Python_exercise1

1) The format of the column "Time" is not correct. Replace all the '.' by ':' in time and the format should be hh/mm/ss dd/mm/yyyy. Store the new format in a new column called "time_new".

2) Find and display the different values in the column "Level", "Firewall", "User", "Service" and its distribution.

3) Based on the column Service, if it is HTTPS and the Sent Bytes or Received Bytes are greater than 20000 set a flag as "a_000x8". If Service has any other value and the Sent Bytes or Received Bytes are greater than 10000 set a flag as "v_000x2".

4) How manu unqiue devices are there in the dataset?

5) The field ReportToManager has CN="XXXXXXX" which describes the user's manager. Find the manager to whom most users report to.

6) Extract the fist component from the field enName and store is as a new column named enName1

Python_exercise2

1) Extract the fist component from the field dcName (wsaddc000XX should be extracted) using RE and store as a new column named dcName1

2) Each user is described in the field memberOf by CN="XXXXXXX". Derive the user name for each row of the dataset and store it as a new column.

3) Find the total logonCount for each user and raise set a flag of 1 if the total is greater than 50 and 0 otherwise.

4) Find and display the different type of accountExpires for each user and the percentage for each value.

5) pwdLastSet, lastLogon, lastLogonSynced are used to display the date and time but the format is not correct. Format the dates in each of the column to "hh:mm:ss AM,PM, dd/mm/yyyy Day" and store it as new column with original column name concatenated with "_formated". Try to incoporate the use of loops to reduce the number of lines of code.

6) Find the total badPwdCount for each user and raise set a flag of 1 if the total is greater than 80% when comared with logonCount.

7) The field ReportToManager has CN="XXXXXXX" which describes the user's manager. Find the manager to whom most users report to. Find the value distribution of column dcName1 for the top manager.

Python_exercise3

1) Extract the component wsaddc000XX from the field dcName and store the last two digits XX in a new column.

2) Each user is described in the field memberOf by CN="XXXXXXX". Derive the type of user for each row (wether c or d account)

3) Find the total logonCount for each user and raise set a flag of 1 if the total count is even and 0 if count is odd.

4) The users in the dataset are d789451,d951754,d321745,d986574,d327416,c456789,d456785,d123456,d789465,d321654,c654875,c999999,c587469,c195753,c465789. For each of the type of users (c or d type) find the ratio of total badPwdCount to logonCount. Store the result in a new dataframe.

5) pwdLastSet, lastLogon, lastLogonSynced are used to display the date and time but the format is not correct. Format the dates in each of the column to "hh:mm:ss AM/PM, dd/mm/yyyy Day" and store it as new column with original column name concatenated with "_formated". Try to incoporate the use of loops to reduce the number of lines of code. Pay attention to the case of the letters.

6) The field ReportToManager has CN="XXXXXXX" which describes the user's manager. For each manager find the ratio of the user accounts (c or d type) that reports the Manager.