Primary Dataset: Indicators of economic prosperity

GDP per capita in US$
Public health expenditure (% GDP)
The poverty gap, in international-$
Annual healthcare expenditure per capita
Death rate from malnutrition
Infant mortality rate

First Secondary Dataset: Indicators of woman empowerment

Proportion of seats held by women in national Parliaments,
Proportion of women in senior and middle management positions,
Youth literacy rate (female),
Share of firms with female top managers,

In fact, we have more metrics than rows and that is why we will be picking only some of them so as not to get into the "curse of dimensionality" problem.

Second Secondary Dataset: Indicators of Educational Development

Govt. expenditure on education (as percentage of GDP),
Completion rate of lower secondary education,
Gross enrollment ratio in primary education,
Gross enrollment ratio in secondary education,
In this case as well, we have lot more metrics than rows and we will need to choose from these.

Target Variable

Homicide rate

In most cases, the data for the above predictors is mostly available for 2012-2014 time frame. We choose one year and work for that year (say 2012). Most predictors are continuous variables. Notice that we have data for multiple years. If we have a data value missing, we can impute from the trends (from previous years). WE need for about 100 countries (some flexibility on that is possible). So, if we were to think in terms of training test split: 75 for training and 25 for test. However, we can use cross validation. However, let us do for 100 countries. The datasets are already in excel files in most of the cases and pretty well organized. So, it should not take long to put it in dataframe.