Claims_selection_VF

October 2, 2020

[1]:
```
# import necessaries modules:
import numpy as np
import datetime
import pandas as pd
import gc
```

0.0.1 Step 1.1: Reading data related to the tblClaims

[2]:
```
# pkl file related to the selected columns and rows from tblClaims:
df_claims_raw = pd.read_pickle('openIMIS csv/Claims2020_sel.pkl')

memStats_claims = (df_claims_raw.memory_usage()/1024/1024).sum()
shape_claims = df_claims_raw.shape
```

0.0.2 Step 1.2: Reading the data related to tblClaimAdmins

[3]:
```
# csv file related to the tblClaimAdmins:
filename = 'openIMIS csv/claim_admins2020.csv'

# selection of columns (the entire table has 30 columns)
cols = ['ClaimAdminId', 'HFId', 'ClaimAdminUUID',\
        'ValidityFrom','ValidityTo']

# read the csv file
df_claimadmins_raw = pd.read_csv(filename,low_memory=False,usecols=cols,\
                                 parse_dates = ['ValidityFrom','ValidityTo'])
df_claimadmins_raw = df_claimadmins_raw.iloc[:-2,:]

# rename the columns in order to have similar name as the items related dataset
df_claimadmins_raw.rename(columns = {'HFID': 'ClaimAdminHFID',\
                                     'ValidityFrom':␣
 ↪'ClaimAdminValidityFrom',\
                                     'ValidityTo': 'ClaimAdminValidityTo'},␣
 ↪inplace = True)
df_claimadmins_raw['ClaimAdminId'] = df_claimadmins_raw['ClaimAdminId'].
 ↪astype(float)
```

```
memStats_claimadmins = (df_claimadmins_raw.memory_usage()/1024/1024).sum()
shape_claimadmins = df_claimadmins_raw.shape
```

0.0.3 Step 1.3: Concatenation of the tblClaims and tblClaimAdmins (based on ClaimAdminId)

[4]:
```
# Concatenate the dataframes, based on ClaimAdminID column
df_claims= pd.merge(df_claims_raw,df_claimadmins_raw,on='ClaimAdminId')

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

[5]:
```
del [[df_claims_raw,df_claimadmins_raw]]
df_claims_raw = pd.DataFrame()
df_claimadmins_raw = pd.DataFrame()
gc.collect()
```

[5]:
```
40
```

0.0.4 Step 2.1: Reading the file related to the tblInsurees, tblFamilies and tblLocations

[6]:
```
# read the pkl file related to the tblInsurees, tblFamilies and tblLocations
df_insuree_fmlies_locs = pd.read_pickle('openIMIS csv/
 ↪Insurees_Fmlies_Loc2020_sel.pkl')

df_insuree_fmlies_locs['InsureeID']=df_insuree_fmlies_locs['InsureeID'].
 ↪astype(int)

memStats_fam_loc = (df_insuree_fmlies_locs.memory_usage()/1024/1024).sum()
shape_fam_loc = df_insuree_fmlies_locs.shape
```

0.0.5 Step 2.2: Adding the insurees, families, locations to the previous dataframe

[7]:
```
df_claim_si_concat = pd.merge(df_claims,df_insuree_fmlies_locs,on='InsureeID')

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

0.0.6 Step 3.1: Reading the file related to the tbHFs and tblLocations

[8]:
```
# read the pkl file related to the tblHFs and tblLocations
df_HF_locations = pd.read_pickle('openIMIS csv/HF_Locations2020_sel.pkl')

df_HF_locations['HFID'] = df_HF_locations['HFID'].astype(float)
```

```
memStats_HF_loc = (df_HF_locations.memory_usage()/1024/1024).sum()
shape_HF_loc = df_HF_locations.shape
```

0.0.7 Step 3.2: Adding the HFs and locations to the previous dataframe

[9]:
```
df_claim_si_concat = pd.merge(df_claim_si_concat,df_HF_locations,on='HFID')

memStats_concat3 = (df_claim_si_concat.memory_usage()/1024/1024).sum()
shape_concat3 = df_claim_si_concat.shape
```

0.0.8 Step 4.1: Reading the files related to the tblDiagnosis

[10]:
```
# csv file related to the tblDiagnosis:
filename = 'openIMIS csv/diagnosis2020.csv'
cols = ['ICDID', 'ICDCode', 'ICDName','ValidityFrom','ValidityTo']
df_diagnosis_raw = pd.read_csv(filename,low_memory=False, usecols = cols,\
                               parse_dates =['ValidityFrom','ValidityTo'] )
df_diagnosis_raw = df_diagnosis_raw.iloc[:-2,:]

df_diagnosis_raw.rename(columns = {'ValidityFrom': 'ICDValidityFrom',\
                                   'ValidityTo': 'ICDValidityTo'}, inplace =␣
↪True)


df_diagnosis_raw['ICDID'] = df_diagnosis_raw['ICDID'].astype(float)

memStats_diag = (df_diagnosis_raw.memory_usage()/1024/1024).sum()
shape_diag = df_diagnosis_raw.shape
```

0.0.9 Step 4.2: Adding the diagnosis data to the previous dataframe

[11]:
```
df_claims_final= pd.merge(df_claim_si_concat,df_diagnosis_raw,on='ICDID')

memStats_concat4 = (df_claims_final.memory_usage()/1024/1024).sum()
shape_concat4 = df_claims_final.shape
```

[12]:
```
df_claims_final.to_pickle('openIMIS csv/ClaimsPlus2020_sel.pkl')
#df_claims_raw.to_csv('openIMIS csv/ClaimsPlus2020_sel.csv')
```

0.0.10 Summary

[13]:
```
print(f'''Summary of the concatenation process:
- tblClaims has : {shape_claims[0]} rows; {shape_claims[1]} columns;\
{round(memStats_claims,2)} MB memory consumption;
```

```
- tblClaimAdmins has : {shape_claimadmins[0]} rows; {shape_claimadmins[1]}␣
 ↪columns;\
{round(memStats_claimadmins,2)} MB memory consumption;
- Concatenation of tblClaims and tblClaimAdmins has : {shape_concat1[0]} rows;\
{shape_concat1[1]} columns; {round(memStats_concat1,2)} MB memory consumption;
- tblInsurees/Families/Locations has : {shape_fam_loc[0]} rows;␣
 ↪{shape_fam_loc[1]} columns;\
{round(memStats_fam_loc,2)} MB memory consumption;
- Concatenation of tblClaims/tblClaimAdmins with tblInsurees/Families/Locations␣
 ↪has : \
{shape_concat2[0]} rows; {shape_concat2[1]} columns;\
{round(memStats_concat2,2)} MB memory consumption;
- tblHFs/Locations has : {shape_HF_loc[0]} rows; {shape_HF_loc[1]} columns;\
{round(memStats_HF_loc,2)} MB memory consumption;
- Concatenation of tblClaims/tblClaimAdmins/tblInsurees/Families/Locations with␣
 ↪\
tblHFs/Locations has : \
{shape_concat3[0]} rows; {shape_concat3[1]} columns;\
{round(memStats_concat3,2)} MB memory consumption;
- tblDiagnosis has : {shape_diag[0]} rows; {shape_diag[1]} columns;\
{round(memStats_diag,2)} MB memory consumption;
- Concatenation of tblClaims/tblClaimAdmins/tblInsurees/Families/Locations/
 ↪tblHFs/Locations \
with tblDiagnosis has : \
{shape_concat4[0]} rows; {shape_concat4[1]} columns;\
{round(memStats_concat4,2)} MB memory consumption;
''')
```

```
Summary of the concatenation process:
- tblClaims has : 5965605 rows; 20 columns;910.28 MB memory consumption;
- tblClaimAdmins has : 913 rows; 5 columns;0.03 MB memory consumption;
- Concatenation of tblClaims and tblClaimAdmins has : 5953640 rows;24 columns;
1135.57 MB memory consumption;
- tblInsurees/Families/Locations has : 3790789 rows; 25 columns;751.96 MB memory
consumption;
- Concatenation of tblClaims/tblClaimAdmins with tblInsurees/Families/Locations
has : 5953640 rows; 48 columns;2225.71 MB memory consumption;
- tblHFs/Locations has : 780 rows; 10 columns;0.07 MB memory consumption;
- Concatenation of tblClaims/tblClaimAdmins/tblInsurees/Families/Locations with
tblHFs/Locations has : 5953640 rows; 57 columns;2634.51 MB memory consumption;
- tblDiagnosis has : 1959 rows; 5 columns;0.07 MB memory consumption;
- Concatenation of
tblClaims/tblClaimAdmins/tblInsurees/Families/Locations/tblHFs/Locations with
tblDiagnosis has : 5953640 rows; 61 columns;2816.21 MB memory consumption;
```