# R Notebook

## Introduction

– What is the aim of the project? - Summarize the problem - State your goals – What do you do in this report? - Offer a roadmap of the project

# Statement of the problem

# Goals

Using information from SWTS the objective of this project is to identify the characteristics that young people need to have in order to find a job. Furthermore, to identify the characteristics that employers are looking for in a youth to employ her/him will be extremely useful. This will provide policy elements for potential interventions to link young people with labor market from both sides, supply and demand.

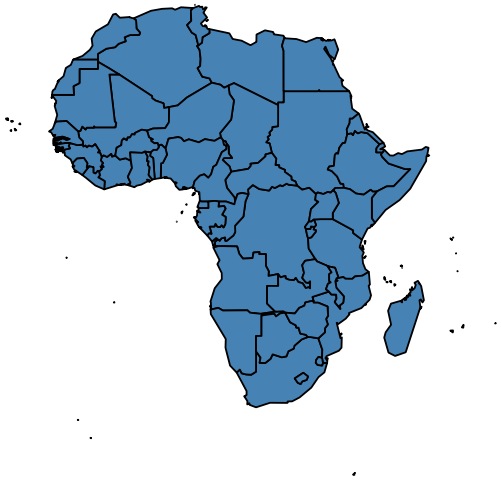
## Problem Statement and Background

– Give a clear and complete statement of the problem and/or aim of your analysis. – Include a brief summary of any related work that has tried to tackle a project similar to yours (i.e. a light literature review)

Youth employment is a major concern of governments across the developing world, where young people are up to four times more likely than adults to be unemployed. In these countries, 85% of the jobs taken by youth belong to informal markets, and a quarter of the youth cannot find a salary above the extreme poverty line, $1.25 per day (Goldin et al., 2015). In addition, the majority of youth are dissatisfied with their employment situation, and want to change jobs, even more so in rural settings and in the agricultural sector (OECD, 2018). For all sectors, the main reason for wanting to change jobs is low pay, followed by the temporary nature of employment and poor working conditions.

To improve youth employment, interventions from the supply and demand side can be done. From the supply side, interventions focus mainly in increasing employment rate and earnings. Programs on skills training and support self-employment have largest effects, although the second one works better. On the demand side, programs have positive effects, although they are small. Those interventions that support the small and medium entrepreneurship (SMEs) are more effective in creating jobs than interventions that address micro-firms.

### Africa plot



### Data

– Where does the data come from? – What is the unit of observation? – What are the variables of interest? – What steps did you take to wrangle the data?

## Source of the data

With the aim to portray which are the elements that allow a youth to transit from unemployment to employment, the International Labour Organization (ILO) designed the School-to-Work Transition Survey (SWTS). As ILO defined, the SWTS is a unique survey instrument that generates relevant labor market information on young people aged 15 to 29 years, including longitudinal information on transitions within the labor market. The SWTS thus serves as a unique tool for demonstrating the increasingly tentative and indirect paths to decent and productive employment that today's young men and women are facing.

The databases for 33 countries was shared by the ILO for this final project. The databases corrreponds to countries for different countries, however, for this project, 9 countries from the non Sub-Saharan Africa were choosen. The reason for choosing those is because currently I am working in my capstone project in which we are trying to identify the determinants of youth employment in Burkina Faso and the countries selected preserve similar income conditions as Burkina.

The unit of observation in this project is each youth among 15 and 29 year old that answer the survey in each of the selected countries. So far, a country-youth unit of analysis is out from the scope of this project.

# Variables of interest

In average, each survey has around 400 questions. Considering that my objective is to find which characteristics determine that a youth get a job, I chose as dependent the dummy variable "employed" which values 1 if the youth have a job at the time of the survey and 0 if not. This will allow me to structure models that measure the probability that a youth get a youth based on the chosen independent variables.

As independet variables, I chose 11 variables: age, sex, education level, marital status, education level, if work while study, if internship, if have a reserve wage, if live in rural or urban areas, if have childrens, and financial situation at home.

# Wrangling data

The most challeging part in wrangling the data was to match each variable of interest in the different databases. Even though each database follows in general the same SWTS strucrure and almost all the questions are the same, those are named and labeled differentyl. This implied to go over each database, along the more than 400 questions in the 9 countries, to match the variables of interest.

After that, the path was easier. Each base was in Stata format, so changed that to R format and applied a "remove labels" command because labels merged with the variable tittle complicating the subsequent merge. Then for each database, I selected just the variables of interest mentioned above and relabeled. Then create dummies

Finally, I merged the 9 databases in one.

```
##
##    0    1
## 1689 1587
```

```
##
##    0    1
## 3160  116
```

```
##
##    0    1
## 2493  701
```

```
##
##   0   1
## 494 207
```

```
##
##    0    1
## 2137 1139

##
##   0   1
## 174 368

##
##   0   1
## 682 139

##
##    0    1
## 2083 1193

##
##    0    1
## 1657 1440

##
##    0    1
## 1906 1191

##
##    0    1
## 2287  675

##
##   0   1
## 587  88

##
##    0    1
##  994 2103

##
##   0   1
## 150 234

##
##   0   1
## 546 542
```

```
##
##    0    1
## 1741 1356

##
##    0    1
## 1728 1497

##
##    0    1
## 2639  586

##
##    0    1
## 2423  502

##
##   0   1
## 345 157

##
##    0    1
## 1418 1807

##
##   0   1
## 362 387

##
##   0   1
## 542 283

##
##    0    1
## 2214 1011

##
##    0    1
## 1831 1469

##
##    0    1
## 2217 1082
```

```
##
##    0    1
## 1660 1073

##
##    0    1
## 2688   55

##
##    0    1
##  686 2614

##
##    0    1
##  907 1667

##
##   0   1
## 262 426

##
##    1
## 3285

##
##    0    1
## 1003  985

##
##    0    1
## 1731  257

##
##    0    1
## 1597  296

##
##    0    1
## 1738  155

##
##    0    1
## 1219  769
```

```
##
##   0   1
## 465 185

##
##   0   1
## 326 231

##
##    0    1
## 1563  425

##
##    0    1
## 1659 1390

##
##    0    1
## 2430  618

##
##    0    1
## 2115  760

##
##   0   1
## 673  76

##
##    0    1
## 1088 1961

##
##   0   1
## 216 233

##
##   0   1
## 163 343

##
##    0    1
## 1758 1235
```

## Analysis

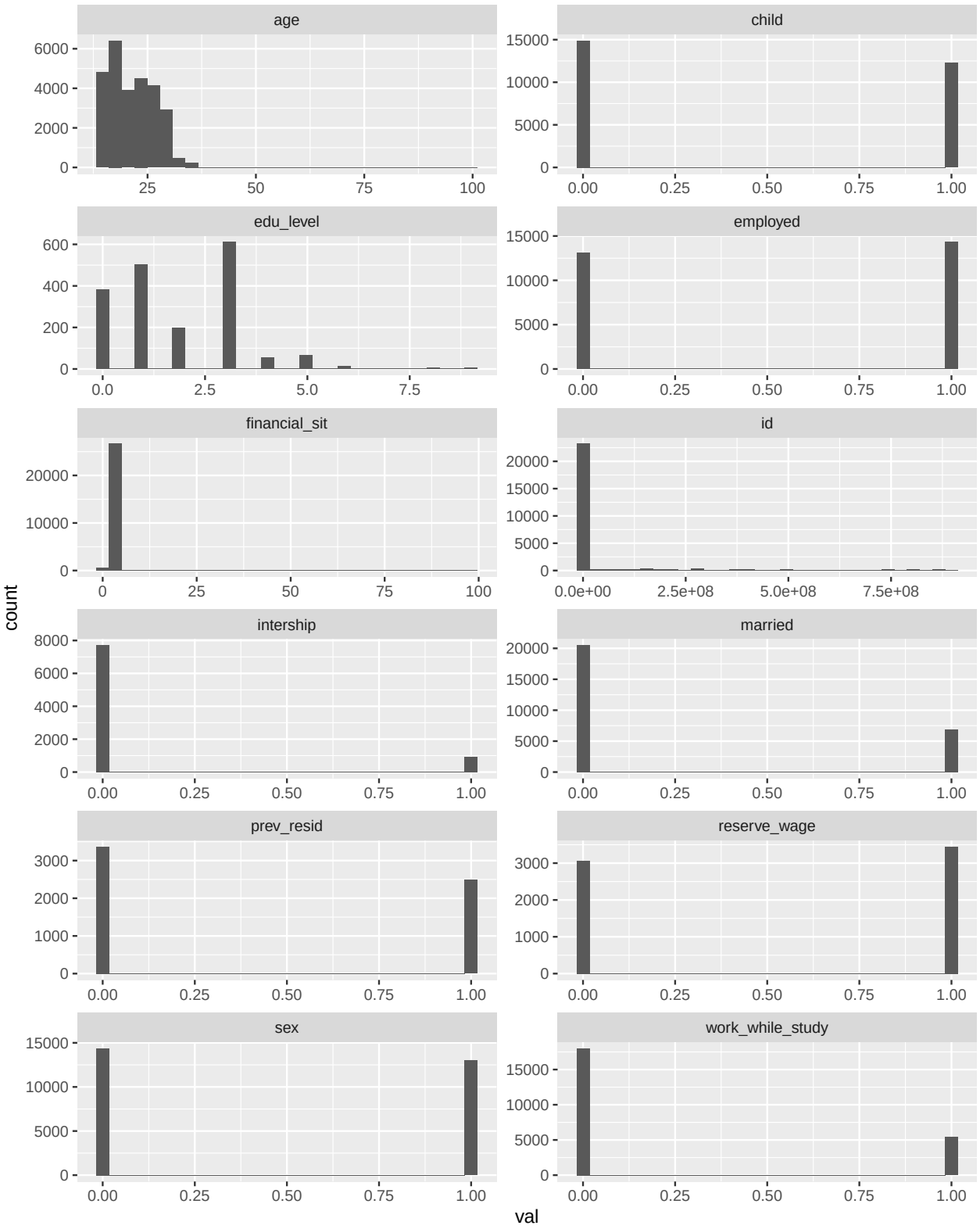
– Describe the methods/tools you explored in your project. – Outline in detail our entire analysis. - Justify the tools/methods that you used. - Assume the reader is smart but doesn't know R/Machine Learning well. That is, be crystal clear about what you're doing and why.

In the process of running ML techniques, I identify some outliers in my data: - Remove age>29 - Remove financial_sit>5

First, let's look at the variables.

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```

```
## Warning: Removed 90886 rows containing non-finite values (stat_bin).
```




Extra wrangling

```
##
##    14    15    16    17    18    19    20    21    22    23    24    25    26    27    28
```

```
##    2 2626 2168 2136 2372 1897 2314 1575 1726 1365 1422 1659 1190 1283 1451
##   29
## 1170


##
##    1    2    3    4    5
##  649 2649 8985 8469 5542
```

## 3. The model

The objective of the project is to find a model that predicts the probability of young people to find a job, base on certain characteristics, in Sub-Saharan Africa countries.

$$Y = \beta_0 + \beta_n X_n + \epsilon$$

Where:

- $Y$ is the probability of a youth to being employed
- $X_n$ is a matrix of predictors based on the information from the survey

## Logit regression

This is a simple logit regression to see the the results of the expected model

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -26.5660722 | 648623.50 | -0.0000410 | 0.9999673 |
| sex | -0.0000001 | 125844.28 | 0.0000000 | 1.0000000 |
| age | 0.0000001 | 28914.78 | 0.0000000 | 1.0000000 |
| married | 0.0000000 | 272152.39 | 0.0000000 | 1.0000000 |
| edu_level | -0.0000002 | 57510.11 | 0.0000000 | 1.0000000 |
| work_while_study | 0.0000010 | 188357.25 | 0.0000000 | 1.0000000 |
| reserve_wage | 0.0000007 | 163941.78 | 0.0000000 | 1.0000000 |
| intership | -0.0000006 | 443727.64 | 0.0000000 | 1.0000000 |
| child | 53.1321366 | 413743.74 | 0.0001284 | 0.9998975 |
| prev_resid | -0.0000001 | 142853.82 | 0.0000000 | 1.0000000 |
| financial_sit | 0.0000006 | 82822.81 | 0.0000000 | 1.0000000 |

### Machine learning Techniques

## Split the Sample: Training and test data

Before event looking at the data, let's split the sample up into a training and test dataset. We'll completely hold off on viewing the test data, so as not to bias our development of the learning model.

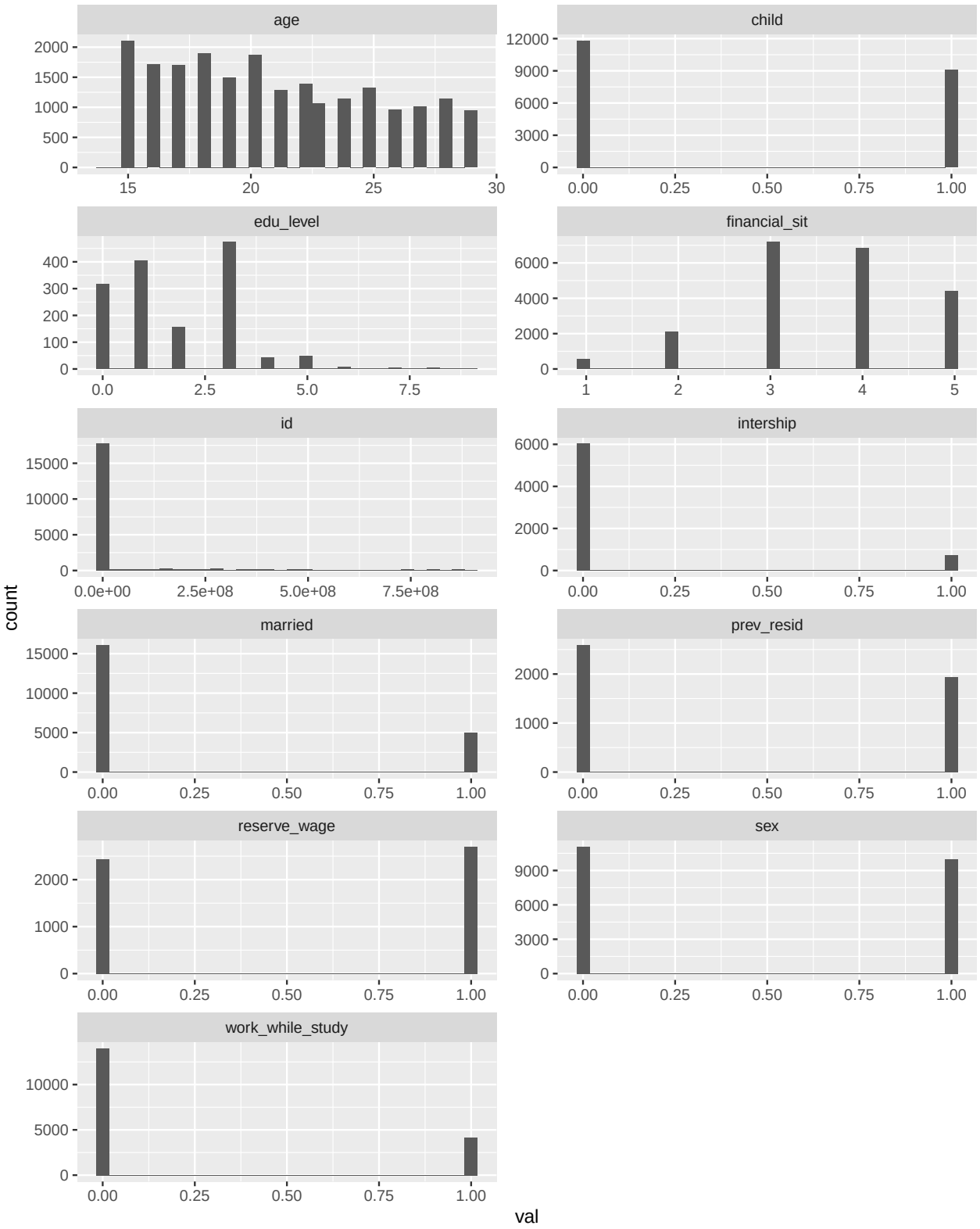
```
## [1] 21036    13
```

```
## [1] 5258   13
```

Visualize the distribution for each variable.

First, let's look at the variables.

```
## Warning: Removed 69332 rows containing non-finite values (stat_bin).
```


age
child
edu_level
financial_sit
id
intership
married
prev_resid
reserve_wage
sex
work_while_study
count
val

# Pre-process the Data

Missing values:

```
## [1] 86670
```

Remove missing values:

```
## [1] 0
```

Check that our pre-processing worked.

```
## # A tibble: 6 x 20
##        id   sex   age prev_resid married child financial_sit edu_level
##     <dbl> <dbl> <dbl>      <dbl>   <dbl> <dbl>         <dbl>     <dbl>
## 1 0.          1 0.4           NA       0     0             1        NA
## 2 2.23e-9     1 0.2           NA       0     0             1        NA
## 3 3.34e-9     1 0.133         NA       0     0             1        NA
## 4 4.46e-9     1 0.6           NA       1     0             1        NA
## 5 5.57e-9     0 0.4           NA       1     0             1        NA
## 6 6.69e-9     0 0.933         NA       1     1             1        NA
## # ... with 12 more variables: work_while_study <dbl>, intership <dbl>,
## #   reserve_wage <dbl>, employed <fct>, country_Congo <dbl>,
## #   country_Liberia <dbl>, country_Madagascar <dbl>, country_Malawi <dbl>,
## #   country_Tanzania <dbl>, country_Togo <dbl>, country_Uganda <dbl>,
## #   country_Zambia <dbl>
```

# Cross-validation

When comparing different machine learning models, we want to make sure we're making a fair and equal comparison. One way that random chance can sneak into our assessments of model fit is through cross-validation. We want to make sure that we're cross-validating the data on the exact same data partitions.

`caret` makes this ease to do. Let's use the k-fold cross-validation method with 10 folds in the data.

```
## Fold1 Fold2 Fold3 Fold4 Fold5
##  4206  4208  4207  4207  4208
```

Now, let's use the `trainControl()` function from `caret` to set up our validation conditions. An important changes from last week: we have to tell `caret` that we are dealing with a classification problem. We can do this by adding `summaryFunction = twoClassSummary` and `classProbs = TRUE` to the cross-validation function.

We'll now use this same cross-validation object for everything that we do.

# Models

Let's explore the different models that we covered in the lecture using the same package framework. As we saw last time in class, `caret` facilitates this task nicely.

## Logistic Regression

Recall we have issues with some of the variable categories (i.e. do to insufficient variation), so we'll use a retricted model here and just include _some_of the variables.

## Results

– Give a detailed summary of your results. – Present your results clearly and concisely. – Please use visualizations and tables whenever possible.

## Discussion

– Speak on the "success" of your project (as defined in your proposal). - Did you achieve what you set out to do? If not why? – What tools/methods did you consider but not use in the final analysis? – How would you expand the analysis if given more time?

In this sense, a success in the project will be if I can provide evidence of which is the youth profile that successfully enroll the job market and which is the weight of each characteristic in the probability of success. For example, if years of education influence the probability to find a job, in which proportion, in average the impact is.

# References:


- Goldin, N. & M. Hobson with P. Glick, M. Lundberg, S. Puerto (S4YE). 2015. "Toward Solutions for Youth Employment: A Baseline for 2015." Solutions for Youth Employment, Washington D.C
- OECD (2018), The Future of Rural Youth in Developing Countries: Tapping the Potential of Local Value Chains, OECD Publishing, Paris.