

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO
PAULO

DEPARTAMENTO DE INFORMÁTICA E TURISMO
TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

HENRIQUE ORSONI VAZ
JHONATAN MARCELO DE SOUSA SANTOS
JÚLIA INOSCÊNCIA OLIVEIRA DOS SANTOS
LENNON DE OLIVEIRA MANHOLETO

PROJETO - ANÁLISE EXPLORATÓRIA DE DADOS

São Paulo
2025

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO
PAULO

DEPARTAMENTO DE INFORMÁTICA E TURISMO
TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

HENRIQUE ORSONI VAZ
JHONATAN MARCELO DE SOUSA SANTOS
JÚLIA INOSCÊNCIA OLIVEIRA DOS SANTOS
LENNON DE OLIVEIRA MANHOLETO

PROJETO - ANÁLISE EXPLORATÓRIA DE DADOS

Trabalho apresentado ao curso de
Análise e Desenvolvimento de
Sistemas, para a disciplina Estatística
e Probabilidade.

São Paulo
2025

INTRODUÇÃO

A pandemia de COVID-19 destacou a importância da análise de dados para orientar políticas públicas e compreender disparidades regionais. Este relatório analisa dados de **mortes por milhão de habitantes** e **hospitalizações**, com foco em identificar padrões geográficos e temporais, além de discutir limitações metodológicas, como a subnotificação em regiões de baixa renda.

METODOLOGIA

1. Dados Utilizados

- **Dados de Mortes:** O conjunto de dados *total_deaths_per_million.csv* contém informações sobre o total de mortes por milhão de habitantes para diversos países e continentes, ao longo do tempo. São dados diários de países e continentes.
- **Dados de Hospitalizações:** O conjunto de dados "covid-hospitalizations.csv" contém informações sobre hospitalizações relacionadas à COVID-19, incluindo ocupação de leitos de UTI, para diversos países e ao longo do tempo. Dados de ocupação de UTIs, com lacunas significativas em países como por exemplo: Brasil, Índia, Geórgia e Montenegro.

2. Tratamento de Dados

- **Tradução e Padronização:**
 - Nomes de países e continentes foram traduzidos para português (ex: "North MacedOnia" → "Macedônia do Norte").
- **Valores Faltantes:**
 - **Mortes:** Dados completos para todos os países/continentes selecionados.
 - **Hospitalizações:** Valores ausentes foram substituídos pela mediana da série temporal correspondente.

3. Técnicas de Análise Descritiva

Para descrever os dados, foram utilizadas as seguintes técnicas:

- **Estatísticas Descritivas:** Cálculo de medidas de centralidade (média, mediana) e dispersão (desvio padrão), além de valores mínimo e máximo, para resumir a distribuição dos dados.

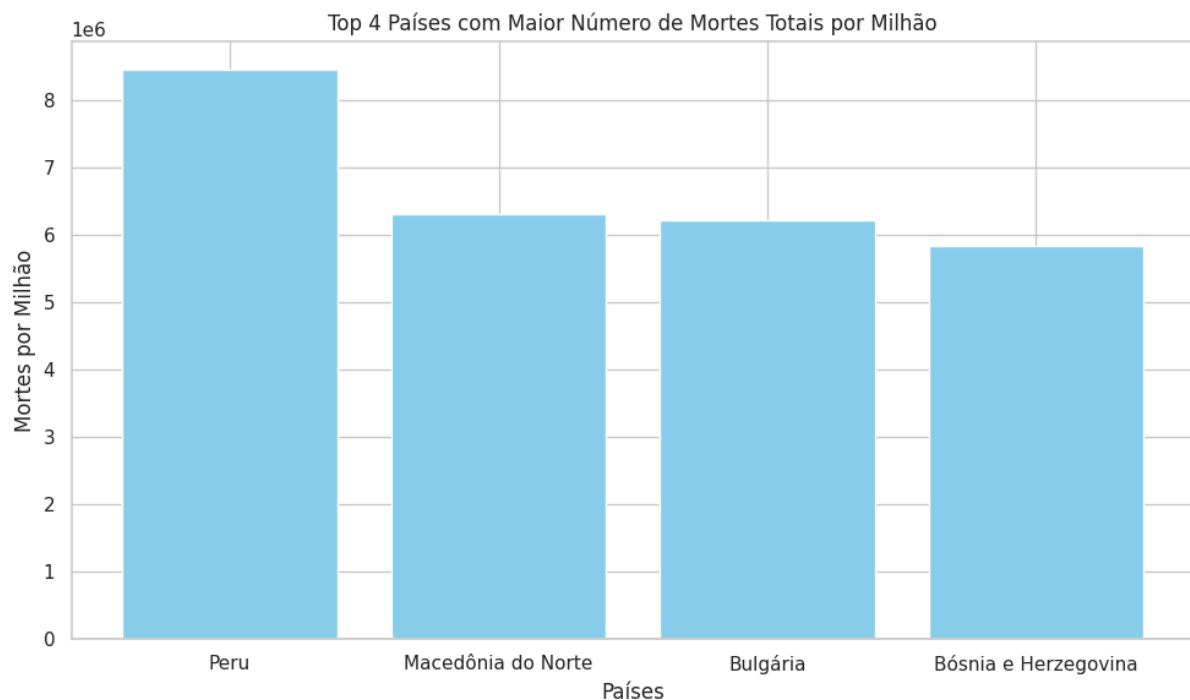
- **Visualizações Gráficas:** Criação de gráficos de barras, linhas e boxplots para visualizar os padrões e tendências nos dados.
- **Tabelas de Frequência:** Criação de tabelas para apresentar a distribuição de frequências de variáveis categóricas.
- **Análise Temporal:** Investigação da evolução dos dados ao longo do tempo, por meio de gráficos de linhas e análise de séries temporais.
- **Correlação:** Cálculo da correlação entre variáveis para identificar possíveis relações entre elas.
- **Probabilidade:** Cálculo de probabilidades para estimar a chance de ocorrência de eventos específicos.

ANÁLISES E RESULTADOS

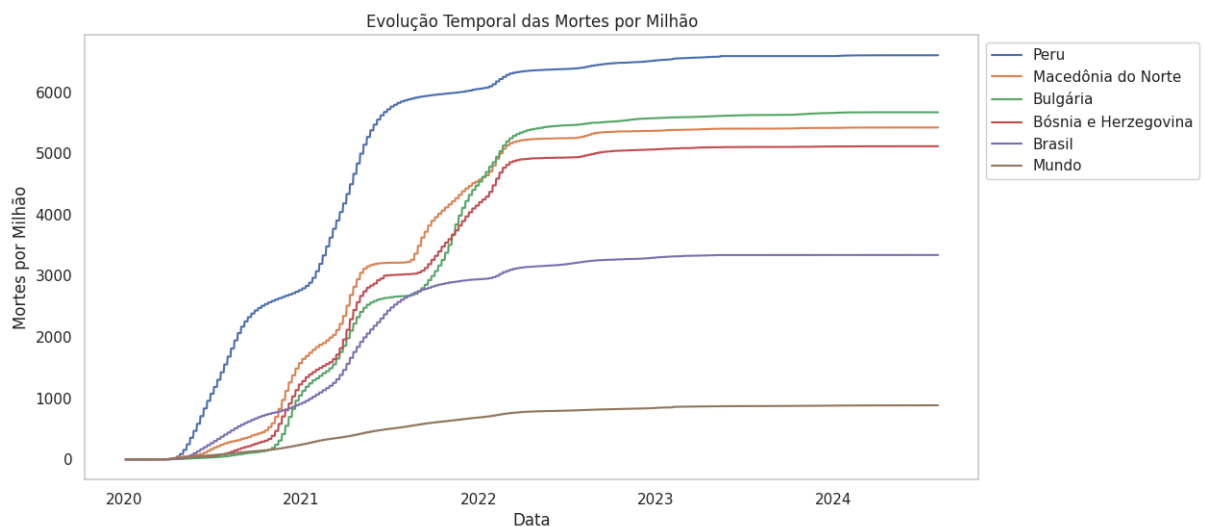
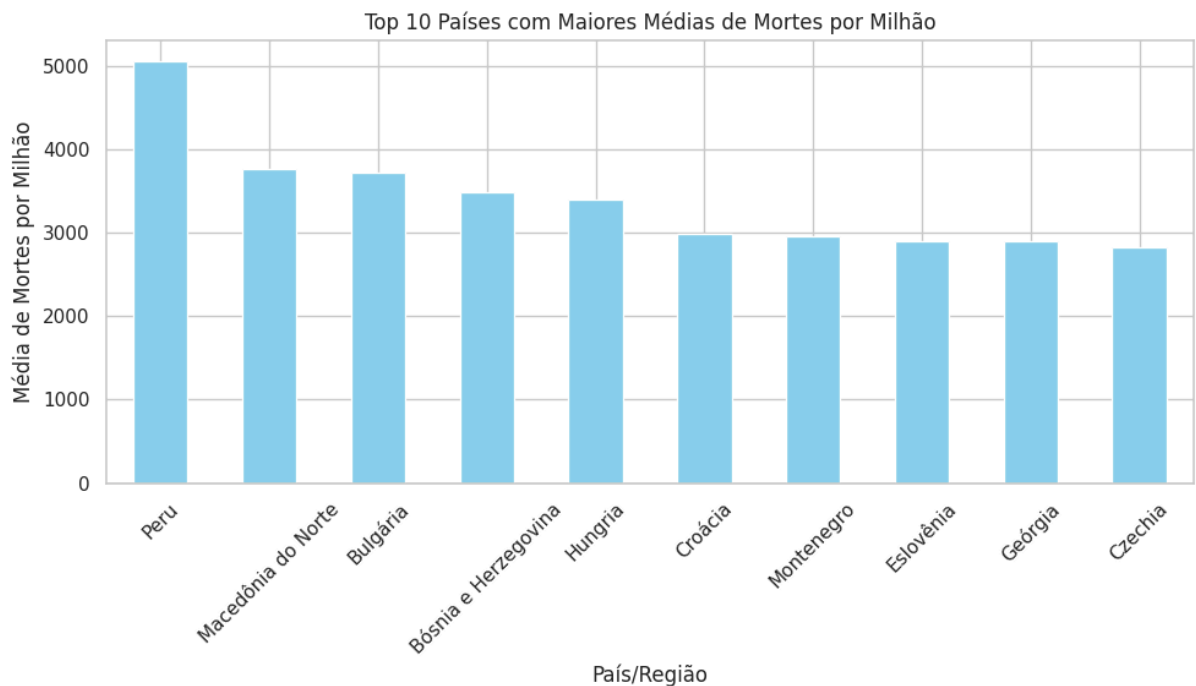
1. Mortes por COVID-19

- **Total de Mortes por Milhão:**

Os quatro países com o maior número total de mortes por milhão foram Peru, Macedônia do Norte, Bulgária e Bósnia e Herzegovina. **Essa informação foi obtida através da soma das mortes registradas ao longo do tempo em cada país e dividindo pelo número de habitantes, expresso por milhão.**

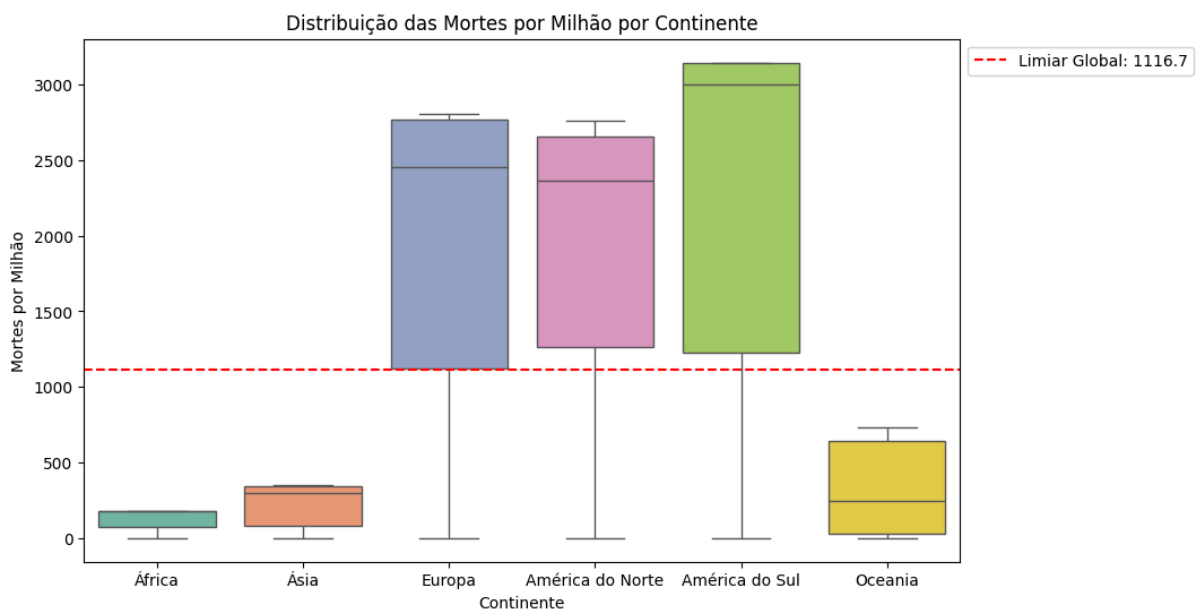


- Foi observada uma evolução temporal das mortes por milhão, com tendências distintas entre os países analisados. **Para visualizar essa evolução, foram criados gráficos de linhas que mostram a variação do número de mortes ao longo do tempo para cada país.**
- Os países com as maiores médias de mortes por milhão foram Peru, Bulgária, Bósnia e Herzegovina e Hungria. **A média de mortes por milhão para cada país foi calculada considerando todo o período de tempo disponível nos dados.**

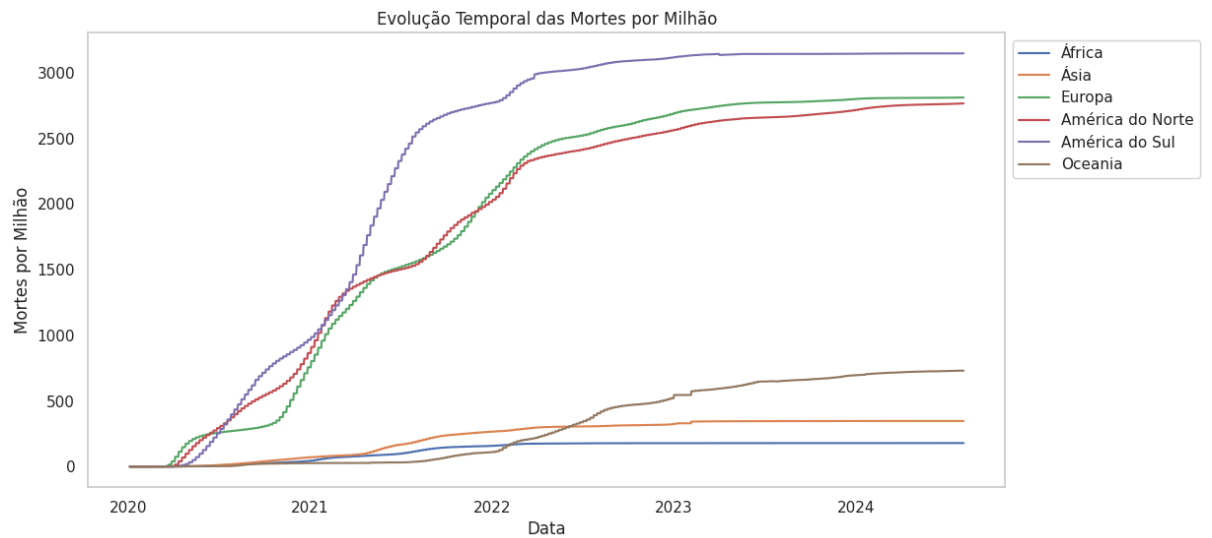


- **Mortes por Continente:**

- O boxplot das mortes por milhão por continente evidenciou a variabilidade entre as regiões, com a América do Sul apresentando valores mais elevados. **O boxplot permitiu visualizar a distribuição dos dados, incluindo a mediana, os quartis e a presença de outliers.**



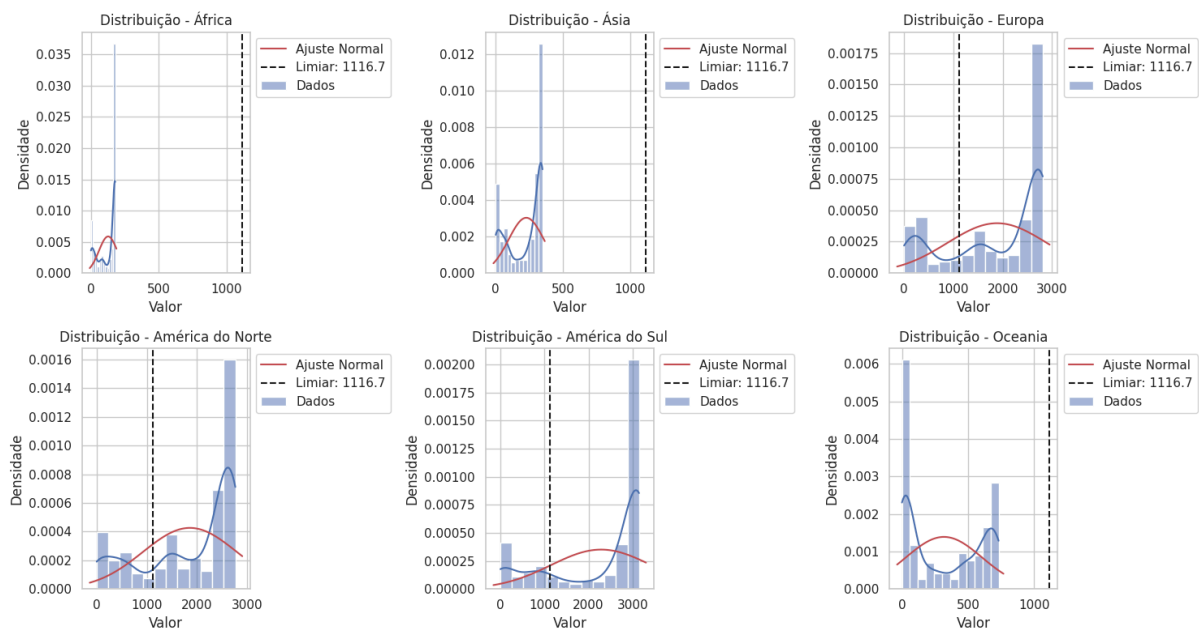
- A América do Sul apresentou o maior número total de mortes por milhão, seguida da Europa e América do Norte. **Essa informação foi obtida agregando os dados de mortes por país, de acordo com o continente ao qual pertencem.**
- A análise temporal das mortes por milhão por continente revelou tendências específicas para cada região. **Gráficos de linhas foram utilizados para visualizar a variação do número de mortes por continente ao longo do tempo.**



- A América do Sul teve a maior média de mortes por milhão entre os continentes. **A média de mortes por milhão para cada continente foi calculada considerando todo o período de tempo disponível nos dados.**

| Continente | Média de Mortes por Milhão |
|------------------|----------------------------|
| África | 126.781959 |
| América do Norte | 1858.901126 |
| América do Sul | 2285.347781 |
| Ásia | 226.849660 |
| Europa | 1884.892241 |
| Oceania | 317.368711 |

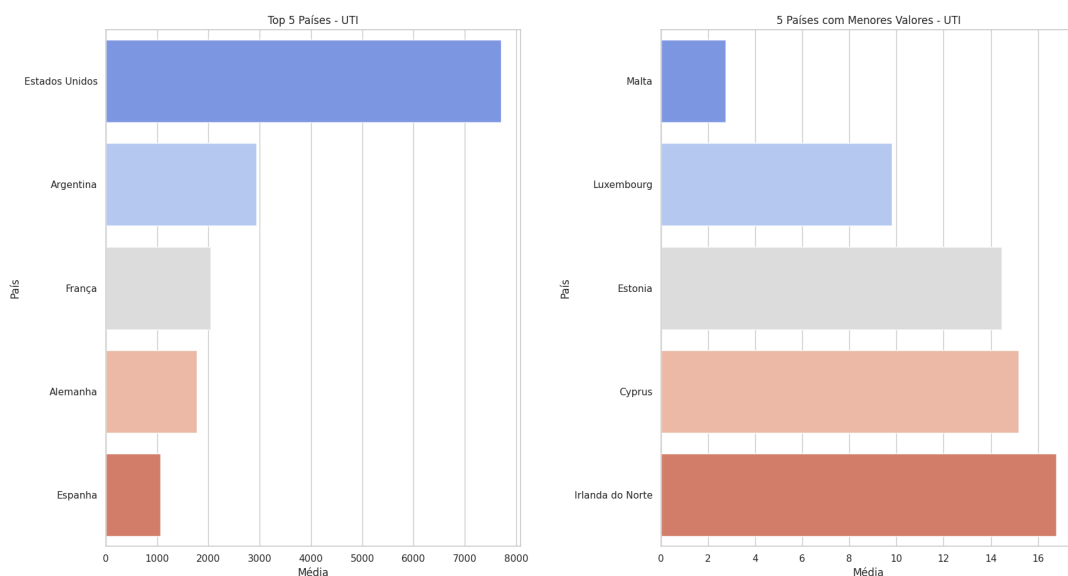
- A análise da distribuição de continente pela limiar global indicou a probabilidade de cada continente ultrapassar o limiar médio de mortes. **Essa análise foi realizada utilizando métodos probabilísticos, considerando a distribuição dos dados.**



2. Hospitalizações por COVID-19

● Evolução das Hospitalizações:

- A análise da evolução das hospitalizações ao longo dos anos, para os cinco países com os maiores números, demonstrou as tendências médias. **Gráficos de linhas foram utilizados para visualizar a variação do número médio de hospitalizações para cada país.**

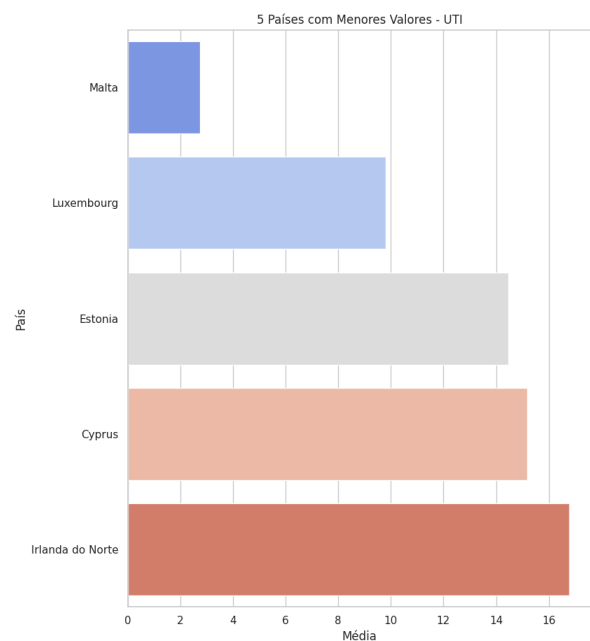
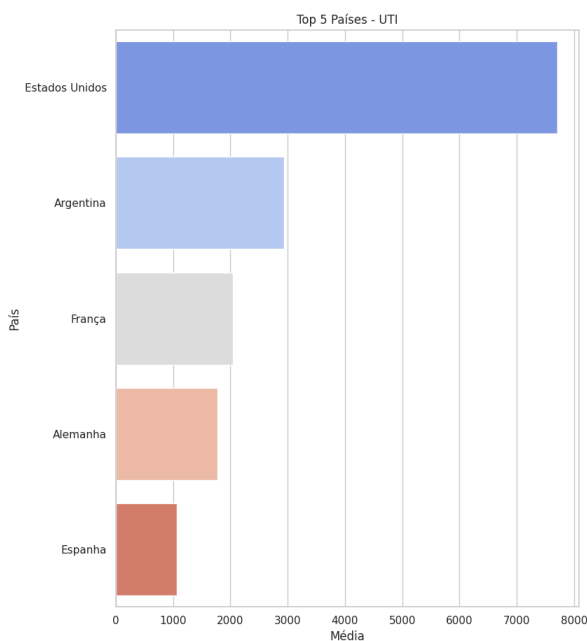


- **Ocupação de UTI:**

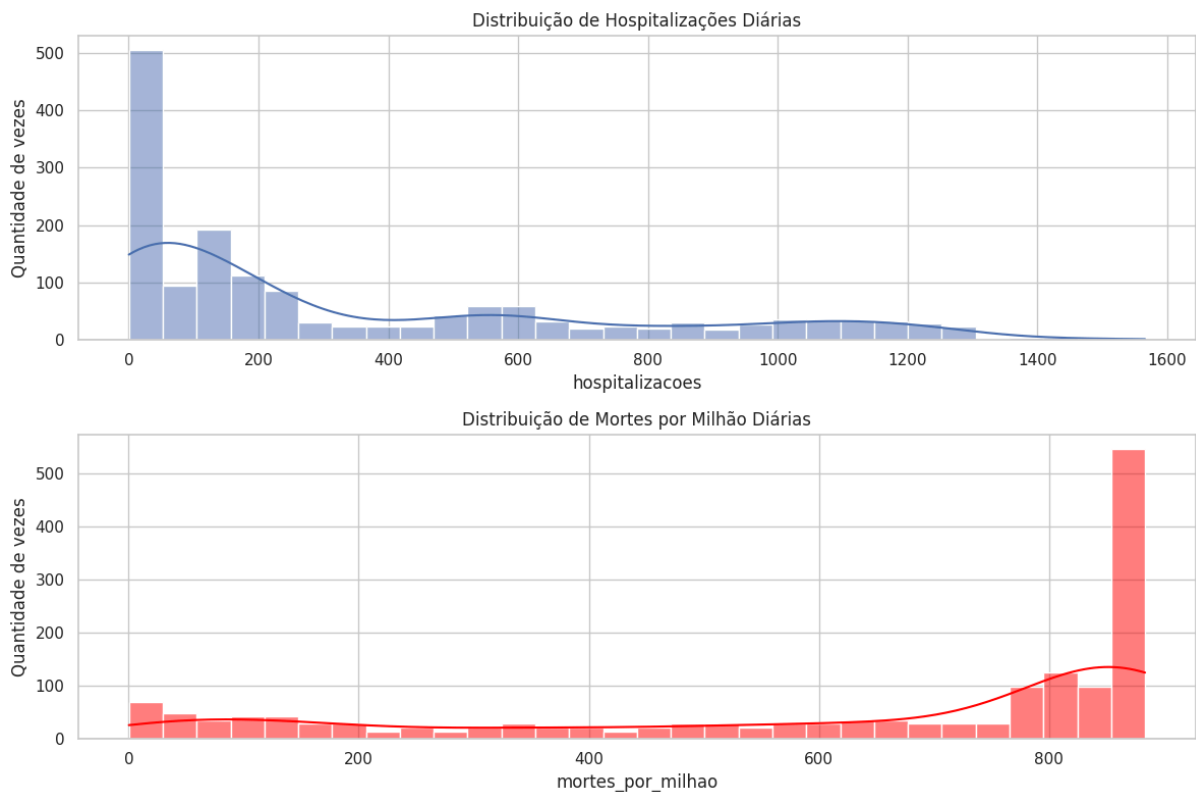
- Foram analisadas as estatísticas descritivas da ocupação diária de UTI e da ocupação diária de UTI por milhão. **Foram calculadas medidas como média, mediana, desvio padrão, mínimo e máximo para descrever a distribuição dos dados.**

| UTI | |
|---------------|--------------|
| Contagem | 39116.000000 |
| Média | 660.971418 |
| Desvio Padrão | 2139.615532 |
| Mínimo | 0.000000 |
| 25% | 21.000000 |
| 50% | 90.000000 |
| 75% | 413.000000 |
| Máximo | 28891.000000 |

- Os países com maior e menor ocupação de UTI foram identificados. As distribuições da ocupação de UTI e da ocupação de UTI por milhão foram visualizadas. **Essas informações foram obtidas comparando os valores das estatísticas descritivas para cada país.**



- **Correlação entre Hospitalizações e Mortes:**



Previamente, foi realizado um filtro entre as tabelas, para que fossem considerados apenas os dados referentes aos países que constem em ambas as tabelas, para que os dados representem o cenário mais realista. Dessa forma, considerando o indicador "Daily ICU occupancy per million", referente à Hospitalizações por Covid, e o total de mortes por milhão, pode-se obter as seguintes análises:

- Foi calculada a correlação entre hospitalizações e mortes por milhão.

O coeficiente de correlação de Pearson foi utilizado para medir a força e a direção da relação linear entre as variáveis.

- **Correlação entre hospitalizações e mortes por milhão: -0.51**

Esse valor indica uma correlação negativa moderada entre o número de hospitalizações e o número de mortes por milhão. Isso sugere que, conforme o número de hospitalizações aumenta, o número de mortes por milhão tende a diminuir, e

vice-versa. É importante interpretar essa correlação com cautela, pois pode haver outros fatores em jogo, como a capacidade do sistema de saúde, tratamento disponível e características da população, que influenciam essa relação.

- Um teste qui-quadrado de independência foi realizado para verificar a relação estatística entre as variáveis. **Esse teste avaliou se existe uma associação significativa entre as variáveis.**

- **Teste qui-quadrado de independência: p-valor = 0.00648**

O p-valor, menor que 0.05, indica uma relação estatisticamente significativa entre as variáveis hospitalizações e mortes por milhão. Rejeitamos, portanto, a hipótese nula de que não há associação entre essas duas variáveis. Isso sugere que há alguma dependência entre as hospitalizações e as mortes, e que a relação observada é improvável de ser devida ao acaso.

- Probabilidade de um pico de hospitalização resultar em um pico de mortes: 0.00%
 - Nos dados analisados, não houve casos em que um pico de hospitalizações (definido como acima do percentil 75) coincidiu com um pico de mortes por milhão (também definido como acima do percentil 75). Isso sugere que, nas situações analisadas, quando as hospitalizações estavam em alta, as mortes não seguiram o mesmo padrão. Esse resultado pode ser influenciado por diversos fatores, como medidas de controle da pandemia, disponibilidade de tratamentos e a própria evolução da doença.

Conclusão

A análise descritiva dos dados de COVID-19 revelou padrões e tendências importantes relacionados a mortes e hospitalizações, utilizando os métodos e dados descritos na metodologia. Observou-se uma variação significativa entre países e continentes em relação ao número total de mortes por milhão e à ocupação de UTI. A análise temporal demonstrou a evolução da pandemia ao longo do tempo, com tendências específicas para cada região. A correlação entre hospitalizações e mortes indicou uma possível relação entre as variáveis. A probabilidade de um pico de hospitalizações levar a um pico de mortes foi estimada, fornecendo insights sobre a dinâmica da pandemia. É fundamental destacar que este relatório se limita à descrição dos dados, sem realizar inferências causais ou preditivas. Para aprofundar a análise e obter conclusões mais robustas, recomenda-se a realização de estudos adicionais, com métodos estatísticos mais avançados.