

UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

CURSO DE ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS

PROJETO INTEGRADOR DO MÓDULO 1

INTRODUÇÃO

O presente documento tem o objetivo de descrever o projeto envolvendo as disciplinas estudadas durante o módulo 1 do curso de especialização de Ciência de Dados pela UTFPR.

O projeto tem a proposta transversal de trabalhar todo o conteúdo abordado nas disciplinas de *Ambiente de Ensino e Aprendizagem a Distância*, *Análise e Modelagem de Banco de Dados*, *Introdução à Análise e Ciência de Dados* e *Introdução ao Gerenciamento de Bancos de Dados*.

O trabalho foi desenvolvido pela Equipe 3, composta por:

- Arturo Vaine
- Otávio Teixeira
- Robson Mamede

CONTEXTO

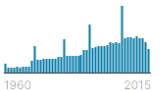
A proposta desenvolvida aqui objetiva a utilização da base de dados indicadores de desenvolvimento mundial em um período que compreende 1960 a 2015, com granularidade nacional (países) e periodicidade anual.

A base utilizada está publicada no Kaggle, sob o título [World Development Indicators](#), e tem como fonte o Banco Mundial, cujo tamanho total é de 1.9GB.

Os formatos disponíveis dos dados são:

Arquivos CSV com informações sobre os países, indicadores, classificação dos indicadores, notas sobre os indicadores etc.

Um arquivo .sqlite apropriado para o banco de dados SQLite que geralmente é utilizado em apps de dispositivos móveis.

| Indicators.csv (547.7 MB) | | | | | |
|---|----------------------|---|-----------------------|---|----|
| About this file | | | | | |
| Values of different indicators for all the countries. | | | | | |
| CountryName | CountryCode | IndicatorName | IndicatorCode | Year | # |
| 247 unique values | 247 unique values | 1344 unique values | 1344 unique values |  | -6 |
| Arab World | ARB | Adolescent fertility rate (births per 1,000 women ages 15-19) | SP.ADO.TFRT | 1960 | 11 |
| Arab World | ARB | Age dependency ratio (% of working-age population) | SP.POP.DPND | 1960 | 8 |

A base conta com informações de 217 nações distintas e mais de 1340 indicadores, com os quais objetivamos realizar análises e comparações, a seguir:

- A evolução de alguns indicadores Brasil em comparação a alguns países ao longo de um dado período;

Indicadores:

| | |
|-----------------------------|---|
| <i>SL.TLF.CACT.FE.NE.ZS</i> | <i>Labor force participation rate, female (% of female population ages 15+) (national estimate)</i> |
| <i>SP.MTR.1519.ZS</i> | <i>Teenage mothers (% of women ages 15-19 who have had children or are currently pregnant)</i> |
| <i>SP.DYN.LE00.IN</i> | <i>Life expectancy at birth, total (years)</i> |

- Crescimento médio a cada 10 anos da população do Brasil.

Indicador:

| | |
|--------------------|-------------------------------------|
| <i>SP.POP.GROW</i> | <i>Population growth (annual %)</i> |
|--------------------|-------------------------------------|

- População Urbana x Rural em determinado período no Brasil.

Indicadores:

| | |
|--------------------------|---|
| <i>SP.URB.TOTL.IN.ZS</i> | <i>Urban population (% of total)</i> |
| <i>SP.RUR.TOTL.ZS</i> | <i>Rural population (% of total population)</i> |

- Crescimento da taxa de mortalidade infantil abaixo de 5 anos (a cada 1000 nascidos vivos) ao longo de um período no Brasil comparado com alguns países:

Indicador:

| | |
|--------------------|--|
| <i>SH.DYN.MORT</i> | <i>Mortality rate, under-5 (per 1,000 live births)</i> |
|--------------------|--|

- **Curiosidades:** aqui é uma miscelânea de indicadores excêntricos e sobre eles, falaremos mais na seção que trata da disciplina de *Introdução à Análise e Ciência de Dados*, onde explicamos os motivos de terem sido escolhidos e as observações feitas sobre os dados.

Indicadores

| | |
|-----------------------|---|
| <i>SH.MED.PHYS.ZS</i> | <i>Physicians (per 1,000 people)</i> |
| <i>SG.VAW.REAS.ZS</i> | <i>Women who believe a husband is justified in beating his wife (any of five reasons) (%)</i> |
| <i>SG.VAW.ARGU.ZS</i> | <i>Women who believe a husband is justified in beating his wife when she argues with him (%)</i> |
| <i>SG.VAW.BURN.ZS</i> | <i>Women who believe a husband is justified in beating his wife when she burns the food (%)</i> |
| <i>SG.VAW.GOES.ZS</i> | <i>Women who believe a husband is justified in beating his wife when she goes out without telling him (%)</i> |
| <i>SG.VAW.NEGL.ZS</i> | <i>Women who believe a husband is justified in beating his wife when she neglects the children (%)</i> |

SG.VAW.REFU.ZS *Women who believe a husband is justified in beating his wife when she refuses sex with him (%)*

DISCIPLINA: AMBIENTE DE ENSINO E APRENDIZAGEM A DISTÂNCIA

Tema: explorar, analisar, manipular e gerar visualizações de dados referentes ao desenvolvimento mundial.

Contexto: já apresentado na contextualização.

Estratégia do grupo:

Por ser uma base de ampla interdisciplinaridade, a equipe optou por explorar dados de indicadores sociais, que são de mais direto entendimento. Optamos também focar no Brasil em comparação com outros países quando conveniente.

A estratégia do grupo foi a de que, a partir dos dados, fossem avaliados requisitos e ideias sobre o que explorar e perguntas a serem respondidas. Esse foi um trabalho que acompanhou toda a execução, pois as ideias surgiam à medida que ficávamos mais íntimos dos dados utilizados.

Na a execução, dividimos as tarefas conforme os integrantes fossem se sentindo mais à vontade em dadas tarefas. Por exemplo, quem estivesse confortável em R, ficaria com o grosso das tarefas nessa área; quem estivesse à vontade com SGBD, tomaria a frente em PostgreSQL e SQL etc.

Quanto à comunicação, pensamos em pontos de controles a cada 3 dias quando possível, utilizando Google Meet, Whatsapp e Discord. Neste último, utilizamos para organizar toda a produção de conhecimento, não apenas para este Projeto Integrador, mas para futuros trabalhos dentro do curso de especialização. Lá, criamos canais especializados para cada tipo de interação: conversas gerais; materiais; links; recursos etc.

DISCIPLINA: ANÁLISE E MODELAGEM DE BANCO DE DADOS

Requisitos

Boa parte dos requisitos já foram apresentados na contextualização deste projeto.

A partir dos arquivos, iniciamos o trabalho de abstração e modelagem de dados. Percebemos que havia muita redundância de dados, exigindo uma normalização para obtermos um modelo mais enxuto.

Por exemplo, vimos que no arquivo *Indicators.csv* o nomes dos países se repetiam demais, bem como o nome dos indicadores. Isso poderia ser resolvido aplicando as formas normais.

Ao explorar os dados a fim de ganhar familiaridade, além da redundância, vimos que muitas colunas e arquivos que não nos seriam úteis.

O arquivo *.sqlite*, que provavelmente poderia nos dar algo estruturado em termos de objetos de bancos de dados relacional, também foi descartado por ter problema de redundância, apenas refletindo os arquivos *.csv*.

Assim, decidimos trabalhar com os seguintes arquivos/colunas:

Country.csv – contém as informações sobre os países à parte dos indicadores. Utilizamos as colunas *CountryCode*, *ShortName*, *LongName*, *Region*, *CurrencyUnit* das mais de 30 disponíveis.

< Country.csv (128.75 KB)

Detail

Compact

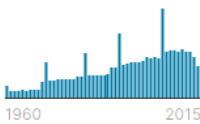
Column

About this file

US

| CountryCode | ShortName | TableName | LongName |
|----------------------|----------------------|----------------------|---|
| 247 unique values | 247 unique values | 247 unique values | 247 unique values |
| AFG | Afghanistan | Afghanistan | Islamic State of Afghanistan |
| ALB | Albania | Albania | Republic of Albania |
| DZA | Algeria | Algeria | People's Democratic Republic of Algeria |

Indicators.csv – é o arquivo central com as informações de países, indicadores aferidos e o ano do registro. *CountryCode*, *IndicatorCode*, *IndicatorName*, *Year* e *Value* foram as colunas aproveitadas.

| < Indicators.csv (547.7 MB) | | | | |
|---|----------------------|---|-----------------------|---|
| <div> Detail Compact Column </div> <div>6 of 6 columns</div> | | | | |
| About this file Values of different indicators for all the countries. | | | | |
| CountryName | CountryCode | IndicatorName | IndicatorCode | Year |
| 247 unique values | 247 unique values | 1344 unique values | 1344 unique values |  |
| Arab World | ARB | Adolescent fertility rate (births per 1,000 women ages 15-19) | SP.ADO.TFRT | 1960 |

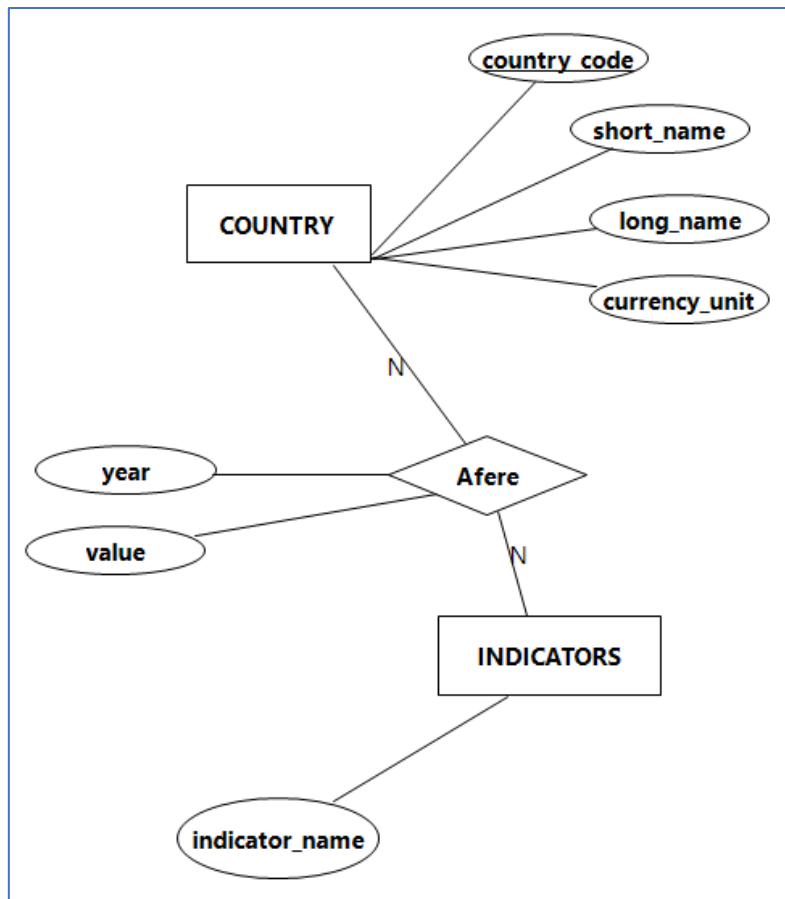
Series.csv – este arquivo contém informações detalhadas sobre os indicadores, se são econômicos, sociais, governamentais, financeiros, produtivos etc. Dele, nos interessa as colunas *SeriesCode* (Código do Indicador), *IndicatorName*, *Topic* (Classificação do indicador).

| < Series.csv (3.16 MB) | | | |
|---|--|--|---|
| <div> Detail Compact Column </div> | | | |
| About this file Information about indicators used for the entire data set | | | |
| SeriesCode | Topic | IndicatorName | ShortDefinition |
| 1345 unique values | Social Protection & ... 6% Economic Policy & ... 4% Other (1214) 90% | 1345 unique values | [null] Net official flo Other (105) |
| BN.KLT.DINV.CD | Economic Policy & Debt: Balance of payments: Capital & financial account | Foreign direct investment, net (BoP, current US\$) | |
| BX.KLT.DINV.WD.GD.ZS | Economic Policy & Debt: Balance of payments: Capital & financial account | Foreign direct investment, net inflows (% of GDP) | |

A partir da seleção acima, realizamos o diagrama de Entidade-Relacionamento e o Projeto Lógico de banco de dados.

Diagrama Entidade Relacionamento:

Da nossa análise, conseguimos extrair três entidades importantes REGION, COUNTRY e INDICATORS. Do relacionamento entre COUNTRY e INDICATORS, **Afere** vai nos fornecer os dados principais do modelo.



Projeto Lógico:

Do modelo anterior, e aplicando as formas normais, temos o seguinte resultado:

REGION{region_code, region_name}

TOPIC{topic_code, topic_description}

COUNTRY{country_code, short_name, long_name, currency_unit, region_code}

INDICATORS{indicator_code, indicator_name, topic_code}

INDICATORS_COUNTRY{indicator_code, country_code, year_indicator, value_indicator}

DISCIPLINA: INTRODUÇÃO AO GERENCIAMENTO DE BANCO DE DADOS

Os dados a serem utilizados nas queries desta seção foram obtidos dos arquivos .csv a partir do processo de importação utilizado a linguagem R. Os scripts encontram-se no arquivo *Script_Importacoes.R* e serão entregues com os demais artefatos do projeto. Portanto, antes de executar os SQLs desta seção, é necessária a execução do script no RStudio.

Optamos por reduzir o escopo dos dados importados para ganhar tempo. Assim, os dados utilizados dizem respeito a países das Américas e Caribe.

As queries construídas visam atender os requisitos exigidos pelo projeto integrador, bem como atender os comandos das seguintes consultas:

CONSULTA 1: PROJEÇÃO E SELEÇÃO

Objetivo da query: recuperar informações do indicador Taxa de Mortalidade Infantil – por cada 1000 pessoas nascidas vivas – (indicador *SP.DYN.IMRT.IN*) para o país Brasil por toda a série histórica disponível desde 1960 até 2015.

```
8 SELECT c.short_name PAÍS, i.indicator_code CODIGO_INDICADOR,  
9      i.indicator_name NOME_INDICADOR,  
10     ic.year_indicator ANO, ic.value_indicator VALOR  
11 FROM indicators_country ic  
12     INNER JOIN indicators i  
13         ON i.indicator_code = ic.indicator_code  
14     INNER JOIN country c  
15         ON c.country_code = ic.country_code  
16 WHERE 1 = 1  
17 AND i.indicator_code = 'SP.DYN.IMRT.IN'  
18 AND c.country_code = 'BRA'  
19 ORDER BY ic.year_indicator;
```

Resultado:

| Data Output | | Explain | Messages | Notifications | | |
|-------------|------------------------|---------|------------------------------------|-----------------------------------|-----------------|---------------------------|
| | país character (50) | | codigo_indicador character (30) | nome_indicador character (200) | ano smallint | valor double precision |
| 1 | Brazil | ... | SP.DYN.IMRT.IN | Mortality rate, infant ... | 1960 | 129.4 |
| 2 | Brazil | ... | SP.DYN.IMRT.IN | Mortality rate, infant ... | 1961 | 126.1 |
| 3 | Brazil | ... | SP.DYN.IMRT.IN | Mortality rate, infant ... | 1962 | 122.9 |
| 4 | Brazil | ... | SP.DYN.IMRT.IN | Mortality rate, infant ... | 1963 | 119.9 |
| 5 | Brazil | ... | SP.DYN.IMRT.IN | Mortality rate, infant ... | 1964 | 117.1 |
| 6 | Brazil | ... | SP.DYN.IMRT.IN | Mortality rate, infant ... | 1965 | 114.5 |
| 7 | Brazil | ... | SP.DYN.IMRT.IN | Mortality rate, infant ... | 1966 | 112.1 |

CONSULTA 2: JUNÇÃO EXTERNA

Objetivo da query: quais os indicadores foram aferidos (país, ano, valor não nulos) e quais não foram aferidos em 1960 no Brasil.


```

27 SELECT i.indicator_code CODIGO_INDICADOR,
28        i.indicator_name NOME_INDICADOR,
29        ic2.short_name PAIS,
30        ic2.year_indicator ANO, ic2.value_indicator VALOR
31 FROM indicators i
32     LEFT JOIN (
33         SELECT *
34         FROM indicators_country ic
35              INNER JOIN country c
36                   ON c.country_code = ic.country_code
37              WHERE 1 = 1
38                 AND ic.country_code = 'BRA'
39                 AND ic.year_indicator = 1960
40     ) ic2
41     ON ic2.indicator_code = i.indicator_code
42 ORDER BY i.indicator_code

```

Resultado:

| Data Output | Explain | Messages | Notifications | | |
|-------------|------------------------------------|---|------------------------|-----------------|---------------------------|
| | codigo_indicador character (30) | nome_indicador character (200) | pais character (50) | ano smallint | valor double precision |
| 79 | DC.DAC.AUSL.CD | Net bilateral aid flows from DAC donors, Aus... | [null] | [null] | [null] |
| 80 | DC.DAC.AUTL.CD | Net bilateral aid flows from DAC donors, Au... | Brazil | 1960 | -130000 |
| 81 | DC.DAC.BELL.CD | Net bilateral aid flows from DAC donors, Bel... | [null] | [null] | [null] |
| 82 | DC.DAC.CANL.CD | Net bilateral aid flows from DAC donors, Ca... | [null] | [null] | [null] |
| 83 | DC.DAC.CECL.CD | Net bilateral aid flows from DAC donors, Eur... | [null] | [null] | [null] |
| 84 | DC.DAC.CHEL.CD | Net bilateral aid flows from DAC donors, Swi... | [null] | [null] | [null] |
| 85 | DC.DAC.CZEL.CD | Net bilateral aid flows from DAC donors, Cze... | [null] | [null] | [null] |
| 86 | DC.DAC.DEUL.CD | Net bilateral aid flows from DAC donors, Ger... | Brazil | 1960 | 1260000 |
| 87 | DC.DAC.DNKL.CD | Net bilateral aid flows from DAC donors, De... | [null] | [null] | [null] |

CONSULTA 3: CONSULTA COM UNION, INTERSECT OU EXCEPT

Objetivo da query: quais os indicadores foram aferidos (país, ano, valor não nulos) e quais não foram aferidos em 1960 no Brasil.







OBSERVAÇÃO: O RESULTADO É O MESMO DA QUERY ANTERIOR.

```

51 SELECT i.indicator_code CODIGO_INDICADOR,
52        i.indicator_name NOME_IDICADOR,
53        c.short_name PAIS,
54        ic.year_indicator ANO,
55        ic.value_indicator VALOR
56 FROM indicators i
57      INNER JOIN indicators_country ic
58            ON ic.indicator_code = i.indicator_code
59      INNER JOIN country c
60            ON c.country_code = ic.country_code
61 WHERE 1 = 1
62 AND ic.country_code = 'BRA'
63 AND ic.year_indicator = 1960
64
65 UNION -----
66
67 SELECT i.indicator_code CODIGO_INDICADOR,
68        i.indicator_name NOME_IDICADOR,
69        --Como já sabemos que teremos apenas os indicadores que
70        --não tenham apuração em 1960, por conta da condição WHERE,
71        --os demais campos ficam como nulos
72        NULL PAIS,
73        NULL ANO,
74        NULL VALOR
75 FROM indicators i
76 WHERE NOT EXISTS (
77     SELECT 1
78     FROM indicators_country ic2
79     WHERE 1 = 1
80     AND ic2.indicator_code = i.indicator_code
81     AND ic2.country_code = 'BRA'
82     AND ic2.year_indicator = 1960
83 )

```

Resultado:

| | Data Output | Explain | Messages | Notifications | | |
|----|---|---|--|--|--|---|
| |  codigo_indicador character (30) |  nome_idicador character (200) |  pais character |  ano smallint |  valor double precision |  |
| 20 | DC.DAC.SWEL.CD | ... | Net bilateral aid flow... | [null] | [null] | [null] |
| 21 | NE.CON.TOTL.CD | ... | Final consumption e... | Brazil | 1960 | 12190115073.3641 |
| 22 | EN.POP.DNST | ... | Population density (p... | [null] | [null] | [null] |
| 23 | NE.EXP.GNFS.CN | ... | Exports of goods an... | Brazil | 1960 | 7.27491e-05 |
| 24 | CM.MKT.TRAD.CD | ... | Stocks traded, total v... | [null] | [null] | [null] |
| 25 | NY.GDP.MKTP.KN | ... | GDP (constant LCU) ... | Brazil | 1960 | 192725712900 |
| 26 | NV.IND.MANF.CN | ... | Manufacturing, value... | Brazil | 1960 | 0.0002617763 |
| 27 | FS.AST.PRVT.GD.ZS | ... | Domestic credit to pr... | [null] | [null] | [null] |
| 28 | DC.DAC.ISU.CD | ... | Net bilateral aid flow... | [null] | [null] | [null] |

CONSULTA 4: DIVISÃO RELACIONAL

Objetivo da query: quais os países que possuem todos os indicadores sobre Taxa de Alfabetização aferidos no ano de 1990.

São sete indicadores:

SE.ADT.LITR.FE.ZS

Literacy rate, adult female (% of females ages 15 and above)

| | |
|----------------------|--|
| SE.ADT.LITR.MA.ZS | Literacy rate, adult male (% of males ages 15 and above) |
| SE.ADT.LITR.ZS | Literacy rate, adult total (% of people ages 15 and above) |
| SE.ADT.1524.LT.FM.ZS | Literacy rate, youth (ages 15-24), gender parity index (GPI) |
| SE.ADT.1524.LT.FE.ZS | Literacy rate, youth female (% of females ages 15-24) |
| SE.ADT.1524.LT.MA.ZS | Literacy rate, youth male (% of males ages 15-24) |
| SE.ADT.1524.LT.ZS | Literacy rate, youth total (% of people ages 15-24) |

```

101 SELECT c.country_code, c.short_name, c.long_name
102 FROM country c
103 WHERE NOT EXISTS (
104     (SELECT DISTINCT 1
105     FROM indicators_country ic
106     WHERE 1 = 1
107     AND ic.indicator_code IN (
108         'SE.ADT.LITR.FE.ZS', 'SE.ADT.LITR.MA.ZS',
109         'SE.ADT.LITR.ZS', 'SE.ADT.1524.LT.FM.ZS',
110         'SE.ADT.1524.LT.FE.ZS', 'SE.ADT.1524.LT.MA.ZS',
111         'SE.ADT.1524.LT.ZS'
112     )
113     AND ic.year_indicator = 1990)
114     EXCEPT -----
115     (SELECT 1
116     FROM indicators_country ic
117     WHERE 1 = 1
118     AND ic.indicator_code IN (
119         'SE.ADT.LITR.FE.ZS', 'SE.ADT.LITR.MA.ZS',
120         'SE.ADT.LITR.ZS', 'SE.ADT.1524.LT.FM.ZS',
121         'SE.ADT.1524.LT.FE.ZS', 'SE.ADT.1524.LT.MA.ZS',
122         'SE.ADT.1524.LT.ZS'
123     )
124     AND ic.year_indicator = 1990
125     AND ic.country_code = c.country_code)
126 );

```

Resultado:

| Data Output | Explain | Messages | Notifications |
|------------------------------------|------------------------------|------------------------------|---------------|
| country_code [PK] character (3) | short_name character (50) | long_name character (150) | |
| 1 ECU | Ecuador | Republic of Ecuador | |
| 2 PAN | Panama | Republic of Panama | |
| 3 PRI | Puerto Rico | Puerto Rico | |
| 4 VEN | Venezuela | República Bolivarian... | |
| 5 MEX | Mexico | United Mexican Stat... | |
| 6 TTO | Trinidad and Tobag... | Republic of Trinidad | |

CONSULTA 5: AGREGAÇÃO + GROUP BY






Proósito: Consolidar o indicador Taxa Crescimento Populacional Anual do Brasil de forma que sejam realizadas as médias de cada década dentro da série histórica disponível. Série histórica: 1960-2015:

```

SELECT PAIS, DECADA || '0', MEDIA FROM (
  SELECT ic.country_code PAIS,
    SUBSTRING(ic.year_indicator::varchar(4),1,3) DECADA,
    avg(ic.value_indicator) MEDIA
  FROM indicators_country ic
  WHERE 1 = 1
  AND ic.country_code IN ('BRA')
  AND ic.indicator_code = 'SP.POP.GROW'
  GROUP BY ic.country_code, SUBSTRING(ic.year_indicator::varchar(4),1,3)
  ORDER BY ic.country_code, SUBSTRING(ic.year_indicator::varchar(4),1,3)
) SUBQUERY

```

Resultado:

| | Data Output | Explain | Messages | Notifications |
|---|---|--|---|---|
| |  pais character (3)  |  ?column? text |  media double precision |  |
| 1 | BRA | 1960 | 2.8501226236427515 | |
| 2 | BRA | 1970 | 2.431105763414003 | |
| 3 | BRA | 1980 | 2.138836334093512 | |
| 4 | BRA | 1990 | 1.58303682569185 | |
| 5 | BRA | 2000 | 1.2751034637282852 | |
| 6 | BRA | 2010 | 0.931358373736448 | |

DISCIPLINA: INTRODUÇÃO À ANÁLISE E CIÊNCIA DE DADOS

Nesta seção do projeto, vamos mostrar algumas análises realizadas a partir dos cenários elencados na seção de contextualização, bem como os scripts produzidos para viabilizá-las. Gráficos foram criados para ajudar na leitura.

Como parte dos requisitos do projeto, utilizado a Linguagem R nos trabalhos. Os scripts encontram-se no arquivo *Script_Manipulacao_Tibbles.R*, que vai compor os demais artefatos da entrega.

A seguir seguem os cenários, os insumos, visualizações e observações:

- **CENÁRIO 1:** a evolução de alguns indicadores sociais Brasil em comparação a alguns países ao longo de um dado período;

Indicadores:

SL.TLF.CACT.FE.NE.ZS Labor force participation rate, female (% of female population ages 15+) (national estimate)

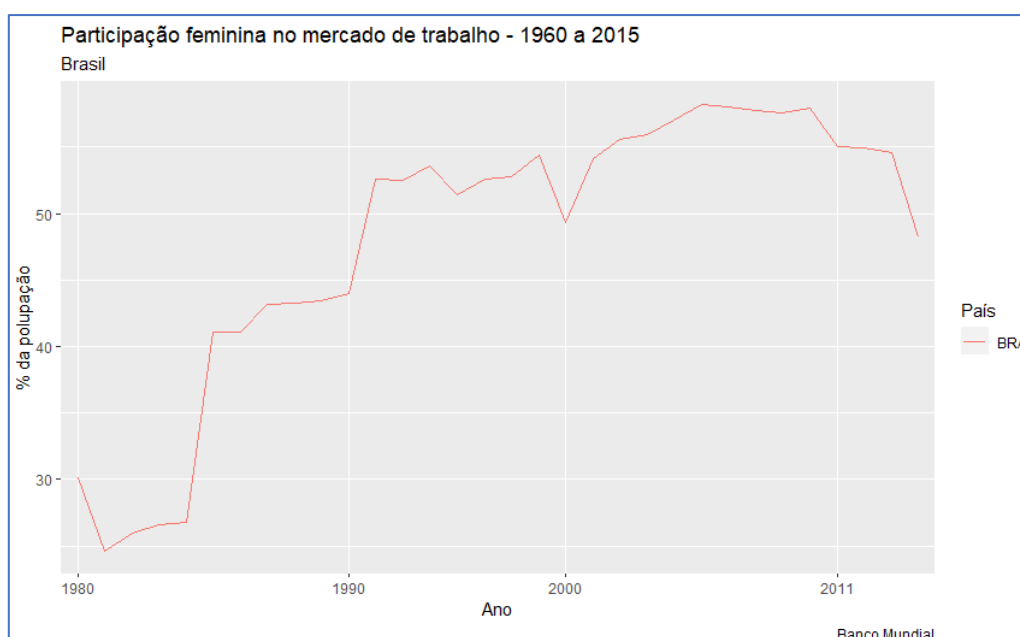
SP.MTR.1519.ZS Teenage mothers (% of women ages 15-19 who have had children or are currently pregnant)

SP.DYN.LE00.IN Life expectancy at birth, total (years)

Para o indicador da participação feminina no mercado de trabalho, utilizamos algumas funções de manipulação (select, filter, arrange) e para apresentação, um gráfico de linhas.

```
trabFeminino <- select(indicatorsCountries, CountryCode, IndicatorCode, Year, Value) %>%  
  filter(IndicatorCode == 'SL.TLF.CACT.FE.NE.ZS' &  
         CountryCode %in% c('BRA'))  
  ) %>%  
  arrange(CountryCode, Year)  
  
ggplot(data = trabFeminino) +  
  geom_line(aes(x=as.factor(Year), y=value, group=CountryCode, colour=CountryCode)) +  
  labs(x = 'Ano', y = '% da população',  
       colour='País',  
       title = 'Participação feminina no mercado de trabalho - 1960 a 2015',  
       subtitle = 'Brasil',  
       caption = 'Banco Mundial') +  
  scale_x_discrete(breaks = c(1960,1970,1980,1990,2000,2011,2015))
```

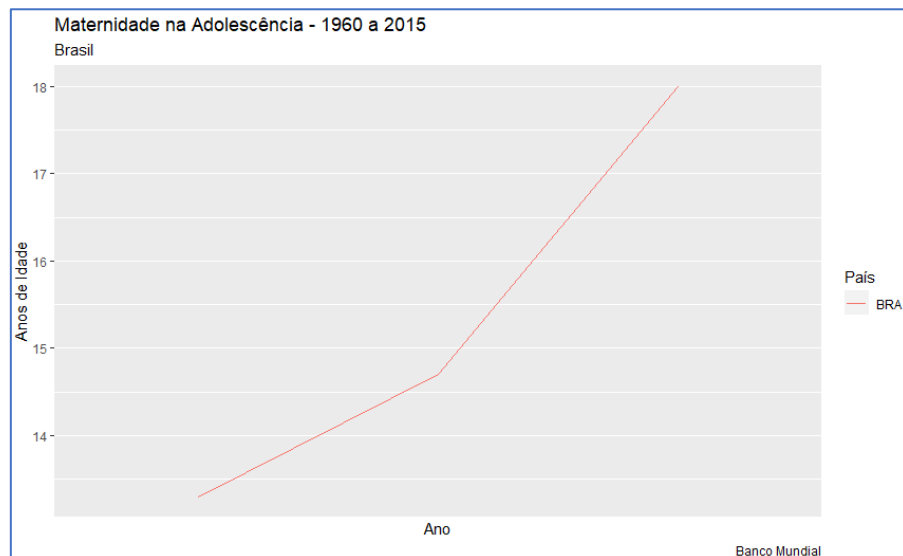
Gráfico:



Observações: por volta de meados dos anos 80, algo fez com que a participação saltasse mais de 10 pontos percentuais. O mesmo nota-se no início dos anos 90 com quase 8 pontos. A partir de 2010 a 2014 percebe-se um descenso significativo de aproximadamente 9 pontos, com tendência a voltar para os números de 1990.

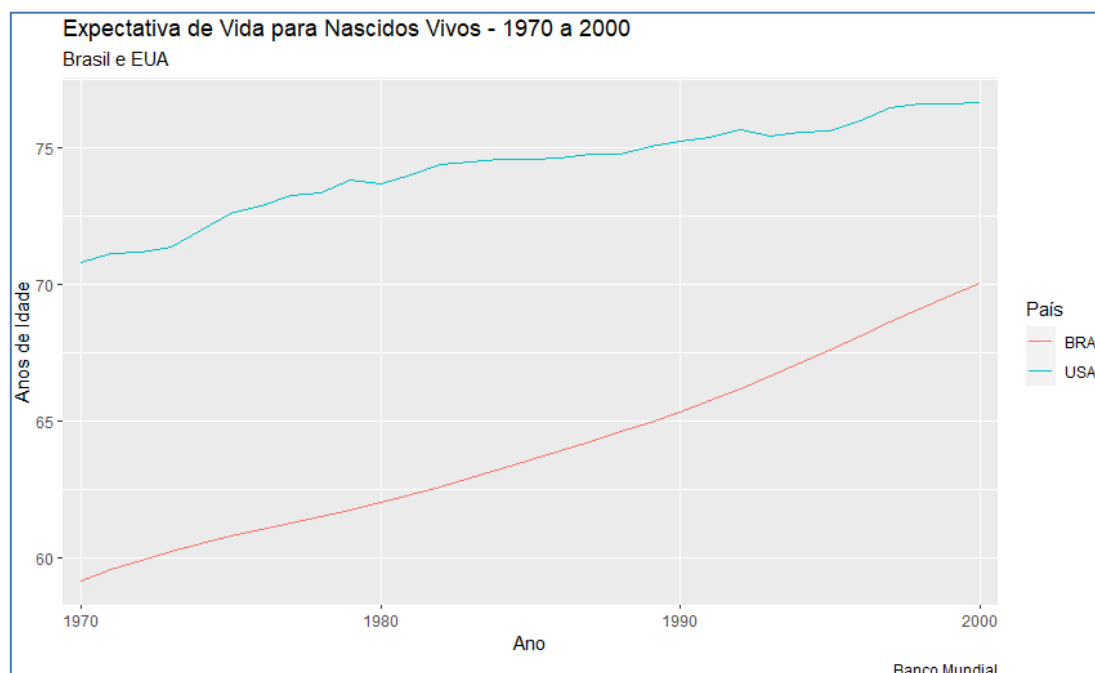
Para o indicador sobre maternidade na adolescência, o trabalho foi quase o mesmo do cenário anterior, alterando algumas variáveis.

Gráfico:



Observações: é um indicador com poucas aferições. Destaque para países subdesenvolvidos como Brasil e Bolívia que tem alguns dados. Em países desenvolvidos, esses dados praticamente não existem.

Para o indicador sobre expectativa de vida, também fizemos praticamente uso das mesmas funções e gráfico.



Observação: o Brasil atingiu a expectativa de vida de 70 anos de idade por volta do ano 2000. Essa mesma expectativa já era uma realidade nos EUA em 1970.

- **CENÁRIO 2:** crescimento médio por decênio da população do Brasil a partir de 1960 a 2015.

Indicador:

SP.POP.GROW *Population growth (annual %)*

Para este cenário foi necessária a criação de uma função `getDecade()` para auxiliar a consolidação de dados anuais em médias decenais.

```
getDecade <- function(lista) {
  if (is.null(lista) || length(lista) == 0) {
    stop('Parâmetro nulo ou vazio.')
  }
  if (any(is.na(lista))) {
    warning('Existe valores NA no dataset passado.')
  }

  retorno <- c()
  referencia <- substring(lista[1],1,3)

  for(elemento in lista) {
    ano <- substring(elemento,1,3)
    if(ano != referencia) {
      referencia <- substring(elemento,1,3)
    }
    decada <- paste0(ano,'0')
    retorno <- c(retorno, decada)
  }
  retorno
}
```

Na manipulação, um pipe com funções `select`, `filter`, `arrange`, `mutate`, `group_by` e `summarise` foram utilizados para geração do tibble.

```
crescPopBrasil <-
  select(indicatorsCountries, CountryCode, IndicatorCode, Year, value) %>%
  filter(IndicatorCode == 'SP.POP.GROW' &
         CountryCode == 'BRA'
  ) %>%
  mutate(Decade=getDecade(Year)) %>%
  arrange(CountryCode, Year) %>%
  select(Decade, value) %>%
  group_by(Decade) %>%
  summarise(Mean=mean(value))
```

Para visualização, utilizamos as funções `tableGrob()` e `grid.arrange()` das libs *gridExtra*, *grid* e *gtable*

| Decade | Mean |
|--------|-----------|
| 1960 | 2.8501226 |
| 1970 | 2.4311058 |
| 1980 | 2.1388363 |
| 1990 | 1.5830368 |
| 2000 | 1.2751035 |
| 2010 | 0.9313584 |

Observação: aqui, nota-se a diminuição da taxa de crescimento populacional com o passar das décadas. A informação nos anos 2010 está incompleta por haver dados até 2014.

- **CENÁRIO 3:** populações urbana x rural comparativo entre 1960 e 2015.

Indicadores:

SP.URB.TOTL.IN.ZS Urban population (% of total)

SP.RUR.TOTL.ZS Rural population (% of total population)

Os tibbles foram montados a partir da criação de um para população urbana (*popUrbana*) e outro para a rural (*popRural*), aplicando filtros, seleções e ordenamento. Daí unificou-se ambos em um só (*popsRuralUrbana*), e foi este último o tibble de interesse.

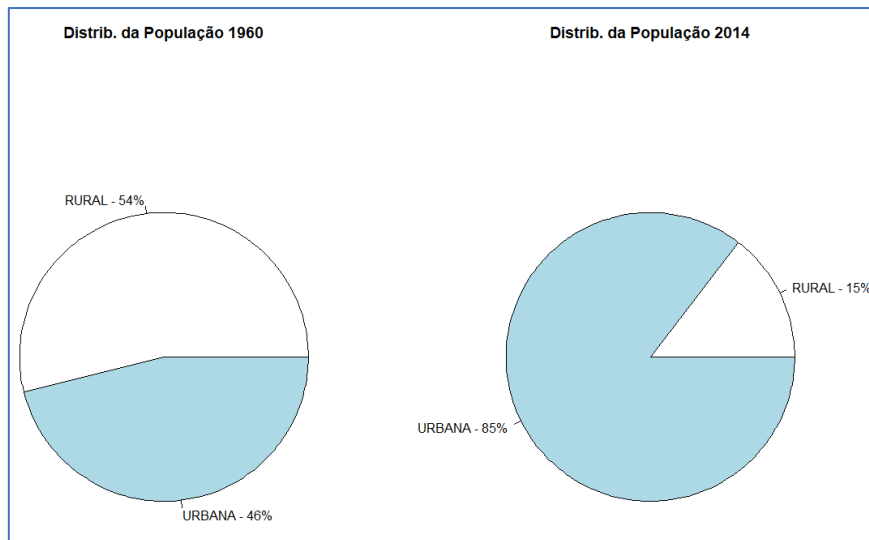
```
popurbana <- select(indicatorsCountries, CountryCode, IndicatorCode, Year, value) %>%
  filter(IndicatorCode == 'SP.URB.TOTL.IN.ZS' &
        CountryCode %in% c('BRA') &
        Year %in% c(1960, 2014))
) %>%
  arrange(CountryCode, Year)
popRural <- select(indicatorsCountries, CountryCode, IndicatorCode, Year, value) %>%
  filter(IndicatorCode == 'SP.RUR.TOTL.ZS' &
        CountryCode %in% c('BRA') &
        Year %in% c(1960, 2014))
) %>%
  arrange(CountryCode, Year)

popsRuralUrbana <- union(popRural, popurbana) %>%
  mutate(Zona = ifelse(substr(IndicatorCode, 4, 6) == 'RUR', 'RURAL', 'URBANA')) %>%
  arrange(Year)
```

Para apresentação, fizemos uso de dois gráficos setoriais.

```
par(mfrow=c(1,2))
#values absolutos
valuesurbana <- popsRuralUrbana[1:2,]$value
#values percentuais
percenturbana <- round(valuesurbana/sum(valuesurbana)*100)
#Captura as legendas
labelsurbana <- popsRuralUrbana[1:2,]$Zona
#Concatena as legendas como os percentuais
labelsurbana <- paste0(labelsurbana, ' - ', percenturbana, '%')
#Plota o gráfico de setores
pie(
  valuesurbana,
  labels=labelsurbana,
  main="Distrib. da População 1960"
)

#values absolutos
valuesRural <- popsRuralUrbana[3:4,]$value
#values percentuais
percentRural <- round(valuesRural/sum(valuesRural)*100)
#Captura as legendas
labelsRural <- popsRuralUrbana[3:4,]$Zona
#Concatena as legendas como os percentuais
labelsRural <- paste0(labelsRural, ' - ', percentRural, '%')
#Plota o gráfico de setores
pie(
  valuesRural,
  labels=labelsRural,
  main="Distrib. da População 2014"
)
```

- **CENÁRIO 4:** Crescimento da taxa de mortalidade infantil abaixo de 5 anos (a cada 1000 nascidos vivos) ao longo de um período no Brasil comparado com alguns países:

Indicador:

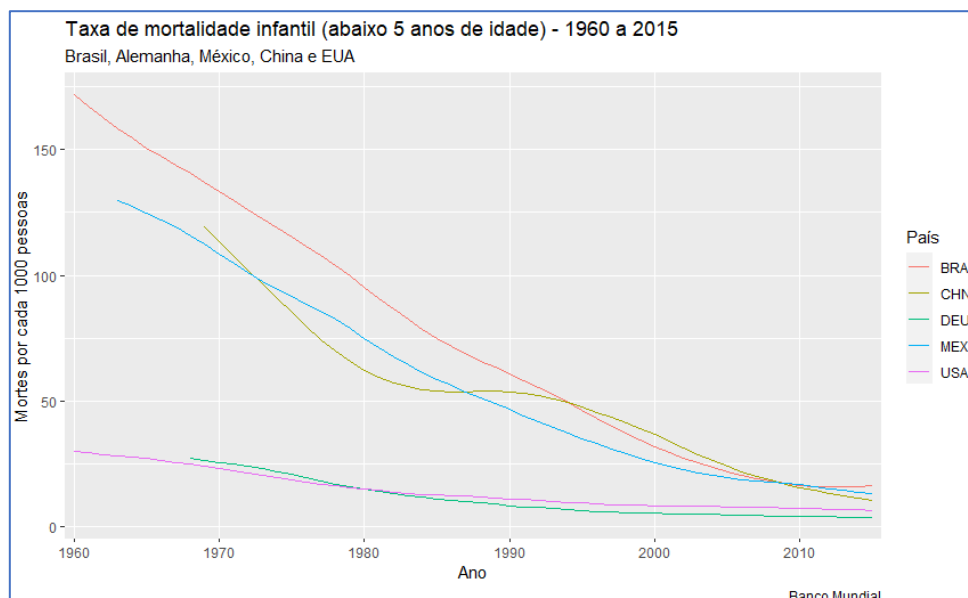
SH.DYN.MORT *Mortality rate, under-5 (per 1,000 live births)*

Aqui, como nos tibbles anteriores, aplicamos praticamente as mesmas funções de manipulação de dados.

```
txMortesBrasilOutros <-
  select(indicatorsCountries, CountryCode, IndicatorCode, Year, value) %>%
  filter(IndicatorCode == 'SH.DYN.MORT' &
         CountryCode %in% c('MEX', 'DEU', 'BRA', 'CHN', 'USA')) %>%
  arrange(CountryCode, Year)
txMortesBrasilOutros

ggplot(data = txMortesBrasilOutros) +
  geom_line(aes(x=as.factor(Year), y=value, group=CountryCode, colour=CountryCode)) +
  labs(x = "Ano", y = "Mortes por cada 1000 pessoas",
       colour="País",
       title = "Taxa de mortalidade infantil (abaixo 5 anos de idade) - 1960 a 2015",
       subtitle = "Brasil, Alemanha, México, China e EUA",
       caption = "Banco Mundial") +
  scale_x_discrete(breaks = c(1960, 1970, 1980, 1990, 2000, 2010))
```

Na apresentação, utilizamos um gráfico de linha, que se mostrou bem efetivo para mostrar comparativos ao longo do tempo.



Observação: a taxas dos países subdesenvolvidos estão muito próximas dos países ricos desde o ano de 2010.

- **CURIOSIDADES:** aqui selecionamos alguns indicadores curiosos dentre os muitos que haviam disponíveis.

Indicadores:

| | |
|----------------|--|
| SH.MED.PHYS.ZS | Physicians (per 1,000 people) |
| SG.VAW.REAS.ZS | Women who believe a husband is justified in beating his wife (any of five reasons) (%) |
| SG.VAW.ARGU.ZS | Women who believe a husband is justified in beating his wife when she argues with him (%) |
| SG.VAW.BURN.ZS | Women who believe a husband is justified in beating his wife when she burns the food (%) |
| SG.VAW.GOES.ZS | Women who believe a husband is justified in beating his wife when she goes out without telling him (%) |
| SG.VAW.NEGL.ZS | Women who believe a husband is justified in beating his wife when she neglects the children (%) |
| SG.VAW.REFU.ZS | Women who believe a husband is justified in beating his wife when she refuses sex with him (%) |

Para o indicador de físicos formados para cada 1000 pessoas, usamos as seguintes manipulações, sem muitas alterações dos demais cenários:

```
fisicosPaises <- select(indicatorsCountries, CountryCode, IndicatorCode, Year, value) %>%
  filter(IndicatorCode == 'SH.MED.PHYS.ZS' &
         CountryCode %in% c('BRA', 'CHN', 'USA', 'DEU'))
) %>%
  arrange(CountryCode, Year)

ggplot(data = fisicosPaises) +
  geom_line(aes(x=as.factor(Year), y=value, group=CountryCode, colour=CountryCode)) +
  labs(x = "Ano", y = "Quant. por 1000 pessoas",
       colour="País",
       title = "Físicos - Quantidade - 1960 a 2015",
       subtitle = "Brasil, China, EUA, Alemanha",
       caption = "Banco Mundial") +
  scale_x_discrete(breaks = c(1960, 1970, 1980, 1990, 2000, 2010))
```

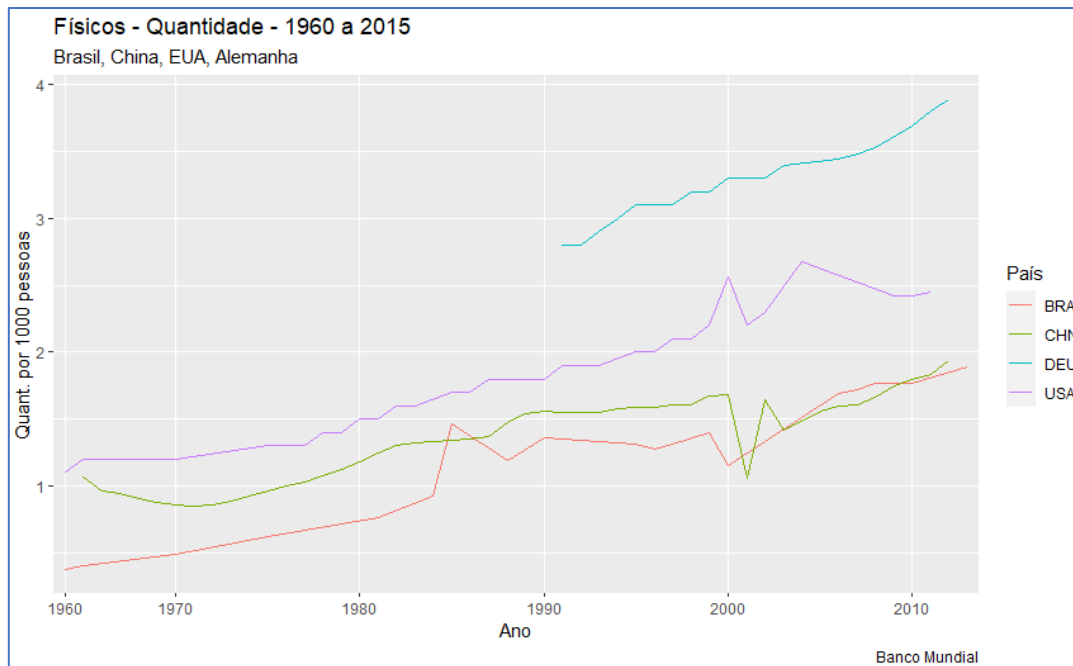
Além dos físicos, há apenas outro profissional em destaque nos indicadores são os que trabalham com saúde.

Comparativo: Brasil, China, EUA, Alemanha

O comparativo teve como critério saber como o Brasil se compara com dois países tradicionalmente desenvolvidos (sendo um nas Américas e outro na Europa), e a China, que se desenvolveu recentemente.

Observações: O resultado se mostrou interessante, pois achávamos que os EUA estariam muito à frente do Brasil nesse campo. A comparação com a China se mostra equilibrada. Ponto para a Ciência brasileira! :D

A Alemanha é que não se esperava que estivesse tão à frente dos EUA.



Por fim, escolhemos um indicador sobre a violência contra a mulher.

```
violenciaMulher <- filter(indicatorsCountries,
                           value > 70 &
                           IndicatorCode == 'SG.VAW.REAS.ZS') %>%
  select(CountryName, Year, value) %>%
  arrange(desc(value))

#Plota uma tabela como gráfico
ss <- tableGrob(violenciaMulher, rows=NULL)
grid.arrange(ss)
```

A curiosidade aqui é quanto a mulheres que acreditam que seus maridos têm razão em bater nelas¹. Isso reflete o grau de cultura que determinados países têm em relação ao papel dos sexos na sociedade. Aqui, selecionamos os casos que consideramos mais notáveis (acima de 70%).

| CountryName | Year | Value | | | |
|--------------|------|-------|------------------|------|------|
| Guinea | 2012 | 92.1 | Somalia | 2006 | 75.7 |
| Mali | 2001 | 88.8 | Congo, Dem. Rep. | 2007 | 75.6 |
| Timor-Leste | 2010 | 86.2 | Mali | 2006 | 75.2 |
| Guinea | 2005 | 85.6 | Congo, Dem. Rep. | 2014 | 74.8 |
| Zambia | 2002 | 85.4 | Tajikistan | 2005 | 74.4 |
| Sierra Leone | 2005 | 85.0 | Gambia, The | 2006 | 74.0 |
| Ethiopia | 2000 | 84.5 | Burundi | 2010 | 72.9 |
| Ethiopia | 2005 | 81.0 | Burkina Faso | 2006 | 71.4 |
| Uganda | 2001 | 76.5 | Burkina Faso | 2003 | 71.1 |
| Mali | 2013 | 76.3 | Eritrea | 2002 | 70.7 |
| | | | Uganda | 2006 | 70.2 |
| | | | Niger | 2006 | 70.1 |

Observações: muitos dos países com dados a este respeito são países africanos, asiáticos, muçulmanos, pobres ou uma combinação de alguns destes.

¹ Não sabemos se o número se refere ao percentual das entrevistadas ou em relação à população total.

Não se sabe se países mais avançados econômica, política e socialmente possuem dados a este respeito e não foram divulgados.

Há alguns outros indicadores semelhantes a este (citados no início deste cenário) diferenciando tão somente nos motivos pelos quais a violência ocorre.