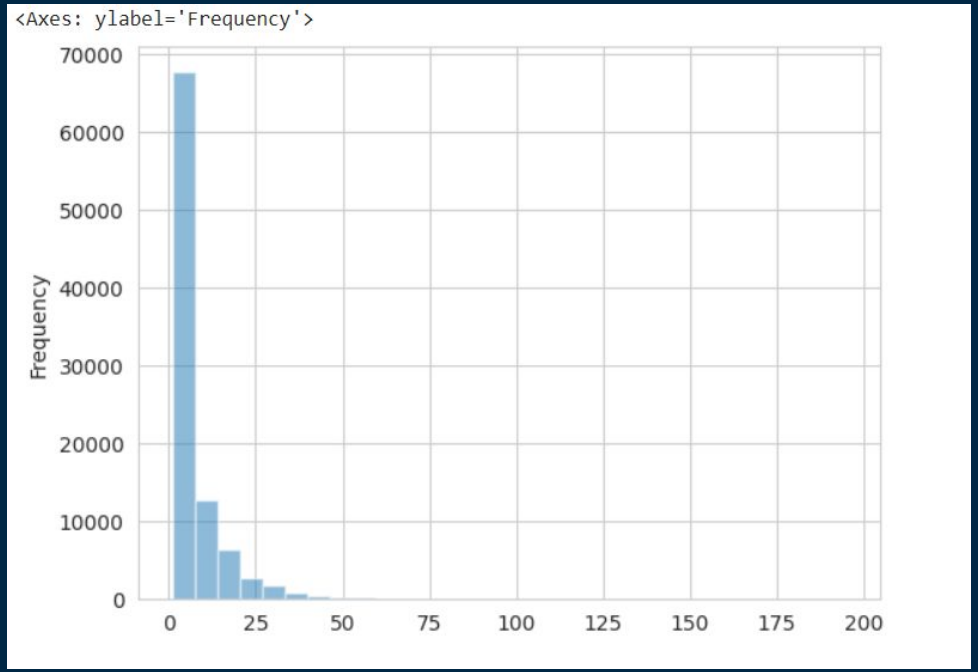
Name: ScanCount, dtype: int64

# UPC per Visit Number

- 我們發現沒有缺失值
- 有93086的購買次數，我們創建一個Series，統計每次購買的商品的種類數量
- Walmart這家大賣場，平均每次購買銷售 7 種的商品

```
count    93086.000000
mean         6.697774
std          8.418281
min          1.000000
25%          2.000000
50%          4.000000
75%          8.000000
max        195.000000
Name: unique_count, dtype: float64
```

購買項目數量的頻率分布

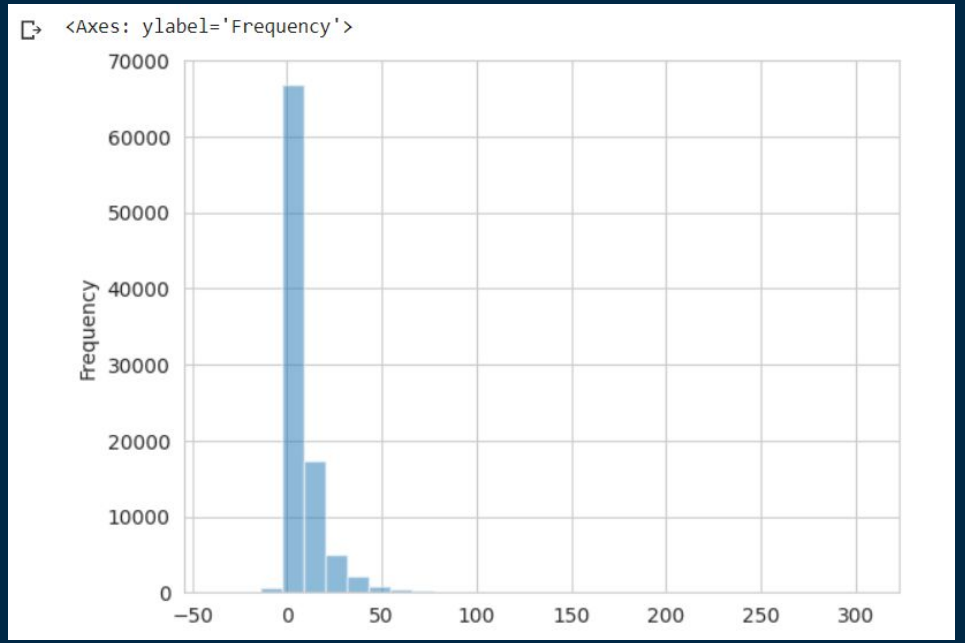


# ScanCount per VisitNumber

- 我們創建Series統計每次購買的商品數量
- 平均每次購買銷售約8個商品

```
count    93086.000000
mean         7.623327
std         10.267037
min        -37.000000
25%          2.000000
50%          4.000000
75%          9.000000
max        306.000000
Name: item_sum, dtype: float64
```

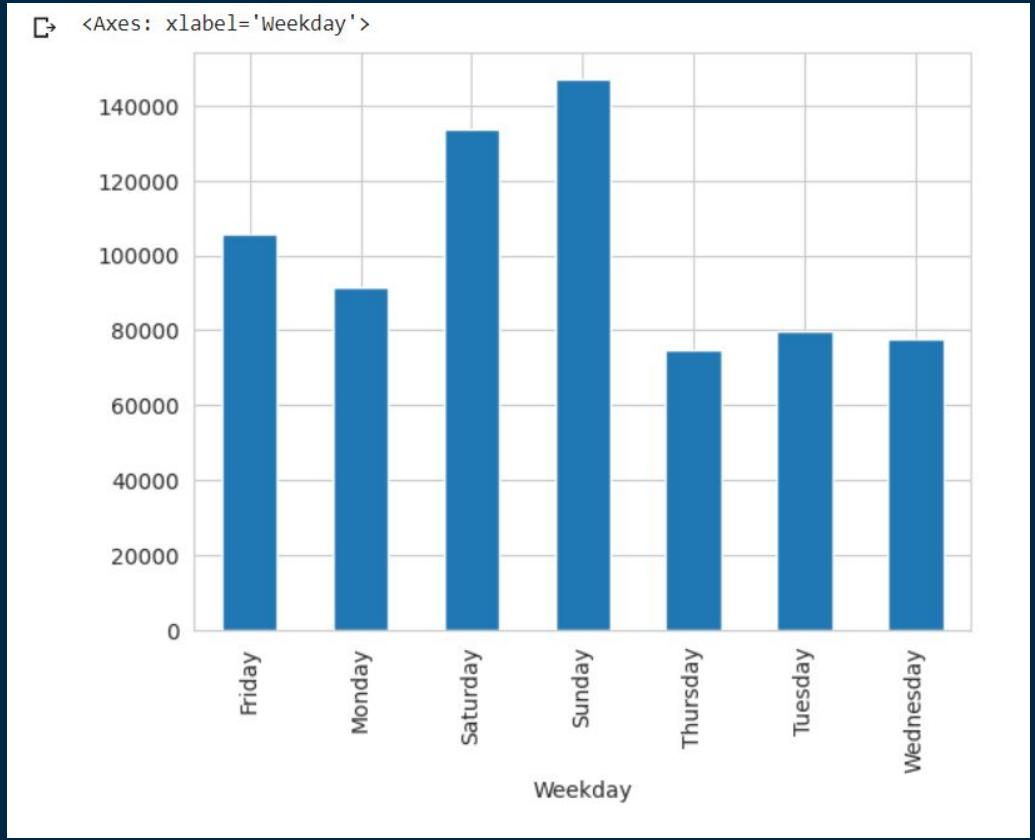
購物數量的頻率分布



# Weekday

- 我們觀察到周末的銷售量最大

每天的購物數量分布



# Department Description

- 這個資料有68種部門

## 銷售數量前十名的部門

```
GROCERY DRY GOODS            69016
DSD GROCERY                  66342
PRODUCE                      49563
DAIRY                        43006
PERSONAL CARE                41232
IMPULSE MERCHANDISE          27791
HOUSEHOLD CHEMICALS/SUPP     24352
PHARMACY OTC                 22772
FROZEN FOODS                 20726
HOUSEHOLD PAPER GOODS        15963
Name: DepartmentDescription, dtype: int64
```

## 銷售數量倒數十名的部門

```
1-HR PHOTO                   337
MENSWEAR                     302
CAMERAS AND SUPPLIES         207
PHARMACY RX                  143
OPTICAL - LENSES              85
LARGE HOUSEHOLD GOODS         77
CONCEPT STORES                35
OTHER DEPARTMENTS             29
SEASONAL                      29
HEALTH AND BEAUTY AIDS         2
Name: DepartmentDescription, dtype: int64
```

# FinelineNumber

- 這個資料有5188種部門

FinelineNumber的值出現的次數

```
5501.0    8150
1508.0    4904
135.0     4440
808.0     4331
0.0       3725
          ...
6345.0       1
4314.0       1
7160.0       1
3430.0       1
7313.0       1
Name: FinelineNumber, Length: 5188, dtype: int64
```

- 我們認為FinelineNumber太過detail，對資料預測幫助甚小，所以將這一列捨棄掉

| | DepartmentDescription | FinelineNumber | ScanCount |
|---|---|---|---|
| **8666** | PRODUCE | 5501.0 | 8098 |
| **1859** | DAIRY | 1508.0 | 5623 |
| **4740** | IMPULSE MERCHANDISE | 808.0 | 4703 |
| **4713** | IMPULSE MERCHANDISE | 135.0 | 4634 |
| **2775** | FINANCIAL SERVICES | 0.0 | 3717 |
| **3339** | GROCERY DRY GOODS | 3120.0 | 3710 |
| **2113** | DSD GROCERY | 4606.0 | 3617 |
| **2250** | DSD GROCERY | 9546.0 | 3173 |
| **1839** | DAIRY | 1407.0 | 2947 |
| **4711** | IMPULSE MERCHANDISE | 115.0 | 2837 |

# Feature Engineering

02

# Create feature from department detail

- 從前面VisitNumber的分析，我們得到每趟採買的物品數量及物品種數

| | VisitNumber | unique_count | item_sum | TripType | Weekday |
|---|---|---|---|---|---|
| **0** | 5 | 1 | -1 | 999 | Friday |
| **1** | 7 | 2 | 2 | 30 | Friday |
| **2** | 8 | 20 | 27 | 26 | Friday |
| **3** | 9 | 3 | 3 | 8 | Friday |
| **4** | 10 | 3 | 3 | 8 | Friday |

# Create feature from department detail

- 建立pivot table, 由此查看每趟採買在各個department共買了幾個東西(退貨也會記錄)

| DepartmentDescription | VisitNumber | 1-HR PHOTO | ACCESSORIES | AUTOMOTIVE | BAKERY | BATH AND SHOWER | BEAUTY | BEDDING | BOOKS AND MAGAZINES | BOYS WEAR | BRAS & SHAPEWEAR | CAMERAS AND SUPPLIES | CANDY, TOBACCO, COOKIES | CELEBRATION | COMM BREAD | CONCEPT STORES | COOK AND DINE | DAIRY |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 1 | 7 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | 8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.0 |
| 3 | 9 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 4 | 10 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

| DSD GROCERY | ELECTRONICS | FABRICS AND CRAFTS | FINANCIAL SERVICES | FROZEN FOODS | FURNITURE | GIRLS WEAR, 4-6X AND 7-14 | GROCERY DRY GOODS | HARDWARE | HEALTH AND BEAUTY AIDS | HOME DECOR | HOME MANAGEMENT | HORTICULTURE AND ACCESS | HOUSEHOLD CHEMICALS/SUPP | HOUSEHOLD PAPER GOODS | IMPULSE MERCHANDISE | INFANT APPAREL | INFANT CONSUMABLE HARDLINES |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.0 | 0.0 | -1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

# Create feature from department detail

- 合併前述兩張圖表，得到所需資料

| | VisitNumber | unique_count | item_sum | TripType | Weekday | 1-HR PHOTO | ACCESSORIES | AUTOMOTIVE | BAKERY | BATH AND SHOWER | BEAUTY | BEDDING | BOOKS AND MAGAZINES | BOYS WEAR | BRAS & SHAPEWEAR | CAMERAS AND SUPPLIES | CANDY, TOBACCO, COOKIES | CELEBRATION | COMM BREAD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 5 | 1 | -1 | 999 | Friday | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **1** | 7 | 2 | 2 | 30 | Friday | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **2** | 8 | 20 | 27 | 26 | Friday | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **3** | 9 | 3 | 3 | 8 | Friday | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **4** | 10 | 3 | 3 | 8 | Friday | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |

| CONCEPT STORES | COOK AND DINE | DAIRY | DSD GROCERY | ELECTRONICS | FABRICS AND CRAFTS | FINANCIAL SERVICES | FROZEN FOODS | FURNITURE | GIRLS WEAR, 4-6X AND 7-14 | GROCERY DRY GOODS | HARDWARE | HEALTH AND BEAUTY AIDS | HOME DECOR | HOME MANAGEMENT | HORTICULTURE AND ACCESS | HOUSEHOLD CHEMICALS/SUPP | HOUSEHOLD PAPER GOODS | IMPULSE MERCHANDISE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | -1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 2.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |
| 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| 0.0 | 0.0 | 0.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

# Encoding Days

- 將weekdays改為數字，Monday為0，Tuesday為1，依此類推
- 將數字改為sin與cos，使其具有循環性
- 資料新增sin_day與cos_day欄位，刪除Weekday欄位

| | VisitNumber | sin_day | cos_day |
|---|---|---|---|
| 0 | 5 | -0.433884 | -0.900969 |
| 1 | 7 | -0.433884 | -0.900969 |
| 2 | 8 | -0.433884 | -0.900969 |
| 3 | 9 | -0.433884 | -0.900969 |
| 4 | 10 | -0.433884 | -0.900969 |

# EDA2

03

從購買商品的department判斷trip type的目標，以trip type 39為例：

| | VisitNumber | unique_count | item_sum | TripType | BEAUTY | CANDY, TOBACCO, COOKIES | COOK AND DINE | DAIRY | DSD GROCERY | FROZEN FOODS | FURNITURE | GIRLS WEAR, 4-6X AND 7-14 | GROCERY DRY GOODS | HOUSEHOLD CHEMICALS/SUPP | HOUSEHOLD PAPER GOODS | IMPULSE MERCHANDISE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **13** | 26 | 9 | 12 | 39 | 0.0 | 0.0 | 2.0 | 6.0 | 0.0 | 2.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| **41** | 79 | 8 | 9 | 39 | 0.0 | 3.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 3.0 | 0.0 | 0.0 | 1.0 |
| **72** | 138 | 7 | 7 | 39 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 2.0 | 0.0 | 1.0 | 0.0 |
| **108** | 224 | 14 | 18 | 39 | 0.0 | 5.0 | 0.0 | 0.0 | 7.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 6.0 | 0.0 |
| **155** | 314 | 5 | 5 | 39 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 |

type 39看起來是grocery trip，接著看trip type 5：

| | VisitNumber | unique_count | item_sum | TripType | PHARMACY OTC |
|---|---|---|---|---|---|
| **58** | 105 | 3 | 4 | 5 | 4.0 |
| **103** | 218 | 1 | 1 | 5 | 1.0 |
| **139** | 285 | 5 | 5 | 5 | 3.0 |
| **193** | 382 | 2 | 2 | 5 | 2.0 |
| **215** | 418 | 1 | 1 | 5 | 1.0 |

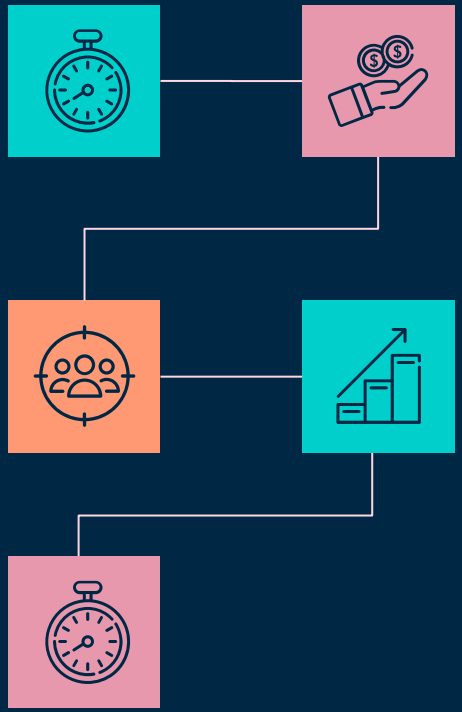
type 5看起來是pharmacy trip

# Modeling

04

# Model

Split test set



Random forest

SVM

KNN

Naive Bayes

# Split test set

- 我們將特徵數據X和目標數據y分割成訓練集和測試集，並將他們存儲在四個變量中:X_train、X_test、y_train 和 y_test。
- 將測試集的大小設定為總樣本數的 20%，也就是將資料的 20% 作為測試集，剩餘的 80% 作為訓練集。

```
1 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

# Random forest

n_estimators =10

Log Loss: 4.406240841282578
Accuracy: 0.9200059085781812

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 3 | 0.787810 | 0.968100 | 0.868699 | 721.000000 |
| 4 | 0.000000 | 0.000000 | 0.000000 | 71.000000 |
| 5 | 0.633292 | 0.738686 | 0.681941 | 685.000000 |
| 6 | 0.667969 | 0.665370 | 0.666667 | 257.000000 |
| 7 | 0.619469 | 0.673077 | 0.645161 | 1144.000000 |
| 44 | 0.997840 | 0.965517 | 0.981413 | 957.000000 |
| 999 | 0.989506 | 0.789052 | 0.877983 | 5736.000000 |
| accuracy | 0.920006 | 0.920006 | 0.920006 | 0.920006 |
| macro avg | 0.932963 | 0.881793 | 0.901634 | 74468.000000 |
| weighted avg | 0.924433 | 0.920006 | 0.919688 | 74468.000000 |

n_estimators =100

Log Loss: 1.7228307286316096
Accuracy: 0.9288687758500296

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 3 | 0.791572 | 0.963939 | 0.869293 | 721.000000 |
| 4 | 0.000000 | 0.000000 | 0.000000 | 71.000000 |
| 5 | 0.660256 | 0.751825 | 0.703072 | 685.000000 |
| 6 | 0.703252 | 0.673152 | 0.687873 | 257.000000 |
| 7 | 0.687122 | 0.694930 | 0.691004 | 1144.000000 |
| 44 | 1.000000 | 1.000000 | 1.000000 | 957.000000 |
| 999 | 0.994752 | 0.793061 | 0.882530 | 5736.000000 |
| accuracy | 0.928869 | 0.928869 | 0.928869 | 0.928869 |
| macro avg | 0.943319 | 0.893547 | 0.912054 | 74468.000000 |
| weighted avg | 0.933499 | 0.928869 | 0.928569 | 74468.000000 |

# Random forest

n_estimators =500

Log Loss: 1.3368782753346702
Accuracy: 0.928882204436805

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 3 | 0.790724 | 0.969487 | 0.871028 | 721.000000 |
| 4 | 0.111111 | 0.014085 | 0.025000 | 71.000000 |
| 5 | 0.663239 | 0.753285 | 0.705400 | 685.000000 |
| 6 | 0.697211 | 0.680934 | 0.688976 | 257.000000 |
| 7 | 0.689746 | 0.687937 | 0.688840 | 1144.000000 |
| 44 | 1.000000 | 1.000000 | 1.000000 | 957.000000 |
| 999 | 0.995185 | 0.792713 | 0.882484 | 5736.000000 |
| accuracy | 0.928882 | 0.928882 | 0.928882 | 0.928882 |
| macro avg | 0.944341 | 0.892711 | 0.911897 | 74468.000000 |
| weighted avg | 0.933746 | 0.928882 | 0.928605 | 74468.000000 |

# KNN

## K=1

Log Loss: 15.32697410471805
accuracy: 0.9077590374389

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 3 | 0.780702 | 0.864078 | 0.820276 | 721.000000 |
| 4 | 0.086022 | 0.112676 | 0.097561 | 71.000000 |
| 5 | 0.596932 | 0.624818 | 0.610556 | 685.000000 |
| 6 | 0.555160 | 0.607004 | 0.579926 | 257.000000 |
| 7 | 0.590444 | 0.604895 | 0.597582 | 1144.000000 |
| 44 | 1.000000 | 1.000000 | 1.000000 | 957.000000 |
| 999 | 0.874790 | 0.816074 | 0.844412 | 5736.000000 |
| accuracy | 0.907759 | 0.907759 | 0.907759 | 0.907759 |
| macro avg | 0.892001 | 0.888818 | 0.889102 | 74468.000000 |
| weighted avg | 0.908791 | 0.907759 | 0.907805 | 74468.000000 |

## K=5

Log Loss: 6.583590821200494
accuracy: 0.7141993876564431

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 3 | 0.765502 | 0.941748 | 0.844527 | 721.00000 |
| 4 | 0.032258 | 0.014085 | 0.019608 | 71.00000 |
| 5 | 0.599767 | 0.750365 | 0.666667 | 685.00000 |
| 6 | 0.670782 | 0.634241 | 0.652000 | 257.00000 |
| 7 | 0.583888 | 0.690559 | 0.632759 | 1144.00000 |
| 44 | 0.604982 | 0.177638 | 0.274637 | 957.000000 |
| 999 | 0.972048 | 0.727510 | 0.832187 | 5736.000000 |
| accuracy | 0.714199 | 0.714199 | 0.714199 | 0.714199 |
| macro avg | 0.626572 | 0.541460 | 0.560260 | 74468.000000 |
| weighted avg | 0.712604 | 0.714199 | 0.702763 | 74468.000000 |

# KNN

K=100

Log Loss: 1.7499422024420705
accuracy: 0.6409330182091637

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 3 | 0.775882 | 0.945908 | 0.852500 | 721.000000 |
| 4 | 0.000000 | 0.000000 | 0.000000 | 71.000000 |
| 5 | 0.638353 | 0.724088 | 0.678523 | 685.000000 |
| 6 | 0.740741 | 0.544747 | 0.627803 | 257.000000 |
| 7 | 0.705549 | 0.544580 | 0.614702 | 1144.000000 |
| 44 | 0.041667 | 0.001045 | 0.002039 | 957.000000 |
| 999 | 0.998761 | 0.702406 | 0.824770 | 5736.000000 |
| accuracy | 0.640933 | 0.640933 | 0.640933 | 0.640933 |
| macro avg | 0.507197 | 0.406626 | 0.426470 | 74468.000000 |
| weighted avg | 0.630745 | 0.640933 | 0.614879 | 74468.000000 |

# SVM

## C=1(default)

Log Loss: 1.1011347109839278
Accuracy: 0.6857844980394263

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 3 | 0.779408 | 0.928718 | 0.847537 | 2918.000000 |
| 4 | 0.000000 | 0.000000 | 0.000000 | 275.000000 |
| 5 | 0.595231 | 0.792382 | 0.679801 | 2678.000000 |
| 6 | 0.677233 | 0.461690 | 0.549065 | 1018.000000 |
| 7 | 0.699853 | 0.617973 | 0.656369 | 4607.000000 |
| 44 | 0.600402 | 0.312435 | 0.410997 | 957.000000 |
| 999 | 0.998757 | 0.700488 | 0.823445 | 5736.000000 |
| accuracy | 0.685784 | 0.685784 | 0.685784 | 0.685784 |
| macro avg | 0.618580 | 0.473923 | 0.491628 | 74468.000000 |
| weighted avg | 0.687898 | 0.685784 | 0.667751 | 74468.000000 |

## C=10

Log Loss: 1.0384631841763812
Accuracy: 0.7191142504162862

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 3 | 0.797131 | 0.933173 | 0.859804 | 2918.000000 |
| 4 | 0.500000 | 0.003636 | 0.007220 | 275.000000 |
| 5 | 0.623586 | 0.823376 | 0.709688 | 2678.000000 |
| 6 | 0.695890 | 0.748527 | 0.721249 | 1018.000000 |
| 7 | 0.723022 | 0.698068 | 0.710326 | 4607.000000 |
| 44 | 0.789244 | 0.567398 | 0.660182 | 957.000000 |
| 999 | 0.995825 | 0.706939 | 0.826876 | 5736.000000 |
| accuracy | 0.719114 | 0.719114 | 0.719114 | 0.719114 |
| macro avg | 0.647676 | 0.543052 | 0.556259 | 74468.000000 |
| weighted avg | 0.722973 | 0.719114 | 0.708872 | 74468.000000 |

# Naive Bayes (GaussianNB)

var_smoothing =1e-9(default)

Log Loss: 30.291361284655817
Accuracy: 0.10604554976634259

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 3 | 0.985075 | 0.022618 | 0.044221 | 2918.000000 |
| 4 | 0.014205 | 0.785455 | 0.027905 | 275.000000 |
| 5 | 0.847368 | 0.120239 | 0.210595 | 2678.000000 |
| 6 | 0.122917 | 0.115914 | 0.119312 | 1018.000000 |
| 7 | 0.552674 | 0.074018 | 0.130551 | 4607.000000 |
| 44 | 0.269436 | 0.264368 | 0.266878 | 957.000000 |
| 999 | 0.730769 | 0.003312 | 0.006595 | 5736.000000 |
| accuracy | 0.106046 | 0.106046 | 0.106046 | 0.106046 |
| macro avg | 0.308950 | 0.245208 | 0.146449 | 74468.000000 |
| weighted avg | 0.454671 | 0.106046 | 0.128823 | 74468.000000 |

var_smoothing =0.001

Log Loss: 6.254887019268576
Accuracy: 0.39165816189504216

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 3 | 0.165993 | 0.984578 | 0.284090 | 2918.000000 |
| 4 | 0.057531 | 0.400000 | 0.100594 | 275.000000 |
| 5 | 0.217070 | 0.120612 | 0.155065 | 2678.000000 |
| 6 | 0.285324 | 0.307466 | 0.295981 | 1018.000000 |
| 7 | 0.373406 | 0.502279 | 0.428360 | 4607.000000 |
| 44 | 0.303185 | 0.248694 | 0.273249 | 957.000000 |
| 999 | 0.605096 | 0.066248 | 0.119422 | 5736.000000 |
| accuracy | 0.391658 | 0.391658 | 0.391658 | 0.391658 |
| macro avg | 0.366103 | 0.376693 | 0.318155 | 74468.000000 |
| weighted avg | 0.499927 | 0.391658 | 0.378683 | 74468.000000 |

# Naive Bayes (GaussianNB)

var_smoothing =0.01

Log Loss: 3.4628437007772037
Accuracy: 0.4096658967610249

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 3 | 0.805025 | 0.505141 | 0.620762 | 2918.000000 |
| 4 | 0.250000 | 0.003636 | 0.007168 | 275.000000 |
| 5 | 0.167022 | 0.205004 | 0.184074 | 2678.000000 |
| 6 | 0.251989 | 0.093320 | 0.136201 | 1018.000000 |
| 7 | 0.221776 | 0.287389 | 0.250355 | 4607.000000 |
| 44 | 0.326019 | 0.108673 | 0.163009 | 957.000000 |
| 999 | 0.803864 | 0.239365 | 0.368888 | 5736.000000 |
| accuracy | 0.409666 | 0.409666 | 0.409666 | 0.409666 |
| macro avg | 0.441873 | 0.265045 | 0.289799 | 74468.000000 |
| weighted avg | 0.503884 | 0.409666 | 0.388346 | 74468.000000 |

var_smoothing =1

Log Loss: 2.79306097416673
Accuracy: 0.2280711177955632

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 3 | 0.000000 | 0.000000 | 0.000000 | 2918.000000 |
| 4 | 0.000000 | 0.000000 | 0.000000 | 275.000000 |
| 5 | 0.000000 | 0.000000 | 0.000000 | 2678.000000 |
| 6 | 0.000000 | 0.000000 | 0.000000 | 1018.000000 |
| 7 | 0.000000 | 0.000000 | 0.000000 | 4607.000000 |
| 44 | 0.000000 | 0.000000 | 0.000000 | 957.000000 |
| 999 | 1.000000 | 0.000349 | 0.000697 | 5736.000000 |
| accuracy | 0.228071 | 0.228071 | 0.228071 | 0.228071 |
| macro avg | 0.084328 | 0.064888 | 0.048182 | 74468.000000 |
| weighted avg | 0.179925 | 0.228071 | 0.122846 | 74468.000000 |

# Reference

https://www.kaggle.com/code/naksungp/quick-eda-plus-model#EDA!

# Thanks