Hw_1

Assignment 1

- Overview of Hw 1 using R, Python and STATA

A.1 Import and create folders

A.1.1 Import libraries

```
# R
library(foreign)
```

```
# Python
import pandas as pd
import os
```

```
* Stata
* Not necessary for stata
```

A.1.2 Create folders (directory) and change directory

1. Find working director
2. Make new folder
3. Change working directory
4. Create dta folder
5. Return to hw folder

```
# R
#1 Find working directory
path = getwd()
#2 Make new folder
dir.create('hw1')
#3 Change working directory
setwd('hw1')
#4 Create dta folder
dir.create('dta')
#5 Return to hw folder
setwd(path+'/hw1')
```

```
# Python
#1 Find working directory
path = os.getcwd()
#2 Make new folder
os.mkdir('hw1')
#3 Change working directory
os.chdir('hw1')
#4 Create dta folder
os.mkdir('dta')
#5 Return to hw folder
os.chdir(path+'/hw1')
```

```
* Stata
local directory : pwd
mkdir hw1
cd "`directory'\hw1\"
local hw_fol "`directory'\hw1"
mkdir dta
local dta_fol "`hw_fol'\dta\"
cd ..
```

A.1.3 Upload files

- Important step! Place the dta files into the dta folder we created
    - This folder is now accessible:
        - In Stata as local variable: `dta_fol'
        - In R & Python as local variable: hw_fol

```
# R
dta_fol = path+'/hw1/dta/'
df = read.dta(dta_fol+'c_ls.dta')
```

```
# Python
hw_fol = path+'/hw1/dta/'
df = pd.read_stata(dta_fol+'c_ls.dta')
df.set_index(['folio', 'ls'], inplace=True) # This step is important for merging
```

```
* Stata
use "`direct_folder'\c_ls.dta"
```

A.1.3.1 Merge data

```
# R
df_size = read.dta(dta_fol+'c_portad.dta')
merged <- merge(df,df_size,by=c("ls"))
```

```
# Python
df_size = pd.read_stata(dta_fol+'c_portad.dta')
```

```
df_size.set_index(['folio','ls'], inplace=True) #This step is important for merging
merged = pd.merge(df_size, df, left_index=True, right_index=True)
```

```
* Stata
merge m:m folio ls using "`dta_fol'\c_portad"
```

A.2 Examine data

A.2.1 Show first 5 rows

```
# R
head(df, 5)
```

```
# Python
df.head(5)
```

```
* Stata
list in 1/5
```

A.2.2 Show information on variables

```
# R
str(df)
```

```
# Python
df.info()
```

```
* Stata
describe
```

A.3 Drop ages other than 5-14

```
# R
ages_5to14 <- df[which(df$ls02_2>4 & df$ls02_2<15),]
```

```
# Python
mask = (df['ls02_2']>4) & (df['ls02_2']<15)
ages_5to14 = df.loc[mask]
```

```
* Stata
keep if ls02_2>4 & ls02_2<15
```

A.3.1 Replace non-attendance number 3 to 0

```
# R
ages_5to14$ls16[ages_5to14$ls16==3]=0
```

```python
# Python
merged['enrolled_dummy'] = merged['ls16'].map({3:0, 1:1})
```

```stata
* Stata
replace ls16=0 if ls16==3
```

Q.1

What proportion of children between the ages of 5 and 14 are enrolled in school?

```r
# R
q1_answer <- as.data.frame.matrix(table(ages_5to14$ls02_2, ages_5to14$ls16))
colnames(q1_answer)=c("Attend","Non_attend")
q1_answer$enrolled_pct <- paste(round(q1_answer$Attend/(q1_answer$Attend+q1_answer$Non_attend)*100, 2), "%")
q1_answer
```

```python
# Python
ages_5to14['enrolled_dummy'] = ages_5to14.ls16.map({3:0, 1:1})
q1_answer = ages_5to14.groupby('ls02_2')[('enrolled_dummy')].mean().to_dict()
pd.DataFrame({'Age': list(q1_answer.keys()),'Pct_attendance': list(q1_answer.values())})
```

```stata
* Stata
tab ls02_2 ls16, row nofreq
```

Q.2

How does the proportion enrolled in school differ by gender?

```r
# R
q2_answer <- as.data.frame.matrix(table(ages_5to14$ls04, ages_5to14$ls16))
colnames(q2_answer)=c("Attend","Non-attend")
q2_answer$enrolled_pct <- paste(round(q2_answer$Attend/(q2_answer$Attend+q2_answer$Non_attend)*100, 2), "%")
rownames(q2_answer) <- c("Male","Female")
q2_answer
```

```python
# Python
ages_5to14['Sex_dummy'] = ages_5to14.ls04.map({1:'Male', 3:'Female'})
q2_answer = ages_5to14.groupby('Sex_dummy')[('enrolled_dummy')].mean().to_frame()
q2_answer.reset_index(inplace=True)
q2_answer.columns = ['Gender', 'Pct_attendance']
q2_answer
```

```stata
* Stata
sort ls04
```

```
by ls04: tab ls16
```

Q.3

How does the proportion enrolled in school differ by size of zone of residence?

- Provide a table or graph.

Tip: Estrato contains the size of community and is located in c_portad.dta from the Control Book.

- Estrato=1 Households located in localities with more than 100,000 inhabitants.
- Estrato=2 Households located in localities with populations between 15,000 and 100,000.
- Estrato=3 Households located in towns with a population between 2,500 and 15,000.

```
# R
q3_answer <- as.data.frame.matrix(table(merged$estrato, merged$ls16))
colnames(q3_answer)=c("Attend","Non-attend")
q3_answer$enrolled_pct <- paste(round(q3_answer$Attend/(q3_answer$Attend+q3_answer$Non-attend)*100, 2), "%")
rownames(q3_answer) <- c('100k', '100k-15k', '15k-2.5k', '<2.5k')
q3_answer
```

```
# Python
merged['enrolled_dummy'] = merged['ls16'].map({3:0, 1:1})
q3_answer = merged.groupby('estrato')['enrolled_dummy'].mean().to_dict()
q3_answer = pd.DataFrame(list(q3_answer.values()),
index=['100k', '100k-15k', '15k-2.5k', '<2.5k'],
columns= ['Pct_attendance'])
q3_answer.index.name = 'City size'
q3_answer
```

```
* Stata
tab estrato ls16, row
```

Q.5

What proportion of children are working outside the home?

- Provide evidence on whether children who work outside the home less likely to attend school.
- Compare this evidence for boys and girls and by size of zone of residence.

```
# R
q5_answer <- as.data.frame.matrix(table(merged$ls12, merged$ls16))
colnames(q5_answer)=c("Attend","Non-attend")
q5_answer$enrolled_pct <- paste(round(q5_answer$Attend/(q5_answer$Attend+q5_answer$Non-attend)*100, 2), "%")
```

```
rownames(q3_answer) <- c('100k', '100k-15k', '15k-2.5k', '<2.5k')
q3_answer
```

```
# Python
merged['Worked_dummy'] = merged.ls12.map({1:1, 3:0})
q5_answer = merged.groupby('Worked_dummy')['enrolled_dummy'].mean().to_dict()
q5_answer = pd.DataFrame(list(q5_answer.values()), index=['Worker','Non-worker'],columns=['Pct_attendance'])
q5_answer
```

```
* Stata
tab ls12 ls16, row
bysort ls04: tab ls12 ls16, row
bysort estrato: tab ls12 ls16, row
```