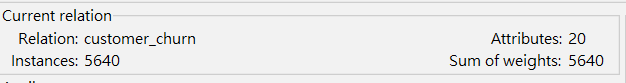
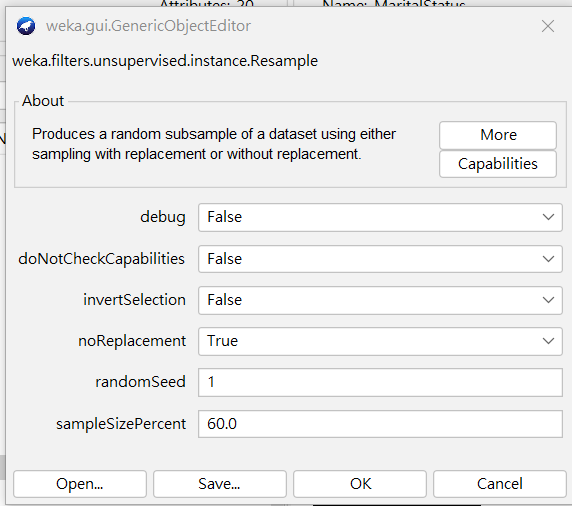
電子商務技術\_HW5

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

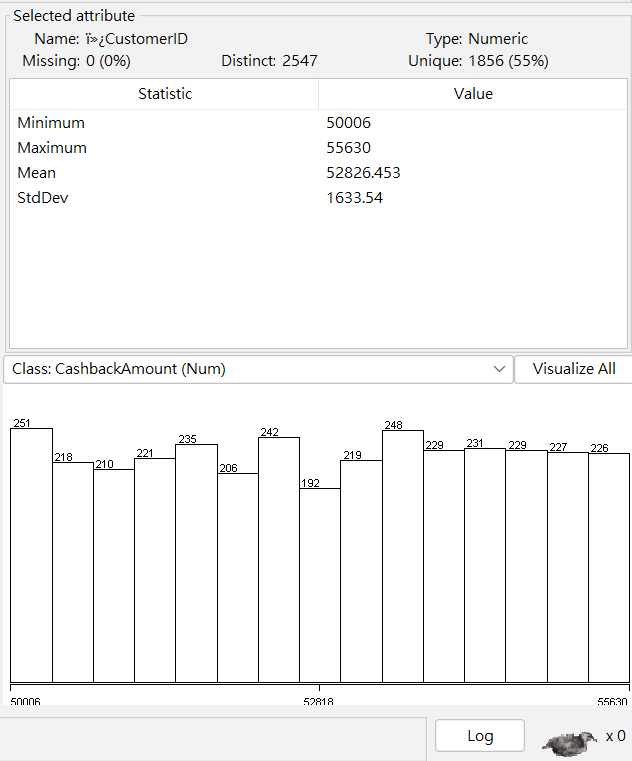
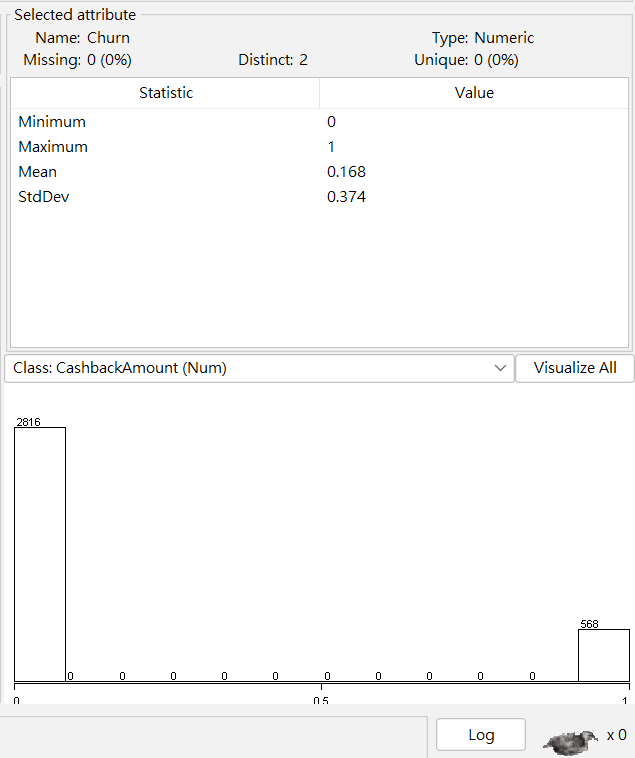
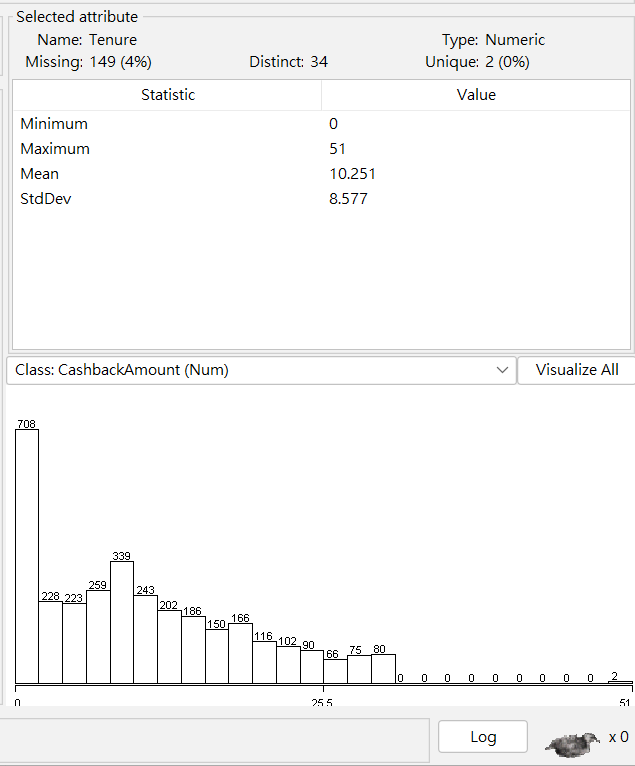


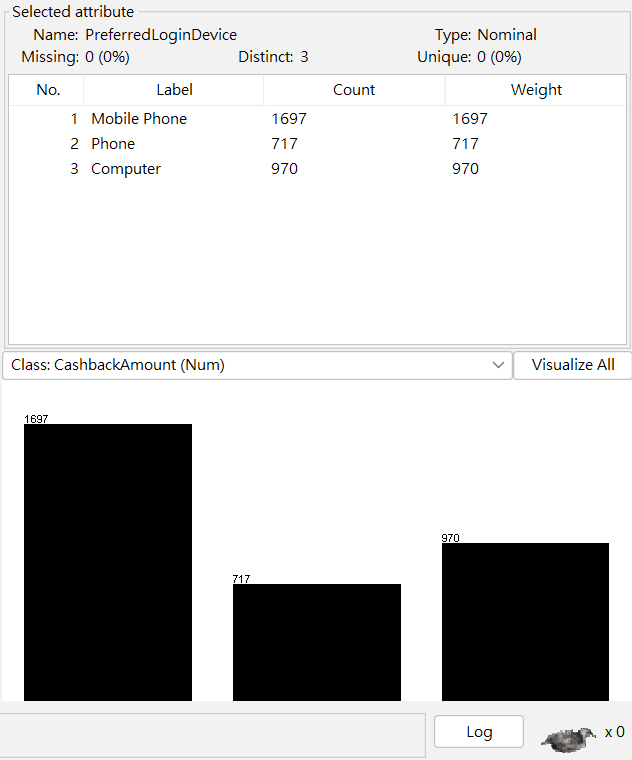
初始資料量



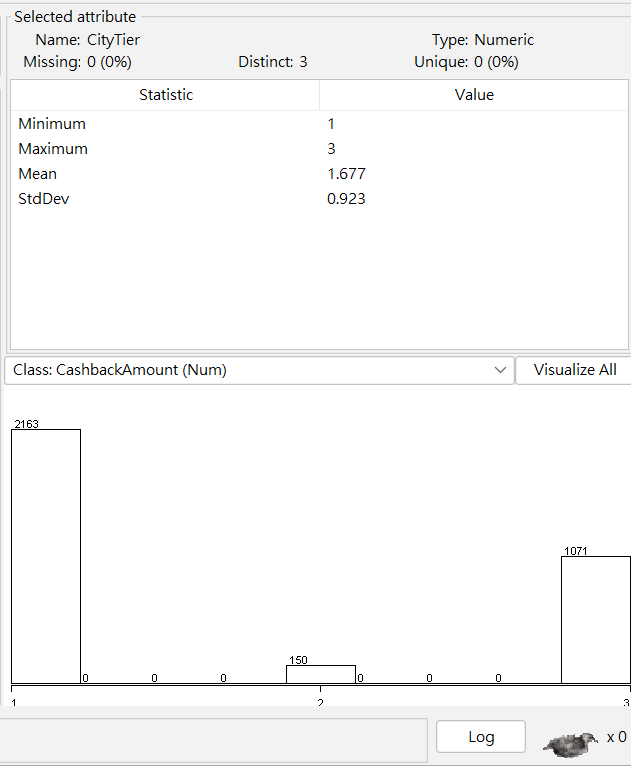
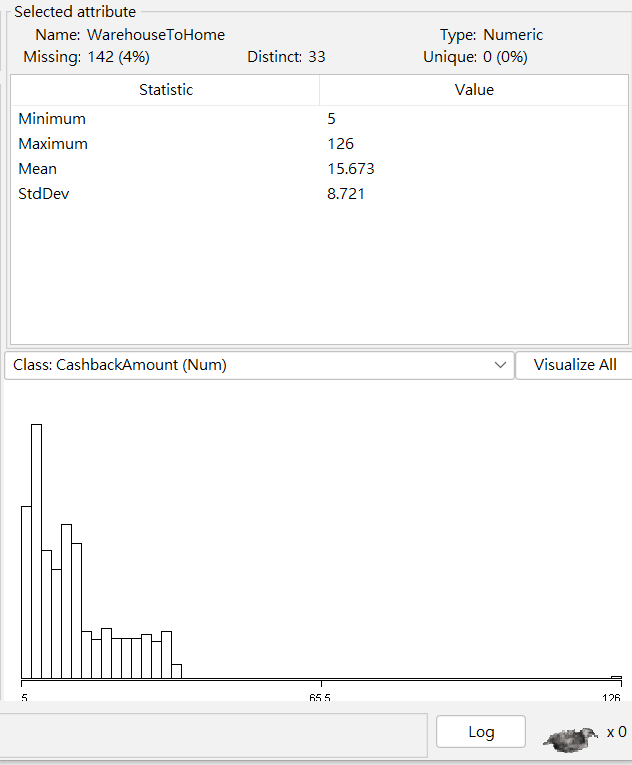
使用Resample取60%的資料

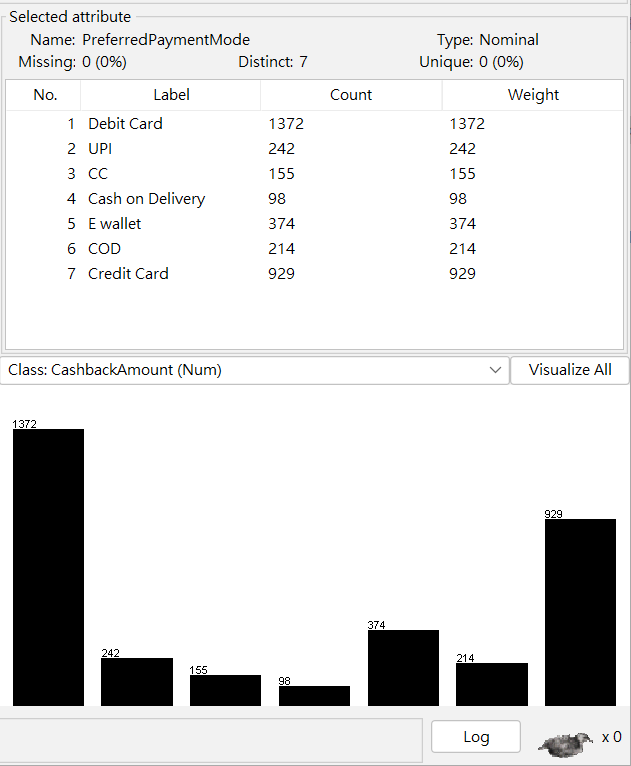
2. 顯示取樣後各類別的資料數量

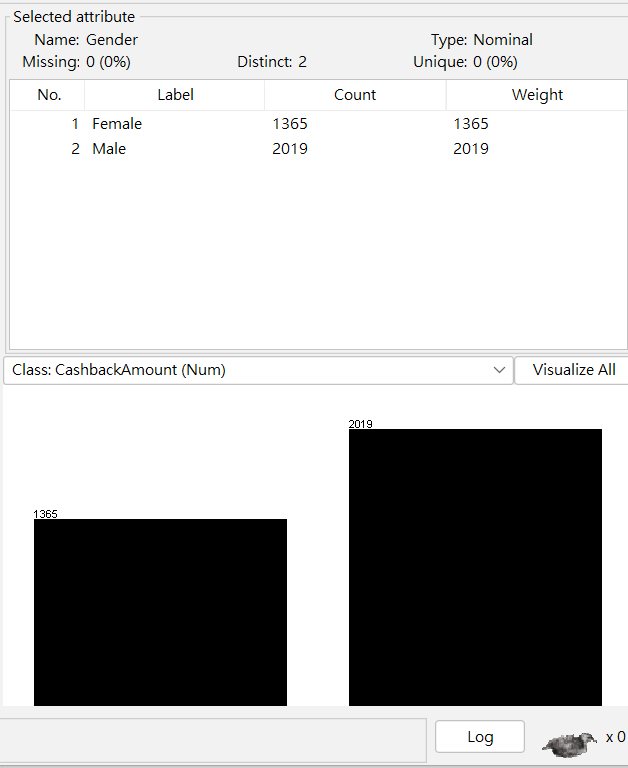


CustomerID Churn

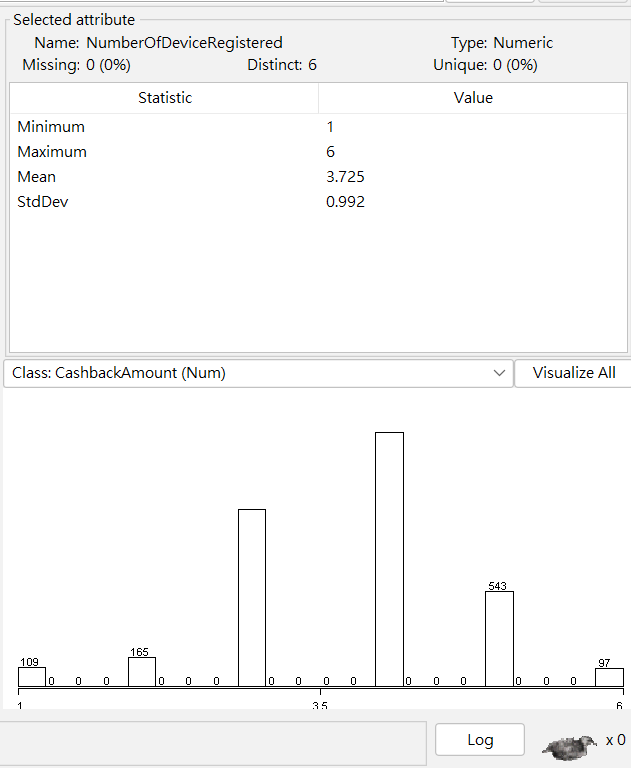
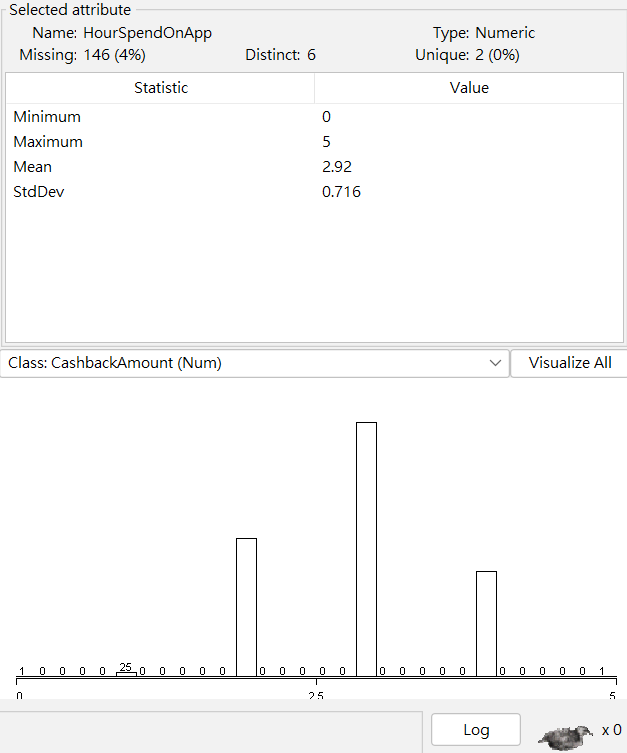
Tenure PreferredLoginDevice



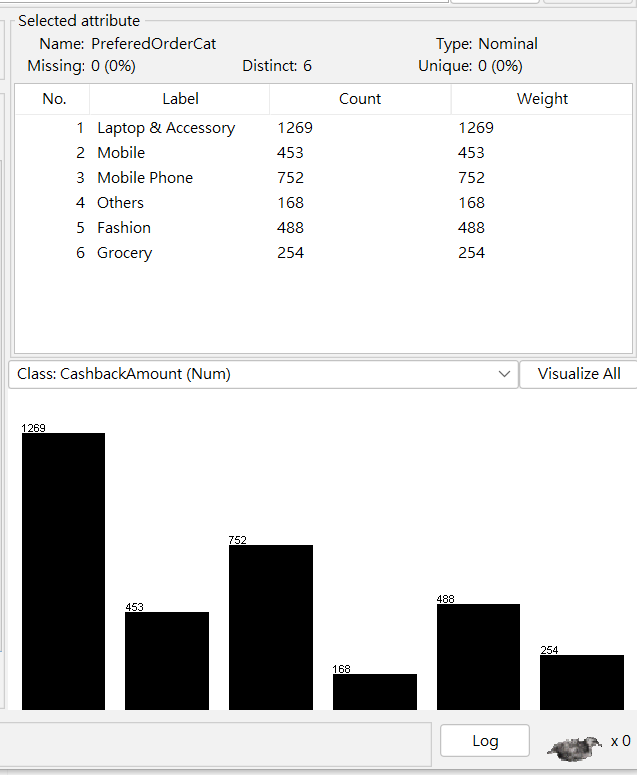
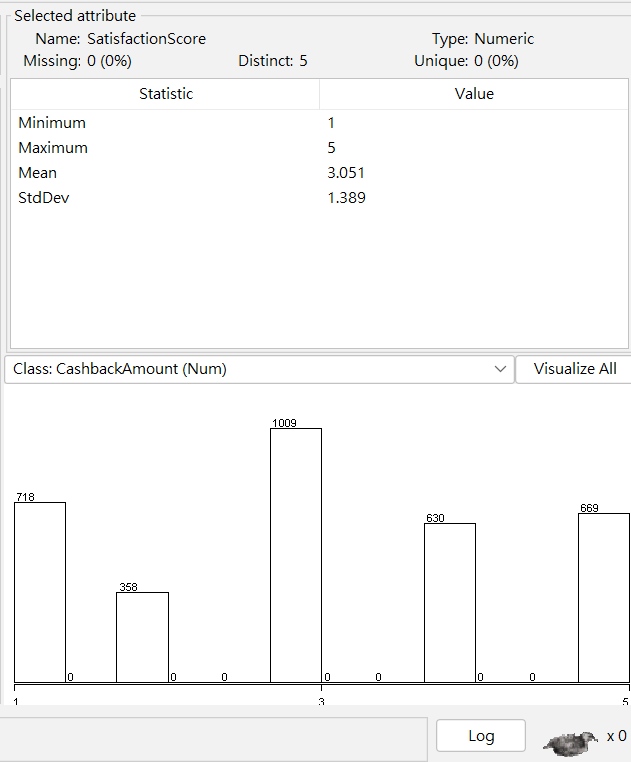
CityTier WarehouseToHome



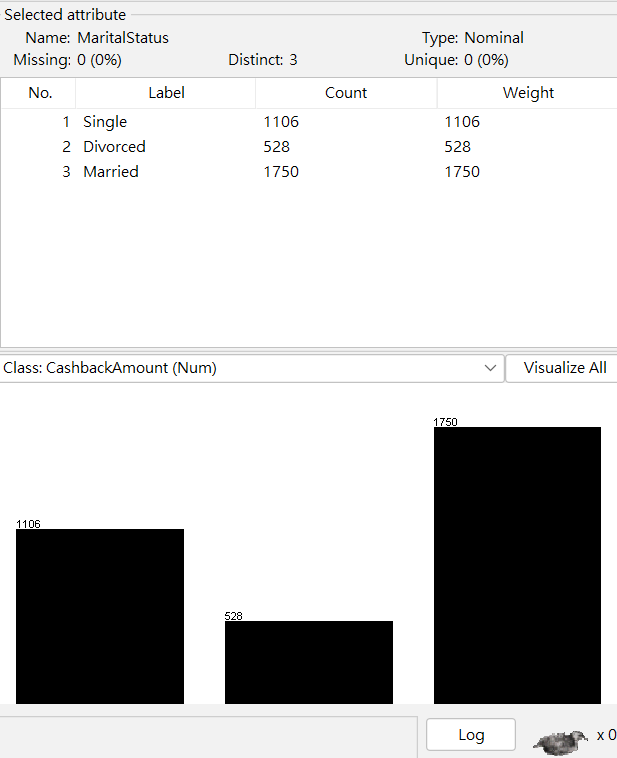
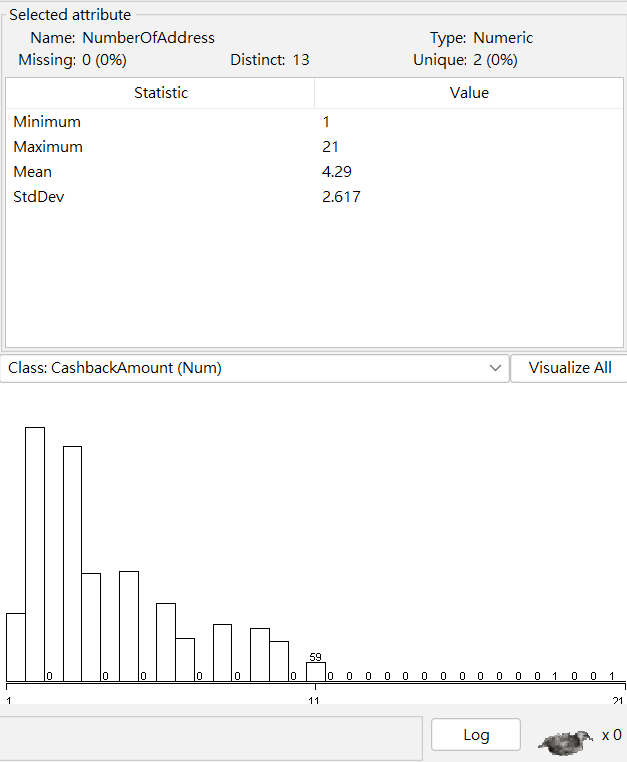
PreferredPaymentMode Gender



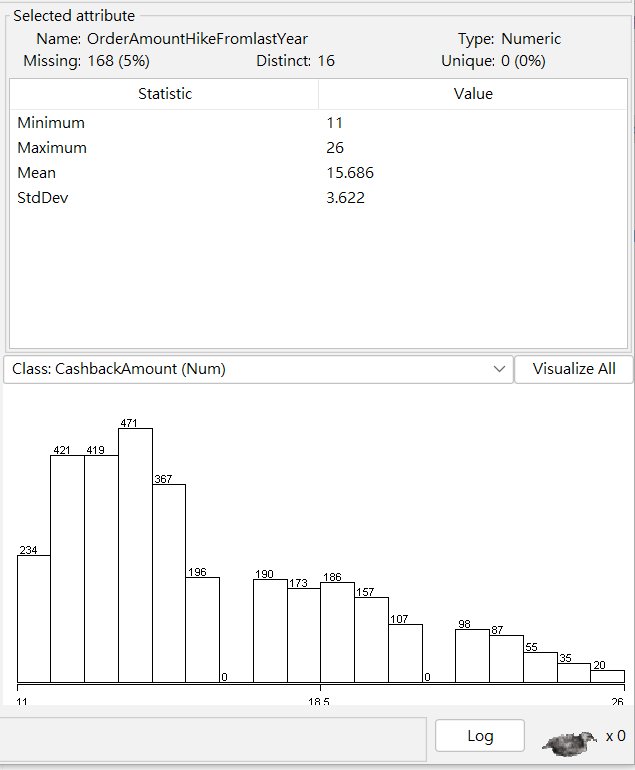
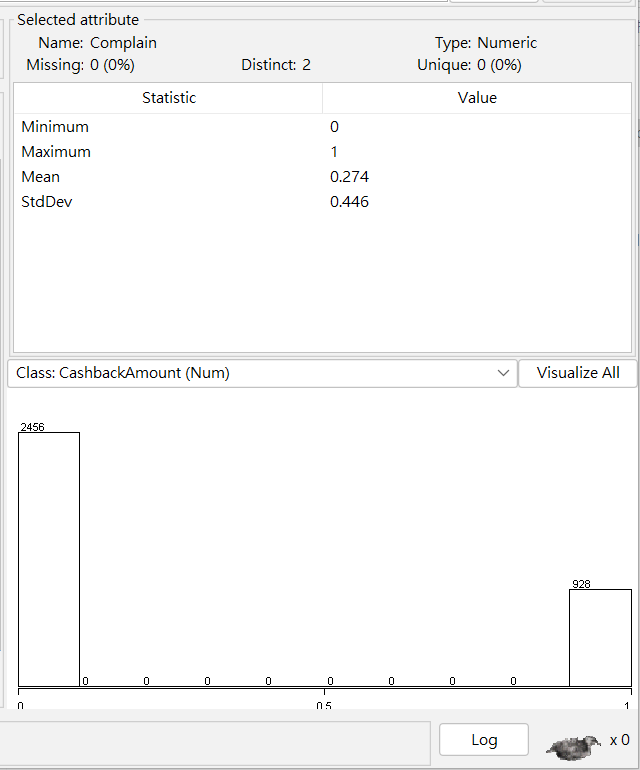
HourSpendOnApp NumberOfDeviceRegistered



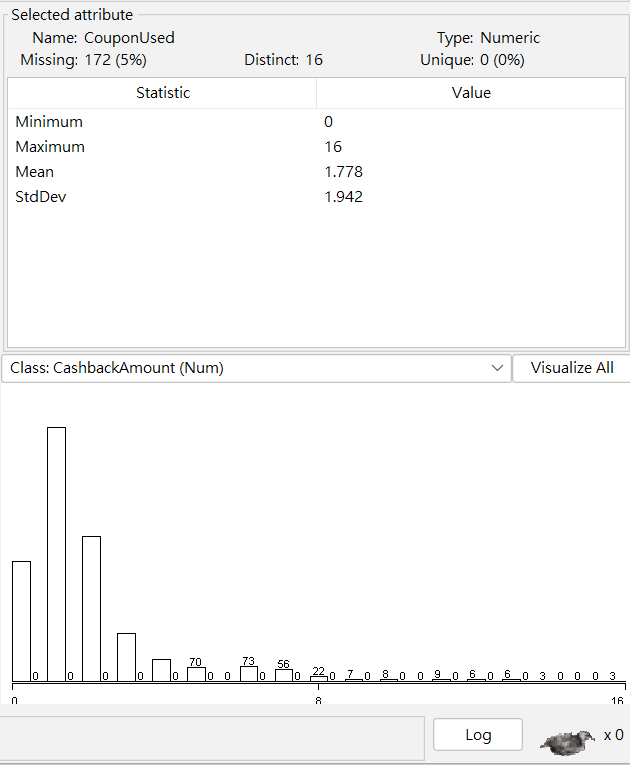
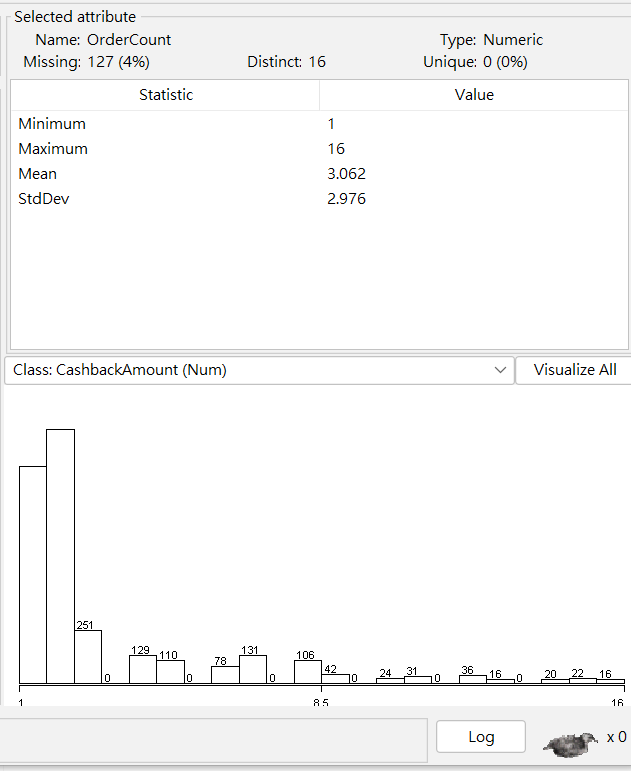
PrederredOrderCat SatisfactionScore

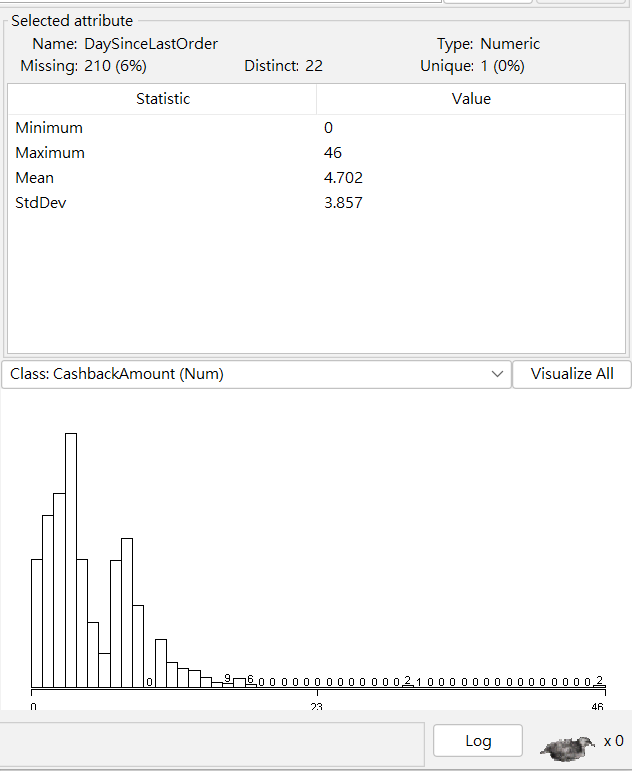
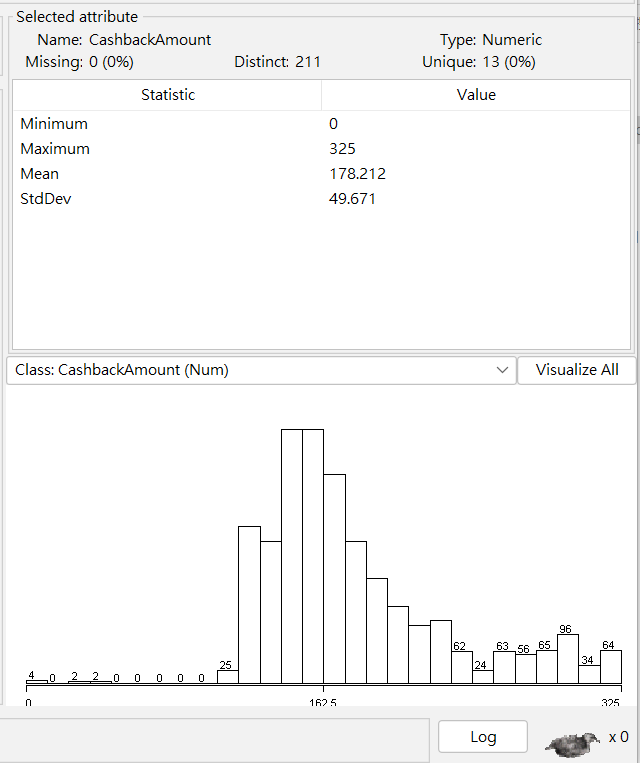


MaritalStatus NumberOfAddress



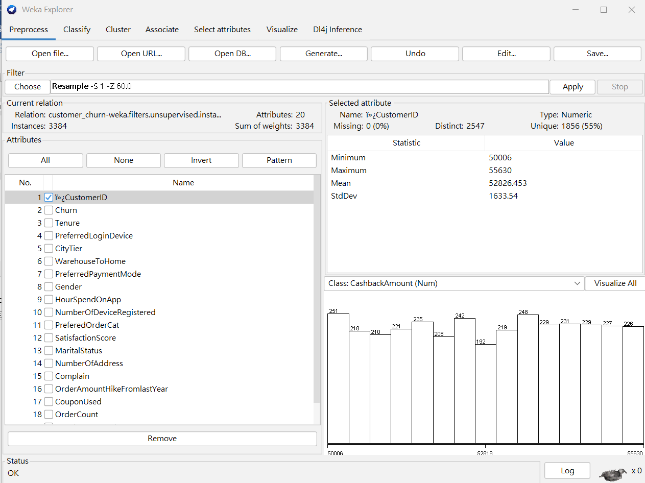
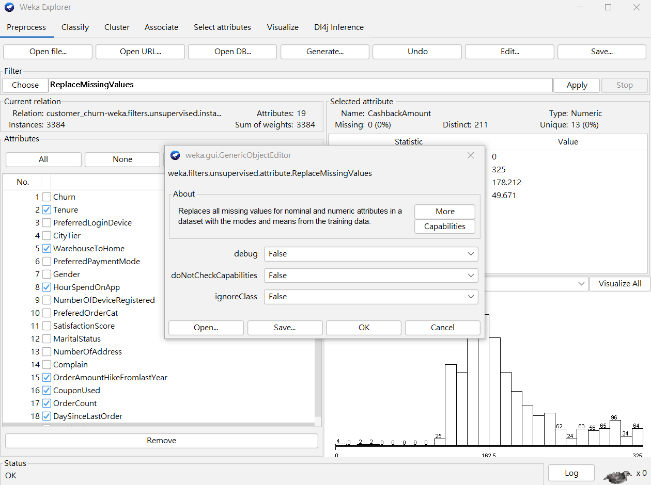
Complain OrderAmountHikeFromlastYear

CouponUsed OrderCount

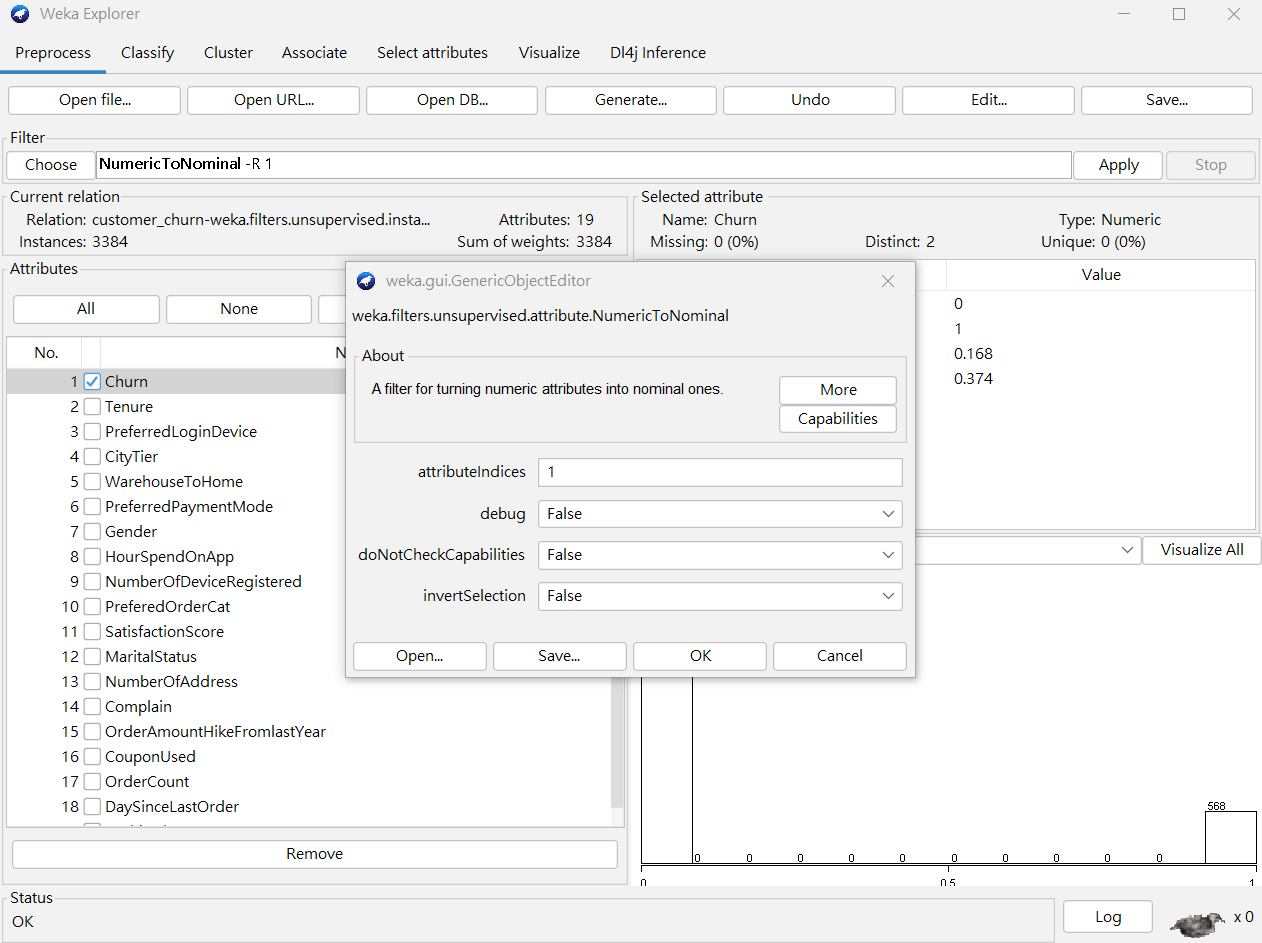


DaySinceOrder CashbackAmount

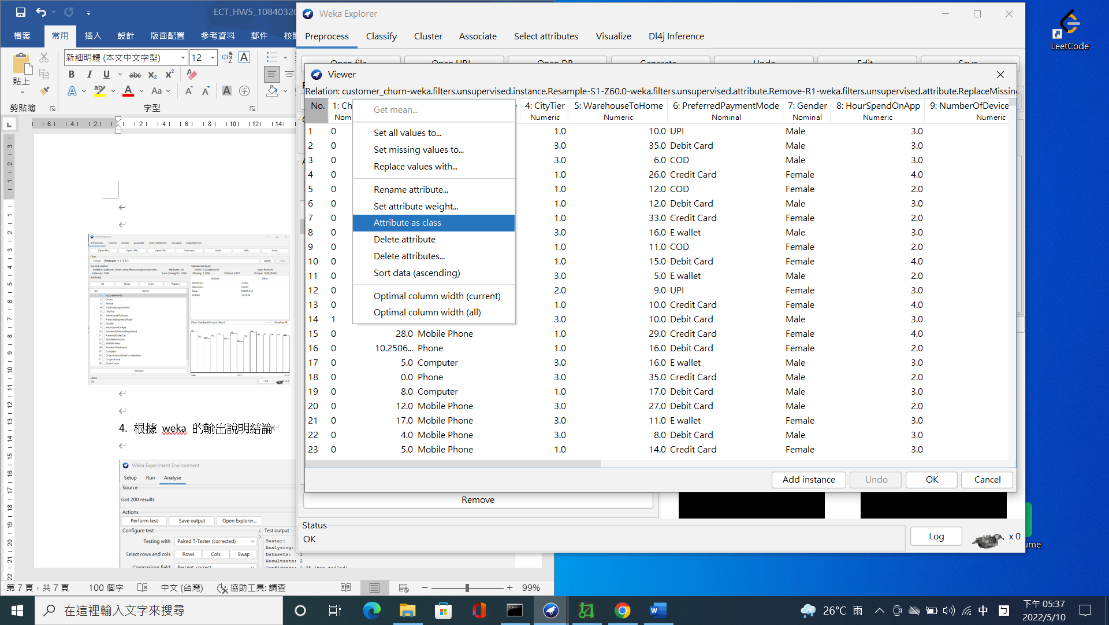
3. 資料前處理，並以 repeated 10 folds cross-validation (重複 10 次) Paired t-test 比 較 Logistic Regression 及 SVM 模型



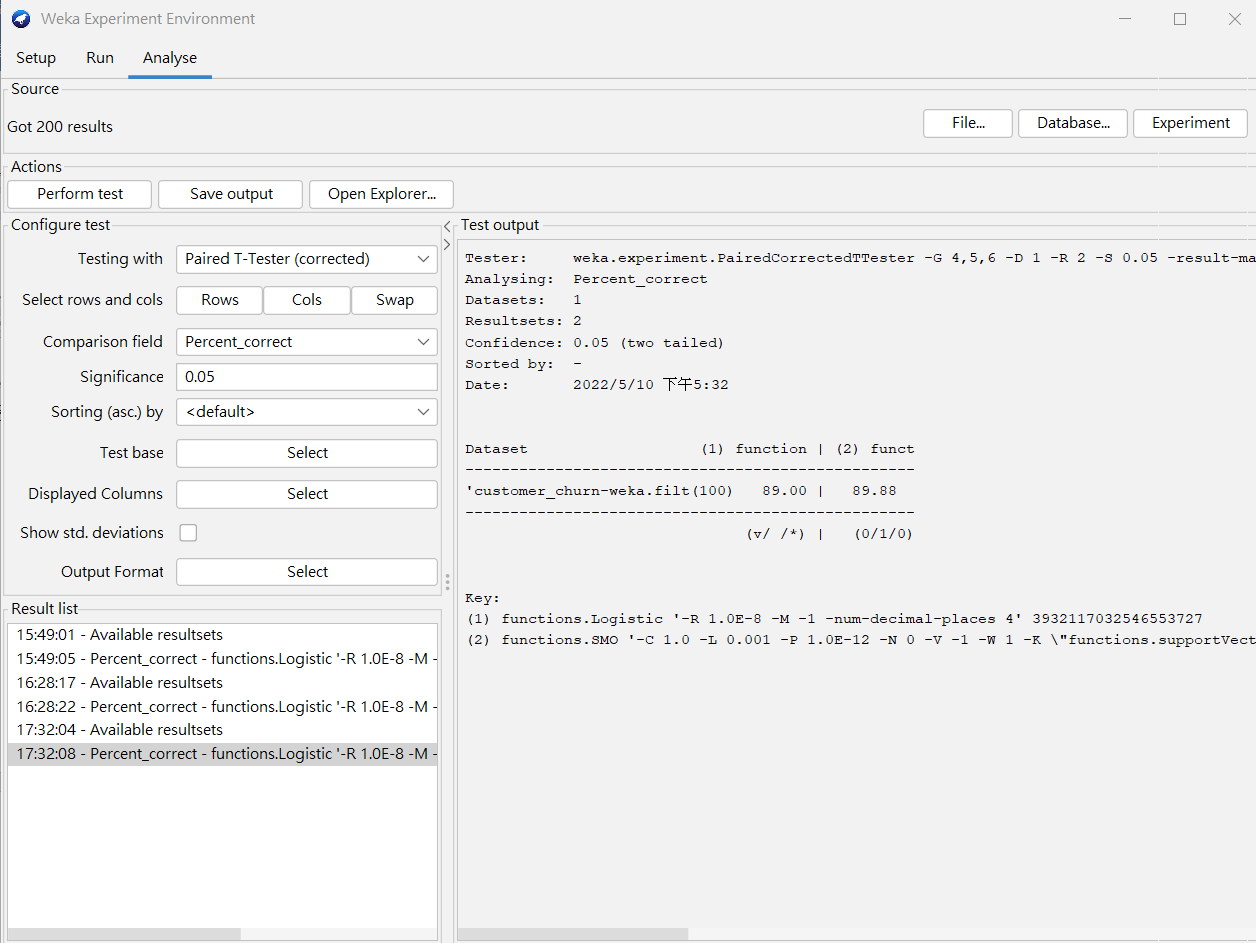
刪除CustomerID 填補空值



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