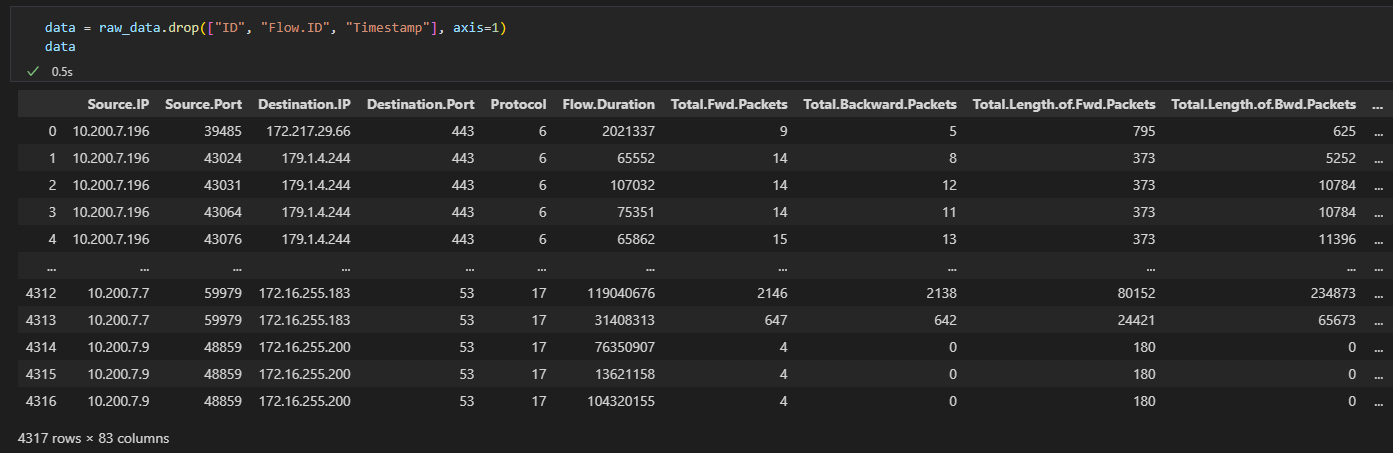
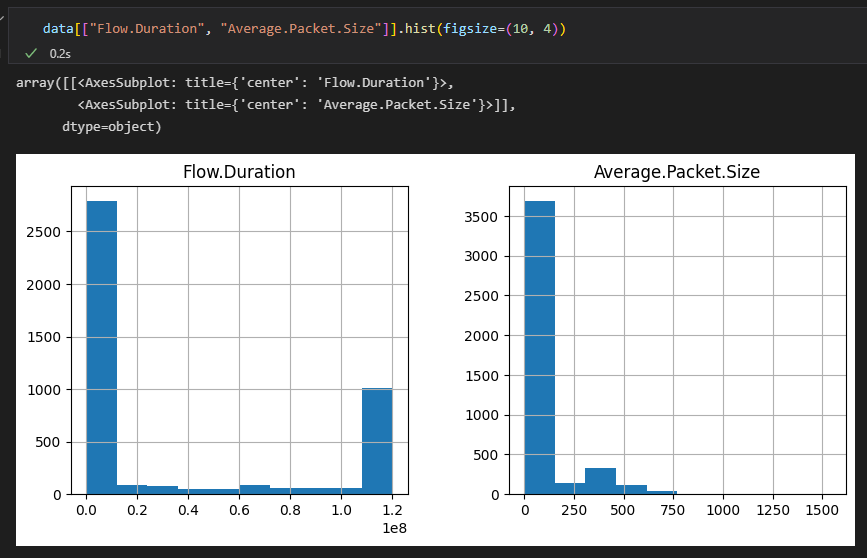
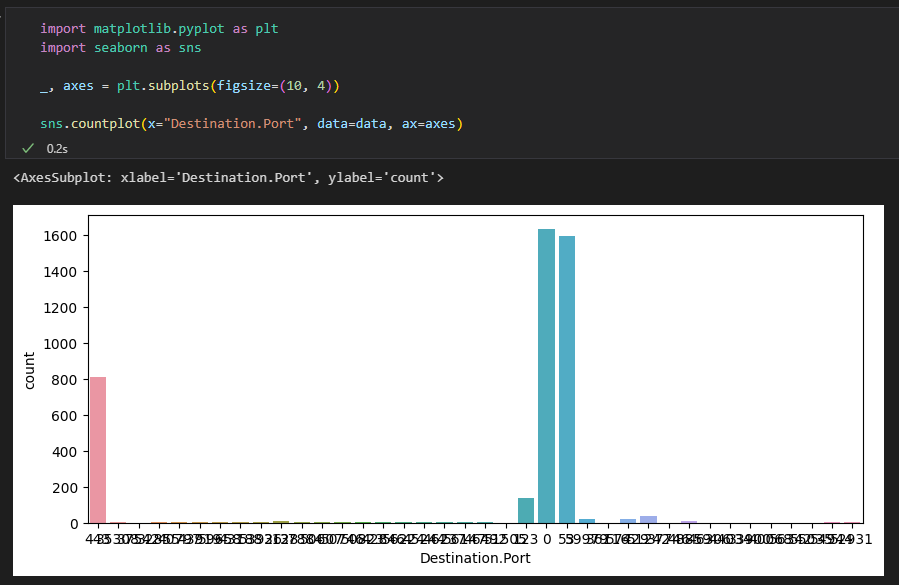
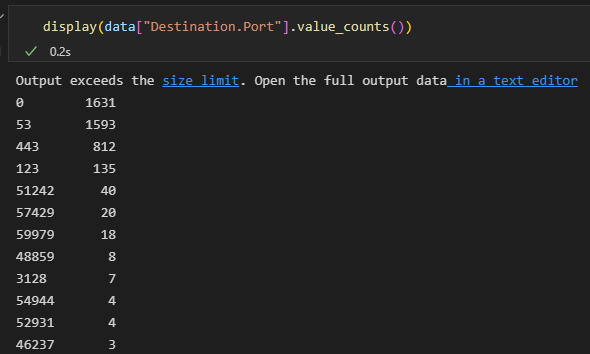
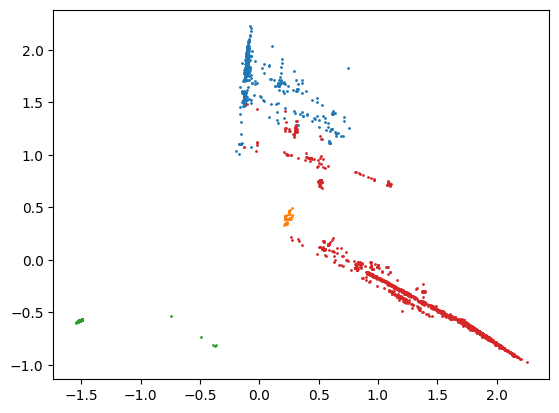
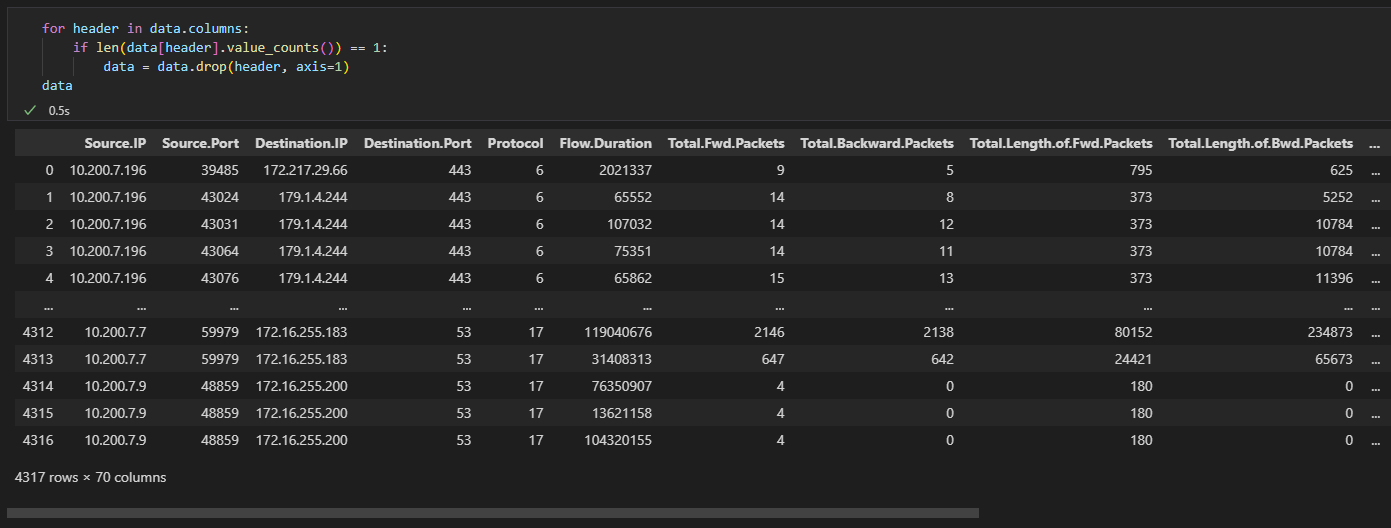
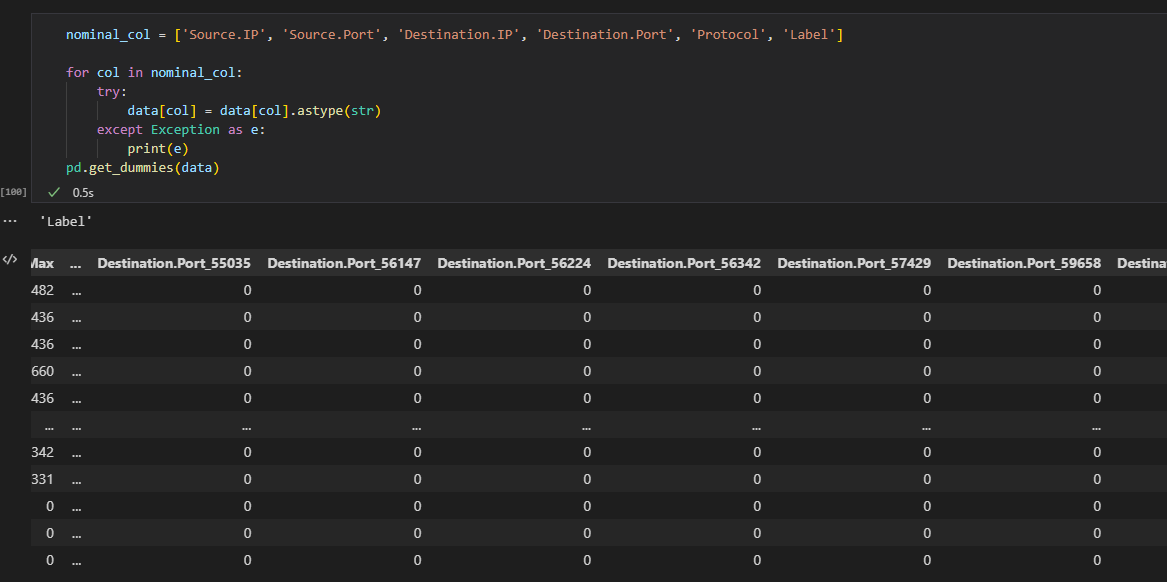
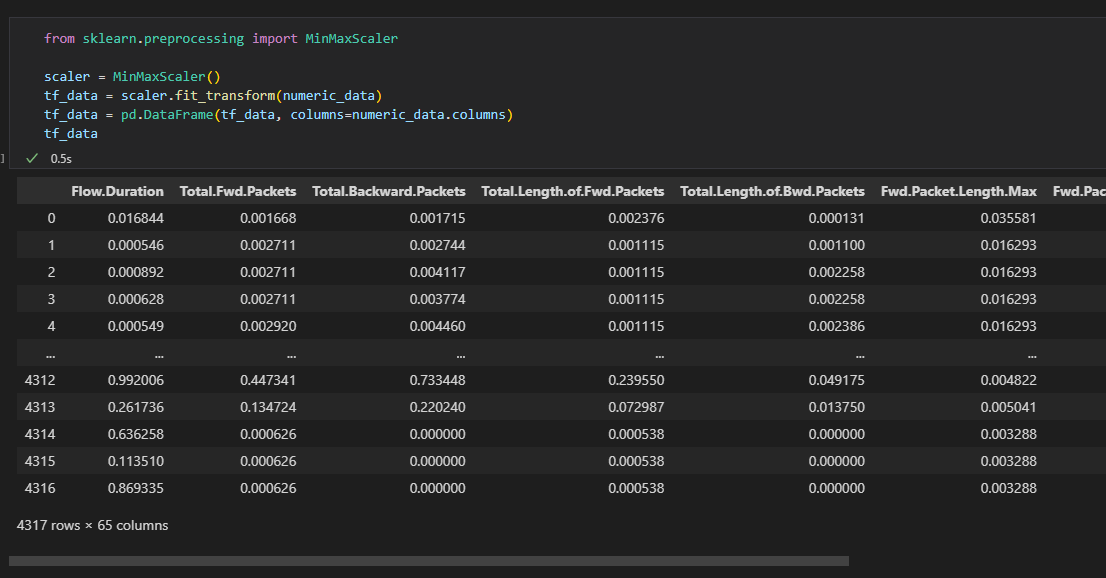
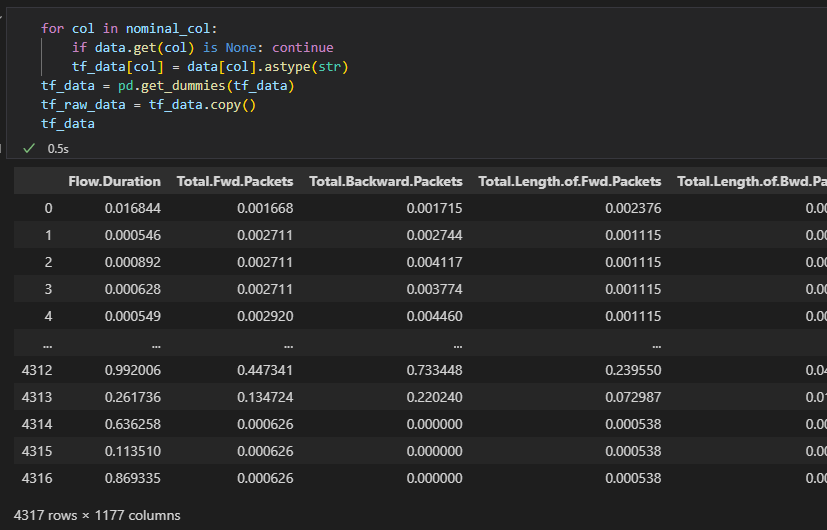
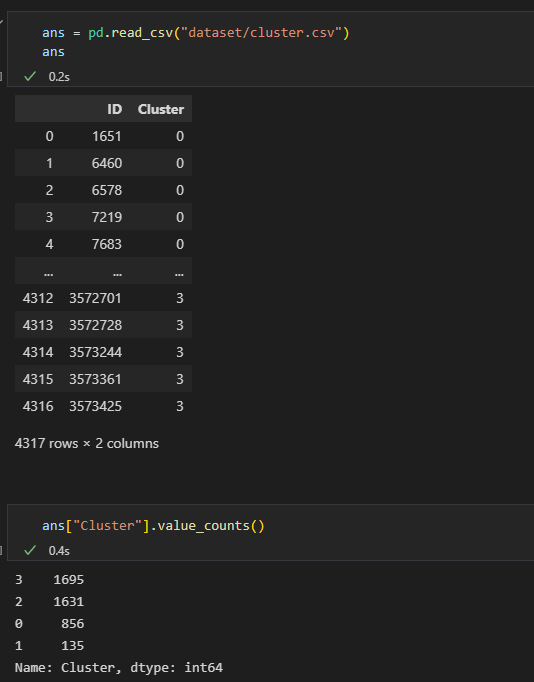
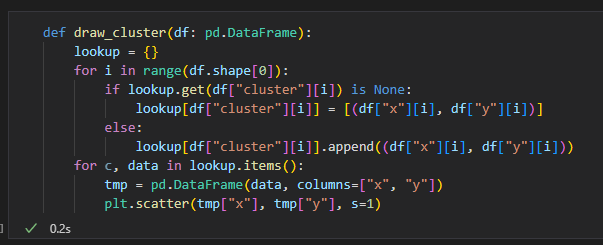
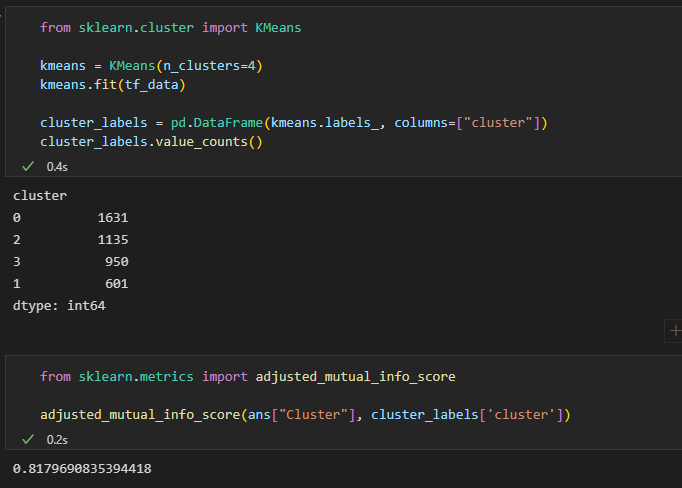
MLFN Lab2 Report

資科工一 游庭瑋

1. Data processing:  
   首先使用pandas讀取csv資料集，會發現有部分欄位為ID等不重複資料，因此必須先將其丟棄  
   
2. Data visualization:  
   這裡選用一些比較顯著的特徵來視覺化  
     
   從Flow Duration可以看出，絕大部分的Flow時間佔用都是比較少的，而從average packet size也可以看到，絕大多數的封包都小於250bytes  
     
   從Destination Port這邊則可以看出，大部分的Flow也都集中在特定的幾個Port上。並從value\_counts函數可以看出，這些Flow主要的port如下:  
     
   另外，就Protocol來統計則有以下結果  
     
   最後將所有numerical attribute正規化並將nominal attribute做one-hot-encoding後，使用PCA分析，將維度降至二維後比較cluster.csv中的正確答案可以得到下圖:  
     
   因此可以看出，此問題是可以被妥善分群的。
3. Feature engineering  
   在前面讀取資料時有發現到，有部分的attribute對於整個資料集的數據都相同，因此將有符合此特性的欄位皆刪除。留下共70個欄位。  
   接著將所有非數值資料做one-hot-encding，目前共有1177個欄位。  
     
   接著對丟棄所有非數值資料做Min Max Scaling，方便後面分群的距離計算  
     
   將以上正規化後的數值欄位及one hot encoding完的非數值欄位合併，即得到能夠直接分群用的資料集  
     
   同時匯入cluster.csv作為參考  
   
4. Clustering:  
   先定義好用來繪製分群結果的函數  
   
   1. KMeans  
        
      分群結果及視覺化  
      