**Aim**: find the difference between two csv file.

Input:

* File1: Tap file
* File2: Month file

Programs (to be run in order):

* Prog1.py: this program is used to filter the data of month file. It will generate 6 output files i.e. monthname\_na\_data\_single.csv, monthname\_dup\_with\_na.csv, monthname\_dup\_without\_na.cvs, monthname\_not\_null.csv, monthname\_merge.cvs and monthname\_finalfile.csv
* Prog2: this program is used to convert the date format of tap and month file to use it further for comparison and sort the month file.
* Prog3: this program is used for comparison of monthname\_finalfile.csv (generated by Something.py) to compare with tap file by using the set properties, A-B is used to find the set difference, where A and B are the csv file. A = Tap file and B = Monthname\_finalfile. A-B gives the data present only in file A but not in file B and vice versa.

**Prog1 .py** : use pandas and csv module

Input: name of the month file to be clean.

Output: returns the final file.

**Prog2.py** : use pandas, csv and datetime module

df.Flt\_Dt = pd.to\_datetime(df.Flt\_Dt, format='%d/%m/%Y')

the above line use the datetime module to compare the date format of tap file and convert it into desired form. If the file is already in the desired format then the prog will show an error.

df=pd.read\_csv(nof1+"\_finalfile.csv")  
df=df.rename(columns={ "Flight Departure Date" : "Flight\_Departure\_Date"})  
df.Flight\_Departure\_Date = pd.to\_datetime(df.Flight\_Departure\_Date)  
df.to\_csv(nof1+"\_finalfile.csv",index=False)

the above code is use to convert the date format of month file in desired way.

new=open(nof1+'\_finalfile.csv','r')  
new=''.join([i for i in new]).replace("AI9","9I0")  
x=open(nof1+'\_finalfile.csv','w')  
x.writelines(new)  
x.close()

the above code is used to convert the Flight Number of “AI9XXX” to “9I0XXX” format.

df = pd.read\_csv(nof1+'\_finalfile.csv')  
df.sort\_values(['Flight Number','Flight Departure Date'],axis = 0, ascending = True,  
 inplace = True)  
print(df)  
df.to\_csv(nof1+'\_finalfile.csv',index=False)

the above code is used to sort the month file w.r.t ‘Flight Number’ and ‘Flight Departure Date’.

**Prog3.py**: this program is used for comparison

Input file: tap file i.e. (Tap\_Apr19)- **File A** and month file i.e. (Apr2019\_finalfile) – **File B**.

User input: user need to enter the name of month and the year (only last two digits i.e. ‘19’ for year 2019)

Output: the program will return the two-output file.

1. Tap\_Apr19\_A-B.csv
2. Apr2019\_finalfile\_B-A.csv
3. A.csv and b.csv two files will be generated in process.

nof= input("Enter the name of file ", )  
yr=input("Enter the yr(only last 2 digits) ", )

the above code is for user input for the name of file i.e. nof and year i.e. yr. the nof and yr will be use in such a way (as shown below).

df2=pd.read\_csv("Tap\_"+nof+yr+".csv")

df2 = pd.DataFrame(df2,columns=['Flt\_No','Flt\_Dt','AC\_Reg','Dep\_Stn','Arr\_Stn'])  
df2['key1'] = df2['Flt\_No'] + ','+df2['Flt\_Dt']+ ','+df2['AC\_Reg']+ ','+ df2['Dep\_Stn']+ ','+df2['Arr\_Stn']

the above code is used to concatenate the five column i.e. ‘Flt\_No, Flt\_Dt, AC\_reg, Dep\_Stn,Arr\_Stn’ and make a key column name ‘Key1’ for tap file which is stored in A.csv file.

df2.to\_csv('a.csv',index=None)

same procedure is used to concatenate the five columns for month file, which is later saved in b.csv file.

df=pd.read\_csv(nof+"20"+yr+"\_finalfile.csv")  
df = pd.DataFrame(df,columns=['Flight Number','Flight Departure Date','Aircraft Registration Number','Departure Airport Code','Arrival Airport Code'])  
df['key1'] = df['Flight Number'] +',' + df['Flight Departure Date']+ ','+df['Aircraft Registration Number']+','+df['Departure Airport Code']+','+df['Arrival Airport Code']  
#df['key1'].astype(int)  
#print(df.dtypes)  
df.to\_csv('b.csv',index=None)

the code shown below is use to compare two file for A-B.

import csv  
f1=open("Tap\_"+nof+yr+"\_A-B.csv","w")  
c1=csv.writer(f1)  
print("here start A-B ")  
A = set(pd.read\_csv("a.csv", index\_col=False,header=None)[5]) #reads the tap csv file , takes only the key column and creates a set out of it.  
B = set(pd.read\_csv("b.csv", index\_col=False,header=None)[5])#same for the month file  
C= list(A-B)  
for i in range(len(C)):  
 res=C[i].split(",")  
 print(res)  
 c1.writerow(res)  
f1.close()

in the above code A is the set created for Tap file using key column

similarly set B is created for month file.

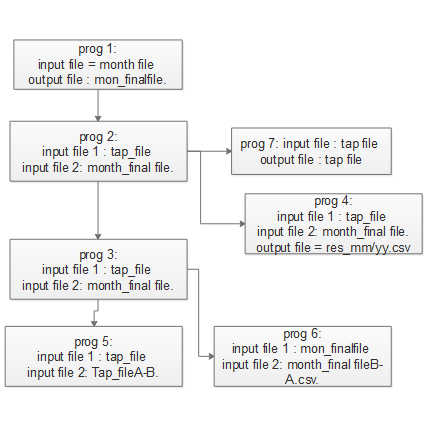
the same procedure is used for B-A comparison the result file will consist of data that is present in month file only.

print("here start B-A \n ")  
f2=open(nof+'20'+yr+'\_finalfile\_B-A.csv','w')  
c2=csv.writer(f2)  
D=list(B-A)  
for j in range(1,len(D)):  
 res\_n=D[j].split(",")  
 print(res\_n)  
 c2.writerow(res\_n)  
f2.close()

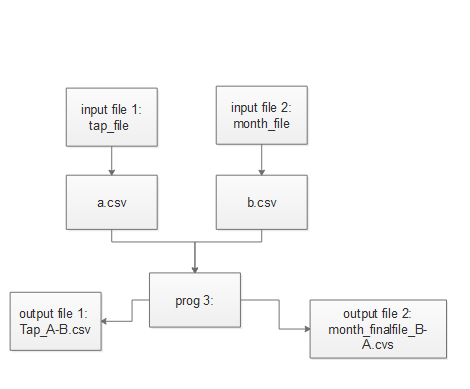
the both files created are unsorted and does not consist of headers which are included using the below code.

data=pd.read\_csv('Tap\_'+nof+yr+'\_A-B.csv')  
data.columns=['Flt\_No','Flt\_Dt','AC\_Reg','Dep\_Stn','Arr\_Stn']  
data.sort\_values(['Flt\_No','Flt\_Dt'],axis = 0, ascending = True,inplace = True)  
data.to\_csv('Tap\_'+nof+yr+'\_A-B.csv',index=False)  
#print(data)  
  
data1=pd.read\_csv(nof+"20"+yr+"\_finalfile\_B-A.csv")  
data1.columns=['Flight Number','Flight Departure Date','Aircarft Registration Number','Departure Airport Code','Arrival Airport Code']  
data1.sort\_values(['Flight Number','Flight Departure Date'],axis = 0, ascending = True,inplace = True)  
data1.to\_csv(nof+"20"+yr+"\_finalfile\_B-A.csv",index=False)  
#print(data1)

**flow chart:**



**Flow chart (prog 3):**



**Flow chart (prog 3):** to concatenate multiple columns and generate key column. The key column is later used to find the difference between two files A and B.

