Análise da Poluição do Ar com Dados do GBD

João Silva, Marco Pereira, Mariana Ribeiro

2025-10-14

Contents

1 Resumo

Este trabalho analisa os impactos da poluição do ar na saúde pública utilizando dados do Global Burden of Disease (GBD)(Bennitt et al., n.d.). Através de técnicas de preparação e análise exploratória de dados (EDA), são identificadas tendências, padrões e potenciais problemas de qualidade nos dados.

2 Introdução

A poluição do ar é um dos principais fatores de risco para doenças respiratórias e cardiovasculares. Este estudo visa explorar os dados do GBD para compreender melhor os efeitos da poluição do ar em diferentes países, faixas etárias e géneros.

3 Metodologia

3.1 Fonte de Dados

- Dados extraídos do GBD 2021

- Ferramenta utilizada: GBD Results Tool

- Link da pesquisa e ficheiro: 1990 + Global + SDI + Health System https://vizhub.healthdata.org/gbd-results?params=gbd-api-2021-permalink/4ea5715918446e5a6d9b154d62e0cc4a IHME-GBD_2021_DATA-4835a3dc-1.csv 2000 + Global + SDI + Health System https://vizhub.healthdata.org/gbd-results?params=gbd-api-2021-permalink/64781d061f111ef4af5ad17974b6eb98 IHME-GBD_2021_DATA-c56a3848-1.csv 2010 + Global + SDI + Health System https://vizhub.healthdata.org/gbd-results?params=gbd-api-2021-permalink/88ae5e347d197231fa598e3dfc8e219a IHME-GBD_2021_DATA-1923af35-1.csv 2020 + Global + SDI + Health System https://vizhub.healthdata.org/gbd-results?params=gbd-api-2021-permalink/2d070aac165f1d6a455d41df6c34e501 IHME-GBD_2021_DATA-d14075a8-1.csv 2021 + Global + SDI + Health System https://vizhub.healthdata.org/gbd-results?params=gbd-api-2021-permalink/3fad5e919cb9822381a5b56543978c2a IHME-GBD_2021_DATA-840155c6-1.csv 1990 + Países todos + Age (all + standardized) + Sex (Both) https://vizhub.healthdata.org/gbd-results?params=gbd-api-2021-permalink/8dcef8a43426d16927c72f4d2a96147c IHME-GBD_2021_DATA-7baf5a43-1.csv 2000 + Países todos + Age (all + standardized) + Sex (Both) https://vizhub.healthdata.org/gbd-results?params=gbd-api-2021-permalink/137d422bc7e27adb3442cd2a0c699368 IHME-GBD_2021_DATA-db69c1e8-1.csv 2010 + Países todos + Age (all + standardized) + Sex (Both) https://vizhub.healthdata.org/gbd-results?params=gbd-api-2021-permalink/be61825c6c4ec111b07e9d9847613cd0 IHME-GBD_2021_DATA-dcf93c30-1.csv 2020 + Países todos + Age (all + standardized) + Sex (Both) https://vizhub.healthdata.org/gbd-results?params=gbd-api-2021-permalink/972411dc3a931543fdecc15424db9019 IHME-GBD_2021_DATA-03b79351-1.csv 2021 + Países todos + Age (all + standardized) + Sex (Both) https://vizhub.healthdata.org/gbd-results?params=gbd-api-2021-permalink/337e92e538d09c5d3d80640ee324032e IHME-GBD_2021_DATA-ce582a59-1.csv Todos os anos + Países GBD exceto costum + Age standardized + Sex (Both) https://vizhub.healthdata.org/gbd-results?params=gbd-api-2021-permalink/0afce2e9094f891f22453111c6986ffb IHME-GBD_2021_DATA-382c4db5-1.csv

- Em termos de temas (features) temos: *Fatores de risco (poluição e subdivisões e tabagismo)* Causa (doenças respiratórias crónicas e suas subdivisões) *Sexo* Idade *Locais (regiões who, continentes (?), SDI e sistemas de saúde)* Anos (1990, 2000, 2010,2020 e 2021)

3.2 Preparação dos Dados

- Seleção de colunas relevantes (e.g., país, ano, sexo, idade, medida, causa)
- Tratamento de dados faltantes e outliers
- Verificação de formatos e consistência

```
#install.packages("tidyverse")
library(tidyverse) #Já inclui dplyr + ggplot2 + forcat etc...
```

```
## Warning: package 'ggplot2' was built under R version 4.5.1

## Warning: package 'tidyr' was built under R version 4.5.1

## Warning: package 'purrr' was built under R version 4.5.1

## Warning: package 'forcats' was built under R version 4.5.1
```

```r
#Importar os dados
gbd_data <- read.csv("IHME-GBD_2021_DATA-4835a3dc-1.csv")
gbd_data2 <- read.csv("IHME-GBD_2021_DATA-c56a3848-1.csv")
gbd_data3 <- read.csv("IHME-GBD_2021_DATA-1923af35-1.csv")
gbd_data4 <- read.csv("IHME-GBD_2021_DATA-d14075a8-1.csv")
gbd_data5 <- read.csv("IHME-GBD_2021_DATA-840155c6-1.csv")
gbd_data6 <- read.csv("IHME-GBD_2021_DATA-7baf5a43-1.csv")
gbd_data7 <- read.csv("IHME-GBD_2021_DATA-db69c1e8-1.csv")
gbd_data8 <- read.csv("IHME-GBD_2021_DATA-dcf93c30-1.csv")
gbd_data9 <- read.csv("IHME-GBD_2021_DATA-03b79351-1.csv")
gbd_data10 <- read.csv("IHME-GBD_2021_DATA-ce582a59-1.csv")
gbd_data11 <- read.csv("IHME-GBD_2021_DATA-382c4db5-1.csv")
#Juntar as bases de dados
gbd_total <- bind_rows(gbd_data, gbd_data2,gbd_data3,gbd_data4,gbd_data5,gbd_data6,
        gbd_data7,gbd_data8,gbd_data9,gbd_data10,gbd_data11)
#Mudar para um formato mais tidy
gbd_total_wide <- gbd_total %>%
        pivot_wider(
                names_from = c(measure, metric),
                values_from = c(val, upper, lower),
                names_sep = "_",
                names_vary = "slowest"
  )
```

```r
#Sumário dos dados
glimpse(gbd_total)
```

```
## Rows: 438,015
## Columns: 11
## $ measure  <chr> "Deaths", "Deaths", "Deaths", "Deaths", "Deaths", "Deaths", "~
## $ location <chr> "Global", "Global", "Global", "Global", "Global", "Global", "~
## $ sex      <chr> "Male", "Female", "Both", "Male", "Female", "Both", "Male", "~
## $ age      <chr> "All ages", "All ages", "All ages", "All ages", "All ages", "~
## $ cause    <chr> "All causes", "All causes", "All causes", "All causes", "All ~
## $ rei      <chr> "Air pollution", "Air pollution", "Air pollution", "Air pollu~
## $ metric   <chr> "Number", "Number", "Number", "Rate", "Rate", "Rate", "Number~
## $ year     <int> 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1990, 1~
## $ val      <dbl> 3.920050e+06, 3.420586e+06, 7.340636e+06, 1.459577e+02, 1.291~
## $ upper    <dbl> 4.542068e+06, 3.998895e+06, 8.540006e+06, 1.691177e+02, 1.510~
## $ lower    <dbl> 3.292876e+06, 2.809383e+06, 6.120949e+06, 1.226058e+02, 1.060~
```

```r
# Converter colunas para factor
gbd_total <- gbd_total %>%
        mutate(across(c(measure, location, sex, age, cause, rei, metric), as.factor))
```

```r
names(gbd_total)
```

```
## [1] "measure"  "location" "sex"      "age"      "cause"    "rei"
## [7] "metric"   "year"     "val"      "upper"    "lower"
```

```r
# Devolve um vetor com todos os valores únicos (os nomes das categorias)
measure_classes <- unique(gbd_total$measure)
location_classes <- unique(gbd_total$location)
sex_classes <- unique(gbd_total$sex)
age_classes <- unique(gbd_total$age)
cause_classes <- unique(gbd_total$cause)
rei_classes <- unique(gbd_total$rei)
metric_classes <- unique(gbd_total$metric)
year_classes <- unique(gbd_total$year)
# Criar uma lista com todas as variáveis
classes_list <- list(
        measure = measure_classes,
        location = location_classes,
        sex = sex_classes,
        age = age_classes,
        cause = cause_classes,
        rei = rei_classes,
        metric = metric_classes,
        year = year_classes
)
# Ver a estrutura da lista de modo a ficar mais compacto.
str(classes_list)
```

```
## List of 8
##  $ measure : Factor w/ 4 levels "DALYs (Disability-Adjusted Life Years)",..: 2 1 3 4
##  $ location: Factor w/ 237 levels "Advanced Health System",..: 83 96 95 137 123 124 1 20 121 138 ...
##  $ sex     : Factor w/ 3 levels "Both","Female",..: 3 2 1
##  $ age     : Factor w/ 9 levels "0-14 years","15-49 years",..: 9 2 8 5 6 1 3 7 4
##  $ cause   : Factor w/ 4 levels "All causes","Asthma",..: 1 4 3 2
##  $ rei     : Factor w/ 7 levels "Air pollution",..: 1 5 3 4 7 2 6
##  $ metric  : Factor w/ 2 levels "Number","Rate": 1 2
##  $ year    : int [1:5] 1990 2000 2010 2020 2021
```

3.3 Ferramentas

- Linguagem: R
- Ambiente: RStudio
- Pacotes: tidyverse

4 Análise Exploratória dos Dados (EDA)

4.1 1. Impacto Geral da Poluição do Ar

4.2 2. Doenças Respiratórias Específicas

4.3 3. Tendências Temporais

4.4 4. Disparidades Regionais e Sociais

4.4.1 Quais regiões têm maior carga de doenças respiratórias atribuíveis à poluição do ar?

```r
# Filtrar linhas
gbd_filtered <- gbd_total %>% filter(
        measure == "Deaths" &
        sex == "Both" &
        age == "Age-standardized" &
        cause == "Chronic respiratory diseases" &
        rei == "Air pollution" &
        metric == "Rate" &
        year == "2021")
# Ordenar as regiões por valor
gbd_filtered<-mutate(gbd_filtered,location = fct_reorder(location, val))
# Fazer Barplot
ggplot(data = gbd_filtered, aes(x = location, y = val)) +
        geom_bar(stat = "identity", width = 0.6, fill = "steelblue") +
        geom_errorbar( aes(x=location, ymin=lower, ymax=upper),
                width=0.4, colour="orange", alpha=0.9, size=0.5) +
        coord_flip() +
        labs(title = "Taxa de Mortes por Doenças respiratórias por Local",
                x = "Local", y = "Taxa") +
        theme_minimal()
```

```
## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```

Taxa de Mortes por Doenças respiratórias por Loc...
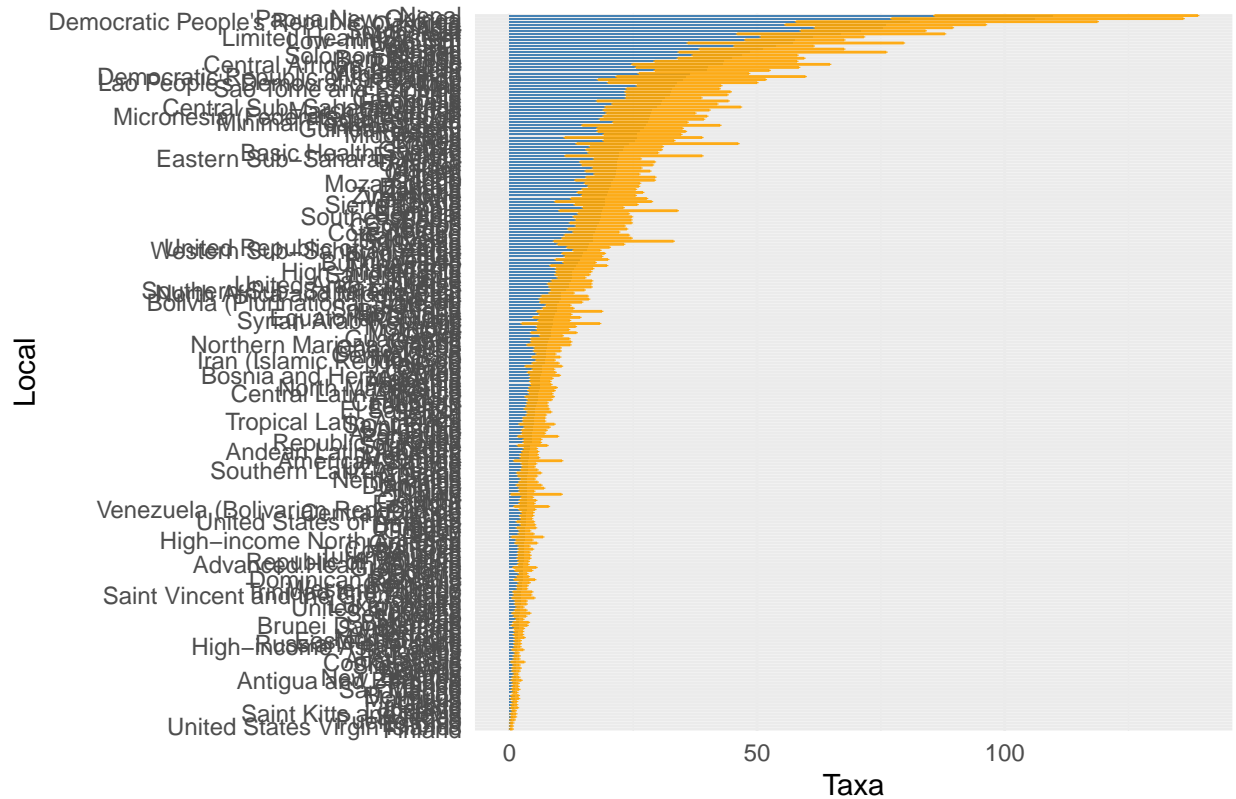
```r
# Filtrar linhas
gbd_deaths <- gbd_total %>% filter(
    measure == "Deaths" &
    location %in% gbd_data11$location &
    sex == "Both" &
    age == "Age-standardized" &
    cause == "Chronic respiratory diseases" &
    rei == "Air pollution" &
    metric == "Rate" &
    year == "2021")
gbd_dalys <- gbd_total %>% filter(
    measure == "DALYs (Disability-Adjusted Life Years)" &
    location %in% gbd_data11$location &
    sex == "Both" &
    age == "Age-standardized" &
    cause == "Chronic respiratory diseases" &
    rei == "Air pollution" &
    metric == "Rate" &
    year == "2021")
gbd_ylds <- gbd_total %>% filter(
    measure == "YLDs (Years Lived with Disability)" &
    location %in% gbd_data11$location &
    sex == "Both" &
    age == "Age-standardized" &
    cause == "Chronic respiratory diseases" &
    rei == "Air pollution" &
```

```
        metric == "Rate" &
        year == "2021")
gbd_ylls <- gbd_total %>% filter(
        measure == "YLLs (Years of Life Lost)" &
        location %in% gbd_data11$location &
        sex == "Both" &
        age == "Age-standardized" &
        cause == "Chronic respiratory diseases" &
        rei == "Air pollution" &
        metric == "Rate" &
        year == "2021")
# Ordenar as regiões por valor
gbd_deaths<-gbd_deaths %>% arrange(desc(val))
gbd_dalys<-gbd_dalys %>% arrange(desc(val))
gbd_ylds<-gbd_ylds %>% arrange(desc(val))
gbd_ylls<-gbd_ylls %>% arrange(desc(val))
# Fazer tabela
cbind(
  Deaths = as.character(gbd_deaths$location),
  DALYs = as.character(gbd_dalys$location),
  YLDs = as.character(gbd_ylds$location),
  YLLs = as.character(gbd_ylls$location)
)
```

```
##       Deaths                        DALYs
##  [1,] "Oceania"                     "Oceania"
##  [2,] "South Asia"                  "South Asia"
##  [3,] "East Asia"                   "Central Sub-Saharan Africa"
##  [4,] "Central Sub-Saharan Africa"  "East Asia"
##  [5,] "Eastern Sub-Saharan Africa"  "Eastern Sub-Saharan Africa"
##  [6,] "Southeast Asia"              "Western Sub-Saharan Africa"
##  [7,] "Western Sub-Saharan Africa"  "Southeast Asia"
##  [8,] "Southern Sub-Saharan Africa" "Southern Sub-Saharan Africa"
##  [9,] "North Africa and Middle East" "North Africa and Middle East"
## [10,] "Central Asia"                "Central Asia"
## [11,] "Central Latin America"       "Caribbean"
## [12,] "Caribbean"                   "Central Latin America"
## [13,] "Tropical Latin America"      "Tropical Latin America"
## [14,] "Andean Latin America"        "Central Europe"
## [15,] "Southern Latin America"      "Andean Latin America"
## [16,] "Central Europe"              "Southern Latin America"
## [17,] "High-income North America"   "High-income North America"
## [18,] "Western Europe"              "Western Europe"
## [19,] "Eastern Europe"              "Eastern Europe"
## [20,] "High-income Asia Pacific"    "High-income Asia Pacific"
## [21,] "Australasia"                 "Australasia"
##       YLDs                          YLLs
##  [1,] "South Asia"                  "Oceania"
##  [2,] "Central Sub-Saharan Africa"  "South Asia"
##  [3,] "Oceania"                     "Central Sub-Saharan Africa"
##  [4,] "Eastern Sub-Saharan Africa"  "East Asia"
##  [5,] "Western Sub-Saharan Africa"  "Eastern Sub-Saharan Africa"
##  [6,] "East Asia"                   "Southeast Asia"
```

```
## [7,] "Southern Sub-Saharan Africa"   "Western Sub-Saharan Africa"
## [8,] "Southeast Asia"                 "Southern Sub-Saharan Africa"
## [9,] "North Africa and Middle East"  "North Africa and Middle East"
## [10,] "Central Asia"                  "Central Asia"
## [11,] "Andean Latin America"          "Caribbean"
## [12,] "Central Europe"                "Central Latin America"
## [13,] "Caribbean"                     "Tropical Latin America"
## [14,] "Central Latin America"         "Central Europe"
## [15,] "High-income Asia Pacific"      "Southern Latin America"
## [16,] "High-income North America"     "Andean Latin America"
## [17,] "Tropical Latin America"        "High-income North America"
## [18,] "Southern Latin America"        "Western Europe"
## [19,] "Eastern Europe"                "Eastern Europe"
## [20,] "Western Europe"                "Australasia"
## [21,] "Australasia"                   "High-income Asia Pacific"
```

4.4.2 Existe correlação entre o Índice Sociodemográfico (SDI) e a carga de doença?

```r
# Filtrar linhas
gbd_filtered <- gbd_total %>% filter(
        measure == "Deaths" &
        location %in% c("Low SDI","Low-middle SDI","Middle SDI","High-middle SDI","High SDI") &
        sex == "Both" &
        age == "Age-standardized" &
        cause == "All causes" &
        rei == "Air pollution" &
        metric == "Rate" &
        year == "2021")
# Fazer Barplot
ggplot(data = gbd_filtered, aes(x = factor(location,
                levels = c("Low SDI", "Low-middle SDI", "Middle SDI","High-middle SDI", "High SDI")),
                y = val)) +
        geom_bar(stat = "identity", width = 0.6, fill = "steelblue") +
        geom_errorbar( aes(x = factor(location,
                levels = c("Low SDI", "Low-middle SDI", "Middle SDI", "High-middle SDI", "High SDI")),
                ymin=lower, ymax=upper),
                width=0.4, colour="orange", alpha=0.9, size=0.5) +
        coord_flip() +
        labs(title = "Taxa de Mortes por SDI",
                x = "Local", y = "Taxa") +
        theme_minimal()
```

Taxa de Mortes por SDI
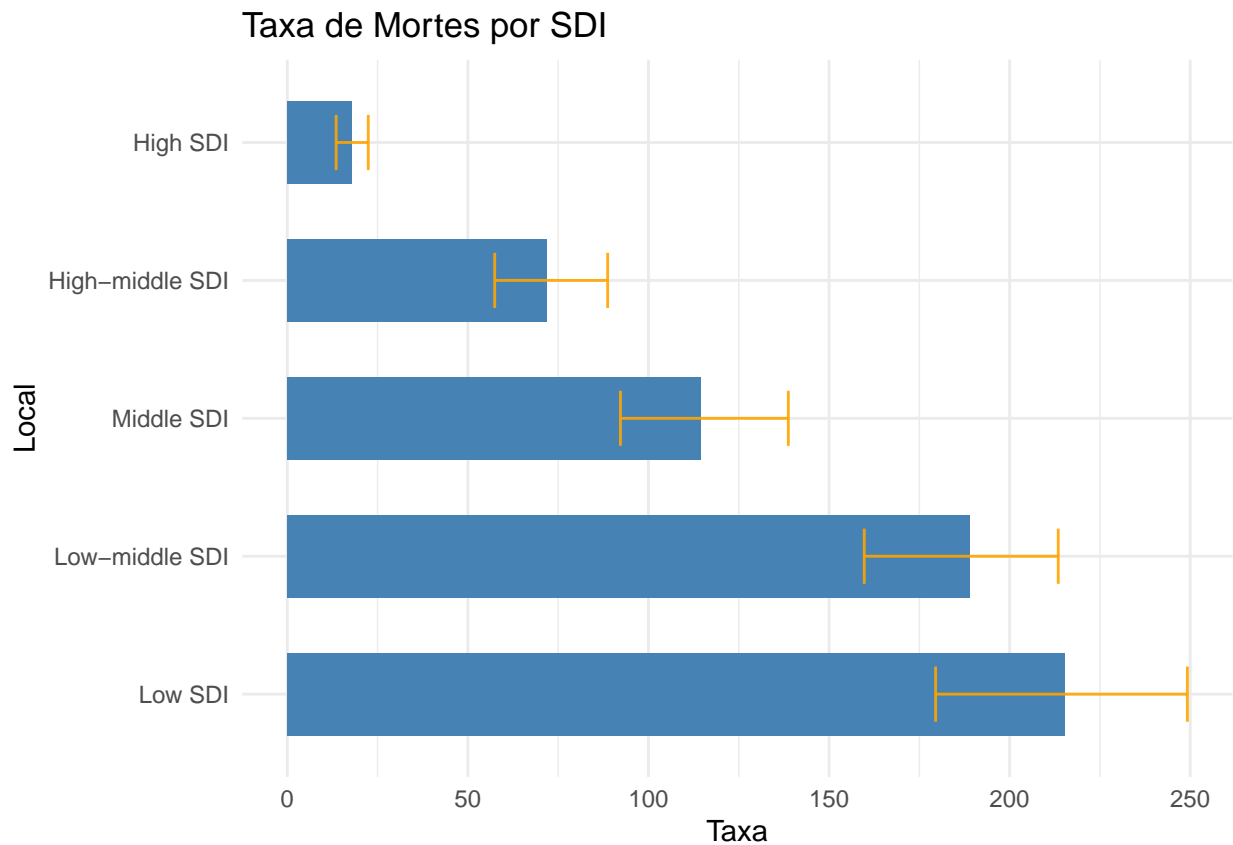

```r
# Filtrar linhas
gbd_filtered <- gbd_total %>% filter(
        measure == "Deaths" &
        location %in% c("Low SDI","Low-middle SDI","Middle SDI","High-middle SDI","High SDI") &
        sex == "Both" &
        age == "Age-standardized" &
        cause == "All causes" &
        rei == "Air pollution" &
        metric == "Rate")
# Fazer Barplot
ggplot(data = gbd_filtered, aes(x = year,
                y = val,color=factor(location,
                levels = c("Low SDI", "Low-middle SDI", "Middle SDI","High-middle SDI", "High SDI")))) +
        geom_line(stat = "identity", width = 0.6, fill = "steelblue") +
        geom_errorbar( aes(x = year,
                ymin=lower, ymax=upper),
                width=0.4, colour="orange", alpha=0.9, size=0.5) +
        labs(title = "Taxa de Mortes por ano por SDI",
                x = "Local", y = "Taxa", color="Nível de SDI") +
        theme_minimal()
```

```
## Warning in geom_line(stat = "identity", width = 0.6, fill = "steelblue"):
## Ignoring unknown parameters: 'width' and 'fill'
```

Taxa de Mortes por ano por SDI
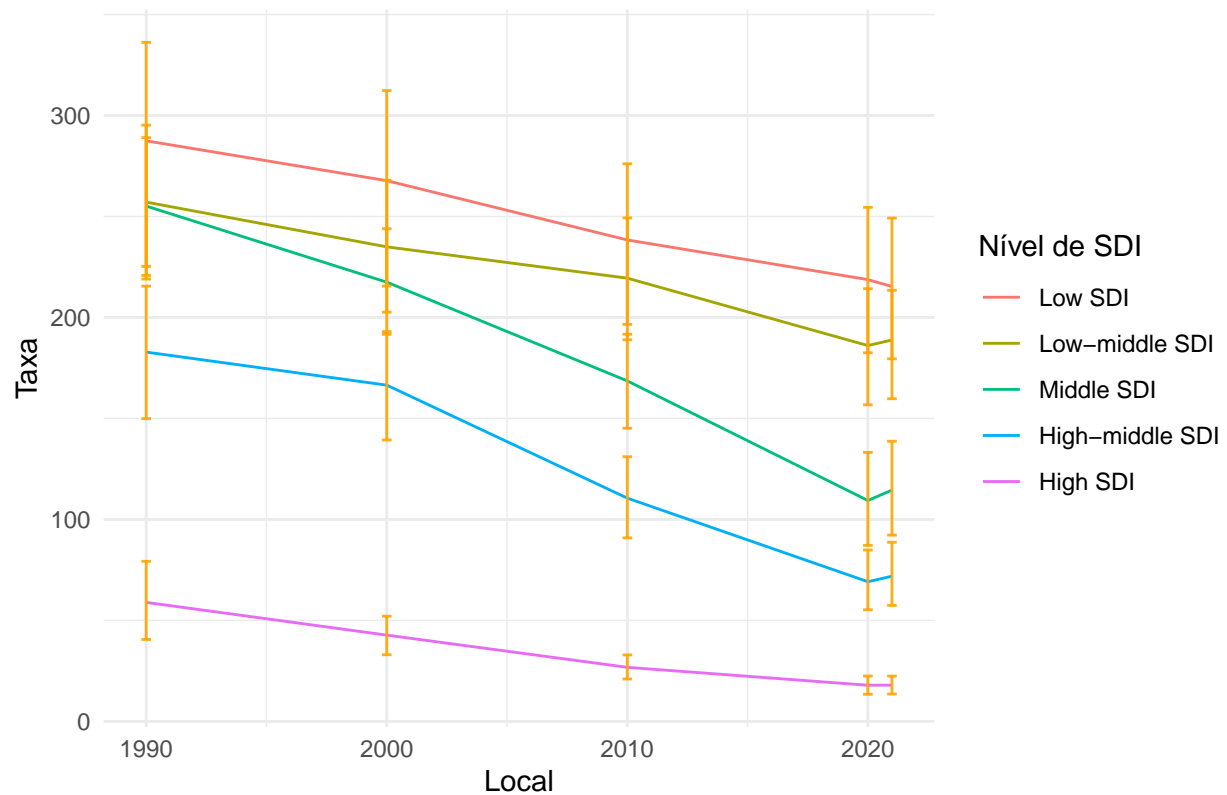
4.4.3 Como variam os impactos por sexo (boxplot/violinplot) e faixa etária (barras/histograma)?
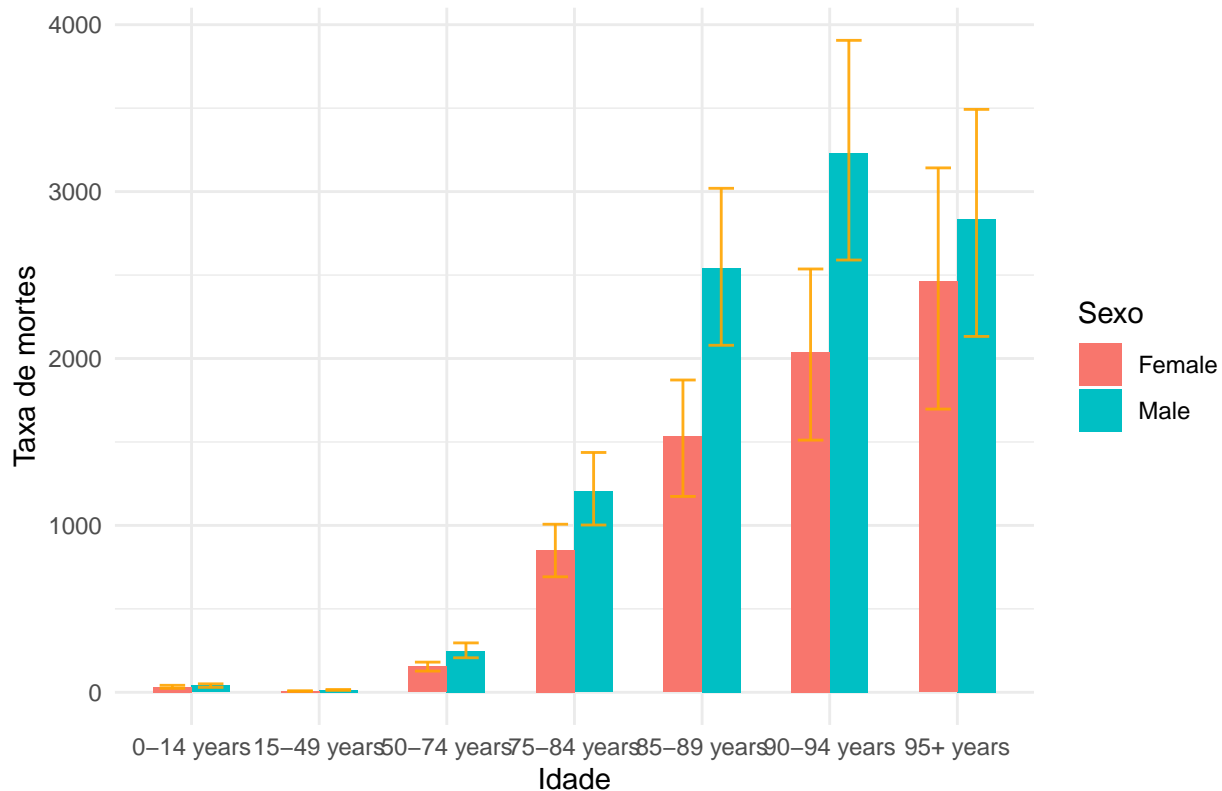
```r
# Filtrar apenas linhas com metric == "Rate" e measure == "Deaths"
gbd_filtered <- gbd_total %>% filter(
        measure == "Deaths" &
        location == "Global" &
        sex == "Both" &
        !age %in% c("All ages", "7-27 days", "Age-standardized") &
        cause == "All causes" &
        rei == "Air pollution" &
        metric == "Rate" &
        year == "2021")
# Fazer Barplot
ggplot(data = gbd_filtered, aes(x = age, y = val)) +
        geom_bar(stat = "identity", width = 0.6, fill = "steelblue") +
        geom_errorbar( aes(x=age, ymin=lower, ymax=upper),
                width=0.4, colour="orange", alpha=0.9, size=0.5) +
        labs(title = "Taxa de Mortes por SDI",
                x = "Local", y = "Taxa") +
        theme_minimal()
```

Taxa de Mortes por SDI
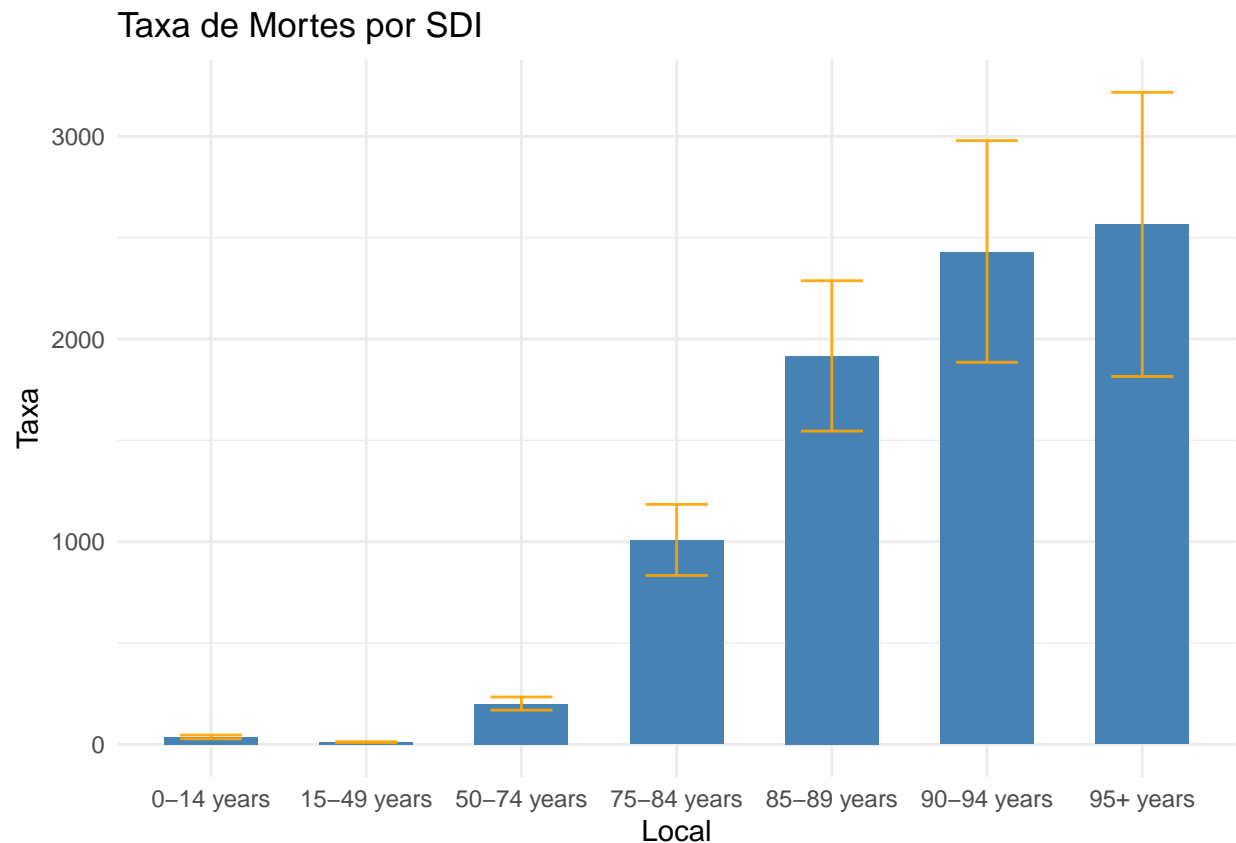

```r
# Filtrar apenas linhas com metric == "Rate" e measure == "Deaths"
gbd_filtered <- gbd_total %>% filter(
        measure == "Deaths" &
        location == "Global" &
        sex %in% c("Male", "Female") &
        !age %in% c("All ages", "7-27 days", "Age-standardized") &
        cause == "All causes" &
        rei == "Air pollution" &
        metric == "Rate" &
        year == "2021")
# Fazer Barplot
ggplot(data = gbd_filtered, aes(x = age, y = val, fill = sex)) +  # Adicionado fill = sex para agrupame
        geom_bar(stat = "identity", width = 0.6, position = position_dodge(width = 0.6)) +  # position_
        geom_errorbar(aes(ymin = lower, ymax = upper),
                position = position_dodge(width = 0.6),  # Alinhar barras de erro com as barras dodged
                width = 0.4, colour = "orange", alpha = 0.9, size = 0.5) +
        labs(title = "Taxa de Mortes por Idade e Sexo (Poluição do Ar, Global, 2021)",
                x = "Idade", y = "Taxa de mortes", fill = "Sexo") +
        theme_minimal() +
        theme(legend.position = "right")
```

Taxa de Mortes por Idade e Sexo (Poluição do Ar, Global, 2021)


```r
# Filtrar linhas
gbd_filtered <- gbd_total %>% filter(
        measure == "Deaths" &
        location == "Global" &
        sex %in% c("Male", "Female") &
        age %in% c("Age-standardized") &
        cause == "All causes" &
        !rei %in% c("Air pollution","All risk factors","Particulate matter pollution") &
        metric == "Rate") %>%
                mutate(year = factor(year, levels = c(1990, 2000, 2010, 2020, 2021)))
# Fazer Barplot
ggplot(data = gbd_filtered, aes(x = year, y = val, fill = rei)) +
        geom_bar(stat = "identity", width = 0.7, position = position_dodge(width = 0.8)) +  # position_
        geom_errorbar(aes(ymin = lower, ymax = upper),
                position = position_dodge(width = 0.8),  # Alinhar barras de erro com as barras dodged
                width = 0.4, colour = "orange", alpha = 0.9, size = 0.5) +
        facet_grid(sex ~ ., scales = "fixed") +
        labs(title = "Taxa de Mortes por Idade e Sexo (Poluição do Ar, Global, 2021)",
                x = "Idade", y = "Taxa de mortes", fill = "Fator de risco") +
        theme_minimal() +
        theme(legend.position = "right")
```

Taxa de Mortes por Idade e Sexo (Poluição do Ar, Global, 2021)
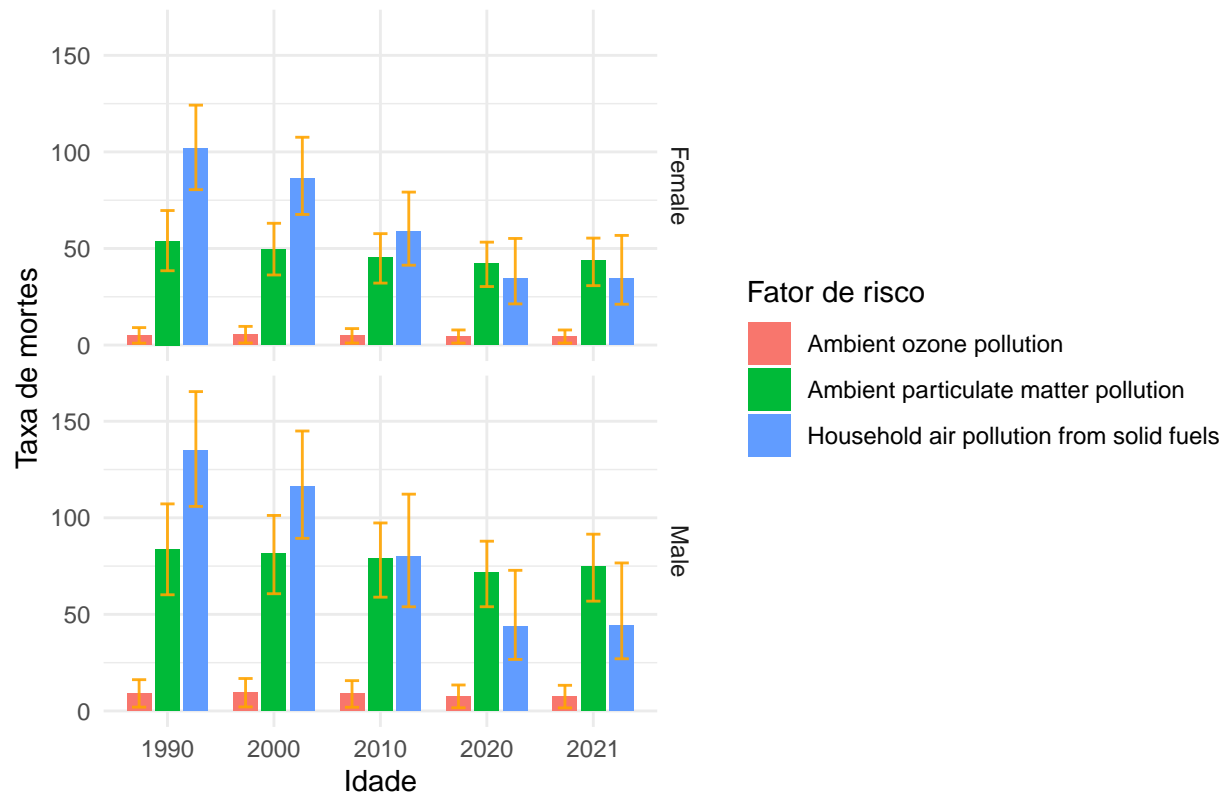
O Nitrogen dioxide pollution não aparece porque só tem a medição DALYs.

5 Discussão

Interpretação dos resultados Comparação com fontes externas (e.g., WHO, dados governamentais) *Limitações dos dados e da análise

6 Conclusão

Principais conclusões sobre o impacto da poluição do ar Sugestões para futuras investigações

7 Referências

Anexos

Bennitt, F. B., Wozniak, S., Causey, K., Spearman, S., Okereke, C., Garcia, V., Hashmeh, N., Ashbaugh, C., Abdelkader, A., Abdoun, M., Abdurebi, M. J., Abedi, A., Zuñiga, R. A. A., Aboagye, R. G., Abubakar, B., Abu-Zaid, A., Adane, M. M., Adegboye, O. A., Adekanmbi, V., . . . Burkart, K. (n.d.). *Global, regional, and national burden of household air pollution, 1990–2021: A systematic analysis for the global burden of disease study 2021.* Retrieved October 1, 2025, from https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(24)02840-X/fulltext?uuid/x3duuid%3A40008a0f-2266-4637-8209-d6041e2790e4