Project review

Andrei Teixeira and Rafael Gouveia
Visualização Avançada de Dados
Mestrado em Engenharia e Ciência de Dados (MECD)

Estrutura do doc:

I. Autores

Andrei Fokin Teixeira
- email: andrei.fokin.teixeira@gmail.com
- #UC: 2022135701

Rafael Santos Gouveia
- email: rafaelgou1110@gmail.com
- #UC: 2022130518

II. Introdução

A História da humanidade está repleta de eventos de grande impacto como guerras, crises econômicas e sanitárias ou alterações políticas das mais diversas naturezas e conforme o tempo passa, maior a frequência de eventos por conta dos avanços tecnológicos e sociais que aumenta a velocidade das interações e mais ampla é a reação em diferentes povos e países devido à maior inter-relação entre estes.

Ser capaz de entender como um evento impacta determinada estatística de uma população é interessante para uma melhor tomada de decisão de políticas no nível público e de investimentos no nível privado. Muitas vezes, ações são tomadas baseadas em ideias pré-concebidas sobre um determinado tema, em ideologias e/ou preconceitos, em teorias ou por mera repetição de ideias populares.

O mundo também está cada vez mais repleto de dados e existe um potencial enorme para buscar tais padrões e extrair valor das informações obtidas. E para tal, usar não apenas comunicação verbal, mas também comunicação visual em novos estudos e plataformas.

Nossa calculadora, "InfoCountry", é uma plataforma para visualizar a evolução de dados socioeconômicos dos países ao longo do tempo e permite a comparação entre um ou mais países, podendo-se escolher um deles como referência. Nesta versão, é disponibilizada uma lista dos principais eventos nacionais de 39 países europeus desde o ano de 1945 (ano do fim da Segunda Guerra Mundial).

A ferramenta permite explorar o comportamento e comparar a performance média de uma variável socioeconômica de um ou mais países em relação a um país de referência em dois

períodos diferentes de tempo, tendo como marco separador um um evento marcante para a história do país-referência, por exemplo, a Reunificação da Alemanha, a entrada/saída de um país na União Europeia ou a desintegração da antiga Iugoslávia. Logo, o principal papel é ser uma ferramenta de análise exploratória voltada para um maior potencial visual.

Por isso, este relatório se propõe a descrever com detalhes técnicos as características e a viabilidade da ferramenta a nível de código e de visualização.

III. Trabalhos relacionados

Quando se trata de desenvolvimento dos países, comparações são feitas com frequência, ainda mais para destacar exemplos de sucessos a se seguir. No Brasil, por exemplo, muito se comenta da Coreia do Sul quando o assunto é renda *per capita*. A Imagem 1 ilustra a diferença dessa variável para os dois países. A fonte é o site do Countryeconomy, que compara países 2 a 2, mas que trabalha com informações de longo prazo.

Outro exemplo, na Imagem 2, cuja fonte é o site do Banco Mundial, comparando N países, mas com menos dados disponíveis, é o de Portugal e Espanha em relação a alguns países bálticos, como Lituânia e Estônia, onde o flip da renda *per capita* é mais recente.

Visualizações como estas são bem comuns de se encontrar e geralmente estão no formato de gráfico de linhas, o que ajuda muito na percepção da tendência e mostra a aproximação ou o distanciamento dos países. Para complementar essa visualização, seria interessante uma visão comparativa e que mostrasse algum evento importante que explica o começo de divergência/ convergência de tendências entre séries temporais.

Já quando se trata de comparações no mesmo instante do tempo, geralmente se observam os valores brutos das variáveis de todos os países, como por exemplo em mapas-mundi. É possível extrair muita informação dessa maneira, principalmente se variáveis estão ao redor de zero ou de um valor de referência que é conhecido e usado por um grupo de pessoas. O problema aqui é que se perde a dimensão temporal. A Imagem 3 ilustra uma comparação de valores absolutos em torno de zero.

Com relação à interação, um dos sites mais conhecidos é o Gapminder, que com sua ferramenta Gapminder Tools, têm-se a dinâmica dos dados ao longo do tempo para diferentes países e em diferentes tipos de gráficos. É algo bem completo, mas a informação fica perdida pela alta performance visual. A Imagem 4 mostra a quantidade de possibilidades oferecidas pelo site.

Um tipo de comparação menos comum e que é um dos objetos deste trabalho, são comparações em termos percentuais entre poucos países e para isso, é necessário adotar um país como referência. Na nossa “calculadora”, intitulada "InfoCountry" será adicionada informação sobre os principais eventos dos países para que o utilizador tenha um tipo adicional de iteração: a iteração buscando relacionar os números com acontecimentos do mundo real. Para ilustrar o interesse deste trabalho, foi criada uma visualização simples em Tableau e olhando apenas uma variável: crescimento anual do PIB (%). O Mockup 1, de alta fidelidade, mostra a comparação da evolução da variável de interesse para Alemanha e Suíça em relação à Áustria e tomando como marco o ano de 1995, quando este último entrou na UE.

IV. Requisitos de design

De maneira geral, gostávamos de ter uma visualização em que seja possível escolher diferentes países para consulta,utilizando botões onde poderá ser selecionados até 5 países, e observar eventos nacionais que são marcos para os países selecionados e assim utilizar dois mapas como referência visual para comparação entre diferentes variáveis socioeconômicas, poderá ser selecionada no máximo de 3 dessas variáveis, e também utilizar a média como termo de

comparação entre os países selecionados ou todos os países na nossa linha de pesquisa.

Além dos dois mapas seria utilizado gráficos de linhas, onde seria mostrado a evolução do países conforme o passar do tempo. Quanto a variável tempo seria selecionada períodos por meio de slider e assim poderíamos fazer a seleção de diferentes momentos para capa mapa, feito de maneira separada. Todas as seleções, alterações, etc que o usuário poderá fazer irá ter efeito em todas as visualizações disponíveis no dashboard. Logo, há três grandes blocos de visualização, que são: (i) os eventos; (ii) os países comparados; e (iii) os países individualizados.

De maneira específica, assim poderia ser para países teríamos botão para tornar um país como referência (selector) e para eventos será utilizado scroll para intervalo de tempo dos eventos (seletor) e uma tabela para mostrar eventos dos países (resultado).

- Específicos para o comparador de países:

>> Como seletor será utilizado dropdown para países scroll para selecionar o intervalo de tempo e checkboxes para variáveis.

>> Como visualizador de resultados teremos mapa para mostrar a comparação das médias, gráfico de barras horizontal para comparar a média da variável e comparar a diferença média entre países, tomando tendo um como referência.

- Específicos para a informação individual de países:

>> Será utilizado checkboxes para variáveis, dropdown para países e scroll para selecionar o intervalo de tempo como seletores. E para resultados terá gráfico de linhas para mostrar a evolução individual de cada país.

Se for tecnicamente viável, outro requisito será o seguinte: ao fazer o select de em uma nova variável, mostrar um novo par de gráficos e assim por diante para cada novo select (todos submetidos ao mesmo intervalo de tempo do Mockup 1). O Mockup 2 (também de alta fidelidade) ilustra a proposta para duas ou mais variáveis. Pode-se perceber que aparecem novos gráficos e novos mapas. Durante a fase de código, serão descobertas as potencialidades da ferramenta.

<Mockup 1: foi disponibilizado na fase de proposta>

<Mockup 2: está disponibilizado em documento Tableau nesta fase de revisão>

V. Fonte dos dados

O projeto conta quatro variáveis de interesse em frequência anual: crescimento do PIB (%), PIB *per capita* (US$), taxa de inflação (%) e expectativa de vida. São extraídos do site do Banco Mundial, em formato tabular e pronto para serem trabalhados posteriormente em Python enquanto dataframe. Já os dados dos eventos são obtidos de duas maneiras: (i) a partir de Web Scraping realizado em páginas da Wikipédia onde o dado estava devidamente tabelado [válido para 25 países]; e (ii) criados manualmente a partir de páginas da Wikipédia em que a informação estava disponível, mas não em formato tabular [válido para 2 países].

A Tabela 1 mostra um resumo da origem dos dados.

VI. Análise exploratória dos dados

Em documento *Jupyter Lab* compartilhado, foram realizados dois tipos de análise:

- uma estática: um espaço onde são apresentados insights sobre os dados extraídos; e
- uma dinâmica: um código onde o próprio usuário que está a ler pode escolher valores para os parâmetros e ver o que acontece → permite os professores explorarem conosco.

Análise estática:

1. análise geral de métricas e missing values;

2. evolução das variáveis e suas médias ao longo do tempo (na Imagem 5a, o exemplo é com a variável "*gdp_growth*");

3. tendência das variáveis ao longo do tempo em grupos diferentes: países do Leste e do Oeste europeu (na Imagem 5b, o exemplo da variável "*gdp_growth*" mostra que a tendência para ambos os grupos de países é de queda no crescimento; o gráfico da esquerda vem desde 1960 com a totalidade dos dados e a da direita desde 1996 com a maior quantidade de dados por país);

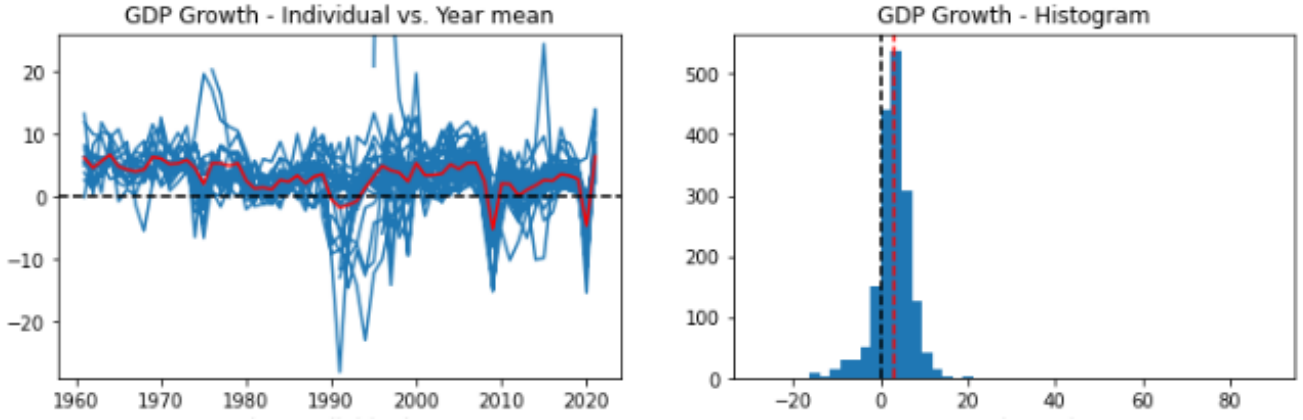


Imagem 5a. Exemplo de evolução de variável (todos os países e média geral + histograma).

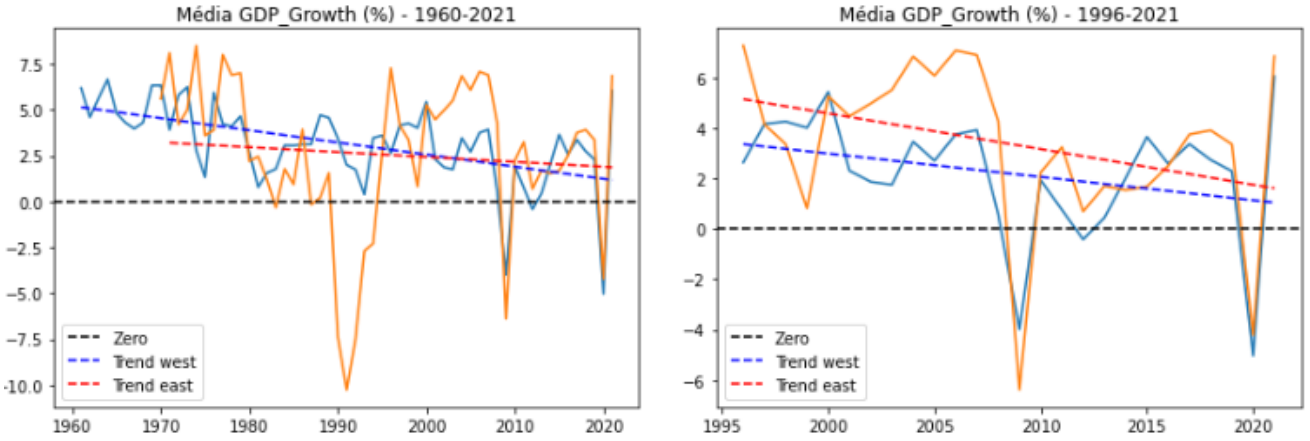


Imagem 5b. Exemplo de evolução de variável (média dos países do Leste e do Oeste, em dois intervalos de tempo).

4. valor médio da variável ao longo do tempo em dois momentos do tempo (na Imagem 6a, o exemplo é com a variável "*gdp_per_capita*" desde 1960 → 16 países estão acima da média; já na Imagem 6b, desde 1998, com a maioria dos dados → 14 países estão acima da média e os valores médios são maiores).

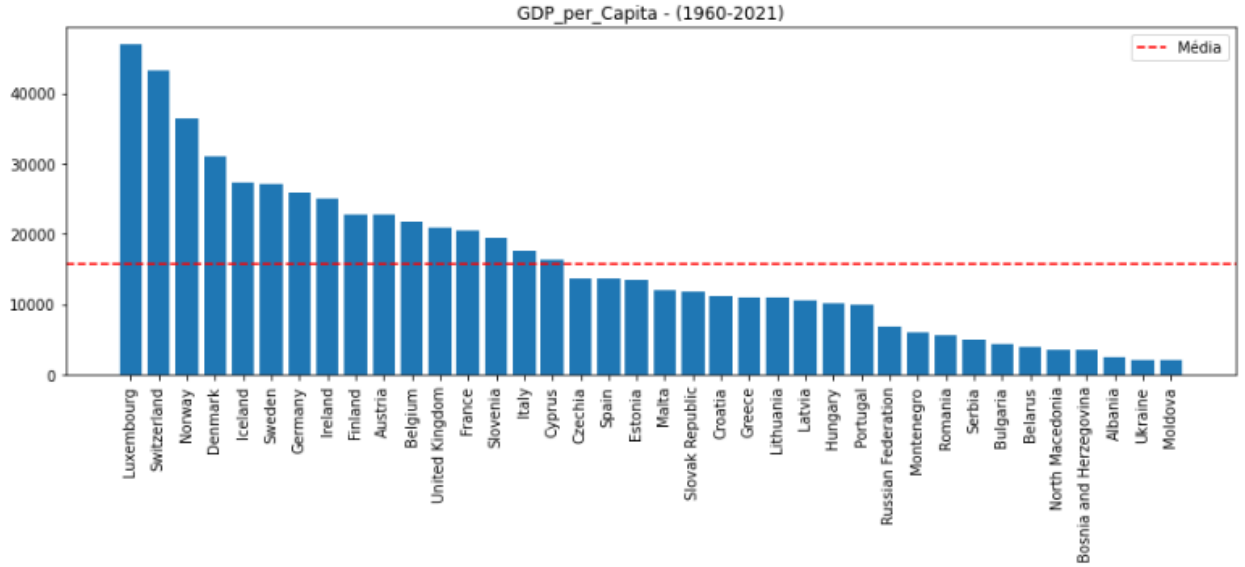


Imagem 6a. Exemplo de evolução de variável (média dos países do Leste e do Oeste, em dois intervalos de tempo).

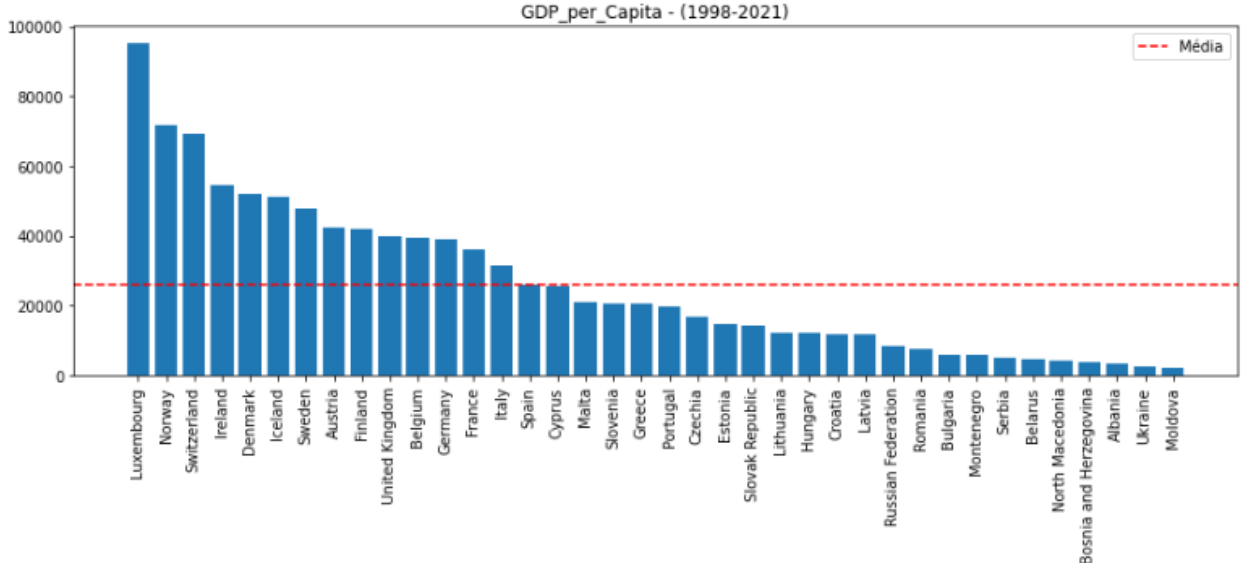


Imagem 6b. Exemplo de evolução de variável (média dos países do Leste e do Oeste, em dois intervalos de tempo).

5. Correlação:
- de todo o período para todos os países;
- de todo o período para Oeste e para Leste (Imagem 7a);
- de dois blocos temporais para todos os países (Imagem 7b);
- de dois blocos temporais para Oeste e para Leste;
- década a década para todos os países (Imagem 7c);
- década a década para Oeste e para Leste.

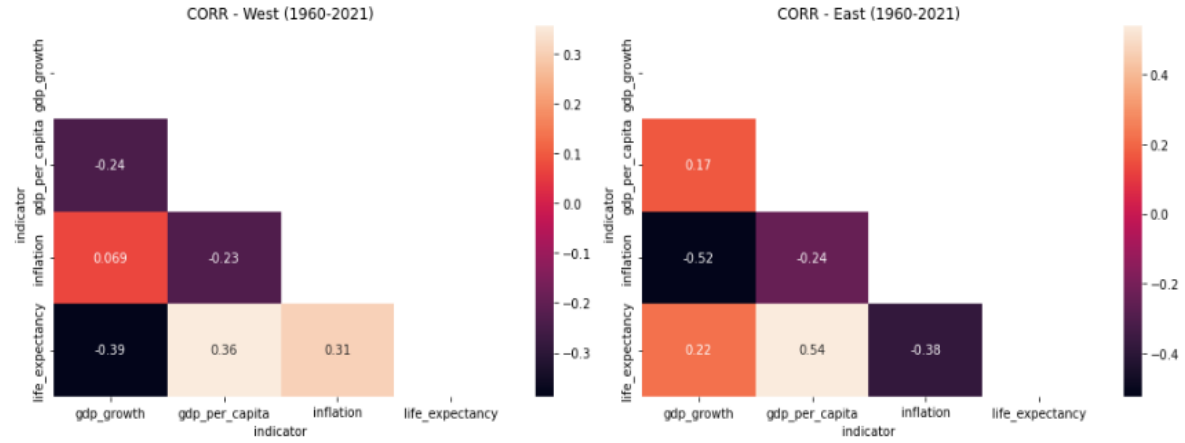


Imagem 7a. Correlação de dois grupos de países, no mesmo momento do tempo.

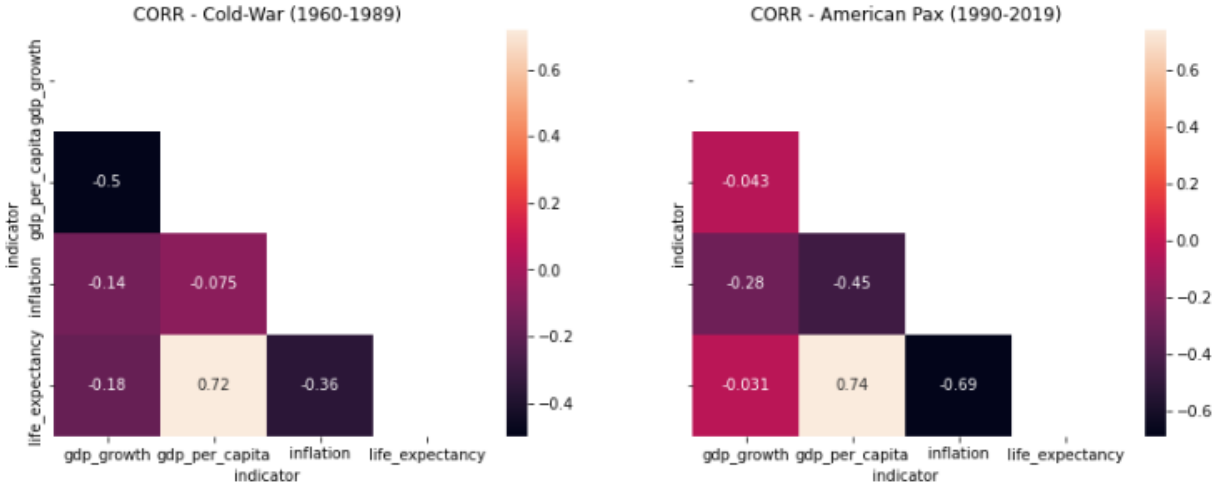


Imagem 7b. Correlação de todos os países, em momentos diferentes do tempo.

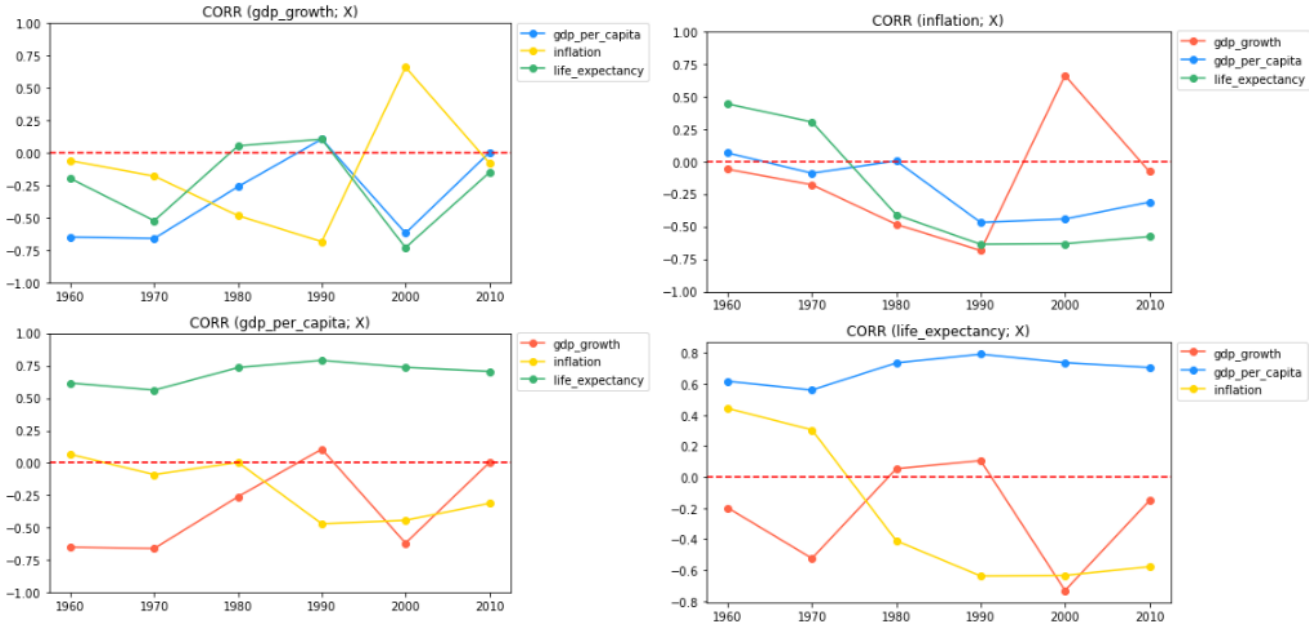


Imagem 7c. Evolução das correlações de todas as variáveis em relação às demais e ao longo do tempo.

6. Distribuição:
- de dois blocos temporais para todos os países;
- de dois blocos temporais para Oeste e para Leste;

- década a década para todos os países (na Imagem 8a, o exemplo é a variável "*gdp_per_capita*", onde percebe-se que a renda tem aumentado junto com a desigualdade);
- década a década para Oeste e para Leste (na Imagem 8b, o exemplo é com a variável "life_expectancy", onde percebe-se um aumento da idade média em ambos os grupos).

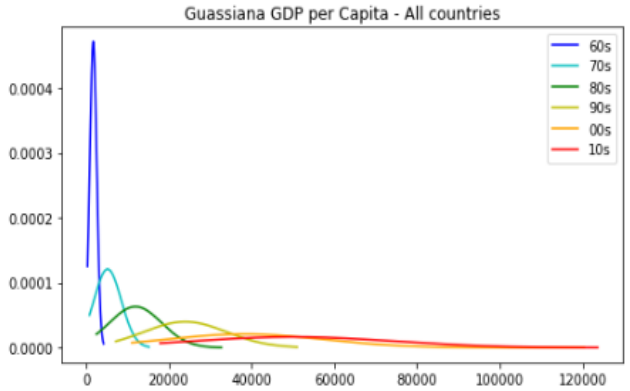


Imagem 8a. Evolução das distribuições da renda per capita para todos os países.

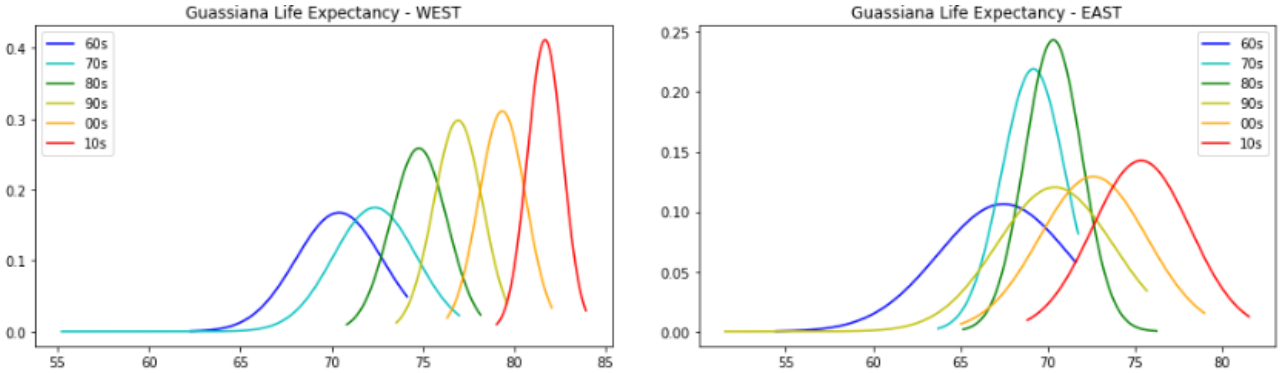


Imagem 8b. Evolução das distribuições da expectativa de vida para os países do Oeste e os do Leste.

Análise dinâmica:

1. valor médio de cada variável para um país em dois momentos do tempo: um lag de anos antes e um depois do ano de um evento-marco do país (exemplo na Imagem 7: inflação alemã 10 anos antes e 10 anos depois em relação à 1989, ano da queda do Muro de Berlim); → 4 variáveis.
2. idem acima, mas para dois ou mais países; → 5+ variáveis!
3. para este segundo cenários, é visualizada a comparação entre o valor absoluto, a diferença absoluta e a diferença percentual (imagem 8).

Para GDP per Capita

```
# Repetindo, mas agora para mais países, tomando um deles como referência
event = 'Queda do Muro de Berlim'
country_ref = 'Germany'
year = 1989
country_2 = 'France'
country_3 = 'Italy'
lag = 10
```

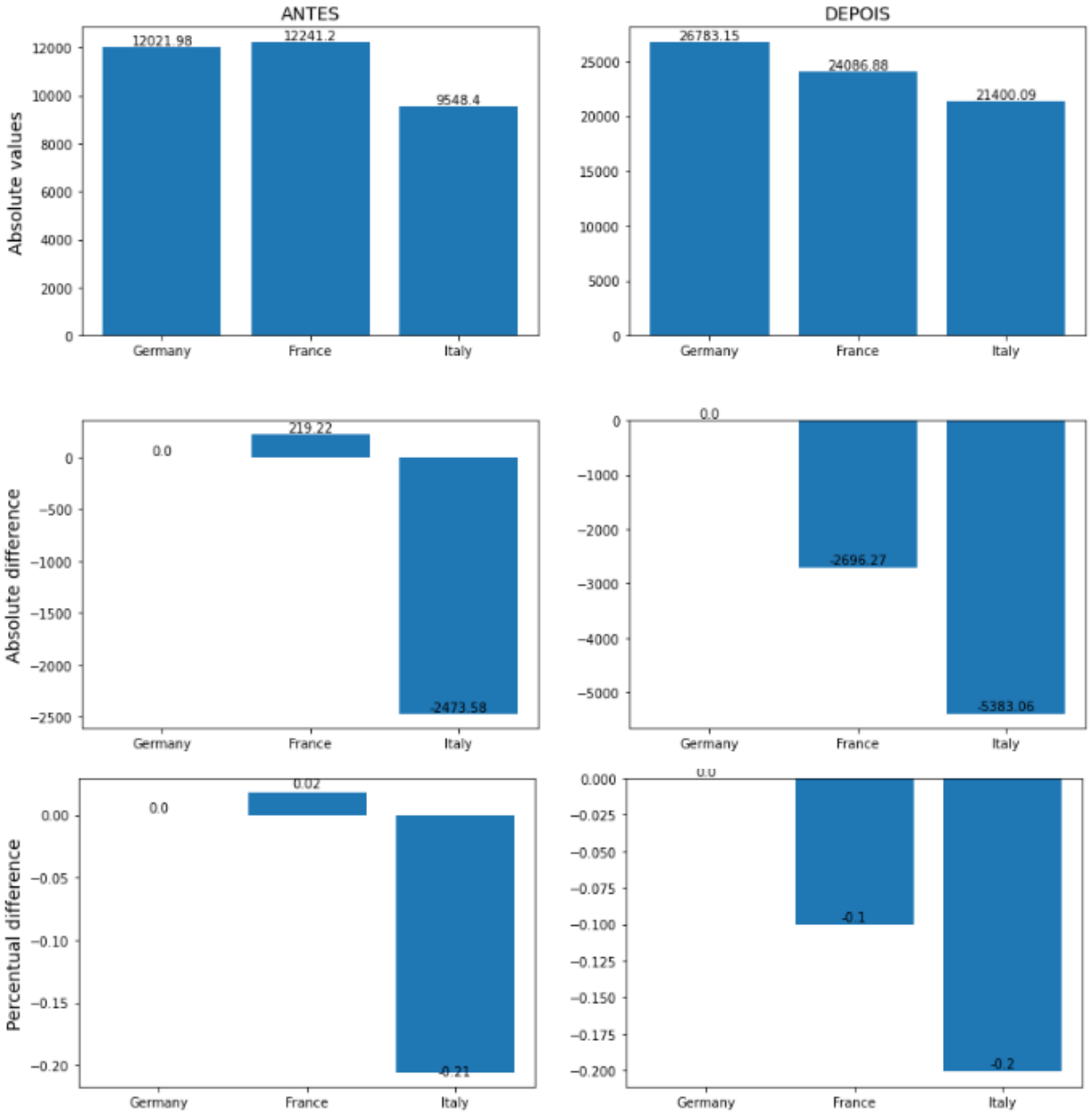


Imagem 8. Comparação de valores entre países e entre dois intervalos do tempo

VII. APÊNDICE

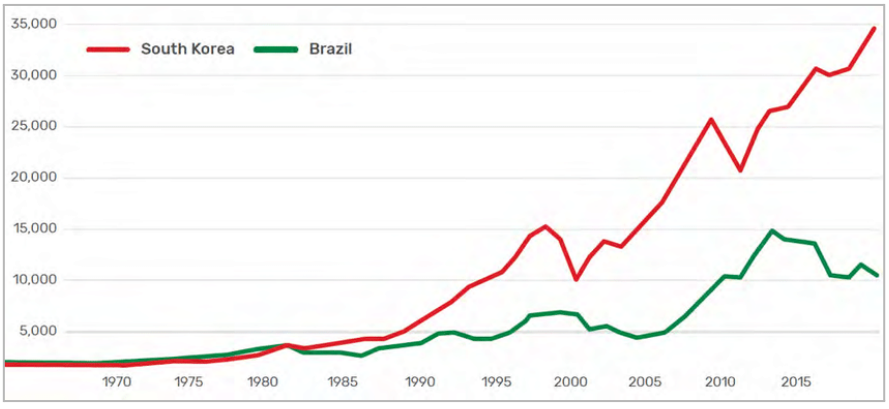


Imagem 1. Comparação do PIB per capita do Brasil e da Coreia do Sul (Fonte: Countryeconomy).

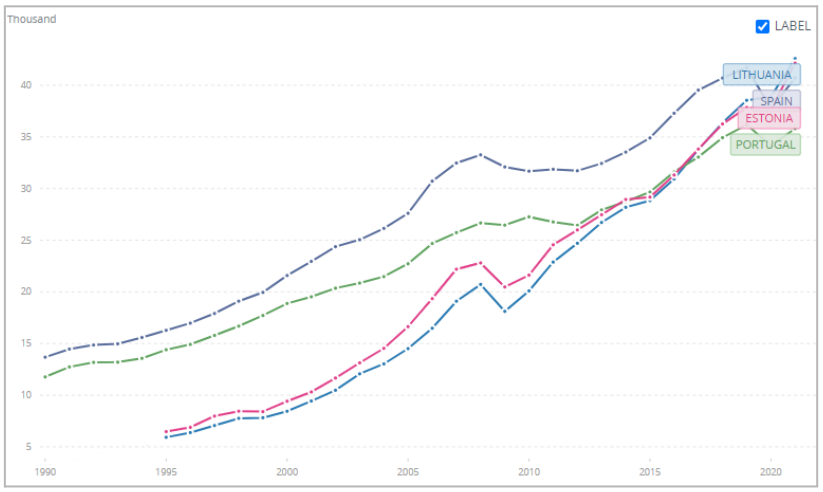


Imagem 2. Comparação do PIB per capita de Portugal, Espanha, Estônia e Lituânia (Fonte: The World Bank Data).

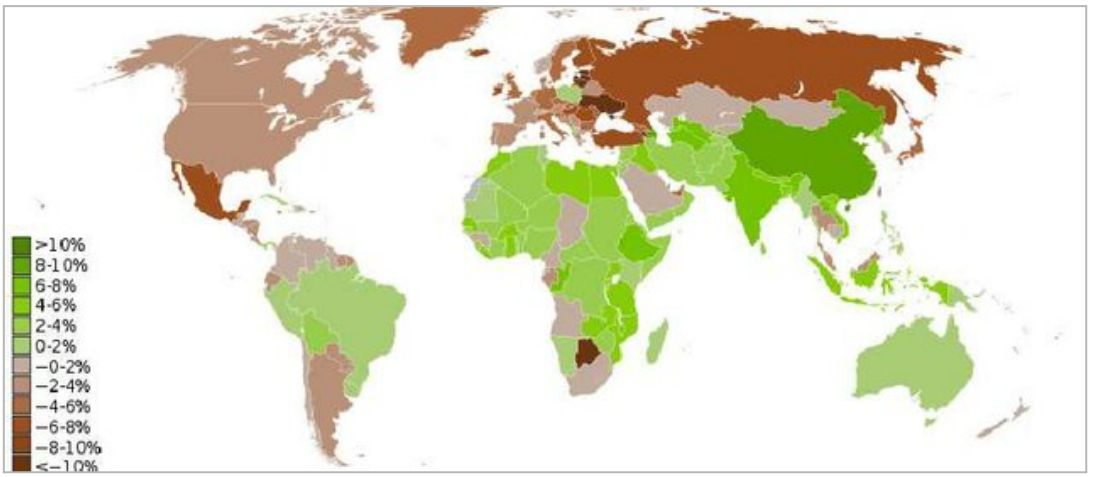


Imagem 3. Crescimento do PIB dos países em 2009 (Fonte: ResearchGate).

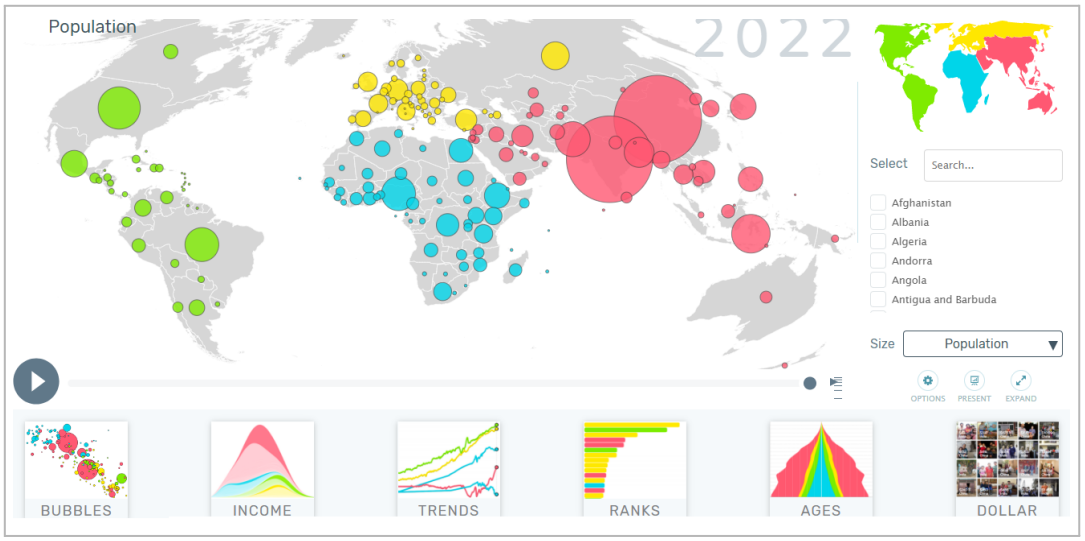
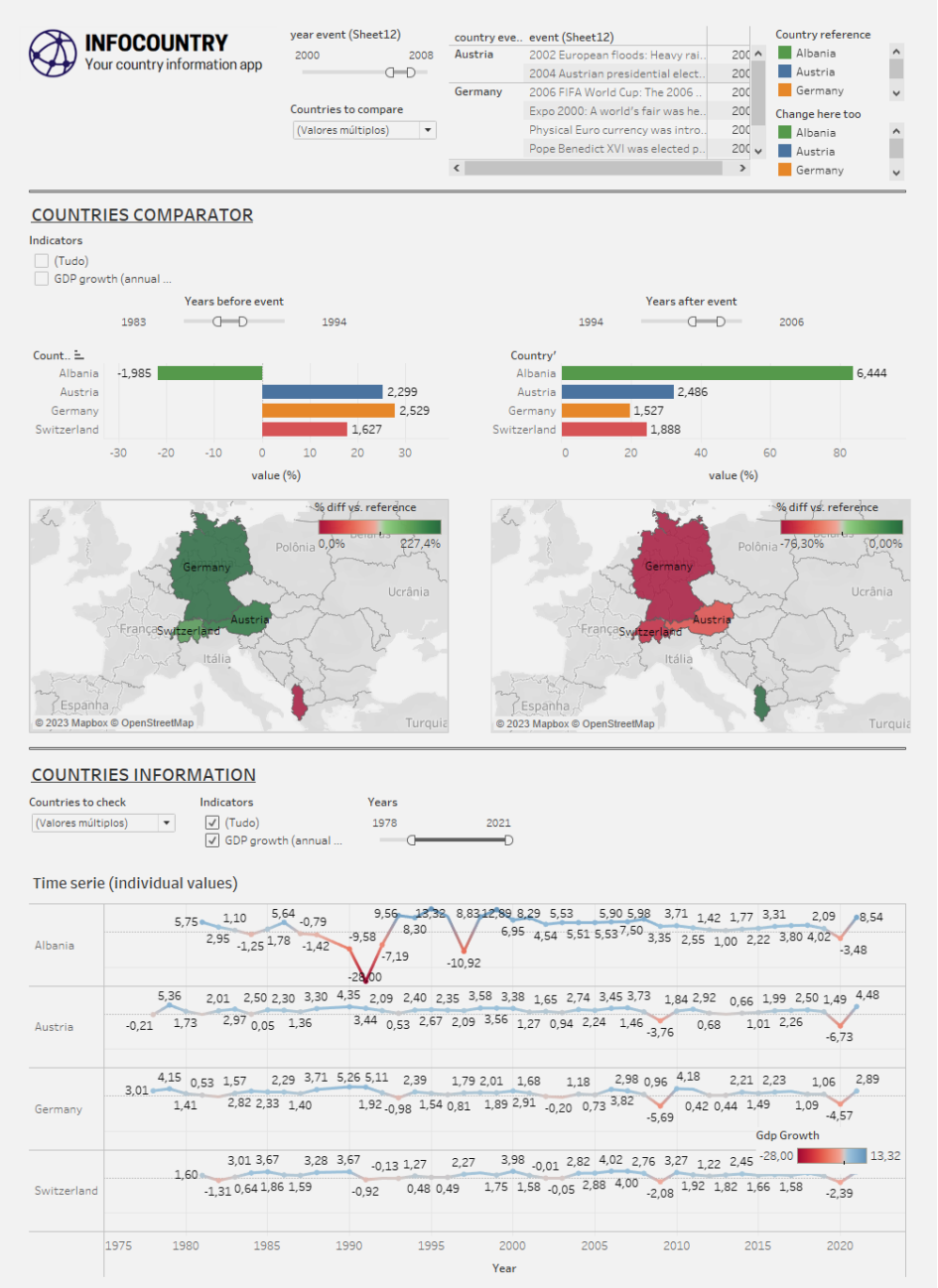
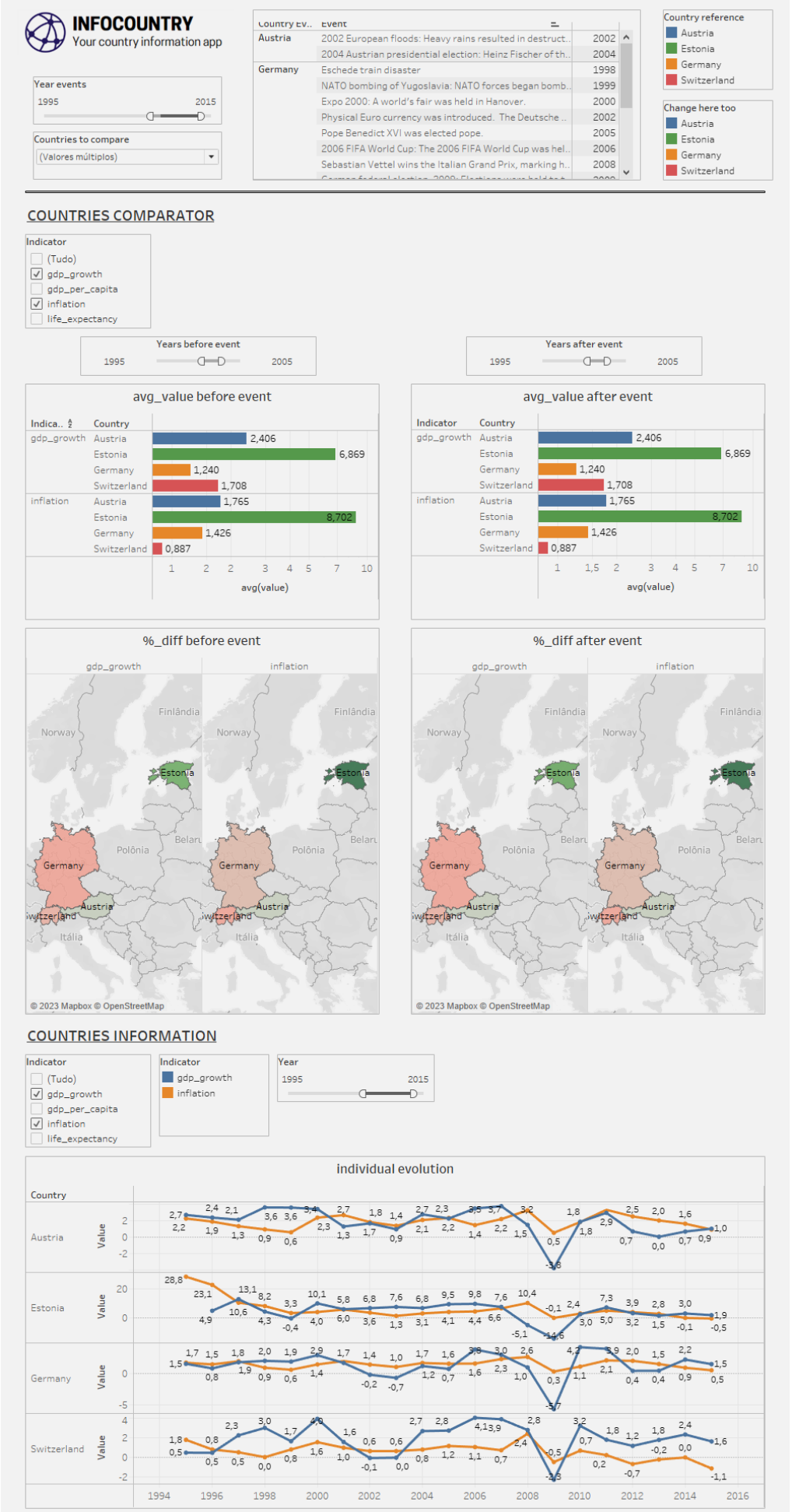


Imagem 4. Visualizações dinâmicas (Fonte: Gapminder Tools).



Mockup 1. Comparação de uma variável para quatro países (Fonte:Autores)



Mockup 2. Comparação de duas ou mais variáveis para quatro países (Parte 2 - Fonte:Autores).

DADOS	variáveis de interesse	eventos tabulados	eventos não-tabulados	fora da Wikipédia
fonte	site do Banco Mundial	páginas da Wikipédia, cujos dados estavam tabulados	páginas da Wikipédia, cujos dados não estavam tabulados	páginas suporte: <em pesquisa>
tipo	quantitativas	temporal (ano) + qualitativa (evento)	temporal (ano) + qualitativas (evento)	
estrutura	tabular (excel)	tabular (HTML)	escrito listado → tabular (excel)	
países	todos os 39 países de interesse	Portugal, Espanha, França, Suíça, Alemanha, Áustria, Eslovênia, Itália, Bélgica, Reino Unido, Irlanda, Islândia, Suécia, Finlândia, Rússia, Estônia, Letônia, Polônia, Romênia, Sérvia, Bulgária, Croácia, Albânia, Chipre e Malta (25)	Grécia e Luxemburgo (2)	Dinamarca, Noruega, Lituânia, Belarus, Ucrânia, Moldávia, Eslováquia, Chéquia, Montenegro, Bósnia, Macedônia do Norte e Países Baixos (12)
recursos/ extração	download para o PC e upload para Jupyter Notebook	técnicas de web scraping das URL das páginas; no código Python	extração manual para um Excel e upload para Jupyter Notebook	extração manual para um Excel e upload para Jupyter Notebook
código	Python			
visual	Pyplot/Dash			

Tabela 1. Características dos dados a serem usados.