Insurees_Families_Locations

October 2, 2020

[1]:

```
# import necessaries modules:
import numpy as np
import datetime
import pandas as pd
import gc
```

0.1 Step 1 : Reading tblInsurees and tblFamilies and concatenate them

0.1.1 Step 1.1 : Reading tblInsurees

Reading the file, selecting the necessary fields and concatenate it to the previous table

[2]:

```
# read the csv file and selecting the necessary columns
filename = 'openIMIS csv/insurees2020.csv'
cols = ['InsureeID','InsureeUUID','FamilyID','CHFID','HFID',\
         'Relationship','IsHead','Marital',
        'DOB','PhotoDate','ValidityFrom','ValidityTo',\
        'TypeOfId','Gender']
df_insuree_raw = pd.read_csv(filename,low_memory=False,usecols=cols,\
                             parse_dates =␣
 ↪['PhotoDate','DOB','ValidityFrom','ValidityTo'])
df_insuree_raw = df_insuree_raw.iloc[:-2,:]

# rename columns name as there will be several columns ValidityFrom and␣
 ↪VaidityTo
df_insuree_raw.rename(columns = {'ValidityFrom': 'InsureeValidityFrom',\
                                    'ValidityTo': 'InsureeValidityTo'}, inplace␣
 ↪= True)

memStats_insurees = (df_insuree_raw.memory_usage()/1024/1024).sum()
shape_insurees = df_insuree_raw.shape
```

0.1.2 Step 1.2: Reading tblFamilies

[3]:

```
# read the csv file and selecting the necessary columns
filename = 'openIMIS csv/families2020.csv'
cols = ['FamilyID', 'InsureeID', 'LocationId','FamilyUUID',\
        'Poverty','ValidityFrom','ValidityTo']
df_families_raw = pd.read_csv(filename,low_memory=False,usecols=cols,\
                             parse_dates = ['ValidityFrom','ValidityTo'])

df_families_raw = df_families_raw.iloc[:-2,:]

df_families_raw['FamilyID'] = df_families_raw['FamilyID'].astype(float)

# rename columns name as there will be several columns ValidityFrom and␣
 ↪VaidityTo
df_families_raw.rename(columns = {'ValidityFrom': 'FamilyValidityFrom',\
                                 'ValidityTo': 'FamilyValidityTo'}, inplace =␣
 ↪True)

memStats_families = (df_families_raw.memory_usage()/1024/1024).sum()
shape_families = df_families_raw.shape
```

0.1.3 Step 1.3 Concatenation of tblInsurees and tblFamilies

[4]:

```
df_insuree_fmlies = pd.merge(df_insuree_raw,df_families_raw,on=['FamilyID'])

memStats_concat1 = (df_insuree_fmlies.memory_usage()/1024/1024).sum()
shape_concat1 = df_insuree_fmlies.shape
```

0.2 Step 2. Read the tblLocations and concatenate to the previous dataframe

0.2.1 Step 2.1 Reading the tblLocations

[5]:

```
# read the csv file and selecting the necessary columns
filename = 'openIMIS csv/locations2020.csv'
cols =␣
 ↪['LocationId','LocationName','LocationType','LocationUUID','ValidityFrom','ValidityTo']
df_location_raw = pd.read_csv(filename,low_memory=False,usecols=cols,\
                             parse_dates = ['ValidityFrom','ValidityTo'])
df_location_raw = df_location_raw.iloc[:-2,:]

df_location_raw['LocationId'] = df_location_raw['LocationId'].astype(int)

# rename columns name as there will be several columns ValidityFrom and␣
 ↪VaidityTo
df_location_raw.rename(columns = {'ValidityFrom': 'LocationValidityFrom',\
                                 'ValidityTo': 'LocationValidityTo'}, inplace =␣
 ↪True)

memStats_locs = (df_location_raw.memory_usage()/1024/1024).sum()
shape_locs = df_location_raw.shape
```

0.2.2 Step 2.2: Concatenate the tblLocation to the previous dataframe

[6]:
```
df_insuree_fmlies_locs = pd.
 ↪merge(df_insuree_fmlies,df_location_raw,on='LocationId')

# rename columns in the dataframe
df_insuree_fmlies_locs.rename(columns = {'LocationID': 'InsureeLocationID',
                                  'LocationName': 'InsureeLocationName',
                                  'LocationType': 'InsureeLocationType',
                                  'HFID': 'InsureeHFID',
                                  'HFUUID': 'InsureeHFUUID',
                                  'InsureeID_x': 'InsureeID',
                                  'InsureeID_y': 'FamHeadInsuree'
                                 }, inplace = True)

memStats_concat2 = (df_insuree_fmlies_locs.memory_usage()/1024/1024).sum()
shape_concat2 = df_insuree_fmlies_locs.shape
```

[7]:
```
#Save data in a pkl file: (or csv)
df_insuree_fmlies_locs.to_pickle('openIMIS csv/Insurees_Fmlies_Loc2020_sel.pkl')
#df_insuree_fmlies_locs.to_csv('openIMIS csv/Insurees_Fmlies_Loc2020_sel.csv')
```

0.2.3 Summary

[8]:
```
print(f'''Summary of the concatenation process:
- tblInsurees has : {shape_insurees[0]} rows; {shape_insurees[1]} columns; \
{round(memStats_insurees,2)} MB memory consumption;
- tblFamilies has : {shape_families[0]} rows; {shape_families[1]} columns; \
{round(memStats_families,2)} MB memory consumption;
- Concatenation of tblInsurees and tblFamilies has : {shape_concat1[0]} rows; \
{shape_concat1[1]} columns ; {round(memStats_concat1,2)} MB memory consumption;
- tblLocations has : {shape_locs[0]} rows; \
{shape_locs[1]} columns ; {round(memStats_locs,2)} MB memory consumption;
- Concatenation of tblInsurees,tblFamilies and tblLocations has :␣
 ↪{shape_concat2[0]} rows; \
{shape_concat2[1]} columns; {round(memStats_concat2,2)} MB memory consumption;
''')
```

```
Summary of the concatenation process:
- tblInsurees has : 3790789 rows; 14 columns; 404.9 MB memory consumption;
- tblFamilies has : 977860 rows; 7 columns; 52.22 MB memory consumption;
- Concatenation of tblInsurees and tblFamilies has : 3790789 rows; 20 columns ;
607.35 MB memory consumption;
- tblLocations has : 10350 rows; 6 columns ; 0.47 MB memory consumption;
- Concatenation of tblInsurees,tblFamilies and tblLocations has : 3790789 rows;
25 columns; 751.96 MB memory consumption;
```