## Take Home Exam: Youths Not in Employment, Education, or Training

ECON 258 Data Analytics with R — Winter 2022

**Important to read until the end first before starting!**

**Due Tuesday, March 1, 11:59 PM**

**Motivation:**

Hypothetical scenario: A wealthy donor is interested in investing in youth labor market programs in a number of Asian countries. This client tasked you to provide a brief report on the characteristics of youths who are not in employment, education, or training in Indonesia.

If you want some background knowledge on the issue, here are optional background reading:

1) Generation Jobless from The Economist
2) Report on Youth Unemployment from ILO

**Exam Goal:**

⇒ **Conduct preliminary descriptive statistics on who are the youth neets in Indonesia**

Age, gender, marital status, and highest education level achieved will be the main focus.

**Disclaimer:** proper descriptive statistics should take into account survey weights. For this exam, we will ignore this for now and get descriptive statistics for the sample and not care about representative descriptive statistics.

## Part 1: Setting Up Directory, Importing Data, and Data Prep

Download data from the Indonesia Family Life Survey (IFLS) 2014 provided on canvas, "Exam1_Data". The folder contains:

- b3a_cov includes basic information on an individual's basic characteristics
- b3a_dl1 includes information on highest education
- b3a_tk1 includes information on weekly activities and employment status

To aid with understanding the data, the folder includes the questionnaire for these datasets. Open the module and go to section TK (employment) and DL (education). This will help you understand the variables.

The main variables that you will need:

- hhid14 is the unique household identifier
- pid14 is the unique person ID within the household
- From b3a_cov: age, sex, marital status
- From b3a_dl1: dl04, dl06, dl07a (careful dl06 is answered only for those who answered yes to dl04, the rest is missing)

- From b3a_tk1: tk01a, tk02, tk03, tk04

**A Few Data Prep Instructions:**[1]

**1) Need to create a variable indicating highest education attainment. The Indonesian data is much more detailed than what we would like.**

Create a variable that shows a general highest education category that is based on variable dl06. This new variable will have values:

0 - no schooling (if dl06 is empty or the person never went to school (dl04))

1 - elementary school (dl06 of 2 and 72)

2 - junior high (dl06 of 3, 73, 4)

3 - high school general (dl06 of 5, 74)

4 - high school vocational (dl06 of 6)

5 -higher education (dl06 of 13, 60, 61, 62, 63)

Any other dl06 should be coded as missing for this variable and not used in the data analysis. These other ones are 11, 12, 15, 17, 14, and bigger than 90.

**2) Need to create a variable categorizing whether an individual is a NEET or not.**

Use tk01a, tk02, tk03, tk04, dl07a, and age to specify whether a person is a NEET or not.

A NEET is a person age 15 - 29 who is currently not in school and currently not employed or in training. Therefore, a person who is **NOT** a NEET will answer yes to either dl07a, tk01a, tk02, tk03, or tk04. Otherwise, label them as a NEET. We will focus on just this age group. So you can ignore or filter out other age groups.

**3) Create a variable of different age categories.**

1 - those age 15 - 19

2 - those age 20 - 24

3 - those age 25 - 29

**4) Handling missing variables.**

IFLS usually uses the code 8, 9, 98, 99, 998, or 999 (patterns of 8 and 9s) to specify missing or when an individual does not know the answer. Check to make sure the **variables you used** in your data analysis does not contain these numbers. Missings should be deleted. You can also print the variable labels to see how missings/not known are coded. Keep these code checks when you submit.

To summarize: the sample we are using for analysis are those where information on age, gender, schooling, work, and marital status are not missing. Be careful distinguishing actual missing (due to human error) versus missings because the answer is not applicable. For example, a person's education is missing but because he previously did not go to school. This should have been coded already in data prep part 1.

## Part 2: Getting Summary Statistics and Some Plots

1. Look at the number of observations for highest level of education category. No schooling has very few observations, thus drop these individuals.
2. Find the proportion of individuals who are NEETS by gender **and** highest education. The **and** here is highlighted because I want them combined: Female of education 1, 2, 3 etc then Male of education 1, 2, 3 etc. For example, the denominator for female education 1 is all female of education 1. The numerator for female education 1 is the proportion in that category who are NEETS.
3. Find the proportion of individuals who are NEETS by the different age category **and** gender.
4. Find the proportion of individuals who are NEETS by the different age category **and** marital status.
5. For one of the above proportions, create a ggplot for them. You can see reference visualization book link here. See section "categorical vs categorical" and also about 5.1 grouping and 5.2 faceting when using ggplot. Label everything properly! Make sure an audience understand what this graph is.
6. Graph frequency histogram of individuals who are considered NEET by education level and gender category. Label everything properly.
7. Do not forget to have a code to save your graph.

## Part 3: Write Down Brief Notes Given Your Results

Answer the following questions briefly, use numbers found in part 2 to make your arguments more convincing. Play around with the data more if you want to find more evidence to convince yourself.

1. Looking at gender, education, and age, which group has the highest proportion of NEET rates?
2. How large is the gap in NEET rates between men and women?
3. Open question: Did you find any surprising or not surprising patterns concerning the NEET rates by group categories? Explain.

---

**Some hints:**

1. **First,** come up with a strategy on how you want to answer the questions in Part 3. Look at the data and look at the homework 1 and lab 1 for inspiration for steps to do.
2. You should be able to use codes and packages discussed in class to answer these questions. Other commands or packages not used in class are fine if you prefer them but not necessary.