In [3]:

```
root_path = 'C:/Users/rodri/Dropbox/Malawi/SIEG2021 (1)/2022 July'
path_19 = 'C:/Users/rodri/Dropbox/Malawi/Chied_Field_June_19/Data/'

import numpy as np
import pandas as pd
import os

# Set the working directory
os.chdir(root_path+'/Data/Clean data/Phase 1 - Roster')

## Display  set-up
pd.options.display.float_format = '{:,.2f}'.format
pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)

percentiles = [0.05, 0.1, .25, .5, .75, 0.8, 0.9, 0.95, 0.99]


# Village19 variable names roster
'''
'hhid', 'intervieweename', 'wave','head_name', 'village', 'subvillage', 'hh_size', 'head_
 'head_educ', 'head_religion', 'head_female', 'head_divorced', 'head_widowed', 'head_sep
 'head_belowprimary4', 'head_belowprimary7', 'head_belowsecond3', 'head_secondary', 'hea
'''

# ============================================================================
# Import data: Data from the field and conversion rates (ISA-LSMS price conversions)
# ============================================================================

roster = pd.read_stata(root_path+"/Data/Raw data/SIEG-Phase 1-Household Listing 2022-DC.

## Look at duplicates:
dupl_roster = pd.value_counts(roster['hhid'])
print('================================================================')
print('These households are duplicate')

print(dupl_roster[dupl_roster>1])
print('duplicated household has been removed')
dupl = roster.loc[roster.hhid==1317]

print('================================================================')
print('Households interviewed in 2019')
print(pd.value_counts(roster['hh_inter']))


roster2 = roster.drop_duplicates(subset='hhid')

roster = roster[roster.index!=104]

roster.rename(columns={'hh_vill':'invillage_19','hh_inter':'interviewed_19','inter1_full

roster.replace(['6 Years and Over','Less 6 Years',],0, inplace=True)

max_head = max(roster['hh_size'])

for i in range(1,max_head):
    roster['age_years_'+str(i)] = pd.to_numeric(roster['age_years_'+str(i)])
```

```
pd.value_counts(roster['rel_hhhead_1'])

roster[['head_age','head_name','head_gender','head_marital', 'head_nickname']] = roster[


for i in  range(2,max_head):
    #print(i)
    roster[['head_age']] = np.where(roster['rel_hhhead_'+str(i)] == 1, roster['age_years
    roster[['head_nickname']] = np.where(roster['rel_hhhead_'+str(i)] == 1, roster['nick
    roster[['head_name']] = np.where(roster['rel_hhhead_'+str(i)] == 1, roster['mem_name
    roster[['head_gender']] = np.where(roster['rel_hhhead_'+str(i)] == 1, roster['gender
    roster[['head_marital']] = np.where(roster['rel_hhhead_'+str(i)] == 1, roster['matst

check_head = roster[['rel_hhhead_1','rel_hhhead_2','rel_hhhead_3','age_years_1','mem_nam
#head age, name, gender, and marital status properly assigned


## Gender, marital-status, religion, and educ dummies
pd.value_counts(roster['head_gender'], normalize=True)
roster['head_female'] = 1*(roster['head_gender']==2)


## Marital status head
print('================================================')
print('Summary head marital status')
print(pd.value_counts(roster['head_marital'], normalize=True))
roster['head_married_mono'] = 1*(roster['head_marital']==1)
roster['head_married_poly'] = 1*(roster['head_marital']==2)
roster['head_divorced'] = 1*(roster['head_marital']==4)
roster['head_widowed'] = 1*(roster['head_marital']==5)
roster['head_separated'] = 1*(roster['head_marital']==3)
roster['head_nevermarried'] = 1*(roster['head_marital']==6)


# head characteristics
print('================================================')
print('Summary household size and household head characteristics')
print(roster[['hh_size','head_age','head_female','head_marital','head_divorced','head_wi


# head education
print('================================================')
print('Summary household head education')
print(pd.value_counts(roster['head_educ'], normalize=True))
roster['head_noeduc'] = 1*(roster['head_educ']=='No education')
roster['head_belowprimary4'] = 1*((roster['head_educ']=='Primary Standard 1')|(roster['h
roster['head_belowprimary7'] = 1*((roster['head_educ']=='Primary Standard 4')|(roster['h
roster['head_belowsecond3'] = 1*((roster['head_educ']=='Secondary form 1')|(roster['head
roster['head_secondary'] = 1*((roster['head_educ']=='Secondary form 3')|(roster['head_ed

roster['head_educ_countin'] = 0
roster.loc[ roster['head_educ']=='Primary Standard 1' , 'head_educ_countin'] = 1
roster.loc[ roster['head_educ']=='Primary Standard 2' , 'head_educ_countin'] = 2
roster.loc[ roster['head_educ']=='Primary Standard 3' , 'head_educ_countin'] = 3
roster.loc[ roster['head_educ']=='Primary Standard 4' , 'head_educ_countin'] = 4
roster.loc[ roster['head_educ']=='Primary Standard 5' , 'head_educ_countin'] = 5
roster.loc[ roster['head_educ']=='Primary Standard 6' , 'head_educ_countin'] = 6
roster.loc[ roster['head_educ']=='Primary Standard 7' , 'head_educ_countin'] = 7
roster.loc[ roster['head_educ']=='Primary Standard 8' , 'head_educ_countin'] = 8

roster.loc[ roster['head_educ']=='Secondary form 1' , 'head_educ_countin'] = 9
```

```
roster.loc[ roster['head_educ']=='Secondary form 2' , 'head_educ_countin'] = 10
roster.loc[ roster['head_educ']=='Secondary form 3' , 'head_educ_countin'] = 11
roster.loc[ roster['head_educ']=='Secondary form 4' , 'head_educ_countin'] = 12
roster.loc[ roster['head_educ']=='Training college year 2', 'head_educ_countin'] = 11
roster.loc[ roster['head_educ']=='University 4', 'head_educ_countin'] = 12

# head/family religion, ethnicity

roster[['head_christian']] = 1*(roster[['head_religion']]=='Christian')
roster[['chief_related']] = 1*(roster[['chief_related']]=='Yes')
roster[['village_born']] = 1*(roster[['village_born']]=='Yes')
roster['elder_yes'] =  1*(roster[['elder_yes']]=='Yes')
roster['elders_related'] = 1*(roster[['elders_related']]=='Yes')

print('================================================')
print('Summary religion, village background, chiefs and elders')
print(roster[['head_christian','village_born','village_years','chief_related','chief_rel

print(roster[['village','subvillage','head_religion','ethnic','mlanguage','chief_relatio

# other variables

roster['wave'] = '2022'

roster = roster[['hhid','wave','invillage_19','interviewed_19','oldhhid', 'intervieweena
 'head_educ', 'head_religion', 'head_female', 'head_married_mono','head_married_poly','h
 "spouse_educ", "ethnic",  "mlanguage",'village_born','village_years','chief_related','c
 'head_belowprimary4', 'head_belowprimary7', 'head_belowsecond3', 'head_secondary', 'hea


print('================================================')
roster.to_csv(root_path+"/Data/Clean data/Phase 1 - Roster/roster_22.csv",index=False)
print('final dataset saved in clean data/phase 1/roster.csv')
print('================================================')
print("Containts the following variables: 'hhid','wave','invillage_19','interviewed_19',
```

```
============================================================
These households are duplicate
1317    2
Name: hhid, dtype: int64
duplicated household has been removed
============================================================
Households interviewed in 2019
Yes    217
No      80
Name: hh_inter, dtype: int64
==============================================
Summary head marital status
1.00   0.62
4.00   0.20
2.00   0.09
6.00   0.05
3.00   0.02
5.00   0.02
Name: head_marital, dtype: float64
==============================================
Summary household size and household head characteristics
       hh_size  head_age  head_female  head_marital  head_divorced  \
count   296.00    296.00       296.00        296.00         296.00
mean      4.36     42.66         0.35          2.05           0.20
std       1.86     17.69         0.48          1.54           0.40
min       1.00     10.00         0.00          1.00           0.00
25%       3.00     27.00         0.00          1.00           0.00
50%       4.00     40.00         0.00          1.00           0.00
75%       6.00     52.25         1.00          4.00           0.00
max      11.00     93.00         1.00          6.00           1.00

       head_widowed  head_separated
count        296.00          296.00
mean           0.02            0.02
std            0.13            0.15
min            0.00            0.00
25%            0.00            0.00
50%            0.00            0.00
75%            0.00            0.00
max            1.00            1.00
==============================================
Summary household head education
Primary Standard 8        0.16
Primary Standard 7        0.14
Primary Standard 5        0.11
No education              0.10
Primary Standard 4        0.09
Primary Standard 6        0.07
Primary Standard 3        0.07
Primary Standard 2        0.06
Primary Standard 1        0.05
Secondary form 4          0.04
Secondary form 2          0.04
Secondary form 3          0.03
Secondary form 1          0.02
Training college year 3   0.00
Training college year 2   0.00
University 5 and above    0.00
Name: head_educ, dtype: float64
==============================================
Summary religion, village background, chiefs and elders
```

```
       head_christian  village_born  village_years  chief_related  elder_y
es  \
count          296.00        296.00         296.00         296.00     296.
00
mean             0.16          0.73          13.96           0.62       0.
12
std              0.37          0.44          14.06           0.48       0.
33
min              0.00          0.00           0.00           0.00       0.
00
25%              0.00          0.00           3.00           0.00       0.
00
50%              0.00          1.00           9.00           1.00       0.
00
75%              0.00          1.00          20.00           1.00       0.
00
max              1.00          1.00          59.00           1.00       1.
00

       elders_related
count          296.00
mean             0.47
std              0.50
min              0.00
25%              0.00
50%              0.00
75%              1.00
max              1.00
                               village subvillage head_religion ethnic
\
count                              296        139           296    296
unique                               5          7             3      8
top     Geradi (different sub-villages).     Geradi        Muslim    Yao
freq                               139         38           245    229

       mlanguage chief_relation      elders_relation
count        296            185                  139
unique         5             15                   16
top          Yao    Grandparent  Maternal aunt/uncle
freq         179             48                   29
================================================
final dataset saved in clean data/phase 1/roster.csv
================================================
Containts the following variables: 'hhid','wave','invillage_19','interview
ed_19','oldhhid', 'intervieweename', 'wave','head_name', 'village', 'subvi
llage', 'head_religion','key_landmark','mosque_church','hh_size', 'hh_phon
e', 'head_gender', 'head_marital', 'head_age', 'head_nickname','head_edu
c', 'head_religion', 'head_female', 'head_married_mono','head_married_pol
y','head_nevermarried', 'head_divorced', 'head_widowed', 'head_separated',
'head_christian', 'head_noeduc', 'spouse_educ', 'ethnic',  'mlanguage','vi
llage_born','village_years','chief_related','chief_relation','elder_ye
s','elders_related','elders_relation','head_belowprimary4', 'head_belowpri
mary7', 'head_belowsecond3', 'head_secondary', 'head_educ_countin',  'gps_
lat', 'gps_long'
```