# Web Scraping e Análise Exploratória das Taxas de Mortalidade por COVID-19

Maria José García Montes
UFPE - Tópicos Especiais em Estatística Computacional

Septembro 2025

**Resumo**

Este trabalho apresenta uma análise exploratória de web scraping aplicada ao site *Statista*, onde foram coletados dados sobre as taxas de mortalidade por COVID-19 em diferentes países. A análise foi realizada no Google Colab utilizando o pacote `rvest`, e os dados foram examinados por meio de uma abordagem exploratória que inclui medidas descritivas e visualizações.

## 1 Introdução

A pandemia de COVID-19 teve um impacto significativo na saúde e na economia global, tendo surgido no final de 2019. Trata-se de uma doença respiratória altamente contagiosa, cujos sintomas podem variar de leves a muito graves, podendo inclusive levar à morte. Essa situação evidenciou a importância de dispor de dados confiáveis e atualizados para compreender seu impacto em escala mundial. Uma das plataformas que disponibiliza informações estatísticas é o *Statista*, que reúne dados provenientes de organismos internacionais e nacionais. Neste trabalho, foram aplicadas técnicas de web scraping para extrair informações do site Statista, especificamente da página sobre taxas de mortalidade por COVID-19 (Statista, 2025). Em seguida, foi realizada uma análise exploratória com o objetivo de observar o comportamento das variáveis, seguindo as diretrizes metodológicas sugeridas por Shilees (Ferreira, 2025).

## 2 Marco Teórico

O web scraping é uma técnica que permite extrair dados de sites da internet de forma automatizada por meio de código. Essa abordagem tornou-se fundamental para obter informações atualizadas em tempo real *rvest*.

A análise exploratória de dados (EDA, na sigla em inglês) é uma etapa inicial do processo estatístico que tem como objetivo identificar padrões, tendências e possíveis anomalias nas informações coletadas. Según (Tukey, 1977), o EDA é fundamental para compreender a natureza dos dados.

No caso da COVID-19, diversos organismos coletaram dados em tempo real, porém, a disponibilidade, qualidade e comparabilidade dessas estatísticas variam conforme a fonte. Por esse motivo, ferramentas como o Statista tornaram-se repositórios úteis para apresentar indicadores de interesse, como as taxas de mortalidade (Statista, 2025).

# 3 Metodología

O presente trabalho foi desenvolvido utilizando o **Google Colab**, com o uso da linguagem R e do pacote `rvest`. Os dados foram obtidos da plataforma *Statista*, especificamente sobre as taxas de mortalidade por COVID-19 (Statista, 2025).

Foram aplicadas técnicas de *web scraping* utilizando a linguagem R para extração de informações de interesse, seguidas por um processo de limpeza dos dados. Posteriormente, realizou-se uma análise exploratória por meio de estatísticas descritivas e visualizações gráficas, seguindo as recomendações metodológicas propostas por (Ferreira, 2025).

A Tabela 1 apresenta os nomes das variáveis incluídas na análise, juntamente com uma breve descrição de seus significados.

| **Nome da variável** | **Descrição** |
|---|---|
| Characteristic | Nome do país |
| Confirmed | Casos confirmados |
| Last7Days | Casos reportados nos últimos 7 dias |
| DailyCases | Casos diários |
| Deaths | Total de mortes |
| DailyDeaths | Mortes diárias |
| DeathRate | Taxa de mortalidade (%) |

Tabela 1: variáveis analisadas e descrições

# 4 Resultados

Este conjunto de dados é composto por 221 linhas e 7 colunas. Na Tabela 2, observa-se que a variável Confirmed apresenta os valores mais elevados, pois representa o número de casos confirmados de COVID-19, com um intervalo entre 2641 y 80 442 894, o que reflete uma magnitude considerável.

| **Variable** | **Mínimo** | **1er Cuartil** | **Mediana** | **Media** | **3er Cuartil** | **Máximo** |
|---|---|---|---|---|---|---|
| Confirmed | 2641 | 30752 | 200990 | 2303078 | 1101968 | 80442894 |
| Last7Days | 0 | 17 | 250 | 18657 | 2892 | 679183 |
| DailyCases | 0 | 0 | 13 | 2593 | 215 | 136798 |
| Deaths | 1 | 233 | 2201 | 28123 | 14059 | 986896 |
| DailyDeaths | 0.00 | 0.00 | 0.00 | 10.91 | 1.00 | 341.00 |
| DeathRate (%) | 0.02 | 0.55 | 0.99 | 1.49 | 1.95 | 19.11 |

Tabela 2: Resumo estatístico das variáveis relacionadas à COVID-19

Figura 1 representa os casos confirmados de COVID-19, na qual se observa a presença de numerosos valores atípicos, correspondentes a países com alta incidência da doença. Esses valores exercem uma influência significativa sobre a distribuição geral. Como evidenciado na Tabela 2, o valor mínimo registrado é de apenas $2,641$ casos, enquanto o máximo alcança $80,442,894$, refletindo uma enorme dispersão.

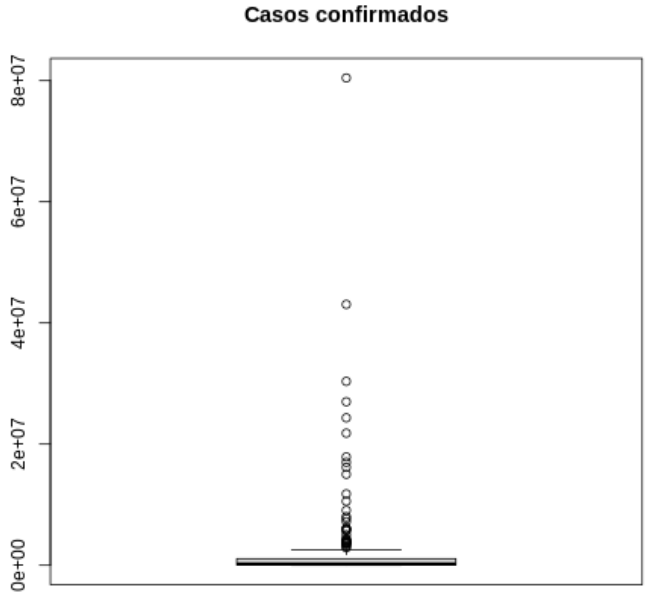


Figura 1: Casos Confirmados de Covid-19

A Figura 2 apresenta uma comparação entre os países com maior e menor número de mortes registradas por COVID-19. Observa-se que os Estados Unidos lideram a lista com 986, 896 óbitos, seguidos pelo Brasil 662, 964, Índia 522, 223, Rússia 367, 521 e México 324, 134. Esses países concentram um número expressivo de mortes, o que reflete a magnitude do impacto da pandemia em nações com elevada densidade populacional.

Em contraste, países como Nova Zelândia, Saint Pierre e Miquelon, Palau, Saint-Barthélemy e Anguila registraram números mínimos de mortes (entre 1 e 9 casos), o que evidencia uma gestão mais eficaz no controle da propagação do vírus.

A comparação evidencia a grande diferença no impacto da pandemia: enquanto em alguns países o número de mortes ultrapassou centenas de milhares, em outros foram registrados apenas alguns óbitos. Isso reforça a ideia de que as características próprias de cada nação — como o tamanho da população, a capacidade do sistema de saúde e a aplicação de medidas preventivas — desempenharam um papel fundamental na evolução da crise sanitária.

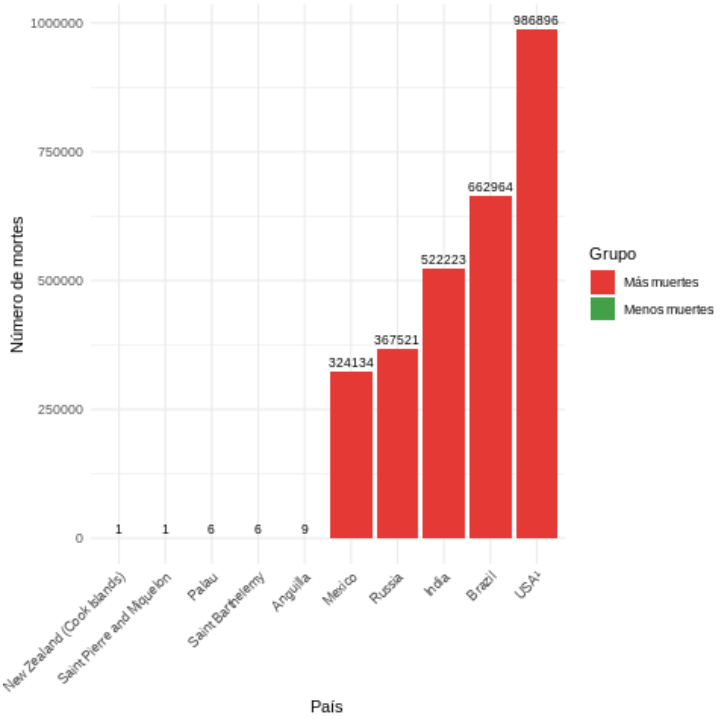


Figura 2: Países com mais e menos mortes por COVID-19

A Figura 3 compara o número absoluto de mortes com a taxa de mortalidade. Observa-se que países como Reunion e Yemen apresentam uma taxa de mortalidade elevada, embora não estejam entre aqueles que registram o maior número total de óbitos. Por outro lado, Estados Unidos apresentam o maior número de mortes, mas não exibem uma taxa de mortalidade elevada.

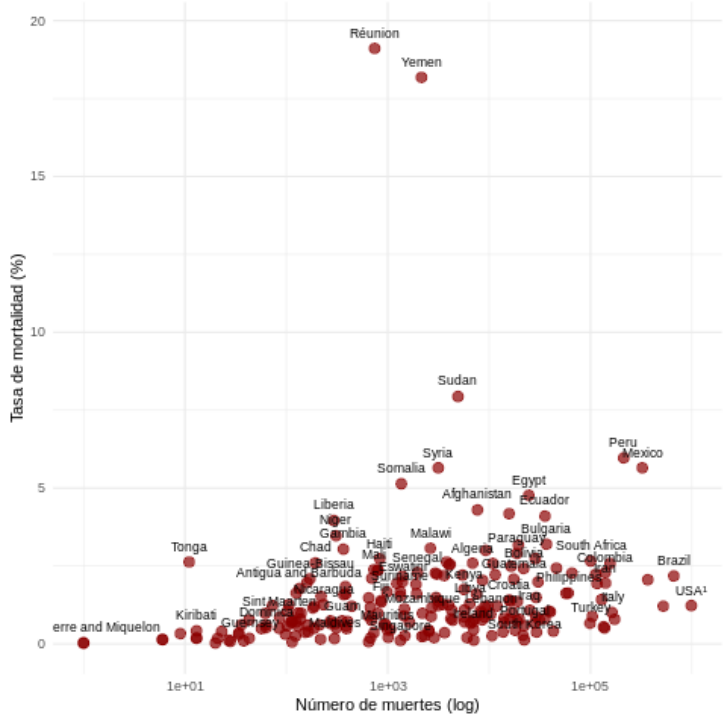


Figura 3: Comparação entre mortes absolutas e taxa de mortalidade por COVID-19

A Figura 6 analisa o aumento diário de casos de COVID-19. Observa-se que a Alemanha apresentou o maior crescimento diário, seguida pela Coreia, enquanto a Tailândia registrou o menor incremento.

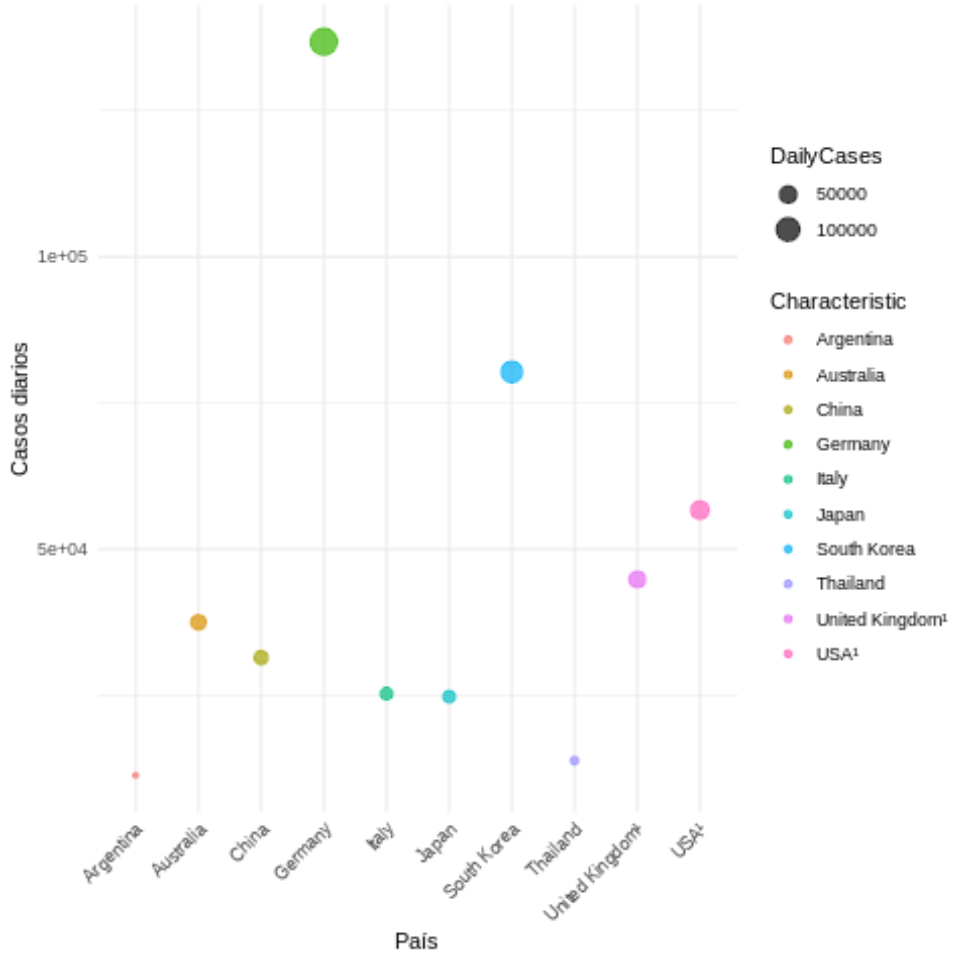


Figura 4: Casos diários confirmados de COVID-19 por país

A Figura 5 mostra a relação entre os casos confirmados e as mortes por COVID-19 em diferentes países, com ambos os eixos representados em escala logarítmica. Observa-se uma tendência positiva clara: à medida que aumenta o número de casos confirmados, também cresce o número de mortes. No entanto, a dispersão dos pontos indica que alguns países apresentam mais mortes do que o esperado em relação ao número de casos, enquanto outros registram menos. O uso da escala logarítmica permite visualizar adequadamente tanto os países com poucos casos quanto aqueles com números extremamente elevados.

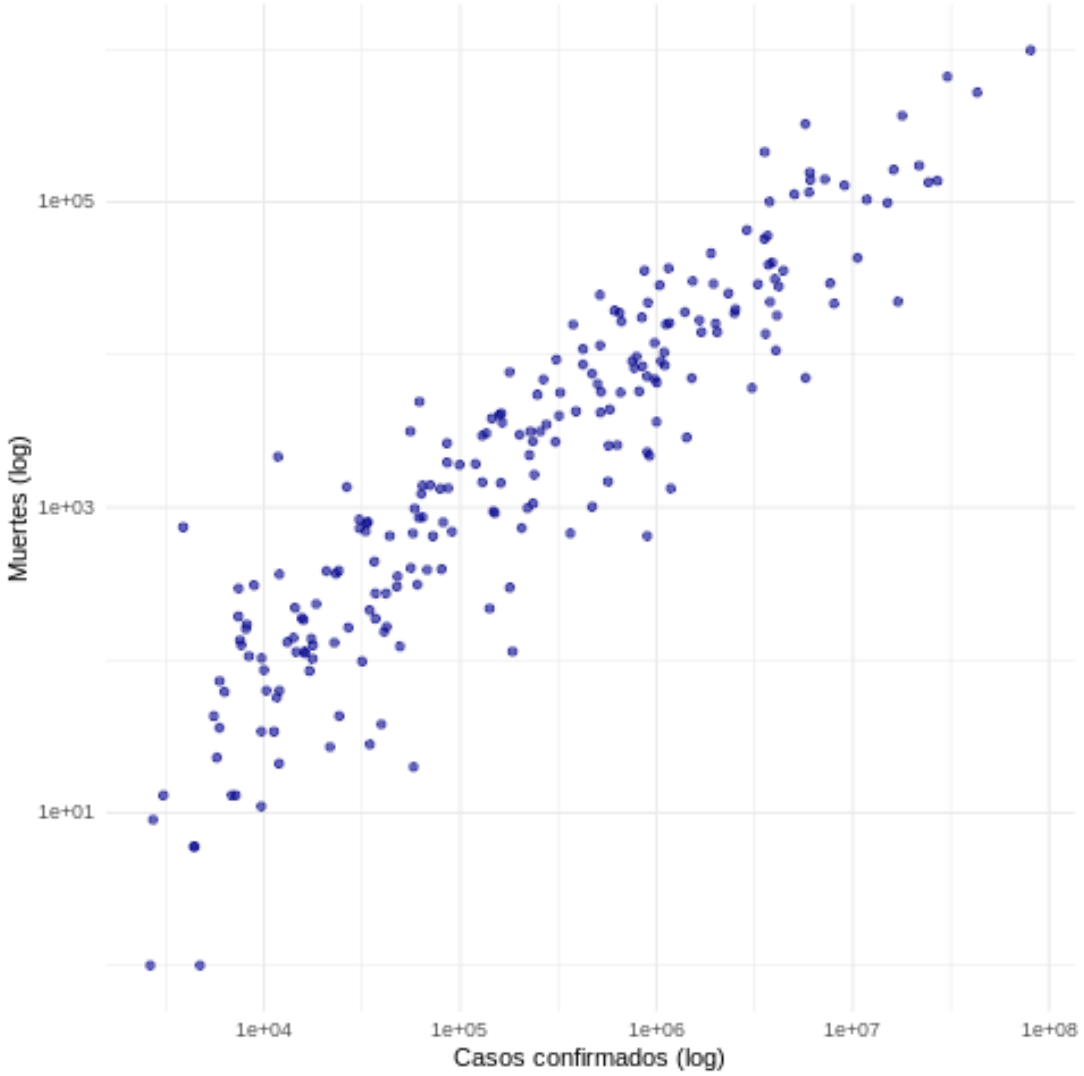


Figura 5: Casos confirmados de COVID-19 vs. mortes por COVID-19

A Figura 6 mostra a correlação existente entre as variáveis, destacando-se que Confirmed e Deaths apresentam uma correlação próxima de 1, o que indica que os países com maior número de casos confirmados tendem a registrar também um número proporcionalmente elevado de mortes. Em seguida, observa-se uma correlação igualmente alta entre Last7Days e DailyCases, o que é esperado, uma vez que Last7Days representa a soma dos casos diários registrados nos últimos sete dias.

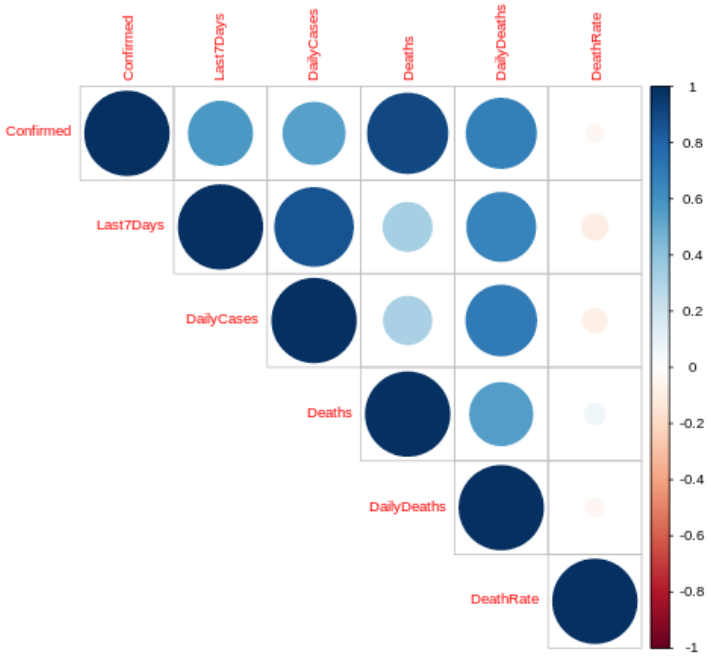


Figura 6: Correlação de variáveis

# 5 Conclusão

Este trabalho evidenciou a utilidade do web scraping como método para coletar informações atualizadas de forma automatizada a partir de plataformas estatísticas online, neste caso, o site Statista. A análise exploratória permitiu identificar diferenças significativas entre os países quanto ao impacto da COVID-19, refletidas no número de casos, óbitos e taxas de mortalidade. Além disso, ficou evidente a importância do web scraping, aliado à análise exploratória de dados, como ferramenta valiosa para a detecção de padrões globais.

# Referências

Ferreira, J. (2025). Material de apoyo de la asignatura: Tópicos especiais em estatística computacional. Disponible en: `https://jodavid.github.io/computationalstatisticstopics/`.

Statista (2025). Coronavirus death rates worldwide. Disponible en: `https://www.statista.com/statistics/1105914/coronavirus-death-rates-worldwide/`.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.