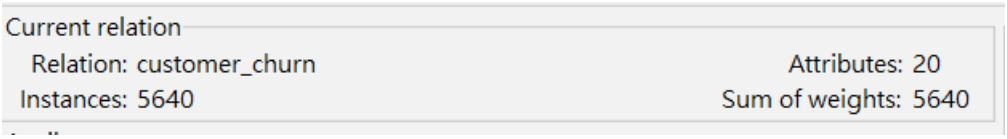
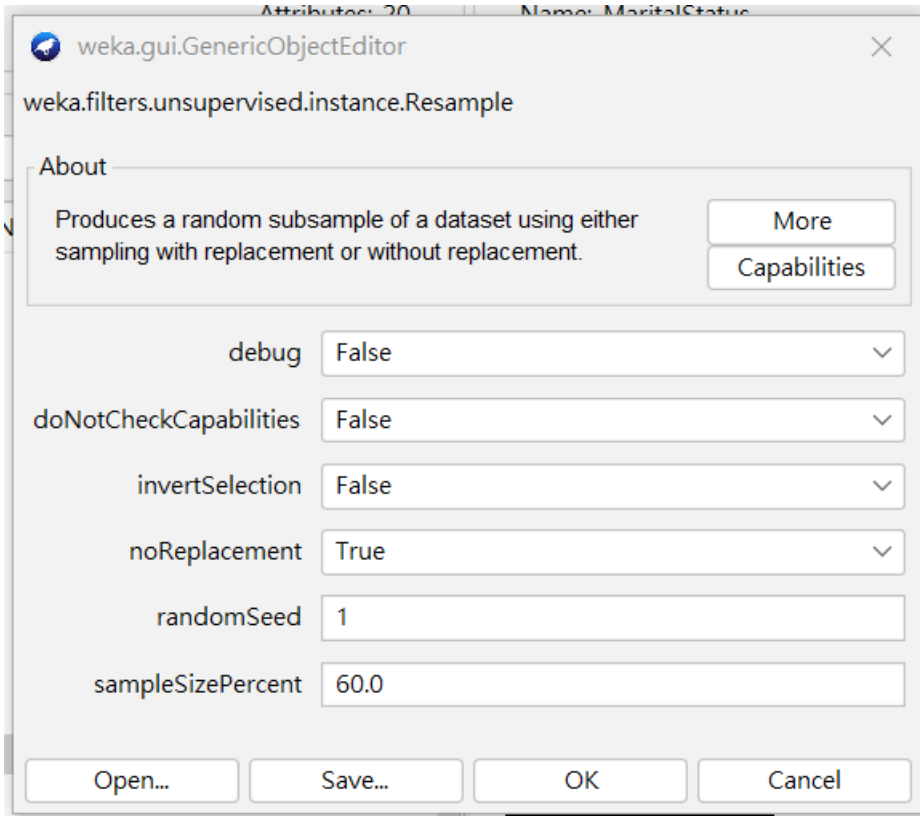
電子商務技術_HW5

108403201 資管三 A 黃名揚

1. 使用 Stratified sampling 從原本的資料集中取 60%的資料

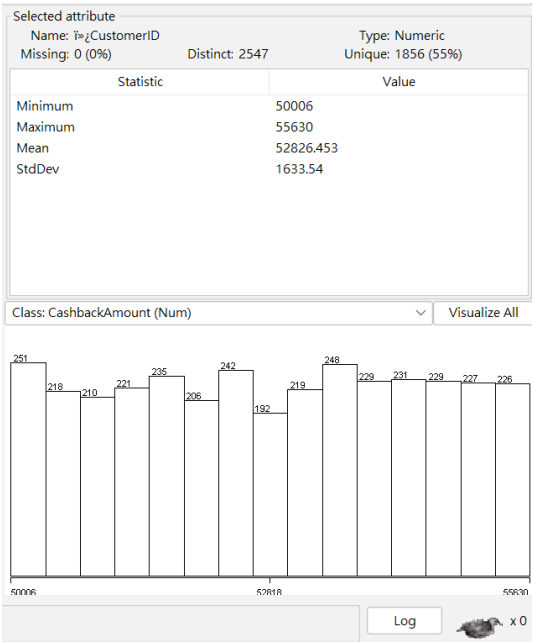


初始資料量

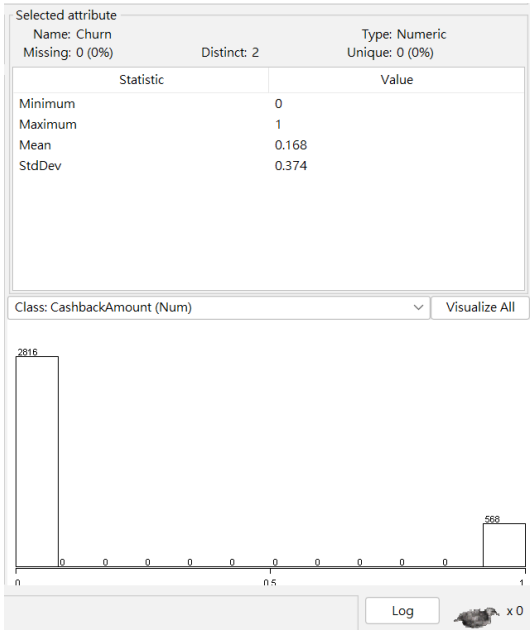


使用 Resample 取 60%的資料

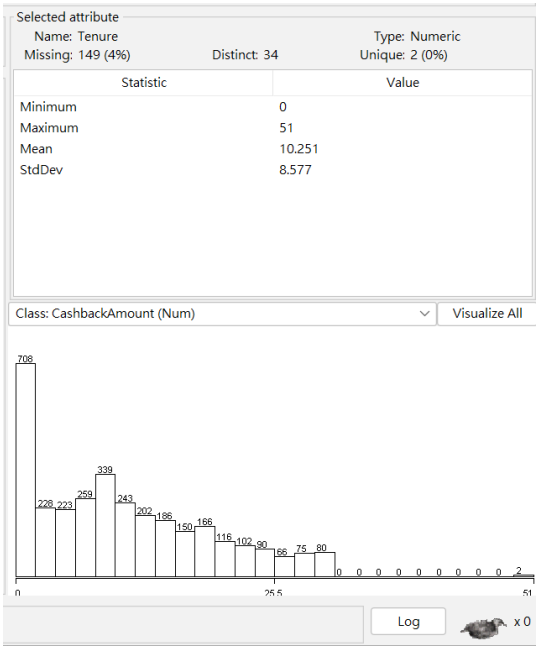
2. 顯示取樣後各類別的資料數量



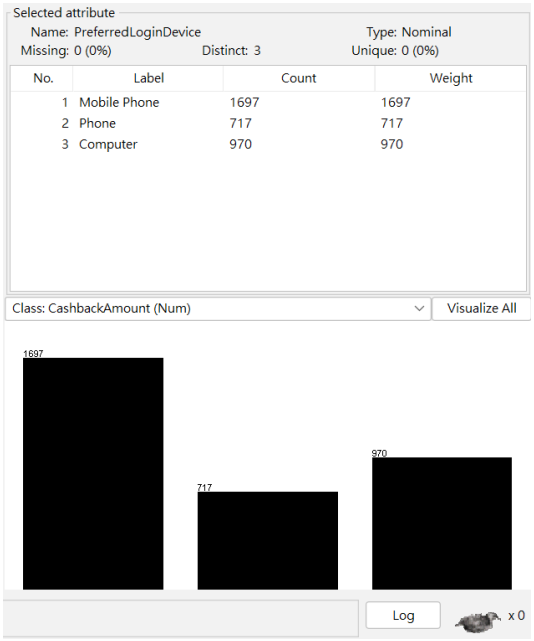
CustomerID



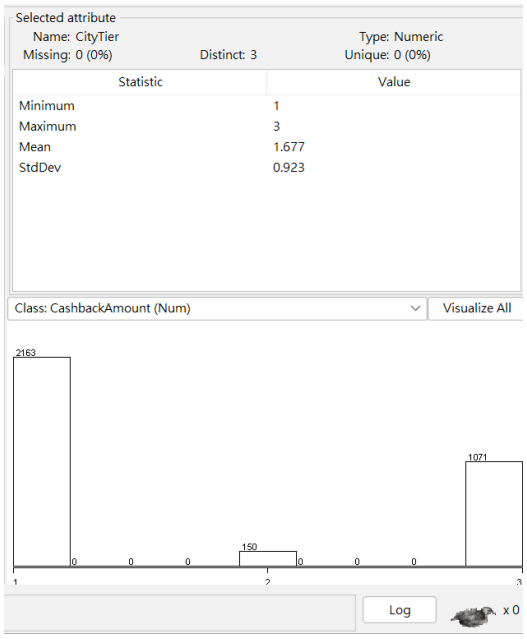
Churn



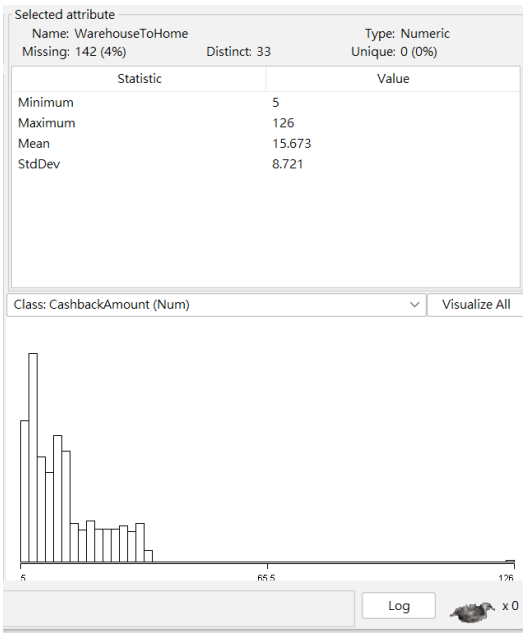
Tenure



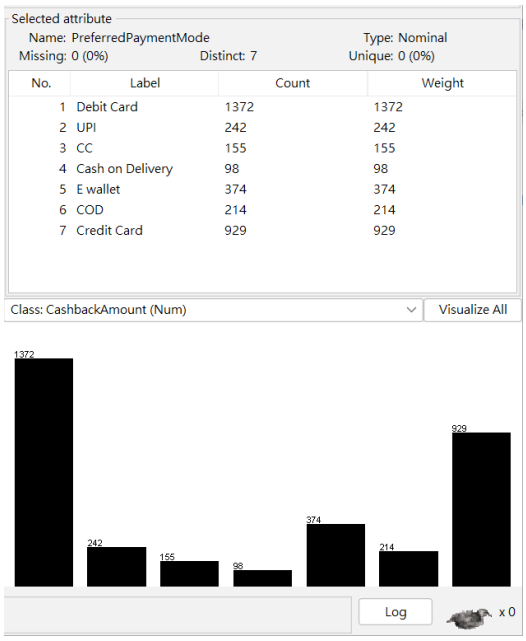
PreferredLoginDevice



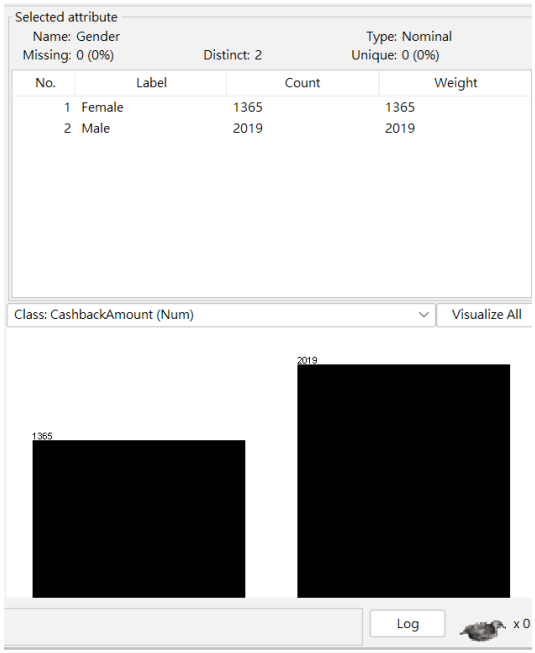
CityTier



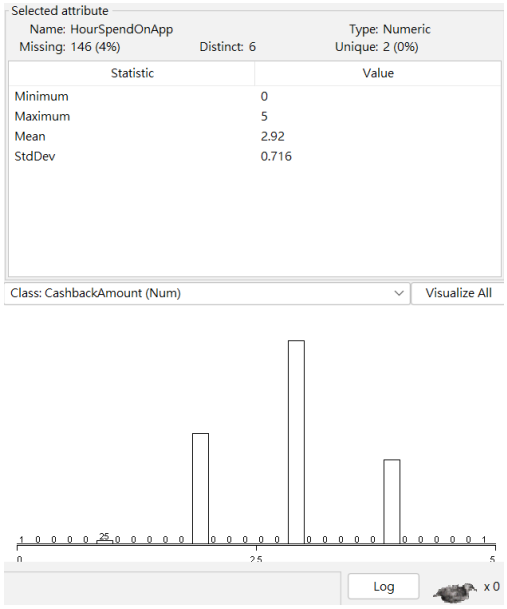
WarehouseToHome



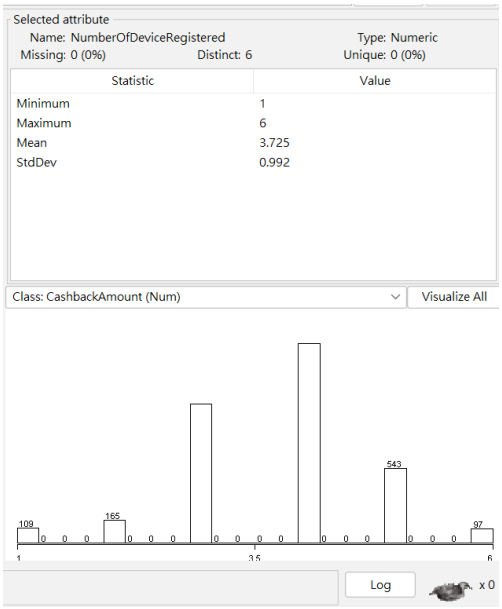
PreferredPaymentMode



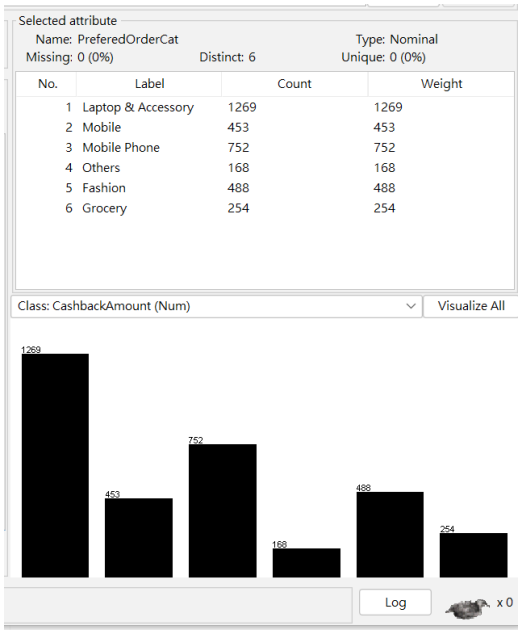
Gender



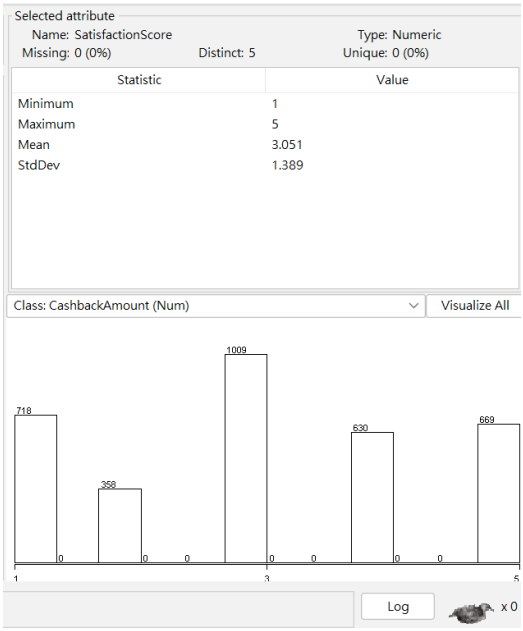
HourSpendOnApp



NumberOfDeviceRegistered



PrederredOrderCat



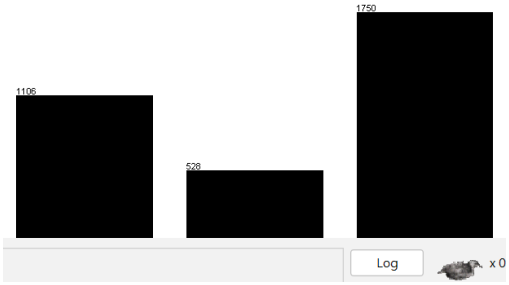
SatisfactionScore

Selected attribute

Name: MaritalStatus Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	Single	1106	1106
2	Divorced	528	528
3	Married	1750	1750

Class: CashbackAmount (Num) Visualize All



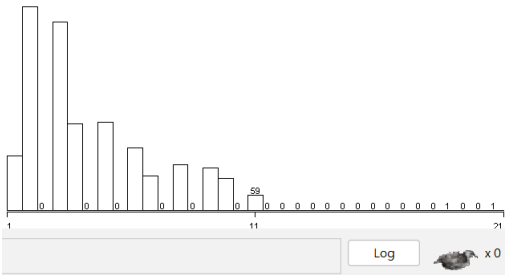
MaritalStatus

Selected attribute

Name: NumberOfAddress Type: Numeric
Missing: 0 (0%) Distinct: 13 Unique: 2 (0%)

Statistic	Value
Minimum	1
Maximum	21
Mean	4.29
StdDev	2.617

Class: CashbackAmount (Num) Visualize All



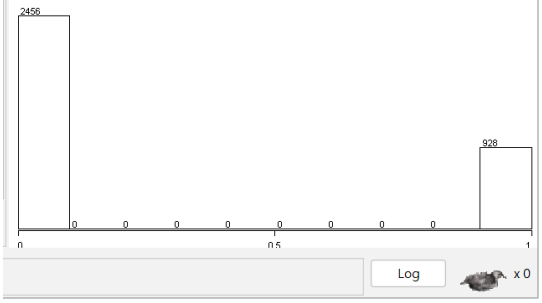
NumberOfAddress

Selected attribute

Name: Complain Type: Numeric
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.274
StdDev	0.446

Class: CashbackAmount (Num) Visualize All



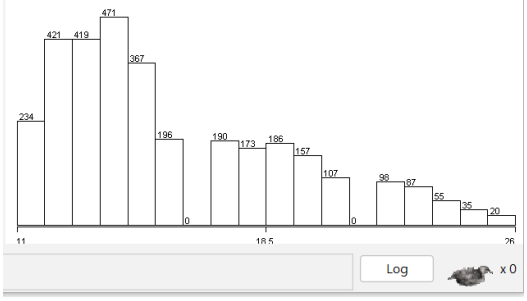
Complain

Selected attribute

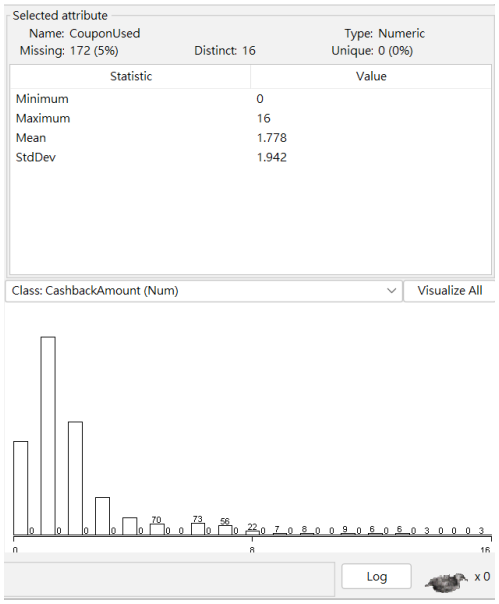
Name: OrderAmountHikeFromlastYear Type: Numeric
Missing: 168 (5%) Distinct: 16 Unique: 0 (0%)

Statistic	Value
Minimum	11
Maximum	26
Mean	15.686
StdDev	3.622

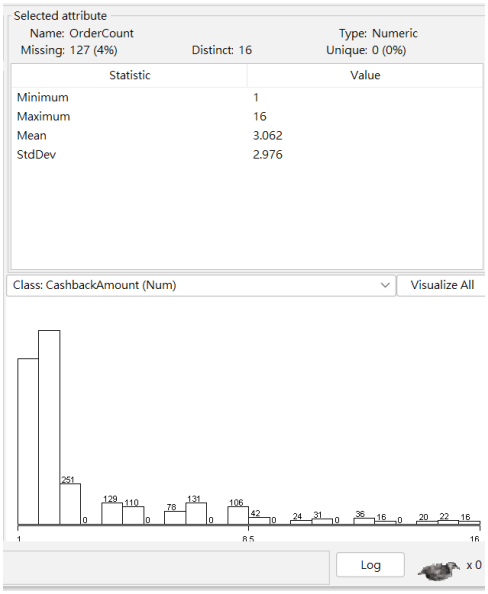
Class: CashbackAmount (Num) Visualize All



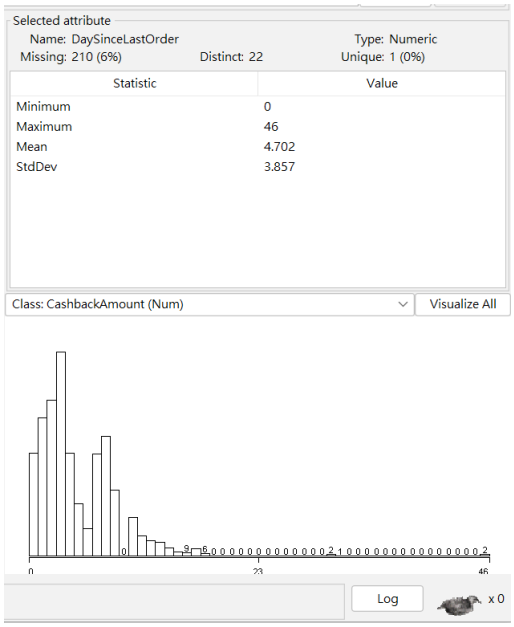
OrderAmountHikeFromlastYear



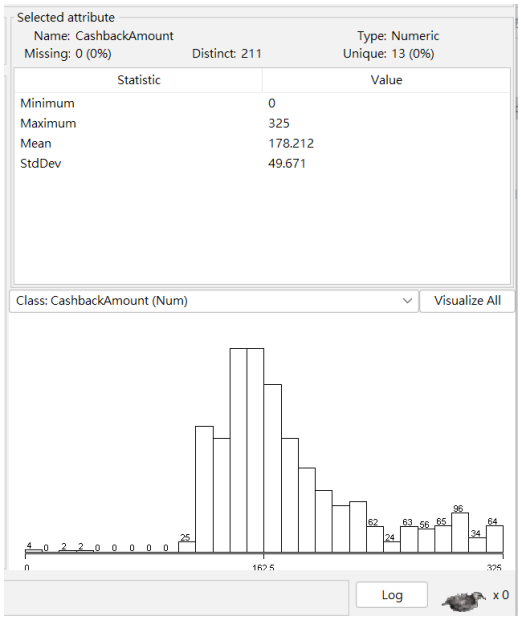
CouponUsed



OrderCount

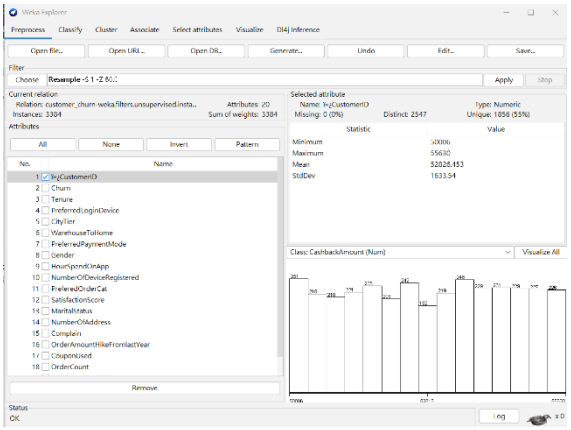


DaySinceOrder

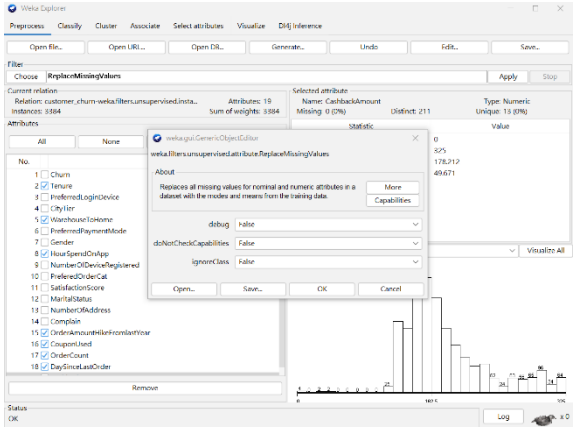


CashbackAmount

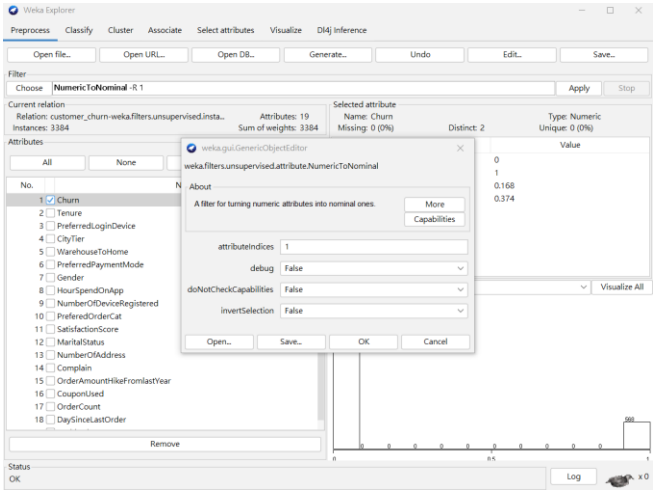
3. 資料前處理，並以 repeated 10 folds cross-validation (重複 10 次) Paired t-test 比 較 Logistic Regression 及 SVM 模型



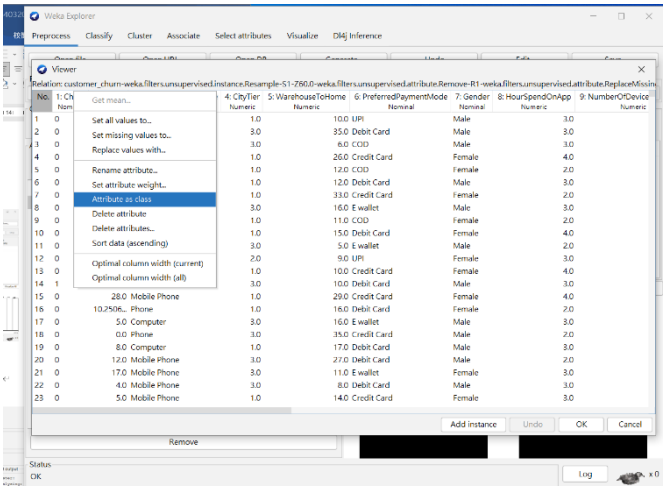
刪除 CustomerID



填補空值

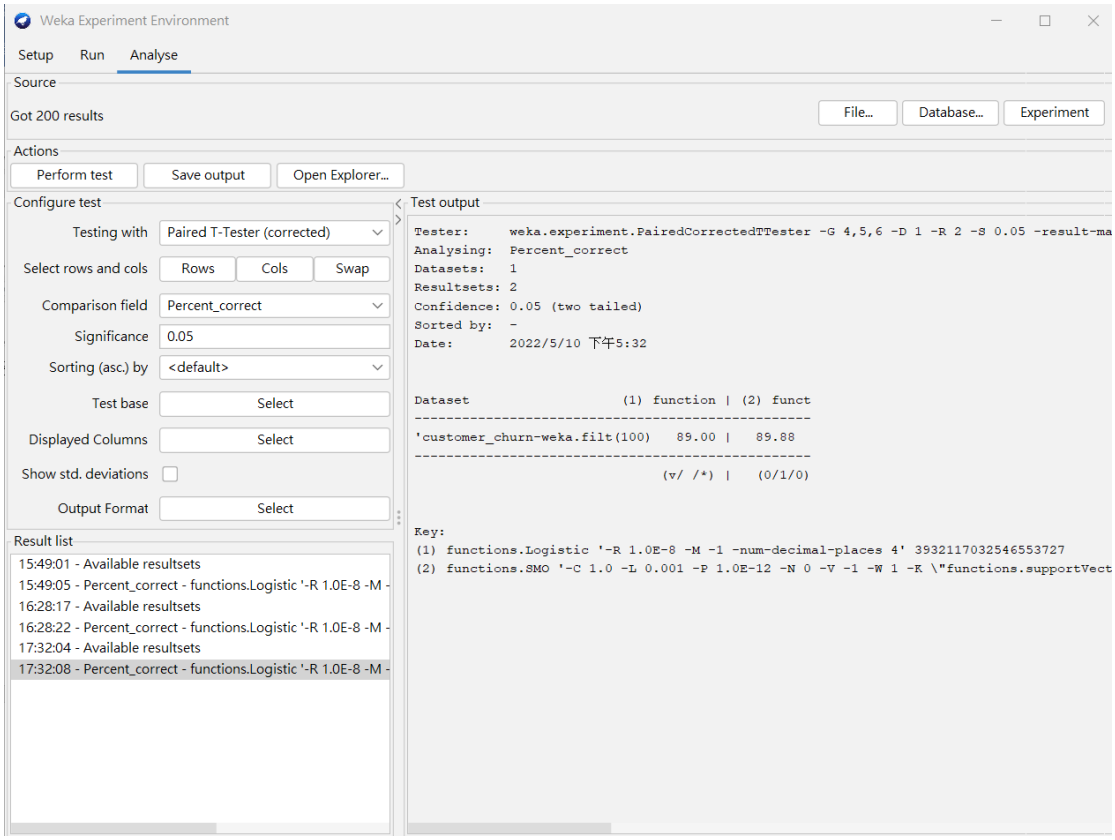


將 Churn 轉為 Nominal



將 Churn 設為 class

4. 根據 weka 的輸出說明結論



信賴區間:0.05

採用雙尾檢定，logistic model 與 SVM model 準確率皆約為 89%，故此訓練集在兩模型的預測下差異不太，無法拒絕虛無假設