Definição do projeto

Tiago da Silva Henrique

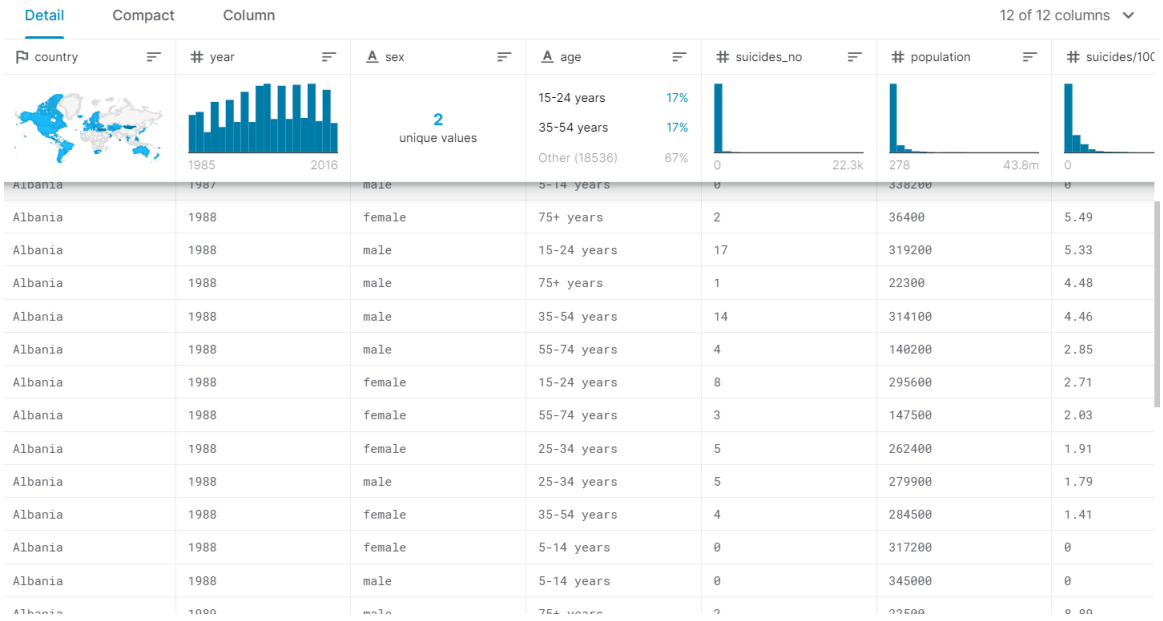
Junho de 2020

1 Base de dados e perguntas

Esse projeto avaliará a taxa de suicídio de alguns países entre 1987 e 2010. Para isso, as seguintes perguntas servirão de guias:

- Há alguma relação entre a taxa de suicídio e o PIB per capta de um país?
- Qual é a distribuição sexual dessa propriedade?
- Países com maior IDH tendem a ter uma menor taxa de suicídio? Ou não há relação evidente?
- Os suicídios aumentaram ou diminuíram nos últimos anos?
- Qual é a faixa etária com maior pretensão a tirar a própria vida?

A base de dados que embasará todas as visualizações está disponível em Suicide Rates Overview (Kaggle); ela comtém as colunas country, year, sex, age group, count of suicides, population, suicide rate, country-year composite key, HDI for year, gdpforyear, gdppercapita e generation; parte deles é mostrada na Figura 1.



Albania	1987	male	5-14 years	0	338200	0
Albania	1988	female	75+ years	2	36400	5.49
Albania	1988	male	15-24 years	17	319200	5.33
Albania	1988	male	75+ years	1	22300	4.48
Albania	1988	male	35-54 years	14	314100	4.46
Albania	1988	male	55-74 years	4	140200	2.85
Albania	1988	female	15-24 years	8	295600	2.71
Albania	1988	female	55-74 years	3	147500	2.03
Albania	1988	female	25-34 years	5	262400	1.91
Albania	1988	male	25-34 years	5	279900	1.79
Albania	1988	female	35-54 years	4	284500	1.41
Albania	1988	female	5-14 years	0	317200	0
Albania	1988	male	5-14 years	0	345000	0
Albania	1989	male	75+ years	2	22500	8.89

Figura 1: Base de dados de suicídio

2 Visualizações de referências

Inicialmente, a Figura 2[1] pode ser adaptada para representação[2] de suicídios; por exemplo, trocamos os valores do eixo horizontal pela proporção de pessoas que tiraram a prórpria vida. Poderíamos ver, assim, se,

[1] Ela foi retirdada desse trabalho de mestrado.

[2] Talvez o resultado não traga muitas informações; no entanto, um gráfico desse estilo pode ser feito com `geom_point(aes(x = country, y = suicide.rate, alpha = year, color = sex))`; para a ordenação, usamos `mutate(country = reorder(country, suicide.rate, FUN = mean))`; para as cores e os títulos, recorremos às técnica usuais de temas do `ggplot2`.

um, há diferença significativa entre os gêneros; dois, se a taxa diminuiu ou reduziu ao longo dos anos; e três, quais são os países que, nesse caso, estão em pior situação. Se funcionar adequadamente, essa técnica pode responder a duas das cinco perguntas.

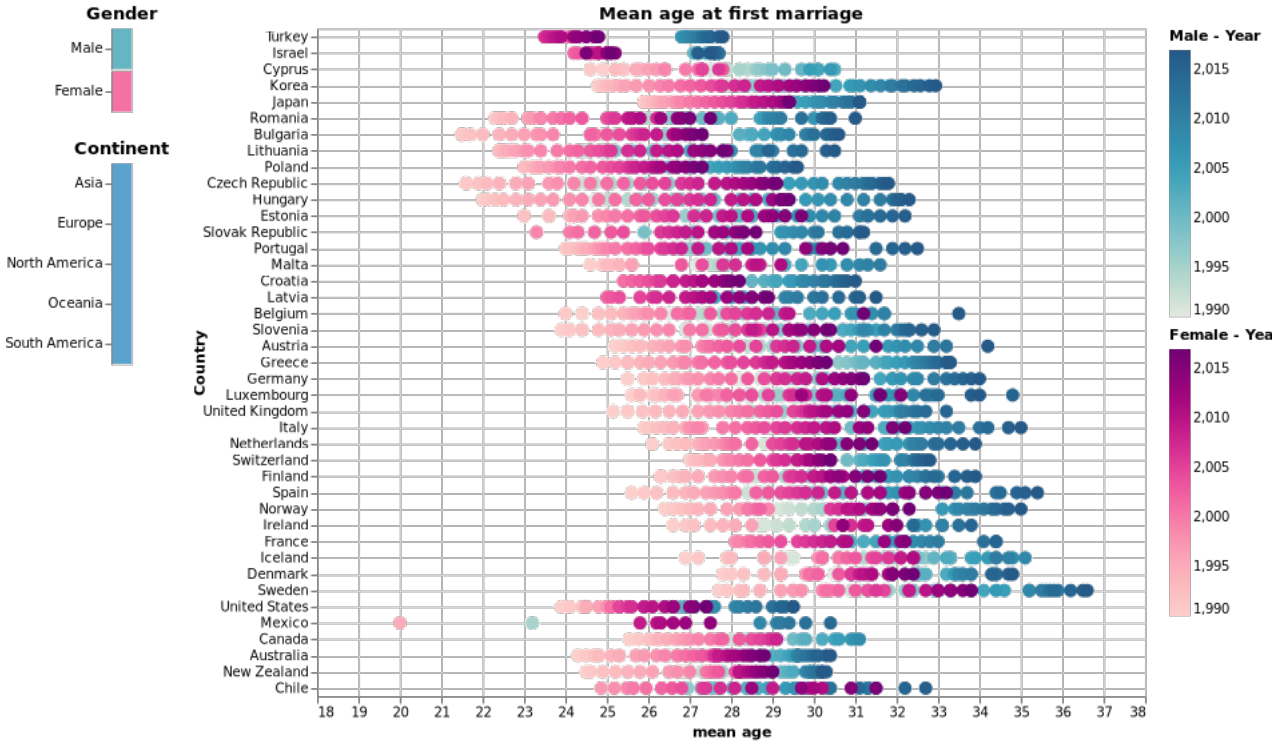


Figura 2: Quantidade de casamentos por sexo, por idade e por ano.

Para analisar o PIB, o IDH e a taxa de suicídio, podemos buscar as visualizações grandiosas de Hans Rosling, como na Figura 3. No eixo y, colocamos a variável que queremos analisar; no eixo x, o PIB (ou o IDH) - o tamanho da bolha pode representar o IDH (ou o PIB). Responderemos, assim, a outras duas perguntas.

Por fim, podemos, para medir a distribuição etária dos suicidas, podemos fazer um gráfico de colunas; no eixo x, colocamos as idades; no y, o número de suicídios. Alternativamente, podemos agrupar os dados por geração; e.g., boomer e silent.

3 Alguns vislumbres e definições

Podemos utilizar o código

```
suicide %>%
  group_by(country, sex, year) %>%
  summarise(suicide.rate = sum(suicides_no)/sum(population)) %>%
  filter(!suicide.rate == 0,
         country %in% c("Brazil", "United States", "Argentina")) %>%
  ungroup() %>%
  mutate(country = reorder(country, suicide.rate, FUN = mean)) %>%
  ggplot(aes(x = country, y = suicide.rate, color = sex, alpha = year)) +
  geom_point() +
  scale_y_log10() +
  coord_flip()
```

para produzir o código da Figura 4; logo observamos algumas tendências. Várias modificações, claro, serão feitas até período de entrega; devo selecionar os países, ordená-los, escolher melhor as cores (apesar de o padrão do `ggplot2` ser bastante sugestivo!), adicionar títulos, modificar as fontes e diversas outras escolhas temáticas.

Além disso, um simples gráfico de colunas pode indicar que, quanto mais próxima da morte, maior a probabilidade de uma pessoa matar a si mesma. No entanto, restringimos nossos dados ao ano de 2010; podemos fazer alguma espécie de rastro, como na Figura 2, para representar todos os anos; de modo alternativo, podemos selecionar algumas épocas e facetar os gráficos; um gráfico de linhas colorido pela faixa etária também pode servir, como na Figura 5.

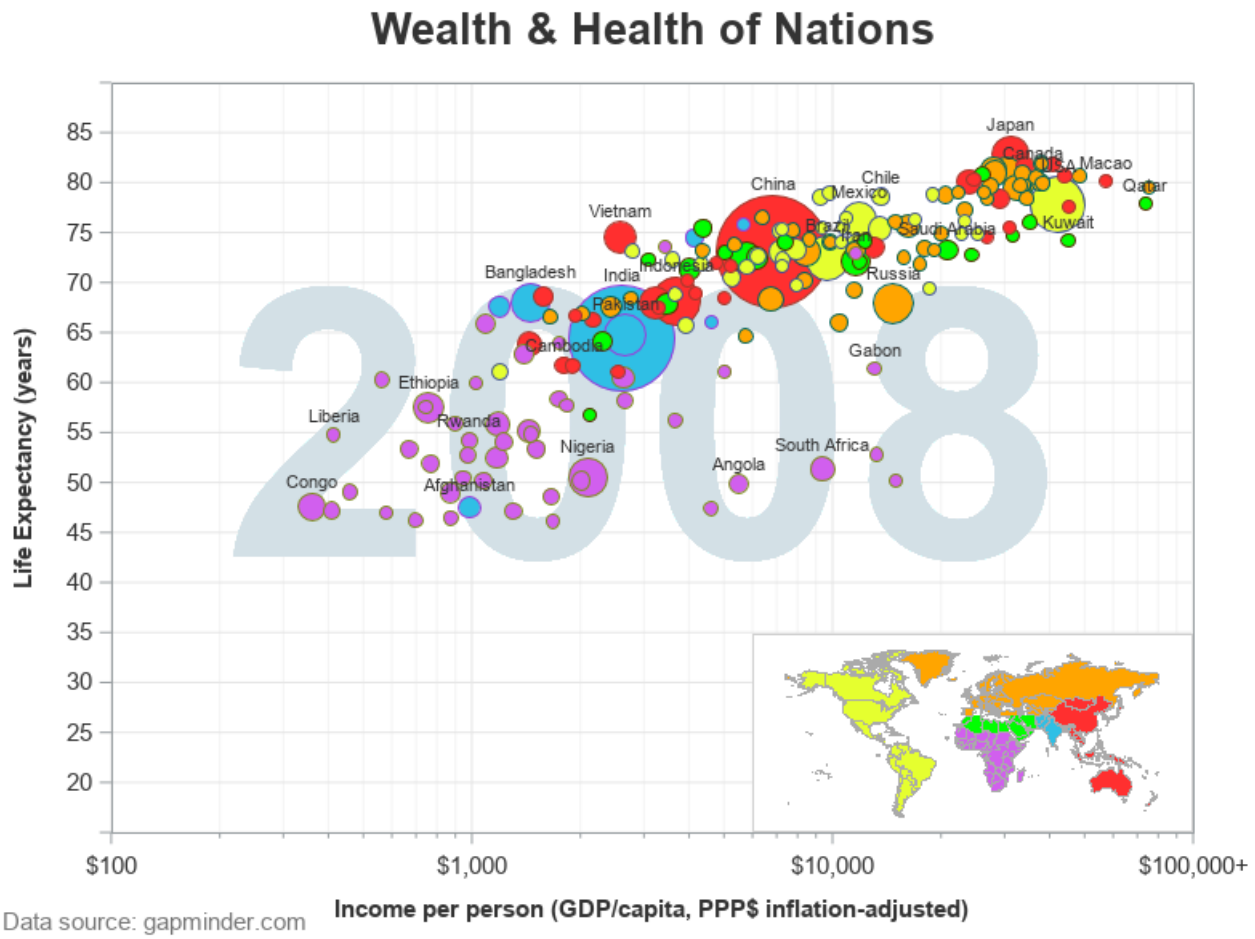


Figura 3: Expectativa de vida em relação ao PIB.

Por fim, alguns comentários, seguindo as orientações do livro de Andy Kirk. Um, esse trabalho será feito, na medida do possível, inteiramente por uma pessoa - eu - e sua audiência pode ser qualquer pessoa interessada em uma análise exploratória dos dados de suicídios. Dois, ele deve ser concluído em, no máximo, 14 dias, para que possa ser entregue antes da semana de provas. Três, as visualizações serão puramente estáticas - ferramentas de gráficos dinâmicos serão dispensadas. Quatro, todas as figuras setão feitas com as ferramentas disponíveis na linguagem R, mais especialmente, as disponíveis em `tidyverse` e as derivadas de `ggplot2`. A despeito disso, devo usar o Excel para fazer uma rápida limpeza na base de dados; e.g., mudar o formato do PIB per capta, que está bastante inutilizável. Se for necessário, posso utilizar o MySQL para criar a variável continente; isso, no entanto, exigiria um trabalho razoável e deve ser bem ponderado, porquanto eu teria de descobrir, para cada país, sua região; e há mais de uma centena deles.

As referências serão devidamente adicionadas na versão definitiva do projeto.

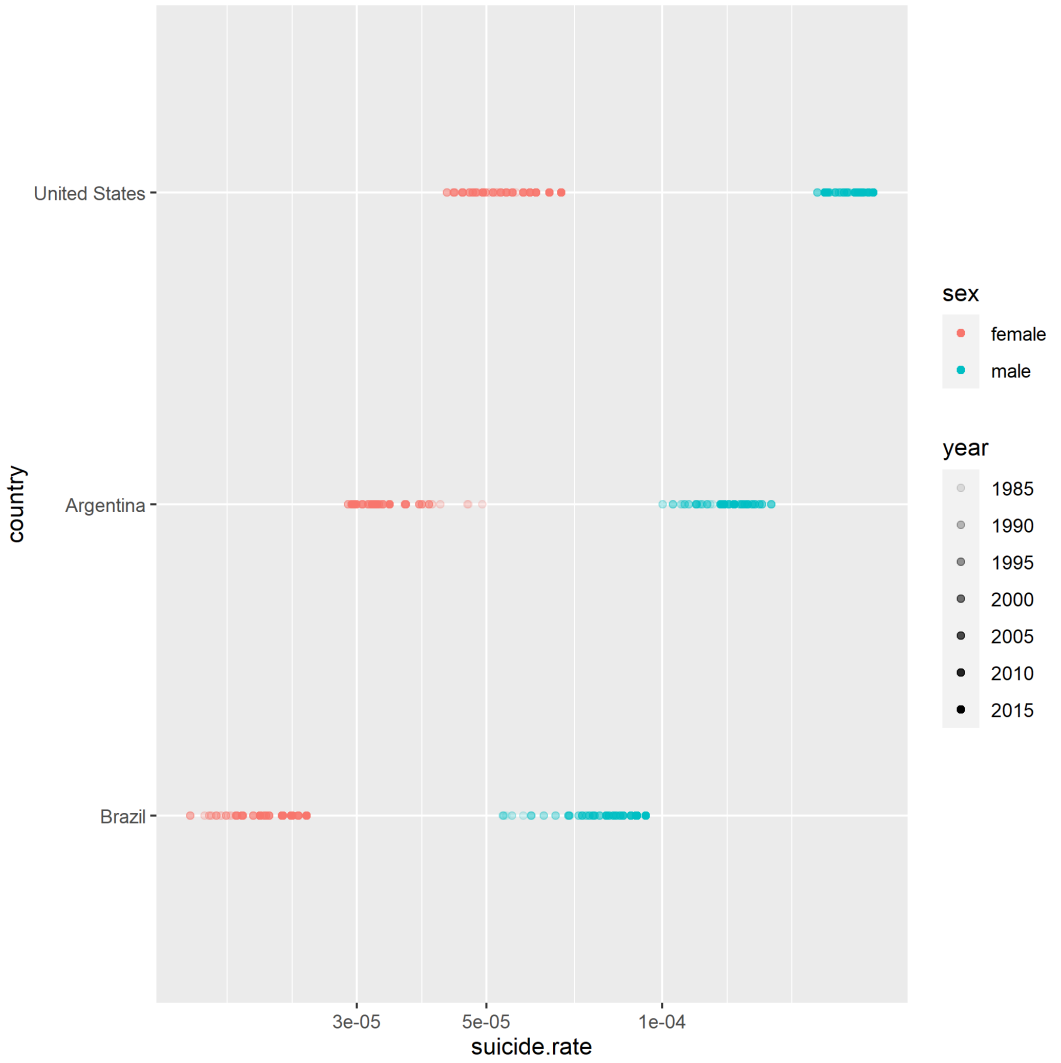


Figura 4: Um esboço para a avaliação temporal por sexo e por localidade da taxa de suicídios. Talvez seja interessante utilizar alguma crominância para distinguir os continentes; quiçá há alguma tendência entre eles.

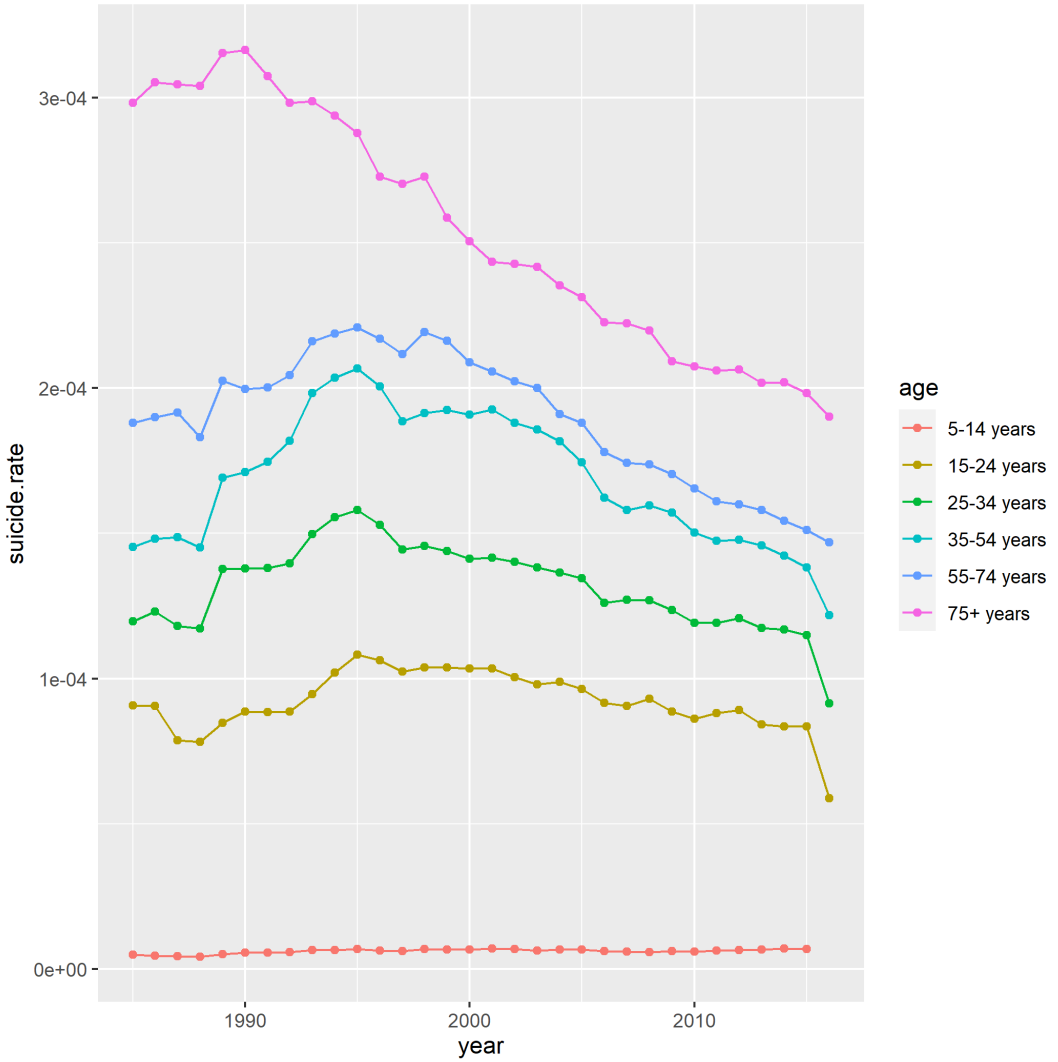


Figura 5: Este é apenas um esboço; a escala está ruim, apesar de podermos observar que houve uma redução da taxa de suicídios; as cores são arbitrárias; o fundo está cinza; enfim, muitos detalhes a serem modificados. Em adição, a queda brusca no ano de 2016 inidica uma possível incompletude nos dados; por exemplo, uma coleta parcial.