**Lab02-**Mining Frequent Itemsets and Association Rules

**Data Mining -Term I /2020-2021**

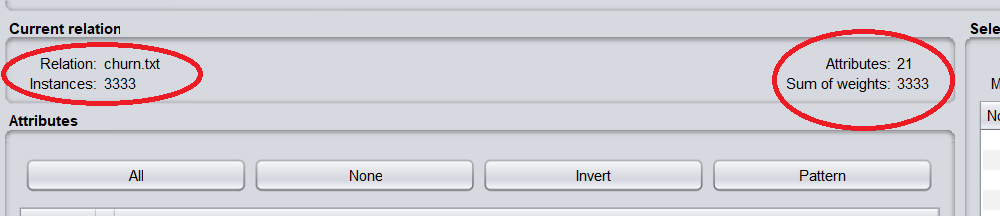
**Thông tin thành viên**

|  |  |
| --- | --- |
| **Họ Tên** | **MSSV** |
| Trần Ngọc Tịnh | 18120597 |
| Nguyễn Ngọc Năng Toàn | 18120600 |

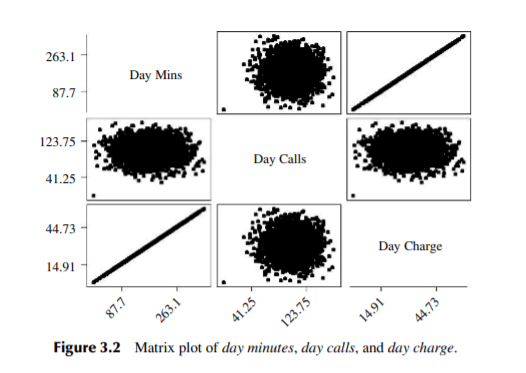
**I Báo cáo Data 1 :**

**1 Nhận biết Dữ liệu ban đầu thông qua file Mô tả dữ liệu**

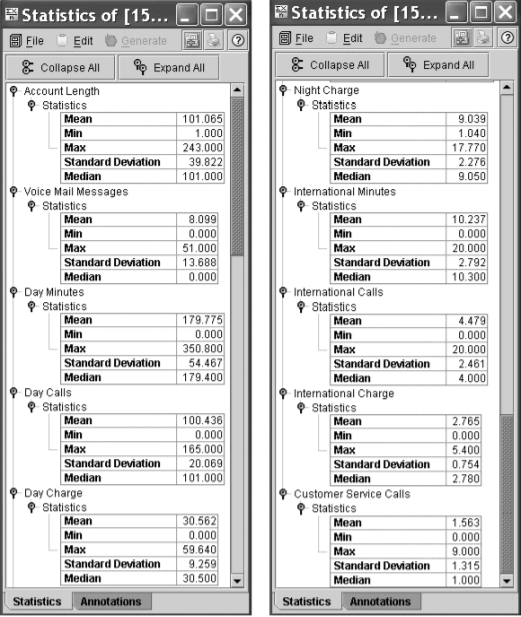
***Tập dữ liệu churn.txt có 21 thuộc tính & 3333 mẫu***

****  **+State**: Categorical .Mô tả tên 50 tiểu bang và quận của Columbia  
  
 **+Account length:** Integer-valued .Tài khoản đã hoạt động được bao lâu   
  
 **+Area code:** Categorical . Mã vùng   
  
 **+Phone number**: Categorical . Số điện thoại của khách hàng -> Đại diện cho ID của khách hàng  
  
 **+International Plan**: Dichotomous Categorical. Phân loại nhị phân có 2 giá trị của thuộc tính “ Có” Và “Không “ tham gia   
  
 **+VoiceMail Plan:** Dichotomous Categorical. Phân loại nhị phân có 2 giá trị của thuộc tính “ Có” Và “Không “ tham gia   
  
 **+Number of voice mail messages:** Interger-Valued . Số lượng thư thoại  
  
 **+Total day minutes:** Continuous . Tổng số phút khách hàng sử dụng trong ngày  
  
 **+Total day calls:** Integer-valued . Tổng số cuộc gọi trong ngày   
  
 **+Total day charge:** Continuous . Tổng số phí trong ngày  
  
 **+Total evening minutes:** Continuous . Tổng số phút khách hàng sử dụng trong tối   
  
 **+Total evening calls**: Integer-valued . Tổng số cuộc gọi trong tối  
  
 **+Total evening charge:** Continuous . Tổng số phí trong tối  
  
 **+Total night minutes:** Tổng số phút khách hàng sử dụng ban đêm  
  
 **+Total night calls:** Integer-valued . Tổng số cuộc gọi ban đêm  
  
 **+Total night charge:** Continuous . Tổng số phí sử dụng ban đêm  
  
 **+Total international minutes:** Tổng số phút khách hàng sử dụng liên lạc quốc tế  
  
 **+Total international calls:** Tổng số cuộc gọi quốc tế   
  
 **+Total international charge:** Tổng số phí cuộc gọi quốc tế  
  
 **+Number of calls to customer service**: Integer-valued . Số cuộc gọi đến dịch vụ khách hàng

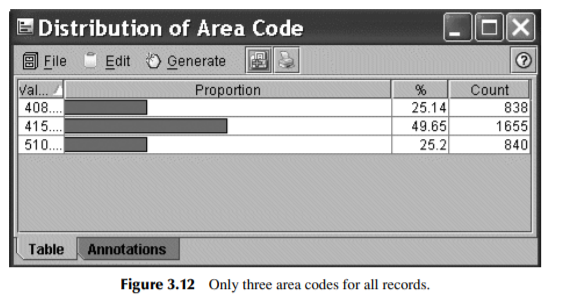
**2 Phân tích các thuộc tích của dữ liệu để Tiền Xử Lí dữ liệu**

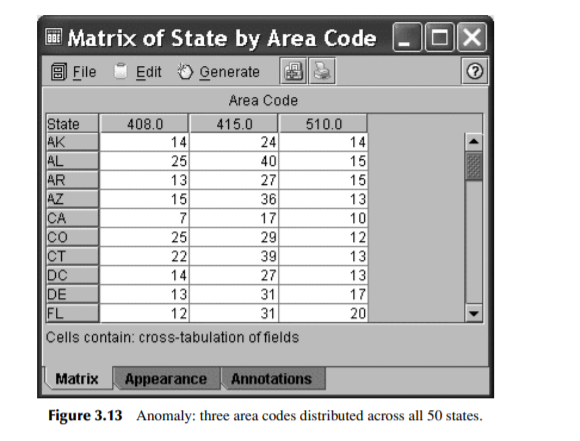
Đồ thị biểu thị mối quan hệ giữa 3 thuộc tính Day Mins , Day Calls , Day Charge  
Dựa vào độ thị ta thấy ****không có bất kì mối quan hệ nào giữa số phút trong ngày và cuộc gọi trong ngày hoặc giữa các cuộc gọi trong ngày với phí trong ngày   
Mặt khác : Có một mối quan hệ tuyến tính giữa số phút trong ngày và phí trong ngày  
Phí = 0,000613 + 0.17 \* Phút

***⇨ Vì phí trong ngày tương quan với số phút trong ngày ,chúng ta sẽ loại bỏ 1 trong 2 biến ( ở đây em loại bỏ thuộc tính Day Charge ) .Các thuộc tính tương tự như tối , đêm , quốc tế cũng tương tự nên tiếp tục loại bỏ thuộc tính Eve Eve Charge','Night Charge','Intl Charge.Chúng ta đã loại được 4 thuộc tính***

Các số liệu tóm tắt các thuộc tính Numeric như  
Mean,Min,Max,SD,Median  
  
Từ bảng ta có thể hình dung sơ lược dãi phân bố của các thuộc tính Numeric  
Vd:  
- Thời gian tài khoản dài nhất là 234 và thấp nhất là 1,Thời gian trung bình của 3333 thuộc tính là 101  
-Tin nhắn thư thoại trong ngày ít nhất là 0,nhiều nhất là 51.Đặc biệt có giá trị Median=0.Nghĩa là có ít nhất một nữa các khách hàng có thư thoại là 0

-Các thuộc tính như Day Minutes,Day Calls,Day Charge,Night Charge,Int Minutes,Int Calls,Int Charge,Customer Service Calls có giá trị Mean khá gần với giá trị Median -> Phân bố các giá trị đều  
 **Tiếp đến chúng ta sẽ phân tích về các thuộc tính có kiểu là Categorical là State và Area Code**

  
Có kiểu dữ liệu là số nhưng em lại xếp nó vào Categorical là bởi vì mục đích chính của nó là dùng để phân loại các vùng địa lí.Điều kì lạ ở đây là bộ giá trị của Area Code chỉ gồm có 3 giá trị đó là ( 408 , 415 , 510 ) và theo thông tin miêu tả từ file mô tả tệp tin thì 3 giá trị này đều thuộc về Califonia

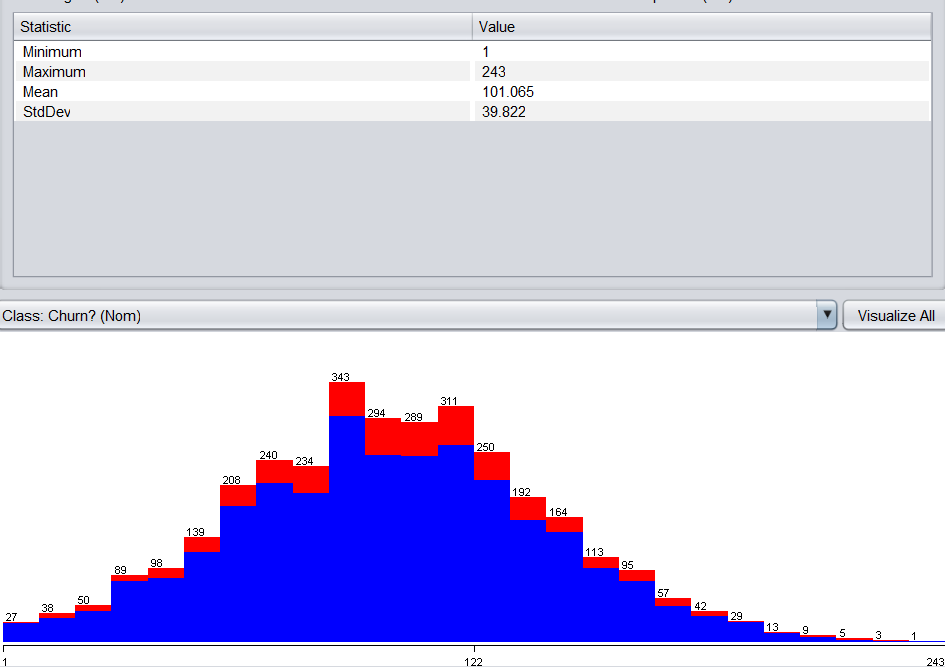
Tuy nhiên theo hình ở phía dưới đây ba vùng được phân bố nhiều hơn hoặc ít hơn đồng đều trên tất cả các Tiểu Bang và Đặc khu Columbia, đây có thể là trường chỉ chứa các giá trị xấu.Do đó nếu chúng ta đi xa có thể không giúp làm đầu vào cho các mô hình tiếp theo -> Có thể gây ra lỗi

***⇨ Kết Luận : Quyết định loại bỏ 2 thuộc tính State ( Mô tả 50 tiểu bang) , Area Code vì đây là 2 thuộc tính không bình thường như đã phân tích ở trên***

**3 Định nghĩa các khái niệm phân cấp**

Các thuộc tính Numeric đã cho thì em thấy việc sử dụng phân cấp Binning là hợp lí nhất (also called banding)

+ Account Length

Nhìn vào sự phân bố thì chúng ta có thể thấy giá trị Min là 1,Max là 243 và Mean là 101,065.Đồ thị có sự phân bố khá đều nên em quyết định gom các thuộc tính này về 3 giá trị đó là :

- Short ( 0->100 )

- Middle Length ( 100->100 )

- Long ( 200->250 )

**Data**

Dữ liệu sau khi được Binning

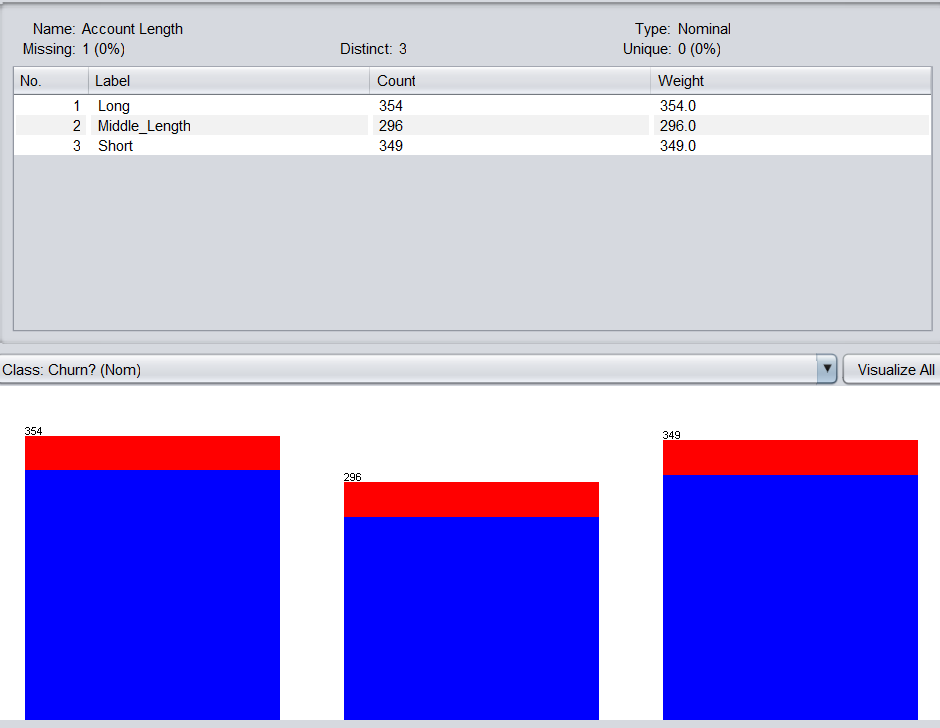
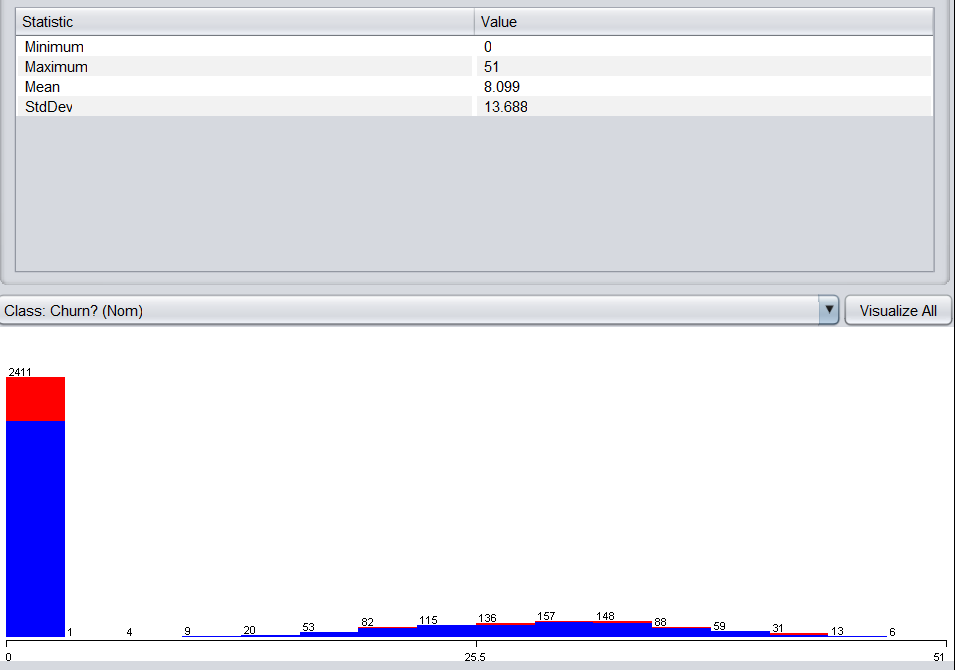
*Nhận xét : Dữ liệu khá đồng đều*

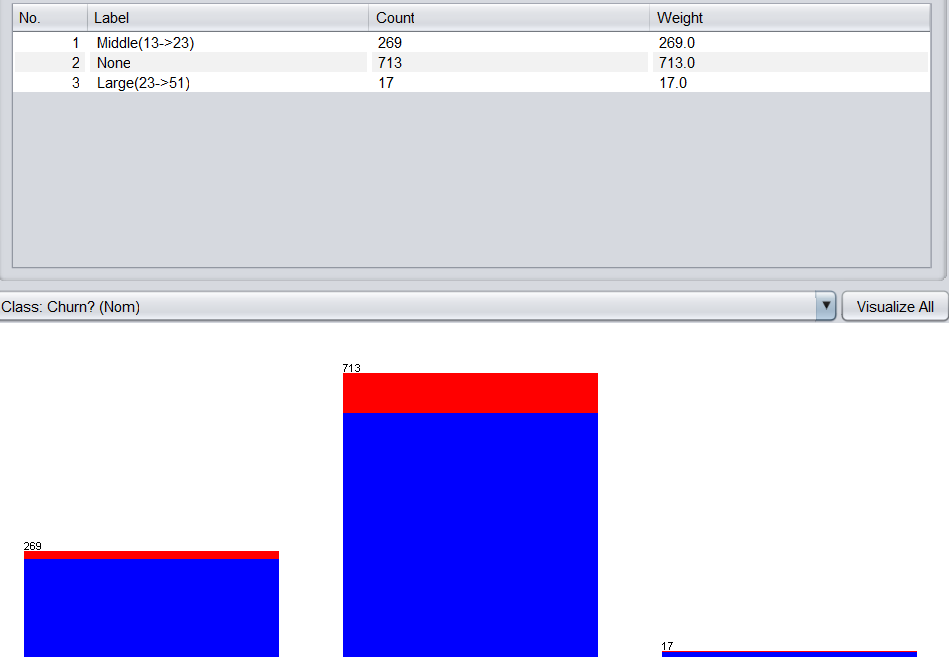
Figure Sau Khi Phân Cấp

+ VMail Message :

Nhìn vào sự phân bố ta có thể thấy giá trị 0 chiếm đa số.Từ đây có thể nhận xét rất ít khách hàng nhắn tin. Giá trị Max cũng chỉ là 51. Đồ thị phân bố không đồng đều.

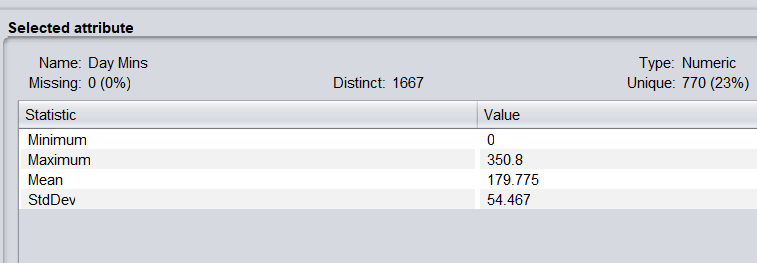
Quyết định gom đồ thị này về 3 giá trị   
+ None ( 0->13 )  
+ Middle (13 -> 23)   
+Large (23->53)

Dữ liệu thuộc tính VMail Message sau khi được Binning

*Nhận xét : Dữ liệu tin nhắn đa số nằm ở None & Middle*

Các thuộc tính còn lại như Day Mins , Day Calls , Eve Mins , Eve Calls , Night Mins , Night Calls , Intl Mins , Intl Calls , CustSery Calls như đã phân tích ở trên là các kiểu dữ liệu phân bố khá bình thường và ổn định nên mỗi kiểu dữ liệu em sẽ dùng rang(min,max) và chia làm 4 phần -> 4 cấp độ khác nhau.

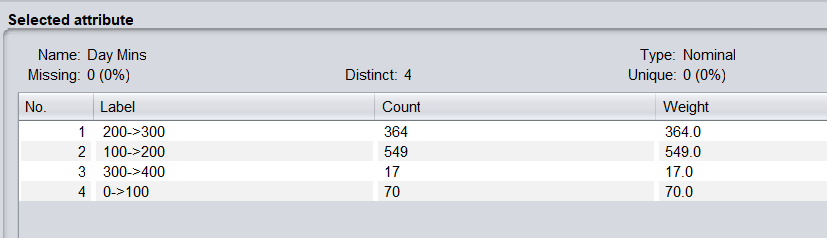
Ví dụ



Trước khi Binning

Min : 0  
 Max : 350   
 -> Chia làm 4 phần

0 -> 100  
100 ->200  
200->300  
300->400



Sau Khi Binning

**II Code**

Cú pháp để chạy code python3 main.py churn.txt data.txt 2000   
Trong đó : main.py : Tên file python  
 churn.txt: File Input  
 data.txt : File Output  
 2000 : Có ý nghĩa là sẽ lấy tập dữ liệu con gồm 2000 dòng đầu tiên

Trong phần Code em sử dụng

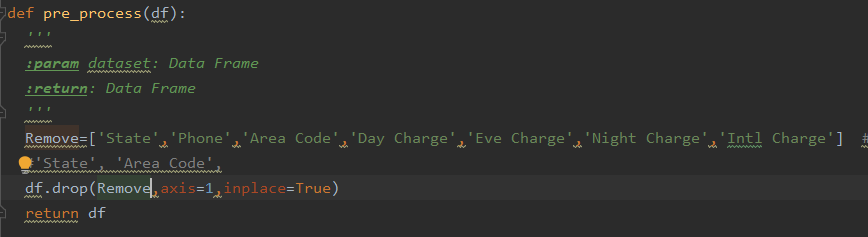
+ Thư viện Pandas

- Pandas dùng để đọc , ghi dữ liệu   
 - Binning các thuộc tính

+ 3 Hàm

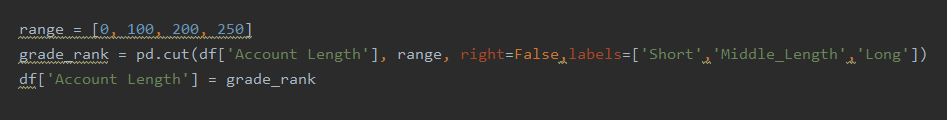
* def pre\_process(df): Dùng để tiền xử lí ( Loại bỏ các thuộc tính không cần thiết )
* def hierarchies(df): Phân cấp các thuộc tính
* def main(fileinput,output,nrows): Đọc, ghi, xử lí dữ liệu

**1 Hàm pre\_process(df)**



Hàm Pre\_Process nhận tham số là 1 Data Frame và trả về một Data Frame  
Hàm sẽ giúp chúng ta xóa các cột không có tác động vào Churn mà chúng ta đã phân tích ở trên thông qua Method drop.

**2 Hàm** hierarchies(df):

  
Hàm hierarchies nhận tham số là 1 Data Frame và trả về một Data Frame   
Hàm sẽ giúp chúng ta phân cấp ( Binning ) thông qua Method cut của thư viện Pandas  
Chúng ta sẽ gán các giá trị trong mảng Range để lưu các đoạn giá trị và đặt các đoạn đó vào Labels

**3 Hàm** main(fileinput,output,nrows)

Hàm main sẽ nhận vào 3 tham số   
fileinput : là file chúng ta chọn dùng để thực hiện.Ở đây là churn.txt  
output : là file chúng ta chọn để xuất dữ liệu sau khi tiền xử lí và phân cấp  
nrows : là số nguyên dương, có ý nghĩa chọn ra nrows hàng đầu tiên để xử lí ( tập con của churn.txt )