1. Tem-se uma base de dados sobre tratamentos de saúde, o problema é que querem relacionar as condições médicas e tempo de recuperação à idade, gênero, tipo sanguíneo e qual medicamento foi ministrado. O objetivo final é conseguir saber informações como: Qual a faixa etária está mais sujeita a X doença, ou qual medicamento trata mais rápido Y doença.
2. Os dados foram obtidos através da plataforma do Kaggle (<https://www.kaggle.com/datasets/prasad22/healthcare-dataset/code>) em formato CSV. Os dados são fictícios.
3. Para utilizar os dados, primeiramente foi removido os nomes das pessoas, já que não serão relevantes para as análises. Após isso, linhas com valores inválidos, datas faltando ou outras informações incompletas, que não permitem uma análise completa, foram excluídas.
4. Visualização dos dados:

**import** pandas **as** pd  
  
**data** = pd.read\_csv('**healthcare\_dataset**.**csv'**)  
  
# Estatísticas numéricas  
**print**(**data**.describe())    
**import** matplotlib.pyplot **as** plt  
**import** seaborn **as** sns  
  
**numeric\_data** = **data**.select\_dtypes(**include**=['**number'**])  
  
# Histograma  
**data**['Age'].hist(**bins**=30)  
**plt**.show()  
  
# Boxplot  
**sns**.boxplot(x=**data**['Age'])  
**plt**.show()  
  
# Correlação entre apenas dados numéricos  
**sns**.heatmap(numeric\_data.corr(), annot=True)    
**plt**.show()

1. Para ver média, mínima, máxima, 25%, 50%, 75% e desvio padrão da idade:

**import** pandas **as** pd  
  
**data** = pd.read\_csv('**healthcare\_dataset**.**csv'**)  
  
# Estatísticas numéricas  
**print**(**data**.describe())

1. Feito, arquivo no github.
2. Modelagem de predição não aplicável.
3. Como os dados são fictícios e foram gerados aleatoriamente, sem nenhum critério, a análise deles de tendencias e hipóteses são irrelevantes. Mas aqui estão algumas hipóteses que surgiram:
   1. Será que as pessoas mais velhas ficam mais tempo no hospital que pessoas mais novas? Apesar da aleatoriedade dos dados, pessoas de 13 anos ficam, em média, mais tempo no hospital quando comparados com outras faixas etárias.

Não consegui identificar nenhuma outra tendencia ou comprovar outra hipótese devido a distribuição normalizada dos dados (por serem gerados aleatoriamente).

1. Como dito previamente, é bem complicado tomar decisões em dados completamente aleatórios já que apresentam pouca ou nenhuma tendencia além de normalizados. Logo não tem como tomar nenhuma decisão específica baseado na análise desses dados.
2. A única parte do processo inteiro que foi um pouco desanimadora foi perceber que, após a análise, os dados não apresentam nenhuma tendencia e, no fim das contas, nenhum significado. Com isso, a melhoria que eu faria foi numa escolha melhor da base de dados, já que o meu objetivo final era a análise dos mesmos.

Análises Referentes ao item 5**:** Arquivo do colab disponível no github do projeto.